Principles of time series

New and intuitive introduction to working with time series

Kamil Kovar[1]
Moody's Analytics

May 1, 2025

[1]Email: kovarkamil@gmail.com.

Chapter 8

Model building

In previous chapters we have discussed various types of time series models, both univariate and multivariate, focusing on interpretation and behavior of the models. While this provides us with knowledge to be informed model *users*, it is not sufficient knowledge for model *builders*. Rather it only forms the background knowledge necessary for model builders, and we will need to add to this knowledge if we want to become model builders.

It is true that we have partially addressed the perspective of model builders by always dedicating a section to discussion of **model selection: the task of selecting particular type of model from given model class**. This provided the bare minimum knowledge for somebody who is tasked with creating time series morel, not just using it. However, rather than making the discussion of model selection the final word on model creation process, as is customary, this chapter complements the discussion with discussion of **model building: the *iterative* and *interactive* process of creating times series models based on information about (quality of) their behavior**.[1]

What do we mean by iterative and interactive process of creating times series models? **Iterative process** signifies that the models are created in incremental steps, with each model specification formed by improving the previous model specification. This is in a sense inspired by modern machine-learning algorithms, which are also built around incremental improvement. The key idea is that such process of perfecting model is more likely to yield a good model, rather than a process which aims to determine good model more directly.

[1] While this chapter is written with model builders in mind, most of the discussion is valuable also to model users: knowing how to evaluate model from perspective its behavior is important if model users need to know how much confidence they should place in given model, as is often the case.

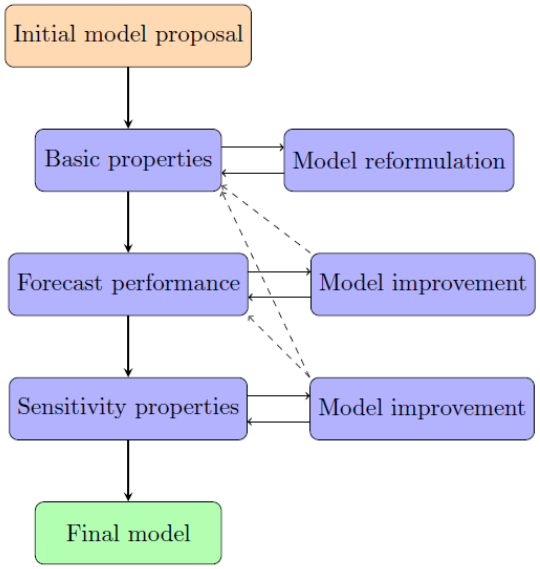
Unlike machine-learning, model building is not relying on computer algorithm to make suggestions for improving the model, but rather on human mind. This is what it meant by **interactive process**. The key idea here is that for some types of modelling problems a human mind is more capable of coming up with suggestions for model improvements as it can rely on outside sources information. For example, in macroeconomic times series modeling one can often leverage some underlying macroeconomic theory. Similarly, human mind might be more capable in learning from bad model performance in particular parts of the sample, as it can link it to real-world events. This is especially true for modelling problems with small data samples. Another reason why human mind might be better is because unlike algorithm it can also leverage information that is not quantifiable, and can create mental constructs based on the observed model performance. Ultimately, what is at heart of the interactive model process is learning on the part of the model builder about the data and resulting model.

The first section of this chapter provides brief introduction to (automatic) model selection approach, together with discussion of its shortcomings. This will motivate our exploration of our alternative approach based on model building. Ultimately, the difference between model selection and model building is matter of role played by the model creator in defining the possible model universe: Is their role passive, focused only on choosing particular class of models? Or is their role more active, deciding which model improvements to explore one at a time? As we will argue in this chapter, in many situations more active approach leads to better results than the passive approach commonly taken by many model creators.

The key to the model building process outlined above is the information used. Correspondingly, this chapter will effectively be an answer to question '**Base on what information are models evaluated and improvements suggested?**'. The discussion of this is divided into three parts. First, we look at information contained in standard **estimation output** such as coefficient estimates and t-statistics, together with R-squared and other model statistics. While this information does not need introduction - advanced modelers might wish to skip this section - it is useful to discuss each item from the perspective of its value for model building, given that such perspective is rarely provided as part of time series courses. Moreover, the shortcomings of the standard regression information will provide us blackboard on which we will discuss the more sophisticated information.

The other two categories of information about time series models are information about **forecast performance** of the models and **sensitivity to shocks**. The former should be at the heart of any model building endeavor, since whether model does or does not work in forecasting during observed

Figure 8.1: Model building workflow



historical sample is a key aspect indicating whether the model will be useful for forecasting future values.[2] Meanwhile, it is often suitable to complement forecasting performance by information about sensitivity to shocks. This is especially true when primary goal of the time series model under consideration is scenario analysis, since proper scenario behavior relies on correct sensitivity to shocks which are at the heart of the scenario.[3]

The last two sections then leverage the knowledge we will gain throughout this chapter to outline model building philosophy and illustrated it on extended application. The former is really just outlining the steps in the iterative and interactive modeling process and linking them to the individual information discussed in previous section. As a preview, Figure H.1 shows the outline of the model building process, which separates the process into several steps that align with the structure of this chapter. The application then serves as an detailed recapitulation of the lessons learnt.

Before proceeding it is worth noting that at the heart of the fourth and fifth section of this chapter is the large number of figures included, since the model building process relies heavily on graphical

[2]Of course, if forecasting is the ultimate goal, then forecasting performance is the obvious yardstick. However, forecasting performance should be relevant information for model evaluation even if forecasting is not the primary objective, since forecasting performance can reveal aspects along which the model is not valid even for non-forecasting purposes. Afterall, model that is not good in forecasting given variable can hardly be considered correct representation of the data generating process of that variable. [XXX CITATION HERE ON PAPER ABOUT PREDICTIONS]

[3]Analysis of sensitivity to shocks is valuable even if scenario analysis is not the primary goal of given model. This is obvious once one realizes that analysis of shock responses is one of the primary tools of evaluating econometric validity of VAR models. In some sense the discussion in the section of this topic is just bringing this standard practice from VAR models to single-equation models. As discussed in the introduction, the absence of the shock-response analysis from introduction to single-equation time series models is hard to understand in view of modern approach to teaching VAR models.

information. This means that if one is to follow the model building process outlined here, one needs to have these figures readily available. While some of these figures are readily available irrespective of the statistical program one uses, most of them are not. Nevertheless, in the specific case of EViews used in this textbook there is user add-in called SpecEval that was developed to facilitate the model building process outlined here. Correspondingly majority of the figures displayed below were created by the SpecEval add-in.

8.1 Automatic model selection and its limits

Before we dive into the discussion of model building, it is useful to clarify what do we mean by model selection and why in many situations it might not be sufficient approach to time series modelling.

How do we understand model selection? Model selection is the process of selecting chosen model from list of considered candidate models. While model selection can be performed "manually", the advent of fast computer processing led to shift towards automatic model selection, under which the selection is performed (almost completely) by algorithm. We will therefore focus on *automatic* model selection. One example of such automatic model selection is the best subset model selection algorithm. Under this approach the modeler specifies list of all acceptable regressors and then runs algorithm that estimates all the possible models. In this case the universe of models is defined by all possible combinations of the acceptable regressors. For example, in case of ARMA models it is all possible combinations of AR and MA components. Once all models are estimated then the algorithm evaluates each one of them based on chosen model evaluation criterion and selects the best model from according to this criterion.

The key aspect of automatic model selection is that the role of modeler is focused on selecting model from list of candidates, rather than on creating the list of candidate models. Rather than spending time on determining the list of candidate models, the list is typically defined by complete universe of models belonging to particular model class, or well defined subset of this and the modeler just decides which class of models to focus on.[4]

[4]Notice that this means that the most important choice happens in some sense before the modelling and is often done not fully consciously: it is the choice of class of models that will be considered.

8.1.1 Limits of model selection

Dimensionality. The first drawback of the automatic model selection – indeed the drawback that mainly occupied the proponents of this approach - is the curse of dimensionality. The number of potential models grows very fast with the possible dimensions of the model. For example, when one considers simple static regressions with several independent variables then the number of possible models grows exponentially: for two variables you have 3 possible models, for 3 variables it is 9, for 4 it is 17 and so on. In some applications this exponential growth can become punishing even with the computational power we have at hand.

Related to this problem is the issue of model dimensionality versus the data dimensionality: sometimes some plausible models are impossible to estimate given the amount of data. For example, when you have more candidate variables than data points for dependent variables than model cannot be estimated. In both cases the key idea of model selection - estimating large number of models and selecting best one - directly leads us into a blind alley. Nevertheless, well-known solutions to this problem exist, even if those solutions sometimes create problems of their own.[5]

Narrowness of selection criterion. The second drawback that received large amount of attention is the choice of model selection criterion. Model selection criterion is the characteristic based on which we evaluate the various models and ultimately select the best one. Of course, there are many postulated selection criterions, and which one is chosen might depend on particular situation. That said, in almost all the situation one ultimately relies on some measure of model fit. This is quite logical, as good model fit is a pre-requisite of good forecasting model: if model is not able to explain observed data, then expecting it to be able to predict future data is hopeless.

Which measure of model fit should be used? The departing point is the realization that the two standard measures of model fit – R-squared and log-likelihood – are monotonic functions of number of parameters, which means that model with more components will always have higher R-squared/lower log-likelihood (more on this later). This then means that if one would use these as selection criterion then one would always choose the most complex model of all the possible models. To address this issue, one needs to introduce penalization for number of model components/parameters. Different selection criteria then differ in the exact form of penalty, which then reflects the initial considerations of the particular criterion. These wary from ad-hoc parameter penalization of adjusted R-squared to

[5]For example, one solution to these drawbacks is using stepwise regression instead of the best subset selection. In stepwise regression one does not estimate all candidate models but rather starts from initial model and add or subtracts regressors from this model based on pre-defined rules. However, these pre-defined rules often create problems of their own. XXX citation

theoretically derived penalizations of Akaike and Schwarz information criteria.[6]

While there can be a discussion about which measure of model fit is the most appropriate, this discussion is akin to discussing which ice cream flavor is the best: while important, it can easily distract from the fact that there are other important considerations than just model fit, like there are other types of desert than just ice cream. Correspondingly, it is important to realize what all the model fit criteria are. They are just a single number summarizing the fit of the model to the data.

There are two problems with basing all modelling on a single summary of model fit. First, a single number summarizing the overall performance can hide important heterogeneity across the whole sample. Different models can perform better or worse in different parts of sample; a typical example from macroeconomics are models that perform well during or outside of recessions. Not only is summary statistic not informative about such heterogeneity, but it can be downright misleading if different parts of sample have different importance from the perspective of model. Continuing with our example from macroeconomics, the fact that recessions are relatively rare means that bad forecast performance during recessions can be easily hidden by good performance outside of recessions; moreover, good forecasts during recessions are typically much more valuable than good forecasts outside of recessions.[7]

Second, the model fit provides information only about particular type of model behavior. To see this, it is just enough to realize that model fit can be thought of as forecast performance in terms one-step-ahead in-sample forecasts (more on this later on). But one-step-ahead forecasts are in many situations not the only, or even the primary forecasts of interest. Therefore, focusing on model fit is like drafting a NBA player based on how good they are at shooting without considering other aspects of his game skills: while player with terrible shooting is probably not a good player, player that is good only in shooting and nothing else might not be your primary drafting target, especially if you want somebody to improve your defense.[8]

One way to address the second issue is to use cross-validation approach: in this approach the selection criterion is actual forecast performance in hold-out sample, say last 20% of observations. The use of actual forecasts, rather than just model fit, means that one can tailor the selection criterion to the specific model behavior that is of interest, say quality of forecasts at 8-steps ahead. However, note that while this address second issue - the focus on potentially unimportant aspect of model

[6] XXX Citation for more information.

[7] Of course, one could try to address this by particular schemes which would still rely on summary statistics in some form – by for example using a weighted average of model performance metrics, where bigger weight would be given to particular parts of the sample – but such approach is cumbersome and rarely taken given that the heterogeneity of the sample is issue not considered by the proponents of automatic model selection.

[8] Enough with the metaphors!

behavior – it actually makes things worse in terms of the first issue: evaluating model based on forecasting performance in particular part of the sample is more extreme version of ignoring potential heterogeneity in the data. [9]

The issue of non-representativeness of summary of model performance is just one of the big drawbacks of focus of model selection on model fit as the selection criterion. Another related one is the absence of the analysis of, or even knowledge of, sensitivity to shocks when automatic model selection is used. As mentioned earlier, sensitivity to shocks can be important, or even most important model aspect in many applications. However, sensitivity to shocks never plays any role in automatic model selection. One reason for this is that sensitivity to good shocks is hard to measure quantitatively. This then prevents it from being used in algorithmic approaches to modelling, and one rather has to rely on more hands-on model building approach, something we return to later.

Unexplored and unexplained models. The above discussion of limits of automatic model selection focused on the fact that when modelers use automatic model selection, they are likely to focus on a too narrow information set when selecting between different models. This can easily lead to selection of model that is not optimal for given objective. However, even bigger limitation of automatic model selection lies in potentially excluding better models a-priory, rather than just not choosing the best model from the list of considered models.

By necessity the automatic model selection has to be performed over relatively limited subsets of models, since the model space can explode relatively quickly. The limitations typically take form of selecting only some classes of models, limiting the set of independent variables and/or limiting the transformation of dependent and independent variables. Even worse, these limitations are often done unconsciously, with modelers not realizing that they might be missing important model alternatives.[10] For example, a model selection approach applied to modeling interbank interest rates might not a-priori consider using a spread as dependent variable, even though such approach might lead to superior model behavior/performance. Similarly, modeler might limit the model search to models with lagged dependent variables, but not with autoregressive errors.

[9] This is most obvious in macroeconomic data, where the last 20% of data can be very different from the rest of the data. Prior to pandemic the last 20% of data were period of relatively stable macroeconomic environment which is hardly representative of the macroeconomic data over last 25 years. In contrast, after the pandemic the last 20% of data are period of extreme macroeconomic volatility, which again is not representative of the overall data.

This is also reason why hold-out sample cross-validation is more suitable in cross-sectional data than in time series data: randomly selected 20% of observations are more likely to be representative than last 20% of time series data.

[10] Related point is the possibility that model selected by the automatic selection will not satisfy some theoretical constraints. Since the model creator was not actively participating in the model creation – and since they did not explore the model behavior - then it is possible that some aspects of the model remain unknown to the modeler. Simply, the lack of detailed knowledge of the model increases the likelihood that model is wrong from some perspective.

In contrast, model building, which is focused on analysis of the behavior and the performance of the proposed model, is much less likely to suffer from these limitations. By analyzing the model behavior in detail, the modeler will identify instances of bad model behavior/performance and link them to model structure. Therefore, in many situations the modeler is likely to identify alternative model structures or classes that are appropriate for the data at hand, and which are not obvious a-priori. In this way the model building approach is likely to lead to broader set of possible models compared with the model selection approach.[11]

Apart from widening the model universe, the model building approach has advantages over automatic model selection in terms of understanding of the data and resulting model. In many situations, forecasting model should not be thought of as just a tool to create forecasts. Instead, it is tool for understanding the historical data and for forming perspectives on their future evolution. Such understanding does not easily arise from automatic model selection, since in automatic model selection the modeler is not learning much herself.

In terms of understanding the data, the automatic model selection does not include repeated interaction between the modeler and the data, especially in the context of multiple models. The modeler simply sets up the algorithm and then uses the model which was selected as best. They do not know in which periods did the unselected models worked worse, in what way and why. If one understands models as story-telling tools, then observing models that tell wrong stories about the historical data is useful in understanding the nature of the data. In contrast, in model building, there is repeated interaction between the modeler and the data, during which the modeler learns to understand the modeled data themselves. This means that modeler can act not only as subject matter expert on the model itself, but also on the data that are being modeled.

In terms of understanding the model, modeler relying solely on automatic model selection is typically unable to justify the aspects of the structure of the chosen model, such as use of specific variables, transformations and lags. They can only refer to optimality of the chosen model in terms of the selection criteria. In contrast, the interactive and iterative nature of the model building process means that the final model is based on empirically driven modifications of the original (theory-based) model. Simply, the model structure reflects the nature of the modeled data, and the issues in particular part of the sample, as encountered during the modeling. As such, the modeler is able to justify the particular

[11] A possible counter is that by being more ad-hoc the model building approach is likely to lead to less detailed model universe than the systematic model selection approach. However, one can leverage model selection throughout the model building process. For example, one can use automatic model selection when deciding on particular model aspect like number and/or timing of lags of particular variable.

model aspects by relating them to the original motivation for their inclusion, such as periods of bad performance of models which lacked these aspects.

8.2 Estimation output information

All statistical programs provide basic information about the estimated model in what we will call the estimation output. What exactly is included differs, but at minimum the estimation output includes coefficient estimates and associated standard errors, as well as some measure of model fit, most typically R-squared. These pieces of information are included in estimation output for two reasons. First, they provide most basic understanding of the model, as they capture the relationship between dependent and independent variables and strength of this relationship. However, this is not the only or maybe not even the main reason for the typical structure of estimation output. The other reason is that the information included is useful in answering questions that regression methods are mostly used to answer when doing econometric analysis: *'What is the effect of increase in independent variable on dependent variable?'*, *'Is this effect statistically distinguishable from no effect at all?'*, and *'How much of variation in dependent variable do the independent variable(s) explain?'*. All of these questions can be answered (for better or worse) from the estimation output.

Importantly, all of these questions are firmly in the domain of econometric analysis, as opposed to being motivated by forecasting. This does not mean that they are not useful for building forecasting models. Indeed, in many contexts the information included in the estimation output is the most natural starting point in analyzing forecasting model. Correspondingly, the first goal of this section is exposition of how estimation output information can be useful in developing forecasting models.

That said, after starting with outlining the value of each piece of estimation output, we will always proceed to discussing its limitations and drawbacks. Therefore, the second, and in a sense more dominant, message will be that estimation output should rarely be the *last* piece of information one looks at when developing forecasting models: given that their inclusion was not motivated by forecasting considerations, they provide only limited information about the suitability of given model for such purpose.

The rest of this section is separated into three parts. First, we will discuss the informational value of coefficient estimates and associated statistics for the purpose of building forecasting models. In second part we turn to the parts of estimation output which relate to the whole model, rather than to individual independent variables. Finally, the third part will illustrate the points made throughout

the first two parts.

8.2.1 Coefficients and associated statistics

The first numbers one should look at in the estimation output are the coefficient estimates for each independent variable. After all, a model is a postulated relationship between dependent and independent variables, and coefficients tell us what the estimated relationship is. Understanding this estimated relationship can be separated into two steps. First, one can focus solely on the direction of this relationship, which is captured in the coefficient sign. Second, one can also think about the coefficient size together with coefficient significance, to judge whether the relationship is reasonable also from quantitative perspective, not just directionally. Finally, the quantitative perspective will then lead us to discuss associated, less commonly used coefficient statistics.

Coefficient sign. Coefficient sign indicates the direction of effects of movements in in independent variables on the dependent variable. As such, coefficient signs can often be of crucial importance in model building process. In many applications theory provides strong predictions about direction of relationship between dependent and independent variable. However, it can often happen that estimated coefficient implies opposite relationship.

Wrong coefficient sign can be problematic for two reasons. On practical level, the wrong relationship suggests that, if theory is to be trusted, the model itself is wrong and hence is unlikely to work well in forecasting. As such, ruling out models that fail to conform to theoretical predictions for the direction of relationship between dependent and independent variables can help us decrease the negative influence of noise in our data sample. At more pedestrian level, wrong relationship is not acceptable if there are strong demands on part of model users for behavior to comfort to theoretical predictions. For example, model users are unlikely to accept model and its predictions if the estimated coefficients imply that unemployment rate decreases when output decreases. Correspondingly, the first step in model building process in theory-heavy environments is typically checking the coefficient signs and ruling out models that have wrong coefficient signs.

Why would an estimated relationship be different from the one predicted by theory and hence presumably "wrong"? One reason is straight out of econometrics textbook: omitted variable bias. Given independent variable might be correlated with some omitted factor, which biases its associated coefficient away from its true value, and such bias might be so strong so as to cause the coefficient sign to switch. While this is a standard econometrics problem, one should remember that in applied work

focused on forecasting one is much more likely to encounter and even work with models that feature omitted variable bias, simply because some factors might not be observable. From this perspective ruling out models with wrong coefficient signs is akin to ruling out models in which omitted variable bias is too strong.

More benign reason for wrong coefficient sign has to do with the translation of theory into actual model. Theory commonly provides predictions on direction of relationship between two variables but does not provide guidance into the exact form of this relationship. For example, it is not often completely clear which transformation should be relevant: is the level of interest rates that matter for macroeconomic variables, or is it the change from previous period? Similarly, the timing of the relationship is often uncertain and coefficient signs can change from one lag to another. Hence the wrong coefficient sign can be due to wrong exact specification of generally correct relationship. From this perspective, ruling out models with wrong coefficient sign is akin to ruling out models where the exact postulated relationship is "wrong".

The issue of timing of the individual regressors raises one related question: Does individual wrong coefficient sign pose problem when the model includes multiple lags of given regressor? Here it should be clear from discussion in Chapter **??** that individual negative coefficient does not mean negative relationship between dependent and independent variables, as long as there are positive coefficients on earlier lags. Since we typically care only about intermediate and/or total effects, we should not care about individual coefficients but rather at the implications of the coefficients for the dynamic response. Hence, models with wrong coefficient sign on individual lag should not be discarded without consideration of the coefficients on other lags.[12]

Coefficient size and significance. Coefficient sign is the most important piece of information about the relationship between dependent variable and given independent variable. However, it still provides only limited information because it does not speak to the quantitative nature of this relationship. In addition to questions about direction, we also want to answer questions such as *'Does dependent variable move a lot or little in response to changes in the independent variable?'*, and *'Which independent variables are important for determining the movements in dependent variable?'*. These are questions related to coefficient size and significance. Once we have a model that conforms to theory in terms of coefficient signs, one can shift focus to the coefficient size and significance, and what they tell us from perspective of the behavior of the model.

[12]Of course, sometimes even single wrong coefficient sign might suggest problematic dynamic response profile from theoretical perspective, e.g., hump shaped profile associated with shock amplification, but this is not in general the case.

From basic econometrics we know that regression coefficient tells us how much does dependent variable change on average when independent variable changes by one unit.[13] This is an elementary quantitative information about the relationship between the dependent and independent variable and as such is central to analyzing such relationship. Is this information sufficient from perspective of model building and evaluation? That is, does it help to answer our questions *'Is the effect of independent variable appropriately large?'*, and *'What is the relative importance of regressors?'*.

While in principle regression coefficients provide information about both of these questions, in practice they provide information that is far from ideal from model building perspective. The key issue is that regression coefficients reflect the units of the independent variables; they tell the effect of one unit increase in independent variable, whatever are the units in which the independent variable is measured.[14]

This nature of regression coefficients is problematic from model building perspective for two reasons. First, one cannot interpret the regression coefficients without thinking about their units. Thus, regression coefficient equal to, say, 1570 can be appropriately large, or it can be way too small, depending on the units of the independent variable. Moreover, the regression coefficient also reflects the units of dependent variable. In time series context this can often lead to seemingly "small" coefficients in situation when dependent variable is a log-difference while independent variable is not, given that log-differences are typically very small numbers. Bottom line is that whether regression coefficients have appropriate size - that is whether the effects of independent variable are reasonable - is not straightforward to say using the reported coefficients.

Second related issue is that comparing regression coefficients across individual independent variables is not very informative about their relative importance. Since individual independent variables are measured in units that have possibly very different scale, the differences in coefficients often reflect differences in measurement units rather than differences in relative importance. This means that one cannot conclude which variables are more important - or have larger effect - based on comparison of regression coefficients.

If coefficient size alone is imperfect measure tool to judge whether the independent variable has appropriate effect on dependent variable, can we rely on standard errors and associated t-statistics/p-

[13] Note that here we will talk from perspective of least squares estimation, rather than output from general estimation procedure. This is because the interpretation of coefficient size does not necessarily apply to methods that do not rely on least squares estimation. For example, in quantile regression the interpretation is slightly different.

[14] This is the essence of the basic econometrics result that regression coefficients automatically re-scale when we change the units of independent variable. For example, coefficients become 1000 times smaller when we switch measurement from meters to kilometers.

values to rectify this issue? At first glance it would seem so: standard error is measure of uncertainty surrounding our coefficient estimate, and correspondingly t-statistic/p-value is measure of confidence we have that the coefficient is different from zero. Since zero coefficient would mean that given independent variable does not have any effect on dependent variable, it seems reasonable to rely on these statistics to answer our question. Simply, it seems that we could equate large t-statistic/small p-value, what indicates we are certain that the effect is not zero, with conclusion that the effect of given variable is appropriately large.

Alas, things are not so straightforward. T-statistic is ratio of coefficient and standard error. As such it is smaller when coefficient is smaller, but also smaller when standard error is larger. While the former suggest that t-statistic is a good measure of importance of independent variable, the latter complicates the picture. Its presence means that t-statistic can be small even if the effect implied by the coefficient is large, or vice versa.

Low t-statistic despite large coefficient occurs when we think that the coefficient is large, but we are really not sure this is the case. From basic econometrics we know one special case when this can easily happen, namely the situation when we have strong multi-collinearity among the independent variables. In such situation we are uncertain about assigning the effect to either of the collinear variables, while being sure that their combined effect is potentially large.[15] The reverse - significant t-statistic despise small coefficient - just corresponds to situation when given independent variable does not have large effect on independent variable, but we are very sure about the size of the effect. While somewhat unusual, this can still happen, as we illustrate later.

The conclusion is then simple: while providing some additional information about whether given variable has appropriately large effect on dependent variable, t-statistics are imperfect tools for answering questions we want to answer during model building. This should not surprise us given that the motivation for inclusion of t-statistics in regression output was *not* building models for forecasting, but rather answering questions such as 'Can we be sure the effect of given variable is statistically significantly different from zero?'. The limited meaning of statistical significance is also why most econometric papers also talk about economic significance, which is exactly what we meant by coefficients being appropriately large.

[15]Note that this is a key reason why model selection based on variable significance is not a great idea in context of forecasting - you might eliminate variables that have large effects but are insignificant because of strong collinearity among them. Note that from forecasting perspective multicollinearity is almost irrelevant problem, since you care only about the forecast and hence the combined effect of independent variables, and not about their individual effects and whether they are significant.

Standardized coefficients. If size and t-statistics of regression coefficients provide only imperfect information from model building perspective, is there something else we could be using? The answer is yes. The underlying problem we have identified is that coefficients depend on units in which variables are measured, what means that we need to think about these units when interpreting them. The solution is then simple: we just need to change the units in which we measure the independent variables! What units should we use so that the coefficients are immediately meaningful from our model building perspective? The answer is also simple: we should use statistical units such as standard deviation. The whole idea of statistical units is that they are immediately comparable across variables that come from very different distributions. Hence one standard deviation has similar meaning whatever the units in which given variable is measured; and it has the same meaning whether the variable is very volatile or very stable.

This is the idea behind using what we will call standardized coefficients[16] : standardized coefficients say the same thing as normal coefficients, but instead of saying it in terms of natural units of the (in)dependent variables they say it in terms of statistical units, and specifically in terms of standard deviations. Therefore, standardized coefficient tells us by how many standard deviations does dependent variable increase when independent variable increases by one standard deviation.[17] To achieve this interpretation, we just need to multiply normal regression coefficient by the standard deviation of dependent variable and divide it by standard deviation of independent variable:

$$\tilde{\beta} = \beta \frac{\sigma_y}{\sigma_x} \tag{8.1}$$

Thus calculated standardized coefficients provide us with significant improvement over normal coefficients. First, we can immediately understand whether some effect is appropriately large or not, since the effects are expressed in terms of standard-deviation-to-standard-deviation. Thus, if the coefficient is close to 1, we know that the effect of independent variable substantial or even very large given that one standard deviation change in independent variable leads to one standard deviation change in dependent variable. Meanwhile, if the standardized coefficient is close to zero - say less than 0.01 - then we know the variations in independent variable do not cause almost any movement in dependent variable.

[16] Standardized coefficients are also referred to as XXX

[17] This should immediately alert us to limitations of standardized coefficients: standard deviation is not a sensible quantity for many regressors. This includes dummy variables, variables where distribution is far from symmetric or variable that are trending over time, as we can encounter in co-integration regressions. In all of these situations standardized coefficients should be either interpreted with grain of salt or ignored altogether.

What values should we expect for standardized coefficients? To answer this question, we need to realize that regression is fundamentally about explaining the variations in dependent variable. If we have model with single independent variable and the standardized coefficient is close to 0 - i.e., dependent variable does not respond to movement in independent variable - then it cannot be that we are explaining significant portion of variance in the dependent variable. Conversely, when in such situation we have coefficient close to 1 then we would expect that we are explaining very large share of variation in dependent variable. Since R^2 is measure of explained variance in dependent variable, we would expect standardized coefficients to be sum up to somewhere around R^2, and hence individual standardized coefficients to be somewhere around $\frac{1}{K}$ of R^2. More generally, standardized coefficients above 0.1 can be considered meaningful in most situations, and ones above 0.3 can be considered large.

Second, we can immediately compare standardized coefficients across variables because they are all measured in the same statistical units. Thus, variables with larger standardized coefficients will be responsible for a larger share of movements in dependent variable in our forecasts. Here it is important to provide a word of caution about interpreting standardized coefficients, what will lead us to second piece of information we should be looking at when building models.

Relative importance of regressors. While standardized coefficients allow for immediate comparison and understanding where movements in dependent variable will come from in our forecasts, they are not suitable for answering more specific question of *'What share of explained variance does each regressor explain?*'. Since the standardized coefficients do not sum up to R^2 they cannot be used to answer such answer with any precision. Instead, to answer this question we need to rely on methods that are aimed at decomposing share of variance of the dependent variable explained by the model. In other words, we need methods that decompose R^2.

The problem of decomposing R^2 into contribution of individual independent variables is not as simple as it sounds: given that individual independent variables are in general correlated together, then assigning "explanation" to individual variables is complicated. To see this consider determining shares of explained variance by running series of regressions in which dependent variable is always regressed on single independent variable. Since R^2 is the share of variance explained, then it is reasonable to expect R^2 from such regressions with single explanatory variable could be considered as share of variance explained by that variable. Alas, this approach will not lead us to our desired goal due to likely correlation among independent variables: in presence of such correlation the explanatory variable in single variable regression will pick up the effect of other correlated explanatory variables as well,

and hence R^2 will not reflect only the explanatory power of given independent variable.[18] In practice this means that the sum of R^2 from this series of regressions will be higher, potentially significantly higher, than the R^2 from the actual regression with multiple independent variables. Simply, there will be some double counting caused by correlation among independent variables.

An alternative approach also based on running multiple regressions is to look how much does R^2 decrease if we remove each individual variable one by one. The idea here is that the share of explained variance could be equated with the increase in R^2 caused by including the variable. Alas, here the problem is the reverse of problem encountered in previous solution: the correlation among independent variables means that if given variable is missing from the model then some of its explanatory power will be appropriated by the independent variables present in the model. In practice this means that the sum of R^2 obtained from such regressions will be lower, potentially significantly lower, than the actual R^2.

This discussion should teach us one key lesson: decomposition of R^2 into shares assigned to each variable is not uniquely defined, given that the share assigned to given variable by particular decomposition scheme depends on which other variables are already included in the model. The solution to this problem is to try all possible combination of models and then average the contributions over all of these combinations. In practice this means that to decompose R^2 into shares assigned to each independent variables we need to rely on a set of auxiliary regression. Specifically, we will run regression with and without our variable of interest for each possible combination of the remaining independent variables. We then calculate the increase in R^2 caused by inclusion of our variable in each individual auxiliary regression and then calculate their average. This will then be our portion of variance in dependent variable explained by given independent variable. We can than repeat this for all independent variables to get portions of variance explained for each independent variable. Finally, if we want things expressed as share of R^2 we can also divide these average increases in R^2 by the actual R^2.[19]

[18] This effect is something we know from basic econometrics course. It is the effect that lies behind omitted variable bias.

[19] To clarify this procedure, consider example of regression with constant and three independent variables, x_1, x_2 and x_3. To assign share of explained variance to x_3 we would run 4 pairs of regressions, with one that would not include x_3 and one that would. First, we would run regression only with constant, and then regression with constant and x_3. Next, we would run regression with constant and x_1, with and without x_3. Then two regressions with constant and x_2, and finally two regressions with constant, x_1 and x_2. For each pair we would calculate the difference in R^2 between the regression that includes x_3 and that does not. Finally, we would calculate the average difference in R^2,which would be then our share of variance explained by x_3. To get the same for x_1 and x_2 we would just repeat the above procedure for each of these variables.

8.2.2 Model statistics

The second category of information provided in typical estimation output relates to the whole model, rather than to individual impendent variables. Correspondingly we will call the information in this category *model statistics*. The primary such statistic is of course R-squared, but we will also discuss other statistics.

R^2 is included in standard estimation output as a measure of explanatory power of the model. The idea is simple. The dependent variable of given model varies over our sample, being high in some periods and low in other periods. The goal of the model is to explain this variation. We can then evaluate the model by looking at what share of the variance does the model explain. And this is what R^2 does: it is the ratio of explained sum of squares divided by total sum of squares. Since total sum squares is basically the variance of dependent variable, then ratio of explained sum of squares to total sum of squares is really just ratio of explained variance to total variance.

In time series context the explanatory variables are collection of information available for forecasting dependent variable. For example, in univariate time series models the explanatory variables are past values of dependent variable. Therefore, R^2 is the share of variance in dependent variable that can be explained by its past values, which means that it is a measure of how well we can forecast *current* value of dependent variable using its *past* values. In multivariate context, where we potentially also use concurrent variables, the interpretation is slightly different, but R^2 still measures how well can we forecast current value of dependent variable given our information set.

This then is the most important informational content of R^2 in context of building times series models for forecasting: It is effectively a specific measure of particular type of forecasting performance. However, in next section of this chapter we will see that R^2 is rather limited measure of forecasting performance, what will be our motivation for alternative, more sophisticated measures of forecasting performance.

Another commonly reported model statistics is the **standard error of the regression**. This is the standard deviation of the residuals of the estimated equation. Given that residuals are one-step-ahead forecast errors of the equation - a residual is a difference between the value predicted by the equation and the actual value - then the standard error of regression can be interpreted as measure of quality of forecasts. Specifically, the statistic can easily tell us what is the uncertainty around our forecasts. In other words, the standard error of the regression tells us the typical width of the range

within which we should expect our forecasts to lie.[20] This reveals why the statistic can sometimes be more valuable than R^2. Since R^2 is a share statistic, it does not contain any distance information, and hence cannot help us answer questions which start with *'how far?'*. Relatedly, the standard error of the regression has also the advantage that it reflects the units of dependent variable, whatever is its transformation, and hence allows for easy interpretation. Of course, the flip side to this is that if the modeler does not have intuitive feeling what the standard error should be, then single value might not provide much information (even though comparison of values for multiple models would).

A similar statistic to R^2 is a log-likelihood of the estimated model. This is a transformation of the value of the likelihood of the model, that is, approximately the probability that we would observe given data if the estimated model is true. However, this statistic does not have a straightforward interpretation like R^2 as it does not have meaningful units, and hence its value for model building is limited. That said, it is the building blog of information criteria, which are another common model statistic reported by the statistical packages. These are one of the most common statistics used in model *selection*. However, as derivates of log-likelihood they share its drawbacks, and hence their value in model *building* is limited.

Statistical packagers also commonly report test statistic closely related to R^2, namely the F-statistic. This is the statistic for joint hypothesis test that all coefficients in regression are zero. Hence, it can be seen as the formal statistical test of hypothesis that the model does not provide any useful information about the dependent variable. This can be useful in model building in so far as to judge whether the model is useful, statistically speaking, but ultimately it does not provide much additional information in addition to information contained in R^2.

Finally, apart from various measures of model fit, the reported model estimates also often contain statistic that is related to the statistical nature of the residuals, the Durbin-Watson statistic. This is a statistic that measures the first-order autocorrelation in the residuals. Since residuals are forecast errors, then Durbin-Watson statistic can tell us whether forecast errors will be correlated over time; in other words, it tells us whether we should expect positive or negative forecast error next period, if we observed positive forecast error in current period. Since correlated forecast errors are not optimal, this tells us that our model is missing something in order to be optimal forecasting model. Given that the value of 2 means absence of serial correlation, then values far away from 2 – e.g., values below 1

[20] This formulation of course violates the statistical language a bit. A more precise - but less intuitive - formulation would be to say that the standard error of the regression tells us the size of region in which around two thirds of predicted values would fall if we would add it to the predicted value and assume that the errors are distributed normally.

or above 3 - suggest that we should think about including additional components to our model. These additional components can take form of ARMA errors or lagged dependent variable, as we discussed in Chapter **??**. Or they can take form of additional regressors which are themselves correlated, and hence absence of which cause correlation in forecast errors. In either case, the Durbin-Watson statistic can be a signal for modeler to go back to drawing board because the model is missing something.[21]

8.2.3 Illustrations

Table H.1 shows prototype of estimation output produced by SpecEval EViews add-in. The original output is taken from EViews but adjusted to make it more suitable for model building purposes. This includes several things. First, many of the values are color-coded so that the modeler can quickly and easily extract the basic information about the values, given that human brain processes graphical information much faster than numerical information. The coefficient values are color coded based on their sign, with positive values assigned dark green color and negative values assigned red color. Similarly, the p-values are color coded by the implied significance of the coefficients, with significance indicated by light green, insignificance by orange and border cases by yellow. In the same spirit, the R^2, F-statistic and Durbin-Watson statistic are all assigned colors to indicate values that are either ok or potentially problematic, even though in case of R^2 the classification is highly imprecise for some applications. All of these numerical values are also set to format which will be easiest for user to process, decreasing the number of digits where appropriate.

Second, the table includes additional information. In terms of regression coefficients, it also includes the standardized coefficients, which are often very useful for model building. These are also color coded, this time according to their absolute value. Meanwhile, at the bottom of the table is added information about the underlying independent values, which is useful information for outside evaluators of the model which might not know the variable names used for the variables.

[21] It is important to keep in mind that Durbin-Watson statistic close to 2 does not mean we are out of the woods: we might have autocorrelation of higher degree, even though that is less likely and more complicated to deal with in terms of model adjustments.

Table 8.1: Illustration of estimation output

Dependent Variable: DVAR
Method: Least Squares
Sample (adjusted): 2000M01 2019M12
Included observations: 240 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.	Std. coef.
C	−0.58	−0.029	−20.00	0.0000	
IVAR1	10.5	7.0	1.50	0.084	0.13
IVAR2	158	316	−0.50	0.22	−0.072
IVAR3	−0.0032	−0.0013	2.50	0.00	−0.73

R-squared	0.32	Mean dependent var	4.03
Adjusted R-squared	0.24	S.D. dependent var	1.44
S.E. of regression	0.93	Akaike info criterion	2.71
Sum squared resid	153	Schwarz criterion	2.79
Log likelihood	-240	Hannan-Quinn criter.	2.74
F-statistic	85.0	Durbin-Watson stat	0.14
Prob(F-statistic)	0.00		

DVAR	Dependent variable description
IVAR1	Independent variable 1 description
IVAR2	Independent variable 2 description
IVAR3	Independent variable 3 description

Table H.1 is not based on actual data; it is just a prototype with numbers. To illustrate some of the points discussed in this section, we will turn to several examples of regression outputs estimated either on real-world data or on simulated data.[22]

First, Table H.2 provides example of how coefficient signs can be sensitive to what is included in our model. It relates to classic macroeconomic questions of the effect of changes in exchange rate on inflation. Theory is pretty clear that depreciation should lead to higher inflation rate, as costs of imported goods rise. Hence, we expect a negative coefficient on exchange rate. However, when we estimate model linking inflation rate to changes in Czechia exchange rate against euro, we obtain positive coefficient (Model 1). A possible solution in form of using current and lagged changes in exchange rate, or just lagged changes, does not help us (Model 2&3). However, once we include change in oil prices, which are important driver of inflation, then the coefficient turns negative as expected. This is especially true if we include not only current change in oil prices, but also two lags. This illustrates how coefficient on our variable of interest can easily have wrong coefficient sign, and why checking coefficient signs should be the first step in analyzing our estimated model.

[22]The below examples should not be treated as examples of good models. Instead, they are illustrations of particular problems one can encounter when using regression outputs during model building .

Table 8.2: Illustration of coefficient sign sensitivity

Dependent Variable: @PC(CPI)

Variable	Model 1	Model 2	Model 3	Model 4	Model 5
C	0.17	0.16	0.17	0.17	0.17
@PC(EURO_FX)	0.01	0.01		−0.01	−0.02
@PC(EURO_FX(-1))		0.01			
@PC(EURO_FX(-2))		0.02	0.02		
@PC(EURO_FX(-3))		0.02	0.02		
@PC(EURO_FX(-4))		0.01	0.01		
@PC(OIL_PRICE)				0.01	0.01
@PC(OIL_PRICE(-1))					0.00
@PC(OIL_PRICE(-2))					0.00
R-squared	0.00	0.03	0.02	0.07	0.11

Next, Table H.3 shows a typical example of how interpreting coefficients size requires some care. The table shows regression output for model linking inflation rate, modeled as log-differences in CPI, to oil prices, import prices and output gap. First thing to notice is that all the coefficients are small numbers. Does this mean that the effect of given independent variable on inflation is correspondingly small? The answer in all three cases is no.

Table 8.3: Illustration of coefficient size sensitivity

Dependent Variable: DLOG(CPI)
Method: Least Squares
Sample: 1998Q1 2019Q2
Included observations: 86

Variable	Coefficient	Std. Error	t-Statistic	Prob.	Std. coef.
C	0.0062	0.00068	9.11	0.000	
@MOVAV(DLOG(OIL_PRICE),4)	0.011	0.0077	1.40	0.17	0.14
@MOVAV(DLOG(IMPORT_PRICES(-1)),4)	0.21	0.066	3.19	0.00	0.31
OUTPUT_GAP	0.00087	0.00032	2.75	0.01	0.27

R-squared	0.215	Mean dependent var	0.01
Adjusted R-squared	0.187	S.D. dependent var	0.01
S.E. of regression	0.0061	Akaike info criterion	−7.30
Sum squared resid	0.0031	Schwarz criterion	−7.19
Log likelihood	318	Hannan-Quinn criter.	−7.25
F-statistic	7.5	Durbin-Watson stat	1.17
Prob(F-statistic)	0.000		

OUTPUT_GAP

Start with the simplest reason that applies to case of output gap. Output gap is measured in

percentages and its values range from around -5 to around +5. In contrast, log-differences are very small numbers, in our case varying between -0.005 and 0.025. Therefore, the dependent and independent variables are measured in units that have very different scale, which can be seen in right panel of Figure H.2. When we put both variables on the same axis, the log-differences in CPI actually look like flat line. And this is the explanation for the small coefficient associated with output gap. If instead of the coefficient associated with natural units, we look at the coefficient associated with statistical units - the standardized coefficient - then we can see that the coefficient on output gap is far from being small. This "problem" will be present whenever we use log-differences for dependent variable and some other transformation for independent variable. [23]

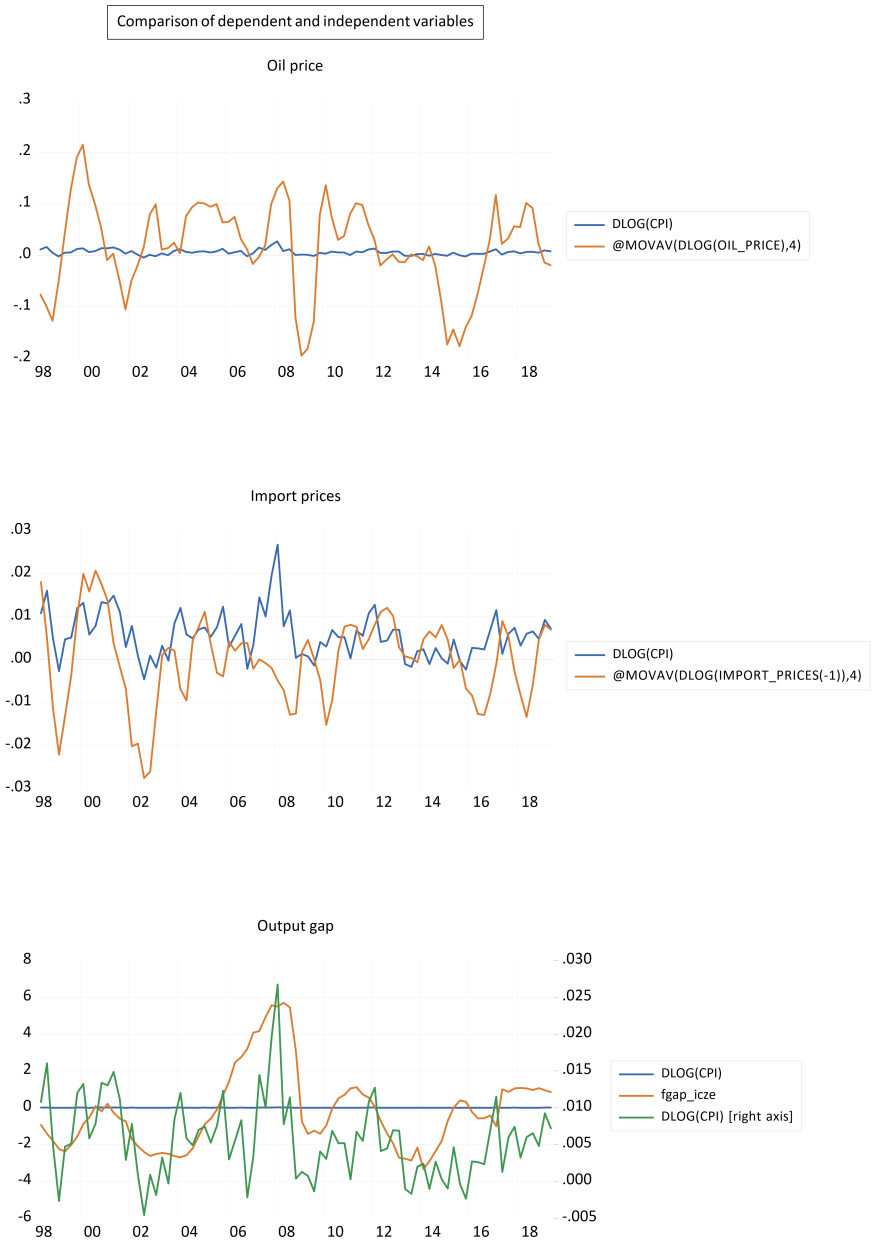
Nevertheless, this explanation for seemingly small coefficient applies only to the output gap. The other two independent variables are measured in log-differences, and hence small coefficient is not a problem of different units. Still, it does not mean that the coefficients are small. Take first the import prices. When interpreting the coefficient of 0.21 - which states that when import prices increase by 1% CPI increases by 0.21% - we need to take into account the relative variability of import and consumer prices. Middle panel of Figure H.2 shows that while both variables have *roughly* the same scale, import prices are roughly 1.5-times more variable. This is then reflected in the standardized coefficient, which is 1.5 time as large as the unadjusted coefficient. In other words, when import prices increase by 1 standard deviation, consumer prices increase by 0.3 standard deviations. This is reasonably large coefficient.

The same logic applies to oil prices, just at a much bigger scale. Left panel of Figure H.2 shows that the variability of oil prices is many times larger than variability of consumer prices. Hence, the unadjusted coefficient of 0.011, which seems very small, actually captures relatively strong effect of oil prices on consumer prices. As indicated by the standardized coefficient, one standard deviation leads to 0.14 standard deviation change in consumer prices. Simply, a one percentage change in import prices is much "bigger" than one unit change in oil prices, and this difference has to be taken into account.

Final observation based on this example is the comparison of relative importance of regressors. Even though import prices and oil prices are measured in the same units, and even though the coefficient on

[23]Log-differences are very similar to percentage changes in terms of what they measure - a change from period to period. However, the units are very different. That said, there is a simple rule-of-thumb: to get units comparable with percentage changes, just multiply log-differences by 100. In our example here, using this rule of thumb suggest that when output gap is 1% higher than quarterly inflation is 0.12% lower. In annualized inflation terms this amounts to 0.5% lower inflation rate, which is a sizable effect.

Figure 8.2: Illustration of regressors with different variance



the import prices is 20-times larger than on oil prices, the importance of import prices is only 2-times larger, as indicated by the standardized coefficients. In other words, oil prices explain almost as much variance in the dependent variable even though they have coefficient 20-times smaller. Simply, while the effect of same move as measured in log-differences is 20-times smaller, oil prices typically face so much larger moves that this almost completely compensates for this difference. Looking at unadjusted coefficients would not provide any information about this.

This is also confirmed if we look at relative importance of regressors via decomposition of R^2. Table H.4 shows the decomposition using three different methods: (1) when only given regressor is included, (2) when given regressor is dropped, (3) averaged over all orderings of regressors. The table confirms what we already concluded from looking at standardized coefficients (include in the last column for comparison purposes). The share of variance explained by the import prices is only around 3-times as large as share explained by import prices, despite the associated coefficient being 20-times as large.

Table 8.4: Illustration of relative importance of regressors

	Statistic			
	Included	Dropped	Average	Std. coef.
Oil price	0.04	0.019	0.03	0.14
Import prices	0.11	0.097	0.1	0.31
Output gap	0.09	0.07	0.08	0.27

Turning to coefficient significance, Table H.5 shows example of model which features statistically insignificant regressors that are nevertheless important for the dependent variable. Specifically, the model links log-returns of Czechia stock prices to log-returns in German and French stock prices. Coefficients for both are statistically insignificant at usual significance levels. However, this does not mean that the German and French stock prices are useless for explaining movements in Czechia stock prices. Coefficients of 0.38 and 0.47 mean that 1% move in German and French stock prices causes 0.38% and 0.47% move in Czechia stock prices. This is clearly not small. This can also be seen in the standardized coefficients, which are clearly not small. Or for the matter the of fact also in R^2, which is reasonably high, especially given that we are talking about stock prices, which should be hard to predict.

Figure 8.3: Illustration of regressors with different variance

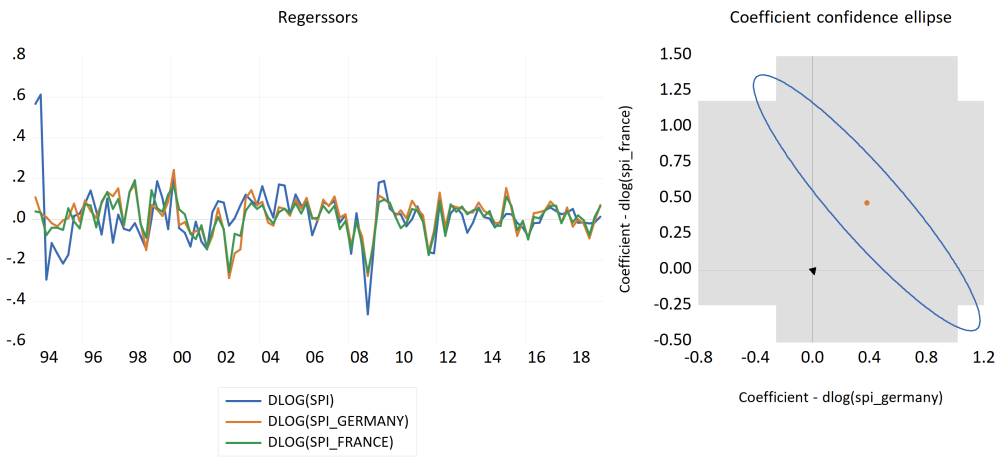


Table 8.5: Illustration of coefficient significance sensitivity

Dependent Variable: DLOG(SPI)
Method: Least Squares
Date: 03/12/23 Time: 12:02
Sample (adjusted): 1993Q4 2019Q2
Included observations: 103 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.	Std. coef.
DLOG(SPI_GERMANY)	0.38	0.32	1.19	0.24	0.25
DLOG(SPI_FRANCE)	0.47	0.36	1.32	0.19	0.28

R-squared	0.275	Mean dependent var	0.01
Adjusted R-squared	0.268	S.D. dependent var	0.13
S.E. of regression	0.11	Akaike info criterion	−1.49
Sum squared resid	1.30	Schwarz criterion	−1.44
Log likelihood	78.9	Hannan-Quinn criter.	−1.47
Durbin-Watson stat	1.130		

What explains the insignificance of the coefficients is strong multicollinearity between the two regressors: the correlation between them is 0.92, as left panel of Figure H.3 attests. What this means is that any regression will have problem distinguishing between effect of movements in German stock prices and movement in French stock prices, as these occur together. This makes it basically impossible to determine which one is "causing" movement in Czechia stock prices. It could be the former, it could be the latter, or it could be both, and the regression is unable to tell which one of these is the case.

Statistically this is reflected in large standard errors. And hence, the insignificant coefficients are

not caused by small coefficients, but rather by large standard errors. This in turn reflect the uncertainty about the exact value of the coefficient due to the uncertainty about distributing the effects between the two variables. This, however, in no way means that the two variables do not have effect. One way to see that is to visualize the joint confidence ellipse, shown in right panel of Figure H.3. While the individual confidence intervals for both coefficients - visualized by shaded area - clearly contain zero value, reflecting their statistical insignificance, the confidence ellipse clearly excludes the possibility that both are zero at *the same time*. The multicollinearity of the two regressors is reflected in the shape of ellipse, which is very elongated and different from circle, which would be the case if the value of the two coefficients would be independent. In other words, if we would tell the regressions that value of one coefficient is lower, then the regression would suggest that the value of the other regressor is correspondingly larger.

This example then highlights the importance of treating coefficient significance with care. If the modeler would not be careful, he could conclude that neither variable belongs to the model. In reality, all that the regression output is saying is that the regression is uncertain about the effect of each regressor on the dependent variable. Note that from perspective of forecasting, this uncertainty is not an issue - it mostly pertains to the standard errors, not the coefficients themselves. And even more importantly, in so far it pertains to the coefficients, it has almost no bearing on the *sum* of the two coefficients. Assuming that movements in German and French stock prices remain correlated into the future, the uncertainty about individual coefficients will be almost irrelevant. This then leads us to our conclusion: If forecasting is the only use of the model then statistical insignificance of coefficients is pretty much irrelevant on its own.

Meanwhile, tables H.6 and H.7 provide examples of situation when significance coefficient does not mean large effect. In the first example the coefficient is seemingly large, and is clearly significant, but as the standardized coefficient shows, movements in the independent variable cause only small movements in the dependent variable. This is also reflected in the very low R^2 value, highlighting that x is unable to explain much variation in y. This is a typical example of significant but irrelevant regressor: the variance of y caused by other factors than x swamps the variance caused by x. Meanwhile, the second example shows situation where we have two regressors with similar coefficients and both being statistically significant, but only one of them can be considered "large" in the sense of having large influence over variation in y - that of x_2, as indicated by standardized coefficients. Again, while x_1 does influence y, its variation relative to variation of y is just too small to have any real influence. In

contrast, variation in x_2 is an important determinant of variations in y.

Table 8.6: Illustration of statistically significant regressor with small effect (Example 1)

Dependent Variable: Y
Method: Least Squares

Variable	Coefficient	Std. Error	t-Statistic	Prob.	Std. coef.
C	14.8	0.17	85.6	0.000	
X	2.38	0.3	7.96	0.000	0.079

R-squared	0.006	Mean dependent var	15.97
Adjusted R-squared	0.006	S.D. dependent var	8.64
F-statistic	63.4	Durbin-Watson stat	1.97
Prob(F-statistic)	0.000		

Table 8.7: Illustration of statistically significant regressor with small effect (Example 2)

Dependent Variable: Y
Method: Least Squares

Variable	Coefficient	Std. Error	t-Statistic	Prob.	Std. coef.
C	0.49	0.0093	52.4	0.000	
X1	0.052	0.01	5.21	0.000	0.046
X2	0.019	0.001	18.8	0.000	0.17
X3	0.51	0.01	50.3	0.000	0.44

R-squared	0.225	Mean dependent var	0.86
Adjusted R-squared	0.225	S.D. dependent var	0.33
F-statistic	967.7	Durbin-Watson stat	1.97
Prob(F-statistic)	0.000		

Both examples then highlight the point that we can have regressors that have significant but not meaningfully large impact on dependent variable, which occurs when the variance explained by these regressors is small fraction of the overall variance of dependent variables. To see this what is behind this consider the data-generating processed used to simulated the data. The first regression is based on data simulated according to $y = 2 * x + \epsilon$, where the variance of the ϵ is 1000-times bigger than variance of x, so that the other stuff than x is much more important than x. The second regression is based on data simulated according to $y = 0.02 * x_1 + 0.02 * x2 + 0.5 * x3 + \epsilon$, where x_1 has variance of 1 while x_2 has variance of 100. This means that x_2 is much more variable than x_1, and since their coefficient is the same, also much more important for y.

8.3 Forecasting performance information

It is natural that when building models used for forecasting, forecasting performance is at the heart of the model building process. That said, it is still useful to stop and think in what way can studying historical forecasting performance be useful in our model building process.

Broadly speaking, the value of forecasting performance can be divided into three categories. First, and most obvious, is the **signal value**: good forecasting performance on observed historical sample is likely to be signal of good forecasting performance on future values. This is quite natural; a-priori, it would be strange to expect negative correlation between historical and future forecasting performance, even if one can imagine situations where models with better historical forecasting performance are likely to have worse future forecasting performance.[24] Second value of studying forecasting performance is the **selection value**. The goal of any modelling is constructing model that best approximates the true data-generating process (DGP) of the time series. It is logical that models that are closer to the true DGP should have better forecasting performance. Hence using forecasting performance as guiding principle in selecting the final model should mean that we will be more likely to select the "correct" model.

The final, and in some sense most important, value of forecasting performance is the **indicative value**. When we will evaluate forecasting performance in detail, we are likely to observe periods of bad forecasting performance. Such periods can be indicative of missing factors or particular form of model misspecification, and hence can directly lead us to improvements for the model. This goes to the heart of our model building process: the fact that it is iterative and interactive means that we can rely on detailed information about forecasting performance to propose improvements to the current considered model.

The rest of this section is divided several parts. The first part discusses one-step forecasting performance and how it relates to model statistics discussed in previous section. In doing so it also provides additional critical discussion of these model statistics. The second part switches focus to multi-step forecasts. The third part then returns to some issues glossed over in the first two parts and provides more in-depth discussion.

[24] One such situation is when model building leads to overfitting. We will discuss the problem of overfitting in dedicated section.

8.3.1 One-step forecasts and model fit information

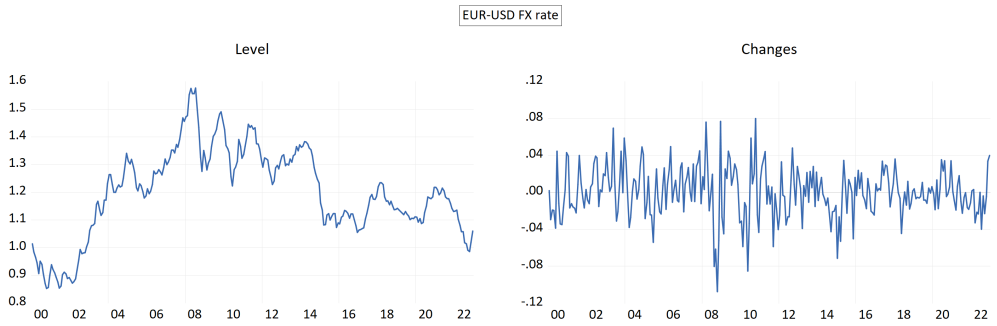
One-step ahead forecasts are the most common forecasts used in practice, which makes them a natural starting point in our discussion. Apart from their prominence, one-step forecasts are also different from multi-step forecasts because some information about their precision is readily available: as we have discussed in previous section, some of the model statistics in estimation output can be viewed as measures of precision of one-step forecasts. R^2 is a share of variance of dependent variable explained by the independent variables, which means it is measure of how much variance of dependent variable is captured by the one-step ahead predictions. Similarly, the standard error of the regression is measure how far on average are one-step predictions from the actual values. This means that both of these model statistics are measures of one-step ahead forecasting performance. More generally, for any model the model residuals are readily available, and model residuals are just a difference between one step forecasts and actual values. The question is whether we can fully rely on R^2 and related statistics in model building process, or whether we need any additional information about the one-step forecasting performance?

It turns out that the answer is that we cannot rely on R^2 alone and therefore that we need additional information to draw more robust conclusions. The model residuals and the associated model statistics have substantial drawbacks that limit their use in terms of model building. The key problem is that model residuals are measured in terms of units of dependent variable. This means that they change when the dependent variable has different transformation, and hence that the model statistics are sensitive to such changes in transformation of dependent variable. This then limits the usefulness of the model statistics in model building when different transformations of dependent variable are considered.

Take for example situation when one is considering models both in terms of levels and in terms of first differences, possibly because the stationarity of the series is in question, as in Figure H.4. The key aspect is that levels of such series are much easier to forecast than its first differences, in the sense that one can "explain" the variance of levels easier than variance of first differences. In extreme case, the value of random walk can be well forecasted by its past value, while its first differences are unforecastable. This them means that any model that has level of the series as dependent variable will likely have much higher R^2 than model that has first differences as dependent variable.

This can even take absurd proportions, when the two models are identical. Table H.8 shows example when one estimates an identical random walk model for U.S. NAIRU, just differently expressed.

Figure 8.4: Example of borderline nonstationary series



Specifically, the first model has first differences are dependent variable, while the second model has levels as dependent variables, and includes past value of NAIRU on the right-hand side with forced coefficient equal to 1, so that it is simply the same equation with past value of NAIRU with given coefficient on left- or right-hand side of the equation. Even though the models are identical, their R^2 values are completely different, with first model having R^2 of close to 1, while second having R^2 of zero. A modeler not aware of this sensitivity of R^2 to different transformations of dependent variable might be lead astray when following R^2 as its guide, since he might conclude that the former model is better than the latter model. Of course, this was just and extreme example, but it points towards a general problem: the model residuals and associated model statistics are hard to compare across models with different transformation of dependent variable.[25]

Another related issue is the sensitivity of model statistics to inclusion of dynamics terms. In so far as time series are persistent, inclusion of dynamic terms improves the precision of one-step forecasts: persistence means that adjacent values are relatively close to each other, so that models that anchor current value to previous value(s) will be led to smaller residuals and hence higher R^2. While this improvement is not a mirage, as in the case of sensitivity to dependent variable transformations, there are important caveats to it. First, the overall improvement in one-step ahead forecasting performance can mask worsening of performance in periods of increased volatility. During such periods times series can display significantly less persistence - think about central banks throwing their caution in the air during periods of recessions or financial stress. Correspondingly, models with dynamic terms might have significantly lower precision of forecasts.

Prime example of this phenomenon are models for policy rates, which often include lagged policy

[25]In general, the other model statistics share these weaknesses, given that they are also functions of model residuals, even though in the specific situation here this is not the case.

Table 8.8: Illustration of R^2 sensitivity (Random walk for FX)

Dependent Variable: D(EURUSD)
Method: Least Squares
Sample (adjusted): 1999M02 2016M03
Included observations: 206 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.	Std. coef.
C	−0.00035	0.0023	−0.15	0.88	

R-squared	0.000	Mean dependent var	−0.00
Adjusted R-squared	0.000	S.D. dependent var	0.03
S.E. of regression	0.033	Akaike info criterion	−4.00
Sum squared resid	0.22	Schwarz criterion	−3.98
Log likelihood	413	Hannan-Quinn criter.	−3.99
Durbin-Watson stat	1.450		

Variable	Description

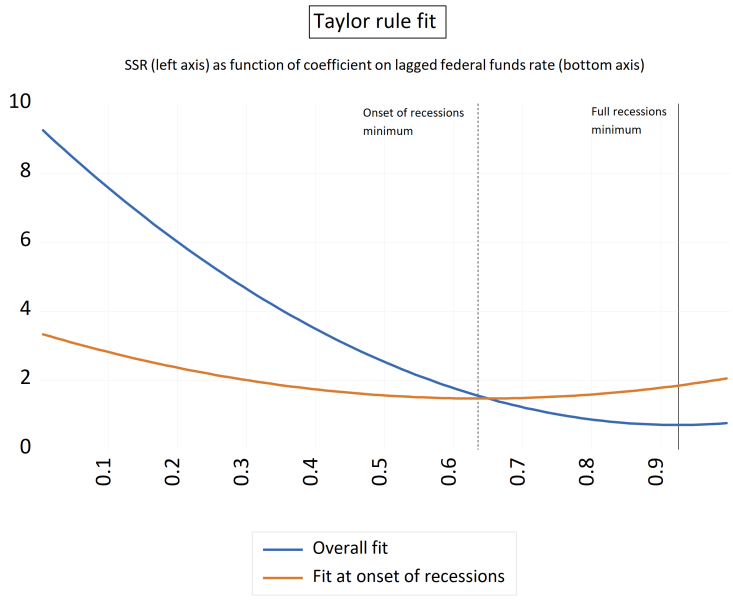
Dependent Variable: EURUSD
Method: Least Squares (Gauss-Newton / Marquardt steps)
Sample (adjusted): 1999M02 2016M03
Included observations: 206 after adjustments
EURUSD = EURUSD(-1)+C(1)

	Coefficient	Std. Error	t-Statistic	Prob.	Std. coef.
C(1)	−0.00035	0.0023	−0.15	0.88	

R-squared	0.958	Mean dependent var	1.19
Adjusted R-squared	0.958	S.D. dependent var	0.16
S.E. of regression	0.033	Akaike info criterion	−4.00
Sum squared resid	0.22	Schwarz criterion	−3.98
Log likelihood	413	Hannan-Quinn criter.	−3.99
Durbin-Watson stat	1.450		

Variable	Description

Figure 8.5: Illustration of trade-off between better overall fit and better sub-sample fit



rate as one of the regressors. While this greatly improves the overall forecasting performance, and correspondingly increases the overall R^2, it can lead to worse performance during onset of recessions, and correspondingly lowers the R^2 for these periods. Figure H.5 illustrates this point: While the fit for the whole sample is maximized when the coefficient on lagged dependent variable is equal to 0.93, the fit periods corresponding to onset of recessions is maximized at much lower value of 0.64. [26] In other words, there is a trade-off between fit for the whole sample and fit for periods when the series is more volatile. Of course, in many applications it is the latter periods which are of particular interest. Notice that for the periods corresponding to onset of recessions the value of even including dynamic terms is much smaller than for the whole sample performance, as sum of squared residuals only declines from around 3.5 to around 2 when the coefficient from lagged dependent variable goes from 0 to 0.64.

Meanwhile, the second caveat of model statistics in presence of dynamic terms is connected to their limited focus: they reflect the forecasting performance in terms of one-step forecasts. So even if inclusion of dynamic terms leads to improvement in one-step forecasting, as captured in higher R^2, this might not be - and often is not - without costs in terms of multi-step forecasting performance. Therefore, focusing on model statistics might lead to suboptimal model choice if longer forecasts are

[26] Specifically, the Figure shows the model fit for the standard Taylor rule model given by $ffr = \rho * ffr_{t-1} + (1-\rho) * [r_t^* + 2 + 1.5 * (inflation_t - 2) + 0.5 * output\_gap_t]$, where ρ is allowed to vary from 0 to 0.99.

also of interest. This is issue we return to in next subsection.

The discussion so far focused on R^2, which is model statistic that gets most of the attention. Another common tool for analyzing time series models, this time graphical one, is the time series plot of model residuals, or its analog of time series plot that includes actual and fitted values as well as residuals. Few examples of such plot are in Figure H.6. How is inspecting such plot different from looking at R^2, give that R^2or standard error of the regression, given that those statistics are based on residuals? It turns out that the difference can be large, like a difference between inspecting plot of values of given variable and just looking at the summary statistics. The key thing to realize is that the time series plot of residuals provides us with much more information than the model statistics, because it shows all the values, rather than just one representative statistic. In situations when given summary statistic is not representative of all the values, the plot can provide information that goes above and beyond the information contained in the statistic.

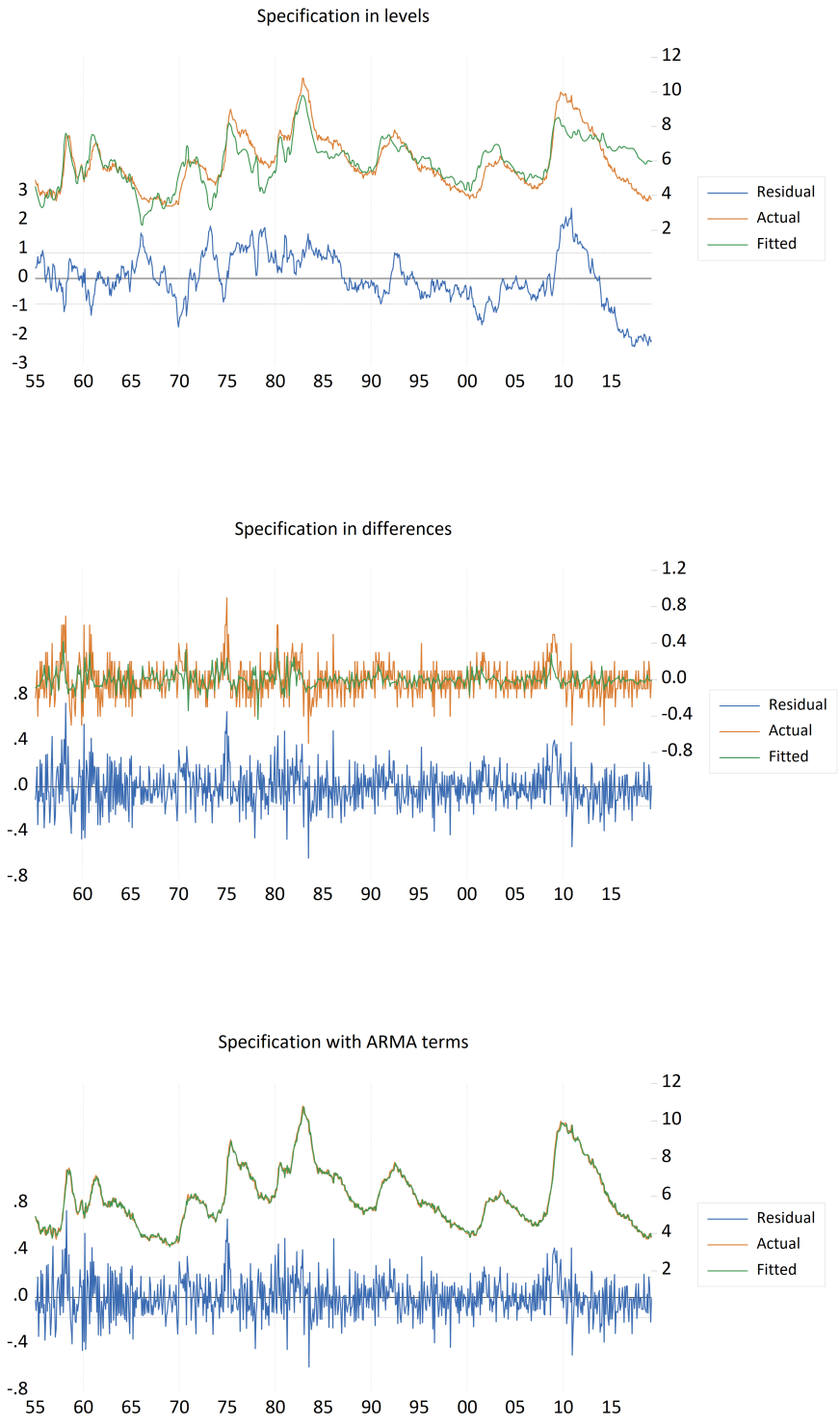
For example, the plot can reveal that a model for unemployment rate fits well during normal times but poorly during periods of recession, or that it fits much worse in one part of the sample than in the rest of the sample. Such nuanced conclusions are impossible to draw based on model statistics, which are by nature calculated for whole sample.[27]

While the time series plot of actual and fitted values together with residuals is potentially useful tool, there are many situations in which its value is much diminished. For it to be valuable, one needs to be able to discern periods and/or conditions when the model is making good forecasts, and when it is not. This is unfortunately not always the case. At benign level, if the sample is very large, counting hundreds of observations, then the individual periods will not be discernible. Hence, the plot is much more useful for macroeconomic data with quarterly frequency than for financial data with daily frequency.

At more serious level, the plot suffers in much the same situations when value of $R^2$is limited: in presence of transformations and dynamic terms. First, the plot displays values in terms of the actual dependent variable, be it levels, differences, log-differences or even ratio of the underlying equation variable. Often, using transformations almost completely eliminates any informational value of the plot, as in center panel of Figure H.6. Second, for persistent variables that are modelled with ARMA terms, the actual and fitted values are so close together, that distinguishing them is not possibly in

[27]Some of these nuances would be picked up by model tests. For example, bad fit in later half of sample would show in residuals being systematically larger or systematically biased in that part of the sample, something that model stability tests would likely detect.

Figure 8.6: Examples of actual, fitted, residuals graph



time series plot. Again, very little can be learned from inspecting the plot of actual-vs-fitted values, as can be seen in the bottom panel of Figure H.6. Note, however, that in both cases the time series plot of residuals can still be used to identify observations with abnormally large residuals, even if it is harder than in situations with equations without transformation or dynamic terms.

The varying value of the plot with fitted values, actual values and residuals suggests that one should use it more strategically: rather than using it in *all* situations, one can leverage it in *early stage* of the model building process, when one is still exploring relationships between modeled variable and potential regressors. In such situation one can focus on specifications without dynamics terms and transformations, even if the final specification will be using those. And during such exploratory phase the plot of actual and fitted values with residuals can be invaluable.

As a final note in this section, it is worth mentioning that $R^2$and other model statistics based on residuals have one additional drawback. Specifically, the model estimation has almost always as objective minimization of some function of residuals. This means that including additional regressor without changing anything else will always lead to decrease of residuals, simply because we gave the model additional degree of flexibility. This then means that model statistics which are function of these residuals will always improve when we include additional regressor, e.g., R^2 will always increase in such situation. Of course, this makes these model statistics very problematics from perspective of model selection, since they would always suggest to use the most complex model.

This of course an issue we have already encountered in Chapter **??**, when discussing selection of univariate time series models. There we have also discussed the solution in form of using adjusted R^2 or the information criteria. Briefly, the logic of those alternatives is simple: they use the initial model statistic - R^2 or log-likelihood - and add to it a penalty term that penalizes given model for number of estimated coefficients. This effectively addresses this drawback of these model statistics. It does not, however, solve the other issues discussed in this section.

8.3.2 Multi-step forecasts and measuring forecasting performance

Model residuals and associated model fit statistics provide us information about very short-term forecasting performance, as they are measures linked to one step ahead forecasts. *Do we have at disposal any analogical information also about longer term forecast performance? Do the model residuals, or their associated statistics, capture model performance in terms of multi-step forecasts?*

Unfortunately, the answer is '*Mostly no*'. While it is natural to expect that good one-step forecasts

are indicative of good multi-step forecasts, this is not guaranteed. Since in many applications we are interested in multi-step forecasts, we need to develop different tools for analyzing quality of those. An added bonus is that the tools we will develop do not suffer from the same limitations as the model residuals and associated model statistics, which we encountered in previous subsection. This also means that the tools will be useful even in the context of one-step forecasts.

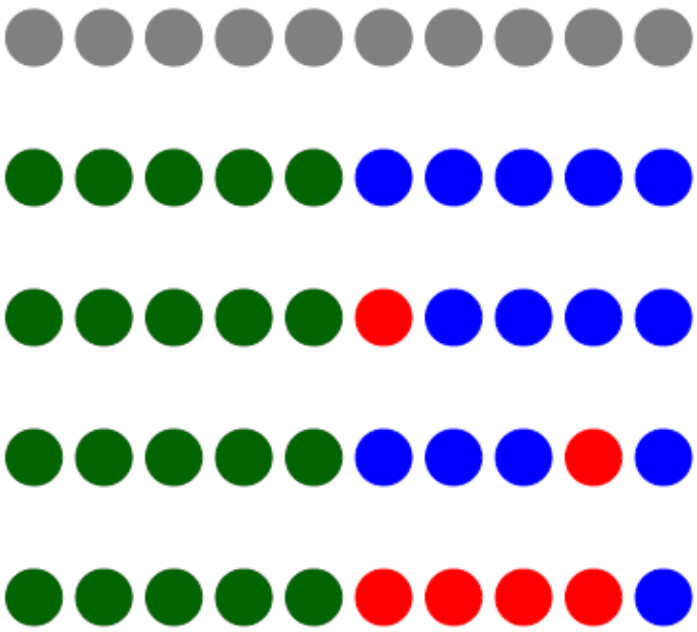
So how do we analyze forecasting performance of our model? Unsurprisingly, the best way to analyze forecasting performance is to analyze the forecasts themselves. Of course, this leads to a bit of a chicken-and-egg problem: to build a good model, we need to know it's forecasts; but to know the forecasts, we would need to have the model. The way to solve this problem is to use the method of **backtesting**, which refers to evaluating the performance of a model using historical data.[28] More specifically, backtesting involves using past observations to obtain forecasts that the model would had made if it was used at given historical periods. We can then compare these forecasts to the actual values that occurred during the same time period to evaluate our model. This will provide us with *approximation* of the *future* forecasting performance of the model.

The starting point of this process is then obtaining the historical forecasts. This can be done in several different ways, which differ in how one treats different parts of the available historical information. In the main text we will focus on the most simple implementation and relegate the discussion of further complications to later sections. Figure H.7 provides visualization of the simplest implementation. The top panel represents all the available observations, with each observation represented by one grey ball. The next panel separates these observations into two categories: the green balls are the ones which were available before period 6, and blue balls are those that were not available before given period. The idea of backtesting is to try to use the information contained in the sample consisting of the green balls to make forecast for the all or part of sample of the blue balls.

Of course, there is the question of length of forecasts. A one-step ahead backtest forecast would amount to forecasting the single red ball in the third panel of Figure H.7. However, more typically we will consider ourselves with either forecasts for multiple periods ahead, or with forecasts lasting multiple periods. Last two panels illustrate both cases for the example of 4 periods. The second-to-last panel shows forecast for 4 periods, or steps, ahead. Of course, if we are interested in forecasts covering multiple forecast horizons, not just one, it makes sense to create forecast covering periods from first

[28]The term should be understood in opposition to term of *forecasting*: backtesting is focused on past instead of future, and hence the preposition *back*. Meanwhile, it is not focused on forecasting, but rather on *testing*, and hence the word testing.

Figure 8.7: Backtesting scheme (one sample split)



period after the start to the last period after the start, something captured in the last panel.

Figure H.7 concerned itself with a single forecast defined by a single set of green balls containing our available information. Of course, we will want to create as many of such forecasts as possible, in order to make the set of observations on which we will evaluate our model as large (and robust) as possible. Figure H.8 takes the last panel from Figure H.7 and expands it to the other possible partitions of the sample. The idea is simple. We start with the minimum number of green balls and make our forecast. We then proceed to next round, in which we add one more ball to green balls (and correspondingly subtract one from blue balls), increasing our available information by one observation. We then create a new forecast, which this time around will start one period later. This will provide second backtest forecast for our model evaluation. We then proceed again in the same fashion, until we reach a point when only one blue ball is left.[29] This will yield a sequence of observations used for model evaluation, with each one corresponding to forecast starting in one successive point in time.

To summarize, here is more formal description of the algorithm we have just described:

1. Starting from the minimum sample size, use your observations $1, 2, 3, \ldots, t$ to initialize your forecast.

[29] Of course if we are interested only in more than one-step-ahead forecasts then we stop sooner. Specifically, the last partition we make is the one which leaves the number of blue balls equal to the minimum length of forecast horizon.

Figure 8.8: Backtesting scheme (multiple sample splits)

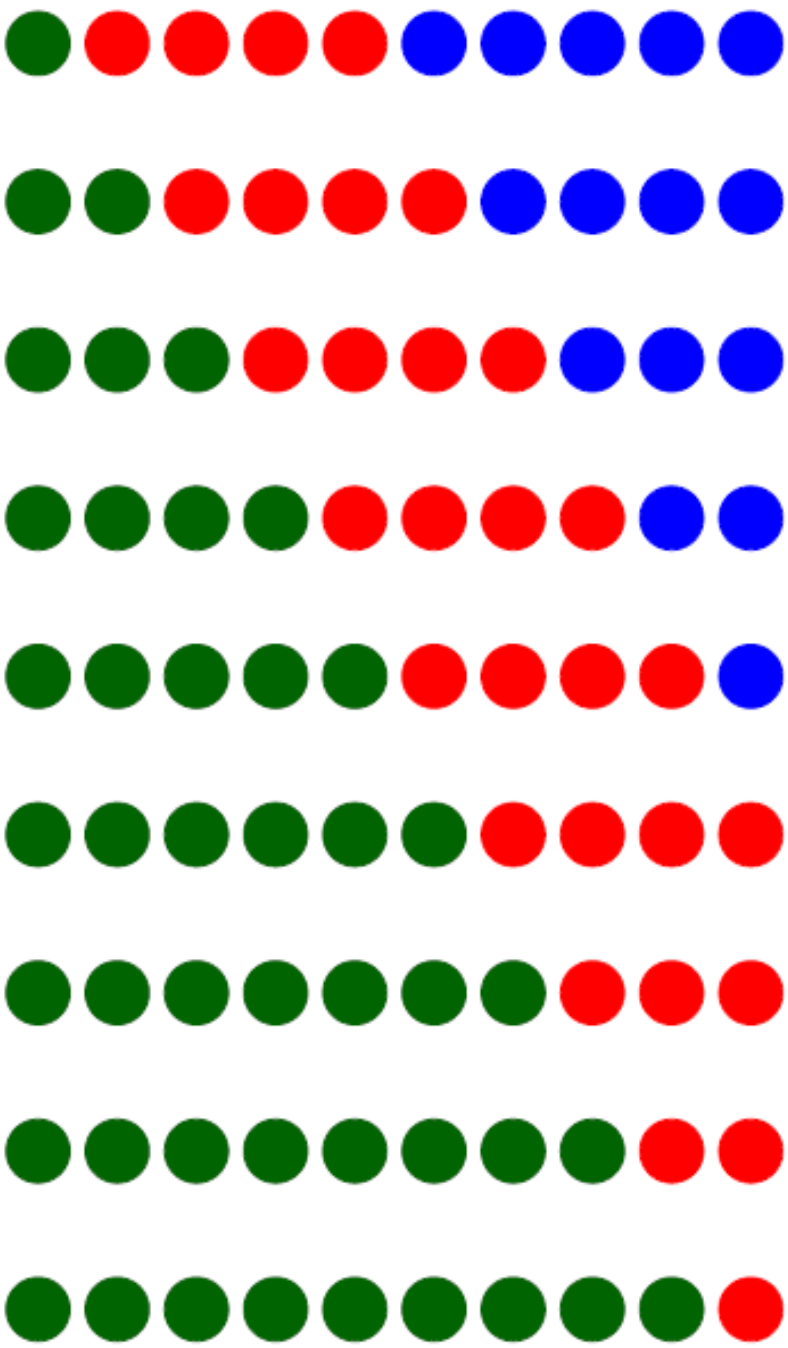
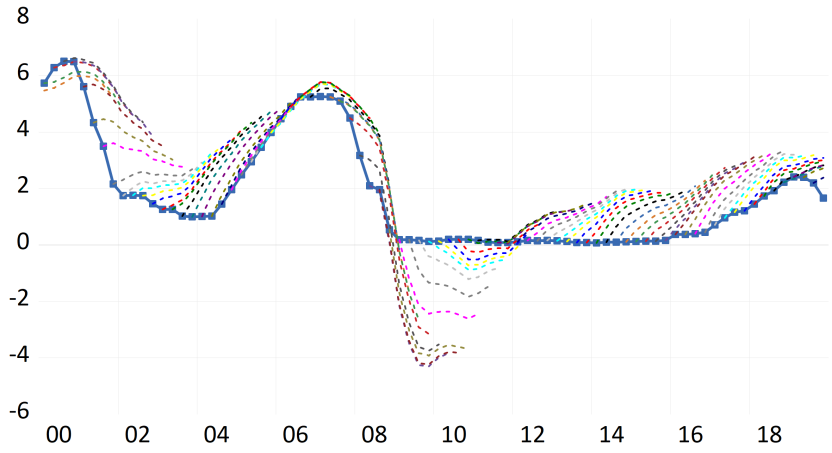


Figure 8.9: Example of output from backtesting scheme



2. Make forecast for periods $t+1, t+2, \ldots, t+h$.
    - For past values of variables use actual observed values (and for errors use values implied by observed values).
    - For future values of dependent variable use forecast values made in previous step (aka multi-step forecasts).
3. Store your forecast and proceed into next iteration be adding one more observation to your initialization sample.

As a result of this algorithm, we will be left with set of forecasts. Figure H.9 shows an example. Blue line with squares captures our time series which we are trying to model. Meanwhile, the individual dashed lines are all historical (backtest) forecasts, with each line corresponding to one forecast (which in turn corresponds to one line from Figure H.9). Therefore, as a result of our work in this section we are left with large number of historical forecasts which we can compare with the actual observed historical values.

Once we have created the backtest forecasts, it becomes a question what we should do with them. Of course, our goal is to get sense of whether the historical forecasts were precise or not. Answer to such question can be provided either in terms of numerical information or in terms of graphical information. The numerical information is more common, mostly because it lends itself to precise answers and exact comparisons. However, as we will see, in many situations graphical information can be superior, for similar reasons we encountered during discussion of one-step-ahead forecasts. We will discuss each in turn.

Numerical information

Once we have our historical forecasts, it is straightforward to obtain numerical information about the overall precision of these forecasts: we simply need to calculate the forecast errors. For single historical forecast we can calculate the forecast error as

$$e_t = a_t - f_t \tag{8.2}$$

where e_t is the forecast error in period t, a_t is the actual observed value in this period and f_t is the forecast for that period. Such forecast error tells us whether our period t featured an upside surprise relative to our forecast, corresponding to positive forecast error, or downside surprise, corresponding to negative forecast error. In this sense it is analogical to residuals, which tell us whether relative to our model given period featured upside or downside shock.

Of course, we do not have a forecast just for single period t, but also other forecast for other periods $t+1, t=2$ and so on. And since there is little sense to judge model based on how well it did for single period - our goal is to know how our model does *overall* - we need to summarize errors across multiple periods. This leads to the classical problem of positive and negative errors canceling each other, which means we need to use function that will prevent that. The most basic function that will does this is the absolute value function.[30] If we use this function, then our measure of overall forecasting performance is given by:

$$MAE = \frac{1}{T-h}\sum_{t=1}^{T-h}|e_t| \tag{8.3}$$

Here we are summing forecast errors over all available historical forecasts, which is defined as number of available periods T minus the length of the forecast horizons h. Each forecast error enters in its absolute value, so that what matters is how far was our forecast from actual value, not whether it was too high or too low. In other words, our measure of overall forecast precision is the average of absolute value of forecast errors for all the forecasts we have. We will refer to this value as the mean absolute error, or MAE for short.

This kind of information is useful on its own as long as modeler has a clear idea what is a good mean absolute error. This is indeed sometimes the case, but more commonly, it is not. In such situation the information is valuable only in comparison with MAE values for alternative models, which are either

[30] Of course, absolute value is not the only function with the desired properties. A common alternative is square function. Later in this section we will return to which functions are useful when.

alternatives that are being considered, or alternatives that are used as a benchmark. The comparison with alternative models allows us to make statements like "Model 1 is better than Model 2" or even "Model 1 is twice as good as Model 2".[31] We simply need to compare the MAE values for the two models.

In model building process we will often consider more than just two models. In such situation the natural thing to do is to put all the models' MAEs into a table so that they can be compared. Moreover, to aid quick comparison, especially when the number of considered models is large, it is useful to color code the numbers. Table H.9 provides an example of such color-coded table, produced by SpecEval. [32]

Table 8.9: Illustration of table with forecast performance metric values for several models

Description	**MAE**
Model 1	0.31
Model 2	0.49
Model 3	0.61
Model 4	0.28
Model 5	0.60
Model 6	0.75
Model 7	0.31
Model 8	0.28
Model 9	0.53

This concludes the basic introduction of the numerical information on forecast performance. The additional discussion will cover many modifications and intricacies of this kind of information. However, before we turn to graphical information we need to address the issue of multi-step forecasts. So far in our discussion we have considered the situation that we have one forecast for period t which in the s in the simplest case of one-step ahead forecasts would be made in period $t-1$. Meanwhile, forecast for period $t+1$ would be made in period t, and so on. But if we consider multistep forecasts forecast, then forecast for period t could be also made in period $t-2$ (two-step forecast) and $t-3$ (three step forecast), and so on. In other words, when considering multistep forecasts we have multiple forecasts for the same period of time which differ in their starting period.

[31] There is also a formal way to test such statements mathematically. Specifically, the Diebold-Mariano tests can be used to formally test whether set of forecasts from one model is statistically significantly smaller than corresponding errors from another model. See XXX for more details.

[32] The exact color-coding scheme is to some degree an arbitrary choice. It can be done based on distance from the mean value across models, much like Excel does color coding. However, this can cause problems when the set of models includes one or few outliers with abnormally large numbers. In such situation all the low values will have similar color and won't be distinguishable from each other. An alternative employed by SpecEval is to instead rely only on ranking, so that the colors capture ordering rather than magnitudes. By including a detailed color scale, it still highlights which model is best.

Of course, forecasts made over different forecast horizons are not directly comparable. Typically, making short horizon forecasts is easier than making long horizon forecasts, and hence we cannot mix these together. Moreover, models are often used for multiple forecast horizons. For example, often we care about value of our variable *over the course* of next 8 periods, rather than just its value *in* 8 periods from now. In such situation we need to evaluate the model for multiple forecast horizons. Table H.9 then turns into table H.10.[33]

Table 8.10: Illustration of table with forecast performance metric values for several models and horizons

	Forecast horizons (# of steps ahead)					
Description	**1**	**3**	**12**	**24**	**48**	**Avg.**
Model 1	0.19	0.2	0.24	0.3	0.31	**0.24**
Model 2	0.26	0.27	0.33	0.41	0.49	**0.36**
Model 3	0.17	0.19	0.29	0.41	0.61	**0.35**
Model 4	0.088	0.2	0.25	0.3	0.28	**0.22**
Model 5	0.084	0.21	0.44	0.51	0.6	**0.38**
Model 6	0.091	0.2	0.35	0.5	0.75	**0.4**
Model 7	0.058	0.1	0.24	0.32	0.31	**0.19**
Model 8	0.051	0.093	0.22	0.3	0.28	**0.18**
Model 9	0.051	0.1	0.24	0.36	0.53	**0.27**

Note that the table includes only selected forecast horizons, as in our example we are considering horizons ranging from 1 period to 60 periods and including 60 numbers would be impractical. In such situations one should think about particular forecast horizons as being representative for horizons of similar length. Typically, you will want to include forecast horizons which are representative of short, medium and long forecasts.

Also note that when investigating forecast performance for multiple forecast horizons then often encounter situation when one model outperforms another model for some horizons, but not for all horizons. Consider for example Model 7 and Model 9, with Model 9 outperforming Model 7 for the 1-period forecast horizon, matching for the 3- and 12-period horizons, and performing substantially worse for the two longest forecast horizons. In such situation either the average over forecast horizons can be used as guiding principle, which is why the Table H.10 includes the average in the last column. Or the model builder needs to make a decision on which forecast horizons are more important than others. This will often be the case.

[33]There is an alternative to having one number per forecast horizon. We could do double averaging: first calculate MAE for one given forecast for all the 8 periods; and then average it over all the 8-period forecasts. While strictly speaking this is valid approach in practice it is rarely used. The reason seems to be the poor statistically properties of such statistic. Of course, in case we are not performing statistical tests, this should not matter much.

What we have outlined in this subsection is a powerful way of gauging whether given model has good historical forecast performance or not. Such process will give you a good idea which model would have performed well in history, and how well would it perform. It takes just a small leap of faith to use this information to also make inferences about performance in also future. How big is this leap of faith is something we will return to in additional discussion at the end of this section. However, before we proceed it is important to highlight different issue: A MAE is a summary statistic of forecast performance. By design it averages across the whole sample. Of course, as we discussed earlier, such summary statistic can hide important heterogeneity across the sample.

There are two ways to address this. One way is to stick to a summary statistic but to select few sub-samples of interest and compute the summary statistic also for those sub-samples. This means that instead of having one table like Table H.10 we will have several such tables, one for the whole sample and one for each sub-sample. A typical choice of sub-samples for macroeconomic data are periods of recessions or other types of crisis such as the Great Recession of 2008-2009. The idea is that we want our model to perform well in particular types of situations in future and hence we pay additional attention to its performance during such situations in history.

While this is a possible solution, especially when we are considering only few such sub-samples, there is potentially a better alternative. Moreover, this alternative has additional advantages. We now turn to this alternative.

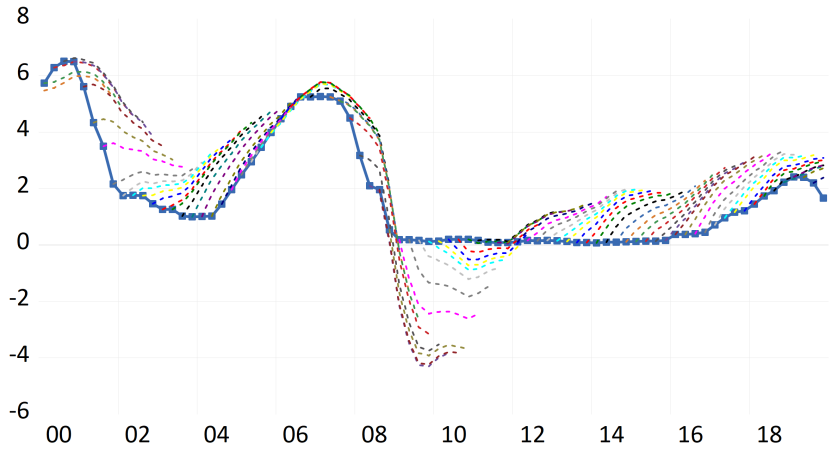
Graphical information

An alternative way of evaluating forecast performance is to approach the problem graphically rather than numerically. The idea is to take the historical forecasts and put all of them into a graph together with the actual historical series. Figure H.10 provides an example.

Inspection of this figure allows us to very quickly get a sense of whether our model would have produced accurate forecast in history or not. Note that in contrast to numerical value, the graphical information lends itself more naturally to absolute interpretations of "good vs bad" forecasts, rather than just relative "better or worse than", since we can get immediate sense of (lack of) precision. Of course, it also comes with its limitations. Mainly, such judgments are in their nature rather rough, and there is no way to do quick comparison across more than few models.

Ultimately, there are important advantages of graphical analysis over numerical analysis, and vice versa, which mean that the numerical and graphical methods should be both used as they are comple-

Figure 8.10: Example of output from backtesting scheme



mentary. This relies on the same logic as the logic behind the value of model building in comparison to automatic model selection. As we mentioned in end of section on numerical information, the main drawback information is its inherent nature as a summary statistic: It can hide a lot of important heterogeneity. And this is where graphical information can help greatly: By visualizing all the forecasts together it is immediately clear whether the model works well always, or just in some periods; and what are the periods when it does not work well.

Even more importantly from the perspective of the model building process, the information on heterogeneity in the forecast performance can be crucial source of ideas for improvements of the model. Does your macroeconomic model perform poorly during 2008-2009? Then probably we are missing some factors relating to financial stress. Or does it perform poorly during the sovereign debt crisis? Then you might be missing factors related to stress in government bond markets. Alternatively, it can help you figure out when to disregard poor forecasting performance. Does your model work poorly during the pandemic? Then it means that it's structure does not apply to this highly unusual period. While this might mean you need a different model, often this is not the case, assuming you do not plan to use your model for such unusual periods in future.

Similarly, it is not just the fact that forecast performance during particular period is bad that is informative. Even more informative can be the knowledge of in what way is it bad. Is it bad because your forecast does not drop as much as the actual values? Or is it bad because it drops too much? Both of these situations have potentially different implication for your model building process. Ultimately,

the key idea is that graph of your forecasts contains a sea of useful information compared with a single MAE number (or even several MAE numbers). And this sea of information can *sometimes* be usefully leveraged. Of course, there is a flip side to this statement: sometimes a sea of information is just too much. And that is why numerical information also plays a very useful role.

As an example of usefulness of graphical information take Figure H.10 from beginning of this section, which showed historical performance of model for U.S. monetary policy rate (the federal funds rate). In grand scheme of things the model seems to be doing relatively *ok* job, with forecasts decreasing in periods when actual values did, and increasing when actual values did. But there are specific periods when it is not doing a *great* job. One such period is from 2012 to 2015 when the model kept predicting increase in policy rates while in reality they stayed at 0. This clearly calls for investigation and potential adjustment of the model.

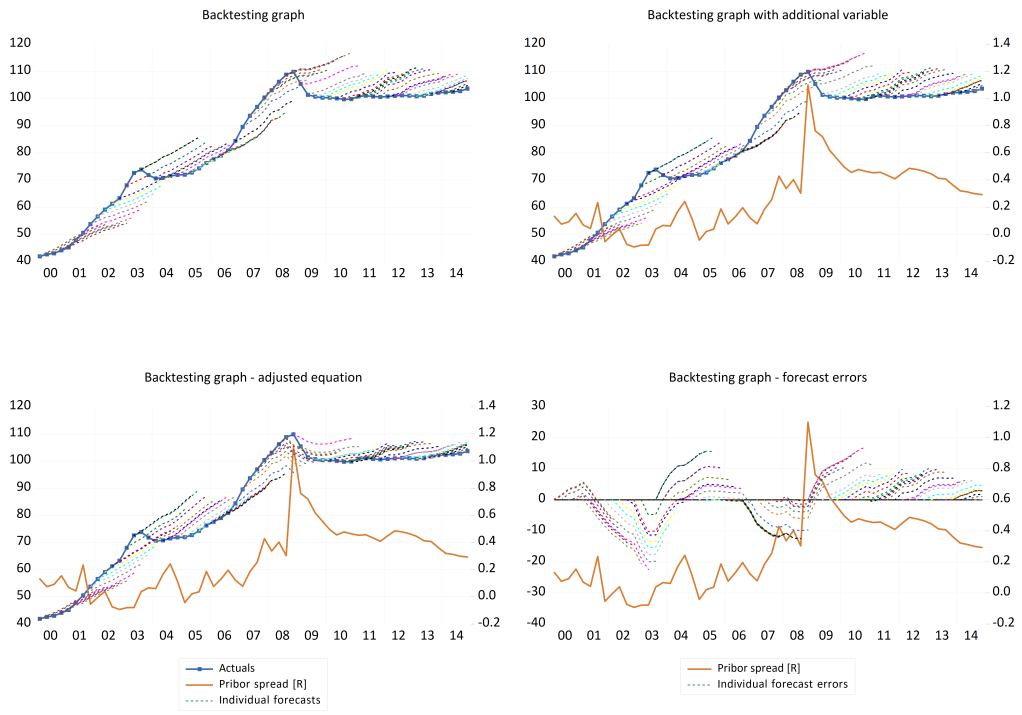
Meanwhile, the historical performance is also suboptimal during period of 2009 to 2010, when model predicts policy rates that are deeply negative. In contrast with the previous shortcoming, this one is much less likely to call for outright model adjustment: The central bank actually wished to lower the policy rates, but was prevented by the zero lower bound on interest rates. Rather than adjusting the model to prevent it from forecasting deeply negative interest rates, at the potential cost of making the forecast worse in some other part of the sample, it is better to keep the model intact and account for this constraint outside of the model.[34] Note that this kind of nuances are impossible to conclude form numerical information alone.

Building on this perspective it is good idea to talk about few ways in which this graphical analysis can be augmented to make it even more useful. One simple trick is to include additional variables in the chart in order to see whether periods of bad forecasting performance correlates with movements of such variable. This a useful way to explore whether given variable or possibly its transformation could be a way to improve the model. For example, top left panel of Figure H.11 shows forecast performance for model for Czechia house prices. It can be seen that the model is struggling with forecasting house prices during the period from late 2008 to 2014: It first fails to forecast drop in house prices and then consistently forecasts rapid increases, while house prices actually stagnated.

In such case our job as modelers is to figure out how to improve the model so that its behavior in this period is improved, possibly by including additional variables. Top right panel of the figure shows how in exploratory stage one can benefit from simply including a candidate variable in the graph with

[34]For example, one could enrich the single equation model by including additional identity that would take the form of zero lower bound.

Figure 8.11: Example of backtest graph with added variable



historical forecasts. In addition to the actual value and the individual forecasts, the graph also includes money market rate spread, a measure of stress in banking system. This reveals two things. First, it shows that the period with extreme stress in banking system coincides with period when house prices dropped (2008-2009). Similarly, the period of elevated stress in banking system in 2009-2014 coincided with period of stagnation in house prices. This is a first signal that given variable could be useful for forecasting house prices. The second signal is in form of correlation between given variable and periods when our model's forecasts have subpar quality: it is exactly when the variable is elevated that our model has tendency to forecast too high growth. Confirming these two observations, the bottom left panel of Figure H.11 shows that once the variable is included the performance in given period improves significantly.

This example showed that the inclusion of additional variable in our chart can be done for either of two purposes. Either as a way of checking whether movements in this additional variable are correlated with movements in our dependent variable. Or to check whether they correlated with forecast errors in particular part of the sample. For this second analysis it is sometimes useful to use graph that captures the *forecast errors* rather than the forecasts themselves. This is shown in bottom right panel

Figure 8.12: Example of backtest graph with growth transformation

of Figure H.11, in which is easier to see that the forecast errors are positive when our variable is high.

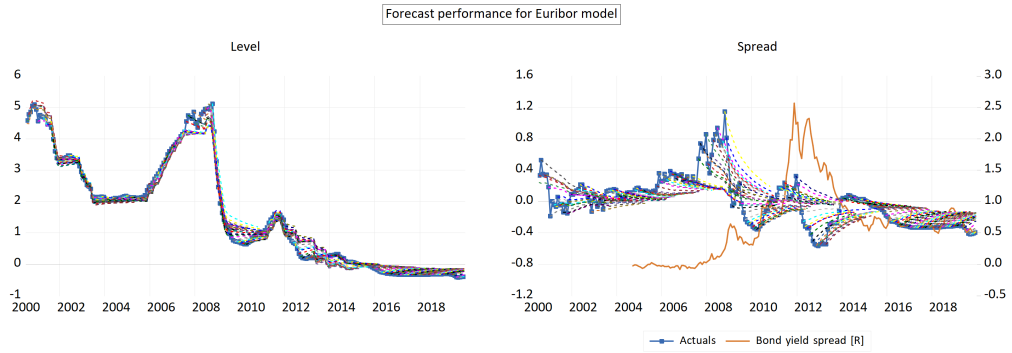
Another useful way to adjust this graphical analysis is to vary the transformation of the forecasts. It is not by necessity that we should be looking only on the level of our underlying base variable. Often times looking at other transformation is more useful. At the most basic case often looking at growth rates rather than (or in addition to) levels might be more revealing. Figure H.12 continues in the previous example, but this time shows growth rate of house prices, both for actual values and forecasts from our model. Given that house prices are a trending variable, the growth rate is more revealing about their movements as can be seen in the figure.[35]

Similarly, for many financial variables it is spreads that are most useful unit of analysis. For example, in left panel of Figure H.13 we capture the forecast performance for model for Euribor interest rate. Base on this graph the model seems to be performing really well. However, this is misleading, because vast majority of variation in the Euribor is due to variation in monetary policy rates. To show this the left panel includes also the relevant policy rate. As can be seen, the two series move almost perfectly together. This means that if we use these the policy rate as our explanatory variable and then visualize our forecasts in terms of level of Euribor, we are effectively setting for ourselves too low bar.

This is confirmed by looking at right panel of Figure H.13. That panel shows the same variable and same set of forecasts as the right panel, but this time we are showing the spread between Euribor and the policy rate. In this view the model still does a good job in 2007-2008 period, but not such a good job in period 2011-2012, when the forecasts for spread fail to increase. In other words, by zooming

[35]In principle we could model growth rate of house prices, rather than level of house prices, and then the question of transformation would become moot. But this misses the point. The point is that we benefit from having the flexibility to choose transformation in which visualize our forecasts. In this view limiting ourselves to only growth rate transformation would be suboptimal. Similarly, not always do we model trending variables in growth rates, and ye we might benefit from the ability to visualize growth rates even in such case.

Figure 8.13: Example of backtest graph with spread transformation



in on behavior of Euribor *relative to its main driver*, we are able to uncover periods of suboptimal performance that were not visible when looking at level of Euribor.[36] And by identifying this period of suboptimal forecasts we can also immediately start focusing on them. For example, the right panel includes one variable that might help us with forecasting the Euribor spread: the spread between average Eurozone bond yield and German bond yield, which can be considered a measure of stress in Eurozone bond markets. The panel shows that this variable increases during the period of 2011-2012, exactly when our model fails to produce an increase in the Euribor spread.

There are other transformation that can be useful, like index or ratio, but the idea is the same as in above examples: Simply, we should use the transformation(s) that will be most informative about the quality of forecast from our model. Note that this is independent of the transformation actually used for the dependent variable in the model. For example in many models one uses log-differences as the dependent variable, but rarely if ever would one want to be visualizing log-differences. Instead, using the closely related growth rates would be preferable, as the numbers will be easier to interpret.

The bottom line is that the graphical information about forecast performance can be augmented in several ways. And these kind of augmentations are a prototypical example of model building process which is centered on the interactive and iterative model improvement. The idea is at each step in the model building process the modeler is trying to learn something about the data and about the models being considered. And not only is this learning typically easier based on graphical analysis; but this kind of learning often requires looking at the data in multiple different ways, which is possible with augmentations of graphical information.

[36] In a sense this is a matter of axis range. If for example policy rates would not move a lot in our sample then the suboptimal forecast performance would be visible even when graphing the level of Euribor. An alternative to graphing spread would be graphing the forecast errors directly, which would show that in the period 2011-2012 they are larger than in other periods.

8.3.3 Additional discussion

In-sample vs. out-of-sample forecasts

Previous discussion was intentionally vague on one important question.We have used the wording of *initializing forecast* based on the data that would have been available at given time, without really saying what it means. It does clearly mean that we are not using data on dependent variable past that point in time. But what does it mean about the model itself? One crucial question is on which part of data should our model be estimated when we are creating historical forecasts.

One rather obvious option is to use our final estimated model. This means that we estimate the model coefficients on the full available data sample and then use this estimated model to create the historical forecasts. Of course, this is in some sense cheating, because our model coefficients are based on data which would not be available to econometrician at the time of making the forecast. So the alternative option is to retain the model structure, but to estimate the model coefficients only on the data available before the start of the forecast. The motivation here is to approximate the situation of the forecaster in given historical time.

We can illustrate this two options by returning to Figure H.8 which illustrated our backtesting algorithm. The latter option would mean that we estimate the model coefficients using only the green balls. In that way the information contained in blue balls does not influence our model estimates in any way. In contrast, the former option would mean we would use both the green and blue balls to estimate the coefficients, so that part of the information contained in blue balls is used in making forecasts for them.

The difference between the two approaches then rests on the question whether the estimation and forecast samples are completely separate or whether they overlap. This then yields the **in-sample vs. out-of-sample terminology**. The sample to which the terminology refers is the *estimation sample*, and hence the full description would be forecast evaluation *inside of the estimation sample* or *outside of the estimation sample*. And hence if we use data available only before the start of the estimation sample then we are creating out-of-sample forecasts, while if we use the full available dataset for estimation, then we are creating in-sample forecasts.[37]

These two approaches are different in their philosophy. The **out-of-sample forecasts** aim to approximate how well would forecaster be served in given historical period by the knowledge of the model *structure* linking the dependent variable to independent variables. It is a hypothetical thought

[37]The options are not limited to these two options. For more details on alternative schemes see XXX.

exercise that imagines that you have traveled back in time but all you brought with you was the model in its general form without the estimated coefficients. You would then estimate the coefficients on the data you would have available at that point in time and make forecasts.

In-sample forecasts perform a slightly different hypothetical thought exercise. In contrast to the out-of-sample thought experiment, in this experiment you are allowed to bring with yourself not only the model in its general form, but also the coefficients that were estimated on data that would not be known at given historical period.

Effectively, the two approaches try to answer slightly different questions, focusing on evaluating a different objects. In case of out-of-sample forecasts we are asking about the usefulness of the *general model structure*, which to some degree is independent of the data. In contrast, in-sample forecasts are evaluating the *actual estimated model*. Since the coefficients are estimated on the specific dataset, the resulting object is data specific, while model structure is not. This then means that by evaluating model in-sample we answering the question *'Does your estimated model explain observed historical data well?'*. This is related but somewhat different question than '*Is your model a good model in the sense of approximating well the true data generating process*'. [38]

So which question of these two questions should we focus in **model development**? Well, both. As a starting point it seems logical to focus on the question answered by in-sample forecasts; after all, in future forecasting exercises we will be using the final estimated model, so we should be interested how this model works. However, things are not that simple. Ultimately, we are interested in building a forecasting model that is going to be forecasting well *in the future*. The key thing to realize is that the in-sample forecasting performance will be a downwardly biased measure of the future forecasting exercise, and potentially more so for some model specifications than for others. This is due to the simple fact that a model estimated on a given historical sample is to some degree optimized to work well in that sample. Once it encounters future data on which it will not be optimized, it will work worse. In most extreme this is the problem of overfitting, which we discuss below.

So if we want to know answer to question '*How well will the model work in forecasting future observations?*', then we need to focus on out-of-sample forecasting performance. That said, the answer provided by out-of-sample forecasts can be rather noisy. This is because we are basing our evaluation of versions of our model estimated on subset of our data, and there is the possibility that the subset

[38] Of course, the difference between out-of-sample vs in-sample exercise is not that sharp: Unless the model under consideration is suffering from problem of overfitting discussed below then the explaining observed historical data and approximating the true data generating process should be very closely related.

of data is insufficient to give us reasonable coefficient estimates. This leaves us with a trade-off between the in-sample and out-of-sample forecasts in the context of model development: in-sample performance leverages all available information efficiently, but leads to potentially overly optimistic and hence misleading results; meanwhile, the out-of-sample performance provides on average better signal, but the signal can be very noisy.[39]

The good news is that the choice is not either/or, since we can leverage both types of exercises, with possibly different importance assigned to each, or with each exercise used in different stages of the process. For example, in model building it is common to start with analysis of in-sample forecast performance, which is more robust and hence suitable for exploratory stage of model building when large number of models are being considered. Once the model space has been narrowed down, one can shift focus to out-of-sample forecasting performance. Of course, different situations will call for different choices with respect to the importance of in-sample vs. out-of-sample performance.

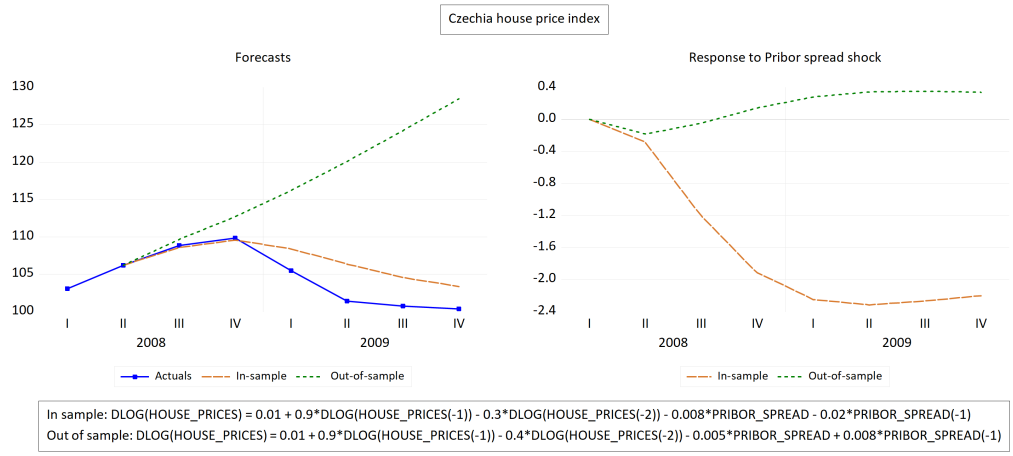
Apart from these perspectives focused on the model development, it is important not to forget that forecast performance is not only used in model development. Instead, it is often also used for the purposes of **model evaluation** and **model understanding**. And here, the case for in-sample forecasts is further strengthened. This is for two reasons. Philosophically, the problem of overfitting is much diminished once model development is done and we switch to model evaluation. And it is not present at all in context of model understanding. Fundamentally, given that it is the final estimated model which will be used in future forecasting, then it makes much more sense to use in-sample forecasts to study and evaluate its behavior, rather than out-of-sample forecast.

All these points are illustrated in Figure H.14, left panel of which shows two forecasts from model linking Czechia house price index to measure of financial stress, the Pribor spread. One forecast is done in-sample, one forecast is done out-of-sample. The bottom of the figure also includes the two estimated equations used to create the forecasts, showing how they feature very different coefficients on and hence effect of Pribor spread, something that is illustrated also in the right panel.[40] The equation used for in-sample forecast features large negative coefficients and correspondingly very large negative effect of Pribor spread on house prices. In contrast, the equation used for out-of-sample forecasts features small positive effect of Pribor spread on house prices, corresponding to sum of coefficients that are positive. This is difference should not be surprising, given that the estimation sample used

[39]Note that the trade-off has the form of trade-off between bias and variance that is omnipresent in econometrics: in-sample is biased measure with lower variance, while out-of-sample is unbiased measure with higher variance.

[40]The right panel shows the effect of one-period one-standard deviation shock to Pribor spread. More details on the nature of this graph is provided in section on sensitivity to shocks.

Figure 8.14: Illustration of difference between in-sample and out-of-sample forecasts



for the out-of-sample forecast did not include period of financial stress.

Unsurprisingly then the out-of-sample forecast is rather poor, with house prices failing to decline at all despite rise in Pribor spread during given period. The question is how much should one count this against the model in light of the fact that in-sample forecast is pretty good. Sure, having this model structure back in 2008 would not have been useful. But that does not mean that the model is a bad model, especially going forward. And while the in-sample forecast does flatter the model in terms of forward-looking performance, it still suggests that the model should be usefull *in similar future situations*. Simply, while the in-sample forecast is biased signal of quality of the model, the out-of-sample forecast in this case is a very noisy signal.

Similarly, Figure H.14 speaks to the relative value of in-sample and out-of-sample forecasts in model evaluation and understanding: it is not clear what can one take from the bad out-of-sample forecast during great recession in terms of evaluating the model and understanding how it behaves, given that the model used to create the forecast is very different from the final estimated model, and the final estimated model actually works well. Moreover, from perspective of model user, rather than model developer, it is the in-sample forecast that contains useful information, while the out-of-sample forecast would if anything be misleading: Based on the in-sample forecast the model user can conclude that the model features a significant response to financial stress; in contrast, the out of sample forecast would suggest that there is no response at all.

To summarize, both in-sample and out-of-sample forecasts have their informational value, be it in model development or model evaluation. This is because they each answer similar but somewhat

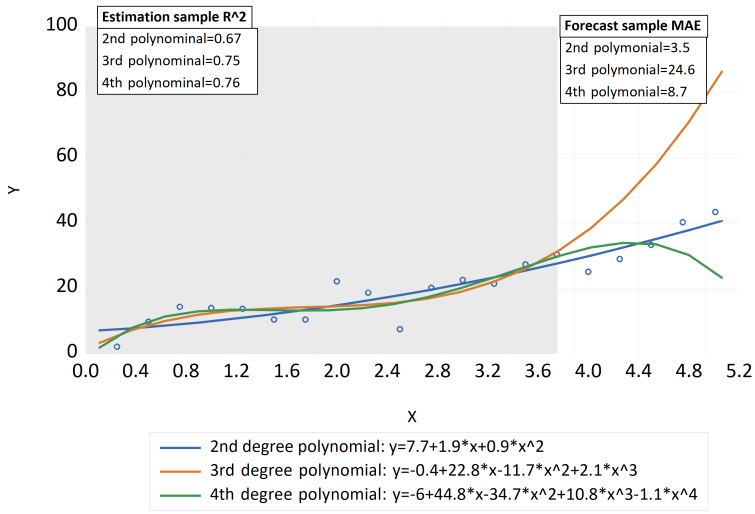
different questions. Simply, if you want to know how precise will be forecast from your model in future, and/or your model development process is solely based on optimization of such performance, then you should focus more in out-of-sample forecasts, as otherwise you are in danger of overfitting. On the other hand, if you want to know how does your estimated model behave in different historical situations then you should use in-sample forecasts. If you are doing something in between, then you should leverage both. In particular, since model building is a lot about evaluating and understanding when and why the model works, and when it does not - and not that much about optimizing forecast performance - then in-sample forecasts are much more suitable in our model development processes. This is simply because we are at less of a danger originating in overfitting, to which we turn now.

Overfitting. As we mentioned few times, the choice between out-of-sample and in-sample forecast evaluation is closely related to the problem of overfitting. To appreciate this problem, return back to our discussion of in-sample forecasts as a measure of future forecasting performance. As we already mentioned earlier, in-sample model performance is a biased measure of future forecasting performance. This is because of the nature of the estimation process: all estimation techniques are aimed at maximizing some measure of the model fit. Therefore, the forecast performance will always seem better when done inside of the estimation sample, simply because the goal of estimation is to make the model correspond to the data on which it is being estimated. However, in doing so, they invariably lead to coefficient estimates that are influenced by the data we have, and hence reflect their idiosyncratic aspects, in addition to the general systematic relationships the model is meant to capture. Of course, once we use the estimated model on data outside of the estimation sample, which will feature different idiosyncratic aspects of data, the forecasting performance will inevitably be worse than inside of the estimation sample.

How does this then cause the problem of overfitting? When model development is primarily driven by historical forecast performance, and when these forecasts are created within estimation sample, we are in danger of picking model that works well on the historical data, but will not work well on future data. Simply, we have overfitted the model to the idiosyncratic aspects of the historical data, its noise or random fluctuations, rather than capturing the underlying patterns and trends.

To visualize this, consider Figure H.15, where data were generated according to a square function, $y = x^2 + \epsilon$, where $V(\epsilon) = 400$, giving a parabola pattern, but with a significant amount of noise. Given that we know that the data were generating by the square function, we know that the model should not include any additional transformations. However, the data would be better fitted with model that

Figure 8.15: Illustration of overfitting



also includes higher order terms, such as x^3 and x^4. But this would be overfitting the model, as the additional terms are picking up idiosyncratic aspects of the data, rather than the true relationship. Correspondingly the forecast outside of the range of our estimation data would be worse than if we would not include these additional terms.

This problem can be guarded against by using out-of-sample forecasts.[41] Indeed, the problem of overfitting is why there is usually a dominant focus on out-of-sample forecasting: most forecasting textbooks take it for granted that one should use out-of-sample forecast performance when doing model development, and then fail to even mention in-sample forecasting, despite its potential value in development and elsewhere.

The focus on out-of-sample forecasts is appropriate when model development is primarily centered around historical forecast performance, and especially when such performance is being optimized by algorithmic model selection, in which case using in-sample forecast performance would clearly lead to selection of overfitted model. But as we already suggested above, this conclusion does not necessarily generalize to other development strategies. In context of model building, which is different from simple automatic model selection, the dominant focus on out-of-sample forecasting is not so clear cut. Not only is the likelihood of overfitting much lower when selection is not done algorithmically. Even more importantly, model evaluation and model understanding, for which in-sample forecasting is more appropriate, are much bigger part of the process.

[41] Albeit only to some degree, especially in context of time series modelling, as discussed below.

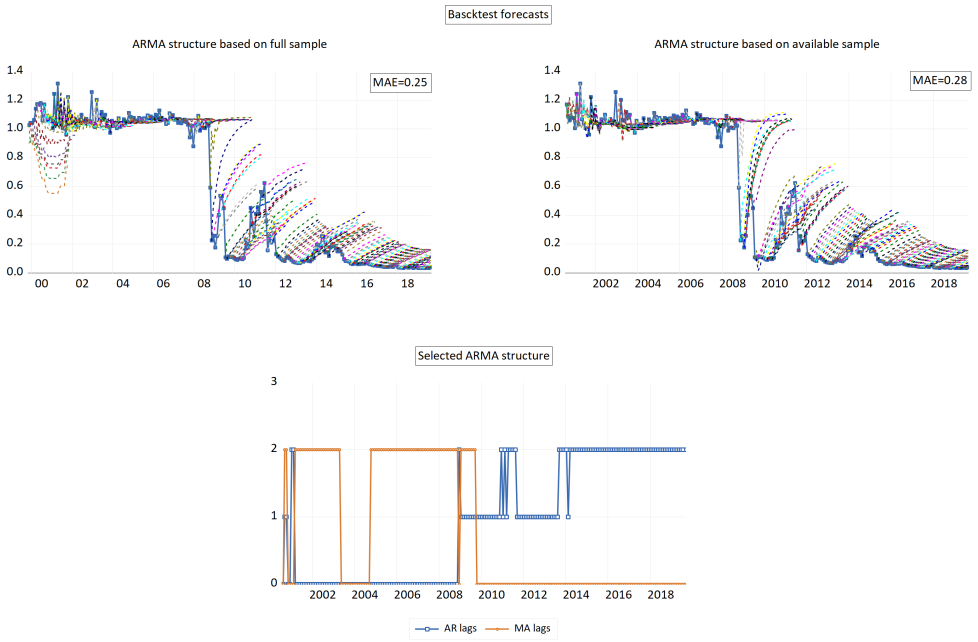
Moreover, the case for dominant focus on out-of-sample forecasting becomes even weaker in situations when multi-step forecasts are the primary interest of the model. The overfitting problem arises mostly due to estimation maximizing the fit of the model. But this in almost all the cases means maximizing the model fit in terms of one-step-ahead forecasts; for example, the usual OLS procedure minimizes the sum of squared residuals which are the difference between actual values and one-step-ahead forecasts. While precision of one-step-ahead forecasts and multi-step forecasts is likely positively related, the link between those can often be rather weak. So the fact that the former is being maximized does not mean that the latter is. Therefore, you are much less likely to end up with overfitted model when you rely on in-sample *multi-step* forecasting performance when developing your model, than when you rely on in-sample *one-step-ahead* forecasting performance.

True and pseudo out of sample forecasts. The out-of-sample thought exercise imagined that you have traveled back in time and brought with you only the model structure, but not the model coefficients; it is out-of-sample in that the forecast are made outside of the estimation sample. However, note that the thought exercise still imagined you did bring something back into the past, your model structure! While this might at first sight seem innocuous, it is not without its significance, because model structure can be heavily influenced by knowledge of developments not known back in history. In other words, the model structure is an embodiment of future knowledge. For example, before the Global financial crisis of 2007-2009 most macroeconomic models would not include indicators of stress in financial markets because such stress was considered to be mostly unimportant to macroeconomic fluctuations.

For this reason the proper full term for the forecasts created in such way is *pseudo* out-of-sample forecasts, which is in contrast to *true* out-of-sample forecasts: the true out of sample forecasts do not reflect information about future developments in any way shape or form, not even in the structure of the model. Of course, the only true way to achieve this from model building perspective is to create future forecasts and wait for the future to happen.

One way to illustrate this point is consider the example of evaluating forecasting performance of ARMA model for the case of Euribor interest rate. Imagine that you have run automatic ARMA model selection, yielding an ARMA(2,0) model. Now you want to evaluate how well does that model work in terms of forecasting. Trying to get proper answer to such question you of course perform out-of-sample forecast evaluation, which means that you always estimate your ARMA model only on data available before the start of each forecast. Top left panel of Figure H.17 gives you the results.

Figure 8.16: Illustration of pseudo out-of-sample forecasting



However, the question is whether this exercise really provided you with the answer to your question. Why? The point is that your ARMA model order was chosen based on the whole dataset. But if a modeler would be performing the automatic model selection without the knowledge of later datapoints, he might have ended up with different ARMA model. This is indeed confirmed in the bottom panel of Figure H.17, which shows the ARMA orders chosen by automatic ARMA selection mechanism at any given point in time based on the data available up until that point in time. As can be seen, the ARMA model chosen varies over time very significantly: Initially the algorithm chosen a MA model of low order, but later it switches to AR model structure. This highlights how the model structure has embodied the knowledge from future, which would not be available back in the time. Top right panel then illustrates that this that such knowledge is indeed valuable and skews our results.[42] When we determine the model structure solely on data available at each point in time, the overall forecast performance becomes significantly worse.[43]

Relation to cross-validation methodology and terminology. This chapter is dedicated to topic of time series modelling, even though some of the aspects also apply to cross-sectional models.

[42]Note that this does not have to be the case: Using model structure that changes with estimation sample has the added advantage that for earlier parts of sample it might be significantly better, outweighing the drawbacks.

[43]SpecEval allows you to perform both types of exercises, one where the ARMA structure is fixed, and one where it varies.

In that field there is well established methodology and terminology of cross-validation: a model is estimated on part of the dataset (typically 80% of the full dataset), and then tested on the remaining part of the dataset. The two parts of the dataset are commonly referred to as "training" and "test" data/samples, or simply as estimation dataset and "hold-out dataset", where the latter is indicating that the data where held out for the purpose of evaluating the model. One can also encounter the terminology of "in-sample data" and "out-of-sample data", which makes it clear that the concepts are analogical to what we have encountered in our discussion, just that the terminology is different.

However, the fact that cross-validation methodology is conceptually similar to in-sample and out-of-sample forecasting does not mean that one can use the methodology suitably for time series modelling. The primary problem lies in the division of the dataset into training and test data. In cross-sectional econometrics this division is typically done randomly, with 20% of observations being randomly selected and removed from the model estimation/selection stage. In context of time series models this is not feasible, given that the order of observations is important: for estimation and forecasting purposes we require continuous samples (unless we use only static models).

As a result one sometimes encounters implementation of cross-validation methodology to time series model development that selects first 80% of observations as the training data and the remaining 20% as the test data. While this might seem sensible, one runs the huge risk of testing his model on data that are not completely representative of the whole data set. Most typically this problem arises in macroeconomic time series, or for that matter any time series dependent on macroeconomic conditions, where the test sample might include only period of relative calm in terms of macroeconomic developments. For example, one might end up testing his model only on period which does not include any recession. Similarly, in financial data one might test only on observations that do not include any periods of financial stress. In either case, the methodology leads to a training dataset that is not representative of all the possible conditions.[44]

This is the reason why our approach relied on different backtesting scheme: we tried to evaluate our model on as many observations as possible. Indeed, some authors refer to this approach as "time series cross-validation", see e.g. XXX. This suggests that cross-validation methodology is not completely unsuitable for time series model development. That said, the terminology of train and test samples lends itself less suitably for the discussion of in-sample vs. out-of-sample model evaluation: it sounds

[44]Of course one could shorten the train sample to make the test sample much longer. However, that runs into the problem of sensitivity of coefficient estimates to estimation samples: if the training dataset is very short then your model estimates might b very different from the final model estimates and you might be evaluating model that is behaving very differently from your final model. This again relates to the bias-variance trade-off discussed above.

strange to say that you will evaluate your model on your train data, given that those are just the train data. In other words, the cross-validation terminology nudges the modeler towards using the out-of-sample model evaluation, and in-sample model evaluation is typically not even mentioned.[45] For this reason this chapter relied on the in-sample vs. out-of-sample terminology.

Ex-ante vs ex-post forecasts

The question of in-sample vs out-of-sample forecasts is all about on what sample do we estimate our forecasting model. However, when creating multi-step forecasts from multivariable models, this is not the only question we need to grapple with. We also need to determine which values will we use for future observations of our right-hand side variables.

This might sound strange, but upon consideration of example it should become pretty obvious. Imagine you are forecasting a variable y, say inflation, based on model that links it to variable x, say unemployment rate. And say that the model postulates that current value of inflation depends on previous value of unemployment rate, as in:

$$inflation_t = \beta_0 + \beta_1 unemployment_{t-1} \tag{8.4}$$

When you want to create forecast for inflation for next period, $inflation_{t+1}$ you just use the current period value of unemployment rate, $unemployment_t$, which is known already. But what if you want to create forecast multiple periods into the future? In such situation you encounter the problem of future values of independent variables. For example, to forecast $inflation_{t+2}$ you need to know the value of $unemployment_{t+1}$. Since this is currently not known, you need to first forecast that, before you can forecast inflation itself. As we discussed in Chapter XXX, there are two ways of going about this. If you also have equation for unemployment rate, as is the case in multi-equation models, then you can use forecast from this equation to forecast unemployment rate, and then in turn forecast other inflation based on this forecast. Or you can use some satellite model for unemployment rate, such as simple univariate model for forecasting.

However, this discussion applies when we are forecasting the true future. If we are instead backtesting, i.e. creating historical forecasts to evaluate their performance, then we have alternative solution:

[45] This is even reflected in the available functionality of different software. For example the dominant software for data analysis - R - does not lend itself easily to in-sample forecasts, because forecast is always assumed to start after the end of estimation sample. In contrast, Eviews allows for flexible control of estimation and forecasting samples, which are treated as completely independent. Similarly, SpecEval gives user complete control over evaluation sample and whether forecasts are created in-sample or out-of-sample.

Use the actual historical values for unemployment rate, given that from the current vintage point you know those. Thus, for example, if you are trying to evaluate how your model would forecast during the great recession, i.e. starting in October 2008 and lasting until end of 2010, then you can use actual values for unemployment rate observed for this period.

Of course, this is a bit of cheating: back in October 2008 a forecaster would not know what the unemployment rate will be with certainty, so using the actual values for our independent variable removes major source of uncertainty from our forecasts. For this reason we distinguish between the two types of forecasts - one using actual values for independent variables, one using forecasts for independent variables - by using a different adjective for them. A most typical terminology is ex-ante and ex-post forecasts. **Ex-ante forecasts are those that only use the values of independent variables which were known at the time of the actual forecast**. In this sense they are true forecasts.[46] Meanwhile, **ex-post forecasts use values for independent variables that became available only after the starting period of the forecast**. In our example ex-post forecasts use the future values of unemployment rate that would not be known in October 2008.

We can again illustrate this using Figure H.8. This time around the distinction between green and blue balls applies to values of independent variables. In case of ex-ante forecasts we use only information contained in green balls. To create forecasts for our dependent variable we create forecasts for values of independent variables corresponding to the blue balls, and based on those we create forecasts for our dependent variable. In contrast, for ex-post forecasts we use actual values of independent variables corresponding to blue balls. In that sense, some information from blue balls is used for our forecasts for them. That said, we still do not use information about values of our dependent variable.

Now you may wonder, why would one study ex-post forecasts given that they are not true forecasts? There are two answers to this question, one benign and one substantial. On benign level the answer is because sometimes making ex-ante forecasts is not possible. If you are developing model for one variable, and one variable only, then you might not be in position to make past forecasts also for the independent variables, even if you have forecasts for them for the future. Simple, you do not have any equations for them. In this situation ex-post forecasts are the only possible forecasts you can make.

That said, there is also more substantial reason for the use of ex-post forecasts: Isolating forecast errors originating in your model. To appreciate this return back to our example. Imagine you first

[46] This applies only to independent variables; we still do not use any future information about the dependent variable.

created forecast for unemployment rate and then based on this you crated forecast for inflation, as in:

$$inflation^f_{t+2} = \beta_0 + \beta_1 unemployment^f_{t+1} \tag{8.5}$$

If your forecast is not particularly good you are faced with a question whether your forecast error is due to bad model linking inflation to unemployment rate or whether your forecast error is due to bad forecast for unemployment rate itself being wrong. Consider first the latter the case. Imagine situation when your model produces perfect forecast for inflation when you feed it the actual values for unemployment rate, but when you feed it forecast values for unemployment rate it produces poor forecast. In this case your model for inflation is fine, it is the model for unemployment rate which is faulty. Hence there is no need for corrective action with respect to your model for inflation.

In contrast, when your model for inflation produces poor forecast for inflation even when fed the true values for unemployment rate then you do have a problem and you need to adjust your model, such as using 4 lags of unemployment rate rather than just 1 or something else. In other words, your forecast error is a result of two errors, one originating in your model and one originating in forecasts for your independent variables. This complicates the evaluation of your model.[47]

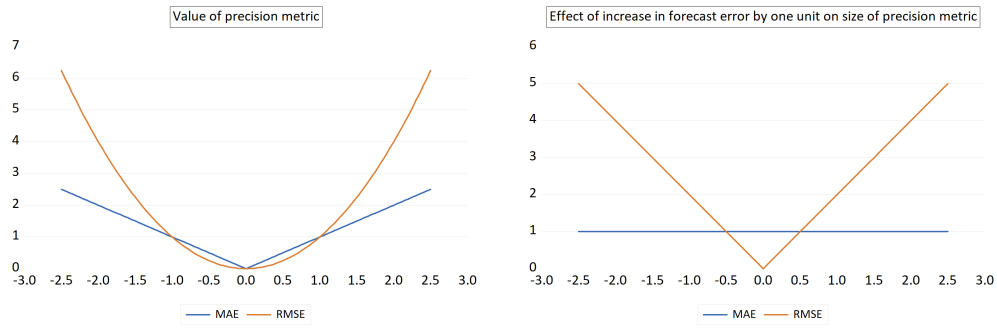
For this reason ex-post forecasts are actually a logical starting point. This is especially true when working on single equation multivariable models, but applies even to multi-equation models whenever one is interested in performance of particular equation of the model. Simply, as a first step when evaluating particular model it makes sense to focus on the forecast errors originating only in the given model. Only when we want an estimate of expected overall future forecasting performance do we really need to switch to ex-ante forecasts. In this sense the ex-ante vs ex-post forecasts debate is analogical to the out-of-sample vs in-sample debate. They are answers to different question, one more limited (ex-post and in-sample) and one more broad (ex-ante and out-of-sample). Given that the more broad question brings with it further complications in terms of the nature of the answer, it pays of to first focus on the more limited question, and only if the broader question is of interest, then switch to it.

Forecast precision metrics

RMSE and penalty function. When measuring forecast performance we have relied on one specific metric, the mean absolute error (MAE). However, it turns out there are several alternatives to this

[47] Of course, for one specific forecast you can disentangle these errors by looking at the forecasts errors for independent variables, but this becomes impossible once we are studying multiple forecasts, especially if we are relying on summary statistics like MAE.

Figure 8.17: Illustration of RMSE precision metric



metric. Here we discuss the most common ones and highlight what are the differences between these metrics.[48] The most common alternative to MAE is metric called root mean squared error (RMSE). As the name suggests, RMSE takes an alternative approach to dealing with the problem of positive and negative errors canceling each other: Rather than using absolute value function, it relies on the quadratic (square) function, which also translates negative errors into positive errors. Specifically, the formula for RMSE is following:

$$RMSE = \sqrt{\frac{1}{T-h}\sum_{t=1}^{T-h} e_t^2} \tag{8.6}$$

While this solution addresses the underlying motivation of avoiding canceling positive and negative errors, it is important to realize that the alternative function is not without important implications. Chiefly, in contrast to the absolute value function, which is linear, the quadratic function is non-linear, and more precisely convex. What does this mean in our case? Well, a non-linear convex function increases faster and faster as the we get further away from zero. So in our case this means that the *penalty* for forecast errors increases faster and faster as the forecast error becomes larger. Or in terms of marginal perspective, increase in forecast error by additional unit increases the penalty more than the previous unit increase. Figure H.17 illustrates.

At first sight this might seem like a strange idea, but once we think it through it actually becomes quite appealing. There are two arguments why square penalty function of RMSE might be preferable to absolute value function of MAE, one practical and one theoretical. On practical level it sounds like a good idea to give bigger weight to large forecast errors, because making very large error might be *disproportionately* more costly than making medium-sized error. This is exactly what RMSE does, so

[48]The discussion here is meant as an introduction to this topic. For more involved treatment, see for example, XXX.

when this characteristic is desirable, RMSE is indeed preferable for MAE.

On theoretical level, under certain assumptions a square penalty function is optimal for a decision-maker.[49] While this is on its own interesting, is also teaches us a broader lesson: the optimal penalty function might depend on the situation. And we do not need to be limited to absolute value and quadratic functions. Indeed, in some cases there are pretty good reasons to consider non-symmetric functions, so that errors with one sign are more costly than errors with another sign. For example, if you are seller of fresh fruit and you are forecasting demand, then it is more costly to forecast too high demand, which will lead to spoiled produce, than to forecast too low demand, which just results into forgone profit. So the lesson really is that it pays off to think which penalty function is appropriate to use given your modelling context.

So far we have explained why using square penalty function might make sense, given us the MSE of RMSE. On top of squaring and averaging, RMSE also takes square root of the overall number. The idea is really to reverse the distortion of units caused by the squaring (without negating the non-linear effect, of course). Like this, the final number is in units of the dependent variable and comparable with the range of the dependent variables. For example, when working with interest rates ranging from 0 to 5 then the final number will be likely be somewhere in the region of 0 to 10. This allows us to say things like "on average the forecast is 1.5 units from the actual value", while keeping in mind that this language is somewhat statistically and mathematically imprecise.

MAPE and relative errors. Apart from using a different penalty function, in some situations it is also suitable to use relative errors. What do we mean by this? Consider the MAE and its relative-error cousin Mean Average Percentage Error (MAPE).MAPE is very similar to MAE, but rather than using forecast errors measured in units of the dependent variable, it uses forecast errors measures in percentages of the dependent variable:

$$\tilde{e_t} = \frac{a_t - f_t}{a_t} \tag{8.7}$$

When using MAE the forecast precision statistic basically tells us how far on average is our forecast from the actual value. If, for example, our dependent variable is GDP measured in millions and MAE is 1.2, then we cans say that the average forecast is 1.2 million away from the actual value (in absolute sense). In contrast, if we would instead use MAPE, then it would be telling us that our forecast is on average 1.2% from the actual value. In other words, while MAE measures errors in natural units of

[49] See XXX for more discussion.

the dependent variable, MAPE measures them in percentages of the dependent variable.

The question whether one should use normal and percentage errors is basically a question of whether it is more natural to think about given variable in terms of its natural units, or in terms of its percentages. Upon introspection this should be in most cases clear, but there are few things that can be said in general.

First, the main distinction can be made between variables that are trending and variables that are stable. A trending variable can have values that are vastly different in different parts of sample; typically they can be very low in beginning of sample and very high at the end. This often means that the variance of movements in these variables, as measured in their natural units, is also increasing, as it scales with the current value of the series. Say you have variable that has value 10 in beginning of the sample and value 100 at the end of sample. It is natural to expect that a typical change in beginning of sample is something like 1, but that at the end of the sample it is something like 10. This is then why it is more natural to think about given variable in terms of percentage changes, rather than changes in natural units; both changes are equal to 10% of the current value of the variable.

Of course, all of this then has bearing on measurement of forecast errors. Imagine that you use MAE or RMSE as your forecast precision metric. This will effectively mean that you give more weight to forecasts for periods at the end of the sample, compared with the forecasts for periods in beginning of the sample. To see this, realize that your forecasts are likely to be proportional to the typical movement in the variable, and we already said that these are larger towards the end of the sample. This would favor models that have better forecasting performance towards the end of the sample. If instead we use precision metrics based on percentage errors than all parts of the sample will have equal weight.[50]

In contrast, MAPE is often unsuitable for rate variables. Consider the example interest rates, which typically vary from values close to zero, to values below 10. To appreciate would be the effect of using MAPE in such situation imagine that you make error equal to 1 in two different situation: when interest rates are very low, say 0.1, and when they are very high, say 10. This would translate into MAPE of 10 in the first case, and to MAPE 0.1 in the second case. In other words, periods when interest rates are very low would get much higher weight in our evaluation, while periods when rates are high would get very low weight. Again, in most cases such weighting scheme is not desirable. This

[50] Of course, giving more weight to more recent periods is a very natural thing to do. The problem is that here we are not doing it in conscious, controlled way. If such weighting is indeed desirable, it should be done explicitly via weighting scheme.

applies doubly so to the cousin of MAPE, the root mean square percentage error RMSPE, which is the relative error version of RMSE. Since RMSE puts more weight on larger errors then in our example here it would compound the effect coming from using relative errors.[51] The bottom line is that one should be mindful of choosing between aboslute and relative error precision metrics, and mostly use the relative version for trending variables and absolute version for rates.

Transformations of dependent variable. Another question relating to trending time series is whether we should be measuring forecast precision in terms of their levels or in terms of their growth rates. This is especially pertinent in situations when the dependent variable of the model is a growth rate in given variable. Here, the answer is much more ambiguous as it depends on the use case. If the modeler or model user is specifically interested in exact movements in the growth rate of given variable, then it makes sense to calculate forecast prevision based on growth rates. But in many situations it is not the exact movements in growth rate that matters, but the overall movements in the underlying variables. Which means we are left with a choice of what to do.

That said, there are important implications of either choice. Consider the example of measuring forecast performance for GDP during a period of recession illustrated in Figure H.18. In that figure GDP recorded negative growth in 1st and 3rd periods, and positive growth in 2nd and 4th, illustrated in blue line in Figure H.18. Meanwhile, imagine that our model predicts the correct movements, but has the timing wrong, with declines in 2nd and 4th periods, rather than 1st and 3rd, shown in orange line in Figure H.18. If we measure forecast performance in terms of growth, then our model will be evaluated as very bad. Indeed, it will be evaluated as worse than mode that will predict no contractions at all! Such conclusion feels strange, and indeed in many contexts will be wrong. In contrast, when looking at levels the model that predicts recession is substantially better than model that does not.

The message of this extreme example is that when evaluating models in terms of growth rates, we are putting focus on getting the timing of movements right, potentially more so than on getting the direction and magnitudes right. If instead we would measure forecast precision in terms of levels, than our model would be considered ok: it is able to predict the recession and its magnitude, it just gets the timing wrong.

Meanwhile, one should appreciate that when measuring forecast precision in terms of levels, there is the potential drawback of "multiple punishments". Specifically, Figure H.19 shows illustration in which GDP drops in 1st period and then remains unchanged, while our model does not predict this

[51] Otherwise, the difference between MAPE and RMSPE is the same as between their absolute error versions, MAE and RMSE.

Figure 8.18: Illustration of sensitivity of precision metrics to transformations

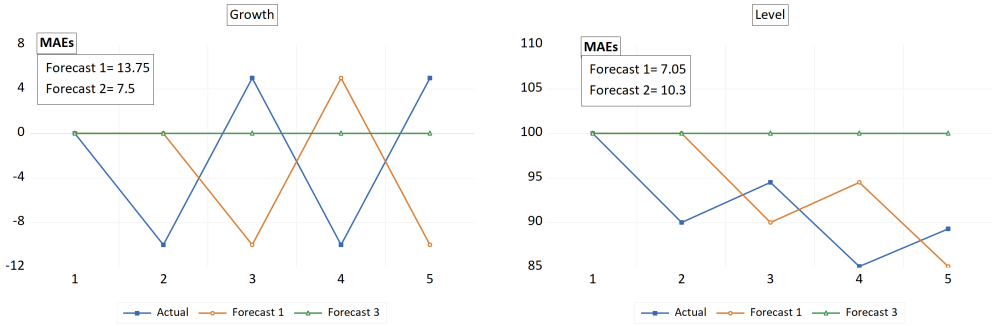
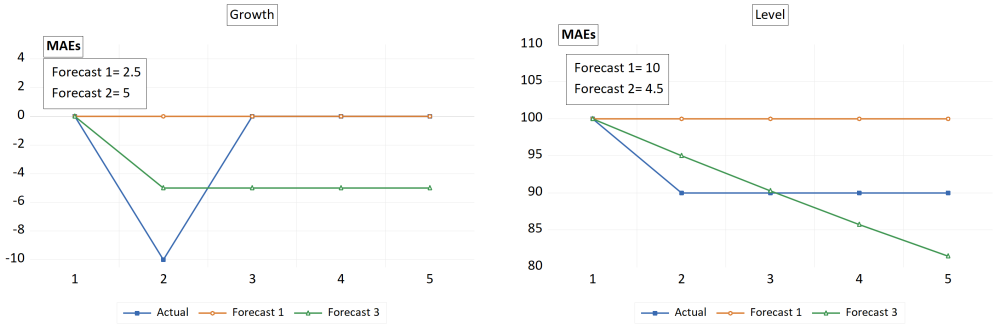


Figure 8.19: Illustration of sensitivity of precision metrics to transformations



decline, but does correctly predict that GDP remains unchanged afterwards. If we measure forecast precision in terms of level then the model will be deemed bad because after missing the drop in first period the level for following periods is wrong. This is despite the fact that it gets 3 out of 4 movements correctly. If instead we would use growth rate then the model will be considered ok. Sometimes this is appropriate, and sometimes it is not. As with other things in measuring forecast prevision, it depends on what are the forecasts used for.

These considerations are also related more generally to models that use some transformations as dependent variable. Should one measure forecast performance in terms of the underlying variable, or in terms of the transformed variable? Again, there is no correct answer to this question. That said, in most situations it is likely that modeler is interested in the base variable, not the transformed variable. Also, for the most simple transformation, a spread between two variables, it really does not make a difference.

This section added a lot of extra complexity to our initial, relatively simple discussion of measuring

historical forecast performance. So it is worth highlighting what is the key message here. The key message is that one evaluating a forecasting model, there are many different ways of doing so. These vary in terms of how the forecasts are created to how their precision is measured. And it is the role of modeler to choose between those different ways, choice which is sometimes of great importance. Moreover, sometimes these choice are not made consciously, what can have significant and undesirable implications for your modeling. So do not treat these choices lightly.

8.4 Sensitivity to shocks

The previous section focused on information about the *overall* forecasting performance of given model. This should be the primary source of information for evaluating forecasting model as it provides very good *overview* of the model performance. However, there are three reasons why this overall performance sometimes does not provide *fully sufficient* information about the model.

The first two reasons relate to different types of use cases for our forecasting model. First, rather than being interested in how will the model do in future in an environment *that is average from historical perspective*, we might be interested in how will the model behave in *abnormal situations* which feature large movements in some regressors, such as during recessions or other crises. Since by definition abnormal periods are relatively rare, the overall forecasting performance might hide the most relevant information behind the abundance of less relevant information. Second reason, applies to situations when forecasting models are used for particular type of forecasting. In some cases the main application of the forecasting model is not "best guess forecasting", but rather alternative scenario forecasting. Since these scenarios typically explore the more extreme outcomes, a good overall forecasting performance might not provide sufficient information to evaluate whether give forecasting model will perform well in such scenarios.

Final reason relates specifically to multivariate time series models. In those models a particular forecast profile is a result of influence of multiple drivers, each corresponding to one regressor. Even if we know that forecast during particular abnormal period displayed appropriate movement, it does not tell us that responses to movement *in each* regressor are appropriate. If we are only interested in the combined forecast, this is of course fine. However, for proper understanding of forecasts, and whether a good forecasting performance in future abnormal periods can be expected, it is better to also study the role played by each individual regressor in forecasts, and especially to what degree is the model incorporating shocks to each regressor.

For all these reasons it is sometimes suitable to analyze the forecasting model from different perspective than just overall forecasting performance. While the specific perspective will depend on the purpose and structure of the model, in general we can broadly classify these perspectives as information about sensitivity of the model to shocks: information that helps us answer questions like 'Is the model able to produce/explain *extreme* movements in the dependent variable?' or 'Does the model respond sufficiently to shocks in individual regressors?'. While overall forecasting performance does provide some information to answer these questions, in many situations this information is insufficient, and we need to use more specialized tools.

There are multiple ways we can asses sensitivity of models to shocks, which differ in terms of how easy they are to obtain and interpret. Some approaches will be almost readily available and will be very straightforward to interpret, while others will be harder to obtain and interpret. We will start with the simple approaches and proceed to the more complicated ones. The reasons why we will discuss also the more complicated approaches is because their complexity comes with a benefit: they either provide more general and/or robust answer than the simple approaches, or they provide answer to very specific, but useful question about sensitivity to shocks.

8.4.1 One shock at a time approach

Throughout this textbook we have heavily relied on one tool for understanding various time series models: impulse-response graph/function. We have introduced a single shock - either a random shock or a shock originating in exogenous variable - and studied how the dependent variable evolves after this shock. Therefore unsurprisingly, the starting point of our discussion of analyzing sensitivity to shocks is relying on this approach of introducing one shock at a time and analyzing the responses of dependent variable.

The subsection is divided into two parts. First, we will briefly discuss a numerical tools, before we turn to the graphical tools. As in previous cases, both tools have their strengths and weaknesses. Numerical tools are easy to obtain, straightforward to interpret and as such lent themselves to automatization. In contrast, graphical tools can me more complicated to obtain and require human interpretation, which can sometimes be too costly. That said, the graphical tools are provide much more comprehensive view of model sensitivity to shocks that simply cannot be obtained using numerical tools, which is why they are sometimes valuable.

Numerical tools

In case of numerical tools our approach to studying sensitivity of model to shocks is pretty straightforward. We simply want to know answer to following question: *If the regressor changes, does the dependent variable change a lot or a little.*

The only factors we need to determine is how much do we move the regressors, and how do we measure the responses of the dependent variable. The answers to these questions are motivated by desire for interpretability: We want to introduce shocks that are comparable across variables, and we want to measure the responses in a way that makes it easy to interpret and judge whether the response is appropriately large. Both of these considerations lead us to focus on statistical units, and specifically on standard deviation. Hence our numerical measure of sensitivity to shocks is response of dependent variable, measured in terms of standard deviations, to standard deviation shock in regressor.

This results into something we already know from earlier parts of this chapter: **standardized coefficients**. We have introduced them as a tool for analyzing whether coefficients have appropriate size and for comparing coefficients across variables. The idea was that standardized coefficients translate regression coefficients expressed in natural units, that are not comparable across variables, into coefficients expressed in statistical units, that are comparable. Standardized coefficients are the answer to question 'By how many standard deviations does dependent variable move, if independent variable moves by one standard deviation?'. But this is exactly an answer to our most basic question about model sensitivity! Therefore, in addition to helping with comparing importance of regressors, standardized coefficients are also useful for studying sensitivity to shocks. This is another reason why they should feature prominently in model building process.

Table H.11, replication of Table H.3 from section H.2, illustrates the value of using standardized coefficients. The model includes three variables with very different coefficients: the largest coefficient is 200-times larger than the smallest coefficient. However, these differences in un-adjusted coefficients measured in natural units do not imply different sensitivity to shocks in these three variables. This becomes clear when looking the standardized coefficients, which are very similar to each other. In other words, when we consider comparable shocks for all three variables then the responses to these shocks are very similar.

Table 8.11: Illustration of coefficient size sensitivity

Dependent Variable: DLOG(CPI)
Method: Least Squares
Sample: 1998Q1 2019Q2
Included observations: 86

Variable	Coefficient	Std. Error	t-Statistic	Prob.	Std. dev.	Std. coef.
C	0.0062	0.00068	9.11	0.000		
@MA(DLOG(OIL_PRICE),4)	0.011	0.0077	1.40	0.17	0.09	0.14
@MA(DLOG(IMPORT_PRICES(-1)),4)	0.21	0.066	3.19	0.00	0.0021	0.31
OUTPUT_GAP	0.00087	0.00032	2.75	0.01	2.2	0.27

So how can we then use standardized coefficients to determine whether model is sensitive to shocks in particular variable? The idea is relatively simple. Sufficient sensitivity to shocks means that shock of standard size leads to reasonably large response in the dependent variable. What is reasonably large? As we discussed in section H.2, standardized coefficients are typically somewhere between 0 a 0.5, with 0 implying no effect on dependent variable and 0.5 implying large effect. To conclude that model has sufficient sensitivity to shock in given variable we would typically want standardized coefficient of at least 0.1, and ideally larger than 0.2. At these values we can conclude that move standard size cause measurable response in the dependent variable.

What does our example suggests in regards to this? We can see that shocks to log-difference of import prices of standard size, which is equal to 0.01, leads to response of 0.0021 in log-differences of CPI ($0.01 * 0.21$). Since standard move in log-differences of CPI is 0.007, then the response to shock is equal to 31% of the standard move in log-differences of CPI (0.0021/0.007), yielding the standardized coefficient of 0.31. This is rather large sensitivity to shocks in import prices. More interestingly, the response to oil price shocks is also reasonably large, as indicated by standardized coefficient of 0.14. This is despite the un-adjusted coefficient being 20-times smaller than that of import prices. To see why, we just need to realize that the standard-sized shock in log-difference of oil prices, which is equal to 0.09, is almost 10-times larger than the standard-sized shock in import prices, which we said is 0.01. This then partly compensates for the smaller unadjusted coefficient, yielding still-large standardized coefficient. At 0.14 this coefficient says a standard move in oil prices causes response that is equal to 14% of standard move in consumer prices.

Limitations. Standardized coefficients are extremely simple and yet useful source of information about the sensitivity of the model to shocks. However, it is important to be aware of their limitations.

Table H.12 illustrates the main limitation. It captures a model that links current unemployment rate gap to previous values of unemployment rate gap and recession dummy. This correspondents to basic macroeconomic idea that when economy is in recession then unemployment rate gap should be higher. While the coefficient associated with the dummy variable is indeed positive, the size of standardized coefficient seems to suggests that the effect is small.

However, in this case standardized coefficient is potentially misleading, if not interpreted with care. The key realization is that the model at hand is a dynamic model. As we learned in Chapter **??**, in such models immediate shock response can be very different from interim and/or long-term shock response, because the initial affect can be propagated and amplified via the lagged dependent variables.[52] Indeed, that is the case in our example here: the coefficients on two lags of dependent variable are such that there is both very long-lasting propagation and very strong amplification, as illustrated in Figure H.20. And while standardized coefficients provide information about the immediate response, which in this case is rather small, they do not provide any information on the size of responses later on. This is especially important because shock responses for this model have strong accumulation feature: If the shocks will last say 4 periods, then the peak effect will be almost 4 times the initial effect (something we will illustrate in next subsection). Therefore, reliance on standardized coefficients can lead to misleading conclusions, unless one is careful not to forget that they provide information only about the immediate shock responses. [53]

[52]Of course, this distinction is relevant only for models that are dynamic. If models are completely static then there is little ambiguity of what to do: we can simply rely immediate response, since that is equal to interim and long-term response. However, if models are dynamic, then we need to distinguish between the immediate response on one hand, and the interim and long-term responses on the other hand.

[53]Note that this was much less of an issue in our previous use of standardized coefficients as a measure of relative importance of regressors, since the dynamic aspect due to autoregressive structure of the model applies to all regressors equivalently.

Table 8.12: Illustration of limits of standardized coefficients for analyzing shock responses

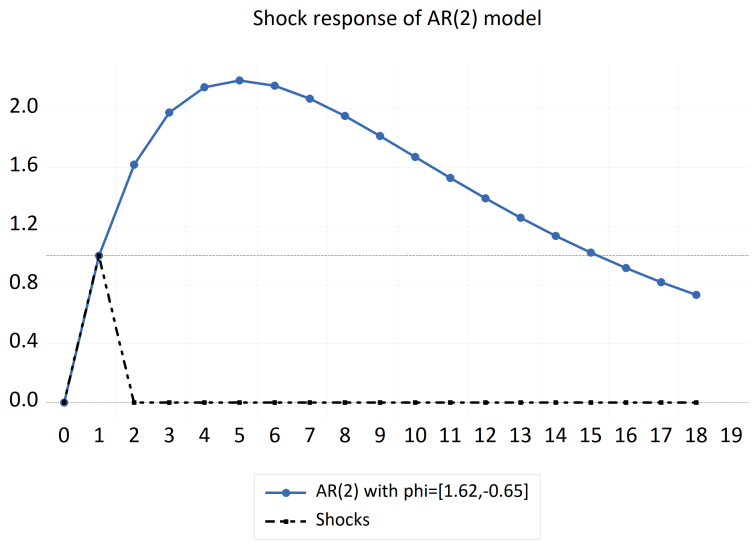
Dependent Variable: UNP_GAP
Method: Least Squares
Date: 04/07/23 Time: 10:49
Sample (adjusted): 1996Q2 2019Q2
Included observations: 93 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.	Std. coef.
C	−0.067	0.024	−2.79	0.01	
UNP_GAP(-1)	1.56	0.075	20.8	0.000	1.54
UNP_GAP(-2)	−0.59	0.078	−7.63	0.000	−0.57
RECESS_DUMMY	0.22	0.056	4.02	0.000	0.068

R-squared	0.981	Mean dependent var	−0.38
Adjusted R-squared	0.980	S.D. dependent var	1.37
S.E. of regression	0.19	Akaike info criterion	−0.41
Sum squared resid	3.31	Schwarz criterion	−0.30
Log likelihood	23.2	Hannan-Quinn criter.	−0.37
F-statistic	1528.5	Durbin-Watson stat	2.15
Prob(F-statistic)	0.000		

Variable	Description
UNP_GAP	Unemployment rate gap
RECESS_DUMMY	Recession dummy

Figure 8.20: Shock response profile for model from Table H.12



Are there any solutions to this problem? In principle yes, but in practice they do not seem to be completely satisfactory. Instead of measuring only the immediate shock response, any alternative solution would need to calculate the response over a longer horizon. For example, imagine introducing a shock in the model captured in Table H.12, and then calculating not by how much does dependent variable increase *immediately*, but by how many standard deviations does it increase *after 10 or 20 periods*. This could then be a measure of our sensitivity to shocks. Or even better, we could calculate the *average* difference between the path of dependent variable in absence of a shock and in presence of a shock. This would be more suitable for transitory shocks, which by definition do not have permanent effects. While such approaches are plausible they unavoidably become too complex for numerical approaches to be really valuable. Instead we will shift to graphical tools for analyzing sensitivity.

Graphical tools

The alternative to numerical tools for analyzing sensitivity to shocks are graphical tools. As in other areas, the graphical tools have the advantage of providing more complete information, but at a cost of requiring human mind to interpret them. Sometimes, this cost is too large to go down this road. But in other times this cost is justified or even necessary, in which case it is useful to know what graphical tools to focus on.

The starting point of our discussion will be the last example from previous subsection. This example

considered model that featured dynamic terms and highlighted the shortcoming of the numerical tools in presence of dynamic responses. Figure H.21 provides an illustration for this example. The top left panel captures the shock we are introducing to the model, in this case a one period increase in the recession dummy, with size of shock equal to one standard deviation. We have seen in previous section that the immediate response of unemployment rate is relatively small. But the top right panel confirms what we concluded in previous subsection, highlighting the difference between the limited immediate response to the shock and the more significant dynamic response. We can see that the immediate effect in the period of the shock (here 0.09 in 2010Q1) is less than half as large the peak effect (0.19 in 2010Q4).

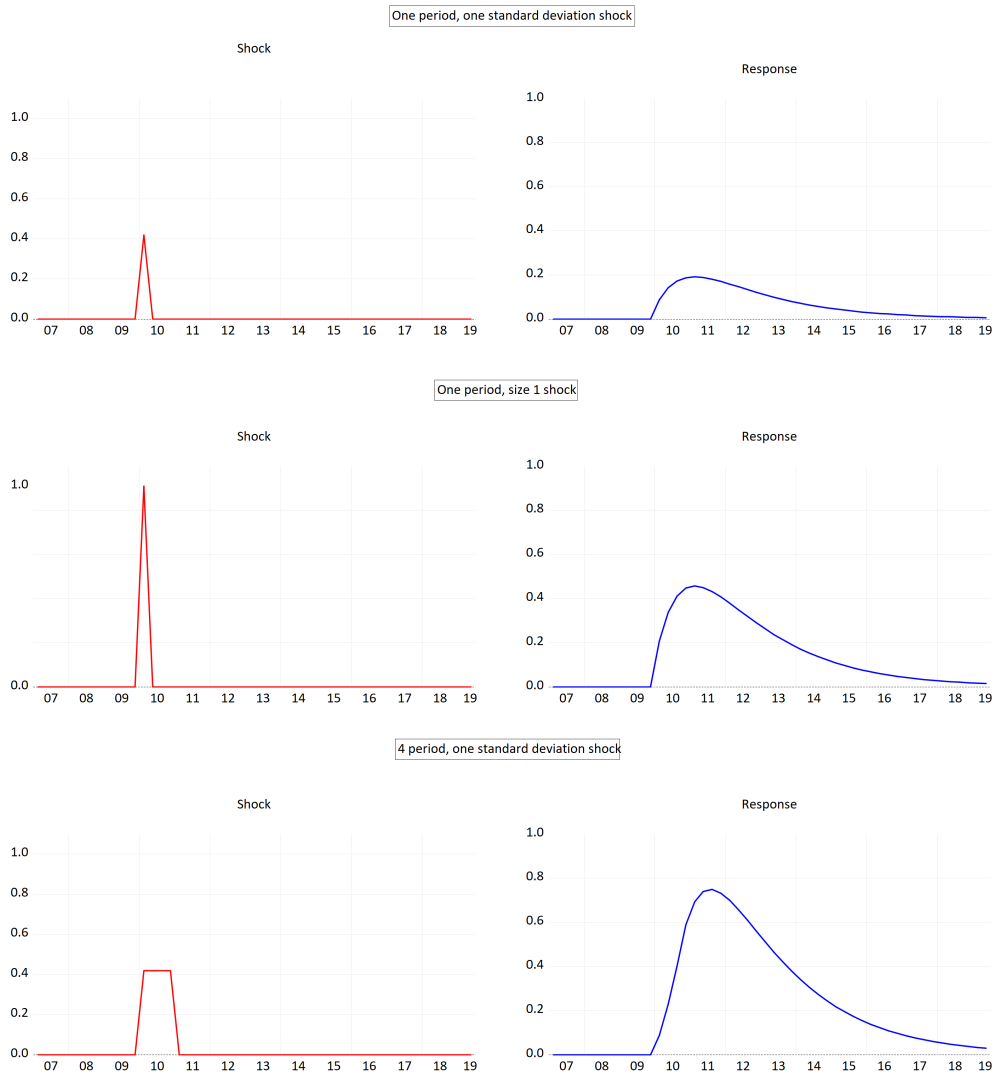
This highlights the first advantage of graphical analysis. The second advantage is that graphical analysis gives us clear idea of the whole profile of response to our shock: how long it takes us to reach peak, how long the peak effect lasts, and how long before the effect becomes negligible. In most cases all of these answers are impossible to see from analysis of regression output.

Modifications. Apart from highlighting the difference between immediate and dynamic responses, this example also raises the question of the specification of the shock we are introducing. The top row of Figure H.21 used one standard deviation shock lasting for one period. However, in the case of our regressor, which is the dummy variable, both the size and length of the shock are somewhat strange. Dummy variables take only two values, 0 and 1, which means that the natural shock would be increasing the dummy variable from 0 to 1, not by its standard deviation. Even more importantly, our regressor captures a presence of recession, and specifying one-period shock amounts to assuming that we have recession that lasts one period (quarter). Since typical recessions last longer than that, it seems that studying effect of a longer shock would be appropriate.

Middle and bottom row of Figure H.21 rectify these shortcomings by showing sensitivity to shocks with different specification. The middle row shows the response to shock of size 1. Changing the shock size to the more natural shock size significantly changes the conclusions. Not only is the peak effect proportionally larger (0.46 instead of 0.19), but even the immediate effect is now meaningfully large (0.21 rather than 0.09). This highlights how considering "appropriate" size of shocks can play important role.

Meanwhile, the bottom panels capture the effect of shock of one standard deviation lasting for 4 quarters, instead of the 1 quarter from the first panel. Again, this change demonstrates how length of shocks can play important role in our conclusions about sensitivity to shocks. Since the model is

Figure 8.21: Shock response graphs



dynamic in nature, the effect of shocks accumulates and hence shock lasting several periods will have very different dynamic effect from shock lasting only one period. Here, given the strong accumulation tendency of the model the peak effect of 4-quarter shock is almost 4-times as large as the peak effect of 1-quarter shock (0.75 vs 0.19).

This example thus highlights that question of size and length of shocks are potentially important when studying sensitivity of models to shocks via graphical tools: Different specifications can lead different conclusions. In terms of size, the main distinction is between statistical and natural units. In most cases you will want to rely on statistical shocks. This is for two reasons. First, we want the shock to be representative of shocks we could encounter in future. Second, we want the shocks to be comparable across variables. The natural candidate is therefore a standard deviation, which is a measure of variability of given variable, so that the shocks should have a size of one (or more) standard deviation of give variable.

That said, a standard deviation is a *summary* statistic of variability, which means that it is based on the whole distribution of given variable. This works well when given variable has distribution that is more or less similar to normal distribution, but can potentially be misleading when given variable has distribution that features extreme values. In that case, a one- or even two-standard deviation move might not be representative of true extreme movements. Rather than relying on standard deviation, we can rely on alternative statistical units which are not calculated based on the full distribution of given variable: percentiles. Therefore, an alternative way to determine the size of the shock is to use difference between median and an extreme percentile of given variable, such as the 99th percentile.

Meanwhile, in terms of length of shocks, the main distinction is between *transitory* and *permanent* shocks. In most situations we want to study the sensitivity to transitory shocks, as transitory (one- or multi-period) shocks are most natural shocks to imagine. However, the graphical tools allow us also to study effect of permanent shocks as well, which are sometimes of interest. Figure H.22 shows one such example: The model links long-term interest rates to level of debt relative to nominal GDP (see Table H.13). Of particular interest with regards to such link is the question *'What is the effect of permanently higher level of debt on long-term interest rates?'*. The figure makes it clear that while the immediate impact of 10% higher debt level is very small, the long-term impact of permanently higher government is substantial at 0.35%.

Table 8.13: Model for Czechia 10-yr yield

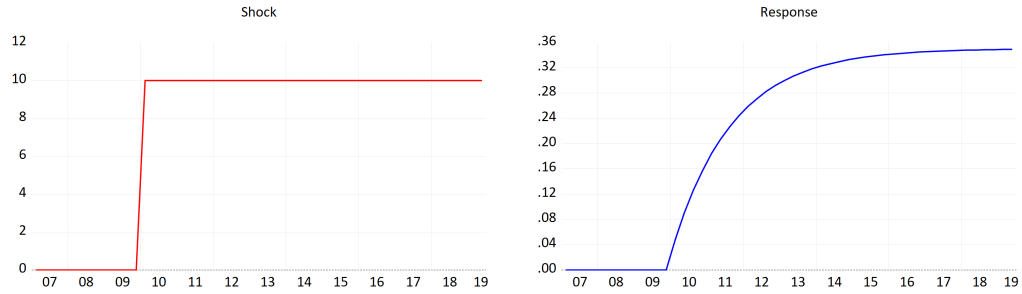
Dependent Variable: YIELD_10Y-YIELD_10Y_GERMANY
Method: Least Squares
Date: 05/14/23 Time: 12:58
Sample (adjusted): 2001Q1 2019Q2
Included observations: 74 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.	Std. coef.
C	-0.084	0.15	−0.55	0.58	NA
YIELD_10Y(-1)-YIELD_10Y_GERMANY(-1)	0.86	0.055	15.6	0.000	0.88
@MOVAV(DEBT_TO_GDP(-1),4)	0.0048	0.0045	1.09	0.28	0.061

R-squared	0.77	Mean dependent var	0.55
Adjusted R-squared	0.77	S.D. dependent var	0.59
S.E. of regression	0.29	Akaike info criterion	0.37
Sum squared resid	5.80	Schwarz criterion	0.47
Log likelihood	-10.8	Hannan-Quinn criter.	0.41
F-statistic	122.18	Durbin-Watson stat	1.51
Prob(F-statistic)	0.00		

Variable	Description
YIELD_10Y	Government 10-yr yield
YIELD_10Y_GERMANY	Government 10-yr yield (Germany)
DEBT_TO_GDP	Debt-to-GDP (%)

Figure 8.22: Shock response graphs - permanent shock



Note: Shock applied to regressor $@MOVAV(DEBT\_TO\_GDP(-1), 4)$ from model from Table H.13.

Finally, the questions surrounding the specification of shock do not end with the size and timing of the shocks, but rather can be further complicated by also asking *'What should be shocked?'*. At first this question seems to have obvious answer: we should shock the right-hand side variables of our model, and indeed that is what we did so far. However, the problem arises when we realize that many

models contain regressors that are transformations of single or even multiple independent variables. Should we study the responses to changes in the independent variables, or changes in the regressors? And, of course, this question also has an analogy on the left-hand side of the equation: should we study the response of dependent variable, more properly called the regressand, or of the underlying base variable which the model determines?[54]

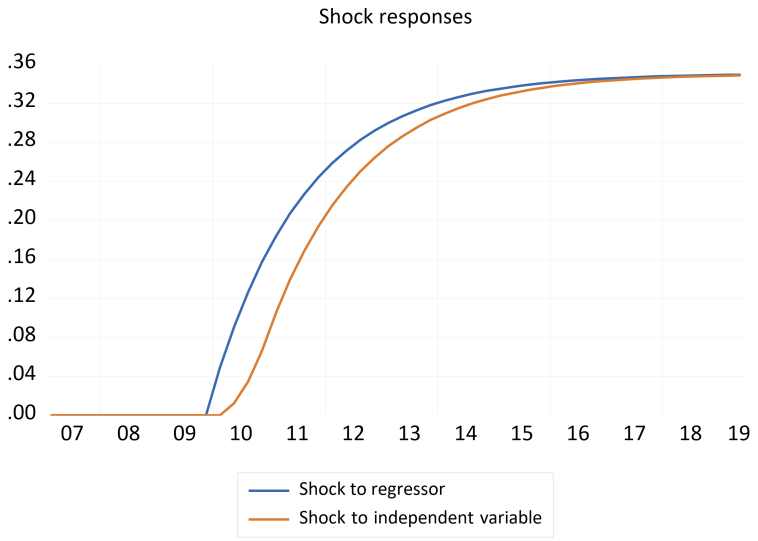
There is no right answer to these questions, and either approaches can be valuable depending on situation. True, most modelers will restrict their attention to situation when shocks are applied to regressors and responses in the regressand are studied. This is effectively the case when we think of the model as being formed of new, transformed variables, rather than of the original variables from the dataset. However, this restricted attention often has more to do with the fact that most statistical packages operate in way that doing something else would be complicated, instead of what is actually valuable. Indeed, most often it is more natural to think in terms of responses of the underlying base variable, rather than the transformed dependent variable, since the underlying variable is what we are trying to forecast. And since neither approach is always better, it typically makes sense to include both types of shock response exercise in our analysis, so that either can be called upon when needed.

To appreciate this distinctions, we provide here few examples. Returning to our last example studying effect of debt on interest rate, notice that our shock was applied to moving average of debt-to-GDP. Of course, in some sense this is misleading: if there is a one-time shock to debt-to-GDP, then it takes 4 quarters before this shock is fully reflected the moving average. Therefore, it might be more appropriate to study the effect on change in the independent variable, $DEBT\_TO\_GDP$, rather than in the regressor, $@MOVAV(DEBT\_TO\_GDP(-1), 4)$. Figure H.23 compares the responses to the two different specifications of shocks side-by-side. While the long-run response is of course unaffected, the short- and even medium-term profiles are measurably different. Simply, if we shock the independent variable, not the regressor, then the effect of shocks is delayed, as it takes time for shock in the independent variable to fully reflect in the moving-average regreesor.

Distinguishing between shocks to debt-to-GDP and and shocks to moving average of debt-to-GDP does not have the be the end of the story. One could also unpack the debt-to-GDP variable, which really is composed of two variables, debt and GDP. Therefore, we could study the effect of shock to

[54] As a reminder, the language we used in this textbook is that regressor is a term on right-hand side of an equation that can be thought of a standalone term, i.e. is additively separable form the other terms. Independent variable is then an actual variable from the original dataset used as part of regressors. For example, $y - z = \beta_0 + \beta_1 x_1/x_2$ has one underlying variable, y, one dependent variable/regressand, $y - z$, three independent variables, z, x_1 and x_2, and one regressor, x_1/x_2.

Figure 8.23: Shock response graphs - permanent shock



Note: Shock applied to regressor $@DEBT\_TO\_GDP(-1)$ from model from Table H.13.

GDP and effect of shock to debt separately. This can be especially relevant when we are interested in the effects of such underlying variables, rather than the composite variables, or when the composite variable is more complex than just simple ratio or difference.

The above is an example where shocking independent variable was for most purposes more appropriate. Meanwhile, to provide example when the opposite is the case, consider model in Table H.14 that links house prices to disposable income per capita.

Our focus will be on the second regressor, $DISPOSABLE\_INCOME(-1)/POPULATION(-1)$, which captures the idea that house prices should be higher when people have higher disposable incomes. In this case, while we could look at the effect shock in disposable income, it could easily lead to confusion, because disposable income measures the total disposable income in the economy, not the average person's disposable income. In other words, if total disposable income increases by 10%, it raises a question whether average person has disposable income that is higher by 10%, or whether there are 10% more people in the economy. Similar thing applies to the third regressor, $@MOVAV(UNP\_RATE - NAIRU, 8)$: if unemployment rises, is it because there is a recession, or because there was structural change in the economy? These two options have different implications for the model, and hence it might be more appropriate to shock the regressor than the independent variable.

Table 8.14: Model for Czechia house prices

Dependent Variable: @PC(HOUSE_PRICES)
Method: Least Squares
Date: 06/03/23 Time: 15:30
Sample (adjusted): 2000Q2 2019Q2
Included observations: 77 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.	Std. coef.
C	9.59	3.33	2.88	0.01	NA
HOUSE_PRICES(-1)	-0.25	0.05	−4.97	0.000	-2.76
DISPOSABLE_INCOME(-1)/POPULATION(-1)	0.086	0.03	2.84	0.01	1.49
@MOVAV(UNP_RATE-NAIRU,8)	-1.57	0.34	−4.60	0.000	-0.85
YIELD_10Y	-0.52	0.23	−2.28	0.03	-0.43

R-squared	0.35	Mean dependent var	1.65
Adjusted R-squared	0.32	S.D. dependent var	2.19
S.E. of regression	1.81	Akaike info criterion	4.08
Sum squared resid	235	Schwarz criterion	4.24
Log likelihood	-152	Hannan-Quinn criter.	4.14
F-statistic	9.85	Durbin-Watson stat	0.58
Prob(F-statistic)	0.00		

Variable	Description
HOUSE_PRICES	House price index
DISPOSABLE_INCOME	Total disposable income
POPULATION	Czechia population
UNP_RATE	Unemployment rate
NAIRU	Natural rate of unemployment
YIELD_10Y	Government 10-yr yield

Drawbacks

Our first solution to studying sensitivity to shocks was pretty straightforward: we simply introduced one shock of a statistically standard size and studied how the dependent variable changes as a result. For immediate shock response, or for models that do not feature any dynamic behavior, we could simply rely on the standardized coefficient we have already introduced earlier; here, we have just given them an additional meaning. For dynamic responses it is best to rely on graphing the responses.

There are two problems with the one-shock-at-a-time approach. First, using standardized shocks inherits the limits of standard deviation as measure of variability. While for variables whose distribution is fairly close to normal distribution using standard deviation is reasonable, this is no longer case when variable's distribution deviates significantly from normal distribution. The most important deviation is presence of extreme shocks, which can be fairly common in economic time series. If given regressor features extreme values then shocks that will have size of one or even two standard deviations might not be representative of what the extreme shocks might look like. Since sensitivity to extreme shocks is what we are most interested in, we might be missing the point, especially when comparing across variables.[55]

As mentioned before, this shortcoming might have relatively simple solution: use of percentile shocks. However, even this solution fails to address the problem of variables that don't have stable distribution. For example, in co-integration regressions we can encounter regressors that are non-stationary in the sense that they don't have stable mean or that they have trending mean. In such case taking any statistical measure is problematic. One solution is to determine whether series is or is not stationary and take first differences of series that are not.[56][57]

The second drawback is inherent in the approach focusing on one shock at a time: it is the fact that we are introducing one shock at a time. While this is a logical starting point, it ignores the fact that shocks to regressors can and do occur at the same time. Indeed, the most extreme movements in dependent variable typically do not happen because of one regressor moving a lot, but rather because of multiple regressors moving at the same time. Therefore, looking at responses to one shock at a time

[55] Note that this is a problem only in terms of analyzing the responses to extreme shocks. In terms of knowing whether given model responds to shocks, the size of shock is in itself not important, since in linear models the responses scales linearly with shocks, as we have seen in Chapter 6.

[56] SpecEval does something in this direction, even though it classifies series according to whether they are trending or not, rather than based on their stationarity. The reason is to avoid relying on statistical tests for stationarity, which are known to have low power.

[57] Another, more benign problem is posed by discrete variables, such as dummy variables, as we pointed out earlier. In this case we are probably most interested in the effect of moving from value of 0 to value of 1, rather than increasing by one standard deviation - and increasing by two standard deviations might not even make sense.

might provide us with misleading picture, because the responses will look small relatively to the range of movements observed in history.

The natural solution to this problem would seem to be to introduce shocks in correlated fashion: not introduce one shock, but rather a set of shocks corresponding to some historical measure of correlation between shocks. Indeed, this is similar to approach followed by modern literature on VAR models. However, this approach has a limitation closely related to the limitation of standardized shocks. Namely, using overall correlation between regressors as a measure of correlation between our shocks might be problematic, because correlation between regressors might vary with the magnitude or direction of shocks. For example, two regressors might on average be uncorrelated, but they might have tendency to co-move in extreme environments.[58] Accounting for this is not straightforward. Rather, we will take a different approach in following subsection.

8.4.2 Multiple shocks approach

Since extreme movements in dependent variable are typically result of large correlated movement in independent variables, rather than large movement in single dependent variable, we need have ability to study responses to multiple correlated shocks. Only by introducing multiple shocks at a time will we be able to determine whether the model is able to produce movements in dependent variable that are sufficiently extreme.

How do we introduce multiple shocks at a time? To circumvent the problem of determining correlation between the shocks, we can rely on two similar but slightly different approaches. First approach uses shocks observed during particular historical period, such as the global financial crisis or great recession. We will call this historical shocks approach Second approach is to use shocks entailed in particular alternative extreme scenario, assuming such scenario is available. We will refer to this as scenario shocks approach. We discuss these approaches in turn.

Historical shocks. Consider situation of building a model for some macroeconomic variable, such as unemployment rate or stock price index. It might be of the utmost importance to know whether your model is able to produce forecasts that feature large movements in given variable, such as movements observed during severe recessions or during period of turmoil in financial markets. Luckily, typically we have several such periods available in our historical sample. Therefore, it seems logical to look at those specific periods to know whether your proposed model is able to produce large movements in

[58] Alternative approach of introducing standardized shocks all together is clearly not optimal, as that amounts to assuming perfect correlation between shocks.

dependent variable. This is exactly what we will do here.

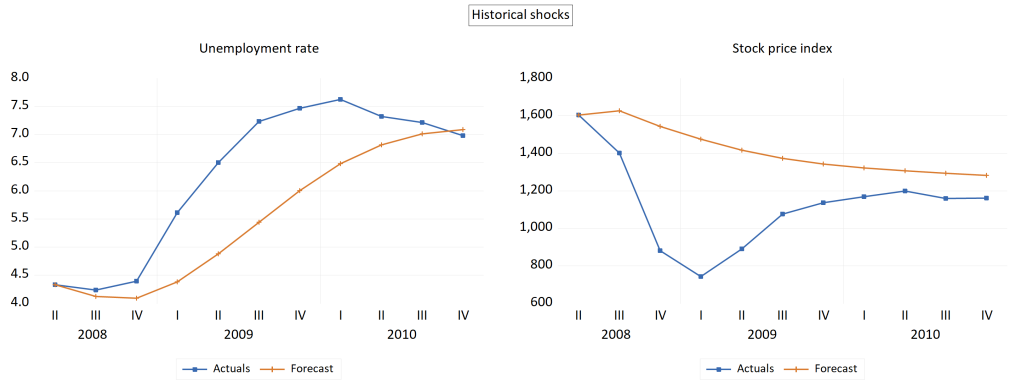
The historical shock approach amounts to using values for our independent variables actually observed in history and based on these values creating forecast for the independent variable. This forecast can then be analyzed from the perspective of responses to these shocks. Effectively, this is just analyzing one particular backtest forecasts, so producing this does not require any new skills or tools on top of what we have discussed in section H.3. The difference here is in the focus of the analysis: we are specifically interested in the magnitude of the movements, and not necessarily in the precision of the forecast per se. While these two are closely related, they are not the same thing.

Once we have obtained our stress period forecast, we need to evaluate how well did the model respond to shocks observed during the stress period. In general, to judge the magnitude of predicted movements, we need to know what is the appropriate magnitude. A priori, one does not know how much should unemployment rate increase during particular recession, or how much should stock prices drop in period of financial crisis. Should unemployment rate rise by 2% or 6%? Should stock prices drop by 20% or more like 50%?

Of course, in case of historical shocks we do have natural benchmark: the actual observed movement, which tells us how much did unemployment rise and stock markets drop. Therefore, we can benchmark the movement in dependent variable produced by the model to the actual movement observed. Figure H.24 illustrates the approach: It shows forecasts for unemployment rate in the left panel and for stock price index in the right panel, both for the period of great recession. You can notice that the model for unemployment rate is able to produce a large increase in unemployment rate, with unemployment rate rising from 4% to 7%. In contrast the model for stock prices is not: While stock price do decrease, their decrease is not only too slow compared with the actuals, but also significantly smaller than the observed peak drop of 50%. This suggests that the model for unemployment rate is fine as far as shock responses are considered, but the model for stock prices should not be expected to produce large movements in stock prices during periods of financial market turmoil.

Of course, there are limits to such reasoning. First, there are reasons why the movements should be smaller, or larger, than what the model produced: the presence of random shocks originating in the dependent variable itself. For example, during the pandemic government took specific actions to prevent unemployment rate from rising. From a perspective of model developed before the pandemic, this amounts to negative random shocks to unemployment rate that result in lower unemployment rate than what the model would produce.

Figure 8.24: Example of historical shocks graphs



Second, and related, we are effectively trying to make conclusions based on sample size of one (or few, in case we study few such periods). Even a basic knowledge of statistics makes it clear this is dangerous. Therefore, one should treat this approach with a degree of skepticism: It can provide useful information, but should not be relied on religiously.

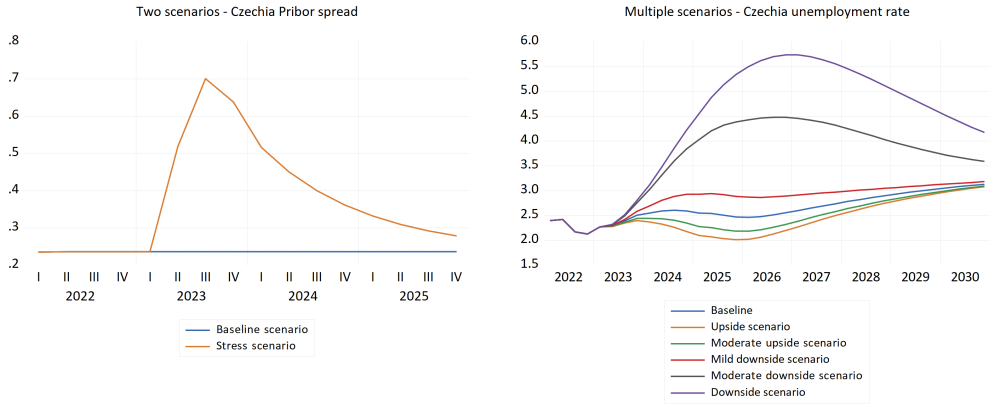
Scenario shocks. An alternative to using observed historical values for independent variables is to use values from particular scenario or scenarios. In many situations companies and institutions do not produce just single baseline forecast, that captures the most likely outcome, but also many alternative scenarios. These typically focus on downside or upside risks to the baseline forecast, and as such feature large shocks. Moreover, these shocks have been curated, which means that they create a set of internally consistent and plausible shocks. As such, using forecasts from these scenarios provides a suitable alternative to either historical shocks or to introducing correlated shocks.

The approach then is very similar to the approach in case of historical shocks. One just needs to use the scenario values for all independent variables and create associated forecast for dependent variable.[59] Left panel of Figure H.25 illustrates this approach. It shows the scenario forecast for Czechia Pribor spread in a stress scenario. This scenario features financial market shock and corresponding increase in the Pribor spread, which measures the stress in interbank markets.

[59]This if of course only possible if such values are available, which means we need to be performing model development in environment where multiple scenarios exist.

An example of such alternative scenarios are macroeconomic scenarios used for stress testing, either produced internally, such as IFRS 9 scenarios, or prepared by regulators, such as the CCAR scenarios from the Federal reserve.

Figure 8.25: Example of scenario shocks graphs



While such scenario forecasts can be studied individually, it is typically more useful to study them as a set. At minimum it is good to include the baseline forecast, since the relative comparison to baseline forecast highlights the effect of the shocks that have been introduced. This is why the left panel included also a baseline scenario forecast. Moreover, sometimes it is also useful to include multiple alternative scenarios with varying severity, which then allows the model builder to judge how the increasing severity of shocks translates in movements in dependent variable. Right panel of Figure H.25 provides illustration for the case of the unemployment rate.

Compared to the historical shocks approach, the scenario shocks approach has its advantages and disadvantages. The advantage is the ability to study shock responses in more controlled environment that features more systematic and less random shocks. Any particular historical experience features a large degree of uniqueness, what can make it hard to draw firm conclusions from such experience. Meanwhile, the maim disadvantage is the fact that we do not know what the appropriate shock responses look like, unlike in historical shocks where we can rely on the observed path for the dependent variable. This absence of any reliable benchmark means that the evaluation is to a large degree subjective.[60]

Additional reason when scenario shocks are useful comes into play when model that is being developed will later be used for creating scenario forecasts. In such situation it is often the case that scenario forecasts are typically of special interest - they are often more important that the best-guess forecasts. In such situation scenario shocks approach is particularly useful: Studying scenario shocks

[60]That said, one can rely on comparison with similar historical experiences. For example, the magnitude of a shock response in a severe downside scenario can be partly judged by comparing it with the movement during the Great recession.

is not only valuable for understanding responsiveness of the model to shocks, but also as a way of knowing how will the model behave in actual scenario forecasts.

8.4.3 Forecast decomposition

While the multiple shock approaches are useful in answering the question whether the model is able to produce/explain extreme movements in the dependent variable, at a first glance they seem unsuitable to answering the question whether the model responds sufficiently to shocks *in individual regressors*. The problem is that the response in the dependent variable is a result of multiple shocks in independent variables, so it is hard to know how much weight does each individual regressor pull. In a sense, the multiple shock approach gave us answer to one question we are interested in at the cost of not answering another question.

Luckily it turns out that we can recover the needed information, at least to some degree. Simply, to answer the question about responses to individual regressors, we just need to decompose the response of the dependent variable into the effect of individual independent variables. And exactly such forecast decomposition is the topic of this subsection.

Consider a simple multivariate model of the ARDL family, such as $y_t = \beta_0 + \beta_1 x_t + \beta_2(w_{t-1} - z_{t-2}) + \beta_3 y_{t-1}$. As explained in Chapter 6, this model is a linear model in that the right-hand side of the equation is additively separable: each individual regressor, such as $(w_{t-1} - z_{t-2})$, is simply added to the other regressors to produce the forecast for the dependent variable. Of course, we can reverse this addition: we can decompose the forecast for the dependent variable in each period into the contributions of the individual regressors. In our example, $y^f_{t+\tau}$ can be decomposed into four independent components: β_0, $\beta_1 x_{t+\tau}$, $\beta_2(w_{t+\tau-1} - z_{t+\tau-2})$ and $\beta_3 y_{t+\tau-1}$. This is true for any τ. We will call these four components the forecast drivers, which are simply the regressors multiplied by their associated coefficients.

Armed with this decomposition, we can proceed to analyzing responses to individual regressor shocks in the context of our historical and scenario shocks approach. To know whether the model responds appropriately to shocks in a particular regressor we simply need to decompose our forecast, and then study whether the individual forecast driver is important factor behind the forecast or not. If its values are close to zero, compared with the movement in dependent variable, then clearly the associated regressor is not an important driver of the forecast. This can either mean that the regressor has not recorded significant movement in given period, or, more interestingly, that the coefficient is

just too small to make the regressor an important driver of forecasts. If it is the latter then it is a signal that the model does not feature sufficient responses to shocks in given regressor, which might be a cause for model re-specifications.

Figure H.26 provides first illustration of the forecast decomposition tool. It decomposes a downside scenario forecast for house price growth into individual components.[61] We can see that house price growth collapses in beginning of the scenario and then recovers later on. What is behind this pattern? The most important driver, the one with largest movements, is the second driver, change in unemployment rate gap. This driver explains more than half of the decrease in the dependent variable, as well as the subsequent increase. Meanwhile, the first driver, growth in disposable income per capita, explains most of the remainder in the movement in the dependent variable. Finally, the third driver, change in he 10-yr yield has more varied effect. Initially it has sizable negative effect, which then turns briefly positive before converging to virtually zero effect. The initially positive effect corresponds to the fact that this downside scenario features an initial spike in interest rates, and spike in interest rates leads to lower house prices. As interest rates decline later on, house prices are lifted. Finally, there is also a constant as a driver, but of course that does not show any movements; it acts as a level shift for the house price growth.

Figure H.26 was an example of decomposition of a particular scenario forecast. Meanwhile, Figure H.27 illustrates forecast decomposition for a historical forecasts, specifically forecast covering the Great recession. The main difference from the previous illustration is that this time around we need to distinguish between the actual historical values of the dependent variable and the forecasts made by the model.[62] That is why the figure includes two separate series for the dependent variable: the actual values in blue with squares, and the model forecast that corresponds to the sum of the individual drivers in purple line with pluses. It is visible that the two are quite different, corresponding to the fact that the model fails to predict the large drop in house prices in the first two quarters of 2009. Otherwise the decomposition graph works the same. For example, we can see that by far the largest effect on house price growth in forecast covering the great recession comes from second driver, the unemployment rate gap, but first driver also contributes.

[61] The exact model from which forecasts were made is following:

$$@pc(house_prices) = \quad c \quad @movav(@pc(disposable_income(-1)/population(-1)),4) \quad @d(@movav(unp_rate-nairu,8)) \quad d(yield_10y)$$

[62] In case of scenario forecasts this made little sense, but even there it can be relevant in situation when the scenario forecast is not the same as model forecast. For example, the forecast can include expert judgment adjustments.

Figure 8.26: Illustration of forecast decomposition

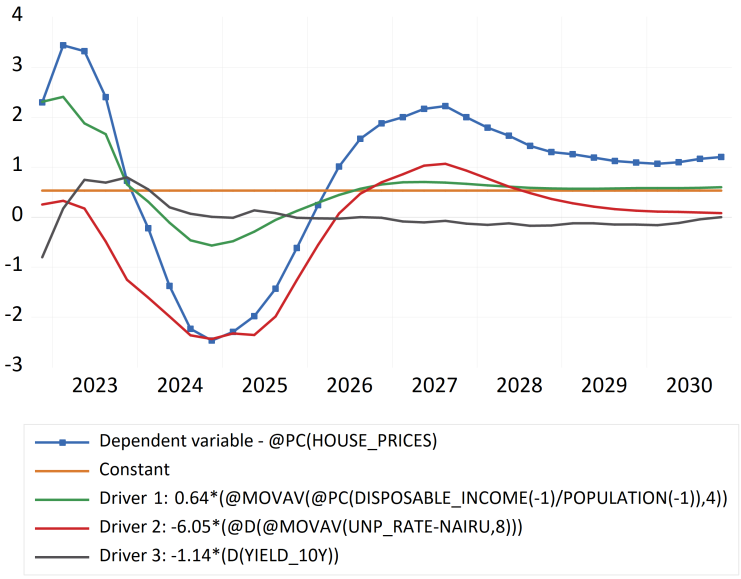
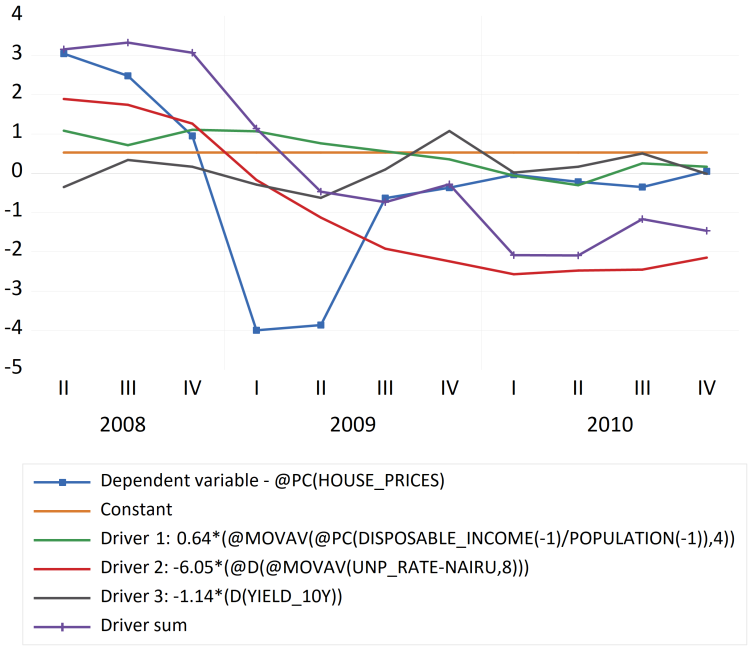


Figure 8.27: Illustration of forecast decomposition - Historical forecast



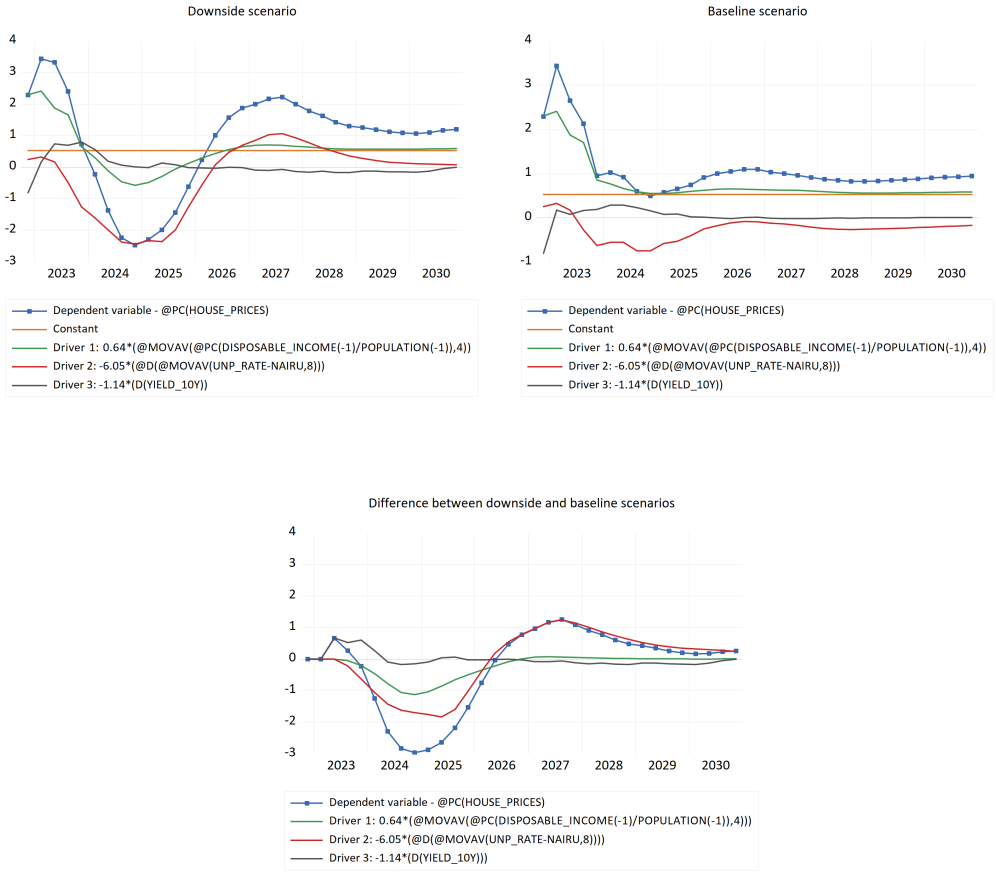
While this basic decomposition can be very useful in analyzing the shock responses of our model,

in the context of scenario shocks there is a modification that can be even more useful. Specifically, rather than decomposing and analyzing the scenario forecast itself, we might be more interested in decomposing the scenario forecast relative to some baseline forecast. In other words, we might wish the decompose the difference between two forecasts, $y_t^{scenario} - y_t^{baseline}$, into the individual forecast drivers. This is straightforward extension of the forecast decomposition tool. We first decompose the two individual forecasts into their forecast drivers, and then we subtract the values of the forecast drivers in the two scenarios from each other, e.g. $\beta_1 x_t^{scenario} - \beta_1 x_t^{baseline}$. One can call this term the *forecast difference driver*.

To appreciate why is this beneficial, return to Figure H.26. The figure shows a decomposition of a downside scenario, in which one would expect that house prices decline as a result of negative effect of all (or most) drivers. But that is not immediately visible in the figure: the dependent variable is positive in beginning and in the end; similarly, the first driver has positive influence in beginning and in the end. But this does not mean that the downside scenario is *relatively* "good", or that the first driver has a *relatively* positive effect. To investigate such *relative* perspective we need to compare the downside scenario with some reference scenario. This should be obvious even from Figure H.26: one of the drivers is a constant, which has a positive effect in the downside scenario; of course, effect of constant is the same in all scenarios, so relatively the effect of constant term is zero.

Figure H.28 illustrates this. Top left panel reproduces the decomposition of downside scenario from Figure H.26. Meanwhile, top right panel provides decomposition of the reference (baseline) scenario. The bottom panel then shows the difference between these two, what is the forecast difference decomposition.

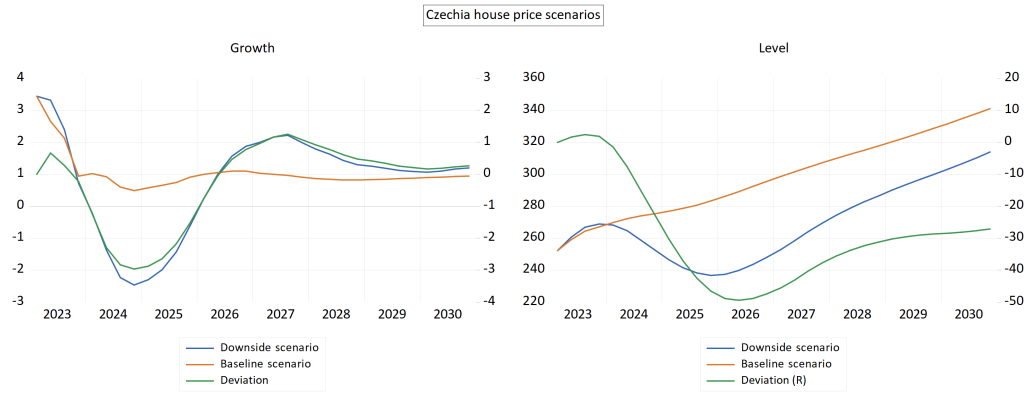
Figure 8.28: Illustration of forecast difference decomposition



From this figure we can see that most of the positive values we saw in Figure H.26 are also present baseline scenario. Hence, relatively speaking, downside scenario is indeed a negative scenario. This is especially true when looking at some of the drivers. While the first driver was initially positive in the downside scenario, in relative terms it is immediately negative, and never really becomes positive. But there are also opposite conclusions: the second driver is not as negative relatively to baseline, as it was when looking at the downside scenario alone. This means that some of the negative effect was not specific to the downside scenario, but rather was feature of the baseline scenario. These aspects highlight how in case of reference and alternative scenarios it can be important to look at difference in forecast drivers, not just their absolute values, to appreciate the relative effect of each driver.

Discussion. As a final word, it should be made clear what kind of decomposition we can and we cannot do. Specifically, in the context of models that include time series transformation of dependent variable, such as a growth rates or log-differences, we can decompose the movements in this transfor-

Figure 8.29: Illustration of forecast difference decomposition (different transformations)



mation. However, there is no way how to simply decompose the movements in the underlying variable. For example, since undoing log-difference leads to a multiplicative relationship, the forecast drivers are no longer linearly additive. Hence, forecast decomposition is limited to the actual transformation of the underlying dependent variable used.

This also explains why in the bottom panel of Figure H.28 the dependent variable was still positive in middle of the sample even though we are looking at downside scenario: Simply following an initial drop the scenario forecast features some recovery in house prices, which means above baseline growth. This does not mean that the level of house prices rise above the baseline level; rather that they *partially* converge back to it. Figure H.29 illustrates this point, showing growth rates and levels of house prices, and for both deviation between downside and baseline scenarios on the right.

Similarly, when the model features lagged dependent variable(s) then delayed effects of individual forecast drivers are captured as the effect of the lagged dependent variable(s). In other words, forecast decomposition captures only the immediate effects. This might make it sometimes look like that the forecast drivers have small effects, even if in reality they have large, delayed effects.

Finally, note that the value of forecast decomposition tool is not limited the studying of shocks responses. The tool can be also leveraged whenever understanding the shape of particular forecast is not easy. In such situation, using the forecast decomposition allows the forecaster to identify the individual sources of fluctuations in the forecast for the dependent variable and hence understand the forecast itself. Similarly, one can use the decomposition into forecast driver differences to understand drivers of the difference between two alternative forecasts. In either case, the tool is used for interpreting some

existing forecasts.[63]

8.4.4 Multiple-equation models

The discussion so far focused on single-equation time series models, in line with the focus of the whole textbook. While the single equation perspective is also useful in context of multiple equation models, there are some specifics that apply only to multiple equation models. Therefore, it is worth briefly discussing how things are different in multiple-equation models.

The main difference between single equation and multiple equation models from the perspective of studying sensitivity to shocks is the possibility for dynamic feedback. In single equation models there is single dependent/underlying variable. In Chapter **??** we have seen that in multivariable models an initial shock can lead to further delayed shocks: since the original shock changes the dependent variable, and since the regressors can include lagged dependent variable, then current changes to dependent variable lead to future changes in dependent variable. Meanwhile, in appendix to Chapter **??** we have briefly seen that in multiequation models the scope for such feedbacks is much greater, since we have multiple dependent variables that can change and cause delayed effects, not only for themselves, but also for other variables. So shock affecting one variable can gradually propagate across the system of variables.

This introduces additionally layer of complexity to shock responses, and hence their study. For example, if initial shock to one variable causes only small response in another variable, is it because the link between the shocked variable and the response variable is weak? Or is it because the propagation of shocks in the shocked variable is weak? Or is it because the link to third variable, which is responsible for transmitting shocks between first and second variable, is itself weak?

The complex feedback channels also raise a question of how do we treat the shocked variable. In single-equation models this question was irrelevant since the independent variables were always determined outside of our model. In multi-equation models this is no longer necessarily true, as the shocked variable might be one of the dependent variables. It is then question whether the shocked variable should be allowed to respond to the feedback caused by shock originating in it. For example, if we are studying responses of multiple equation system to unemployment rate shock, do we want the unemployment rate to reflect the changes in other variables, or do we want to keep unemployment rate

[63] In this use of the tool there might be a need for additional driver on top of the drivers corresponding to regressors: the add-factor series. This is a series used for explicit adjustments to particular forecast, and hence becomes a possibly important source of fluctuations in the forecast.

path as given?

In most cases, we are interested in shock responses that account for the feedbacks to the shocked variable. This then means that shocks are introduced as innovation shocks: we specify by how much does unemployment rate change in one period and the let it evolve (i.e. continue treating it as endogenous), rather than specifying the full path for unemployment rate. However, in some situations we do not want to allow for feedbacks back to the shocked variable. In such situation we treat it as exogenous - i.e. no longer determined by the associated model equation - and specify the shock in terms of specifying the full path for the variable.

Finally, the use of innovation shocks also has a bearing on the issue of single vs. multiple shocks approaches. In many multi-equation models the model residuals/innovations are expected to be correlated. While in principle one can still study responses of the model to single innovation shock, in practice in most cases it is most appropriate to study responses to whole vector of shocks. In other words, one accounts for the correlation between the innovation shocks. This is what is the standard practice of the VAR literature.

8.5 Putting it all together: Model building process

We now have understanding of all the tools in our toolkit for model building. That leaves us in position of a chef or a doctor who has learned about all the tools there are in kitchen or hospital and how should they be used. While this is a logical starting point, it is not on its own enough. Apart from knowing what the tools are and how to use them, we also have to know *when* to use them and *to what effect*. This section will provide some guidance on this.

However, to avoid over-promising and under-delivering, it is useful to stress from the very beginning that there is no single answer to the question of when and to what effect should which tools be used. As in cooking or operating, different tools are useful in different moments, and there is no single plan that can be followed. Chefs have different recipes for different meals and doctors have different orders of tests to run and procedures to perform depending on the situation of the patient. Indeed, the situation of a model builder is more similar to doctor than to a chef: while a given recipe provides a well thought-though plan from which there is usually no reason to deviate, doctors situation is more complicated, as they have to decide on what is the appropriate next step after observing results of the previous step. Similarly, building time series models cannot follow fixed pre-scheduled plan, but rather has to adjust and adapt based on how previous stages of model building worked out.

With this qualification in mind, we can proceed to discussing how a typical modelling process might look like. This is really about when and how should which tools should be used, and how to know what should be done next. The imagined situation is situation when you have been tasked with developing new, single-equation multivariate model for a time series which you have not encountered before. How should you proceed from the initial stage of not knowing anything about the time series to having a forecasting model for it? What should you do first, and when next?

As the very first step, before any modelling starts, you should always familiarize yourself with the time series at hand. This really means figuring out how the time series behaves, both theoretically and empirically. This constitutes a stage of model building process that can be called **background analysis**. In terms of theoretical perspective, this is about trying to obtain knowledge from other people's work relating to your time series. If your time series is fairly standard time series, such as some common macroeconomic variable, there likely is a basic theory that will help you understand how the series behaves. For example, there is a rich theoretical literature that links exchange rates to various macroeconomic factors, from prices, to interest rates to trade balance. If it is more quirky time series, there might not be anything of that sort, but you still might gain by searching the internet, or consulting the all-powerful ChatGPT, and seeing if there is something that can help you understand how your time series behaves.

Together with understanding theoretically how your time series behaves, you should also be investigating it empirically. At the most basic level this means looking at it. For a low frequency time series such as GDP or inflation there is no better way than to visualize the basic time series plot. This plot can reveal a lot of patterns and interesting features, from tendency for persistence, through existence of time-varying mean or variance, to presence of abnormal observations. For a high-frequency time series, such as daily variance of returns observed over period of several years, the number of observations is typically too large to make many useful conclusions based on time series plot. Rather, you will have to rely on distributional plots, ranging from simple probability distribution function or cumulative density function of the data, to the ACF and PACF plots we have seen throughout this book. These are of course useful also for the low-frequency series.

Once you are done with your background analysis, you are ready to start modelling. As stressed by this chapter, the model building process should be **iterative**. This means that you can start with the simplest model you can think of, knowing full well that this will not be the actual model. Rather, it will serve you as a starting point on which you will improve in iterative fashion. The second aspect

of the model building process is that it should be **interactive**. You take your model, proceed with analyzing it, and then based on the results of your analysis you propose improvements to the model. Like this there is an interaction between you, the modeler, and the results of your analysis of the current model.

What is the first thing you look at when analyzing your model, be it your initial model, or later generation model? In most cases, modeler will first be interested in the basic properties of the model summarized in the **estimation output**. They will want to check whether the coefficient signs correspond to the expectations, whether their magnitudes are appropriate and whether the values such as R^2 or Durbin-Watson statistic suggest some problems. As discussed in the dedicated section, this information is mostly used in form of yes-or-no decision-making, with some models ruled out. Or it is used as a signal of existence of problems, but not much in terms of suggestion how these could be fixed.

If given model passes the checks based on the estimation output, the most common next point of focus is the **forecast performance**. If single model is under analysis, then graphical analysis of overall forecasting performance is best place to start. The graphical analysis will often reveal periods when model does not perform satisfactory well. This might either lead to ideas how to enrich the model, such as realizing that some financial factor might be valuable given the poor performance during periods of financial stress. Or it might result in realization of inappropriate model structure, such as realization that one needs to use different dependent variable or different transformation of independent variable. As an example of the former one might often realize that spread is more appropriate dependent variable. As for the latter, one often concludes that given independent variable needs to be used in a lag or a moving average, to pick up the effects.

In this stage, modeler also needs to rely on the full toolbox with respect to analyzing forecast performance. The overall forecast performance can be possibly complemented by the analysis of performance during specific periods of interests, such as the global financial crisis. Good choice of the transformation in which forecasts are shown can often be a key to realizing model improvements. Alternatively, inclusion of potential independent variables in the same graph as the individual forecasts can reveal correlation between such variable and forecast performance and hence directly lead to suggestions of what else should be included in the model.

Similarly, the information from forecasting performance analysis should be looked at through the prism of information contained in the estimation output. Most importantly, coefficients sign and

size can be directly linked to forecasting behavior. Another example is the information in Durbin-Watson statistic, which can suggest problems with autocorrelated residuals. This will often manifest as insufficient or too strong persistence in forecasts.

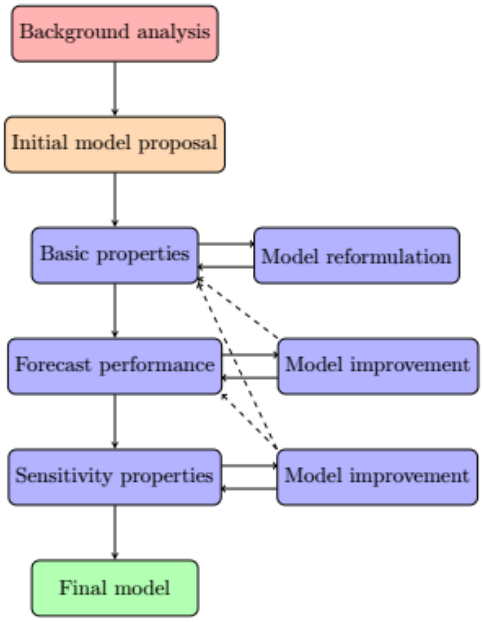
While there is no beating of this graphical analysis, sometimes this approach is not feasible. This is the case when working with high frequency data, where there are just too many forecasts to visualize. Similarly, graphical analysis becomes very inefficient when comparing more than few models at a given time. For example, when one wants to know how does model performance change when model includes moving average with increasing length for some independent variable, graphs for different model specifications become too similar to distinguish. For this reason, graphical analysis can always be complemented by numerical analysis. Numerical analysis can provide quick summary of which model is better, potentially with different answer for different horizon or period. This on its own can sometimes produce suggestions for model improvement.

Once model modification/improvement is identified, and new model is estimated, typically one should return to the first stage of the modelling process and analyze the basic properties of the new model, given that coefficient sign and size are typically necessary conditions which the model should pass. If it does, then one can proceed to the analysis of forecasting performance, and specifically to whether the modification yielded the expected/desired improvement in model performance. This either results into negative conclusion, in which case further model modifications in the same spirit might be performed. Or it can result in positive conclusion, meaning that the model performance did improve. In either case, one proceeds with further model modifications based on the analysis. This is then the essence of the interactive and interactive modelling process: in any given iteration modeler considers one or several candidate models; these models are evaluated using the tools discussed in this chapter, and based on this analysis, modeler comes up with modifications/improvements to the model.[64]

Finally, the information from estimation output and forecast performance can be often usefully complemented by information on **sensitivity to shocks**. This provides much more detailed look on the model behavior then either of the other two parts of model building process. In effect it explains how the model translates inputs, i.e. movements in independent variables, into outputs, i.e. the

[64]Note that in essence this is not much different from the principles on which machine-learning is based. Machine learning is also at heart an iterative process, where the algorithm goes through large number of iterations. Each iteration, in turn, consists of learning and improvement. The difference here is the nature of learning and improvements: In machine learning the learning necessarily based on numerical factors and improvements follow some pre-defined pattern; here, the learning is not limited to numerical factors and improvements are relying on creativity of human brain. Insofar as such non-numerical information and human creativity are superior to a numerical learning and pre-defined pattern *in context of given modelling task*, this process is superior to a machine-learning process. Of course, there are plenty of applications where the opposite is true. Which is just saying that the modelling approach discussed in this chapter is not well suited to all possible applications.

Figure 8.30: Model building workflow



individual forecasts. As such, it can often point towards problematic aspects of the model which can then be linked to subpar forecast performance. This is true either for the individual shock responses, or for the historical and scenario approaches to multiple shock responses. The latter can often be also source of information about problematic model structure: a scenario forecast that is un-intuitive can often be linked to inappropriate model structure.

We have outlined how the the tools we discussed in our three sections of this chapter fit into our iterative and interactive model building process. The whole process is schematically captured in Figure H.30. The figure includes all the main steps of the process and captures also the key aspect of the process: While model building can in some sense be consider a linear process, starting with background analysis and ending with a final model, in practice it is anything but. The process is interactive and iterative. It is interactive because there is a lot of learning on part of the modeller, be it from the basic model properties entailed in estimation output or from the detailed analysis of model behavior in sensitivity to shocks. It is iterative in that there are lot of model reformulations which often put us back into the initial position, and there are lot of dead ends.

The key thing to take away is that at this point we have many tools at our disposal and we know when and how to use them to analyze our model, and based on such analysis propose improvements to the model. We will hammer home this knowledge in next section, which will try to apply all of it

to particular model building task.

8.6 Application: Interbank interest rates

8.7 Appendix: Statistical tests and their role in model building