Chapter 3

Predictive Regression: Review and Interpretation

Ideas that fall under the general heading of "**regression analysis**" are crucial for building forecasting models, using them to produce forecasts, and evaluating those forecasts. Here we provide a linear regression refresher. Again, be warned: this chapter is no substitute for a full-introduction to regression, which you should have had already.

3.1 Regression as Curve Fitting

3.1.1 Simple Regression

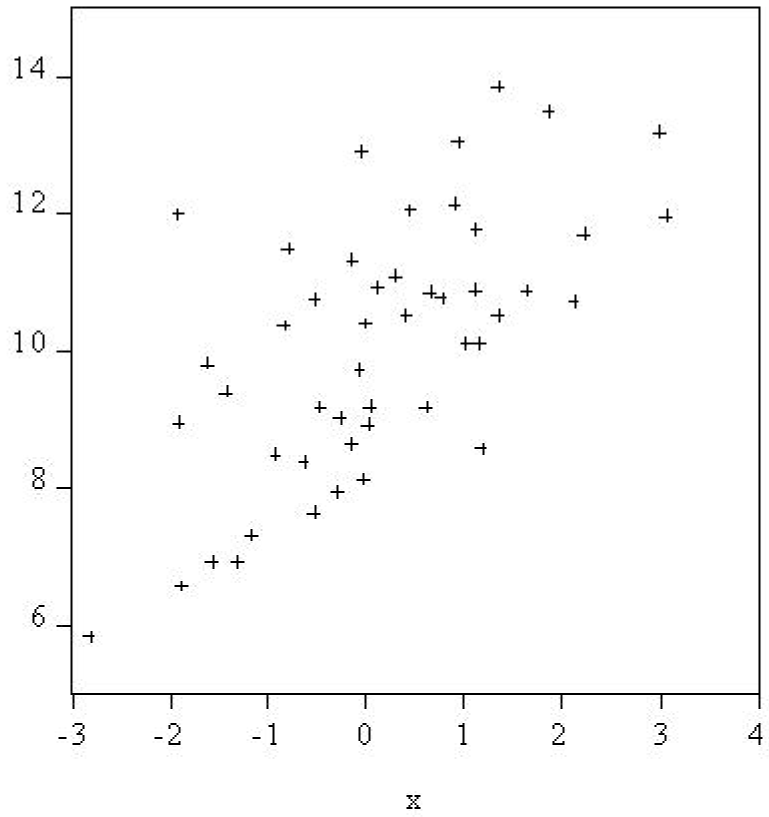
Suppose that we have data on two variables ("simple," or "bivariate," regression), y and x, as in Figure 1, and suppose that we want to find the linear function of x that best fits the data points, in the sense that the sum of squared vertical distances of the data points from the fitted line is minimized. When we "run a regression," or "fit a regression line," that's what we do. The estimation strategy is called **least squares**.

In Figure 2, we illustrate graphically the results of regressing y on x, which we sometimes denote by $y \rightarrow c, x$.[1] The best-fitting line slopes upward,

[1] The "c" denotes inclusion of a constant, or intercept, term.

Figure 1

Scatterplot of y versus x



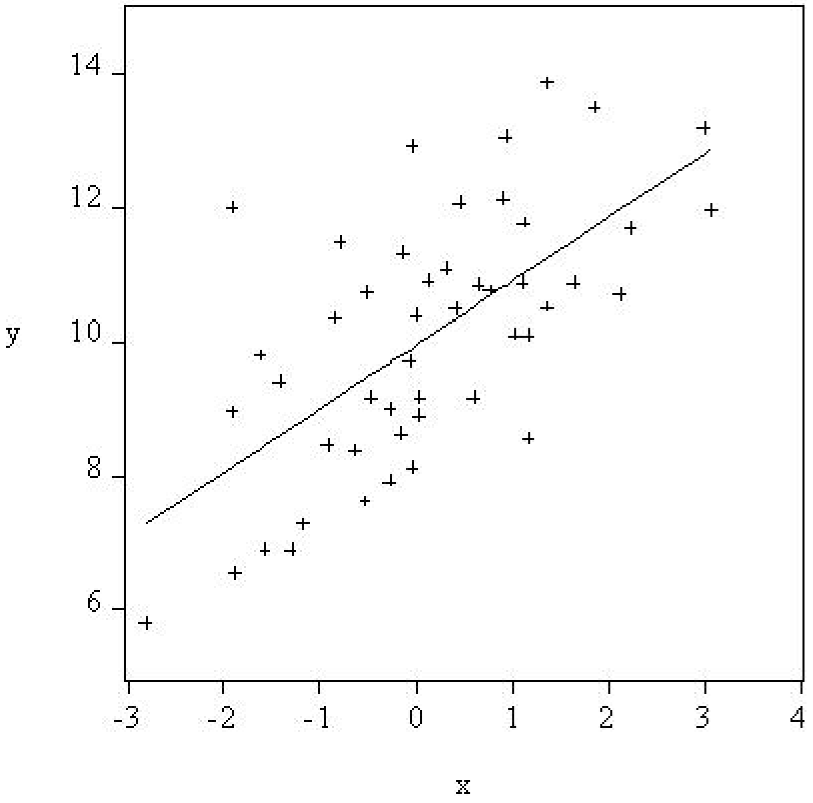
reflecting the positive correlation between y and x. Note that the data points don't satisfy the fitted linear relationship exactly; rather, they satisfy it on average.

Let us elaborate on the fitting of regression lines, and the reason for the name "least squares." When we run the regression, we use a computer to fit the line by solving the problem

$$\min_{\beta} \sum_{t=1}^{T} (y_t - \beta_1 - \beta_2 x_t)^2,$$

where β is shorthand notation for the set of two parameters, β_1 and β_2. We denote the set of estimated, or fitted, parameters by $\hat{\beta}$, and its elements by

Figure 2
Scatterplot of y versus x
Regression Line Superimposed



$\hat{\beta}_1$ and $\hat{\beta}_2$.

The regression **fitted values** are

$$\hat{y}_t = \hat{\beta}_1 + \hat{\beta}_2 x_t,$$

$t = 1, ..., T$. The regression **residuals** are simply the difference between actual and fitted values. We write

$$e_t = y_t - \hat{y}_t,$$

$t = 1, ..., T$.

In in all linear regressions (even with multiple RHS variables, to which we turn shortly), the least-squares estimator has a simple formula. We use a computer to evaluate the formula, simply, stably, and instantaneously.

3.1.2 Multiple Regression

Extension to the general multiple linear regression model, with an arbitrary number of right-hand-side variables (K, including the constant), is immediate. We simply run $y \rightarrow c, x_2, ..., x_K$, again picking the parameters to minimize the sum of squared residuals, and everything goes through as in the case of simple regression.

The least squares estimator is

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y, \tag{3.1}$$

where X is a $T \times K$ matrix,

$$X = \begin{pmatrix} 1 & x_{21} & x_{31} & \dots & x_{K1} \\ 1 & x_{22} & x_{32} & \dots & x_{K2} \\ \vdots & & & & \\ 1 & x_{2T} & x_{3T} & \dots & x_{KT} \end{pmatrix}$$

and y is a $T \times 1$ vector, $y' = (y_1, y_2, ..., y_T)$. The time-$t$ fitted value is

$$\hat{y}_t = x_t'\hat{\beta},$$

where $x_t' = (x_{1t}, ..., x_{Kt})$ is the time-t vector of x's, and the time-t residual is

$$e_t = y_t - \hat{y}_t.$$

The vector of fitted values is

$$\hat{y} = X\hat{\beta},$$

and the vector of residuals is

$$e = y - \hat{y}.$$

3.2 Regression as a Probability Model

We work with the full multiple regression model; simple regression is of course a special case.

3.2.1 A Population Model and a Sample Estimator

Thus far we have *not* postulated a probabilistic model that relates y_t and x_t; instead, we simply ran a mechanical regression of y_t on x_t to find the best fit to y_t formed as a linear function of x_t. It's easy, however, to construct a probabilistic framework that lets us make statistical assessments about the properties of the fitted line. Assume, for example, that y_t is linearly related to an exogenously-determined x_t, with an independent and identically distributed zero-mean (iid) Gaussian **disturbance**:

$$y_t = \beta_1 + \beta_2 x_{2t} + ... + \beta_K x_{Kt} + \varepsilon_t = x_t' \beta + \varepsilon_t$$

$$\varepsilon_t \sim iid N(0, \sigma^2),$$

$t = 1, ..., T$. The intercept of the line is β_1, the slope parameters are the β_i's, and the variance of the disturbance is σ^2.[2] Collectively, we call the β's the model's **parameters**. The index t keeps track of time; the data sample begins at some time we've called "1" and ends at some time we've called "T", so we write $t = 1, ..., T$. (Or, in cross sections, we index cross-section units by i and write $i = 1, ..., N$.)

In this linear regression model the expected value of y_t conditional upon x_t taking a particular value, say x_t^*, is

$$E(y_t | x_t = x_t^*) = {x_t}^{*\prime} \beta.$$

That is, the **regression function** is the **conditional expectation** of y_t.

[2] We speak of the **regression intercept** and the **regression slope**.

We assume that the the linear model sketched is true in population; that is, it is the **data-generating process (DGP)**. But in practice, of course, we don't know the values of the model's parameters, β_1, β_2, ..., β_K and σ^2. Our job is to *estimate* them using a sample of data from the population. We estimate the β's precesiely as before, using the computer to solve $\min_\beta \sum_{t=1}^T \varepsilon_t^2$.

3.2.2 Notation, Assumptions and Results: The Full Ideal Conditions

The discussion thus far was intentionally a bit loose, focusing on motivation and intuition. Let us now be more precise about what we assume and what results obtain.

A Bit of Matrix Notation

One reason that vector-matrix notation is useful is because the probabilistic regression model can be written very compactly using it. We have written the model as

$$y_t = \beta_1 + \beta_2 x_{2t} + ... + \beta_K x_{Kt} + \varepsilon_t, \;\; t = 1, ..., T.$$

$$\varepsilon_t \sim iid\, N(0, \sigma^2)$$

Now stack $\varepsilon_t, t = 1, ..., T$, into the vector ε, where $\varepsilon' = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_T)$. Then we can write the complete model over all observations as

$$y = X\beta + \varepsilon \tag{3.2}$$

$$\varepsilon \sim N(\underline{0}, \sigma^2 I). \tag{3.3}$$

This concise representation is very convenient.

Indeed representation (3.2)-(3.3) is crucially important, not simply because it is concise, but because the various assumptions that we need to make to get various statistical results are most naturally and simply stated on X and ε in equation (3.2).

The most restrictive set of assumptions is known as the "full ideal conditions" (FIC), which are so strict as to be nearly preposterous in economic contexts, and most of econometrics is devoted to confronting various *failures* of the FIC. But before we worry about FIC failures, it's useful first to recall what happens when they hold.

Assumptions: The Full Ideal Conditions (FIC)

1. The DGP is (3.2)-(3.3), and the fitted model matches the DGP exactly.

2. X is fixed in repeated samples.

3. X is of full column rank (K).

FIC 1 has many important sub-conditions embedded. For example:

1. Linear relationship, $E(y) = X\beta$

2. Fixed coefficients, β

3. $\varepsilon \sim N$

4. ε has constant variance σ^2

5. The ε's are uncorrelated.

FIC 2 says that re-running the world to generate a new sample y^* would entail simply generating new shocks ε^* and running them through equation (3.2):

$$y^* = X\beta + \varepsilon^*.$$

That is, X would stay fixed across replications.

FIC 3 just says “no multicollinearity” – i.e., no redundancy among the variables contained in X (more precisely, no regressor is a perfect linear combination of the others).

Results Under the FIC

The least squares estimator remains

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y,$$

but in a probabilistic interpretation under the FIC, we can say a great deal about its statistical properties. Among other things, it is miniumum-variance unbiased (MVUE) and normally distributed with covariance matrix $\sigma^2(X'X)^{-1}$. We write

$$\hat{\beta}_{OLS} \sim N\left(\beta,\ \sigma^2(X'X)^{-1}\right).$$

We estimate the covariance matrix $\sigma^2(X'X)^{-1}$ using $s^2(X'X)^{-1}$, where

$$s^2 = \frac{\sum_{t=1}^{T} e_t^2}{T-K}.$$

3.3 A Typical Regression Analysis

Consider a typical regression output, which we show in Table 1. We do so dozens of times in this book, and the output format and interpretation are always the same, so it’s important to get comfortable with it quickly. The output is in Eviews format. Other software will produce more-or-less the same information, which is fundamental and standard.

The results begin by reminding us that we’re running a least-squares (LS) regression, and that the left-hand-side variable is y. It then shows us the sample range of the historical data, which happens to be 1960 to 2007, for a total of 48 observations.

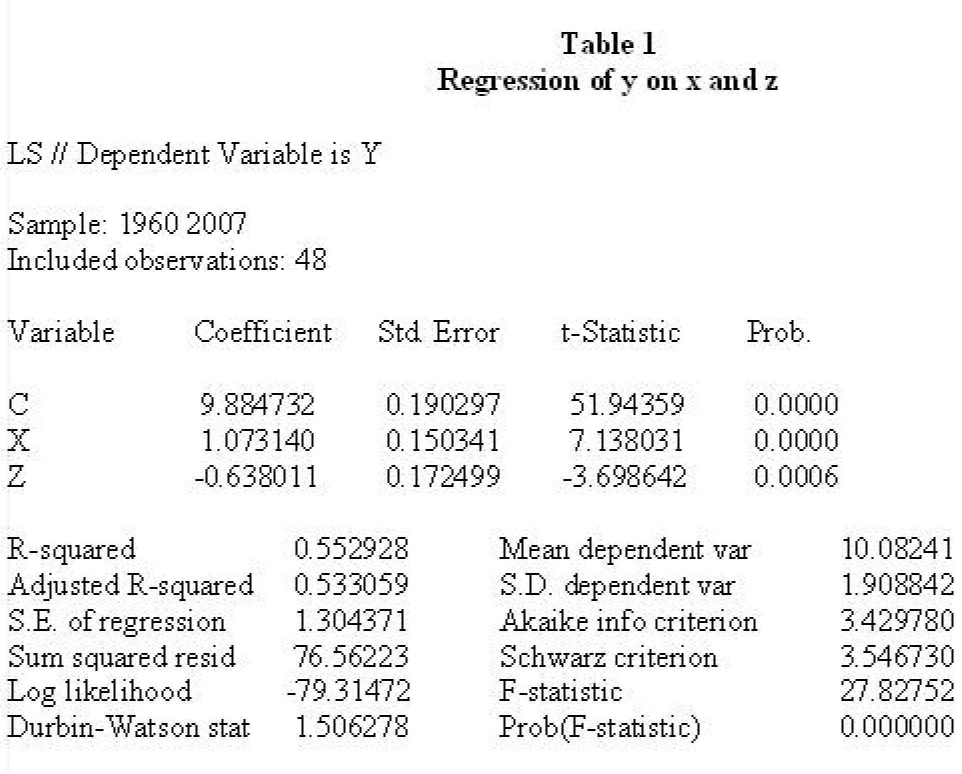


Table 1
Regression of y on x and z

```
LS // Dependent Variable is Y

Sample: 1960 2007
Included observations: 48

Variable      Coefficient    Std. Error    t-Statistic    Prob.

C              9.884732      0.190297      51.94359      0.0000
X              1.073140      0.150341      7.138031      0.0000
Z             -0.638011      0.172499     -3.698642      0.0006

R-squared             0.552928    Mean dependent var      10.08241
Adjusted R-squared    0.533059    S.D. dependent var      1.908842
S.E. of regression    1.304371    Akaike info criterion   3.429780
Sum squared resid     76.56223    Schwarz criterion       3.546730
Log likelihood       -79.31472    F-statistic             27.82752
Durbin-Watson stat    1.506278    Prob(F-statistic)       0.000000
```

Next comes a table listing each right-hand-side variable together with four statistics. The right-hand-side variables x and z need no explanation, but the variable c does. c is notation for the earlier-mentioned constant variable. The c variable always equals one, so the estimated coefficient on c is the estimated intercept of the regression line.[3]

3.3.1 Coefficient Estimates, Standard Errors, t Statistics and p-Values

The four statistics associated with each right-hand-side variable are the estimated coefficient ("Coefficient"), its standard error ("Std. Error"), a t statistic, and a corresponding probability value ("Prob.").

The "coefficients" are simply the regression coefficient estimates. Per the OLS formula that we introduced earlier in equation (3.1), they are the elements of the $(K \times 1)$ vector, $(X'X)^{-1}X'y$.

[3] Sometimes the population coefficient on c is called the **constant term**, and the regression estimate is called the estimated constant term.

The **standard errors** of the estimated coefficients indicate their sampling variability, and hence their reliability. In line with result (**??**) above, the ith standard error is

$$s\sqrt{(X'X)^{-1}_{ii}},$$

where $(X'X)^{-1}_{ii}$ denotes the ith diagonal element of $(X'X)^{-1}$, and s is an estimate (defined below) of σ.

The estimated coefficient plus or minus one standard error is approximately a 68% confidence interval for the true but unknown population parameter (contribution to the conditional expectation), and the estimated coefficient plus or minus two standard errors is approximately a 95% confidence interval, assuming that the estimated coefficient is approximately normally distributed.[4] Thus large coefficient standard errors translate into wide confidence intervals.

Each t **statistic** provides a test of the hypothesis of variable irrelevance: that the true but unknown population parameter (contribution to the conditional expectation) is zero, so that the corresponding variable contributes nothing to the conditional expectation and can therefore be dropped. One way to test this variable irrelevance, with, say, a 5% probability of incorrect rejection, is to check whether zero is outside the 95% confidence interval for the parameter. If so, we reject irrelevance. The t statistic is just the ratio of the estimated coefficient to its standard error, so if zero is outside the 95% confidence interval, then the t statistic must be bigger than two in absolute value. Thus we can quickly test irrelevance at the 5% level by checking whether the t statistic is greater than two in absolute value.[5]

Finally, associated with each t statistic is a **probability value**, which is the probability of getting a value of the t statistic at least as large in

[4] Coefficients will be approximately normally distributed in large samples quite generally, and exactly normally distributed in samples of any size if the regression disturbance is normally distributed.

[5] In large samples the t statistic is distributed $N(0,1)$ quite generally. In samples of any size the t statistic follows a t distribution if the regression disturbances are Gaussian.

absolute value as the one actually obtained, assuming that the irrelevance hypothesis is true. Hence if a t statistic were two, the corresponding probability value would be approximately .05 (asstuming large T and/or Gaussian disturbances). The smaller the probability value, the stronger the evidence against irrelevance. There's no magic cutoff, but typically probability values less than 0.1 are viewed as strong evidence against irrelevance, and probability values below 0.05 are viewed as very strong evidence against irrelevance. Probability values are useful because they eliminate the need for consulting tables of the t or z distributions. Effectively the computer does it for us and tells us the significance level at which the irrelevance hypothesis is just rejected.

Now let's interpret the actual estimated coefficients, standard errors, t statistics, and probability values. The estimated intercept is approximately 10, so that conditional on x and z both being zero, we expect y to be 10. Moreover, the intercept is very precisely estimated, as evidenced by the small standard error relative to the estimated coefficient. An approximate 95% confidence interval for the true but unknown population intercept is $10 \pm 2(.19)$, or [9.62, 10.38]. Zero is far outside that interval, so the corresponding t statistic is huge, with a probability value that's zero to four decimal places.

The estimated coefficient on x is 1.07, and the standard error is again small in relation to the size of the estimated coefficient, so the t statistic is large and its probability value small. Hence at conventional levels we reject the hypothesis that x contributes nothing to the conditional expectation $E(y|x,z)$. The estimated coefficient is positive, so that x contributes positively to the conditional expectation; that is, $E(y|x,z)$ is larger for larger x, other things equal.

The estimated coefficient on z is -.64. Its standard error is larger relative to the estimated parameter; hence its t statistic is smaller than those of the other coefficients. The standard error is nevertheless small, and the absolute value of the t statistic is still well above 2, with a small probability value of .06%. Hence at conventional levels we reject the hypothesis that z contributes nothing to the conditional expectation $E(y|x, z)$. The estimated coefficient is negative, so that z contributes negatively to the conditional expectation; that is, $E(y|x, z)$ is smaller for larger z, other things equal.

3.3.2 Residual Plot

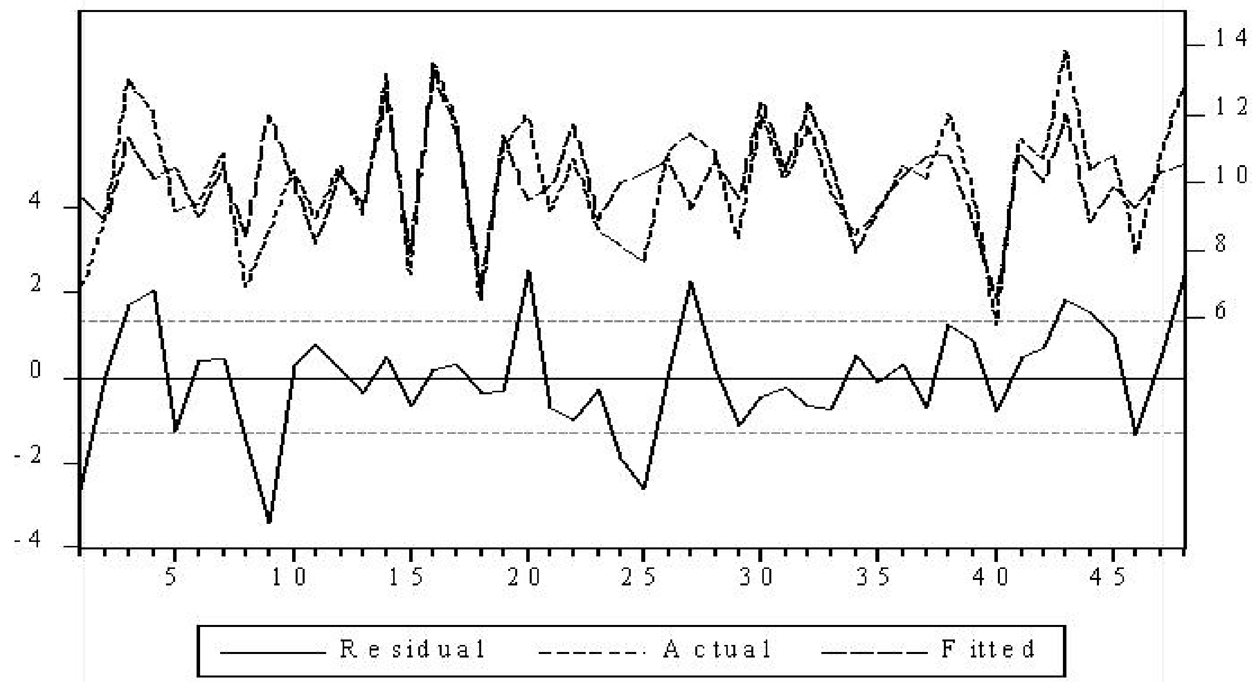
After running a time-series regression, it's usually a good idea to assess the adequacy of the model by plotting over time and examining the actual data (y_t's), the fitted values ($\hat{y}_t$'s), and the residuals (e_t's). Often we'll refer to such plots, shown together in a single graph, as a **residual plot**.[6] In Figure 4 we show the residual plot for the regression of $y \rightarrow c, x, z$. The actual (short dash) and fitted (long dash) values appear at the top of the graph; their scale is on the right. The fitted values track the actual values fairly well. The residuals appear at the bottom of the graph (solid line); their scale is on the left. It's important to note that the scales differ; the e_t's are in fact substantially smaller and less variable than either the y_t's or the $\hat{y}_t$'s. We draw the zero line through the residuals for visual comparison. There are no obvious patterns in the residuals.

Residual plots are obviously useful in time-series perspective, and not useful in cross sections, for which there is no natural ordering of the data. In cross sections, however, we often examine **residual scatterplots**, that is, scatterplots of y vs. $\hat{y}$ for all observations in the cross section, with special attention paid to the general pattern of deviations from the forty-five degree line.

A variety of diagnostic statistics follow; they help us to evaluate the adequacy of the regression. Here we review them very briefly.

[6] Sometimes, however, we'll use "residual plot" to refer to a plot of the residuals alone. The intended meaning will be clear from context.

Figure 4
Residual Plot
Regression of y on x and z



3.3.3 Mean dependent var

The **sample mean of the dependent variable** is

$$\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$$

It measures the central tendency, or location, of y.

3.3.4 S.D. dependent var

The **sample standard deviation of the dependent variable** is

$$SD = \sqrt{\frac{\sum_{t=1}^{T}(y_t - \bar{y})^2}{T-1}}.$$

It measures the dispersion, or scale, of y.

3.3.5 Sum squared resid

Minimizing the **sum of squared residuals** is the objective of least squares estimation. It's natural, then, to record the minimized value of the sum of squared residuals. In isolation it's not of much value, but it serves as an input to other diagnostics that we'll discuss shortly. Moreover, it's useful for comparing models and testing hypotheses. The formula is

$$SSR = \sum_{t=1}^{T} e_t^2.$$

3.3.6 F−statistic

We use the F **statistic** to test the hypothesis that the coefficients of all variables in the regression except the intercept are jointly zero.[7] That is, we test whether, taken jointly as a set, the variables included in the model contribute nothing to the expectation of y conditional on the variables. This contrasts with the t statistics, which we use to examine the contributions of the variables one at a time.[8] If no variables contribute, then if the regression disturbances are Gaussian the F statistic follows an F distribution with $K-1$ and $T-K$ degrees of freedom. The formula is

$$F = \frac{(SSR_{res} - SSR)/(K-1)}{SSR/(T-K)},$$

where SSR_{res} is the sum of squared residuals from a *restricted* regression that contains only an intercept. Thus the test proceeds by examining how much SSR increases when all the variables except the constant are dropped. If it increases by a great deal, there's evidence that at least one of the variables

[7] We don't want to restrict the intercept to be zero, because under the hypothesis that all the other coefficients are zero, the intercept would equal the mean of y, which in general is not zero.

[8] In the case of only one right-hand-side variable, the t and F statistics contain exactly the same information, and one can show that $F = t^2$. When there are two or more right-hand-side variables, however, the hypotheses tested differ, and $F \neq t^2$.

contributes to the conditional expectation.

3.3.7 Prob(F−statistic)

The probability value for the F statistic gives the significance level at which we can just reject the hypothesis that the set of right-hand-side variables makes no contribution to the conditional expectation. Here the value is indistinguishable from zero, so we reject the hypothesis overwhelmingly.

3.3.8 S.E. of regression

We'd like an estimate of σ^2, because σ^2 tells us whether the regression "fit" is good. The observed residuals, the e_t's , are effectively estimates of the unobserved population disturbances, the ε_t's. Thus the sample variance of the e's, which we denote s^2 (read "**s-squared**"), is a natural estimator of σ^2:

$$s^2 = \frac{\sum_{t=1}^{T} e_t^2}{T-K}.$$

s^2 is an estimate of the dispersion of the regression disturbance and hence is used to assess goodness of fit. The larger is s^2, the worse the model's fit. s^2 involves a degrees-of-freedom correction (division by $T-K$ rather than by T or $T-1$), which, but whether one divides by T ot $T-K$ is of no asymptotic consequence.

The **standard error of the regression** (SER) conveys the same information; it's an estimator of σ rather than σ^2, so we simply use s rather than s^2. The formula is

$$SER = s = \sqrt{s^2} = \sqrt{\frac{\sum_{t=1}^{T} e_t^2}{T-K}}.$$

The standard error of the regression is easier to interpret than s^2, because its units are the same as those of the e's, whereas the units of s^2 are not. If the e's are in dollars, then the squared e's are in dollars squared, so s^2 is in dollars squared. By taking the square root at the end of it all, SER converts the units back to dollars.

3.3.9 R-squared

If an intercept is included in the regression, as is almost always the case, R-squared must be between zero and one. In that case, R-squared, usually written R^2, is the percentage of the variance of y explained by the variables included in the regression. R^2 is widely used as an easily-interpreted check of **goodness of fit**. Here the R^2 is about 55% – good but not great.

The formula is for R^2 is

$$R^2 = 1 - \frac{\sum_{t=1}^{T} e_t^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2}.$$

The key is the ratio on the right of the formula. First, note that the ratio must be positive (it exclusively involves sums of squares) and less than one (if the regression includes an intercept). Hence R^2, which is one *minus* the ratio, must be in $[0, 1]$. Second, note what the ratio involves. The numerator is SSR from a regression on *all variables*, and the denominator is the is SSR from a regression on an intercept alone. Hence the ratio is the fraction of variation in y *not* explained by the included x's, so that R^2 – which, again, is one *minus* the ratio – is the fraction of variation in y that *is* explained by the included x's.

We can write R^2 in a more roundabout way as

$$R^2 = 1 - \frac{\frac{1}{T}\sum_{t=1}^{T} e_t^2}{\frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})^2}.$$

This proves useful for moving to *adjusted* R^2, to which we now turn.

3.3.10 Adjusted R$-$squared

The interpretation is the same as that of R^2, but the formula is a bit different. Adjusted R^2 incorporates adjustments for the K degrees of freedom used in fitting the full model to y (numerator of the ratio), and for the 1 degree of freedom used in fitting the a mean to y (denominator of the ratio). As long as there is more than one right-hand-side variable in the model fitted, adjusted $\bar{R}^2$ is smaller than R^2; here, however, the two are typically very close (in this case, 53% vs. 55%). The formula for $\bar{R}^2$ is

$$\bar{R}^2 = 1 - \frac{\frac{1}{T-K}\sum_{t=1}^{T} e_t^2}{\frac{1}{T-1}\sum_{t=1}^{T}(y_t - \bar{y})^2},$$

where K is the number of right-hand-side variables, including the constant term. The numerator in the large fraction is precisely s_e^2, and the denominator is precisely s_y^2.

3.3.11 Durbin-Watson stat

We're always interested in examining whether there are patterns in residuals; if there are, the model somehow missed something systematic in the y data. The **Durbin-Watson statistic** tests for a certain kind of pattern, correlation over time, called **serial correlation**.

The Durbin-Watson test works within the context of the model

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 z_t + \varepsilon_t$$

$$\varepsilon_t = \phi\varepsilon_{t-1} + v_t$$

$$v_t \overset{iid}{\sim} N(0, \sigma^2).$$

The regression disturbance is serially correlated when $\phi \neq 0$. The hypothesis of interest is $\phi = 0$. When $\phi \neq 0$, the disturbance is serially correlated. More specifically, when $\phi \neq 0$, we say that ε_t follows an autoregressive process of order one, or $AR(1)$ for short.[9] If $\phi > 0$ the disturbance is positively serially correlated, and if $\phi < 0$ the disturbance is negatively serially correlated. **Positive serial correlation** is typically the relevant alternative in the economic and financial applications that will concern us.

The formula for the Durbin-Watson (DW) statistic is

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}.$$

DW takes values in the interval [0, 4], and if all is well, DW should be around 2. If DW is substantially less than 2, there is evidence of positive serial correlation. As a rough rule of thumb, if DW is less than 1.5, there may be cause for alarm, and we should consult the tables of the DW statistic, available in many statistics and econometrics texts. Here DW is very close to 1.5. A look at the tables of the DW statistic reveals, however, that we would not reject the null hypothesis at the five percent level. (Why Eviews neglects to print a p-value is something of a mystery.)

Note well that DW and its good properties involve several strict assumptions. Gaussian disturbances are required, and the $AR(1)$ alternative is the only one explicitly entertained, whereas in reality much richer forms of serial correlation may arise, and disturbances may of course be non-Gaussian. Subsequently we will introduce much more flexible approaches to testing/assessing residual serial correlation.

3.3.12 Akaike info criterion and Schwarz criterion

The Akaike and Schwarz criteria are used for model selection, and in certain contexts they have provable optimality properties in that regard. The

[9] The Durbin-Watson test is designed to be very good at detecting serial correlation of the $AR(1)$ type. Many other types of serial correlation are possible; we'll discuss them extensively later.

formulas are:

$$AIC = e^{\left(\frac{2K}{T}\right)} \frac{\sum_{t=1}^{T} e_t^2}{T}$$

and

$$SIC = T^{\left(\frac{K}{T}\right)} \frac{\sum_{t=1}^{T} e_t^2}{T}.$$

Both are penalized versions of the mean-squared residual, where the penalties are functions of the degrees of freedom used in fitting the model. For both AIC and SIC, "smaller is better." We will have much more to say about them in section **??** below.

3.3.13 Log Likelihood

The **likelihood function** is tremendously important in statistics, as it summarizes all the information contained in the data. It is simply the joint density function of the data, viewed as a function of the model parameters.

The number reported is the maximized value of the log of the likelihood function under the assumption of Gaussian disturbances.[10] Like the sum of squared residuals, SIC and AIC, it's not of much use alone, but it's useful for comparing models and testing hypotheses. We will sometimes use the maximized log likelihood function directly, although we'll often focus on the minimized sum of squared residuals.

3.4 Regression From a Forecasting Perspective

3.4.1 The Key to Everything (or at Least Many Things)

Linear least squares regression, by construction, is consistent under very general conditions for "the linear function of x_t that gives the best approximation

[10] Throughout, "log" refers to a natural (base e) logarithm.

to y_t under squared-error loss," which is the linear projection,

$$P(y_t|x_t) = x_t'\beta.$$

If the conditional expectation $E(y_t|x_t)$ is linear in x_t, then the linear projection and the conditional expectation coincide, and OLS is consistent the for conditional expectation $E(y_t|x_t)$.

Hence to forecast y_t for any given value of x_t, we can use the fitted line to find the value of y_t that corresponds to the given value of x_t. In large samples that "linear least squares forecast" of y_t will either be the conditional mean $E(y_t|x_t)$, which as we mentioned earlier in Chapter 2 is is the optimal forecast under quadratic loss, or the best linear approximation to it, $P(y_t|x_t)$.

One leading case in which the linear projection and conditional mean coincide (that is, $E(y_t|x_t)$ is linear in x_t) is joint normality. In particular, suppose that

$$\begin{pmatrix} y \\ x \end{pmatrix} \sim N\left(\mu, \Sigma\right)$$

where

$$\mu = \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix}.$$

Then it can be shown that:

$$y|x \sim \quad N\left(\mu_{y|x}\ ,\ \Sigma_{y|x}\right)$$

where

$$\mu_{y|x} \ = \ \mu_y \ + \ \Sigma_{yx}\, \Sigma_{xx}^{-1}\, (x \ - \ \mu_x)$$

$$\Sigma_{y|x} \ = \ \Sigma_{yy} \ - \ \Sigma_{yx}\, \Sigma_{xx}^{-1}\, \Sigma_{xy}$$

In what follows we'll often casually speak of linear regression as estimating the conditional expectation function. You can think of a Gaussian situation, or your can just mentally replace "conditional expectation" with "linear projection." We'll also implicitly assume quadratic loss, which is why we're interested in the conditional mean in the first place.

3.4.2 Why Take a Probabilistic Approach to Regression, as Opposed to Pure Curve Fitting?

We want conditional mean point forecasts, and the conditional mean is a probabilistic concept. We also may want to test hypotheses regarding which variables actually enter in the determination of the conditional mean. Last and not at all least, we also want to quantify the *uncertainty* associated with our forecasts – that is, we want interval and density forecasts – and doing so requires probabilistic modeling.

3.4.3 Regression For Estimating Conditional Means is Regression for Forecasting

We already introduced this, but we repeat for emphasis: In our regression model, the expected value of y_t conditional on x_t is

$$E(y_t|x_t) = x_t'\beta.$$

That is, the **regression function** is the **conditional expectation** of y_t given x_t.

This is crucial for forecasting, because the expectation of future y conditional upon available information is a particularly good forecast. In fact, under quadradic loss, it is the best possible forecast (i.e., it minimizes expected loss). The intimate connection between regression and optimal forecasts makes regression an important tool for forecasting.

3.4.4 LS and Quadratic Loss

Quadratic loss is routinely invoked for prediction, in which case the conditional mean is the optimal forecast, as mentioned above. OLS optimizes quadratic loss in estimation, and it's good to have the model estimation criterion match the predictive criterion.

3.4.5 Estimated Coefficient Signs and Sizes

The "best fit" that OLS delivers is effectively a best (in-sample) forecast. Each estimated coefficient gives the weight put on the corresponding x variable in forming the best linear in-sample forecast of y.

3.4.6 Standard Errors, t Statistics, p-values, F Statistic, Log Likelihood, etc.

These let us do formal statistical inference as to which regressors are relevant for forecasting y.

3.4.7 Fitted Values and Residuals

The fitted values are effectively in-sample forecasts:

$$\hat{y}_t = x_t'\hat{\beta},$$

$t = 1, ..., T$. The in-sample forecast is automatically unbiased if an intercept is included, because the residuals must then sum to 0 (see EPC 2).

The residuals are effectively in-sample forecast errors:

$$e_t = y_t - \hat{y}_t,$$

$t = 1, ..., T$.

Forecasters are keenly interested in studying the properties of their forecast errors. Systematic patterns in forecast errors indicate that the forecasting model is inadequate – as we will show and explore later in great depth, forecast errors from a good forecasting model must be unforecastable! And again, residuals are in-sample forecast errors.

3.4.8 Mean and Variance of Dependent Variable

An obvious benchmark forecast is the sample, or historical, mean of y, an estimate of the *unconditional* mean of y. It's obtained by regressing y on an intercept alone – no conditioning on other regressors!

The sample standard deviation of y is a measure of the (in-sample) accuracy of the unconditional mean forecast under quadratic loss.

It's natural to compare the accuracy of our conditional-mean forecasts to naive unconditional-mean forecasts. R^2 and $\bar{R}^2$, to which we now turn, do precisely that.

3.4.9 R^2 and $\bar{R}^2$

Hopefully conditional-mean forecasts that condition on regressors other than just an intercept are better than naive unconditional-mean forecasts. R^2 and $\bar{R}^2$ effectively *compare* the in-sample accuracy of conditional-mean and unconditional-mean forecasts.

$$R^2 = 1 - \frac{\frac{1}{T}\sum_{t=1}^{T} e_t^2}{\frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})^2}.$$

$$\bar{R}^2 = 1 - \frac{\frac{1}{T-K}\sum_{t=1}^{T} e_t^2}{\frac{1}{T-1}\sum_{t=1}^{T}(y_t - \bar{y})^2},$$

3.4.10 SSR (or MSE), SER (or Residual s^2), AIC and SIC

Each attempts an estimate of *out-of-sample* forecast accuracy (which is what we really care about) on the basis of in-sample forecast accuracy, with an eye toward selecting forecasting models. (That is, we'd like to select a forecasting model that will perform well for out-of-sample forecasting, quite apart from its in-sample fit.) Everything proceeds by inflating the in-sample mean-squared error (MSE), in various attempts to offset the deflation from regression fitting, to obtain a good estimate of out-of-sample mean-squared error. We have:

$$MSE = \frac{\sum_{t=1}^{T} e_t^2}{T}$$

$$s^2 = \left(\frac{T}{T-K}\right) MSE$$

$$AIC = \left(e^{\left(\frac{2K}{T}\right)}\right) MSE$$

$$SIC = \left(T^{\left(\frac{K}{T}\right)}\right) MSE.$$

We will have much more to say about AIC and SIC in section **??** below.

3.4.11 Durbin-Watson

We mentioned earlier that we're interested in examining whether there are patterns in our forecast errors, because errors from a good forecasting model should be unforecastable. The Durbin−Watson statistic tests for a particular and important such pattern, serial correlation. If the errors made by a forecasting model are serially correlated, then they are forecastable, and we could improve the forecasts by forecasting the forecast errors! We will subsequently discuss such issues at great length.

3.4.12 Residual Plots

Residual plots are useful for visually flagging neglected things that impact forecasting. Residual serial correlation indicates that point forecasts could be improved. Residual volatility clustering indicates that interval forecasts and density could be improved. (Why?) Evidence of structural change in residuals indicates that *everything* could be improved.

3.5 Exercises, Problems and Complements

1. Regression, regression diagnostics, and regression graphics in action.

   At the end of each quarter, you forecast a series y for the next quarter. You do this using a regression model that relates the current value of y to the lagged value of a single predictor x. That is, you regress

   $$y_t \to c, x_{t-1}.$$

   (In your computer workfile, y_t is called Y, and x_{t-1} is called XLAG1. So you run

   $$Y \to c, XLAG1.$$

   (a) Why might include a *lagged*, rather then current, right-hand-side variable?

   (b) Graph Y vs. XLAG1 and discuss.

   (c) Regress Y on XLAG1 and discuss (including related regression diagnostics that you deem relevant).

   (d) Consider as many variations as you deem relevant on the general theme. At a minimum, you will want to consider the following:

      i. Does it appear necessary to include an intercept in the regression?

ii. Does the functional form appear adequate? Might the relationship be nonlinear?

iii. Do the regression residuals seem completely random, and if not, do they appear serially correlated, heteroskedastic, or something else?

iv. Are there any outliers? If so, does the estimated model appear robust to their presence?

v. Do the regression disturbances appear normally distributed?

vi. How might you assess whether the estimated model is structurally stable?

2. Least-squares regression residuals have zero mean.

   Prove that least-squares regression residuals must sum to zero, and hence must have zero mean, if an intercept is included in the regression. Hence in-sample regression "forecasts" are unbiased.

3. Conditional mean and variance

   Consider the regression model,

$$y_t = \beta_1 + \beta_2 \, x_t + \beta_3 x_t^2 + \beta_4 z_t + \varepsilon_t$$

$$\varepsilon_t \overset{iid}{\sim} (0, \sigma^2).$$

   (a) Find the mean of y_t conditional upon $x_t = x_t^*$ and $z_t = z_t^*$. Does the conditional mean vary with the conditioning information (x_t^*, z_t^*)? Discuss.

   (b) Find the variance of y_t conditional upon $x_t = x_t^*$ and $z_t = z_t^*$. Does the conditional variance vary with the conditioning information (x_t^*, z_t^*)? Discuss.

4. Conditional means vs. linear projections.

   Consider a scalar y and a vector x, with joint density $f(y, x)$.

   (a) The conditional mean $E(y|x)$ is not *necessarily* linear in x. Give an example of a non-linear conditional mean function.

   (b) Consider such a non-linear conditional mean situation. You assume (incorrectly) that a linear regression model holds. You estimate the model by OLS. We say that you are estimating "**linear projection weights**." Linear projection weights are best linear approximations (under quadratic loss) to conditional expectations.

   (c) Consider again such a non-linear conditional mean situation. In large samples and under quadratic loss, what can be said about the comparative merits of conditional-mean vs. linear-projection forecasts?

   (d) What factors influence whether your prediction will actually perform well?

5. Squared residual plots for time series.

   Consider a time-series plot of *squared* residuals rather than "raw" residuals. Why might that be useful?

6. HAC Standard Errors

   Recall that OLS linear regression is consistent for the linear projection under great generality. Also recall, however, that if regression disturbances are autocorrelated and/or heteroskedastic, then OLS standard errors are biased and inconsistent. This is an issue if we want to do inference as to which predictors (x variables) are of relevance. HAC methods (short for "heteroskedasticity and autocorrelation consistent") provide a quick and sometimes-useful fix. The variance of the OLS estimator is:

$$\Sigma = (X'X)^{-1}E(X'\varepsilon\varepsilon'X)(X'X)^{-1}.$$

For *iid* ε_t this collapses to the usual $\sigma^2(X'X)^{-1}$, but otherwise we need the full formula. Write it as:

$$\Sigma = (X'X)^{-1}E(X'\Omega X)(X'X)^{-1}.$$

The question is what estimator to use for $E(X'\Omega X)$. In the case of pure heteroskedasticty (Ω diagonal but not scalar), we can use the White estimator,

$$E(\widehat{X'\Omega X}) = X'diag(e_1^2, ..., e_T^2)X = \sum_{t=1}^{T} e_t^2 x_t' x_t.$$

In the case of heteroskedasticity and autocorrelation, we can use the Newey-West estimator,

$$E(\widehat{X'\Omega X}) = \sum_{t=1}^{T} e_t^2 x_t' x_t + \sum_{l=1}^{m}\left(1 - \frac{l}{m+1}\right)\sum_{t=l+1}^{T} e_t e_{t-l}(x_t' x_{t-l} + x_{t-l}' x_t),$$

where the so-called "truncation lag" m is chosen by the user. The first Newey-West term is the White estimator, and the second Newey-West term is an additional adjustment for autocorrelation.

There is a strong case against HAC estimators in forecasting contexts: They achieve robust inference for predictor relevance, but they don't *exploit* any heteroskedasticity present (to improve interval forecasts) or serial correlation present (to improve point forecasts). Nevertheless we introduce them here because (1) they are often produced by regression software, and (2) they can be of use, as we will see, in exploratory modeling en route to arriving at a complete forecasting model.