

1 Variance Estimation

The regression assumptions give us a baseline to evaluate the adequacy of the model. Do they hold, do they not hold? If so then good, if not, then what are our alternatives? The assumptions are also useful in allowing us to connect our estimated regression function back to the population regression function, for through the assumptions, we're saying something about the nature of the data and the nature of ϵ in the population.

But we need more precision in connecting our estimates back to the population parameters. Since our estimates are from sample data, it is very likely that the estimates themselves will change from sample to sample. Because the estimates can (and probably will) vary, we need a measure of the estimate's *precision* or its reliability. That is, we want to say, how much variability is there in our estimates of the slope (and intercept) coefficients?

The usual measure of precision in statistics is the *standard error*. What *is* a standard error? It is simply the standard deviation of the sampling distribution of the estimator. What is the sampling distribution? It is simply a probability distribution of the set of values of the estimator obtained from all possible samples of the same size from a given population. Remember the fundamentals you learned in POL 582? Here they are again. We're squarely going to rely on notions of sampling distributions and standard errors in order to lead us toward statistical inference.

That you understand what a sampling distribution is, is something I'm going to assume; therefore, consult an introductory statistics text if you need a refresher. But the basic idea is simple: given that our estimator has a probability distribution (for a given sample size from a given population), it is natural to ask what the variance is *of* that distribution (again, just like basic statistics when you learned to make inferences on \bar{X} !). This leads directly to the consideration of the *variance of the estimators*.

1.1 Bivariate Case

Let's consider the simple bivariate model first (extension to the n -variable case is straightforward). The variance of the regression slope, \hat{b} is given by

$$\text{var}(\hat{b}) = \frac{\sigma^2}{\sum(X_i - \bar{X})^2},$$

and so the standard error is obtained by taking the square root of the variance, giving us

$$\text{se}(\hat{b}) = \frac{\sigma}{\sqrt{\sum(X_i - \bar{X})^2}}.$$

The variance of the regression intercept $\hat{\beta}_0$ is given by

$$\text{var}(\hat{\beta}_0) = \left(\frac{\sum X_i^2}{n \sum(X_i - \bar{X})^2} \right) \sigma^2,$$

and the standard error is given by

$$\text{se}(\hat{\beta}_0) = \sqrt{\left(\frac{\sum X_i^2}{n \sum (X_i - \bar{X})^2}\right) \sigma}.$$

In general, we will be more interested in the precision around the slope coefficient than the intercept (Why?).

Where have we seen σ^2 before? Recall the assumption of homoscedasticity. Under that assumption, it was stated that the variance of the disturbance term ϵ_i was equal to some positive constant, σ^2 . Hence, the term in the variances denoted by σ^2 is precisely this homoscedastic variance of the disturbance.

Note, however, that we usually will not directly observe this term (why?). Instead we have to estimate it directly from the data. Yet take note that apart from not directly estimating σ^2 , we *can* directly estimate the other components of the standard error (as it involves the variance in X_i). Now, if σ^2 corresponds to the variance of ϵ_i , then what would be our best estimate of this quantity? (A: the variance of the residuals, that is $\text{var}(e_i)$.)

This corresponds to what component of the regression output you obtain from Stata? Recall from class and from the book that the variance of the residuals was given by

$$\begin{aligned} \text{var}(e_i) &= \frac{\sum (Y_i - \hat{Y}_i)^2}{n - 2} \\ &= \frac{\sum (e_i)^2}{n - 2} \\ &= \frac{SSE}{n - 2}, \end{aligned}$$

which, after taking the square root, gave us

$$\begin{aligned} \text{se}(e_i) &= \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n - 2}} \\ &= \sqrt{\frac{\sum (e_i)^2}{n - 2}} \\ &= \sqrt{\frac{SSE}{n - 2}}, \end{aligned}$$

which corresponded to what? (A: The **standard error of the regression estimate**.) Recall that we referred to the variance of the residuals as the mean square error, or the MSE, and we referred to the square root of the variance—the standard error of the estimate—as the root mean square error, or the RMSE. These statistics, both generated from the variance components, are directly given to you in your Stata output. Now we see that our estimate of σ^2 is given by the MSE.

To illustrate, let's work through an example using the Davis data (I sent this to you). First, using data on women only, regress the measured weight variable on the reported weight variable. In the Stata output, what is the numerator for the standard error on the slope

coefficient? (It is the RMSE: 2.0569). What is the denominator? (It is the square root of the sum of mean deviations of the variable reported weight). To obtain this, compute the variance of X and multiply it by $n - 1$ (why?). The variance of reported weight is 45.39 and so the sum of the squared deviations is $45.39 \times 100 = 4539.307$. Taking the square root of this gives us $\sqrt{\sum(X_i - \bar{X})^2} = 67.37$. Substituting these numbers into the formula for the standard error (note that we replace σ^2 with the MSE) gives us $\frac{2.06}{67.37} = .0305$, which is of course equivalent to the standard error reported in the `regress` output. Again, the numbers do not come out of thin air, but are a logical extension of the variance components and values of the independent variable.

For the intercept, the relevant information can be directly gotten from Stata. For the numerator, we need the sum of the squared values of X . Using `egen` in Stata, I find that this value is 329,731. For the denominator, I multiply $n = 101$ by the sum of the squared mean deviations of X , that is $101 \times 4539.31 = 458,470.31$. Substituting these numbers into the formula given previously, I obtain $\sqrt{\frac{329731}{458470.31}} \times RMSE \approx 1.744$, again, equivalent to the Stata output.

It is instructive to understand what helps to improve the precision of your estimates. Anything that lowers your standard error increases the precision of your estimates. Consequently, it should be clear that increasing the variance in X will help to lower the standard errors. (An obvious extension of principles learned in research design: maximize variance on your covariates!). Note also the close connection between precision of the estimates and the standard errors. As the sample size increases, the standard error must decrease? (Why?) (A: increases in sample sizes increase variability in X ; also, increase in sample sizes help to lower the RMSE (note that you're dividing the MSE by $n - 2$ and so as n increases, the RMSE must decrease, thus lower the standard error).

1.2 Multiple Regression Case

The extension to multiple regressors is straightforward (although like the least squares estimators, the presentation in scalar form gets ugly). Suppose we have two independent variables and an intercept, the variances of the intercepts are given by

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum(X_1 - \bar{X}_1)^2(1 - r_{1,2}^2)},$$

for $\hat{\beta}_1$,

$$\text{var}(\hat{b}_2) = \frac{\sigma^2}{\sum(X_2 - \bar{X}_2)^2(1 - r_{1,2}^2)},$$

for \hat{b}_2 , and

$$\text{var}(\hat{\beta}_0) = \left[\frac{1}{n} \frac{\bar{X}_1^2 \sum(X_2 - \bar{X}_2)^2 + \bar{X}_2^2 \sum(X_1 - \bar{X}_1)^2 - 2\bar{X}_1\bar{X}_2 \sum(X_1 - \bar{X}_1)(X_2 - \bar{X}_2)}{\sum(X_1 - \bar{X}_1)^2 \sum(X_2 - \bar{X}_2)^2 - (\sum(X_1 - \bar{X}_1)(X_2 - \bar{X}_2))^2} \right] \cdot \sigma^2,$$

for $\hat{\beta}_0$. The standard errors, analogous to the bivariate case, are given by the square roots of the variance functions, which yields (for the slopes)

$$\text{se}(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum(X_1 - \bar{X}_1)^2(1 - r_{X_1, X_2}^2)}},$$

for $\hat{\beta}_1$,

$$\text{se}(\hat{b}_2) = \frac{\sigma}{\sqrt{\sum(X_2 - \bar{X}_2)^2(1 - r_{X_1, X_2}^2)}},$$

for \hat{b}_2 . The term, $1 - r_{1,2}^2$, is known as the “auxiliary regression” where the r^2 is obtained when one estimates X_1 on X_2 . Equivalently, the square root of the r^2 term gives you the correlation coefficient between X_2 and X_1 . Think about what it is: it is a measure of how collinear the covariates are. Under conditions of perfect collinearity, what happens to the standard error? Under conditions of high collinearity, what happens to the standard error? (It must increase).

All of the remarks regarding the bivariate case hold for the multiple regression setting (that is, all the stuff about increasing the variance in X and so forth). Like the bivariate case, we don’t observe σ^2 , but instead must estimate it using the *RMSE*.

Let’s work through an example. Using the Duncan data, we estimate a regression of occupational prestige on education and income. To see where the standard errors come from, we need only compute the relevant information required in the formulae given out previously. First, to derive the variances, this requires us to estimate σ^2 . This estimate is given by the *MSE* and equals 178.73. For the denominator (let’s consider $\hat{\beta}_1$ first), we need to compute the sum of the squared mean deviations for X_1 (the education variable). This is equal to the variance times $n - 1$ (why?) which is equal to $885.7071 \times 44 = 38971.11$. To derive the auxiliary regression, we could regress education on income and take the r^2 . This would give us $r^2 = .5249$. Equivalently, we could square the simple correlation coefficient and obtain the same number: $(r_{education, income})^2 = (.7245)^2 = .5249$. The auxiliary regression would be $1 - .5249 = .4751$. Now we have all the pieces to compute the variance: $\frac{178.73}{38971.11 \times .4751} = .00965$. The standard error is gotten by taking the square root of the variance, which gives us .09825, which is equivalent to the `regress` output.

The variance of \hat{b}_2 is computed similarly (the only term that changes is the variance term in the denominator), which is $\sum(X_2 - \bar{X}_2)^2 = 26,271.199$. Putting the pieces together gives me $\frac{178.73}{26271.20 \times .4751} = .0143$ for the variance and .1196 for the standard error, again, numbers equivalent to the Stata output. (I am going to omit computing the standard error for the intercept, though I encourage you to try).

Computing the standard errors are a breeze. Interpreting the standard errors is absolutely necessary in order to make inferences about the model. However, to go one step further—toward hypothesis testing—we now, finally, have to say something about the distribution of e_i .

2 The Normality Assumption

No assumptions regarding the error term have been made. The only assumptions regarding e_i are what?

1. conditional mean is 0.
2. variance is homoscedastic.
3. 0 covariance with x_i .

We can get far without saying anything about the distribution. However, if we're interested in saying something about the population parameters, then we need to go beyond point estimation and enter into the world of hypothesis testing. This absolutely requires us to say something about the distribution of the error term.

Why?

The regression coefficients are a linear function of e_i (recall the least squares estimator). Therefore, the sampling distribution of our least squares estimator will depend on the sampling distribution of ϵ . That is, we have to say something about the distribution of the error term in order to say something about the distribution of the regression parameters. Note that Gauss-Markov isn't helpful here. We need to make some assumptions.

In the classic linear model, the usual assumption that is made is that the disturbance term is **normally distributed**. Specifically, we are going to assume (using the standard notation) that

$$\begin{aligned}E(\epsilon_i) &= 0 \\E(\epsilon_i^2) &= \sigma^2 \\E(\epsilon_i, \epsilon_j) &= 0, i \neq j,\end{aligned}$$

which of course are the usual assumptions. *But in addition to this*, we're going to assume the ϵ is *normally distributed*. This leads to the following assumption (again, in notational form):

$$\epsilon_i \sim N(0, \sigma^2),$$

which says that ϵ is normally distributed with mean 0 and variance σ^2 . We can state this more forcefully by recognizing that *for any two normally distributed random variables, a zero covariance between them implies independence*. This means that if ϵ_i and ϵ_j have a 0 covariance (which they do by assumption), then they can be said to be **independently distributed**, which leads to the following statement:

$$\epsilon_i \sim \text{NID}(0, \sigma^2),$$

where NID means *normally and independently distributed*. Now why do we assume the normal? There are a wide variety of possible distributions out there ... so why the normal? The principle reason is given by the **central limit theorem** (again, another POL 582 concept emerging). What does the central limit theorem say here? If there are large number of independently and identically distributed random variables (identically distributed means they have the same probability distribution defined on them), the distribution of their sum

will tend to a normal distribution as n increases. So it is the central limit theorem that provides us with a strong justification to assume normality (again, the assumptions don't come out of thin air: statistics class is cumulative! The stuff you learned last semester leads to a result that helps us keep on the path to statistical inference).

Once we get ourselves to the normal, the problem of making inferences gets resolved. An important result of the normal distribution is that *any linear function of normally distributed random variables is itself, normally distributed*. Since the regression coefficients are linear functions of e_i , and we're assuming the sampling distribution of e_i is normal, then the sampling distributions for the regression estimates are *also normally distributed*.

Eureka! Now all the stuff you learned about hypothesis testing and confidence intervals—which was taught to you in terms of the normal distribution assumption—fall right into the lap of regression analysis.

So under conditions of the normal, what are the properties of our regression estimates (note that we're assuming the conditions leading to Gauss-Markov hold here, so we're building on them)? We know the following. First, the mean of the slope coefficient is

$$E(\hat{\beta}_1) = \beta_1;$$

second, the variance of the slope coefficient is

$$\sigma_{\hat{\beta}_1}^2 = \frac{\sigma^2}{\sum(X_i - \bar{X})^2},$$

for the bivariate case, and

$$\sigma_{\hat{\beta}_1}^2 = \frac{\sigma^2}{\sum(X_1 - \bar{X}_1)^2(1 - r_{X_1, X_2}^2)},$$

for the multiple regression setting with two independent variables (note that I am omitting the variance of $\hat{\beta}_0$, as it is trivially different from the previous equation). We can state this more compactly, then, as

$$\hat{\beta}_1 \sim N(\beta_1, \sigma_{\hat{\beta}_1}^2),$$

which reads, “ $\hat{\beta}_1$ is normally distributed with mean β_1 and variance $\sigma_{\hat{\beta}_1}^2$.” By the properties of the normal distribution, we can create a z -variable defined as

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}},$$

which follows the standardized normal distribution, that is

$$Z \sim N(0, 1).$$

For the intercept, we can proceed in the exact same way (which I won't do here) by writing out its mean and variance, and we would show that

$$\hat{\beta}_0 \sim N(\alpha, \sigma_{\hat{\beta}_0}^2),$$

and so

$$Z = \frac{\hat{\beta}_0 - \alpha}{\sigma_{\hat{\beta}_0}}.$$

In short, getting to the normal allows us to make use, at least in principle, of the normal distribution (I say in principle because we're going to see that the actual distribution we'll rely on for hypothesis testing is the t -distribution (why?)).

Also under the normal distribution, we can define a distribution for our estimator $\hat{\sigma}^2$ (this is what? MSE) as

$$\frac{(n - k - 1)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-k-1}^2,$$

where χ^2 denotes the chi-square distribution with $n - k - 1$ degrees of freedom. Use of the χ^2 statistic will allow us to compute confidence intervals around the estimator σ^2 .

Finally, it is important to point out that under the normal distribution, the regression estimates *have minimum variance in the entire class of unbiased estimators*. This takes the “L” out of BLUE, because we're extending beyond the class of linear estimators. Hence, under normality, the least squares estimators are the **best unbiased estimators**; that is, they are **BUE**. This result is due to a statistician named C.R. Rao (1965).

One last thing before we move to interval estimation and hypothesis testing. If e_i is distributed normally, then Y_i itself must be normally distributed; that is,

$$Y_i \sim N(\alpha + \beta_k X_i, \sigma^2).$$

This result stems from the fact that any linear function of a normally distributed variable is itself, normally distributed.

3 Interval Estimation: Statement Confidence Coefficients

With our assumption of normality in hand, we can now consider interval estimation. Last semester in POL 582, you learned how to compute confidence intervals around, say, the mean. The logic of interval estimation was simple: since the parameter of interest is unknown, it is natural to consider “how close” your estimate is to the population parameter. In the context of POL 582, you probably paid attention to the relationship between \bar{X} and μ (or some variation on this theme). When you did interval estimation, you assumed normality because that opened up the door for you to use the standard normal distribution to define the bounds on the confidence interval (that is, under the normal, you could easily compute the upper and lower probability for a given confidence level). To be precise, you assumed normality because of the central limit theorem and the law of large numbers. This got you to the normal, just like in regression.

In the regression model, the logic is the same. We're going to be asking how precise our estimator is by computing a confidence interval around the point estimator. Conceptually, we are going to compute an interval such that the probability the interval contains the true parameter *in repeated samples* is equal to the probability $1 - \alpha$. In this probability statement,

α is referred to as the **level of significance** while the probability value is known as the **confidence level** or the **confidence coefficient**.

Clearly, the confidence level will be determined by the significance level, that is, α such that lower levels of α give a higher probability that the interval, in repeated samples, will contain the true parameter. However, the trade-off here is that lower α levels (as we will see) create larger intervals. Higher α levels produce tighter intervals, but with lower levels of confidence. Ideally, we want small intervals and high levels of confidence.

At the start, I want you to make sure you understand how to interpret the confidence interval. Suppose we're interested in making an inference about a slope parameter, β_k . Since it is assumed that in the population β_k is a *fixed number*, it must be the case that for *any single interval that you compute, the probability that β_k lies in the interval is either 0 or 1*. The interpretation of the interval only makes sense in terms of repeated sampling.

3.1 Interval Estimation for Regression Coefficients

Under the normality condition, we can specify $Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}}$; however, a fundamental problem exists because σ is usually unknown. In its place, we estimate σ by using the standard error of $\hat{\beta}_1$, leading to the consideration of the statistic

$$t = \frac{\hat{\beta}_1 - \beta_1}{\text{s.e.}(\hat{\beta}_1)},$$

which follows the *t distribution* with $n - k - 1$ degrees of freedom. Recall from introductory statistics the close connection between the *t* distribution and the standard normal distribution. In the limit, they are identical. Also, recall in POL 581 that it was generally the case that you couldn't observe directly the standard error of μ , so you estimated it using the estimated standard error of \bar{Y} . The slope estimate, $\hat{\beta}_1$, like \bar{Y} , is a linear combination of the observations Y_i (why?); the only difference here than when compared to \bar{y} is that the degrees of freedom differ for the *t* statistic.

Now since $\frac{\hat{\beta}_1 - \beta_1}{\text{s.e.}(\hat{\beta}_1)} \sim t(n - k - 1)$, we can use the *t* distribution to establish a confidence interval in the following way:

$$\Pr(-t_{\alpha/2} \leq t \leq t_{\alpha/2}) = 1 - \alpha.$$

The term $t_{\alpha/2}$ denotes our critical value and α denotes the significance level. The level $\alpha = .05$ is common, but $.01$ or $.10$ levels are also commonly used as well.

Substituting terms for the interval, we can rewrite the previous statement as

$$\Pr(-t_{\alpha/2} \leq \frac{\hat{\beta}_1 - \beta_1}{\text{s.e.}(\hat{\beta}_1)} \leq t_{\alpha/2}) = 1 - \alpha,$$

and rearranging, gives

$$\Pr[\hat{\beta}_1 - t_{\alpha/2}\text{s.e.}(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2}\text{s.e.}(\hat{\beta}_1)]$$

which is the $100(1 - \alpha)$ **percent confidence interval**. Hence, $\alpha = .05$ yields a 95 percent confidence interval. We can display this more compactly as

$$\hat{\beta}_1 \pm t_{\alpha/2} \text{s.e.}(\hat{\beta}_1).$$

One important thing to note is the fact that we're dividing the significance level by two. Why is this the case? (A: confidence intervals give upper and lower bounds; therefore, area from both tails of the t -distribution is used. This requires us to split the probability area given by α into two halves.) Note also that the width of the c.i. is proportional to the standard error of the coefficient. We can now see why the standard error is a measure of *precision*: it directly effects the interval in which the population parameters will probabilistically reside (over repeated samples).

3.2 Interval Estimation for σ^2

In addition to computing confidence intervals for the regression estimates, it also may make sense to compute a confidence interval for the estimator, $\hat{\sigma}^2$. Recall that earlier we said that the variance estimator, under the condition of normality, could be expressed in terms of a χ^2 variable,

$$\chi^2 = \frac{(n - k - 1)\hat{\sigma}^2}{\sigma^2},$$

where

$$\frac{(n - k - 1)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-k-1}^2.$$

We can use the χ^2 distribution to establish a confidence interval around $\hat{\sigma}^2$ (recall the close connection between the χ^2 and the normal distributions):

$$\Pr(\chi_{1-\alpha/2}^2 \leq \chi^2 \leq \chi_{\alpha/2}^2) = 1 - \alpha.$$

The terms $\chi_{1-\alpha/2}^2$ and $\chi_{\alpha/2}^2$ give the two critical χ^2 values that cut off the $100(\alpha/2)$ percent areas of the χ^2 distribution. By substitution, we can rewrite the confidence interval as

$$\Pr\left[(n - k - 1)\frac{\sigma^2}{\chi_{\alpha/2}^2} \leq \sigma^2 \leq (n - k - 1)\frac{\sigma^2}{\chi_{1-\alpha/2}^2}\right] = 1 - \alpha,$$

thus giving us the $100(1 - \alpha)$ **percent confidence interval for σ^2** . To illustrate, consider the Duncan data. For these data, our estimate of σ^2 is 178.73, which is the mean square error. Our degrees of freedom equal 42. The χ^2 critical values are found by consulting a χ^2 table. Setting $\alpha = .05$, the first cut off is equal to 61.777 (i.e. $\chi_{\alpha/2}^2$) and the second cut off is equal to 25.999 (i.e. $\chi_{1-\alpha/2}^2$). Note that in Stata, the critical values can be obtained by typing:

```
display invchi(df,p),
```

where df denotes the degrees of freedom (42 in this example) and p denotes the probability value (.025 for the first cut off and .975 for the second cut off). Substituting these numbers into the formula gives us the following interval

$$121.38 \leq \sigma^2 \leq 288.73$$

(Verify for yourself these are the correct numbers). The interpretation of this interval is that *the true value of σ^2 will lie in this interval in 95 percent of all possible samples of size n (in this example $n = 45$) from the same population; that is, in the long run, over repeated sampling, we would be correct 95 percent of the time.*

4 Hypothesis Testing

Closely related to the concept of interval estimation is the concept of hypothesis testing. The basic premise of hypothesis testing is to ask (and answer) the question (or some variation of it): what is the probability my result could have occurred by random chance alone? Since our data are probabilistic and the model has a (sometimes large) stochastic component to it, the best we can do in answering this question is only relative to some prespecified probability level. With confidence intervals, we choose α and as such, if we make α very small, the confidence coefficient is very large *but* the confidence intervals also get very large (thus allowing the true parameter to reside in a wide space over repeated samples). In contrast, we can set α to be a relatively large number (.10, .15, .20) and we will obtain smaller confidence intervals; however, in the long run, the probability of an incorrect inference increases because the confidence coefficient decreases as α increases.

Classical hypothesis testing usually begins with the statement of a null hypothesis and an alternative hypothesis. Standard hypothesis testing then proceeds by evaluating the veracity or the likelihood of the null hypothesis by using observed data and statistical theory. We can consider two complementary means to test hypotheses.

4.1 Interval Estimation and Hypothesis Testing

To illustrate, we can link hypothesis testing back to interval estimation. Suppose that in the occupational prestige model (the Duncan data), we wanted to ascertain whether or not education had a statistically significant relationship to prestige. Under conditions of the null, we might want to evaluate the claim that education had *no impact* on prestige (note that the “null” is called the “null” because of the premise that there is no effect). In terms of regression analysis, this would imply that the slope coefficient for education would be 0, leading to the following null hypothesis:

$$H_0 : \beta_1 = 0.$$

The alternative to this hypothesis could then be expressed as

$$H_a : \beta_1 \neq 0.$$

Under the null, we state there exists no relationship; under the alternative, we posit that the relationship is *something other than 0*. The way the alternative is written, then, the relationship between education and prestige could be positive *or* it could be negative. Expressed in this way, we have a **two-sided hypothesis**.

To test the null hypothesis, we want to ask whether or not the estimated regression coefficient for education is compatible with the null hypothesis. Return to the confidence

interval computed earlier. We found that the 95 percent confidence interval around the education slope was bounded by .348 and .744 (or similarly, $.546 \pm .198$). The logic of the hypothesis test would be to *fail to reject the null hypothesis if the interval “covered” the condition of the null*. Or to put this another way, since we know that in the long run 95 percent of all samples will contain the true parameter in this interval, if β_1 under the null falls within this interval, then we cannot reject the null hypothesis. If β_1 under the null falls outside the interval, then we can reject the null *with 95 percent confidence*. If we rejected the null, then we would be essentially saying that the probability that $\hat{\beta}_1$ could have occurred by chance alone is 5 percent. Since this probability is small, we reject the null and accept the alternative hypothesis.

Returning to our example, it is clear that β_1 under the null is *not* contained within the 95 percent confidence interval: it is outside the interval. Thus, we would reject the null and accept the alternative hypothesis and therefore conclude that the education coefficient is, with 95 percent confidence, *significantly different from 0*.

The alternative hypothesis specified above is known as a two-tail test. This is because we specified the condition as being *either* greater than 0 *or* less than 0. In terms our hypothesis test, if $\alpha = .05$, then we would divide α by 2 (i.e. $\alpha/2$) to account for the probability area in both tails of the distribution. Sometimes one-tail tests are more appropriate, however. Under a one-tail test, we might specify the alternative hypothesis as

$$H_a : \beta_1 > 0.$$

This would require us to recover the probability area under only 1-tail of the distribution. To see the differences more clearly, let us consider what some call, the “test-of-significance” approach.

4.2 Test-of-Significance Approach

The basic idea here is similar to that just discussed. Specifically our aim is to judge the veracity of the condition of the null hypothesis using a *test statistic* and the observed data. Further, we define a probability distribution on the test statistic and use the distributional properties to probabilistically evaluate the null hypothesis. The theory is straightforward. Recall that if we assume normality, then we can specify a variable t denoted as

$$t = \frac{\hat{\beta}_1 - \beta_1}{\text{s.e.}(\hat{\beta}_1)}.$$

This variable follows the t distribution with $n - k - 1$ degrees of freedom. To test some hypothesis, we specify the value of β_1 under the condition of the null. As before, we could state the null as

$$H_0 : \beta_1 = 0,$$

but we could easily specify β_1 under the null as being equal to *any* hypothetical value (i.e. 1, .5, 3.14, etc.). Suppose we define β_1^* as the value of β_1 under the null. Then we could rewrite t as

$$t = \frac{\hat{\beta}_1 - \beta_1^*}{\text{s.e.}(\hat{\beta}_1)}.$$

where β_1^* now reflects the condition of the null (and the $t_{\alpha/2}$ denote the critical t values, as before). Since this test statistic follows the t distribution, we could establish a statement confidence interval,

$$\Pr(-t_{\alpha/2} \leq \frac{\hat{\beta}_1 - \beta_1^*}{\text{s.e.}(\hat{\beta}_1)} \leq t_{\alpha/2}) = 1 - \alpha,$$

and rearranging, gives

$$\Pr[\beta_1^* - t_{\alpha/2}\text{s.e.}(\hat{\beta}_1) \leq \hat{\beta}_1 \leq \beta_1^* + t_{\alpha/2}\text{s.e.}(\hat{\beta}_1)],$$

which looks an awful lot like the confidence interval computed for the regression coefficient earlier. The only difference here is the term at the center of the interval. This will give us the interval in which $\hat{\beta}_1$ will fall (in repeated samples) with $1 - \alpha$ probability, **given that** $\beta_1 = \beta_1^*$, that is, the condition of the null. If the estimated parameter *falls outside the interval*, then we can reject the null hypothesis. This is the case because the interval is based on the condition of the null hypothesis. If it were “true,” this interval should contain (up to a level of probability) the estimated coefficient. If the interval doesn’t contain the estimated parameter, then we can reject the null. Hence, area outside the interval is called the **region of rejection**, because we reject H_0 . The **region of acceptance** of the null is thus given by area *inside* the interval. If our estimated parameter falls within this region, we say that our estimate is *not* significantly different from the condition of the null. Hence, if the null states that $\beta_1 = 0$, then we would say that our estimated regression coefficient is *not* significantly different from 0. Our level of confidence would be given by $100(1 - \alpha)$.

This approach to hypothesis testing is clearly connected to the confidence interval approach. In the latter, we’re constructing an interval around β_1 and ascertaining the probability that over repeated samples, the true parameter would be found inside this interval; in the former, we’re hypothesizing some value of β_1 and ascertaining the probability that the estimated coefficient lies within the confidence limits around the hypothesized value of β_1 .

To illustrate, return to the Duncan data. Suppose that we hypothesize for the null that the true value of the education slope is 0. Following the test of significance approach and setting $\alpha = .05$, the confidence interval around $\hat{\beta}_1$ is given by

$$\Pr(-.198 \leq \hat{\beta}_1 \leq .198) = .95.$$

Since our estimated coefficient of .546 falls outside this interval, we reject the null hypothesis that the slope is 0 and conclude the alternative hypothesis (with 95 percent confidence).

In applied work, one needs not compute this interval, for there is a shortcut. All we need to do to test the hypothesis is compute the term in the middle of double inequality statement given by

$$\Pr(-t_{\alpha/2} \leq \frac{\hat{\beta}_1 - \beta_1^*}{\text{s.e.}(\hat{\beta}_1)} \leq t_{\alpha/2}) = 1 - \alpha.$$

That is, we can compute the t value directly and see whether it falls inside or outside the interval bounded by the critical t values under the condition of the null. To illustrate, the t value for the education slope is given by $\frac{.546-0}{.098} = 5.57$; clearly, $t = 5.57$ falls outside the interval bounded by $\pm .198$. Hence, we can reject the null accept the alternative with 95 percent confidence.

Note from this shortcut an important result: the further away the estimated coefficient is from the hypothesized value, the larger the t -value becomes. In absolute value, large t values are *evidence against the null*. The t -value reported in your Stata output is premised on the assumption that the null hypothesis is 0; this need not be the case. Suppose that we hypothesize that the null was not 0 but, say, .3. Then the t value would be given by $\frac{.546-.3}{.098} = 2.51$. The confidence interval under this version of the null would be bounded by $-.10$ and $.498$ (verify this) and so we would reject the null and accept the alternative that the true slope parameter is different from .3 (with 95 percent confidence).

Again, though, we don't need to go through and compute the confidence interval. Instead, we can utilize p values to determine the probability of a t value. In consulting a t table, we can look up the appropriate degrees of freedom and derive the probability for a given t value. Suppose we have 8 degrees of freedom and obtain a t value of 2.306. In looking at the t table, we see that the probability of obtaining a t value of 2.306 or greater is 5 percent. This means that this result could have occurred by chance alone only about 5 percent of the time. Since this probability is sufficiently low, we are 95 percent confident in our result. In our example, the probability of obtaining a t value of 5.57 with 47 degrees of freedom is extremely low (not even on the table!). In this case, we conclude that the probability of obtaining this value is *at least* as small as .001 (and certainly smaller).

Using the t table to establish a probability value in this manner is known as a t test. To be precise, a statistic is said to be statistically significant if the value of the test statistic lies in the region of rejection, discussed above (that is, outside the interval bounded by the critical t under the null). If this occurs, we reject the null with $100(1 - \alpha)$ confidence. A test is statistically insignificant, however, if the t value lies within the region of acceptance (i.e. falls within the interval). Here, we fail to reject the null.

To this point, we've only concerned ourselves with two-tail tests. Sometimes, our hypotheses will be more precise such that we specify the alternative in directional terms (greater than or less than). If we can be sufficiently precise, then a one-tail test is appropriate. Why? Because we need only look at 1 tail of the distribution. The test procedure is identical as before except the critical t value is t_α , *not* $t_{\alpha/2}$. One-tail tests are less conservative than two-tail tests. To illustrate, suppose that we obtain a t value of 1.860 on 8 degrees of freedom. Further, suppose that $\alpha = .05$. The critical t for a *two-tail* test is 2.306 (why? Because $.05/2 = .025$); however, for a *one-tail* test, the critical t value is 1.86. Thus, if we obtained a t value of 1.86, we would reject the null at the 5 percent level for a one-tailed test, but *fail* to reject the null at the 5 percent level for a two-tailed test.

This raises the question of which test is appropriate? If you can be precise in your hypotheses, then use the one-tail test. Stata, and most statistical software by default gives the p value for a two-tailed test at $\alpha = .05$. You should practice one-tail tests at different significance levels.

5 Confidence Intervals for Prediction

Commonly, we're interested in making predictions about Y , or analogously, making inferences based on predicted values generated from a regression model. However, just as we were concerned with uncertainty in our estimates of the regression coefficients, we should be

equally concerned with uncertainty in our predictions. After all, the predictions are based on estimates, which are themselves, uncertain.

5.1 Mean Prediction

To fix ideas, consider the bivariate regression model

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X.$$

The interpretation of this model is that as X changes by a unit amount, the $E(Y | X)$ changes by $\hat{\beta}$ amount. The notion of the expected value stems from the fact that we're really estimating, in the context of the regression model, the mean response of Y conditional on some value of X .

The problem is, we don't actually observe $E(Y | X)$, but instead must estimate it. Hence, just as the intercept is our best estimator of α and our slope coefficient is our best estimator of β , our \hat{Y} is our best estimator of $E(Y | X)$. This is the natural and obvious choice for the estimator of the conditional mean.

In treating this as our estimator, some interesting facts emerge. First, since the regression coefficients, under Gauss-Markov, are BLUE estimators, so to is \hat{Y} as our estimator of $E(Y | X)$. Further, under the normality assumption, since \hat{Y} is a linear function of the regression coefficients which are themselves normally distributed, then \hat{Y} is *also* normally distributed. This opens up the door for interval estimation in the exact same way the door opened up for interval estimation (and hypothesis testing) when we assumed normality on the regression estimates.

Hence, since we recognize that \hat{Y} is an estimator, we must also recognize that there exists variance around the estimate. Suppose that we define \hat{Y}_o as the predicted value of Y when $X = X_o$, that is, some specific value. What we want to know is, what is the variability around this predicted value, which is our estimator of the mean response, or mean value of Y , when $X = X_o$.

Based on the normal, we know that the *mean* of the sampling distribution of \hat{Y}_o is given by,

$$E(\hat{Y}_o) = E(Y | X_o).$$

The question is, what is the variance around \hat{Y}_o ? Omitting the proof, it turns out that the variance (in the bivariate case) is given by (if you want the proof, I can give you some cites)

$$\sigma^2(\hat{Y}_o) = \sigma^2 \left[\frac{1}{n} + \frac{(X_o - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]$$

which must be estimated by

$$var(\hat{Y}_o) = MSE \left[\frac{1}{n} + \frac{(X_o - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right].$$

(Why?) The standard error around \hat{Y}_o is then the square root of the above, hence,

$$s.e.(\hat{Y}_o) = RMSE \left[\frac{1}{n} + \frac{(X_o - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]^{\frac{1}{2}}.$$

Because we are using the *MSE* in our variance estimates, we must use the *t* distribution to construct a confidence interval for $E(Y | X_o)$. From the *t*, we can define the following:

$$t = \frac{\hat{Y}_o - E(Y | X_o)}{s.e.(\hat{Y}_o)},$$

which can then be used to construct the following confidence interval:

$$\Pr[\hat{Y}_o - t_{\alpha/2}se(\hat{Y}_o) \leq E(Y | X_o) \leq \hat{Y}_o + t_{\alpha/2}se(\hat{Y}_o)] = 1 - \alpha.$$

If $\alpha = .05$, then we obtain the $100(1 - \alpha)$ percent confidence interval for $E(Y | X_o)$. The interpretation: in repeated samples of the same size from the same population, 95 percent of all samples will give the true conditional mean value, that is $E(Y | X_o)$.

To illustrate, consider the Stata handout. In using the Duncan data, we regress prestige on education. We obtain the following model:

$$\hat{Y} = .284 + .902(\text{Education}).$$

Each of the \hat{Y} may be obtained in Stata by typing

```
predict xb,
```

where **xb** is the name of the variable I created storing the predicted values. Now suppose I was interested in estimating the mean prediction error for the case when $X = 86$. Plugging 86 into the regression model, I obtain a predicted prestige score of 77.86 (note the **list** command in the Stata output).

All the pieces for computing the variance of \hat{Y}_o are available in the data. (Note that I use the Stata **summarize** command to get me the relevant information I need for X). Putting the pieces together for the variance, I obtain

$$var(\hat{Y}_o) = 278.63 \left[\frac{1}{45} + \frac{(86 - 52.56)^2}{38971.11} \right],$$

which gives me an estimate of the variance as 14.12. Taking the square root, I obtain the standard error, which is 3.76. Now, to construct a 95 percent confidence interval, I set $\alpha = .05$ and compute the critical *t* value on 43 degrees of freedom. In Stata, I type

```
display invt(43, .95)
```

and this returns the critical *t* value as 2.017 (see bottom of page 1 on Stata handout). Substituting these numbers into the formula for the confidence interval, I obtain the following interval:

$$77.86 \pm 7.60,$$

which renders the following interpretation: in repeated samples, 95 percent of the samples will contain the true value of $E(Y | X_o)$ in the interval as given above. Another way of looking at this is that the error on the prediction interval around the mean response could

be as high as plus or minus 7.6 (for a 95 percent confidence interval). At the top of the second page of the Stata handout, I create the upper and lower 95 percent confidence limits using the `gen` command and I list these bounds for $X = 86$ using the `list` command.

It is important to note that usually, you will want to compute this mean prediction interval *for all predicted values*. Clearly, this would be time-consuming; however, it turns out, Stata will compute for you the standard error of the mean prediction if you type:

```
predict meanpred, stdp
```

where `meanpred` is the variable I created that stores the standard errors of the mean prediction. From these, you can compute the *95 percent confidence bands* for the mean prediction. On the Stata output, I do this (see middle part of page 2). I use the critical t computed earlier and the variable `xb` (which contains all the predicted values) to derive the upper and lower 95 percent confidence bands. Finally, I graph the bands using the `graph` command in Stata. The first graph in your handout corresponds to the graph I created. I draw reference lines through the points that correspond to the upper and lower confidence limits we computed for $X = 86$. We can verify that we computed them correctly, as the lines go right through the points given by 77.86 ± 7.60 .

5.2 Individual Prediction

Suppose that instead of predicting the mean response (and its error), we were interested in forecasting or predicting *an individual outcome*? To derive the error around the forecast, or individual prediction, we require a different procedure. Why? When we're dealing with the mean response, we're predicting the conditional mean—that is, the mean of the distribution of Y given X . But for the case of individual prediction, or forecasting, we are predicting an individual outcome which is itself, drawn from the conditional distribution of Y . Since individual outcomes deviate from the mean response, the prediction interval around the forecast of an individual outcome will be wider than for the mean prediction.

As an analogy, the difference here is roughly similar to forecasting the percentage of votes a Democrat will get from a particular election as compared to predicting the average number of votes Democrats get across all elections (the analogy isn't quite perfect, but close enough). Clearly, forecasting a single election will be more error prone.

For individual prediction, we are interested in Y_o , as opposed to $E(Y | X_o)$. As before, our best estimator of Y_o is \hat{Y} .

The variance of the individual prediction (given without proof) is

$$\text{var}(Y_o - \hat{Y}_o) = E[Y_o - \hat{Y}_o]^2 = \text{MSE} \left[1 + \frac{1}{n} + \frac{(X_o - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right].$$

The standard error is then the square root of the above, hence,

$$\text{RMSE} \left[1 + \frac{1}{n} + \frac{(X_o - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right]^{\frac{1}{2}}.$$

Note that this term must be larger than the mean prediction. Because we are using the *MSE* in our variance estimates, we must use the *t* distribution to construct a confidence interval for Y_o). From the *t*, we can define the following:

$$t = \frac{Y_o - \hat{Y}_o}{s.e.(Y_o - \hat{Y}_o)},$$

which can then be used to construct the following confidence interval:

$$\Pr[\hat{Y}_o - t_{\alpha/2}se(\hat{Y}_o) \leq Y_o \leq \hat{Y}_o + t_{\alpha/2}se(\hat{Y}_o)] = 1 - \alpha.$$

In the Stata handout, I go through an example following the same steps I followed to derive the mean prediction. Like the mean prediction error, it is usually best to compute individual prediction errors for the full data set. This will give you another set of confidence bands. So you know, Stata will compute the standard error of the forecast for you by typing

```
predict ipred, stdf
```

where `ipred` corresponds to the variable I created storing the standard error of the forecast (denoted by Stata as `stdf`).

Finally, I graph the forecast bands and that corresponds to the second graph in your handout.