Econ 452 Section 6 - STATA

Connor Cole

October 21, 2015

Regression, Continued

Regression Inference

Having gone over the basics of interpreting regression output, let's be more specific about regression inference. First, it's worth noting that regression inference is a further step with regression analysis which is, in some senses, dissociable from other useful properties of regression estimation. Regression fits a model that minimizes the residual sum of squares, and, assuming that the errors are *iid* and homoskedastic and as long as $\mathbb{E}[\epsilon|X] = 0$, these estimates $\hat{\beta}$ are the best linear unbiased (BLUE) estimates of the coefficients on the model we are estimating. Remember that 'best' in the acronym BLUE refers to the fact that these estimates are efficient, or have the lowest possible variance among the class of linear estimators, and 'linear estimators' refers to estimators of our regression coefficients β that can be expressed as a linear function of the dependent variable y. Even if error terms are not homoskedastic, it's still true that our regression estimates will be unbiased and consistent estimators of the true value of β as long as $\mathbb{E}[\epsilon_i|X] = 0$. However, just because an estimator of β is unbiased and consistent doesn't mean that the coefficients we observe with our given data, $\hat{\beta}$, are 'close' to the true values of β. Unbiasedness implies that if we had repeated draws of data that we would, on average, get the 'true' values of β, and consistency implies that as we increase the sample size we would approach the 'true' value of β in probability limit. We don't know, however, whether or not we are 'close' to certain potential values of β for our particular realization of data without inference. We might be especially worried that the true value of a coefficient on some variable is in fact 0, but we observe some non-zero coefficient because of sampling variation in the particular sample at which we are looking.

Regression inference opens up a whole new set of tools that we can use to test hypotheses about the values on certain coefficients, and construct confidence intervals for various parameters that provide useful information about what the values of β might be beyond our simple point estimate of $\hat{\beta}$. Wooldridge motivates regression inference by claiming that we have assumed $\epsilon \sim N(0, \sigma^2)$. We will relax this assumption later when we learn more about the asymptotic properties of regression (or, what happens to our estimator as we send the number of observations to infinity) and many of the basic results we will discuss will continue to hold.

Let's return to our regression output example and interpret it further.

Figure 1: Regression Output

```
. regress VariableName1 VariableName2
```

Source	SS	df	MS	Number of obs	=	5,516
				F(1, 5514)	=	755.27
Model	4.9528e+11	1	4.9528e+11	Prob > F	=	0.0000
Residual	3.6159e+12	5,514	655770269	R-squared	=	0.1205
				Adj R-squared	=	0.1203
Total	4.1112e+12	5,515	745457975	Root MSE	=	25608

VariableName1	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
VariableName2	.3223284	.0117286	27.48	0.000	.2993356	.3453211
_cons	12779.37	638.8523	20.00	0.000	11526.97	14031.78

We previously discussed what each of the individual components of this table are, now let's focus on the table at the bottom describing estimated coefficients on variables. We see that regressing $VariableName1$ on $VariableName2$ results in an estimated coefficient on $VariableName2$ of around 0.3223. This value is our value of $\hat{\beta}_1$ in estimating the equation $VariableName1 = \beta_0 + \beta_1 VariableName2$. Since we are in a single variate regression with a constant, this estimated coefficient is:

$$\hat{\beta}_{\text{VariableName2}} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Next, we see a standard error term of 0.0117. Since we are in a single variable linear regression, this value will be:

$$se(\hat{\beta}_{\text{VariableName2}}) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Remember that $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u_i}^2}{n-k-1}$.

Now, let's turn to thinking about these results in the context of hypothesis testing. We first need to specify a null hypothesis and an alternative hypothesis. STATA reports results in this table under the t and $P > |t|$ columns that come from individually testing for each coefficient the null hypothesis that the coefficient on the variable is 0, and the alternative hypothesis that the coefficients on these variables are non-zero. We will look at hypothesis tests regarding the values on multiple coefficients simultaneously later, but for the time being we will only consider these

individual null hypotheses about coefficients on variables. Note that specifying the alternative hypothesis in this way implies a two-sided hypothesis test. Furthermore, remember that, assuming the errors to be normally distributed, then our test statistic T has the following exact statistical distribution:

$$T = \frac{\hat{\beta}_{\text{VariableName2}} - \beta^{True}_{\text{VariableName2}}}{se(\hat{\beta}_{\text{VariableName2}})} \sim t_{n-k-1}$$

Or, a t distribution with $n - k - 1$ degrees of freedom. So, to do hypothesis testing, we would apply our null hypothesis to our test statistic, and calculate the probability of observing an outcome as large as T or greater assuming the null to be true. If T falls into some predetermined rejection region given by a certain level of size α (generally set to be 0.05 by convention), we would reject our null hypothesis at the α level of confidence and report the probability of observing a test statistic as great as T or greater as the p-value from our hypothesis test.

Let's apply our step-by-step hypothesis testing procedure and rederive the results that STATA reports in the regression table above regarding the coefficient on $VariableName2$.

1. **Define H_0.**
   STATA in the test above implicitly tests $H_0 : \beta_{VariableName2} = 0$.

2. **Define H_1.**
   STATA in the test above implicitly tests $H_1 : \beta_{VariableName2} \neq 0$.

3. **Define α = Probability of Type I Error.**
   While STATA does not set a level of α, by convention most researchers set $\alpha = 0.05$.

4. **Define the test statistic and test.**
   STATA applies a t-test, using the fact that $T = \frac{\hat{\beta}_{\text{VariableName2}} - \beta^{Null}_{\text{VariableName2}}}{se(\hat{\beta}_{\text{VariableName2}})} \sim t_{n-k-1}$ if the null hypothesis were true.

5. **Define a rejection region given the test.**
   We are in a two-sided t-test and α has been set as 0.05, and the degrees of freedom are $5516 - 1 - 1 = 5514$. We need to find some value of C such that $P(|T| \geq C|H_0) = \alpha$, or the value C such that the probability of observing some absolute value of a test statistic $|T|$ greater than C assuming the null is true is α. To find this value, we type in:

   ```
   display invttail(5514, .025)
   ```

   This command reports the value of C such that the probability of observing a random variable T from the t distribution with 5514 degrees of freedom that is greater than C is .025, that is: $P(T > C) = 0.025$. Then, by the symmetry of the t distribution, we can claim that C is the value such that the probability of observing a random variable T from the t distribution with 5514 degrees of freedom that in *absolute value* is greater than C is 0.05, or $P(|T| > C) = 0.05$. This value of C for this example is 1.9603, a value that is very close to what we would find in

the normal distribution, 1.9599. The closeness of these two numbers is a reminder that the t distribution approaches the normal distribution as we increase sample size.

Therefore, our rejection region is:

$$RR = (-\infty, -1.9603) \cup (1.9603, \infty)$$

6. **Calculate and report the test statistic.** STATA reports that our test statistic is $T = 27.48$. We can derive this by hand as follows:

$$\begin{aligned} T &= \frac{\hat{\beta}_{\text{VariableName2}} - \beta^{Null}_{\text{VariableName2}}}{se(\hat{\beta}_{\text{VariableName2}})} \\ &= \frac{0.322328 - 0}{0.01172} \\ &= 27.48 \end{aligned}$$

7. **Do inference.** STATA does not report conclusions about inference because inference depends on the size α that we as STATA users assume. Under our choice of $\alpha = 0.05$, we see that the test statistic is very far into the rejection region. Therefore, we reject the null hypothesis at the 0.05 confidence level.

8. **Report p-values**. The p-value is $P(|X| > |T||H_0)$, or the probability of observing an outcome as great or greater than the test statistic we observe given that the null is true. Note here that the p-value is two-sided because we have done a two-sided hypothesis test. STATA reports this probability under the column titled `P>|t|`. We see there that the p-value for this test listed in the table is essentially 0. We could derive this result by hand by typing in:

```
display 2*ttail(5514, 27.48)
```

Where the `ttail` function reports the probability of observing a value at or above 27.48 in a t distribution with 5514 degrees of freedom, and we multiply it by 2 using the symmetry of the t-distribution to account for the fact that we want two-sided p-values.

We have covered every piece of output and filled in the hypothesis testing implicit in the regression table above except for the final columns, which report a 95% confidence interval. 95% confidence intervals are often reported simultaneously with results from hypothesis testing, but the 95% confidence interval uses slightly different approach to inference. In classical hypothesis testing, we test a specific hypothesis about the value of certain coefficients. With 95% confidence intervals, we observe an outcome from a procedure that doesn't require a specific null and alternative hypothesis and provides a general indication of the potential value of the true value of β.

Remember that for the 'true' value of $\beta_{\text{VariableName2}}$ our previous test statistic has a t distribution with 5514 degrees of freedom. Therefore, using the critical values we established earlier such that the probability of observing a value from this t distribution outside of a certain boundary is 0.05, we know that:

$$P(-1.9603 < \frac{\hat{\beta}_{\text{VariableName2}} - \beta_{\text{VariableName2}}}{se(\hat{\beta})} < 1.9603) = 0.95$$

By rearranging, we have:

$$\begin{aligned}
0.95 &= P(-1.9603 < \frac{\hat{\beta}_{\text{VariableName2}} - \beta_{\text{VariableName2}}}{se(\hat{\beta})} < 1.9603) \\
&= P(-1.9603 \cdot se(\hat{\beta}) < \hat{\beta}_{\text{VariableName2}} - \beta_{\text{VariableName2}} < 1.9603 \cdot se(\hat{\beta})) \\
&= P(\hat{\beta}_{\text{VariableName2}} - 1.9603 \cdot se(\hat{\beta}) < \beta_{\text{VariableName2}} < \hat{\beta}_{\text{VariableName2}} + 1.9603 \cdot se(\hat{\beta}))
\end{aligned}$$

Therefore, we can expect the 'true' value $\beta_{\text{VariableName2}}$ to fall in this interval *before we compute* $\hat{\beta}$ 95% of the time. After we compute the confidence interval, of course, either the 'true' value of $\beta_{\text{VariableName2}}$ is or isn't in the interval we've computed. Some students get mixed on this point of interpretation and believe that the probability of the 'true' $\beta_{\text{VariableName2}}$ falling in the confidence interval *after* we compute $\hat{\beta}$ is 95%. It is best to read confidence intervals as the outcome of a procedure that gives us a rough set of bounds for the 'true' value of a parameter.

STATA reports that the confidence interval for $\hat{\beta}_{\text{VariableName2}}$ is (0.2993, 0.3453). We can derive this finding by hand by using the formula established above:

$$\begin{aligned}
95\ \%\ \text{CI} &= (\hat{\beta}_{\text{VariableName2}} - 1.9603 \cdot se(\hat{\beta}), \hat{\beta}_{\text{VariableName2}} + 1.9603 \cdot se(\hat{\beta})) \\
&= (0.322328 - 1.9603 \cdot 0.01172, 0.322328 + 1.9603 \cdot 0.01172) \\
&= (0.2993, 0.3453)
\end{aligned}$$

Testing Other Null Hypotheses on Regression Coefficient Values

As described previously, STATA's default regression output reports results from hypothesis testing on coefficient values assuming default H_0 that the value of a coefficient in question is 0 and assumes a default H_1 that the value on a coefficient is not 0. If we want to test other null hypotheses about coefficient values, then we have all the tools to do it by hand if we'd like to, or we can use the `lincom` command. Let's say that I wanted to test the null hypothesis that the coefficient on $VariableName2$ is 1 with an alternative hypothesis that the coefficient on $VariableName2$ is not 1. Then, we would type after the relevant regression:

```
lincom VariableName2-1
```

From the regression above, that would give us the following output in figure 2:

Figure 2: `Lincom` Output

```
. lincom VariableName2-1

 ( 1)  VariableName2 = 1
```

VariableNa~1	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	-.6776716	.0117286	-57.78	0.000	-.7006644	-.6546789

Note that the standard error is the same as before, but the 'coefficient' has changed, along with the test statistic t and the 95% confidence interval. This result makes sense, as all we have done is change the null hypothesis. Again, let's apply our step-by-step hypothesis testing procedure and rederive the results that STATA reports from this particular hypothesis test.

1. **Define H_0.**
   We have explicitly stated the null hypothesis $H_0 : \beta_{VariableName2} = 1$.

2. **Define H_1.**
   STATA assumes that our alternative hypothesis is $H_1 : \beta_{VariableName2} \neq 1$.

3. **Define α = Probability of Type I Error.**
   Again, by convention, we set $\alpha = 0.05$.

4. **Define the test statistic and test.**
   Just as before, STATA applies a t-test, using the fact that $T = \frac{\hat{\beta}_{\text{VariableName2}} - \beta^{Null}_{\text{VariableName2}}}{se(\hat{\beta}_{\text{VariableName2}})} \sim t_{n-k-1}$ if the null hypothesis were true.

5. **Define a rejection region given the test.**
   Again, as previous, we are in a two-sided t-test and α has been set as 0.05, and the degrees of freedom are $5516 - 1 - 1 = 5514$. We need to find some value of C such that $P(|T| \geq C|H_0) = \alpha$, or the value C such that the probability of observing some absolute value of a test statistic $|T|$ greater than C assuming the null is true is α. By the same calculation as previous, the critical value C is 1.9599.

   Therefore, our rejection region is:

$$RR = (-\infty, -1.9603) \cup (1.9603, \infty)$$

6. **Calculate and report the test statistic.** STATA reports two figures of note. First, STATA reports a 'coefficient' of -0.6776. We can derive this number by subtracting off the observed coefficient on $VariableName2$ by 1: $0.322328 - 1 = -.6776$. Then, STATA reports a test statistic of -57.78. Again, we can derive this value by hand:

$$
\begin{aligned}
T &= \frac{\hat{\beta}_{\text{VariableName2}} - \beta^{Null}_{\text{VariableName2}}}{se(\hat{\beta}_{\text{VariableName2}})} \\
&= \frac{0.322328 - 1}{0.01172} \\
&= -57.78
\end{aligned}
$$

7. **Do inference.** As previous, under our choice of $\alpha = 0.05$, we see that the test statistic is very far into the rejection region. Therefore, we reject the null hypothesis at the 0.05 confidence level.

8. **Report p-values**. STATA reports this probability under the column titled `P>|t|`. We see there that the p-value for this test is essentially 0. We could derive this result by hand by typing in:

   ```
   display 2*ttail(5514, 57.78)
   ```

   Where the `ttail` function reports the probability of observing a value at or above 57.78 in a t distribution with 5514 degrees of freedom, and we multiply it by 2 using the symmetry of the t-distribution to account for the fact that we want two-sided p-values.