Introductory Applied Econometrics
EEP/IAS 118 Steven Buck
Spring 2015 Lecture #13

Statistical Inference with Regression Analysis

Next we turn to calculating confidence intervals and hypothesis testing of a regression coefficient ($\hat{\beta}$). Fortunately, $\hat{\beta}$ is a random variable similar to $\bar{y}$. Just like the estimated $\bar{y}$s, the estimated $\hat{\beta}$s have a distribution with some mean, $\bar{\hat{\beta}}$, and variance, $\sigma^2_{\hat{\beta}}$. Provided that these estimated $\hat{\beta}$s are Normally distributed, then we can use the same method for calculating confidence intervals and steps for hypothesis testing as we did when drawing sample means to make statistical inferences about the population mean. In order to maintain this proviso, we just need to impose a couple of assumptions in addition to MLR1-MLR5 (Gauss-Markov Assumptions). First, let's review the G-M Assumptions.

Gauss-Markov Assumptions Review:

1. What assumptions do we need for our $\hat{\beta}$ estimators to be unbiased, i.e. $\mathbb{E}\left[\hat{\beta}_j\right] = \beta_j$?
   - MLR.1: **Linearity in parameters**.
   - MLR.2: **Random sampling**.
   - MLR.3: **No perfect multicollinearity**.
   - MLR.4: **Zero conditional mean** Satisfying this assumption can be difficult and violation of it is often the cause of omitted variable bias. One way to think about this assumption is that there should be nothing in the error term that is correlated with both our explanatory variable (x) of interest and our outcome variable (y). If there is a factor (z) in the error term that is correlated with both x and y then the zero conditional mean assumption will not be satisfied and our estimator $\hat{\beta}_j$ will be biased.
2. What assumptions do we need to have that $Var\left(\hat{\beta}_j\right) = \frac{\sigma^2}{SST_j\left(1-R_j^2\right)}$? We are interested in this because it provides a notion for the precision of the OLS estimator over repeated sampling.
   (a) MLR.1-4 and MLR.5: MLR.1-MLR.4 as before and MLR5 which is **homoskedasticity**. The assumption of homoskedasticity tells us that for each value of x the error terms have the same variance. The assumption is important because it allows us to calculate analytic standard errors without much fuss (analytic just means there exists a formula for the calculation). Why do we care about being able to calculate a variance for $\hat{\beta}_j$? Well, without it then we wouldn't be able to proceed to statistical inference. Why? Because in order to calculate confidence intervals and to perform hypothesis testing, we need estimates of the standard error of $\hat{\beta}_j$, which is just the square root of $\widehat{var(\hat{\beta}_j)}$.
   (b) **Note:** In my view, the assumption of homoskedasticity is one of convenience and based on the historical fact that we didn't have a ton of computing power so it was nice to have a simple assumption that permitted us to write down the standard error of $\hat{\beta}_j$ with an explicit equation. In practice, we are often suspect of this assumption and we'll spend some time in the remainder of the course on how to proceed when homoskedasticity isn't likely satisfied.
3. MLR.1-MLR.5 are collectively referred to as the Gauss-Markov assumptions.

MLR.6 The Normality Assumption

In order to construct confidence intervals and conduct hypothesis testing about $\hat{\beta}$ we need to know the full distribution of $\hat{\beta}$, not just its mean and variance. To understand why, let's draw another picture.

Life would be a lot easier for us if we could assume that the $\hat{\beta}$s were Normally distributed. Why? Because if a RV is Normally distributed and we know the mean and variance, then we know the full distribution of the RV (remember, we write $X \sim N(\mu, \sigma^2)$).

In order to have $\hat{\beta} \sim N(\beta, var(\hat{\beta}))$ we require:

- MLR.1-MLR.5
- MLR.6: (Normality) The population error is independent of the explanatory variables $x_1, x_2, ..., x_k$ and is Normally distributed with zero mean and variance σ_u^2: $u \sim Normal(0, \sigma_u^2)$.

Under MLR.1-MLR.6 (aka, the Classical Linear Model assumptions), we know that (conditional on the sample values of the independent variables):

$$\hat{\beta} \sim N(\beta, var(\hat{\beta}))$$

Therefore, we can standardized our estimated $\hat{\beta}$s so that:

$$\frac{\hat{\beta} - \beta}{\sigma_{\hat{\beta}}} \sim N(0, 1)$$

Of course, we don't actually know $var(\hat{\beta})$ or $s.e.(\hat{\beta})$, so then $\frac{\hat{\beta}-\beta}{SE(\hat{\beta})} \sim t_{n-k-1}$.

→ What is k in the degrees of freedom formula for a regression coefficient?

Why this works?

- **Basic Idea:** Without going through all the details once can write: $\hat{\beta} = \beta + \frac{1}{s_x^2} \sum ((x_i - \bar{x}) \cdot u_i)$. Thus, $\hat{\beta}$ can be written as a constant plus a linear combination of the u_i's, where under MLR.6 these errors are independent and identically distributed $N(0, \sigma^2)$. An important fact is that a linear combination of Normally distributed random variables is also a normally distributed random variable (Appendix B of Wooldridge).
- In fact, in small samples this doesn't always work.
- However, in large samples this approximately holds thanks to the Central Limit Theorem.
- Don't get too hung up on this.

Hypothesis tests for one parameter: Guide to Handout 12b

Thanks the Normality assumption we can apply the same formula and steps that we used for sample means to find confidence intervals and test hypotheses for regression parameters, with a couple notable changes.

Earlier we [hopefully] convinced ourselves that under the Normality assumption we have:

$$\frac{\hat{\beta}-\beta}{\sigma_{\hat{\beta}}} \sim N(0,\,1)$$

However, when we are using a sample [& particularly, the sample variance of $\hat{\beta}$], then the above expression no longer holds. Instead, we have:

$$\frac{\hat{\beta}-\beta}{\hat{\sigma}_{\hat{\beta}}} \sim t_{n-k-1}$$

- Note the degrees of freedom adjustment: k is the number of explanatory variables in the regression equation.
- $\hat{\sigma}_{\hat{\beta}} = s.e.(\hat{\beta}_j)=\frac{\hat{\sigma}_u}{\sqrt{SST_j\cdot(1-R_j^2)}}$

Example 1. The wage equation Returning to our data from the 2006 Current population survey let us consider the Stata regression output in Example 1 of Handout #12b. I estimated the following model:

$$lwage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 female + \beta_4 nonwhite + u$$

Here are the results:

```
. reg lwage educ exper female nonwhite

      Source |       SS       df       MS              Number of obs =    2000
-------------+------------------------------           F(  4,  1995) =  186.11
       Model |   182.76916     4  45.6922899           Prob > F      =  0.0000
    Residual |  489.807708  1995  .245517648           R-squared     =  0.2717
-------------+------------------------------           Adj R-squared =  0.2703
       Total |  672.576867  1999  .336456662           Root MSE      =   .4955

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1166907   .0053149    21.96   0.000     .1062673    .1271141
       exper |   .0108903   .0008687    12.54   0.000     .0091867    .0125939
      female |  -.2533517   .0222186   -11.40   0.000    -.2969257   -.2097776
    nonwhite |  -.0374722   .0311433    -1.20   0.229     -.098549    .0236047
       _cons |   1.062023   .0758905    13.99   0.000     .9131902    1.210856
------------------------------------------------------------------------------
```

Given the Stata output we approximate that an additional year of experience increase wages by roughly 1%, which seems small. Let's formally test the hypothesis that exper has no positive effect lwage.

Step 1: Define null and alternative

$$\begin{aligned} H_0: \quad & \beta_{exper} = 0 \\ H_1: \quad & \beta_{exper} > 0 \end{aligned}$$

Under the null hypothesis, the true population parameter $\beta_{exper} = 0$, which, in words, says: Years experience has no effect on lwages once controlling for educ, female and nonwhite.

When this null hypothesis holds, we know that $t = \frac{\hat{\beta}_{exper} - 0}{SE(\hat{\beta}_{exper})} \sim t_{n-k-1}$.

Step 2: Compute the test statistic. We can plug in the values that Stata gives us into the formula above for the t-stat.

$$t = \frac{\hat{\beta}_{exper} - \beta_{exper}}{SE\left(\hat{\beta}_{exper}\right)} = \frac{.0109 - 0}{.00087} = 12.53$$

Step 3: Choose the significance level and the critical value of the test. Find c:

1. Significance level is $\alpha = .01$
2. Two sided test or One sided test
3. Degrees of freedom: n-k-1 = 2000 - 4 -1 = 1995, which is close to ∞
4. Check the t-table to see that $c_{1\%}$=

Step 4: Reject/Faill to reject. We reject the null hypothesis or fail to reject the null hypothesis.

Step 5: Interpret in a "reader friendly" way.

At the 1% significance level, we reject the null hypothesis that the experience has no positive impact on lwage controlling for education level, sex and non-white race-ethnicity. ~~We find statistical evidence that experience has a positive effect on wages.~~

To be perfectly clear with you all, the striked out line above is not technically wrong; in fact, the Wooldridge textbook uses similar language. The concern raised in lecture is that we want to be very careful that *evidence of a positive effect* is not misconstrued as an *unambiguous claim of a positive [causal] effect* of experience on wages. It is easy for consumers of empirical results (e.g. policy-makers, bloggers, lawyers, etc.) to confuse these two. To avoid the possibility of such confusion to the audience of readers, we are strongly encouraging language such as the following:

"Under MLR 1-6 assumptions, we find statistical evidence that β_{exper} is positive after controlling for educ, female and non-white race-ethnicity. The validity of our statistical inference rests on the validity of our assumptions."

By writing our statement in terms of β_{exper}, we obviate the issue of explicitly stating what interpretation we attach to our estimator of it. Of course, providing an interpretation of $\hat{\beta}_{exper}$ (e.g. whether we should attach a causal interpretation to our estimate of $\hat{\beta}_{exper}$) is an important component of an empirical analysis. However, there is no harm in distinguishing these tasks in a written summary.

Thus, we might add to the above bold-faced language with:

"The interpretation of $\hat{\beta_{exper}}$ and whether is it should be considered an estimator of a causal effect is an important one that can be addressed as a separate issue from hypothesis testing and is, therefore, left for a separate discussion."

This whole discussion relates to one of the first lectures in which I claimed that it is always safe to attach a descriptive interpretation to the regression coefficients, but moving beyond that to attach a causal interpretation gets us into murky water pretty quick.

We'll have some additional discussion and material (e.g. DA 9 and solution) to share on this topic as it is an important one.

Is this positive effect economically significant?

What is the p-value?

- To avoid arbitrariness of choosing a significance level report the p-value for a test.
- The p-value is the most stringent α level significance test for which we would still fail to reject H_0.
- The p-value is the probability that we would find a $\hat{\beta}$ as far or farther from the hypothesized value if μ_0 were true.
- The p-value reported in Stata output corresponds tot he p-value fo the test: $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$. Not necessarily relevant to your hypothesis test so be careful!

Example 2. Influence of school size on test score

Step 1.

Step 2.

Step 3.

Step 4.

Step 5.

Notes:

Example 3. Pollution by paper mills

Step 1.

Step 2.

Step 3.

Step 4.

Step 5.

Notes: