NYU Department of Politics - Quant I
Fall 2013 - Prof. Patrick Egan

16 Lecture 16

16.1 Confounds, revisited

[Go over "identify the potential confound."]

16.2 Omitted variable bias

- We've just explored several different ways that one might go about controlling for a variable. We are about to go into detail on one of the simplest (and perhaps least satisfying) way to do this: including an additive term with the potential confound, Z, in the linear model.
- Although this technique may seem overly simple, it can still provide us with unbiased estimates of the *ceteris paribus* relationship between X and Y if certain assumptions hold. To see this, let's first analyze what happens when we don't control for Z:
- Assume that the true model is

$$y = \beta_0 + \beta_1 x + \beta_2 z + u$$

- Notice that we're making a big assumption here about z: no interaction between x and z, and z enters into the DGP in a linear fashion.
- Because this model is properly specified, ν is an error term that does not covary with either x or z conditional on the other variable: i.e., $cov\,(u|x,z) = cov\,(u|z,x) = 0$.
- But let's say instead we regress y only on x, falsely assuming that the model is

$$y = \beta_0 + \beta_1 x + \nu,$$

- and thus incorrectly assuming that $cov\,(\nu, x) = 0$.

- then what we are really doing is moving $\beta_2 z$ to the error term, ν :

$$\begin{aligned} y &= \beta_0 + \beta_1 x + (\beta_2 z + u), \\ \text{where } \nu &= (\beta_2 z + u). \end{aligned}$$

- You'll recall that in the bivariate case that our estimator is

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{cov\,(x, y)}{var\,(x)}$$

(here writing ν instead of u):

$$\widehat{\beta}_1 = \beta_1 + \frac{\sum (x_i - \overline{x})\, \nu_i}{SST_x},$$

- Here rely on the assumption that the covariance of x and ν is zero to make the final term dissappear, and thus say that $E\left(\widehat{\beta}_1\right) = \beta_1$. But now consider

$$\widehat{\beta}_1 = \beta_1 + \frac{\sum (x_i - \overline{x})\, (\beta_2 z_i + u_i)}{SST_x}.$$

- Taking expectations, we now have

$$\begin{aligned} E\left(\widehat{\beta}_1\right) &= E\,(\beta_1) + E\left[\frac{\sum (x_i - \overline{x})\, (\beta_2 z_i + u_i)}{SST_x}\right] \\ &= \beta_1 + \frac{E\,(\sum x_i \beta_2 z_i + x_i u_i - \overline{x} \beta_2 z_i - \overline{x} u_i)}{SST_x} \end{aligned}$$

- Now we do two things. We (1) assume the z's are fixed (just as we do the x's in the bivariate case) and (2) we invoke the (correct) assumption that $E\,(u|x, z) = 0$. Now we can write:

$$\begin{aligned} &= \beta_1 + \frac{\sum x_i \beta_2 z_i - \overline{x} \beta_2 z_i}{SST_x} \text{or more helpfully,} \\ E\left(\widehat{\beta}_1\right) &= \beta_1 + \beta_2 \left[\frac{\sum z_i\, (x_i - \overline{x})}{SST_x}\right]. \end{aligned}$$

- With a little manipulation, we see that

$$\frac{\sum z_i \left(x_i - \overline{x}\right)}{SST_x} = \frac{\sum z_i x_i - \sum z_i \overline{x}}{\sum \left(x_i - \overline{x}\right)^2} = \frac{\sum z_i x_i - n\overline{zx}}{\sum \left(x_i - \overline{x}\right)^2} = \frac{S_{xz}}{S_{xx}} = \frac{cov(x,z)}{var(x)}$$

- And so it turns out that $z_i \frac{\sum (x_i - \overline{x})}{SST_x} = \frac{cov(x,z)}{var(x)}$, which is the slope coefficient we would obtain if we simply regressed z on x! So quite simply, we can write:

$$E\left(\widehat{\beta}_1\right) = \beta_1 + \beta_2 \frac{cov(x,z)}{var(x)},$$

- and thus

$$BIAS\left(\widehat{\beta}_1\right) = E\left(\widehat{\beta}_1\right) - \beta_1 = \beta_2 \frac{cov(x,z)}{var(x)}.$$

- What if we wanted to say something about the sign of the bias? Well, note that $sign\left[\frac{cov(x,z)}{var(x)}\right] = sign\left[cov(x,z)\right]$. So if we omit z from our equation, we can now say that sign of $\widehat{\beta}_1$'s bias is

$$sign\left[cov(x,z) \times \beta_2\right]$$

- What does this mean in practice? Consider a regression in which you model feelings toward Barack Obama as a function of Democratic Party identification. You omit a dummy variable for whether an individual is African-American. In what direction is your estimate of β_1 almost assuredly biased?

- That is, you assume the model is

$$\begin{aligned} \text{ObamaFT} &= \beta_0 + \beta_1 (\text{DEM}) + \nu, \text{ when the true model is} \\ \text{ObamaFT} &= \beta_0 + \beta_1 (\text{DEM}) + \beta_2 (\text{BLACK}) + u. \end{aligned}$$

- Well, we're pretty sure that $\beta_2 > 0$ and $cov\left(\text{DEM, BLACK}\right) > 0$.

- So our estimate of β_1 will have a bias greater than zero. A.k.a., it is "biased upward,": we will overestimate the effect of Democratic Party identification because we are not accounting

for African-American racial identity.

- What happens if $cov(x, z) = 0$? What happens if $\beta_2 = 0$?
  - That's right: as we've said before, when a variable is omitted, TWO problems must be present in order for it to cause bias:
    1. it is correlated with one or more x's in your model.
    2. its partial effect on y is not zero.
  - Why, then, do we love randomly assigning individuals to x? Because by construction, $cov(x, z)$ (for any omitted z you can think of) is zero, making $\widehat{\beta}_1$unbiased.
- This is a nice simple example, but it gets more complicated in a multivariate context. You'll see this shortly.
  - [If the class asks: that's because the term $\beta_2 \left[\frac{\sum z_i(x_i - \overline{x})}{SST_x}\right]$ becomes $\beta_2 \left[\frac{1}{N}(X'X)^{-1}(X'z)\right]$, which takes into account the extent to which the omitted variable (z) is collinear with all the included x's in the model. In practice, the sign of this bias is hard to consider in such a back-of-the-envelope fashion.]
- Take-home point: if you leave out a variable that is BOTH correlated with included x's and has a separate effect on y, your estimates will suffer from omitted variable bias.
- If this omitted variable enters into the true DGP in an additive linear fashion, we can obtain unbiased estimates of β_1 and β_2–that is, the *ceteris paribus* relationships of y and x, and y and z, respectively–by moving to multiple regression. But to do that, we need a little matrix algebra.

16.3 Revisiting matrix algebra

- Here, go over:
  - Matrix algebra handout I, pp. 1-3;
  - Handout IV (entire)

17 Lecture 17

17.1 The sampling distribution of $\widehat{\beta}$

- Be sure to talk about the intepretation of a t-statistic:
  - Hypothesized mean of zero;
  - two-tailed tests
  - asterisks with p-values.
- There may be times when we want to know something different than whether some β is equal to zero; perhaps $\beta_1 = 1$ in the model [NEXT YEAR, NEW EXAMPLE. THIS ONE DOESN'T QUITE WORK; or elaborate.

$$FT\_othergroup = \beta_0 + \beta_1 (FT\_owngroup) + \mathbf{Z}\boldsymbol{\beta} + u,$$

where $\mathbf{Z}$ is a matrix of covariates. Here,

$$\begin{aligned} H_0 \quad &: \quad \beta = 1; H_A : \beta \neq 1. \\ \text{test statistic is } t \quad &= \quad \frac{\widehat{\beta} - 1}{\widehat{se\left(\widehat{\beta}\right)}}. \end{aligned}$$

- Note that you can find $\widehat{se\left(\widehat{\beta}\right)}$ by looking at the appropriate diagonal entry of the variance-covariance matrix of the vector $\widehat{\beta}$. It contains $\widehat{var\left(\widehat{\beta}\right)}$, and so $\widehat{se\left(\widehat{\beta}\right)}$ is the square root of this entry.

17.2 Interpreting an OLS regression equation

- Consider the estimated equation

$$\widehat{income} = -17,431 + 2,708(educyears) + 1,050(income16)$$

- Here, the estimates $\widehat{\beta}_1 = 2,708$ and $\widehat{\beta}_2 = 1,050$ have partial effect, or ceteris paribus, inter-

pretations. Note that

$$\frac{\partial \widehat{income}}{\partial educyears} = 2,708 \text{ and } \frac{\partial \widehat{income}}{\partial income16} = 1,050$$

- Thus we can write a statement like, "Holding family income at age 16 constant, each additional year of education is associated with an addition $2,700 in income." Other ways to say this:
  - Holding family income *fixed*
  - Education has a ceteris paribus association of $2,708 in income for each additional year of education
- Note that we can also use the equation to generate predictions about y for different values of x.

17.3 Goodness of Fit

- Just as in the simple regression case,

$$R^2 = \frac{\sum (\widehat{y}_i - \overline{y})^2}{\sum (y_i - \overline{y})^2} = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

- R^2 never decreases–and usually increases–whenever we add additional x's to the model by construction. This is because the denominator doesn't change, but the numerator usually increases with the addition of x's.
- For this reason, when comparing the goodness of fit statistics across models, it is better to compare their adjusted R-squared, which is calculated

$$R^2_{adj} = 1 - \left[\frac{\frac{SSR}{(n-k-1)}}{\frac{SST}{(n-1)}} \right] = 1 - \frac{\frac{\widehat{\sigma}^2_u}{(n-k-1)}}{\frac{SST}{(n-1)}}$$

- By construction, R^2_{adj} increases with the introduction of a new regressor into a model if and only if the t-statistic on the new variable's coefficient is greater than one in absolute value.

- Simple algebra gives
$$R^2_{adj} = 1 - \frac{\left(1 - R^2\right)(n-1)}{(n-k-1)}.$$

- When are R^2 and R^2_{adj} likely to be different? When are they likely to be similar?

17.4 Too many variables

- [for this and next subsection, draw familiar diagram of X Y Z on the board; now include W [here] and remove correlation between X and Z [below]

- we call a variable **irrelevant** if it has no partial effect on y in the population. that is, a variable W is irrelevant if:
$$\frac{\partial y}{\partial W} = 0$$

- the inclusion of an irrelevant variable in your model (aka "overspecifying the model") has no effect on the unbiasedness of the estimates of any of the betas, as it does not violate any of the assumptions 1 through 4.

- so if the true model is
$$y = \beta_0 + \beta_1 X + u$$
but you estimate
$$y = \beta_0 + \beta_1 X + \beta_2 W + u,$$
the estimates generated of all the betas (including β_2) will be unbiased: $E(\widehat{\beta}_0) = \beta_0$, etc.

- however, including an irrelvant variable in the model is harmful to the extent that this variable is collinear with other X's in the model. Recall that
$$Var\left(\widehat{\beta}_j\right) = \frac{\sigma^2}{n \cdot var(x_j) \cdot \left(1 - R^2_j\right)},$$
where R^2_j is the R-squared obtained from regressing x_j on all other independent variables in the model. Well, to the extent that the irrelevant variable W covaries with x's included in your model, the variances of the estimated coefficients associated with those x's will become

inflated.

- The result is that your estimates become less efficient: that is, their statistical power decreases, which increases the likelihood of falsely accepting the null that $\beta_j = 0$.
- To assess the threat of multicollinearity, you can generate the R_j^2 yourself for each variable x_j.
- One way to think about the size of this threat is a quantity known as the variance inflation factor (VIF) associated with each of the x's in your model. It is calculated as

$$VIF\left(\widehat{\beta}_j\right) = \frac{1}{1 - R_j^2},$$

  VIF ranges between unity (when $R_j^2 = 0$) and approaches infinity as R_j^2 approaches 1.
- Now we can re-write $Var\left(\widehat{\beta}_j\right)$ as

$$Var\left(\widehat{\beta}_j\right) = \frac{\sigma^2}{n \cdot var(x_j)} VIF\left(\widehat{\beta}_j\right).$$

- And so VIF_j is the factor by which $Var\left(\widehat{\beta}_j\right)$ is increased due to the fact that x_j is correlated with the other x's in the model.

- When displaying results, it is not always the case that we should take variables that have zero effect on y out of a model. Sometimes we include a variable x_j that we know to have zero effect in order to show our readers that we have controlled for it, and thus we are certain that the estimates on the x's we care about are not confounded by x_j.

17.5 Correlated with y, but not with x

- Recalling that

$$Var\left(\widehat{\beta}_j\right) = \frac{\sigma^2}{n \cdot var(x_j) \cdot \left(1 - R_j^2\right)},$$

  note that we will often want to include covariates Z that predict y in a model even if we don't think they are confounds with the x's we care about. That is, it can often be wise to

include predictors Z where

$$\frac{\partial x}{\partial Z} = 0 \text{ but } \frac{\partial y}{\partial Z} \neq 0$$

- Why? To the extent that they help explain y, they improve the model's predictive power and thus lower σ^2, making all of our estimates more efficient, even estimates on predictors uncorrelated with these covariates.
  - Country/state/regional fixed effects–which we often include to control for potential confounders–can also be a good example of this.

18 Lecture 18

18.1 BLUE

- Go over handout.
- Furthermore, if we add the (recall, troubling) assumption that the population errors u are distributed Normal, then $\widehat{\beta}$ is not only the BLUE of β, it is also the **minimum variance unbiased estimator (MVUE)** of β. That is, no other unbiased estimator of β exists–whether linear or not–that has a lower variance (i.e., is more efficient).

18.2 Interpreting Categorical Dummy Variables

- Go over handout.

18.3 Interaction terms

- When the effect of one variable x1 changes the effect of another x2 on y, we say that an interaction effect exists between x1 and x2.
- We model an interaction effect by creating a new variable that is the product of x1 and x2 and including it in our equation.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 \cdot x_2) + u$$

- Note that

$$\frac{\partial y}{\partial x_1} = \beta_1 + \beta_3 x_2$$

So that the slope of the line describing the relationship between each x and y varies by the value of the other x. Consider the cases where we estimate:

$$y = 1 + 3x_1 + 2x_2 + 4\left(x_1 \cdot x_2\right) + u$$

- Here's what y looks like at three different values of x2 (2, 5, and 10):
- In this case, increasing values of x2 *amplify* the effect of x1 on y, increasing the magnitude of the slope in the signed direction.
- Now consider

$$y = 1 + 3x_1 + 2x_2 - 4\left(x_1 \cdot x_2\right) + u$$

- Here's what y looks like at three different values of x2 (2, 5, and 10):
- In this case, increasing values of x2 *dampen* the effect of x1 on y, increasing the magnitude of the slope in the signed direction.
- To get a sense of interaction effects, you generally need to plot predicted probabilities of y by holding the value of one of the x's constant while varying the other value. You should label each plot accordingly.
- Statistical software programs can't tell the difference between constituitive terms and interaction terms, and so they blindly spit back incorrect standard errors. We have calculated

$$\frac{\partial y}{\partial x_1} = \widehat{\beta}_1 + \widehat{\beta}_3 x_2$$

and so we are interested in

$$\widehat{\sigma}_{\frac{\partial y}{\partial x_1}} = \sqrt{var\left(\widehat{\beta}_1 + \widehat{\beta}_3 x_2\right)} = \sqrt{var\left(\widehat{\beta}_1\right) + x_2^2 var\left(\widehat{\beta}_3\right) + 2x_2 cov\left(\widehat{\beta}_1, \widehat{\beta}_3\right)}$$

- Why? Recall that $var\left(x + ay\right) = var(x) + a^2 var(y) + 2acov(x, y)$.

- Note how the variance is non constant across values of x_2.
- Not easily calculated, but is definitely available from Stata (go over handout).
- What do we do with $\widehat{\sigma}_{\frac{\partial y}{\partial x_1}}$? We typically use it to see at what values of x_2 the variable x_1 has a non-zero ceteris paribus effect on y. That is, since $T = \frac{\widehat{\beta}_1 + \widehat{\beta}_3 x_2 - 0}{\sqrt{var\left(\widehat{\beta}_1 + \widehat{\beta}_3 x_2\right)}}$ is distributed t with $N - K - 1$ degrees of freedom, we can run the usual significance tests across the entire range of x_2. The values of $x_2$for which T surpasses the significance threshold is where x_1 has a significant effect on y. This may be all, some or only a range of the values of x_2.

18.4 Partialling out

- Here's another way to calculate the OLS estimator of, say, β_1 :

$$\widehat{\beta}_1 = \frac{\sum \widehat{u}_{i1} y_i}{\sum \left(\widehat{u}_{i1}\right)^2} = \frac{cov\left(\widehat{u}_{i1}, y_i\right)}{var\left(\widehat{u}_{i1}\right)},$$

  where $\widehat{u}_{i1}$ are the residuals from a regression of x_1 on all the other x's in the model–that is the variation in x_1 that is not explained by a linear combination of the other x's.
- So one way to think about this estimate is that its numerator is the proportion of this unexplained variation in x_1 that covaries with y.

18.5 Hypotheses about Parameters

- So far, we have focused on hypothesis tests about one parameter (a $\widehat{\beta}_j$) at a time. But there are instances in which you want to test hypotheses involving more than one parameter. Your book has the example where researchers are interested whether the effect on income of an additional year of education at a junior college is as much as the effect of an additional year of education at four-year university. The idea here is that jc's are lower status in the U.S. than universities, so maybe employers value these years of education less. (A complementary hypothesis would be that a jc education may be of lower quality.) The model assumed is

$$\log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 work + u,$$

where

$$\begin{aligned} jc &= \text{\# years attending a junior college} \\ univ &= \text{\# years attending a university} \\ work &= \text{\# months in the workforce.} \end{aligned}$$

- If we are interested in whether there is a difference in return to education from junior colleges and universities, what is an appopriate null hypothesis? It's

$$H_0 : \beta_1 = \beta_2.$$

- And an appropriate alternative is

$$H_1 : \beta_1 < \beta_2.$$

- Is a one-sided test appropriate here? Yes: theory justifies this hypothesis.
- So in the case of, say whether two groups have different means, what kind of tests did we run? (Encourage class to come up with them.)
- Rewrite null and alternative as

$$\begin{aligned} H_0 &: \beta_1 - \beta_2 = 0 \\ H_1 &: \beta_1 - \beta_2 < 0 \end{aligned}$$

- We are interested in hypotheses about the quantity $\beta_1 - \beta_2$. The statistic

$$\frac{\widehat{\beta}_1 - \widehat{\beta}_2}{se(\widehat{\beta}_1 - \widehat{\beta}_2)}$$

is distributed t.

- But how do we get $se(\widehat{\beta}_1 - \widehat{\beta}_2)$? Well,

$$\begin{aligned} se(\widehat{\beta}_1 - \widehat{\beta}_2) &= \sqrt{var(\widehat{\beta}_1 - \widehat{\beta}_2)}, \text{ and} \\ var(\widehat{\beta}_1 - \widehat{\beta}_2) &= var\left(\widehat{\beta}_1\right) + var\left(\widehat{\beta}_2\right) - 2cov\left(\widehat{\beta}_1, \widehat{\beta}_3\right). \text{ So} \\ se(\widehat{\beta}_1 - \widehat{\beta}_2) &= \sqrt{var\left(\widehat{\beta}_1\right) + var\left(\widehat{\beta}_2\right) - 2cov\left(\widehat{\beta}_1, \widehat{\beta}_3\right)} \end{aligned}$$

- Here we reject $H_0 : \beta_1 = \beta_2$ if $\widehat{\beta}_1 - \widehat{\beta}_2 + t_{crit}\left[se(\widehat{\beta}_1 - \widehat{\beta}_2)\right] < 0$.

- We can pull these from the variance-covariance matrix of the estimated betas as we did when estimating the standard errors associated with interaction effects (an example in a minute).

- But there is a much easier way to do this, as described on page 142 of your text. We care about $\beta_1 - \beta_2$, so let's call this a parameter, $\theta = \beta_1 - \beta_2$, and thus $\beta_1 = \theta + \beta_2$. Now

$$\begin{aligned} \log(wage) &= \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 work + u \text{ becomes} \\ &= \beta_0 + (\theta + \beta_2)\, jc + \beta_2 univ + \beta_3 work + u \\ &= \beta_0 + \theta jc + \beta_2\,(univ + jc) + \beta_3 work + u. \end{aligned}$$

- So if we create a new variable, $univ + jc$, and run the regression

```
reg lnwage jc univplusjc work
```

- the coefficient on jc will be the parameter we care about, and its standard error will be exactly that calculated by Stata.

- Go over example from handout.

18.6 Multiple Linear Restrictions

- The tests we've described so far are about what we call single *restrictions.* That is, we are testing whether the data justify rejecting a hypothesized restriction that $\beta_k = 0$ (in the single parameter case) or β_k is equal to, greater than, or less than some other β_j.

- But there are times when one will wish to conduct tests with multiple linear restrictions, as well. The most common such test is whether a group of variables has no effect on the dependent variable, y.

- Write the *unrestricted* model as

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u \qquad \text{(UR)}$$

- The number of variables we decide to restrict as equal to zero is q, and for convenience we assume that the restricted variables are included in the model after the unrestricted variables, then we can state the null as

$$H_0 : \beta_{k-q+1} = \beta_{k-q+2} = ... = \beta_k = 0$$

- Thus the *restricted* model is

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_{k-q} x_{k-q} + u \qquad \text{(R)}$$

- If we define the F-statistic as

$$F \equiv \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)},$$

- it has an F-distribution with q "numerator d.f." and $n-k-1$ "denominator d.f."

- Note that because $SSR_r \geq SSR_{ur}$ (since ur has more variables than r), F is always non-negative. And what we're really testing here is whether the explanatory power of ur is significantly greater than r. Under the null,

$$F \tilde{} F_{q,n-k-1}$$

- The F-distribution is the ratio of two independent chi-square random variables, divided by their respective degrees of freedom. (Recall that a chi-square is the sum of the squares of independent standard normal RVs, which is what the SSRs are.)

- Like any distribution, we can use the F's density to determine the likelihood that we'd get the F-statistic we see due to chance variation in our data. We reject the null if it's extremely unlikely that we'd obtain the F-statistic by chance.

- If H_0 is rejected, we say that the excluded variables are **jointly significant**. If the null is not rejected, then we say that they are **jointly insignificant**. It's also common to report the *p*-value associated with an F-test.

- A tidbit: the *F* statistic obtained by testing the exclusion of a single variable is equal to the square of the the *t*-statistic obtained on its coefficient via OLS:

$$F_{\text{exclude } \widehat{\beta}_k} = \left(\frac{\widehat{\beta}_k}{\widehat{\sigma}_{\widehat{\beta}_k}} \right)^2$$

- Note that Stata's regression output includes a statistic it calls F, along with "Prob >F." These are the (hardly ever used) F-statistic and p-value associated with the test that all coefficients associated with the variables you've included in your regression are zero.

18.7 Quadratics

- When the relationship is curvilinear between a Y and an X, we create the square of X, and include it in our model:

- (see "handout on polynomials")

19 Lecture 19

19.1 Heteroskedasticity and What to Do About It

- As discussed earlier in the course, we assume homoskedasticity of the errors across all observations in order to vastly simplify our calculation of $Var\left(\widehat{\beta}_j\right)$. By assumiming that $\sigma_i^2 = \sigma^2$

for all i, we can then write

$$\begin{aligned} Var\left(\widehat{\beta}_j\right) &= \frac{\sigma^2}{n \cdot var(x_j) \cdot \left(1 - R_j^2\right)} \\ \widehat{Var\left(\widehat{\beta}_j\right)} &= \frac{\widehat{\sigma}^2}{n \cdot var(x_j) \cdot \left(1 - R_j^2\right)} \end{aligned}$$

- We also needed the homoskedasticity assumption in order for the Gauss-Markov theorem to hold that OLS is the best linear unbiased estimator of the parameters of a linear population model.

- What to do? There are two approaches:

  - Heteroskedasticity of unknown form (the safe, but ignorant and often inefficient approach)
  - Modeling heteroskedasticity (requires more assumptions, but if assumptions are correct the efficient approach)

19.1.1 Heteroskedasticity of unknown form: use robust ("White") standard errors

- In the simple bivariate case, we of course write the model

$$y_i = \beta_0 + \beta_1 x_{1i} + u_i$$

- The presence of heteroskedasticity means that we can no longer write

$$VAR(u_i|x_i) = \sigma^2.$$

- We of course need to write instead

$$VAR(u_i|x_i) = \sigma_i^2,$$

because the value of σ^2 now depends on the value of x_i.

- Recall that in our final step of deriving the OLS estimator in scalar form we write

$$\widehat{\beta}_1 = \beta_1 + \frac{\sum (x_i - \overline{x})\, u_i}{\sum (x_i - \overline{x})^2}.$$

- So now consider

$$\begin{aligned} VAR\left(\widehat{\beta}_1\right) &= VAR\left[\frac{\sum (x_i - \overline{x})\, u_i}{\sum (x_i - \overline{x})^2}\right] \\ &= \left[\frac{1}{SST_x}\right]^2 \sum (x_i - \overline{x})^2 \, VAR(u_i|x_i) \\ &= \frac{\sum (x_i - \overline{x})^2 \sigma_i^2}{SST_x^2}. \end{aligned}$$

- What to do? Well, in 1980 (in the most cited economics paper in the past 35 years), Halbert White showed that a valid estimator for $VAR\left(\widehat{\beta}_1\right)$ in the presence of heteroskedasticity (if the other Gauss-Markov assumptions hold) is

$$\widehat{VAR\left(\widehat{\beta}_1\right)} = \frac{\sum (x_i - \overline{x})^2 \, \widehat{u}_i^2}{SST_x^2},$$

  where $\widehat{u}_i^2$ is simply the squared residual associated with each observation i.

- A similar formula holds in the multiple regression model, where we write

$$\widehat{VAR\left(\widehat{\beta}_1\right)} = \frac{\sum \widehat{r}_{ij}^2 \widehat{u}_i^2}{SSR_j^2},$$

  where

  - $\widehat{r}_{ij}$ is the residual obtained for observation i when regressing x_j on all the other x's, and
  - SSR_j^2 is the sum of squared residuals from this regression.
  - Note the similarities to the formula in the bivariate case.

- I hope it is obvious then that the estimated standard error of $\widehat{\beta}_1$ is

$$\sqrt{\frac{\sum \widehat{r}_{ij}^2 \widehat{u}_i^2}{SSR_j^2}}.$$

- These standard errors have lots of different names:
  - "White standard errors"
  - "Huber-White standard errors"
  - "Robust standard errors" (because the se's are "robust" in the presence of heteroskedasticity)
  - "Heteroskedasticity-robust standard errors"
  - These all mean the same thing.
- It is often–but not always–the case that robust standard errors are larger than OLS standard errors.

19.1.2 Testing for heteroskedasticity

- We can blithely report robust standard errors to be sure that our hypothesis tests are correct in the presence of heteroskedasticity.
- But, remember that if heteroskedasticity is present, OLS is no longer the best linear unbiased estimator. As we will see, you can obtain a better estimator when the form of heteroskedasticity is known.
- We are interested in tests that detect error variance that depends on the value of x. We start with the linear model

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u.$$

- Now let's specify the null hypothesis

$$H_0 : VAR(u|x_1, x_2, ...x_k) = \sigma^2$$

- Under the zero condition mean assumption this is equivalent to

$$H_0 : E(u^2|x_1, x_2, ...x_k) = E\left(u^2\right) = \sigma^2.$$

- So how do we test whether u^2 is related to the x's? How about assuming a linear function

$$u^2 = \delta_0 + \delta_1 x_1 + ... + \delta_k x_k + \nu$$

(Why am I using delta's instead of beta's here?)

- The null hypothesis now becomes

$$H_0 : \delta_0 = \delta_1 = ...\delta_k = 0.$$

- We of course do not have u^2 - these are population values that we never see. But we have estimates of u^2–our squared residuals, the $\widehat{u}^2$. So if we estimate the equation

$$\widehat{u}^2 = \delta_0 + \delta_1 x_1 + ... + \delta_k x_k + error$$

we now have a test to see the extent to which the errors in the population are related to one or more of the x's.

- One approach would be to see if any of the delta's are statistically significant. But what might be a better way?

- Look at the F-statistic from this regression, which tells us whether the x's are *jointly* significant in explaining the squared residuals. (For once, the dumb F-stat provided by typical OLS output is helpful here!) You'll recall that we defined the F-statistic as

$$F \equiv \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

- Noting that $SST = SSE + SSR$ and $R^2 = \frac{SSE}{SST}$ and $1 - R^2 = \frac{SSR}{SST}$, we can write $SSR =$

$SST\left(1-R^2\right)$. Now rewrite F as

$$\begin{aligned} F &\equiv \frac{\left[SST_r\left(1-R_r^2\right)-SST_{ur}\left(1-R_{ur}^2\right)\right]/q}{\left[SST_{ur}\left(1-R_{ur}^2\right)\right]/(n-k-1)} \\ &= \frac{\left[\left(1-R_r^2\right)-\left(1-R_{ur}^2\right)\right]/q}{\left(1-R_{ur}^2\right)/(n-k-1)} \text{ [since } SST_r = SST_{ur}\text{]} \\ &= \frac{\left(R_{ur}^2-R_r^2\right)/q}{\left(1-R_{ur}^2\right)/(n-k-1)}. \end{aligned}$$

- In the case where we are testing the joint significance of all the coefficients in a model, the restricted equation is

$$y = \beta_0 + u.$$

- Note that this equation explains none of the variation in y, as there is nothing on the right-hand side that varies except u. Thus $R_r^2 = 0$, and so the F-statistic in this case (and in any case where the test is that all the variables in the model are jointly insignificant) is equal to

$$F = \frac{\left(R_{ur}^2\right)/k}{\left(1-R_{ur}^2\right)/(n-k-1)}.$$

- Note that in this case $q = k$, as we have restricted the values of each of the k x's to be zero.

- This statistic as approximately a $F_{k,n-k-1}$ distribution under H_0, and so the p-value associated with this F-statistic is the probability that we could have obtained the coefficients we see by chance if there were no heteroskedasticity. So where $p < .05$ (or as Stata puts it, "Prob $>$ F" is less than .05, we reject H_0 at the .05 level and decide that heteroskedasticity is present.

- There are lots of other tests for heteroskedasticity. They all follow the same general pattern but with more complexity. Read about them if you like on pages. 271-276 of your text.

19.1.3 Modeling Heteroskedasticity

- We don't have time to cover the ways heteroskedasticity is modeled and corrected for using what is called generalized least squares. (Neal is likely to pick this topic up in Week 1 or Week 2 of Quant II.) The preview is this:

- Model the heteroskedasticity using versions of the linear model we used above.
- Determine the extent to which the errors change with each observation.
- Instead of running OLS, which counts each observation the same when it minimizes the sum of squared residuals...
- ...run weighted least squares, which *downweights* those observations with a higher error variance when minimizing the sum of squared residuals.
- If you have modeled the heteroskedasticity correctly, you now have estimators that are BLUE.
- If you haven't modeled it correctly, you have biased estimates of the βs.

19.1.4 What to do

- Generally you want to be able to say that your results are robust to the threat of heteroskedasticity.
- By presenting robust standard errors, you can assure your reader that the statistical significance of a particular $\widehat{\beta}_k$ is not due to an improperly estimated $VAR\left(\widehat{\beta}_k\right)$.
- Notice that because robust se's are (generally) larger than OLS se's, you're taking the safe route.
- HOWEVER, what if your paper relies on the idea that β_k is zero–a failure to reject the null? Then you'll probably want to venture into modeling heteroskedasticity, because proper modeling yields results that are more efficient–that is, less likely to get a false negative result.
- This is complicated. You'll learn more in Quant II.

19.2 Transformations of Variables

- Go over "Tranforming Nonlinearity" from Fox.