Econ 452 Section 13 - STATA

Connor Cole

December 4, 2015

Linear Probability Models

So far, we have considered models where the outcome variable is arguably a (somewhat) continuous measure, for example income. However, often we consider outcomes that are not continuous - for example whether or not an individual is a high school graduate, or some Likert scale measure of job satisfaction. For this class, we will consider one tool that we can use for situations where a dependent variable is an indicator variable taking on values 0 and 1. It is worth noting that there are other tools available that make more structural assumptions about the relationship of the dependent variable and the independent variables (e.g. probit, logit) and other tools that look at categorical variables that take on more than two values (e.g. multinomial logit, ordered probit, multinomial probit, etc.).

Suppose that we had data on whether or not an individual chooses to go to college, where the *attendcollege* variable takes on value 1 when an individual attends and 0 otherwise. Now consider estimating the following equation:

$$attendcollege = \beta_0 + \beta_1 x_1 .... + \beta_k x_k + \epsilon$$

Obviously, the dependent variable can only take on two values: 0 and 1. Note that:

$$\mathbb{E}[attendcollege|X] = \beta_0 + \beta_1 x_1 .... + \beta_k x_k$$

Note that since *attendcollege* takes on either values 0 or 1, the expectation of that variable is by definition the probability that *attendcollege* takes on value 1. Therefore, the predicted values of *attendcollege* can re reconceptualized as the 'predicted' probability of the dependent variable taking on value 1.

We can estimate this model using the usual regression tools assuming, as always that $\mathbb{E}[\epsilon|X] = 0$.

```
regress attendcollege x1 x2.....  xk
```

As was discussed in class, for this model to make sense, the predicted probabilities should be between 0 and 1. Whether or not the predicted values fall in this interval is a useful test of the model.

Also, as we discussed in class, a linear probability model by construction results in heteroskedastic errors. Suppose that conditional on X_1, we find that $\mathbb{E}[attendcollege|X] = .3$. Therefore, the probability that $attendcollege = 1$ is .3, and hence the residuals take on either value 0.7 with probability 0.3 or value -0.3 with probability 0.7. Note then that:

$$\begin{aligned}\sigma^2_{\epsilon,X_1} &= \mathbb{E}[\epsilon^2|X_1] - \mathbb{E}[\epsilon|X_1]^2 \\ &= 0.7^2 \cdot 0.3 + (-0.3)^2 \cdot .7 - (.3 \cdot .7 - .3 \cdot .7) \\ &= .21\end{aligned}$$

Now, suppose that conditional on X_2, we find that $\mathbb{E}[attendcollege|X] = .1$. Therefore, the probability that $attendcollege = 1$ is .1, and hence the residuals take on either value 0.9 with probability 0.1 or value -0.1 with probability 0.9. Note then that:

$$\begin{aligned}\sigma^2_{\epsilon,X_1} &= \mathbb{E}[\epsilon^2|X_1] - \mathbb{E}[\epsilon|X_1]^2 \\ &= 0.9^2 \cdot 0.1 + (-0.1)^2 \cdot 0.9 - (.1 \cdot .9 - .1 \cdot .9) \\ &= .09\end{aligned}$$

Thus $\sigma^2_{\epsilon,X_1} \neq \sigma^2_{\epsilon,X_2}$.

Heteroskedasticity

We have generally assumed that the variance of the error term is constant across all observations, or that $V(\epsilon|X) = \sigma^2$. If this assumption is true, then Ordinary Least Squares (OLS) is the best linear unbiased estimator (BLUE) of β, and our usual estimates of standard errors are consistent. If the variance of the error term varies across observations as a function of X, or if $V(\epsilon|X) = f(X)$, then OLS is no longer BLUE and our usual non-robust standard errors are no longer consistent estimators of the standard error. However, note that OLS estimates of $\hat{\beta}$ are still unbiased and consistent as long as $\mathbb{E}[\epsilon|X] = 0$. The only elements of our estimation process that go awry are the standard errors and hence the inference.

Testing for Heteroskedasticity

Heteroskedasticity implies that there is some relationship between the variance of the error terms for observations and the values of covariates X. One way to test for heteroskedasticity would be to regress some measure of the variance of the error terms on covariates X. If the covariates matter in hypothesis testing, then that would be a sign that the error term's variance depends on X. As an estimate of the variance, consider taking the residuals from the regression and squaring them. Since $\sigma_\epsilon^2 = \mathbb{E}[\epsilon^2|X] - \mathbb{E}[\epsilon|X]^2$, and since $\mathbb{E}[\epsilon|X]$ by assumption, then $\sigma_\epsilon^2 = \mathbb{E}[\epsilon^2|X]$, and if σ_ϵ^2 depends on X, then an estimate of the variance of the error term would be:

$$\hat{\sigma}_\epsilon^2(X) = \hat{\mathbb{E}}[\epsilon^2|X] = \hat{u}^2$$

Therefore, we can just use the residuals squared as an estimate of the variance of the error term. We covered three basic tests of heteroskedasticity: the Breusch-Pagan test, the White Test and a third 'unnamed test'. All three of these tests work similarly by regressing this estimate of the variance of the error term on some function of the covariates X and using hypothesis testing to measure whether or not these measures related to X seem to matter in our estimate of the variance.

Let's say we've estimated a regression of $VariableName1$ on $VariableName2$ and $VariableName3$.

The Breusch-Pagan test works by pulling out the residuals from the regression, squaring them, and then regressing squared residuals on all the covariates from our regression and testing the joint null hypothesis that all the coefficients have value 0. We can implement this test long-hand by typing:

```
regress VariableName1 VariableName2 VariableName3

predict residuals, resid

generate residualssq = residuals^2

regress residualssq VariableName2 VariableName3

test (VariableName2=0) (VariableName3=0)
```

We can implement the Breusch-Pagan test using hard coded STATA routines as:

```
estat hettest, fstat rhs
```

The White test works similarly by pulling out the residuals from the regression, squaring them, and then regressing squared residuals on all the covariates from our regression as well as their squares and their linear interactions and testing the joint null hypothesis that all the coefficients have value 0. We can implement this test long-hand by typing:

```
regress VariableName1 VariableName2 VariableName3
```

```
predict residuals, resid

generate residualssq = residuals^2

generate VariableName2sq = VariableName2^2

generate VariableName3sq = VariableName3^2

generate VariableNameinteract = VariableName2*VariableName3

regress residualssq VariableName2 VariableName2sq VariableName3 VariableName3sq
VariableNameinteract

test (VariableName2=0) (VariableName3=0) (VariableName3=0) (VariableName3sq=0)
(VariableNameinteract=0)
```

STATA's version of the White test is different than this procedure, as STATA's version of the White test uses the R^2 measure from the last regression times N as the test statistic, resulting in a different distribution of test statistics and hence a different p-value. You may implement this alternate procedure using hard coded STATA routines as:

```
estat imtest, white
```

Lastly, consider the 'unnamed' test that we discussed in class. Again, we regress squared residuals on some measure of the X variables to test whether or not these variables seem to matter. Here, unlike before, we use the fitted values of the dependent variable as our covariates. We can implement this test long-hand by typing:

```
regress VariableName1 VariableName2 VariableName3

predict residuals, resid

generate residualssq = residuals^2

predict fittedvalue, xb

generate fittedvaluesq=fittedvalue^2

regress residualssq fittedvalue fittedvaluesq

test (fittedvalue=0) (fittedvaluesq=0)
```

Responses to Heteroskedasticity

We have three general responses to heteroskedasticity: an altered standard error estimation process in the normal OLS estimation framework (e.g. Huber-White 'robust' standard errors), reweighting

the data by a naive estimate of the variance (weighted least squares), and estimating a structural relationship between the variance of the error term and the X values (feasible generalized least squares). We will cover the third method next week.

Response One: 'Robust' Standard Errors

Since OLS is still unbiased and consistent , one natural response to heteroskedasticity is to simply alter the standard error estimation process to make it consistent. Then, even though OLS is no longer the lowest variance method of estimating β, both our estimates $\hat{\beta}$ and our standard errors will be consistent.

There are actually more than a few estimates of the standard errors that are still consistent under hetereoskedasticity, but we will only consider one: the Huber-White standard error. Let's say we're regressing $VariableName1$ on $VariableName2$ and $VariableName3$ and we want to use Huber-White 'robust' standard errors.

```
regress VariableName1 VariableName2 VariableName3, vce(robust)
```

Changing the standard error estimation process does not change the estimated coefficients $\hat{\beta}$, but it does change the standard errors STATA reports for us. Note that these heteroskedasticity robust standard errors are consistent under *any* form of error structure as long as errors are indpendent. Hence, if we use this option even if errors are in fact homoskedastic, our estimates of the standard error will continue to be consistent. Note, however, that when the error term is in fact homoskedastic the normal standard error estimation process is in fact the least variance way of estimating the standard errors.

As a practical matter, most researchers just apply robust standard errors by default given that the robust standard error estimator is consistent for any form of error variance as long as the errors are independent.

We will consider the other two possible responses next week.