MLE: Lab 3

Shahryar Minhas

January 29, 2014

Setup

Muller & Seligson Regression

Functional Form

Omitted Variable Bias

Consequences of omitted variable bias?

- Biased parameter estimates and standard errors

In cases of more than two predictors direction of bias becomes hard to assess.

```
msrepBiRegData=na.omit(msrep[,c('deaths75ln', 'upper20', 'sanctions75ln')])
cor(msrepBiRegData)

# Estimate coefficients
coeftest( lm(deaths75ln ~ upper20 + sanctions75ln, data=msrepBiRegData) )

# If we excluded sanctions75ln from the model what would you expect of the
# coefficient on upper20
coeftest( lm(deaths75ln ~ upper20, data=msrepBiRegData) )

# If we excluded upper20 from the model what would you expect of the
# coefficient on sanctions75ln
coeftest( lm(deaths75ln ~ sanctions75ln, data=msrepBiRegData) )

# More complicated model
ivs=c('upper20', 'energypcln', 'intensep',
      'sanctions70ln', 'sanctions75ln', 'deaths70ln')
msrepRegData=na.omit(msrep[,c('deaths75ln', ivs)])
olsForm1=formula(paste0('deaths75ln ~ ', paste(ivs, collapse=' + ')))
olsForm2=formula(paste0('deaths75ln ~ ', paste(ivs[c(2:length(ivs))], collapse=' + ')))
mod1 = lm(olsForm1, data=msrepRegData)
mod2 = lm(olsForm2, data=msrepRegData)

# View model results
summary(mod1)
summary(mod2)
```

F-test

A partial F-test (also called an incremental F-test or an extra sum of squares F-test) is the appropriate test to use when the simultaneous test of the statistical significance of a group of variables is needed.

A partial F-test requires fitting two regression models.

- A full model (F) that includes all the variables currently of interest
- A reduced model (R) that includes all the variables currently of interest except those whose statistical significance we wish to test.

The partial F-test assesses whether the improvement in model fit (as assessed by a reduction in prediction error) using the full model is too large to be ascribed to chance alone.

- To determine this we look at the average difference in SSE between the two models and compare this to the average prediction error in the best model we have (the full model).
- If these two measures of error are roughly the same, the added regressors in the full model are not contributing very much.
- But if the the average decrease in prediction error due to adding the regressors is quite large, then these regressors are contributing a good deal of explanatory power and thus should be retained.

The formula for the of the partial F-test is:

```
# How would we do a partial F-test
ssr1=sum(resid(mod1)^2)
ssr2=sum(resid(mod2)^2)
chgReg=ncol(model.matrix(mod1)) - ncol(model.matrix(mod2))
# F statistic
Fstat=((ssr1-ssr2)/chgReg)/(ssr1/df.residual(mod1))
1-pf(abs(Fstat), chgReg, df.residual(mod1))

# Using anova function from base R
anova(mod1, mod2)
```

Now $F_{2,40,.05} = 3.23$. Since $19.53 > 3.23$ the result is statistically significant. So we conclude that even after controlling for education and income, occupational type has a significant effect on prestige due. Another way of saying this is that even after controlling for education and income, the mean prestige values accorded the occupation types by the survey respondents differ.

Likelihood Ratio Test

In statistics, a likelihood ratio test is a statistical test used to compare the fit of two models, one of which (the null model) is a special case of the other (the alternative model). The test is based on the likelihood ratio, which expresses how many times more likely the data are under one model than the other.

```
# How would we run a likelihood ratio test
ltest=nrow(model.matrix(mod2))*log(ssr1/ssr2)
pchisq(abs(ltest), df=chgReg, lower.tail=FALSE)

# Using lrtest from lmtest library
lrtest(mod1, mod2)
```

Non-Constant Coefficients

The Chow test is a statistical and econometric test of whether the coefficients in two linear regressions on different data sets are equal. The Chow test was invented by economist Gregory Chow in 1960. In econometrics, the Chow test is most commonly used in time series analysis to test for the presence of a structural break.

The Chow test is used to see if it makes sense to run two separate regressions on two mutually exclusive subsets of your data (divided by a break point) by comparing the results of the two "unrestricted" regressions versus the "restricted" regression that pools all the data together.

The procedure is as follows: * Run a "restricted" regression on all your data (pooled). * Divide your sample into to groups, determined by your breakpoint (e.g. a point in time, or a variable value). * Run an "unrestricted" regression on each of your subsamples. You will run two "unrestricted" regressions with a single breakpoint.

Partial Residual Plots

Non-linearity in parameters. Parabolic relationships?

Wages-Education.

BIT signings - Level BITs example.

Heteroskedasticity

Errors may increase as the value of an IV increases. For example, consider a model in which annual family income is the IV and annual family expenditures on vacations is the DV. Families with low incomes will spend relatively little on vacations, and the variations in expenditures across such families will be small. But for families with large incomes, the amount of discretionary income will be higher. The mean amount spent on vacations will be higher, and there will also be greater variability among such families, resulting in heteroscedasticity

Similar examples: Error terms associated with very large firms might have larger variances than error terms associated with smaller firms. Sales of larger firms might be more volatile than sales of smaller firms.

Errors may also increase as the values of an IV become more extreme in either direction, e.g. with attitudes that range from extremely negative to extremely positive. This will produce something that looks like an hourglass shape:

Measurement error can cause heteroscedasticity. Some respondents might provide more accurate responses than others. (Note that this problem arises from the violation of another assumption, that variables are measured without error.)

Heteroscedasticity can also occur if there are subpopulation differences or other interaction effects (e.g. the effect of income on expenditures differs for whites and blacks). (Again, the problem arises from violation of the assumption that no such differences exist or have already been incorporated into the model.) For example, in the following diagram suppose that Z stands for three different populations. At low values of X, the regression lines for each population are very close to each other. As X gets bigger, the regression lines get further and further apart. This means that the residual values will also get further and further apart.

Other model misspecifications can produce heteroskedasticity. For example, it may be that instead of using Y, you should be using the log of Y. Instead of using X, maybe you should be using X2, or both X and X2. Important variables may be omitted from the model. If the model were correctly specified, you might find that the patterns of heteroskedasticity disappeared.

Consequences

- Heteroscedasticity does not result in biased parameter estimates.
- However, OLS estimates are no longer BLUE. That is, among all the unbiased estimators, OLS does not provide the estimate with the smallest variance. Depending on the nature of the heteroscedasticity, significance tests can be too high or too low. As Allison puts it: "The reason OLS is not optimal when heteroskedasticity is present is that it gives equal weight to all observations when, in fact, observations with larger disturbance variance contain less information than observations with smaller disturbance variance."
- In addition, the standard errors are biased when heteroskedasticity is present. This in turn leads to bias in test statistics and confidence intervals.

Ways of testing: Breusch-Pagan Breusch-Pagan / Cook-Weisberg tests the null hypothesis that the error variances are all equal versus the alternative that the error variances are a multiplicative function of one or more variables. For example, in the default form of the hettest command shown above, the alternative hypothesis states that the error variances increase (or decrease) as the predicted values of Y increase, e.g. the bigger the predicted value of Y, the bigger the error variance is. A large chi-square would indicate that heteroscedasticity was present. In this example, the chisquare value was small, indicating heteroskedasticity was probably not a problem (or at least that if it was a problem, it wasn't a multiplicative function of the predicted values).

```
# How to calculate the Breusch-Pagan test statistic
residMod1=resid(mod1)^2
bpForm=formula(paste0('residMod1 ~', paste(ivs, collapse=' + ')))
bpMod=lm(bpForm, data=msrepRegData)
bpStat=summary(bpMod)$r.squared*nrow(msrepRegData)
1-pchisq(bpStat, df=length(ivs))

# Breusch-Pagan test: using bptest from lmtest library
bptest(mod1)
```

Ways of testing: Visualization

```
# Can use the sandwich package in combination with lmtest to help
# us calculate robust standard errors
coeftest(mod1, vcov=vcovHC(mod1, type='HC1'))
coeftest(mod1)
```

Dealing with heteroskedasticity: Robust standard errors