JBS Quantitative Research Methods Module MPO1

Michaelmas 2010
Thilo Klein
http://thiloklein.de

Computer Lab Session 3
Model Selection; Inference; Dummy Variables

Contents

Exercise 1. Multicollinearity; Omitted Variable Bias

Create y, z, x1, x2 and x3, generated as follows. Let n=10000, let $\varepsilon_i, \omega_i, \eta_i, \varsigma_i \sim N(0,1)$ be independent random variables with standard normal distribution, *i=1, ..., n.* Define:

$$x_{1i} = 5 + \omega_i + 0.3 \cdot \eta_i$$
$$x_{2i} = 10 + \omega_i$$
$$x_{3i} = 5 + \eta_i$$
$$y_i = 20 + x_{1i} + x_{2i} + \varepsilon_i$$
$$z_i = 30 + x_{2i} + x_{3i} + \varsigma_i$$

Then, see Exercise 4 in Lab Session 3, $Var(x_1) = Var(x_2) = 1$; $\rho_{X1,X2} = 0.958$; $\rho_{X2,X3} = 0$.

a) Create a sample of 10000 observations and generate the variables.
b) Regress y on a constant, x1 and x2. Comment on the outcome.
c) Regress z on a constant, x2 and x3. Comment on the outcome.
d) Regress y on a constant and x1, compare this with the regression of y on a constant, x1 and x2. Regress z on a constant and x2, compare this with the regression of z on a constant, x2 and x3.

Help:

To create variables with a standard normal distribution:

- **varname <- qnorm(runif (n=1000, min=0, max=1))**

or simply:

- **rnorm(n=1000, mean=1, sd=1)**

a)

```
epsilon <- rnorm(1000)
omega <- rnorm(1000)
eta <- rnorm(1000)
zeta <- rnorm(1000)

x1 <- 5 + omega + 0.3* eta
x2 <- 10 + omega
x3 <- 5 + eta
y <- 20+ x1 + x2 + epsilon
z <- 30+ x2 + x3 + zeta

cor(cbind(x1, x2, x3))
```

b)

- **lm1b <- lm(y ~ x1 + x2) robust?**
- **vif(lm1b)**

Note all estimators are very close to the population values. This is what we would have expected, as they are *unbiased*, when no relevant variables are omitted (there is no correlation between either x1 or x2 and epsilon). In this case, however, they're not spot on, though. Why would this be? The problem is one of multicollinearity, as both regressors are highly correlated. The impact that this problem has on the estimated coefficients is not to bias them, but to increase their variance, thus increasing the range of the confidence interval. This is the only reason why the parameters are not as close to 1 as in part c, below.

Variance Inflation Factor (VIF)

Under Gauss-Markov assumptions, the variance of the OLS estimator for a typical regression coefficient can be shown to be the following

$$\text{Population variance of } b_2 = \sigma^2_{b_2} = \frac{\sigma^2_u}{n\,Var(X_i)} \times \frac{1}{1-R^2_i}$$

where R_i is the unadjusted R^2 when you regress X_i against all the other explanatory variables in the model, that is, against a constant, X_2, …., X_{i-1}, X_{i+1}, ., X_k.

If there is no linear relation between X_i and the other explanatory variables in the model, R_i will be zero. Obviously, the diagnostic used for multicollinearity is related to R_i

$$Variance\ Inflation\ Factor_i = \frac{1}{\left(1-R^2_i\right)}$$

The VIF shows us how much the variance of the coefficient estimate is being inflated by multicollinearity. The bigger R^2_i is (i.e. the more highly correlated X_j is with the other regressors in the model), the bigger the standard error will be. Indeed, if X_i is perfectly correlated with the other regressors $\left(R^2_i = 1\right)$, the standard error will equal infinity. This is referred to as the problem of perfect multicollinearity.

As the Xs become more highly correlated, it becomes more and more difficult to determine which X is actually producing the effect on Y. A R^2_i close to 0 means there is little multicollinearity, whereas higher values suggest that multicollinearity may be a threat. The square root of the VIF tells you how much larger the standard error is, compared with what it would be if that variable were uncorrelated with the other X variables in the equation. For example, if VIF for a variable were 9, its standard error would be three times as large as it would be if its VIF was 1. In such a case, the coefficient would have to be 3 times as large to be statistically significant. VIF-statistic ranges from 1.0 to infinity. VIFs greater than 10.0 for any variable are generally seen as indicative of severe multicolinearity.

In this case, the value for VIF is very high, confirming what we observed in the correlation matrix above.

c)

- **lm1c <- lm(z ~ x2 + x3)**

Note all estimators are spot on.

d)

lm(y ~ x1)
lm(y ~ x1 + x2)
lm(z ~ x2)
lm(z ~ x2 + x3)

We observe then that, while in the first regression with y, the estimator b_2 is biased (the bias being equal to $\beta_2 \cdot \frac{\text{cov}(X_1, X_2)}{Var(X_1)} = 1$)[1], in the first regression with z the estimator b_2 is unbiased because $\frac{\text{cov}(X_2, X_3)}{Var(X_2)} = 0$. The constant in the first regression with z is biased, though. Why?

[1] This is a formulae for the bias presented in Wooldridge, equivalent to the one given in class for large samples, therefore the magnitude of the inconsistency, in the case when $u = \beta_2 X_2$.

Note that according to the true model $\bar{z} = 30 + \bar{x}_2 + \bar{x}_3$, but we are not including x_3, so the constant in the new model is equal to $\bar{z} - \bar{x}_2$, which is actually equal to $30 + \bar{x}_3 = 35$. The same or worse occurs with the first model with y. Now, not only is there an omitted variable bias for b_2, but the estimated constant is also biased.

Exercise 2. Multicollinearity; Variance Inflation Factor

Use salary.txt to estimate *log(salary)* on *education, starting salary (in logs), gender* and if the individual is member of a minority. Use an indicator to assess the problem of multicollinearity.

Remember:

$$Var(b_j) = \frac{\sigma^2}{n \cdot Var(X) \cdot (1 - R_j^2)}, \quad (j = 2,3,\ldots,k)$$

Then, the correlation between the regressors, assessed by R_j^2, is used to assess how much this correlation inflates the variance of the estimator b_j. $1/(1 - R_j^2)$ are the variance inflation factors.

Answer:

- **salary <- read.table("salary.txt", header=T, sep=" ")**

1st question: Is n big?

- **str(salary)**

n=474, therefore it is big. You can use robust standard errors.

- **lm2 <- lm(LOGSAL ~ EDUC + LOGSALBEGIN + GENDER + MINORITY, data=salary)**
- **vif(lm2)** # remember: library(car)

None of the variables' vif is higher than 10, so no problem identified this way.

In order to confirm vif results above, I calculate them without the built in command in the following part of the solution. It may be also useful if you wanted to apply similar methods to compute other statistics.

- **lm2ed <- lm(EDUC ~ LOGSALBEGIN + GENDER + MINORITY, data=salary)**

After each regression R keeps several of the statistics calculated for further use. You can see which statistics are kept by typing **str(lm2) and summary(lm2)** and scrolling down to the list of saved results.

```
print( paste( "The R2 is",  summary(lm2ed)$r.squared ) )
print( paste( "variance inflation factor is",  1/(1- summary(lm2ed)$r.squared) ) )
```

The VIFs estimated for each regressor are not that big. We calculate the R_j^2 (0.47, 0.33, and 0.07, 0.59). To calculate the impact on the standard deviation of the estimators we calculate vif:

$$vif = \frac{1}{(1 - R_j^2)}$$

The highest result is 2.45 (for logsalbegin), which doesn't seem to be so serious. Several present results which are over 1.5, so we want to do another test as well.

- **attach(salary)**
- **cor(cbind(EDUC, LOGSALBEGIN, GENDER, MINORITY))**
- **detach(salary)**

logsalbegin and educ are quite highly correlated (very close to .7, which could be the limit of the 'red zone'; being worryingly high over, say, 0.6). Could we drop any of these variables or combine them somehow? Probably not, so we may just leave the model as it stands. Especially so, considering that with this issue not much can be done.

Exercise 3. The effects of having highly correlated regressors

Use eaef21.csv.

- **eaef21 <- read.csv("eaef21.csv", header=T)**

SM and *SF* represent years of schooling corresponding to each of the individuals' parents.

a) Investigate the determinants of family size by regressing *SIBLINGS* on *SM* and *SF* for different ethnic groups. *SM* and *SF* are likely to be highly correlated (find the correlation in your data set) and the regression may be subject to multicollinearity. Check this.

b) Run a regression of SIBLINGS on SM and SF for the whole sample. Test and if possible introduce the restriction that the theoretical coefficients of *SM* and *SF* are equal. Run the regression a second time, replacing *SM* and *SF* by their sum, *SP*. Evaluate the regression results. Do you think proceeding like this would improve in any sense the estimations?

First we ask ourselves whether we have big samples for all ethnic groups.

```
sum( ifelse(eaef21$ETHWHITE==1, 1, 0))
sum(eaef21$ETHBLACK)
sum(eaef21$ETHHISP)
```

The samples for the white and black ethnic groups are reasonably big. We can thus use robust standard errors for regressions with these two samples. The sample size for the Hispanic ethnic category is 33, which is perhaps too small to allow us to use robust standard errors.

a)

```
lm3w <- lm(SIBLINGS ~ SM + SF, data=eaef21[eaef21$ETHWHITE==1, ])
cor( cbind(eaef21$SM, eaef21$SF)[eaef21$ETHWHITE==1, ] )
vif(lm3w)

lm3b <- lm( SIBLINGS ~ SM + SF, data=subset(eaef21, ETHBLACK ==1) )
cor( subset(eaef21, ETHBLACK ==1, select=c(SM, SF)) )
vif(lm3b)

lm3h <- lm( SIBLINGS ~ SM + SF, data= subset(eaef21, ETHHISP==1) )
cor( subset(eaef21, ETHHISP ==1, select=c(SM, SF)) )
vif(lm3h)
```

Clearly the greater the education of the parents, the smaller the number of children in the household. The opportunity cost of dedicating more human capital away from obtaining incomes may be a good reason. Similarly, it could explain women's behaviour regarding fertility, as they tend to have their first children later in life in order to be able to invest in education.

SF and SM are quite highly correlated for all groups, though the variance inflation factors are not very big.

SF is always non-significant. As we include SM as well it is possible that this variable is capturing family background and so in this case there is the classic ambiguity caused by multicollinearity. It is possible that father's education has no effect on family size. While we expect that it actually does, the correlation between SM and SF, being 0.58 or higher, combined with a relatively small sample size (for this kind of work) conspire to make the standard error so large that the coefficient is insignificant in all cases.

b) The model is:

$$Siblings = \beta_0 + \beta_1 \cdot SM + \beta_2 \cdot SF + u$$

- **lm3.1 <- lm(SIBLINGS ~ SM + SF, data=eaef21)**

The test of hypothesis could be equal to: $Ho : \beta_1 = \beta_2$

- **linearHypothesis(model=lm3.1, "SM=SF")**

the restriction is not rejected at 5%, so the model could be estimated as follows:

$$Siblings = \beta_0 + \beta_1 \cdot SM + \beta_2 \cdot SF + u = \beta_0 + \beta_1 \cdot (SM + SF) + u$$

(Remember that by imposing restrictions which are not rejected we produce more efficient estimators).

- **eaef21$SP <- eaef21$SF + eaef21$SM**
- **lm3.2 <- lm(SIBLINGS ~ SP, data=eaef21)**

The new estimation is between both previous coefficient estimations, and the standard deviation of the new estimator is smaller than any of the previous estimators. We have dealt with the issue of multicollinearity.

Exercise 4. OLS assumptions; dummy variables (optional)

What are the main assumptions underlying the classical linear regression model. What do they mean? Do this in your own time if you wish.

Answer: There are 5 main assumptions:

1) $E(\varepsilon_i) = 0$ – The errors have zero mean.
2) $Var(\varepsilon_i) = \sigma^2 < \infty$ – The variance of the errors is constant and finite over all values of x_i. This is the homoscedasticity assumption.
3) $Cov(\varepsilon_i, \varepsilon_j) = 0$ – The errors are statistically independent of one another. This is the no autocorrelation assumption.
4) $Cov(\varepsilon_i, x_i) = 0$ – There is no relationship between the error and the corresponding x.
5) $\varepsilon_i \sim N(0, \sigma^2)$ – The errors are normally distributed. This is the normality assumption

Dummy variables (optional)

Yet another possible problem related to the *ceteris paribus* assumption can motivate the discussion of the use dummy variables. This one arises when we define

$$Y_i = \beta_2 \cdot D_{1i} + \beta_3 \cdot D_{2i} + u_i ,$$

when D_{1i} and D_{2i} are dummy variables collecting two mutually exclusive or disjoint states (those defined as 'the result of the state is either one or the other'), which, together define all

possible situations. For instance the D_1 is the first half of the year and D_2 the second half of the year, and of course for any individual we cannot change the value of variable D_1 (say for 1 to 0) without changing the value of variable D_2.

This is an important reason to do the following transformation to this model. We know that $D_{1i} + D_{2i} = 1$. Then, $D_{1i} = 1 - D_{2i}$, and so the model above is changed into $Y_i = \beta_2 \cdot (1 - D_{2i}) + \beta_3 \cdot D_{2i} + u_i$ which is the same, after operating to run:

$$Y_i = \beta_2 + (\beta_3 - \beta_2) \cdot D_{2i} + u_i$$

It is evident that when you include all dummies except for one, the constant collects the effect corresponding to the default state and the included dummies collect the effect due to the dummy wise specified state as an increment on the constant. In our example, β_2 collects the effect of the first half of the year and $(\beta_3 - \beta_2)$ what the second half of the year adds to the first.

Note that you can do the same with the slopes, i.e., not only with the constant but with all other independent variables as well. Example: consider the model:

$$Y_i = \beta_1 + \beta_2 \cdot X_{1i} + u_i .$$

Is β_2 different for the second half of the year? The new model would be:

$$Y_i = \beta_1 + \beta_2 \cdot D_{1i} \cdot X_{1i} + \beta_3 \cdot D_{2i} \cdot X_{1i} + u_i .^2$$

In this case, as we did previously, we know that $D_{1i} + D_{2i} = 1$. Then, $D_{1i} = 1 - D_{2i}$, and the model is changed into: $Y_i = \beta_1 + \beta_2 \cdot (1 - D_{2i}) \cdot X_{1i} + \beta_3 \cdot D_{2i} \cdot X_{1i} + u_i$, or equivalently

$$Y_i = \beta_1 + \beta_2 \cdot X_{1i} + (\beta_3 - \beta_2) \cdot D_{2i} \cdot X_{1i} + u_i .$$

As before, we just include one of the 2 dummies, which collect the additional effect of the second half of the year now on the coefficient of X_1.

Variable transformations which consist of multiplying a regressor by a dummy variable, as above, are called *interaction terms*.

Exercise 5. Bank wages

bank.csv includes information on salaries in a US bank. Describe and summarize to see the contents.

- **bank <- read.csv("bank LS4.csv", header=T)**
- **str(bank)**

i) Regress the log of salaries on a constant, education, the log of the starting salary, and define a way to capture percentage differences due to gender and belonging to a minority.

ii) Create a dummy variable for each job category and estimate a model to observe effects generated by these job-categories. Is there a significant difference in income between the job categories?

[2] Note that if the observation is in the first semester $D_{1i}=1$ but $D_{2i} =0$ so the model becomes: $Y_i = \beta_1 + \beta_2 \cdot D_{1i} \cdot X_{1i} + u_i$. And it's the other way around in the second semester.

iii) Estimate the same model for those employees with custodial jobs (jobcat=2) and for those with managerial jobs (jobcat=3). Why do you think that some variables are dropped? (Hint: do: **sum(bank$male[bank$jobcat==2**])

iv) Are there significant differences between minority and non-minority employees? By gender? Is there any variation due to being simultaneously female and also minority?

v) Test for the joint significance of gender and if they are part of minorities. What's the effect of imposing these restrictions?

vi) Test the hypothesis of the returns to education being = 7%. Then test this hypothesis jointly which the hypothesis of female and minority having the same discriminatory effect.

vii) salbegin is the salary received by the individual when starting his position at the same bank. Test whether the log of salary when beginning has a significantly different effect for managerial jobs. What does this mean?

i) First of all we check whether we have a big or a small sample. It's big, so we use robust standard errors.

First things first. They ask us to work with males and log(salary). They give us these variables already. Note that differences of variables in log multiplied by 100 are used to estimate percentage increments, if these are low. In this exercise we estimate:

$$\log(salary) = \beta_0 + \beta_1.education_i + \beta_2.\log salbegin_i + \beta_3.male_i + \\ + \beta_4.\min ority_i + u_i$$

- **logsalbegin <- log(salbegin)**
- **lm5i <- lm(logsal ~ educ + logsalbegin + male + minority, data=bank)**

Then given all the rest of the characteristics, the difference between two typical individuals, a female and a male would be given by β_3, which indicates how much more (if negative, less) men earn than women – the results suggest that males earn 8% more on average. Likewise, β_4 indicates differences against (if negative) of minorities – minorities appear to earn 8% less.

ii) There are two ways to generate dummy variables in R:

1. with command **ifelse**

- **bank$jobcat.cler <- ifelse(bank$jobcat==1, 1, 0)**
- **bank$jobcat.cust <- ifelse(bank$jobcat==2, 1, 0)**
- **bank$jobcat.man <- ifelse(bank$jobcat==3, 1, 0)**
- **lm(logsal ~ educ + male + minority + jobcat.cler + jobcat.cust, data=bank)**

This command creates a variable per category. You should use information about which category corresponds to which variable, e.g. 1="clerical", 2="custodial", 3="managerial". Remember to leave one category aside!!

2. with command **factor**

- **bank$jobcat.fac <- factor(bank$jobcat)**
- **levels(bank$jobcat.fac) <- c("Clerical", "Custodial", "Managerial")**
- **lm(logsal ~ educ + male + minority + jobcat.fac, data=bank)**

This command creates a variable per category and includes them in the regression. As before, it's up to you to check which category corresponds to which variable.

There are significant differences between both categories explicitly included with variables and the third category, which acts as default (t-values higher than 1.96).

➢ **linearHypothesis(model=lm5iia, "jobcat.cler = jobcat.cust", vcov=hccm)**

There are also differences between the two included as we reject the null of both coefficients being equal. Note that since we estimated with robust, this test calculates the ***heteroscedastic-robust F-statistic*** which, for big samples, tends to be distributed as χ_1^2 .

iii)

➢ **table(bank$jobcat)**

We observe that category custodial (2) does not represent a large subsample. The inferences regarding this category should rely on the disturbances being *homoscedastic* and *normally distributed* (assumptions which we will learn how to test soon).

```
lm(logsal ~ educ + male + minority, data=subset(bank, jobcat==2) )
lm(logsal ~ educ + male + minority, data=subset(bank, jobcat==3) )
sum(bank$male[bank$jobcat==2])
```

No females do custodial jobs. So male should not have been included. This does not happen with managerial jobs.

iv)

Yes and yes. The null hypotheses of the coefficients being zero are rejected. To capture specific differences for female pertaining to the minority we do

```
bank$female <- ifelse(bank$male==0, 1, 0)
bank$femaleandminority <- bank$female * bank$minority
lm5iv <- lm(logsal ~ educ + male + minority + femaleandminority, data=bank)
```

As we can see, there is a special negative effect (discrimination) this group seems subject to.

v)

➢ **linearHypothesis(model=lm5iv, "male = minority")**

We want to know if at least one is statistically different from 0. (We knew it already, I know).

vi)

```
lm5vi <- lm(logsal ~ educ + female + minority , data=bank)
linearHypothesis(lm5vi, "educ = 0.07")
```

It is not rejected.

➢ **linearHypothesis(lm5vi, c("educ = 0.07", "female = minority"))**

$Ho: \beta_{educ} = 0.07 \; and \; \beta_{minority} = \beta_{female}$

$Ha: either \; \beta_{educ} \neq 0 \; or \; \beta_{minority} \neq \beta_{female}$

The null of these hypotheses is rejected, the effect is statistically higher for female, and that's the reason to reject that both hypotheses apply, despite the fact that on its own, the first hypothesis is **not** rejected.

vii)

➢ **lm(logsal ~ educ + female + minority + logsalbegin + logsalbegin:jobcat.man, data=bank)**

logsalbegin is significant indicating the importance of negotiating a good salary when starting. The interaction term with the managerial category is significant indicating a different in the

slope for this category when compared with the rest of the categories. This means that for managers, an initial salary being 10% higher corresponds to a 0.19% higher current salary, compared to non-managers (a low impact, significant, yes, but low).

Exercise 6. NO2 pollution

Nitrogen dioxide (NO_2) is a pollutant that attacks the human respiratory system and increases the likelihood of respiratory illness. One common source of nitrogen dioxide is automobile exhaust. File NO2pollution.csv contains a subset of 500 hourly observations made from October 2001 to August 2003. Variables are:

LNO2	: ln(concentration of NO_2)
lcars	: ln(number of cars per hour)
temp	: temperature 2 meters above ground
wndspd	: wind speed
tchng23	: temperature difference between 25 and 2 meters above ground.
WNDDIR	: wind direction
HOUR	: hour of day
DAYS	: day number from October 1, 2001

a) Regress the log of NO_2 concentration on the log of the number of cars, the two temperature variables, the two wind variables and the time index (days). Test whether wind speed and direction have the same impact.

b) The sample has 500 observations. Does the validity of the F-test described in (a) rest heavily on having normally, or almost normally distributed disturbances?

- **no2poll <- read.csv("no2pollution.csv", header=T)**

a) The sample is big, so we run with robust.

- **lm6 <- lm(lno2 ~ lcars + temp + tchng23 + wndspd + wnddir + day, data=no2poll)**
- **linearHypothesis(lm6, "wndspd = wnddir")**

We reject this hypothesis.

b)

Given the sample is large with n=500, the validity of the test does NOT rest heavily on having normally distributed variances.

Exercise 7. Programming in R (strictly optional)

We use the dataset employed in the first lab-session: dataset.csv. Do larger companies (those with higher *rev)* grow at a faster rate than small ones? Create a Script-file to analyse this question.

a) Create a Script-file with the recommended header (see the hand out on R programming).
b) Load dataset.csv
c) Create a variable enumerating the companies.
d) Sort the data considering comp_name and year. Inspect the result of your sorting.
e) Are all observations reported for consecutive years? Or is it the case that years are missing in the span for some companies?
f) Create a new variable with the rate of growth for each company defined as:

$$revgrow = \frac{rev\_n - rev\_n{-}1}{rev\_n{-}1}.$$

g) List whatever variables you need to inspect that you did this correctly. Why would this variable lead to problems?

h) Create a new variable named:

$$revgrowcorrected = \left(\frac{\left(rev\_n - rev\_n{-}1\right) \Big/ \left(year\_n - year\_n{-}1\right)}{rev\_n{-}1} \right)$$

Include comments to remind yourself the reasons for creating this variable.

i) Visually inspect the variables involved.

j) Estimate the regression for the whole sample.

k) Divide the sample in 3 sub-samples: up to 1998; up to 2001; the rest. Observe results. Estimate the regression for each sub-sample.

Some help:

c) Look up the command **as.numeric** in the R help.

d) Look up the command **order**, in the R help.

e) Look at company ACE Technosoft. Also, assertions are useful commands when programming in R to deal with these sorts of questions. If you state an assertion, R responds with whether this assertion is correct. In this case we'd say that for each individual the difference between years is 1.

f) Here you can make use of R's vector structure. In the next lab session, we will have a closer look at how to generate lagged variables to calculate first differences.