JBS Masters in Finance Econometrics Module

Michaelmas 2010
Thilo Klein
http://thiloklein.de

Computer Lab Session 2
Model Selection; Inference; Non-linear Models

Contents

Exercise 1. Confidence intervals of regression coefficients

Use oilprice1.csv. This is an example that can be found expanded in the very good book by Murray (2006). It has to do with a trial, where the judge has to decide if the price differential charged to oil suppliers to a pipe because of differences in the quality of oil is fair (based on the market premium for quality). The quality is measured in API degrees (the more the more the quality). Up to the trial the additional price charged is 0.15$ per API degree of oil. The challengers want a price between 3 and 5 cents. In this data set you have information on the crude oil's quality and price per barrel.

a) observe by a scatter-plot if quality has any impact on prices.

b) use regression analysis to quantify this relationship.

c) construct a 95% confidence interval for how much the price of barrel changes when API increases in one degree.

d) Is then the price charged fair?

e) Perform the same results in cents. For this create a new variable multiplying the price by 100. Does the fit of the regression change? What happens to the coefficients estimated?

Answers:

a)

- **str()**
- **plot(price ~ api, data=oilprice1)** # there's a strong relationship

b)

- **lm1 <- lm(price ~ api, data= oilprice1)**
- **summary(lm1)**

c)

$$P\left(b_2 - se(b_2)\cdot t_{0.975} \le \beta_2 \le b_2 + se(b_2)\cdot t_{0.975}\right) = 0.95$$

Lookup 0.975-quantile of t-distribution with n-2 degrees of freedom:

- **qt(p=0.975, df=13-2)** # = 2.201

$$0.09493 - 0.00827*2.201 \le \beta_2 \le 0.09493 + 0.00827*2.201$$

$$0.08 \le \beta_2 \le 0.11$$

d) The results in the last part indicate that the price that the market pays is statistically within the interval 9 cents and 11 cents per API degree, more would be unfair. The price suggested by the challenger is too low.

e) This is another example of Exercise 5. The only change is in the scale of the results. R^2 does not change.

Exercise 2. Omitted variable bias and highly correlated regressors

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, $i=1, \ldots, n$. 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)

- **lm2b <- lm(y ~ x1 + x2) robust?**
- **vif(lm2b)**

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

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

$$Variance\ Inflation\ Factor_i = \frac{1}{(1-R_i^2)}$$

The VIF shows us how much the variance of the coefficient estimate is being inflated by multicollinearity. The bigger R_i^2 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 $(R_i^2 = 1)$, 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_i^2 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)

- **lm2c <- 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?

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.

[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 $u = \beta_2 X_2$.

Exercise 3. Non-linear models

Are any of the following models linear in the parameters?

i) $Y_i = \beta_1 \cdot X_i + \beta_2\beta_3 \cdot Z_i + u_i$

ii) $\ln(Y_i) = \beta_1 \cdot X_i + \beta_2 \cdot Z_i + u_i$

iii) $Y_i = \beta_1 \cdot X_i^2 + \beta_2 \cdot Z_i + u_i$

iv) $Y_i = \beta_1 \cdot X_i + \beta_1^2 \cdot Z_i + u_i$

v) $Y_i^4 = \beta_1 \cdot X_i + \beta_1^2 \cdot Z_i + u_i$

Answer: only models ii) and iii) are linear in the parameters.

If the model is linear ***in the parameters***: the parameters should not be squared, logged, inverted or transformed by any function. The variables that multiply these parameters, however, may be transformed – this allows us much greater flexibility in capturing aspects of the relationship. The following table describes some of these possibilities.

Model	**Dependent variable**	**Independent variable**	**Algebraic interpretation of β_1**	**Conceptual interpretation of β_1**
level-level	Y	X	$\Delta Y = \beta_1 \Delta X$	A constant level change after change in one unit of X
Semi-log log-level	log(Y)	X	$\%\Delta Y = (100*\beta_1)\Delta X$	A constant % change in Y after change in one unit of X
Double-log log-log	log(Y)	log (X)	$\%\Delta Y = \beta_1 \% \Delta X$	A constant % change in Y after change of X in 1%

Examples:

<u>Level-level</u>: Example: exercise 3 in session 1. If *height* changes by 1 unit (one inch, as it is measured in inches), how much does *weight* increase – in pounds?

<u>Semi-log (or Log-level)</u>: In this case an increase of X in one unit always leads to the same increment *in percentage* in Y. For instance: $log(wage) = \beta_0 + \beta_1 education + u$.
In this case, β_1 gives us the percentage by which wages change with a change of one unit (say one more year) in education. Concept: "rate of return" to education.

Note that in this model the assumption is that the rate of return is identical for all the education levels. A uniform rate of return is estimated for any additional year in school or any additional year in college.

What would the β_1 capture in the following model? $log(profit\ level) = \beta_0 + \beta_1 capital + u$

<u>Double-log or Log-log</u>: In this case β_1 gives you the percentage change in variable Y after a change of X by 1%. Due to an increase (or decrease) of one per cent in X, by how many percentage points will Y change? This is the concept of elasticity.

Take, for instance: $log(supply\ of\ labour) = \beta_0 + \beta_1 log(wages) + u$. This is the elasticity of labour supply.

$log(demand\ of\ cars) = \beta_0 + \beta_1 log(car\ prices) + \beta_2 log(household\ income) + u$. Now β_1 captures the price elasticity of the demand for cars and β_2 collects the income elasticity of the demand for cars.

Exercise 4. Hypotheses testing in the Log-log model

Use gasoline.csv. Suppose you want to estimate the demand for gasoline in a given country (Uruguay) as a function of the GDP and the price of gasoline in that country:

$$log(gasoline_t) = \beta_0 + \beta_1 log(GDP_t) + \beta_2 \ log(price_t)$$

a) Run the previous regression. Interpret the coefficients for GDP and price of gasoline.
b) Is the "income" elasticity equal to one? Test this hypothesis.

a) Read the data and estimate the log-log model:

- **gas <- read.csv("gasoline.csv", header=T)**
- **lm4 <- lm(log(gasoline) ~ log(gdp) + log(price), data=gas); summary(lm4)**

The coefficients capture the income and price elasticity of demand for gasoline, respectively. A 1% increase in GDP produces a 0.86% increment in the demand of gasoline. A 1% increment in the price of gasoline produces the effect of reducing the demand in 0.35%.[2]

b) We want to test the hypothesis:

$$H0: \beta_1 = 1; \quad Ha: \beta_1 \neq 1$$

The function linearHypothesis from R's car library computes a F-statistic for carrying out a Wald-test-based comparison between our original model lm4 and a linearly restricted model, where $\beta_1 = 1$.

- **library(car)**
- **linearHypothesis(model=lm4, "log(gdp)=1")**

The p-value of 0.02 leads us to reject the null hypothesis of constant income elasticity (b_1=1) at the 5% level. There is little evidence for constant income elasticity of demand for gasoline in Uruguay.

Exercise 5. Non-linear models. Production function. Multiple hypotheses.

Use usmetal.txt. Dataset: production data for the year 1994; n=26; US firms in the sector of primary metal industries. For each firm, values are given of production (*y*, value added in millions of dollars), labour (*L*, total payroll in millions of dollars), and capital (*K*, capital stock in millions of 1987 dollars).

a) Generate new variables as logs of the old variables. Inspect the variables. (graph with histogram and scatter)
b) Using a double log specification, estimate a production function. (This has been called the Cobb-Douglas production function). Comment on the coefficients.
c) Test the hypothesis that the sum of the coefficients is equal to 1.
d) Impose the restriction and re-estimate. Compare the standard error for the estimator of β_3.

Help for c): Cobb-Douglas functions

The Cobb-Douglas function is defined as follows:

[2] In strict terms the t-Statistics reported here cannot be compared with the t distribution since the variables are non-stationary (this is an advanced problem that is treated in time series courses). But just for the sake of the exercise let's pretend that the p-values are still valid.

$$Y_i = \beta_1 \cdot K_i^{\beta_2} \cdot L_i^{\beta_3}$$

therefore:[3]

$$\log(Y_i) = \beta_1 + \beta_2 \cdot \log(K_i) + \beta_3 \cdot \log(L_i)$$

d) **linearHypothesis(model=lm1c, "lK=lL")**. The hypothesis is rejected at the 5% level.

e) In order to impose the restriction take into account: $Ho : \beta_2 + \beta_3 = 1$
Explanation: CRS is such that: f(λK, λL)=λf(K,L). Then:

$$Y_i = \beta_1 \cdot K_i^{\beta_2} \cdot L_i^{\beta_3}$$

therefore, if CRS,

$$\beta_1 \cdot (\lambda K_i)^{\beta_2} \cdot \lambda L_i^{\beta_3} = \lambda \cdot \beta_1 \cdot K_i^{\beta_2} \cdot L_i^{\beta_3}$$

then,

$$\lambda^{\beta_2+\beta_3} \cdot \beta_1 \cdot (K_i)^{\beta_2} \cdot (L_i)^{\beta_3} = \lambda \cdot \beta_1 \cdot K_i^{\beta_2} \cdot L_i^{\beta_3}$$

and these expressions are equivalent if: $\lambda^{\beta_2+\beta_3} = \lambda$ or, equivalently: $\beta_2 + \beta_3 = 1$

- **linearHypothesis(model=lm1c, "lk+ll=1")**

f) To impose CRS first note that: $\beta_2 + \beta_3 = 1$; so: $\beta_2 = 1 - \beta_3$.

$$\log(Y_i) = \beta_1 + \beta_2 \cdot \log(K_i) + \beta_3 \cdot \log(L_i) + u_i$$

$$\log(Y_i) = \beta_1 + (1-\beta_3) \cdot \log(K_i) + \beta_3 \cdot \log(L_i) + u_i$$

$$\log(Y_i) - \log(K_i) = \beta_1 + \beta_3 \cdot [\log(L_i) - \log(K_i)] + u_i$$

In R there is no need to transform the variables. To subtract log(K) from log(Y) on the left hand side of the formula, we use R's **offset** command. To inhibit misinterpretation of the subtraction lL – lK, we use the function **I()**.

- **lm1f <- lm(lY ~ I(lL - lK), offset=lK, data=metal)**
- **anova(lm1c, lm1f)**

The final F-test, using the anova-function, is equivalent to the linear hypothesis of CRS and confirms that the models lm1c and lm1f are almost equal.

Exercise 6. Bank wages

bank.csv includes information on salaries in a US bank. Describe and summarize.

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

[3] Remember ln(AB)=ln(A)+ln(B); $\log(A^{\alpha})=\alpha$ * ln(A)

iii) 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.

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)**
- **lm6i <- lm(logsal ~ educ + logsalbegin + male + minority, data=bank)**
- **shccm(lm6i)**

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)

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**
- **lm6ii <- lm(logsal ~ educ + male + minority + femaleandminority, data=bank)**

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

iii)

- **lm6iii <- lm(logsal ~ educ + female + minority , data=bank)**
- **linearHypothesis(lm6iii, "educ = 0.07")**

It is not rejected.

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

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

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

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.