JBS Masters in Finance Econometrics Module

Michaelmas 2010
Thilo Klein
http://thiloklein.de

Computer Lab Session 3
The Generalized Linear Regression Model

Contents

Exercise 1. Heteroskedasticity (1)

a) Use the data in hprice1.csv to obtain the heteroskedasticity-robust standard errors and homoskedastic-only standard errors for equation:

$price = \beta_1 + \beta_2 lotsize + \beta_3 sqrft + \beta_4 bdrms + u$. Discuss any important difference with the usual homoskedasticity-only standard errors.

b) Repeat part a) for $log(price) = \beta_1 + \beta_2 log(lotsize) + \beta_3 log(sqrft) + \beta_4 bdrms + u$

c) What does this example suggest about heteroskedasticity and the transformation used for the dependent variable?

d) Apply the full White test for heteroskedasticity to part b). Which variables does it apply? Using the chi-squared form of the statistic, obtain the p-value. What do you conclude?

Answer:

a)

- **lm1a <- lm(price ~ lotsize + sqrft + bdrms, data=house)**
- **summary(lm1a)**
- **shccm(lm1a)**

The estimated equation with both sets of standard errors (heteroskedasticity-robust standard errors in brackets) is:

price_hat = -21.77 + 0.00207 lotsize + 0.123 sqrft + 13.85 bdrms
(29.48) (0.00064) (0.013) (9.01)
[36.28] [0.0012] [0.017] [8.28]
N=88 R^2=0.672

The robust standard error on lotsize is almost twice as large as the homoskedastic-only standard error, making lotsize much less significant (the t-statistic falls from about 3.22 to about 1.65). The t-statistic on sqrft also falls, but it is still very significant. The variable bdrms actually becomes somewhat more significant but is still barely significant. The most important change is in the significance of lotsize.

b)

- **lm1b <- lm(lprice ~ llotsize + lsqrft + bdrms, data=house)**
- **summary(lm1b); shccm(lm1b)**

For the log-log model:

log(price_hat) = -1.30 + 0.0168 log(lotsize) + 0.700 log(sqrft) + 0.037 bdrms
(0.65) (0.038) (0.093) (0.028)
[0.76] [0.041] [0.10] [0.030]
N=88 R^2=0.643

Here, the heteroscedasticity-robust standard error is always slightly greater than the corresponding usual standard error, but the differences are relatively small. In particular, log(lotsize) and log(sqrft) still have very large t-statistics, and the t-statistic on bdrms is not significant at the 5% level against a one-sided alternative using either standard error.

c) Using the logarithmic transformation of the dependent variable often mitigates, if not entirely eliminates, heteroskedasticity. (see Wooldridge section 6.2, Dougherty in chapter 7, section about non-linear models). This is certainly the case here, as no important conclusions in the model for log(price) depend on the choice of standard error. (We have also transformed two of the independent variables to make the model of the constant elasticity variety in lotsize and sqrft).

d) After estimating the equation in part b) we obtain squared OLS residuals. The full White-test is based on the R^2 from the auxiliary regression (with an intercept) on log(lotsize), log(sqrft), bdrms, $\log^2$(lotsize), $\log^2$(sqrft), $bdrms^2$, log(lotsize)·log(sqrft), log(lotsize)·bdrms, log(sqrft)·bdrms

- **house$lm1b.sqres <- lm1b$residuals^2**
- **lm1b.white.test <- lm(lm1b.sqres ~ llotsize*lsqrft*bdrms - llotsize:lsqrft:bdrms + I(llotsize^2) + I(lsqrft^2) + I(bdrms^2), data=house); shccm(lm1b.white.test)**
- **T <- summary(lm1b.white.test)$r.squared * nrow(house)**
- **pchisq(q=T, df=9, lower.tail=F)**

With 88 observations, the nR^2 version of the White statistic is 9.55, and this is the outcome of an (approximately) chi-squared random variable with 9 degrees of freedom. The p-value is about 0.385, which provides little evidence against the homoskedasticity assumption.

Exercise 2. Heteroskedasticity (2)

Use data training.csv.

a) Consider the simple regression model *log(scrap)*=$\beta_1 + \beta_2$ *grant* + *u*, where *scrap* is the firm scrap rate, and *grant* is a dummy variable indicating whether a firm received a job training grant. Can you think of some reasons why the unobserved factors in *u* might be correlated in *grant*?

b) Estimate a simple regression model. Does receiving a job-training grant significantly lower a firm's scrap rate?

c) Now, add as an a explanatory variable *log(scrap_1)* (this variable is the scrap rate of the previous year). How does this change the estimated effect of *grant*? Is it statistically significant at the 5% level?

d) Test the null hypothesis that the parameter on *log(scrap_1)* is 1 against the two-sided alternative. Report the p-value for the test.

e) Repeat parts c) and d) using heteroscedasticity-robust standard errors, and briefly discuss any notable differences.

Answer:

a) If the grants were awarded to firms based on firm or worker characteristics, grant could easily be correlated with factors that affect productivity. In the simple regression model presented, these are contained in *u*.

b) The simple regression estimates are obtained by

- **lm2b <- lm(log(scrap) ~ grant, data=training); summary(lm2b)**

The coefficient on grant is positive, but not statistically different from 0.

c) When we add log(scrap_1) to the equation, we obtain

- **lm2c <- lm(log(scrap) ~ grant + log(scrap_1), data=training); summary(lm2c)**

At the 5%-level, the coefficient pertaining to grant is not significant. (At the 10% it is significant but negative).

d) Test the linear hypothesis:

- **linearHypothesis(model=lm2c, "log(scrap_1) = 1")**

We strongly reject H0.

e) Heteroscedasticity-consistent coefficient variance-covariance matrix for model lm2c yields the summary

- **shccm(lm2c)**
- **linearHypothesis(model=lm2c, "log(scrap_1) = 1", white.adjust=T)**

Note that the standard errors for the variable grant do not change at all. Its coefficient therefore remains insignificant at 5%.

Linear hypothesis-test and regression summary using the White-adjusted coefficient variance-covariance matrix show that the standard error for the coefficient pertaining to scrap_1 is far higher now. P-values are higher as a consequence of it and the linear hypothesis is no longer significant at 5%-level.

Exercise 3. Autocorrelation

Use bonds.csv. It contains data on returns for AAA bonds and interest rates from US Treasury Bills from January, 1950 to December, 1999.

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

a) Regress changes in AAA bond returns (daaa) on US Treasury Bill interest rates (dus3mt). Plot the residuals. Are the residuals distributed evenly across time?

b) Investigate serial autocorrelation in residuals. Use the Breusch-Godfrey Serial Correlation LM Test and the Durbin-Watson test for auto-correlated errors.

Answer:

a) Regress changes in AAA bond returns on US Treasury Bill interest rates.

- **lm3a <- lm(daaa ~ dus3mt, data=bond); shccm(lm3a)**

To investigate serial autocorrelation in residuals, create and examine the residuals for this analysis, showing the residuals over time. To do this, first generate the variable to use to define the date:

- **bond$paneldate <- as.yearmon(bond$paneldate, format="%Ym%m")**

Then plot the residuals against this time-variable:

```
e <- lm3a$res
plot(e ~ bond$paneldate, type="l")
```

Note the following pattern in the residual series: high volatility (variance) is followed by high volatility and vice versa. Probably the residuals are correlated.

b) Let's compute the correlation of the residuals with the residuals in the previous periods (autocorrelation). You can do this in the following way

```
N <- length(e)
e1 <- c(NA, e[1:(N-1)])
e2 <- c(NA, NA, e[1:(N-2)])
plot(e ~ e1)
cor(e, e1, use="complete")
abline(a=0, b=0.2761491, col="red", lwd=2)
```

As you can see the residuals in period t are correlated with the residuals in the previous period. In fact the correlation is +0.278.

Interpretation: the *positive* correlation indicates that if the model under-predicts in one period it does the same the following time. This is because the adjustment to equilibrium is not achieved automatically, and therefore errors are followed by errors of the same sign.

Let's test formally for serial correlation:

Breusch-Godfrey Serial Correlation LM Test

Note that if the R-square is high in the next regression this means that the residuals in period t depends in a meaningful way of the residuals in previous periods and/or the dependant variable, and therefore we can reject the null of no serial correlation.

- **lm3bBG <- lm(bond$daaa ~ bond$dus3mt + e1 + e2); shccm(lm3bBG)**

As you can see the residuals are correlated with the residuals in the previous periods. The formal test indicates that we can reject the null hypothesis that the residuals are not correlated. Ways of dealing with autocorrelation in residual will be analysed in the next term.

Durbin-Watson Test for Autocorrelated Errors

- **?durbinWatsonTest**
- **durbinWatsonTest(lm4, max.lag=1, alternative="positive")**

This command computes the Durbin-Watson statistic to test for positive (alternative="positive"), first-order (max.lag=1) serial correlation in the disturbances when all the regressors are strictly exogenous. durbinWatsonTest values: if there were no autocorrelation, the value of the Durbin-Watson statistic would be around 2, and the closer the value is to 0 or to 4, the greater the autocorrelation.

In our case the lower and upper bound critical values at 5% are 1.86257 and 1.86925 respectively (http://www.stanford.edu/~clint/bench/dw05d.htm with T=600 and K=2). If the test statistic is below the lower bound critical value, this is evidence of positive autocorrelation. If it is between the lower and upper bound critical values, the test is inconclusive. If it is above the upper bound critical value, this is evidence of the error terms not being positively correlated. To test for negative autocorrelation, follow the same logic but use option alternative="negative" with (4 – DW statistic) as your test statistic.

In our case, the test statistic of 1.45 is lower than the lower bound critical value and so we can conclude that the model does suffer from autocorrelation in the residuals.

Exercise 4. Non-linearity in variables

In linear regression, in general the relationship between the response variable and the predictors is linear. If this assumption is violated, trying to fit a straight line to data that does not follow a straight line will be a mis-specification, and furthermore, may lead to violate the assumption of disturbances being iid.

We estimate: $y_i = \beta_0 + \beta_1 \cdot x_1 + ... + \beta_k \cdot x_k + \varepsilon_i$, to try for non-linearities, we could do:
$y_i = \beta_0 + \beta_1 \cdot x_1 + ... + \beta_k \cdot x_k + \gamma_{11} x_1^2 + \gamma_{22} x_2^2 + ... + \gamma_{kk} x_k^2 + \gamma_{12} x_1 x_2 + ... + \varepsilon_i$
A test of non-linearity would consist just on testing that each of the gammas is equal to 0.

Open the file nysevolume.csv and examine the data.

a. Fit a regression model of **volume** on **t** (a time trend)
b. Examine the residuals. Assess the linearity of the relation between volume and the time trend.
c. Generate a log transformation of the variable volume. Run the regression in a. with this new variable. What happens now with the residual?

d. Now run the following model (and analyse again the residuals):
$log(volume) = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot t^2 + \varepsilon_i$.

Answer:

a)

- **lm4a <- lm(volume ~ t, data=nyse); shccm(lm4a)**

b) To produce four summary-plots of the fitted model in 2 rows and 2 columns:

- **par(mfrow=c(2,2))**
- **plot(lm4a)**

Clearly it is impossible to try to fit a straight line if the original series is a curve. Clearly the residuals are not random values around 0. A scatter plot of volume and t will show us something similar.

- **plot(volume ~ t, data=nyse)**
- **abline(lm4a, col="red", lwd=2)**

c)

```
lm4c <- lm(log(volume) ~ t, data=nyse); shccm(lm4c)
par(mfrow=c(2,2))
plot(lm4c)
```

The residuals show that we still have problems.

d) Now run:

- **lm4d <- lm(log(volume) ~ t + I(t^2), data=nyse); shccm(lm4d)**

Now it is much better. Even though we can hardly say that the residuals are completely random. But we will leave the exercise here. (Note: if you include additional powers of ***t*** you will get better and better fit).

Exercise 5. Normality

Use bonds.csv. Normality of residuals is required for valid hypothesis testing, The normality assumption assures that the p-values for the t-tests and F-test will be valid. Normality is not required in order to obtain unbiased estimates of the regression coefficients.

a) Regress changes in AAA bond returns (daaa) on US Treasury Bill interest rates (dus3mt). Obtain the histogram of the residuals.
b) Analyse the Jarque-Bera test of normality results.

Answer:

a)

- **lm5a <- lm(daaa ~ dus3mt, data=bond); shccm(lm5a)**
- **hist(lm5a$res, breaks=30)**

Note that the histogram of the residuals has a "bell shape" (as the normal distribution has).

- **library(timeDate)**
- **skewness(lm5a$res, method="moment")**
- **kurtosis(lm5a$res, method="moment")**

But note that a normal distribution has Kurtosis = 3 and Skewness = 0, and here we have Kurtosis = 8. This means that too many residuals are concentrated very close or around zero.

b) Normal distribution has Kurtosis = 3 and Skewness = 0. The Jarque-Bera is a test statistic for testing whether the series is normally distributed. The test statistic measures the difference of the skewness (S) and kurtosis (K) of the series with those from the normal distribution.

$$JB = \frac{T-k}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$$

- k represents the number of estimated coefficients
- JB is distributed as a χ^2 with 2 degrees of freedom
- If JB > $\chi^2_{0.95}(2)$, which is at 0.05 significance, we reject the null hypothesis of a normal distribution

The test statistic is obtained by

- **library(tseries)**
- **jarque.bera.test(lm5a$res)**

The null hypothesis is that the residual series is normal. Because the p-value < 0.05 we reject the null of normality and therefore the residual series is not normal. We have to do something to solve this problem, but this is out of the scope of this term. Next term you learn some ways of solving this.

Exercise 6: Outliers

Unusual and Influential data

A single observation that is substantially different from all other observations can make a large difference in the results of your regression analysis. If a single observation (or small group of observations) substantially changes your results, you would want to know about this and investigate further. There are three ways that an observation can be unusual.

Outliers: In linear regression, an outlier is an observation with large residual. In other words, it is an observation whose dependent-variable value is unusual given its values on the predictor variables. An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.
Leverage: An observation with an extreme value on a predictor variable is called a point with high leverage. Leverage is a measure of how far an *independent variable* deviates from its mean. These leverage points can have an effect on the estimate of regression coefficients.

Influence: An observation is said to be influential if removing the observation substantially changes the estimate of coefficients. Influence can be thought of as the product of leverage and outlierness.

How can we identify outlying observations? Let's look at an example dataset called **crime**. Use crime.csv. The variables that we will work with are violent crimes per 100,000 people (**crime**), the percent of the population living in metropolitan areas (**pctmetro**), percent of population living under poverty line (**poverty**), and percent of population that are single parents (**single**). (This dataset appears in *Statistical Methods for Social Sciences, Third Edition* by Alan Agresti and Barbara Finlay).

a. A regression model for **crime** might have **pctmetro**, **poverty**, and **single** as independent variables. Run this regression and plot the residuals.

b. Identify what is the observation that is an outlier.

c. Try to solve this problem adding to the regression an impulse dummy variable.

Answer:

a)

```
lm6a <- lm(crime ~ pctmetro + poverty + single, data=crime); shccm(lm6a)
par(mfrow=c(2,2))
plot(lm6a)
```

Observations number 9, 25 and 51 seem to be outliers. R automatically identifies them in the plot, because the absolute values of these residuals (426, 523 and 412) are larger than 2 times the standard error of the regression (182.1).

b) We can examine the studentized residuals as a first means for identifying outliers. Use the **rstudent** command to generate studentized residuals. Studentized residuals are a type of standardized residual that can be used to identify outliers.

- `crime$rstudent <- rstudent(lm6a)`

Sort the data on the residuals and show the 10 largest and 10 smallest residuals along with the state id and state name.

```
crime <- crime[order(crime$rstudent), ]
head(crime)
tail(crime)
```

We should pay attention to studentized residuals that exceed +2 or -2, and get even more concerned about residuals that exceed +2.5 or -2.5 and even yet more concerned about residuals that exceed +3 or -3. These results show that DC and MS are the most worrisome observations followed by FL.

c) Generate an indicator variable for the outlier "dc"

- `crime$DC <- ifelse(crime$state=="dc", 1, 0)`

and include this indicator variable in the original regression:

- `lm6c <- lm(crime ~ pctmetro + poverty + single + DC, data=crime); shccm(lm6c)`

Note that this is equivalent to dropping the outlier:

- `lm6c <- lm(crime ~ pctmetro + poverty + single, data=subset(crime, state!="dc")) ; shccm(lm6c)`

Note that the variable DC is significant. Moreover, note how estimated parameters and the standard errors change (with respect to our first estimation). This is why it is very important to "neutralise" the effect of outliers.

The coefficient for **single** dropped from 132.4 to 89.4. Graphical assessment of this result:

```
plot(crime ~ single, data=crime, col="white"); text(x=crime$single, y=crime$crime, labels=crime$state)

m.pctmetro <- mean(crime$pctmetro); m.poverty <- mean(crime$poverty)
r.single <- seq(min(crime$single),max(crime$single),.1)

myReg <- function(x, model){
        coef(model)%*%c(1, m.pctmetro, m.poverty, x)
}

y <- sapply(r.single, myReg, model=lm6a); lines(x=range.single, y=y, col="red")
y <- sapply(r.single, myReg, model=lm6c); lines(x=range.single, y=y, col="blue")
legend("topleft", legend=c("with DC","without DC"), fill=c("red","blue"))
```