JBS Quantitative Research Methods Module MPO1

Michaelmas 2010
Thilo Klein
http://thiloklein.de

Computer Lab Session 4
The Generalized Linear Regression Model

Contents

Functional forms of regression - logarithms

Model	**Dependent variable**	**Independent variable**	**Algebraic interpretation of β_1**	**Conceptual interpretation of b_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	$\%\Delta Y = (100 \cdot \beta_1 \Delta X$	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%

<u>Note</u>: log (A) – log (B) = log (A/B) and this is approximately equal to the percent increase from B to A divided by 100, if the increase is low. Imagine for instance A=105, B=100. Do:

- **log(105/100)**

very close to 5%. Now A =110,

- **log(110/100)**

there is already an important difference with 10%. Finally, let A be 130.

- **log(130/100)**

Now it is quite different. Differences in logs are used as proxies for percent increases between variables, but sometimes these approximations are not that exact.

Examples:

<u>Level-level:</u> Example: exercise 3 in session 2. 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: $log(wage) = \beta_0 + \beta_1 \cdot education + u$.
In this case, $100 \cdot \beta_1$ captures 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? $log(profit\ level) = \beta_0 + \beta_1 \cdot capital + u$

<u>Double-log or Log-log:</u> In this case β_1 captures 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: $log(supply\ of\ labour) = \beta_0 + \beta_1 \cdot log(wages) + u$
This is the elasticity of labour supply.

$log(demand\ of\ cars) = \beta_0 + \beta_1 \cdot log(car\ prices) + \beta_2 \cdot log(household\ income) + u$
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 1. Non-linear models. Production function. Multiple hypotheses.

Dataset: production data for the year 1994; n=26; US firms in the sector of primary metal industries. (Gray, NBER, Technical Working Paper 205).

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)** Load usmetal.txt with **read.table** (R programming, page 1).
- **b)** Generate new variables as logs of the old variables. Inspect the variables. (summary and graph with histogram and scatter-plot)
- **c)** Using a double log specification, estimate a production function. (This is the Cobb-Douglas production function). Comment on the coefficients.
- **d)** Test the hypothesis that the coefficients are equal.
- **e)** **[Optional]** Test the hypothesis of constant returns to scale (CRS).
- **f)** Impose the restriction and re-estimate.

Help for c): Cobb-Douglas functions

The Cobb-Douglas function is defined as follows:

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

therefore:

$$log(Y_i) = \beta_1 + \beta_2 \cdot log(K_i) + \beta_3 \cdot log(L_i)$$ [1]

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

e) In order to impose the restriction take into account: $H0\!: \beta_2 + \beta_3 = 1$
Explanation: CRS is such that: f(λ·K, λ·L) = λ·f(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.

[1] Remember: $\ln(AB) = \ln(A) + \ln(B)$; $\log(A^{\alpha}) = \alpha \cdot \ln(A)$

Exercise 2. Model selection

How do growing weather and a wine's age influence a Bordeaux wine's price? Data: wineweather1.csv. It contains average 1989 prices for Bordeaux wines for the vintages from 1951 to 1989, together with data on conditions when each vintage was being grown. Forget about typical problems with time series variables. Variables in the file:

logprice: natural log of the price of Bordeaux wines relative to the price of the 1961 vintage.
degrees: average temperature in the growing season.
hrain: rainfall in the harvest season.
wrain: winter rainfall
time_sv: time from 1989 back to the wine's vintage year.

a) Regress log(price) on growing-season temperatures, harvest-season rainfall, off-season rainfall, and the age of a wine. Use the R^2 to compute the F-statistic to test the null hypothesis that none of the variables in the regression matter for the price of wine.
b) Test at the 5% significance level the null hypothesis that the intercept changes in the decades after the 50s with respect to the corresponding one in the 50s.
c) The regression used for b) should EITHER include a dummy variable for each decade and no constant OR a dummy for each of three decades and a constant. Explain.
d) Drop time_sv. Include lagged values (up to 2) for the average temperature in the growing season. Select the best model with BIC and AIC criteria.

```
➢ wine <- read.csv("wineweather1.csv", header=T)
```

a)

The sample is small, we have to rely on assumptions of homoscedasticity and normality of the disturbances.

```
➢ lm2a <- lm(logprice ~ degrees + hrain + wrain + time_sv, data=wine)
```

R performs the F-test for you and reports the test statistic and p-value of the test in the last line on the regression output. If you wish to perform the test manually, the instructions are below:

Remember: $F = \dfrac{R^2/(k\text{-}1)}{(1\text{-}R^2)/(n\text{-}k)}$

```
qf( p=0.95, df1=4, df2=22 ) # for the critical value.
paste("R^2 is", summary(lm2a)$r.squared)s
paste("df residuals = n-k =", 27-5)
paste("Restrictions=k–1=df of the model=", 5-1)
```

there are 2 restrictions.

```
paste("F-stat =", summary(lm2a)$r.squared / (5-1)
        / ((1- summary(lm2a)$r.squared) / (27-5))
```

or simply:

```
summary(lm2a)$fstatistic
```

F=[R^2/(k-1)]/ [(1-R^2)/(n-k)]=26.38>2.81 from an F(4,22) at 5% level of significance. Reject null that none of the variables matter for the price of wine.

b)

```
wine$vint
```

```
wine$sixties <- ifelse(wine$vint>1959 & wine$vint<1970, 1, 0)
wine$seventies <- ifelse(wine$vint>1969 & wine$vint<1980, 1, 0)
wine$eighties <- ifelse(wine$vint>1979, 1, 0)
lm2b <- lm(logprice ~ degrees + hrain + wrain + sixties + seventies + eighties,
        data=wine)
linearHypothesis(sixties seventies eighties)
```

We reject the null that all three coefficients are equal to 0. Since each of them represents the difference with the default categories (observations picked up in the 50s) we conclude that at least in one of the decades there was a significant difference with the intercept in the 50s.

c) The regression used for b) should EITHER include a dummy variable for each decade and no constant OR a dummy for each of three decades and a constant. To see this, realize that the intercept is obtained by adding up all dummies. This leads to problems of perfect multicollinearity in the model. R will, by default, drop one dummy and the coefficients of the remaining dummies are interpreted with respect to the dropped dummy (referred to as the reference category).

d)

Sort data by vintage if it isn't already

- `wine <- wine[order(wine$vint, decreasing=T), ]`
- `wine$deglag1 <- c(wine$degrees[-1], NA)`
- `wine$deglag2 <- c(wine$degrees[-c(1,2)], rep(NA,2))`

Type `wine[c(1:10, 30:38), ]` to see what the above command does.

```
lm2di <- lm(logprice ~ degrees + hrain + wrain + deglag1 + deglag2, data=wine)
lm2dii <- lm(logprice ~ degrees + hrain + wrain + deglag1, data=wine)
lm2diii <- lm(logprice ~ degrees + hrain + wrain, data=wine)

myIC <- function(model){
  print(model$call)
  print( paste("AIC:", AIC(model, k=2) ))                          # Akaike's An IC
  print( paste("BIC:", AIC(model, k=log(length(model$res))) )) # Bayes IC
  print( paste("R2 :", summary(model)$adj.r.squared ))          # adjusted R^2
 }
 myIC(lm2di); myIC(lm2dii); myIC(lm2diii)
```

R2-adjusted, BIC and AIC give different results. BIC is consistent, but we have a small sample so there's no clear cut solution for this problem. With the adjusted R2 we would have picked the model with no lags. With AIC we would have selected the model with 2 lags (the one with lowest AIC), with BIC the model without any lag.

Exercise 3. Heteroskedasticity.

a) Use the data in <u>hprice1.csv</u> 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?

a)

```
➢ lm3a <- lm(price ~ lotsize + sqrft + bdrms, data=house)
➢ summary(lm3a)
➢ shccm(lm3a)
```

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

$$
\begin{array}{llllll}
\text{price\_hat} = & -21.77 & + 0.00207\ \text{lotsize} & + 0.123\ \text{sqrft} & + 13.85\ \text{bdrms} \\
 & (29.48) & (0.00064) & (0.013) & (9.01) \\
 & [36.28] & [0.0012] & [0.017] & [8.28] \\
 & & N=88 & R^2=0.672 &
\end{array}
$$

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)

```
➢ lm3b <- lm(lprice ~ llotsize + lsqrft + bdrms, data=house)
➢ summary(lm3b); shccm(lm3b)
```

For the log-log model:

$$
\begin{array}{llllll}
\log(\text{price\_hat}) = & -1.30 & + 0.0168\ \log(\text{lotsize}) & + 0.700\ \log(\text{sqrft}) & + 0.037\ \text{bdrms} \\
 & (0.65) & (0.038) & (0.093) & (0.028) \\
 & [0.76] & [0.041] & [0.10] & [0.030] \\
 & & N=88 & R^2=0.643 &
\end{array}
$$

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$lm3b.sqres <- lm3b$residuals^2
```

```
lm3b.white.test <- lm(lm3b.sqres ~ llotsize*lsqrft*bdrms - llotsize:lsqrft:bdrms
+ I(llotsize^2) + I(lsqrft^2) + I(bdrms^2), data=house); shccm(lm3b.white.test)
T <- summary(lm3b.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 4. Autocorrelation (optional)

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

```
bond <- read.csv("bond_int_rates.csv", header=T)
```

Generate the variable to use to define the date:

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

regress changes in AAA bond returns on US Treasury Bill interest rates.

```
lm4 <- lm(daaa ~ dus3mt, data=bond); shccm(lm4)
```

Investigate serial autocorrelation in residuals. For this purpose, create and examine the residuals for this analysis, showing the residuals over time.

Some useful definitions:

Are the residuals distributed evenly across time?

```
e <- lm4$res
plot(e ~ bond$paneldate, type="l")

N <- length(e)
e1 <- c(NA, e[1:(N-1)])
plot(e ~ e1)
cor(e, e1, use="complete")
```

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.

This could have been done using the definition of the lag operator. Lag operator:

$$L \cdot y_t = y_{t-1}$$
$$L^2 \cdot y_t = L \cdot (L \cdot y_t) = L \cdot y_{t-1} = y_{t-2}$$
$$L^n \cdot y_t = y_{t-n}$$

Are the residuals independent over time?

```
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.

Linearity

In linear regression, the assumption is that 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 violate the assumption of disturbances being iid. We saw in the lectures RESET as a test to check for relevant omitted variables, it can also be used to test for non-linearities.

We estimate: $y_i = \beta_0 + \beta_1 \cdot x_1 + \ldots + \beta_k \cdot x_k + \varepsilon_i$, to try for non-linearities, we could do:

$$y_i = \beta_0 + \beta_1 \cdot x_1 + \ldots + \beta_k \cdot x_k + \gamma_{11} x_1^2 + \gamma_{22} x_2^2 + \ldots + \gamma_{kk} x_k^2 + \gamma_{12} x_1 x_2 + \ldots + \varepsilon_i$$

A test of non-linearity would consist just on testing that each of the gammas is equal to 0. RESET (regression specification error test) consist of doing something simpler:

$$y_i = \beta_0 + \beta_1 \cdot x_1 + \ldots + \beta_k \cdot x_k + \gamma_y \cdot \hat{y}_i^2 + \varepsilon_i.$$

Note, however, that $\hat{y}_i^2$ is stochastic (β is in it) and so gamma should only be valid for big samples. (Also higher order terms of $\hat{y}$ could be added). RESET test is, then, a misspecification test, but if the null is rejected it doesn’t tell us how to solve this problem.

Otherwise, we should see for each of the plots just a random scatter of points.

Exercise 5. Linearity

Open the nations.csv dataset and examine the data.

Fit a regression model of **birth** on **gnpcap** (GNP per capita) and **urban** (the proportion of urban population)

Collect the residuals

```
➢ e <- lm5$resid
```

Examine the scatter plot of the residuals against the different independent variables in the model. Use the help command to understand any of the commands below you are not familiar with.

```
➢ plot(e ~  lm5$model[,2]); lines(lowess(cbind(lm5$model[,2], e), f=1), col=2)
➢ plot(e ~ lm5$model[,3]); lines(lowess(cbind(lm5$model[,3], e), f=1), col=2)
```

Notice the clear deviation from linearity with respect to **gnpcap**

The **resettest** command performs a test of regression model specification. It performs a regression specification error test (RESET) for omitted variables. It creates new variables based on the predictors and refits the model using those new variables to see if any of them would be significant. Execute an **resettest** and assess results.

Let's look at the relationship between these variables more closely.

```
➢ plot(subset(nations, select=c("birth","gnpcap","urban")))
```

Assess the linearity of the relation between birth rate and per capita gross national product and between birth rate and urban population.

It is possible that **gnpcap** is very skewed in its distribution. This may affect the linearity in the relationship.

Examine the variable **gnpcap**, using **summary**. Examine its distribution. **density** gives a kernel density estimate. It can be thought of as a histogram with narrow bins and moving average. Kernel density is the smoothed out contribution of each observed data point over a local neighbourhood of that data point.

```
plot(density(nations$gnpcap))
grid.x <- seq(-10000, 20000, 1)
grid.y <- dnorm(grid.x, sd=sd(nations$gnpcap))
lines(grid.x, grid.y, col="blue", lwd=2)
legend("topright", legend=c("Density of gnpcap","Normal density"),
fill=c("green","blue"))
```

The distribution of **gnpcap** is very skewed. This suggests that some transformation of the variable may be necessary. A commonly used transformation is log transformation.

```
➢ nations$lgnp <- log(nations$gnpcap)
```

Does the transformation help reduce the skewness of the variable?

Fit a regression model replacing **gnpcap** by **lgnp**, and examine linearity

```
lm5b <- lm(birth ~ lgnp + urban, data=nations)
```

Assess the deviation from linearity. Execute an **resettest** and assess results.

Normality

Normality of residuals is only 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. OLS regression merely requires that the residuals (errors) be identically and independently distributed. There is no assumption or requirement that the predictor variables be normally distributed. After regression analysis, we can use the **model$residuals** command to create residuals, and then use commands such as **density**, **qnorm** and **pnorm** to check the normality of the residuals.

Exercise 6. Normality

1. Use the earnings **eaef21** dataset from session 3 (see Annex 1 for variable descriptions) and regress **EARNINGS** on **S** and **ASVABC**
2. Use the **model$res** command to generate residuals.
3. Use the **density** command to produce a kernel density plot. Overlay the plot with a normal density.

The **qqnorm** command graphs a standardized normal probability plot. In a normal probability plot, the data are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line. Departures from this straight line indicate departures from normality. The procedure **qqnorm** does the following:

a) The data are arranged from smallest to largest.
b) The percentile of each data value is determined.

c) From these percentiles, normal calculations are done to determine their corresponding z-scores.

d) Each z-score is plotted against its corresponding data value

qqnorm is sensitive to non-normality in the middle range of data

4. Execute the command **qqnorm** and assess normality of residuals.
5. Examine the results from **density** and **qqnorm** applied to the residuals after a semi-log (log-level) regression

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 a 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.

We will focus on outliers in this module. How can we identify outlying observations? Let's look at an example dataset called **crime**. This dataset appears in *Statistical Methods for Social Sciences, Third Edition* by Alan Agresti and Barbara Finlay (Prentice Hall, 1997).

The variables are state name (**state**), violent crimes per 100,000 people (**crime**), murders per 1,000,000 (**murder**), the percent of the population living in metropolitan areas (**pctmetro**), the percent of the population that is white (**pctwhite**), percent of population with a high school education or above (**pcths**), percent of population living under poverty line (**poverty**), and percent of population that are single parents (**single**).

Exercise 7: Outliers

Load the dataset crime.csv and inspect. A regression model for **crime** might have **pctmetro**, **poverty**, and **single** as independent variables.

Look at the scatter plots of crime against each of the predictor variables before the regression analysis so we will have some ideas about potential problems.

```
➢ plot(subset(crime, select=c("crime", "pctmetro", "poverty", "single")))
```

Notice the data point that is far away from the rest of the data points. Generate individual graphs of **crime** with **pctmetro** and **poverty** and **single** to get a better view of these scatter plots. Use the 'add text' option to label each marker with the state name to identify outlying states.

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

Which is the state that requires extra attention?

Fit a regression model for **crime** with **pctmetro**, **poverty**, and **single** as independent variables.

```
lm7 <- lm(crime ~ pctmetro + poverty + single, data=crime); shccm(lm7)
```

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(lm7)
```

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(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.

Show all of the variables in our regression where the studentized residual exceeds +2 or -2, i.e., where the absolute value of the residual exceeds 2.

```
subset(crime, abs(rstudent) > 2)
```

We see the data for the three potential outliers we identified, namely Florida, Mississippi and Washington D.C.

Now, let's run the analysis omitting DC by including **data=subset(crime, state != "dc")** on the **lm** command (here **!=** stands for "not equal to").

```
lm7b <- lm(crime ~ pctmetro + poverty + single,
data=subset(crime, state!="dc")); shccm(lm7b)
```

What has happened to the results? The coefficient for **single** dropped from 132.4 to 89.4. 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=lm7)
lines(x=range.single, y=y, col="red", lwd=2)

y <- sapply(r.single, myReg, model=lm7b)
lines(x=range.single, y=y, col="blue", lwd=2)

legend("topleft", legend=c("with DC","without DC"), fill=c("red","blue"))
```

Annex 1. Variables in eaef.csv

Personal variables

AGE	C	age in 1994
S	C	years of schooling (highest grade completed as of 1994)
MALE	D	sex of respondent (1 if male, 0 if female)
ETHBLACK	D	ethnicity: black
ETHHISP	D	hispanic
		score on a component of the *ASVAB* battery (scaled with mean 50, standard deviation 10)
ASVAB2	C	arithmetic reasoning
ASVAB3	C	word knowledge
ASVAB4	C	paragraph comprehension
ASVABC	C	composite of *ASVAB2* (with double weight), *ASVAB3* and *ASVAB4*
CHILDREN	C	number of children in the household
YOUNGEST	C	age of youngest child
CHILDL06	C	presence of a child age < 6 in the household
CHILDL16	C	presence of a child age < 16, but no child age < 6, in the household
MARISTAT	T	marital status, coded as: 1 never married; 2 married, spouse present; 3 other
MARRIED	D	married (*MARISTAT*=2)

Work-related variables

EARNINGS	C	current hourly earnings in $ reported at 1994 interview
WORKING	D	working (has recorded earnings)
EMPSTAT	T	employment status, coded as: 1 employed; 2 unemployed; 3 out of the labor force