Lunds University
Mathematical Sciences/Mathematical Statistics 2017

Computer lab 6

Multivariate regression is useful when investigating how a response may be modelled as a function of several different factors. In this lab we are interested in a linear multivariate regression. We will study two different data sets, one with three regressors and one with six regressors. At the end we will construct the linear regression model from the factorial design.

1 Introductory example

We start out by considering a baby example with just three regressors. The data set considered is murders per annum in a town, (Y), as a function of the total number inhabitants, (X_1), poverty, (X_2), and unemployment, (X_3). The data set consists of 20 measurements. We seek a model of the form

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 \tag{1}$$

Start by installing the R-packages *car*, *MASS*, and *gdata*. This is done under *Tools->Install Packages*.

Incorporate the libraries you just installed by typing

```
library(gdata)
library(car)
library(MASS)
```

The data is stored in the file murder.dat. Along with the .dat file there is a descriptive file called murder.txt where you can get some background concerning the data. Import the data into your script by typing:

```
mydata <- read.table("murder_01.dat",header=FALSE)
```

Next we wish to name the data and attach it,

```
names(mydata)<-c("I","X1","X2","X3","Y")
attach(mydata)
```

The first line in the data set is just an index and we will not use it in our analysis.

Start by taking a first look at the data using the commands

```
plot(X1,Y)
plot(X2,Y)
plot(X3,Y)
```

Which regressors seem to have a significant effect on the number of murders?

We continue by forming the linear regression model

```
mymodel <- lm(Y ~ X1+X2+X3, data = mydata)
summary(mymodel) # show results
```

What can you say about the model? Which regressors seem to have the largest impact on the number of murders?

Take a further look on the estimated regression coefficients and residuals using the commands

```
coefficients(mymodel) # model coefficients
confint(mymodel, level=0.95) # CIs for model parameters
fitted(mymodel) # predicted values
residuals(mymodel) # residuals
plot(residuals(mymodel))
```

What can you say about the regression coefficients? How do the residuals look?

Also, take at the covariance matrix.

```
vcov(mymodel) # covariance matrix for model parameters
```

What conclusions can you draw?

Next, we wish to try a simpler model, excluding X_1,

```
# compare models
mymodel2 <- lm(Y~ X1 + X2 + X3, data=mydata)
mymodel3 <- lm(Y ~ X2 + X3, data=mydata)
anova(mymodel3, mymodel2)
detach(mydata)
```

You can also use AIC criterion try

```
    step <-stepAIC(mymodel3, direction = "both")
```

What can you say about the simplified model? Which model would you choose?

2 Example with many regressors

Our next example of multivariate regression comes from the economic sciences. We consider the number of people employed as a function of other economic and social factors. The data looks as follows

- I, the index;
- X_1, the percentage price deflation;
- X_2, the GNP in millions of dollars;
- X_3, the number of unemployed in thousands;
- X_4, the number of people employed by the military;
- X_5, the number of people over 14;
- X_6, the year
- Y, the number of people employed.

Once again we want a model of the form

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta_4 \cdot X_4 + \beta_5 \cdot X_5 + \beta_6 \cdot X_6 \tag{2}$$

We repeat the analysis steps from the previous section

```
mydata <- read.table("economic_01.dat",header=FALSE)
names(mydata)<-c("I","X0","X1","X2","X3", "X4", "X5", "X6", "Y")
attach(mydata)
```

Where X_0 is just a vector of 1s.

Continue by forming the model and looking at the residuals

```
mymodel <- lm(Y ~ X1+X2+X3 +X4 +X5 +X6, data = mydata)
summary(mymodel) # show results
coefficients(mymodel) # model coefficients
confint(mymodel, level=0.95) # CIs for model parameters
fitted(mymodel) # predicted values
residuals(mymodel) # residuals
plot(residuals(mymodel))
vcov(mymodel) # covariance matrix for model parameters
```

Which regressors seem to be significant?

Next we wish to do variable selection, i.e. choose which regressors should be included in the model. This can be done in many different ways; forward, backward or both. In R we use the function *stepAIC()* which uses Akaikes information criterion to determine which model should be chosen. Do forward, backward and forward-backward model selection on our data set What can you say about the final model? Is the suggested final model the same in all methods or do they differ?

3 Linear Regression and Factorial Design

In 2^k factorial design the results can be expressed in terms of a linear regression in the following way. For example, let us assume there are two quantitative variables in 2^2 factorial design. The regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon,$$

where β_0 is the half of the grand average of all observations, all other coefficients are one-half the corresponding factor effect estimates and the regressors x_i are equal to

$$x_i = \frac{X_i - (X_{i,low} + X_{i,high})/2}{(X_{i,high} - X_{i,low})/2}.$$

For more information, look at Chapter 6 in [2]. Construct the linear model for the exercise from the Computer Lab 4.

Referenser

[1] Box, G. E. P., Hunter, J. S. and Hunter, W. G. (2005). Statistics for Experiments John Wiley & Sons.

[2] Montgomery, D. C. (2009). Statistics for Experiments John Wiley & Sons.