Statistics with R – Intermediate Level

Section 3

Predictive Techniques

Lesson 17 - Multiple Linear Regression – Basics

```
stud = read.csv("students.csv")

View(stud)

#########
### how to perform the multiple regression analysis
#########

#########
### Basic assumptions:

# the relationship between the variables is liniar
# the variables do not present important outliers*
# there is independence of errors*
# there is no important multicollinearity*
# there is homoskedasticity*
# the residuals are normally distributed*

### we will check only the assumptions marked with an
asterisk (*)
##########
```

```
### dependent variable: test score (score)
### explainers: iq and hours of study (hours)

### how to get the goodness-of-fit (R squared)
### the ANOVA table and the regression coefficients

fit <- lm(score~iq+hours, data = stud)
summary(fit)
```

Lesson 18 - Multiple Linear Regression - Testing Assumptions

```
stud = read.csv("students.csv")

View(stud)

#########
### the multiple regression analysis - checking the
assumptions
#########

#########
### Basic assumptions:

# the relationship between the variables is liniar
# the variables do not present important outliers*
# there is independence of errors*
# there is no important multicollinearity*
# there is homoskedasticity*
# the residuals are normally distributed*

### we will check only the assumptions marked with an
asterisk (*)
##########

### run the regression again

fit <- lm(score~iq+hours, data = stud)

###############

### to detect the outliers, we get the standardized
residuals
```

```
### and check whether there are values greater than 3

res <- residuals(fit)

zres <- scale(res)

View(zres)

#################

### to check the independence of errors we use the Durbin-
Watson test

### we can find it in the car package

require(car)

durbinWatsonTest(fit)

### alternatively, we can find the Durbin-Watson test
### in the lmtest package

require(lmtest)

dwtest(fit)

################

### to check for multicollinearity we compute the VIF
### (variance inflation factor)

### first we create a new data frame with the independents
only

x <- data.frame(stud$iq, stud$hours)

View(x)

## load the usdm package

require(usdm)

### use the vif function
```

```
vif(x)

###################

### to check for homoskedasticity, we must plot the
### residuals against the fitted (predicted) test score values

### we will use ggplot for that

require(ggplot2)

### we already have the residuals stored in the variable res

### now we get the predicted values of the response variable

pred <- fitted(fit)

### create a new data frame with the residuals and the fitted values

dat <- data.frame(pred, res)

View(dat)

### build the chart

ggplot()+geom_point(data=dat, aes(x=res, y=pred))

##################

### finally, we check for the normality of the residuals

shapiro.test(res)
```

Lesson 19 - Multiple Regression with Dummy Variables

```
stud = read.csv("students.csv")

View(stud)
```

```
#########
### how to perform the multiple regression analysis with
DUMMY variables
#########

### dependent variable: test score (score)
### explainers: iq, hours of study (hours) and gender

### the procedure is the same

fit <- lm(score~iq+hours+gender, data = stud)
summary(fit)
```

Lesson 20 - Sequential Regression

```
stud = read.csv("students.csv")

View(stud)

#########
### how to perform the hierarchical regression analysis
#########

### dependent variable: test score (score)
### explainers: iq, hours of study (hours) and gender

### the independent variables will be introduced in blocks

### block 1: iq
### block 2: iq and hours of study
### block 3: iq, hours of study (hours) and gender

### we run a separate regression for each block
### using the lm function

fit1 <- lm(score~iq, data = stud)

fit2 <- lm(score~iq+hours, data = stud)

fit3 <- lm(score~iq+hours+gender, data = stud)
```

```
### to get the results of each regression analysis
separately
### we use the summary function

summary(fit1)

summary(fit2)

summary(fit3)

### to get the ANOVA table for the whole model
### we run the following

anova(fit1, fit2, fit3)

### the last two columns tell us whether the model improved
by adding new variables
### i.e. whether the R square increases are statistically
significant
```