QMB Exercise 3 - Regression for Wal Mart Dataset

Karin Gryzlak & Dino Nienhold

Thursday, May 27, 2015

Introduction

The following report is based on the QMB Exercise 3 - Multiple Regression. The task description pdf file is Exercise-Walmart-20150520.pdf.

Requirements

Please make sure that you the following packages loaded in your workspace.

```
library("dplyr")
library("ggplot2")
library("leaps")
library("car")
library("QuantPsyc")
```

Data Set

Please make sure you have the file housingrents.csv in the subdirectoy Data in your workspace.

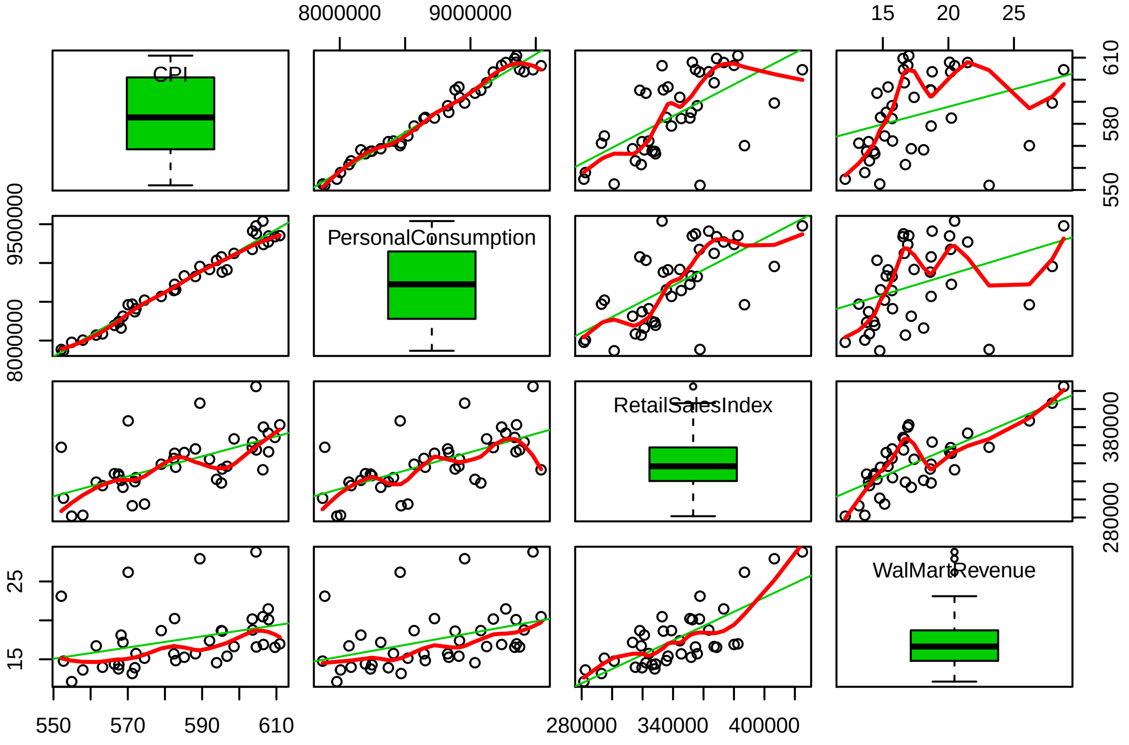
```
walmart <- read.csv("./data/Wal-Mart_revenue.csv",sep=";")
```

Problem 1

Task 1

Create a scatterplot matrix with the variables in the data set. Comment!

```
scatterplotMatrix(~CPI+PersonalConsumption+RetailSalesIndex+WalMartRevenue,
  reg.line=lm, smooth=TRUE, spread=FALSE, span=0.5, id.n=0, diagonal = 'boxplot', data=walmart)
```



CPI and PersonalConsumption have a high positive correlation, tis means the variables have a strong relationship between each other and an almost normal distribution. RetailStatesIndex and WalMartRevenue have a low positive correlation. Both have outliers which are shown in the upper part of the boxplot. The other varialbes have no or a low positive correlation.

Task 2

Fit a regression with response WalMartRevenue and explanatory variables RetailSalesIndex, PersonalConsumption and CPI.

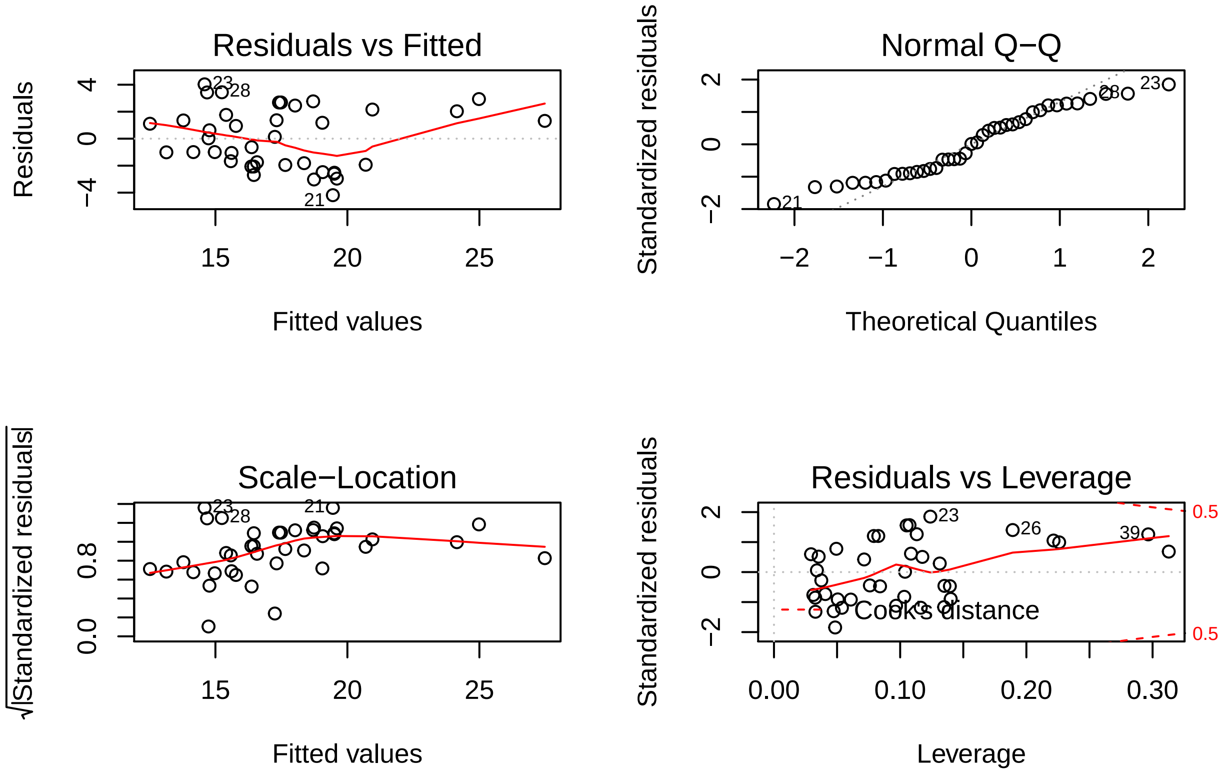
```
walmartLm <- lm(WalMartRevenue~CPI+PersonalConsumption+RetailSalesIndex,
                data=walmart)
summary(walmartLm)
```

```
## 
## Call:
## lm(formula = WalMartRevenue ~ CPI + PersonalConsumption + RetailSalesIndex, 
##     data = walmart)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.1894 -1.9418  0.0239  1.8960  4.0275 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          8.701e+01  3.360e+01   2.590   0.0139 *  
## CPI                 -3.448e-01  1.203e-01  -2.865   0.0070 ** 
## PersonalConsumption  1.108e-05  4.403e-06   2.518   0.0165 *  
## RetailSalesIndex     1.032e-04  1.546e-05   6.674 1.01e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.327 on 35 degrees of freedom
## Multiple R-squared:  0.6665, Adjusted R-squared:  0.638 
## F-statistic: 23.32 on 3 and 35 DF,  p-value: 1.794e-08
```

Task 3

Comment the diagnostic plots and identify outliers and leverage points!

```r
par(mfrow = c(2, 2))
plot(walmartLm)
```



On the plot residuals vs. fitted values there are at least 3 outliers (23 / 28 / 21) which should be checked in detail. The data has to be checked to see what causes the outliers. The Norm Q-Q plots shows an almost normal distribution, there are 3 outliers (23 / 28 / 21) at the bottom and top of the distriubution. There are leverage points like 23, 26 and 39 which can have a negative influence on the regression and "attract" the regression line to them.

Task 4

Does it seem that Wal-Marts's revenue is closely related to the general state of the economy? Yes, the F statistic and Rˆ2 is rather large. The p-value for the F-statistic allows to reject that $H_0 = 0$.

Task 5

Calculate the standardised regression coefficients for the three explanatory variables and discuss the relevance of the variables.

```
lm.beta(walmartLm)
```

```
##                 CPI PersonalConsumption    RetailSalesIndex
##          -1.6192494           1.4422733           0.8505235
```

An increase in the consumer price index has a negative impact on Wal-Mart's revenue. Both the personal consumption and retail sales index increase have a positive impact.

Problem 2

In Task 2 we continue the multiple regression but introduce an indicator variable December.

1.

Calculate the regression model with the four explanatory variables RetailSalesIndex, PersonalConsumption, CPI and December. Discuss the individual coefficients including their significance and explain what the coefficient for December means.

```
walmartLm <- lm(WalMartRevenue~RetailSalesIndex+PersonalConsumption+CPI+December,
                data = walmart)
summary(walmartLm)
```

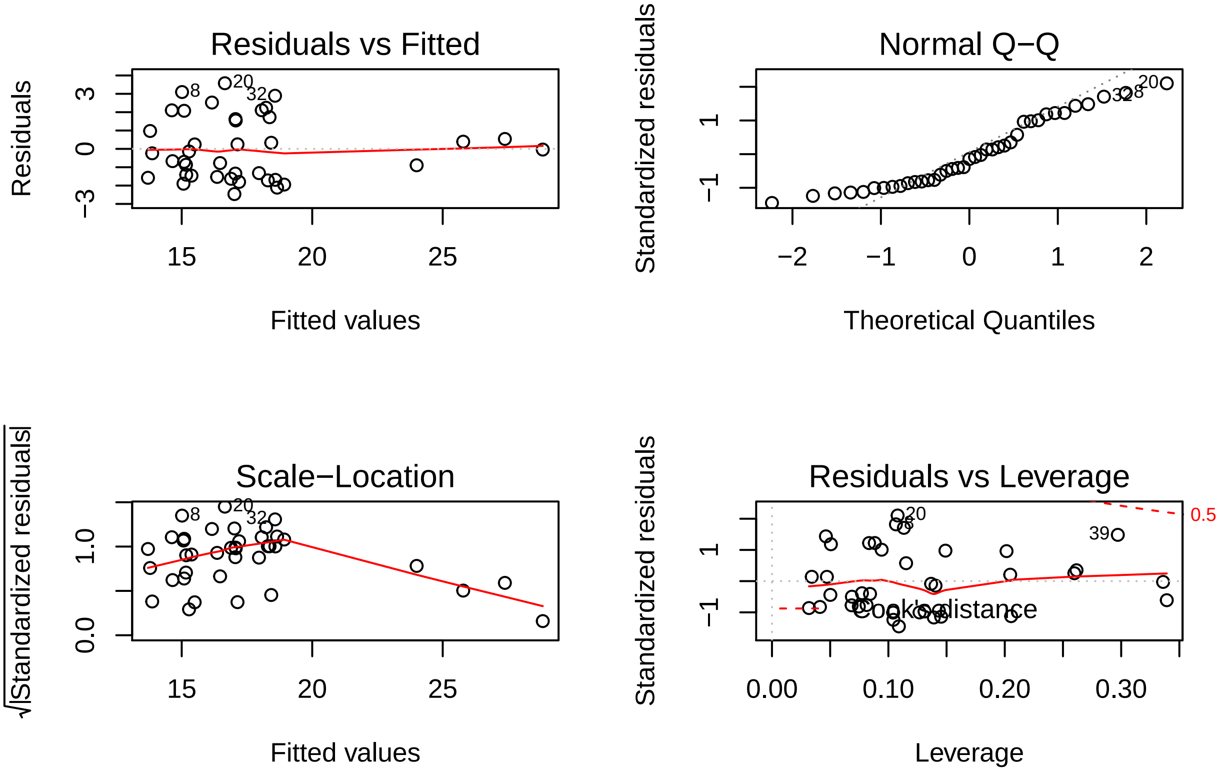
```
##
## Call:
## lm(formula = WalMartRevenue ~ RetailSalesIndex + PersonalConsumption +
##     CPI + December, data = walmart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4639 -1.4941 -0.2417  1.5851  3.5764
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -2.094e+01  3.395e+01  -0.617    0.542
## RetailSalesIndex     1.471e-05  2.152e-05   0.684    0.499
## PersonalConsumption  1.145e-06  3.956e-06   0.289    0.774
## CPI                  3.851e-02  1.212e-01   0.318    0.753
## December             9.385e+00  1.898e+00   4.944 2.03e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.801 on 34 degrees of freedom
## Multiple R-squared:  0.806,  Adjusted R-squared:  0.7832
## F-statistic: 35.32 on 4 and 34 DF,  p-value: 1.146e-11
```

The coefficients for RetailSalesIndex, PersonalConsumption and CPI are in the Range of 0.00001 to 0.03 which means that a small change in these indexes have an impact on Wal Marts sales revenue. However their significance is far above 0.05, which means that the coefficients value for this sample could be due to chance. The indicator variable December adds 9.39 to the Intercept, which means that the revenue is nearly 10 units higher in December than for non December month. The adjusted Rˆ2 is 0.7832, which means that the model explains the variance to a large degree.

2.

Check the diagnostic plots. Comments?

```r
par(mfrow = c(2, 2))
plot(walmartLm)
```



The Normal Q-Q plot shows that the distribution is not normal distributed and has some outliers. The residual vs. fitted plot are sparse on the right bottom. Additionally outliers 23, 28 and 21 pull the fitted line up respectively down, which means they have a high leverage. In the residuals vs.leverage plot the points 23, 26 and 39 are shown again. These outliers would have to be checked and dealt with.

3.

Does it seem that Wal-Marts's revenue is closely related to the general state of the economy? The general state of the economy here is represented by the variables in the regression equation?

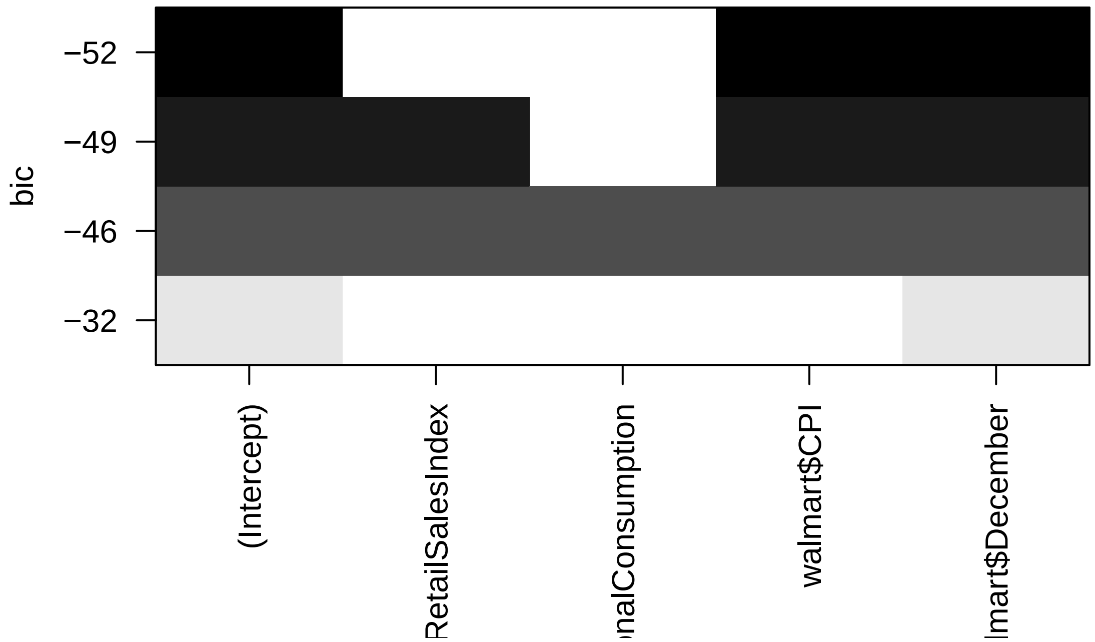
Yes, the F statistic and Rˆ2 is rather large. The p-value for the F-statistic allows to reject that $\mathsf{H}_0 = 0$.

Problem 3

1.

*Use the full model with all variables to find a best subset model with the BIC criterion (Use Models/subset model selection . . .).

```
le <- regsubsets(walmart$WalMartRevenue~walmart$RetailSalesIndex+walmart
                  $PersonalConsumption+walmart$CPI+walmart$December,data=walmart)
plot(le,scale="bic")
```



The model with the coeficient CPI and December do have the smallest BCI value and is therefore the best suited model measured with the Bayesian Information Criteria.

2.

Drop the variables RetailSalesIndex and PersonalConsumption from the regression, i.e. recalculate the regression with just CPI and December as explanatory variables.

```
walmartLm <- lm(WalMartRevenue~CPI+December, data = walmart)
summary(walmartLm)
```

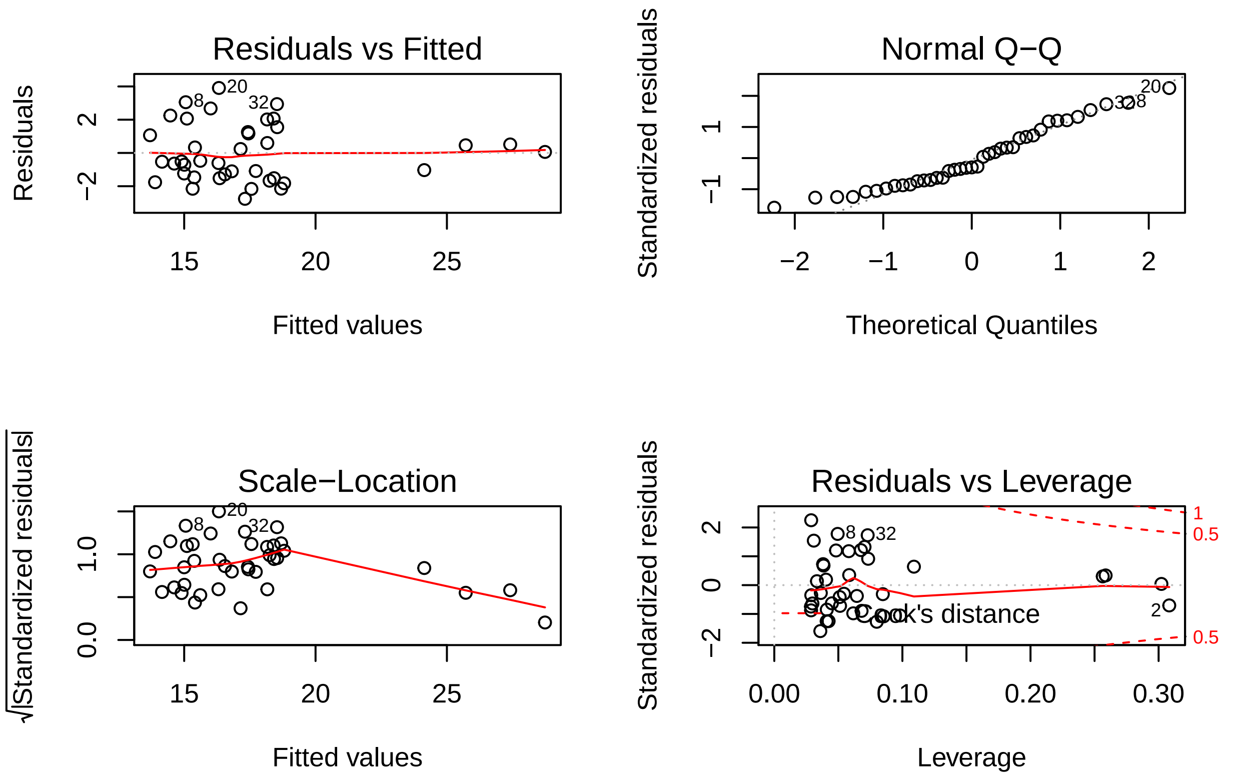
```
##
## Call:
## lm(formula = WalMartRevenue ~ CPI + December, data = walmart)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7586 -1.3798 -0.5137  1.2146  3.9075
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -34.77553    9.23825  -3.764 0.000596 ***
## CPI           0.08771    0.01580   5.550 2.77e-06 ***
## December     10.49054    0.93367  11.236 2.53e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.762 on 36 degrees of freedom
## Multiple R-squared:  0.8033, Adjusted R-squared:  0.7924
## F-statistic: 73.53 on 2 and 36 DF,  p-value: 1.937e-13
```

3.

Check the diagnostic plots. Comments?

```r
par(mfrow = c(2, 2))
plot(walmartLm)
```



The plots do show a much better fit of residuals to fitted respectively leverage. Additionally the model follows much more a normal distribution than the previous model.

4.

Does it seem that Wal-Marts's revenue is closely related to the general state of the economy? Yes, compared to the previous models this model coefficients show high significance. The Rˆ2 value is very high, meaning the model explains the variability ad the F-Statistic is large and also highly significant.

5.

*Compare this last model with two explanatory variables with the full model containing four explanatory variables with an F-test.

```r
walmartBestLm <- lm(WalMartRevenue~CPI+December, data = walmart)
walmartFullLm <- lm(WalMartRevenue~RetailSalesIndex+PersonalConsumption+
                      CPI+December, data = walmart)
anova(walmartBestLm,walmartFullLm)
```

```
## Analysis of Variance Table
##
## Model 1: WalMartRevenue ~ CPI + December
## Model 2: WalMartRevenue ~ RetailSalesIndex + PersonalConsumption + CPI +
##     December
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     36 111.75
## 2     34 110.22  2    1.5264 0.2354 0.7915
```

Given the F value of 0.2354 and the high p-value of 0.7915. We cannot reject the H_0 so that the full model with the variables RetailSalesIndex and PersonalConsumption do contribute additional information.

6.

Use the Durbin-Watson test to see whether the residuals exhibit autocorrelation.

```r
durbinWatsonTest(walmartBestLm)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.4395516      2.830718   0.014
##  Alternative hypothesis: rho != 0
```

Given that the D-W Statistic is outside the [1.5-2.5] and the p-value is < 0.05, we fail to reject H_0 that the residuals are not autocorrelated. Therefore a time series model might be a better fit.