Task p. 51-1 / Anwendung S. 51-1

FK

automatic

Working directory

```
> setwd("D:/kronthafranz/Documents/01Lehre/06Quantitative Forschungsmethoden
dt en")
```

Load data

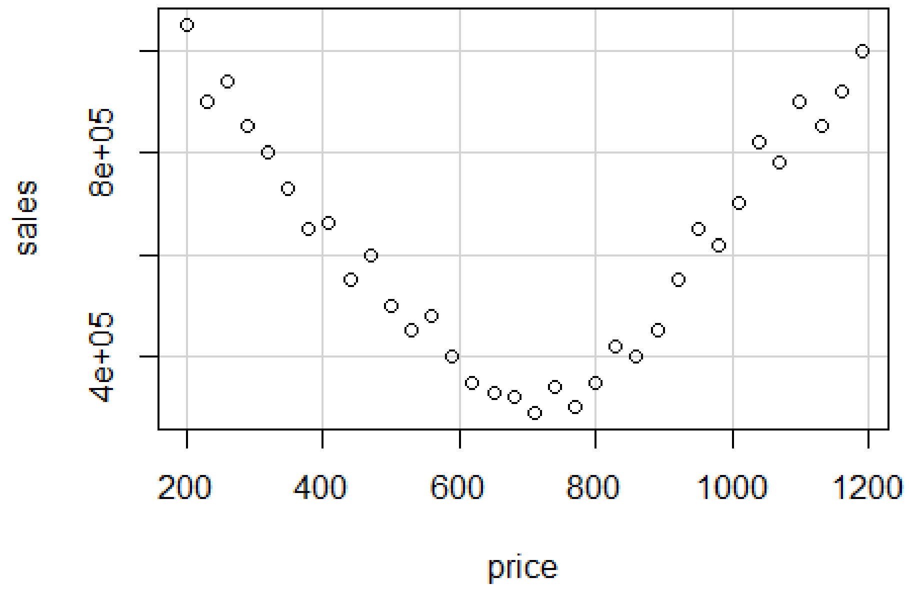
```
> load("D:/kronthafranz/Documents/01Lehre/06Quantitative Forschungsmethoden
dt en/07Regression Diagnostic/smartphone.RData")
```

Descriptive statistics

```
> summary(smartphone)

     price            sales        
 Min.   : 200.0   Min.   : 290000  
 1st Qu.: 447.5   1st Qu.: 405000  
 Median : 695.0   Median : 610000  
 Mean   : 695.0   Mean   : 614706  
 3rd Qu.: 942.5   3rd Qu.: 815000  
 Max.   :1190.0   Max.   :1050000  
```

Scatterplot

```
> scatterplot(sales~price, reg.line=FALSE, smooth=FALSE, spread=FALSE,
+   boxplots=FALSE, span=0.5, ellipse=FALSE, levels=c(.5, .9),
data=smartphone)
```



--> clearly a non-linear relationship, u-shaped

--> modelling a u-shaped relationship: y-estimated=b0+b1*X+b2*X^2

Correlation coefficients

```
> cor(smartphone[,c("price","sales")], use="complete")

            price       sales
price  1.00000000 -0.01201927
sales -0.01201927  1.00000000
```

Generate squared independent variable

```
> smartphone$price_sq <- with(smartphone, price * price)
```

Estimate model

```
> RegModel.1 <- lm(sales~price+price_sq, data=smartphone)
> summary(RegModel.1)


Call:
lm(formula = sales ~ price + price_sq, data = smartphone)
```

```
Residuals:
   Min     1Q Median     3Q    Max 
-88928 -43648    940  28233 116014 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.753e+06  5.378e+04   32.59   <2e-16 ***
price       -3.977e+03  1.716e+02  -23.17   <2e-16 ***
price_sq     2.854e+00  1.213e-01   23.52   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54740 on 31 degrees of freedom
Multiple R-squared:  0.947, Adjusted R-squared:  0.9435 
F-statistic: 276.7 on 2 and 31 DF,  p-value: < 2.2e-16
```

--> Model is significant

--> R2 is 94.7% explained

--> Independent variable and squared independent variable are significant

--> Sign of coefficient of independent variable is negative

--> Sign of coefficient of squared independent variable is positive

--> Because of signs and significance we can conclude that there is a u-shaped relationship

Evaluate GM assumptions

Add regression statistics

```
> smartphone<- within(smartphone, {
+   fitted.RegModel.1 <- fitted(RegModel.1)
+   residuals.RegModel.1 <- residuals(RegModel.1)
+   rstudent.RegModel.1 <- rstudent(RegModel.1)
+   hatvalues.RegModel.1 <- hatvalues(RegModel.1)
+   cooks.distance.RegModel.1 <- cooks.distance(RegModel.1)
+   obsNumber <- 1:nrow(smartphone)
+ })
```

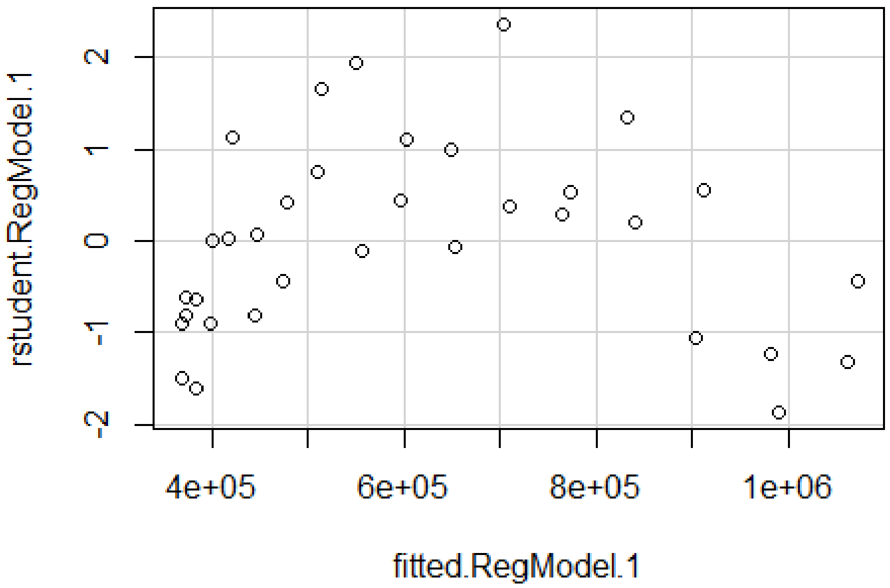
GM1: Linearity and complete specification

Only one independent variable: model fully specified?

Non-linearity is already considered

GM2: Expected value = 0

```
> scatterplot(rstudent.RegModel.1~fitted.RegModel.1, reg.line=FALSE,
+   smooth=FALSE, spread=FALSE, boxplots=FALSE, span=0.5, ellipse=FALSE,
+   levels=c(.5, .9), data=smartphone)
```

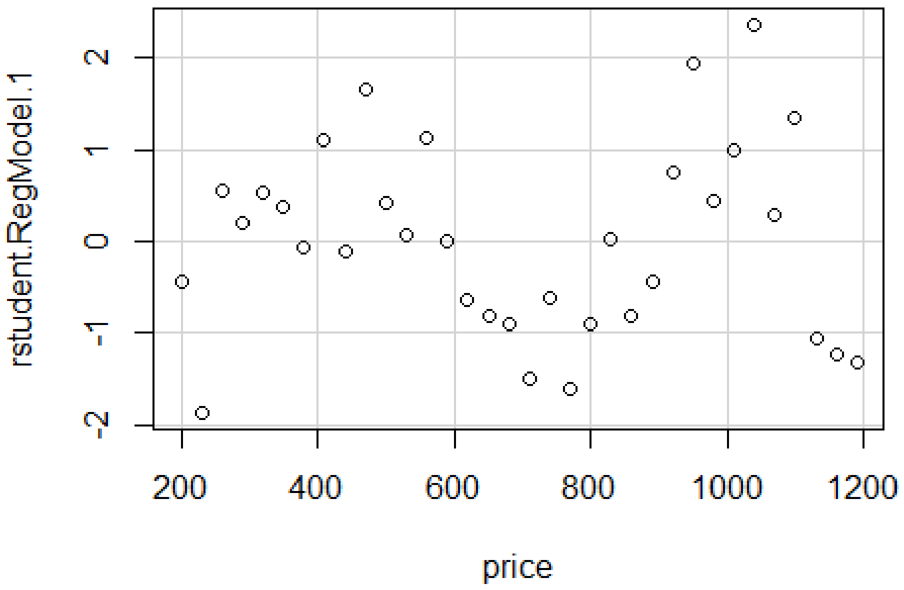


Is violated

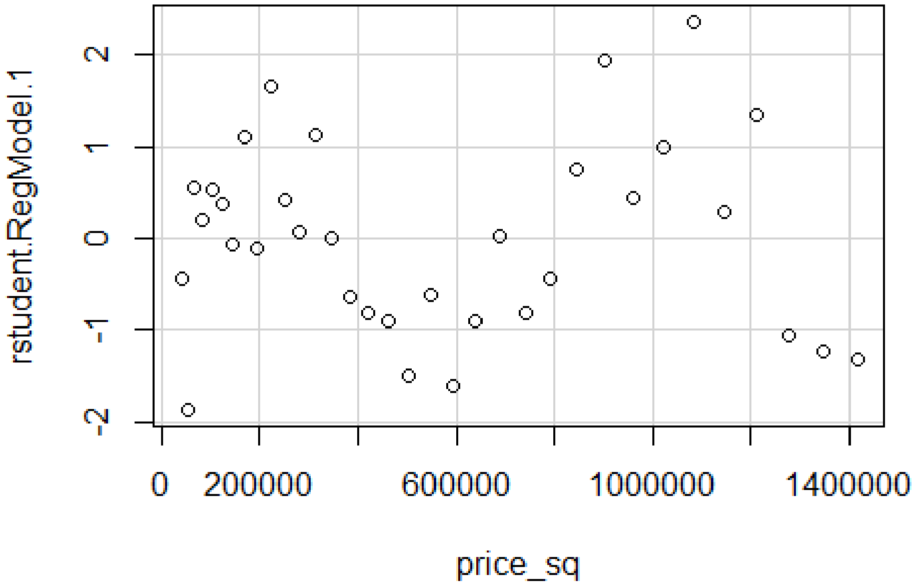
Important independent variable missing?

GM3: Error term correlated with independent variables?

```
> scatterplot(rstudent.RegModel.1~price, reg.line=FALSE, smooth=FALSE,
+   spread=FALSE, boxplots=FALSE, span=0.5, ellipse=FALSE, levels=c(.5, .9),
+   data=smartphone)
```



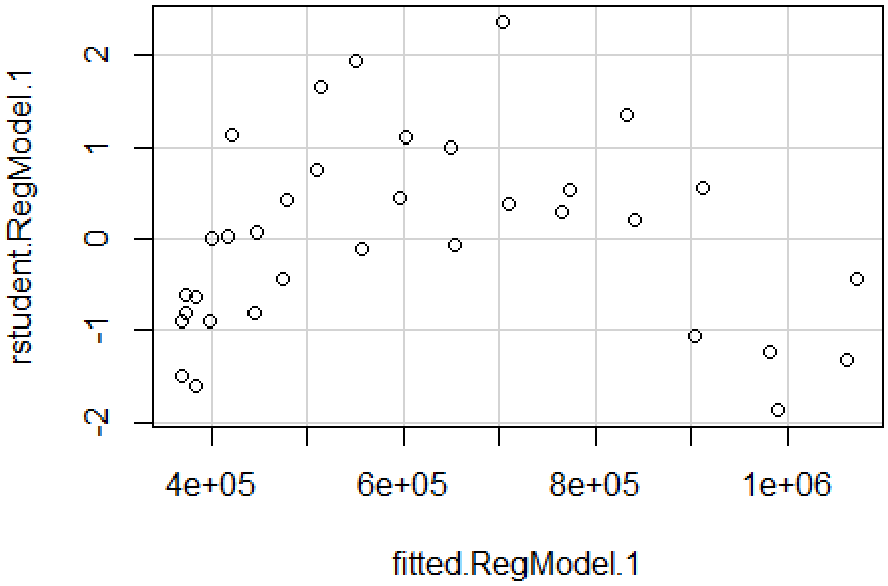
```
> scatterplot(rstudent.RegModel.1~price_sq, reg.line=FALSE, smooth=FALSE,
+   spread=FALSE, boxplots=FALSE, span=0.5, ellipse=FALSE, levels=c(.5, .9),
+   data=smartphone)
```



Seems not to be the case

GM4: Heteroscedasticity?

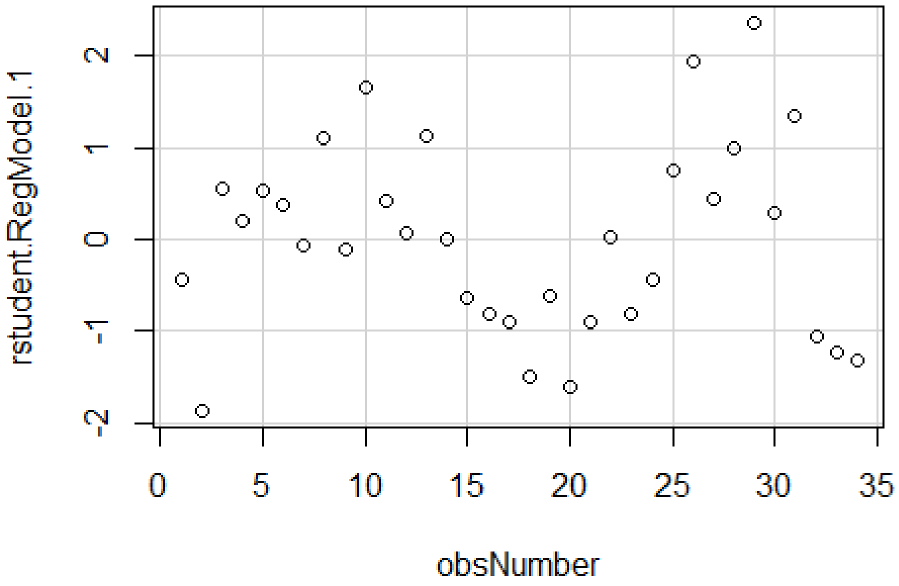
```
> scatterplot(rstudent.RegModel.1~fitted.RegModel.1, reg.line=FALSE,
+   smooth=FALSE, spread=FALSE, boxplots=FALSE, span=0.5, ellipse=FALSE,
+   levels=c(.5, .9), data=smartphone)
```



Difficult to evaluate (GM2 is violated)

GM5: Autocorrelation?

```
> scatterplot(rstudent.RegModel.1~obsNumber, reg.line=FALSE, smooth=FALSE,
+   spread=FALSE, boxplots=FALSE, span=0.5, ellipse=FALSE, levels=c(.5, .9),
+   data=smartphone)
```



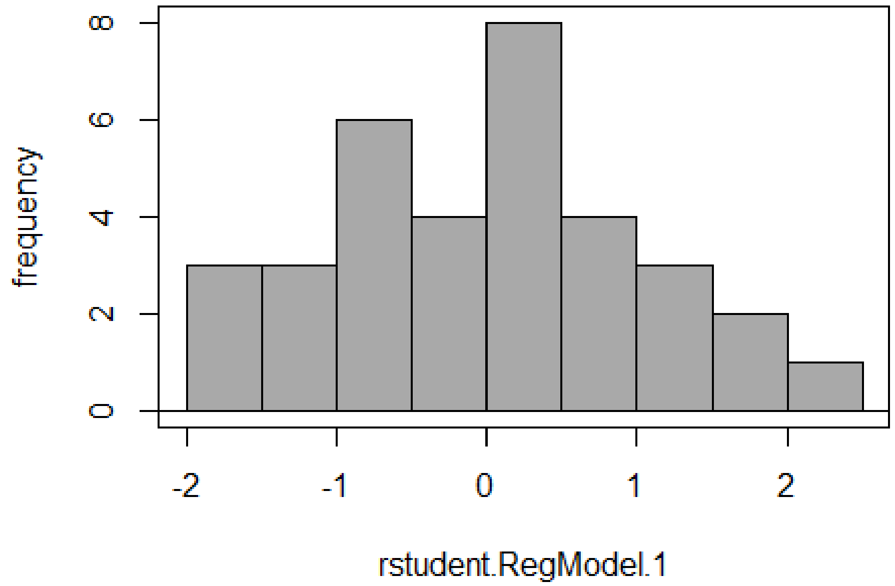
Pattern, autocorrelation seems to be problematic

GM6: Multicollinearity?

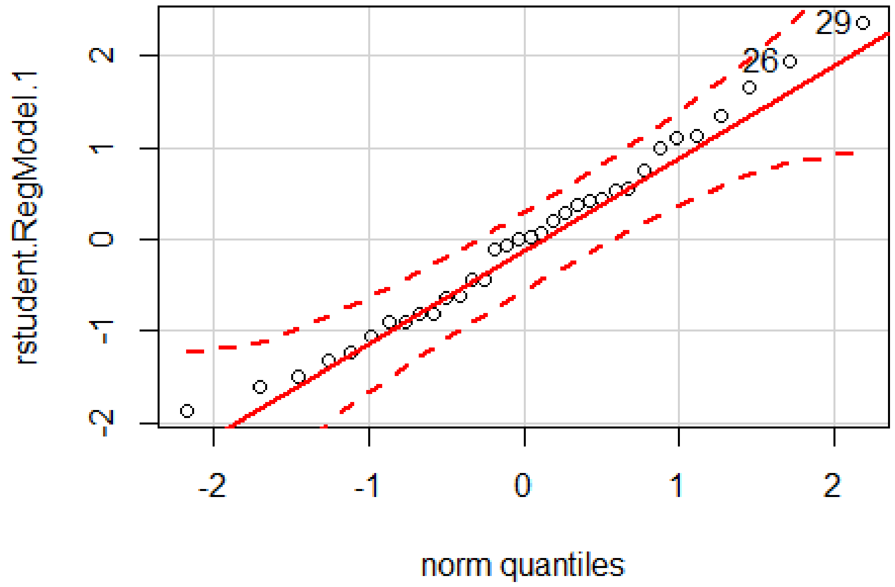
Not a problem, because there is only one independent variable

GM7: Normality?

```
> with(smartphone, Hist(rstudent.RegModel.1, scale="frequency",
+   breaks="Sturges", col="darkgray"))
```



```
> with(smartphone, qqPlot(rstudent.RegModel.1, dist="norm", id.method="y",
+   id.n=2, labels=rownames(smartphone)))
```



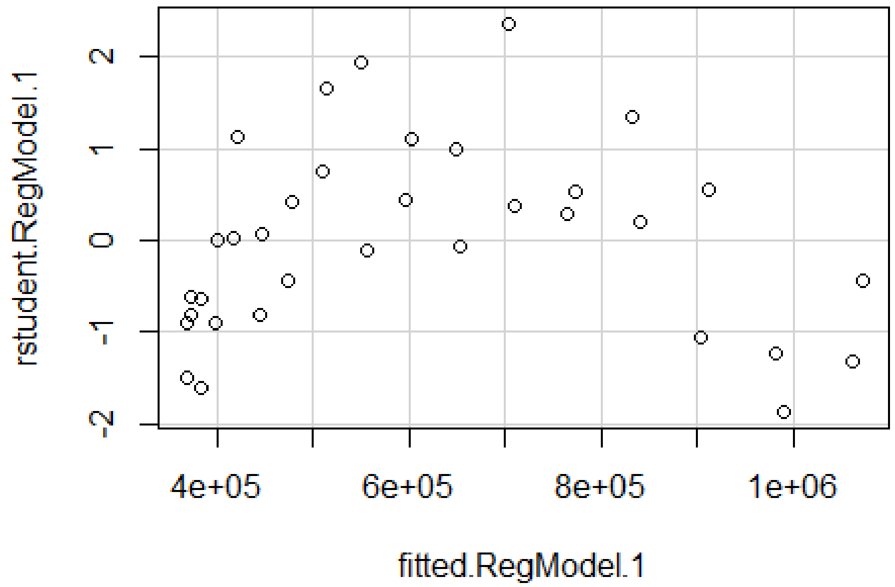
```
29 26
34 33
```

```
> library(nortest, pos=14)

> with(smartphone, shapiro.test(rstudent.RegModel.1))


    Shapiro-Wilk normality test

data:  rstudent.RegModel.1
W = 0.98304, p-value = 0.8625
```

Normal distribution is not violated

Outliers?

```
> scatterplot(rstudent.RegModel.1~fitted.RegModel.1, reg.line=FALSE,
+   smooth=FALSE, spread=FALSE, boxplots=FALSE, span=0.5, ellipse=FALSE,
+   levels=c(.5, .9), data=smartphone)
```



Studentised residuals are not substantial larger than 2

Summary:

Several assumptions are violated

Most important problem is probably the missing of important independent variables

Course of action would be to solve this problem first and then estimate and evaluate the model again