Case Study 4 - Group 4

Annika Janson h11829506
Jan Beck h11814291
Franz Uchatzi h1451890

13.12.2020

2 Model

2.1 Model estimation

2.1.1 and 2.1.2

(See page 2 for model comparison and regression output.)

The R^2 value of model 1 and 2 is **0.348** and **0.828** respectively.

The estimates and standard errors for the non-brand explanatory variables of model 1 and 2 are identical.

The estimates for `rq`, `vo`, `wa`, `ju`/intercept, `education`, `income`, `age` and `price` are significant at the 5%-level.

2.2

Model 1: the estimate for `kr` is **-0.287950**, which means that on average the rating is changing by **-0.2887950** c.p. In other words, we shift the regression line down by 0.2887950.

Model 2: the estimate for `kr` is **20.560087**, this is the intercept for `kr`. On average, if the brand kr and all other variables were 0, the rating would be **20.560087** c.p.

2.3

We can calculate the regression parameter associated with `kr` in Model 1 by subtracting the value of `ju` in Model 2 from the value of `kr`in Model 2.

This is because `ju` was our reference group, so the intercept of Model 1 is equivalent to the intercept of `ju`, which is also shown in Model 2. Model 1 shows us the difference between choosing "kr" or any other group and Model 2 shows us each groups intercept.

Table 1: Model comparison

	Dependent variable:	
	rating	
	(1)	(2)
rq	3.884***	24.732***
	(0.312)	(0.478)
vo	3.557***	24.405***
	(0.312)	(0.478)
wa	0.596*	21.444***
	(0.312)	(0.478)
kr	−0.288	20.560***
	(0.312)	(0.478)
ju		20.848***
		(0.478)
education	−0.257	−0.257
	(0.218)	(0.218)
gender	−0.107	−0.107
	(0.200)	(0.200)
income	−0.641***	−0.641***
	(0.205)	(0.205)
age	0.012**	0.012**
	(0.006)	(0.006)
price	−0.303***	−0.303***
	(0.008)	(0.008)
Constant	20.848***	
	(0.478)	
Observations	3,195	3,195
R^2	0.348	0.828
Adjusted R^2	0.346	0.828
Residual Std. Error (df = 3185)	5.584	5.584
F Statistic	188.881*** (df = 9; 3185)	1,537.900*** (df = 10; 3185)
Note:		*p<0.1; **p<0.05; ***p<0.01

2.4

H0: $\beta_{wa} = 0$
H1: $\beta_{wa} \neq 0$

In model 1, the p-value for β_{wa} is **0.05641**. Therefore, for $\alpha = 0.05$, we can not reject the null hypothesis. We conclude, that there is no difference in the average rating between the brands `ju` and `wa` c.p.

Bonus question:

```
## Linear hypothesis test
##
## Hypothesis:
## wa - ju = 0
##
## Model 1: restricted model
## Model 2: rating ~ 0 + rq + vo + wa + kr + ju + education + gender + income +
##     age + price
##
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1   3186 99433
## 2   3185 99320  1    113.58 3.6425 0.05641 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-test shows that the p-value again is **0.05641**, which is exactly the p-value we expected, as it was the one we could see in the results of `wa` in Model 1.

2.5

2.5.1 To check whether the brand information is helpful to determine the rating of mineral water, we perform an F-test for Model 1 with the following H0 and H1. However, we need to exclude the variable `ju` as it acts as the baseline for the brand effect in Model 1.

H0: $\beta_{rq} = \beta_{vo} = \beta_{wa} = \beta_{kr} = 0$
H1: H_0 is not true.

```
## Linear hypothesis test
##
## Hypothesis:
## rq = 0
## vo = 0
## wa = 0
## kr = 0
##
## Model 1: restricted model
## Model 2: rating ~ rq + vo + wa + kr + education + gender + income + age +
##     price
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   3189 109650
## 2   3185  99320  4     10331 82.823 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After we run the test we find that the p-value is $5.5040372 \times 10^{-67}$ and the F-statistic is **82.8229128**. Therefore, we reject the null hypothesis that the coefficients for `rq`, `vo`, `wa` and `kr` are equal to 0 and should keep them in the model, c.p.

Bonus question: For the bonus question, we take same approach as for Model 1 with the difference that now, all brands of mineral water are included in the H0. In Model 2 the intercept β_0 is excluded.

$H_0 : \beta_{ju} = \beta_{rq} = \beta_{vo} = \beta_{wa} = \beta_{kr} = 0$
$H_1 : H_0$ is not true.

```
## Linear hypothesis test
##
## Hypothesis:
## rq = 0
## ju = 0
## vo = 0
## wa = 0
## kr = 0
##
## Model 1: restricted model
## Model 2: rating ~ 0 + rq + vo + wa + kr + ju + education + gender + income +
##     age + price
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   3190 192342
## 2   3185  99320  5     93023 596.61 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We run an F-test test and find that the p-value is **0** and the F-statistic is **596.613813**. Therefore, we reject the null hypothesis that the coefficients for `ju`, `rq`, `vo`, `wa` and `kr` are equal to 0 and should keep them in the model, c.p.

2.5.2 For our Model 3, we remove all brand variables from Model 1.

Table 2: Model comparison

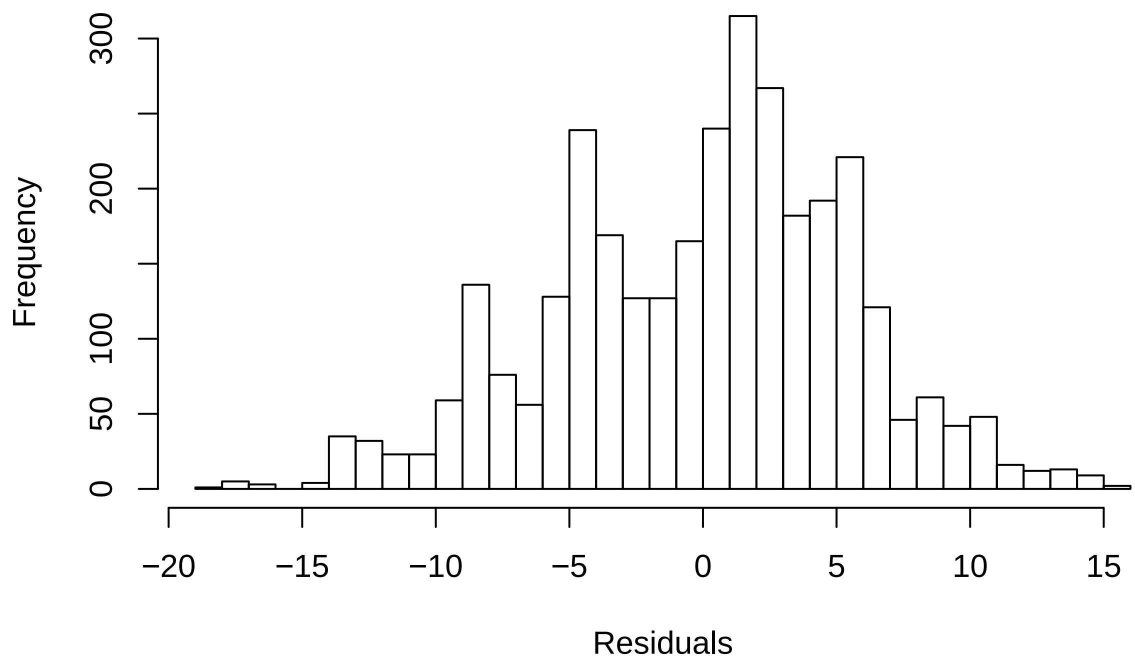
	K	R-squared	Adj. R-squared	AIC	BIC
Model 1	9	0.348	0.346	20,069.450	20,136.210
Model 3	5	0.280	0.346	20,377.610	20,420.100

The table above shows various model selection criteria for Model 1 and Model 3. We see that R-squared of Model 1 is **0.0678192** larger than for Model 3, suggesting that Model 1 explains **6.7819**% more variation in rating can be explained with variation of the independent variables. However, Model 1 consists of 4 more explanatory variables than Model 3 and the R-squared increases for each additional explanatory variable added to the model.

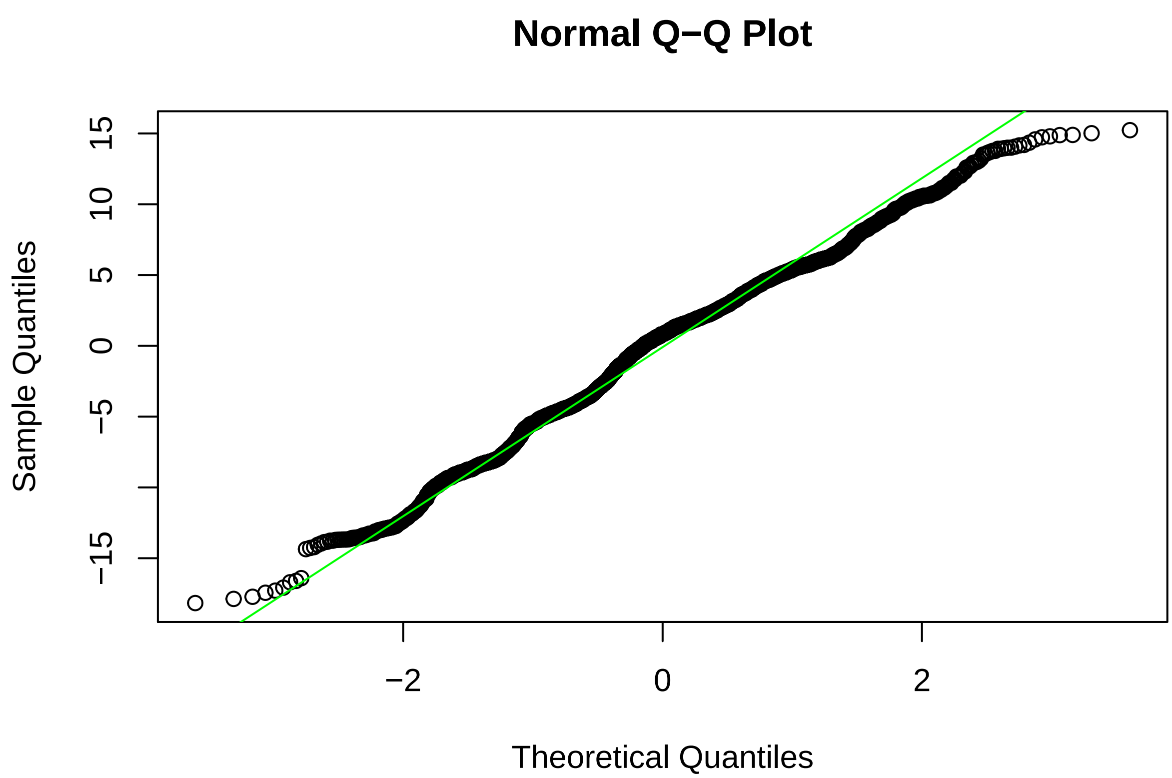
We therefore look at the adjusted R-squared next, which penalizes extra variables added to the model. Its values is the same for Model 1 and Model 3 respectively with **0.3461517** . This criterion suggests, that adding the brand variables does not increase goodness of fit.

Lastly, we compare the AIC and BIC values for each model and see that for Model 1 the AIC is **308.1600653** and the BIC is **283.8826959** units smaller than for model 3. The smaller AIC and BIC values of Model 1 indicate a better fit of the model in comparison to Model 3. By this criterion, Model 1 explains the changes in rating better than Model 3.

2.6



Frequency
0
50
100
200
300
−20
−15
−10
−5
0
5
10
15
Residuals



Normal Q−Q Plot
Sample Quantiles
−15
−5
0
5
10
15
−2
0
2
Theoretical Quantiles

```
##
##  Jarque Bera Test
##
## data:  resids
## X-squared = 36.524, df = 2, p-value = 1.172e-08
```

H0: Residuals are normally distributed H1: Residuals are not normally distributed

Histogram: Looking at the histogram, it does not look like a symmetric normal distribution around 0. The distribution seems slightly left-skewed and there are less values at the center than we would expect for a normal distribution.

QQ-Plot: Till 1.5 it seems the residuals follow a normal distribution. But for values higher than 1.5, they seem to differ from normal distribution.

Jarque-Bera-Test: The test confirms our observations from the histogram and the QQ-Plot. With X-squared = **36.525** it is bigger than **6**, which is the limit. Additional the p-value is **1.172e-08**, so very small. At a 5%-level, the residuals are not normally distributed and we reject the H0.

Summarizing our observations, our error term is not normally distributed, we have a problem with our model.

2.7

We add interactions between dummy variables and continuous explanatory variables in two steps. First, the interaction between `kr` and `age`. Second, interaction added is between the variables `wa` and `price`. The results between each step are shown in 2.8.

2.8

Table 3:

	R-squared	Adj. R-squared	AIC	BIC
Model 1	0.348	0.346	20,069.450	20,136.210
Step1 (kr:age)	0.348	0.346	20,071.400	20,144.230
Step2 (wa:price)	0.348	0.346	20,072.990	20,151.890

The table above shows the addition of each interaction between a pair of selected variables and the effect on R-squared, adjusted R-squared, AIC and BIC. As a reference we compare each change in the parameters with the respective parameters of Model 1. Furthermore, the next table shows the respective p-values as well as t-test results for each interaction term and step respectively.

Table 4:

	p-values (1)	p-values (2)	t-test (1)	t-test (2)
Step1 (kr:age)	0.817	0.523	0.232	0.817
Step2 (wa:price)	0.817	0.232	−0.639	0.817

2.8.1

```
##
## Call:
## lm(formula = rating ~ rq + vo + wa + kr + education + gender +
##     income + age + price + kr:age + wa:price, data = marketing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.1010  -4.1185   0.8302   3.9312  15.1589
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.779705   0.511606  40.617  < 2e-16 ***
## rq           3.884194   0.312487  12.430  < 2e-16 ***
## vo           3.557121   0.312487  11.383  < 2e-16 ***
## wa           1.064896   0.797149   1.336  0.18168
## kr          -0.414490   0.629351  -0.659  0.51020
## education   -0.256875   0.218174  -1.177  0.23913
## gender      -0.106798   0.199940  -0.534  0.59328
## income      -0.641044   0.204740  -3.131  0.00176 **
## age          0.011408   0.006677   1.709  0.08761 .
## price       -0.299911   0.009205 -32.580  < 2e-16 ***
## kr:age       0.003347   0.014449   0.232  0.81684
## wa:price    -0.013161   0.020594  -0.639  0.52283
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.586 on 3183 degrees of freedom
## Multiple R-squared:  0.3481, Adjusted R-squared:  0.3458
## F-statistic: 154.5 on 11 and 3183 DF,  p-value: < 2.2e-16
```

2.8.2 We interpret the interaction term between `kr` and `age`. The estimated coefficient in our model for the interaction term is **0.003347**. This means, that for every additional year a consumer would rate the mineral water of brand `kr` on average **0.003347** higher. This could be the result of a marketing strategy that primarily targets older consumers.

It's important to note that this additional effect only applies to mineral waters of the brand `kr`. `age` and brand `kr` still have separate effects on the average rating, therefore we can not say "all else equal" in respect to the interaction.

2.9

From the output in 2.8.1 and Table 3 we can see that the two interaction terms we added are not significant at the 5%-level and did not increase goodness of fit by any measure. We therefore drop these terms and return to Model 1 and inspect and p-values of the included terms.

```
##
## Call:
## lm(formula = rating ~ rq + vo + wa + kr + education + gender +
##     income + age + price, data = marketing)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -18.167  -4.118   0.827   3.931  15.232
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.848037   0.477726  43.640  < 2e-16 ***
## rq           3.884194   0.312412  12.433  < 2e-16 ***
## vo           3.557121   0.312412  11.386  < 2e-16 ***
## wa           0.596244   0.312412   1.909  0.05641 .
## kr          -0.287950   0.312412  -0.922  0.35675
## education   -0.256875   0.218121  -1.178  0.23902
## gender      -0.106798   0.199892  -0.534  0.59319
## income      -0.641044   0.204691  -3.132  0.00175 **
## age          0.012078   0.006017   2.007  0.04483 *
## price       -0.302541   0.008232 -36.750  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.584 on 3185 degrees of freedom
## Multiple R-squared:  0.348,  Adjusted R-squared:  0.3462
## F-statistic: 188.9 on 9 and 3185 DF,  p-value: < 2.2e-16
```

Since gender has the highest p-value of the remaining variables, we drop it from the model next.

```
##
## Call:
## lm(formula = rating ~ rq + vo + wa + kr + education + income +
##     age + price, data = marketing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.096  -4.096   0.814   3.954  15.186
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.795786   0.467557  44.478  < 2e-16 ***
## rq           3.884194   0.312377  12.434  < 2e-16 ***
## vo           3.557121   0.312377  11.387  < 2e-16 ***
## wa           0.596244   0.312377   1.909  0.05639 .
## kr          -0.287950   0.312377  -0.922  0.35670
## education   -0.257095   0.218096  -1.179  0.23856
## income      -0.637549   0.204563  -3.117  0.00185 **
## age          0.011828   0.005999   1.972  0.04871 *
## price       -0.302541   0.008232 -36.754  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.584 on 3186 degrees of freedom
## Multiple R-squared:  0.3479, Adjusted R-squared:  0.3463
## F-statistic: 212.5 on 8 and 3186 DF,  p-value: < 2.2e-16
```

Finally, we remove `education`:

```
##
## Call:
## lm(formula = rating ~ rq + vo + wa + kr + income + age + price,
##     data = marketing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.0259  -4.1470   0.8092   3.9519  15.1029
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.589149   0.433480  47.497  < 2e-16 ***
## rq           3.884194   0.312396  12.434  < 2e-16 ***
## vo           3.557121   0.312396  11.387  < 2e-16 ***
## wa           0.596244   0.312396   1.909 0.056401 .
## kr          -0.287950   0.312396  -0.922 0.356730
## income      -0.694788   0.198729  -3.496 0.000478 ***
## age          0.013631   0.005801   2.350 0.018844 *
## price       -0.302541   0.008232 -36.752  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.584 on 3187 degrees of freedom
## Multiple R-squared:  0.3477, Adjusted R-squared:  0.3462
## F-statistic: 242.6 on 7 and 3187 DF,  p-value: < 2.2e-16
```

The final improved model has a BIC of 2.0121754×10^4 (recall that the BIC from Model 4, with which we started, was 2.0151888×10^4). We have therefore improved goodness of fit even though we dropped 4 variables in the process. The variables that we kept are all significant at the 5%-level except for `wa` and `kr`.

3 Theorie

3.1

That is true. R^2 is always increasing with each additional variable, no matter how good the new variable is. In general SSR are always smaller than TSS, and R^2 is close to 1 the smaller SSR is. If SSR = 0, then $R^2 = 1$. In this case we don't make any errors and were able to explain the variance of our model completely. In a model with a fixed number of observations N, R^2 will be always 1 if we add N-1 explanatory variables, no matter how useful they are.

For example:

```
##
## Call:
## lm(formula = log(consum) ~ log(income) + log(pchick) + log(pbeef) +
##     log(ppork), data = chick1)
##
## Residuals:
## ALL 5 residuals are 0: no residual degrees of freedom!
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.6347         NA      NA       NA
## log(income)  -1.0030         NA      NA       NA
## log(pchick)  -0.7657         NA      NA       NA
## log(pbeef)   -2.9596         NA      NA       NA
## log(ppork)    2.6654         NA      NA       NA
##
## Residual standard error: NaN on 0 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:    NaN
## F-statistic:   NaN on 4 and 0 DF,  p-value: NA
```

We used the chicken data set to show that R^2 is increasing to 1, if we set the numbers of observations to explanatory variables + 1. We created a new data frame including all 4 explanatory variables (`income`, `pbeef`, `pchick`, `ppork`) and 5 observations. The result shows us the expected R^2 of 1.

The adjusted R^2 in comparison, is taking in to account how good the new variable is. So the $ R^2adj $ is only increasing, if the change in R^2 is large.

The formula: $R^2_{adj} = 1 - \frac{N-1}{N-K-1} * (1 - R^2)$ So with increasing "K", the term 1 $\frac{N-1}{N-K-1}$ gets bigger and $R^2 adj$ smaller, but with the term $(1 - R^2)$ it is still increasing if the change is large.

3.2

We consider the model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + u$

First, we should test if we should include the quadratic term or not:

H0: $\beta_2 = 0$
H1: $\beta_2 \neq 0$

We can use a t-test to that end. If $\beta_0 \neq 0$, non-linearity is given in our model and we should keep the quadratic term.

Next, to check whether the sign changes at 1 for X, we use the following formula

$$\frac{\partial \mathbb{E}(Y \mid X = x)}{\partial x} = \beta_1 + 2\beta_2 x = 0$$

which yields that sign change lies at

$$X_0 = -\beta_1/(2\beta_2)$$

Since we test for $X_0 = 1$, our null hypothesis is:

$$\beta_1 + 2\beta_2 = 0 \quad \text{or:} \quad \beta_2 + \frac{\beta_1}{2} = 0$$

Only if β_1 and β_2 have different signs, the vertex can be positive, as in our case where $X_0 = 1$.

We can test this using an F-test.

3.3

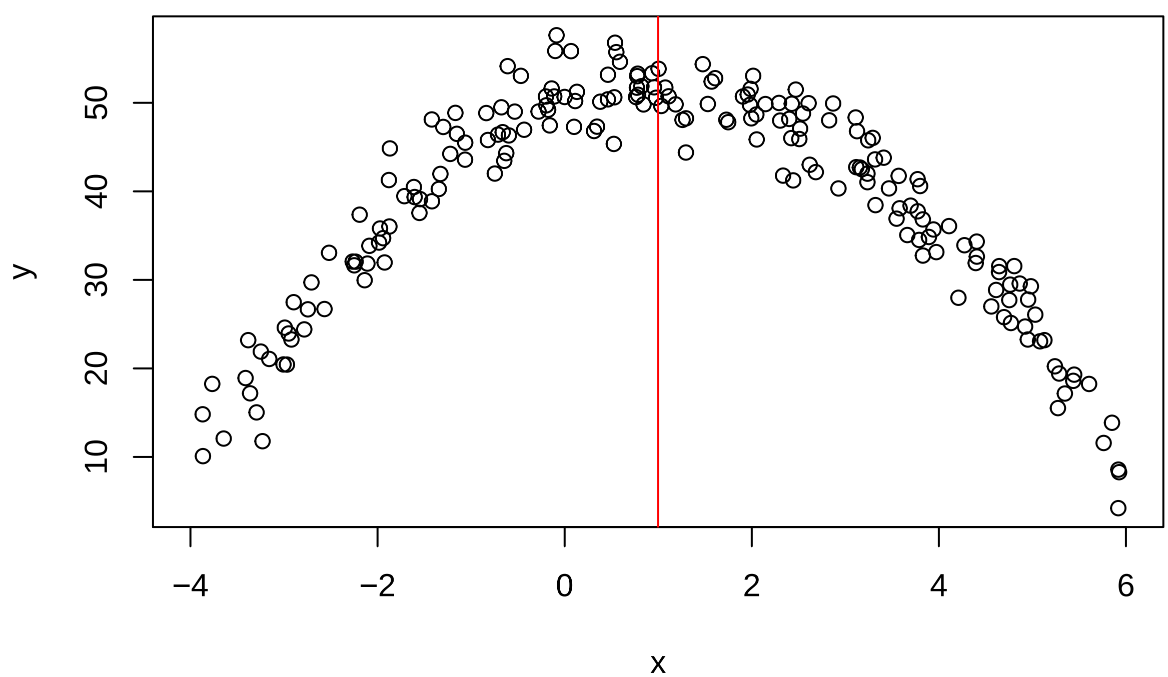
We simulate non-linear data where the sign change occurs at $X_0 = 1$ and plot the result:

```
set.seed(1)

# our parameters
N <- 200
beta0 <- 50
beta1 <- 3.5
beta2 <- -beta1/2 # sign change at 1
mu <- 0
sigma <- 3
minX = -4
maxX = 6

# our model
x <- runif(N, min = minX, max = maxX)
u <- rnorm(N, sd = sigma, mean = mu)
y <- beta0 + beta1*x + beta2*x^2 + u

# plot data
plot( x, y, xlim = c(minX, maxX) )
abline(v=1, col="red")
```



Next, we try to estimate a model with a quadratic term

```
qm <- lm(y ~ x + I(x^2))
summary(qm)
```

```
##
## Call:
## lm(formula = y ~ x + I(x^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8611 -1.8162 -0.1815  1.9085  8.0432
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  49.9007     0.2968  168.15   <2e-16 ***
## x             3.4597     0.1049   32.97   <2e-16 ***
## I(x^2)       -1.7343     0.0311  -55.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.998 on 197 degrees of freedom
## Multiple R-squared:  0.9409, Adjusted R-squared:  0.9403
## F-statistic:  1568 on 2 and 197 DF,  p-value: < 2.2e-16
```

The regression output shows that the p-value of the quadratic term is very low (significant) and the adjusted R-squared is **0.943** (the model fits the data well).

For comparison, we estimate a model without a quadratic term:

```
lm <- lm(y ~ x)
summary(lm)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.002  -8.340   3.635  10.042  18.056
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.5424     0.9460  41.801   <2e-16 ***
## x            -0.3925     0.3228  -1.216    0.225
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.25 on 198 degrees of freedom
## Multiple R-squared:  0.007411,   Adjusted R-squared:  0.002398
## F-statistic: 1.478 on 1 and 198 DF,  p-value: 0.2255
```

The p-value of β_1 alone is now **0.758** and the adjusted R-squared **-0.009225**. Therefore, keeping the quadratic term is a good idea.

Lastly, we run an F-test to see if the hypothesis $\beta_1 + 2\beta_2 = 0$ holds:

```r
linearHypothesis(qm, c("x+2*I(x^2)=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## x  + 2 I(x^2) = 0
##
## Model 1: restricted model
## Model 2: y ~ x + I(x^2)
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    198 1770.5
## 2    197 1770.3  1   0.11496 0.0128 0.9101
```

From the F-test we obtain an F-statistic of **0.0128** and a p-value of **0.9101**. Therefore, we find little evidence in the data that we should reject the hypothesis that the sign chance occurs at $X_0 = 1$.