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**. We find little evidence in the data that we should reject the null hypothesis that the coefficients for `rq`, `vo`, `wa` and `kr` are equal to 0 and therefore can be jointly excluded from 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.

H0: $H_0 : \beta_{ju} = \beta_{rq} = \beta_{vo} = \beta_{wa} = \beta_{kr} = 0$
H1: $H_1 : H_0$ 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**. We find little evidence in the data that we should reject the null hypothesis that the coefficients for `ju`, `rq`, `vo`, `wa` and `kr` are equal to 0 and therefore can be jointly excluded from the model, c.p.

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

Table 2: Model comparison

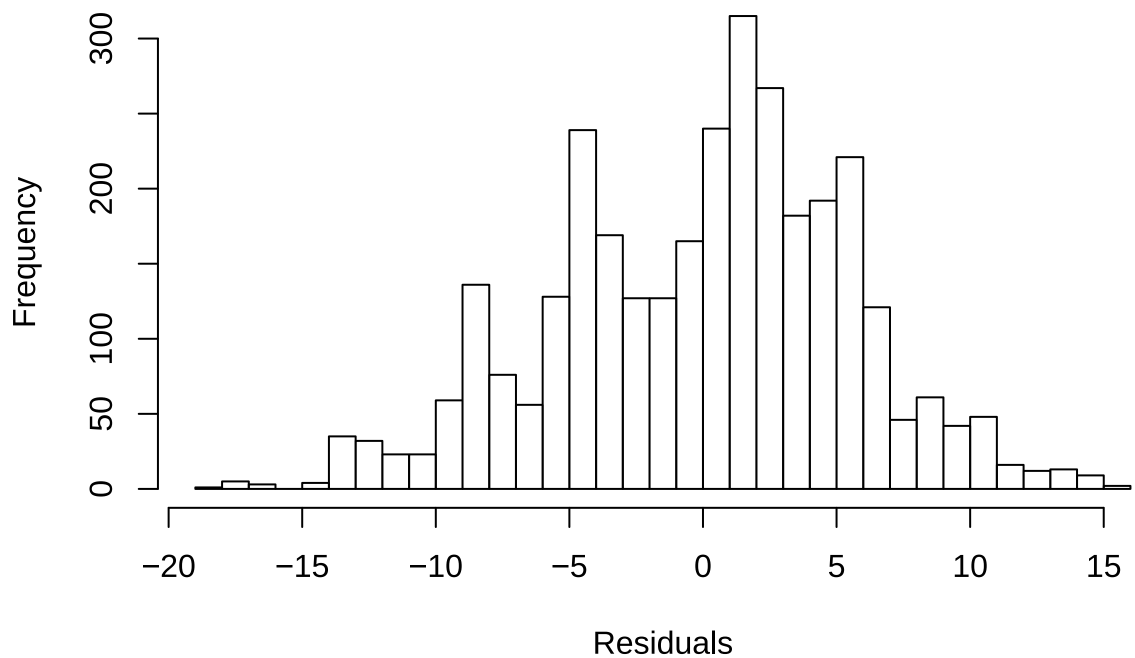
	K	R-squared	Adj. R-squared	AIC	BIC
Model 1	9	0.348	0.346	$20,069.450$	$20,136.210$
Model 3	5	0.280	0.346	$20,377.610$	$20,420.100$

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.

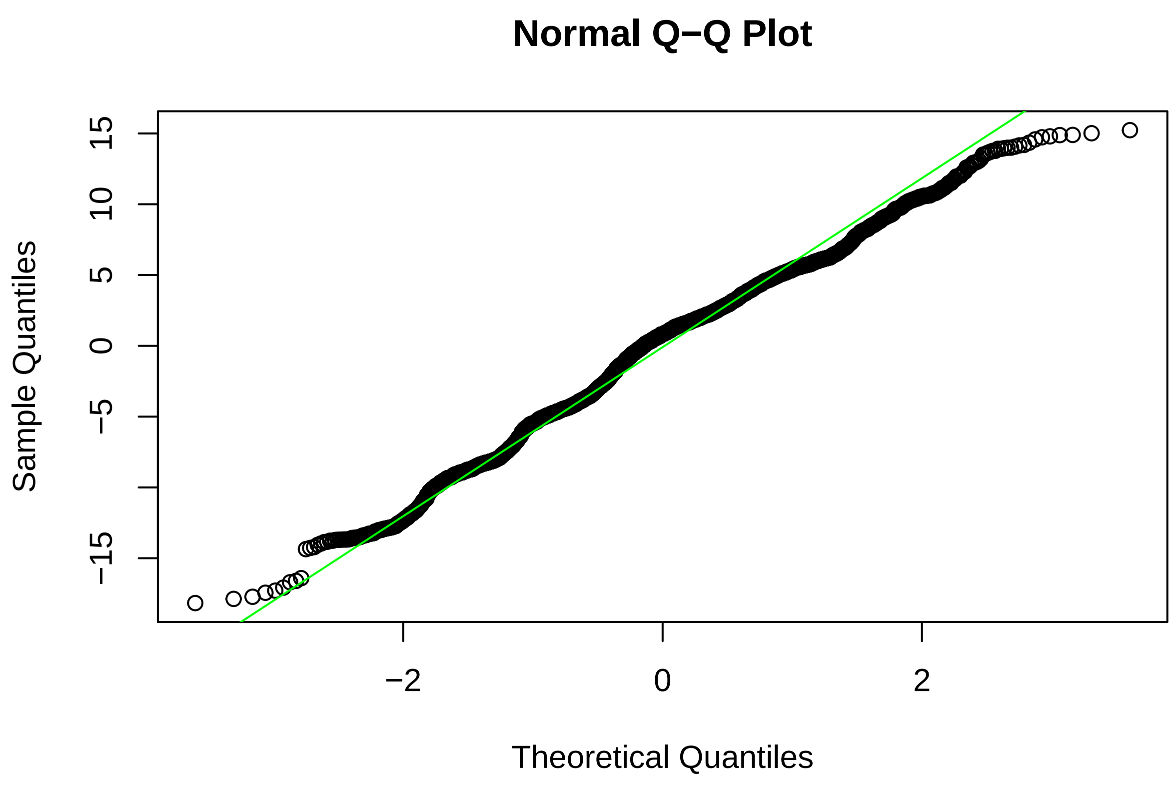
We therefore look at the adjusted R-squared next, which penalizes extra variables added to the model. Its values is the same for Model 1 and Model 3 respectively with **0.3461517** . This criterion suggests, that adding the brand variables does not increase goodness of fit.

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 distribution around 0. It seems that the residuals are not normally distributed, as they are located around 1.5 and not around 0. Additionally they have outliers on both sides.

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 three steps. First, the interaction between `kr` and `age`. Second, between the variables `vo` and `income`. The last 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 (vo:income)	0.348	0.346	20,071.730	20,150.630
Step3 (wa:price)	0.348	0.346	20,073.320	20,158.290

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.

2.8.1

```
##
## Call:
## lm(formula = rating ~ rq + vo + wa + kr + education + gender +
##     income + age + price + kr:age + vo:income + wa:price, data = marketing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.3439  -4.1330   0.7936   3.9543  15.0952
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.716642   0.513885  40.314   <2e-16 ***
## rq           3.884194   0.312455  12.431   <2e-16 ***
## vo           3.856810   0.389425   9.904   <2e-16 ***
## wa           1.064896   0.797066   1.336   0.1816
## kr          -0.398865   0.629402  -0.634   0.5263
## education   -0.256875   0.218151  -1.178   0.2391
## gender      -0.106798   0.199920  -0.534   0.5932
## income      -0.513376   0.227407  -2.258   0.0240 *
## age          0.011491   0.006676   1.721   0.0853 .
## price       -0.299911   0.009204 -32.584   <2e-16 ***
## kr:age       0.002934   0.014451   0.203   0.8391
## vo:income   -0.638339   0.495079  -1.289   0.1974
## wa:price    -0.013161   0.020591  -0.639   0.5228
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.585 on 3182 degrees of freedom
## Multiple R-squared:  0.3484, Adjusted R-squared:  0.346
## F-statistic: 141.8 on 12 and 3182 DF,  p-value: < 2.2e-16
```

2.8.2

2.9

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 (inocme, 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^2adj smaller, but with the term $(1 - R^2)$ it is still increasing if the change is large.

3.2

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

The null hypothesis for a statistical test that the point where the effect of a marginal increase in X on the conditional expectation E(Y |X) changes its sign is 1:

H0: $\beta_2 = 0$
H1: $\beta_2 \neq 0$ We can use a t-test, to test if we should include the quadratic part of the function or not. If $\beta_0 \neq 0$ non-linearity is given in our model and we should not exclude the quadratic term.

We are looking for the point where the marginal increase in X on the conditional expectation $E(Y|X = 1) = 0$.
H0: $1 = -\beta_1/(2\beta_2)$

We can calculate the point where the signs change with $X_0 = -\beta_1/(2\beta_2)$. If β_1 *and* β_2 have different signs, the vertex can be positive. So only for different signs of β_1 *and* β_2 the vertex can be 1. In our case this is true if we set $ X_0= 1$, so $-\beta_1/(2\beta_2) = 1$

??? Welcher Test

3.3