WS 2021/22 1048 - Econometrics I
Case Study 4

Anja Kulagic (h12100543), Jovana Mileusnic (h12100542), Ema Vargova (h11914081)

Contents

List of Figures

List of Tables

1 Data Description

```
data <- read.csv("marketing.csv")
data$rating <- as.numeric(data$rating)
data$age <- as.numeric(data$age )
metric.index <- which(lapply(data, class)=="numeric")
factor.index <- which(lapply(data, class)=="integer")
```

```
op <- par(mfrow=c(2, 6), cex=.5, mar = c(1.5, 2.5, 1.5, 1), mgp = c(1.3, .5, 0))
for(i in 1:ncol(data)) {
  if(i %in% metric.index) {
    hist(data[[i]], freq=FALSE, main=names(data)[i], xlab="")
  } else {
    barplot(table(data[[i]]), main=names(data)[i], ylab="Frequency") }}
```

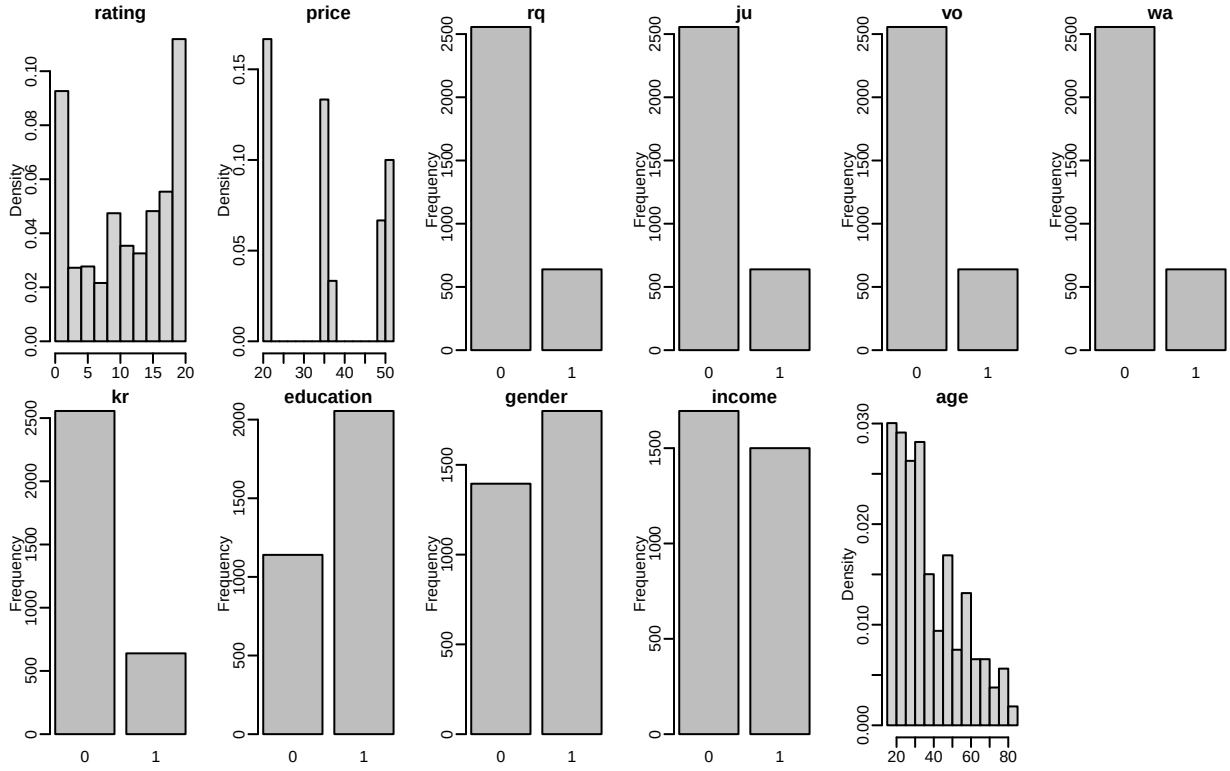

Figure 1: Distribution of variables

The provided dataset contains 11 variables where we can distinguish two groups of dummy variables, along with the variables price and age as one continuous and one discrete variable, respectively.

The first group of dummy variable is related to the brand of mineral water considering five different brands, sorting data into mutually exclusive categories (meaning that each consumer rating should be based on one out of five observed brands of mineral water, $rq_i + ju_i + vo_i + wa_i + kr_i = 1$).

This should be taken into consideration if we are building a multiple linear regression model as one of main Gauss-Markov conditions could be violated (*dummy variable trap*). Note that dataset includes 639 observations for each observed brand of mineral water.

However, so-called *zero-one variables* related to education, gender and income are not complementary, as there is not any kind of relationship between them, thus we can face with man consumer with no high school diploma and below average income, meaning that all previously mentioned dummy variables equal zero. If we look briefly at the price and age histograms, we can spot existence of three price categories up to a certain point, whereas we could notice that there is positive asymmetry in age distribution.

2 Model

2.1 Model estimation

2.1.1

```
model1 <- lm(rating ~ . - rq, data)
```

2.1.2

```
model2 <- lm(rating ~ . - 1, data=data)
```

```
library(stargazer)
stargazer(model1, model2, single.row=TRUE, header=FALSE, model.numbers=FALSE,
          column.labels = c("Model 1", "Model 2"), title="Multiple OLS Regression Models")
```

Table 1: Multiple OLS Regression Models

	Dependent variable:	
	rating	
	Model 1	Model 2
price	−0.303*** (0.008)	−0.303*** (0.008)
rq		24.732*** (0.478)
ju	−3.884*** (0.312)	20.848*** (0.478)
vo	−0.327 (0.312)	24.405*** (0.478)
wa	−3.288*** (0.312)	21.444*** (0.478)
kr	−4.172*** (0.312)	20.560*** (0.478)
education	−0.257 (0.218)	−0.257 (0.218)
gender	−0.107 (0.200)	−0.107 (0.200)
income	−0.641*** (0.205)	−0.641*** (0.205)
age	0.012** (0.006)	0.012** (0.006)
Constant	24.732*** (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

Table 1 was created using stargazer v.5.2.2 by Hlavac (2018).

In order to avoid so called *dummy variable trap*, solution is to drop one of the categorical variables, or in other words to let one mineral water brand to be a baseline. Using *rq* as a baseline for the brand effect, we consider this brand to the *reference value* while the values of the remaining categories represent the change from this reference. In other words, the brand rq stands for the group against which comparison is made. All other variables are considered as self-explanatory variables.

Alternative way to avoid dummy variable trap is to drop the intercept constant from the multiple regression model. In our case, this is illustrated with the model 2, which now does not use a brand as a baseline but includes all brand dummies in the model, while excluding the intercept. However, after comparing OLS estimates of coefficients related to self-explanatory variables from both models, it is clear that this change does not have any influence on their values. To rephrase it, they are exactly the same in both models. Even though including g dummies without an overall intercept is sometimes useful, it has two practical drawbacks which we should be aware of. Firstly, it makes it more cumbersome to test for differences relative to a base group and additionally, we will face with the uncentered R-squared, which will be discussed later.

2.2

Model 1: To interpret the coefficients on the dummy variables, we must remember that the base group, or so-called benchmark, is the mineral water brand rq. Thus, the estimates on the four dummy variables measure the proportionate difference in consumer rating relative to the brand rq. To be specific, the brand vo is estimated to have 0.327 lower rating than the brand rq, with the same levels of the other variables.

Model 2: Looking at the estimator related to the brand vo, we can conclude that if the dummy variable of the brand vo is equal to one, all other dummy variables related to mineral water brands must be zero. Furthermore, it will have a positive effect of 24.405 on consumer rating, with all other variables being constant.

2.3

We can calculate the regression parameter associated with vo in Model 1 (denoted $\hat{\beta}_3$) from the regression parameters of Model 2 (denoted $\tilde{\beta}_4$) in the following way:

$$\hat{\beta}_3 = \tilde{\beta}_2 + \tilde{\beta}_4$$

$$\hat{\beta}_3 = 24.732 - 24.405 \approx 0.327$$

To be specific, in Model 1 the coefficient on the mineral water brand rq would be represented by the intercept, whereas the estimates on the remaining brands of mineral water would measure the proportionate difference in consumer rating relative to the brand rq. Therefore, the straightforward way of computing the estimate on the brand vo in Model 1 is by subtracting the estimation on vo from the estimation on rq in Model 2.

2.4

We are interested in hypothesis that average rating between brands rq and wa are identical, with the same levels of the other variables, at a significance level of $\alpha = 0.05$.

For the usual t test to be valid, we must assume that the homoskedasticity assumption holds, which means that the population variance in ratings is the same among the different brands of mineral water, namely, the brand rq and the brand wa.

For Model1:

```
library(car)
kable(linearHypothesis(model1, c("wa=0")),
      caption = "Testing the null hypothesis: There is no difference in the average rating
      between the brands rq and wa, ceteris paribus.")
```

Table 2: Testing the null hypothesis: There is no difference in the average rating between the brands rq and wa, ceteris paribus.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
3186	102773.54	NA	NA	NA	NA
3185	99319.55	1	3453.991	110.7633	0

Under the null:

$$t = \frac{\hat{\beta}_4}{se(\hat{\beta}_4|X)}|X \sim t_{df}$$

$$t = \frac{-3.288}{0.312} \approx -10.538$$

The estimated difference, -3.288, has a t statistic of -10.538, which is very statistically significant, and p-value of 2e-16 thus we conclude that the estimated difference between the brands rq and wa is very statistically significant and reject the null.

To conclude, when we are doing the testing with dummy variables, the easiest thing to do it to choose one of these groups to be the base group in order to get needed estimates and its standard error directly.

For Model2:

```
kable(linearHypothesis(model2, c("rq=wa")),
      caption = "Testing the null hypothesis: There is no difference in the average rating between the
```

Table 3: Testing the null hypothesis: There is no difference in the average rating between the brands rq and wa, ceteris paribus.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
3186	102773.54	NA	NA	NA	NA
3185	99319.55	1	3453.991	110.7633	0

Running a test on Model 2 would lead us to the same conclusion. On contrary, the F-test would be the preferred method for test in joint hypothesis which requires running two regressions, one regression on the unrestricted model and one regression on the restricted model, which we have imposed the restrictions of the null.

Manually, we can compute the F-statistic as follows:

$$F = \frac{(SSR_R - SSR_{UR})/g}{SSR_{UR}/(n-k-1)}$$

and the intuition behind it is:

$$F = \frac{\text{Average loss in explanatory power under } H_0}{\text{Average unexplained variation under } H_A}$$

We obtain F=110.76, which is high value, thus we can conclude that we lose a lot of explanatory power by our restrictions.

```
op <- par(mar = c(2.3, 2, 1, 1), mgp = c(1.1, .5, 0), cex.lab=0.8, cex.axis=0.6, cex.main=0.8)
confidenceEllipse(model2, fill = T, lwd = 0,which.coef = c("rq", "wa"),
                  ylab=bquote(paste(beta)[5]), xlab=bquote(paste(beta)[2]))
```

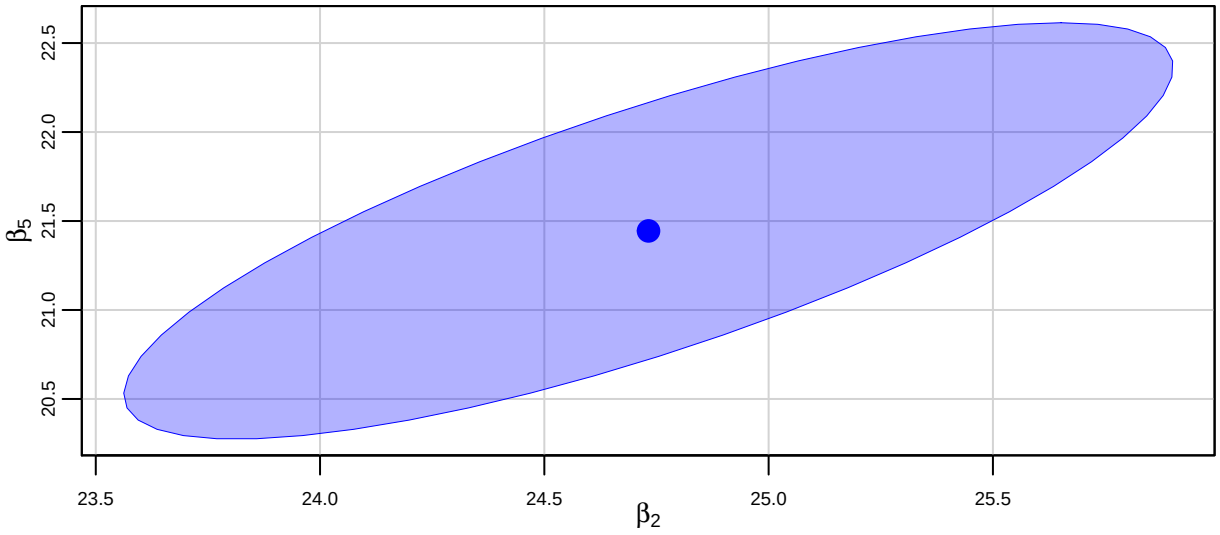

Figure 2: Model 2 - 95% Confidence Set

The 95% confidence set for coefficients on brands rq and wa (namely β_2 and β_5) is an ellipse which contains the pairs of values of β_2 and β_5 that cannot be rejected using F-statistic at the 5% significance level.

2.5

Building a reduced Model 3:

```
model3 <- lm(rating ~ price + education + gender + income + age, data=data)
```

2.5.1

```
kable(linearHypothesis(model1, c("ju=0", "vo=0", "wa=0", "kr=0")),
      caption = "Testing the null hypothesis: The brand information
      is not helpful to determine the rating of mineral water.")
```

Table 4: Testing the null hypothesis: The brand information is not helpful to determine the rating of mineral water.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
3189	109650.39	NA	NA	NA	NA
3185	99319.55	4	10330.84	82.82291	0

```
kable(linearHypothesis(model2, c("rq=0", "ju=0", "vo=0", "wa=0", "kr=0")), caption = "Testing the null
```

Table 5: Testing the null hypothesis: The brand information is not helpful to determine the rating of mineral water.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
3190	192342.18	NA	NA	NA	NA
3185	99319.55	5	93022.63	596.6138	0

In both cases, the obtained results lead us to the same conclusion. For the given p-value, we would reject null hypothesis which states that the brand information is not helpful to determine the rating of mineral water, as it is smaller than any significance level commonly used in practice.

However, we can notice that the obtained F-test differ between Model 1 and Model 2, due to high value of the sum of squared residuals (SSR) of restricted model, which may point out that if we exclude the brand information from our model, we will lose a lot of explanatory power due to our restriction. The loss is mainly noticeable in Model 2 caused by excluding intercept and the brand information which are obviously significant for our Model.

2.5.2

Table 6: Comparing models in terms of model selection criteria

	Model 1	Model 2	Model 3
R-squared	0.348	0.828	0.280
Adjusted R-squared	0.346	0.828	0.279
AIC	20069.451	20069.451	20377.611
BIC	20136.213	20136.213	20420.096

As we can notice, the value of R^2, along with the value R^2_{adj}, differs significantly (compering Model 1 and Model 2), which may lead us to the inaccurate conclusion that Model 2 has stronger explanatory power than Model 1. When we excluded intercept, we faced two practical drawbacks and one of them especially noticeable in this case. Namely, regression packages usually change the way R-squared is computed when an overall intercept is not included by simply replacing SST with a total sum of squares that does not center y_i about its mean ($SST_0 = \sum_{i=1}^{n} y_i^2$), so-called uncentered R-squared. Moreover, the resulting R-squared, say $R_0^2 = 1 - SSR/SST_0$, is rarely suitable as a goodness of fit measure, as its high value is an artifact of non centering the total sum of squares in the calculation. However, when compering the measures related to "information criteria," AIC and SC(BIC), their identical value tell us that the out-of-the-sample performance of Model 1 and Model 2 is same, thus we would be indifferent if we chose between these two models.

In line with everything previously said, the higher value of AIC/BIC for Model 3, along with lower value of R^2_{adj}, makes it to be inferior choice in comparison to Model 1. We would end up with the exact conclusion when compering Model2 and Model3, as the values of AIC and BIC are the same as for Model1.

We obtain "information criteria" AIC and SC(BIC) using following formula which includes complexity of out-of-sample performance:

$$AIC/BIC = log(\frac{SSR}{n}) + \frac{m}{n}(k-1)$$

where for AIC $m = 2$ and for BIC $m = log(n)$.

BIC is especially useful tool/measure if we are building a model with aim of making predictions, as it penalizes increase in k more than the AIC.

2.6

```
op <- par(mfrow = c(1,2), mar = c(2.2, 2, 1.5, 1), mgp = c(1.2, .5, 0),
          cex.lab=0.7, cex.axis=0.7, cex.main=0.8)
hist(model1$residuals, breaks=20, xlab="Residuals", main="Histogram of Residuals")
qqnorm(model1$residuals,pch=19,cex=0.4)
qqline(model1$residuals, col="red")
```

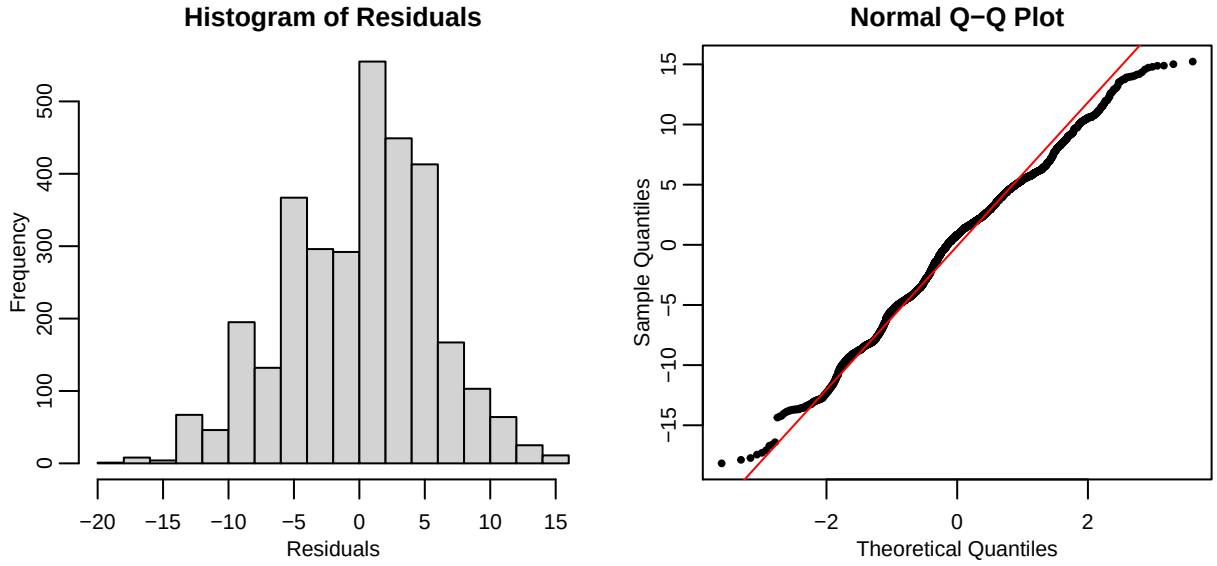

Figure 3: Model 1 - Histogram of Residuals and Normal Q-Q Plot

Based on the histogram in Figure 3, we see that the distribution of residuals appears to not follow normal distribution. Moreover, the Normal Q-Q plot in Figure 3 suggests the same conclusion as the dots lie on the line only in a small region of the between 0 and -2 standard deviation.

Jarque-Bera test for normality of residuals:

$$J = \frac{N-K}{6}(m_3^2 + \frac{1}{4}(m_4 - 3)^2)$$

H_0 : The errors follow a normal distribution.

H_A : The errors do not follow a normal distribution.

```
JB.test <- tseries::jarque.bera.test(model1$residuals)
JB.test.res <- matrix(c(JB.test$statistic,round(JB.test$parameter,0),JB.test$p.value),
                      dimnames=list(c("X-squared","df","p-value"),"Value"))
kable(round(JB.test.res,3), caption="Model 1 - Jarque Bera Test")
```

Table 7: Model 1 - Jarque Bera Test

	Value
X-squared	36.524
df	2.000
p-value	0.000

Based on the Jaque Bera Test, we can reject the null hypothesis H_0 that the residuals are normally distributed since $J > \chi^2_{2,0.95}$ and the p-value of J smaller than 5%. This means that residuals are not normally distributed.

Checking for correct model specification and homoskedacity:

```
op <- par(mar = c(2.5, 2.5, 2.5, 1), mgp = c(1.3, .5, 0), cex=0.7)
plot(model1$fitted.values, model1$residuals, pch=19, cex=0.4, xlab="Fitted values", ylab="Residuals")
abline(h=mean(model1$residuals), lty=2, col="red")
legend("topright", legend=c("Mean"), lty=2, col="red")
```

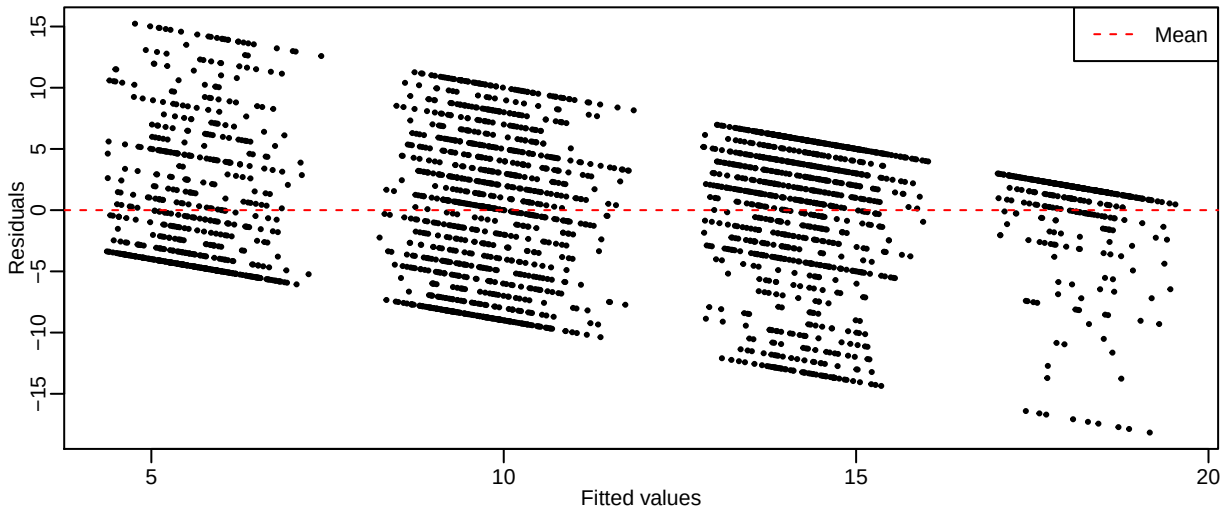

Figure 4: Model 1 - Fitted Values vs. Residuals

Based on the Fitted Values vs. Residuals plot in Figure 4, we see that the model specification is correct as $\mathbb{E}(u|X_1, ..., X_K) = 0$. However, the residuals seem to be dependent on the fitted values of the model which means that we need to check further the homoskedacity assumption.

Breusch-Pagan Test for homoskedacity assumption:

$$H_0 : \text{Homoskedasticity is present.}$$
$$H_A : \text{Heteroskedasticity is present.}$$

```
library(lmtest)
bp.test <- bptest(model1)
bp.test.res <- matrix(c(bp.test$statistic,bp.test$parameter,bp.test$p.value),
                      dimnames=list(c("BP","df","p-value"),"Value"))
kable(round(bp.test.res,3), caption="Model 1 - Breusch-Pagan test")
```

Table 8: Model 1 - Breusch-Pagan test

	Value
BP	88.102
df	9.000
p-value	0.000

Based on Breusch-Pagan test, we can reject the null hypothesis that homoskedasticity is present meaning that meaning that the model is heteroskedastic. This causes a violation of the standard assumption regarding homoskedasticity of OLS model.

Formally, this problem can be written as:

$$\mathbb{V}(u|X_1, ..., X_K) = \sigma^2 \Rightarrow \mathbb{V}(u_i|X_{1,i}, ..., X_{K,i}) = \sigma_i^2$$

Checking for no perfect multicollinearity assumption:

```
library(corrplot)
corrplot(cor(data[-3]), type="upper", diag=FALSE, mar = c(0, 0, 0, 0), tl.col="black",
         tl.cex=0.6, method="ellipse", addCoef.col='black', number.cex=0.5, cl.cex=0.6)
```

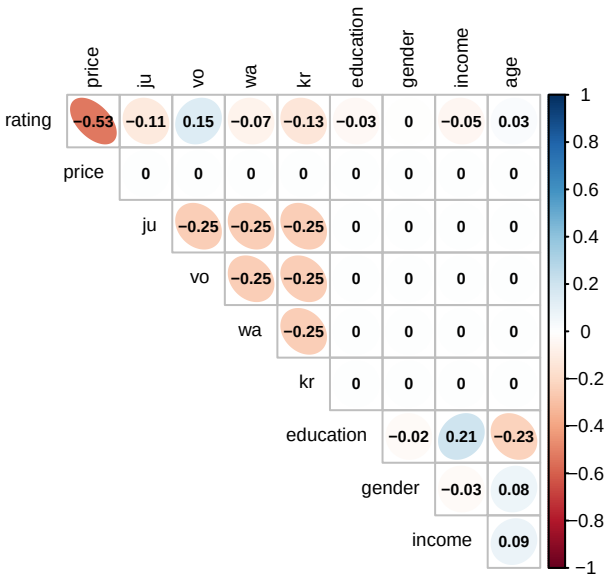

Figure 5: Model 1 - Correlation matrix between regressors

Based on correlation matrix in Figure 5, we see that the variables are not strongly correlated, hence, no perfect multicollinearity is present. We managed to ensure this by not including the variable *rq* and using it as baseline. If the variable *rq* was included along with the constant β_0, perfect multicollinearity would become an issue.

2.7

a) Adding interaction terms to the model:

First of all, we need to create a linear model with all possible interaction effects and select only the relevant ones. The relevant interaction terms are the ones which are a combination of numeric and factor variables meaning that they do not produce NAs and they do not include *rq* since it is included as the model's baseline.

```
model4 <- lm(rating ~ .*. - rq, data)
rel.vars <- model4$coefficients[which(is.na(model4$coefficients)==FALSE)]
rel.vars <- rel.vars[as.vector(grep(":", names(rel.vars)))]
rel.vars <- names(rel.vars[-as.vector(grep("rq", names(rel.vars)))])
rel.vars <- rel.vars[rel.vars!="price:age"]
```

Now, we can use the obtained relevant interaction terms and add them one by one to Model 1 in order to see how they impact the goodness-of-fit criteria, in other words, the R-squared and the Adjusted R-squared values.

```
models <- list()
inter.terms <- c()
r.squared <- c()
r.adj.squared <- c()
for (i in seq_along(rel.vars)) {
  form <- formula(paste(c("rating ~ . - rq", rel.vars[c(1,i)]), collapse="+"))
  model <- lm(form, data)
  inter.terms[i] <- rel.vars[i]
  r.squared[i] <- summary(model)$r.squared
  r.adj.squared[i] <- summary(model)$adj.r.squared
}
inter.res <- matrix(round(c(r.squared,r.adj.squared),3), ncol=2,
                    dimnames=list(inter.terms,c("R-squared","Adjusted R-squared")))
kable(inter.res, caption="Observing how the R-squared change by adding the given interation term")
```

Table 9: Observing how the R-squared change by adding the given interation term

	R-squared	Adjusted R-squared
price:ju	0.348	0.346
price:vo	0.348	0.346
price:wa	0.348	0.346
price:education	0.348	0.346
price:gender	0.349	0.346
price:income	0.349	0.346
ju:education	0.352	0.350
ju:gender	0.349	0.346
ju:income	0.348	0.346
ju:age	0.350	0.347
vo:education	0.348	0.346
vo:gender	0.349	0.346
vo:income	0.348	0.346
vo:age	0.351	0.349
wa:education	0.349	0.347

	R-squared	Adjusted R-squared
wa:gender	0.348	0.346
wa:income	0.348	0.346
wa:age	0.349	0.347
education:gender	0.349	0.347
education:income	0.349	0.347
education:age	0.348	0.346
gender:income	0.350	0.347
gender:age	0.351	0.348
income:age	0.348	0.346

As we can see in Table 9, none of the interaction terms significantly increased the R^2 and the Adjusted R^2 values after being added to the model.

b) Reporting the full output of the OLS Model 4:

```
model4.form <- formula(paste(c("rating ~ . - rq", rel.vars), collapse="+"))
model4 <- lm(model4.form, data)
```

See Model 4 in Table 10 for the full output.

c) Interpretation of the *education:income* coefficient:

For people without high school diploma or higher education, increase of income by 1 increases the rating by 0.546, $B_{\text{education}} = 0.546$. For people with high school diploma or higher education, increase of 1 in income decreases the rating by 0.447, $B_{\text{education}} + B_{\text{educationXincome}} = 0.546 - 0.993 = -0.447$. This means that the interaction effect decreases the impact of income and it becomes negative for people with higher education.

d) Improving Model 4 by omitting terms which are not helpful:

For improving Model 4, we will try to restrict it by excluding interaction terms which do not carry any significance at $\alpha = 0.1$

```
p.vals <- summary(model4)$coefficients[,4]
sig.p.vals <- p.vals[which(p.vals<0.1)]
sig.p.vals <- sig.p.vals[as.vector(grep(":", names(sig.p.vals)))]
```

Now, we are ready to include only significant interaction terms.

```
inter.terms <- formula(paste(c("rating ~ . -rq",names(sig.p.vals)), sep="+", collapse="+"))
model4.red <- lm(inter.terms, data)
stargazer(model4, model4.red, single.row=TRUE, header=FALSE, model.numbers=FALSE,
          column.labels = c("Model 4", "Restricted Model 4 "), title="Multiple OLS Regression Models wi
```

Based on the Restricted Model 4, we see that the R^2 and Adjusted R^2 values decreases by very small amount compared to the full model which means that we managed to exclude variables which are not helpful to explain the variation in our data. See Restricted Model 4 in Table 10 for the full output.

Table 10 was created using stargazer v.5.2.2 by Hlavac (2018).

Table 10: Multiple OLS Regression Models with interaction terms

	Dependent variable:	
	rating	
	Model 4	Restricted Model 4
price	−0.265*** (0.020)	−0.303*** (0.008)
ju	−3.264*** (1.193)	−3.509*** (0.815)
vo	2.708** (1.196)	1.208* (0.669)
wa	−4.544*** (1.186)	−4.504*** (0.669)
kr	−4.172*** (0.309)	−4.172*** (0.310)
education	−0.211 (0.890)	0.546* (0.303)
gender	1.979** (0.882)	1.845*** (0.506)
income	2.401** (0.937)	0.439 (0.414)
age	0.032** (0.014)	0.035*** (0.012)
price:ju	0.004 (0.022)	
price:vo	−0.033 (0.023)	
price:wa	−0.020 (0.022)	
price:education	−0.005 (0.017)	
price:gender	−0.024 (0.016)	
price:income	−0.025 (0.017)	
ju:education	−1.905*** (0.592)	−1.991*** (0.525)
ju:gender	−0.627 (0.542)	
ju:income	−0.459 (0.555)	
ju:age	0.027* (0.016)	0.024 (0.016)
vo:education	−0.588 (0.592)	
vo:gender	0.883 (0.542)	
vo:income	−0.620 (0.555)	
vo:age	−0.045*** (0.016)	−0.041*** (0.016)
wa:education	0.781 (0.592)	
wa:gender	0.526 (0.542)	
wa:income	−0.537 (0.555)	
wa:age	0.037** (0.016)	0.032** (0.016)
education:gender	0.667 (0.441)	
education:income	−1.118** (0.452)	−0.993** (0.435)
education:age	0.017 (0.013)	
gender:income	−0.955** (0.414)	−0.872** (0.402)
gender:age	−0.036*** (0.012)	−0.040*** (0.012)
income:age	−0.019 (0.013)	
Constant	22.179*** (1.005)	23.190*** (0.630)
Observations	3,195	3,195
R^2	0.365	0.360
Adjusted R^2	0.359	0.357
Residual Std. Error	5.530 (df = 3161)	5.538 (df = 3178)
F Statistic	55.166*** (df = 33; 3161)	111.816*** (df = 16; 3178)
Note:		*p<0.1; **p<0.05; ***p<0.01

3 Theory

3.1

In a general sence, SSR or Sum of Squared Residuals is a good measure of variation that is *not explained* by the given model. However, as models become more complex, the drawbacks of this measure are getting visible and more serious. The main disadvantage of this measure is that SSR automatically decreases when the number of explanatory variables increase. Consequently, it affects R^2 in an opposite way bearing in mind the relation between those measures: $R^2 = 1 - \frac{SSR}{SST}$. Thus, just relying on the values of the SSR or R^2, we would always choose to have as much variables in our model as we can. Unfortunately, by following this way of reasoning we could fall into the trap. Namely, we could face the *bias-variance trade off* which states that higher number of variables can definitely increase the precision of the model inside of the sample, but not necessarily out of the sample.

All things considered, one way to overcome the above mentioned problem is to include additional criteria based on which we can choose the model with *optimal* number of variables. BIC criterion is defined as following:

$$BIC = log(\frac{SSR}{N}) + \frac{m}{N}(k-1)$$

where $m = log(n)$

Therefore, the main purpose of BIC is to decide which model can best describe Y with respect to the number of variables.

In our particular example, considering the possibility of having different explanatory variables in given models, we can not state the same values for R^2 and BIC in both models, even though the response variable Y is the same. However, they will definitely be consistent, meaning that if the conslusion based on R^2 is that one model better explained Y than another, BIC criterion would state the same. The main reason is that models are different just in terms of explanatory variables but with *same amount of them* ($k_1 = k_2 = k$). To rephrase it, in this situation additional criterion will state the same as R^2 because BIC is valuable only when we have to go for one from two models, with respect to *different* values of k or explanatory variables.

To prove our statements mathematically, we should consider following formulas :

$$\text{Model 1: } R_1^2 = 1 - \frac{SSR_1}{SST}, \quad BIC_1 = log(\frac{SSR_1}{N}) + \frac{m}{N}(k-1)$$

$$\text{Model 2: } R_2^2 = 1 - \frac{SSR_2}{SST}, \quad BIC_2 = log(\frac{SSR_2}{N}) + \frac{m}{N}(k-1)$$

In view of models given in the task, both R^2 and BIC are identical in terms of all variables in equations except SSR. Therefore, the model with the smaller value of SSR or higher value of R^2 will have better in-sample performance. In addition, smaller value of SSR will have an influence on BIC as well, lowering the observed value. Having in mind that we want BIC value to be small in order to choose desired model, both measures lead us to the same, consistent conclusion: model with smaller SSR will be preferable using both criteria.

With the aim of proving what has been written before, we used the given dataset and run regression of rating on price and education, denoted Model1, and, as Model2, run regression of rating on age and income. The given requirements are met as we are having the same number of observations and explanatory variables.

```
lin.model1 <- lm(rating ~ price + education, data)
lin.model2 <- lm(rating ~ age + income, data)
```

Table 11: Proof for 3.1.

	Model 1	Model 2
R-squared	0.278	0.003
Adjusted R-squared	0.277	0.003
AIC	20382.827	21411.203
BIC	20407.105	21435.481

3.2

True: In the regression model

$$Y = \beta_0 + \beta_1 D + \beta_2 D^2 + u, \quad \mathbb{E}[u|D] = 0$$

a quadratic term involves squaring the predictor, where the predictor is dummy-coded variable.

Generally, when it comes to Quadratic regression models, the rule is that perfect multicolinearity assumption is not violated. The reason is because violation of the assumption implies the presence of exact linear relation between any predictors X_i and Xj when $i \neq j$. Consequently, even though $X_2 = X_1^2$ while $X_1 = X$ in example $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + u$, variable X_2 can not be written as a linear function of X_1 but rather non-linear, hence multicolinearity assumption still holds. However, in this particular example, when the explanatory variables are *dummy variables*, meaning that they can take either values 1 or 0, the general rule for Quadratic regression models is violated. Namely, having in mind that taking 0 or 1 to the power of 2 does not make any difference in comparison to non-squared dummy variable, this model should be considered as a Multiple regression model rather than the Quadratic. Thus, we will have our perfect multicolinearity assumption to be violated which bring us to the following consequences :

(1) The matrix $X'X$ is not invertible, hence the OLS estimator does not exist
(2) There are infinitely many parameter values γ having the same minimal sum of squared residuals($SSR(\gamma)$)
(3) The parameters in the regression model are not identified (not available: NA)

```
model5 <- lm(rating ~ education + I(education^2), data)
stargazer(model5, header=FALSE, single.row=TRUE, title="Model 5 - Multiple OLS Regression Model")
```

Table 12: Model 5 - Multiple OLS Regression Model

	Dependent variable:
	rating
education	−0.493* (0.255)
I(education^2)	
Constant	11.872*** (0.204)
Observations	3,195
R^2	0.001
Adjusted R^2	0.001
Residual Std. Error	6.903 (df = 3193)
F Statistic	3.745* (df = 1; 3193)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 12 was created using stargazer v.5.2.2 by Hlavac (2018).

With the aim of proving what has been written before, we used the given dataset and run regression of rating on education and where education is a dummy variable which takes value of one if consumer has 1 high school diploma or higher (see Table 12).

3.3

We are considering the following quadratic model:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + u$$

where $\mathbb{E}[u|X] = 0$.

The null hypothesis that states that the point where the effect of a marginal increase in X on the conditional expectation $\mathbb{E}[Y|X]$ changes its sign is 1 (in other words, **the value of vertex** equals one) would be written as follows:

$$H_0 : \frac{-\beta_1}{2\beta_2} = 1 \quad \longrightarrow \quad H_0 : \beta_1 + 2\beta_2 = 0$$
$$H_1 : \beta_1 + 2\beta_2 \neq 0$$

The distribution of $\beta_1 + 2\beta_2$ under the null is given by the following univariate normal distribution:

$$\beta_1 + 2\beta_2 | X \sim \mathcal{N}(0, \mathbb{V}(\beta_1 + 2\beta_2 | X))$$

where $Var(\hat{\beta}_1 + 2\hat{\beta}_2) = Var(\hat{\beta}_1) + 4Var(\hat{\beta}_2) + 4Cov(\hat{\beta}_1, \hat{\beta}_2)$.

The test statistic

$$t = \frac{\hat{\beta}_1 + 2\hat{\beta}_2}{se(\hat{\beta}_1 + 2\hat{\beta}_2)} | X \sim t_{df}$$

follows the t_{df} distribution with $df = (N - K - 1)$ when σ^2 is substituted by $\hat{\sigma}^2$.

However, the test statistic requires the standard error which is not reported in a standard regression.

The needed standard error is given by the square root of the variance:

$$se(\hat{\beta}_1 + 2\hat{\beta}_2) = \sqrt{Var(\hat{\beta}_1 + 2\hat{\beta}_2)}$$

The variance is given by:

$$Var(\hat{\beta}_1 + 2\hat{\beta}_2) = Var(\hat{\beta}_1) + 4Var(\hat{\beta}_2) + 4Cov(\hat{\beta}_1, \hat{\beta}_2)$$

The estimator of the standard error is thus given by:

$$se(\hat{\beta}_1 + 2\hat{\beta}_2) = \sqrt{[se(\hat{\beta}_1)]^2 + [se(2\hat{\beta}_2)]^2 + 2s_{12}}$$

The sample covariance s_{12} of the parameters is not given in a standard regression.To get the standard error of a linear combination you either need:

(1) To rewrite the model so that the needed standard error is given in a standard regression.
(2) A statistical software that computes the sample covariance between the parameters
(3) **A statistical software that allows direct testing of linear combinations**.

Under these circumstances, it is crucial to pay attention to the position of the vertex relative to the range of data X, meaning that if the range includes the vertex, the sign would change.

3.4

```
set.seed(1)

b0 <- 2
b1 <- -1.5
b2 <- 3
iters <- c(10)

models <- list()
for (i in seq_along(iters)) {
x <- runif(iters[i], -3, 3)
x1 <- x^2
u <- rnorm(iters[i], 0, 2^2)
y <- b0 + b1*x + b2*x1 + u
models[[i]] <- lm(y ~ x + x1) # or: models[[i]] <- lm(y ~ x + I(x^2))
}
```

```
stargazer(models[[1]], header=FALSE, single.row=TRUE, title="Simulated Regression Model")
```

Table 13: Simulated Regression Model

	Dependent variable:
	y
x	−1.538* (0.745)
x1	3.872*** (0.497)
Constant	−0.541 (2.073)
Observations	10
R^2	0.897
Adjusted R^2	0.867
Residual Std. Error	4.113 (df = 7)
F Statistic	30.350*** (df = 2; 7)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 13 was created using stargazer v.5.2.2 by Hlavac (2018).

```
kable(t(range(x)), col.names = c("Min", "Max"), caption = "The range of explanatory variable")
```

Table 14: The range of explanatory variable

Min	Max
-2.629282	2.668052

Note that the vertex is in the range of explanatory variable.

```
kable(linearHypothesis(models[[1]], c("x=-2*x1")),
      caption = "Testing the null hypothesis:
      The vertex of a quadratic equation equals one.")
```

Table 15: Testing the null hypothesis: The vertex of a quadratic equation equals one.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
8	663.4941	NA	NA	NA	NA
7	118.4034	1	545.0907	32.22572	0.0007532

The answer is p-value = 0.0007532, which fails to reject the null hypothesis. Thus, the conjecture that a point where the effect of a marginal increase in X on the conditional expectation $\mathbb{E}[Y|X]$ changes its sign is 1 is not supported by the data.

References

Hlavac, Marek. 2018. *Stargazer: Well-Formatted Regression and Summary Statistics Tables.* Bratislava, Slovakia: Central European Labour Studies Institute (CELSI). https://CRAN.R-project.org/package=stargazer.