Tampereen yliopisto

ASSIGNMENT 3

Vilho Koivunen
151153479
Tijo.220

1. For this report, R was used as the analysis tool. The dependent variable is health care costs, and the independent variables are region, gender, age, marital status, alcohol consumption, cigarette consumption and exercise. We will be using multiple linear regression since the dependent variable is continuous and since we have multiple independent variables to be analysed.

2. In data preprocessing, 4 rows were removed for having missing values: two in cigarette data and two in exercise data. In addition, the survey date column was removed to exclude redundant data.

   colSums(is.na(data))

id	*svdate*	*region*	*Gender*	*age*	*marit*	*alco*	*cigs*	*exer*	*costs*
0	*0*	*0*	*0*	*0*	*0*	*0*	*2*	*2*	*0*

   > data <- na.omit(data)

   > data <- data[, -c(2)]

   For dealing with outliers, the first step was examining the minimum and maximum ranges of each variable. Since we only have 278 datapoints left, the goal is to be conservative in outlier removal to retain as much data as possible. It was detected that there is one considerably older participant in the data of 105 years and one person who consumes relatively a considerable amount of alcohol.

   In more scientific approach IQR or interquartile range was used to detect outliers in all dependent variables, using the code (example with age):

   > age = data$age

   > iqr_age = IQR(age)

   *> age_lower <- quantile(age, 0.25) - 1.5 * iqr_age*

   *> age_upper <- quantile(age, 0.75) + 1.5 * iqr_age*

   > age_ourliers <- age[age < age_lower | age > age_upper]

   > age_outliers <- age[age < age_lower | age > age_upper]

   > print(age_outliers)

   [1] 91.14156 93.55578 91.60099 91.23666 90.14151 90.99018 90.35706 90.08974 91.08080 91.85396 93.13486 104.89480

   [13] 90.39875

13 highest values for age were selected. For the purpose of conserving data, only the datapoint with the substantially higher age will be removed.

The same method (IQR) is used for all independent variables. For alcohol, the model considered the amounts >24 (16 in total) to be potential outliers. Again, only the substantially higher value was removed to conserve data. For cigarettes, 4 top values were detected as potential outliers and in exercise, the bottom two and top two were detected as potential outliers. Since these are the examined variables in this study, all these potential outliers are removed.

For checking the skewness of the data, the function skewness() is used from package “e1071”.

> columns <- c("age", "alco", "cigs", "exer")

> skew_values <- sapply(data_cleaned[columns], skewness)

> print(skew_values)

age alco cigs exer

0.87375969 0.93233712 0.73037659 0.08431107

The alcohol, age and cigarette columns seem to have a positive skew. Log transformation is done in attempt to normalise data. A constant is applied to log_cigs to avoid errors in log(0).

> colnames(data_cleaned)[colnames(data_cleaned) == "age_exer"] <- "exer_log"

> columns <- c("age_log", "cigs_log", "exer_log")

> log_skew <- sapply(data_cleaned[columns], skewness)

> print(log_skew)

age_log cigs_log exer_log

0.7128304 NaN -0.8352016

> data_cleaned$cigs_log <- NULL

> data_cleaned$cigs_log <- log(cigs + 0.001)

Error: object 'cigs' not found

> data_cleaned$cigs_log <- log(data_cleaned$cigs +1)

> columns <- c("age_log", "cigs_log", "exer_log")

```
> log_skew <- sapply(data_cleaned[columns], skewness)
> print(log_skew)
   age_log   cigs_log   exer_log
 0.7128304 -0.5323606 -0.8352016
```

The log transformation did not help considerably with skewed values, so the regular values will be used.

Using the linear model in R, here is the summary:

```
Call:
lm(formula = costs ~ exer + age + cigs + alco + factor(region) +
    factor(Gender) + marit, data = data_cleaned)

Residuals:
    Min      1Q  Median      3Q     Max
-2878.2  -983.5  -171.8   792.2  5245.0

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2718.52    1507.73  -1.803  0.07253 .
exer             -229.76      57.26  -4.013 7.84e-05 ***
age               107.30      17.15   6.258 1.59e-09 ***
cigs              121.97      42.16   2.893  0.00414 **
alco               73.93      23.68   3.122  0.00200 **
factor(region)2  -532.32     299.43  -1.778  0.07661 .
factor(region)3  -475.73     285.21  -1.668  0.09652 .
factor(region)4  -539.26     274.88  -1.962  0.05085 .
factor(region)5  -477.31     265.33  -1.799  0.07319 .
factor(Gender)1   563.17     182.45   3.087  0.00224 **
marit            -113.91     108.60  -1.049  0.29517
```

---

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Residual standard error: 1482 on 261 degrees of freedom

Multiple R-squared: 0.3977, Adjusted R-squared: 0.3747

F-statistic: 17.24 on 10 and 261 DF, p-value: < 2.2e-16

Based on the model, age, alcohol consumption, exercise, gender 1, cigarettes correlate with healthcare costs, and cigarettes also seem to have a connection. Next, let's do a simplified linear model without factor of region, and without marital status, since it doesn't seem to have an effect.

Call:

lm(formula = costs ~ exer + age + cigs + alco + factor(Gender),

data = data_cleaned)

Residuals:

Min 1Q Median 3Q Max

-3018.2 -994.6 -196.9 712.7 5520.4

Coefficients:

Estimate Std. Error t value Pr(>|t|)

*(Intercept) -3359.81 1459.12 -2.303 0.02207 **

*exer -240.19 57.31 -4.191 3.78e-05 ***

*age 108.28 16.91 6.402 6.90e-10 ***

*cigs 109.81 41.91 2.620 0.00930 **

*alco 78.42 23.59 3.324 0.00101 **

*factor(Gender)1 565.38 182.91 3.091 0.00221 **

---

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Residual standard error: 1488 on 266 degrees of freedom

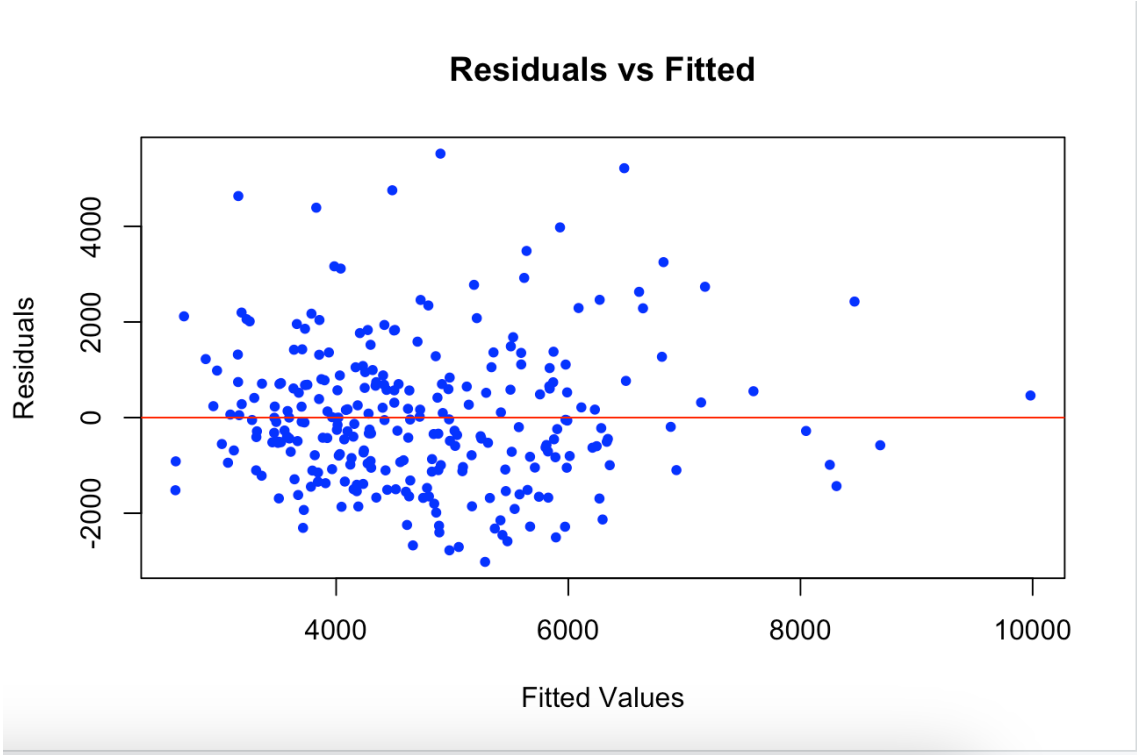
Multiple R-squared: 0.3811, Adjusted R-squared: 0.3695

F-statistic: 32.76 on 5 and 266 DF, p-value: < 2.2e-16

The results of the linear model make sense and align with the hypothesis: exercise reduces healthcare costs significantly, and age, cigarette and alcohol consumption increase healthcare costs. The factor of gender 1 implies that being gender 1 increases healthcare costs significantly, which is surprising.

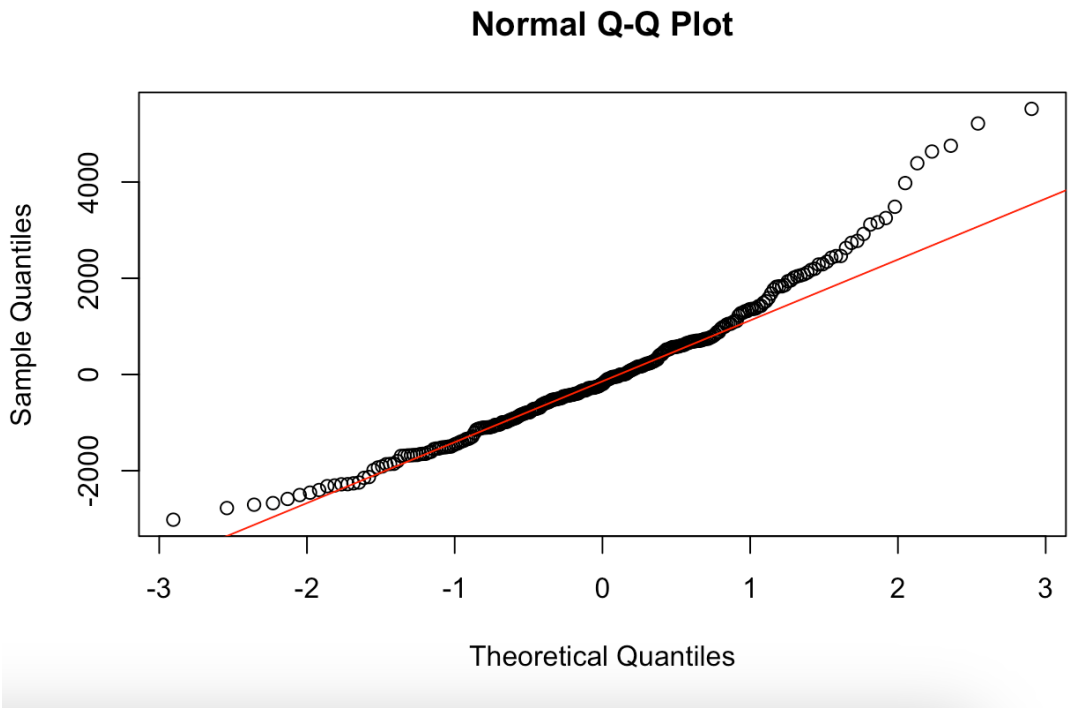
3. Next, an assumption check will be done. First, a linearity check will be done using residual errors:

```
plot(model$fitted.values, model$residuals,
+     main = "Residuals vs Fitted",
+     xlab = "Fitted Values",
+     ylab = "Residuals",
+     pch = 20, col = "blue")
> abline(h = 0, col = "red")
```



There are is no significant discrepancy in values and no patterns to be found: the data acts as linear. Although, there are a few datapoints with significant positive

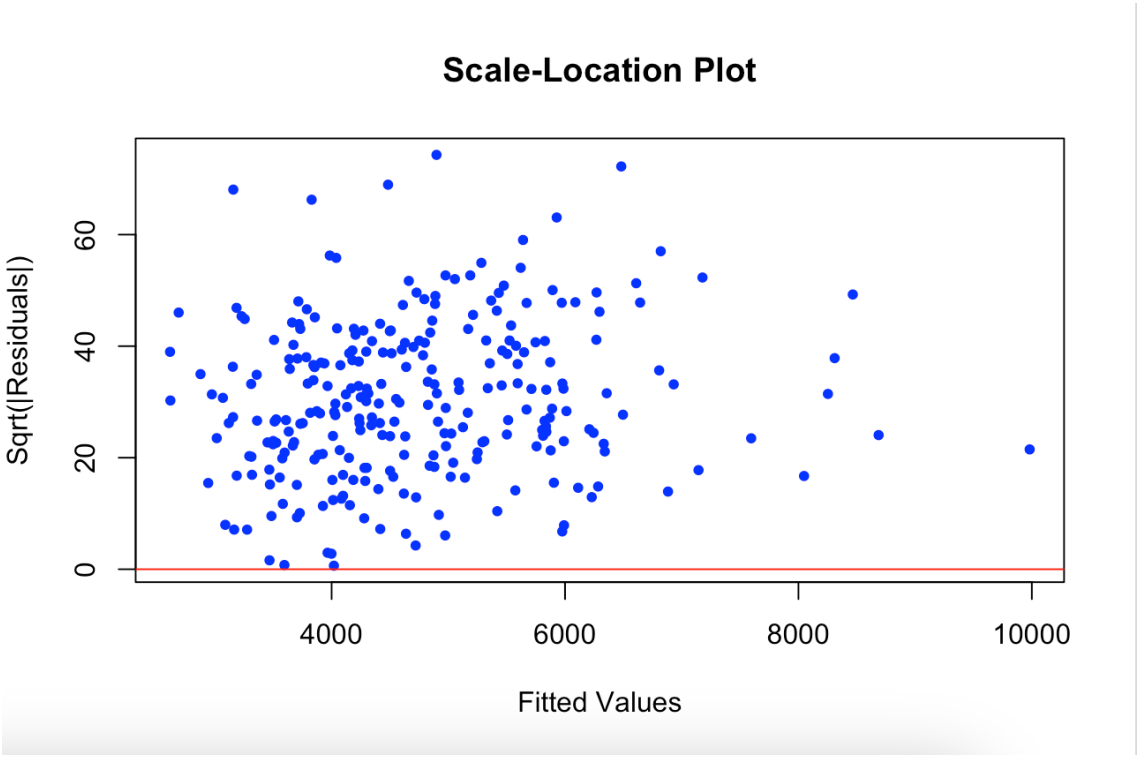
residuals, and the plot reveals the presence of a few outliers in the healthcare cost dimension in the upper end, that could have been removed. This can also been seen in the normal q-q plot:



There is a clear indication that too many outliers were left in the data, since the residuals do not follow a normal distribution in the extremes. This might be why we saw a surprising result in the model in the case of gender.

Next, with a scale-location plot, we can observe if there is constant variance in our residuals:

```
plot(model$fitted.values, sqrt(abs(model$residuals)),
+     main = "Scale-Location Plot",
+     xlab = "Fitted Values",
+     ylab = "Sqrt(|Residuals|)",
+     pch = 20, col = "blue")
> abline(h = 0, col = "red")
```



There seems to be approximately constant variance within the residuals, with a couple of datapoints with relatively high values.

4. The results show that cigarette consumption and exercise both have big effects on healthcare costs. For cigarette consumption, the number 109.81 means that each extra unit of smoking raises healthcare costs by around 109.81 units, while keeping other factors the same. This increase is important ($p = 0.0093$) and shows how smoking adds to health problems and makes treatments more expensive. Smoking causes illnesses like lung disease, heart problems, and cancer, which are very costly to treat. This suggests that reducing smoking could lower healthcare costs a lot.

   Exercise, on the other hand, helps lower healthcare costs. The number -240.19 means that each extra unit of exercise reduces healthcare costs by about 240.19 units. This effect is very strong and shows that exercise is good for health. Regular physical activity helps prevent diseases like diabetes and high blood pressure, which are expensive to treat. These results suggest that promoting exercise is a good way to make people healthier and reduce medical costs.

5. The model has an R-squared value of 0.3811, meaning it explains about 38% of the variation in healthcare costs, which isn't very high. The Adjusted R-squared

is slightly lower at 0.3695, showing that adding the predictors didn't overfit the model much. The p-value for the model is highly significant, meaning the predictors collectively have a strong relationship with healthcare costs. But the relatively low R-squared means other variables, not in the model, might better explain healthcare costs. Overall, the predictive power of the model is moderate.