Homework7

Dima Mikhaylov

10/21/2021

1. Questions based on `swiss` dataset:

```
library(datasets)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.4     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.3     ✓ stringr 1.4.0
## ✓ readr   2.0.1     ✓ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(corrplot)
```

```
## corrplot 0.90 loaded
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
data(swiss)
head(swiss)
```

```
##              Fertility Agriculture Examination Education Catholic
## Courtelary        80.2        17.0          15        12     9.96
## Delemont          83.1        45.1           6         9    84.84
## Franches-Mnt      92.5        39.7           5         5    93.40
## Moutier           85.8        36.5          12         7    33.77
## Neuveville        76.9        43.5          17        15     5.16
## Porrentruy        76.1        35.3           9         7    90.57
##              Infant.Mortality
## Courtelary               22.2
## Delemont                 22.2
## Franches-Mnt             20.2
## Moutier                  20.3
## Neuveville               20.6
## Porrentruy               26.6
```

a. In the previous homework, you fit a model with the fertility measure as the response variable and used all the other variables as predictors. Now, consider a simpler model, using only the last three variables as predictors: Education, Catholic, and Infant.Mortality. Carry out an appropriate hypothesis test to assess which of these two models should be used. State the null and alternative hypotheses, find the relevant test statistic, p-value, and state a conclusion in context. (For practice, try to calculate the test statistic by hand.)

Below is the original full model:

```
full_lm.fit <- lm(Fertility ~ ., data=swiss)
summary(full_lm.fit)
```

```
##
## Call:
## lm(formula = Fertility ~ ., data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2743  -5.2617   0.5032   4.1198  15.3213
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
## Agriculture      -0.17211    0.07030  -2.448  0.01873 *
## Examination      -0.25801    0.25388  -1.016  0.31546
## Education        -0.87094    0.18303  -4.758 2.43e-05 ***
## Catholic          0.10412    0.03526   2.953  0.00519 **
## Infant.Mortality  1.07705    0.38172   2.822  0.00734 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.165 on 41 degrees of freedom
## Multiple R-squared:  0.7067, Adjusted R-squared:  0.671
## F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

This is reduced subset model:

```
reduced_lm.fit <- lm(Fertility ~ Education + Catholic + Infant.Mortality, data=swiss)
summary(reduced_lm.fit)
```

```
##
## Call:
## lm(formula = Fertility ~ Education + Catholic + Infant.Mortality,
##     data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4781  -5.4403  -0.5143   4.1568  15.1187
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      48.67707    7.91908   6.147 2.24e-07 ***
## Education        -0.75925    0.11680  -6.501 6.83e-08 ***
## Catholic          0.09607    0.02722   3.530  0.00101 **
## Infant.Mortality  1.29615    0.38699   3.349  0.00169 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.505 on 43 degrees of freedom
## Multiple R-squared:  0.6625, Adjusted R-squared:  0.639
## F-statistic: 28.14 on 3 and 43 DF,  p-value: 3.15e-10
```

In order to compare these two models one needs to test H0: B(Agriculture) = B(Examination) = 0 and alternative Ha: B(Agriculture) and B(Examination) together are not 0. Comparing these two models with partial F test below:

```
anova(reduced_lm.fit, full_lm.fit)
```

```
## Analysis of Variance Table
##
## Model 1: Fertility ~ Education + Catholic + Infant.Mortality
## Model 2: Fertility ~ Agriculture + Examination + Education + Catholic +
##     Infant.Mortality
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     43 2422.2
## 2     41 2105.0  2     317.2 3.0891 0.05628 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on these results p-value is not statistically significant at 5% and thus we fail to reject H0. In other words, the data suggests that we can drop variables Agriculture and Examinbation. Below is additional calculations “by hand”:

```
test_lm.fit <- lm(Fertility ~ Education + Catholic + Infant.Mortality + ., data=swiss)
anova(test_lm.fit)
```

```
## Analysis of Variance Table
##
## Response: Fertility
##                  Df Sum Sq Mean Sq F value    Pr(>F)
## Education         1 3162.7  3162.7 61.6004 1.073e-09 ***
## Catholic          1  961.1   961.1 18.7187 9.478e-05 ***
## Infant.Mortality  1  631.9   631.9 12.3080  0.001109 **
## Agriculture       1  264.2   264.2  5.1454  0.028641 *
## Examination       1   53.0    53.0  1.0328  0.315462
## Residuals        41 2105.0    51.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Checking Sum of Squares and F-stat:

```
F0_stat = ((264.2+53)/2)/(2105/41)
F0_stat
```

```
## [1] 3.089121
```

Checking significance of the F-stat by computing p-value:

```
1-pf(F0_stat, 2, 41)
```

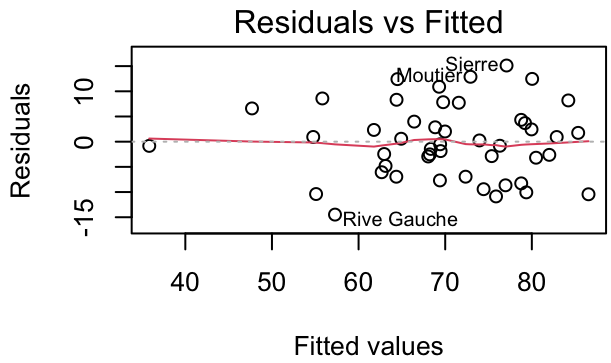
```
## [1] 0.05628116
```

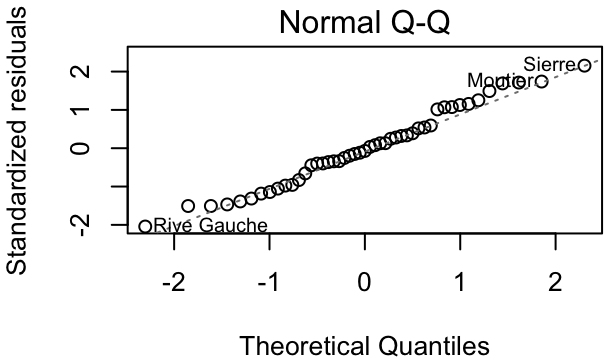
Conclusion: based on these manual calculations one arrives to the similar conclusion - failing to reject H0 at 95% significance and therefore Agriculture and Examinbation can be dropped from the model.

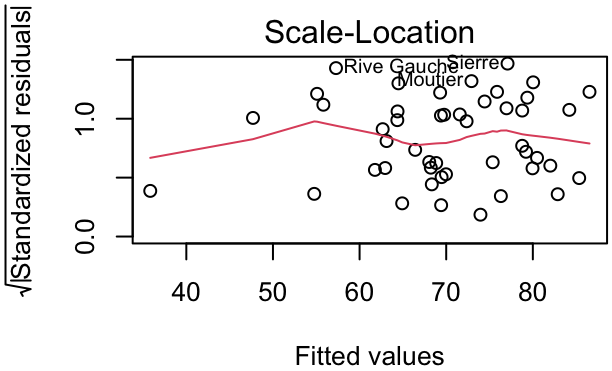
b. For the model you decide to use from part 1a, assess if the regression assumptions are met.

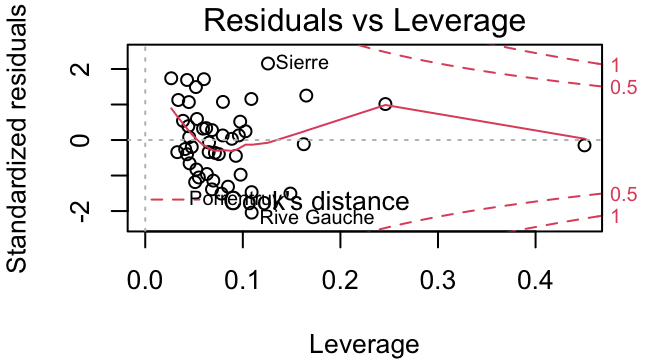
Overall, linear regression model assumptions are met.

```
par(mfrow=c(2,2))
plot(reduced_lm.fit)
```









Focusing on the Residuals plot, variances seems to be constant and overall mean tends to be zero. Moreover, from the QQ plot, theoretical quantiles were matched. Standartized residuals are well distributed and there are not to many high leverage points (still note presence of leverage from "Rive Gauche", "Moutier", and "Sierre" points)

2. Questions based on the data from 113 hospitals.

a. Based on the t statistics, which predictors appear to be insignificant?

Based on the t statistics only "Stay" and "Culture" slopes are significant. The other three predictors, "Age", "Census", and "Beds" appear to be insignificant (given the presence of the other predictors).

b. Based on your answer in part 2a, carry out the appropriate hypothesis test to see if those predictors can be dropped from the multiple regression model. Show all steps, including your null and alternative hypotheses, the corresponding test statistic, p-value, critical value, and your conclusion in context.

HO: B("Age") = B("Census") = B("Beds") = 0 and Ha: B("Age"), B("Census"), B("Beds") not 0, given the presence of the other predictors.

Compute partial F statistic:

```
F_stat = ((0.136 + 5.101 + 0.028) / 3) / (105.413 / 107)
F_stat
```

```
## [1] 1.781422
```

Compute p-value of F statistic:

```
1-pf(F_stat, 3, 107)
```

```
## [1] 0.1550925
```

Based on these results, we fail to reject H0 at 95% significance and thus we can drop these predictors from the model.

c. Suppose we want to decide between two potential models:

- Model 1: using x1, x2, x3, x4 as the predictors for InfctRsk
- Model 2: using x1, x2 as the predictors for InfctRsk Carry out the appropriate hypothesis test to decide which of models 1 or 2 should be used. Be sure to show all steps in your hypothesis test.

Does removing x3 and x4 from the full model (`InfctRsk~x1+x2+x3+x4`) result in a statistically significant increase in the sum of squared errors, i.e. residual sum of squares? In other words, does removing these two variables result in increased error and thus decreased predictive power of the model? All esle equal, simplier model should always be preferable.

H0: there is no difference in SSE of full and reduced models, i.e. models do not significantly differ.

Ha: full model (`InfctRsk~x1+x2+x3+x4`) has significantly lower SSE than the reduced model(`InfctRsk~x1+x2`)

F-stat - (SSres(reduced)-SSres(full)/(chage in # of params)) / MSres(full)

- SSr(x1) = 57.305
- SSr(x2|x1) = 33.397
- SSr(x3|x1, x2) = 0.136
- SSr(x4|x1, x2, x3) = 5.101

```
Original_SSt <- 57.305 + 33.397 + 0.136 +  5.101 + 0.028 + 105.413 # Total sum of squared from 5-predictor model

Model2_SSr <- 57.305 + 33.397 # reduced Model 2 sum or squares, explained by the reduced subset
Model1_SSr <- 57.305 + 33.397 + 0.136 +  5.101 # full Model 1 sum or squares, explained by the full model

Model2_SSres <- Original_SSt - Model2_SSr # Res. - variations not explained by Model 2, the reduced subset
Model1_SSres <- Original_SSt  - Model1_SSr # Res. - variations not explained by Model 1, the full model

#F_stat = ((Model2_SSres-Model1_SSres)/2) / ((0.028 + 105.413)/107) #Alternative formulae
F_stat = ((Model1_SSr-Model2_SSr)/2) / ((0.028 + 105.413)/107)

1-pf(F_stat, 2, 107)
```

```
## [1] 0.07477035
```

Concslusion: p-value is high and thus fail to reject H0 at 95% significance, there is no difference in SSE of full and reduced models, and thus we can drop these predictors from the model. In other words, epthasising simplicity we should fo with Model 2: using x1, x2 as the predictors for InfctRsk

3. Questions based on a data set seen in Homework Set 4.

Explain how this output indicates the presence of multicollinearity in this regression model.

The output suggests the presence of collinearity based on the following tests:

- The individual t tests suggests none of the two predictors are significant (given the presence of the other predictors)
- The F test suggests that overall the model is useful in predicting the response because p-value of F-stats is very small.
- Business logic suggests that the two predictors are strongly correlated and thus removing one of them or combining both into one variable will help increase significance of t statistics.