Homework6

Dima Mikhaylov

10/12/2021

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)
```

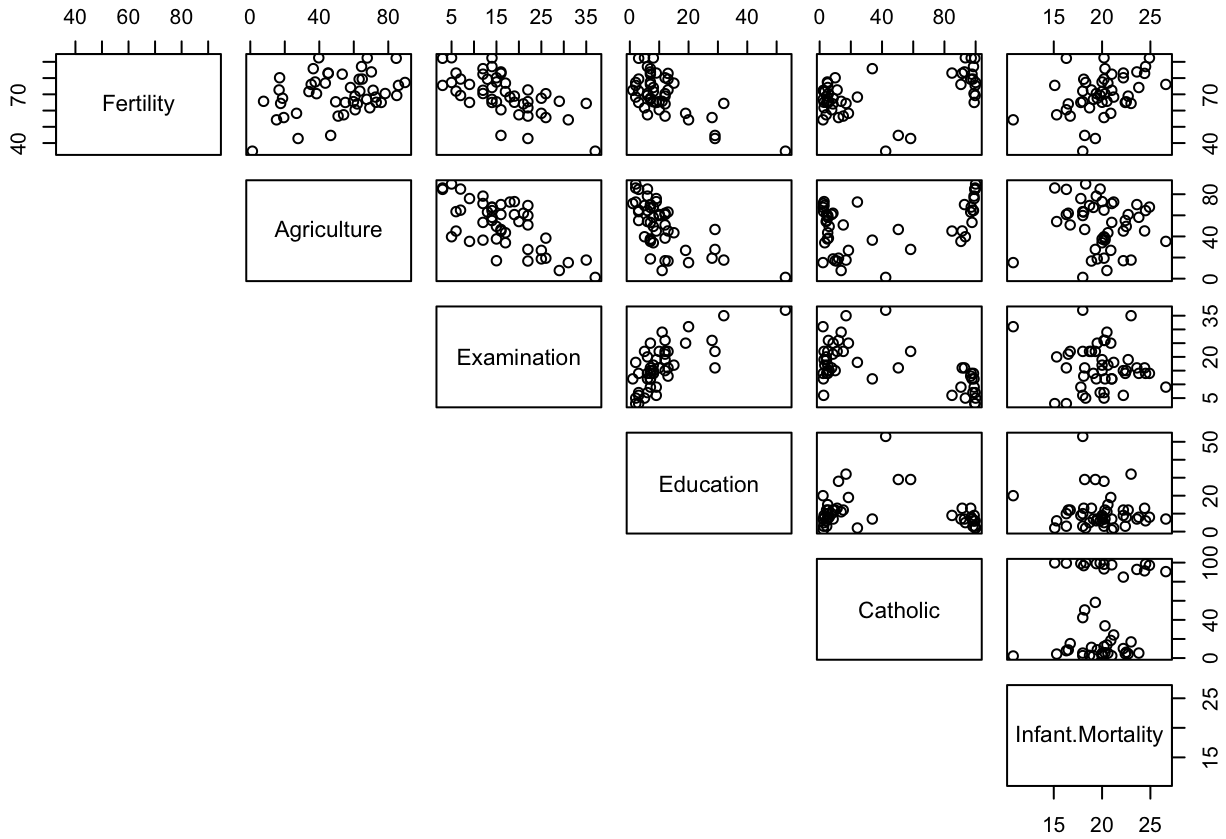
```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

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

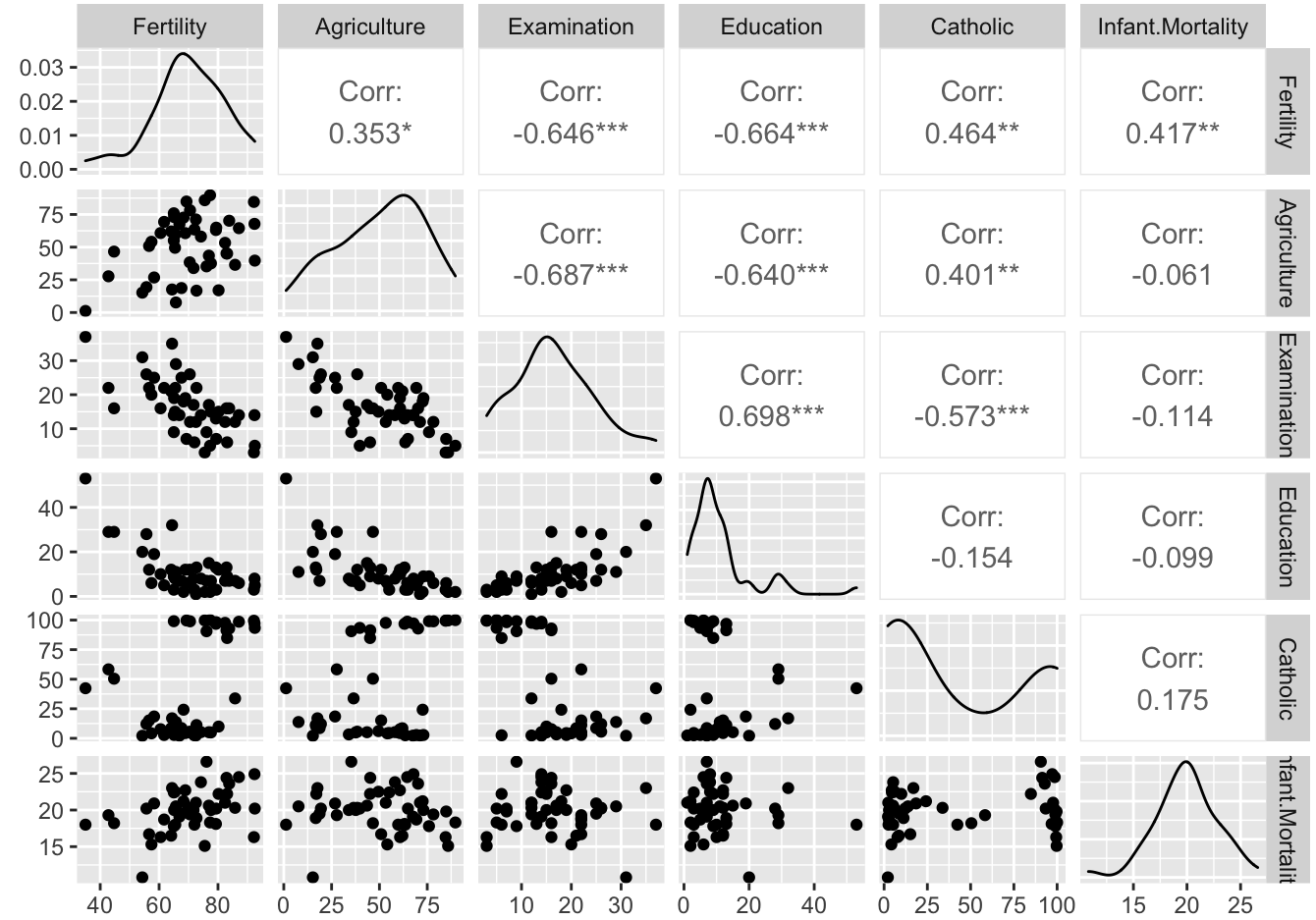
```
##              Fertility Agriculture Examination Education Catholic
## Courtelary        80.2        17.0          15        12     9.96
## Delemont          83.1        45.1           6         9    84.84
## Franches-Mnt      92.5        39.7           5         5    93.40
## Moutier           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. Create a scatterplot matrix and find the correlation between all pairs of variables for this data set.

```
pairs(swiss, lower.panel = NULL)
```




```
ggpairs(swiss)
```



Answer the following questions

based on the output:

i. Which predictors appear to be linearly related to the fertility measure?

Comment: it seems that some predictors may be linearly related to the fertility, for example variables Argiculture, Examination, Education, and Catholic.

ii. Do you notice if any of the predictors are highly correlated with one another? If so, which ones?

```
round(cor(swiss), 2)
```

```
##                  Fertility Agriculture Examination Education Catholic
## Fertility             1.00        0.35       -0.65     -0.66     0.46
## Agriculture           0.35        1.00       -0.69     -0.64     0.40
## Examination          -0.65       -0.69        1.00      0.70    -0.57
## Education            -0.66       -0.64        0.70      1.00    -0.15
## Catholic              0.46        0.40       -0.57     -0.15     1.00
## Infant.Mortality      0.42       -0.06       -0.11     -0.10     0.18
##                  Infant.Mortality
## Fertility                    0.42
## Agriculture                 -0.06
## Examination                 -0.11
## Education                   -0.10
## Catholic                     0.18
## Infant.Mortality             1.00
```

Comment: Examination and Education are more (negatively) correlated (r=-0.64 and -0.66, respectively). Catholic and Infant.Mortality are moderately (positively) correlated (r=0.46 and r=0.42, respectively). Argiculture is somewhat (positively) correlated (r=0.35).

b. Fit a multiple linear regression with the fertiliy measure as the response variable and all the other variables as predictors. Use the summary() function to obtain the estimated coefficients and results from the various hypothesis tests for this model.

```
result <- lm(Fertility ~ Agriculture + Examination + Education + Catholic + Infant.Mortality, data=swiss)
summary(result)
```

```
##
## Call:
## lm(formula = Fertility ~ Agriculture + Examination + Education +
##     Catholic + Infant.Mortality, 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
```

i. What is being tested by the ANOVA F statistic? What is the relevant conclusion in context?

Comment: F statistic shows the overall significance of the model, i.e. if all the predictors taken together explain the dependent variable. In other words, H0: B1 = B2 = B3 = B4 = B5 = 0 and Ha: at least one of the slopes is not zero.

ii. Look at the numerical values of the estimated slopes as well as their p-values. Do they seem to agree with or contradict with what you had written in your answer to part 1a? Briefly explain what do you think is going on here.

Comment: Examination variable, although correlated, appears to be not significant. Having said that, Education was correlated and also appears to be significant. Moreover, Agriculture, although not very correlated, appears to be statistically significant. This behaviour can be a result of correlated variables preventing each other from having high significance in MLR, when taken together.

2. Questions based on the data from 113 hospitals.

a. What is the value of the estimated coefficient of the variable Stay? Write a sentence that interprets this value.

Comment: estimate of slope for Stay is 0.24, this is the expected change in the response variable per unit of change of Slope holding all over regressor variables constant.

b. Derive the test statistic, p-value, and critical value for the variable Age. What null and alternative hypotheses are being evaluated with this test statistic? What conclusion should we make about the variable Age?

Comment: using two-tailed test to check if the value is different from zero, H0: B=0 and Ha: B!=0. Given the large p value of 0.5367937 (not significant) we can't reject H0 that the slope for Age is zero.

```
test_stats <- -0.014071 / 0.022708
df <- 108
p_value <- pt(test_stats, df) * 2
cat("The p_value is", p_value)
```

```
## The p_value is 0.5367937
```

c. A classmate states: "The variable Age is not linearly related to the predicted infection risk." Do you agree with your classmate's statement? Briefly explain.

Comment: no, I disagree. Form the analysis above we can only conclude that given all other regressors, the slope of Age is statistically indistinguishable from zero. It may still show a very significant linear relationship, even when measured by estimating the slope coefficient, when regressing with other predictors or solo.

d. Using the Bonferroni method, construct 95% joint confidence intervals for β1, β2, and β3. Calculating at least 95% confidence with the formula: Bi +/- [t (1-a/2g); n-p] * SE_Bi

```
g = 3 # number of intervals
a = 0.05 # original alpha
df = 108 # degrees of freedom
multiplier = qt(1-(a/2*g), df)
B1 = 0.23 # estimate of variable "Stay"
SE_B1 = 0.06 # error of estimate B1
B2 = -0.01 # estimate of variable "Age"
SE_B2 = 0.02 # error of estimate B2
B3 = 0.02 # estimate of variable "Xray"
SE_B3 = 0.01 # error of estimate B3
cat("B1 95% interval is [", B1-multiplier*SE_B1,",", B1+multiplier*SE_B1, "]")
```

```
## B1 95% interval is [ 0.1430094 , 0.3169906 ]
```

```
cat("B2 95% interval is [", B2-multiplier*SE_B2,",", B2+multiplier*SE_B2, "]")
```

```
## B2 95% interval is [ -0.03899687 , 0.01899687 ]
```

```
cat("B3 95% interval is [", B3-multiplier*SE_B3,",", B3+multiplier*SE_B3, "]")
```

```
## B3 95% interval is [ 0.005501566 , 0.03449843 ]
```

e. Fill in the values for the ANOVA table for this regression model.

ANOVA table:

Degrees of freedome

Regression_df = k = 4 Error_df = n - k - 1 = 113 - 4 - 1 = 108 Total_df = n - 1 = 113 - 1 = 112

Sum of Squares

Regression_SS = Regression_MS* Regression_df = 21.1561 * 4 = 84.6244 Error_SS = Error_df * (Residual_std) ** 2 = 108 * ((1.04) ** 2 = 116.8128 Total_SS = 84.6244 + 116.8128 = 201.4372

Mean Square

Regression_MS = F_stats * Error_MS = 19.56 * 1.0816 = 21.1561 Error_MS = Error_SS / Error_df = 116.8128 / 108 = 1.0816 F-stats = Regression_MS / Error_MS = 21.1561 / 1.0816 = 19.56 (check)

f. What is the R2 for this model? Write a sentence that interprets this value in context.

From the missing ANOVA table

R-squared is around 42% what indicates thath nearly half of variations in dependent variable were explained by 4 independent variables in this model.

R2 = Regression_SS / Total_SS = 84.6244 / 201.4372 = 0.4201031 or 42%

g. What is the R2_adj for this model?

R2_adj would always be smaller than R2 due to a penalty of using additional variables:

R2_adj = 1-(1-R2) * (n-1)/(n-k-1) = 1 - (1-0.4201031) * (113-1)) / (113-4-1) = 0.3986254 or 40%

3. Questions based on the data from 55 college students.

A classmate points out that there appears to be a contradiction in the R output, namely, while the ANOVA F statistic is significant, the t statistics for both predictors are insignificant. Is your classmate's concern warranted? Briefly explain.

Comment: all individual slope coefficients (after being regressed together) may be statistically insignificant, as measured by individual t statistics, when the overall predictive power of the model, measured by F stats, is high. This is due to the fact that F stats uses H0 that at ALL the slopes are zeros at the same time. If at least one slope is not zero, the model could be VERY useful in predicting the response. At the same time, individual insignificant slope only means that it can be droped in the presence of other predictors, but this, however, will change the estimates of the remaining slopes.