HW 07

Matt Scheffel

2022-10-17

Matt Scheffel

mcs9ff

Module 7 HW

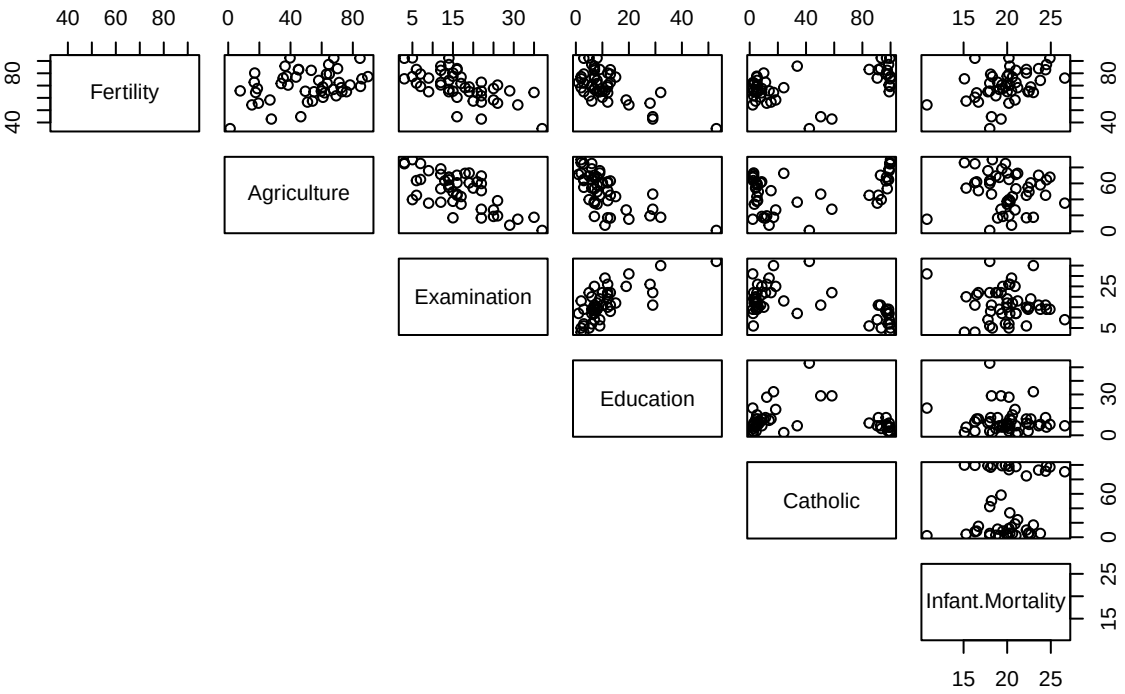
```
library(datasets)
data(swiss)
head(swiss)
```

```
##              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
```

```
?swiss
```

1A

```
pairs(swiss, lower.panel = NULL, main="Scatterplot of All Variables")
```



i. The Examination and Education predictors appear to be linearly related to the Fertility measure. (Perhaps an argument could be made for Agriculture.)

ii. Yes, some of the predictors appear to be highly correlated with one another:

Examinations and Education, Examinations and Agriculture, Agriculture and Education

1B

```
fertilityMLR<-lm(swiss$Fertility~swiss$Agriculture+swiss$Examination+swiss$Education+swiss$Catholic+swis

summary(fertilityMLR)
```

```
##
## Call:
## lm(formula = swiss$Fertility ~ swiss$Agriculture + swiss$Examination +
##     swiss$Education + swiss$Catholic + swiss$Infant.Mortality)
##
## 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 ***
## swiss$Agriculture        -0.17211    0.07030  -2.448  0.01873 *
## swiss$Examination        -0.25801    0.25388  -1.016  0.31546
## swiss$Education          -0.87094    0.18303  -4.758 2.43e-05 ***
## swiss$Catholic            0.10412    0.03526   2.953  0.00519 **
## swiss$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. The ANOVA F-statistic is testing if at least one of the coefficients from the regression slopes differ from 0 (or do not), i.e. do they = 0? In context, the relevant conclusion is that the F-test (=19.76) and p-value (5.59*10^-10) < 0.05, so we reject the null hypothesis. This means the overall model is significant.

ii. The p-value does not agree with the answer to 1A in regards to the Exam variable. It does agree with the answer regarding the Education variable. The p-value and slope does not agree with the answers to part 2 of 1A (Examinations and Education, Examinations and Agriculture, Agriculture and Education).

There is likely a strong presence of multicolinearity in the regressor variables, causing the model fit to be unreliable.

10/17/22

Module 06 Homework

#1) On Rmd. file.

#2) a) Est. coefficient of "Stay" = 0.237209.

If all other variables are kept constant, then a 1 unit increase in the average stay length is associated w/ a 0.237209% increase in infection risk percentage.

b.) t-stat:

$$t = \frac{-0.014071}{0.022708} = \boxed{-0.6196}$$

p-value:

$$p = P(t_{df=113-5\ (=108)} \leq -0.6196) \cdot 2 = \boxed{0.5368}$$

Critical value:

$$CV = \pm t_{0.025, df=108} = \boxed{\pm 1.9821}$$

Hypotheses

$H_0: \beta_{age} = 0$, $H_a: \beta_{age} \neq 0$

Since our p-value is > 0.05, we fail to reject the null hypothesis.

c) I agree with my classmate's statement that the variable Age is not linearly related to the predicted infection rate. This is because the large p-value caused us to fail to reject the null hypothesis.

d) Significance level:

$$s.l. = \frac{0.05}{3} = 0.01667$$

$$t_{cr, df=108, 0.5-0.01667 (=0.0083)} = 2.4318$$

95% joint confidence intervals

(coeff) (t) (SE)

β_1: $0.237209 \pm 2.4318 \times 0.060957 = (0.0889, 0.3854)$

β_2: $-0.014071 \pm 2.4318 \times 0.022708 = (-0.0693, 0.0412)$

β_3: $0.020383 \pm 2.4318 \times 0.005524 = (0.0069, 0.03382)$

e) ANOVA Table

Source of Variation	df	SS	MS
Regression	4	84.588	21.147
Error	108	116.791	1.081
Total	112	201.38	****

R^2:

f) $R^2 = \frac{84.586}{201.380} = 0.42 = 42\%$ $\left(R^2 = \frac{SSR}{SST}\right)$

The R^2 value indicates that the 4 different variables in the model explain 42% of the observed variation in hospital infection risk.

g) R^2_{adj}:

$$R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-p-1} = 1 - \frac{(1-0.42)(113-1)}{(113)-(4)-(1)}$$

$$R^2_{adj} = 1 - \left(\frac{64.96}{108}\right) = 1 - 0.601 = 0.399 = 39\%$$

3) Contradiction: ANOVA F-stat is significant, but t-stats for both predictors are insignificant.

The t-statistics are insignificant due to incredibly high correlation between the 2 variables (Left Foot & Right Foot.) This is not surprising as most people's bodies will be highly proportional.

This results in a statistical concept known as multicollinearity, when several independent variables in a model are correlated. A common sign of multicollinearity is when the t-stats are non-significant for each variable, but the overall F-test is significant.

---

4) $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = (X'X)^{-1}X'y$ | $\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y = Hy$ w/ $H = X(X'X)^{-1}X'$

Show $HH = H$:

$$HH = X(X'X)^{-1}X'X(X'X)^{-1}X' \quad \text{by definition}$$

$$= X[(X'X)^{-1}X'X](X'X)^{-1}X'$$

$$= XI(X'X)^{-1}X' \quad \text{w/ } I = \text{identity matrix}$$

$$= X(X'X)^{-1}X'$$

$$= H$$