hw08

Matt Scheffel

2022-10-24

Matt Scheffel

mcs9ff

Module 7 HW

1A

```
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
```

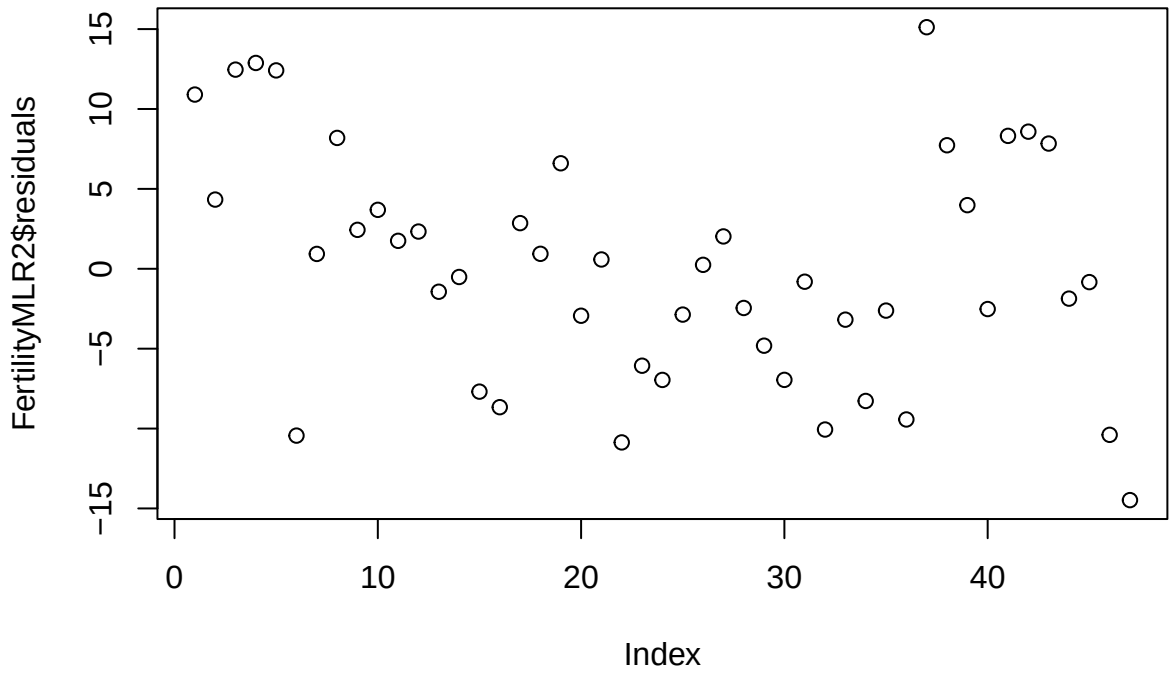
```
data<-swiss
y<-data$Fertility
x1<-data$Education
x2<-data$Catholic
x3<-data$Infant.Mortality
x4<-data$Agriculture
x5<-data$Examination

FertilityMLR2<-lm(y~x1+x2+x3)
summary(FertilityMLR2)
```

```
##
## Call:
```

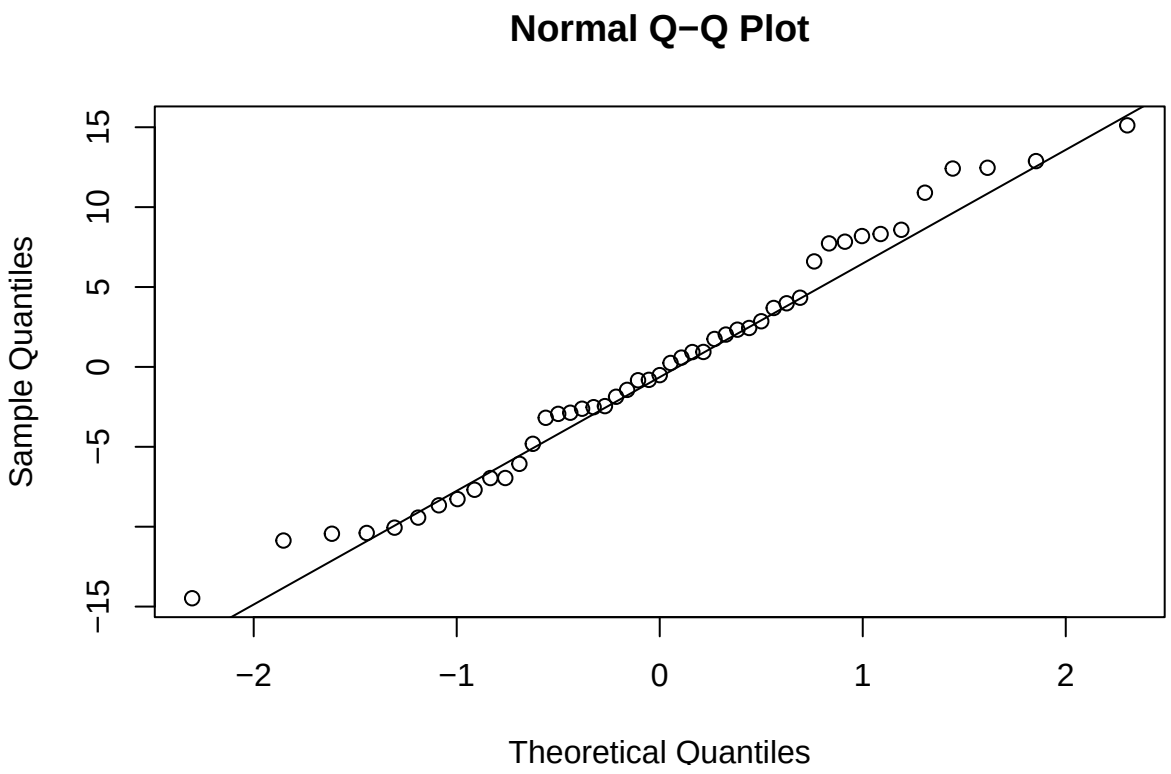
```
## lm(formula = y ~ x1 + x2 + x3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4781  -5.4403  -0.5143   4.1568  15.1187
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 48.67707    7.91908   6.147 2.24e-07 ***
## x1          -0.75925    0.11680  -6.501 6.83e-08 ***
## x2           0.09607    0.02722   3.530  0.00101 **
## x3           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
```

1B

```
plot(FertilityMLR2$residuals)
```




```
qqnorm(FertilityMLR2$residuals)
qqline(FertilityMLR2$residuals)
```




```
round(1-pf(1.7814,3,107),4)
```

```
## [1] 0.1551
```

```
round(qf(0.95,3,107),4)
```

```
## [1] 2.6895
```

```
round(1-pf(2.658,2,107),4)
```

```
## [1] 0.0747
```

```
round(qf(0.95,2,107),4)
```

```
## [1] 3.0812
```

10/24/22

Module 07 Homework

1)

a) 1st model:

$$y = \beta_0 + \sum_{i=1}^{5} \beta_i X_i + e \quad \text{w/ 5 predictors}$$

2nd model:

$$y = \beta_0 + \sum_{i=1}^{3} \beta_i X_i + e \quad \text{w/ 3 predictors}$$

Hypotheses:

H_0: $\beta_4 = \beta_5 = 0$

H_a: β_4 and/or $\beta_5 \neq 0$

F-test:

$$= \frac{\left(\frac{SSE_2 - SSE_1}{df_2 - df_1}\right)}{SSE_1 / df_1}$$

$$= \frac{\left(\frac{2422.56 - 2105.29}{43 - 41}\right)}{2105.29 / 41}$$

$$= 3.089$$

w/ $SSE = (1 - R^2) \times (SST)$

$SST = 7177.96$

$SSE_2 = (1 - 0.6625) \times 7177.96$

$\rightarrow SSE_2 = 2422.56$

$SSE_1 = (1 - 0.7067) \times 7177.96$

$\rightarrow SSE_1 = 2105.29$

$df_1 = 41$

$df_2 = 43$

P-value

$P(F_{2,41} \geq 3.089) = 0.056 < 0.05$

As a result, we fail to reject the null hypothesis.

The 2nd model (the simpler model w/ 3 predictors) is the better model to use.

b) From R code:

Used R code to create a Normal QQ Plot and a QQ Line Plot.

Based of the results of the 2 graphs, we can see that the simpler model meets the regression assumptions for linearity, independence, and normality.

2) a) Based on t-statistics, the predictors that appear to be insignificant are:

- Age: $0.4519 > 0.05$
- Beds: $0.8676 > 0.05$
- Census: $0.7686 > 0.05$

b) Hypothesis test:

β_3 = true regression coeff. for Age

β_4 = true regression coeff. for Census

β_5 = " " for Beds

H_0 : $\beta_3 = \beta_4 = \beta_5 = 0$

H_a : at least 1 of $\beta_3, \beta_4, \beta_5 \neq 0$

$H_0 \rightarrow SSE_{H_0} = 105.413 + 0.136 + 5.101 + 0.028 = 110.678$ w/ df = 110

$SSE = 105.413$ w/ df = 107

$SSH_0 = SSE_{H_0} - SSE = 0.136 + 5.101 + 0.028 = 5.265$ w/ df = 3

$$F_{obs} = \frac{SSH_0/3}{SSE/107} = \frac{5.265/3}{105.413/107} = 1.7814$$

P-value

$$P(F > 1.7814 \mid F \sim F_{3,107}) = 0.1551$$

Critical value

$$F_{crit} = F_{0.05,3,107} = 2.6895$$

$F_{critical} > F_{observed}$ & p-value > 0.05

→ Fail to reject H_0 at 5% significance level. Thus, we can drop these predictors from the model.

c) Model 1: x_1, x_2, x_3, x_4 as predictors for InfctRisk

Model 2: x_1, x_2 as predictors for InfctRisk

Hypothesis test

β_3 = true reg. coeff. for Age

β_4 = true reg. coeff. for Census

H_0: $\beta_3 = \beta_4 = 0$

H_a: β_3 and/or $\beta_4 \neq 0$

SSE = 105.413 w/ df = 107

$SSE_{H_0} = 105.413 + 0.136 + 5.101 = 110.65$ w/ df = 109

$SSH_0 = SSE_{H_0} - SSE = 0.136 + 5.101 = 5.237$ w/ df = 2

$$F_{obs} = \frac{SSH_0/2}{SSE/107} = \frac{5.237/2}{105.413/107} = 2.658$$

$$P\text{-value} = P(F > 2.658 \mid F \sim F_{2,107}) = 0.0747$$

$$F_{crit} = F_{0.05,2,107} = 3.0812$$

$F_{critical} > F_{observed}$; p-value > 0.05

→ Fail to reject H_0 at 5% significance level, thus, these predictors can be dropped from the model.

Model 2 will likely be the better option.

3) Presence of multicollinearity

From R:

$\hat{y} = 11.7104 + 0.3519 x_1 + 0.1850 x_2$

w/ x_1 = Left Foot, x_2 = Right Foot, $\hat{y}$ = Left Arm

→ p-values show us x_1 & x_2 are <u>not</u> significant

→→ $\hat{y} = 11.7104 + 0.74 x_1 + 0.185 (x_2 - 2x_1)$

From rewriting the equation in this format, we see that $\hat{y}$ contains a confounding relationship between x_1 & x_2. This confounding r-ship can lead to the variables being insignificant.

This means x_1 & x_2 have an correlational relationship. The variables cannot be statistically significant until the correlation is removed.

The insignificant variables/predictors indicate that multicollinearity exists in this regression model.