Module 7

Tuesday, March 18, 2025 12:28 AM

Homework Set 7

Michael Puchalski

2025-03-18

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become e
```

```
library(datasets)
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.4.3
```

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

```
data<-swiss
```

1.

(a) Fit a simple model using just education, catholic and infant.morality

```
result<- lm(Fertility~., data=data)
summary(result)
```

```
##
## Call:
## lm(formula = Fertility ~ ., data = data)
##
## 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
```

```
simple.result<-lm(Fertility~Education+Catholic+Infant.Mortality, data=data)
anova(simple.result)
```

```
## Analysis of Variance Table
##
## Response: Fertility
##                  Df Sum Sq Mean Sq F value    Pr(>F)
## Education         1 3162.7  3162.7  56.145 2.505e-09 ***
## Catholic          1  961.1   961.1  17.061 0.0001637 ***
## Infant.Mortality  1  631.9   631.9  11.218 0.0016938 **
## Residuals        43 2422.2    56.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the findings of this comparison, it would suggest that in this model these coefficients all have a significant test statistic as well as a significant P-value. and a highly significant p-value for the model as a whole

```
summary(simple.result)
```

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

```
anova(simple.result, result)
```

```
## 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
```

Null Hypothesis: At least one of the coefficients from the full model is 0

Alternate: none of these values are 1

Homework Set 7

2a) Based on ~~predictors~~ t statistics, x_3, x_4, x_5 all seem to be insignificant based on a low t stat followed by p values that are lower than 0.05

2b) Using General linear F test to determine if all ~~the~~ predictors chosen can be removed.

$H_0: \beta_3 = \beta_4 = \beta_5 = 0$ H_a: at least one coeff is not zero.

Looking for $F_0 =$

F_0 compared @ $f_{r, n-p}$:

r denotes parameter dropped = 3

$n = 113$ $n - p = 108$

$p = 5$

$F_{3, 50}$

$$F_0 = \frac{SS_{res}(R) - SS_{res}(F)}{SS_{res}(F)/(n-p)} / r$$

~~$SS_{res}(R) = 33.97$~~

$SS_{res} = SST - SSR$

~~$SS_{res}(F) = 105.2413$~~

$$F_0 = \frac{SS_R(F) - SS_R(R)/r}{SS_{res}(f)/(n-p)}$$

~~$SS_R(F) =$~~

~~$\Delta SS_R(F) \Delta SS_R(R) =$~~

$$0.136 + 5.101 + 0.028$$

$$F_0 = \frac{5.265/3}{105.413/108}$$

~~$F_0 = 0.83244$~~ 1.79807

~~P-value = 1 - pf(0.83244, 3,~~

$= 1 - pf(1.79807, 3, 107)$

P-value $= 0.1519456$

Critcal value $= qf(1-0.05, 3, 107)$

$F_0 = 1.79807$

$p\text{value} = 0.1919456$

$df = 2.68949$

Conclusion: We cannot reject the null. This means none of these values do not equal zero also meaning that we can remove these predictors from the model and have a significant model where we appropriately explain variance w/ the model.

2(c)

Model 1 = x_1, x_2, x_3, x_4

Model 2 = x_1, x_2

3) This output is suggesting multi-colinearity because although the p-value for each individual coeff, is insignificant, the full model has a p-value that is highly significant and a high Fstat.