# uwa6xv_M06_HW

Alanna Hazlett

2024-03-12

## Problem 1

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

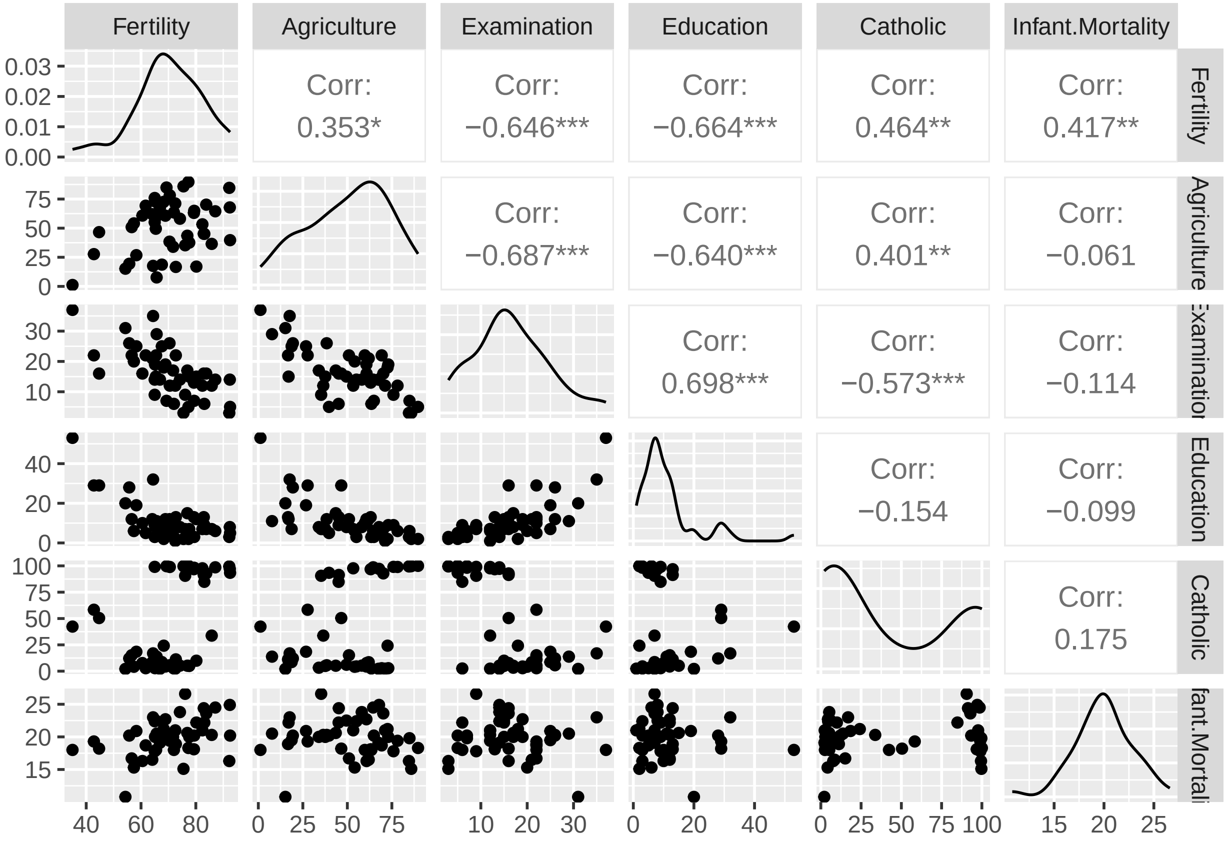
```
## Loading required package: ggplot2

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

```
Data<-swiss
```

### (a)

Create a scatterplot matrix and find the correlation between all pairs of variables for this data set.

```
ggpairs(Data)
```



**(i)**
Which predictors appear to be linearly related to the fertility measure?
* Examination
* Education

**(ii)**
Do you notice if any of the predictors are highly correlated with one another? If so, which ones?
* Examination and Agriculture
* Education and Agriculture
* Education and Examination
* Catholic and Examination

**(b)**
Fit a multiple linear regression with the fertility 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.

```r
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
```

**(i)**
What is being tested by the ANOVA F statistic? What is the relevant conclusion in context?

$$H_0 : \hat{\beta}_0 = \hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_3 = \hat{\beta}_4 = \hat{\beta}_5 = 0 H_a : at\ least\ one\ coefficient\ in\ H_0 \neq 0$$

The F statistic helps us to determine if the MLR model is useful. Here our F statistic is 19.76 on 5 and 41 degrees of freedom and we will compare that to a F (p-1,n-p) distribution critical value.

```r
qf(0.95,5,41)
```

```
## [1] 2.443429
```

Our F statistic is greater than our critical value (our p-value is less than our significance level), so we reject the null hypothesis. The data support the claim that our model with the five predictors is useful in predicting the fertility.

**(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.
If we conducted t tests for each predictor variable the null hypothesis would be that the beta for that predictor would equal zero and the alternative would be that it would not equal zero. We can see that none of the predictor variables have a beta value of zero. This is in agreeance with what I found in part i of part b. Agriculture, Education, Catholic, and Infant Mortality all have p-values that are less than the level of significance of 0.05. This indicates that for these variables we reject the null hypothesis and the data supports keeping these variables in the MLR model. For Examination, the p-value is greater than our significance level, so we would fail to reject the null hypothesis and this supports dropping this variable from our MLR model. This is contradictory to our statement from part i in part b. The explanation for this is multicollinearity. The Examination variable has high correlation values to other predictors, Education, Catholic, and Agriculture.

## Problem 2

**(a)**
What is the value of the estimated coefficient of the variable Stay? Write a sentence that interprets this value. The estimated coefficient of the variable Stay is 0.237209. This denotes the change in the predicted response, Infection Risk, per unit change in the length of stay, when the other predictors are held constant. For each unit of increase in length of stay, the infection risk increases by 0.24%, while the other predictors are 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?
The test statistic is:

$$t = \frac{\hat{\beta}_{age}}{se(\hat{\beta}_{age})}$$

$$t = \frac{-0.014071}{0.022708} = -0.619649$$

The p-value is:

```
2*pt(-0.619649, df=113-5)
```

```
## [1] 0.536794
```

The critical value is:

```
qt(0.95, 113-5)
```

```
## [1] 1.659085
```

$$H_0 : \hat{\beta}_{age} = 0$$

$$H_a : \hat{\beta}_{age} \neq 0$$

Our test statistic is less than our critical value (in magnitude) and our p-value is greater than our significance level, so we fail to reject the null hypothesis. The data supports dropping this predictor, age, from the MLR model.
**(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.
We can not distinctively determine this based on our t test alone. Just because the data supports dropping this predictor from the MLR model does not mean that it is not linearly related. It could display multicollinearity with the other predictor variables. In order to determine if age is linearly related to our response variable, infection risk, we would need to graph a scatterplot of infection risk against age and evaluate it.

Alanna Hazlett
3/18/24

MØ6 Continued

2.) (d.)

| Source of variation | df | SS | MS |
|---|---|---|---|
| Regression | 5-1 = 4 | 84.6244 | 21.1561 |
| Error | 113-5 = 108 | 120.0576 | 1.0816 |
| Total | 113-1 = 112 | 204.6820 | *** |

$S = 1.04 \quad S^2 = \boxed{1.0816 = MS_{res}}$

$MS_{res} = \frac{SS_{res}}{113-2} \quad 1.0816 \cdot 111 = \boxed{120.0576 = SS_{res}}$

$F\ stat = 19.56 \quad F\text{-}Stat = \frac{MS_R}{MS_{res}} \quad 19.56 \cdot 1.0816 = \boxed{21.1561 = MS_R}$

$SS_R = MS_R \cdot df_R = 21.1561 \cdot 4 = \boxed{84.6244 = SS_R}$

$SS_T = SS_R + SS_{res} = 84.6244 + 120.0576 = \boxed{204.6820 = SS_T}$

(e.) What is $R^2$ for this model? Interpret in context.

$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{res}}{SS_T} = \frac{84.6244}{204.6820} = \boxed{0.4134}$

This is the proportion of variance in the response variable that is explained by the predictor variables.
About 41.34% of the variance in infection risk can be explained by Stay, Age, Xrays, and Services.

(f.) What is $R^2_{adj}$ for this model?

$R^2_a = 1 - \left(\frac{n-1}{n-p}\right)\left(\frac{SS_{res}}{SS_T}\right) = 1 - \left(\frac{113-1}{113-5}\right)\left(\frac{120.0576}{204.6820}\right) = 1 - 0.60828$
$= \boxed{0.3917}$

3.) ANOVA F statistic is significant, t-statistics for both predictors are insignificant. Does this warrant concern?

- If the F statistic is significant then we reject the null hypothesis of $\hat{\beta}_0 = \hat{\beta}_1 = \hat{\beta}_2 = 0$, our data support alternative hypothesis that at least one coefficient $\neq 0$.
- If t statistic is insignificant for the predictors we fail to reject the null hypothesis of $\hat{\beta}_j = 0$ for each predictor.

This is Concerning, as these two tests should draw the same conclusions, either all reject the null hypotheses or all fail to reject the null hypotheses.

4.) $H = X(x'x)^{-1}X'$

Show that $H$ is idempotent, so $HH = H$.

$$=\overset{H}{(X(X'X)^{-1}X')}\cdot\overset{H}{(X(X'X)^{-1}X')}$$

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

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

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

$H \times H = H$ $\quad (X(X'X)^{-1}X')(X(X'X)^{-1}X') = (X(X'X)^{-1}X')$