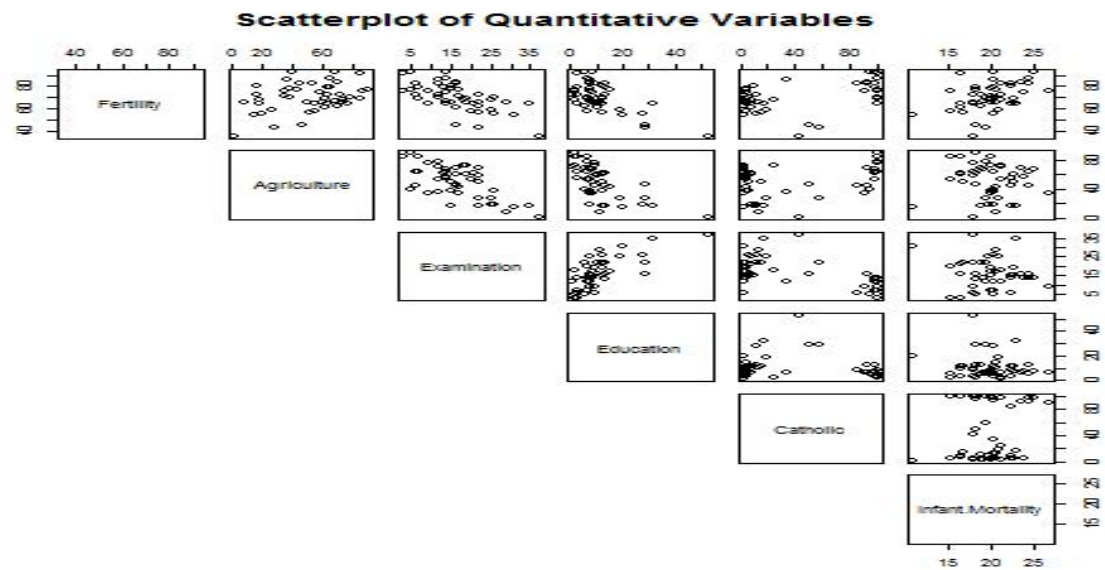


# Stat 6021: Homework Set 6 Solutions

- (a) The scatterplot matrix is displayed below:



The correlation matrix is shown below:

|                  | Fertility        | Agriculture | Examination | Education | Catholic |
|------------------|------------------|-------------|-------------|-----------|----------|
| Fertility        | 1.000            | 0.353       | -0.646      | -0.664    | 0.464    |
| Agriculture      | 0.353            | 1.000       | -0.687      | -0.640    | 0.401    |
| Examination      | -0.646           | -0.687      | 1.000       | 0.698     | -0.573   |
| Education        | -0.664           | -0.640      | 0.698       | 1.000     | -0.154   |
| Catholic         | 0.464            | 0.401       | -0.573      | -0.154    | 1.000    |
| Infant.Mortality | 0.417            | -0.061      | -0.114      | -0.099    | 0.175    |
|                  | Infant.Mortality |             |             |           |          |
| Fertility        | 0.417            |             |             |           |          |
| Agriculture      | -0.061           |             |             |           |          |
| Examination      | -0.114           |             |             |           |          |
| Education        | -0.099           |             |             |           |          |
| Catholic         | 0.175            |             |             |           |          |
| Infant.Mortality | 1.000            |             |             |           |          |

- i. The predictors Examination and Education have the highest (negative) correlations with the fertility measure.
- ii. A number of predictors are correlated with one another. For example, the percent of males involved in agriculture is correlated Examination and Education.

(b) The R output from the MLR:

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  | **  |

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 whether our MLR model is useful in predicting the response variable. With a small p-value, our data supports the claim that our model is useful in predicting the fertility measure.
  - ii. A couple of “contradictions”:
    - We noted in the previous part that the response variable was most correlated with the predictors Examination and Education. However, in the MLR, the variable Examination is insignificant. This informs us that we could drop Examination from the model. This apparent contradiction is due to the fact that Examination is highly correlated with a number of other predictors, so when those predictors are present, we do not need to add Examination to the model.
    - Also note that the relationship between the percent of males in agriculture and the fertility measure is positive, but in the MLR, the slope is negative. This apparent contradiction is due to the scatterplot and correlation not taking the effect of the predictors into account, whereas the MLR model is taking the other predictors into account.
2. (a) The estimated coefficient of *Stay* is 0.237209. The estimated *InfctRsk* increases by 0.237209 percent, for every one unit increase in the average length of stay in the hospital, when the other predictors are held constant.
- (b) The hypothesis statements are
- $H_0 : \beta_2 = 0.$
- $H_0 : \beta_2 \neq 0.$

The  $t$  statistic is  $\frac{-0.014071}{0.022708} = -0.6196$ . The pvalue is  $2 \times pt(-0.6196, 108)$  which is 0.5368.

Since p-value is greater than 0.05, we fail to reject the null hypothesis. Our data suggests that *Age* is not useful in predicting the response *InfctRsk*, when the other predictors are already in the model.

- (c) I disagree. The  $t$  test in a multiple linear regression does not test if a predictor is linearly related to the response, on its own.
- (d)
- Residual standard error =  $s = 1.04$ , so  $MSE = s^2 = 1.04^2$ , and  $SSE = 108 \times 1.04^2 = 116.8128$ .
  - $F = \frac{MSR}{MSE} \implies MSR = F \times MSE = 19.56 \times 1.04^2 = 21.1561$ . So  $SSR = 21.1561 \times 4 = 84.6244$ .
  - $SST = SSR + SSE = 84.6244 + 116.8128 = 201.4372$ .
  - DF for regression =  $p - 1 = 5 - 1 = 4$ .
  - DF for error =  $n - p = 113 - 5 = 108$ .
  - DF for total =  $n - 1 = 113 - 1 = 112$ .

| Source of Variation | SS       | df  | MS      |
|---------------------|----------|-----|---------|
| Regression          | 84.6244  | 4   | 21.1561 |
| Error               | 116.8128 | 108 | 1.0816  |
| Total               | 201.4372 | 112 | *****   |

(e)  $R^2 = \frac{SS_R}{SS_T} = \frac{84.6244}{201.4372} = 0.4201$ .

About 42% of the variation in the response variable *InfctRsk* is explained by our model.

(f)  $R_{adj}^2 = 1 - \frac{MS_{res}}{MS_T} = 1 - \frac{1.0816}{201.4372/112} = 0.3986$ .

3. The classmate's concern is not warranted. For the ANOVA  $F$  test, the hypothesis statements are

$$H_0 : \beta_1 = \beta_2 = 0.$$

$H_a$  : not all  $\beta_1, \beta_2$  equal zero.

Since the p-value is less than 0.05, we reject the null hypothesis. Our data suggests that our model is useful in predicting the response *LeftArm*.

Based on the  $t$  tests, we see that both of them do not have statistically significant results. We make the following conclusions:

*LeftFoot* can be removed from the model, when *RtFoot* is already in the model.

Likewise, *RtFoot* can be removed from the model, when *LeftFoot* is already in the model.

The conclusion for each  $t$ -test is that the variable is not a useful predictor given the presence of the other predictor in the model. Combined with the result of the  $F$  test, this means that we should only need one predictor, not both, to adequately predict left forearm length.

4.

$$\begin{aligned} HH &= (X(X'X)^{-1}X') (X(X'X)^{-1}X') \\ &= X(X'X)^{-1}X'X(X'X)^{-1}X' \\ &= XI(X'X)^{-1}X' \\ &= X(X'X)^{-1}X' \\ &= H \end{aligned}$$