HW4

Robert Chandler

2024-06-07

Setup

Begin by reading the dataset for this set of problems:

```
le <- read.csv("../../datasets/life_expectancy.csv")
```

```
y <- le$X2015Life.expectancy
x1 <- le$Medical.doctors
x2 <- le$Nurses
x3 <- le$Pharmacists
```

Problem 1

Complete this question with only the two ANOVA tables for $Y \sim X_1 + X_2 + X_3$, $Y \sim X_1$ generated in R.

1.a

Compute the following using the appropriate formulas of ESS and SS terms. For terms that can be found directly on the R output, please highlight them.

First, compute the ANOVA tables for each of the models specified:

SLR ANOVA table:

```
model_slr <- lm(y ~ x1)
anova_slr <- anova(model_slr)
anova_slr
```

```
## Analysis of Variance Table
##
## Response: y
##            Df Sum Sq Mean Sq F value    Pr(>F)
## x1          1 6428.4  6428.4  216.78 < 2.2e-16 ***
## Residuals 176 5219.2    29.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

MLR ANOVA table:

```
model_mlr <- lm(y ~ x1 + x2 + x3)
anova_mlr <- anova(model_mlr)
anova_mlr
```

```
## Analysis of Variance Table
##
```

```
## Response: y
##             Df Sum Sq Mean Sq  F value    Pr(>F)
## x1           1 6428.4  6428.4 231.4630 < 2.2e-16 ***
## x2           1  200.9   200.9   7.2325  0.007856 **
## x3           1  185.8   185.8   6.6897  0.010515 *
## Residuals 174 4832.5    27.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.a.i

$SSR(X_1)$ can be found directly from the `x1` row of the `Sum Sq` column of the SLR ANOVA table above.

```
ssr_x1 <- anova_slr["x1", "Sum Sq"]
```

$$SSR(X_1) = 6428.4367928$$

1.a.ii

$SSE(X_1)$ can be found directly from the `Residuals` row of the `Sum Sq` column of the SLR ANOVA table above.

```
sse_x1 <- anova_slr["Residuals", "Sum Sq"]
```

$$SSE(X_1) = 5219.1755668$$

1.a.iii

$SSR(X_2|X_1)$ can be found directly from the MLR ANOVA table at the `x2` row of the `Sum Sq` column. This is because each sequential row of the `Sum Sq` column in the ANOVA table gives the SSR given that the previous factors are already included in the model, so the `x2` row would give $(X_2|X_1)$:

```
ssr_x2_x1 <- anova_mlr["x2", "Sum Sq"]
```

$$SSR(X_2|X_1) = 200.8688839$$

1.a.iv

$SSE(X_2|X_1)$ is equivalent to $SSR(X_2|X_1)$ as discussed in the lecture notes. This is proven by starting with $SSE(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2)$, and then the latter term is:

$$SSE(X_1, X_2) = SSTO - SSR(X_1, X_2) = SSE(X_1) + SSR(X_1) - SSR(X_1, X_2)$$

so

$$\begin{aligned} SSE(X_2|X_1) &= SSE(X_1) - (SSE(X_1) + SSR(X_1) - SSR(X_1, X_2)) \\ &= SSR(X_1, X_2) - SSR(X_1) \\ &= SSR(X_2|X_1) \end{aligned}$$

Therefore,

$$SSE(X_2|X_1) = 200.8688839$$

1.a.v

We can find $SSR(X_1, X_2)$ using the formula:

```
ssr_x12 <- ssr_x2_x1 + ssr_x1
```

$$SSR(X_1, X_2) = SSR(X_2|X_1) + SSR(X_1) = 200.8688839 + 6428.4367928 = 6629.3056767$$

1.a.vi

We can find $SSE(X_1, X_2)$ using the following equation:

```
ssto <- sse_x1 + ssr_x1
sse_x12 <- ssto - ssr_x12
```

$$SSE(X_1, X_2) = SSTO - SSR(X_1, X_2)$$

where we can find $SSTO$ from the sum

$$SSTO = SSR(X_1) + SSE(X_1) = 1.1647612 \times 10^4$$

so then

$$SSE(X_1, X_2) = 1.1647612 \times 10^4 - 6629.3056767 = 5018.3066828$$

1.b

Test whether X_2 can be dropped from a model with X_1 (i.e., the marginal effect of X_2). Define H_0, H_a, compute the test statistic, denote and define the critical value, and state the conclusion.

The hypotheses are:

$$H_0 : \beta_2 = 0$$
$$H_a : \beta_2 \neq 0$$

assuming that the model already contains X_1.

This makes the full model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

and the reduced model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$$

So, the general linear test statistic is:

```
n <- length(y)
f_star_x2_x1 <- ssr_x2_x1 / (sse_x12 / (n - 3))
```

$$F^* = \frac{SSR(X_2|X_1)}{1} \div \frac{SSE(X_1, X_2)}{n-3}$$
$$F^* = 200.8688839 \div \frac{5018.3066828}{178-3}$$
$$F^* = 7.0047641$$

Assuming $\alpha = 0.05$, the critical value is:

```
alpha <- 0.05
p_reduced <- 2
p_full <- 3
df_reduced <- n - p_reduced
df_full <- n - p_full
f_crit <- qf(1 - alpha, df_reduced - df_full, df_full)
```

$$F(1-\alpha; df_R - df_F; df_F) = F(1-0.05; 1; 175) = 3.8951461 < 7.0047641 = F^*$$

Therefore, we reject H_0 and conclude that X_2 cannot be dropped from a model with X_1.

1.c

Test whether X_1, when used alone, has any significant linear impact on Y. Define H_0, H_a, compute the test statistic, denote and define the critical value, and state the conclusion.

The hypotheses are:

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

assuming that the model does not already contain any other variables.

This makes the full model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$$

and the reduced model

$$Y_i = \beta_0 + \varepsilon_i$$

So, the general linear test statistic is:

```
f_star_x1 <- ssr_x1 / (sse_x1 / (n - 2))
```

$$F^* = \frac{SSR(X_1)}{1} \div \frac{SSE(X_1)}{n-2}$$
$$F^* = 6428.4367928 \div \frac{5219.1755668}{178-2}$$
$$F^* = 216.7784665$$

Assuming $\alpha = 0.05$, the critical value is:

```r
p_reduced <- 1
p_full <- 2
df_reduced <- n - p_reduced
df_full <- n - p_full
f_crit <- qf(1 - alpha, df_reduced - df_full, df_full)
```

$$F(1-\alpha; df_R - df_F; df_F) = F(1-0.05; 1; 176) = 3.894838 < 216.7784665 = F^*$$

Therefore, we reject H_0 and conclude that X_1, when used alone, does have a significant linear impact on Y.

1.d

Test whether X_2 and X_3 can be dropped from a model with X_1. Define H_0, H_a, compute the test statistic, denote and define the critical value, and state the conclusion.

The hypotheses are:

$$H_0 : \beta_2 = \beta_3 = 0$$
$$H_a : \beta_2 \neq 0 \text{ or } \beta_3 \neq 0$$

assuming that the model already contains X_1.

This makes the full model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

and the reduced model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$$

So, the general linear test statistic is:

```r
p_reduced <- 2
p_full <- 4
df_reduced <- n - p_reduced
df_full <- n - p_full
ssr_x23_x1 <- sum(anova_mlr[c("x2", "x3"), "Sum Sq"])
sse_x123 <- sum(anova_mlr["Residuals", "Sum Sq"])
f_star_x23_x1 <- (
  (ssr_x23_x1 / (df_reduced - df_full))
  / (sse_x123 / (df_full))
)
```

$$F^* = \frac{SSR(X_2, X_3|X_1)}{df_R - df_F} \div \frac{SSE(X_1, X_2, X_3)}{n - df_F}$$
$$F^* = \frac{386.6617982}{2} \div \frac{4832.5137686}{4}$$
$$F^* = 6.9610927$$

Assuming $\alpha = 0.05$, the critical value is:

```r
f_crit <- qf(1 - alpha, df_reduced - df_full, df_full)
```

$$F(1-\alpha; df_R - df_F; df_F) = F(1-0.05; 2; 174) = 3.0479065 < 6.9610927 = F^*$$

Therefore, we reject H_0 and conclude that X_2 and X_3 cannot both be dropped from a model with X_1.

Problem 2

Use R to find the ANOVA table for $Y \sim X_1 + X_2 + X_3$, then...

```
anova_mlr
```

```
## Analysis of Variance Table
##
## Response: y
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## x1          1 6428.4  6428.4 231.4630 < 2.2e-16 ***
## x2          1  200.9   200.9   7.2325  0.007856 **
## x3          1  185.8   185.8   6.6897  0.010515 *
## Residuals 174 4832.5    27.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.a

Compute the coefficient of partial determination of X_2 given X_1 is included in the model.

Using the ANOVA table above:

```
ssr_x1 <- anova_mlr["x1", "Sum Sq"]
ssr_x12 <- sum(anova_mlr[c("x1", "x2"), "Sum Sq"])
ssr_x2_x1 <- ssr_x12 - ssr_x1
ssto <- sum(anova_mlr[, "Sum Sq"])
sse_x1 <- ssto - ssr_x1
r_sq_2_1 <- ssr_x2_x1 / sse_x1
```

$$R^2_{Y2|1} = \frac{SSR(X_2|X_1)}{SSE(X_1)} = \frac{200.8688839}{5219.1755668} = 0.0384867$$

2.b

Compute the coefficient of partial determination of X_3 given X_1 and X_2 are included in the model.

```
ssr_x3_x12 <- anova_mlr["x3", "Sum Sq"]
sse_x12 <- ssto - ssr_x12
r_sq_3_12 <- ssr_x3_x12 / sse_x12
```

$$R^2_{Y3|12} = \frac{SSR(X_3|X_1, X_2)}{SSE(X_1, X_2)} = \frac{185.7929142}{5018.3066828} = 0.037023$$

Problem 3

Analyze the multicollinearity in the MLR $Y \sim X_1 + X_2 + X_3$.

3.a

Use R to compute the linear correlation coefficient between each pair of the predictors, show the pairwise scatter plot, and comment on the level of multicollinearity.

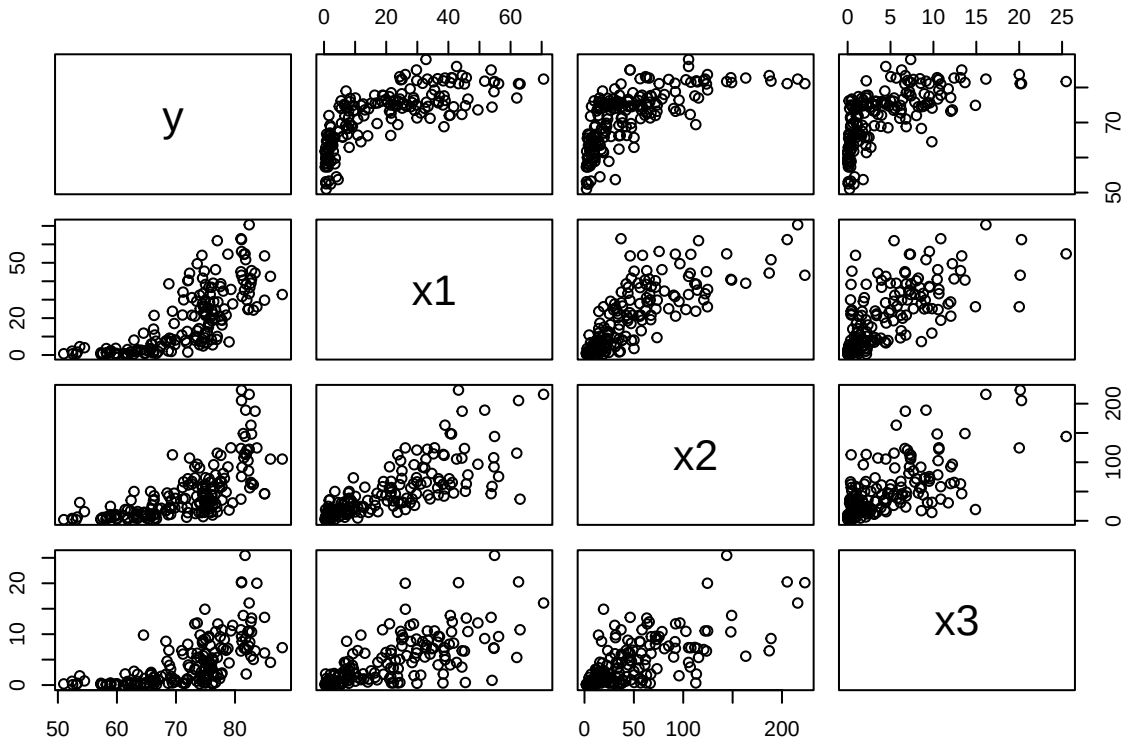
Correlation matrix showing the linear correlation coefficient between each pair of the predictors:

```
df <- data.frame(y, x1, x2, x3)
cm <- cor(df)
cm
```

```
##            y        x1        x2        x3
## y  1.0000000 0.7429066 0.6612558 0.6370947
## x1 0.7429066 1.0000000 0.7793268 0.7038282
## x2 0.6612558 0.7793268 1.0000000 0.6864924
## x3 0.6370947 0.7038282 0.6864924 1.0000000
```

Pairwise scatter plot:

```
plot(df)
```



The correlation matrix shows that the predictors are all fairly positively correlated with one another, with the lowest correlation coefficient being 0.6864924 between X_2 and X_3 and the highest being 0.7793268 between X_1 and X_2. The scatter plots agree with this, showing an obvious positive linear correlation between each of the predictor variables, the strongest being between X_1 and X_2.

In conclusion, there is certainly a fair amount of multicollinearity since the predictor variables are consistently correlated amongst themselves at significant levels. This makes sense, considering that we would expect the number of doctors, pharmacists, and nurses to increase in tandem with each other; it wouldn't make sense to have a very large number of doctors in a country but then only have a very small number of nurses.

3.b

Use R to compute the Type I and II ANOVA table, compare and then comment on the level of multicollinearity.

Type I ANOVA table

```
anova_mlr
```

```
## Analysis of Variance Table
##
## Response: y
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## x1          1 6428.4  6428.4 231.4630 < 2.2e-16 ***
## x2          1  200.9   200.9   7.2325  0.007856 **
## x3          1  185.8   185.8   6.6897  0.010515 *
## Residuals 174 4832.5    27.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Type II ANOVA table

```
library(car)
anova_ii <- Anova(model_mlr, type = "II")
anova_ii
```

```
## Anova Table (Type II tests)
##
## Response: y
##           Sum Sq  Df F value    Pr(>F)
## x1         983.1   1 35.3989 1.441e-08 ***
## x2          85.6   1  3.0805   0.08100 .
## x3         185.8   1  6.6897   0.01051 *
## Residuals 4832.5 174
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Beginning with the Type I table, we observe each row of the `Sum Sq` column. The `x1` row denotes $SSR(X_1)$, and we note its value of 6428.4367928. Next, we observe the `x2` row, which is $SSR(X_2|X_1)$, the marginal contribution of X_2 in the reduction of the error sum of squares when X_1 has already been considered. At a value of 200.8688839, it is much lower relative to the value for $SSR(X_1)$. This indicates that X_1 has already accounted for much of the same reduction in SSE that X_2 accounts for. This is an indication that these two variables share much of the same information and that there is multicollinearity between them. The same goes for the `x3` row, which is $SSR(X_3|X_1, X_2)$, and is also much lower relatively with a value of 185.7929142.

Moving onto the Type II table, we once again notice that the `Sum Sq` values for the different `x` rows are fairly small relative to the `Residuals` row, which is the SSE. Each row here represents the marginal decrease in the SSE attributed to the variable in the row when all other variables are already considered in the model. For example, when X_1 and X_3 are already considered, X_2 does not reduce SSE much at all in comparison, so we conclude that it is highly correlated with the other factors. We could also calculate the partial coefficients of determination/correlation with these values and we would see that these are fairly low. For example:

$$R^2_{Y2|1,3} = \frac{SSR(X_2|X_1, X_3)}{SSR(X_2|X_1, X_3) + SSE} = \frac{85.5542932}{85.5542932 + 4832.5137686} = 0.0173959$$

Which shows that marginal increase in SSR/decrease in SSE that results from including X_2 in the model is

low after the other variables have already been included because they have already accounted for most of the information in X_2 due to the level of multicollinearity between the input variables.

Problem 4

Use R to conduct a General Linear Test (GLT) to investigate whether X_1, X_2, and X_3 have an equivalent linear effect on Y. Specify H_0, H_a, establish a Reduced model and a Full model, calculate the test statistic, determine the critical value, calculate the p-value, and state your conclusion.

The hypotheses are:

$$H_0 : \beta_1 = \beta_2 = \beta_3$$
$$H_a : \text{not all } \beta_i \text{ are equal}, \quad i = \{1, 2, 3\}$$

The reduced model assumes that all β_i are equal, so the model is:

$$Y_i = \beta_0 + \beta_c(X_{i1} + X_{i2} + X_{i3}) + \varepsilon_i$$

where β_c is the common coefficient for $\beta_1, \beta_2, \beta_3$ under H_0 and the sum of X_i terms is the new X variable corresponding to the coefficient.

The full model is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

So, the full model has 4 parameters while the reduced model only has 2 since three of them have been collapsed to 1.

```
p_reduced <- 2
p_full <- 4
df_reduced <- n - p_reduced
df_full <- n - p_full
```

This makes $df_F = 174$ and $df_R = 176$.

We use the typical calculation for the GLT test statistic:

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F}$$

We can find the respective SSE terms using the ANOVA tables for each model:

```
anova_mlr
```

```
## Analysis of Variance Table
##
## Response: y
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## x1          1 6428.4  6428.4 231.4630 < 2.2e-16 ***
## x2          1  200.9   200.9   7.2325  0.007856 **
## x3          1  185.8   185.8   6.6897  0.010515 *
## Residuals 174 4832.5    27.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
model_reduced <- lm(y ~ I(x1 + x2 + x3))
anova_reduced <- anova(model_reduced)
anova_reduced
```

```
## Analysis of Variance Table
##
## Response: y
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## I(x1 + x2 + x3)    1 6105.8  6105.8  193.91 < 2.2e-16 ***
## Residuals        176 5541.8    31.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
sse_reduced <- anova_reduced["Residuals", "Sum Sq"]
sse_full <- anova_mlr["Residuals", "Sum Sq"]
f_star <- (sse_reduced - sse_full) /
  (df_reduced - df_full) /
  (sse_full / df_full)
f_crit <- qf(1 - alpha, df_reduced - df_full, df_full)
```

$SSE(R)$ can be pulled directly from the `Residuals` row of the `Sum Sq` column from the ANOVA table of the reduced model and the same goes for $SSE(F)$ with the full model. Then the test statistic is:

$$F^* = \frac{5541.8273859 - 4832.5137686}{176 - 174} \div \frac{4832.5137686}{174} = 12.7698104$$

The critical value (assuming $\alpha = 0.05$) is:

$$F(1-\alpha; df_R - df_F; df_F) = F(1 - 0.05; 2; 174) = 3.0479065 < 12.7698104 = F^*$$

The p-value is:

```r
p <- 1 - pf(f_star, df_reduced - df_full, df_full)
```

$$p = 1 - P\{F^* > F(df_R - df_F; df_F)\} = 1 - P\{12.7698104 > F(176 - 174; 174)\} = 6.6872751 \times 10^{-6}$$

We can check our results by using the two models in the `anova()` function directly:

```r
anova(model_reduced, model_mlr)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ I(x1 + x2 + x3)
## Model 2: y ~ x1 + x2 + x3
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1    176 5541.8
## 2    174 4832.5  2    709.31 12.77 6.687e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The values are in agreement, and since F^* is greater than the critical value and $p < \alpha$, we reject the null hypothesis and conclude that **the input variables do not have an equivalent linear effect on** Y.