HW5

Robert Chandler

2024-07-14

Setup

Begin by reading the dataset for this set of problems:

```
df <- read.csv("../../datasets/senic.csv")
```

Problem 1

The length of stay Y is to be regressed on X_4 (infection) and $X_{11(new)}$ (the availability of nurses). Note that the categorical variable, $X_{11(new)}$, is created from the original continuous variable, nurse in the data. Specifically, $X_{11(new)} = 0$ if the number of nurses is not less than the average value (173), and 1 otherwise. That is:

```
x11_new <- ifelse(df$nurse >= 173, 0, 1)
```

1.a

Consider the MLR model:

$$Y = \beta_0 + \beta_4 X_4 + \beta_{11(new)} X_{11(new)} + \beta_{4,11(new)} X_4 X_{11(new)} + \varepsilon$$

Use R to complete the model summary

```
y <- df$length
x4 <- df$infection
model <- lm(y ~ x4 + x11_new + x4 * x11_new)
model
```

```
##
## Call:
## lm(formula = y ~ x4 + x11_new + x4 * x11_new)
##
## Coefficients:
## (Intercept)           x4      x11_new   x4:x11_new
##      5.2257       1.0359       1.5468      -0.4174
```

1.b

The general form of the mean response functions for Y at the baseline level when $X_{11(new)} = 0$ is

$$Y = \beta_0 + \beta_4 X_4 + \beta_{11(new)} \cdot 0 + \beta_{4,11(new)} X_4 \cdot 0 = \beta_0 + \beta_4 X_4$$

Find the estimated mean response function by finding the value of the parameter estimate, i.e., b_0 and b_4 from the model summary.

Using the values from the `Coefficients` field of the model summary, the equation of the estimated mean response is:

```
b0 <- model$coefficients["(Intercept)"]
b4 <- model$coefficients["x4"]
```

$$\hat{Y} = 5.2257403 + 1.035904X_4$$

1.c

The general form of the mean response functions at the level where $X_{11(new)} = 1$ is

$$Y = \beta_0 + \beta_4 X_4 + \beta_{11(new)} \cdot 1 + \beta_{4,11(new)} X_4 \cdot 1 = (\beta_0 + \beta_{11(new)}) + (\beta_4 + \beta_{4,11(new)}) X_4$$

Find the estimated form.

Using the values from the `Coefficients` field of the model summary, the equation of the estimated mean response is:

```
b11_new <- model$coefficients["x11_new"]
b4_11_new <- model$coefficients["x4:x11_new"]
```

$$\hat{Y} = (5.2257403 + 1.5468466) + (1.035904 + -0.4174181)X_4 = 6.7725869 + 0.6184859X_4$$

Problem 2

Refer to Problem 1, complete the following hypothesis with a GLT F-test. Specifically, define the full model, reduced model, compute the test statistic, critical value, p-value, and state the conclusion.

2.a

Infection has no impact on Y, whether the number of nurses is less than the average or not.

In the case where the number of nurses is greater than the average ($X_{11(new)} = 0$), the term corresponding to the linear impact of infection X_4 on length of stay Y is β_4, so under this hypothesis, $\beta_4 = 0$

In the other case ($X_{11(new)} = 1$), the term corresponding to the linear impact of infection X_4 on length of stay Y is $\beta_4 + \beta_{4,11(new)}$, so under this hypothesis, $\beta_4 = \beta_{4,11(new)} = 0$.

Therefore, the reduced model, under $H_0 : \beta_4 = \beta_{4,11(new)} = 0$, is:

$$Y_i = \beta_0 + \beta_{11(new)} X_{i11(new)} + \varepsilon_i$$

and the full model under $H_a : \beta_4 \neq 0 \neq \beta_{4,11(new)}$ is:

$$Y_i = \beta_0 + \beta_4 X_{i4} + \beta_{11(new)} X_{i11(new)} + \beta_{4,11(new)} X_{i4} X_{i11(new)} + \varepsilon_i$$

```
model_reduced <- lm(y ~ x11_new)
```

The test statistic is:

```r
anova_res <- anova(model_reduced, model)
anova_res
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x11_new
## Model 2: y ~ x4 + x11_new + x4 * x11_new
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    111 383.79
## 2    109 283.06  2    100.72 19.393 6.233e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
f_star <- anova_res[2, "F"]
p <- anova_res[2, "Pr(>F)"]
```

$$F^* = 19.3930699$$

and the p-value is

$$p = 6.2331443 \times 10^{-8}$$

The p-value is well below an α of 0.05, so we reject H_0 and conclude that infection **does** impact Y irrespective of the category of nurses.

2.b

Infection has the same impact on Y when both when the number of nurses is less than the average and when the number of nurses is greater than the average. In other words, there is no interaction impact between X_4 and $X_{11(new)}$ on Y.

The parameter terms corresponding to the linear impact of X_4 on Y for the two different cases (identified in the previous part) will be equal to each other if the linear impact is equal regardless of the nurse category. Therefore:

$$\begin{aligned} \beta_4 &= \beta_4 + \beta_{4,11(new)} \\ 0 &= \beta_{4,11(new)} \end{aligned}$$

So the reduced model under $H_0 : \beta_{4,11(new)} = 0$ is

$$Y_i = \beta_0 + \beta_4 X_{i4} + \beta_{11(new)} X_{i11(new)} + \varepsilon_i$$

and the full model under $H_a : \beta_{4,11(new)} \neq 0$ is

$$Y_i = \beta_0 + \beta_4 X_{i4} + \beta_{11(new)} X_{i11(new)} + \beta_{4,11(new)} X_{i4} X_{i11(new)} + \varepsilon_i$$

```r
model_reduced <- lm(y ~ x4 + x11_new)
```

The test statistic is:

```r
anova_res <- anova(model_reduced, model)
anova_res
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x4 + x11_new
## Model 2: y ~ x4 + x11_new + x4 * x11_new
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1    110 288.89
## 2    109 283.06  1    5.8223 2.242 0.1372
```

```
f_star <- anova_res[2, "F"]
p <- anova_res[2, "Pr(>F)"]
```

$$F^* = 2.241993$$

and the p-value is

$$p = 0.137198$$

The p-value is greater than $\alpha = 0.05$, so we fail to reject H_0 and conclude that we cannot say with statistical significance that there is any interaction impact between X_4 and $X_{11(new)}$ on Y.

Problem 3

Consider the multiple linear regression (MLR) model that regresses the response on infection, region, and the interaction between infection and region. Region has four levels and can be modeled with 3 dummy variables: $X_{9,1}$, $X_{9,2}$, and $X_{9,3}$. The default baseline chosen in R is NC based on alphabetical order, with

$X_{9,1} = X_{9,2} = X_{9,3} = 0$, if the region is NC

$X_{9,1} = 1, X_{9,2} = X_{9,3} = 0$ if the region is NE

$X_{9,1} = 0, X_{9,2} = 1, X_{9,3} = 0$ if the region is S

$X_{9,1} = 0, X_{9,2} = 0, X_{9,3} = 1$ if the region is W.

The MLR full model can be written as follows:

$$Y = \beta_0 + \beta_4 X_4 + \beta_{9,1} X_{9,1} + \beta_{9,2} X_{9,2} + \beta_{9,3} X_{9,3} + \beta_{4\cdot 9,1} X_4 X_{9,1} + \beta_{4\cdot 9,2} X_4 X_{9,2} + \beta_{4\cdot 9,3} X_4 X_{9,3}$$

The response function can be specifically defined based on the 4 levels of the categorical variable, region. For the baseline (NC): $Y = \beta_0 + \beta_4 X_4$, which is estimated as $b_0 + b_4 X_4$.

For Level 2 (NE): $Y = \beta_0 + \beta_{9,1} + (\beta_4 + \beta_{4\cdot 9,1}) X_4$

For Level 3 (S): $Y = \beta_0 + \beta_{9,2} + (\beta_4 + \beta_{4\cdot 9,2}) X_4$

For Level 4 (W): $Y = \beta_0 + \beta_{9,3} + (\beta_4 + \beta_{4\cdot 9,3}) X_4$

3.a

Obtain the model summary and ANOVA table using R. Then, using the model summary, write the estimated response function for each level of the four regions.

```
region_map <- data.frame(nc = 1, ne = 2, s = 3, w = 4)
x91 <- x92 <- x93 <- rep(0, nrow(df))
x91[df$region == region_map$ne] <- 1
x92[df$region == region_map$s] <- 1
```

```r
x93[df$region == region_map$w] <- 1

model <- lm(y ~ x4 + x91 + x92 + x93 + x4 * x91 + x4 * x92 + x4 * x93)
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x4 + x91 + x92 + x93 + x4 * x91 + x4 * x92 +
##     x4 * x93)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1887 -0.7172 -0.2014  0.6821  6.2617
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.56051    0.83255   9.081 7.04e-15 ***
## x4           0.48317    0.18149   2.662  0.00898 **
## x91         -3.02259    1.32302  -2.285  0.02435 *
## x92         -0.43117    1.05413  -0.409  0.68335
## x93          0.47754    1.96420   0.243  0.80838
## x4:x91       0.86458    0.27371   3.159  0.00207 **
## x4:x92       0.04191    0.23841   0.176  0.86079
## x4:x93      -0.46589    0.43802  -1.064  0.28993
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.353 on 105 degrees of freedom
## Multiple R-squared:  0.5301, Adjusted R-squared:  0.4987
## F-statistic: 16.92 on 7 and 105 DF,  p-value: 8.413e-15
```

```r
anova_res <- anova(model)
anova_res
```

```
## Analysis of Variance Table
##
## Response: y
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## x4          1 116.446 116.446 63.5820 1.997e-12 ***
## x91         1  43.566  43.566 23.7881 3.843e-06 ***
## x92         1   2.400   2.400  1.3104 0.2549182
## x93         1  26.001  26.001 14.1969 0.0002721 ***
## x4:x91      1  25.894  25.894 14.1388 0.0002796 ***
## x4:x92      1   0.532   0.532  0.2903 0.5911519
## x4:x93      1   2.072   2.072  1.1313 0.2899308
## Residuals 105 192.300   1.831
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
b0 <- model$coefficients["(Intercept)"]
b4 <- model$coefficients["x4"]
b91 <- model$coefficients["x91"]
b92 <- model$coefficients["x92"]
b93 <- model$coefficients["x93"]
b4x91 <- model$coefficients["x4:x91"]
```

```r
b4x92 <- model$coefficients["x4:x92"]
b4x93 <- model$coefficients["x4:x93"]
```

The estimated response function for NC is:

$$\hat{Y} = b_0 + b_4 X_4 = 7.5605062 + 0.4831707 X_4$$

For NE:

$$\begin{aligned}\hat{Y} &= b_0 + b_{9,1} + (b_4 + b_{4\cdot 9,1}) X_4 \\ &= 7.5605062 + -3.0225854 + (0.4831707 + 0.8645752) X_4 \\ &= 4.5379208 + 1.3477459 X_4\end{aligned}$$

For S:

$$\begin{aligned}\hat{Y} &= b_0 + b_{9,2} + (b_4 + b_{4\cdot 9,2}) X_4 \\ &= 7.5605062 + -0.4311688 + (0.4831707 + 0.041912) X_4 \\ &= 7.1293374 + 0.5250827 X_4\end{aligned}$$

For W:

$$\begin{aligned}\hat{Y} &= b_0 + b_{9,3} + (b_4 + b_{4\cdot 9,3}) X_4 \\ &= 7.5605062 + 0.477542 + (0.4831707 + -0.4658921) X_4 \\ &= 8.0380482 + 0.0172786 X_4\end{aligned}$$

3.b

Consider the MLR model $Y \sim \text{infection} \cdot \text{Region}$ instead of $Y \sim \text{infection} \cdot \text{factor(Region)}$ in R, which causes R to treat "region" as a continuous variable. Compare the model summary and ANOVA output for this incorrect model to those in the previous questions. Be sure to specify in detail the differences between the two models, particularly in terms of the beta coefficients, MSE and dfE.

```r
model_wrong <- lm(y ~ x4 * df$region)
summary(model_wrong)
```

```
##
## Call:
## lm(formula = y ~ x4 * df$region)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8758 -0.8720 -0.2408  0.5230  8.1609
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.36820    1.33881   4.757 6.07e-06 ***
## x4            0.99601    0.29101   3.423 0.000875 ***
## df$region     0.06567    0.53685   0.122 0.902869
## x4:df$region -0.12112    0.11965  -1.012 0.313643
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.56 on 109 degrees of freedom
## Multiple R-squared:  0.352,  Adjusted R-squared:  0.3342
## F-statistic: 19.74 on 3 and 109 DF,  p-value: 2.717e-10
```

```
anova_wrong <- anova(model_wrong)
anova_wrong
```

```
## Analysis of Variance Table
##
## Response: y
##                Df  Sum Sq Mean Sq F value    Pr(>F)
## x4              1 116.446 116.446 47.8674 3.271e-10 ***
## df$region       1  25.110  25.110 10.3221  0.001729 **
## x4:df$region    1   2.493   2.493  1.0247  0.313643
## Residuals     109 265.161   2.433
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mse <- anova_res["Residuals", "Mean Sq"]
mse_wrong <- anova_wrong["Residuals", "Mean Sq"]
```

As shown in the output above, there are only 4 parameters (including the intercept) in this incorrect model compared to the 8 in the original one. This is because the original model had one continuous input and one categorical input with four categories. We broke this into three binary pseudo-inputs and encoded the categories in these variables, giving one parameter for the intercept, one for infection, one for each of the three pseudo-inputs, and one for the interaction between infection and each of the three pseudo-inputs. In the incorrect model, we only have parameters for the intercept, each of the two continuous inputs, and the single interaction between those two inputs. The numeric values of the categorical variable have no intrinsic meaning beyond identifying which category they are in, but we are regressing on them as if they do. We might still get lucky on the order of the values and the way they are arranged and still find some trends in the data, but it just does not have the same meaning it would as if we were using a truly continuous variable and we should not draw any legitimate conclusions from it.

We can see in the model summary that the p-value for the `region` parameter is significant, so we did happen to get lucky and the values for the region happen to line up in an order that actually looks like a linear response. While this is indicative that the `region` does at least have an impact on `Y`, we should not interpret meaning from the values of the parameters themselves here, since the numbers only indicate categories. We do note a lack of significance in most of the interaction parameters in both the correct and incorrect models. We also note that the intercept values are fairly close to each other.

In terms of the MSE values, we have a value of 1.8314273 in the original model and 2.4326735 in the incorrect model, which is a 1.3282938 increase in the estimation of σ^2, which points toward the model being a worse fit to the data.

Regarding df_E, we have a value of 105 for the original model and 109 for the incorrect model. This is due simply to the number of parameters being greater for the original model as discussed previously; as the number of parameters decreases, the degrees of freedom increases.

Problem 4

What is the implication of centering X variable to simplify the mean response prediction? If we center the X variable,

$$X' = (X - \bar{X}) \therefore X \triangleq \bar{X} \implies X' = 0$$

We can then fit the model $Y \sim \beta_0 + \beta_1 X'$. The mean response prediction at the mean level $X = \bar{X}$ simplifies to $Y_h = \beta_0$, because there is only the intercept value in the equation.

Now center the infection variable and denote the centered variable as $X_{4(center)}$, where $X_{4(center)} = X_4 - \bar{X}_4$ (infection minus the mean of infection). Then $X_{4(center)} = 0$ whenever $X_4 = \bar{X}_4$ (whenever it takes on the average value). Refit the MLR model with $X_{4(center)}$ (centered infection), X_9 (region, represented with three dummy variables $X_{9,1}$, $X_{9,2}$, $X_{9,3}$ and the interaction between them.

```
x4 <- x4 - mean(x4)
model <- lm(y ~ x4 + x91 + x92 + x93 + x4 * x91 + x4 * x92 + x4 * x93)
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x4 + x91 + x92 + x93 + x4 * x91 + x4 * x92 +
##     x4 * x93)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1887 -0.7172 -0.2014  0.6821  6.2617
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.66465    0.23934  40.381  < 2e-16 ***
## x4           0.48317    0.18149   2.662 0.008985 **
## x91          0.74252    0.36528   2.033 0.044603 *
## x92         -0.24865    0.33340  -0.746 0.457456
## x93         -1.55136    0.41456  -3.742 0.000298 ***
## x4:x91       0.86458    0.27371   3.159 0.002070 **
## x4:x92       0.04191    0.23841   0.176 0.860791
## x4:x93      -0.46589    0.43802  -1.064 0.289931
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.353 on 105 degrees of freedom
## Multiple R-squared:  0.5301, Adjusted R-squared:  0.4987
## F-statistic: 16.92 on 7 and 105 DF,  p-value: 8.413e-15
```

4.a

Predict the mean response value, Y_h, when infection takes the average value and fill in the blanks. The purpose of this question is to interpret the meaning of $\beta_0, \beta_4, \beta_{9,1}, \beta_{9,2}, \beta_{9,3}, \beta_{4\cdot 9,1}, \beta_{4\cdot 9,2}, \beta_{4\cdot 9,3}$ when the predictor takes the average value.

When infection takes on the average value, all X_4 terms go to 0:

$$Y_h = \beta_0 + \beta_{9,1}X_{9,1} + \beta_{9,2}X_{9,2} + \beta_{9,3}X_{9,3}$$

4.a.i

The average Y for NC is represented by β_0 and has a predicted value of

```
b0 <- model$coefficients["(Intercept)"]
```

$$Y_{h,NC} = b_0 = 9.6646505$$

4.a.ii

The average Y for NE is different from NC by $\beta_{9,1}$ and has a predicted value of

```
b91 <- model$coefficients["x91"]
```

$$Y_{h,NE} = b_0 + \beta_{9,1}X_{9,1} = 9.6646505 + 0.7425248X_{9,1}$$

4.a.iii

The average Y for S is different from NC by $\beta_{9,2}$ and has a predicted value of

```
b92 <- model$coefficients["x92"]
```

$$Y_{h,S} = b_0 + \beta_{9,2}X_{9,2} = 9.6646505 + -0.2486476X_{9,2}$$

4.a.iv

The average Y for W is different from NC by $\beta_{9,3}$ and has a predicted value of

```
b93 <- model$coefficients["x92"]
```

$$Y_{h,W} = b_0 + \beta_{9,3}X_{9,3} = 9.6646505 + -0.2486476X_{9,3}$$

4.b

Now consider the impact of infection on Y.

4.b.i

```
t_star <- summary(model)$coefficients["x4", "t value"]
p <- summary(model)$coefficients["x4", "Pr(>|t|)"]
```

The impact of infection on Y is $\beta_{4(center)}$ for NC.

N.B.: abbreviating $4(center)$ *as* 4_c *henceforth*

This impact is **significant** because (from the model summary above) the parameter has a t-value of 2.6621852 corresponding to a p-value of 0.0089846, which is less than $\alpha = 0.05$.

4.b.ii

```
b4x91 <- model$coefficients["x4:x91"]
t_star <- summary(model)$coefficients["x4:x91", "t value"]
p <- summary(model)$coefficients["x4:x91", "Pr(>|t|)"]
```

The impact of infection on Y for NE is different from NC by $\beta_{4_c \cdot 9,1}$ and has a predicted value of

$$b_{4_c} + b_{4_c \cdot 9,1} = 0.4831707 + 0.8645752 = 1.3477459$$

The difference is **significant** because (from the model summary above) it has a t-value of 3.1586777 corresponding to a p-value of 0.0020701, which is less than $\alpha = 0.05$.

4.b.iii

```
b4x92 <- model$coefficients["x4:x92"]
t_star <- summary(model)$coefficients["x4:x92", "t value"]
p <- summary(model)$coefficients["x4:x92", "Pr(>|t|)"]
```

The impact of infection on Y for S is different from NC by $\beta_{4_c \cdot 9,2}$ and has a predicted value of

$$b_{4_c} + b_{4_c \cdot 9,2} = 0.4831707 + 0.041912 = 0.5250827$$

The difference is **insignificant** because (from the model summary above) it has a t-value of 0.1757988 corresponding to a p-value of 0.8607906, which is greater than $\alpha = 0.05$.

```
b4x93 <- model$coefficients["x4:x93"]
t_star <- summary(model)$coefficients["x4:x93", "t value"]
p <- summary(model)$coefficients["x4:x93", "Pr(>|t|)"]
```

The impact of infection on Y for W is different from NC by $\beta_{4_c \cdot 9,3}$ and has a predicted value of

$$b_{4_c} + b_{4_c \cdot 9,3} = 0.4831707 + -0.4658921 = 0.0172786$$

The difference is **insignificant** because (from the model summary above) it has a t-value of -1.0636424 corresponding to a p-value of 0.2899308, which is greater than $\alpha = 0.05$.

4.c

Based on the result in part 4.b, can you judge whether the impact of infection on Y for NE is significantly different from for S? If not, find a way to compare and then conclude.

Testing whether the impact of infection on Y for NE and S are the same is equivalent to testing whether the coefficients corresponding to the linear impact of the respective parameters on Y are equal. Those coefficients are $\beta_{4_c} + \beta_{4_c \cdot 9,1}$ for NE and $\beta_{4_c} + \beta_{4_c \cdot 9,2}$ for S. If they are equal, then:

$$\begin{aligned} \beta_{4_c} + \beta_{4_c \cdot 9,1} &= \beta_{4_c} + \beta_{4_c \cdot 9,2} \\ \beta_{4_c \cdot 9,1} &= \beta_{4_c \cdot 9,2} \end{aligned}$$

Based on the analysis in 4.b, we see that the values for these two parameters look to be fairly different, but we cannot make any statistical conclusions without performing a test. We can perform a GLT where the reduced model under $H_0 : \beta_{4_c \cdot 9,1} = \beta_{4_c \cdot 9,2} = \beta'$ is

$$\begin{aligned} Y &= \beta_0 + \beta_{4_c} X_{4_c} + \beta_{9,1} X_{9,1} + \beta_{9,2} X_{9,2} + \beta_{9,3} X_{9,3} + \beta' X_{4_c} X_{9,1} + \beta' X_{4_c} X_{9,2} + \beta_{4_c \cdot 9,3} X_{4_c} X_{9,3} \\ &= \beta_0 + \beta_{4_c} X_{4_c} + \beta_{9,1} X_{9,1} + \beta_{9,2} X_{9,2} + \beta_{9,3} X_{9,3} + \beta' \left(X_{4_c} X_{9,1} + X_{4_c} X_{9,2} \right) + \beta_{4_c \cdot 9,3} X_{4_c} X_{9,3} \end{aligned}$$

and the full model under $H_0 : \beta_{4_c \cdot 9,1} \neq \beta_{4_c \cdot 9,2}$ is

$$Y = \beta_0 + \beta_{4_c} X_{4_c} + \beta_{9,1} X_{9,1} + \beta_{9,2} X_{9,2} + \beta_{9,3} X_{9,3} + \beta_{4_c \cdot 9,1} X_{4_c} X_{9,1} + \beta_{4_c \cdot 9,2} X_4 X_{9,2} + \beta_{4_c \cdot 9,3} X_{4_c} X_{9,3}$$

We can test the two against each other using `anova()`:

```
model_full <- model
model_reduced <- lm(
  y ~ x4 + x91 + x92 + x93 + I(x4 * x91 + x4 * x92) + x4 * x93
)
```

```r
anova_res <- anova(model_reduced, model_full)
anova_res
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x4 + x91 + x92 + x93 + I(x4 * x91 + x4 * x92) + x4 * x93
## Model 2: y ~ x4 + x91 + x92 + x93 + x4 * x91 + x4 * x92 + x4 * x93
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    106 211.11
## 2    105 192.30  1    18.814 10.273 0.001788 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
p <- anova_res[2, "Pr(>F)"]
```

The test has a p-value of $0.0017884 < 0.05 = \alpha$, so we reject H_0 and we conclude that the linear impact of infection on Y for NE **is significantly different** from that of S, confirming our suspicion observed by a glance at the estimates.

Problem 5

Perform a hypothesis test to see whether the interaction effect between infection and region on Y is significant. Define the hypothesis with appropriate notation. Then complete the hypothesis with a GLT F-test. Specifically, define the full model, reduced model, compute the test statistic, critical value, p- value, and state the conclusion.

If the interaction effect is *not* significant, then the coefficients on the interaction terms will be zero. That is:

$$H_0 : \beta_{4_c \cdot 9,1} = \beta_{4_c \cdot 9,2} = \beta_{4_c \cdot 9,3} = 0$$

under which the reduced model is:

$$Y = \beta_0 + \beta_{4_c} X_{4_c} + \beta_{9,1} X_{9,1} + \beta_{9,2} X_{9,2} + \beta_{9,3} X_{9,3}$$

The alternative hypothesis is:

$$H_a : \text{not all } (\beta_{4_c \cdot 9,1}, \beta_{4_c \cdot 9,2}, \beta_{4_c \cdot 9,3}) = 0$$

under which the full model is:

$$Y = \beta_0 + \beta_{4_c} X_{4_c} + \beta_{9,1} X_{9,1} + \beta_{9,2} X_{9,2} + \beta_{9,3} X_{9,3} + \beta_{4_c \cdot 9,1} X_{4_c} X_{9,1} + \beta_{4_c \cdot 9,2} X_4 X_{9,2} + \beta_{4_c \cdot 9,3} X_{4_c} X_{9,3}$$

Modeling these two in R and using `anova()` to test them against each other:

```r
model_full <- lm(y ~ x4 + x91 + x92 + x93 + x4 * x91 + x4 * x92 + x4 * x93)
model_reduced <- lm(y ~ x4 + x91 + x92 + x93)
anova_res <- anova(model_reduced, model_full)
anova_res
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x4 + x91 + x92 + x93
```

```
## Model 2: y ~ x4 + x91 + x92 + x93 + x4 * x91 + x4 * x92 + x4 * x93
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1    108 220.8
## 2    105 192.3  3    28.498 5.1868 0.002215 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
df_reduced <- anova_res[1, "Res.Df"]
df_full <- anova_res[2, "Res.Df"]

alpha <- 0.05
f_star <- anova_res[2, "F"]
f_crit <- qf(1 - alpha, df_reduced - df_full, df_full)
p <- anova_res[2, "Pr(>F)"]
```

The critical value is:

$$F(1-\alpha; df_R - df_F; df_F) = F(1-0.05; 3; 105) = 2.6911329$$

The test statistic, as shown in the table, is

$$F^* = 5.1868171$$

with a corresponding p-value of

$$p = 0.0022151$$

which is less than $\alpha = 0.05$, so we reject H_0 and conclude that the interaction effect between infection and region on Y **is significant**.

Problem 6

Consider the life expectancy data, and the MLR model given by $Y \sim X_1 + X_2 + X_3$.

Implement the best subsets regression algorithm. Look at all 7 possible models (all but the one with just an intercept). Then answer each of the following questions, providing R output to support your answers.

```r
le <- read.csv("../../datasets/life_expectancy.csv")
model <- lm(X2015Life.expectancy ~ ., data = le[, 2:5])

best_sub <- ALSM::BestSub(x = le[, 3:5], y = le$X2015Life.expectancy, n = 3)
```

```
##   p 1 2 3      SSEp         r2     r2.adj        Cp      AICp      SBCp    PRESSp
## 1 2 1 0 0  5219.176  0.5519103  0.5493643  13.922185  605.3394  611.7030  5346.345
## 1 2 0 1 0  6554.586  0.4372592  0.4340618  62.005130  645.8923  652.2559  6731.332
## 1 2 0 0 1  6919.968  0.4058896  0.4025140  75.161083  655.5481  661.9117  7114.980
## 2 3 1 0 1  4918.068  0.5777617  0.5729361   5.080477  596.7620 *606.3073  5092.511
## 2 3 1 1 0  5018.307  0.5691558  0.5642318   8.689679  600.3534  609.8988  5188.805
## 2 3 0 1 1  5815.650  0.5007002  0.4949939  37.398924  626.6013  636.1467  6066.149
## 3 4 1 1 1 *4832.514 *0.5851069 *0.5779536  *4.000000 *595.6383  608.3654 *5049.472
```

N.B.: the row with the best value has been marked with an asterisk for each column

6.a

Which model is best according to R_a^2?

```
r_adj_best <- max(best_sub[, "r2.adj"])
```

The model including all three inputs, $Y \sim X_1 + X_2 + X_3$, is best. It has the highest R_a^2 value at $R_a^2 = 0.5779536$.

6.b

Which model is best according to C_p?

```
c_p_best <- best_sub[7, "Cp"]
```

The model including all three inputs, $Y \sim X_1 + X_2 + X_3$, is best. It has both the lowest C_p value and the value of C_p that perfectly equals p at $C_p = 4$.

6.c

Which model is best according to AIC_p?

```
aic_best <- min(best_sub[, "AICp"])
```

The model including all three inputs, $Y \sim X_1 + X_2 + X_3$, is best. It has the lowest AIC_p value at $AIC_p = 595.6382548$.

6.d

Which model is best according to BIC_p?

```
bic_best <- min(best_sub[, "SBCp"])
```

The model $Y \sim X_1 + X_3$, is best. It has the lowest BIC_p value at $BIC_p = 606.3073277$.

6.e

Which model is best according to $PRESS$?

```
press_best <- min(best_sub[, "PRESSp"])
```

The model including all three inputs, $Y \sim X_1 + X_2 + X_3$, is best. It has the lowest $PRESS_p$ value at $PRESS_p = 5049.4717769$.

6.f

Based on your answers in parts 6.a through 6.e, which model do you believe is best? Explain the reasoning behind your answer.

All of the criteria above agree that $Y \sim X_1 + X_2 + X_3$ is the best model aside from BIC_p, so we conclude that it is the best since a variety of metrics each testing slightly different properties point to it being the best model of the subsets. However, it is worth noting that the model chosen by BIC_p, $Y \sim X_1 + X_3$, was a very close second in all metrics, with $Y \sim X_1$ by itself not far behind.