Benji_HW3

April 10, 2018

1 Benji Wagner

2 BIOS705 Homework 3

3 Question 1

Suppose we collected measurements on 70 subjects, where the response Y was SBP, and the predictors were age (X as continuous), gender (Z as categorical $0 =$ male, $1 =$ female), and education level (high-school, BS, MS and PhD).

3.0.1 Part A

Using dummy variables, specify the appropriate single regression model relating SBP to age, gender, and education level includes the following interactions: age × gender, age × education, and gender × education.

Let education level be denoted with 3 dummy variables $W_1 = 1$ for high school and 0 otherwise, $W_2 = 1$ for BS and 0 otherwise, $W_3 = 1$ for MS and 0 otherwise. If W_1, W_2, and W_3 are all 0, then the education level is PhD.

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 W_1 + \beta_4 W_2 + \beta_5 W_3 + \beta_6 XZ + \beta_7 XW_1 + \beta_8 XW_2 + \beta_9 XW_3 + \beta_{10} ZW_1 + \beta_{11} ZW_2 + \beta_{12} ZW_3 +$$

3.0.2 Part B

In terms of regression parameters from a), state the null hypothesis to be tested for the following:

Part 1 All regression equations coincide.

$$H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = 0$$

Part 2 All regression equations are parallel.

$$\beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$$

Part 3 All gender equations are parallel, controlling for education level.

$$\beta_6 = 0$$

Part 4 All education equations are parallel, controlling for gender.

$$\beta_7 = \beta_8 = \beta_9 = 0$$

3.0.3 Part C

For each of the hypotheses given in part a) specify the degrees of freedom and the test statistic.

There are 13 parameters in the full model and the sample size is 70. Therefore, the full model has degrees of freedom $70 - 13 = 57$.

Part 1:

$$F = \frac{SSR(Z, W_1, W_2, W_3, XZ, XW_1, XW_2, XW_3, ZW_1, ZW_2, ZW_3/X)/11}{MSE} \sim F_{11,57}$$

Part 2:

$$F = \frac{SSR(XZ, XW_1, XW_2, XW_3/X, Z, W_1, W_2, W_3, ZW_1, ZW_2, ZW_3)/4}{MSE} \sim F_{4,57}$$

Part 3:

$$F = \frac{SSR(XZ/Z, W_1, W_2, W_3, XZ, XW_1, XW_2, XW_3, ZW_1, ZW_2, ZW_3)/1}{MSE} \sim F_{1,57}$$

Part 4:

$$F = \frac{SSR(XW_1, XW_2, XW_3/Z, W_1, W_2, W_3, XZ, ZW_1, ZW_2, ZW_3)/3}{MSE} \sim F_{3,57}$$

4 Question 2

An investigator is interested in whether Bayer aspirin (treatment "A"), Tylenol (acetaminophen) (treatment "B") or Aleve (naproxen) (treatment "C") works more quickly to relieve the pain of a common headache. She recruits n individuals with frequent headaches, randomly assigns them to one of the three pain killers, asks them to take the medication upon first signs of the headache, and to record the time until the pain is gone Y.

4.0.1 Part A

Using indicator (dummy) variables in regression, set up the model for the experiment above. Assume n_i $i = 1, 2, 3$ subjects are randomized to group i, where $n = n_1 + n_2 + n_3$.

Let $X_1 = 1$ for Treatment A (Bayer aspirin) and 0 otherwise.

Let $X_2 = 1$ for Treatment B (Tylenol acetaminophen) and 0 otherwise.

If both X_1 and X_2 are 0, then the group is Treatment C (Aleve naproxen).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

4.0.2 Part B

Write your model in part a) in a matrix form as $Y = X\beta + \epsilon$. Identify the design matrix X as well as Y, β and ϵ vectors.

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} Y_{11} \\ \vdots \\ Y_{n_1} \\ Y_{21} \\ \vdots \\ Y_{n_2} \\ Y_{31} \\ \vdots \\ Y_{n_3} \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1_{n_1} & 1_{n_1} & 0_{n_1} \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1_{n_2} & 0_{n_2} & 1_{n_2} \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1_{n_3} & 0_{n_3} & 0_{n_3} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{n_1} \\ Y_{21} \\ \vdots \\ Y_{n_2} \\ Y_{31} \\ \vdots \\ Y_{n_3} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1_{n_1} & 1_{n_1} & 0_{n_1} \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1_{n_2} & 0_{n_2} & 1_{n_2} \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1_{n_3} & 0_{n_3} & 0_{n_3} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

4.0.3 Part C

State the null hypothesis of no difference among the means of the three groups. How would you carry-out this test and what is the distribution of the proposed test statistic?

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

I would use a partial F-statistic to see if at least one of the means is different.

$$F = \frac{MS_{trt}}{MSE} = \frac{SS_{trt}}{2} \frac{n-3}{SSE} \sim F_{2,\, n-3}$$

4.0.4 Part D

Provide the least-squares estimates $\hat{\beta}$ algebraically. Let n = 30 such that $n_1 = n_2 = n_3 = 10$

Using our design matrix X:

$$X^T X = \begin{bmatrix} n & n_1 & n_2 \\ n_1 & n_1 & 0 \\ n_2 & 0 & n_2 \end{bmatrix} = \begin{bmatrix} 30 & 10 & 10 \\ 10 & 10 & 0 \\ 10 & 0 & 10 \end{bmatrix}$$

$$X^T Y = \begin{bmatrix} y_{..} \\ y_{1.} \\ y_{2.} \end{bmatrix}$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y = \begin{bmatrix} \frac{y_{3.}}{10} \\ \frac{y_{1.}}{10} \\ \frac{y_{2.}}{10} \end{bmatrix}$$

4.0.5 Part E

Provide an expression for Var-Cov matrix for $\hat{\beta}$

$$\begin{bmatrix} Var(\hat{\beta_0}) & Cov(\hat{\beta_0}\hat{\beta_1}) & Cov(\hat{\beta_0}\hat{\beta_2}) \\ Cov(\hat{\beta_0}\hat{\beta_1}) & Var(\hat{\beta_1}) & Cov(\hat{\beta_1}\hat{\beta_2}) \\ Cov(\hat{\beta_0}\hat{\beta_2}) & Cov(\hat{\beta_1}\hat{\beta_2}) & Var(\hat{\beta_2}) \end{bmatrix}$$

$$Var(\hat{\beta}) = \hat{\sigma^2}(X^T X)^{-1} = \hat{\sigma^2} \begin{bmatrix} 0.1 & -0.1 & -0.1 \\ -0.1 & 0.2 & 0.1 \\ -0.1 & 0.1 & 0.2 \end{bmatrix}$$

5 Question 3

For the experiment in question 2 above, express the regression model as a "cell-mean" model, that is $Y_{ij} = \mu_i + \epsilon_{ij}$, where μ_i is the mean of the *ith* treatment group

5.0.1 Part A

Write the "cell-mean" model above in a matrix form as $Y = X\mu + \epsilon$, identify the design-matrix X and the parameters vector μ for the "cell-mean" model.

$$Y = \begin{bmatrix} Y_{11} \\ \vdots \\ Y_{n_1} \\ Y_{21} \\ \vdots \\ Y_{n_2} \\ Y_{31} \\ \vdots \\ Y_{n_3} \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1_{n_1} & 0_{n_1} & 0_{n_1} \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0_{n_2} & 1_{n_2} & 0_{n_2} \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0_{n_3} & 0_{n_3} & 1_{n_3} \end{bmatrix}$$

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}$$

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{n_1} \\ Y_{21} \\ \vdots \\ Y_{n_2} \\ Y_{31} \\ \vdots \\ Y_{n_3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1_{n_1} & 0_{n_1} & 0_{n_1} \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0_{n_2} & 1_{n_2} & 0_{n_2} \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0_{n_3} & 0_{n_3} & 1_{n_3} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

5.0.2 Part B

State the null hypothesis of no difference among the means of the three groups.

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

5.0.3 Part C

Provide the least-squares estimates $\hat{\mu}$ algebraically.

$$\hat{\mu_1} = \frac{Y_{11} + \cdots + Y_{n_1}}{n_1}$$

$$\hat{\mu_2} = \frac{Y_{21} + \cdots + Y_{n_2}}{n_2}$$

$$\hat{\mu_3} = \frac{Y_{31} + \cdots + Y_{n_3}}{n_3}$$

6 Question 4

A rehabilitation center researcher was interested in examining the relationship between physical ftness level prior to surgery of persons undergoing corrective knee surgery and time required in physical therapy until successful rehabilitation (in days). Patient records from the rehabilitation center were examined, and 24 male subjects ranging in age from 18 to 30 years who had undergone similar corrective knee surgery during the past year were selected for the study. The number of days required for successful recovery and fitness level (1=below average, 2=average and 3=above average) were recorded. The patient's age at the time of surgery was also recorded because previous studies have shown that younger patients tend to recover more qucikly. Data is given in HW3-prob4.dat file on Sakai.

```
In [4]: library(dplyr);
```

```
In [5]: df <- read.delim(file = "/home/jovyan/work/705/HW3-prob4.dat.txt", header = TRUE, sep =
```

```
In [3]: df$f1 <- 0
        df$f2 <- 0
        df[df$status == 1, ]$f1 <- 1
        df[df$status == 2, ]$f2 <- 1
```

6.0.1 Part A

Fit the ANCOVA model. Based on this model fit, is there evidence that time to recovery differs across fitness levels after accounting for a patient's age? Please provide the hypotheses, test statistic, and p-value of the test used to address this question.

$$Y = \beta_0 + \beta_1 X + \beta_2 F_1 + \beta_3 F_2$$

where X = age, $F_1 = 1$ for below-average fitness and 0 otherwise, $F_2 = 1$ for average fitness and 0 otherwise, and $F_1 = F_2 = 0$ indicates an above-average fitness level.

```
In [7]: summary(lm(days ~ age + f1 + f2, data = df))
```

```
Call:
lm(formula = days ~ age + f1 + f2, data = df)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.03891 -0.36892  0.05891  0.33098  0.89991

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.29121    0.72975  -1.769   0.0921 .
age          1.16729    0.03201  36.461  < 2e-16 ***
f1           8.72289    0.33296  26.198  < 2e-16 ***
f2           6.87551    0.28838  23.842 3.68e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.5552 on 20 degrees of freedom
Multiple R-squared:  0.9943,Adjusted R-squared:  0.9935
F-statistic:  1170 on 3 and 20 DF,  p-value: < 2.2e-16
```

It seems there is evidence supporting the claim that recovery time differs across fitness levels.

$$H_0 : \beta_0 = \beta_2 = \beta_3$$

$$F_{3,20} = 1170 P - value < 2.2 \times 10^{-16}$$

6.0.2 Part B

Using the model fit in part (a), provide the estimated regression model for each fitness level group.

Below Average:

$$\hat{Y} = \hat{\beta_0} + \hat{\beta_2} + \hat{\beta_1} X$$

$$\hat{Y} = 7.48168 + 1.16729(age)$$

Average:

$$\hat{Y} = \hat{\beta_0} + \hat{\beta_3} + \hat{\beta_1} X$$

$$\hat{Y} = 5.5843 + 1.16729(age)$$

Above Average:

$$\hat{Y} = \hat{\beta_0} + \hat{\beta_1} X$$

$$\hat{Y} = -1.29121 + 1.16729(age)$$

6.0.3 Part C

Estimate the overall slope beween days to recovery and patient age by removing fitness level from the model fit in part (a). How does this slope estimate compare to the slope estimates reported in part (b)?

```
In [8]: summary(lm(days ~ age, data = df))
```

```
Call:
lm(formula = days ~ age, data = df)

Residuals:
   Min     1Q Median     3Q    Max 
-7.540 -1.906  1.344  1.990  4.760 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -2.6820     4.1207  -0.651    0.522    
age           1.4711     0.1723   8.538 1.97e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.386 on 22 degrees of freedom
Multiple R-squared:  0.7682,Adjusted R-squared:  0.7576 
F-statistic: 72.89 on 1 and 22 DF,  p-value: 1.972e-08
```

$$\hat{Y} = -2.6820 + 1.4711(age)$$

The slope estimate is now slightly higher.

6.0.4 Part D

Compute the unadjusted mean days to recovery for each fitness level group. Using this numerical ouput, what factor level differences appear to be driving the global signal from fitness level if one exists?

```
In [17]: df %>% filter(f1 == 1) %>% summarize("Below Average Fitness Mean Days to Recovery" = me
```

	Below Average Fitness Mean Days to Recovery
1	38

```
In [18]: df %>% filter(f2 == 1) %>% summarize("Average Fitness Mean Days to Recovery" = mean(day
```

	Average Fitness Mean Days to Recovery
1	32

```
In [19]: df %>% filter(f1 == 0 & f2 == 0) %>% summarize("Above Average Fitness Mean Days to Reco
```

	Above Average Fitness Mean Days to Recovery
1	24

It seems the global signal from fitness level is driven by this linear relationship between fitness level and days to recovery.

6.0.5 Part E

Compute the mean days to recovery for each fitness level adjusted for average patient age. Using this numerical ouput, what factor level difference appear to be driving the global signal from fitness level if one exists? Does your conclusion differ from the conclusion you reached in part (d)? If so, why does the discrepancy exist?

Below Average:

```
In [31]: 7.48168 + 1.16729 * mean(df$age) # Mean patient age overall
```

```
35.00054175
```

Average:

```
In [32]: 5.5843 + 1.16729 * mean(df$age) # Mean patient age overall
```

```
33.10316175
```

Above Average:

```
In [34]: -1.29121 + 1.16729 * mean(df$age) # Mean patient age overall
```

```
26.22765175
```

When we adjust for average overall patient age, the difference between mean days to recovery becomes smaller. This is especially true for the change from Below Average Fitness to Average Fitness. The difference is very small, leading us to conclude that global signal from fitness arises from the difference of having Above Average Fitness.

6.0.6 Part F

Was it beneficial to include patient age in this analysis when the primary goal was to assess the relationship between days to recovery and fitness level? Explain your reasoning.

I would say so. Although the relationship remained relatively the same after adjusting for age, a clinician may tell us that there is evidence supporting the claim that age affects recovery rate. Thus, if we didn't adjust for age, it could be a confounding variable in our model.