1. Review - Please skip questions

This two weeks we began discussing multiple linear regression. The model that we fit is an extension of that fit in simple linear regression, and is given by:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip} + \varepsilon_i,$$

We assume that the observations Y_is are independent from each other, $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$ and that $p < n$.

The β_j's are interpreted as the the change in the expected response (i.e., E(Y)) per unit change in X_j, holding the other X_i ($i \neq j$) constant.

1. What is the multiple linear regression model in matrix form?
2. What are each of the pieces of the model representing?
3. What is the least squares estimate for $\beta = (\beta_0, \beta_1, \dots, \beta_p)'$ in matrix form?

In addition to including multiple covariates, there are several reasons for using a multiple linear regression model. These include:

- Creating a model with a predictor that is described by several dummy variables
  - $E(Y_i) = \beta_0 + \beta_2 I_{i2} + \dots + \beta_5 I_{ip}$
- Incorporating nonlinear effects by including polynomial terms of a predictor.
  - $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X^2_i + \dots \beta_p X^p_i + \varepsilon_i$
- Adjusting for confounding.
- Incorporating interactions.

2. Example

The data set contains information from a study of 25 patients with cystic fibrosis. The investigators were interested in assessing predictors of PEmax, a measure of malnutrition. The data set contains a new categorical variable labeled FEV_2 that we will examine more closely this week. The categorical variable FEV_2 has three ordinal levels: 1, 2 and 3. The data set named cf2.sas7bdat is posted on the course web page in Moodle in the folder "Lab 2" under topic 3.

2.1 Multiple Linear Regression with Categorical Predictors

We will begin by considering the impact of the new variable in the data set, FEV_2 on PEmax.

- Create binary indicator variables to represent FEV_2, using level 1 of FEV_2 as the reference level. How many binary indicator variables do you need?

Level 1 is used as the reference level, so levels 2 and 3 will be used as indicator variables relative to level 1. Level 2 will be represented as a 1 if FEV_2 is level 2, and 0 otherwise. For the other binary indicator variable, level 3 will be represented as a 1 if FEV_2 is level 2, and 0 otherwise.

- Write the multiple linear regression model for prediction PEmax from FEV_2, using level 1 of FEV_2 as the reference level.

PEmax = $\beta_0 + \beta_1$*FEV_2L2 + β_2*FEV_2L3 + ε_i

The model is calculating the predicted value of PEmax when FEV_2 is level 1. β_0 is the intercept, or the level of PEmax when FEV_2 is at level 1. The regression coefficients, represented by β_1 and β_2, represent the change in PEmax when moving to level 2 or level 3 from level 1. FEV_2L2 and FEV_2L3 are the binary indicator variables. ε_i is the random error term.

- We will now fit this model in SAS. Interpret the regression coefficients in this model?

```r
#Install and import necessary libraries
#install.packages("readr")
#install.packages("dplyr")
#install.packages("stats")
library(readr)    # For reading CSV file
library(dplyr)    # For data manipulation
library(stats)    # For linear regression

#Import data
cf2 <- read.csv("C:\\Users\\acrot\\Downloads\\cf2.csv")

#View structure of the data
str(cf2)
head(cf2)

#Convert FEV2 to a factor variable
cf2$FEV2 <- factor(cf2$FEV2)

#Create binary indicator variables
predictors <- model.matrix(~ FEV2, data = cf2)

# Convert predictors matrix to data frame
predictors <- as.data.frame(predictors)

#Prepare the response variable
response <- cf2$PEmax

#Fit the multiple linear regression model
model <- lm(response ~ ., data = predictors)

#View the summary of the regression model
summary(model)
```

```
> summary(model)

Call:
lm(formula = response ~ ., data = predictors)

Residuals:
    Min      1Q  Median      3Q     Max 
-24.714  -8.571   0.455   6.429  40.286 

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     78.571      5.548  14.163 1.55e-12 ***
`(Intercept)`       NA         NA      NA       NA    
FEV22           20.974      7.097   2.955  0.00731 ** 
FEV23           76.143      7.846   9.705 2.08e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.68 on 22 degrees of freedom
Multiple R-squared:  0.8234,    Adjusted R-squared:  0.8073 
F-statistic: 51.27 on 2 and 22 DF,  p-value: 5.226e-09
```

The intercept represents the expected value of PEmax when all predictor variables (Level 2 and Level 3) are zero, meaning it is at level 1. This means that when FEV_2 is at level 1, the expected level of PEmax is 78.571, with a standard error of 5.548. For level 2, the test indicates that the value of PEmax will increase from level 1 by 20.974. Level 2 has a p-value of 0.00731, making it statistically significant and indicates a strong relationship between the two variables. For level 3, the test indicates that the value of PEmax will increase from level 1 by 76.143. Level 3 has a p-value of 2.08e-09, indicating an extremely strong relationship between the two variables. The R-squared value indicates that 82.34% of the variation in the model can be explained by predictor variables in the model. The F-statistic's p-value of 5.226e-09 is very low, indicating the statistical significance of the overall model and the strong relationship of the predictor variables to the response variable.

2.2 Confounding

We are interested in examining the impact of Age and FEV_2 on PEmax. In this example, our primary interest is with Age, but we also want to investigate if FEV_2 is a confounder.

First we will investigate confounding. There are two ways to do it.

One way by looking at the association between these three variables directly.

- Calculate Pearson correlation coefficient for continuous variables Age and PEmax. Is r significantly different from 0? Is there association between Age and PEmax?

```
#Calculate Pearson correlation coefficient
correlation <- cor(cf2$Age, cf2$PEmax)

#Print the Pearson correlation coefficient
print(correlation)
```

```
> #Calculate Pearson correlation coefficient
> correlation <- cor(cf2$Age, cf2$PEmax)
>
> #Print the Pearson correlation coefficient
> print(correlation)
[1] 0.6134741
>
```

```
# Perform a hypothesis test
cor_test <- cor.test(cf2$Age, cf2$PEmax)

# Print the hypothesis test results
print(cor_test)
```

```
> # Print the hypothesis test results
> print(cor_test)

	Pearson's product-moment correlation

data:  cf2$Age and cf2$PEmax
t = 3.7255, df = 23, p-value = 0.001109
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2882048 0.8118182
sample estimates:
      cor
0.6134741
```

The Pearson correlation coefficient for Age and PEmax is 0.6134741. A hypothesis test was performed using a two-sided t-test to determine its significant difference from 0. Using a significance level of 0.05, we can assess the p-value. The test shows a p-value of 0.001109, below the 0.05 significance level. This allows us to reject the null hypothesis that that r is not significantly different from 0. Therefore, we can conclude that there is an association between Age and PEmax.

- Investigate the association between FEV_2 and PEmax. Notice that FEV_2 is a categorical variable with 3 levels and PEmax is continuous. What test should we use?

```
#Fit the linear regression model
model2 <- lm(PEmax ~ FEV2, data = cf2)

#Perform ANOVA
anova_result <- anova(model2)

#Print the ANOVA table
print(anova_result)
```

```
> print(anova_result)
Analysis of Variance Table

Response: PEmax
          Df  Sum Sq Mean Sq F value    Pr(>F)    
FEV2       2 22092.8 11046.4  51.272 5.226e-09 ***
Residuals 22  4739.9   215.4                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

An appropriate test for a continuous variable and a categorical variable is an ANOVA, which tests two independent variables on one dependent variable. The results of the ANOVA showed a p-value of 5.226e-09, indicating strong evidence to reject the null hypothesis and that there is a strong relationship between FEV_2 and PEmax.

- Investigate the association between FEV_2 and Age. What test should we use?

```
#Perform one-way ANOVA
anova_1 <- aov(Age ~ FEV2, data = cf2)

#Print the ANOVA table
summary(anova_1)
```

```
> #Print the ANOVA table
> summary(anova_1)
            Df Sum Sq Mean Sq F value  Pr(>F)   
FEV2         2  213.8   106.9   5.873 0.00904 **
Residuals   22  400.4    18.2                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

An appropriate test for the relationship between categorical variable FEV_2 and continuous variable Age would be a one-way ANOVA, which enables testing between three or more means. This will allow us to test the differences in mean Age among the three levels of FEV_2. Based on the p-value of 0.00904, the test indicates strong evidence to reject the null hypothesis and that there is a strong relationship between FEV_2 and Age.

- Assuming that there is no causal relationship between Age and FEV_2, do we think that FEV_2 is a confounder of the relationship between Age and PEmax? Why?

Within linear regression models, a variable can be a confounder if it is associated with X and causally related to the Y, but is not a consequence of X. Since there is no causal relationship between Age and FEV_2, it is less likely that it would be a confounding variable because the confounding variable would likely be associated with PEmax and causally related to Age.

We can also compare the unadjusted β for Age with the adjusted β for Age after controlling for FEV_2 to see if FEV_2 confounds the association between Age and PEmax. Usually, we conclude that FEV_2 is a confounder when we see a change in β of 10% or more.

- To begin, we fit a simple linear regression model with Age alone.

```
#Fit the simple linear regression model for Age alone
model2 <- lm(PEmax ~ Age, data = cf2)

#Print the summary of the regression model
summary(model2)
```

```
> summary(model2)

Call:
lm(formula = PEmax ~ Age, data = cf2)

Residuals:
    Min      1Q  Median      3Q     Max 
-48.666 -17.174   6.209  16.209  51.334 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   50.408     16.657   3.026  0.00601 **
Age            4.055      1.088   3.726  0.00111 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.97 on 23 degrees of freedom
Multiple R-squared:  0.3764,	Adjusted R-squared:  0.3492 
F-statistic: 13.88 on 1 and 23 DF,  p-value: 0.001109
```

- Then we fit the multiple linear regression model with both Age and FEV_2 included.

```
#Fit the multiple linear regression model for Age and FEV2
model3 <- lm(PEmax ~ Age + FEV2, data = cf2)

#Print the summary of the regression model
summary(model3)
```

```
Call:
lm(formula = PEmax ~ Age + FEV2, data = cf2)

Residuals:
    Min      1Q  Median      3Q     Max 
-23.126  -9.979   0.530   9.265  37.109 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  69.3827    10.0984   6.871 8.62e-07 ***
Age           0.7941     0.7305   1.087   0.2893    
FEV22        19.4787     7.2003   2.705   0.0133 *  
FEV23        70.2439     9.5131   7.384 2.90e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.62 on 21 degrees of freedom
Multiple R-squared:  0.8328,	Adjusted R-squared:  0.8089 
F-statistic: 34.86 on 3 and 21 DF,  p-value: 2.438e-08
```

- Assuming that there is no causal relationship between Age and FEV_2, do we think that FEV_2 is a confounder of the relationship between Age and PEmax, after looking at the output from the two above models? Why?

Usually, we conclude that X2 is a confounder when we see a change in beta of 10% (X1) or more. In the regression for only Age, the Estimate was 4.055, the Std. Error was 1.088, and the p-value was 0.00111. In the regression for Age and FEV_2, the estimate was 0.7941, the Std. Error was 0.7305, and the p-value was 0.2893. In addition to there being a change of more than 10% between the two models, the coefficient for Age's p-value changes significantly between the two models. This demonstrates that FEV_2 had an impact on the interaction between Age and PEmax, suggesting it is a confounder of the relationship.

- What is the expected (or average) PEmax score from someone who is Age 16 and has FEV_2 score of 1? FEV_2 score of 2? FEV_2 score of 3?

```
#Convert 'FEV2' to a factor variable
cf2$FEV2 <- as.factor(cf2$FEV2)

#Fit the multiple linear regression model for Age and FEV2
model4 <- lm(PEmax ~ Age + FEV2, data = cf2)

#Create a new data frame with Age 16 and FEV2 levels
new_df <- data.frame(Age = rep(16, 3), FEV2 = factor(c(1, 2, 3), levels = levels(cf2$FEV2)))

#Predict the PEmax scores using the model
predictions <- predict(model4, newdata = new_df)

#Print the predictions
print(predictions)
```

```
> #Print the predictions
> print(predictions)
        1         2         3 
 82.08812 101.56678 152.33201 
> 
```

Based on an age of 16 and an FEV_2 score of 1, the average PEmax score is 82.08812. Based on an age of 16 and an FEV_2 score of 2, the average PEmax score is 101.56678. Based on an age of 16 and an FEV_2 score of 3, the average PEmax score is 152.33201.

2.3 Interactions

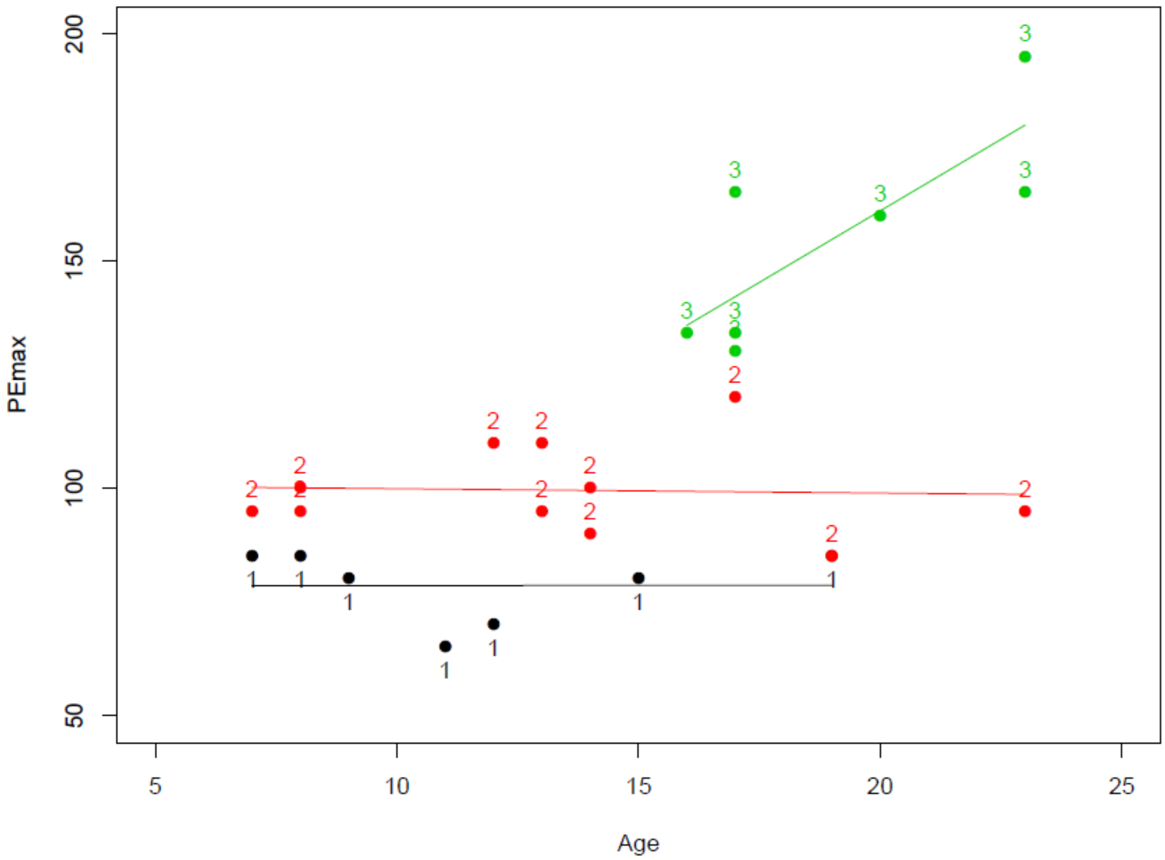
- Using PEmax as a response variable, write out the full model for Age and each level of FEV_2, as well as interaction terms between Age and FEV_2.

The full model would be:

PEmax = β_0 + β_1Age + β_2FEV2L1 + β_3FEV2L2 + β_4FEV2L3 + β_5*(AgeFEV2L1) + β_6(AgeFEV2L2) + β_7(Age*FEV2L3) + ε_i

The model is calculating the predicted value of PEmax as a response variable for age and each level of FEV_2 and the interaction terms between Age and FEV_2. β_0 represents the intercept, or the value of PEmax when Age is equal to zero and FEV_2 is at the reference level, level 1. β_1Age represents the Age coefficient, or the expected change in PEmax when Age increases by one year. β_2FEV2L1, β_3FEV2L2, and β_4FEV2L3 represent the coefficients for FEV_2 at its different levels, or the expected change in PEmax when the level changes from the reference level to the specified level. β_5*(AgeFEV2L1), β_6(AgeFEV2L2), and β_7(Age*FEV2L3) are the coefficients that represent the interaction terms between Age and FEV_2 at its different levels. ε_i represents the error term within the model.

- We will look at this relationship graphically. What do you notice from the plot?



Based on this plot, the line of best fit for level 1 basically has a slope of zero. This makes sense considering that level 1 is the reference level, so the coefficient term would not register a difference between the reference level and the provided level. The line of best fit for level 2 appears to have a very slight downward trend, but also has a slope very close to zero. This was more unexpected, but may indicate no relationship between Age and PEmax at FEV_2 level 2. The third level appears to have a linear relationship between Age and PEmax based on the slope of the line, indicating a potential relationship.

- We will now fit the model with the interaction terms. What do you conclude from the model?

```
#Fit the model with PEmax, Age, FEV2, and Age and FEV2 interactions
model5 <- lm(PEmax ~ Age + FEV2 + Age:FEV2, data = cf2)

#Print the summary of the model
summary(model5)
```

```
Call:
lm(formula = PEmax ~ Age + FEV2 + Age:FEV2, data = cf2)

Residuals:
    Min      1Q  Median      3Q     Max 
-14.899  -8.122  -1.011   6.452  22.878 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  78.494695  13.763484   5.703  1.7e-05 ***
Age           0.006631   1.126469   0.006  0.99536    
FEV22        22.331657  17.478673   1.278  0.21676    
FEV23       -43.410039  33.506791  -1.296  0.21065    
Age:FEV22    -0.101833   1.357011  -0.075  0.94097    
Age:FEV23     6.289665   1.949380   3.226  0.00444 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.69 on 19 degrees of freedom
Multiple R-squared:  0.9032,	Adjusted R-squared:  0.8777 
F-statistic: 35.46 on 5 and 19 DF,  p-value: 5.353e-09
```

Based on the output of this model, we can conclude a few things. First, based on the p-value for Age alone, 0.99536, the relationship between Age and PEmax is not significant when considering it with all other variables in the model. The p-values for both FEV_2 levels, 2 and 3, are 0.21676 and 0.21065, respectively. This is also not statistically significant, indicating no significant difference between the PEmax score of Level 1 and the PEmax score of levels 2 or 3. The interaction between Age and FEV_2 of level 2 was also not statistically significant with a p-value of 0.94097. However, the interaction between Age and FEV_2 of level 3 had a p-value of 0.00444, indicating statistical significance, or that the relationship between Age and PEmax may be affected by an FEV_2 of level 3. The R-squared value indicated the model accounts for 90.32% of the variance in the scores, and the F-test p-value of 5.353e-09 shows that the overall model is statistically significant.

- What is the expected PEmax score from someone who is Age 16 and has FEV_2 score of 1? FEV_2 score of 2? FEV_2 score of 3?

```
#Predict the PEmax scores using the model
predictions2 <- predict(model5, newdata = new_df)

#Print the predictions
print(predictions2)
```

```
> #Print the predictions
> print(predictions2)
        1         2         3 
 78.60080  99.30312 135.82540 
> 
```

Based on the new model, the predicted value of someone with Age 16 and FEV_2 score of 1 is 78.60080. Based on the new model, the predicted value of someone with Age 16 and FEV_2 score of 2 is 99.30312. Based on the new model, the predicted value of someone with Age 16 and FEV_2 score of 3 is 135.82540.

- How does this compare to your previous estimate?

Overall, the predictions are pretty close for levels 1 and 2. Level 1 had an about 4.3% difference. Level 2 had a difference of 2.3%. Level 3 had a larger difference of 11.5%. But overall, the scores appear to fall close together, with around the same amount of difference between levels as well.