BIOST 546 - Homework 1

John Schoof

1/25/2021

Question 1

Question 1.A.

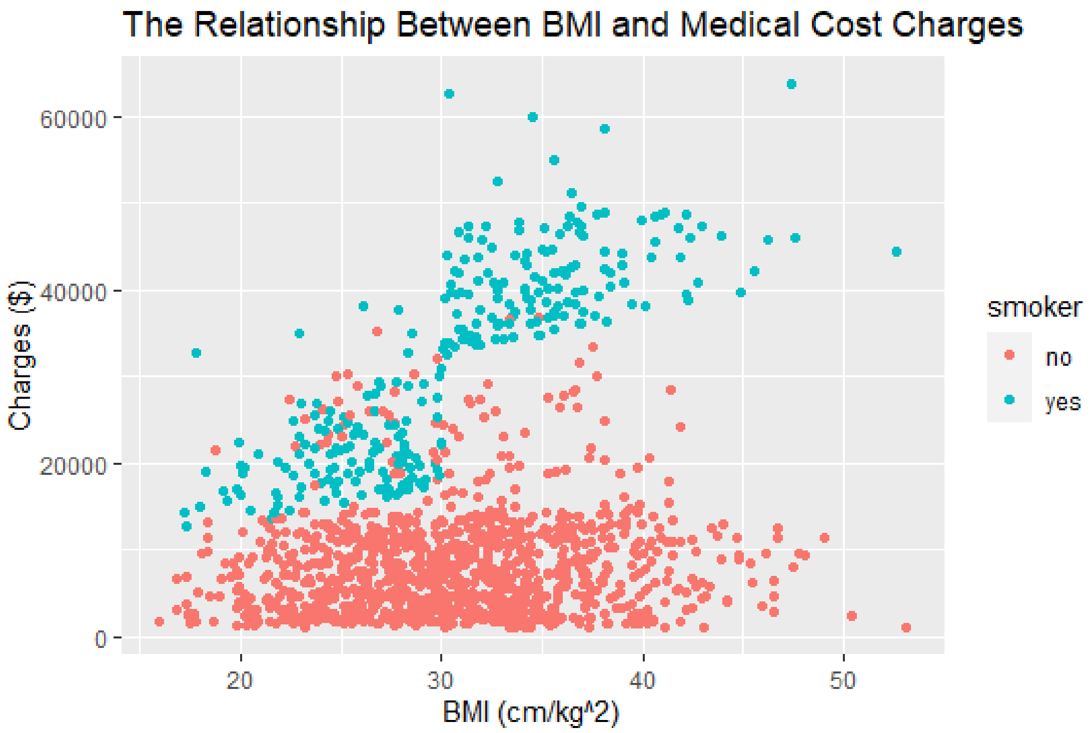
The below code reads in the “Medical_Cost” data and checks for any missing data. There is no missing data.

```
## read in medical cost data
  load("C:/Users/jscho/OneDrive - UW/Winter 2021-LAPTOP-7K6NFTGB/BIOST
546/Homework/Medical_Cost.RData")
  cost <- df
  glimpse(cost)

## Rows: 1,338
## Columns: 7
## $ age      <int> 19, 18, 28, 33, 32, 31, 46, 37, 37, 60, 25, 62, 23, 56, 27...
## $ sex      <fct> female, male, male, male, male, female, female, female, ma...
## $ bmi      <dbl> 27.9, 33.8, 33.0, 22.7, 28.9, 25.7, 33.4, 27.7, 29.8, 25.8...
## $ children <int> 0, 1, 3, 0, 0, 0, 1, 3, 2, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0...
## $ smoker   <fct> yes, no, no, no, no, no, no, no, no, no, no, yes, no, no, ...
## $ region   <fct> southwest, southeast, southeast, northwest, northwest, sou...
## $ charges  <dbl> 16885, 1726, 4449, 21984, 3867, 3757, 8241, 7282, 6406, 28...

  which(is.na(cost))

## integer(0)
```

Question 1.B.

The below scatter plot of charges in dollars on BMI in cm/kg^2. The blue dots represent smokers while the red dots represent non-smokers. The plot suggests that there may be effect modification between smoking and BMI.

```
ggplot(data = cost) +
  geom_point(aes(x=bmi, y= charges, color = smoker)) +
  labs(title = "The Relationship Between BMI and Medical Cost Charges",
       x = "BMI (cm/kg^2)",
       y = "Charges ($)")
```



###

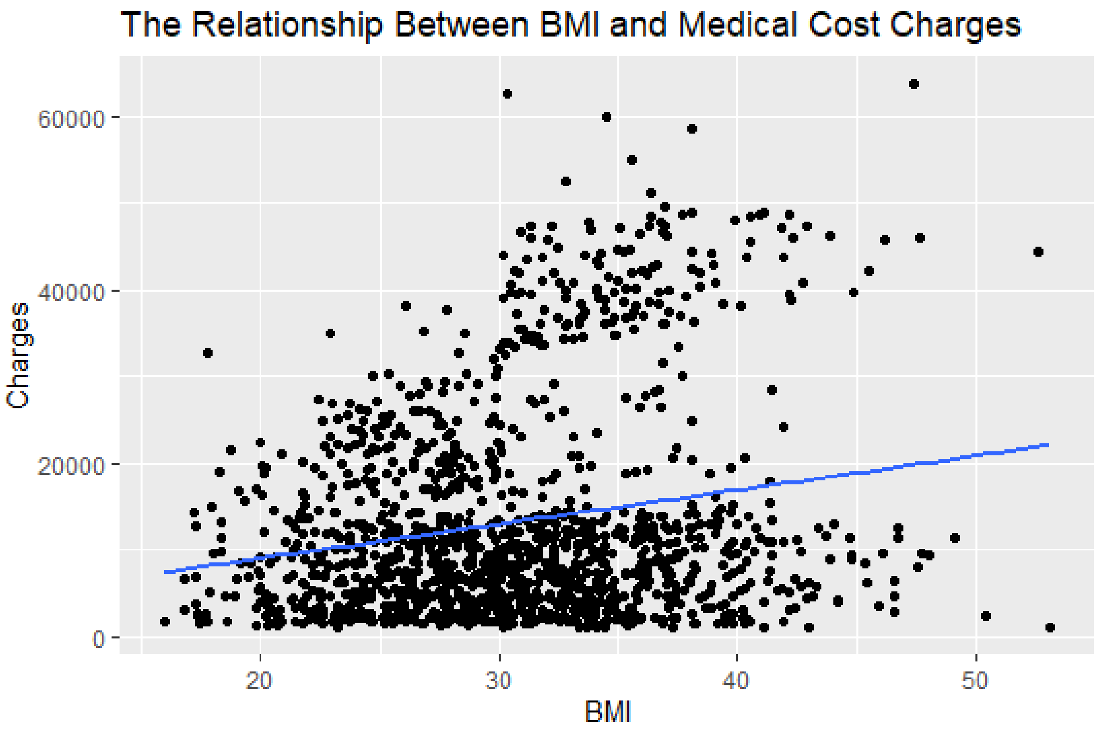
Question 1.C. Fit the following three models

Model 1. The first model regresses BMI on charges.

$$\mathrm{P}(Charges = Y|BMI = x) = \beta_0 + \beta_1 * BMI$$

Below is the table of output and scatter plot with the line of best fit. The Mean Squared Error of the training set is $140,777,900. The predicted charges for someone with a BMI of 32 is $13,797.

Coefficient	Estimate	95% C.I.	P-value	Interpretation
β_0	1192.9	(-2010, 4396)	0.47	The average charges for someone with a BMI equal to zero (not scientifically relevant).
β_1	393.9	(281, 507)	2< 0.001	The average change in charges that corresponds to a one unit change in BMI.



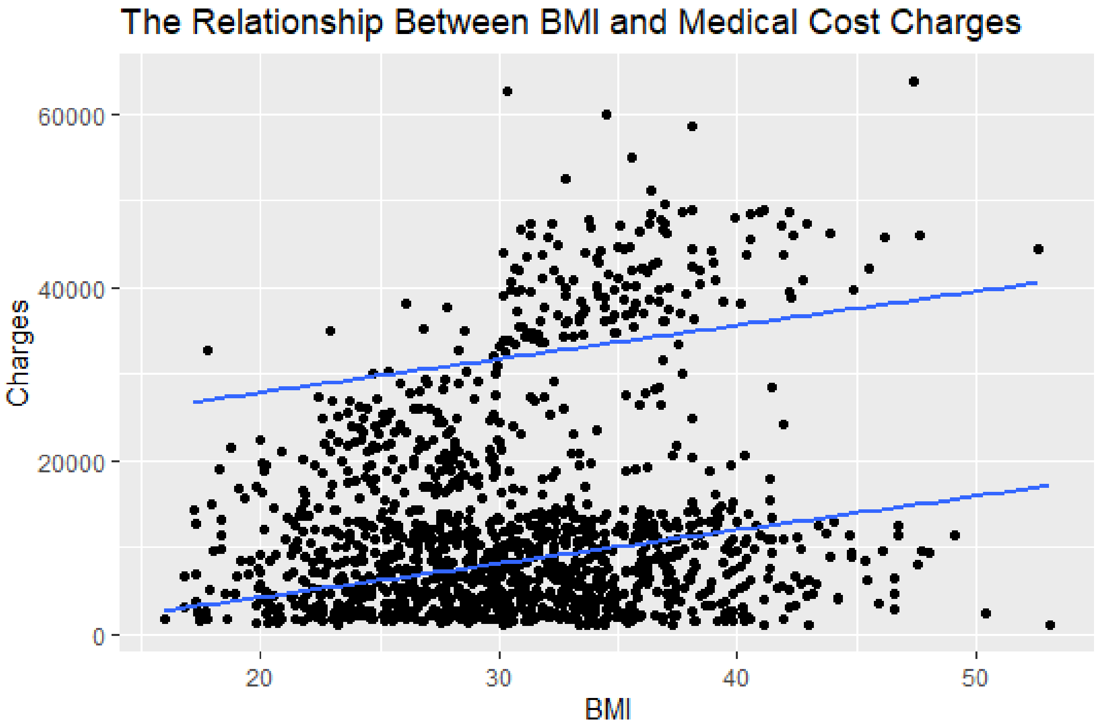
Model 2. The second model regresses medical charges on BMI and smoking status.

$$\mathrm{P}(Charges = Y|BMI, Smoker) = \beta_0 + \beta_1 * BMI + \beta_2 * Smoker$$

Below is the table of output and scatter plot with the line of best fit. The Mean Squared Error of the training set is $50,126,126. The predicted amount of charges for someone with a BMI of 32 is $32,551.

Coefficient	Estimate	95% C.I.	P-value	Interpretation
β_0	-3459	(-5488, -1430)	< 0.001	The average charges for someone who does not smoke and has a BMI equal to zero (not scientifically relevant).
β_1	388	(322, 454)	< 0.001	The average change in charges that corresponds to a one unit change in BMI among individuals with the same smoking status.
β_2	23594	(22390, 24797)	< 0.001	The average change in charges comparing smokers to non-smokers

				among individuals with the same BMI.



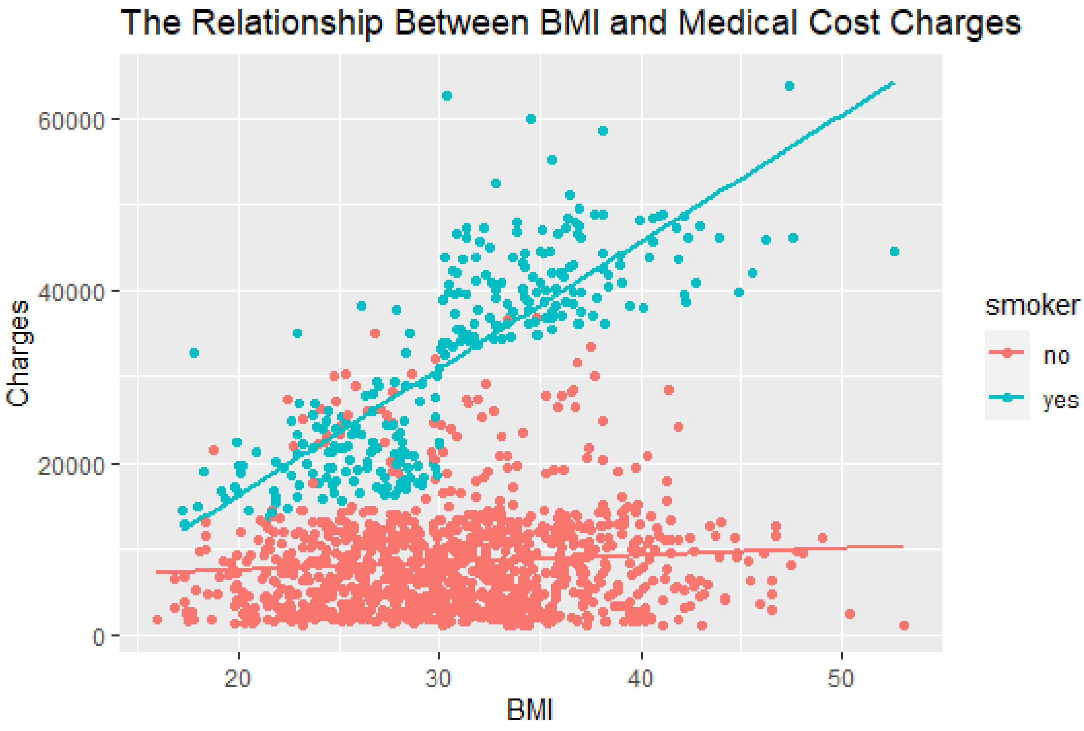
Model 3. The third model regresses medical charges on BMI and smoking status and includes an interaction term between BMI and smoking status.

$$\mathrm{P}(Charges = Y|BMI, Smoker, BMI * Smoker)$$
$$= \beta_0 + \beta_1 * BMI + \beta_2 * Smoker + \beta_3 * Smoker * BMI$$

Below is the table of output and scatter plot with the line of best fit. The Mean Squared Error of the training set is 37,841,585. The predicted charges for someone with a BMI of 32 is $33,954.. If someone were to lower their BMI from 32 to 28 we would expect to see an average decrease in their medical costs of $5,892.

Coefficient	Estimate	95% C.I.	P-value	Interpretation
β_0	5879	(-5488, -1430)	< 0.001	The average charges for someone who does not smoke and has a BMI equal to zero (not scientifically relevant).

β_1	83	(28, 139)	0.007	The average change in charges that corresponds to a one unit change in BMI among individuals with the same smoking status.
β_2	-19066	(-23644, -14488)	< 0.001	The average change in charges comparing smokers to non-smokers among individuals with the same BMI.
β_3	1390	(1239, 1540)	< 0.001	The change in the average change in charges that corresponds to a one unit change in BMI among individuals that smoke.



Question 1.D.

The below model regresses medical charges on BMI, smoking status, smoker_bmi30p and the interaction terms of BMI and smoker_bmi30p and smoking status and smoker_bmi30p.

$$\mathrm{P}(Charges = Y|BMI, Smoker, SmokerBMI30p, BMI * SmokerBMI30p, Smoker * SmokerBMI30p) = \beta_0 + \beta_1 * BMI + \beta_2 * Smoker + \beta_3 * SmokerBMI30p + \beta_4 * BMI * SmokerBMI30p + \beta_5 * Smoker * SmokerBMI30p$$

Below is the table of output.

Coefficient	Estimate	P-value	Interpretation
β_0	5481	< 0.001	The average charges for someone who does not smoke and has a BMI equal to zero (not scientifically relevant).
β_1	96	< 0.001	The average change in charges that corresponds to a one unit change in BMI among individuals with the same smoking status.
β_2	13445	< 0.001	The average change in charges comparing smokers to non-smokers among individuals with the same BMI.
β_3	4690	0.27	The change in the average charges comparing those who smoke and have a BMI > 30 to those who either don't smoke or have a BMI < 30.
β_4	412	< 0.001	Change in the average change in charges that corresponds to a one unit change in BMI among those who smoke and have a BMI > 30.
β_5	NA	NA	No estimate possible due to multicollinearity.

Predictor Coefficients:

BMI: Given that the p-value is much less than 0.05, there is strong evidence to reject the null hypothesis that there is no linear association.

Smoker: Given that the p-value is much less than 0.05, there is strong evidence to reject the null hypothesis that there is no linear association.

Smoker_BMI30p: Given that the p-value is greater than 0.05, evidence to reject the null hypothesis that there is no linear association does not exist.

BMI x Smoker_BMI30p: Given that the p-value is much less than 0.05, there is strong evidence to reject the null hypothesis that there is no linear association.

Smoker x Smoker_BMI30p: Due to multicolieanrity, a coefficient cannot be estimated for this parameter.

Non-significant Predictor Variables:

Smoker_BMI30p: There is no evidence of a linear association between. Our sample would not be unusual if the true value of this coefficient was zero. If this variable was removed from the model, and assuming that means that the interaction term with this variable is also removed from the model, the figure would look like the second model from question 1c.

Question 2

Question 2a - Regression Problems

- **Blood glucose levels as an effect of fat content in diet**: The outcome would be a continuous measure of blood glucose levels and the main predictor of interest would be proportion of calories from fat. Other predictor variables included in the model would be carbohydrate intake, protein intake, level of physical activity, and common demographic variables. This data would likely be low-dimensional.
- **Tumor size as an effect of gene presentation**: The outcome would be a continuous measure of tumor size. The predictors would be tens of thousands of genes in the human genome. This data would be high-dimensional.
- **Health care costs among medicare patients**: The outcome variable would be a continuous measure of health care cost such as dollars per year. The data would likely use electronic medical record data and there would be many predictors from many different doctors visit. This would likely be high dimensional data.

Question 2b - Classification Problems

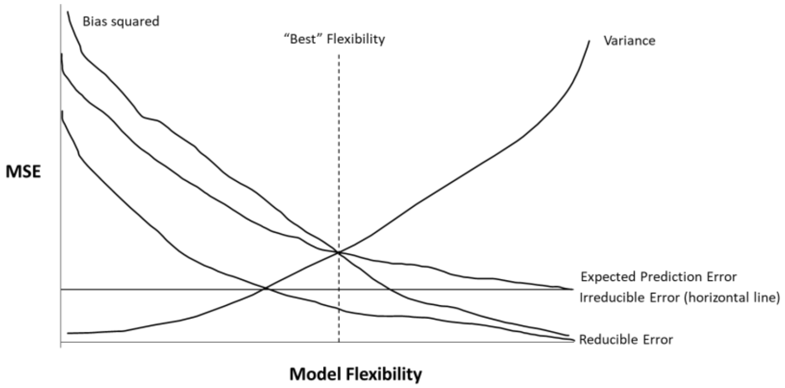
- **Diagnosing breast cancer tumor malignancy using medical imaging**: The outcome variable would be a binary measure of tumor malignancy. The predictors of interest would be each of the pixels in the MRI. This data would likely be high-dimensional.
- **Assessing risk factors for myocardial infarction:** The outcome variable would be a binary diagnosis of a heart attack. The predictors would be all potential risk factors for cardiovascular disease. This data would be low-dimensional.
- **Randomized controlled trial of PrEP and HIV infection:** The outcome is a binary diagnosis of HIV. The predictor is assignment to PrEP or placebo. This data would be low-dimensional.

Question 2c - Unsupervised Learning

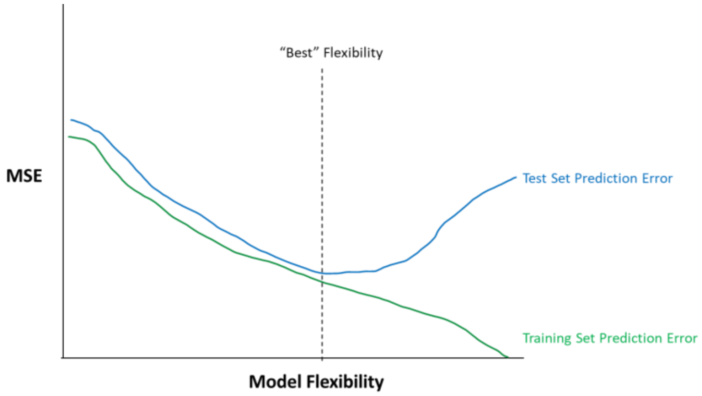
- **Wearable heart monitor data on arrhythmias in the heart to determine most predictive risk factors among younger adults.**: The outcome variable is arrythmias. The predictor is the presence of an irregular readout from the monitor. This data would be high-dimensional
- **Neighborhood risk factors and the risk of injury from gun violence:** The outcome would be the risk of injury from gun violence and the predictors would be many different demographic and spatial indicators of a neighborhood. This data would be high-dimensional.
- **Diagnosing covert stroke with medical imaging.**: The outcome variable would be binary measure of the presence of a covert stroke or not. The predictors of interest would be each of the pixels in the CT or MRI as well as risk factors critical for the occurrence of cardiovascular disease and stroke, such as body mass index (BMI), hemoglobin A1c (HbA1c), systolic and diastolic blood pressure, and smoking status.

Question 3

Question 3a



Question 3b



Question 3c

A simple linear regression model where the number of parameters equals 1 (p=1) would have a very low variance and a very high level of bias. The model could be depicted by a straight line. Therefore, it would have a very low model flexibility and would lie on the far left of the above plot. A K-nearest neighbor model where the K equals 1 (K=1) would have a very high variance and very low level of bias. A depiction of the model would be a line that connects all of the points in the data set. Therefore, it would have a very low model flexibility and would lie on the far left of the above plot.

Question 4

- If interpretability of the analysis is more important than complexity, one would likely decide to use a linear model instead of a non-parametric model.
- If the data includes more than four predictor variables, I would always choose a linear model as opposed to a non-parametric model. This is because with more predictor variables the more flexible, non-parametric model is much more prone to overfitting.

Question 5

Question 5a

Below is the equation for the fitted model and the table of the regression output. The interpretation of the intercept is the average medical charges for individuals from the northwest region. The dummy variable coefficients are interpreted as the difference in average medical charges for individuals from the northeast, southeast, or southwest region and the average medical charges for individuals from the northwest region.

$$\mathrm{P}(Charges = Y|NE01, SE01, SW01) = \beta_0 + \beta_1 * NE01 + \beta_2 * SE01 + \beta_3 * SW01$$

```
cost <- cost %>% mutate(northeast01 = ifelse(region == "northeast", 1, 0),
                        southeast01 = ifelse(region == "southeast", 1, 0),
                        southwest01 = ifelse(region == "southwest", 1, 0))
```

Question 5b

Below is the equation for the fitted model and the table of the regression output. The intercept has no valid interpretation in this case. The dummy variable coefficients are interpreted as the difference in average medical charges for individuals from the northeast, southeast, or southwest region and the average medical charges for individuals from the northwest region.

$$\mathrm{P}(Charges = Y|NE.5, SE.5, SW.5) = \beta_0 + \beta_1 * NE.5 + \beta_2 * SE.5 + \beta_3 * SW.5$$

```
cost <- cost %>% mutate(northeast.5 = ifelse(region == "northeast", 0.5, -0.5),
                        southeast.5 = ifelse(region == "southeast", 0.5, -0.5),
                        southwest.5 = ifelse(region == "southwest", 0.5, -0.5))
```