BIOST546 Homework 1

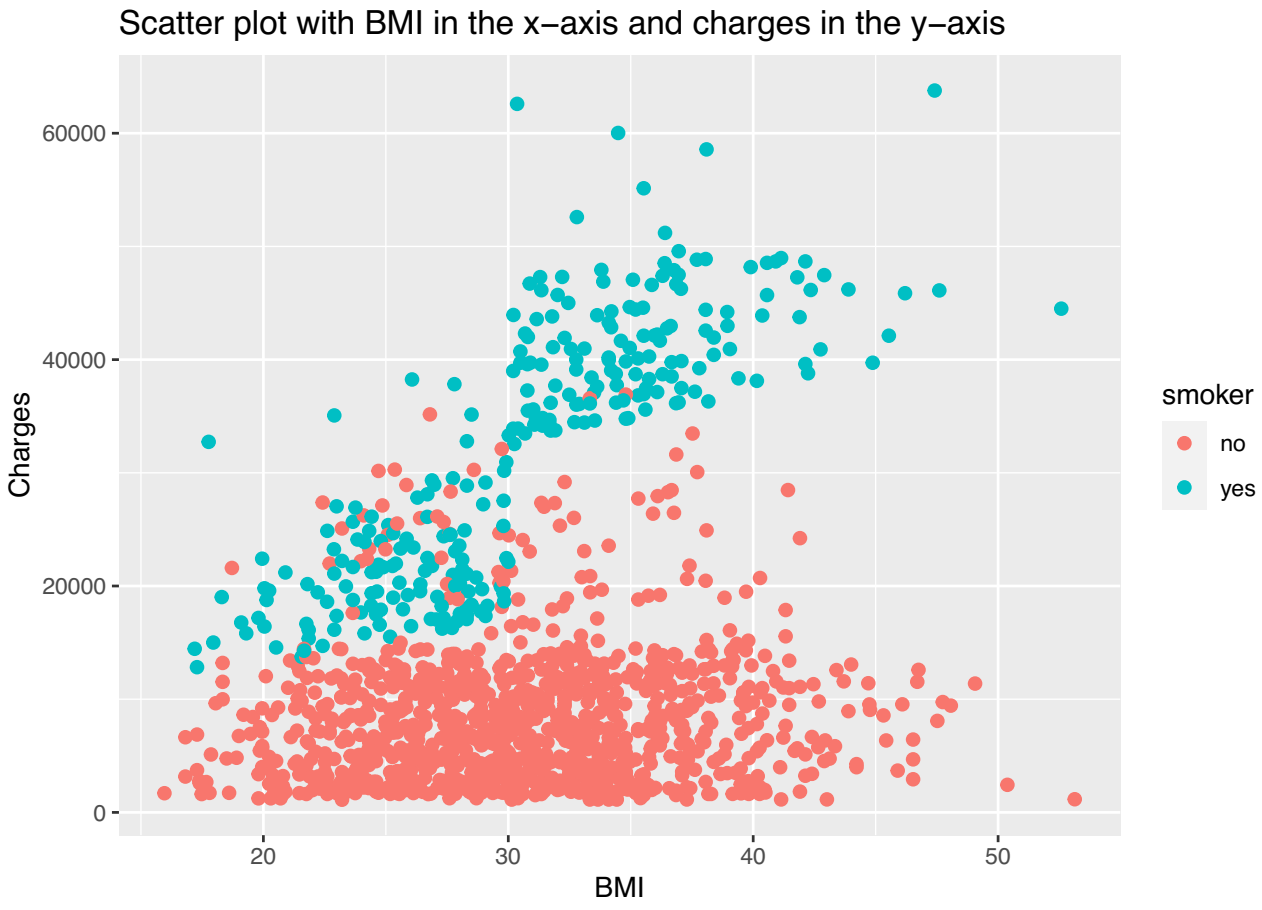
Ivy Zhang

1/12/2022

Problem 1

(a) We load data and check whethere there is null value inside the dataset. There is zero null value in the dataset, therefore we don't need to worry about the missing data.

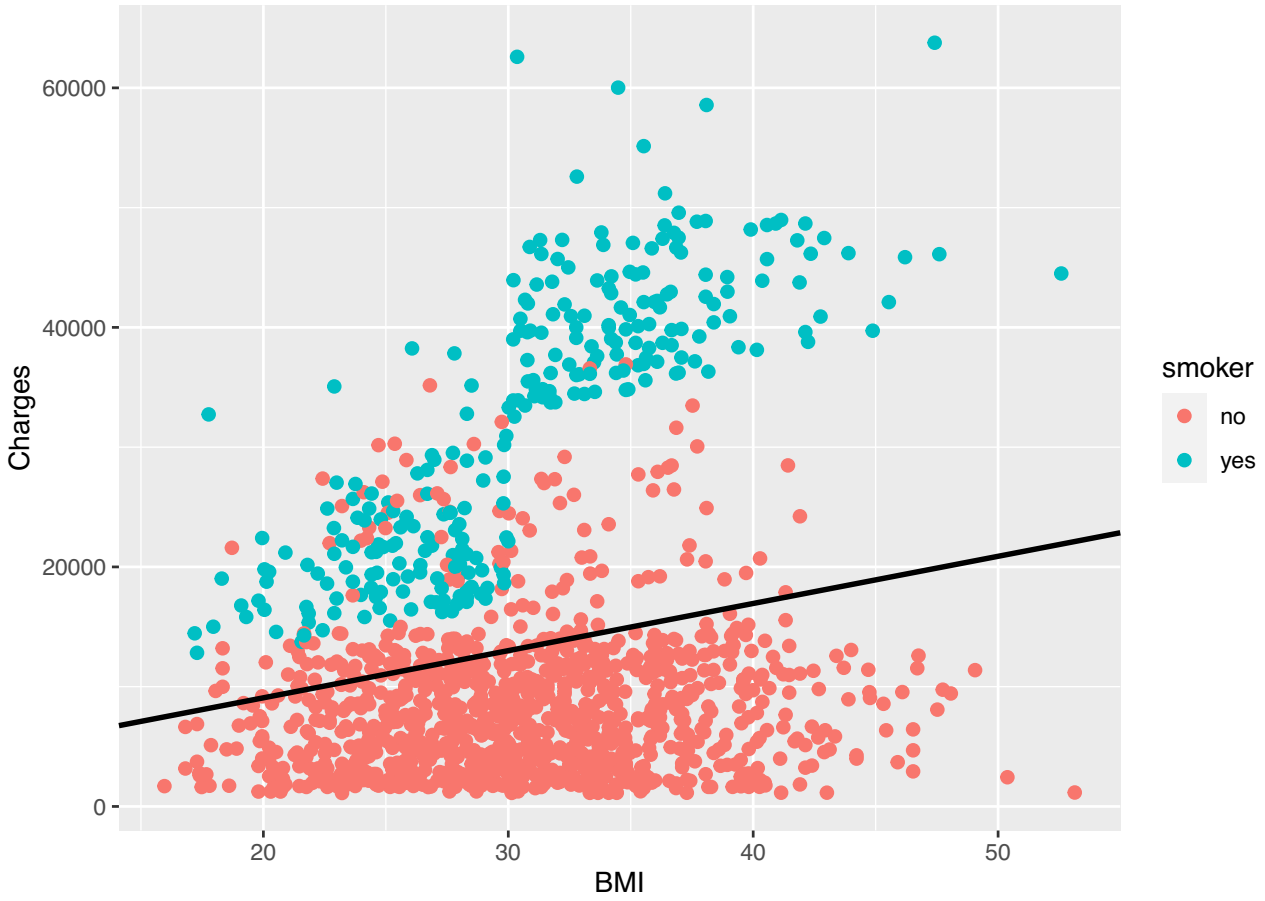
(b)



(c)

Model: Charges using bmi as the only predictor

	Estimated	Intepretation
(Intercept)	1192.94	Estimated charges when BMI equals to 0
bmi	393.87	Estimated change of Charges when BMI increase by 1

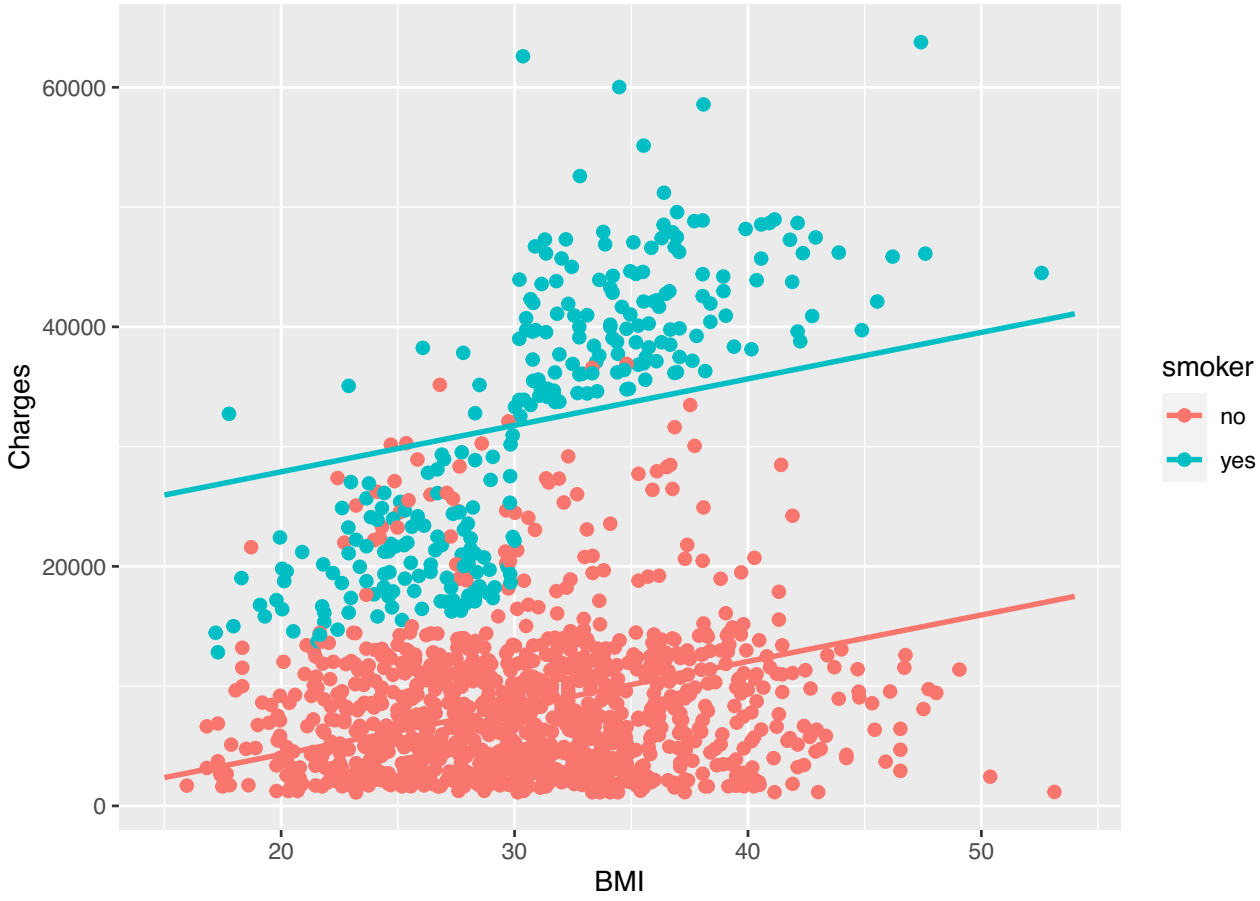


Based on the regression model we fit, we estimated that the 95 confident interval of variable BMI is [289.49, 498.33]. We are 95 percent confidence that the true bmi parameter lies in [289.49, 498.33]. Our data won't be suprising if the true bmi parameter lies in [289.49, 498.33].The mean squared error of the model is 140777900.

We also estimated that a smoker with BMI 29 will be charged for 12615.26, and a smoker with BMI 33 will be charged for 14190.75 with a difference of 1575.49 dollars (lower BMI with lower cost).

Model: Charges using bmi and smoker as predictors

	Estimated	Intepretation
Intercept	-3459.10	Estimated charges when BMI equals to 0
BMI	388.02	Estimated change of Charges when BMI increase by 1 when smoking status remains same
Smoker	23593.98	Estimated difference in charges from non-smoker to smoker with same BMI

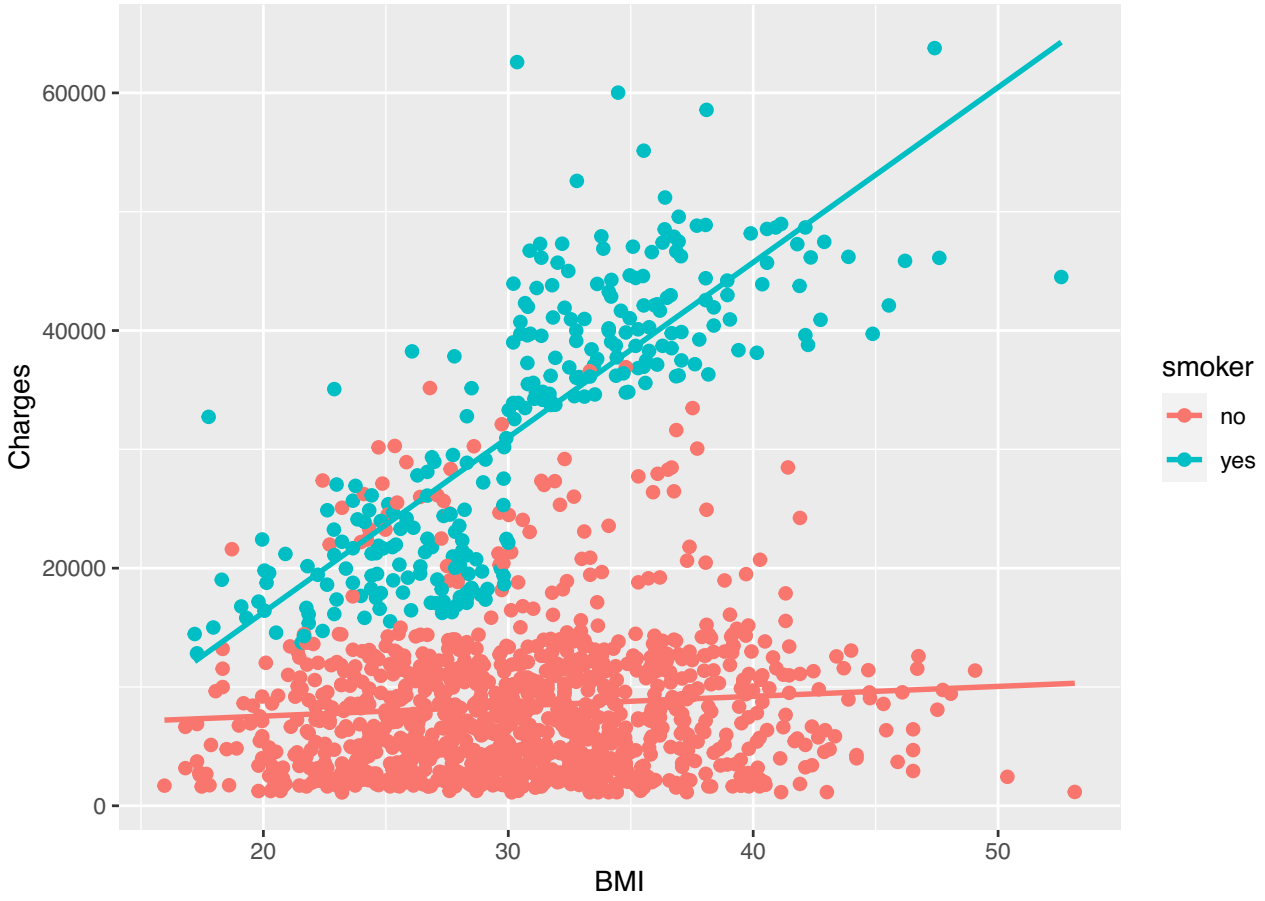


Based on the regression model we fit, we estimated that the 95 confident interval of variable BMI is [325.67, 450.37]. We are 95 percent confidence that the true bmi parameter lies in [325.67, 450.37]. Our data won't be suprising if the true bmi parameter lies in [325.67, 450.37].The mean squared error of the model is 50126126.

We also estimated that a smoker with BMI 29 will be charged for 31387.33, and a smoker with BMI 33 will be charged for 32939.39 with a difference of 1552.06 dollars (lower BMI with lower cost).

Model: Charges using bmi and smoker as predictors and an interaction term between them

Variable	Estimated	Intepretation
Intercept	5879.42	Estimated charges when BMI equals to 0
BMI	83.35	Estimated change of Charges when BMI increase by 1 when smoking status remains same
Smoker	- 19066.00	Estimated difference in charges from non-smoker to smoker when BMI equals zero
BMI X Smoker	1389.76	Estimated change in difference in charges from non-smoker to smoker when BMI increase by 1



Based on the regression model we fit, we estimated that the 95 confident interval of variable BMI is [22.01, 144.69]. We are 95 percent confidence that the true bmi parameter lies in [22.01, 144.69]. Our data won't be surprising if the true bmi parameter lies in [22.01, 144.69].The mean squared error of the model is 37841585.

We also estimated that a smoker with BMI 29 will be charged for 29533.51, and a smoker with BMI 33 will be charged for 35425.93 with a difference of 5892.42 dollars (lower BMI with lower cost).However, for non-smoker, if the participant lower his or her BMI from 33 to 29, it only cause decreasing of 333.402 dollars in charges.

(d)

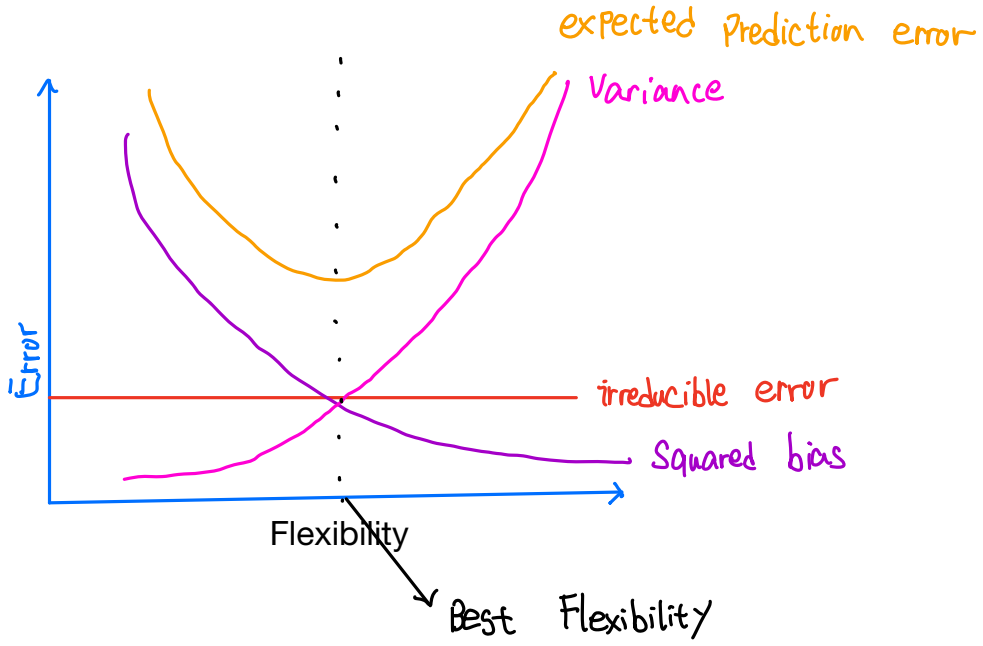
	Estimates	P_Value
Intercept	5879.42	0.000
Smoker&BMI>30	14546.03	0.013
BMI	83.35	0.005
Smoker	3191.77	0.451
Smoker&BMI>30 X BMI	23.43	0.906
Smoker X BMI	401.75	0.015

The previous table shows the estimated coefficients and Wald p-value by fitting model using boolean variable smoker_bmi30p that is true only if the subject is a smoker and has a bmi greater than 30, together with bmi and smoking status to predict the charges. For smoker variable, the new boolean variable smoker_bmi30p, and the interaction term between smoker and bmi, at the significance level of 0.05, we can reject the null

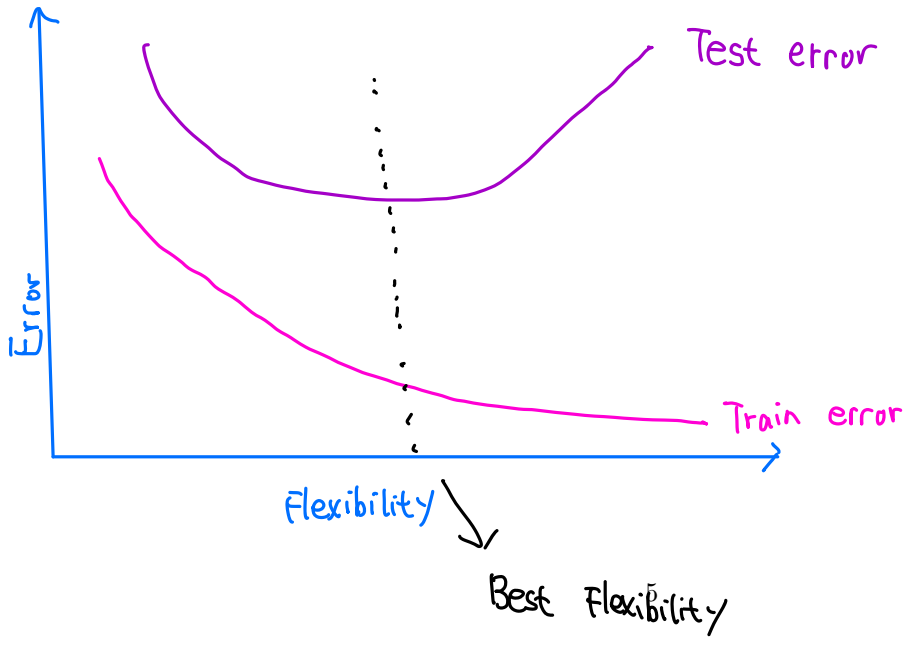
hypothesis there is no (linear) association between each of these variables and charges, conditional on the other predictors in the model. For smoker variable and the interaction term between smoker_bmi30p and bmi variable, we fail to reject the null hypothesis there is no (linear) association between each of these variables and charges, conditional on the other predictors in the model at the significance level of 0.05. For smoker variable, it is interpret as average difference in charges of two individuals whose bmi equals to zero but having different smoking status. If we drop this variable, then the estimated charges with people in different smoking status but bmi equals to 0 will be the same. For the interaction term between smoker_bmi30p and bmi variable, it can be interpreted as the change in two rate of charges in bmi between bmi over 30 group and bmi lower than 30 group for smokers. If we are going to drop out this term, then the slope of the change over bmi will be same for smoker with bmi over 30 group and smoker with bmi below 30 group.

Problem 2

(a)



(b)



Problem 3

For this problem, I will show my code inside my assignment as part of my answer, but will also attach code at the code appendix.

(a)

```
##Problem 3
###(a)
set.seed(0)
X = rnorm(30)
epsilon = rnorm(30)
```

(b)

```
###(b)
Y = 3-2*X+3*(X^2)+epsilon
```

(c)

```
###(c)
lm3c1 = lm(Y~X)
lm3c2 = lm(Y~X+I(X^2))
lm3c3 = lm(Y~X+I(X^2)+I(X^3)+I(X^4))
summary(lm3c1)
summary(lm3c2)
summary(lm3c3)
```

$f(X) = \beta_0 + \beta_1 \times X$: We estimate that β_0 equals to 5.3697 and β_1 equals to -0.6614.

$f(X) = \beta_0 + \beta_1 \times X + \beta_2 \times X^2$: We estimate that β_0 equals to 3.0242 and β_1 equals to -1.9953, and β_2 equals to 2.9369.

$f(X) = \beta_0 + \beta_1 \times X + \beta_2 \times X^2 + \beta_3 \times X^3 + \beta_4 \times X^4$: We estimate that β_0 equals to 3.0761 and β_1 equals to -1.5030, β_2 equals to 2.6677, β_3 equals to -0.3147, and β_4 equals to 0.1401.

(d)

```
#(d)
(mse_lm3c1 = mean(summary(lm3c1)$residuals^2))
(mse_lm3c2 = mean(summary(lm3c2)$residuals^2))
(mse_lm3c3 = mean(summary(lm3c3)$residuals^2))
```

$f(X) = \beta_0 + \beta_1 \times X$ train MSE: 10.86177

$f(X) = \beta_0 + \beta_1 \times X + \beta_2 \times X^2$ train MSE:1.019437

$f(X) = \beta_0 + \beta_1 \times X + \beta_2 \times X^2 + \beta_3 \times X^3 + \beta_4 \times X^4$ train MSE:0.9855839

From the previous result, we can clearly see, if we include more parameters into our models and give more flexibility to the model, the training MSE will keep decreasing.

(e)

```
X_test = rnorm(10000)
epsilon_test = rnorm(10000)
Y_test= 3-2*X_test+3*(X_test^2)+epsilon_test
X_test_pred_lm3c1 = predict(lm3c1, data.frame(X = X_test))
X_test_pred_lm3c2 =  predict(lm3c2, data.frame(X = X_test))
X_test_pred_lm3c3 =  predict(lm3c3, data.frame(X = X_test))
(mse_X_test_lm3c1 = mean((X_test_pred_lm3c1 - Y_test)^2))
(mse_X_test_lm3c2 = mean((X_test_pred_lm3c2 - Y_test)^2))
(mse_X_test_lm3c3 = mean((X_test_pred_lm3c3 - Y_test)^2))
```

$f(X) = \beta_0 + \beta_1 \times X$ test MSE: 19.77607

$f(X) = \beta_0 + \beta_1 \times X + \beta_2 \times X^2$ test MSE:1.004828

$f(X) = \beta_0 + \beta_1 \times X + \beta_2 \times X^2 + \beta_3 \times X^3 + \beta_4 \times X^4$ test MSE:2.44185

For test MSE, we can see model 2 has lowest MSE. As parameter number increase, test MSE first decrease and then increase. Test MSE do not keep decreasing when we give more flexibility to the model.

(g)

```
##(g)
X_E = matrix(rnorm(30*40), ncol = 40)
epsilon_e = matrix(rnorm(30*40), ncol = 40)
Y_E = 3-2*X_E+3*(X_E^2)+epsilon_e
predict_3g = matrix(NA, ncol = 3, nrow =40 )
for(i in 1:40){
  X_temp = X_E[,i]
  lm_g1 = lm(Y_E[,i]~X_temp)
  lm_g2 = lm(Y_E[,i]~X_temp+I(X_temp^2))
  lm_g3 = lm(Y_E[,i]~X_temp+I(X_temp^2)+I(X_temp^3)+I(X_temp^4))
  predict_3g[i,1] = predict(lm_g1, data.frame(X_temp = 0.3))
  predict_3g[i,2] = predict(lm_g2, data.frame(X_temp = 0.3))
  predict_3g[i,3] = predict(lm_g3, data.frame(X_temp = 0.3))
}
true_x0 = 3 - 0.3*2+(0.3^2)*3
for(i in 1:3){
  print((mean(predict_3g[,i])-true_x0)^2)
  print(var(predict_3g[,i]))
}
```

The estimated bias for the three models are 6.552, 0.00029, 0.0007.

The estimated variance for the three models are 0.82, 0.045, 0.072.

From the calculation, we can see bias decrease and increase a very small amount when we give more flexibility to the model by adding more parameters.However, the variance decrease at first and increase again when we adding estimated parameters. The difference of bias and variance between the first model and second model is much larger than the differences between the second model and the third model. Since MSE is positively associated with the square of bias and variance, it fits the trend of test MSE what we have calculated at 3(e).

Code Appendix

```
#Set Up
knitr::opts_chunk$set(echo =FALSE, warning = FALSE, message = FALSE,
                      fig.height =5, fig.width = 7)
###Question 1
##(a)
#Loading Data
load("/Users/ivyyuezhang/Desktop/R hw/Medical_Cost.RData")
#Checking Missing Data
sum(is.na(df))
##(b)
library(ggplot2)
ggplot(data = df)+geom_point(aes(x = bmi, y = charges, col = smoker),
                             size = 2)+
  labs(title = "Scatter plot with BMI in the x-axis and charges in the y-axis",
       x ="BMI", y = "Charges")
##(c)
#Model: Charges using bmi as the only predictor
library(knitr)
library(dplyr)
lm_c1 = lm(charges~bmi, data = df)
t1 = data.frame(Estimated = round(lm_c1$coefficients,2),
                Intepretation = c("Estimated charges when BMI equals to 0",
                                  "Estimated change of Charges when BMI increase by 1"))
kable(t1)
ggplot(data = df)+geom_point(aes(x = bmi, y = charges, col = smoker),
                             size = 2) +
    geom_abline(slope = coef(lm_c1)[2], size = 1,
              intercept = coef(lm_c1)[1])+
  labs( x ="BMI", y = "Charges")
confint(lm_c1)
mean(summary(lm_c1)$residuals^2)
predict(lm_c1 , data.frame(bmi = c(29,33)))
#Model: Charges using bmi and smoker as predictors
lm_c2 = lm(charges~bmi+smoker, data = df)
t2 = data.frame(Estimated = round(lm_c2$coefficients,2),
                Intepretation = c("Estimated charges when BMI equals to 0",
                                  "Estimated change of Charges when BMI increase by 1 when smoking statu
                                  "Estimated difference in charges from non-smoker to smoker with same
rownames(t2) = c("Intercept","BMI","Smoker")
kable(t2)

covariates_pred = data.frame(bmi = seq(from = 15, to = 54, by = 1), smoker = "yes")
covariates_pred = rbind(covariates_pred,data.frame(bmi = seq(from = 15, to = 54, by = 1),
                                                   smoker = "no"))
pred = predict(lm_c2, covariates_pred)
DatawithPred = cbind(covariates_pred, pred)


ggplot(data = df,aes(x = bmi, y = charges, col = smoker))+
  geom_point(size = 2) +
  geom_line(mapping=aes(x = bmi, y=pred),size=1, data = DatawithPred)+
```

```
  labs( x ="BMI", y = "Charges")
confint(lm_c2)
mean(summary(lm_c2)$residuals^2)
predict(lm_c2 , data.frame(bmi = c(29,33),smoker = "yes"))
#Model: Charges using bmi and smoker as predictors and an interaction term between them
lm_c3 = lm(charges~bmi*smoker, data = df)
t3 = data.frame(Variable = c("Intercept","BMI","Smoker","BMI X Smoker"),
                Estimated = round(lm_c3$coefficients,2),
                Intepretation = c("Estimated charges when BMI equals to 0",
                                  "Estimated change of Charges when BMI increase by 1 when smoking statu
                                  "Estimated difference in charges from non-smoker to smoker when BMI e
                                  "Estimated change in difference in charges from non-smoker to smoker w
kable(tibble(t3))
ggplot(data = df,aes(x = bmi, y = charges, col = smoker))+
  geom_point(size = 2) +
  geom_smooth(method = "lm",se = F)+
  labs( x ="BMI", y = "Charges")
confint(lm_c3)
mean(summary(lm_c3)$residuals^2)
predict(lm_c3 , data.frame(bmi = c(29,33),smoker = "yes"))
predict(lm_c3 , data.frame(bmi = c(29,33),smoker = "no"))
##1(d)
df  = df %>% mutate(smoker_bmi30p = (smoker == "yes" & bmi >30))
lm_d = lm(charges~smoker_bmi30p*bmi+smoker*bmi,data = df)
td = data.frame(Estimates = round(coef(lm_d),2),
                P_Value = round(summary(lm_d)$coefficients[,4],3))
rownames(td) = c("Intercept","Smoker&BMI>30","BMI","
                           Smoker","Smoker&BMI>30 X BMI","Smoker X BMI")
kable(td)
##Problem 3
###(a)
set.seed(0)
X = rnorm(30)
epsilon = rnorm(30)
###(b)
Y = 3-2*X+3*(X^2)+epsilon
###(c)
lm3c1 = lm(Y~X)
lm3c2 = lm(Y~X+I(X^2))
lm3c3 = lm(Y~X+I(X^2)+I(X^3)+I(X^4))
summary(lm3c1)
summary(lm3c2)
summary(lm3c3)
#(d)
(mse_lm3c1 = mean(summary(lm3c1)$residuals^2))
(mse_lm3c2 = mean(summary(lm3c2)$residuals^2))
(mse_lm3c3 = mean(summary(lm3c3)$residuals^2))

X_test = rnorm(10000)
epsilon_test = rnorm(10000)
Y_test= 3-2*X_test+3*(X_test^2)+epsilon_test
X_test_pred_lm3c1 = predict(lm3c1, data.frame(X = X_test))
X_test_pred_lm3c2 =  predict(lm3c2, data.frame(X = X_test))
```

```
X_test_pred_lm3c3 =  predict(lm3c3, data.frame(X = X_test))
(mse_X_test_lm3c1 = mean((X_test_pred_lm3c1 - Y_test)^2))
(mse_X_test_lm3c2 = mean((X_test_pred_lm3c2 - Y_test)^2))
(mse_X_test_lm3c3 = mean((X_test_pred_lm3c3 - Y_test)^2))

##(g)
X_E = matrix(rnorm(30*40), ncol = 40)
epsilon_e = matrix(rnorm(30*40), ncol = 40)
Y_E = 3-2*X_E+3*(X_E^2)+epsilon_e
predict_3g = matrix(NA, ncol = 3, nrow =40 )
for(i in 1:40){
  X_temp = X_E[,i]
  lm_g1 = lm(Y_E[,i]~X_temp)
  lm_g2 = lm(Y_E[,i]~X_temp+I(X_temp^2))
  lm_g3 = lm(Y_E[,i]~X_temp+I(X_temp^2)+I(X_temp^3)+I(X_temp^4))
  predict_3g[i,1] = predict(lm_g1, data.frame(X_temp = 0.3))
  predict_3g[i,2] = predict(lm_g2, data.frame(X_temp = 0.3))
  predict_3g[i,3] = predict(lm_g3, data.frame(X_temp = 0.3))
}
true_x0 = 3 - 0.3*2+(0.3^2)*3
for(i in 1:3){
  print((mean(predict_3g[,i])-true_x0)^2)
  print(var(predict_3g[,i]))
}
```