Machine Learning Homework 1

Latera Tesfaye Olana

25 January, 2023

Answers:

Question 1: The data has some missing values. Table 1 and Figure 1 provide a summary of missing values across variables. There were 1338 observations before *na.omit* and 1278 observations after.

Table 1: Missing Patterns

Variables	Number of missing observations
bmi	10
region	10
charges	9
sex	9
smoker	9
age	8
children	5

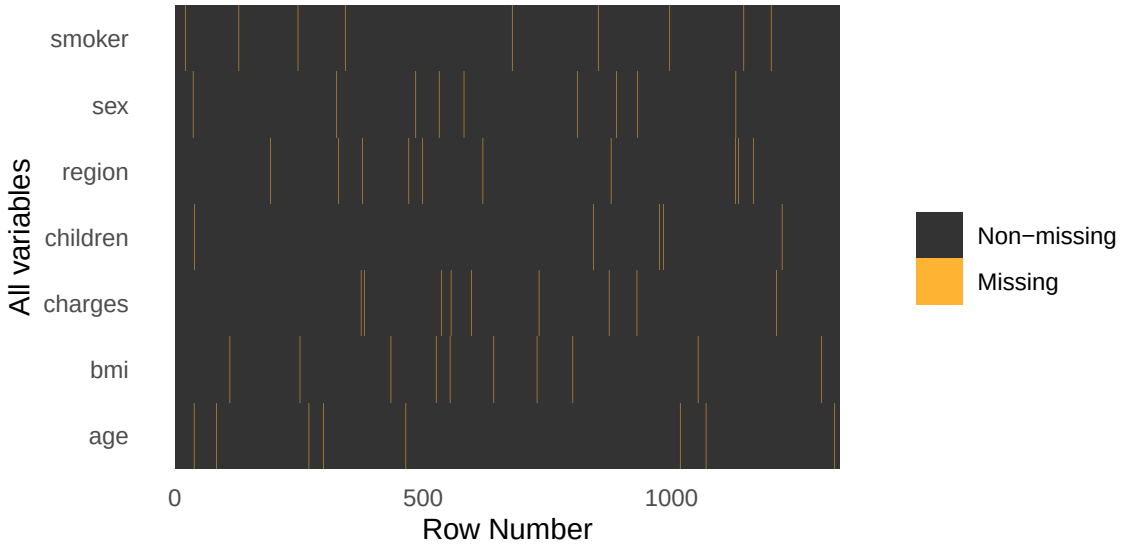


Figure 1: Plotting missing values

Figure 2 shows the scatter plot with body mass index - BMI (kg/m^2) on the x-axis and charges (dollars) on the y-axis. The graph is also stratified by smoking status. I believe there is a linear relationship between

charges and BMI. In other words, from the given sample data, with an increase in BMI there seems to be an increase in charges. Accordingly, the first order trend suggestive of a tendency for higher average charges in higher BMI groups. As shown on the figure, there seems to be some suggestion of greater variability in charges in higher BMI groups than there is in the smaller BMI groups. Even though, reduced the variability still exists after the stratification by smoking. There are no striking outliers.

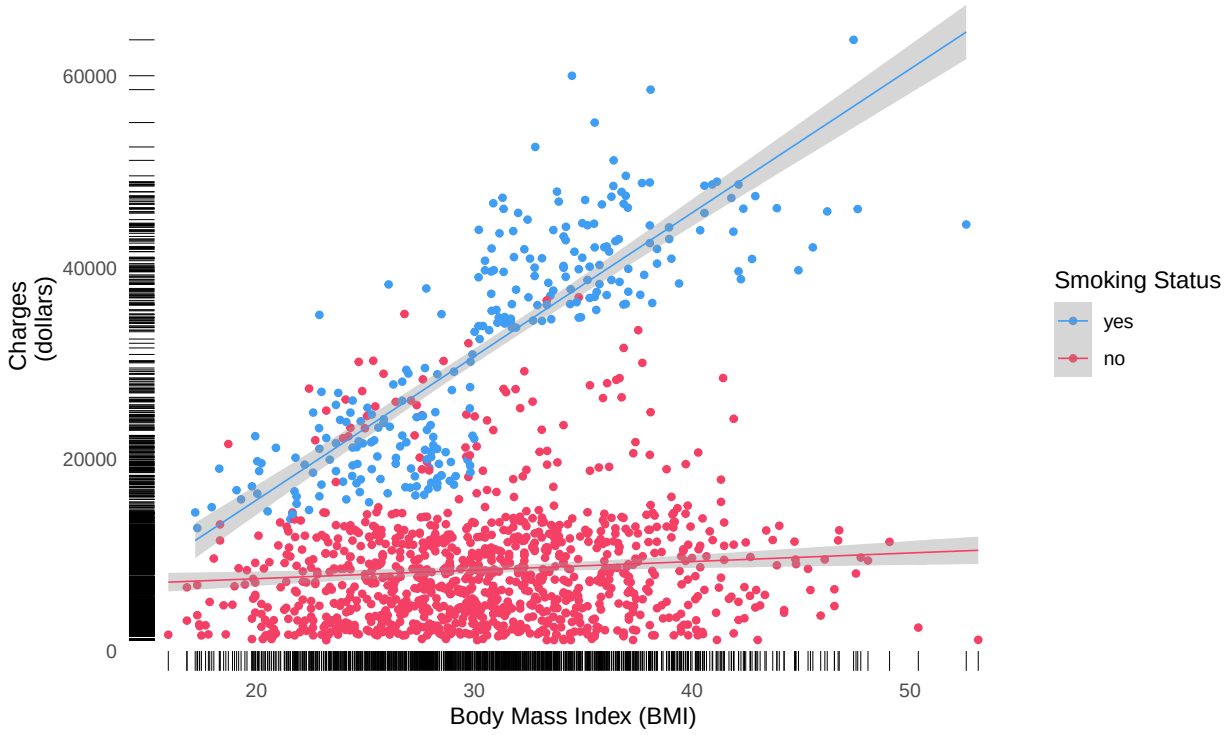


Figure 2: Body Mass Index (BMI) and charges scatter plot

Fitting models:

The First model -

$$charges = \beta_0 + \beta_1 * bmi + e$$

Based on a simple linear model, we estimate that the difference in mean charges in our study group differing by one unit of BMI is 402.65, with the higher BMI groups having higher mean charges (95% confidence interval [287.85; 517.45]). In this study group, we found evidence of an association between charges and BMI ($p < 0.05$). Accordingly, we reject the null hypothesis of non linear trend in the expected value of charges as a function of BMI. The estimated mean value for individuals with zero BMI is 938.44. *The intercept is just a mathematical construct allowing us to fit a line over the range of our data* and it might not be scientifically meaningful. The mean squared error is calculated as:

$$\text{MSE}_{\text{Train}} = \text{MSE}(\hat{f}, \text{Train Data}) = \frac{1}{n_{\text{Tr}}} \sum_{i \in \text{Train}} \left(y_i - \hat{f}(\mathbf{x_i})\right)^2$$

The calculated mean squared error is 1.3835837×10^8. Table 2, provides summary of the first simple model. Figure 3, shows the fitted line over the given data for this first model.

Table 2: OLS - Linear fit of charges and body mass index (BMI)

	Estimate	Naive SE	Robust SE	95%L	95%H	t value	Pr(>|t|)
(Intercept)	938.4422	1682.0009	1656.0268	-2310.3925	4187.277	0.5667	0.571
bmi	402.6474	53.8462	58.5178	287.8458	517.449	6.8808	0.000

According to this model, a smoker with a BMI of 31.5 would be charged 1.3621835×10^4 dollars, where as a smoker with a BMI of of 29 would be charged 1.2615217×10^4. Reducing their BMI to 29 is associated with a change in mean estimated cost of 1006.62 dollars. Figure 3, shows the plot for this model.

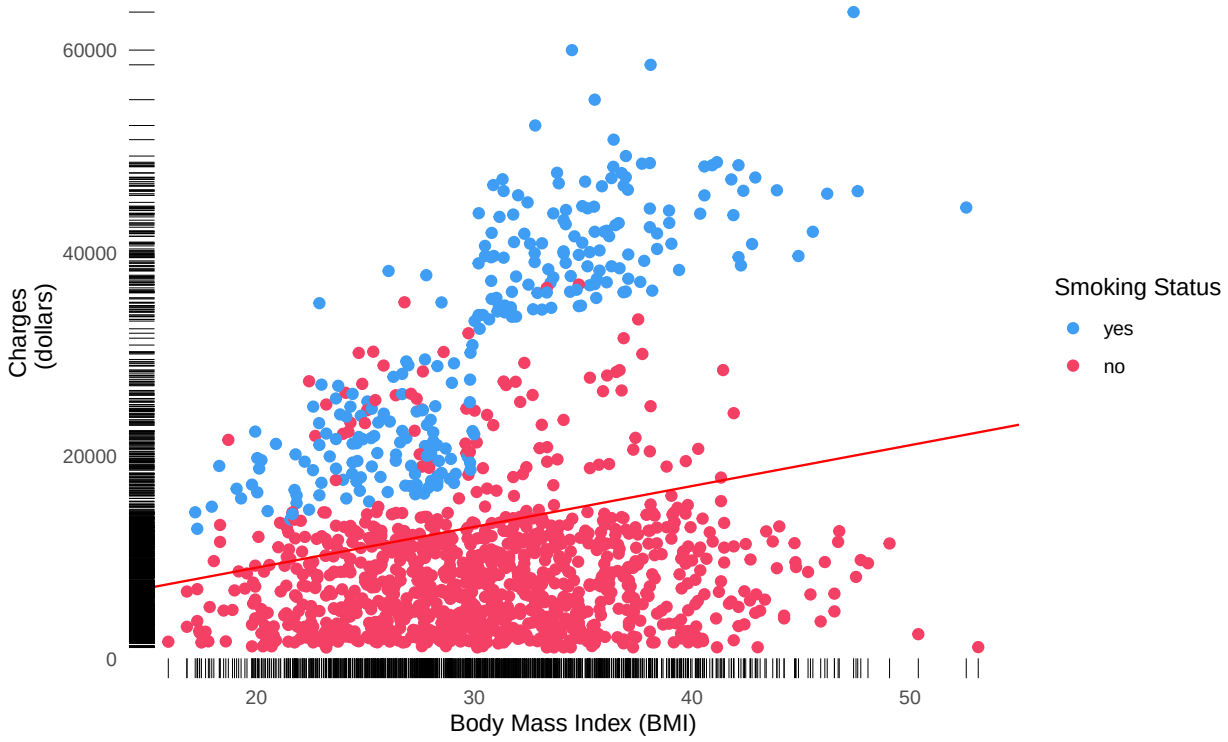


Figure 3: Model 1, plot

The second model -

$$charges = \beta_0 + \beta_1 * bmi + \beta_2 * smoker + e$$

The following table show the summary of linear fit for this model.

Table 3: OLS - Linear fit of charges with body mass index (BMI) and smoking status

	Estimate	Naive SE	Robust SE	95%L	95%H	t value	Pr(>|t|)
(Intercept)	-3711.6846	1018.7759	1061.4637	-5794.0920	-1629.277	-3.4968	5e-04
bmi	398.4665	32.4620	34.7006	330.3901	466.543	11.4830	0e+00
smokeryes	23218.7734	491.0418	618.0295	22006.3069	24431.240	37.5690	0e+00

The estimate difference in the expected value of charges in members of individuals with the same BMI: differing by smoking status is 2.321877×10^4 (with 95% confidence interval 22006.31; 24431.24) dollars. The estimate difference in the expected value of charges in study group differing in smoking status is 398.47 (with 95% confidence interval 330.39; 466.54) dollars. Since the estimated p-value in both cases are less than 0.05, the data is consistent with the above stated statements. To summarize, in this study group, we found significant relationships between BMI and charges and the habit of smoking and charges for medical insurance ($p < 0.001$ for each). Specifically we found a 398.47% increase in the amount of dollars charged for every 1% increase in the value of BMI, and a 2.321877×10^4% increase in the amount of dollars charged for every 1% increase in smoking. This model has a mean squared of 5.0246296×10^7. According to this model, a smoker with a BMI of 31.5 would be charged 3.205878×10^4 dollars, where as a smoker with a BMI of of 29 would be charged 3.106262×10^4. A non-smoker with a BMI of 31.5 would be charged 8840.01. Where as, a non smoker with a BMI of 29 would be charged 7843.84. Reducing a BMI of smokers to 29 is associated with a change in mean estimated cost of 996.17 dollars. Where as, for a non- smokers this reduction would result in 996.17 dollars. Figure 4, shows the plot of the model. The mean squared error for this model is 5.0246296×10^7.

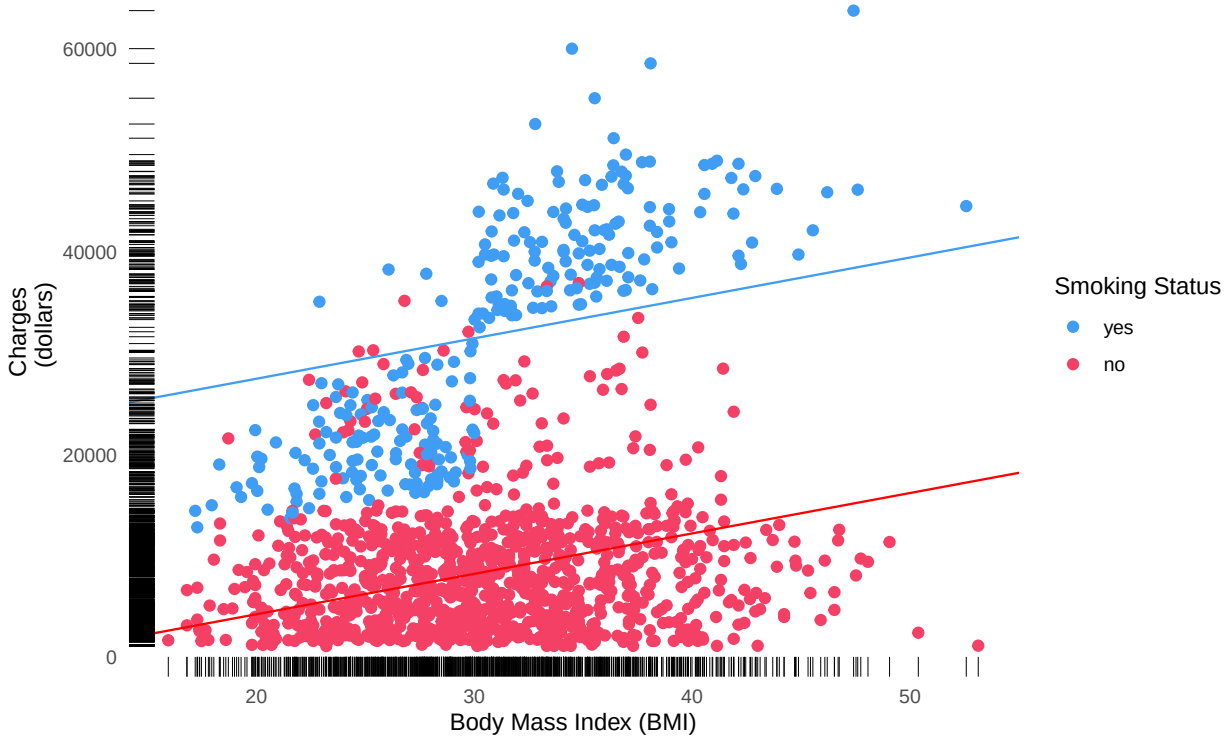


Figure 4: Model 2, plot:

The third model -

$$charges = \beta_0 + \beta_1 * bmi + \beta_2 * smoker + \beta_3 * bmi * smoker + e$$

. The following table shows the summary fit of this model.

Table 4: OLS - Linear fit of charges with body mass index (BMI) and smoking status

	Estimate	Naive SE	Robust SE	95%L	95%H	t value	Pr(>|t|)
(Intercept)	5750.97	991.45	907.22	3971.16	7530.79	6.34	0
bmi	89.47	31.76	29.33	31.94	147.01	3.05	0
smokeryes	-20008.39	2122.67	2283.14	-24487.51	-15529.27	-8.76	0
bmi:smokeryes	1410.05	67.84	76.40	1260.17	1559.94	18.46	0

According to Model 3 calculations, a unit increase in Body Mass Index (BMI) is projected to result in an additional medical expense of 89.47 (95% confidence interval of 29.33; 31.94). Furthermore, individuals who smoke can expect to incur an additional cost of 1410.05 (95% confidence interval of 76.4; 1260.17) per unit increase in BMI. The data suggests that transitioning from being a non-smoker to a smoker is associated with a decrease in cost of 2.000839×10^4 (95% confidence interval 2283.14; 2.448751×10^4). According to this model, a smoker with a BMI of 31.5 would be charged 3.297771×10^4 dollars, where as a smoker with a BMI of of 29 would be charged 2.922889×10^4. A non-smoker with a BMI of 31.5 would be charged 8569.4, where as, a non smoker with a BMI of 29 would be charged 8345.72 dollars. Reducing a BMI of smokers to 29 is associated with a change in mean estimated cost of 3748.82 dollars. Where as, for a non- smokers this reduction would result in 223.68 dollars. Figure 5, shows the plot of the model. The mean squared error for this model is 3.7522941×10^7.

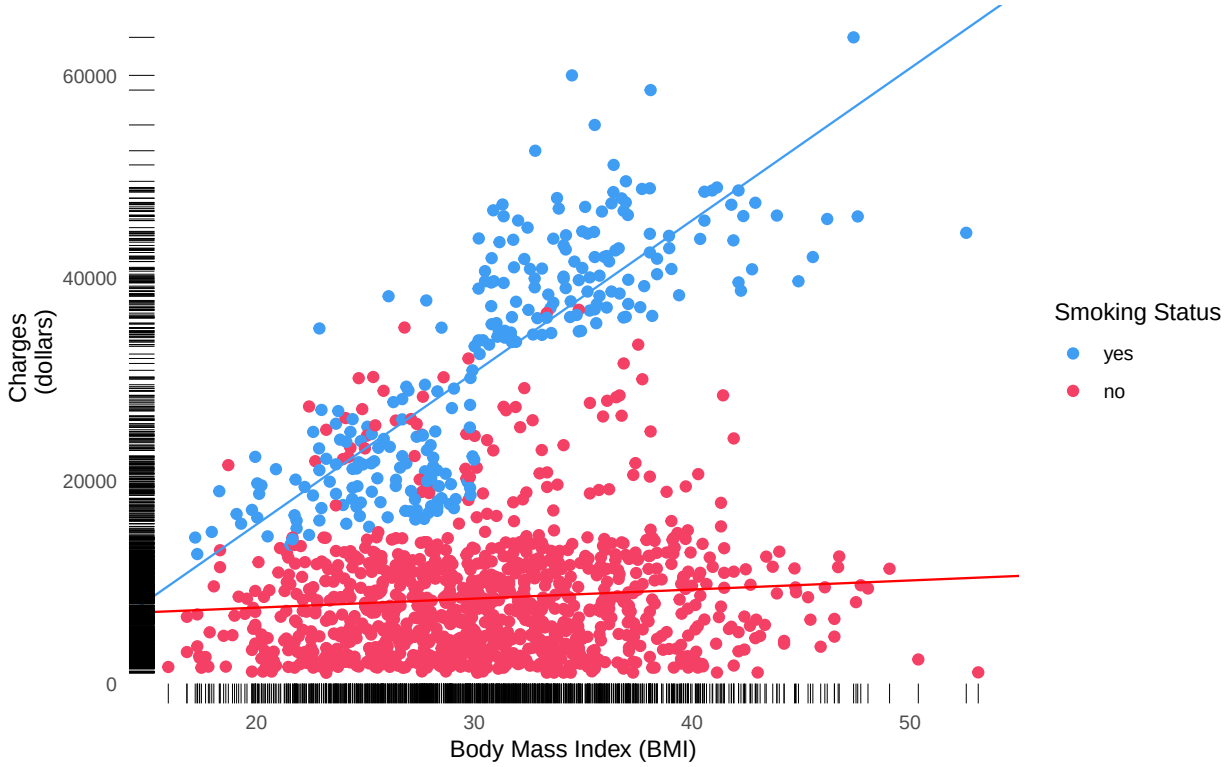


Figure 5: Model 3, plot: with interaction

Brief summary about the models

The MSE provides an absolute measure of lack of fit of model to the data, the R Squared (R^2) statistic provides an alternative measure of fit, which a proportion of variance. The first simple model has (R^2) =

0.04. The second model where smoking status was added has (R^2) = 0.65. For the multiplicative interaction model, between BMI and smoking has (R^2) = 0.74. The (R^2) across each model, the third model explaining the relationship between BMI and charges more accurately. But caution should be taken while interpreting (R^2), as it increases as more variables are added. Therefore, we should also look into whether the p-value of all the variables are significant for each model. However, for our particular question all the p-values of the variables are significant across each model. The p-value for the interaction term, BMI×smoker, is extremely low, indicating that there is strong evidence for rejecting the null hypothesis. This also suggests the true relation might not be additive. 0.09 % of the variability in charges that remained after fitting the second (additive) model can be explained by this interaction. The supplementary plots provide multiple figures for comparing the three models.

Model 4: Adding a variable smokers who are above BMI 30.

$$charges = bmi * (smoker + smokerBMI_{30p}) + e$$

The summary of the model after after adding a variable for those who are smokers and have BMI above 30, the summary of the model is shown in the following table.

Table 5: Summary of model for smokers and BMI above 30

Variable	Estimate	Standard Error	t-statistcis	P-value	95% CI Lower	95% CI Higher
(Intercept)	5846.08	923.73	6.33	0.00	4033.91	7658.25
bmi	85.35	29.58	2.89	0.00	27.33	143.37
smokeryes	1264.79	4325.76	0.29	0.77	-7221.43	9751.00
smoker_bmi30pTRUE	13861.22	5964.13	2.32	0.02	2160.88	25561.55
bmi:smokeryes	469.27	167.86	2.80	0.01	139.97	798.57
bmi:smoker_bmi30pTRUE	21.38	202.63	0.11	0.92	-376.14	418.89

After fitting this model, for the following predictors, bmi, smoker_bmi30p, and for the interaction term between smoker and bmi, we found a strong evidence of linear association (p-value < 0.05). On the other hand, we fail to reject the null hypothesis, which specifies, there are no linear relationship between smoker and the interaction term of smoker and BMI above 30 (bmi * smoker_bmi30p) and outcome variable charges. Here having a p-value > 0.05 for the interaction term, means that the effect of BMI on the the amount of charges in dollars is not different among groups above 30 with different smoking habit. *Keep in mind the hierarchical principle, which dictates that when incorporating an interaction term in a model, the corresponding main effects must also be included, regardless of their coefficient p-values.* Accordingly, the variables that can be dropped is the interaction term between smoker and BMI above 30 (bmi * smoker_bmi30p). Before dropping this interaction term the mean squared error is calculated to be 3.3428165×10^7.

Figure 6 shows the plotting of the data separated and fitted across different groups (above or below 30 BMI and smoking status).

After dropping the interaction term between smoker and bmi above 30 (bmi * smoker_bmi30p) the new linear fit look like:

$$charges = \beta_0 + \beta_1 * bmi + \beta_2 * smoker + \beta_3 * smokerBMI_{30p} + \beta_4 * bmi * smoker + e$$

The summary of the new model is given in table 6.

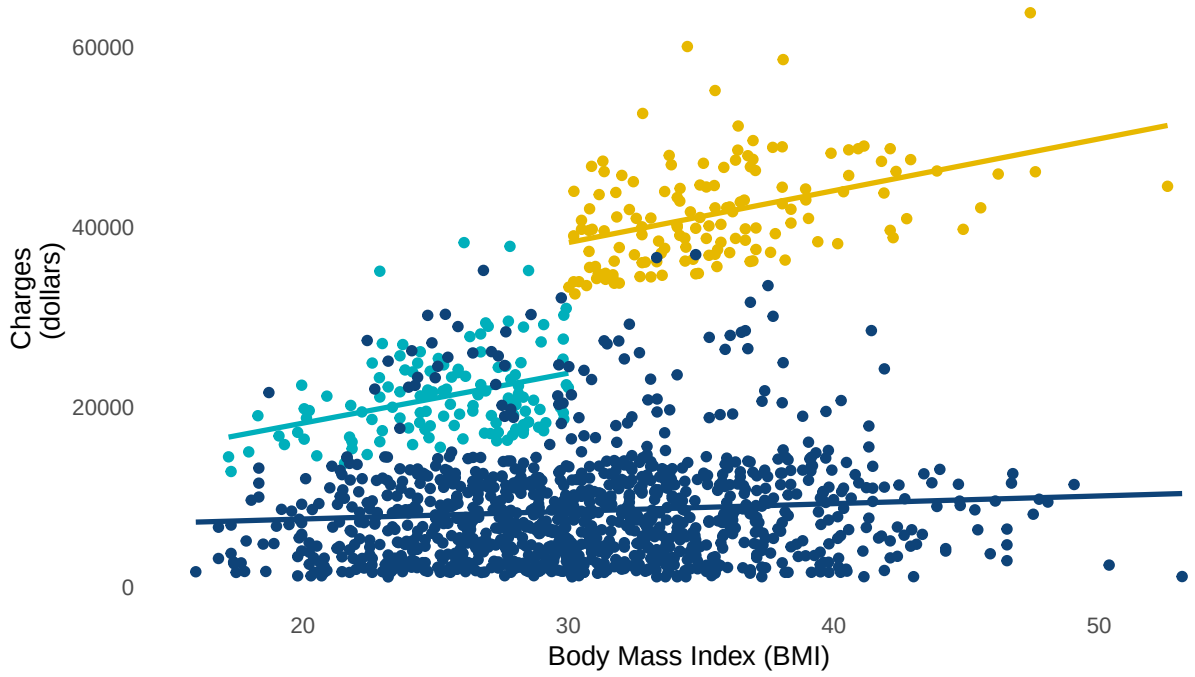


Figure 6: Plotting the data by dividing it into multiple groups

Table 6: Summary of model for smokers and BMI above 30: After filtering for non significant indicators

Variable	Estimate	Standard Error	t-statistcis	P-value	95% CI Lower	95% CI Higher
(Intercept)	5846.08	923.38	6.33	0.00	4034.60	7657.56
bmi	85.35	29.56	2.89	0.00	27.35	143.35
smokeryes	903.90	2646.91	0.34	0.73	-4288.76	6096.56
smoker_bmi30pTRUE	14477.38	1208.68	11.98	0.00	12106.22	16848.55
bmi:smokeryes	483.49	100.07	4.83	0.00	287.16	679.81

According to this new model, a smoker with a BMI of 31.5 would be charged 3.914563×10^4 dollars, where as a smoker with a BMI of of 29 would be charged 2.324616×10^4. A non-smoker with a BMI of 31.5 would be charged 8534.55, where as, a non smoker with a BMI of 29 would be charged 8321.18 dollars. Reducing a BMI of smokers to 29 is associated with a change in mean estimated cost of 1.589947×10^4 dollars. Where as, for a non- smokers this reduction would result in 213.37 dollars. The implication of dropping the interaction term would be, the charges for smokers with a BMI higher than 30, will be irrespective of their BMI. After dropping the interaction term the mean squared error was calculated to be 3.342845×10^7.

Question 2: Figure 7 shows the bias-variance trade off plot. I used R to generate the graph. Figure 8 shows training and test error with model flexibility. The label of underfitting and overfitting are relative to the optimal model. Any model with increased complexity and a higher test MSE would be considered overfitting. Conversely, a model with decreased complexity and a higher Test MSE would be considered underfitting.

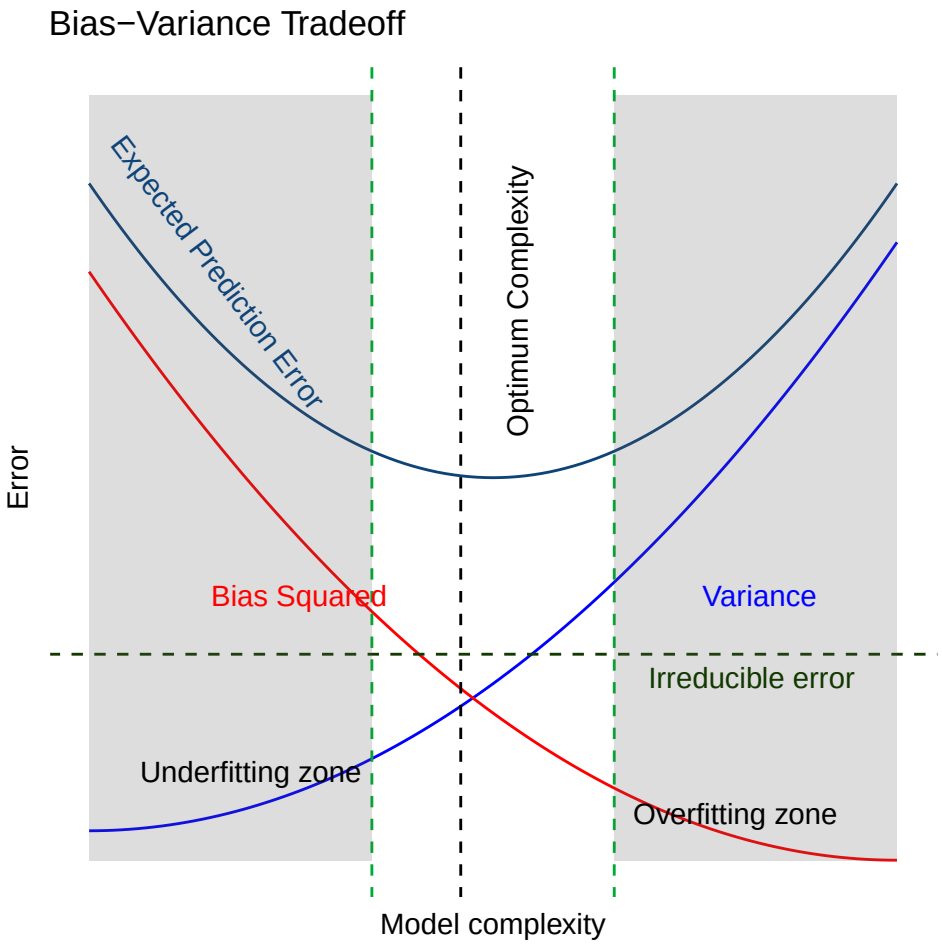


Figure 7: Generating variance trade-off graph

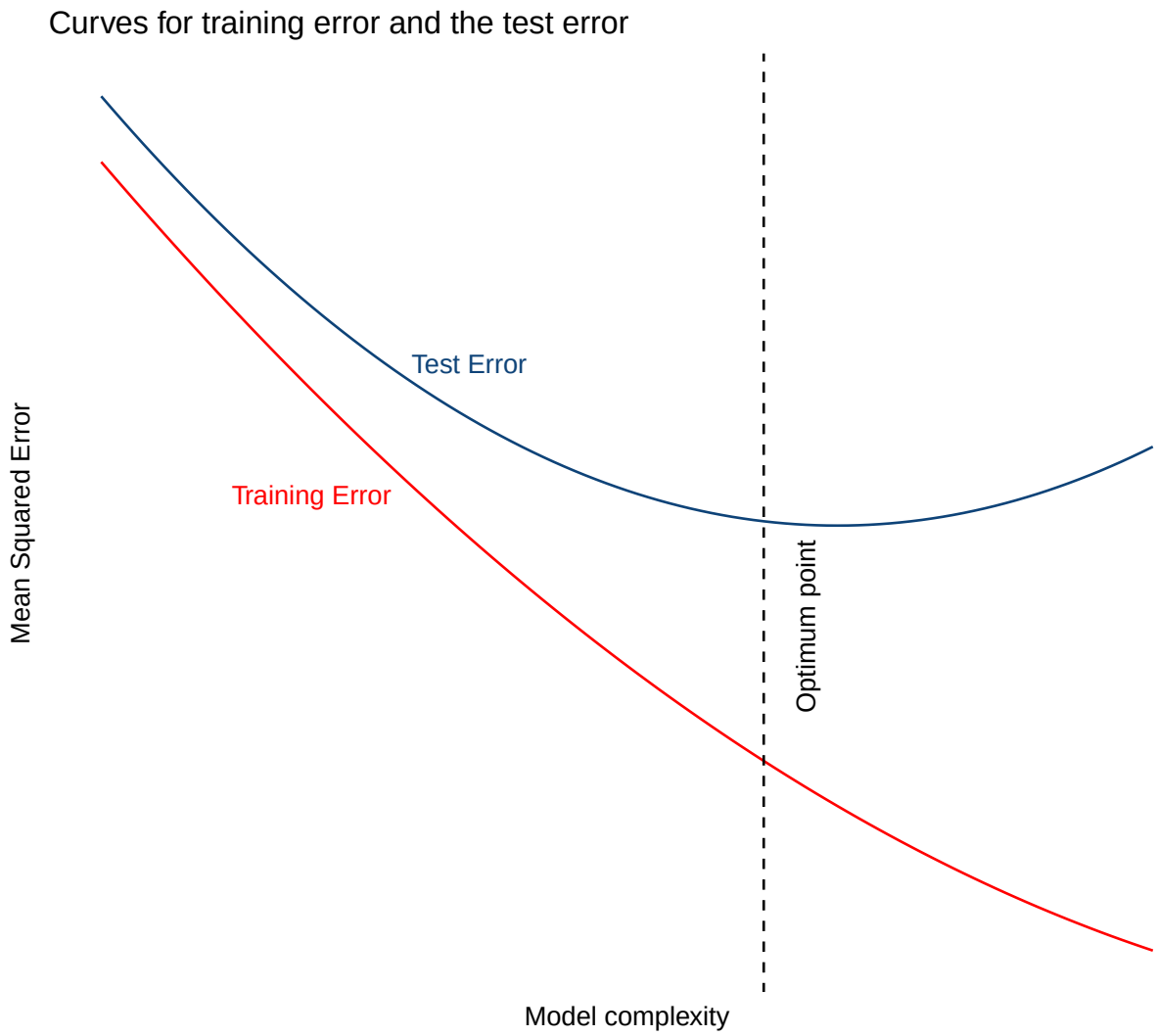


Figure 8: Training error and test error plot

Question 3

Table 8 to 11 of the supplementary results shows the summary of all the fitted models. The given models are listed below:

$$\mathrm{f}(X) = \beta_0 + \beta_1 * X$$

$$\mathrm{f}(X) = \beta_0 + \beta_1 * X + \beta_2 * X^2$$

$$\mathrm{f}(X) = \beta_0 + \beta_1 * X + \beta_2 * X^2 + \beta_3 * X^3 + \beta_4 * X^4$$

$$\mathrm{f}(X) = \beta_0 + \beta_1 * X + \beta_2 * X^2 + \beta_3 * X^3$$

The mean squared error of the training set are 26.58,13.2, 0.06, and 0.06, respectively, for the above functions. As we progress (except the 3rd and 4th), we will be fitting models that are increasingly flexible, incorporating more polynomial terms. As a result, the training error will decrease with each iteration of increased model flexibility. The third order polynomial and the fourth order has almost similar mean square error. Models that are more complex (higher power, in this case) tend to have lower bias but higher variance, while models that are simpler often have higher bias but lower variance. For the given fits the mean squared error of the test set are 49.01, 78.88, 0.12, and 0.1, respectively. The MSEs for the test dataset have increased (slightly, in fit 3 and 4 - but scale still matters). Comparing the (R^2) of the models are, 0.72, 0.86, 1, and 1, respectively. As the model complexity increases, the value of (R^2) also increases. However, this should not be an implication of model's performance. For the first model the p-value for X variable was estimated to be significant ($\mathrm{p} < 0.05$). For the second model, the first order and the second order terms were also found to be significant ($\mathrm{p} < 0.05$). However, for the third model, the second and fourth order variables were found to be insignificant, with a p-values of 0.43 and 0.64. Essentially, the inclusion of this variables and their orders provide no impact in the model fit to the training data, and there consideration may lead to issues of overfitting.

The training and test MSEs of the true regression function $\mathrm{f}^{true}(X) = 3 + 2 * X + 3 * X^3$ are 0.34 and 0.33. On both training and test datasets the value of the MSEs are almost equal for true function.

BONUS

Model Complexity	Mean Squared Error	Bias Squared	Variance
Model 1	9.71293	8.09280	1.62014
Model 2	8.26813	5.37530	2.89282
Model 3	0.23855	0.23473	0.00381
Model 4	0.24278	0.23457	0.00821

What are our expectations? As complexity of the models increase, bias decreases and variance increases. The mean squared error, which is a function of the bias and variance, has to be selected so that a balance between bias and variance is maintained. In this example I ordered the models depending on their complexity (from least order to higher order - so model three is now model four and vice-versa). As expected the bias squared decreases when complexity of the model increases. From model one to model two the variance increased. However, it decreased from model two to three but increased again on model four. Model 1 and 2 are biased as they assume the wrong order of the regression fit. Model 4 (the fourth order fit) has the lowest MSE and squared bias, however model 3 (the third order fit) has the smallest variance and it will outperform model 4 on unobserved data. The mean squared in this test (40 generated dataset - each 30 observations) for model 1 and 2 decreased as compared to the test MSE of 10000 observations (3e). However, for model 3 and 4 the MSE increased.

Supplementary results - Tables

Table 8: Summary of fit for the first order

Variable	Estimate	Standard Error	t-statistcis	P-value	95% CI Lower	95% CI Higher
(Intercept)	4.486082	0.9746096	4.602953	8.21e-05	2.489685	6.48248
X	9.125529	1.0840532	8.417972	0.00e+00	6.904947	11.34611

Table 9: Summary of fit for the second order

Variable	Estimate	Standard Error	t-statistcis	P-value	95% CI Lower	95% CI Higher
(Intercept)	1.751871	0.8733328	2.005960	0.0549750	-0.0400601	3.543802
X	7.617166	0.8298337	9.179147	0.0000000	5.9144875	9.319844
I(X^2)	3.423729	0.6546784	5.229634	0.0000164	2.0804397	4.767018

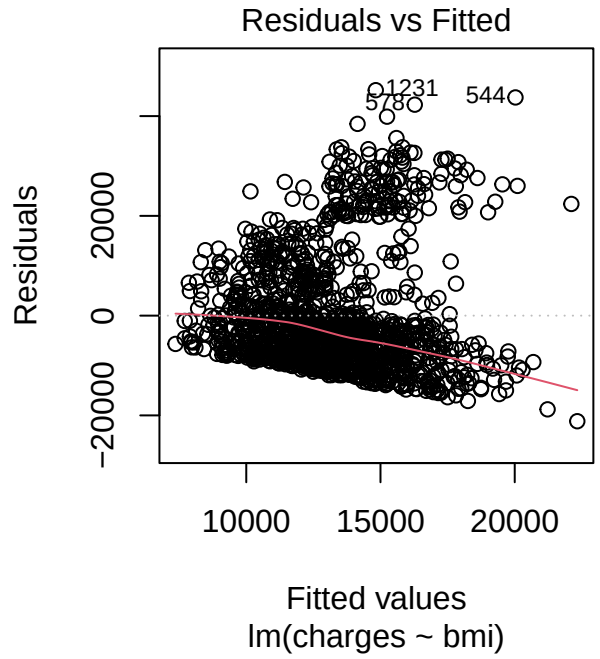
Table 10: Summary of fit for the fourth order

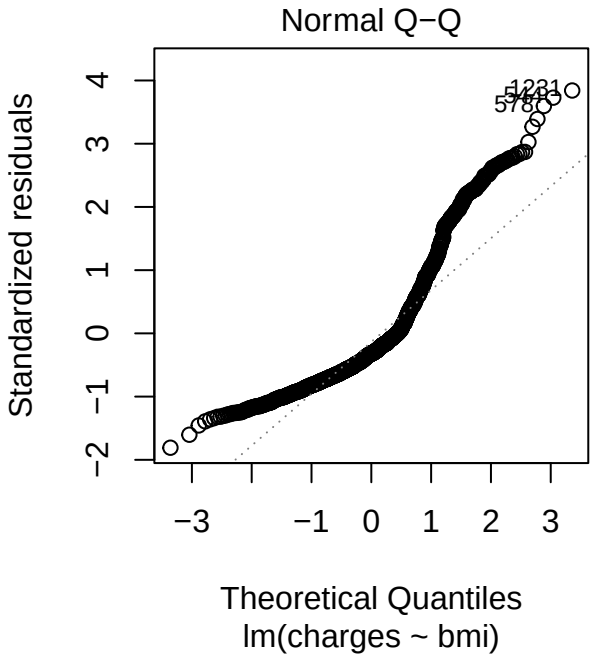
Variable	Estimate	Standard Error	t-statistcis	P-value	95% CI Lower	95% CI Higher
(Intercept)	3.5808075	0.0705248	50.7737416	0.0000000	3.4355589	3.7260560
X	1.9915852	0.1285715	15.4901002	0.0000000	1.7267873	2.2563831
I(X^2)	-0.1037411	0.1286861	-0.8061563	0.4277551	-0.3687752	0.1612929
I(X^3)	2.9591530	0.0805285	36.7466439	0.0000000	2.7933014	3.1250046
I(X^4)	0.0205549	0.0429045	0.4790842	0.6360427	-0.0678086	0.1089183

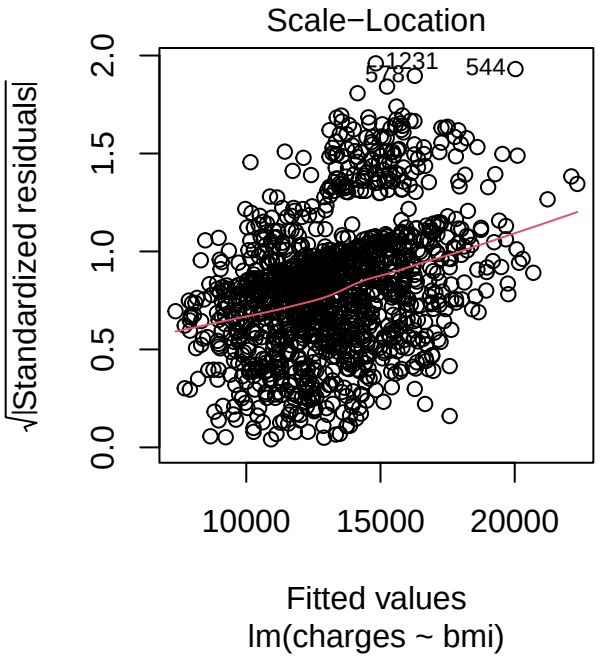
Table 11: Summary of fit for the third order

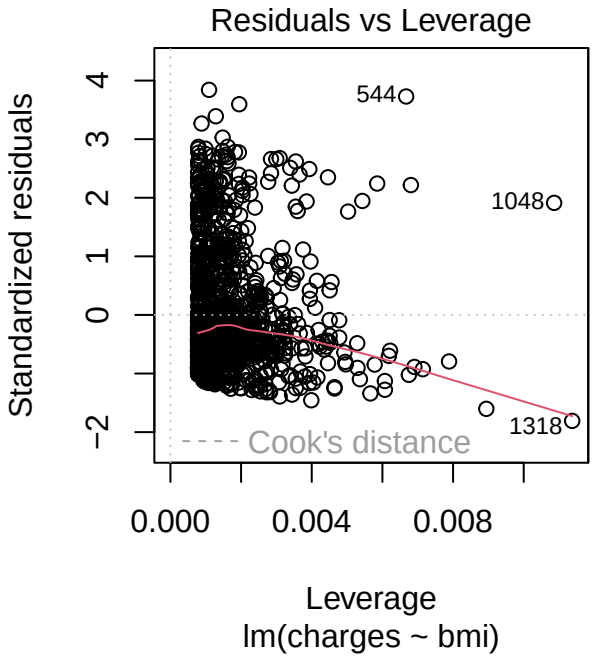
Variable	Estimate	Standard Error	t-statistcis	P-value	95% CI Lower	95% CI Higher
(Intercept)	3.532911	0.0467106	75.63403	0	3.437069	3.628754
X	1.979533	0.0816661	24.23935	0	1.811968	2.147098
I(X^3)	2.970969	0.0264887	112.15999	0	2.916619	3.025320

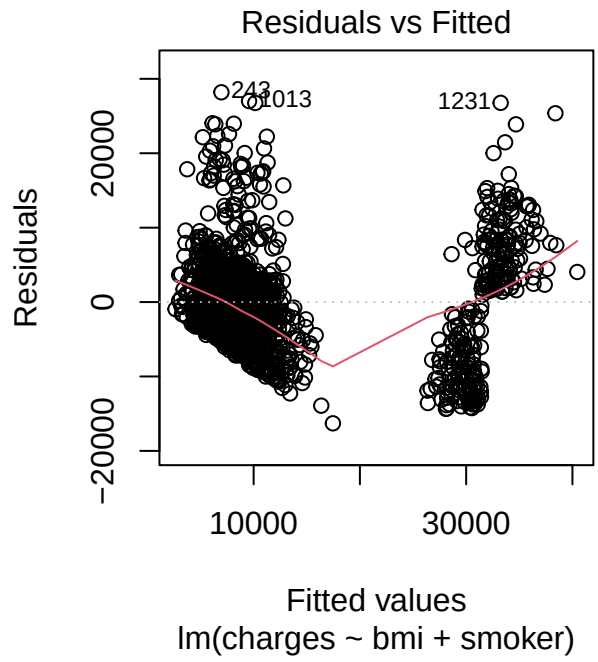
Supplementary results - Plots for question 1 models

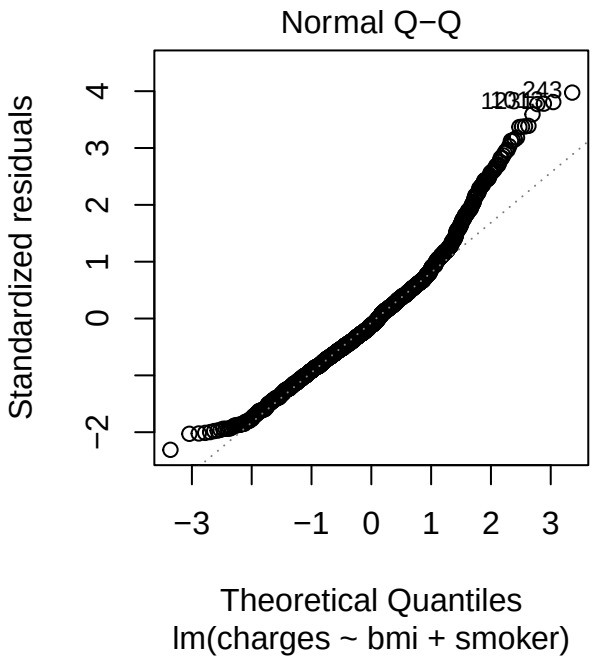


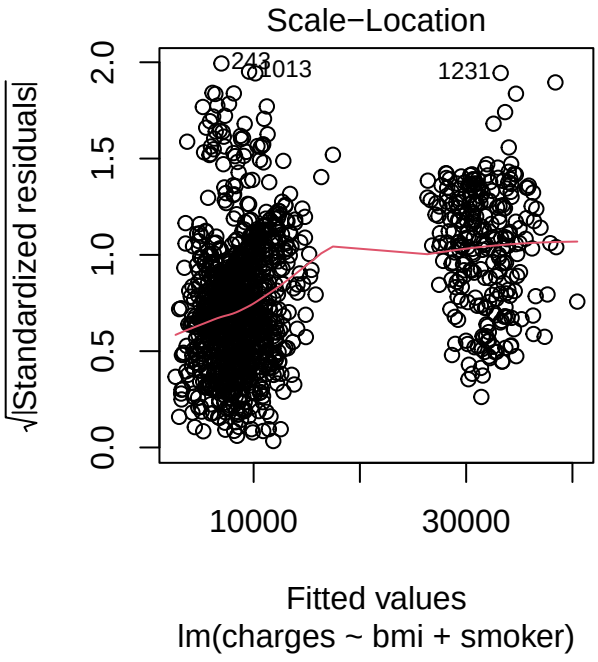


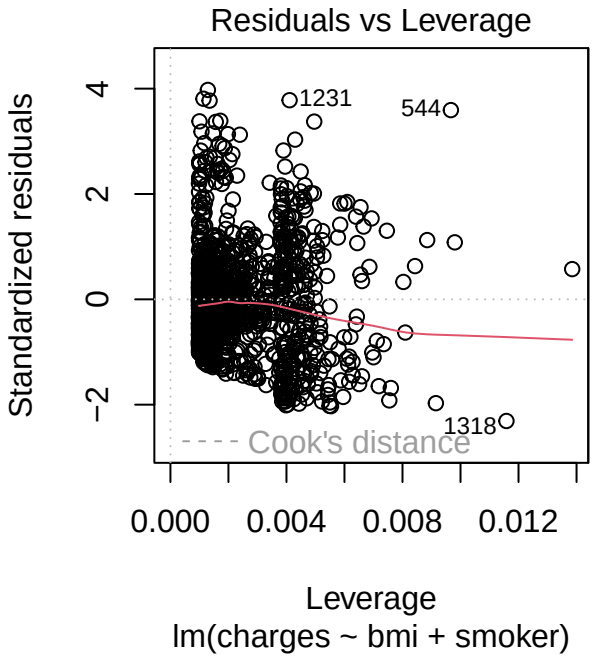


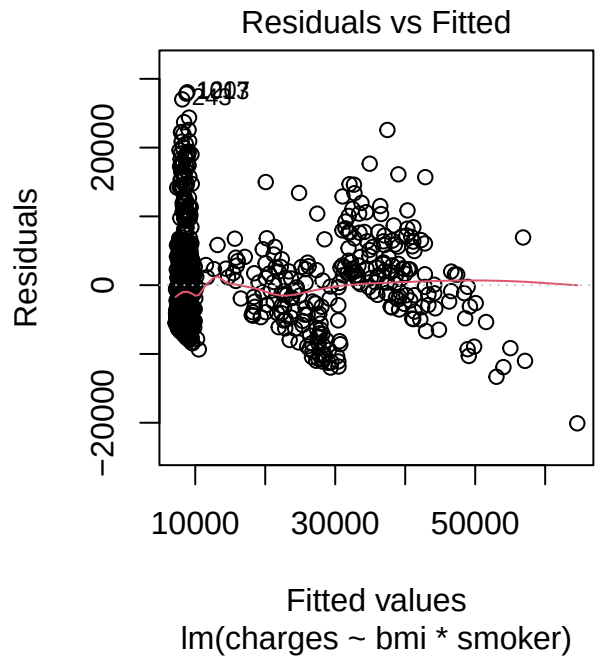


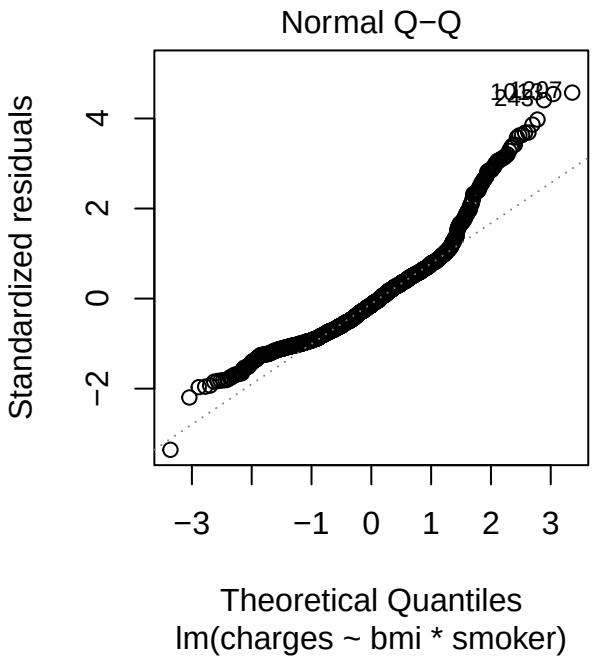

Residuals vs Fitted
Residuals
Fitted values
lm(charges ~ bmi + smoker)

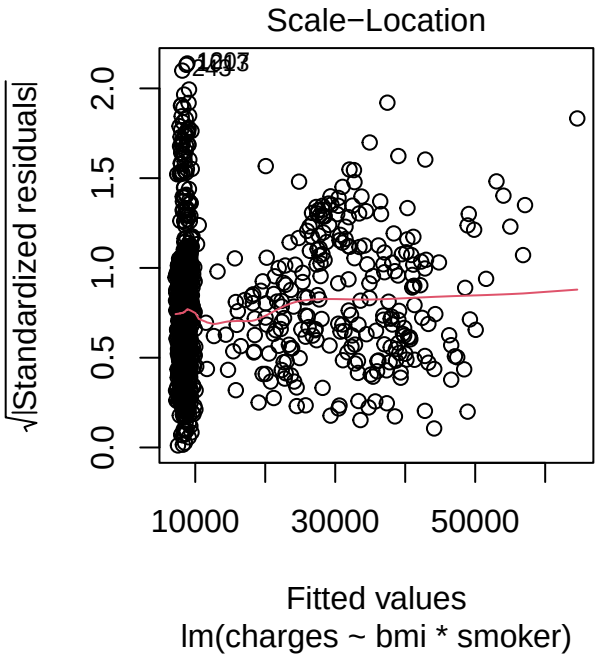

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(charges ~ bmi + smoker)

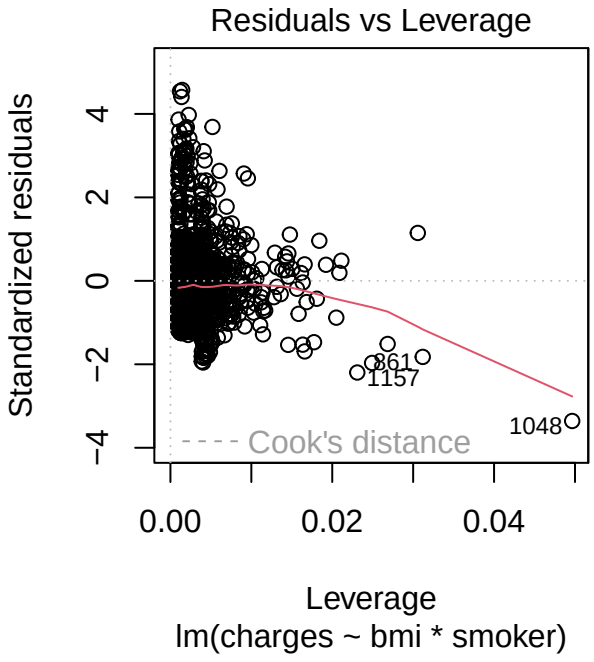

Scale−Location
Fitted values
lm(charges ~ bmi + smoker)


Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
lm(charges ~ bmi + smoker)


Residuals vs Fitted
Residuals
Fitted values
lm(charges ~ bmi * smoker)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(charges ~ bmi * smoker)


Scale−Location
√|Standardized residuals|
Fitted values
lm(charges ~ bmi * smoker)


Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
lm(charges ~ bmi * smoker)

Reference

1. Chapter 8 Bias–Variance Tradeoff: https://daviddalpiaz.github.io/r4sl/biasvariance-tradeoff.html

Code Appendix

```
knitr::opts_chunk$set(fig.pos = 'H')
### Setting up the packages
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
# check if packages are installed; if not, install them
packages <- c("tidyverse", "readr", "ggExtra", "plotly",
              "ggplot2","ggstatsplot","ggside","rigr","nlme","lmtest",
              "sandwich","gridExtra","broom")
not_installed <- setdiff(packages, rownames(installed.packages()))
if (length(not_installed)) install.packages(not_installed)

# load packages
library(sandwich)
library(readr)
library(lmtest)
library(nlme)
library(broom)
library(ggstatsplot)
library(ggside)
library(rigr)
library(ggExtra)
library(gridExtra)
library(plotly)
library(ggplot2)
library(tidyverse)

### ------------------------------------------------------------
#Loading working directory of the raw data

#Please load your data/directory by changing it with your work directory
#Throughout this code module you will see a tone of places, where
#data is read and written, so please make sure to change them to your
#working directory folder format

working_directory_data <- setwd("C:/Users/latera/Desktop/ML_ass")

#loads the data on a variable df
load("data/Medical_Cost_2.RData")
study_data <- df

missing.values <- study_data %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing))
#study_data_clean <- study_data[complete.cases(study_data), ]
study_data_clean <- na.omit(study_data)
```

```r
# fitting: linear regression with charges as the response variable


lm_first <- regress("mean", charges ~ bmi,
                    data = study_data_clean)
coef(lm_first)[,c('Estimate','Naive SE',
                  'Robust SE','95%L',
                   '95%H','Pr(>|t|)')]
as.data.frame(coef(lm_first)[,c('Estimate',
                                'Naive SE',
                   'Robust SE','95%L','95%H','Pr(>|t|)')])

coef(lm_first)


# Predictions for new observations
lm_first_predict <- lm(charges ~ bmi, data = study_data_clean)
predict(lm_first_predict, data.frame(bmi = c(29)))


#Robust fit
#Generate summary statistics for Robut SE
#tests<-coeftest(lm(charges ~ bmi,
                    #data = study_data_clean),
#                vcov=vcovHC(lm(charges ~ bmi,
                #data = study_data_clean),
#                            type = "HC0"))


# fitting: linear regression with charges as the response variable


lm_second <- regress("mean", charges ~ bmi+smoker,
                    data = study_data_clean)
coef(lm_second)[,c('Estimate','Naive SE',
                  'Robust SE','95%L',
                   '95%H','Pr(>|t|)')]
as.data.frame(coef(lm_second)[,c('Estimate',
                                'Naive SE',
                   'Robust SE','95%L','95%H','Pr(>|t|)')])

coef(lm_second)

# Predictions for new observations
lm_second_predict <- lm(charges ~ bmi+smoker, data = study_data_clean)
predict(lm_second_predict, data.frame(bmi = c(29), smoker="yes"))
predict(lm_second_predict, data.frame(bmi = c(31.5), smoker="yes"))
predict(lm_second_predict, data.frame(bmi = c(29), smoker="no"))
predict(lm_second_predict, data.frame(bmi = c(31.5), smoker="no"))
# fitting: linear regression with charges as the response variable


lm_third <- regress("mean", charges ~ bmi*smoker,
```

```
                    data = study_data_clean)
coef(lm_third)[,c('Estimate','Naive SE',
                  'Robust SE','95%L',
                   '95%H','Pr(>|t|)')]
as.data.frame(coef(lm_third)[,c('Estimate',
                                'Naive SE',
                   'Robust SE','95%L','95%H','Pr(>|t|)')])

coef(lm_third)

lm_third_predict <- lm(charges ~ bmi*smoker, data = study_data_clean)
predict(lm_third_predict, data.frame(bmi = c(29), smoker="yes"))
predict(lm_third_predict, data.frame(bmi = c(31.5), smoker="yes"))
predict(lm_third_predict, data.frame(bmi = c(29), smoker="no"))
predict(lm_third_predict, data.frame(bmi = c(31.5), smoker="no"))


#Generating tables


kable(missing.values, col.names =
      c('Variables', 'Number of missing observations'),
      caption = "Missing Patterns")


row.plot <- study_data %>%
  mutate(id = row_number()) %>%
  gather(-id, key = "key", value = "val") %>%
  mutate(isna = is.na(val))

ggplot(row.plot,aes(key, id, fill = isna)) +
    geom_raster(alpha=0.8) +
    scale_fill_manual(name = "",
        values = c('#010101', '#fea000'),
        labels = c("Non-missing", "Missing")) +
    labs(x = "All variables",
           y = "Row Number") +
  theme_bw() +
  theme(axis.line = element_line(colour = "white"),
        axis.ticks = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank())+
    coord_flip()

#Plotting scatter plot between weight and height (inches)

p <- ggplot(study_data_clean, aes(x=bmi, y=charges, color = smoker)) +
      geom_point(size=1.2)+
      xlab("Body Mass Index (BMI)") + ylab("Charges \n (dollars)")+
  scale_color_manual(name="Smoking Status",breaks=c('yes', 'no'),
                     values=c('yes'='#409df4', 'no'='#f54066'))+
```

```
  geom_rug(col="black",linewidth=0.20)+
  theme_bw() +
  theme(axis.line = element_line(colour = "white"),
        axis.ticks = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank())+
  theme(legend.background = element_rect
        (fill = "transparent"))

p+geom_smooth(method='lm',se=TRUE,level=0.95,
              linewidth = 0.35,formula = y~x)

#Generating tables
kable(coef(lm_first)%>%round(4), caption = "OLS - Linear fit of charges
      and body mass index (BMI)")


ggplot(study_data_clean, aes(x=bmi, y=charges, color = smoker))+
  geom_point(size=2)+
  geom_abline(slope = lm_first_predict$coefficients[2],
              intercept = lm_first_predict$coefficients[1], col = 'red')+
   xlab("Body Mass Index (BMI)") + ylab("Charges \n (dollars)")+
  scale_color_manual(name="Smoking Status",breaks=c('yes', 'no'),
                     values=c('yes'='#409df4', 'no'='#f54066'))+
  geom_rug(col="black",linewidth=0.20)+
  theme_bw() +
  theme(axis.line = element_line(colour = "white"),
        axis.ticks = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank())+
  theme(legend.background = element_rect
        (fill = "transparent"))

kable(coef(lm_second)%>%round(4), caption = "OLS - Linear fit of charges
      with body mass index (BMI) and smoking status")


ggplot(study_data_clean, aes(x=bmi, y=charges, color = smoker))+
  geom_point(size=2)+
  geom_abline(slope = lm_second_predict$coefficients[2],
              intercept = lm_second_predict$coefficients[1],
              col = 'red')+
  geom_abline(slope = lm_second_predict$coefficients[2],
              intercept = (lm_second_predict$coefficients[1]+
                             lm_second_predict$coefficients[3]),
              col = '#409df4')+
   xlab("Body Mass Index (BMI)") + ylab("Charges \n (dollars)")+
  scale_color_manual(name="Smoking Status",breaks=c('yes', 'no'),
                     values=c('yes'='#409df4', 'no'='#f54066'))+
```

```
  geom_rug(col="black",linewidth=0.20)+
  theme_bw() +
  theme(axis.line = element_line(colour = "white"),
        axis.ticks = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank())+
  theme(legend.background = element_rect
        (fill = "transparent"))

kable(coef(lm_third)%>%round(4), caption = "OLS - Linear fit of charges
      with body mass index (BMI) and smoking status",digits = 2)


ggplot(study_data_clean, aes(x=bmi, y=charges, color = smoker))+
  geom_point(size=2)+
  geom_abline(slope = lm_third_predict$coefficients[2],
              intercept = lm_third_predict$coefficients[1], col = 'red')+
  geom_abline(slope = lm_third_predict$coefficients[2] +
                lm_third_predict$coefficients[4],
              intercept = lm_third_predict$coefficients[1] +
                lm_third_predict$coefficients[3], col = '#409df4')+
   xlab("Body Mass Index (BMI)") + ylab("Charges \n (dollars)")+
  scale_color_manual(name="Smoking Status",breaks=c('yes', 'no'),
                     values=c('yes'='#409df4', 'no'='#f54066'))+
  geom_rug(col="black",linewidth=0.20)+
  theme_bw() +
  theme(axis.line = element_line(colour = "white"),
        axis.ticks = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank())+
  theme(legend.background = element_rect
        (fill = "transparent"))
new_data <- df %>% mutate(smoker_bmi30p = smoker == "yes" & bmi > 30)
lm_fourth <- lm(charges ~ bmi* (smoker + smoker_bmi30p), data = new_data)


kable(tidy(lm_fourth, conf.int = T),
      col.names = c('Variable',"Estimate","Standard Error",
                    "t-statistcis", "P-value","95% CI Lower","95% CI Higher"),
      caption = "Summary of model for smokers and BMI above 30",digits = 2)


above_30_smoker <- new_data %>%
  filter(smoker_bmi30p == "TRUE")
below_30_smoker <- new_data %>%
  filter(smoker_bmi30p == "FALSE" & smoker == "yes")
none_smoker <- new_data %>%
  filter(smoker == "no")
##############################################
```

```
ggplot() +
  geom_point(data = above_30_smoker,aes(x = bmi, y = charges),
             color = '#E7B800') +
  geom_smooth(data = above_30_smoker,aes(x = bmi, y = charges),
              color = '#E7B800',method = lm, se = FALSE) +
  geom_point(data = below_30_smoker,aes(x = bmi, y = charges),
             color = "#00AFBB") +
  geom_smooth(data = below_30_smoker,aes(x = bmi, y = charges),
              color = "#00AFBB", method = lm, se = FALSE) +
  geom_point(data = none_smoker,aes(x = bmi, y = charges),
             color = "#0e4378") +
  geom_smooth(data = none_smoker,aes(x = bmi, y = charges),
              color = "#0e4378",method = lm, se = FALSE)+
  xlab("Body Mass Index (BMI)") + ylab("Charges \n (dollars)")+
  scale_color_manual(name="Smoking Status",breaks=c('yes', 'no'),
                     values=c('yes'='#409df4', 'no'='#f54066'))+
  geom_rug(col="black",linewidth=0.20)+
  theme_bw() +
  theme(axis.line = element_line(colour = "white"),
        axis.ticks = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank())+
  theme(legend.background = element_rect
        (fill = "transparent"))

lm_fifth <- lm(charges ~ bmi + smoker + smoker_bmi30p + bmi*smoker, data = new_data)

predict(lm_fifth, data.frame(bmi = c(29), smoker=c("yes"),
                             smoker_bmi30p=c(F)))
predict(lm_fifth, data.frame(bmi = c(31.5), smoker=("yes"),
                             smoker_bmi30p=c(T)))
predict(lm_fifth, data.frame(bmi = c(29), smoker=("no"),
                             smoker_bmi30p=c(F)))
predict(lm_fifth, data.frame(bmi = c(31.5), smoker=("no"),
                             smoker_bmi30p=c(F)))

kable(tidy(lm_fifth, conf.int = T), col.names = c('Variable',
                                                  "Estimate",
                                                  "Standard Error",
                                                  "t-statistcis",
                                                  "P-value",
                                                  "95% CI Lower",
                                                  "95% CI Higher"),
      caption = "Summary of model for smokers and BMI above 30:
      After filtering for non significant indicators",digits = 2)

x <- seq(0, 1, 0.01)
bias_squared <- 0.05+x^2
variance <- (1 - x)^2
generalization_error <- bias_squared + variance + 0.1
```

```
# Create the plot
ggplot() +
  geom_line(aes(x, bias_squared), color = "blue") +
  geom_line(aes(x, variance), color = "red") +
  geom_line(aes(x, generalization_error), color = "#0e4378") +
  geom_vline(xintercept = 0.46, color = "black", linetype = "dashed") +
  geom_vline(xintercept = 0.65, color = "#00aa34", linetype = "dashed")+
  geom_hline(yintercept = 0.35, color = "#123a00", linetype = "dashed")+
  annotate("rect", xmin = 0.65, xmax = 1, ymin = 0, ymax = 1.3,
           alpha = .2)+
  geom_vline(xintercept = 0.35, color = "#00aa34", linetype = "dashed")+
  annotate("rect", xmin = 0.0, xmax = 0.35, ymin = 0, ymax = 1.3,
           alpha = .2)+
  annotate("text",x = 0.20, y = 0.15,
           label = "Underfitting zone", color = "black") +
  annotate("text",x = 0.80, y = 0.08,
           label = "Overfitting zone", color = "black") +
  xlab("Model complexity") +
  ylab("Error") +
  ggtitle("Bias-Variance Tradeoff") +
  scale_color_manual(name = "Error", values =
                       c("blue" = "variance", "red" = "bias squared",
                                "#0e4378" = "generalization error"))+
  annotate("text", x = 0.83, y = 0.45,
           label = "Variance",color='blue') +
  annotate("text", x = 0.26, y = 0.45,
           label = "Bias Squared",color='red')+
  annotate("text", x = 0.16, y = 1,
           label = "Expected Prediction Error",
           color='#0e4378',angle = -53)+
  annotate("text", x = 0.53, y = 0.95,
           label = "Optimum Complexity",color='black',angle = 90)+
  annotate("text", x = 0.82, y = 0.31,
           label = "Irreducible error",color='#123a00')+

  theme_bw()+
  theme(axis.line = element_line(colour = "white"),
        axis.ticks = element_blank(),
        axis.text = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.border = element_blank(),
        panel.background = element_blank())
library(readr)
x <- seq(0, 1, 0.01)
bias_squared <- 0.05+x^2
variance <- (1.4 - x)^2
generalization_error <- bias_squared + variance + 0.1

# Create the plot
ggplot() +

  geom_line(aes(x, variance), color = "red") +
```

```
  geom_line(aes(x, generalization_error), color = "#0e4378") +
  geom_vline(xintercept = 0.63, color = "black", linetype = "dashed") +

  xlab("Model complexity") +
  ylab("Mean Squared Error") +
  ggtitle("Curves for training error and the test error") +
  scale_color_manual(name = "Error",
                     values = c( "red" = "Training MSE",
                                 "#0e4378" = "generalization error"))+
  annotate("text", x = 0.20, y = 1.2, label = "Training Error",color='red')+
  annotate("text", x = 0.35, y = 1.5, label = "Test Error",color='#0e4378')+
  annotate("text", x = 0.67, y = 0.90,
           label = "Optimum point",color='black',angle = 90)+
  theme_bw()+
  theme(axis.line = element_line(colour = "white"),
        axis.ticks = element_blank(),
        axis.text = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.border = element_blank(),
        panel.background = element_blank())


# Generating training dataset
set.seed(0)
n <- 30
X<-matrix(rnorm(n),n,1)
eps <- runif(n)
Y=3 + X[,1]*2 + (X[,1]^3)*(3) + eps

df_training <- data.frame(X,Y)

fit_1 <- lm(Y ~ X, data = df_training)

fit_2 <- lm(Y ~ X + I(X^2), data = df_training)

fit_3 <- lm(Y ~ X + I(X^2) + I(X^3) + I(X^4), data = df_training)

fit_4 <- lm(Y ~ X + I(X^3), data = df_training)

#Generating test dataset

n <- 10000
X<-matrix(rnorm(n),n,1)
eps <- runif(n)
Y<-3 + X[,1]*2 + (X[,1]^3)*(3) + eps

df_test<-data.frame(X,Y)
#mean((df_test$Y - predict(fit_1, df_test)) ^ 2)
#mean((df_test$Y - predict(fit_2, df_test)) ^ 2)
#mean((df_test$Y - predict(fit_3, df_test)) ^ 2)
#mean((df_test$Y - predict(fit_4, df_test)) ^ 2)
```

```r
#Fitting the true f

#To generate for the true f, lets repeat the process
set.seed(0)
n <- 30
X <- rnorm(n)
eps <- runif(n)
Y <- 3 + 2*X + 3*X^3 + eps

X_test <- rnorm(10000)
eps_test <- runif(10000)
Y_test <- 3 + 2*X_test + 3*X_test^3 + eps_test

# Training MSE
f_true <- 3 + 2*X + 3*X^3
training_mse <- mean((Y - f_true)^2)

# Test MSE
f_true_test <- 3 + 2*X_test + 3*X_test^3
test_mse <- mean((Y_test - f_true_test)^2)


set.seed(0)
data_generate = 40
models = 4
predictions = matrix(0, nrow = data_generate, ncol = models)

generator = function(sample_size = 30) {
  x = rnorm(sample_size)
  eps <- runif(sample_size)
  y = 3 + 2*x + 3*x^3 + eps
  return(data.frame(x, y))
}

for (new_data in 1:data_generate) {

  #generate data
  datasets = generator()

  # fit models
  fit_0_bonus = lm(y ~ x, data = datasets)
  fit_1_bonus = lm(y ~ x + I(x^2), data = datasets)
  fit_2_bonus = lm(y ~ x + I(x^2) + I(x^3) + I(x^4), data = datasets)
  fit_3_bonus = lm(y ~ x + I(x^3), data = datasets)

  # get predictions
  predictions[new_data, 1] = predict(fit_0_bonus, data.frame(x = 0.30))
  predictions[new_data, 2] = predict(fit_1_bonus, data.frame(x = 0.30))
  predictions[new_data, 4] = predict(fit_2_bonus, data.frame(x = 0.30))
  predictions[new_data, 3] = predict(fit_3_bonus, data.frame(x = 0.30))
}

mse_f = function(f_true, estimate) {
```

```
  mean((estimate - f_true) ^ 2)
}

bias_f = function(estimate, f_true) {
  mean(estimate) - f_true
}

var_f = function(estimate) {
  mean((estimate - mean(estimate)) ^ 2)
}

bias = apply(predictions, 2, bias_f, f_true = 3 + 2*0.3 + 3*(0.3^3))
variance = apply(predictions, 2, var_f)
mse = apply(predictions, 2, mse_f, f_true = 3 + 2*0.3 + 3*(0.3^3))


specs = data.frame(
  poly_degree = c("Model 1", "Model 2", "Model 3", "Model 4"),
  round(mse, 5),
  round(bias ^ 2, 5),
  round(variance, 5)
)
colnames(specs) = c("Model Complexity", "Mean Squared Error", "Bias Squared", "Variance")
rownames(specs) = NULL
kable(specs)

#Generating tables

kable(tidy(fit_1, conf.int = T), col.names =
        c('Variable','Estimate','Standard Error','t-statistcis',
          'P-value','95% CI Lower','95% CI Higher'),
      caption = "Summary of fit for the first order")

kable(tidy(fit_2, conf.int = T),
      col.names = c('Variable','Estimate','Standard Error','t-statistcis',
                    'P-value','95% CI Lower','95% CI Higher'),
      caption = "Summary of fit for the second order")

kable(tidy(fit_3, conf.int = T),
      col.names = c('Variable','Estimate','Standard Error','t-statistcis',
                    'P-value','95% CI Lower','95% CI Higher'),
      caption = "Summary of fit for the fourth order")

kable(tidy(fit_4, conf.int = T), col.names = c('Variable','Estimate',
                                               'Standard Error',
                                               't-statistcis',
                                               'P-value',
                                               '95% CI Lower',
                                               '95% CI Higher'),
      caption = "Summary of fit for the third order")


#kable((tests_cm)[,], caption = "Linear fit of weight and height
```

```
 #     (centimeters) - RSE")

#checking OLS fit (residuals Vs fitted plots)
plot(lm(charges ~ bmi, data = study_data_clean))


#checking OLS fit (residuals Vs fitted plots)
plot(lm(charges ~ bmi+smoker, data = study_data_clean))


#checking OLS fit (residuals Vs fitted plots)
plot(lm(charges ~ bmi*smoker, data = study_data_clean))
```