BIOST 546 HW 1

My-Anh Doan

2023-01-09

```
# set global options for code chunks
knitr::opts_chunk$set(message = FALSE, warning = FALSE, collapse = TRUE)
knitr::opts_knit$set(root.dir = rprojroot::find_rstudio_root_file())

library(dplyr)
library(knitr)
```

Q1

Q1. In this problem, we will make use of the data set `Medical_Cost_2.RData`.

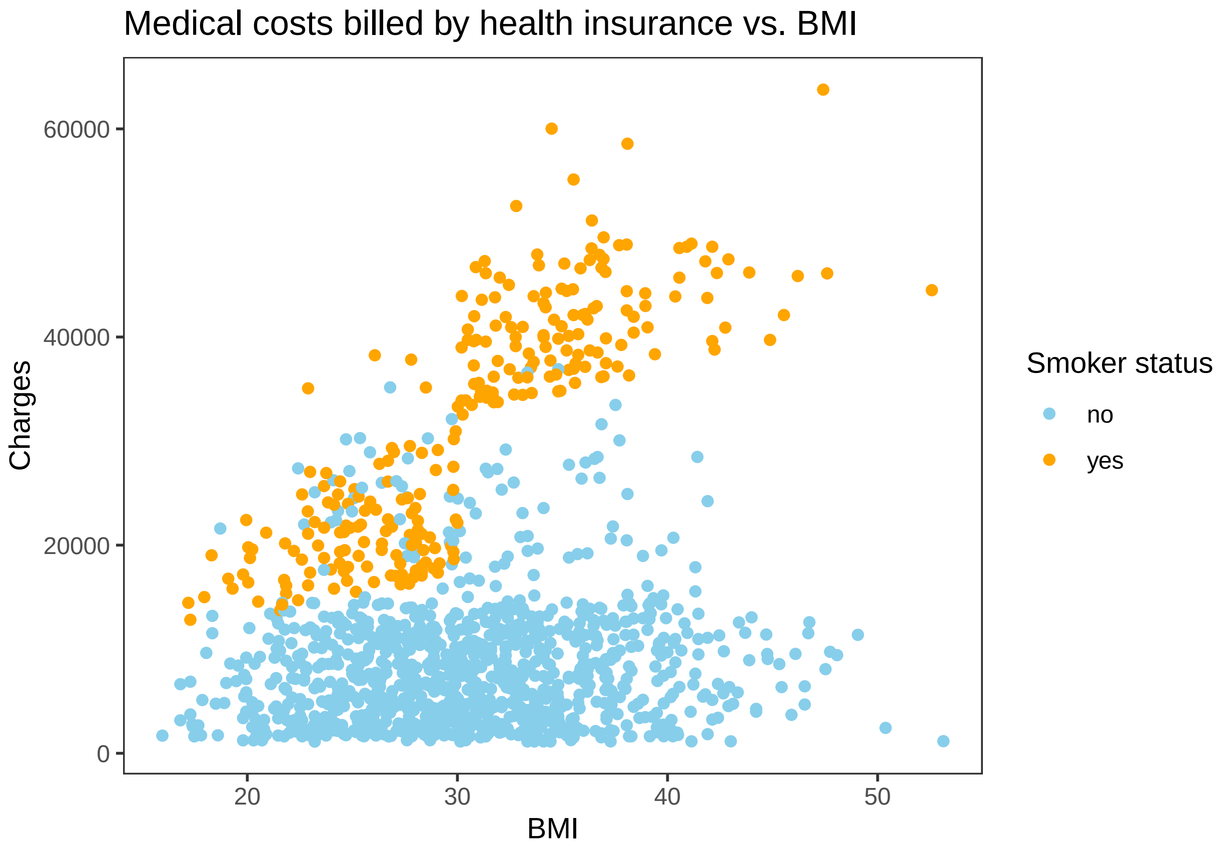
(a) Load the data set with the command `load` and check if there are missing data
(b) If any, remove the missing data using the command `na.omit`.

```
# load data
load("./dataset/Medical_Cost_2.RData")

# check for missing data in `df` data frame
# if returned value is non-zero, there are missing data
sum(is.na(df) == TRUE)
## [1] 60

# remove the missing data from `df`
df <- na.omit(df)
str(df)
## 'data.frame':    1278 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
##  - attr(*, "na.action")= 'omit' Named int [1:60] 22 37 39 40 84 111 129 193 248 252 ...
##   ..- attr(*, "names")= chr [1:60] "22" "37" "39" "40" ...
```

Initial loading of data, we had a data frame with 1338 observations and p = 7 variables. After using the command `na.omit`, the resulting data frame contains 1278 observations and p = 7 variables.

(c) We decide to focus on the outcome variable `charges` (individual medical costs billed by health insurance) and the predictors `bmi` (body mass index) and `smoker` (whether the subjects is a smoker or not). Make a scatter plot with `bmi` on the x-axis, `charges` on the y-axis, and with the color of each dot representing whether the subject is a smoker or not.

```r
library(ggplot2)

q1c_plot <- ggplot(df, aes(x = bmi, y = charges, color = smoker)) +
  geom_point() +
  scale_color_manual(values = c("no" = "sky blue",
                                "yes" = "orange")) +
  labs(title = "Medical costs billed by health insurance vs. BMI",
       x = "BMI",
       y = "Charges",
       color = "Smoker status") +
  theme_bw() +
  theme(panel.grid.minor = element_blank(),
        panel.grid.major = element_blank())
q1c_plot
```



(d) Fit a least-squares linear model, with intercept, in order to predict:

- `charges` using `bmi` as the only predictor;
- `charges` using `bmi` and `smoker` as predictors;

- charges using bmi and smoker as in the previous model; but allowing for an interaction term between the variables bmi and smoker

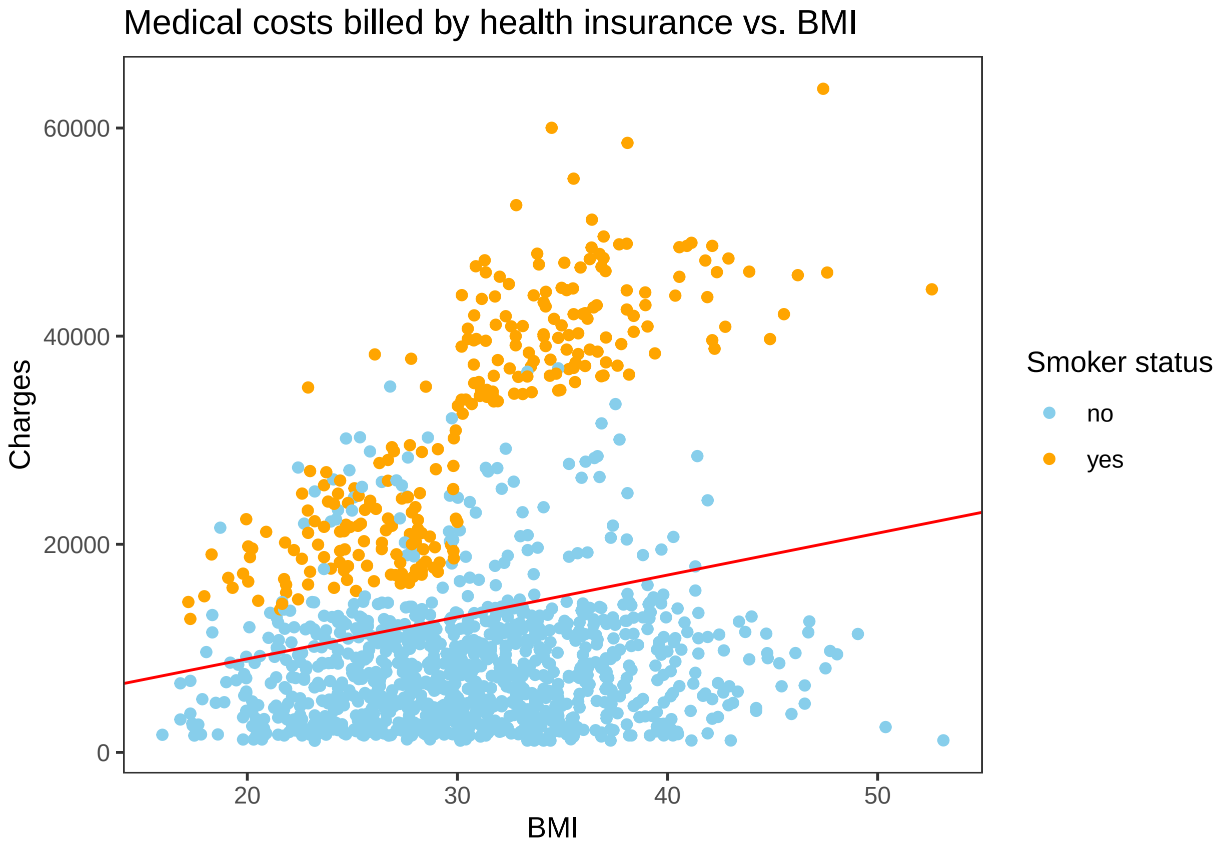
For each of the three models:

- Present your results in the form of a table where you report the estimated regression coefficients and their interpretations (be careful with the dummy variables);
- Report the 95% confidence interval for the coefficient of the variable bmi, and provide a sentence explaining the meaning of this confidence interval;
- Draw the regression line(s) of the model on the scatter plot produced in point (c);
- Report the (training set) mean squared error of the model;
- Predict the medical costs billed by the health insurance company to a smoker with a bmi that is 29 and 31.5;
- Compute the predicted difference in charges between a smoker with bmi 31.5 and one with bmi 29. Do the same for non-smokers. Comment on the results

```r
# functions for answering modeling and prediction questions in Q1

# returns a table containing model coefficient estimates, standard errors, t-values,
# p-values, and upper and lower limits of 95% CIs
results_table <- function(linear_model) {
  coeffs <- cbind(as.data.frame(summary(linear_model)$coefficients),
                  confint(linear_model))
  colnames(coeffs) <- c("Estimate", "Std. Error", "t-value", "p-value",
                        "95% CI LL", "95% CI UL")
  coeffs <- coeffs %>%
    mutate(Estimate = round(Estimate, 2),
           `Std. Error` = round(`Std. Error`, 2),
           `t-value` = round(`t-value`, 2),
           `p-value` = format(`p-value`, scientific = TRUE, digits = 3),
           `95% CI LL` = round(`95% CI LL`, 2),
           `95% CI UL` = round(`95% CI UL`, 2))
  coeffs
}

# calculates the mean squared error of the training model
calc_mse <- function(linear_model) {
  x <- mean(linear_model$residuals^2)
  print(paste("The training set mean squared error of the model is:", round(x, 0)))
}

# calculates the predicted charges for smokers and non-smokers at different BMIs
pred_costs <- function(linear_model, smoker, bmi) {
  round(predict(linear_model, data.frame(smoker, bmi)), 2)
}

# calculates the difference in charges at different BMIs
diff_costs <- function(linear_model, smoker, bmi) {
  round(max(pred_costs(linear_model, smoker, bmi)) -
          min(pred_costs(linear_model, smoker, bmi)), 2)
}
```

Model 1: `charges ~ bmi`

```r
# linear model using `bmi` as the only predictor
charges_lm1 <- lm(charges ~ bmi, data = df)
kable(results_table(charges_lm1))

q1c_plot +
  geom_abline(slope = coef(charges_lm1)[2],
              intercept = coef(charges_lm1)[1],
              col = "red")
```



	Estimate	Std. Error	t-value	p-value	95% CI LL	95% CI UL
(Intercept)	938.44	1682.00	0.56	5.77e-01	-2361.35	4238.23
bmi	402.65	53.85	7.48	1.40e-13	297.01	508.28

The general model for linear regression models with only one predictor is defined as:

$$y = \beta_0 + \beta_1 \times x_1$$

Thus, the model for both smokers and non-smokers is given as:

$$charges = \beta_0 + \beta_1 \times bmi,$$

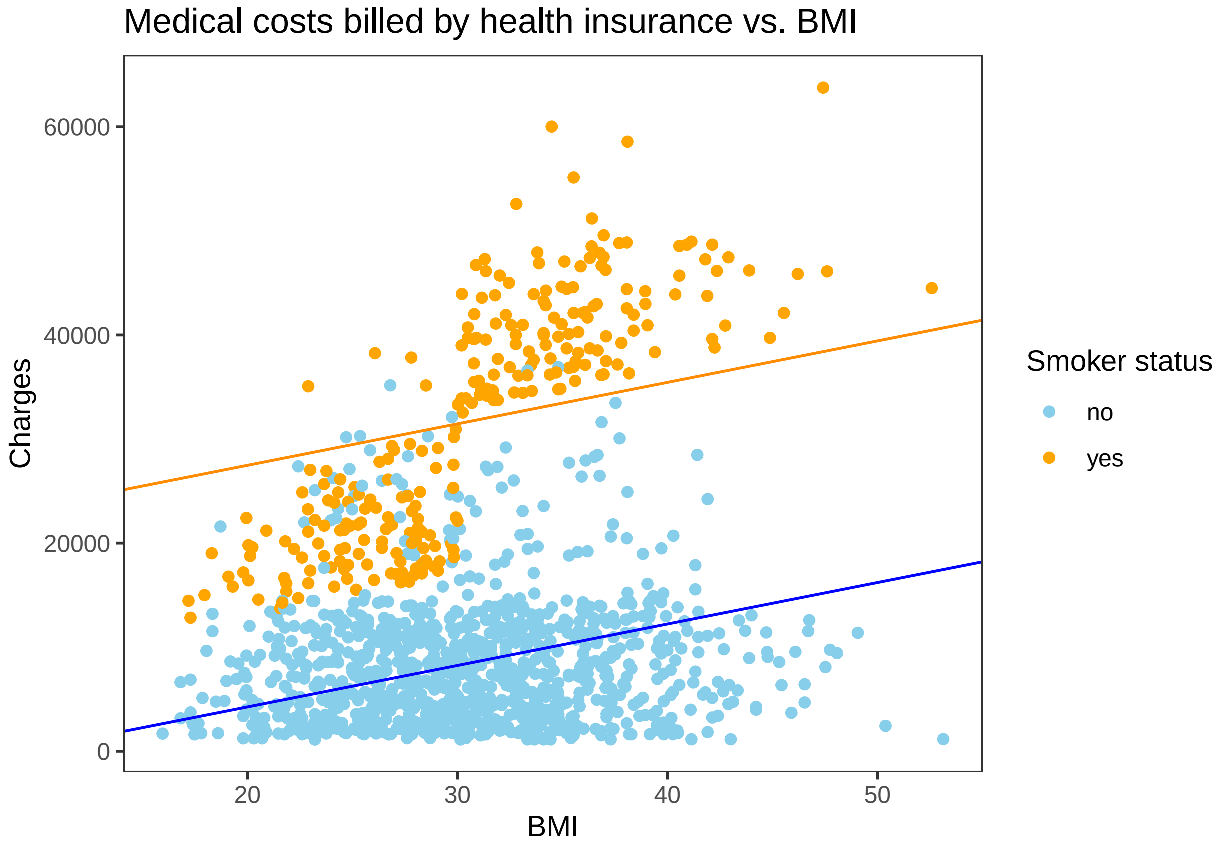
where β_0 (the intercept) is equal to 938.44 and β_1 (the slope) is equal to 402.65. The coefficient of variable `bmi` was estimated to be 402.65 with a 95% confidence interval of (297.01, 508.28). In other words, there is a 95% chance that the interval (297.01, 508.28) contains the true value of the coefficient of variable `bmi`. Additionally, for every unit increase in `bmi`, there will be an average increase in medical `charges` of between $297.01 and $508.28.

```
calc_mse(charges_lm1)
## [1] "The training set mean squared error of the model is: 138358366"

# predict the medical costs billed by the health insurance company to a smoker
# with a `bmi` that is 29 and 31.5 and compute the difference in charges.
pred_costs(charges_lm1, smoker = "yes", bmi = c(29, 31.5))
##        1        2
## 12615.22 13621.84
diff_costs(charges_lm1, smoker = "yes", bmi = c(29, 31.5))
## [1] 1006.62


# Do the same for non-smokers. Comment on the results
pred_costs(charges_lm1, smoker = "no", bmi = c(29, 31.5))
##        1        2
## 12615.22 13621.84
diff_costs(charges_lm1, smoker = "no", bmi = c(29, 31.5))
## [1] 1006.62
```

The predicted difference in charges between individuals with BMI 31.5 ($13,621.84 for both smokers and non-smokers) and BMI 29 ($12615.22 for both smokers and non-smokers) is the same between smokers and non-smokers, which is a difference of $1006.62. This makes sense because the model is not dependent on smoking status and only uses BMI as the predictor (i.e. the model treats smokers and non-smokers as the same and only looks at BMI).

Model 2: Using `bmi` and `smoker` as predictors

```r
# linear model using `bmi` and `smoker` as predictors
charges_lm2 <- lm(charges ~ bmi + smoker, data = df)
kable(results_table(charges_lm2))

q1c_plot +
  # smoker model
  geom_abline(slope = coef(charges_lm2)[2],
              intercept = coef(charges_lm2)[1] + coef(charges_lm2)[3],
              col = "dark orange") +
  # non-smoker model
  geom_abline(slope = coef(charges_lm2)[2],
              intercept = coef(charges_lm2)[1],
              col = "blue")
```



	Estimate	Std. Error	t-value	p-value	95% CI LL	95% CI UL
(Intercept)	-3711.68	1018.78	-3.64	2.80e-04	-5710.35	-1713.02
bmi	398.47	32.46	12.27	8.12e-33	334.78	462.15
smokeryes	23218.77	491.04	47.28	1.02e-282	22255.43	24182.11

The general model for linear regression models with two predictors is defined as:

$$y = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2$$

or in this case:

$$charges = \beta_0 + \beta_1 \times bmi + \beta_2 \times smoker$$

Thus the models are as follows:

- If the individual is a smoker: $charges = \beta_0 + \beta_2 + \beta_1 \times bmi$
- If the individual is a non-smoker: $charges = \beta_0 + \beta_1 \times bmi$

The smoker model is has a slope of β_1 and intercept of $\beta_0 + \beta_2$, whereas the non-smoker model has a slope of β_1 and intercept of β_0, where $\beta_0 =$ -3711.68, $\beta_1 =$ 398.47, and $\beta_2 = 2.32 \times 10^4$. The coefficient of variable `bmi` was estimated to be 398.47 with a 95% confidence interval of (334.78, 462.15). For every unit increase in `bmi`, there will be an average increase in medical `charges` of between $334.78 and $462.15.

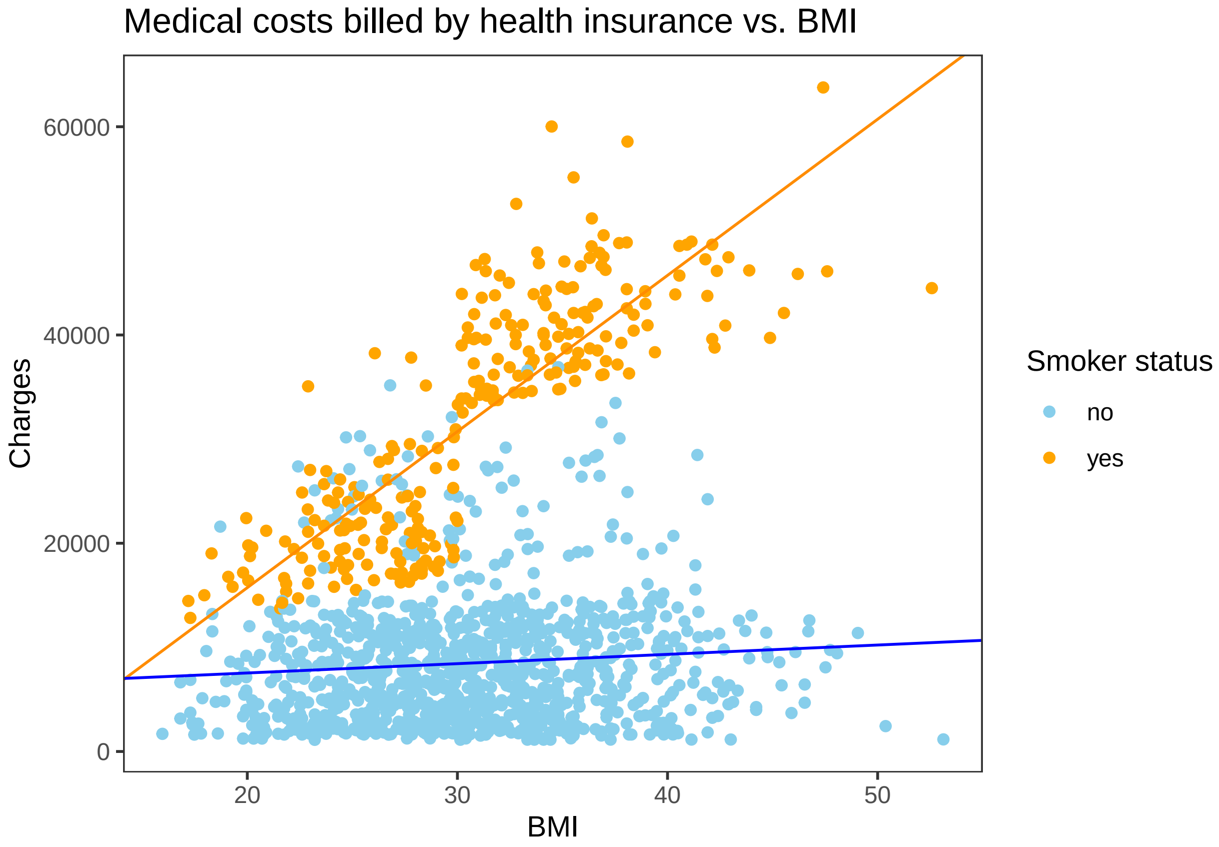
```
calc_mse(charges_lm2)
## [1] "The training set mean squared error of the model is: 50246296"

# predict the medical costs billed by the health insurance company to a smoker
# with a `bmi` that is 29 and 31.5 and compute the difference in charges.
pred_costs(charges_lm2, smoker = "yes", bmi = c(29, 31.5))
##        1        2
## 31062.62 32058.78
diff_costs(charges_lm2, smoker = "yes", bmi = c(29, 31.5))
## [1] 996.16


# Do the same for non-smokers. Comment on the results
pred_costs(charges_lm2, smoker = "no", bmi = c(29, 31.5))
##       1       2
## 7843.84 8840.01
diff_costs(charges_lm2, smoker = "no", bmi = c(29, 31.5))
## [1] 996.17
```

The predicted difference in charges between individuals with BMI 31.5 and BMI 29 is the same between smokers and non-smokers, which is $996.17. This makes sense because the smoker and non-smoker models have the same slope of $\beta_1 =$ 398.47, so the change in charges will be the same with each unit increase in BMI.

However, the difference in charges between smokers and non-smokers at BMI 29 is $23,218.78 while the difference in charges between smokers and non-smokers at BMI 31.5 is $23,218.77. This is a difference of about $23,000, which is the approximate value of β_2 (the difference between the two models).

Model 3: Using `bmi` and `smoker` as predictors with an interaction term between `bmi` and `smoker`

```
# linear model using `bmi` and `smoker` as predictors but also allowing for an
# interaction term between predictors `bmi` and `smoker`
charges_lm3 <- lm(charges ~ bmi + smoker + bmi*smoker , data = df)
kable(results_table(charges_lm3))

q1c_plot +
  # smoker model
  geom_abline(slope = coef(charges_lm3)[2] + coef(charges_lm3)[4],
              intercept = coef(charges_lm3)[1] + coef(charges_lm3)[3],
              col = "dark orange") +
  # non-smoker model
  geom_abline(slope = coef(charges_lm3)[2],
              intercept = coef(charges_lm3)[1],
              col = "blue")
```



	Estimate	Std. Error	t-value	p-value	95% CI LL	95% CI UL
(Intercept)	5750.97	991.45	5.80	8.33e-09	3805.92	7696.03
bmi	89.47	31.76	2.82	4.92e-03	27.17	151.78
smokeryes	-20008.39	2122.67	-9.43	1.94e-20	-24172.70	-15844.07
bmi:smokeryes	1410.05	67.84	20.78	7.42e-83	1276.96	1543.15

The general model with two predictors and an interaction term between the two predictors is defined as:

$$y = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \beta_3 \times x_1 x_2$$

or in this case:

$$charges = \beta_0 + \beta_1 \times bmi + \beta_2 \times smoker + \beta_3 \times bmi \times smoker$$

Thus the models are as follows:

- If the individual is a smoker: $charges = \beta_0 + \beta_2 + bmi \times (\beta_1 + \beta_3)$
- If the individual is a non-smoker: $charges = \beta_0 + \beta_1 \times bmi$

The smoker model has a slope of $\beta_1 + \beta_3$ and intercept of $\beta_0 + \beta_2$ while the non-smoker model has a slope of β_1 and intercept of β_0, where $\beta_0 = 5750.97$, $\beta_1 = 89.47$, $\beta_2 = -2 \times 10^4$, and $\beta_3 = 1410.05$. The coefficient of variable `bmi` was estimated to be 89.47 with a 95% confidence interval of (27.17, 151.78). For every unit increase in `bmi`, there will be an average increase in medical `charges` of between \$27.17 and \$151.78.

```
calc_mse(charges_lm3)
## [1] "The training set mean squared error of the model is: 37522941"

# predict the medical costs billed by the health insurance company to a smoker
# with a `bmi` that is 29 and 31.5 and compute the difference in charges.
pred_costs(charges_lm3, smoker = "yes", bmi = c(29, 31.5))
##        1        2
## 29228.89 32977.71
diff_costs(charges_lm3, smoker = "yes", bmi = c(29, 31.5))
## [1] 3748.82


# Do the same for non-smokers. Comment on the results
pred_costs(charges_lm3, smoker = "no", bmi = c(29, 31.5))
##       1       2
## 8345.72 8569.40
diff_costs(charges_lm3, smoker = "no", bmi = c(29, 31.5))
## [1] 223.68
```

The predicted difference in charges between individuals with BMI 31.5 and BMI 29 is different between smokers and non-smokers. This makes sense because the models do not have the same slope so charges will not increase at the same rate with each unit increase in BMI. For smokers, the difference in charges at BMI 31.5 (\$32,977.71) and BMI 29 (\$29,228.89) is \$3,748.82. For non-smokers, the difference in charges at BMI 31.5 (\$8,569.40) and BMI 29 (\$8,345.72) is \$223.68. The change in charges has a smaller increase for every unit increase in BMI for non-smokers compared to smokers.

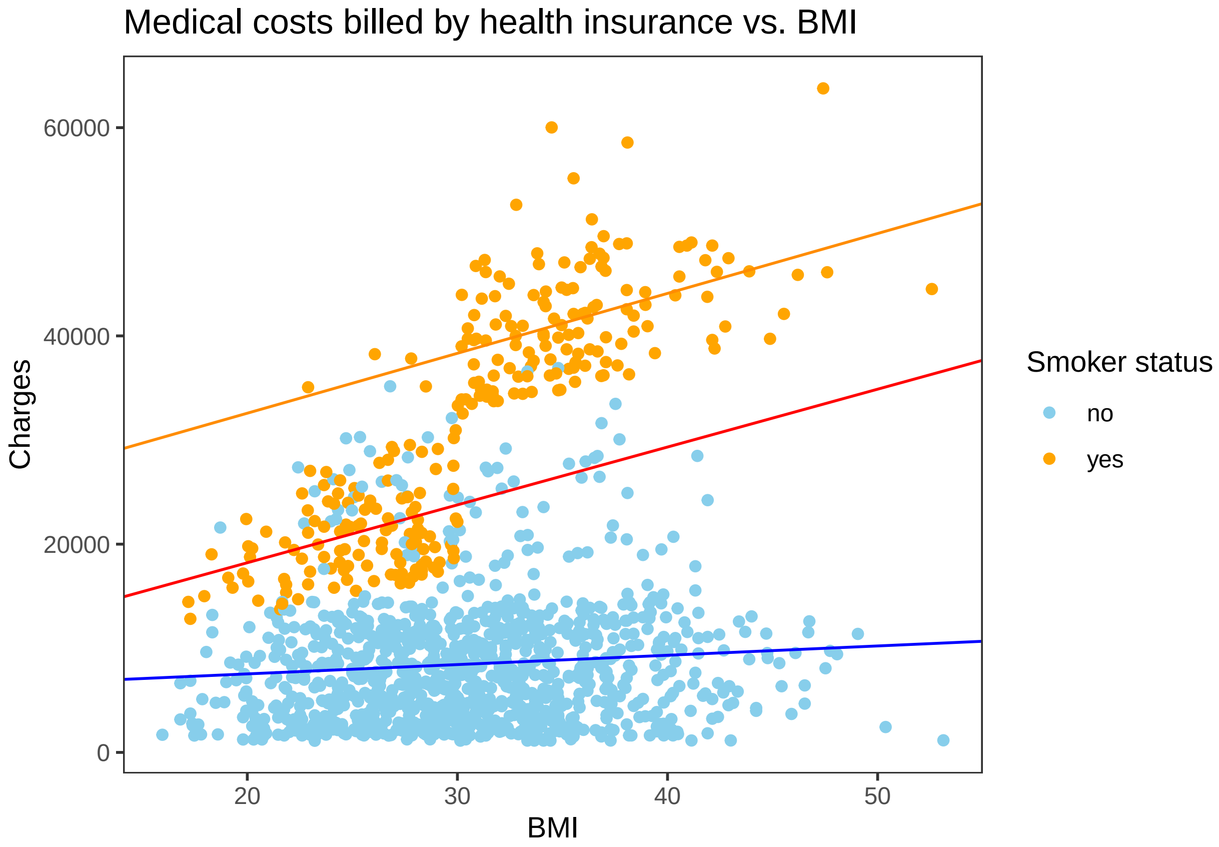
The difference in charges between smokers and non-smokers at BMI 29 is \$20,883.17 while the difference in charges between smokers and non-smokers at BMI 31.5 is \$24,408.31.

Model 4: Using bmi, smoker, and smoker_bmi30p as predictors with interaction terms

(e) Now define and add to the data set a new Boolean variable smoker_bmi30p that is True only if the subject is a smoker **and** has a bmi greater than 30. Use this newly defined variable, together with bmi and smoker, to fit the linear model represented in Figure 1 by carefully defining the interaction terms (allow each of the three straight lines to have their own intercept and slope, but use the command lm only once).

- Present your results in the form of one table where you report the estimated coefficients of the model.
- For each predictor, comment on whether you can reject the null hypothesis that there is no (linear) association between that predictor and charges, conditional on the other predictors in the model.
- Explain the interpretation of the non-significant variables in the model ($p > 0.05$) and explain how Figure 1 would change if we were to discard those variables, i.e., perform variable selection.
- According to this newly defined model, compute the predicted difference in charges between a smoker with bmi 31.5 and one with bmi 29. Do the same for non-smokers. Compare the analogous results in point (d) and comment on the results.

```
# create new variable `smoker_bmi30p` that is TRUE when subject is both a smoker
# and has a BMI greater than 30, and FALSE when subject is either not a smoker
# or is a smoker with a BMI less than 30
df <- df %>%
  mutate(smoker_bmi30p = ifelse(smoker == "yes" & bmi > 30, TRUE, FALSE))

# linear model using `bmi`, `smoker`, and `smoker_bmi30p` as predictors but also
# allowing for interaction terms
charges_lm4 <- lm(charges ~ bmi + smoker + smoker_bmi30p + bmi*smoker +
                    bmi*smoker_bmi30p, data = df)
kable(results_table(charges_lm4))

q1c_plot +
  # non-smoker model
  geom_abline(slope = coef(charges_lm4)[2],
              intercept = coef(charges_lm4)[1],
              col = "blue") +
  # smoker & bmi <= 30 model
  geom_abline(slope = sum(coef(charges_lm4)[c(2, 5)]),
              intercept = sum(coef(charges_lm4)[c(1, 3)]),
              col = "red") +
  # smoker & bmi > 30 model
  geom_abline(slope = sum(coef(charges_lm4)[c(2, 5, 6)]),
              intercept = sum(coef(charges_lm4)[c(1, 3, 4)]),
              col = "dark orange")
```



	Estimate	Std. Error	t-value	p-value	95% CI LL	95% CI UL
(Intercept)	5750.97	941.29	6.11	1.32e-09	3904.33	7597.62
bmi	89.47	30.15	2.97	3.06e-03	30.32	148.63
smokeryes	1353.55	4350.79	0.31	7.56e-01	-7181.97	9889.07
smoker_bmi30pTRUE	13966.45	6022.46	2.32	2.05e-02	2151.40	25781.51
bmi:smokeryes	466.10	168.81	2.76	5.84e-03	134.92	797.29
bmi:smoker_bmi30pTRUE	19.70	204.39	0.10	9.23e-01	-381.28	420.68

The three models are as follows:

- Non-smokers: $charges = \beta_0 + \beta_1 \times bmi$
- Smoker & BMI ≤ 30: $charges = \beta_0 + \beta_2 + (\beta_1 + \beta_4) \times bmi$
- Smoker & BMI > 30: $charges = \beta_0 + \beta_2 + \beta_3 + (\beta_1 + \beta_4 + \beta_5) \times bmi$

Null hypothesis: there is no association between the predictor and `charges`
Alternative hypothesis: there is an association between the predictor and `charges`

- `bmi`: We reject the null hypothesis because p-value $= 0.00306 < 0.05$
- `smoker`: We fail to reject the null hypothesis because p-value $= 0.75577 > 0.05$
- `smoker_bmi30p`: We reject null hypothesis because p-value $= 0.02055 < 0.05$
- `bmi` and `smoker` interaction: We reject the null hypothesis because p-value $= 0.00584 < 0.05$
- `bmi` and `smoker_bmi30p` interaction: We fail to reject the null hypothesis because p-value $= 0.92324$ > 0.05

If we were to discard the non-significant variables `smoker` and the interaction between `bmi` and `smoker_bmi30p`, then the models will then become:

- Non-smokers: $charges = \beta_0 + \beta_1 \times bmi$ (unchanged)
- Smoker & BMI ≤ 30: $charges = \beta_0 + (\beta_1 + \beta_3) \times bmi$
- Smoker & BMI > 30: $charges = \beta_0 + \beta_2 + (\beta_1 + \beta_3) \times bmi$

The model for smokers with BMI ≤ 30 has the same intercept as the non-smoker model (β_0).

```
calc_mse(charges_lm4)
## [1] "The training set mean squared error of the model is: 33769007"

# predict the medical costs billed by the health insurance company to a smoker
# with a `bmi` that is 29 and 31.5 and compute the difference in charges.
smoker_bmi29 <- round(predict(charges_lm4, data.frame(smoker = "yes",
                                                       smoker_bmi30p = FALSE,
                                                       bmi = 29)), 2)
smoker_bmi31.5 <- round(predict(charges_lm4, data.frame(smoker = "yes",
                                                         smoker_bmi30p = TRUE,
                                                         bmi = 31.5)), 2)

# Do the same for non-smokers. Comment on the results
nonsmoker_bmi29 <- round(predict(charges_lm4, data.frame(smoker = "no",
                                                       smoker_bmi30p = FALSE,
                                                       bmi = 29)), 2)
nonsmoker_bmi31.5 <- round(predict(charges_lm4, data.frame(smoker = "no",
                                                       smoker_bmi30p = FALSE,
                                                       bmi = 31.5)), 2)
```

The difference in charges between smokers and non-smokers at BMI 29 is \$14,870.47 while the difference in charges between smokers and non-smokers at BMI 31.5 is \$30,622.63. The difference in charges between smokers and nonsmokers at BMI greater than 30 is more than twice the difference in charges between smokers and nonsmokers at BMI less than 30.

Q2

Q2. This problem has to do with the notation of bias-variance trade-off. For (a) and (b), it's okay to submit hand-sketched plots: this is a conceptual exercise.

(a) Make a plot, like the one we saw in class, with "flexibility" on the x-axis. Sketch the following curves: squared bias, variance, irreducible error, expected prediction error. Be sure to label each curve. Indicate which level of flexibility is *best*.

(b) Make a plot with "flexibility" on the x-axis. Sketch curves corresponding to the training error and the test error. Be sure to label each curve. Indicate which level of flexibility is "best".

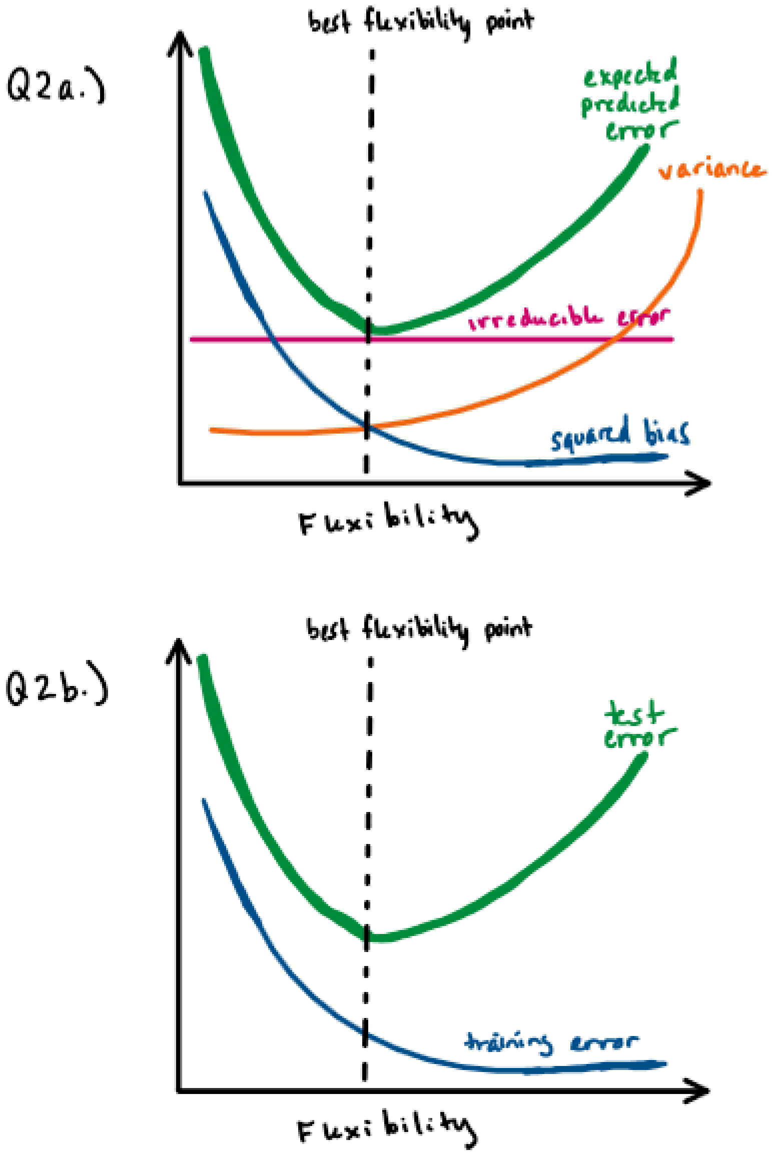


Figure 1: Plots of (a) squared bias, variance, irreducible error, and expected prediction error with respect to flexibility and (b) training and test error with respect to flexibility

Q3

Q3. This problem has to do with numerical explorations of the bias-variance trade-off phenomenon. You will generate simulated data, and will use these data to perform **linear regression**. Set the seed with `set.seed(0)` before you begin.

(a) Use the `rnorm()` function to generate a predictor vector `X` of length $n = 30$, and use `runif()` to generate a noise vector ϵ of length $n = 30$.
(b) Generate a response vector Y of length $n = 30$ according to the model: $Y = f^{true}(X) + \epsilon$, with $f^{true}(X) = 3 + 2X + 3 * X^3$

```
set.seed(0)

X <- rnorm(30)
e <- runif(30)
f_true <- 3 + 2*X + 3*X^3
Y <- f_true + e

train_set <- data.frame(Y, X)
```

(c) Fit the model $Y = f(X) + \epsilon$ to the data (using the `lm()` function), for the following choices of f:

1. $f(X) = \beta_0 + \beta_1 \times X$
2. $f(X) = \beta_0 + \beta_1 \times X + \beta_2 \times X^2$
3. $f(X) = \beta_0 + \beta_1 \times X + \beta_2 \times X^2 + \beta_3 \times X^3 + \beta_4 \times X^4$
4. $f(X) = \beta_0 + \beta_1 \times X + \beta_3 \times X^3$

(d) For each of the models above, compute the training mean squared error (MSE). Comment on the results.

```
Q3c1 <- lm(Y ~ X, train_set)
Q3c2 <- lm(Y ~ X + I(X^2), train_set)
Q3c3 <- lm(Y ~ X + I(X^2) + I(X^3) + I(X^4), train_set)
Q3c4 <- lm(Y ~ X + I(X^3), train_set)

kable(data.frame(`Model` = c(1, 2, 3, 4),
                 `Training MSE` = round(c(mean(Q3c1$residuals^2),
                                          mean(Q3c2$residuals^2),
                                          mean(Q3c3$residuals^2),
                                          mean(Q3c4$residuals^2)), 3)))
```

Model	Training.MSE
1	26.580
2	13.205
3	0.055
4	0.057

As the models increases in complexity, the training MSE decreases until it reaches close to 0.

(e) Now generate 10K (new) **test** observations following steps 3(a) and 3(b). Compute the test MSE of the models fitted in 3(c) on these **test** observations. Report and comment on the results.

```r
set.seed(0)

X_test <- rnorm(10000)
e_test <- runif(10000)
Y_test <- 3 + 2*X_test + 3*X_test^3 + e_test
test_set <- data.frame(Y_test, X_test)
colnames(test_set) <- c("Y", "X")

Q3c1_pred <- predict(Q3c1, test_set)
Q3c1_mse <- mean((Y_test - Q3c1_pred)^2)

Q3c2_pred <- predict(Q3c2, test_set)
Q3c2_mse <- mean((Y_test - Q3c2_pred)^2)

Q3c3_pred <- predict(Q3c3, test_set)
Q3c3_mse <- mean((Y_test - Q3c3_pred)^2)

Q3c4_pred <- predict(Q3c4, test_set)
Q3c4_mse <- mean((Y_test - Q3c4_pred)^2)

kable(data.frame(`Model` = c(1, 2, 3, 4),
                 `Training MSE` = round(c(mean(Q3c1$residuals^2),
                                          mean(Q3c2$residuals^2),
                                          mean(Q3c3$residuals^2),
                                          mean(Q3c4$residuals^2)), 3),
                 `Test MSE` = round(c(Q3c1_mse, Q3c2_mse,
                                      Q3c3_mse, Q3c4_mse), 3)))
```

Model	Training.MSE	Test.MSE
1	26.580	49.027
2	13.205	78.753
3	0.055	0.116
4	0.057	0.098

As previously mentioned, the training MSE decreases until it reaches close to 0 as the fitted models increases in complexity. Model 3, $f(X) = \beta_0 + \beta_1 \times X + \beta_2 \times X^2 + \beta_3 \times X^3 + \beta_4 \times X^4$, has the smallest training MSE and is the model that has the highest complexity/flexibility. We do not see the same trend with the test MSE values because there is a bias-variance trade-off for minimizing MSE. At a certain point, increasing model complexity will result in an increase in test MSE instead of decrease. We see that Model 4, $f(X) = \beta_0 + \beta_1 \times X + \beta_3 \times X^3$, has the lower test MSE compared to Model 3.

(f) Compute the training and test MSEs of the true regression function f^{true}. Compare to those of the models fitted in 3(c). Comment on the results.

```r
ft_trainMSE <- mean((Y - f_true)^2) # training MSE of f_true

f_true_test <- 3 + 2*X_test + 3*X_test^3

ft_testMSE <- mean((Y_test - f_true_test)^2) # test MSE of f_true
```

The training MSE of the true regression function $f^{true}(X) = 3 + 2X + 3 * X^3$ is 0.3381281 and the test MSE is 0.3285228. The training MSE value of f^{true} is not as small as the training MSE values obtained for fitted models $f(X) = \beta_0 + \beta_1 \times X + \beta_2 \times X^2 + \beta_3 \times X^3 + \beta_4 \times X^4$ and $f(X) = \beta_0 + \beta_1 \times X + \beta_3 \times X^3$, however this makes sense since training MSE will continue to decrease as model complexity increases.

It is interesting to note that the test MSE for $f^{true} = 0.33 >$ test MSE value for Model 4, $f(X) = \beta_0 + \beta_1 \times X + \beta_3 \times X^3$ even though they have the same form. Looking at the coefficient estimates for Model 4 shows that they are very close to the coefficients of $f^{true}(X) = 3 + 2X + 3 * X^3$.

```
kable(results_table(Q3c4))
```

	Estimate	Std. Error	t-value	p-value	95% CI LL	95% CI UL
(Intercept)	3.53	0.05	75.63	5.67e-33	3.44	3.63
X	1.98	0.08	24.24	7.40e-20	1.81	2.15
I(X^3)	2.97	0.03	112.16	1.41e-37	2.92	3.03