Problem Set 8 Key

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

1 Problem 1

The dataset penguins preloaded in R includes data on the size and sex of adult penguins in the Palmer Archipelago. Suppose we are interested in testing for a correlation between body mass and flipper length.

```
glimpse(penguins)
```

```
Rows: 344
Columns: 8
$ species     <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Ad~
$ island      <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgersen, Tor~
$ bill_len    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, 42.0, ~
$ bill_dep    <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, 20.2, ~
$ flipper_len <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186, 180,~
$ body_mass   <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, 4250, ~
```

```
$ sex          <fct> male, female, female, NA, female, male, female, male, NA, ~
$ year         <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007~
```

a. Estimate the correlation coefficient for the sample and provide 95% confidence interval. [2 pts.]

```
ctest <- cor.test(penguins$body_mass,penguins$flipper_len)
ctest
```

```
    Pearson's product-moment correlation

data:  penguins$body_mass and penguins$flipper_len
t = 32.722, df = 340, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.843041 0.894599
sample estimates:
      cor
0.8712018
```

b. Perform a hypothesis test that tests whether there is a correlation between body mass and flipper length. What is the value of the test statistic? What is the p-value? [2pts.]

```
ctest$statistic
```

```
       t
32.72223
```

```
ctest$p.value
```

```
[1] 4.370681e-107
```

Or <2.2e-16

c. Is there evidence of a correlation between body mass and flipper length? Use a significance level of 0.01. [1 pt.]

> **Note**
>
> We reject the null hypothesis at the significance level of 0.01. There is sufficiet evidence of a correlation between body mass and flipper length.

2 Problem 2

The Pima Indian Diabetes Dataset, uploaded as PimaIndiansDiabetes.csv on Canvas, contains information of 768 women from a population of Pima Indians near Phoenix, Arizona. Suppose we want to estimate the glucose levels based on BMI levels.

```
pima <- read.csv("datasets/PimaIndiansDiabetes.csv")
glimpse(pima)
```

```
Rows: 768
Columns: 9
$ Pregnancies              <int> 6, 1, 8, 1, 0, 5, 3, 10, 2, 8, 4, 10, 10, 1, ~
$ Glucose                  <int> 148, 85, 183, 89, 137, 116, 78, 115, 197, 125~
$ BloodPressure            <int> 72, 66, 64, 66, 40, 74, 50, NA, 70, 96, 92, 7~
$ SkinThickness            <int> 35, 29, NA, 23, 35, NA, 32, NA, 45, NA, NA, N~
$ Insulin                  <int> NA, NA, NA, 94, 168, NA, 88, NA, 543, NA, NA,~
$ BMI                      <dbl> 33.6, 26.6, 23.3, 28.1, 43.1, 25.6, 31.0, 35.~
$ DiabetesPedigreeFunction <dbl> 0.627, 0.351, 0.672, 0.167, 2.288, 0.201, 0.2~
$ Age                      <int> 50, 31, 32, 21, 33, 30, 26, 29, 53, 54, 30, 3~
$ Outcome                  <int> 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, ~
```

a. Estimate the best fit line using regression methods. What is the equation of the best fit line? [1pt.]

```
mod1 <- lm(Glucose~BMI,data=pima)
summary(mod1)
```

```
Call:
lm(formula = Glucose ~ BMI, data = pima)

Residuals:
   Min     1Q Median     3Q    Max
-70.28 -21.41  -4.11  18.32  81.80
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  88.5775     5.2046  17.019  < 2e-16 ***
BMI           1.0280     0.1568   6.555 1.04e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.78 on 750 degrees of freedom
  (16 observations deleted due to missingness)
Multiple R-squared:  0.05418,   Adjusted R-squared:  0.05292
F-statistic: 42.96 on 1 and 750 DF,  p-value: 1.037e-10
```

The equation of the best fit line is $\hat{y} = 88.5775 + 1.028 * BMI$.

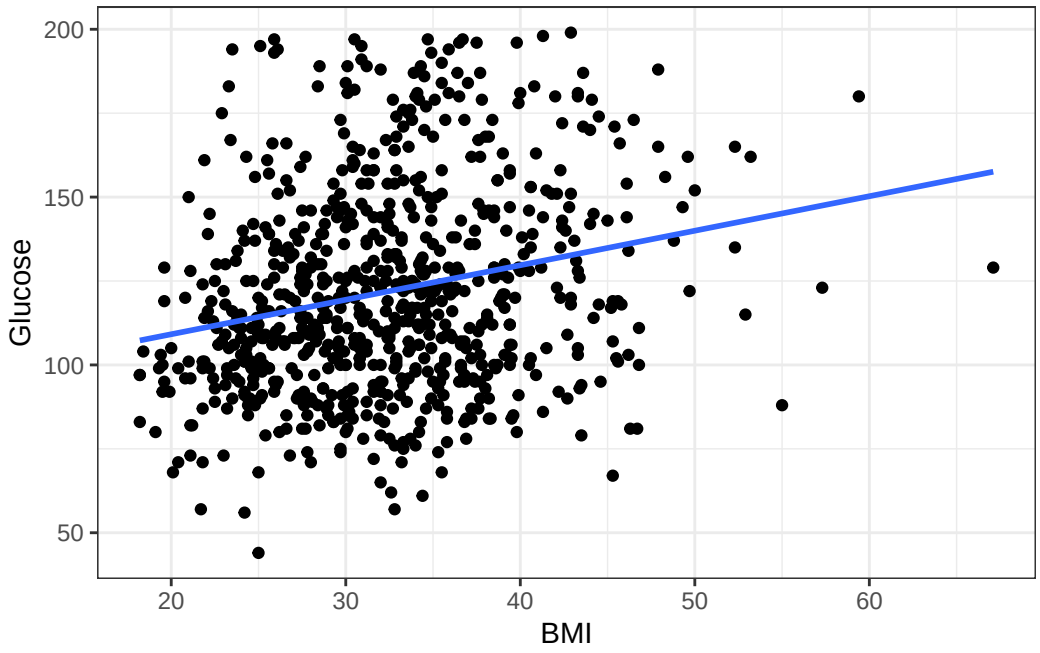
b. Create a scatter plot of the data using ggplot2. Include an overlay of the best fit line using geom_smooth(). Comment on the fit of the regression line on the data. [1pt.]

```
ggplot(data=pima,aes(x=BMI,y=Glucose)) +
  geom_point() +
  geom_smooth(method="lm",se=F) +
  theme_bw()
```

```
`geom_smooth()` using formula = 'y ~ x'
```

```
Warning: Removed 16 rows containing non-finite outside the scale range
(`stat_smooth()`).
```

```
Warning: Removed 16 rows containing missing values or values outside the scale range
(`geom_point()`).
```



c. Is there evidence of an association between the glucose levels and BMI? Provide the value of the test statistic, p-value, and coefficient estimate. [2 pts.] `t=6.555`, p=1.04e-10, estimate $\beta = 1.028$.

H_0 : There is no association between glucose and BMI OR the coefficient of BMI in the regression model is zero. ($\beta = 0$)

H_a : There is an association between glucose and BMI OR the coefficient of BMI in the regression model is not zero. ($\beta \neq 0$)

At significance level of 0.05, we reject the null hypothesis. We have sufficient evidence that there is an association between glucose levels and BMI.

d. What is the coefficient of determination? [1 pt.]

Coefficient of determination is the R^2. Based on the summary output, $R^2 = 0.05418$.

e. What is the predicted glucose level of an individual with a BMI of 25? Provide the 95% prediction interval. [2pts.]

```
newdata <- data.frame(BMI=c(25))

 predict(mod1,newdata,interval="prediction",level=0.95)
```

```
      fit      lwr      upr
1 114.278 55.73112 172.8248
```

Predicted glucose level is 114.28, 95% prediction confidence interval: (55.7,172.8)

3 Problem 3

The file FirstYearGPA.csv includes data from 219 first year students from a university in the United States. Suppose we want to predict the first-year GPA in college (variable GPA) based on gender (variable Male; 1=Male), if the student is FirstGen (1 = student is first in family to attend college), SAT verbal score (variable SATV), and SAT math score (variable SATM).

```
first <- read.csv("datasets/FirstYearGPA.csv")
glimpse(first)
```

```
Rows: 219
Columns: 11
$ Obs          <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17~
$ GPA          <dbl> 3.06, 4.15, 3.41, 3.21, 3.48, 2.95, 3.60, 2.87, 3.67, 3.4~
$ HSGPA        <dbl> 3.83, 4.00, 3.70, 3.51, 3.83, 3.25, 3.79, 3.60, 3.36, 3.7~
$ SATV         <int> 680, 740, 640, 740, 610, 600, 710, 390, 630, 680, 380, 63~
$ SATM         <int> 770, 720, 570, 700, 610, 570, 630, 570, 560, 670, 470, 67~
$ Male         <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, ~
$ HU           <dbl> 3.0, 9.0, 16.0, 22.0, 30.5, 18.0, 5.0, 10.0, 8.5, 16.0, 1~
$ SS           <dbl> 9.0, 3.0, 13.0, 0.0, 1.5, 3.0, 19.0, 0.0, 15.5, 12.0, 7.0~
$ FirstGen     <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, ~
$ White        <int> 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, ~
$ CollegeBound <int> 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
```

a. Estimate the best fit line using multiple linear regression. What is the equation of the best fit line? [1pt.]

```
mod1 <- lm(GPA~Male+FirstGen+SATV + SATM,data=first)
sum_mod <- summary(mod1)
sum_mod
```

```
Call:
lm(formula = GPA ~ Male + FirstGen + SATV + SATM, data = first)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.15833 -0.32630  0.04695  0.34933  1.11795

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.0551881  0.2927031   7.021 2.87e-11 ***
Male        -0.0076621  0.0651461  -0.118 0.906483
FirstGen    -0.1202098  0.0981175  -1.225 0.221861
SATV         0.0014533  0.0004351   3.340 0.000988 ***
SATM         0.0002821  0.0005042   0.560 0.576375
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4456 on 214 degrees of freedom
Multiple R-squared:  0.1005,    Adjusted R-squared:  0.08373
F-statistic: 5.981 on 4 and 214 DF,  p-value: 0.00014
```

The best fit line is: $\hat{y} = 2.055 - 0.0077 \times Male - 0.1202 \times FirstGen + 0.0014 \times SATV + 0.0003 \times SATM$.

b. What is the result of the test of significance of regression? Is there evidence that at least one of the coefficients is non-zero? Use a significance level of 0.05. Provide the test statistic and the p-value [2 pts.]

```
sum_mod$fstatistic
```

```
    value      numdf      dendf
 5.980545   4.000000 214.000000
```

The p-value is 0.00014. We reject the null hypothesis. There is sufficient evidence that at least one of the coefficients is non-zero.

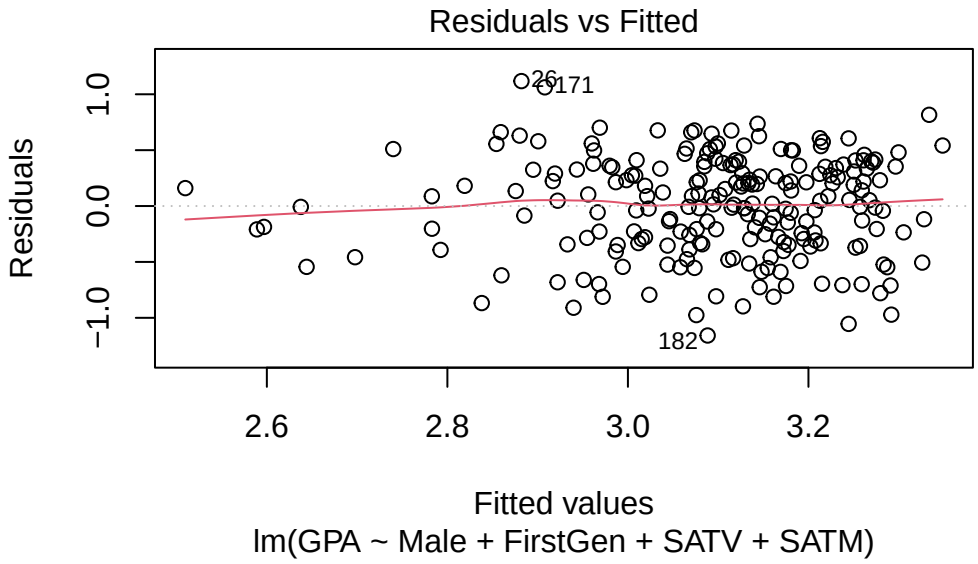
c. If applicable, which predictors are associated with first year GPA? Use a significance level of 0.05. [2pts.]

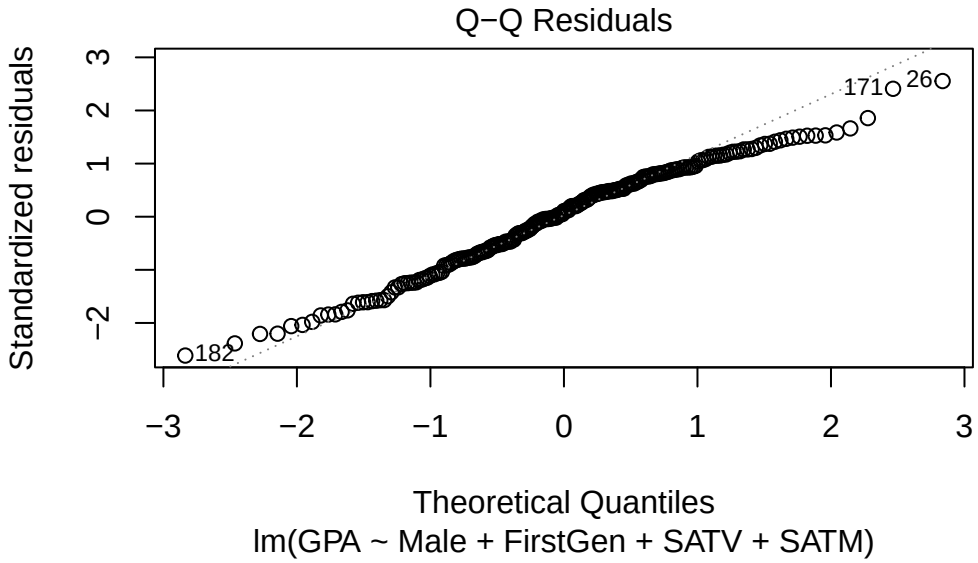
Only SATV yielded sufficient evidence of an association with first year GPA ($\beta = 0.0014$, p=0.0010).

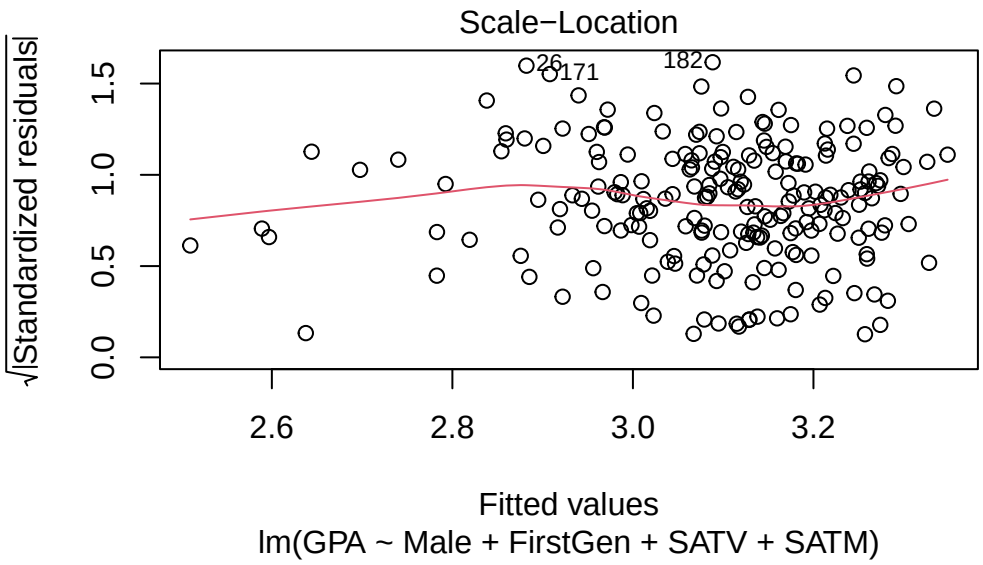
d. What is the adjusted coefficient of determination (adjusted R2) value? [1pt.] Adjusted R^2 is 0.08373.

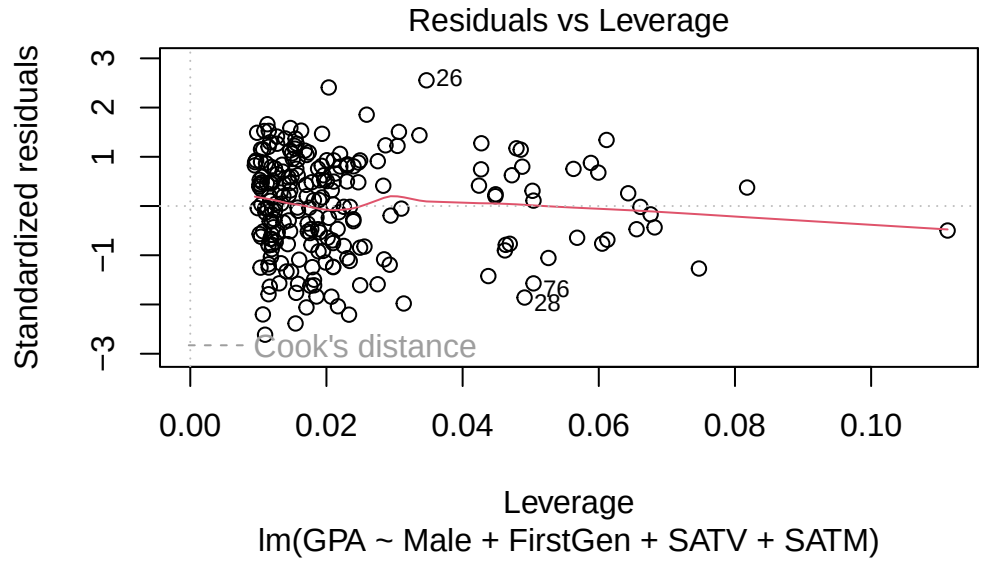
e. Comment on the model fit based on model diagnostic plots. [2 pts.]

```
plot(mod1)
```


Residuals vs Fitted
Residuals
1.0
0.0
−1.0
26
171
182
2.6
2.8
3.0
3.2
Fitted values
lm(GPA ~ Male + FirstGen + SATV + SATM)


Q−Q Residuals
Standardized residuals
3
2
1
0
−2
171
26
182
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(GPA ~ Male + FirstGen + SATV + SATM)





Adjusted R^2 is low, indicative of poor model fit. Model assumptions seem to hold (no influential points, not a big deviation from normality, no evidence of unequal variance and nonlinearity).