Problem Set 1

Zia Hassan

These exercises cover Modules 1 through 3 of Machine Learning Methods and Applications. There is a point value for each exercise. There are 24 total points across two exercises. Submit your `.qmd` and your rendered PDF or HTML with your name in both filenames. You can do this! Good luck!

Table of contents

Instructions

For each exercise, write your answers in a well-formatted combination of text and code blocks. You are required to submit your answers in Quarto because of its ability to combine text and code seamlessly. Answer the question first, then follow that with any supporting code. Commenting your code will make it easier for me to grade your work as you intended it to be understood.

The guidance within the document uses R, but you can use R or Python as you wish. Note that while specific functions may be recommended, you won't be penalized for utilizing alternatives that produce similar results. Points may be deducted for untidy formatting. **You will submit your complete `.qmd` file and your rendered .pdf or .html document.**

> Use line breaks (\) to make sure your code appears in full on the page. Help your instructor grade faster by not having to consult the .qmd every time you don't use line breaks!
>
> Compare the following:
> aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
> with
> aaaaa
> aaaaa
> aaaaa

You may collaborate with your classmates, but every keystroke that goes into your final work must be your own. Do not copy or paste another student's exact language or substantive or numeric examples. You can talk about the assignment and study together, but I expect you to submit your own work. I understand that there are only so many ways to code simple mathematical operations. But if I see substantive examples, critical reasoning/interpretations/responses, or vectors of example numbers that are the same, this can be problematic from an academic integrity standpoint. Make this your own work.

I assume that some of you will find using an LLM like Google Gemini or ChatGPT helpful in completing this assignment. I do not object to you using these tools to find information or generate ideas. However, academic integrity and good professional practice require that you not copy the results of an LLM query blindly or uncritically. **If you use an LLM, be prepared to describe your queries and how you adapted the responses to represent your own work.** You should know that LLMs include a lot of comments in their code when you ask it a coding question. If I see an unnatural level of documentation in your code, I may assume that you copied it directly from an LLM. Similarly, many LLMs are quite verbose and explain every step when you ask them a question. I don't need you to document every line in painstaking detail. Answer the questions presented directly and concisely. This presentation may factor into the points you earn on a question.

Objectives

These exercises will help you develop your skills with the following:

- Calculate and interpret descriptive statistics. (CLO 1)
- Calculate and interpret regression results. (CLO 1, 3)

Recommended Timeline

I highly recommend completing this problem set over the course of several weeks. Here's one possible timeline you might follow:

- Exercise 1 (1 week)
- Exercise 2 (1 week)

Exercise 1 - 11 points

This exercise is modeled off Exercises 2.9 and 4.14 in **An Introduction to Statistical Learning (with Applications in R) (second edition)**.

Download `sim_data.csv`, which is posted on Canvas. You should subset your data to include only `absentee` as your outcome variables and four quantitative predictor variables. Note that you may find it helpful to use the data.frame() function to create a single data set containing both absentee and the four predictor variables. Make sure that the missing values have been removed from the data.

```r
# loading the data
sim_data <- read.csv("sim_data.csv")
```

```r
# exploring the data
names(sim_data)
```

```
[1] "X"        "absentee" "nonwhite" "lunch"    "IEP"      "GPA"      "change"
[8] "income"
```

```r
# checking data types and summary
summary(sim_data)
```

```
       X             absentee          nonwhite         lunch      
 Min.   :   1.0   Min.   : 0.5646   Min.   :35.04   Min.   :19.80  
 1st Qu.: 250.8   1st Qu.: 5.3206   1st Qu.:50.00   1st Qu.:38.09  
 Median : 500.5   Median : 6.3702   Median :54.24   Median :44.39  
 Mean   : 500.5   Mean   : 6.3408   Mean   :54.55   Mean   :44.55  
 3rd Qu.: 750.2   3rd Qu.: 7.3542   3rd Qu.:58.94   3rd Qu.:51.09  
 Max.   :1000.0   Max.   :11.3776   Max.   :73.26   Max.   :74.98  
                                    NA's   :2       NA's   :1      
      IEP              GPA            change          income      
 Min.   : 3.471   Min.   :1.052   Min.   :2.591   Min.   : 12335  
 1st Qu.:12.778   1st Qu.:1.775   1st Qu.:3.690   1st Qu.: 50519  
 Median :14.878   Median :1.980   Median :3.990   Median : 59333  
 Mean   :14.890   Mean   :1.994   Mean   :3.984   Mean   : 59812  
 3rd Qu.:17.157   3rd Qu.:2.220   3rd Qu.:4.298   3rd Qu.: 69707  
 Max.   :23.530   Max.   :3.077   Max.   :5.245   Max.   :100574  
                                  NA's   :1       NA's   :1       
```

```r
# subsetting the data
subset <- data.frame(
  absentee = sim_data$absentee,
  GPA = sim_data$GPA,
  income = sim_data$income,
  change = sim_data$change,
  IEP = sim_data$IEP
)
```

```r
# clean the data for missing values
subset <- na.omit(subset)
```

```r
# Package installation
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

> Note
>
> These data are made up, but their distributions may correspond to somewhat realistic parameters; however, don't update any of your beliefs based on their values. The unit of analysis is the individual classroom. `absentee`, `nonwhite`, `lunch`, `IEP`, and `change` are percentages, and, respectively, they denote percentage of a classroom that are chronically absent; are nonwhite; receive free or reduced lunch; have an individual education plan; and have changed schools in the last year. `GPA` denotes mean grade point average. `income` denotes parental income. Assume that values in each observation are independent from values in other classrooms.

1.1 - 1 point

What is the range of each quantitative predictor? You can answer this using the range() function.

Your answer here.

The range of GPA is 1.05-3.08. The range of income is 12,334.57-100,573.68. The range of change is 2.59-5.24. The range of IEP is 3.47-23.52.

```r
### Your code here.
range(subset$GPA)
```

```
[1] 1.052324 3.076501
```

```r
range(subset$income)
```

```
[1]  12334.57 100573.68
```

```r
range(subset$change)
```

```
[1] 2.590961 5.244809
```

```r
range(subset$IEP)
```

```
[1]  3.471439 23.529557
```

1.2 - 1 point

What is the mean and standard deviation of each quantitative predictor?

Note

The mean of GPA is 1.99, and it's SD is .38. The mean of income is $59,834 and it's SD is $14,160. The mean of change is 3.98% and its SD is .42%. The mean of IEP is 14.89% and its SD is 3.23%.

```r
### Your code here.
mean(subset$GPA)
```

```
[1] 1.993669
```

```r
mean(subset$income)
```

```
[1] 59834.13
```

```r
mean(subset$change)
```

```
[1] 3.983662
```

```r
mean(subset$IEP)
```

```
[1] 14.89108
```

```r
sd(subset$GPA)
```

```
[1] 0.3378689
```

```r
sd(subset$income)
```

```
[1] 14159.53
```

```r
sd(subset$change)
```

```
[1] 0.422509
```

```r
sd(subset$IEP)
```

```
[1] 3.232009
```

1.3 - 2 points

Now take a random sample of at least 100 observations. What is the range, mean, and standard deviation of each predictor in this sample? Do these differ from the values in the full sample?

Your answer here.

The mean of GPA is 2.069 and its SD is 0.364 The mean of IEP is 15.183 and its SD is 3.646 The mean of income is 61,861.44 and its SD is 17,362.17 The mean of change is 3.950 and its SD is 0.464

```r
### Your code here.
sample_data <- subset %>% slice_sample(n = 100)
dim(sample_data)
```

```
[1] 100   5
```

```r
head(sample_data)
```

```
  absentee      GPA   income   change       IEP
1 6.664005 2.087277 59039.42 3.630982 17.265419
2 5.869007 1.914791 46993.37 4.030992  9.687595
3 6.888344 1.920527 60033.57 3.645913 14.163586
4 7.031543 2.560119 86354.42 3.273140 20.702230
5 5.254406 2.620626 61254.25 3.571512 13.122172
6 7.946277 1.472981 44171.73 4.354382 14.775000
```

```r
mean(sample_data$GPA)
```

```
[1] 1.92133
```

```r
mean(sample_data$IEP)
```

```
[1] 14.79114
```

```r
mean(sample_data$income)
```

```
[1] 59467.05
```

```r
mean(sample_data$change)
```

```
[1] 4.027231
```

```r
sd(sample_data$GPA)
```

```
[1] 0.3655285
```

```r
sd(sample_data$IEP)
```

```
[1] 3.317137
```

```r
sd(sample_data$income)
```

```
[1] 14503.61
```

```r
sd(sample_data$change)
```

```
[1] 0.4388671
```

1.4 - 2 points

Using the full data set, investigate the predictors and outcome graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

Note

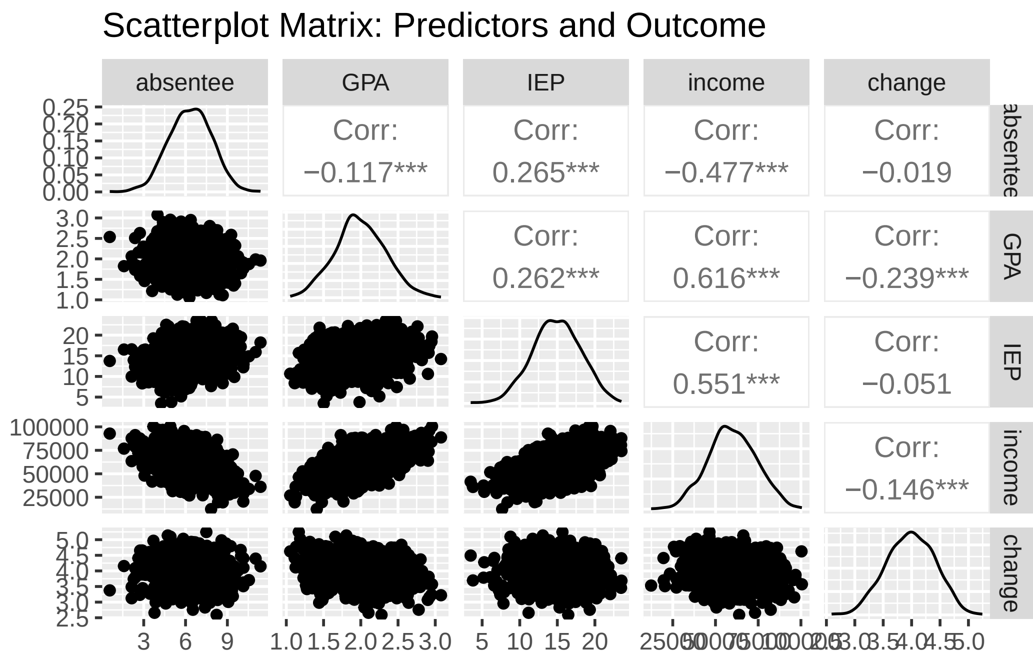
I notice:
A positive, strong, correlation between GPA and Income (.616), a positive but moderate to weak correlation between GPA and IEP (.262) , a moderate to strong positive correlation between IEP and income (.551). When it comes to how change relates to the other 3 continuous variables, the shape looks much more cloud like, and the numbers reflect this as most are close to 0 (ex -.14, .2, etc). While there is some evidence of a relationship it does appear to be as strong as the other relationships discussed here. What surprised me most was the correlation between GPA and IEP, though it is somewhat weak. I suppose that is explained by the fact that having support for whatever disability or condition you have would help with grades, as it is meant to do.
For the predictor (absentee), the strongest relationship seems to be negative for income (-.477). The others are far weaker correlations, with IEP being the only close second (though still weak). As a former educator, this seems to make sense.

```r
### Your code here.
# install.packages("GGally")
# install.packages("ggplot2")
library(GGally)
```

```
Loading required package: ggplot2

Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2
```

```r
library(ggplot2)
df <- subset
# scatterplot of all vars
ggpairs(subset,
        columns = c("absentee", "GPA", "IEP", "income", "change"),
        title = "Scatterplot Matrix: Predictors and Outcome")
```



1.5 - 2 points

Suppose that we wish to predict `absentee` on the basis of the other variables. Do your plots suggest that any of the variables might be useful in predicting `absentee`? Justify your answer.

Your answer here.

Income and IEP are probably the most useful for predicting absentee rates, with income showing the strongest relationship (-0.477) and IEP a close second. GPA has a weak positive relationship with absentee (.11) so it might be useful. Change, being close to 0, does not seem particularly useful.

```r
cor(subset$absentee, subset$income, use = "complete.obs")
```

```
[1] -0.4772694
```

```r
cor(subset$absentee, subset$IEP, use = "complete.obs")
```

```
[1] 0.2648619
```

```r
cor(subset$absentee, subset$GPA, use = "complete.obs")
```

```
[1] -0.1171887
```

```r
cor(subset$absentee, subset$change, use = "complete.obs")
```

```
[1] -0.01927614
```

1.6 - 1 point

Create a binary variable, `absentee_bin`, that contains a 1 if `absentee` contains a value above its median, and a 0 if `absentee` contains a value below its median. You can compute the median using the median() function.

Note

The code to create the binary variable is provided below.

```r
# calculate median
absentee_median <- median(subset$absentee)

# create binary
subset$absentee_bin <- ifelse(subset$absentee > absentee_median, 1, 0)
```

1.7 - 2 points

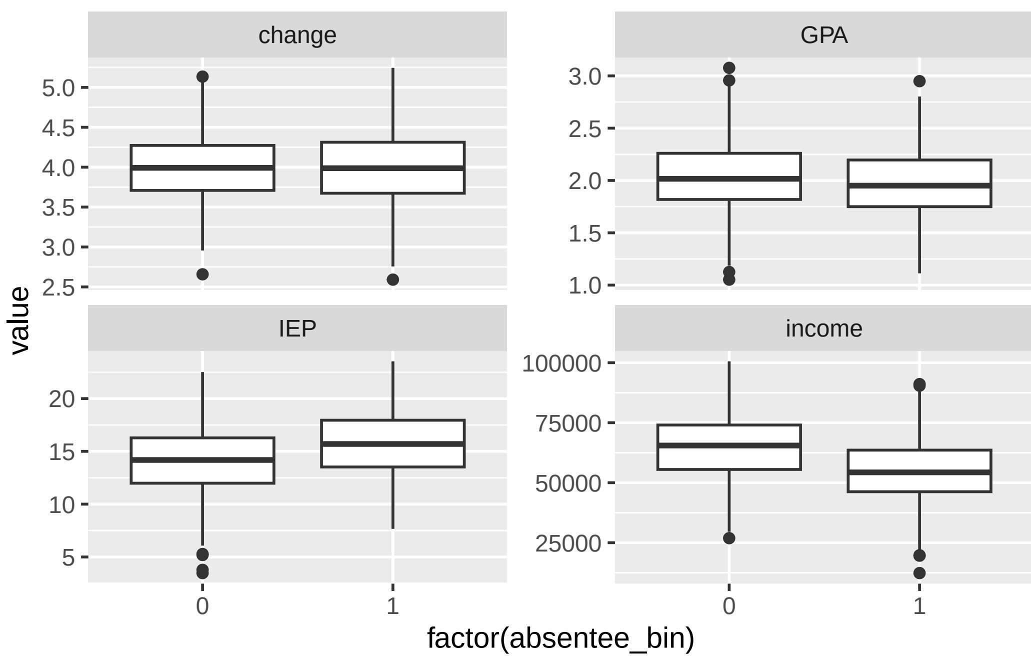
Explore the data graphically in order to investigate the association between `absentee_bin` and the four selected predictors/features. Which of the other features seem most likely to be useful in predicting `absentee_bin`? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

Note

The boxplots confirm my assertions in the previous answer. Even with the binary variable, income and IEP are the best predictors. However, the visual is making me think that GPA might be worth including as well. Although it is not a gigantic difference, it is enough that it might make sense to include it as a predictor. I could always remove it later.

```r
# pivot to long
library(tidyr)
subset_long <- subset %>%
  pivot_longer(cols = c(GPA, income, IEP, change),
               names_to = "predictor", values_to = "value")

ggplot(subset_long, aes(x = factor(absentee_bin), y = value)) +
  geom_boxplot() +
  facet_wrap(~predictor, scales = "free_y")
```



Exercise 2 - 13 points

This exercise is modeled off Exercises 3.8 and 3.9 in **An Introduction to Statistical Learning (with Applications in R) (second edition)**.

Download `sim_data.csv`, which is posted on Canvas. You should subset your data to include only `absentee` as your outcome variables and four quantitative predictor variables. Avoid selecting predictor variables that could not have a plausible causal effect on the response. Note that you may find it helpful to use the data.frame() function to create a single data set containing both absentee and the four predictor variables. Make sure that the missing values have been removed from the data.

```r
# Create subset2 for Problem 2, just to be clean!
subset2 <- data.frame(
  absentee = sim_data$absentee,
  GPA = sim_data$GPA,
  income = sim_data$income,
  change = sim_data$change,
  IEP = sim_data$IEP
)

# Remove missing values
subset2 <- na.omit(subset2)
```

2.1 - 2 points

Use the lm() function to perform a simple bivariate linear regression with `absentee` as the response and only one of the four quantitative variables as the predictor. Use the summary() function to print the results. Comment on the output. For example:

- Is there a relationship between the predictor and the response?
- How strong is the relationship between the predictor and the response?
- Is the relationship between the predictor and the response positive or negative?
- What is the predicted `absentee` associated with a specific value of the predictor of your choosing? What are the associated 95% confidence and prediction intervals?

Your answer here.

There is a strong relationship between the predictor and response. The p-value is less than .001, which means is statistically significant at almost any threshold, and certainly under the .05 threshold that is standard in social science. Additionally, the adjusted r squared is high for a bivariate regression at .22. The relationship between income and absentee is negative, given that the coefficient is negative.
I used a $50,000 value for income to predict a 6.84% absentee rate.
6.74% to 6.95% confidence interval 4.22% to 9.47% prediction interval

```r
### Your code here.
# bivirate regression
model1 <- lm(absentee ~ income, data = subset2)
summary(model1)
```

```
Call:
```

```
lm(formula = absentee ~ income, data = subset2)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.0868 -0.8961  0.0668  0.9576  4.0674 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.406e+00  1.838e-01   51.18   <2e-16 ***
income      -5.123e-05  2.989e-06  -17.14   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.336 on 996 degrees of freedom
Multiple R-squared:  0.2278,    Adjusted R-squared:  0.227 
F-statistic: 293.8 on 1 and 996 DF,  p-value: < 2.2e-16
```

```r
# 95% confidence interval
predict(model1, newdata = data.frame(income = 50000),
        interval = "confidence")
```

```
      fit      lwr      upr
1 6.84419 6.743114 6.945266
```

```r
# 95% prediction interval
predict(model1, newdata = data.frame(income = 50000),
        interval = "prediction")
```

```
      fit      lwr      upr
1 6.84419 4.220046 9.468334
```

2.2 - 2 points

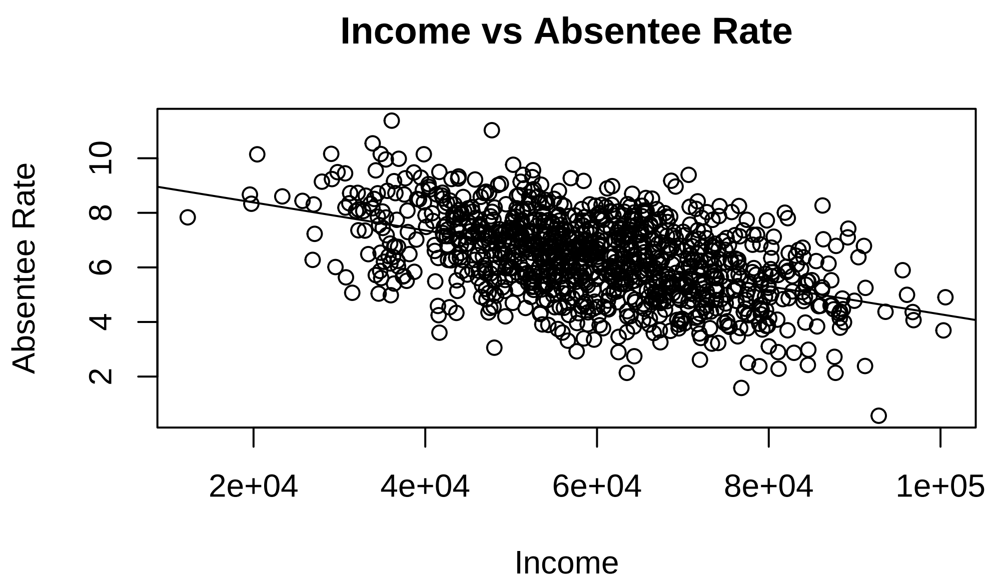
Plot the relationship between the response and the predictor. Use the abline() function to display the least squares regression line.

Note

Plot is below.

```
### Your code here.
plot(subset2$income, subset2$absentee,
     xlab = "Income",
     ylab = "Absentee Rate",
     main = "Income vs Absentee Rate")

# use abline for the regression
abline(model1)
```



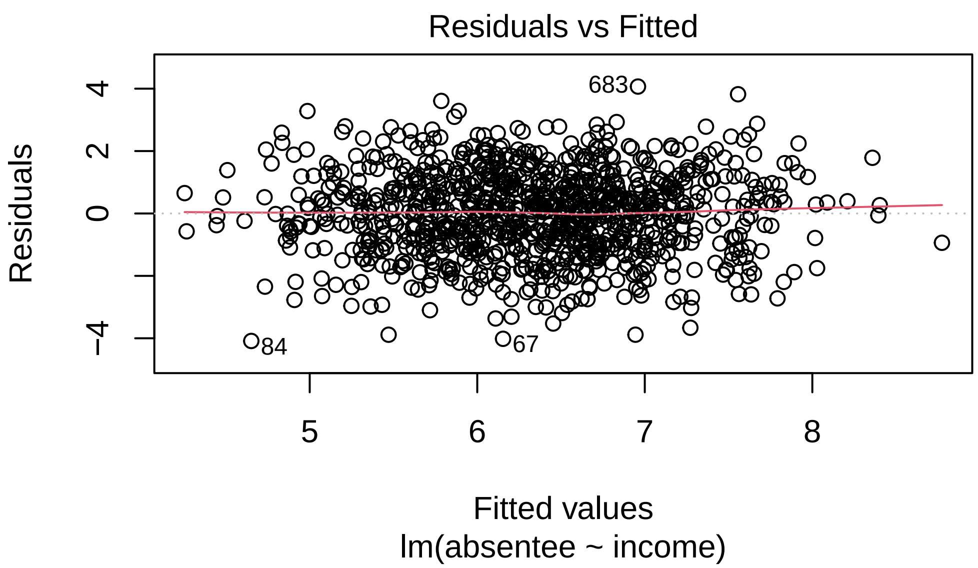
2.3 - 2 points

Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit. Do residual plots suggest any unusually large outliers? Does a leverage plot identify any observations with unusually high leverage?

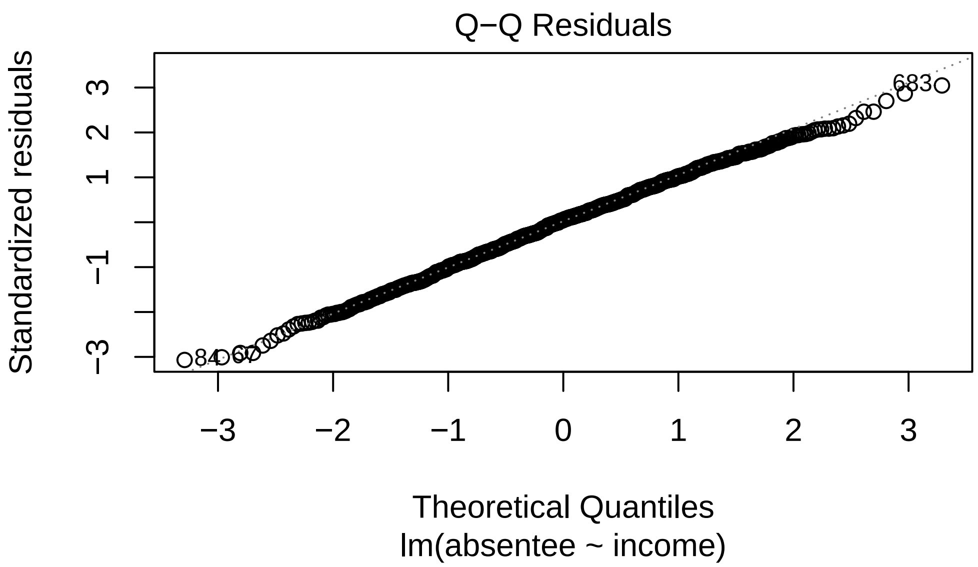
Note

Yes, there are some unusually large outliers. All the plots show that observations 67, 84, and 683 are outliers. Observation 84 has unusually high leverage, but does not appear to be influential enough to change the model. All of the high residual observations are still within Cook's threshold.

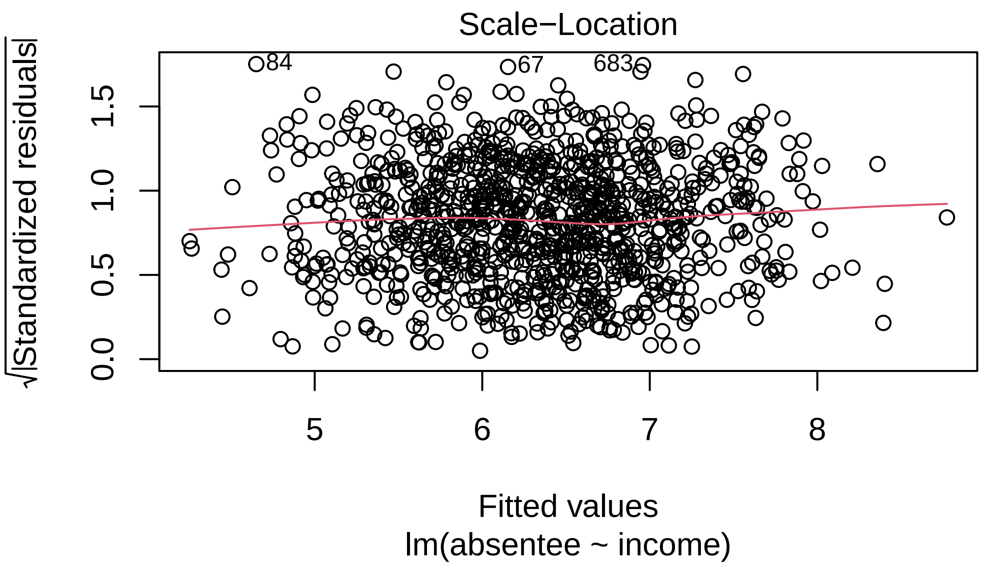
```
plot(model1)
```

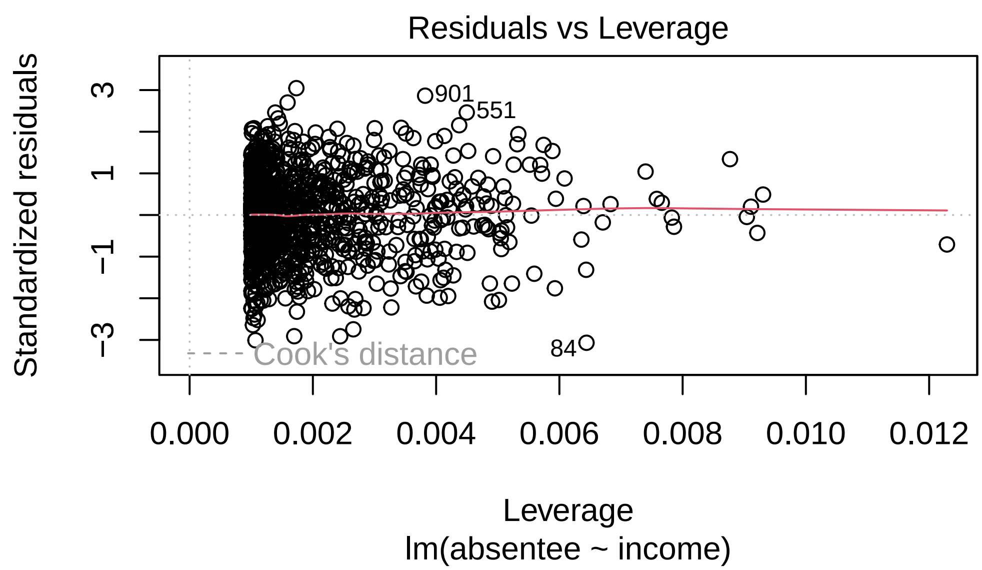


Residuals vs Fitted
Residuals
Fitted values
lm(absentee ~ income)
683
84
67



Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(absentee ~ income)
683
84
67


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
84
67
683
5
6
7
8
Fitted values
lm(absentee ~ income)


Residuals vs Leverage
Standardized residuals
−3
−1
1
3
901
551
84
Cook's distance
0.000
0.002
0.004
0.006
0.008
0.010
0.012
Leverage
lm(absentee ~ income)

2.4 - 1 point

Compute the matrix of correlations between all five selected variables using the function cor().

Note

The matrix is provided below.

```
### Your code here.
cor(subset2)
```

```
            absentee        GPA     income      change         IEP
absentee  1.00000000 -0.1171887 -0.4772694 -0.01927614  0.26486193
GPA      -0.11718872  1.0000000  0.6158060 -0.23945780  0.26190757
income   -0.47726940  0.6158060  1.0000000 -0.14637093  0.55097584
change   -0.01927614 -0.2394578 -0.1463709  1.00000000 -0.05079061
IEP       0.26486193  0.2619076  0.5509758 -0.05079061  1.00000000
```

2.5 - 2 points

Identify two quantitative variables that are plausible confounders for the relationship explored above in 2.1. Describe your reasoning for choosing these variables as potential confounders.

Note

I would choose GPA and IEP as potential confounders. Both variables have somewhat moderate/strong relationships with the variable of interest, income, and also with the outcome variable. And, both could reasonably explain both income and absentee rates.
For example, GPA could have a relationship with income because higher-income families often provide more resources for their kids to help boost grades. It could also affect absentee rates becuase high GPA students may feel more motivated to attend class, and vice versa.
IEP rate could have a relationship with income because perhaps lower income families have less access to to the types of systems/resources that identify disabilities early, meaning that richer families will have a higher rate of IEPs. And if one has an IEP, they are subject to pull outs, medical appointments, and other related items, which could have a relationship with absentee rates.

```
### No code needed for this section.
```

2.6 - 2 points

Use the lm() function to perform a multiple linear regression with `absentee` as the response and with the same variable from 2.1 and the two plausible confounders from 2.6 as predictors. Use the summary() function to print the results. Comment on the output. For example:

- Is there a relationship between the predictors and the response?
- Which predictors appear to have a statistically significant relationship to the response?
- How did the coefficient for the variable from 2.1 change? Does this change constitute evidence that one or more predictor from 2.6 was a confounder?
- Calculate the variance inflation factor for each of the three predictors. Should any of these predictors be dropped from the model?

Note

Yes, all three predictors have a relationship with the response, which can be seen by the tiny p-values, meaning all are statistically significant.
The coefficient got larger, meaning that at least one of the confounders I added were suppressing the true effect, but likely both since they are both statistically significant.
Given that the VIFs for each are below 5, I'll keep them in the model. There is no evidence of significant multicollinearity.

```r
### Your code here.
# multiple regression
model2 <- lm(absentee ~ income + GPA + IEP, data = subset2)
summary(model2)
```

```

Call:
lm(formula = absentee ~ income + GPA + IEP, data = subset2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.96807 -0.57705  0.02727  0.54801  2.81835 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.705e+00  1.753e-01   26.85   <2e-16 ***
income      -1.240e-04  2.671e-06  -46.43   <2e-16 ***
GPA          1.730e+00  9.678e-02   17.87   <2e-16 ***
IEP          3.765e-01  9.552e-03   39.41   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.8079 on 994 degrees of freedom
Multiple R-squared:  0.7183,	Adjusted R-squared:  0.7175 
F-statistic:   845 on 3 and 994 DF,  p-value: < 2.2e-16
```

```r
#vif
# Install and load car package if needed
# install.packages("car")
library(car)
```

```
Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:dplyr':

    recode
```

```r
# Calculate VIF
vif(model2)
```

```
  income      GPA      IEP 
2.184649 1.633496 1.456073 
```

2.7 - 2 points

Provide a one-paragraph, high-level, non-technical summary of your analysis that's targeted toward a specific stakeholder. Identify an interest the stakeholder might have in your results and connect the results to a recommended action they could take in pursuit of this interest.

Note

There is a strong relationship between low-income classrooms and high absentee rates, even when controlling for GPA and IEP rates. In fact, controlling for GPA and IEP only increases the strength of class income being a predictor for high absentee rates. The prediction effects of GPA and IEP are not insignificant, however, and they are part of the puzzle. This means that low income classrooms also have students who are lower-performing and who have a higher rate of IEPs. My recommendation is to specifically target low income classrooms with resources that 1) encourage and reward high attendance, 2) resources for special education students, and 3) provide overall support for students who may not have an IEP but are struggling with academic performance anyway.

```
### Your code here.
```

Reflection on LLM Usage

Transparency is an important objective in data analytics. Use this section to describe whether and how you used an LLM to help you complete this problem set.

Your Answer Here

My usage of LLMs was mostly for recalling and constructing specific commands in R. The class I took in R previously did not cover all of the functions in this assignment, though my Stata-based classes did. So, it was convenient to pull up those commands with an LLM. I double checked a few of my responses to make sure they made sense, but did the interpretation myself first before checking. I believe in all cases my logic matched with what the LLM thought the answer should be. I also tried to re-type the commands even though I could copy and paste, 1) because the instructions forbade copy and pasting and 2) because it helps solidify those commands in my memory.