DS 3010 Homework 1

Instructions: Homework is to be submitted on Canvas by the deadline stated above. Please clearly print your name and student ID number on your HW.

Show your work (including calculations) to receive full credit. Please work hard to make your submission as readable as you possibly can - **this means no raw `R` output or code** (unless it is asked for specifically or needed for clarity).

Code should be submitted with your homework *as a separate file* (for example, a .R file, text file, Word file, or .Rmd are all acceptable). You should mark sections of the code that correspond to different homework problems using comments (e.g. `##### Problem 1 #####`).

Problem 1: `R` review

Load in the `Auto.data` from Module 'Introduction to DS 301'.

(a) How many rows and columns are in this dataset? Note that in supervised learning, we refer to the rows in the dataset as 'observations' and columns as 'variables' (or predictors/response).

Load in dataset and look at a snapshot:

```
auto = read.table('/Users/lchu/Dropbox/Teaching/DS301/Lectures/0-Introduction/Auto.data',
 header=TRUE, na.strings="?")
head(auto)
str(auto)
dim(auto)
```

397 observations and 9 variables

(b) Remove any observations with missing values in the dataset. Now how many rows and columns are in this dataset?

```
> auto=na.omit(auto)
> dim(auto)
[1] 392   9
```

392 observations and 9 variables

(c) Which columns are stored as `factors`? Define what `factors` are in `R`.

None are currently stored as `factors`. Factors are a data class used to work with categorical variables. In `R`, factors have a fixed and limited number of possible values, which we refer to as levels. An example of a variable that could be treated as a factor would be letter grade (i.e. A, B, C, D, and F).

(d) Should the variable `name` be encoded as a `factor`? Justify why or why not.

I would not recommend converting the variable `name` *to a* `factor`*. Although it appears categorical, the number of levels is quite large because each observation has a unique name. In this case, it makes more sense to leave the variable as a* `chr`*.*

(e) Print a snapshot of the first 10 observations in the dataset. Which `R` command did you use?

```
> head(auto)
  mpg cylinders displacement horsepower weight acceleration year
1  18         8          307        130   3504         12.0   70
2  15         8          350        165   3693         11.5   70
3  18         8          318        150   3436         11.0   70
4  16         8          304        150   3433         12.0   70
5  17         8          302        140   3449         10.5   70
6  15         8          429        198   4341         10.0   70
  origin                       name
1      1 chevrolet chevelle malibu
2      1         buick skylark 320
3      1        plymouth satellite
4      1             amc rebel sst
5      1               ford torino
6      1          ford galaxie 500
```

(f) Extract observations 10, 14, and 29 using one line of code. Print your code/output here.

```
> auto[c(10,14,29),]
   mpg cylinders displacement horsepower weight acceleration year
10  15         8          390        190   3850          8.5   70
14  14         8          455        225   3086         10.0   70
29   9         8          304        193   4732         18.5   70
   origin                    name
10      1      amc ambassador dpl
14      1 buick estate wagon (sw)
29      1                hi 1200d
```

(g) Extract the displacement and horsepower values for observations 10, 14, and 29 using one line of code. Print your code/output here.

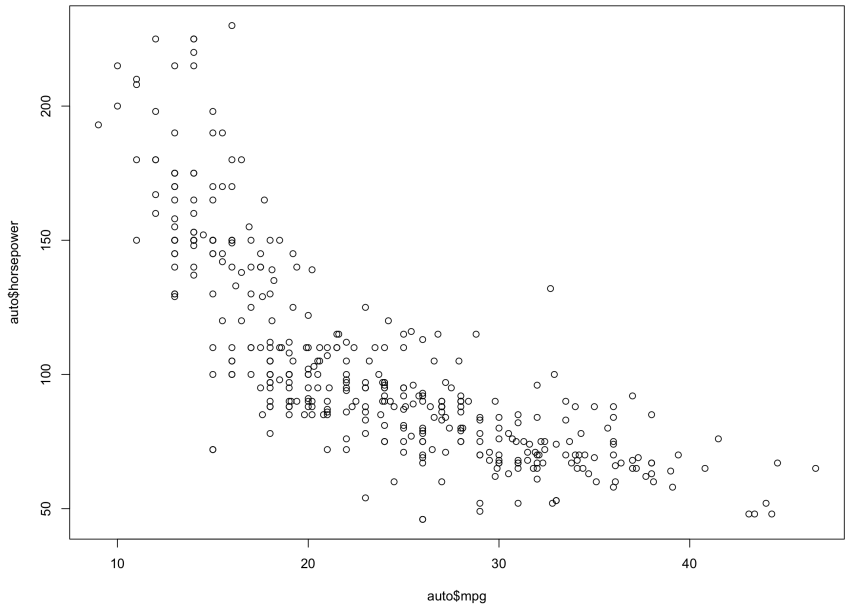
```
> auto[c(10,14,29),c(3,4)]
   displacement horsepower
10          390        190
14          455        225
29          304        193
```

(h) Find the average mpg for all observations in the dataset that have a horsepower less than 200. Can you compute that using one line of code?

```
> mean(auto$mpg[auto$horsepower<200])
[1] 23.75801
```

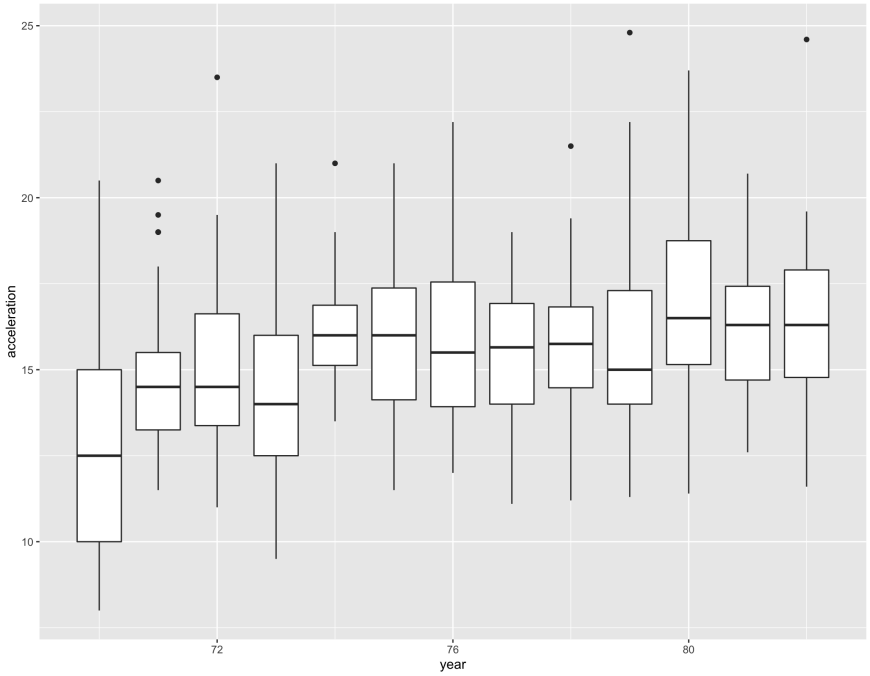
(i) What kind of plot would be appropriate to examine the relationship between mpg and horsepower? Show that plot here.

Since these are two continuous variables, we would use a scatterplot here.



(j) What kind of plot should be appropriate to examine the relationship between year and acceleration? Show that plot here. Note you should not be using the same type of plot as the previous question.

Since year takes on discrete values, a scatterplot would not be appropriate here. An alternatives would be using a boxplot grouped by year.



(k) Many functions in `R` are **vectorized**. Explain in plain language (to someone with no coding background) what that means.

A vectorized function is a function that automatically operates on all elements of a vector without needing to loop through and act on each individual element. This can improve computational efficiency and make code easier to read.

Problem 2: Multiple linear regression

For this problem, we will use the `Boston` data set which is part of the `ISLR2` package. To access the data set, install the `ISLR2` package and load it into your `R` session:

```
install.packages("ISLR2") #you only need to do this one time.
library(ISLR2) #you will need to do this every time you open a new R session.
```

To get a snapshot of the data, run `head(Boston)`. To find out more about the data set, we can type `?Boston`.

We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response (Y), and the other variables are the predictors (X).

(a) How many rows (n) are in the data set? How many columns are in the data set?

There are 506 observations or rows in the data set and 13 columns.

(b) What does the variable `lstat` represent? (Hint: check `?Boston`)

The variable `lstat` represents a percent of lower status in the population.

(c) Obtain the average per capita crime rate across all suburbs in the data set. Report that here.

```
mean(Boston$crim)
```

The average per capita crime rate is 3.613.

(d) Obtain the average crime rate only for those suburbs who are not near the Charles river (`chas ==0`) and those suburbs who are near the Charles river (`chas ==1`). Report both values here. Is it safer to be near or away from the Charles river?

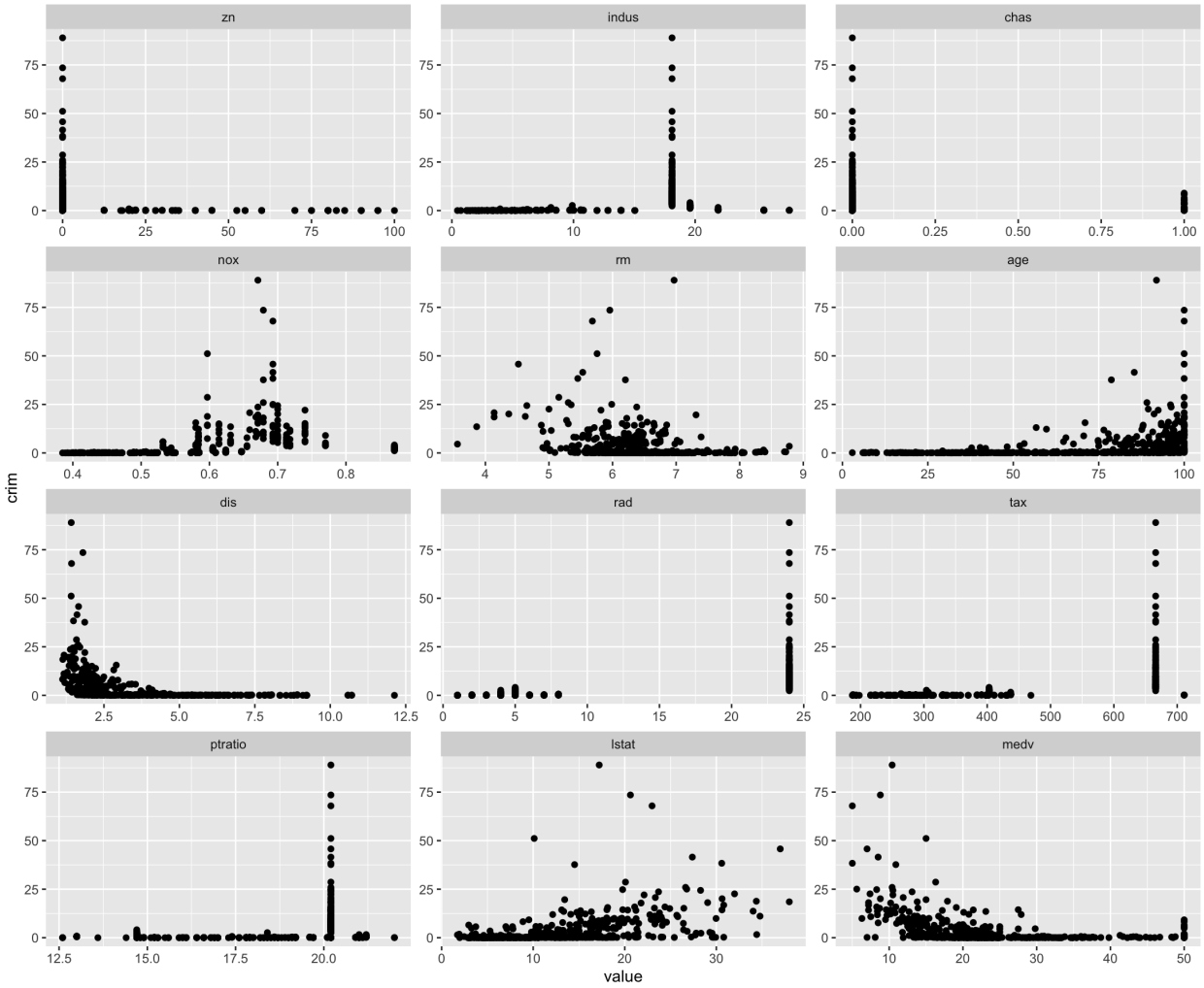
The average crime rate for those suburbs near the Charles river is 1.85 while the average crime rate for those not near the Charles river is 3.74. It seems that neighborhoods near the Charles river appear to be safer.

(e) Do any of the suburbs of Boston appear to have particularly high crime rates? Define what a 'high' crime rate is and provide some summary statistics on the crime rate.

Students could have answered this a number of different ways. At minimum, students should report the summary statistics from `crim` *or better yet, plot the variable.*

(f) Are any of the other predictors in the data set associated with per capita crime rate? Use your exploratory data analysis skills from DS 202 to uncover insights. Describe your findings.

At minimum, students should demonstrate an attempt to explore the data by reporting plots or summary statistics. They can discuss whether or not the associations are positive, negative, strong, weak, etc. Based on my plot (below), it seems `crim` *is associated with* `age, dis, rad, tax, ptratio`*, and* `medv`*.*



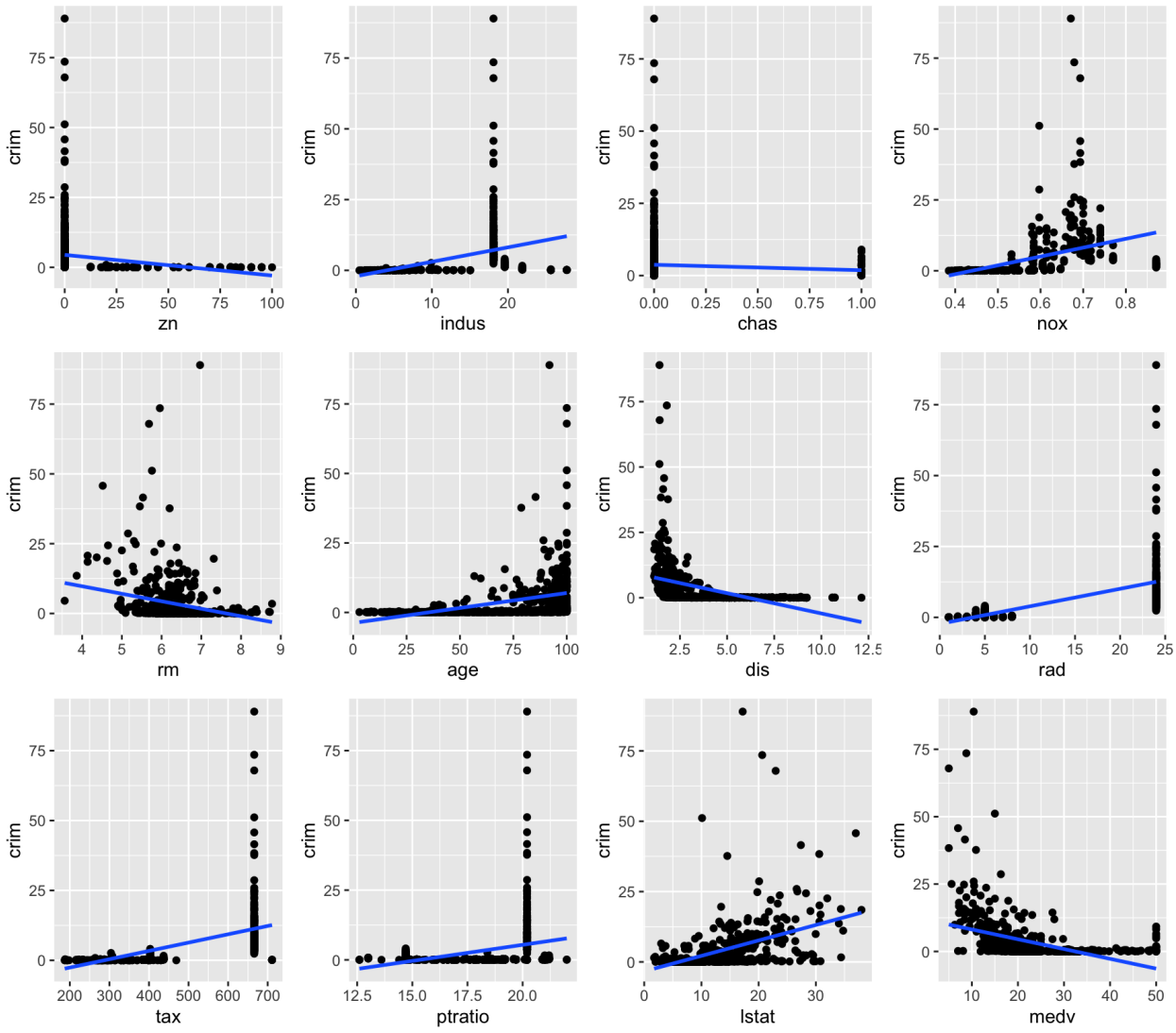
(g) Fit a simple linear regression model with `crim` as the response and `lstat` as the predictor. Describe your results. What are the estimated coefficients from this model? Report them here.
Note: a simple linear regression is just a regression model with a single predictor.

The simple linear regression model with `lstat` *and* `crim` *had an estimated intercept of -3.331 and estimated regression coefficient for* `lstat` *of 0.54*

(h) Repeat part (g) for *each predictor in the dataset*. That means for each predictor, fit a simple linear regression model to predict the response. Describe your results and **organize them in a table**. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

Variable	$\hat{\beta}_0$	$\hat{\beta}_1$	p-value for predictor
zn	4.45	-0.07	5.51×10^{-06}
indus	-2.06	0.50	2×10^{-16}
chas	3.74	-1.89	0.209
nox	-13.72	31.24	2×10^{-16}
rm	20.48	-2.68	6.35×10^{-7}
age	-3.77	0.10	2.85×10^{-16}
dis	9.49	-1.59	2×10^{-16}
rad	-2.28	0.61	2×10^{-16}
tax	-8.52	0.29	2×10^{-16}
ptratio	-17.64	1.15	2.94×10^{-11}
lstat	-3.33	0.54	2×10^{-16}
medv	11.79	-0.36	2×10^{-16}

Out of the 12 predictors, 11 of them showed significance (setting $\alpha = 0.05$) when running a simple linear regression model with `crim` *as the response variable. Only* `chas` *(surprisingly) did not show up as significant. Students could have looked at pairwise comparisons of crim and the significant predictors to see if they could observe a relatively strong association. Indeed it looks like* `chas` *only takes 2 values: 0 or 1, so a linear relationship is not appropriate here.*

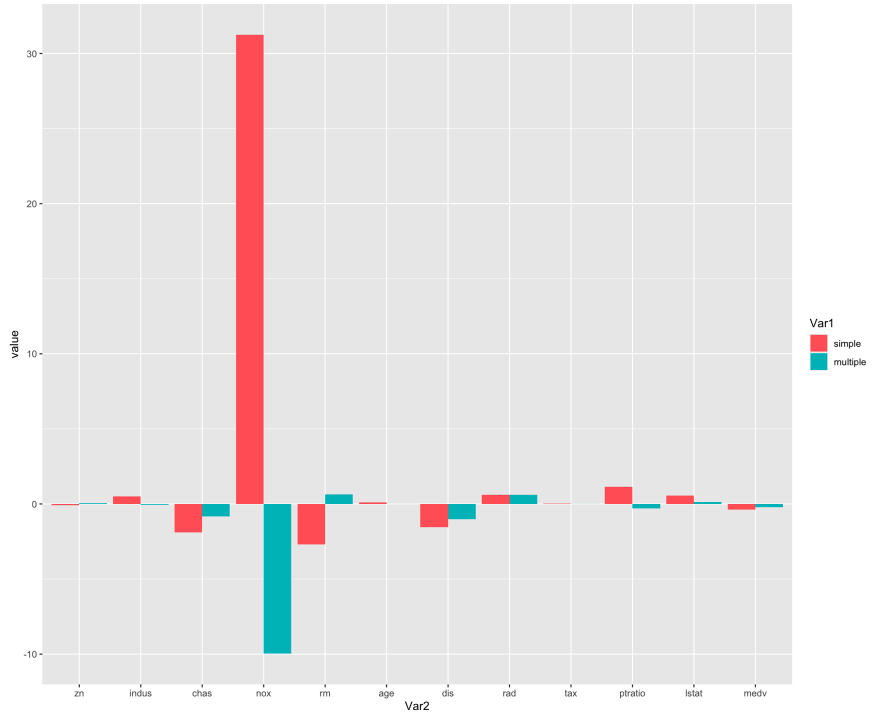


(i) Fit a multiple regression model to predict the response using all of the predictors. Summarize your results neatly in a table.

Students should neatly summarize their results. Raw `lm` *output should result in point deductions.*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	13.7783938	7.0818258	1.946	0.052271 .
zn	0.0457100	0.0187903	2.433	0.015344 *
indus	-0.0583501	0.0836351	-0.698	0.485709
chas	-0.8253776	1.1833963	-0.697	0.485841
nox	-9.9575865	5.2898242	-1.882	0.060370 .
rm	0.6289107	0.6070924	1.036	0.300738
age	-0.0008483	0.0179482	-0.047	0.962323
dis	-1.0122467	0.2824676	-3.584	0.000373 ***
rad	0.6124653	0.0875358	6.997	8.59e-12 ***
tax	-0.0037756	0.0051723	-0.730	0.465757
ptratio	-0.3040728	0.1863598	-1.632	0.103393
lstat	0.1388006	0.0757213	1.833	0.067398 .
medv	-0.2200564	0.0598240	-3.678	0.000261 ***

(j) How do your results from (h) compare to your results from (i)? Create a plot comparing the simple linear regression coefficients from (h) to the multiple regression coefficients from (i). Describe what you observe.



It's clear from the plot that the values for the estimated regression coefficients differ quite a bit. The most obvious standout is **`nox`***: it is 31 in the simple linear regression setting but -9.95 in the multiple linear regression setting. Not only is the magnitude different, but there is a complete sign flip!.*

(k) Explain why your results from (j) provide evidence that using many simple linear regression models is not sufficient compared to a multiple linear regression model. What information does a *multiple linear regression model* capture that many simple linear regression models cannot capture?

The reason why fitting many simple linear regressions is not sufficient is because simple linear regression models do not take into account the relationship ***between*** *predictors. They simply look at each predictor (as a standalone), without taking into account the dynamics between predictors and how these dynamics might affect the response* Y*. For example, the relationship of GPA* (X_1) *and salary* (Y) *is affected by your major* (X_2)*. A multiple linear regression model properly takes into account how the interactions between GPA and major might affect salary. Recall that the estimated regression coefficient* ($\hat{\beta}$) *is the effect of a predictor on the (average) of* Y*,* ***controlling for all other predictors. Students need to (1) provide a plot and (2) provide reasoning as to why the SLR is not sufficient to MLR in order to receive full credit.***

Problem 3: Interpreting Multiple Linear Regression Models

Suppose we have a data set with five predictors, X_1 = GPA, X_2 = IQ, and X_3 = Level (1 for College and 0 for High School). The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit a multiple linear regression model on our data set and get $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$ and $\hat{\beta}_3 = 35$.

(a) Which answer is correct, and why?

i. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates.

ii. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.

iii. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.

iv. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates provided that the GPA is high enough.

(ii.) If IQ and GPA are fixed, the average salary is higher for college graduates than high school graduates (note that $\hat{\beta}_3 = 35$ is positive).

(b) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

Plug and chug into our estimated regression line:

$$\begin{aligned}\hat{Y} &= \hat{\beta}_0 + \hat{\beta}_1 * 4 + \hat{\beta}_2 * 110 + \hat{\beta}_3 * 1 \\ &= 50 + 20 * 4 + 0.07 * 110 + 35 * 1 = 172.7 \text{ } \textit{(in thousands of dollars)}\end{aligned}$$

(c) True or false: Since the coefficient of IQ is very small, the effect of IQ effect on salary is not very important. Justify your answer.

False. The magnitude of the coefficient does not determine its importance. The magnitude is only affected by the scale of the predictor. To assess the importance of a predictor in a model there are other metrics we can consider (that we'll discuss next week).

End of assignment.