IMT 573: Problem Set 7 - Regression

LEE CHEN HSIN

Due: Tuesday, Nov 30, 2021

Collaborators:

Instructions: Before beginning this assignment, please ensure you have access to R and RStudio.

1. Download the `problemset7.Rmd` file from Canvas. Open `problemset7.Rmd` in RStudio and supply your solutions to the assignment by editing `problemset7.Rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment. In particular, note that Stack Overflow is licenses as Creative Commons (CC-BY-SA). This means you have to attribute any code you refer from SO.
4. Partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. But please **DO NOT** submit pages and pages of hard-to-read code and attempts that is impossible to grade. That is, avoid redundancy. Remember that one of the key goals of a data scientist is to produce coherent reports that others can easily follow. Students are *strongly* encouraged to attempt each question and to document their reasoning process even if they cannot find the correct answer. If you would like to include R code to show this process, but it does not run without errors you can do so with the `eval=FALSE` option as follows:

```
a + b # these object dont' exist
# if you run this on its own it with give an error
```

6. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF`, rename the knitted PDF file to `ps6_ourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.
7. Collaboration is often fun and useful, but each student must turn in an individual write-up in their own words as well as code/work that is their own. Regardless of whether you work with others, what you turn in must be your own work; this includes code and interpretation of results. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.

Setup In this problem set you will need, at minimum, the following R packages.

```
# Load standard libraries
library(tidyverse)
library(dplyr)
library(MASS) # Modern applied statistics functions
```

Problem 1: Housing values in Boston

[TOTAL = 15pts]

In this problem we will continue using the Boston housing dataset. This dataset contains information about median house value for 506 neighborhoods in Boston, MA.

The variables in the data are:

- [crim] per capita crime rate by heighborhood.
- [zn] proportion of residential land zoned for lots over 25,000 sq.ft.
- [indus] proportion of non-retail business acres per neighborhood.
- [chas] Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
- [nox] nitrogen oxides concentration (parts per 10 million).
- [rm] average number of rooms per dwelling.
- [age] proportion of owner-occupied units built prior to 1940.
- [dis] weighted mean of distances to five Boston employment centres.
- [rad] index of accessibility to radial highways.
- [tax] full-value property-tax rate per $10,000.
- [ptratio] pupil-teacher ratio by neighborhood.
- [black] $1000(B_k - 0.63)^2$ where B_k is the proportion of blacks by neighborhood.
- [lstat] lower status of the population (percent).
- [medv] median value of owner-occupied homes in $1000s.

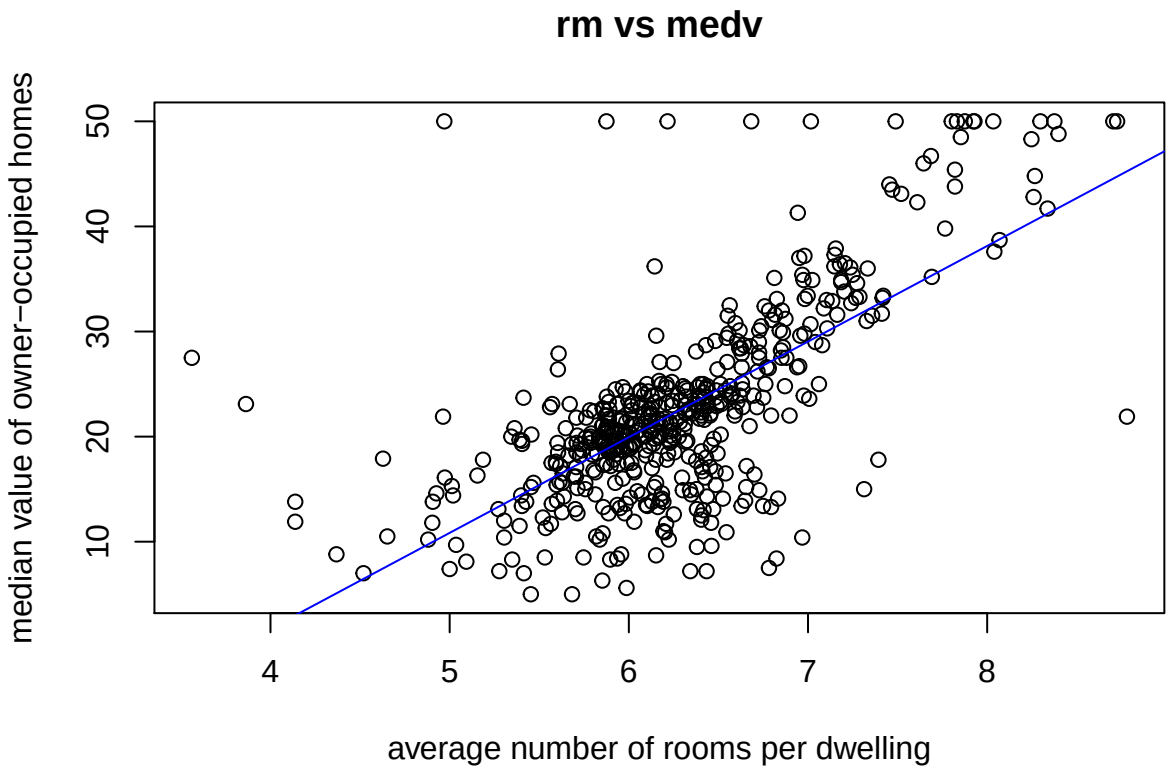
We are modeling the neighborhood median house price `medv`.

Part a)

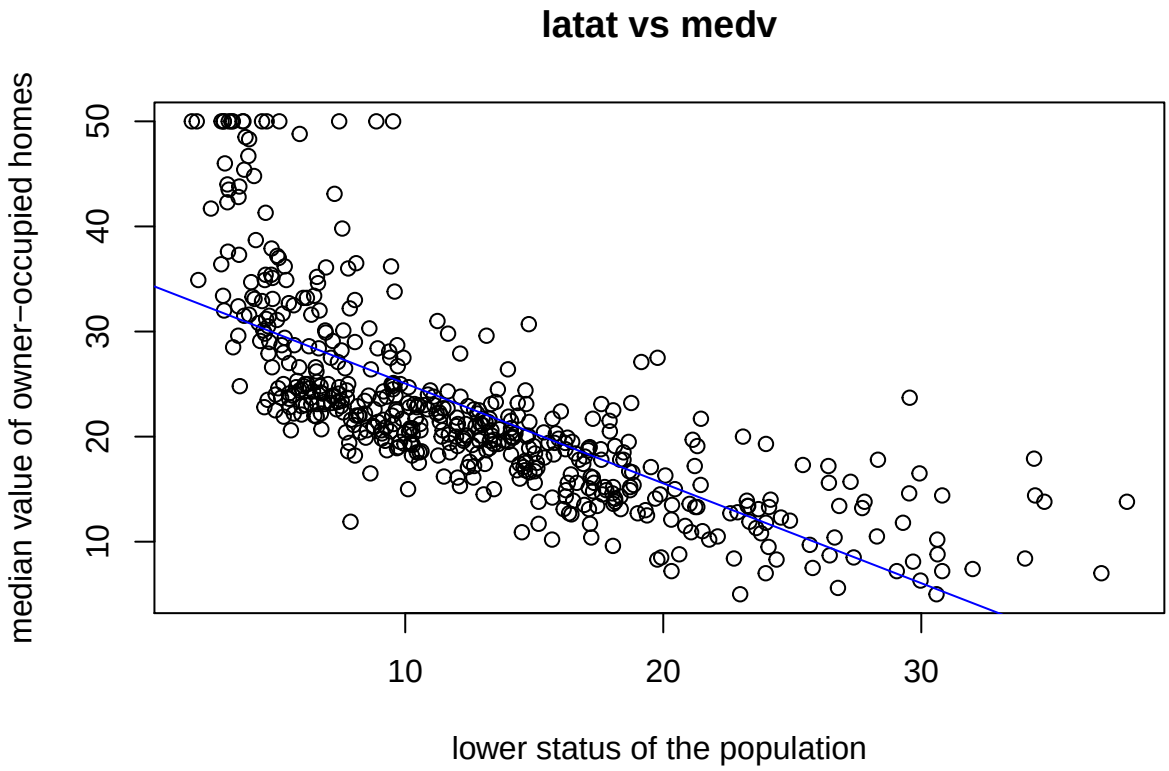
Use the following predictors: `rm`, `lstat`, and add an additional predictors of your choice, something that you consider might be interesting to analyze. Provide a rationale for your choice For each predictor do the following:

1. (3 X 1pts) Make a scatterplot that displays how `medv` is related to that predictor and add regression line to that plot. Comment on the result: do you see any relationship?

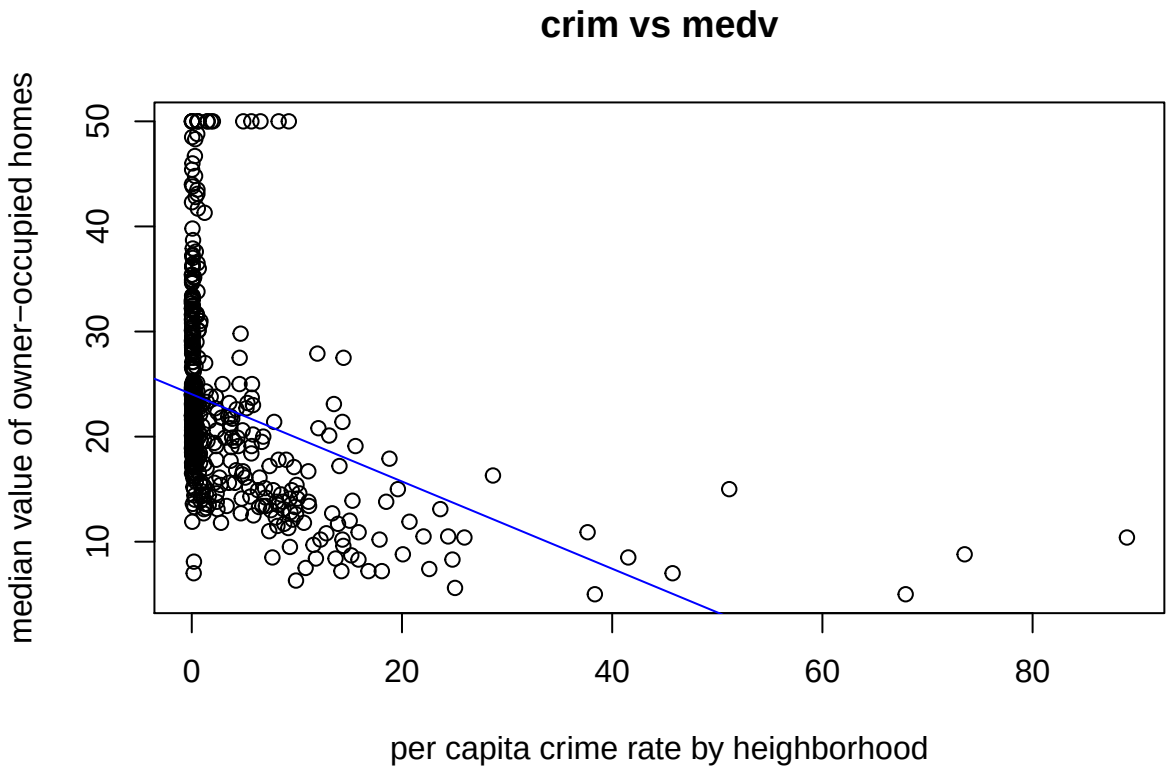
```
data(Boston)

plot(x = Boston$rm,y = Boston$medv,
   xlab = "average number of rooms per dwelling",
   ylab = "median value of owner-occupied homes",
   main = "rm vs medv"
)

abline(lm(medv ~ rm, data = Boston), col = "blue")
```




```
#positive relationship

plot(x = Boston$lstat,y = Boston$medv,
   xlab = "lower status of the population",
   ylab = "median value of owner-occupied homes",
   main = "latat vs medv"
)

abline(lm(medv ~ lstat, data = Boston), col = "blue")
```




```
#negative relationship

plot(x = Boston$crim,y = Boston$medv,
   xlab = "per capita crime rate by heighborhood",
   ylab = "median value of owner-occupied homes",
   main = "crim vs medv"
)

abline(lm(medv ~ crim, data = Boston), col = "blue")
```




```
#negative relationship
```

Hint: add regression line with `geom_smooth` *or* `abline` *methods*

2. (3 X 1pts) Now fit a simple linear regression model to predict the response. Show the regression output.

```
f1=lm(formula=medv~rm,data=Boston)
summary(f1)
```

```
##
## Call:
## lm(formula = medv ~ rm, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.346  -2.547   0.090   2.986  39.433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -34.671      2.650  -13.08   <2e-16 ***
## rm             9.102      0.419   21.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.616 on 504 degrees of freedom
## Multiple R-squared:  0.4835, Adjusted R-squared:  0.4825
## F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
f2=lm(formula=medv~lstat,data=Boston)
summary(f2)
```

```
##
## Call:
## lm(formula = medv ~ lstat, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.168  -3.990  -1.318   2.034  24.500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.55384    0.56263   61.41   <2e-16 ***
## lstat       -0.95005    0.03873  -24.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.216 on 504 degrees of freedom
## Multiple R-squared:  0.5441, Adjusted R-squared:  0.5432
## F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

3. (3 X 1pts) Interpret the slope. Explain why do you think you see (or don't see) the relationship on the figure or the model. Try to think about the possible social processes that make certain neighborhoods more or less expensive.

```
#in the regression of medv and rm, the p-value is smaller then 0.05 so there is
#a significant relationship between medv and rm. As a result, average number of
#rooms per dwelling make neighborhoods more expensive.
#And the slope of the medv and rm is  9.102, it means that one unit of rm can
#lead to the increase of one unit of medv.

#in the regression of medv and lstat, the p-value is smaller then 0.05 so there
#is a significant relationship between medv and lstat. AS a result,
#lower status of the population make neighborhoods less expensive.
#And the slope of the medv and lstat is -0.95005, it means that one unit of
#lstat can lead to the decrease of one unit of medv.
```

Part b)

1. (3 X 2pts) Compare simple and multiple regression results: In PS6, question 4, you had built a kitchen-sink model by fitting a multiple regression model to predict the response using all of the predictors. Now compare the results for `rm`, `lstat` and `indus` across the multiple regression and the simple regressions that you just built. Interpret your results. Explain why do the values differ.

```
model1<-lm(medv ~ rm, data = Boston)
model2<-lm(medv~lstat, data = Boston)
model3<-lm(medv~indus, data = Boston)

summary(model1)
```

```
##
## Call:
## lm(formula = medv ~ rm, data = Boston)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -23.346  -2.547   0.090   2.986  39.433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -34.671      2.650  -13.08   <2e-16 ***
## rm             9.102      0.419   21.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.616 on 504 degrees of freedom
## Multiple R-squared:  0.4835, Adjusted R-squared:  0.4825
## F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = medv ~ lstat, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.168  -3.990  -1.318   2.034  24.500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.55384    0.56263   61.41   <2e-16 ***
## lstat       -0.95005    0.03873  -24.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.216 on 504 degrees of freedom
## Multiple R-squared:  0.5441, Adjusted R-squared:  0.5432
## F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
summary(model3)
```

```
##
## Call:
## lm(formula = medv ~ indus, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.017  -4.917  -1.457   3.180  32.943
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.75490    0.68345   43.54   <2e-16 ***
## indus       -0.64849    0.05226  -12.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.057 on 504 degrees of freedom
## Multiple R-squared:  0.234,  Adjusted R-squared:  0.2325
## F-statistic:   154 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
#in the simple liner regression model,the p-value of rm,lstat and indus are
#smaller than 0.5 so there is a significant relationship between them.

model5 <- lm(medv ~ rm+lstat+indus, data = Boston)
summary(model5)
```

```
##
## Call:
## lm(formula = medv ~ rm + lstat + indus, data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.3179  -3.5006  -0.9387   2.0762  28.8816
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.96865    3.18168  -0.304    0.761
## rm           5.07379    0.44428  11.420   <2e-16 ***
## lstat       -0.60671    0.05046 -12.025   <2e-16 ***
## indus       -0.06364    0.04506  -1.412    0.159
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.535 on 502 degrees of freedom
## Multiple R-squared:   0.64,  Adjusted R-squared:  0.6378
## F-statistic: 297.5 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
#in the multiple liner regression model, the p-value of rm and lstat are smaller
#than 0.5 so there are  significant relationship between them. However, the
#p-value of indus is larger than 0.5 so there is not a significant relationship
#between them. The reasons of why the values differ may because of that
#variables will affect each other so it leads to the p-value of indus be increased.
```

Problem 2: Neighborhoods and Social Capital

Vanholm & monaghan 2020 paper analyze how evictions influence social capital across neighborhoods. If you'd like, you can checkout the paper pdf on Canvas. They proxy social capital with number of 311 calls. These are little bit like 911 emergency calls, just for non-urgent purposes (such as garbage or potholes on street). They estimate the model in the form:

$$311calls_i = \beta_0 + \beta_1.evictions_i + \beta_2.demographics_i + \beta_3.urbanCharacter_i + \epsilon$$

Here `evictions` is the number of evictions in neighborhood i, `demographics` is a vector of neighborhood demographic characteristics and `urban character` is a vector of urban environment specific variables. β_1 is the variable of interest, the effect of evictions on social capital. Note, the paper actually use logs of a number of variables. But, ignore logarithms when attempting this particular question. Their results are in Table 3 (page 8). Let us focus on model 3 (the column labeled as "(3)'') and ignore the other two models. We stress here to ignore the logs, assume the variables are not logged!

Now answer the following questions and for each write your interpretation:

[TOTAL = 20 pts]

(a) (4pt) Do neighborhoods with more evictions see more or less 311 calls? By how much?

```
#the third model slope is 0.048 and its p-value is smaller than 0.05
#so it proves that a correlation between evictions and 311 calls.
```

```
#neighborhoods that see more evictions have smaller numbers of 311 calls.

#slope is 0.048 so it means that the more eviction,
#the less 311 calls.
```

(b) (4pt) How is poverty rate associated with 311 calls? How much more (or less) calls there are in neighborhoods with 10% point more poverty?

```
#the third model slope is -0.14 and its p-value is smaller than 0.05 so
#it proves that a correlation between poverty rate and 311 calls.

#neighborhoods that see less poverty have higher 311 calls.


#slope is -0.14 so it means that the more calls,
#the less 10% point more poverty.
```

(c) (2pt) What can you tell about association of race (*white*) and calls?

```
#in the table3, we can see the third model slope is logged-0.038, and its
#p-value is smaller than 0.05 so it proves that a correlation between
#race and calls.


#slope is logged-0.038, which is smaller than 0.05
```

(d) (4pt) Is older median age associated with more or less 311 calls? At which level is this statistically significant?

```
#the third model slope is 0.0067 and its p-value is smaller
#than 0.01 so it is highly rejected the H0 hypothesis(the coefficient is 0) and it
#proves that the significantly important correlation between median
#age and 311 calls.

#Older median age can see less 311 calls.
```

(e) (2pt) The value for housing density is `-0.13`. What does this number mean?

```
#it means that the third model slope is -0.13 which is smaller
#than 0.01 so it proves that the significantly important correlation between
#housing density and 311 calls.
```

(f) (4pt) The omitted category for city is Austin, TX. Are there more or fewer calls in similar neighborhoods in Philadelphia, compared to Austin? By how much?

```
#the third model slope of Philadelphia is -0.56, and its p-value is smaller
#than 0.01 so it proves that the significantly important correlation between
#Philadelphia and 311 calls.


#So compared to austin, there are fewer calls in similar neighborhoods
#in Philadelphia by 0.56.
```

Problem 3: Extra Credit (10 pts)

Repeat the previous question, but now take into account the fact that some of the variables are logged. Respond the questions accordingly.

Problem 4: Price of Meal in Italian Restaurants in NYC

The Italian restaurants in New York City are legendary, and it's time to put your newly developed regression modeling skills to work to understand how they operate. What are the factors that contribute to the price of a meal at Italian restaurants in New York City?

```
a<-read.csv(file = '/Users/leechenhsin/Desktop/study@USA/07_UW_School/IMT573/nyc.csv')

summary(a)
```

```
##       Case                Restaurant      Price           Food
##  Min.   :  1.00   Amarone      :  1   Min.   :19.0   Min.   :16.0
##  1st Qu.: 42.75   Anche Vivolo:  1   1st Qu.:36.0   1st Qu.:19.0
##  Median : 84.50   Andiamo      :  1   Median :43.0   Median :20.5
##  Mean   : 84.50   Arno         :  1   Mean   :42.7   Mean   :20.6
##  3rd Qu.:126.25   Artusi       :  1   3rd Qu.:50.0   3rd Qu.:22.0
##  Max.   :168.00   Baci         :  1   Max.   :65.0   Max.   :25.0
##                   (Other)      :162
##      Decor          Service          East
##  Min.   : 6.00   Min.   :14.0   Min.   :0.000
##  1st Qu.:16.00   1st Qu.:18.0   1st Qu.:0.000
##  Median :18.00   Median :20.0   Median :1.000
##  Mean   :17.69   Mean   :19.4   Mean   :0.631
##  3rd Qu.:19.00   3rd Qu.:21.0   3rd Qu.:1.000
##  Max.   :25.00   Max.   :24.0   Max.   :1.000
##
```

```
#restaurant records the restaurant name
#price records the US dollars of a meal
#food records the food Zagat rating, scale 1-30
#decor records the decoration Zagat rating, scale 1-30
#services records the service Zagat rating, scale 1-30
#east records whether the restauratn is located east or west of Fifth Avenue
```

You will need to address this question with a series of multiple regression models. The Zagat guide is an influential review of restaurants. You will be looking at the numeric reviews posted on the Zagat review. Each restaurant is rated on a scale of 0 to 30 for the quality of its food, decor, and service.The data comes in the form of Zagat reviews from 168 Italian restaurants in New York City from 2001.

Part a)

[TOTAL = 17 pts]

1. You plan to visit an Italian restaurant for lunch. You check the Zagat review for three different restaurants that your colleagues have been suggesting and you find that Zagat has rated the quality of food for those restaurant as 20, 25, 35. What's your best estimate of the price of a meal that you would need to pay for lunch at each of these restaurants? *Hint: Before coming up with your best estimate you need to build the model and interpret your results and also explain the choice of your model* (pts: 1 build + 1 explain choice + 4 interpret + 6 estimate)

```
model <- lm(Price ~Food, data = a)
model
```

```
##
## Call:
## lm(formula = Price ~ Food, data = a)
##
## Coefficients:
```

```
## (Intercept)         Food
##     -17.832        2.939
```

```
predict(model,newdata=data.frame(Food=20))
```

```
##        1
## 40.94705
```

```
predict(model,newdata=data.frame(Food=25))
```

```
##        1
## 55.64185
```

```
predict(model,newdata=data.frame(Food=35))
```

```
##        1
## 85.03144
```

```
#I choose liner regression model to use food to predict the relationship with
#price. This is because there is only one variable, food to be chosen for
#prediction of price. By selecting different quality of food for those
#restaurant, 20,25 and 35, I can make estimate of the price of a meal.
```

2. Your office offers you a $100 reimbursement coupon for your lunch that you are only allowed to use as much as possible in one go at a lunch meal. Given what you know about the relationship between food quality and price, the three restaurant suggestions, and that you need to provide at least 15% tip for your meal, which restaurant would you pick and why to stay within budget? (5 pts)

```
price=predict(model,newdata=data.frame(Food=35))
price_new=price*1.15
price_new
```

```
##        1
## 97.78616
```

```
#i will pick the restaurant of FELIDIA since its food quality if the highest
#among all the restaurant.
```

Part b)

[TOTAL = 24 pts]

1. (2pts + 2pts + 4 pts build and interpret.) Based on your knowledge of the restaurant industry, do you think that the quality of the food as well as the service in a restaurant are important determinants of the price of a meal at that restaurant? How will you prove your intuition through regression modeling? Build and interpret model output.

```
model1 <- lm(Price ~ Food, data = a)
summary(model1)
```

```
##
## Call:
## lm(formula = Price ~ Food, data = a)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.8860  -3.9470   0.2056   4.2513  26.9919
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -17.8321     5.8631  -3.041  0.00274 **
## Food          2.9390     0.2834  10.371  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.261 on 166 degrees of freedom
## Multiple R-squared:  0.3932, Adjusted R-squared:  0.3895
## F-statistic: 107.6 on 1 and 166 DF,  p-value: < 2.2e-16
```

```
model2 <- lm(Price ~ Service, data = a)
summary(model2)
```

```
##
## Call:
## lm(formula = Price ~ Service, data = a)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.6646  -4.7540  -0.2093   4.3368  26.2460
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.9778     5.1093  -2.344   0.0202 *
## Service       2.8184     0.2618  10.764   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.153 on 166 degrees of freedom
## Multiple R-squared:  0.4111, Adjusted R-squared:  0.4075
## F-statistic: 115.9 on 1 and 166 DF,  p-value: < 2.2e-16
```

```
#Based on my knowledge of the restaurant industry, I think that the quality of
#the food and service in a restaurant are important determinants of the price
#of a meal.
#In model1 and model2, the p-value of them are all lower than 0.05 so
#there is a significant liner regression with the price of food and service.
```

2. (3pts + 2pts for model + 5 justify choice.) Another important consideration in dining out is the decor. Are people willing to pay more for better restaurant decor? More interestingly, are people willing to pay more for fancy Decor, irrespective of the quality of food? How much more? Now answer this question with an appropriate model. Justify the choice of your model and variables that go into the model.

```
model3 <- lm(Price ~ Decor, data = a)
summary(model3)
```

```
##
## Call:
## lm(formula = Price ~ Decor, data = a)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9578  -4.4862  -0.4673   4.0422  18.5138
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.362      3.292  -0.414     0.68
```

```
## Decor          2.490      0.184  13.537   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.426 on 166 degrees of freedom
## Multiple R-squared:  0.5247, Adjusted R-squared:  0.5218
## F-statistic: 183.2 on 1 and 166 DF,  p-value: < 2.2e-16
```

```
#The perceived quality of decoration has a p-value 2e-16, which is smaller than
#0.05, so it means that quality of decoration has a strong correlation with the
#price of a meal.

model5 <- lm(Price ~ Decor+Food, data = a)
summary(model5)
```

```
##
## Call:
## lm(formula = Price ~ Decor + Food, data = a)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.945  -3.766  -0.153   3.701  18.757
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -24.5002     4.7230  -5.187 6.19e-07 ***
## Decor         1.8820     0.1919   9.810  < 2e-16 ***
## Food          1.6461     0.2615   6.294 2.68e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.788 on 165 degrees of freedom
## Multiple R-squared:  0.6167, Adjusted R-squared:  0.6121
## F-statistic: 132.7 on 2 and 165 DF,  p-value: < 2.2e-16
```

```
#Yes,based on the multiple liner regression model,the p-value of decoration and
#food are smaller than 0.05 so it means that people are still willing to pay
#more for fancy decoration irrespective of the quality of food.

#Fancy decoration and the quality of food have a correlation with 1.8820 and
#1.6461.
```

3. (3 X 2pts.) Among the three considerations of food quality, decor and service, compare their effects to determine what are important considerations.

```
model2 <- lm(Price ~ Decor+Food+Service, data = a)
summary(model2)
```

```
##
## Call:
## lm(formula = Price ~ Decor + Food + Service, data = a)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.8440  -3.7039  -0.1525   3.6218  19.0576
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -24.6409     4.7536  -5.184 6.33e-07 ***
## Decor         1.8473     0.2176   8.491 1.17e-14 ***
## Food          1.5556     0.3731   4.170 4.93e-05 ***
## Service       0.1350     0.3957   0.341    0.733
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.803 on 164 degrees of freedom
## Multiple R-squared:  0.617,  Adjusted R-squared:   0.61
## F-statistic: 88.06 on 3 and 164 DF,  p-value: < 2.2e-16
```

```
#According to the multiple liner regression model,the p-value of decoration and
#food are smaller than 0.05 so they have important regression with the price.
#However,the p-value of service is larger than 0.05 so it doesn't have important
#regression with the price.
```

4. [**EXTRA CREDIT** (2pts + 2pts model + 4 pts interpret)]. For the restaurants in NYC, does food quality affect the price of a meal for a certain level of service? You need to explain the choice of your model, build the model, and interpret the results. *Hint: Think interactions*

```
model2 <- lm(Price ~ Food*Service, data = a)
summary(model2)
```

```
##
## Call:
## lm(formula = Price ~ Food * Service, data = a)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.6653  -4.5058   0.6272   3.6307  27.3283
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   59.9584    48.9550   1.225   0.2224
## Food          -2.4334     2.3967  -1.015   0.3115
## Service       -2.5773     2.6003  -0.991   0.3231
## Food:Service   0.2057     0.1233   1.668   0.0972 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.905 on 164 degrees of freedom
## Multiple R-squared:  0.4578, Adjusted R-squared:  0.4479
## F-statistic: 46.16 on 3 and 164 DF,  p-value: < 2.2e-16
```

```
#I choose a liner regression model with the interaction of food and service.
#In the interaction of the food quality and a certain level of service,
#the p-value of them is 0.0972 which is larger than 0.05 so it doesn't affect
#the price of a meal.
```

Problem 5: Mario Kart

[TOTAL = 24 pts]

In our regression labs you worked with the Mario Kart dataset. Recall that to load the data you had to use the `openintro` library. You should checkout the regression labs to figure out how to get the `mario_kart`

data.

(a) (2pts) Inspect the data using your usual inspect data functions to get a sense of how the variables are encoded and what values they typically take on. Describe the data and variables.

```
library(openintro)
```

```
## Loading required package: airports
```

```
## Loading required package: cherryblossom
```

```
## Loading required package: usdata
```

```
##
## Attaching package: 'openintro'
```

```
## The following objects are masked from 'package:MASS':
##
##     housing, mammals
```

```
summary(mariokart)
```

```
##        id                duration          n_bids        cond
##  Min.   :1.104e+11   Min.   : 1.000   Min.   : 1.00   new :59
##  1st Qu.:1.404e+11   1st Qu.: 1.000   1st Qu.:10.00   used:84
##  Median :2.205e+11   Median : 3.000   Median :14.00
##  Mean   :2.235e+11   Mean   : 3.769   Mean   :13.54
##  3rd Qu.:2.954e+11   3rd Qu.: 7.000   3rd Qu.:17.00
##  Max.   :4.001e+11   Max.   :10.000   Max.   :29.00
##
##     start_pr         ship_pr          total_pr            ship_sp
##  Min.   : 0.010   Min.   : 0.000   Min.   : 28.98   standard  :33
##  1st Qu.: 0.990   1st Qu.: 0.000   1st Qu.: 41.17   upsGround :31
##  Median : 1.000   Median : 3.000   Median : 46.50   priority  :23
##  Mean   : 8.777   Mean   : 3.144   Mean   : 49.88   firstClass:22
##  3rd Qu.:10.000   3rd Qu.: 4.000   3rd Qu.: 53.99   parcel    :16
##  Max.   :69.950   Max.   :25.510   Max.   :326.51   media     :14
##                                                     (Other)   : 4
##   seller_rate     stock_photo     wheels
##  Min.   :     0   no : 38     Min.   :0.000
##  1st Qu.:   109   yes:105     1st Qu.:0.000
##  Median :   820               Median :1.000
##  Mean   : 15898               Mean   :1.147
##  3rd Qu.:  4858               3rd Qu.:2.000
##  Max.   :270144               Max.   :4.000
##
##                                                        title
##  BRAND NEW NINTENDO MARIO KART WITH 2 WHEELS              :23
##  Mario Kart Wii (Wii)                                    :19
##  BRAND NEW NINTENDO 1 WII MARIO KART WITH 2 WHEELS +GAME: 8
##  Mario Kart Wii (GAME ONLY/NO WHEEL) - Nintendo Wii Game: 4
##  Mario Kart Wii (Wii) Nintendo Wii game *--WOW --AWESOME: 4
##  (Other)                                                 :84
##  NA's                                                    : 1
```

```
mariokart
```

```
## # A tibble: 143 x 12
##              id duration n_bids cond  start_pr ship_pr total_pr ship_sp seller_rate
```

```
##           <dbl>    <int>  <int> <fct>    <dbl>   <dbl>    <dbl> <fct>         <int>
##  1 150377422259        3     20 new       0.99    4        51.6 standa~        1580
##  2 260483376854        7     13 used      0.99    3.99     37.0 firstC~         365
##  3 320432342985        3     16 new       0.99    3.5      45.5 firstC~         998
##  4 280405224677        3     18 new       0.99    0        44   standa~           7
##  5 170392227765        1     20 new       0.01    0        71   media            820
##  6 360195157625        3     19 new       0.99    4        45   standa~      270144
##  7 120477729093        1     13 used      0.01    0        37.0 standa~        7284
##  8 300355501482        1     15 new       1       2.99     54.0 upsGro~        4858
##  9 200392065459        3     29 used      0.99    4        47   priori~          27
## 10 330364163424        7      8 used     20.0     4        50   firstC~         201
## # ... with 133 more rows, and 3 more variables: stock_photo <fct>,
## #   wheels <int>, title <fct>
```

```
#id means Auction ID assigned by Ebay
#duration means Auction length, in days
#n_bids means number of bids
#cond means Game condition, either new or used
#start_pr means Start price of the auction
#ship_pr means Shipping price
#total_pr means Total price, which equals the auction price plus the shipping
#price
#ship_sp means Shipping speed or method
#seller_rate means The seller's rating on Ebay( the number of positive ratings
#minus the number of negative ratings for the seller)
#stock_photo means Whether the auction feature photo was a stock photo or not
#wheels means Number of Wii wheels included in the auction
#title means The title of the auctions
```

(b) (2 + 2pts) Does the duration of the auction effect the price of a MarioKart? You need to build an a). appropriate model and b). interpret the results to answer the questions.

```
d1=lm(formula = total_pr ~ duration, data = mariokart)
summary(d1)
```

```
##
## Call:
## lm(formula = total_pr ~ duration, data = mariokart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.035  -8.116  -3.015   3.209 277.953
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  51.4246     3.8168   13.47   <2e-16 ***
## duration     -0.4097     0.8360   -0.49    0.625
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.76 on 141 degrees of freedom
## Multiple R-squared:  0.0017, Adjusted R-squared:  -0.00538
## F-statistic: 0.2402 on 1 and 141 DF,  p-value: 0.6249
```

```
#Since the p-value of duration is larger than 0.05 so it means that the duration
#and total price doesn't have a correction relationship. In other words,
```

```
#the duration of the auction not affect the price of a MarioKart.
```

(c) Experiment with other variables you see fit for this task, that is to see how they effect the price of MarioKart. Do other variables change your results in a major way? Did you have to remove any variables before fitting the model? Make sure that you build an 1). appropriate model while explaining your choice and 2). interpret the results to answer the questions. (pts: 2 model choice + 2 build + 3 interpret + 1 why/whynot remove)

```
d4=lm(formula = total_pr ~ ship_pr , data = mariokart)
summary(d4)
```

```
##
## Call:
## lm(formula = total_pr ~ ship_pr, data = mariokart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.786 -10.095  -3.684   6.781 179.622
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.2455     2.5379  14.282  < 2e-16 ***
## ship_pr       4.3372     0.5656   7.669 2.58e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.66 on 141 degrees of freedom
## Multiple R-squared:  0.2943, Adjusted R-squared:  0.2893
## F-statistic: 58.81 on 1 and 141 DF,  p-value: 2.583e-12
```

```
d6=lm(formula = total_pr ~ wheels, data = mariokart)
summary(d6)
```

```
##
## Call:
## lm(formula = total_pr ~ wheels, data = mariokart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.401  -6.411  -2.417   0.579 268.093
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   38.405      3.433  11.188  < 2e-16 ***
## wheels        10.006      2.411   4.151  5.7e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.34 on 141 degrees of freedom
## Multiple R-squared:  0.1089, Adjusted R-squared:  0.1026
## F-statistic: 17.23 on 1 and 141 DF,  p-value: 5.704e-05
```

```
#in ds4, the p-value of ship_pr is 2.58e-12, which is smaller than 0.05 so
#it means that ship_pr has an effect on the price of MarioKart.

#in ds6, the p-value of wheels is 5.7e-05, which is smaller than 0.05 so
```

```
#it means that  wheel has an effect on the price of MarioKart.
```

(d) Now let's check for interactions. Does duration effect the price of MarioKart based on the condition being new or used? You need to a). explain the choice of your model, b). build the model, c). interpret model results to answer this question. d). draw appropriate visual to confirm your interpretation. *Hint: You should think about plotting price versus duration, colored by condition* (pts: 2 choice + 2 build + 3 interpret + 3 visual)

```
d5=lm(formula = total_pr ~ duration*cond , data = mariokart)
summary(d5)
```

```
##
## Call:
## lm(formula = total_pr ~ duration * cond, data = mariokart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.102  -7.198  -2.323   2.002 276.427
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         58.268      5.032  11.580   <2e-16 ***
## duration            -1.966      1.653  -1.189   0.2363
## condused           -17.564      7.981  -2.201   0.0294 *
## duration:condused    3.305      2.014   1.641   0.1030
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.5 on 139 degrees of freedom
## Multiple R-squared:  0.03544,    Adjusted R-squared:  0.01463
## F-statistic: 1.703 on 3 and 139 DF,  p-value: 0.1693
```

```
#use liner regression model
#the interaction between condition and duration is not statistically significant
#since its p-value is 0.6249, which is not corresponded to any significant level
#so it means that the duration will not affect the price of MarioKart.

mariokart%>%
    group_by(mariokart$duration)%>%
       summarize(mean_total=mean(mariokart$total_pr))
```

```
## # A tibble: 5 x 2
##   `mariokart$duration` mean_total
##                  <int>      <dbl>
## 1                    1       49.9
## 2                    3       49.9
## 3                    5       49.9
## 4                    7       49.9
## 5                   10       49.9
```

```
mariokart$mean_total=mean(mariokart$total_pr)
str(mariokart)
```

```
## tibble [143 x 13] (S3: tbl_df/tbl/data.frame)
##  $ id         : num [1:143] 1.5e+11 2.6e+11 3.2e+11 2.8e+11 1.7e+11 ...
##  $ duration   : int [1:143] 3 7 3 3 1 3 1 1 3 7 ...
```

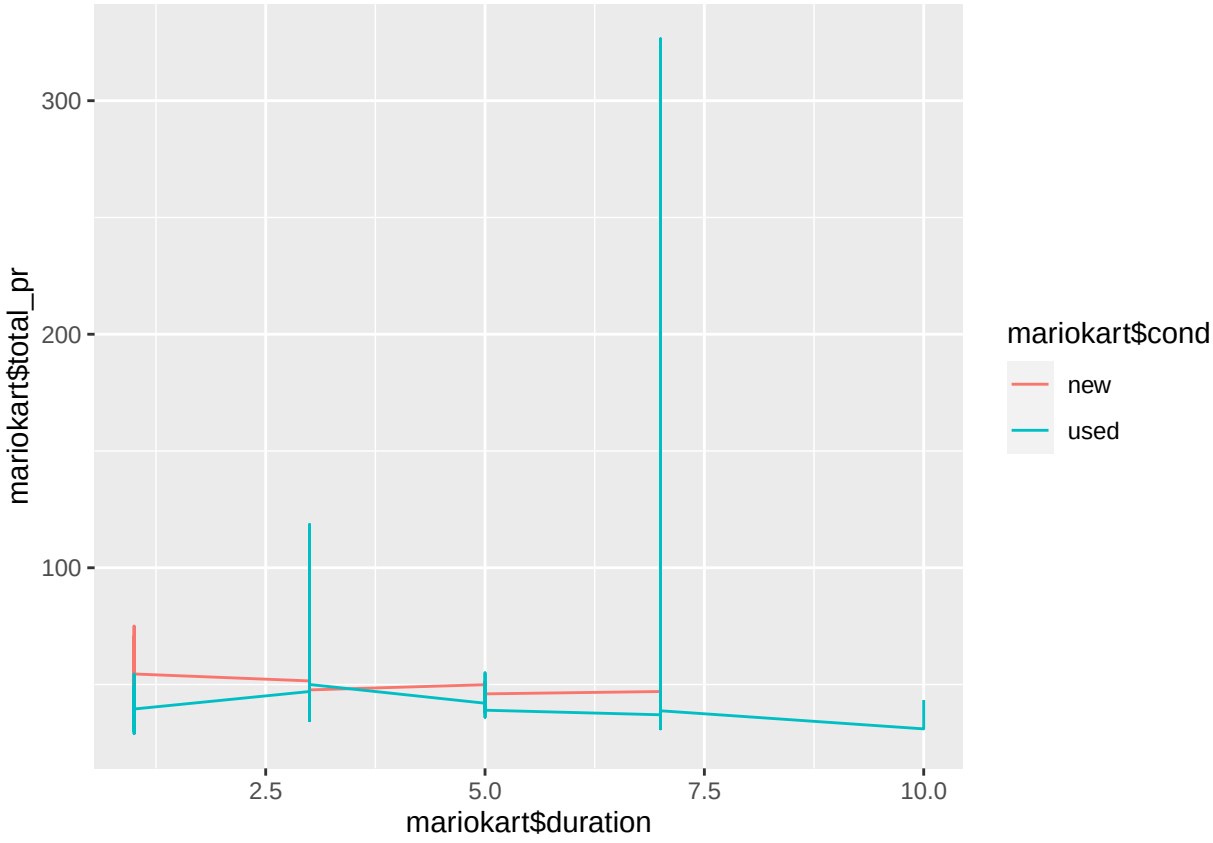
```
##  $ n_bids     : int [1:143] 20 13 16 18 20 19 13 15 29 8 ...
##  $ cond       : Factor w/ 2 levels "new","used": 1 2 1 1 1 1 2 1 2 2 ...
##  $ start_pr   : num [1:143] 0.99 0.99 0.99 0.99 0.01 ...
##  $ ship_pr    : num [1:143] 4 3.99 3.5 0 0 4 0 2.99 4 4 ...
##  $ total_pr   : num [1:143] 51.5 37 45.5 44 71 ...
##  $ ship_sp    : Factor w/ 8 levels "firstClass","media",..: 6 1 1 6 2 6 6 8 5 1 ...
##  $ seller_rate: int [1:143] 1580 365 998 7 820 270144 7284 4858 27 201 ...
##  $ stock_photo: Factor w/ 2 levels "no","yes": 2 2 1 2 2 2 2 2 2 1 ...
##  $ wheels     : int [1:143] 1 1 1 1 2 0 0 2 1 1 ...
##  $ title      : Factor w/ 80 levels " Mario Kart Wii with Wii Wheel for Wii (New)",..: 80 60 22 7 4 1
##  $ mean_total : num [1:143] 49.9 49.9 49.9 49.9 49.9 ...
```

```
plot1=ggplot(mariokart,aes(mariokart$duration,mariokart$total_pr, colour=mariokart$cond))+geom_line()
plot1
```

```
## Warning: Use of `mariokart$duration` is discouraged. Use `duration` instead.
```

```
## Warning: Use of `mariokart$total_pr` is discouraged. Use `total_pr` instead.
```

```
## Warning: Use of `mariokart$cond` is discouraged. Use `cond` instead.
```




```
#according to the plot, the interaction between condition and duration is not
#statistically significant.
```