IMT 573: Problem Set 7 - Regression

Jay Kuo

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 don't 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 Ames, Iowa

In this problem we will continue using the Ames housing dataset. This dataset contains information about home sales in Ames, IA. You used this dataset in the previous problem set.

As before, load the package and use `make_ames()` to store the dataset. Filter the data to only contain observations where the Sale_Condition was "Normal." As before, the variables of interest are: Lot_Frontage, Lot_Area, Bldg_Type, Overall_Qual, Overall_Cond, Year_Built, Gr_Liv_Area, TotRms_AbvGrd, Fireplaces, Garage_Cars, Garage_Area, Wood_Deck_SF, Total_Bsmt_SF, Full_Bath, Half_Bath, Year_Sold, and Sale_Price. You will be modeling Sale_Price.

Part a Use the following predictors: `Garage_Cars`, `Gr_Liv_Area`, `Lot_Area`, and add an additional explanatory variable of your choice - something that you think might be interesting to analyze. Provide rationale for your choice of additional explanatory variable. For each predictor do the following:

1. Make a scatterplot that displays how `Sale_Price` is related to that predictor and add a line of best fit to that plot. Comment on the results - do you see any relationship?

Hint: add the line with `geom_smooth` or `abline` methods

Answer

```
library(magrittr)

# import the data set
library(AmesHousing)
dt <- make_ames()

dt %<>%
  filter(Sale_Condition == "Normal") %>% #subset(Sale_Condition == "Normal")
  dplyr::select(Lot_Frontage, Lot_Area, Bldg_Type, Overall_Qual, Overall_Cond, Year_Built, Gr_Liv_Area,

# adjust some variables
dt_new <- dt

# Assign `Bldg_Type = 1` if Bldg_Type == "OneFam"
dt_new$Bldg_Type <- ifelse(dt$Bldg_Type == "OneFam", 1, 0)

# Assign `Overall_Qual = 1` if Overall_Qual > Average
dt_new$Overall_Qual <- ifelse(dt$Overall_Qual %in% c("Above_Average", "Good", "Very_Good", "Excellent",

# Assign `Overall_Cond = 1` if Overall_Cond > Average
dt_new$Overall_Cond <- ifelse(dt$Overall_Cond %in% c("Above_Average", "Good", "Very_Good", "Excellent",

# Assign `Year_Built = 1` if Year_Built >= 1971
dt_new$Year_Built <- ifelse(dt$Year_Built >= 1971, 1, 0)

# Assign `Year_Sold = 1` if Year_Sold >= 2008
dt_new$Year_Sold <- ifelse(dt$Year_Sold >= 2008, 1, 0)
```

Question

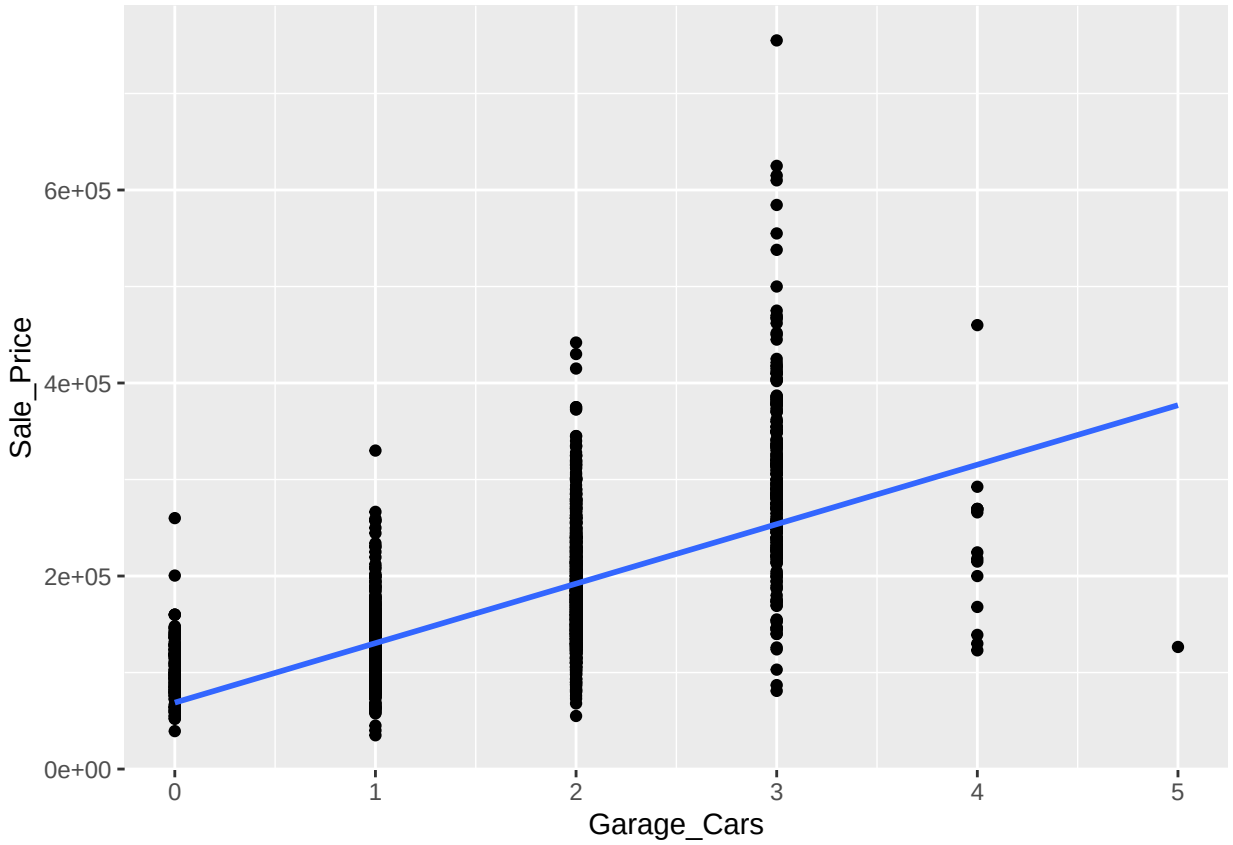
- Makes scatterplot for each predictor with regression line and comments on results.

Answer

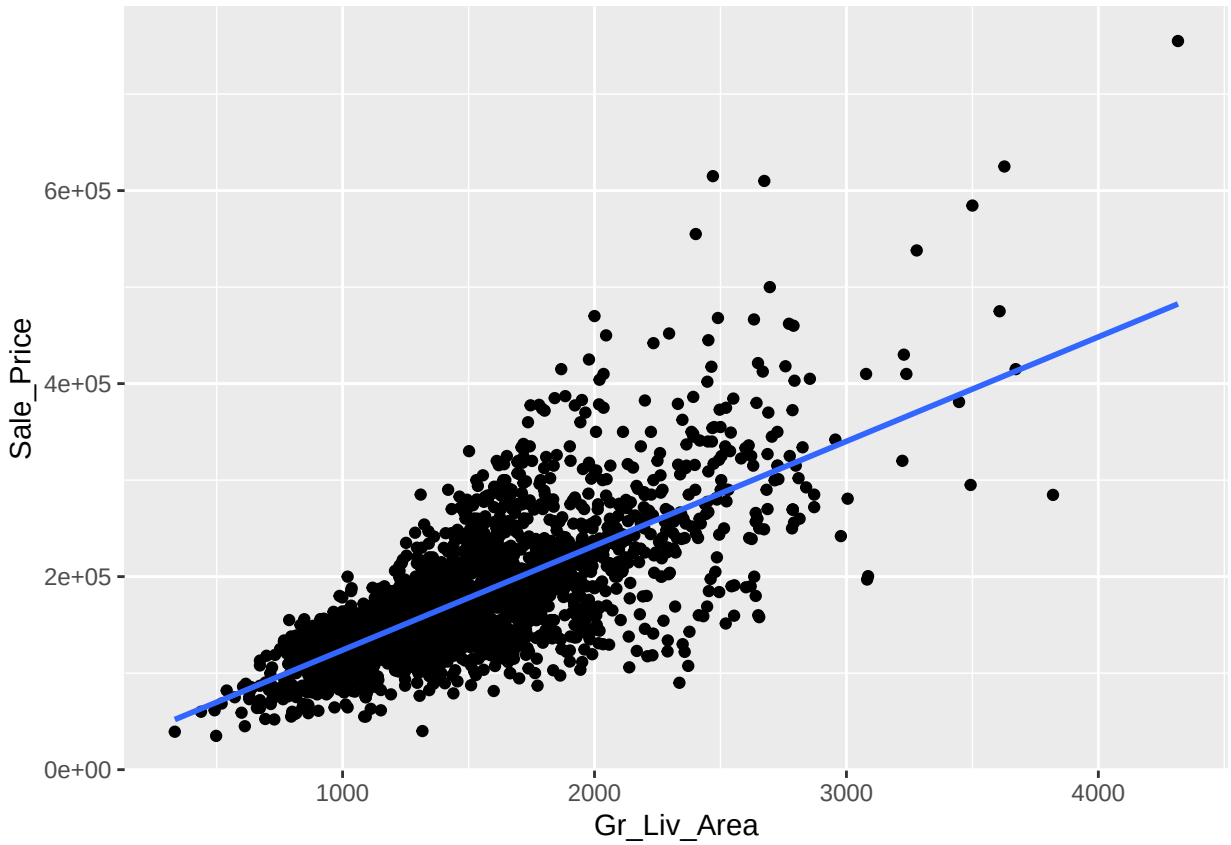
- `Garage_Cars`: positive relation with `Sale_Price`
- `Gr_Liv_Area`: positive relation with `Sale_Price`
- `Lot_Area`: positive relation with `Sale_Price`
- `Lot_Frontage`: positive relation with `Sale_Price`; the slope is closer to zero than other predictors

```
#`Garage_Cars`, `Gr_Liv_Area`, `Lot_Area` + `Lot_Frontage`
ggplot(data = dt_new, aes(x = Garage_Cars, y = Sale_Price)) +
```

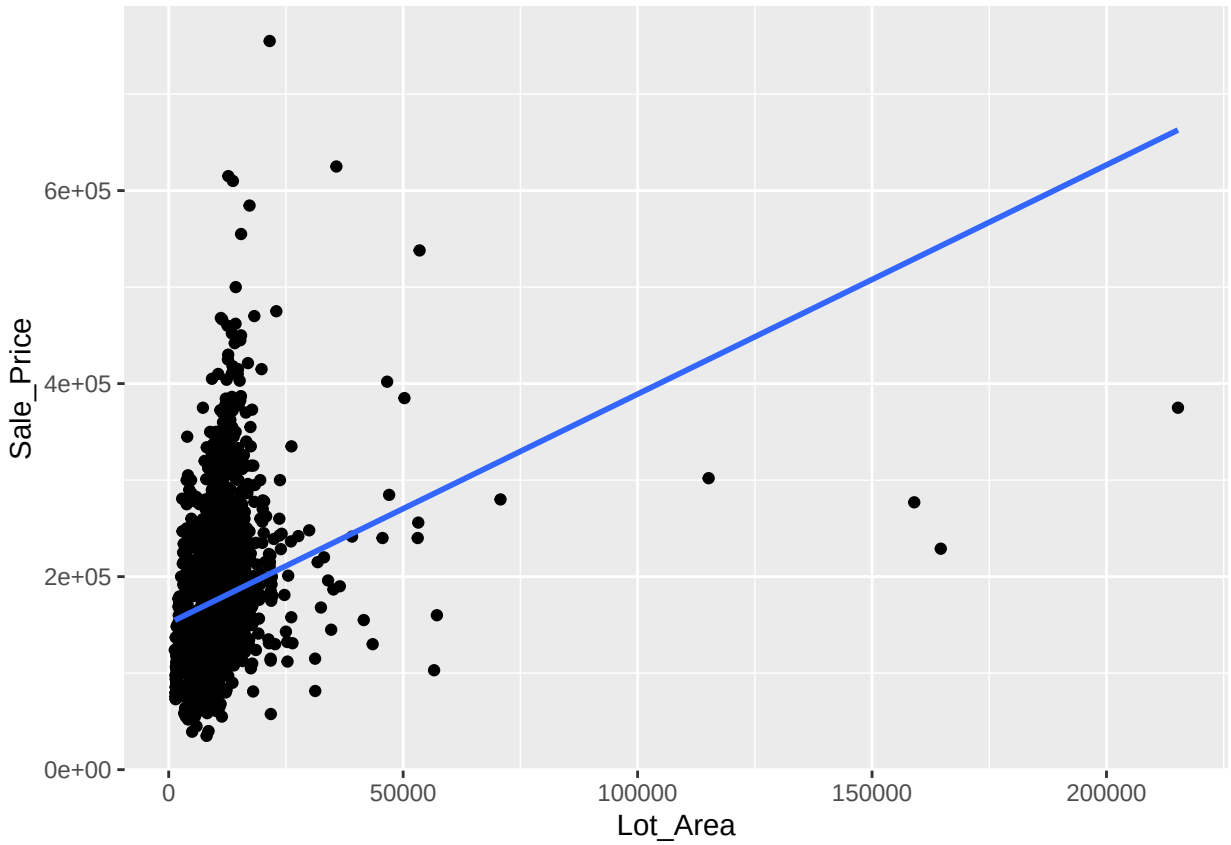
```
  geom_point() +
  geom_smooth(method = lm, se = FALSE)
```



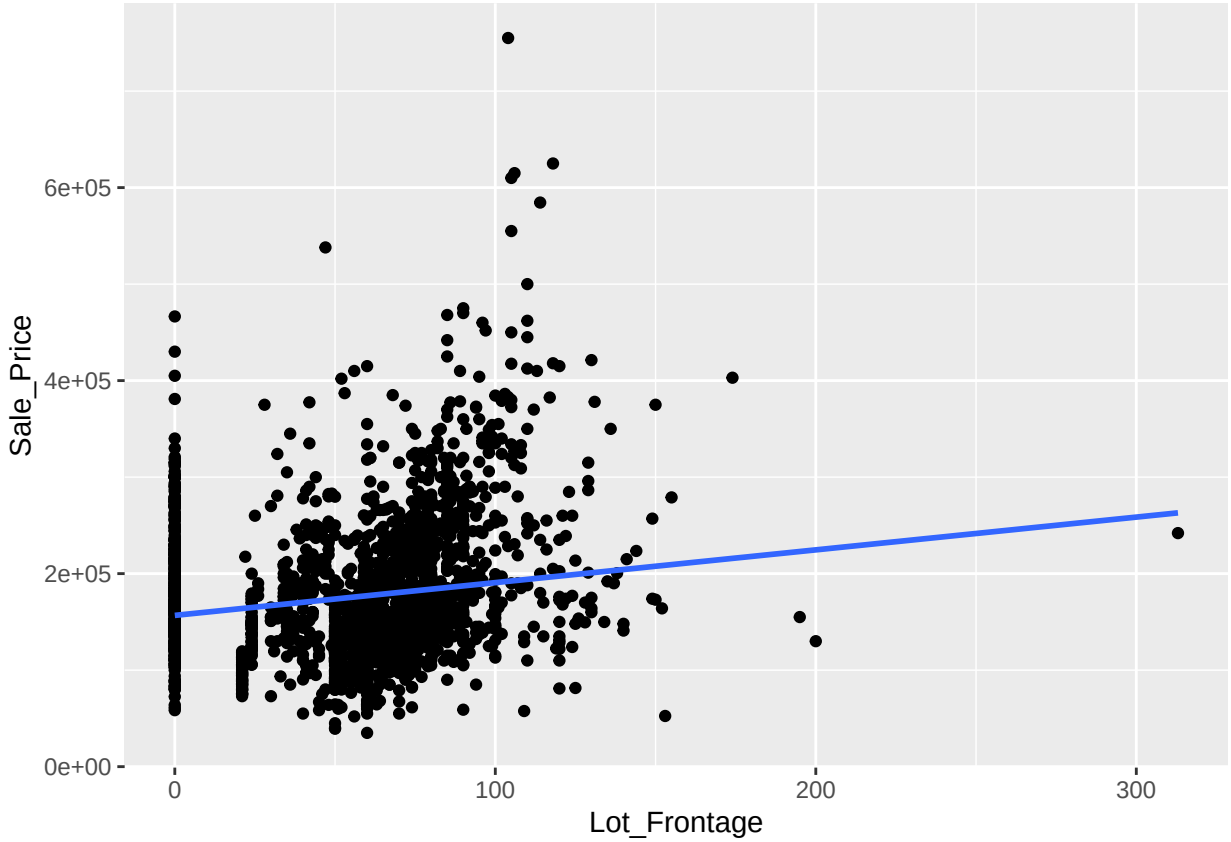

```
ggplot(data = dt_new, aes(x = Gr_Liv_Area, y = Sale_Price)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE)
```




```
ggplot(data = dt_new, aes(x = Lot_Area, y = Sale_Price)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE)
```




```
ggplot(data = dt_new, aes(x = Lot_Frontage, y = Sale_Price)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE)
```



2. Now fit a multiple regression to model the response using the three required predictors and your explanatory variable of choice. Show the regression output.

Question

- Fit a simple linear regression model and shows output for each variable.

Answer

```
HW7Q1A2 <- dt_new %>%
  lm(Sale_Price ~ Garage_Cars + Gr_Liv_Area + Lot_Area + Lot_Frontage,
             data = .)

summary(HW7Q1A2)
```

```
##
## Call:
## lm(formula = Sale_Price ~ Garage_Cars + Gr_Liv_Area + Lot_Area +
##     Lot_Frontage, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -204088  -21842   -1367   19117  310519
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.324e+04  3.088e+03  -4.288 1.87e-05 ***
## Garage_Cars   3.441e+04  1.348e+03  25.530  < 2e-16 ***
## Gr_Liv_Area   7.930e+01  2.072e+00  38.279  < 2e-16 ***
```

```
## Lot_Area      6.797e-01  1.080e-01   6.295 3.65e-10 ***
## Lot_Frontage  9.415e+01  2.587e+01   3.639 0.000279 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42070 on 2408 degrees of freedom
## Multiple R-squared:  0.6493, Adjusted R-squared:  0.6487
## F-statistic:  1115 on 4 and 2408 DF,  p-value: < 2.2e-16
```

3. Interpret the coefficients. Explain why you think you see (or don't see) the relationship on the figures or the model. Try to think about the possible processes that make certain neighborhoods more or less expensive.

Question

- Interprets intercept for each variable.

Answer

-Every intercept is expected and positively correlated.

- `Garage_Cars` (Size of garage in car capacity)
  - Other things being equal, if `Garage_Cars` increases by 1 car, `Sale_Price` is expected to increase by $\$3.4411337 \times 10^4$.
- `Gr_Liv_Area` (Above grade (ground) living area square feet)
  - Other things being equal, if `Gr_Liv_Area` increases by 1 square feet, `Sale_Price` is expected to increase by \$79.2967329.
- `Lot_Area` (Lot size in square feet)
  - Other things being equal, if `Lot_Area` increases by 1 square feet, `Sale_Price` is expected to increase by \$0.6796519.
- `Lot_Frontage` (Linear feet of street connected to property)
  - Other things being equal, if `Lot_Frontage` increases by 1 feet, `Sale_Price` is expected to increase by \$94.1468106.

Part b

1. In Problemset6, question 5, 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 `Garage_Cars`, `Gr_Liv_Area` and `Lot_Area` across the multiple regression and the simple regressions that you just built. Interpret your results. Explain why the values differ.

Question

- Explains differences in regression results across simple and multiple regression for rm, lstat, and indus.

Answer

- All three coefficients of multiple regression are larger than kitchen-sink model.

```
library(jtools); library(huxtable)

y <- c("Sale_Price")
controls <- colnames(dt)[-17]

HW6Q1B <- dt_new %>%
  lm(formula = as.formula(paste(y, paste(c(controls), collapse = " + "), sep = " ~ ")), data = .)

export_summs(HW7Q1A2, HW6Q1B,
             coefs = c("Garage_Cars", "Gr_Liv_Area", "Lot_Area", "Lot_Frontage"))
```

	Model 1	Model 2
Garage_Cars	34411.34 ***	5276.26 *
	(1347.86)	(2087.14)
Gr_Liv_Area	79.30 ***	73.98 ***
	(2.07)	(3.12)
Lot_Area	0.68 ***	0.44 ***
	(0.11)	(0.09)
Lot_Frontage	94.15 ***	113.16 ***
	(25.87)	(20.32)
N	2413	2413
R2	0.65	0.80

*** p < 0.001; ** p < 0.01; * p < 0.05.

Problem 2: Neighborhoods and Social Capital

A paper by Vanholm & Monaghan in 2020 analyzed how evictions influence social capital across neighborhoods. The paper is available in Canvas under readings. They proxy social capital with number of 311 calls. These are similar to 911 emergency calls, but 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 `urbanCharacter` 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 uses the logs of a number of variables. For now, ignore logarithms when attempting to answer this particular question. The researchers' results are in Table 3 (page 8). There are three models shown and let's focus on model 3 (the column labeled as "(3)'') and ignore the other two models. Again, ignore the fact that some variables are on a log scale by simply assuming they are not.

Answer the following questions and include your interpretation of the model outputs:

Part a .

Question

1. Do neighborhoods with more evictions see more or less 311 calls? By how much?

Answer

- At 5% level, the effect is significant.
- Neighborhoods with more evictions see more 311 calls.
- Other things being equal, `311calls` is expected to increase by 0.048 unit if 1 more eviction.

Question

2. How is poverty rate associated with 311 calls? How many more (or fewer) calls there are in neighborhoods with 10% point more poverty?

Answer

- At 5% level, the effect is significant.
- They are negative correlated.
- Other things being equal, `311calls` is expected to decrease $0.14 \times 10 = 1.4$ unit with 10% point more poverty.

Question

3. What can you tell about association of race (*white*) and calls?

Answer

- At 5% level, the effect is not significant.
- They may be negative correlated.

Question

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

Answer

- At 5% level, the effect is significant.
- Older median age is associated with more 311 calls.

Question

5. The value for housing density is `-0.13`. What does this number mean?

Answer

- Other things being equal, if housing density increases by 1 unit, `311calls` is expected to decrease by 0.31 unit.

Question

6. 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?

Answer

- At 5% level, the effect is significant.
- There are fewer calls in similar neighborhoods in Philadelphia, compared to Austin.
- Other things being equal, `311calls` in Philadelphia is 0.56 unit less than in Austin.

Part b .

Question

1. Extra Credit: repeat the previous question, but now take into account the fact that some of the variables are on a log scale. Answer the questions accordingly.

Answer

(log y-log x: x $\uparrow$ 1%, y $\uparrow$ $\widehat{\beta}$%) A1: Other things being equal, if `eviction` increases by 1%, `311calls` is expected to increase by 0.048%.

(log y-linear x: x $\uparrow$ 1, y $\uparrow$ $\widehat{\beta} \times 100\%$) A2: Other things being equal, if `poverty rate` increases by 1 unit, `311calls` is expected to decrease by $0.14 \times 100 = 14\%$.

(log y-linear x: x $\uparrow$ 1, y $\uparrow$ $\widehat{\beta} \times 100\%$) A3: Other things being equal, if `white` increases by 1 unit, `311calls` is expected to decrease by $0.038 \times 100\% = 3.8\%$.

(log y-linear x: x ↑ 1, y ↑ $\widehat{\beta} \times 100\%$) A4: Other things being equal, if `Median Age` increases by 1 unit, `311calls` is expected to increase by $0.0067 \times 100\% = 0.67\%$.

(log y-log x: x ↑ 1%, y ↑ $\widehat{\beta}\%$) A5: Other things being equal, if `Housing Density` increases by 1%, `311calls` is expected to decrease by 0.13%.

(log y-linear x: x ↑ 1, y ↑ $\widehat{\beta} \times 100\%$) A6: Other things being equal, `311calls` in Philadelphia is $0.56 \times 100\% = 56\%$ less than in Austin.

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? You will 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. The data is contained in the `nyc.csv` file on Canvas.

Part a

1. You plan to visit an Italian restaurant for lunch. You check the Zagat review for three different restaurants and you find that Zagat has rated the quality of food for those restaurants 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 a model and interpret your results. Also explain your choice of model*

Question

- Build model for part a.
- Explain model for part a.
- Interpret model for part a.

Answer

```
nyc <- read.csv("/Users/Jay/Desktop/nyc.csv")

HW7Q4A1 <- nyc %>%
  lm(Price ~ Food, data = .)

summary(HW7Q4A1)
```

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

```
# predicted price
predict(HW7Q4A1, data.frame(Food = c(20, 25, 35)))
```

```
##        1        2        3
## 40.94705 55.64185 85.03144
```

2. Your office offers you a $100 reimbursement coupon for your lunch that you are only allowed to use in one go for 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?

Question

- Estimate meal price for each restaurant.
- Correctly answers question and explains choice.

Answer

```
# keep in budget
nyc_new <- nyc %>%
  filter(Price * 1.15 <= 100)

# predict the price with three restaurant suggestions
HW7Q4A2 <- nyc_new %>%
  lm(Price ~ Food + Decor + Service + East, data = .)

summary(HW7Q4A2)
```

```
##
## Call:
## lm(formula = Price ~ Food + Decor + Service + East, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.0465  -3.8837   0.0373   3.3942  17.7491
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -24.023800   4.708359  -5.102 9.24e-07 ***
## Food          1.538120   0.368951   4.169 4.96e-05 ***
## Decor         1.910087   0.217005   8.802 1.87e-15 ***
## Service      -0.002727   0.396232  -0.007   0.9945
## East          2.068050   0.946739   2.184   0.0304 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.738 on 163 degrees of freedom
## Multiple R-squared:  0.6279, Adjusted R-squared:  0.6187
## F-statistic: 68.76 on 4 and 163 DF,  p-value: < 2.2e-16
```

```
# save the predicted values
nyc_new$PriceHat <- predict(HW7Q4A2)

# show the final choice
nyc_new %>%
```

```
  filter(PriceHat == max(PriceHat))
```

Case	Restaurant	Price	Food	Decor	Service	East	PriceHat
88	Erminia	54	25	24	24	1	62.3

Part b

1. 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 a model and interpret the model output.

Question

- Determines whether quality of food and service are important determinants of meal price.

Answer

- Yes, both variables are important determinants of meal price.

```
HW7Q4B1 <- nyc %>%
  lm(Price ~ Food + Service, data = .)

summary(HW7Q4B1)
```

```
##
## Call:
## lm(formula = Price ~ Food + Service, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1333  -4.7053   0.4169   3.5992  27.0728
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -21.1586     5.6651  -3.735 0.000258 ***
## Food          1.4954     0.4462   3.351 0.000997 ***
## Service       1.7041     0.4185   4.072 7.22e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.942 on 165 degrees of freedom
## Multiple R-squared:  0.4486, Adjusted R-squared:  0.4419
## F-statistic: 67.12 on 2 and 165 DF,  p-value: < 2.2e-16
```

2. 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.

Question

- Answers questions on decor.

Answer

- `decor` is an important factor of meal price.

```
HW7Q4B2 <- nyc %>%
  lm(Price ~ Food + Service + Decor, data = .)

summary(HW7Q4B2)
```

```
##
## Call:
## lm(formula = Price ~ Food + Service + Decor, data = .)
##
## 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 ***
## Food          1.5556     0.3731   4.170 4.93e-05 ***
## Service       0.1350     0.3957   0.341    0.733
## Decor         1.8473     0.2176   8.491 1.17e-14 ***
## ---
## 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
```

Problem 5: Mario Kart

For this problem, use the Mario Kart that is part of the `openintro` library (load the library and access the data using the `mariokart` command. See here for more info on the data: https://www.openintro.org/data/index.php?data=mariokart.

1. 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.

Question

- Inspects and describes data.

Answer

- Only 1 title is missing.

```
library(openintro)
data(mariokart)

# structure
str(mariokart)
```

```
## tibble [143 x 12] (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
```

```
# summary statistics
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
```

```
# check missing
is.na(mariokart) %>%
  apply(2, sum)
```

```
##          id    duration      n_bids        cond    start_pr     ship_pr
##           0           0           0           0           0           0
##    total_pr     ship_sp seller_rate stock_photo      wheels       title
##           0           0           0           0           0           1
```

2. Does the duration of the auction effect the price of a MarioKart? To answer this, build an appropriate model and interpret the results to answer this question.

Question

- Builds model for part b.

Answer

- Answer = Total price - ship price

```
HW7Q52 <- mariokart %>%
  lm(I(total_pr - ship_pr) ~ duration, data = .)

summary(HW7Q52)
```

```
##
## Call:
## lm(formula = I(total_pr - ship_pr) ~ duration, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.808  -8.300  -3.307   2.212 256.657
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  49.5292     3.5720  13.866   <2e-16 ***
## duration     -0.7408     0.7823  -0.947    0.345
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.11 on 141 degrees of freedom
## Multiple R-squared:  0.006319,   Adjusted R-squared:  -0.0007279
## F-statistic: 0.8967 on 1 and 141 DF,  p-value: 0.3453
```

Question

- Interprets model for part b.

Answer

- If `duration` increases by 1 day, `total_pr` is expected to decrease 0.7408443.

3. Experiment with other variables you see fit for this task, that is to see how they effect the price of an auction. Do other variables change your results in a major way? Did you have to exclude any variables before fitting the model? Make sure that you build an appropriate model while explaining your choice and interpret the results to answer the questions.

Question

- Builds model for part c.

Answer

```
#cor(mariokart[, -c(1, 4, 8, 10, 12)])

HW7Q53 <- mariokart %>%
  lm(I(total_pr - ship_pr) ~ duration + n_bids + seller_rate + wheels, data = .)

summary(HW7Q53)
```

```
##
## Call:
## lm(formula = I(total_pr - ship_pr) ~ duration + n_bids + seller_rate +
##     wheels, data = .)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -18.945  -7.511  -2.547   3.244 236.612
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.001e+01  7.562e+00   2.647  0.00908 **
## duration    7.261e-01  7.943e-01   0.914  0.36225
## n_bids      7.570e-01  3.293e-01   2.299  0.02302 *
## seller_rate 4.780e-05  3.804e-05   1.257  0.21103
## wheels      1.132e+01  2.409e+00   4.698 6.27e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.45 on 138 degrees of freedom
## Multiple R-squared:  0.1568, Adjusted R-squared:  0.1324
## F-statistic: 6.417 on 4 and 138 DF,  p-value: 9.136e-05
```

Question

- Interprets model for part c

Answer

- Other things being equal, if `wheels` increases by 1 unit, `auction price` is expected to increase by 11.3157985.

Question

- Explains whether variables should be removed.

Answer

- `duration` and `seller_rate` may be considered to remove, but it may impact other values of other coefficients.