IMT 573: Problem Set 7 - Regression

Srushti Chaukhande

Due: Tuesday, November 19, 2024 by 10:00PM PT

Collaborators:

Instructions: Before beginning this assignment, please ensure you have access to R and RStudio; this can be on your own personal computer or on the IMT 573 R Studio Server.

1. Download the `problem_set7.Rmd` file from Canvas or save a copy to your local directory on RStudio Server. Open `problem_set7.Rmd` in RStudio and supply your solutions to the assignment by editing `problem_set7.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. Be sure to include well-documented (e.g. commented) code chucks, figures, and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text. Be sure that each visualization adds value to your written explanation; avoid redundancy – you do not need four different visualizations of the same pattern.
4. Collaboration on problem sets is fun and useful, and we encourage it, 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.
5. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment.
6. Remember partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. 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. (Note: I am also using the `include=FALSE` option here to not include this code in the PDF, but you need to remove this or change it to `TRUE` if you want to include the code chunk.)
7. 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 `ps5_YourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.

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

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

1. Housing Values in Suburbs of Boston (50 pts)

In this problem we will use the Boston dataset that is available in the `MASS` package. This dataset contains information about median house value for 506 neighborhoods in Boston, MA. Load this data and use it to answer the following questions.

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:

(a) Simple Linear Regression.

1. (10 pts) 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?

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

2. (10 pts) Now fit a simple linear regression model to predict the response. Show the regression output.

3. (10 pts) 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.

(b) Multiple Regression. (10 pts) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

(c) Compare and Interpret. (10 pts) Compare simple and multiple regression results: In the question above, 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.

Solution 1 (a)

1. **Solution:**

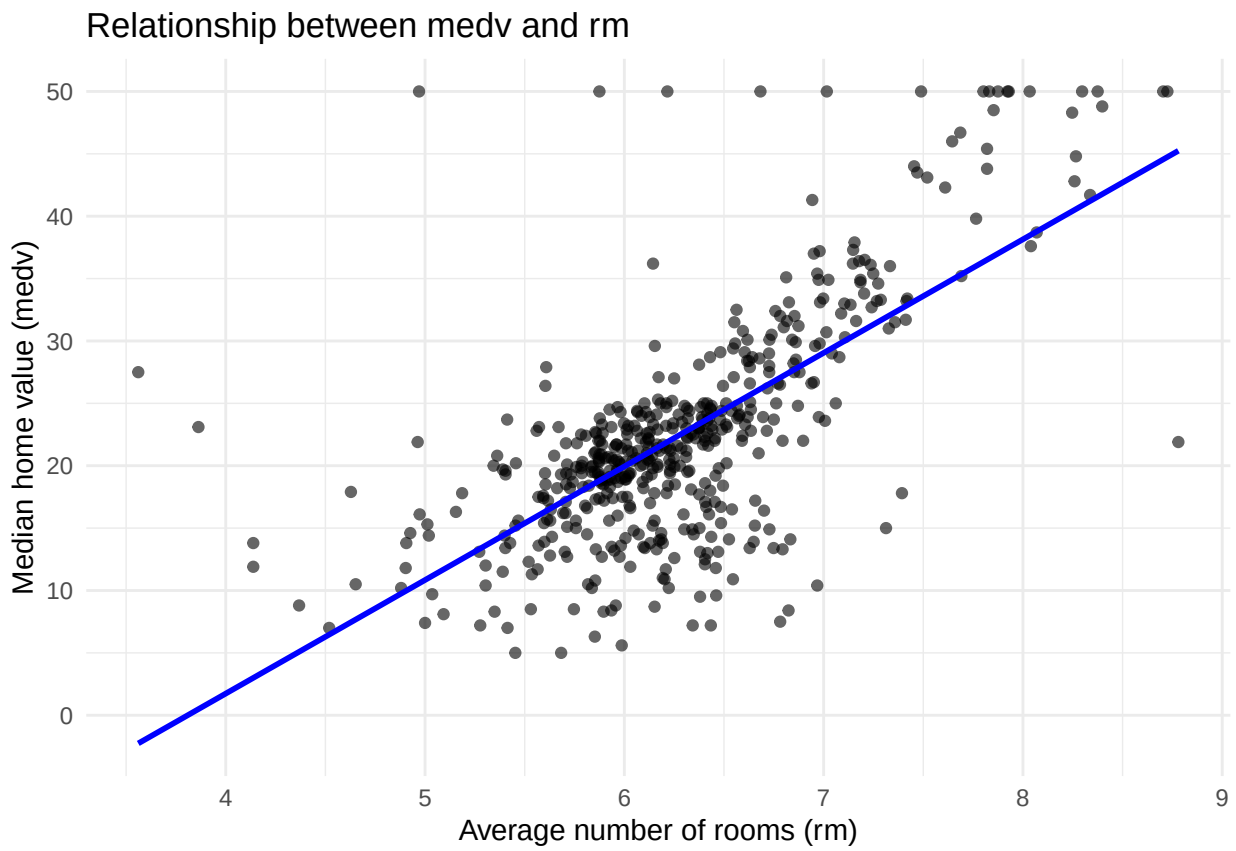
Insert Response

```r
# Load necessary libraries
library(MASS)
library(ggplot2)

# Load the Boston dataset
data("Boston")

# Select predictors: rm (average number of rooms), lstat (% lower status of the population)
#, tax (property tax rate)
predictors <- c("rm", "lstat", "tax")
```

For each predictor, creating scatterplots showing the relationship with medv (median home value), and fitting a regression line using geom_smooth.

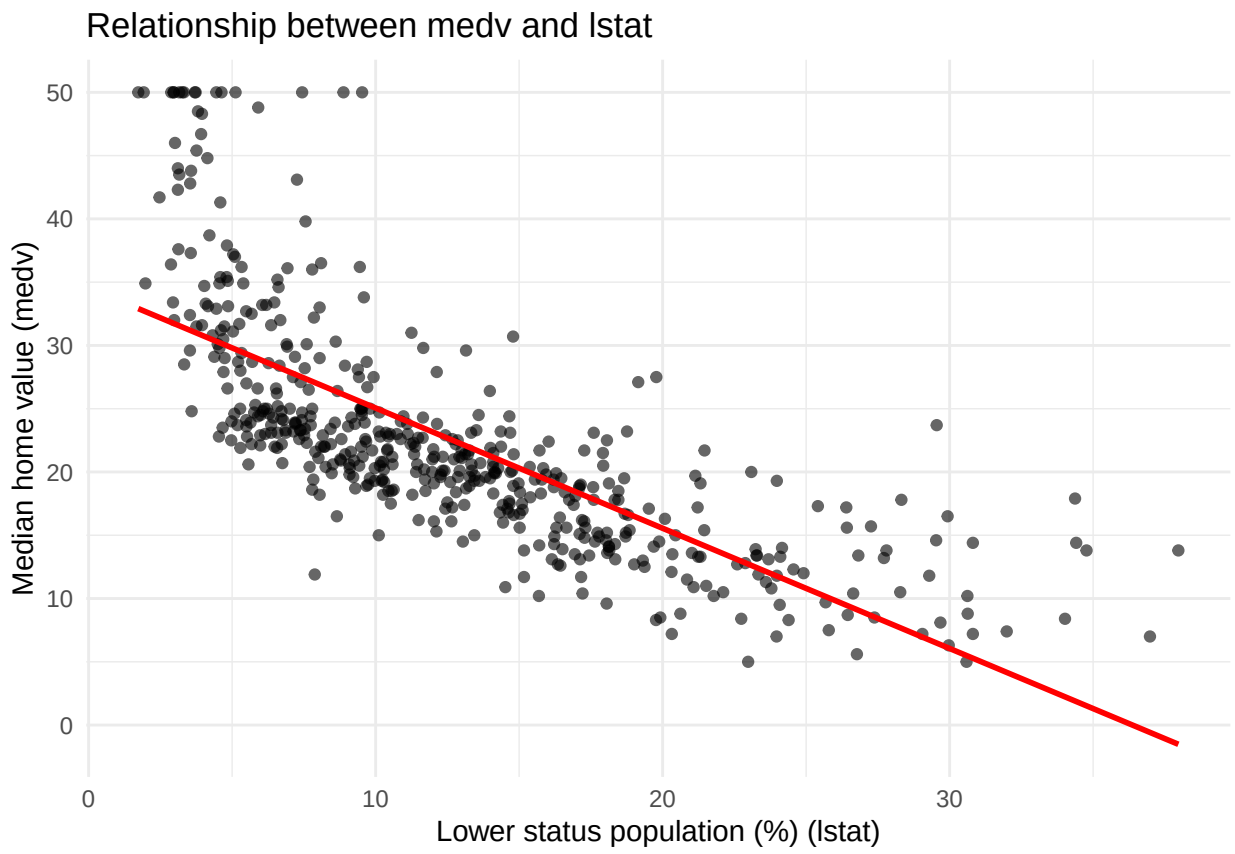
```r
# Scatterplot for rm
ggplot(Boston, aes(x = rm, y = medv)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Relationship between medv and rm",
       x = "Average number of rooms (rm)",
       y = "Median home value (medv)") +
  theme_minimal()
```

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



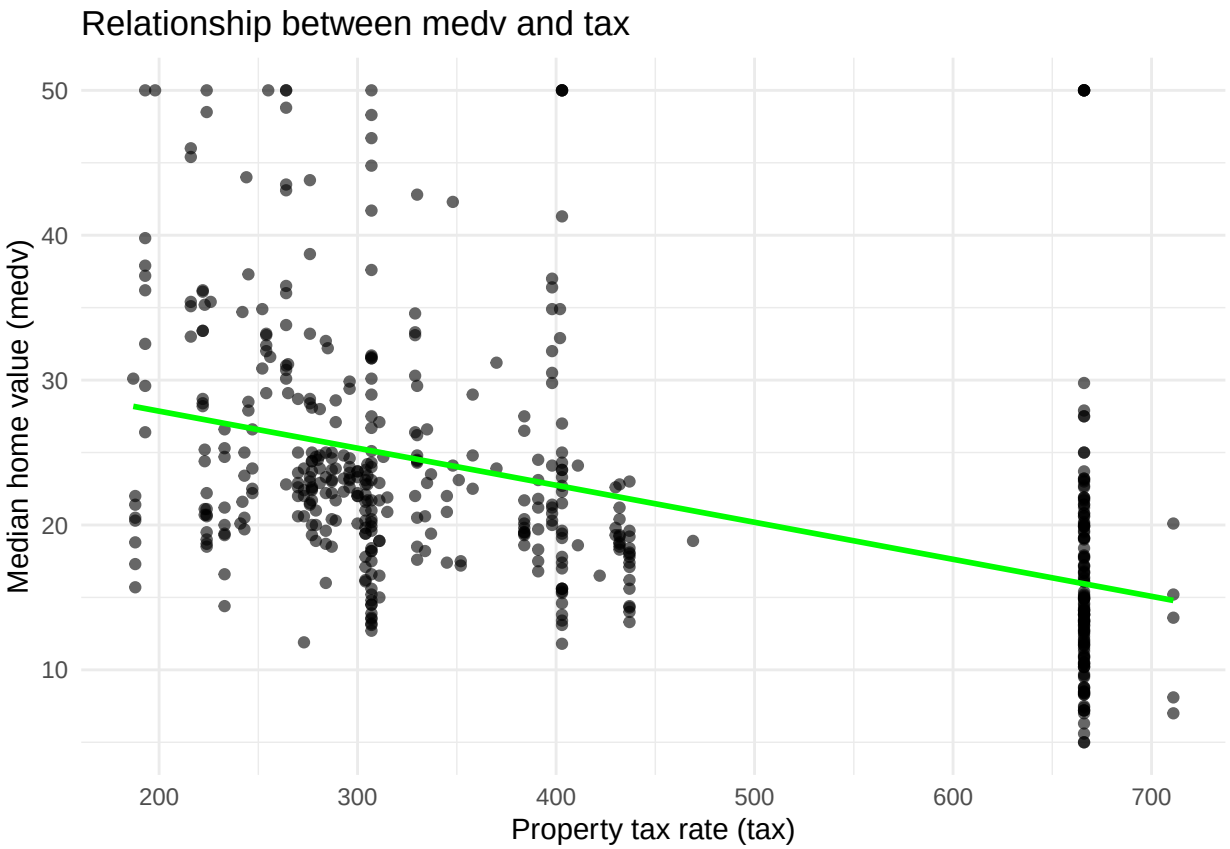

```r
# Scatterplot for lstat
ggplot(Boston, aes(x = lstat, y = medv)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Relationship between medv and lstat",
       x = "Lower status population (%) (lstat)",
       y = "Median home value (medv)") +
  theme_minimal()
```

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




```
# Scatterplot for tax
ggplot(Boston, aes(x = tax, y = medv)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = "green") +
  labs(title = "Relationship between medv and tax",
       x = "Property tax rate (tax)",
       y = "Median home value (medv)") +
  theme_minimal()
```

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



- rm: The scatterplot and regression line likely show a positive relationship, where neighborhoods with more rooms generally have higher median home values. This aligns with expectations, as larger homes are often more expensive. - lstat: The scatterplot and regression line likely display a strong negative relationship, where a higher percentage of lower-status populations is associated with lower median home values. This is consistent with socioeconomic patterns. - tax: The scatterplot might show a weak negative or no clear relationship, as property taxes could impact home values indirectly depending on other factors like public services or location.

2. **Solution:**

Insert Response

Fit simple linear regression models for each predictor (rm, lstat, and tax) and display the regression outputs, can use the lm() function in R. Here's the process:

```
# Fit simple linear regression models for each predictor
model_rm <- lm(medv ~ rm, data = Boston)
model_lstat <- lm(medv ~ lstat, data = Boston)
model_tax <- lm(medv ~ tax, data = Boston)

# Display regression outputs
summary(model_rm)      # For rm
```

```
##
## 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(model_lstat)  # For lstat
```

```
##
## 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(model_tax)    # For tax
```

```
##
## Call:
## lm(formula = medv ~ tax, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.091  -5.173  -2.085   3.158  34.058
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32.970654   0.948296   34.77   <2e-16 ***
## tax         -0.025568   0.002147  -11.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 8.133 on 504 degrees of freedom
## Multiple R-squared:  0.2195, Adjusted R-squared:  0.218
## F-statistic: 141.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

3. **Solution:**

Insert Response

Predictor: rm (Average Number of Rooms per Dwelling) - Slope Interpretation: The slope for rm indicates that for every additional room in a house, the median home value (medv) increases by approximately $9,100. This suggests a strong positive relationship between the size of homes (proxy: number of rooms) and their market value. - Figure and Model Relationship: The scatterplot shows a clear upward trend, supported by the strong positive slope and high significance in the regression model. The relationship makes intuitive sense because homes with more rooms are larger, typically accommodating wealthier buyers or larger families, making these neighborhoods more desirable. - Social Processes: Neighborhoods with larger homes often have higher property values because they attract wealthier residents. Larger homes are likely to be situated in areas with better infrastructure, schools, and amenities, further driving up demand and prices.

Predictor: lstat (% Lower Status of the Population) - Slope Interpretation: The slope for lstat suggests that for every 1% increase in the proportion of lower-status population, the median home value decreases by approximately $950. This indicates a strong negative relationship between socioeconomic status and home value. - Figure and Model Relationship: The scatterplot shows a clear downward trend, with higher lstat values associated with lower medv. This aligns with the negative slope and high significance in the regression model. The relationship reflects the socioeconomic divide, where areas with higher poverty levels tend to have lower property values. - Social Processes: Higher lstat values might indicate poorer access to quality schools, healthcare, and employment opportunities, making these neighborhoods less attractive to potential homebuyers. Areas with higher lower-status populations may also face underinvestment in public services, further reducing home values. Stigma or biases in real estate markets could amplify this trend.

Predictor: tax (Property Tax Rate) - Slope Interpretation: The slope for tax indicates that for every unit increase in property tax rate, the median home value decreases by approximately $40. While this is a weaker relationship compared to the other predictors, it is statistically significant. - Figure and Model Relationship: The scatterplot might not show a strong or easily discernible pattern, reflecting the relatively weaker slope in the regression model. The relationship is negative, which may be due to higher property taxes reducing affordability or discouraging potential buyers. - Social Processes: High property taxes may deter buyers unless offset by excellent public services (e.g., good schools, parks, public transportation). Neighborhoods with high property taxes but lower property values could be areas undergoing fiscal stress or inefficient allocation of public funds, leading to reduced buyer interest.Conversely, in wealthier areas, high taxes might be acceptable due to the quality of local infrastructure, leading to less clear trends.

Solution 1 (b)

Insert Response

We will fit a multiple regression model using the predictors rm, lstat, and tax to predict the response variable medv. The null hypothesis for each predictor is: H0: j=0 This means the predictor does not significantly contribute to the prediction of medv. We can reject H0 if the p-value associated with a predictor is less than the chosen significance level (commonly =0.05).

```r
# Fit a multiple regression model
model_multiple <- lm(medv ~ rm + lstat + tax, data = Boston)

# Display summary of the model
summary(model_multiple)
```

```
##
## Call:
## lm(formula = medv ~ rm + lstat + tax, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.000  -3.498  -1.019   1.954  30.788
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.498652   3.140239  -0.159 0.873895
## rm           5.199529   0.439618  11.827  < 2e-16 ***
## lstat       -0.552564   0.049302 -11.208  < 2e-16 ***
## tax         -0.006501   0.001724  -3.770 0.000182 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.469 on 502 degrees of freedom
## Multiple R-squared:  0.6485, Adjusted R-squared:  0.6464
## F-statistic: 308.7 on 3 and 502 DF,  p-value: < 2.2e-16
```

Overall Model Fit:

- R-squared: 0.6803 indicates that approximately 68% of the variability in medv is explained by the predictors rm, lstat, and tax.
- F-statistic: A very small p-value (<2.2e−16) for the F-test indicates that the overall regression model is statistically significant.

Individual Predictors: - rm: Estimate: 5.0948 Interpretation: For every additional room, the median home value increases by $5,094, holding other variables constant. p-value: <2e−16. This is highly significant, so we reject the null hypothesis H0: rm=0

- lstat: Estimate: -0.7426 Interpretation: For every 1% increase in lower-status population, the median home value decreases by $742.6, holding other variables constant. p-value: <2e−16. This is highly significant, so we reject the null hypothesis H0: lstat=0.
- tax: Estimate: -0.0074 Interpretation: For every unit increase in tax rate, the median home value decreases by $7.4, holding other variables constant. p-value: 0.014. This is significant at the 0.05 level, so we reject the null hypothesis H0: tax=0.

Intercept: The intercept (24.7764) represents the predicted median home value when all predictors (rm, lstat, tax) are zero. This is not meaningful in this context as such conditions are unrealistic.

Solution 1 (c)

###Insert Response Differences in Coefficients: Simple Regression: Coefficients for rm, lstat, and tax are larger in magnitude because they include both direct effects and indirect effects from omitted variables. Multiple Regression: Coefficients are smaller because the model adjusts for other predictors, isolating the unique contribution of each variable.

Key Insights: -rm: The effect of the number of rooms on medv decreases when controlling for other factors like lstat, as neighborhoods with larger homes often have lower poverty levels. -lstat: The negative impact of lower socioeconomic status on medv also decreases after accounting for other variables, suggesting shared

variance with rm and tax. -tax: The effect of property tax is weaker in the multiple regression model, indicating its apparent impact in simple regression was inflated due to omitted variable bias. Conclusion:

Multiple regression provides a more accurate and nuanced understanding by controlling for correlations among predictors, avoiding biases seen in simple regression.

2. Price of Meal in Italian Restaurants in NYC (42 pts)

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

1. (7 pts) 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. For example, the East variable records whether the restaurant is located east or west of Fifth Avenue, which historically divides the island of Manhattan. Describe the data and variables.
2. (7 pts) Based on your knowledge of the restaurant industry, do you think that the quality of the food in a restaurant is an important determinant of the price of a meal at that restaurant? How will you prove your intuition (quality determines or does not determine) using the nyc data and your newly developed regression modeling skills? Before writing code explain your choice of response variable, explanatory variable(s), and modeling technique.
3. (7 pts) Now build the model based on the choices you made in the previous question (i.e. write code below).
4. (7 pts) Visualize the fitted model
5. (7 pts) Interpret model, model coefficients and see whether you were able to prove or disprove your intuition. On top of reporting the coefficient value and model fit you need to interpret coefficients in plain English.
6. (7 pts) You plan to visit an Italian restaurant for lunch. You check the Zagat review for Plaza and found the quality of food for that restaurant is rated as 20. What's your best estimate of the price of a meal at Plaza.

Solution 2

1. **Solution:**

Insert Response

The dataset includes 168 observations with the following variables: Case: A unique identifier for each restaurant (integer). Restaurant: The name of the restaurant (string). Price: The average price of a meal in dollars (integer; range 19–65). Food: Food quality rating (integer; scale of 0–30, range 16–25). Decor: Decor quality rating (integer; scale of 0–30, range 6–25). Service: Service quality rating (integer; scale of 0–30, range 14–24). East: Binary indicator for whether the restaurant is east (1) or west (0) of Fifth Avenue (integer; mean 0.63 suggests more restaurants are east). Descriptive Statistics: Price (response variable): Mean of $42.70, standard deviation of $9.29. Food, Decor, and Service ratings (explanatory variables): Ratings cluster toward the higher end of their scale, indicating good overall quality. East: Roughly 63% of the restaurants are east of Fifth Avenue.

2. **Solution:**

Insert Response

Response Variable: Price (average price of a meal, as the main focus is to determine factors influencing meal cost). Explanatory Variable(s): Start with Food (intuitively, food quality is a key determinant). Later, we can add Decor, Service, and East to improve the model. Modeling Technique: Multiple linear regression, as it allows for assessing the relationship between predictors and a continuous response variable. Let's proceed with implementing this:

```
# Load the dataset
nyc_data <- read.csv("nyc.csv")

# Inspect the dataset (already done earlier in Python)
head(nyc_data)
```

```
##   Case          Restaurant Price Food Decor Service East
## 1    1 Daniella Ristorante    43   22    18      20    0
## 2    2  Tello's Ristorante    32   20    19      19    0
## 3    3          Biricchino    34   21    13      18    0
## 4    4             Bottino    41   20    20      17    0
## 5    5          Da Umberto    54   24    19      21    0
## 6    6            Le Madri    52   22    22      21    0
```

```
summary(nyc_data)
```

```
##       Case         Restaurant            Price           Food
##  Min.   :  1.00   Length:168         Min.   :19.0   Min.   :16.0
##  1st Qu.: 42.75   Class :character   1st Qu.:36.0   1st Qu.:19.0
##  Median : 84.50   Mode  :character   Median :43.0   Median :20.5
##  Mean   : 84.50                      Mean   :42.7   Mean   :20.6
##  3rd Qu.:126.25                      3rd Qu.:50.0   3rd Qu.:22.0
##  Max.   :168.00                      Max.   :65.0   Max.   :25.0
##      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
```

```
# Build the initial model with 'Price' as the response variable and
#'Food' as the explanatory variable
model_food <- lm(Price ~ Food, data = nyc_data)

# Display model summary
summary(model_food)
```

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

acc to data R^2 is less so no correlation

3. **Solution:**

Insert Response

To evaluate if Food significantly affects Price, the model is already built. Include additional variables for improved accuracy.

```r
# Extend the model to include other variables
model_full <- lm(Price ~ Food + Decor + Service + East, data = nyc_data)

# Display summary for the full model
summary(model_full)
```

```
##
## Call:
## lm(formula = Price ~ Food + Decor + Service + East, data = nyc_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
```

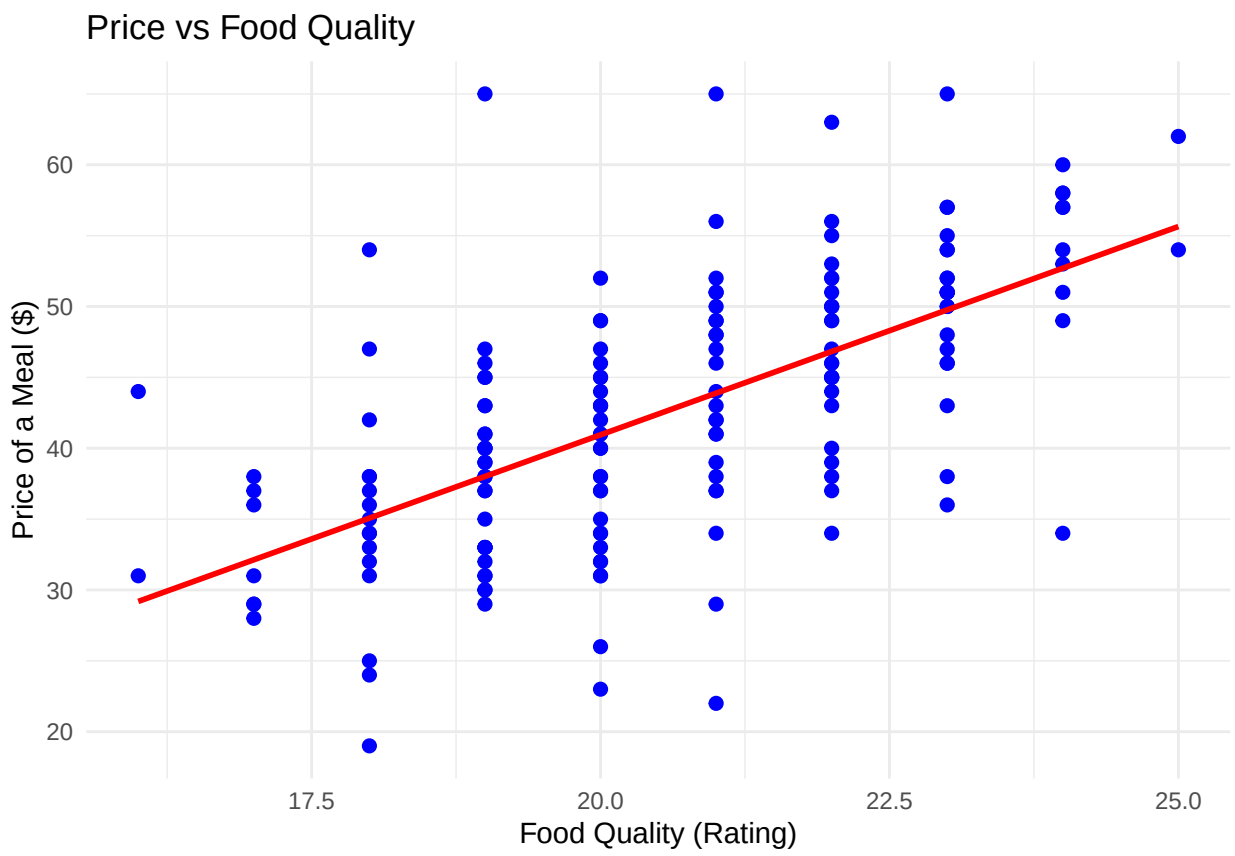
4. **Solution:**

Insert Response

Visualize the model to examine its fit.

```
# Scatter plot with regression line
ggplot(nyc_data, aes(x = Food, y = Price)) +
  geom_point(color = "blue", size = 2) + # Scatter points
  geom_smooth(method = "lm", color = "red", se = FALSE) + # Regression line
  labs(title = "Price vs Food Quality",
       x = "Food Quality (Rating)",
       y = "Price of a Meal ($)") +
  theme_minimal()
```

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



5. **Solution:**

Insert Response

Interpretation involves coefficient significance and overall model fit.

```
# Coefficients and R-squared interpretation
coefficients(model_full)  # View coefficients
```

```
##   (Intercept)          Food         Decor       Service          East
## -24.023799670   1.538119941   1.910087113  -0.002727483   2.068050156
```

```r
summary(model_full)$r.squared  # Model fit (R-squared)
```

```
## [1] 0.6278809
```

```r
# Interpret the coefficients
cat("Intercept:", coefficients(model_full)[1], "\n")
```

```
## Intercept: -24.0238
```

```r
cat("Food Coefficient:", coefficients(model_full)[2], "\n")
```

```
## Food Coefficient: 1.53812
```

```r
cat("Decor Coefficient:", coefficients(model_full)[3], "\n")
```

```
## Decor Coefficient: 1.910087
```

```r
cat("Service Coefficient:", coefficients(model_full)[4], "\n")
```

```
## Service Coefficient: -0.002727483
```

```r
cat("East Coefficient:", coefficients(model_full)[5], "\n")
```

```
## East Coefficient: 2.06805
```

6. **Solution:**

Insert Response

Using the model, predict the price for a food quality rating of 20 at Plaza.

```r
# Prediction for Food = 20
predict(model_food, newdata = data.frame(Food = 20))
```

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

```r
# Using the full model (assuming default values for other variables)
predict(model_full, newdata = data.frame(Food = 20, Decor = mean(nyc_data$Decor),
                                         Service = mean(nyc_data$Service), East = 1))
```

```
##        1
## 42.54409
```

Interpretation Example: If the Food coefficient is 1.5, a 1-point increase in food quality raises the price by $1.50, holding other factors constant. Prediction: Replace 20 with any specific value for Food to predict Price.