MISSISSIPPI STATE UNIVERSITY

Data Analysis One

Emily Hubbard

September 17, 2025

HW 2

Executive Summary:

A. Introduction:
Linear Regression is a way to investigate whether or not two variables or more have a linear relationship. As one increases or decreases, does the other increase or decrease? When testing the importance of more than one covariate, the process is known as Multiple Linear Regression. This allows us to understand the effects of multiple variables on a specific response variable. Depending on the data, this can allow for a more precise model. The most crucial part of MLR is to find the correct covariates to consider. Linear Regression loses its usefulness if the model becomes too complicated. One of the motivations to us LR is to have an easy interpretation of coefficients. This characteristic becomes less impactful when there are too many variables to make reasonable interpretations. Therefore, choosing the most valuable covariates is the most important step in creating a MLR model. However, if this is done correctly, it can allow for a more well-rounded interpretation of covariates and prediction of the response variable. The same as simple LR, you test the reliability of the model based on its satisfaction of the base assumptions: 1. Linearity, 2. Normality, 3. Constant Variance, 4. Independence.

This data set includes 12 variables on 1,266 separate orders received in the month of May 2014 at an Italian restaraunt offering home delivery. This report will outline the process of finding the best fit MLR model, checking assumptions, and drawing impactful interpretations.

A simple linear model and a multiple linear model are based on the following formulas, respectively:

$$Y = \alpha + \beta_1 X + \epsilon \text{ (Simple LR)}$$
$$Y = \alpha + \sum_{i=0}^{p} \beta_i X_i + \epsilon \text{ (Multiple LR)}$$

B. Data Collection
Consider the pizza delivery data described below: The pizza delivery data is a simulated data set. The data refers to an Italian restaurant which offers home delivery of pizza. It contains the orders received during a period of one month: May 2014. There are three branches of the restaurant. The pizza delivery is centrally managed: an operator receives a phone call and forwards the order to the branch which is nearest to the customer's address. One of the five drivers (two of whom only work part time at the weekend) delivers the order. The data set captures the number of pizzas ordered as well as the final bill which may also include drinks, salads, and pasta dishes. The owner of the business observed an increased number of complaints, mostly because pizzas arrive too late and too cold. To improve the service quality of his business, the owner wants to measure:

- the time from call to delivery and
- the pizza temperature at arrival (which can be done with a special device)

Ideally, a pizza arrives within 30 min of the call; if it takes longer than 40 min, then the customers are promised a free bottle of water. The temperature of the pizza should be above 65^C at the time of delivery. The analysis of the data aims to determine the factors which influence delivery time and temperature of the pizzas.

B. Summary Information

The best fit model when considering two or more covariates (multiple regression) is given when running a stepAIC function on the response variable time given all covariates. This will then rule out any covariates that do not have a significant effect on the response variable. The covariates that were outputted by the stepAIC function were the following: temperature, branch, day, driver, bill, and number of pizzas. This model gives us the highest R^2_{adj} compared to the first model tested in this report.

The R^2_{adj} is still quite low, around 31 percent, but there is still a significant effect on the response given these 6 covariates. To clean this model, the next steps would be to consider combining or getting rid of some of the covariates due to correlation. It is also worth considering implementing some transformations on certain variables to see if that increases R^2_{adj}. Overall, this model needs work because as seen when predicting the response variable in question 10, it was not exact. There was at least 4 percent error. Whether this is good or bad, I would have to know more about the industry. However, I think there is potential to find a

more precise model working with the same covariates that are altered in some way.

Now we will look at these results in more detail.

Data Analysis:

1. Read the data into R. Fit a multiple linear regression model with delivery time as the outcome and temperature, branch, day, operator, driver, bill, number of ordered pizzas, and discount customer as covariates. Give a summary of the coefficients.

   Covariate Coefficients:
   The covariates that are **bolded** are those that have significant p-values. This is to visually see the important covariates that we may want to consider keeping in the model. This will also offer us a comparison when we run the stepAIC function.

   Temperature = -0.208
   Branch East = -1.603
   Branch West = -0.119
   Day Monday = -1.159
   Day Saturday = 0.882
   Day Sunday = 1.017
   Day Thursday = 0.789
   Day Tuesday = 0.793
   Day Wednesday = 0.258
   Operator Melissa = -0.158
   Driver Domenico = -2.593
   Driver Luigi = -0.809
   Driver Mario = -0.395
   Driver Salvatore = -0.504
   Bill = 0.141
   Pizzas = 0.556
   Discount Customer = -0.283

   Explanation:
   Using the summary output, it is easy to see the significance of some of the covariate coefficients in predicting pizza delivery time. Whenever the covariate is increased by one unit, its corresponding coefficient gives the increase(+) or decrease(-) that will be seen in pizza delivery time. There are 5 coefficients that are extremely significant in

predicting the response variable: temperature, branchEast, driverDomenico, bill, and number of pizzas. According to the p-values (which are also significantly lower than an $\alpha = 0.05$), these variables are important in this linear regression model. This is also supported by the p-value result of the F-test. Having such a low p-value indicates that the full model (as compared to the reduced model) is a better predictor for the response variable. What this means is that the covariates are important and at least some of them should be kept in the model to allow a more accurate prediction.

2. Use R to calculate the 95 percent confidence intervals of all coefficients.

Confidence Intervals:

Temperature = $(-0.259, -0.157)$
Branch East = $(-2.433, -0.772)$
Branch West = $(-0.851, 0.613)$
Day Monday = $(-2.400, 0.083)$
Day Saturday = $(-0.102, 1.866)$
Day Sunday = $(-0.084, 2.117)$
Day Thursday = $(-0.251, 1.829)$
Day Tuesday = $(-0.434, 2.020)$
Day Wednesday = $(-0.932, 1.448)$
Operator Melissa = $(-0.831, 0.515)$
Driver Domenico = $(-4.034, -1.152)$
Driver Luigi = $(-1.961, 0.343)$
Driver Mario = $(-1.252, 0.462)$
Driver Salvatore = $(-1.357, 0.349)$
Bill = $(0.110, 0.172)$
Pizzas = $(0.326, 0.786)$
Discount Customer = $(-1.006, 0.440)$

Explanation:
Notice the sign of the values within the confidence intervals. These values give us another indication of the most significant variables. All 5 that showed significant p-values also have a range that is fully positive or fully negative. This makes sense because when a confidence interval includes 0, it is possible for the covariate to have zero affect on the response variable. When ranges do not include zero, it can be concluded that 95 percent of the time, a test will conclude the covariate has an effect on the response, even if it is higher or lower than the defined coefficient value.

3. Reproduce the least squares estimate of σ^2.

   Explanation:
   The folowing formulas are a walk through of how to calculate the Least Squares Estimate (LSE) of σ^2 AKA Residual Standard Error(RSE) by using the Residual Sum of Squares (RSS) and degrees of freedom(df).

   Calculating RSE:

$$\text{RSE} = \sqrt{\text{RSS}/\text{df}}$$
$$\text{RSS} = 36028.96$$
$$\text{df} = 1248$$
$$\text{RSE} = \sqrt{36028.96/1248}$$
$$\text{RSE} = 5.373$$

   This RSE or LSE is the same as given by the Model One summary output.

4. Now use R to estimate both R^2 and R^2_{adj} with the results of model output from part 1. Interpret the results.

   Explanation:
   From the summary information, we see that R^2 is 0.3178 and R^2_{adj} is 0.3085. These R-squared values tell us a lot about the accuracy of the current model. With an adjusted R-squared of 0.3085, it is reasonable to say that we are missing out on a lot of information that this data could be giving us. We are losing almost 70 percent accuracy with the current model. This encourages me to rework the model to find the covariates that will give us the highest amount of prediction precision.

5. Use backward selection by means of the stepAIC function from the library MASS to find the best model according to AIC.
   Explanation:
   According to the AIC step function conducted in R, the best model includes: temperature, branch, day, driver, bill, and number of pizzas. This was supported by the p-values and confidence intervals found for the first model. These covariates were the only ones that held significant values. Therefore, it makes perfect sense that they would show up in the best fit multiple regression model.

6. Obtain R^2_{adj} from the model identified in 5. and compare it to the full model from 1.

Explanation:
The R^2_{adj} barely changed in Model Two compared to Model One. it only increased by 0.0007. That is minuscule and indicates this model has the same level of accuracy as the first model. I think that to get a model with a greater R^2_{adj}, you would have to specify which driver and which branch to include in your prediction.

7. Identify whether the model assumptions are satisfied or not.

Assumptions Check:

Linearity: Given that the F-test in the summary output of Model Two indicates that the included covariates are helpful in the model, I would subjectively say that this data meets the assumption of linearity. Because using the full model compared to the reduced does indicate a slight linear relationship. If it had no impact, you would simply drop on or more of the covariates.

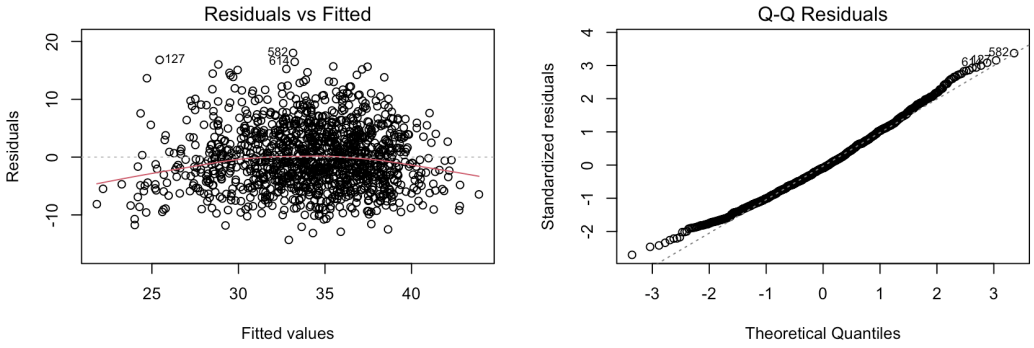

Figure 0.1: Assumption Check for Constant Variation and Normality Model Two

Normality: By using the lillie.test in R and observing the Q-Q plot in Figure 0.1, a conclusion can be made about the data's normality- it is not normal. The P-value associated with the lillie.test is well below 5 percent meaning that we reject the null. The Q-Q plot indicates outliers towards the end and beginning of the observations. This implies non-normality of the data.

Constant Variance: The Residuals vs Fitted plot in Figure 0.1 shows an curved variation of the residuals. Towards the middle it levels out slightly but bends at the beginning and end of the observations. This means that the assumption for constant variation is not satisfied. This also supports the conclusion that the residuals diverge from normal.

Figure 0.2: ACF Graph Model 2

Independence:
Of Residuals:
The ACF function in Figure 0,2 tells us that the residuals are mostly independent. However, there are 3 bars around the 23 to 32 mark that slightly cross the blue-dotted line. This indicates that there could be some correlation between some of the residuals. This is a similar finding to the independence of the covariates. They are practically independent, but there are a few that need to be checked for overlap. With this being said, I will conservatively conclude that the residuals are not independent.

Of Covariates:
Because some of the covariates are numerical and some are categorical, there are seperate checks that we have to run. For the numerical covariates alone, it was simply to put together a correlation chart and use that to visualize any possible correlations.

Numerical Variable Check:

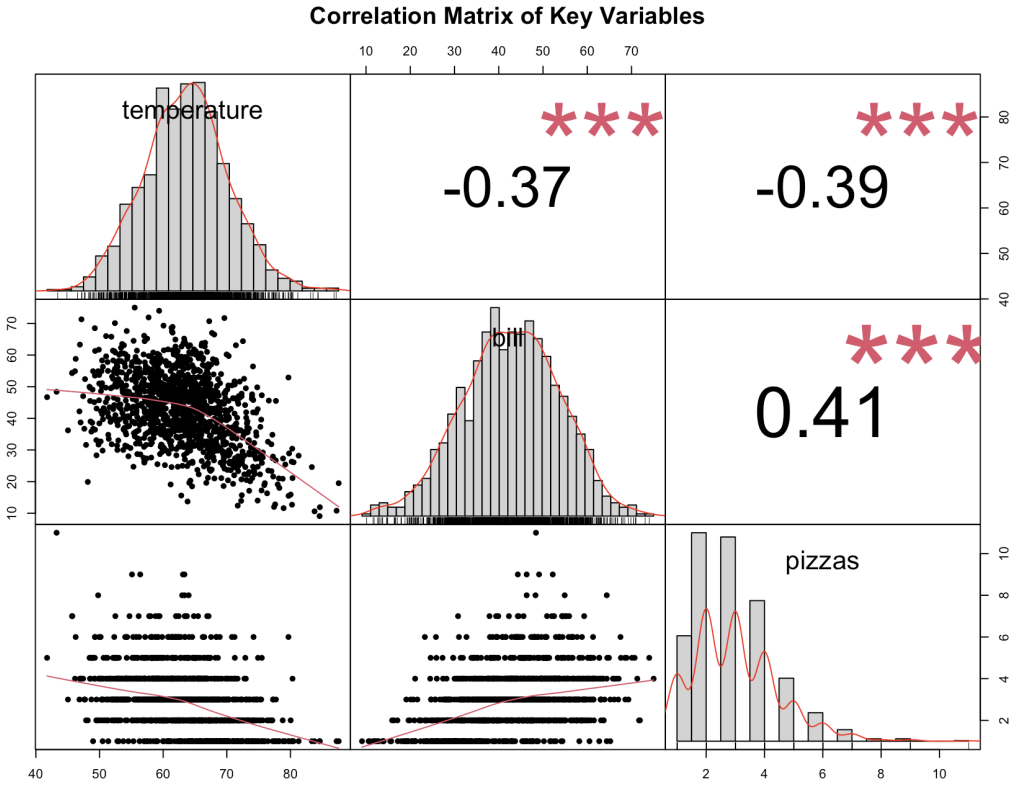

Figure 0.3: Correlation Matrix of Key Numeric Variables

Figure 0.3 Correlation Matrix tells us so much about the numerical covariates (temperature, bill, and pizzas) and their independence or lack there of. There is a slight negative correlation between temperature and bill, and also temperature and pizzas. Probably for the same reason: more pizzas (and higher price) means a larger order which implies a longer prep time leading to more opportunity for pizza to lose heat. There is a pretty strong correlation between bill and pizzas. This makes sense because you spend more money on each pizza that you add to your order. This correlation might cause overemphasis of certain information. Because of these things, I conclude that the numerical covariates are not independent, at least not all of them.

Categorical Variable Check:
For the categorical variables (day, branch, and driver) I ran a correlation test comparing each of them with the other using a χ^2 test. Using a Cramer V correlation value, we can see if each of the variables have correlation with the other.

Branch v. Driver:

Correlation Value ≥ 0.30 indicates moderate to extreme correlation

Correlation Value < 0.30 indicates weak to nonexistent correlation

Correlation Value = 0.1498

Results: The value indicates a weak positive correlation between the covariates branch and driver.

Branch v. Day:

Correlation Value ≥ 0.30 indicates moderate to extreme correlation

Correlation Value < 0.30 indicates weak to nonexistent correlation

Correlation Value = 0.0680

Results: This value indicates nonexistent correlation AKA independence between branch and day.

Day v. Driver:

Correlation Value ≥ 0.30 indicates moderate to extreme correlation

Correlation Value < 0.30 indicates weak to nonexistent correlation

Correlation Value = 0.3532

Results: The value indicates a moderate level of correlation between the covariates day and driver.

Based on these test results, we can conclude that, similar to the residuals, most of these covariates are independent of each other but there is some overlap. Therefore, we must conclude that these covariates do not satisfy the assumption of independent covariates.

Assumptions Summary:

Linearity = TRUE

Normality = FALSE

Constant Variance = FALSE

Independence:

Of Residuals = FALSE

Of Covariates = FALSE

8. Are all variables from the model in 5. causing the delivery time to be either delayed or improved?

Explanation:
Temperature, Domenico the driver, and the East branch cause the delivery time to be shortened (aka improve). Whereas the bill and number of pizzas cause the order time to increase (aka delay). This is gathered from the coefficients listed on page 3 of this report.

9. Test whether it is useful to add a quadratic polynomial of temperature to the model.

Explanation:
Using R to plug in the squared version of temperature into a new model allowed us to determine whether or not a quadratic polynomial of temperature was helpful or not in making the model more precise. The summary shows an increase in the adjusted R-squared which immediately makes this a better model than our previous one. This means that keeping the quadratic transformation of temperature in the model helps with its predictive power. This quadratic transformation being helpful also allows us to say that the relationship between time and temperature is not strictly linear (because a curved shape provides a better fit). This is further supported by the results of the Anova Test: a significantly low p-value that indicates the quadratic model is the superior choice.

10. Use the model identified in 5. to predict the delivery time of the last captured delivery (i.e. number 1266). Use the predict() command to ease the calculation of the prediction.

Explanation:
Using the last captured delivery as our "new data" in the predict function gave us the estimated time of delivery strictly based on the model we constructed. The calculation and interpretation are as follows:

Prediction Output and Percent Error:

$$\begin{aligned}
&\text{Predicted Delivery Time} = 34.2296 \text{ minutes} \\
&\text{Actual Delivery} = 35.7375 \text{ minutes} \\
&\text{Percent Error} = \frac{\text{Predicted-Actual}}{\text{Actual}} \\
&\text{Percent Error} = 0.0422, 4.22 \text{ percent}
\end{aligned}$$

Interpretation:
Using the latest model, we predict the time of the latest delivery to be 34.2296 minutes. There is a percent error of 4.22 percent in this prediction. I am not sure of its significance based on context, but it would not come as a shock to me if this percent error was significant because of the low value of the R^2_{adj} of our model.

APPENDIX
R code:

```
#### HW 2 ####
data <- read.csv("pizza_delivery.csv",header = TRUE)
data
View(data)
#### 1 Model One ####
ml <- lm(time~ temperature + branch + day + operator + driver + bill + pizzas + discount
ml

summary(ml)

#Using the summary output, it is easy to see the significance of some of these coefficie
#Whenever the covariate is increased by one unit, its corresponding coefficient gives th
#There are 5 coefficients that are extremely significant in predicting the response vari
#According to the p-values (which are also significantly lower than and alpha of 0.05),
#This is also supported by the p-value result of the F-test. Having such a low p-value i
#What this means is that the covariates are important and at least some of them should b

#### 2 ####
confint(ml, level = 0.95)  # 95% CI

#These confidence intervals also give us an indication of the most significant variables
#When a confidence interval includes 0, this means that it is possible for that covariat
#it can be concluded that 95% of the time, a test will conclude the covariate has an eff

#### 3 ####
#Least Squares Estimate of sigma^2 AKA Residual Standard Error is calculated by the ollo
#Sqrt(Residual sum of squares/degrees of freedom)

#To calculate RSS, we have to square the residuals and then sum them which is produced b

residuals <- resid(ml) #residuals
RSS <- sum(residuals^2) #squaring residuals and then summing them to produce RSS

df <- ml$df.residual #retrieve degrees of freedom from the model

LSE <- sqrt(RSS / df) #calculating Least Squares Estimate (LSE)
LSE

#This gives us 5.373 which matches the Residual Standard Error that the summary output g
```

```
#### 4 ####
#From the summary information, we see that R-squared is 0.3178 and Ajusted R-squared is
#These R-squared values tell us a lot about the accuracy of the current model. With an a
#it is reasonable to say that we are missing out on a lot of information that this data
#We are losing almost 70 percent accuracy with this current model. This encourages me to
#covariates that will give us the highest amount of prediction precision.

#### 5 ####
library(MASS)
step <- stepAIC(m1, direction="backward")
step$anova

#According to AIC, the best model includes: temperature, branch, day, driver, bill, and

#### 6 Model Two ####
m2 <- lm(time~ temperature + branch + day + driver + bill + pizzas, data = data)
m2

summary(m2)

#The adjusted  R-squared barrely changed in Model Two compared  to Model One. it only in
#That is minuscule  and indicates the same level of accuracy that is lost in this model.
#I think that to get a model with a greater Adjusted  R-squared, you would have to speci

#### 7 Checking Assumptions ####
###Checking for normality of residuals (which will apply to the normality of the respons
install.packages("nortest")
library(nortest)

m2$residuals
lillie.test(m2$residuals)
#P-value is well below 5% meaning that we reject the null.
#In this case, it means that the residuals vary significantly from normality.
#It does not meet the normality assumption.

###Checking for constant variation
mean(m2$residuals)
par(mfrow=c(2,2))
plot(m2)
#The Residuals vs Fitted plot shows an increasing variation. This means that the assumpt
#constant variation is not met. This also supports the conclusion that the residuals div
#The Q-Q plot matches the result of the lillie.test- not normal.
#The plot indicates outliers towards the end and beginning of the observations.
```

```r
###Checking for independence of residuals
acf(m2$residuals)
#The acf function tells us that the residuals are mostly independent.However, there are
#cross the blue-dotted line. This indicates that there could be some correlation between
#This is a similar finding to the independency of the covariates. They are moslty indepe
#that need to be checked for overlap.

### Checking for independence of covariates

##Numerical covariate independence check
install.packages("PerformanceAnalytics")
library(PerformanceAnalytics)

nums <- data[, c("temperature", "bill", "pizzas")]

chart.Correlation(nums, method = "spearman", histogram = TRUE, pch = 16, main = "Correla
#This correlation chart tells us so much about the numerical covariates and their indepe
#There is a slight negative correlation between temperature and bill, and also temperatu
#for the same reason- more pizzas (higher price also) means a larger order -> longer pre
#There is a pretty strong correlation between bill and pizzas. This makes sense because
#This correlation might cause overemphasis of certain information.

##Categorical covariate independence check
branch.v.driver <- table(data$branch, data$driver)
chisq.test(branch.v.driver)
DescTools::CramerV(branch.v.driver)

#Result:0.1498
#Interpretation: The correlation value indicates a weak positive correlation between the

branch.v.day <- table(data$branch, data$day)
chisq.test(branch.v.day)
DescTools::CramerV(branch.v.day)

#Result:0.0680
#Interpretation: The correlation value indicates an independency between the covariates

day.v.driver <- table(data$day, data$driver)
chisq.test(day.v.driver)
DescTools::CramerV(day.v.driver)

#Result:0.3532
#Interpretation: The correlation value indicates a moderate level of correlation between
```

```r
#Based on these observations of independence or lack there of, we can conclude that ther
#we should consider merging or not including due to correlation. This could possibly hel
#certain information and making a more reliant model.

##Checking for continuous response variable
class(data$time)
#The outcome of this function is "numeric" which most likely indicates a continuous vari

#Overall, the assumptions of normality, independence of residuals, independence of cova
#are unmet, possibly met, unmet, and unmet respectively.

#### 8 ####
#Temperature, Domenico the driver, and the East branch cause the delivery time to be sho
#Whereas the bill and number of pizzas cause the order time to increase (aka delayed).

#### 9 ####
m_quad <- lm(time ~ temperature + I(temperature^2) + branch + day + driver + bill + pizz
m_quad

summary(m_quad)
#This summary shows an increase in the adjusted R-squared which immediately makes this a
#This means that keeping the quadratic transformation of temperature in the model helps
#This quadratic transformation being helpful also allows us to say that the relationship
#(because a curved shape provides a better fit)

anova(m2, m_quad)    # if p < 0.05, the quadratic term significantly improves fit
#This better fit is also supported in the results of the anova test: significantly low p


#### 10 ####
last.captured.delivery <- data[nrow(data), ]
View(last.captured.delivery)

predict(ml, newdata = latest_delivery)

#Result: 34.2296 minutes

#Points of comparison
#Actual time of last delivery: 35.7375

mean(data$time)
#Result: 36.5063 minutes
```

```
#Difference between actual and predicted: 1.5079
#Percent Error: 1.5079/35.7375 = 0.0422 aka 4.22%

#Interpretation: Using our latest model, we predict the time of the latest delivery to b
#There is a percent error of 4.22 percent in this prediction. I am not sure of its signi
#it would not come as a shock to me if this percent error was significant because of the
#low value of the adjusted R-squared of our model.
```