Linear Regression on Cafe Yelp Data

Joseph de Leon

Loading the Dataset

```
cafe_df <- read.csv("cafe_data.csv")
```

Linear Regression Model with 13 chosen parameters

```
cafe_model <- lm (formula = stars ~ review_count + weekly_hours + accepts_credit_cards + price_range + w
        takeout + outdoor_seating + bike_parking + caters + delivery +
        noise_level + valet_parking + other_parking, data = cafe_df)

summary(cafe_model)
```

```
##
## Call:
## lm(formula = stars ~ review_count + weekly_hours + accepts_credit_cards +
##     price_range + wifi + takeout + outdoor_seating + bike_parking +
##     caters + delivery + noise_level + valet_parking + other_parking,
##     data = cafe_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.25250 -0.41591  0.03074  0.41940  1.75498
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.8161903  0.2451638  19.645  < 2e-16 ***
## review_count          0.0003702  0.0001823   2.030 0.042874 *
## weekly_hours         -0.0108822  0.0008514 -12.781  < 2e-16 ***
## accepts_credit_cards -0.1680632  0.1978517  -0.849 0.396060
## price_range           0.0063352  0.0601725   0.105 0.916194
## wifi                 -0.1672627  0.0673890  -2.482 0.013404 *
## takeout              -0.3293370  0.0887796  -3.710 0.000232 ***
## outdoor_seating       0.0610648  0.0659027   0.927 0.354606
## bike_parking          0.2151923  0.0723012   2.976 0.003064 **
## caters                0.0148202  0.0687757   0.215 0.829480
## delivery             -0.0827281  0.0649830  -1.273 0.203608
## noise_level          -0.2221058  0.0781473  -2.842 0.004672 **
## valet_parking        -0.3517840  0.2169386  -1.622 0.105549
## other_parking         0.2294401  0.0777332   2.952 0.003316 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.6247 on 480 degrees of freedom
##   (59 observations deleted due to missingness)
## Multiple R-squared:  0.4672, Adjusted R-squared:  0.4528
## F-statistic: 32.38 on 13 and 480 DF,  p-value: < 2.2e-16
```

Dealing with Review Count

Review count is obviously not going to linearly predict the rating, so we tested a log transformation, square root transformation, binning approach, categorical approach, and polynomial approach to deal with this parameter. We went with a log transformation as this resulted in the greatest adjusted R squared value.

```
# 1. Log transformation of review_count
cafe_df$log_review_count <- log(cafe_df$review_count + 1)  # Adding 1 to handle any zeros

# 2. Square root transformation
cafe_df$sqrt_review_count <- sqrt(cafe_df$review_count)

# 3. Binning Approach
cafe_df$review_count_cat <- cut(cafe_df$review_count,
                               breaks = c(0, 50, 200, 500, Inf),
                               labels = c("very_few", "few", "moderate", "many"),
                               include.lowest = TRUE)

# 4. Categorical Approach
cafe_df$hours_category <- cut(cafe_df$weekly_hours,
                             breaks = c(0, 40, 70, Inf),
                             labels = c("low_hours", "avg_hours", "high_hours"),
                             include.lowest = TRUE)

cafe_df$hours_category <- relevel(cafe_df$hours_category, ref = "low_hours")

# Run models with different transformations
# Log transformation model
model_log <- lm(formula = stars ~ log_review_count + weekly_hours + accepts_credit_cards + price_range
        takeout + outdoor_seating + bike_parking + caters + delivery +
        noise_level + valet_parking + other_parking, data = cafe_df)

# Square root transformation model
model_sqrt <- lm(formula = stars ~ sqrt_review_count + weekly_hours + accepts_credit_cards + price_rang
        takeout + outdoor_seating + bike_parking + caters + delivery +
        noise_level + valet_parking + other_parking, data = cafe_df)

# Categorical review count model
model_cat_reviews <- lm(formula = stars ~ review_count_cat + weekly_hours + accepts_credit_cards + pric
        takeout + outdoor_seating + bike_parking + caters + delivery +
        noise_level + valet_parking + other_parking, data = cafe_df)

# Polynomial model
model_poly <- lm(formula = stars ~ review_count + I(review_count^2) + weekly_hours + accepts_credit_car
        takeout + outdoor_seating + bike_parking + caters + delivery +
        noise_level + valet_parking + other_parking, data = cafe_df)
```

```r
# Model with Categorical Hours
model_cat_hours <- lm(formula = stars ~ review_count + weekly_hours + accepts_credit_cards + price_range
        takeout + outdoor_seating + bike_parking + caters + delivery +
        noise_level + valet_parking + other_parking, data = cafe_df)

# Compare models using adjusted R-squared and AIC
models <- list(cat_hours = model_cat_hours, log = model_log, sqrt = model_sqrt,
               cat_reviews = model_cat_reviews, poly = model_poly)

# Comparison table sorted by desending adjusted R-squared
comparison <- data.frame(
  Model = names(models),
  Adj_R_squared = sapply(models, function(m) summary(m)$adj.r.squared),
  AIC = sapply(models, AIC)
)

comparison <- comparison[order(-comparison$Adj_R_squared),]
print(comparison)
```

```
##                   Model Adj_R_squared      AIC
## log                 log     0.4550750 950.8013
## sqrt               sqrt     0.4547800 951.0686
## poly               poly     0.4534068 953.2810
## cat_hours     cat_hours     0.4527724 952.8842
## cat_reviews cat_reviews     0.4505174 956.8531
```

```r
summary(models[[comparison$Model[1]]])
```

```
##
## Call:
## lm(formula = stars ~ log_review_count + weekly_hours + accepts_credit_cards +
##     price_range + wifi + takeout + outdoor_seating + bike_parking +
##     caters + delivery + noise_level + valet_parking + other_parking,
##     data = cafe_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.18997 -0.39987  0.03634  0.41212  1.67948
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.6375086  0.2474317  18.743  < 2e-16 ***
## log_review_count      0.0713841  0.0287432   2.484 0.013350 *
## weekly_hours         -0.0108006  0.0008505 -12.699  < 2e-16 ***
## accepts_credit_cards -0.1673895  0.1973297  -0.848 0.396709
## price_range           0.0046723  0.0595916   0.078 0.937538
## wifi                 -0.1797846  0.0670185  -2.683 0.007557 **
## takeout              -0.3294538  0.0885876  -3.719 0.000224 ***
## outdoor_seating       0.0568705  0.0658084   0.864 0.387919
## bike_parking          0.1894755  0.0737517   2.569 0.010497 *
## caters               -0.0019455  0.0690382  -0.028 0.977530
## delivery             -0.0836294  0.0643181  -1.300 0.194141
## noise_level          -0.2258804  0.0779574  -2.897 0.003934 **
## valet_parking        -0.3904181  0.2178876  -1.792 0.073789 .
## other_parking         0.1867472  0.0813930   2.294 0.022199 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.6234 on 480 degrees of freedom
##   (59 observations deleted due to missingness)
## Multiple R-squared:  0.4694, Adjusted R-squared:  0.4551
## F-statistic: 32.67 on 13 and 480 DF,  p-value: < 2.2e-16
```

Dealing with Categorical Parameters

We take the log of weekly hours as it is right skewed. We then convert it into a categorical variable: low, average, and high. The average category is the middle 50% of weekly hours.

Price range is also converted to a factor as it is categorical. Noise level is also converted to a factor with values 0-1 being low noise and 2-3 being high noise.

```r
# check distribution
summary(cafe_df$weekly_hours)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.50   47.00   66.00   75.55   98.00  168.00
```

```r
# Take log of weekly hours
cafe_df$log_weekly_hours <- log(cafe_df$weekly_hours)
summary(cafe_df$log_weekly_hours)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.6931  3.8501  4.1897  4.1782  4.5850  5.1240
```

```r
# Convert log weekly hours to categorical: low, average, high based on middle 50%
cafe_df$log_hours_category <- cut(cafe_df$log_weekly_hours,
                             breaks = c(0, 3.8501, 4.5850, Inf),
                             labels = c("low_hours", "avg_hours", "high_hours"),
                             include.lowest = TRUE)
cafe_df$log_hours_category <- relevel(cafe_df$log_hours_category, ref = "low_hours")

# Convert price_range to a factor since it's categorical (1,2,3,4)
cafe_df$price_range <- as.factor(cafe_df$price_range)
cafe_df$price_range <- relevel(cafe_df$price_range, ref = "1")

# Noise groups instead of original factor (0, 1, 2, 3)
cafe_df$noise_level <- as.character(cafe_df$noise_level)
# Create a new factor with two levels - "low_noise" for 0 to 1; "high_noise" for 2 and 3
cafe_df$noise_group <- factor(
  ifelse(cafe_df$noise_level %in% c("0", "0.863636363636364", "1"), "low_noise", "high_noise"),
  levels = c("low_noise", "high_noise")
)
cafe_df$noise_group <- relevel(cafe_df$noise_group, ref = "low_noise")
```

```r
# New model with above changes
new_cafe_model <- lm(formula = stars ~ log_review_count + log_hours_category +
                    accepts_credit_cards + price_range + wifi + takeout +
                    outdoor_seating + bike_parking + caters + delivery +
                    noise_group + valet_parking + other_parking,
                    data = cafe_df)
summary(new_cafe_model)
```

```
## 
## Call:
## lm(formula = stars ~ log_review_count + log_hours_category + 
##     accepts_credit_cards + price_range + wifi + takeout + outdoor_seating + 
##     bike_parking + caters + delivery + noise_group + valet_parking + 
##     other_parking, data = cafe_df)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.04405 -0.41949  0.02084  0.41751  1.57921 
## 
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                  4.151444   0.220745  18.807  < 2e-16 ***
## log_review_count             0.091921   0.028648   3.209 0.001423 ** 
## log_hours_categoryavg_hours -0.298909   0.073520  -4.066  5.6e-05 ***
## log_hours_categoryhigh_hours -1.162235   0.091054 -12.764  < 2e-16 ***
## accepts_credit_cards        -0.306081   0.197945  -1.546 0.122698    
## price_range2                 0.008742   0.062511   0.140 0.888840    
## price_range3                -0.318830   0.326564  -0.976 0.329403    
## wifi                        -0.179023   0.068482  -2.614 0.009227 ** 
## takeout                     -0.299111   0.088777  -3.369 0.000815 ***
## outdoor_seating              0.089325   0.066650   1.340 0.180814    
## bike_parking                 0.150339   0.074301   2.023 0.043591 *  
## caters                       0.019583   0.068931   0.284 0.776457    
## delivery                    -0.168553   0.063175  -2.668 0.007890 ** 
## noise_grouphigh_noise       -0.374381   0.174440  -2.146 0.032361 *  
## valet_parking               -0.484068   0.218208  -2.218 0.026999 *  
## other_parking                0.169284   0.081899   2.067 0.039276 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.6255 on 477 degrees of freedom
##   (60 observations deleted due to missingness)
## Multiple R-squared:  0.468,  Adjusted R-squared:  0.4513 
## F-statistic: 27.97 on 15 and 477 DF,  p-value: < 2.2e-16
```

Excluding Insignificant Parameters

We removed parameters in which its coefficient p-value is greater than 0.05.

```r
refined_cafe_model <- lm(formula = stars ~ log_review_count + log_hours_category +
                    wifi + takeout + bike_parking + delivery + noise_group
                    + valet_parking + other_parking, data = cafe_df)
```

```
summary(refined_cafe_model)
```

```
##
## Call:
## lm(formula = stars ~ log_review_count + log_hours_category +
##     wifi + takeout + bike_parking + delivery + noise_group +
##     valet_parking + other_parking, data = cafe_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9888 -0.4097  0.0416  0.4547  1.4305
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     4.10219    0.11930  34.386  < 2e-16 ***
## log_review_count                0.07604    0.02630   2.891  0.00399 **
## log_hours_categoryavg_hours    -0.32178    0.06868  -4.685 3.55e-06 ***
## log_hours_categoryhigh_hours   -1.28153    0.08066 -15.887  < 2e-16 ***
## wifi                           -0.12805    0.06267  -2.043  0.04151 *
## takeout                        -0.34717    0.07981  -4.350 1.63e-05 ***
## bike_parking                    0.11656    0.06768   1.722  0.08557 .
## delivery                       -0.15993    0.05824  -2.746  0.00623 **
## noise_grouphigh_noise          -0.43177    0.17477  -2.471  0.01380 *
## valet_parking                  -0.32590    0.18893  -1.725  0.08509 .
## other_parking                   0.15425    0.07514   2.053  0.04057 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6384 on 541 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4562, Adjusted R-squared:  0.4461
## F-statistic: 45.38 on 10 and 541 DF,  p-value: < 2.2e-16
```

Excluding Insignificant Parameters 2

We removed parameters in which its coefficient p-value is greater than 0.05 once again. Now all parameters are significant at our 0.05 significance level.

```
# Excluding insignificant parameters again (p-value > 0.05)

refined_cafe_model2 <- lm(formula = stars ~ log_review_count + log_hours_category + wifi +
                      takeout + delivery + noise_group + other_parking, data = cafe_df)

summary(refined_cafe_model2)
```

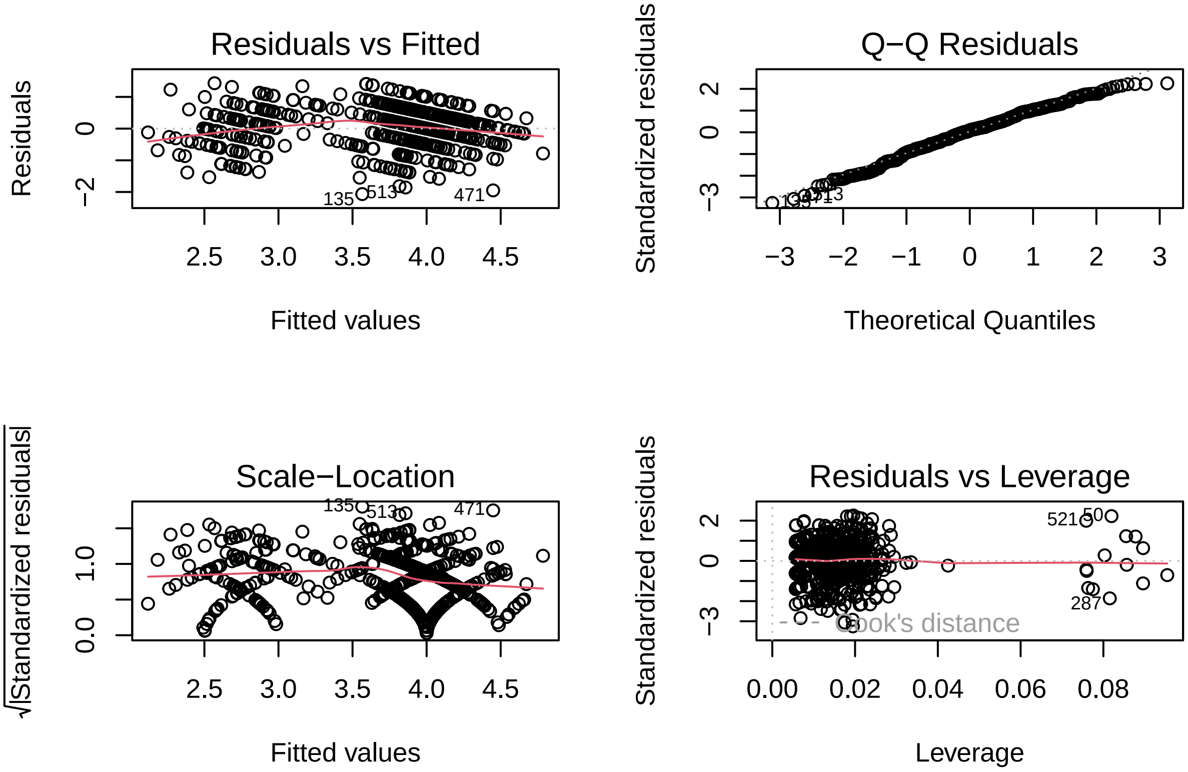
```
##
## Call:
## lm(formula = stars ~ log_review_count + log_hours_category +
##     wifi + takeout + delivery + noise_group + other_parking,
##     data = cafe_df)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.06572 -0.40411  0.05769  0.44942  1.43143
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   4.10604    0.11976  34.286  < 2e-16 ***
## log_review_count              0.08497    0.02540   3.346 0.000877 ***
## log_hours_categoryavg_hours  -0.30843    0.06874  -4.487 8.82e-06 ***
## log_hours_categoryhigh_hours -1.28278    0.08082 -15.871  < 2e-16 ***
## wifi                         -0.13133    0.06291  -2.088 0.037283 *
## takeout                      -0.31231    0.07775  -4.017 6.72e-05 ***
## delivery                     -0.16032    0.05824  -2.752 0.006112 **
## noise_grouphigh_noise        -0.41734    0.17542  -2.379 0.017699 *
## other_parking                 0.15839    0.07532   2.103 0.035951 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6412 on 543 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4495, Adjusted R-squared:  0.4414
## F-statistic: 55.42 on 8 and 543 DF,  p-value: < 2.2e-16
```

Plotting the Model

We plot the model to check assumptions and outliers. Although assumptions are not fully satisfied, we remove outliers to increase the accuracy of our linear regression model.

```r
par(mfrow=c(2,2))
plot(refined_cafe_model2)
```



Removing the Outliers

Here is the final linear regression model after removing the outliers from our data. This is our highest Adjusted R-sqaured value of all the previous regression models, even though it can use more improvement.

```r
# Outliers from the plots
outliers <- c(135, 513, 471, 50, 521, 287)

# New data set without outliers
cafe_df_clean <- cafe_df[-outliers, ]

# New model without outliers
improved_model <- lm(formula = stars ~ log_review_count + log_hours_category +
    wifi + takeout + delivery + noise_group,
    data = cafe_df_clean)

summary(improved_model)
```

```
##
## Call:
## lm(formula = stars ~ log_review_count + log_hours_category +
##     wifi + takeout + delivery + noise_group, data = cafe_df_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.78999 -0.40450  0.04607  0.43698  1.41377
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  4.16375    0.11654  35.729  < 2e-16 ***
## log_review_count             0.10463    0.02224   4.704 3.25e-06 ***
## log_hours_categoryavg_hours  -0.32601    0.06730  -4.844 1.67e-06 ***
## log_hours_categoryhigh_hours -1.31976    0.07915 -16.675  < 2e-16 ***
## wifi                        -0.14109    0.06103  -2.312 0.021158 *
## takeout                     -0.28161    0.07532  -3.739 0.000204 ***
## delivery                    -0.17195    0.05658  -3.039 0.002489 **
## noise_grouphigh_noise       -0.53736    0.19041  -2.822 0.004946 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.6212 on 538 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4723, Adjusted R-squared:  0.4654
## F-statistic: 68.78 on 7 and 538 DF,  p-value: < 2.2e-16
```

The Linear Regression Equation

$$\begin{aligned}&Rating_i = \beta_0 + \beta_1 \cdot \log(ReviewCount) + \beta_2 \cdot I(Hours = average) + \beta_3 \cdot I(Hours = high)\\&+ \beta_4 \cdot I(Wifi = True) + \beta_5 \cdot I(Takeout = True) + \beta_6 \cdot I(Delivery = True) + \beta_7 \cdot I(Noise = High)\end{aligned}$$

With the actual values:

$$\begin{aligned}&Rating_i = 4.16375 + 0.10463 \cdot \log(ReviewCount) - 0.32601 \cdot I(Hours = average) - 1.31976 \cdot I(Hours = high)\\&- 0.14109 \cdot I(Wifi = True) - 0.28161 \cdot I(Takeout = True) - 0.17195 \cdot I(Delivery = True) - 0.53736 \cdot I(Noise = High)\end{aligned}$$