605 Wk 12 Discussion - Predicting Diamond Price

Jen Abinette

2023-04-20

Assignment

Build a multiple regression model for data that interests you. Include in this model at least one quadratic term, one dichotomous term, and one dichotomous vs. quantitative interaction term. Interpret all coefficients. Conduct residual analysis. Was the linear model appropriate? Why or why not?

Dichotomous variable

```
# New data frame that remove cases where cut is Fair
df <- subset (diamonds, cut != 'Fair')
# Create dichotomous variable with Good and Very Good as 0 and Premium and Ideal as 1
df$cut_dich <- ifelse(grepl("Good",df$cut),0,1)
# Quadratic variable
df$carat.sq <- df$carat**2
```

Regression Modeling

Research Question - How can we predict diamond price?

```
df2.lm <- lm(price ~ carat + carat.sq + cut_dich + carat*cut_dich, data=df)
summary(df2.lm)
```

```
##
## Call:
## lm(formula = price ~ carat + carat.sq + cut_dich + carat * cut_dich,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20868.3   -677.6    -39.6    397.6  13054.7
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1766.02      29.56 -59.739  < 2e-16 ***
## carat           6170.55      51.97 118.724  < 2e-16 ***
## carat.sq         772.89      21.79  35.471  < 2e-16 ***
## cut_dich         120.44      27.77   4.337 1.45e-05 ***
## carat:cut_dich   140.87      29.85   4.719 2.38e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1483 on 52325 degrees of freedom
## Multiple R-squared:  0.8627, Adjusted R-squared:  0.8626
## F-statistic: 8.216e+04 on 4 and 52325 DF,  p-value: < 2.2e-16
```

What about using squareroot of carat instead of carat squared?

```
df.lm <- lm(price ~ carat + sqrt(carat) + cut_dich + carat*cut_dich, data=df)
summary(df.lm)
```

```
##
## Call:
## lm(formula = price ~ carat + sqrt(carat) + cut_dich + carat *
##     cut_dich, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19591.9   -559.5    -38.7    279.7  13153.6
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1740.52      77.48  22.464  < 2e-16 ***
## carat          12836.40      93.21 137.720  < 2e-16 ***
## sqrt(carat)    -9513.59     169.59 -56.096  < 2e-16 ***
## cut_dich         108.50      27.29   3.975 7.04e-05 ***
## carat:cut_dich   135.39      29.34   4.614 3.95e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1457 on 52325 degrees of freedom
## Multiple R-squared:  0.8673, Adjusted R-squared:  0.8673
## F-statistic: 8.552e+04 on 4 and 52325 DF,  p-value: < 2.2e-16
```

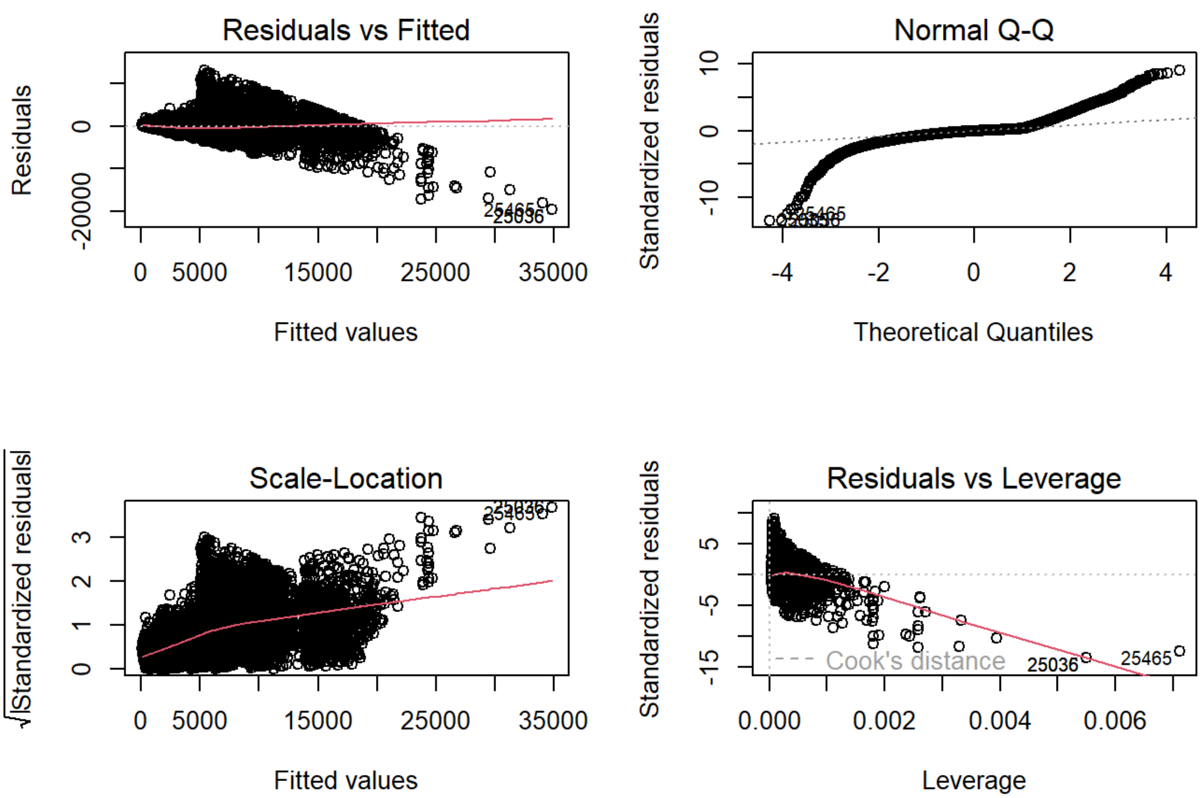
$$\hat{price} = -1740.52 + 12836.40 \times carat - 9513.59 \times \sqrt{carat} + 108.5 \times cut.dich + 135.39 \times carat \times cut.dich$$

Since the effect size is slightly larger and the residuals more normally distributed using the square root instead of quadratic version of carat, I chose to move forward with that model. Based on the summary of the model above, the coefficients for the model are all statistically significant with p-values less than .001 and minimal variability given the standard error for the intercept and slopes are much smaller than the coefficients. Additionally, Multiple R-squared indicates our model predicts 87% of the variability in diamond price.

Residual Analysis

To assess whether the linear model is reliable, we need to check for (1) linearity, (2) nearly normal residuals, and (3) constant variability

```
par(mfrow=c(2,2))
plot(df.lm)
```



An analysis of the residuals plots indicates a problem with our model assumptions. The plots above show a lack of constant variability and normal distribution. In particular, the residuals are not uniformly scattered above and below zero on the residuals plot and the 2 ends diverge on the Q-Q plot. Although our model has a large effect size, it is still not the best model as it is not fully explaining the variability in price.