STAT 101A HW4

Brandon Montez

February 2023

1

```
legos <- read.table("Legos.txt", header = TRUE)

lm1 <- lm(price ~ bricks, data = legos)

anova(lm1)

Analysis of Variance Table

Response: price
           Df Sum Sq Mean Sq F value    Pr(>F)
bricks      1 460348  460348  2933.7 < 2.2e-16 ***
Residuals 155  24322     157
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

2

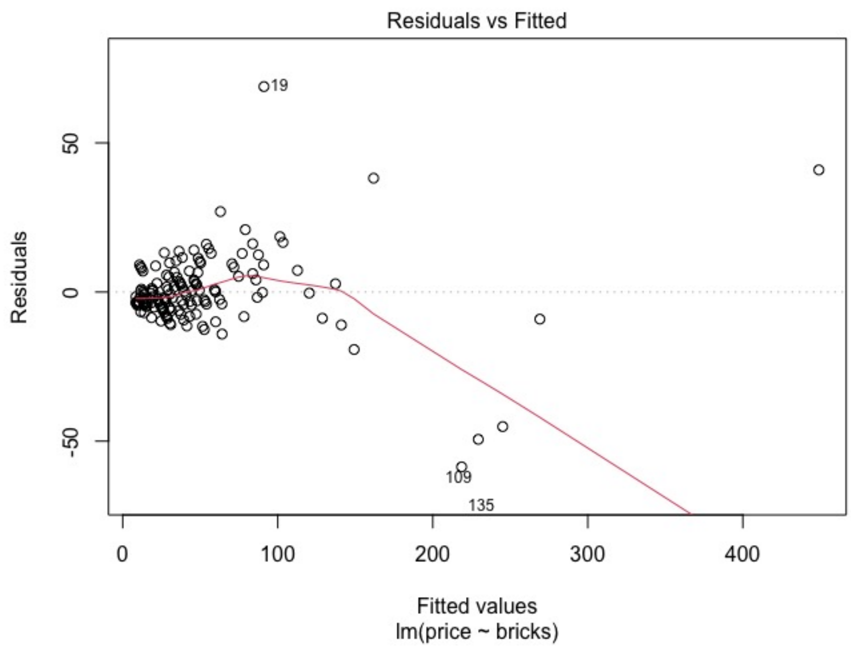
Null hypothesis: The mean sales price of Legos is the same across all levels of the number of bricks in the set. Alternative hypothesis: The mean sales price of Legos is different across at least one level of the number of bricks in the set.

3

Based on the ANOVA summary, the p-value is less than 0.05, which suggests evidence to reject the null hypothesis that there is no relationship between the number of bricks in a Lego set and its price. Hence, we can conclude that the number of bricks in a Lego set is a significant predictor of its price.

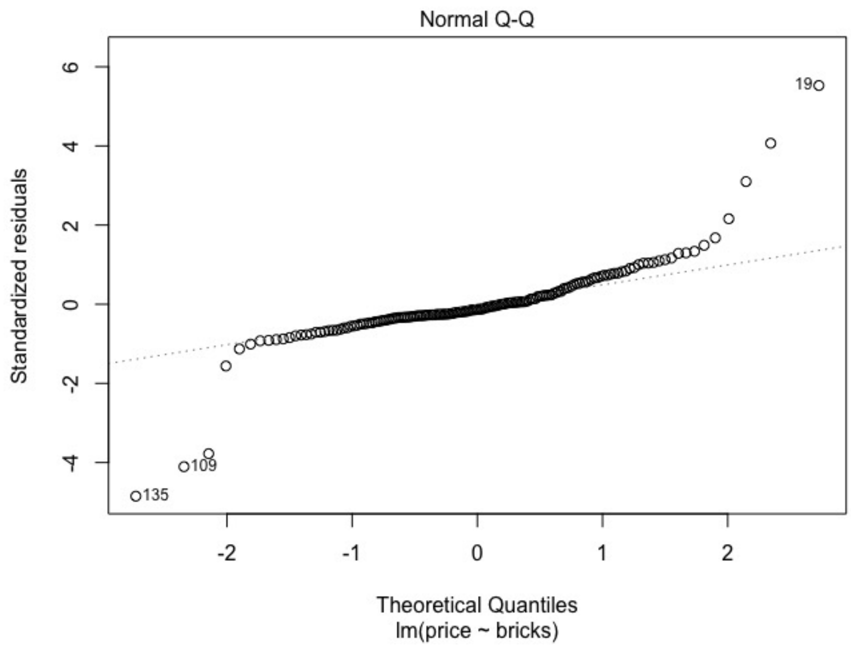
4

```
# Residuals vs Fitted plot
plot(lm1, which = 1)
```



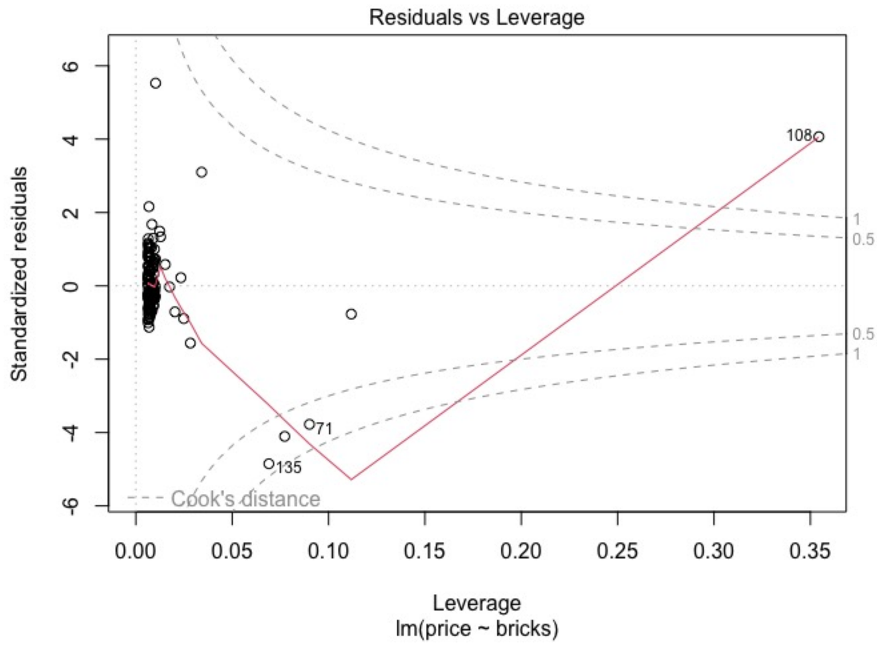
The residuals appear to have random and equal variance for sets predicted to cost less than \$200. The variability of residuals increases with the price and doesn't appear to be normally distributed across the entire sample.

```
# Normal QQ plot
plot(lm1, which = 2)
```



The QQ plot appears to be heavy-tailed, with a relatively fitted line except for some edge cases.

```
# Residuals vs Leverage plot
plot(lm1, which = 5)
```



108 appears to be a significant observation or an outlier. 71 and 135 also appear to hold moderate significance. Another observation falls near but is not labeled.

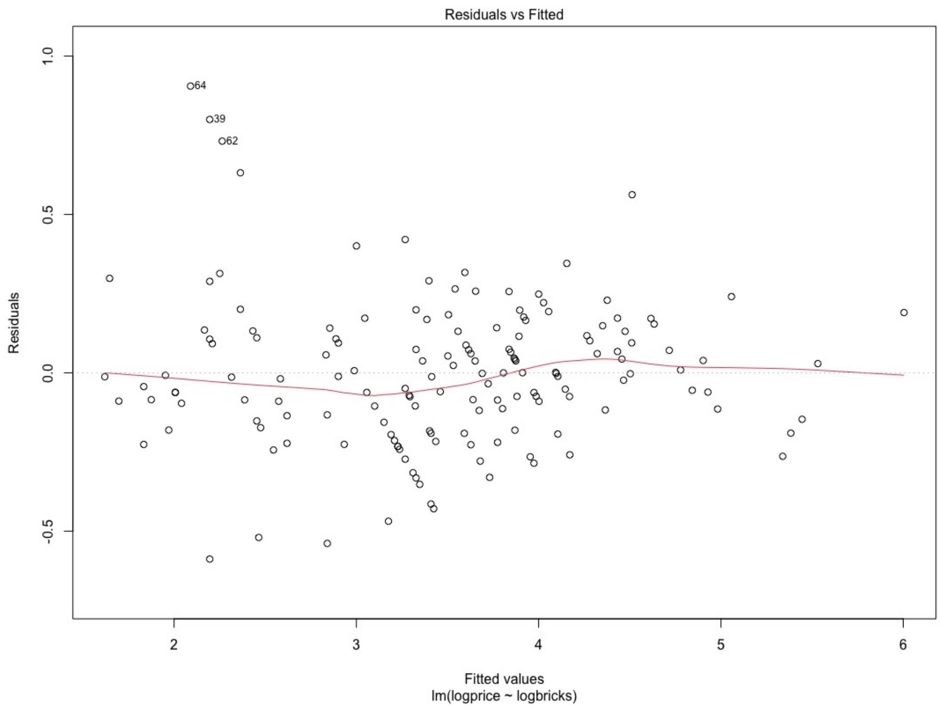
5

```
legos$logbricks <- log(legos$bricks)
legos$logprice <- log(legos$price)
```

6

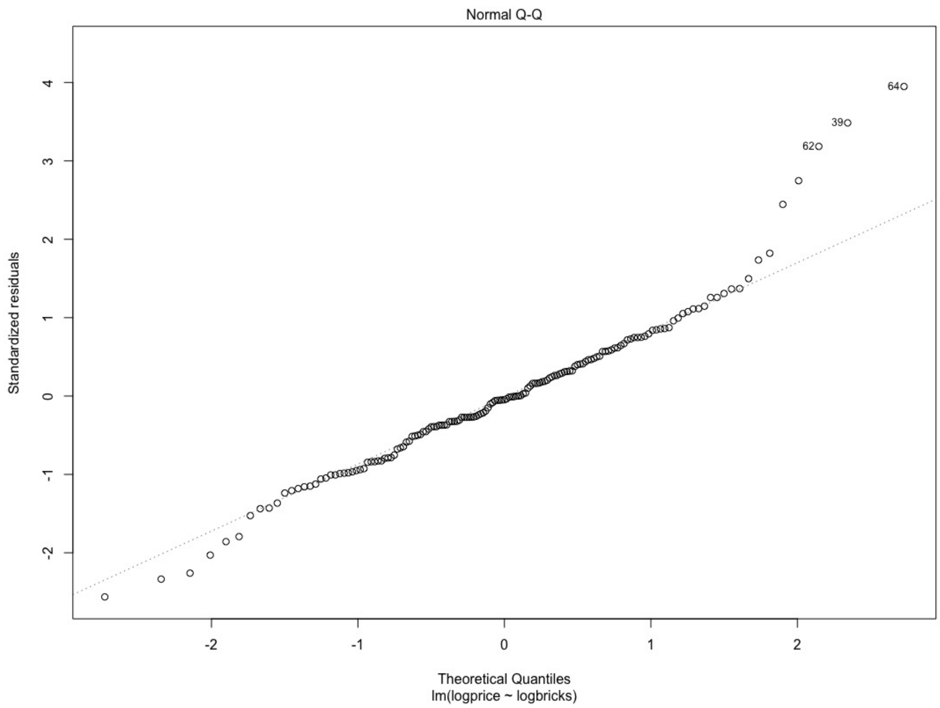
```
lm2 <- lm(logprice ~ logbricks, data = legos)

# Residuals vs Fitted Plot
plot(lm2, which = 1)
```



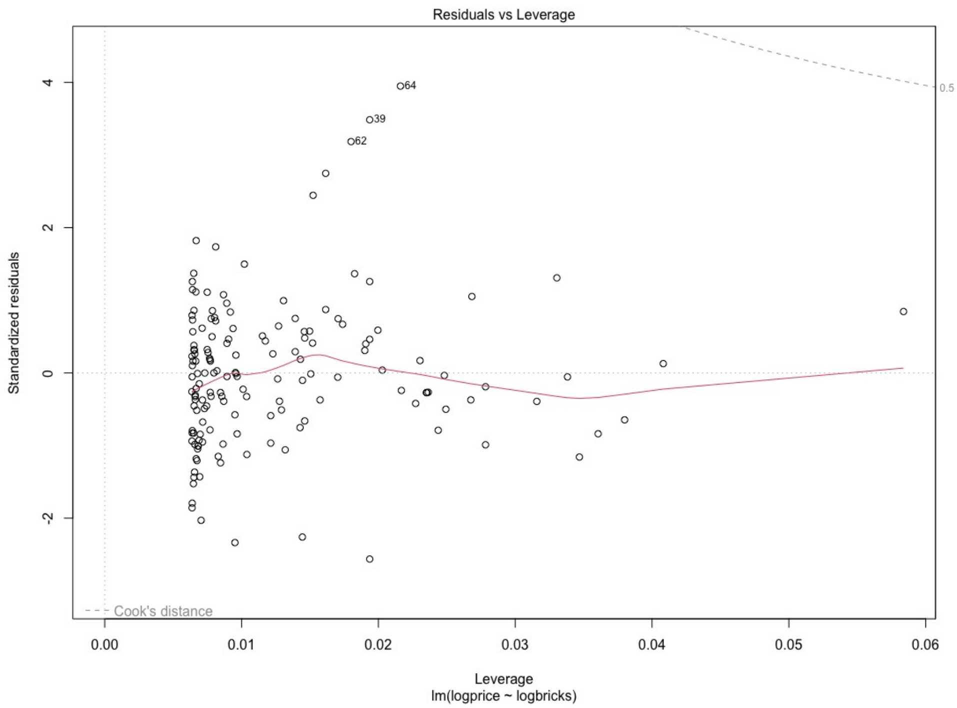
The observations now appear to be normal and randomly distributed across the log-transformed data set.

```
# Normal QQ Plot
plot(lm2, which = 2)
```



The normal QQ plot appears to be soft-tailed, and suggests that the log transformation has improved the normality of the residuals.

```
# Residuals vs Leverage Plot
plot(lm2, which = 5)
```



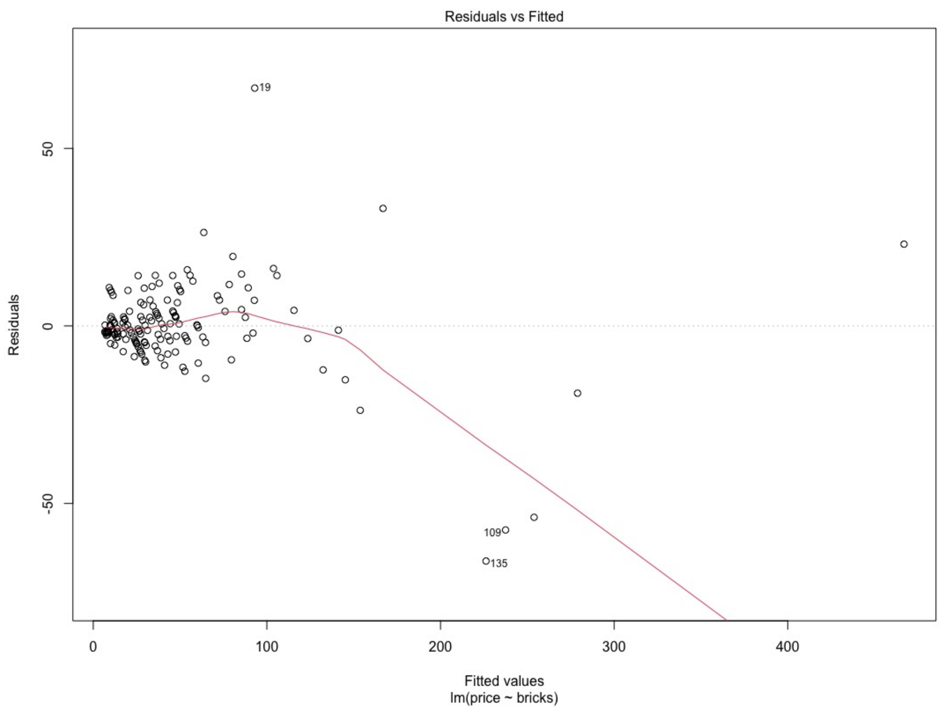
There no longer appears to be any significant outliers or heavily leveraged observations.

7

```
lm3 <- lm(price ~ bricks, data = legos, weight = 1 / bricks)
```

8

```
plot(lm3, which = 1)
```



9

It does not appear that the weighted model changed the fit, but the log transformation greatly effected the heteroscedasticity of our model. This seems intuitive because the price of lego sets within our data seems to follow a logarithmic distribution - there are much more sets at a cheaper price than at a higher price.

10

```
library(readr)
armspans <- read_csv("armspans2022_gender.csv")

lm4 <- lm(armspan ~ is.female, data = armspans)

summary(lm4)

Call:
lm(formula = armspan ~ is.female, data = armspans)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-9.7586 -2.0248  0.2414  2.2414  8.2414

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  69.7586     0.7399  94.284  < 2e-16 ***
is.female    -7.7338     1.2408  -6.233 1.68e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.984 on 43 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4746,Adjusted R-squared:  0.4624
F-statistic: 38.85 on 1 and 43 DF,  p-value: 1.676e-07
```

69.7586 - 7.7338 * (is.female = 1)

11

The average armspan of males is 69.7586.

12

The average armspan of females is 7.7338 less than males, about 62.0248.