STA471 - Homework 8

Richard McCormick

2023-11-24

Problem B. For the given set of response data, find the best transformation. Perform all the usual regression analysis for your best Lambda, including examination of residuals.

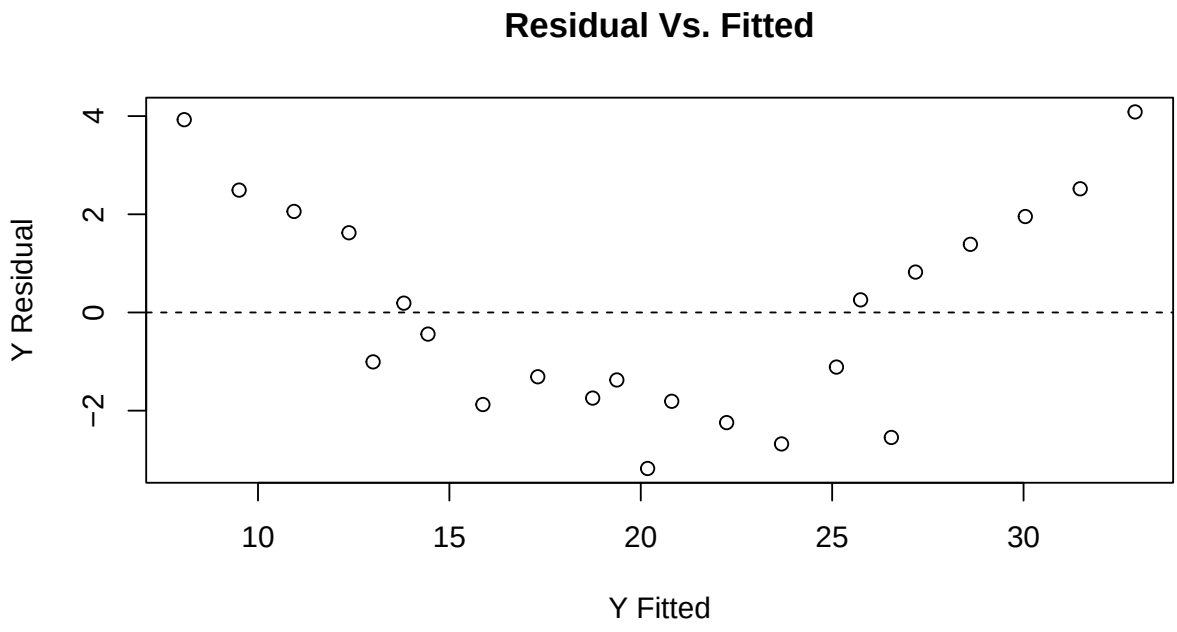
```
# Data Entry
p <- c( 0, 10, 20,
        0, 10, 20, 30,
        0, 10, 20, 30,
        0, 10, 20, 30,
        0, 10, 20, 30,
        0, 10, 20, 30 )

f <- c( 0, 0, 0,
        12, 12, 12, 12,
        24, 24, 24, 24,
        36, 36, 36, 36,
        48, 48, 48, 48,
        60, 60, 60, 60 )

Y <- c( 26, 18, 12,
        28, 19, 14, 12,
        30, 20, 14, 12,
        32, 21, 16, 13,
        34, 24, 17, 14,
        37, 24, 17, 14 )

# Creating a dataframe from the data.
hw8.data <- data.frame( p, f, Y )
# Creating a fitted linear model from the data.
model <- lm( Y ~ p + f, data=hw8.data )
```

I. Run regression analysis on non-transformed model.

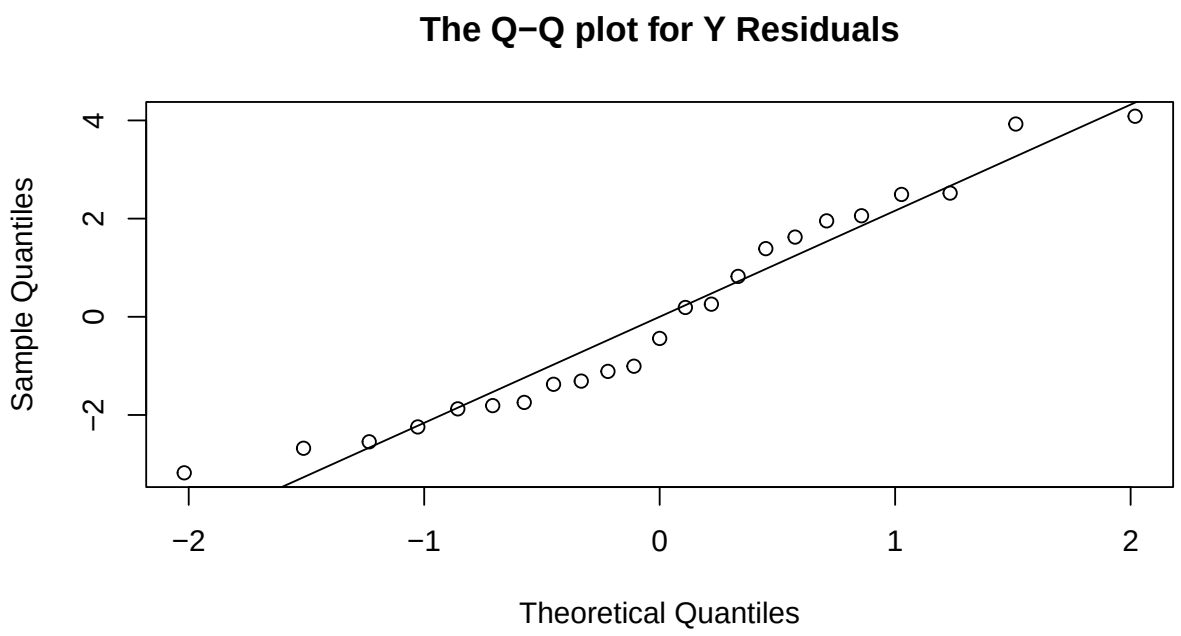
```
# Model Summary
summary( model )
```

```
##
## Call:
## lm(formula = Y ~ p + f, data = hw8.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1784 -1.7766 -0.4411  1.7891  4.0855
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.74288    1.03842   24.79  < 2e-16 ***
## p           -0.63681    0.04329  -14.71 3.45e-12 ***
## f            0.11953    0.02381    5.02 6.57e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.267 on 20 degrees of freedom
## Multiple R-squared:  0.9199, Adjusted R-squared:  0.9119
## F-statistic: 114.9 on 2 and 20 DF,  p-value: 1.083e-11
```

```
# Residual Plot of Model Residuals
plot( fitted( model ), resid( model ), xlab='Y Fitted', ylab='Y Residual', main='Residual Vs. Fitted')
abline(0, 0, lty = 2)
```



```r
# Q-Q Plot of Model Residuals
qqnorm( resid( model ), main = "The Q-Q plot for Y Residuals" )
abline( mean( resid( model ) ), sd( resid( model ) ) )
```

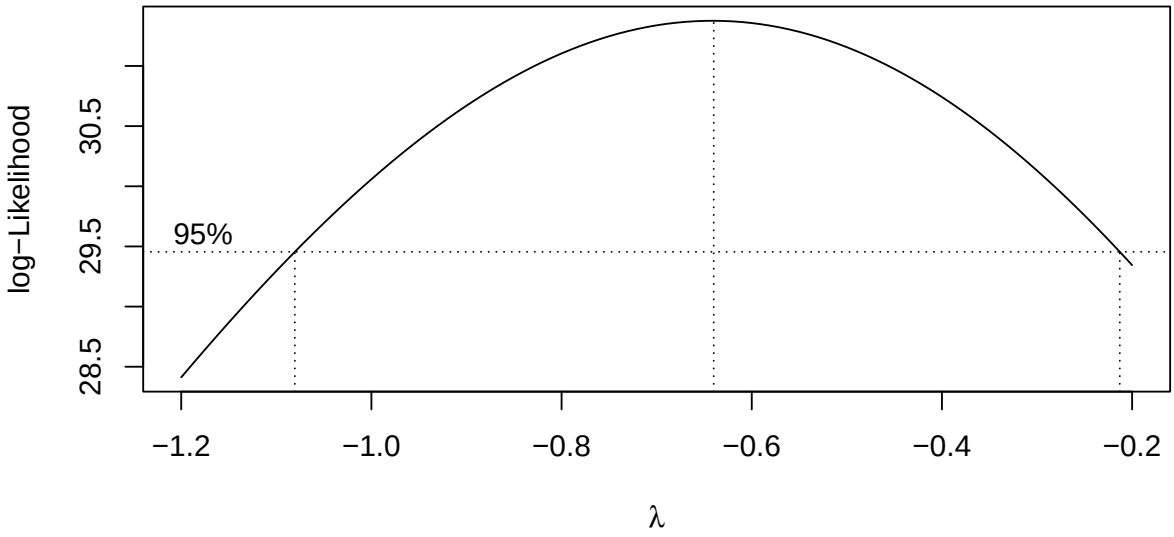



```r
# Shapiro-Wilk Test
shapiro.test( resid( model ) )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.94216, p-value = 0.1999
```

II. Run Box-Cox analysis and transformations.

```
# Plot Box-Cox analysis
boxcox( model, plotit=T, lambda=seq( -1.2, -0.2, by=0.01 ) )
```




```
z <- boxcox( model, plotit=F, lambda=seq(-1.2, -0.2,by=0.01))
lambda <- z$x[ which( z$y == max( z$y ) ) ]
best.model <- lm( Y^( lambda ) ~ p + f, data=hw8.data )
summary( best.model )
```

```
##
## Call:
## lm(formula = Y^(lambda) ~ p + f, data = hw8.data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0077648 -0.0030336 -0.0006629  0.0020281  0.0154624
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.295e-01  2.537e-03  51.020  < 2e-16 ***
## p            2.947e-03  1.058e-04  27.857  < 2e-16 ***
## f           -5.290e-04  5.818e-05  -9.092 1.53e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.005538 on 20 degrees of freedom
## Multiple R-squared:  0.9761, Adjusted R-squared:  0.9738
## F-statistic: 409.3 on 2 and 20 DF,  p-value: < 2.2e-16
```

```r
print( paste( "The best Lambda value is:", lambda ) )
```

```
## [1] "The best Lambda value is: -0.64"
```

III. Regression Analysis & Model Comparison

1. P-Value

The p-value of the transformed model is 2.2 * 10^{-16}. Compared to the original model's p-value of 1.083 * 10^{-11}, the new model has a much lower (better) p-value.

2. R^2 Value

The R^2 value of the new model is 0.9761, which shows an extremely strong positive correlation. This value is much higher than the original model's R^2 value of 0.9199, indicating that much more variation is explained by the new model.

3. R_A^2 Value

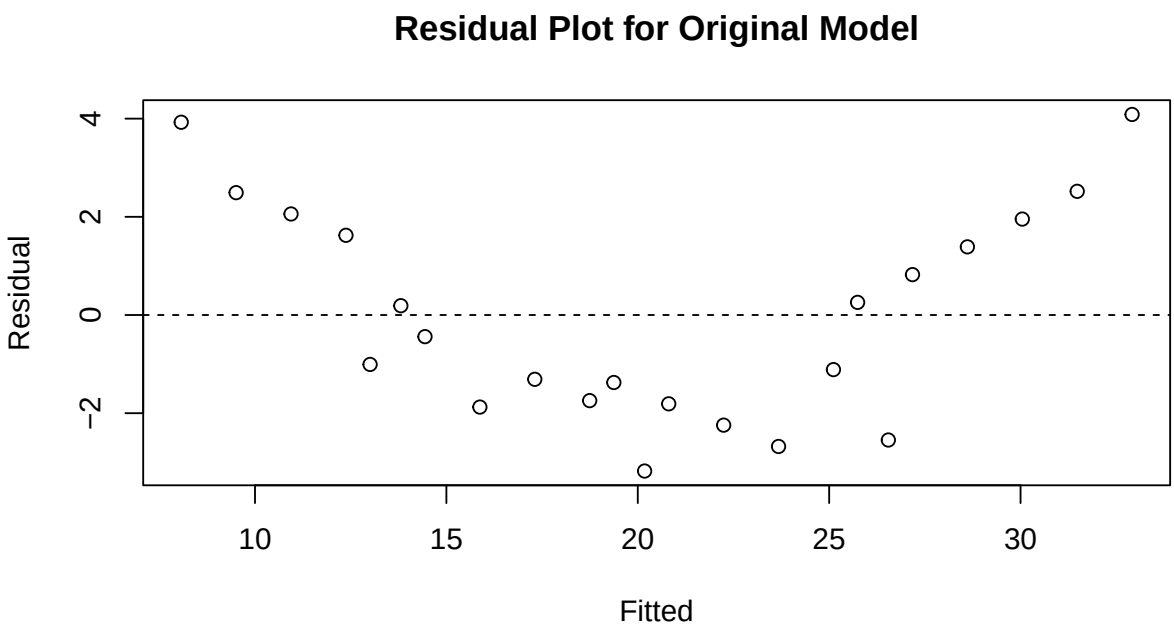
The R_A^2 value of the new model is 0.9738, which is also significantly higher than the original model's R_A^2 value of 0.9119. This indicates that the new model shows a stronger positive correlation than the old model.

4. Residual Standard Error (S Value)

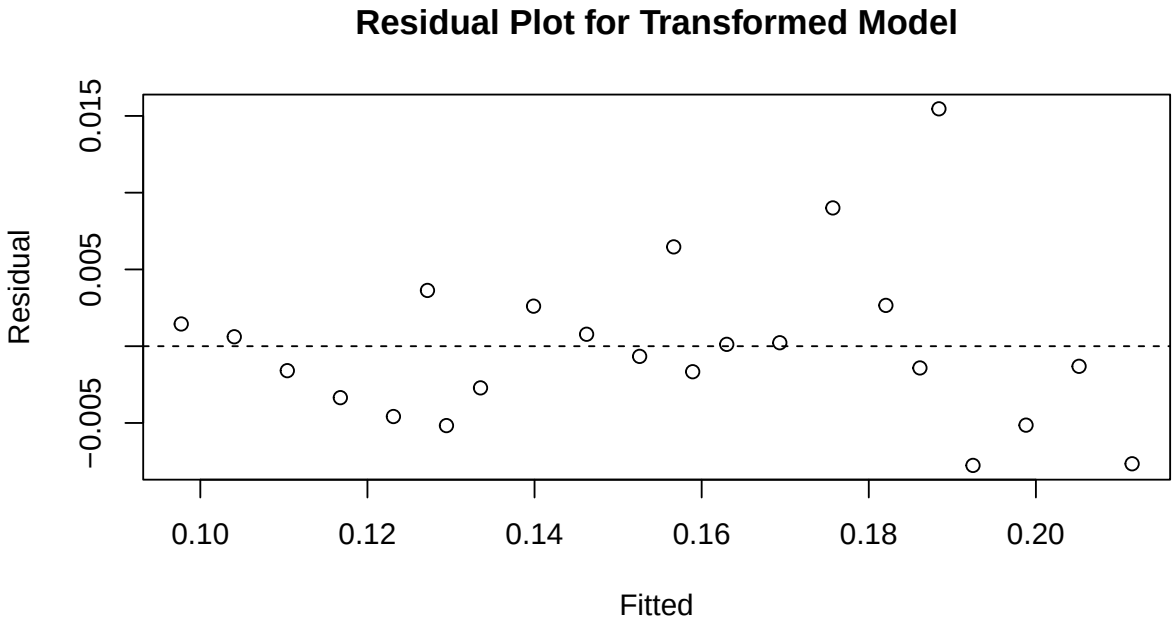
The S value of the new model is 0.005538, compared to the old model's value of 2.267. This shows a **significant** decrease in standard error after the Box Cox transformation has been applied.

5. Residual Scatter Plots

```r
plot( fitted( model ), resid( model ), xlab='Fitted', ylab='Residual', main='Residual Plot for Original
abline(0, 0, lty = 2)
```



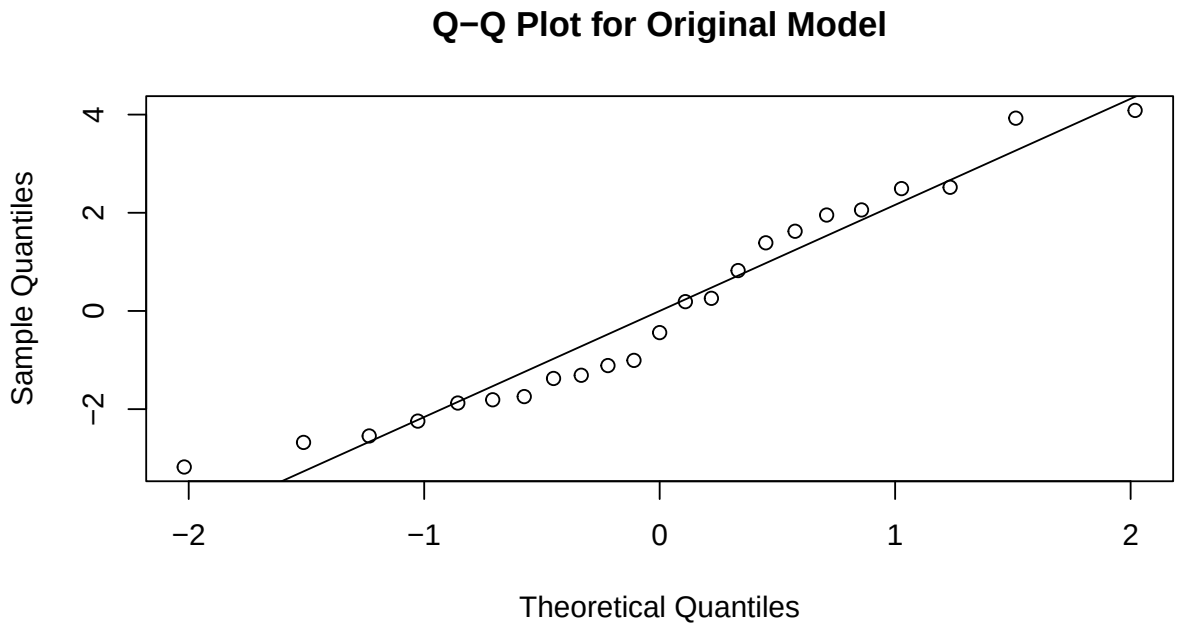

```
plot( fitted( best.model ), resid( best.model ), xlab='Fitted', ylab='Residual', main='Residual Plot fo
abline(0, 0, lty = 2)
```



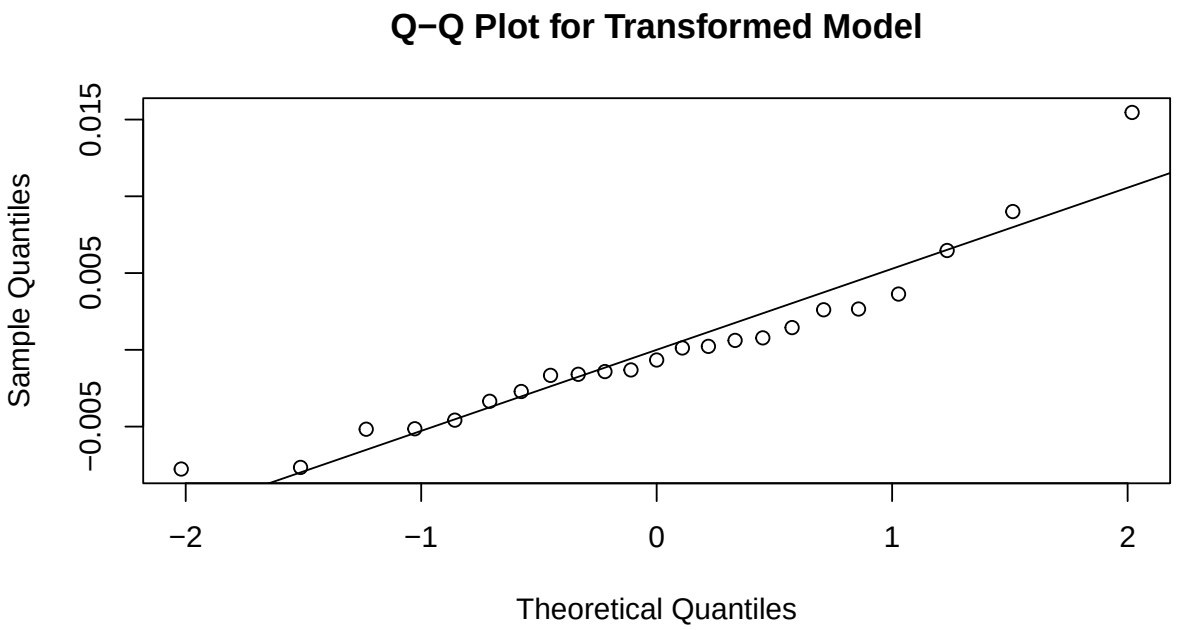
The new, transformed model clearly shows a more evenly distributed plot of residuals, compared to the U-shaped distribution which was prevelant in the original model. As such, it is approrpiate to say that the new model is better in this respect.

6. Residual Q-Q Plots

```
qqnorm( resid( model ),
        main="Q-Q Plot for Original Model" )
abline( mean( resid( model ) ),
        sd( resid( model ) ) )
```




```
qqnorm( resid( best.model ),
        main="Q-Q Plot for Transformed Model")
abline( mean( resid( best.model ) ),
        sd( resid( best.model ) ) )
```



The Q-Q Plot of residuals for both models indicates that the second (transformed) model has a stronger linear correlation along the Q-Q line. Thus, we can say that the new model is more appropriate in this respect than the original model.

7. Shapiro & Wilk Test

```
shapiro.test( resid( model ) )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.94216, p-value = 0.1999
```

```
shapiro.test( resid( best.model ) )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(best.model)
## W = 0.92277, p-value = 0.07637
```

According to the Shapiro and Wilk test, the new model residuals are less likely to be normally distributed than the original model's residuals. However, despite having a lower p-value, the value itself is not past the threshold of statistical significance. As such, we can still say that the new model produces normally distributed residuals.

IV. Conclusion

Using the Box Cox transformation, it was possible to create a model for the data which performed significantly better than a model which did not use a transformation. In all areas except for the Shaprio and Wilk test, the new model performed better. While the new model did worse on the Shapiro and Wilk test, it still did not violate any assumptions of the regression. As such, the Box Cox transformation produced a better model overall.