STAT-615 Regression Exam 1

Yunting Chiu

2021-03-15

Part 2 (80 points): Exercises

1. (Use R for data analysis) The 1974 Motor Trend US magazine contained data on fuel consumption of 32 automobiles (1973-74 models). These data are in dataset "mtcars" which is already loaded in R. You can look at it with commands attach(mtcars), names(mtcars), summary(mtcars), mtcars. Your task is to study the effect of the number of carburetors (variable carb) on the fuel consumption in miles per gallon (variable mpg).

```
mtcars
```

```
##                      mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4           21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag       21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710          22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive      21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout   18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant             18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360          14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D           24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230            22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280            19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
## Merc 280C           17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
## Merc 450SE          16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
## Merc 450SL          17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
## Merc 450SLC         15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
## Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
## Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
## Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
## Fiat 128            32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
## Honda Civic         30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
## Toyota Corolla      33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
## Toyota Corona       21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
## Dodge Challenger    15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
## AMC Javelin         15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
## Camaro Z28          13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
## Pontiac Firebird    19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
## Fiat X1-9           27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
## Porsche 914-2       26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
## Lotus Europa        30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
## Ford Pantera L      15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
## Ferrari Dino        19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
## Maserati Bora       15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
## Volvo 142E          21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

(a) Fit a linear regression model that can be used to predict miles per gallon based on the number of carburetors. Is the number of carburetors significant in this prediction? Report the estimated regression equation, the p-value testing significance of carburetors, and state your conclusion.

- With the p-value 0.001084, we have evidence to reject the null hypothesis in favor of an alternative hypothesis. That is, if the number of carburetors adds one unit, the miles per gallon will decrease by 2.0557 gallons.

```
reg <- lm(mpg~carb, data = mtcars) # the fuel consumption in miles per gallon ~ number of carburetors
summary(reg)
```

```
##
## Call:
## lm(formula = mpg ~ carb, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.250 -3.316 -1.433  3.384 10.083
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.8723     1.8368  14.085 9.22e-15 ***
## carb         -2.0557     0.5685  -3.616  0.00108 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.113 on 30 degrees of freedom
## Multiple R-squared:  0.3035, Adjusted R-squared:  0.2803
## F-statistic: 13.07 on 1 and 30 DF,  p-value: 0.001084
```

(b) Conduct a lack-of-fit test to decide whether the relation between the fuel consumption and the number of carburetors is linear. State the test statistic, the p-value, and your conclusion. What does this test statistic measure?

- reduced model is the usual linear regression model, SSE(Reduced) = 784.27
- full model is treating `X` as categorical and fitting the mean at each carb. SSE(Full) = 625.49 = SSE(pure error)
- The lack of fit SSE(lack of fit) = SSE(reduced) - SSE(Full) = 784.27-625.49 = 158.78
- F = (158.78/4) / (625.49)/26 = 39.695 / 24.05731 = 1.650018
- We conclude that the p-value is 0.1918, we fail to reject the H0, meaning that there is no evidence of lack of fit. Thus, using the linear regression is almost as good as using separate means at the each level of the number of carburetors.

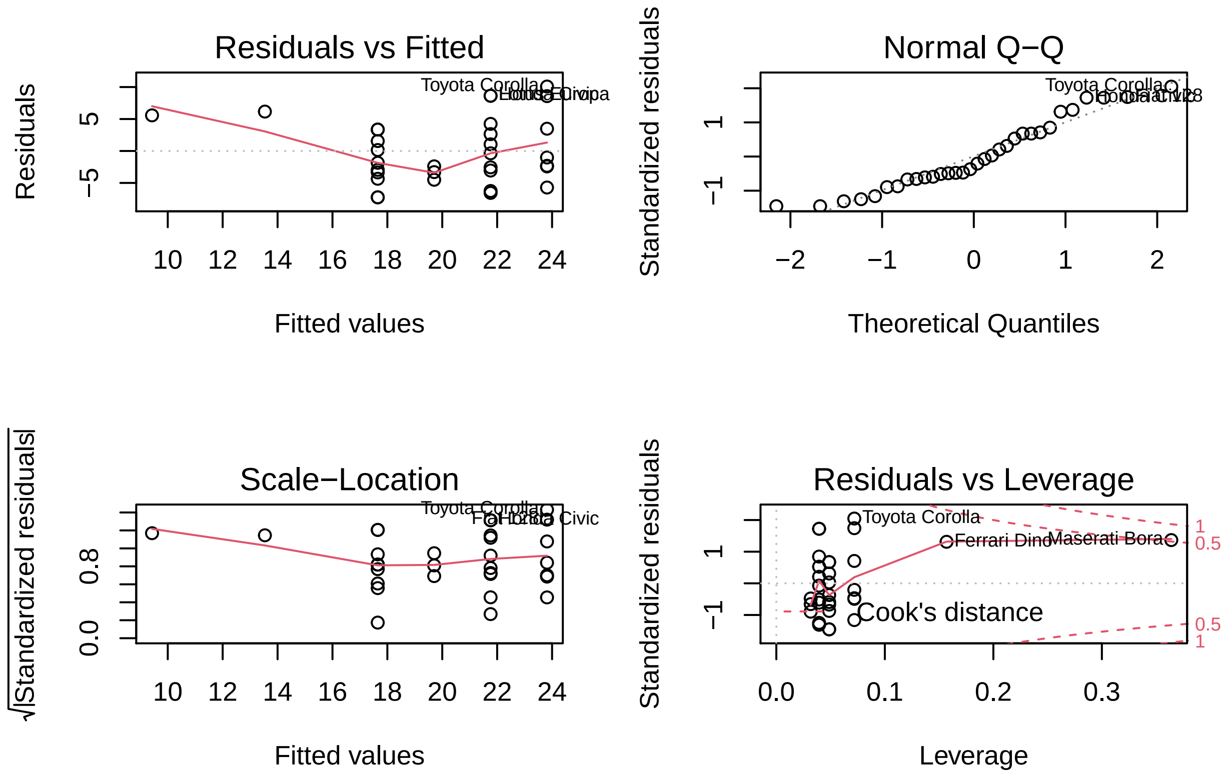
```
reduced <- lm(mpg ~ carb, data = mtcars) # simple linear regression predicting Y in terms of X
full <- lm(mpg ~ as.factor(carb), data = mtcars) # using group means to predict Y for each value of X,

anova(reduced, full)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ carb
## Model 2: mpg ~ as.factor(carb)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     30 784.27
## 2     26 625.49  4    158.78 1.6501 0.1918
```

(c) Are there any outliers in this regression analysis? Test each residual keeping the familywise error rate at a 5% level. Explain how you did the test, report the numbers that lead to your conclusion.

- At the individual level $\alpha = 0.05$, there is a potential outlier - observation *Toyota Corolla* with the studentized residual t = 2.169892. Then, keeping the familywise error rate at the same level and using `outlierTest` for testing,

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




```r
# Studentized residuals and testing for outliers
#t <- rstudent(reg)
#par(mfrow=c(1,1)) # Return to the 1x1 plot window
# plot(t)
#t[ abs(t) > 2 ]
#n = length(carb)
#qt( 0.025/n, n-2 )
#t[ abs(t) > abs(qt( 0.025/n, n-2 ))]
```

```r
outlierTest(reg)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##                rstudent unadjusted p-value Bonferroni p
## Toyota Corolla 2.169892           0.038349           NA
```

2. (Use R for data analysis) The purpose of this experiment was to assess the influence of calcium in solution on the contraction of heart muscle in rats. The left auricle of 21 rat hearts was isolated and on several occasions a constant length strip of tissue was electrically stimulated and dipped into various concentrations of calcium chloride solution, after which the shortening of the strip was accurately measured as the response.

The data are stored in R package MASS. You can look at them with commands attach(muscle), names(muscle), summary(muscle), muscle. A linear regression model is used to predict the change in length of the strip (variable Length, in mm) based on the concentration of calcium chloride solution (variable Conc, in multiples of 2.2 mM).

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```r
muscle
```

```
##     Strip Conc Length
## 3     S01 1.00   15.8
## 4     S01 2.00   20.8
## 5     S01 3.00   22.6
## 6     S01 4.00   23.8
## 9     S02 1.00   20.6
## 10    S02 2.00   26.8
## 11    S02 3.00   28.4
## 12    S02 4.00   27.0
## 13    S03 0.25    7.2
## 14    S03 0.50   15.4
## 15    S03 1.00   22.8
## 16    S03 2.00   27.4
## 19    S04 0.25    2.2
## 20    S04 0.50    9.0
## 21    S04 1.00   16.6
## 25    S05 0.25    2.0
## 26    S05 0.50    6.0
## 27    S05 1.00   15.2
## 31    S06 0.25    5.0
## 32    S06 0.50    9.2
## 33    S06 1.00   14.2
## 39    S07 1.00   28.0
## 40    S07 2.00   32.0
## 43    S08 0.25    5.6
## 45    S08 1.00   26.0
## 50    S09 0.50   15.4
## 51    S09 1.00   23.2
## 55    S10 0.25   11.8
## 57    S10 1.00   29.0
## 61    S11 0.25   11.0
## 62    S11 0.50   18.8
## 63    S11 1.00   26.2
## 69    S12 1.00   26.0
## 70    S12 2.00   33.8
## 75    S13 1.00   24.2
## 76    S13 2.00   28.8
## 80    S14 0.50   15.0
## 81    S14 1.00   24.0
## 86    S15 0.50   20.8
## 87    S15 1.00   29.0
```

```
## 93    S16 1.00   18.2
## 94    S16 2.00   25.8
## 95    S16 3.00   30.0
## 96    S16 4.00   32.2
## 99    S17 1.00   21.5
## 100   S17 2.00   28.4
## 101   S17 3.00   32.0
## 102   S17 4.00   29.6
## 105   S18 1.00   15.4
## 106   S18 2.00   19.0
## 107   S18 3.00   19.4
## 111   S19 1.00   29.0
## 112   S19 2.00   34.0
## 113   S19 3.00   37.0
## 117   S20 1.00   22.2
## 118   S20 2.00   29.0
## 119   S20 3.00   32.2
## 123   S21 1.00   23.0
## 124   S21 2.00   27.4
## 125   S21 3.00   30.4
```

(a) Calculate the equation of the sample regression line that predicts Length based on Conc.

- According to the summary table below, we focus on $\beta 1$. If the concentration of calcium add one unit, the change in length of the strip will increase 5.4030 mm.

```r
reg2 <- lm(Length~Conc, data = muscle)
summary(reg2)
```

```
##
## Call:
## lm(formula = Length ~ Conc, data = muscle)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.884  -4.097   1.060   4.487  10.064
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.5330     1.4229   9.511 1.93e-13 ***
## Conc          5.4030     0.7653   7.060 2.32e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.411 on 58 degrees of freedom
## Multiple R-squared:  0.4622, Adjusted R-squared:  0.4529
## F-statistic: 49.85 on 1 and 58 DF,  p-value: 2.322e-09
```

(b) Complete the ANOVA table and estimate the variance of Length.

- At the α - 0.05, we set H0: $\beta 1 = 0$ v.s. H:a $\beta 1$!= 0.
- We tested the F-value is 49.847. However, in the significant level tα - 0.05, the F-stat is 4.006873.
- Because 49.847 > 4.006873 so p-value is less than 0.05, the H0 can be rejected, meaning that the linear relation between `Conc`and `Length` are found significant.

```r
anova(reg2)
```

```
## Analysis of Variance Table
##
## Response: Length
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Conc       1 2048.7  2048.7  49.847 2.322e-09 ***
## Residuals 58 2383.7    41.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
qf(0.95, df1 = 1, df2 = 58)
```

```
## [1] 4.006873
```

(c) Compute a 95% confidence interval for the regression slope $\beta 1$

- The 95% confidence interval for the slope (5.4030) is between 3.871132 to 6.934835

```r
confint(reg2, "Conc", level = 0.95)
```

```
##         2.5 %   97.5 %
## Conc 3.871132 6.934835
```

(d) Test whether the slope is zero or not.

- The p-value of slope $\beta 1$ was found significant in the summary table (p-value: 2.32e-09). That is, the slope is not equal to zero.

(e) Calculate the percent of total variation explained by this regression model.

- The r-square is 0.4622014, so the linear regression model has 46 % of the variance for a dependent variable `Length` that's explained by an independent variable `Conc` in the regression model.

```r
summary(reg2)$r.square
```

```
## [1] 0.4622014
```

(f) Compute a 90% confidence interval for the mean Length when the concentration of calcium is 2.5.

```r
muscle %>%
  filter(Conc == 2.5) -> tmpDf
# Confidence intervals for regression coefficients
confint(reg2, level = 0.9)
```

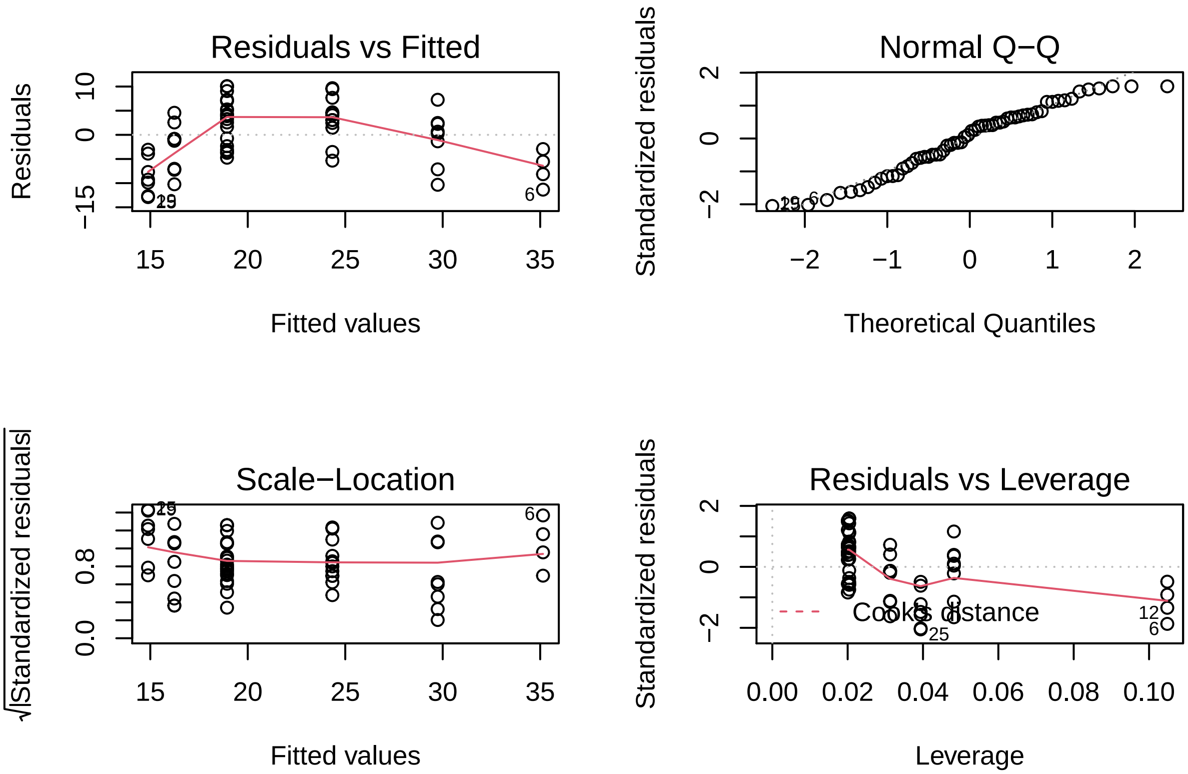
```
##                   5 %     95 %
## (Intercept) 11.154494 15.91148
## Conc         4.123797  6.68217
```

(g) Compute a 90% prediction interval for Length if the concentration of calcium is 2.5.

(h) Verify the standard regression assumptions - normality and homoscedasticity. Report p-values and state your conclusions.

Here are the assumptions of simple linear regression model: 1. independent observation 2. Normally distribution 3. Equal variances 4. No influential outliers 5. Linear association between (mean) y and x. That is, residual : ri = yi - yhat i. ## Normality - using Normal Q-Q plot - According to the normal QQ plot, there are some potential outliers in the upper extremity and lower extremity

```r
par(mfrow=c(2,2))
plot(reg2)
```



Normality - Shapiro-Wilk normality test - With large p-value 0.07566, we fail to reject the null, meaning that the data may not be non-normal.

```r
t <- rstudent(reg2)
shapiro.test(t)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  t
## W = 0.9642, p-value = 0.07566
```

Homoscedasticity (constant variance)

- With a high p-value 0.57094,there is no evidence of non-constant variance.

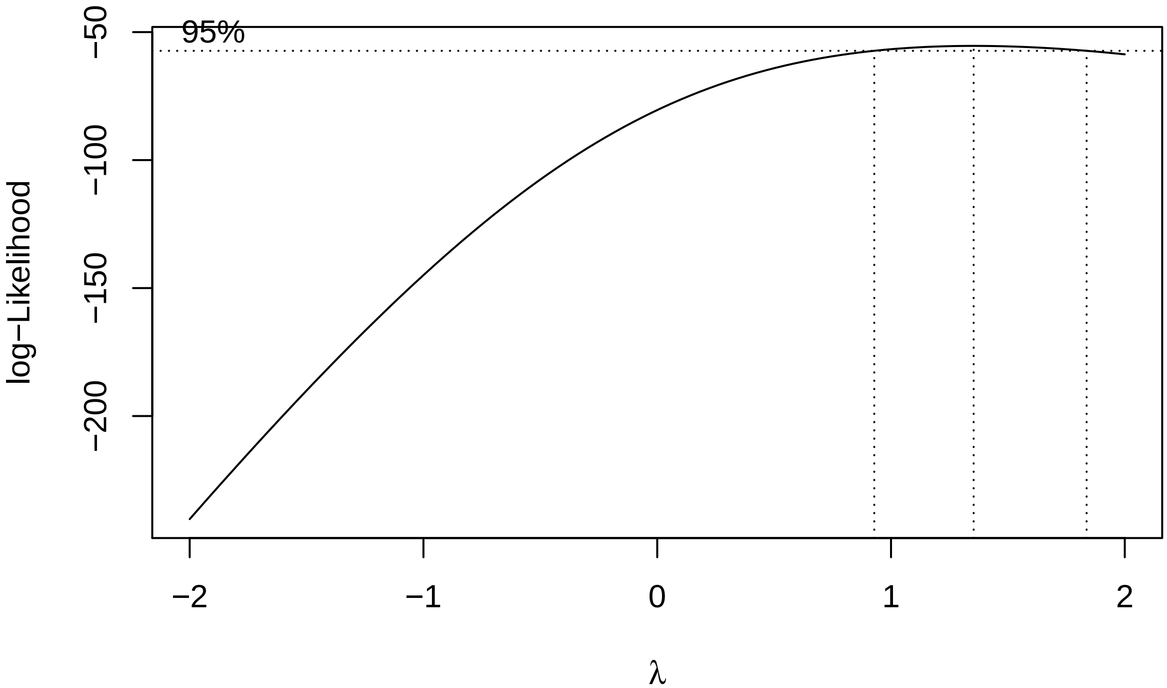
```r
ncvTest(reg2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.3211136, Df = 1, p = 0.57094
```

(i) **(Graduate only)** Find the optimal Box-Cox transformation. Does it improve normality of residuals?

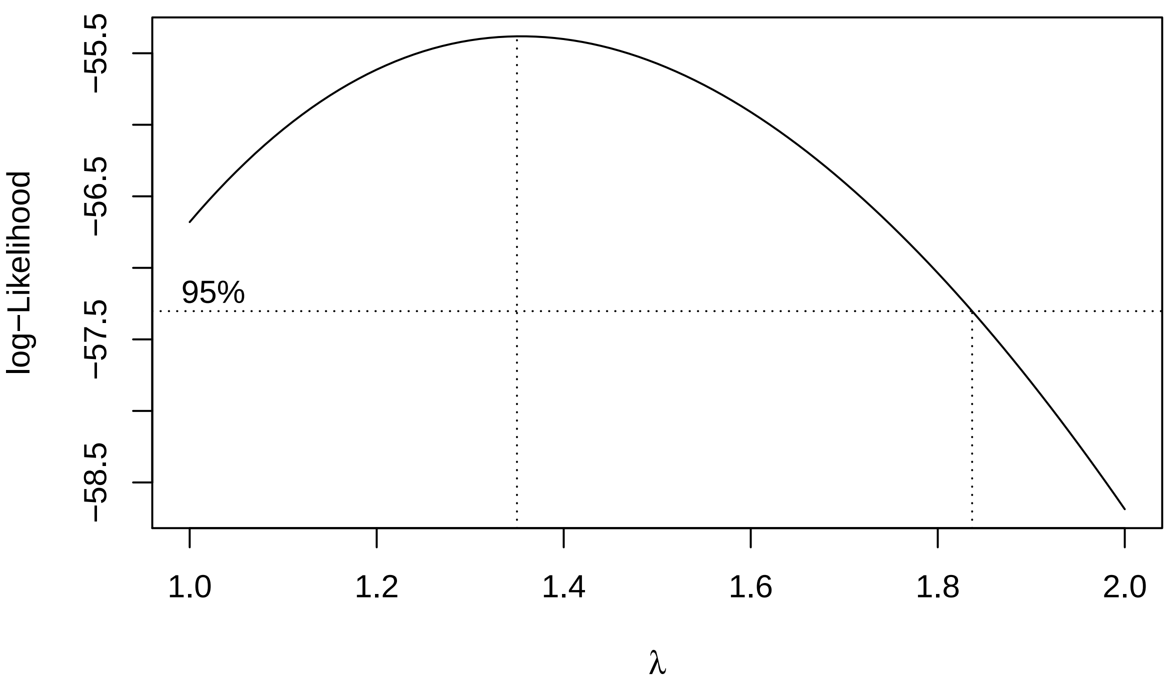
- A Box Cox transformation is a transformation of a non-normal dependent variables into a normal shape. In this case, we need to focus on **the largest Y-value** mapping to the X position. Thus, the optimal lambda is somewhere between 1 to 2. Then, we zoom in the 1:2 domain with the step 0.01

```r
boxcox(reg2)
```



- Now, we can see that the best lambda is approximately close to 1.4 on x-axis (the peak spot). Let's introduce a variable that is the corresponding power transform of our response Y, fit this new regression, and check residuals for normality.

```
boxcox(reg2, lambda = seq(1, 2, 0.01))
```



- Recalled: the normality test p-value of original model is **0.07566** - According to the Shapiro-Wilk normality test table below, the p-value is **0.1533** - Because 0.1533 > 0.07566, also the p-value is far away to the α level. We can conclude that the Box-Cox transformation improves residual normality.

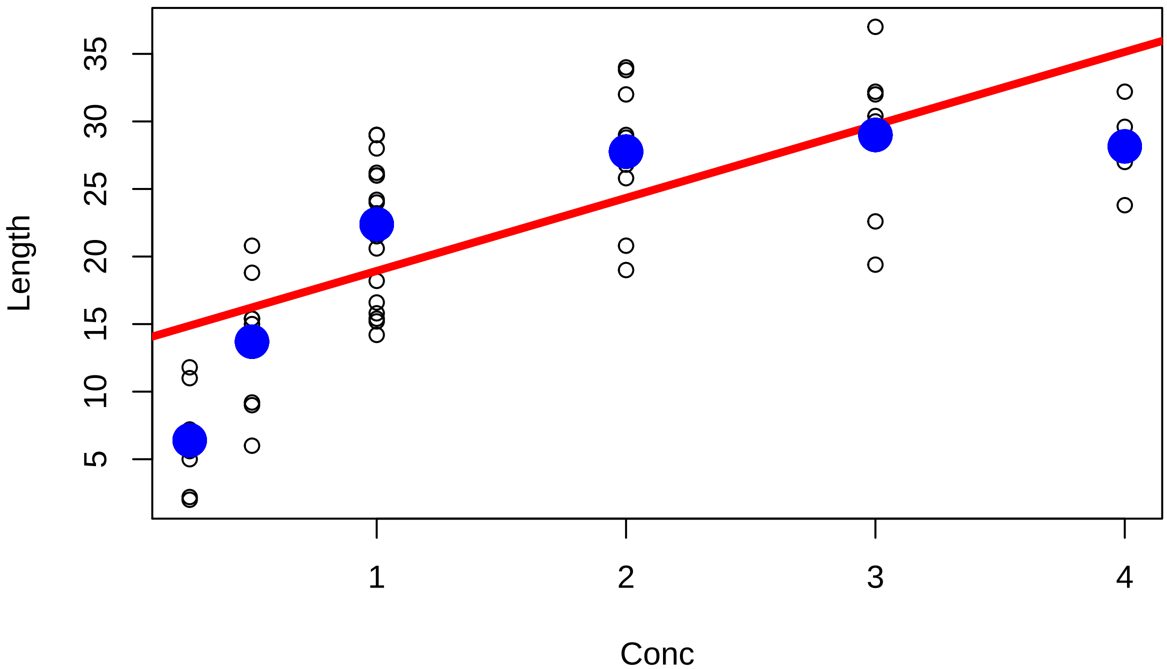
```
attach(muscle)
z <- Length^(1.4)
newReg2 <- lm(z~Conc)
```

```r
shapiro.test(rstudent(newReg2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstudent(newReg2)
## W = 0.97044, p-value = 0.1533
```

(j) (Graduate only) Test the model for the lack of fit.

```r
reduced2 <- lm(Length ~ Conc, data = muscle)
full2 <- lm(Length ~ as.factor(Conc), data = muscle)

plot(Conc, Length)
abline(reduced2,col="red",lwd = 4)
points(Conc, predict(full2), col="blue", lwd = 10 )
```



A rigorous F-test for the lack of fit - `reduced2` is the usual linear regression model, SSE(Reduced) = 784.27 - `full2` is treating `X` as categorical and fitting the mean at each carb. SSE(Full) = 625.49 = SSE(pure error) - The lack of fit SSE(lack of fit) = SSE(reduced) - SSE(Full) = 784.27-625.49 = 158.78 - F = (158.78/4) / (625.49)/26 = 39.695 / 24.05731 = 1.650018 - We conclude that the p-value is 0.1918, we fail to reject the H0, meaning that there is no evidence of lack of fit. Thus, using the linear regression is almost as good as using separate means at the each level of the number of carburetors.

```r
anova(reduced2, full)
```

```
## Warning in anova.lmlist(object, ...): models with response '"mpg"' removed
## because response differs from model 1

## Analysis of Variance Table
##
## Response: Length
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Conc       1 2048.7  2048.7  49.847 2.322e-09 ***
## Residuals 58 2383.7    41.1
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```