STAT-615 Regression Exam 1

Yunting Chiu

2021-03-21

Part 1 (20 points): Concept problems

True or False. Justify your answer

1. The sum of the residuals is equal to zero.
   Answer: True

- Half of the residuals will equal exactly half of the remaining residuals. Half are positive, half are negative, and they eliminate each other out.

2. A significant positive correlation between X and Y implies that changes in X cause Y to change.
   Answer: False

- The strength of the linear association is measured by correlation. `r` is always a number between -1 and 1. If r is close to 0, it means there is no relationship between the variables (1 means prefect positive correlation; -1 means prefect negative correlation).

3. The residual is the difference between the observed value of the dependent variable and the predicted value of the dependent variable. In mathematical notation this is given by $Y - E\{Y\}$.
   Answer: True

- The difference between the observed Y and the predicted Y $(Y - \hat{Y})$ is called a residual.

4. If MSR and MSE are of the same order of magnitude, this would suggest that $\beta 1 \neq 0$.
   Answer: False

- If the MSE and the MSR are of the same order of magnitude, this suggests that $\beta 1 = 0$. If the MSR is significantly greater than the MSE, this suggests that $\beta 1 \neq 0$.

5. When using simple regression analysis, if there is a strong correlation between the independent and dependent variable, then we can conclude that an increase in the value of the independent variable causes an increase in the value of the dependent variable.
   Answer: False

- We need to focus on $\beta 1$ to know the association between independent variable and dependent variable.

6. The least squares regression line minimizes the sum of the squared differences between actual and predicted Y values.
   Answer: True

- The least squares regression line minimizes the sum of the residuals squared.

7. The correlation coefficient takes values between 0 and 1.
   Answer: False

- The range of correlation coefficient is between -1 to +1.

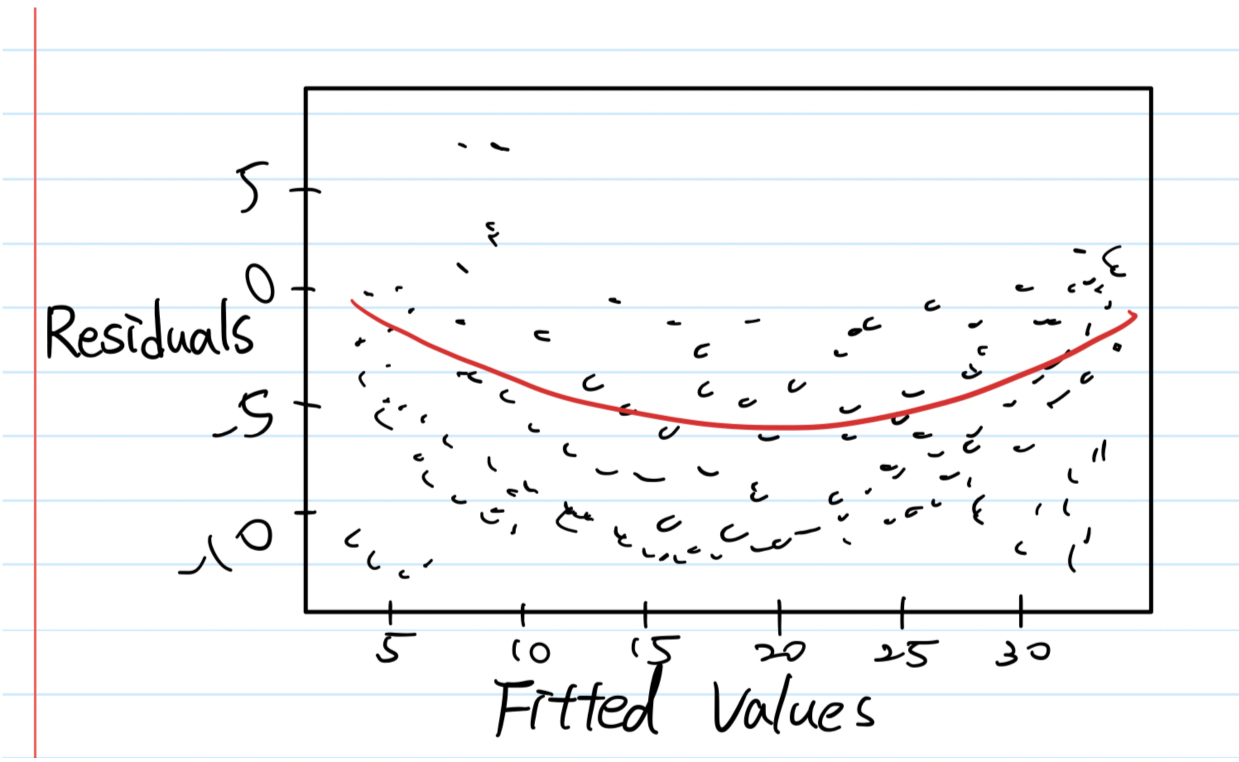
8. The coefficient of determination is interpreted as the proportion of observed variation in X that can be explained by the simple linear regression model.
   Answer: False

- In statistics, the coefficient of determination, denoted R^2
- R-squared gives us the percentage variation in y explained by x-variable
- The usual way of interpreting the coefficient of determination R^2 is to see it as the percentage of the variation of the dependent variable y (Var(y)) can be explained by our model.

9. One way to study the normality of the error is by histograms.
   Answer: True

- If the graph has a bell-shaped shape and is symmetric about the mean, the data may follow a normal distribution.

10. Draw a fitted versus residuals plot where we see that the constant variance assumption is not met and the linearity assumption is not violated.



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.

$$mpg = 25.8723 - 2.0557carb$$

```r
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)
```

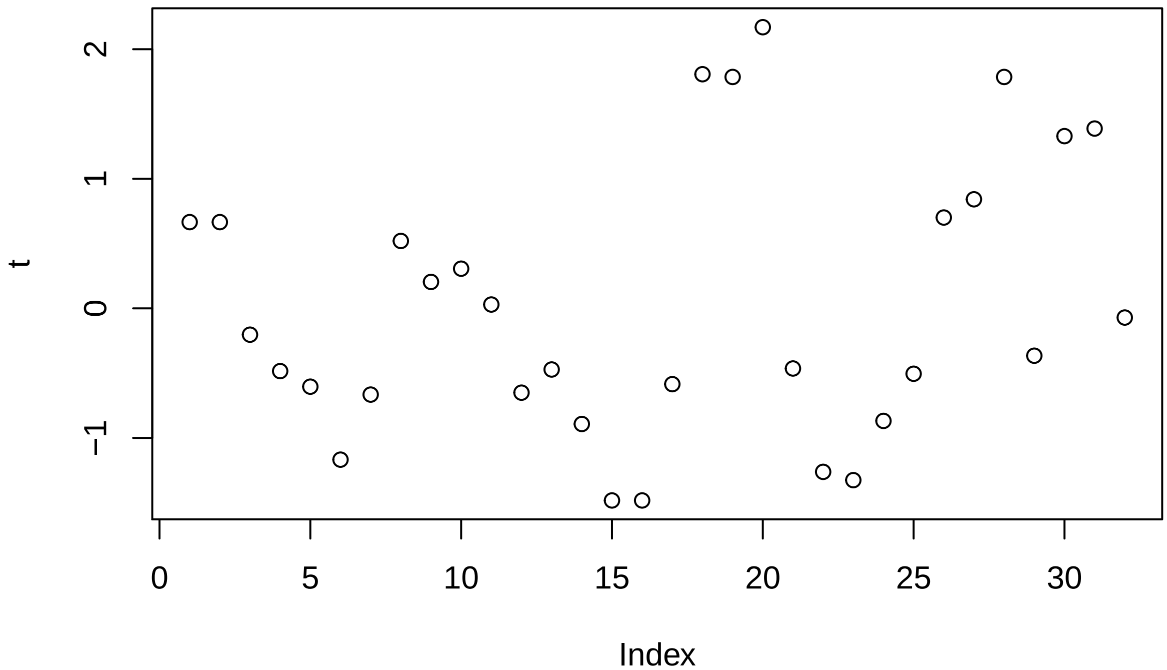
```
## Analysis of Variance Table
##
## Model 1: mpg ~ carb
## Model 2: mpg ~ as.factor(carb)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     30 784.27
## 2     26 625.49  4    158.78 1.6501 0.1918
```

(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, we are going to keep the familywise error rate at the same level and using `outlierTest` for testing.

- The test provided no outiers

```
# 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]
```

```
## Toyota Corolla
##       2.169892
```

```
attach(mtcars)
```

```
## The following object is masked from package:ggplot2:
##
##     mpg
```

```
n = length(carb)
qt( 0.025/n, n-2 ) # -3.478736
```

```
## [1] -3.478736
```

```
t[ abs(t) > abs(qt( 0.025/n, n-2 ))]
```

```
## named numeric(0)
```

- According to Quantitative Research Methods for Political Science, Public Policy and Public Administration for Undergraduates, they indicate the `outlierTest` is: " The Bonferroni Outlier Tests uses a t distribution to test whether the model's largest studentized residual value's outlier status is statistically different from the other observations in the model. A significant p-value indicates an extreme outlier that warrants further examination."
- According to the conclusion of the `outlierTest`, the Bonferroni p-value for the largest (absolute) residual is not statistically significant (No Studentized residuals with Bonferroni $p < 0.05$). Thus, There is no evidence of any outliers.
- Reference: https://bookdown.org/wwwehde/qrm_textbook_updates/ols-assumptions-and-simple-regression-diagnostics.html

```
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.

$$Length = 13.5330 + 5.4030 * Conc$$

```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.

- Estimate the variance = S^2 = MSE = 41.1

Extra explanation: - 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.

```r
summary(reg2)$coefficients[2,] # b_1
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 5.402983e+00 7.652686e-01 7.060245e+00 2.321930e-09
```

(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.

- 90% confidence interval for `Length` expected values at `Conc` = 2.5 is:

```r
# muscle
predict(reg2, data.frame(Conc = 2.5), interval = "confidence", level = 0.90)
```

```
##        fit      lwr      upr
## 1 27.04045 25.16706 28.91383
```

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

- 90% prediction interval for `Length` expected values at `Conc` = 2.5 is:

```
predict(reg2, data.frame(Conc = 2.5), interval = "prediction", level = 0.90)
```

```
##        fit      lwr      upr
## 1 27.04045 16.16187 37.91902
```

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

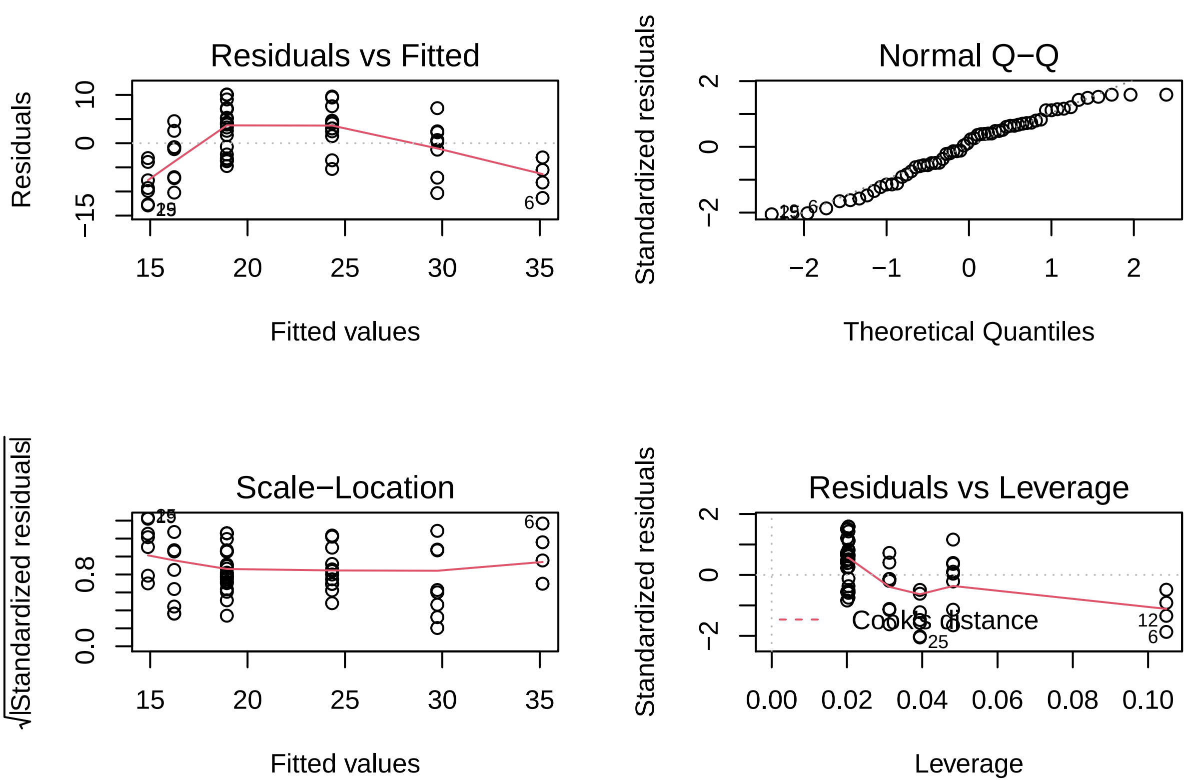
Here are the assumptions of simple linear regression model:

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

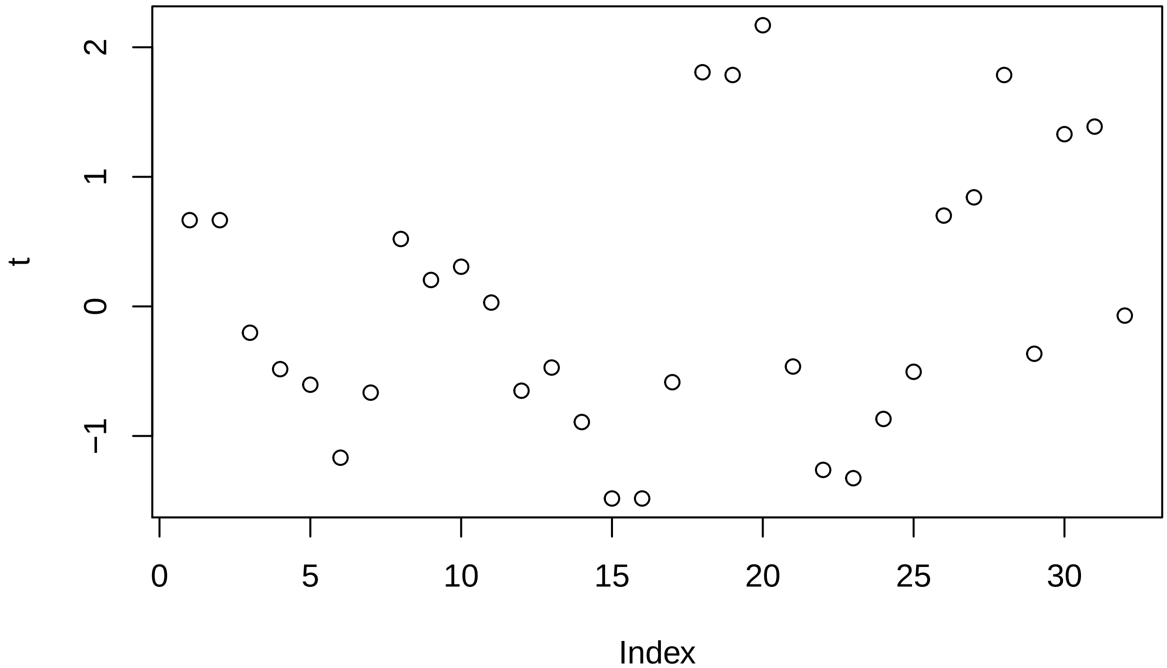
```
par(mfrow=c(2,2))
plot(reg2)
```



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.

```
tReg2 <- rstudent(reg2)
par(mfrow=c(1,1)) # Return to the 1x1 plot window
plot(t)
```




```
shapiro.test(tReg2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  tReg2
## 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.

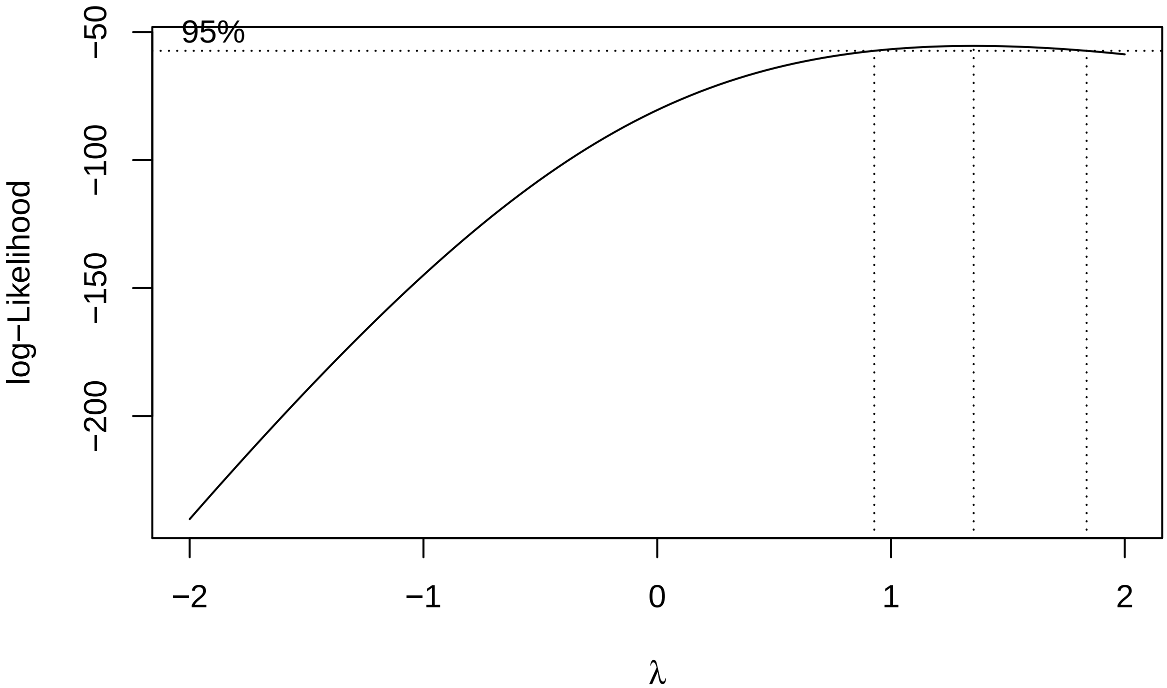
```
ncvTest(reg2)
```

```
## 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?

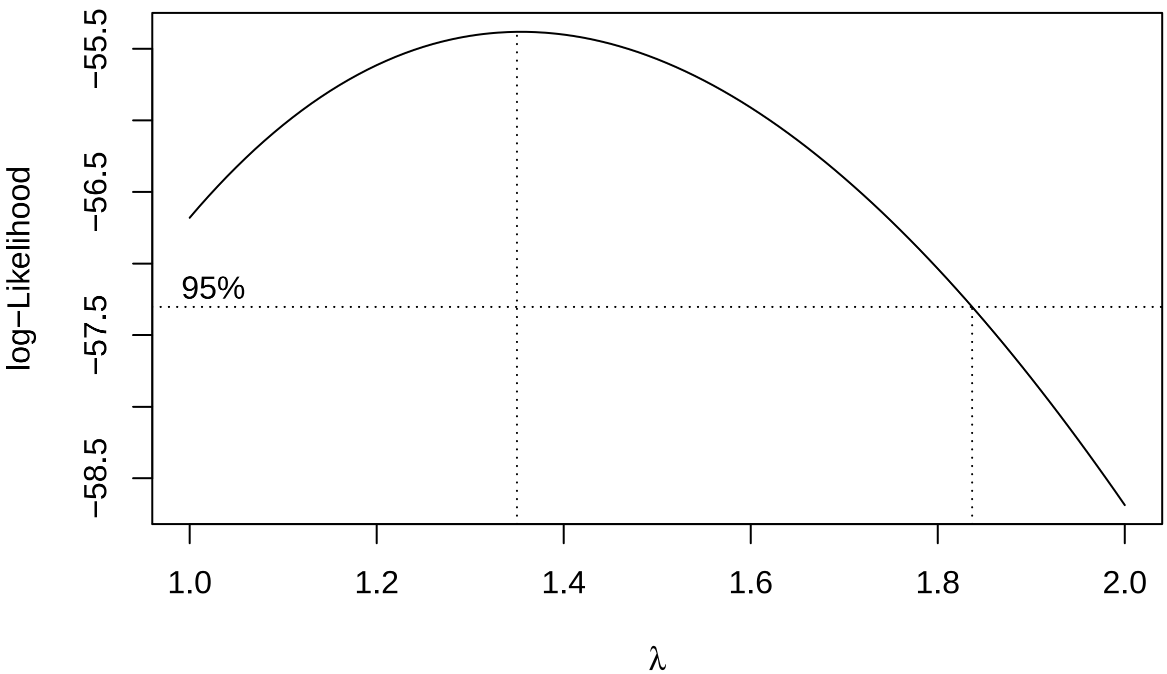
- A Box Cox transformation is a transformation of a non-normal dependent variables into a normal shape. In this case, we need to focus on **the largest Y-value** mapping to the X position. Thus, the optimal lambda is somewhere between 1 to 2. Then, we zoom in the 1:2 domain with the step 0.01

```
boxcox(reg2)
```

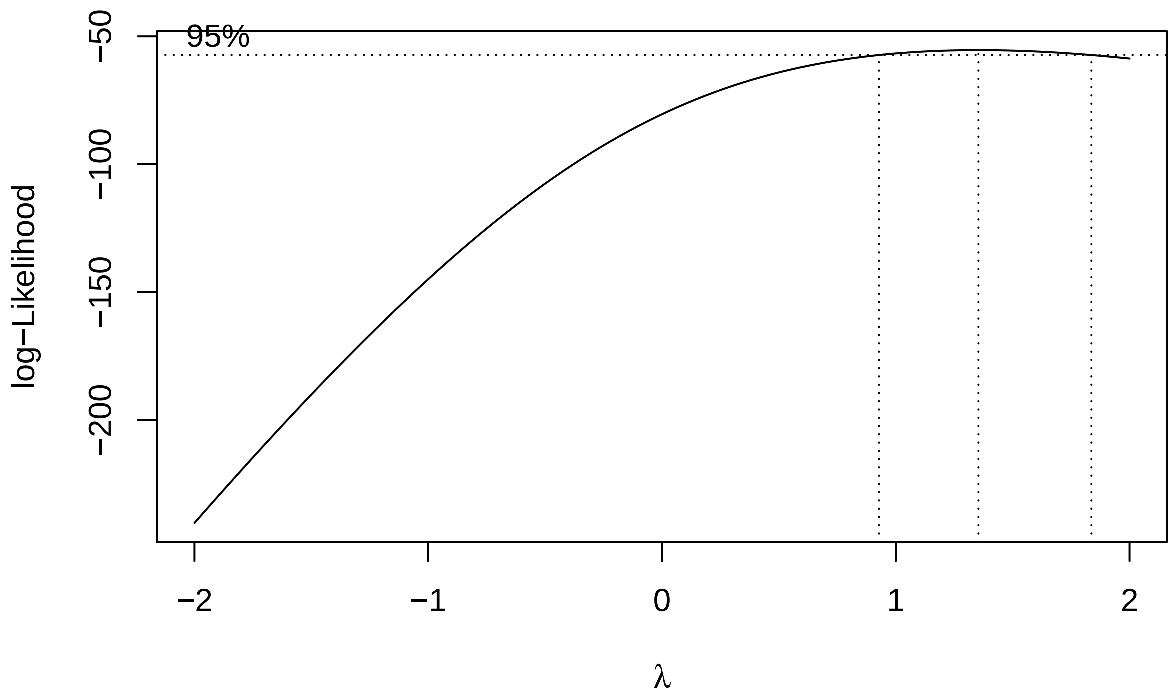


- Now, we can see that the best lambda is approximately close to 1.4 on x-axis (the peak spot). Let's introduce a variable that is the corresponding power transform of our response Y, fit this new regression, and check residuals for normality.

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




```
# find the max lambda, and we get the value is 1.353535
bc <- boxcox(reg2)
```




```
spot <- bc$x[which.max(bc$y)]
spot
```

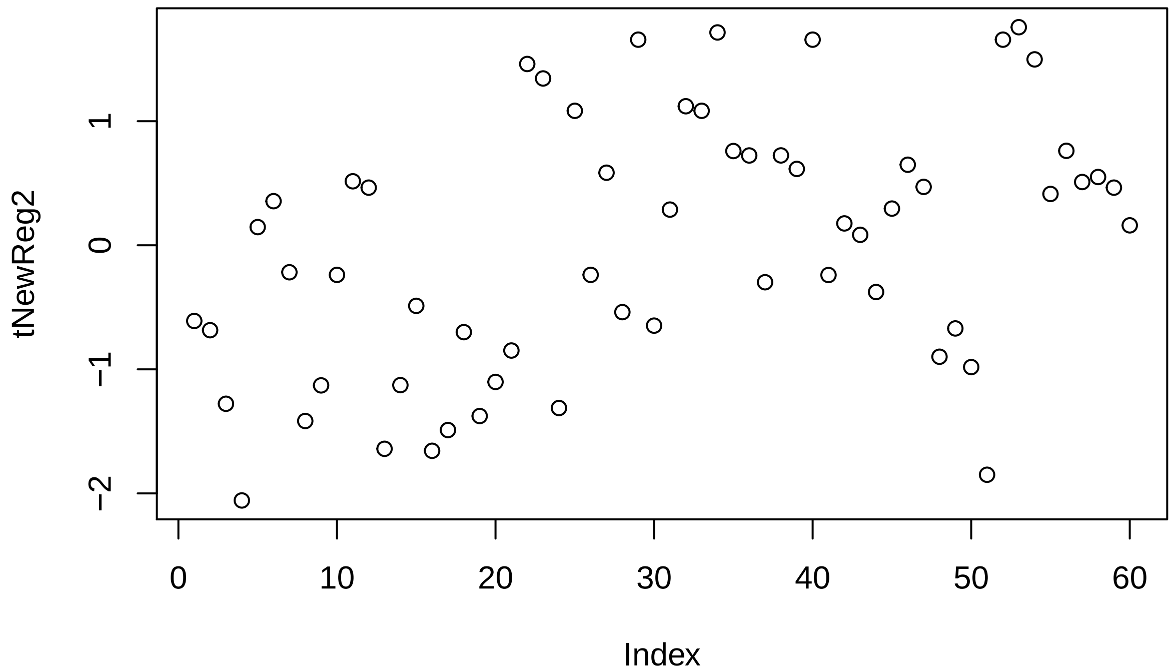
```
## [1] 1.353535
```

- 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 **0.1378**
- Because $0.1378 > 0.07566$, and the p-value is far away to the α level. Plus, the below residual plot indicates that the Box-Cox transformation improves residual normality.

```
attach(muscle)
z <- Length^(1.353535)
newReg2 <- lm(z~Conc)
shapiro.test(rstudent(newReg2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstudent(newReg2)
## W = 0.96949, p-value = 0.1378
```

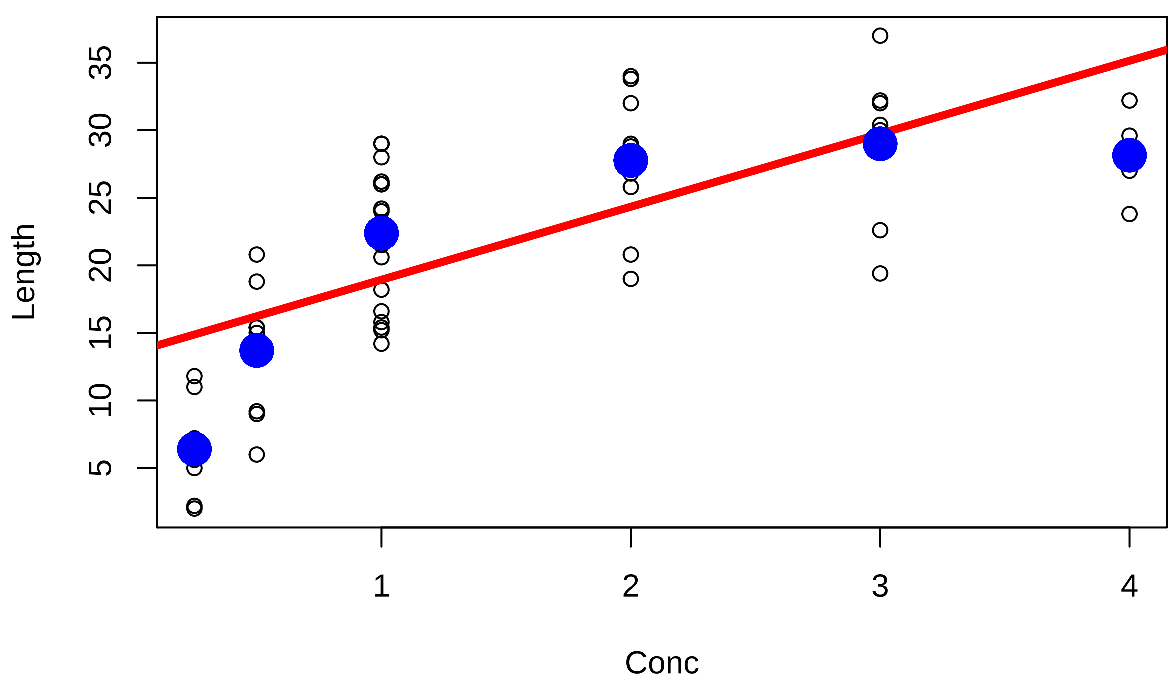
```
tNewReg2 <- rstudent(newReg2)
par(mfrow=c(1,1)) # Return to the 1x1 plot window
plot(tNewReg2)
```



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

```
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) = 2383.7

- `full2` is treating `X` as categorical and fitting the mean at each `Y`. SSE(Full) = 1237.5 = SSE(pure error)
- The lack of fit SSE(lack of fit) = SSE(reduced) - SSE(Full) = 2383.7-1237.5 = 1146.2
- F = 12.504
- With the small p-value 2.873e-07, there is evidence of lack of fit.

```
anova(reduced2, full2)
```

```
## Analysis of Variance Table
##
## Model 1: Length ~ Conc
## Model 2: Length ~ as.factor(Conc)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     58 2383.7
## 2     54 1237.5  4    1146.2 12.504 2.873e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3. (By hand: show all steps)

A sample of size n = 100 contains two variables, X and Y . Sample statistics are: X_bar = 50, Y_bar = 10, S_X = 10, S_Y = 4, r_XY = 0.2.

(a) Calculate the equation of the sample regression line that predicts Y based on X. Predicted values:

$$\hat{Y}i = 6 + 0.08 * X_i$$

(a) $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

$E\{Y_i\} = E\{\beta_0 + \beta_1 X_i\}$

sample of sd:

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$$

s = sample standard deviation
N = the number of observations
x_i = the observed values of a sample item
$\overline{x}$ = the mean value of the observations

Correlation coefficient $= r = \frac{S_{xy}}{S_x S_y}$

So $0.2 = \frac{S_{xy}}{S_x S_y}$ $\quad 0.2 = \frac{S_{xy}}{40}$ $\quad S_{xy} = 8$

$$r = \frac{s_{xy}}{s_x s_y} \implies b_1 = \frac{s_{xy}}{s_x^2} = r\,\frac{s_y}{s_x}.$$

Sample regression slope $\quad b_1 = \frac{S_{XY}}{S_{XY}} = \frac{s_{xy}}{s_x^2}$

Sample regression intercept $\quad b_0 = \overline{Y} - b_1 \cdot \overline{X}$

$b_1 = \frac{8}{S_X^2} = \frac{8}{100} = 0.08$ or $b_1 = r \cdot \frac{S_y}{S_x} = 0.2 \cdot \frac{4}{10} = 0.2 \cdot 0.4 = 0.08$

$b_0 = \overline{Y} - b_1 . \overline{X} = 10 - 0.08 \times 50 = 10 - 4 = 6$

Predict Value $: \hat{Y}_i = b_0 + b_1 . X_i$

$= \hat{Y}_i = 6 + 0.08 \cdot X_i$

or

$E(Y|X) = 6 + 0.08 . X$ #

(b) Complete the ANOVA table and estimate the variance of Y.
Include sum of squares, degrees of freedom, mean squares and the ANOVA F-statistic.

$n = 100$

	sample mean	sd	sample correlation coeff. (r_{xy})
x	50	10	0.2
Y	10	4	

(b)

ANOVA Table

Sum of squares	DF	Mean Squares	F-ratio
① SSReg = 63.36	1	MSReg = 63.36	4.083335
SSErr = 1520.64	$n-2=98$	MSErr = 15.51673	
SSTot = 1584	$n-1=99$		

$SSTot = (n-1) \cdot s_y^2 = 99 \cdot 16 = 1584$

$SSErr = SSTot - SSReg = 1584 - 63.36 = 1520.64$

$SSReg = r^2 \cdot SSTot = (0.2)^2 \cdot 1584 = 0.04 \cdot 1584 = 63.36$

$MSR = \frac{SSReg}{1} = 63.36$

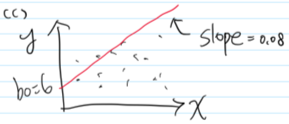
$MSErr = \frac{SSE}{n-2} = \frac{1520.64}{98} = 15.51673$

$F\text{-stat} = \frac{MSR}{MSE} = \frac{63.36}{15.51673} = 4.083335$

② Estimate $var(Y) \Rightarrow S^2 = MSErr = 15.51673$

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

(c)



S = sd of the residuals

0.975

95%, df = 100 − 2 = 98, t = 1.984

$0.08 \pm t^* \cdot$

critical t value

$$S = S_y\sqrt{\frac{n-1}{n-2}(1-r^2)} = 4\sqrt{\frac{99}{98}(1-0.04)} = 4\sqrt{1.01\times 0.96} = 4\cdot\sqrt{0.97} = 3.94$$

$$S_{b_1} = \frac{S}{\sqrt{S_{xx}}} = \frac{S}{S_x\sqrt{n-1}} = \frac{3.94}{10\sqrt{99}} = \frac{3.94}{99.5} = 0.0396$$

$$\Rightarrow b_1 \pm t_{0.025} S_{b_1} = 0.08 \pm (1.984)(0.0396) = 0.08 \pm 0.0785664 = [\,0.0014336\ ,\ 0.1585664\,]$$

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

(d) $H_0: \beta_1 = 0$ v.s. $H_a: \beta_1 \neq 0$

We know F-ratio is 4.08 (F^*)

According to F-table $F(df_1 = 1, df_2 = 98)$ is 3.938 (F)

$\because F^* > F \Rightarrow 4.08 > 3.938$, the p-value is $P = P\{F > 4.08\} < 0.05 \Rightarrow$ reject null.

$\therefore$ we reject H_0, we have evidence conclude that the slope is not zero.

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

(e) $r = 0.2$, $r^2 = 0.04$ or $\frac{63.36}{1584} = 0.04$

$\therefore$ we have 4% explained the total variation of the model.

(f) Compute a 90% confidence interval for the mean response when $X = 35$.

(f) When $x = 35$

$\hat{Y}_i = b_0 + b_1 X_i \Rightarrow \hat{Y} = 6 + (0.08)(35) = 6 + 2.8 = 8.8$ $\qquad 1458 = s^2$

$$90\%\,CI: \hat{Y} \pm t(\alpha/2, df = n-2) \cdot s\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}} = 8.8 \pm (1.661)(3.94)\sqrt{\frac{1}{100} + \frac{\overbrace{(\cancel{35-50})}^{15}{}^2}{9900}}$$

$\frac{0.1}{2} = 0.05$

$df = 98$

$t = 1.661$

$$= 8.8 \pm (6.54434) \cdot \sqrt{\frac{1}{100} + \frac{225}{9900}}$$

$$= 8.8 \pm (6.54434) \cdot \sqrt{\cancel{\frac{9900 + 22500}{990000}}\ 0.0327272}$$

$$= 8.8 \pm (6.54434) \cdot (0.1809068) = 8.8 \pm 1.183916$$

$$= [7.616084,\ 9.983916]$$

Ans: $7.616084 \le E\{y_i\} \le 9.983916$ (90% CI)#

$\frac{S_X^2}{1} = \frac{S_{XX}}{n-1}$

$S_{XX} = S_X^2 \times 99$

$S_{XX} = 100 \times 99$

$= 9900$

(g) Compute a 90% prediction interval for the response Y_0 if the corresponding independent variable is $X_0 = 35$

(g) When $x = 35$

$$90\%\,PI: \hat{Y} \pm (t\alpha/2, df = n-2) \cdot s\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}} = 8.8 \pm (6.54434) \cdot \sqrt{\cancel{\frac{990000 + 9900 + 22500}{990000}}\ 1.0327272}$$

$$= 8.8 \pm (6.54434) \cdot (1.016232)$$

$$= 8.8 \pm 6.650568 = [2.149432,\ 15.45057]$$

Ans: $2.149432 \le E\{y_i\} \le 15.45057$ (90% PI)#

References:

- http://www.r-tutor.com/elementary-statistics/numerical-measures/correlation-coefficient#
- https://web.njit.edu/~wguo/Math644_2012/Math644_Chapter%201_part2.pdf
- https://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm