- **Homework 3 (50) Due Friday 9/28**

Name: Yinuo Sun
Section 10 (13:30-14:45pm)
Date: 09/27/2018

1. (15) A chemist studied the concentration of a solution (Y) over time (X). Fifteen identical solutions were prepared. The 15 solutions were randomly divided into five sets of three, and the five sets were measured, respectively, after 1,3,5,7 and 9 hours. The results is in concentration.csv

   R code for Question 1 is attached in the Appendix.

   a) (1) Fit a linear regression function.

   $\bar{X} = 5$

   $\bar{Y} = 0.955333$

   $SS_{XY} = \sum(X_i - \bar{X})(Y_i - \bar{Y}) = -38.88$

   $SS_X = \sum(X_i - \bar{X})^2 = 120$

   $b_1 = \frac{SS_{XY}}{SS_X} = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2} = \frac{-38.88}{120} = -0.324$

   $b_0 = \bar{Y} - b_1\bar{X} = 0.955333 - (-0.324)*5 = 2.575333$

   $\hat{Y} = b_0 + b_1X = 2.575333 - 0.324X$

   b) (3) Complete the ANOVA table, what is the SSE?

Source	Degrees of freedom	Sum of Squares	Mean Square	F-value
Model	1	12.59712	12.59712	55.99467
Error	13	2.924653	0.22497	
Corrected Total	14	15.521773		

   $SSR = \sum(\hat{Y}_i - \bar{Y})^2 = 12.59712$

   $MSR = SSR/1 = 12.59712$

   $SSE = \sum(Y_i - \hat{Y}_i)^2 = 2.924653$

   $MSE = \frac{SSE}{(n-2)} = \frac{2.924653}{13} = 0.22497$

   $F = \frac{MSR}{MSE} = \frac{12.59712}{0.22497} = 55.99467$

   SSE is the error sum of squares, which is 2.924653.

c) (3) Perform the F test to determine whether or not there is lack of fit of a linear regression function; use a significant level of 0.025. Compute the test statistic, reject region, and estimate the p-value. Is the model lack of fit for the data?

From data, the 15 solutions were randomly divided into five sets of three. Hence, there are 5 level of X. i.e. c = 5.

$H_0: E\{Y\} = \beta_0 + \beta_1 X$

$H_a: E\{Y\} \neq \beta_0 + \beta_1 X$

SSE(R)=SSE=2.924653, $df_R = n - 2 = 13$

SSE(F)=SSPE=$\sum\sum(Y_{ij} - \bar{Y}_j)^2$=0.1574, $df_F = n - c = 15 - 5 = 10$

$$F^* = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}} = \frac{\frac{SSE - SSPE}{n - 2 - (n - c)}}{\frac{SSPE}{n - c}} = \frac{\frac{2.924653 - 0.1574}{13 - 10}}{\frac{0.1574}{10}} = 58.60341$$

Since $F^* = 58.60341 > F(1 - \alpha; c - 2, n - c) = F(0.975; 3,10) = 4.83$, reject H_0.

Rejection Region: Reject H_0 if $F^* > 4.83$

To estimate P-value,

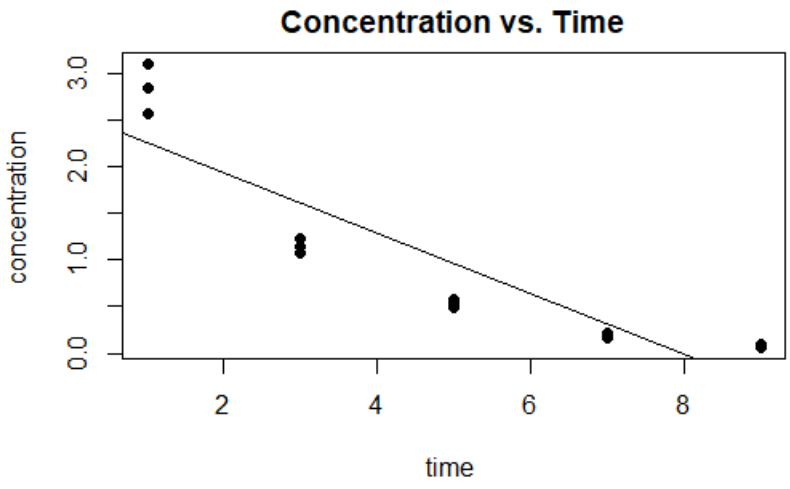
$F^* > F(0.999,3,10) = 12.55$, so P-value < 0.001, so reject H_0.

Since we reject H_0, the current model does not fit the data, i.e., the model is lack of fit for the data.

d) (3) Perform the diagnostics on the data

i. Plot the dependent variable versus the explanatory variable and comment on the shape and any unusual points.

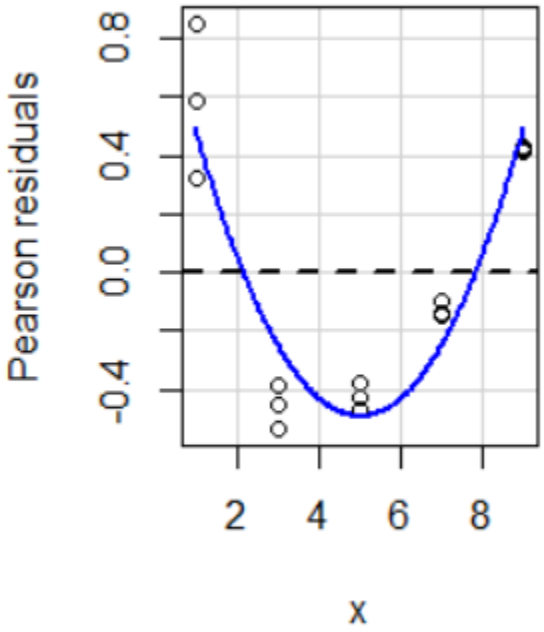
Figure 1.



From figure 1, X and Y are associated with each other, but it looks like a nonlinear relationship between time and concentrations.

ii. Plot the residuals versus the explanatory variable and briefly describe the plot noting any unusual patterns or points.

Figure 2.



From Figure 2, the distribution of residuals has a certain pattern, which means the error terms are non-normal and have non-constant variances.

iii. Examine the distribution of the random error. What do you conclude?

1. From Figure 2 in the last question, fortunately, it is also a plot of residuals versus time (x), we can easily find there is a certain pattern in the plot. Hence, the residuals are dependent with time.

2. H_0: Data follows normal distribution

   H_a: Data violates normal distribution

   Refer to Figure 4., P-value = 0.05634 > significance level $\alpha = 0.05$, but it is slightly bigger than the significance level. If we compare it to significance level $\alpha = 0.1$, then the p-value < $\alpha = 0.1$, so we could reject H_0. we are 90% confident that the data violates normal distribution. Also, from the Normal probability plot, the residual points do not fit the line very well.

Figure 3.

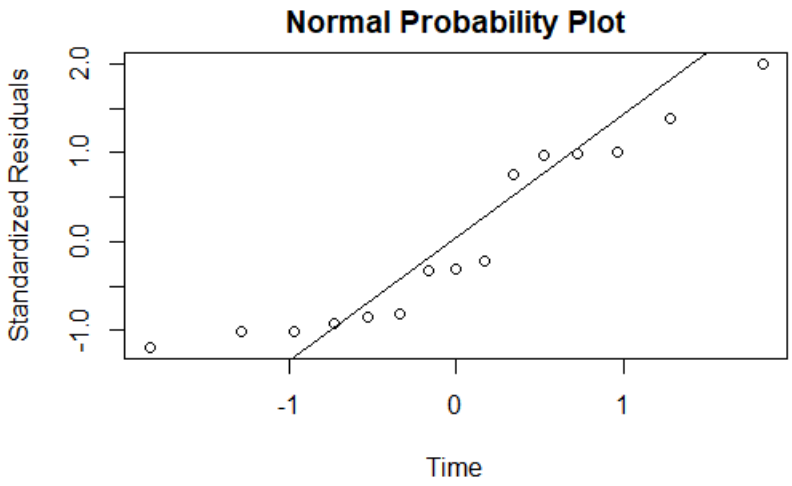


Figure 4.

```
        Shapiro-Wilk normality test

data:  concentration.stdres
W = 0.88498, p-value = 0.05634
```

3. H_0: Residuals have constant variances

   H_a: Residuals have non-constant variances

   Refer to Figure 5., P-value = 0.662261> significance level $\alpha = 0.05$, so we do not reject H_0. Hence, we are 95% confident that the residuals have constant variances.

   Figure 5.

```
        t.value  P.Value alpha df
[1,] 0.4469581 0.662261  0.05 13
```

iv. Do we need to do a transformation on X or Y? Why or why not?

We do not need to do a transformation on X, because explanatory variable and dependent variable are correlated with each other, even though it is a nonlinear relationship. However, we need to do a transformation on Y, because the error terms are non-normal from the Shapiro Test, and the distributions of residuals has a certain pattern from the residual plot. Both the distribution can be changed to a normal distribution with constant variances and the relationship can be improved to linear by the transformation of Y.

e) (2) Using the automated Box-Cox Procedure, determine which transformation of Y would be appropriate (if any)?

By using the biggest log-likelihood, I got $\lambda = 0.03030303$

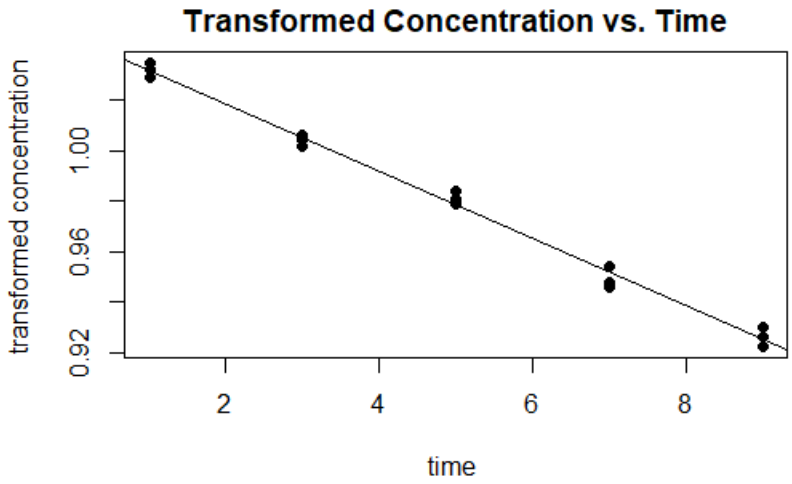
By using the smallest SSE, I got $\lambda = 0$. when $\lambda = 0$, the Y is transformed to lnY.

I would pick $\lambda = 0.03030303$.

So, raise Y to 0.03030303, or Y'=$Y^{0.03030303}$ would be appropriate transformation of Y.

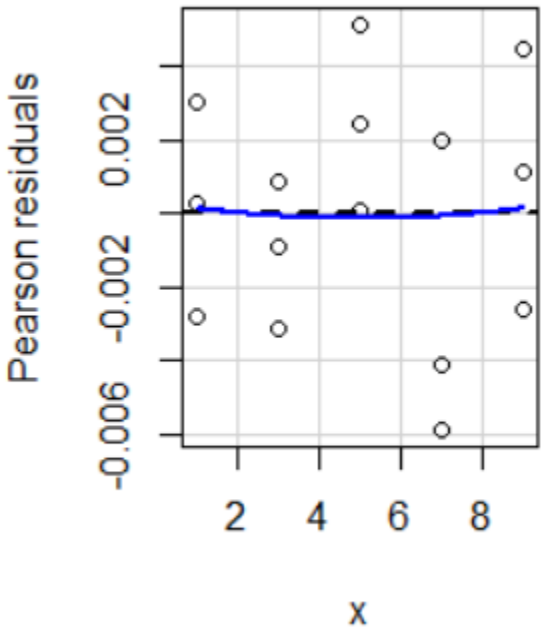
f) (3) Using the appropriate transformation, fit a new model and repeat the diagnostics in part d)

i. Figure 5.



From Figure 5, there is a strong linear relationship between time and transformation of concentration and there are no unusual points.

ii. Figure 6.



From Figure 6, the distribution of residuals has a random pattern around 0 line, which means the relationship is linear is reasonable. And no one residual stands out from the basic random pattern of residuals. This suggests that there are no outliers.

iii.

1. From Figure 6 in the last question, fortunately, it is also a plot of residuals versus time (x), we can easily find there is a random pattern in the plot. Hence, the residuals are independent with time.

2. H_0: Data follows normal distribution

   H_a: Data violates normal distribution

   Refer to Figure 8., P-value = 0.8324 > significance level $\alpha = 0.05$, and it is much bigger than the significance level, so the distribution of the residuals is a normal distribution. Also, from the Normal probability plot, the residual points fit the line very well.

Figure 7.

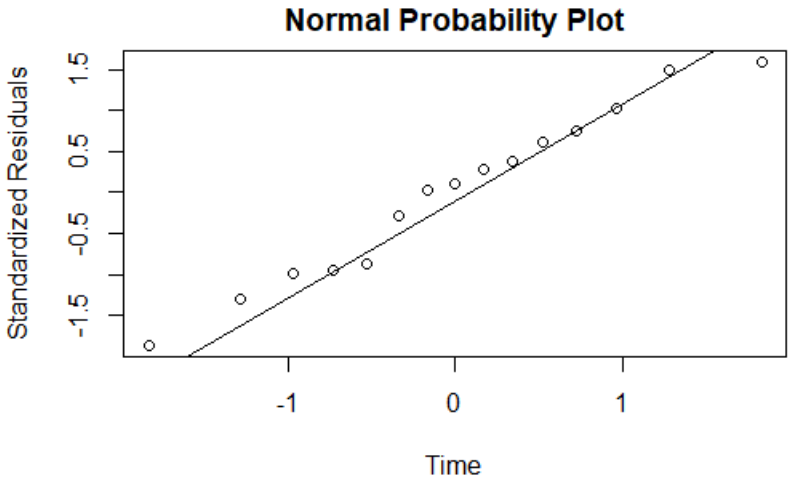


Figure 8,

```
        Shapiro-Wilk normality test

data:  concentration.stdres1
W = 0.96832, p-value = 0.8324
```

3. H_0: Residuals have constant variances

   H_a: Residuals have non-constant variances

   Refer to Figure 9., P-value = 0.13685> significance level $\alpha = 0.05$, so we do not reject H_0. Hence, we are 95% confident that the residuals have constant variances.

   Figure 9.

```
      t.value   P.Value alpha df
[1,] 1.585583 0.1368489  0.05 13
```

i.

We do not need to do a transformation on X, because the relationship between explanatory variables and dependent variables are strongly linear.

We do not need to do a transformation on Y, because the distributions of residuals has a random pattern around 0 line, and there are no outliers, from the Shapiro test and BF test, we can also see that the error terms are normal and have constant variances.

g) With back-transformation, compute **and interpret** the confidence interval for $\hat{Y}_h$ when $X_h = 7.5$.

From R output, the confidence interval for $\hat{Y}_h$ when $X_h = 7.5$ for the transformed Y should be (0.9427026, 0.9476571)

	x <dbl>	Fit <dbl>	Lower.Band <dbl>	Upper.Band <dbl>
1	7.5	0.9451798	0.9427026	0.9476571

1 row

Then, we need to do the back-transformation, i.e.,

since Y'=$Y^{0.03030303}$, Y = $Y'^{\frac{1}{0.03030303}}$, and we get the confidence interval for $\hat{Y}_h$ when $X_h = 7.5$ for Y is:

(0.142680707, 0.1696251627)

We conclude with confidence coefficient 0.95 that the prediction in concentration is somewhere between 0.14268 and 0.16963 for $X_h = 7.5$.

2. (3) If the error terms in a regression model are independent and Normally distributed, $N(0, \sigma)$, what can be said about the error terms after transformation $X' = 1/X$ is used? Is the situation the same after transformation $Y' = 1/Y$ is used?

Since the error terms are independent and normally distributed, which means the error terms has a normal distribution and have constant variance, the transformation of X may have almost no influence on the distribution of error terms. However, the transformation of Y may change the shape of the distribution of the error terms from the normal distribution and may also lead to substantially differing error term variances.

3. (12) The following data were obtained in a study of the relation between diastolic blood pressure (Y) and age (X) for boys 5 to 13 years old

i	1	2	3	4	5	6	7	8
X	5	8	11	7	13	12	12	6
Y	63	67	74	64	75	69	90	60

a) (3) Assuming SLM is appropriate, obtain the estimated regression function and plot the residuals against X. What does your residual plot show?

$\bar{X}$ = 9.25

$\bar{Y}$ = 70.25

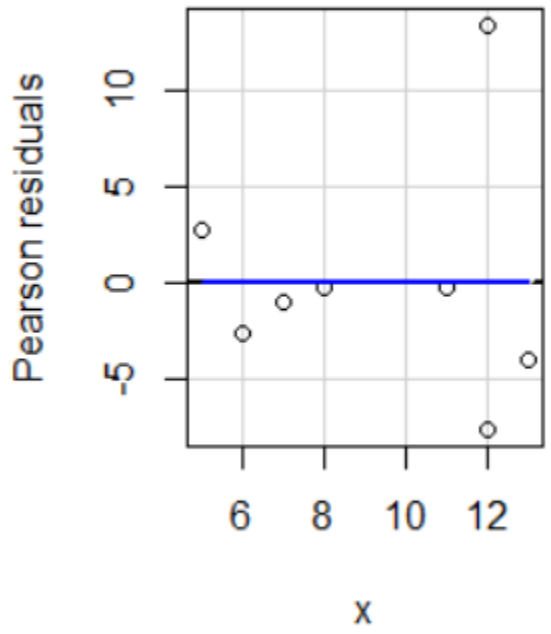
$SS_{XY} = \sum(X_i - \bar{X})(Y_i - \bar{Y})$ = 157.5

$SS_X = \sum(X_i - \bar{X})^2$ = 67.5

$b_1 = \frac{SS_{XY}}{SS_X} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{157.5}{67.5}$ = 2.33333

$b_0 = \bar{Y} - b_1\bar{X}$ = 70.25 –2.33333*9.25 = 48.6667

$\hat{Y} = b_0 + b_1X$ = 48.6667 + 2.33333X



The plot shows a random pattern, but it has an obvious usual point at X=12.

b) (2) Omit case 7 from the data and obtain the regression function based on the remaining cases. Compare the estimated regression function to a). What can you conclude about the effect of case 7?

$\bar{X}$ = 8.857143

$\bar{Y}$ = 67.42857

$SS_{XY} = \sum(X_i - \bar{X})(Y_i - \bar{Y})$ = 95.42857

$SS_X = \sum(X_i - \bar{X})^2$ = 58.85714

$b_1 = \frac{SS_{XY}}{SS_X} = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2} = \frac{95.42857}{58.85714}$ = 1.62136

$b_0 = \bar{Y} - b_1\bar{X}$ = 67.42857 –1.62136*8.857143 = 53.06795

$\hat{Y} = b_0 + b_1X$ = 53.06795+ 1.62136X

Both b_0 and b_1 are smaller than that in part (a). The case 7 highers them a lot and it is the obvious unusual point as we can see from the residual plot in part (a).

c). (2) use your fitted regression function in b), obtain a 99% prediction interval for a new Y observation at X=12.

X = X_h = 12

$\hat{Y}_h = b_0 + b_1 X_h$ = 53.06795 + 1.62136*12 = 72.52427

$MSE = s^2 = \frac{\sum(Y_i-\hat{Y}_i)^2}{n-2} = \frac{34.99029}{7-2}$ =6.998058

$s^2_{\{pred\}} = s^2\{\hat{Y}_h\} + s^2 = MSE[\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2} + 1] = s^2[\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2} + 1]$ = 6.998058 *[$\frac{1}{7}$ + $\frac{9.87755}{58.85714}$ + 1]= 9.1722

so $s_{\{pred\}} = \sqrt{9.1722}$ = 3.0286

The 99% confidence interval is

$\hat{Y}_h \pm t(1 - \frac{\alpha}{2}; n-2)s_{\{pred\}}$= 72.52427 $\pm$ 4.032*3.0286 = (60.31295, 84.73559)

We conclude with confidence coefficient 0.99 that the mean blood pressure is somewhere between 60.31295and 84.73559 when X=12.

d) Use on the data in b),

i. (2) Calculate and interpret the 90% **individual** confidence intervals for β_0 and then for β_1.

Intercept:

$s\{b_0\} = \sqrt{MSE[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i-\bar{X})^2}]} = \sqrt{6.998058 * [\frac{1}{7} + \frac{78.44898}{58.85714}]}$ = 3.2136

Hence the 90% confidence interval for β_0 is

$b_0 \pm t\,(1-\frac{\alpha}{2}; n-2)\, s\{b_0\}$ = 53.06795 $\pm$ 2.015 * 3.2136=(46.592546, 59.543354)

We are 90% confident that the intercept parameter β_0 falls between 46.592546 and 59.543354.

Slope:

$s\{b_1\} = \sqrt{\frac{MSE}{\sum(X_i-\bar{X})^2}} = \sqrt{\frac{6.998058}{58.85714}} = 0.3448$

Hence the 90% confidence interval for β_1 is

$b_1 \pm t(1-\frac{\alpha}{2}; n-2)\, s\{b_1\} = 1.62136 \pm 2.015 * 0.3448 = (0.926588, 2.316132)$

It means we are 90% confident that the blood pressure will increase by at least 0.926588 and most 2.316132 when increase one year old.

ii. (2) Obtain the Bonferroni **joint** confidence intervals for β_0 and β_1 using an $\alpha = 0.10$. Interpret your confidence intervals.

$B = t(1-\frac{\frac{\alpha}{2}}{2}; n-2) = t(1-\frac{\alpha}{4}; n-2) = t(0.975;5) = 2.571$

Hence the 90% Bonferroni **joint** confidence interval for β_0 is

$b_0 \pm B\, s\{b_0\} = 53.06795 \pm 2.571 * 3.2136 = (44.80578, 61.33012)$

Hence the 90% Bonferroni **joint** confidence interval for β_1 is

$b_1 \pm B\, s\{b_1\} = 1.62136 \pm 2.571 * 0.3448 = (0.73488, 2.50784)$

Thus, we conclude that β_0 is between 44.80578 and 61.33012 and β_1 is between 0.73488, 2.50784.

(The family confidence coefficient is at least 0.90 that the procedure leads to correct pairs of interval estimates.)

iii. (1) Compare your answers in part i) and ii). Which one of the parts has a larger confidence interval? Why?

Part ii) has a larger confidence interval.

The $1 - \alpha = 1 - 0.1 = 0.9$ joint(family) confidence interval is done by estimating β_0 and β_1 separately with individual confidence level of $1 - \frac{\alpha}{2} = 1 - \frac{0.1}{2} = 0.95$ each.

So, the confidence interval in part ii) is calculated by 95% confidence level, whereas the confidence interval in part i) is calculated by 90% confidence level.

Hence, the 95% confidence interval is larger than 90% confidence interval.

4. (10 pts.) Based on the following small data set, **construct by hand** the design matrix, **X**, its transpose **X'**, and the matrices **X'X**, $(\mathbf{X'X})^{-1}$, **X'Y**, $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$, the variance-covariance matrix of $\Sigma\{b\}$ and $\Sigma\{\hat{Y}\}$ (i.e., $\boldsymbol{s^2\{b\}}$ and $\mathbf{s^2}\{\widehat{\boldsymbol{Y}}_{\boldsymbol{h}}\}$ in matrix form.

X	Y
2	1
3	2
6	4
7	5
9	7

4. $X = \begin{pmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 6 \\ 1 & 7 \\ 1 & 9 \end{pmatrix} \quad Y = \begin{pmatrix} 1 \\ 2 \\ 4 \\ 5 \\ 7 \end{pmatrix}$

$X^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 6 & 7 & 9 \end{pmatrix}$

$X^TX = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 6 & 7 & 9 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 6 \\ 1 & 7 \\ 1 & 9 \end{pmatrix} = \begin{pmatrix} 5 & 27 \\ 27 & 179 \end{pmatrix}$

$[A \quad I] = [I \quad A^{-1}]$

$\left(\begin{array}{cc|cc} 5 & 27 & 1 & 0 \\ 27 & 179 & 0 & 1 \end{array}\right) \rightarrow \left(\begin{array}{cc|cc} 1 & \frac{27}{5} & \frac{1}{5} & 0 \\ 27 & 179 & 0 & 1 \end{array}\right) \rightarrow \left(\begin{array}{cc|cc} 1 & \frac{27}{5} & \frac{1}{5} & 0 \\ 0 & \frac{166}{5} & -\frac{27}{5} & 1 \end{array}\right)$

$\rightarrow \left(\begin{array}{cc|cc} 1 & \frac{27}{5} & \frac{1}{5} & 0 \\ 0 & & \frac{-27}{166} & \frac{5}{166} \end{array}\right) \rightarrow \left(\begin{array}{cc|cc} 1 & 0 & \frac{179}{166} & \frac{-27}{166} \\ 0 & 1 & \frac{-27}{166} & \frac{5}{166} \end{array}\right)$

$\underbrace{\qquad\qquad}_{(X^TX)^{-1}}$

$(X^TX)^{-1} = \begin{pmatrix} \frac{179}{166} & \frac{-27}{166} \\ \frac{-27}{166} & \frac{5}{166} \end{pmatrix}$

$X^TY = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 6 & 7 & 9 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 4 \\ 5 \\ 7 \end{pmatrix} = \begin{pmatrix} 19 \\ 130 \end{pmatrix}$

$b = (X^TX)^{-1}X^TY = \begin{pmatrix} \frac{179}{166} & \frac{-27}{166} \\ \frac{-27}{166} & \frac{5}{166} \end{pmatrix} \begin{pmatrix} 19 \\ 130 \end{pmatrix} = \begin{pmatrix} \frac{-109}{166} \\ \frac{137}{166} \end{pmatrix}$

$H = X(X^TX)^{-1}X^T$

$= \begin{pmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 6 \\ 1 & 7 \\ 1 & 9 \end{pmatrix} \begin{pmatrix} \frac{179}{166} & \frac{-27}{166} \\ \frac{-27}{166} & \frac{5}{166} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 6 & 7 & 9 \end{pmatrix}$

$= \begin{pmatrix} \frac{125}{166} & \frac{-17}{166} \\ \frac{49}{83} & \frac{-6}{83} \\ \frac{17}{166} & \frac{3}{166} \\ \frac{-5}{83} & \frac{4}{83} \\ \frac{-32}{83} & \frac{9}{83} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 6 & 7 & 9 \end{pmatrix} = \begin{pmatrix} \frac{91}{166} & \frac{37}{83} & \frac{23}{166} & \frac{3}{83} & \frac{-14}{83} \\ \frac{37}{83} & \frac{31}{83} & \frac{13}{83} & \frac{7}{83} & \frac{-5}{83} \\ \frac{23}{166} & \frac{13}{83} & \frac{35}{166} & \frac{19}{83} & \frac{22}{83} \\ \frac{3}{83} & \frac{7}{83} & \frac{19}{83} & \frac{23}{83} & \frac{31}{83} \\ \frac{-14}{83} & \frac{-5}{83} & \frac{22}{83} & \frac{31}{83} & \frac{49}{83} \end{pmatrix}$

$$e = (I-H)Y = \begin{pmatrix} \frac{75}{166} & \frac{-37}{83} & \frac{-23}{166} & \frac{-3}{83} & \frac{14}{83} \\ \frac{-37}{83} & \frac{52}{83} & \frac{-13}{83} & \frac{-7}{83} & \frac{5}{83} \\ \frac{-23}{166} & \frac{-13}{83} & \frac{131}{166} & \frac{-19}{83} & \frac{-22}{83} \\ \frac{-3}{83} & \frac{-7}{83} & \frac{-19}{83} & \frac{60}{83} & \frac{-31}{83} \\ \frac{14}{83} & \frac{5}{83} & \frac{-22}{83} & \frac{-31}{83} & \frac{34}{83} \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 4 \\ 5 \\ 7 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{166} \\ \frac{15}{83} \\ \frac{-49}{166} \\ \frac{-10}{83} \\ \frac{19}{83} \end{pmatrix}$$

$$SSE = e^T e = \begin{pmatrix} \frac{1}{166} & \frac{15}{83} & \frac{-49}{166} & \frac{-10}{83} & \frac{19}{83} \end{pmatrix} \begin{pmatrix} \frac{1}{166} \\ \frac{15}{83} \\ \frac{-49}{166} \\ \frac{-10}{83} \\ \frac{19}{83} \end{pmatrix} = \frac{31}{166} \approx 0.18675$$

$$MSE = \frac{SSE}{5-2} = \frac{\frac{31}{166}}{3} \approx 0.062249$$

$$\Sigma\{b\} = \sigma^2 (X^T X)^{-1} = MSE(X^T X)^{-1} = 0.062249 \begin{pmatrix} \frac{179}{166} & \frac{-27}{166} \\ \frac{-27}{166} & \frac{5}{166} \end{pmatrix}$$

$$= \begin{pmatrix} 0.06712392 & -0.01012484 \\ -0.01012484 & 0.00187497 \end{pmatrix}$$

let $X_h = 3$

$$\Sigma\{\hat{Y}\} = X_h' \Sigma\{b\} X_h$$

$$= \begin{pmatrix} 1 & 3 \end{pmatrix} \begin{pmatrix} 0.06712392 & -0.01012484 \\ -0.01012484 & 0.00187497 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \end{pmatrix}$$

$$= \begin{pmatrix} 0.0367494 & -0.00449993 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \end{pmatrix}$$

$$= 0.02324961$$

5. (10) Describe your project and describe the response variable and (what you think) the most relevant exploratory variable (X). Use R to build a SLR between your response variable. Include and assessment of the key regression assumptions. If the assumptions are not met, include and justify appropriate remedial measures. Use the final model to predict $\hat{Y}_h$ when X_h = the sample mean. Discuss the strengths and weaknesses of the final model.

R code for Question 1 is attached in the Appendix.

Our project topic is aimed at understanding what influences happiness in the world, and we import the 2017 WHR collected data from 3,000 people on average for 155 countries using the question based on life evaluation and the countries are ranked by their correlated happiness levels. This rank is then compared to six key variables: Economy (GDP capita), Family (Social support), Health (Life Expectancy), Freedom (to make life choices), Trust (Absence of Government Corruption), and Generosity. Our research is to find out what and how of the variables affect the happiness score of a country. Among these variables, the most relevant exploratory variable (X) is Economy (GDP capita).

Let Y = happiness score, X = Economy

From R output:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.2032     0.1356   23.62   <2e-16 ***
x             2.1842     0.1267   17.24   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6617 on 153 degrees of freedom
Multiple R-squared:  0.6601,    Adjusted R-squared:  0.6579
F-statistic: 297.1 on 1 and 153 DF,  p-value: < 2.2e-16
```

the fitted regression line:

$\hat{Y} = b_0 + b_1 X = 3.2032 - 2.1842X$

The plot of Happiness score versus Economy is:

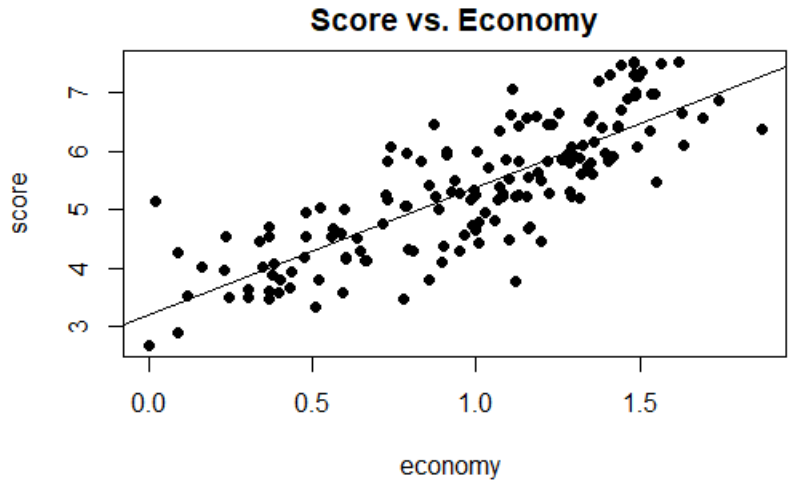


Figure 1.

From Figure 1, there is a linear relationship between Economy and Happiness score and there are almost no unusual points.

The plot of the residuals versus the explanatory variable and the plot of the residuals versus the fitted value:

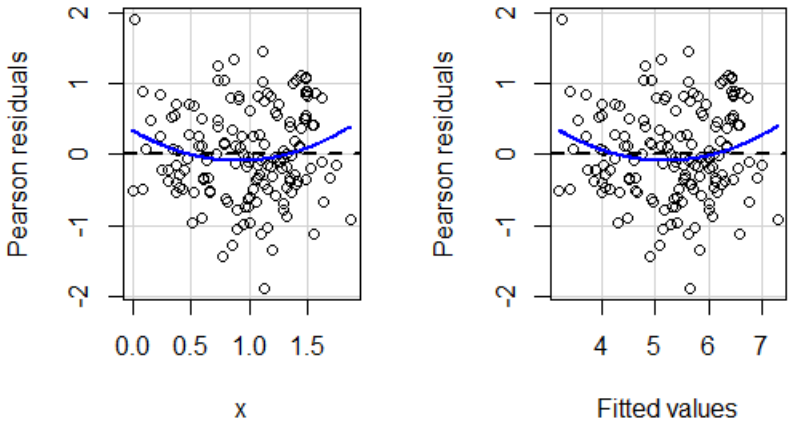


Figure 2.

From Figure 2, the distribution of residuals has a random pattern around 0 line, which means the relationship is linear is reasonable. And no one residual stands out from the basic random pattern of residuals. This suggests that there are no outliers.

The plot of normal probability of the residuals:

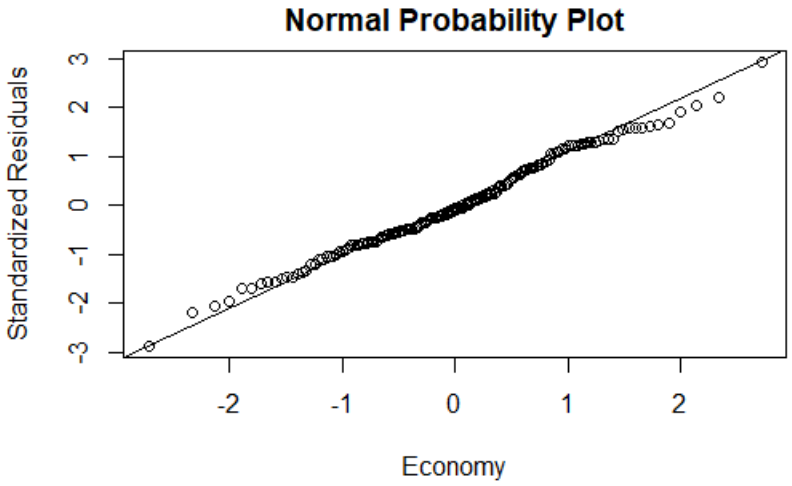


Figure 3.

```
        Shapiro-Wilk normality test

data:  score.stdres
W = 0.992, p-value = 0.5398
```

Figure 4.

H_0: Data follows normal distribution

H_a: Data violates normal distribution

Refer to Figure 4., P-value = 0.5398 > significance level $\alpha = 0.05$, so we do not reject H_0. Hence, we are 95% confident that the data follows normal distribution, the normality assumption appears to be reasonable here.

Brown-Forsythe test

```
       t.value   P.Value alpha  df
[1,] 0.4644873 0.6429589  0.05 153
```

Figure 5.

H_0: Residuals have constant variances

H_a: Residuals have non-constant variances

Refer to Figure 5., P-value = 0.6429589> significance level $\alpha = 0.05$, so we do not reject H_0. Hence, we are 95% confident that the residuals have constant variances, and this support my preliminary findings in Figure 2., where we can see the residuals scattered randomly around zero, which means they have constant variance.

Because the relationship between Happiness score and Economy is linear, and the distribution of residual is normal and the error terms have constant variance, we do not need to do transformation on X or Y. Hence, I used the original Y as the final model.

```
Economy..GDP.per.Capita.
Min.   :0.0000
1st Qu.:0.6634
Median :1.0646
Mean   :0.9847
3rd Qu.:1.3180
Max.   :1.8708
```

From R output, X_h = the sample mean=0.9847

	x <dbl>	Fit <dbl>	Lower.Band <dbl>	Upper.Band <dbl>
1	0.9847	5.35398	5.248985	5.458975

From R output, the confidence interval for $\hat{Y}_h$ when $X_h = 0.9847$

should be (5.248985, 5.45875)

The strength of the model is that the error terms are normal distributed and have constant variances.

The weakness of the model is that the $R^2 = 0.6601$ is not very close to 1.

Appendix ---R code

HW3 Q1

```
concentration<-read.table("C:/Users/candi/Desktop/STAT 512/concentration.csv",header=TRUE,sep = ",")
concentration
```

```
##       Y X
## 1  0.07 9
## 2  0.09 9
## 3  0.08 9
## 4  0.16 7
## 5  0.17 7
## 6  0.21 7
## 7  0.49 5
## 8  0.58 5
## 9  0.53 5
## 10 1.22 3
## 11 1.15 3
## 12 1.07 3
## 13 2.84 1
## 14 2.57 1
## 15 3.10 1
```

```
library(alr4)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

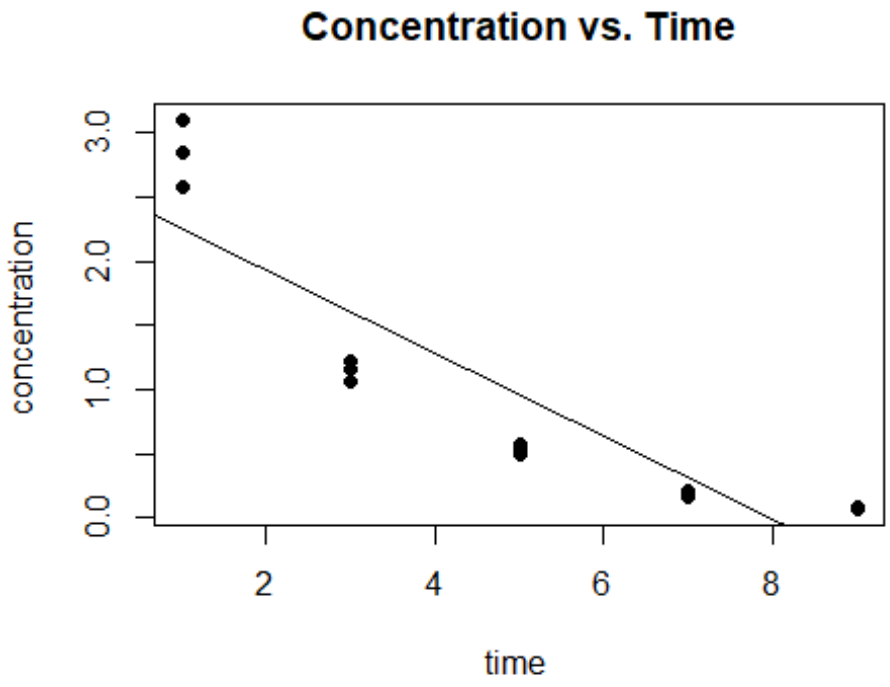
```
## Loading required package: effects
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

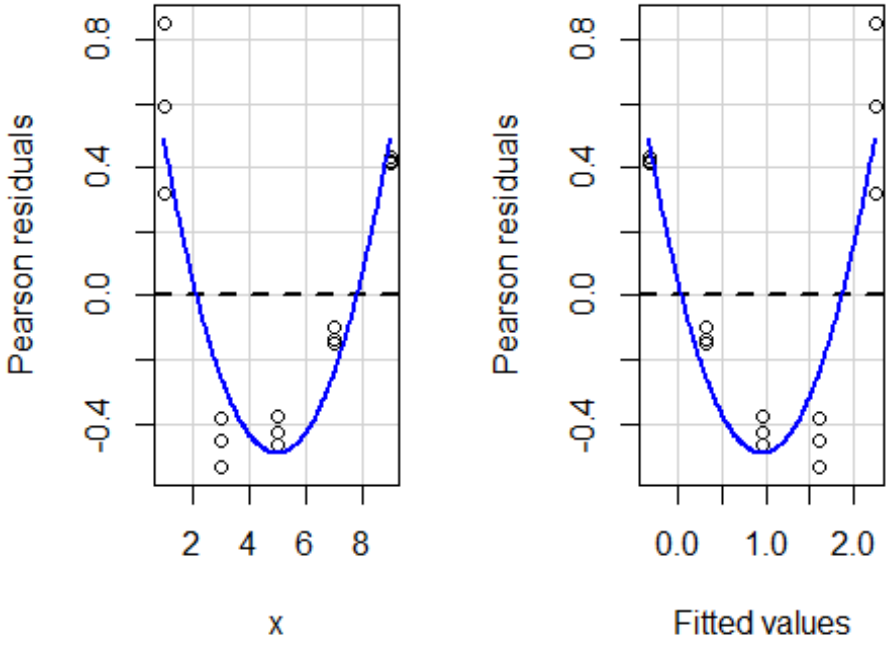
```
x<-concentration$X
y<-concentration$Y
fit=lm(y~x,data = concentration)
concentration.stdres=rstandard(fit)
fit
```

```
##
## Call:
## lm(formula = y ~ x, data = concentration)
##
## Coefficients:
## (Intercept)            x
##       2.575       -0.324
```

```
#Question 1 (d)i
plot(x,y, main="Concentration vs. Time",
xlab="time ", ylab="concentration ", pch=19)
abline(fit)
```




```
#Question 1 (d)ii
residualPlots(fit, tests=TRUE, quadratic=TRUE, smooth=FALSE)
```



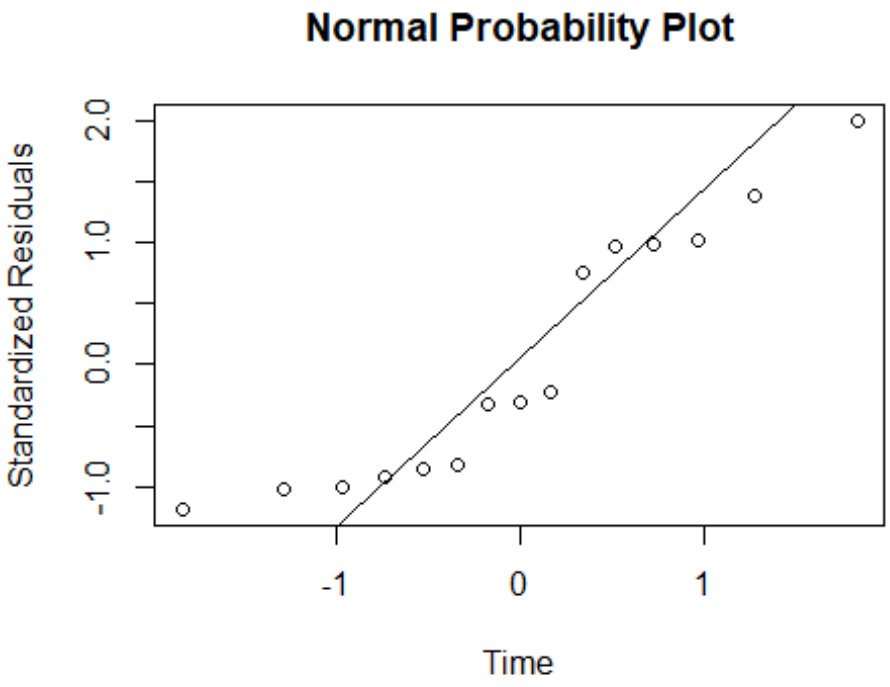

```
##            Test stat Pr(>|Test stat|)
## x             8.8458        1.325e-06 ***
```

```
## Tukey test    8.8458        < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Question 1 (d)iii
shapiro.test(concentration.stdres)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  concentration.stdres
## W = 0.88498, p-value = 0.05634
```

```
qqnorm(concentration.stdres, ylab="Standardized Residuals",
       xlab="Time", main="Normal Probability Plot")
qqline(concentration.stdres)
```




```
library(ALSM)
```

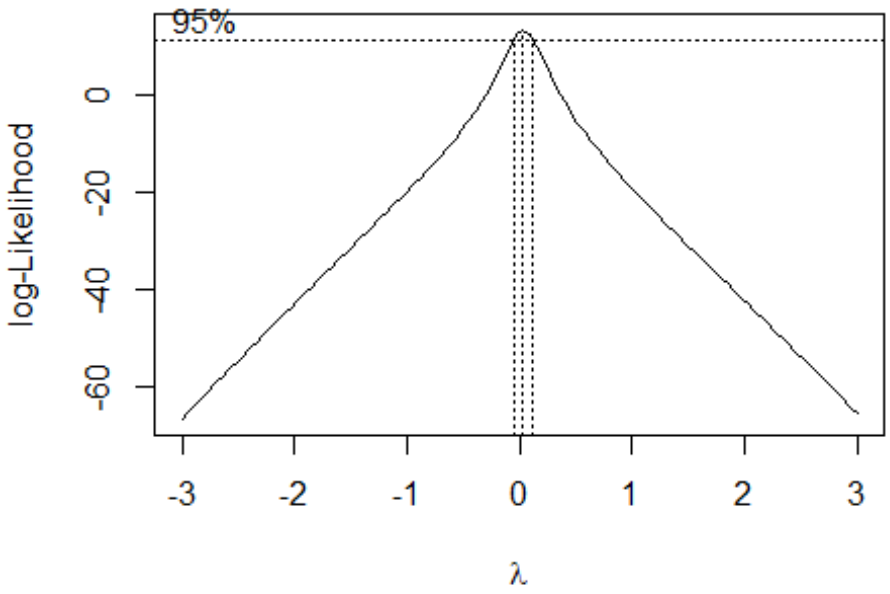
```
## Loading required package: leaps
```

```
## Loading required package: SuppDists
```

```
g<-rep(1,15)
g[x<=5]=0
bftest(fit,g)
```

```
##        t.value  P.Value alpha df
## [1,] 0.4469581 0.662261  0.05 13
```

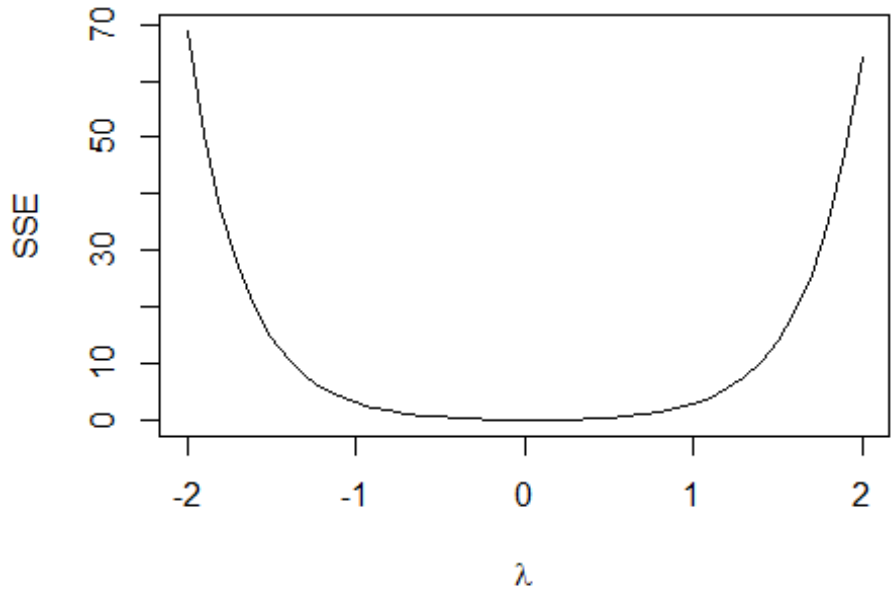
```r
#Question 1 (e)
library(MASS)
bcmle<-boxcox(fit,lambda=seq(-3,3, by=0.1))
```




```r
lambda<-bcmle$x[which.max(bcmle$y)]
lambda
```

```
## [1] 0.03030303
```

```r
library(ALSM)
bcsse<-boxcox.sse(concentration$X,concentration$Y,l=seq(-2,2,0.1))
```



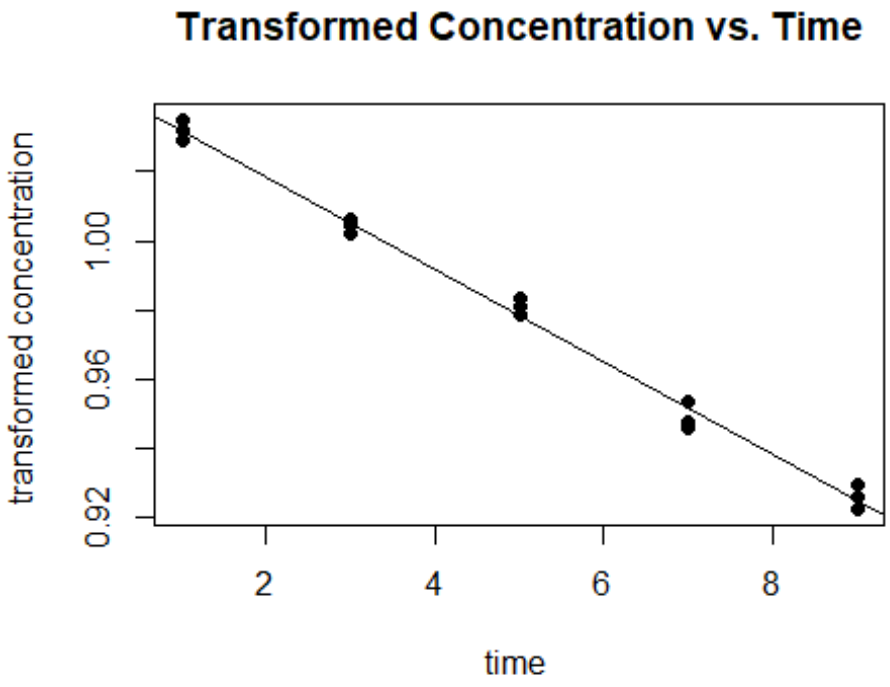

```
lambda<-bcsse$lambda[which.min(bcsse$SSE)]
lambda
```

```
## [1] 0
```

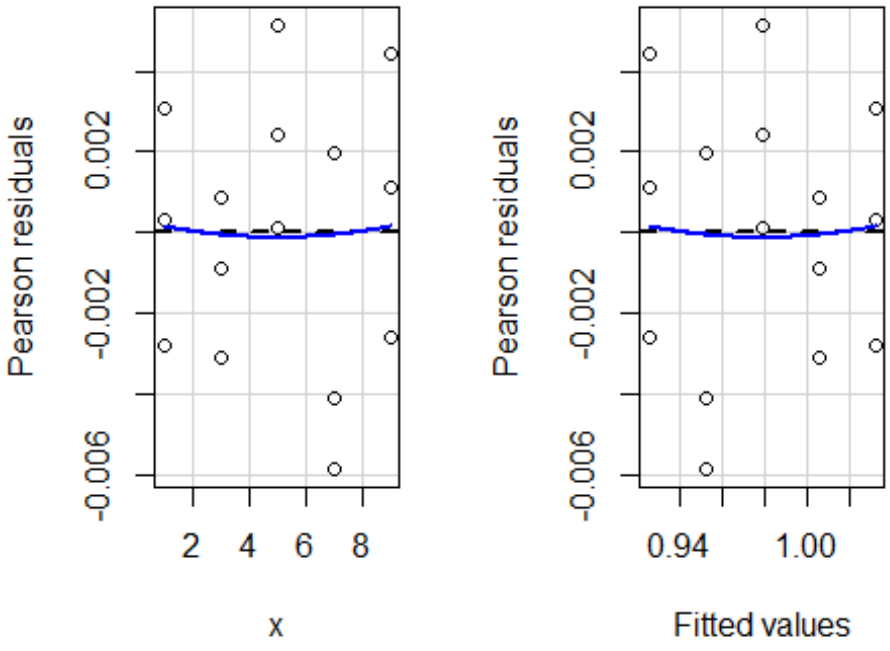
```
#Question 1 (f) i
y1<-y^0.03030303
y1
```

```
##  [1] 0.9225777 0.9296305 0.9263184 0.9459810 0.9477205 0.9538085 0.9786153
##  [8] 0.9836286 0.9809451 1.0060440 1.0042442 1.0020524 1.0321360 1.0290162
## [15] 1.0348794
```

```
fit1=lm(y1~x,data = concentration)
plot(x,y1, main="Transformed Concentration vs. Time",
     xlab="time ", ylab="transformed concentration ", pch=19)
abline(fit1)
```




```r
#Question 1 (f) ii
residualPlots(fit1, tests=TRUE, quadratic=TRUE, smooth=FALSE)
```

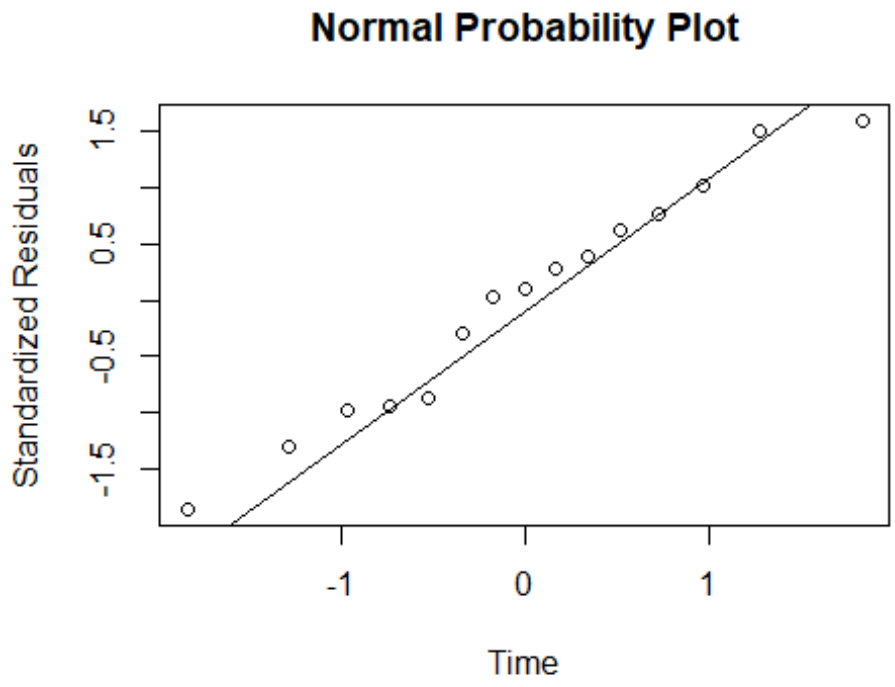


```
##            Test stat Pr(>|Test stat|)
## x             0.1288           0.8997
## Tukey test    0.1288           0.8976
```

```
#Question 1 (f)iii
concentration.stdres1=rstandard(fit1)
shapiro.test(concentration.stdres1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  concentration.stdres1
## W = 0.96832, p-value = 0.8324
```

```
qqnorm(concentration.stdres1, ylab="Standardized Residuals",
       xlab="Time", main="Normal Probability Plot")
qqline(concentration.stdres1)
```




```
#summary(concentration)
library(ALSM)
g<-rep(1,15)
g[x<=5]=0
bftest(fit1,g)
```

```
##       t.value   P.Value alpha df
## [1,] 1.585583 0.1368489  0.05 13
```

```
#Question 1 (g)
summary(fit1)
```

```
##
## Call:
## lm(formula = y1 ~ x, data = concentration)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0058642 -0.0027096  0.0003068  0.0022010  0.0051221
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.0451599  0.0017450  598.96  < 2e-16 ***
## x           -0.0133307  0.0003038  -43.89 1.62e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.003328 on 13 degrees of freedom
## Multiple R-squared:  0.9933, Adjusted R-squared:  0.9928
## F-statistic:  1926 on 1 and 13 DF,  p-value: 1.619e-15
```

```
library(ALSM)
fit1
```

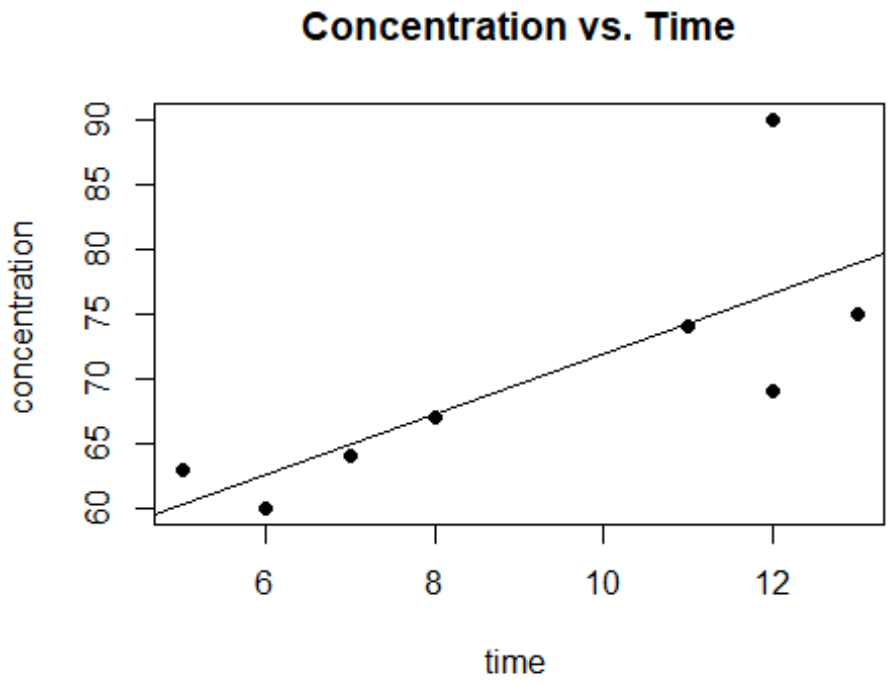
```
##
## Call:
## lm(formula = y1 ~ x, data = concentration)
##
## Coefficients:
## (Intercept)            x
##     1.04516     -0.01333
```

```
new<-data.frame(x=7.5)
ci.reg(fit1, new, type='m',alpha=0.05)
```

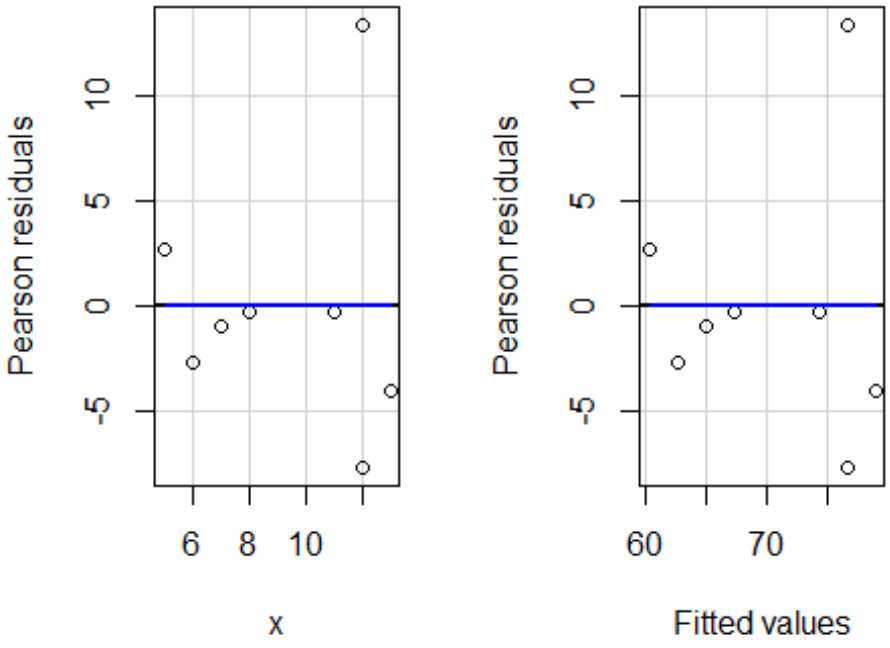
```
##     x       Fit Lower.Band Upper.Band
## 1 7.5 0.9451798  0.9427026  0.9476571
```

HW3 Q3

```
blood<-read.table("C:/Users/candi/Desktop/STAT 512/HWQ3.csv",header=TRUE,sep = ",")
library(alr4)

## Loading required package: car

## Loading required package: carData

## Loading required package: effects

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.

y<-blood$Y
x<-blood$X
fit=lm(y~x,data = blood)
blood.stdres=rstandard(fit)
fit

##
## Call:
## lm(formula = y ~ x, data = blood)
##
## Coefficients:
## (Intercept)            x
##      48.667        2.333

#Question 3
plot(x,y, main="Concentration vs. Time",
xlab="time ", ylab="concentration ", pch=19)
abline(fit)
```




```r
#Question 3 (a)
residualPlots(fit, tests=TRUE, quadratic=TRUE, smooth=FALSE)
```

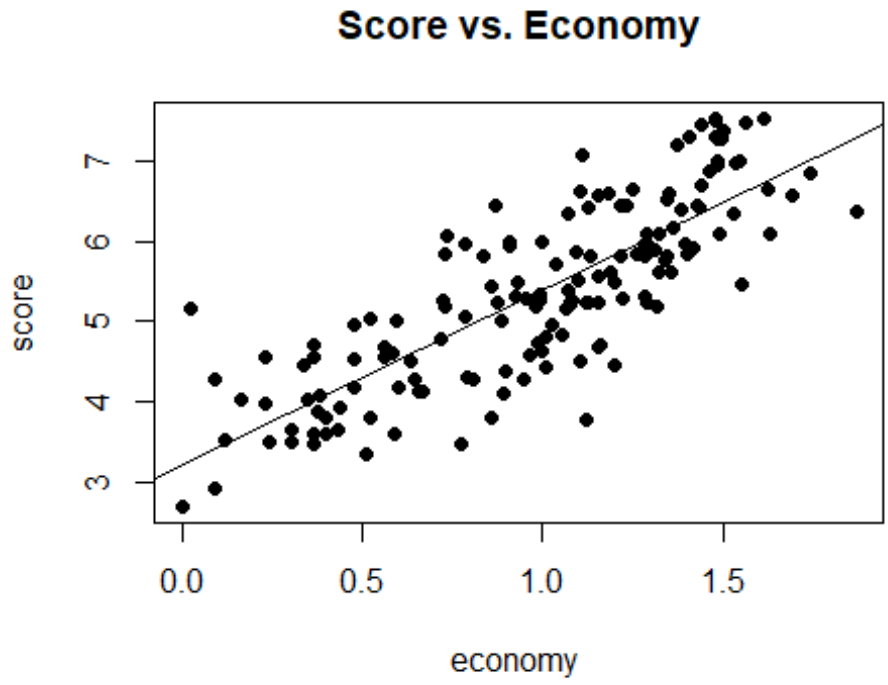


```
##            Test stat Pr(>|Test stat|)
## x                  0                1
## Tukey test         0                1
```

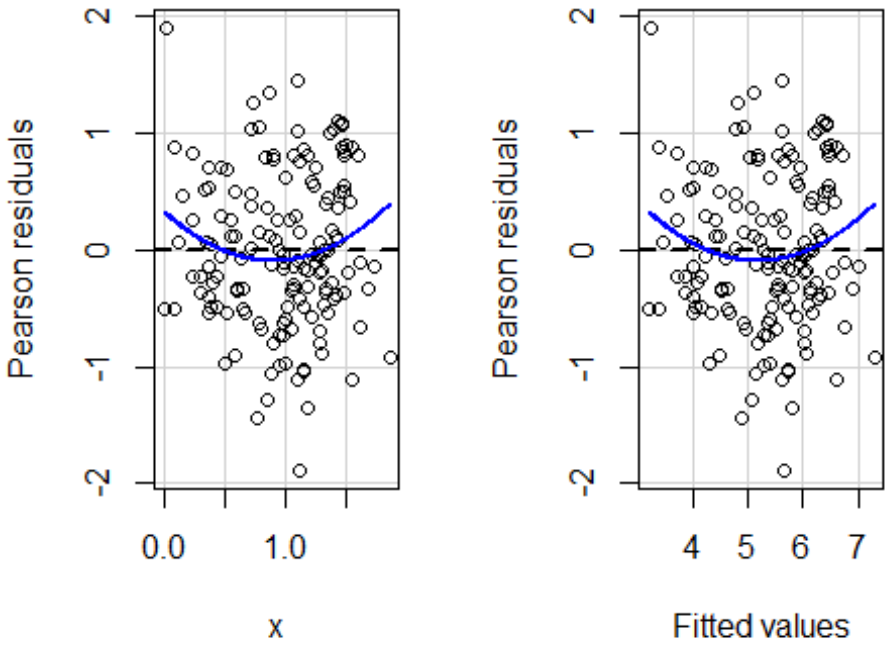
HW3 Q5

```
score<-read.table("C:/Users/candi/Desktop/STAT 512/2017.csv",header=TRUE,sep = ",")
library(alr4)

## Loading required package: car

## Loading required package: carData

## Loading required package: effects

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.

y<-score$Happiness.Score
x<-score$Economy..GDP.per.Capita.
fit=lm(y~x,data = score)
score.stdres=rstandard(fit)
fit

##
## Call:
## lm(formula = y ~ x, data = score)
##
## Coefficients:
## (Intercept)            x
##       3.203        2.184

summary(fit)

##
## Call:
## lm(formula = y ~ x, data = score)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.88807 -0.45200 -0.05328  0.49425  1.89833
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2032     0.1356   23.62   <2e-16 ***
## x             2.1842     0.1267   17.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6617 on 153 degrees of freedom
## Multiple R-squared:  0.6601, Adjusted R-squared:  0.6579
## F-statistic: 297.1 on 1 and 153 DF,  p-value: < 2.2e-16
```

```
plot(x,y, main="Score vs. Economy",
xlab="economy ", ylab="score", pch=19)
abline(fit)
```




```
residualPlots(fit, tests=TRUE, quadratic=TRUE, smooth=FALSE)
```

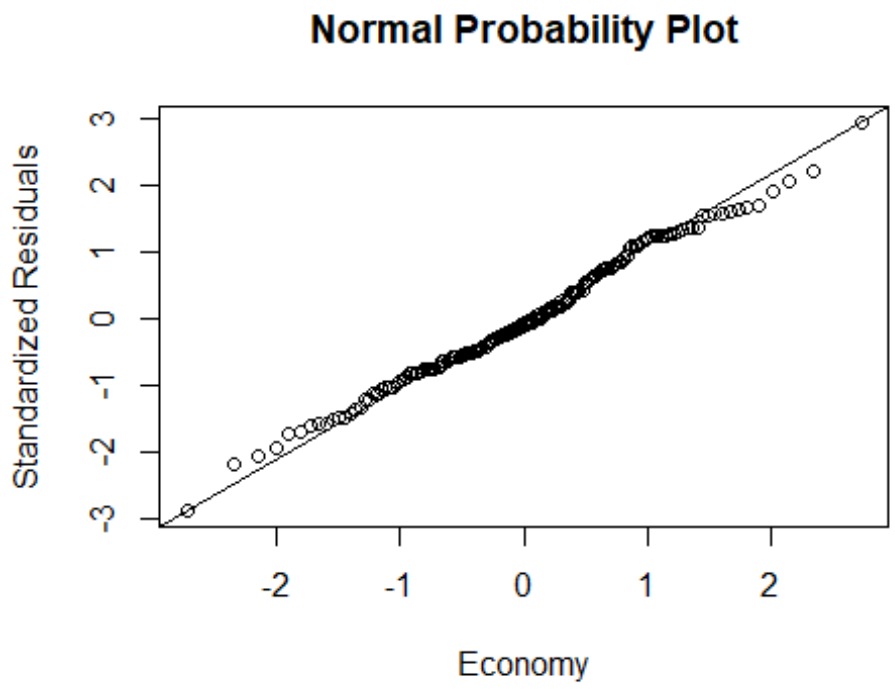


```
##            Test stat Pr(>|Test stat|)
## x             1.8847          0.06138 .
## Tukey test    1.8847          0.05948 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
shapiro.test(score.stdres)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  score.stdres
## W = 0.992, p-value = 0.5398
```

```
qqnorm(score.stdres, ylab="Standardized Residuals",
       xlab="Economy", main="Normal Probability Plot")
qqline(score.stdres)
```




```
summary(score)
```

```
##         Country    Happiness.Rank  Happiness.Score  Whisker.high
##  Afghanistan:  1   Min.   :  1.0   Min.   :2.693   Min.   :2.865
##  Albania    :  1   1st Qu.: 39.5   1st Qu.:4.505   1st Qu.:4.608
##  Algeria    :  1   Median : 78.0   Median :5.279   Median :5.370
##  Angola     :  1   Mean   : 78.0   Mean   :5.354   Mean   :5.452
##  Argentina  :  1   3rd Qu.:116.5   3rd Qu.:6.101   3rd Qu.:6.195
##  Armenia    :  1   Max.   :155.0   Max.   :7.537   Max.   :7.622
##  (Other)    :149
##   Whisker.low    Economy..GDP.per.Capita.     Family
```

```
##  Min.   :2.521   Min.   :0.0000           Min.   :0.000
##  1st Qu.:4.375   1st Qu.:0.6634           1st Qu.:1.043
##  Median :5.193   Median :1.0646           Median :1.254
##  Mean   :5.256   Mean   :0.9847           Mean   :1.189
##  3rd Qu.:6.007   3rd Qu.:1.3180           3rd Qu.:1.414
##  Max.   :7.480   Max.   :1.8708           Max.   :1.611
##
##  Health..Life.Expectancy.    Freedom         Generosity
##  Min.   :0.0000           Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.3699           1st Qu.:0.3037   1st Qu.:0.1541
##  Median :0.6060           Median :0.4375   Median :0.2315
##  Mean   :0.5513           Mean   :0.4088   Mean   :0.2469
##  3rd Qu.:0.7230           3rd Qu.:0.5166   3rd Qu.:0.3238
##  Max.   :0.9495           Max.   :0.6582   Max.   :0.8381
##
##  Trust..Government.Corruption. Dystopia.Residual
##  Min.   :0.00000               Min.   :0.3779
##  1st Qu.:0.05727               1st Qu.:1.5913
##  Median :0.08985               Median :1.8329
##  Mean   :0.12312               Mean   :1.8502
##  3rd Qu.:0.15330               3rd Qu.:2.1447
##  Max.   :0.46431               Max.   :3.1175
##
```

```
library(ALSM)
```

```
## Loading required package: leaps
```

```
## Loading required package: SuppDists
```

```
g<-rep(1,155)
g[x<=1.0646]=0
bftest(lm(y~x,data = score),g)
```

```
##        t.value   P.Value alpha  df
## [1,] 0.4644873 0.6429589  0.05 153
```

```
library(ALSM)
fit
```

```
##
## Call:
## lm(formula = y ~ x, data = score)
##
## Coefficients:
## (Intercept)            x
##       3.203        2.184
```

```
new<-data.frame(x=0.9847)
ci.reg(fit, new, type='m',alpha=0.05)
```

```
##        x     Fit Lower.Band Upper.Band
## 1 0.9847 5.35398   5.248985   5.458975
```