

hw2

Hanbei Xiong

Problem 1

Show that in a simple linear regression model, F statistic for the F test for regression and t statistic for the test of $H_0: \beta_1 = 0$ are equivalent, that is, $F = t^2$. Show this by using the formulas for the test statistics.

Answer:

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

Since (\bar{X}, \bar{Y}) must be on the estimated regression line.

$$SSR = \sum (\hat{\beta}_0 + \hat{\beta}_1 X_i - \hat{\beta}_0 - \hat{\beta}_1 \bar{X})^2$$

$$SSR = \hat{\beta}_1^2 \sum (X_i - \bar{X})^2$$

$$F = \frac{MSR}{MSE} = \frac{\frac{SSR}{1}}{\frac{SSE}{n-2}} = \frac{\hat{\beta}_1^2 \sum (X_i - \bar{X})^2}{MSE}$$

$$SE(\hat{\beta}_1) = \frac{\sqrt{MSE}}{\sqrt{\sum (X_i - \bar{X})^2}}$$

Hence:

$$F = \frac{\hat{\beta}_1^2}{SE(\hat{\beta}_1)^2} = \left(\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \right)^2 = t^2$$

Problem 2

Consider the simple linear regression model with normal errors, $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\varepsilon_i \sim$ independent $N(0, \sigma^2)$, $i = 1, \dots, n$. A **linear transformation of a variable** w involves replacing w with a new variable $w^* = d + cw$, where c and d are constants. Explain how the quantities $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$, R^2 and the test of $H_0: \beta_1 = 0$ are affected by the following linear transformations. Provide mathematical justification for your answers.

- Each value of the predictor x_i is replaced by cx_i , where c is a non-zero constant. For example, the data set has age in months and we convert this to age in years by dividing by 12.
- Each value of x_i is replaced by $x_i + d$. For example, we subtract the mean \bar{x} from each x_i .
- Each value of the response y_i is replaced by ky_i , for a non-zero constant k . For example, suppose that we convert income from units of \$1 to units of \$1000, by dividing by 1000.
- Each value of y_i is replaced by $y_i + d$. For example, we subtract the mean \bar{y} from each y_i .
- Verify your answers to (a)-(d) by fitting models that regress HEIGHT on AGE in the SPIROMETRY data set. Fit models that: convert age in months to age in years; subtract mean age; convert height in cm to height in inches; subtract mean height. Compare to the model using the original scaling of the variables.

Note:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum (Y_i - \bar{Y})^2$$

$$R^2 = \frac{SSR}{SSTO} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 \sqrt{\sum (X_i - \bar{X})^2}}{\hat{\sigma}}$$

(a) Answer:

Given x_i is replaced by cx_i ,

$$\hat{\beta}_0^* = \bar{Y} - \hat{\beta}_1^* (c\bar{X}) = \hat{\beta}_0$$

$$\hat{\beta}_1^* = \frac{\sum (cX_i - c\bar{X})(Y_i - \bar{Y})}{\sum (cX_i - c\bar{X})^2} = \frac{c \sum (X_i - \bar{X})(Y_i - \bar{Y})}{c^2 \sum (X_i - \bar{X})^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{c \sum (X_i - \bar{X})^2} = \frac{1}{c} \hat{\beta}_1$$

$$\hat{\sigma}^{*2} = \frac{1}{n-2} \sum (Y_i - \bar{Y})^2 = \hat{\sigma}^2$$

$$R^{*2} = \frac{\sum (\hat{Y}_i^* - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = \frac{\sum (\hat{\beta}_0^* + \hat{\beta}_1^* cX_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = \frac{\sum (\hat{\beta}_0 + \frac{1}{c} \hat{\beta}_1 cX_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = R^2$$

$$t^* = \frac{\hat{\beta}_1^* \sqrt{\sum (cX_i - c\bar{X})^2}}{\hat{\sigma}^*} = \frac{\frac{1}{c} \hat{\beta}_1 \sqrt{c^2 \sum (X_i - \bar{X})^2}}{\hat{\sigma}} = \frac{\hat{\beta}_1 \sqrt{\sum (X_i - \bar{X})^2}}{\hat{\sigma}} = t$$

(b) Answer:

Given x_i is replaced by $x_i + d$,

$$\begin{aligned}
\hat{\beta}_0^* &= \bar{Y} - \hat{\beta}_1^*(\bar{X} + d) = \bar{Y} - \hat{\beta}_1(\bar{X} + d) = \bar{Y} - \hat{\beta}_1\bar{X} - \hat{\beta}_1d = \hat{\beta}_0 - \hat{\beta}_1d \\
\hat{\beta}_1^* &= \frac{\sum((X_i+d)-(\bar{X}+d))(Y_i-\bar{Y})}{\sum((X_i+d)-(\bar{X}+d))^2} = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2} = \hat{\beta}_1 \\
\hat{\sigma}^{*2} &= \frac{1}{n-2} \sum(Y_i - \bar{Y}_i)^2 = \hat{\sigma}^2 \\
R^{*2} &= \frac{\sum(\hat{Y}_i^* - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{\sum(\hat{\beta}_0^* + \hat{\beta}_1^*(X_i + d) - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{\sum(\hat{\beta}_0^* + \hat{\beta}_1^*X_i + \hat{\beta}_1^*d - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{\sum(\hat{\beta}_0 - \hat{\beta}_1d + \hat{\beta}_1X_i + \hat{\beta}_1d - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = R^2 \\
t^* &= \frac{\hat{\beta}_1^* \sqrt{\sum((X_i+d)-(\bar{X}+d))^2}}{\hat{\sigma}^*} = \frac{\hat{\beta}_1 \sqrt{\sum(X_i-\bar{X})^2}}{\hat{\sigma}} = t
\end{aligned}$$

(c) Answer:

Given y_i is replaced be ky_i ,

$$\begin{aligned}
\hat{\beta}_0^* &= k\bar{Y} - \hat{\beta}_1^*\bar{X} = k\bar{Y} - k\hat{\beta}_1\bar{X} = k(\bar{Y} - \hat{\beta}_1\bar{X}) = k\hat{\beta}_0 \\
\hat{\beta}_1^* &= \frac{\sum(X_i-\bar{X})(kY_i-k\bar{Y})}{\sum(X_i-\bar{X})^2} = \frac{k\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2} = k\hat{\beta}_1 \\
\hat{\sigma}^{*2} &= \frac{1}{n-2} \sum(kY_i - k\bar{Y}_i)^2 = \frac{k^2}{n-2} \sum(Y_i - \bar{Y}_i)^2 = k^2\hat{\sigma}^2 \\
R^{*2} &= \frac{\sum(\hat{Y}_i^* - k\bar{Y})^2}{\sum(kY_i - k\bar{Y})^2} = \frac{\sum(\hat{\beta}_0^* + \hat{\beta}_1^*X_i - k\bar{Y})^2}{\sum(kY_i - k\bar{Y})^2} = \frac{\sum(k\hat{\beta}_0 + k\hat{\beta}_1X_i - k\bar{Y})^2}{\sum(kY_i - k\bar{Y})^2} = \frac{k^2\sum(\hat{\beta}_0 + \hat{\beta}_1X_i - \bar{Y})^2}{k^2\sum(Y_i - \bar{Y})^2} = R^2 \\
t^* &= \frac{\hat{\beta}_1^* \sqrt{\sum(X_i-\bar{X})^2}}{\hat{\sigma}^*} = \frac{k\hat{\beta}_1 \sqrt{\sum(X_i-\bar{X})^2}}{k\hat{\sigma}} = \frac{\hat{\beta}_1 \sqrt{\sum(X_i-\bar{X})^2}}{\hat{\sigma}} = t
\end{aligned}$$

(d) Answer:

Given y_i is replaced be $y_i + d$,

$$\begin{aligned}
\hat{\beta}_0^* &= \bar{Y} + d - \hat{\beta}_1^*\bar{X} = \bar{Y} + d - \hat{\beta}_1\bar{X} = \hat{\beta}_0 + d \\
\hat{\beta}_1^* &= \frac{\sum(X_i-\bar{X})(Y_i+d-\bar{Y}-d)}{\sum(X_i-\bar{X})^2} = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2} = \hat{\beta}_1 \\
\hat{\sigma}^{*2} &= \frac{1}{n-2} \sum(Y_i + d - \bar{Y}_i - d)^2 = \frac{1}{n-2} \sum(Y_i - \bar{Y}_i)^2 = \hat{\sigma}^2 \\
R^{*2} &= \frac{\sum(\hat{Y}_i^* - \bar{Y} - d)^2}{\sum(Y_i + d - \bar{Y}_i - d)^2} = \frac{\sum(\hat{\beta}_0^* + \hat{\beta}_1^*X_i - \bar{Y} - d)^2}{\sum(Y_i - \bar{Y})^2} = \frac{\sum(\hat{\beta}_0 + d + \hat{\beta}_1X_i - \bar{Y} - d)^2}{\sum(Y_i - \bar{Y})^2} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = R^2 \\
t^* &= \frac{\hat{\beta}_1^* \sqrt{\sum(Y_i+d-\bar{Y}_i-d)^2}}{\hat{\sigma}^*} = \frac{\hat{\beta}_1 \sqrt{\sum(Y_i-\bar{Y})^2}}{\hat{\sigma}} = t
\end{aligned}$$

(e) Answer:

I will list the associated quantities below in order $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2, R^2, t)$

Original scaling of variables: 75.92256, 0.53609, 19.372, 0.858, 20.42

Convert age in month to age in year: 75.92256, 6.43307, 19.372, 0.858, 20.42

Subtract mean age: 104.61972, 0.53609, 19.372, 0.858, 20.42

Convert height in cm to height in inches: 29.89077, 0.21106, 3.0027, 0.858, 20.42

Subtract mean height: -28.69716, 0.53609, 19.372, 0.858, 20.42

Each quantities after fitting model with transformation matches with my mathematical derivation in part a), b), c), d).

Problem 3

For the SENIC data, consider the variables risk, nurses, length and svcs. Obtain summary statistics and univariate plots of these variables (e.g., histograms) and describe briefly. Construct scatterplots of the following pairs of variables to examine their relationships (first variable is Y, second is X) and fit **loess curves**:

- Risk and nurses
- Risk and length
- Nurses and svcs

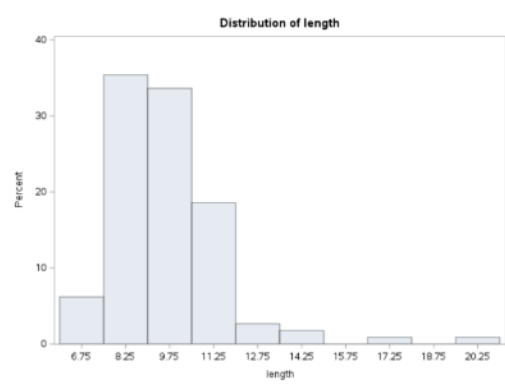
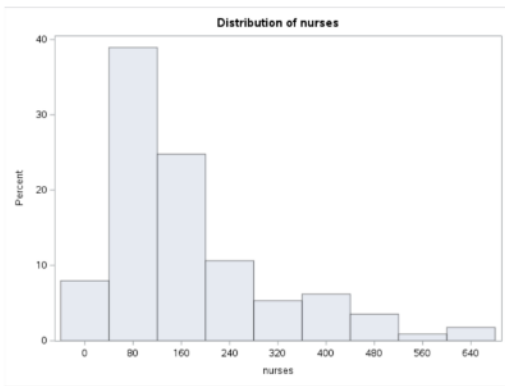
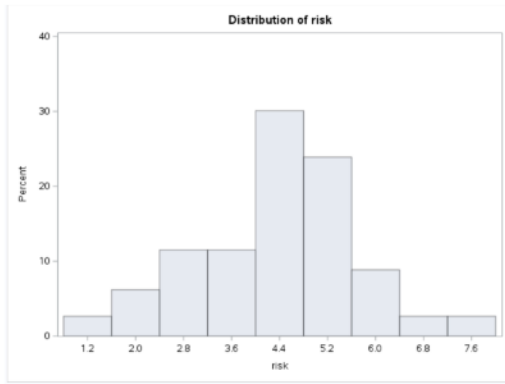
If the relationship in a scatterplot is nonlinear, attempt to make the relationship linear by transforming one or both variables, or explain why a **power/root transformation** to linearity is not an appropriate strategy. For each pair, are you able to find transformations such that the assumptions of the normal error regression model are approximately met? Conduct residuals analysis to determine this. What violations seem to remain and be difficult to address, if any? *Note: Often, the default level of smoothing in a software routine is too low; be sure to try different levels of smoothing.*

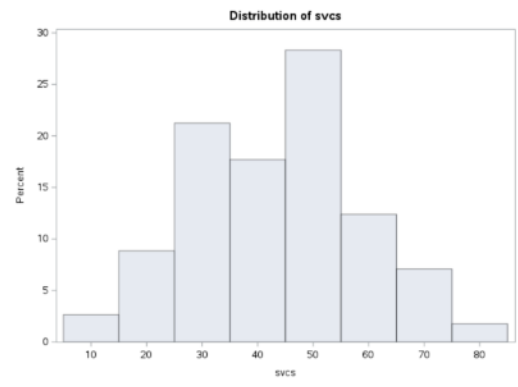
Answer:

Here are summary statistics generated in SAS:

The SUMMARY Procedure							
Variable	Mean	Std Dev	Median	Kurtosis	Skewness	Lower Quartile	Upper Quartile
risk	4.3548673	1.3409080	4.4000001	0.1823554	-0.1197582	3.7000000	5.1999998
nurses	173.2477876	139.2653897	132.0000000	1.5535566	1.3787710	66.0000000	218.0000000
length	9.6483186	1.9114560	9.4200001	8.0774894	2.0689174	8.3400002	10.4700003
svcs	43.1592918	15.2008613	42.9000015	-0.4182831	0.0741808	31.3999996	54.2999992

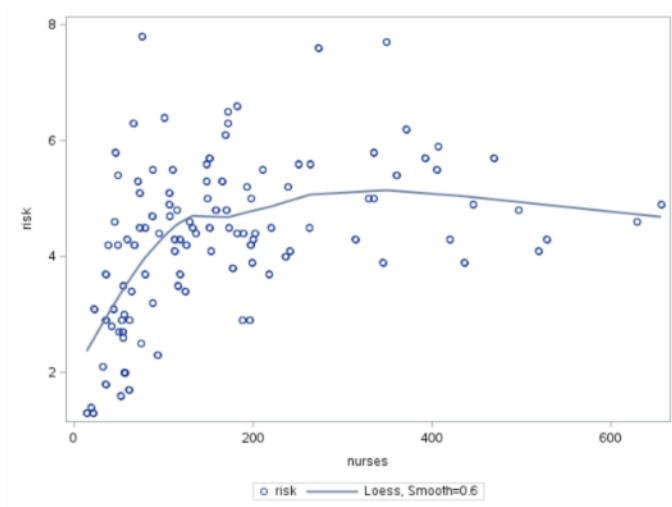
Here are histograms for each variable:



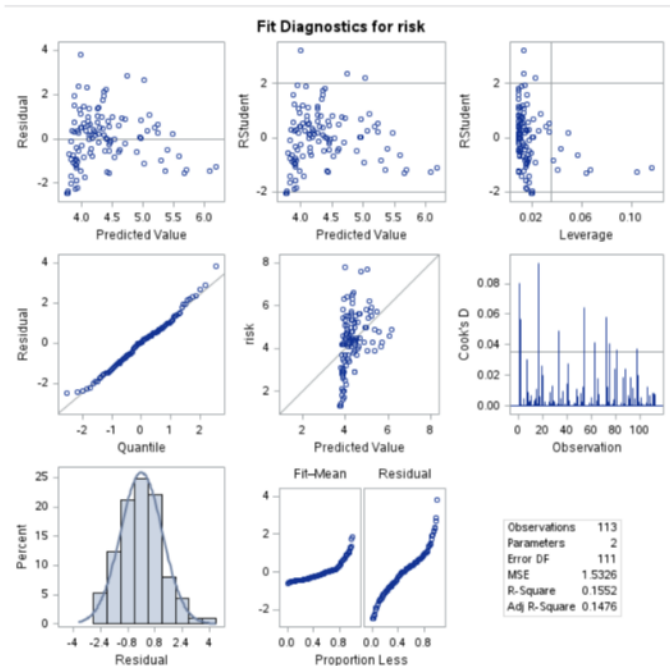


By observing the summary statistics, the mean and median of each variable except “nurses” are close to each other, which indicates that the distribution of each variable is close to normal distribution. For variable “nurses”, the mean is greater than the median which means there are right skewness in the distribution. By observing the histograms, the distribution of each variable except “nurses” is close to normal distribution. The distribution of “nurses” is right skewed. The histograms verify our interpretation of summary statistics.

Here is scatter plot and fitted loess curve for a) between Risk and nurses:

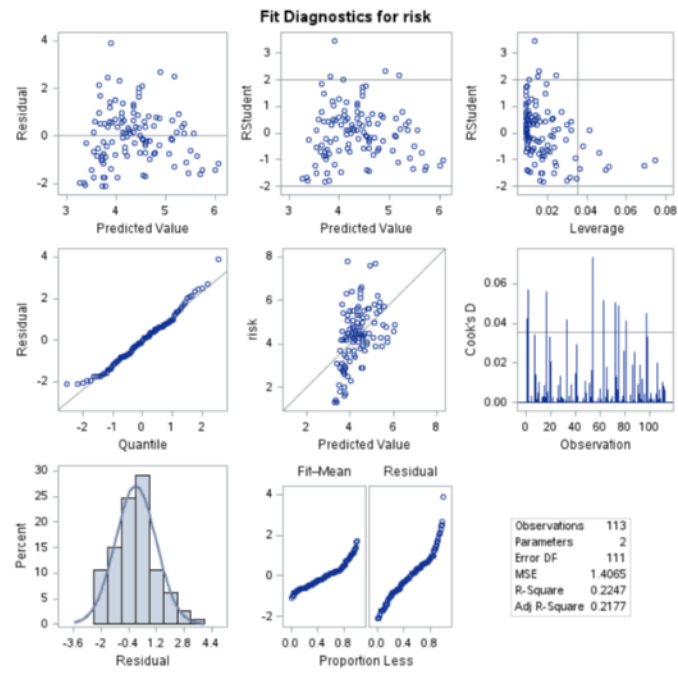


Here is the original residuals analysis to check the assumptions:



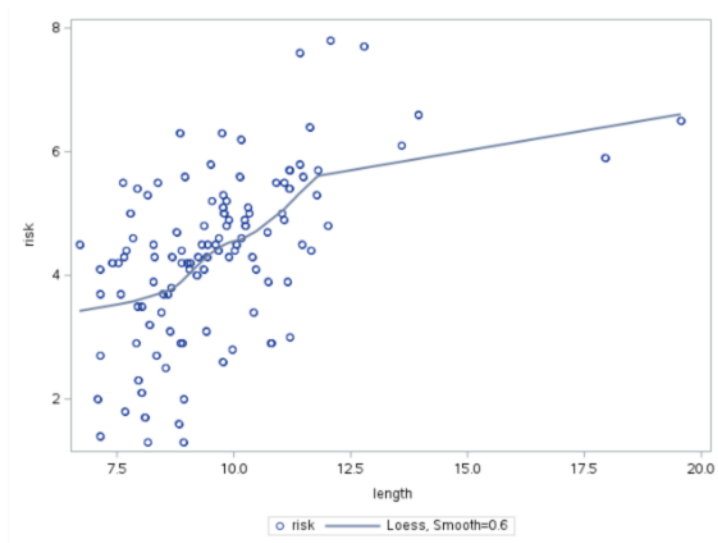
We can see residuals are more scattered in right of predicted value and more narrow to the right. The constancy of error does not seem to meet. The linearity assumption is closely to meet. Residuals seem to distribute normally.

Since the loess curve is relatively monotone and simple from fitted loess curve, we apply squared root transformation on x. Here is the updated residual analysis:

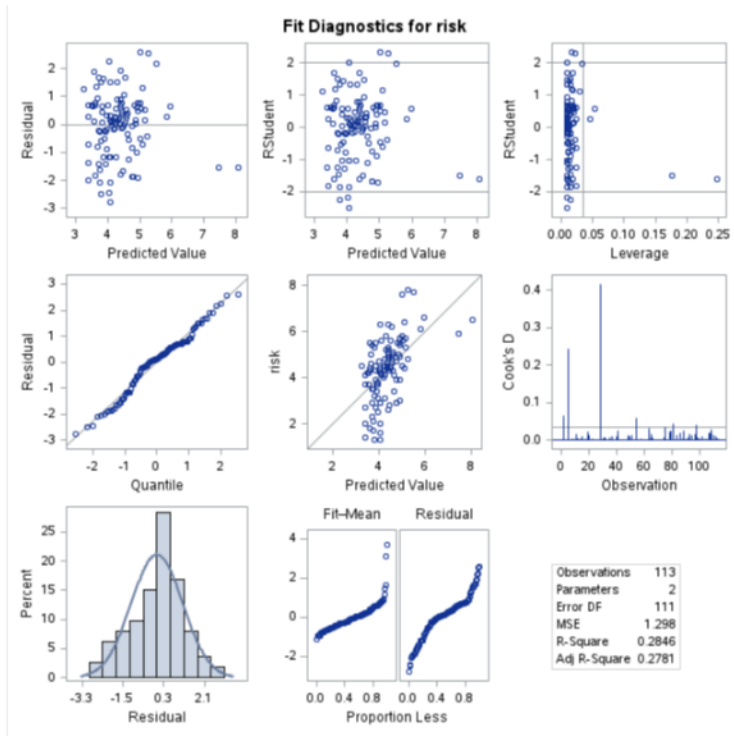


The constancy of error seems to be verified after transformation. The linearity is a little off on both end side but show linearity in overall pattern. The distribution of residuals are a little right skewed but show normal distribution roughly.

Here is scatter plot and fitted loess curve for b) between Risk and length:

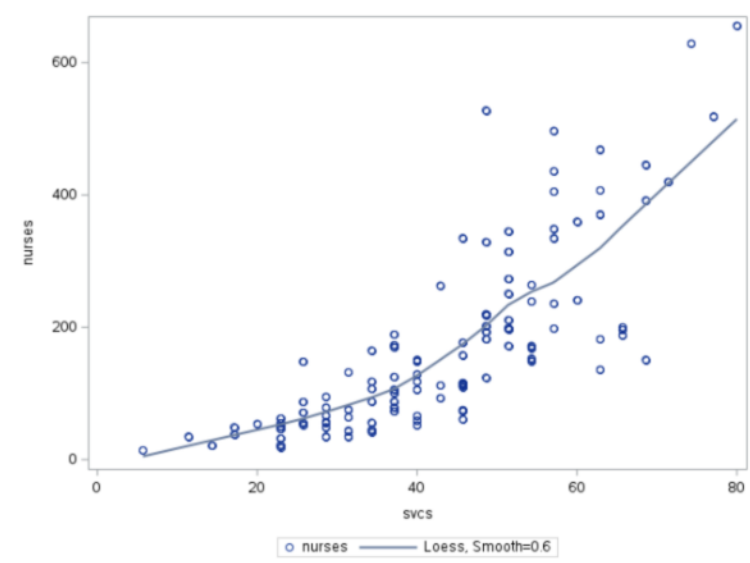


Here is the original residuals analysis to check the assumptions:

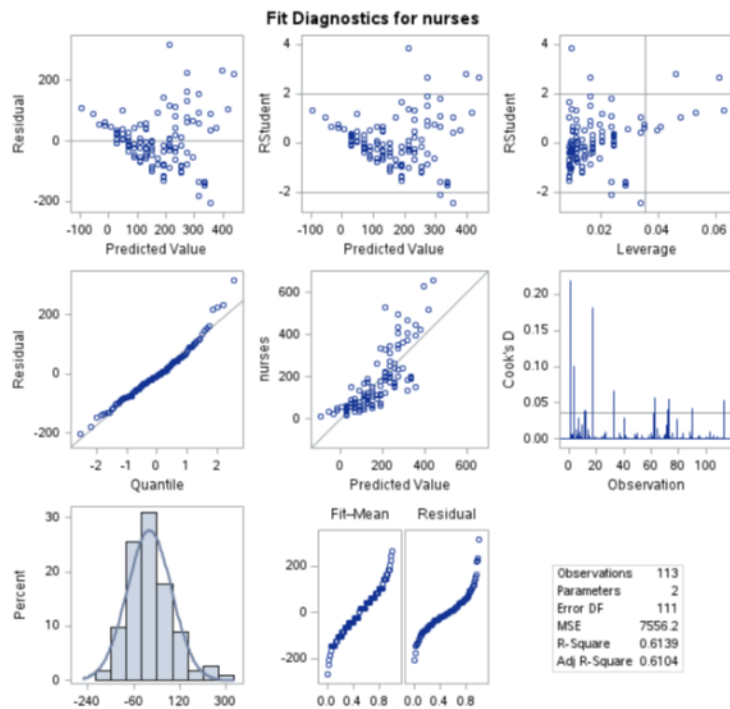


The constancy of error seems to meet as we ignore the outlier. The linearity assumption is closely to meet. Residuals seem to distribute normally. Therefore, we can apply linear regression model on this data.

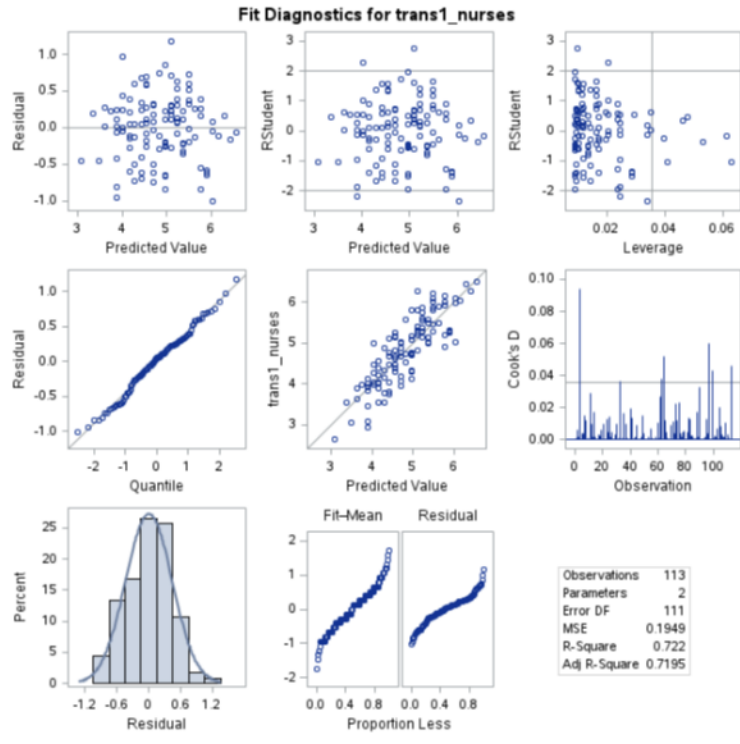
Here is scatter plot and fitted loess curve for c) between nurses and svcs:



Here is the original residuals analysis to check the assumptions:



The assumption of constant error is obviously violated. The linearity assumption is closely to meet. Residuals show normal distribution. We apply log transformation on y. Here is the updated residual analysis:



All assumptions seem to meet after this transformation.

Problem 4

For the simple linear regression model, show that (a) $\frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{\sum(X_i - \bar{X})Y_i}{\sum(X_i - \bar{X})^2}$ and (b) $Cov(\bar{Y}, \hat{\beta}_1) = 0$.

(a) **Answer:**

$$\begin{aligned} \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} &= \frac{\sum(X_i Y_i - X_i \bar{Y} - \bar{X} Y_i + \bar{X} \bar{Y})}{\sum(X_i - \bar{X})^2} \\ \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} &= \frac{\sum(X_i Y_i - X_i \bar{Y} - \bar{X} Y_i) + n \frac{\sum X_i}{n} \bar{Y}}{\sum(X_i - \bar{X})^2} \\ \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} &= \frac{\sum(X_i Y_i - X_i \bar{Y} - \bar{X} Y_i + X_i \bar{Y})}{\sum(X_i - \bar{X})^2} \\ \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} &= \frac{\sum(X_i Y_i - \bar{X} Y_i)}{\sum(X_i - \bar{X})^2} \\ \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} &= \frac{\sum(X_i - \bar{X}) Y_i}{\sum(X_i - \bar{X})^2} \end{aligned}$$

(b) **Answer:**

Let $c_i = \frac{X_i - \bar{X}}{\sum_{j=1}^n (X_j - \bar{X})^2}$

$$Cov(\bar{Y}, \hat{\beta}_1) = Cov(\frac{1}{n} \sum Y_i, \sum c_i Y_i) = \frac{1}{n} Cov(\sum Y_i, \sum c_i Y_i) = \frac{1}{n} Cov(Y_1 + Y_2 + \dots + Y_n, c_1 Y_1 + c_2 Y_2 + \dots + c_n Y_n)$$

Since Y_i are independent, $Cov(Y_i, Y_j) = 0$ for $i \neq j$

$$Cov(\bar{Y}, \hat{\beta}_1) = \frac{1}{n} (Cov(Y_1, c_1 Y_1) + Cov(Y_2, c_2 Y_2) + \dots + Cov(Y_n, c_n Y_n)) = \frac{1}{n} (\sum c_i Var(Y_i)) = \frac{\sigma^2}{n} \sum c_i$$

Since $\sum c_i = \frac{\sum (X_i - \bar{X})}{\sum_{j=1}^n (X_j - \bar{X})^2} = 0$

$$Cov(\bar{Y}, \hat{\beta}_1) = 0$$