hw2

AUTHOR
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.

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

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

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

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

Hence:

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

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 $N(0, \sigma^2),\ i = 1, \ldots, n$. 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.

a. 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.
b. Each value of x_i is replaced by $x_i + d$. For example, we subtract the mean $\bar{x}$ from each x_i.
c. 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.
d. Each value of y_i is replaced by $y_i + d$. For example, we subtract the mean $\bar{y}$ from each y_i.
e. 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:

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

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

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

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

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

(a) Answer:

Given x_i is replaced be cx_i,

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

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

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

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

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

In this linear transformation, the $\hat{\beta}_0$, $\hat{\sigma}^2$, t-statistic and R^2 are invariant. The $\hat{\beta}_1$ is scaled by $\frac{1}{c}$.

(b) Answer:

Given x_i is replaced be $x_i + d$,

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

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

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

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

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

In this linear transformation, the $\hat{\beta}_1$, $\hat{\sigma}^2$, t-statistic and R^2 are invariant. The $\hat{\beta}_0$ is shifted by $-\hat{\beta}_1 d$.

(c) Answer:

Given y_i is replaced be ky_i,

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

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

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

$R^{*2} = \frac{\sum(\hat{Y}_i^* - k\bar{Y})^2}{\sum(kY_i - k\bar{Y})^2} = \frac{\sum(\hat{\beta}_0^* + \hat{\beta}_1^* X_i - k\bar{Y})^2}{\sum(kY_i - k\bar{Y})^2} = \frac{\sum(k\hat{\beta}_0 + k\hat{\beta}_1 X_i - k\bar{Y})^2}{\sum(kY_i - k\bar{Y})^2} = \frac{k^2\sum(\hat{\beta}_0 + \hat{\beta}_1 X_i - \bar{Y})^2}{k^2\sum(Y_i - \bar{Y})^2} = R^2$

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

In this linear transformation, $\hat{\beta}_0$ and $\hat{\beta}_1$ are both scaled by k, $\hat{\sigma}^2$ is scaled by k^2, t-statistic and R^2 are invariant.

(d) Answer:

Given y_i is replaced be $y_i + d$,

$\hat{\beta}_0^* = \bar{Y} + d - \hat{\beta}_1^* \bar{X} = \bar{Y} + d - \hat{\beta}_1 \bar{X} = \hat{\beta}_0 + d$

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

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

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

$t^* = \frac{\hat{\beta}_1^* \sqrt{\sum(Y_i + d - \bar{Y}_i - d)^2}}{\hat{\sigma^*}} = \frac{\hat{\beta}_1 \sqrt{\sum(Y_i - \bar{Y})^2}}{\hat{\sigma}} = t$

In this linear transformation, $\hat{\beta}_0$ is shifted by d, $\hat{\beta}_1$, $\hat{\sigma}^2$, t-statistic and R^2 are invariant.

(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

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

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

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

d) 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**:

a. Risk and nurses
b. Risk and length
c. 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*.

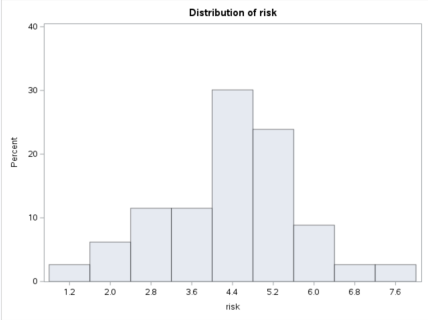
Answer:

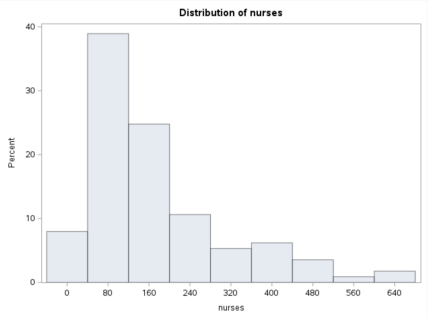
Here are summary statistics generated in SAS:

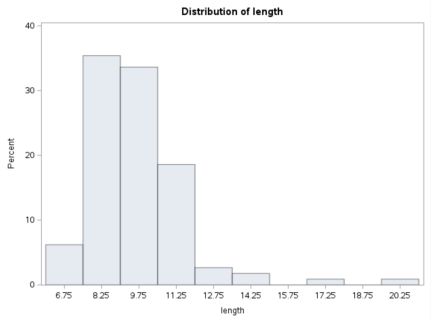
The SUMMARY Procedure

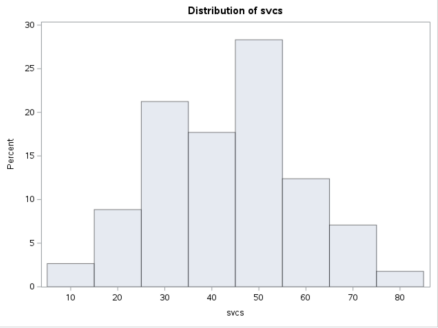
Variable	Mean	Std Dev	Median	Kurtosis	Skewness	Lower Quartile	Upper Quartile
risk	4.3548673	1.3409080	4.4000001	0.1823554	-0.1197582	3.7000000	5.1999998
nurses	173.2477876	139.2653897	132.0000000	1.5535566	1.3787710	66.0000000	218.0000000
length	9.6483186	1.9114560	9.4200001	8.0774894	2.0689174	8.3400002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

Here are histograms for each variable:





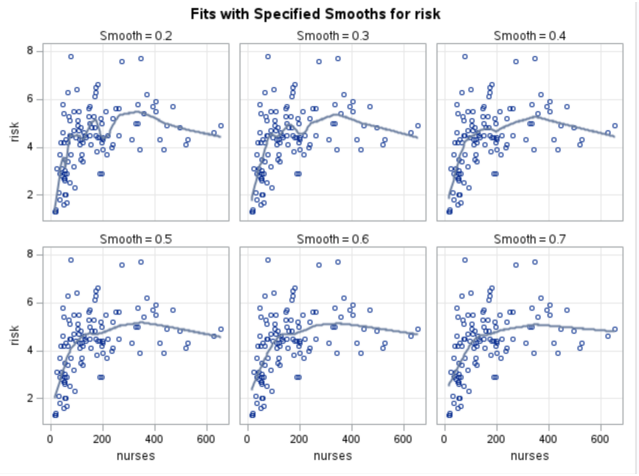




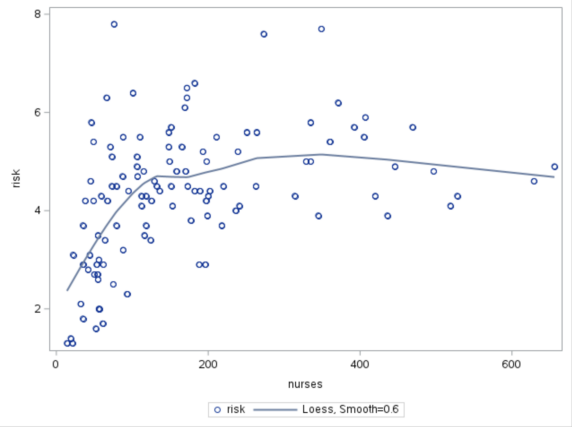
By observing the summary statistics, the mean and median of each variable except “nurses” are close to each other, which indicates that the distribution of each variable is close to normal distribution. For variable “nurses”, the mean is greater than the median which means there are right skewness in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

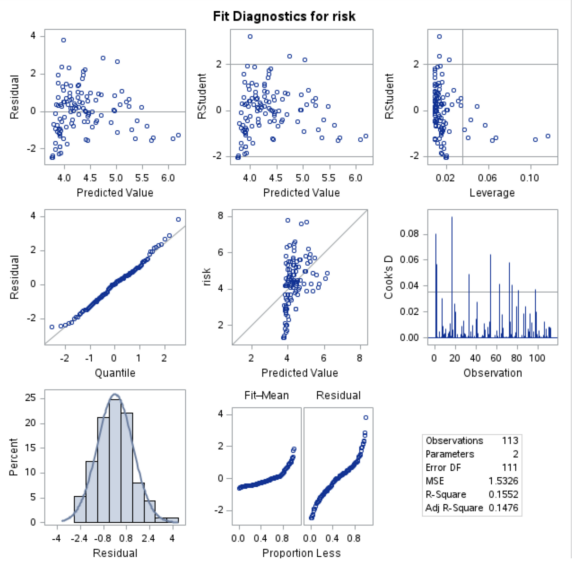
Here is scatter plot and fitted loess curves with different parameters for a) between Risk and nurses:



We set parameter to be 0.6 and here is the plot we used to examine residuals:

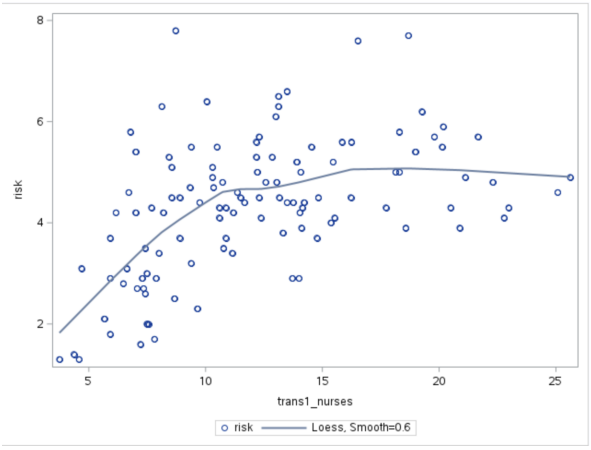


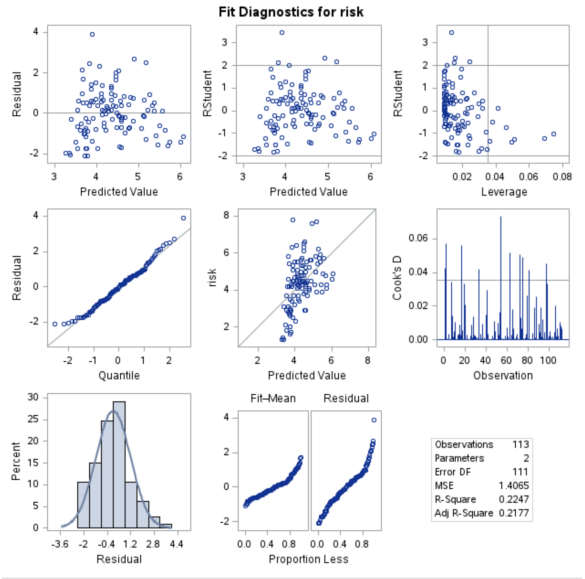
Here is the original residuals analysis to check the assumptions:



We can see residuals are more spread on left and more narrow to the right. Therefore, the constancy of error does not seem to meet. The linearity assumption is closely to meet. Residuals seem to distribute normally.

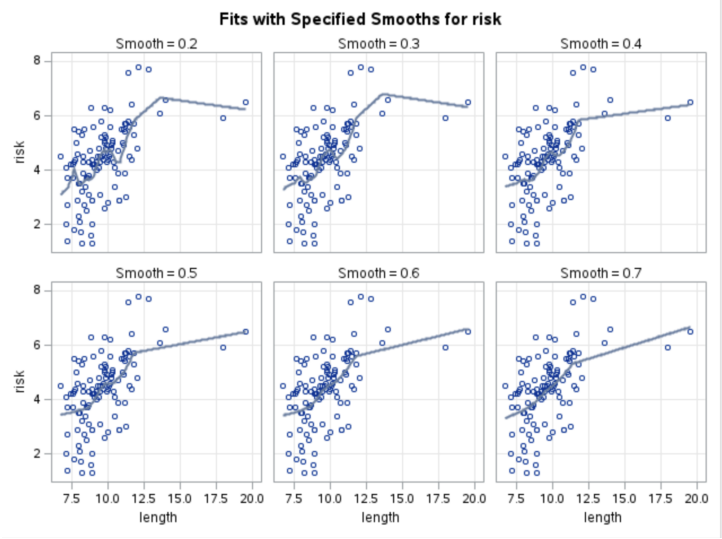
Since the loess curve is not showing a linear relationship between variables, we apply squared root transformation on x. Here is the updated scatter plot with fitted loess curve and residual analysis:



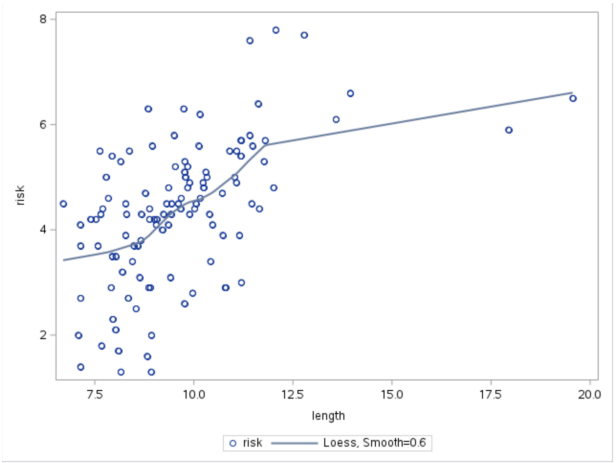


The plot after transformation still shows a non-linear relationship. Since we observe there is decreasing pattern at the end of original plot (before transformation), linear transformation might not be appropriate since data is not following a monotone pattern. However, it could also due to the fact that there are not enough data being collected on the right. Regarding the assumptions, the constancy of error seems to be verified after transformation. The linearity exist in the overall pattern. The distribution of residuals show normal pattern roughly.

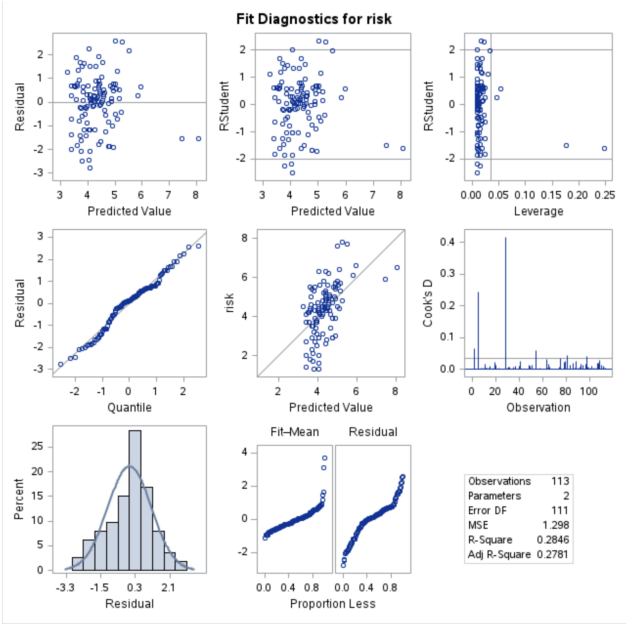
Here is scatter plot and fitted loess curves with different parameters for b) between Risk and length:



We set parameter to be 0.6 and here is the plot we used to examine residuals:

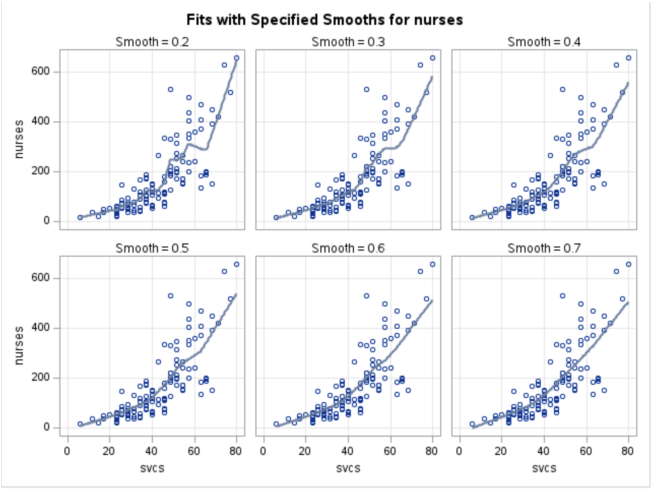


Here is the original residuals analysis to check the assumptions:

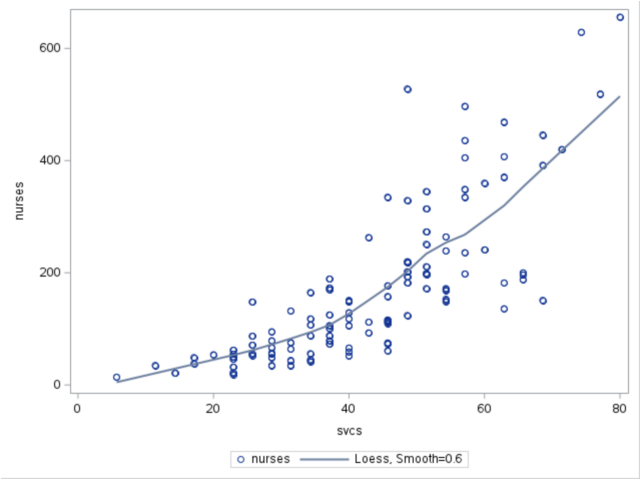


The constancy of error seems to meet as we ignore the outlier. The linearity assumption is closely to meet. Residuals seem to distribute normally. If we remove the outliers as we observe in the scatter plot, it should show a linear regression relationship. Therefore, we do not need to transform variables.

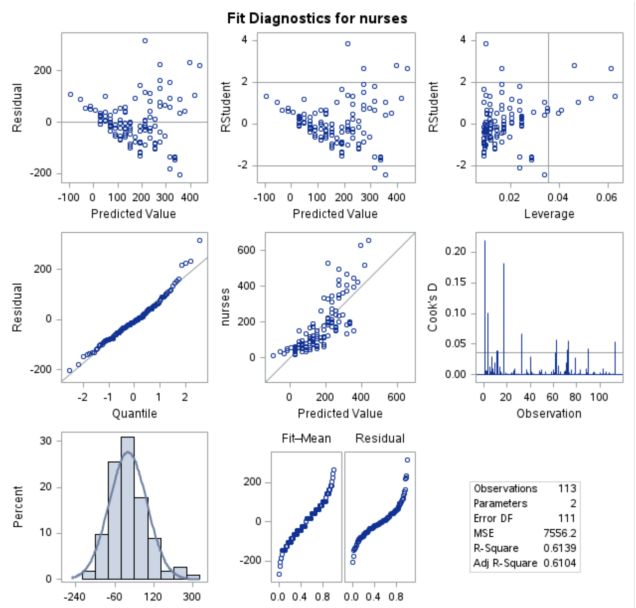
Here is scatter plot and fitted loess curves with different parameters for c) between nurses and svcs:



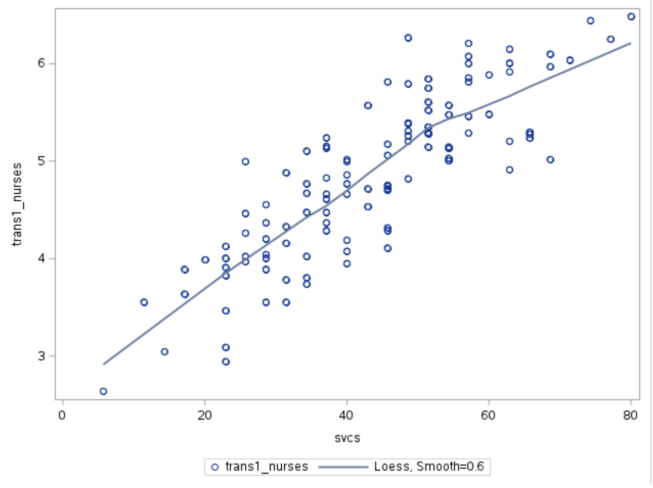
We set parameter to be 0.6 and here is the plot we used to examine residuals:

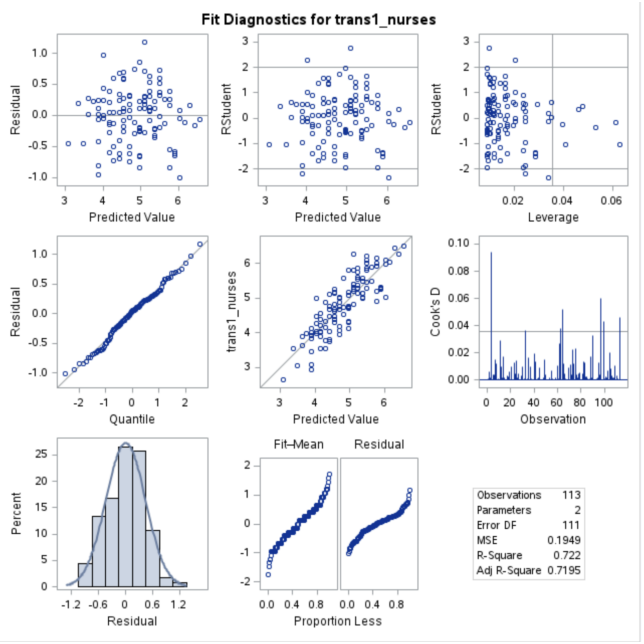


Here is the original residuals analysis to check the assumptions:



The assumption of constant error is obviously violated. The linearity assumption is closely to meet. Residuals show normal distribution. Since the fitted loess curve is not showing an obvious linear relationship between variables. We apply log transformation on y. Here is the updated scatter plot with fitted loess curve and residual analysis:





Variables show a linear relationship after log transformation on nurses. Regarding assumptions, 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:

$$\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2} = \frac{\sum(X_iY_i-X_i\bar{Y}-\bar{X}Y_i+\bar{X}\bar{Y})}{\sum(X_i-\bar{X})^2}$$

$$\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2} = \frac{\sum(X_iY_i-X_i\bar{Y}-\bar{X}Y_i)+n\frac{\sum X_i}{n}\bar{Y}}{\sum(X_i-\bar{X})^2}$$

$$\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2} = \frac{\sum(X_iY_i-X_i\bar{Y}-\bar{X}Y_i+X_i\bar{Y})}{\sum(X_i-\bar{X})^2}$$

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

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

(b) Answer:

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

$Cov(\bar{Y},\hat{\beta}_1) = Cov(\frac{1}{n}\sum Y_i, \sum c_iY_i) = \frac{1}{n}Cov(\sum Y_i, \sum c_iY_i) = \frac{1}{n}Cov(Y_1+Y_2+\cdots+Y_n, c_1Y_1+c_2Y_2+\cdots+c_nY_n)$

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

$Cov(\bar{Y},\hat{\beta}_1) = \frac{1}{n}(Cov(Y_1,c_1Y_1)+Cov(Y_2,c_2Y_2)+\cdots+Cov(Y_n,c_nY_n)) = \frac{1}{n}(\sum c_iVar(Y_i)) = \frac{\sigma^2}{n}\sum c_i$

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

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