Biostat 200B HW2

1. t statistics^2 = F statistics

$$t=\frac{\hat\beta_1}{SE(\hat\beta_1)}=\frac{S_{xy}/S_{xx}}{\sigma/\sqrt{S_{xx}}}=\frac{S_{xy}/S_{xx}}{\sqrt{MSE}/\sqrt{S_{xx}}}\Rightarrow t^2=\frac{S_{xy}^2/S_{xx}^2}{MSE/S_{xx}}=\frac{S_{xy}^2}{MSE\cdot S_{xx}}=\frac{MSReg}{MSE}=F$$

$$\left(SSReg=\Sigma(\hat y_i-\bar y)^2=\Sigma(\hat\beta_0+\hat\beta_1x_i-\hat\beta_0-\hat\beta_1\bar x)^2=\hat\beta_1^2\Sigma(x_i-\bar x)^2=\frac{S_{xy}^2}{S_{xx}}\right)$$

$$MSReg=\frac{SSReg}{df_{reg}}=\frac{SSReg}{1}=\frac{S_{xy}^2}{S_{xx}}$$

$$\left(\hat\beta_1=S_{xy}/S_{xx},\quad \sigma=\sqrt{MSE},\quad SE(\hat\beta_1)=\sigma/\sqrt{S_{xx}}\right)$$

2. For the quantities $\hat\beta_0=\bar y-\hat\beta_1\bar x$; $\hat\beta_1=\frac{\Sigma(x_i-\bar x)(y_i-\bar y)}{\Sigma(x_i-\bar x)(x_i-\bar x)}$; $\hat\sigma^2=\frac{1}{n-2}\Sigma(y_i-\hat y_i)^2$; $R^2=\frac{\Sigma(\hat y_i-\bar y)^2}{\Sigma(y_i-\bar y)^2}$

$$F=\frac{\Sigma(\hat y_i-\bar Y)^2/1}{\Sigma(y_i-\hat y_i)^2/n-2}$$ (F test of $H_0:\beta_1=0$, we only show F statistics because $t^2=F$)

(a) $Y_i=\beta_0+\beta_1cX_i+\varepsilon_i$, $\varepsilon_i\overset{iid}{\sim}N(0,\sigma^2)$

$$\hat\beta_1^a=\frac{\Sigma(cx_i-c\bar x)(y_i-\bar y)}{\Sigma(cx_i-c\bar x)(cx_i-c\bar x)}=\frac{1}{c}\times\hat\beta_1\ ;\ \hat\beta_0^a=\bar y-\frac{1}{c}\hat\beta_1c\bar x=\bar y-\hat\beta_1\bar x=\hat\beta_0\ \text{(remains the same)}$$

$$\hat\sigma^{2a}=\frac{1}{n-2}\Sigma(y_i-\hat y_i)^2=\frac{1}{n-2}\Sigma(y_i-(\hat\beta_0+\tfrac{1}{c}\hat\beta_1cx_i))^2=\frac{1}{n-2}\Sigma(y_i-\hat y_i)^2=\hat\sigma^2\ \text{(remains the same)}$$

Since we know $\hat y_i$ remains the same when the predictor x_i is replaced by cx_i,
we can know that $R^2=\frac{\Sigma(\hat y_i-\bar y)^2}{\Sigma(y_i-\bar y)^2}$ and $F=\frac{\Sigma(\hat y_i-\bar Y)^2}{\Sigma(y_i-\hat y_i)^2/n-2}$ remain the same

(b) $Y_i=\beta_0+\beta_1(x_i+d)+\varepsilon_i$, $\varepsilon_i\overset{iid}{\sim}N(0,\sigma^2)$

$$\hat\beta_1^b=\frac{\Sigma(x_i+d-\bar x-d)(y_i-\bar y)}{\Sigma(x_i+d-\bar x-d)(x_i+d-\bar x-d)}=\frac{\Sigma(x_i-\bar x)(y_i-\bar y)}{\Sigma(x_i-\bar x)\Sigma(x_i-\bar x)}=\hat\beta_1\ \text{(remains the same)}$$

$$\hat\beta_0^b=\bar y-\hat\beta_1(\bar x+d)=\bar y-\hat\beta_1\bar x-\hat\beta_1d=\hat\beta_0-\hat\beta_1d$$

$\hat y$ will be $(\hat\beta_0-\hat\beta_1d)+\hat\beta_1(x_i+d)=\hat\beta_0\cancel{-\hat\beta_1d}+\hat\beta_1x_i\cancel{+\hat\beta_1d}=\hat y$, remains the same

Therefore, we know that $\hat\sigma^2=\frac{1}{n-2}\Sigma(y_i-\hat y_i)^2$, $R^2=\frac{\Sigma(\hat y_i-\bar y)^2}{\Sigma(y_i-\bar y)^2}$, $F=\frac{\Sigma(\hat y_i-\bar Y)^2}{\Sigma(y_i-\hat y_i)^2/n-2}$

will all remain the same because $\hat y$ remains the same and there're no other Xs involved

(c) $Ky_i=\beta_0+\beta_1x_i+\varepsilon_i$, $\varepsilon_i\overset{iid}{\sim}N(0,\sigma^2)$

$$\hat\beta_1^c=\frac{\Sigma(x_i-\bar x)(Ky_i-K\bar y)}{\Sigma(x_i-\bar x)(x_i-\bar x)}=K\hat\beta_1\ ;\ \hat\beta_0^c=K\bar y-K\hat\beta_1\bar x=K(\bar y-\hat\beta_1\bar x)=K\hat\beta_0$$

$$\hat\sigma^{2c}=\frac{1}{n-2}\Sigma(Ky_i-K\hat y_i)^2=K^2\cdot\frac{1}{n-2}\Sigma(y_i-\hat y_i)^2=K^2\hat\sigma^2$$

$$R^{2c} = \frac{\Sigma(k\hat{y}_i - k\bar{y})^2}{\Sigma(ky_i - k\bar{y})^2} = \frac{k^2}{k^2}\frac{\Sigma(\hat{y}_i-\bar{y})^2}{\Sigma(y_i-\bar{y})^2} = R^2 \text{ (remains the same)}$$

$$F = \frac{\Sigma(\hat{y}_i - \bar{Y})^2}{\Sigma(y_i - \hat{y}_i)^2/n-2}$$ will also remain the same because y_i exist both in denominator and numerator

(d) $y_i + d = \beta_0 + \beta_1 x_i + \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$

$$\hat{\beta}_1^d = \frac{\Sigma(x_i-\bar{x})(y_i+d-\bar{y}+d)}{\Sigma(x_i-\bar{x})(x_i-\bar{x})} = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})(x_i-\bar{x})} = \hat{\beta}_1 \text{ (remains the same)}$$

$$\hat{\beta}_0^d = (\bar{y}+d) - \hat{\beta}_1\bar{x} = (\bar{y}-\hat{\beta}_1\bar{x}) + d = \hat{\beta}_0 + d$$

$$\hat{\sigma}^{2d} = \frac{1}{n-2}\Sigma((y_i+\cancel{d}) - (\hat{\beta}_0 + \cancel{d} + \hat{\beta}_1 x_i))^2 = \frac{1}{n-2}\Sigma(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 = \hat{\sigma}^2 \text{ (remains the same)}$$

$$R^{2d} = \frac{\Sigma(\hat{\beta}_0 + d + \hat{\beta}_1 x_i - \bar{y} - d)^2}{\Sigma(y_i + d - \bar{y} - d)^2} = \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2} = R^2 \text{ (remains the same)}$$

We can know $F = \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \hat{y}_i)^2/n-2}$ will also remain the same by other quatitles

(e) Will be shown in the last page of this homework

4. (a) $$\frac{\Sigma(x_i-\bar{x})(Y_i-\bar{Y})}{\Sigma(x_i-\bar{x})^2} = \frac{\Sigma(x_i-\bar{x})\cdot Y_i - \overbrace{\Sigma(x_i-\bar{x})\cdot\bar{Y}}^{\to 0}}{\Sigma(x_i-\bar{x})^2} = \frac{\Sigma(x_i-\bar{x})Y_i}{\Sigma(x_i-\bar{x})^2} \#$$

$$\left(\Sigma(x_i-\bar{x})\bar{Y} = (\Sigma x_i - \Sigma\bar{x})\bar{Y} = (\Sigma x_i - n\cdot\frac{\Sigma x_i}{n})\bar{Y} = 0\cdot\bar{Y} = 0\right)$$

(b) $$\mathrm{Cov}(\bar{Y},\hat{\beta}_1) = \mathrm{Cov}\left(\frac{\Sigma y_i}{n}, \Sigma c_i y_i\right), \text{ where } c_i = \frac{\Sigma(x_i-\bar{x})}{\Sigma(x_i-\bar{x})^2}$$

$$= \frac{1}{n}\Sigma c_i \mathrm{Cov}(\Sigma y_i, \Sigma y_i)$$

$$= \frac{1}{n}\Sigma c_i \mathrm{Var}(y_i) = \frac{\sigma^2}{n}\overbrace{\Sigma c_i}^{\to 0} = 0 \#$$

$$\left(c_i = \frac{\Sigma(x_i-\bar{x})}{\Sigma(x_i-\bar{x})^2} = \frac{\Sigma x_i - n\frac{\Sigma x_i}{n}}{\Sigma(x_i-\bar{x})^2} = 0\right)$$

3.

summary statistics and univariate plots

```
proc univariate data=hw.senic;
var risk nurses length svcs;
histogram;
run;
```

(a) risk

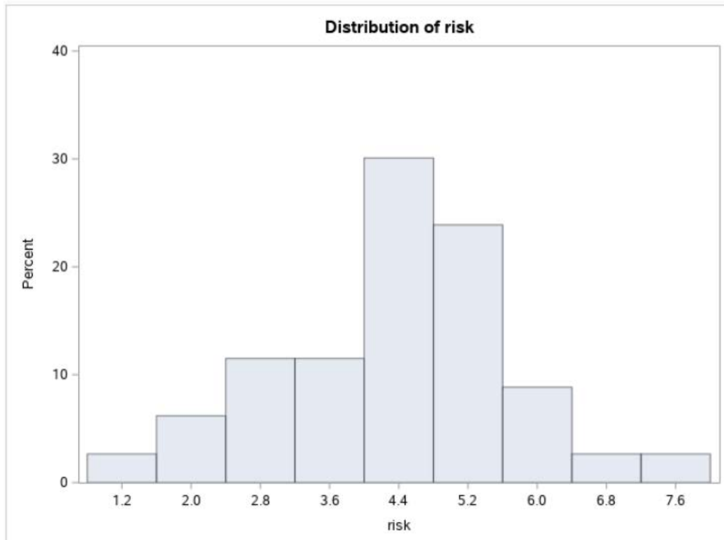
The UNIVARIATE Procedure
Variable: risk

Moments			
N	113	Sum Weights	113
Mean	4.35486727	Sum Observations	492.100001
Std Deviation	1.34090795	Variance	1.79803413
Skewness	-0.1197582	Kurtosis	0.18235536
Uncorrected SS	2344.41001	Corrected SS	201.379823
Coeff Variation	30.7910177	Std Error Mean	0.12614201

Basic Statistical Measures			
Location		Variability	
Mean	4.354867	Std Deviation	1.34091
Median	4.400000	Variance	1.79803
Mode	4.300000	Range	6.50000
		Interquartile Range	1.50000

Note: The mode displayed is the smallest of 2 modes with a count of 7.

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	34.52353	Pr > |t|	<.0001
Sign	M	56.5	Pr >= |M|	<.0001
Signed Rank	S	3220.5	Pr >= |S|	<.0001



The summary statistics and the histogram of variable **risk** are shown above. From the histogram, we can find the distribution is likely normal. We also get the idea from the skewness (-0.11 is very close to 0, normal distribution) from summary statistics table.

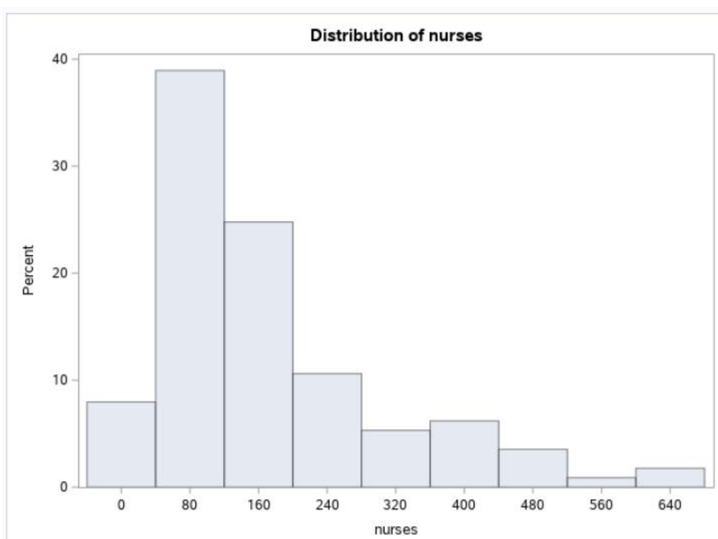
(b) nurses

The UNIVARIATE Procedure
Variable: nurses

Moments			
N	113	Sum Weights	113
Mean	173.247788	Sum Observations	19577
Std Deviation	139.26539	Variance	19394.8488
Skewness	1.37877104	Kurtosis	1.55355665
Uncorrected SS	5563895	Corrected SS	2172223.06
Coeff Variation	80.3850898	Std Error Mean	13.1009858

Basic Statistical Measures			
Location		Variability	
Mean	173.2478	Std Deviation	139.26539
Median	132.0000	Variance	19395
Mode	35.0000	Range	642.00000
		Interquartile Range	152.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	13.22403	Pr > |t|	<.0001
Sign	M	56.5	Pr >= |M|	<.0001
Signed Rank	S	3220.5	Pr >= |S|	<.0001



The summary statistics and the histogram of variable **nurses** are shown above. From the histogram, we can find the distribution is very right skewed. We also get the idea from the skewness (1.38 > 0, right skewed distribution) from summary statistics table.

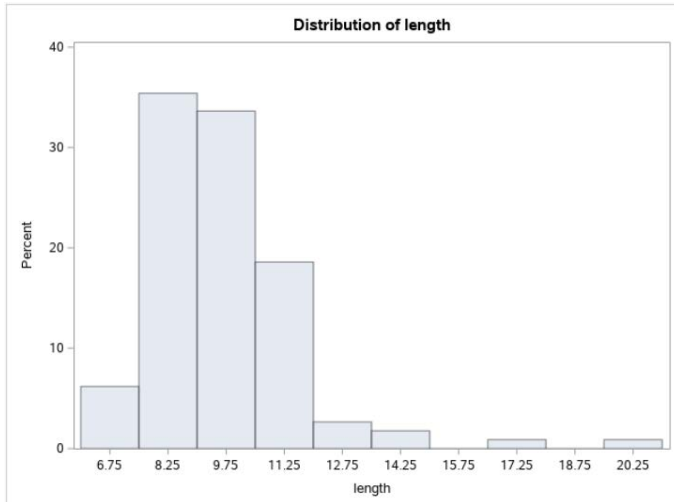
(c) length

The UNIVARIATE Procedure
Variable: length

Moments			
N	113	Sum Weights	113
Mean	9.64831856	Sum Observations	1090.26
Std Deviation	1.91145602	Variance	3.6536641
Skewness	2.06891738	Kurtosis	8.07748944
Uncorrected SS	10928.3861	Corrected SS	409.210379
Coeff Variation	19.8112863	Std Error Mean	0.17981466

Basic Statistical Measures			
Location		Variability	
Mean	9.648319	Std Deviation	1.91146
Median	9.420000	Variance	3.65366
Mode	7.140000	Range	12.86000
		Interquartile Range	2.13000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	53.65702	Pr > |t|	<.0001
Sign	M	56.5	Pr >= |M|	<.0001
Signed Rank	S	3220.5	Pr >= |S|	<.0001



The summary statistics and the histogram of variable **length** are shown above. From the histogram, we can find the distribution is very right skewed. We also get the idea from the skewness (2.07 > 0, right skewed distribution) from summary statistics table.

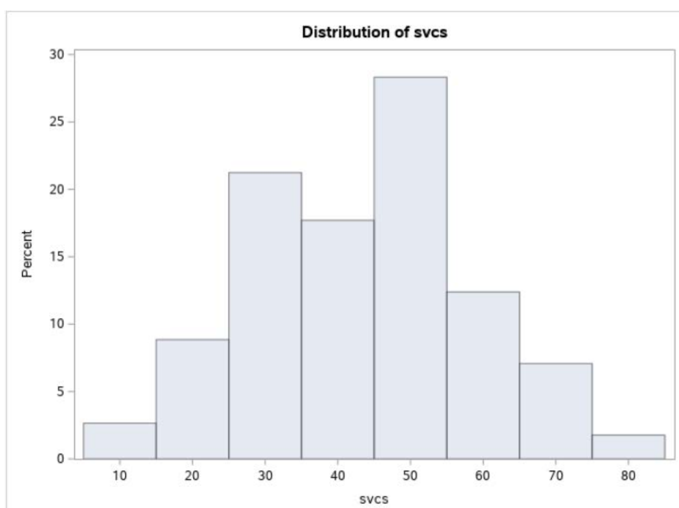
(d) svcs

The UNIVARIATE Procedure
Variable: svcs

Moments			
N	113	Sum Weights	113
Mean	43.1592918	Sum Observations	4876.99998
Std Deviation	15.2008613	Variance	231.066183
Skewness	0.07418083	Kurtosis	-0.4182831
Uncorrected SS	236367.278	Corrected SS	25879.4125
Coeff Variation	35.2203676	Std Error Mean	1.42997674

Basic Statistical Measures			
Location		Variability	
Mean	43.15929	Std Deviation	15.20086
Median	42.90000	Variance	231.06618
Mode	45.70000	Range	74.30000
		Interquartile Range	22.90000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	30.18181	Pr > |t|	<.0001
Sign	M	56.5	Pr >= |M|	<.0001
Signed Rank	S	3220.5	Pr >= |S|	<.0001



The summary statistics and the histogram of variable **svcs** are shown above. From the histogram, we can find the distribution is likely normal. We also get the idea from the skewness (0.07 is very close to 0, normal distribution) from summary statistics table.

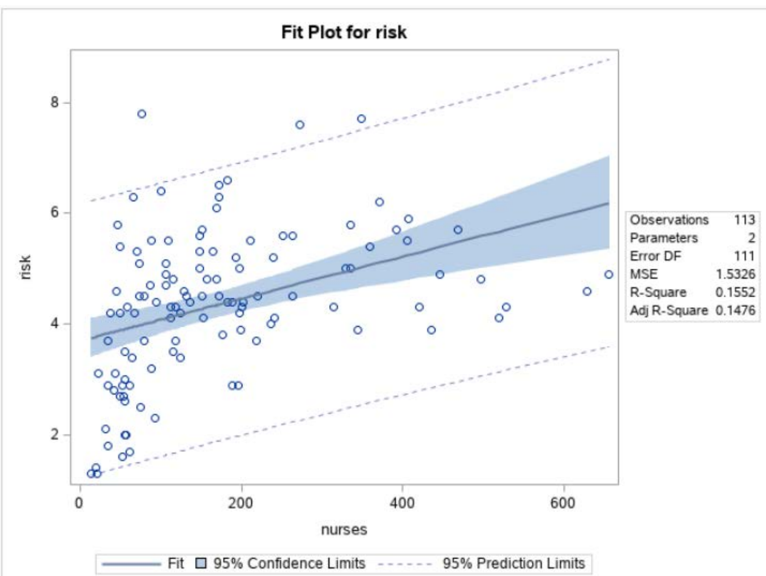
(a) Risk and nurses

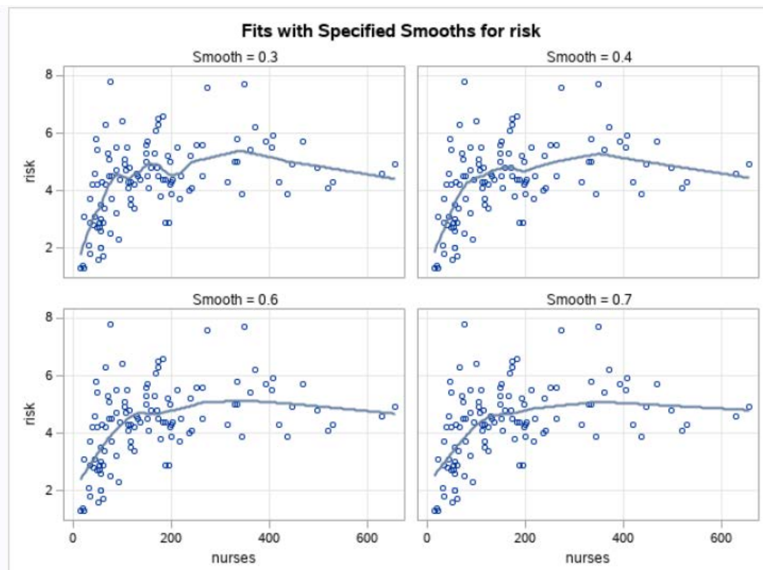
```
proc reg data=hw.senic;
model risk = nurses;
run;
proc loess data=hw.senic;
model risk = nurses / smooth=0.3 0.4 0.6 0.7;
run;
```

Analysis of Variance

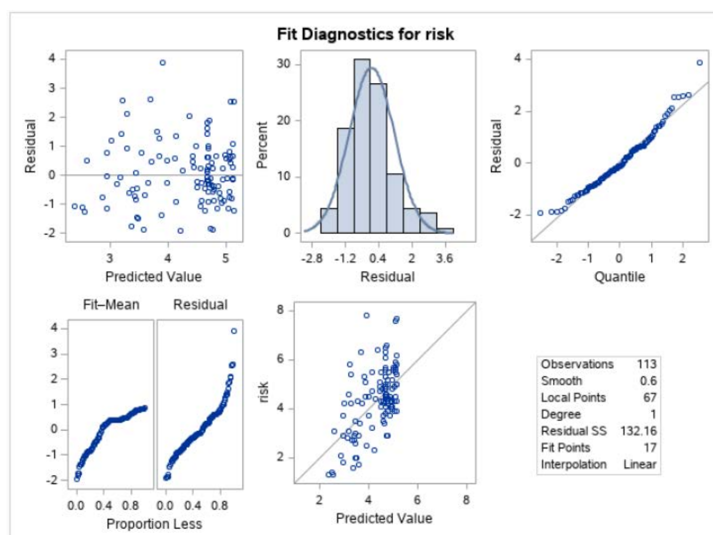
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	31.25844	31.25844	20.40	<.0001
Error	111	170.12139	1.53263		
Corrected Total	112	201.37982			

Root MSE	1.23799	R-Square	0.1552
Dependent Mean	4.35487	Adj R-Sq	0.1476
Coeff Var	28.42779		





The scatterplot with regression line and the table of Analysis of Variance of **risk and nurses** are shown above. We fit different loess curves with different levels of smoothing, and find that the relationship between **risk and nurses** is nonlinear. We choose the loess curves with levels of smooth = 0.6; from this curve we can see it is not monotone, which means the association between **risk and nurses** is not either positive or negative everywhere. Therefore, a power/root transformation to linearity is not an appropriate strategy to apply here.



We can check the assumptions of the normal error regression model via the fit diagnostics plot shown above. From the residuals analysis, we can find the regression function is nonlinear, and the variance are not constant; it seems to meet the assumption of normality of residual according to Q-Q plot, but we may need further inspection such as Shapiro – Wilk test. We need to do a different transformation other than power/root transformation to meet the assumptions of linearity and constant variance.

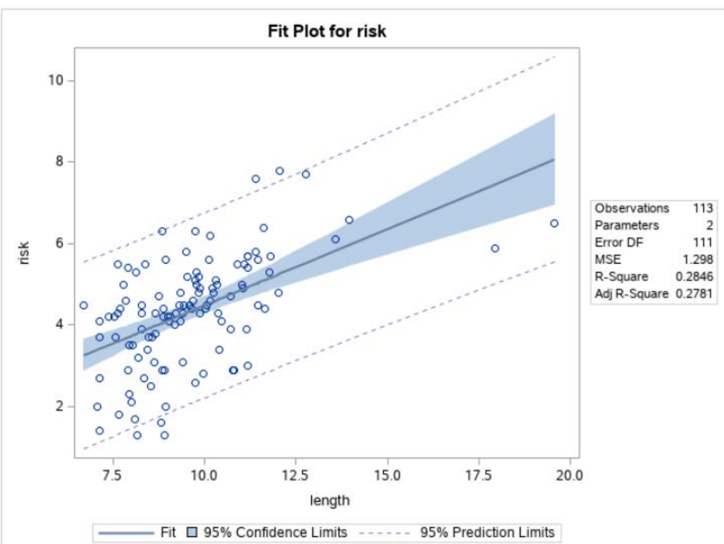
(b) Risk and length

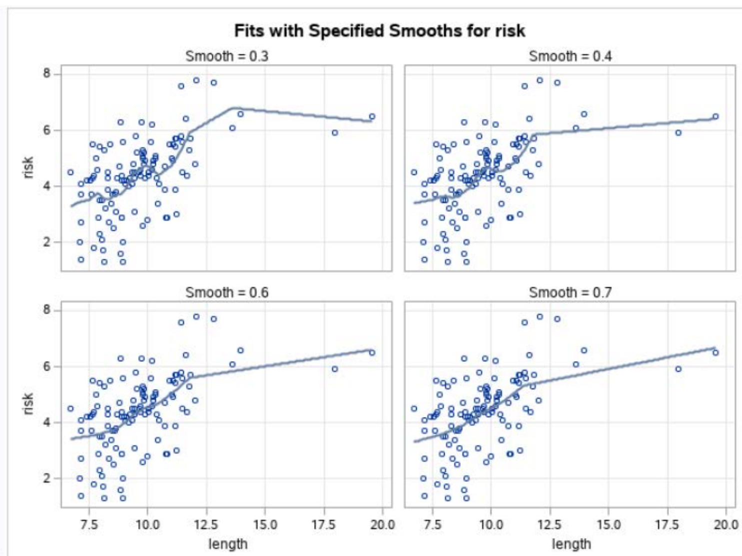
```
proc reg data=hw.senic;
model risk = length;
run;
proc loess data=hw.senic;
model risk = length/ smooth=0.3 0.4 0.6 0.7;
run;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	57.30511	57.30511	44.15	<.0001
Error	111	144.07472	1.29797		
Corrected Total	112	201.37982			

Root MSE	1.13929	R-Square	0.2846
Dependent Mean	4.35487	Adj R-Sq	0.2781
Coeff Var	26.16119		

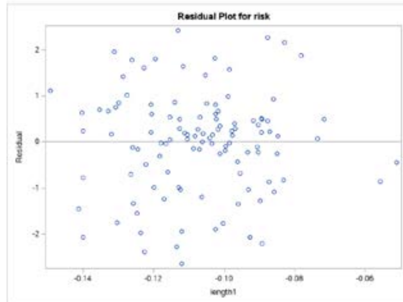


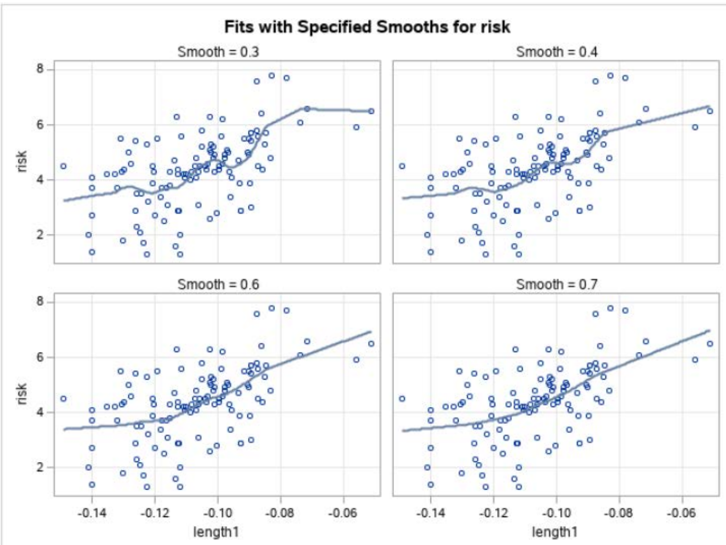


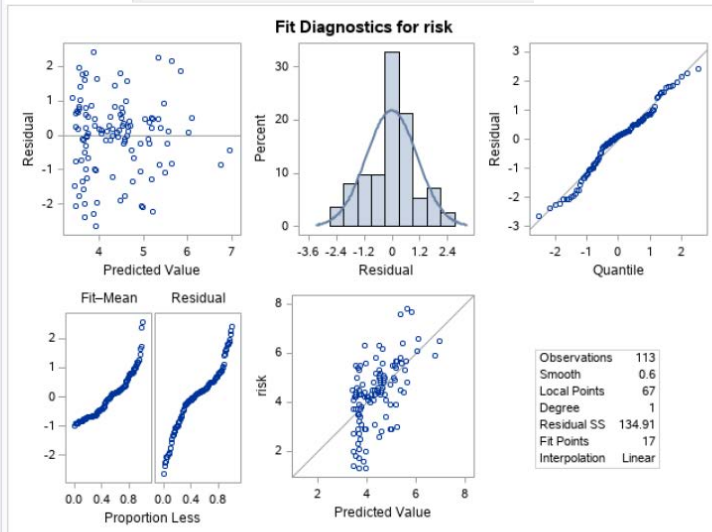
The scatterplot with regression line and the table of Analysis of Variance of **risk and length** are shown above. We fit different loess curves with different levels of smoothing, and find that the relationship between **risk and length** is nonlinear. We choose the loess curves with levels of smooth = 0.6; from this curve we can see it concave slightly downward. Therefore, we can try root transformation to linearity by doing $log(x)$, $\sqrt{x}$ or $-1/x$. After trying all those transformation, we find $-1/x$ be the best transformation to meet the assumptions of the normal error regression model.

```
data hw.senic2;
set hw.senic;
```

```
length1 = -1/length;
run;
proc loess data=hw.senic2;
model risk = length1/ smooth=0.3 0.4 0.6 0.7;
run;
```







We can check the assumptions of the normal error regression model via those residuals analysis shown above. After transformation, the loess curves are more linear; by the residual plot, we can also find that the error variances are constant enough. It also seems to meet the assumption of normality of residual according to Q-Q plot, but we may need further inspection such as Shapiro – Wilk test.

(c) Risk and length

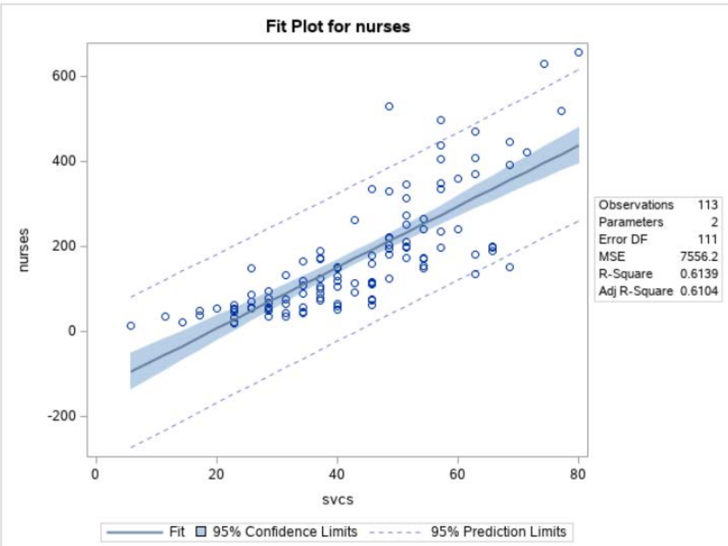
```
proc reg data=hw.senic;
model nurses = svcs;
run;
proc loess data=hw.senic;
model nurses = svcs / smooth=0.3 0.4 0.6 0.7;
run;
```

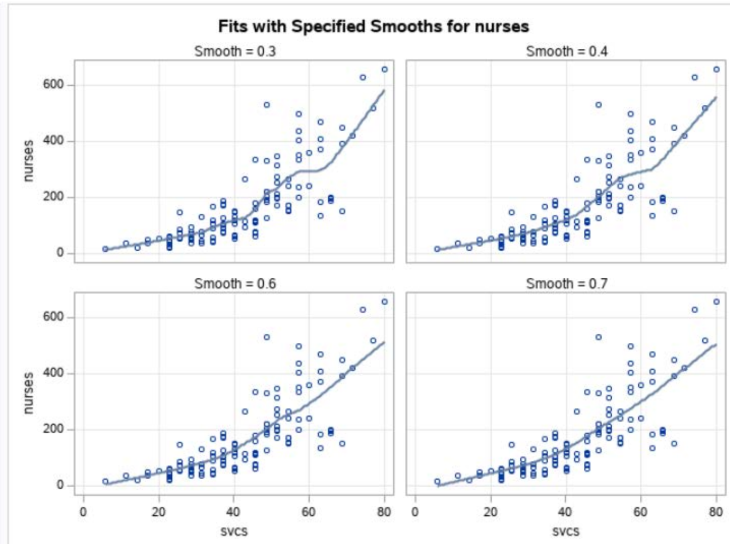
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1333486	1333486	176.48	<.0001
Error	111	838737	7556.18811		
Corrected Total	112	2172223			

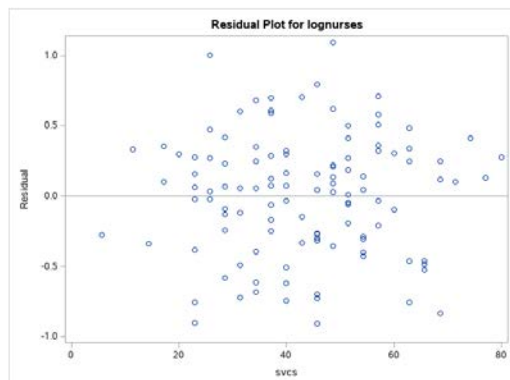
Root MSE	86.92634	R-Square	0.6139
Dependent Mean	173.24779	Adj R-Sq	0.6104
Coeff Var	50.17457		

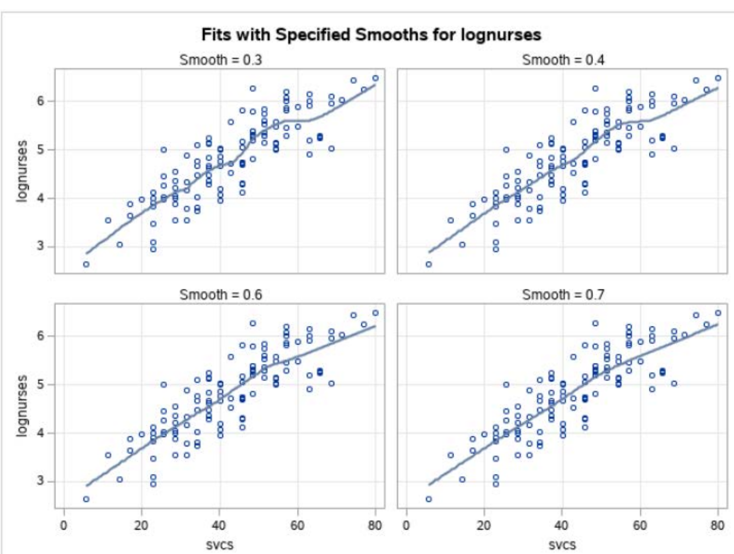
The scatterplot with regression line and the table of Analysis of Variance of **risk and length** are shown below. We fit different loess curves with different levels of smoothing, and find that the relationship between **risk and length** is nonlinear. We choose the loess curves with levels of smooth = 0.6; from this curve we can see it concave slightly upward. Therefore, we can try power transformation to linearity by doing x^2, x^3 or e^x. After trying all those transformation, we find transforming x hard to meet the assumptions of the normal error regression model. Therefore, we try to do root transformation on y. We find $log(y)$ best to meet those assumptions.

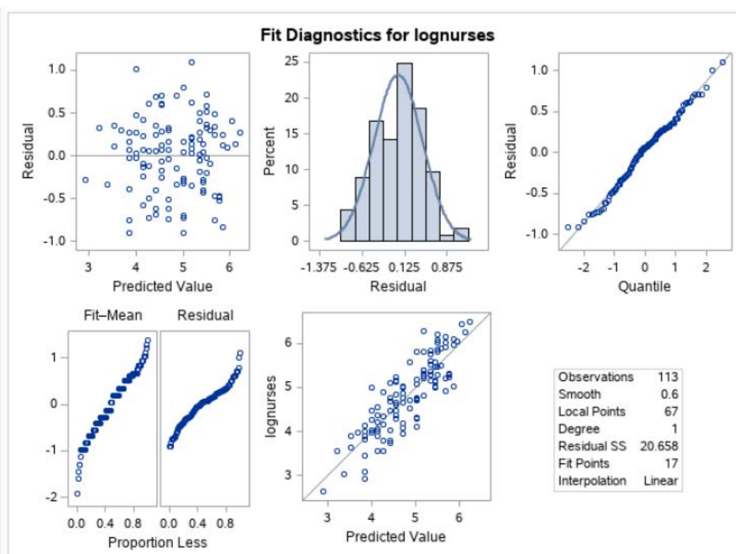





```
data hw.senic3;
set hw.senic;
lognurses = log(nurses);
run;
proc loess data=hw.senic3;
model lognurses = svcs / smooth=0.3 0.4 0.6 0.7;
run;
```







We can check the assumptions of the normal error regression model via those residuals analysis shown above. After transformation, the loess curves are more linear; by the residual plot, we can also find that the error variances are constant. It also meets the assumption of normality of residual according to Q-Q plot

2.(e)

```
proc means data=hw.spirometry;
var age height;
run;

data hw.spirometry1;
set hw.spirometry;
age_year = age/12;
age_mean = age-53.5305493;
height_inches = height*0.393701;
height_mean = height-104.6197183;
run;

proc reg data=hw.spirometry1;
model height = age;
run;
```

Original scaling of model that regress **height** on **age** in the SPIROMETRY data set

$$Y_i = \beta_0 + \beta_1 * X_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$$

We get $\widehat{\beta_0} = 75.92,\ \widehat{\beta_1} = 0.54,\ \widehat{\sigma^2} = 4.40^2,\ R^2 = 0.86,\ F\ \ Value = 416.89$

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	8076.05675	8076.05675	416.89	<.0001
Error	69	1336.67564	19.37211		
Corrected Total	70	9412.73239			

Root MSE	4.40138	R-Square	0.8580
Dependent Mean	104.61972	Adj R-Sq	0.8559
Coeff Var	4.20702		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	75.92256	1.49942	50.63	<.0001
AGE	Age (Months)	1	0.53609	0.02626	20.42	<.0001

(i)

```
proc reg data=hw.spirometry1;
model height = age_year;
run;
```

Model that regress **height** on **age**: convert to age in years by dividing by 12 ($c = \frac{1}{12}$)

$$Y_i = \beta_0 + \beta_1 * X_i/12 + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$$

We get $\widehat{\beta_0} = 75.92,\ \widehat{\beta_1} = 6.43,\ \widehat{\sigma^2} = 4.40^2,\ R^2 = 0.86,\ F\ \ Value = 416.89$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	8076.05675	8076.05675	416.89	<.0001
Error	69	1336.67564	19.37211		
Corrected Total	70	9412.73239			

Root MSE	4.40138	R-Square	0.8580
Dependent Mean	104.61972	Adj R-Sq	0.8559
Coeff Var	4.20702		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	75.92256	1.49942	50.63	<.0001
age_year		1	6.43307	0.31507	20.42	<.0001

$$\widehat{\beta_1}^* = \frac{1}{c} * \widehat{\beta_1} = 12 * 0.54 \approx 6.43, which\ \ verifies\ \ 2(a).$$

(ii)

```
proc reg data=hw.spirometry1;
model height = age_mean;
run;
```

Model that regress **height** on **age**: subtract the mean $\bar{x}$ from each x_i (d $= -53.53$)

$$Y_i = \beta_0 + \beta_1 * (X_i - 53.53) + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$$

We get $\widehat{\beta_0} = 104.62,\ \widehat{\beta_1} = 0.54,\ \widehat{\sigma^2} = 4.40^2,$ $R^2 = 0.86,\ F\ \ Value = 416.89$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	8076.05675	8076.05675	416.89	<.0001
Error	69	1336.67564	19.37211		
Corrected Total	70	9412.73239			

Root MSE	4.40138	R-Square	0.8580
Dependent Mean	104.61972	Adj R-Sq	0.8559
Coeff Var	4.20702		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	104.61972	0.52235	200.29	<.0001
age_mean		1	0.53609	0.02626	20.42	<.0001

$$\widehat{\beta_0}^* = \widehat{\beta_0} - \widehat{\beta_1} * d = 75.92 - 0.54 * (-53.53) \approx 104.62, which\ \ verifies\ \ 2(b).$$

(iii)

```
proc reg data=hw.spirometry1;
model height_inches = age;
run;
```

Model that regress **height** on **age**: convert to height in inches by multiplying by 0.394 ($k = 0.394$)

$$0.394 * Y_i = \beta_0 + \beta_1 * X_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$$

We get $\widehat{\beta_0} = 29.89,\ \widehat{\beta_1} = 0.21,\ \widehat{\sigma^2} = 1.73^2,\ R^2 = 0.86,\ F\ \ Value = 416.89$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1251.79265	1251.79265	416.89	<.0001
Error	69	207.18536	3.00269		
Corrected Total	70	1458.97801			

Root MSE	1.73283	R-Square	0.8580
Dependent Mean	41.18889	Adj R-Sq	0.8559
Coeff Var	4.20702		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	29.89079	0.59032	50.63	<.0001
AGE	Age (Months)	1	0.21106	0.01034	20.42	<.0001

$$\widehat{\beta_0}^* = k^2 * \widehat{\beta_0} = 0.394 * 75.92 \approx 29.89$$

$$\widehat{\beta_1}^* = k * \widehat{\beta_1} = 0.394 * 0.54 \approx 0.21$$

$$\widehat{\sigma^2}^* = 0.394^2 * 4.40^2 \approx 1.73^2, which\ \ verifies\ \ 2(c)$$

(iv)

```
proc reg data=hw.spirometry1;
model height_mean = age;
run;
```

Model that regress **height** on **age**: subtract the mean $\bar{y}$ from each y_i (d $= -104.62$)

$$Y_i - 104.62 = \beta_0 + \beta_1 * X_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$$

We get $\widehat{\beta_0} = -28.70,\ \widehat{\beta_1} = 0.54,\ \widehat{\sigma^2} = 4.40^2,\ R^2 = 0.86,\ F\ \ Value = 416.89$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	8076.05675	8076.05675	416.89	<.0001
Error	69	1336.67564	19.37211		
Corrected Total	70	9412.73239			

Root MSE	4.40138	R-Square	0.8580
Dependent Mean	9.859153E-9	Adj R-Sq	0.8559
Coeff Var	44642538851		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-28.69716	1.49942	-19.14	<.0001
AGE	Age (Months)	1	0.53609	0.02626	20.42	<.0001

$$\widehat{\beta_0}^* = \widehat{\beta_0} + d = 75.92 - 104.62 \approx -28.70, which\ \ verifies\ \ 2(d).$$