Tutorial #4

Exercise 1:

The table below lists the USA social security costs for 7 specific years between 1965 and 1992

Year	**1965**	**1970**	**1975**	**1980**	**1985**	**1990**	**1992**
X=year-1960	5	10	15	20	25	30	32
Y= social security costs($)	17.1	29.6	63.6	117.1	186.4	246.5	285.1

a. Plot the data using y against x (hand-drawn graph is acceptable).

b. Compute $\sum_{i=1}^{7} X_i$, $\sum_{i=1}^{7} Y_i$, $\sum_{i=1}^{7} X_i^2$, $\sum_{i=1}^{7} X_i Y_i$, $\sum_{i=1}^{7} Y_i^2$ .use these figure to fit the data with the simple linear regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$.

c. Test the hypothesis $H_0: \beta_1 = 0 \; vs. \beta_1 > 0$. at the 5% significance level, what can be concluded about social security costs from this test?

d. Plot the residual against x (hand-drawn graph is acceptable). Is the model fitted well? If not, discuss what might try to achieve a better fit.

Solution:

a. Minitab Result: **Regression Analysis: y versus x**

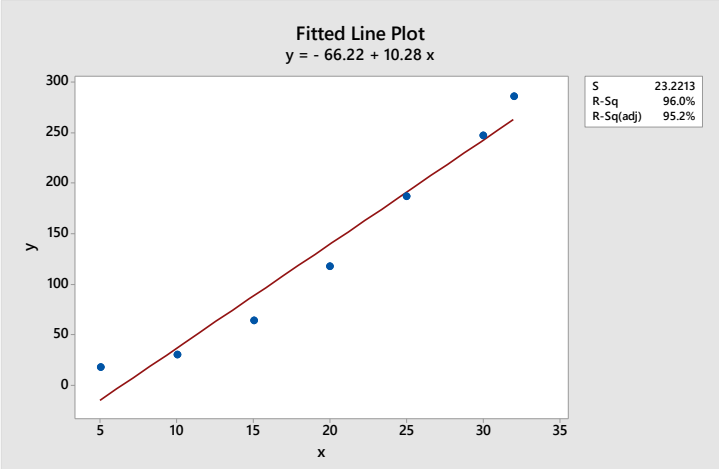
The regression equation is
y = - 66.22 + 10.28 x

Model Summary

S	R-sq	R-sq(adj)
23.2213	96.04%	95.24%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	65336.0	65336.0	121.17	0.000
Error	5	2696.1	539.2		
Total	6	68032.1			



b.

i	X_i	Y_i	x_iy_i	y_i^2	x_i^2	$\hat{y}$	$(y-\hat{y})^2$	$(X_i-\bar{X})^2$
1	5	17.1	85.5	292.41	25	-14.82	1018.89	212.33
2	10	29.6	296	876.16	100	36.58	48.72	91.61
3	15	63.6	954	4044.96	225	87.98	594.38	20.90
4	20	117.1	2342	13712.41	400	139.38	496.40	0.18
5	25	186.4	4660	34744.96	625	190.78	19.18	29.47
6	30	246.5	7395	60762.25	900	242.18	18.66	108.76
7	32	285.1	9123.2	81282.01	1024	262.74	499.97	154.47
sum	137	945.4	24855.7	195715.16	3299	944.82	2696.21	617.7143

$$b_1 = \widehat{\beta_1} = \frac{\sum_{i=1}^n x_iy_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i /n}{\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2/n} = \frac{\sum_{i=1}^n x_iy_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

$$b_1 = \frac{24855.7 - 7\left(\frac{137}{7}\right)\left(\frac{945.4}{7}\right)}{(3299 - (137^2/7))} = 10.284$$

$$b_0 = \bar{y} - b_1\bar{x} = \frac{\sum_{i=1}^n y_i}{n} - b_1\left(\frac{\sum_{i=1}^n x_i}{n}\right) = \frac{945.4}{7} - 10.284 * \left(\frac{137}{7}\right) = -66.215$$

$$\boldsymbol{\hat{Y} = -66.2 + 10.284\,X}$$

***this indicates that social security cost increases by about $10.284 every year.**

c. Test , using $\alpha = 0.05$

We first have to estimate σ using $\widehat{\sigma^2}$ = MSE $= \frac{\sum(y-\hat{y})^2}{n-2} = \frac{2696.21}{5} = 539.242$

$\hat{\sigma} = 23.2216$

Then we have to find standard error of β_1 using

$$S.E(\hat{\beta}_1) = S(\hat{\beta}_1) = \sqrt{S^2(\hat{\beta}_1)} = \sqrt{\frac{MSE}{\sum_{i=1}^n (X_i-\bar{X})^2}} = \frac{23.2216}{\sqrt{617.71}} = 0.934$$

1. Hypothesis

$H_0: \beta_1 = 0$ v.s $H_1: \beta_1 > 0$

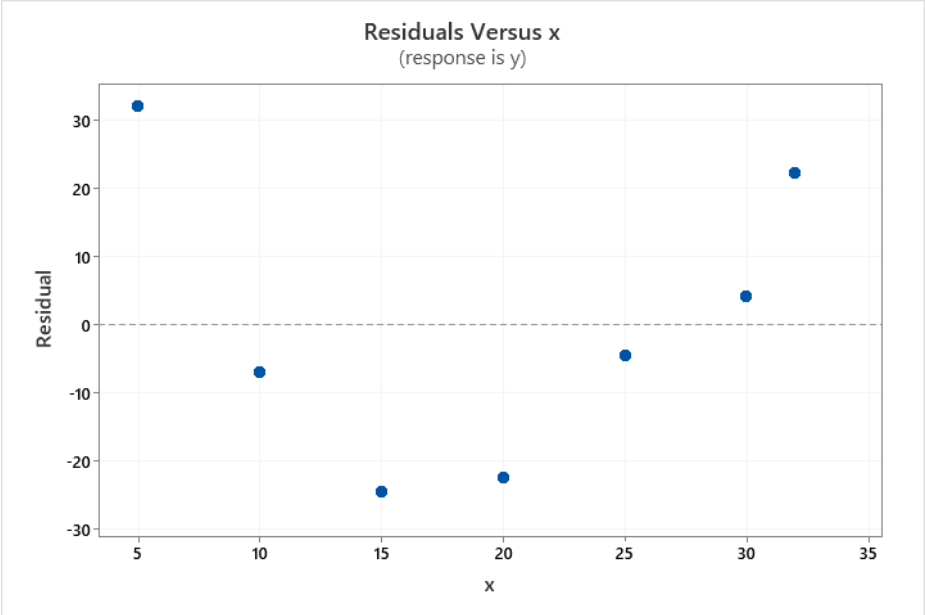
2. Test statistic

$$T_0 = \frac{b_1 - \beta_{10}}{s(b_1)} = \frac{b_1}{s(b_1)} = \frac{10.284}{0.934} = 11.0107$$

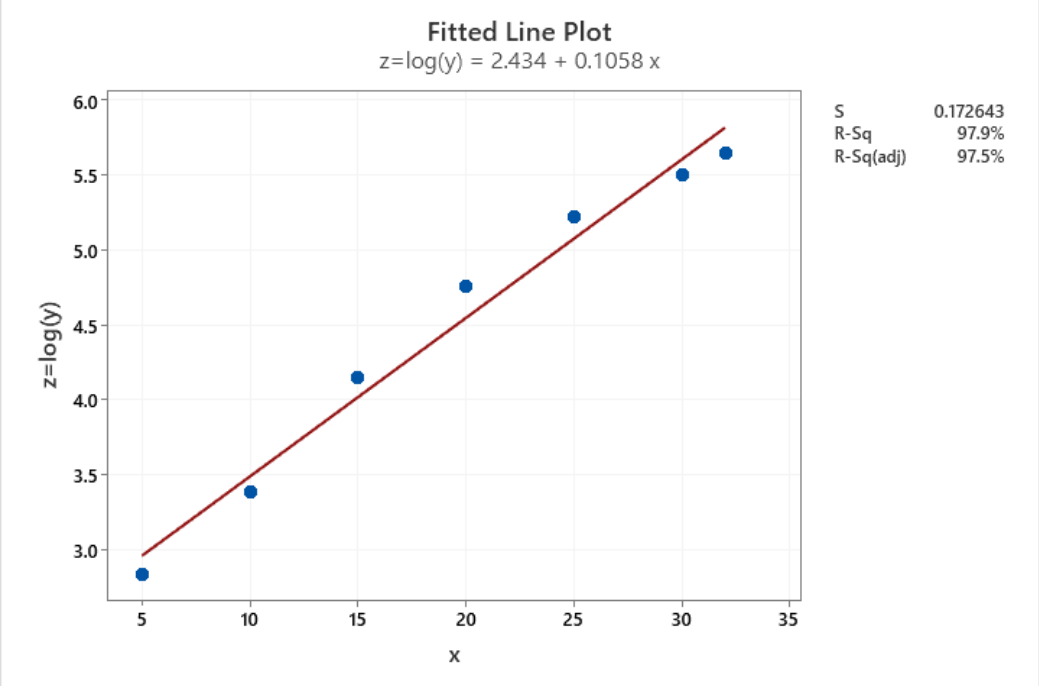
3. Decision: Reject H_0 if $T_0 > t_{(1-\alpha,\ n-2)}$, $11.0107 > t_{(0.95\,,5)} = 2.01505$

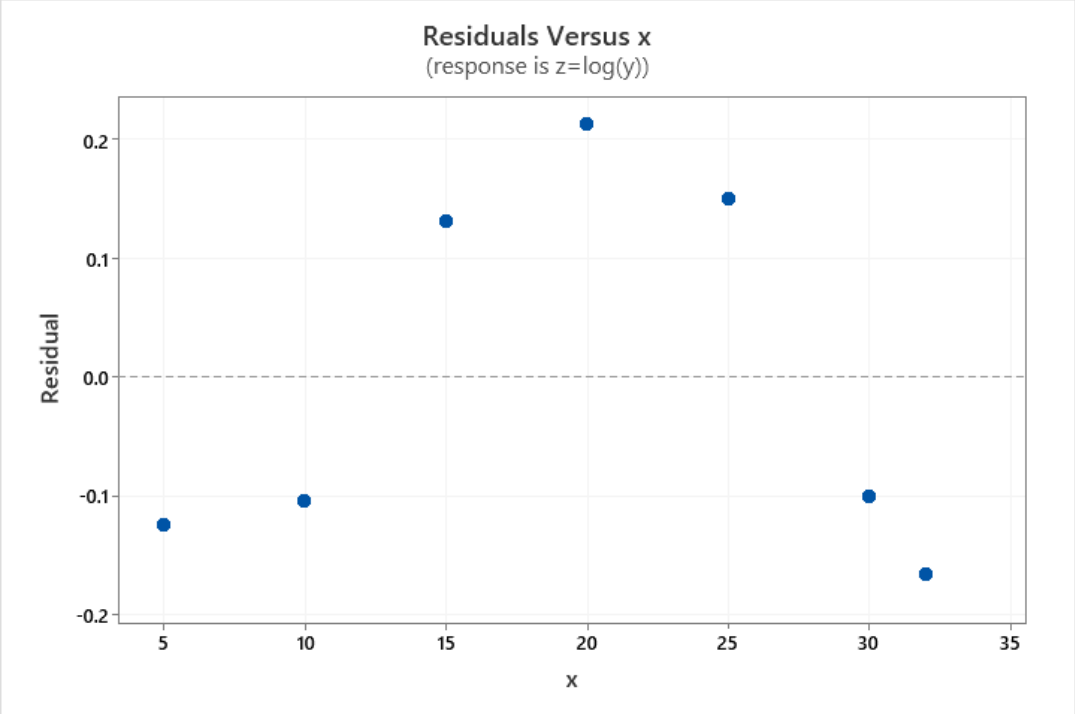
Then we reject H_0 at 5% significance level, which means that social security cost increases over time.

d.



The graph shows clear non-random pattern, indicating inadequacy of linear model. By looking to graph in (a) we would apply **log-transformation**, i.e. let $Z = log(y)\ in$ order to accommodate non-linear relationship between y and x.





Note that, unfortunately, the non-linearity in this dataset cannot be removed using a simple log-transformation. Fitting a linear regression of Z=log(y) on x , results in the estimated model $\hat{z} = \mathbf{2.434} + \mathbf{0.1058}\, x$ with $\hat{\sigma} = \sqrt{\text{MSE}} = \sqrt{0.02981} = 0.173$. The plots below indicate that there is still some pattern in the residuals.

Minitab Result: **Regression Analysis: z=log(y) versus x**

The regression equation is
z=log(y) = 2.434 + 0.1058 x

Model Summary

S	R-sq	R-sq(adj)
0.172643	97.89%	97.47%

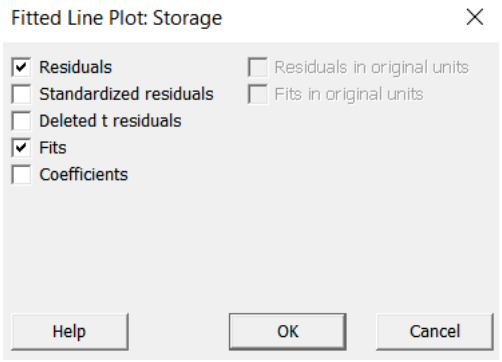
Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	6.91006	6.91006	231.84	0.000
Error	5	0.14903	0.02981		
Total	6	7.05908			

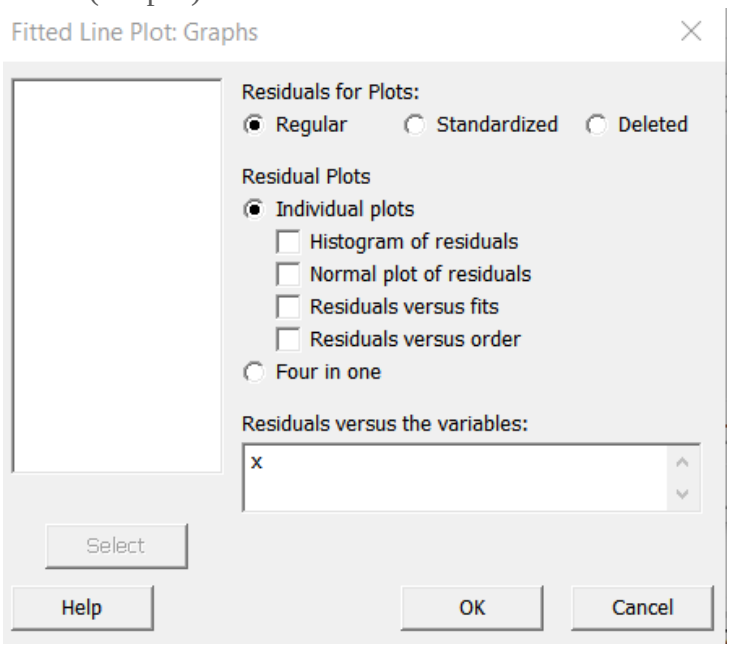
Steps by Minitab:

Stat>> Regression >> Fitted line plot >> Response(Y): y
Predictor (X): x

Click (Storge) >>



Click (Graphs) >>



To Calculate log (y) by Minitab:

Calc >> Calculator >>

Store result in variable: z=log(y)
Expression: log('y') >> ok

Exercise 2: The stopping distance $(y, in\ feet)$ of a car was studied in relation to its velocity $(x, in\ miles\ per\ hour(mph))$. The table below lists the stopping distances at 6 different velocities.

Velocity (mph): x	20.5	20.5	30.5	40.5	48.8	57.8
stopping distances (ft): y	15.4	13.3	33.9	73.1	113	142.6

a) Plot $\boldsymbol{y}$ against $\boldsymbol{x}$, and $\boldsymbol{z} = \sqrt{\boldsymbol{y}}$ against $\boldsymbol{x}$.
b) Compute the sample correlation coefficients of $\boldsymbol{y}$ with $\boldsymbol{x}$, and $\boldsymbol{z}$ with $\boldsymbol{x}$.
c) Fit a linear regression model to $\boldsymbol{y}$ on $\boldsymbol{x}$ and examine the residuals (the differences between the $\hat{\boldsymbol{y}}$ values and the $\boldsymbol{y}$ values, i.e $\hat{e}_i = y_i - \hat{y}_i$). The estimate of $\boldsymbol{\sigma}$ in this case is $\hat{\boldsymbol{\sigma}} = \mathbf{7.563}$
d) Fit a linear regression model to $\boldsymbol{z}$ *on* $\boldsymbol{x}$ and examine the residuals. The estimate of σin this case is $\hat{\boldsymbol{\sigma}} = \mathbf{0.322}$
e) Compute prediction intervals with a coverage probability of **0.95** for $\boldsymbol{y}$ and $\boldsymbol{z}$ when $\boldsymbol{x} = \mathbf{35.}$
f) Which model is better? Briefly explain why.

Solution:

a) To Calculate log (y) by Minitab:

Calc >> Calculator >> Store result in variable: Z Expression: SQRT('Y') >> ok

X	Y	$Z = \sqrt{y}$
20.5	15.4	3.924283
20.5	13.3	3.646917
30.5	33.9	5.822371
40.5	73.1	8.549854
48.8	113	10.63015
57.8	142.6	11.94152

Regression Analysis: Y versus X

The regression equation is
Y = - 62.05 + 3.493 X

Model Summary

S	R-sq	R-sq(adj)
7.56312	98.42%	98.03%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	14262.5	14262.5	249.34	0.000
Error	4	228.8	57.2		
Total	5	14491.3			

Regression Analysis: Z versus X

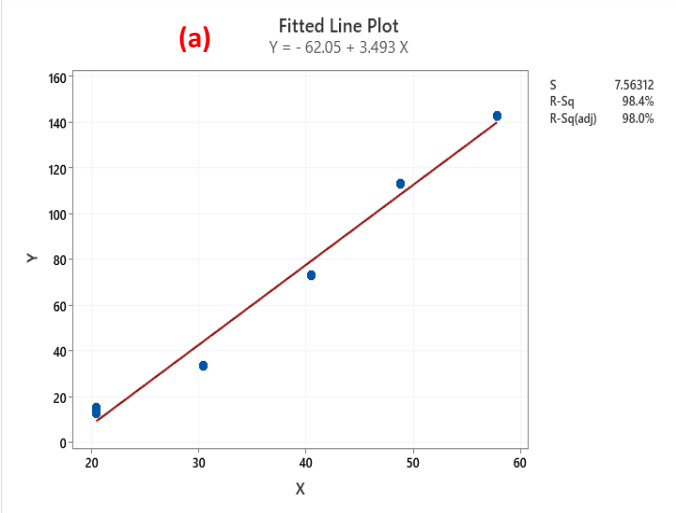
The regression equation is
z = - 0.8776 + 0.2277 X

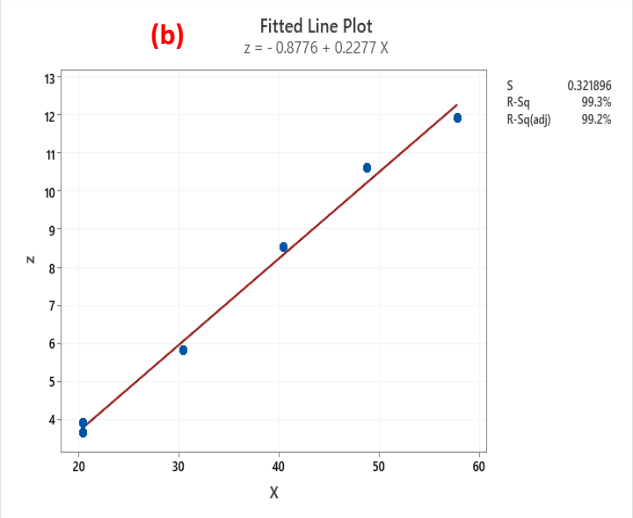
Model Summary

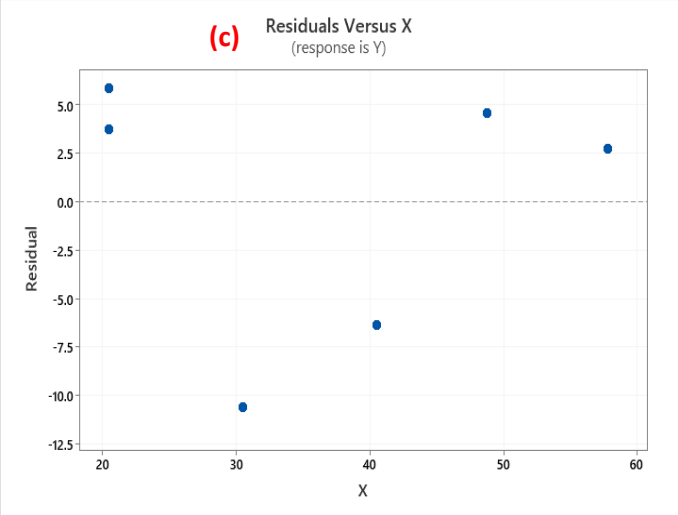
S	R-sq	R-sq(adj)
0.321896	99.32%	99.15%

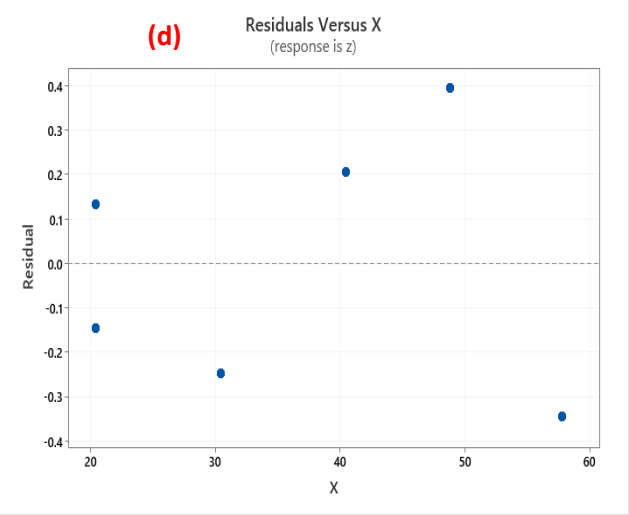
Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	60.6199	60.6199	585.04	0.000
Error	4	0.4145	0.1036		
Total	5	61.0344			









b) by Minitab: stat >> basic statstic >> correlation

The sample correlation between X and Y = 0.992

The sample correlation between X and Z = 0.997

$r_{XY} = \sqrt{R_{XY}^2} = \sqrt{0.9842} = 0.992$; $r_{XZ} = \sqrt{R_{XZ}^2} = \sqrt{0.9932} = 0.997$

c) To fit linear regression model, we must find b_0 and b_1

$b_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$; $b_0 = \bar{Y} - b_1\bar{X}$; $\bar{X} = 36.4333$; $\bar{Y} = 65.2166$

$b_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \frac{4083.167}{1168.953} = 3.4930$

$b_0 = \bar{Y} - b_1\bar{X} = 65.2167 - 3.4930 * 36.4333 = -62.0454$

$\hat{y} = -62.0454 + 3.4930x$

$\hat{\sigma} = \sqrt{\text{MSE}} = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}} = \sqrt{\frac{228.8029}{4}} = \sqrt{57.2} = 7.5631$

X	Y	$Z = \sqrt{Y}$	$\hat{Y}$	$e = Y - \hat{Y}$
20.5	15.4	3.924283	9.56	5.84
20.5	13.3	3.646917	9.56	3.74
30.5	33.9	5.822371	44.49	-10.59
40.5	73.1	8.549854	79.42	-6.32
48.8	113	10.63015	108.41	4.59
57.8	142.6	11.94152	139.85	2.75

The line was added in Figure (a) and the residuals are Plotted against (x) in Figure (c).

d) To fit linear regression model, we must find b_0 and b_1

$b_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Z_i-\bar{Z})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$; $b_0 = \bar{Z} - b_1\bar{X}$; $\bar{X} = 36.4333$; $\bar{Z} = 7.4192$

$b_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Z_i-\bar{Z})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \frac{266.1989}{1168.953} = 0.2277$

$b_0 = \bar{Z} - b_1\bar{X} = 65.2167 - 3.4930 * 36.4333 = -0.8776$

$\hat{z} = -0.8776 + 0.2277x$

$\hat{\sigma} = \sqrt{\text{MSE}} = \sqrt{0.1036} = 0.322$

X	Y	Z	$\hat{Z}$	$e = Z - \hat{Z}$
20.5	15.4	3.924283	-5.55	9.47
20.5	13.3	3.646917	-5.55	9.19
30.5	33.9	5.822371	-7.82	13.65
40.5	73.1	8.549854	-10.10	18.65
48.8	113	10.63015	-11.99	22.62
57.8	142.6	11.94152	-14.04	25.98

The line was added in Figure (b) and the residuals are Plotted against (x) in Figure (d).

e) $When\ x = 35$

$t_{1-\alpha/2,n-2} = t_{0.975,4} = 2.7764\ , n = 6$

For Prediction interval for Y: $\widehat{Y_h} \pm t_{\left(1-\frac{\alpha}{2};n-2\right)} s\left(\widehat{Y_{new}}\right)$

$\widehat{Y_h} = -62.0454 + 3.4930\ (\mathbf{35}) = 60.21\ ; \quad MSE = 7.563^2$

$S^2\left(\widehat{Y_{new}}\right) = \boldsymbol{MSE}\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]$

$= 7.563^2\left[1 + \frac{1}{6} + \frac{(35-36.433)^2}{1168.953}\right] = 7.563^2 * \mathbf{1.1684} = 66.833$

$S\left(\widehat{Y_{new}}\right) = 8.1751$

$\hat{y} \pm t_{\left(1-\frac{\alpha}{2},n-2\right)} * S\left(\widehat{Y_{new}}\right) = 60.21 \pm 22.6974$

$\mathbf{95\%PI\ for\ Y\ is: (37.513, 82.907)}$

For Prediction interval for Z: $\widehat{Z_h} \pm t_{\left(1-\frac{\alpha}{2};n-2\right)} s\left(\widehat{Z_{new}}\right)$

$\widehat{Z_h} = -0.8776 + 0.2277\ (\mathbf{35}) = 7.0919\ ; \quad MSE = 0.322^2$

$s^2\left(\widehat{Z_{new}}\right) = \boldsymbol{MSE}\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] = 0.322^2(\mathbf{1.1684}) = 0.1746$

$S\left(\widehat{Z_{new}}\right) = 0.3481$

$\hat{z} \pm t_{\left(1-\frac{\alpha}{2},n-2\right)} S\left(\widehat{Z_{new}}\right) = 7.0919 \pm 0.9663$

$\mathbf{95\%PI\ for\ Z\ is: (6.126, 8.058)}$

by using Minitab:

Stat > Regression > Regression> predict

Prediction of Y

Fit	SE Fit	95% CI	95% PI
60.2100	3.10387	(51.5923, 68.8277)	(37.5119, 82.9082)

Prediction of Z

Fit	SE Fit	95% CI	95% PI
7.09278	0.132104	(6.72600, 7.45956)	(6.12672, 8.05884)

f) since the sample correlation coefficient between z and x greater than that between y and x suggested a stronger linear relationship between z and x .also, the fitted model for **z** has much smaller errors as that reflected by value of $\sigma = 0.322$ which leads to more accurate prediction interval for **Stopping distance**.

Hence it seems a better option to fit a linear regression model to $z = \sqrt{y}$ instead of y.

Note: The improvement from using the z model would be mor evident with larger sample sizes.