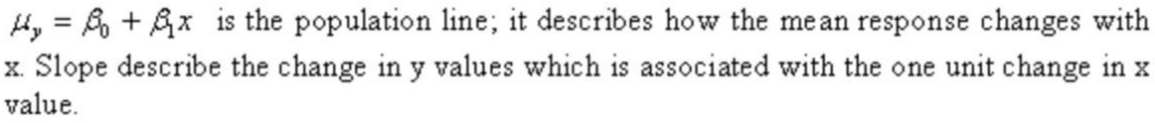
ST305-hw8

(Total: 49 points)

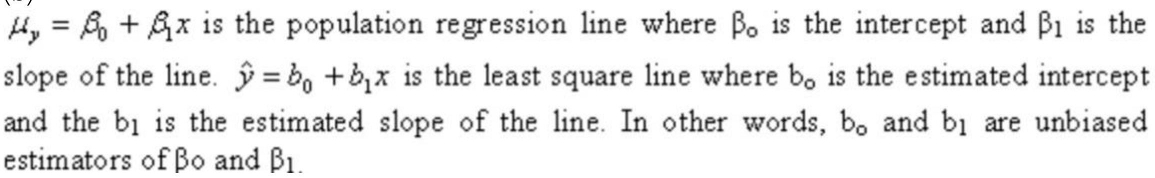
10.6 (3 points, one for each)

(a)



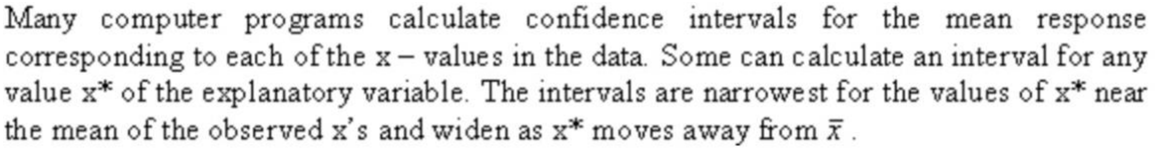
$\mu_y = \beta_0 + \beta_1 x$ is the population line; it describes how the mean response changes with x. Slope describe the change in y values which is associated with the one unit change in x value.

(b)



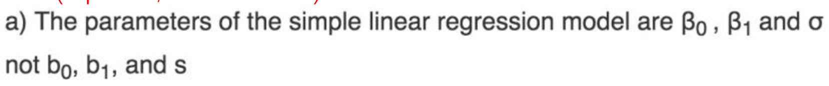
$\mu_y = \beta_0 + \beta_1 x$ is the population regression line where β_o is the intercept and β_1 is the slope of the line. $\hat{y} = b_0 + b_1 x$ is the least square line where b_o is the estimated intercept and the b_1 is the estimated slope of the line. In other words, b_o and b_1 are unbiased estimators of βo and β_1.

(c)

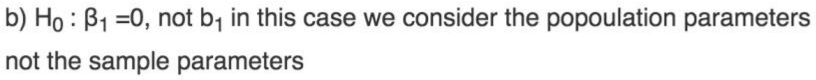


Many computer programs calculate confidence intervals for the mean response corresponding to each of the x – values in the data. Some can calculate an interval for any value x* of the explanatory variable. The intervals are narrowest for the values of x* near the mean of the observed x's and widen as x* moves away from $\bar{x}$.

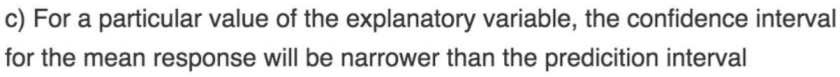
10.7 (3 points, one for each)



a) The parameters of the simple linear regression model are β_0 , β_1 and σ not b_0, b_1, and s



b) H_0 : β_1 =0, not b_1 in this case we consider the popoulation parameters not the sample parameters



c) For a particular value of the explanatory variable, the confidence interval for the mean response will be narrower than the predicition interval

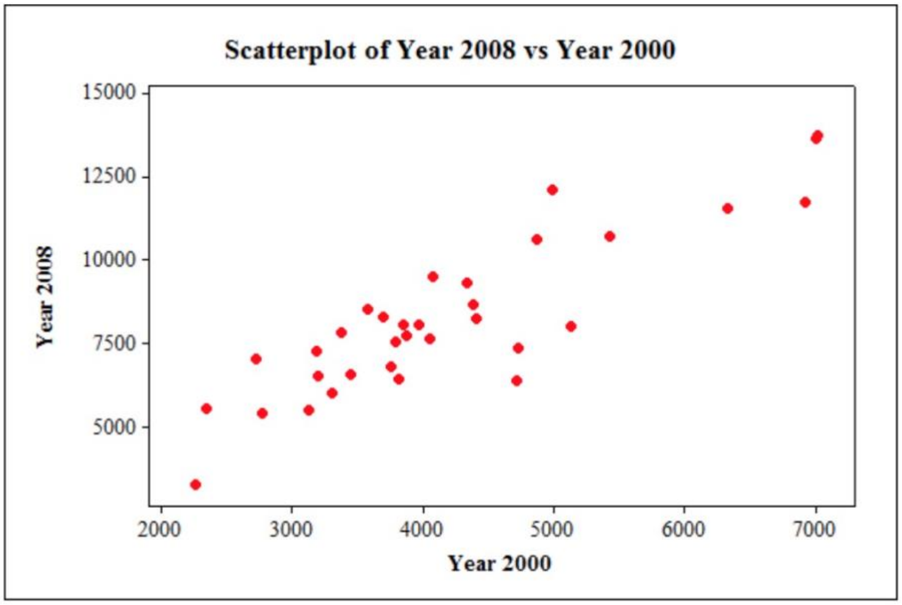
10.10 (6 points, one for each)

a)

There is information about public university tuition in 2000 and 2008.

Take the variable "2000 tuition" on the *x* axis and "2008 tuition" on the *y* axis. Plot the data on two-dimensional plane.

Scatter plot of "2000 tuition" versus "2008 tuition":

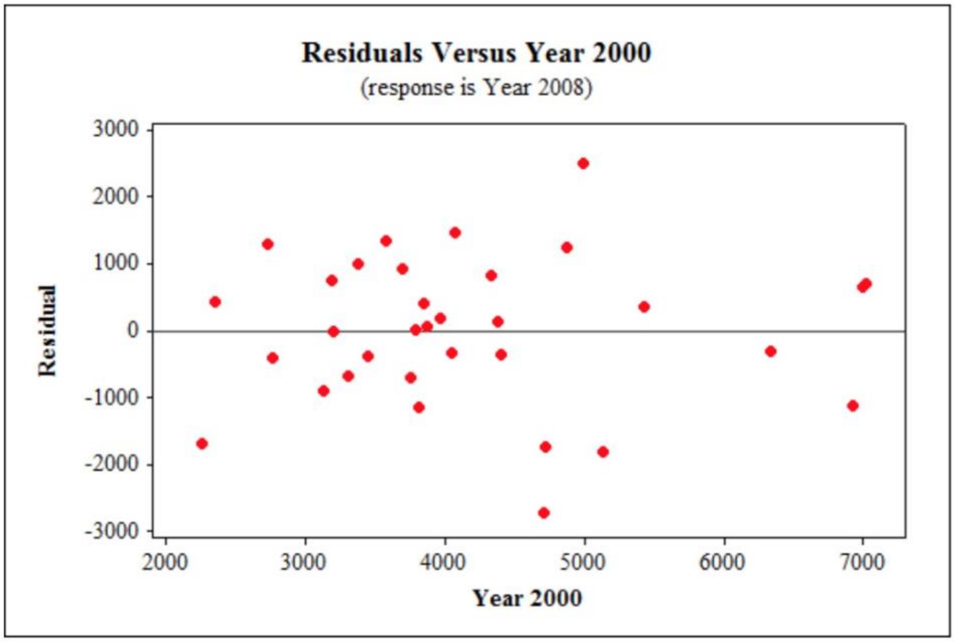


From the scatter plot we reveal that there is no point unusual and moreover, there is strong linear relationship exists between the tuition in 2000 and 2008.

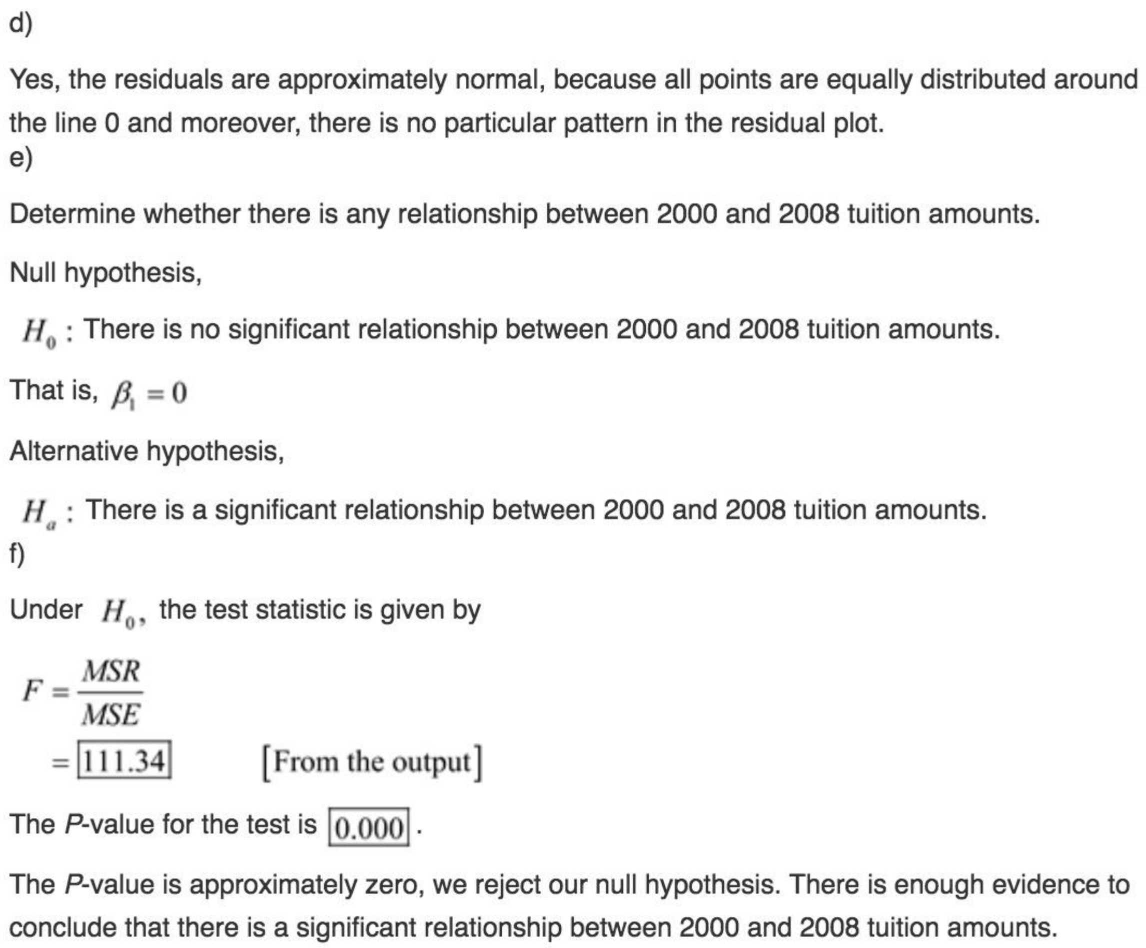
(b) $\hat{y} = 1133 + 1.69\,x$

(c)

Plot of residuals versus the 2000 tuition amount:



From the above plot there is no observation seems to be unusual.



d)

Yes, the residuals are approximately normal, because all points are equally distributed around the line 0 and moreover, there is no particular pattern in the residual plot.

e)

Determine whether there is any relationship between 2000 and 2008 tuition amounts.

Null hypothesis,

H_0 : There is no significant relationship between 2000 and 2008 tuition amounts.

That is, $\beta_1 = 0$

Alternative hypothesis,

H_a : There is a significant relationship between 2000 and 2008 tuition amounts.

f)

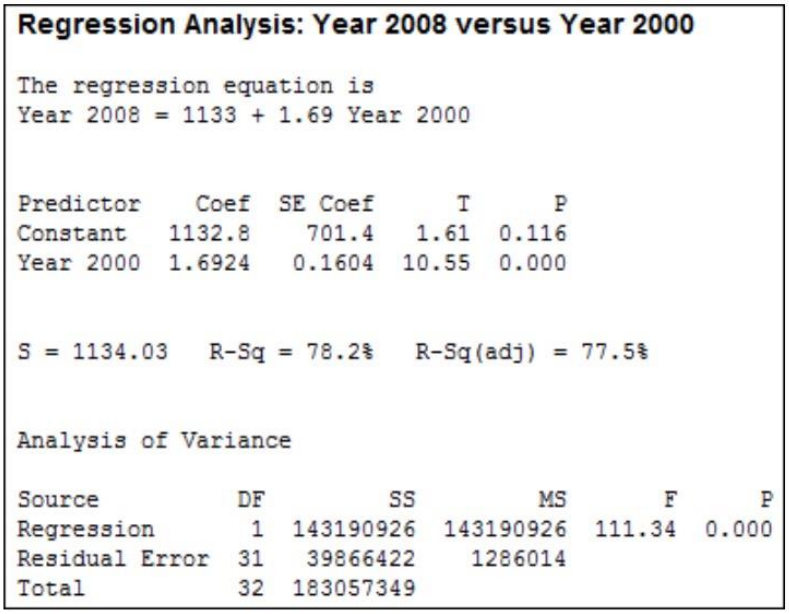
Under H_0, the test statistic is given by

$$\begin{aligned} F &= \frac{MSR}{MSE} \\ &= \boxed{111.34} \qquad \left[\text{From the output}\right] \end{aligned}$$

The *P*-value for the test is $\boxed{0.000}$.

The *P*-value is approximately zero, we reject our null hypothesis. There is enough evidence to conclude that there is a significant relationship between 2000 and 2008 tuition amounts.

10.11 (5 points, one for each)



```
Regression Analysis: Year 2008 versus Year 2000

The regression equation is
Year 2008 = 1133 + 1.69 Year 2000


Predictor     Coef  SE Coef      T      P
Constant    1132.8    701.4   1.61  0.116
Year 2000   1.6924   0.1604  10.55  0.000


S = 1134.03   R-Sq = 78.2%   R-Sq(adj) = 77.5%


Analysis of Variance

Source          DF         SS         MS       F      P
Regression       1  143190926  143190926  111.34  0.000
Residual Error  31   39866422    1286014
Total           32  183057349
```

From the output, the least-squares regression line is,

$\hat{y} = 1132.8 + 1.6924\,x$

a)

Determine 95% confidence interval for the slope.

$100(1-\alpha)$ percent confidence interval for β_1 is $b_1 \pm t^* \mathrm{SE}_{b_1}$

From the output, $b_1 = 1.6924,\ \mathrm{SE}_{b_1} = 0.1604,\ \text{and}\ \ n = 33$

Degrees of freedom,

$$\begin{aligned} df &= n-2 \\ &= 33-2 \\ &= 31 \end{aligned}$$

At 95% confidence with $df = 31$, the tabulated value is $t^* = 2.04$.

The required confidence interval is,

$$\begin{aligned} b_1 \pm t^* \mathrm{SE}_{b_1} &= 1.6924 \pm (2.04)(0.1604) \\ &= 1.6924 \pm 0.3272 \\ &= \boxed{(1.3652,\ 2.0196)} \end{aligned}$$

Therefore, the 95% confidence interval for the slope is between 1.3652 and 2.0196.

Moreover, there is a one dollar difference in the tuition in year 2008 from 2000.

b)

From the output, the coefficient of determination is $R^2 = 78.2\%$. It means that $\boxed{78.2}$ percent of the variability in 2008 tuition is explained by a linear regression model using the 2000 tuition.

c)

Suppose the tuition at university S in year 2000 is $5100. Then, the estimated predicted tuition in year 2008 is,

$$\begin{aligned}\hat{y} &= 1132.8 + 1.6924\ (5100)\\ &= 1132.8 + 8631.24\\ &= \boxed{\$9,764.04}\end{aligned}$$

d)

Suppose the tuition at university M in year 2000 is $8700. Then, the estimated predicted tuition in year 2008 is,

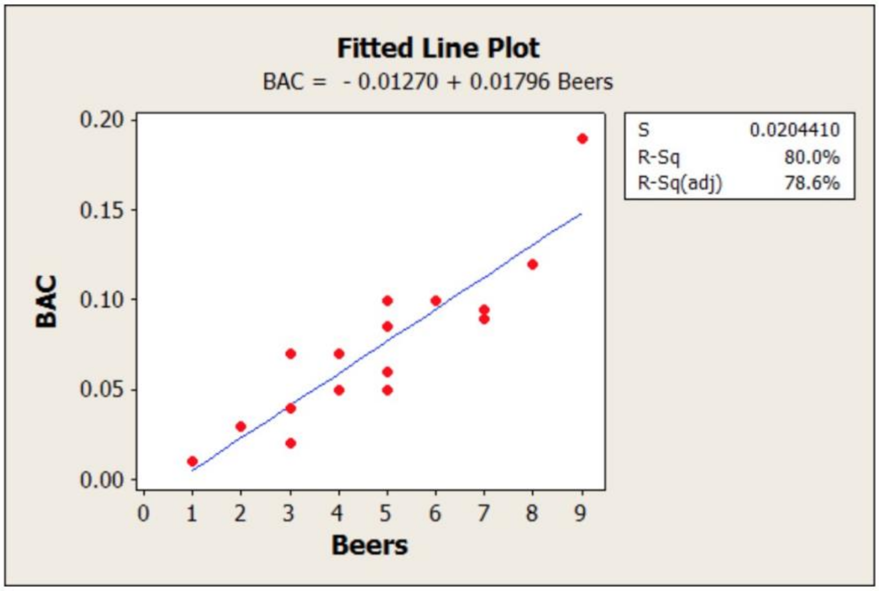
$$\begin{aligned}\hat{y} &= 1132.8 + 1.6924\ (8700)\\ &= 1132.8 + 14723.88\\ &= \boxed{\$15,856.68}\end{aligned}$$

e)

The predicted value for university S is $9,764 which is in reasonable limits because it lies within the range, but for university U ($ 15,856.68) does not lies within the range.

10.13 (3 points, one for each)

(a)



the least squares regression equation is,

$$\text{BAC} = -0.01270 + 0.01796\ \text{Beers}$$

The coefficient of determination is,

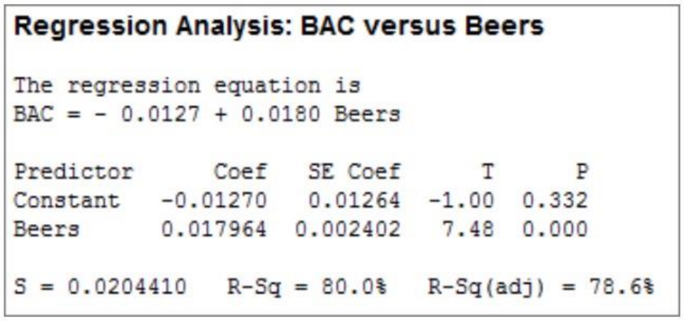
$$r^2 = 80\%$$

About 80% of the variation in BAC is explained by the beers.

(b)

The null and alternative hypotheses are,

$$H_0 : \beta = 0$$
$$H_a : \beta > 0$$



Regression Analysis: BAC versus Beers

```
The regression equation is
BAC = - 0.0127 + 0.0180 Beers

Predictor       Coef   SE Coef      T      P
Constant    -0.01270   0.01264  -1.00  0.332
Beers       0.017964  0.002402   7.48  0.000

S = 0.0204410    R-Sq = 80.0%    R-Sq(adj) = 78.6%
```

From the obtained output, the test statistic value corresponding to the beers is $t = 7.48$ and the two tailed P-value is 0. Obviously one tailed P-value is 0. The P-value is less than any level of significance. Reject the null hypothesis. Therefore, it can be concluded that drinking more beers increases BAC.

(c)

Find the 90% prediction interval for Steve's BAC is 0.08.

The estimated mean balance is,

$$\begin{aligned}\hat{\mu} &= \beta_0 + \beta_1 x \\ &= -0.0127 + 0.0179(0.08) \\ &= 0.077118\end{aligned}$$

The standard error is $SE_{\hat{y}} = 0.0024$

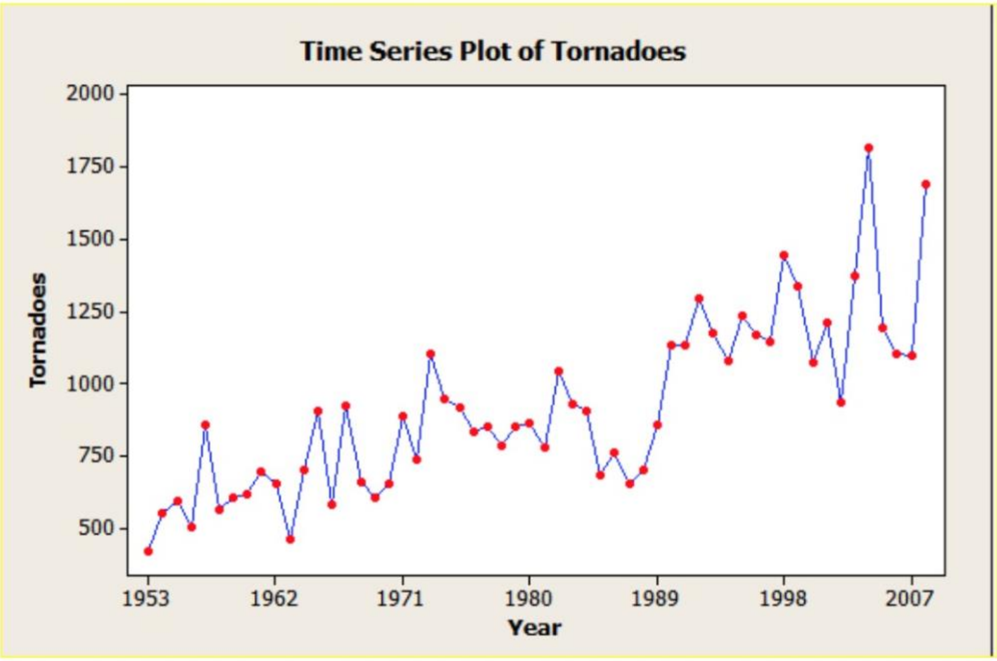
Therefore, the 90% prediction interval is,

$$\begin{aligned}\hat{\mu} \pm t^{*} SE_{\hat{\mu}} &= 0.077 \pm 15.42(0.0024) \\ &= 0.077 \pm 0.0370 \\ &= (0.040, 0.114)\end{aligned}$$

10.19 (5 points, one for each)

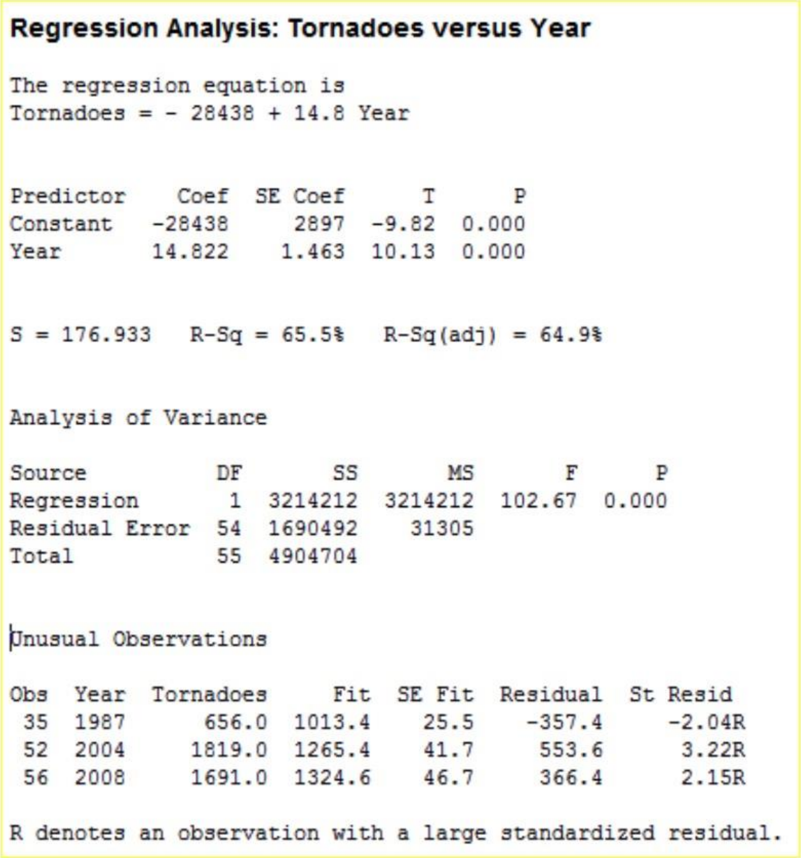
(a)

Find the line plot of tornadoes.



From the linear plot, observe that there exists a strong upward relation between Year and Number of tornadoes. That is, the number of tornadoes is increasing year over year. Also, the observation corresponding to year 2004 is high compared to all other observations.

(b)



Regression Analysis: Tornadoes versus Year

```
The regression equation is
Tornadoes = - 28438 + 14.8 Year


Predictor    Coef  SE Coef      T      P
Constant   -28438     2897  -9.82  0.000
Year       14.822    1.463  10.13  0.000


S = 176.933   R-Sq = 65.5%   R-Sq(adj) = 64.9%


Analysis of Variance

Source          DF       SS       MS       F      P
Regression       1  3214212  3214212  102.67  0.000
Residual Error  54  1690492    31305
Total           55  4904704


Unusual Observations

Obs  Year  Tornadoes     Fit  SE Fit  Residual  St Resid
 35  1987      656.0  1013.4    25.5    -357.4    -2.04R
 52  2004     1819.0  1265.4    41.7     553.6     3.22R
 56  2008     1691.0  1324.6    46.7     366.4     2.15R

R denotes an observation with a large standardized residual.
```

From the output, the regression line between Year and annual tornadoes is

$$\text{Tornadoes} = -28,438 + 14.822\ (\text{Year})$$

The slope of the regression line 14.822 is the annual increment in the tornadoes. From the output, the Standard error of the corresponding slope is 1.463. There are 56 observations.

That is, $b = 14.822, s_b = 1.463 \text{ and } n=56$

The critical value of ‘t’ at (56–2=) 54 degrees of freedom corresponding to 0.05 significance level is 2.005. That is, $t_{\frac{0.05}{2},54} = 2.005$

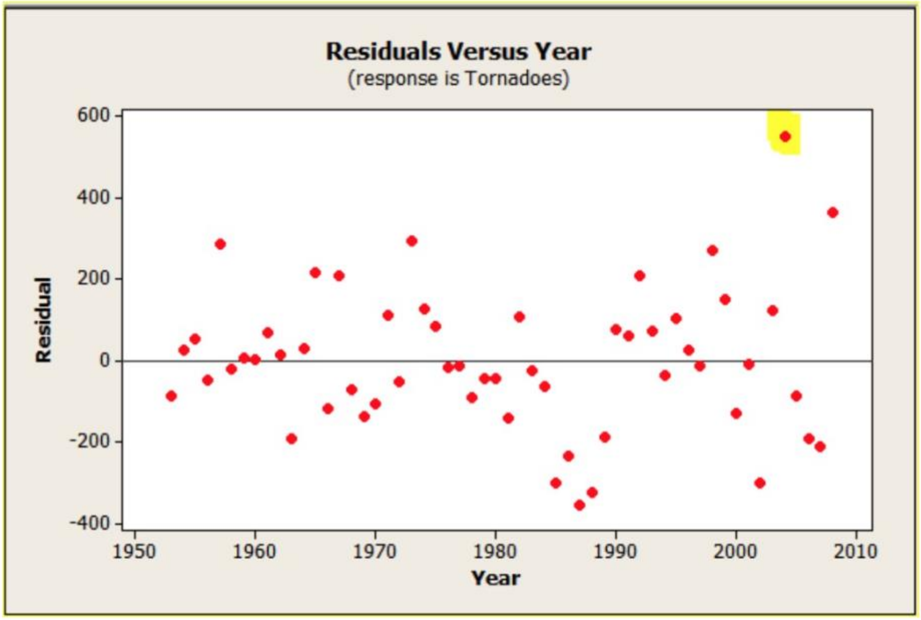
Then, the 95% confidence interval for the annual increment in the tornadoes is given by

$$\begin{aligned} b \pm t_{\alpha/2,(n-2)} \times s_b &= 14.822 \pm 2.005 \times 1.463 \\ &= 14.822 \pm 2.933 \\ &= (11.89, 17.76) \end{aligned}$$

The 95% confidence interval for the annual increment in the tornadoes is (11.89, 17.76) c)

Find the predicted value for each year using the regression equation and graph residuals versus year plot.

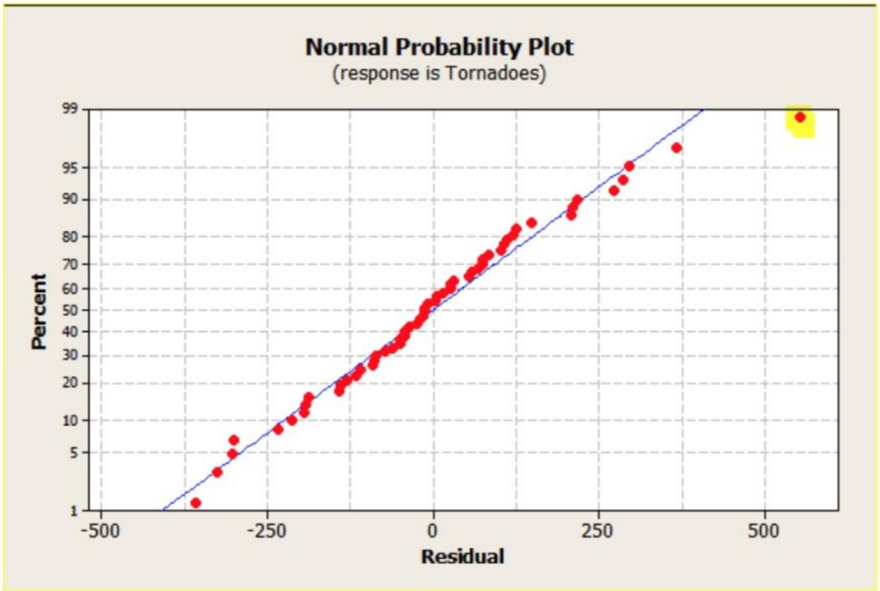
Find the residuals versus year plot:



The observations in the above plot are random and are not showing any pattern on either side of zero line. Also, observe that the highlighted point corresponds to year 2004 which is the highest points and seems to be an outlier.

(d)

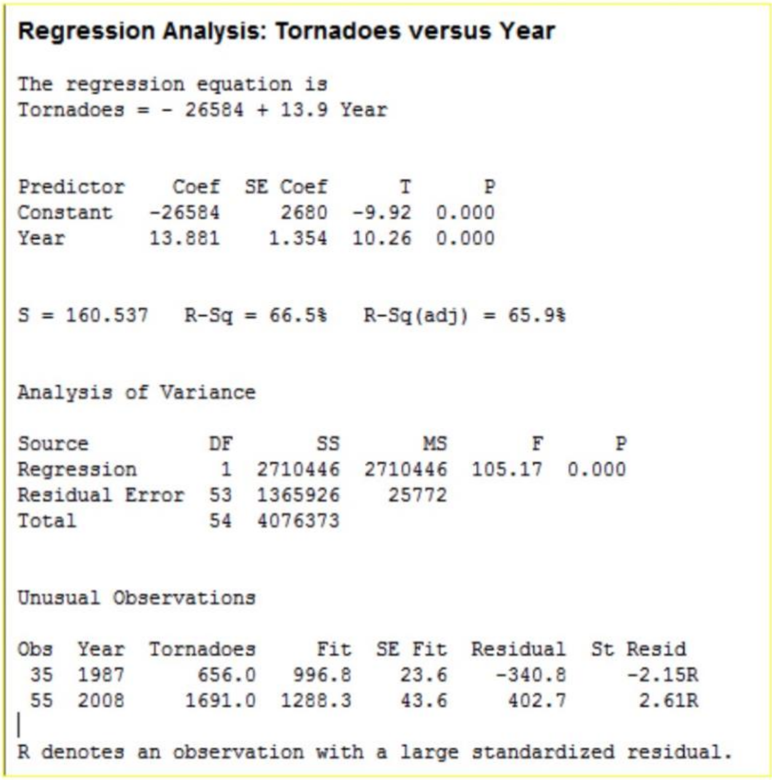
Find the normal probability plot of the residuals.



From above normal plot, observe that all points are very close to straight line except one point. Ignoring the highest point, the remaining observations are normally distributed.

e)

Remove the highest observation in tornadoes column and run the regression on remaining observations.



Regression Analysis: Tornadoes versus Year

```
The regression equation is
Tornadoes = - 26584 + 13.9 Year


Predictor    Coef  SE Coef      T      P
Constant   -26584     2680  -9.92  0.000
Year       13.881    1.354  10.26  0.000


S = 160.537   R-Sq = 66.5%   R-Sq(adj) = 65.9%


Analysis of Variance

Source          DF       SS       MS       F      P
Regression       1  2710446  2710446  105.17  0.000
Residual Error  53  1365926    25772
Total           54  4076373


Unusual Observations

Obs  Year  Tornadoes     Fit  SE Fit  Residual  St Resid
 35  1987      656.0   996.8    23.6    -340.8    -2.15R
 55  2008     1691.0  1288.3    43.6     402.7     2.61R

R denotes an observation with a large standardized residual.
```

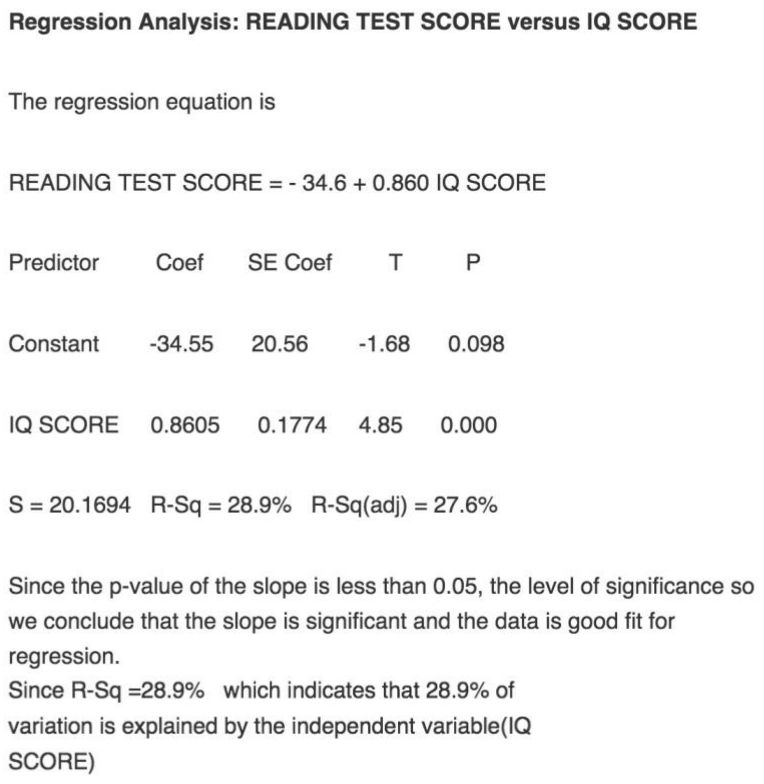
From the output, the regression line between Year and annual tornadoes is

$$\text{Tornadoes} = -26{,}584 + 13.9\ (\text{Year})$$

Observe that the slope of the regression line for modified data reduced to 13.9 which is the annual increment in the tornadoes.

10.28. (2 points, one for each)

(a)

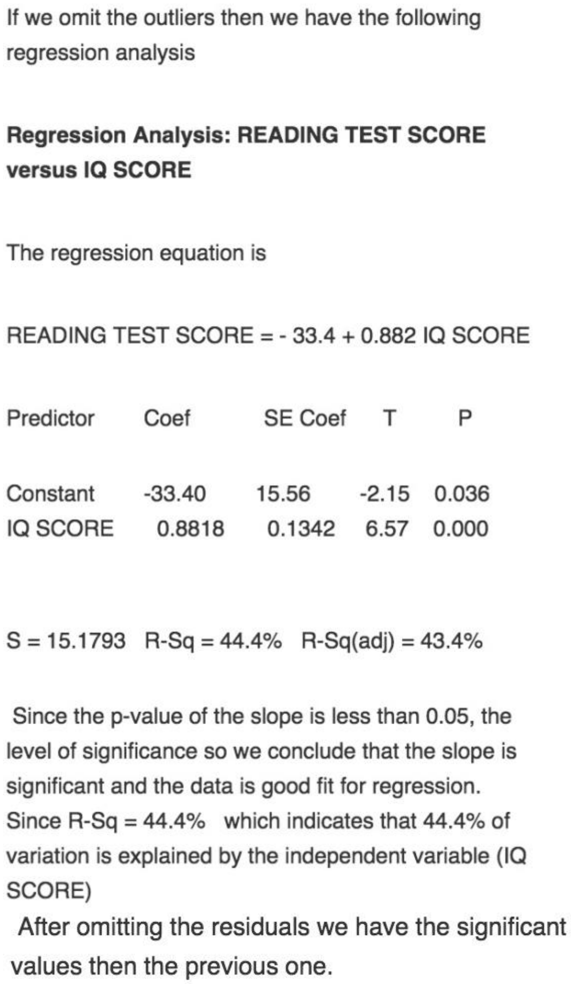


Regression Analysis: READING TEST SCORE versus IQ SCORE

```
The regression equation is

READING TEST SCORE = - 34.6 + 0.860 IQ SCORE

Predictor      Coef     SE Coef      T       P

Constant      -34.55     20.56     -1.68    0.098

IQ SCORE     0.8605     0.1774    4.85    0.000

S = 20.1694   R-Sq = 28.9%   R-Sq(adj) = 27.6%
```

Since the p-value of the slope is less than 0.05, the level of significance so we conclude that the slope is significant and the data is good fit for regression.
Since R-Sq =28.9%   which indicates that 28.9% of variation is explained by the independent variable(IQ SCORE)

(b)



If we omit the outliers then we have the following regression analysis

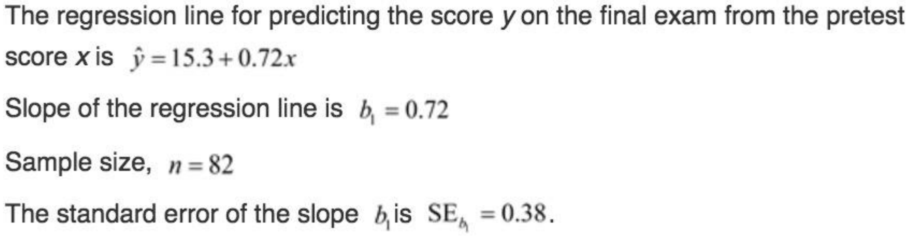
Regression Analysis: READING TEST SCORE versus IQ SCORE

```
The regression equation is

READING TEST SCORE = - 33.4 + 0.882 IQ SCORE

Predictor    Coef      SE Coef    T      P

Constant     -33.40     15.56    -2.15   0.036
IQ SCORE      0.8818     0.1342   6.57   0.000


S = 15.1793   R-Sq = 44.4%   R-Sq(adj) = 43.4%
```

Since the p-value of the slope is less than 0.05, the level of significance so we conclude that the slope is significant and the data is good fit for regression. Since R-Sq = 44.4% which indicates that 44.4% of variation is explained by the independent variable (IQ SCORE)

After omitting the residuals we have the significant values then the previous one.

10.34 (2 points, one for each)



The regression line for predicting the score *y* on the final exam from the pretest score *x* is $\hat{y} = 15.3 + 0.72x$

Slope of the regression line is $b_1 = 0.72$

Sample size, $n = 82$

The standard error of the slope b_1 is $\text{SE}_{b_1} = 0.38$.

a)

Null hypothesis,

$H_0 : \beta_1 = 0$

[There is no linear relationship between the pretest score and final exam score]

Alternative hypothesis,

$H_0 : \beta_1 \neq 0$

[There is a linear relationship between the pretest score and final exam score]

Let level of significance be $\alpha = 0.05$

Under H_0, the test statistic is given by

$$\begin{aligned} t &= \frac{b_1 - 0}{\mathrm{SE}_{b_1}} \\ &= \frac{0.72 - 0}{0.38} \\ &\approx 1.895 \end{aligned}$$

Degrees of freedom,

$$\begin{aligned} df &= n - 2 \\ &= 82 - 2 \\ &= 80 \end{aligned}$$

At 95% confidence with $df = 80$, *t* distribution critical values table gives, $t^* = 1.99$.

Since our test is two tailed, we reject H_0 if $t < -1.99$ or if $t > 1.99$.

The calculated value (1.895) is not greater than critical value (1.99), so we fail to reject our null hypothesis, $H_0 : \beta_1 = 0$.

We conclude that there is no linear relationship between the pretest score and the final exam score.

b)

Null hypothesis,

$H_0 : \beta_1 = 0$

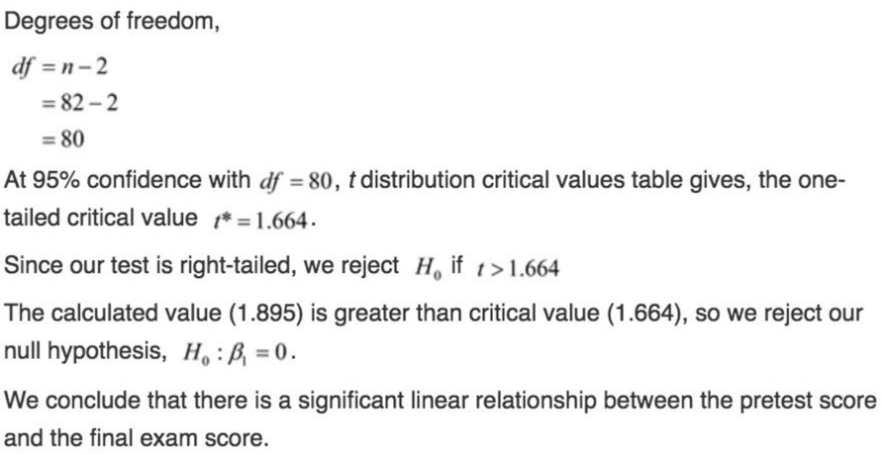
Alternative hypothesis,

$H_0 : \beta_1 > 0$

Let level of significance be $\alpha = 0.05$

Under H_0, the test statistic is given by

$$\begin{aligned} t &= \frac{b_1 - 0}{\mathrm{SE}_{b_1}} \\ &= \frac{0.72 - 0}{0.38} \\ &\approx 1.895 \end{aligned}$$



Degrees of freedom,

$$\begin{aligned} df &= n-2 \\ &= 82-2 \\ &= 80 \end{aligned}$$

At 95% confidence with $df = 80$, *t* distribution critical values table gives, the one-tailed critical value $t^* = 1.664$.

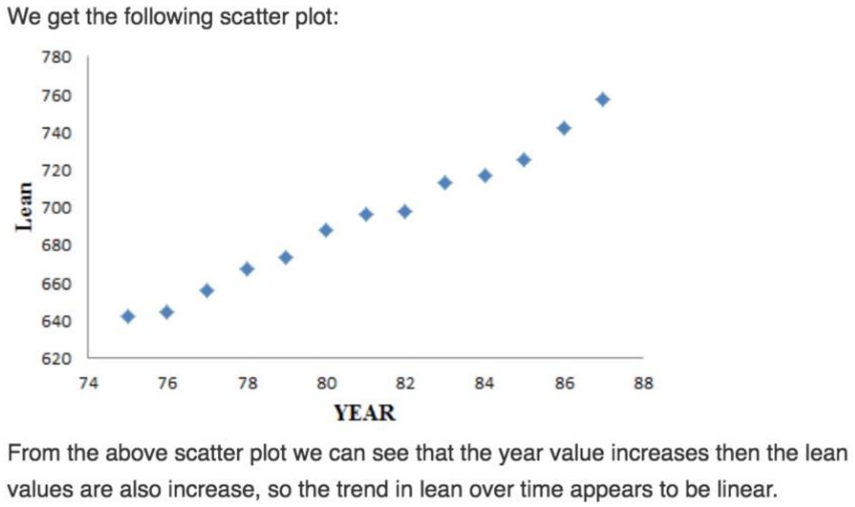
Since our test is right-tailed, we reject H_0 if $t > 1.664$

The calculated value (1.895) is greater than critical value (1.664), so we reject our null hypothesis, $H_0 : \beta_1 = 0$.

We conclude that there is a significant linear relationship between the pretest score and the final exam score.

10.37 (3 points, one for each)

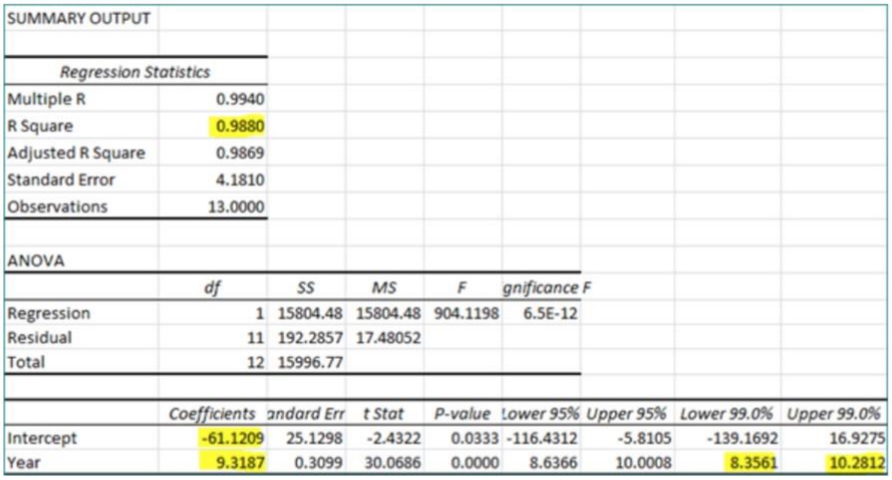
(a)



We get the following scatter plot:

From the above scatter plot we can see that the year value increases then the lean values are also increase, so the trend in lean over time appears to be linear.

(b)



SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.9940							
R Square	0.9880							
Adjusted R Square	0.9869							
Standard Error	4.1810							
Observations	13.0000							
ANOVA								
	df	*SS*	*MS*	*F*	*gnificance F*			
Regression	1	15804.48	15804.48	904.1198	6.5E-12			
Residual	11	192.2857	17.48052					
Total	12	15996.77						
	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	-61.1209	25.1298	-2.4322	0.0333	-116.4312	-5.8105	-139.1692	16.9275
Year	9.3187	0.3099	30.0686	0.0000	8.6366	10.0008	8.3561	10.2812

From the regression output,

The least square line is,

$$\widehat{\text{Lean}} = -61.1209 + 9.32\text{YEAR}$$

The value of R-square is 0.9880.

Therefore, the 98.9% of variation in lean is explained by the least square line.

c) Using the regression output in part (b),

The 99% confidence interval for the average rate of change of the lean is lies between 8.3561 and 10.2812.

10.45 (6 points, one for each blank)

From the available information,

The regression equation is

$$\text{EAFE} = -2.58 + 0.775\ \text{S\&P}$$

Number of observations, $n = 20$

Degrees of freedom for the total, $\text{DFT} = 19$

Sum of squares for the model, $\text{SSM} = 4560.6$

Sum of squares for the total, $\text{SST} = 8556$

Degrees of freedom for the model, $\text{DFM} = 1$

Degrees of freedom for the residual error,

$$\begin{aligned}\text{DFE} &= n - 2 \\ &= 20 - 1 \\ &= 18\end{aligned}$$

Sum of squares for the residual error,

$$\begin{aligned}\text{SSE} &= \text{SST} - \text{SSM} \\ &= 8556 - 4560.6 \\ &= 3995.4\end{aligned}$$

Mean sum of squares for the model,

$$\begin{aligned}\text{MSM} &= \frac{\text{SSM}}{\text{DFM}}\\ &= \frac{4560.6}{1}\\ &= 4560.6\end{aligned}$$

Mean sum of squares for the residual error,

$$\begin{aligned}\text{MSE} &= \frac{\text{SSE}}{\text{DFE}}\\ &= \frac{3995.4}{18}\\ &\approx 221.97\end{aligned}$$

The test statistic is given by

$$\begin{aligned}\text{F} &= \frac{\text{MSM}}{\text{MSE}}\\ &= \frac{4560.6}{221.97}\\ &\approx 20.55\end{aligned}$$

Therefore, the analysis of variance table is

Source	DF	SS	MS	F
Regression	1	4560.6	4560.6	20.55
Residual Error	18	3995.4	221.97	
Total	19	8556.0		

10.46 (1 point)

From the available information, the analysis of variance table is

Source	DF	SS	MS	F
Regression	1	4560.6	4560.6	20.55
Residual Error	18	3995.4	221.97	
Total	19	8556.0		

From the table, $\text{MSE} = 221.97$

The value of the regression standard error is

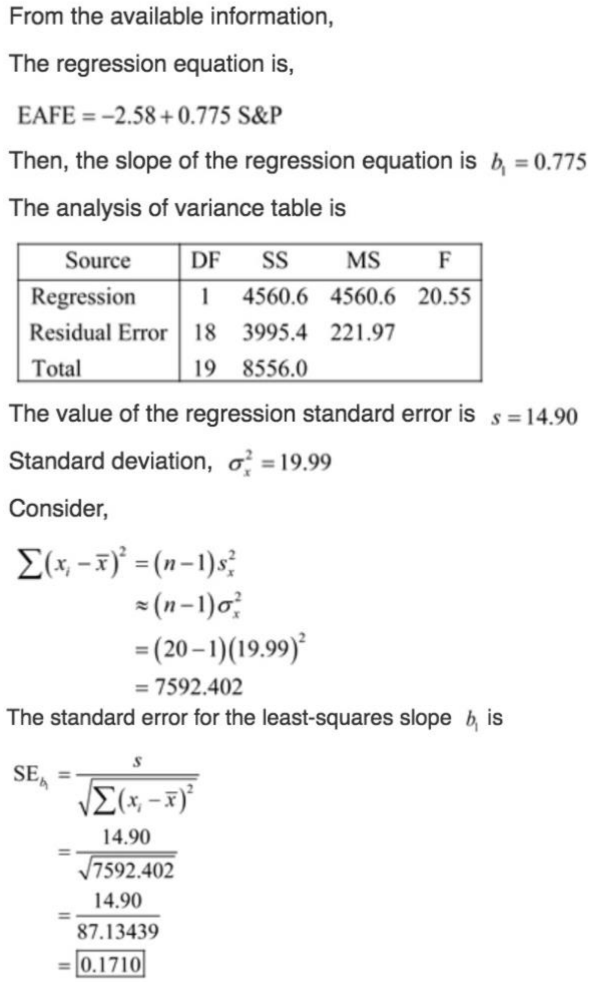
$$\begin{aligned}s &= \sqrt{\text{MSE}}\\ &= \sqrt{221.97}\\ &= 14.89866\\ &\approx \boxed{14.90}\end{aligned}$$

The value of the regression squared correlation is

$$\begin{aligned}r^2 &= \frac{\text{SSM}}{\text{SST}}\\ &= \frac{4560.6}{8556.0}\\ &= 0.533029\\ &\approx \boxed{53.3\%}\end{aligned}$$

Therefore, the coefficient of determination for the model is 53.3 percent.

10.47 (2 points, one for each)



From the available information,

The regression equation is,

$\text{EAFE} = -2.58 + 0.775\ \text{S\&P}$

Then, the slope of the regression equation is $b_1 = 0.775$

The analysis of variance table is

Source	DF	SS	MS	F
Regression	1	4560.6	4560.6	20.55
Residual Error	18	3995.4	221.97	
Total	19	8556.0		

The value of the regression standard error is $s = 14.90$

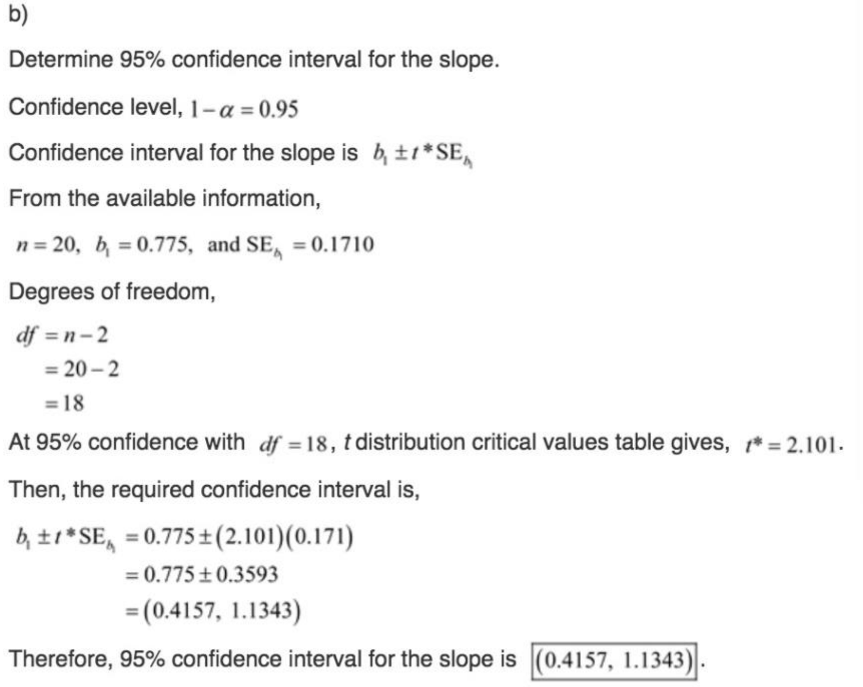
Standard deviation, $\sigma_x^2 = 19.99$

Consider,

$$\begin{aligned}\sum(x_i - \bar{x})^2 &= (n-1)s_x^2 \\ &\approx (n-1)\sigma_x^2 \\ &= (20-1)(19.99)^2 \\ &= 7592.402\end{aligned}$$

The standard error for the least-squares slope b_1 is

$$\begin{aligned}\text{SE}_{b_1} &= \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}} \\ &= \frac{14.90}{\sqrt{7592.402}} \\ &= \frac{14.90}{87.13439} \\ &= \boxed{0.1710}\end{aligned}$$



b)

Determine 95% confidence interval for the slope.

Confidence level, $1 - \alpha = 0.95$

Confidence interval for the slope is $b_1 \pm t^* \text{SE}_{b_1}$

From the available information,

$n = 20,\ b_1 = 0.775,\ \text{and SE}_{b_1} = 0.1710$

Degrees of freedom,

$$\begin{aligned}df &= n-2 \\ &= 20-2 \\ &= 18\end{aligned}$$

At 95% confidence with $df = 18$, t distribution critical values table gives, $t^* = 2.101$.

Then, the required confidence interval is,

$$\begin{aligned}b_1 \pm t^* \text{SE}_{b_1} &= 0.775 \pm (2.101)(0.171) \\ &= 0.775 \pm 0.3593 \\ &= (0.4157,\ 1.1343)\end{aligned}$$

Therefore, 95% confidence interval for the slope is $\boxed{(0.4157,\ 1.1343)}$.