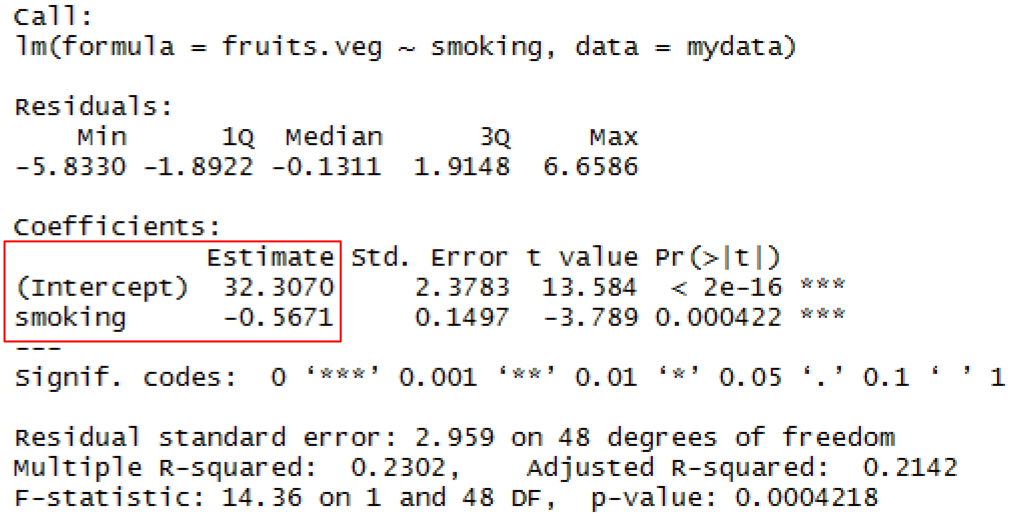
COMM 581 - Assignment #3
Simple Linear Regression – Part 2

Name: _____Gurpal Bisra____________ **Total: 20 marks**
Due date: Monday Sept. 26, 2015 (11pm)

Background: Continuation from Assignment #2 (Subtotal: 5 marks)
Explanatory variable: % of people who smoke every day
Response variable: % of people who eat at least 5 servings of fruits and vegetables every day

1. State the estimates of the co-efficients (b_0, b_1) and calculate their 95% confidence intervals. (**1 mark**)

The estimates of Bo and B1 are 32.307 and -0.5671, respectively. In R, I used the summary( ) command and verified my Bo and B1 coefficients were correct as illustrated below:



```
Call:
lm(formula = fruits.veg ~ smoking, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.8330 -1.8922 -0.1311  1.9148  6.6586 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  32.3070     2.3783  13.584  < 2e-16 ***
smoking      -0.5671     0.1497  -3.789 0.000422 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.959 on 48 degrees of freedom
Multiple R-squared:  0.2302,	Adjusted R-squared:  0.2142 
F-statistic: 14.36 on 1 and 48 DF,  p-value: 0.0004218
```

I calculated the coefficients’ 95% confidence intervals below. In my Excel spreadsheet, from assignment 2, I had calculated the following pieces of information: (1) MSE = 8.756953416; (2) SSx = 390.9832; (3) $\bar{x}$ = 15.644. I addition, I looked up the following t value using a student’s t-distribution: $t_{1-\frac{\alpha}{2},n-2}$ = 2.0106.

$$s_{b_1} = \sqrt{\frac{MSE}{SSx}} = \sqrt{\frac{8.756953416}{390.9832}} = 0.14965715029$$

$$s_{b_0} = \sqrt{MSE(\frac{1}{n} + \frac{\bar{x}^2}{SSx})} = \sqrt{(8.756953416)(\frac{1}{50} + \frac{15.644^2}{390.9832})} = 2.378345481$$

$$For\ B_0 = b_0 \pm (t_{1-\frac{\alpha}{2},n-2}) * s_{b_0}$$

$$For\ B_1 = b_1 \pm (t_{1-\frac{\alpha}{2},n-2}) * s_{b_1}$$

Hence, the 95% confidence intervals are as follows:

b_0 [27.525, 37.089]

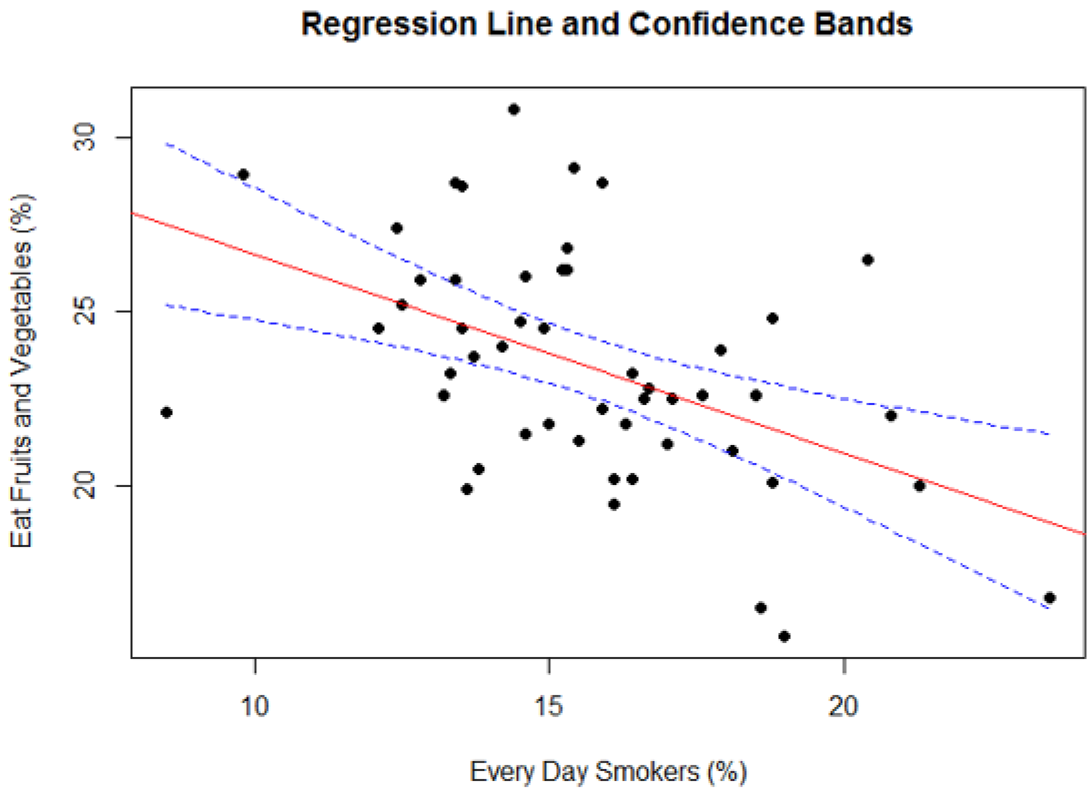
b_1 [-0.8680 , -0.2662]

2. State the model (b_0, b_1) in the form $\hat{y}_i = b_0 + b_1 x_i$, replacing b_0 and b_1 with their estimates. **(0.5 marks)**

$\hat{y_i} = 32.307 - (0.5671 * x_i)$

3. Create a plot of the data including the regression line and the confidence bands. Why are the confidence bands wider at the ends and narrower in the middle? **(1.5 marks)**

My plot of the data including the regression line and confidence bands is found below.



The confidence bands are wider at the ends and narrower in the middle because the 95% confidence intervals must envelop the true best-fit linear regression line. Since both the slope and intercept of the regression line were calculated form sampled data (i.e. xi and $\bar{x}$), the best estimate of the population would occur if the mean went through both ($\bar{x}$, $\bar{y}$). One must picture trying to fit straight lines through ($\bar{x}$, $\bar{y}$). By only changing the slopes of these straight lines, a fan pattern emerges. From a mathematical viewpoint, the calculation of the confidence bands considers the square root of errors in the form of (x - $\bar{x}$). Such functions would appear hyperbolic if plotted which is exactly what is observed by the confidence intervals. Therefore, the confidence bands are wider at the ends and fan out at the ends.

4. Since the District of Columbia is not represented by any of the states, and they were only able to obtain data on the smoking habits (13.4 %), the CDC wants to try to predict the dietary habits of people in this region. Calculate the predicted value for the response variable (show your calculations). Calculate a 95% prediction interval for this point estimate. (**1 mark**)

Here, I will use my equation: $\hat{y_i} = 32.307 - (0.5671 * x_i)$ with x_h = 13.4.

$\hat{y}_{(new)} | \; x_h = 32.307 - (0.5671 * 13.4) =$ 24.708. Hence, the predicted response variable, when 13.4% of the population smokes every day, is 24.7%. This means 24.7% of people eat at least 5 servings of fruits and vegetables every day.

$$s_{\hat{y}_{(new)}|x_h} = \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_h - \overline{x})^2}{SSx}\right)}$$

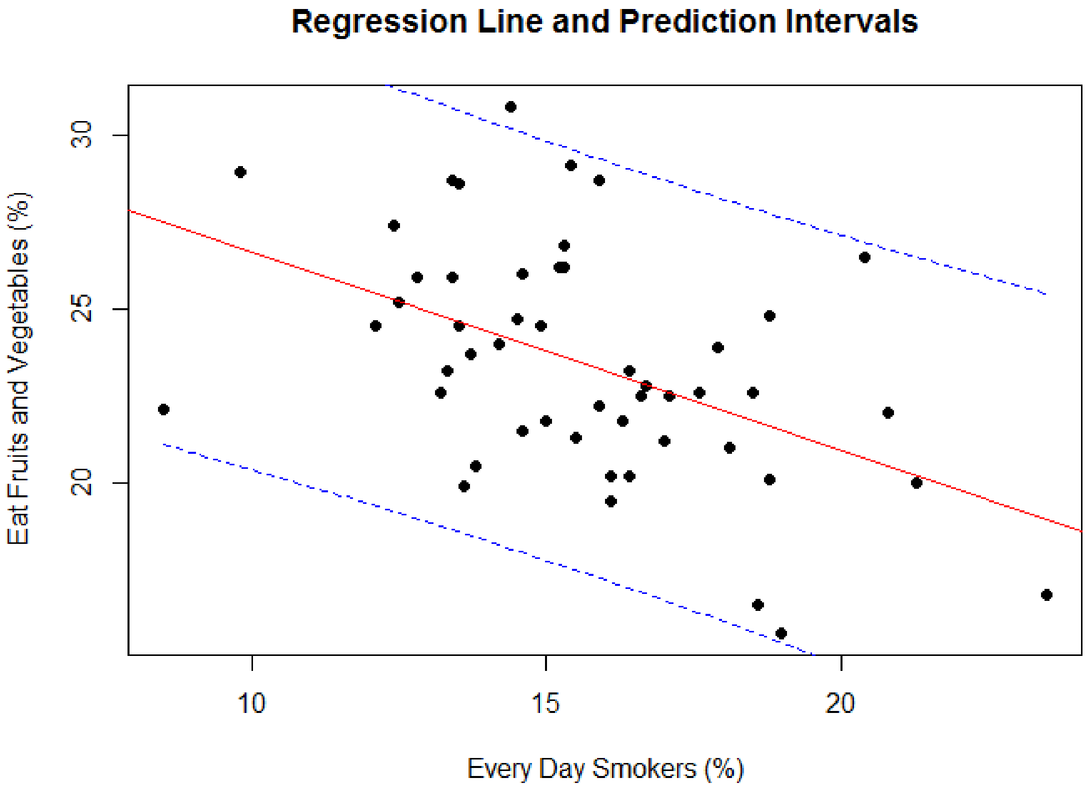
$$= \sqrt{(8.756953416)(1 + \frac{1}{50} + \frac{(13.4-(15.644))^{\,2}}{390.9832})} = 2.996929066$$

$$\hat{y}_{(new)} \mid x_h \pm \; (t_{1-\frac{\alpha}{2},n-2}) * s_{\hat{y}_{(new)}|x_h} = 24.70786 \pm 2.0106*(2.996929066)$$

$$= [18.68223442 \, , 30.73348558]$$

$\hat{y}_{(new)} \mid x_h$ [18.682, 30.733]

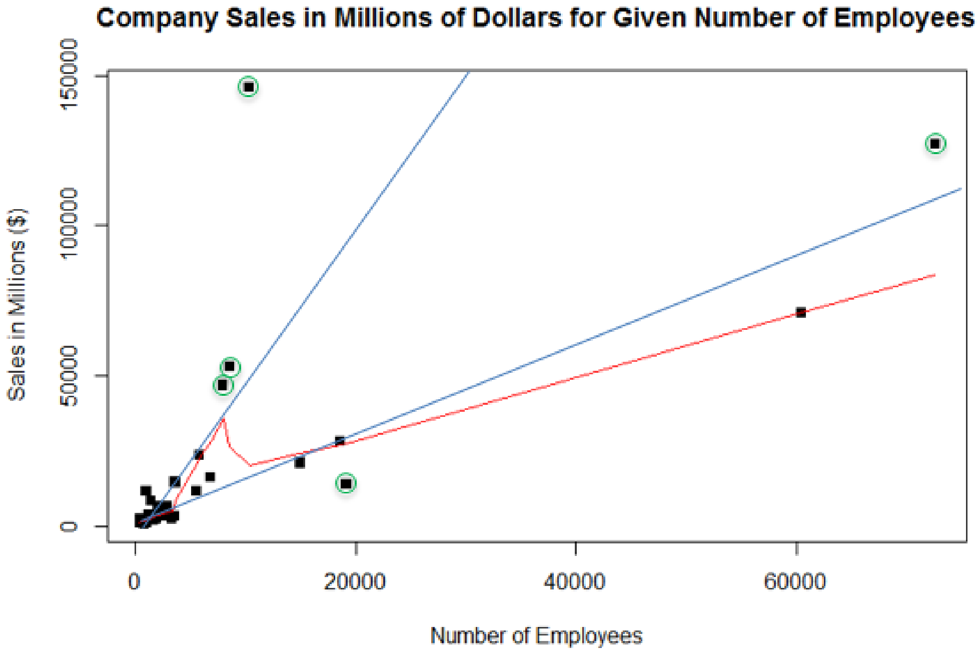
5. Create a plot of the data including the regression line and lines for the prediction intervals. (**1 mark**)



Background: Sales and number of employees (Subtotal: 15 marks)
You are trying to determine the relationship between sales and number of employees based on information from different companies.

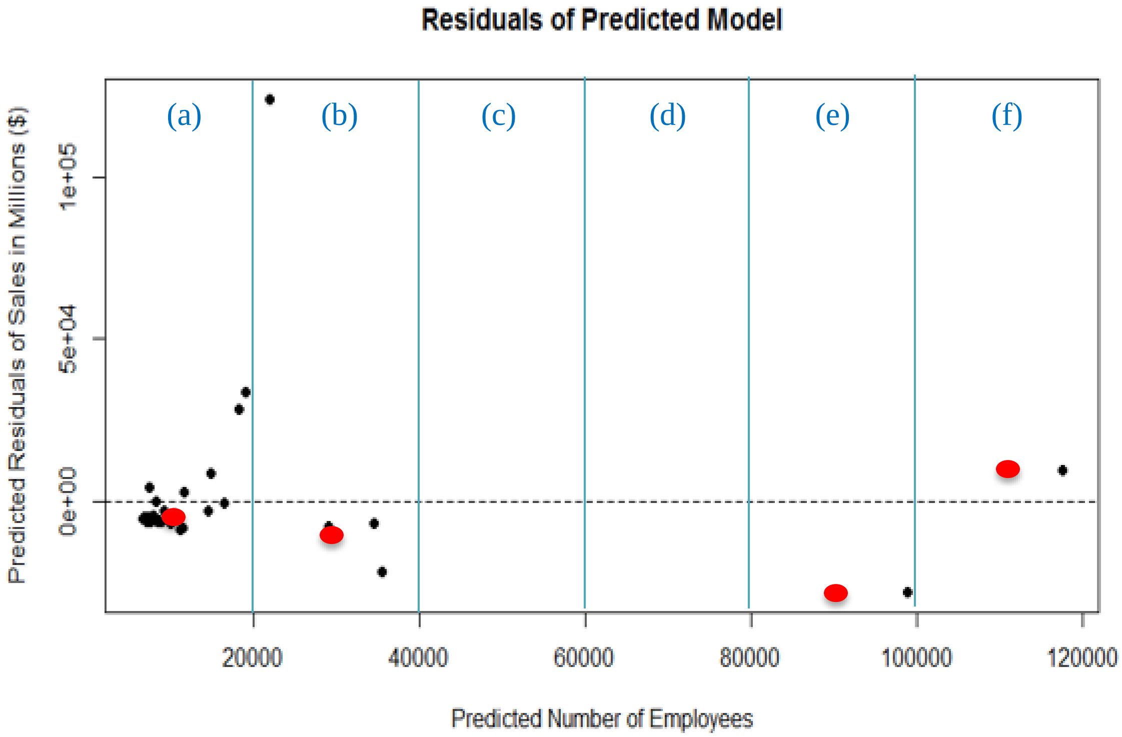
1. Graph the relationship between sales and number of employees. Does the relationship look linear? Are there any outliers? Which companies do these outlying points represent? (**1.5 marks**)

My graph of the sales in millions of dollars (i.e. response variable) against the number of employees (i.e. explanatory variable) in a company is plotted below. Upon initial inspection, while the graph doesn't appear linear, there appears to be a correlation between the response and explanatory variable. Perhaps the graph is linear but contains outliers. This makes sense because only a company that is selling to make large profits can pay to employ more employees or the employees can produce more products, or services, to yield a larger income for the company. The relationship appears positive, strong (i.e. clustered) for companies with less than 10,000 employees, and outliers do exist. For example, assuming a linear relationship was supposed to be present with no outliers in the data, a linear line could be drawn in 2 directions as illustrated in the figure with blue lines. In order to assist me determine any outliers, I added a smoothed red curve through the data. When my delta parameter equals 0.1, as shown in the graph below, there does appears to be 5 outlines which correspond to the following companies: (1) Costco Wholesale; (2) Starbucks; (3) Amazon.com; (4) Nordstrom; and (5) Weyerhaeuser.



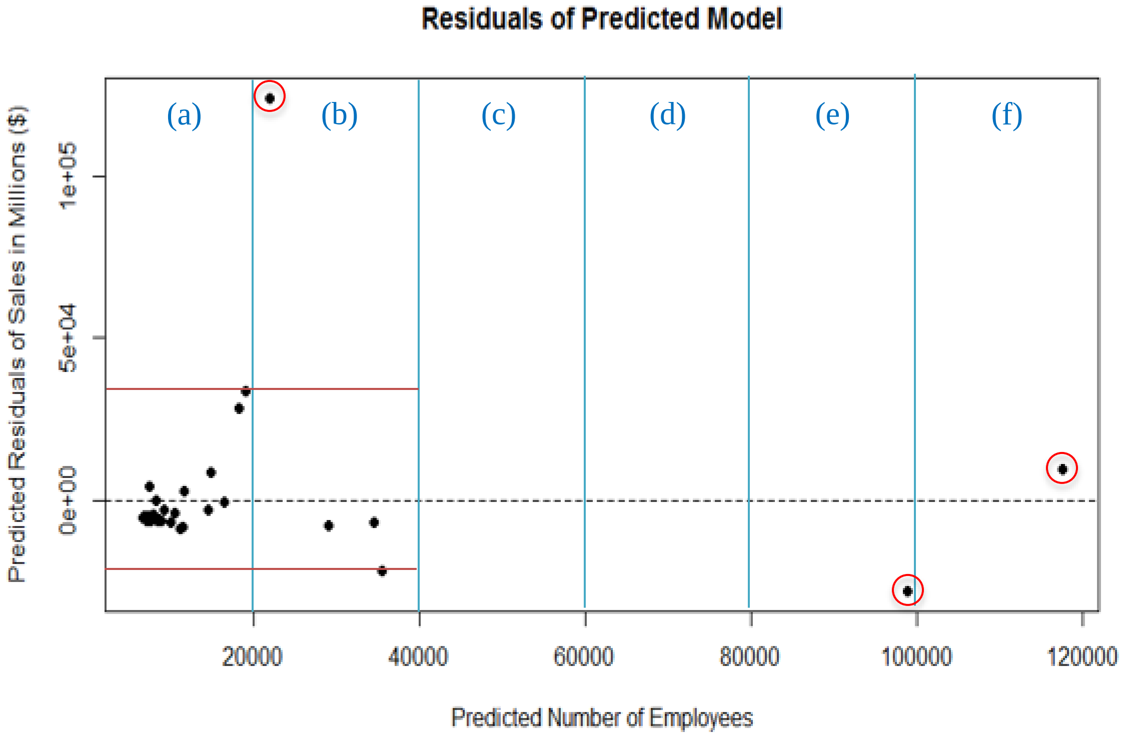
2. Use a residual plot to help you assess the assumptions of linearity and equal variance. (**1 mark**)

The assumptions of linearity and equal variance are required to be met in order for one to fit a linear regression line into the data well. To test these assumptions, I plotted the residuals of my linear model against my explanatory variable, the number of employees, below. Next, I divided my plot of the residuals into 6 segments labelled a, b, c, d, e and f. I made the following graph below to illustrate my visual test for the assumption of linearity.



I am trying to detect whether enough points appear evenly spread above and below the line of zero residuals. First, I predicted the mean values of each segment and plotted a red dot of that mean in the segments with data points. Segments c, d, e, and f do not have enough data points to predict whether the assumption of linearity is met. In contrast, sections a and b have several points. However, there does not appear to be an even spread of points above and below the line of zero residuals. If I were to draw a line connecting the red points, the line would deviate well below the line of zero residuals. Hence, I conclude that the assumption of linearity has not been met. This means the regression line would not fit into my data well and the estimates of my coefficients and standard errors would be biased.

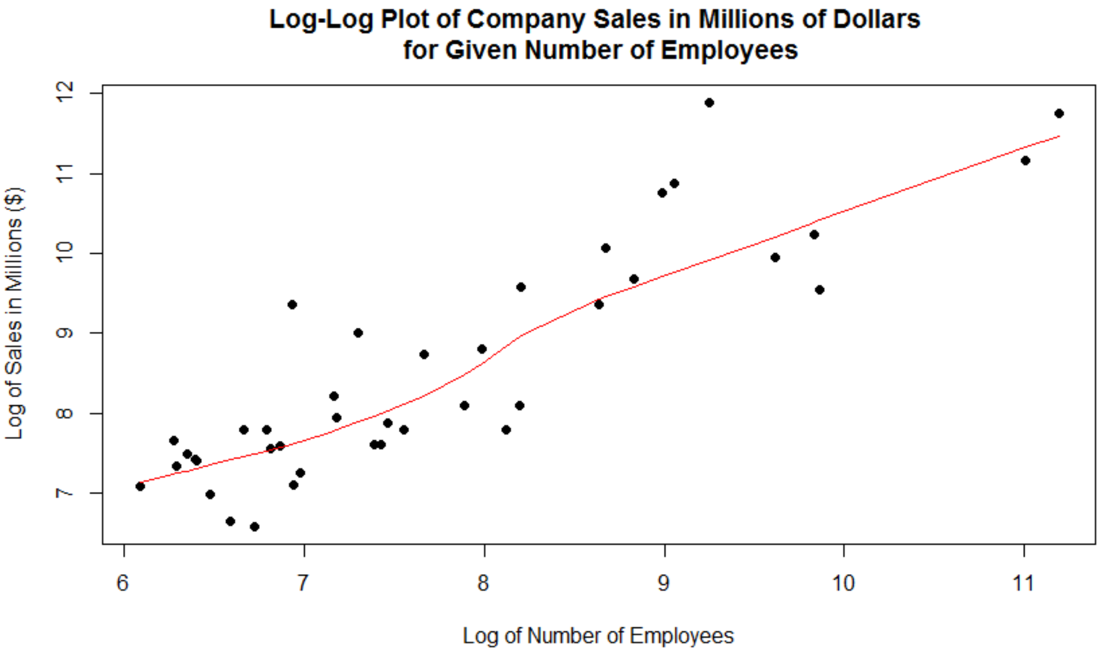
While I would normally not continue my analysis since the test of linearity has failed, I am asked to also test the assumption of equal variance. I used the following plot below to test this assumption.



The assumption of equal variance, or the measure of spread, can be verified if the residuals fall above and below zero residuals evenly. Once again, segments c, d, e and f do not have enough data points to predict whether the assumption of equal variance is met. On the other hand, segments a and b have some data points. I drew 2 straight red lines in the graph to illustrate approximate equal variance. In addition, I have circled the points which do not contribute to determining whether the assumption of equal variance is met. Given my analysis above, I concluded the assumption of equal variance cannot be met without additional data. At present, I cannot calculate the confidence intervals (CI's) or test the significance of the explanatory variable. In addition, while my regression co-efficients will be unbiased, the estimates of standard errors of co-efficients would be biased.

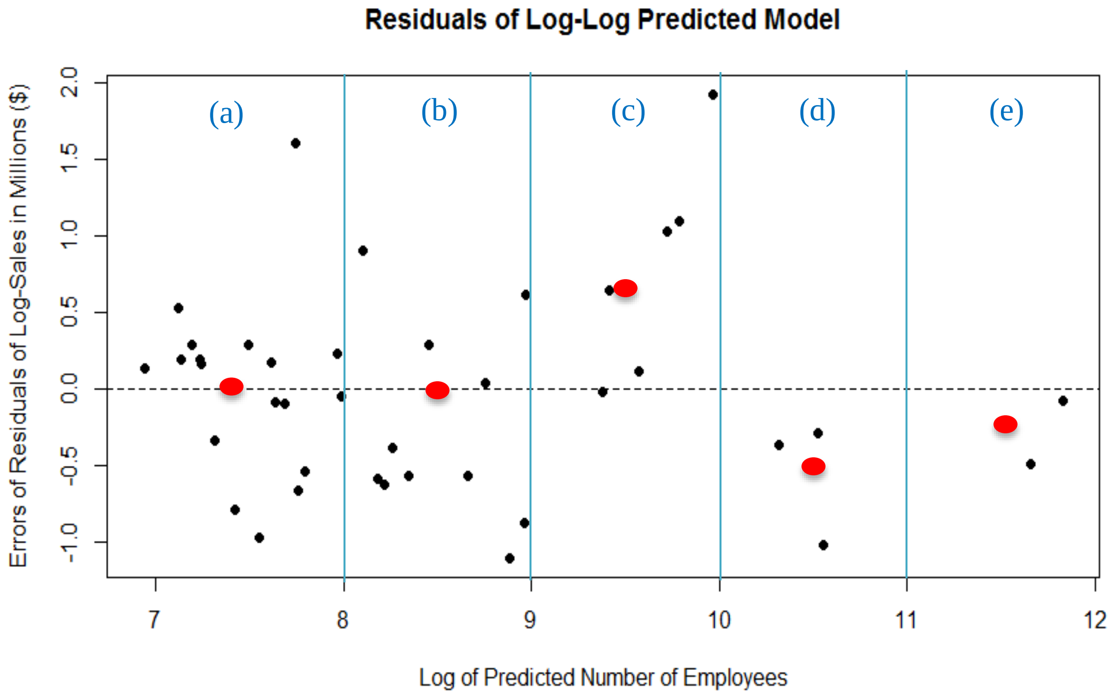
3. Try transforming the x and y variable using a natural logarithm. Graph the relationship between these new variables. Does the relationship look linear? (**1 mark**)

My graph of the log of sales in millions of dollars (i.e. response variable) against the log of the number of employees (i.e. explanatory variable) in a company is plotted below. As expected, one would assume that a company which sells more to make large profits can pay to retain more employees or the employees can produce more products, or services, to yield a larger income for the company. The log-log relationship appears positive, strong for all log values of explanatory variable, and no clear outliers exist. In order to assist me determine any outliers, I added a smoothed red curve through the data. When my delta parameter equals 0.1, as shown in the graph below, there does not appear to be any outliers now. Thus, I conclude that my log-log relationship does appear linear.



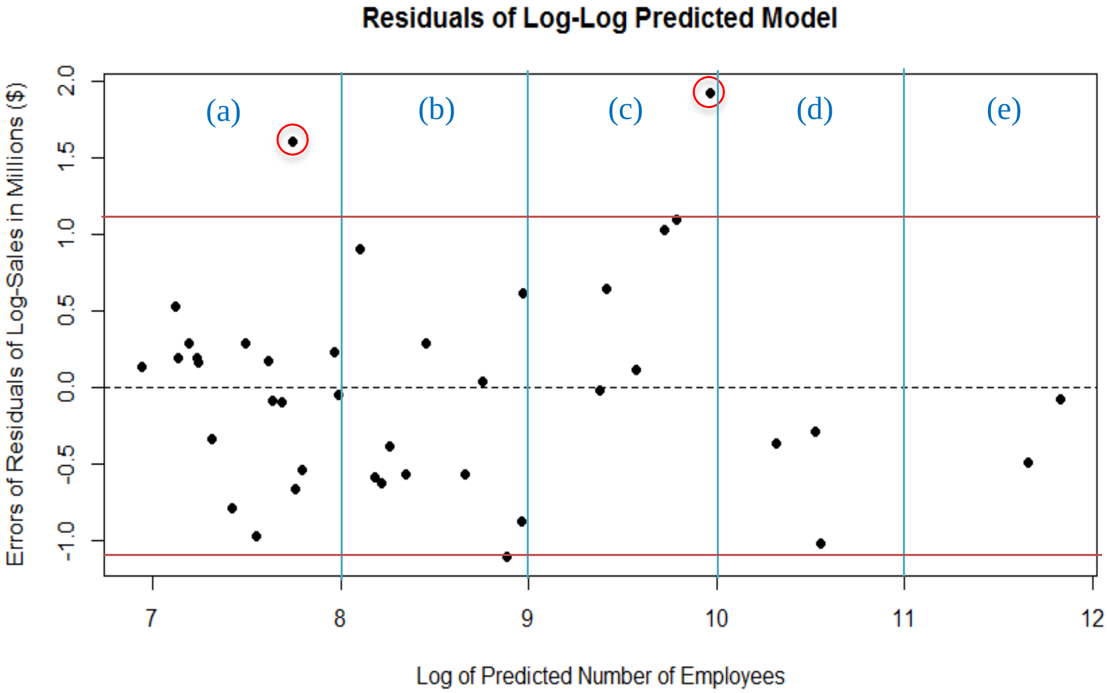
4. Use a residual plot to help you assess the assumptions of linearity and equal variance for this new model. State any concerns you have and their consequences. **(1 mark)**

The assumptions of linearity and equal variance are required to be met in order for one to fit a linear regression line into the log-log data well. To test these assumptions, I plotted the residuals of my log-log linear model against my explanatory variable, the log of the number of employees, below. Next, I divided my plot of the residuals into 6 segments labelled a, b, c, d, and e. I made the following graph below to illustrate my visual test for the assumption of linearity.



I am trying to detect whether enough points appear evenly spread above and below the line of zero residuals. First, I predicted the mean values of each segment and plotted a red dot of that mean in the segments with data points. While it would be nice for segments c, d, and e to have more data points, I still tried to predict whether the assumption of linearity is met. If I were to draw a line connecting the red points, the line would not deviate far from the line of zero residuals. Hence, I conclude that the assumption of linearity has been met. This means the regression line would likely fit into my data well and the estimates of my coefficients and standard errors would not be biased.

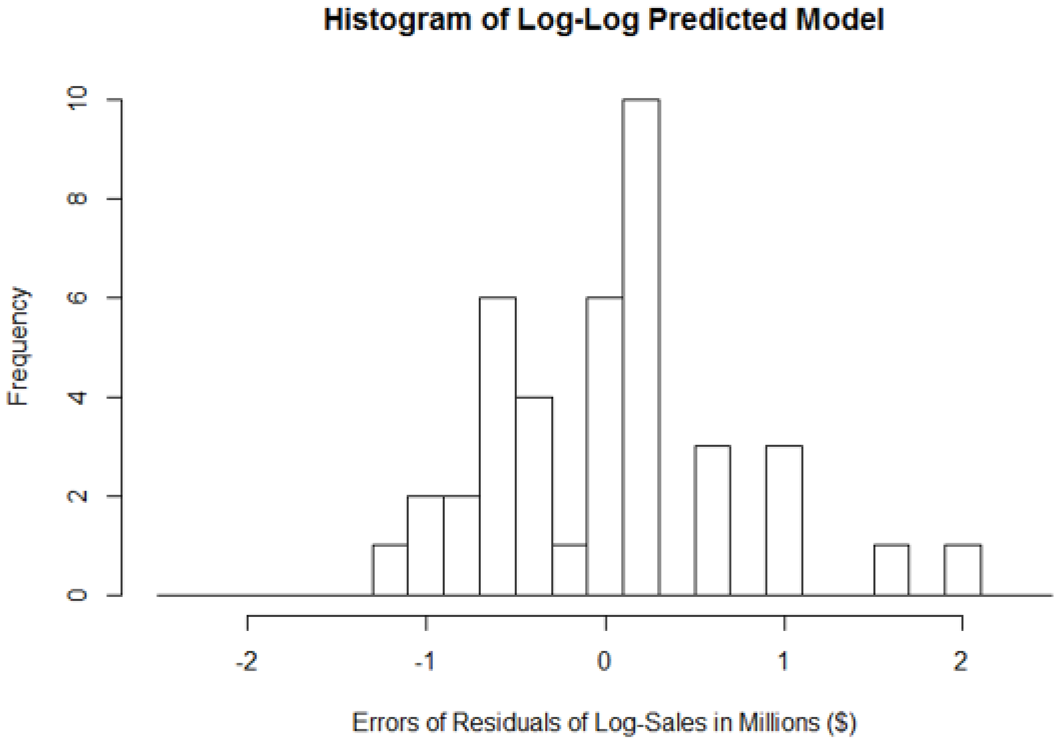
Next, I used the following plot below to test this assumption of equal variance.



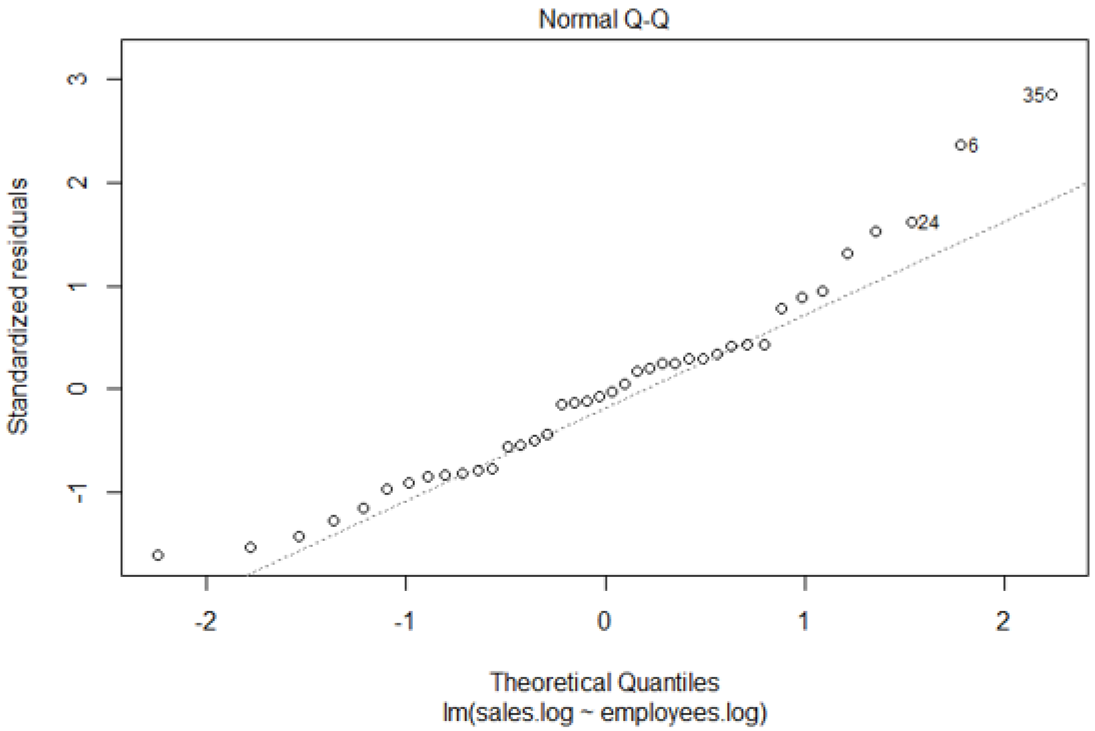
The assumption of equal variance, or the measure of spread, can be verified if the residuals fall above and below zero residuals evenly. Once again, it would be nice for segments c, d, and e to have more data points to predict whether the assumption of equal variance is met there. I have circled the points which do not contribute to determining whether the assumption of equal variance is met. Given my analysis above, I concluded the assumption of equal variance has not failed and can be possibly met with more confidence with additional data. At present, it appears I can possibly calculate the confidence intervals (CI's) or test the significance of the explanatory variable. In addition, while my regression co-efficients will likely be unbiased, the estimates of standard errors of co-efficients might be biased.

5. Check the other assumptions: normality (histogram, normality plot, normality tests), independence of observations, and assumptions related to sampling. State any concerns you have and their consequences. (**2 marks**)

In order to fulfill the assumption of normality of errors, the errors must be normally distributed. I plotted a histogram of the residual errors from my log-log model, as seen below, and it appears approximately normally distributed upon visual assessment. In particular, I noticed the tails of the histogram are either missing or appear to be very small.



Next, I plotted the Q-Q plot of my log-log data as seen below. Since the standardized residuals change linearly by the theoretical quantiles (i.e. lm(sales.log ~ employees.log)), there is further evidence that the residuals errors are normally distributed for each of the explanatory variable's log values. In particular, the Q-Q plot exhibits a right-skew.



Furthermore, I performed four normality tests whose results are summarized below. My hypothesis is:

H0: Errors of log-log model are normally distributed.

H1: Errors of log-log model are not normally distributed.

Test	**Statistic**	**p Value**		**Accept or Reject H0**
Shapiro-Wilk normality test	W = 0.95419	p < W	p = 0.1057	Accept H0
Lilliefors (Kolmogorov-Smirnov) normality test	D = 0.13459	p > D	p = 0.06569	Accept H0
Cramer-von Mises normality test	W = 0.072813	p > W	p = 0.2501	Accept H0
Anderson-Darling normality test	A = 0.49627	p < A	p = 0.2014	Accept H0

In my testing, I am using an alpha value of 0.05. Given that 100% of the tests accept the null hypothesis (i.e. $p > 0.05$), more evidence is given that the errors of the log-log data are normally distributed. Therefore, I will continue with my model.

Most importantly, statistical tests do not always work. If it were the case that the tests only passed because I had data that mimicked a normal distribution, then it would mean I could not calculate the CI's or test the significance of the explanatory variable, log of the number of employees, because I wouldn't know what probabilities to use. Hence, then the estimated coefficients would then no longer equal to the maximum likelihood solutions.

The assumption of independence of observations cannot be verified because we are not provided any information on when or where the data was taken. It is possible that there are repeated observations on the same individual. For example, companies across several industries may downscale or lay off employees. Such employees might get rehired at another company – possibly within a company provided in the data set. Therefore, if the data for one company was taken at one point in time, one of its former employees could be accounted for in another company's data at another time. Perhaps that employee was exceptional at reducing the total sales in a company. Then, the sales produced by a company might need to be explained by another variable instead of number of employees etc. Unfortunately, one cannot determine whether this assumption is verified from the regular residual plots. The provided dataset only specifies a company's sales in millions of dollars and the number of employees they have. Furthermore, the data set lists companies from a variety of industries. Costco has the highest number of employees because it requires labour to set out all of their products to be sold. On the other hand, Microsoft, which has the second highest number of employees, obtains its profits from selling its products which do not need to be continually restocked in the same fashion as Costco's stores. Hence, the nature of industries requires a different number of employees regardless of their sales success.

If by chance the same individual worked in 2 more of the companies listed in the data set, the assumption of independence of observations would not be met. This would mean the coefficients describing the model will be unbiased but estimates of the standard errors of coefficients will be biased. Failing this assumption would mean one cannot calculate the CI's or test the significance of the explanatory variable. This would occur because the outcome in any one trial has an effect on the error term for any other trial.

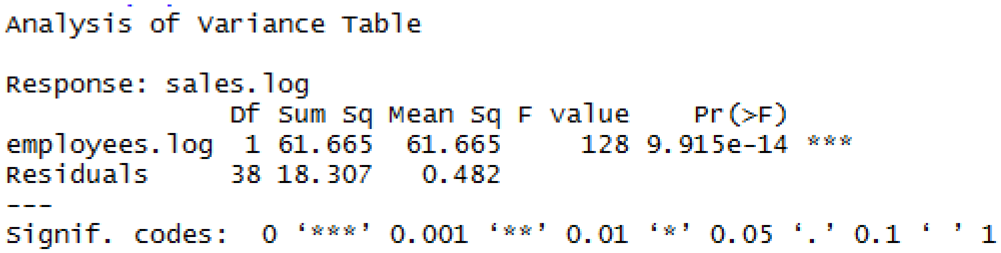
6. Test the significance of the regression. (**1 mark**)

Step 1: My hypothesis is:

H0: Regression is not significant. $B_1 \neq 0$

H1: Regression is significant. $B_1 \neq 0$

Step 2: In R, I used the anova ( ) command and determined my p-value to be $9.915 * 10^{-14}$, and F-value to be 128, as shown below:



```
Analysis of Variance Table

Response: sales.log
              Df Sum Sq Mean Sq F value    Pr(>F)    
employees.log  1 61.665  61.665     128 9.915e-14 ***
Residuals     38 18.307   0.482                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Step 3: $F_{critical} = F_{1,38,0.95} = 4.10$. Hence:

(F-value = 128) > ($F_{critical}$ = 4.10)

(p-value = $9.915 * 10^{-14}$) < (alpha = 0.05)

Step 4: I reject my null hypothesis. Therefore, my regression is significant.

Since my null hypothesis was that the errors are normally distributed, and p < 0.05 (i.e. alpha = 0.05), I would reject my null hypothesis. This means the regression is significant.

7. Calculate the predicted values (in log-transformed units), then back-transform these. Use these back-transformed values to calculate the errors in the original units. Calculate the SSE, SSY and SSR in original units (show your calculations, you are welcome to do this in R or Excel) (**1.5 marks**)

First, I calculated the predicted values (in log-transformed units) using the following command in R: predict2 <- predict(z2)

Next, I back-transformed the predicted values (in log-transformed units) to original units in R by: yhat.original <- exp(predict2)

My list of back-transformed $\widehat{yi}$ values are shown below.

	yhat.original
1	7854.857773
2	38340.66781
3	3716.033222
4	1241.518433
5	2073.230404
6	2316.54483
7	2950.218194
8	137123.0175
9	3303.157246
10	6356.717961
11	11865.01258
12	1498.790954
13	1385.939821
14	2431.287734
15	2890.147191
16	2177.65324
17	2027.473499
18	4209.280037
19	4689.494662
20	12282.32746
21	115179.1678
22	1896.485921
23	37307.0765
24	17739.03315
25	2346.877998
26	1031.323498
27	30264.85393
28	3582.18672
29	3860.265935
30	14311.99273
31	7227.152726
32	1399.236735
33	7811.711219
34	5795.758949
35	21312.15558
36	1800.177142
37	1328.254466
38	1257.105486
39	1675.074804
40	16637.32477

Then, using the back-transformed values, I calculated the errors in original units as follows:

- $SSE = \sum_{i=1}^{n}(y_i - \widehat{y_i})^2$ therefore, in R I used the command:
  SSE <- sum((mydata.2$sales - mydata.2$yhat.original)^2)
  and calculated my SSE = 20813830042 = 2.081 * 10^{10}

- $SSY = \sum_{i=1}^{n}(y_i - \bar{y})^2$ therefore, in R I used the command:
  SSY <- sum((mydata.2$sales - mean(mydata.2$sales))^2)
  and calculated my SSY = 39721185007 = 3.972 * 10^{10}

- $SSR = \sum_{i=1}^{n}(\bar{y} - \hat{y}_i)^2$ therefore, in R I used the command:

  SSR <- sum((mean(mydata.2$sales) - mydata.2$yhat.original)^2)

  and calculated my SSR = 30679867116 = 3.068 * 10^{10}

8. Using these values, calculate the pseudo-r^2 (or I^2) value and the standard error of the estimate (SE_E') in the original units. (**1 mark**)

I calculated my pseudo-r^2 value as shown below.

$$pseudo - r^2 = I^2 = 1 - \frac{SSE}{SSY} = 1 - \frac{20813830042}{39721185007}$$

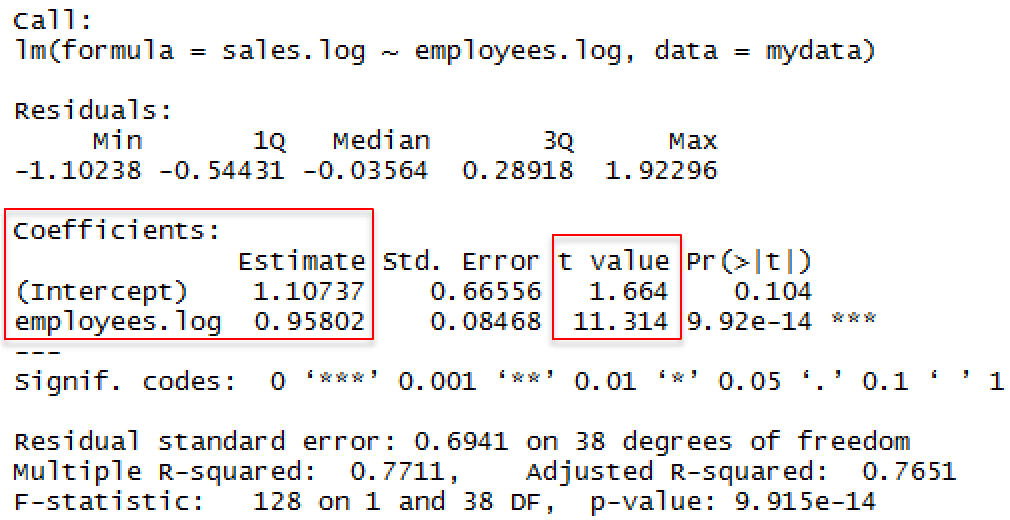
$$pseudo - r^2 = 0.47600178498 \qquad \text{(original units)}$$

The standard error of the estimate (SE_E) is a measure of the variation not accounted for in the model. I calculated SE_E', as shown below:

$$SE_E' = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{20813830042}{40-2}} = 23403.683 \qquad \text{(original units)}$$

9. State the estimates of the co-efficients (b_0, b_1) and calculate their 95% confidence intervals. Back-transform these values into their original units. (**2 marks**)

In R, I used the summary( ) command to verify my Bo and B1 coefficients in log-transformed units as illustrated below:



```
Call:
lm(formula = sales.log ~ employees.log, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.10238 -0.54431 -0.03564  0.28918  1.92296 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    1.10737    0.66556   1.664    0.104    
employees.log  0.95802    0.08468  11.314 9.92e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6941 on 38 degrees of freedom
Multiple R-squared:  0.7711,	Adjusted R-squared:  0.7651 
F-statistic:   128 on 1 and 38 DF,  p-value: 9.915e-14
```

My estimates of the Bo and B1 coefficients are 1.10737 and 0.95802, respectively, in log-transformed units. I obtained each value, in original units, by doing the following calculations: (1) Bo.original = $e^{1.10737}$ = 3.026; (2) B1.original = $e^{0.95802}$ = 2.607.

I calculated the 95% confidence intervals below. In my Excel spreadsheet, I calculated the following pieces of information in log-transformed units: (1) MSE = 0.481759982 ; (2) SSx = 67.18852567; (3) $\bar{x}$ = 7.752363287; (4) $t_{1-\frac{\alpha}{2},n-2}$ = 2.024.

$$s_{b_1} = \sqrt{\frac{MSE}{SSx}} = \sqrt{\frac{0.481759982}{67.18852567}} = 0.08467745666$$

$$s_{b_0} = \sqrt{MSE(\frac{1}{n} + \frac{\bar{x}^2}{SSx})} = \sqrt{(0.481759982)(\frac{1}{40} + \frac{7.752363287^2}{67.18852567})} = 0.665560767$$

$$For\ B_0 = b_0 \pm (t_{1-\frac{\alpha}{2},n-2}) * s_{b_0}$$

$$For\ B_1 = b_1 \pm (t_{1-\frac{\alpha}{2},n-2}) * s_{b_1}$$

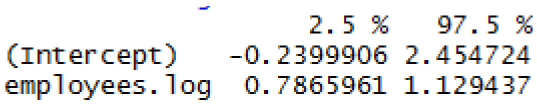
Hence, the 95% confidence intervals are as follows:

b_0	[-0.2397249, 2.4544649]	(Log-transformed units)
b_1	[0.7866328, 1.1294071]	(Log-transformed units)

The 95% confidence intervals were calculated, in original units, by back-transforming with the exponential function as follows:

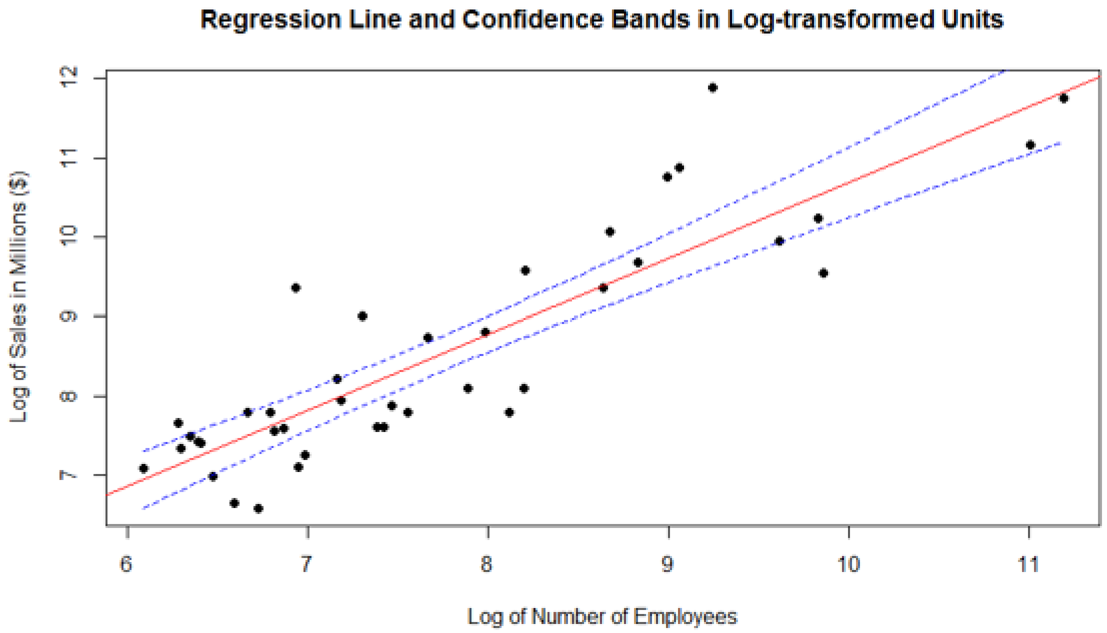
b_0	[0.786844219, 11.64020430]	(Original units)
b_1	[2.195989686, 3.093821852]	(Original units)

In addition, I verified my results in R using the command confint(z2, level=0.95). My result was as follows:

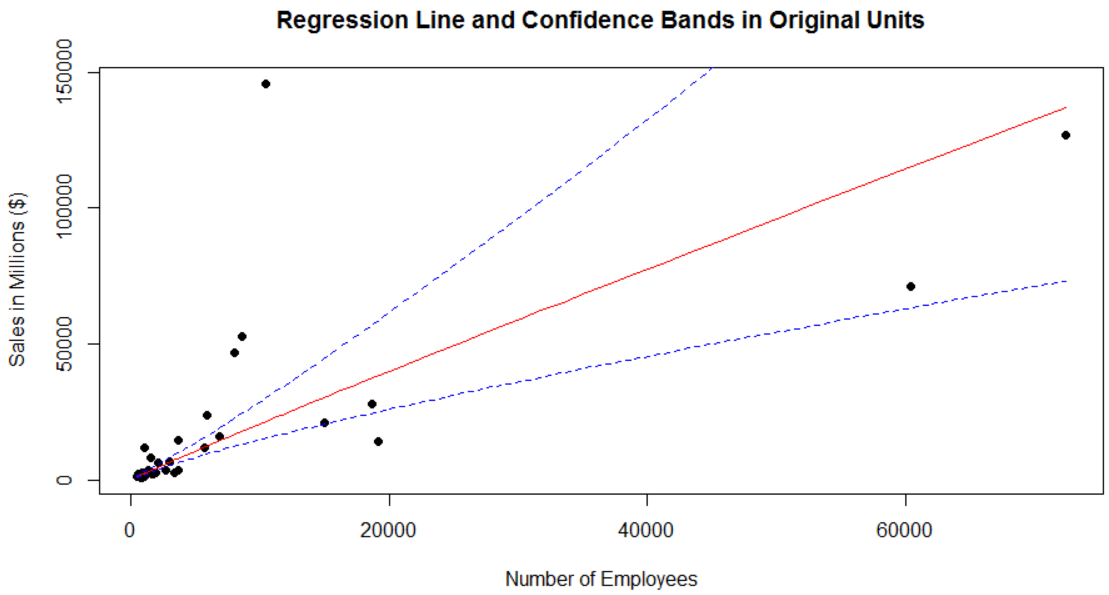


```
                  2.5 %   97.5 %
(Intercept)   -0.2399906 2.454724
employees.log  0.7865961 1.129437
```

10. Create a plot of the data including the regression line and the confidence bands in the log units. (**1 mark**)



11. Create a plot of the data including the regression line and the confidence bands in the original units. (**2 marks**)


Regression Line and Confidence Bands in Original Units
Sales in Millions ($)
Number of Employees