COMM 581 - Assignment #2
Simple Linear Regression – Part 1

Name: ____Gurpal Bisra_____ **Total: 20 marks**
Due date: Monday Sept. 19, 2016 (11pm)

Background:
The Centers for Disease Prevention and Control (CDC) have collected data about health conditions and risk behaviors. They are trying to develop a profile of the states where people eat the recommended quantities of fruits and vegetables so that a national health initiative can develop a plan to improve health.

Explanatory variable: % of people who smoke every day
This explanatory variable is being used as a proxy for how concerned people are with their health. Information about this variable is easy to obtain because people who have health insurance have to answer this question, and these responses can be provided anonymously to the CDC.

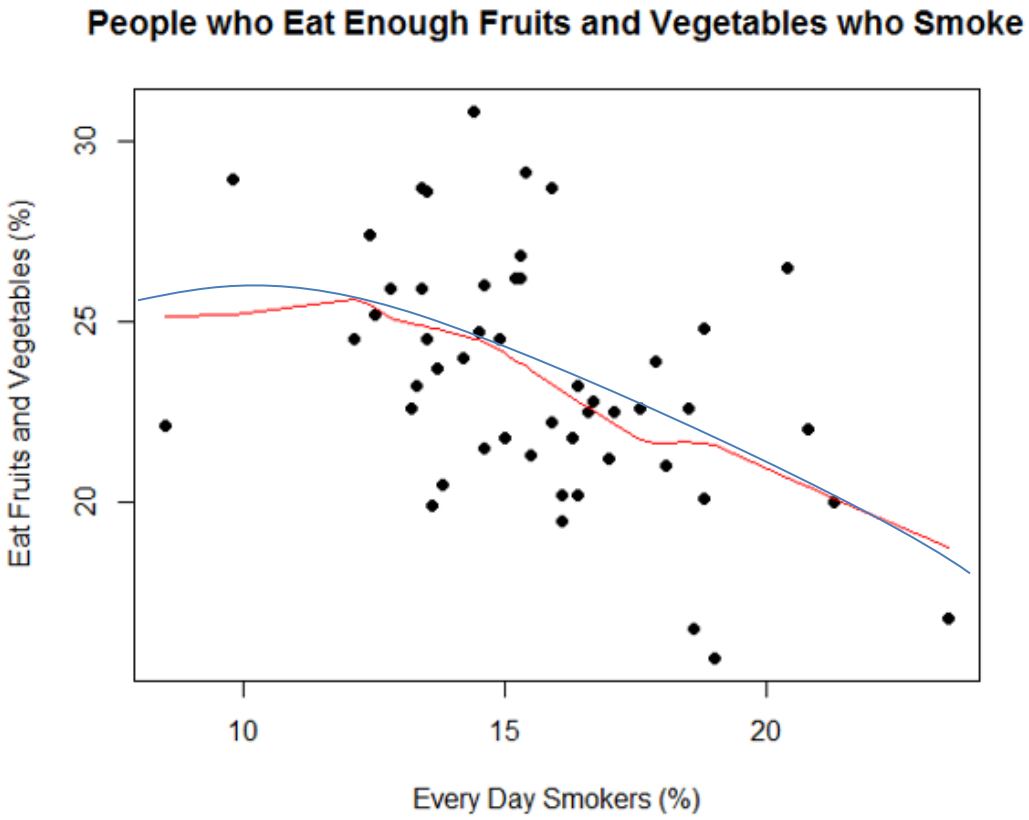
Response variable: % of people who eat at least 5 servings of fruits and vegetables every day
Information about this variable is difficult to get and can only be obtained from a survey specifically asking about health-related habits.

Questions (include all graphs mentioned in the questions in your final assignment):
Assessing assumptions

1. Make a scatterplot of % of people who eat enough fruits and vegetables vs. % of people who smoke every day. Fit a smoothing curve to the plot. **Describe the relationship shown on the graph – is this the type of association you would expect? Does the relationship look linear? (2 marks)**

My graph of the percentage of people who eat 5 servings of fruits and vegetables every day (i.e. response variable) plotted against the percentage of people who smoke every day (i.e. explanatory variable) is shown below. Upon initial inspection, the relationship between the response and explanatory variables appear to be a decreasing function and does not seem linear. A decreasing association was expected. For instance, eating enough fruits and vegetables is generally considered a healthy lifestyle choice while smoking every day is not. It seems logical to assume people who smoke every day would select other non-healthy lifestyle choices such as not eating enough fruits and vegetables.

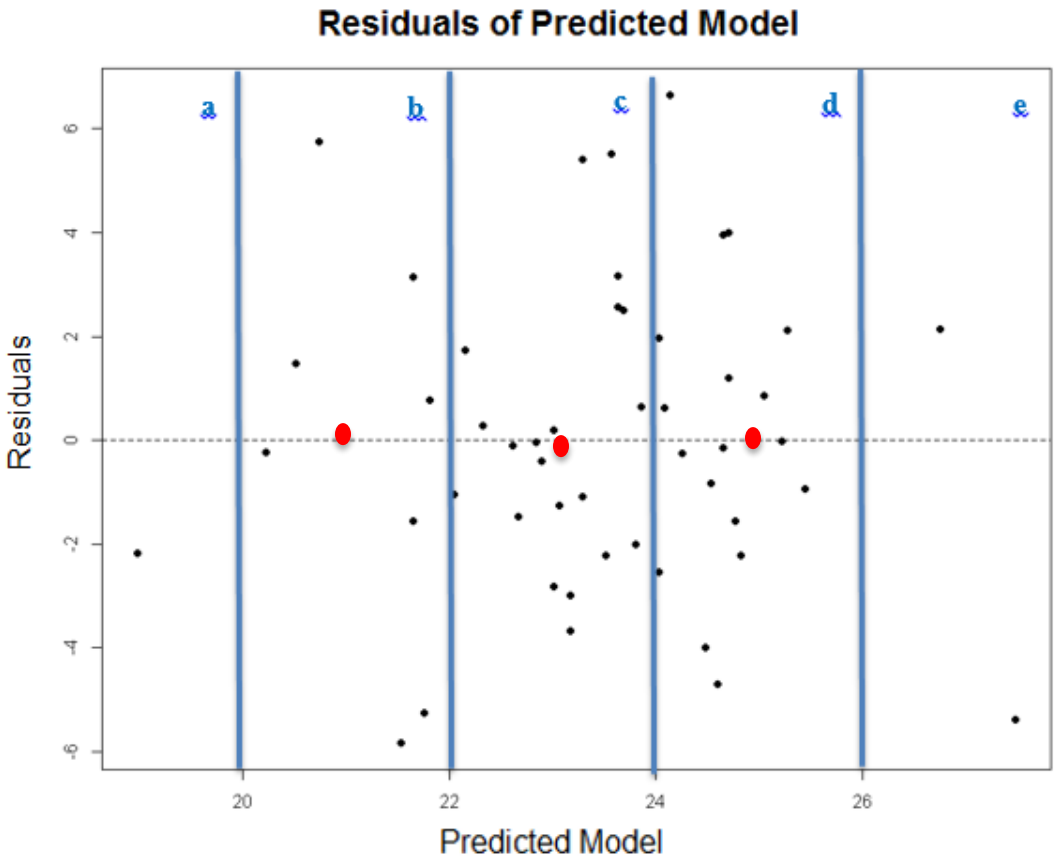
Next, I added a smoothed curve through the data. When my delta parameter equals 0.1, as shown in the graph below, the fit does not seem linear. When comparing the two lines, I feel the data could show a linear relationship for the percentage of every day smokers between 12 – 22%.



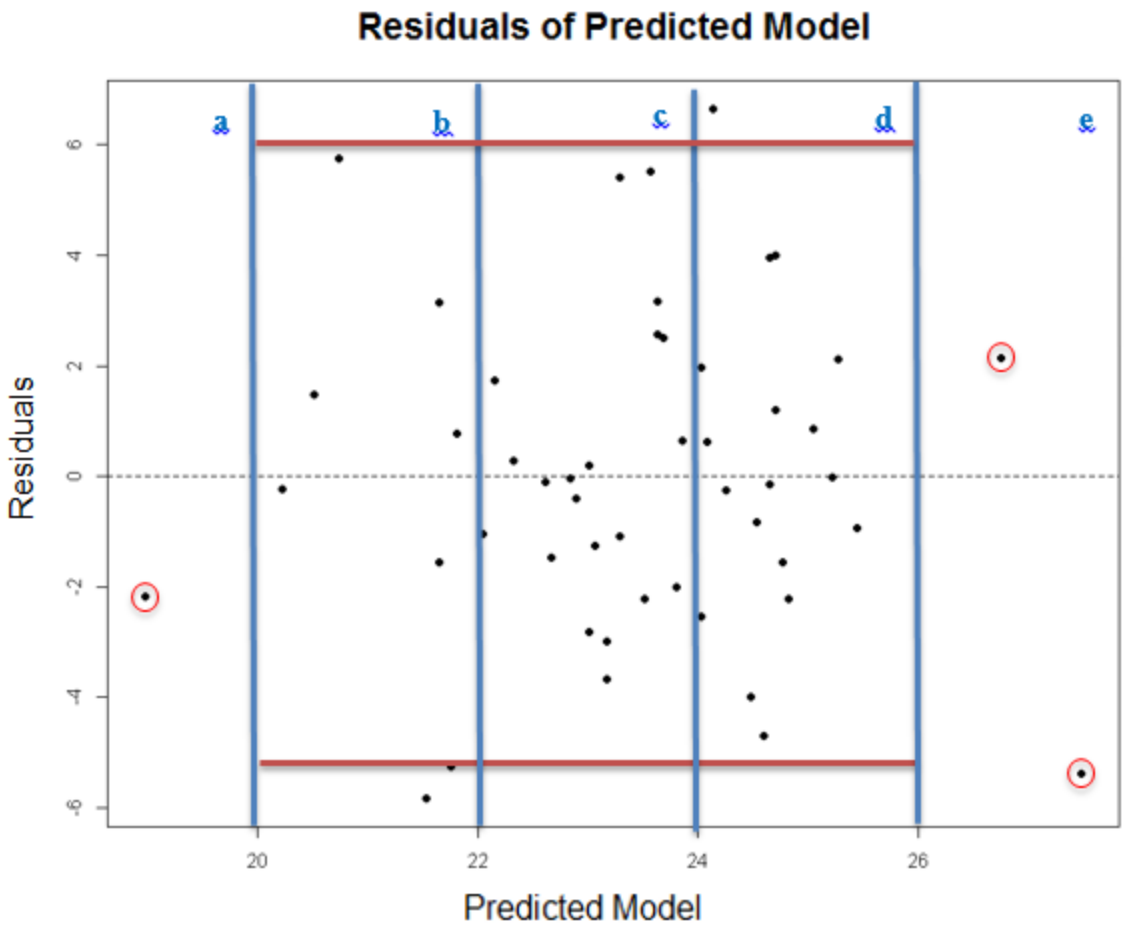
2. Create a linear model to fit this data. Extract the residuals and predicted values. Create the **residual plot** and examine it to see if the **assumptions of linearity and equal variance** are met. Describe how well you think these assumptions are met; is it sufficient to continue with the model? State any concerns you have and their possible consequences. **(2 marks)**

I plotted the residuals of my linear model against the predicted y-values below. 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 this assumption, I divided my plot of the residuals into 5 segments labelled a, b, c, d, and e. Segments a and e do not have enough data points to predict whether the assumption of linearity is met. Fortunately, sections b, c, and d have enough points which appear approximately evenly spread above and below the line of zero residuals. Using visual inspection, I added red points at the means of each of the sections which have enough points. Since the red dots appear linear at the line, formed by residuals equaling 0, I conclude assumption of linearity is met (i.e. has not failed). I will continue with my model.

If I received more data points and segments a and e observed a lack of linearity, I would be concerned that the regression line wouldn't fit the data well and then the estimates of my coefficients and standard errors would be biased. In that instance, I could not continue my analysis.



I will continue and test the 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. For instance, in sections a and e below, it would be difficult to draw a line for comparing equal variance because there are not enough points. I have circled the points which do not contribute to determining whether the assumption of equal variance is met. On the other hand, for sections b, c, and d the assumption of equal variance is met. Hence, I drew a straight red lines in the graph below to illustrate approximate equal variance.



Suppose I receive more points in segments a and e. If the assumption of equal variance was not met given the additional data, then that would mean I could not 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.

Thus, based on the graphs above, the feel the assumptions of linearity and equal variance are met in sections b, c, and d. I will continue with the model. If more points were given for sections a and e, then I could determine the assumptions for them.

3. Is there any reason to think that the **assumption of independence of observations** is violated by this data? How would you determine if this assumption was violated? State any concerns you have and their possible consequences. **(2 marks)**

The assumption of independence of observations means there is no repeated observation on the same individual. For example, this means the same individual was not sampled at the same time, or location etc. Unfortunately, one cannot determine whether this assumption is verified from the regular residual plots alone. Instead, one must plot the residuals against what they think is causing the dependency.

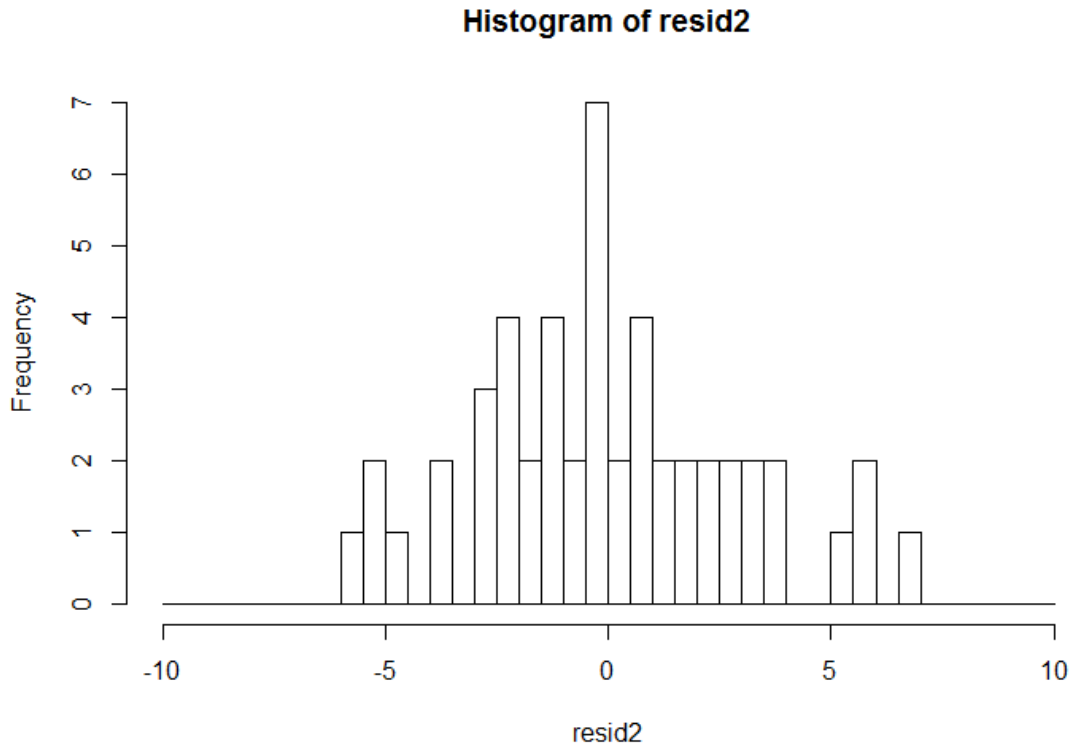
The provided dataset only specifies the percent of people who smoke per day, and the percent of people who eat at least 5 servings of fruits and vegetables per day, for each American state. Since we have no further information, for time or location per se, we cannot plot residuals against anything we believe may be causing the dependency. I would be interested in knowing more about the data methodology and how the study was conducted. For example, the time of year would likely have a profound effect on the data collected. In the past decade, there have been several anti-smoking campaigns while promoting healthy eating. If some states collected data 10-years-ago, and others collected data within the past year, there could be great variation in the data.

If by chance the same individual took the survey, 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.

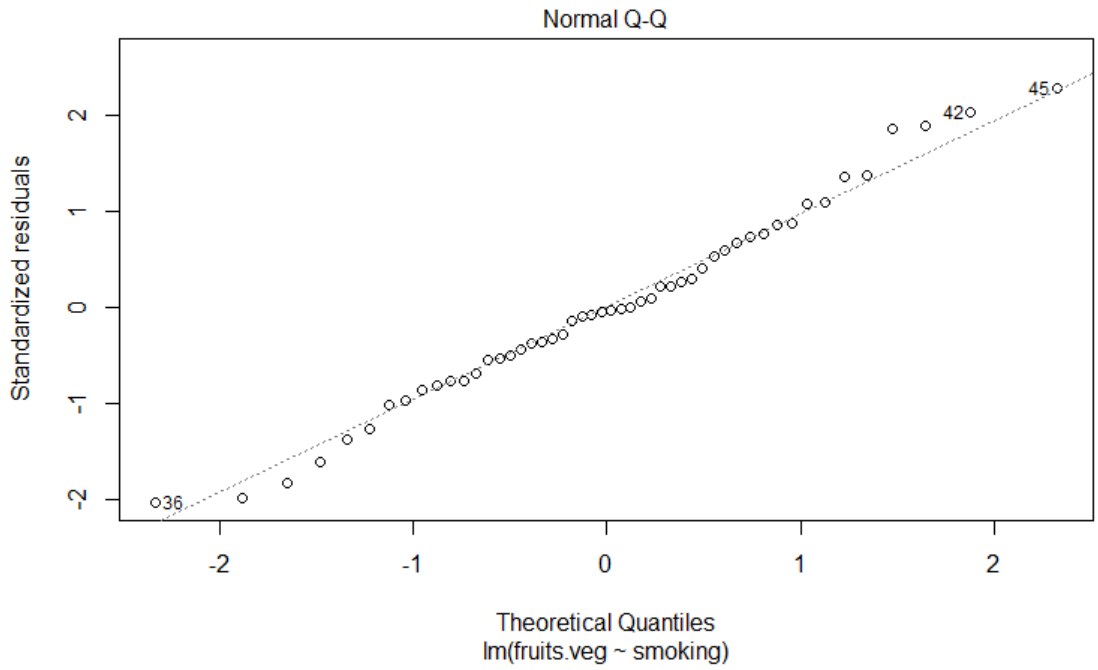
To summarize, while there are reasons to suspect the assumption of independence of observations could be violated, we are not provided any additional required information to test them.

4. Use a **histogram of the residuals, normality tests and the normality plot (Q-Q plot)** to determine if the **assumption of normality of errors** is met. Describe how well you think this assumption was met; is it sufficient to continue with the model? State any concerns you have and their possible consequences. **(4 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, 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 as seen below. Since the standardized residuals change linearly by the theoretical quantiles (i.e. lm(fruits.veg ~ smoking)), there is further evidence that the residuals errors are normally distributed for each of the x values. In particular, the Q-Q plot exhibits light tails.



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

H0: Errors are normally distributed

H1: Errors are not normally distributed.

Test	Statistic	p Value		Accept or Reject H0
Shapiro-Wilk normality test	W = 0.98485	p < W	p = 0.7651	Accept H0
Lilliefors (Kolmogorov-Smirnov) normality test	D = 0.065657	p > D	p = 0.8524	Accept H0
Cramer-von Mises normality test	W = 0.029276	p > W	p = 0.8532	Accept H0
Anderson-Darling normality test	A = 0.2061	p > A	p = 0.8626	Accept H0

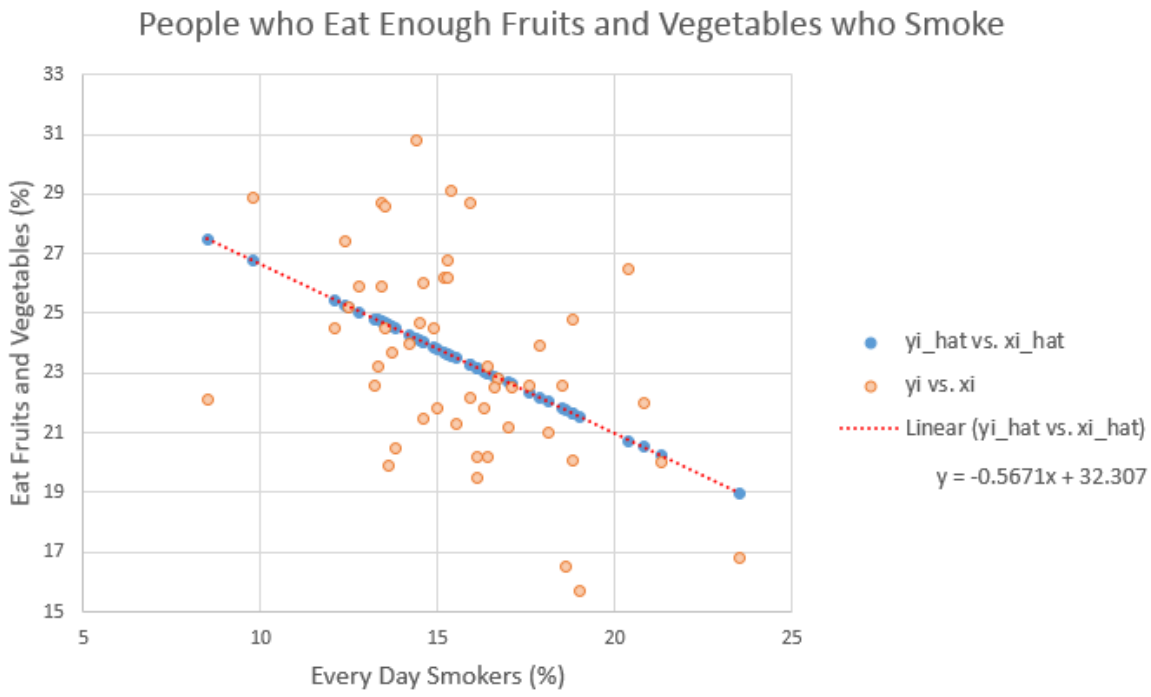
In my testing, I am using an alpha value of 0.05. Given that 100% of the tests accept the null hypothesis (i.e. $p > 0.05$), more evidence is given that the errors are normally distributed. Thus, I will continue with my model.

If it were the case that the majority of the tests failed (i.e. $p < 0.05$, or more data indicated heavy-tails), that would mean that I could not calculate the CI's or test the significance of the explanatory variable, smoking, because I wouldn't know what probabilities to use. Hence, the estimated coefficients would then no longer equal to the maximum likelihood solutions.

Creating and assessing the model

5. Using Excel, where each column represents a different step in your calculations, calculate the slope and intercept for your linear model. Label each column with a clear description of its contents. **(5 marks)**

Using Excel, I calculated my linear regression model to have a slope of -0.5671 (i.e. B1) and intercept (i.e. Bo) of 32.307, respectively. My plot of the percent of people who eat at least 5 fruits and vegetables versus the percent of people who smoke every day is shown below. In particular, I added the original data points and fitted linear regression curve, formed after I performed my calculations in Excel, to my graph.



6. Check your results for the coefficients using R.

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
```

7. Is the regression significant? Remember to state all 4 steps of your hypothesis test! **(1 mark)**

In R, I used the anova ( ) command and determined my p-value to be 0.0004218 as shown below:

```
Analysis of Variance Table

Response: fruits.veg
          Df Sum Sq Mean Sq F value    Pr(>F)
smoking    1 125.72 125.721  14.357 0.0004218 ***
Residuals 48 420.33   8.757
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

My hypothesis is:

H0: Errors are normally distributed

H1: Errors are not normally distributed.

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. Moreover, the $F_{critical}$ or F statistic for 48 degrees of freedom (DOF) in the denominator, and 1 DOF in the numerator, for the 0.05 significance level is 4.04. Since my calculated F value, of 14.357, is greater than 4.04, then I have further evidence the regression is significant.

8. Using Excel, calculate the co-efficient of determination (r^2) and the standard error of the estimate (SE_E). **(2 marks)**

Using Excel, I calculated that my co-efficient of determination (r^2) to be 0.230 and standard error the estimate (SE_E) to be 2.959.

9. Check your results for these measures of goodness of fit using R. Do these measures indicate a good model? **(1 mark)**

In R, I used the summary( ) command and verified measures of goodness of fit were correct as illustrated below:

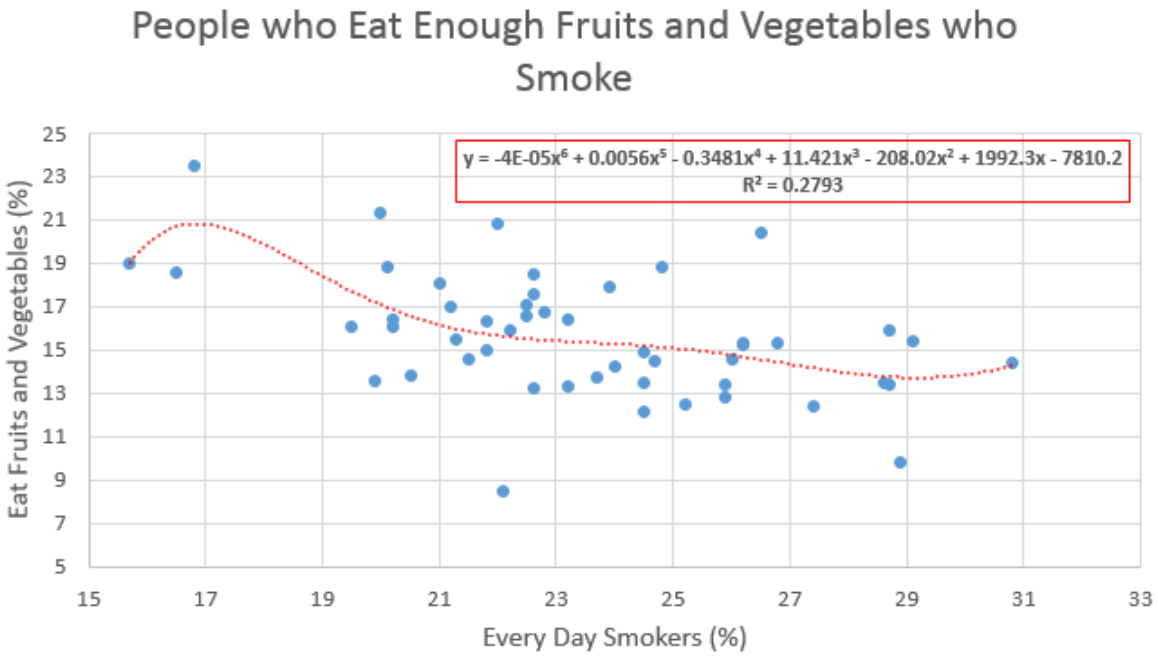
```
Call:
lm(formula = fruits.veg ~ smoking, data = mydata)

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
```

An R-squared value is proportional to the total variation described by the regression. While the interpretation of the R-squared value depends on the field one works in, and the cost of making a mistake, a relatively low value of 0.2302 may suggests my measures do not indicate a good model. Given that the CDC is trying to predict whether people eat at least 5 fruits and vegetables per day given they smoke every day, I personally feel a 20% error in a model does not seem bad. On the other hand, I cannot state whether an R-squared of 0.230 would allow the CDC to make a decision regarding a healthcare initiative.

The standard error of the estimate (SE_E) is a measure of the variation not accounted for in the model. Hence, it indicates the percentage of deviation of error. While a 2.96% error seems low, it accounts for a 10% error in our model since our explanatory and response variables only range ~ 20%. Again, the SE_E suggests my measures do not indicate a good model. I require more context, regarding what is acceptable in the healthcare field, before I can say anything with more certainty.

10. Overall, what do you think could be done to improve your model? **(1 mark)**

Overall, I believe the model could be better approximated by using a different model or by getting more data points. As stated in question 1 from visual inspection, I believe the fitting function is not linear. In fact, I fitted a polynomial function which had higher R-squared value indicating a better fit than the linear model. However, this model, as shown below is too complex and not straightforward to use for predicting information.



Since removing points, which I may suspect to be outliers which bias the data, I will instead think about how to change the study parameters to get more data points. For example, it may be the case that certain states have higher prices of food. If an every day smoker lived in a "food desert," then they would eat less fresh fruits and vegetables because it may cost too much. In addition, there may be certain states which promote smoking. Such advertisement could have a big impact on the percentage of every day smokers. Moreover, one can simply just collect more data. This method has the tradeoff that it will be costly and require more resources (i.e. time) to obtain. Alternatively, a different explanatory variable like the percentage of people who sleep 8 hours per night, or exercise regularly, might yield a better linear relationship between the explanatory and response variables. Thus, a better model might be achieved if the aforementioned considerations are factored into the study design.