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

Name: ____________________________ **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):
<u>Assessing assumptions</u>

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)**

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)**

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)**

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)**

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)**

6. Check your results for the co-efficients using R.

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

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

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

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