Name: _____   **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**)

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**)

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**)

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**)

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**)

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

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**)

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**)

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**)

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

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**)

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**)

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**)

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**)