Stat 311 Homework 3

This assignment will use the plasma retinol data set (PlasmaRetinolData.csv) and focuses on simple linear regression. Be sure to check the data dictionary (PlasmaRetinolDataDescription.pdf) to know what the variables stand for and to get their corresponding units for axis labels and interpretations in the context of the problem.

1. Some basic EDA

   a) Read in the data and use the ggplot2 package to make histograms of BetaDiet, ln (BetaDiet), BetaPlasma, and ln (BetaPlasma). Make sure your histograms are on one figure, so they are easy to compare. Compare the log transformed distribution to the unlogged distribution for each variable.

   b) You should see the following warning, "**Warning: Removed 1 rows containing non-finite values (stat_bin).**" in the output on the R Console page after rendering the histograms. You are getting this error because there are one or more zeros in BetaDiet or BetaPlasma; ln (0) is undefined. I have included the code in the template to pull out these observations. How many observations are impacted by applying a natural log transformation?

   c) Create a subset of data that removes any points with zero in x or y. Use this subset for the remaining problems.

2. Making scatterplots. Use the ggplot2 function to create the four scatterplots given below Add a regression line to each scatterplot [code for first scatterplot provided in the template]. Arrange the scatterplots in one figure.

   a) Make a scatterplot of BetaPlasma (y) on BetaDiet (x) and calculate the correlation coefficient, r. Report the correlation and describe the joint relationship between the two variables.

   b) Make a scatterplot of ln (BetaPlasma (y)) on BetaDiet (x) and calculate the correlation coefficient, r. Report the correlation and describe the joint relationship between the two variables.

   c) Make a scatterplot of BetaPlasma (y) on ln (BetaDiet (x)) and calculate the correlation coefficient, r. Report the correlation and describe the joint relationship between the two variables.

   d) Make a scatterplot of ln (BetaPlasma) on ln (BetaDiet) and calculate the correlation coefficient, r. Report the correlation and describe the joint relationship between the two variables.

   e) Based on your scatterplots and the histograms from Problem 1, do you recommend trying a natural log transformation for either variable or both to better meet assumptions for simple linear regression modeling? Briefly explain.

3. Fit a linear regression for ln (BetaPlasma) on ln (BetaDiet). Write out the regression equation.

4. Interpret the estimated slope of the regression line in the context of the problem. Remember that the interpretation is a different when using ln-ln transformations.

5. What is the expected value for ln (BetaPlasma) when BetaDiet is 2100?

6. What is the expected value for BetaPlasma when BetaDiet is 2100? To back transform to the original units of y, use the equation $\hat{y} = e^{a+b\ln(x)}$.

7. Row 42 in the data set has a BetaDiet value of 2100. What is the residual based on the fitted regression line?

8. Do you think that BetaDiet is useful for estimating BetaPlasma? Explain.