**Question 2.**

The data here are numbers of visits to Hawaii from the other 49 states in the year 2000 along with population, income and distance to Hawaii, that might help “explain” the variation in visits.

In the data step we compute the natural logarithm of visits: LVISITS; Dist: Ldist. Pop: Lpop, and Income: Lincome.

1. Run the procedure PROC CORR on the log transformed data

2. Which explanatory variable is most highly correlated with LVISITS?
   a. Using only that variable, what is the predicting equation of LVISITS?
   b. List the R-square for comparison to the next model.

3. Regress the log transformed visits on similarly transformed income, population, and distance.
   a. Write out the fitted equation.
   b. What proportion of variation is explained by the multiple regression?
   c. How much improvement in R-square did you get (versus the simple linear model)?
   d. Test the residuals for normality. Use alpha=0.05.

Our model contains the variable population. You would expect a state with a large population to produce more visits to Hawaii than a small state. Perhaps we should have looked at the proportion of people visiting Hawaii, that is, log(visits/pop).

Notice that log(visits/pop) = log(visits) - log(pop), as is true for the log of any ratio. So if

\[
\log(\text{visits/pop}) = \_\_ + \_\_ \log(\text{income}) + \_\_ \log(\text{distance})
\]

this becomes theoretically

\[
\log(\text{visits}) = \_\_ + \_\_ \log(\text{income}) + \_\_ \log(\text{distance}) + 1 \log(\text{pop}).
\]

In other words the coefficient on log(pop) would be a 1 in the regression in part 3. The estimate is close to 1, but of course not exactly 1.

4. Compute a t test for testing the null hypothesis that this coefficient is 1. Write down the null hypothesis, the t test and its degrees of freedom, and your decision. **[hint: The estimate and standard error can be taken right from your printout and EVERY t test has the form \( t = (\text{estimate-hypothesized value}) / (\text{standard error}) \).]**

5. Switching to GLM, add all the two way interactions into your model. Note how ALL the statistics change!
   a. Test the null hypothesis that the coefficients on all the interaction terms are simultaneously 0 (that is, test to see if you can leave them all out). As a reminder, you can difference either the model or error sums of squares from the full and reduced model, divide this by 3 df (3 interaction terms) and then divide that mean square by the MSE from the full model to get your F test.
   b. State the null and alternative hypotheses in terms of your model parameters, compute F and its degrees of freedom, and state your conclusion.
   c. State your conclusion.