DSO 529 Homework 1

due Tuesday, February 7 in class

1. A marketing research firm wishes to compare the prices charged by two supermarket chains – Miller's and Albert's. The research firm, using a standardized one-week shopping plan (grocery list), makes purchases at 10 of each chain's stores. The stores are randomly selected and all purchases are made in a single week. The shopping expenses obtained at the two chains are as follows:

Miller's:	$119.25	$121.32	$122.34	$120.14	$122.19
	$123.71	$121.72	$122.42	$123.63	$122.44
Albert's:	$111.99	$114.88	$115.11	$117.02	$116.89
	$116.62	$115.38	$114.40	$113.91	$111.87

Because the stores in each sample are different stores in different chains, it is reasonable to assume that the samples are independent and that the variances are similar at the two stores. We also assume the weekly expenses are normally distributed.

a. Letting μ_M be the mean weekly expense for the shopping plan at Miller's and letting μ_A be the mean weekly expense for the shopping plan at Albert's, test $H_0 : \mu_M - \mu_A = 0$ versus $H_a : \mu_M - \mu_A \neq 0$ at the 0.05 level of significance.

b. Create a 95% confidence interval for the differences between the population means of the two stores.

c. Set up the null and alternative hypotheses needed to attempt to establish that the mean weekly expense for the shopping plan at Miller's exceeds the mean weekly expense at Albert's by more than $5 and test the hypotheses at $\alpha = 0.05$. What is your conclusion (in real words, not jut "reject" or "do not reject"!)?

2. **LA Real Estate.** (dataset: *LARealEstate.xlsx*, found on Blackboard) The data are semi-recent properties for sale in Los Angeles and you are to determine how *List Price* is related to the other variables measured in the dataset.

List Price: Price the property is currently listed for
Home Type: Options are *Condo/Coop* or *Single Family Residential*
Beds: Number of bedrooms
Baths: Number of bathrooms
Location: Options are *Santa Monica, Beverly Hills*, and *Downtown Los Angeles*
Sqft: Square footage of the living space
Lot Size: Square footage of the lot the property is on
Year Built: Year the structure was built
Parking Spots: Number of parking spots
Days on Market: Number of days the property has been listed for sale

a) First account for the categorical nature of the *Home Type* and *Location* variables by creating separate independent dummy variables appropriately. Now create a correlation matrix and decide which variable would be the best predictor of *List Price.* (**JMP**: *Analyze* → *Multivariate Methods* → *Multivariate*, enter all variables in the *Y* section)

b) Run a regression using that variable. (**JMP**: *Analyze* → *Fit Model*, enter the appropriate *Y* and *X* variables) What is your regression equation?

c) What is the standard error of the estimate? How does this compare to the standard deviation of the *Y* variable by itself? Why is that?

d) Now create a regression model using *all* the available variables. Be sure to include the Variance Inflation Factors in the output. (**JMP**: *Analyze* → *Fit Model* and get your output. Hover over the *Parameter Estimates* section and right click to the *Columns* option, then choose VIF.) Are there any issues with multicollinearity? If so, what would you do to address this?

e) Create a *Residual by Predicted Plot* and assess it for potential heteroscedasticity and non-linearity. (**JMP**: *Analyze* → *Fit Model* and get your output. Click on the red triangle, select *Row Diagnostics,* and choose *Plot Residual by Predicted*.) Do you notice any issues? What step(s) should be taken?

f) After adjusting for any issues above in part e, create a "good" regression model using the available variables, making sure to account for multicollinearity and a 5% significance level. Include a short table with your choices and reasons (like the ones I included in the slides for our Lab Session on January 26th). What is your final model?

g) Just humor me here… Assess any non-linearity between your chosen *Y* variable and your remaining *X* variables. (**JMP**: *Analyze* → *Fit Y by X,* put your *Y* variable in and then insert all your *X* variables in the *X* dialog box together. This will create separate scatterplots between each pair of *Y* and *X*. Now look at potential curvature. ☺) What recommendations, if any, do you have?

h) Incorporate any ideas from part g and continue refining your model by assessing p-values. Include a short table with your choices and reasons (like the ones I included in the slides for our Lab Session on January 26th). What is your final model?

i) Give a brief interpretation of your resulting slope coefficients and their relationships to *List Price.*

3. **Education Expenditures.** Refer to *Education.xlsx* on Blackboard with information for 38 countries.
EDUC: Expenditures on education in millions of U.S. dollars
GDP: Gross Domestic Product in millions of U.S. dollars
POP: Country population in millions of people

a) Estimate a linear model for the data. What are the resulting equation and relevant output values (i.e. *F* statistic, *t* values, and R^2)?

b) Now attempt to estimate a log-linear model (where both the dependent and independent variables are in the natural log format).

c) With the log-linear model, what does the coefficient of the GDP variable indicate about Education? What about the Population variable?

d) Which model is more appropriate? Why?