**Problem Set 8**

**Business Analytics Fall, 2016**

1. **Physicians**. The Physicians Data spreadsheet contains information on the numbers of physicians in the US as well as the total US population for selected years.
2. Draw a scatterplot of the number of physicians versus the year.
3. Fit a linear model ***(# physicians = a*** + ***b*** x ***year)***  to the data. Be sure to include residual output. Do a scatterplot of the residuals versus year. You should copy the year column from the data next to the residuals column in your spreadhseet to make this easy. Do you see any evidence that the linear model does not fit the situation? Explain why or why not.
4. Fit quadratic ***(# physicians = a*** + ***b*** x ***year*** + ***c*** x ***year2 )*** and cubic ***(# physicians = a*** + ***b*** x ***year*** + ***c*** x ***year2*** + ***d*** x ***year3 )*** models to the data. Hint: add columns for both year squared and year cubed to the data so you can use those as additional explanatory variables. Be sure to include residual output in each case.
5. Based solely on the R-squared values, ignoring all other information, which is the best model?
6. If you look at the standard error term alone, ignoring all other information, which is the best model?
7. Do a residual analysis of the three models. Which model appears to have the best behaved residuals and why? Hint: look at a scatterplot of residuals as in part b.
8. Based upon all information, R-squared values, standard error, residuals, and p-values for both intercept and slope(s), which is the best model. Explain your reasoning
9. Run a regression with total MD’s as the y-variable and both year and US population as the explanatory variables. Compare this model to one in which the only explanatory variable is year (that is, the model considered in part b). What are the weaknesses in the two models and which, if you had to choose one of the two, would be preferable? Explain why.
10. **Schools and Poverty** A series of articles in the *Des Moines Register*, Sunday November 26, 2000, describes an analysis of scores of third and fourth grade students on the Iowa Test of Basic Skills (ITBS) versus a measure of the poverty rate in Iowa elementary school districts. The articles contain data for third grade students from school districts in Iowa City, Cedar Rapids, Waterloo, Davenport, and Des Moines, and data for fourth grade students for the above cities plus Sioux City. The articles and data are available in the poverty and schools folder.
11. To start, run a regression of the ITBS scores versus the poverty rate for the fourth grade group. Give the regression output including residuals, and provide a histogram of the residuals. How well does the poverty rate predict the ITBS scores? How would you interpret the intercept and slope coefficient in this regression?
12. Group residuals by city. Is there any systematic misprediction by city? One way of answering this is to see whether or not the average residual for each city is statistically different from zero. Perform T-tests for each group of residuals to answer this question. Another way of looking at this is to see whether or not the average variance in residuals for each city is different from the overall variance across all observations. After computing the sample variance, use chi-square tests of variance to see if the variance within each city is different from the overall sample variance.
13. Try adding in indicator variables for each city. The indicator variable for Iowa City, for example, is equal to 1 for all data entries coming from Iowa City and zero for other cities. Which indicator variables seem to be useful at predicting scores and which do not? How does your answer relate (or does it?) to question b?
14. **Artsy.** Do a regression analysis for Artsy using, sex, grade, and tingrade to predict rate. Do your results agree with those in the case? Are there significant mispredictions by sex? Hint: do a pivot table with sex as the rows and residuals as the values. Do a t-test to see if the average male/female residuals are significantly different from zero.

Do a second regression using an indicator variable for sex as an additional variable. Are there mispredictions by sex this time around? Can you explain your answer?