Regression Analysis

Assignment 1A

Note: "Computer help" refers to the numbered items in the software information files available from the book website.

Problem 1.1 on page 29. This exercise uses the NBASALARY dataset and covers histograms (see page 3), QQ-plots (page 8), and the normal distribution (page 5).

Problem 1.1. The **NBASALARY** data file contains salary information for 214 guards in the National Basketball Association (NBA) for 2009-2010 (obtained from the online USA Today NBA Salaries Database).

a. Construct a histogram of the *Salary* variable, representing 2009-2010 salaries in thousands of dollars [computer help #14].
b. What would we expect the histogram to look like if the data were normal?
c. Construct a QQ-plot of the *Salary* variable [computer help #22].
d. What would we expect the QQ-plot to look like if the data were normal?
e. Compute the natural logarithm of guard salaries (call this variable *Logsal*) [computer help #6], and construct a histogram of this *Logsal* variable [computer help #14].

   Hint: The "natural logarithm" transformation (also known as "log to base-e", or by the symbols $\log_e$ or ln) is a way to transform (rescale) skewed data to make them more symmetric and normal.

f. Construct a QQ-plot of the *Logsal* variable [computer help #22].
g. Based on the plots in parts (a), (c), (e), and (f), say whether salaries or log-salaries more closely follow a normal curve, and justify your response.

[7 points]

Problem 1.5 on page 30. This exercise uses the COUNTRIES dataset and covers sample statistics (page 4), random sampling (page 11), and confidence intervals (page 17).

Problem 1.5. *Gapminder* is a "non-profit venture promoting sustainable global development and achievement of the United Nations Millennium Development Goals." It provides related time series data for all countries in the world at the website www.gapminder.org. For example, the **COUNTRIES** data file contains the 2010 population count (variable *Pop* in millions) of the 55 most populous countries together with 2010 life expectancy at birth (variable *Life* in years).

a. Calculate the sample mean and sample standard deviation of *Pop* [computer help #10].
b. Briefly say why calculating a confidence interval for the population mean *Pop* would *not* be useful for understanding mean population counts for all countries in the world.
c. Consider the variable *Life*, which represents the average number of years a newborn child would live if current mortality patterns were to stay the same. Suppose that for *this* variable, these 55 countries *could* be considered a random sample from the population of all countries in the world. Calculate a 95% confidence interval for the population mean of *Life* [computer help #23].

   Hint: Calculate by hand (using the fact that the sample mean of Life is 69.787, the sample standard deviation is 9.2504, and the 97.5th percentile of the t-distribution with 54 degrees of freedom is approximately 2.005) and check your answer using statistical software.

[6 points]

Problem 1.7 on page 31. This exercise also uses the COUNTRIES dataset and covers hypothesis tests (use the rejection region method at the top of page 21), and prediction intervals (page 27).

Problem 1.7. Consider the **COUNTRIES** data file from Problem 1.5.

a. A journalist speculates that the population mean of *Life* is at least 68 years. Based on the sample of 55 countries, a smart statistics student thinks there is insufficient evidence to conclude this.. Do a hypothesis test to show who is correct based on a significance level of 5% [computer help #24].

   Hint: Make sure you lay out all the steps involved—as on page 21—and include a short sentence summarizing your conclusion; that is, who do you think is correct, the journalist or the student?

b. Calculate a 95% prediction interval for the variable *Life* . Discuss why this interval is so much wider than the confidence interval calculated in Problem 1.5 part (c).

   Hint: Calculate by hand (using the fact that the sample mean of Life is 69.787, the sample standard deviation is 9.2504, and the 97.5th percentile of the t-distribution with 54 degrees of freedom is approximately 2.005) and check your answer using statistical software (if possible—see page 27).

[5 points]