**LAB 1: SIMPLE LINEAR REGRESSION**

The **CDI** data set in Appendix C.2 provides selected county demographic information (CDI) for 440 of the most populous counties in the United States. Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county. Counties with missing data were deleted from the data set. The information generally pertains to the years 1990 and 1992.

The number of active physicians in a CDI *(Y)* is expected to be related to the total population, number of hospital beds, and total personal income. Assume that the normal error linear regression model is appropriate for each of the three predictor variables.



1. Regress the number of active physicians in turn on each of the three predictor variables. State the estimated regression functions. Discuss the good fit of each model.



1. Plot the three estimated regression functions and data on separate graphs. Does a linear regression relation appear to provide a good fit for each of the three predictor variables?
2. Calculate *MSE* for each of the three predictor variables. Which predictor variable leads to the smallest variability around the fitted regression line?
3. Using coefficient of determination as the criterion, which predictor variable accounts for the largest reduction in the variability in the number of active physicians? Discuss.
4. Obtain a separate interval estimate of for each model. Use a 90 percent confidence coefficient in each case.
5. Using various residual plots and test for constant variance to check the adequacy of each model? Summarize your conclusions. Is the standard normal linear regression model (2.1) more appropriate in one case than in the others?
6. Are there any outliers? If Yes, Does the model change much if we omit the outliers?