FINAL PROJECT

Official Due Date: Friday, December 2nd

Basic Description of the Project

The goal of the project is to analyze a real data set, appropriately interpret your findings, and to write up your results in the form of a consulting report. You may choose one of two data sets which are described in more detail below or pick a data set of your own, although if you wish to do the latter you need to check with me to make sure your data set has the appropriate characteristics. Of the data sets which I am providing, the first, SENIC, deals with length of stay and infection rates in U.S. hospitals. The second, CDI, focuses on crime rates in U.S. counties. The two data sets have somewhat different issues but I do not consider one to be significantly harder than the other overall. It should also be noted that there isn't a single "right" answer and there are a myriad of different issues that you could discuss, so use your imagination and think about all the topics we have covered over the course of the quarter. Whatever data set you pick, you will need to analyze one continuous outcome variable using multiple linear regression and one dichotomous outcome using logistic regression. The multiple linear regression is the primary analysis since we will not have spent as much time on logistic regression. Do the MLR first and use it to inform your choices for the logistic regression. There is also an optional component which you may do for extra credit if you wish (but do it only after completing the rest of the project!) Details of what I expect you to include in your analyses and write-up up are given below along with descriptions of the two data sets I am providing. Note that for each data set there are some specific issues which I have listed along with the variable descriptions.

Instructions For Statistical Modeling

Whichever data set you are using, the following approaches and issues may be helpful to consider:

- **Exploratory Analyses:** It is advisable to start with a set of exploratory analyses to examine the distributions of the variables and the relationships between pairs of variables. This will help you identify potential problems including possible outliers, multicollinearity and the need for transformations.
- **Preliminary Models:** Likewise it is a good idea to fit simple linear regressions of your outcome on each of the main predictors and a global model of the outcome on all the predictors together to get a feel for what is going on. In particular notice whether there are any interesting differences between your simple linear regressions and the overall model.
- **Transformations and Deriving New Variables:** Transformations will be very important in this project for several reasons. In addition to using them to get the right shaped relationships or to fix other problems with the regression assumptions you may want to create new variables because they will be more relevant, more interpretable, or will help you to deal with multicollinearity. You should think about this BEFORE you start building your final model.
- **Regression Diagnostics:** Validity of the regression assumptions and handling of outliers will also be important in this project. Make sure you create an appropriate set of diagnostics plots and statistics for your final model as well as describing any intermediate checks you performed to help you identify problems with your preliminary models. Carefully justify your decisions about whether or not to exclude outliers. (Note: Whether something is an outlier depends on what model you are fitting, so you will have to examine this iteratively; you can't make the decision about what to include or exclude right at the beginning nor can you wait until after you have chosen the final model!)

- **Interactions, Confounding, Mediation, etc.:** There are a number of possible interactions in these data sets. Think carefully about how the relationships of the various predictors with Y might depend on one another. Also consider any possible cases of mediation or confounding and how these might affect your interpretation of your models.
- **Model Selection:** The ultimate goal of the project is to come up with a pair of models (one linear, one logistic) that best describe the relationships between your predictors and your outcomes of interest. Be sure you describe carefully the set of procedures you use to identify your final model. This may represent a mixture of techniques.
- **Prediction:** Frequently one uses statistical models to make predictions for new observations. As part of your write-up use each of your final models to make predictions for what you think would be an interesting or important combination of your predictor values, briefly explaining your choices.

Write-up

Your findings should be written up as a statistical consulting report, aimed at a public health audience, in which you describe the data set and the goals of the analysis, explain the procedures you used to arrive at your final models, and present interpretations of the parameters of those models in an accurate and useful manner, along with appropriate predictions based on those models. You should also describe any problems you encountered and how you resolved them, carefully justifying your choices. Finally, your report will be enhanced by discussing the implications of your models in the context of the problem and any limitations of the data set for answering the questions of interest. Your report will be evaluated in part based on your ability to communicate insights about how the predictor variables (hospital or population characteristics) relate to the outcomes (length of hospital stay/risk of infection, crime rate/high poverty rate). It is therefore important to be clear about your motivation for carrying out the analyses as well as about the technical interpretations of the models.

Your main report should be on the order of 5-7 pages. You may include an appendix with computer output and graphics if you feel they will provide a useful reference. However, keep in mind that your audience (myself included!) does NOT want to have to sift through reams of printouts. Include ONLY what is genuinely useful and excise any parts of the printouts to which you will not refer. Any really critical graphics should be included in the main text. Some key information may be better summarized in tables than by including the printouts (e.g. it may be enough to give regression coefficients and p-values without all the accompanying information provided by STATA or SAS). Make sure that all tables, graphs, etc., are clearly labeled. The body of your report should include the following sections:

- **Introduction:** Provide a brief descrption of the data set and the goals of your analysis for someone who is unfamiliar with the project. Note that there are multiple possible motivations for these analyses; pick one that seems reasonable and make it clear from what point of view you are examining the data.
- **Statistical Methods:** State what analysis strategies you used, noting relevant models or assumptions. While it may be helpful in some cases to describe your analysis sequentially, do not take the reader down every false path you tried. It is more important that the order of presentation be consistent with your ultimate conceptual view of the problem than with your initial approach.
- **Results:** Give a summary of your main findings regarding relationships between the predictor and outcome variables along with interpretations relating to the magnitude, direction, statistical significance and interdependencies of those relationships.
- **Discussion/Conclusion:** Give highlights of interesting or important conceptual findings and discuss any other important issues that arose in the course of the analysis and their implications for your conclusions. This is where you get to be creative and combine your public health knowledge with your statistical skills!

Data Set I: SENIC

This data set comes from the Study on the Efficacy of Nosocomial Infection Control (SENIC study). The original objective of the study was to determine whether infection surveillance and control programs reduced rates of nosocomial (hospital-acquired) infections in U.S. hospitals. We will focus on a narrower set of questions about how hostpial characterstics relate to length of patient stay and risk of infection. In the data set available to us, there are 113 hospitals (indexed by ID number), with 11 quantities measured on each:

VARIABLE NAME	DESCRIPTION
id	Hopsital ID number ranging from 1-113
length	Average length of stay (in days) of all patients in the hospital
age	Average age (in years) of all patients in the hospital
risk	Infection risk measured as the estimated probability (in percent) of acquiring an infection in the hospital
culture	Ratio of number of cultures performed to number of patients without signs or symptoms of hospital acquired infections, times 100
xray	Ratio of number of x-rays performed to number of patients without signs or symptoms of pneumonia, times 100
beds	Average number of beds in hospital during study period
msch	Whether the hospital was affiliated with a medial school: 1 = Yes, 2 = No
region	Geographic area of the U.S. in which the hospital was located 1 = Northeast, 2 = North Central, 3 = South, 4 = West
census	Average number of patients in hospital per day during the study period
nurses	Average number of full-time equivalent registered and licensed practical nurses during study period (number full plus half the number part time)
svcs	Percent of a list of 35 potential facilities and services that were available at the hospital

Specific Notes for Analyzing the SENIC Data Set:

- Treat the average length of stay as the outcome variable, Y, for your multiple regression analysis.
- Define a new variable "HighRisk" which is 1 if the hospital has a risk of infection of 5% or higher and 0 otherwise and use this as the outcome variable for your logistic regression model.
- One question of particular interest for this data set is whether there are regional differences in the relationships between the various predictors and length of stay or risk of infection.
- It may be necessary to define a number of new variables for this analysis including transformations, interactions, conversion of categorical variables to dummy variables and even creation of derived variables that will be more appropriate for answering the questions of interest. Think about these choices carefully before beginning your analysis.

Data Set II: CDI

Your second option is to analyze the "County Demographic Information" (CDI) data set, which contains characteristics of 440 counties in the United States collected from 1990-1992. The primary objective of this investigation is to develop insight about how population characterstics of the counties relate to the crime rate (which you might want to summarize as the number of serious crimes per 1,000 people, CRM_{1000}) and to whether the county has a high poverty rate. The variables in the CDI data set are as follows:

VARIABLE NAME	DESCRIPTION
ID	An ID number for the county, ranging from 1-440
cty	A character string giving the name of the county
state	The two letter abbreviation for the name of the state continaining the county
area	Land area of the county in square miles
pop	The estimated population of the county in 1990
pop18	Percentage of total population in county aged 18-34 in 1990
pop65	Percentage of total population aged 65 or older in 1990
docs	Number of professionally active nonfederal physicians in 1990
beds	Number of available hospital beds (beds, cribs and bassinets) in 1990
crimes	Total number of serious crimes in 1990 as reported by law enforcement including rape, murder, robbery, aggravated assault, burglary, larceny-theft and motor vehicle theft
hsgrad	Percentage of persons 25 years or older who completed 12 or more years of school
bagrad	Percentage of persons 25 years or older who had a bachelor's degree
poverty	Percentage of 1990 total population with income below the poverty line
unemp	Percentage of labor force that was unemployed in 1990
pcincome	Per capita income, in dollars per person among those in the 1990 total population
totalinc	Total personal income in millions of dollars among those in the 1990 total population
region	Geographic area of the United States, according to the U.S. Census Bureau classified as 1 = Northeast, 2 = Northcentral, 3 =South 4 = West

Specific Notes for Analyzing the CDI Data Set:

- Treat the crime rate per thousand people as the outcome variable, Y, for your multiple regression analysis. (You will have to create this variable.)
- Define a new variable "HighPoverty" which is 1 if the county has a poverty rate of 10% or higher and 0 otherwise and use this as the outcome variable for your logistic regression model.
- In this data set, inter-relationships among the predictor variables will have a big effect on your analysis. Before beginning, think about how you might want to define new variables that would both provide better insight about the questions of interest and might eliminate some of the problems caused by the existing relationsips.

Optional Component

If you are feeling ambitious you can get bonus credit by writing a one-page abstract in which you describe a data set of interest to you in your own research, explain briefly how you might analyze it in light of what we have learned in the class, and describe some of the conceptual issues or problems you might encounter. Note: If you decide to actually analyze your own data set the optional component would need to describe an analysis separate from the one you present in your main report.