

Fall BioEpi 691F: Practical Data Management and Statistical Computing 2011

Assignment 9: Cleaning Data and Preparing a Preliminary Descriptive Report

Due: 01 DEC 2011

Reading in Course Notes:

- SAS Notes Unit 3: Introduction to the Data Step
- SAS Notes Unit 4: More on Working with Data in SAS
- SAS Notes Unit 5: Procedures for Data Description

Fetal Lung Maturity Study

Background:

The data for this assignment come from a study of an assay of amniotic fluid, used to try to determine fetal lung maturity. Many babies delivered prematurely are at risk of severe respiratory problems. Of course, lung maturity is associated with gestational age (the age, in weeks, of the fetus). However, factors causing stress for the fetus can also cause early maturing of lungs so that gestational age alone, which is usually an estimate, is not always a good predictor of lung maturity. Additionally, gestational age is unknown for some women when they come into the hospital to deliver.

There would be many advantages to patient management if it were possible to predict with some accuracy which babies are more likely to develop respiratory problems at birth once premature labor has begun, or when there are pressing reasons to induce labor or to perform a cesarean section before labor has begun naturally. The standard test available for fetal lung maturity is an expensive and time-consuming one. This new fetal lung maturity (FLM) assay is a much faster and simpler test to perform, requiring a small sample of amniotic fluid, obtained from either amniocentesis or vaginal pooling.

Over a period of time, investigators have been using the new assay on samples of amniotic fluid from women

- in preterm labor
- with gestational problems requiring labor induction
- having scheduled c-sections
- in labor with pregnancies of unknown gestational age.

If delivery occurred within 72 hours of the sampling, the data were included in the study.

After delivery, infants were evaluated for presence or absence of respiratory distress -- specifically respiratory distress syndrome (RDS) and transient tachypnea of the newborn (TTN).

High values for the FLM assay are indicative of lung maturity (good), and low values of immature lungs (bad). Investigators are interested in determining a cutoff value of FLM that will be highly sensitive in identifying cases of respiratory distress - either RDS or TTN, but will not

have too many false positives. The particular cutoff points for predicting lung immaturity that are under consideration are FLM values less than 50, or values less than 70.

There is some concern that the presence of blood in the sample will affect the accuracy of the assay, so information on the presence of blood in the amniotic fluid sample is also included.

The Study Data

The data are in text file format, in the file [flmraw.txt](#). Data available are (in order):

- FLM assay value
- an indicator of respiratory outcome at delivery: 1=RDS, 2=TTN, 0=no illness
- an indicator of presence of blood in the sample, 1=blood, 0=none
- gestational age (GA) at delivery, in weeks
- birthweight (missing!)
- a patient ID number

The data (~ 20 yrs old) were originally entered into the computer using the program STATVIEW on the MAC, and then translated into IBM DOS format. **Birthweight is missing in all cases** - it was not appropriately entered as numeric data, and did not translate across systems in readable format. Note: there are additional data on each record -- repeat information, and more missing data on the time and date of assay, time and date of birth, and method of sampling. **These data can be ignored since most of it is missing.**

1. Save a copy of the data and write a SAS program to:

a. Read the data into a SAS data set. (*? INFILE and INPUT or SET ?*)

TIP: The data has problem values – strange characters appeared in transition across platforms; to avoid seeing endless error listing in your log use:

OPTIONS errors=3; (or other small number to limit the number of errors printed).

b. Create a variable to indicate if the delivery is:

preterm (GA<37 weeks)
fullterm (GA 37-42 weeks)
postdates (GA >42 weeks)
or of unknown gestational age

c. Create two indicator variables: A variable to indicate whether the FLM value is below the cutoff value of 50 for prematurity, and the second indicator variable for FLM value less than 70.

d. Label all variables, create a format data set and assign formats to the variables, as appropriate. Save the data in a permanent SAS data set. Use PROC CONTENTS to document the data set.

2. **Data cleaning:** Check for and report on missing data, duplicate records, unusual values. Make a decision on how to handle the problem records and problem values.
3. Write a preliminary report of **no more than 4 pages including tables** for the investigator (a physician, ***not a statistician***), describing the data. **This is not an analysis report** (no hypotheses, statistical tests, inferences), but a preliminary report with descriptive information on the data that is available for use in analysis.

As appropriate, **your SAS output should be incorporated into the report.** Your report should address the following questions:

- How many records are there?
- Are there any duplicate records?
- Are there any values that appear unusual or suspicious to you?
- How have you handled any problem records? ***What was your rationale?***
- How have you handled any problem values? ***What was your rationale?***

Include in your report simple frequency tables of respiratory distress, blood in sample, maturity at delivery (preterm/fullterm/postdates/unknown), and the FLM cutoff indicators.

As appendices to the report, include

- 1) **results of PROC CONENTS on your final data set, and**
- 2) **all of your SAS logs from final runs of your programs.**

Try to be creative in using PUT statements, DO loops, arrays, ... as you check for problem values and create SAS missing values.

Your programs should have comments to make the purpose of each step clear.