Homework 4

STAT 448 - Advanced Data Analysis

Due: October 20, 2018 12:00:00 AM

Submitting your work to Compass

You are to submit two files for your homework submission.

1. Your SAS program file which should be saved as `HW#_YourNetID.sas`. For example, my file for the HW1 assignment would be `HW1_kinson2.sas`. All program statements and code should be included in one program file.
2. Your Report including all relevant code and output to address the exercises which should be saved as `HW#_YourNetID.pdf`. For example, my file for HW1 would be `HW1_kinson2.pdf`.

You have an unlimited number of submissions, but only the last submission (which contains those two files) will be viewed and graded. To submit, click on the title of the assignment in Compass. The homework questions begin below the line. Be sure to attach the relevant files as dictated in that week's assignment.

Starting SAS program for this assignment

To complete this assignment, you will need to analyze the data sets in the `Program_HW4_Data_Fall2018.sas` file on Compass.

The **birth2007** dataset is a public use data file of registered births in the US collected the Center for Disease Control and Prevention's (CDC) National Center for Health Statistics (NCHS). This natality dataset has been partitioned for storage capacity reasons and the combined dataset contains various measures about the mother, medical care, and infant for over a million births from the year 2007. The exact dimensions of **birth2007** are 1718600 observations and 21 variables . The table below includes descriptions of the variables.

Variable Name	Description
DBWT	infant's birth weight (in grams, g)
BWTRC	infant's birth weight as category; 1=1499 g or less, 2=1500-2499 g, 3=2500 g or more
MRACE	mother's race; 1=not white, 0=white
MARR	mother's marital status; 1=not married, 0=married
SEX	sex of the infant; M=male, F=female
MAGER	mother's age in years
MAGERC	mother's age as category; 1=15-19 yrs, 2=20-24 yrs, 3=25-29 yrs, 4=30-34 yrs, 5=45-39 yrs, 6=4-44 yrs, 7=45-49 yrs
CIG_REC	0=mother is a non-smoker, 1=mother is a smoker
CIG_1	number of cigarettes (per day) during 1st trimester
CIG_2	number of cigarettes (per day) during 2nd trimester
CIG_3	number of cigarettes (per day) during 3rd trimester
WTGAIN	mother's weight gain (in pounds) during pregnancy
PRECARE_REC	period of months that prenatal care began; 1=1st-3rd month, 2=4th-6th month, 3=7th-final month, 4=no prenatal care
MEDUC	mother's education level; 1=middle school, 2=some high school, 3=high school graduate or GED, 4=some college, 5=associate degree, 6=bachelor's degree, 7=master's degree, 8=doctorate or professional degree
BFACIL	birth place facility; 1=in hospital, 0=not in hospital
RDMETH_REC	birth delivery method; 1=vaginal, 2=vaginal after previous cesarean, 3=primary cesarean, 4=repeat cesarean
ATTENDC	type of birth attendant; 1=doctor, 2=certified nurse midwife, 3=other midwife, 4=other
APGAR5	five minute APGAR score between 0 and 10 used to assess infant's health (the higher the better)
APGAR5R	five minute APGAR as category; 1=score of 0-3, 2=score of 4-6, 3=score of 7-8, 4=score of 9-10
DPLURAL	plurality or the number of babies born for this pregnancy; 1=single, 2=twin, 3=triplet, 4=quadruplet, 5=quintuple or more

The problems below will be restricted to births to a single baby (not n-tuples) from mothers between the ages of 18 and 45 and will investigate models for counts and categorical responses. (10 problems)

1. What is the frequency of the birth weight categories among the apgar score categories? Create a contingency table that shows the cross-tabulation. Does there appear to be an association between the two categorical variables (without running a test)?

2. What is the frequency of the birth weight categories among the mother's age categories? Create a contingency table that shows the cross-tabulation. Does there appear to be an association between the two categorical variables (without running a test)?

3. Dichotomize the infant birth weight as low (=1 if less than 2500 grams) vs normal (=0 if at least 2500 grams). Using this dichotomized variable as the response (reference level is 0), fit a logistic regression using all of the following predictors in the model: mother's race, mother's marital status, mother's weight gain, delivery method, mother's age (categorical), beginning prenatal care period, and mother's education level. Which predictors are significant?

4. Starting with the model in *Problem 3*, determine the best set of predictors for the model and comment on any issues with influential points. If any extreme influential points exist, remove them and perform model selection again, before choosing a final model.

5. Using your final chosen model from *Problem 4*, discuss the relevant model results. Include comments on the significance of parameter estimates, the goodness of fit, and model diagnostics. Interpret what the model tells us about relationships between the predictors and the odds of an infant having low birth weight.

6. The test called, Apgar Scores, was created in 1953 by Virginia Apgar, an anesthesiologist. The score is recorded at 5 minutes to summarize how well the baby is doing outside of the mother's womb. Now, set the **APGAR5R** variable as the response. Fit a cumulative logit model (such that the parameter estimates are in favor of lower order direction rather than the higher order direction) using all of the following predictors in the model: mother's race, mother's marital status, mother's weight gain, delivery method, mother's age (categorical), beginning prenatal care period, and mother's education level. Which predictors are significant?

7. Starting with the model in *Problem 6*, determine the best set of predictors for the model and comment on any issues with influential points. If any extreme influential points exist, remove them and perform model selection again, before choosing a final model.

8. Using your final chosen model from *Problem 7*, discuss the relevant model results. Include comments on the significance of parameter estimates, the goodness of fit, and model diagnostics. Interpret what the model tells us about relationships between the predictors and the ordinal response.

9. Using the DATA step, create a new dataset called **cigbirth2007** (based on **birth2007**) that contains only the mothers who are smokers (**CIG_REC**=1). There should be observations in this new dataset. Create a new variable called **DAILYCIG_AVG** that is the rounded average of the daily cigarette variables. In other words, take the sum of the 3 cigarette trimester variables (**CIG_1**, **CIG_2**, **CIG_3**) divided by 3, then use the **round()** function to compute the rounded integer value. Treating this **DAILYCIG_AVG** as a count response, fit a Poisson log-linear model with the following predictors: mother's race, mother's marital status, mother's weight gain, mother's age (categorical), beginning prenatal care period, and mother's education level. Should we use an overdispersed model instead of the traditional Poisson model? Discuss why or why not? Which predictors are significant?

10. If we should use an overdispersed model, use that model for this problem. Otherwise, use the model from *Problem 9.* Determine the best set of predictors for the model and discuss the relevant model results. *If there are issues with influential points, resolve them and refit the model, before discussing relevant model results.* Include comments on the significance of parameter estimates, the goodness of fit, and model diagnostics. Interpret what the model tells us about relationships between the predictors and the count response.