ERHS 642 Logistic Regression Spring 2016
Final Exam

Discussing any part of the exam with others is not permitted.
Students caught cheating or attempting to cheat will be reported to the CSU Conflict Resolution and Student Conduct Services (CRSCS) office and will receive a grade of 0 on the exam.

The website Mags4u.com is planning an e-mail marketing campaign. Emails will be sent to customers who have previously bought a magazine subscription at Mags4u.com and who have not opted out of receiving e-mails. Mags4u.com wants to target the ads to specific customers and needs to know which combinations of variables explain purchase probabilities for different magazines.

One of the magazines to be advertised is "Kid Creative" whose target audience are children. Mags4u.com sends "experimental" emails containing the ad for "Kid Creative" to 500 randomly selected customers who have previously bought a magazine subscription at Mags4u.com. For each customer, Mags4u.com records whether or not he/she buys a subscription for "Kid Creative" after receiving the ad. For each customer, Mags4u.com combines this information with data collected during the most recent prior subscription purchase and with third party data which Mags4u.com acquired. The following variables are available:

- Customer number (Obs_No_)
- Purchased "Kid Creative" (Buy = 1 if purchased "Kid Creative," 0 otherwise)
- Household Income (Income; rounded to the nearest $1,000)
- Gender (Is_Female = 1 if the person is female, 0 otherwise)
- Marital Status (Is_Married = 1 if married, 0 otherwise)
- College Educated (Has_College = 1 if one or more years of college education, 0 otherwise)
- Employed in a Profession (Is_Professional = 1 if employed in a profession, 0 otherwise)
- Retired (Is_Retired = 1 if retired, 0 otherwise)
- Not employed (Unemployed = 1 if not employed, 0 otherwise)
- Length of Residency in Current City (Residence_Length; in years)
- Dual Income if Married (Dual_Income = 1 if dual income, 0 otherwise)
- Children (Minors = 1 if children under 18 are in the household, 0 otherwise)
- Home ownership (Own = 1 if own residence, 0 otherwise)
- Resident type (House = 1 if residence is a single family house, 0 otherwise)
- Race (White = 1 if race is white, 0 otherwise)
- Language (English = 1 is the primary language in the household is English, 0 otherwise)
- Previously purchased a children's magazine (Prev_Child_Mag = 1 if previously purchased a children's magazine, 0 otherwise)
- Previously purchased a parenting magazine (Prev_Parent_Mag = 1 if previously purchased a parenting magazine, 0 otherwise)

(Note: Replicate is the number of your data set. For example, if your data set is kc_1, then replicate=1)

1. Assume "Minors" is a risk factor of particular interest. Estimate the power to detect an OR of 2.5 for "Minors" in a model containing 3 additional dichotomous variables. Assume (i) that Minors=1 for 30% of controls but that the additional dichotomous variables are equal to 1 for only 5% of controls; (ii) that 15% of ad recipients buy "Kid Creative"; and (iii) that α=0.05. Note that controls are households who did not buy "Kid Creative" after receiving the ad. *(3 points)*

2. Build a model based on goal 1. You must include the following steps (each step must be included but you can change the order or add other steps):

   a. Assessment of the frequencies of the categorical study variables stratified by the outcome variable; description of sparse cells and how you are dealing with them *(2 points)*

   b. Assessment of the descriptive statistics of the continuous study variables stratified by the outcome variable *(1 point)*
   Assessment of the extreme observations of the continuous study variables stratified by the outcome variable *(1 point)*
   Description of unusual values and how you are dealing with them *(1 point)*

   c. Univariate scale assessment (spline plots and fp) including conclusions *(4 points)*

   d. Univariate statistical significance assessment including conclusions *(2 points)*
   Multivariate statistical significance/confounding assessment including your reasoning along the way and conclusions *(4 points)*

   e. If your main effects model includes continuous variables: Multivariate scale assessment (fp or design variable plots) including conclusions *(-1 point if needed but missing)*

   f. Presentation of the final main effects model (no interpretation of the results necessary) *(1 point)*

3. Perform best subsets selection to determine if you missed any important model covariates (main effects only). You can use collapsed variables and scale assessment results from question 2. If necessary, make changes to your final model from question 2. *(4 points)*

4. Starting with your main effects model (from question 2 or, if you made changes, from question 3), perform a complete assessment of interactions. YOUR FINAL MODEL MUST INCLUDE AT LEAST ONE INTERACTION TERM. *(4 points)*

5. Present your final model: Present two tables, one containing the model coefficients, their standard errors and p-values and the other containing appropriate ORs and 95% confidence intervals. *(4 points)*

6. Interpret the ORs and 95% confidence intervals. *(2 points)*

7. Recall that Mags4u.com wants to target ads to specific customers. Describe the customers who should receive the ad for "Kid Creative". *(2 points)*

8. For your final model (from question 5), use the Pearson chi-square test, deviance test, Hosmer-Lemeshow test and Osius-Rojek test to assess overall model fit. Comment on the appropriateness of each test and draw conclusions regarding model fit. *(5 points)*

9. Use the Stukel test to test the tails assumption. Draw conclusions. *(2 points)*

10.
    a. Use all 4 logistic regression diagnostics to identify outliers. *(4 points)*
    b. Describe why the outliers are outliers in the model. *(2 points)*
    c. Show the effect deletion of the outliers has on the model ORs and p-values. *(2 points)*
    d. Propose how to deal with the outliers. *(1 point)*

11. Keeping all outliers in the data set, determine how well your model predicts the outcome.
    a. Plot the ROC curve and determine the area under the ROC curve. *(1 point)*
    b. Plot sensitivity and specificity vs. possible cutpoints and select the “best” cutpoint *(2 points)*
    c. Based on this cutpoint, calculate sensitivity and specificity. *(2 points)*
    d. Calculate the positive and negative predictive value assuming the prevalence of the outcome in the population of interest is 10%. *(2 points)*
    e. Draw conclusions keeping in mind the limitations of the above methods (show relevant SAS output). *(2 points)*