STAT 645: Biostatistics - Assignment 5
Due Monday, October 5, 11:55pm CT

1. Suppose that a pilot study was conducted to assess the feasibility of patient recruitment, the ability of patients and clinicians to comply with study protocols, and the use of data collection instruments to collect cost-effective data, and to obtain variability estimates for sample-size calculations for a full-scale trial. Suppose that twenty patients were randomized into the study with treatment and control group. Out of twelve patients in the treatment, 8 showed substantial improvements in the main patient-rated outcomes at the end of the 12-week intervention phase. Let $\pi$ be the proportion patients who showed substantial improvements in the main patient-rated outcomes at the end of the 12-week intervention phase in the treatment group.

   (a) Construct two-sided 95% confidence interval for $\pi$ using Agresti-Coull, Jeffreys, Wilson, Clopper-Pearson methods.

   (b) What would be the required sample size for the actual study if we want to test $H_0 : \pi = 0.6$ versus $H_a : \pi > 0.6$ at the 5% level, and we desire to have 90% power to reject $H_0$ when in fact $\pi = 0.7$?

   (c) Recalculate the needed sample size for the above scenario considering that there is a possibility of 35% drop-out or study non-compliance.

2. Suppose in an observational study on PTSD we have obtained the following data. Test at the 5% level if there is any association between PTSD and gender. Write the hypothesis, do the analysis, and write your conclusion. Use the both methods, the chi-square test of independence and the odds ratio approach.

| PTSD | Gender | |
|---|---|---|
| | M | F |
| Y | 40 | 60 |
| N | 280 | 156 |

3. Consider the Pima.tr dataset in library(MASS). This dataset contains information on some 200 Pima Indian women who were all at least 21 years old. Please look at `https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/Pima.tr.html` for details of the data. Suppose that interest is finding association between type and npreg, glu, bp, skin, bmi, ped, age. Before the analysis, tranform glu, bp, bmi, ped, age into variables that have zero mean and standard deviation one.

   (a) Test if age is positively associated with the disease (chances of the disease).

   (b) Test $H_0 : \beta_{skin} = \beta_{bp} = \beta_{bmi} = 0$ at the 5% level. Use both the likelihood ratio and Wald test approaches.

   (c) Provide a Cook's distance plot and check if there is any influential observation.

   (d) Consider the model containing, npreg, glu, ped, age, age$^2$, ped$\times$ age, glu$\times$ age, glu$\times$ ped as explanatory variables. Do a stepwise regression to find the the best fitted model for this data based on the above specified explanatory variables [Hint use the step(obj) function].