Homework 4, STAT 7630, SPRING 2024

D. Pluta

Due: 2024-04-30

Homework 4

1. Provide a one paragraph description of the data set you will be using for the final project. Make sure to mention the sample size, the outcome measure, and the covariates measured.
2. What questions will you be trying to answer using a linear regression model? Explain the context of the data and the question in terms that can be understood by a non-expert.
3. Provide the title, authors, journal, issue, and year of publication for three useful papers you have found as part of your literature review.
4. What covariates in the data set do you think will need to be included as potential confounders? (Describe your current thoughts and plan, it's okay if this doesn't exactly match the model you use in the final report.)
5. Describe/summarize the method of data collection for your data set. What is an appropriate population we can generalize to from these data?
6. Provide a scatterplot of your data with the outcome on the y-axis and one of the covariates of interest on the x-axis. Label the axes with English words that clearly indicate what the axis is quantifying and the scale used.

---

7. For the Framingham Heart Study dataset you used in Homework 3, we wish to determine if there is an interaction effect of current smoking status and BMI on systolic blood pressure.

a. Fit the baseline main effect model:

$$sysBP \sim TenYearCHD + male$$

Provide a nicely formatted table of the regression results, including the upper and lower limits for the 95% confidence intervals for the coefficients.

b. Fit interaction model:

$$sysBP \sim TenYearCHD + male + TenYearCHD : male$$

Provide a nicely formatted table of the regression results similar to (a).

c. Compare and interpret the results of the two models. How does the interpretation of the effect of **male** differ in the two models?

d. Produce a boxplot of **sysBP** stratified by **male** on the x-axis, and with different colors for the value of **TenYearCHD**. Interpret this plot, and describe how it shows a significant interaction effect.

8. For the linear model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} * x_{i2} + \varepsilon_i, \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, assume that x_2 is a binary covariate. Calculate the variance of the estimator for the difference in means of two observations, where observation A has covariate value $x_{A2} = 1$, and observation B has $x_{B2} = 0$. Assume $x_{A1} = x_{B1} - 1$.
9. Use the provided R code to construct a power plot for testing a main effect $H_0 : \beta_1 = 0$ with parameter settings $\beta_0 = -1, \beta_1 = 0.25, \sigma^2 = 0.7$. Vary over the sample size n in increments of 50, from $n = 50$ up to $n = 250$. Run 500 replicates for each setting of n and plot the results.

10. Use the provided R code to construct a power plot for testing an interaction effect $H_0 : \beta_3 = 0$ with parameter settings $\beta_0 = -1, \beta_1 = 0.25, \beta_2 = 0.1, \beta_3 = 0.2, \sigma^2 = 0.7$. Vary over the sample size n in increments of 50, from $n = 50$ up to $n = 250$. Run 500 replicates for each setting of n and plot the results.