**Biostat/Epi 536 2024**

**HW 8 (7 questions)**

In 1960-1961, 3,154 healthy, middle-aged men were entered into the WesternCollaborative Group Study (WCGS), a long-term study of coronary heart disease (CHD). The risk of coronary heart disease mortality wasstudied for several variables measured at baseline, i.e., Type A/Bbehavior, systolic blood pressure, serum cholesterol level, cigarettesmoking status, and age. Although there is no codebook for the WCGS dataset, the variable names in the “**wcgsdata.Rdata**” dataset are largely informative. As done in Lecture 9, we will pretend that the variable chd69 is a scientifically valid binary indicator of coronary heart disease, ignoring the important issue of differential length of follow-up.

**1.** Use logistic regression to investigate the question: Does smoking status modify the association between systolic blood pressure (sbp) and CHD. Fit an appropriate logistic regression model that addresses this question using a binary smoking exposure indicator variable (0 if ncigs==0, 1 if ncigs>0).

a. Write your fitted logistic regression model.

b. Explain *briefly* how to use the model to test the null hypothesis that smoking does not modify the SBP-CHD association.

c. Perform an appropriate statistical test corresponding to b. Write the results in a sentence suitable for a scientific publication.

d. Regardless of the results in (c), use your fitted model to describe the SBP association with CHD for smokers and non-smokers and include confidence intervals.

**2.**

a. Use a regression technique that summarizes associations with relative risks. Investigate an association between systolic blood pressure (sbp) and having a CHD event. Write out the fitted model, then write 1-2 sentences summarizing results in language suitable for a scientific publication.

b. (part b is optional) In class we discussed three different variants of relative risk regression. Use a different variant than the one you used for 3a and compare/contrast results.

**3.** Does being a smoker modify the association between systolic blood pressure (sbp) and CHD as measured with the relative risk? Fit an appropriate model that addresses this question using a binary smoking exposure variable (0 if ncigs==0, 1 if ncigs>0).

a. Write your fitted model.

b. Perform an appropriate statistical test to answer the question. Write the results in a sentence suitable for a scientific publication.

**4**. Comment on differences or similarities between results in Q1 and Q3

**5.** Now use a regression technique that summarizes associations with risk differences. Investigate an association between systolic blood pressure (sbp) and having a CHD event, adjusting for age, body mass index (bmi) and cholesterol (chol). Write out the fitted model, then write 1-2 sentences summarizing results in language suitable for a scientific publication.

**6.** We will now revisit the Framingham data and consider modeling CHD incidence and its association with high cholesterol. We will use the “framingham\_HW8.Rdata” dataset.

**Hint:** Incidence=#chd events/person-time at risk. Assume there are 365.25 days in a year use R to generate the variable logpyears = log(days/365.25)

Also, create a binary CHDbin variable which is 0 if CHD=”Censored” and 1 if CHD=”CHD”

a) Use Poisson regression to estimate the association between high cholesterol (chol200=1) and **incidence** of CHD, adjusted for sex, age (agegrp) and body mass index (BMIgrp). State your conclusion in a sentence or two suitable for an abstract in a medical journal.

b) From the model in 6a, estimate the incidence and 95% CI of CHD (per 1000 person-years) for males over 60 with BMI > 30 and cholesterol over 200 (chol200=1), and then similarly for calculate the incidence for the same group but with cholesterol <=200 (chol200=0). For this dataset, the model and robust variance estimates are similar, so you can use the estimable command to get good approximations of the confidence intervals you need.

**7.** Fit a slightly different model compared to Q1. Use a different coding for the binary smoking variable (-1 if ncigs==0, 1 if ncigs>0). With this coding, fit a model with a main effect for sbp, main effect for smoking (coded -1/1), and an interaction term.

a. Explain how to use this model to test the null hypothesis that smoking does not modify the SBP association with CHD.

b. Compare the p-value for the interaction term from this model to the p-value for the interaction term for Q1. Are they the same or different? Why does this make sense?