Andrew J. Spieker, PhD
BIOS 6312 - Modern Regression Analysis
Spring 2020
Homework #4 Key

1. We consider the data set set `mri.csv`.

   (a) Perform an analysis to evaluate the association between age and odds of diabetes.

   Ans: These data do not provide sufficient evidence of an association between age and diabetes ($p = 0.59$). Comparing subgroups differing in age by one year, we estimate the older subgroup to have a 1.20% lower odds of diabetes as compared to the younger subgroup. Based on a 95% confidence interval, this estimate would not be deemed atypical if in truth the older subgroup had between a 5.45% lower and 3.24% higher odds of diabetes.

   (b) Repeat part (a); adjust for sex and race and allow a two-way interaction between age/sex.

   Ans: We simultaneously test the age coefficient and the age-sex interaction term. These data do not provide sufficient evidence of an overall association between age and odds diabetes ($p = 0.493$).

   (c) Use the model of part (b) to evaluate whether sex modifies the (adjusted) association between age and odds of diabetes.

   Ans: We estimate the interaction term (i.e., the *ratio* of odds ratios comparing *males* of the same race and differing in age by one year to *females* of the same race and differing in age by one year) to be 1.04; based on a 95% confidence interval, this estimate would not be deemed atypical if in truth the ratio of odds ratios were between 0.952 and 1.14. This study does not provide sufficient evidence of effect modification ($p = 0.366$)

   (d) Is it possible to use the model of part (b) to predict the odds of diabetes among black 70-year-old females? If so, do so; if not, briefly explain why not (maximum: two sentences).

   Ans: It is possible. We estimate the odds to be 0.161 in this subgroup:

$$\widehat{\text{Odds}} = \exp(0.6705363 + 70 \times (-0.0454722) + 0.6883816) = 0.161.$$

   (e) Is it possible to repeat part (b) on the *risk* scale instead of the odds scale? If so, do so; if not, briefly explain why not (maximum: two sentences).

   Ans: This is possible because we are in the setting of a cohort study (i.e., not one of outcome-dependent sampling). We simultaneously test the age coefficient and the age-sex interaction term. These data do not provide sufficient evidence of an overall association between age and risk of diabetes ($p = 0.497$).

2. A case-control study examines risk-factors for esophageal cancer. Load the data set `esoph.csv`.

   (a) Perform an analysis to evaluate the association between alcohol consumption and odds of esophageal cancer, adjusting for tobacco use and age.

   Ans: This study provides strong evidence of an association between alcohol consumption and odds of esophageal cancer ($p < 0.001$). We estimate odds ratios of 3.20, 4.93, and 9.23, each of which compares the odds of esophageal cancer between alcohol consumption groups 40-79 g/day, 80-119 g/day, and 120+ g/day to the reference group of the same age and tobacco consumption groups, and an alcohol consumption of 0-30 g/day. The respective 95% confidence intervals for each adjusted odds ratio (i.e., the sets of values of the odds ratios for which our own estimates would not be judged unusual) are [2.01, 5.09], [2.96, 8.22], and [5.18, 16.4].

   (b) Suppose it was believed *a priori* that tobacco use was associated with both esophageal cancer and alcohol consumption. In a maximum of two sentences, justify the choice of including tobacco use in the model of part (a).

   Ans: Including a variable for tobacco use serves to adjust for potential confounding.

   (c) Suppose it was believed *a priori* that age was associated with esophageal cancer but unrelated to alcohol consumption. In a maximum of two sentences, justify the choice of including age in the model of part (a).

   Ans: Failure to adjust for age in this setting would likely attenuate the target parameters (i.e., bring them closer to the null).

   (d) Is it possible to use part (a) to predict the odds of esophageal cancer among 60-year-olds who do not smoke or drink alcohol? If so, do so; if not, briefly explain why not (maximum: two sentences).

   Ans: This is not possible because we are in the setting of a case-control study. The odds ratios comparing the odds of an outcome across subgroups of the exposure may be estimable from mathematical equivalence to the (estimable) odds ratio comparing the odds of *exposure* across the outcome, but that does not mean the subgroup-specific odds are estimable themselves (they are not).

   (e) Is it possible to repeat part (b) on the *risk* scale instead of the odds scale? If so, do so; if not, briefly explain why not (maximum: two sentences). Additionally, if you answered that it is *not* possible, what information, if true, would allow you to use part (b) to *approximate* the desired risk ratio?

   Ans: This is not possible because we are in the setting of a case-control study. While odds ratios that compare odds of the outcome across subgroups of the exposure have convenient properties that allow them to be estimated, no such equivalence exists for risk ratios; if the disease is rare (it is in this case), the odds ratios do, however, approximate the desired risk ratios.

3. A cross-sectional study is conducted to examine the association between kidney stone history and coronary artery calcification (CAC). For problems (a)-(d), work on the scale of odds ratios. Load the data set `cac.csv`.

   (a) Perform an analysis to evaluate the association between kidney stone history and CAC, treating each as binary and adjusting for age, gender, and race category.

   Ans: Based on this analysis, there is insufficient evidence to conclude an association between kidney stone history and odds of CAC ($p = 0.697$). Comparing subgroups with a history of kidney stones to subgroups of the same age, gender, and race but with no history of kidney stones, we estimate that the subgroup with a history of kidney stones has a 24.5% higher odds of CAC. Based on a 95% confidence interval, this estimate would not be judged unusual if in truth the kidney stone group had between a 58.7% lower and 275% higher odds of CAC.

   (b) Repeat part (a), instead treating kidney stone history as a continuous variable.

   Ans: There is insufficient evidence to conclude an association between kidney stone history and odds of CAC ($p = 0.655$). Comparing subgroups of the same age, gender, and race differing in their kidney stone category by one unit, we estimate that the subgroup with more extensive kidney stone history has a 21.6% higher odds of CAC. Based on a 95% confidence interval, this estimate would not be judged unusual if in truth the group with more extensive kidney stone history had between a 48.3% lower and 186% higher odds of CAC.

   (c) Repeat part (a), instead treating both CAC and kidney stone history nominally (categorically, with unordered categories).

   Ans: We employ a multinomial logistic model for this purpose, selecting the group with *no* CAC as the reference outcome, and selecting the group with no kidney stone history as the reference exposure category. In turn, we obtain six separate odds ratios. First consider the odds ratios comparing the odds of mild CAC (1-99 HU) relative to no CAC. We estimate that the odds ratio comparing those with one prior kidney stone to those with none (but of the same age, gender, and race) to be 0.855, and comparing those with two or more prior kidney stones to those with none (but of the same age, gender, and race) to be 0.818; the 95% confidence intervals for these odds ratios are given by [0.220, 3.32] and [0.0750, 8.92], respectively. Now, consider the odds ratios comparing the odds of moderate CAC (100-399 HU) relative to no CAC. We estimate that the odds ratio comparing those with one prior kidney stone to those with none (but of the same age, gender, and race) to be 1.55, and comparing those with two or more prior kidney stones to those with none (but of the same age, gender, and race) to be 1.02; the 95% confidence intervals for these odds ratios are given by [0.410, 5.83] and [0.0919, 11.4], respectively. Now, consider the odds ratios comparing the odds of severe CAC (400+ HU) relative to no CAC. We estimate that the odds ratio comparing those with one prior kidney stone to those with none (but of the same age, gender, and race) to be 4.91, and comparing those with two or more prior kidney stones to those with none (but of the same age, gender, and race) to be 24.9; the 95% confidence intervals for these odds ratios are given by [0.425, 56.7] and [1.11, 556], respectively. There is not sufficient evidence in these data to suggest an association between kidney stone history and CAC category ($p = 0.299$).

(d) Repeat part (a), instead treating CAC ordinally and kidney stone history nominally.

We employ an ordinal logistic model for this purpose, selecting the group with *no* CAC as the reference outcome, and selecting the group with no kidney stone history as the reference exposure category. In this case, our odds ratios compare the odds of belonging to a CAC outcome category one level higher between subgroups. We estimate that the odds ratio comparing those with one prior kidney stone to those with none (but of the same age, gender, and race) to be 1.77, and comparing those with two or more prior kidney stones to those with none (but of the same age, gender, and race) to be 5.13; the 95% confidence intervals for these odds ratios are given by [0.757, 4.12] and [0.687, 38.3], respectively. There is not sufficient evidence of an association between kidney stone history and CAC category ($p = 0.162$).

(e) Provide a one-sentence advantage and one-sentence limitation of *each* model you fit—(a), (b), (c), and (d).

Ans: My responses are as follows. There may be other correct answers.

- Advantage of (a): This is extraordinarily simple and easily explainable.
- Disadvantage of (a): There is a loss of information in the outcome the exposure.
- Advantage of (b): The ordered nature of the exposure is accounted for in this model, even if that is not so for the outcome.
- Disadvantage of (b): The difference in means across neighboring exposure categories is presumed constant, which may not be reasonable.
- Advantage of (c): This requires the fewest assumptions of all the models considered and does not further coarsen the data to achieve this.
- Disadvantage of (c): It is cumbersome to have to interpret six separate coefficients and there is no accounting for ordering of the exposure/outcome.
- Advantage of (d): The ordered nature of the outcome is taken into account here.
- Disadvantage of (d): The proportional odds assumption might not be satisfied in practice (this could at least be inspected with the multinomial logistic model).

4. Consider the data set `mri.csv`. One measure taken from the MRI scan included the variable `numinf`, the number of distinct regions identified on the MRI scan suggestive of infarcts.

(a) Perform an analysis to evaluate whether cerebrovascular events (`stroke`) are associated with number of distinct regions identified on the MRI scan suggestive of infarcts. You will need to make decisions in this analysis; state them upfront in a couple of sentences and briefly justify them.

Ans: We treat stroke category categorically to avoid unfounded assumptions regarding constant rate ratios; we employ robust standard errors to allow for misspecification of the mean model and the mean-variance relationship. There is strong evidence of an association between stroke history and rate of infarct regions ($p < 0.001$). We estimate the rate ratio comparing those with a prior TIA to those with no stroke history to be 1.36 (95% CI: [0.814, 2.27]). We further estimate the rate ratio comparing those with a prior stroke to those with no stroke history to be 1.93 (95% CI: [1.53, 2.45]).

(b) Bill argues that we should include the variable `volinf` as an offset. In a maximum of two sentences, state whether or not you agree with Bill and why. If you agree, perform this analysis. If you disagree, state what you would *prefer* to do if you had access to any variables of your choosing (regardless of whether they were actually measured)—in the latter case, you would not be expected to perform the analysis.

Ans: I disagree. Instead, a measure of brain volume may better serve the purpose of an offset.