STA 2201/442 Assignment 10

1. This question explores the practice of "centering" quantitative explanatory variables in a regression by subtracting off the mean.

   (a) Consider a simple experimental study with an experimental group, a control group and a single quantitative covariate. Independently for $i = 1, \ldots, n$ let

   $$Y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \epsilon_i,$$

   where x_i is the covariate and d_i is an indicator dummy variable for the experimental group. If the covariate is "centered," the model can be written

   $$Y_i = \beta'_0 + \beta'_1(x_i - \overline{x}) + \beta'_2 d_i + \epsilon_i,$$

   where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

   i. Express the β' quantities in terms of the β quantities.

   ii. If the data are centered, what is $E(Y|x)$ for the experimental group compared to $E(Y|x)$ for the control group?

   iii. By the invariance principle, what is $\widehat{\beta}_0$ in terms of $\widehat{\beta}'$ quantities? Assume ϵ_i normal if you wish.

   (b) In this model, there are $p-1$ quantitative explanatory variables. The un-centered version is

   $$Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_i,$$

   and the centered version is

   $$Y_i = \beta'_0 + \beta'_1(x_{i,1} - \overline{x}_1) + \ldots + \beta'_{p-1}(x_{i,p-1} - \overline{x}_{p-1}) + \epsilon_i,$$

   where $\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}$ for $j = 1, \ldots, p-1$.

   i. What is β'_0 in terms of the β quantities?

   ii. What is β'_j in terms of the β quantities?

   iii. By the invariance principle, what is $\widehat{\beta}_0$ in terms of the $\widehat{\beta}'$ quantities? Assume ϵ_i normal if you wish.

   iv. Show that $\widehat{\beta}'_0 = \overline{Y}$. Hint: Differentiate the log likelihood.

   (c) Now consider again the study with an experimental group, a control group and a single covariate. This time the interaction is included.

   $$Y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \beta_3 x_i d_i + \epsilon_i$$

   The centered version is

   $$Y_i = \beta'_0 + \beta'_1(x_i - \overline{x}) + \beta'_2 d_i + \beta'_3(x_i - \overline{x})d_i + \epsilon_i$$

i. For the un-centered model, what is the difference between $E(Y|X = \overline{x})$ for the experimental group compared to $E(Y|X = \overline{x})$ for the control group?

ii. What is the difference between intercepts for the centered model?

(d) Suppose that in the study with an experimental group, a control group and a single covariate, the response variable is binary and we are doing a logistic regression.

i. Under the un-centered model, if there is no interaction, the odds of $Y = 1$ are ___ times as great for the experimental group, for any fixed value of x.

ii. Under the *centered* model, if there is no interaction, the odds of $Y = 1$ are ___ times as great for the experimental group, for any fixed value of x.

iii. If there *is* an interaction and $x = \overline{x}$, the odds of $Y = 1$ for the experimental group are ___ times as great. Express the answer in terms of β values, and also in terms of β' values.

2. If two events have equal probability, the odds ratio equals ___.

3. Consider a logistic regression in which the cases are newly married couples with both people from the same religion, the explanatory variable is religion (A, B, C and None – let's call "None" a religion[1]), and the response variable is whether the marriage lasted 5 years (1=Yes, 0=No).

(a) Make a table with four rows, showing how you would set up indicator dummy variables for Religion, with None as the reference category.

(b) Add a column showing the odds of the marriage lasting years. The symbols for your dummy variables should not appear in your answer, because they are zeros and ones, and different for each row.

(c) What is the ratio of the odds of lasting 5 years or more for religion C to the odds of lasting 5 years or more for No Religion? Answer in terms of the β symbols of your model.

(d) What is the ratio of the odds of lasting 5 years or more for religion A to the odds of lasting 5 years or more for Religion B? Answer in terms of the β symbols of your model.

(e) You want to test whether Religion is related to whether the marriage lasts 5 years. State the null hypothesis in terms of one or more β values.

(f) You want to know whether marriages from Religion A are more likely to last 5 years than marriages from Religion C. State the null hypothesis in terms of one or more β values.

(g) You want to test whether marriages between people of No Religion have a 50-50 chance of lasting 5 years. State the null hypothesis in terms of one or more β values.

[1] It is, for some people.

4. In a logistic regression, what does $\beta_0 = 0$ mean in terms of the probability that $Y = 1$?

5. Recall that the *null deviance* refers to the -2 log likelihood for a model with all regression coefficients equal to zero except β_0. Suppose that $n = 100$ and $\overline{Y} = 0.40$. What is the null deviance? The answer is a single number. There is a smart way to answer this question, a medium way and a dumb way. The smart way and the medium way use the invariance principle. Here is the dumb way, so you can check your answer.

```
> y = c(numeric(40)+1,numeric(60))
> glm(y~1,family=binomial)$null.deviance
[1] 134.6023
```

6. People who raise large numbers of birds inhale potentially dangerous material, especially tiny fragments of feathers. Can this be a risk factor for lung cancer, controlling for other possible risk factors? Which of those other possible risk factors are important? Here are the variables in the file `birdlung.data`.

Variable	**Values**
Lung Cancer	1=Yes, 0=No
Gender	1=Female, 0=Male
Socioeconomic Status	1=High, 0=Low
Birdkeeping	1=Yes, 0=No
Age	
Years smoked	
Cigarettes per day	

Please use SAS. You may check your answer with R if you like, but bring the SAS log and list file to the quiz. *A calculator might be useful this time.* For this assignment, do not include any interactions among explanatory variables, even though it's something you would investigate in a more thorough data analysis.

(a) The primary question is this. Controlling for all other variables, is birdkeeping related to the chance of getting lung cancer? Perform a likelihood ratio test as well as a Wald test. What do you conclude? Presence or absence of a relationship is not enough. Say what happened.

(b) All other things being equal, the estimated odds of cancer for a birdkeeper are _____ times as great as the estimated odds of cancer for a non-birdkeeper.

(c) For a birdkeeping woman of low socioeconomic status who is average on all the other independent variables, what is the estimated probability of lung cancer? This answer should be based on the full model.

(d) For a *non*-birdkeeping woman of low socioeconomic status who is average on all the other independent variables, what is the estimated probability of lung cancer? This answer should be based on the full model.

(e) There are lots of other meaningful questions. Be able to interpret the estimates and hypothesis tests for the other variables, too. For example, controlling for all other variables, are the chances of cancer different for men and women?

(f) As another example, all other things being equal, when a person smokes 10 more cigarettes a day (ten, not one), the estimated odds of cancer are _____ times as great.

Again, please bring your `SAS` log file and list file to the quiz. Remember, the log file is not the same as a plain listing of the program. Please be sure that the log file and list file come from the same SAS run. *Do not write anything on your printouts in advance except maybe your name* (and student number, if you wish).