STA 2101/442 Assignment 7[1]

1. In the *Heart attack data* (which you will analyze later), a sample of middle-aged men who had heart attacks were classified into three groups. Either they died of the first heart attack, or they died during the next 10 years, or they were still alive 10 years after the first attack. This is the response variable. Potential explanatory variables include age, blood pressure, and family history of heart disease (Yes-No). Let's just consider these for now. For interpretability, make the probability of being alive 10 years later the denominator in each generalized logit.

   (a) Write the multinomial logit model for these data. How many generalized logits do you have? Of course you must have a regression equation for each one.

   (b) Solve for the probabilities in terms of the *beta* values in your model. Show your work.

   (c) Make a table with two rows, one for Family history = Yes, and one for Family history = No. In each row, write *two* probability ratios. Let's call then "relative risks." (The relative risk of dying in a particular way is the probability of dying that way divided by the probability of living.)

   (d) Controlling for age and blood pressure, the relative risk of dying in the first heart attack is ________ times as great for those with a family history of coronary heart disease.

   (e) Controlling for age and blood pressure, the relative risk of dying in the next 10 years after the first heart attack is ________ times as great for those with a family history of coronary heart disease.

2. The file `http://www.utstat.toronto.edu/∼brunner/data/illegal/attack.data.txt`. contains the *Heart attack data*, in which a sample of middle-aged men who had heart attacks were classified into three groups. Either they died of the first heart attack, or they died during the next 10 years, or they were still alive 10 years after the first attack. This is the response variable. Please make the probability of being alive 10 years later the denominator in your generalized logits.

   The variables are

   - AGE AT ENTRY TO STUDY
   - AVERAGE DIASTOLIC BLOOD PRESSURE
   - SERUM CHOLESTEROL
   - NUMBER OF CIGARETTES PER DAY (Self report)
   - HEIGHT IN INCHES
   - WEIGHT IN POUNDS
   - FAMILY HISTORY OF CORONARY HEART DISEASE
   - EDUCATION
   - OUTCOME

   Instead of height and weight, let's use Body Mass Index (BMI), defined as

   $$\text{BMI} = 703 \times \frac{\text{weight}}{\text{height}^2}.$$

   A BMI under 18.5 suggests that the person is underweight, while a value over 25 may indicate that the person is overweight. The first full model (the biggest one) will include all available explanatory variables, except that height and weight will be replaced by BMI.

[1] This assignment was prepared by Jerry Brunner, Department of Statistics, University of Toronto. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License. Use any part of it as you like and share the result freely. The LaTeX source code is available from the course website: `http://www.utstat.toronto.edu/~brunner/oldclass/appliedf18`

(a) Fit the model, meaning estimate the parameters.

  i. Test whether *any* of the explanatory variables are useful in predicting the response variable. This is one big test. Give the value of the test statistic, the degrees of freedom, and the p-value. The test statistic and p-value are on your printout, but the degrees of freedom are not. In plain language, what do you conclude?

  ii. We should probably just give up, but let's proceed anyway for practice. If there is any hope, it looks like a model with just age, cholesterol level, and family history of heart disease. So carry out a simultaneous test of all the other explanatory variables. What is your full model? What is your reduced model? Give the value of the test statistic, the degrees of freedom, and the p-value. In plain language, what do you conclude?

(b) Based on the results of the last test, I am willing to consider the model with just age, cholesterol level, and family history of heart disease. For that model, it is possible to reject the null hypothesis that the regression coefficients for all the explanatory variables equal zero? What is your full model? What is your reduced model? Give the value of the test statistic, the degrees of freedom, and the p-value. In plain language, what do you conclude?

(c) Now for this model with three explanatory variables, test each of the explanatory variables controlling for the other two. That's three tests. For each one, what is your reduced model? Give the value of the test statistic, the degrees of freedom, and the p-value. In plain language, what do you conclude?

(d) Overall, what is your assessment of this analysis?

3. Independently for $i = 1, \ldots, n$, let

$$y_i = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i + \epsilon_i$$

where

β_0 (the intercept) is an unknown scalar constant.

$\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown parameters.

$\mathbf{x}_i$ is a $k \times 1$ random vector with expected value $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}_x$.

ϵ_i is a scalar random variable with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$.

$cov(\mathbf{x}_i, \epsilon_i) = \mathbf{0}$.

Is $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\Sigma}}_x^{-1} \widehat{\boldsymbol{\Sigma}}_{xy}$ a consistent estimator of $\boldsymbol{\beta}$? Answer Yes or No and show the calculations. You may use the consistency of sample variances and covariances without proof.

4. This question explores the consequences of ignoring measurement error in the explanatory variable when there is only one explanatory variable. Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} Y_i &= \beta X_i + \epsilon_i \\ W_i &= X_i + e_i \end{aligned}$$

where all random variables are normal with expected value zero, $Var(X_i) = \phi > 0$, $Var(\epsilon_i) = \sigma^2_\epsilon > 0$, $Var(e_i) = \sigma^2_e > 0$ and ϵ_i, e_i and X_i are all independent. The variables W_i and Y_i are observable, while X_i is latent (unobservable, like true number of calories eaten). Error terms are never observable.

(a) What is the parameter vector $\boldsymbol{\theta}$ for this model?

(b) Denote the variance-covariance matrix of the observable variables by $\boldsymbol{\Sigma} = [\sigma_{ij}]$. The distribution of the observable data is completely determined by $\boldsymbol{\Sigma}$. Calculate the $\boldsymbol{\Sigma}$, expressed as a function of the model parameters.

(c) Here, identifiability means that the parameter can be recovered from $\boldsymbol{\Sigma}$ – that is, one can express the parameter as a function of the σ_{ij} values. Are there any points in the parameter space where the parameter β is identifiable? Are there infinitely many, or just one point?

(d) The naive estimator of β is $\widehat{\beta}_n = \frac{\sum_{i=1}^n W_i Y_i}{\sum_{i=1}^n W_i^2}$. Is $\widehat{\beta}_n$ a consistent estimator of β? Why can you answer this question without doing any calculations?

(e) Go ahead and do the calculation. To what does $\widehat{\beta}_n$ converge?

(f) Are there any points in the parameter space for which $\widehat{\beta}_n$ converges to the right answer? Compare your answer to the set of points where β is identifiable.