# STA 2201/442 Assignment 5

1. Men and women are calling a technical support line according to independent Poisson processes with rates $\lambda_1$ and $\lambda_2$ per hour. Data for 144 hours are available, but unfortunately the sex of the caller was not recorded. All we have is the number of callers for each hour, which is distributed Poisson$(\lambda_1 + \lambda_2)$. Here are the data, which are also available in the file `poisson.data` on the course website:

```
12  9 25 11 22 16 17  8 14 17 14 17  8 14 18 16 13 17 13 11  9 15 18
14 16 17 16 19 13 18 12 12 13 10  8 15 13 11 15 15  6 10 13 13 11 13
 9 15 16  9  5 10  8 18 13 17  7 13 13 17 12 17 14 16  6 12 17 10  9
14 11 19 13 17 15 20 14 10 13 14 17  9 13 14  7 16 16  9 25 10 10  9
17  7 15 12 14 21 14 18 14 12 13 15 12 11 16 14 15 16  8 19 13 17 15
11 18 13 12 11 19 14 16 17 13 13 19 19 11 19 10 12  9 18 11 14  9 14
14 14 13  9 13 18
```

   (a) The parameter in this problem is $\boldsymbol{\theta} = (\lambda_1, \lambda_2)'$. Try to find the MLE analytically. Show your work. Are there any points in the parameter space where both partial derivatives are zero?

   (b) Now try to find the MLE numerically with R's `nlm` function. The Hessian is interesting; ask for it. Try two different starting values. Compare the minus log likelihoods at your two answers. What seems to be happening here?

   (c) Try inverting the Hessian to get the asymptotic covariance matrix. Any comments?

   (d) To understand what happened in the last item, calculate the Fisher information in a single observation from the definition. That is, letting $\ell = \log f(Y; \boldsymbol{\theta})$, calculate the elements of the $2 \times 2$ matrix whose $(i, j)$ element is

$$-E\left(\frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j}\right)$$

   (e) The Fisher information in the sample is just $n$ times the Fisher information in a single observation. Using the numerical MLEs from one of your `nlm` runs, estimate this quantity (a $2 \times 2$ matrix). Compare it to the Hessian. Now do you see what happened when you tried to calculate the asymptotic covariance matrix?

   Most good homework problems have a lesson. The lesson here is that it's possible for a model to be perfectly be correct and the sample size to be large, but the data are still not adequate to allow successful estimation of the model parameters by maximum likelihood (or, it turns out, by any other method[1]). And the main clue is a Hessian matrix that is not positive definite.

[1] When two or more sets of parameter values give rise to exactly the same probability distribution for the observed data, using the data to decide which one is correct is a hopeless task, and all reasonable methods of estimation will fail. In such cases, the parameter vector is said to be *not identifiable*.

2. The usual univariate multiple regression model with independent normal errors is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{X}$ is an $n \times p$ matrix of known constants, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants, and $\boldsymbol{\epsilon}$ is multivariate normal with mean zero and covariance matrix $\sigma^2 \mathbf{I}_n$, with $\sigma^2 > 0$ an unknown constant. But of course in practice, the independent variables are random, not fixed. Clearly, if the model holds *conditionally* upon the values of the independent variables, then all the usual results hold, again conditionally upon the particular values of the independent variables. The probabilities (for example, $p$-values) are conditional probabilities, and the $F$ statistic does not have an $F$ distribution, but a conditional $F$ distribution, given $\mathbf{X} = \mathbf{x}$.

   (a) Show that the least-squares estimator $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is conditionally unbiased.

   (b) Show that $\widehat{\boldsymbol{\beta}}$ is also unbiased unconditionally.

   (c) A similar calculation applies to the significance level of a hypothesis test. Let $F$ be the test statistic (say for an $F$-test comparing full and reduced models), and $f_c$ be the critical value. If the null hypothesis is true, then the test is size $\alpha$, conditionally upon the independent variable values. That is, $P(F > f_c | \mathbf{X} = \mathbf{x}) = \alpha$. Find the *unconditional* probability of a Type I error. Assume that the independent variables are discrete, so you can write a multiple sum.

3. It is perfectly natural to assume that something like response to a drug might be approximately linear over some range of dosage values, but that each person in the population might have his or her own slope. Thus each time you select a random sample you'll get a different collection of slopes, and the regression coefficient corresponding to the slope would be a random variable. Here is a simple model illustrating this situation. Let

$$Y_i = B_i x_i + \epsilon_i,$$

where $x_1, \ldots, x_n$ are known constants, and independently for $i = 1, \ldots, n$,

   $B_i$ is a random variable with expected value $\beta$ and variance $\sigma^2_\beta$,

   $\epsilon_i$ is a random variable with expected value zero and variance $\sigma^2_\epsilon$, and

   $B_i$ and $\epsilon_i$ are independent.

   (a) What would happen if you tried to estimate $\beta$ in the usual way with

$$\widehat{\beta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}$$

   Under what conditions on the $x_i$ values is this estimator consistent?

   (b) Find another estimator of $\beta$ by calculating $E(\overline{Y}_n)$. Why does the Law of Large Numbers *not* apply here? Okay, anyway, propose an estimator, and describe the conditions on the $x_i$ values that will make it consistent for $\beta$.

(c) Now suppose that all the $x_i$ values are equal to one. To make things as easy as possible, assume everything is normally distributed.

i. What is the distribution of $Y_i$ in this situation?

ii. Propose an estimator of $\beta$ that should satisfy anyone.

iii. Give an *exact* $(1-\alpha)100\%$ confidence interval for $\beta$; you don't have to show any work.

iv. Now suppose that you want to estimate $\sigma^2_\beta$ and $\sigma^2_\epsilon$. Does Problem 1 tell you anything about your chances of success?

This last little example shows you two things. First, whether the parameters of a model can be estimated depends on how you collect the data; this is a matter of experimental design (assuming the $x_i$ values are under the control of the investigator). Second, it is possible that some parameters can be estimated very successfully, while others cannot be estimated at all.

4. For most configurations of $x_1, \ldots, x_n$, the variance parameters in Question 3 can be estimated successfully — but it's not so easy to see how. So we'll do it numerically with maximum likelihood. Again, suppose everything is normally distributed.

Some data from the model of Question 3 are available from the class website, in the file `randslope.data`.

(a) Make a scatterplot of the data and bring it to the quiz. Does it look funny? You're guaranteed that the model is correct. *Why* does the scatterplot look the way it does? How would it look if there were also a range of negative $x_i$ values?

(b) Estimate the parameters numerically. Your answer to this part is a set of three numbers. Show the definition of the function you're minimizing, as well as all the other input and output leading to your answer. Wondering about starting values? Well, at least you know where to start looking for $\widehat{\beta}$.

(c) Carry out a simple 2-sided $Z$-test of $H_0 : \beta = 0$. Your output should include the computed value of $Z$ and the two-tailed $p$-value. Do you reject $H_0$ at $\alpha = 0.05$?

5. The slides on Wald-like tests have been fixed up a bit. Please take a look at the most recent version.

(a) What, apart from the focus on computer ownership, is the connection between the first example of a Wald-like test, and Question 4 of Assignment 4?

(b) In both examples of Wald-like tests, the objective is to test for differences among two or more expected values. In terms of the structure of the data, what is the main difference between the two examples?