Stat-340/341/342 – Assignment 10
2015 Spring Term

Part I - Making babies via Maximum Likelihood Estimation

In this part of the assignment, you will learn how to:

- create function to evaluate the log-likelihood for a problem;
- plot the log-likelihood against the single parameter to see where the MLE occurs;
- find the MLE numerically, along with an estimate of precision;
- find confidence intervals for MLE based on large sample theory;
- find profile confidence intervals in the case of a single parameter.

For many couples, it is a joyful moment when they decide to try to and have children. However, not every woman becomes immediately pregnant on the first attempt to conceive a child and it may take many months before the woman becomes pregnant.

Fertility scientists are interested in estimating the probability of becoming pregnant in a given month with unprotected intercourse. A sample of couples are enrolled in a study and each couple records the number of months prior to becoming pregnant. Here is the raw data:

> 2; 7; 5; 0; 0; 3; 0; 4; 10+

where the value of 2 indicates that the couple became pregnant on the 3rd month (i.e. there were two months where the pregnancy did not occur PRIOR to becoming pregnant on the 3rd month). The value 10+ indicates that it took longer than 10 months to get pregnant but the exact time is unknown because the experiment terminated.

If the exact value were known for all couples (i.e. including the last couple), then the sample average time (in months) PRIOR to becoming pregnant would be a simple estimator for the average number of months PRIOR to becoming pregnant, and the intuitive estimator for the probability of becoming pregnant on each month would be $\frac{1}{1+\overline{Y}}$. The extra '1' in the denominator accounts for the fact that you become pregnant in the NEXT months after being unsuccessful for Y months. For example, if the average time PRIOR to becoming pregnant was 4 months, then the probability of becoming pregnant in each month is $1/(4+1) = 0.20$.

But what should be done if you have incomplete data (the last couple). This is an example of censored data[1] where there is information, but it is not obvious how to use it. Just using the value of 10 for the last couple will lead to an underestimate of the average time to become pregnant and an overestimate of the probability of becoming pregnant in a month. You could substitute in a value for the last couple before computing an average, but there is no obvious choice for a value to use – should you use 11, 12, 15, 27, etc.?

This problem is amenable to Maximum Likelihood Estimation. A common probability distribution to model this type of data is the geometric distribution with parameter p representing the probability of becoming pregnant in any month.

Let Y be the number of months PRIOR to becoming pregnant. Then

$$P(Y = y|p) = (1-p)^y \times p$$

i.e. there are y "failures to get pregnant" followed by a "success". For censored data, we add together the probability of becoming pregnant over all the months greater than or equal to the censored value:

$$P(Y \geq y|p) = \sum_{i=y}^{\infty} (1-p)^i \times p$$

[1] Another example of censored data are water quality readings where the concentration of a chemical is below the detection limit (denoted as < 25) where the detection limit (25) provides an upper bound on the actual concentration.

which after some algebra reduces to

$$= (1-p)^y$$

We can now construct the likelihood function as the product of the individual terms:

$$L = \prod_{\text{non-censored}} (1-p)^{Y_i} p \times \prod_{\text{censored}} (1-p)^{Y_i}$$

The log-likelihood is then:

$$log(L) = \sum_{\text{non-censored}} [Y_i log(1-p) + log(p)] + \sum_{\text{censored}} Y_i log(1-p)$$

Note that the log(p) term is added FOR EVERY NON-CENSORED value.

If there were no censoring, the above equation can be used to derive the maximum likelihood estimator by finding the first derivate with respect to p, equating the first derivate to 0, and then solving for p. This gives

$$\widehat{p}_{\text{if no censoring}} = \frac{1}{\overline{Y}+1}$$

which has a nice interpretation as the average months prior to becoming pregnant + 1 month when you became pregnant.

Unfortunately, in the presence of censoring, there is NO explicit solution and the MLE MUST be solved numerically.

1. Create a data frame that has two variables – the months PRIOR to becoming pregnant and a variable indicating if the data value is censored with 0/1 indicating if the data value is not/is censored.

2. Create a function that computes and returns the log-likelihood of the above data. It should start:

   ```
   preg.in.month.log.lik <- function(p, data, return.negll=FALSE){
   #  Compute the log-likelihood for becoming pregnant in a month
   #  The return.negll returns the negative of the log-likeiihood if set to TRUE
      ...
       if(return.negll){log.like <- -log.like} # if want the negative of the log likelihoo
       return(log.like)
   }
   ```

Why do we want the *return.negll* argument?

Note that the FIRST argument to your function MUST be the value at which you want the likelihood function to be computed so that the ... argument in the optimization functions can be used to pass the other arguments (i.e. in this case, the data) to your function..

There is NO need to use a *for* loop in the function – remember that most *R* functions are vectorized. Also use the "trick" of

$$...expression + (logical\ expression) * (second\ expression)$$

to add the second expression to the result only when the logical expression is true. For example, suppose you want to add 5 only when the value is more than 20. Then you would have

$$result < -values + (values > 20) * 5$$

where *values* is a VECTOR.

Test your function to see that it is computing the log-likelihood properly.

```
 %2015 values
> preg.in.month.log.lik (.2, my.data)
[1] -19.79295
> preg.in.month.log.lik (.2, my.data, return.negll=TRUE)
[1] 19.79295
```

3. Use a function from the *plyr* package to compute the log-likelihood between 0 and 0.5 in steps of .01 and plot the log-likelihood function. Where does the maximum appear to be?

   Hint: Use the *seq(start,end,by)* to generate the points between 0 and 0.5 and don't forget the ... option in the *ldply()* function to pass the data to your function.

4. Use the *nlm()* function to find the MLE. Note that *nlm()* function finds the location that results in the MINIMUM of the function, so you need to tell your log-likelihood function to return the NEGATIVE of the log-likelihood (Why?)

   Read the help file for the *nlm()* function to see where the additional arguments (the ...) appear.

   Don't forget to specify the *hessian=TRUE* argument so you get back the estimated variance-covariance matrix of your estimates. Again, don't forget to pass your data to your (modified log-likelihood) function using the ... argument.

5. Add the MLE to the previous plot by plotting a filled circle at the maximum of the likelihood and the returned MLE. Hint: don't forget that the

nlm() function used the NEGATIVE of the log-likelihood function. The different plotting symbols are found by reading the help on the *pch* attribute of the *plot()* function in Base *R* graphics and the *shape* argument in the *ggplot()* function.

6. The standard error of an MLE is formally estimated by the square root of the diagonal elements of the inverse of the negative of the matrix of second derivatives (the hessian), but the hessian returned by *nlm()* is based on the negative of the log-likelihood (why?) and so you don't need to take the negative of the *$hessianz* component returned by the *nlm()* function (why?)

7. Find the asymptotic 95% confidence interval by taking the $mle \pm z \times se$ and plot these points on the graph using a open circle. You should compute the z value using the *qnorm()* function, and not simply plug in the value of 1.96. Notice you can find the point on the curve which is the maximum using the results from your optimization of the log-likelihood function. Hint: don't forget that you modified you log-likelihood function to return the negative of the log-likelihood.

8. You will notice from the plot that while the upper and lower asymptotic confidence interval endpoints are symmetric about the MLE (why?), they are not at the same relative distance from the maximum of the likelihood function (i.e. they are not the same distance below the maximum).

9. Profile confidence intervals are an alternate way to find confidence intervals that use the log-likelihood function directly. A likelihood ratio test computes the test-statistic as twice the value of the difference between the log-likelihood at the maximum and at a specified point. This is compared to a chi-square distribution with 1 degree of freedom (in the case of testing a single parameter).

   Load the *bbmle* package and repeat the optimization using the *mle2()* function. You should get the same results as your direct optimization above.

   The *mle2()* function returns an object for which many methods are available – in particular the *coef()*, the *vcov()*, and the *confint()* methods.

10. Plot the profile interval endpoints on your graph as well. Notice that the interval is no longer symmetric about the MLE but both points (the upper and lower confidence bounds) are the same same distance below the MLE peak.

11. Put the actual value of the MLE, its se, and the two confidence intervals on the graph. Don't forget to round these values to a sensible number of decimal places.

You may be curious to know that the commonly "accepted" value for the probability of becoming pregnant in a month with unprotected intercourse is about 25%. This value was obtained using methods very similar to what was shown above. For information on the current "state of the art" for these types of studies, see:

Scheike, T.H. and Keiding, N. (2006).
Design and analysis of time-to-pregnancy.
Statistical Methods in Medical Research, 15, 127-140.
`http://dx.doi.org/10.1191/0962280206sm435oa`

Hand in the following using the electronic assignment submission system:

- Your *R* code that did the above analysis.
- An HTML file containing all of your *R* output.
- A one page (maximum) double spaced PDF file containing a short write up on this analysis explaining the results of this analysis suitable for a manager who has had one course in statistics. You should include the following:
  - A (very) brief description of the how the experiment was run.
  - Your graph of the likelihood function, that includes the value of the MLE, its standard errors, and the two confidence intervals..

  You will likely find it easiest to do the write up in a word processor and then print the result to a PDF file for submission. Pay careful attention to things like number of decimal places reported and don't just dump computer output into the report without thinking about what you want.

Part II - Never underestimate the p-o-w-e-r of the *Orange* side

Many people find it annoying when a cell phone goes off at the exact climax of a film.[2]

When I was visiting England in September 2005, I happened to go to a movie and noticed a series of ads that played before the movie started asking patrons to turn off their cell phone. The premise of these advertisements are pitches by various celebrities to the *Orange Film Funding Board*, a fictitious agency, for films they would like to produce. The ads were sponsored by the Orange Cell Phone company, formerly, one of the largest mobile phone companies in the United Kingdom.[3]

You can view some of the advertisements at (don't forget to press the Play button beneath each ad):

1. `https://www.youtube.com/watch?v=wbtlv2cImxM` - my favorite
2. `https://www.youtube.com/watch?v=L70pRx_eexo`
3. `https://www.youtube.com/watch?v=nTWrvb37i9s&index=2&list=PL69B11E7A51C47B15` - my second favorite
4. `https://www.youtube.com/watch?v=h9_5eh_lJ8g`

These advertisements have made it into Wikipedia at `http://en.wikipedia.org/wiki/Orange_UK` as *Orange Gold Spots*. But do these commercials actually work?

In this part of the assignment, you will learn how to:

- analyze the results from an experiment with a categorical response
- estimate population marginal parameters
- plot these items on a suitable plots.

[2] `http://www.cnn.com/2005/TECH/10/17/wireless.manners/index.html`
[3] More details at `http://www.orange.com/english/default.asp`.

1. How would you perform an experiment as a completely randomized design. The four ads are to be compared (with a control of no ads). There are 10 screens, five showings per day (morning, early afternoon, late afternoon, early evening, and late evening identified by the numbers 1 to 5), seven days per week (1=Sunday, 2=Monday, etc), and a 4 week test period.

2. You can download some data from `http://www.stat.sfu.ca/~cschwarz/Stat-340/Assignments/CellPhone/cellphone.txt`. The variables in the dataset are the week, day, showing, screen, ad used, number of tickets sold, and the number of cell phones that went off.

   Convert the number of cell phones that went off to a simple yes/no variable.

   Don't forget to check your recoding using the *xtabs()*function.

3. Reorder the *ad* factor to have the control group first so that the plots (below) have the control group on the left of the graph.

4. Fit a suitable model for this data. Hint: recall the the type of Y variable (continuous or categorical) is often helpful in deciding if the *lm()* or *glm()* function should be used.

5. Test the hypothesis that the probability of a cell phone interruption is the same for all ads (including the control).

   Recall that the *anova()* function should be used to test hypotheses. However, the default method for *glm()* objects doesn't give p-values. Read the help on the ANOVA method for *glm()* objects by looking at the help for the *anova.glm()* function. This convention (i.e. method name followed by a period followed by the class of the object) is often used where the general method (*anova()*) has specialized functions for different classes of objects (in this case the *glm* class).

   What do you conclude?

   Note that because there is ONLY one term in the model, the Type I and III tests are equivalent and so there is no problem in using the *anova()* function directly.

6. Estimate the probability of a cell phone interrupting the movie for each ad using the *lsmeans* package. Note that despite the name, the estimates are NOT MEANS! So, if they are NOT means, what do the methods in the *lsmeans* package return and how do you back transform them to a probability between 0 and 1.

   Convert the marginal estimates to the probability of an interruption. You CANNOT find the SE on the back-transformed scale by simply taking the back transform of the SE.

   Don't forget to control for multiplicity

7. Draw a suitable graph showing the results on the proportion scale. What does this graph show? Which ad seems to be the most effective?

8. Estimate the differences in the log-odds between the pairs of marginal population values and and an approximate 95% confidence interval. Hint: use the methods in the *lsmeans* package along with a suitable large sample result.

   Convert this to an odds ratio along with a 95% confidence interval. Interpret this odds-ratio.

   What do you conclude about the effect of heavy breathing on having all cell phones turned off relative to no advertisement.

   Truly the *The Phone is Strong Here.*

In more advanced courses, you will learn how to use the actual number of cell phone calls as the response variable and how to adjust it for the number of tickets sold for that showing. This technique, called Poisson Regression is typically used for count data.

Hand in the following using the electronic assignment submission system:

- Your *R* code that did the above analysis.
- An HTML file containing all of your *R* output.
- A one page (maximum) double spaced PDF file containing a short write up on this analysis explaining the results of this analysis suitable for a manager who has had one course in statistics. You should include the following:
  - A (very) brief description of the how the experiment was run.
  - Your graph of the results.
  - How much more effect is heavy breathing compared to doing nothing on persuading patrons to turn off their cellphones prior to a movie?

  You will likely find it easiest to do the write up in a word processor and then print the result to a PDF file for submission. Pay careful attention to things like number of decimal places reported and don't just dump computer output into the report without thinking about what you want.

Part 3 - Why you should check residuals - Challenging

I've mentioned several times the importance of model assessment in Statistics. One important part of model assessment is looking at residual plots.

In this part of the assignment, you will learn how to:

- create interesting residual plots.

Let us begin.

1. Download the dataset *interesting.txt* from the website at `http://www.stat.sfu.ca/~cschwarz/Stat-340/Assignments/Residuals/interesting2.csv`.
2. Read in the dataset using the (very) uninteresting variables Y, $X1$, etc.
3. Create a scatterplot matrix of all variables. What do you conclude based this plot?
4. Do a multiple regression of Y vs. all of the X variables.

   What do you conclude based on the output from the *summary()* method? Are you surprised given the results of the scatterplot matrix?
5. Extract the residuals and predicted values using the appropriate methods. Plot the residual against the predicted value. Use -2 to $+2$ for the range of the X axis. What do you conclude?
6. An interesting article on how to do this is found at:

   `http://www4.stat.ncsu.edu/~stefanski/NSF_Supported/Hidden_Images/Residual_Surrealism_TAS_2007.pdf` and `http://www4.stat.ncsu.edu/~stefanski/NSF_Supported/Hidden_Images/stat_res_plots.html`.

   Who says that Statisticians are no fun!

Hand in the following using the online submission system:

- Your *R* code.
- An HTML files containing the output from your *R* program.
- The residual plot and your interpretation of the plot.

Part IV - Integration of course material

View the lecture at http://www.youtube.com/watch?v=jbkSRLYSojo and be prepared to explain the concepts outlined on the final exam.

Comments from the marker

Part I

- Most people got the MLE and the asymptotic CI.
- Most of the deductions were for not including a caption on figures.
- Another point that was missed by many was discussing which CI is more appropriate for this scenario.
- Some people described the analysis in terms of *R* packages - don't do this! There are many, many *R* packages. You can't assume that your reader knows what every function does. It is also totally unnecessary.
- Some submissions lost points for not clearly stating the parameter that is being estimated (probability of a pregnancy in a given month)
- Another big error was saying that the profile CI and the asymptotic CI are equivalent! This is not true, and you can see it in your results! If they are equivalent, they should give the same CI. They are asymptotically equivalent (as n grows, the profile CI approaches the asymptotic CI, hence the name.)
- When you're using data to estimate something, by far the MOST important thing to communicate in your writeup is what you are estimating. Seems kind of obvious, but a lot of people never explained what p was.

  How true... you will find that when you get into the "real world" that communication skills will make or break you. Technical skills are important, but those who can communicate their work (both orally and written) get the next promotion.
- Maximum Likelihood is a fascinating method, made even more fascinating by the fact that turns an estimation problem into a well-posed optimization problem, which is, in turn, a pleasure to solve with the use of computers. I'm very happy that many of you considered it an important enough method to include, in detail, in your writeups. But I'm a statistics nerd, and not your uber-important boss, so what pleases me would not please her.

  Regardless of whether you use a clever method in computing your results, the boss is usually interested in hearing one thing: results. The equation of the likelihood function ranks far, far, FAR below in importance to the golden phrase 'The probability of getting pregnant in any month is equal to *xxx* (*SE yyy*)." And yet I saw a lot of the former and not nearly enough of the latter.

- Most students got the two different CIs correctly, and some even went so far as to label them accordingly, "asymptotic" and "profile". And then most stopped right there.

  Now, as a boss, whenever I see two CIs for the same estimate, I immediately have several logical questions: why do I need two? what's the difference? and which one do I use? Those are the questions I sought answers to, and only very few actually answered them.

  Here is a good answer from a student to the second question:

  > "The asymptotic 95% ci is calculated based on the asymptotically normal assumption of MLE estimators which results in an interval in which the lower and upper limit are equal distance from the MLE estimator. The profile 95% ci is calculated based on the assumption that twice the difference between the log-likelihood at the maximum and a specified point follows chi-square distribution with 1 degrees of freedom, which results in an interval in which the two limits are the same height below the MLE peak. "

  And this explanation would have been perfect, if it were accompanied with the mention that the profile CI is preferred for small samples such as this ($n = 9$), but for larger samples the two CIs are expected to be the same.

  In your more advanced courses in Statistics you will find that there are actually several ways to compute confidence intervals:

  - the asymptotic interval (mle +/- z se)
  - the profile interval (uses the likelihood function and can be computationally expensive
  - Wald intervals based on the score function (the first derivative of the log-likelihood).

  As a statistician, you need to know, as the marker pointed out, when to use each of the intervals and their advantages and disadvantages.

Part II

- Almost everybody got the correct results for this part, but sometimes the interpretation was lacking.
- Most deductions were for not quantifying the effect of "heavy breathing" (the Darth Vader ad). This could've been quantified using an odds ratio comparing the odds of a call when no ad was played compared to the odds of a call when the Darth Vader ad was played.

- Lots of students said that the Darth Vader ad resulted in the lowest probability, which is technically true, but there is no evidence that the Darth Vader is ad is more effective than the John Cleese or Steven Segal ads.

- There were lots of deductions for not including a caption for figures.

- Once again, introduction is important – it lets the person know what the data is all about, and the questions that could theoretically be posed in context. So, when I see an introductory sentence that says something to the effect of "we study how ads influence cell phone use", I get a very insufficient information regarding where or why the data was collected. Are these ads for cellphones? Are we measuring the minutes per month people spend on their cells? I took points off whenever it wasn't clear to me exactly what the experiment entailed.

- A fair number of students (a large number considering this is towards the end of the course) had plotted values without their accompanying CIs. Naked values in a plot, even if they're accompanied by correct conclusions, are a really REALLY bad thing. There is a huge possibility of misleading people when you report naked values. I hope that everyone knows better now!

  ## NO NAKED ESTIMATES; NO NAKED P-VALUES!

- Another common mistake was drawing incorrect conclusions from the data. The most common incorrect conclusions were:

  - "*Darth Vader* is the most effective ad' – incorrect, as there is no evidence that it is more effective than the *Steven Segal* or *John Cleese* ads
  - "having ads is better than not having ads" – incorrect, as there is no evidence that *Daryl Hannah* ad is more effective than no ads.

  The moral of this is, CIs in the plots are not just for decoration: they provide valuable information regarding whether the differences in the plotted data are statistically significant or not.