Homework1

Conceptual Exercises

1. Suppose we plan to use data to estimate one parameter, p_B.

- **When using a likelihood to obtain an estimate for the parameter, which is preferred a large or a small likelihood value? Why?**
  - A relatively large likelihood value, because under the fit model our probability of seeing the data we did in fact see will be larger.
- **The height of a likelihood curve is the probability of the data for the given parameter. The horizontal axis represents different possible parameter values. Does the area under the likelihood curve for an interval from .25 to .75 equal the probability that the true probability of a boy is between 0.25 and 0.75?**
  - No, as stated the likelihood is the probability of the data for a certain parameter, not the probability of the parameter. So certainly the integral would not be a probability for a range of parameter values. However, while the likelihood curve is not a PDF, it seems intuitive that its integral should mean something.

2. Suppose the families with an "only child" were excluded for the Sex Conditional Model. How might the estimates for the three parameters be affected? Would it still be possible to perform a Likelihood Ratio Test to compare the Sex Unconditional and Sex Conditional Models? Why or why not?

Removing the families with an "only child" from the Sex Conditional Model lowers the amount of $p_{B|N}$ terms multiplied into the likelihood function. Each parameter is estimated by solving for the maximum of the log likelihood function, which is linear. So the only parameter affected would be $p_{B|N}$ itself.

In the NLSY there are 930 "B" families, and 951 "G" families. Removing these occurrences removes more girls than boys from the dataset. So we would expect the proportion of boys born into neutral families to increase, i.e. $p_{B|N}$ should inreas.

In fact, we can calculate exactly what the new parameter estimate would be. Removing the "only child" counts, we are left with the following log-likelihood function:

$$\begin{aligned} logLik(p_{B|N}, p_{B|B_{bias}}, p_{B|G_{bias}}) = 2231log(p_{B|N}) + 2168log(1-p_{B|N}) + \\ 1131log(p_{B|B_{bias}}) + 1164log(1-p_{B|B_{bias}}) + \\ 1124log(p_{B|G_{bias}}) + 973log(1-p_{B|G_{bias}}) \end{aligned}$$

Differentiating with respect to $p_{B|N}$ and solving yields $\hat{p}_{B|N} = \frac{2231}{2231+2168} = 0.5072$. This is larger than the previous $p_{B|N}$ of 0.5033, as expected.

A Likelihood Ratio Test would no longer be appropriate to compare the two models, as the maximum likelihood for the Sex Conditional Model would be smaller simply due to the function using less data, biasing the ratio.

Guided Exercises

2. In Case 1 we used hypothetical data with 30 boys and 20 girls. Case 2 was a much larger study with 600 boys and 400 girls. Consider Case 3, a hypothetical data set with 6000 boys and 4000 girls.

- **Use the methods for Case 1 and Case 2 and determine the MLE for $\mathbf{p_B}$ for the independence model. Compare your result to the MLEs for Cases 1 and 2.**
  - With 6000 boys and 4000 girls, the likelihood function for the Sex Unconditional Model becomes:

    $$Lik(p_B) = p_B^{6000}(1-p_B)^{4000}$$

    With such high exponents, the likelihoods becomes too small for numerical precision to capture. However, we can still use calculus to solve for the minimum. Taking the log, and solving for when the derivative is 0, gives $\hat{p}_B = 0.6$, the same estimate as in Cases 1 and 2.

- **Describe how the graph of the log-likelihood for Case 3 would compare to the log-likelihood graphs for Cases 1 and 2.**
  - The larger sample size results in less variation in our estimate. Thus the graph of the log-likelihood for Case 3 will be narrower around the estimate of 0.6.

- **Compute the log-likelihood for Case 3. Why is it incorrect to perform an LRT comparing Cases 1, 2, and 3?**
  -

    $$logLik(p_B) = 6000log(p_B) + 4000log(1-p_B)$$

    This log-likelihood is exactly the same as for the other cases, only the coefficients of the two terms have changed due to the difference in sample size. This leads to wildly different orders of magnitude in the likelihoods at the shared optimum estimate, $\hat{p}_B = 0.6$. Therefore a Likelihood Ratio Test comparing Cases 2 or 3 to Case 1 will always determine Case 1 to be superior.

    But even more fundamentally than that, the LRT is meant to be used for model selection. However here the model used in all three cases is actually the same: the Sex Unconditional Model. So in this situation the LRT serves no purpose.

3. Write out an expression for the likelihood of seeing our data (5,416 boys and 5,256 girls) if the true probability of a boy is:

a. $\mathbf{p_B = 0.5}$

Assuming that we are using the Sex Unconditional Model:

$$(0.5)^{5416} * (1-0.5)^{5256} = \frac{1}{2^{10672}}$$

b. $\mathbf{p_B = 0.45}$

$$(0.45)^{5416} * (0.55)^{5256}$$

c. $\mathbf{p_B = 0.55}$

$$(0.55)^{5416} * (0.45)^{5256}$$

d. $\mathbf{p_B = 0.5075}$

$$(0.5075)^{5416} * (0.4925)^{5256}$$

These values are very small, and experience numerical underflow.

- **Compute the value of the log-likelihood for each of the values of p_B above.**

  Using base-2 logs:

a. $\mathbf{p_B = 0.5}$

$$5416log(0.5) + 5256log(0.5) = -10672$$

b. $\mathbf{p_B = 0.45}$

$$5416log(0.45) + 5256log(0.55) \approx -10772.5$$

c. $\mathbf{p_B = 0.55}$

$$5416log(0.55) + 5256log(0.45) \approx -10726.2$$

d. $\mathbf{p_B = 0.5075}$

$$5416log(0.5075) + 5256log(0.4925) \approx -10670.3$$

- **Which of these four possibilities, $p_B = 0.45$, $p_B = 0.5$, $p_B = 0.55$, or $p_B = 0.5075$ would be the best estimate of p_B given what we observed (our data)?**

  In MLE we always want to choose parameter values which maximize the probability of observing our data. Log is a strictly increasing function, so to maximize the likelihood, we should choose the maximum log likelihood value. Of the four p_B possibilities, $\hat{p_B} = 0.5075$ gives the largest log likelihood value, so it is our best estimate of p_B given our data.