

Political Science 150C/350C
Spring 2007
Problem Set 1

1. Verify that $\hat{\theta}$ maximizes the likelihood function $\mathcal{L}(\theta; y)$ if and only if it also maximizes the log of the likelihood function, $\ln\mathcal{L}(\theta; y)$.
2. A five member committee votes 3-2 in favor of a proposal. Assume voting is independent. Let π be the probability that a committee member votes for the proposal.
 - (a) We have no information with which to distinguish committee members (in the language of Bayesian statistics, we'd say that the committee members are *exchangeable*, but I digress). What is the maximum likelihood estimate (MLE) of π , the probability that any particular committee member votes for the proposal?
 - (b) What is the log-likelihood of $\pi = .5$? Compare this value of the log-likelihood function with that attained at the MLE with a likelihood ratio test. What does this say about the plausibility of $H_0 : \pi = .5$?
 - (c) How would your conclusion about the plausibility of $H_0 : \pi = .5$ change if we observed
 - i. a 10 person committee splitting 6-4 in favor of the proposal?
 - ii. a 50 person assembly splitting 30-20 in favor of the proposal?i.e., what is happening to the likelihood function and/or the log-likelihood function in these cases relative to the case of a five person committee? In particular, what is happening the 2nd derivative of the log-likelihood function in the neighborhood of the MLE?
3. Download the data on Al Gore's share of the 2-party vote for president by congressional district in the 2000 election, [available from my web site](#). Denote this variable y_i . Assume (perhaps implausibly), $y_i \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$.
 - (a) What are the MLEs of μ and σ^2 ?
 - (b) Use graphical means to verify that your estimate of σ^2 is in fact a MLE. That is, over a grid of values for σ^2 that includes the MLE of σ^2 , compute the *conditional* MLE for μ and note the value of the conditional log-likelihood function at the MLE of μ (conditional, in that for the purposes of this exercise, we are treating each assumed value for σ^2 on the grid as a known constant, instead of a parameter to be estimated). Plot the values of the log-likelihood against the assumed grid of values for σ^2 . Comment briefly on the shape of this *profiled* log-likelihood in the neighborhood of the MLE for σ^2 . Hint: in R, you may need to write a loop over the grid of assumed values for σ^2 , and you'll also find the `dnorm` function helpful.
4. Consider the *Titanic* survival data. These data are part of R, or [available from my web site](#). The data set contains 2,201 observations (for each known passenger and crew member) on 4 variables:

- `class`: class of service, 0 = Crew, 1 = 1st class, 2 = 2nd class, 3 = 3rd class.
 - `adult`: 1 if adult, 0 otherwise.
 - `male`: 1 if male, 0 otherwise.
 - `survive`: 1 if survive, 0 otherwise.
- (a) Cross-tabulate `survival` and `male`. Test the hypothesis that there is no association between survival and gender. In R, use the `chisq.test` function.
 - (b) Estimate the logistic regression of `survival` on `male`. In R, use the `glm` function, with the option, `family=binomial`. Use a z-test to test whether men were more or less likely to survive the disaster. Compare your results with the results from the cross-tabulation in the first part of this question.
 - (c) Repeat the test you performed in the previous question, but via likelihood ratio test. That is, estimate a 2nd, restricted `glm` and compare it with the unrestricted model. In R, use the `anova` command, with the `test="Chi"` option.
5. Download the file `nagler.asc` from [my web site](#). This file contains 98,857 cases (welcome to large n research!) from the 1984 Current Population Survey, analyzed by Jonathan Nagler in two articles: “The Effects of Registration Laws and Education on Voter Turnout” *American Political Science Review*, 1991, 85:1393--1405; ”Scobit: an alternative estimator to logit and probit” *American Journal of Political Science*, 1994, 38:230--255.

The data in the file comprise the following variables (in column order):

<code>turnout</code>	1 if the respondent reports turning out to vote in the 1984 presidential election, 0 otherwise.
<code>educ</code>	1 for 0-4 yrs education; 2 for 5-7 yrs; 3 for 8 yrs; 4 for 9-11 yrs; 5 for 12 yrs; 6 for 1-3 yrs college; 7 for 4 yrs college; 8 for 5+ yrs college
<code>age</code>	age of respondent, in years
<code>south</code>	1 if respondent line in the South, 0 otherwise.
<code>govelec</code>	1 if a gubernatorial election coincided with the presidential election
<code>closing</code>	number of days before election day that voter registration closes in the respondent’s state

The following questions ask to you to estimate a series of logistic regression models. Construct a publication-quality table with the parameter estimates and standard errors for each the models, along with some summary information (e.g., goodness-of-fit, deviance, etc).

- (a) Estimate a logit model predicting turnout with the predictors `educ` and `age` and the square of each of these predictors. Provide a brief write-up of the parameter estimates (i.e., assess statistical significance and substantive implications) and the goodness-of-fit of the logistic regression model.

- (b) How many unique predicted probabilities are produced by this model? Explain how you derived your answer.
- (c) Compare the predicted probabilities from the logit model with the corresponding predicted probabilities from a probit model. How and why do they differ, if at all? Is there any statistical basis for preferring logit over probit or vice-versa?
- (d) Augment your logit model from the first part of this question with the following **additional** “contextual” predictors: `south`, `govelec`, and `closing`, and *interactions* between the two education variables (`educ` and `educ2`) and the `closing` date variable (i.e., make the effects of closing date quadratically conditional on the categorical education measure). Discuss the estimates and goodness-of-fit of this model in contrast with those obtained from the model for the previous question. Report a likelihood ratio test of the joint significance of the new predictors.
- (e) Using the estimates from the second model, plot the implied coefficient for `closing` as a function of education, given the interaction effects estimated above. Overlay 95% confidence intervals around the point estimates. Offer a substantive interpretation of what this plot reveals.
- (f) Using the estimates from the second model, consider a hypothetical non-southerner, in a state without a gubernatorial election, who has 12 years of education and has the median age of a non-southerner with 12 years of education. Plot the predicted probability of turnout for this person, as the closing date requirement varies over the range of closing date requirements observed in non-southern states. Overlay 95% confidence intervals around the point estimates.
- (g) Using the estimates from the second model, consider a hypothetical non-southerner, in a state without a gubernatorial election, who has 5+ years of college and has the median age of a non-southerner with 5+ years of college. Plot the predicted probability of turnout for this person, as the closing date requirement varies over the range of closing date requirements observed in non-southern states. Overlay 95% confidence intervals around the point estimates. Briefly compare the answers from this question with those from the previous question.

Due in class, Monday, April 16, 2007.