Worksheet 2

Due Monday, February 2nd at 5 PM

Please refer to the standing instructions on Ed for submission guidelines.

1 Log-Odds

You've seen this function in logistic regression in Data 100. But here, let's work in the situation where are no predictor variables, just a sample of zeros and ones.

Return to the frequentist world and assume that p is fixed and in the interval $(0, 1)$. If an event A has probability p, then $\text{odds}(p) = p/(1-p)$ is called the *odds in favor of* A or simply the odds of A. We will call this function the *odds function*.

The *odds against* A, defined as the odds of A^c, is the same as $1/\text{odds}(p)$. In common usage, odds are most commonly expressed as ratios of whole numbers, e.g. "5 to 2" for odds(5/7).

(a) Suppose you and your friend bet on an event that has probability p. If the event occurs, you give your friend 1 dollar. If the event does not occur, your friend gives you x dollars. The value of x that makes the game fair is the one for which neither you nor your friend expects to make money. Find that value in terms of the odds function.

(b) What is the range of odds function? Is the function increasing, decreasing, or neither?

[Please don't differentiate. Write $\text{odds}(1-p)$ as a function of $1/p$. No further calculation should be needed.]

(c) Define the *logit* or *log-odds* function as $\text{logit}(p) = \log(\text{odds}(p))$. What is the range of the log-odds function? What is the relation between $\text{logit}(p)$ and $\text{logit}(1-p)$? Sketch its graph.

(d) Let $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli (p) and let $\bar{X}$ be the sample mean. Name at least three useful properties of $\bar{X}$ as an estimator of p.

(e) Create a plug-in estimator of $\text{logit}(p)$ based on $\bar{X}$. Assume that n is large and use the delta method to find the asymptotic distribution of the plug-in estimator.

(f) Construct an approximate 95% confidence interval for $\text{logit}(p)$ when $n = 100$ and the observed value of $\bar{X}$ is 0.8. Explain where the approximations are.

2 The Exponential Rate, Revisited

Let $X_1, X_2, \ldots, X_n$ be i.i.d. exponential with rate λ_0. The density is $f(x \mid \lambda) = \lambda e^{-\lambda x}$ for $x > 0$. You know that the MLE of λ_0 is $\hat{\lambda}_n = 1/\bar{X}_n$.

(a) Assume n is large. Last week, we used the delta method to show why the distribution of $\hat{\lambda}_n$ is approximately normal. Now perform the delta method calculation to find the parameters of the asymptotic distribution.

(b) Now use the MLE theory of this week to see whether or not it agrees with the delta method result. That is, find the score function and the Fisher information, and see if you end up with the same normal approximation as in part (a).

3 MLE and Cross-Entropy Loss

Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables, and for each i, let Y_i have the Bernoulli (p_i) distribution. This is the model in logistic regression, with p_i equal to the sigmoid function evaluated at a linear combination of the predictor variables of individual i. For our purposes, all we need to remember is that each indicator has a different success probability.

(a) Write the likelihood function $\text{Lik}(p_1, \ldots, p_n)$.

[Refer to Lecture 3 for a useful expression for the likelihood function of a Bernoulli variable.]

(b) The maximum likelihood estimates of $p_1, \ldots, p_n$ must maximize the log-likelihood function $L(p_1, \ldots, p_n)$. Write the log-likelihood function.

(c) In Data 100, logistic regression is performed by minimizing the empirical risk based on cross-entropy loss. That empirical risk is given by

$$R(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \left(y_i \log(\sigma(\mathbb{X}_i \cdot \theta)) + (1 - y_i)(\log(1 - \sigma(\mathbb{X}_i \cdot \theta))) \right)$$

Compare this procedure with the MLE procedure described in part (b).

4 Jensen's Inequality

We used a version of this in the seminar to show the consistency of the MLE.

Let X be a random variable and g a convex function.

(a) Show that $\text{E}(g(X)) \geq g(\text{E}(X))$. This is *Jensen's inequality*, widely used in statistics.

[Start by sketching a graph of g and drawing the tangent line at the point $\text{E}(X)$. Compare the graph and the line, and remember how we started the proof of Markov's inequality in Data 140.]

(b) What does Jensen's inequality imply for a concave function g? Prove your answer.

(c) Compare the following.

(i) $\text{E}(X^2)$ and $(\text{E}(X))^2$.

(ii) $\text{E}(|X|)$ and $|\text{E}(X)|$.

(iii) $\text{E}(1/X)$ and $1/\text{E}(X)$ for a positive random variable X.

(iv) $\text{E}(\log X)$ and $\log \text{E}(X)$ for a positive random variable X.

(d) In Data 140/EECS 126, you showed that if $X_1, X_2, \ldots, X_n$ are i.i.d., then the *sample variance*

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

is an unbiased estimator of the *population variance* $\text{Var}(X_1)$.

Step back from these details, and suppose V is any unbiased estimator of a variance parameter. Suppose you decide to use $\sqrt{V}$ as an estimator of the standard deviation. Discuss whether $\sqrt{V}$ is an unbiased estimator of the standard deviation. If you think it is biased, what can you say about the bias without making any further assumptions about the underlying probability model?

5 Pareto Shape Parameter

For a positive parameter θ, the density $f(x \mid \theta) = \theta x^{-(\theta+1)}$ for $x \geq 1$ defines the Pareto family with shape parameter θ. You can assume that $\theta > 2$ so that all the necessary variances exist. If X has density $f(x \mid \theta)$ then

$$\text{E}(X) = \frac{\theta}{\theta - 1} \qquad \text{and} \qquad \text{Var}(X) = \frac{\theta}{(\theta-1)^2(\theta-2)}.$$

You can check these facts by integration but you don't have to.

(a) Find the score function $S(\theta; x)$ and the Fisher information $I(\theta)$.

(b) Let $X_1, X_2, \ldots, X_n$ be i.i.d. with density $f(x \mid \theta_0)$. Find $\hat{\theta}_n$, the MLE of θ_0.

(c) For this part and those that follow, assume n is large. Find the approximate distribution of $\hat{\theta}_n$.

(d) Suppose $n = 625$ and the value of $\hat{\theta}_n$ is 2.7. Construct an approximate 95% confidence interval for θ_0.

(e) Propose two different approximately normal estimators for the Pareto mean $\theta_0/(\theta_0-1)$ based on $X_1, X_2, \ldots, X_n$. Each estimator should be unbiased or approximately unbiased. Find or approximate the variance of each estimator analytically (that is, without simulation or the bootstrap). We'll compare the estimators next week.