Ec240a, Fall 2023

Professor Bryan Graham

Problem Set 2

Due: November 9th, 2023

Problem sets are due at 5PM. The GSI will provide instructions on how to turn in your problem set. You may work in groups, but each student should turn in their own write-up (including a "printout" of a narrated/commented and executed Jupyter Notebook if applicable). Please also e-mail a copy of any Jupyter Notebook to the GSI (if applicable).

1 Loan problem

This problem is adapted from (that late) Gary Chamberlain's undergraduate class at Harvard. Consider a sub-population of borrowers. To make this problem more realistic you may imagine these borrows are homogenous in a vector of observable attributes. Let $Y = 1$ if a borrower repays their loan and $Y = 0$ if they default. We have

$$\Pr\left(Y = 1|\, \theta\right) = \theta, \quad \Pr\left(Y = 0|\, \theta\right) = 1 - \theta.$$

You work at Proxima Centauri Bank (PCB). PCB is the a largest bank on a generation starship with tens of thousands of passengers traveling to a planetary system 100 light-years from earth. If the borrower pays back the loan the gain to PCB is g, whereas if they default the loss to the bank is l. Expected profits therefore equal

$$g \Pr\left(Y = 1|\, \theta\right) - l \Pr\left(Y = 1|\, \theta\right) = g\theta - l\left(1 - \theta\right).$$

[a] Assume that θ is known. What is the minimal repayment rate (i.e., value of θ) such that is profitable to lend to this group of borrowers?

[b] Let $\mathbf{Y} = (Y_1, \ldots, Y_N)'$ be a vector of past repayment outcomes for a random sample of borrowers. For a given θ what is the ex ante probability of the event $\mathbf{Y} = \mathbf{y}$ (i.e., find an expression for the likelihood $\Pr\left(\mathbf{Y} = \mathbf{y}|\, \theta\right) = f\left(\mathbf{y}|\, \theta\right)$)?

[c] You have been asked to consider whether it is profitable to continue to lend to this subpopulation (you may base your decision on the dataset introduced in part [b]). You may take one of two actions

$$\mathcal{A} = \left\{a_1, a_2\right\}.$$

Action a_1 corresponds to continuing to approve loans for this subpopulation. Whereas action a_2 corresponds to no longer lending to this group (which yields a payoff of zero under all states of nature). Write down the loss function associated these two actions (i.e., an expression for $L\left(\theta, a_1\right)$ and $L\left(\theta, a_2\right)$). Assume that loss equals the negative of expected profits calculated 'as if' θ were known.

[d] You attended many applied microeconomics seminars as a graduate student. Based on this experience you decide it is best to "let the data speak". Specifically you decide that you will construct a decision rule which maps the data into actions: $d{:}\mathbb{Y} \rightarrow \mathcal{A}$. The data will speak and you, as the ultimate decision-maker, will decide. Here $\mathbb{Y} = \left\{0, 1\right\}^N$ is the set of possible repayment patterns in your sample. Define *risk* and

explain why it equals

$$\begin{aligned} R(\theta, d) &= \mathbb{E}\left[L(\theta, d(\mathbf{Y})) \middle| \theta \right] \\ &= \sum_{\mathbf{y} \in \{0,1\}^N} L(\theta, d(\mathbf{y})) f(\mathbf{y}|\theta). \end{aligned}$$

[e] Prior to boarding the starship you worked in a bank in Idaho. From this experience you formed a prior about θ with density $\pi(\theta)$. Using this prior show that average risk equals

$$\begin{aligned} r(\theta, d) &= \int R(\theta, d) \pi(\theta) \mathrm{d}\theta \\ &= \sum_{\mathbf{y} \in \{0,1\}^N} \left[\int L(\theta, d(\mathbf{y})) f(\mathbf{y}|\theta) \pi(\theta) \mathrm{d}\theta \right], \end{aligned}$$

and argue that you can solve for the average-risk-minimizing decision rule “sample-wise”:

$$d_0(\mathbf{y}) = \arg\min_{a \in \mathcal{A}} \int L(\theta, d(\mathbf{y})) \pi(\theta|\mathbf{y}) \mathrm{d}\theta,$$

where

$$\pi(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)\pi(\theta)}{\int f(\mathbf{y}|\theta)\pi(\theta)\mathrm{d}t}.$$

[e] Compute the posterior expected loss from making the loan. From your answer deduce the (average risk) optimal decision rule (i.e., the “Bayes’ rule”).

[f] Show that the likelihood can be written as

$$f(\mathbf{y}|\theta) = \theta^{s_N}(1-\theta)^{N-s_N}$$

with $S_N = \sum_{i=1}^N Y_i$. Assume further a prior on θ of

$$\theta \sim \mathrm{Beta}(\alpha_1, \alpha_2)$$

and hence show that

$$\theta|\mathbf{z} \sim \mathrm{Beta}(S + \alpha_1, N - S + \alpha_2).$$

What is the average risk optimal decision rule under this prior? Why does this decision rule only depend on the data through $S_N = s_N$?

2 Bayesian Bootstrap

For this part of the problem set you may find the article by Chamberlain & Imbens (2003) helpful. David Card’s Fisher-Schultz lecture is a useful overview of the literature on estimating the return to schooling (Card, 2001).

The file `nlsy97ss.csv` is in the problem sets folder on GitHub. This is a comma delimited text file It includes a measure of average annual earnings (`avg_earn_2014_to_2018`), years of schooling (`hgc_ever`), ‘AFQT’ score (`asvab`), a female dummy and two ethnicity dummies for a sub-sample of respondents in the

National Longitudinal Survey of Youth 1997 cohort. Earnings equals average annual earnings over the 2014, 2016 and 2018 calendar years in 2012 prices. Define LogEarn to be the natural logarithm of Earnings.

1. Construct a sub-sample of non-black, non-hispanic, non-female respondents with positive earnings. Construct the LogEarn variable. Create a table of summary statistics for avg_earn_2014_to_2018, LogEarn, hgc_ever and asvab for this sub-sample.

2. Compute the least squares fit of LogEarn onto a constant and hgc_ever. Report the point estimate on the schooling variable as well as it heteroscedastic robust asymptotic standard error (you may use the StatsModels implementation of OLS to do this; later in the course we will construct our own program for these calculations).

3. Compute the least squares fit of LogEarn on a constant, hgc_ever and asvab. Does the estimate coefficient on hgc_ever change?

4. Estimate the parameters of the following linear regression model by the method of least squares

$$\mathbb{E}^*[\texttt{LogEarn}|\,\texttt{X}] = \alpha_0 + \beta_0\texttt{hgc\_ever} + \gamma_0\texttt{hgc\_ever} \times (\texttt{asvab} - 50) + \delta_0\texttt{asvab}$$

   where $X = (\texttt{hgc\_ever}, \texttt{hgc\_ever} \times (\texttt{asvab} - 50), \texttt{asvab})$'.

   (a) Provide a semi-elasticity interpretation of β_0.

   (b) Provide a semi-elasticity interpretation of $\beta_0 + \gamma_0 (\texttt{asvab} - 50)$.

5. Construct a plot with the OLS estimate of $\beta_0 + \gamma_0 (\texttt{asvab} - 50)$ on the y-axis and a grid of asvab values on the x-axis.

6. Using the Bayes' Bootstrap to approximate a posterior distribution for $\beta_0 + \gamma_0 (\texttt{asvab} - 50)$ at each value of asvab shown in your plot. Add (estimates of) the 0.025 and 0.975 quantiles, as well as the mean, of the posterior distribution of $\beta_0 + \gamma_0 (\texttt{asvab} - 50)$ to your plot.

7. Summarize what you have learned about the relationship between earnings, schooling and AFQT among white male millennials?

References

Card, D. (2001). Estimating the return to schooling: progress on some persistent econometric problems. *Econometrica*, 69(5), 1127 – 1160.

Chamberlain, G. & Imbens, G. W. (2003). Nonparametric applications of bayesian inference. *Journal of Business and Economic Statistics*, 21(1), 12 – 18.