MaCSS222, Spring 2025

Professor Bryan Graham

Problem Set 2

Due: March 3rd, 2025

Problem sets are due at 11:59PM. The reader 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 include a list of classmates you collaborated with when you turn in your problem set. Please also e-mail a copy of any Jupyter Notebook to the GSI (if applicable).

1 Logistic regression for college completion

The comma delimited text file **nlsy97ss.csv**, available on GitHub, will be used for this problem set. This is the same dataset used for Problem Set 1.

The variables of interest for what follows are schooling (**yrssch**), race (**black**, **hispanic**), sex (**female**) and Armed Services Vocational Aptitude Battery test score (**asvab**). While not accurate, as a convenient shorthand, I will reference to non-Black, non-Hispanic respondents as 'white' (in practice this group is a mixture of respondents with European, Asian, Native American and other ancestries).

1. Drop all units in the dataset with less than 12 years of schooling (i.e., we will confine our analysis to the population of high school graduates). Create a dummy variable for completing a 4-year college degree (**college**); this corresponds to 16 or more years-of-schooling. Drop any records with item non-response.

2. Construct a categorical variable for asvab *quintiles* (note code for a quartiles division is provided in the notebook stub, so you will need to modify this). Construct a cross-tab of **college**-by-**asvab** categories. Discuss this cross-tab? How does the frequency of college completion vary across sub-populations with different asvab scores? How does the distribution of asvab scores differ across college and non-college graduates?

3. Compute a logistic regression fit of **college** onto a **constant**, **female**, **black**, **hispanic** and **asvab.** Let $Y = \textbf{college}$, $X = (1, \textbf{female}, \textbf{black}, \textbf{hispanic}, \textbf{asvab})' = (1, X_1, X_2, \ldots, X_K)'$ for $K = 4$ and $\gamma = (\alpha, \beta_1, \ldots, \beta_K)'$:

$$\Pr\left(Y = 1 \ X = x\right) = \frac{\exp\left(x'\gamma\right)}{1 + \exp\left(x'\gamma\right)} = \frac{\exp\left(\alpha + \beta_1 x_1 + \cdots + \beta_K x_K\right)}{1 + \exp\left(\alpha + \beta_1 x_1 + \cdots + \beta_K x_K\right)}.$$

4. Use Efron's bootstrap to construct 95% confidence intervals for each of the coefficients entering your logit model (you may use the regular percentile bootstrap or the reverse percentile one). Report these confidence intervals in a table along with the coefficient estimates and correspond variable names (see notebook stub for some relevant code snippets).

5. Construct bootstrap standard error estimates as follows. Let $\left[\underline{\beta}_k, \bar{\beta}_k\right]$ be your $(1 - \alpha) \times 100$ percent

confidence interval for β_k. A bootstrap standard error is given by

$$\mathrm{se}\left(\hat{\beta}_k\right) = \frac{\left(\bar{\beta}_k - \underline{\beta}_k\right)}{\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) - \Phi^{-1}\left(\frac{\alpha}{2}\right)},$$

with $\Phi\left(\cdot\right)$ the CDF of a standard normal random variable.

6. Construct these standard errors for each of your logit coefficients and add them to your results table. Why is this a valid method of standard error construction (Hint: imagine you knew $\mathrm{se}\left(\hat{\beta}_k\right)$ and were asked to construct a $(1-\alpha) \times 100$ percent confidence interval for β_k, then work in reverse).

7. Verify the following derivative expression:

$$\frac{\partial \Pr\left(Y = 1\ X = x\right)}{\partial x_k} = \frac{\exp\left(\alpha + \beta_1 x_1 + \cdots + \beta_K x_K\right)}{1 + \exp\left(\alpha + \beta_1 x_1 + \cdots + \beta_K x_K\right)} \beta_k.$$

Argue that

$$\delta_k^{\mathrm{APE}} = \mathbb{E}_X\left[\frac{\partial \Pr\left(Y = 1\ X\right)}{\partial X_k}\right] = \mathbb{E}_X\left[\frac{\exp\left(\alpha + \beta_1 X_1 + \cdots + \beta_K X_K\right)}{1 + \exp\left(\alpha + \beta_1 X_1 + \cdots + \beta_K X_K\right)} \beta_k\right],$$

provides a summary estimate of the (average) effect of a unit increase in X_k on our *prediction* that $Y = 1$ given X. Why do we need to average over X to compute this effect, which we shall call the *average partial effect* (APE)?

8. Compute the estimate of the APE of **asvab** on **college**:

$$\hat{\delta}_k^{\mathrm{APE}} = \frac{1}{N}\sum_{i=1}^{N} \frac{\exp\left(\hat{\alpha} + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_K X_{iK}\right)}{1 + \exp\left(\hat{\alpha} + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_K X_{iK}\right)} \hat{\beta}_k$$

Construct a bootstrap 95 percent confidence interval as well as a bootstrap standard error estimate for δ_k^{APE} (here $k = 4$).

9. The State of California announces a plan to boost asvab scores of all high school graduates by 10 percentage points through an innovative college readiness program. The governor asks you to predict the likely effect of this program on college completion. You suggest two possible measures

$$\gamma_1^{\mathrm{CR}} = \mathbb{E}_X\left[\frac{\exp\left(\alpha + \beta_1 X_1 + \cdots + \beta_K X_K\right)}{1 + \exp\left(\alpha + \beta_1 X_1 + \cdots + \beta_K X_K\right)} \beta_k\right] \times 10$$

and

$$\gamma_2^{\mathrm{CR}} = \mathbb{E}_X\left[\frac{\exp\left(\alpha + \beta_1 X_1 + \cdots + \beta_K\left(X_K + 10\right)\right)}{1 + \exp\left(\alpha + \beta_1 X_1 + \cdots + \beta_K\left(X_K + 10\right)\right)}\right] - \mathbb{E}_X\left[\frac{\exp\left(\alpha + \beta_1 X_1 + \cdots + \beta_K X_K\right)}{1 + \exp\left(\alpha + \beta_1 X_1 + \cdots + \beta_K X_K\right)}\right].$$

Provide rationales for these two expressions. Construct estimates of them as well as bootstrap confidence intervals and standard error estimates. Do they differ? Why? Provide two reasons for why either one of these might be poor estimates of the effect of the college readiness program on changes in college completion.