Chapter 1

Reweighting Methods and the ”Propensity Score” Using Logistic Regression

Review from last session: We want to run regression on two different groups within our population. We run counterfactuals: What is the outcome for group A if we had the characteristics of group B (and vice-versa)? We can distill our measured gap into the gap between one counterfactual and another. This is a very popular pparoch in labor econmics because we take nonlinear functions and allows us to... Review **Oaxaca Decomposition**.

1.1 Re-weighted counterfactual

We have two groups a and b. Group a is the reference group.

$$\begin{aligned}\bar{y}^a &= \text{mean of outcome for group } a\\ \bar{y}^b &= \text{mean of outcome for group } b\\ N^a &= \text{num obs in group a}\\ \bar{p}^a_g - \frac{N^a_g}{N_a} &= \text{fraction of group } a \text{ in category } g\end{aligned}$$

We define the conunterfactual for group b if this group has the *same distribution across categories* as group a:

$$\bar{y}^b_{counterf} = \sum_g \bar{p}^a_g \bar{y}^b_g = (\bar{x}^a)' \hat{\beta}^b$$

We can erwright the counterfactual mean as a ”reweighted mean” where hte weight for a person i who is a category g is $w_i = w_{g(i)} = N^a_g / N^b_g$

$$\bar{y}_{counterf} = \frac{\sum_{i \in b} w_{g_i} y_i}{\sum_{i \in b} w_{g_i}}$$

Consider crazy regression for combined sample:

$$m_i = x_i'\theta + n\text{where } m_i = 1[i \in a] \text{ and } x_i' = (D_{1i}, D_{2i}, \ldots, D_{Gi})$$

We showed that:

$$w_{g(i)} = \frac{\hat{m}_i}{1 - \hat{m}_i} = \frac{N_g^a}{N_g^b}$$

When we stack up observations from 2 groups, and fit a model to rpedict who belongs to goup a given the covariates, the predicted porbability is called the *propensity score* or p-score for membership in a. So the crazy regression is just a model for the propensity score. Summary of reweighted conunterfactual method: 1. combine groups and fit a model for $p_i = p(m_i = 1|x - i)$

Method can be used even though the x's are not really discrete categories. Suppose for ample that we have info on eductation and age. We would not necsssarily want to divide the data into ’buckets’ with only one year of age and each possible value of education, b/c some of the buckets will be empty. We’d want to smooth across categories in some ways to perform the reweighting. It turns out that using a *flexible estimated propensity score* is the right approach. Fit a model for the propensity score that has dummies for each education, linear and quadratic terms in age, and interactions of the liner and quadaratic age terms with the education dummies.

Card’s rule of thumb: $\frac{N}{k} > 1000$. One problem we cna run into is that a simple linear model can predict values outside our (0,1) range. We want to run a logit model, thus, to get reasonable results.

A linear regression model for a 0/1 variable like m_i is known as a linear probability model. Given $m_i \in$ 0, 1,

$$E[m_i, x_i] = P(m_i = 1|x_i) = x_i'\theta$$

When x_i has a large range, a model like this is less attractive. So instead, assume that

$$E[m_i|x_i] = P(m_i = 1|x_i) = G(x_i'\theta)$$

for some function G that maps from $(-\infty, \infty)$ to (0, 1). Pick a distribution function; we will run with the logistic model.

1.2 Logistic Model

suppose there is a ”latex index”

$$m_i* = x_i'\theta + \epsilon_i\text{where } \epsilon_i \text{ is a r.v. with logit dist}$$

$m_i = 1 \leftrightarrow m_i^* > 0$. In this setting, m_i^* represents the tendency of person i to have $m_i = 1$.

$$\begin{aligned} P(m_i = 1|x_i) &= P(m_i * > 0|x_i) \\ &= P(x_i'\theta + \epsilon_i > 0) \\ &= P(\epsilon_i > -x_i'\theta) \\ &= 1 - G(-x_i'\theta) \\ &= G(x_i'\theta) \\ &= \frac{e^{x_i\theta}}{1 + e^{x_i\theta}} \end{aligned}$$

To find the optimal weights $\hat{\theta}$, we use the log-maximum-likelihood:

$$\frac{1}{N}\sum_i \log P(m_i|x_i, \theta)$$

$$\begin{aligned} P(m_i|x_i, \theta) &= G(x_i'\theta)^{m_i}(1 - G(x_i'\theta)^{m_i})^{1-m_i} \\ \log(P) &= m_i \cdot \log G(x_i'\theta) + (1 - m_i) \cdot \log(1 - G(x_i'\theta) \end{aligned}$$

$$\delta L/\delta\theta = \frac{1}{N}$$

counterfactual standard deviation:

$$\sigma^b_{counterf} = \frac{1}{N-1}\frac{\sum_{i\in b} w_i(y_i - \bar{y}^b_{counterf})^2}{\sum_{i\in b} w_i}$$

We can use this approach to estimate the effect of demographic changes on wage inequality.