Assessment 1 Review

February 22, 2023

Recap: Linear Probability Model Predictions

In last review session, we discussed a peculiar feature of the linear probability model, that it can lead to predicted values that are less than 0 or greater than 1.

This is a result of modeling assumptions on the conditional expectation of Y.

For example, suppose we are interested in the conditional probability of US military service given whether an individual is foreign- or US-born and their gender: $\Pr[Served_i = 1|Foreign_i, Female_i]$. $Foreign_i$ and $Female_i$ are both binary variables, so $\Pr[Served_i = 1|Foreign_i, Female_i]$ can only take one of 4 values, each corresponding to the probability of service for one pair of $Foreign_i$ and $Female_i$. We can estimate these probabilities using the conditional sample proportions without making any assumptions on the relationship between the variables.

We can also use regression to characterize the conditional expectation.

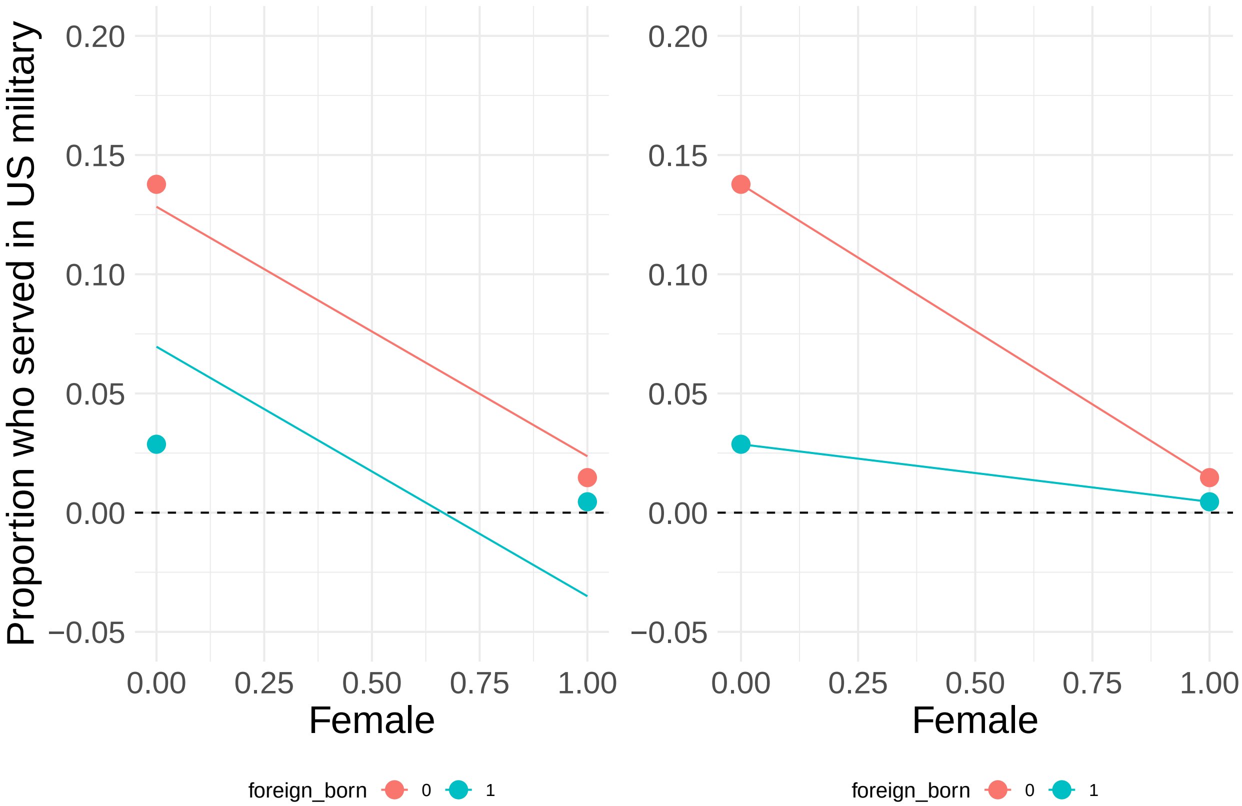
$$Served_i = \beta_0 + \beta_1 Foreign_i + \beta_2 Female_i + u_i \tag{1}$$

$$Served_i = \beta_0 + \beta_1 Foreign_i + \beta_2 Female_i + \beta_3 Foreign_i \times Female_i + u_i \tag{2}$$

In Regression 1, we get that $\widehat{Served}_i = -0.035$, a negative number! In Regression 2, we get a more reasonable positive prediction.

The below figure demonstrates what is going on: Regression 1 imposes that the relationship between military service and gender is the same regardless of the value of $Foreign_i$. This misspecification of the regression allows for predicted values outside of $[0, 1]$. On the other hand, Regression 2 allows the relationship between military service and gender to vary with the value of $Foreign_i$. The proportions are “pinned down” because the model allows for sufficient flexibility to capture all 4 conditional probabilities.

Note that this doesn’t mean that Regression 1 is a “bad” model. β_1 gives the average difference in service rates between the foreign- and US-born conditional on gender. However, we likely would not want to rely on Regression 1 for accurate predictions of military service because we can certainly do better than a negative probability.



Proportion who served in US military
0.20
0.15
0.10
0.05
0.00
−0.05
0.00
0.25
0.50
0.75
1.00
Female
Female
foreign_born
0
1
foreign_born
0
1

Review Questions

Frisch-Waugh-Lovell

Suppose we are interested in the relationship between years of education ($Educ_i$) and earnings ($Earn_i$). We also have data on years of experience: $Expr_i$. Consider the following SRFs:

$$Educ_i = \hat{\alpha}_0 + \hat{\alpha}_1 Expr_i + \hat{v}_i \tag{3}$$

$$Earn_i = \hat{\gamma}_0 + \hat{\gamma}_1 \hat{v}_i + \eta_i \tag{4}$$

1. What does the Frisch-Waugh-Lovell Theorem say about $\hat{\gamma}_1$?
2. Suppose we estimate Regression 3 and we find that $v_i = 0$ for all i. What would this mean for estimating Regression 4? What would it mean about the relationship between education and experience?

Interpretation of Non-linearities

In the United States, there is massive variation in health care spending across geographies. Consider the following sample regression functions that model a state's health care spending per capita, $HealthSpendPercap_i$ as a function of the state's average age, $AvgAge_i$. Let $Expand_i$ be a dummy variable that equals 1 if a state expanded its Medicaid program under the Affordable Care Act and 0 otherwise.[1]

$$HealthSpendPercap_i = -20000 + 1000 AvgAge_i - 8 AvgAge_i^2 + \hat{u}_i \tag{5}$$

$$\log(HealthSpendPercap_i) = 7.5 + 0.035 AvgAge_i + \hat{u}_i \tag{6}$$

$$\log(HealthSpendPercap_i) = 7.5 + 0.04 AvgAge_i + 0.2 Expand_i - 0.004 Expand_i \times AvgAge_i + \hat{u}_i \tag{7}$$

1. In Regression 5, health care spending per capita is **increasing/decreasing** at an **increasing/decreasing** rate.
2. What is the change in predicted health care spending per capita that is associated with a one year increase in average age in a state with average age of 40?
   - a. In Regression 5.
   - b. In Regression 6 — you can give the percent change if applicable.
3. In Regression 7, what is the predicted percent change in health spending per capita that is associated with a one year increase in average age in a non-expansion state? An expansion state?

R-squared

1. If you add a new covariate to a multiple regression, the R-squared will **increase/decrease/stay the same**?
2. Suppose we estimated the following regression: $Y_i = \beta_0 + \beta_1 Y_i + u_i$. What is the R-squared?

[1] Note that these coefficients are all made up!

Standard errors

Suppose you partner with a film studio to study the effect of movie director quality on actors' labor market outcomes. You estimate the following regression:

$$Quote_{ij} = \beta_0 + \beta_1 IMDB_j + u_{ij}$$

For actor i and director j, $Quote_{ij}$ is the actor's rate. This is the amount they get paid even if they do a bad job. It depends on the director's IMDB score and an error term u_{ij}. Many actors work with each director and other director-specific factors may impact actors' rates similarly.

1. What kind of standard errors should you use?
2. Suppose you use a random sample of 1000 actors working for the top 50 directors on IMDB, but your estimates are imprecise. Which will likely improve the precision of your estimate more, doubling the number of actors or doubling the number of directors?

Types of average treatment effects

You have been put in charge of a randomized evaluation of a micro-finance program. After recruiting a sample of 1000 study participants, you randomly give 500 of them access to micro-credit ($D_i = 1$). The rest of the participants do not have access to micro-credit ($D_i = 0$). After 3 months, you measure the weekly earnings of all participants, Y_i.

1. Let Y_{1i} be i's weekly earnings if they have access to micro-credit and Y_{0i} be their earnings if they do not. In words, what is $E[Y_{1i} - Y_{0i} | D_i = 1]$? Can you estimate it in this setting?
2. You claim to be able to estimate the average effect of treatment on the untreated (ATU). Your colleague claims this is impossible because none of the untreated receive micro-credit and therefore their treated potential outcomes are unobservable. Who is correct?