Stats 201A: Final Study Guide

Melody Y. Huang

Note: I am many magnitudes more tired for the final than I was for the midterm, which means this will be a lot shorter because I have less energy to expend pressing keys on my keyboard.

1. Regression

The general summary of regression is that we are often times interested in something called the conditional expectation function, which is the best predictor of some Y, given some X. We often cannot calculate or estimate the CEF, so we have to estimate it. This is done using something known as the best linear predictor (BLP), which takes the form of:

$$Y = \alpha + \beta X + \varepsilon$$

The best linear predictor will be the best linear predictor of the CEF. [1] The issue is that BLP relies on estimating the true values of α and β from the population. We need a sample analog of this. This is going to be what we usually refer to as our ordinary least squares regression (OLS).

$$\hat{Y} = \hat{\alpha} + \hat{\beta} X$$

TL;DR: We want to estimate a conditional expectation function to tell us what Y should be, given some X values. We approximate this using a linear model. The sample analog of the linear model is OLS.

1.1. Assumptions

There are a lot of assumptions that we have to impose within the regression framework. These assumptions build on properties that are inherent to the conditional expectation function and the best linear predictor.

To begin, we know that the conditional expectation has exogeneity (i.e., $\mathbb{E}(\varepsilon|X) = 0$). This is because:

$$\begin{aligned}
\mathbb{E}(\varepsilon|X) &= \mathbb{E}(Y - \mathbb{E}(Y|X)|X) \\
&= \mathbb{E}(Y|X) - \mathbb{E}(\mathbb{E}(Y|X)|X) \\
&= \mathbb{E}(Y|X) - \mathbb{E}(Y|X) \\
&= 0
\end{aligned}$$

[1] Yes, the redundancy is not lost upon me there...

This is great because with exogeneity, we have the implied properties of orthogonality ($\mathbb{E}(\varepsilon X) = 0$), zero correlation ($cov(X, \varepsilon) = 0$), and zero average error ($\mathbb{E}(\varepsilon) = 0$). Therefore, it is a little misleading to put them as separate properties, but I will list them as such to be consistent with other material.

As a summary:

Properties of the Conditional Expectation Function

Property 1. Exogeneity - $\mathbb{E}(\varepsilon|X) = 0$

Property 2. Orthogonality - $\mathbb{E}(\varepsilon X) = 0$

Property 3. Zero average error - $\mathbb{E}(\varepsilon) = 0$

Property 4. Zero correlation - $cov(X, \varepsilon) = 0$

With the linear model, we *lose the property of exogeneity*. In other words, we cannot assume that the conditional expectation of our error term is zero. However, we still have orthogonality, zero average error, and zero correlation:

Properties of the Best Linear Predictor

Property 1. Orthogonality - $\mathbb{E}(\varepsilon X) = 0$

Property 2. Zero average error - $\mathbb{E}(\varepsilon) = 0$

Property 3. Zero correlation - $cov(X, \varepsilon) = 0$

With ordinary least squares, we impose several assumptions.

Assumption 1. Linearity - the data generating process will be linear (i.e., $Y = \alpha + \beta X + \varepsilon$)

Assumption 2. Exogeneity - $\mathbb{E}(\varepsilon|X) = 0$
Note once again that this implies orthogonality and zero correlation.

Assumption 3. Spherical Error Variance
1. Homoskedasticity - $\mathbb{E}(\varepsilon_i^2|X) = \sigma^2$
2. No serial correlation - $\mathbb{E}(\varepsilon_i \varepsilon_j|X) = 0$

I have simply made a table of everything.

	Estimate	$\mathbb{V}(\cdot)$	$\hat{\mathbb{V}}(\cdot)$
β	$\frac{cov(X,Y)}{\mathbb{V}(X)}$	—	—
$\hat{\beta}$	$\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$	$\frac{\sigma^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$	$\frac{\frac{1}{n-2}\sum_{i=1}^{n}e_i^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$
	$(X^\top X)^{-1}X^\top Y$	$\sigma^2(X^\top X)^{-1}$	$\frac{1}{n-k}\cdot e^\top e(X^\top X)^{-1}$
α	$\mathbb{E}(Y)-\beta\mathbb{E}(X)$		
$\hat{\alpha}$	$\bar{Y}-\hat{\beta}\bar{X}$	$\sigma^2\left(\frac{1}{n}+\frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)$	

1.2. Causality

We can use regression to estimate treatment effects from experimental data. We use an indicator variable that denotes treatment assignment (T_i), and estimate and it's effect on an outcome Y_i (i.e., by estimating the coefficient preceding T_i).

Note: There are some differences in how we think about this framework, versus the sampling (and Neyman) framework. The primary difference is that Y_i (our outcomes) are no longer fixed. Recall that within sampling, the only thing that was random within our framework was the treatment assignment. However, in the regression framework, we can no longer make this claim. Instead, we make a distinction between the sample we are working with, and some imaginary super-population. The outcomes in the super-population are still fixed, but the outcomes we observe are assumed to be randomly sampled from them. Therefore, when we estimate our regression, depending on how we specify our regression, we run into potential issues, such as finite-sample bias.

No Covariate Case

To begin simply:

$$Y_i(T_i) = \alpha + \tau_{SR}T_i + \varepsilon_i$$

$$\hat{Y}_i = \hat{\alpha} + \hat{\tau}_{SR}T_i$$

Notational note: I will denote the estimated coefficient of T_i here as $\hat{\tau}_{SR}$ to minimize confusion with the true treatment, difference in means estimate ($\hat{\tau}_{diff}$), and the later estimated coefficient from multiple regression.

$\hat{\tau}_{SR}$ will take on the same form as $\hat{\beta}$ from the simple regression standpoint. However, instead of X_i

and $\bar{X}$, we use T_i:

$$\hat{\tau}_{SR} = \frac{\sum_{i=1}^n (T_i - \bar{T})(Y_i - \bar{Y})}{\sum_{i=1}^n (T_i - \bar{T})^2}$$

What's kind of neat and an utter algebraic nightmare to show is that $\hat{\tau}_{SR}$ can be rewritten to:

$$\frac{1}{n_t} \sum_{i \in T} Y_i - \frac{1}{n_c} \sum_{i \in C} Y_i \equiv \hat{\tau}_{diff}$$

Therefore, when we regress only on the treatment indicator variable T_i, the estimate of $\hat{\tau}_{SR}$ is identical to that of the difference in means estimator. We know from our extensive hours of studying for the midterm that the difference in means estimator is an unbiased estimator for SATE. Therefore, $\hat{\tau}_{SR}$ is equivalently an unbiased estimator for SATE. *This is the only time in regression for which this is true!* When we introduce other covariates into the regression, we will induce finite-sample bias.

We can use the estimator of variance from OLS to come up with a variance estimator for $\hat{\tau}_{SR}$. We can come up with a slightly different formulation of $\hat{\sigma}^2$ that can help us tie this estimator back to the more familiar Neyman variance. The least squares formulation of the variance estimator is:

$$\hat{\mathbb{V}}(\hat{\tau}_{SR}) = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_i - \bar{X})^2},$$

where:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2,$$

and e_i are the residuals from our regression We can rewrite $\hat{\sigma}^2$:

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n-2} \sum_{i=1}^n e_i^2 \\
&= \frac{1}{n-2} \left(\sum_{i \in T} e_i^2 + \sum_{i \in C} e_i^2 \right) \\
&= \frac{1}{n-2} \left(\sum_{i \in T} (Y_i - \hat{Y}_i)^2 + \sum_{i \in C} (Y_i - \hat{Y}_i)^2 \right)
\end{aligned}$$

Note that $\hat{Y}_i$ can only take on two values: $\hat{\alpha}$ when $T_i = 0$, and $\hat{\alpha} + \hat{\tau}$ when $T_i = 1$. This is equivalent to the mean of the treatment and control groups (i.e., $\bar{Y}_t = \hat{\alpha} + \hat{\tau}$, and $\bar{Y}_c = \hat{\alpha}$. Therefore:

$$= \frac{1}{n-2} \left(\sum_{i \in T} (Y_i - \bar{Y}_t)^2 + \sum_{i \in C} (Y_i - \bar{Y}_c)^2 \right)$$

This is equivalent to the common variance across two potential outcome distributions. Plugging this all in:

$$\hat{\mathbb{V}}(\hat{\tau}_{SR}) = \hat{\sigma}^2 \cdot (\frac{1}{n_c} + \frac{1}{n_t})$$

This is what we commonly refer to as the *pooled variance*. If we relax the assumption of homoskedasiticty and allow for heteroskedasticity, we can u se the standard robust sampling variance estimator:

$$\hat{\mathbb{V}}_{hetero} = \frac{\sum_{i=1}^{n} e_i^2 (T_i - \bar{T})^2}{\left(\sum_{i=1}^{n} (T_i - \bar{T})^2\right)^2}$$

We can do a bunch of algebra and show that this becomes:

$$\hat{\mathbb{V}}_{hetero} = \frac{1}{n_t} \sum_{i \in T} (Y_i - \bar{Y}_t)^2 + \frac{1}{n_c} \sum_{i \in C} (Y_i - \bar{Y}_c)^2 \equiv \hat{\mathbb{V}}_{Neyman}$$

Long story short: when we have no additional covaraites, $\hat{\tau}_{SR}$ is an unbiased estimator for SATE, and if we allow for heteroskedasticity, the variance for $\hat{\tau}_{SR}$ will be equivalent to that of the Neyman variance.

2. Appendix

Derivation of $\hat{\tau}_{SR} = \hat{\tau}_{diff}$:

We begin with:

$$\hat{\tau}_{SR} = \frac{\sum_{i=1}^{n} (T_i - \bar{T})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (T_i - \bar{T})^2}$$

We begin by expanding out the denominator:

$$\begin{aligned}
\sum_{i=1}^{n} (T_i - \bar{T})^2 &= \sum_{i=1}^{n} (T_i^2 - 2T_i\bar{T} + \bar{T}^2) \\
&= \sum_{i=1}^{n} T_i^2 - 2\bar{T} \sum_{i=1}^{n} T_i + n\bar{T}^2 \\
&\text{We note that } T_i^2 = T_i\text{:} \\
&= \sum_{i=1}^{n} T_i - 2n\bar{T}^2 + n\bar{T}^2 \\
&= n\bar{T} - n\bar{T}^2 \\
&= \frac{n \cdot n_t}{n} - \frac{n_t^2}{n} \\
&= \frac{n_t(n - n_t)}{n} \\
&= \frac{n_t \cdot n_c}{n}
\end{aligned}$$

Now expanding out the numerator:

$$\begin{aligned}\sum_{i=1}^{n}(T_i-\bar{T})(Y_i-\bar{Y}) &= = \sum_{i=1}^{n}(T_iY_i - T_i\bar{Y} - \bar{T}Y_i + \bar{T}\bar{Y}) \\ &= \sum_{i=1}^{n} T_iY_i - n\bar{T}\bar{Y} - \bar{T}\sum_{i=1}^{n} Y_i + n\bar{T}\bar{Y} \\ &= \sum_{i=1}^{n} T_iY_i - \bar{T}\sum_{i=1}^{n} Y_i\end{aligned}$$

We note that we can rewrite the first summation as just the summation of all of our outcomes in the treatment group. Furthermore, the second summation can be expanded out:

$$\begin{aligned}&= \sum_{i\in T} Y_i - \bar{T}\sum_{i=1}^{n}(Y_i(1)T_i + (1-T_i)Y_i(0)) \\ &= \sum_{i\in T} Y_i - \bar{T}\left(\sum_{i\in T} Y_i + \sum_{i\in C} Y_i\right) \\ &= \sum_{i\in T} Y_i - \frac{n_t}{n}\left(\sum_{i\in T} Y_i + \sum_{i\in C} Y_i\right) \\ &= \frac{n-n_t}{n}\sum_{i\in T} Y_i - \frac{n_t}{n}\sum_{i\in C} Y_i \\ &= \frac{n_c}{n}\sum_{i\in T} Y_i - \frac{n_t}{n}\sum_{i\in C} Y_i\end{aligned}$$

Therefore, combining the numerator and denominator:

$$\begin{aligned}\hat{\tau}_{SR} &= \frac{\frac{n_c}{n}\sum_{i\in T} Y_i - \frac{n_t}{n}\sum_{i\in C} Y_i}{\frac{n_t \cdot n_c}{n}} \\ &= \frac{1}{n_t}\sum_{i\in T} Y_i - \frac{1}{n_c}\sum_{i\in C} Y_i \\ &\equiv \hat{\tau}_{diff}\end{aligned}$$

Derivation of $\hat{Y} \perp e$:

$$\begin{aligned}P_xY \cdot M_xY &= (M_xY)^\top P_xY \\ &= Y^\top M_x^\top P_xY \\ &= Y^\top (I-P_x)^\top P_xY \\ &= Y^\top (P_x - P_x^\top P_x)Y \\ &= Y^\top (P_x - P_x)Y\end{aligned}$$