# MAS 420: OLS Results and Inference

Adam Rohde

Department of Statistics
University of California, Los Angeles

October 10, 2022

# Recommended resources

We'll go through the details of how to derive some key OLS results as well as discuss common types of standard errors for OLS coefficient estimates.

Useful resources

1. Foundations of Agnostic Statistics - Aronow and Miller (2019) [link]
2. Elements of Statistical Learning - Hastie et al (2009) [link]
3. Econometrics - Hansen (2014)

# Running Example

We'll use Lalonde (1986) as an example dataset. The authors are trying to understand whether a job training program increased earnings. We have high-level information on the participants (e.g., age and education) as well as their earnings from 1975 and their earnings from 1978.

We want to estimate how earnings compare for participants that were in the job training program relative to those that were not.

We'll use OLS regression to estimate the best linear predictor (BLP) as a simple but interpertable approximation to the conditional expectation function (CEF) and a way to understand how the average 1978 earnings compares between the treated group and the control group. (We're not going to really worry about causal inference, but this is a RCT.)

# Set up

We have random variables $Y_i$ (i.e., 1978 earnings) and $\mathbf{X}_i = (X_{i,1}, X_{i,2}, X_{i,3}, X_{i,4}, X_{i,5})$ (i.e., treatment indicator, age, education in years, married indicator, 1975 earnings, so $p = 5$) for each of $n = 445$ participants (indexed by $i$). Suppose that $Y, \mathbf{X}$ are drawn from a joint density $p(Y, \mathbf{X})$ and our observations are IID.

We'll arrange the $X$'s into a $n \times (p+1)$ matrix with $n$ rows for the 445 participants and $p+1$ columns for the 5 explanatory variables as well as a column of 1's for the intercept.

$$\mathbb{X}_{n\times(p+1)} = \begin{pmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,p} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,p} \end{pmatrix} \text{ and } \mathbf{Y}^\top = (Y_1, Y_2, \ldots, Y_n)$$

[R demo]

# Using a linear model

So in reality 1978 earnings ($Y$) is not necessarily a linear function of treatment indicator, age, education in years, married indicator, 1975 earnings ($\mathbf{X}$), but we can get a simple and useful summary of how the treatment (the job training program $X_1$) relates to 1978 earnings using a linear model.

We can write $Y_i = \mathbf{X}_i\beta + \epsilon_i$, where $\epsilon_i$ is the error $Y_i - \mathbf{X}_i\beta$ and $\beta^\top = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p) \in \mathbb{R}^{p+1}$ is a vector of coefficients.

We can use our data to estimate $\beta$. We'll call our estimate $\hat{\beta}$.
And our residuals are $\hat{e}_i = Y_i - \mathbf{X}_i\hat{\beta}$. We'd like these to be small to approximate $Y$ well.

OLS regression estimates $\hat{\beta}$ by minimizing the sum of squared residuals across all obs.
That is, finding the $\hat{\beta}$ that minimizes $\sum_{i=1}^n \hat{e}_i^2$.

# OLS

$\hat{\mathbf{e}} = \mathbf{Y} - \mathbb{X}\hat{\beta}$ is the vector of these residuals.
The OLS problem is to find the $\hat{\beta}$ that minimizes $\sum_{i=1}^{n} \hat{e}_i^2$:

$$\begin{aligned}
\hat{\beta} &= \arg\min_{\mathbf{b}\in\mathbb{R}^{p+1}} \sum_{i}^{n} \hat{e}_i^2 \\
&= \arg\min_{\mathbf{b}\in\mathbb{R}^{p+1}} \|\hat{\mathbf{e}}\|_2^2 \\
&= \arg\min_{\mathbf{b}\in\mathbb{R}^{p+1}} \|\mathbf{Y} - \mathbb{X}\mathbf{b}\|_2^2 \\
&= \arg\min_{\mathbf{b}\in\mathbb{R}^{p+1}} (\mathbf{Y} - \mathbb{X}\mathbf{b})^\top (\mathbf{Y} - \mathbb{X}\mathbf{b})
\end{aligned}$$

# OLS

We can minimize $(\mathbf{Y} - \mathbb{X}\mathbf{b})^\top(\mathbf{Y} - \mathbb{X}\mathbf{b})$ by taking the derivative with respect to $\mathbf{b}$ and setting it to zero:

$$\begin{aligned}
\frac{\delta}{\delta \mathbf{b}}(\mathbf{Y} - \mathbb{X}\mathbf{b})^\top(\mathbf{Y} - \mathbb{X}\mathbf{b}) &= -2\mathbb{X}^\top(\mathbf{Y} - \mathbb{X}\mathbf{b}) \\
0 &= -2\mathbb{X}^\top(\mathbf{Y} - \mathbb{X}\mathbf{b}) = -2(\mathbb{X}^\top\mathbf{Y} - \mathbb{X}^\top\mathbb{X}\mathbf{b}) \\
\implies \mathbb{X}^\top\mathbb{X}\mathbf{b} &= \mathbb{X}^\top\mathbf{Y} \\
\implies \hat{\beta} = \mathbf{b} &= (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\mathbf{Y}
\end{aligned}$$

[R demo]

# $\beta$ and $\hat{\beta}$

Recall that $Y = \mathbb{X}\beta + \epsilon$.
We can show that

$$\begin{aligned}
\hat{\beta} &= (\mathbb{X}^\top \mathbb{X})^{-1}\mathbb{X}^\top \mathbf{Y} \\
&= (\mathbb{X}^\top \mathbb{X})^{-1}\mathbb{X}^\top (\mathbb{X}\beta + \epsilon) \\
&= (\mathbb{X}^\top \mathbb{X})^{-1}\mathbb{X}^\top \mathbb{X}\beta + (\mathbb{X}^\top \mathbb{X})^{-1}\mathbb{X}^\top \epsilon \\
&= \beta + (\mathbb{X}^\top \mathbb{X})^{-1}\mathbb{X}^\top \epsilon
\end{aligned}$$

It turns out that $\sqrt{n}(\hat{\beta} - \beta)$ converges in distribution to a multivariate normal distribution with expectation 0.

# Expectation of OLS estimator

Without dealing with asymptotics, we can still investigate $\mathbb{E}[\hat{\beta}|\mathbb{X}]$ and $\mathbb{E}[\hat{\beta}]$:

$$\begin{aligned}\mathbb{E}[\hat{\beta}|\mathbb{X}] &= \mathbb{E}[\beta + (\mathbb{X}^\top \mathbb{X})^{-1}\mathbb{X}^\top \epsilon|\mathbb{X}] \\ &= \beta + (\mathbb{X}^\top \mathbb{X})^{-1}\mathbb{X}^\top \mathbb{E}[\epsilon|\mathbb{X}]\end{aligned}$$

So unless $\mathbb{E}[\epsilon|\mathbb{X}] = 0$, $\mathbb{E}[\hat{\beta}|\mathbb{X}] \neq \beta$. Here we are taking the data matrix $\mathbb{X}$ to be random, and so it only can be pulled out of the expectation when we condition on it.

Often people treat $\mathbb{X}$ as fixed (not random) in this case, $\mathbb{E}[\hat{\beta}] = \beta + (\mathbb{X}^\top \mathbb{X})^{-1}\mathbb{X}^\top \mathbb{E}[\epsilon] = \beta$, since $\mathbb{E}[\epsilon] = 0$ for the BLP.

# Variance of OLS estimator

Once we have our OLS estimate, we want to quantify our uncertainty about our estimate. We do this by looking at the sampling variance (and it's square root, the standard error) of our estimator. We can then use these to estimate confidence intervals or p-values. E.g., we might be interested in whether we can statistically distinguish our estimates from zero.

So let's investigate $V[\hat{\beta}|\mathbb{X}]$, the sampling variance of our OLS estimate:

$$\begin{aligned} V[\hat{\beta}|\mathbb{X}] &= V[\beta + (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\epsilon|\mathbb{X}] \\ &= V[(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\epsilon|\mathbb{X}] \\ &= (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top V[\epsilon|\mathbb{X}]\mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1} \quad \text{since } \mathbb{X}^\top\mathbb{X} \text{ is symmetric} \\ &= (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\Sigma\mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1} \end{aligned}$$

where $V[\epsilon|\mathbb{X}] = \Sigma = \epsilon\epsilon^\top$ is the conditional covariance matrix of the errors.
We want to find a way to estimate $V[\hat{\beta}|\mathbb{X}]$.

# Assumptions about errors

Recall that $Y_i = \mathbf{X}_i\beta + \epsilon_i$.

$$\Sigma = \epsilon\epsilon^\top = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,n} \\ \sigma_{1,2} & \sigma_2^2 & \cdots & \sigma_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,n} & \sigma_{2,n} & \cdots & \sigma_n^2 \end{pmatrix}$$ is the conditional covariance matrix of the errors, $\epsilon_i$.

So we are now considering how the errors for each observation covary or not.

It turns out that a simple plug in estimator using $\hat{\mathbf{e}}\hat{\mathbf{e}}^\top$ (i.e., using the regression residuals) for $\Sigma$ will not converge to $V[\hat{\beta}|\mathbb{X}]$ as $n \to \infty$.

There are a few common assumptions about these errors that we'll discuss: homoscedastic errors, heteroscedastic errors, clustered errors. These are assumptions about the distribution and relationship between the errors across observations that allow us to estimate $V[\hat{\beta}|\mathbb{X}]$.

# Homoscedastic errors

An assumption of homoscedasticity means that we assume the $\epsilon_i$'s are independent (i.e., have zero covariance) and have a common variance $\sigma^2$. That is, $\Sigma = \mathbb{I}_n \sigma^2$:

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}$$

Under this assumption, we see that

$$V[\hat{\beta}|\mathbb{X}] = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \Sigma \mathbb{X} (\mathbb{X}^\top \mathbb{X})^{-1} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbb{I}_n \sigma^2 \mathbb{X} (\mathbb{X}^\top \mathbb{X})^{-1} = \sigma^2 (\mathbb{X}^\top \mathbb{X})^{-1}$$

We can estimate $\sigma^2$ with $\hat{\sigma}^2 = \frac{1}{n-p} \sum \hat{e}_i^2$.

So we estimate the standard errors of $\hat{\beta}$ as $\sqrt{\hat{V}[\hat{\beta}|\mathbb{X}]} = \sqrt{\hat{\sigma}^2 (\mathbb{X}^\top \mathbb{X})^{-1}}$.
[R demo]

# Heteroscedastic errors

An assumption of heteroscedasticity means that we assume the $\epsilon_i$'s are independent (i.e., have zero covariance) but each has its own variance $\sigma_i^2$. That is,

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

Under this assumption, we see that

$$V[\hat{\beta}|\mathbb{X}] = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\boldsymbol{\Sigma}\mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1} = (\mathbb{X}^\top\mathbb{X})^{-1}\left(\sum_{i=1}^{n}\mathbf{X_i}\mathbf{X_i}^\top\sigma_i^2\right)(\mathbb{X}^\top\mathbb{X})^{-1}$$

It can be shown that $\boldsymbol{\Sigma} = \mathbb{E}[\epsilon\epsilon^\top|\mathbb{X}] = \mathbb{E}[D_0|\mathbb{X}]$, where $D_0$ is a diagonal matrix with $\epsilon_i^2$'s on the diagonal. We can estimate the $\epsilon_i^2$'s with $\hat{e}_i^2$'s. So we estimate the standard errors of $\hat{\beta}$ as $\sqrt{\hat{V}[\hat{\beta}|\mathbb{X}]} = \sqrt{(\mathbb{X}^\top\mathbb{X})^{-1}\left(\sum_{i=1}^{n}\mathbf{X_i}\mathbf{X_i}^\top\hat{e}_i^2\right)(\mathbb{X}^\top\mathbb{X})^{-1}}$.
This is the simplest form of ”robust” SE.

# Robust standard errors

There are a few flavors of robust standard error.

1. HC0: $\hat{V}[\hat{\beta}|\mathbb{X}] = (\mathbb{X}^\top \mathbb{X})^{-1} \left( \sum_{i=1}^{n} \mathbf{X_i} \mathbf{X_i}^\top \hat{e}_i^2 \right) (\mathbb{X}^\top \mathbb{X})^{-1}$
2. HC1: We could scale the variance estimate by $\frac{n}{n-p}$ to get an unbiased estimate.
3. HC2: We could use standardized residuals rather than the residuals $\hat{e}_i$.
4. HC3: We could use the prediction errors from leave one out cross validation rather than the residuals $\hat{e}_i$.

[R demo]

# Clustered errors

An assumption of clustered errors means that we assume the $\epsilon_i$'s are not all independent. That is, within in clusters of observations, the errors are correlated, but across clusters the errors are not clustered. In this case, the covariance matrix is block diagonal. For example, with four units and two clusters

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & 0 & 0 \\ \sigma_{12} & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & \sigma_{34} \\ 0 & 0 & \sigma_{34} & \sigma_4^2 \end{pmatrix}$$

Think of a study looking at a new type of teaching method for elementary schools. Students are exposed or unexposed to the new teaching method based on which classroom they are in. But the outcomes for the students in the same class are also probably correlated.

We can use plug in estimates of the cluster covariance matrices to get a consistent estimator for $V[\hat{\beta}|\mathbb{X}]$, as the number of clusters becomes large (e.g., more than like 40).

# The Bootstrap

**Idea**

The bootstrap is another way to do inference for estimators. It is based on using the empirical distribution (of $Y, \mathbf{X}$) as an approximation to the true distribution.

If we knew the true distribution, then we could simulate draws from that distribution to calculate the sampling distribution of any estimator to arbitrary precision by literally taking repeated simulated samples of size $n$ and each time calculating our estimate and then looking at the distribution of these estimates.

Of course, we don't know the true distribution. **The bootstrap says pretend that the distribution that we are sampling looks exactly like the sample we have. Plug in the empirical distribution for the true distribution and sample from that.**

# The Bootstrap

**Procedure**

To get standard errors for our estimator,

1. Take a sample, *with replacement*, of size $n$ from our data sample.
2. Calculate the estimate for that bootstrap sample.
3. Repeat 1. and 2. many many times.
4. Calculate the standard deviation of the resulting collection of bootstrap estimates.

The result of 4. is our estimate for the standard error of our estimator. This will consistently estimate the standard error.

The bootstrap assumes IID observations and plug in regularity conditions. But notice that this is a very general procedure that can be applied not just for regression but for any estimator. We can also look at quantiles of the distribution of bootstrap estimates to estimate confidence intervals. **"When in doubt, use the bootstrap."** (paraphrasing Aronow and Miller (2019)) [R demo]

# Any remaining time

Group work on PSET / questions