Lab 2: (Yet another) Intro to Regression II

Info 251 — Applied Machine Learning
University of California, Berkeley ∘ Spring 2026
authors: Simón Ramírez Amaya
{simonra}@berkeley.edu

Objectives

1. Understand how sampling variation affects estimation in regression problems.
2. Use linear regression estimators in Python and interpret their output.

Sampling variation

So far, we have largely ignored the fact that our estimates of the linear parameters depend on the random sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ drawn from the joint distribution of Y and $\mathbf{X}$. In other words, we can expect to get a different answer every time we sample. For instance, even if we are sampling from distribution A, most of our estimates will still be negative or positive. This suggests that in order to make meaningful interpretations of our estimates, we need to somehow account for sampling variation and device a mechanism that allows us to distinguish whether the observed estimates are consistent with some hypothesis of interest (tipically $\beta = 0$).

In large samples, the behavior of sampling variation is well understood. We will discuss three main results. The first two tell us that, as n grows large, our estimators converge to the slopes that we would use if we had perfect knowledge of the joint distribution (also called *best linear predictors*, BLPs).

1. In the context of simple linear regression, $\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \to \frac{\text{Cov}(X,Y)}{Var(X)}$.
2. In the context of multiple linear regression, $\hat{\boldsymbol{\beta}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y} \to \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]^{-1}\mathbb{E}[\boldsymbol{X}Y]$

The third result states that, for large n, the sampling distribution of our estimators is going to be approximately normal, centered at the true slope (BLP if CMF is non-linear), and with a variance that is going to relate the sample size n, the unexplained variation in Y and the variation in X. In the context of simple linear regression:

$$\hat{\beta}_1 \sim \mathcal{N}\Big(\beta_1, \frac{1}{n}\frac{\text{noise in Y}}{\text{spread in X}}\Big)$$

There are a number of important considerations for defining and estimating the sampling variance. Since this is AML and not econometrics, we will not go into the details. However, there is one applied tip worth mentioning: *always* use heteroskedasticity-robust estimation (`cov_type=HC1` in the `statsmodels.api.OLS.fit` method).

Exercise 1

Head to the companion notebook and implement the following procedure. Draw $m = 1,000$ different random samples of size $n = 500$ from distribution C (positive correlation between income and years of schooling, discontinuities at x=12 and x=16.). For each sample, estimate a simple linear regression model and use the results to construct a figure with two subplots. In the first subplot show a histogram of the slope estimates with a vertical dashed line at the value of the true slope. In the second, display the regression lines with the true CMF and the BLP overlayed.

Hypothesis Testing in Simple Linear Regression

We are going to conduct hypothesis testing about β using the large sample approximation to the sampling distribution of $\hat{\beta}_1$. We present the plain vanilla case in the simple linear regression case.

1. **Null and alternative hypotheses**

   We wish to test a hypothesis about the slope coefficient β_1:

   $$H_0 : \beta_1 = 0 \qquad \text{vs} \qquad H_1 : \beta_1 \neq 0,$$

2. **Estimator**

   The OLS estimator of the slope is

   $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

3. **Standard error**

   The standard error of $\hat{\beta}_1$ is defined as the square root of its estimated conditional variance:

   $$\widehat{\mathrm{SE}}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}},$$

   where

   $$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2, \qquad \hat{\varepsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i.$$

4. **Test statistic**

   The t–statistic for testing H_0 is

   $$t = \frac{\hat{\beta}_1 - 0}{\widehat{\mathrm{SE}}(\hat{\beta}_1)}.$$

5. **Distribution of the test statistic**

   Under H_0:

   $$t \sim t_{n-2}.$$

6. **Rejection region**

   For a significance level $\alpha \in (0, 1)$, reject H_0 if

   $$|t| > t_{n-2}(1 - \alpha/2),$$

   where $t_{n-2}(1-\alpha/2)$ denotes the $(1-\alpha/2)$ quantile of the Student t distribution with $n-2$ degrees of freedom.

Exercise 2

Head to the companion notebook. Draw a sample of size n=500 from distribution A and estimate a simple linear model using `statsmodels.api.OLS`. Print the results and interpret the regression table. Can you map the elements of the hypothesis test outlined above? Please discuss with the person next to you. Repeat this analysis for samples drawn from distributions B and C.

OLS with a binary regressor (time allowing)

Exercise 3

Consider a randomized experiment with a binary treatment indicator W and an outcome Y. It is often stated that the average treatment effect can be measured by the difference in mean outcomes between the treated and control groups. Show that the ordinary least squares (OLS) estimator from a regression of Y on W yields exactly this difference in means. This is:

$$(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y} = \begin{bmatrix} \frac{1}{n_0}\sum_{i:T_i=0} Y_i \\ \frac{1}{n_1}\sum_{i:T_i=1} Y_i - \frac{1}{n_0}\sum_{i:T_i=0} Y_i \end{bmatrix}, \tag{1}$$

where n_0 and n_1 is the number of observations in the sample assigned to control and treatment conditions, respectively.

Hint:

$$\begin{bmatrix} n & n_1 \\ n_1 & n_1 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{n_0} & \frac{-1}{n_0} \\ \frac{-1}{n_0} & \frac{n}{n_1 n_0} \end{bmatrix} \tag{2}$$