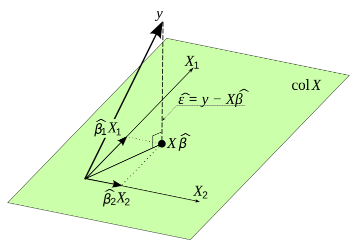
Problem Set 1

PPHA41430/ECMA31110
Due: Thursday, October 17, 2024

1. Geometry of OLS

Consider the drawing on of the orthogonal projection interpretation of OLS.



a. Can you illustrate the case of perfect fit, e.g., when $p \geq n$.

b. Produce the equivalent drawing to illustrate the omitted variable bias phenomenon.

hint: Think of $Y_i = X_{i,1}\beta_1 + X_{i,2}\beta_2 + \varepsilon_i$ as the long regression, $Y_i = X_{i,1}\beta_1 + \varepsilon_i$ as the short regression, and produce a drawing that has the projection of the vector $\mathbf{Y}$ on the span of $\mathbf{X}_1 = (X_{1,1}, \ldots, X_{n,1})^T$ and $\mathbf{X}_2 = (X_{1,2}, \ldots, X_{n,2})^T$, the projection of $\mathbf{Y}$ on the span of $\mathbf{X}_1$, and one other projection.

c. bonus: Can you produce the same drawing to illustrate an IV regression?

2. The best linear predictor is a best linear approximator

Consider the linear predictor

$$E^*[Y|X] := X\beta^* := \arg\min_{X\beta} E\left[(Y - X\beta)^2\right].$$

Show that

$$E^*[Y|X] = \arg\min_{X\beta} E\left[(X\beta - E[Y|X])^2\right],$$

and explain why this is an important result.

3. Variance of 2SLS

Consider the two-stage least squares estimator for

$$\mathbf{Y} = \mathbf{D}\beta + \varepsilon$$

and

$$\mathbf{D} = \mathbf{Z}\gamma + \eta,$$

where $\mathbf{D} \in \mathbb{R}^n$ is the vector of endogenous variables and $\mathbf{Z} \in \mathbb{R}^{n\times d}$ is the matrix of d instruments. Specifically $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_d)$ and $\mathbf{Z}_j = (Z_{1j}, \ldots, Z_{nj})^T$ for all

j. Assume the observations and errors are iid and homoskedastic. Show that the variance –asymptotic or under distributional assumptions– of $\hat{\beta}$ is smaller when d is larger. Specifically, consider the case of using additional instruments Z_{d+1},...,$Z_{d'}$, for some $d < d' < n$, and using $\mathbf{Z}' = (\mathbf{Z}_1, \ldots, \mathbf{Z}_d, \ldots, \mathbf{Z}_{d'})$.

4. Partial identification of the linear regression coefficient with singular design matrix

1. Give the pseudocode of the algorithm for Gaussian elimination with rectangular matrices. Discuss the difference, if any, between "tall" and "wide" matrices.

2. Construct an example with $\mathbf{X} \in \mathbb{R}^{n \times p}$, $p = 3$, where exactly one covariate is identified.

3. Can you construct an example with $\mathbf{X} \in \mathbb{R}^{n \times p}$, $p = 3$, where exactly two covariates are identified?

4. Produce user-friendly software which outputs the lm output when the design matrix is full rank and, when the design matrix is singular, outputs point estimates (with standard errors, p-values, etc) only for the identified coefficients, as well as an intuitive description of the null space of $\mathbf{X}$ that suggests to the user which covariate is in a linear dependent relationship with which.

You may, instead of using Gaussian elimination, consider using the singular value decomposition. Explain your methodology and reasoning.

bonus: Suggest an extension of the method for thin and nearly-collinear design matrices.