ECON 480-3
LECTURE 1: LINEAR REGRESSION

Ivan A. Canay
Northwestern University

Interpreting Linear Regression

Let (Y, X, U) be a random vector where Y and U take values in $\mathbf{R}$ and X takes values in $\mathbf{R}^{k+1}$. Assume further that the first component of X is a constant equal to one, i.e., $X = (X_0, X_1, \ldots, X_k)'$ with $X_0 = 1$. Let $\beta = (\beta_0, \beta_1, \ldots, \beta_k)' \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \; .$$

β_0 is an intercept parameter and the remaining β_j are slope parameters. There are several ways to interpret β depending on the assumptions imposed on (Y, X, U). We will study three such ways.

Three Interpretations of Linear Regression

- Linear Conditional Expectation
- Best Linear Approximation
- Causal Model

1: Linear Conditional Expectation

- Suppose that:

$$E[Y|X] = X'\beta$$

  and define $U = Y - E[Y|X]$.

- This has several implications:

- **Descriptive:** β is a convenient way of summarizing a feature of the joint distribution of Y and X.

- **Question:** can we interpret β_j as the ceteris paribus (i.e., holding X_{-j} and U constant) effect of a one unit change in X_j on Y?

QUESTIONS?

2: "Best" Linear Approximation

- In general, the conditional expectation is probably NOT linear.

- Suppose that: $E[Y^2] < \infty$ and $E[XX'] < \infty$ (or, $E[X_j^2] < \infty$ for $1 \leqslant j \leqslant k$)

- Under these assumptions, one may consider what is the "best" linear approximation (i.e., function of the form $X'b$ for some choice of $b \in \mathbf{R}^{k+1}$) to the conditional expectation.

- To this end, consider the minimization problem

$$\min_{b \in \mathbf{R}^{k+1}} E\left[(E[Y|X] - X'b)^2\right]$$

  and denote by β a solution to this minimization problem.

- **Descriptive:** β is a convenient way of summarizing a feature of the joint distribution of Y and X.

- **Question:** can we interpret β_j as the ceteris paribus (i.e., holding X_{-j} and U constant) effect of a one unit change in X_j on Y?

Best Linear Predictor

Claim

$$\beta \in \underset{b \in \mathbf{R}^{k+1}}{\text{argmin}}\, E\left[(Y - X'b)^2\right],$$

so β is also a convenient way of summarizing the “best” linear predictor of Y given X.

Proof:

Two in one

- **Two interpretations from equivalent optimization problems:**

$$\beta \in \underset{b \in \mathbf{R}^{k+1}}{\operatorname{argmin}} E\left[(E[Y|X] - X'b)^2\right] \qquad \text{and} \qquad \beta \in \underset{b \in \mathbf{R}^{k+1}}{\operatorname{argmin}} E\left[(Y - X'b)^2\right].$$

- Note $E[(Y - X'b)^2]$ is convex (as a function of b) and this has the following implications.

QUESTIONS?

3: Causal Model

- Suppose that: $Y = g(X, U)$, where X are the observed determinants of Y and U are the unobserved determinants of Y.
- Such a relationship is a model of how Y is determined and may come from physics, economics, etc.
- The effect of X_j on Y holding X_{-j} and U constant (i.e., *ceteris paribus*) is determined by g.
- **If** g is differentiable, then it is given by
$$D_{X_j} g(X, U) \ .$$
- **If** we assume further that
$$g(X, U) = X'\beta + U \ ,$$
then the *ceteris paribus* effect of X_j on Y is simply β_j. We may normalize U so that $E[U] = 0$ (by replacing U with $U - E[U]$ and β_0 with $\beta_0 + E[U]$ if this is not the case).
- On the other hand, $E[U|X]$, $E[U|X_j]$ and $E[UX_j]$ for $1 \leqslant j \leqslant k$ may or may not equal zero.

Potential Outcomes

- **Potential outcomes**: easy way to think about causal relationships.
- **illustration:** randomized controlled experiment where individuals are randomly assigned to a treatment (a drug) that is intended to improve their health status.
- **Notation:** Let Y denote the observed health status and $X \in \{0, 1\}$ denote whether the individual takes the drug or not.
- The causal relationship between X and Y can be described using the so-called *potential outcomes*:

$$\begin{array}{ll} Y(0) & \text{potential outcome in the absence of treatment} \\ Y(1) & \text{potential outcome in the presence of treatment} \end{array}.$$

- Thus, we imagine two health status variables $(Y(0), Y(1))$ where $Y(0)$ is the value of the outcome that would have been observed if (possibly counter-to-fact) X were 0; and $Y(1)$ is the value of the outcome that would have been observed if (possibly counter-to-fact) X were 1.

Treatment Effects

- The difference $Y(1) - Y(0)$ is called the treatment effect.
- The quantity $E[Y(1) - Y(0)]$ is usually referred to as the average treatment effect.
- Using this notation, we may rewrite the observed outcome as:

Interpretation

$$Y = \beta_0 + \beta_1 X + U \quad \text{with} \quad \beta_1 = Y(1) - Y(0)\ .$$

- Not quite "the" linear model: the coefficient β_1 is random.
- For β_1 to be constant, we need to assume that $Y(1) - Y(0)$ is constant across individuals.
- Under all these assumptions: we end up with a *linear constant effect causal model* with $U \perp\!\!\!\perp X$ (from the nature of the randomized experiment), $E[U] = 0$, and so $E[XU] = 0$.
- Without assuming constant treatment effects it can be shown that a regression of Y on X identifies the average treatment effect,

$$\beta = \frac{\text{Cov}[Y, X]}{\text{Var}[X]} = E[Y(1) - Y(0)]$$

which is often called a causal parameter given that it is an average of causal effects.

QUESTIONS?

Linear Regression when $E[XU] = 0$

- Let (Y, X, U) be a random vector where Y and U take values in $\mathbf{R}$ and X takes values in $\mathbf{R}^{k+1}$. Assume further that $X = (X_0, X_1, \ldots, X_k)'$ with $X_0 = 1$ and let $\beta = (\beta_0, \beta_1, \ldots, \beta_k)' \in \mathbf{R}^{k+1}$ be such that
$$Y = X'\beta + U\,.$$
Suppose 1 $E[XU] = 0$ 2 $E[XX'] < \infty$, and 3 that there is *no perfect collinearity* in X.

- The justification of 1 varies depending on which of the three preceding interpretations we invoke.

- 2 ensures that $E[XX']$ exists.

- 3 is equivalent to the assumption that the matrix $E[XX']$ is in fact invertible. Since $E[XX']$ is positive semi-definite, invertibility of $E[XX']$ is equivalent to $E[XX']$ being positive definite.

Invertibility

Definition

There is *perfect collinearity* or *multicollinearity* in X if there exists nonzero $c \in \mathbf{R}^{k+1}$ such that $P\{c'X = 0\} = 1$, i.e., if we can express one component of X as a linear combination of the others.

Lemma

Let X be such that $E[XX'] < \infty$. Then $E[XX']$ is invertible **iff** *there is no perfect collinearity in X.*

Solving for β

- $E[UX] = 0$ implies that $E[X(Y - X'\beta)] = 0$, i.e.,

$$E[XY] = E[XX']\beta \ .$$

- Since $E[XX']$ is invertible, there is a unique solution to this system of equations, namely,

$$\beta = E[XX']^{-1}E[XY] \ .$$

- If $E[XX']$ is not invertible there will be more than one solution to this system of equations. Any two solutions β and $\tilde{\beta}$ will necessarily satisfy $P\{X'\beta = X'\tilde{\beta}\} = 1$.

- **In this important?:** It depends on the interpretation. For instance, in the second interpretation, each such solution corresponds to the same "best" linear predictor of Y given X, whereas in the third interpretation different values of β could have wildly different implications for how X affects Y holding U constant.

Estimating β: OLS

- Let (Y, X, U) be as described and let P the marginal distribution of (Y, X).
- Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an i.i.d. sequence of random vectors with distribution P.
- A natural estimator of $\beta = (E[XX'])^{-1}E[XY]$ is simply

$$\hat{\beta}_n = \left(\frac{1}{n} \sum_{1 \leqslant i \leqslant n} X_i X_i' \right)^{-1} \left(\frac{1}{n} \sum_{1 \leqslant i \leqslant n} X_i Y_i \right) .$$

- This estimator is called the ordinary least squares (OLS) estimator of β because it can also be derived as the solution to the following minimization problem:

$$\min_{b \in \mathbf{R}^{k+1}} \frac{1}{n} \sum_{1 \leqslant i \leqslant n} (Y_i - X_i' b)^2 .$$

Estimating β: OLS

Claim

$\hat{\beta}_n$ solves the following minimization problem: $\min_{b \in \mathbf{R}^{k+1}} \frac{1}{n} \sum_{1 \leqslant i \leqslant n} (Y_i - X_i' b)^2$.

QUESTIONS?

Matrix Notation

Define

$$\begin{aligned}
\mathbb{Y} &= (Y_1, \ldots, Y_n)' \\
\mathbb{X} &= (X_1, \ldots, X_n)' \\
\hat{\mathbb{Y}} &= (\hat{Y}_1, \ldots, \hat{Y}_n)' \\
&= \mathbb{X}\hat{\beta}_n \\
\mathbb{U} &= (U_1, \ldots, U_n)' \\
\hat{\mathbb{U}} &= (\hat{U}_1, \ldots, \hat{U}_n)' \\
&= \mathbb{Y} - \hat{\mathbb{Y}} \\
&= \mathbb{Y} - \mathbb{X}\hat{\beta}_n \, .
\end{aligned}$$

In this notation,

$$\hat{\beta}_n = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}$$

and may be equivalently described as the solution to

$$\min_{b \in \mathbf{R}^{k+1}} |\mathbb{Y} - \mathbb{X}b|^2 \, .$$

Hence, $\mathbb{X}\hat{\beta}_n$ is the vector in the column space of $\mathbb{X}$ that is closest (in terms of Euclidean distance) to $\mathbb{Y}$.

Projection Matrices

$$\mathbb{X}\hat{\beta}_n = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}$$

is the orthogonal projection of $\mathbb{Y}$ onto the ($(k+1)$-dimensional) column space of $\mathbb{X}$.

The matrix

$$\mathbb{P} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$$

is known as a projection matrix. It projects a vector in $\mathbf{R}^n$ (such as $\mathbb{Y}$) onto the column space of $\mathbb{X}$. Note that $\mathbb{P}^2 = \mathbb{P}$, which reflects the fact that projecting something that already lies in the column space of $\mathbb{X}$ onto the column space of $\mathbb{X}$ does nothing.

The matrix $\mathbb{P}$ is also symmetric. The matrix

$$\mathbb{M} = \mathbb{I} - \mathbb{P}$$

is also a projection matrix. It projects a vector onto the ($(n-k-1)$-dimensional) vector space orthogonal to the column space of $\mathbb{X}$. Hence, $\mathbb{M}\mathbb{X} = 0$. Note that $\mathbb{M}\mathbb{Y} = \hat{\mathbb{U}}$. For this reason, $\mathbb{M}$ is sometimes called the “residual maker” matrix.