M1 INTERMEDIATE ECONOMETRICS

Linear regression model

Koen Jochmans François Poinas

2024 — 2025

Overview

This deck of slides set up the linear regression model and goes through a selection of elementary results that are considered prerequisites for this course.

The corresponding chapters in Hansen are 2 to 5.

(Linear) conditional-expectation function (H2.14-H2.15)

Dependent variable $Y \in \mathbb{R}$ and (column) vector of regressors $X \in \mathbb{R}^k$

The conditional-expectation function (CEF) is

$$m(x) = \mathbb{E}(Y|X = x).$$

The regression error is

$$e = Y - m(X).$$

Then

$$Y = m(X) + e, \qquad \mathbb{E}(e|X = x) = 0.$$

The CEF is interesting from a prediction perspective as it is optimal under expected squared loss, i.e.,

$$\mathbb{E}[(Y - m(X))^2] \leq \mathbb{E}[(Y - \tilde{m}(X))^2]$$

for any function $\tilde{m}$.

The linear CEF model restricts the conditional mean function to be

$$\mathbb{E}(Y|X = x) = x_1\beta_1 + x_2\beta_2 + \cdots + x_k\beta_k = x'\beta.$$

This is linear in the coefficient vector β.

Nonlinearities in the regressors can be accommodated.

Alternatively (and equivalently),

$$Y = X'\beta + e, \qquad \mathbb{E}(e|X = x) = 0. \tag{1}$$

Observe that, by the law of iterated expectations, $\mathbb{E}(e|X = x) = 0$ implies (among other things) that

$\mathbb{E}(Xe) = 0$

and also that $\mathbb{E}(e) = 0$.

We only restrict the conditional mean, not the conditional variance. So

$$\mathbb{E}(e^2|X = x)$$

may be a function of x.

One sometimes assumes that

$$\mathbb{E}(e^2|X = x) = \sigma^2,$$

which does not change with x.

This is an assumption of homoskedasticity.

Best linear approximation (H2.18 and H2.25)

When the CEF m is nonlinear we can still look for the best linear approximation $x'\beta$ to it.

Using expected squared loss, this mean solving

$$\min_{\beta} \mathbb{E}[(m(X) - X'\beta)^2].$$

The first-order condition for this problem is

$$\mathbb{E}(XX')\beta = \mathbb{E}(Xm(X)),$$

and is linear in β.

If $\operatorname{rank} \mathbb{E}(XX') = k$ we find the unique solution

$$\beta = \mathbb{E}(XX')^{-1}\,\mathbb{E}(Xm(X)) = \mathbb{E}(XX')^{-1}\,\mathbb{E}(XY)$$

where we used $\mathbb{E}(XY) = \mathbb{E}[\mathbb{E}(XY|X)] = \mathbb{E}[X\mathbb{E}(Y|X)] = \mathbb{E}(Xm(X))$ in the last step.

With this definition of β we can define the projection error

$$e = Y - X'\beta$$

and observe that

$$\mathbb{E}(Xe) = \mathbb{E}(XY) - \mathbb{E}(XX')\beta = \mathbb{E}(XY) - \mathbb{E}(XX')\,\mathbb{E}(XX')^{-1}\,\mathbb{E}(XY) = 0$$

so that we can always write

$$Y = X'\beta + e, \qquad \mathbb{E}(Xe) = 0. \tag{2}$$

If a constant term is included this further implies that $\mathbb{E}(e) = 0$.

Equations (1) and (2) only concern the same parameter vector β and error e when the CEF is linear. In the nonlinear CEF case, (1) is meaningless, but (2) is not.

Random (i.i.d.) sampling (H4.2)

$(Y_1, X_1), \ldots, (Y_n, X_n)$ is a random sample (of size n) on (Y, X) if and only if the (Y_i, X_i) are

independent across i and,

identically distributed according to the distribution of (Y, X).

Both parts can be relaxed considerably, but the random-sampling framework suffices for our purposes.

Matrix notation

We have $Y_i = X_i'\beta + e_i$ for all $i = 1, \ldots, n$.

Can vertically stack these n equations to arrive at

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{e}.$$

Here, $\boldsymbol{X}$ is an $n \times k$ matrix while $\boldsymbol{Y}$ and $\boldsymbol{e}$ are each $n \times 1$ vectors.

Now wish to construct an estimator $\hat{\beta}$ of β.

An estimator is just a function of the sample; many such functions can be taken.

Here, focus on ordinary least squares (OLS) estimator.

Ordinary least-squares estimator (H3.4 and H3.10)

Derivation 1: As the minimizer of the sum of squared residuals (SSR) or sum of squared errors (SSE),

$$\text{SSE}(b) = \sum_{i=1}^{n}(Y_i - X_i'b)^2 = (\boldsymbol{Y} - \boldsymbol{X}b)'(\boldsymbol{Y} - \boldsymbol{X}b).$$

The first-order condition for a minimum is

$$\sum_{i=1}^{n} X_i(Y_i - X_i'b) = \boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}b) = 0,$$

this is a linear system of k equations in k unknowns.

A unique solution exists if $\operatorname{rank} \boldsymbol{X}'\boldsymbol{X} = k$; it equals

$$\hat{\beta} = \left({}^1\!/_n \sum_{i=1}^{n} X_i X_i'\right)^{-1} \left({}^1\!/_n \sum_{i=1}^{n} X_i Y_i\right) = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y}).$$

The motivation for this approach to deriving OLS comes from the fact that we can characterize β to be the solution to the corresponding *population* problem

$$\mathbb{E}[(Y - X_i'b)^2].$$

Let $e_i(b) = Y_i - X_i'b$. Then

$$\min_b \text{SSE}(b) = \min_b \sum_{i=1}^{n} e_i^2(b),$$

and so we are minimizing the *variance* (if a constant term is included) of the squared errors.

The least-squares residuals are

$$\hat{e}_i = Y_i - X_i'\hat{\beta} = e_i(\hat{\beta}).$$

Collecting these in the $n \times 1$ vector $\hat{\boldsymbol{e}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\beta}$ we easily observe that

$$\sum_{i=1}^{n} X_i \hat{e}_i = \boldsymbol{X}'\hat{\boldsymbol{e}} = \boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{X}\hat{\beta} = \boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y}) = 0.$$

This motivates Derivation 2: As the estimator that yields residuals that are orthogonal (in sample) to the regressors,

$$\text{argsolve}_b \sum_{i=1}^{n} X_i e_i(b) = 0.$$

In matrix notation this is equal to solving the OLS estimating equations

$$\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}b) = 0$$

and this yields $\hat{\beta}$. This problem is a sample version of (2).

Derivation 3: Finally, given that we know that

$$\beta = \mathbb{E}(XX')^{-1}\mathbb{E}(XY),$$

a ratio of expectations, we may consider a direct sample counterpart to this.

A sample counterpart of $\mathbb{E}(XY)$ is $1/n \sum_{i=1}^{n} X_i Y_i = \boldsymbol{X}'\boldsymbol{Y}/n$.

A sample counterpart of $\mathbb{E}(XX')$ is $1/n \sum_{i=1}^{n} X_i X_i' = \boldsymbol{X}'\boldsymbol{X}/n$.

This again yields

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y}).$$

Collinearity

The estimating equations

$$\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}b) = 0$$

have a unique solution when the $k \times k$ design matrix $\boldsymbol{X}'\boldsymbol{X}$ is invertible.

As, for any $k \times 1$ vector $\boldsymbol{a}$,

$$\boldsymbol{a}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a} = \|\boldsymbol{X}\boldsymbol{a}\|^2$$

the design matrix is invertible if and only if the columns of $\boldsymbol{X}$ are linearly independent.

This is a no-multicollinearity condition.

In absence of this there are many solutions to the least-squares problem.

Projection matrices (H3.11-H3.16)

The matrix

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

projects on the space spanned by the columns of $\boldsymbol{X}$.

The matrix

$$\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{P} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

projects on the orthogonal complement.

Observe that

$$\boldsymbol{P}\boldsymbol{Y} = \boldsymbol{X}\hat{\beta}, \qquad \boldsymbol{M}\boldsymbol{Y} = \boldsymbol{Y} - \boldsymbol{X}\hat{\beta} = \hat{\boldsymbol{e}}$$

and that $\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{Y} + \boldsymbol{M}\boldsymbol{Y}$, decomposing $\boldsymbol{Y}$ into two orthogonal parts.

We derived an expression for β from (2). It is valid under both (1) and (2).

Under (1) we have $\mathbb{E}(e|X = x) = 0$ and so also

$$\mathbb{E}[g(X)\, e] = \mathbb{E}[g(X)(Y - X'\beta)] = 0$$

by the law of iterated expectations.

Hence, for any function g such that $\mathbb{E}[g(X)X']$ is an invertible matrix,

$$\beta = \mathbb{E}[g(X)X']^{-1}\mathbb{E}[g(X)Y]$$

is equally true.

Leads to many possible linear estimators of β.

Not true under (2) alone!

Let $\boldsymbol{G}$ be $n \times k$ matrix collecting all $g(X_i)$. Then the corresponding estimator is

$$\tilde{\beta} = (\boldsymbol{G}'\boldsymbol{X})^{-1}(\boldsymbol{G}'\boldsymbol{Y}).$$

Immediately clear that

$$\mathbb{E}(\tilde{\beta} - \beta | \boldsymbol{X}) = 0$$

and that

$$\text{var}(\tilde{\beta}|\boldsymbol{X}) = (\boldsymbol{G}'\boldsymbol{X})^{-1}(\boldsymbol{G}'\boldsymbol{D}\,\boldsymbol{G})(\boldsymbol{X}'\boldsymbol{G})^{-1},$$

where the $n \times n$ diagonal matrix

$$\boldsymbol{D} = \mathbb{E}(\boldsymbol{e}\boldsymbol{e}'|\boldsymbol{X}) = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$$

is the conditional variance matrix of the error vector.

For OLS, $\boldsymbol{G} = \boldsymbol{X}$. Under homoskedasticity, $\boldsymbol{D} = \sigma^2 \boldsymbol{I}_n$ and so the variance of the OLS estimator becomes

$$\text{var}(\hat{\beta}|\boldsymbol{X}) = \sigma^2\,(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Under (1) and homoskedastic errors, OLS is the best linear unbiased estimator (BLUE) of β. That is, we use $\boldsymbol{G} = \boldsymbol{X}$.

No other linear estimator that is unbiased can achieve a variance that is smaller than $\hat{\beta}$.

No longer true under heteroskedasticity. If

$$\boldsymbol{e}|\boldsymbol{X} \sim (0, \boldsymbol{D}),$$

then $\boldsymbol{D}^{-1/2}\boldsymbol{e}|\boldsymbol{X} \sim (0, \sigma^2 \boldsymbol{I}_n)$, and so the regression model

$$\boldsymbol{D}^{-1/2}\boldsymbol{Y} = \boldsymbol{D}^{-1/2}\boldsymbol{X}\beta + \boldsymbol{D}^{-1/2}\boldsymbol{e}$$

has homoskedastic errors. OLS applied to the transformed data yields

$$(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{Y}),$$

the generalized/weighted least-squares estimator (GLS/WLS). That is, we use $\boldsymbol{G} = \boldsymbol{D}^{-1}\boldsymbol{X}$. The variance of the GLS estimator is $(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}$.

Sampling properties (H5.5-H5.12)

When

$$\boldsymbol{e}|\boldsymbol{X} \sim N(0, \sigma^2\, \boldsymbol{I}_n),$$

we have that

$$\hat{\beta}|\boldsymbol{X} \sim N(\beta, \sigma^2\, (\boldsymbol{X}'\boldsymbol{X})^{-1}),$$

for finite n.

Follows from $\hat{\beta}$ being linear in $\boldsymbol{e}$ and the normal distribution being a local-scale family.

In this case,

we can construct hypothesis tests on linear contrasts of β whose size can be controlled exactly.

This does not extend beyond the normal model.

To get general results we will work under asymptotic approximations.