POLSCI 630 Introduction to Regression Methods Spring 2026

Lecture 3 — Finite Sample Property

Jiawei Fu, Duke University *Scribe: Jiawei Fu*

1 Overview

In the last lecture, we completed the OLS algebra. A natural next question is whether the OLS estimator is a *good* estimator. We will answer this from two perspectives: finite-sample properties and large-sample (asymptotic) properties. In this lecture, we focus on the finite-sample results. In particular, we study unbiasedness, the variance of the estimator, efficiency, and the finite-sample distribution under strong additional assumptions.

2 Unbiasedness

The OLS estimator is computed from a random sample, so it is itself a random variable. If we repeatedly drew samples from the same population, we would obtain different estimates each time. A basic desirable property is that, on average, the estimator equals the true parameter value. This is *unbiasedness*.

Definition 1. *An estimator $\hat{\beta}$ is unbiased if $\mathbb{E}[\hat{\beta}] = \beta$.*

In linear regression we often work with *conditional* moments given the regressor matrix X. Saying that an estimator is unbiased *conditional on X* means that, for every (fixed) realization of the regressor matrix, the estimator has mean equal to the true parameter value.

Theorem 2. *Consider the linear regression model $Y = X\beta + e$ and assume $\mathbb{E}[e \mid X] = 0$. Then the OLS estimator satisfies $\mathbb{E}[\hat{\beta} \mid X] = \beta$.*

The key assumption $\mathbb{E}[e \mid X] = 0$ is the *zero conditional mean* (exogeneity) condition. It implies that the conditional expectation function (CEF) is linear: $\mathbb{E}[Y \mid X] = X\beta$.

Proof. In this proof we work in the sample: X is the $n \times k$ regressor matrix, Y is the $n \times 1$ outcome vector, and the model is $Y = X\beta + e$. The OLS estimator is $\hat{\beta} = (X'X)^{-1}X'Y$ (assuming $X'X$ is invertible).

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

Now, the bias is $(X'X)^{-1}X'\mathbb{E}[e|X]$. If $\mathbb{E}[e|X] = 0$, we conclude no bias: $\mathbb{E}[\hat{\beta}|X] = \beta$. $\square$

Taking expectations again yields the unconditional result: $\mathbb{E}[\mathbb{E}[\hat{\beta} \mid X]] = \mathbb{E}[\hat{\beta}] = \beta$.

3 Variance Estimator

Because the estimator is random, we also care about its precision, which is summarized by its variance. Conditional on X,

$$\begin{aligned} Var(\hat{\beta}|X) &= Var[(X'X)^{-1}X'Y|X] \\ &= (X'X)^{-1}X'Var(Y|X)X(X'X)^{-1} \\ &= (X'X)^{-1}X'Var(e|X)X(X'X)^{-1} \end{aligned}$$

The third line uses that, conditional on X, the only random component of $Y = X\beta + e$ is e.

Now consider the middle term $Var(e \mid X)$, the $n \times n$ conditional variance–covariance matrix of the error vector e:

$$Var[e|X] = \mathbb{E}[ee'|X] = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

The ith diagonal element is $\mathbb{E}\left[e_i^2 \mid X\right] = \sigma_i^2$. The diagonal form above assumes *conditional uncorrelatedness* across observations: $\mathbb{E}\left[e_i e_j \mid X\right] = 0$ for $i \neq j$.[1] Let D denote this diagonal matrix. Then

$$Var(\hat{\beta} \mid X) = (X'X)^{-1}X'DX(X'X)^{-1}.$$

It is also useful to note that $X'DX = \sum_{i=1}^{n} X_i X_i' \sigma_i^2$, where X_i is the $k \times 1$ regressor vector for observation i.

3.1 Homoskedasticity

The classical linear model assumes *homoskedasticity*: the conditional error variance does not depend on X.

Assumption 3 (Homoskedasticity)**.** $\mathbb{E}\left[e^2|X\right] = \sigma^2(X) = \sigma^2$

Under homoskedasticity, $Var(e \mid X) = \sigma^2 I_n$, so

$$\begin{aligned} Var(\hat{\beta}|X) &= (X'X)^{-1}X'DX(X'X)^{-1} \\ &= (X'X)^{-1}\sigma^2 X'X(X'X)^{-1} \\ &= (X'X)^{-1}\sigma^2 \end{aligned}$$

The unconditional variance can be decomposed as

$$\begin{aligned} Var(\hat{\beta}) &= \mathbb{E}\left[Var(\hat{\beta}|X)\right] + Var[\mathbb{E}\left[\hat{\beta}|X\right]] \\ &= \mathbb{E}\left[(X'X)^{-1}\sigma^2\right] + Var[\beta] \\ &= \mathbb{E}\left[(X'X)^{-1}\right]\sigma^2 \end{aligned}$$

[1] This can fail under clustering, serial correlation, or other dependence, in which case $Var(e \mid X)$ is not diagonal and the "sandwich" form must be modified.

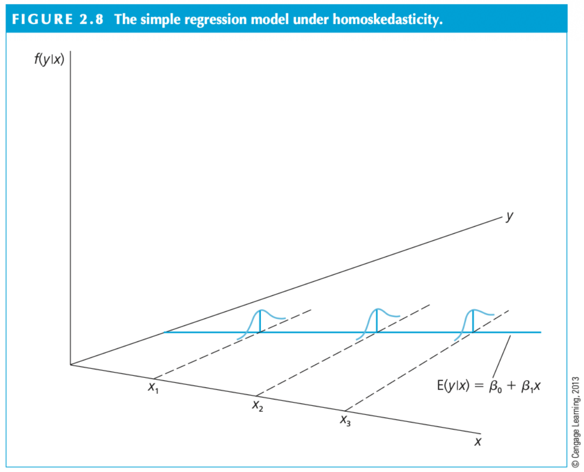

FIGURE 2.8 The simple regression model under homoskedasticity.

Figure 1: Homoskedasticity: From Wooldridge (2010)

where $Var(\mathbb{E}[\hat{\beta} \mid X]) = 0$ because $\mathbb{E}[\hat{\beta} \mid X] = \beta$ under exogeneity.

This variance contains the unknown parameter σ^2, which we must estimate. Since we cannot observe the error term e, a natural choice is the sample analogue

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i^2.$$

Is it unbiased? Let us check.

Recall that $\hat{e} = MY = M(X\beta + e) = Me$. Therefore,

$$\hat{\sigma}^2 = \frac{1}{n}\hat{e}'\hat{e} = \frac{1}{n}e'Me.$$

Since $e'Me$ is a scalar, it equals its own trace: $e'Me = \mathrm{tr}(e'Me)$. Using the cyclic property of the trace, $\mathrm{tr}(e'Me) = \mathrm{tr}(Mee')$. Thus,

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}^2|X] &= \frac{1}{n}tr(\mathbb{E}[Mee'|X]) \\
&= \frac{1}{n}tr(M\mathbb{E}[ee'|X]) \\
&= \frac{1}{n}tr(M\sigma^2 I) \\
&= \frac{1}{n}\sigma^2 tr(M)
\end{aligned}$$

What is the trace of the residual-maker matrix M? Since $M = I - P$ and $\mathrm{tr}(P) = k$, we have $\mathrm{tr}(M) = n - k$.

Therefore, $\mathbb{E}[\hat{\sigma}^2 \mid X] = \frac{n-k}{n}\sigma^2$, so $\hat{\sigma}^2$ is downward biased. The unbiased estimator is

$$s^2 = \frac{1}{n-k}\hat{e}'\hat{e} = \frac{1}{n-k}\sum_{i=1}^{n} \hat{e}_i^2.$$

In summary, under homoskedasticity the conditional covariance matrix takes the simple form $V_{\hat{\beta}}^0 = (X'X)^{-1}\sigma^2$. The classic conditional covariance matrix estimator is $\hat{V}_{\hat{\beta}}^0 = (X'X)^{-1}s^2$.

Remark 1. *Wooldridge (2010) shows that in the simple linear regression $Y = \beta_0 + X_1\beta_1 + e$, $Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{\sigma^2}{SST_x}$. In the multiple regression, $Var(\hat{\beta}_j|X) = \frac{\sigma^2}{SST_j(1-R_j^2)}$, where SST_j is the total sample variation in X_j, and R_j^2 is the R-squared from regressing X_j on all other independent variables. Now, you should be clear what affects the variance of OLS estimator.*

3.2 Heteroskedasticity

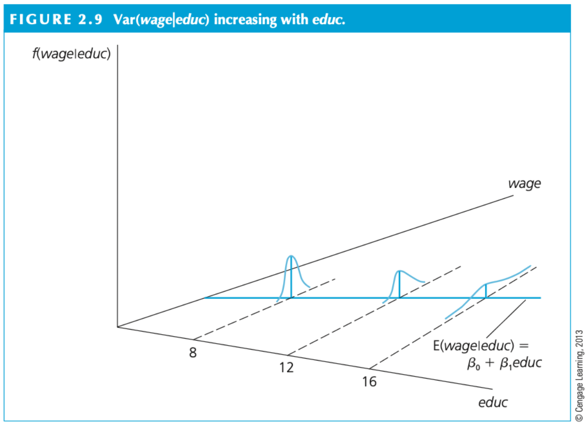

Figure 2: Heteroskedasticity: From Wooldridge (2010)

Homoskedasticity is a strong assumption and is often violated in practice. Let us return to the general variance formula

$$V_{\hat{\beta}} = Var(\hat{\beta} \mid X) = (X'X)^{-1}X'DX(X'X)^{-1}.$$

The formula contains the unknown diagonal matrix D. How can we estimate it? If the errors e_i were observable, then since $\mathbb{E}\left[ee' \mid X\right] = D$, a natural "oracle" (infeasible) estimator would be

$$\hat{V}_{\hat{\beta}}^{ideal} = (X'X)^{-1}\Big(\sum_{i=1}^n X_i X_i' e_i^2\Big)(X'X)^{-1}.$$

$$\begin{aligned}
\mathbb{E}\left[\hat{V}_{\hat{\beta}}^{ideal}|X\right] &= (X'X)^{-1}(\sum_{i=1}^n X_i X_i' \mathbb{E}\left[e_i^2|X\right])(X'X)^{-1} \\
&= (X'X)^{-1}(\sum_{i=1}^n X_i X_i' \sigma_i^2)(X'X)^{-1} \\
&= (X'X)^{-1}X'DX(X'X)^{-1} \\
&= V_{\hat{\beta}}
\end{aligned}$$

Replacing e_i^2 with the squared residuals $\hat{e}_i^2$ yields the heteroskedasticity-consistent covariance matrix estimator HC0:

$$\hat{V}_{\hat{\beta}}^{HC0} = (X'X)^{-1}(\sum_{i=1}^{n} X_i X_i' \hat{e}_i^2)(X'X)^{-1}$$

Is it a good estimator? Even under homoskedasticity, $\hat{e}_i^2$ is a biased proxy for e_i^2 because residuals are "shrunk" by leverage. A common degrees-of-freedom adjustment yields HC1:

$$\hat{V}_{\hat{\beta}}^{HC1} = \frac{n}{n-k}(X'X)^{-1}(\sum_{i=1}^{n} X_i X_i' \hat{e}_i^2)(X'X)^{-1}$$

One important point is that we generally cannot estimate each σ_i^2 (and hence D) unbiasedly, because each observation is only observed once. With a single realization per i, there is not enough information to recover the conditional variance for each individual.

HC1 is an ad hoc adjustment. We can do better. Under homoskedasticity, $\mathbb{E}[e_i^2 \mid X] = \sigma^2$, and

$$\mathbb{E}[\hat{e}\hat{e}' \mid X] = \mathbb{E}[Mee'M \mid X] = M\mathbb{E}[ee' \mid X]M = M\sigma^2.$$

The ith diagonal element of M is $1 - h_{ii}$, so $\mathbb{E}[\hat{e}_i^2 \mid X] = (1 - h_{ii})\sigma^2$.

Remark 2 (Leverage values)**.** *Recall $M = 1 - P$. Therefore, $h_{ii} = X_i'(X'X)^{-1}X_i$ is actually the diagonal elements of the projection matrix. We call it leverage values for the regressor matrix X. Heuristically, h_{ii} measures how "far" X_i is from the center of the regressor cloud. Observations with large h_{ii} have high leverage and can exert substantial influence on fitted values and OLS coefficients.*

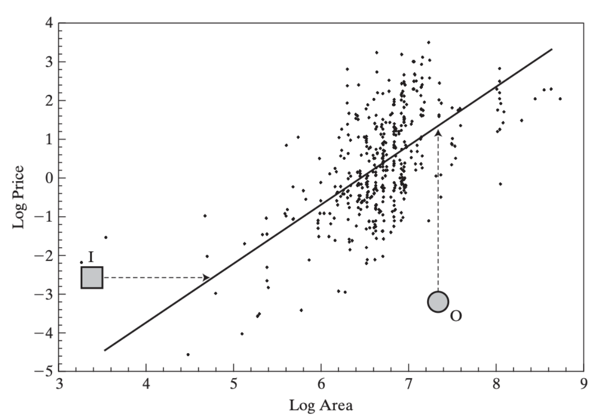

Figure 3: From Greene (2000)

Remark 3 (Outlier)**.** *We usually refer to point O in the figure as an outlier, and point I as an influential observation with high leverage. In principle, an "outlier" is an observation that appears to be outside the reach of the model, perhaps because it arises from a different data-generating process.*

This suggests defining standardized residuals $\overline{e}_i = (1 - h_{ii})^{-1/2}\hat{e}_i$, for which $\mathbb{E}[\overline{e}_i^2 \mid X] = \sigma^2$ under homoskedasticity. Replacing $\hat{e}_i$ with $\overline{e}_i$ in the variance estimator yields HC2.

$$\hat{V}_{\hat{\beta}}^{HC2} = (X'X)^{-1}(\sum_{i=1}^{n}(1-h_{ii})^{-1}X_iX_i'\hat{e}_i^2)(X'X)^{-1}$$

In general, $\hat{V}_{\hat{\beta}}^{HC0} < \hat{V}_{\hat{\beta}}^{HC2}$. Under homoskedasticity, $\mathbb{E}\,[\hat{V}_{\hat{\beta}}^{HC2}|X] = (X'X)^{-1}\sigma^2$, and $\mathbb{E}\,[\hat{V}_{\hat{\beta}}^{HC0}|X] < (X'X)^{-1}\sigma^2$.

Unfortunately, many regression packages default to $\hat{V}_{\hat{\beta}}^{HC0}$, so users must intentionally select a robust covariance matrix estimator.

A much more conservative estimator is HC3:

$$\hat{V}_{\hat{\beta}}^{HC3} = (X'X)^{-1}(\sum_{i=1}^{n}(1-h_{ii})^{-2}X_iX_i'\hat{e}_i^2)(X'X)^{-1}$$

It is weakly larger than the correct variance for any realization of X: $\mathbb{E}\,[\hat{V}_{\hat{\beta}}^{HC3}|X] \geq V_{\hat{\beta}}$.

4 Distribution

If we additionally assume normality, $e \mid X \sim N(0, \sigma^2 I_n)$, then $\hat{\beta} \mid X$ is exactly normal in finite samples:

$$\hat{\beta} \mid X \sim N\big(\beta,\ \sigma^2(X'X)^{-1}\big).$$

This is a strong assumption; in most applications we instead rely on large-sample (asymptotic) approximations, which will be covered in the next lecture.

To see the distribution of $\hat{\beta}$, note that $\hat{\beta} - \beta = (X'X)^{-1}X'e$. Therefore

$$\begin{aligned}
\hat{\beta} - \beta|X &\sim (X'X)^{-1}X'N(0, I_n\sigma^2) \\
&= N(\sigma^2(X'X)^{-1}X'X(X'X)^{-1}) \\
&= N(0, \sigma^2(X'X)^{-1})
\end{aligned}$$

5 Gauss-Markov Theorem

Why do we love OLS? The classical Gauss–Markov theorem gives rise to the description of OLS as the *best linear unbiased estimator* (BLUE).

Theorem 4 (Gauss-Markov)**.** *In the linear regression model $Y = X\beta + e$ where $\mathbb{E}\,[e|X] = 0$ and $Var[e|X] = \sigma^2 I_n$, if $\tilde{\beta}$ is an unbiased estimator of β, then $Var[\tilde{\beta}|X] \geq \sigma^2(X'X)^{-1}$*

It provides a lower bound on the covariance matrix of linear unbiased estimators under the assumption of homoskedasticity. It states that no linear unbiased estimator can have a variance matrix smaller (in the positive semidefinite sense) than the variance of OLS. Consequently, OLS is efficient within the class of linear unbiased estimators.

The theory also shows that $\sigma^2(X'X)^{-1}$ is the semiparametric efficiency bound, but that result is beyond the scope of this course. Instead, we show that OLS $\hat{\beta}$ is the minimum-variance *linear* unbiased estimator of β.

Proof. Let $\tilde{\beta} = CY$ be any linear estimator, where C is $k \times n$ matrix.

We have $\mathbb{E}[\tilde{\beta} \mid X] = C\,\mathbb{E}[Y \mid X] = CX\beta$. Because $\tilde{\beta}$ is unbiased for β for all β, we must have $CX = I_k$.

For the variance, $Var[\tilde{\beta} \mid X] = Var[CY \mid X] = C\,Var(Y \mid X)\,C' = \sigma^2 CC'$.

Now define $B = C - (X'X)^{-1}X'$, Then, $C = B + (X'X)^{-1}X'$,

$$Var[\tilde{\beta}|X] = \sigma^2[(B + (X'X)^{-1}X')(B + (X'X)^{-1}X')'].$$

We know that $CX = I_k = BX + (X'X)^{-1}X'X$, so $BX = 0$. Therefore,

$$Var[\tilde{\beta}|X] = \sigma^2(X'X)^{-1} + \sigma^2 BB' = Var[\hat{\beta}|X] + \sigma^2 DD' \geq Var[\hat{\beta}|X]$$

□

References

Greene, W. H. (2000). Econometric analysis 4th edition. *International edition, New Jersey: Prentice Hall*, (pp. 201–215).

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.