

## LECTURE 2 LINEAR REGRESSION MODEL AND OLS

### Definitions

A common objective in econometrics is to study the effect of a group of variables,  $X_i$ , called the *regressors*, on a variable  $Y_i$ , called the *dependent variable*. The econometrician observes data:

$$(Y_1, X_1), (Y_2, X_2), \dots, (Y_n, X_n), \quad (1)$$

where for  $i = 1, \dots, n$ ,  $Y_i$  is a random variable and  $X_i$  is a random  $k$ -vector:

$$X_i = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ik} \end{pmatrix}.$$

A pair  $(Y_i, X_i)$  is called an *observation*, and the collection of observations in (1) is called the *sample*. The vector  $X_i$  collects the values of  $k$  variables for observation  $i$ .

The joint distribution of the sample (1) is called the *population*. The population does not correspond to any physical population; rather, it refers to a probability space. In a *cross-sectional* framework (each observation is a different individual, firm, etc.), it is often natural to assume that all observations are independently drawn from the same distribution. In this case, the population is described by the distribution of a single observation  $(Y_1, X_1)$ . Equivalently, the observations  $(Y_i, X_i)$  are iid for  $i = 1, \dots, n$ . The iid assumption does not imply that  $Y_i$  and  $X_i$  are independent, but rather that the random vector  $(Y_i, X_i)$  is independent of  $(Y_j, X_j)$  for  $i \neq j$ . At the same time,  $Y_i$  and  $X_i$  can still be related.

In cross-sections, the relationship between the regressors and the dependent variable is modeled through the conditional expectation  $E[Y_i | X_i]$ . The deviation of  $Y_i$  from its conditional expectation is called the *error* (or the *disturbance*):

$$U_i = Y_i - E[Y_i | X_i]. \quad (2)$$

Unlike  $X_i$  and  $Y_i$ , the error  $U_i$  is not observable, since the conditional expectation function is unknown to the econometrician.

In the *parametric* framework, it is assumed that the conditional expectation function depends on a number of unknown constants or *parameters*, and that the functional form of  $E[Y_i | X_i]$  is known. In the linear regression model, it is assumed that  $E[Y_i | X_i]$  is *linear in the parameters*:

$$\begin{aligned} E[Y_i | X_i] &= \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} \\ &= X_i^\top \beta, \end{aligned} \quad (3)$$

where

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$$

is a  $k$ -vector of unknown constants. The linearity of  $E[Y_i | X_i]$  can be justified, for example, by assuming that  $(Y_i, X_i)$  is jointly multivariate normal.

Letting  $m(x) := E[Y_i | X_i = x] = x^\top \beta$ , we have  $\beta_j = \partial m(x) / \partial x_j$ , so  $\beta$  is a vector of *marginal effects*:  $\beta_j$  gives the change in the conditional mean of  $Y_i$  per unit change in  $X_{ij}$ , holding the other regressors  $X_{il}$  ( $l \neq j$ ) fixed. One of the objectives is *estimation* of the unknown  $\beta$  from the sample (1).

Combining (2) and (3) gives

$$Y_i = X_i^\top \beta + U_i. \quad (4)$$

By definition (2),

$$E[U_i | X_i] = 0.$$

This implies that the regressors contain no information about the deviation of  $Y_i$  from its conditional expectation. By the law of iterated expectations (LIE), the errors have zero mean:

$$E[U_i] = E[E[U_i | X_i]] = E[0] = 0.$$

If  $(Y_i, X_i)$  are iid, then since  $U_i$  is the same measurable function of  $(Y_i, X_i)$  for every  $i$ , the errors  $\{U_i : i = 1, \dots, n\}$  are iid as well.

In the *classical regression model*, the variance of the error  $U_i$  is assumed to be independent of the regressors and to be constant across observations:

$$\text{Var}(U_i | X_i) = \sigma^2,$$

for some constant  $\sigma^2 > 0$ . This property is called *homoskedasticity*.

## Assumptions

In this section, we formally define the linear regression model. Set

$$\begin{aligned} X &= \begin{pmatrix} X_1^\top \\ X_2^\top \\ \vdots \\ X_n^\top \end{pmatrix} \\ &= \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1k} \\ X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nk} \end{pmatrix}, \\ Y &= \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \text{ and} \\ U &= \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix}. \end{aligned}$$

The four classical regression assumptions are as follows:

- (A1)  $Y = X\beta + U$  for some  $\beta \in \mathbb{R}^k$ .
- (A2)  $E[U | X] = 0$  a.s.
- (A3)  $\text{Var}(U | X) = \sigma^2 I_n$  a.s., for some  $\sigma^2 > 0$ .
- (A4)  $\text{rank}(X) = k$  a.s. (in particular,  $n \geq k$ ).<sup>1</sup>

<sup>1</sup>The column (row) rank of a matrix is the maximum number of linearly independent columns (rows). For any matrix, the column and row ranks are equal. If  $A$  is an  $n \times k$  matrix, then  $\text{rank}(A) \leq \min(n, k)$ . If  $\text{rank}(A) = n$  (or  $\text{rank}(A) = k$ ), we say that  $A$  has full row (column) rank. Properties:

$$\begin{aligned} \text{rank}(A) &= \text{rank}(A^\top) = \text{rank}(A^\top A) = \text{rank}(AA^\top), \\ \text{rank}(AB) &\leq \min(\text{rank}(A), \text{rank}(B)), \\ \text{rank}(AB) &= \text{rank}(A) \text{ if } B \text{ is square and of full rank.} \end{aligned}$$

In the classical regression model, instead of conditioning on the regressors, one can assume that  $X$  is non-random; that is, the values of  $X$  are fixed in repeated samples. In this case, (A2) becomes  $E[U] = 0$  and (A3) becomes  $\text{Var}(U) = \sigma^2 I_n$ . Since conditioning on  $X$  is equivalent to treating  $X$  as fixed, the two sets of assumptions lead to the same results.

For inference purposes, sometimes it is assumed that:

**(A5)**  $U | X \sim N(0, \sigma^2 I_n)$ .

In the case of fixed regressors, it is assumed instead that the unconditional distribution of the errors is  $N(0, \sigma^2 I_n)$ . Assumptions (A1)–(A5) define the *classical normal regression model*, under which

$$Y | X \sim N(X\beta, \sigma^2 I_n).$$

Since the covariance matrix  $\sigma^2 I_n$  in (A5) is diagonal and the conditional distribution is jointly normal, the errors are independent (conditional on  $X$ ). Assumptions (A1)–(A4) alone do not imply independence across observations. For many important results, independence across observations is not required. Nevertheless, sometimes we would like to assume independence without normality:

**(A6)**  $\{(Y_i, X_i) : i = 1, \dots, n\}$  are iid.

In the case of fixed regressors, (A6) can be replaced with the assumption that  $\{U_i : i = 1, \dots, n\}$  are iid.

Assumption (A2) says that  $U$  is mean-independent of  $X$ , which is a strong requirement (sometimes called *strict exogeneity* when stated for the full vector  $U$  given the full design  $X$ ). Mean independence is difficult to justify in many economic applications. However, many important results hold under the weaker condition of uncorrelatedness:

**(A2\*)** For  $i = 1, \dots, n$ ,  $E[U_i X_i] = 0$  and  $E[U_i] = 0$  (with implicit moment existence:  $E\|X_i\| < \infty$  and  $E|U_i| < \infty$ ).

Under (A2\*), the expression  $X_i^\top \beta$  does not have the interpretation of the conditional expectation. In this case, one can treat (4) as a *data-generating process*.

Assumption (A3) implies that the errors  $U_i$  have the same conditional variance  $\text{Var}(U_i | X) = \sigma^2$  a.s. for all  $i$ , and are conditionally uncorrelated:  $E[U_i U_j | X] = 0$  a.s. for  $i \neq j$ . If desired, independence of the errors can be achieved through (A5) or through Assumptions (A1) and (A6).

Assumption (A4) says that the columns of  $X$  are linearly independent (almost surely). If (A4) is violated, then one or more regressors are exact linear combinations of the others (*perfect multicollinearity*) and should be dropped from the regression.

Often, the first column of the matrix  $X$  is a column of ones; its coefficient  $\beta_1$  is then called the *intercept* and gives the conditional mean  $E[Y_i | X_{i,2} = \dots = X_{i,k} = 0]$  of the dependent variable when all non-intercept regressors are zero.

## Estimation by the method of moments

One of the objectives of econometric analysis is *estimation* of unknown parameters  $\beta$  and  $\sigma^2$ . An *estimator* is any function of the sample  $\{(Y_i, X_i) : i = 1, \dots, n\}$ . An estimator cannot depend directly on the unobserved errors  $U_i$  or unknown parameters such as  $\beta$ ; it can depend on unobservables only through the observed variables  $Y$  and  $X$ . An estimator is not unique in general: several estimators may exist for the same parameter.

One of the oldest methods of finding estimators is called the *method of moments* (MM). It replaces population moments (expectations) with the corresponding sample moments (averages). Assumptions (A2) or (A2\*) imply that at the true value of  $\beta$ ,

$$\begin{aligned} 0 &= E[U_i X_i] \\ &= E[(Y_i - X_i^\top \beta) X_i]. \end{aligned} \tag{5}$$

Let  $\widehat{\beta}$  be an estimator of  $\beta$ . According to the MM principle, we replace the expectation in (5) with the sample average:

$$\begin{aligned} 0 &= n^{-1} \sum_{i=1}^n (Y_i - X_i^\top \widehat{\beta}) X_i \\ &= n^{-1} \sum_{i=1}^n X_i Y_i - n^{-1} \sum_{i=1}^n X_i X_i^\top \widehat{\beta}. \end{aligned}$$

Here  $Y_i - X_i^\top \widehat{\beta}$  is a scalar, so it commutes past  $X_i$ , and  $(X_i^\top \widehat{\beta}) X_i = X_i X_i^\top \widehat{\beta}$  (a  $k$ -vector). Solving for  $\widehat{\beta}$ , one obtains:

$$\begin{aligned} \widehat{\beta} &= \left( n^{-1} \sum_{i=1}^n X_i X_i^\top \right)^{-1} n^{-1} \sum_{i=1}^n X_i Y_i \\ &= \left( \sum_{i=1}^n X_i X_i^\top \right)^{-1} \sum_{i=1}^n X_i Y_i \\ &= (X^\top X)^{-1} X^\top Y. \end{aligned} \tag{6}$$

The matrix  $\sum_{i=1}^n X_i X_i^\top = X^\top X$  is invertible under Assumption (A4). To show that  $X^\top X = \sum_{i=1}^n X_i X_i^\top$ , expand

$$\begin{aligned} X^\top X &= \begin{pmatrix} X_1^\top \\ X_2^\top \\ \vdots \\ X_n^\top \end{pmatrix}^\top \begin{pmatrix} X_1^\top \\ X_2^\top \\ \vdots \\ X_n^\top \end{pmatrix} \\ &= (X_1 \quad X_2 \quad \dots \quad X_n) \begin{pmatrix} X_1^\top \\ X_2^\top \\ \vdots \\ X_n^\top \end{pmatrix} \\ &= X_1 X_1^\top + X_2 X_2^\top + \dots + X_n X_n^\top \\ &= \sum_{i=1}^n X_i X_i^\top. \end{aligned}$$

The expression  $X_i^\top \widehat{\beta}$  gives the *estimated regression line*, with  $\widehat{Y}_i = X_i^\top \widehat{\beta}$  being the *predicted value* of  $Y_i$ , and  $\widehat{U}_i = Y_i - X_i^\top \widehat{\beta}$  being the *sample residual*,

$$\widehat{U} = Y - X\widehat{\beta}.$$

The vector  $\widehat{U}$  is a function of the estimator of  $\beta$ . For the MM estimator, the sample residuals satisfy the *normal equation*:

$$\begin{aligned} 0 &= X^\top \widehat{U} \\ &= \sum_{i=1}^n \widehat{U}_i X_i \\ &= \begin{pmatrix} \sum_{i=1}^n \widehat{U}_i X_{i1} \\ \sum_{i=1}^n \widehat{U}_i X_{i2} \\ \vdots \\ \sum_{i=1}^n \widehat{U}_i X_{ik} \end{pmatrix}. \end{aligned} \tag{7}$$

If the model contains an intercept, that is,  $X_{i1} = 1$  for all  $i$ , then the normal equation implies that  $\sum_{i=1}^n \widehat{U}_i = 0$ .

To estimate  $\sigma^2$ , write

$$\begin{aligned}\sigma^2 &= \mathbb{E}[U_i^2] \\ &= \mathbb{E}[(Y_i - X_i^\top \beta)^2].\end{aligned}$$

The MM analog replaces the population expectation with the sample average:

$$n^{-1} \sum_{i=1}^n (Y_i - X_i^\top \beta)^2.$$

Since  $\beta$  is unknown, we replace it with its MM estimator  $\widehat{\beta}$ :

$$\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^n (Y_i - X_i^\top \widehat{\beta})^2.$$

## Least Squares

Let  $b$  be a candidate value for  $\beta$ . The *ordinary least squares* (OLS) estimator of  $\beta$  minimizes the *sum of squared errors*:

$$\begin{aligned}S(b) &= \sum_{i=1}^n (Y_i - X_i^\top b)^2 \\ &= (Y - Xb)^\top (Y - Xb).\end{aligned}$$

The MM estimator  $\widehat{\beta}$  coincides with the OLS estimator. To verify this, write

$$\begin{aligned}S(b) &= (Y - Xb)^\top (Y - Xb) \\ &= (Y - X\widehat{\beta} + X\widehat{\beta} - Xb)^\top (Y - X\widehat{\beta} + X\widehat{\beta} - Xb) \\ &= (Y - X\widehat{\beta})^\top (Y - X\widehat{\beta}) \\ &\quad + (X\widehat{\beta} - Xb)^\top (X\widehat{\beta} - Xb) \\ &\quad + 2(Y - X\widehat{\beta})^\top (X\widehat{\beta} - Xb) \\ &= (Y - X\widehat{\beta})^\top (Y - X\widehat{\beta}) \\ &\quad + (\widehat{\beta} - b)^\top X^\top X (\widehat{\beta} - b) \\ &\quad + 2\widehat{U}^\top X (\widehat{\beta} - b) \quad (\text{zero by the normal equation, since } X^\top \widehat{U} = 0 \text{ implies } \widehat{U}^\top X = 0^\top) \\ &= (Y - X\widehat{\beta})^\top (Y - X\widehat{\beta}) + (\widehat{\beta} - b)^\top X^\top X (\widehat{\beta} - b).\end{aligned}$$

Since the first term  $(Y - X\widehat{\beta})^\top (Y - X\widehat{\beta})$  does not depend on  $b$ , minimizing  $S(b)$  is equivalent to minimizing  $(\widehat{\beta} - b)^\top X^\top X (\widehat{\beta} - b)$ . If  $X$  is of full column rank, as assumed in (A4), then  $X^\top X$  is positive definite: for any  $v \neq 0$ , full column rank implies  $Xv \neq 0$ , so  $v^\top X^\top X v = \|Xv\|^2 > 0$ . Therefore

$$(\widehat{\beta} - b)^\top X^\top X (\widehat{\beta} - b) \geq 0,$$

with equality if and only if  $\hat{\beta} - b = 0$ . Hence  $S(b) \geq S(\hat{\beta})$ , with equality only at  $b = \hat{\beta}$ , so  $\hat{\beta}$  is the unique minimizer of  $S(b)$ .

Alternatively, one can obtain  $\hat{\beta}$  by taking the derivative of  $S(b)$  with respect to  $b$  and solving the first-order condition  $\partial S(b)/\partial b = 0$ . Write

$$S(b) = Y^\top Y - 2b^\top X^\top Y + b^\top X^\top X b.$$

Using the fact that for a symmetric matrix  $A$ ,

$$\frac{\partial x^\top A x}{\partial x} = 2Ax,$$

and that  $X^\top X$  is symmetric (since  $(X^\top X)^\top = X^\top X$ ), the first-order condition is

$$\left. \frac{\partial S(b)}{\partial b} \right|_{b=\hat{\beta}} = -2X^\top Y + 2X^\top X \hat{\beta} = 0. \quad (8)$$

Solving for  $\hat{\beta}$ , one obtains:

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y. \quad (9)$$

The second-order condition is satisfied since the Hessian  $\partial^2 S(b)/\partial b \partial b^\top = 2X^\top X$  is positive definite under (A4), confirming that  $\hat{\beta}$  is the unique minimizer of  $S(b)$ .

The first-order condition (8) can also be written as  $X^\top(Y - X\hat{\beta}) = 0$ , which is the normal equation (7). Thus the FOC for the OLS criterion is identical to the MM normal equation, giving an alternative direct proof that the OLS and MM estimators coincide.

## Properties of $\hat{\beta}$

1.  $\hat{\beta}$  is a *linear* estimator: an estimator  $b$  is called linear if it can be written as  $b = AY$ , where  $A$  is a matrix that depends on  $X$  alone and does not depend on  $Y$ . In the case of the OLS,  $A = (X^\top X)^{-1} X^\top$ .
2. Under Assumptions (A1), (A2), and (A4),  $\hat{\beta}$  is an *unbiased* estimator, that is,

$$\mathbb{E}[\hat{\beta}] = \beta.$$

To show unbiasedness, substitute the expression for  $Y$  from Assumption (A1) into equation (9):

$$\begin{aligned} \hat{\beta} &= (X^\top X)^{-1} X^\top (X\beta + U) \\ &= \beta + (X^\top X)^{-1} X^\top U. \end{aligned} \quad (10)$$

Next, consider the conditional expectation of  $\hat{\beta}$  given  $X$ :

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

Now,

$$\begin{aligned} \mathbb{E}[(X^\top X)^{-1} X^\top U | X] &= (X^\top X)^{-1} X^\top \mathbb{E}[U | X] \\ &= 0, \end{aligned}$$

since, by Assumption (A2),  $\mathbb{E}[U | X] = 0$ . Therefore,

$$\mathbb{E}[\hat{\beta} | X] = \beta. \quad (11)$$

Finally, the LIE implies that

$$\begin{aligned} \mathbb{E}[\widehat{\beta}] &= \mathbb{E}[\mathbb{E}[\widehat{\beta} \mid X]] \\ &= \beta. \end{aligned}$$

Equation (11) shows that  $\widehat{\beta}$  is conditionally unbiased given  $X$ . Conditional unbiasedness does not hold under Assumption (A2\*) alone.

3. Under Assumptions (A1), (A2), and (A4),

$$\text{Var}(\widehat{\beta} \mid X) = (X^\top X)^{-1} X^\top \mathbb{E}[UU^\top \mid X] X (X^\top X)^{-1}.$$

Under homoskedastic errors, that is, under Assumption (A3), the expression for the conditional variance simplifies to

$$\text{Var}(\widehat{\beta} \mid X) = \sigma^2 (X^\top X)^{-1}.$$

To show this, we start with the definition of the variance:

$$\begin{aligned} \text{Var}(\widehat{\beta} \mid X) &= \mathbb{E} \left[ (\widehat{\beta} - \mathbb{E}[\widehat{\beta} \mid X]) (\widehat{\beta} - \mathbb{E}[\widehat{\beta} \mid X])^\top \mid X \right] \\ &= \mathbb{E} \left[ (\widehat{\beta} - \beta) (\widehat{\beta} - \beta)^\top \mid X \right] \quad (\text{by (11)}) \\ &= \mathbb{E} \left[ (X^\top X)^{-1} X^\top UU^\top X (X^\top X)^{-1} \mid X \right] \quad (\text{by (10)}) \\ &= (X^\top X)^{-1} X^\top \mathbb{E}[UU^\top \mid X] X (X^\top X)^{-1}. \end{aligned}$$

Under homoskedasticity,  $\mathbb{E}[UU^\top \mid X] = \sigma^2 I_n$ , and

$$\begin{aligned} (X^\top X)^{-1} X^\top \mathbb{E}[UU^\top \mid X] X (X^\top X)^{-1} &= (X^\top X)^{-1} X^\top (\sigma^2 I_n) X (X^\top X)^{-1} \\ &= \sigma^2 (X^\top X)^{-1} X^\top X (X^\top X)^{-1} \\ &= \sigma^2 (X^\top X)^{-1}. \end{aligned}$$

In the case of fixed regressors,  $\text{Var}(\widehat{\beta}) = \sigma^2 (X^\top X)^{-1}$ .

4. If we add normality of the errors, that is, under Assumptions (A1)–(A5), we obtain the following result:

$$\widehat{\beta} \mid X \sim N \left( \beta, \sigma^2 (X^\top X)^{-1} \right).$$

It is sufficient to show that, conditional on  $X$ , the distribution of  $\widehat{\beta}$  is normal. Then,  $\widehat{\beta} \mid X \sim N(\mathbb{E}[\widehat{\beta} \mid X], \text{Var}(\widehat{\beta} \mid X))$ . To see this, observe that conditional on  $X$ , the matrix  $(X^\top X)^{-1} X^\top$  is non-random, so  $\widehat{\beta} = (X^\top X)^{-1} X^\top Y$  is a linear function of  $Y$  given  $X$ . Combined with Assumption (A5), this implies that  $Y \mid X$  is normal (see the section on the normal distribution in Lecture 1), so  $\widehat{\beta} \mid X$  is normal. Again, in the case of fixed regressors, we simply omit conditioning on  $X$ ,

$$\widehat{\beta} \sim N \left( \beta, \sigma^2 (X^\top X)^{-1} \right).$$

5. *Efficiency* (the Gauss–Markov theorem): Under Assumptions (A1)–(A4), the OLS estimator is the Best Linear Unbiased Estimator of  $\beta$  (BLUE), where “best” means having the smallest conditional variance in the matrix (positive-semidefinite) sense: for any linear and unbiased estimator  $b$ ,  $\text{Var}(b \mid X) - \text{Var}(\widehat{\beta} \mid X)$  is a positive semidefinite matrix:

$$\text{Var}(b \mid X) - \text{Var}(\widehat{\beta} \mid X) \geq 0.$$

Furthermore, if  $\tilde{\beta}$  is a linear unbiased estimator and  $\text{Var}(\tilde{\beta} | X) = \text{Var}(\hat{\beta} | X)$ , then  $\tilde{\beta} = \hat{\beta}$  almost surely.

Since the theorem concerns the conditional variance, the unbiasedness condition refers to conditional unbiasedness:  $\text{E}[b | X] = \beta$ .

**Proof:** Let  $b$  be a linear and unbiased estimator of  $\beta$ :

$$\begin{aligned} b &= AY, \\ \text{E}[b | X] &= \beta. \end{aligned}$$

These conditions imply that  $AX = I_k$  almost surely. Indeed,

$$\begin{aligned} \text{E}[b | X] &= \text{E}[A(X\beta + U) | X] \\ &= AX\beta + A\text{E}[U | X]. \end{aligned}$$

By Assumption (A2),  $\text{E}[U | X] = 0$ , so  $\text{E}[b | X] = AX\beta$ . For unbiasedness to hold for every  $\beta \in \mathbb{R}^k$ , we need  $AX = I_k$  almost surely. Next, we show that  $\text{Cov}(\hat{\beta}, b | X) = \text{Var}(\hat{\beta} | X)$ :

$$\begin{aligned} \text{Cov}(\hat{\beta}, b | X) &= \text{E}[(\hat{\beta} - \beta)(b - \beta)^\top | X] \\ &= \text{E}[(X^\top X)^{-1} X^\top U U^\top A^\top | X] \\ &= (X^\top X)^{-1} X^\top \text{E}[U U^\top | X] A^\top \\ &= \sigma^2 (X^\top X)^{-1} X^\top A^\top \quad (\text{since, by Assumption (A3), } \text{E}[U U^\top | X] = \sigma^2 I_n) \\ &= \sigma^2 (X^\top X)^{-1} \quad (\text{since } AX = I_k \text{ implies } X^\top A^\top = I_k) \\ &= \text{Var}(\hat{\beta} | X). \end{aligned}$$

Finally,

$$\begin{aligned} \text{Var}(\hat{\beta} - b | X) &= \text{Var}(\hat{\beta} | X) - \text{Cov}(\hat{\beta}, b | X) - \text{Cov}(b, \hat{\beta} | X) + \text{Var}(b | X) \\ &= \text{Var}(b | X) - \text{Var}(\hat{\beta} | X). \end{aligned} \tag{12}$$

Since any variance-covariance matrix is positive semidefinite,

$$\text{Var}(b | X) - \text{Var}(\hat{\beta} | X) \geq 0.$$

To show uniqueness, suppose that there is a linear unbiased estimator  $\tilde{\beta}$  such that  $\text{Var}(\tilde{\beta} | X) = \text{Var}(\hat{\beta} | X)$ . Then, by (12) applied with  $b = \tilde{\beta}$ ,  $\text{Var}(\hat{\beta} - \tilde{\beta} | X) = 0$ . Since a random vector with zero conditional variance equals its conditional mean almost surely,  $\tilde{\beta} = \hat{\beta} + c(X)$  for some  $k$ -vector-valued function  $c(X)$  that depends only on  $X$ . Since both  $\hat{\beta}$  and  $\tilde{\beta}$  are conditionally unbiased given  $X$ , taking conditional expectations gives  $c(X) = 0$  almost surely. Combined with  $\tilde{\beta} = \hat{\beta} + c(X)$ , this yields  $\tilde{\beta} = \hat{\beta}$  almost surely.  $\square$

Assumption (A3),  $\text{E}[U U^\top | X] = \sigma^2 I_n$ , plays a crucial role in the proof: without it, the cancellation  $X^\top \text{E}[U U^\top | X] A^\top = \sigma^2 X^\top A^\top$  used above would fail. Under known heteroskedasticity, the generalized least squares (GLS) estimator achieves a strictly smaller conditional variance in the positive-semidefinite sense.