

# 3 Geometric Interpretation of OLS, Mean Variance of OLS, Partitioned Regression

## 3.1 Geometric Interpretation

Consider estimation of  $\beta$  in the model:

$$y_i = x_i' \beta + \varepsilon_i, \quad i = 1, \dots, n$$

This is equivalent in matrix form to:  $Y = X\beta + \varepsilon$

The OLS estimator is:  $\hat{\beta} = (X'X)^{-1}X'Y$

### Definition 3.1.1

The Projection Matrix is defined as:

$$P_X = X(X'X)^{-1}X'$$

The Residual Maker Matrix is defined as:

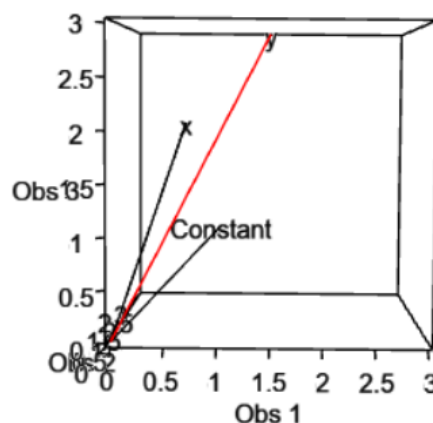
$$M_X = I - P_X$$

Then

$$\hat{Y} = X\hat{\beta} = P_X Y$$

$$\hat{\varepsilon} = Y - \hat{Y} = M_X Y$$

**Claim 3.1.1.**  $P_X$  and  $M_X$  are symmetric and idempotent.



Thus,  $\hat{Y} = X\hat{\beta}$  is the orthogonal projection of the  $n$ -dimensional vector  $Y$  onto the subspace spanned by the columns of  $X$ . Each column of  $X$  represents the  $n$  values that each regressor takes for every observation.

The "subspace" spanned by the columns of  $X$  is the set of all linear combinations of the columns of  $X$ . The orthogonal projection of  $Y$  onto this subspace is the closest point in the subspace to  $Y$ . This is because we solve:

$$\hat{\beta} = \underset{b}{\operatorname{argmin}} \sum (y_i - x'_i b)^2 = \underset{b}{\operatorname{argmin}} (Y - Xb)'(Y - Xb) = \underset{b}{\operatorname{argmin}} \|Y - Xb\|^2$$

**Example.**  $k = n$

Clearly if we had  $k=n$  regressors, then the columns of  $X$  would span the entire  $n$ -dimensional space and the projection would be the identity matrix. In this case,  $\hat{Y} = Y$ , and the residuals would be zero.

### 3.1.1 The Residual Vector

The difference between  $Y$  and the projection of  $Y$  onto the subspace is the residual vector  $\hat{\varepsilon}$ .

**Claim 3.1.2.** The residual vector is orthogonal to the subspace spanned by the columns of  $X$  and so is orthogonal to each column of  $X$   $X'\hat{\varepsilon} = 0$

**Proof.** Intuitively: This is because the projection of  $Y$  onto the subspace is the closest point in the subspace to  $Y$ . If the residual vector were not orthogonal to the subspace, then we could move the projection of  $Y$  onto the subspace along the residual vector and get a point that is closer to  $Y$ . This would contradict the fact that the projection of  $Y$  onto the subspace is the closest point in the subspace to  $Y$ .

Algebraically:

$$X'\hat{\varepsilon} = X'(Y - \hat{Y}) = X'(Y - P_X Y) = X'(Y - X(X'X)^{-1}X'Y) = 0$$

□

## 3.2 Conditional Mean and Variance of OLS

### 3.2.1 Conditional Mean

**Claim 3.2.1.**  $\hat{\beta}$  is a conditionally unbiased estimator of  $\beta$

$$\mathbb{E}[\hat{\beta}|X] = \beta$$

**Proof.**

$$\hat{\beta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + \varepsilon) = \beta + (X'X)^{-1}X'\varepsilon$$

$$\mathbb{E}[\hat{\beta}|X] = \beta + (X'X)^{-1}X'\mathbb{E}[\varepsilon|X] \stackrel{1}{=} \beta$$

1. via strict exogeneity  $\mathbb{E}[\varepsilon|X] = 0$ , do not need iid (e.g. can have a regressor  $x_i = i$ )

□

Also only need strict exogeneity for a causal interpretation of  $\beta$ .

**Claim 3.2.2.**  $\hat{\beta}$  is an unconditionally unbiased estimator of  $\beta$ , provided expectations exist

$$\mathbb{E}[\hat{\beta}|X] = \beta$$

**Proof.** via law of iterated expectations

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}[\mathbb{E}[\hat{\beta}|X]] = \mathbb{E}[\beta] = \beta$$

□

### 3.2.2 Conditional Variance

**Theorem 3.2.1.**

$$\text{Var}(\hat{\beta}|X) = \sigma^2(X'X)^{-1}$$

**Lemma 3.2.1.** Unconditional Variance of a vector:

$$\text{Var}(z) = \mathbb{E}[(z - \mathbb{E}[z])(z - \mathbb{E}[z])'] = \mathbb{E}[zz'] - \mathbb{E}[z]\mathbb{E}[z']$$

**Corollary 3.2.1.** Conditional Variance of a vector:

$$\text{Var}(z|X) = \mathbb{E}[zz'|X] - \mathbb{E}[z|X]\mathbb{E}[z'|X]$$

Thus for  $z = A(X)w$  where  $A$  is a matrix that depends on  $X$  we have:

$$\begin{aligned} \text{Var}(z|X) &= \mathbb{E}[A(X)ww'A(X)'|X] - \mathbb{E}[A(X)w|X]\mathbb{E}[w'A(X)'|X] \\ &= A(X)\mathbb{E}[ww'|X]A(X)' - A(X)\mathbb{E}[w|X]\mathbb{E}[w'|X]A(X)' \\ &= A(X)\text{Var}(w|X)A(X)' \end{aligned}$$

Therefore:

$$\text{Var}(\hat{\beta}|X) = \text{Var}(\beta + (X'X)^{-1}X'\varepsilon|X) = (X'X)^{-1}X'\text{Var}(\varepsilon|X)X(X'X)^{-1}$$

Then assuming homoskedasticity and no serial correlation:  $\text{Var}(\varepsilon|X) = \sigma^2 I_n$

$$= (X'X)^{-1}X'\sigma^2 I_n X(X'X)^{-1} = \sigma^2(X'X)^{-1}$$

### 3.3 Partitioned Regression

To find formulae for conditional variances of component of  $\hat{\beta}$  we can partition  $X$  and  $\beta$  into two parts:

$$X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$$

,  $X_1$  is  $n \times k_1$ ,  $X_2$  is  $n \times k_2$ ,  $k_1 + k_2 = k$

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

Then:  $Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$

**Theorem 3.3.1.**

$$Var(\hat{\beta}_1|X) = \sigma^2(X_1' M_2 X_1)^{-1}$$

**Proof.** Recall that

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (X'X)^{-1}X'Y$$

$$\begin{bmatrix} X_1 & X_2 \end{bmatrix}' \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} X_1 & X_2 \end{bmatrix}' Y$$

thus

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} X_1'Y \\ X_2'Y \end{bmatrix}$$

this yields two equations in two unknowns:

$$X_1'X_1\hat{\beta}_1 + X_1'X_2\hat{\beta}_2 = X_1'Y$$

$$X_2'X_1\hat{\beta}_1 + X_2'X_2\hat{\beta}_2 = X_2'Y$$

Expressing  $\hat{\beta}_1$  in terms of  $\hat{\beta}_2$  and substituting into the second equation yields:

$$\begin{aligned} (X_2'X_1)(X_1'X_1)^{-1}(X_1'Y - (X_1'X_2)\hat{\beta}_2) + (X_2'X_2)\hat{\beta}_2 &= X_2'Y \\ ((X_2'X_2) - (X_2'X_1)(X_1'X_1)^{-1}(X_1'X_2))\hat{\beta}_2 &= (X_2'Y - (X_2'X_1)(X_1'X_1)^{-1}X_1'Y) \\ X_2'(I - X_1(X_1'X_1)^{-1}X_1')X_2\hat{\beta}_2 &= X_2'(I - X_1(X_1'X_1)^{-1}X_1')Y \end{aligned}$$

Recalling the definition of the residual maker matrix,  $M_x$ , we define  $M_1$  as the residual maker matrix for  $X_1$ :

$$M_1 = I - X_1(X_1'X_1)^{-1}X_1'$$

Therefore,

$$\hat{\beta}_2 = (X_2'M_1X_2)^{-1}X_2'M_1Y$$

and similarly

$$\hat{\beta}_1 = (X_1'M_2X_1)^{-1}X_1'M_2Y$$

$$\begin{aligned} Var(\hat{\beta}_1|X) &= Var((X_1'M_2X_1)^{-1}X_1'M_2Y|X) \\ &= (X_1'M_2X_1)^{-1}X_1'M_2Var(Y|X)M_2X_1(X_1'M_2X_1)^{-1} \\ &= (X_1'M_2X_1)^{-1}X_1'M_2\sigma^2I_nM_2X_1(X_1'M_2X_1)^{-1} \\ &= \sigma^2(X_1'M_2X_1)^{-1} \end{aligned}$$

Similarly,

$$Var(\hat{\beta}_2|X) = \sigma^2(X_2'M_1X_2)^{-1}$$

If  $X_1$  and  $X_2$  are 'almost' colinear, projection of  $X_1$  onto spaces orthogonal to  $X_2$  is almost zero. Thus  $X_1' M_2 X_1$  is almost zero and so  $Var(\hat{\beta}_1|X)$  is very large. This is an example of multicollinearity. □

### 3.3.1 FRISCH-WAUGH-LOVELL THEOREM

**Theorem 3.3.2.** The OLS estimator of  $\beta_1$  in the regression of  $Y$  on  $X$  is the same as the OLS estimator of  $\beta_1$  in the regression of  $M_2 Y$  on  $M_2 X_1$ .

This is from a two step procedure:

1. Obtain  $M_2 Y$  by regressing  $Y$  on  $X_2$  and forming residuals. This is the portion of  $Y$  not correlated with  $X_2$ .

$$\hat{e} = Y - X_2(X_2' X_2)^{-1} X_2' Y = M_2 Y$$

Obtain  $M_2 X_1$  by regressing  $X_1$  on  $X_2$ . This is the portion of  $X_1$  not correlated with  $X_2$ .

$$\hat{v} = X_1 - X_2(X_2' X_2)^{-1} X_2' X_1 = M_2 X_1$$

2. Then regress  $M_2 Y$  on  $M_2 X_1$ , equivalently  $\hat{e}$  on  $\hat{v}$ . This measures the effect of  $X_1$  on  $Y$  after controlling for  $X_2$ .

**Proof.** Comparing the OLS estimators:

$$\begin{aligned} \hat{\beta}_1 &= (X_1' M_2 X_1)^{-1} X_1' M_2 Y = (X_1' M_2' M_2 X_1)^{-1} X_1' M_2' M_2 Y \\ &= [(M_2 X_1)' (M_2 X_1)]^{-1} (M_2 X_1)' M_2 Y \end{aligned}$$

Thus the OLS estimator of  $\beta_1$  in the regression of  $Y$  on  $X$  is the same as the OLS estimator of  $\beta_1$  in the regression of  $M_2 Y$  on  $M_2 X_1$ .

Then comparing regression residuals:

$$\hat{\varepsilon} = Y - X\hat{\beta} = Y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2$$

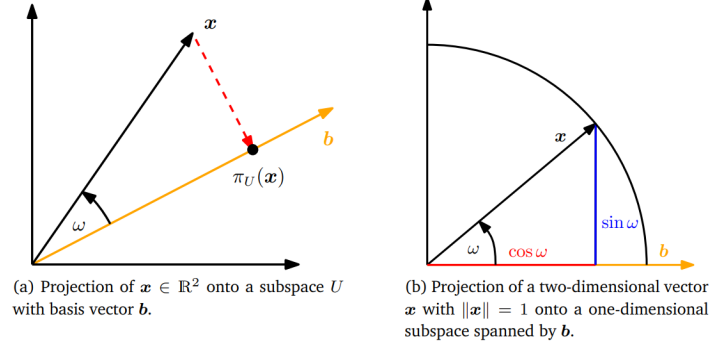
Residual from step 2 of the partitioned regression is:

$$\tilde{\varepsilon} = M_2 Y - M_2 X_1 \hat{\beta}_1 = M_2 (Y - X_1 \hat{\beta}_1) = M_2 (Y - X_1 \hat{\beta}_1 - X_2 \hat{\beta}_2) = M_2 \hat{\varepsilon} = \tilde{\varepsilon}$$

This third equality holds because  $M_2 X_2 = 0$ . Thus the residuals from the two regressions are the same and so the regression procedures are identical. □

## 3.4 Appendix: Projection Onto a Line

Assume inner product is the dot products, defined as  $x'y = \sum_{i=1}^n x_i y_i$



where  $x$  is projected onto a one-dimensional subspace  $U \subseteq \mathbb{R}^n$  spanned by basis vector  $b$ . This goes through the origin.

When projecting  $x \in \mathbb{R}^n$  onto  $U$ , we want to find the vector  $\pi_U(x) \in U$  that is closest to  $x$ .

**Proposition 3.4.1.** As before we minimise  $\|x - \pi_U(x)\|^2$ . This implies that  $x - \pi_U(x)$  is orthogonal to  $U$  and thus also orthogonal to the basis vector  $b$ .

$$\langle x - \pi_U(x), b \rangle = 0$$

**Proposition 3.4.2.** Further, the projection  $\pi_U(x)$  must be an element of  $U$  and so is a scalar multiple of  $b$ , which spans  $U$ . Hence:

$$\pi_U(x) = \lambda b$$

for some  $\lambda \in \mathbb{R}$

### 3.4.1 Finding $\lambda$

Substituting Prop 1.4.2 into 1.4.1 we get:

$$\langle x - \lambda b, b \rangle = 0$$

Exploiting the bilinearity of the inner product:

$$\begin{aligned} \langle x, b \rangle - \lambda \langle b, b \rangle &= 0 \\ \Rightarrow \lambda &= \frac{\langle x, b \rangle}{\langle b, b \rangle} = \frac{\langle x, b \rangle}{\|b\|^2} = \frac{x'b}{b'b} \end{aligned}$$

### 3.4.2 Finding $\pi_U(x)$

Since  $\pi_U(x) = \lambda b$ , we have:

$$\pi_U(x) = \frac{x'b}{b'b} b$$

The length of  $\pi_U(x)$  is:

$$\|\pi_U(x)\| = \|\lambda b\| = |\lambda| \|b\|$$

Thus the projection acts as a coordinate of  $\pi_U(x)$  in the direction of  $b$ .

Using the dot product as the inner product we have:

$$= \frac{|x'b|}{\|b\|^2} \|b\| = |\cos(\theta)| \|x\| \|b\| \frac{\|b\|}{\|b\|^2} = |\cos(\theta)| \|x\|$$

### 3.4.3 The Projection Matrix $P_\pi$

As projection is a linear mapping, there exists a matrix  $P_\pi$  such that:

$$\pi_U(x) = P_\pi x$$

With the dot as the inner product and

$$\pi_U(x) = \lambda b = b\lambda = b \frac{b'x}{||b||^2} = \frac{bb'}{||b||^2} x$$

Thus

$$P_\pi = \frac{bb'}{||b||^2}$$

## 3.5 Projection Onto a General Subspace

We find a projection of  $x \in \mathbb{R}^n$  onto a subspace  $U \subseteq \mathbb{R}^n$  with  $\dim(U) = m \geq 1$ . Assume that  $b_1, \dots, b_m$  is an ordered basis for  $U$ . Any projection  $\pi_U(x)$  onto  $U$  can be written as a linear combination of the basis vectors: such that  $\pi_U(x) = \sum_{i=1}^m \lambda_i b_i$ . We follow the same three step procedure as before:

### 3.5.1 Finding $\lambda_1, \dots, \lambda_m$

We find coordinates  $\lambda_1, \dots, \lambda_m$  such that the linear combination

$$\pi_U(x) = \sum_{i=1}^m \lambda_i b_i = \mathbf{B} \vec{\lambda}$$

$$\mathbf{B} = \begin{bmatrix} \vec{b}_1 & \dots & \vec{b}_m \end{bmatrix}, \in \mathbb{R}^{n \times m}, \vec{\lambda} = \begin{bmatrix} \lambda_1 \\ \dots \\ \lambda_m \end{bmatrix} \in \mathbb{R}^m$$

is such that  $\pi_U(x)$  is the closest point in  $U$  to  $x$ . This implies that  $x - \pi_U(x)$  is orthogonal to  $U$  and thus also orthogonal to each basis vector  $b_i$ . Thus we obtain simultaneous equations:

$$\langle x - \pi_U(x), b_1 \rangle = b'_1(x - \pi_U(x)) = 0$$

$$\vdots$$

$$\langle x - \pi_U(x), b_m \rangle = b'_m(x - \pi_U(x)) = 0$$

as  $\pi_U(x) = \mathbf{B} \vec{\lambda}$  we have:

$$b'_1(x - \mathbf{B} \vec{\lambda}) = 0$$

$$\vdots$$

$$b'_m(x - \mathbf{B} \vec{\lambda}) = 0$$

thus we obtain a homogeneous system of linear equations:

$$\begin{bmatrix} b'_1 \\ \vdots \\ b'_m \end{bmatrix} (x - \mathbf{B} \vec{\lambda}) = 0$$

$$\Leftrightarrow \mathbf{B}'(x - \mathbf{B}\vec{\lambda}) = 0$$

$$\Leftrightarrow \mathbf{B}'\mathbf{B}\vec{\lambda} = \mathbf{B}'x$$

$$\Leftrightarrow \vec{\lambda} = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'x$$

where we require that  $\mathbf{B}'\mathbf{B}$  is invertible, which is true if and only if  $\mathbf{B}$  has full column rank, which is true if and only if the basis vectors  $b_1, \dots, b_m$  are linearly independent.

### 3.5.2 Finding $\pi_U(x)$

We have that  $\pi_U(x) = \mathbf{B}\vec{\lambda}$  and so:

$$\pi_U(x) = \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'x$$

### 3.5.3 The Projection Matrix $P_\pi$

As projection is a linear mapping, there exists a matrix  $P_\pi$  such that:

$$\pi_U(x) = P_\pi x$$

Thus

$$P_\pi = \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$$

## 3.6 Appendix: OLS Estimator Equivalence

### Claim 3.6.1.

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'Y = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \\ &\Leftrightarrow X\end{aligned}$$

includes a constant

Let us take the case for  $k = 1$ , i.e.  $X$  is a vector of length  $n$ . Then: suppose  $X$  includes a constant,

i.e.  $X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$ . Then let  $\tilde{x}_i = (1, x_i)'$  Then  $X = (\tilde{x}_1, \dots, \tilde{x}_n)'$  Thus:

$$\begin{aligned}(X'X)^{-1}X'Y &= \left(\sum_{i=1}^n \tilde{x}_i \tilde{x}_i'\right)^{-1} \sum_{i=1}^n \tilde{x}_i y_i \\ &= \left[\sum_{i=1}^n \begin{bmatrix} 1 & x_i \\ x_i & x_i^2 \end{bmatrix}\right]^{-1} \sum_{i=1}^n \begin{bmatrix} 1 \\ x_i \end{bmatrix} Y_i \\ &= \left[n \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \frac{1}{n} \sum_{i=1}^n x_i^2 \end{bmatrix}\right]^{-1} \begin{bmatrix} \bar{y} \\ \frac{1}{n} \sum_{i=1}^n x_i y_i \end{bmatrix} \\ &= \frac{1}{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2} \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \begin{bmatrix} \bar{y} \\ \frac{1}{n} \sum_{i=1}^n x_i y_i \end{bmatrix}\end{aligned}$$



The second component is the estimate for the slope coefficient, and the first component is the estimate of the intercept coefficient. Thus we have:

$$\hat{\beta} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$