Econometrics II

OLS in Matrix Notation

Lasha Chochua

2026

Dataset

- Dataset components:
  - **Random variables**: (Y, X).
  - **Observations**: (Y_i, X_i) for $i = 1, \ldots, n$.
  - **Sample size**: n.

Assumptions

- Homogeneity of data:
  - Observations drawn from an identical population F.
  - **Assumption 1**: Variables $\{(Y_1, X_1), \ldots, (Y_n, X_n)\}$ are identically distributed.
- View data as:
  - **Population**: Underlying common distribution.
  - **Sample**: Observed realizations.

Population vs. Sample

- Distinction:
  - **Population**: Theoretical, common distribution F.
  - **Sample or Dataset**: Observed data $\{(Y_i, X_i)\}$.
- Modeling:
  - Treat datasets as random draws from F.
  - Example: Tax records modeled as observations from an infinitely large population.

The Linear Model

- Model equation:

$$Y = X'\beta + e \tag{1}$$

- Components:
  - Y: Dependent variable.
  - X: Independent variables (including a constant) – A $k \times 1$-dimensional vector.
  - β: Coefficients – A $k \times 1$-dimensional vector.
  - e: Error term.

Linear Model

- Objective:
  - Minimize the expected squared error.
- Formula:

$$\beta = \arg\min_{b \in \mathbb{R}^k} S(b) \tag{2}$$

where:

$$S(b) = \mathbb{E}\left[(Y - X'b)^2\right] \tag{3}$$

Moment Estimators

- Estimating β involves population expectations:
  - Rewrite β as a function of sample moments.
  - Use estimators for these moments.
- Population mean:

$$\mu = \mathbb{E}[Y] = \int_{-\infty}^{\infty} y dF(y) \tag{4}$$

  - Sample mean:

$$\hat{\mu} = \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{5}$$

- $\bar{Y}$ is a natural estimator of μ.

General Moment Estimators

- For a vector of functions $h(Y)$:

$$\mu = \mathbb{E}[h(Y)], \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} h(Y_i)$$

- Example:
  - First moment: $h(y) = y$.
  - Second moment: $h(y) = y^2$.

Variance as a Moment Estimator

- Variance formula:

$$\sigma^2 = \text{var}[Y] = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2$$

- Estimator for variance:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)^2$$

- $\hat{\sigma}^2$ is a plug-in estimator derived from sample moments.

Least Squares Estimator

- Linear model coefficient β minimizes the expected squared error $S(\beta)$:

$$S(\beta) = \mathbb{E}\left[(Y - X'\beta)^2\right]$$

- Moment estimator for $S(\beta)$ (sample average):

$$\hat{S}(\beta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\beta)^2 = \frac{1}{n}\mathsf{SSE}(\beta) \tag{6}$$

- Sum of Squared Errors (SSE) function:

$$\mathsf{SSE}(\beta) = \sum_{i=1}^{n}(Y_i - X_i'\beta)^2$$

Definition of the Least Squares Estimator

- The least squares estimator is defined as:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^k}{\arg\min}\, \hat{S}(\beta)$$

- Equivalently, $\hat{\beta}$ minimizes $\mathsf{SSE}(\beta)$:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^k}{\arg\min}\, \mathsf{SSE}(\beta)$$

Properties of the Least Squares Estimator

- $\widehat{\beta}$ is also known as the **ordinary least squares (OLS)** estimator.
- The hat "$\widehat{\ }$" indicates that $\widehat{\beta}$ is an estimator for β.
- Notations:
  - $\widehat{\beta}_n$: Specifies dependence on sample size n.
  - $\widehat{\beta}_{\text{ols}}$: Refers explicitly to OLS estimation.
- **Distinction**:
  - β: Population parameter (non-random, fixed).
  - $\widehat{\beta}$: Sample estimator (random, varies with sample).

Solving for Least Squares with One Regressor

- Consider the case $k = 1$ (one scalar regressor X and scalar coefficient β).
- Sum of Squared Errors (SSE):

$$\text{SSE}(\beta) = \sum_{i=1}^{n} (Y_i - X_i\beta)^2 = \left(\sum_{i=1}^{n} Y_i^2\right) - 2\beta\left(\sum_{i=1}^{n} X_i Y_i\right) + \beta^2\left(\sum_{i=1}^{n} X_i^2\right)$$

- Quadratic function of β:
  - The OLS estimator $\hat{\beta}$ minimizes $\text{SSE}(\beta)$.

Minimizing SSE

- From elementary algebra:
  - For a quadratic function $a - 2bx + cx^2$, the minimizer is $x = b/c$.
  - For SSE(β), the minimizer is:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \tag{7}$$

- **Special case:** Intercept-only model ($X_i = 1$):

$$\hat{\beta} = \frac{\sum_{i=1}^{n} 1 \cdot Y_i}{\sum_{i=1}^{n} 1^2} = \frac{1}{n} \sum_{i=1}^{n} Y_i = \bar{Y} \tag{8}$$

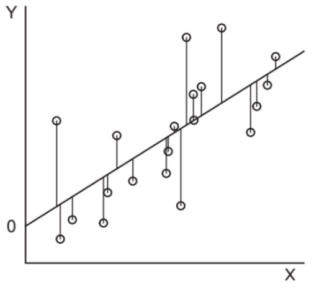
Visual Representation - Regression With One Regressor

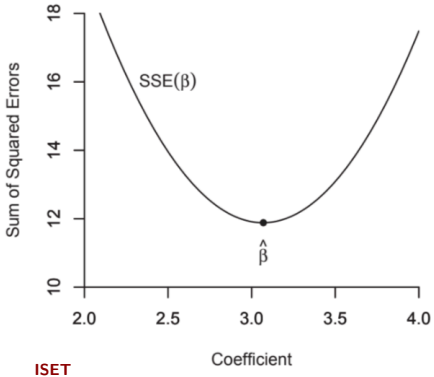
1. Deviation from Fitted Line:

- Scatterplot of 20 pairs (Y_i, X_i).
- Vertical lines represent errors $Y_i - X_i\beta$.

2. Sum of Squared Errors Function:

- SSE(β) is a quadratic function.
- $\hat{\beta}$ minimizes SSE(β).





Solving for Least Squares with Multiple Regressors

- Consider $k > 1$ where $\beta \in \mathbb{R}^k$ (a vector of coefficients).
- Regression function:

$$x'\beta = x_1\beta_1 + x_2\beta_2$$

  - Represents a 2D surface in 3D space.
- Sum of Squared Errors (SSE):

$$\text{SSE}(\beta) = \sum_{i=1}^{n} Y_i^2 - 2\beta' \sum_{i=1}^{n} X_i Y_i + \beta' \sum_{i=1}^{n} X_i X_i' \beta.$$

Why $2\beta' \sum_{i=1}^{n} X_i Y_i$?

- The middle term from expanding the square is:

$$-2Y_i X_i' \beta$$

- $Y_i X_i' \beta$ is a scalar, so transpose it:

$$Y_i X_i' \beta = (Y_i X_i' \beta)' = \beta' X_i Y_i$$

- So both writings are the same:

$$Y_i X_i' \beta = \beta' X_i Y_i$$

- Sum over i and collect the factor of 2:

$$2 \sum_{i=1}^{n} Y_i X_i' \beta = 2\beta' \sum_{i=1}^{n} X_i Y_i$$

Why $(X_i^{'}\beta)^2 = \beta^{'} X_i X_i^{'} \beta$?

- $X_i'\beta$ is a scalar, so squaring it means multiplying it by itself:

$$(X_i'\beta)^2 = (X_i'\beta)(X_i'\beta)$$

- Transpose the first copy. For any scalar a, we have $a = a'$, so:

$$(X_i'\beta)(X_i'\beta) = (X_i'\beta)'(X_i'\beta)$$

- Use the transpose rule $(AB)' = B'A'$:

$$(X_i'\beta)' = \beta' X_i$$

- Substitute back:

$$(X_i'\beta)(X_i'\beta) = \beta' X_i \cdot X_i'\beta = \beta' X_i X_i'\beta$$

Differentiation Rules for Vectors

- Three rules needed. Let $\beta \in \mathbb{R}^k$, $c \in \mathbb{R}^k$, M be $k \times k$ symmetric.
- **Rule 1**: Constant term – disappears:

$$\frac{\partial}{\partial \beta} \sum_i Y_i^2 = 0$$

- **Rule 2**: Linear term – analog of $\frac{d}{d\beta}(c\beta) = c$:

$$\frac{\partial}{\partial \beta} c'\beta = c$$

- **Rule 3**: Quadratic term – analog of $\frac{d}{d\beta}(m\beta^2) = 2m\beta$:

$$\frac{\partial}{\partial \beta} \beta' M \beta = 2M\beta$$

What Does $\frac{\partial}{\partial \beta}$ Mean?

- $\beta \in \mathbb{R}^k$ is a vector, so we cannot take "one derivative" – we take k of them:

$$\frac{\partial}{\partial \beta} f(\beta) \equiv \begin{pmatrix} \frac{\partial f}{\partial \beta_1} \\ \frac{\partial f}{\partial \beta_2} \\ \vdots \\ \frac{\partial f}{\partial \beta_k} \end{pmatrix}$$

- This is just notation: stack all partial derivatives into a $k \times 1$ vector.
- The three rules follow immediately from scalar calculus:

$f(\beta)$	$\frac{\partial f}{\partial \beta_j}$	$\frac{\partial f}{\partial \beta}$
$\sum_i Y_i^2$	0	0
$c'\beta$	c_j	c
$\beta' M \beta$	$2\sum_l M_{jl}\beta_l$	$2M\beta$

First-Order Conditions

- Sum of Squared Errors (SSE):

$$\text{SSE}(\beta) = \sum_{i=1}^{n} Y_i^2 - 2\beta' \sum_{i=1}^{n} X_i Y_i + \beta' \sum_{i=1}^{n} X_i X_i' \beta.$$

- To minimize $\text{SSE}(\beta)$, solve first-order conditions:

$$\frac{\partial}{\partial \beta} \text{SSE}(\beta) = -2 \sum_{i=1}^{n} X_i Y_i + 2 \sum_{i=1}^{n} X_i X_i' \beta = 0. \tag{9}$$

- Solution:

$$\sum_{i=1}^{n} X_i X_i' \widehat{\beta} = \sum_{i=1}^{n} X_i Y_i. \tag{10}$$

Explicit Formula for $\widehat{\beta}$

- Using matrix algebra:

$$\widehat{\beta} = \left(\sum_{i=1}^{n} X_i X_i' \right)^{-1} \left(\sum_{i=1}^{n} X_i Y_i \right). \tag{11}$$

- $\widehat{\beta}$ is the natural estimator of the linear projection coefficient.

Verification of Minimization

- Second-order condition:

$$\frac{\partial^2}{\partial\beta\partial\beta'}\mathsf{SSE}(\beta) = 2\sum_{i=1}^{n} X_i X_i'.$$

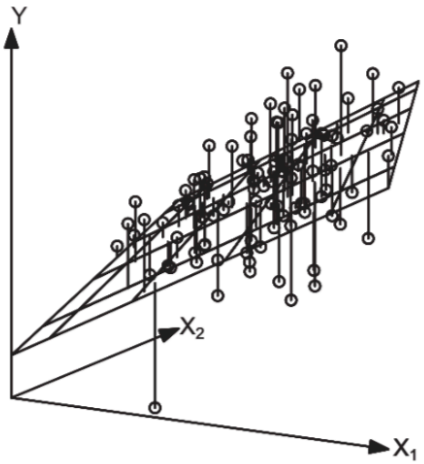
  - Positive semi-definite matrix ensures $\hat{\beta}$ minimizes $\mathsf{SSE}(\beta)$.

- $\hat{\beta}$ minimizes $\mathsf{SSE}(\beta)$ and corresponds to the low spot in contour plots.

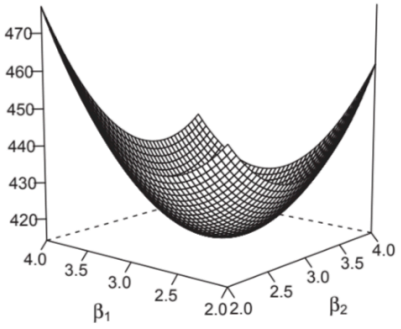
Visual Representation

3D Regression Surface:

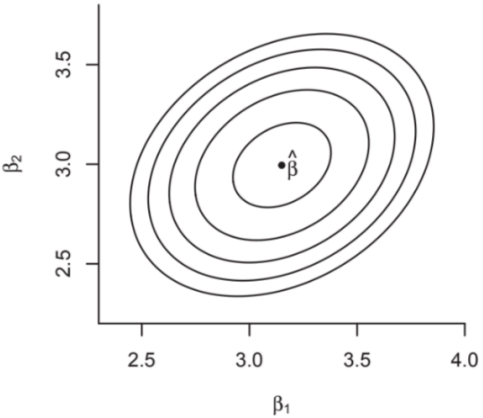
- Vertical lines represent residuals $e_i = Y_i - X_i'\beta$.
- Surface represents the fitted regression plane.



SSE with Two Regressors



(a) Sum of Squared Error Function



(b) SSE Contour

Moment Estimators

- $\hat{\beta}$ can also be expressed using sample moments:

$$\hat{\mathbf{Q}}_{XY} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i, \quad \hat{\mathbf{Q}}_{XX} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i'.$$

- Substituting:

$$\hat{\beta} = \hat{\mathbf{Q}}_{XX}^{-1}\hat{\mathbf{Q}}_{XY}.$$

Uniqueness of $\widehat{\beta}$

- $\widehat{\beta}$ is unique if $\sum_{i=1}^{n} X_i X_i'$ is invertible.

Theorem 1

$$\widehat{\beta} = \left(\sum_{i=1}^{n} X_i X_i' \right)^{-1} \left(\sum_{i=1}^{n} X_i Y_i \right),$$

provided $\sum_{i=1}^{n} X_i X_i' > 0$.

Illustration: Using the CPS Data

- Dataset: March 2009 Current Population Survey (CPS).
  - Focus: Sub-sample of married Black female wage earners with 12 years of potential work experience.
  - Sample size: 20 observations.
- Variables:
  1. **wage**: Total annual earnings divided by hours worked.
  2. **log(wage)**: Natural logarithm of the wage.
  3. **education**: Years of education.
  4. **education^2**: Square of education.
  5. **log(wage) $\times$ education**: Product of log(wage) and education.

Illustration: Using the CPS Data

Observation	wage	log(wage)	education	education^2	education×log(wage)
1	37.93	3.64	18	324	65.44
2	40.87	3.71	18	324	66.79
3	14.18	2.65	13	169	34.48
4	16.83	2.82	16	256	45.17
5	33.17	3.50	16	256	56.03
6	29.81	3.39	18	324	61.11
7	54.62	4.00	16	256	64.00
8	43.08	3.76	18	324	67.73
9	14.42	2.67	12	144	32.03
10	14.90	2.70	16	256	43.23
11	21.63	3.07	18	324	55.44
12	11.09	2.41	16	256	38.50
13	10.00	2.30	13	169	29.93
14	31.73	3.46	14	196	48.40
15	11.06	2.40	12	144	28.84
16	18.75	2.93	16	256	46.90
17	27.35	3.31	14	196	46.32
18	24.04	3.18	16	256	50.76
19	36.06	3.59	18	324	64.53
20	23.08	3.14	16	256	50.22
Sum	515	62.64	314	5010	995.86

Data Summary

- Let Y_i be log(wage) and X_i be Education and an Intercept. Then:

Data Summary

- Let Y_i be log(wage) and X_i be Education and an Intercept. Then:

$$\sum_{i=1}^{n} X_i Y_i = \begin{pmatrix} 995.86 \\ 62.64 \end{pmatrix}, \quad \sum_{i=1}^{n} X_i X_i' = \begin{pmatrix} 5010 & 314 \\ 314 & 20 \end{pmatrix}.$$

Data Summary

- Let Y_i be log(wage) and X_i be Education and an Intercept. Then:

$$\sum_{i=1}^{n} X_i Y_i = \begin{pmatrix} 995.86 \\ 62.64 \end{pmatrix}, \quad \sum_{i=1}^{n} X_i X_i' = \begin{pmatrix} 5010 & 314 \\ 314 & 20 \end{pmatrix}.$$

- Matrix inversion:

$$\left(\sum_{i=1}^{n} X_i X_i' \right)^{-1} = \begin{pmatrix} 0.0125 & -0.196 \\ -0.196 & 3.124 \end{pmatrix}.$$

Regression Results

- Solving:

$$\widehat{\beta} = \begin{pmatrix} 0.0125 & -0.196 \\ -0.196 & 3.124 \end{pmatrix} \begin{pmatrix} 995.86 \\ 62.64 \end{pmatrix} = \begin{pmatrix} 0.155 \\ 0.698 \end{pmatrix}.$$

- Regression equation:

$$\log(\text{wage}) = 0.155 \cdot \text{education} + 0.698. \tag{12}$$

Interpretation and Prediction

- Interpretation:
  - Each additional year of education is associated with a **16% increase in mean wages**.
- Predictions:
  - **12 years of education**:

$$\log(\text{wage}) = 0.155 \cdot 12 + 0.698 = 2.56.$$

  - **16 years of education**:

$$\log(\text{wage}) = 0.155 \cdot 16 + 0.698 = 3.18.$$

Multiple Regression Example

- Including experience and its square:

$$\log(\text{wage}) = 0.143 \cdot \text{education} + 0.036 \cdot \text{experience} - 0.071 \cdot \frac{\text{experience}^2}{100} + 0.575. \quad (13)$$

- Interpretation:
  - **14% increase in mean wages per year of education**, holding experience constant.

Least Squares Residuals

Definition

Fitted value:

$$\hat{Y}_i = X_i' \hat{\beta}.$$

Residual:

$$\hat{e}_i = Y_i - \hat{Y}_i = Y_i - X_i' \hat{\beta}. \tag{14}$$

- **Key Points**:
  - $\hat{Y}_i$ is often called the "predicted value" but is more accurately described as the "fitted value."

$$Y_i = \hat{Y}_i + \hat{e}_i. \tag{15}$$

- **Distinction**:
  - **Error** (e_i): Unobservable.
  - **Residual** ($\hat{e}_i$): Estimator of the error.

Properties of Residuals

- Residual orthogonality:

$$\sum_{i=1}^{n} X_i \hat{e}_i = 0. \tag{16}$$

- **Explanation**:
  - The residuals $\hat{e}_i$ are uncorrelated with the regressors X_i.
  - This property arises because $\hat{\beta}$ minimizes the sum of squared errors.
  - Ensures no systematic relationship between residuals and regressors, validating the regression model fit.

Direct Calculation:

Using equations (14) and (11):

$$\sum_{i=1}^{n} X_i \hat{e}_i = \sum_{i=1}^{n} X_i (Y_i - X_i' \hat{\beta}) = \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i X_i' \hat{\beta}$$

$$= \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i X_i' \left(\sum_{i=1}^{n} X_i X_i' \right)^{-1} \left(\sum_{i=1}^{n} X_i Y_i \right)$$

$$= \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i Y_i = 0.$$

- When X_i contains a constant, an implication is:

$$\frac{1}{n} \sum_{i=1}^{n} \hat{e}_i = 0. \tag{17}$$

Insights

- **Sample Properties**:
  - Residuals have a sample mean of zero.
  - Sample correlation between regressors and residuals is zero.
- **Applicability**:
  - These properties are algebraic results.
  - They hold for all linear regression estimates.

Demeaned Regressors

- Sometimes it is useful to separate the constant from the other regressors and write the linear projection equation in the format:

$$Y_i = X_i'\beta + a + e_i,$$

where a is the intercept and X_i does not contain a constant.

- The least squares estimates and residuals can be written as:

$$Y_i = X_i'\hat{\beta} + \hat{a} + \hat{e}_i.$$

- In this case, the system becomes:

$$\sum_{i=1}^{n}(Y_i - X_i'\hat{\beta} - \hat{a}) = 0,$$

$$\sum_{i=1}^{n} X_i(Y_i - X_i'\hat{\beta} - \hat{a}) = 0.$$

Solving for $\hat{a}$ and $\hat{\beta}$

- From the first equation:
$$\hat{a} = \bar{Y} - \bar{X}'\hat{\beta}.$$
- Subtracting from the second equation:
$$\sum_{i=1}^{n} X_i \left[(Y_i - \bar{Y}) - (X_i - \bar{X})'\hat{\beta} \right] = 0.$$
- Solving for $\hat{\beta}$:
$$\hat{\beta} = \left(\sum_{i=1}^{n} X_i (X_i - \bar{X})' \right)^{-1} \left(\sum_{i=1}^{n} X_i (Y_i - \bar{Y}) \right).$$
- Simplified as:
$$\hat{\beta} = \left(\sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})' \right)^{-1} \left(\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \right). \quad (18)$$

Demeaned Regressors - Conclusion

- The OLS estimator for the slope coefficients is OLS with demeaned data and no intercept.
- Representation (18) is known as the demeaned formula for the least squares estimator.

Model in Matrix Notation

- For computational purposes, it is convenient to write the model in matrix notation.
- The n linear equations:

$$Y_i = X_i'\beta + e_i, \quad i = 1, 2, \ldots, n,$$

can be stacked as:

$$\mathbf{Y} = \underset{n\times 1}{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}, \quad \mathbf{X} = \underset{n\times k}{\begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix}}, \quad \beta_{k\times 1}, \quad \mathbf{e} = \underset{n\times 1}{\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}}.$$

- Compact single equation:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}. \tag{19}$$

Matrix Expressions

- Sample sums:

$$\sum_{i=1}^{n} X_i X_i' = \mathbf{X}'\mathbf{X}, \quad \sum_{i=1}^{n} X_i Y_i = \mathbf{X}'\mathbf{Y}.$$

- Least squares estimator:

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

- Residual vector:

$$\hat{\mathbf{e}} = \mathbf{Y} - \mathbf{X}\hat{\beta}.$$

- Orthogonality property:

$$\mathbf{X}'\hat{\mathbf{e}} = 0.$$

Important Matrix Expressions

Theorem 2 The least squares estimation framework includes:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}),$$

$$\hat{\mathbf{e}} = \mathbf{Y} - \mathbf{X}\hat{\beta},$$

$$\mathbf{X}'\hat{\mathbf{e}} = 0.$$

Matrix Notation: A Simple Example

- Suppose $n = 3$, two regressors X_{1i}, X_{2i} plus a constant:

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 + e_i, \quad i = 1, 2, 3$$

Matrix Notation: A Simple Example

- Suppose $n = 3$, two regressors X_{1i}, X_{2i} plus a constant:

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 + e_i, \quad i = 1, 2, 3$$

- Write out all three equations:

$$Y_1 = \beta_1 X_{11} + \beta_2 X_{21} + \beta_3 + e_1$$

$$Y_2 = \beta_1 X_{12} + \beta_2 X_{22} + \beta_3 + e_2$$

$$Y_3 = \beta_1 X_{13} + \beta_2 X_{23} + \beta_3 + e_3$$

Matrix Notation: A Simple Example (Cont.)

- Stack into matrices:

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}}_{\mathbf{Y}} = \underbrace{\begin{pmatrix} X_{11} & X_{21} & 1 \\ X_{12} & X_{22} & 1 \\ X_{13} & X_{23} & 1 \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}}_{\mathbf{e}}$$

Matrix Notation: A Simple Example (Cont.)

- Stack into matrices:

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}}_{\mathbf{Y}} = \underbrace{\begin{pmatrix} X_{11} & X_{21} & 1 \\ X_{12} & X_{22} & 1 \\ X_{13} & X_{23} & 1 \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}}_{\mathbf{e}}$$

- Compact notation: $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$.

Reading

- **Wooldridge, Introductory Econometrics.**
  - Appendix: E-1 (pp. 760 - 762)