

450A Political Methodology I

Lecture 6: Multiple Regression

1. Mechanics

Stanford University

Preparation

Matrix Representation of Regression

Solving Linear Systems using Matrices

The OLS Estimator in Matrix Algebra

Frisch–Waugh–Lovell Theorem

Simple regression in Vector Form

If we define the following column vectors:

- $\mathbf{1}$ denotes a vector with n ones
- $\mathbf{y}$ denotes an n element data vector for y
- $\mathbf{x}$ denotes an n element data vector for x
- $\hat{\mathbf{u}}$ denotes an n element vector of residuals
- $\hat{\mathbf{y}}$ denotes an n element vector of fitted values

Then we can write the fitted regression line for the data in the following manner.

$$\mathbf{y} = \hat{\beta}_0 \mathbf{1} + \hat{\beta}_1 \mathbf{x} + \hat{\mathbf{u}}$$

and ...

$$\hat{\mathbf{y}} = \hat{\beta}_0 \mathbf{1} + \hat{\beta}_1 \mathbf{x}$$

Example: Simple Regression with Three Observations

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} ; \mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} ; \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}$$

$$\mathbf{y} = \hat{\beta}_0 \mathbf{1} + \hat{\beta}_1 \mathbf{x} + \hat{\mathbf{u}}$$

$$\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \hat{\beta}_0 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \hat{\beta}_1 \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix} + \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \end{bmatrix}$$

Multiple Regression in Matrix Form

Let $\mathbf{X}$ be our $n \times (k + 1)$ data matrix for k predictors (plus the constant),

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

Let $\hat{\beta}$ be the vector $(k + 1) \times 1$ of least squares estimates (intercept and k slopes): $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$

Then $\mathbf{X}\hat{\beta}$ is the $n \times 1$ vector of fitted values:

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \mathbf{X}\hat{\beta} = \begin{bmatrix} 1\hat{\beta}_0 + x_{11}\hat{\beta}_1 + x_{12}\hat{\beta}_2 + \cdots + x_{1k}\hat{\beta}_k \\ 1\hat{\beta}_0 + x_{21}\hat{\beta}_1 + x_{22}\hat{\beta}_2 + \cdots + x_{2k}\hat{\beta}_k \\ \vdots \\ 1\hat{\beta}_0 + x_{n1}\hat{\beta}_1 + x_{n2}\hat{\beta}_2 + \cdots + x_{nk}\hat{\beta}_k \end{bmatrix}$$

Explanatory Data Matrix

The $n \times (k + 1)$ matrix **X** is often called the **design matrix**. Example:

R Code

```
> fit <- lm(vote1 ~ fem + educ + age, data = d)
> X <- model.matrix(fit)
> dim(X)
[1] 1703    4
> head(X, 8)
      (Intercept) fem educ age
1             1    0    1  65
2             1    0    3  29
3             1    1    1  38
4             1    1    1  49
5             1    1    2  23
6             1    1    1  28
7             1    0    3  26
8             1    1    2  24
```

In R, the design matrix can be extracted from a fitted `lm` object by the `model.matrix` function.

Source: national survey by FLACSO/Chile, April–May 1988, before the plebiscite (`carData::Chile`); we keep the 1,703 of 2,700 respondents who intended to vote Yes or No and answered every question. `vote1` = 1 if the respondent intends to vote Yes (for Pinochet), 0 if No; `fem` = 1 if female; `educ`: 1 = primary, 2 = secondary, 3 = post-secondary.

Solving Systems of Equations Using Matrices

Matrices are very useful to solve linear systems of equations, such as the first order conditions for our least squares estimates.

Here is an example with three equations and three unknowns:

$$x + 2y + z = 3$$

$$3x - y - 3z = -1$$

$$2x + 3y + z = 4$$

How would one go about solving this? There are various techniques, including substitution, and multiplying equations by constants and adding them to get single variables to cancel.

Solving Systems of Equations Using Matrices

An easier way is to use matrix algebra. Note that the system of equations

$$x + 2y + z = 3$$

$$3x - y - 3z = -1$$

$$2x + 3y + z = 4$$

can be written as follows:

$$\begin{bmatrix} 1 & 2 & 1 \\ 3 & -1 & -3 \\ 2 & 3 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 3 \\ -1 \\ 4 \end{bmatrix} \iff \mathbf{Au} = \mathbf{b}$$

How do we solve this for $\mathbf{u} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$? Let's look at the scalar case first.

Solving Equations with Inverses (scalar case)

Let's go back to the scalar world of 8th grade algebra. How would you solve the following for u ?

$$au = b$$

We multiply both sides of by the reciprocal $1/a$ (the **inverse of a**) and get:

$$\begin{aligned}\frac{1}{a} a u &= \frac{1}{a} b \\ u &= \frac{b}{a}\end{aligned}$$

(Note that this technique only works if $a \neq 0$. If $a = 0$, then there are either an infinite number of solutions for u (when $b = 0$), or no solutions for u (when $b \neq 0$).)

So to solve our multiple equation problem in the matrix case we need a matrix equivalent of the inverse. This equivalent is the **inverse matrix**. The inverse of $\mathbf{A}$ is written as $\mathbf{A}^{-1}$.

Inverse of a Matrix

The **inverse** $\mathbf{A}^{-1}$ of $\mathbf{A}$ has the property that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix.

- The inverse $\mathbf{A}^{-1}$ exists only if $\mathbf{A}$ is invertible or **nonsingular** (more on this soon)
- The inverse is unique if it exists and then the linear system has a unique solution.
- There are various methods for finding/computing the inverse of a matrix

The **inverse matrix** allows us to solve linear systems of equations.

$$\mathbf{A}\mathbf{u} = \mathbf{b}$$

$$\mathbf{A}^{-1}\mathbf{A}\mathbf{u} = \mathbf{A}^{-1}\mathbf{b}$$

$$\mathbf{I}\mathbf{u} = \mathbf{A}^{-1}\mathbf{b}$$

$$\mathbf{u} = \mathbf{A}^{-1}\mathbf{b}$$

Given $\mathbf{A}$ we find that $\mathbf{A}^{-1}$ is:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 1 \\ 3 & -1 & -3 \\ 2 & 3 & 1 \end{bmatrix}; \mathbf{A}^{-1} = \begin{bmatrix} 8 & 1 & -5 \\ -9 & -1 & 6 \\ 11 & 1 & -7 \end{bmatrix}$$

We can now solve our system of equations:

$$\mathbf{u} = \mathbf{A}^{-1}\mathbf{b} = \begin{bmatrix} 8 & 1 & -5 \\ -9 & -1 & 6 \\ 11 & 1 & -7 \end{bmatrix} \begin{bmatrix} 3 \\ -1 \\ 4 \end{bmatrix} = \begin{bmatrix} 3 \\ -2 \\ 4 \end{bmatrix}$$

So the solution vector is $x = 3$, $y = -2$, and $z = 4$. Verifying:

$$x + 2y + z = 3 + 2 \cdot -2 + 4 = 3$$

$$3x - y - 3z = 3 \cdot 3 - -2 - 3 \cdot 4 = -1$$

$$2x + 3y + z = 2 \cdot 3 + 3 \cdot -2 + 4 = 4$$

Computationally, this method is very convenient. We “just” compute the inverse, and perform a single matrix multiplication.

Singularity of a Matrix

If the inverse of $\mathbf{A}$ exists, then the linear system has a unique solution. If it exists, we say that $\mathbf{A}$ is **nonsingular** or **invertible** (these statements are equivalent).

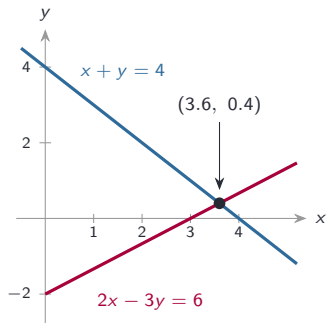
$\mathbf{A}$ must be square to be invertible, but not all square matrices are invertible. More precisely, a square matrix $\mathbf{A}$ is invertible iff its column vectors (or equivalently its row vectors) are **linearly independent**.

The column **rank** of a matrix $\mathbf{A}$ is the largest number of linearly independent columns of $\mathbf{A}$. If the rank of $\mathbf{A}$ equals the number of columns of $\mathbf{A}$, then we say that $\mathbf{A}$ has **full column rank**. This implies that all its column vectors are linearly independent.

If a column of $\mathbf{A}$ is a linear combination of the other columns, there are either **no solutions** to the system of equations or infinitely **many solutions** to the system of equations. With no solution, the system is called **inconsistent**; with infinitely many, the equations do not pin down a single solution.

Geometric Example in 2D

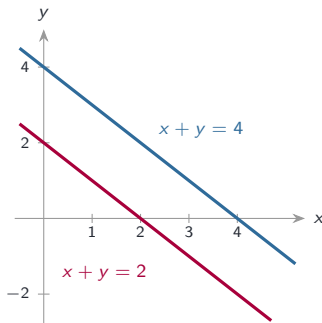
One solution



$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 2 & -3 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 4 \\ 6 \end{bmatrix}$$

$\mathbf{A}$ invertible

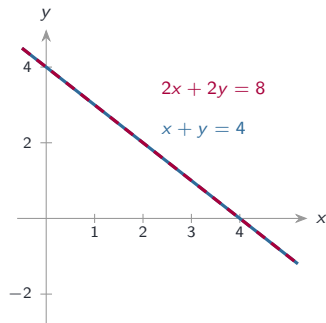
No solution



$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$$

$\mathbf{A}$ singular

Infinitely many solutions



$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 4 \\ 8 \end{bmatrix}$$

$\mathbf{A}$ singular

Why do we care about invertibility?

We have seen that OLS regression is defined by a system of linear equations

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \mathbf{X}\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1\hat{\beta}_0 + x_{11}\hat{\beta}_1 + x_{12}\hat{\beta}_2 + \cdots + x_{1k}\hat{\beta}_k \\ 1\hat{\beta}_0 + x_{21}\hat{\beta}_1 + x_{22}\hat{\beta}_2 + \cdots + x_{2k}\hat{\beta}_k \\ \vdots \\ 1\hat{\beta}_0 + x_{n1}\hat{\beta}_1 + x_{n2}\hat{\beta}_2 + \cdots + x_{nk}\hat{\beta}_k \end{bmatrix}$$

with our data matrix

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

We have also learned that $\hat{\boldsymbol{\beta}}$ is obtained by solving **normal equations**, a linear system of equations. It turns out that to solve for $\hat{\boldsymbol{\beta}}$, we need to invert $\mathbf{X}'\mathbf{X}$, a $(k+1) \times (k+1)$ matrix.

Some Non-invertible Explanatory Data Matrices

$\mathbf{X}'\mathbf{X}$ is invertible iff $\mathbf{X}$ is full column rank (see Blackwell, ch. 7.4), so the collection of predictors need to be linearly independent (no perfect collinearity).

Some example of $\mathbf{X}$ that are not full column rank:

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & -2 \\ 1 & 3 & -3 \\ 1 & 4 & -4 \\ 1 & 5 & -5 \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & 54 & 54,000 \\ 1 & 37 & 37,000 \\ 1 & 89 & 89,000 \\ 1 & 72 & 72,000 \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Preparation

The OLS Estimator in Matrix Algebra

- Vector Derivatives

- Derivation of OLS Estimator

- OLS Assumptions in Matrix Notation

- Properties of OLS Estimator

Frisch–Waugh–Lovell Theorem

The Linear Model in Matrix Form

Let's consider the OLS estimator. The general multiple regression model can be written as,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

where:

- $\mathbf{y}$ is the $n \times 1$ vector of responses
- $\mathbf{X}$ is the $n \times (k + 1)$ design matrix of predictor variables plus constant
- $\boldsymbol{\beta}$ is the $(k + 1) \times 1$ vector of regression coefficients (intercept and slopes)
- $\mathbf{u}$ is the $n \times 1$ vector of errors.

Sum of Squared Residuals

Recall that we derive the OLS estimator by minimizing the sum of squared residuals (SSR) with respect to β .

Let's define the $n \times 1$ vector of residuals for a hypothetical value of β , $\tilde{\beta}$, as $\tilde{u} = y - X\tilde{\beta}$.

Then we can write SSR as,

$$\begin{aligned} S(\tilde{\beta}, X, y) &= \tilde{u}'\tilde{u} \left(= \sum_{i=1}^n \tilde{u}_i^2 \right) \\ &= (y - X\tilde{\beta})'(y - X\tilde{\beta}) \\ &= (y' - (X\tilde{\beta})')(y - X\tilde{\beta}) \\ &= (y' - \tilde{\beta}'X')(y - X\tilde{\beta}) \\ &= y'y - \tilde{\beta}'X'y - y'X\tilde{\beta} + \tilde{\beta}'X'X\tilde{\beta} \\ &= y'y - (X\tilde{\beta})'y - y'X\tilde{\beta} + \tilde{\beta}'X'X\tilde{\beta} \\ &= y'y - y'(X\tilde{\beta}) - y'X\tilde{\beta} + \tilde{\beta}'X'X\tilde{\beta} \\ &= y'y - 2y'X\tilde{\beta} + \tilde{\beta}'X'X\tilde{\beta} \end{aligned}$$

Once we have defined the OLS estimator of β , how do we find the vectors of estimates $\tilde{\beta}$ that minimizes the SSR?

We compute the vector of partial derivatives and solve the resulting system of equations with an inverse matrix.

Gradient

Let $v = v(\mathbf{u})$ be a scalar-valued function $\mathbb{R}^n \rightarrow \mathbb{R}$ where $\mathbf{u}$ is a $(n \times 1)$ column vector. For example: $v(\mathbf{u}) = \mathbf{c}'\mathbf{u}$ where

$$\mathbf{c} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix} \text{ and } \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$$

Definition (gradient)

We can define the column vector of partial derivatives

$$\frac{\partial v(\mathbf{u})}{\partial \mathbf{u}} = \begin{bmatrix} \partial v / \partial u_1 \\ \partial v / \partial u_2 \\ \vdots \\ \partial v / \partial u_n \end{bmatrix}$$

This vector of partial derivatives is called the gradient.

Vector Derivative Rule I (linear functions)

Theorem (differentiation of linear functions)

Given a linear function $v(\mathbf{u}) = \mathbf{c}'\mathbf{u}$ of an $(n \times 1)$ vector $\mathbf{u}$, the derivative of $v(\mathbf{u})$ w.r.t. $\mathbf{u}$ is given by

$$\frac{\partial v}{\partial \mathbf{u}} = \mathbf{c}$$

This also works when $\mathbf{c}$ is a matrix and therefore v is a vector-valued function.

For example, let $v(\mathbf{u}) = \mathbf{c}'\mathbf{u}$ where $\mathbf{c} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}$ and $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$, then

$$v = \mathbf{c}'\mathbf{u} = 0 \cdot u_1 + 1 \cdot u_2 + 3 \cdot u_3$$

and

$$\frac{\partial v}{\partial \mathbf{u}} = \begin{bmatrix} \partial v / \partial u_1 \\ \partial v / \partial u_2 \\ \partial v / \partial u_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix} = \mathbf{c}$$

Hence,

$$\frac{\partial v}{\partial \mathbf{u}} = \mathbf{c}$$

Vector Derivative Rule II (quadratic form)

Theorem (quadratic form)

Given a $(n \times n)$ symmetric matrix $\mathbf{A}$ and a scalar-valued function $v(\mathbf{u}) = \mathbf{u}'\mathbf{A}\mathbf{u}$ of $(n \times 1)$ vector $\mathbf{u}$, we have

$$\frac{\partial v}{\partial \mathbf{u}} = \mathbf{A}'\mathbf{u} + \mathbf{A}\mathbf{u} = 2\mathbf{A}\mathbf{u}$$

For example, let $\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 5 \end{bmatrix}$ and $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$. Then $v(\mathbf{u}) = \mathbf{u}'\mathbf{A}\mathbf{u}$ is equal to

$$\begin{aligned} v &= [3 \cdot u_1 + u_2, \quad u_1 + 5 \cdot u_2] \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \\ &= 3u_1^2 + 2u_1u_2 + 5u_2^2 \end{aligned}$$

and

$$\frac{\partial v}{\partial \mathbf{u}} = \begin{bmatrix} \partial v / \partial u_1 \\ \partial v / \partial u_2 \end{bmatrix} = \begin{bmatrix} 6u_1 + 2u_2 \\ 2u_1 + 10u_2 \end{bmatrix} = 2 \cdot \begin{bmatrix} 3u_1 + 1u_2 \\ 1u_1 + 5u_2 \end{bmatrix} = 2\mathbf{A}\mathbf{u}$$

Suppose v is a scalar-valued function $v = f(\mathbf{u})$ of a $(k + 1) \times 1$ column vector $\mathbf{u} = \begin{bmatrix} u_1 & u_2 & \cdots & u_{k+1} \end{bmatrix}'$

Definition (Hessian)

The $(k + 1) \times (k + 1)$ matrix of second-order partial derivatives of $v = f(\mathbf{u})$ is called the **Hessian matrix** and denoted

$$\frac{\partial^2 v}{\partial \mathbf{u} \partial \mathbf{u}'} = \begin{bmatrix} \frac{\partial^2 v}{\partial u_1 \partial u_1} & \cdots & \frac{\partial^2 v}{\partial u_1 \partial u_{k+1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 v}{\partial u_{k+1} \partial u_1} & \cdots & \frac{\partial^2 v}{\partial u_{k+1} \partial u_{k+1}} \end{bmatrix}$$

Note: The Hessian is symmetric.

We will use the above rules and definitions to derive the OLS estimator next.

$$\begin{aligned} S(\tilde{\beta}, \mathbf{X}, \mathbf{y}) &= (\mathbf{y} - \mathbf{X}\tilde{\beta})'(\mathbf{y} - \mathbf{X}\tilde{\beta}) \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\tilde{\beta} + \tilde{\beta}'\mathbf{X}'\mathbf{X}\tilde{\beta} \end{aligned}$$

$$\frac{\partial S(\tilde{\beta}, \mathbf{X}, \mathbf{y})}{\partial \tilde{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\tilde{\beta}$$

- The first term does not contain $\tilde{\beta}$
- The second term is an example of rule I from the derivative section
- The third term is an example of rule II from the derivative section

And while we are at it the Hessian is:

$$\frac{\partial^2 S(\tilde{\beta}, \mathbf{X}, \mathbf{y})}{\partial \tilde{\beta} \partial \tilde{\beta}'} = 2\mathbf{X}'\mathbf{X}$$

$$\frac{\partial S(\tilde{\beta}, \mathbf{X}, \mathbf{y})}{\partial \tilde{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\tilde{\beta}$$

Setting the vector of partial derivatives equal to zero and substituting $\hat{\beta}$ for $\tilde{\beta}$, we can solve for the OLS estimator.

$$\begin{aligned} \mathbf{0} &= -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\beta} \\ -2\mathbf{X}'\mathbf{X}\hat{\beta} &= -2\mathbf{X}'\mathbf{y} \\ \mathbf{X}'\mathbf{X}\hat{\beta} &= \mathbf{X}'\mathbf{y} \\ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ \mathbf{I}\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ \hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \end{aligned}$$

This needs $\mathbf{X}'\mathbf{X}$ to be invertible, which holds when $\mathbf{X}$ has full column rank. Then the Hessian $2\mathbf{X}'\mathbf{X}$ is positive definite, so this solution minimizes the SSR.

Worksheet Q1

With $n = 3$, regress y on an intercept and x where

$$\mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 2 \end{bmatrix}.$$

- 1 Compute $\hat{\beta}$, fitted values $\hat{\mathbf{y}}$, residuals $\hat{\mathbf{u}}$.
- 2 Verify the normal equations $\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}$.
- 3 Compute $SSR = \hat{\mathbf{u}}'\hat{\mathbf{u}}$ and $\hat{\sigma}^2 = SSR/(n - (k + 1))$.

Regression Assumptions in Matrix Notation

- **I. Linearity:** The population relationship between $\mathbf{X}$ and $\mathbf{y}$ is correctly specified and linear in parameter, $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$
- **II. Random Sampling:** The observed data represent a random sample from the population described by the model.
- **III. No Perfect Collinearity:** $\mathbf{X}$ is a $n \times (k + 1)$ matrix with rank $k + 1$ (full column rank).
- **IV. Zero Conditional Mean:** $E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$, which implies $E[\mathbf{u}] = \mathbf{0}$ and that each u_i is uncorrelated with every element of $\mathbf{X}$.
- **V. Homoskedasticity:** $V[\mathbf{u}|\mathbf{X}] = \sigma^2 \mathbf{I}_n$ the conditional variance of the error term is constant for any value of $\mathbf{X}$.
- **VI. Normality:** $\mathbf{u}|\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ which implies that $\mathbf{y}|\mathbf{X} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$.

Expectation of a Random Vector

The concept of expectation naturally extends to vectors of random variables:

Definition (expected value of random vector)

For a $(n \times 1)$ random vector $\mathbf{u}$, the expected value of $\mathbf{u}$, denoted $E[\mathbf{u}]$, is the vector of expected values:

$$E[\mathbf{u}] = \begin{bmatrix} E[u_1] & E[u_2] & \dots & E[u_n] \end{bmatrix}'$$

For non-random $\mathbf{A}$ and $\mathbf{B}$, the expectation operator works as in the scalar case:

$$E[\mathbf{A}\mathbf{u} + \mathbf{B}] = \mathbf{A}E[\mathbf{u}] + \mathbf{B}$$

If $\mathbf{A}$ and $\mathbf{B}$ are functions of $\mathbf{X}$, the same holds conditional on $\mathbf{X}$: $E[\mathbf{A}\mathbf{u} + \mathbf{B} \mid \mathbf{X}] = \mathbf{A}E[\mathbf{u} \mid \mathbf{X}] + \mathbf{B}$.

Is $E[\hat{\beta}] = \beta$?

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \text{ (linearity and no collinearity)}$$

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'(\mathbf{X}\beta + \mathbf{u})$$

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u}$$

$$\hat{\beta} = \mathbf{I}\beta + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u}$$

$$\hat{\beta} = \beta + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u}$$

$$E[\hat{\beta}|\mathbf{X}] = E[\beta|\mathbf{X}] + E[(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u}|\mathbf{X}]$$

$$E[\hat{\beta}|\mathbf{X}] = \beta + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'E[\mathbf{u}|\mathbf{X}]$$

$$E[\hat{\beta}|\mathbf{X}] = \beta \text{ (zero conditional mean)}$$

So, yes: by the law of iterated expectations, $E[\hat{\beta}] = E[E[\hat{\beta}|\mathbf{X}]] = \beta$.

Variance-Covariance Matrix of Random Vectors

Let's unpack the **homoskedasticity** assumption $V[\mathbf{u}|\mathbf{X}] = \sigma^2 \mathbf{I}_n$.

Definition (variance-covariance matrix)

For a $(n \times 1)$ random vector $\mathbf{u} = \begin{bmatrix} u_1 & u_2 & \dots & u_n \end{bmatrix}'$, its **variance-covariance matrix**, denoted $V[\mathbf{u}]$ or also $\Sigma_{\mathbf{u}}$, is defined as:

$$V[\mathbf{u}] = \Sigma_{\mathbf{u}} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & & \dots & \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{bmatrix}$$

where $\sigma_j^2 = V[u_j]$ and $\sigma_{ij} = \text{Cov}[u_i, u_j]$.

Notice that this matrix is always symmetric.

Homoskedasticity in Matrix Notation

If $V[u_i] = \sigma^2$ for all $i = 1, \dots, n$ and the units are independent then $V[\mathbf{u}] = \sigma^2 \mathbf{I}_n$.

More visually:

$$V[\mathbf{u}] = \sigma^2 \mathbf{I}_n = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

So homoskedasticity $V[\mathbf{u}|\mathbf{X}] = \sigma^2 \mathbf{I}_n$ implies that:

- 1 $V[u_i|\mathbf{X}] = \sigma^2$ for all i (the variance of the errors u_i does not depend on $\mathbf{X}$ and is constant across observations)
- 2 $\text{Cov}[u_i, u_j|\mathbf{X}] = 0$ for all $i \neq j$ (the errors are uncorrelated across observations). This holds under our random sampling assumption.

Variance of Linear Function of Random Vector

Definition (Variance of Linear Transformation of Random Vector)

Recall that for a linear transformation of a random variable X we have $V[aX + b] = a^2 V[X]$ with constants a and b .

There is an analogous rule for linear functions of random vectors. Let $v(\mathbf{u}) = \mathbf{A}\mathbf{u} + \mathbf{B}$ be a linear transformation of a random vector $\mathbf{u}$ with non-random vectors or matrices $\mathbf{A}$ and $\mathbf{B}$. Then the variance of the transformation is given by:

$$V[v(\mathbf{u})] = V[\mathbf{A}\mathbf{u} + \mathbf{B}] = \mathbf{A}V[\mathbf{u}]\mathbf{A}' = \mathbf{A}\Sigma_{\mathbf{u}}\mathbf{A}'$$

If $\mathbf{A}$ and $\mathbf{B}$ are functions of $\mathbf{X}$, the same holds conditional on $\mathbf{X}$: $V[\mathbf{A}\mathbf{u} + \mathbf{B} \mid \mathbf{X}] = \mathbf{A}V[\mathbf{u} \mid \mathbf{X}]\mathbf{A}'$.

Worksheet Q2

Under homoskedasticity and independent sampling, $V[\mathbf{u}|\mathbf{X}] = \sigma^2 \mathbf{I}_n$.

- 1 Prove that $V[\hat{\beta}|\mathbf{X}] = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$.
- 2 For any fixed $(k+1) \times 1$ vector $\mathbf{a}$, derive $V[\mathbf{a}'\hat{\beta}|\mathbf{X}]$ and the plug-in standard error.
- 3 Return to the data in Worksheet Q1 and report the estimated variance–covariance matrix $\hat{V}[\hat{\beta}] = \hat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1}$.

Hint: To estimate $V[\hat{\beta}|\mathbf{X}]$, we replace σ^2 with its unbiased estimator $\hat{\sigma}^2$, which is now written using matrix notation as:

$$\hat{\sigma}^2 = \frac{SSR}{n - (k + 1)} = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n - (k + 1)}$$

Variance-covariance matrix of $\hat{\beta}$

The **variance-covariance matrix** of the OLS estimators is given by:

$$V[\hat{\beta}|\mathbf{X}] = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} =$$

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\dots$	$\hat{\beta}_k$
$\hat{\beta}_0$	$V[\hat{\beta}_0]$	$\text{Cov}[\hat{\beta}_0, \hat{\beta}_1]$	$\text{Cov}[\hat{\beta}_0, \hat{\beta}_2]$	$\dots$	$\text{Cov}[\hat{\beta}_0, \hat{\beta}_k]$
$\hat{\beta}_1$	$\text{Cov}[\hat{\beta}_0, \hat{\beta}_1]$	$V[\hat{\beta}_1]$	$\text{Cov}[\hat{\beta}_1, \hat{\beta}_2]$	$\dots$	$\text{Cov}[\hat{\beta}_1, \hat{\beta}_k]$
$\hat{\beta}_2$	$\text{Cov}[\hat{\beta}_0, \hat{\beta}_2]$	$\text{Cov}[\hat{\beta}_1, \hat{\beta}_2]$	$V[\hat{\beta}_2]$	$\dots$	$\text{Cov}[\hat{\beta}_2, \hat{\beta}_k]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\hat{\beta}_k$	$\text{Cov}[\hat{\beta}_0, \hat{\beta}_k]$	$\text{Cov}[\hat{\beta}_k, \hat{\beta}_1]$	$\text{Cov}[\hat{\beta}_k, \hat{\beta}_2]$	$\dots$	$V[\hat{\beta}_k]$

Variance-covariance matrix of $\hat{\beta}$

```
R Code
> fit <- lm(vote1 ~ fem + educ + age, data = d)
> summary(fit)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4917751  0.0525717   9.354  < 2e-16 ***
fem          0.1284292  0.0236221   5.437 6.21e-08 ***
educ        -0.1051514  0.0164543  -6.391 2.13e-10 ***
age          0.0033573  0.0008235   4.077 4.77e-05 ***
---
```

We can pull out the variance-covariance matrix $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ in R from the `lm()` object:

```
R Code
> V <- vcov(fit)
> V
              (Intercept)              fem              educ              age
(Intercept)  2.763788e-03 -3.717198e-04 -6.543766e-04 -3.324904e-05
fem          -3.717198e-04  5.580053e-04  3.781282e-05  9.631845e-07
educ         -6.543766e-04  3.781282e-05  2.707446e-04  3.860967e-06
age          -3.324904e-05  9.631845e-07  3.860967e-06  6.780985e-07

> sqrt(diag(V))
(Intercept)              fem              educ              age
0.0525717389 0.0236221360 0.0164543200 0.0008234674
```

These are the conventional SEs. They assume homoskedasticity (Assumption V), which a 0/1 outcome makes doubtful. Here robust SEs (Lecture 6.3) are almost the same: 0.0237 instead of 0.0236 for `fem`.

Worksheet Q3

Suppose

- the true model is $y_i = \mathbf{x}_i\boldsymbol{\beta} + u_i$, where $\mathbf{x}_i$ is row i of $\mathbf{X}$ (a $1 \times (k + 1)$ row vector);
- with $\mathbb{E}[u_i|\mathbf{x}_i] = 0$ and $\mathbb{E}[u_i^2] < \infty$;
- the sample $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ is i.i.d.;
- $\mathbb{E}[\mathbf{x}_i'\mathbf{x}_i]$ is finite and nonsingular.

① Show that $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$.

② Assume further that $E[u_i^2|\mathbf{x}_i] = \sigma^2 < \infty$, derive the asymptotic distribution of $\hat{\boldsymbol{\beta}}$.

Theorem

Under Assumptions I–VI, the $(k + 1) \times 1$ vector of OLS estimators $\hat{\beta}$, conditional on $\mathbf{X}$, follows a *multivariate normal distribution* with mean β and variance-covariance matrix $\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$:

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

- Each element of $\hat{\beta}$ (i.e. $\hat{\beta}_0, \dots, \hat{\beta}_k$) is normally distributed, and $\hat{\beta}$ is an unbiased estimator of β as $E[\hat{\beta}] = \beta$
- Variances and covariances are given by $V[\hat{\beta}|\mathbf{X}] = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$
- An unbiased estimator for the error variance σ^2 is given by

$$\hat{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n - (k + 1)}$$

- With a large sample, $\hat{\beta}$ approximately follows the same distribution under Assumptions I–V only, i.e., without assuming the normality of $\mathbf{u}$.

Preparation

The OLS Estimator in Matrix Algebra

Frisch–Waugh–Lovell Theorem

Projection and Annihilator Matrices

Frisch–Waugh–Lovell Theorem

Worksheet Q4

Assume the classical linear regression model satisfying Assumptions I–V:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u},$$

where $\mathbf{X}$ is an $n \times (k + 1)$ matrix of full column rank, and $\mathbf{y}$ is an $n \times 1$ vector of responses.

- ❶ What does $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ represent in the context of OLS estimation?
- ❷ What is the interpretation of $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$?
- ❸ What does $\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ represent?
- ❹ What does $[\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}]'[\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}]$ compute?
- ❺ What does $[\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}]'[\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}]/(n - (k + 1))$ estimate?
- ❻ What is the interpretation of

$$(\mathbf{X}'\mathbf{X})^{-1} \frac{[\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}]'[\mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}]}{(n - (k + 1))} ?$$

Projection Matrix and Annihilator Matrix

$$\begin{aligned}\hat{\mathbf{u}} &= \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \equiv \mathbf{y} - \mathbf{P}\mathbf{y} \\ &= (\mathbf{I} - \mathbf{P})\mathbf{y} \equiv \mathbf{M}\mathbf{y}\end{aligned}$$

$\mathbf{P}$ is called the **projection matrix**:

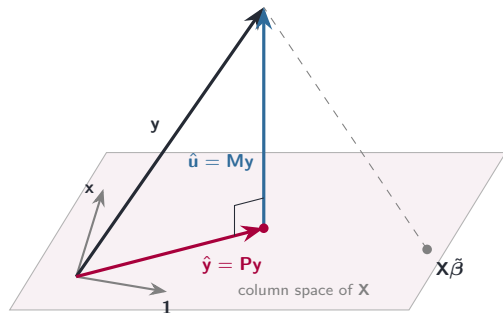
- The above shows $\hat{\mathbf{y}} = \mathbf{P}\mathbf{y}$, i.e., premultiplying a vector by $\mathbf{P}$ turns the vector into its linear fit with respect to $\mathbf{X}$.
- $\mathbf{P}$ is therefore also known as the **hat matrix**

$\mathbf{M}$ is called the **annihilator matrix** (because $\mathbf{MX} = \mathbf{0}$) and more intuitively the “residual maker” matrix:

- The above shows $\hat{\mathbf{u}} = \mathbf{M}\mathbf{y}$, i.e., pre-multiplying a vector by $\mathbf{M}$ turns the vector into the residual from its regression on $\mathbf{X}$.
- $\mathbf{M}$ is therefore a “residual creator” matrix

$\mathbf{P}$ and $\mathbf{M}$ are both symmetric and **idempotent** (i.e. $\mathbf{PP} = \mathbf{P}$ and $\mathbf{MM} = \mathbf{M}$).

The Geometry of Least Squares



With $n = 3$ observations, $\mathbf{y}$, $\mathbf{1}$ and $\mathbf{x}$ are vectors in 3-D space.

- Every fit $\mathbf{X}\tilde{\boldsymbol{\beta}}$ lies in the plane spanned by the columns of $\mathbf{X}$: the **column space** of $\mathbf{X}$.
- OLS picks the point of the plane closest to $\mathbf{y}$: $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{P}\mathbf{y}$.
- The shortest path from $\mathbf{y}$ to the plane meets it at a right angle. So $\hat{\mathbf{u}} = \mathbf{M}\mathbf{y}$ is orthogonal to every column of $\mathbf{X}$: $\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}$.
- $\mathbf{y} = \mathbf{P}\mathbf{y} + \mathbf{M}\mathbf{y}$: $\mathbf{P}$ keeps the part of $\mathbf{y}$ in the plane; $\mathbf{M}$ keeps the part perpendicular to it.

Source: Adapted from Blackwell, *A User's Guide to Statistical Inference and Regression*, ch. 7

Worksheet Q5

Assume the classical linear regression model satisfying Assumptions I–V: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, where $\mathbf{X}$ is an $n \times (k + 1)$ matrix of full column rank, and $\mathbf{y}$ is an $n \times 1$ vector of responses. Define:

$$\mathbf{Q} = \mathbf{X}'\mathbf{X}, \quad \mathbf{A} = \mathbf{Q}^{-1}\mathbf{X}', \quad \mathbf{P} = \mathbf{X}\mathbf{A}, \quad \mathbf{M} = \mathbf{I} - \mathbf{P}.$$

- 1 Express the residuals from regressing $\mathbf{y}$ on $\mathbf{X}$ as a function of $\mathbf{M}$.
- 2 Show that $\mathbf{P}$ is symmetric and idempotent.
- 3 Show that $\mathbf{M}$ is idempotent (verify $\mathbf{M}'\mathbf{M} = \mathbf{M}$ and $\mathbf{M}^2 = \mathbf{M}$).
- 4 Show that $\mathbf{AP} = \mathbf{A}$, $\mathbf{AM} = \mathbf{0}$, $\mathbf{MP} = \mathbf{0}$, and $\mathbf{PM} = \mathbf{0}$.
- 5 Show that $\mathbf{PX} = \mathbf{X}$, $\mathbf{MX} = \mathbf{0}$.
- 6 Show that $\mathbf{P}\hat{\mathbf{y}} = \hat{\mathbf{y}}$, $\mathbf{P}\hat{\mathbf{u}} = \mathbf{0}$.
- 7 Show that $\mathbf{M}\hat{\mathbf{y}} = \mathbf{0}$, $\mathbf{M}\hat{\mathbf{u}} = \hat{\mathbf{u}}$.
- 8 Show that $\mathbf{X}'\hat{\mathbf{y}} = \mathbf{X}'\mathbf{y}$.
- 9 Show that $\hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{y}'\mathbf{M}\mathbf{y} = \mathbf{y}'\mathbf{y} - \hat{\mathbf{y}}'\hat{\mathbf{y}}$.

Frisch–Waugh–Lovell Theorem

- In multiple regression

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{u},$$

we often want to interpret the coefficient β_2 as the effect of $\mathbf{X}_2$ *controlling for* $\mathbf{X}_1$.

- The FWL theorem shows that we can obtain $\hat{\beta}_2$ in two equivalent ways:
 - ① Regress $\mathbf{y}$ directly on both $\mathbf{X}_1$ and $\mathbf{X}_2$ (full regression), or
 - ② First remove from $\mathbf{y}$ and $\mathbf{X}_2$ the parts explained by $\mathbf{X}_1$ (“partial out” $\mathbf{X}_1$), then regress the residuals of $\mathbf{y}$ on those of $\mathbf{X}_2$.
- The theorem provides a clean geometric interpretation using projection and residual-maker matrices.

Frisch–Waugh–Lovell Theorem

The coefficient $\hat{\beta}_2$ from regressing $\mathbf{y}$ on $(\mathbf{X}_1, \mathbf{X}_2)$ equals that from regressing the residualized $\mathbf{y}$ on residualized $\mathbf{X}_2$:

$$\hat{\beta}_2 = (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{y}, \quad \mathbf{M}_1 = \mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$$

The two regressions also leave the same residuals.