Linear Regression: Introduction

Le Wang

Math Preliminaries

Transpose

If a is a scalar, $a = a'$.

Three examples used below:

1. $x'\beta = (x'\beta)' = \beta' x$
2. $yx'\beta = (yx'\beta)' = \beta' xy$
3. $Y'(X\beta) = (Y'(X\beta))' = \beta' X' Y$

Matrix Calculus

Let x be a $k \times 1$ vector and $g(x) = g(x_1, x_2, \ldots, x_k) : R^k \to R$.
The vector derivative is

$$\frac{\partial}{\partial x} g(x) = \begin{pmatrix} \frac{\partial}{\partial x_1} g(x) \\ \frac{\partial}{\partial x_2} g(x) \\ \vdots \\ \frac{\partial}{\partial x_k} g(x) \end{pmatrix}$$

Matrix Calculus

Lets look at the scalar case

$$
\begin{array}{ccc}
f(x) & \rightarrow & \frac{\mathrm{d}f}{\mathrm{d}x} \\
\hline
bx & \rightarrow & b \\
bx & \rightarrow & b \\
x^2 & \rightarrow & 2x \\
bx^2 & \rightarrow & 2bx
\end{array}
$$

The corresponding vector deriviatives look like the following

$$
\begin{array}{rl}
f(\mathbf{x}) \rightarrow \dfrac{\mathrm{d}f}{\mathrm{d}\mathbf{x}} & \\
\hline
& \\
\mathbf{x}'\mathbf{B} \rightarrow \mathbf{B} & \\
\mathbf{x}'\mathbf{b} \rightarrow \mathbf{b} & \\
\mathbf{x}'\mathbf{x} \rightarrow 2\mathbf{x} & \\
\mathbf{x}'\mathbf{B}\mathbf{x} \rightarrow 2\mathbf{B}\mathbf{x} & (\mathbf{B} + \mathbf{B}')\mathbf{x}
\end{array}
$$

Solution (Math Preliminaries)

Special Cases:

$$
\begin{aligned}
\frac{\partial}{\partial x}(b'x) = \frac{\partial}{\partial x}(x'b) = a &\implies \frac{\partial}{\partial \beta}(\beta' xy) = xy \\
\frac{\partial}{\partial x}(x'Bx) = (B + B')x &\implies \frac{\partial}{\partial \beta}(\beta' xx' \beta) = (xx' + (xx')') = 2xx' \\
\frac{\partial}{\partial x'}(Bx) &= B \\
\frac{\partial^2}{\partial x \partial x'}(x'Bx) &= (B + B')
\end{aligned}
$$

Multiple Regression

Multiple Regression

Lets now consider a more general case. **Suppose** that our model is given by

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i$$

How do we choose $\beta_0, \beta_1, \ldots, \beta_k$?

Multiple Regression

We can re-write the linear regression in a more compact form

$$y_i = x_i'\beta + \epsilon_i$$

where

$$x = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

Solution Techniques

There are two (numerically identical) approaches (there are actually more, but we will discuss them later):

1. Method of Moments (MOM) Estimator: Specify the objective as a **population subject**, solve for the anlytical solution, and then derive the **sample estimator**.

2. OLS Estimator: Specify the objective as a **sample subject**, and directly derive the **sample estimator**.

Estimation (1): Method of Moment Estimator

MOM Estimator

We will choose $\beta \in R^k$ to minimize the mean squared error

$$S(\beta) = \mathbb{E}\left[\epsilon^2\right] = \mathbb{E}\left[(y - x'\beta)^2\right]$$

The optimal β^* sets the partial derivatives to zero

$$\frac{\partial}{\partial \beta} S(\beta) = 0$$

Solution

$$
\begin{aligned}
S(\beta) &= \mathbb{E}\left[(y - x'\beta)(y - x'\beta)\right] \\
&= \mathbb{E}\left[(y - \beta'x)(y - x'\beta)\right] \\
&= \mathbb{E}\left[y^2 - \beta'xy - yx'\beta + \beta'xx'\beta\right] \\
&= \mathbb{E}\left[y^2 - \beta'xy - \beta'xy + \beta'xx'\beta\right] \\
&= \mathbb{E}\left[y^2 - 2\beta'xy + \beta'xx'\beta\right]
\end{aligned}
$$

Solution

$$
\begin{aligned}
\frac{\partial}{\partial \beta} S(\beta) &= \frac{\partial}{\partial \beta} \mathbb{E}\left[(y - x'\beta)(y - x'\beta)\right] \\
&= \frac{\partial}{\partial \beta} \mathbb{E}\left[y^2 - 2\beta' xy + \beta' xx' \beta\right] \\
&= \mathbb{E}\left[-2xy + [xx' + (xx')']\beta\right] \\
&= \mathbb{E}\left[-2xy + [xx' + (x')'(x)']\beta\right] \\
&= \mathbb{E}\left[-2xy + 2xx'\beta\right]
\end{aligned}
$$

Solution

Setting $\frac{\partial}{\partial\beta}S(\beta) = 0$, we obtain

$$\begin{aligned}
\mathbb{E}\left[-2xy + 2xx'\beta^*\right] &= 0 \\
\mathbb{E}\left[xx'\beta^*\right] &= \mathbb{E}\left[xy\right] \\
\beta^* &= \left(\mathbb{E}\left[xx'\right]\right)^{-1}\mathbb{E}\left[xy\right]
\end{aligned}$$

Implicit Regularity Assumptions

1. $\mathbb{E}[y^2] < \infty$
2. $\mathbb{E}[||x||^2] = \mathbb{E}[x'x] < \infty$
3. $\mathbb{E}[xx']$ is positive definite (equivalent to assuming that the matrix is invertible, or that the columns of the matrix are linearly independent).

Estimation of the population regression coefficients

$$\beta^* = \left(\mathbb{E}\left[xx'\right]\right)^{-1}\mathbb{E}\left[xy\right]$$

These are still abstract concepts, and how do we obtain these **population** parameters using the **samples**?

There are two (numerically identical) approaches (there are actually more, but we will discuss them later):

1. Method of Moments (MOM) Estimator: Sample version of the coefficient

Estimation

1. Sampling design
2. MOM Estimator

Estimation: Sampling

If we are interested in estimating some linear relationship between wage and education, we would collect a set of observations on wages and education for many individuals.

We wish to distinguish observations from the underlying random variables. The convention in econometrics is to denote observations by appending a subscript i which runs from 1 to n.

$$\{(y_i, x_{1i}); i = 1, \ldots, N\}$$

Estimation: Sampling

We also want to make sure that the expectations over the random variables are common across the observations. The most elegant approach to ensure this is to assume the following

Assumption The observations $\{(y_i, x_{1i}); i = 1, \ldots, N\}$ are identically distributed; they are draws from a common distribution F.

This assumption does not need to be viewed as literally true, rather it is a useful modeling device so that parameters such as βs are well defined. This assumption should be interpreted as how we view an observation a priori, before we actually observe it.

Example: cross-sectional data are assumed to be i.i.d (identically, independently distributed).

Estimation: MOM

Question: What would be a good estimator of population expectation from the sample of observations?

$$\mathbb{E}[\cdot]$$

Answer: An appropriate estimator is the **sample moments**

Estimation: MOM

Example 1: Suppose that we are interested in the population mean μ of a random variable y_i with distribution function, F,

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

To estimate μ given a sample of $\{y_1, \ldots, y_N\}$, a natural estimator is the sample mean

$$\widehat{\mu} = \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

Note that the notation with the hat $\widehat{\cdot}$ is conventional in econometrics to denote an estimator of the parameter.

Estimation: MOM

Example 2: Suppose that we are intersted in a set of population means of possibly non-linear functions of a random vector, $\mathbf{y}_i$, say that $\mu = \mathbb{E}[\mathbf{h}(\mathbf{y_i})]$

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} \mathbb{E}[y_i] \\ \mathbb{E}[y_i^2] \end{bmatrix}$$

A natural MOM estimator would be

$$\widehat{\mu} = \begin{bmatrix} \widehat{\mu}_1 \\ \widehat{\mu}_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} y_i \\ \frac{1}{N}\sum_{i=1}^{N} y_i^2 \end{bmatrix}$$

Estimation: MOM

General Principle: Suppose that we are interested in a non-linear **function** of a set of **moments**

$$\beta = \mathbf{g}(\mu)$$
$$\mu = \mathbb{E}[\mathbf{h}(\mathbf{y_i})]$$

A moment-based estimator of β,

$$\widehat{\beta} = \mathbf{g}(\widehat{\mu})$$
$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{h}(\mathbf{y_i})$$

Methods of Moment Estimator

True linear regression:

$$\beta^* = \left(\mathbb{E}\left[xx'\right]\right)^{-1} \mathbb{E}\left[xy\right]$$

MOM Estimator:

$$\begin{aligned}
\beta^* &= \left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i'\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} x_i y_i\right) \\
&= \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1} \left(\sum_{i=1}^{N} x_i y_i\right)
\end{aligned}$$

Numerical Example

$$Y = \begin{pmatrix} 73 \\ 50 \end{pmatrix}, X = \begin{pmatrix} 1 & 30 \\ 1 & 20 \end{pmatrix}$$

$$\begin{aligned}
\sum_{i=1}^{N} x_i x_i' &= \begin{pmatrix} 1 \\ 30 \end{pmatrix} \begin{pmatrix} 1, 30 \end{pmatrix} + \begin{pmatrix} 1 \\ 20 \end{pmatrix} \begin{pmatrix} 1, 20 \end{pmatrix} \\
&= \begin{pmatrix} 1^2 & 1 \cdot 30 \\ 30 \cdot 1 & 30^2 \end{pmatrix} + \begin{pmatrix} 1^2 & 1 \cdot 20 \\ 20 \cdot 1 & 20^2 \end{pmatrix} \\
&= \begin{pmatrix} 1^2 + 1^2 & 1 \cdot 30 + 1 \cdot 20 \\ 30 \cdot 1 + 20 \cdot 1 & 30^2 + 20^2 \end{pmatrix}
\end{aligned}$$

$$X = \begin{pmatrix} 1 & 30 \\ 1 & 20 \end{pmatrix}$$

$$\begin{aligned} X'X &= \begin{pmatrix} 1 & 30 \\ 1 & 20 \end{pmatrix}' \begin{pmatrix} 1 & 30 \\ 1 & 20 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 1 \\ 30 & 20 \end{pmatrix} \begin{pmatrix} 1 & 30 \\ 1 & 20 \end{pmatrix} \\ &= \begin{pmatrix} 1^2 + 1^2 & 1 \cdot 30 + 1 \cdot 20 \\ 30 \cdot 1 + 20 \cdot 1 & 30^2 + 20^2 \end{pmatrix} \end{aligned}$$

Methods of Moment Estimator

MOM Estimator (Matrix Form):

$$\beta^* = \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1} \left(\sum_{i=1}^{N} x_i y_i\right)$$
$$= (X'X)^{-1} X'Y$$

where

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, X = \begin{pmatrix} (x_{11}, x_{21}, \ldots, x_{k1}) \\ (x_{12}, x_{22}, \ldots, x_{k2}) \\ \vdots \\ (x_{1N}, x_{21}, \ldots, x_{kN}) \end{pmatrix} = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_N' \end{pmatrix}$$

Estimation (2): OLS

Estimation: Ordinary Least Squares (OLS)

Instead of solving the mean (or the expectation of) squared error first for a closed-form population coefficient and then find the sample-moment version of it, we can take a different route by solving the sample-moment version of the **objective function** – mean squared error!

Regression Coefficient One criterion is to choose the values that minimize the mean squared error

$$\beta = \arg\min \mathbb{E}[(y - x'\beta)^2]$$

OLS Coefficient One criterion is to choose the values that minimize the **sample** mean squared error

$$\widehat{\beta}^{OLS} = \arg\min \frac{1}{N} \sum_{i=1}^{N} (y_i - x_i'\beta)^2$$

Estimation: OLS

The sample-moment version of the objective function can be thought of as

$$\begin{aligned}\widehat{S}(\beta) &= \frac{1}{N}\sum_{i=1}^{N}\left(y_i - x_i'\beta\right)^2 \\ &= \frac{1}{N}SSE(\beta)\end{aligned}$$

where $SSE(\beta) = \sum_{i=1}^{N}\left(y_i - x_i'\beta\right)^2$ is called **sum-of-squared-errors** function.

Multiple Regression: OLS

$$\widehat{S}(\beta) = \sum_{i=1}^{N} \epsilon_i^2 = \begin{pmatrix} \epsilon_1, \ldots, \epsilon_N \end{pmatrix} \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix}$$

$$= \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix}' \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix}$$

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix} = \begin{pmatrix} y_1 - x_1'\beta \\ y_2 - x_2'\beta \\ \vdots \\ y_2 - x_2'\beta \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} - \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_N' \end{pmatrix} \beta$$

$$
\begin{aligned}
\widehat{S}(\beta) &= \sum_{i=1}^{N} \epsilon_i^2 = \epsilon'\epsilon \\
&= \left[\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} - \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_N' \end{pmatrix} \beta \right]' \left[\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} - \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_N' \end{pmatrix} \beta \right] \\
&= (Y - X\beta)'(Y - X\beta)
\end{aligned}
$$

where

$$
\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix}, Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, X = \begin{pmatrix} (x_{11}, x_{21}, \ldots, x_{k1}) \\ (x_{12}, x_{22}, \ldots, x_{k2}) \\ \vdots \\ (x_{1N}, x_{21}, \ldots, x_{kN}) \end{pmatrix} = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_N' \end{pmatrix}
$$

Multiple Regression: OLS

The optimal value, β, sets the first order conditions to zeros

$$\begin{aligned}
\frac{\partial}{\partial\beta}\epsilon'\epsilon &= \frac{\partial}{\partial\beta}(Y - X\beta)'(Y - X\beta) \\
&= \frac{\partial}{\partial\beta}(Y' - \beta'X')(Y - X\beta) \\
&= \frac{\partial}{\partial\beta}(Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta) \\
&= \frac{\partial}{\partial\beta}(Y'Y - 2\beta'X'Y + \beta'X'X\beta) \Leftarrow [(Y'X\beta)' = \beta'X'Y] \\
&= -2X'Y + (X'X + (X'X)')\beta \Leftarrow [\text{rule:}\frac{\partial}{\partial x}(x'a) = a] \\
&\qquad \Leftarrow [\text{rule:}\frac{\partial}{\partial x}(x'Ax) = (A + A')x]
\end{aligned}$$

Proof

OLS Estimator

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

A Numerical Example

A numerical example

Suppose that we have total wages and the number of hours worked as follows

$$Y = \begin{pmatrix} 73 \\ 50 \\ 128 \\ 170 \\ 87 \\ 108 \\ 135 \\ 69 \\ 148 \\ 132 \end{pmatrix}, X = \begin{pmatrix} 1 & 30 \\ 1 & 20 \\ 1 & 60 \\ 1 & 80 \\ 1 & 40 \\ 1 & 50 \\ 1 & 60 \\ 1 & 30 \\ 1 & 70 \\ 1 & 60 \end{pmatrix}$$

We can calculate the following quantities

$$(X'X)^{-1} = \begin{pmatrix} 0.83529412 & -0.01470588 \\ -0.01470588 & 0.00029412 \end{pmatrix}, X'Y = \begin{pmatrix} 1100 \\ 61800 \end{pmatrix}$$

We can calculate the following quantities

$$(X'X)^{-1} = \begin{pmatrix} 0.83529412 & -0.01470588 \\ -0.01470588 & 0.00029412 \end{pmatrix}, X'Y = \begin{pmatrix} 1100 \\ 61800 \end{pmatrix}$$

$$\widehat{\beta} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} = \begin{pmatrix} 0.83529412 & -0.01470588 \\ -0.01470588 & 0.00029412 \end{pmatrix} \begin{pmatrix} 1100 \\ 61800 \end{pmatrix} = \begin{pmatrix} 10.0 \\ 2.0 \end{pmatrix}$$

Estimation in Stata:

Stata code: -reg y x-

Lets look at some programming examples, if time permits.

Appendix

Summary of Appendix

1. Single-Variable Linear Regression
2. Connection between multiple regression results and single-variable results
3. Alternative Proof of OLS Estimator (Skip)

Single-Variable Linear Regression (Model Set-up)

Suppose that we have the following model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

How do we choose β_0, β_1?

One criterion is to choose the values that minimize the mean squared error

$$(\beta_0, \beta_1) = \arg\min \mathbb{E}[(y - \beta_0 - \beta_1 \cdot x_1)^2]$$

The regression coefficients:

$$\beta_1^* = \frac{\mathbb{E}[(y - \mu_y)(x - \mu_x)]}{\mathbb{E}[(x - \mu)^2]} = \frac{cov(y, x)}{var(x)}$$

$$\beta_0^* = \mu_y - \beta_1^* \cdot \mu_x$$

The optimal values, β_0^*, β_1^* set the partial derivatives to zero

$$
\begin{aligned}
\frac{\partial S}{\partial \beta_0} &= 0 \\
\mathbb{E}[-2(y - \beta_0^* - \beta_1^* \cdot x_1)] &= 0 \\
&\implies \\
\beta_0^* &= \mathbb{E}[y] - \beta_1^* \cdot \mathbb{E}[x_1] \\
\beta_0^* &= \mu_y - \beta_1^* \cdot \mu_x
\end{aligned}
$$

Proof: (Part II): β_1

$$\frac{\partial S}{\partial \beta_1} = 0$$

$$\frac{\partial}{\partial \beta_1}\mathbb{E}[(y - \beta_0^* - \beta_1^* \cdot x_1)^2] = 0$$

$$\frac{\partial}{\partial \beta_1}\mathbb{E}[(y - (\mu_y - \beta_1^* \cdot \mu_x) - \beta_1^* \cdot x_1)^2] = 0$$

$$\frac{\partial}{\partial \beta_1}\mathbb{E}[((y - \mu_y) - \beta_1^* \cdot (x - \mu_x))^2] = 0$$

$$\frac{\partial}{\partial \beta_1}\mathbb{E}\Big[\Big((y - \mu_y)^2 - 2\beta_1^* \cdot (y - \mu_y) \cdot (x - \mu_x) + (\beta_1^*)^2(x - \mu_x)^2\Big] = 0$$

$$\Longrightarrow$$

$$\mathbb{E}[-2 \cdot (y - \mu_y)(x - \mu_x) + 2\beta_1^* \cdot (x - \mu_x)^2] = 0$$

$$
\begin{aligned}
\mathbb{E}[-2 \cdot (y - \mu_y)(x - \mu_x) + 2\beta_1^* \cdot (x - \mu_x)^2] &= 0 \\
\beta_1^* \cdot \mathbb{E}[(x - \mu)^2] &= \mathbb{E}[(y - \mu_y)(x - \mu_x)] \\
\beta_1^* &= \frac{\mathbb{E}[(y - \mu_y)(x - \mu_x)]}{\mathbb{E}[(x - \mu)^2]}
\end{aligned}
$$

Solution connecting to the single regression

We can compare the vector form of the solution to the previous one in the case of one covariate and an intercept

$$x = \begin{pmatrix} 1 \\ x_1 \end{pmatrix}$$

$$xx' = \begin{pmatrix} 1 \\ x_1 \end{pmatrix} \cdot \begin{pmatrix} 1, x_1 \end{pmatrix} = \begin{pmatrix} 1, & x_1 \\ x_1, & x_1^2 \end{pmatrix}$$

$$\left(\mathbb{E}[xx']\right)^{-1} = \begin{pmatrix} \mathbb{E}[1], & \mathbb{E}[x_1] \\ \mathbb{E}[x_1], & \mathbb{E}[x_1^2] \end{pmatrix}^{-1} = \frac{1}{\mathbb{E}[x_1^2] - \left(\mathbb{E}[x_1]\right)^2} \begin{pmatrix} \mathbb{E}[x_1^2], & -\mathbb{E}[x_1] \\ -\mathbb{E}[x_1], & 1 \end{pmatrix}$$

$$\mathbb{E}[xy] = \mathbb{E}\left[\begin{pmatrix} 1 \\ x_1 \end{pmatrix} y\right] = \mathbb{E}\left[\begin{pmatrix} y \\ x_1 y \end{pmatrix}\right] = \begin{pmatrix} \mathbb{E}[y] \\ \mathbb{E}[x_1 y] \end{pmatrix}$$

Solution connecting to the single regression

Two results

$$\begin{aligned}
\mathbb{E}[(x_1 - \mu_x)^2] &= \mathbb{E}[x_1^2 - 2x_1\mu_x + \mu_x^2] \\
&= \mathbb{E}[x_1^2] - 2\mu_x\mathbb{E}[x_1] + \mathbb{E}[\mu_x^2] \\
&= \mathbb{E}[x_1^2] - (\mathbb{E}[x_1])^2
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[(x_1 - \mu_x)(y - \mu_y)] &= \mathbb{E}[x_1 y - x_1\mu_y - \mu_x y + \mu_x\mu_y] \\
&= \mathbb{E}[x_1 y] - 2\mu_x\mu_y + \mu_x\mu_y \\
&= \mathbb{E}[x_1 y] - \mu_x\mu_y
\end{aligned}$$

Solution connecting to the single regression

$$
\begin{aligned}
\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} &= \left(\mathbb{E}[xx']\right)^{-1} \mathbb{E}[xy] \\
&= \frac{1}{\mathbb{E}[x_1^2] - \left(\mathbb{E}[x_1]\right)^2} \begin{pmatrix} \mathbb{E}[x_1^2], & -\mathbb{E}[x_1] \\ -\mathbb{E}[x_1], & 1 \end{pmatrix} \cdot \begin{pmatrix} \mathbb{E}[y] \\ \mathbb{E}[x_1 y] \end{pmatrix} \\
&= \frac{1}{\mathbb{E}[x_1^2] - \left(\mathbb{E}[x_1]\right)^2} \begin{pmatrix} \mathbb{E}[x_1^2] \cdot \mu_y - \mu_x \cdot \mathbb{E}[x_1 y] \\ \mathbb{E}[x_1 y] - \mu_x \mu_y \end{pmatrix} \\
&= \frac{1}{\mathbb{E}[x_1^2] - \mu_x^2} \begin{pmatrix} \mathbb{E}[x_1^2] \cdot \mu_y \color{red}{- \mu_{\mathbf{x}}^{\mathbf{2}} \cdot \mu_{\mathbf{y}} + \mu_{\mathbf{x}}^{\mathbf{2}} \cdot \mu_{\mathbf{y}}} - \mu_x \cdot \mathbb{E}[x_1 y] \\ \mathbb{E}[x_1 y] - \mu_x \mu_y \end{pmatrix} \\
&= \frac{1}{\mathbb{E}[x_1^2] - \mu_x^2} \begin{pmatrix} \left(\mathbb{E}[x_1^2] - \mu_x^2\right) \cdot \mu_y - \mu_x \cdot \left(\mathbb{E}[x_1 y] - \mu_x \mu_y\right) \\ \mathbb{E}[x_1 y] - \mu_x \mu_y \end{pmatrix} \\
&= \begin{pmatrix} \mu_y - \mu_x \cdot \frac{\mathbb{E}[(y - \mu_y)(x_1 - \mu_x)]}{\mathbb{E}[(x_1 - \mu_x)^2]} \\ \frac{\mathbb{E}[(y - \mu_y)(x_1 - \mu_x)]}{\mathbb{E}[(x_1 - \mu_x)^2]} \end{pmatrix}
\end{aligned}
$$

Estimation: MOM

Recall that the regression coefficients:

$$\beta_1^* = \frac{\mathbb{E}[(y-\mu_y)(x-\mu_x)]}{\mathbb{E}[(x-\mu)^2]} = \frac{cov(y,x)}{var(x)}$$

$$\beta_0^* = \mu_y - \beta_1^* \cdot \mu_x$$

Sample moments:

$$\widehat{\mu}_y = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$\widehat{\mu}_x = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\widehat{cov}(y,x) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{\mu}_y)\cdot(x_i - \widehat{\mu}_x)$$

$$\widehat{var}(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \widehat{\mu}_x)^2$$

Estimation: MOM

the regression coefficients:

$$\beta_1^* = \frac{\mathbb{E}[(y - \mu_y)(x - \mu_x)]}{\mathbb{E}[(x - \mu)^2]} = \frac{cov(y, x)}{var(x)}$$
$$\beta_0^* = \mu_y - \beta_1^* \cdot \mu_x$$

The MOM estimator:

$$\widehat{\beta}_1^* = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{\mu}_y) \cdot (x_i - \widehat{\mu}_x)}{\frac{1}{N}\sum_{i=1}^{N}(x_{1i} - \widehat{\mu}_x)^2} = \frac{\sum_{i=1}^{N}(y_i - \widehat{\mu}_y) \cdot (x_{1i} - \widehat{\mu}_x)}{\sum_{i=1}^{N}(x_{1i} - \widehat{\mu}_x)^2}$$
$$\widehat{\beta}_0^* = \widehat{\mu}_y - \widehat{\beta}_1^* \cdot \widehat{\mu}_x$$

Estimation (2): Ordinary Least Squares (OLS) Estimator

Estimation: Ordinary Least Squares (OLS)

Instead of solving the mean (or the expectation of) squared error first for a closed-form population coefficient and then find the sample-moment version of it, we can take a different route by solving the sample-moment version of the **objective function** – mean squared error!

Regression Coefficient One criterion is to choose the values that minimize the mean squared error

$$(\beta_0, \beta_1) = \arg\min \mathbb{E}[(y - \beta_0 - \beta_1 \cdot x_1)^2]$$

OLS Coefficient One criterion is to choose the values that minimize the **sample** mean squared error

$$(\widehat{\beta}_0^{OLS}, \widehat{\beta}_1^{OLS}) = \arg\min \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 \cdot x_{i1})^2$$

Estimation: OLS

The sample-moment version of the objective function can be thought of as

$$\begin{aligned}\widehat{S}(\beta_0, \beta_1) &= \frac{1}{N}\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 \cdot x_{1i})^2 \\ &= \frac{1}{N} SSE(\beta_0, \beta_1)\end{aligned}$$

where $SSE(\beta_0, \beta_1) = \sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 \cdot x_{i1})^2$ is called **sum-of-squared-errors** function.

Estimation: OLS

The OLS coefficients that solve the minimization problem are equivalent to setting the partial derivatives to zeros

$$\beta_0^{OLS} = \bar{y} - \beta_1^{OLS} \cdot \bar{x}_1$$

$$\beta_1^{OLS} = \frac{\sum_{i=1}^{N}(y_i - \bar{y}) \cdot (x_{1i} - \bar{x})}{\sum_{i=1}^{N}(x_{1i} - \bar{x})^2}$$

MOM and OLS Estimators are simply the same!

$$\frac{\partial}{\partial\beta_0} SSE(\beta_0, \beta_1) = 0$$

$$\frac{\partial}{\partial\beta_0} \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 \cdot x_{1i})^2 = 0$$

$$-2 \cdot \sum_{i=1}^{N} \left(y_i - \beta_0^{OLS} - \beta_1^{OLS} \cdot x_{1i}\right) = 0$$

$$\beta_0^{OLS} = \frac{1}{N} \sum_{i=1}^{N} y_i - \beta_1^{OLS} \cdot \frac{1}{N} \sum_{i=1}^{N} x_{1i}$$

$$\beta_0^{OLS} = \bar{y} - \beta_1^{OLS} \cdot \bar{x}_1$$

Estimation: OLS

$$
\begin{aligned}
\frac{\partial}{\partial \beta_1} SSE(\beta_0, \beta_1) &= 0 \\
\frac{\partial}{\partial \beta_0} \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 \cdot x_{1i})^2 &= 0 \\
-2 \cdot \sum_{i=1}^{N} \left(y_i - (\bar{y} - \beta_1^{OLS} \cdot \bar{x}_1) - \beta_1^{OLS} \cdot x_{1i} \right) &= 0 \\
\vdots &= \vdots \\
\beta_1^{OLS} &= \frac{\sum_{i=1}^{N} (y_i - \bar{y}) \cdot (x_{1i} - \bar{x})}{\sum_{i=1}^{N} (x_{1i} - \bar{x})^2}
\end{aligned}
$$

Estimation (2): Alternative Proof of OLS Estimator (Skip)

Alternative Proof (Using the trace trick)

We will skip this proof, but I supply it here for your references.

Math Preliminaries

For an $N \times N$ square matrix A, the trace of A is defined to be the sum of its diagonal elements:

$$tr(A) = \sum_{i=1}^{N} a_{ii}$$

where

$$A = \begin{pmatrix} a_{11}, & a_{12}, & \dots, & a_{1N} \\ a_{21}, & a_{22}, & \dots, & a_{2N} \\ \vdots, & \vdots, & \ddots, & \vdots \\ a_{N1}, & \dots, & \dots, & a_{NN} \end{pmatrix}$$

Alternative Proof (Math Preliminaries)

facts

1. $tr(a) = tr(a') = a$ for a constant a
2. $tr(A) = tr(A')$
3. $tr(A + B) = tr(B + A) = tr(A) + tr(B)$
4. $tr(a \cdot A) = a \cdot tr(A)$
5. $tr(ABC) = tr(CAB) = tr(BCA)$ (**cyclic permutation property**)
6. $tr(ABCD) = tr(DABC) = tr(CDAB) = tr(BCDA)$

Alternative Proof (Math Preliminaries)

trace trick (derived from cyclic permutation property). We use it to reorder the scalar inner product $x'Ax$ as follows

$$x'Ax = tr(x'Ax) = tr(xx'A) = tr(Axx')$$

Alternative Proof (Math Preliminaries)

For a function $f : R^{m \times n} \mapsto R$ mapping from $m \times n$ matrices to the real numbers, we define the derivative of f with respect to A to be:

$$\frac{\partial}{\partial A} f(A) = \begin{bmatrix} \frac{\partial}{\partial a_{11}} f(A), & \cdots, & \frac{\partial}{\partial a_{1n}} f(A) \\ \frac{\partial}{\partial a_{21}} f(A), & \cdots, & \frac{\partial}{\partial a_{2n}} f(A) \\ \vdots, & \ddots, & \vdots \\ \frac{\partial}{\partial a_{m1}} f(A), & \cdots, & \frac{\partial}{\partial a_{mn}} f(A) \end{bmatrix}$$

Alternative Proof (Math Preliminaries)

facts

1. $\frac{\partial}{\partial A'} f(A) = (\frac{\partial}{\partial A} f(A))'$
2. $\frac{\partial}{\partial A} |A| = |A| \cdot (A^{-1})'$ (e.g., the variance-covariance for the multivariate normal distribution)
3. $\frac{\partial}{\partial A} \log |A| = (A^{-1})'$ (e.g., use in MLE)

Alternative Proof (Math Preliminaries)

Example

Suppose that $A = \begin{bmatrix} a_{11}, & a_{12} \\ a_{21}, & a_{22} \end{bmatrix}$ and the function is given by

$$f(A) = \frac{3}{2} a_{11} + 5a_{12}^2 + a_{21} \cdot a_{22}$$

Then, we have

$$\frac{\partial}{\partial A} f(A) = \begin{bmatrix} \frac{3}{2}, & 10a_{12} \\ a_{22}, & a_{21} \end{bmatrix}$$

Alternative Proof (Math Preliminaries)

facts

1. $\frac{\partial}{\partial A} tr(AB) = \frac{\partial}{\partial A} tr(BA) = B'$
2. $\frac{\partial}{\partial A} tr(ABA'C) = CAB + C'AB'$

Alternative Proof

Recall that our goal is to find β that sets the first order conditions to zero. First note that the inner product $\epsilon'\epsilon$ is a constant

$$\begin{aligned}
\frac{\partial}{\partial\beta}\epsilon'\epsilon &= \frac{\partial}{\partial\beta}tr(\epsilon'\epsilon) = \frac{\partial}{\partial\beta}tr\left((Y - X\beta)'(Y - X\beta)\right) \\
&= \frac{\partial}{\partial\beta}tr\left(Y'Y - 2Y'X\beta + \beta'X'X\beta\right) \\
&= -2\cdot\frac{\partial}{\partial\beta}tr(Y'X\beta) + \frac{\partial}{\partial\beta}tr(\beta'X'X\beta) \\
&= -2\cdot\frac{\partial}{\partial\beta}tr(Y'X\beta) + \frac{\partial}{\partial\beta}tr(\beta\cdot I\cdot\beta'X'X) \quad [I : \text{Identity Matrix}] \\
&= -2(Y'X)' + (X'X)\cdot\beta\cdot I + (X'X)'\cdot\beta\cdot I' \\
&= -2(Y'X)' + 2(X'X)\beta \\
&= 0 \implies \beta = (X'X)^{-1}X'Y
\end{aligned}$$