Session 4
$\mathbb{Q}$uantitative Analysis of Financial Markets
Model k and OLS

Christopher Ting

http://www.mysmu.edu/faculty/christophert/

Christopher Ting

✉: christophert@smu.edu.sg

☎: 6828 0364

LKCSB 5036

October 29, 2018

Broad Lesson Plan

What's the Variance of Sample Mean?

- The mean estimate $\widehat{\mu}$ is a random variable as a result of **random sampling**.
- Assume that X_i is i.i.d. for each i, $i = 1, 2, \ldots n$. With $\mathbb{C}\left(X_i, X_j\right) = 0$ for $i \neq j$,

$$\mathbb{E}\left(\left(\widehat{\mu} - \mu\right)^2\right) \equiv \mathbb{V}\left(\widehat{\mu}\right) = \mathbb{V}\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = \sum_{i=1}^{n} \mathbb{V}\left(\frac{X_i}{n}\right)$$
$$= \sum_{i=1}^{n} \frac{\mathbb{V}\left(X_i\right)}{n^2} = \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{V}\left(X_i\right) = \frac{1}{n^2} \cdot n\sigma^2$$
$$= \frac{\sigma^2}{n}.$$

- **What is the implication of this result?**

You think you know basic statistics?

- What is the information the following estimator extracts from the sample $\{x_i\}_{i=1}^{n}$?

$$\widehat{\beth} := \frac{1}{n(n-1)} \sum_{i<j} (x_i - x_j)^2$$

- Sample mean: $\widehat{\mu} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$.

Answer:

$$\sum_{i<j}(x_i - x_j)^2 = \sum_{j=2}^{n}\sum_{i=1;i<j}^{n-1}(x_i - x_j)^2 = \sum_{j=2}^{n}\sum_{i=1;i<j}^{n-1}\left(x_i^2 + x_j^2 - 2x_ix_j\right)$$

$$=(n-1)\sum_{i=1}^{n}x_i^2 - 2\sum_{j=2}^{n}\sum_{i=1;i<j}^{n-1}x_ix_j$$

$$=n\sum_{i=1}^{n}x_i^2 - \sum_{i=1}^{n}x_i^2 - \sum_{j=1}^{n}\sum_{i=1;i\neq j}^{n}x_ix_j = n\sum_{i=1}^{n}x_i^2 - \sum_{j=1}^{n}\sum_{i,j=1}^{n}x_ix_j$$

$$=n\sum_{i=1}^{n}x_i^2 - \sum_{i=1}^{n}x_i\sum_{j=1}^{n}x_j = n\sum_{i=1}^{n}x_i^2 - n^2\widehat{\mu}^2$$

Therefore

$$\widehat{\beth} = \frac{1}{n(n-1)}\left(n\sum_{i=1}^{n}x_i^2 - n^2\widehat{\mu}^2\right) = \frac{\displaystyle\sum_{i=1}^{n}x_i^2 - n\widehat{\mu}^2}{n-1}$$

More on Covariance

Proposition 1

Let X_i and Y_j be random variables. Then

$$\mathbb{C}\left(X_i, Y_j\right) = 0, if i \neq j.$$

Proof:

- Suppose Y and X are related by a mapping, i.e., $Y_j = f(X_j)$.
- The mapping involves the paired copies because each Y_j is independent of Y_i due to random sampling.
- Otherwise, if $Y_j = f(X_i, X_j)$, then Y_j may depend on Y_i indirectly since $Y_i = f(X_h, X_i)$. □

An Estimator of Covariance

Definition 1: Sample Covariance

Given the paired data, $(x_i, y_i), i = 1, 2, \ldots, n$, the sample covariance is defined as

$$\widehat{\sigma}_{XY} := \frac{1}{n-1}\sum_{i=1}^{n}\big(x_i - \widehat{\mu}_X\big)\big(y_i - \widehat{\mu}_Y\big).$$

Show that

1. $\displaystyle\sum_{i=1}^{n}\big(X_i - \widehat{\mu}_X\big)\big(Y_i - \widehat{\mu}_Y\big) = \sum_{i=1}^{n} X_iY_i - n\overline{X}_n\overline{Y}_n.$
2. Use these results to show that the sample covariance is unbiased.

A Problem about Mean Squared Errors

Problem

Prove that, if one predicts a random variable y by a constant a, the constant which gives the best mean squared errors MSE is $a = \mathbb{E}[y]$, and the best MSE one can get is $\mathbb{V}[y]$.

Answer

$\mathbb{E}[(y-a)^2] = \mathbb{E}[y^2] - 2a\,\mathbb{E}[y] + a^2$. Differentiate with respect to a and set zero to get $a = \mathbb{E}[y]$. □

Assumptions

- A total of n sets of observations (x_t, y_t) are given. The observation noise or error term is denoted by $\boldsymbol{\varepsilon} = \begin{pmatrix}\varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_n\end{pmatrix}'$.

  (A1) $\mathbb{E}[\varepsilon_t] = 0$ for every t

  (A2) $\mathbb{E}[\varepsilon_t^2] = \sigma_\varepsilon^2$ for every t

  (A3) $\mathbb{E}[\varepsilon_t \, \varepsilon_u] = 0$ for every $t \neq u$

  (A4) x_t is non-stochastic “constants”.

- A compact way of representing these assumptions is
  - ★ $\mathbb{E}[\boldsymbol{\varepsilon}] = \boldsymbol{o} := \begin{pmatrix}0 & 0 & \cdots & 0\end{pmatrix}'$ (the null vector)
  - ★ $\mathbb{V}[\boldsymbol{\varepsilon}] = \sigma_\varepsilon^2 \boldsymbol{I}$, where $\boldsymbol{I}$ is the n by n identity matrix.

Model 0: Sample Average

- Model 0 is the simplest estimation problem with n independent observations $y_1, \ldots, y_n$ from the same distribution, for which mean μ and variance σ_ε^2 exist.

- Define $\varepsilon_i = y_i - \mu$, and the vectors $\boldsymbol{y} := \begin{pmatrix} y_1 & y_2 & \cdots & y_n \end{pmatrix}'$ and $\boldsymbol{\iota} = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}'$,

- Then Model 0 in the vector form is

$$\boldsymbol{y} = \boldsymbol{\iota}\mu + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma_\varepsilon^2 \boldsymbol{I}) \tag{1}$$

- μ is the deterministic part of y_i, and ε_i is the random part.

Model 1: Simple Regression

- Each element of $\boldsymbol{y}$ is a constant α plus a constant multiple of the corresponding element of the nonrandom vector $\boldsymbol{x}$ plus a random error term ε_t:

$$y_t = \alpha + x_t\beta + \varepsilon_t, \qquad t = 1, \ldots, n.$$

- In vector-matrix form, Model 1 is written as

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \alpha + \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{2}$$

or

$$\boxed{\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma_\varepsilon^2 \boldsymbol{I})} \tag{3}$$

Model M: Multiple Linear Regression

- If y depends on multiple variables, then we have **multiple linear regression**, Model k.
- Mathematically, multiple linear regression has the same form (3), except $\boldsymbol{X}$ now it has more than two columns.
- The additional assumption is that all columns are linearly independent.
- Model M has Models 0 and 1 as special cases.
- Summary

	observed	unobserved
random	$\boldsymbol{y}$	$\boldsymbol{\varepsilon}$
nonrandom	$\boldsymbol{X}$	$\boldsymbol{\beta}$, σ_{ε}^2

Interchangeability of Inverse and Transpose

Proposition

For a square matrix $\boldsymbol{X}$ that has an inverse, the order of transpose and inverse operations is interchangeable.

$$\left(\boldsymbol{X}'\right)^{-1} = \left(\boldsymbol{X}^{-1}\right)'.$$

Proof: By the definition of inverse,

$$\boldsymbol{X}'\left(\boldsymbol{X}'\right)^{-1} = \boldsymbol{I}.$$

Transpose both sides to yield

$$\left(\left(\boldsymbol{X}'\right)^{-1}\right)'\boldsymbol{X} = \boldsymbol{I}.$$

Multiply both sides by $\boldsymbol{X}^{-1}$ from the right to yield

$$\left(\left(\boldsymbol{X}'\right)^{-1}\right)' = \boldsymbol{X}^{-1}.$$

Transpose both sides again to yield $\left(\boldsymbol{X}'\right)^{-1} = \left(\boldsymbol{X}^{-1}\right)'$ as required. □

Residual as Projection

Problem

Assume that $\boldsymbol{X}$ has full column rank. Show that $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M}\boldsymbol{y}$ where $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Show that $\boldsymbol{M}$ is symmetric and idempotent.

Answer

- $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \left(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)\boldsymbol{y}$.
- To prove idempotent, i.e. $\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$, we compute

$$\begin{aligned}
\boldsymbol{M}\boldsymbol{M} &= \left(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)\left(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)\\
&= \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\\
&\qquad + \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\\
&= \boldsymbol{I} - 2\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\\
&= \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{M}.
\end{aligned}$$

- To show symmetric, $\boldsymbol{M}' = \boldsymbol{I}' - \left(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right) = \boldsymbol{M}$.

Matrix Calculus

- Let x be a column k-vector. Consider the function

$$g(\boldsymbol{x}) = g(x_1, x_2, \ldots, x_k) : \Re^k \longrightarrow \Re.$$

- The vector derivative is

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

and

$$\frac{\partial}{\partial \boldsymbol{x}'} g(\boldsymbol{x}) = \begin{pmatrix} \dfrac{\partial}{\partial x_1} g(\boldsymbol{x}) & \dfrac{\partial}{\partial x_2} g(\boldsymbol{x}) & \cdots & \dfrac{\partial}{\partial x_k} g(\boldsymbol{x}) \end{pmatrix}$$

Matrix Calculus: Basic Properties

For constant vector $\boldsymbol{a}$ and matrix $\boldsymbol{A}$,

$$\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{a}'\boldsymbol{x}) = \frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x}'\boldsymbol{a}) = \boldsymbol{a}, \quad \text{or} \quad \frac{\partial}{\partial \boldsymbol{x}'}(\boldsymbol{a}'\boldsymbol{x}) = \boldsymbol{a}'$$

$$\frac{\partial}{\partial \boldsymbol{x}'}(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}$$

$$\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}) = (\boldsymbol{A} + \boldsymbol{A}')\boldsymbol{x}, \quad \text{or} \quad \frac{\partial}{\partial \boldsymbol{x}'}(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{x}'(\boldsymbol{A} + \boldsymbol{A}')$$

$$\frac{\partial^2}{\partial \boldsymbol{x}\partial \boldsymbol{x}'}(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A} + \boldsymbol{A}'$$

Sum of Squared Errors (SSE) and OLS

- In the model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim (\boldsymbol{o}, \sigma_\varepsilon^2 \boldsymbol{I})$, the OLS-estimate $\hat{\boldsymbol{\beta}}$ is defined to be that value $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ that minimizes

$$\mathrm{SSE} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}. \tag{4}$$

- Note that $\boldsymbol{X}'\boldsymbol{X}$ is a square symmetric matrix. Hence, the first-order condition is

$$\frac{\partial \mathrm{SSE}}{\partial \boldsymbol{\beta}'} = -2\boldsymbol{y}'\boldsymbol{X} - 2\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{0}.$$

- Applying the transpose operation, we obtain $\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$
- Due to our assumption that all columns of $\boldsymbol{X}$ are linearly independent, $\boldsymbol{X}'\boldsymbol{X}$ has an inverse and one can premultiply both sides by $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ to obtain

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Reduction to Model 0: Sample Average

- Sample average is "regression" with an intercept only, i.e.,

$$\boxed{y_t = \alpha + \varepsilon_t.} \tag{5}$$

- To see how Model 0 comes from Model k, suppose $\boldsymbol{X} = \boldsymbol{\iota}$.
- Therefore $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{\iota}'\boldsymbol{\iota} = n$, and $(\boldsymbol{X}'\boldsymbol{X})^{-1} = \dfrac{1}{n}$.
- Consequently,

$$\widehat{\alpha} = \frac{1}{n}\boldsymbol{\iota}'\boldsymbol{y} = \frac{1}{n}\sum_{t=1}^{n} y_t =: \bar{y}.$$

Reduction to Model 1: Simple Regression

- Simple regression is regression with an intercept and one explanatory variable only, i.e.,

$$\boxed{y_t = \alpha + \beta x_t + \varepsilon_t.} \tag{6}$$

- Here $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{\iota} & \boldsymbol{x} \end{pmatrix}$ and $\boldsymbol{\beta} = \begin{pmatrix} \alpha & \beta \end{pmatrix}'$.

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} \boldsymbol{\iota}' \\ \boldsymbol{x}' \end{pmatrix} \begin{pmatrix} \boldsymbol{\iota} & \boldsymbol{x} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\iota}'\boldsymbol{\iota} & \boldsymbol{\iota}'\boldsymbol{x} \\ \boldsymbol{x}'\boldsymbol{\iota} & \boldsymbol{x}'\boldsymbol{x} \end{pmatrix} = \begin{pmatrix} n & \sum x_t \\ \sum x_t & \sum x_t^2 \end{pmatrix}$$

$$\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} = \frac{1}{n\sum x_t^2 - (\sum x_t)^2} \begin{pmatrix} \sum x_t^2 & -\sum x_t \\ -\sum x_t & n \end{pmatrix}$$

$$\boldsymbol{X}'\boldsymbol{y} = \begin{pmatrix} \boldsymbol{\iota}'\boldsymbol{y} \\ \boldsymbol{x}'\boldsymbol{y} \end{pmatrix} = \begin{pmatrix} \sum y_t \\ \sum x_t y_t \end{pmatrix}$$

OLS Estimates for Simple Regression

- Therefore $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ gives $\hat{\boldsymbol{\beta}} = \begin{bmatrix}\hat{\alpha} & \hat{\beta}\end{bmatrix}'$:

$$\hat{\alpha} = \frac{\sum x_t^2 \sum y_t - \sum x_t \sum x_t y_t}{n\sum x_t^2 - (\sum x_t)^2} \tag{7}$$

$$\hat{\beta} = \frac{n\sum x_t y_t - \sum x_t \sum y_t}{n\sum x_t^2 - (\sum x_t)^2} \tag{8}$$

Formula for the Slope Estimate

- Since $\sum x_i = n\bar{x}$ and $\sum y_i = n\bar{y}$, we have

$$\hat{\beta} = \frac{n\sum x_t y_t - n^2 \bar{x}\bar{y}}{n\sum x_t^2 - n^2\bar{x}^2} = \frac{\sum x_t y_t - n\bar{x}\bar{y}}{\sum x_t^2 - n\bar{x}^2}. \tag{9}$$

- Now, note that

$$\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y}) = \sum_{t=1}^{n} x_t y_t - n\bar{x}\bar{y}. \tag{10}$$

- Hence

$$\boxed{\hat{\beta} = \frac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}.} \tag{11}$$

Formula for the Intercept Estimate

- For simple regression $y_t = \alpha + \beta x_t + \varepsilon_t$, or $\varepsilon_t = y_t - \alpha - \beta x_t$, the sum of squared errors can also be written as

$$\mathrm{SSE} = \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} = \sum_{t=1}^{n} \varepsilon_t^2 = \sum_{t=1}^{n} \left(y_t - \alpha - \beta x_t\right)^2.$$

- The first-order condition for the intercept α is

$$\frac{\partial\,\mathrm{SSE}}{\partial \alpha} = -2 \sum_{t=1}^{n} \left(y_t - \alpha - \beta x_t\right) = 0.$$

- The solution is

$$\sum_{t=1}^{n} \hat{\alpha} = \sum_{t=1}^{n} y_t - \hat{\beta} \sum_{t=1}^{n} x_t = n\bar{y} - n\hat{\beta}\bar{x},$$

which is

$$\boxed{\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}.} \tag{12}$$

True Parameter Values of Simple Regression

- If y is predicted by an affine expression of the form $a + bx$, then the lowest mean squared error MSE is obtained. Moreover,

$$b = \frac{\mathbb{C}[x, y]}{\mathbb{V}[x]}, \qquad a = \mathbb{E}[y] - b\,\mathbb{E}[x].$$

- The MSE is variance plus squared bias, therefore

$$\begin{aligned} \mathrm{MSE}[a + bx - y] &= \mathbb{V}[a + bx - y] + \left(\mathbb{E}[a + bx - y]\right)^2 \\ &= \mathbb{V}[bx - y] + \left(a - \mathbb{E}[y] + b\,\mathbb{E}[x]\right)^2. \end{aligned} \tag{13}$$

- Choose a so that the second term is zero, and then minimize the first term with respect to b. Since

$$\mathbb{V}[bx - y] = b^2\,\mathbb{V}[x] - 2b\,\mathbb{C}[x, y] + \mathbb{V}[y], \tag{14}$$

the first order condition of equation (14) results in b.

Mean Squared Errors

- Plugging the optimal a into (13), annuls the last term of (13) so that the MSE is given by (14).
- If one inserts the optimal $b = \dfrac{\mathbb{C}[x, y]}{\mathbb{V}[x]}$ into (14), then

$$\mathrm{MSE} = \left(\frac{\mathbb{C}[x, y]}{\mathbb{V}[x]}\right)^2 \mathbb{V}[x] - 2\frac{\mathbb{C}[x, y]}{\mathbb{V}[x]}\,\mathbb{C}[x, y] + \mathbb{V}[y] \tag{15}$$

$$= \mathbb{V}[y] - \frac{(\mathbb{C}[x, y])^2}{\mathbb{V}[x]}. \tag{16}$$

- Since $y = a + bx + e$ where e is the error term, we have

$$\mathbb{C}[x, y - a - bx] = \mathbb{C}[x, y] - b\,\mathbb{C}[x, x] = 0.$$

- Therefore, the prediction error is uncorrelated with the observed x.

In-Class Exercise

Question in Class

The affine form is $y = a + bx + e$. Show that the mean square error, equation (16), is the variance of the error e, i.e.,

$$\sigma_e{}^2 := \mathbb{V}\left[e\right] = \mathrm{MSE}\,.$$

Insight!

- The first-order conditions can be written as

$$\boldsymbol{X}'\left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)=\mathbf{0},$$

which is

$$\boldsymbol{X}'\widehat{\boldsymbol{u}}=\mathbf{0}. \tag{17}$$

The OLS residuals are orthogonal to $\boldsymbol{X}'$.

- Consequently, if $\boldsymbol{X}$ has a column vector of ones, then the average of the residuals is zero. This is because there is one row of ones in $\boldsymbol{X}'$ and hence

$$\mathbf{1}'\,\widehat{\boldsymbol{u}}=\sum_{i=1}^{n}\widehat{u}_i=0.$$

OLS Algorithm for Model M

1. Estimate the model by OLS: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$
2. Compute the fitted values of $\boldsymbol{y}$: $\widehat{\boldsymbol{y}}\,\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$
3. Compute the residuals or “surprise”: $\widehat{\boldsymbol{u}} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$
4. Compute the residual sum of squares (RSS)

$$\mathrm{SSE} \equiv \text{RSS} = \widehat{\boldsymbol{u}}'\widehat{\boldsymbol{u}} = \sum_{i=1}^{n} \widehat{u_i}^2$$

5. The variance of the residuals is

$$\widehat{\sigma}_u^2 = \frac{1}{n-K}\widehat{\boldsymbol{u}}'\widehat{\boldsymbol{u}}$$

6. Let $\boldsymbol{\Omega} := \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$. The variance of $\widehat{\beta}_i$ is

$$\mathbb{V}(\widehat{\beta}_i) = \widehat{\sigma}_u^2 \Omega_{ii}.$$

Summary of OLS Regression

- The estimates are unbiased, i.e., $\mathbb{E}\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}$.

- The variance of the residuals is unbiased $\mathbb{E}\left(\widehat{\sigma}_u^2\right) = \sigma_u^2$.

- Efficiency: According to the Gauss-Markov theorem, among the classical linear regression models, the OLS estimator is the linear unbiased estimator of $\boldsymbol{\beta}$ with the minimum variance.

- Conditional normality

$$\widehat{\boldsymbol{\beta}}\big|\boldsymbol{X} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma_u^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right) \tag{18}$$

Statistical Inference

- For all $j = 1, 2, \ldots, K$, the t test statistic for $\widehat{\beta}_j$ is, given the null-hypothesized value β_j:

$$\frac{\widehat{\beta}_j - \beta_j}{\widehat{\sigma}_u \sqrt{\Omega_{jj}}} \sim t_{n-K} \tag{19}$$

- The α% significance level for β_j is, assuming two-tail test,

$$\widehat{\beta}_j - q\widehat{\sigma}_u \sqrt{\Omega_{jj}} \leq \beta_j \leq \widehat{\beta}_j + q\widehat{\sigma}_u \sqrt{\Omega_{jj}},$$

where q is the $\big(1 - \alpha/2\big)$-th quantile of the t_{n-K} distribution.

Confidence Interval for Mean Response

- A given observation $\boldsymbol{x}$, which is a $K \times 1$ vector, i.e., $\boldsymbol{x} := \begin{bmatrix} 1 & x_2 & x_3 & \cdots & x_K \end{bmatrix}'$
- The **in-sample** mean response is $\widehat{\boldsymbol{\beta}}'\boldsymbol{x}$.
- Given the unbiased estimate $\widehat{\boldsymbol{\beta}}$, the variance of $\widehat{\boldsymbol{\beta}}'\boldsymbol{x}$ is

$$\mathbb{V}\left(\widehat{\boldsymbol{\beta}}'\boldsymbol{x}\right) = \mathbb{V}\left(\boldsymbol{x}'\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{x}'\mathbb{V}\left(\widehat{\boldsymbol{\beta}}\right)\boldsymbol{x} = \sigma_u^2\,\boldsymbol{x}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}. \tag{20}$$

- Hence, a $100 \times (1-\alpha)$% confidence interval for the in-sample mean response $\widehat{\boldsymbol{\beta}}'\boldsymbol{x}$ is, with $q := t_{N-K,1-\alpha/2}$,

$$\widehat{\boldsymbol{\beta}}'\boldsymbol{x} \pm q\sigma_u\sqrt{\boldsymbol{x}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}} \tag{21}$$

Prediction Interval for a New Observation

- Suppose a future observation of $\boldsymbol{x}$ is obtained. Then, by assumption of $u \sim \mathcal{N}(0, \sigma_u^2)$, we have

$$\mathbb{V}\left(y - \widehat{\boldsymbol{\beta}}'\boldsymbol{x}\right) = \mathbb{V}(u) + \mathbb{V}\left(\widehat{\boldsymbol{\beta}}'\boldsymbol{x}\right) \tag{22}$$

- Hence, a $100 \times (1 - \alpha)\%$ prediction interval for y is

$$\widehat{\boldsymbol{\beta}}'\boldsymbol{x} \pm q\sigma_u \sqrt{1 + \boldsymbol{x}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}}. \tag{23}$$

Akaike Information Criterion (1973)

- For i.i.d. normally distributed errors,

$$\text{AIC} = T \ln\left(\frac{\text{RSS}}{T}\right) + 2K.$$

- For small sample sizes ($T/K \leq 40$), use the second-order AIC:

$$\text{AIC}_c = \text{AIC} + \frac{2K(K+1)}{T-K-1}$$

- The smaller AIC is, the better is the model in not over-fitting the data.

Multiple Linear Regression and the Constant Term

- Model M: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$
  For $t = 1, 2, \ldots, T$,

$$y_t = \beta_1 + \beta_2 X_{2,t} + \beta_3 X_{3,t} + ... + \beta_K X_{K,t} + u_t,$$

  There are K parameters, $\beta_1, \beta_2, \ldots, \beta_K$.

- The first parameter is the y-intercept in Model 1. $X_{1,t} = 1$ being a constant for all t.

$$\boldsymbol{X}_1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

Model 1: Simple Linear Regression

- If $K = 2$, we are back to Model 1

$$\underset{T\times 1}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}} = \underset{T\times 2}{\begin{bmatrix} 1 & x_{2,1} \\ 1 & x_{2,2} \\ \vdots & \vdots \\ 1 & x_{2,T} \end{bmatrix}} \underset{2\times 1}{\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}} + \underset{T\times 1}{\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{bmatrix}}$$

- Notice that the matrices written in this way are **conformable**.

Numerical Illustration: Data

- When $K = 3$,

$$y_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + u_t$$

for $t = 1, 2, \ldots, 15$.

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 2.0 & 3.5 & -1.0 \\ 3.5 & 1.0 & 6.5 \\ -1.0 & 6.5 & 4.3 \end{bmatrix}, \ \boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} -3.0 \\ 2.2 \\ 0.6 \end{bmatrix}$$

- The residual sum of sqauares (RSS) is $\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}$ = 10.96

Calculations in Simple Regression

- To calculate the coefficients, just multiply the matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ by the vector $\boldsymbol{X}'\boldsymbol{y}$ to obtain $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$

$$\begin{bmatrix} 2.0 & 3.5 & -1.0 \\ 3.5 & 1.0 & 6.5 \\ -1.0 & 6.5 & 4.3 \end{bmatrix} \begin{bmatrix} -3.0 \\ 2.2 \\ 0.6 \end{bmatrix} = \begin{bmatrix} 1.10 \\ 4.40 \\ 19.88 \end{bmatrix}$$

- To calculate the standard errors, we need an estimate of σ_u^2.

$$s^2 = \frac{\text{RSS}}{T-K} = \frac{10.96}{15-3} = 0.91$$

- The variance-covariance matrix of $\hat{\beta}$ is given by

$$s^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = 0.91(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 1.82 & 3.19 & -0.91 \\ 3.19 & 0.91 & 5.92 \\ -0.91 & 5.92 & 3.91 \end{bmatrix}$$

Standard Errors and Estimated Model

- The variances are on the leading diagonal:

$$\begin{aligned} \operatorname{var}(\hat{\beta}_1) &= 1.82 & & \mathrm{SE}(\hat{\beta}_1) = 1.35 \\ \operatorname{var}(\hat{\beta}_2) &= 0.91 & \Longleftrightarrow \quad & \mathrm{SE}(\hat{\beta}_2) = 0.95 \\ \operatorname{var}(\hat{\beta}_3) &= 3.91 & & \mathrm{SE}(\hat{\beta}_3) = 1.98 \end{aligned}$$

- We write

$$\begin{array}{llll} \hat{y} & = 1.10 - 4.40 x_2 & + 19.88 x_3 \\ & \quad (1.35) \; (0.96) & \quad (1.98) \end{array}$$

Ordinary Least Squares (OLS)'s Goodness of Fit

- Explained sum of squares (ESS) is

$$\text{ESS} := \sum_{i=1}^{n} \left(\widehat{y}_i - \overline{y}\right)^2$$

- Total sum of squares (TSS) is, with $\boldsymbol{y}_* := \boldsymbol{y} - \overline{y}\mathbf{1}$,

$$\begin{aligned} \text{TSS} := \boldsymbol{y}_*'\boldsymbol{y}_* = \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2 &= \sum_{i=1}^{n} \left(\widehat{y}_i - \overline{y}\right)^2 + \sum_{i=1}^{n} \widehat{u}_i^2 && (24) \\ &= \qquad \text{ESS} \qquad + \quad \text{RSS}. && (25) \end{aligned}$$

- The cross term $\sum_{i=1}^{n} \left(\widehat{y}_i - \overline{y}\right)\left(y_i - \widehat{y}_i\right)$ is zero because of the orthogonality: $\boldsymbol{X}'\widehat{\boldsymbol{u}} = \mathbf{0}$.

Proof

$$
\begin{aligned}
\sum_{i=1}^{n} \big(\widehat{y}_i - \overline{y}\big)\big(y_i - \widehat{y}_i\big) &= \sum_{i=1}^{n} \big(\widehat{y}_i - \overline{y}\big)\widehat{u}_i \\
&= \sum_{i=1}^{n} \widehat{y}_i \widehat{u}_i \qquad \left(\text{since } \overline{y}\sum_{i=1}^{n} \widehat{u}_i = 0\right) \\
&= \sum_{i=1}^{n} \left(\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)_i \widehat{u}_i \\
&= \left(\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)' \widehat{\boldsymbol{u}} = \widehat{\boldsymbol{\beta}}' \boldsymbol{X}' \widehat{\boldsymbol{u}} = 0.
\end{aligned}
$$

Coefficient of Determination R^2

- In summary,

$$\begin{aligned} \text{TSS} &= \text{ESS} + \text{RSS} \\ \sum_t (y_t - \bar{y})^2 &= \sum_t (\hat{y}_t - \bar{y})^2 + \sum_t \hat{u}_t^2 \end{aligned}$$

- Our goodness of fit statistic is $R^2 = \dfrac{\text{ESS}}{\text{TSS}}$
- But since TSS = ESS + RSS, we can also write

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

- R^2 must always lie between zero and one. To understand this, consider two extremes

$$\text{RSS} = \text{TSS} \quad \text{i.e.} \quad \text{ESS} = 0 \quad \text{so} \quad R^2 = \frac{\text{ESS}}{\text{TSS}} = 0$$

$$\text{ESS} = \text{TSS} \quad \text{i.e.} \quad \text{RSS} = 0 \quad \text{so} \quad R^2 = \frac{\text{ESS}}{\text{TSS}} = 1$$

Problems with R^2 as a Goodness of Fit Measure

1. R^2 is defined in terms of variation about the mean of y so that if a model is reparameterised (rearranged) and the dependent variable changes, R^2 will change.

2. R^2 never falls if more regressors are added. to the regression, e.g. consider:

$$\begin{aligned} \text{Regression1}: y_t &= \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + u_t \\ \text{Regression2}: y_t &= \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + \beta_4 x_{4t,} + u_t \end{aligned}$$

3. R^2 will always be at least as high for regression 2 relative to regression 1.

Solution: Adjusted R^2

- To get around these problems, a modification is made to take into account the **loss of degrees of freedom** associated with adding extra variables. This modification is known as $\overline{R}^2$, or adjusted R^2:

$$\overline{R}^2 = 1 - \left[\frac{T-1}{T-K} \left(1 - R^2\right) \right].$$

- So if we add an extra regressor, K increases and unless R^2 increases by a more than offsetting amount, $\bar{R}^2$ will actually fall.

- But there are still problems with the criterion: No distribution for $\bar{R}^2$ or R^2

R^2 and Adjusted R^2

- The coefficient of determination

$$R^2 := \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{26}$$

- Denoted by $\overline{R}^2$, the adjusted R^2 is based on the unbiased variances:

$$\overline{R}^2 = 1 - \frac{\frac{\text{RSS}}{n-K}}{\frac{\text{TSS}}{n-1}} \tag{27}$$

Takeaways

- Vector-matrix approach is “simpler”.
- Important concepts
  - ★ Variance-covariance matrix of K-vector $\widehat{\boldsymbol{\beta}}$
  - ★ Diagonal elements of the variance-covariance matrix $\Longrightarrow$ variance of $\widehat{\beta}_i$
  - ★ Inference by standard error, t statistic, and p value
  - ★ Confidence interval
  - ★ Adjusted R^2, AIC
- R^2 increases with K, which is the total number of parameter estimates
- Important to select the model with minimum AIC value for a given data set.