5. Linear Regression

- linear least-squares/regression
- BLUE property
- distribution of LS estimators
- weighted least-squares

Linear regression

- linear regression is the simplest type of *parametric* model
- it explains a relationship between variables y and β using a linear function:

$$y = X\beta$$

  where $y \in \mathbf{R}^m$, $X \in \mathbf{R}^{m \times n}$, $\beta \in \mathbf{R}^n$

- y contains the measurement variables and is called the *regressed variable* or *regressand*
- each row vector X_k^T in matrix X is called *regressor*
- the matrix X is sometimes called *the design matrix*
- β is the *parameter vector*. Its element β_k is often called *regression coefficients*

Example : A Polynomial trend

assume the model is the form of a polynomial of degree n

$$y(t) = a_0 + a_1 t + \ldots + a_r t^n$$

with unknown coefficients $a_0, \ldots, a_n$

this can be written in the form of linear regression as

$$\begin{bmatrix} y(t_1) \\ y(t_2) \\ \vdots \\ y(t_m) \end{bmatrix} = \begin{bmatrix} 1 & t_1 & \ldots & t_1^n \\ 1 & t_2 & \ldots & t_2^n \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_N & \ldots & t_m^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix}$$

given the measurements $y(t_i)$ for $t_1, t_2, \ldots, t_m$, we want to estimate the coefficents a_k

Solving linear regressions

- the problem is to find an estimate of β from the measurements y and X
- if we choose the number of measurements, m to be equal to n, then x can be solved by
$$\beta = X^{-1}y,$$
provided that X is *invertible*
- in practice, in the presence of noise and disturbance, more data should be collected in order to get a better estimate
- this leads to overdetermined linear equations where an exact solution does not usually exist
- however, it can be solved by **linear least-squares** formulation

Definition of Linear least-squares

Overdetermined linear equations

$$X\beta = y, \quad X \text{ is } m \times n \text{ with } m > n$$

for most y cannot solve for β

Linear least-squares formulation

$$\text{minimize } \|X\beta - y\|_2 = \left(\sum_{i=1}^{m} (\sum_{j=1}^{n} x_{ij}\beta_j - y_i)^2 \right)^{1/2}$$

- $r = X\beta = y$ is called *the residual error*
- β with smallest residual norm $\|r\|$ is called *the least-squares solution*
- equivalent to minimizing $\|X\beta - y\|^2$

Example: Data fitting

fit a function

$$y = g(t) = \beta_1 g_1(t) + \beta_2 g_2(t) + \ldots + \beta_n g_n(t)$$

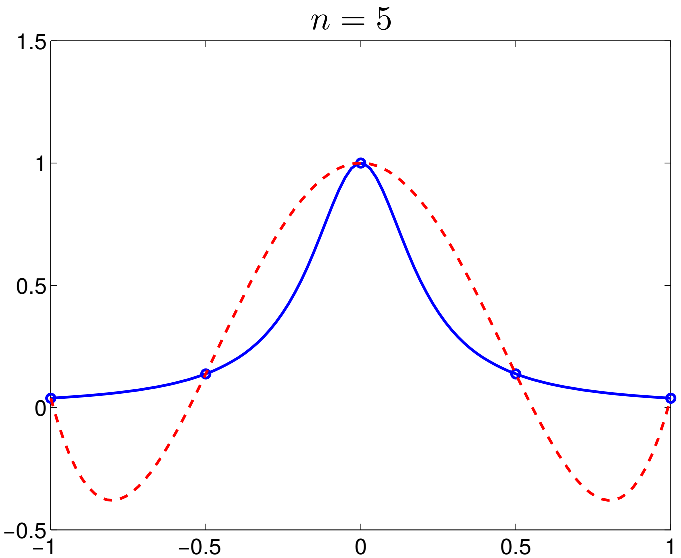
to data (t_1, y_1), (t_2, y_2), . . . , (t_m, y_m), i.e., choose the coefficients β_k so that

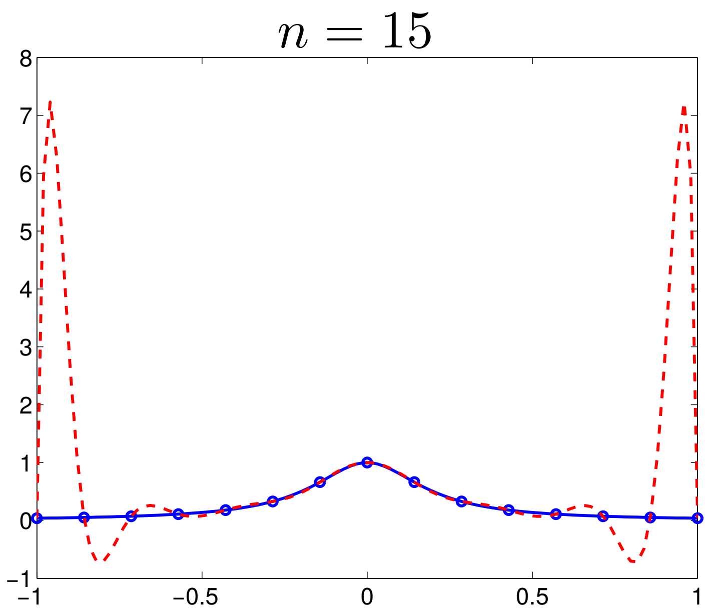
$$g(t_1) \approx y_1, \quad g(t_2) \approx y_2, \quad , g(t_m) \approx y_m$$

- $g_i(t) : \mathbf{R} \to \mathbf{R}$ are given functions (*basis functions*)
- problem variables: the coefficients $\beta_1, \beta_2, \ldots, \beta_n$
- usually $m \gg n$, hence no exact solution with $g(t_i) = y_i$ for all i
- applications: developing simple, approximate model of observed data

Example: fit a polynomial to $f(t) = 1/(1 + 25t^2)$ on $[-1, 1]$

- pick $m = n$ points t_i in $[-1, 1]$ and calculate $y_i = 1/(1 + 25t_i^2)$
- interpolate by solving $X\beta = y$



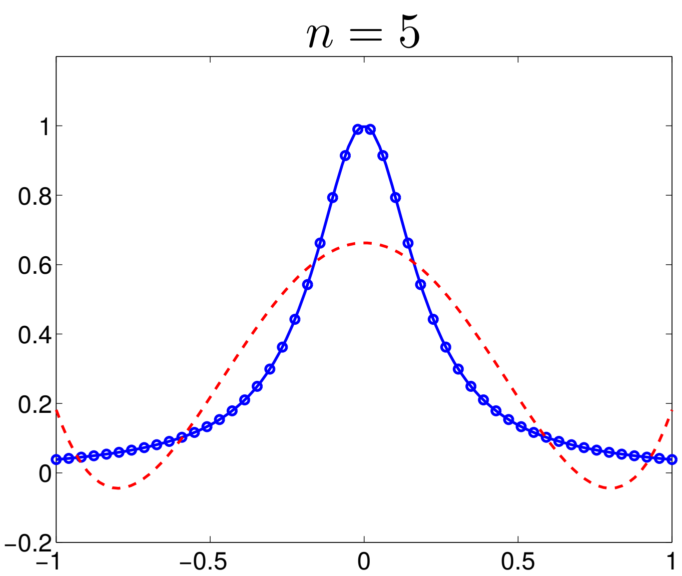


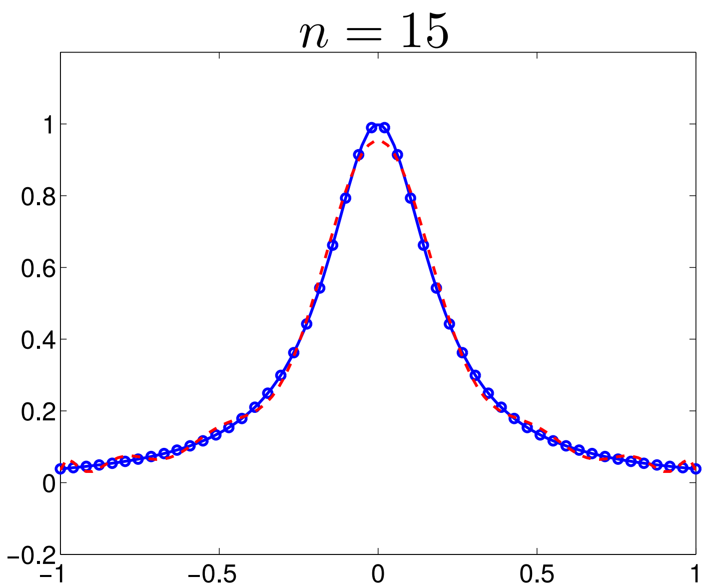
(blue solid line: f; red dashed line: polynomial g)

increase n does not improve the overall quality of the fit

Same example by approximation

- pick $m = 50$ points t_i in $[-1, 1]$
- fit polynomial by minimizing $\|X\beta - y\|$





(blue solid line: f; red dashed line: polynomial g)

much better fit overall

Properties of full rank matrices

suppose X is an $m \times n$ matrix; we always have

$$\mathbf{rank}(X) \leq \min(m, n)$$

if X is **full rank with** $m \geq n$

- $\mathbf{rank}(X) = n$ and $\mathcal{N}(X) = \{0\}$ $(X\beta = 0 \Leftrightarrow \beta = 0)$
- $X^T X$ is positive definite: for any $\beta \neq 0$ then

$$\beta^T X^T X \beta = \|X\beta\|^2 > 0$$

similarly, if X is **full rank** with $m \leq n$

- $\mathbf{rank}(X) = m$ and $\mathcal{N}(X^T) = \{0\}$
- XX^T is positive definite

Normal equations

$$X^T X\beta = X^T y$$

- equivalent to the zero gradient condition:

$$\frac{d}{d\beta}\|X\beta - y\|_2^2 = X^T(X\beta - y) = 0$$

if X has a zero nullspace:

- least-squares solution can be found by solving the normal equations
- n equations in n variables with a positive definite coefficient matrix
- the closed-form solution is $\beta = (X^T X)^{-1} X^T y$
- $(X^T X)^{-1} X^T$ is a left inverse of X

Analysis of LS estimate

- linear regression model in estimation
- analysis of LS estimate
  - LS model with deterministic/fixed regressor
  - LS model with stochastic regressor
- identification
- consistency
- asymptotic ditribution

General regression model

the general regression model with additive errors is given by

$$y = \mathbf{E}[y|X] + u$$

- the data are (y, X) where y is observation and X is a matrix of explanatory variables
- $\mathbf{E}[y|X]$ is considered as a conditional function that gives the average value of y given X
- u is a vector of unknown random errors/noise/disturbances

a linear regression model is obtained when $\mathbf{E}[y|X]$ is linear in X

Linear regression model

a linear regression model is

$$y_i = x_i^T \beta + u, \quad i = 1, 2, \dots, N$$

in matrix notation

$$y = X\beta + u$$

- $X \in \mathbf{R}^{N \times n}$ is regression or sensor matrix
- $y \in \mathbf{R}^N$ is the measurement, also called dependent variable or endogenous variable
- $\beta \in \mathbf{R}^n$ is the parameter vector (to be estimated)
- $u \in \mathbf{R}^N$ is the error vector
- each row vector of X, x_i^T is referred to as regressors/predictors or covariates

common terminology: from the model $y = X\beta + u$

variables y and X are commonly known as

y	X
endogenous variable	exogenous variable
dependent variable	independent variable
explained variable	explanatory variable
response variable	predictor
observable variable	regressor
	covariates
	manipulated variable

Least-squares estimation

from the linear regression model

$$y = X\beta + u$$

the method is to choose an estimate $\hat{\beta}$ that minimizes

$$\|X\hat{\beta} - y\|$$

$i.e.$, minimize the deviation between what we actually observed (y), and what we would observe if $\beta = \hat{\beta}$, and there were no noise ($u = 0$)

the LS estimate is given by

$$\hat{\beta}_{\mathrm{ls}} = (X^T X)^{-1} X^T y$$

provided that X is full rank

Analysis of the LS estimate (static case)

assumptions:

- u is *white noise* with zero mean and covariance matrix Σ
- the least-square estimate is given by

$$\hat{\beta} = \operatorname{argmin} \|X\beta - y\|$$

- the regressor X is *deterministic*

then the following properties hold:

- $\hat{\beta}$ is an unbiased estimate of β ($\mathbf{E}\hat{\beta} = \beta$, or $\hat{\beta} = \beta$ when $u = 0$)
- the covariance matrix of $\hat{\beta}$ is given by

$$\mathbf{cov}(\hat{\beta}) = (X^T X)^{-1} X^T \Sigma X (X^T X)^{-1}$$

short proof: we can write the LS estimate as

$$\hat{\beta} = (X^T X)^{-1} X^T y = (X^T X)^{-1} X^T (X\beta + u) = \beta + (X^T X)^{-1} X^T u$$

- since X is deterministic and u is zero mean, we have $\mathbf{E}\hat{\beta} = \beta$
- the covariance of $\hat{\beta}$ is derived by

$$\mathbf{cov}(\hat{\beta}) = \mathbf{E}[(\hat{\beta} - \mathbf{E}\hat{\beta})(\hat{\beta} - \mathbf{E}\hat{\beta})^T]$$

but $\mathbf{E}\hat{\beta} = \beta$ and that $\hat{\beta} - \mathbf{E}\hat{\beta} = (X^T X)^{-1} X^T u$, hence,

$$\begin{aligned} \mathbf{cov}(\hat{\beta}) &= \mathbf{cov}[(X^T X)^{-1} X^T u] \\ &= (X^T X)^{-1} X^T \mathbf{cov}(u) X (X^T X)^{-1} \\ &= (X^T X)^{-1} X^T \Sigma X (X^T X)^{-1} \end{aligned}$$

if $\Sigma = \sigma^2 I$, then it reduces to $\mathbf{cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$

BLUE property

assumptions: u is white noise with zero mean and **unit** covariance ($\mathbf{cov}(u) = I$)

the estimator defined by

$$\hat{\beta}_{\rm ls} = (X^T X)^{-1} X^T y$$

is the **optimum unbiased linear least-mean-squares** estimator of β

assume $\hat{\beta} = By$ is any other linear estimator of β

- require $BX = I$ in order for $\hat{z}$ to be unbiased
- $\mathbf{cov}(\hat{\beta}) = BB^T$
- $\mathbf{cov}(\hat{\beta}_{\rm ls}) = BX(X^T X)^{-1} X^T B^T \quad$ (apply $BX = I$)

Using $I - X(X^T X)^{-1} X^T \succeq 0$, we conclude that

$$\mathbf{cov}(\hat{\beta}) - \mathbf{cov}(\hat{\beta}_{\rm ls}) = B(I - X(X^T X)^{-1} X^T) B^T \succeq 0$$

- BLUE property is also known as **Gauss-Markov theorem**
- the assumption that $\mathbf{cov}(u) = I$ (or could be $\sigma^2 I$) is equivalent to
  - $\mathbf{var}(u_i) = \sigma^2$ for all i, $i.e.$, the error terms have the same variance (**homoskedasticity**)
  - $\mathbf{cov}(u_i, u_j) = 0$ for $i \neq j$, $i.e.$, the error terms are uncorrelated
- the proof on the optimality use the fact that $P = X(X^T X)^{-1} X^T$ is an **orthogonal projection** matrix with
  - $P^T = P$
  - $P^2 = P$
  - $\|Px\| \leq \|x\|$ for all $x \in \mathbf{R}^n$

  these properties imply that $I - P \succeq 0$ (see more on page 2-23)

Analysis of the LS estimate (stochastic case)

X is not a deterministic matrix (e.g. LS estimate of time series model)

we will explore the following properties of LS estimate

- identification
- consistency
- asymptotic distribution

Identification of LS estimate

the ability of LS etimate to permit identification of $\mathbf{E}[y|X]$ is follows

for the linear model, β is identified if

1. $\mathbf{E}[y|X] = X\beta$
2. $X\alpha = X\beta$ if and only if $\alpha = \beta$

- 1st assumption: the conditional mean is correctly specified ensures that β is of intrinsic interest
- 2nd assumption: equivalent to $\mathcal{N}(X) = \{0\}$ or X is full rank

Consistency of LS estimate

assumptions:

1. the data generating process (dgp) is actually the linear model on page 5-13
2. $\text{plim}(N^{-1}X^TX)^{-1}$ converges in probability to a finite nonzero matrix
3. $\text{plim}\, N^{-1}X^Tu = 0$

the LS estimate can be expressed as

$$\hat{\beta}_{\text{ls}} = \beta + (X^TX)^{-1}X^Tu = \beta + (N^{-1}X^TX)^{-1}N^{-1}X^Tu$$

apply the Slutsky's theorem and use the assumptions

$$\text{plim}\, \hat{\beta}_{\text{ls}} = \beta + \text{plim}(N^{-1}X^TX)^{-1} \cdot \text{plim}\, N^{-1}X^Tu = \beta$$

Distribution of LS estimator

assumptions:

1. the dgp model is $y = X\beta + u$ or $y_i = x_i^T \beta_i + u_i$ for $i = 1, \ldots, N$
2. data are **independent** over i (but not i.i.d.) with $\mathbf{E}[u|X] = 0$ and $\mathbf{E}[uu^T|X] = D = \mathbf{diag}(\sigma_i^2)$
3. X is full rank
4. $\Sigma_x = \operatorname{plim} N^{-1} X^T X$ exists and finite nonsingular
5. by CLT, $\frac{1}{\sqrt{N}} X^T u \xrightarrow{d} \mathcal{N}(0, \Sigma_{ux})$ where $\Sigma_{ux} = \operatorname{plim} N^{-1} X^T u u^T X$

then the LS estimate $\hat{\beta}_{\text{ls}}$ is consistent for β and

$$\sqrt{N}(\hat{\beta}_{\text{ls}} - \beta) \xrightarrow{d} \mathcal{N}(0, \Sigma_x^{-1} \Sigma_{ux} \Sigma_x^{-1})$$

the asymptotic distribution is applicable in *large samples*

Proof. with rescaling from page 5-22, the LS estimate can be expressed as

$$\sqrt{N}(\hat{\beta}_{\rm ls} - \beta) = \left(\frac{1}{N}X^T X\right)^{-1} \frac{1}{\sqrt{N}} X^T u$$

- assumption 2: $x_i u_i$ are independent, so by CLT (on page 4-36) and weak LLN

$$(1/\sqrt{N})X^T u = (1/\sqrt{N})\sum_{i=1}^{N} x_i u_i \xrightarrow{d} \mathcal{N}(0, \Sigma_{ux}), \quad \text{where}$$

$$\begin{aligned}\Sigma_{ux} &= \operatorname{plim} \frac{1}{N} X^T u u^T X = \operatorname{plim} \frac{1}{N}\sum_i u_i^2 x_i x_i^T = \lim \mathbf{E}\left[\frac{1}{N}\sum_i u_i^2 x_i x_i^T\right] \\ &= \lim \frac{1}{N}\sum_i \mathbf{E}[\mathbf{E}[u_i^2 x_i x_i^T | x_i]] = \lim \frac{1}{N}\sum_i \mathbf{E}[\mathbf{E}[u_i^2 | x_i] x_i x_i^T] \\ &= \lim \frac{1}{N}\sum_i \mathbf{E}[\sigma_i^2 x_i x_i^T] = \lim \frac{1}{N}\mathbf{E}[X^T D X]\end{aligned}$$

- assumption 3,4 and by weak LLN (on page 4-12)

$$\frac{1}{N}X^TX = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T \xrightarrow{p} \Sigma_x, \quad \text{where}$$

$$\Sigma_x = \operatorname{plim} \frac{1}{N}X^TX = \lim \frac{1}{N}\sum_{i=1}^{N} \mathbf{E}[x_i x_i^T]$$

- by Slutsky's theorem (on page 4-10) and that the inverse operator is continuous on the space of invertible matrices

$$\left(\frac{1}{N}X^TX\right)^{-1} \xrightarrow{p} \Sigma_x^{-1}$$

- by product limit normal rule (on page 4-16), we obtained the desired result where

$$\sqrt{N}(\hat{\beta}_{\rm LS} - \beta) \xrightarrow{d} \mathcal{N}(0, \Sigma_x^{-1}\Sigma_{ux}\Sigma_x^{-1})$$

Error assumptions

we explore the variance of LS estimate under two conditions on the error, u

- (conditional) homoskedasticity: u_i has the same variance for all i, σ^2

$$\mathbf{E}[uu^T|X] = D = \sigma^2 I$$

- (conditional) heteroskedasticity: u_i may have different variance, σ_i^2

$$\mathbf{E}[uu^T|X] = D = \mathbf{diag}(\sigma_i^2)$$

for both cases, it means u_i's are uncorrelated, *i.e.*, D is diagonal

if u_i's are correlated, then D is only symmetric

Asymptotic Variance Matrix of LS estimate

the asymptotic variance matrix of the distribution and the estimate are

$$P = \Sigma_x^{-1} \Sigma_{ux} \Sigma_x^{-1}, \quad \mathbf{Avar}(\hat{\beta}) = N^{-1} P$$

where

$$\Sigma_{ux} = \lim \frac{1}{N} \mathbf{E}[X^T D X], \quad \Sigma_x = \lim \frac{1}{N} \mathbf{E}[X^T X], \quad D = \mathbf{diag}(\sigma_i^2)$$

define the LS residual

$$\hat{u} = y - X\hat{\beta}$$

the asymptotic variance matrices can be substituted by their estimates

$$\hat{\Sigma}_{ux} = \frac{1}{N} X^T \hat{D} X, \quad \hat{\Sigma}_x = \frac{1}{N} X^T X, \quad \hat{D} = \mathbf{diag}(\hat{u}^2)$$

homoskedascity assumption: the estimated variance matrix can be simplified

if we assume homoskedasticity, $\mathbf{E}[u_i^2|x_i]$ is the same across i, $i.e.$, $D = \sigma^2 I$

hence, $\Sigma_{ux} = \sigma^2 \Sigma_x$ and the asymptotic variance matrix reduces to

$$\mathbf{Avar}(\hat{\beta}_{\rm ls}) = N^{-1} P = N^{-1} \sigma^2 \Sigma_x^{-1}$$

its estimate is given by

$$\hat{\sigma}^2 = \|\hat{u}\|_2^2/(N-n), \quad \widehat{\mathbf{Avar}}(\hat{\beta}_{\rm ls}) = N^{-1}\hat{\sigma}^2 \hat{\Sigma}_x^{-1} = \hat{\sigma}^2 (X^T X)^{-1}$$

- compare with the result on page 5-16
- $\hat{\sigma}^2$ is a consistent estimate of σ^2, regardless of the normalization $N-n$
- many computer packages use this as the *default* OLS variance estimate

heteroskedascity assumption: the asymptotic variance matrix is

$$\mathbf{Avar}(\hat{\beta}_{\mathrm{ls}}) = N^{-1}\Sigma_x^{-1}\Sigma_{ux}\Sigma_x^{-1}$$

and its estimate is

$$\widehat{\mathbf{Avar}}(\hat{\beta}_{\mathrm{ls}}) = N^{-1}\hat{\Sigma}_x^{-1}\hat{\Sigma}_{ux}\hat{\Sigma}_x^{-1} = (X^T X)^{-1}X^T\hat{D}X(X^T X)^{-1}$$

where $\hat{D} = \mathbf{diag}(\hat{u}^2)$ and $\hat{u} = y - X\hat{\beta}$

- $\widehat{\mathbf{Avar}}(\hat{\beta}_{\mathrm{ls}})$ is called **heteroskedastic-consistent** estimate of $\mathbf{Avar}(\hat{\beta}_{\mathrm{ls}})$
- many names for the standard errors, the square roots of the diagonals of $\widehat{\mathbf{Avar}}(\hat{\beta}_{\mathrm{ls}})$
  - White standard errors
  - heteroskedasticity-robust standard errors
  - Huber standard errors

Weighted least-squares

given a positive definite matrix, W that can be factorized as $W = L^T L$

a weighted least-squares problem is

$$\underset{\beta}{\text{minimize}} \;\; (X\beta - y)^T W (X\beta - y)$$

equivalent formulation:

$$\underset{\beta}{\text{minimize}} \;\; \|L(X\beta - y)\|_2^2$$

- if L is diagonal, we penalize each residual $u_i = y_i - X_i^T \beta_i$ by L_{ii}
- large L_{ii} means the ith residual should be much reduced
- a weighted LS reduces to OLS problem when $W = I$

Solution to Weighted LS Problem

normal equation: derive the zero gradient condition

$$X^T W X \beta = X^T W y$$

solution: if X is full rank and $W \succ 0$, then $X^T W X$ is invertible

$$\hat{\beta}_{\mathrm{wls}} = (X^T W X)^{-1} X^T W y$$

Generalized Least-Squares Estimator

Revisit BLUE property of LS: suppose $\mathbf{cov}(u)$ is *not* I, says $\mathbf{E}[uu^T] = \Sigma \succ 0$

scale the equation $y = X\beta + u$ by $\Sigma^{-1/2}$

$$\Sigma^{-1/2} y = \Sigma^{-1/2} X\beta + \Sigma^{-1/2} u$$

the optimal unbiased linear least-mean-squares estimator of β is

$$\hat{\beta}_{\mathrm{gls}} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} y$$

this is a special case of weighted least-squares solution when $W = \Sigma^{-1}$

- if Σ is known the weighted LS estimate is BLUE if $W = \Sigma^{-1}$
- large Σ_{ii} means u_i is more uncertain, so we should put less penalty on this residual
- this solution is known as **generalized least-squares estimator**

Feasible Generalized Least-Squares Estimator

the GLS estimator cannot be implemented because $\mathbf{cov}(u) = \Sigma$ is not known

if we replace Σ by a $\hat{\Sigma}$ in GLS estimator then it yields

$$\hat{\beta}_{\mathrm{fgls}} = (X^T \hat{\Sigma}^{-1} X)^{-1} X^T \hat{\Sigma}^{-1} y$$

known as the **feasible generalized least-squares (FGLS) estimator**

let us specify that $\Sigma = \Sigma(\gamma)$ where γ is a parameter vector

$$\sqrt{N}(\hat{\beta}_{\mathrm{fgls}} - \beta) \xrightarrow{d} \mathcal{N}\left[0, \left(\operatorname{plim} N^{-1} X^T \Sigma^{-1} X\right)^{-1}\right]$$

if we use $\hat{\Sigma} = \Sigma(\hat{\gamma})$ and $\hat{\gamma}$ is consistent for γ

conclusion: FGLS estimator is a special case of weighted LS estimator

Analysis of the WLS estimate (static case)

assumptions:

- the dgp is $y = X\beta + u$
- u is *white noise* with zero mean and covariance matrix Σ
- the weighted least-square estimate is given by $\hat{\beta} = (X^T W X)^{-1} X^T W y$
- the regressor X is *deterministic*

then the following properties hold:

- $\hat{\beta}$ is an unbiased estimate of β ($\mathbf{E}\hat{\beta} = \beta$, or $\hat{\beta} = \beta$ when $u = 0$)
- the covariance matrix of $\hat{\beta}$ is given by

$$\mathbf{cov}(\hat{\beta}) = (X^T W X)^{-1} X^T W \Sigma W X (X^T W X)^{-1}$$

Asymptotic asymptotic covariance matrix of WLS

assumptions: (dynamic case)

- the dgp is $y = X\beta + u$
- u is *white noise* with zero mean and covariance matrix Σ
- the weighted least-square estimate is given by $\hat{\beta} = (X^T W X)^{-1} X^T W y$
- the regressor X is *stochastic*

then the **estimated asymptotic covariance matrix** of WLS estimator is

$$\widehat{\mathbf{Avar}}(\hat{\beta}_{\mathrm{wls}}) = (X^T W X)^{-1} X^T W \hat{\Sigma} W X (X^T W X)^{-1}$$

where $\hat{\Sigma}$ (estimated covariance matrix of error) is such that

$$\operatorname{plim} N^{-1} X^T W \hat{\Sigma} W X = \operatorname{plim} N^{-1} X^T W \Sigma W X$$

conclusion: W must be chosen to be a good estimate of Σ^{-1}

References

Chapter 4 in

J.M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data*, the MIT press, 2010

Chapter 4 in

A.C. Cameron and P.K. Trivedi, *Microeconometircs: Methods and Applications*, Cambridge, 2005

Chapter 6 in

R.A. DeFusco, D.W. McLeavey, J.E.Pinto and D.E. Runkle, *Quantitative Investment Analysis*, 2nd edition, Wiley, 2004

Appendix in

W.H. Greene, *Econometric Analysis*, Prentice Hall, 2008