6. Linear Regression

- linear least-squares/regression
- solving linear least-squares
- BLUE property
- distribution of LS estimators
- weighted least-squares and other variants

Linear regression

- a linear relationship between variables y and x_k using a linear function:

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \triangleq x^T \beta$$

  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 often called the *regressed/response/explained/dependent variable*
- x_k's are the input variables that explain the behavior of y; called the *predictor/explanatory/independent variables*
- β is the *regression coefficient*
- example: product sale amount (unit) is explained by advertising costs (USD)

$$\text{Sales} = \beta_1 \cdot \text{TV} + \beta_2 \cdot \text{Radio} + \beta_3 \cdot \text{News paper}$$

  β_1 gives the average sale increase for one unit increase in TV ads (others fixed)

- given a data set: $\{(x_i, y_i)\}_{i=1}^m$ we can form a matrix form

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \quad \triangleq \quad y = X\beta$$

- the matrix X is sometimes called *the design/regressor matrix*

- given y and X, one would like to estimate β that gives the linear model output match best with y

- in practice, in the presence of noise and disturbance, more data should be collected in order to get a better estimate – leading to *overdetermined* linear equations

- an exact solution to $y = X\beta$ does not usually exist; however, it can be solved by **linear least-squares** formulation

Problem statement

overdetermined linear equations:

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

for most y cannot solve for β

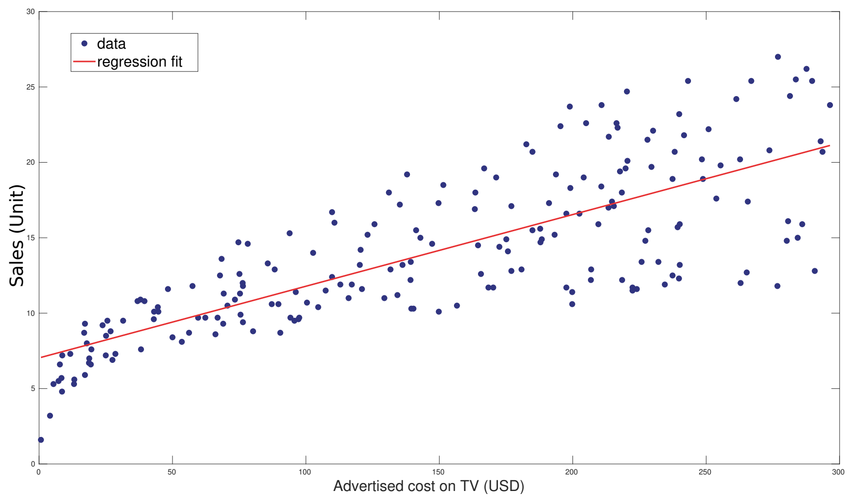
linear least-squares formulation:

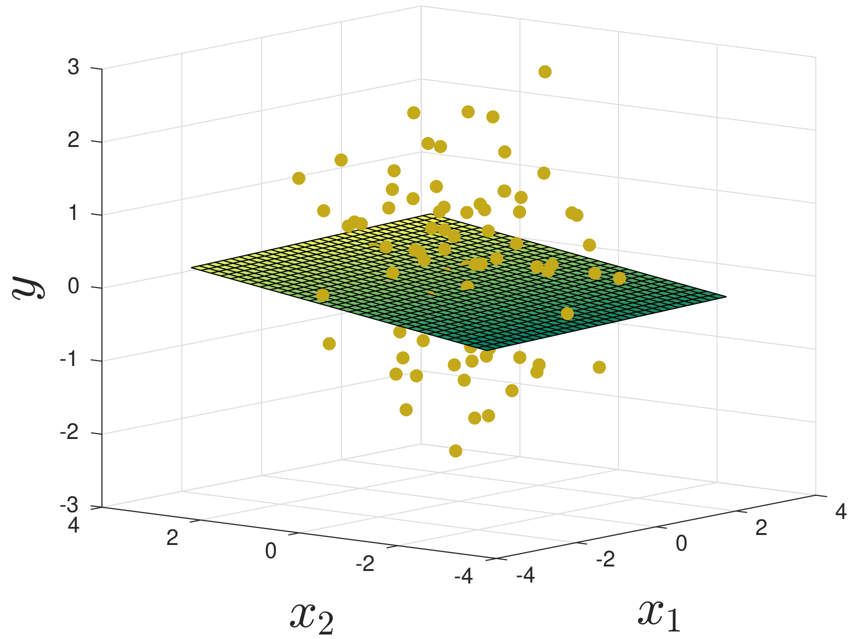
$$\underset{\beta}{\mathbf{minimize}} \quad \|y - X\beta\|_2 = \left(\sum_{i=1}^{m} (\sum_{j=1}^{n} X_{ij}\beta_j - y_i)^2 \right)^{1/2}$$

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

Fitting linear least-squares

left: explain the sale amount by advertising on TV





- left: sum squared distance of data points to the line is minimum (this line fits best)
- right: for two predictors, LS solution is the normal vector of hyperplane that lies closest to all data points of y

Example 1: data fitting

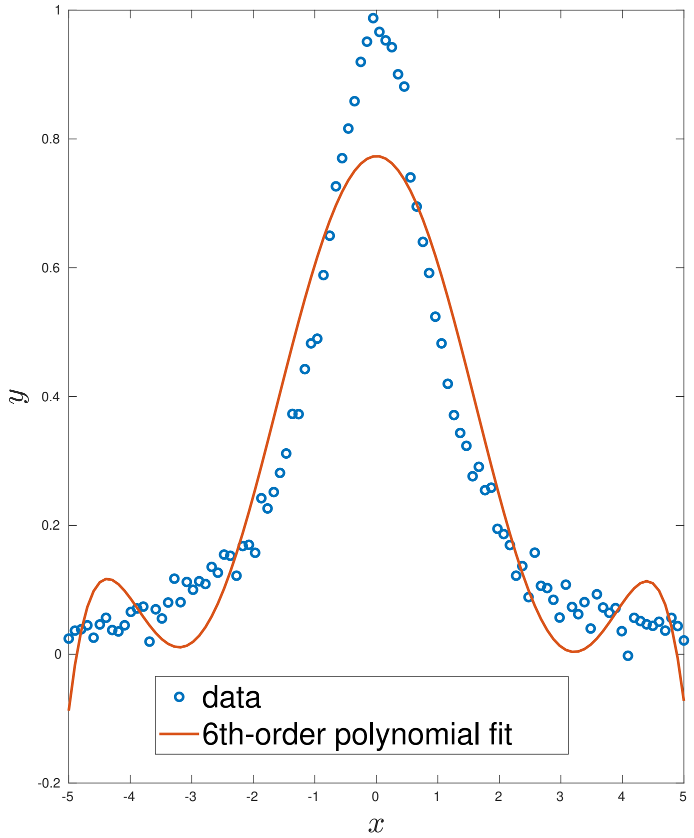
given data points $\{(t_i, y_i)\}_{i=1}^m$, we aim to approximate y using a function $g(t)$

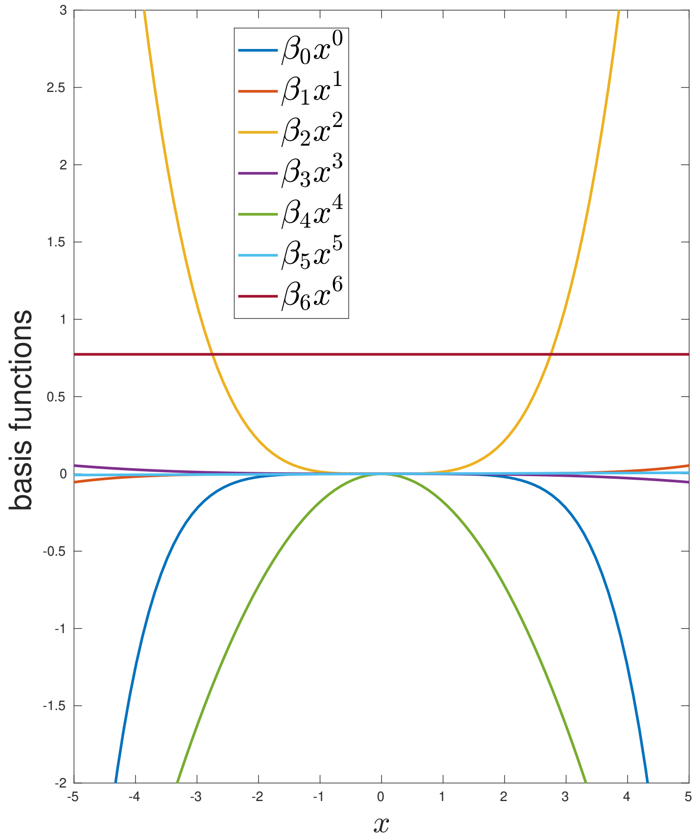
$$y = g(t) := \beta_1 g_1(t) + \beta_2 g_2(t) + \cdots + \beta_n g_n(t)$$

- $g_k(t) : \mathbf{R} \to \mathbf{R}$ is a basis function
  - polynomial functions: $1, t, t^2, \ldots, t^n$
  - sinusoidal functions: $\cos(\omega_k t), \sin(\omega_k t)$ for $k = 1, 2, \ldots, n$
- the linear regression model can be formulated as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} g_1(t_1) & g_2(t_1) & \cdots & g_n(t_1) \\ g_1(t_2) & g_2(t_2) & \cdots & g_n(t_2) \\ \vdots & & & \vdots \\ g_1(t_m) & g_2(t_m) & \cdots & g_n(t_m) \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \quad \triangleq \quad y = X\beta$$

- often have $m \gg n$, *i.e.*, explaining y using a few parameters in the model fitting a 6th-order polynomial to data points generated from $f(t) = 1/(1+t^2)$





- (right) the weighted sum of basis functions (x^k) is the fitted polynomial
- the ground-truth function f is nonlinear, but can be decomposed as a sum of polynomials

Example 2: scalar first-order model

given data set: $\{(u(t), y(t)\}_{t=1}^{N}$, we aim to estimate a scalar ARX model

$$y(t) = ay(t-1) + bu(t-1) + e(t)$$

$y(t)$ is linear in model parameters: a, b

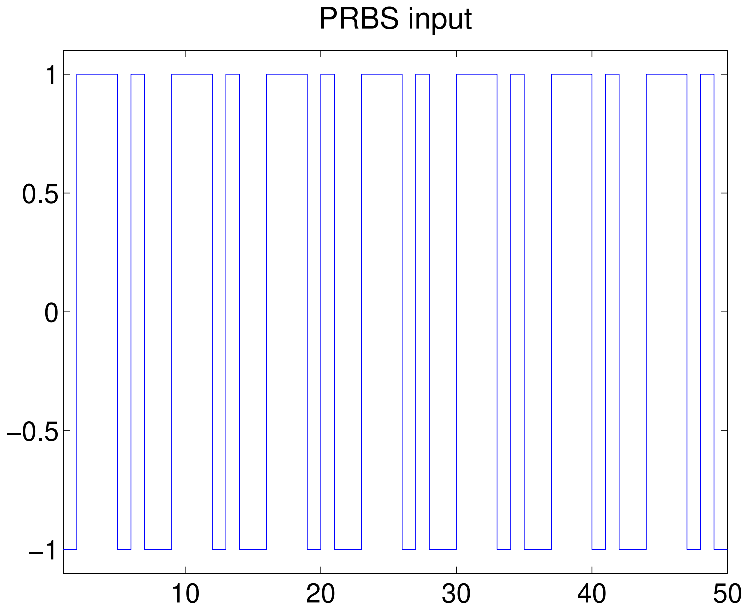
$$\begin{bmatrix} y(2) \\ y(3) \\ \vdots \\ y(N) \end{bmatrix} = \begin{bmatrix} y(1) & u(1) \\ y(2) & u(2) \\ \vdots & \vdots \\ y(N-1) & u(N-1) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

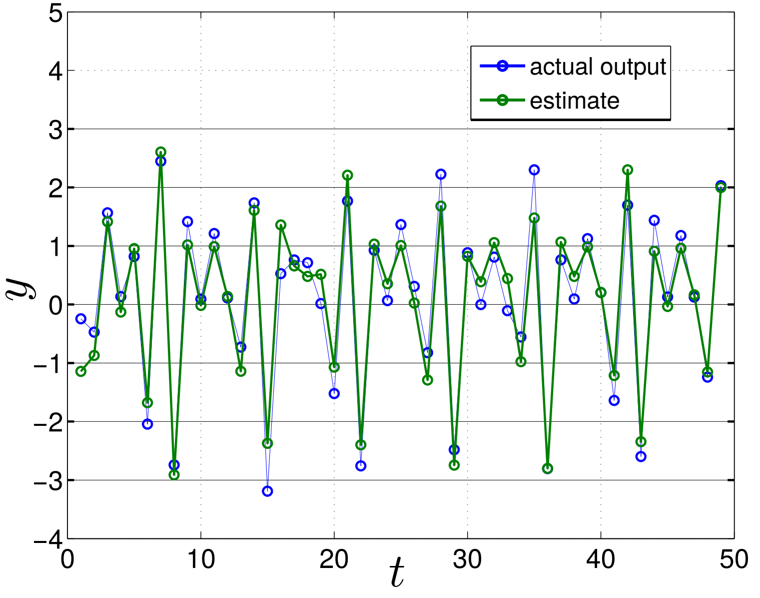
- the model is first-order, the equation is initialized with $y(1), u(1)$
- the model can be generalized to

$$y(t) = a_1 y(t-1) + \cdots + a_p y(t-p) + b_1 u(t-1) + \cdots + b_m u(t-m) + e(t)$$

where $\theta = (a_1, a_2, \ldots, a_p, b_1, b_2, \ldots, b_m)$ is the parameter vector data generation:

- $a = 0.8, b = 1$ are true parameters
- e is white noise with variance 0.1
- PRBS input





estimated parameters: $\hat{a} = 0.75, \hat{b} = 1.08$

Closed-form of least-squares estimate

the zero gradient condition of LS objective is

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

which is equivalent to the **normal equation**

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

if X is **full rank**:

- 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

Properties of full rank matrices

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

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

if X is full rank with $m \geq n$ (tall matrix)

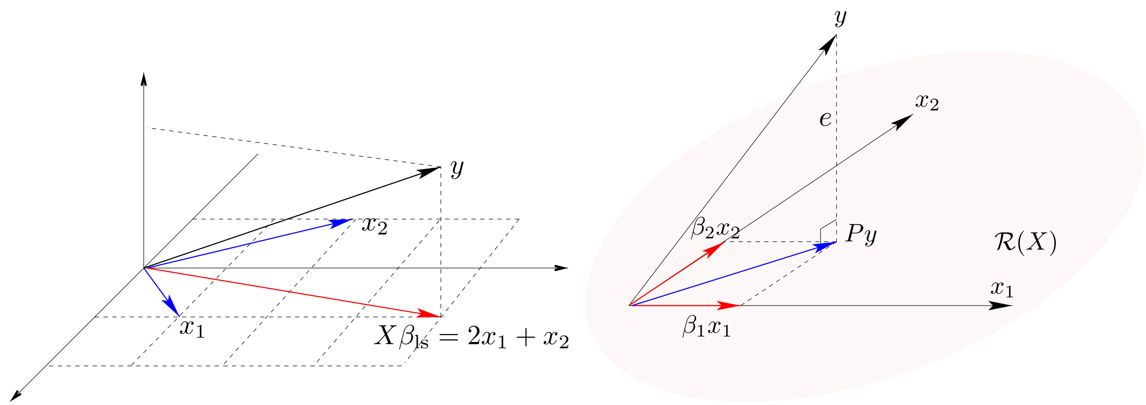
- $\mathbf{rank}(X) = n$ and $\mathcal{N}(X) = \{0\}$ ($Xz = 0 \Leftrightarrow z = 0$)
- $X^T X$ is positive definite: for any $z \neq 0$ then

$$z^T X^T X z = \|Xz\|^2 > 0$$

similarly, if X is full rank with $m \leq n$ (fat matrix)

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

Geometric interpretation of a LS problem



- $\|y - X\beta\|_2$ is the distance from y to

$$X\beta = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

- solution β_{ls} gives the linear combination of the columns of X closest to y
- $X\beta_{\text{ls}}$ is the **orthogonal projection** of y to the range of X
- Py gives the best approximation; for any $\hat{y} \in \mathcal{R}(X)$ and $\hat{y} \neq Py$

$$\|y - Py\| < \|y - \hat{y}\|$$

Numerical computation

we can solve a least-squares problem via

- Cholesky factorization: factor $X^T X \succ 0$ into LL^T where L is lower triangular
- QR factorization

most programming languages provide built-in commands

returned output	MATLAB	Python
$\hat{\beta}$	`X\y`	scipy.linalg.lstsq
estimated model	`fitlm`	sklearn.linear_model.LinearRegression

the closed-form $\hat{\beta} = (X^T X)^{-1} X^T y$ is for analysis purpose

we do not actually compute $\hat{\beta}$ from this expression

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

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

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

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

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

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

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

- 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 $\hat{\beta} - \mathbf{E}\hat{\beta} = (X^TX)^{-1}X^Tu$, hence,

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

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

BLUE property

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

the estimator defined by

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

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
- $\mathbf{cov}(\hat{\beta}) = BB^T$
- $\mathbf{cov}(\hat{\beta}_{\mathrm{ls}}) = BX(X^T X)^{-1} X^T B^T$ (apply $BX = I$)

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

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

- 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 $P = X(X^TX)^{-1}X^T$ 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 $I - P \succeq 0$

Properties of estimation errors

under the homoskedastic assumption $u_i \sim \mathcal{N}(0, \sigma^2)$ and define

$$\hat{u} = y - X\hat{\beta}_{\rm ls}, \quad \mathrm{RSS} = \sum_{i=1}^{N} \hat{u}_i^2, \quad s^2 = \mathrm{RSS}/(N-n) = (N-n)^{-1} \sum_{i=1}^{N} \hat{u}_i^2$$

Facts:

- s^2 is an unbiased estimate for σ^2
- $(N-n)s^2/\sigma^2 \sim \chi^2(N-n)$ (require Gaussian assumption of u_i)

proof sketch:

- unbiased property of s^2
  - $\hat{u} = (I - P)y \triangleq My$ where M is also an orhogonal projection matrix
  - $\hat{u} = Mu$ from the dgp: $y = X\beta + u$ and that $MX = 0$
  - since $M = I - X(X^TX)^{-1}X^T$ we have and $\mathbf{tr}(M) = \mathbf{tr}(I_N) - \mathbf{tr}(I_n)$
  - use $\mathbf{E}\|\hat{u}\|_2^2 = \mathbf{E}[u^TMu] = \mathbf{E}[\mathbf{tr}(u^TMu)]$

- chi-square distribution of s^2
  - $(N - n)s^2/\sigma^2 = \hat{u}^T\hat{u}/\sigma^2 = u^TMu/\sigma^2$
  - use that u_i/σ is standard Gaussian and that M is idempotent

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 6-16
2. $\mathbf{plim}(N^{-1}X^TX)^{-1}$ converges in probability to a finite nonzero matrix
3. $\mathbf{plim}\, N^{-1}X^Tu = 0$

the LS estimate can be expressed as

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

apply rules of limit in probability and use the assumptions

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

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 identically distributed) with

$$\mathbf{E}[u|X] = 0, \quad \mathbf{E}[uu^T|X] = D = \mathbf{diag}(\sigma_i^2)$$

3. X is full rank
4. $\Sigma_x = \mathbf{plim}\, N^{-1} X^T X$ exists and finite nonsingular
5. by CLT, $\frac{1}{\sqrt{N}} X^T u \xrightarrow{d} \mathcal{N}(0, \Sigma_{ux})$ where $\Sigma_{ux} = \mathbf{plim}\, N^{-1} X^T u u^T X$

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

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

Proof. with rescaling from page 6-26, 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 5-43) and weak LLN

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

$$\begin{aligned}\Sigma_{ux} &= \lim \frac{1}{N} \sum_{i=1}^{N} \mathbf{E}[x_i x_i^T u_i^2] \quad \text{(note: } \mathbf{E}[u_i x_i] = 0\text{)} \\ &= \lim \frac{1}{N} \sum_i \mathbf{E}[\mathbf{E}[u_i^2 x_i x_i^T | x_i]] = \lim \frac{1}{N} \sum_i \mathbf{E}[\mathbf{E}[u_i^2 | x_i] x_i x_i^T] \\ &= \lim \frac{1}{N} \sum_i \mathbf{E}[\sigma_i^2 x_i x_i^T] = \lim \frac{1}{N} \mathbf{E}[X^T D X]\end{aligned}$$

- assumption 3,4 and by weak LLN (on page 5-12)

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

- by continuous mapping theorem 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 5-17), we obtained the desired result where

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

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 covariance 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

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

its estimate is given by

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

- compare with the result on page 6-18
- $\hat{\sigma}^2$ is a consistent estimate of σ^2, regardless of the normalization $N-n$
- many computer packages use this as the *default* OLS variance estimate

consistency proof of $\hat{\sigma}^2$

- apply the definition and dgp: $y = X\beta + u$ where u is homoskedastic

$$\hat{\sigma}^2 = \frac{1}{N-n} u^T M u = \frac{N}{N-n} \left[\frac{u^T u}{N} - \left(\frac{u^T X}{N} \right) \left(\frac{X^T X}{N} \right)^{-1} \left(\frac{X^T u}{N} \right) \right]$$

- apply the limit in probability and the product limit rule
  - $\lim_{N\to\infty} N/(N-n) = 1$
  - $\mathbf{plim}(1/N)u^T u = \sigma^2$ (weak LLN)
  - $\mathbf{plim}(1/N)X^T X = \Sigma_x$ (assume to obtain positive matrix in large samples)
  - $\mathbf{plim}(1/N)X^T u = 0$ (assume $\mathbf{E}[u|X] = 0$)

$$(1/N)X^T u = (1/N)\sum_{i=1}^{N} u_i x_i \xrightarrow{p} \mathbf{E}[u_i x_i] = \mathbf{E}_x[\mathbf{E}[u_i x_i|x_i]] = \mathbf{E}_x[x_i \mathbf{E}[u_i|x_i]] = 0$$

- note that the proof follows even when the division is not $N - n$ (*e.g.*, N)

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

a weighted least-squares (WLS) problem is

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

- equivalent formulation: $\textbf{minimize}_x \;\; \|L(X\beta - y)\|^2$
- can be solved from the modified normal equation

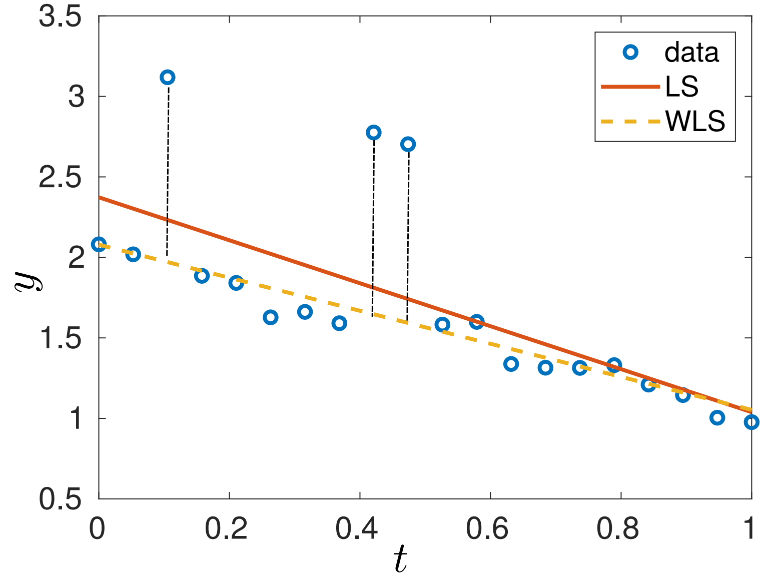
$$X^T W X \beta = X^T W y$$

- the solution is $\hat{\beta}_{\mathrm{wls}} = (X^T W X)^{-1} X^T W y$ (if X is full rank)
- $X\beta_{\mathrm{wls}}$ is the *orthogonal projection* on $\mathcal{R}(X)$ w.r.t the new inner product

$$\langle x, y \rangle_W = \langle Wx, y \rangle$$

Interpretation of WLS

when m-measurements contain some outliers (samples 3,9,10)



using $W = \mathbf{diag}(w_1, w_2, \ldots, w_m)$ gives WLS objective: $\sum_{i=1}^m w_i(y_i - x_i^T \beta)^2$

- use relatively **low** w_3, w_9, w_{10} to penalize **less** on those samples
- the linear model tends not to adapt to outliers – making WLS a more robust method than LS

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

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

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

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

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

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

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

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

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

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

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

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

MATLAB functions

- `fitlm` fits a linear regression
- `glmfit` fit a generalized linear model (linear regression is a special case and the default option)
- `fgls` solve feasible generalized least squares
- `robustfit` fit robust regressions

References

Chapter 3 in

G.James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, Springer, 2013

Chapter 4 and Appendix in

W.H. Greene, *Econometric Analysis*, Prentice Hall, 2008 Chapter 4 in

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

Chapter 4 in

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