II. Least Squares Estimation

Data $(Y_i, X_{i\cdot}^T)^T$, $i = 1, \dots, n$ not necessarily iid

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad X = \begin{pmatrix} X_{1\cdot}^T \\ \vdots \\ X_{n\cdot}^T \end{pmatrix} \begin{matrix} \\ \\ n \end{matrix} = (X^0, \dots, X^{k-1})$$

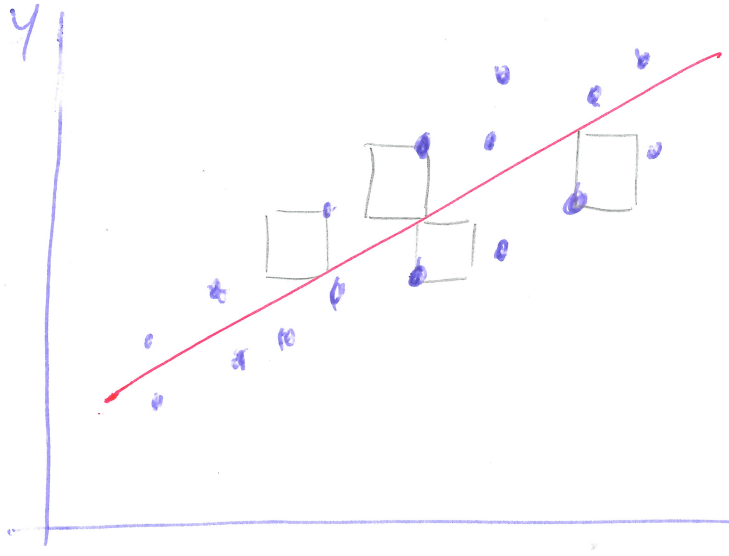
$X^0, \dots, X^{k-1}$: regressors

k

$Y | X \sim (X\beta, \sigma^2 I)$, $\operatorname{rank}(X) = r = k < n$ (almost surely)

β: unknown parameters

2.1 Sum of squares, LSE & normal equations



MODEL: $E[Y | Z = z] = \beta_0 + \beta_1 z = x^T\beta$, $x = (1, z)^T$

$\hat{\beta} = ?$

TAKE $= \operatorname{argmin}_\beta \underbrace{\sum_{i=1}^n (Y_i - X_{i\cdot}^T\beta)^2}_{SS(\beta)}$

Def. 2.1 Sum of squares

Consider a linear model $Y | X \sim (X\beta, \sigma^2 I_n)$. The function $SS: \mathbb{R}^k \to \mathbb{R}$ given as

$$SS(\beta) = \sum_{i=1}^n (Y_i - X_{i\cdot}^T\beta)^2 = \|Y - X\beta\|^2 = (Y - X\beta)^T(Y - X\beta),$$

$\beta \in \mathbb{R}^k$ will be called the sum of squares of the model.

Lemma 2.1 **Least squares estimator** | Let $r=k$.

There exist a unique minimizer to $SS(\beta)$ given as $\hat{\beta} := (X^TX)^{-1}(X^TY)$.

Proof: $SS(\beta) = \|Y - X\beta\|^2 = (Y-X\beta)^T(Y-X\beta) =$

$= Y^TY - 2Y^TX\beta + \beta^TX^TX\beta.$

$\frac{\partial SS}{\partial \beta}(\beta) = -2X^TY + 2X^TX\beta$

X^TX is $k \times k$, $\operatorname{rank}(X^TX) = k$

$\frac{\partial SS}{\partial \beta}(\beta) = 0_k \iff X^TX\beta = X^TY$

$\iff$

$\beta = (X^TX)^{-1}X^TY$

($(X^TX)^{-1}$ always exists, it is unique)

Do we have a minimum?

$\frac{\partial^2 SS}{\partial \beta \partial \beta^T}(\beta) = 2X^TX$

↳ for any $\beta \in \mathbb{R}^k$ this is positive definite matrix,

hence $\hat{\beta} := (X^TX)^{-1}X^TY$ is a unique minimizer to $SS(\beta)$

□

Note: $\hat{\beta}$ solves a linear system

$X^TX\beta = X^TY.$

Def 2.2 Least squares estimator, normal equations

Consider a linear model $Y|X \sim (X\beta, \sigma^2 I_n)$, $\operatorname{rank}(X_{n\times k}) = k$. The quantity $\hat{\beta} := (X^TX)^{-1}X^TY$ will be called the least squares estimator (LSE) of the vector of regression coeffs. β. The linear system $X^TX\beta = X^TY$ will be called the system of normal equations.

Lemma 2.2 Moments of LSE

Let $Y|X \sim (X\beta, \sigma^2 I_n)$, $\operatorname{rank}(X_{n\times k}) = k < n$. Then $E(\hat{\beta}|X) = \beta$, $E(\hat{\beta}) = \beta$,

$\operatorname{var}(\hat{\beta}|X) = \sigma^2(X^TX)^{-1}$.

Proof:

$$E(\hat{\beta}|X) = E((X^TX)^{-1}X^TY|X) =$$

$$= (X^TX)^{-1}X^T \underbrace{E(Y|X)}_{X\beta} = (X^TX)^{-1}X^TX\beta = \beta.$$

$$E(\hat{\beta}) = E(E(\hat{\beta}|X)) = E(\beta) = \beta$$

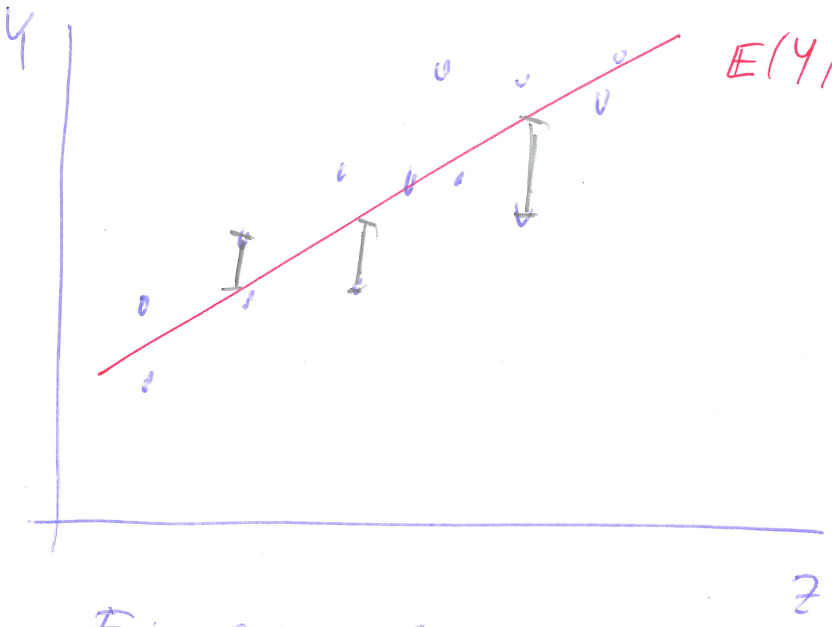
$$\operatorname{var}(\hat{\beta}|X) = \operatorname{var}((X^TX)^{-1}X^TY|X) =$$

$$= (X^TX)^{-1}X^T \underbrace{\operatorname{var}(Y|X)}_{\sigma^2 I_n} X(X^TX)^{-1} =$$

$$= \sigma^2 (X^TX)^{-1}X^TX(X^TX)^{-1} = \sigma^2(X^TX)^{-1}.$$

Remark: LSE $\hat{\beta}$ is unbiased estimator of β, both conditionally (given X) and unconditionally. → what does this mean?

Remark: Other methods exist on how to estimate β.



For example:

$$\tilde{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} |Y_i - X_i^T \beta|.$$

$\equiv$ L_1 regression

Will not be covered by this lecture.

2.2 Fitted values, residuals, projections

Notation: For a matrix $\mathbb{X} = (\underbrace{X^0, \ldots, X^{k-1}}_{k})_{n>k}$

$$\mathcal{M}(\mathbb{X}) := \{\underset{\sim}{v} : \underset{\sim}{v} = \underbrace{\sum_{j=0}^{k-1} b_j X^j}_{= \mathbb{X}b}, \; b = (b_0, \ldots, b_{k-1})^T \in \mathbb{R}^k\}$$

$\equiv$ linear span of columns of $\mathbb{X}$

$$\mathcal{M}(\mathbb{X})^\perp := \{\underset{\sim}{u} : u \in \mathbb{R}^n, \; v^T u = 0 \; \forall v \in \mathcal{M}(\mathbb{X})\}$$

$\equiv$ orthogonal complement to $\mathcal{M}(\mathbb{X})$

$\mathbb{X}_{n\times k}$: $\underset{\text{(column)}}{\text{rank}}(\mathbb{X}) = r \leq k < n$

$\left.\begin{array}{ll} \mathcal{M}(\mathbb{X}) \subset \mathbb{R}^n, & \text{vector dimension } r \\ \mathcal{M}(\mathbb{X})^\perp \subset \mathbb{R}^n, & n-r \end{array}\right\}$ a.s.

$$\mathcal{M}(\mathbb{X}) \cap \mathcal{M}(\mathbb{X})^\perp = \{0_n\}$$

$$\cup \qquad\qquad = \mathbb{R}^n$$

Def. 2.3 Regression and residual space of a linear model

Consider a linear model $Y | \mathbb{X} \sim (\mathbb{X}\beta, \sigma^2 I_n)$, $\text{rank}(\mathbb{X}_{n\times k}) = r \leq k < n$. The regression space of the model is a vector space $\mathcal{M}(\mathbb{X})$. The residual space of the model is the orthogonal complement of the regression space, i.e., a vector space $\mathcal{M}(\mathbb{X})^\perp$.

Def 2.4 Fitted values, residuals

Consider a linear model $\vec{Y}|X \sim (X\beta, \sigma^2 I_n)$, $\operatorname{rank}(X_{n\times k}) = k < n$. The vector $\hat{\vec{Y}} := X\hat{\beta} = X(X^TX)^{-1}X^T\vec{Y}$ will be called the vector of fitted values of the model, the vector $\hat{U} := \vec{Y} - \hat{\vec{Y}}$ will be called the vector of residuals of the model.

Notes:

$\hat{\vec{Y}} = X\hat{\beta} \in \mathcal{M}(X)$

$\left(= \sum_{j=0}^{k-1} \hat{\beta}_j X^{\vec{j}}\right)$

at the same time $\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|\vec{Y} - X\beta\|^2$

That is, $\hat{\vec{Y}} = X\hat{\beta}$ is the closest point (in L_2 norm) to $\vec{Y}$ in the regression space $\mathcal{M}(X)$

Linear algebra: $\hat{\vec{Y}}$ is projection of $\vec{Y}$ to $\mathcal{M}(X)$ (w.r.t. L_2 norm)

Since $\hat{\vec{Y}} = X(X^TX)^{-1}X^T\vec{Y}$, the matrix $H := X(X^TX)^{-1}X^T$ must be the respective projection matrix. (which is unique, see LA).

Residuals: $\vec{u} = \vec{y} - \hat{\vec{y}} = \vec{y} - H\cdot\vec{y} = (I_n - H)\vec{y}$

$M = I_n - H$

Basic properties of matrices H and M:

- $H\cdot H = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = H$

  $M\cdot M = (I-H)(I-H) = I - H - H + \underbrace{H\cdot H}_{H} =$

  $= I - H = M$

  Both H and M are idempotent.

- $H^T = (X(X^TX)^{-1}X^T)^T = X(X^TX)^{-1}X^T = H$

  $M^T = (I-H)^T = I - H^T = I - H = M$

  Both H and M are symmetric.

- $H\cdot M = H\cdot(I-H) = H - H\cdot H = H - H = \mathbb{O}_{n\times n}$

  $M\cdot H = \ldots = \mathbb{O}_{n\times n}$

$\Rightarrow$ Fitted values and residuals are mutually perpendicular:

$$\vec{u}^T\hat{\vec{y}} = (M\vec{y})^T H\vec{y} = \vec{y}^T \underbrace{M\cdot H}_{\mathbb{O}_{n\times n}}\vec{y} = 0$$

That is, $\vec{u} \perp \hat{\vec{y}}$, where $\hat{\vec{y}}$ is projection of $\vec{y}$ into $\mathcal{M}(X)$.

LA: $\vec{u} \in \mathcal{M}(X)^\perp$, $\vec{u} = M\vec{y}$ is projection of $\vec{y}$ into $\mathcal{M}(X)^\perp$.

$U = MY \quad (= (I-H)Y = Y - \hat{Y})$

LA: projection matrix is unique, hence M must be the respective projection matrix into $\mathcal{M}(X)^\perp$

- $HX = X(X^TX)^{-1}X^TX = X$

  $MX = (I-H)X = X - HX = X - X = 0_{n\times k}$

Lemma 2.3 Algebraic properties of fitted values and residuals and related projection matrices

(i) $\hat{Y} = HY$ and $U = MY$ are projections of Y into $\mathcal{M}(X)$ and $\mathcal{M}(X)^\perp$ respectively.

(ii) $\hat{Y} \perp U$ (almost surely).

(iii) H and M are projection matrices into $\mathcal{M}(X)$ and $\mathcal{M}(X)^\perp$, respectively.

(iv) $H^T = H, \quad M^T = M.$

(v) $H \cdot H = H, \quad M \cdot M = M.$

(vi) $H \cdot X = X, \quad MX = 0_{n\times k}.$

Proof: Previous derivations. ∎

Alternative expression of projection matrices

Now: not necessarily full-rank, i.e.

$X_{n\times k}$, $\operatorname{rank}(X) = r \leq k < n$

- $\mathcal{M}(X) \equiv$ vector space, vec-dim $= r$
- $\mathcal{M}(X)^\perp \equiv$ $= n-r$

LA: There exist an orthonormal vector basis of $\mathbb{R}^n$:

$$n\,(\underbrace{q_1, \ldots, q_r}_{r} \,/\, \underbrace{n_1, \ldots, n_{n-r}}_{n-r}) = P$$

Q — basis of $\mathcal{M}(X)$; N — basis of $\mathcal{M}(X)^\perp$

properties of orth. basis:

$$Q^T Q = I_r$$
$$N^T N = I_{n-r}$$
$$Q^T N = 0_{r\times(n-r)}$$
$$N^T Q = 0_{(n-r)\times r}$$

$P^T P = I_n$, P^T is inverse to P,

$\Rightarrow P \cdot P^T = I_n$

$$(Q, N) \cdot \begin{pmatrix} Q^T \\ N^T \end{pmatrix} = \underbrace{Q \cdot Q^T}_{=:\tilde{H}} + \underbrace{N N^T}_{=:\tilde{M}}$$

i.e. $I_n = \tilde{H} + \tilde{M}$

Properties of $\tilde{H}$ (and $\tilde{M}$):

$\tilde{H}^T = (Q \cdot Q^T)^T = Q \cdot Q^T = \tilde{H}$ (symmetry)

$\tilde{H} \cdot \tilde{H} = Q \cdot \underbrace{Q^T \cdot Q}_{I} \cdot Q^T = Q \cdot Q^T = \tilde{H}$ (idempotent)

(the same for $\tilde{M}$)

$\tilde{H} \cdot \tilde{M} = Q \cdot \underbrace{Q^T N}_{0} \cdot N^T = 0$

For any $y \in \mathbb{R}^n$

$\tilde{H} y = Q \cdot \underbrace{Q^T y}_{b = (b_1, \dots, b_k)^T} = Q \cdot b = \sum_{j=1}^{k} b_j \cdot q_j \in \mathcal{U}(X)$

$\tilde{M} y = N \cdot \underbrace{N^T y}_{c = (c_1, \dots, c_{n-k})^T} = N \cdot c \in \mathcal{U}(X)^{\perp}$

$y = I \cdot y = (\tilde{H} + \tilde{M}) y = \underbrace{\tilde{H} y}_{\in \mathcal{U}(X)} + \underbrace{\tilde{M} y}_{\in \mathcal{U}(X)^{\perp}}$

LA: • projections are unique, i.e.

$\tilde{H} y$ must be the projection of y into $\mathcal{U}(X)$

$\tilde{M} y$ — " — into $\mathcal{U}(X)^{\perp}$.

• also projection matrices are unique and hence (in case $k = k$):

$$\boxed{Q \cdot Q^T = \tilde{H} = H = X (X^T X)^{-1} X^T}$$

$$\underline{N \cdot N^T} = \tilde{M} = M = I_n - X (X^T X)^{-1} X^T$$

↑ They do not depend on choices of ~~the~~ orthonormal bases.

Summary:

Fitted values $\hat{\mathbb{Y}}$ = projection of $\mathbb{Y}$ into $\mathcal{M}(X)$

i.e. $\hat{\mathbb{Y}} = \underset{\tilde{\mathbb{Y}} \in \mathcal{M}(X)}{\operatorname{argmin}} \|\mathbb{Y} - \tilde{\mathbb{Y}}\|^2$

$\hat{\mathbb{Y}} = H \cdot \mathbb{Y}$, $H = X(X^T X)^{-1} X^T$ (if $\operatorname{rank}(X_{n \times k}) = k$)

$= Q \cdot Q^T$,

Q = orthonormal basis of $\mathcal{M}(X)$

Residuals $U = \mathbb{Y} - \hat{\mathbb{Y}} =$

= projection of $\mathbb{Y}$ into $\mathcal{M}(X)^\perp$

$U = M \cdot \mathbb{Y}$, $M = I_n - H =$

$= I_n - X(X^T X)^{-1} X^T$ (if $\operatorname{rank}(X) = k$)

$= N \cdot N^T$,

N = orthonormal basis of $\mathcal{M}(X)^\perp$

Terminology:

- H: hat matrix (projection matrix to the ~~the~~ regression space)
- M: residual projection matrix

Remarks: if $\operatorname{rank}(X_{n\times k}) = r < k < n$

- $SS(\beta) = \|Y - X\beta\|^2$ is minimized for any solution to normal equations $X^TX\beta = X^TY$

- All solutions to normal equations are $b = (X^TX)^- X^TY$

- Projection matrices can be expressed as $H = X(X^TX)^- X^T$, $M = I_n - X(X^TX)^- X^T$, where a pseudoinverse $(X^TX)^-$ is not unique, but matrices H and M are unique.

- $\hat{Y} = HY = X\underbrace{(X^TX)^- X^TY}_{b} = Xb$ is unique even if b is not unique

- $U = Y - \hat{Y}$ is also unique.