(8)

- P projects the observations in a space where exists "live" only linear combinations of the colums of X.
  - P is symmetric ($P' = P$) and idempotent $P^2 = P$

Notice also that

$$\hat{\varepsilon} = y - \hat{y}$$

$$= y - Py = (I - P)y = My$$

$M = I - P$ is also a projection matrix which ($M^2 = M$ and $M' = M$) projects the observation into a space orthogonal to V. ⟹ the matrix creates the residuals

or projects the theoretical residuals to the estimated ones since

$$\hat{\varepsilon} = y - X\hat{\beta} = My = M(X\beta + \varepsilon) = \cancel{MX\beta}^{\,0} + M\varepsilon$$

$$MX = (I - P)X = X - PX = X - X = 0$$

projects X on itself.

M is also called annihilator matrix.

Finally, note that $y = \hat{y} + \hat{\varepsilon} = Py + My$

that OLS decomposes y in two orthogonal vectors. ($\hat{y}'\hat{\varepsilon} = 0$)

- Goodness of fit
  - ↳ Sum of Squares (SS) decomposition the same as in the simple linear regression case
  - Check that

$$\underbrace{SST}_{\sum(y_i-\bar{y})^2} = \underbrace{SSR}_{\sum(\hat{y}_i-\bar{y})^2} + \underbrace{SSE}_{\sum(y_i-\hat{y}_i)^2 = \sum\hat{\varepsilon}_i^2}$$

Set $R^2 = \underbrace{1 - \frac{SSR}{SST}}_{\text{ANOVA interpretation}} = \frac{SSE}{SST}$

↳ Adjusted $R^2$

↳ since in multiple linear reg. we can add remove covariates we need to account about overfitting

thus we set $\bar{R}^2 = 1 - \frac{SSR/n-k}{SST/n-1} =$

$$= 1 - \frac{n-1}{n-k}(1-R^2)$$

- For larger $k$ we have larger $R^2$ but the $n-k$ term penalizes overparametrization

- Estimation of the residual variance $\sigma^2$

Recall that $\sigma^2 = Var(\varepsilon_i | X)$

From the variance decomposition formula we have that

$$Var(\varepsilon_i) = Var(\underbrace{E(\varepsilon_i | X)}_{0}) + E(V(\varepsilon_i|X)) = \sigma^2$$

* Note that this implies $\sigma^2 I_n = Cov(\varepsilon) = E[(\varepsilon - E(\varepsilon))(\varepsilon - E(\varepsilon))'] = E(\varepsilon\varepsilon')$

This suggest to estimate $\sigma^2$ by

$\frac{\sum \varepsilon_i^2}{n}$ but $\varepsilon_i$ are unobserved and therefore

we set

$s^2 = \frac{\sum \hat{\varepsilon}_i^2}{n-k}$ to be an estimator for $\sigma^2$.

* $n-k$ instead of $n$ since we have $k$ restrictions: the OLS residuals satisfy $k$ equations:

$$\underset{k\times n}{X'}\ \underset{n\times 1}{\hat{\varepsilon}} = 0_{k\times 1}$$

* Standard error of the regression: $s = \sqrt{s^2}$

→ Finite Sample Properties of OLS estimators:

1) OLS estimators are unbiased (minimum assumptions • GM1 – GM3)

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

Proof: We have that

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

$$= (X'X)^{-1}X'(X\beta + \varepsilon)$$

$$= \beta + (X'X)^{-1}X'\varepsilon$$

(Notice that $\underbrace{\hat{\beta} - \beta}_{\text{sampling error}} = (X'X)^{-1}X'\varepsilon$

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

$$= \beta + (X'X)^{-1}X'\underbrace{\cancel{E(\varepsilon|X)}}_{0} = \beta.$$

2) Covariance matrix of OLS

$$\mathrm{Cov}(\hat{\beta}|X) = \sigma^2 (X'X)^{-1}$$

Proof:

$$\mathrm{Cov}(\hat{\beta}) = E[(\hat{\beta} - \underbrace{\cancel{E(\hat{\beta}|X)}}_{\beta})(\hat{\beta} - \underbrace{\cancel{E(\hat{\beta}|X)}}_{\beta})' ]$$

$$= E((X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}|X]$$

$$= (X'X)^{-1}X'\underbrace{\cancel{E(\varepsilon\varepsilon'|X)}}_{\sigma^2 I_n}X(X'X)^{-1} = \sigma^2(X'X)^{-1}$$

3) $\operatorname{Cov}(\hat{\beta}, \hat{\varepsilon} \mid X) = 0$

$$\operatorname{Cov}(\hat{\beta}, \hat{\varepsilon} \mid X) = E\left[(\hat{\beta} - \underbrace{\cancel{E(\hat{\beta}\mid X)}}_{\to \beta})(\hat{\varepsilon} - \underbrace{\cancel{E(\hat{\varepsilon}\mid X)}}_{\to 0})'\right]$$

$$= E\left[(\hat{\beta} - \beta)\hat{\varepsilon}' \mid X\right] =$$

$$= E\left((X'X)^{-1}X'\varepsilon(M\varepsilon)' \mid X\right) =$$

$$= (X'X)^{-1}X' E(\varepsilon\varepsilon' M' \mid X)$$

$$= (X'X)^{-1}X'\sigma^2 I_n M'$$

$$\overset{MX=0}{=} 0_{k \times n}$$

- Gauss-Markov Theorem:

Under GM1 - GM4 the OLS $\hat{\beta}$ is the Best Linear Unbiased Estimator (BLUE)

Proof ....

What follows: Introduce new assumption that $\varepsilon \mid X \sim N(0, \sigma^2 I_n)$

↳ Hypothesis testing

# 3rd Lecture:

- Gauss-Markov Theorem:

Reminder:
- GM1: Linearity
- GM2: Multicollinearity
- GM3: $E(\varepsilon_i | X) = 0 \quad \forall i = 1, \dots, n$
- GM4: $Var(\varepsilon_i | X) = \sigma^2 \quad \forall i = 1, \dots, n$
  $Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j$

Under GM1-GM4 the OLS estimator is the $\underbrace{\text{Best Linear Unbiased Estimator}}_{\text{BLUE}}$ of $\beta$

↳ $Var(\hat{\beta}|X)$ is the ~~minimum~~ smallest one among the class of all linear (in $y$ ↳ $\hat{\beta} = Ay$) and unbiased estimators.

↳ Let $b$ an other estimator for $\beta$
Then $Var(b|X) - Var(\hat{\beta}|X)$ is a positive semi-definite matrix. $(\geq 0)$.

* A matrix $\Lambda$ is positive semi-definite if $z'\Lambda z \geq 0$ for any non-zero and real column matrix $z$.

- Gauss-Markov Theorem

Proof:

Let $b = Cy$ an other (linear in $y$) estimator

i.e. $C \neq (X'X)^{-1}X'$ and theis we can

assume that

$$C = \underset{k\times n}{D} + (X'X)^{-1}X'$$

Then $b = (D + (X'X)^{-1}X')y = Dy + \hat{\beta} =$

$$= D(\underbrace{Xb + \varepsilon}_{y}) + \hat{\beta}$$

$$= DXb + D\varepsilon + \hat{\beta} \quad (1)$$

Therefore $E(b|X) = DX\beta + D\cancel{E(\varepsilon|X)}^{\to 0} + E(\hat{\beta}|X) =$

$$= DX\beta + \hat{\beta}$$

Since $b$ needs to be unbised (in oreder to be BLUE)

We need $\underset{k\times n}{D}\underset{n\times k}{X} = \underset{k\times k}{0}$

Then from (1) we have that

$$b = D\varepsilon + \hat{\beta}$$

$$= D\varepsilon + (X'X)^{-1}X'(X\beta + \varepsilon)$$

$$= D\varepsilon + \beta + (X'X)^{-1}X'\varepsilon$$

$$= \beta + (D + (X'X)^{-1}X')\varepsilon \quad (2)$$

From (2) we have that

$$Var(b|X) = Var(\overbrace{D + (X'X)^{-1}X'}^{(\dots)})\varepsilon|X)$$

Var Properties

$$= (\dots)\,Var(\varepsilon|X)\,(\dots)' =$$

$$= (\dots)\,\sigma^2 I_n\,(\dots)' =$$

$$\overset{DX=0}{=} \sigma^2\,(\dots)(\dots)' =$$

$$= \sigma^2(DD' + (X'X)^{-1})$$

$$= \sigma^2 DD' + \sigma^2(X'X)^{-1} = \sigma^2 DD' + Var(\hat{\varepsilon}|X)$$

But $DD'$ is positive semi-definite

and thus $Var(b|X) - Var(\hat{\varepsilon}|X)$ is positive semi-definite

Let $c$ a non-zero vector

the $c'DDc = (D'c)'Dc$ = dot product of a vector with itself =

$$= \sum z_i^2 \geq 0$$ as sum of squares.

Finite Sample Properties of $\hat{\sigma}^2 = \frac{\sum \hat{\varepsilon}_i^2}{n-k}$

- $E(\hat{\sigma}^2 | X) = \sigma^2$ (unbiased estimator)

Proof:

We have that $\hat{\sigma}^2 = \frac{\sum \hat{\varepsilon}_i^2}{n-k} = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}$

Then $E(\hat{\varepsilon}'\hat{\varepsilon} | X) = E(\varepsilon' M \varepsilon | X) = (*)$

where $M = I_n - P$ is the annihilator matrix

(Reminder: $\hat{\varepsilon} = y - X\hat{\beta} = y - \hat{y} = y - Py =$
$= (I_n - P)y = My =$
$= M(X\beta + \varepsilon)$
$\overset{MX=0}{=} M\varepsilon$ )

Therefore, $\hat{\varepsilon}'\hat{\varepsilon} = \varepsilon' M' M \varepsilon \overset{M'=M}{=} \varepsilon' M M \varepsilon \overset{M^2=M}{=} \underbrace{\underset{1\times n}{\varepsilon'} \underset{n\times n}{M} \underset{n\times 1}{\varepsilon}}_{1\times 1}$

- Note thus that $\varepsilon' M \varepsilon = \text{trace}(\varepsilon' M \varepsilon)$

Then,

$(*) = E(\text{trace}(\varepsilon' M \varepsilon) | X) =$

$\overset{\text{trace}(AB) = \text{trace}(BA)}{=} E(\text{trace}(M\varepsilon\varepsilon') | X)$

$\overset{\text{trace is sum}}{=} \text{trace}(E(M\varepsilon\varepsilon' | X))$

$= \text{tracce}(M E(\varepsilon\varepsilon' | X)) = \text{trace}(M \, Cov(\varepsilon | X)) =$

$= \text{trace}(M\sigma^2 I_n) = \sigma^2\, \text{trace}(M)$

So we have so far that $E(\hat{\varepsilon}'\hat{\varepsilon} \mid X) = \sigma^2\, \text{trace}(M)$ (3)

But $\text{trace}(M) = \text{trace}(I_n) - \text{trace}(P) =$

$= n - \text{trace}(X(X'X)^{-1}X') =$

$\overset{\text{trace}(AB) = \text{trace}(BA)}{=} n - \text{trace}((X'X)(X'X)^{-1})$

$= n - \text{trace}(I_k) =$

$= n - k$

From (3) $E\left(\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k} \mid X\right) = \sigma^2$

or $E(\hat{\sigma}^2 \mid X) = \sigma^2$.

if $x \sim N(0, I_n)$ and $A$ an idempotent with rank $m$ then $x'Ax \sim \chi^2_m$

Remark 1: If $\varepsilon \mid X \sim N(0, \sigma^2 I_n)$

from distr. theory

then $\frac{\hat{\varepsilon}'\hat{\varepsilon}}{\sigma^2} = \frac{\varepsilon'}{\sigma} M \frac{\varepsilon}{\sigma} \sim \chi^2_{\text{rank}(M)}$

But since $M$ is idempotent $\text{rank}(M) = \text{trace}(M) =$
$= n - k$.

Which means also that

Remark 2: $\hat{\sigma}^2 (X'X)^{-1}$ is an estimator of $Var(\hat{\beta}|X) = \sigma^2 (X'X)^{-1}$

Remark 3: $\sigma_{\hat{\beta}_j} = \sqrt{Var(\hat{\beta}|X)_{jj}}$ is the standard error of $\hat{\beta}_j$ and thus

$\hat{\sigma}_{\hat{\beta}_j} = \hat{\sigma}\sqrt{(X'X)^{-1}_{jj}}$ is its estimator.

---

- An extra assumption:

$$\boxed{\varepsilon | X \sim N(0, \sigma^2 I_n)}$$

Remark 4:
Recall that $\hat{\beta} = \beta + (X'X)^{-1} X'\varepsilon$

But linear transformations preserve normality and thus

$$\hat{\beta}|X \sim N(\beta, \underbrace{\sigma^2 (X'X)^{-1}}_{\sigma^2_{\hat{\beta}_j}})$$

$$*\left( (X'X)^{-1} X' X (X'X)^{-1} = (X'X)^{-1} \right)$$ proof later on

Therefore, $\frac{\hat{\beta} - \beta_j}{\sigma_{\hat{\beta}_j}} \sim N(0,1)$ and if $\sigma_{\hat{\beta}_j}$ replaced by its estimator we have that Normal is $t_{(n-k)}$

Remark 5: $\hat{\sigma}$ and $\frac{\hat{\varepsilon}'\hat{\varepsilon}}{\sigma^2}$ are independent random variables

(See Greene Theorem B.12).

- Hypothesis testing

Consider the statistical test:

$$H_0: \beta_j = \beta_{c,j} \quad \text{vs} \quad H_1: \beta_j \neq \beta_{c,j}, \quad j = 1, \dots, k$$

where $\beta_{c,j}$ is a given constant (e.g. 0).

We rely on the statistic

$$t = \frac{\hat{\beta}_j - \beta_{c,j}}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{(X'X)^{-1}_{jj}}}$$

For $n > 100$

we reject $H_0$ if $|t| > 1.96$ for significance level 5%

significance $\rightarrow$ type I error $\rightarrow$ reject $H_0$ by mistake

* The statistic $t$ is know as t-statistic and

$t \sim t_{n-k}$ under $H_0$.

Proof:

We have that $t = \dfrac{\hat{\beta}_j - \beta_{j,c}}{\hat{\sigma}\sqrt{[(X'X)^{-1}]_{jj}}} \cdot \sqrt{\dfrac{(n-k)\sigma^2}{(n-k)\sigma^2}}$

$$= \frac{\dfrac{\hat{\beta}_j - \beta_{j,c}}{\sigma\sqrt{[(X'X)^{-1}]_{jj}}}}{\sqrt{\dfrac{(n-k)\hat{\sigma}^2}{\sigma^2(n-k)}}} \sim \frac{N(0,1)}{\chi^2_{n-k}/n-k} = t_{n-k}$$

Note: independent as functions of independent random variables.

see distribution theory.

- Confidence interval for $\beta_j$ of significance level $a$:

$$\left[\hat{\beta}_j \pm t_{\frac{a}{2},n-k}\, se(\hat{\beta}_j)\right]$$

if $a = 0.05$ then $(1-a)\%$ of these interval will contain the true value.

roughly: the probability that the true value is included in the interval is 95%.

• P-value is the probability p defined (9)

as $p = 2\,\text{Prob.}\left( t_{(n-k)} > |t| \right)$

↳ the probability that the t-statistic takes an extreme value given under $H_0$. Thus high p-value implies no rejection of the null.

Example: If for significance level $\alpha = 0.05$

p-value = 0.65 then we cannot reject $H_0$

if p-vale = 0.01 (i.e. $< \alpha$) then we reject $H_0$.