III. Basic Regression Diagnostics

DATA: $(Y_i, Z_i^T)^T$, $i = 1, \dots, n$

$\longrightarrow$ regressors $X_i = t(Z_i)$

$$\mathbb{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad \mathbb{Z} = \begin{pmatrix} Z_1^T \\ \vdots \\ Z_n^T \end{pmatrix} \to \mathbb{X} = \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix} =: t(\mathbb{Z})$$

Is it reasonable to assume, for chosen t:

- $E(\mathbb{Y} | \mathbb{Z}) = \mathbb{X}\beta$ for some β $\equiv E(\mathbb{Y}|\mathbb{Z}) \in \mathcal{M}(\mathbb{X})$
- $\text{var}(\mathbb{Y} | \mathbb{Z}) = \sigma^2 I_n$
- $\mathbb{Y} | \mathbb{Z} \sim \mathcal{N}$?

(not needed for many things)

3.1 (Normal) linear model assumptions

1-2

(A1) $E(Y_i | X_i = x) = x^T\beta$ for some β and (almost all) $x \in \mathcal{X}$

$\equiv$ correct regression function

(A2) $\text{var}(Y_i | X_i = x) = \sigma^2$ for some $\sigma^2 > 0$ irrespective of (almost all) values of $x \in \mathcal{X}$

$\equiv$ homoscedasticity

(A3) $\text{cov}(Y_i, Y_l | \mathbb{X} = \mathbb{x}) = 0$, $i \neq l$, for (almost all) $\mathbb{x} \in \mathcal{X}^n$

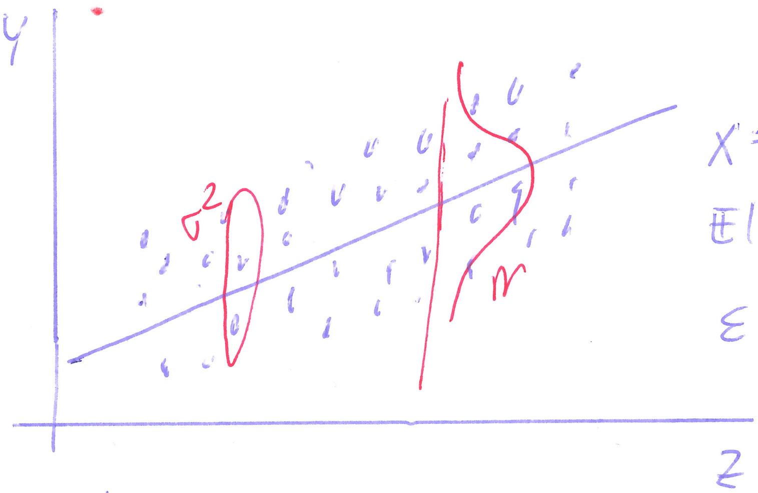
$\equiv$ the responses are conditionally uncorrelated

(A4) $Y_i | X_i = x \sim \mathcal{N}(x^T\beta, \sigma^2)$ for (almost all) $x \in \mathcal{X}$

$\equiv$ normality

Discussion of assupt.

Assumptions in terms of the errors ε



$X = (1, Z)$

$\mathbb{E}(Y \mid Z = z) = \beta_0 + \beta_1 z$ (MODEL)

$\varepsilon = Y - (\beta_0 + \beta_1 Z)$

(A1) $\mathbb{E}(\varepsilon_i \mid X_i = x) = 0$ for (almost all) $x \in \mathcal{X}$

$(\Rightarrow \mathbb{E}\varepsilon_i = 0, \ i = 1, \dots, n)$

$\equiv$ the regression function of the model is correctly specified

(A2) $\mathrm{var}(\varepsilon_i \mid X_i = x) = \sigma^2$ for some $\sigma^2 > 0$ which is constant irrespective of (almost all) values of $x \in \mathcal{X}$

$(\Rightarrow \mathrm{var}(\varepsilon_i) = \sigma^2, \ i = 1, \dots, n)$

$\equiv$ homoscedasticity of the errors

(A3) $\mathrm{cov}(\varepsilon_i, \varepsilon_l \mid X = x) = 0, \ i \neq l$ for (almost all) $x \in \mathcal{X}^n$

$(\Rightarrow \mathrm{cov}(\varepsilon_i; \varepsilon_l) = 0, \ i \neq l$

$\equiv$ the errors are uncorrelated

(A4) $\varepsilon_i \mid X_i = x \sim \mathcal{N}(0, \sigma^2)$ for (almost all) $x \in \mathcal{X}$

$(\Rightarrow \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \ i = 1, \dots, n)$

$\equiv$ ~~the~~ the errors are normally distributed and owing to previous assumptions,

$\varepsilon_1, \dots, \varepsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

Most checks performed through residuals

$$U = MY \quad , \quad M = I_n - H = I_n - X(X^T X)^{-} X^T$$
$$= Y - \hat{Y}$$

Assumptions and residual properties

(A1) $\Rightarrow E(U|X) = 0_n$

(A1) & (A2) & (A3) $\Rightarrow var(U|X) = \sigma^2 M.$

(A1) & (A2) & (A3) & (A4) $\Rightarrow U|X \sim N_n(0, \sigma^2 M)$

Strategy

RHS of implication checked

- if not satisfied then also LHS not valid
- if satisfied, still no guarantee that the LHS valid

---

one more complication

$$Y - X\hat{\beta} = U \equiv \hat{\varepsilon} \quad , \quad \varepsilon = Y - X\beta$$

$var(U|X) = \sigma^2 M$ $\quad\quad$ $var(\varepsilon|X) = \sigma^2 \underline{I_n}$

- not diagonal matrix
- not const diagonal

- uncorrelated
- homoscedast.

3.2 Standardized residuals

We know: $\operatorname{var}(U_i \mid X_i) = \sigma^2 m_{ii}$, $i = 1, \dots, n$

$$\Rightarrow \operatorname{var}\left(\frac{U_i}{\sqrt{\sigma^2 m_{ii}}} \middle| X_i \right) = 1$$

LATER: $m_{ii} = 0 \iff \operatorname{rank}(\underline{X_{(-i)}}) = \operatorname{rank}(X) - 1$

matrix X without row i

$$\boxed{? \quad \operatorname{var}\left(\frac{U_i}{\sqrt{MSe\, m_{ii}}} \middle| X_i \right) = 1 \quad ?}$$

6

Def 3.1 Standardized residuals

The standardized residuals or the vector of standardized residuals of the model is the vector $U^{std} = (U_1^{std}, \dots, U_n^{std})^T$, where

$$U_i^{std} = \begin{cases} \dfrac{U_i}{\sqrt{MSe\, m_{ii}}}, & m_{ii} > 0, \\ \text{undefined}, & m_{ii} = 0, \end{cases} \qquad i = 1, \dots, n.$$

Lemma 3.1 Moments of standardized residuals under normality

Let $Y | X \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$ and let for chosen $i \in \{1, .., n\}$, $m_{ii} > 0$. Then

$$\mathbb{E}(U_i^{std} | X) = 0, \quad var(U_i^{std} | X) = 1.$$

Proof: See the notes (not for exam).

First Lemma B.2 $Z \sim \mathcal{N}_n(0, \sigma^2 I_n)$

$T: \mathbb{R}^n \to \mathbb{R}$ measurable fun.,

scale transformations invariant: $\forall c > 0 \ \forall z \in \mathbb{R}^n$

$T(cZ) = T(Z) \implies T(Z) \perp\!\!\!\perp \|Z\|$.

$$U_i^{std} = \frac{U_i}{\sqrt{MS_e \, m_{ii}}} = \frac{U_i}{\sqrt{SS_e}} \sqrt{\frac{n-r}{m_{ii}}} = \frac{U_i}{\|U\|} \cdot \underbrace{\sqrt{\frac{n-r}{m_{ii}}}}_{\substack{= \text{const} \\ \text{given } X}}$$

$U = M \cdot Y = N N^T Y$, N = orthonormal basis of $\mathcal{M}(X)^\perp$

given X: $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n) \implies$

$$N^T Y \sim \mathcal{N}(\underbrace{N^T X}_{0} \beta, \sigma^2 \underbrace{N^T N}_{I_{n-r}}), \quad r = rank(X)$$

(columns of N,
scalar products with cols. of X,
cols of N = basis of $\mathcal{M}(X)^\perp$

That is, $\underbrace{N^T Y}_{Z \text{ from Lemma}} | X \sim N(0, \sigma^2 I_{n-r})$

$$T(Z) := \frac{j_i^T (NZ)}{(\|NZ\|)} \qquad NZ = NN^TY = U, \quad \|NZ\| = \|NN^TY\| = \|U\|$$

$j_i = (0, \dots, 1, \dots 0)^T$
place i

$$T(cZ) = \frac{j_i^T N c Z}{\|cNZ\|} = T(Z)$$

$\overset{\text{Lemma}}{\Longrightarrow}$ $T(Z) = \dfrac{U_i}{\|U\|}$ and $\|Z\| = \|N^TY\| =$

$$= \sqrt{Y^T \underbrace{NN^T}_{M = MM} Y} = \|U\|$$

are independent (given X)

All expectations given X:

$$E\left(\frac{U_i}{\|U\|} \cdot \|U\|\right) = E\left(\frac{U_i}{\|U\|}\right) \cdot \underbrace{E\|U\|}_{>0}$$

$\| $
$EU_i = 0$
we know

since $\|U\| \geq 0$ and $\frac{1}{\sigma^2}\|U\|^2 \sim \chi^2_{n-r}$, hence it cannot be $\|U\| = 0$ a.s.

That is $0 = E\left(\dfrac{U_i}{\|U\|} \,\Big|\, X\right)$

$$\Rightarrow \underline{E(U_i^{std} | X) = 0}$$

$$E\left(\frac{U_i^2}{\|U\|^2} \cdot \|U\|^2\right) = E\left(\frac{U_i^2}{\|U\|^2}\right) \cdot \underbrace{E\|U\|^2}_{\sigma^2 (n-r)}$$

$\| $
$EU_i^2 = \sigma^2 m_{ii}$
we know

That is, $E\left(\dfrac{U_i^2}{\|U\|^2} \,\Big|\, X\right) = var\left(\dfrac{U_i^2}{\|U\|^2} \,\Big|\, X\right)$

$$= \frac{\sigma^2 m_{ii}}{\sigma^2 (n-r)} = \frac{m_{ii}}{n-r}$$

$$\Rightarrow \underline{var(U_i^{std} | X) = 1}$$

3.3 Graphical tools of regression diagnostics

Model matrix $\mathbb{X} = (X^0, \ldots, X^{k-1}) = t(Z)$

mostly $X^0 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$

often: additional regressors are available

$\mathbb{V} := (V^1, \ldots, V^m) = t_V(Z)$

(A1) $\mathbb{E}(\mathbb{Y}|\mathbb{Z}) = \mathbb{X}\beta$ for some $\beta \in \mathbb{R}^k$?

- perhaps $= \mathbb{X}\beta + \mathbb{V}\gamma \quad (\gamma \in \mathbb{R}^m)$

(A2) $\mathrm{var}(\mathbb{Y}|\mathbb{Z}) = \sigma^2 I_n$

- perhaps $\sigma^2 = \sigma^2(\mathbb{X}), \sigma^2(\mathbb{V}), \sigma^2(\mathbb{X}, \mathbb{V})$

(A1) plots $\quad \mathbb{E}(\mathbb{U}|\mathbb{Z}) = \mathbb{O}$

- does not depend on $\mathbb{X}, \mathbb{V}$
any function of them
$(\hat{\mathbb{Y}}), \ldots$

(A2) homoscedasticity - plots + ... 11-13

Problems when evaluating plots:

$var(U_i / Z) = \sigma^2 m_{ii}$, raw residuals are NOT homoscedast. even if (A2) OK

$var(U_i^{std} / Z) = 1$ (only under $\mathcal{N}$)

scale-location plot: $\sqrt{|U_i^{std}|}$ on y-axis

(A3) uncorrelated errors 14-15

(A4) normality 16

- if $\mathcal{N}$ holds $U / Z \sim \mathcal{N}(0, \sigma^2 M)$

  - not diagonal matrix
  - diagonal not constant (heteroscedasticity)

$U^{std} / Z \sim \boxed{?}(0, \begin{pmatrix} 1 & & * \\ & \ddots & \\ * & & 1 \end{pmatrix})$

- not necessarily normal

not diagonal matrix

- overview of three the most important plots 17

- Examples 18-23