XI. Unusual Observations

Whole chapter → M: $Y|X \sim (X\beta, \sigma^2 I_n)$
$\text{rank}(X_{n\times k}) = k$

Interest: EXPLORATORY identification of unusual observations

- summary of (standard) notation

→ all probabilistic statements will be conditional given X

$$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}$$

(both capital and small letters used)

as usual $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ = vector of outcomes.

11.1 Leave-one-out and outlier model

NOTATION: for chosen $t \in \{1, \dots, n\}$

$Y_{(-t)} = Y$ without Y_t

x_t^T = row t of the model matrix X

$X_{(-t)}$ = matrix X without row t

$j_t = (0, \dots, \underset{\text{place } t}{1}, \dots, 0)^T \in \mathbb{R}^n$

Def 11.1 Leave-one-out model

The t^{th} leave-one-out model is a linear model

$$M_{(-t)}: Y_{(-t)} \,|\, X_{(-t)} \sim (X_{(-t)}\beta, \sigma^2 I_{n-1}).$$

Def 11.2 Outlier model

The t^{th} outlier model is a linear model

$$M_t^{out}: Y \,|\, X \sim (X\beta + j_t \gamma_t^{out}, \sigma^2 I_n).$$

Meaning of the outlier model

$$E(Y_s \,|\, X_s = x_s) = x_s^T \beta, \quad s \neq t$$

$$E(Y_t \,|\, X_t = x_t) = \underbrace{x_t^T \beta + \gamma_t^{out}}$$

can be whatever

$\equiv$ t^{th} observation does not follow model M

Lemma 11.1 Three equivalent statements

While assuming $\text{rank}(X_{n\times k}) = k$, the following three statements are equivalent:

(i) $\text{rank}(X) = \text{rank}(X_{(-t)})$, i.e. $x_t \in \mathcal{M}(X_{(-t)}^T)$;

(ii) $m_{tt} > 0$;

(iii) $\text{rank}(X, j_t) = k+1$.

Further remarks:

(ii) $\Leftarrow$ $h_{tt} < 1$

(iii) $\Leftarrow$ The t^{th} outlier model is of full-rank.

Proof: see the notes (exercise in linear algebra), not requested for exam

---

NOTATION (once conditions of Lemma 11.1 satisfied):

$M_{(-t)}$: $Y_{(-t)} | X_{(-t)} \sim (X_{(-t)}\beta, \sigma^2 I_{n-1})$

$\rightarrow \hat{\beta}_{(-t)}, \hat{y}_{(-t)}, SSe_{(-t)}, MSe_{(-t)}, \ldots$

M_t^{out}: $Y | X \sim (X\beta + j_t \mu_t^{out}, \sigma^2 I_n)$

$\rightarrow \hat{\beta}_t^{out}, \hat{y}_t^{out}, SSe_t^{out}, MSe_t^{out}, \ldots$

$\left((\hat{\beta}_t^{out})^T, \hat{\mu}_t^{out}\right)^T = LSE$ of $\left((\beta_t^{out})^T, \mu_t^{out}\right)$

Lemma 11.2 Equivalence of the outlier model and the leave-one-out model

1. The residual sums of squares in models $M_{(t)}$ and M_t^{out} are the same, i.e., $SS_{e,(-t)} = SS_{e,t}^{out}$.

2. Vector $\hat{\beta}_{(-t)}$ solves the normal equations of model $M_{(-t)}$ if and only if a vector $((\hat{\beta}_t^{out})^T, \hat{\gamma}_t^{out})^T$ solves the normal equations of model M_t^{out}, where $\hat{\beta}_t^{out} = \hat{\beta}_{(-t)}$,
$\hat{\gamma}_t^{out} = Y_t - x_t^T \hat{\beta}_{(-t)}$.

Proof: Remember: solution to normal equations must minimize the corresponding sum of squares.

M_t^{out}: $SS_t^{out}(\beta, \gamma_t^{out}) = \|Y - X\beta - d_t \gamma_t^{out}\|^2 =$

$= \sum_{s \neq t} (Y_s - x_s^T \beta)^2 + (Y_t - x_t^T \beta - \gamma_t^{out})^2$

$\|Y_{(-t)} - X_{(-t)}\beta\|^2 = SS_{(-t)}(\beta)$ = sum of squares to be minimized in $M_{(-t)}$

That is, $SS_t^{out}(\beta, \gamma_t^{out}) = SS_{(-t)}(\beta) + \underbrace{(Y_t - x_t^T \beta - \gamma_t^{out})^2}_{\geq 0}$

- For arbitrary $\beta \in \mathbb{R}^k$, $(Y_t - x_t^T \beta - \gamma_t^{out})^2 = 0$ if $\gamma_t^{out} = Y_t - x_t^T \beta$.
- Hence $\min_{\beta, \gamma_t^{out}} SS_t^{out}(\beta, \gamma_t^{out}) = \min_{\beta} SS_{(-t)}(\beta)$

and $\hat{\beta}_{(-t)}$ minimizes $SS_{(-t)}(\beta)$

$\Leftrightarrow$ $((\underbrace{\hat{\beta}_{(-t)}}_{\hat{\beta}_t^{out}})^T, \underbrace{Y_t - x_t^T \hat{\beta}_{(-t)}}_{\hat{\gamma}_t^{out}})^T$ minimizes $SS_t^{out}(\beta, \gamma_t^{out})$.

NOTATION: Leave-one-out least squares estimators of the response expectations

If $m_{tt} > 0$ for all $t = 1, \dots, n$:

- $\hat{Y}_{[t]} := x_t^T \hat{\beta}_{(-t)}$ = LSE of $E(Y_t / X_t = x_t)$ based on a model $M_{(-t)}$

- $\hat{\mathbb{Y}}_{[0]} := (\hat{Y}_{[1]}, \dots, \hat{Y}_{[n]})^T$
  = LSE of $\mu = \mathbb{X}\beta = E(\mathbb{Y}/\mathbb{X})$, where each element μ_t of μ is estimated using model $M_{(-t)}$ (= data without obs. t)

$M_t^{out}: \mathbb{Y}/\mathbb{X} \sim (\mathbb{X}\beta + j_t \gamma_t^{out}, \sigma^2 I_n)$
= model with added regressor to model
$M: \mathbb{Y}/\mathbb{X} \sim (\mathbb{X}\beta, \sigma^2 I_n)$

Lemma 9.1 (under the condition $m_{tt} > 0$ = everything of full rank)

$$\hat{\gamma}_t^{out} \overset{9.1}{=} (\underbrace{j_t^T M j_t}_{m_{tt}})^{-1} \underbrace{j_t^T U}_{U_t} = \frac{U_t}{m_{tt}}$$

$$\overset{11.2}{=} Y_t - x_t^T \hat{\beta}_{(-t)}$$

$$\overset{\text{notation}}{=} Y_t - \hat{Y}_{[t]}$$

$$\hat{\beta}_{(-t)} \overset{11.2}{=} \hat{\beta}_t^{out} \overset{9.1}{=} \hat{\beta} - \frac{U_t}{m_{tt}} (X^T X)^{-1} x_t$$

$$SSe_{(-t)} \overset{11.2}{=} SSe_t^{out} \overset{9.1}{=} SSe - \frac{U_t^2}{m_{tt}}$$

$$\overset{\text{algebra}}{=} SSe - MSe\,(U_t^{std})^2$$

remember $U_t^{std} = \dfrac{U_t}{\sqrt{MSe\; m_{tt}}}$

Further $MSe_{(-t)} = \dfrac{SSe_{(-t)}}{n-k-1}$, $MSe_t^{out} = \dfrac{SSe_t^{out}}{n-(k+1)}$

$$\Rightarrow \frac{MSe_{(-t)}}{MSe} \overset{11.2}{=} \frac{MSe_t^{out}}{MSe} \overset{\text{algebra}}{=} \frac{n-k-(U_t^{std})^2}{n-k-1}$$

Finally: $\hat{y}_t^{out} \overset{9.1}{=} \hat{y} + \dfrac{U_t}{m_{tt}} m_t$

↑ row (or column) of matrix $\mathbb{M}$.

Lemma 11.3 Quantities of the outlier and leave-one-out model expressed using quantities of the original model

Suppose that for given $t \in \{1, \dots, n\}$ $m_{tt} > 0$.
The following quantities of the outlier model M_t^{out} and the leave-one-out model $M_{(-t)}$ are expressable using the quantities of the original model M as follows.

$$\hat{\varepsilon}_t^{out} = Y_t - x_t^T \hat{\beta}_{(-t)} = Y_t - \hat{Y}_{[t]} = \frac{\hat{u}_t}{m_{tt}},$$

$$\hat{\beta}_{(-t)} = \hat{\beta}_t^{out} = \hat{\beta} - \frac{\hat{u}_t}{m_{tt}} (X^T X)^{-1} x_t,$$

$$SSe_{(-t)} = SSe_t^{out} = SSe - \frac{\hat{u}_t^2}{m_{tt}} = SSe - MSe \cdot (\hat{u}_t^{stud})^2$$

$$\frac{MSe_{(-t)}}{MSe} = \frac{MSe_t^{out}}{MSe} = \frac{n-k-(\hat{u}_t^{stud})^2}{n-k-1}.$$

Proof: Previous calculations while using Lemmas 9.1 and 11.2. ∎

TO REMEMBER:

- Lemma 11.3 provides quantification of influence of obs. t on $\hat{\beta}$, MSe, ...
- all quantities can be calculated from the original fitted model (no need to estimate models $M_{(-t)}$ or M_t^{out} for each t)

$\Rightarrow$ speciality of linear model

Def 11.3 **Deleted residual**

If $m_{tt} > 0$, then the quantity

$$\hat{\gamma}_t^{out} = Y_t - \underbrace{\hat{Y}_{[t]}}_{x_t^T \hat{\beta}_{(t)}} = \frac{u_t}{m_{tt}}$$

$= \hat{E}(Y_t | X_t = x_t)$ based on model without observation t

is called the t^{th} deleted residual of model M.

11.2 Outliers

Observation t is outlier of model M

if $\mathbb{E}(Y_t | X_t = x_t) \neq x_t^T \beta$ (for any β)

! outlier is a relative notion with respect to some model !

That is, outlier $\Leftrightarrow$ $\mathbb{E}(Y_t | X_t = x_t) = x_t^T \beta + \gamma_t^{out}$ for some β, γ_t^{out}

= correction of the model based expectation

~~obs~~ observation with unusual Y value

Possible reasons for being outlier
- data error, really unusual, ...

SOME CALCULATIONS:

- $m_{tt} > 0$ $\Leftrightarrow$ M_t^{out} is of full rank and LSE of γ_t^{out} can be calculated
- t^{th} obs. is not outlier $\Leftrightarrow$ $H_0: \gamma_t^{out} = 0$
- under normality (but ...)

$$T_t = \frac{\hat{\gamma}_t^{out}}{\sqrt{\widehat{var}(\hat{\gamma}_t^{out})}} \sim t_{n-(k+1)}$$

(LSE properties in model M_t^{out})

- $var(\hat{\lambda}_t^{out} | X) = var\left(\frac{U_t}{m_{tt}} | X\right) =$

$$= \frac{1}{m_{tt}^2} \underbrace{var(U_t | X)}_{\sigma^2 m_{tt}} = \frac{\sigma^2}{m_{tt}}$$

$var(U_t | X) = \sigma^2 m_{tt}$ in any model (M_t, M_t^{out}) where we assume $var(Y|X) = \sigma^2 I_n$ since $U = MY$, $var(U|X) = M \underbrace{var(Y|X)}_{\sigma^2 I_n} M^T = \sigma^2 M$.

Hence $\widehat{var}(\hat{\lambda}_t^{out} | X) = \frac{\hat{\sigma}^2}{m_{tt}} = \frac{MSe_{,t}^{out}}{m_{tt}} = \frac{MSe_{,(-t)}}{m_{tt}}$ ← $\hat{\sigma}^2$ from model M_t^{out}

$$\Rightarrow T_t = \frac{\hat{\lambda}_t^{out}}{\sqrt{\frac{MSe_{(-t)}}{m_{tt}}}} = \frac{U_t}{m_{tt}} \cdot \sqrt{\frac{m_{tt}}{MSe_{(-t)}}} =$$

$$= \frac{U_t}{\sqrt{MSe_{(-t)} m_{tt}}} = \text{almost standardized residual}$$

where is the difference?

12

Def 11.4 **Studentized residual**

If $m_{tt} > 0$ then the quantity

$$T_t = \frac{y_t - \hat{y}_{[t]}}{\sqrt{MSe_{(-t)}}} \sqrt{m_{tt}} = \frac{U_t}{\sqrt{MSe_{(-t)} \cdot m_{tt}}}$$

is called the studentized residual of model M.

Remark: Previous calculations have shown:

$$\frac{MSe_{(-t)}}{MSe} = \frac{n-k-(U_t^{std})^2}{n-k-1}$$

$$\overset{\text{algebra}}{\Rightarrow} T_t = \sqrt{\frac{n-k-1}{n-k-(U_t^{std})^2}} \cdot U_t^{std}$$

$\rightarrow$ T_t can be calculated without fitting M_t^{out} or $M_{(-t)}$ ∎

Lemma 11.4 **On studentized residuals**

Let $Y|X \sim N_n(X\beta, \sigma^2 I_n)$, $rank(X_{n\times k}) = k < n$.
Let further $n > k+1$. Let for given $t \in \{1, \dots, n\}$ $m_{tt} > 0$.
Then

1. The t^{th} studentized residuals T_t follows the Student t-distribution with $n-k-1$ degrees of freedom.
2. If additionally $n > k+2$ then $E(T_t) = 0$.
3. If additionally $n > k+3$ then $var(T_t) = \frac{n-k-1}{n-k-3}$.

Proof: 1. previous calculations.

2. & 3. standard properties of t-distribution.

Test for outliers

15

$M_t^{out}: y|X \sim N_n(X\beta + j_t\gamma_t^{out}, \sigma^2 I_n)$

$H_0: \gamma_t^{out} = 0$ = t^{th} not outlier of M

$H_1: \gamma_t^{out} \neq 0$ = t^{th} is outlier of M

$M: y|X \sim N_n(X\beta, \sigma^2 I_n)$

- Under H_0: $T_t \sim t_{n-k-1}$.
- Multiple testing problem (see later)
  - → e.g. Bonferroni corrections of p-values

$p_t^{Bonf} = \min\{1, p_t \cdot n\}$

p_t = standard p-value for a single test

Example: Cars 2004. 16-19

→ Hummer H2 and Toyota Prius are both outliers w.r.t. considered model

some facts about outliers 20-22

Many "outliers" = (usually) a wrong model

11.3 Leverage points

≡ observations with unusual x-values

$$H = X(X^TX)^{-1}X^T = (h_{ts})_{t,s=1,\dots,n}$$

24

Leverage

a diagonal element h_{tt} $(t=1,\dots,n)$ of the hat matrix H is called the **leverage** of the t^{th} observation.

25

Consider now a **model with intercept**

$$X = (\mathbb{1}, x^1, \dots, x^{k-1}) = \begin{pmatrix} 1 & x_{11} & \dots & x_{1,k-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{n,k-1} \end{pmatrix}$$

$$\bar{x}^1 = \frac{1}{n}\sum_{i=1}^{n} x_{i1} \quad \dots \quad \overline{x_{k-1}} = \frac{1}{n}\sum_{i=1}^{n} x_{i,k-1}$$

→ center non-intercept columns

$$\tilde{X} = (x^1 - \bar{x}^1\mathbb{1}, \dots, x^{k-1} - \bar{x}^{k-1}\mathbb{1})$$

→ $\mathcal{M}(X) = \mathcal{M}(\mathbb{1}, \tilde{X})$ & $\mathbb{1}^T\tilde{X} = 0^T$

(sum of values in each column is equal to 0)

$$H = X(X^TX)^{-1}X^T = (\mathbb{1}, \tilde{X})\left((\mathbb{1}, \tilde{X})^T(\mathbb{1}, \tilde{X})\right)^{-1}(\mathbb{1}, \tilde{X})^T$$

$\mathbb{1}^T\mathbb{1} = n$, $\mathbb{1}^T\tilde{X} = \mathbb{0}^T$, $\tilde{X}^T\mathbb{1} = \mathbb{0}$

$$= (\mathbb{1}, \tilde{X})\begin{pmatrix} n & \mathbb{0}^T \\ \mathbb{0} & \tilde{X}^T\tilde{X} \end{pmatrix}^{-1}\begin{pmatrix} \mathbb{1}^T \\ \tilde{X}^T \end{pmatrix} =$$

$$= (\mathbb{1}, \tilde{X})\begin{pmatrix} \frac{1}{n} & \mathbb{0}^T \\ \mathbb{0} & (\tilde{X}^T\tilde{X})^{-1} \end{pmatrix}\begin{pmatrix} \mathbb{1}^T \\ \tilde{X}^T \end{pmatrix} =$$

$$= \frac{1}{n}\mathbb{1}\mathbb{1}^T + \tilde{X}(\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T$$

$$\Rightarrow h_{tt} = \frac{1}{n} + (x_{t,1} - \bar{x}^1, \dots, x_{t,k-1} - \bar{x}^{k-1})(\tilde{X}^T\tilde{X})^{-1}\begin{pmatrix} x_{t,1} - \bar{x}^1 \\ \vdots \\ x_{t,k-1} - \bar{x}^{k-1} \end{pmatrix}$$

$\equiv$ square of a <u>generalized distance</u> of row t $(x_{t,1}, \dots, x_{t,k-1})^T$ from <u>centroid</u> over rows of X $(\bar{x}^1, \dots, \bar{x}^{k-1})^T$.

27

Remember: $H = QQ^T$ | Q = orthonormal vector basis of $M(X)$

$$\sum_{i=1}^{n} h_{ii} = tr(H) = tr(QQ^T) = tr(Q^TQ) =$$
$$= tr(I_k) = k.$$

$$\Rightarrow \bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{k}{n}.$$

R software

t^{th} observation is a leverage point

if $h_{tt} > \frac{3k}{n}$.

28

Leverage can be INFLUENTIAL ("dangerous"):

$$var(U_t|X) = var(Y_t - \hat{Y}_t|X) = \sigma^2 m_{tt} = \sigma^2(1-h_{tt})$$
$$t = 1, \dots, n$$

- High leverage $\Rightarrow$ low $var(U_t|X)$
$= var(Y_t - \hat{Y}_t|X)$.

$\Rightarrow$ the t^{th} fitted value is forced to be close to the observed response value

Example: Cars2004

29-32