- residuals are useful for detecting lack of fit and checking model assumptions

(Q: Why residuals can do the works?

① $Y = X\beta + \varepsilon = \hat{Y} + \hat{\varepsilon}$

② true: $Y = X_1\beta_1 + X_2\beta_2 + \varepsilon = (X_1\beta_1 + H_1X_2\beta_2) + ((I-H_1)X_2\beta_2 + \varepsilon) = \hat{Y}_{X_1} + \hat{\varepsilon}^*_{X_1}$

lack of fit → fitted: $Y = X_1\beta_1 + \varepsilon^*$

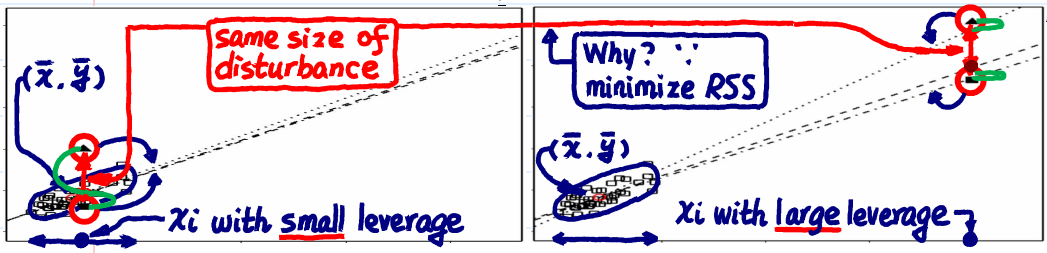
Leverage

- leverage: $h_i \equiv H_{ii}$ (Note 1. $\text{var}(\hat{\varepsilon}_i) = (1-h_i)\sigma^2$. Note 2. h_i is known before observing Y)

x_i whose h_i is large ⇒ $\text{var}(\hat{\varepsilon}_i)$ small ⇒ fitted model has to force to fit close to y_i

x_i whose h_i is small ⇒ $\text{var}(\hat{\varepsilon}_i)$ large ⇒ in this x_i, model cannot fit so well

- h_i roughly determines how close (x_i, y_i) to the regression surface (i.e., $(x_i, \hat{y}_i)$)
- observations with large h_i's should be *paid more attention*. (Q: why?)



Q: why x_i with large leverage has stronger influence on fit?

- for linear model with an intercept, its fitted model must pass the point $(\bar{x}, \bar{y})$

$$E(y) = \beta_0 + \sum_{i=1}^{p-1}\beta_i x_i = \left(\beta_0 + \sum_{i=1}^{p-1}\beta_i\bar{x}_i\right) + \sum_{i=1}^{p-1}\beta_i(x_i - \bar{x}_i) = \beta_0' + \sum_{i=1}^{p-1}\beta_i(x_i - \bar{x}_i)$$

$$\Rightarrow \beta_0' = \bar{y} \Rightarrow E(y) = \bar{y} + \sum_{i=1}^{p-1}\hat{\beta}_i(x_i - \bar{x}_i)$$

Q: why is it called leverage?

- some facts about leverage:

h_i corresponds to a Mahalanobis distance defined by $\mathcal{X}$ (X without intercept),

actually, $h_i = 1/n + [1/(n-1)](x_i - \bar{x})^T \hat{\Sigma}_{\mathcal{X}}^{-1}(x_i - \bar{x})$,

where $\hat{\Sigma}_{\mathcal{X}}$ is the estimated covariance of $\mathcal{X}$

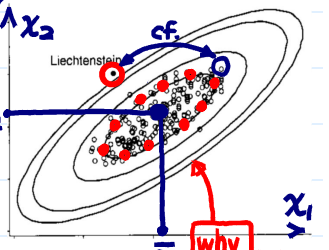
⇒ "extreme" x_i has large leverage

⇒ a little change of y_i value on point with large leverage will change the fit a lot

- $\sum_{i=1}^{n} h_i = p$, and $1 \ge h_i \ge 1/n \ \ \forall i$

(Q: what h_i's are too large? Note: an average leverage is p/n

⇒ large leverage >> p/n ⇒ "rule of thumb": if $h_i > 2p/n$, look closer)



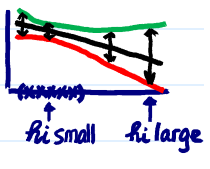
- $\text{var}(\hat{Y}) = \text{var}(HY) = H\sigma^2 \Rightarrow \text{var}(\hat{y}_i) = h_i\sigma^2$ ← Recall C.I. of prediction (LNp.5-4~5)

- (*internally*) *studentized residuals* r_i's:

because $\text{var}(\hat{\varepsilon}_i) = (1-h_i)\sigma^2$, let $r_i = \hat{\varepsilon}_i / [(1-h_i)^{1/2}\hat{\sigma}]$,

then $\text{var}(r_i) \approx 1$ (if model assumptions are correct)

usually different (When can they be identical?)

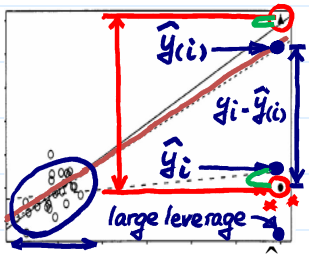


- non-constant variance removed
- sum of r_i's is not zero (cf. $\sum_i \hat{\varepsilon}_i = 0$)
- studentized residuals are preferred in residual plots
- if there is some underlying heteroscedasticity (i.e., violation of $\text{var}(\varepsilon) = \sigma^2 I$) in the errors, studentization cannot correct it
- dependence is very small in practice
- r_i is slightly correlated with $\hat{y}_i$. (cf. $\hat{\varepsilon} \perp \hat{Y}$ and $\text{cov}(\hat{\varepsilon}, \hat{Y}) = 0$)

❖ **Reading**: Faraway (2005, 1st ed.), 4.2.1 ❖ **Further reading**: D&S, 8.1

(check LNp.7-1) unusual observations → **Outlier** → could carry [misleading / useful] information

- an outlier is a point that does not fit the current model (Q: possible cause?)
  ⇒ usually, large residual (Q: why?) ← ①,② in LNp.7-2. ← plot



- typo
- different underlying systems, e.g. most data $Y = X\beta + \varepsilon$, few data $Y = X'\beta' + \varepsilon'$

- Q: is there a problem if (raw or studentized) residuals are used to detect outliers?
  ⇒ outliers may affect the fit (see plot) ← diagnose after fitting (LNp.7-1)
  ← cross-validated, leave-one-out
- **idea**: exclude i^{th} observation and re-compute the estimates to get $\hat{\beta}_{(i)}$ and $\hat{\sigma}^2_{(i)}$, where (i) denotes that the i^{th} case has been excluded. Then, consider $y_i - \hat{y}_{(i)}$, where $\hat{y}_{(i)} = x_i^T \hat{\beta}_{(i)}$. (Q: why is it better in detecting outliers?)
  ∵ $\hat{\beta}_{(i)}$ & $\hat{\sigma}^2_{(i)}$ not influenced by the ith observation

$x_i^T\beta + \varepsilon_i$ → $x_i^T\hat{\beta}_{(i)}$ ← check LNp.4-4

$$\mathrm{var}(y_i - \hat{y}_{(i)}) = \sigma^2(1 + x_i^T(X_{(i)}^T X_{(i)})^{-1}x_i)$$ (Hint. prediction of future observation)

(x_i, y_i) can be regarded as a future obs. for the fitted model developed without using the ith observation

- *jackknife* (or *externally studentized*, or *crossvalidated*) *residuals* ← bias correction

$$t_i = (y_i - \hat{y}_{(i)}) / [(1 + x_i^T(X_{(i)}^T X_{(i)})^{-1}x_i)^{1/2}\hat{\sigma}_{(i)}]$$

null distribution → which are distributed as $t_{(n-1)-p}$ under null, if model is correct and $\varepsilon \sim N(0, \sigma^2 I)$

∵ ith obs. excluded

- a simpler way to calculate t_i (avoid doing n regression)

$H_0^{(i)}$: ith obs. not an outlier
$H_A^{(i)}$: ith obs. is an outlier

raw residual $y_i - \hat{y}_i$

$$t_i = \hat{\varepsilon}_i / [(1-h_i)^{1/2}\hat{\sigma}_{(i)}] = r_i((n-p-1)/(n-p-r_i^2))^{1/2}$$

$i = 1, 2, \cdots, n$

$[X_{(i)}^T X_{(i)}]^{-1} = (X^TX)^{-1} + \frac{(X^TX)^{-1}x_i^T x_i (X^TX)^{-1}}{1-h_i}$

externally / internally → $\hat{\sigma}$ ← studentized residual

- test for outliers
  - ➢ given a specific case i, conclude an outlier if $|t_i| > t_{n-1-p}^{(\alpha/2)}$ ← 2-sided test

- ➢ in practice, a few (or all) t_i's will be tested ⇒ problem of *multiple testing* ← LNp.4-16
  need to adjust the significance level of the test accordingly

$\alpha \le P(\cup RR_i | H_0) = \alpha^* \le \sum_i P(RR_i | H_0) = n\alpha$

Bonferroni correction

H_0: no outlier in the n observations against H_1: at least one outlier

$\cap_{i=1}^n H_0^{(i)}$ ← union-intersection test → $\cup_{i=1}^n H_A^{(i)}$

rejection region $\cup_{i=1}^n RR_i$; acceptance region $\cap_{i=1}^n RR_i^c$

$$1 - \alpha^* = 1 - \mathrm{Prob}(\text{Type I error} | H_0) = \mathrm{Prob}(\text{all tests accept} | H_0)$$

$\cup RR_i$; RR_i

$$= 1 - \mathrm{Prob}(\text{at least one rejected} | H_0) \ge 1 - \sum_i \mathrm{Prob}(\text{test } i \text{ rejects} | H_0) = 1 - n\alpha$$

∵ $\alpha \le \alpha^* \le n\alpha$

⇒ conclude an outlier if $|t_i| > t_{n-p-1}^{(\alpha/2n)}$ ← if $n\alpha = 0.05$ ($\leftrightarrow \alpha = 0.05/n$), then $\alpha^* \le 0.05$

⇒ it's conservative, tends *not* to label points as outliers (especially when n large)

- some problems about outliers:
  - ➢ two or more outliers next to each other can hide each other (see an example in Lab)
  - ➢ an outlier in one model may not be an outlier in another when the variables have been changed or transformed ⇒ reinvestigate outliers when model changed ← change model
  - ➢ the error distribution may not be Normal ⇒ larger residuals may be expected. ← e.g. heavy-tailed error like Cauchy.
  - ➢ for large datasets, individual outliers are usually much less of a problem from the viewpoint of fit. In this case, it's still worth identifying outliers if these types of points are worth knowing about in the particular application. For large datasets, we need only worry about clusters of outliers. ← large n.

$\hat{\beta}, \hat{y}, \hat{\sigma}$

these few units/subjects may carry different information or phenomena

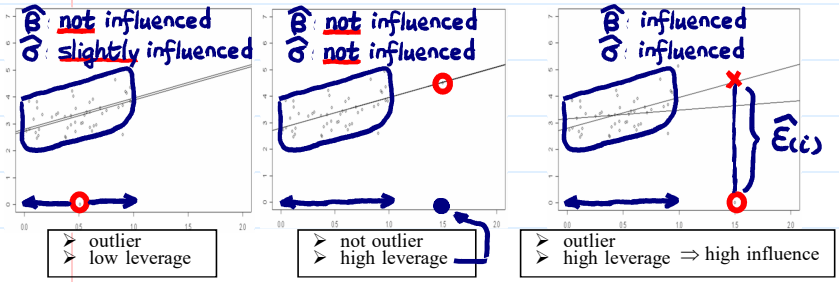
- Q: What should be done if some observations are identified as outliers?
  - check for a data entry error first ← data cleaning, 隨 ⇒ outlier | 規 ⇒ outlier
  - Examine the physical context (sometimes, outliers may have physical significance)
    - exclude the point from the analysis (keep model change data)
      - try to re-include later if model changed
      - if exclude permanently, report
    - use robust regression (future lecture; keep data change model) when outliers cannot be identified as mistakes or aberrations (for heavy-tailed error distribution)
  - dangerous to exclude them in an automatic manner ← e.g. NASA, 1985.

❖ **Reading**: Faraway (2005, 1st ed.), 4.2.2 ❖ **Further reading**: D&S, 8.1

Influential observation

(∵ leverage, residual)

- Each observations have different influence/contribution to the fitted model. Our fitted model should not change too much (i.e., robust) just because of adding/dropping a specific observation. (large or medium leverages; large or medium residuals)
- *influential point*: one whose removal from data would *cause large change* in the *fit*. (individual pattern)
- an influential point *may or may not* be an outlier and *may or may not* have large leverage but it will tend to have at least one of those two properties. ($\hat\beta, \hat y, \hat\sigma$)



Q1: what points are more influential?

Q2: what information should be included in the detection of influential points?

- measures of influence (Q: how to numerically characterize "large change in fit"?) (cross-validation)
  - change in coefficients: $\hat\beta - \hat\beta_{(i)}$ (Q: how large is large?)
  - change in fit: $\hat Y - \hat Y_{(i)}$ (Q: how large is large?)

  Note. ① $\hat\beta - \hat\beta_{(i)}$, $\hat Y - \hat Y_{(i)}$ have units and scales. We may use physical standard to examine whether the change is physically significant. (cf.) ② $(\hat\beta - \hat\beta_{(i)})/\hat\beta$, $(\hat Y - \hat Y_{(i)})/\hat Y$ ③ compare $\hat\beta_{(i)}/se(\hat\beta)$ to $\hat\beta/se(\hat\beta)$
  - Cook's statistics/distances (scale and unit free):

Note. $Var(\hat\beta) = (X^TX)^{-1}\sigma^2$

$$D_i = (\hat\beta - \hat\beta_{(i)})^T (X^TX)(\hat\beta - \hat\beta_{(i)}) / (p\hat\sigma^2)$$

(Standardization. Why? ∵ $\hat\beta$ have different units; cf. Mahalanobis distance)

$$= (\hat Y - \hat Y_{(i)})^T (\hat Y - \hat Y_{(i)}) / (p\hat\sigma^2)$$

$$= (1/p)\, r_i^2\, (h_i/(1-h_i))$$

(residual part; leverage part)

⇒ it's a combination of residual and leverage. (Q: what are the effects of residual and leverage on Cook's statistic?) (like in sampling model (LNp.5-7))

(Di>1 has been suggested) ⇒ Q: how large is large? If assume $\boldsymbol{X}$ is multivariate Normal, can do a test on D_i. However, normality may not be a reasonable assumption in practice. (null: F-dist.)

  - Others: DFFITS, Atkinson's modified Cook's statistics

- suggestion for identifying influential points: use relatively large D_i as an indication of a possible problem, then examine their $\hat\beta - \hat\beta_{(i)}$ and/or $\hat Y - \hat Y_{(i)}$.

($y_X \sim N(\mu_X, \sigma_X^2)$; $X\beta$; $(\sigma_X \leftrightarrow \mu_X)$; $(\sigma_X \leftrightarrow X)$)

❖ **Reading**: Faraway (2005, 1st ed.), 4.2.3 ❖ **Further reading**: D&S, 8.3, 8.4

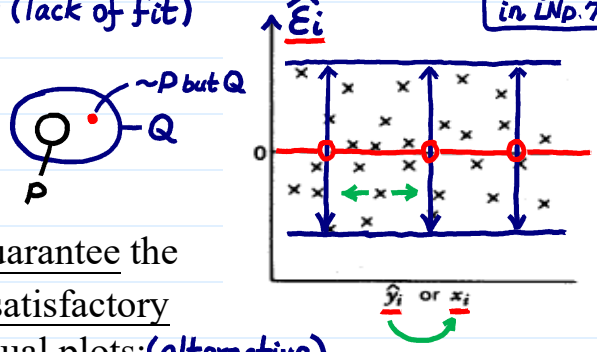
Residual plots

(better to (i) remove outliers & influential obs (ii) use studentized residuals)

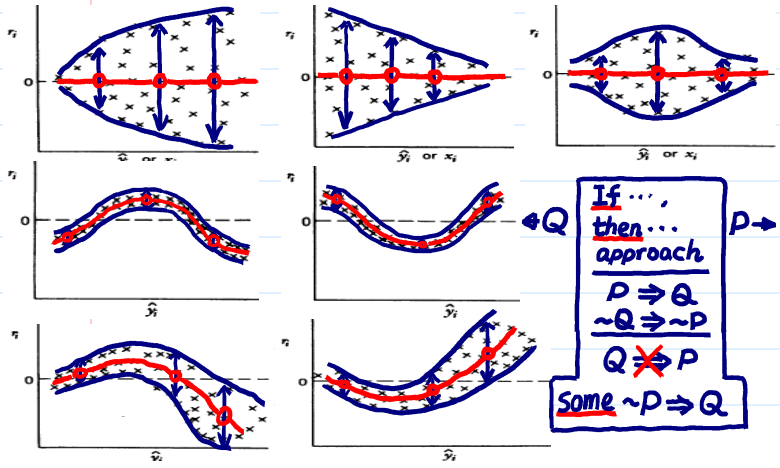
- *residual plots*: plot residuals (or absolute values of residuals) against (i) $\hat y$, (ii) x_k (for predictors in model and not in model), (iii) combination (or transformation) of x_k's, (iv) time order (if available), (v) any other quantities relevant to residuals (Q: why draw residual plots?) ← check ① ② in LNp.7-2

- in residual plots,
  - find *overall* patterns from the shape of *all* points (cf., residuals used in checking outliers or influential points ⇒ identifying *individually unusual* point)
  - check assumptions: (i) non-constant variance; (ii) incorrectly specified mean structure (i.e., $E(\boldsymbol{Y})=\boldsymbol{X}\beta$) ← too simple (lack of fit)
  - rather subjective
- a satisfactory residual plot (null plot)
  - constant variance V
  - no curvature in the mean of residuals V

Note: one satisfactory residual plot cannot guarantee the residual plots for other variables will be satisfactory



- some possible patterns in unsatisfactory residual plots: (alternative)

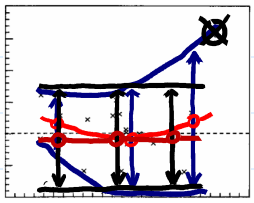


evidence of non-constant variance

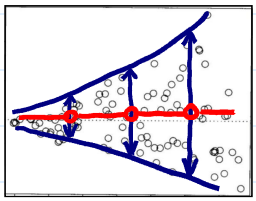
curvature in the mean of residuals ⇒ evidence of incorrectly specified mean structure

evidence of non-constant variance and incorrectly specified mean structure

If ⋯, then ⋯ approach
$P \Rightarrow Q$
$\sim Q \Rightarrow \sim P$
$Q \not\Rightarrow P$
Some $\sim P \Rightarrow Q$



- unfortunately, in real data set, it's rare the pattern is so clear (Q: what will you conclude from the residual plot on the right?)
- in models with many terms or models with complex non-linear mean structure, cannot necessarily associate shapes in a residual plot with a particular problem with the assumptions, e.g.,



(LNp.7-2) ① ②

true model: $E(Y)=|x_1|/[2+(1.5+x_2)^2]$ with constant variance

fitted model: $E(Y)=\beta_0+\beta_1 x_1+\beta_2 x_2$ ← approximate

nonlinear mean structure

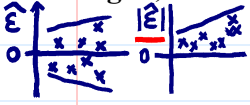
- possible remedies for unsatisfactory residual plots

unsatisfactory residual plot	plot residuals against ...		
	$\hat{y}$	x_k	time order
non-constant variance	1. weighted least square 2. transform y	1. weighted least square 2. transform y	weighted least square
curvature in mean structure	1. add extra term 2. transform y	1. add extra term of x_k 2. transform y	add term of time in model

different philosophy

change current model

❖ **Reading**: F, 4.1.1 ❖ **Further reading**: D&S, 2.5



① (LNp.2) → **Non-constant variance** ← overall pattern

11/24

- if not sure, plot absolute values of residuals against $\hat{y}$, x_k's, time order
- when non-constant variance exists, $\hat{\beta}_{OLS}$ will be more variable than the best estimates ($\hat{\beta}_{OLS}$ unbiased but not BLUE) and $\hat{\sigma}$ wrong (⇒ test and C.I. inaccurate)

$\hat{\beta}_{WLS}$, $\hat{\beta}_{GLS}$