Handout 2

RECAP

Summary of y~x relationship : $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, ε_i uncorrelated

linear

$E(\varepsilon_i) = 0$

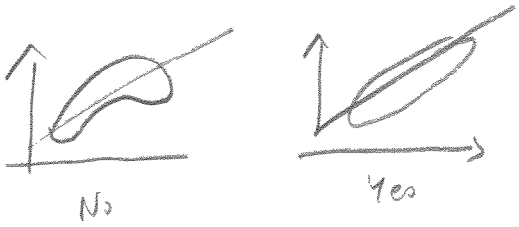
$V(\varepsilon_i) = \sigma^2$ constant

$\Rightarrow$ Fit model (β_0, β_1) to data using Least Squares

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Basic assumptions

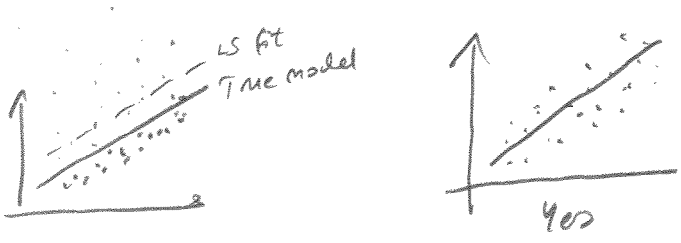
(1) Model sufficiency



No Yes

Tools: graph Fix: data transformations

(2) Symmetry of errors



No Yes

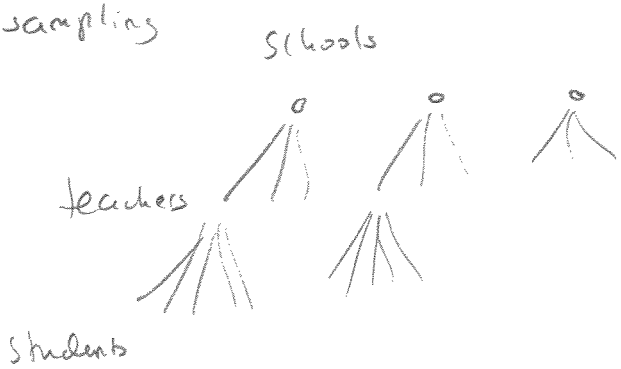
Tools: scatter, residual plots
QQ plot of residuals

Fix: data transformations
or
different fitting criteria
→ median line
ML

(3) Uncorrelated errors

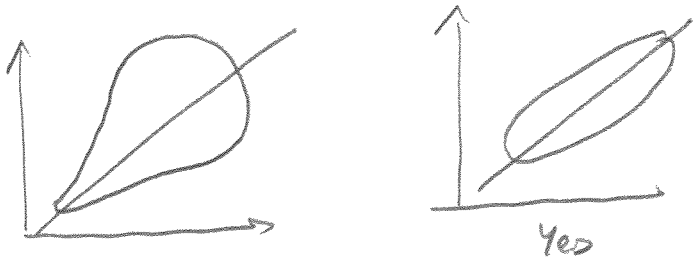
Think about the sampling process

- is there a time component
- are observations followed over time?
- same person, unit of observation measured repeatedly
- clustered sampling



If any of these → need to use other methods like time series, mixed effects modeling, ...

(4) Constant error variance



Tools: scatter, residual plots Fix: data transformations like $\log(y)$, $\sqrt{y}$, $y^{1/3}$

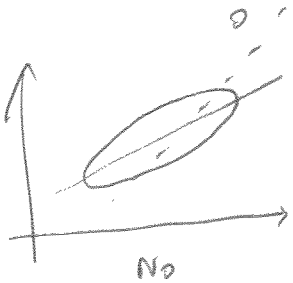
or

Weighted Least Squares (WLS)

$$\sum_{i=1} w_i (y_i - (\beta_0 + \beta_1 x_i))^2$$

(more later)

(5) No outliers



No

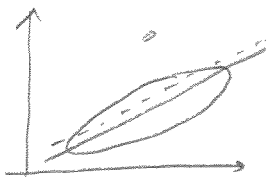
↓

bad fit

No

↓

fit looks "too good"



No

↓

you will overestimate the error scatter

→ impact on inference (more later)

But, careful about dropping too many observations.

Try to figure out why unusual.

Tools: scatter and residual plots.

Other issues: - groups in data

- missing values

omitted variables

- extrapolation

- causal interpretation

⋮

More about least Squares

$$Q = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \quad \rightarrow \text{minimize wrt } (\beta_0, \beta_1)$$

$$\Rightarrow \begin{cases} \frac{\partial Q}{\partial \beta_0} = \sum_{i=1}^{n} -2 \cdot (y_i - (\beta_0 + \beta_1 x_i)) = 0 \\ \frac{\partial Q}{\partial \beta_1} = \sum_{i=1}^{n} -2 x_i \cdot (y_i - (\beta_0 + \beta_1 x_i)) = 0 \end{cases}$$

$$\Rightarrow \boxed{\begin{aligned} \sum_{i=1}^{n} y_i &= n\beta_0 + \left(\sum_{i=1}^{n} x_i\right)\beta_1 \\ \sum_{i=1}^{n} x_i y_i &= \left(\sum_{i=1}^{n} x_i\right)\beta_0 + \left(\sum_{i=1}^{n} x_i^2\right)\beta_1 \end{aligned}}$$

The normal equations

Solve for β_0 : $\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}$ $\left(\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \ \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\right)$

eq(1)

Solve for β_1 : $\sum x_i y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2$

eq(2) w. β_0 solution plugged in

$$= \sum x_i (\bar{y} - \beta_1 \bar{x}) + \beta_1 \sum x_i^2$$

$$\Rightarrow \hat{\beta}_1 = \frac{\sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$= \frac{\text{cov}(x, y)}{\underbrace{\text{cov}(x, x)}_{= V(x)}} = \underbrace{\text{correlation}(x, y)}_{\text{linear dependency measure } \in [-1, 1]} \sqrt{\frac{V(y)}{V(x)}}$$

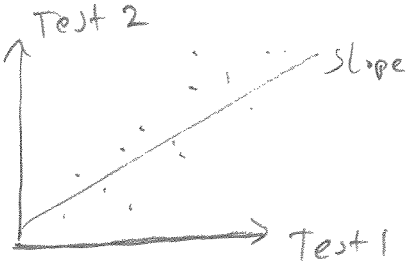
scale factor

So, the slope parameter $\hat{\beta}_1 \sim$ correlation (x, y)

As a consequence;

- regression – linear relationship summary
- If y and x are standardized, $|\beta_1| \leq 1$

(Classical misinterpretation



: meaning Test 2 score < Test 1 score on average

⇒ "Best student – scores worse on Test 2 – slacking off"

"Worst student – scores better on Test 2 – got scared"

wrong! "Regression towards the mean"

Best – nowhere to go but down

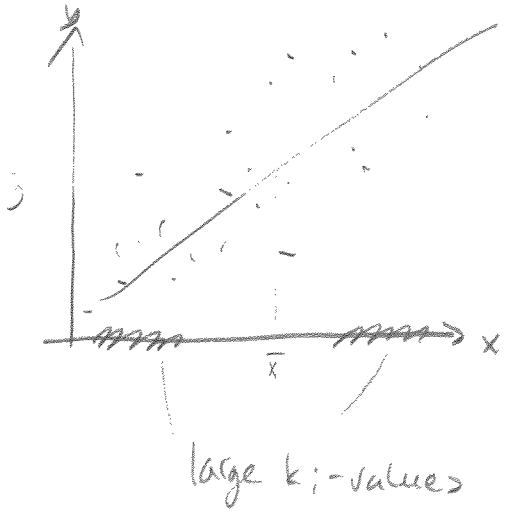
Worst – nowhere to go but up

More about the slope estimate

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\,y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \sum_{i=1}^{n}\left[\frac{(x_i-\bar{x})}{\sum_j (x_j-\bar{x})^2}\right] y_i = \sum_{i=1}^{n} k_i \cdot y_i$$

- That is, $\hat{\beta}_1$ is a linear combination of y-values where weights $k_i = \frac{(x_i-\bar{x})}{\sum(x_j-\bar{x})^2}$ are large for y-values with extreme x-values.

Note, the weights depend only on X so can be pre-computed (not data-dependent (y)) = a.k.a linear model



Properties

$$\hat{\beta}_1 = \sum k_i \cdot y_i$$

$$E(\hat{\beta}_1) = E\left(\sum k_i \cdot y_i\right) = \sum k_i\, E(y_i) = \sum k_i\,(\beta_0+\beta_1 x_i)$$

$$= \sum k_i \cdot \beta_0 + \beta_1 \sum k_i \cdot x_i = \beta_1 \quad \checkmark$$

$$\sum k_i = \sum_i \frac{(x_i-\bar{x})}{\sum_j (x_j-\bar{x})^2} = 0$$

$$\sum k_i x_i = \sum \frac{(x_i-\bar{x})\,x_i}{\sum_j (x_j-\bar{x})^2} = 1$$

So $\hat{\beta}_1$ is an <u>unbiased</u> estimate.

What about the variance?

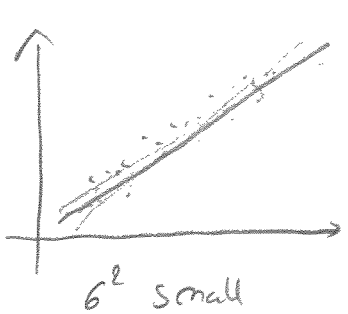
$$V(\hat{\beta}_1) = V\left(\sum k_i y_i\right) = \sum k_i^2 V(y_i) = \sum k_i^2 \sigma^2$$

↑ since y_i's are uncorrelated (because ε_i are)

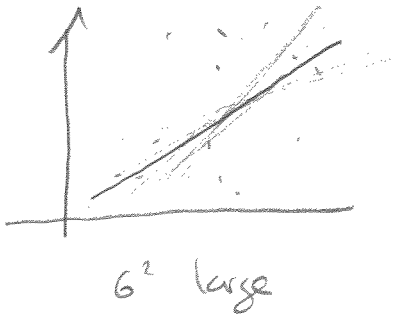
$$= \sum_i \frac{(x_i - \bar{x})^2}{\left(\sum_j (x_j - \bar{x})^2\right)^2} \sigma^2 = \frac{\sigma^2}{\sum_j (x_j - \bar{x})^2}$$

Meaning?

① $V(\hat{\beta}_1)$ increases with noise level in data (σ^2)



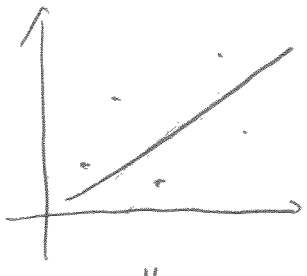
σ^2 small



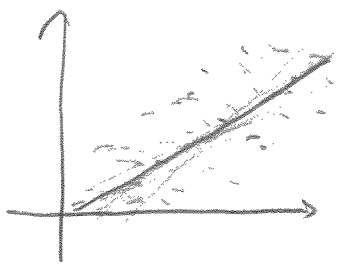
σ^2 large

② $V(\hat{\beta}_1)$ decreases $\sim \frac{1}{n}$ since $V(\hat{\beta}_1) = \frac{\sigma^2}{n\left(\frac{1}{n}\sum_j (x_j - \bar{x})^2\right)} = \frac{\sigma^2}{n \cdot c}$

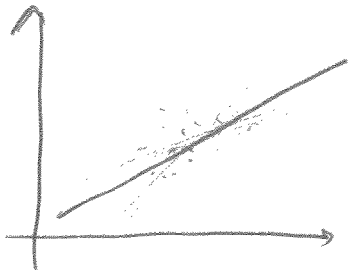
$V(x)$



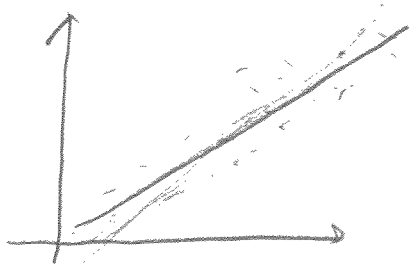
small n



③ $V(\hat{\beta}_1)$ decreases w. variance of x



$\sum(x_i - \bar{x})^2$ small



$\sum(x_i - \bar{x})^2$ large

So, easier to estimate a $y \sim x$ relationship if range of x is wide – opportunity to see how y is affected.

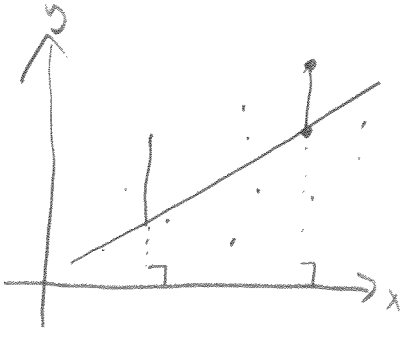
More properties & Diagnostics

- Residuals sum to 0 if intercept included in the model

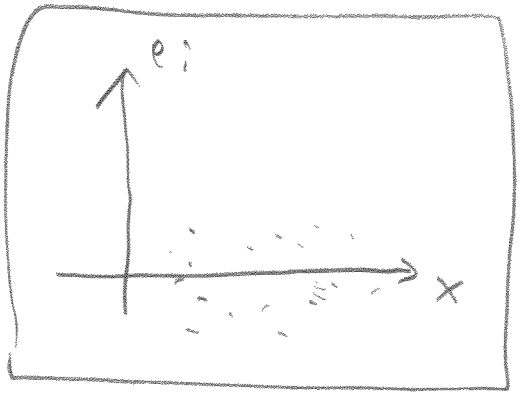
$$\sum e_i = \sum (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) = 0$$

- $\sum x_i e_i = 0$: "orthogonal projection"

Meaning – we 'used up' the linear information in x about y w. our LS fit.



→ diagnostic plot



Should see no pattern ↑ here if linear model sufficient

'proof'

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x})$$

$$= y_i - \bar{y} - \sum_j \frac{(x_j - \bar{x}) y_j}{\sum_t (x_t - \bar{x})^2} (x_i - \bar{x})$$

$$\sum x_i e_i = \sum x_i y_i - n\bar{x}\bar{y} - \sum_j \frac{(x_j - \bar{x}) y_j}{\sum_t (x_t - \bar{x})^2} \sum_i x_i (x_i - \bar{x}) = 0$$

- $\sum \hat{y}_i e_i = 0$, fitted values $\perp$ the residuals

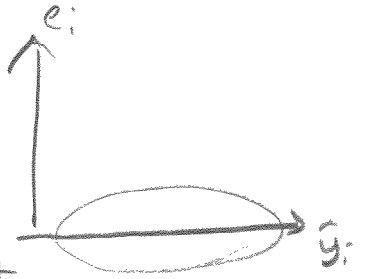
$\hat{y}_i$ = a linear function of x_i — part of y_i explainable from x

e_i = orthogonal to x_i — part of y_i not explainable from x

Diagnostic plot

Good fit ⇒ no patterns here →

o/w model insufficient



Side note — what about $\hat{\beta}_0$?

$$E(\hat{\beta}_0) = \beta_0$$

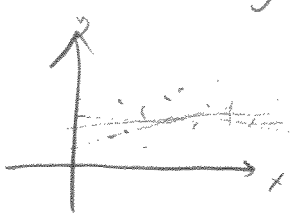
$$V(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_j - \bar{x})^2} \right)$$

↑ if $\bar{x}^2$ large compared w spread $\sum (x_j - \bar{x})^2$, ie if x near constant

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sigma^2 \bar{x}}{\sum (x_j - \bar{x})^2}$$: large negative if x near constant

Why? Because if x near constant models $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \mu_i + \varepsilon_i$

so we don't need 2 parameters



Fitted values

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x}) = \bar{y} + \sum_j k_j y_j (x_i - \bar{x})$$

$$= \sum_j \left(\frac{1}{n} y_j + k_j (x_i - \bar{x}) y_j \right) = \sum_j \left(\frac{1}{n} + k_j (x_i - \bar{x}) \right) y_j = \sum_j h_{ij} \cdot y_j$$

= a weighted average of y-values.

Note, if x is constant, all $h_{ij} = \frac{1}{n}$ and $\hat{y}_i = \bar{y}$!

Let's look at how much y_i contributes to its own fitted value, $\hat{y}_i$:

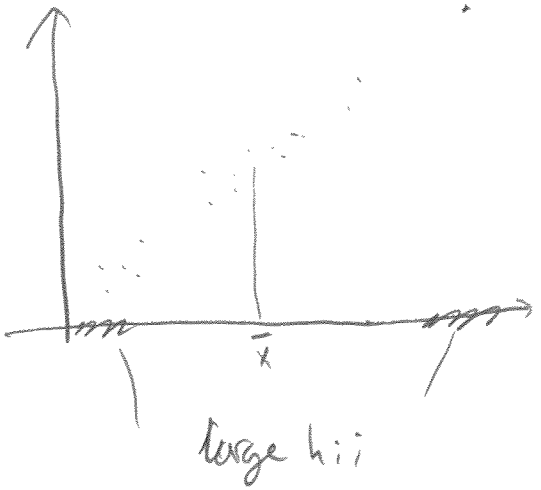
$$h_{ii} = \frac{1}{n} + k_i (x_i - \bar{x}) = \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2} \right]$$ Leverage

$h_{ii} \sim \Theta\left(\frac{1}{n}\right)$

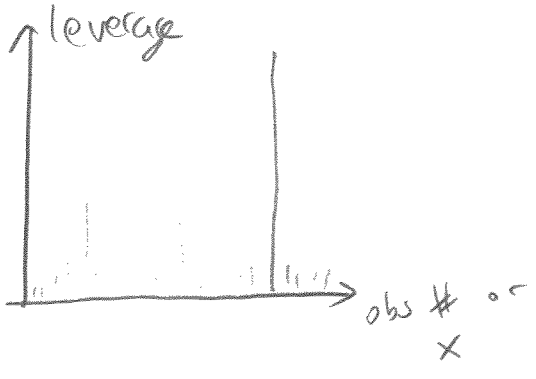
~ is large if x_i 'extreme' - far from $\bar{x}$

~ if $h_{ii} >> h_{ij}$, $\hat{y}_i$ dominated by y_i.

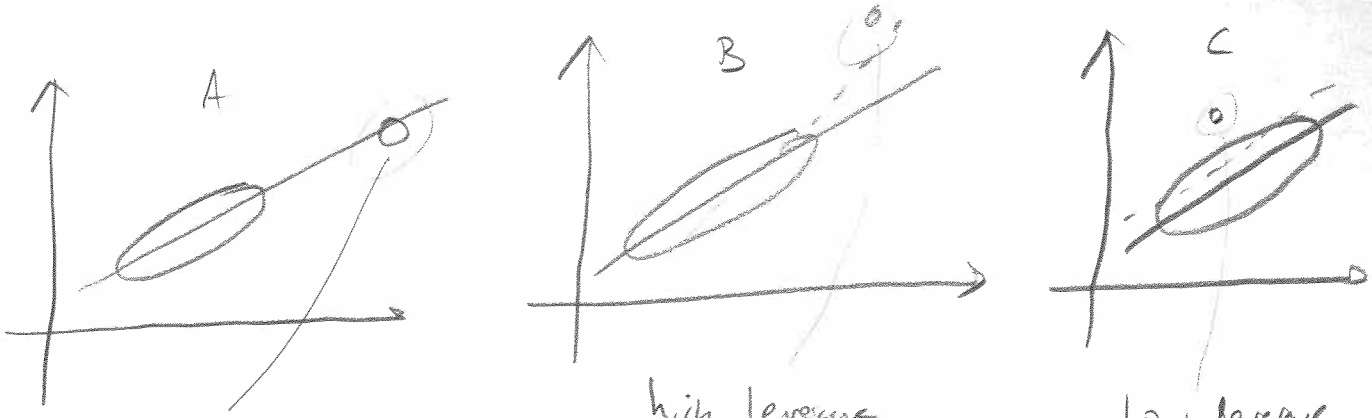
Discuss.



Diagnostic plot



Problem If observation i has a high leverage value, a contamination (outlier) at this location can have a huge impact on the analysis



high leverage, but not influential

high leverage and influential

low leverage but creates a bias in fit "pure outlier"

What to do?

(A) (B) Do analysis w & w/o outlier – discuss impact

(C) – " – " – – check if inference affected

Spotting the problem

(A) – leverage plot, but does not show up in residual!

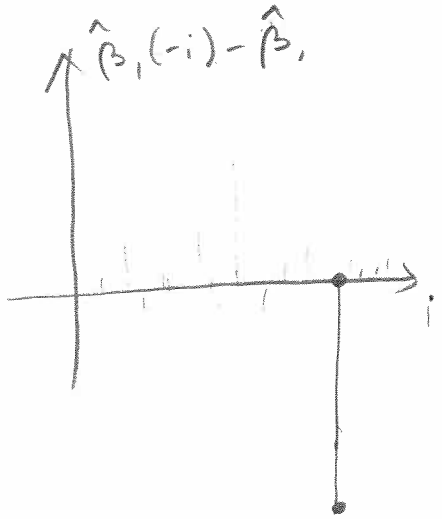
(B) – leverage & residual plot

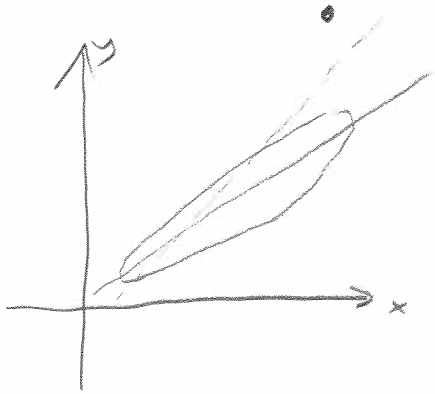
(C) – residual plot

Other diagnostic tools

Dropping an observation - examine refit.

① Impact on slope

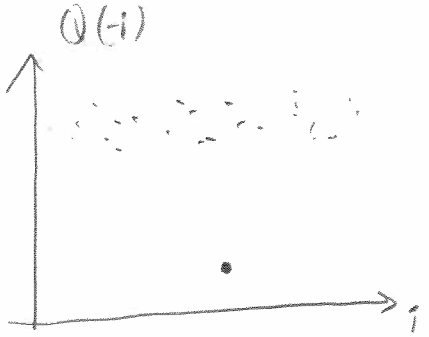


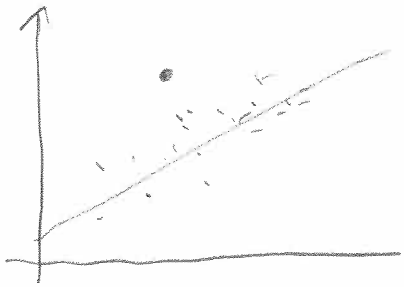


② Impact on LS criterion $\sum (y_i - (\hat{\beta}_0 - \hat{\beta}_1(x_i))^2 = LS = Q$

vs

$\sum (y_i - (\hat{\beta}_0(-i) - \hat{\beta}_1(-i)x_i))^2 = LS(-i) = Q(-i)$





③ Combining residuals, leverage & re-fitting ...

Example Cook's $D = D_i = \left(\frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}\right)^2 \left(\frac{h_{ii}}{1-h_{ii}}\right) \cdot \frac{1}{2}$

(more later)