Handout 2

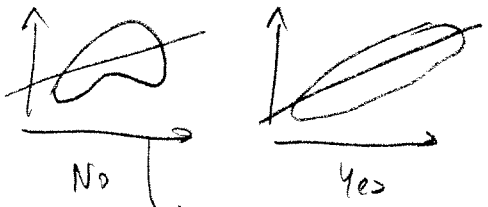
RECAP

Summary $y \sim x$ relationship $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, ε_i uncorrelated, $E(\varepsilon_i) = 0$ and $V(\varepsilon_i) = \sigma^2$

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

- $\sum_{i=1}^{n}$: equal contribution of all terms
- squared error

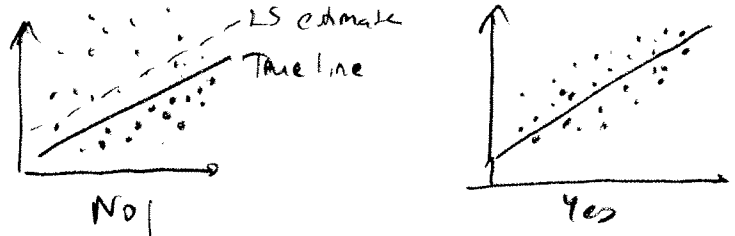
BASIC ASSUMPTIONS ① Model sufficiency



Check? Plot, plot, plot!

→ Solution? Try data transformations (e.g. x^2 instead of x, $\log(y)$ instead of y)

② Symmetry of error scatter



Check? • QQplot of residuals from LS fit.

• Scatter plot y on x may already show a clear asymmetry of course.

→ Solution? Transformation may help → or use a different fitting criteria

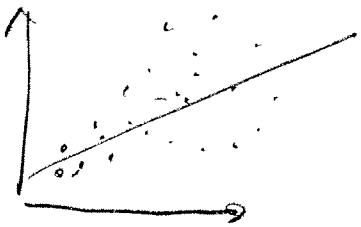
③ Uncorrelated errors

Check? Think about the sampling process.

- Is the data sampled sequentially over time? ⇒ Time Series
- Or samples nested? (e.g. students from the same class / patients from the same hospital)

Solution ⇒ Mixed effects or Hierarchical Linear Model

(4) Constant variance of error scatter



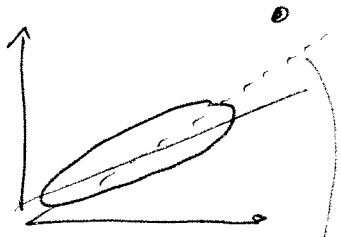
Check? Plot, plot, plot; both y on x and residuals

Solution ⇒ Try some data transformations $\log(y)$, $\sqrt{y}$, $\sqrt[3]{y}$, $\frac{1}{y}$ ....

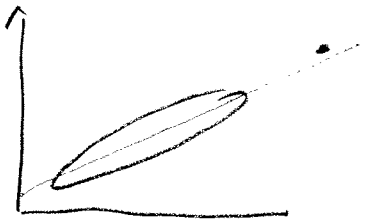
⇒ If that doesn't work ⇒ <u>Weighted</u> Least Squares

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

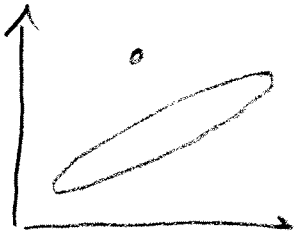
(5) <u>No outliers</u>



Effect: bad fit



Effect: fit looks "too good"



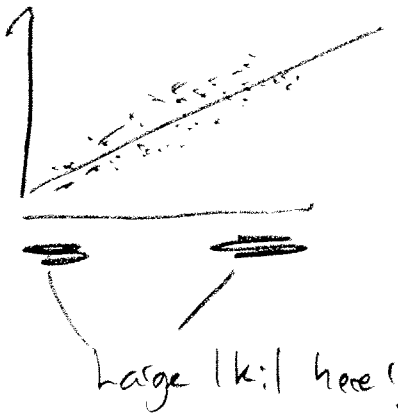
Effect: you will overestimate the error scatter (more on this later)

LS estimates → Properties

Have $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, $\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_j (x_j - \bar{x})^2} = \sum_i k_i y_i$

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

Which observations contribute? The ones w. x_i far from $\bar{x}$!



Note, LS estimates are unbiased

$E(\hat{\beta}_1) = E(\sum k_i y_i) = \sum k_i E(y_i) = \sum k_i (\beta_0 + \beta_1 x_i)$ [Since $E(y_i) = \beta_0 + \beta_1 x_i + E(\varepsilon_i)$]

$= (\sum k_i) \cdot \beta_0 + \beta_1 (\sum k_i x_i) = \beta_1$

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

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

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

So, if we obtain many new data sets (x, y), on average we'll obtain the true regression line.

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

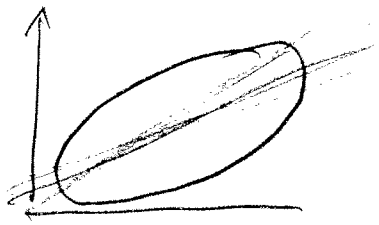
↑ since y_i's are independent (because ε_i's are)

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

Meaning? ① $V(\hat{\beta}_1)$ increases with the noise level

→ more uncertainty, more difficult to identify the slope

σ small

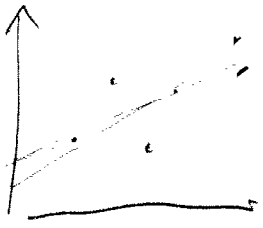


σ large

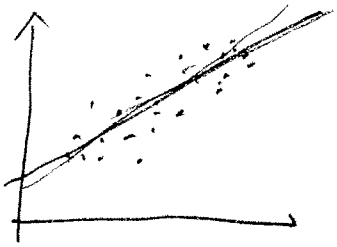
② $V(\hat{\beta}_1)$ decreases $\sim \frac{1}{n}$ w. the sample size n

$\sum_{j=1}^{n} (x_i - \bar{x})^2$ = sum of n terms $\approx n \cdot c$, c some constant

⇒ More data, easier to identify the slope.



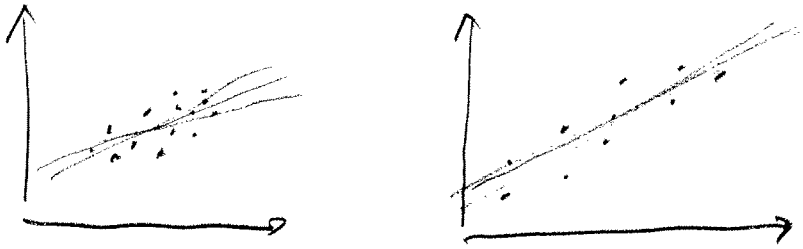
n small



n large

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

⇒ The more spread in x we have, the easier it will be to identify the $y \sim x$ relationship



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

More properties

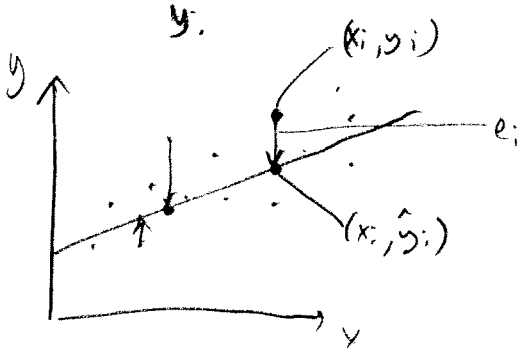
- Residuals sum to 0, $\sum e_i = 0$ (if β_0 included in the model)

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

Meaning ⇒ We used up all the linear information in x about y w. our LS fit

$$\left[e_i = y_i - \bar{y} - \underbrace{\left(\frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} \right) (x_i - \bar{x})}_{\hat{y}_i} \right] = 0$$

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



⊗ projection ↓ is orthogonal (⊥) to the x-axis

More properties

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

$\hat{y}_i$ = linear function of x_i , the part that's explainable from x

e_i = orthogonal to x_i, the part of y_i that's not explainable from x

Side note $V(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right)$ increases if $\bar{x}^2$ large compared w. spread $\sum(x_i-\bar{x})^2$

→ meaning if x near constant

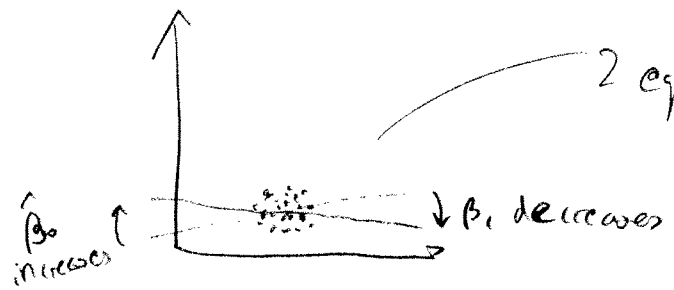
Also $Cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\bar{x}\sigma^2}{\sum(x_i-\bar{x})^2}$, large negative if x near constant

So - i) x near constant, the model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \simeq \mu + \varepsilon_i$$

and we don't need 2 parameters β_0, β_1.

Equally good fits are obtained w. lots of regression line, meaning there is a lot of uncertainty in the estimates of β_0, β_1



What about the fitted values?

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

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

= a weighted average of y-values.

Note, i) x is constant $h_{ij} = \frac{1}{n}$ since $x_i = \bar{x}$ $\forall i$

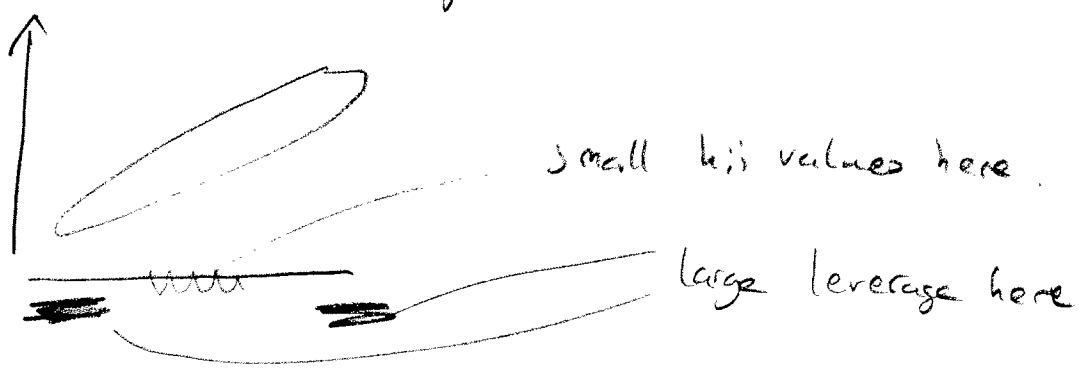
$\Rightarrow \hat{y}_i = \bar{y}$ if x has no information about y

Observation i contributes h_{ii} to its own fitted value

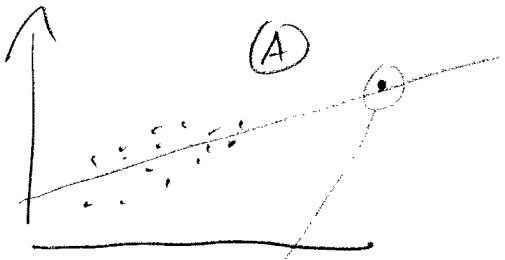
→ Extreme $h_{ii} = 1$ $h_{ij} = 0$ $\forall j \neq i$ $\Rightarrow \hat{y}_i = y_i$

When we use no part of the other observations to help learn the $y \sim x$ relationship

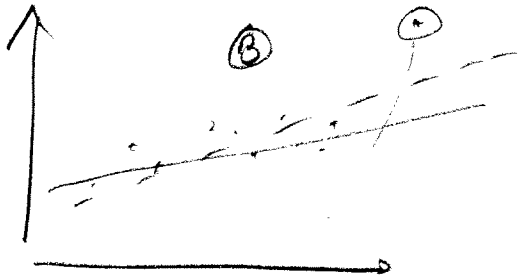
h_{ii} "the leverage" $= \left(\frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2} + \frac{1}{n}\right)$



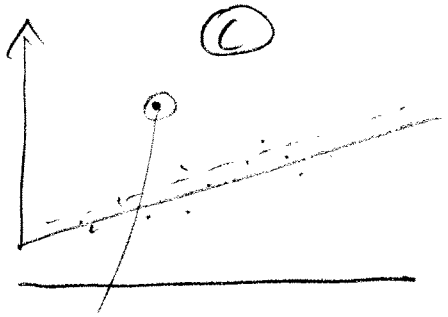
Problem: If observation i has high leverage, a contamination (or outlier) at this location can have a huge impact on the analysis.



High leverage, but not influential



High leverage, and influential



Low leverage, but creates a bias in the fit. "Pure outlier"

Case A: Do the analysis with and without the observation ⊙ & B included → discuss impact

Case C: Do the analysis with & without ... ⊙

→ Check if inferences are affected

Go back to the data source and see if you can figure out **Why** the outlier is present.

More about fitted values;

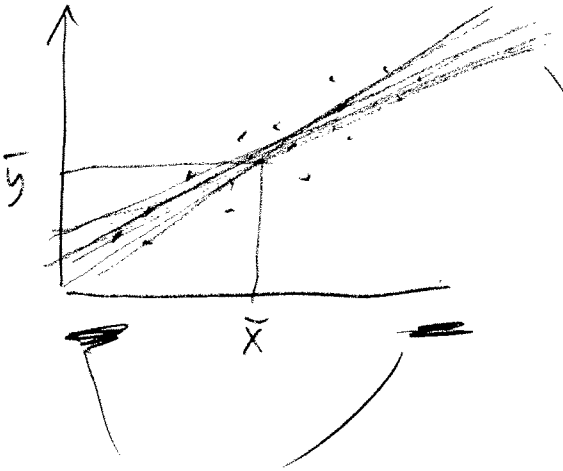
$$\boxed{V(\hat{y}_i) = \sigma^2 h_{ii}}$$

$$\left[V(\hat{y}_i) = V(\hat{\beta}_0 + \hat{\beta}_1 x_i) = V\left(\bar{y} + \hat{\beta}_1 (x_i - \bar{x})\right) = V(\bar{y}) + (x_i - \bar{x})^2 V(\hat{\beta}_1) \right.$$

$$\left(\text{Since } Cov(\bar{y}, \hat{\beta}_1) = Cov\left(\sum \tfrac{1}{n} y_i, \sum k_j y_j\right) = \sum_j \tfrac{1}{n} k_j V(y_j) = \tfrac{\sigma^2}{n} \sum k_j = 0 \right)$$

$$\left. = \frac{\sigma^2}{n} + \frac{\sigma^2 (x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2} = \sigma^2 h_{ii} \right]$$

Meaning; At locations of high leverage, the regression line can fluctuate a lot from sample to sample



Different $(\hat{\beta}_0, \hat{\beta}_1)$ estimates obtained w. samples from $(x_i, y_i = \beta_0 + \beta_1 x_i + \varepsilon_i)$

h_{ii} large here

What about the residuals e_i?

$$V(e_i) = \sigma^2(1-h_{ii})$$

$$e_i = y_i - \hat{y}_i = y_i - \sum h_{ij} y_j$$

$$Cov(e_i, e_i) = Cov(y_i - \hat{y}_i, y_i - \hat{y}_i) = Cov(y_i, y_i) + Cov(\hat{y}_i, \hat{y}_i) - 2Cov(y_i, \hat{y}_i)$$

$$= \sigma^2 + \sigma^2 h_{ii} - 2Cov(\hat{y}_i + e_i, \hat{y}_i) = \sigma^2 + \sigma^2 h_{ii} - 2\sigma^2 h_{ii} = \sigma^2(1-h_{ii})$$

($\perp$ $Cov(e_i, \hat{y}_i) = 0$ — LS properties)

$$Cov(e_i, e_j) = Cov(y_i - \hat{y}_i, y_j - \hat{y}_j) = Cov(y_i, y_j) + Cov(\hat{y}_i, \hat{y}_j) - Cov(y_i, \hat{y}_j) - Cov(y_j, \hat{y}_i)$$

($Cov(y_i, y_j) = 0$)

$$= Cov\left(\sum h_{il} y_l, \sum h_{jl} y_l\right) - Cov(y_i, \sum h_{jl} y_l) - Cov(y_j, \sum h_{il} y_l)$$

$$= \sum h_{il} h_{jl} \sigma^2 - h_{ji}\sigma^2 - h_{ij}\sigma^2 = \sigma^2 h_{ij} - 2\sigma^2 h_{ij} = -\sigma^2 h_{ij}$$

MEANING:

1. residuals are correlated (but errors ε_i are not)
2. residuals have non-constant variance $V(e_i) = \sigma^2(1-h_{ii})$ (errors $V(\varepsilon_i) = \sigma^2$)
3. residual variance is small in high-leverage regions, since observations here drive the fit

Note residuals are not directly comparable since their variances differ

⇒ better to look @ and use

Standardized residuals

$$\tilde{e}_i = \frac{e_i}{\sqrt{1-h_{ii}}}$$

Note $V(\tilde{e}_i) = \sigma^2$

So far

$$E(\hat{\beta}_1) = \beta_1, \quad V(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

$$E(\hat{y}_i) = \beta_0 + \beta_1 x_i, \quad V(\hat{y}_i) = \sigma^2 h_{ii}$$

$$E(e_i) = 0, \quad V(e_i) = \sigma^2 (1-h_{ii})$$

Note, all the variances involve one extra parameter $\boxed{\sigma^2}$

If we don't know it, we can't make inferences

⇒ We the data to estimate this "nuisance" parameter as well.

- RSS = residual sum of squares (or SSE error sum of squares)

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- MSE = mean squared error $= \frac{RSS}{n-2} = \hat{\sigma}^2$

is an unbiased estimate of the error variance σ^2

Note denominator $\underline{\underline{n-2}}$; we're already used the data for two parameter estimates ($\hat{\beta}_0$ & $\hat{\beta}_1$) so only $n-2$ degrees of freedom remains

Next lecture $\Rightarrow$ R^2, and the F & t-tests.