Handout 2

RECAP

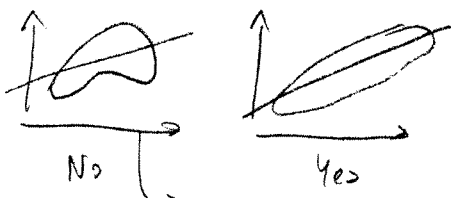
Summary $y \sim x$ relationship $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, ε_i uncorrelated, $E(\varepsilon_i) = 0$ and $V(\varepsilon_i) = \sigma^2$

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

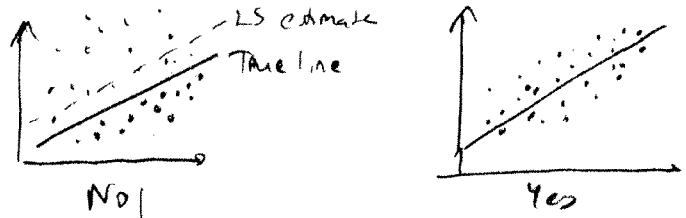
① Model sufficiency



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

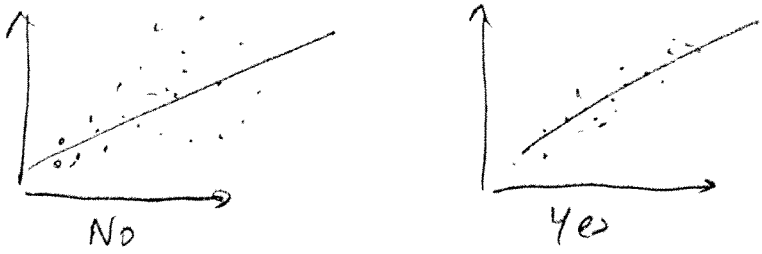
Check? Think about the sampling process.

Solution:

- 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



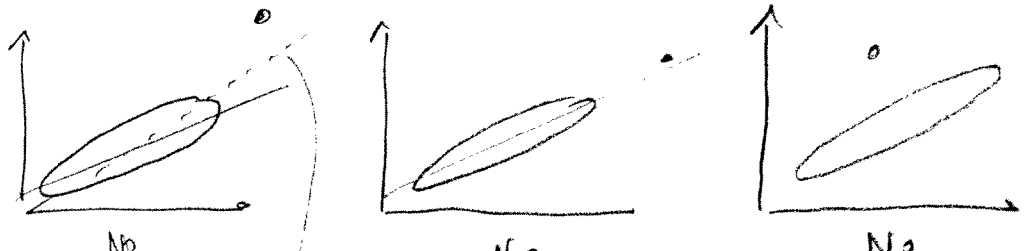
Check? Plot, plot, plot ; both y on x and residuals

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)

Now, can also write errors $e_i = (y_i - (\beta_0 + \beta_1 x_{i1} + \dots \beta_{l-1} x_{i,l-1}))$

in vector form $\underset{\sim}{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \underset{\sim}{y} - \underset{\sim}{\underset{\sim}{X}}\underset{\sim}{\beta}$

$$\Rightarrow Q = \underset{\sim}{e}'\underset{\sim}{e} = (\underset{\sim}{y} - \underset{\sim}{\underset{\sim}{X}}\underset{\sim}{\beta})' (\underset{\sim}{y} - \underset{\sim}{\underset{\sim}{X}}\underset{\sim}{\beta})$$

We want to minimize Q w.r.t respect to $\underset{\sim}{\beta}$.

$\Rightarrow$ Matrix derivative

$$\frac{\partial Q}{\partial \underset{\sim}{\beta}} = -2\underset{\sim}{\underset{\sim}{X}}'(\underset{\sim}{y} - \underset{\sim}{\underset{\sim}{X}}\underset{\sim}{\beta}) = 0$$

inner derivative

$$\Rightarrow \boxed{(\underset{\sim}{\underset{\sim}{X}}'\underset{\sim}{\underset{\sim}{X}})\underset{\sim}{\beta} = \underset{\sim}{\underset{\sim}{X}}'\underset{\sim}{y}}$$ Normal equations on matrix form.

"Interpretation" $(\underset{\sim}{\underset{\sim}{X}}'\underset{\sim}{\underset{\sim}{X}}) \sim Cov(\underset{\sim}{\underset{\sim}{X}})$, where diagonal of $\underset{\sim}{\underset{\sim}{X}}'\underset{\sim}{\underset{\sim}{X}}$ Matrix is the $Var(X_1), Var(X_2) \dots$ and off-diagonal elements are the covariances $Cov(X_1, X_2) \dots$

So if some X's are linearly dependent, or close to
$\rightarrow (\underset{\sim}{\underset{\sim}{X}}'\underset{\sim}{\underset{\sim}{X}})^{-1}$ may not exist, or be numerically unstable

$(\underset{\sim}{\underset{\sim}{X}}'\underset{\sim}{y}) \sim Cov(\underset{\sim}{\underset{\sim}{X}}, \underset{\sim}{y})$ — measure of linear dependence between the

(4) $(\bar{X}'\bar{X})^{-1}$ exists $\Rightarrow \hat{\beta} = (\bar{X}'\bar{X})^{-1}\bar{X}'y$

Problems: If $(\bar{X}'\bar{X})^{-1}$ numerically unstable, meaning X's are related $\not\Rightarrow$ difficult to interpret the meaning of the slopes $\hat{\beta}_1 \ldots \hat{\beta}_{p-1}$ - they are not just measuring the contribution of one x, if x's are related.

Back to simple linear regression $\Rightarrow$ (4a)

Properties

- unbiased estimates

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

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

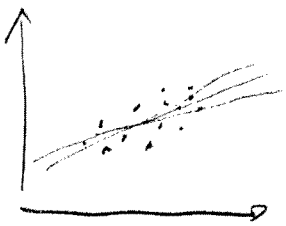
where $k_i = \frac{(x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2}$

That is, $\hat{\beta}_1$ is a weighted average of y-values!

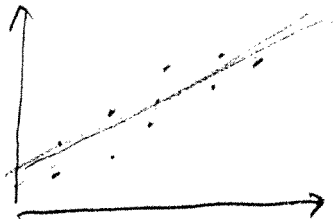
Note, extreme x_i's, far from $\bar{x}$, contribute the most to the slope estimate

③ $V(\hat{\beta}_1)$ decreased 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



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

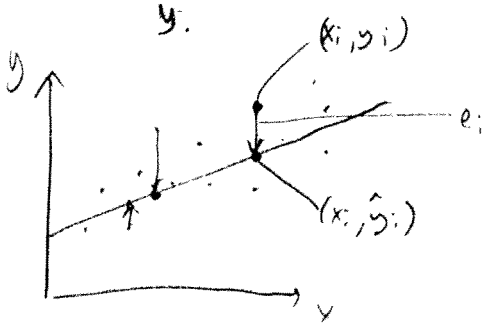
More properties

- 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

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



⊗ 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 : Othogonal 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

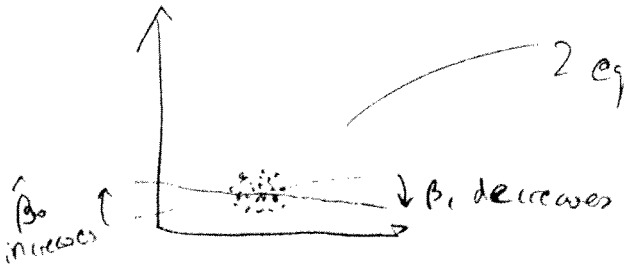
Also $Cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\bar{x}\sigma^2}{\sum (x_i - \bar{x})^2}$, large negative if x near constant

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



More about fitted values;

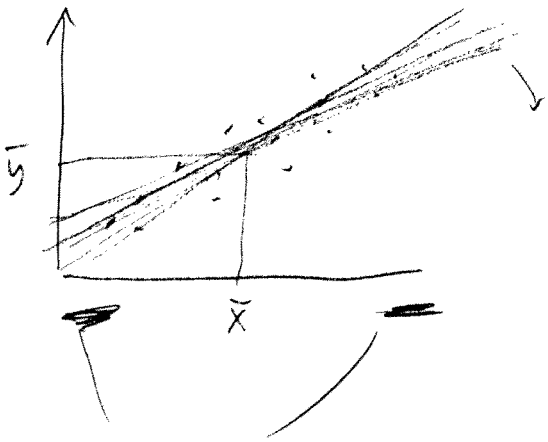
$$\boxed{V(\hat{y}_i) = \sigma^2 h_{ii}}$$

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

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

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

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)

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

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

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

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

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

Now: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \Rightarrow E(y_i) = \beta_0 + \beta_1 x_i$ since $E(\varepsilon_i) = 0$ (assumption 2)

$E(\hat{\beta}_1) = E(\sum k_i y_i) = \sum k_i E(y_i)$ (by uncorrelated ε_i, assumption 3)

$= \sum k_i (\beta_0 + \beta_1 x_i) = \beta_0 \sum k_i + \beta_1 \frac{\sum (x_i - \bar{x}) x_i}{\sum_j (x_j - \bar{x})^2} = \beta_1 \frac{\sum (x_i - \bar{x})(x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2} = \beta_1$ ✓

$\beta_0 \sum k_i$ // 0 since $\sum k_i = \sum \frac{(x_i - \bar{x})}{\sum (x_j - \bar{x})^2} = \frac{\sum x_i - n\bar{x}}{\sum (x_j - \bar{x})^2} = 0$

(and similarly for β_0)

Meaning, on average LS gives us the correct β_0, β_1.

Facts

- residuals sum to 0, $\sum e_i = 0$ (i.e. β_0 estimated)

  Follows from 1st normal equation

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

  Meaning, we used up all the (linear) information that x has on y w.r.t. our LS regression fit.

  Follows from 2nd normal equation (NE).

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

"Decomposition" $y = \hat{y} \oplus e$

$\hat{y}$: explainable; e: orthogonal to x.

Let's look closer at the fitted values $\hat{y}$.

$\hat{y}_i$ (fitted value for the ith observation y_i)

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

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

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

That is, $\hat{y}_i$ is a weighted average of all y_j's

(a linear model, linear in the data y)

Question: how much does y_i impact the overall fit @ x_i?

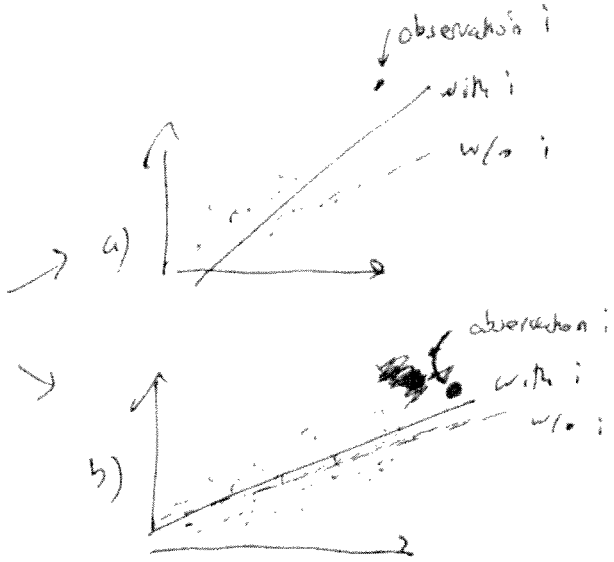
Weight for y_i in $\hat{y}_i$ is $\boxed{h_{ii}}$ (also called the **leverage**)

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

Note
- $h_{ii} \sim o\left(\frac{1}{n}\right)$
- if x_i "extreme", far from $\bar{x}$, h_{ii} is large
- if x_i close to $\bar{x}$, h_{ii} is small.

(NB) $h_{ii} >> h_{ij}$: $\hat{y}_i$ is dominated by ... y_i !!

What does this mean?



These observations drive the fit of the regression line!

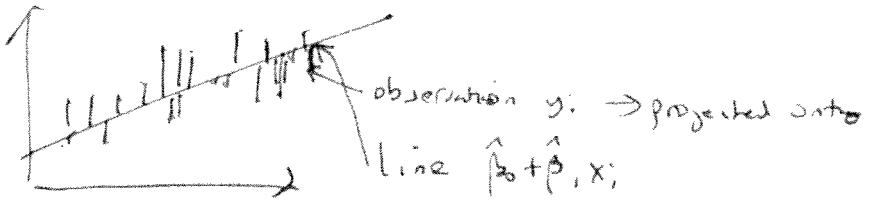
In both a) & b) observation i has high leverage h_{ii}, and could potentially influence the fit. In case a), the ith observation is an ~~influencer~~ influential outlier, producing a poor fit to the rest of the data.
In case b), the slope estimate is not much affected, but we get an exaggerated sense of the goodness of fit of the model. (more later).

We collect the h_{ii}, h_{ij} in a matrix called H, or the hat-matrix

$$\underset{n\times 1}{\hat{y}} = \underset{n\times n}{H}\,\underset{n\times 1}{y} : \begin{pmatrix}\hat{y}_1\\ \hat{y}_2\\ \vdots\\ \hat{y}_n\end{pmatrix} = \begin{pmatrix} h_{11} & h_{12} & h_{13} & \cdots & h_{1n}\\ h_{21} & h_{22} & h_{23} & \cdots & h_{2n}\\ \vdots & & & & \\ h_{n1} & & \cdots & & h_{nn}\end{pmatrix}\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{pmatrix}$$

- This H matrix is a projection matrix.
- It takes the data y, and projects it onto a function of $\vec{x}$

e.g. simple linear regression



- The projection is orthogonal to x ⊥

- Note, it can be precomputed using only x-values.

More LS properties

- Variance $(\hat{\beta}_1) = V(\hat{\beta}_1) = V\left(\sum k_i y_i\right) = \sum k_i^2 V(y_i)$ ($\overset{b}{=}$ uncorrelated ε_i)

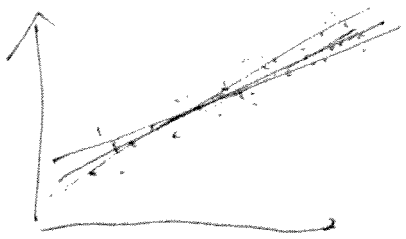
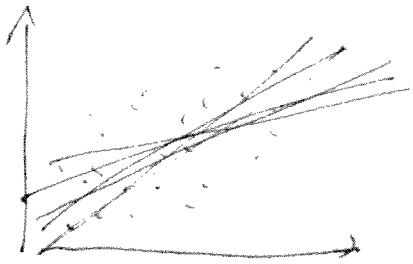
$= \sum k_i^2 \sigma^2 = \sum \frac{(x_i - \bar{x})^2}{\left(\sum_j (x_j - \bar{x})^2\right)^2} \sigma^2$

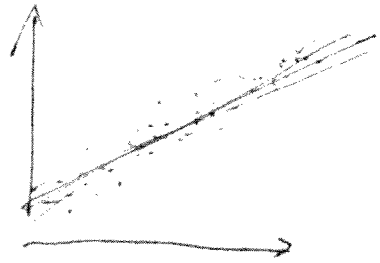
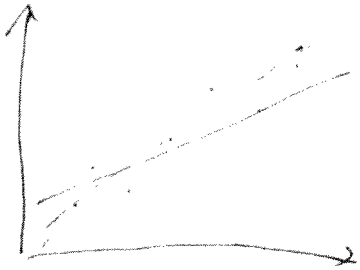
($V(y_i) = \sigma^2$ since the only thing random in y_i is ε_i)

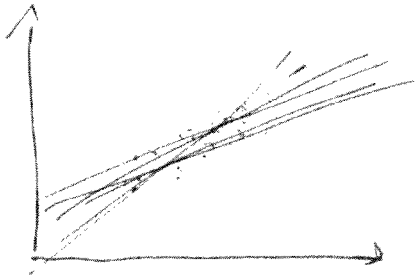
$= \frac{\sigma^2}{\sum_j (x_j - \bar{x})^2}$

Note

- LS is the minimum variance & unbiased linear estimate, any other weights $c_i \neq k_i \Rightarrow$ larger variance.
- 3 factors impact the variance of $\hat{\beta}_1$.
  1) random scatter variance σ^2 (more noise → less precision in $\hat{\beta}_1$ estimate)
  2) sample size n
  3) spread of x

1)  vs 

2)  vs 

3)  vs 

- Variance of fitted values, variance of the regression line

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

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \Rightarrow V(\hat{y}_i) = V(\hat{\beta}_0 + \hat{\beta}_1 x_i) = V(\bar{y} + \hat{\beta}_1 (x_i - \bar{x}))$$

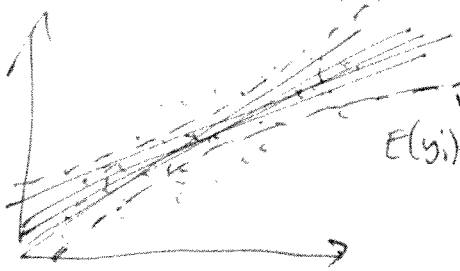
($\bar{y} \perp \hat{\beta}_1$ since $\text{cov}(\bar{y}, \hat{\beta}_1) = \text{cov}\left(\frac{1}{n}\sum y_i, \sum k_j y_j\right) = \sum \frac{1}{n} k_i \text{cov}(y_i, y_i)$

$= \frac{\sigma^2}{n} \sum k_i = 0$) (by uncorrelated ε_i)

$$\Rightarrow V(\bar{y} + \hat{\beta}_1 (x_i - \bar{x})) \overset{\Downarrow}{=} V(\bar{y}) + (x_i - \bar{x})^2 V(\hat{\beta}_1) = \frac{\sigma^2}{n} + \frac{\sigma^2 (x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}$$

$$= \sigma^2 h_{ii}$$

This makes sense: A small change for extreme x_i observations will change the slope. However, since the line always goes through the center of the data $(\bar{x}, \bar{y})$, the observations in the center will have stable fitted values



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

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

  Since $E(e_i) = 0$ $\quad V(e_i) = E(e_i^2)$ or $Cov(e_i, e_i)$

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

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

  $\hat{y}_i, e_i$

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

  $= 0$ by orthogonal projection

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

Meaning: residual variance is small for high leverage observations. That makes sense: for high leverage observations, the regression line is pulled towards the observation, so errors are forced to be small.

One more parameter

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $E(\varepsilon_i) = 0$ (Symmetry) $V(\varepsilon_i) = \sigma^2$ (constant variance)

LS → estimates of β_0, β_1 , but we also want to estimate the amount of scatter around the line ⇒ this is a measure of how well the line summarizes the data y.

Scatter variance σ^2 is estimated by the observed residual variance

⇒ RSS (residual sum of squares) = SSE (Error sum of squares) $= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

(= minimized Q)

⇒ MSE (the mean squared error)

$$= \frac{RSS}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \hat{\sigma}^2$$

why 2?

- Usually, a variance estimate is $\frac{\sum (y_i - \bar{y})}{n-1}$, but here our model is not an average value for $\bar{y}$ (a one-parameter model) like $\beta_0 + \beta_1 x_i = \hat{y}_i$, which is based on 2 fixed parameters!

Variance Decomposition

Use of $\hat{\sigma}^2$.

Total 'variance' $SS_T = \sum(y_i - \bar{y})^2$ of the data

(Total sum of squares)

2 parts

SS_{reg}

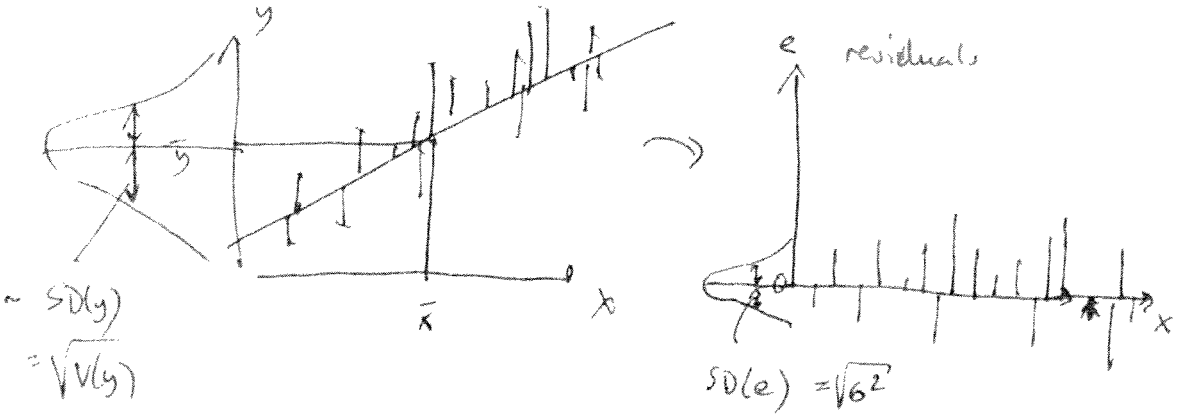
$\sum(\hat{y}_i - \bar{y})^2$

- explainable part
- Variation of line around the mean

RSS (or SS_E)

$\sum(y_i - \hat{y}_i)^2$

- unexplainable part
- scatter around the line



Effect of regression → reducing the unexplained variability

$V(y) \rightarrow \sigma^2$

% of variability explained : $\boxed{R^2 = 1 - \frac{RSS}{SS_T}} = \frac{SS_{reg}}{SS_T} \in [0,1]$

Called the Multiple R-squared