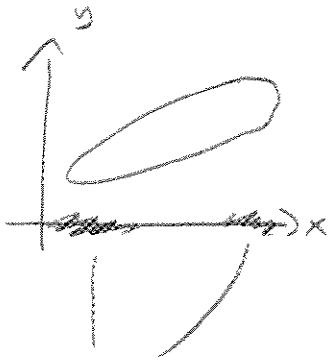
HANDOUT 3

Least Squares estimates

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

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

linear combo of all y-values



Extremes in x dominate the fit

large k_i values

Properties $E(\hat{\beta}_1) = \beta_1$ unbiased - so average across multiple data sets we get the model

but we only see one data in practise — how much can we be off in a typical case?

Variance $V(\hat{\beta}_1) = \frac{\sigma^2}{\sum_j (x_j - \bar{x})^2}$

$\hat{\beta}_1 = \beta_1$ (give or take) $\sqrt{V(\hat{\beta}_1)}$ essentially

captures uncertainty

3 sources of uncertainty

① $V(\hat{\beta}_1)$ ↑ as σ^2 ↑ – i.e. more noise in the data makes estimation more difficult

② $V(\hat{\beta}_1)$ ↓ as $\frac{1}{n}$ – i.e. more data makes estimation easier

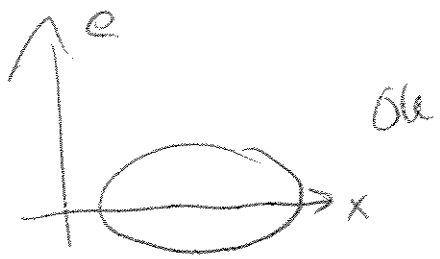
③ $V(\hat{\beta}_1)$ ↓ as $V(X)$ ↑ – i.e. more spread in X (as captured by $\Sigma(x_i - \bar{x})^2$) makes estimation easier.

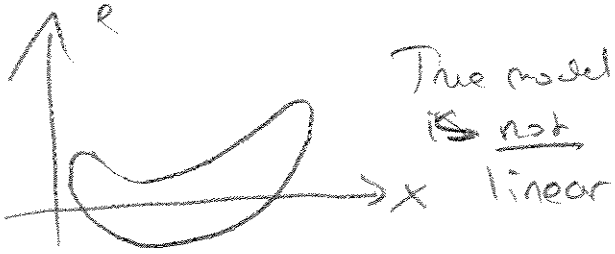
Orthogonality properties

- $\sum_i x_i \cdot e_i = 0$ ⇒ i.e. residuals are orthogonal to x

e and x uncorrelated

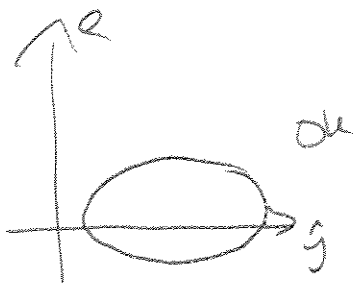
⇒ we used up all the linear information in x about y

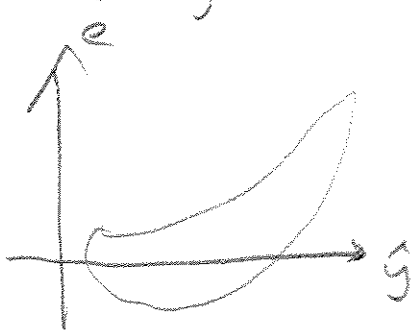




- $\sum_{i=1}^{n} \hat{y}_i \cdot e_i = 0$, i.e. residuals are orthogonal to fitted values

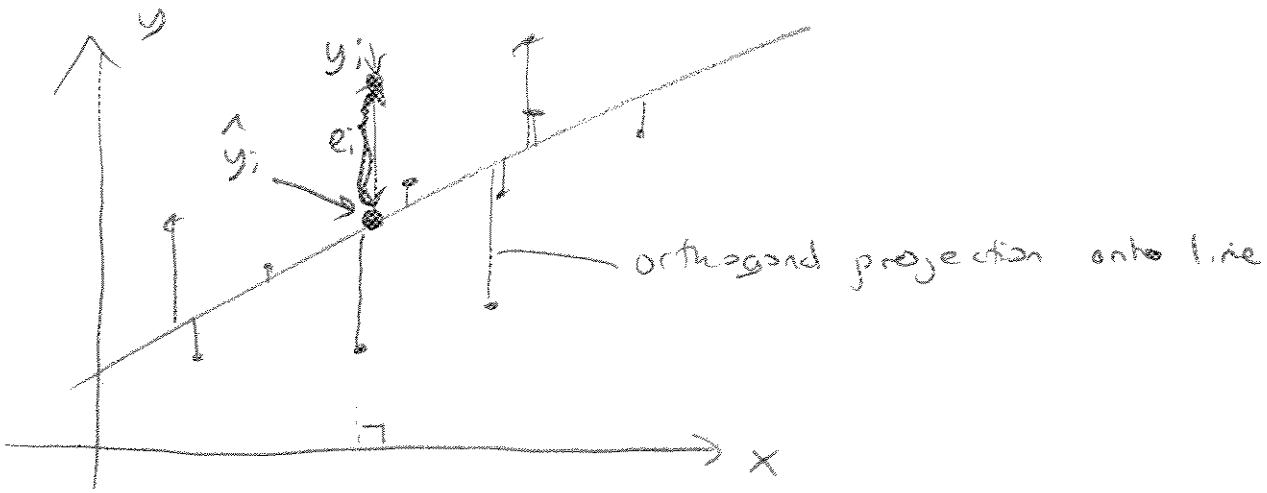
$\Rightarrow$ e and $\hat{y}$ are uncorrelated



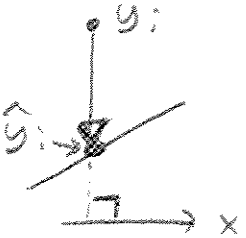


model insufficient.

$\rightarrow$ follows from $\hat{y}$ = a function of x and X is uncorrelated with e.



So $\hat{y}_i = y_i$ projected onto the line orthogonal to x



Leverage

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

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

$$= \sum_j \frac{y_j}{n} + k_j (x_i - \bar{x}) \cdot y_j$$

$$= \sum_j \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_k (x_k - \bar{x})^2} \right) y_j$$

$$= \sum_j h_{ij} \cdot y_j \quad = \text{linear combo of } \underline{\text{all}} \text{ y-values}$$

All x's constant $\Rightarrow h_{ij} = \frac{1}{n} \Rightarrow \hat{y}_i = \bar{y}$

so - no information in x, use mean of y to summarize the data

If x_i is extreme $\Rightarrow h_{ii} >> h_{ij} \Rightarrow \hat{y}_i \approx y_i$

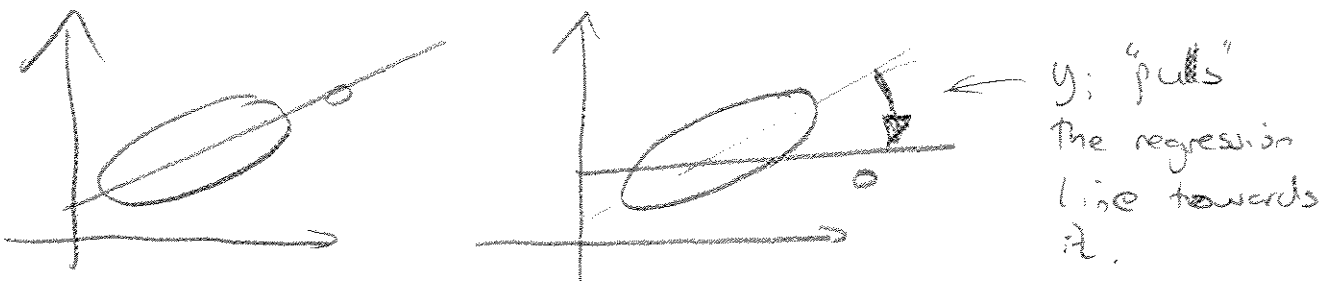
Then we learn nothing from the rest of the data.

Ideally, we want h_{ij} ~ comparable for all j so everyone contributes a little.

If $h_{ii} >> h_{ij}$, i.e. y_i dominates the fit at x_i

- we call x_i a point of <u>high leverage</u>

<u>Problem</u>: If x_i = high leverage <u>and</u> corresponding y_i is extreme ⇒ poor fit



Where does the leverage show up in properties of fitted values?

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \implies E(\hat{y}_i) = E(y_i) = \beta_0 + \beta_1 x_i$ since $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased

$$\downarrow$$

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

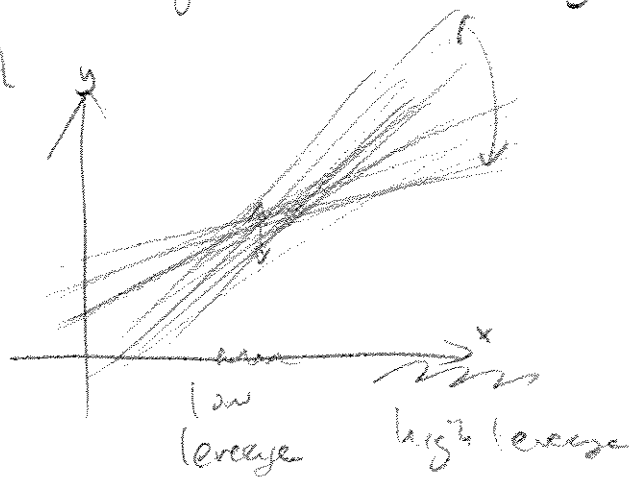
$$= V(\bar{y}) + (x_i - \bar{x})^2 V(\hat{\beta}_1)$$ because $\bar{y}$ and $\hat{\beta}_1$ are uncorrelated

$$\left[Cov(\bar{y}, \hat{\beta}_1) = Cov\left(\frac{1}{n}\sum y_j, \sum k_i \cdot y_i\right) = \sum_i \frac{k_i}{n} Cov(y_i, y_i) \text{ [since } y_i \text{ and } y_j \text{ are uncorrelated]} = \sum_i \frac{k_i}{n} V(y_i) = \frac{\sigma^2}{n} \sum_i k_i = \frac{\sigma^2}{n} \cdot 0 = 0 \right]$$

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

h_{ii}: leverage

So, at points of high leverage, lots of uncertainty in estimate $\hat{y}_i$, since small fluctuations in y_i can really alter the fit



What about the residuals?

$e_i = y_i - \hat{y}_i$, $E(e_i) = 0$

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

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

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

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

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

$Cov(e_i, \hat{y}_i) = 0$ since e_i uncorrelated with $\hat{y}_i$

So, residual variance smaller at points of high leverage since whatever y_i is it pulls the regression line towards it — controls the residual magnitude.

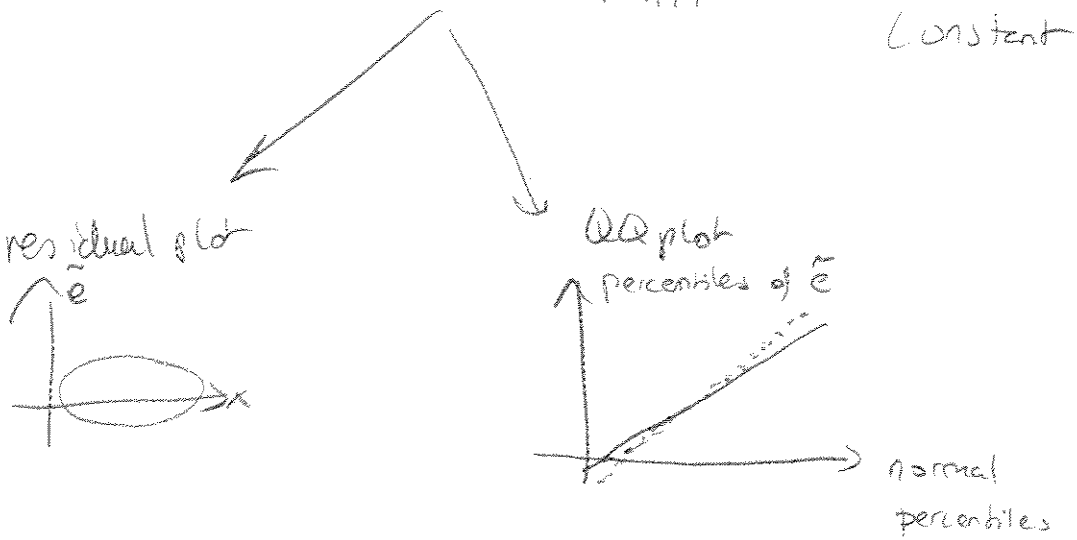
Note also $Cov(e_i, e_j) = -\sigma^2 h_{ij}$ so residuals are correlated

True errors, ε_i	Residuals from fit, e_i
$E(\varepsilon_i) = 0$	$E(e_i) = 0$
$V(\varepsilon_i) = \sigma^2$ constant	$V(e_i) = \sigma^2(1 - h_{ii})$ not constant
$Cov(\varepsilon_i, \varepsilon_j) = 0$ uncorrelated	$Cov(e_i, e_j) = -\sigma^2 h_{ij}$ correlated.

⇒ Careful when looking at residual plots!

Since $V(e_i) = \sigma^2(1-h_{ii})$, can only directly compare residuals if we adjust for the non-constant variance

→ Standardized residuals $\tilde{e}_i = \frac{e_i}{\sqrt{1-h_{ii}}}$, $V(\tilde{e}_i) = \sigma^2$ (constant)



The nuisance parameter, σ^2

$V(\hat{\beta}_1)$, $V(\hat{y}_i)$, $V(e_i)$ all involve σ^2 — so we can't really do inference without knowing it.

Use data to estimate both β and σ^2

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $E(\varepsilon_i)$, $V(\varepsilon_i) = \sigma^2$ assumed model

So σ^2 = variance of <u>true</u> errors ε.

We have the LS fit $\hat{\beta}_0$ and $\hat{\beta}_1$, and the residuals e.

Can we use e to estimate σ^2?

If we knew $\varepsilon_1, \ldots, \varepsilon_n$ (true errors)

(a) $\hat{\sigma}^2 = V(\varepsilon) = \frac{1}{n-1}\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})$

Here, we use Mean Squared Error, MSE

(b) $MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{RSS}{n-2}$

Residual Sum of Squares

Why $n-2$ not $n-1$?

- In (a), the $\varepsilon_i - \bar{\varepsilon}$ share <u>one</u> parameter $\bar{\varepsilon}$, i.e. we constrain the sum of $\varepsilon_i - \bar{\varepsilon}$ to 0 $\Rightarrow$ $n-1$ degrees of freedom
- In (b), $y_i - \hat{y}_i$ share <u>two</u> parameters, $\hat{\beta}_0$ and $\hat{\beta}_1$, $\Rightarrow$ $n-2$ degrees of freedom

MSE $\Rightarrow$ unbiased estimate of σ^2

Key facts used:
- $V(Z) = E(Z^2) - (E(Z))^2$
- $Cov(Z,W) = E(ZW) - E(Z)E(W)$

$$E(RSS) = E\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right) = E\left(\sum e_i^2\right)$$

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

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

$$= \sum_{i=1}^{n} V(y_i) + V(\hat{y}_i) + 2(E(y_i))^2 - 2Cov(\hat{y}_i + e_i, \hat{y}_i) - 2E(y_i)E(\hat{y}_i)$$

$$= \sum_{i=1}^{n} V(y_i) + V(\hat{y}_i) + 2(E(y_i))^2 - 2(E(y_i))^2 - 2Cov(\hat{y}_i, \hat{y}_i) - 2Cov(e_i, \hat{y}_i)$$

$Cov(\hat{y}_i, \hat{y}_i) = V(\hat{y}_i)$; $Cov(e_i, \hat{y}_i) = 0$

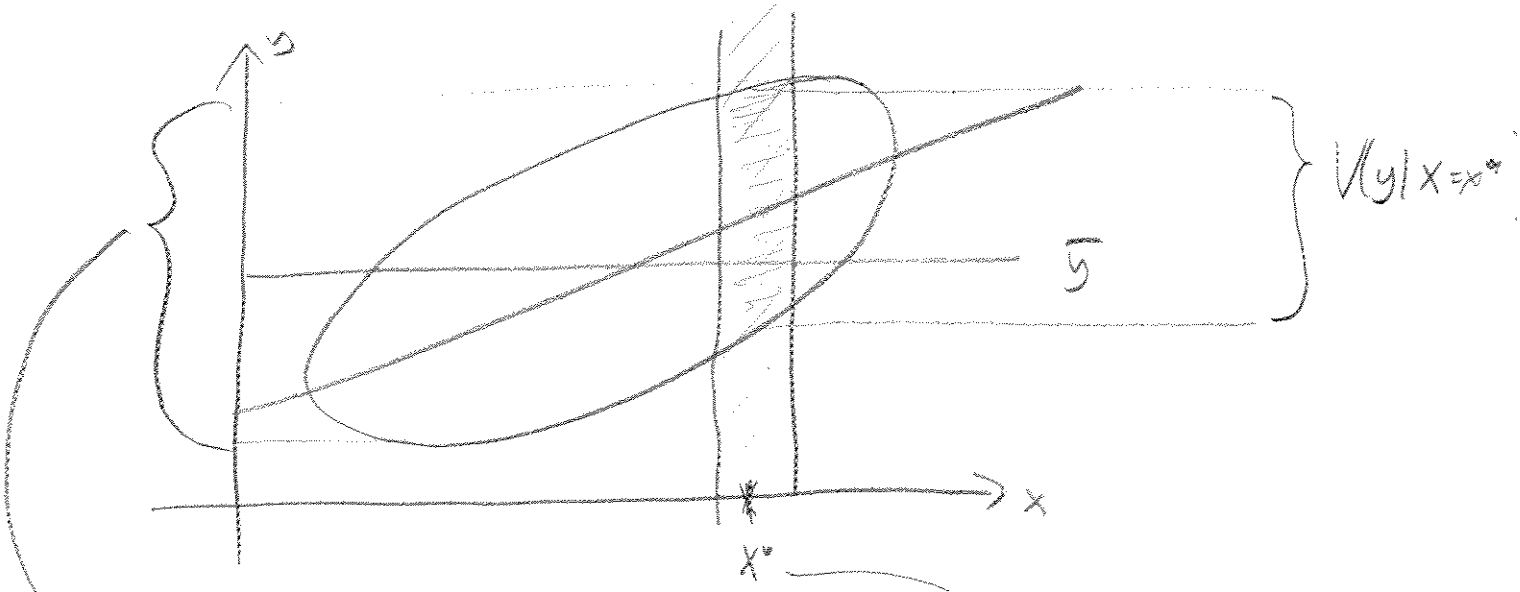
$$= \sum_{i=1}^{n} V(y_i) - V(\hat{y}_i)$$

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

$$= \sigma^2(n-2)$$

$$\Rightarrow \boxed{E(MSE) = \sigma^2} \quad \star$$

Variance Decomposition



Spread among y-values around $\bar{y}$ ⇒ Marginal distribution for y

⇒ Marginal variance $V(y)$

In neighborhood of x^* – spread among y in strip around regression value $\hat{\beta}_0 + \hat{\beta}_1 x^*$

$\sim V(y|x=x^*)$

= Conditional variance

If y and x are unrelated ⇒ $V(y) \approx V(y|x)$

– doesn't help to condition

If $y = \beta_0 + \beta_1 x$ exactly ⇒ $V(y|x) = 0$ – no variability left!

(Note, in LS we assume $V(\varepsilon) = \sigma^2$ constant, or equivalently that $V(y|x=x^*)$ independent of actual x^* value)

Summarize our fit → are we close to

- unrelated x and y

or

- perfect fit $y = \beta_0 + \beta_1 x$ exactly ?

Two quantities to compare:

① $RSS = SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ Spread among y_i around regression line

Residual sum of squares — Error sum of squares

two different names

② $SS_T = \sum_{i=1}^{n} (y_i - \bar{y})$ Spread among y_i around mean of y.

Total sum of squares

Note, since we fit the regression line to data to minimize RSS, RSS is <u>always</u> smaller than SS_T

The R-squared

$$R^2 = \frac{SS_T - RSS}{SS_T} = \frac{\text{reduction in spread of } y}{\text{total spread in } y}$$

= " % of variability in y explained by the regression "