More about fitted values;

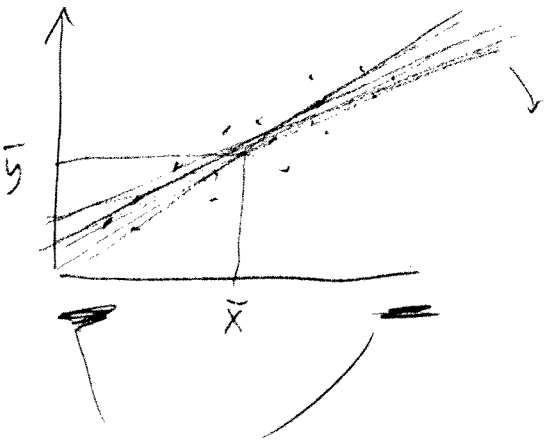
$$\boxed{V(\hat{y}_i) = \sigma^2 h_{ii}}$$

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

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

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

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)

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

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

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

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

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

Now: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \Rightarrow E(y_i) = \beta_0 + \beta_1 x_i$ 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 give 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**)

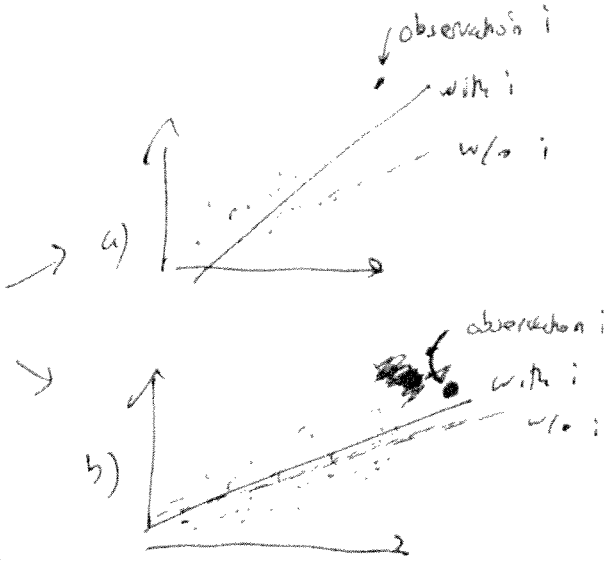
$$h_{ii} = \left(\frac{1}{n} + k_i (x_i - \bar{x})\right) = \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}\right)$$

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.

(N) $h_{ii} >> h_{ij}$ $\hat{y}_i$ is dominated by ... I. II

What does this mean?

These observations drive the fit of the regression line!



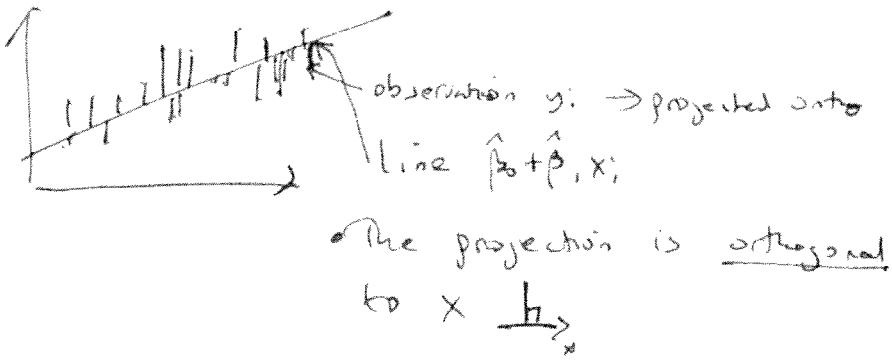
In both a) & b) observation i has high leverage h_{ii}, and could potentially influence the fit. In case a), the ith observation is an ~~influence~~ influential outlier, producing a poor fit to the rest of the data. In case b), the slope estimate is not much affected, but we get an exaggerated sense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

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

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

  e.g. simple linear regression



- The projection is orthogonal to x

- Note, H can be precomputed using only X-values.

More LS properties

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

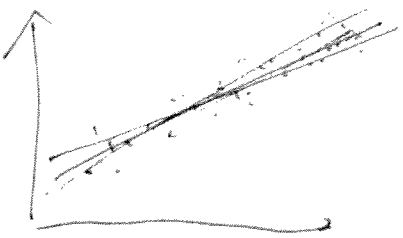
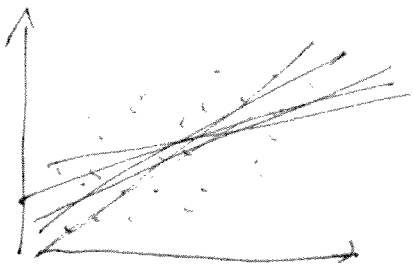
$= \sum k_i^2 \sigma^2 = \dfrac{\sum (x_i - \bar{x})^2}{\left(\sum_j (x_j - \bar{x})^2\right)^2} \sigma^2$

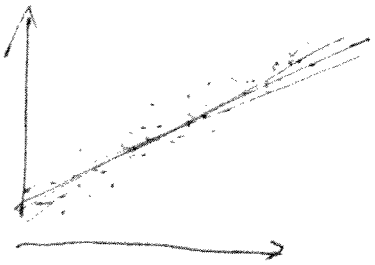
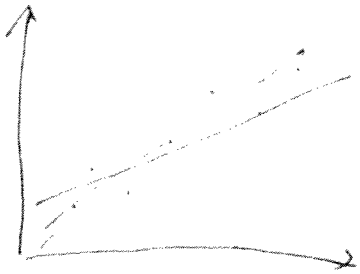
($V(y_i) = \sigma^2$ since the only thing random in y_i is ε_i)

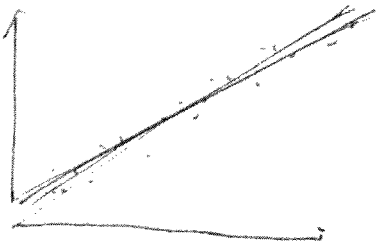
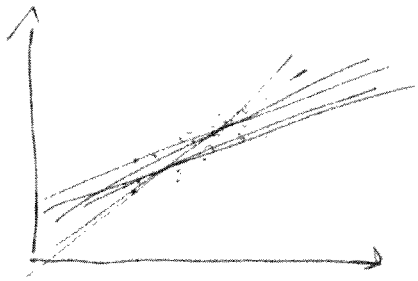
$= \dfrac{\sigma^2}{\sum_j (x_j - \bar{x})^2}$

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

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

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

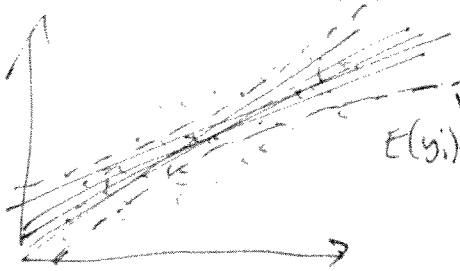
($\bar{y} \perp \hat{\beta}_1$ since $\text{cov}(\bar{y}, \hat{\beta}_1) = \text{cov}\left(\frac{1}{n}\sum y_i, \sum k_j y_j\right) = \sum \frac{1}{n} k_i \text{cov}(y_i, y_i)$

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

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

$= \sigma^2 h_{ii}$

This makes sense: A small change for extreme x_i observation 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 $E(e_i) = 0$ $V(e_i) = E(e_i^2)$ or $Cov(e_i, e_i)$

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

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

  $$= \sigma^2 - \sigma^2 h_{ii} - 2\underbrace{Cov(e_i, \hat{y}_i)}_{=0 \text{ by orthogonal projection}}$$

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

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.

Note residuals are not directly comparable since their variances differ

$\Rightarrow$ 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

$\Rightarrow$ 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've 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.

Diagnostics

$y = X\beta + \varepsilon \Rightarrow$ fit to data $\Rightarrow \hat{\beta} \Rightarrow \hat{y} = X\hat{\beta}$

$e = y - \hat{y}$

Facts

- $V(e_i) = \sigma^2(1 - h_{ii})$ - residuals have non-constant variance

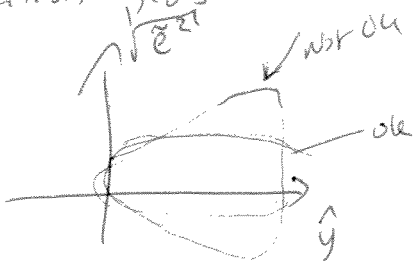
$\Rightarrow$ standardized residuals

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

- Plot (standardized) residuals against
  - fitted values
  - leverage
  - variables included & omitted

} checking basic assumptions (sufficiency of model, constant error var, no outliers, symm. of errors)

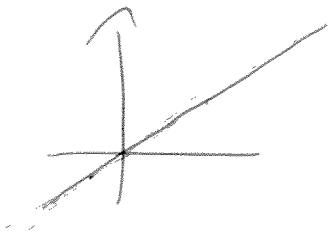
- Scale-location plots $\sqrt{\tilde{e}^2}$ vs $\hat{y}$



- Transformations
- drop observations
- ⋮

Extras

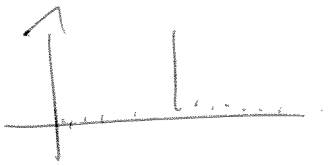
- QQ plot



(check if $e \sim$ Normally distributed
→ needed for some inference

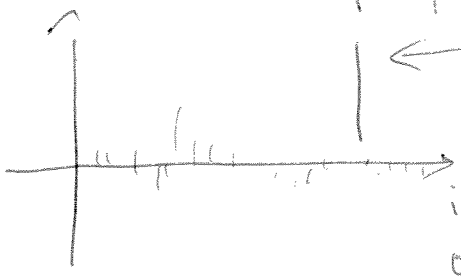
- Contribution by single observation

- leverage

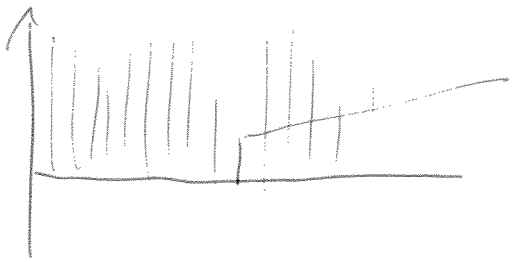


"potential to do harm"

- change in slope — refit w/o obs i and plot $\hat{\beta} - \hat{\beta}_{(i)}$ vs i



- change in MSE — refit w/o obs i and plot $\hat{\sigma}^2_{(i)}$ vs i



Combo of all — Cook's D

- studentized residual — residual for obs i when i is not included in the fit

$$e_i = y_i - \hat{y}_i$$

$$\hat{y}_i = \sum h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j$$

[illegible]

$$\hat{y}_{(i)} = \frac{1}{1-h_{ii}} \sum_{j \neq i} h_{ij} y_j \Rightarrow \hat{y}_{(i)} = \sum_{j \neq i} h_{ij} y_j + h_{ii} \hat{y}_{(i)}$$

$(y_i - \text{LH})^2 = (y_i - \text{RH})^2$

$$\Downarrow \quad y_i - \hat{y}_{(i)} = (y_i - \underbrace{\sum h_{ij} y_j}_{\hat{y}_i} + h_{ii}(y_i - \hat{y}_{(i)})$$

$$\Downarrow \quad \boxed{y_i - \hat{y}_{(i)} = \frac{y_i - \hat{y}_i}{1-h_{ii}}} = e_i^* \text{ or } e_{(i)}$$

$$\Rightarrow \text{Cook's D} = D_i = \frac{e_{(i)}^2}{p \hat{\sigma}^2}, \quad \hat{\sigma}^2 = \frac{\sum e_i^2}{n - df(\text{model})}$$