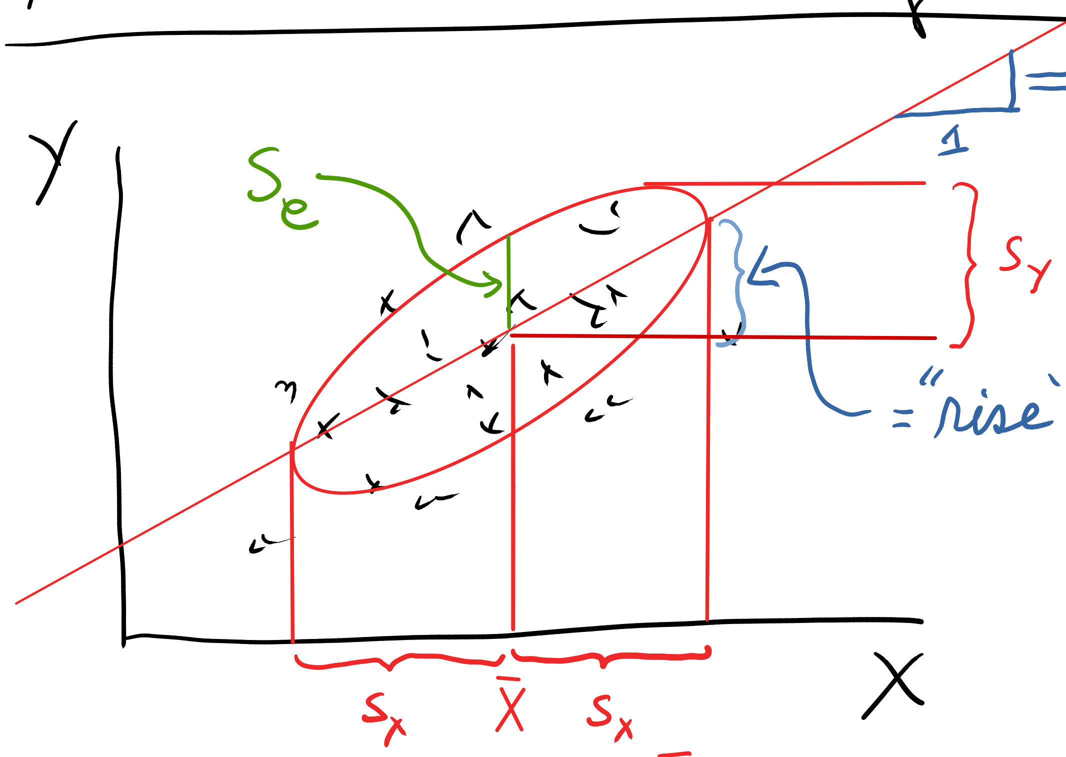
MATH 4939 Regression Recap



$= \frac{S_{xy}}{S_x^2} = \hat{\beta}_x =$ slope of least-squares line

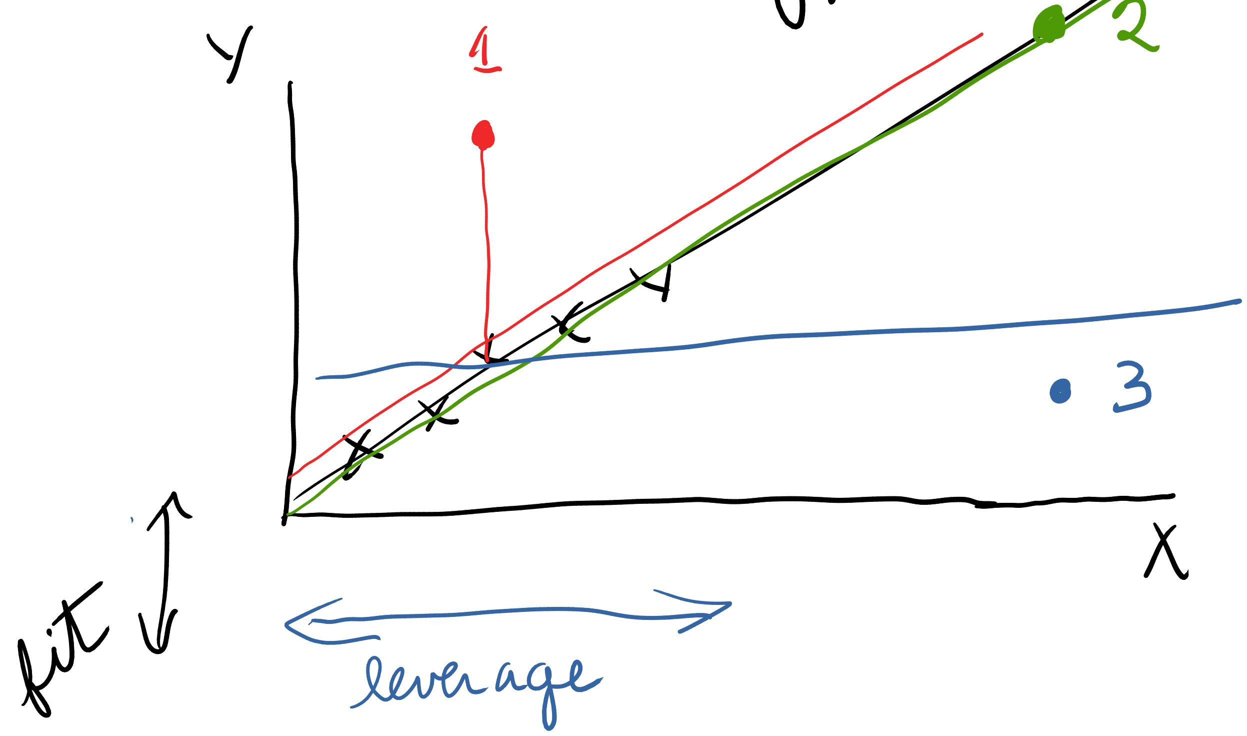
$=$ "rise" $=$ slope × run

$= \hat{\beta}_x \cdot S_x = \frac{S_{xy}}{S_x}$

$$SE(\hat{\beta}_x) = \frac{1}{\sqrt{n}} \frac{S_e}{S_x}$$

C.I. for β_x: $\hat{\beta}_x \pm \frac{D}{\sqrt{n}} \cdot \frac{S_e}{S_x}$

Outlier archetypes

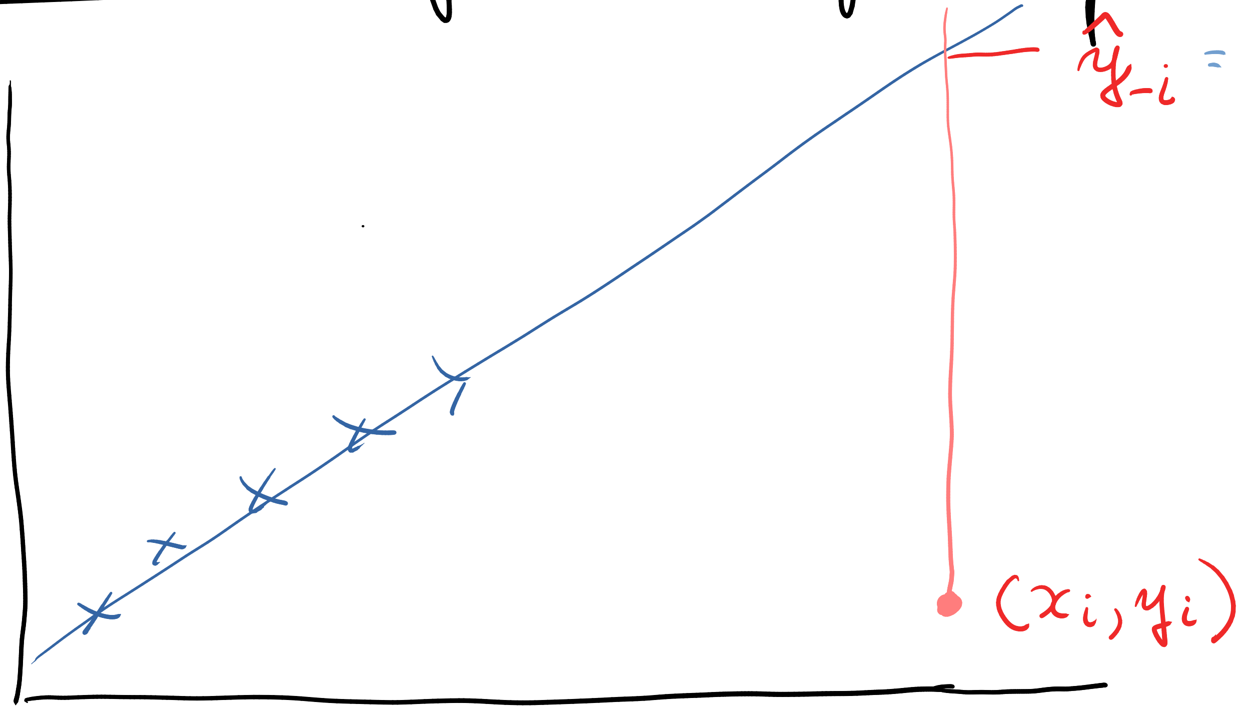


Type	Leverage	Influence on $\hat{\beta}_x$	s_e	s_x	CI width	p value
1	low	=	↑	=	↑	↑
2	high	=	=	↑	↓	↓
3	high	high ↑ or ↓	↑	↑	?	?

loosely { Leverage: How much effect a point can have
Influence: " " " " " does "

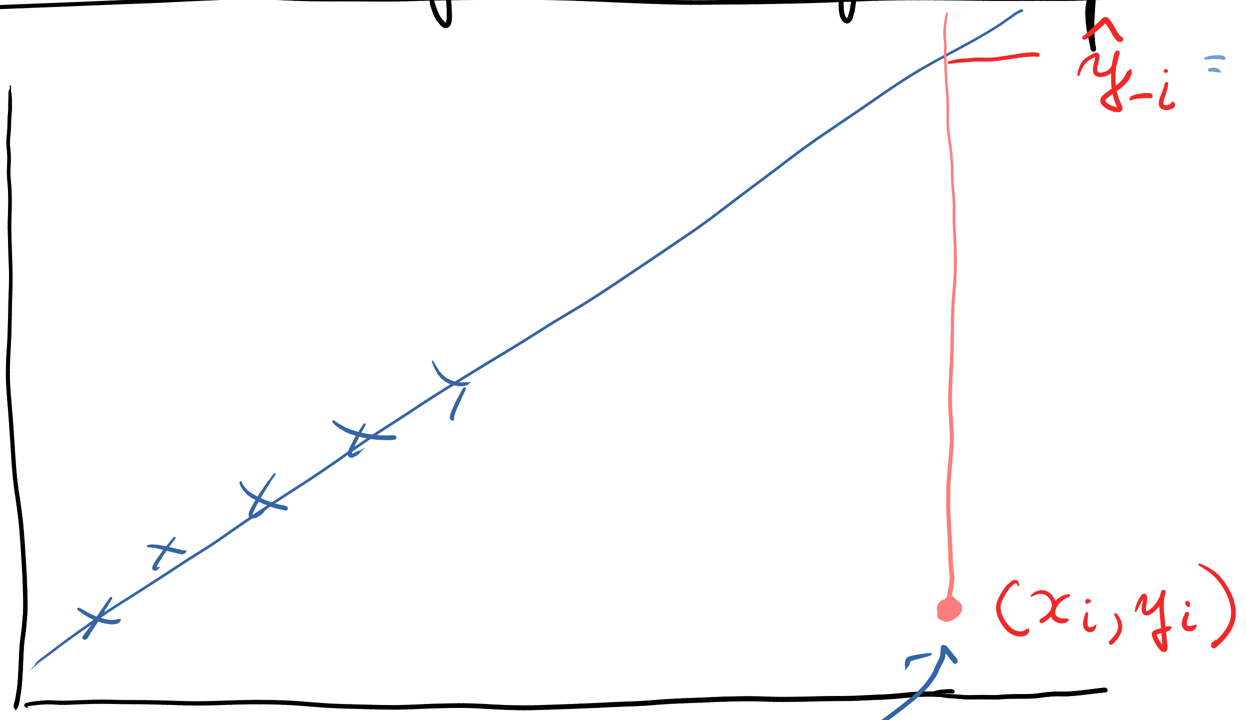
But influence on what? Can differ for different targets of inference.

Influence & leverage in simple regression



$\hat{y}_{-i}$ = fitted value at x_i if point i is omitted

Influence & leverage in simple regression



$\hat{y}_{-i}$ = fitted value at x_i if point i is omitted

How much does point i pull down the regression line?

$$\hat{\underset{\sim}{Y}} = X\hat{\underset{\sim}{\beta}} = X\left[(X'X)^{-1}X'\underset{\sim}{Y}\right]$$

$$= \left(X(X'X)^{-1}X'\right)\underset{\sim}{Y}$$

$$= H\underset{\sim}{Y}$$

<u>H is the matrix of an orthogonal projection in $\mathbb{R}^n$</u>

note: residuals

$$\underset{\sim}{e} = \underset{\sim}{Y} - \hat{\underset{\sim}{Y}} = (I - H)\underset{\sim}{Y}$$

Fact: $\underset{\sim}{e} \perp \text{span}(X)$, i.e. $X'\underset{\sim}{e} = \underset{\sim}{0}$

Proof: $X'\underset{\sim}{e} = X'(I - X(X'X)^{-1}X')\underset{\sim}{Y}$

$$= \big(X' - \underbrace{X'X(X'X)^{-1}}_{I} X' \big) \underset{\sim}{Y}$$

$$= (X' - X') \underset{\sim}{Y} = O \underset{\sim}{Y} = \underset{\sim}{0}$$

zero matrix

zero vector

note: $\hat{Y} \in \text{span}(X)$

since $\hat{Y} = X\hat{\beta}$, i.e $\hat{Y}$ is a linear combination of columns of X

So $\underset{\sim}{\hat{Y}} \perp \underset{\sim}{e}$, i.e. $\underset{\sim}{\hat{Y}}' \underline{e} = 0$

Back to $\hat{y}_i$ and y_i

$$\hat{\underset{\sim}{Y}} = H \underset{\sim}{Y}$$

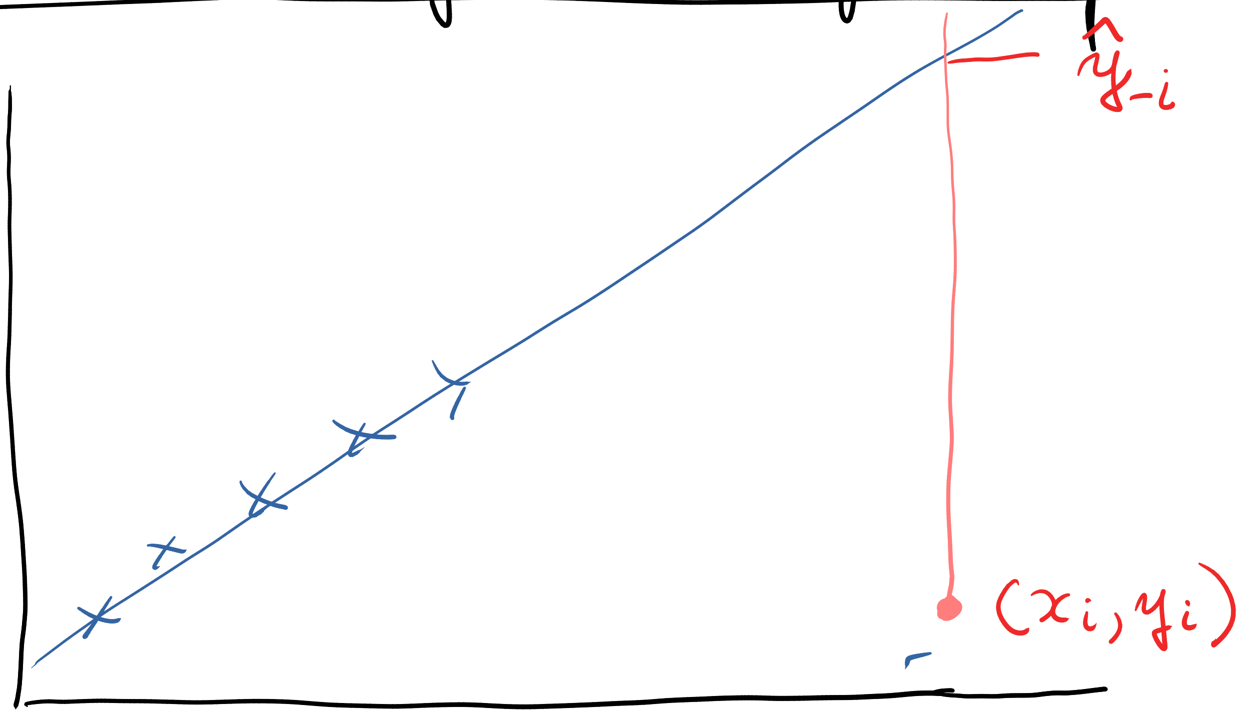
So $\hat{Y}_i$ = ith element of $\hat{\underset{\sim}{Y}}$

= ith row of $H \times \underset{\sim}{Y}$

$$= \sum_{j=1}^{n} h_{ij} Y_j$$

and $\boxed{\dfrac{\partial \hat{Y}_i}{\partial Y_i} = h_{ii}}$

Influence & leverage in simple regression

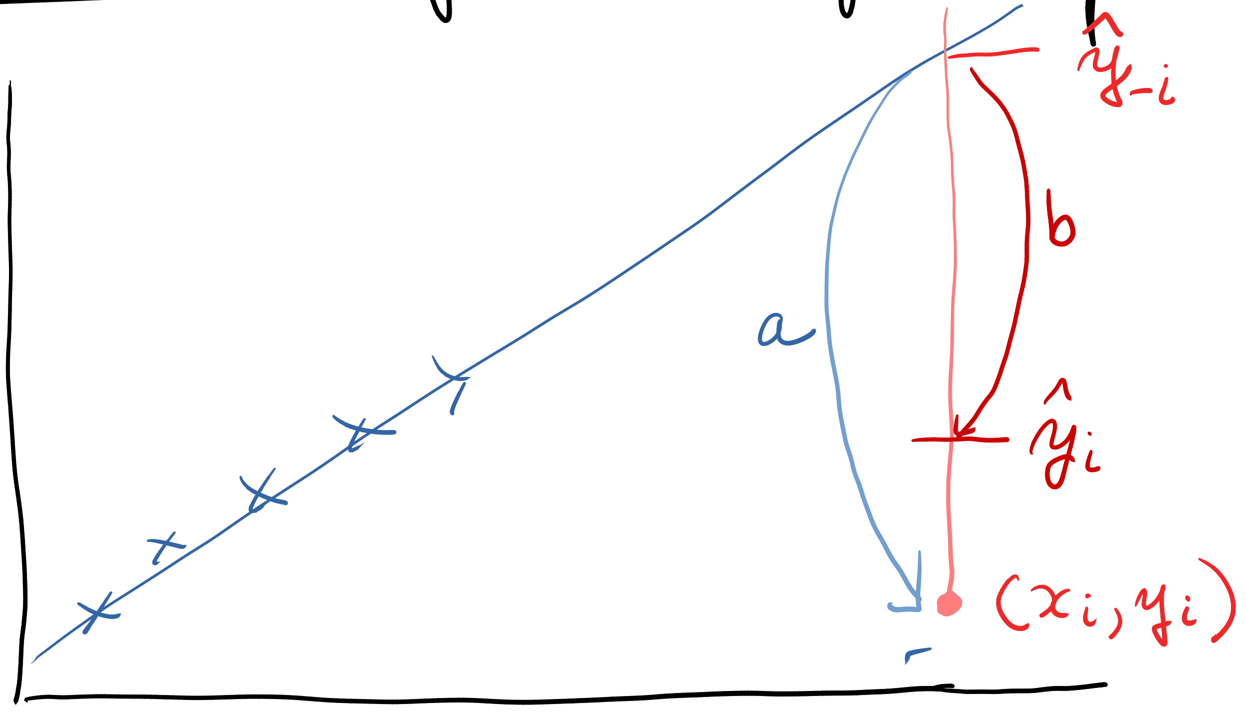


If $y_i = \hat{y}_i$ then LS line is unchanged

now think of pulling y_i down.
How will $\hat{y}_i$ follow y_i?

Answer: $\dfrac{\partial \hat{Y}_i}{\partial Y_i} = h_{ii}$

Influence & leverage in simple regression



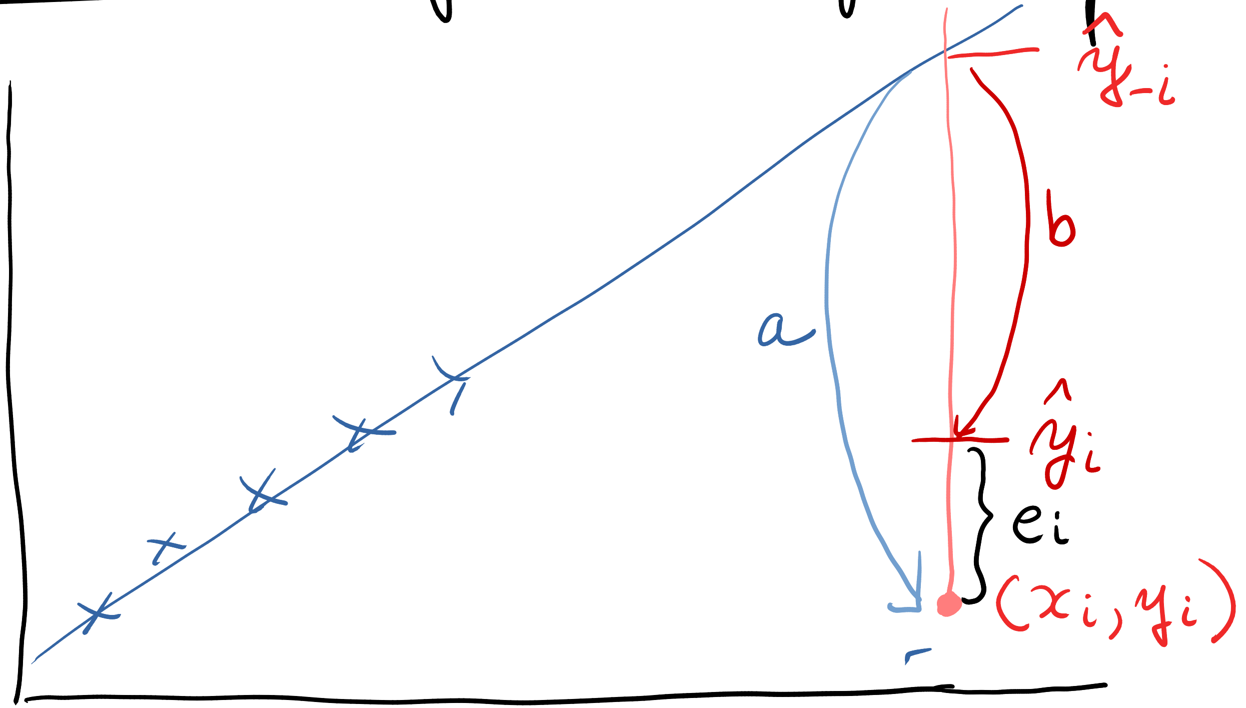
If $y_i = \hat{y}_i$ then LS line is unchanged

now think of pulling y_i down by a. How will $\hat{y}_i$ follow y_i?

Answer: $\dfrac{\partial \hat{Y}_i}{\partial Y_i} = h_{ii}$

$b = h_{ii} \times a$

Influence & leverage in simple regression



If $y_i = \hat{y}_i$ then LS line is unchanged

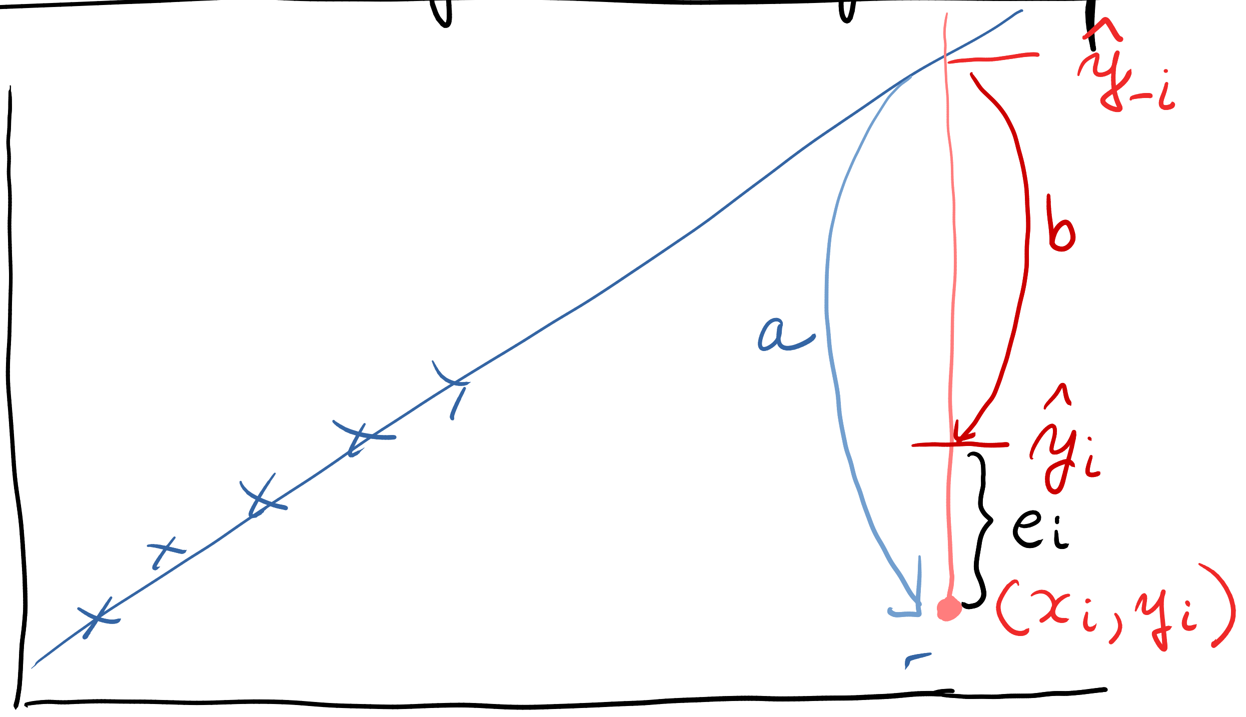
now think of pulling y_i down by a. How will $\hat{y}_i$ follow y_i?

Answer: $\frac{\partial \hat{Y}_i}{\partial Y_i} = h_{ii}$

$$b = h_{ii} \times a$$

"influence on $\hat{y}_i$" "leverage"

Influence & leverage in simple regression



If $y_i = \hat{y}_i$ then LS line is unchanged

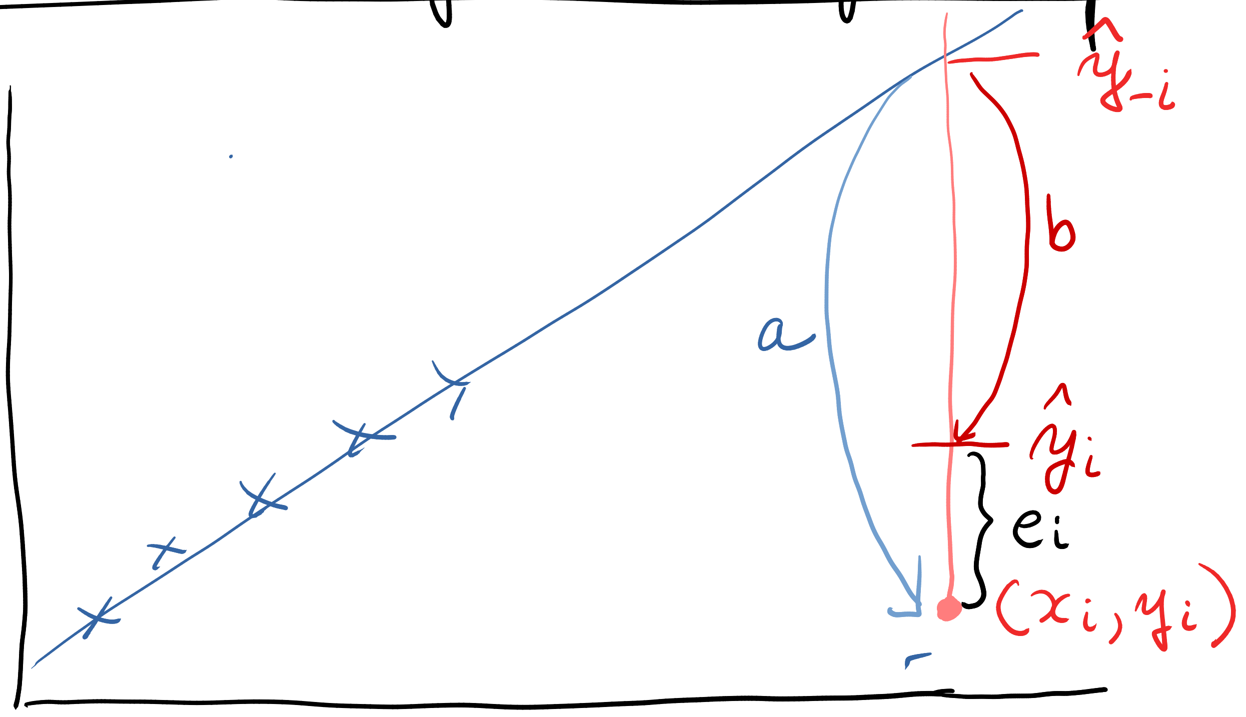
now think of pulling y_i down.
How will $\hat{y}_i$ follow y_i?

Answer: $\frac{\partial \hat{y}_i}{\partial y_i} = h_{ii}$

$$b = h_{ii} \times a$$

$$e_i = (1 - h_{ii}) \times a$$

Influence & leverage in simple regression



If $y_i = \hat{y}_i$ then LS line is unchanged

now think of pulling y_i down.
How will $\hat{y}_i$ follow y_i?

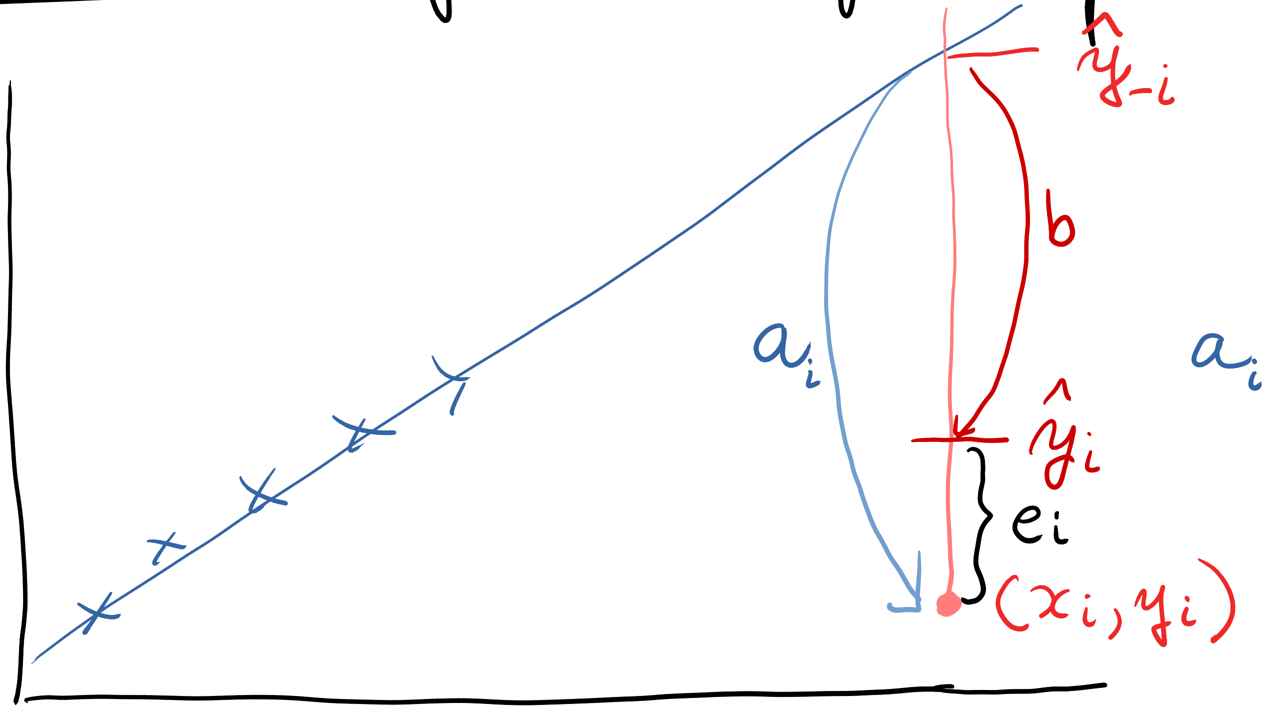
Answer: $\frac{\partial \hat{Y}_i}{\partial Y_i} = h_{ii}$

$$b = h_{ii} \times a$$

$$e_i = (1 - h_{ii}) \times a$$

a high leverage point can hide itself

Influence & leverage in simple regression



Some terminology

$a_i = y_i - \hat{y}_{-i}$

$= $ "deleted residual"

Note: $\operatorname{Var}(\underset{\sim}{e}) = \operatorname{Var}((I-H)\underset{\sim}{Y}) = (I-H)(\sigma^2 I)(I-H)' = \sigma^2 (I-H)$

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

Standardized residual $= \dfrac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$

Deleted residual looks like a better choice to show high influence points:

$$\underset{\sim}{a_i} = \left(\frac{1}{1-h_{ii}}\right) e_i$$

$$Var(a_i) = \frac{1}{(1-h_{ii})^2} Var(e_i) = \frac{1}{(1-h_{ii})^2} \sigma^2 (1-h_{ii})$$

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

$$\underline{\text{"Standardized" } a_i} = \frac{a_i}{\widehat{SD}(a_i)} = \left(\frac{1}{1-h_{ii}}\right) e_i \Big/ \left(\frac{\hat{\sigma}}{\sqrt{1-h_{ii}}}\right)$$

$$= \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}} = \underline{\text{"standardized" } e_i}$$

So standardizing doesn't yield a distinction

What does make a difference is

- internal standardization

  Estimate $\hat{\sigma}$ including ith point

versus

- external standardization:

  Estimate $\hat{\sigma}$ excluding ith point

Personal preference:

Reserve "studentized residual" for externally studentized residuals!

Widely accepted simplistic advice

A newer version of this product documentation is available. You are viewing an older version. View latest

Studentized residual test

Last Updated: 2024-02-29

Studentized residual is computed as regression model residual divided by its adjusted standard error.

Residuals are obtained by subtracting the target value that is predicted by the regression model, from observed target value for each data row. Standard error is given by the square root of the mean square for the error source. The adjustment of the standard error consists in multiplying it by the square root of leverage value that is subtracted from one. Leverage value is computed based on the design matrix and design matrix row for the data row. It adjusts the standard deviation by taking into account predictor values.

An outlier test for studentized residuals is conducted by comparing the absolute value of studentized residual with threshold value 3. Studentized residuals are distributed according to t distribution and the probability of being greater than the threshold is less than 1%.

Points with highest ranking studentized residuals above the threshold value are reported as meaningful differences, that is outliers in this case.

This tends to capture only Type 3 and some Type 1

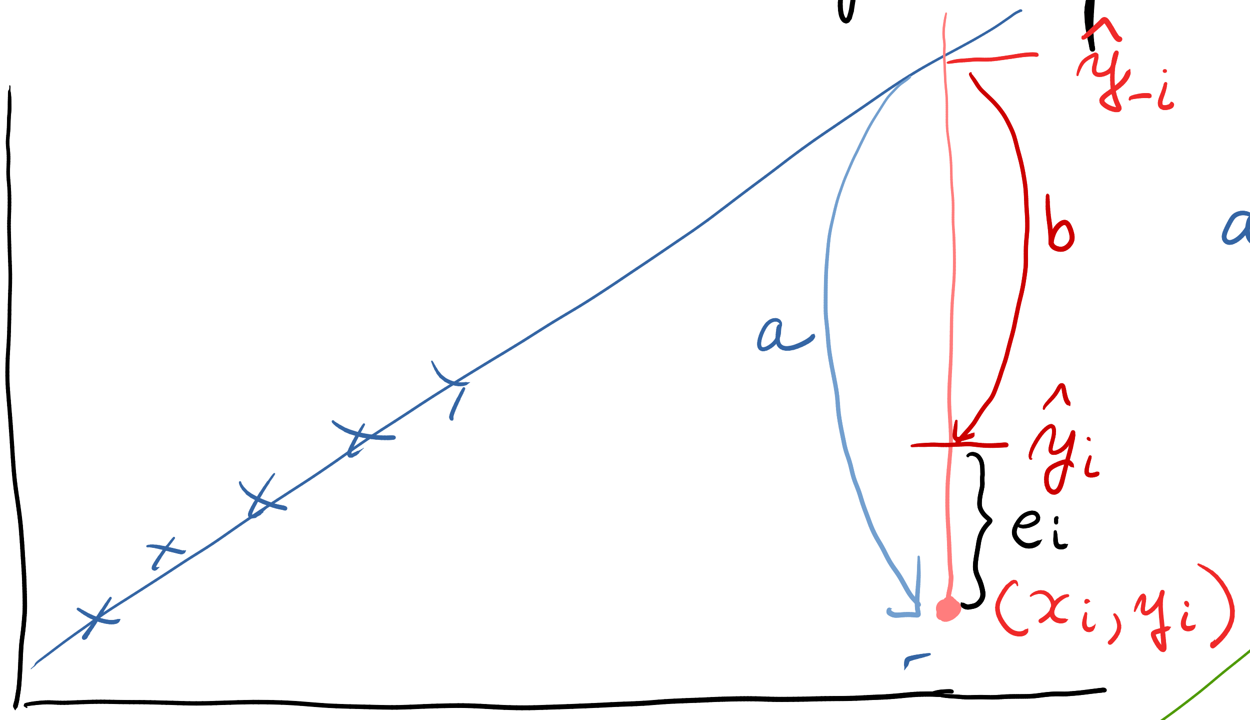
"Outliers" (defined broadly) should be examined for their implications.

- Find out how they occurred
- Do they suggest the need for a different model

Often they suggest new insights and theories

If you treat them mechanically you won't benefit from their often powerful lessons.

Influence & a in simple regression



Some Terminology

$a = y_i - \hat{y}_{-i}$

= "deleted residual"

$\dfrac{a}{\hat{\sigma}_e} = \dfrac{y_i - \hat{y}_{-i}}{\hat{\sigma}_e}$

= "studentized residual"

"internal" or "external" depending on $\hat{\sigma}_e$

$\dfrac{e_i}{\hat{\sigma}_e} = \dfrac{y_i - \hat{y}_i}{\hat{\sigma}_e}$

= "standardized residual".

Relationship between a & e_i

Facts about h_{ii}

Since $H = HH'$

$$h_{ii} = \sum_{j=1}^{n} h_{ij} h_{ij} = \sum_{j=1}^{n} h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$$

So $h_{ii} \leq h_{ii}^2$ and h_{ii} is a sum of squares

$\therefore \quad 0 \leq h_{ii} \leq 1$

We can also show, if X has a constant column

then $\frac{1}{n} \leq h_{ii} \leq 1$

In simple regression

$$h_{ii} = \frac{1}{n}\left(1 + z_i^2\right)$$

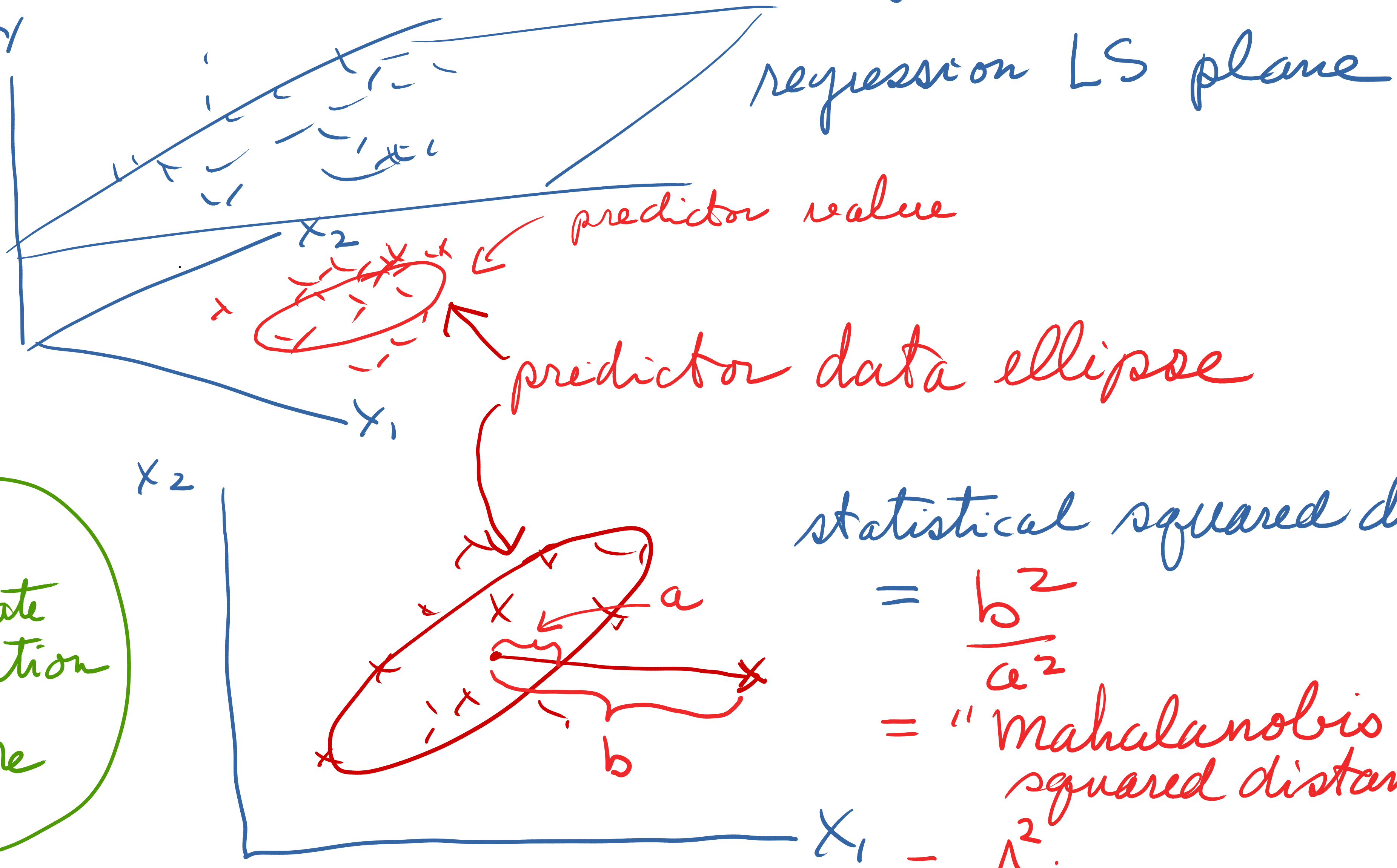
where z_i is X_i expressed as a Z-score

$$z_i = \frac{X_i - \bar{X}}{\hat{\sigma}_x} \qquad \hat{\sigma}_x = \sqrt{\frac{\Sigma (X_i - \bar{X})^2}{n}}$$

Takeaway: a point's leverage on $\hat{y}_i$ depends on how relatively unusual it is in "predictor" space.

Think of z_i^2 as the "statistical squared distance of point i from other points."

This works with multiple regression



Multivariate generalization of Z score

statistical squared distance

$$= \frac{b^2}{a^2}$$

= "Mahalanobis squared distance"

$$= \Delta_i^2$$

$$h_{ii} = \frac{1}{n}\left(1 + \Delta_i^2\right)$$

- So points on the standard data ellipse have leverage $= \frac{2}{n}$
- Points on the data ellipse of radius 2 have leverage $= \frac{5}{n}$

What about influence on $\hat{\beta}_i$?

"Added variable plot" ← unambiguous

AKA "Partial regression leverage plot"

AKA ~~Partial residual~~ plot

~~but bad usage because ambiguous.~~

How? Use. car::avp

• base::plot.

Idea: Regression of Y on $X_1, X_2, \ldots, X_k$

What is influence of ith point of $\hat{\beta}_p$?

- "Account" for all X's except X_p and look at relationship between

residuals of Y regressed on all X's except X_p

versus

residuals of X_p " " " " " "

Plot the result: This is the AVP

Up to first order the AVP gives you the same information about the relationship of

Y and X_p in a multiple regression

as a simple scatterplot gives you about

Y and X_p in a simple regression.

AVP : Mult. regression :: Scatterplot : simple regression

Beware: Simple scatterplot of Y & X_p is not informative for multiple regression

Why? Remember Simpson's Paradox !!

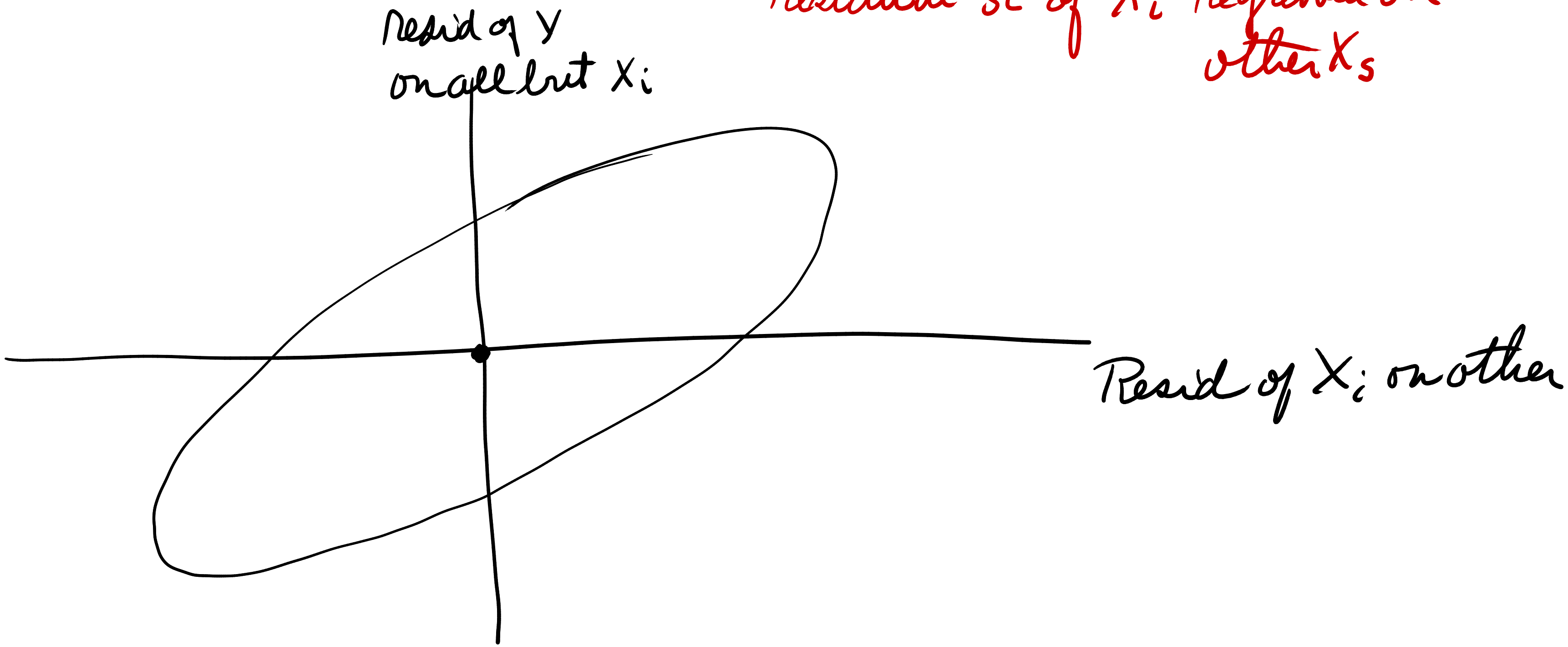
Inference about $\hat{\beta}_i$ in multiple regression

residual SE of Y on all Xs

$$SD(\hat{\beta}_i) = \frac{1}{\sqrt{n}} \frac{S_e}{\underbrace{S_{X_i|\text{others}}}}$$

residual SE of X_i regressed on other Xs

AVP



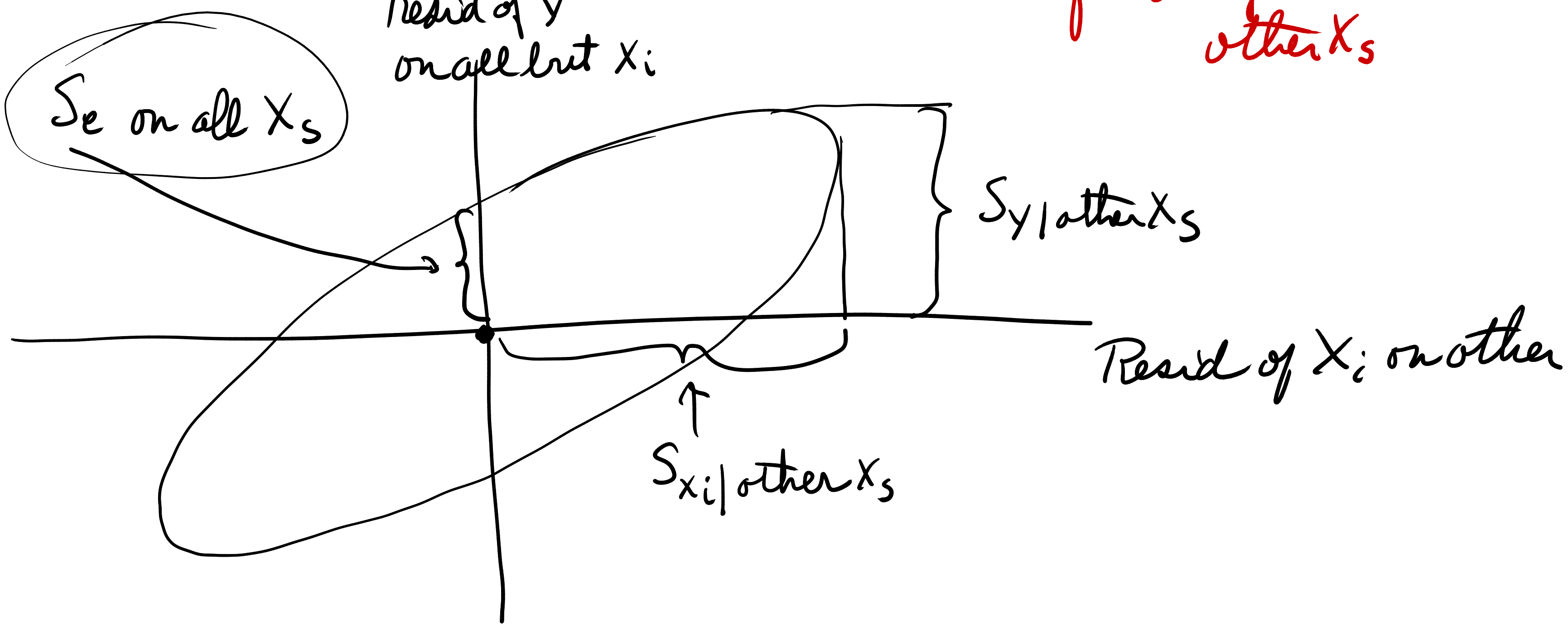
Inference about $\hat{\beta}_i$ in multiple regression

residual SE of Y on all Xs

$$SD(\hat{\beta}_i) = \frac{1}{\sqrt{n}} \frac{S_e}{\underbrace{S_{x_i|\text{others}}}}$$

residual SE of X_i regressed on other Xs

$\boxed{\text{AVP}}$



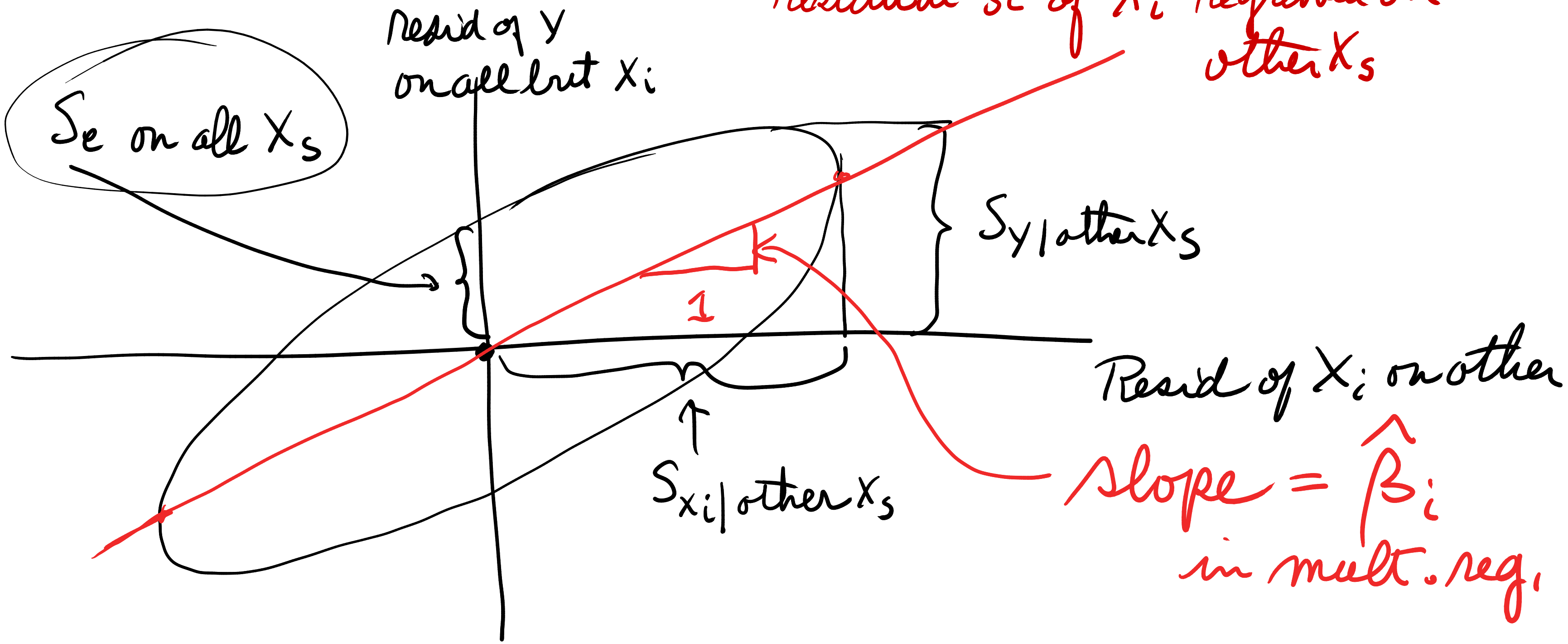
Inference about $\hat{\beta}_i$ in multiple regression

residual SE of Y on all Xs

$$SD(\hat{\beta}_i) = \frac{1}{\sqrt{n}} \frac{S_e}{S_{x_i|\text{others}}}$$

residual SE of X_i regressed on other Xs

$\boxed{\text{AVP}}$



Inference about $\hat{\beta}_i$ in multiple regression

residual SE of Y on all Xs

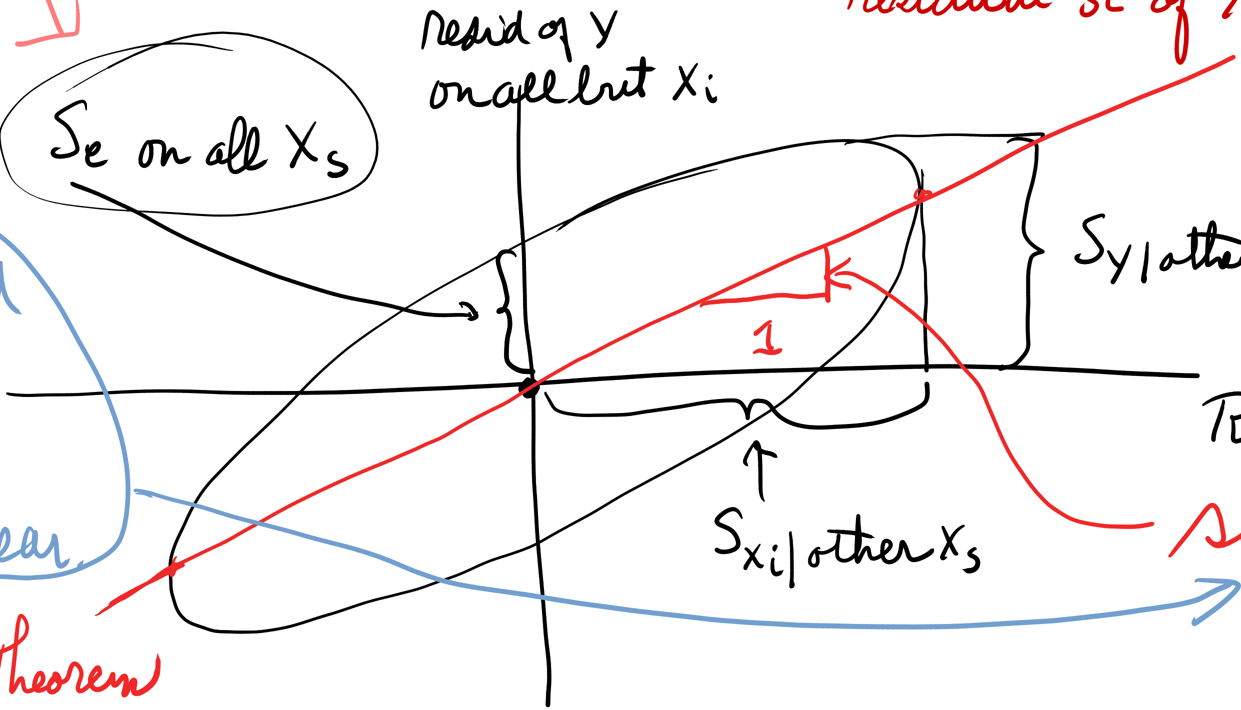
$$SD(\hat{\beta}_i) = \frac{1}{\sqrt{n}} \frac{S_e}{\underbrace{S_{X_i|\text{others}}}}$$

residual SE of X_i regressed on other Xs

proved this

$\boxed{\text{AVP}}$

S_e on all Xs



slope $= \hat{\beta}_i$ in mult. reg.

The guy who proved this got a Nobel Prize the following year.

Frisch-Waugh-Lovell Theorem

Inference about β_i in multiple regression

could be from z, t, F etc.

residual SE of Y on all Xs

$$\text{C.I. for } \beta_i : \quad \hat{\beta}_i \pm \frac{D}{\sqrt{n}} \, \frac{S_e}{\underbrace{S_{x_i | \text{others}}}}$$

residual SE of X_i regressed on other Xs

$\boxed{\text{AVP}}$