Multivariate

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i$

→ if 5 basic assumptions hold ⇒ LS fit OK

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & & x_{2,p-1} \\ & & & & \\ 1 & x_{n1} & x_{n2} & & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\underset{\sim}{y} = \bar{X} \underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

$n \times 1$ $\quad n \times p$ $\quad p \times 1$ $\quad n \times 1$

Minimize LS $Q = \sum (y_i - \underset{\sim}{x_i}' \underset{\sim}{\beta})^2$ where $\underset{\sim}{x_i}' = (1, x_{i1}, \ldots x_{i,p-1})$

$= \sum e_i^2$

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

Scaler 1×1

$\Rightarrow \frac{\partial Q}{\partial \beta} = 0 \Rightarrow$ **Normal Equations** $(\bar{X}'\bar{X})\underset{\sim}{\beta} = \bar{X}'y$

(A) In simple linear regression → solution $\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\mathrm{cov}(x,y)}{\mathrm{cov}(x,x)} \sim \mathrm{corr}(x,y)$

(B) Multiple case → $(\bar{X}'\bar{X}) \sim \mathrm{cov}(\bar{X})$ where $V(x_1), V(x_2) \ldots V(x_{p-1})$ on the diagonal, and cross-covariance $\mathrm{cov}(x_1, x_2), \mathrm{cov}(x_k, x_\ell)$ off-diagonal

$$\rightarrow (X'y) = \begin{pmatrix} \mathrm{cov}(x_1, y) \\ \mathrm{cov}(x_2, y) \\ \\ \mathrm{cov}(x_{p-1}, y) \end{pmatrix}$$

- If $(\bar{X}'\bar{X}) = \Lambda$ (diagonal), all x's are uncorrelated.

→ then $\hat{\beta} = (\bar{X}'\bar{X})^{-1}\bar{X}'y \sim \begin{pmatrix} \text{Corr}(x_1, y) \\ \\ \text{Corr}(x_{p+1}, y) \end{pmatrix}$

ie. each coefficient has a direct interpretation as the dependency between an individual x-variable and y.

- If $(\bar{X}'\bar{X})$ not diagonal → no direct interpretation

→ can't separate influence of one x on y from another (correlated) x.

→ That is, $\hat{\beta}_k$ involves both $\text{Corr}(x_k, y)$ and

$\text{Corr}(x_\ell, y)$, $\ell \neq k$.

$\text{Corr}(x_\ell, x_k)$

- Moreover → if x's are highly correlated, $(\bar{X}'\bar{X})^{-1}$ may ~~not~~ exist! $(\det(\bar{X}'\bar{X}) \to 0)$

or inverse $(\bar{X}'\bar{X})^{-1}$ numerically unstable

→ When this happens, magnitude & even sign of $\hat{\beta}$'s may become meaningless. [Many models fit the data equally well]

Ex: i) $X_2 = a + bX_1$

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

} infinite set of β_0 & β_1 fit these data equally well

The Hat-matrix

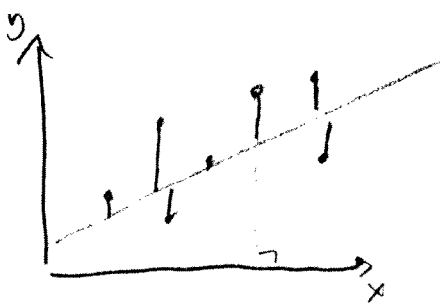
Assume we can solve the normal equations & obtain

$$\hat{\beta} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{y}$$

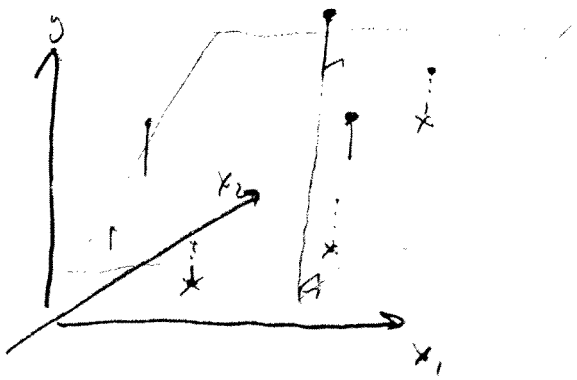
- Filled values $\underset{\sim}{\hat{y}} = \underset{\sim}{X}\hat{\beta} = \underbrace{\underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'}\ \underset{\sim}{y}$

Hat-matrix $H = \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'$

- H is a projection matrix — projects the data $\underset{\sim}{y}$ onto the plane spanned by $\underset{\sim}{X}$



Simple case



projection ⊥ to both X_1 & X_2 in multiple case.

Facts

- $HH = H$ { (H is an idempotent matrix) (NO point in redoing regression)
  ($H' = H$) (Symmetric)
- $e = \hat{\varepsilon} = y - \hat{y} = y - Hy = (I - H)y$
- $X'e = \bar{X}'(I - \bar{X}(\bar{X}'\bar{X})^{-1}\bar{X}')y = 0$
  (orthogonal projection)
- $\hat{y}'e = y'H(I-H)y = 0$
  (fitted values $\perp$ to residuals)

$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & & h_{2n} \\ & & & \\ h_{n1} & h_{n2} & & h_{nn} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

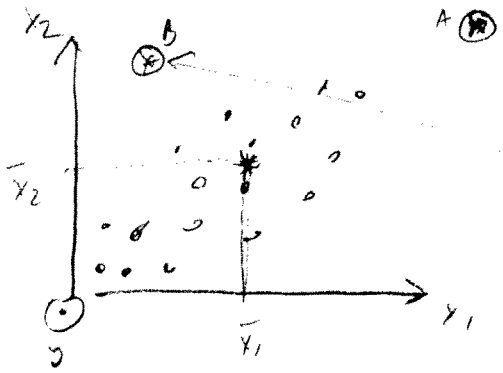
leverage $h_{ii} = \underset{\sim}{x}_i'(\bar{X}'\bar{X})^{-1}\underset{\sim}{x}_i$ (ith row of $\bar{X}$)

Note, h_{ii} is large when $\underset{\sim}{x}_i$ is extreme compared with $\bar{\underset{\sim}{x}}$

$\underset{\sim}{x}_i$: vector of all x-variable values for observation i

$\bar{\underset{\sim}{x}}$: vector, average x-values.

Extreme how? wrt mass of data



Extreme in both x_1 & x_2

Extreme cmp. w. mass of data (extreme in x_2 only.)

Both ⊙ observations have high leverage.

Properties

- As before, $E(\hat{\beta}) = \beta$ (unbiased)

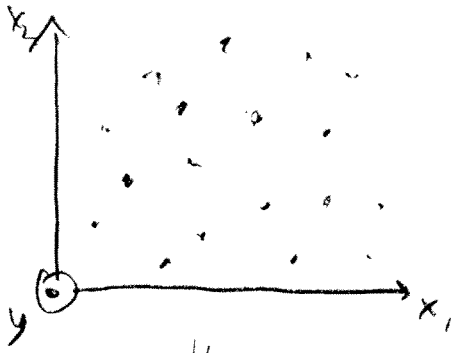
- $V(\hat{\beta}) = V\left(\underbrace{(X'X)^{-1}X'}_{\text{Constant}}y\right) = (X'X)^{-1}X' \cdot \underbrace{V(y)}_{=\sigma^2 I} \cdot X(X'X)^{-1} = \sigma^2(X'X)^{-1}$

Meaning? → More spread in any x ($var(x_u)$ large)
→ better estimate of $\hat{\beta}$

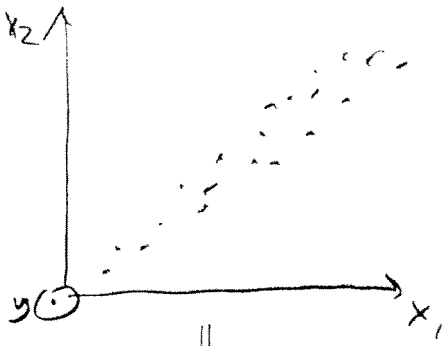
→ $\sigma^2 \uparrow$ → variance ↑

→ Unless $(X'X)$ is diagonal, $\hat{\beta}$'s are correlated ⇒ Think about impact on inference!

→ The more dependent x's are, the larger the $V(\hat{\beta})$.



$\hat{\beta}_1$ & $\hat{\beta}_2$ nearly uncorrelated



$\hat{\beta}_1$ & $\hat{\beta}_2$ heavily dependent

→ large variance associated with estimates

→ Difficult to make direct inferences.

Fitted values $\hat{y} = Hy \Rightarrow E(\hat{y}) = E(y)$

$V(\hat{y}) = H V(y) H = \sigma^2 H$

i.e. large variance @ locations of high leverage)

Residuals $e = (I-H)y \Rightarrow E(e) = 0$

$V(e) = \sigma^2 (I-H)$

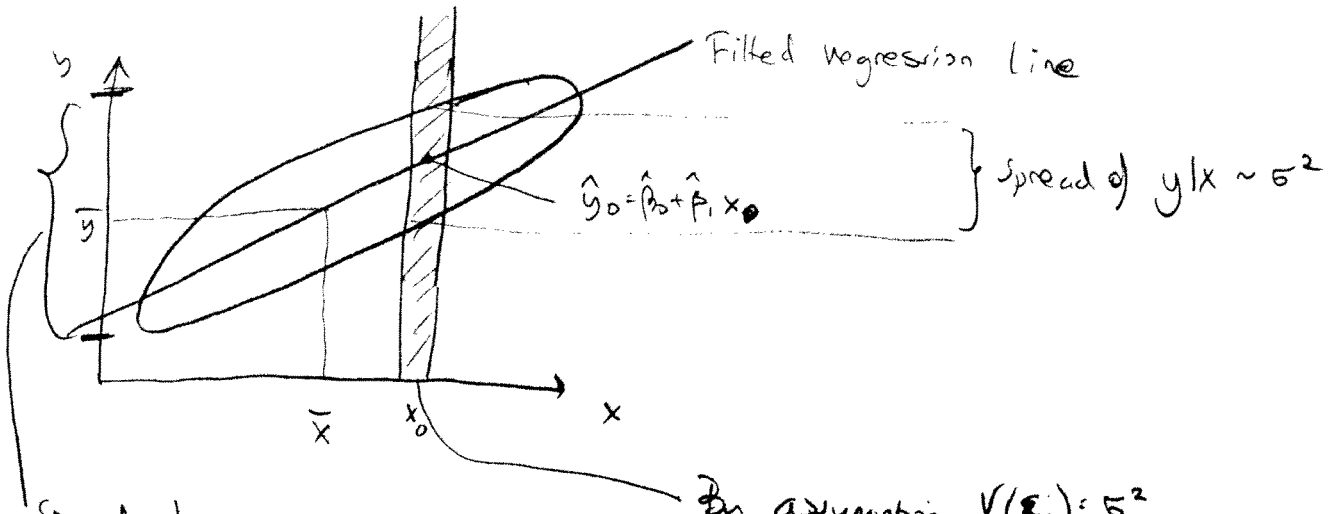
→ smaller variance @ locations w. high leverage

$\downarrow$

e's are correlated!

Variance Decomposition

Assume $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$



Spread of y-values in sample $\sim V(y)$

By assumption $V(\varepsilon_i) = \sigma^2$, the spread of y-values in each vertical strip is $\sim V(y|X=x_0) = \sigma^2$

- Now, if x and y are unrelated, then $V(y) \approx V(y|X=x)$.
- If $y = \beta_0 + \beta_1 x$ exactly, $V(y|X=x) = 0$!

So how much did linear regression reduce the uncertainty in y?

① $RSS = SSE = \sum (y_i - \hat{y}_i)^2$ (spread of y's around the regression line)

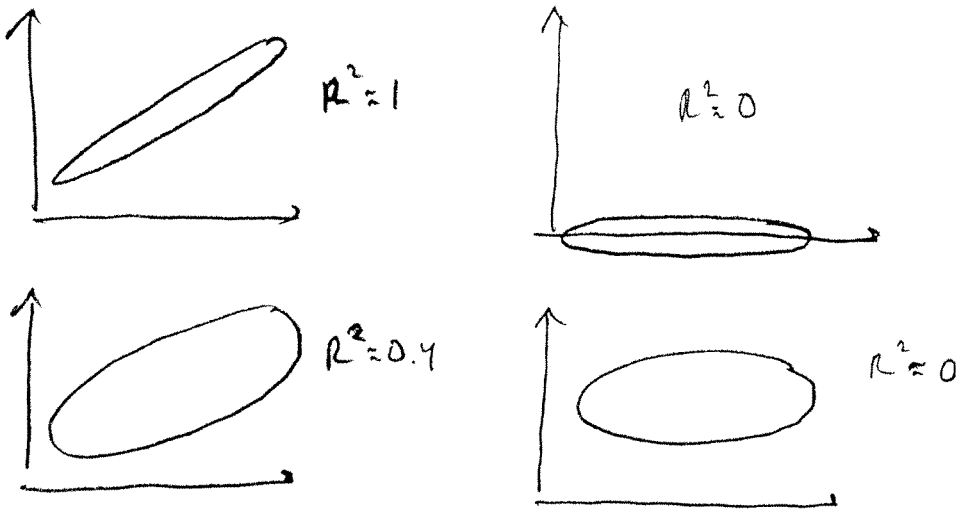
② $SS_T = \sum (y_i - \bar{y})^2$ (spread around mean of y)

- Since $\hat{y}$ is obtained from the data (the best possible line is fit to the n sample points), RSS is <u>always</u> less than SS_T

→ Numerical summary; $R^2 = \frac{SS_T - RSS}{SS_T} = \frac{\text{reduction in spread}}{\text{original spread}}$

$$= 1 - \frac{RSS}{SS_T}$$

R^2 : the % of variance (spread) in y explained by the regression line.



Variance decomposition =)

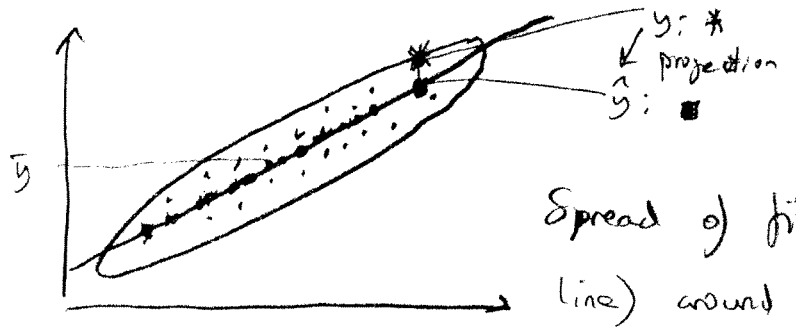
$$SS_T - SS_E = \sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2$$

$$= \cancel{\sum y_i^2} + \sum \bar{y}^2 - 2\bar{y} n \bar{y} - \cancel{\sum y_i^2} - \sum \hat{y}_i^2 + 2\sum y_i \hat{y}_i \quad (y_i = \hat{y}_i + e_i)$$

$$= -n\bar{y}^2 - \sum \hat{y}_i^2 + 2\sum \hat{y}_i^2 + 2\sum e_i \hat{y}_i$$

($\sum e_i \hat{y}_i = 0$ by $\perp$ projection)

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



Spread of fitted values (on the line) around $\bar{y} = \sum (\hat{y}_i - \bar{y})^2 = SS_{reg}$

So, we have $\boxed{SS_T = SS_E + SS_{reg}}$

$$R^2 = 1 - \frac{SS_E}{SS_T} = \frac{SS_{reg}}{SS_T}$$

How large should R^2 be for regression to "make sense"?

- Prediction: ~30%
- Interpretation: ~80%

"Roughly"

More formally

⇒ The F-test Goodness-of-fit test (or Lack-of-fit)

Assume y and x are not (linearly related), i.e. $\beta_1 = 0$!

Then, what would happen to SS_{reg} & SS_E? (R^2)

- Well, if $\beta_1 = 0$ then true model for y is $y_i = \beta_0 + \varepsilon_i$ ($V(\varepsilon_i) = \sigma^2$)
  and our best estimate for β_0 is $\hat{\beta}_0 = \bar{y}$

  The expected value of SS_T under the null ($\beta_1 = 0$, ↑) is

  $$E_0\left(\sum (y_i - \bar{y})^2\right) = (n-1)\sigma^2$$ (just our basic estimate of the variance, remember)

- What if $\beta_1 \neq 0$ (or maybe it is, but we fit the line to the data anyway), then

  $$E\left(\sum (y_i - \hat{y}_i)^2\right) = E\left(\sum e_i^2\right) = \sum_i E(y_i^2) + E(\hat{y}_i^2) - 2E(y_i \hat{y}_i) =$$

  ($y_i = \hat{y}_i + e_i$)

  $$= \begin{bmatrix} V(\hat{y}_i) = \sigma^2 h_{ii} \\ \sum e_i \hat{y}_i = 0 \\ \sum h_{ii} = \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right) = 2 \end{bmatrix} = \sum_i V(y_i) + E(y_i)^2 + V(\hat{y}_i) + E(\hat{y}_i)^2$$

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

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

[Note, if we have p parameters in the regression model $E(SS_E) = \sigma^2(n-p)$]

- What about SSreg?

Under the null / $\beta_1 = 0$, $E(y_i) = \beta_0$

$$\Rightarrow E_0\left(\sum(\hat{y}_i - \bar{y})^2\right) = E_0\left(\sum \hat{y}_i^2 + \bar{y}^2 - 2\bar{y}\hat{y}_i\right) = E_0\left(\sum \hat{y}_i^2\right) + n E_0(\bar{y}^2) - 2E_0(\bar{y}\sum$$

(1) (2) (3)

$$= \underbrace{\sigma^2 \sum h_{ii} + n\beta_0^2}_{(1)} + \underbrace{\sigma^2 + n\beta_0^2}_{(2)} - \underbrace{2\sigma^2 - 2n\beta_0^2}_{(3)} =$$

$\left(\sum V(\hat{y}_i) + \sum E(\hat{y}_i)^2\right)$ $\left(n V(\bar{y}) + E(\bar{y})^2\right)$ $\left(E(\bar{y}\sum\hat{y}_i) = \sum(\mathrm{cov}(y_i, \hat{y}_i) + E(y_i)E(\hat{y}_i))\right)$

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

Note, in general w. p parameters in model, $\sum h_{ii} = p$ and

$$E_0(SSreg) = \sigma^2(p-1)$$

MEANING:

(y) $\beta_1 = 0$ (X and y are not related)

all SS's provide estimates of the same parameter, σ^2 !

Under the null ($\beta_1 = 0$)

$$\frac{SST}{n-1} = \hat{\sigma}^2 \quad , \quad \frac{SSE}{n-p} = \hat{\sigma}^2 \quad \text{and} \quad \frac{SSreg}{p-1} = \hat{\sigma}^2$$

Extra assumption : $\varepsilon_i \sim N(0, \sigma^2)$

Now, i) $\varepsilon_i \sim$ normally distributed and $\beta_1 = 0$

$$\frac{SS_{reg}}{\sigma^2} \sim \chi^2_{p-1} \quad \perp \quad \frac{SS_E}{\sigma^2} \sim \chi^2_{n-p}$$

SS_{reg}: sum of sq's of normals.

$\perp$: independent ($\perp$ projection)

Fact: $\dfrac{\chi^2_{m}/m}{\chi^2_{n}/n} \equiv F_{m,n}$ (indep)

Our test statistic:

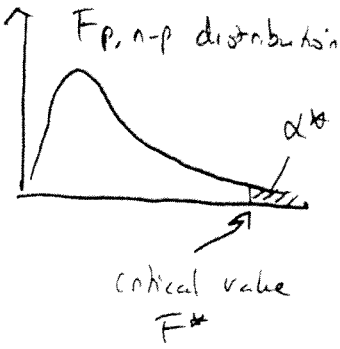
$$F = \frac{SS_{reg}/(p-1)}{SS_E/(n-p)} = \frac{(SS_T - SS_E)/((n-1)-(n-p))}{SS_E/(n-p)}$$

$$= \frac{\left(\dfrac{\text{reduction in SS}}{\text{reduction in \# parameters}}\right)}{(\text{MSE of full model})}$$

{ (1) If $\beta_1 = 0$, $F \sim F_{p-1, n-p}$

(2) $\beta_1 \neq 0$, F is inflated

Hypothesis testing: { Null: $\beta_1 = 0$; Alt: $\beta_1 \neq 0$ (Compare $F_{observed}$ to



We reject $[\beta_1 = 0]$ @ the α^*-level if
$F_{observed}$ exceeds F^*.

Discussion

[Significant is not the same thing as important!
If n is large enough, we reject almost any hypothesis]

Some more distribution theory , assuming $\varepsilon_i \sim N(0, \sigma^2)$

- $\hat{\beta}_1 = \sum k_i y_i$ → (a) if n is large, by CLT $\hat{\beta}_1 \sim$ normally distributed
  → (b) if $\varepsilon_i \sim N(0,\sigma^2) \Rightarrow y_i \sim N \Rightarrow \hat{\beta}_1 \sim$ — " — " —

(case (b)) if $\varepsilon_i \sim N(0,\sigma^2) \Rightarrow y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

$$\Rightarrow \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum (x_j - \bar{x})^2}\right) \quad \text{or} \quad \boxed{\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 / \sum (x_j - \bar{x})^2}} \sim N(0,1)}$$

but we don't know $\sigma^2 \Rightarrow$ use plug-in estimator $\hat{\sigma}^2$

From before we know $\frac{SSE}{n-2} = \hat{\sigma}^2$ and if ε_i's are normal $\frac{SSE}{\sigma^2} \sim \chi^2_{n-2}$ or equivalently $(n-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$

Fact [review basic stats]

$$\frac{N(0,1)}{\sqrt{\chi^2_\nu/\nu}} \equiv t_\nu \quad \text{index}$$

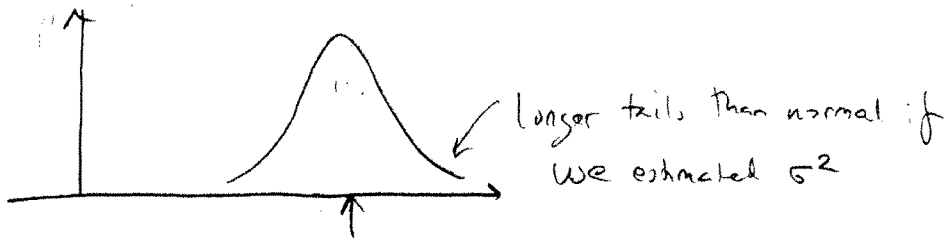
$\Rightarrow$ Here; $\dfrac{\hat\beta_1 - \beta_1}{\sqrt{\sigma^2/\sum(x_i-\bar x)^2}} \Big/ \sqrt{\dfrac{\hat\sigma^2}{\sigma^2}}$

$$= \frac{\hat\beta_1 - \beta_1}{\sqrt{\hat\sigma^2/\sum(x_i-\bar x)^2}} = \frac{\hat\beta_1 - \beta_1}{SE(\hat\beta_1)} = t\text{-statistic}$$

The standard error

$\sqrt{\hat V(\hat\beta_1)}$ ← σ^2 estimated.

So, sampling distribution of estimated $\hat\beta_1$



$\Rightarrow$ Back to Basics now!

- Interval estimate $\left[\hat\beta_1 \pm t_{n-p}(1-\alpha/2)\sqrt{\dfrac{\hat\sigma^2}{\sum(x_i-\bar x)^2}}\right] = I_{\beta_1}$
  - I_{β_1} covers true β_1 w. probability $1-\alpha$.
  - Question: Does I_{β_1} covers 0 (the null)?