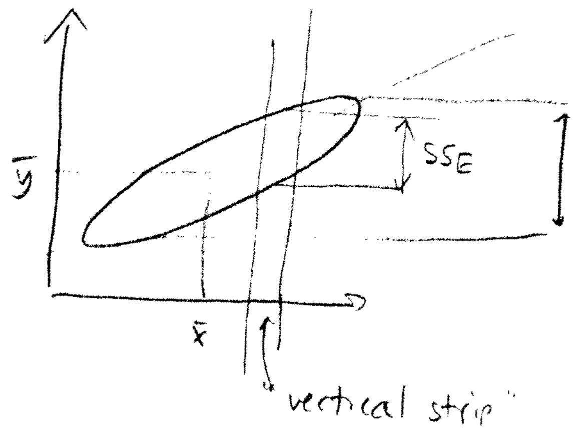
RECAP

Variance Decomp.



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

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

$$SS_T = SS_E + SS_{reg}, \quad \text{where} \quad SS_{reg} = \sum (\hat{y}_i - \bar{y})^2$$

(spread along the regression line)

If $\beta_1 = 0$ (no x~y relationship) $\Rightarrow$

- $\frac{SS_T}{n-1} = \hat{\sigma}^2$
- $\frac{SS_E}{n-p} = \hat{\sigma}^2$
- $\frac{SS_{reg}}{p-1} = \hat{\sigma}^2$

Definitions:

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

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

$$= \frac{\left(\frac{SS(\text{simple model}) - SS(\text{complex model})}{\text{difference in \# of parameters estimated}}\right)}{\frac{SS(\text{complex model})}{n - p(\text{complex model})}}$$

If $\beta_1 = 0 \Rightarrow \cdot\ R^2 \approx 0$

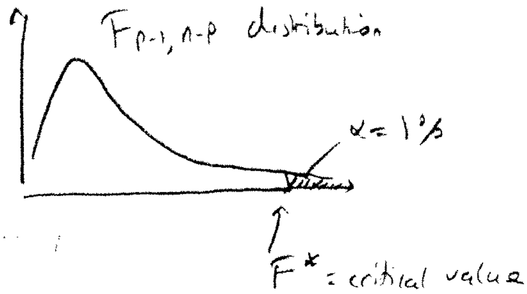
$\cdot\ F_{observed} \approx 1$

If $\beta_1 \neq 0 \Rightarrow \cdot\ R^2 > 0$ (but not nec. very large — important is not the same as significant)

$\cdot\ F_{observed} >> 1$

<u>Distribution theory</u>: If $\varepsilon \sim N(0, \sigma^2)$ and $\beta_1 = 0$

$$F_{observed} \overset{d}{=} F_{(p-1),(n-p)}$$



We reject the <u>null hypothesis</u> $\beta_1 = 0$ if

$$F_{observed} > \underbrace{F_{p,n-p}(1-\alpha)}_{1-\alpha\text{-quantile of the F-distribution}}$$

$\Downarrow$

Conclusion: $y \sim x$ appear to be related

(Discuss. significance vs importance)

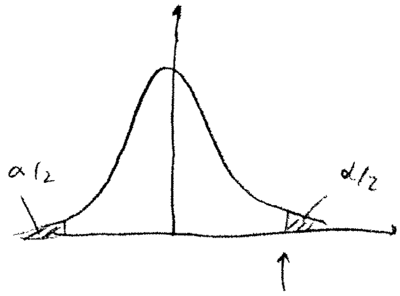
The t-test (if $\varepsilon \sim N(0, \sigma^2)$)

- $$\frac{\hat{\beta}_1 - \beta_1^{true}}{SE(\hat{\beta}_1)} \overset{d}{=} t_{n-p}$$, where $SE(\hat{\beta}_1) = \sqrt{\dfrac{\hat{\sigma}^2}{\sum_j (x_j - \bar{x})^2}}$ (standard error)

- Null hypothesis: $\beta_1 = 0$ $\Rightarrow$ Test-statistic

$$\boxed{t_{observed} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}}$$

$\Rightarrow$ If $\beta_1^{true} = 0$ $\Rightarrow$ $t_{observed} \overset{d}{=} t_{n-p}$

- 

$t_{n-p}(1-\alpha/2)$ critical value

$$\boxed{\begin{array}{l}\text{We reject the null hypothesis } \beta_1 = 0 \text{ if} \\ |t_{observed}| > t_{n-p}(1-\alpha/2)\end{array}}$$

Discuss: one-sided vs two-sided tests!

$\Rightarrow$ p-values!

Duality Testing & Interval estimates

- $I_{\beta_1} = \left[\hat{\beta}_1 \pm t_{n-p}(1-\alpha/2)\, SE(\hat{\beta}_1) \right]$

  is the $100(1-\alpha)\%$ confidence interval for β_1

- If I_{β_1} covers 0, we cannot reject the null hypothesis $\beta_1 = 0$

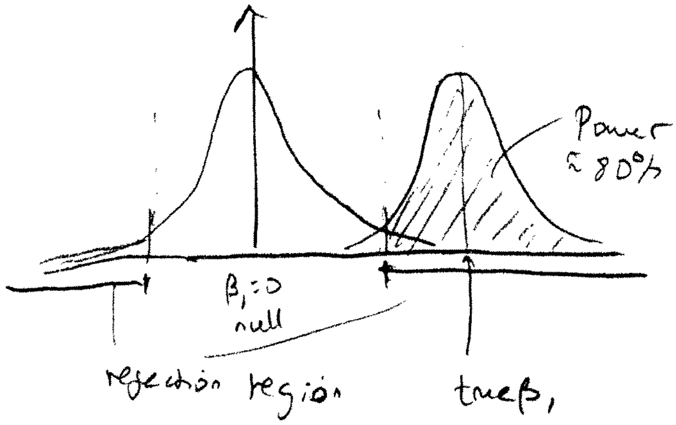
◦ Hypothesis testing

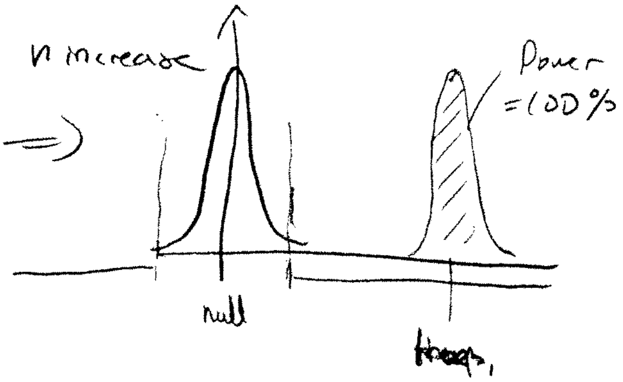
$$\begin{cases} H_0: \beta_1 = 0 \\ H_A: \beta_1 \neq 0 \end{cases} \qquad t_{observed} = \frac{\hat{\beta}_1 - 0}{\sqrt{\dfrac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}}}$$

If $|t_{observed}| > t_{n-p}(1-\alpha/2)$, reject H_0!

[Note, w. prob α we reject H_0 when H_0 is true (type I error)]

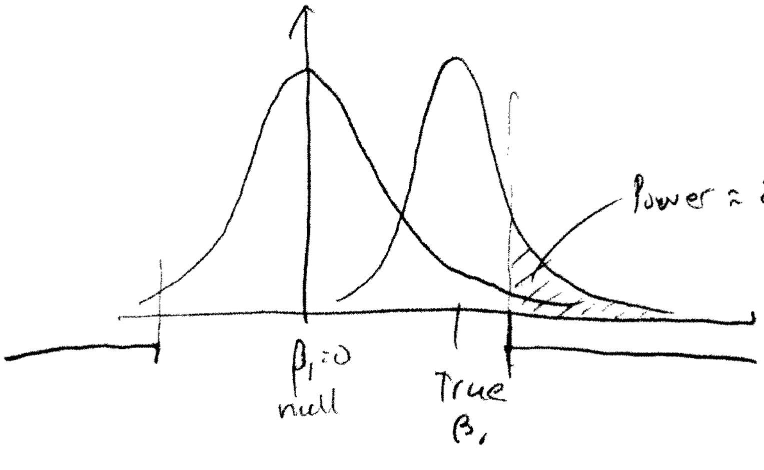
Power – depends on n, the β_1





⇓

β_1 true closer to 0 (null)



Inference on the line

- $\hat{y}_i$ is a statement about the average outcome (or expected) of y @ $X = x_i$

- We know $E(\hat{y}_i) = E(y_i) = \beta_0 + \beta_1 x_i$

- Now, $V(\hat{y}_i) = V(\hat{\beta}_0 + \hat{\beta}_1 x_i) = V(\bar{y} + \hat{\beta}_1 (x_i - \bar{x})) \overset{(*)}{=} V(\bar{y}) + (x_i - \bar{x})^2 V(\hat{\beta}_1)$

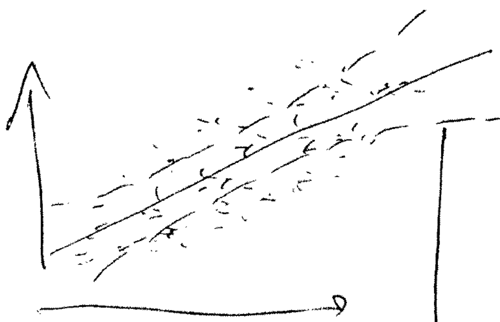
[since $Cov(\bar{y}, \hat{\beta}_1) = Cov\left(\frac{1}{n} \sum y_j, \sum k_i y_i\right) = \sum_j \frac{k_j}{n} Cov(y_j, y_j) = \frac{\sigma^2}{n} \sum k_j = 0$]
so variances in (*) add up

$$\Rightarrow V(\hat{y}_i) = \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_j - \bar{x})^2} \right) = \sigma^2 h_{ii}$$

$\sim \frac{1}{n}$, expected since we're making inferences based on an average

Increases w. leverage since small change to y_i when h_{ii} is large has a huge impact

ie, more uncertainty for extreme x_i values.



pointwise!

CI: $\left[\hat{y}_i \pm t_{n-2}\left(1 - \frac{\alpha}{2}\right) \sqrt{\hat{\sigma}^2 h_{ii}} \right]$

Note, $V(e_i) = \sigma^2 (1 - h_{ii})$ → Variance of residuals ↓ w leverage since regression line forced towards these observations, making residuals small.

residuals do not have equal variance → standardize for direct comparison $\tilde{e}_i = e_i/$

- Predictions - what if we want to make inferences about a new / future observation.

Ex: What is your best interval estimate for the brainweight of an animal w. body weight x_{new}?

Well, one more source of error.

- If we knew the true regression line, our best guess would be $\beta_0 + \beta_1 x_{new}$, give or take $\sqrt{\sigma^2}$
- but we had to use a data set to estimate $\beta_0 + \beta_1 x_{new}$ by $\hat{\beta}_0 + \hat{\beta}_1 x_{new}$, so we're likely to be off by more than just $\sqrt{\sigma^2}$
- Uncertainty of prediction → uncertainty about regression line (estimation uncertainty)

  +

  → irreducible error, the unpredictable part, σ^2

$$\Rightarrow V(\hat{y}(x_{new})) = V(\hat{\beta}_0 + \hat{\beta}_1 x_{new}) + \sigma^2$$

$$= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{\sum (x_j - \bar{x})^2} \right)$$

* CI: $\left[\hat{y}(x_{new}) \pm t_{n-2}(1 - \alpha/2) \sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{\sum (x_j - \bar{x})^2} \right)} \right]$

covers a new observation y_{new} w. prob 95%

Inference in multiple case

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

(Note however, as p gets large, R^2 may be misleading
→ does it really "pay off" to use all x's?

$\Rightarrow R^2_{adjusted} = 1 - \frac{SS_E}{SS_T} \cdot \frac{n-1}{n-p} = 1 - \frac{MSE}{MST}$)

- The t-test

We can construct separate CI's for each $\hat{\beta}_k$ as

$$\left[\hat{\beta}_k \pm t_{n-p}(1-\alpha/2)\, SE(\hat{\beta}_k) \right] = I_{\beta_k}$$

but remember some $\hat{\beta}$'s are correlated, and performing separate t-tests can be misleading.

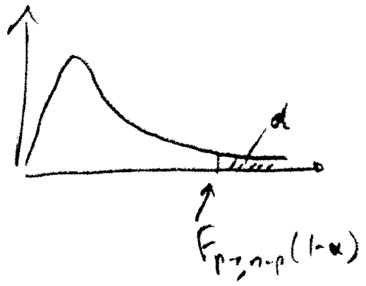
- The F-test ⇒ Now has multiple uses

(A) Lack-of-fit : $H_0: \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0 \Rightarrow SS_T$
H_A: At least one $\beta_k \neq 0 \Rightarrow SS_E$

$$F_{observed} = \frac{(SS_T - SS_E)/((n-1)-(n-p))}{SS_E/(n-p)}$$

If H_0 is true, $F_{observed} =^d F_{p-1,n-p}$

$\rightarrow$ Reject H_0 @ α-level if $F_{observed} > F_{p-1,n-p}(1-\alpha)$



(B) Subset model selection

$$\begin{cases} H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_A: \text{At least one of } \beta_1, \ldots, \beta_k \neq 0 \end{cases}$$

(out of $p-1$ regression coefficients $\{\beta_1, \beta_2, \ldots, \beta_{p-1}\}$)

"Full model" under H_A $\Rightarrow$ SS_f with $df = n-p$

"Reduced model" under H_0 $\Rightarrow$ SS_r with $df = n-(p-k)$

$$F_{observed} = \frac{(SS_r - SS_f) \big/ ((n-(p-k)) - (n-p))}{\frac{SS_f}{n-p}} = \frac{(SS_r - SS_f)/k}{\frac{SS_f}{n-p}}$$

If H_0 is true, $F_{observed} =^d F_{k,n-p}$

k: difference in number of parameters

$n-p$: df of full model

(C) Fixed model (less common)

$$\begin{cases} H_0: \beta_1 = 1, \beta_2 = 5 \\ H_A: \beta_1 \neq 1 \ \& \ \beta_2 \neq 5 \end{cases}$$

H_0 → Fit model w. β_1 & β_2 restricted & compute corresponding $SS = SS_r$

H_A → Fit model without restrictions & compute corresponding $SS = SS_f$

$$F_{observed} = \frac{(SS_r - SS_f)/2}{\frac{SS_f}{n-p}}$$

← Difference in # of parameters estimated

$$\overset{d}{=} F_{2, n-p} \quad \text{if } H_0 \text{ is true}$$

(D) Special cases

$$\begin{cases} H_0: \beta_1 = \beta_2 \Rightarrow \text{Fit model with } \beta_1 = \beta_2 \text{ restriction} \Rightarrow SS_r \\ H_A: \beta_1 \neq \beta_2 \Rightarrow \text{Fit unrestricted model} \Rightarrow SS_f \end{cases}$$

$$F_{obs} = \frac{(SS_r - SS_f)/1}{\frac{SS_f}{n-p}}$$

← only one parameter estimate for β_1 & β_2 need in the restricted model, 2 in the full model.

Discussion

① Why not use t-tests to come up with a subset model $\{k : \hat{\beta}_k$ significantly different from $0\}$?

→ multiple testing

→ dependencies

⟹ simultaneous CI
one solution
or
subset model selection

Ext: P(rej H_0|H_0) $= \alpha = 0.05$

10 test
P(at least one false reject.)
$= 1-(1-\alpha)^{10} = 0.098$

100
≈ 0.63

② Which subset models to consider?

If we have $p-1$ explanatory variables, there are 2^{p-1} possible models

$(p-1)$ # of x's	# of models
1	2
2	4
3	8
10	1024
20	10^6
30	10^9

$\beta_1 = 0, \beta_1 \neq 0$

$(\beta_1 = 0, \beta_2 = 0), (\beta_1 = 0, \beta_2 \neq 0), (\beta_1 \neq 0, \beta_2 = 0), (\beta_1 \neq 0, \beta_2 \neq 0)$

⟹ If $p \gtrsim 30$ ⟹ not even the best software packages are set up to perform all searches.

Alternatives → Directed searches

<u>Backward selection</u>

(0) Fit the full model, with all p parameters $\to SS_f$

(1). Examine the $(p-1)$ subset models corresponding to dropping one of the x's.

- Compute the corresponding RSS $= SS_r(-x_k)$ $k=1,\dots,p-1$

- Identify the x_ℓ for which $SS_r(-x_\ell) = \min_k SS_r(-x_k)$, i.e. the x which increases the SSE the least.

(2) If $F = \dfrac{(SS_r(-x_\ell) - SS_f)/1}{\frac{SS_f}{n-p}} < F_{1,n-p}(1-\alpha)$

(i) → we don't reject null $\beta_\ell = 0$

→ Drop x_ℓ from the model & set

$\circ\ p = p-1$

$\bullet\ SS_f = SS_r(-x_\ell)$

Go to (1)

(ii) If we reject the null $\beta_\ell = 0$ → Stop the search & retain the most recent model.

Forward search → Start with only the intercept

→ Add the x that reduces the RSS the most

→ Stop adding when null is not rejected.

<u>Concerns</u>

- greedy search
- if variable dropped (added) – cannot reverse the decision (⇒ • add random element • stochastic searches)
- Ignores selection uncertainty → more later
- Can lead to models that are difficult to interpret, especially in the case of strong dependencies between variables.