Inference for Simple Linear Regression

independent & identically distributed

Mathematical Model: Theoretical equation to predict *y* from *x* in the population.

$$y = \alpha + \beta x + \varepsilon, \quad \underset{\text{random error (a random variable)}}{\varepsilon} \overset{iid}{\sim} N(\underset{\text{mean}}{0}, \underset{\text{variance}}{\sigma^2})$$

Suppose we have n observations (x_i, y_i), $i = 1, 2, \cdots, n$.

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

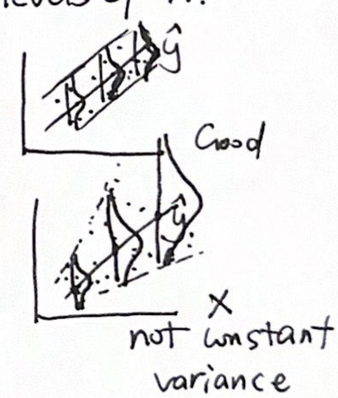
$$\Rightarrow y_i \overset{iid}{\sim} N(\alpha + \beta x_i, \sigma^2) \qquad \mu_{y_i} = \alpha + \beta x_i$$

- Lemma: If r.v. $X \sim N(\mu, \sigma^2)$, then $X + c \sim N(\mu + c, \sigma^2)$

Assumptions needed to make inferences about population from simple linear regression

$$\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$$

- SRS of exp units from population of interest
- Normal distribution of resp. variable in population for each level of x. — NO OUTLIERS
- Constant variance of resp. variable at all levels of x.
  - — NO OUTLIERS
  - — NO FUNNEL SHAPE



Summary of Model:

Model: $y = \alpha + \beta x + \varepsilon$

Assumption: $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$

Parameters & Estimates: For α: $a = \bar{y} - b\bar{x}$

For β: $b = r\dfrac{S_y}{S_x}$

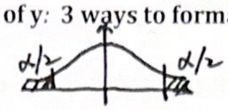
For σ: $s = \sqrt{MSE}$ (pooled stder)

The following 3 questions are equivalent:

1. Is x a good predictor of y? 2. Are x and y ~~independent~~ (associated)? 3. Is ~~sb~~ the slope SIG diff from zero?

Tests to determine if x is a GOOD predictor of y: 3 ways to formally test this.

(1−α) two-sided



1. Confidence Interval for β:

$$\text{estimator} \pm t \cdot \text{Stderror}$$

$$b \pm t_{n-2,\frac{\alpha}{2}} \cdot \cancel{Sd_b}\, s.e._b$$

from t-table w/desired confidence & df from ERROR.

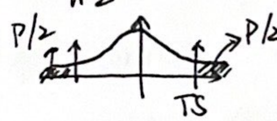
β sig diff from zero if CI does NOT include zero.

2. t-test for β:

H_0: $\beta = 0$ $\quad$ H_a: $\beta \neq 0$

$$TS: \; t = \frac{\text{estimator} - \#}{\text{stderr}} = \frac{b-0}{s.e._b} = \frac{b}{s.e._b}$$

pvalue: $TS \sim t_{n-2}$ when H_0 is True.



β sig diff from zero if p-value $\leq \alpha$.

3. ANOVA Test:

Source	df	SS	MS$\left(=\frac{SS}{df}\right)$	F	p-val
Regression	1	SSR	MSR	$\frac{MSR}{MSE}$	$P(F_{1,n-2} \geq F)$ $(F \sim F_{1,n-2})$
Error	n−2	SSE	MSE		
Total	n−1	SST			

NOTE: Relationship between R^2 and SS: