REGRESSION

HAVE OBSERVATIONS OF A DEPENDENT VARIABLE Y, AND K INDEPENDENT VARIABLES

$$X_1, X_2, \ldots X_K$$

CONSIDER LINEAR RELATIONSHIP

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_K X_K$$

"Y-HAT"

b_0 ↑ INTERCEPT

b_1 ↑ SLOPE COEFF

↓ PREDICTED, OR FITTED VALUE

→ EFFECT OF UNIT INCREASE IN ONE VARIABLE X_i, HOLDING OTHER VARIABLES FIXED

LET $\hat{y}_i$ BE THE FITTED VALUE FOR i^{th} OBSERVATION

LET y_i BE OBSERVED VALUE

LET $e_i = y_i - \hat{y}_i$ BE CALLED OR RESIDUAL

USUAL CRITERION:

MINIMIZE $\sum_{i=1}^{n} e_i^2$

"LEAST SQUARES" REGRESSION

✓ COMPUTATIONALLY TRACTABLE [UP TO REASONABLE SIZE]

✓ PERMITS INFERENCE UNDER APPROPRIATE ASSUMPTIONS

MEASURES OF FIT

$$SSY = SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

"SUM OF SQUARES" — "TOTAL"

$$SS\hat{Y} = SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

"REGRESSION"

$$SSE = \sum_{i=1}^{n} e_i^2$$

ERROR

FACT FOR L-S FIT

$$SSE + SSR = SST$$

$$\frac{SSE}{SST} + \frac{SSR}{SST} = 1$$

"R^2" FRACTION OF VARIATION IN y "EXPLAINED BY" INDEP. VARIABLES *

* USING LINEAR RELATIONSHIP

ANOTHER MEANING OF RSE

$$s_e = \sqrt{\frac{SSE}{n-(k+1)}} \quad \ldots.$$

"STD. ERROR OF ESTIMATE"
(ESSENTIALLY SAMPLE STD. DEV. OF RESIDUALS)

INFERENCE

IDEA: WANT MEASURE FOR POSSIBLE ERROR IN OUTPUTS $b_0, b_1, \ldots b_k$

NEED TO ASSUME A "MODEL"

$$Y = \underbrace{\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k}_{\text{fixed}} + \underbrace{\varepsilon}_{\text{random}}$$

ε IS NORMAL R.V.

$$E[\varepsilon] = 0$$
$$\text{VAR}[\varepsilon] = \sigma_\varepsilon^2$$

ε IS INDEPENDENT OF

$X_1, X_2, \ldots X_k$

UNDER MODEL ASSUMPTION, OUTPUTS OF L-S REGRESSION HAVE KNOWN "SAMPLING DISTRIBUTIONS"....

BOTTOM LINE: CONSTRUCT CONFIDENCE INTERVALS FOR $\beta_0, \beta_1, \ldots \beta_k$

HOW TO EVALUATE REGRESSION MODELS:

(1) INTERPRETATION OF EQUATION

(2) RESIDUALS
- PLOTS AGAINST X_i
- HISTOGRAM