

Details on Inference in Regression

Key Assumptions: 1. - 4.

1. pop of (X, Y) values X = weight of ISU student (or bear neck width),
 & Y = height of ISU student (or bear weight)
- sample n pairs $(X_1, Y_1), \dots, (X_n, Y_n)$ to study pop.

Note: some individuals in pop can have same X -value & Y -values are all independent (e.g. one bear's weight Y doesn't influence a second bear's weight Y)

3. For all individuals in pop with a given X value (e.g. weight), there is a mean/typical value of the response Y :

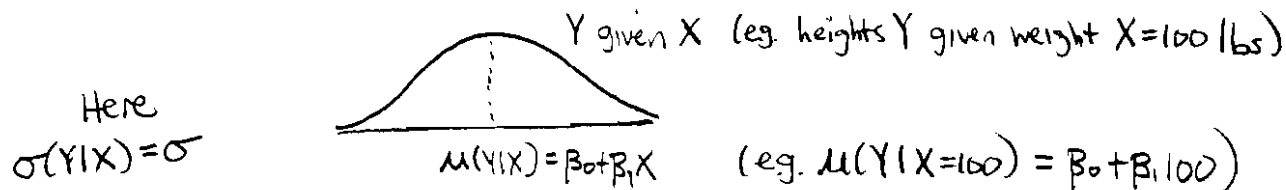
$$\mu(Y | X) = \beta_0 + \beta_1 X \quad \begin{array}{l} \beta_1 = \text{population slope} \\ \beta_0 = \text{population intercept} \end{array}$$

The line above $\mu(Y | X) = \beta_0 + \beta_1 X$ is called the "population regression line".

It tells, *in the entire population*, how the mean value of Y depends on which X we consider (i.e., some Y -values, like bear weights, will be more common given the value of X , given the bear neck width for example).

The goal is to understand this pop regression line using the data.

2. For all individuals in pop. with given X value, the values of Y in the pop are normal with mean/typical value $\mu(Y | X) = \beta_0 + \beta_1 X$



We say the "distribution $Y | X$ " of Y -values given an X -value is normal.

4. The spread (or standard deviation) of Y -values for all individuals in the pop with the same X -value is

$$\sigma(Y | X) = \sigma \quad \leftarrow \begin{array}{l} \text{common value} \\ \text{(same for all } X) \end{array}$$

To state Assumptions 1-4 in another way: for our sample $(X_1, Y_1), \dots, (X_n, Y_n)$, we can write

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

$\mu(Y|X_i)$
 ith Y-value \rightarrow Y_i
 $\underbrace{\beta_0 + \beta_1 X_i}$ mean/typical value of Y for individuals in pop with explanatory variable X_i
 e_i random error or deviation term (normally distributed with mean/typical value 0 + standard deviation σ)

Now for inference on the pop:

• **Important:** we don't know $\beta_0, \beta_1 \Rightarrow$ estimate these with $\hat{\beta}_0, \hat{\beta}_1$ from a sample
 remember: just like \bar{Y} , estimates $\hat{\beta}_0, \hat{\beta}_1$ can change from sample to sample

• Also: we don't observe/know e_1, \dots, e_n (have only (X_i, Y_i) from data)
 but, can use residual \hat{e}_i from i th pair (X_i, Y_i)

$$\hat{e}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \quad \text{to approximate} \quad e_i = Y_i - (\beta_0 + \beta_1 X_i)$$

\uparrow \uparrow \uparrow
 ith residual \hat{e}_i i th observation Y_i \hat{Y}_i i th observation \hat{Y}_i

• Finally: we don't know common spread σ either, but estimate this with

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n \hat{e}_i^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}} = S_Y \sqrt{(1-r^2) \frac{n-1}{n-2}}$$

\uparrow
 kind of resembles a sample standard deviation of residuals (i.e., estimates spread)

• Again $\hat{\beta}_1, \hat{\beta}_0$ are estimates & can vary some sample to sample \Rightarrow
 $\hat{\beta}_1, \hat{\beta}_0$ have a sampling distribution
 (when we collect data, some values of $\hat{\beta}_1, \hat{\beta}_0$ are more likely to occur than others)

▷ Mean($\hat{\beta}_1$) = β_1 & Mean($\hat{\beta}_0$) = β_0

▷ standard errors SE($\hat{\beta}_1$) or SE($\hat{\beta}_0$) (i.e., estimate how precise $\hat{\beta}_1, \hat{\beta}_0$ are)

▷ $t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$ follows t -distribution with $df = n - 2$, so does $t = \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)}$

▷ get tests & CIs for β_1, β_0

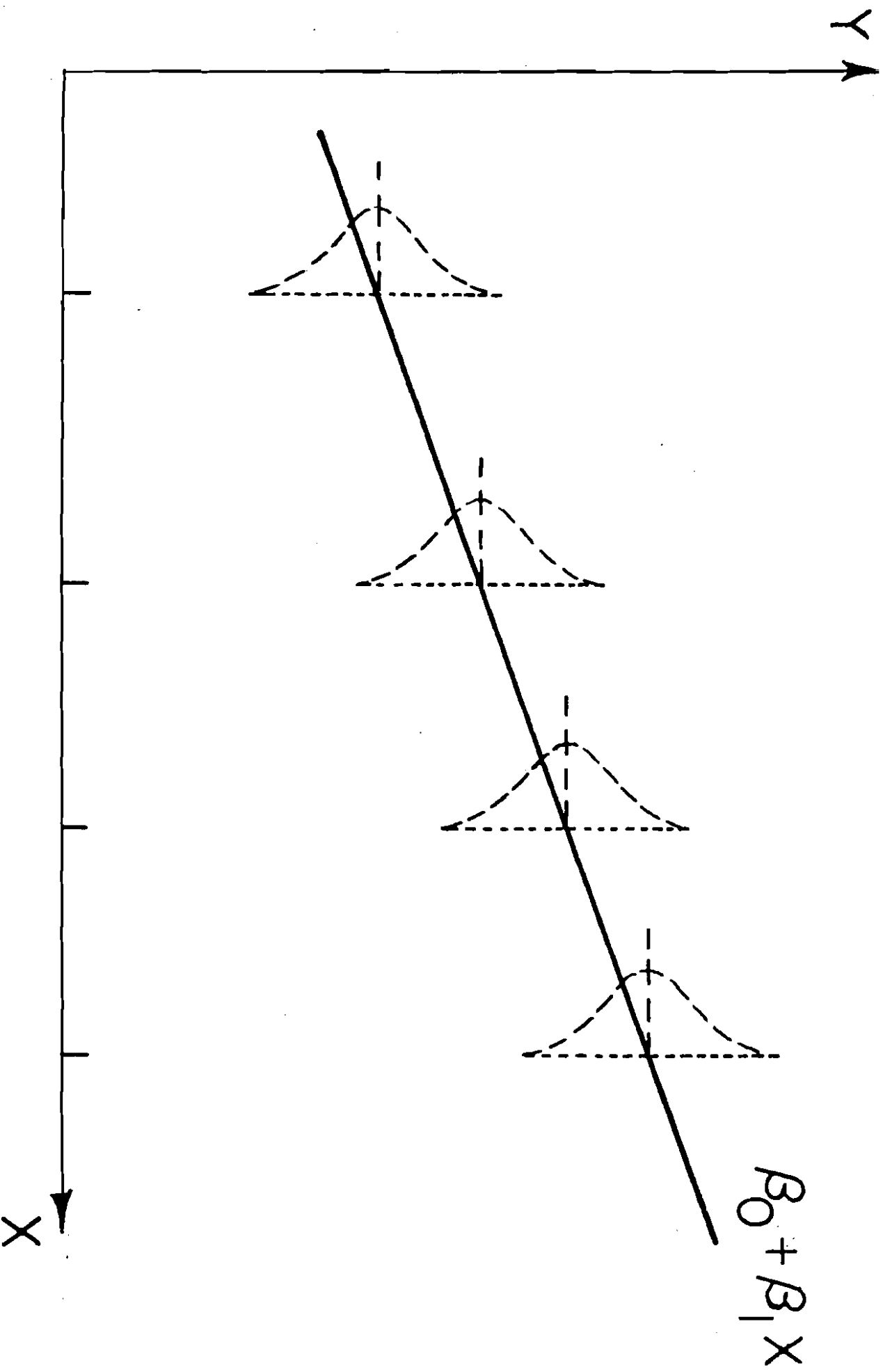


Fig. 9.3.1—Representation of the linear regression model. The normal distribution of Y about the regression line $\beta_0 + \beta_1 X$ is shown for four selected values of X .