In order to be able to assess the "goodness" of our estimators: $\hat{\alpha} = a$, $\hat{\beta} = b$, we need to make some assumptions concerning the distribution of the random variables we are dealing with and about the true nature of the regression line.

Inferences based on the L.S.E

The L.S.E (least square estimates) are used when the relationship between x and the mean of Y (EY) is linear or close enough.

Assumptions:

- Regression is linear
- $Y_1, Y_2, \ldots, Y_n$ are independent
- $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$, $i = 1, 2, \ldots, n$

Mean is the deterministic, i.e <u>no random</u> part of the regression

Variance constant

Statistical Model for Straight Line Regression

$Y_i = \alpha + \beta x_i + \varepsilon_i$, $i = 1, \ldots, n$

where $\varepsilon_i \sim N(0, \sigma^2)$

Show that if $\varepsilon_i \sim N(0, \sigma^2)$ then

$Y_i \sim N(\alpha + \beta x_i, \sigma^2)$

Inference on LSE.

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \underbrace{\left(y_i - (a + bx_i)\right)^2}_{\text{residual sum of squares}}$$

$\hat{\sigma}$ - standard error of estimate

<u>OR</u>

$$\hat{\sigma}^2 = \frac{S_{yy} - (S_{xy})^2 / S_{xx}}{n-2} = MSE$$

Under the previous assumptions

$$t = \frac{\hat{\alpha} - \alpha}{\hat{\sigma}} \cdot \sqrt{\frac{n S_{xx}}{S_{xx} + n(\bar{x})^2}} \sim t(n-2)$$

$$t = \frac{\hat{\beta} - \beta}{\hat{\sigma}} \cdot \sqrt{S_{xx}} \sim t(n-2)$$

These are the statistics we will use to do inference on α and β, that is, confidence intervals and hypothesis tests. Point est. have been already derived.

Confidence Intervals.

For a given confidence level α, a $(1-\alpha)\cdot 100\%$ confidence Interval, (C.I) for the regression parameters α and β is given by:

$$\alpha \in \left(\hat{\alpha} \pm t_{\alpha/2, n-2} \cdot \hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}} \right)$$

$$\beta \in \left(\hat{\beta} \pm t_{\alpha/2, n-2} \cdot \hat{\sigma} \cdot \frac{1}{\sqrt{S_{xx}}} \right)$$

Remember:

$$\hat{\alpha} = a = \bar{y} - b \cdot \bar{x}$$

$$\hat{\beta} = b = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\sigma} = \sqrt{\frac{S_{yy} - S_{xy}^2 / S_{xx}}{n-2}}$$

Confidence Interval for the α

__Example 3 (continue)__ Construct a 95% C.I for the regression coefficient α.

We know from Example 3 that

$n = 10$, $\bar{x} = 200$, $S_{xx} = 132000$, $S_{yy} = 2.13745$

$S_{\cancel{x}y} = 505.40$.

Then:

$$\hat{\sigma}^2 = \frac{2.13745 - \frac{(505.40)^2}{132{,}000}}{8} = \frac{0.20238}{8} = \underline{\underline{0.0253}}$$

$$\Rightarrow \hat{\sigma} = \sqrt{0.0253} = 0.159$$ standard error of estimate.

$t_{0.025,8} = 2.306$, a 95% C.I for α is given by

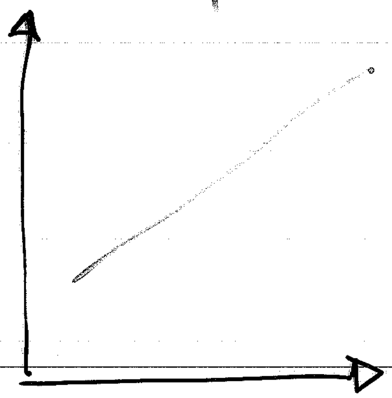
$$\alpha \in \left(\hat{\alpha} \pm t_{\frac{\alpha}{2}, n-2} \cdot \hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}\right) = \left(0.069 \pm 2.306 \cdot 0.159 \cdot \sqrt{\frac{1}{10} + \frac{200^2}{132{,}000}}\right) = (-0.164, 0.302)$$

line could pass through origin!

β - slope of regression = the change in the mean of Y corresponding to a unit increase in x.

$\beta = 0$ line horizontal, EY does not depend on x

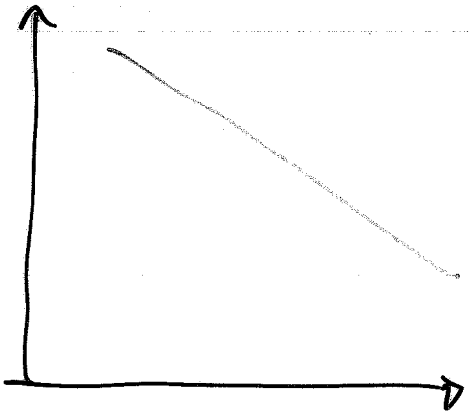
$\beta > 0$



possitive association

large (small) values of Y are associated with large (small) values of x

$\beta < 0$



negative association

Large (small) values of Y are associated with small (large) values of x.

Hence it is of interest to test whether $\beta = 0$.

Hypothesis test on β.

$H_0 : \beta = \beta_0$

$H_a : \beta \neq \beta_0$

$$t_0 = \frac{\hat{\beta} - \beta}{\hat{\sigma}} \cdot \sqrt{S_{xx}} \sim t_{(n-2)}$$

Reject H_0 if $t_0 < -t_{\alpha/2, n-2}$ or $t_0 > t_{\alpha/2, n-2}$.

Example 3 continue II

$H_0 : \beta = 0$

$H_a : \beta \neq 0$

$\alpha = 0.05 \Rightarrow t_{0.025, 8} = 2.306$

$$t_0 = \frac{0.00383 - 0}{0.159} \cdot \sqrt{132000} = 8.75$$

Since $t_0 = 8.75 > 2.306 = t_{0.025, 8}$ we reject the null hypothesis and conclude: that there is a relationship between air velocity and the average evaporation coefficient. Scatter plot tells me is a linear relationship.

Mean Response

We are interested in estimating the regression $\alpha + \beta x$, i.e the mean of the distribution of Y for a given value of x, say x_0. It is reasonable to do so by using

$$\hat{Y} = \hat{\alpha} + \hat{\beta} \cdot x_0 = a + b x_0, \quad \boxed{a, b - \text{L.SE}}$$

A $(1-\alpha) \cdot 100\%$ C.I for the mean response

$$\boxed{Y \in \left(\hat{Y} \pm t_{\alpha/2,\, n-2} \cdot \hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right)}$$

Example 3. Construct a 95% C.I for the mean evaporation coefficient when the air velocity is 190 cm/s.

$$Y \in \left(0.80 \pm 2.306 * 0.159 \sqrt{\frac{1}{10} + \frac{(190-200)^2}{132{,}000}} \right) =$$

$$= (0.68, 0.92)$$

Hence a 95% C.I for the mean evaporation coefficient of $\alpha + 190\beta$ is

$(0.68, 0.92)$

Prediction of new value

Even more important than estimating $\alpha + \beta x_0$ is the prediction of a future value of Y when $x = x_0$ and x_0 is within the range of experimentation.

Method: We want to predict a future obs Y when $x = x_0$ with probability α.

If α and β were known, then since

$$Y \sim N(\alpha + \beta x_0, \sigma^2) \iff Y - \alpha - \beta x_0 \sim N(0, \sigma^2)$$

BUT α, β are NOT known and estimated by $\hat{\alpha} = a$, $\hat{\beta} = b$. A $(1-\alpha).100\%$ C.I

$$Y_{new} \in \left(\underset{\substack{\| \\ a + bx_0}}{\hat{Y}_{new}} \pm t_{\alpha/2, n-2} \cdot \hat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right)$$

Example 3 A 95% "prediction" interval for an observation of the evaporation coefficient when air velocity = 190 cm/s.

$$Y_{new} \in \left(0.80 \pm 2.306 * 0.159 \cdot \sqrt{1 + \frac{1}{10} + \frac{(190-200)^2}{132000}} \right) =$$

$$= (0.41, 1.19)$$

Interval of prediction is __wider__ than the C.I for the mean response!

More rigorous "proofs" of the facts presented here, follow closely the arguments we have already presented in Ch 5 and 6. If you want to have a look they are repeated at Ch 7 of your book!!