

(1)  $\hat{y}_i = \beta_0 + \beta_1(x_i - \bar{x}) + \varepsilon$ ,  $\sum (y_i - \hat{y})^2$  is what we minimize

$y = \beta_0 + \beta_1 x_i - \beta_1 \bar{x} \Rightarrow$  minimize  $\sum (y_i - \beta_0 - \beta_1 x_i + \beta_1 \bar{x})^2$

$\frac{\partial S}{\partial \beta_0} = -2 \sum (y_i - \beta_0 - \beta_1 x_i + \beta_1 \bar{x}) = 0 \Rightarrow \sum y_i - \varepsilon \beta_0 - \sum \beta_1 x_i + \sum \beta_1 \bar{x} = 0$

$n\bar{y} - n\beta_0 - n\bar{x}\beta_1 + n\bar{x}\beta_1 = 0 \Rightarrow \beta_0 = \bar{y}$

$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2)^{1/2}} = \frac{S_{xy}}{(S_{xx} \cdot S_{yy})^{1/2}}$ , Prove  $r^2 = R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$

$r = \frac{\sum y_i (x_i - \bar{x}) - \bar{y} \sum (x_i - \bar{x})}{(S_{xx} \cdot S_{yy})^{1/2}}$ ;  $\sum \bar{y} (x_i - \bar{x}) = \sum \bar{y} x_i - \sum \bar{y} \bar{x} = n\bar{x}\bar{y} - n\bar{x}\bar{y} = 0$

numerator  $r = \sum y_i (x_i - \bar{x}) = \sum x_i (y_i - \bar{y})$

$\sum (\hat{y}_i - \bar{y})^2 = \sum (\beta_0 + \beta_1 x_i - \beta_1 \bar{x} - \bar{y})^2 = \sum (\bar{y} + \beta_1 x_i - \beta_1 \bar{x} - \bar{y})^2$   
 $= \sum (\beta_1 x_i - \beta_1 \bar{x})^2 = \beta_1^2 \sum (x_i - \bar{x})^2$

Solve for  $\beta_1$

$\frac{\partial S}{\partial \beta_1} = -2 \sum (x_i - \bar{x})(y_i - \beta_0 - \beta_1(x_i - \bar{x})) = 0 \Rightarrow \sum [(-x_i + \bar{x})y_i - \beta_0(-x_i + \bar{x}) - \beta_1 x_i(-x_i + \bar{x}) + \beta_1 \bar{x}(-x_i + \bar{x})] = 0$

$\sum (-x_i y_i + \bar{x} y_i + \beta_0 x_i - \beta_0 \bar{x} + \beta_1 x_i^2 - \beta_1 x_i \bar{x} - \beta_1 \bar{x} x_i + \beta_1 \bar{x}^2) = 0$

$\sum x_i y_i - n\bar{x}\bar{y} = n\beta_0 \bar{x} - n\beta_0 \bar{x} + \beta_1 \sum x_i^2 - \beta_1 n\bar{x}^2 - \beta_1 n\bar{x}^2 + n\beta_1 \bar{x}^2$

$\beta_1 = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{S_{xy}}{S_{xx}}$  (same as formula)

$R^2 = \frac{(S_{xy})^2}{S_{xx} \sum (x_i - \bar{x})^2} = \frac{(S_{xy})^2}{S_{xx} S_{xx}} \Rightarrow R^2 = \frac{(S_{xy})^2}{(S_{xx})^2}$

$R^2 = \frac{[\sum (x_i - \bar{x})(y_i - \bar{y})]^2}{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2} = r^2 = \left[ \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{[\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2]^{1/2}} \right]^2$

$\Rightarrow R^2 = r^2$  for  $\hat{y}_i = \beta_0 + \beta_1(x_i - \bar{x}) + \varepsilon$



2.  $R^2 = 0.82$ ,  $S_{yy} = 50$ ,  $n = 25$ , 95% PI of  $y$  for  $x = \bar{x}$

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{res}}{SS_T}, \quad SS_{res} = \sum (y_i - \hat{y})^2, \quad SS_R = \sum (\hat{y}_i - \bar{y})^2$$

$$S_{yy} = \sum (y_i - \bar{y})^2, \quad SS_T = SS_R + SS_{res}, \quad SS_{TOT} = \sum (y_i - \bar{y})^2 = S_{yy} = 50$$

$$R^2 = 1 - \frac{SS_{res}}{SS_T} \Rightarrow .82 = 1 - \frac{SS_{res}}{SS_T} \Rightarrow (SS_T)(.82) = SS_T - SS_{res}$$

$$SS_{res} = SS_T - (SS_T)(.82) = 50 - (50)(.82) = 9$$

$$SS_R = SS_T - SS_{res} = 50 - 9 = 41$$

$$MS_{res} = \hat{\sigma}^2 = SS_{res}/n-2 = 9/23 = .3913$$

$$B_1 = \frac{S_{xy}}{S_{xx}}, \quad R^2 = \hat{B}_1^2 \frac{S_{xx}}{SS_T} \Rightarrow \hat{B}_1 = \sqrt{\frac{R^2 SS_T}{S_{xx}}} = \sqrt{\frac{.82 \cdot 50}{S_{xx}}} \rightarrow \text{still don't know } S_{xx}, \text{ oops}$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \Rightarrow r^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} \Rightarrow$$

$$\psi = y_0 - \hat{y}_0$$

$$E(y|x=\bar{x}) = \hat{B}_0 + \hat{B}_1 \bar{x}, \quad \text{Var}(\psi) = \text{var}(y_0 - \hat{y}_0) = \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

$$\text{Standard error for } \psi = \sqrt{MS_{res} \left( 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}$$

$$\text{for } x_0 = \bar{x} \text{ this simplifies to } \sqrt{MS_{res} \left( 1 + \frac{1}{n} \right)} = \sqrt{MS_{res} \left( 1 + \frac{1}{25} \right)} = \sqrt{1.04 \cdot MS_{res}}$$

$$\Rightarrow SE = .6379$$

$$t_{.05/2, 23} = 2.069$$

$$\text{Prediction Interval: } \hat{y}_0 - (t_{.05, 23})(se) \leq y_0 \leq \hat{y}_0 + (t_{.05, 23}) se$$

$$\hat{y}_0 - 1.32 \leq y_0 \leq \hat{y}_0 + 1.32$$

I tried everything I could think of to find  $B_1/B_0$ , but no  $S_{xx}$  or  $S_{xy}$  term

I do know that OLS passes through  $(\bar{x}, \bar{y}) \Rightarrow \hat{y}_0$  for  $\bar{x} = \bar{y}$

$$\Rightarrow \bar{y} - 1.32 \leq \bar{y} \leq \bar{y} + 1.32$$



(3) (8,10), (8,10), (7,9), (6,10), (3,6), (4,8), (4,8), (5,9), (3,7), (6,9)

$$\bar{x} = \sum x_i / n = (8+8+7+6+3+4+4+5+3+6) / 10 = 5.4$$

$$\bar{y} = \sum y_i / n = (10+10+9+10+6+8+8+9+7+9) / 10 = 8.6$$

$$S_{xx} = \sum (x_i - \bar{x})^2 = (8-5.4)^2 + (8-5.4)^2 + (7-5.4)^2 + (6-5.4)^2 + (3-5.4)^2 + (4-5.4)^2 + (4-5.4)^2 + (5-5.4)^2 + (3-5.4)^2 + (6-5.4)^2 = 32.4$$

$$S_{xy} = \sum y_i (x_i - \bar{x}) = (10)(8-5.4)^2 + 10(8-5.4)^2 + 9(7-5.4)^2 + (10)(6-5.4)^2 + (6)(3-5.4)^2 + (8)(4-5.4)^2 + (8)(4-5.4)^2 + 9(5-5.4)^2 + (7)(3-5.4)^2 + (9)(6-5.4)^2 = 20.6$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{20.6}{32.4} = \boxed{.6358}^{\text{slope}}, \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 8.6 - (.6358)(5.4) = \boxed{5.167}^{\text{y-int}}$$

$$(b) \hat{\text{Var}}(y) = \frac{SS_{\text{res}}}{n-2} = \frac{\sum y_i^2 - n\bar{y}^2 - \hat{\beta}_1 S_{xy}}{n-2}$$

$$\sum y_i^2 = 10^2 + 10^2 + 9^2 + 10^2 + 6^2 + 8^2 + 8^2 + 9^2 + 7^2 + 9^2 = 756$$

$$SS_{\text{res}} = 756 - (10)(8.6)^2 - (.6358)(20.6) = \boxed{3.3031}$$

$$\hat{\sigma}^2 = SS_{\text{res}} / n-2 = 3.303 / 8 = \boxed{.412875} = MS_{\text{res}}$$

(c) Test for statistical confidence of slope

$$H_0: \hat{\beta}_1 = 0, \text{Ita: } \beta_1 \neq 0 \text{ (2 way test)}$$

$$\text{Standard error: } se(\hat{\beta}_1) = \sqrt{MS_{\text{res}} / S_{xx}} = \sqrt{.412875 / 32.4} = .1129$$

$$t_0 = \hat{\beta}_1 / se = .6358 / .1129 = 5.632$$

$$\text{Testing at 95\% significance} \Rightarrow t_{.05, 8} = 1.860$$

Since  $t_0 >> t_{.05, 8} \Rightarrow$  Reject Null hypothesis. I conclude that there is a statistically significant linear relationship between  $x$  and  $y$



(d) Construct 95% CI on slope

100(1- $\alpha$ ) CI given by:  $\hat{\beta}_1 - t_{\alpha/2, n-2} \cdot se(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2, n-2} \cdot se(\hat{\beta}_1)$   
 $t_{0.05/2, 8} = 1.860$ ,  $se(\hat{\beta}_1) = .1129$  (prev problem),  $\hat{\beta}_1 = .6358$

$$CI: (.6358) - (1.860)(.1129) \leq \beta_1 \leq (.6358) + (1.860)(.1129)$$

$$.42581 \leq \beta_1 \leq .8458$$

| (e) Source of Variation | Sum of Squares | Degrees Freedom | Mean Square | F <sub>0</sub> | p-value              |
|-------------------------|----------------|-----------------|-------------|----------------|----------------------|
| Regression              | 13.1           | 1               | 13.1        | 31.76          | $4.9 \times 10^{-4}$ |
| Residual                | 3.3            | 8               | .4125       |                |                      |
| Total                   | 16.4           | 9               |             |                |                      |

$$SS_R = (\hat{\beta}_1) S_{xy} = (.6358)(20.6) = 13.1; SST = SS_R + SS_{res} = 13.1 + 3.303 = 16.4$$

$$MS_R = SS_R / dof = 13.1, MS_{res} = SS_{res} / dof = 3.3/8$$

$$F_0 = MS_R / MS_{res} = 13.1 / .4125 = 31.76; F_0$$

If we test at 95% significance,  $F_{0.05, 1, 8} = 5.32$

Since  $F_0 \gg F_{0.05, 1, 8}$  we reject null hypothesis and say significant

(f)  $\epsilon$  is random variable w/  $\sim N(0, \sigma^2) \rightarrow$  Applicable to all cases

$Y$  is random variable,  $X$  is fixed  $\rightarrow$  Applicable to all cases

Because  $E(\epsilon) = 0$  and  $Var(\epsilon) = \sigma^2$ , OLS estimates are unbiased (a, b, c, d)

LSE are best linear unbiased estimators (minimum variance)

Uncorrelated Choice of  $\hat{X}$  does not affect random error of  $y$  and since mean of  $\epsilon = 0$ , it does not matter what Ind Var we choose (a, b, c)

Regression model is linear in coefficient and error - model must fit linear pattern (a-e)

Error term has constant variance (homoscedasticity) - (a, b)

Since homoscedasticity + no autocorrelation  $\rightarrow$  error term is Independent and identically distributed



Error term is normally distributed - not required to be Normal for unbiased but allows for more accurate testing  $\rightarrow (s, d, e)$

$$(4) \text{Var}(e_i) = \text{Var}(y_i - \hat{y}) = \text{Var}(y_i - (\bar{y} + \beta_1(x - \bar{x}))) = \text{Var}((y_i - \bar{y}) - \beta_1(x - \bar{x}))$$

$$\text{Var}(y_i - \bar{y}) = \text{Var}(y_i) + \text{Var}(\bar{y}) = \sigma^2 + \frac{\sigma^2}{n} \quad (\text{Cov}(y_i, \bar{y}) = 0)$$

$$\text{Var}(\beta_1(x - \bar{x})) = (x - \bar{x})^2 \cdot \text{Var}(\beta_1) = (x - \bar{x})^2 \cdot \frac{\sigma^2}{S_{xx}} \quad (\text{Var}(\beta_1) \rightarrow HW)$$

$$\text{Var}((y_i - \bar{y}) - \beta_1(x - \bar{x})) = \sigma^2 + \frac{\sigma^2}{n} + \frac{\sigma^2}{S_{xx}} \cdot (x - \bar{x})^2 + 2 \cdot \text{Cov}((y_i - \bar{y}), (\beta_1(x - \bar{x})))$$

$$2 \text{Cov}(y_i - \bar{y}, \beta_1(x - \bar{x})) = \left(\frac{1}{n-1}\right) \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$\text{Var}(e_i) = \sigma^2 + \frac{\sigma^2}{n} + \frac{\sigma^2}{S_{xx}} \cdot (x_i - \bar{x})^2 + \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$= \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right] + \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$\text{Var}(E) = \sigma^2 \Rightarrow \text{Var}(e_i) > \text{Var}(E)$  since it gets multiplied by a number bigger than 1, and the covariance term is in there being added

I think this is due to the fact that you have variance contributions for both predicted ( $\hat{y}$ ) and observed ( $y_i$ ) where as the  $E$  term is by itself and contributes to  $y_i$  ( $\beta_0 + \beta_1 x_i + E$ ). So  $y_i$  includes variance from  $E$  and variance from  $\hat{y}$ .

At first I thought  $y_i$  variance would be less since we subtract  $\hat{y}$  from it in  $\text{Var}(y_i - \hat{y})$  but

$$\text{Var}(a - b) = \text{Var}(a) + \text{Var}(b) \text{ not } \text{Var}(a) - \text{Var}(b)$$