

**The Simple Linear Regression Model (Reference: Chapter 12 of Devore's 7<sup>th</sup> Edition)** **Maghsoodloo**

The objective of simple linear regression (SLR or SLREG) is to determine if an output  $y$  is linearly related to a single input  $x$ . The SLREG model is given by  $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ , where  $\epsilon_i$ 's are assumed  $NID(0, \sigma_\epsilon^2)$ .

Regression	Versus	Correlation
Levels of $x$ are fixed		$(x, y)$ is a bivariate random vector

As an example, the tensile strength (TS) of paper is thought to be linearly related to the amount of hardwood concentration in the pulp. In a pilot plant, 10 sample pairs  $(x_i, y_i)$ ,  $i = 1, 2, \dots, 10$ , were produced leading to the following data:

$x_i$ : 10    15    15    20    20    20    25    25    28    30%

$y_i$ : 160   171   175   182   184   181   188   193   195   200 psi

where  $x$  = Hardwood Concentration is fixed, and  $y$  = Paper TS is a rv. Thus,

$$V(y_i) = V(\beta_0 + \beta_1 x_i + \epsilon_i) = V(\mu_i + \epsilon_i) \quad \sigma_\epsilon^2 = \sigma^2,$$

and as a result  $y_i$ 's are assumed  $NID(\mu_i = E(y_i | x_i) = \beta_0 + \beta_1 x_i, \sigma_\epsilon^2)$ . The objective is to estimate the parameters  $\beta_0$  and  $\beta_1$  such that the least squares function (LSF):

$$L(\beta_0, \beta_1) = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \quad (52)$$

is minimized wrt (with respect to)  $\beta_0$  and  $\beta_1$ . For example, at  $(x_1 = 10, y = y_1)$ , the expression for the error before experimentation is given by  $\epsilon_1 = y_1 - \beta_0 - 10\beta_1$ , while at  $(x_2, y_2)$ ,  $\epsilon_2 = y_2 - \beta_0 - 15\beta_1$ , ..., and at  $(x_{10}, y_{10})$  the prior error is given by  $\epsilon_{10} = y_{10} - \beta_0 - 30\beta_1$ . However, after data have been gathered, as shown above, the LSF in Eq. (52) reduces to:

$$L(\beta_0, \beta_1) = (160 - \beta_0 - 10\beta_1)^2 + (171 - \beta_0 - 15\beta_1)^2 + \dots + (200 - \beta_0 - 30\beta_1)^2.$$

In order to minimize the LSF in Eq. (52) wrt the unknown parameters  $\beta_0$  and  $\beta_1$ , we partially differentiate the LSF in Eq. (52) and equate its partial derivatives to zero in order to obtain the optimum point.

$$\frac{\partial L}{\partial \beta_0} = 2(160 - \beta_0 - 10\beta_1)(-1) + 2(171 - \beta_0 - 15\beta_1)(-1) + \dots + 2(200 - \beta_0 - 30\beta_1)(-1) \xrightarrow{\text{Set to}} = 0$$

$$\frac{\partial L}{\partial \beta_1} = 2(160 - \beta_0 - 10\beta_1)(-10) + 2(171 - \beta_0 - 15\beta_1)(-15) + \dots + 2(200 - \beta_0 - 30\beta_1)(-30) \xrightarrow{\text{Set to}} = 0$$

The above 2 equations simplify to

$$(160 - \hat{\beta}_0 - 10\hat{\beta}_1) + (171 - \hat{\beta}_0 - 15\hat{\beta}_1) + \dots + (200 - \hat{\beta}_0 - 30\hat{\beta}_1) = 0, \text{ and}$$

$$(160 - \hat{\beta}_0 - 10\hat{\beta}_1)(10) + (171 - \hat{\beta}_0 - 15\hat{\beta}_1)(15) + \dots + (200 - \hat{\beta}_0 - 30\hat{\beta}_1)(30) = 0.$$

After some algebraic simplification, the above two equations reduce to

$$\sum_{i=1}^n y_i - \hat{\beta}_0 \sum_{i=1}^{10} 1 - \hat{\beta}_1 \sum_{i=1}^{10} x_i = 0$$

$$\sum_{i=1}^n x_i y_i - \hat{\beta}_0 \sum_{i=1}^{10} x_i - \hat{\beta}_1 \sum_{i=1}^{10} x_i^2 = 0.$$

The above heterogeneous system of 2 equations with 2 unknowns when written as

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{10} x_i = \sum_{i=1}^n y_i \quad \rightarrow \quad 10\hat{\beta}_0 + 208\hat{\beta}_1 = 1829, \text{ and}$$

$$\hat{\beta}_0 \sum_{i=1}^{10} x_i + \hat{\beta}_1 \sum_{i=1}^{10} x_i^2 = \sum_{i=1}^n x_i y_i \quad \rightarrow \quad 208\hat{\beta}_0 + 4684\hat{\beta}_1 = 38715,$$

is called the LS (Least-Squares) normal equations. Note that in matrix form, the above system of equations can be written as

$$\begin{bmatrix} 10 & 208 \\ 208 & 4684 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} 1829 \\ 38715 \end{bmatrix}, \quad \text{or} \rightarrow \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

The 1<sup>st</sup> normal equation gives rise to  $\hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} \rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \rightarrow$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i \rightarrow \hat{y}_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x}). \quad (53a)$$

Note that  $\hat{y}_i$  is called the fitted value of the LS model, and if the slope  $\hat{\beta}_1$  is zero, then

the best fit for each  $y_i$  is  $\bar{y}$ . Substituting  $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$  into

the 2<sup>nd</sup> normal equation ( $\hat{\beta}_0 \sum_{i=1}^{10} x_i + \hat{\beta}_1 \sum_{i=1}^{10} x_i^2 = \sum_{i=1}^n x_i y_i$ ) yields

$$(\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^{10} x_i + \hat{\beta}_1 \sum_{i=1}^{10} x_i^2 = \sum_{i=1}^n x_i y_i \quad \longrightarrow$$

$$(-\hat{\beta}_1 \bar{x}) \sum_{i=1}^{10} x_i + \hat{\beta}_1 \sum_{i=1}^{10} x_i^2 = \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^{10} x_i \quad \longrightarrow$$

$$\hat{\beta}_1 \left( \sum_{i=1}^{10} x_i^2 - \bar{x} \sum_{i=1}^{10} x_i \right) = \sum_{i=1}^n x_i (y_i - \bar{y}) \quad \longrightarrow$$

$$\hat{\beta}_1 \sum_{i=1}^{10} x_i (x_i - \bar{x}) = \sum_{i=1}^n x_i (y_i - \bar{y}) - \bar{x} \sum_{i=1}^n (y_i - \bar{y}) \quad \longrightarrow$$

$$\hat{\beta}_1 \sum_{i=1}^{10} (x_i - \bar{x})^2 = \sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})] = S_{xy} \rightarrow$$

$$\hat{\beta}_1 = S_{xy} / S_{xx}, \text{ where } S_{xx} = \sum_{i=1}^{10} (x_i - \bar{x})^2 = \sum_{i=1}^{10} x_i^2 - \left( \sum_{i=1}^{10} x_i \right)^2 / n,$$

$$S_{xx} = 4684 - (208)^2 / 10 = 357.6, \quad \bar{x} = 20.80, \text{ and } \bar{y} = 182.90,$$

$$S_{xy} = \sum_{i=1}^n x_i y_i - \left( \sum_{i=1}^{10} x_i \right) \left( \sum_{i=1}^n y_i \right) / 10 = 38715 - (208)(1829) / 10 \rightarrow$$

$$S_{xy} = 671.8 \quad \rightarrow \quad \hat{\beta}_1 = 671.8 / 357.6 = 1.87864 \rightarrow$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 182.9 - 1.87864 \times 20.8 = 143.824385. \text{ Hence, the fitted regression}$$

$$\text{model is: } \hat{y} = 143.8244 + 1.87864x. \quad (53b)$$

Note that  $\hat{y}$  is an estimate of the (curve) of regression of  $y$  on  $x$ , which is

given by  $E(y | x) = \mu_{y|x} = \beta_0 + \beta_1 x$ . For example, when  $x = 20$ , then  $\hat{E}(y | 20) =$

$$\hat{\mu}_{y|x=20} = \hat{y}_4 = 143.8244 + 1.87864 \times 20 = 181.3971 \rightarrow e_4 = y_4 - \hat{y}_4 = 182 - 181.3971 =$$

0.60291 (the 4<sup>th</sup> residual). Similarly,  $e_1 = -2.6107, \dots, e_{10} = y_{10} - \hat{y}_{10} = -0.1834$  (the

10<sup>th</sup> residual error). Further, it is necessary that  $\sum_{i=1}^n e_i \equiv 0$  for all statistical models.

For example, in the case of SLREG  $\sum_{i=1}^n e_i =$  The sum of all  $n$  residuals =

$$\sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n [(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})] \equiv 0 \text{ because } \hat{y}_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x}),$$

$$\sum_{i=1}^n (y_i - \bar{y}) \equiv 0 \text{ and } \sum_{i=1}^n (x_i - \bar{x}) \equiv 0.$$

A test of Significance in Regression is 1<sup>st</sup> conducted through an ANOVA Table.

To this end, we compute the SS's for the ANOVA.

$$USS = \sum_{i=1}^n y_i^2 = 335825 \text{ (with } n = 10 \text{ df)}, \text{ CF} = \left( \sum_{i=1}^n y_i \right)^2 / n = 1829^2 / 10 = 334524.10 \text{ (with}$$

$$1 \text{ df}) \rightarrow SS_T = SS_{\text{Total}} = USS - CF \text{ (with } 9 \text{ df)} = \sum_{i=1}^n (y_i - \bar{y})^2 = USS - CF = 1300.90.$$

$$SS(\text{Residuals}) = \sum_{i=1}^n e_i^2 = 38.83277405 \text{ (with } n - 2 = 8 \text{ df because there are 2}$$

constraints; try to determine what the 2 constraints are for bonus points).

$$\text{Now, } SS(\text{Residuals}) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n [(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})]^2 =$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 - 2\hat{\beta}_1 \sum_{i=1}^n [(y_i - \bar{y}) (x_i - \bar{x})] + \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 =$$

$$S_{yy} - 2\hat{\beta}_1 S_{xy} + \hat{\beta}_1^2 S_{xx} = S_{yy} - 2\hat{\beta}_1 S_{xy} + \hat{\beta}_1 (S_{xy} / S_{xx}) S_{xx} \rightarrow$$

$$SS_{RES} = SS(\text{Total}) - \hat{\beta}_1 S_{xy} = SS_T - SS_{\text{Model}}.$$

**Exercise 93.** Show that  $SS_T = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 =$

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = SS_{RES} + SS_{REG} \text{ (where REG = Regression),}$$

and as a result  $SS_{REG} = SS(\text{Model}) = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1 S_{xy}.$

For our example,  $SS_{REG} = SS_{\text{Model}} = \hat{\beta}_1 S_{xy} = 1.87864 \times 671.8 = 1262.0672$ . Note that  $SS_{REG} + SS_{RES} = 1262.0672 + 38.8328 = 1300.90 = SS_{\text{Total}} = SS_T$ . The ANOVA Table is given below.

Source	df	SS	MS	F <sub>0</sub>
Total	9	1300.90		
Model or Regression	1	1262.06723	1262.06723	260.0004
Residuals	8	38.83277	4.8541	F <sub>0.01,1,8</sub> = 11.25862

Since the *P-value* of the F-test is almost zero, then strongly reject  $H_0: \beta_1 = 0$ . Thus, the regressor variable  $x$  accounts for 97.015% (=  $SS_{REG}/SS_T$ ) of variability in the response  $y$ , i.e.,  $R_{\text{Model}}^2 = R_{\text{REG}}^2 = 1262.06723/1300.90 = 97.015\%$ .

To determine how well the model  $\hat{y} = 143.8244 + 1.87864x$  fits the 10 ordered pairs  $(x_i, y_i)$ , we proceed as follows:  
 $SS(\text{Pure Error}) = (171^2 + 175^2 - 346^2/2) + (182^2 + 184^2 + 181^2 - 547^2/3) + (188^2 + 193^2 - 381^2/2) = 25.16667$  (with  $1 + 2 + 1 = 4$  df).  $\rightarrow SS(\text{LOF} = \text{Lack of Fit}) = SS_{RES} - SS(\text{PE} = \text{Pure Error}) = 38.832774 - 25.16667 = 13.66611$  (with  $8 - 4 = 4$  df). The augmented ANOVA Table is given atop the next page. Since the *P-value* for LOF is  $\hat{\alpha} = P(F_{4,4} \geq 0.54302) = 0.715623$ , the regression model  $\hat{y} = 143.8244 + 1.87864x$  fits the 10 points extremely well.

The Complete ANOVA Table for the Example on page 156

Source	df	SS	MS	F <sub>0</sub>
Total	9	1300.90		
Model	1	1262.06723	1262.06723	260.0004
Residuals	8	38.83277	4.8541	F <sub>0.01,1,8</sub> = 11.26
Pure Error (PE)	4	25.16667	6.29167	
Lack of Fit (LOF)	4	13.66611	3.41653	0.54302

**Exercise 94.** Let  $Y_1 = \sum_{i=1}^n a_i U_i$  and  $Y_2 = \sum_{j=1}^m b_j W_j$  be two linear combinations

with  $E(U_i) = \mu_{1i}$ ,  $E(W_j) = \mu_{2j}$ , and  $\text{COV}(U_i, W_j) = \sigma_{ij}$ . Show that the  $\text{COV}(Y_1, Y_2) =$

$$\sum_{i=1}^n \sum_{j=1}^m a_i b_j \sigma_{ij}.$$

As an application of the above Exercise, suppose  $U_1, U_2, W_1, W_2,$  and  $W_3$  are variates (or rvs) with  $\text{COV}(U_i, W_j) = (-1/2)^{i-j}$ . Our objective is to compute the  $\text{COV}(-U_1 + 3U_2, 2W_1 - W_2 - 2W_3) = \text{COV}(Y_1, Y_2) = (-1)(2) (-1/2)^0 + (-1) (-1)(-2) + (-1)(-2)(-1/2)^{-2} + (3)(2) (-1/2) + (3)(-1)(1) + (3)(-2)(-1/2)^{-1} \rightarrow$   
 $\text{COV}(Y_1, Y_2) = -2 - 2 + 8 - 3 - 3 + 12 = 10.$

**Exercise 95.** Show the following properties of SLREG estimators: (you must assume that  $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ ). (1) Both  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are unbiased. (2)  $V(\hat{\beta}_1) = \sigma_\epsilon^2 / S_{xx}$ ,  $\text{COV}(\bar{y}, \hat{\beta}_1) = 0$ ,  $V(\hat{\beta}_0) = [\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}] \sigma_\epsilon^2$ ,  $\text{COV}(\hat{\beta}_0, \hat{\beta}_1) = -\bar{x} \sigma_\epsilon^2 / S_{xx}$ ,  $E(\epsilon_i) = 0$ ,  
 $V(\epsilon_i) = [\frac{n-1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}] \sigma_\epsilon^2$ , and  $V(\hat{y}_0) = V(\hat{\beta}_0 + \hat{\beta}_1 x_0) = [\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}] \sigma_\epsilon^2$ . (3)  
 $E(SS_{\text{RES}}) = (n-2) \sigma_\epsilon^2$ , which shows that  $MS_{\text{RES}}$  is an unbiased estimator of  $\sigma_\epsilon^2$  because  
 $MS_{\text{RES}} = SS_{\text{RES}} / (n-2).$

**Exercise 96.** The strength of paper,  $y$ , used in the manufacture of cardboard boxes is related to % of hardwood in the original pulp ( $x$ ). Under controlled conditions, a pilot plant manufactures 14 samples, each from a different batch of pulp and measures the TS. The resulting data is provided atop the next page. (a) Fit a SLR model to the data and check your answer via Minitab. (b) Provide the ANOVA table and conduct all tests of significance. (c) After studying through pp. 163-164 of these notes, obtain the 95% CIs for  $\beta_0$ ,  $\beta_1$  and  $E(Y | x = 2.5)$ .

$x_i$	1.0	1.5	1.5	1.5	2.0	2.0	2.2
$y_i$	101.4	117.4	117.1	106.2	131.9	146.9	146.8
$x_i$	2.4	2.8	2.8	3.0	3.0	3.2	3.3
$y_i$	133.9	135.1	145.2	134.3	144.5	143.7	146.9

**Exercise 97.** Show that in SLREG with  $n_i$  observations at  $m$  distinct levels  $SS_{RES}$

$$\begin{aligned}
 &= \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2 = \sum_{i=1}^m \sum_{j=1}^{n_i} [(y_{ij} - \bar{y}_i) + (\bar{y}_i - \hat{y}_i)]^2 \\
 &= \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^m n_i (\bar{y}_i - \hat{y}_i)^2 = SS(PE) + SS(LOF).
 \end{aligned}$$

Note that the terminology regression is due to Francis Galton who first observed (in the late 19<sup>th</sup> century) that the height of sons tended to be closer to the mean height of species than their fathers, i.e., the height of off-springs tended to regress back toward the mean height of the entire species, no matter what the height of fathers were. This is called regressing toward the mean.

### Statistical Inference in SLREG

Recall that if a rv is  $N(\mu, \text{unknown } \sigma^2)$ , then the statistic

$[rv - E(rv)]/se(rv) \sim T_v$ , where  $v$  is the  $df$  of the  $se(rv)$ .

Therefore, to test  $H_0: \beta_1 = 0$  VS  $H_1: \beta_1 \neq 0$ , we may use the fact that

$$\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MS_{RES}/S_{xx}}}$$

has a Student's t-distribution with  $v = n - 2$   $df$ . We now apply the above t-test to the regression model of the Example on page 156 of these notes. We 1<sup>st</sup> compute the  $se(\hat{\beta}_1) = \sqrt{MS_{RES}/S_{xx}} = \sqrt{4.8541/357.6} = 0.11651$  (see Exercise 95). Since under  $H_0$  the slope,  $\beta_1$ , is hypothesized to be zero, then our null test statistic becomes

$$t_0 = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MS_{RES}/S_{xx}}} = \frac{1.87864 - 0}{0.11651} = 16.12455,$$

which far exceeds  $t_{0.025,8} = 2.3060$ , and hence the null hypothesis of zero slope must be strongly rejected. Note that  $(t_0)^2 = F_0 = 260.0004$  of the ANOVA Table on page 161 of these notes. The discrepancy in the 4<sup>th</sup> decimal is strictly due to rounding.

The next SI (Statistical Inference) is to obtain either a 95% or 99% CI for  $\beta_1$ .

Figure 27 shows the sampling distribution (SMD) of  $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{MS_{RES}/S_{xx}}}$  for a confidence

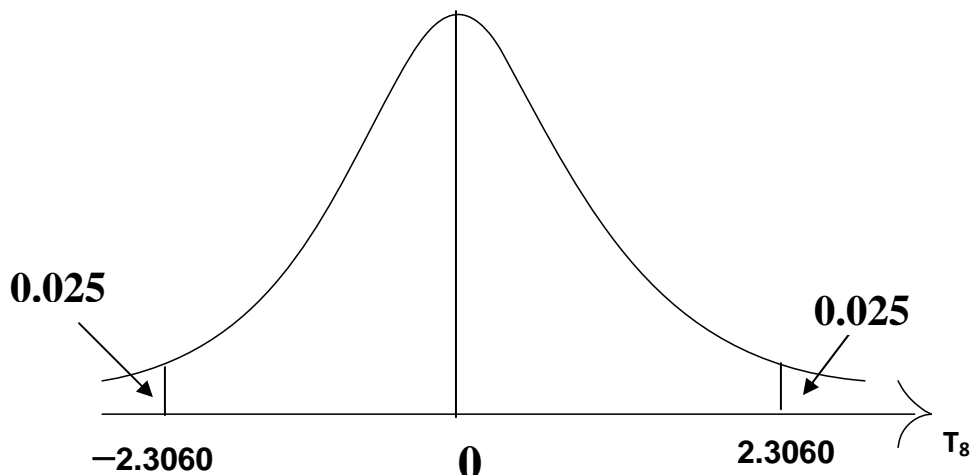
level of  $1 - \alpha = 0.95$ . Figure 27 clearly shows that the  $\Pr(-2.3060 \leq T_8 \leq 2.3060) = \Pr(-2.3060 \leq (\hat{\beta}_1 - \beta_1)/se(\hat{\beta}_1) \leq 2.3060) = 0.95$ , (note that  $2.306 = t_{0.025,8}$ ), or

$$\Pr[-2.3060 se(\hat{\beta}_1) - \hat{\beta}_1 \leq -\beta_1 \leq -\hat{\beta}_1 + 2.3060 se(\hat{\beta}_1)] = 0.95.$$

In order to solve for  $\beta_1$ , we need to multiply the above 2 inequalities by  $-1$ , which results in

$$\Pr[\hat{\beta}_1 - 2.3060 se(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + 2.3060 se(\hat{\beta}_1)] = 0.95.$$

The above equation clearly shows that the 95% confidence limits are  $L(\beta_1) = \hat{\beta}_1 - 2.3060 se(\hat{\beta}_1)$  and  $U(\beta_1) = \hat{\beta}_1 + 2.3060 se(\hat{\beta}_1)$ ; further, the 95% HCIL



**Figure 27.** The SMD of  $(\hat{\beta}_1 - \beta_1)/se(\hat{\beta}_1)$

(Half-CI-Length) is given by  $t_{0.025,8} \times se(\hat{\beta}_1) = 2.3060 \times 0.11651 = 0.2687$  and hence the requisite CI is  $\hat{\beta}_1 \pm 0.2687 = (1.6100, 2.1473)$ . Since this 95% CI:  $1.6100 \leq \beta_1 \leq 2.1473$ , clearly excludes the hypothesized value of zero, it is consistent with the rejection of the above t-test on  $\beta_1 = 0$  at the 5% level.

Applying the same procedure as the above to  $\hat{\beta}_0$ , and using  $V(\hat{\beta}_0) = \left[ \frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}} \right] \sigma_\epsilon^2$ , we obtain the following 95% CI for the parameter  $\beta_0$ .

$$138.00973 \leq \beta_0 \leq 149.63904$$

Since  $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$  is a point unbiased estimate of  $E(Y | x_0)$ , then using one of the results of Exercise 95, the 95% CI for  $\mu_0 = \beta_0 + \beta_1 x_0$  is given by

$$\hat{y}_0 \pm 2.3060 \times se(\hat{y}_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 \pm 2.3060 \times \sqrt{\left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right] \times MS_{RES}} .$$

As an example, if we wish to obtain a 95% CI for  $E(Y|x = 22) = \beta_0 + 22\beta_1$ , the above formula yields  $185.1544 \pm 2.3060 \times 0.7106 = (183.5158, 186.7930)$ , or  $183.5158 \leq \beta_0 + 22\beta_1 \leq 186.7930$ . Note that this last 95% CI does not have a 0.95 Pr of containing the mean of Y at  $x = 22$ ; that Pr is either 0 or 1. The

length of this CI is  $2 \times 2.3060 \times se(\hat{\beta}_0 + \hat{\beta}_1 x_0) = 3.2773$ .

### PREDICTION INTERVAL FOR a FUTURE OBSERVATION at a SPECIFIED $X_0$

As an example, reconsider the regression model of Eq. (53b) on p. 159 of my notes. Suppose we are to make  $N$  (the most common value of  $N$  is one observation in the future) observations at  $x_0 = 22\%$  hardwood concentration in the future. Let  $y_0$  be an actual future observation (not yet observed and hence is a rv). Clearly, the best single point forecast from our model is  $\hat{y}_0 = 143.8244 + 1.87864 \times 22 = 185.1545$ . How do we use our rvs  $y_0$  and  $\hat{y}_0$  to obtain a prediction interval for  $y_0$ ?

To accomplish this task, we must 1<sup>st</sup> define the forecast error rv as  $\Psi = y_0 - \hat{y}_0$ . Because  $E(\Psi) = E(\beta_0 + \beta_1 x_0 + \varepsilon_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0) = 0$ , and the variance of forecast error at  $N = 1$  is given by

$$V(\Psi) = V[y_0 - \bar{y} - \hat{\beta}_1(x_0 - \bar{x})] = V(y_0) + V(\bar{y}) + (x_0 - \bar{x})^2 V(\hat{\beta}_1)$$

$$V(\Psi) = \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right] \sigma_\varepsilon^2. \text{ Since } \Psi = y_0 - \hat{y}_0 \text{ is a LC (Linear Combination) of}$$

NID rvs, then the forecast error  $\Psi$  is also  $N(0, V(\Psi))$ , and as a result the rv

$$\frac{\Psi - E(\Psi)}{se(\Psi)} = \frac{y_0 - \hat{y}_0}{\sqrt{\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right] \times MS_{RES}}}$$

has a student's t-distribution with  $(n - 2)$  *df*. Therefore, with  $N = 1$  future observation at  $x_0 = 22$  we obtain the following 95% prediction Pr statement:

$$\Pr\left[-2.3060 \leq \frac{y_0 - \hat{y}_0}{\sqrt{\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right] MS_{RES}}} \leq 2.3060\right] = 0.95$$

Rearranging the inequality inside the above brackets yields the 95% PI (prediction interval) for  $y_0$ .

$$\Pr[\hat{\beta}_0 + \hat{\beta}_1 x_0 - 2.3060se(\psi) \leq y_0 \leq \hat{\beta}_0 + \hat{\beta}_1 x_0 + 2.3060se(\psi)] = 0.95.$$

To obtain the actual 95% PI, we 1<sup>st</sup> compute the  $se(\psi)$ .

$$se(\psi) = se(y_0 - \hat{y}_0) = \sqrt{\left[1 + \frac{1}{10} + \frac{(22 - 20.8)^2}{357.6}\right] 4.8541} = 2.3150$$

Next, we insert  $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 185.1545$  and the  $se(\psi)$  into the above Pr statement.

$$\Pr[-2.3060 \leq \frac{y_0 - 185.1545}{2.3150} \leq 2.3060] = 0.95$$

Finally, rearranging the inequality inside the above brackets results in  $\Pr[185.1545$

$$-2.3060 \times 2.3150 \leq y_0 \leq 185.1545 + 5.3383] = 0.95 \rightarrow$$

$$\Pr(179.8160 \leq y_0 \leq 190.4927) = 0.95.$$

The length of the above prediction band is 10.6766. Note that, unlike a CI, the above PI actually has a Pr of 0.95 to contain a future observation because  $y_0$  is still a random variable, as it has not yet been observed. Further, the length of the 95% PI is always larger than the corresponding 95% CI because it contains 2 sources of error unlike a CI.

## CORRELATION

Regression is generally applicable if the regressor (or independent) variable on the RHS of the model can be controlled by the experimenter so that  $V(X) = 0$ . In studies where both variables X and Y have to be measured from the same sampling unit, i.e., the  $(x, y)$  pair form a bivariate random vector, then it is best to conduct a correlation analysis than regressing y on x. To this end, let  $[X \quad Y]^T = [\mathbf{X} \quad \mathbf{Y}]'$  be a  $2 \times 1$  bivariate vector; then the population correlation coefficient between X and Y is defined as

$$\rho_{xy} = \frac{\text{COV}(X, Y)}{\sigma_X \times \sigma_Y} = \frac{\sigma_{xy}}{\sigma_X \sigma_Y} = \rho,$$

where  $\sigma_{xy} = E[(X - \mu_x)(Y - \mu_y)] = E(XY) - E(X) \times E(Y)$  is the covariance between the

random variables X and Y. It can be proven that  $-1 \leq \rho \leq 1$ , or  $|\rho| \leq 1$ , which you will be asked to do in the following Bonus Exercise.

**Exercise 98 (5 Bonus Points).** Prove that  $|\rho| \leq 1$  by expanding the  $V(aX + bY)$ , where a and b are arbitrary real constants of your choice.

When  $\rho = \pm 1$ , the two random variables X and Y are said to be perfectly correlated. If X and Y are independent rvs, then  $\rho = 0$ , but  $\rho = 0$  does not always imply that X and Y are independent. However, if X and Y have the joint bivariate normal density function, then  $\rho = 0$  does imply that X and Y are independent.

In practice, the value of the population correlation coefficient,  $\rho$ , is unknown and has to be estimated from a random sample of size n pairs. The sample point estimate of  $\rho$  is given by

$$\hat{\rho} = r = \frac{\hat{\sigma}_{xy}}{S_x S_y} = \frac{S_{xy}/(n-1)}{S_x S_y} \quad (54a)$$

where the numerator of r is the sample covariance  $\hat{\sigma}_{xy} = S_{xy}/(n-1) =$

$\frac{1}{n-1} \sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})]$ ; equation (54a) can also be written as

$$r = \hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{\sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{[\sum_{i=1}^n (x_i - \bar{x})^2] \times [\sum_{i=1}^n (y_i - \bar{y})^2]}} \quad (54b)$$

**Example 45.** The following data give the final averages of 15 randomly selected INSY students in Engineering Statistics (X) and Operations Research (Y = OR).

X: 86% 75 63 64 92 58 78 90 85 77 69 82 84 94 76%

Y: 77% 85 70 57 83 69 76 82 95 87 62 86 83 85 88%

The sample statistics are:  $\sum_{i=1}^{15} x_i = 1173$ ,  $\bar{x} = 78.20$ ,  $\sum_{i=1}^{15} y_i = 1185$ ,  $\bar{y} =$

$$79.00, \sum_{i=1}^n x_i^2 = 93405, \sum_{i=1}^n y_i^2 = 95145, \sum_{i=1}^n x_i y_i = 93755, S_{xx} = 1676.40,$$

$$S_{yy} = 1530, S_{xy} = 1088, S_x = 10.9427, S_y = 10.4540, \hat{\sigma}_{xy} = 77.7143, r = \hat{\sigma}_{xy} / (S_x S_y) =$$

0.67935. Note that if  $y$  is regressed on  $x$  and the resulting model  $R^2$  is computed, then

$$r = \sqrt{R_{\text{Model}}^2}.$$

## TEST OF HYPOTHESIS ABOUT $\rho$

There are two different tests that can be conducted on the population parameter  $\rho$ : (1)  $H_0: \rho = 0$ , (2)  $H_0: \rho = \rho_0$ , where  $\rho_0 \neq 0$ .

(1) Testing  $H_0: \rho = 0$  versus one of the 3 alternatives  $H_1: \rho \neq 0$ , or

$H_1: \rho < 0$ , or  $H_1: \rho > 0$ .

Recall that statistical inference on a parameter is conducted using the sampling distribution of the point estimator. The point estimator of  $\rho$  is the sample correlation coefficient  $r$ . It can be shown that the null sampling distribution (SMD) of the statistic

$$t_0 = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}} \quad (55)$$

follows a Student's  $t$ -distribution with  $(n-2)$  *df*. For the example 45 above, the most appropriate alternative is  $H_1: \rho > 0$ . Therefore, the critical region for testing  $H_0: \rho = 0$  at the LOS  $\alpha = 0.05$  is  $(1.771, \infty)$ . The value of our test statistic is  $t_0 = (0.67935 \sqrt{13}) / \sqrt{1-0.46152} = 3.3379$ , which easily exceeds  $t_{0.05,13} = 1.771$ , leading to the rejection of zero correlation between  $X$  and  $Y$ . The *P-value* (or the *Pr* level) of the test is  $\hat{\alpha} = P(T_{13} \geq 3.33794) = 0.002672$ , which is less than 0.05 as expected because  $H_0$  was rejected at the 5% level of significance. If the assumption of joint bivariate normal density for the  $2 \times 1$  vector  $[X \quad Y]^T$  is indeed tenable, then the rejection of  $H_0: \rho = 0$  implies that  $X$  and  $Y$  are not independent; otherwise, the rejection of  $H_0$  implies that  $X$  and  $Y$  are linearly related.

**Exercise 99.** Use results from regression to show that the statistic  $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

has a Student's  $t$  sampling distribution. Hint: First refer to an

ANOVA table for the regression of  $Y$  on a single regressor  $X$  and then use

the fact that  $F_{1,v_2} = t_{v_2}^2$ .

(2) Testing  $H_0: \rho = 0.50$  versus the alternative  $H_1: \rho \neq 0.50$  at the LOS  $\alpha = 0.05$ . The  $t_0$  given in equation (55) cannot be used to test  $H_0: \rho = 0.50$  because the expression for  $t_0$  is free of  $\rho$ . It has been shown, however, in statistical literature that the sampling distribution (SMD) of the statistic

$$Z = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right) = \operatorname{arctanh}(r) = \tanh^{-1}(r)$$

is approximately normal with  $E(Z|\rho) = \frac{1}{2} \ln\left(\frac{1+\rho}{1-\rho}\right) = \tanh^{-1}(\rho)$  and  $V(Z) \cong 1/(n-3)$ .

For the Example 45 above, the SMD of  $Z$  is depicted in Figure 28, where under the null

hypothesis  $E(Z|\rho=0.50) = \frac{1}{2} \ln\left(\frac{1+0.50}{1-0.50}\right) = 0.54931 = \mu_Z$  and  $V(Z) \doteq 1/(15-3) =$

$0.08333\bar{3}$ . The acceptance interval for testing the 2-sided  $H_0: \rho = 0.50$  is given by  $(A_L,$

$A_U) = (-0.0165, 1.1151)$ . The value of the test statistic is  $Z = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right) =$

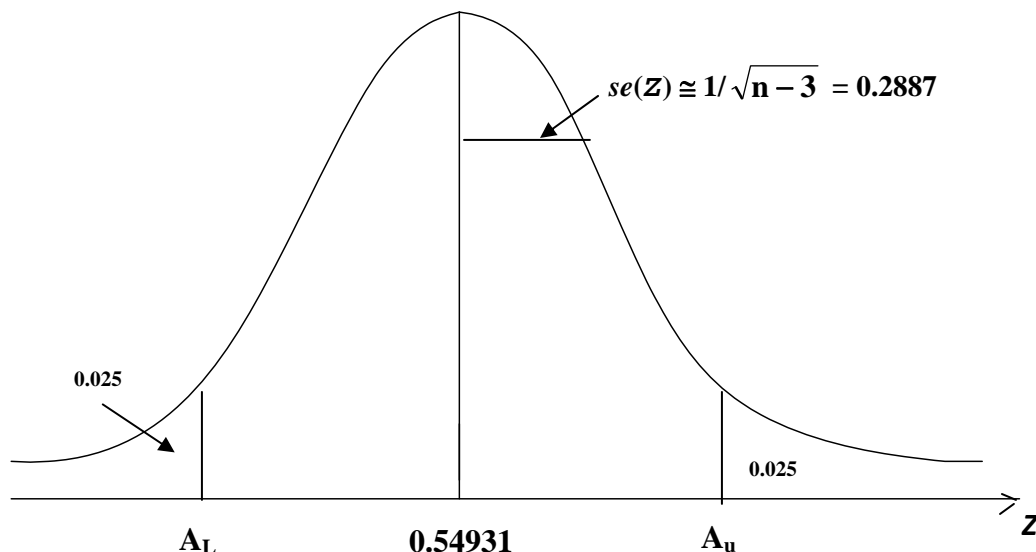
$\frac{1}{2} \ln\left(\frac{1+0.67935}{1-0.67935}\right) = \operatorname{arctanh}(0.67935) = 0.82791$ , which is well inside the AI =

$(-0.0165, 1.1151)$  and hence we cannot reject  $H_0: \rho = 0.50$ . The  $P$ -value of the test is

given by  $\hat{\alpha} = 2 \times \Pr(Z \geq 0.82791) = 2 \times \Pr[Z_0 \sim N(0, 1) \geq 0.96511] = 0.33449$ . Therefore,

we cannot deduce that the data provide sufficient evidence for  $\rho \neq 0.50$  but they do

provide sufficient evidence that  $\rho > 0$ . Note that Kendall and Stuart (Vol. 1, 2<sup>nd</sup>



**Figure 28. The Approximate SMD of  $Z$**

edition, p. 391) give a better approximation for  $V[\frac{1}{2} \ln(\frac{1+r}{1-r})]$  as  $\frac{1}{n-1} + \frac{4-\rho^2}{2(n-1)^2}$ .

For small  $\rho$  and  $n > 10$ , this last approximation is almost equal to  $1/(n-3)$ .

#### OBTAINING a 95% CI FOR $\rho$

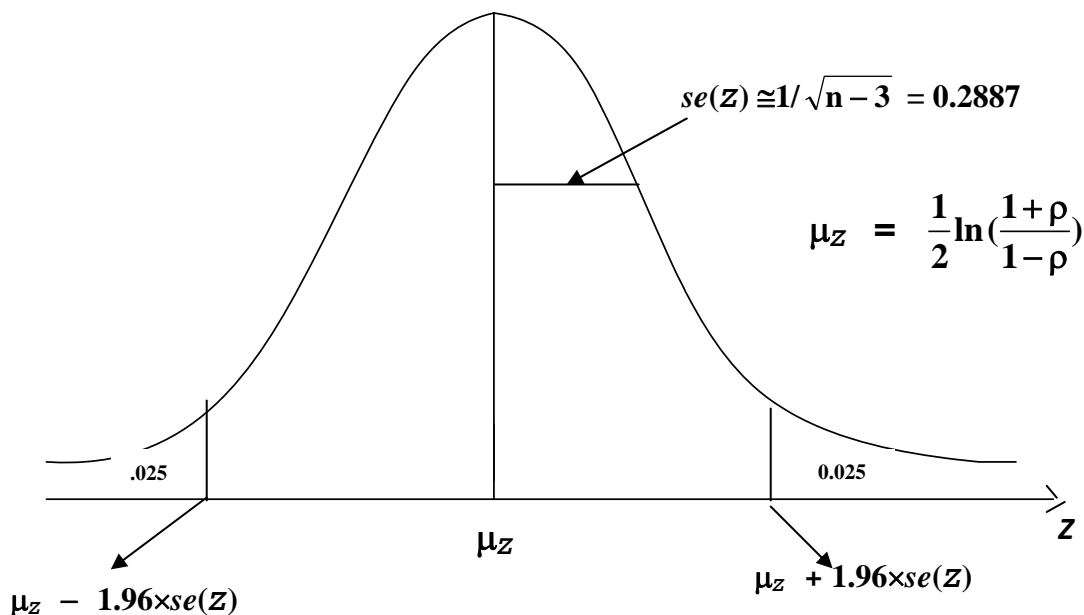
Again, we have to make use of the fact that  $\frac{1}{2} \ln(\frac{1+r}{1-r}) = \operatorname{arctanh}(r) =$

$\tanh^{-1}(r) \sim N[\frac{1}{2} \ln(\frac{1+\rho}{1-\rho}), 1/(n-3)] = N[\tanh^{-1}(\rho), 1/(n-3)]$ , as depicted in Figure

29, which clearly shows that

$$\Pr[\mu_z - 1.96 \times se(Z) \leq \frac{1}{2} \ln(\frac{1+r}{1-r}) \leq \mu_z + 1.96 \times se(Z)] = 0.95$$

$$\Pr[-\frac{1}{2} \ln(\frac{1+r}{1-r}) - 1.96 \times se(Z) \leq -\mu_z \leq -\frac{1}{2} \ln(\frac{1+r}{1-r}) + 1.96 \times se(Z)] = 0.95$$



**Figure 29. The Approximate SMD of**

$$\Pr[-0.82791 - 1.96 \times se(Z) \leq -\mu_Z \leq -0.82791 + 1.96 \times se(Z)] = 0 \text{ or } 1.$$

$$\Pr[-0.82791 - 0.56580 \leq -\mu_Z \leq -0.82791 + 0.56580] = 0 \text{ or } 1.$$

$$\Pr[0.82791 - 0.56580 \leq \mu_Z \leq 0.82791 + 0.56580] = 0 \text{ or } 1.$$

$$0.26210 \leq \frac{1}{2} \ln\left(\frac{1+\rho}{1-\rho}\right) \leq 1.39371 \quad \rightarrow \quad (56)$$

$$0.52420 \leq \ln\left(\frac{1+\rho}{1-\rho}\right) \leq 2.78742 \quad \rightarrow \quad 1.68912 \leq \frac{1+\rho}{1-\rho} \leq 16.23904 \quad \longrightarrow$$

$1.68912(1-\rho) \leq 1 + \rho \leq 16.23904(1-\rho)$ . These last two inequalities when solved separately from the left yield  $\rho_L = 0.25626$  and from the right yield  $\rho_U = 0.88398$ , and hence the 95% CI for population correlation coefficient is given by  $0.256263 \leq \rho \leq 0.883985$ .

The above CI interval could also have been obtained by recognizing

that the inverse function of  $\frac{1}{2} \ln\left(\frac{1+\rho}{1-\rho}\right)$  is  $\tanh(\rho)$  and taking the  $\tanh$  of inequality

(56), i.e.,  $\tanh(0.26210) \leq \rho \leq \tanh(1.39371) \rightarrow 0.25626 \leq \rho \leq 0.88398$ , which agrees

with the above CI for  $\rho$ . As expected, the above CI does include the value of  $\rho = 0.50$  because the null hypothesis  $H_0 : \rho = 0.50$  could not be rejected at the 5% level.

**Exercise 100.** Repeat the above correlation analysis for the data of Exercise 62 on page 493 of Devore. Check your answer for the 95% CI on  $\rho$  by using the formulas

$$\rho_L = \tanh[\operatorname{arctanh}(r) - Z_{\alpha/2}/\sqrt{n-3}] = \tanh\left[\frac{1}{2} \ln \frac{1+r}{1-r} - Z_{\alpha/2}/\sqrt{n-3}\right]$$

and  $\rho_U = \tanh[\operatorname{arctanh}(r) + Z_{\alpha/2}/\sqrt{n-3}]$ , where  $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$  and

$$\tanh^{-1}(x) = \operatorname{arctanh}(x) = \frac{1}{2} \ln \frac{1+x}{1-x}. \quad (\text{b}) \text{ For the data of Exercise 62 p. 493 of Devore,}$$

obtain the 95% lower one-sided CI for  $\rho$ . The Microsoft Excel and Matlab functions for  $\operatorname{arctanh} = \tanh^{-1}$  and hyperbolic tangent are ATANH and TANH, respectively.

**Exercise 101 (10 Bonus Points).** (a) Show that the inverse function of  $f(x) = \frac{1}{2} \ln\left(\frac{1+x}{1-x}\right)$ ,  $|x| < 1$ , is  $f^{-1}(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ , where  $-1 \leq \tanh(x) \leq 1$ . (b) Use

your results in part (a) to derive the following 95% CI for  $\rho$ :  $\tanh\left[\frac{1}{2} \ln\left(\frac{1+r}{1-r}\right) -$

$$\frac{1.96}{\sqrt{n-3}}\right] \leq \rho \leq \tanh\left[\frac{1}{2} \ln\left(\frac{1+r}{1-r}\right) + \frac{1.96}{\sqrt{n-3}}\right], \text{ where}$$

the Excel or Matlab function is  $\operatorname{ATANH}(r) = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right)$ .