CHAPTER 10. SIMPLE REGRESSION AND CORRELATION

In agricultural research we are often interested in describing the change in one variable (Y, the dependent variable) in terms of a unit change in a second variable (X, the independent variable). Regression is commonly used to establish such a relationship. A simple linear regression takes the form of

$$\hat{Y} = a + bx$$

where $\hat{Y}$ is the predicted value of Y for a given value of X, a estimates the intercept of the regression line with the Y axis, and b estimates the slope or rate of change in Y for a unit change in X.

The regression coefficients, a and b, are calculated from a set of paired values of X and Y. The problem of determining the best values of a and b involves the principle of least squares.

10.1 The Regression Equation

To illustrate the principle, we will use the artificial data presented as a scatter diagram in Figure 10-1.

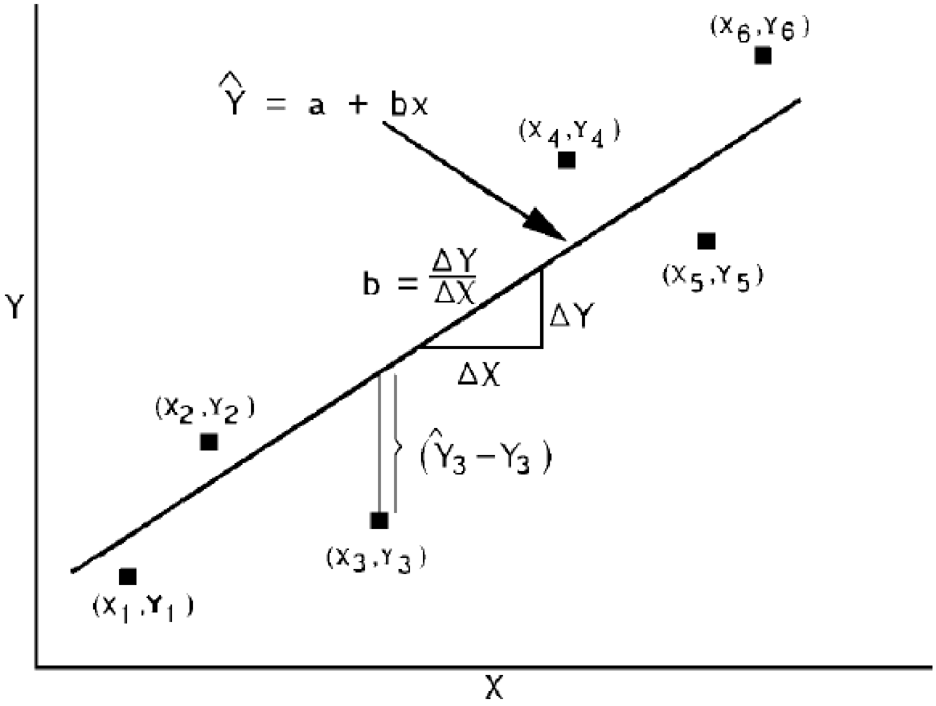

Figure 10-1. A scatter diagram to illustrate the linear relationship between 2 variables.

Because of the existence of experimental errors, the observations (Y) made for a given set of independent values (X) will not permit the calculation of a single straight line that will go through all the points. The least squares line is the line that goes through the points so that the sum of the squares of the vertical deviations of the points from the line is minimal. Those with a knowledge of calculus should recognize that this is a problem of finding the minimum value of a function. That is, set the first derivatives of the regression equation with respect to a and b to zero and solve for a and b. This procedure yields the following formulas for a and b based on k pairs of X and Y: If X is not a random variable, the coefficients so obtained are the best linear unbiased estimates of the true parameters.

$$b = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{\Sigma(X - \overline{X})^2} = \frac{\Sigma XY - (\Sigma X \Sigma Y) / k}{\Sigma X^2 - (\Sigma X)^2 / k}$$

$$a = \frac{(\Sigma X^2)\overline{Y} - \overline{X}(\Sigma XY)}{\Sigma X^2 - (\Sigma X)^2 / k} = \overline{Y} - b\overline{X}$$

Independent Variable - Fixed Design Points

In Chapter 9, we showed that a linear response was appropriate to describe the effect of N fertilizer on the sucrose content of beet roots. Note that the N rates were specifically chosen by the experimenter and, therefore, are considered fixed design points. The differences in the levels are not random. Now we show the computation of the regression equation for this situation. The first step is to complete a scatter diagram of the mean responses of % sucrose to increasing levels of N. The data are given in Table 10-1 and the scatter diagram in Figure 10-2. The construction of the least squares line is as follows:

Table 10-1. Elements necessary to compute the least squares regression for changes in % sucrose associated with changes in N-fertilizer.

X lbs N (acre)	Y mean % (sucrose)	X^2	XY	$\hat{Y}$ predicted (% sucrose)	$\hat{Y}$-Y
0	16.16	0	0	16.22	-0.06
50	15.74	2,500	787	15.78	-0.04
100	15.29	10,000	1,529	15.35	-0.06
150	15.29	22,500	2,293.5	14.92	0.39
200	14.36	40,000	2,872	14.48	-0.12
250	13.94	62,500	3,485	14.05	-0.11
(Total) 750	90.78	137,500	10,966		
(Mean) 125	15.13	22,916.67			

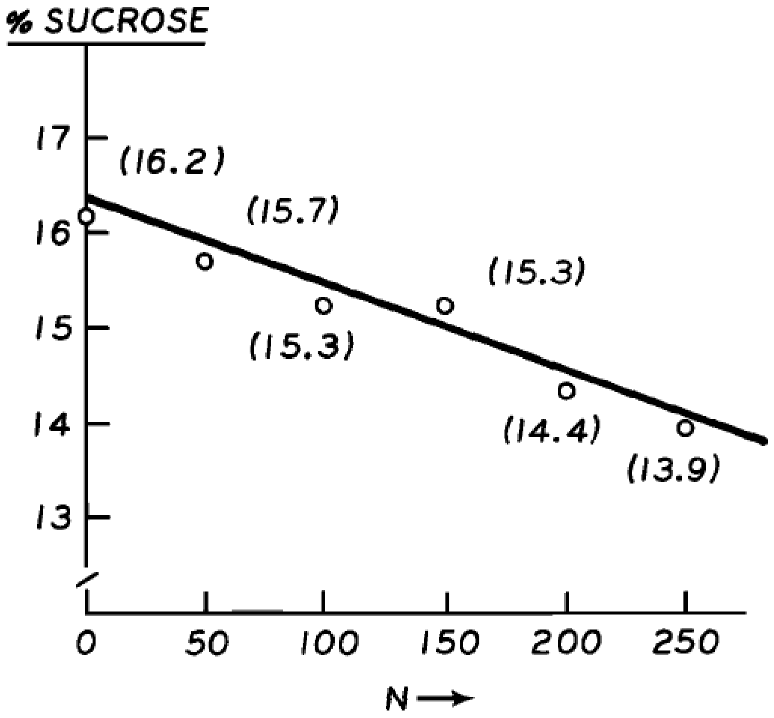

Figure 10-2. The relationship between % sucrose and levels of N-fertilizer.

$$b = \frac{\Sigma XY - (\Sigma X \bullet \Sigma Y)/k}{\Sigma X^2 - (\Sigma X)^2/k} = \frac{10966.5 - (750)(90.78)/6}{137500 - (750)^2/6}$$

$$= \frac{-381.0}{43750} = -0.0087$$

$$a = \overline{Y} - b\overline{X} = 15.13 - (-0.0087)(125) = 16.22$$

The resulting regression equation is, $\hat{Y}$ = 16.22 - - 0.0087X. This equation says that for every additional pound of fertilizer N, % sucrose decreases by 0.0087 sucrose percentage points. Our best estimate of percent sucrose from 0 to 250 lb N/acre is determined by substituting the N rate in the regression equation and calculating Y (the last column of Table 10-1). For example, we may want to estimate % sucrose for 135 lb N/acre, then

$$\hat{Y} = 16.22 - 0.0087(125) = 15.13$$

Independence variable - measurement with error

Sometimes researchers are interested in estimating a quantity that is difficult to measure directly. It is desirable to be able to predict this quantity from another variable that is easier to measure. For example, to predict leaf area from the length and width of leaves, sugar content from percent total solids, or rate of gain from initial body weight.

For a case study we will use data collected to see if it is possible to predict the weight of the livers of mice from their body weights. The data are given in Table 10-2 and the calculation of the regression line is shown below the table.

Table 10-2. Mice body and liver weights (grams) and predicted liver weights from a linear regression of Y on X.

X body wts (x10g)	X^2	Y liver wt.	Y^2	XY	$\hat{Y}$ predicted liver wt.	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
16.4	268.96	2.67	7.13	43.79	2.37	0.30	0.09
17.2	295.84	2.75	7.56	47.30	2.95	-0.20	0.04
17.6	309.76	2.99	8.94	52.62	3.24	-0.25	0.06
18.0	324.00	3.14	9.86	56.52	3.53	-0.39	0.15
18.2	331.24	3.88	15.05	70.62	3.68	-0.20	0.04
18.5	342.25	4.23	17.89	78.25	3.89	0.34	0.12
(Total)							
105.9	1,872.05	19.66	66.44	389.10		0.00	0.50
(Mean)							
17.65		3.28					

$$b = \frac{\Sigma XY - (\Sigma X \bullet \Sigma Y)/k}{\Sigma X^2 - (\Sigma X)^2/k} = \frac{349.10 - (105.9)(19.66)/6}{1872.05 - (105.9)^2/6}$$

$$a = \overline{Y} - b\overline{X} = 3.28 - 0.72(17.65) = -9.43$$

$$\hat{Y} = -9.43 + 0.72X$$

The predicted values of Y are obtained by substituting X's in the regression equation. The values of $\hat{Y}$ in Table 10-2 were calculated to several decimal places and rounded off, and therefore will not be exactly equal to $\hat{Y}$ values by using the regression equation given above.

The relation between body and liver weights and the regression line are plotted in Figure 10-3

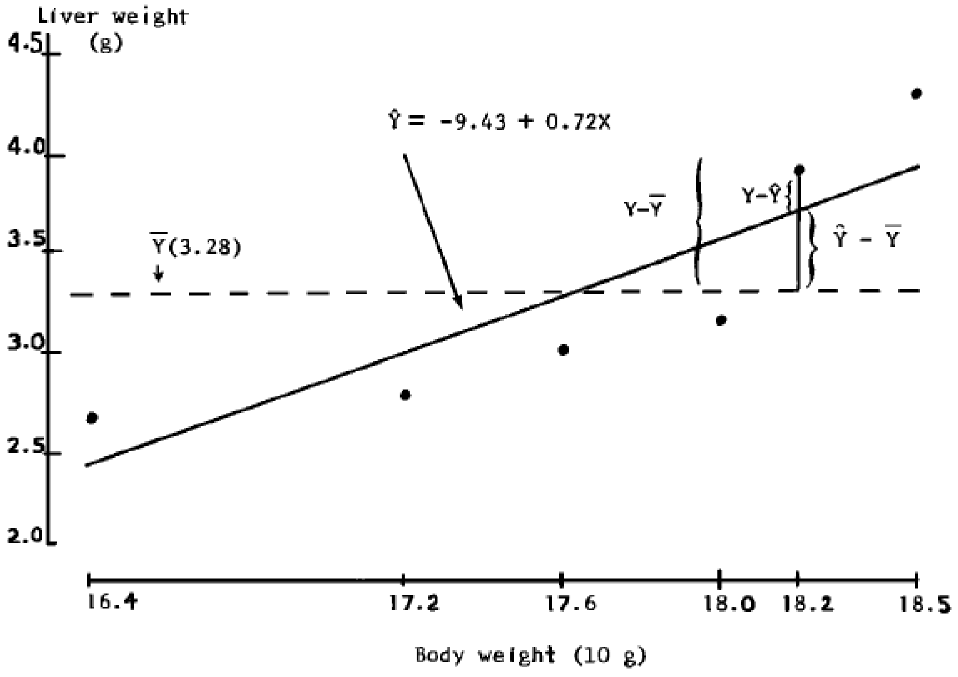

Figure 10-3. Linear regression of liver weight (g.) on body weight (10 g) of mice.

Note that the calculation procedures for determining the regressions of Figures 10-2 and 10-3 are identical. However, in the case where X values are measured with error there are two variances, one associated with measuring Y and the other with measuring X.

The variance in measuring Y is σ_e^2 and the variance in measuring X values σ_x^2 estimated regression coefficient (b) is biased toward zero by a factor of

$$\sigma_x^2 / (\sigma_x^2 + \sigma_e^2)$$

The effect of the error of X on the standard error of b is not always biased one direction, but the ratio of the regression coefficient and the standard error (that t statistic for testing a greater than zero slope) is always smaller in absolute value than the case when X values can be fixed experimentally without error. Therefore, the probability of detecting a nonzero slope is decreased. Thus an experimenter may be justified in selecting a higher probability for rejection of the null hypothesis (e.g. 10% rather than 5%).

We now turn to the consideration of the validity and usefulness of regression equations.

10-2. The analysis of variance of regression

The total sum of squares of the dependent variable (Y) can be partitioned into two components: one due to the regression line and the other due to the sum of squares not explainable by the regression line. The deviation of each Y from $\overline{Y}$ is made up of a deviation due to regression, $\hat{Y} - \overline{Y}$ and a deviation not explainable by regression (or the difference between observation and prediction), Y - $\hat{Y}$ (see Table 10-3). That is,

$$Y - \overline{Y} = (\hat{Y} - \overline{Y}) + (Y - \hat{Y})$$

squaring and summing all terms, gives

$$\Sigma(Y - \overline{Y})^2 = \Sigma(\hat{Y} - \overline{Y})^2 + \Sigma(Y - \hat{Y})^2$$

where $\Sigma(Y - \overline{Y})^2$ is the total sum of squares, SSY.

$\Sigma(\hat{Y} - \overline{Y})^2$ is the sum of squares due to the linear regression SSR, with mean square MSR.

$\Sigma(Y - \hat{Y})^2$ is the sum of squares not explainable by the regression line, and is called the residual sum of squares Ssr, with mean square Msr.

This information can be summarized in an analysis of variance table (Table 10-3).

Table 10-3. Analysis of variance of regression.

Source	DF	SS	MS	F
Total	k-1	$\Sigma(\mathrm{Y}-\overline{\mathrm{Y}})^2$		
Regression	1	$\Sigma(\hat{\mathrm{Y}}-\overline{\mathrm{Y}})^2$	MSR	MSR/MSr
Residual	k-2	$\Sigma(\mathrm{Y}-\hat{\mathrm{Y}})^2$	MSr	

The F test, MSR/MSr provides a test for the null hypothesis which is that the true regression coefficient equals Zero, $\beta = 0$, versus the alternative hypothesis that $\beta \neq 0$. This test is only valid when Msr estimates the variance of the experimental error. However, this condition cannot be tested unless there are replications of Y-values for each X so that the true experimental error can be estimated.

The machine formulas for the sum of squares in Table 10-3 are,

$$\mathrm{SSY} = \Sigma(\mathrm{Y}-\overline{\mathrm{Y}})^2 = \Sigma\mathrm{Y}^2 - (\Sigma\mathrm{Y})^2 / \mathrm{k}$$

$$\mathrm{SSR} = \Sigma(\hat{\mathrm{Y}}-\overline{\mathrm{Y}})^2 = \frac{[\Sigma\mathrm{XY} - (\Sigma\mathrm{X})(\Sigma\mathrm{y}) / \mathrm{k}]^2}{\Sigma\mathrm{X}^2 - (\Sigma\mathrm{X})^2 / \mathrm{k}}$$

$= \mathrm{b}[\Sigma\mathrm{XY} - (\Sigma\mathrm{X})\ (\Sigma\mathrm{Y})/\mathrm{k}]$

or in another form

$= \mathrm{b}^2[\Sigma\mathrm{X}^2 - (\Sigma\mathrm{X})^2/\mathrm{k}]$

$$\mathrm{SSr} = \Sigma(\mathrm{Y}-\hat{\mathrm{Y}})^2$$

$$= [\Sigma\mathrm{Y}^2 - (\Sigma\mathrm{Y})^2 / \mathrm{k}] - \frac{[\Sigma\mathrm{XY} - (\Sigma\mathrm{X})(\Sigma\mathrm{Y}) / \mathrm{k}]^2}{\Sigma\mathrm{X}^2 - (\Sigma\mathrm{X})^2 / \mathrm{k}}$$

For the data of Table 10-2, the AOV for the regression is presented in Table 10-4 with calculations shown below.

Table 10-4. AOV of regression of the liver weight on body weight.

Total	5	2.02		
Regression	1	1.53	1.53	11.77
Residual	4	0.49	0.13	

$\mathrm{SSY} = \Sigma\mathrm{Y}^2 - (\Sigma\mathrm{Y})^2/\mathrm{k} = 66.44 - 19.66^2/6 = 2.02$

$\mathrm{SSR} = \Sigma(\hat{\mathrm{Y}}-\overline{\mathrm{Y}})^2 = (2.37 - 3.28)^2 + ... + (3.89 - 3.28)^2 = 1.53$

$\mathrm{Ssr} = \mathrm{SSY} - \mathrm{SSR} = 2.02 - 1.53 = 0.49$

The significant F-test suggests that there is a nonzero regression coefficient. However, due to the lack of replication, no rigorous assessment of lack of fit to the model can be made.

10.3 Testing Fitness of a Regression Model

In this section, data of nitrogen content in corn crops obtained from a CRD field experiment will be used to illustrate the procedure of testing fitness of a regression model. Five levels of fertilization with a stable isotopic formulation of ammonium sulfate were involved in the experiment. This formulation enabled the researcher to distinguish between nitrogen in the crop derived from the fertilizer and soil. The data are shown in Table 10-5.

Table 10-5. Nitrogen (lb/acre) in a corn crop (green, cobs and fodder) derived from 5 rates of N^{15} depleted ammonium sulfate.

Fertilizer N	Replication			
lb/acre	1	2	3	Mean
50	20.47	20.91	18.15	19.84
100	41.61	44.07	60.03	48.57
150	89.06	86.27	87.16	87.50
200	83.83	116.16	120.67	106.89
250	121.43	250	153.68	133.45

The data in Table 10-5 are plotted in Figure 10-4, a practice that provides a visual examination of the response trend.

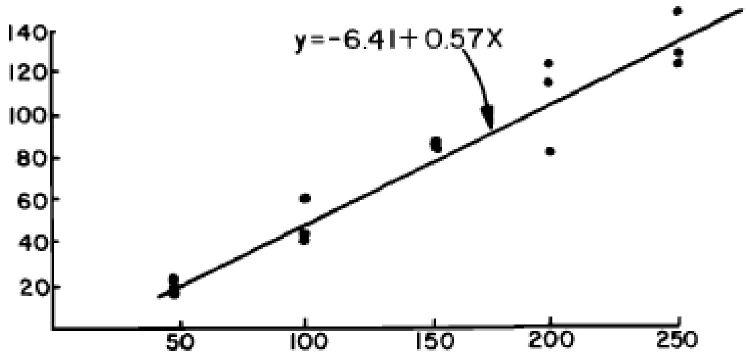

Figure 10-4. A plot of the data in Table 10-5.

The intercept and regression coefficients are calculated as shown in Section 10.1. this can be done by using all observations or by just using the treatment means. The coefficients will be the same in either case. To simplify the calculation, we use means.

$$b = \Sigma(X_i - \overline{X})(\overline{Y}_i - \overline{Y}) / \Sigma(X_i - \overline{X})^2$$

$$= (\Sigma X_i \overline{Y}_i - (\Sigma X_i)(\Sigma \overline{Y}_i) / k) / (\Sigma X_i^2 - (\Sigma X_i)^2 / k)$$

$$= 0.57$$

and

$$a = \overline{Y} - b\overline{X} = -6.41$$

The regression coefficient estimates the rate of fertilizer-N recovery by the crop, that is, 57% of the applied fertilizer-N is taken up by the corn crop. Note that the true intercept should not be less than zero which indicates that the information of this regression should not be extrapolated below 50 lb/acre fertilizer N.

To test how well the regression model fits the data, we proceed with the analysis outlined in Table 10-6.

Table 10-6. AOV to test fitness of a regression model with k levels of treatment and n replications per treatment.

Source	df	SS	MS	F
Total	kn-1	SSY		
Regression	1	SSR	MSR	
Residual	kn-2	Ssr	Msr	
Deviation	k-2	SSD	MSD	MSD/MSE
Exp. error	k(n-1)	SSE	MSE	

$$SSY = \Sigma\Sigma(Y_{ij} - \overline{Y})^2 = \Sigma\Sigma Y_{ij}^{2} - (\Sigma\Sigma Y_{ij})^2 / kn$$

$$SSR = n\Sigma(\hat{Y}_i - \overline{Y})^2 = n\ \{\frac{[\Sigma X_i \overline{Y}_i - (\Sigma X_i)(\Sigma \overline{Y}_i) / k]^2}{\Sigma X_{i}^{2} - (\Sigma X_i)^2 / k}\}$$

$$SSr = \Sigma\Sigma(Y_{ij} - \hat{Y}_i)^2 = SSY - SSR$$

$$SSE = \Sigma\Sigma(Y_{ij} - \overline{Y}_i.)^2 = \Sigma\Sigma\ _{ij}^{2} - n(\Sigma \overline{Y}_{i}^{2}.)$$

The difference between Ssr and SSE measures the deviations of the data points from the regression line that are not due to experimental error. This is frequently called the "lack of fit" sum of squares, and is denoted as sum of squares of deviation, SSD.

$$SSD = Ssr - SSE$$

The ratio of MSD/MSE provides a F test of the lack of fit of the regression model. The nonsignificance of this F value indicates that the deviation from the linear regression is entirely due to random error. Thus a linear regression is a good description of the relationship between the dependent and independent variables. A significant F test would suggest the existence of a non-random deviation from the linear model and that the data may be better described by some other model.

For the data in Table 10-5, the AOV is given in Table 10-7.

Table 10-7. AOV of the nitrogen recovery data.

Source	df	SS	MS	F
Total	14	26,302.81		
Regression	1	24,459.36	24,459.36	172.49
Residual	13	1,843.45	141.80	
Deviation	3	262.09	87.36	0.55
Exp. error	10	1,581.35	158.13	

The nonsignificant lack of fit F test (F = 0.55) indicates that a linear regression is an adequate model to describe the uptake of fertilizer-N by corn. The hypothesis of a zero regression slope is then tested by using the residual mean square to form the test F = MSR/MSr = 172.49. The F is highly significant ($P < 0.01$) indicating the null hypothesis should be rejected. If the lack of fit F test is significant, then MSE should be used instead of Msr to form a F test (F = MSR/MSE) about the hypothesis of a zero regression slope.

10.4 Coefficient of determination and correlation coefficient

The coefficient of determination (R^2) is defined as

$$R^2 = \frac{SSR}{SSY} \; (= \frac{F}{F + k - 2})$$

If a regression model fits perfectly, the total sum of squares is due entirely to regression, then $R^2 = 1$. Thus R^2. (100) indicates the percent of total variability accounted for by the regression. For a simple linear regression R^2 is commonly denoted as r^2. In the example of regressing liver weight on body weight,

$$r^2 = \frac{SSR}{SSY} = \frac{1.53}{2.02} = 0.75$$

$$(\text{also} \quad r^2 = \frac{11.77}{11.77 + 6 - 2} = 0.75)$$

Thus 75% of the variation in liver weight is accounted for by a linear relationship with the body weight. This high percentage of explainable variability gives some assurance that the linear regression equation is a reasonable description of the relationship. The coefficient of determination is particularly useful when the data do not permit a formal lack of fit test.

The correlation coefficient, r, is calculated as,

$$r = \frac{\Sigma(X - \overline{X}) \bullet (Y - \overline{Y})}{[\Sigma(X - \overline{X})^2 \bullet \Sigma(Y - \overline{Y})^2]^{1/2}}$$
$$= \frac{\Sigma XY - (\Sigma X)(\Sigma Y) / k}{[(\Sigma X^2 - (\Sigma X)^2 / k)(\Sigma Y^2 - (\Sigma Y)^2 / k)]^{1/2}}$$

r ranges between -1 and +1. A coefficient of 1 indicates a perfect linear relationship and the sign indicates the direction of the relationship. A negative sign indicates that Y decreases as X increases. Note the square of r is the coefficient of determination, but only one solution of the square root of r^2 is the correct correlation coefficient.

For the example of the animal data of Table 10-2,

$$r = \frac{349.103 - (105.9)(19.66) / 6}{\{[1872.05 - (105.9)^2 / 6][66.44 - (19.66)^2 / 6]\}^{1/2}}$$
$$= 0.87$$

Based on the observed r-value, the hypothesis that the true correlation is zero can be tested by referring to a critical value of Appendix Table A-11 with k-2 degrees of freedom. The tabular value from Appendix Table A-11 for 4 degrees of freedom and at the 5% level is 0.81. Thus we again have evidence for a significant linear relationship.

In calculating r, it is immaterial which variable is designated as dependent or independent. In computing the regression coefficient, b, it is important to recognize which of the two variables depends on the other. r and b are related as follows:

$$r = \sqrt{b_{YX} \; b_{XY}}$$

with the same sign as b_{YX} or b $_{XY}$. The coefficient b_{YX} represents the regression of Y on X and b_{XY} is the regression of X on Y.

From the formula of r, one can easily shown that

$$r = b_{YX}\sqrt{\Sigma(x-\overline{X})^2 / \Sigma(Y-\overline{Y})^2}$$

which is identical to the earlier calculation.

Note that if there is variation among X's and Y's, a nonzero r implies a nonzero b and vice versa. Therefore the F test for a nonzero β is equivalent to the test for a nonzero r (Appendix Table A-11).

Because r and r^2 are related algebraically to regression, the sum of squares for regression and residual in Table 10-3 can be expressed in terms of r^2.

$$SSR = \Sigma(\hat{Y}-\overline{Y})^2 = r^2 \bullet \Sigma(Y-\overline{Y})^2$$
$$SSr = \Sigma(y-\hat{Y})^2 = (1-r^2) \bullet \Sigma(y-\overline{Y})^2$$

where $\Sigma(Y-\overline{Y})^2$ is the total sum of squares among the Y-values. For the AOV in Table 10-5,

SSR =(0.75) • (2.02) = 1.52
Ssr = (0.25) • (2.02) = 0.51

In more complicated regression models, it is often desirable to adjust $R^2 (R^2_{adj})$ for the degrees of freedom of regression by

$$R^2_{adj} = 1 - MSr / MSY$$

For the liver-body weight examples (Table 10-4)

$$R^2_{adj} = 1 - 0.13 / (2.02 / 5) = 0.68$$

Note that the considerable reduction in R^2 (from 0.75 to 0.68) is a result of the large adjustment due to the small number of regression points. In the case of corn data in Table 10-7,

R^2 = 24459.36/26302.81 = 0.93 and
R^2_{adj} = 1 - [141.8/(26302.81/14)] = 0.925

This is a relatively small change due to the greater number of degrees of freedom.

10.5 Confidence limits for the regression coefficient (β)

The mean square for residuals from regression (Msr) can be used as an estimate of the population variance if there is no significant lack of fit. Its square root is used to compute standard errors for various statistics associated with regression. The standard error of the regression coefficient b is S_b

$$S_b = [MSr / \Sigma(X - \overline{X})^2]^{1/2}$$

To construct a confidence interval for the true regression coefficient, we need a t-statistic which is defined as

$$t = (b - \beta/S_b \quad \text{with df} = k\text{-}2$$

where β is the true regression coefficient. This formula provides a t-test for significance for a nonzero β equivalent to the F-test of the AOV for regression.

The confidence limits of β for a given α are,

$$b \pm t_{,\alpha,k-2} \bullet S_b$$

The 95% C. L. for for the regression of liver weight on body weight are computed as follows,

$$MSr = 0.13, \quad \Sigma(X - \overline{X})^2 = 2.92,$$
$$b = 0.72 \quad \text{and } t_{0.05,4} = 2.776$$
$$S_b = \sqrt{0.13/2.92} = 0.21$$

and

$$\begin{matrix} L \\ U \end{matrix} = 0.72 \pm 2.776(0.21) = 0.72 \pm 0.58$$

$$= \begin{matrix} 0.14 \\ 1.30 \end{matrix}$$

Note that the confidence limits do not bracket zero, also suggesting a nonzero β.

10.6 Confidence limits for the intercept (a_0)

The standard error of the estimated Y intercept, a, is

$$S_a = \{MSr\,[\frac{1}{k} + \frac{\overline{X}^2}{\Sigma(X - \overline{X}^2}]\}^{1/2}$$

Again a t statistic for testing the hypotheses related to the true intercept, a_O is,

$$a \pm t_{,df} \bullet S_a$$

For the liver-body weight example,

$$MSr = 0.13, \quad \overline{X} = 17.65, \quad \Sigma(X - \overline{X})^2 = 2.92, \quad a = -9.43$$
$$t_{0.05,4} = 2.776 \qquad \text{and}$$
$$S_a = \{0.13\,(\frac{1}{6} + \frac{17.65^2}{2.92})\}^{1/2}$$

= 3.73

-9.43 ± (2.776)(3.73) = -9.43 ± 10.35
= -19.78
= 0.92

Since this confidence limit brackets zero, we do not have evidence for a nonzero intercept. For the purpose of predicting the liver weight from body weight, the intercept value has little meaning. In other cases, however, inferences about the intercept may have considerable importance.

10.7 Confidence limits for predicted values ($\hat{Y}$)

$\hat{Y}$ can be considered as a predicted value for a mean of Y's for a given X or as a predicted value for an individual Y for a given X. In both cases $\hat{Y}$ will be the same but the reliability will be different.

The standard error for $\hat{Y}$ when it predicts a Y-mean is,

$$S_{\hat{\bar{y}}} = [MSr(\frac{1}{k} + \frac{(X - \overline{X}^2}{\Sigma(X - \overline{X})^2}])^{1/2} \qquad \text{with df = k-2}$$

and the standard error for   when it predicts an individual Y is,

$$S_{\hat{y}} = [MSr(1 + \frac{1}{k} + \frac{(X - \overline{X})^2}{\Sigma(X - \overline{X})^2})]^{1/2} \qquad \text{with df = k-2}$$

The (1-α) confidence limits for a predicted mean are

$$\hat{Y} \pm t_{\alpha,df} \bullet S_{\hat{\bar{y}}}$$

and for an individual $\hat{Y}$ are,

$$\hat{Y} \pm t_{\alpha,df} \bullet S_{\hat{y}}$$

Note that a different standard error is needed for each value of X and that the greater the departure of X from $\overline{X}$ larger the standard error. The reliability of a prediction decreases as X moves away from the central value of the independent variable. Thus extrapolation from a regression model should be done with great caution.

Table 10-8 gives confidence limits for predicted means and individual observations for the liver-body weight data. The calculations for completing the first line of the table are as follows:

$$S_{\hat{\bar{y}}} = [0.13(\frac{1}{6} + \frac{(16.4 - 17.65)^2}{2.92})]^{1/2} = 0.30$$

$$\begin{matrix} L \\ U \end{matrix} = 2.37 \pm 2.776(0.30) = 2.37 \pm 0.83$$

L = 1.54 and U = 3.20

$$S_{\hat{y}} = [0.13(1 + \frac{1}{6} + \frac{(16.4 - 17.65)^2}{2.92})]^{1/2} = 0.47$$

$$\begin{matrix} L \\ U \end{matrix} = 2.37 \pm 2.776(0.47) = 2.37 \pm 1.30$$

L = 1.07 and U = 3.67

Thus, for a given body weight of 164 g., we have 95% confidence that the mean liver weight will fall between 1.54 and 3.20 g. and that an individual liver weight will fall in the range of 1.07 and 3.67 g.

Table 10-8. 95% confidence limits for predicted means and individual observations for the liver - body weight data.

				Limits for predicted mean		Limits for individual observation	
X (10 g)	Y (g)	$S_{\hat{\bar{y}}}$	$S_{\hat{y}}$	L	U	L	U
16.4	2.37	0.30	0.47	1.54	3.20	1.07	3.67
17.2	2.95	0.18	0.40	2.45	3.45	1.84	4.06
17.6	3.24	0.15	0.39	2.82	3.66	2.16	4.32
18.0	3.53	0.16	0.40	3.09	3.97	2.42	4.64
18.2	3.68	0.19	0.41	3.15	4.21	2.54	4.82
18.5	3.89	0.23	0.43	3.25	4.53	2.70	5.08

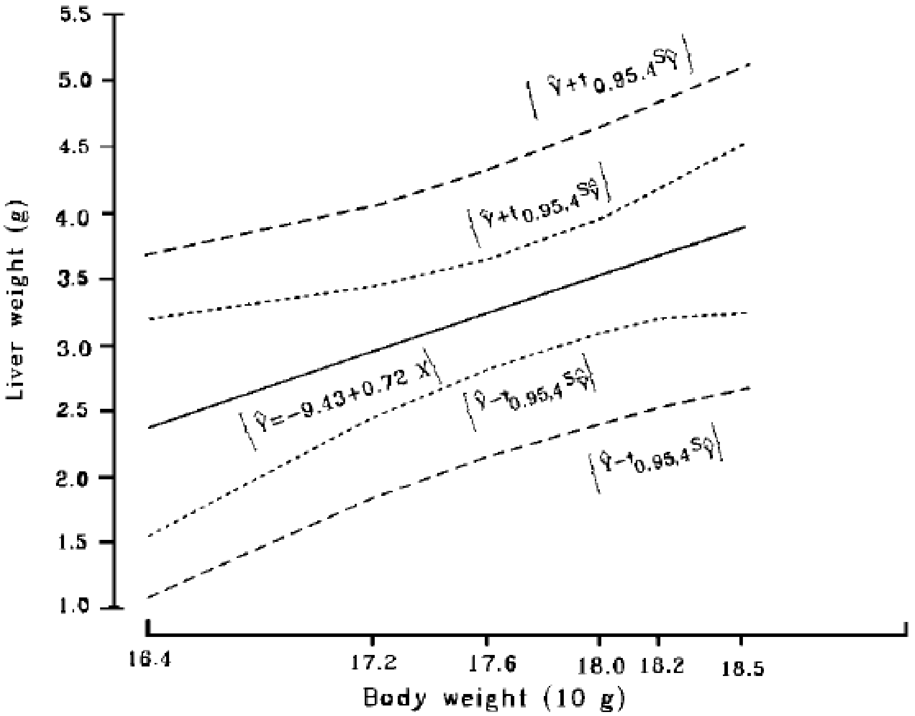

Figure 10-5. 95% confidence bands for the predicted means and predicted individual observations of liver weight from body weight.

10.8 Difference between two regression coefficients or two intercepts

If two regression coefficients b_1 and b_2 are compared to determine whether or not they can be considered estimates of a common , the following t-test can be applied.

$$t = \frac{b_1 - b_2}{[S_p^2(\frac{1}{SSX_1} + \frac{1}{SSX_2})]^{1/2}} \qquad \text{with df} = k_1 + k_2 - 4,$$

SSX_1 and SSX_2 are the sums of squares of the X's of the two regressions, and

$$S_p^2 = \frac{SSr_1 + SSr_2}{(k_1 - 2) + (k_2 - 2)}$$

is a pooled estimate of the variation about regression. SSr_1 is the sum of squares of deviation from the first regression and Ssr2 is the sum of squares of deviation in the second regression analysis.

To illustrate the procedure, we use the % sucrose data from Table 8-4. Suppose we are concerned that the effect of nitrogen levels on % sucrose is the same in block 1 and block 5. The data are presented in Table 10-9. The AOV is presented in Table 10-10.

Table 10-9. % sucrose for two blocks.

X	(block 1) Y_1	(block 5) Y_2
0	16.45	16.40
50	16.30	16.30
100	15.35	14.75
150	15.55	15.40
200	13.90	14.10
250	13.60	14.45

Table 10-10. Regression AOV of % sucrose in blocks 1 and 5.

		Block		Block 5	
Source	df	SS	MS	SS	MS
Total	5	7.167		4.658	
Regression	1	6.428	6.428	3.532	3.52
Residual	4	0.739	0.185	1.126	0.282

$$\hat{Y}_1 = 16.71 - 0.0121 \ X \qquad r_1^2 = 0.90$$

$$\hat{Y}_2 = 16.35 - 0.0090 \ X \qquad r_2^2 = 0.76$$

Now the H_0: $\beta_1 = \beta_2$ versus H_1: $\beta_1 \neq \beta_2$ can be tested by

$$S_p^2 = \frac{0.739 + 1.126}{4 + 4} = 0.233$$

$$t = \frac{-0.0121 - (-0.0090)}{[0.22 \ (\frac{1}{43750.0} + \frac{1}{43750.0})]^{1/2}}$$

$$= \frac{-0.0031}{0.0032} = -0.9775$$

with df = 8 which is not significant at the 5% level. Therefore we conclude there is no evidence that the rate of sucrose reduction in the two blocks is different.

Similarly the difference between two intercepts can also be tested by

$$t = \frac{a_1 - a_2}{\{S_p^2[(\frac{1}{k_1} + \frac{\overline{X}_1^2}{\Sigma(X_1 - \overline{X}_1)^2}) + (\frac{1}{k_2} + \frac{\overline{X}_2^2}{\Sigma(X_2 - \overline{X}_2)^2})]\}^{1/2}}$$

with df = $k_1 + k_2 - 4$.

For the above example,

the H_O: $a_{01} = a_{02}$ versus H_1: $a_{01} \neq a_{02}$

is tested by calculating

$$t = \frac{16.70 - 16.35}{[0.22(2)\,(\frac{1}{6} + \frac{125^2}{43750})]^{1/2}} = \frac{0.25}{0.4801} = 0.73$$

with df = 8. Again the null hypothesis $a_{01} = a_{02}$ cannot be rejected.

10.9 Inverse prediction

Sometimes, it is desirable to use the regression equation to estimate the unknown value of X associated with an observed Y value. This is particularly useful for bioassays, such as to estimate the concentration of a toxic substance in the soil from the comparison of growth of plants in the soil with a standard growth curve. Suppose the following data were obtained in an experiment for establishing the standard curve.

Table 10-11. % growth as control (Y) under herbicide application (X).

	X	Y		$(Y -)^2$
	-2	0.95	0.9	0.0025
	-1	0.75	0.7	0.0025
	0	0.50	0.5	0.0000
	1	0.25	0.3	0.0025
	2	0.05	0.1	0.0025
Total	0	2.50	2.5	0.01
Mean	0	0.5	0.5	

Msr = (0.01)/(5 - 2) = 0.0033

Where X is the coded concentration of a herbicide (say, the concentration is 10^{X+3} ppm) and Y is the observed growth, as a percent of control, the estimate regression equation is

$$\hat{Y} = 0.5 - 0.2X$$

and $S_b^2 = MSr / \Sigma(X - \overline{X})^2 = 0.00033$

The inverse estimation is the estimate, X, for an observed Y value:

$$\hat{X} = (Y - a)/b$$

A 100 • (1 -) confidence interval for the true but unknown X can be constructed as

$$\begin{matrix}L\\U\end{matrix} = \bar{X} + \frac{b(Y-\bar{Y})}{C} \pm \frac{t \bullet \sqrt{MSr}}{C} \bullet [D(Y-\bar{Y})^2 + C(\frac{k+1}{k})]^{1/2}$$

where $C = (b^2 - t^2 \bullet S_b^2)$ and $D = [1/\Sigma(X-\bar{X})]^2$

In our example if we observed a plant growth of 70% in soil which contains an unknown amount of the herbicide, the estimated concentration is

$$\hat{X} = \frac{0.7-0.5}{-0.2} = -1.0$$

(or 10^{-1+3} = 100 ppm)

The 95% confidence interval of the herbicide concentration is:

$$\begin{matrix}L\\U\end{matrix} = 0 + \frac{(-0.2)(0.7-0.5)}{C} \pm \frac{t\sqrt{0.0033}}{C}\sqrt{D(0.7-0.5)^2 + C(6/5)}$$

where $t_{5\%,\,3}$ = 3.18, C = $[(0.2)^2 - (3.18)^2(0.00033)]$ = 0.0367

and $D = 1/\Sigma(X-\bar{X})^2 = 0.1$. Thus

$$\begin{matrix}L\\U\end{matrix} = \begin{matrix}-2.1809\\0.0011\end{matrix} \quad \text{or} \quad \begin{matrix}6.6\text{ ppm}\\1002.54\text{ ppm}\end{matrix}$$

If several observations (say r observations) are made at the unknown level of X, the average value of these r observations should be used in the above formulas. Also replace k by (k + r) and $t_{\ ,(k-2)}$ by $t_{\ ,(k+r-3)}$ in the formula to construct an approximate (1-) confidence interval.

10.10. Relationship between regression and analysis of variance models.

In Chapter 7, 8, and 9, experiments that were designed to study the response of a variable to varying doses of a treatment were analyzed. The total sum of squares for any particular variable is partitioned according to the design of the experiment and the assumed model of the response. For example, in a typical one-way classification experiment from CRD with several treatments, the analysis of variance model is,

$$Y_{ij} = \mu + \ _{i} + \ _{ij}$$

where Y_{ij} is the jth observation in ith treatment, $_{i}$ is the ith treatment effect, and $_{ij}$ is the experimental error associated with the observation Y_{ij}. Based on this model, the sum of squares of Y is partitioned into sum of squares of treatment and sum of squares of error. In a regression analysis, the model may be specified as:

$$Y_{ij} = \ + \ X_{i} + \ _{ij}$$

where and are intercept and regression coefficients, respectively, and X_i is the ith level of the independent variable. The sum of squares of Y is now partitioned into sum of squares of regression and sum of squares of residuals. When the treatments are quantitative levels of a factor and a regression relationship can be defined, the two procedures are closely related.

The example of percent sucrose in sugarbeet shown in chapters 8 and 9 will be used to illustrate the relationship. For the purpose of easy explanation, the means of the experimental plots are presented in Table 10-12.

Table 10-12. Mean % sucrose for two beet samples per plot from an RCBD.

Fertilizer N	Block					Treatment
(lb/acre)	1	2	3	4	5	Mean
0	16.45	16.10	15.50	16.35	16.40	16.16
50	16.30	14.15	16.05	15.90	16.30	15.74
100	15.35	14.65	16.05	15.65	14.75	15.29
150	15.55	14.95	15.55	15.00	15.40	15.29
200	13.90	13.75	14.80	15.25	14.10	14.36
250	13.60	12.55	14.70	14.40	14.45	13.94
Block mean	15.19	14.36	15.44	15.43	15.23	15.13

If the treatment means were analyzed according to the simple regression model, the following analysis of variance result would be obtained.

Table 10-13. Analysis of variance of regression of % sucrose on N levels based on treatment means.

Source	df	SS	MS	F
Total	5	3.494		
Regression	1	3.318	3.318	75.41
Residual	4	0.176	0.044	

$$\Sigma(Y - \overline{Y})^2 = (16.16 - 15.13)^2 + ... + (13.93 - 15.13)^2 = 3.494$$
$$\Sigma(\hat{Y} - \overline{Y})^2 = (16.20 - 15.13)^2 + ... + (14.04 - 15.13)^2 = 3.318$$
$$\Sigma(Y - \hat{Y})^2 = (16.16 - 16.20)^2 + ... + (13.93 - 14.04)^2 = 0.176$$

Recall that the treatment means are calculated from observations of 5 replications with 2 samples per replicate. If we multiply the total sum of squares of Table 10-13 by 10, 5 replicates times 2 samples (e.g., 3.494 (5)(2) = 34.94) we would obtain identical treatment sum of squares as in Table 8-6 or Table 9-7. The term Ssr (sum of squares of residual) in Table 10-13 multiplied by 10 will also yield the same sum of squares of residual of Table 9-7. This illustrates that the AOV for the regression of treatment means is merely a partitioning of the treatment sum of squares.

The estimated regression equation as calculated in section 1 is:

$$\hat{Y} = 16.22 - 0.0087X$$

If all the plot means were used in the regression analysis, the estimated regression equation would still be the same as calculated from regression on treatment means but the AOV would be as shown in Table 10-14.

Table 10-14. AOV of regression of sucrose of sugarbeets based on plot means.

Source	df	SS	MS	F
Source	df	SS	MS	F
Total	29	27.85		
Regression	1	16.57	16.57	41.43
Residual	28	11.28	0.40	

When we compare the analysis of variance results in Table 10-14 and Table 9-7, we find the following:

From Table 10-14, regression AOV of all plots.			From Table 9-7, regression AOV of individual observations.	
df	Sum of squares x	Sample per plot =	Sum of squares	df
29	Sum of squares of total	x 2 =	Sum of squares of plots	29
1	Sum of squares of regression	x 2 =	Linear treatment of squares	1
28	Sum of squares of residual	x 2 =	Sum of squares of (blocks	4
			+	+
			residual	4
			+	+
			exp. error	20

In this case the factor 2 is the sample size per plot, or the number of observations used to calculate plot means in the regression analysis.

Basically, there is no difference in ideas between the regression analysis of variance and the analysis of variance of experiments as discussed in Chapters 7, 8, and 9. It all depends on how the data are organized and which part of the sum of squares are partitioned. The sum of squares of regression for a quantitative treatment is only a part of the treatment sum of squares. In some cases Y and X are paired observations without replicates, and a simple partitioning of sum of squares of Y (dependent variable) into sum of squares of regression and of residual may be adequate. However, in designed experiments where multiple observations of Y are made for each fixed level of X, a complete breakdown of the total sum of squares of individual sample observations as shown in Table 9-7 is recommended.

The relationships of various procedures in partitioning sums of squares for CRD and RCBD experiments are shown in Tables 10-15 and 10-1

Table 10-15. Relationships among various procedures in partitioning sum of squares for a CRD experiment (k levels of treatments and r replications).

Procedures

1. AOV on treatment means (Regression analysis on means)

   Total S.S.{$\Sigma\overline{Y}_{i}.-\overline{Y}..)^2$ with df = k - 1}

   = Regression S.S.{$\Sigma(\hat{Y}_{i}.-\overline{Y}..)^2$ with df = 1}

   + Deviation S.S.{$\Sigma(\hat{Y}_{i}.-\overline{Y}_{i}.)^2$ with df = k - 2}

2. AOV on all data (Design analysis on data)

   Total {$\Sigma\Sigma(Y_{ij}-\overline{Y}..)^2$ with df = kr - 1}

   = Treatment S.S.{$r\Sigma(\overline{Y}_{i}.-\overline{Y}..)^2$ with df = k - 1}

   + Error {$\Sigma\Sigma(Y_{ij}-\overline{Y}_{i}.)^2$ with df = k(r - 1)}

3. AOV on all data (Regression analysis on data)

   Total S.S.{$\Sigma\Sigma(Y_{ij}-\overline{Y}..)^2$ with df = kr - 1}

   = Regression $r\Sigma(\hat{Y}_{i}.-\hat{Y}_{i}.-\overline{Y}..)^2$ with df = 1}

   + Residual S.S.{$\Sigma\Sigma(Y_{ij}-\hat{Y}_{i}.)^2$ with df = kr - 2}

   and

   Residual S.S.{$\Sigma\Sigma(Y_{ij}-\hat{Y}_{i}.)^2$ with df = k - 2}

   = Deviation S.S.{$r\Sigma\hat{Y}_{i}.-\overline{Y}_{i}.-\overline{Y}_{i}.)^2$ with df = k - 2}

   + Error S.S.{$\Sigma\Sigma(Y_{ij}-Y_{i}.)^2$ with df = k(r - 1)}

Relationships among analyses

1. (Total S.S.) of procedure - 1 x ®
   = (Treatment S.S.) of procedure - 2
   = (Regression S.S.) + (Deviation S.S.) of procedure - 3
2. (Regression S.S.) of procedure - 1 x ®
   = (Regression S.S.) of procedure - 3
3. (Deviation S.S.) of procedure - 1 x ®
   = Deviation S.S. of procedure - 3
4. (Total S.S.) of procedure - 2
   = (Total S.S.) of procedure - 3

Table 10-16. Relationships among various procedures in partitioning sum of squares for a RCBD experiment (k levels of treatments and r blocks).

Procedures

1. AOV on treatment means (Regression analysis on means)

   Total $\text{S.S.}\{\Sigma(\overline{Y}_i.-\overline{Y}..)^2$ with df = k - 1}

   = Regression $\text{S.S.}\{\Sigma(\hat{Y}_i.-\overline{Y}..)^2$ with df = 1}

   + Deviation $\text{S.S.}\{\Sigma(\hat{Y}_i.-\overline{Y}_i.)^2$ with df = k - 2}

2. AOV on all data (Design analysis on data)

   Total $\text{S.S.}\{\Sigma\Sigma(Y_{ij}-\overline{Y}..)^2$ with df = kr - 1}

   = Treatment $\text{S.S.}\{r\Sigma(\overline{Y}_i.-\overline{Y}..)^2$ with df = k - 1}

   + block $\text{S.S.}\{k\Sigma(Y._{j_j}-\overline{Y}..)^2$ with df = r - 1}

   + Error $\text{S.S.}\{\Sigma\Sigma(Y_{ij}-\overline{Y}_i.-\overline{Y}._j+\overline{Y}..)$
   with df = (k - 1)(r - 1)}

3. AOV on all data (Reression analysis on data)

   Total $\text{S.S.}\{\Sigma\Sigma(Y_{ij}-\overline{Y}..)^2$ with df = kr - 1}

   = Regression $\text{S.S.}\{k\Sigma(\hat{Y}_i.-\overline{Y}..)^2$ with df = 1}

   + Residual $\text{S.S.}\{\Sigma\Sigma(Y_{ij}-\hat{Y}_i.)^2$ with df = kr - 2}

   and

   Residual $\text{S.S.}\{\Sigma\Sigma(Y_{ij}-\hat{Y}_i.)^2$ with df = kr - 2}

   = Block $\text{S.S.}\{k\Sigma(\overline{Y}._j-\overline{Y}..)^2$ with df = r - 1}

   + Deviation $\text{S.S.}\{r\Sigma\hat{Y}_i.-\overline{Y}_i.)^2$ with df = k - 2}

   + Error $\text{S.S.}\{\Sigma\Sigma(Y_{ij}-\overline{Y}_i.-\overline{Y}._j+\overline{Y}..)^2$
   with df = (k - 1) (r - 1}

Relationships among analyses

1. (Total S.S.) of procedure - 1 x ®
   = (Treatment S.S.) of procedure - 2
   = (Regression S.S.) + (Deviation S.S.) of procedure - 3
2. (Regression S.S.) of procedure - 1 x ®
   = (Regression S.S.) of procedure - 3
3. (Deviation S.S.) of procedure - 1 x ®
   = Deviation S.S. of procedure - 3
4. (Total S.S.) of procedure - 2
   = (Total S.S.) of procedure - 3

SUMMARY

1. Coefficients of a simple regression equation

$$b = \frac{\Sigma XY - (\Sigma X)\,(\Sigma Y)/k}{\Sigma X^2 - (\Sigma X)^2/k}$$

$$a = \overline{Y} - b\overline{X}$$

2. AOV of a regression

Source	df	SS	MS
Total	k-1	$\Sigma(Y-\overline{Y})^2$	
Regression	1	$\Sigma(\hat{Y}-\overline{Y})^2$	MSR
Residual	k-2	$\Sigma(Y-\hat{Y})^2$	MSr

If replications (say, n experimental units with and without s sampling units) of Y's are available for each given X, then the random variation estimate, MSE, can be calculated to compare with Msr which is calculated based on treatment means, i.e.,

$$F = \frac{MSr(n)}{MSE} \qquad \text{without subsamples}$$

$$= \frac{MSr(n)(s)}{MSE} \qquad \text{with subsamples}$$

will provide a test of the lack of fit to the regression model.

3. The coefficient of determination, R^2 or r^2 is defined as

$$R^2 = \frac{SSR}{SSY} \text{ and } R^2_{adj} = 1 - \frac{MSr}{MSY},$$

where SSR is the sum of squares of regression, SSY is the total sum of squares and Msr is the mean square of residual. The simple correlation coefficient is

$$r = \frac{\Sigma XY - (\Sigma X)\,(\Sigma Y)/k}{\{[\Sigma X^2 - (\Sigma X)^2/k][\Sigma Y^2 - (\Sigma Y)^2/k]\}^{1/2}} \text{ with df} = k-2$$

The significance of r can be tested by referring to critical values of Appendix Table A-11.

4. Confidence limits of β are,

$$b \pm t_{\alpha,k\text{-}2} \bullet S_b$$

where $S_b = [MSr/\Sigma(X-\overline{X})^2]^{1/2}$ with df = k - 2

5. Confidence limits of the true intercept (a_0) are

$$a \underline{+} t_{\alpha,df} \bullet S_a$$

where $S_a = [MSr(\frac{1}{k} + \frac{\overline{X}^2}{\Sigma(X - \overline{X})})]^{1/2}$

6. Confidence limits for a predicted mean

$$\hat{\overline{Y}} \pm t_{\alpha,df} \bullet S_{\hat{\bar{y}}}$$

where $\hat{\overline{Y}}$ is the predicted value and

$$S_{\hat{\bar{y}}} = [MSr(\frac{1}{k} + \frac{(X - \overline{X})^2}{\Sigma(X - \overline{X})^2})]^{1/2}$$

7. Confidence limits for a predicted individual observation

$$\hat{Y} \pm t_{\alpha,df} \bullet S_{\hat{Y}}$$

where

$$S_{\hat{y}} = [MSr(1 + \frac{1}{k} + \frac{(X - \overline{X})^2}{\Sigma(X - \overline{X})^2})]^{1/2}$$

EXERCISES

1. Given the following data on the amount of fertilizer (X) and the yield (Y) of wheat:

Fertilizer in Pounds X	
2	8
4	9
6	11
8	11
10	12
12	15

a) Plot the data as a scatter diagram.
b) Find the regression line of Y on X. (a = 6.6, b = 0.62)
c) Predict the wheat yield when 9 and 13 pounds of fertilizer are used.
(y = 12.26 and 14.78)

2. Complete the following table and draw an appropriate conclusion of the linear regression coefficient.

Source	df	Sum of Squares
Total	25	800
Regression		
Residual		600

3. Does the least-square line always pass through the point $(\overline{X}, \overline{Y})$?

4. Use data in exercise 1.

a) Step up the analysis of variance table and test the significance of the regression.
(F = 47.7)

b) Construct 95% confidence intervals for and .
(α: U = 4.632, L = 8.568; β: U = 0.376, L = 0.882)

c) Find 95% confidence intervals for the predicted yields of wheat at fertilizer levels of 9 and 13 pounds.

(for $\overline{y}$, U = 13.26, L = 11.26
U = 16.52, L = 13.03)

5. Show algebraically that

a) $\Sigma(\hat{Y} - \overline{Y})^2 = b\Sigma(X - \overline{X})(Y - \overline{Y}$

b) $\Sigma(Y - \hat{Y})^2 = \Sigma(Y - \overline{Y})^2 - b\Sigma(X - \overline{X})(Y - \overline{Y})$

6. Suppose that wheat yields in exercise 1 are actually means of two replications, i.e.,

Rep-1	7	7.5	10.5	10	11.7	14.5
Rep-2	9	10.5	11.5	12	12.3	15.5
$\overline{Y}$	8	9	11	11	12	15

Test the lack of fit of the regression model that was obtained in exercise 1.

(F = 0.73)

7. Sulfur and selenium are closely related chemically and it has been reported that each can reduce the uptake of the other in animals. Outbreaks of white muscle disease in lambs, a selenium deficiency symptom have been observed following application of sulfur to pastures. Data from an experiment to investigate this problem are shown below.

S application rate lbs/acre	Blood levels of lambs on subclover-grass pasture µg Se/ml
0	0.07
10	0.04
20	0.05
40	0.02
80	0.03

a) Find the regression equation. (a = 0.05475, b = -0.000425)

b) Is there a significant negative linear trend between the rate of sulfur application and the level of selenium in lambs blood? (F = 2.861)

c) Confirm the conclusion reached in b) by calculating a 95% CI.

(-0.0012243, 0.0003743)

8. Two treatments were used in an experiment to determine the N-treatment effect on growth rate of potato tubers. The conventional method used 260 Kg N/ha in four applications and 45 cm irrigations. The improved treatment was intended to minimize leaching of nitrogen. It received 170 Kg N/ha in ten applications and 27 cm irrigations. The following average accumulated dry matter of tubers were obtained.

Days after	Average dry matter	Yield (g/plant)
emergence, X	Conventional, Y_1	Improved, Y_2
35	8	9
49	50	52
71	180	170
91	250	270
104	270	310

a) Find the regression equation of Y_1 on X. (a = -133.204, b = 4.069)

b) Find the regression equation of Y_2 on X. (a = -158.91, b = 4.587)

c) Test the hypothesis that the two regression coefficients are the same.
(t = 1.845, df = 6)

d) Interpret your results.

9.Reduced legume growth under acid soil conditions is often due to the toxicity of aluminum. The effect of aluminum on the growth of soybeans supplied with adequate inorganic nitrogen was investigated. The results are presented in the table below.

	Replication (g/plant)			
Al (m)	1	2	3	Mean
0	0.51	0.63	0.65	0.5967
4	0.56	0.61	0.65	0.6067
7	0.48	0.51	0.55	0.5133
10	0.44	0.49	0.50	0.4767
20	0.28	0.35	0.42	0.3500
30	0.28	0.31	0.34	0.3100
35	0.18	0.19	0.20	0.1900

a) Calculate the regression line. (a = 0.6071, b = -0.01138)

b) Perform the AOV of the regression line, including lack of fit.
(F =416.70; 1.558)

c) Calculate the 95% confidence interval for mean soybean growth.
(0.4939, 0.5269)

10. Consider the regression problem of Y_2 on X in exercise 8 and predict the average dry matter weight at 150 days after emergence. Is the prediction reliable?

(U = 587.33, L = 470.94)

11. Crop residues are often used as soil amendments. A study was conducted to investigate the yield response of continuous corn to removal or addition of crop residue under no-tillage management. The quantities of residue to be returned to each treatment were 0, 50, 100, or 150% of that produced (150% included additional residue over what was produced). The results for 1981 and 1982 are shown below:

	Grain Yield	
Treatment	1981	1982
0	5.36	5.71
50	3.90	6.85
100	4.97	7.72
150	6.01	7.75

a) Calculate the regression equation for 1981. (a = 3.207, b = 0.01804)
b) Perform the AOV for this regression. (F = 104.068)
c) Calculate the regression equation for 1982. (a = 5.956, b = 0.01398)
d) Perform the AOV for this regression. (F = 15.067)
e) Are the two regression coefficients for 1981 and 1982 significantly different?

(t = 1.012, df = 4)

12. Data gathered to estimate the influence of residual phosphorus in the soil on wheat yield is given in the table below:

Phosphorus Applied in 1978	Yield of Wheat 1979	$\overline{Y}$
0	2910	
0	3100	3005.00
25	3006	
25	3320	3163.00
50	3340	
50	3590	3465.00

a) Calculate the regression equation. (a = 2,981; b = 9.2)
b) Perform the AOV, including lack of fit and discuss the results.

(F = 8.02; <1)

c) Calculate a 95% confidence interval for the true regression coefficient, .

(0.1835, 18.217)

13. In determining the relationship of plant growth rate (absolute ratio of dry matter accumulation per plant) and growth rate of the plant tissue (leaf area growth rate), the following data were obtained in the time interval from 22 to 29 days after planting for 10 soybean cultivars.

Growth Rate (g/plant/day) Y	Leaf Area Growth Rate (cm/plant/day) X
0.339	34.33
0.398	51.70
0.386	48.61
0.385	45.64
0.378	42.52
0.368	39.09
0.356	38.43
0.354	36.13
0.353	41.30
0.351	37.36

a) Find the regression line of Y on X. (a = 0.235, b = 0.0032)

b) Perform the AOV of the regression model. (F = 62.32)

c) Find the 99% confidence limits for the regression coefficient.
(U = 0.00456, L = 0.00184)

d) Give some reasons why or why not the leaf area growth rate is a good predictor of the plant growth rate for the data. (R^2 = 0.8862)

14. The California Department of Food and Agriculture publishes reports of agricultural statistics periodically. For purposes of the exercise we divided all crops into three categories: field Crops (A): Fruit and Nut Crops (B); and Vegetable and Melons (C). Data of harvested acreage (millions of acres) of type-B and type-C crops from 1981 to 1985 were obtained from one of the reports and are shown in the following tables.

Table 1. Harvested acreage (millions of acres) of type-B crops from 1981 to 1985 and some calculated values.

Year	Average (Y)	X^2	Y^2	XY
1	1.80	1	3.2400	1.80
2	1.82	4	3.3124	3.64
3	1.86	9	3.4596	5.58
4	1.91	16	3.6481	7.64

	5	1.96	25	3.8416	9.80
Total	15	9.35	55	17.5017	28.46

Table 2. Harvested acreage (millions of acres) of type-C crops from 1981 to 1985 and some calculated values.

	Year (X)	Acreage (Y)	X^2	Y^2	XY
	1	0.87	1	0.7569	0.87
	2	0.94	4	0.8836	1.88
	3	0.95	9	0.9025	2.85
	4	1.00	16	1.0000	4.00
	5	0.99	25	0.9801	4.95
Total	15	4.75	55	4.5231	14.55

Based on the above information, answer the following questions:

a) What are your estimates of the rate of change of harvested acreage of each crop type (B and C)? (0.041 acre/yr, 0.030 acre/yr)

b) Test whether the estimated rates (slopes) are significantly different from zero. What are the R^2-values of the fitted lines? Interpret your results. (F = 129.31; 16.875)

c) Assume that the total harvested crop land was 8.8 million acres in 1986. What is your estimated acreage of type-A crop? Give reasons of why or why not the prediction is reliable. (5.767 acres)

15. Student enrollment figures of the fall quarter on the Davis campus since 1977 are shown below. (Note: carry three digits after the decimal point for the calculation)

	Year X	Students (in 1,000) Y	X^2	Y^2	XY
	1977	17.250	3,908,529	297.562	34,103.250
	1978	17.511	3,912,484	306.635	34,636.758
	1979	17.950	3,916,441	322.203	35,523.050
	1980	18.887	3,920,400	356.719	37,396.260
	1983	18.969	3,932,289	356.823	37,615.527
	1984	19.542	3,936,256	381.890	38,771.328
	1985	19.721	3,940,225	388.918	39,146.185
Total	13,866	129.830	27,466,624	2,413.759	257,192.358

a) Based on this data, what is the rate of increase of student enrollment over the years? (300 students/yr)

b) How would you decide if the regression model is an acceptable description of the relationship between enrollment and year?

c) What is your prediction of student enrollment in the year 2000? Give an indication as to how good this prediction is with a 95% level of confidence.
(24,290 students; CI: 22,118 to 26,462)

16. A new herbicide was tested in the field for weed control. Five levels of the herbicide were applied in the study, with each level applied to four plots by a completely randomized design. The following table shows the percent weeds in each plot and the average and the total of each treatment.

	Herbicide Concentration (lbs/acre)					
	1	2	3	4	5	Total
	26	27	29	12	13	
	30	34	28	10	6	
	27	26	26	14	8	
	34	29	25	11	13	
Ave.	29.25					107
		29.00	27.00	11.75	10.00	
Total	117	116	108	47	40	428

a) Is there a linear relationship between the percent weed and the herbicide concentration? Use average of % weeds for each treatment to estimate the linear regression line and draw a residual plot. (F = 58.038)

b) Give a statistical conclusion about the significance of the lack of fit of the linear regression line. (F = 9.441)

17. For what range of values can the coefficient of correlation of a sample of 52 pairs be considered significantly different from zero at the 1% level of significance?
(range = 0.3541 to 1.00)

18. Test at the 5% level the hypothesis H_0: p = 0.50, if in a sample of 100 pairs the correlation coefficient is found to be equal to 0.80.

19. At the 5% level of significance what is the complete range of values for which the coefficient of correlation of a sample of 21 pairs of values is significantly different from a population value of 0.77? ($r \leq 0.3371$)

20. Explain the differences between regression and correlation analyses.

21. It was hypothesized that the elder brother (E) of twin pairs was slightly taller than the younger one (Y). The following table presents the height data of 8 twin brothers.

Twin	1	2	3	4	5	6	7	8	Total
E	70	67	64	71	66	68	72	65	543
Y	71	70	65	69	70	66	71	63	545
E^2	4900	4489	4096	5041	4356	4624	5184	4225	36,915
Y^2	5041	4900	4225	4761	4900	4356	5041	3969	37,193
ExY	4970	4690	4160	4899	4620	4488	5112	4095	37,034

Test if there is a significant correlation between the heights of the twin brothers.

(r = 0.682)

22. Do the following for the data in exercise 8:

a) Compute the coefficients of correlation between X and Y_1, Y_1 and Y_2, and X and Y_2.

(r = 0.987; 0.7968; 0.9966)

b) Compute coefficients of determination between X and Y_1, and X and Y_2. Interpret these coefficients. (r = 0.974; 0.993)

c) Test the hypothesis that the population correlation coefficient between Y_1 and Y_2 is 0.80 against the alternative that it is not.

23. The following data of yield components were obtained from 10 parental triticale populations in 1972. (Averages of several locations.)

Kernel Weight (mg)	No. Kernels Per Spikelet	No. Spikelets Per Spike	No. Tillers
41	2.9	37	10.6
55	2.8	34	14.7
38	3.2	25	8.8
46	2.8	27	9.0
53	2.7	24	13.4
44	1.9	26	11.6
50	3.1	25	14.9
45	3.0	29	11.4
42	2.6	33	9.3
40	3.3	28	8.9

a) Compute coefficients of correlation between each pair of components.

b) Do the correlations differ significantly from zero?

c) Interpret the results.