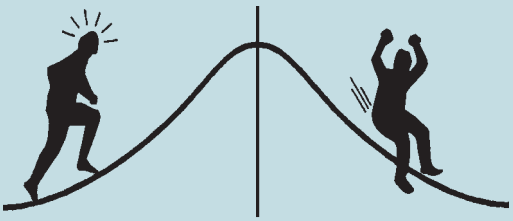


BIOMETRICS INFORMATION

HANDBOOK NO. 1 OCTOBER 1991

Pictures of Linear Models

Biometrics Information Handbook Series

BC Ministry of Forests

Pictures of Linear Models

by
Wendy Bergerud, Series Editor
Forest Science Research Branch
B.C. Ministry of Forests
31 Bastion Square
Victoria, B.C.
V8W 3E7

October 1991

BC
Ministry of Forests

Canadian Cataloguing in Publication Data

Bergerud, W. A. (Wendy Ann), 1954-
Pictures of linear models
(Biometrics information handbook series, ISSN 1183-9759 ; no. 1)

Includes bibliographical references: p.
ISBN 0-7718-9113-X

1. Forests and forestry - Statistical methods. 2. Linear models (Statistics) 3. Forests and forestry - Research. I. British Columbia. Ministry of Forests. II. Title. III. Series.

SD387.M33B47 1991 634.9'01174 C91-092300-0

Published by the
Forest Science Research Branch
Ministry of Forests
31 Bastion Square
Victoria, B.C. V8W 3E7

Copies of this and other Ministry of Forests titles are available from Crown Publications Inc., 546 Yates Street, Victoria, B.C. V8W 1K8.

ACKNOWLEDGEMENTS

I would like to thank all the regional staff who took part in my workshops on *Statistical Analysis Using SAS/PC* in the winter of 1988/1989. The idea for presenting linear models in this fashion developed while I was trying to explain various statistics during the workshops. The explanations reached a particularly high intensity with the Prince George people and I especially thank them for demanding better explanations. Also, I would like to thank Vera Sit, William Roland, Selma Low, Amanda Nemec, and Linda Stordeur for reviewing this document and providing a great many useful suggestions for improvements. I particularly thank David Izard of the Communications and Extension Services section for making the wonderful graphs.

TABLE OF CONTENTS

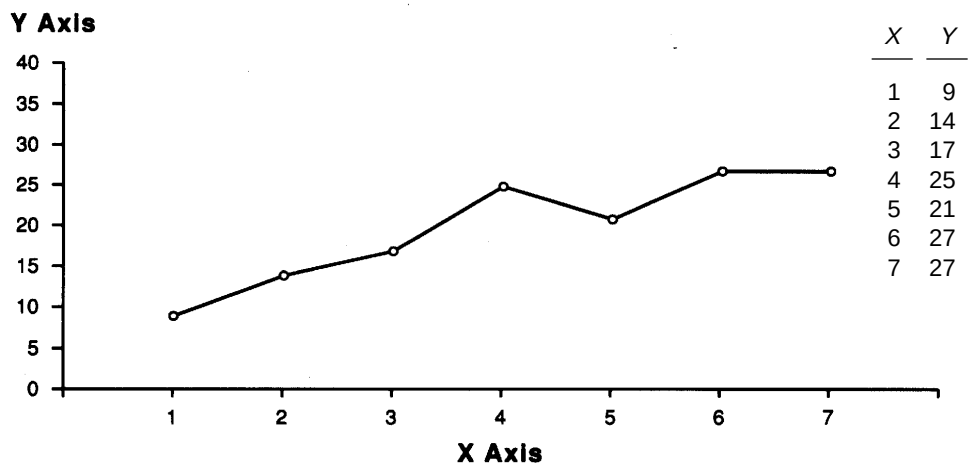
1 INTRODUCTION

This handbook explains various types of statistical models, using simple graphs and simple arithmetic. Most statistical tests used in forestry research are based on the linear models of simple and multiple regression, and analyses of variance and covariance. A linear model is a weighted sum of explanatory variables used to predict or model the response variable. The weights are called parameters and since they are usually unknown, they must be estimated from the data. The differences between the observed values and those fitted by a model are minimized through (ordinary) least squares (OLS or LS) methods. This means that a type of model is chosen for the data by the researcher and then OLS chooses the specific values of the model parameters, such as the intercept and slope for a simple linear regression. The values chosen are those for which the sum of squares of the differences (SSR) between the actual and fitted values is the least. Hence the term ''least squares''.

Several different models are usually fitted to a set of data. Pairs of models are compared where the simpler one is missing one or more parameters of the other. For instance, the mean is a simpler model than a simple linear regression model since the mean has only one parameter (namely the mean) while the regression has two (the intercept and slope)[1]. If certain assumptions about the data are reasonable, statistical tests can be used to compare such models.[2]

2 THE MEAN

The simplest model that can be fit to any set of data is a constant. Mathematically, this is written as $Y = \boldsymbol{a}$, where $\boldsymbol{a}$ is the parameter of the model. As an example, suppose that we have the following set of data:



X	Y
1	9
2	14
3	17
4	25
5	21
6	27
7	27

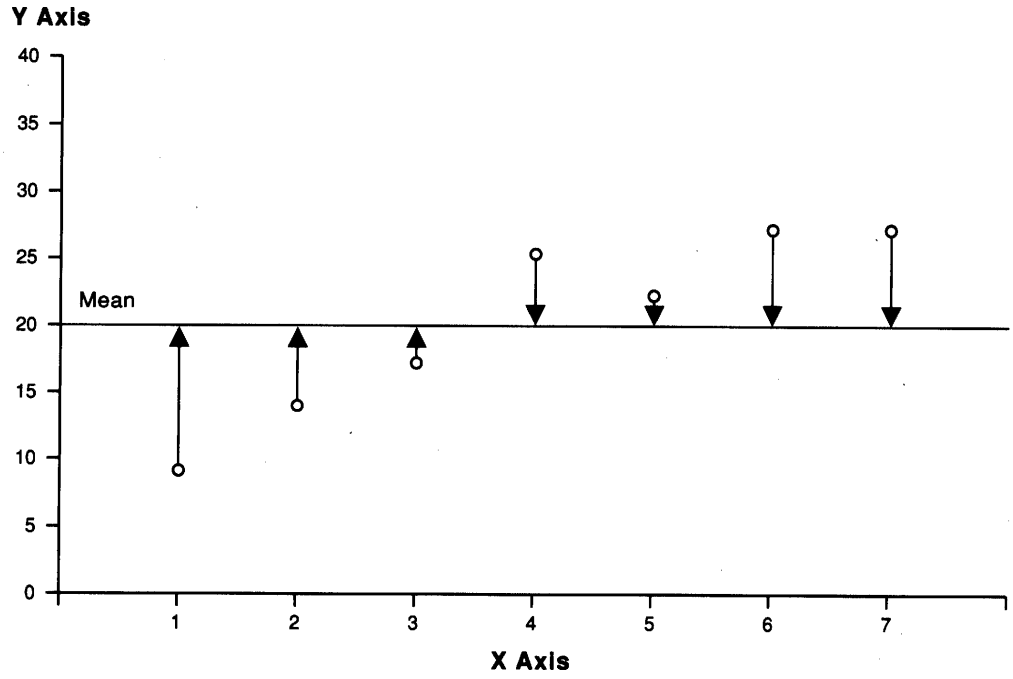
For the simple model of $Y = \boldsymbol{a}$, we must consider which value of $\boldsymbol{a}$ to choose. A sensible approach would be to choose $\boldsymbol{a}$ such that the variability of the data around $\boldsymbol{a}$ is a minimum. A useful measure of this variability is the sums of squares of the differences. Statisticians call these differences **residuals** (they are also called deviations). The residual is defined as the difference between each observed value, Y_i, and the value fitted or

[1] If the slope is zero, then the intercept will be the mean.

[2] This handbook skims over technical statistical details in order to emphasize the deterministic and logical aspects of modelling. The stochastic component of modelling (i.e., the distribution of and assumptions about the residuals) and the foundation it provides for statistical testing of hypotheses are largely ignored. There are many books available which discuss this topic in depth. See the bibliography for some commonly used textbooks.

predicted by the model, $\hat{Y}_i$. Least squares (LS or OLS) fitting methods are used to find those values of the model parameters that will minimize this sum of squares. For the model of $Y = \boldsymbol{a}$, there is only one parameter, namely ***a***.

For this simple model the value of ***a*** that has the smallest sum of squared residuals (SSR) is the mean or average[3]. In this case, the average of the *Y*-values is 20, and hence the simplest model is $Y = 20$ or $\boldsymbol{a} = 20$. This means that $\hat{Y}_i = 20$ regardless of the value of *X* or any other variable we could imagine which might be related to *Y*. The residuals from the mean are shown below.



The sums of squares of the residuals, SSR, is calculated by SSR = $\Sigma(Y_i - \hat{Y}_i)^2$. In this case, the predicted values are the mean, $\overline{Y}$, so that SSR = SSRM[4] = $\Sigma(Y_i - \overline{Y})^2$. The calculations required to obtain this sum are shown below:

X	Observed values (Y_i)	Fitted values ($\hat{Y}_i$)	Residuals ($Y_i - \hat{Y}_i$)	Squared residuals
1	9	20	–11	121
2	14	20	–6	36
3	17	20	–3	9
4	25	20	5	25
5	21	20	1	1
6	27	20	7	49
7	27	20	7	49
Sum:	140	140	0	290 = SSRM
			Degrees of freedom: $df = 7 - 1 = 6$	

Thus the sum of the squared residuals is 290. Note that the sum of the residuals is generally zero and so is of little use in choosing between models.

[3] This is shown in Appendix 2.
[4] Note that SSRM is not standard notation. The equivalence of my notation with standard notation is presented in Appendix 1.

While the simplest model, the mean, is of little interest in itself, it is useful to compare it to other models. The residual sums of squares from any other linear model will be less than, or at most, equal to the residual sums of squares from the mean. Thus, on an intuitive basis, we can say that another model is better than the mean when the residual sums of squares is sufficiently smaller than that for the mean. When certain assumptions[5] about the residuals are made, the usual statistical tests (*F*-tests) can be used to test whether the reduction in the residual sums of squares is significant. These tests will help us decide how simple our model can be, while still providing a good fit to the data.

3 NOTATION

In general, the residual sums of squares will be denoted by SSR. Letters will be added to the end to indicate the residual sums of squares for a particular type of model. For example, SSRM will denote the SSR calculated from the mean of all the data. This mean will be referred to as the grand mean, when necessary, to differentiate it from group means of subsets of the data. In the next section we will define SSRL, the residual sums of squares from a straight line model known as simple linear regression. Appendix 1 summarizes the SSR notation used throughout this handbook.

Summation notation will be used as demonstrated by the following examples.

Example 1: $\Sigma(Y_i - \overline{Y})^2 = (Y_1 - \overline{Y})^2 + (Y_2 - \overline{Y})^2 + (Y_3 - \overline{Y})^2 + \ldots$

Example 2: $\Sigma(Y_{ij} - \overline{Y})^2 = (Y_{11} - \overline{Y})^2 + (Y_{12} - \overline{Y})^2 + (Y_{13} - \overline{Y})^2 + \ldots$

$+ (Y_{21} - \overline{Y})^2 + (Y_{22} - \overline{Y})^2 + (Y_{23} - \overline{Y})^2 + \ldots$

$+ (Y_{31} - \overline{Y})^2 + (Y_{32} - \overline{Y})^2 + (Y_{33} - \overline{Y})^2 + \ldots$

$+ \ldots$

You will note that complete summation occurs over each index in the expression. In the first case, only index *i* is present so that only the different $(Y_i - \overline{Y})^2$ values are summed. In the second case, the indices *i* and *j* occur so that all the squared differences $(Y_{ij} - \overline{Y})^2$ are summed. The use of this notation is demonstrated by the calculations shown in following sections.

4 DEGREES OF FREEDOM

The total degrees of freedom (*df*) in a statistical analysis will be equal to *n*, the number of observations. For most familiar models or analyses, one degree of freedom is always devoted to the grand mean. Other degrees of freedom are associated with the parameters in the model, which restrict the values the data can have. The leftover or residual degrees of freedom are associated with the variability in the rest of the data. This variability is measured by the SSR.

Each SSR has an associated degree of freedom. In general, the value for the degrees of freedom is the number of observations, *n*, minus the number of parameters, *p*, estimated by the associated model (*p* must include a count for the mean or intercept).

Differences between SSR's of two different models can be calculated when one of the associated models is a simpler version of the other. The degrees of freedom for the resulting sums of squares is the difference of the degrees of freedom for the SSR's involved in the subtraction. For instance, the sums of squares produced by a regression model (SSL) can be calculated as the difference of two SSR's, namely:

$$\text{SSL} = \text{SSRM} - \text{SSRL}$$

[5] See Section 6.2.

For the example in Section 2, SSRM has $df = 7 - 1 = 6$ (one parameter) and SSRL has $df = 7 - 2 = 5$ (two parameters). Thus, SSL has $df = 6 - 5 = 1$. Note that any positive difference of SSR's will have a positive *df*, since the model with the greater number of parameters will have the smaller SSR with fewer *df*. Continuing the example, SSRM will be larger than SSRL and the *df* for SSRM (six) will be greater than the *df* for SSRL (five) so that the difference between SSRM and SSRL (i.e., SSL) will be positive and the difference of the corresponding *df*'s will also be positive (i.e., one *df*).[6]

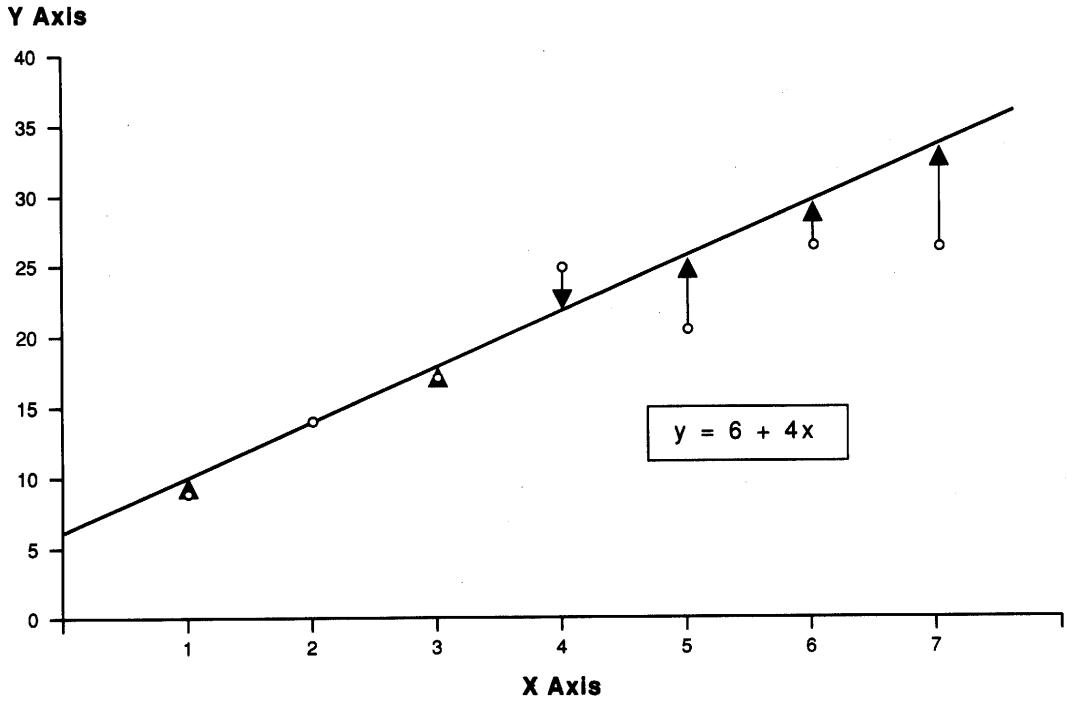
5 SIMPLE REGRESSION

Another model to consider for the example data set described in Section 2 is that of a straight line, where *Y* will increase or decrease a constant amount with a constant change in *X*. This simple regression model has the general mathematical form:

$$Y = \mathbf{a} + \mathbf{b}X$$

where **a** and **b** are the parameters of the model. In particular, **a** is known as the intercept, the value *Y* has when $X = 0$, and **b** is the slope[7] of the line. The slope measures the increase (or decrease if **b** is negative) in *Y* when *X* increases by one unit.

Any line could be drawn and the corresponding SSR calculated. For example, a ruler could be placed on a graph of the data and a line drawn that appeared to provide a good fit. One such line could be:



This line has the equation: $Y = 6 + 4X$. The magnitudes of the residuals are shown by the length of the arrows and the SSR is calculated by $\Sigma(Y_i - \hat{Y}_i)^2$ as shown on the next page.

---

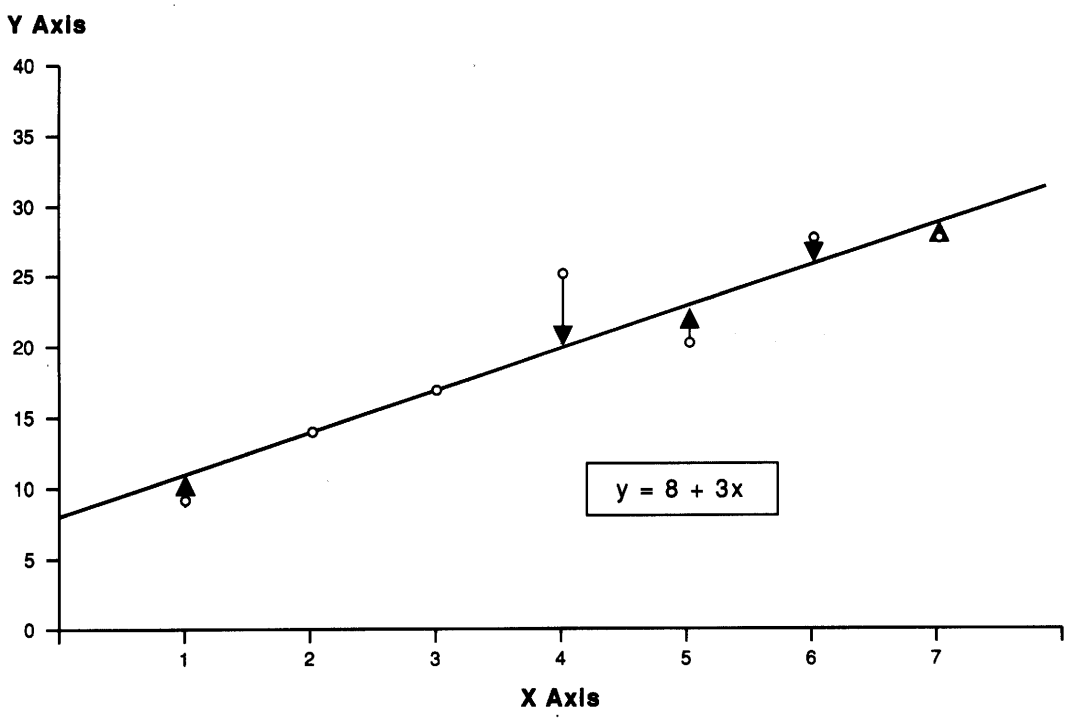
[6] See Biometrics Information Pamphlet No. 21, "What are Degrees of Freedom", for another discussion of this topic. Available from the B.C. Ministry of Forest, Forest Science Research Branch, Biometrics Section, Victoria, B.C.

[7] This parameter is also called a regression coefficient.

X	Observed values (Y_i)	Fitted values ($\hat{Y}_i$)	Residuals	Squared residuals
1	9	10	–1	1
2	14	14	0	0
3	17	18	–1	1
4	25	22	3	9
5	21	26	–5	25
6	27	30	–3	9
7	27	34	–7	49
Sum:	140	154	–14	94 = SSR
			Degrees of freedom: $df = 7 - 2 = 5$	

An obvious definition for a line that best fits the data is to choose that line which has the smallest SSR. The hard way to find this line is to try different values for **a** and **b** until the smallest SSR is found. Mathematicians have solved this problem algebraically[8] so that any computing package can automatically find the values of **a** and **b** for which the corresponding SSR is the minimum. This method is easily generalized to all linear models and is known as fitting by Least Squares (LS) or Ordinary Least Squares (OLS).

The best fit line for the example data set is $Y = 8 + 3X$. This is shown below:



[8] See Appendix 2 for a description of the algebraic solution.

The SSR is denoted by SSRL and is calculated by SSRL = $\Sigma(Y_i - \hat{Y}_i)^2$ as shown by:

X	Observed values (Y_i)	Fitted values ($\hat{Y}_i$)	Residuals ($Y_i - \hat{Y}_i$)	Squared residuals
1	9	11	–2	4
2	14	14	0	0
3	17	17	0	0
4	25	20	5	25
5	21	23	–2	4
6	27	26	1	1
7	27	29	–2	4
Sum:	140	140	0	38 = SSRL
			Degrees of freedom: $df = 7 - 2 = 5$	

You will note that the sum of the fitted values ($\Sigma\ \hat{Y}_i = 140$) is the same as the sum of the observed values and that the sum of the residuals is zero[9]. The residual sums of squares (38) is less than that for the line $Y = 6 + 4X$. In fact, it is the least possible value for any straight line that we might draw for this data. Also note that this value, denoted by SSRL = 38, is smaller than that for the mean, SSRM = 290.

6 CRITERIA FOR COMPARING MODELS

When several models are fit to a set of data, some criteria are needed to help decide which model is best. The definition of best will always be somewhat subjective and will depend on why the models are being fit, that is, the objective of the study. Nevertheless, there are some generally useful measures to use when selecting a model. Those discussed here rely on the Extra Sums of Squares Principle.

6.1 The Extra Sums of Squares Principle

The best model imaginable for any data set would have an SSR of zero. But the model should also be simple with a small number of parameters. These tend to be mutually exclusive criteria since complex models (i.e., those with more parameters) have smaller SSR's than the simple ones. Hence we must choose a model that provides a reasonable compromise.

When a new variable is added to a model we can ask: how much more does this reduce the SSR? This observed difference in SSR is the extra sums of squares due to that variable or term. If this difference is small, then we might decide not to add that variable or term to the model. Alternatively, we could start with a large model containing many variables or terms and examine the change in sums of squares when one of them is removed. This difference in residual sums of squares is also an extra sums of squares. With either approach, the actual value of the extra sums of squares depends on which other terms are also in the model. An extra sums of squares can be calculated and tested for a group of variables or terms in a model. The statistical significance of an extra sums of squares is often determined by an *F*-test.

6.2 The *F*-Test

The extra sums of squares due to a term will be small if the model with that term does not offer a marked improvement over that provided by the model without that term. For example, if the linear regression model is

[9] The sum of the residuals is always zero for the best fitting linear model so long as the model contains a constant or intercept. See Appendix 2 for a proof.

not a marked improvement from the mean, then we would expect the slope to be zero and the difference of the two SSR's, SSL = SSRM − SSRL, to be small (but not zero!). This can be tested by:

$$F = \frac{(\text{SSRM} - \text{SSRL})/1}{\text{SSRL}/(n-2)} \text{ with } df = 1,\ n-2$$

For the example this is:

$$F = \frac{(290 - 38)/1}{38/(7-2)} = \frac{252}{7.6} = 33.16 \text{ with } df = 1,\ 5,\ prob = 0.0022$$

The *F*-test is a ratio of two sums of squares (SS) each divided by their *df*. An SS divided by its *df* is called a mean square (MS). Each *F*-test has two degrees of freedom, one for the numerator and one for the denominator. Their values are the *df*'s associated with the numerator and denominator SS's used in the *F*-test. The ratio of two MS's will follow an *F*-distribution given that:

1. the model chosen is essentially correct; and
2. the residuals from the linear model are independent and normally distributed with constant variance.[10]

If the more complex model being tested is no better at fitting the data than the simpler model, then the observed *F*-value follows the usual *F*-distribution (known to *aficionados* as the central *F*-distribution). The conjecture that the complex model is no better than the simple model, or that one or more parameters of the complex model are zero, is known as the **null hypothesis** (denoted by H_0). When the null hypothesis is reasonable, then the two mean squares in the *F*-ratio are expected to have the same value so that the *F*-value should be about one. On the other hand, if the null hypothesis is not true, the *F*-ratio is expected to be substantially larger than one.

The statistical significance of an observed *F*-value is determined by the probability of obtaining *F*-values exceeding the actual value observed. (This probability can be determined from *F*-tables or calculated by computer.) The smaller the probability or *p*-value (i.e., the larger the *F*-value), the stronger the evidence against the null hypothesis. For the above example, the *p*-value of 0.0022 is quite small if the null hypothesis is true. In other words, only 22 out of 10,000 times would an *F*-value as large as 33.16 be observed if the regression did not, in fact, provide a better overall fit than the grand mean. Thus, we might conclude that the regression is the better model to use for fitting the data.

6.3 Coefficient of Determination, or R^2.

A useful measure of how well a model explains the variability of the data is the R^2 value. This value is calculated by:

$$R^2 = \frac{\text{SSRM} - \text{SSR}}{\text{SSRM}} = \frac{290 - 38}{290} = 0.869 \text{ for the example}$$

This can have any value between 0 and 1, and is the proportion of the total SS (SSRM) that is explained by the model. If the value is zero, then the model does not explain any of the variability in the data. If it is one,

---

[10] This statement is not strictly correct from a statistical point of view. Residuals have observed and theoretical values just as the parameters *a* and *b* in simple regression have. It is the theoretical values that must be independent and normally distributed with constant variance. The observed residuals are not independent, for instance, because they must sum to zero.

then it explains or accounts for all the variability in the data. Obviously, a model is fitting better overall if its R^2 is close to one. For the simple regression example above, we can test if H_0: $R^2 = 0$ by:

$$F = \frac{R^2/1}{(1 - R^2)/(n-2)},\ df = 1,\ n-2$$

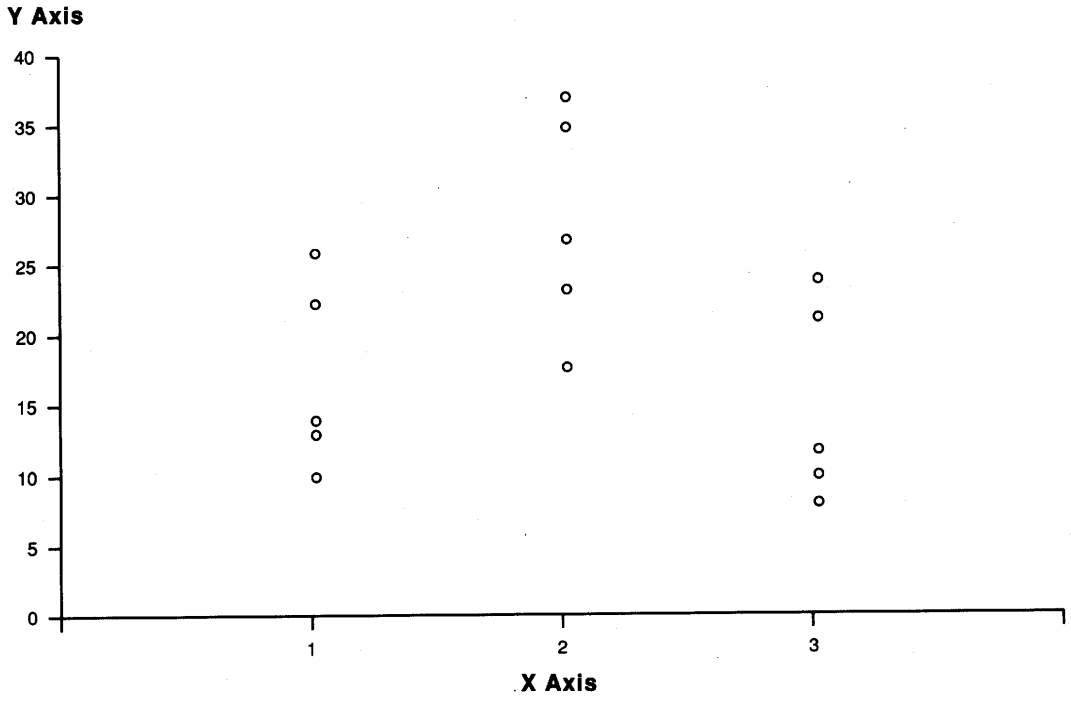
$$F = \frac{0.869/1}{(1 - 0.869)/(5)} = 33.16,\ df = 1,\ 5,\ \text{for the example}$$

In general, this tests whether the whole model is better than the grand mean. For the simple regression case, the *F*-tests described in this section and Section 6.2 are identical.[11]

7 ONE-WAY ANALYSIS OF VARIANCE (ANOVA)

Suppose that there were several groups of numbers instead of individual pairs of data values. These groups may be identified by having been assigned different treatments. Consider the following set of data:

Group	*Y*	Mean
1	13, 14, 10, 26, 22	$\overline{Y}_1 = 17$
2	18, 23, 27, 35, 37	$\overline{Y}_2 = 28$
3	8, 10, 12, 24, 21	$\overline{Y}_3 = 15$
	Grand mean:	$\overline{Y} = 20$

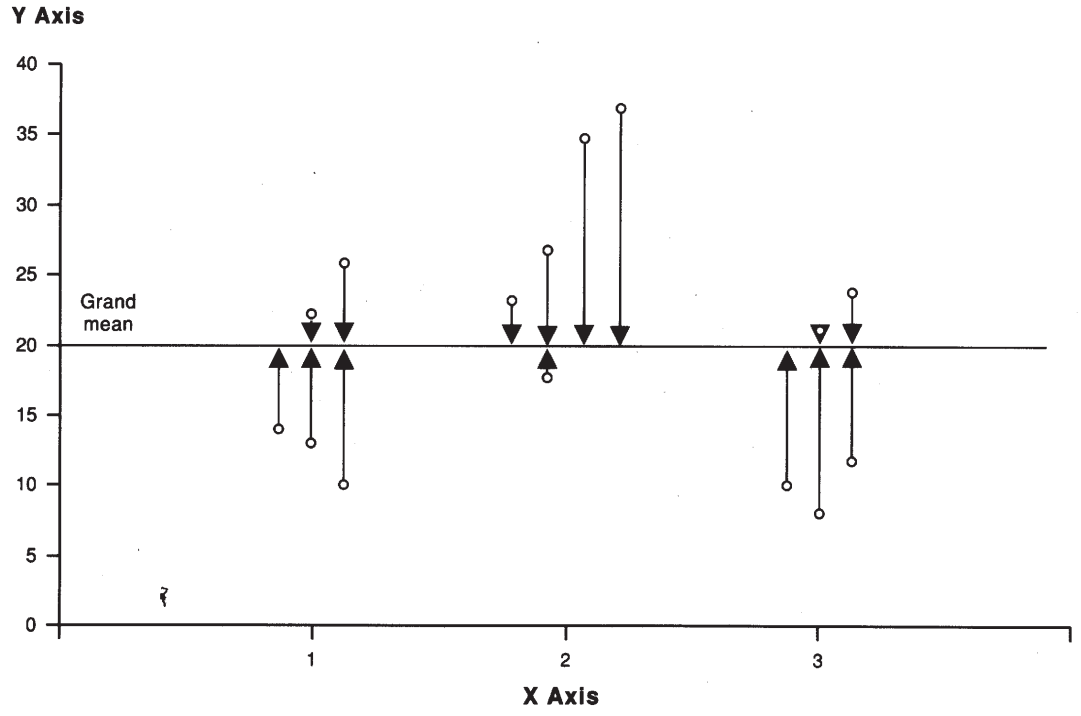


[11] See Biometrics Information Pamphlets Nos. 18 ''Multiple Regression: Selecting the Best Subset,'' and 27, ''When the *t*-test and the *F*-test are equivalent'', for more information. Available from B.C. Ministry of Forests, Forest Science Research Branch, Biometrics Section, Victoria, B.C.

First the grand mean is fit to the data the SSRM is calculated. The sum of squared residuals, SSRM = $\Sigma(Y_{ij} - \hat{Y}_{ij})^2 = \Sigma(Y_{ij} - \bar{Y})^2$, is calculated by:

X	Observed values (Y_{ij})	Fitted values ($\hat{Y}_j$)	Residuals ($Y_{ij} - \hat{Y}_{ij}$)	Squared residuals
1	13	20	–7	49
	14	20	–6	36
	10	20	–10	100
	26	20	6	36
	22	20	2	4
2	18	20	–2	4
	23	20	3	9
	27	20	7	49
	35	20	15	225
	37	20	17	289
3	8	20	–12	144
	10	20	–10	100
	12	20	–8	64
	24	20	4	16
	21	20	1	1
Sum:	300	300	0	1,126 = SSRM
			Degrees of freedom: df = 15 –1 = 14	

This model with its residuals is shown below:

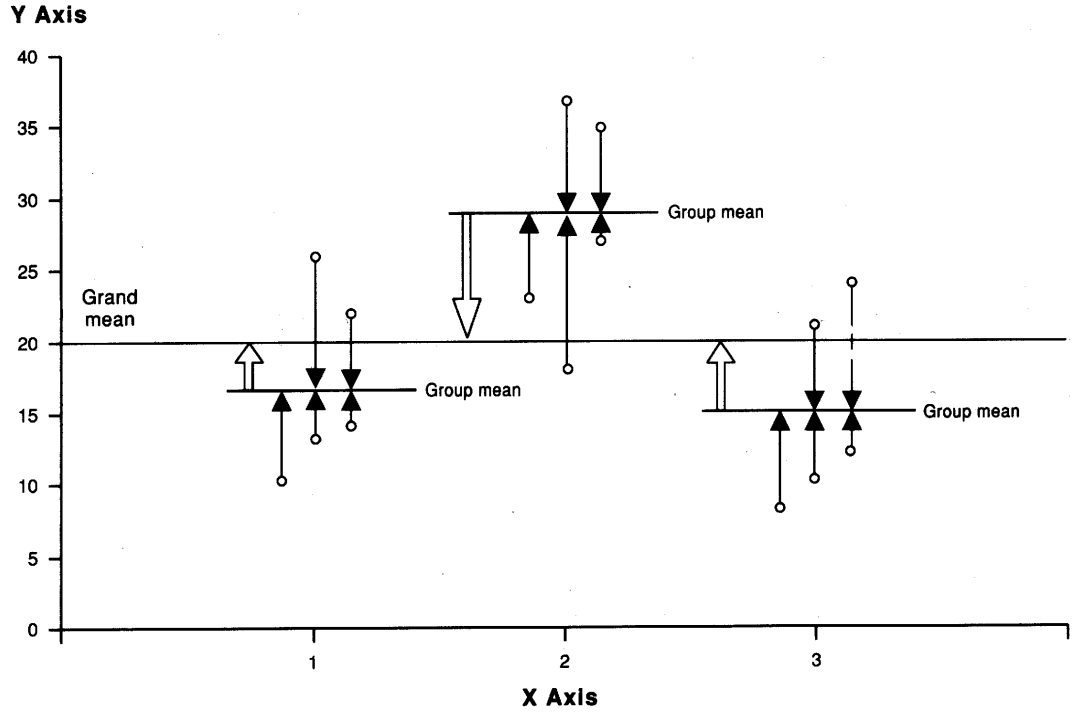


The second model to fit predicts three different means, one for each group. The means are estimated by the observed group means, since they provide the smallest SSR. This SSR is the sum of each within-group residual sums of squares (SSRWG). These sums of squares are calculated by SSRWG = $\Sigma(Y_{ij} - \hat{Y}_{ij})^2 = \Sigma(Y_{ij} - \bar{Y}_i)^2$ as shown below:

Group	Observed values (Y_{ij})	Fitted values ($\hat{Y}_{ij}$)	Residuals ($Y_{ij} - \hat{Y}_{ij}$)	Squared residuals
1	13	17	–4	16
	14	17	–3	9
	10	17	–7	49
	26	17	9	81
	22	17	5	25
2	18	28	–10	100
	23	28	–5	25
	27	28	–1	1
	35	28	7	49
	37	28	9	81
3	8	15	–7	49
	10	15	–5	25
	12	15	–3	9
	24	15	9	81
	21	15	6	36
Sum:	300	300	0	636 = SSRWG

Degrees of freedom: $df = 15 - 3 = 12$

The residuals from this model (thin arrows) can be pictured as:



Notice that the residual sums of squares has been reduced from 1126 to 636. Whether this reduction in sums of squares is large enough to reject the grand mean as an adequate model is tested by:

$$F = \frac{(\text{SSRM} - \text{SSRWG})/(g-1)}{\text{SSRWG}/(n-g)}$$, where n is the number of observations and g is the number of groups

$$F = \frac{(1126 - 636)/(3 - 1)}{636/(4 + 4 + 4)} = \frac{490/2}{636/12} = \frac{245}{53}$$

$F = 4.62,\ df = 2,\ 12,\ prob = 0.033$

In this case, there is good reason to believe that the group means model is better than the grand mean model.

The difference SSRM – SSRWG is the sums of squares between groups (SSRBG)[12] and can also be calculated by SSRBG = $5\ \Sigma(\overline{Y}_i - \hat{Y}_i)^2 = 5\Sigma(\overline{Y}_i - \overline{Y})^2$ as shown by:

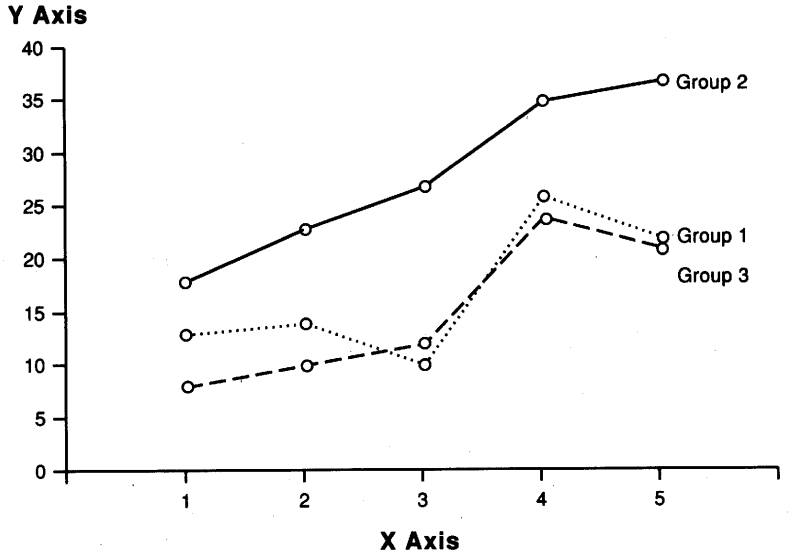
Group	Group means ($\overline{Y}_i$)	Grand mean ($\overline{Y}$)	Residuals ($\overline{Y}_i - \overline{Y}$)	Squared residuals
1	17	20	–3	9
	17	20	–3	9
	17	20	–3	9
	17	20	–3	9
	17	20	–3	9
2	28	20	8	64
	28	20	8	64
	28	20	8	64
	28	20	8	64
	28	20	8	64
3	15	20	–5	25
	15	20	–5	25
	15	20	–5	25
	15	20	–5	25
	15	20	–5	25
Sum:	300	300	0	490 = SSRBG
			Degrees of freedom: $df = 3 - 1 = 2$	

The above calculations show that the SSRBG is the sums of squared residuals of three means about their grand mean, weighted by the sample size or number of observations per mean (five in this case). These residuals were shown by the thick arrows in the previous picture.

[12] Note that this is the extra sums of squares due to adding group means to the model. The SSRBG is usually called the sums of squares between (SSB). See Appendix 1.

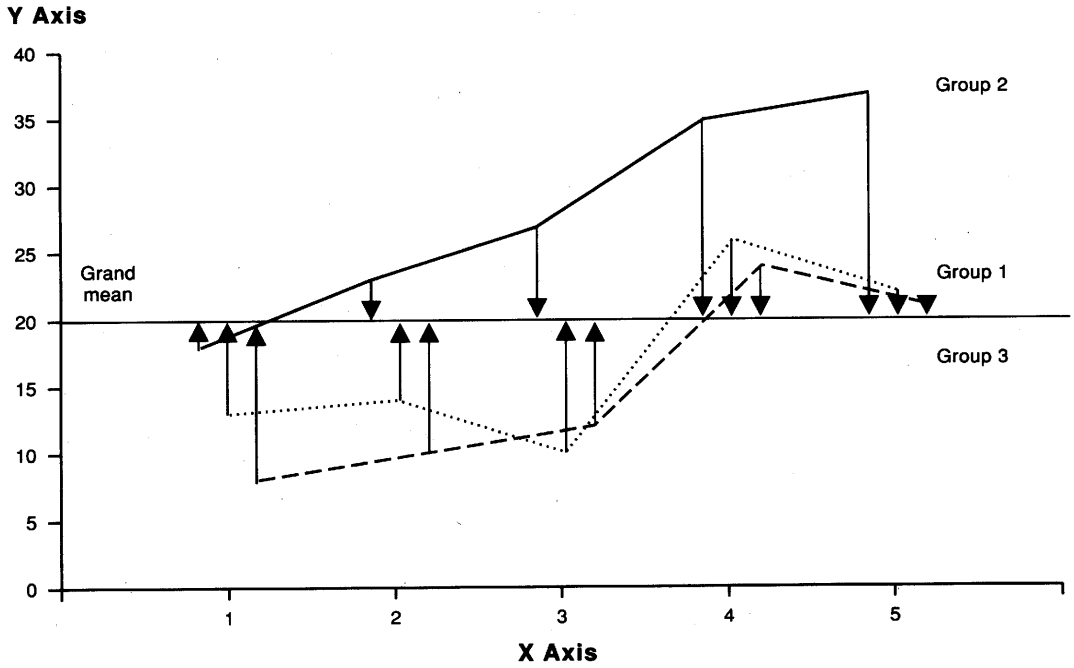
8 ONE-WAY ANALYSIS OF COVARIANCE (ANCOVA)

If each *Y*-value in the previous example had an associated *X*-value that is linearly related to the *Y*, then some of the variability within each group would be explained by its associated *X*-value. We could redraw the data as:



	Group		
X	1 Y	2 Y	3 Y
1	13	18	8
2	14	23	10
3	10	27	12
4	26	35	24
5	22	37	21
$\overline{Y}_i$:	17	28	15

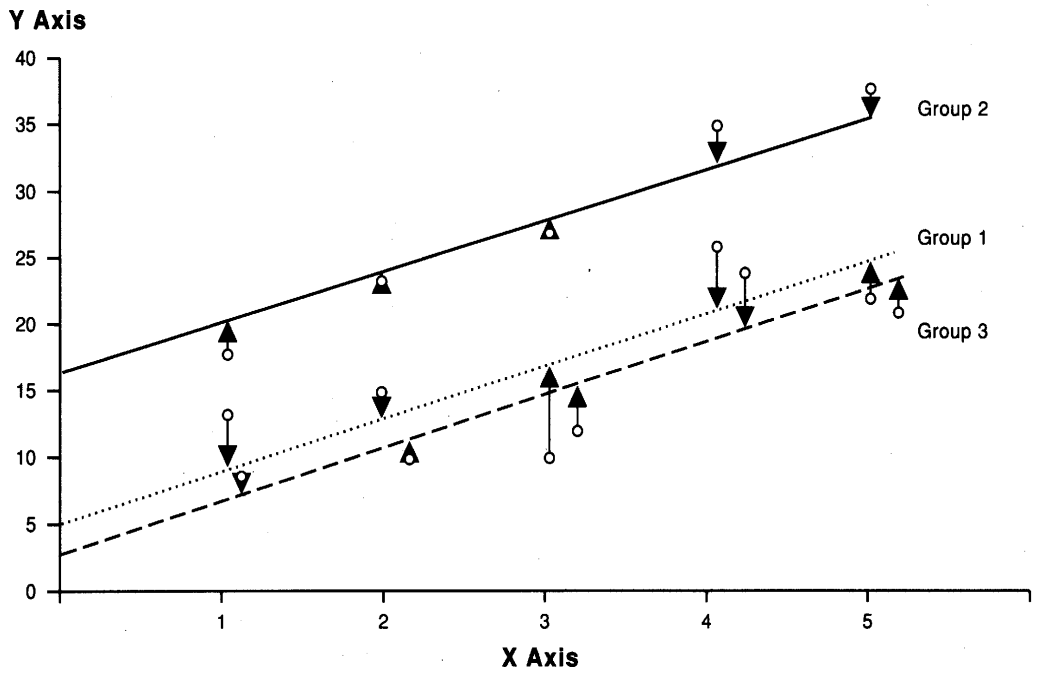
First the grand mean is fitted. The calculations for the residual sums of squares (SSRM) has been done previously. The fit of the grand mean is shown below:



Lines are fitted to the three groups **forcing** them to have the same slope, but allowing them different intercepts. For this example, the fitted equations are:

Group	Equation
1	$Y = 5 + 4X$
2	$Y = 16 + 4X$
3	$Y = 3 + 4X$

The analysis of covariance model looks like:



The calculations for the new residual sums of squares SSR3LP = $\Sigma(Y_{ij} - \hat{Y}_{ij})^2$ are shown below:

Group	Observed values (Y_{ij})	Fitted values ($\hat{Y}_{ij}$)	Residuals ($Y_{ij} - \hat{Y}_{ij}$)	Squared residuals
1	13	9	4	16
	14	13	1	1
	10	17	–7	49
	26	21	5	25
	22	25	–3	9
2	18	20	–2	4
	23	24	–1	1
	27	28	–1	1
	35	32	3	9
	37	36	1	1
3	8	7	1	1
	10	11	–1	1
	12	15	–3	9
	24	19	5	25
	21	23	–2	4
Sum:	300	300	0	156 = SSR3LP

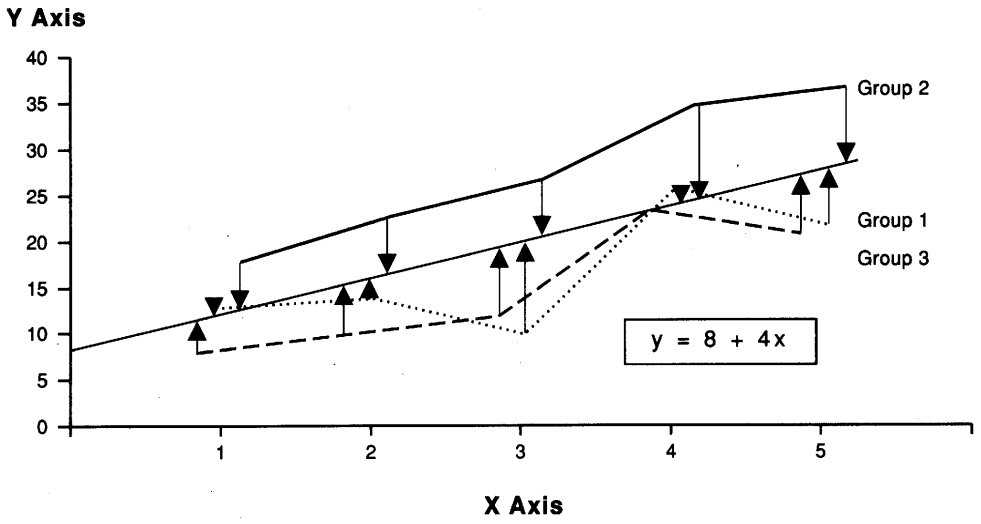
Degrees of freedom: $df = 15 - 4 = 11$

The *F*-test is:

$$F = \frac{(636 - 156)/(12 - 11)}{156/11} = 33.85,\ df = 1,\ 11\ prob = 0.00012$$

This tests whether *X* is worth having in the model. For this example, *X* has substantially reduced the residual sums of squares and so is worth having in the model.

To test for group differences, one line must be fitted through all the data. This fitted line is $Y = 8 + 4X$ and is pictured by:



The SSR are calculated by SSRL = $\Sigma(Y_{ij} - \hat{Y}_{ij})^2$ as shown below:

Group	Observed values (Y_{ij})	Fitted values ($\hat{Y}_{ij}$)	Residuals ($Y_{ij} - \hat{Y}_{ij}$)	Squared residuals
1	13	12	1	1
	14	16	–2	4
	10	20	–10	100
	26	24	2	4
	22	28	–6	36
2	18	12	6	36
	23	16	7	49
	27	20	7	49
	35	24	11	121
	37	28	9	81
3	8	12	–4	16
	10	16	–6	36
	12	20	–8	64
	24	24	0	0
	21	28	–7	49
Sum:	300	300	0	646 = SSRL

Degrees of freedom: $df = 15 - 2 = 13$

The test comparing three parallel lines with only one line is:

$$F = \frac{(646 - 156)/(13 - 11)}{156/11} = \frac{490/2}{156/11} = 17.28,\ df = 2,\ 11,\ prob = 0.0004$$

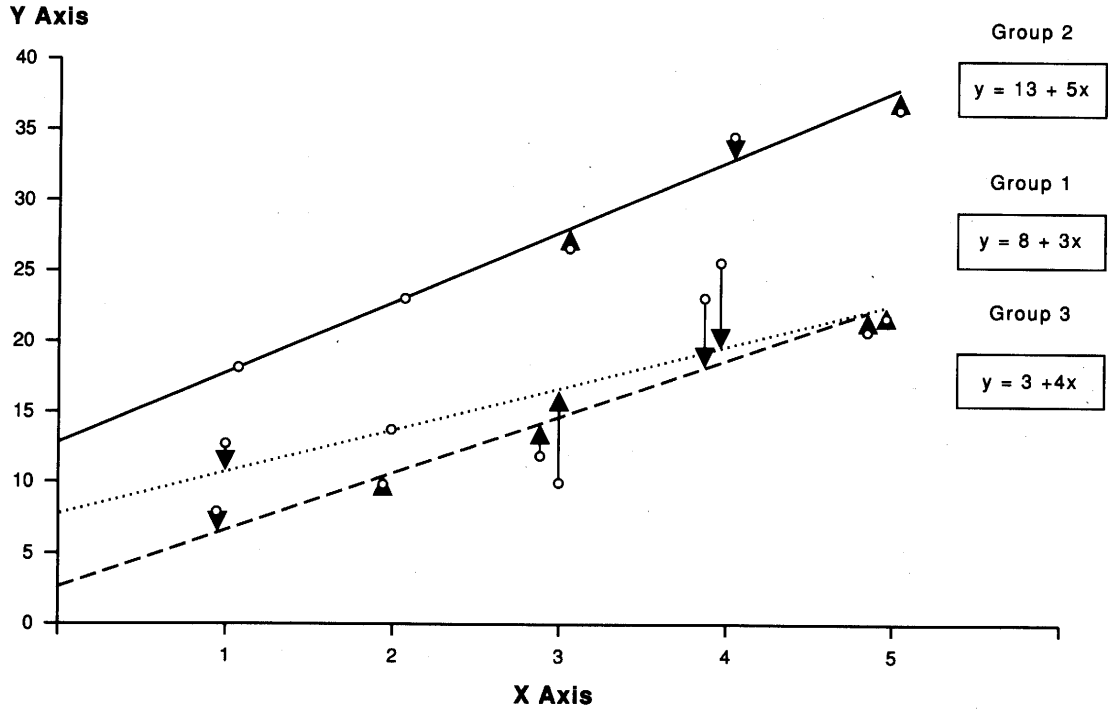
This is, in fact, a test for treatment differences where the treatment effect is to shift the line (i.e., relationship between *Y* and *X*) up or down. This shifting is accomplished with each line being given a different intercept. For the example, the model with three lines provides a much better fit than does the model with only one line.

9 HETEROGENEITY OF REGRESSION, OR COMPARING REGRESSION LINES

Analysis of covariance assumes that the regression lines are parallel, meaning that the slope, ***b***, is the same for each line. This is an assumption that can be tested. The testing procedure involved can also be used to test for specific differences between slopes. For instance, instead of assuming parallel slopes, we may have a specific interest in testing treatment effects on the slopes. In either case, regression lines are fitted to the three groups, allowing them to have separate parameter estimates. This is the same as fitting each line separately. The fitted equations for the example are:

Group	Equation
1	$Y = 8 + 3X$
2	$Y = 13 + 5X$
3	$Y = 3 + 4X$

This new model can be pictured as:



The new residuals are calculated by:

Group	Observed values (Y_{ij})	Fitted values ($\hat{Y}_{ij}$)	Residuals ($Y_{ij} - \hat{Y}_{ij}$)	Squared residuals
1	13	11	2	4
	14	14	0	0
	10	17	–7	49
	26	20	6	36
	22	23	–1	1
2	18	18	0	0
	23	23	0	0
	27	28	–1	1
	35	33	2	4
	37	38	–1	1
3	8	7	1	1
	10	11	–1	1
	12	15	–3	9
	24	19	5	25
	21	23	–2	4
Sum:	300	300	0	136 = SSR3L
			Degrees of freedom: $df = 15 - 6 = 9$	

The test for parallel lines or homogeneity of slopes is:

$$F = \frac{(156 - 136)/(11 - 9)}{136/9} = \frac{20/2}{136/9} = 0.66,\ df = 2, 9,\ prob = 0.54$$

Clearly, in this case, there is little reason to believe that the three lines have different slopes.

10 ANOVA TABLES FOR MODELS FITTED

The previous sections have described various models and how sums of squares (SS's) for various situations are calculated. This section will show how these SS's are used to conduct various tests about the example data. The calculation of the SS's do not require any assumptions about the data, but the tests do. The assumptions required for the tests are discussed in many textbooks (see Bibliography).

10.1 Simple Regression

The residual sums of squares (SSR) that were calculated for the simple regression example (Section 5) are:

Model	Number of parameters	*df*	SSR name	SSR value	Difference	Proportion of SSRM (R^2)
1. Grand Mean	1	6	SSRM	290		1.00
					252	0.87
2. Line: $Y = 8 + 3X$	2	5	SSRL	38		0.13

These SS's are used to calculate the following ANOVA table:

Source of variation	Sums of squares notation	*df*	Sums of squares	Mean square	*F*-value	*p*-value
Regression	SSL = SSRM − SSRL	1	252	252	33.16	0.0022
Error	SSRL	5	38	7.6		
Total	SSRM	6	290			

These results indicate there is strong evidence against the null hypothesis that the slope is zero.

10.2 One-Way ANOVA

The residual SS's calculated for the simple one-way ANOVA (Section 7) are:

Model	Number of parameters	*df*	SSR name	SSR value	Difference	Proportion of SSRM (R^2)
1. Grand Mean	1	14	SSRM	1126		1.00
					490	0.44
2. Group Means	3	12	SSRWG	636		0.56

These SS's are used to calculate the following ANOVA table.

Source of variation	Sums of squares notation	*df*	Sums of squares	Mean square	*F*-value	*p*-value
Between Groups	SSRBG = SSRM − SSRWG	2	490	245	4.62	0.033
Error (Within Groups)	SSRWG	12	636	53		
Total	SSRM	14	1126			

These results provide strong evidence against the null hypothesis. In other words, the group means model provides a better fit to the data than does the grand mean model.

10.3 Analysis of Covariance, or Comparing Regression Lines

The residual SS's for the ANCOVA (Section 8) and comparing regression lines (Section 9) are:

Model	Number of parameters	*df*	SSR name	SSR value	Difference	Proportion of SSRM (R^2)
1. Grand Mean	1	14	SSRM	1126		1.00
					480[a]	0.43
2. One Line	2	13	SSRL	646		0.57
					not meaningful	—
3. Group Means	3	12	SSRWG	636		0.56
					480[a]	0.43
4. Three Lines (same slope)	4	11	SSR3LP	156		0.14
					20	0.02
5. Three Lines (different slopes)	6	9	SSR3L	136		0.12

[a] These two SS's do not usually have the same value. They do here because all three groups have exactly the same *X*-values.

These are used to form an ANCOVA table as follows:

Source of variation	Sums of squares notation	*df*	Sums of squares	Mean square	*F*-value	*p*-value
Between Groups	SSRBG = [a] SSRL − SSR3LP	2	490	245	17.28	0.0004
Covariate	SSCov = [a] SSRWG − SSR3LP	1	480	480	33.85	0.0001
Error	SSR3LP	11	156	14.2		
Total	SSRM	14	1126			

[a] Note that this is the Type III Sums of Squares output by SAS.

This ANCOVA table assumes that the lines for each group have the same slope. This assumption can be tested using a heterogeneity of regression test calculated as follows:

$$F = \frac{(\text{SSR3LP} - \text{SSR3L})/(11 - 9)}{\text{SSR3L}/9} = \frac{20/2}{136/9} = 0.66,\ df = 2, 9,\ p = 0.54$$

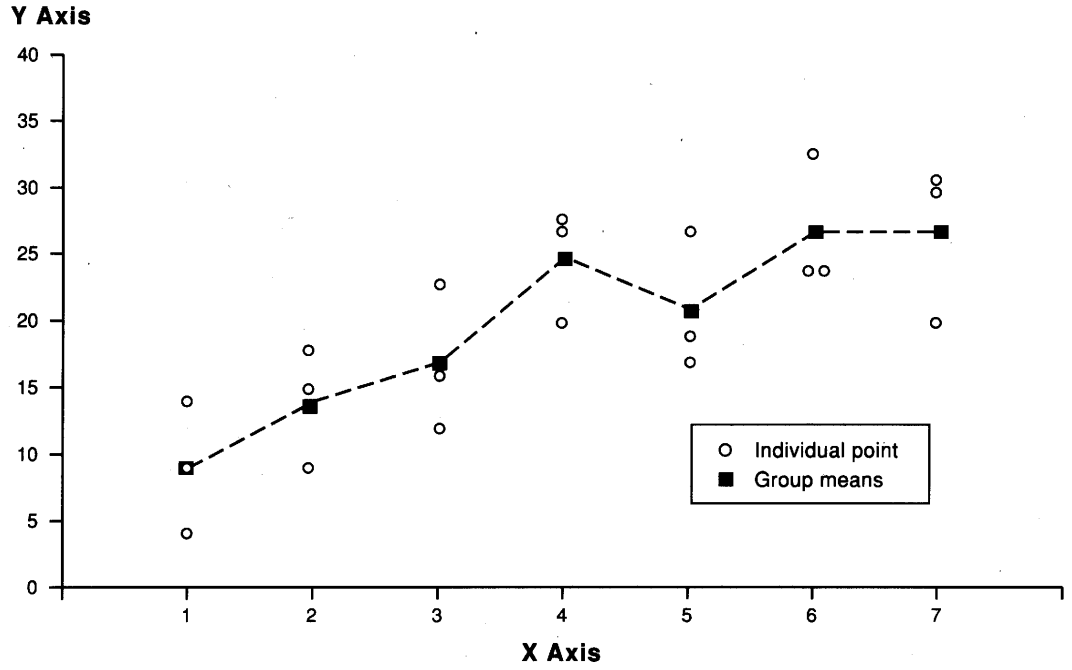
In this case, the assumption of constant slope is tenable.

11 SIMPLE REGRESSION WITH MULTIPLE OBSERVATIONS

Frequently there are multiple observations for each *X*-value in a simple regression. A common example is an experiment where the treatment levels are increasing amounts of fertilizer or herbicide. Having multiple observations is advantageous because it is possible to test whether the linear regression model is appropriate by using a lack-of-fit test. Suppose there were multiple observations for the data set discussed in Sections 2 and 5 so that means were used in the previous calculations, determined from the following individual data:

X	Observed Values (Y_{ij})	Group Means ($\bar{Y}_i$)
1	9, 4, 14	9
2	15, 18, 9	14
3	16, 12, 23	17
4	27, 28, 20	25
5	19, 17, 27	21
6	24, 24, 33	27
7	30, 31, 20	27

These data look like:

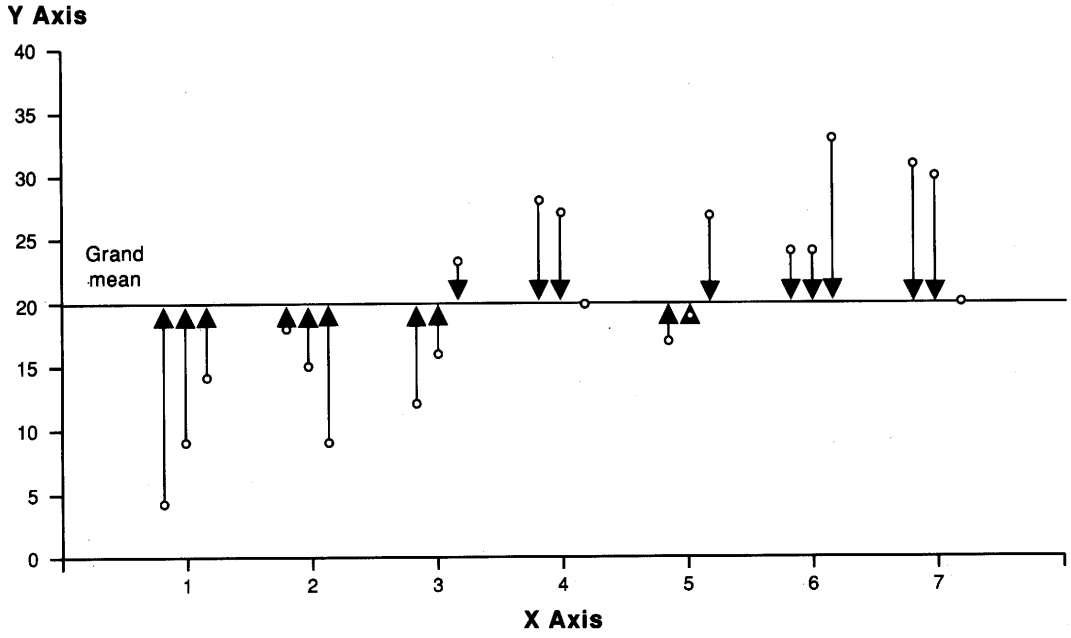


As usual, the first model fit is the grand mean, which is still 20. The SSRM = $\Sigma(Y_{ij} - \bar{Y})^2$ and is calculated as usual:

X	Observed values (Y_{ij})	Fitted values ($\hat{Y}_{ij}$)	Residuals ($Y_{ij} - \hat{Y}_{ij}$)	Squared residuals
1	9	20	–11	121
	4	20	–16	256
	14	20	–6	36
2	15	20	–5	25
	18	20	–2	4
	9	20	–11	121
3	16	20	–4	16
	12	20	–8	64
	23	20	3	9
4	27	20	7	49
	28	20	8	64
	20	20	0	0
5	19	20	–1	1
	17	20	–3	9
	27	20	7	49
6	24	20	4	16
	24	20	4	16
	33	20	13	169
7	30	20	10	100
	31	20	11	121
	20	20	0	0
Sum:	420	420	0	1,246 = SSRM

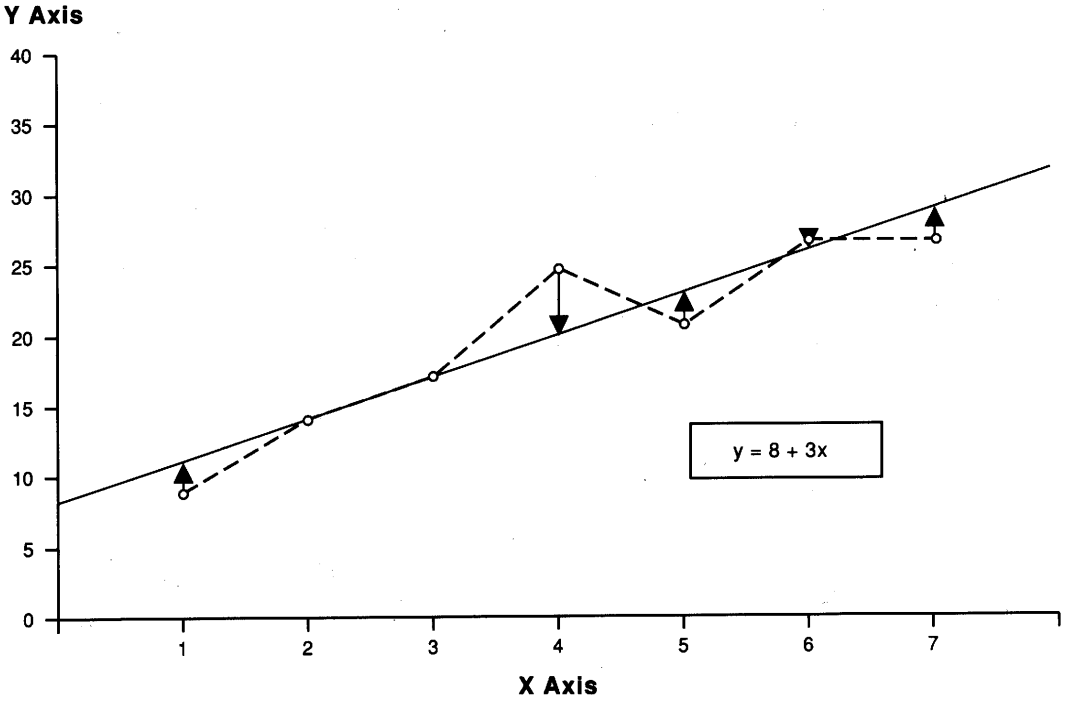
Degrees of freedom: $df = 21 - 1 = 20$

These residuals are pictured below:

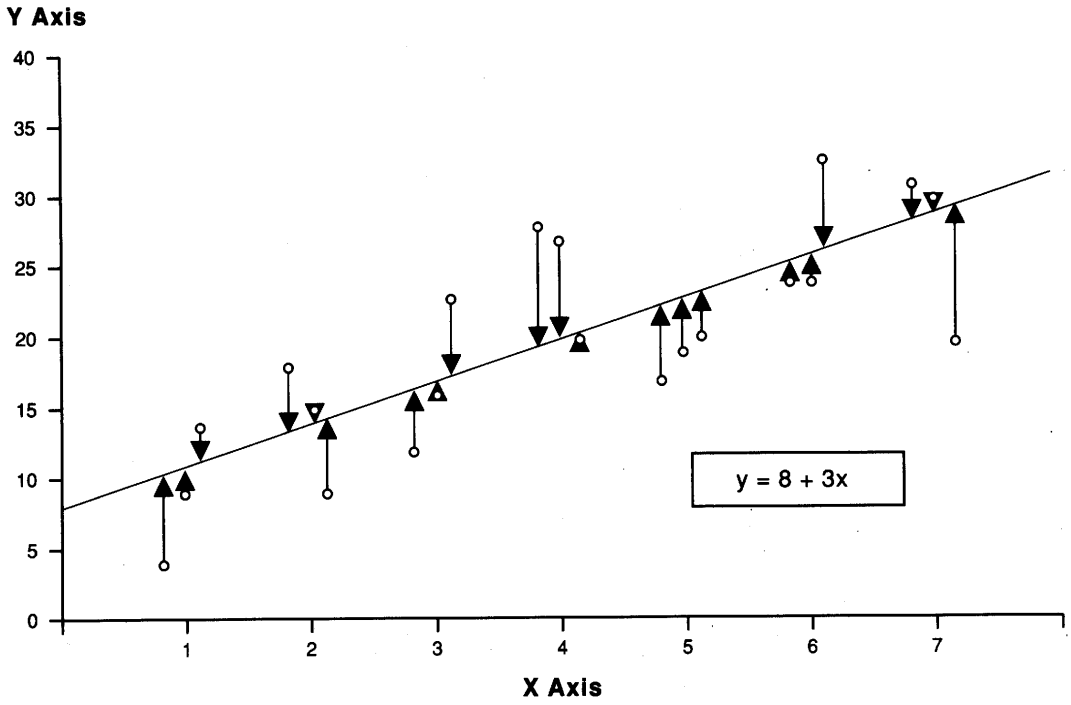


11.1 Regression Approach

The SSR for the regression on the group means (SSRLG) is calculated as before for SSRL in Section 5. In fact, the SSRLG = sample size × previous SSRL, so that SSRLG = 3(38) = 114 (still with 5 *df*). The residuals are shown below. Note that each arrow actually occurs three times, once for each individual value. This SSR measures the lack-of-fit of the linear model.



The residual sums of squares for the regression on the individual values (SSRL) is shown below:



SSRL is calculated by $\Sigma(Y_{ij} - \hat{Y}_{ij})^2$ with the following details:

X	Observed values (Y_{ij})	Fitted values ($\hat{Y}_{ij}$)	Residuals ($Y_{ij} - \hat{Y}_{ij}$)	Squared residuals
1	9	11	–2	4
	4	11	–7	49
	14	11	3	9
2	15	14	1	1
	18	14	4	16
	9	14	–5	25
3	16	17	–1	1
	12	17	–5	25
	23	17	6	36
4	27	20	7	49
	28	20	8	64
	20	20	0	0
5	19	23	–4	16
	17	23	–6	36
	27	23	4	16
6	24	26	–2	4
	24	26	–2	4
	33	26	7	49
7	30	29	1	1
	31	29	2	4
	20	29	–9	81
Sum:	420	420	0	490 = SSRL

Degrees of freedom: $df = 21 - 2 = 19$

When there are multiple observations per X-value, the ANOVA table presented in Section 10.1 for simple regression must be modified. The usual residual SS (SSRL) can be split into two sources of variation: a SS caused by lack-of-fit (SSRLG), and a pure error term (SSRWG). The usual regression model (without replication) assumes that there is no lack-of-fit. This assumption can be tested when there are multiple observations for at least some of the X-values. The SS's that we have calculated so far are:

Model	Number of parameters	*df*	SSR name	Sums of squares	Difference	Proportion of SSRM (R^2)
1. Grand Mean	1	20	SSRM	1246		1.00
					756	0.61
2. Line: $Y = 8 + 3X$	2	19	SSRL	490		0.39
(about individual data)					376	0.30
3. Line: $Y = 8 + 3X$	2	5	SSRLG	114		0.09
(about group means)						

These SS's are used to calculate the following ANOVA table:

Source of variation	Sums of squares notation	*df*	Sums of squares	Mean square	*F*-value	*p*-value
Regression	SSL = SSRM − SSRL	1	756	756	29.3	0.0001
Error	SSRL	19	490	25.8		
i) Lack of fit	SSRLG	5	114	22.8	0.85	0.54
ii) Pure error	SSRL − SSRLG	14	376	26.9		
Total	SSRM	20	1246			

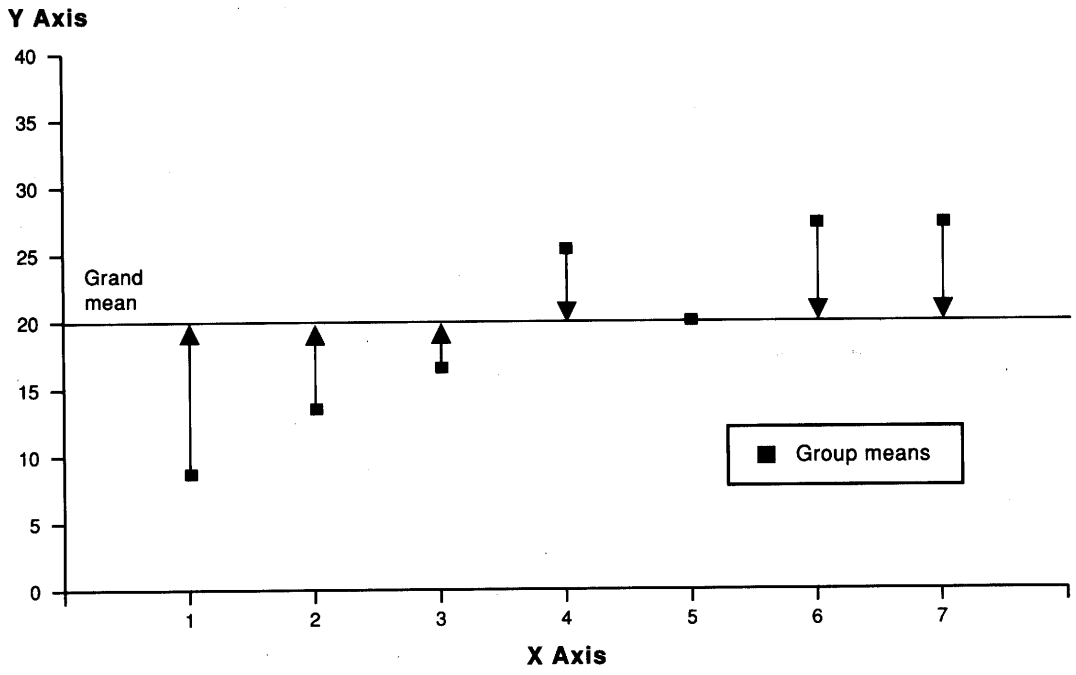
A significant lack-of-fit test suggests that the model of a straight line is inappropriate for the data. When this occurs, other models should be examined.

The above table summarizes the regression approach to this problem.[13]

[13] See Biometrics Information Pamphlet No. 28, ''Simple Regression with Replication: Testing for Lack of Fit'', for more information on the regression approach. Available from B.C. Ministry of Forests, Forest Science Research Branch, Biometrics Section, Victoria, B.C.

11.2 ANOVA Approach

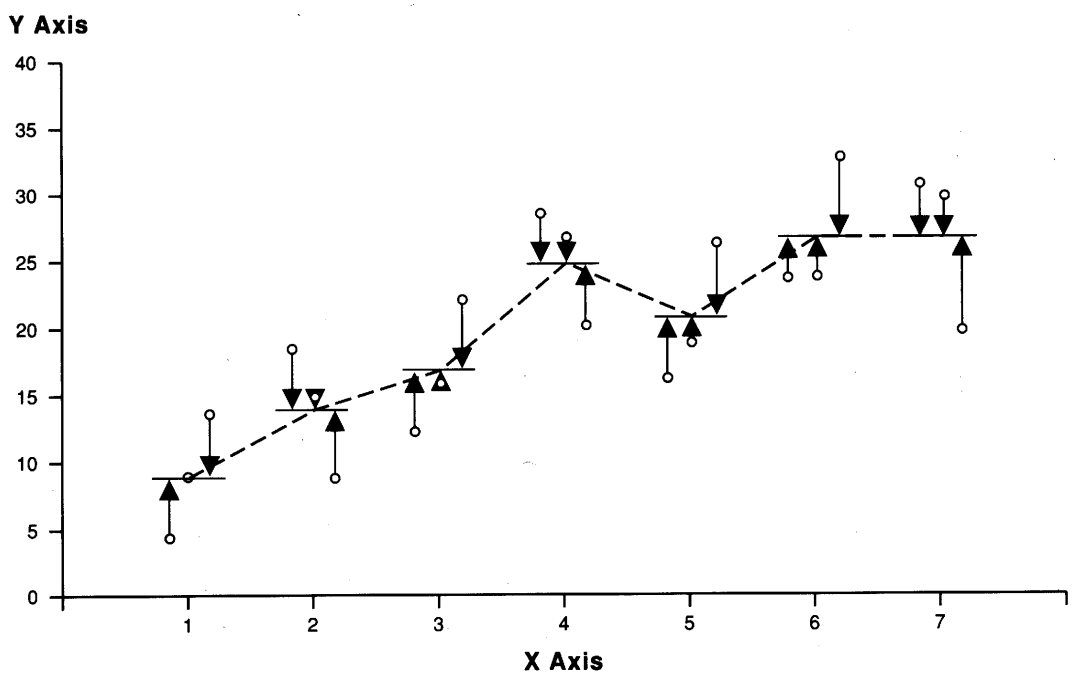
This situation can also be thought of as an ANOVA. The residuals for a group means model is shown by:



The SSR for between groups (SSRBG) is calculated by $\Sigma(Y_{ij} - \hat{Y}_{ij})^2 = 3\Sigma(\overline{Y}_i - \overline{Y})^2$. These calculations are shown below:

X	Group means ($\overline{Y}_i$)	Grand mean ($\overline{Y}$)	Residuals ($\bar{y}_i - \overline{Y}$)	Squared residuals
1	9	20	–11	121
2	14	20	–6	36
3	17	20	–3	9
4	25	20	5	25
5	21	20	1	1
6	27	20	7	49
7	27	20	7	49
Sum:	$3\times140=420$	$3\times7\times20=420$	$3\times0=0$	$3\times290=870=$SSRBG
			Degrees of freedom:	$df = 7 - 1 = 6$

The SSR for individual values within each group (SSRWG) is the SSR around each group mean, similar to the SSRWG calculated for the one-way ANOVA. These residuals are:



The calculations, for SSRWG = $\Sigma(Y_{ij} - \hat{Y}_{ij})^2 = \Sigma(Y_{ij} - \overline{Y}_i)^2$, are:

X	Observed values (Y_{ij})	Group means ($\overline{Y}_i$)	Residuals ($Y_{ij} - \overline{Y}_i$)	Squared residuals
1	9	9	0	0
	4	9	–5	25
	14	9	5	25
2	15	14	1	1
	18	14	4	16
	9	14	–5	25
3	16	17	–1	1
	12	17	–5	25
	23	17	6	36
4	27	25	2	4
	28	25	3	9
	20	25	–5	25
5	19	21	–2	4
	17	21	–4	16
	27	21	6	36
6	24	27	–3	9
	24	27	–3	9
	33	27	6	36
7	30	27	3	9
	31	27	4	16
	20	27	–7	49
Sum:	420	420	0	376 = SSRWG

Degrees of freedom: $df = 21 - 7 = 14$

The SS's required for the ANOVA approach are:

Model	Number of parameters	*df*	SSR name	Sums of squares	Difference	Proportion of SSRM (R^2)
1. Grand Mean	1	20	SSRM	1246		1.00
					870	0.70
2. Three Groups (individual data)	7	14	SSRWG	376		0.30
					not meaningful	—
3. Three Groups (using means)	7	6	SSRBG	870		0.70

These SS's are used to calculate the following ANOVA table:

Source of variation	Sums of squares notation	*df*	Sums of squares	Mean square	*F*-value	*p*-value
Between Groups	SSRBG	6	870	145	5.40	0.0045
i) Regression	SSL = SSRBG − SSRLG	1	756	756	28.15	0.0001
ii) Lack of fit	SSRLG	5	114	22.8	0.85	0.54
Error (Within Groups)	SSRWG	14	376	26.86		
Total	SSRM	20	1246			

Note that the regression SS (SSL) could also have been obtained using a linear contrast.

11.3 Summary of Calculated Sums of Squares

The following are all the SS's that have been calculated for the regression with multiple observations:

Model Used	SS name	SS value	*df*
Calculations using individual data ($n = 21$):			
1. One mean	SSRM	1246	20
2. One line	SSRL	490	19
3. Model SS for one line	SSL = SSRM − SSRL	756	1
4. Seven group means (one for each *X*-value)	SSRWG	376	14
Calculations using group means only ($n = 7$):			
5. One mean	SSRBG	870	6
6. One line	SSRLG	114	5
7. Model SS for one line	SSL = SSRBG − SSRLG	756	1

Note that the two SSL's (3 and 7 above) will not be equal when the sample sizes for each group or the *X*-values are not the same (unless a weighted regression on the means is used).[14]

[14] See Biometrics Information Pamphlet No. 28, Ibid.

Both the regression and the ANOVA approach lead to the same basic analysis and conclusions. The difference lies mainly in whether the problem is seen to be a regression with extra observations at each *X*-value, or an ANOVA where the grouping variable or treatment is quantitative and a linear response to those levels is of interest. An advantage of the ANOVA approach is that SAS can produce all the required SS's in one PROC step (if contrasts can be used), while the regression approach requires two PROC steps.

APPENDIX 1: Summary of Sums of Squares of the Residuals (SSR) Notation

Most of the abbreviations defined below are not in widespread use, although I have tried to keep them as consistent with normal statistical terminology as possible. Each definition below includes the abbreviation and a description of the common name.

SSR: A general term for the sums of squares of the residuals or differences between the actual and the fitted values. It is a measure of the variation of the data about a model.

No common abbreviation is in widespread use.

SSRM: This SSR measures the variation of individual data points about one mean for all the data.

Commonly called the Total Sums of Squares (SST or TSS) or the Corrected Sums of Squares (CSS).

SSRL: This SSR measures the variation of individual data points about the best-fit line.

This is usually just referred to as the residual sums of squares (SSR).

SSL: The difference between SSRM and SSRL. This is the amount of variability in the data which is explained by the best-fit line. This can also be calculated directly. See standard textbooks for the formula.

This term is called the regression or model sums of squares and has no common abbreviation.

SSRWG: This SSR measures the variation of individual data points about their group means. This is the variability Within Groups (hence the WG).

This term is commonly referred to as the Within Sums of Squares (SSW) or the Error Sums of Squares (SSE or ESS).

SSRBG: This SSR measures the variation of the group means about the mean for all the data. It is the SSR of group means fitted by the grand mean and weighted by each group mean's sample size (BG stands for Between Group means).

This term is usually referred to as the Sums of Squares Between groups (SSB) in an ANOVA.

SSRLG: This SSR measures the variation of the group means about the best-fit line. It is the SSR for a line fitted to group means (G for groups and L for line), and is also known as the lack of fit SS.

No distinct common abbreviation in the statistics literature.

SSR3L: This SSR is the sum of the variation of the individual data points about three best-fit lines, where there is one line for each group.

No distinct common name or abbreviation in the statistics literature.

SSR3LP: This SSR is the sum of the variation of the individual data points about three lines which have been fit with the restriction that they all have the same slope (i.e., the lines are parallel, hence the P).

No distinct name for this, but it is the usual error term in Analysis of Covariance.

APPENDIX 2: Mathematical Proofs

This appendix contains some very simple mathematical proofs, but it does require that the reader remember some algebra and calculus. The most useful fact to remember is that the minimum or maximum of a function can be determined by setting the derivative of that function to zero.

1. Proof that the mean provides the best fit for the constant model: $Y = \mathbf{a}$

The type of model under examination is $Y_i = \mathbf{a}$, where **a** is constant for all the data. The residual sums of squares is defined as SSR = sum of $(Y_i - \mathbf{a})^2$ for all data, since **a** is the fitted value. The difference, $Y_i - \mathbf{a}$ is the residual. Using summation notation (and always summing over all values of the index *i*), then:

$$SSR = \Sigma(Y_i - \mathbf{a})^2$$

Now the challenge is to determine which value of **a** will provide the minimum value for this sum. Take the derivative of this function with respect to **a** and set it to zero, namely:

$$\frac{\partial\ SSR}{\partial\ \mathbf{a}} = -2\Sigma(Y_i - \mathbf{a}) = 0$$

Therefore $\Sigma(Y_i - \mathbf{a}) = 0$

implying that $\Sigma Y_i - n\mathbf{a} = 0$

and hence $\Sigma Y_i/n = \overline{Y} = \mathbf{a}$

NOTE: The proof that $\mathbf{a} = \overline{Y}$ = mean is a minimum and not a maximum is left to those who really remember their calculus.

2. Determination of the estimates for a and b in simple regression: $Y = \mathbf{a} + \mathbf{b}X$

The form of the model under examination is $Y = \mathbf{a} + \mathbf{b}X$, where **a** is the intercept and **b** the slope. The residual sums of squares is defined as:

$$SSR = \Sigma(Y_i - \hat{Y}_i)^2 = \Sigma[Y_i - (\mathbf{a} + \mathbf{b}X_i)]^2$$

In this formula, Y_i represents the observed value and $\hat{Y}_i$ represents the predicted or fitted value obtained from the model. The values of **a** and **b** that minimize this sum are the least squares estimates and are denoted by $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$.

Again, differentiate SSR with respect to **a** and **b** and set both derivatives to zero:

$$\frac{\partial\ SSR}{\partial\ \mathbf{a}} = -2\Sigma[Y_i - \hat{Y}_i] = -2\Sigma[Y_i - (\mathbf{a} + \mathbf{b}X_i)] = 0^{15}$$

$$\frac{\partial\ SSR}{\partial\ \mathbf{b}} = -2\Sigma X_i[Y_i - \hat{Y}_i] = -2\Sigma X_i[Y_i - (\mathbf{a} + \mathbf{b}X_i)] = 0$$

[15] Note that this equation can be written as $\Sigma(Y_i - \hat{Y}_i) = 0$. In other words, the sum of the residuals will be zero for the best fitting model.

We can rewrite these as:

$$\Sigma Y_i - n\mathbf{a} - \mathbf{b}\Sigma X_i = 0$$

and $\Sigma X_i Y_i - \mathbf{a}\Sigma X_i - \mathbf{b}\Sigma X_i^2 = 0$

Solve the first equation for **a** by dividing by n and rearranging to get:

$$\frac{\Sigma Y_i}{n} - \mathbf{b}\frac{\Sigma X_i}{n} = \mathbf{a} \text{ so that } \hat{\mathbf{a}} = \bar{Y} - \mathbf{b}\bar{X}$$

Substitute this value for **a** into the second equation to get:

$$\Sigma X_i Y_i - (\bar{Y} - \mathbf{b}\bar{X})\Sigma X_i - \mathbf{b}\Sigma X_i^2 = 0$$

Rearrange this to get all the terms with **b** onto the left side:

$$\mathbf{b}\Sigma X_i^2 - \mathbf{b}\bar{X}\Sigma X_i = \Sigma X_i Y_i - \bar{Y}\Sigma X_i$$

hence, $\mathbf{b}(\Sigma X_i^2 - n\bar{X}^2) = \Sigma X_i Y_i - n\overline{XY}$

and $$\hat{\mathbf{b}} = \frac{\Sigma X_i Y_i - n\overline{XY}}{\Sigma X_i^2 - n\bar{X}^2} = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$$

3. When the sum of the residuals will be zero for the best fitting model

The general linear model is $Y_i = \mathbf{b_1}X_{1i} + \mathbf{b_2}X_{2i} + \ldots \mathbf{b_p}X_{pi}$ with p parameters. Hence the SSR will be:

$$SSR = \Sigma(Y_i - \hat{Y}_i)^2 = \Sigma[Y_i - (\mathbf{b_1}X_{1i} + \mathbf{b_2}X_{2i} + \ldots \mathbf{b_p}X_{pi})]^2$$

Derivatives with respect to the parameters are:

$$\frac{\partial\ SSR}{\partial\ \mathbf{b_1}} = -2\ \Sigma X_{1i}[Y_i - (\mathbf{b_1}X_{1i} + \mathbf{b_2}X_{2i} + \ldots \mathbf{b_p}X_{pi})] = 0$$

$$\frac{\partial\ SSR}{\partial\ \mathbf{b_2}} = -2\ \Sigma X_{2i}[Y_i - (\mathbf{b_1}X_{1i} + \mathbf{b_2}X_{2i} + \ldots \mathbf{b_p}X_{pi})] = 0$$

$$\vdots$$

etc.

The linear model contains a constant if, for example, $X_{1i} = 1$ for all i. In this case, $\mathbf{b_1}$ is an intercept and the first equation above simply states that the sum of the residuals is equal to zero for the best fitting model. Correspondingly, the sum of the fitted values will be the same as the sum of the observed values. This can be seen by examining the first equation above:

$$\frac{\partial \text{ SSR}}{\partial \mathbf{b_1}} = -2 \Sigma X_{1i}[Y_i - (\mathbf{b_1}X_{1i} + \mathbf{b_2}X_{2i} + \ldots \mathbf{b_p}X_{pi})] = 0$$

$$= -2 \Sigma [Y_i - (\mathbf{b_1} + \mathbf{b_2}X_{2i} + \ldots \mathbf{b_p}X_{pi})]$$

$$= -2 \Sigma(Y_i - \hat{Y}_i)$$

Hence, $\Sigma(Y_i - \hat{Y}_i)$ = sum of residuals = 0

When there is no constant in the model, the sum of the residuals is very unlikely to be zero. This situation occurs, for instance, when regressions are forced through the origin (i.e., when the intercept **must** be zero) and for most non-linear models.

APPENDIX 3: Example SAS Programs and Output

The calculations done by hand throughout this handbook can all be done using SAS. The following are example SAS programs with their accompanying output to show how this is done. This should help the translation process from conceptual ideas discussed in the handbook to the actually running of data analyses and understanding of the output.

3.1 Calculations for Sections 2 and 5

```
/* Sections 2 and 5 */

data;
   input x y @@;
cards;
1 9 2 14 3 17 4 25 5 21 6 27 7 27
run;
title '---------------------------------------------- Analysis of first example data ----------------------------------------------';
title2 'Calculation of SSRM - corrected sums of squares for all data';
title3 'Section 2';
proc means n mean css ;
   var y;
run;
title2 'Calculation of SSRL - SSR about best-fit line';
title3 'Section 5';
proc reg;
   model y = x;
run;
```

Output from this program:

---------- Analysis of first example data ----------

Calculation of SSRM - corrected sums of squares for all data

Section 2

Analysis Variable : Y

N Obs	N	Mean	CSS
7	7	20.0000000	290.0000000

---------- Analysis of first example data ----------
Calculation of SSRL - SSR about best-fit lines
Section 5

Model: MODEL1
Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	252.00000	252.00000	33.158	0.0022
Error	5	38.00000	7.60000		
C Total	6	290.00000			

Root MSE	2.75681	R-square	0.8690
Dep Mean	20.00000	Adj R-sq	0.8428
C.V.	13.78405		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	8.000000	2.32992949	3.434	0.0186
X	1	3.000000	0.52098807	5.758	0.0022

3.2 Calculations for Sections 7, 8 and 9

```
/* Sections 7 to 9 */
data;
   input x @;
   do group = 1 to 3;
      input y @;
      output;
   end;
cards;
1 13 18  8
2 14 23 10
3 10 27 12
4 26 35 24
5 22 37 21
run;
title' ---------- Analysis of second example data  ----------';
title2 'Calculation of SSRM - corrected sums of squares for all data';
title3 'Section 7';
proc means n mean css ;
   var y;
run;
title2 'Calculation of SSRWG and SSRBG - One-way ANOVA';
title3 'Section 7';
proc anova;
   class group;
   model y = group;
   means group;
run;
title2 'Calculation of SSRL - One line';
title3 'Section 8';
proc reg;
   model y = x;
run;
title2 'Calculation of SSR3LP and SSRBG - One-way ANCOVA';
title3 'Section 8';
proc glm;
   class group;
   model y = group x / solution;
   means group;
   lsmeans group;
run;
title2 'Calculation of SSR3L - Heterogeneity of Regression';
title3 'Section 9';
proc glm;
   class group;
   model y = group x group*x / solution;
run;
```

Output from this program:

---------- Analysis of second example data ----------
Calculation of SSRM - corrected sums of squares for all data
Section 7

Analysis Variable : Y

N Obs	N	Mean	CSS
15	15	20.0000000	1126.00

---------- Analysis of second example data ----------
Calculation of SSRWG and SSRBG - One-way ANOVA
Section 7

Analysis of Variance Procedure
Class Level Information

Class	Levels	Values
GROUP	3	1 2 3

Number of observations in data set = 15

---------- Analysis of second example data ----------
Calculation of SSRWG and SSRBG - One-way ANOVA
Section 7

Analysis of Variance Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	490.0000000	245.0000000	4.62	0.0325
Error	12	636.0000000	53.0000000		
Corrected Total	14	1126.0000000			

R-Square	C.V.	Root MSE	Y Mean
0.435169	36.40055	7.280110	20.0000000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
GROUP	2	490.0000000	245.0000000	4.62	0.0325

---------- Analysis of second example data ----------
Calculation of SSRWG and SSRBG - One-way ANOVA
Section 7

Analysis of Variance Procedure

Level of GROUP	N	Y Mean	Y SD
1	5	17.0000000	6.70820393
2	5	28.0000000	8.00000000
3	5	15.0000000	7.07106781

---------- Analysis of second example data ----------
Calculation of SSRL - One line
Section 8

Model: MODEL1
Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	480.00000	480.00000	9.659	0.0083
Error	13	646.00000	49.69231		
C Total	14	1126.00000			

Root MSE	7.04928	R-square	0.4263
Dep Mean	20.00000	Adj R-sq	0.3822
C.V.	35.24639		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	8.000000	4.26854926	1.874	0.0836
X	1	4.000000	1.28701603	3.108	0.0083

---------- Analysis of second example data ----------
Calculation of SSR3LP and SSRBG - One-way ANCOVA
Section 8

General Linear Models Procedure
Class Level Information

Class	Levels	Values
GROUP	3	1 2 3

Number of observations in data set = 15

---------- Analysis of second example data ----------
Calculation of SSR3LP and SSRBG - One-way ANCOVA
Section 8

General Linear Models Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	970.0000000	323.3333333	22.80	0.0001
Error	11	156.0000000	14.1818182		
Corrected Total	14	1126.0000000			

R-Square	C.V.	Root MSE	Y Mean
0.861456	18.82938	3.765875	20.0000000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
GROUP	2	490.0000000[a]	245.0000000	17.28	0.0004
X	1	480.0000000	480.0000000	33.85	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
GROUP	2	490.0000000[a]	245.0000000	17.28	0.0004
X	1	480.0000000	480.0000000	33.85	0.0001

Parameter		Estimate		T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT		3.00000000	B	1.13	0.2839	2.66287609
GROUP	1	2.00000000	B	0.84	0.4189	2.38174878
	2	13.00000000	B	5.46	0.0002	2.38174878
	3	0.00000000	B	.	.	.
X		4.00000000		5.82	0.0001	0.68755165

NOTE: The X′X matrix has been found to be singular and a generalized inverse was used to solve the normal equations. Estimates followed by the letter 'B' are biased, and are not unique estimators of the parameters.

[a] In this case, the Type I and III SS for GROUP are the same. This is unusual and occurs here because the X-values for each group are exactly the same.

---------- Analysis of second example data ----------
Calculation of SSR3LP and SSRBG - One-way ANCOVA
Section 8

General Linear Models Procedure

Level of GROUP	N	Y Mean[21]	Y SD	X Mean	X SD
1	5	17.0000000	6.70820393	3.00000000	1.58113883
2	5	28.0000000	8.00000000	3.00000000	1.58113883
3	5	15.0000000	7.07106781	3.00000000	1.58113883

---------- Analysis of second example data ----------
Calculation of SSR3LP and SSRBG - One-way ANCOVA
Section 8

General Linear Models Procedure
Least Squares Means

GROUP	Y LSMEAN[a]
1	17.0000000
2	28.0000000
3	15.0000000

---------- Analysis of second example data ----------
Calculation of SSR3L - Heterogeneity of Regression
Section 9

General Linear Models Procedure
Class Level Information

Class	Levels	Values
GROUP	3	1 2 3

Number of observations in data set = 15

[a] The means and lsmeans are not usually the same. They are in this case because the covariate mean is the same for each group. The lsmeans are called adjusted means in ANCOVA. Each such mean has been adjusted by being moved up or down its covariate line to the *Y*-value at the grand mean of X.

---------- Analysis of second example data ----------
Calculation of SSR3L - Heterogeneity of Regression
Section 9

General Linear Models Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	PR > F
Model	5	990.0000000	198.0000000	13.10	0.0007
Error	9	136.0000000	15.1111111		
Corrected Total	14	1126.0000000			

R-Square	C.V.	Root MSE	Y Mean
0.879218	19.43651	3.887301	20.0000000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
GROUP	2	490.0000000	245.0000000	16.21	0.0010
X	1	480.0000000	480.0000000	31.76	0.0003
X*GROUP	2	20.0000000	10.0000000	0.66	0.5393[2]

Source	DF	Type III SS	Mean Square	F Value	Pr > F
GROUP	2	45.4545455	22.7272727	1.50	0.2732
X	1	480.0000000	480.0000000	31.76	0.0003
X*Group	2	20.0000000	10.0000000	0.66	0.5393[a]

Parameter		Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT		3.00000000 B	0.74	0.4806	4.07703596
GROUP	1	5.00000000 B	0.87	0.4084	5.76579955
	2	10.00000000 B	1.73	0.1169	5.76579955
	3	0.00000000 B	.	.	.
X		4.00000000 B	3.25	0.0099	1.22927259
X*GROUP	1	–1.00000000 B	–0.58	0.5792	1.73845397
	2	1.00000000 B	0.58	0.5792	1.73845397
	3	0.00000000 B	.	.	.

NOTE: The X′X matrix has been found to be singular and a generalized inverse was used to solve the normal equations. Estimates followed by the letter ‘B’ are biased, and are not unique estimators of the parameters.

[a] The test for heterogeneity of regression. This is the only test worth examining on the printout.

3.3 Calculations for Section 11

```
/* Section 11 */

data reg;
   input x @;
   do i = 1 to 3;
      input y @;;
      output;
   end;
cards;
1   9   4 14
2 15 18   9
3 16 12 23
4 27 28 20
5 19 17 27
6 24 24 33
7 30 31 20
;
title ' ---------------------------------------- Analysis of third example data  ----------------------------------------' ;
title2 'Calculation of SSRM - corrected sums of squares for all data' ;
title3 'Section 11' ;
proc means n mean css;
   var y;
run;
title2 'Calculation of group means' ;
title3 'Section 11' ;
proc means mean;
   class x;
   var y;
output out=means mean=ymean n=number;
run;
title2 'Calculation of SSRLG' ;
title3 'Section 11' ;
proc reg data=means;
   weight number;
   model ymean = x;
run;
title2 'Calculation of SSRL' ;
title3 'Section 11' ;
proc reg data=reg;
   model y = x;
run;
title2 'Calculation of SSRBG and SSRWG' ;
title3 'Section 11' ;
proc glm data=reg;
   class x;
   model y = x;
   contrast 'One Line' x –3 –2 –1   0   1   2   3;
run;
```

Output from this program:

---------- Analysis of third example data ----------

Calculation of SSRM - corrected sums of squares for all data
Section 11

Analysis Variable : Y

N Obs	N	Mean	CSS
21	21	20.0000000	1246.00

---------- Analysis of third example data ----------

Calculation of group means
Section 11

Analysis Variable : Y

X	N Obs	Mean
1	3	9.0000000
2	3	14.0000000
3	3	17.0000000
4	3	25.0000000
5	3	21.0000000
6	3	27.0000000
7	3	27.0000000

---------- Analysis of third example data ----------
Calculation of SSRLG
Section 11

Model: MODEL1
Dependent Variable: YMEAN

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	756.00000	756.00000	33.158	0.0022
Error	5	114.00000	22.80000		
C Total	6	870.00000			

Root MSE	4.77493	R-square	0.8690
Dep Mean	20.00000	Adj R-sq	0.8428
C.V.	23.87467		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	8.000000	2.32992949	3.434	0.0186
X	1	3.000000	0.52098807	5.758	0.0022

---------- Analysis of third example data ----------
Calculation of SSRL
Section 11

Model: MODEL1
Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	756.00000	756.00000	29.314	0.0001
Error	19	490.00000	25.78947		
C Total	20	1246.00000			

Root MSE	5.07833	R-square	0.6067
Dep Mean	20.00000	Adj R-sq	0.5860
C.V.	25.39167		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	8.000000	2.47797314	3.228	0.0044
X	1	3.000000	0.55409164	5.414	0.0001

---------- Analysis of third example data ----------
Calculation of SSRBG and SSRWG
Section 11

General Linear Models Procedure
Class Level Information

Class	Levels	Values
X	7	1 2 3 4 5 6 7

Number of observations in data set = 21

---------- Analysis of third example data ----------
Calculation of SSRBG and SSRWG
Section 11

General Linear Models Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	870.0000000	145.0000000	5.40	0.0045
Error	14	376.0000000	26.8571429		
Corrected Total	20	1246.0000000			

R-Square	C.V.	Root MSE	Y Mean
0.698234	25.91194	5.182388	20.0000000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X	6	870.0000000	145.0000000	5.40	0.0045

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X	6	870.0000000	145.0000000	5.40	0.0045

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
One Line	1	756.0000000	756.0000000	28.15	0.0001

ADDITIONAL READING

General Textbooks:

Snedecor, G. W. and W. G. Cochran. 1980. Statistical methods. 7th ed. The Iowa State Univ. Press, Ames, Iowa.

Sokal, R.R. and F.J. Rohlf. 1981. Biometry. W.H. Freeman and Co., San Francisco, Ca.

Steel, R.G.D. and J.H. Torrie. 1980. Principles and procedures of statistics: a biometrical approach. 2nd ed., McGraw-Hill Book Co., New York, N.Y.

More Advanced or Specific Textbooks

Dobson, A. 1983. An introduction to statistical modelling. Chapman and Hall, London, Eng.

Gilchrist, W. 1984. Statistical modelling. John Wiley & Sons, Toronto, Ont.

Huitema, B.E. 1980. The analysis of covariance and alternatives. John Wiley & Sons, New York, N.Y.

Milliken, G.A. and D.E. Johnson. 1984. Analysis of messy data. Vol. I: Designed experiments. Lifetime Learning Public., Belmont, Ca.

Wetherill, G.B. 1981. Intermediate statistical methods. Chapman and Hall, New York, N.Y.

Biometrics Information Pamphlets (Referenced in Text)

Available from the B.C. Ministry of Forests, Forest Science Research Branch, Biometrics Section, Victoria, B.C.

No. 18 Multiple regression: selecting the best subset

No. 21 What are degrees of freedom?

No. 27 When the *t*-test and *F*-test are equivalent

No. 28 Simple regression with replication: testing for lack of fit