R Module 6

Simple Linear Regression

Certificate in EnvlroStats (Non-Award)

This document is part of an online Certificate in EnviroStats (Non-Award) by the University of Canberra. Course enquiries can be directed to the address below. Expressions of interest in the course can be made online through:

http://aerg.canberra.edu.au/envirostats

Copies of this publication are available from:

The Institute for Applied Ecology
University of Canberra ACT 2601
Australia

Email: **bernd.gruber@canberra.edu.au, georges@aerg.canberra.edu.au**

R is an open source statistical programme. It is developed by:
R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL **http://www.R-project.org**.

SPONSORED BY:

Australian Government
Australian Centre for International Agricultural Research

WorldFish CENTER

Australian Government
Fisheries Research and Development Corporation

Materials development team:

Author:	Arthur Georges, 2002, 2006
Instructional designer:	Peter Donnan, 2002
Editor:	Loretta Barnard, 2002
Graphic Design:	Peter Delgado, 2002
Desktop Publishing:	Kristi McDonald, 2004 Sue Bebbington, 2004
FDDU Project Manager:	Deborah Veness, 2002
Dynamic Web Page Design:	TCNI Software Solutions PO Box 47 LATHAM ACT 2615 Australia

First prepared in January, 2002 for Semester 1, 2002.
Reprinted January 2003 for Semester 1, 2003.
Reprinted January 2004 for Semester 1, 2004.
Reprinted November 2004 for Semester 1, 2005.
Revised and reprinted, June 2006
Reprinted February 2007 for Semester 1, 2007
Revised former SAS Module 5 and converted to R, February 2011

Published by Technology & Educational Design Services

(TEDS)
University of Canberra
ACT 2601, AUSTRALIA

Module 5

Simple Linear Regression

Lesson 1: Key Concepts in Regression 5
Overview 5
Simple linear regression in a nutshell 7
The Regression Equation 7
Parameter Estimation 8
Confidence Limits 10
Significance Testing 11
Strength of Result 12
Adequacy of Fit 12
Where have we come? 13
Lesson 2: Regression as an Analysis of Variance 14
Single-factor ANOVA revisited 14
From ANOVA to regression 16
A Fundamental Difference 16
Partition of the Sums of Squares 16
The ANOVA Table 17
Adequacy of Fit 18
Interpreting a Significant Residual 19
Curvilinearity 19
Other influential regressors 21
Reporting the results 22
Special Case: Simple linear regression 23
Where have we come? 25
Lesson 3: Application Notes 27
Assumptions of regression 27
X values fixed 27
Randomness in sampling 28
Linearity 29
Independence 29
Equality of variances 30
Normality 31
Analysis of residuals 32
Non-linearity in the regression function 33
Heterogeneity of the error variance 34
Lack of independence of the errors 35
Presence of aberrant points (outliers) 36
Non-normality in the distribution of errors 38
Caveats on the use of transformations 39
Caveats on the use of R^2 40
Where have we come? 40
Lesson 4: Step-through Examples 42
Example 6-1: Time to pipping in the pig-nosed turtle 42
The problem 42
The data 42
The Analysis 43
Regression analysis 45
Results 48
Discussion 49
Sources 49
Example 6-2: Mercury in Gemfish 50
The problem 50
The data 50
The analysis 50

Data entry and exploratory examination 50
Regression analysis 52
Results 60
Discussion 61
Sources 62
Where have we come? 62
Lesson 5: Exercises 63
Exercise 6-1: Mercury Contamination of Gemfish 63
Analysis -- Mercury 64
Analysis – Length-Weight Relationship 65
Analysis – Size Frequency Distribution 65
Results summary 65
Discussion 65
Exercise 6-2: Zinc Accumulation in Cockles 66
Exercise 6-3: Patterns of butterfly species richness 68
Where have we come? 69

Lesson 1: Key Concepts in Regression

Overview

In **regression**, we estimate the relationship one variable has with another by expressing one as a function of the other.

The growth curve of a weed species is spoken of as a regression of height against age. If we know the age of a stand, we can use this relationship to obtain an estimate of the height of the plants. Turbidity in a lake varies with distance from the primary inflow, and this relationship may be modelled as a regression. Length-weight relationships are of great value in fisheries, where it is useful to obtain predictions of the weight of a fish from a measurement of its length. Mercury concentrations accumulating in the muscle tissue of fish may be strongly related to the age of a fish, and therefore its size, so we may be interested in establishing a regression of mercury concentration on body mass.

Regression analyses can have a number of objectives. Most commonly, regression is used to predict the value of one variable from the value of another, when the two are related.

In regression, the relationship between two variables is expressed as a function, expressed graphically as a line. In **simple linear regression**, the function is a straight line that can be expressed as:

$$Y = B_0 + B_1 X$$

where B_0 is the ***Y* intercept** and B_1 is the **slope** of the line.

You can verify that this is a straight line on a graph by plotting it for selected points, or by noting that a given increment in X yields a given change in Y that does not vary with the value of X. This would not be true of the relationship

$$Y = B_0 + B_1 X^2$$

for example.

The regression equation describes the **fitted line** that minimises the squared deviations of the data points from the line, the residual sums of squares, and as such is sometimes called **least squares regression**.

Regression analysis does not treat the two variables equally. *Y*, the **response** or dependent variable, is treated as a function of *X*, the

regressor or independent variable. The values of *X* are usually under the control of the investigator.

Statistical analyses involve estimating the parameters of the regression, their standard errors and confidence limits. The parameters can be tested for significance.

If the slope is significantly different from zero, we say that there is a significant regression. In other words, for an incremental change in the regressor *X*, we can expect a corresponding incremental change in the response variable *Y*. The two variables are related in a predictable way. A test for a significant slope is a test for a significant regression.

We can use a significant regression to predict a value of the response variable for a given value of the regressor, and set confidence limits for this prediction.

If the slope of the regression is not significant, we have no evidence to refute the suggestion that the two variables are varying independently. An incremental change in the regressor *X* provides no information on the corresponding change likely to occur in the response variable *Y*.

While regression might be very useful for prediction, it is important to note that no matter how strong the regression or how tight the fit of the data to the regression line, causality is not necessarily implied. The number of churches in a city and the crime rate are strongly and positively related. Indeed, if we draw from American towns and cities, a regression between the number of churches and the number of violent crimes yields and R^2 of 0.72. The more churches, the more violent crime.

The number of churches may be used as a regressor to predict the crime rate well in a linear regression, but no one suggests that the churches or those who preach in them are causing the crime. City size is the likely culprit, hidden behind the scenes as a coincident driver of the crime rate and the number of churches in a city. A strong and significant regression does not, on its own, provide evidence of causality. Causality cannot easily be established outside a strict experimental context.

Simple linear regression can be regarded as a natural extension of fixed-model single-factor ANOVA, bringing the two together under the umbrella of a **General Linear Model** (GLM). Modern statistical packages present the output of a regression analysis in the form of an analysis of variance, and interpretation of this output demands an understanding of the theoretical links between regression and ANOVA. Much of the theoretical material presented in this workbook will deal with this link.

Simple linear regression in a nutshell

The Regression Equation

The objective of regression analysis is usually to obtain a predictive relationship between one variable, the regressor or independent variable, and the other, the response or dependent variable. Usually, values of the regressor variable are deliberately chosen or controlled by the investigator.

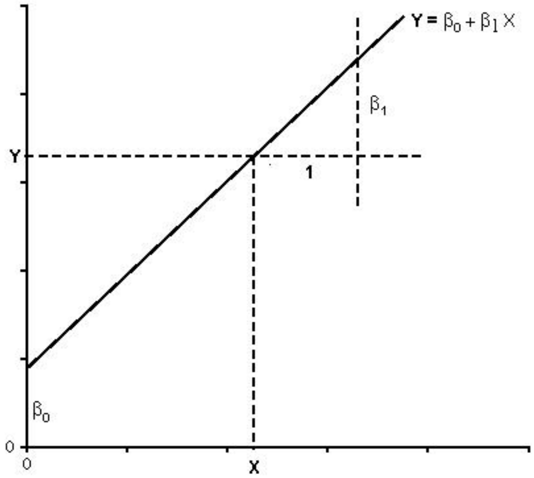
In simple linear regression, we assume that the underlying relationship between two variables, if any, is linear. A linear relationship can be described by the function:

$$\mu_{Y|X} = \beta_0 + \beta_1 X$$

where $\mu_{Y|X}$ is the true parametric mean of response Y for a given value of the regressor X, and β_0 and β_1 are parameters to be estimated.

When this function is graphed (Figure 6-1), the parameters β_0 and β_1 take on intuitive meaning. β_0 is the ***Y* intercept**; it is the value of Y when $X = 0$. β_1 is the **slope** or **regression coefficient**; it is the incremental change in *Y* for a unit change in *X*. If the response variable *Y* increases with increasing *X*, then the slope is positive; if *Y* decreases with increasing *X*, the slope is negative.

Figure 6-1. A linear function of Y on X showing the interpretation of the parameters β_0 (the Y intercept) and β_1 (the slope). Note that the slope is the amount by which Y increases for each unit increase in X.



The problem of quantifying the regression of one variable on the other becomes the problem of estimating the intercept and slope of the relationship between the two variables.

In the real world, variables are not perfectly related (Figure 6-2). There may be a good relationship between the age of a woody weed and its height, but it would be too much to expect that knowledge of a plant's age would enable us to calculate its height exactly. We might expect to get an estimate of its height from our regression of height against age, but the estimate would be subject to natural error. Plants of the same age vary in height.

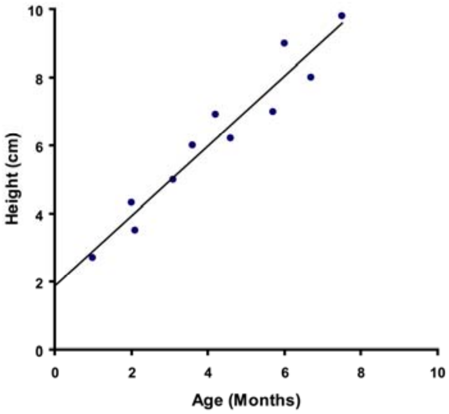


Figure 6-2. True underlying relationship between height and age in young individuals of a woody weed. Note that the relationship is not perfect. Plants vary in height from that predicted from the true underlying relationship.

Hence, the relationship between *Y* and *X* becomes:

$$Y = \beta_0 + \beta_1 X + \varepsilon_i$$

where *Y* is an actual value of the response variable for a given *X* and ε_i is the deviation of the value from that expected from the linear relationship. Height of a woody weed is governed by some underlying average relationship between height and age, and an additional component, the error ε_i, associated with that particular plant.

Parameter Estimation

When faced with data such as those shown in Figure 6-2, we do not know the true parametric relationship between the two variables, but can calculate our best estimate of it, the straight line of best fit:

$$\hat{Y} = B_0 + B_1 X$$

where $\hat{Y}$ is the value on the line corresponding to our given value of *X*. Intercept B_0 and slope B_1 are estimates of the true parameters β_0 and β_1 respectively (Figure 6-3).

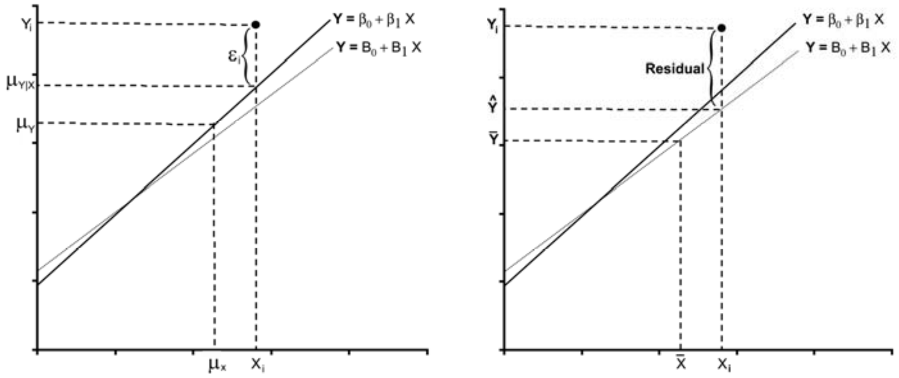


Figure 6-3. Relationship between the true regression line (solid line) and that estimated from the data (dashed line). Note that the sample regression differs from the true regression in slope ($B_1 \neq \beta_1$) and in elevation ($\overline{Y} \neq \mu_Y$). Note also the difference in terminology for describing the true line and its estimate.

In particular, $\hat{Y}$ is the value taken by the estimated line for a given value of X ($\hat{Y} \approx \mu_{Y|X}$).

The regression line we calculate from our data will be our best estimate of the true but unknown linear relationship, but will differ from it because of sampling error. It will differ in two important ways. First, it will differ from the true relationship because the slope is subject to sampling error. B_1 will not equal β_1, but rather will be an estimate of it. Second, our calculated relationship will differ from the true underlying relationship because the bivariate mean through which our line runs will be subject to sampling error. $\overline{Y}$ and $\overline{X}$ will only be estimates of their parametric values μ_Y and μ_X respectively.

But how do we calculate our sample line? The line we seek is the least squares line of best fit. It is the line for which the sum of the squared deviations of the points Y from their corresponding values on the line is a minimum. We write:

$$SS_{residual} = \sum_i \left(Y - \hat{Y}\right)^2$$

We could find the least squares line of best fit by fitting a best guess to the data by eye, then jiggling it about in some systematic way to iteratively find the combination of slope and intercept that minimises the residual sums of squares. This is the approach often taken in non-linear regression (see `nls()` in R). Alternatively, and more usually in cases of linear regression, we can rely on the work of mathematical statisticians who have derived formulae for the least squares solutions for the slope and intercept of the line of best fit.

$$B_1 = \frac{\sum^{n} \left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{\sum^{n} \left(X - \overline{X}\right)^2}$$

$$B_0 = \overline{Y} - B_1\overline{X}$$

Confidence Limits

Confidence limits are used to place bounds on the value of population parameters whose values are typically unknown. Confidence limits, for example, may be used to determine the range within which we can be 95% sure the true parametric population mean lies (refer to Workbook 3).

$$CL_{95} = Statistic \pm t_{0.05[2]n-2} S\tan dard\ Error$$

In regression, three parameters were introduced. There were the parameters of the regression line itself, the intercept β_0 and slope β_1, and there was the parametric value of the true value of *Y* for a given value of *X*, $\mu_{Y|X}$. We can set confidence limits for each of these. We can also set confidence limits for the determination of a single value of Y for a given value of X. When referring to confidence limits in regression, it is important to distinguish among these possibilities.

The 95% confidence limits for the slope provide bounds within which we can be 95% sure the true parametric slope lies. The range of slopes within which you can be 95% sure the true slope lies can be shown graphically as two lines intersecting at the bivariate mean of *X* and *Y*.

Calculation of the 95% confidence limits for the expected value of *Y* (that is, $\overline{Y}$ *)* for a given value of *X* is more complicated. Such estimates will have two sources of error—error in the estimate of the elevation of the regression line (our estimate of $\overline{Y}$ at $\overline{X}$) and error in the slope. Clearly, the impact of error in the slope on our predictions will depend on the value of *X* (Figure 6-4). Thus, the confidence limits for the prediction of *Y* for a given value of *X* will depend on which value of *X* we choose. The further away our *X* value is from $\overline{X}$, the poorer will be the precision in our estimates of *Y* (Figure 6-4).

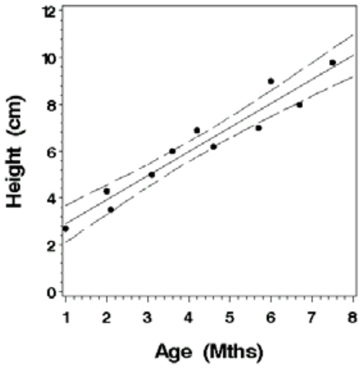


Figure 6-4. Regression of height of young woody weeds against age, showing the 95% confidence limits for the expected value of height (that is, the mean value of height) for a given value of age.

The 95% confidence limits for the expected value of the *Y* intercept are just a special case of the confidence limits for the expected value of *Y* for a given value of *X* (*X* is zero). Of course, great care must be taken in relying on these confidence limits if the *Y* intercept lies outside the range of the data.

The 95% confidence limits for a single value of Y for a given X are very different from the 95% confidence limits for the mean value of Y for a given X. You need to be careful in choosing the limits appropriate to your research question. If, for example, you are wishing to set limits for the average consumption of mercury in fish for the human population, limits based on the 95% confidence limits for the determination of mean mercury content from fish length may well be appropriate. If however you intend to set a maximum limit for mercury in any fish sent to market, then the appropriate confidence limits are the ones that give you 95% assurance that no single fish exceeds your limits. The confidence limits for determination of mercury in an individual fish will be much wider than those for the mean determination for fish of a given size.

Sokal and Rohlf (1994: Box 14.2) provide formulae for standard errors and confidence limits in regression. Most statistical packages will provide output that includes all of the above confidence limits. The distinction between the various confidence limits will be provided in the worked examples.

Significance Testing

Testing the significance of the regression involves testing the significance of the regression coefficient β_1, that is, we test the null hypothesis that the sample value of B_1 comes from a population with a parametric value of $\beta_1 = 0$:

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$$

The sampling distribution of B_1 is normal under not too restrictive assumptions, so the regression can be tested with a t-test:

$$t = \frac{Statistic}{Standard\ Error} = \frac{B_1 - 0}{S_{B_1}}$$

where S_{B_1} is the standard error for the slope *B1*. The degrees of freedom for the test are $\nu = n - 2$ as there are two fitted parameters. The test is a two-tailed test because the parametric slope may be less than, equal to or greater than zero.

Of course, we could also test whether *B1* was different from a parametric value of *β1* other than zero.

Strength of Result

The strength of the regression is given by the magnitude of the regression coefficient, the steepness of the slope, though you need to be careful as it is unit dependent. It is the measure of the magnitude of expected change in *Y* for a given change in *X*. Note that where the data are abundant, a highly significant regression can nevertheless be a weak regression.

Adequacy of Fit

The **coefficient of determination** (*R*2) is a measure of the adequacy of the regression line as a summary of the data upon which it is based.

$$R^2 = \frac{SS_{regression}}{SS_{total}} = \frac{\sum^{n}\left(\hat{Y} - \overline{Y}\right)^2}{\sum^{n}\left(Y - \overline{Y}\right)^2}$$

It is coincidentally equal to the square of the Pearson correlation coefficient, with $0 \le R^2 \le 1$.

If there is wide scatter of the data about the regression line, then the fit is poor, predictions will be imprecise, and the coefficient of determination, *R*2, will be small.

If there is a tight fit of the data to the regression line, then the fit is good, predictions will be precise, and *R*2 will be close to 1. If all the data points lie on the line, then $R^2 = 1$.

What is regarded as a good fit will vary from discipline to discipline and from application to application. When calibrating a piece of equipment, an *R*2 of at least 0.98 might be expected for the regression of the instrument readings against the true value. In an ecological context, *R*2 of 0.70 might be regarded as a good fit.

Note that statistical significance, strength of result and adequacy of fit are not closely coupled. A highly significant result can occur when the regression is weak and the fit poor, if the sample size is large. The fit can be good, and predictions precise, even though the influence of the regressor variable on the response variable is modest (the regression is weak). And of course, a strong regression apparent in the sample data (B_1 large) may not be significant if the sample size is small.

Where have we come?

Much of the above introduction to regression should have been revision for you, as this course assumes that you have done a first course in statistics. You should now appreciate that

- Regression is an analysis that estimates a linear function relating one variable (the response variable Y) to another variable (the independent variable X).
- Regression is used to predict the value of the response variable for a given value of the independent variable.
- Regression does not, on its own, have anything to say on the matter of causality.
- In a regression analysis, we typically estimate the slope of the regression and the Y intercept. These two statistics enable us to construct the regression equation used to predict Y from X.
- A t-test can be used to test the significance of the regression, or more precisely, whether or not the slope of the regression is significantly different from zero. A zero slope means that variation in X is unrelated to variation in Y.
- R^2, the coefficient of determination, provides a measure of the proportion of variation in Y that can be explained by variation in X. If there is a perfect fit of the line to the data, then R^2 will be equal to 1. If there is no regression at all, R^2 will be equal to 0. R^2 is a measure of the scatter (unexplained noise) of the points about the regression line.
- Strength of result is given by the magnitude of the slope, in that the magnitude of the slope tells us by how much Y changes for a given change in X. This assessment needs to take into account the units of measurement.
- Confidence limits of the slope, for the predicted mean value of Y for a given X, and for the prediction of an individual Y estimate for a given value of X are all available to assist in conveying confidence in the results of the analysis.

If any of these matters remain unclear to you, refer to an elementary text in statistics.

Lesson 2: Regression as an Analysis of Variance

Single-factor ANOVA revisited

In a pilot study, Kurt Hammerschmidt collected ten replicate samples of water from each of ten sites in Lake Burley Griffin (Figure 6-5). The sites were specifically chosen at set intervals along the main channel leading from the inflow to the Scrivener Dam wall so that they could be revisited if necessary. Turbidity (in ntu) was measured for each replicate sample taken at each site, and the data are shown in the Table 6-1.

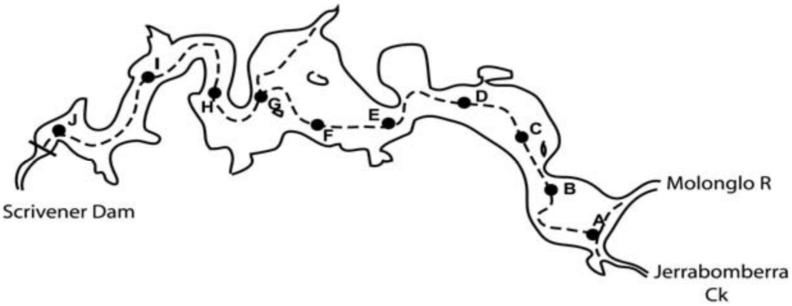


Figure 6-5. A map of Lake Burley Griffin showing the main channel (dashed line) and the sampling stations used for collection of water samples (•).

Table 6-1. Turbidity values (NTU) for each of 10 water samples taken at each of 10 sites in Lake Burley Griffin, Canberra.

				SITE					
A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**1**	**J**
43	25	23	32	17	23	14	13	15	13
28	28	24	32	21	21	18	26	15	15
43	28	30	32	18	17	14	18	14	14
28	25	32	33	17	18	16	15	12	13
42	25	25	32	25	19	14	15	17	16
43	25	28	29	17	24	9	14	19	19
40	26	23	26	18	14	14	17	15	16
35	25	25	38	14	17	26	15	14	15
42	23	26	27	15	17	10	11	16	11
43	25	27	29	15	18	15	14	14	15

We could analyse these data as a single-factor ANOVA to determine if there were differences in turbidity among sites and where those differences lie. Indeed, this was the objective of Exercise 4-1 of Module 4.

It is a fixed model ANOVA, as the sampling sites were chosen specifically and systematically along the drainage channel.

Sites in Lake Burley Griffin differed significantly in turbidity (F=52.62; df=9,90; p<0.0001) (Table 6-2). Site A, closest to the inflows of the Molonglo River and Jerrabomberra Creek had significantly higher turbidity than any other site (Tukey-Kramer Procedure, P<0.05). Sites B, C and D of the east and central basins did not differ significantly in turbidity, and were intermediate. Sites

E, F, G, H, I and J of the western basin, closest to the dam wall, did not differ significantly in turbidity, and collectively had the lowest turbidity (Table 6-2).

Table 6-2. Results of a single-factor ANOVA comparing turbidity (NTU) among sites in Lake Burley Griffin.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Among Sites	9	6025.04	669.448889	52.62	<.0001
Within	90	1145.00	12.722222		
Total	99	7170.04			

Non-significant subsets were obtained from Tukey-Kramer multiple comparisons.

A
38.7

B	C	D
30.5	26.3	26.0

F	E	H	I	G	J
18.8	17.7	15.8	15.1	15.0	14.7

The results are shown graphically in Figure 6-6. The overall pattern is obvious, supported by the Tukey-Kramer analysis. There is a progressive decline in turbidity as turbid water entering the lake moves through the lake toward the outflow. The suspended solids are flocculating out over time.

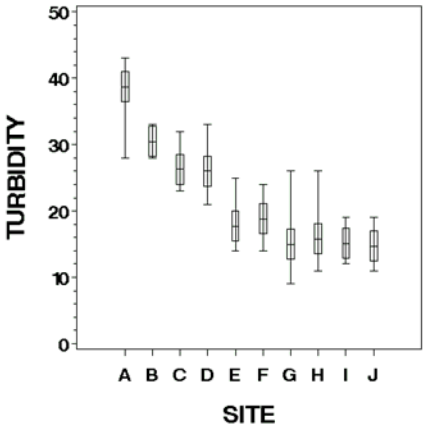


Figure 6-6. Turbidity values (NTU) for each of 10 water samples taken at each of 10 sites in Lake Burley Griffin, Canberra. Error bars show ranges, boxes show ± 2 standard errors (n=10).

From ANOVA to regression

A Fundamental Difference

But might not this analysis be better done using regression? What change do we have to make to convert the problem from a single-factor ANOVA to a regression? It turns out that the only change we need is to convert our factor SITE, measured at the nominal or ordinal scale, to a regressor variable DISTANCE (X), measured at the ratio scale. Distance can be measured along the drainage channel from the inflow. TURBIDITY remains as the response variable (Y). The new dataset is shown in Table 6-3 and graphed in Figure 6-7.

Table 6-3. Turbidity values (NTU) for each of 10 water samples taken at each of 10 sites in Lake Burley Griffin, Canberra. Distance is measured as km from the inflow of Molonglo River.

DISTANCE									
0.82	2.00	3.09	3.91	5.59	6.47	8.00	9.00	10.44	12.65
43	25	23	32	17	23	14	13	15	13
28	28	24	32	21	21	18	26	15	15
43	28	30	32	18	17	14	18	14	14
28	25	32	33	17	18	16	15	12	13
42	25	25	32	25	19	14	15	17	16
43	25	28	29	17	24	9	14	19	19
40	26	23	26	18	14	14	17	15	16
35	25	25	38	14	17	26	15	14	15
42	23	26	27	15	17	10	11	16	11
43	25	27	29	15	18	15	14	14	15

Partition of the Sums of Squares

The linear regression line is clearly a poor model for these data, but let us persist with it for the moment.

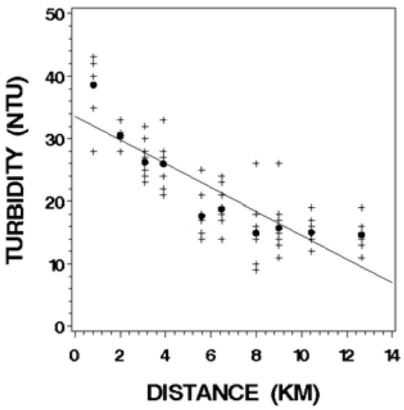


Figure 6-7. A plot of turbidity against distance from the inflow of Lake Burley Griffin, Canberra. Both individual data points (+) and means (•) are shown. The line is for the least squares regression.

In our single-factor ANOVA, we partitioned the total variation in turbidity measurements into two components, a component attributed to variation among the means and a component attributed to variation of individual measurements about their site mean:

$SS_{total} = SS_{among} + SS_{within}$

We can now carry this partitioning one step further. We can ask, how much of the variation among site means can be attributed to an underlying linear relationship with distance and how much remains unexplained?

$SS_{among} = SS_{regression} + SS_{residual}$

This new, more detailed partition of the total sums of squares is illustrated diagrammatically in Figure 6-8. The corresponding ANOVA table is shown in Table 6-4.

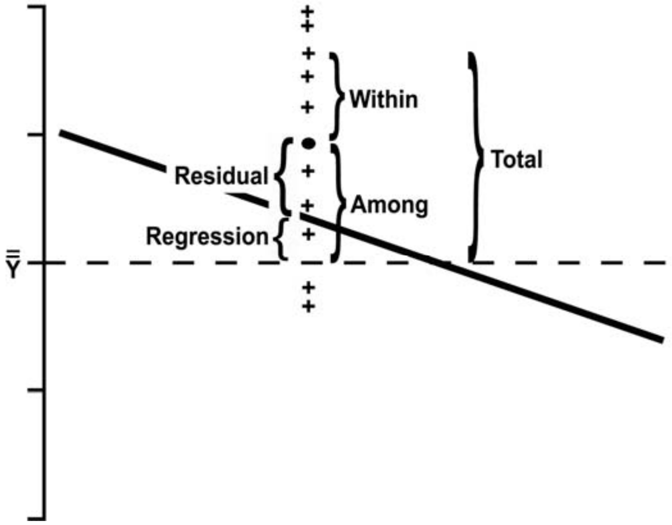


Figure 6-8. A diagramatic representation of the partition of sums of squares for the regression with more than one value of Y for each value of X. The grand mean $\bar{\bar{Y}}$ *is shown as a horizontal dashed line; the individual sample mean* $\bar{Y}$ *is a dot (•); the individual data points (Y) are crosses (+).*

The ANOVA Table

There is much to interpret in this ANOVA table (Table 6-4). There are significant differences in turbidity among the sites (F=52.62; df=9,90; p<0.0001). We knew this from the single-factor ANOVA. The among sites line and within lines in this expanded table are the same as for the single-factor ANOVA. What we can now say, though, is that a significant component of the variation among sites can be explained by a linear regression of turbidity against distance from the inflow (F=31.76; df=1,8; p<0.0005). The regression equation can be calculated separately, and is $\hat{Y} = 33.66 - 1.905X$.

Table 6-4. Results of a regression of turbidity (NTU) versus distance from the inflow for Lake Burley Griffin, Canberra. Note that there is more than one value of Y (turbidity) for each value of X (distance), which allows for a more detailed analysis.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Among Sites	9	6025.04	669.4489	52.62	0.0001
Regression	1	4812.73	4812.730	31.76	0.0005
Residual	8	1212.31	151.5388	11.91	0.0001
Within	90	1145.00	12.72222		
Total	99	7170.04			

Source	Expected MS
Among Sites	$\sigma^2 + n\sigma_A^2$
Regression	$\sigma^2 + n\sigma_{res}^2 + n\sigma_{reg}^2$
Residual	$\sigma^2 + n\sigma_{res}^2$
Within	σ^2

In testing the regression, we use:

$$F = \frac{MS_{reg}}{MS_{res}} \approx \frac{\sigma^2 + n\sigma_{res}^2 + n\sigma_{reg}^2}{\sigma^2 + n\sigma_{res}^2}$$

MS_{reg} always has 1 degree of freedom in simple linear regression.

Note that had we tested the slope by traditional means, using a t-test, we would have obtained t = 5.636 with $n-2=8$ degrees of freedom. This is the third instance of where $F_{0.05[1]1,v_2} = t^2_{0.05[2]v_2}$, ensuring that the two approaches to testing regression always yield the same outcome.

Adequacy of Fit

Just how good is the regression line in summarising the observed variation among the site means? One way of measuring this is to express the sums of squares explained by the regression (SS_{reg}) as a proportion of the total sums of squares among means (SS_{among}). This value is called the **coefficient of determination** and usually represented by the symbol R^2:

$$R^2 = \frac{SS_{regression}}{SS_{among}} = \frac{4812.73}{6025.04} = 0.7899$$

So 79% of the variation among mean turbidity measurements can be explained by a linear regression against distance from the inflow.

If you have computed linear regressions in elementary statistical courses, you will have come across R^2 before. You will have appreciated that it is a measure of the proportion of variation explained by the regression, but it is not until you view regression as a natural extension of ANOVA that you can understand why it gives the proportion of variation explained.

Interpreting a Significant Residual

Once we have taken out the effect of the regressor, that is, the variation explained by the regression, we can ask if the remaining variation among the means, the residual, is greater than we might expect by chance. We are effectively taking the regression line as our base line, and then performing a single-factor ANOVA to compare the means. Such means are referred to as **corrected means**.

In testing the residual, we use:

$$F = \frac{MS_{res}}{MS_{within}} \approx \frac{\sigma^2 + n\sigma_{res}^2}{\sigma^2}$$

Clearly, the variation of the site means about the regression line is greater that we would anticipate on consideration of the variation within sites (F=11.91; df=8,90; p<0.0001).

A significant residual can occur in one of two ways. We have assumed under our null hypothesis, that the true population means $\mu_{Y|X}$ all lie on the regression line, that is, that there is a true underlying linear relationship. Under the null hypothesis, all of the variation in sample means about the regression line would be a result of sampling error. If the true population means are not perfectly linearly related, then we can expect a significant residual variance. This can occur if the underlying relationship is curvilinear rather than linear. It can also happen if there is another factor or regressor, uncorrelated or poorly correlated with distance from the inflow, that is differentially pulling the site means away from the regression line. Often, it is a combination of both influences that leads to a significant residual.

Curvilinearity

Let us explore the curvilinear option first. Curvilinearity certainly appears to be the case in our turbidity example. A significant residual would indicate that our linear model is a **poorly specified model**. Examination of the residuals in graphic form demonstrates this clearly (Figure 6-9). If the regression model was appropriate, we would expect no systematic trend in the residuals across the plot—they should scatter about the reference line randomly. They clearly do not.

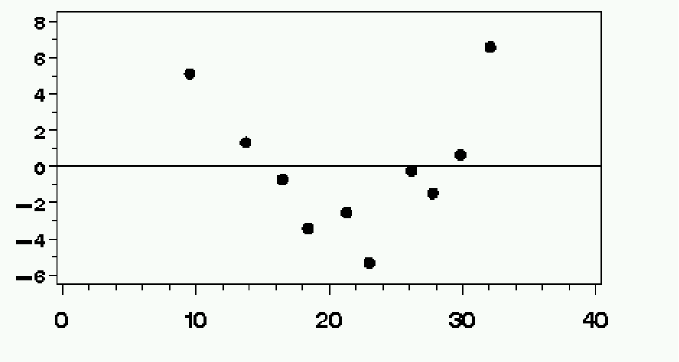


Figure 6-9. A plot of residuals for mean turbidity across 10 sites in Lake Burley Griffin, Canberra. There is clear evidence of curvilinearity.

There are several approaches to overcoming curvilinearity. We may turn to theory for a better model to describe our trend. Length-weight relationships are usually not linear, but follow a power function of the form:

$$l = Aw^B$$

A suitable transformation may bring the relationship into the linear regression fold:

$$Log_{10}l = Log_{10}A + B.Log_{10}w$$

In other cases, no suitable transformation to linearity exists, and we must turn to iterative approaches to fitting the least squares solution to the theoretical model (using `nls()` in R).

In still other cases, there may be no basis for selecting an underlying theoretical curve, and we may need to take an empirical approach—fitting a generic **polynomial model** is one such approach. Here we first fit a linear model as above, then fit a quadratic term to accommodate the curvature and reduce the residuals, and so on, until the residual variation is no longer significant.

Table 6-5. Results of a polynomial regression of turbidity (NTU) versus distance from the inflow for Lake Burley Griffin, Canberra.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Among Sites	9	6025.04	669.4489	52.62	0.0001
Distance	1	4812.73	4812.730	210.35	0.0001
Distance^2	1	1052.15	1052.15	45.99	0.0001
Residual	7	160.157	22.8796	1.80	0.0969
Within	90	1145.00	12.72222		
Total	99	7170.04			

Our revised ANOVA table will look like that shown in Table 6-5. Note that both the linear and the quadratic terms are significant. The residual is no longer significant, so there is no point in carrying the analysis further by including a cubic or higher order terms. The polynomial equation to describe the relationship between turbidity and distance from the inflow is given by:

$$Turbidity = 41.42 - 5.3211.Distance + 0.2596.Distance^2$$

The coefficient of determination is:

$$R^2 = \frac{SS_{regression}}{SS_{among}} = \frac{4812.73 + 1052.15}{6025.04} = \frac{5864.88}{6025.04} = 0.9734$$

So 97.3% of variation in mean turbidity among sites can now be explained by a regression against distance from the inflow, a substantial (and significant) improvement on the 79.0% that could be explained by the linear regression. The regression is plotted in Figure 6-10. It is a descriptive model only, useful for prediction, but not necessarily giving any insight into a functional relationship between turbidity and distance.

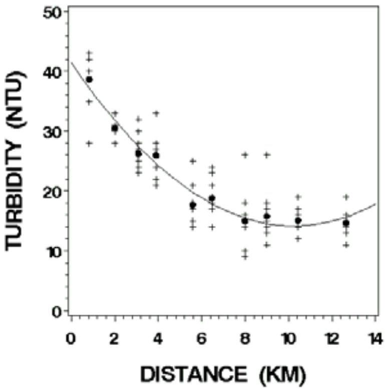


Figure 6-10 A plot of turbidity against distance from the inflow of Lake Burley Griffin, Canberra. Both individual data points (+) and means (•) are shown. The line is for the least squares quadratic regression.

Other influential regressors

The second way that a significant residual can arise is if there is a second regressor (or discrete factor), uncorrelated or poorly correlated with our regressor, that is pulling the site means away from our underlying linear relationship.

For example, distance from the shoreline may influence turbidity if wind action is influential in re-suspending the sediments. Distance from the shoreline is not particularly well correlated with distance from the inflow (Figure 6-5). We might wish to include this second regressor in the analysis. Here we first fit the regressor distance from outflow, then we ask how much of the residual variation can be explained by the second regressor, distance from the shoreline. This approach takes us into the domain of **multiple regression**. The ANOVA table is shown in Table 6-6.

Table 6-6. Results of a multiple regression of turbidity (NTU) versus distance in km from the inflow (DISTLONG) and distance in metres from the shoreline (DISTLAT) for Lake Burley Griffin, Canberra.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Among Sites	9	6025.04	669.4489	52.62	0.0001
DISTLONG	1	4812.73	4812.730	36.54	0.0005
DISTLAT	1	250.6421	250.6421	1.82	0.2188
Residual	7	961.668	137.3812	10.80	0.0001
Within	90	1145.00	12.72222		
Total	99	7170.04			

Clearly, the second regressor does not provide an explanation for the significant residual variation in the earlier analysis (F=1.82; df=1,7; p=0.2188). That is, distance from the shoreline does not explain any significant variation in turbidity over and above what was already explained by distance from the inflow. The residual variance remains significant (F=10.80; df=7,90; p<0.0001).

Reporting the results

We would report the results of the overall analysis as follows:

Mean turbidity differed among sites chosen at intervals along the drainage channel in Lake Burley Griffin (F=52.62; df=9,90; p<0.0001).

Highest turbidity of 38.7 ntu occurred at Site A adjacent to the inflow of Molonglo Creek (Tukey-Kramer Procedure, p<0.05). The six sites above the outflow at Scrivener Dam had the lowest turbidity averaging 16.2 ntu, with the lowest recorded turbidity of 14.7 ntu at Site J adjacent to the Scrivener Dam wall. Sites B, C and D were intermediate in turbidity.

A total of 79% of variation in turbidity could be explained by a significant linear regression of turbidity against distance from the inflow of the Molonglo River (F=31.76; df=1,8; p<0.0005) described by the formula:

$$Turbidity = 33.66 - 1.905.Distance$$

$$R^2 = 0.79$$

where turbidity is in ntu and distance is in km. Some variation remained unaccounted for by the regression (F=11.91; df=8,90; p<0.0001), and a plot of turbidity against distance revealed that this was probably a reflection of curvilinearity in the relationship between the two variables.

We could go on to report the results of the polynomial regression or the multiple regression, but these analyses are beyond the scope of this introductory treatment.

Note that in this description, we have:

- **described the trends** in the data, including a summary of the magnitude of the response variable across the sites, and specification of the linear regression equation *with units of measurement for both variables*;
- provided an indication of the **significance of the results**, including significance of the differences among sites (Tukey-Kramer Procedure) and significance of the regression and residual (ANOVA results);
- provided an indication of the **strength of the regression** (1.905 ntu reduction for every km down the channel);
- provided an indication of the **adequacy of fit** ($R^2 = 0.79$), which is poor in the context of this problem, and reasons why it is so poor.

Special Case: Simple linear regression

It is time now to come back down to earth and the topic of this Workbook. Regression with more than one Y value for each value of X, although an elegant and powerful analysis, is rarely undertaken. More often, we have only one value of Y for each X. What happens to the analysis when we have no replication? What functionality is lost, and what functionality is retained?

Consider the analysis of variance table that would arise had Kurt Hammerschmidt collected not 10 but one sample of water at each site (Table 6-7). With only one value per site, there is no estimate of the within-site variance, so that line in the ANOVA table would contain no useful information. The total sums of squares would now be equal to the sums of squares among sites (with n=1), so the *Among Sites* entry and the *Total* entry would become one and the same.

It would not be possible to conduct any tests with MS_{within} as the error term, so tests of variation among sites and of the residual would no longer be possible. However, a test of the regression would still be possible as the residual mean square is the error term for that test. Hence, in the simplified analysis with only one Y for each X, we can still test for a significant regression.

Table 6-7. Results of a regression of turbidity (NTU) versus distance in km from the inflow for Lake Burley Griffin, Canberra. Only one sample is collected from each of 10 sites,

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Among Sites	9	6025.04	669.4489	.	.
Regression	1	4812.73	4812.730	31.76	0.0005
Residual	8	1212.31	151.5388	.	.
Within	0	.	.		
Total	9	6025.04			

so n=1. Note how the ANOVA table collapses to a much simpler form.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Regression	1	4812.73	4812.730	31.76	0.0005
Residual	8	1212.31	151.5388	.	.
Total	9	6025.04			

Source	Expected MS
Among Sites	$\sigma^2 + n\sigma_A^2$
Regression	$\sigma^2 + n\sigma_{res}^2 + n\sigma_{reg}^2$
Residual	$\sigma^2 + n\sigma_{res}^2$
Within	—

A summary of the results for this simplified analysis might read as follows:

Turbidity declined progressively from 38.7 ntu at the inflow from Molonglo River to 14.7 ntu at the outflow near the Scrivener Dam wall. A linear regression of turbidity against distance from the inflow was significant (F=31.76; df=1,8; p<0.0005) and could be described by the formula:

$$Turbidity = 33.66 - 1.905.Distance$$

$$R^2 = 0.79$$

where turbidity is in ntu and distance is in km. However a plot of turbidity against distance from the inflow showed distinctive curvilinearity (Figure 6-12), and the linear model was clearly a poor description of the variation and inadequate for predictive purposes. Modelling the trend with curvilinear models should be explored.'

The ANOVA table can be included in your report or publication, but it is not customary.

Notice how much weaker this statement is than the equivalent statement for the analysis with more than one *Y* for each *X*. This is the cost of failing to replicate, but one that most investigators are willing to wear in the case of simple linear regression.

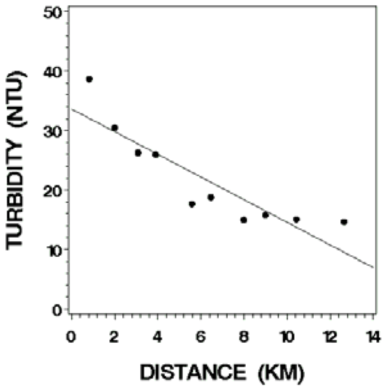


Figure 6-12. A plot of turbidity against distance from the inflow of Lake Burley Griffin, Canberra. One sample is taken from each of 10 sites. The line is the least squares regression.

Where have we come?

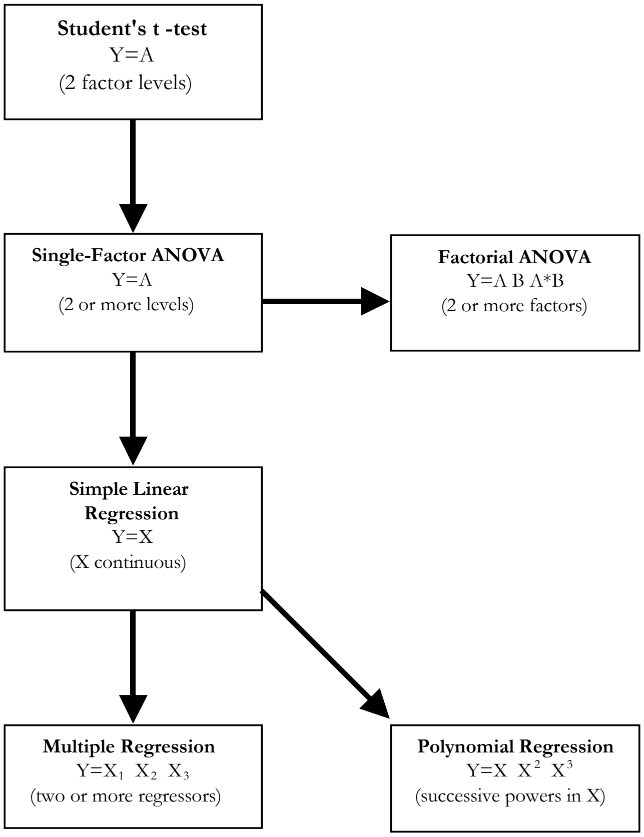
The major message of this lesson is to establish that simple linear regression can be viewed as a natural extension of single-factor ANOVA. Regression and ANOVA are just two forms of a more general class of analyses, called general linear models (GLM). The take home messages from this Lesson are:

- The full analysis involves more than one value of Y for each value of X. In this case, the values of the X variable can be considered levels of a discrete factor, and a single factor ANOVA can be used to explore if there is significant variation among the means. No particular relationship between the means and the X variable is assumed.
- Where there is a significant difference among the means, we can further partition SS_{among} into a component explained by a linear regression and a residual component.
- Where the residual means square is significant, two possible explanations present themselves. One possibility is that the linear model is inappropriate. We need to explore curvilinear models or transformations to render the linear model appropriate.
- The second possibility is that a second regressor (or factor), uncorrelated or poorly correlated with the first, is pulling the points away from the underlying linear relationship with the first regressor and inflating the residual variance. We can account for the effects of such a second regressor using multiple regression.
- Simple linear regression is the case where we have only one value of Y for each X, and this leads to a much simplified ANOVA Table.

A diagram showing where we have come since is shown in Figure 6-11. We link to polynomial regression and multiple regression, but they are beyond the scope of this Module.

We need now to consider some practicalities of regression analysis, when it is applied to real world data.

Figure 6-11. Diagrammatic representation of how far we have come in developing a general framework of linear models.



Lesson 3: Application Notes

Assumptions of regression

Because regression is a natural extension of the fixed model single-factor ANOVA, it should not come as a surprise to discover that regression and fixed model ANOVA share many assumptions.

- **Randomness:** that the entities subject to measurement have been allocated or selected at random.
- **Independence:** that we have achieved independence in sampling, that is, the value of any one measurement has no bearing on the value of any other, relative to the predicted value of the underlying model.
- **Normality:** that the Y values for each X are drawn from a population with a normal distribution.
- **Homogeneity of variances:** that the variances σ^2 of the underlying distributions of Y for each X are equal across the range of X.
- **Linearity:** that the true population means of Y for each X lie on the regression line.

The analysis is derived from a fixed factor ANOVA, so we have the additional constraint that **X is fixed** — the regressor X takes on fixed values that are fully under the control of the investigator. Any error is derived from uncertainty in Y.

Violations of these assumptions involve trade-offs for the interpretation of regression analysis. We need to ascertain whether the assumptions are met before embarking on an analysis, and if they are not, we need to take steps to ensure that they are met.

Below, we look in more detail at the assumptions of regression, at how to check if they are reasonable, and at how to proceed in the face of perceived violations.

X values fixed

Often it is possible to specifically select the values of the regressor in designing a study. For example in manipulative experimental studies, the fire frequency can be manipulated in plots established at a landscape scale and used in a regression of floristic attributes against fire frequency.

Even if the plots are not selected at random and manipulated, it is still possible to control the values of the regressor variable. Fire may not be manipulated, but rather plots selected systematically on the basis of their fire history, and a regression developed between bristle bird abundance and fire frequency. Because the plots have been selected specifically on the basis of particular fire frequency, the regressor is fully under the control of the investigator.

However, regression analyses are frequently applied in observational studies, where the values of both response and regressor variables are drawn at random. Alternatively, the regressor may be measured with non-negligible error. For example, we may wish to establish a predictive regression between fish length and fish weight. Fish are selected at random from the population, weighed and measured. Both the regressor (length) and the response variable (weight) are subject to sampling error and length is measured with non-negligible error. How does this violation of theory compromise the analysis?

The subject is one on which research and controversy are continuing, so a definitive answer is difficult to provide.

The first thing to consider is whether you really need a regression analysis. If your interest is solely on testing whether two variables co-vary, without any interest in quantifying a presumed linear relationship, then a correlation analysis is appropriate (refer to Section 15.1 of Sokal and Rohlf, 1994, for a comparison of correlation and regression). If prediction is your goal, then proceed with the regression.

If we treat *X* as fixed when in fact it is measured with error, and *X* and *Y* are related, then we will have in effect artificially inflated the variance of *Y* for a given *X*. All of the error variance is assigned to the response variable, including that variation in *Y* that can be attributed to random shifts in *X*. The inflated residual variance will reduce the power of our regression, and in marginal cases, lead to a failure to detect an underlying true regression—undesirable but not fatal.

The advice of Neter et al (1996) is that all results on estimation, testing and prediction still apply to regression where *X* is random, provided we make an additional assumptions about the distribution of *X*. The values of *X* must be independent and random with a probability distribution that does not involve the parameters β_0, β_1 or σ^2.

Most people therefore are willing to apply regression techniques to data where *X* is not fixed. We will adopt this approach in this Module.

Randomness in sampling

As with any statistical inference, it is important that the data at hand are representative of the population(s) from which they are drawn. In that we have fixed values of *X*, regression assumes that the items, individuals or entities allocated to each of value of *X* are done so at random. It is important that the only systematic difference between them, if any, is attributable to the differential effect of the regressor. Non-randomness may manifest itself as lack of independence of the entities, unequal variances, non-normality or non-linearity.

Violation of the assumption of randomness in sampling cannot be overcome easily, and typically the data must be discarded, the sampling protocols redesigned and the data recollected. Adequate attention must be paid at the time of designing an experiment, or when sampling from natural populations, to ensure random sampling.

Linearity

The existence of a true underlying linear relationship among the means of Y for each value of X is fundamental to the analysis. In the analysis with more than one value of Y for each X we can test this assumption by asking if there is variation in the deviations of the means for each X from the line, over and above what would be expected by chance. We can test the residual variation for significance.

If we restrict our attention to simple linear regression with one Y for each X, then we can look for violations of this assumption in the distribution of residuals (see Figure 6-9 for an example).

Violations of this assumption will result in an inflation of the $MS_{residual}$, and so greatly reduce the power of the test to detect a significant regression.

Independence

The assumption of independence refers to independence in the error term, and an error term is defined in the context of an underlying model. It makes no sense to ask, ‘Are the measurements of platypus bill length independent?’—independent with respect to what underlying model? Clearly they are not independent entirely, as the mere fact that they are taken from the same species provides some indication of their collective magnitude. Measure the first few, and you have general idea of the magnitude of subsequent measurements.

When we ask, ‘Are the measurements of platypus bill length independent?’, we are asking if they are independent with respect to the mean. Does information on the length of one platypus bill *relative to the mean* provide information on the length of any other platypus bill *relative to the mean*? If we have siblings in our sample, and the bill of one sibling is longer than the average, then there is a high probability that the bill of the other sibling will be longer than average. The error terms for the bills of the two siblings will be correlated, and the assumption of independence violated.

In the context of regression, our underlying model is not the mean but the regression equation. Bill length in a platypus may regress with age. The question of independence relates to whether information on the length of one platypus bill *with respect to the line*, provides any information on the length of any other bill, *with respect to the line*. We will have the same problem of independence with our siblings.

Lack of independence manifests itself in a particularly destructive way when it artificially deflates the mean square error in the F ratios of the regression ANOVA table. The consequences are dire because the analysis will yield significant results apparently at the 0.05 level of significance when in fact the real probability of Type I error may be 0.15, 0.20 or worse. The analysis will yield significant results without foundation. This is most undesirable.

In regression with more than one Y for each value of X, lack of independence among the Y values for a given X may deflate the MS_{within}, which is used as the error mean square in tests of the residual and overall variation among means. We may be led to believe that the regressor does not explain all of the significant variation among the means when in fact it does. This may send us down the path of searching inappropriately for a curvilinear model or additional regressors. The test of regression is not affected by this violation of independence.

Lack of independence in the mean values of Y for a given X, or in the values of Y if there is only one for each X, can lead to deflation of the MS_{residu}al. A common scenario is where the values of Y are serially dependent in time, if X is a time related variable, or in space, if X is a spatial variable. Serial time dependence means that if a value of Y is particularly high with respect to the regression line, then the next value is also likely to be high. The variance in Y values about the line is constrained by this dependence, and so artificially deflated. Regressions that would not otherwise be significant will become so, in the absence of a true underlying relationship. Certainly, given the objective of undertaking statistical analysis, this is an intolerable violation of the underlying basis of the technique.

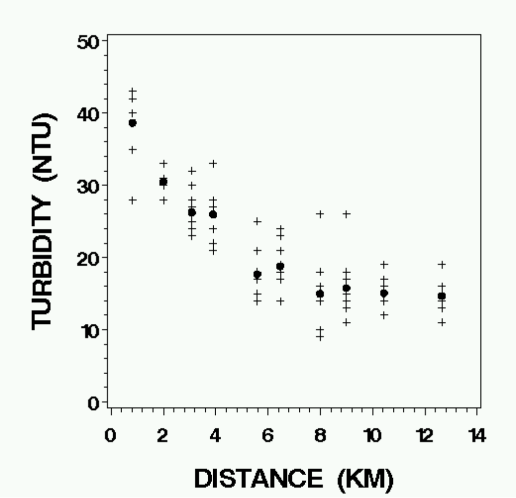
A residual analysis, described in a later section, will reveal cases of serial dependence.

Equality of variances

In developing the rationale of single-factor ANOVA, it was argued that the effect of the factor across samples should act differentially to increase or decrease the sample means, but not to differentially alter the sample variances. An assumption of ANOVA is that the individual sample variances for each factor level estimate a common population variance, that is, that the population variances are equal. This assumption carries through to regression.

Where we have more than one value of Y for each value of X, this assumption is easy to visualise (Figure 6-13). We must be willing to accept that the variation in Y for each value of X, $S^2_{Y|X}$, is estimating a population variance, $\sigma^2_{Y|X}$ common to all X. Perusing the data for turbidity in Lake Burley Griffin would suggest that this is a reasonable assumption.

Figure 6-13. A plot of turbidity against distance from the inflow of Lake Burley Griffin, Canberra. Both individual data points (+) and means (•) are shown.



Where there is one value of Y for each X, it is not possible to assess the validity of this assumption directly. However, if the conditional means $\mu_{Y|X}$ all lie on the parametric regression line, as we require under the assumption of linearity, then the variation among conditional means, $\overline{Y}_{Y|X}$, will be driven entirely by variation in Y for a given X. We can look at the scatter of the means about the regression line to assess whether the variation in Y values is estimating a common population variance. This will work even when there is only one Y for each X (means with n=1).

In regression with only one value of Y for each X, testing the assumption of homogeneity of variances becomes a check of whether there is an even spread in the scatter of points about the regression line. Again, this can be done with an analysis of residuals, as we will see later.

Normality

A single-factor ANOVA assumes that the individual measurements in each sample are normally distributed about the true sample mean. For regression, the assumption is that the Y values for each X are drawn from a normal distribution (Figure 6-14). Again, if the linearity assumption is met, then the mean values of Y across the range of X will be normally distributed. We can examine the residuals to verify this assumption.

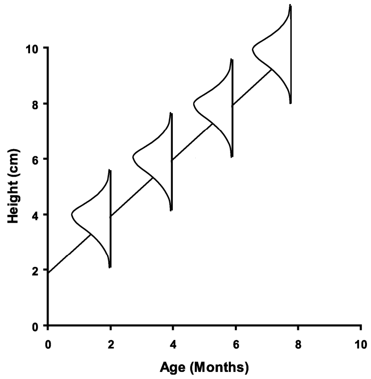


Figure 6-14. An assumption of regression is that, for each given value of X, repeated measurements of Y will be drawn from a normal distribution.

Pooling the residuals for a test of normality is only valid provided there is no evidence of a departure from the assumptions of linearity or homogeneity of variances.

Analysis of residuals

Residual analysis is a graphical approach to diagnosing violations of the assumptions in regression. This approach does not require hypothesis testing, but does require some experience on the part of the researcher.

The ideal residual plot has the points distributed at random with respect to the horizontal reference line across the full domain of the regressor (Figure 6-15).

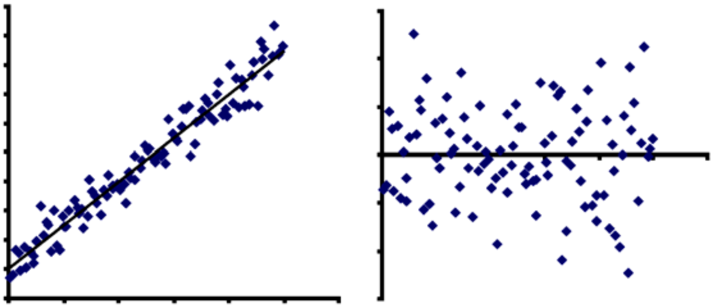


Figure 6-15. A regression with the ideal distribution of residuals—the residuals are distributed at random with respect to the horizontal regression line, with no systematic trend evident.

Residual analysis can detect five important departures from the underlying assumptions of the simple linear regression model:

- Non-linearity in the regression function.
- Heterogeneity of the error variance across the range of *X*, including the presence of aberrant points (outliers).
- Lack of independence of the errors (deviations of the points from the line).
- Non-normality of the distribution of errors.

Non-linearity in the regression function

Take again the example of turbidity in Lake Burley Griffin, where only one sample of water was collected from each site. In Figure 6-16, we have a scatterplot of the data with a simple linear regression line fitted. To get a simple residual plot, we take the regression line to be our X-axis reference line, and show only the deviations of the Y values from the line (Figure 6-16). The curvilinearity is more strongly evident in the residual plot, which is one reason why such plots are used.

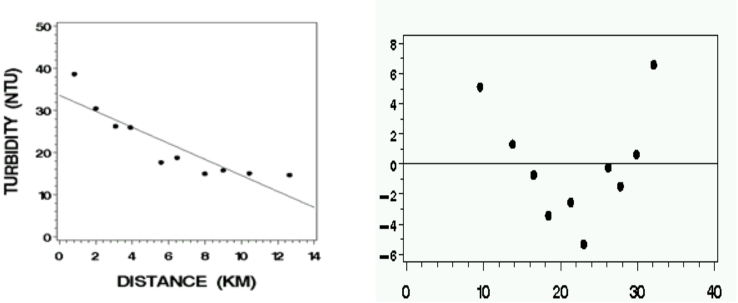


Figure 6-16. A scatterplot and regression of turbidity against distance from the inflow of Lake Burley Griffin, Canberra (left) and a residual plot of the same data (right). Note that the curvilinearity is more strongly evident in the residual plot.

A linear model is clearly poorly specified. If we now look at the residuals following a quadratic regression (Figure 6-17), we see a much more satisfactory spread of residuals. The objective of residual analysis is to have the residuals fall with even spread in a horizontal band centred on zero, with no systematic tendencies to be positive or negative.

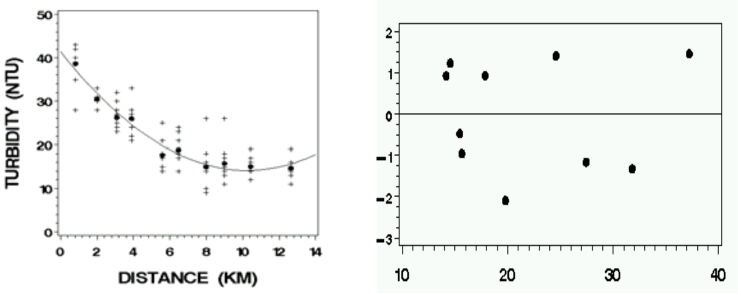


Figure 6-17. A scatterplot and polynomial regression of turbidity against distance from the inflow of Lake Burley Griffin, Canberra (left) and a residual plot of the same data (right).

We have several options to respond to curvilinearity:

a. Identify a theoretical underlying curvilinear function and select a transformation that will linearise the function.

b. Fit the theoretical underlying curvilinear function using non-linear iterative techniques (eg `nls()` in R).

Both options will yield an estimate of the functional relationship between the response variable and regressor, as well as providing a foundation for prediction. Option (b) may be preferred over option (a) depending on the effect of transformation on other assumptions (normality, homogeneity of variances).

c. Select a transformation that yields adequate residual plot, but which has no foundation in theory.

d. Apply a generic polynomial model, as we did in the turbidity example.

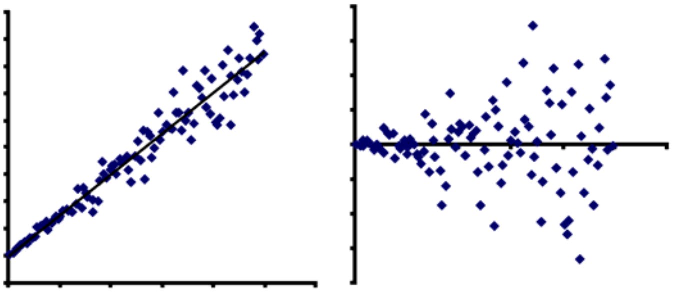
Both of these options will yield a regression function that is an adequate statistical description of the data useful for prediction, but provides little insight into the form of the underlying functional relationship.

Only options (a) and (c) are covered by this Workbook. They will be dealt with in the worked examples and exercises.

Heterogeneity of the error variance

Heterogeneity of the error variance is evident in a residual plot as variation in the spread of values across the domain the residual plot (usually $\hat{Y}$) (Figure 6-18).

Figure 6-18. Regression where the errors are correlated with X.



A common instance of this is when the variance in Y and the X are correlated. Variation in body weight, for example, depends upon body size, so we would expect heterogeneity in the error variance for a length-weight relationship. One can also encounter error variances that decrease with increasing values of the regressor X or sometimes varying in more complex fashion.

The solution of heterogeneous variances is typically transformation, and this option will be demonstrated in the worked examples. Common transformations used for this purpose are:

$$Y' = Log_{10}(Y)$$

$$Y' = \sqrt{Y}$$

In cases where there is more than one value of *Y* for each *X*, it is possible to estimate the variance of *Y* for each *X*. We can then weight the observations, giving less weight to those observations that are more variable. The optimal weight to use for each observation is:

$$W = \frac{1}{S^2_{Y|X}} \sim \frac{1}{\sigma^2_{Y|X}}$$

This is an alternative to transformation that can be used in experimental situations where we can design the data collection to obtain estimates of $\sigma^2_{Y|X}$.

Lack of independence of the errors

Whenever the data are collected in a time sequence or in a spatial sequence, such as for adjacent plots, it is sensible to prepare a **sequential plot of residuals**. Sometimes the regressor *X* is a temporal or spatial variable.

A sequential plot of residuals will reveal if there is any correlation between the error terms for adjacent observations in time or space. Is there any indication that the magnitude or sign of the residual of one observation in the series has an influence on the magnitude or sign of the residual of the next? The **autocorrelation** can be positive, as in Figure 6-19, or more rarely, negative.

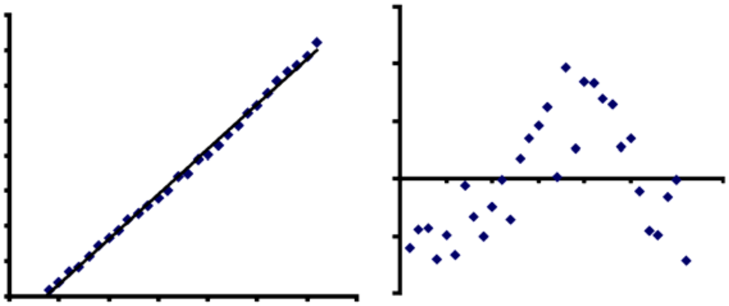


Figure 6-19. Linear regression where the regressor is Time and the errors in Y are serially dependent. Note the progression of the residuals.

The consequences of correlated errors has been discussed earlier, namely deflation, or more rarely inflation, of the error term in the test of significance of the regression coefficient. Deflation leads to significant results that would not otherwise occur, and so is a serious violation of the assumptions of regression.

First order autocorrelation is where the value of *Y* at time *t* is influenced only by the value of *Y* immediately preceding and the regressor *X*. This can be overcome by applying a first-order autocorrelative model defined by:

$$Y_t = \beta_0 + \beta_1 X_t - \beta_2 Y_{t-1} + \varepsilon_t$$

Note here that *Y* at time *t* is determined by *Y* at time *t-1* and *X*. We can convert to this autoregressive model from our standard regression model:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

by subtracting $\beta_2 Y_{t-1}$ from both sides of the equation.

$$Y_t - \beta_2 Y_{t-1} = \beta_0 + \beta_1 X_t - \beta_2 Y_{t-1} + \varepsilon_t$$

Setting $Y_t^* = Y_t - \beta_2 Y_{t-1}$ we have:

$$Y_t^* = \beta_0 + \beta_1 X_t + \varepsilon_t - \beta_2 \left(\beta_0 + \beta_1 X_{t-1} + \varepsilon_{t-1}\right)$$

$$= \beta_0 \left(1 - \beta_2\right) + \beta_1 \left(X_t - \beta_2 X_{t-2}\right) + \left(\varepsilon_t - \beta_2 \varepsilon_{t-1}\right)$$

Setting $X_t^* = X_t - \beta_2 X_{t-1}$, we have:

$$Y_t^* = \beta_0^* + \beta_1 X_t^* + \varepsilon_t^*$$

which is conveniently in linear form. Unbiased estimates of the parameters β_0^* and β_1 can be obtained by linear regression. β_0 can be back-calculated from β_0^*. An estimate of β_2 used in the transformation of *X* can be obtained by regressing Y_t against Y_{t-1}.

A more thorough treatment of **time series regression** can be obtained from the text by Box and Jenkins (1976).

Presence of aberrant points (outliers)

Aberrant points can have a profound effect on the final solution for the regression equation. They also affect the residual variance and therefore the outcome of significance testing in regression.

The effect of a single aberrant point with leverage (some distance from the bivariate mean) on the regression line is shown in Figure 6-20. The aberrant point is circled. Note the dramatic effect on both the least squares solution for the regression and on the residual plot.

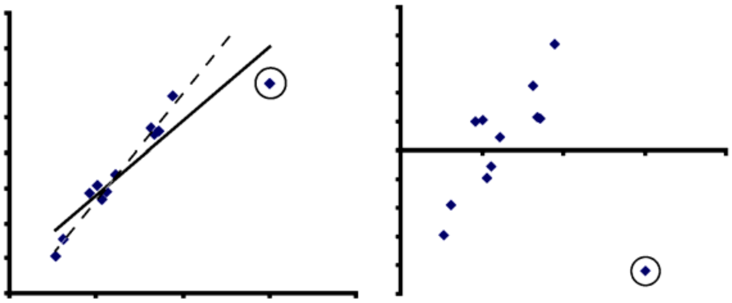


Figure 6-20. A regression where a single outlier (circled) exerts undue influence on the outcome. The dashed line shows the regression in the absence of the outlier. Note the impact on the regression line itself and the dramatic effect on the residuals.

This sensitivity presents the researcher with one of the greatest dilemmas in regression analysis. The aberrant points may arise for a number of reasons. They may be suspect, having been mistyped or recorded incorrectly at the time of measurement. It could be that the entity under measurement is itself aberrant, having come from a dirty test tube or deformed animal, or that some unmeasured regressor is differentially influencing the aberrant data point. Or perhaps the error variance in *Y* is greater for the value of *X* corresponding to the aberrant point than for other values of *X*.

Unfortunately, it could also be that the point is not biased at all, or that the error variance is homogeneous, and that the large error associated with the aberrant point occurred by chance alone. Elimination of outliers from a dataset in order to improve the fit of the regression model is therefore problematic. First port of call is to check the original data sheets, and if possible to re-measure the suspect animal, plant or entity. However, there is often no clear-cut reason to believe that a suspect point should be deleted, and the researcher must use judgment, guided by some form of statistical assessment of the suspect point's probability of occurrence.

The traditional tool for detecting outliers is again an examination of residuals. Instead of considering raw residuals, it is possible to compute standard errors of the residuals and scale the *Y*-axis on the residual plot in terms of units of standard error. Such residuals are called **studentised residuals**. When the error degrees of freedom for the regression exceed 10, a studentised residual of 2.5 or greater is rare, and so forms the basis for identifying outliers. A larger cut-off may be chosen if the number of data points is large.

A plot of the studentised residuals for the data shown in Figure 6-20 is presented in Figure 6-21. Clearly, the suspected outlier (circled) is the only point that is greater than 2.5 standard errors from expected, and so is a candidate for greater scrutiny or omission.

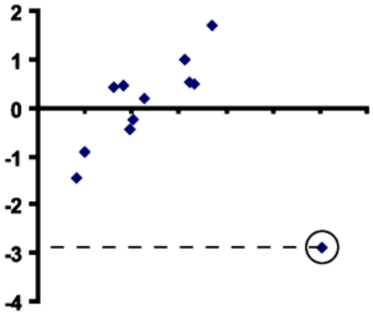


Figure 6-21. A plot of studentised residuals used for the identification of outliers. The circled point is clearly an outlier as it is more than 2.5 standard errors from expected.

A difficulty with the use of studentised residuals for detecting outliers is that the outliers themselves influence the position of the regression line against which they are being assessed. This is clearly evident in Figure 6-20. As a consequence, many researchers prefer to use what are referred to as **influence statistics** in making their judgments.

Residuals are calculated for each point after omitting that point from calculations leading to the regression equation (Figure 6-22). You can see how much more sensitive this approach is for detecting outliers. The cut-off of ≥ 2.5 standard errors still applies.

We would omit this outlier as aberrant, and restrict our predictions from the regression to the domain of the remaining points.

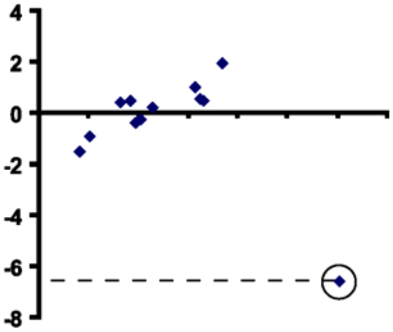


Figure 6-22. A plot of studentised residuals where each point is assessed against the regression calculated with the point under scrutiny omitted. The circled point is clearly an outlier as it is more than 2.5 standard errors from expected.

Non-normality in the distribution of errors

The distribution of residuals when pooled will be influenced by:

- curvilinearity which adds a systematic element to their distribution;
- heterogeneity of the variances which can lead to kurtosis in their distribution;
- the presence of outliers; and
- non-normality in the errors in *Y* for a given value of *X*.

Hence, using residuals for detecting deviations from the assumption of normality of the errors in *Y* for a given *X* is a staged process. We must first ensure that the linear assumption is satisfied, then ensure that the variances are homogeneous, then pool the residuals for a check on the assumption of normality.

The data of Figure 6-15 clearly meet the assumptions of linearity, homogeneity of variances, and there are no obvious outliers, so we can pool the residuals as in Figure 6-23. Clearly, on perusing the histogram, there are no substantive departures from normality, a conclusion we could reinforce with a probability plot or other diagnostics (refer to Module 2).

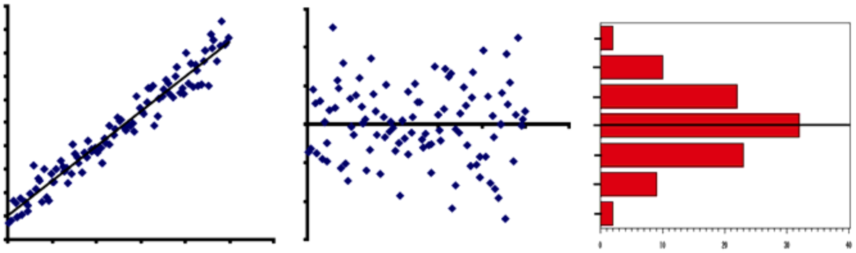


Figure 6-23. A scatterplot, plot of residuals and histogram of residuals as a check on the assumption of normality.

If the errors are not normal, or at least approximately so, remedial action is advised. The most common approach is to transform the *Y* variable using a transformation appropriate to the violation of the normality assumption. Such transformations were introduced in Module 3. Following transformation, the regression analysis should be repeated, and the residuals examined a second time.

As an alternative to transformation, theory has advanced to the stage where regression models have been developed for a range of underlying distributions for the error term—binomial, Poisson, negative-binomial, etc. Referred to as Generalised Linear Models (GLIMs), they are particularly useful in cases where no transformation can normalise the errors (such as Poisson counts of rare events with lots of zeros) or where transformation to rectify the error structure leads to secondary violations of other assumptions.

GLIMs are beyond the scope of this Module.

Caveats on the use of transformations

Transformations were recommended as a method of linearising data as a prelude to regression, to homogenise the variances, to deal with serial dependence of the errors, and to render the errors normal. Transformation to linearise the relationship between two variables as a prelude to regression will implicitly alter the error structure. If the error structure was in good shape before the transformation, it is unlikely to be so after the transformation.

The researcher must learn to draw from the suite of regression options in order to adequately meet the requirements of the various assumptions. Transformation is not always appropriate. The options include:

- Transformation of *X* or *Y* or both to address one or more of the assumptions of linearity, normality of errors, heterogeneity of variances, and independence.
- Use of iterative non-linear least-squares techniques where there is clear evidence of a curvilinear relationship but where the error structure is appropriate (eg `nls()`in R).
- Use of weighted regression to address violations of the assumption of homogeneity of variances, as an alternative to transformation, where such transformation is likely to produce a second violation.
- Use of Generalised Linear Models (GLIMs, distinctive from General Linear Models or GLMs) to cater for non-normal distributions of the error terms (eg or `glm()` in R).

Caveats on the use of R^2

The coefficient of determination, R^2, provides an indication of the proportion of total variation in the response variable *Y* that can be explained by the regressor *X*. It gives an indication of adequacy of fit, of scatter about the regression line.

This said, R^2 is not estimating a true parametric value and so in some respects is subjective. For a given scatter about the regression line, R^2 can be made arbitrarily high by simply increasing the range of *X* subject to experimentation.

Failure to appreciate these qualities of R^2 has led to many misunderstandings.

Where have we come?

You should now have a grasp of another of the most important statistical concepts of use to biologists—simple linear regression. Its fundamental objective is to estimate a linear functional relationship between two variables or to formulate a linear model useful for predicting one variable from the other.

Linear regression can be seen as a natural extension of single-factor ANOVA, where the discrete factor *A* of the ANOVA is converted to a continuous regressor *X*. From this perspective, the traditional approach of testing the regression coefficient with a student's t-test is replaced by testing the regression coefficient in an ANOVA table.

Where there is more than one value of the response variable *Y* for a given value of the regressor *X*, a full ANOVA is possible including a test of the residual variance remaining after the regression has been fitted.

The many assumptions of simple linear regression were presented together with an introduction to analysis of residuals in regression. Analysis of residuals provides a graphic way of verifying that the assumptions of regression are satisfied, and checking the effectiveness of remedial measures taken when violations are detected.

Although the workbook focuses on simple linear regression, links were made to other very important analyses, including polynomial regression, multiple regression, time series analysis, generalised linear models (GLMs) and iterative curvilinear regression. It is important that you see simple linear regression as embedded in a broader class of linear models.

Key concepts with which you need to be broadly familiar include:

- an intuitive meaning for the regression coefficient, the *Y* intercept and the coefficient of determination R^2;
- the distinctive meanings of MS_{within}, MS_{among}, $MS_{regression}$, $MS_{residual}$, their sums of squares and degrees of freedom, and the various *F* ratios used for testing in regression;
- the assumptions of simple linear regression, how to detect violations and how to overcome them, with emphasis on displaying and interpreting residuals.

It is now appropriate to put this knowledge to use via worked examples and exercises.

Lesson 4: Step-through Examples

Example 6-1: Time to pipping in the pig-nosed turtle

This is a simple linear regression provided as an elementary example.

The problem

Laboratory studies of the embryonic development of the pig-nosed turtle revealed that the embryos develop rapidly then, when fully formed, enter diapause—oxygen consumption drops precipitously and the embryos become torpid (Webb et al, 1986). Hatching of diapausal eggs could be stimulated by immersion in water or by displacing the air around the eggs with nitrogen. This strongly suggested to the researchers that depletion of oxygen in nests following early wet-season inundation of torrential rain was the stimulus required for successful hatching.

The eggs of pig-nosed turtles are hard-shelled, so dehydration causes an air pocket to form between the eggshell and the underlying egg membranes. The air pocket of dehydrated eggs could be expected to sustain the embryo for longer following immersion in water compared to turgid eggs. An experiment was mounted to test this.

Eggs were weighed before and after incubation on wire racks at 30^{o}C and a high but unmeasured humidity. Eggs incubated under these conditions progressively dehydrated by amounts not under the control of the experimenter. When the eggs had reached full term, they were immersed in water at 30^{o}C to stimulate hatching. Just before hatching, the hatchlings pip the egg—a small egg tooth on the tip of the snout is used to break the eggshell. The time to pipping after immersion was recorded.

The data

Table 6-8. Percentage weight loss and time to pipping in minutes for eggs of the pig-nosed turtle from the wet-dry tropics of northern Australia.

% Wt Loss	Time	% Wt Loss	Time	% Wt Loss	Time	% Wt Loss	Time
21.10	41.00	12.36	32.66	23.51	53.02	8.72	17.92
21.12	59.33	12.78	32.66	20.15	49.83	7.73	26.33
15.83	42.33	14.65	39.66	16.66	44.83	8.32	9.38
17.62	42.70	15.89	32.66	11.03	27.28	7.72	12.55
17.53	43.34	23.65	55.91	5.86	16.17		

The data are held in a file called PIPPING.DAT as two columns, percent dehydration (`weight`) and time to pipping (`time`) (Table 6-8).

The Analysis

Double click on the Tinn-R icon and launch R from within Tinn-R (Click in the Menu on R->Initiate/Close Rgui->Initiate preferred Rgui)

The first step in the analysis is to read the data in to a data frame called `turtle`, quickly peruse it and compute some basis statistics.

```
> setwd("H:\\Biometry\data")
> turtle <- read.table("PIPPING.DAT",header=T, sep="",
na.strings=".",  header=FALSE)# Read in data
> turtle

   weight  time
1   21.10 41.00
2   21.12 59.33
3   15.83 42.33
4   17.62 42.70
5   17.53 43.34
6   12.36 32.66
7   12.78 32.66
8   14.65 39.66
9   15.89 32.66
10  23.65 55.91
11  23.51 53.02
12  20.15 49.83
13  16.66 44.83
14  11.03 27.28
15   5.86 16.17
16   8.72 17.92
17   7.73 26.33
18   8.32  9.38
19   7.72 12.55

> dim(turtle)

[1] 19  2

> names(turtle)

[1] "weight" "time"
```

Submit the above program for execution.

The resulting data frame turtle should contain two variables—`weight` and `time`. You can peruse the data at this point to see if it has been read as intended.

Look at some simple summary statistics for the variables contained within the data frame `turtle`.

```
> summary(turtle)
```

Box 6-1. Summary statistics for percentage dehydration and time to pipping for the pig-nosed turtle.

```
   weight           time      
Min.   : 5.860   Min.   : 9.38  
1st Qu.: 9.875   1st Qu.:26.80  
Median :15.830   Median :39.66  
Mean   :14.854   Mean   :35.77  
3rd Qu.:18.885   3rd Qu.:44.09  
Max.   :23.650   Max.   :59.33  
```

Percentage dehydration ranges from 5.86% to 23.65% by weight. Time to pipping ranges from 9.38 minutes to 59.33 minutes.

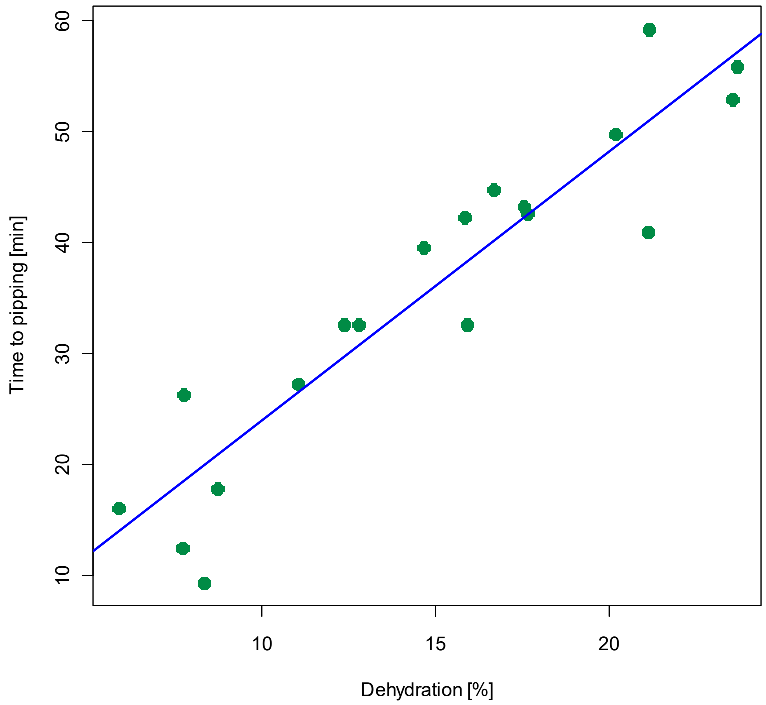
The next step in the analysis is to plot the data to see if there is any indication of a relationship between the two variables. For revision we use several versions of the plot commands to improve the plot. Feel free to use different colors, symbols, labels (Figure 6.24, note the regression line will be added later).

```
> plot(turtle$time, turtle$weight)
> plot(turtle$time, turtle$weight, pch=16)
> plot(turtle$time, turtle$weight, pch=16,
col="springgreen4") #nice filled coloured dots
> plot(turtle$weight, turtle$time, pch=16,
col="springgreen4", xlab="Dehydration [%]", ylab="Time to
pipping [min]") #nice filled coloured dots and lables
```

Submit the above program for execution.

Try to create a plot the ranges from zero to `max(time)` on the x-axis and zero to `max(weight)` on the y-axis (see `?xlim, ?ylim`)

Figure 6-24. A plot of time to pipping (minutes) against percentage dehydration by weight for the pig-nosed turtle (Carettochelys insculpta).



Regression analysis

Perusing the data of Figure 6-24 suggests that a linear model is quite appropriate. To create a simple linear model in R we will use the lm() function just as in the ANOVA case in R module 5. We want to do a regression of weight against time, so our response is time.

```
> lm.turtle <- lm(time ~ weight, data=turtle)
> lm.turtle #brings you the regression coefficients,
intercept and beta-weight

Call:
lm(formula = time ~ weight, data = turtle)

Coefficients:
(Intercept)       weight
    -0.3281       2.4299
```

Before we check the residuals, we would like to draw the proposed regression line into the scatterplot above. A convenient way is to use the `abline()` function. This function can be used to draw a line based on a formula such as the regression coefficients (see output above). (You can also draw horizontal and vertical lines, see `?abline` for all options.)

```
> abline(lm.turtle, col="blue", lwd=2) #line in color and
thicker
```

If you are interested in the regression coefficient only, just use the `coef()` function on the `lm.turtle` object.

```
> coef(lm.turtle)

(Intercept)      weight
 -0.3281128   2.4299123
```

Hence the regression equation is:

```
time = -0.328113 + 2.429912 * weight
```

This can be quite convenient if a full analysis is not required.

No we should examine the residuals for the linear model, and then plot them.

```
> plot(lm.turtle, pch=16) #switch on History->Recording in
the graphics window)
```

Submit the above program for execution.

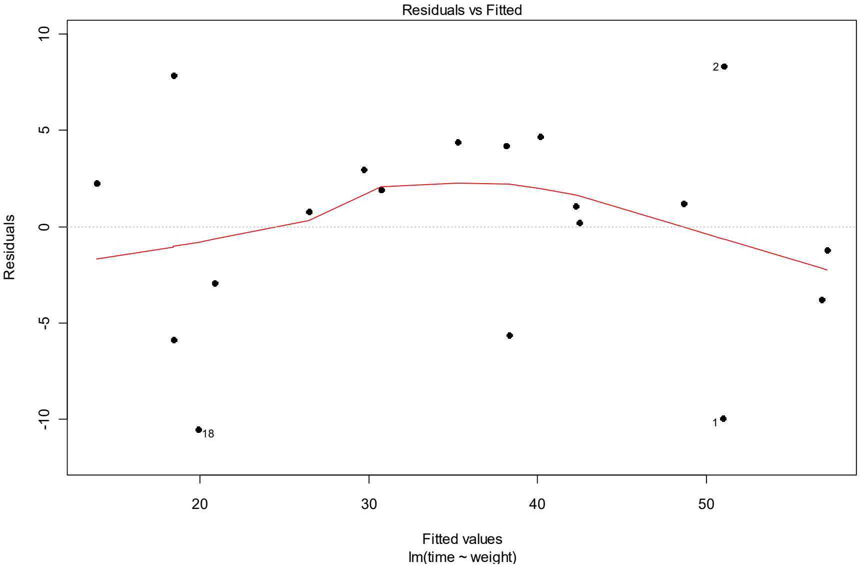


Figure 6-25. A plot of residuals for the linear regression of time to pipping against percent dehydration for the pig-nosed turtle (Carettochelys insculpta).

The plot of the residuals (Figure 6-25) shows an adequate spread of the response variable across the range of the expected values (and therefore the regressor). We can now check the normality of the residuals using the `qqnorm()` function to make a quantile-quantile plot (Figure 6-26).

```
> qqnorm(resid(lm.turtle), pch=16) # Make qq-plot
> qqline(resid(lm.turtle))
```

Submit the above program for execution.

You can also check the normality be plotting a histogram of the residulas, just like in the ANOVA in R module4 and 5.

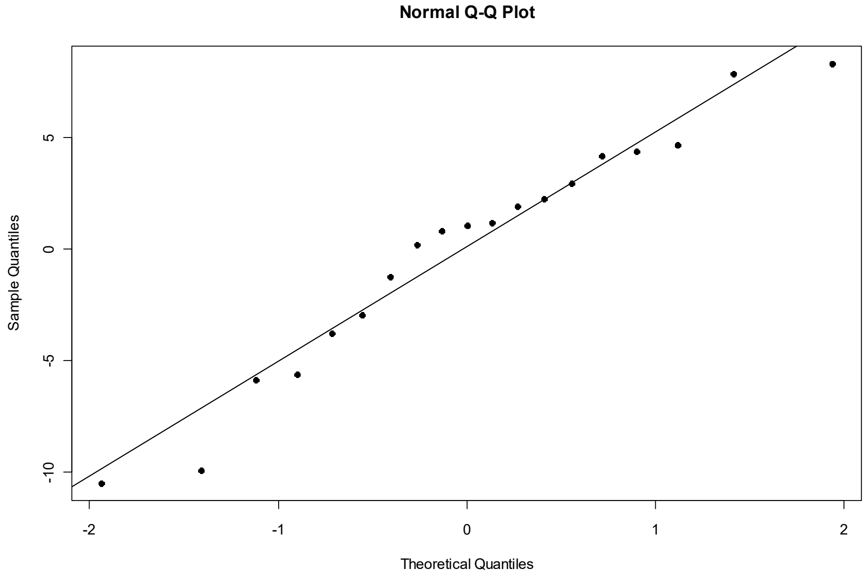


Figure 6-26. A normal probability plot of the residuals. The residuals appear to be approximately normally distributed.

With the assumptions satisfied, we can proceed to interpret the output of the regression. This is found by using the `anova()` and `summary()` functions:

```
> summary(lm.turtle)

Call:
lm(formula = time ~ weight, data = turtle)

Residuals:
    Min      1Q  Median      3Q     Max 
-10.509  -3.360   1.072   3.574   8.338 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -0.3281     3.6449   -0.09     0.93    
weight        2.4299     0.2303   10.55 7.02e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 5.483 on 17 degrees of freedom
Multiple R-squared: 0.8675,     Adjusted R-squared: 0.8597 
F-statistic: 111.3 on 1 and 17 DF,  p-value: 7.016e-09
```

```
> anova(lm.turtle)

Analysis of Variance Table

Response: time
          Df Sum Sq Mean Sq F value    Pr(>F)
weight     1 3346.4  3346.4  111.32 7.016e-09 ***
Residuals 17  511.0    30.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The resulting output gives us everything we need. We have the ANOVA table giving a test of the significance of the regression coefficient. The regression is highly significant (F=111.3; df=1,17; p<0.0001) with a coefficient of determination of $R^2 = 0.87$.

The equation for the regression line is:

```
time = -0.328113 + 2.429912 * weight
```

where time is in minutes and dehydration is a percentage.

Note that while the regression coefficient is significant, the intercept is not significantly different from zero (t=-0.09; p=0.93).

Results

The following is an example of an appropriate results summary, as would appear in your report or publication.

There was a strong linear relationship between time to pipping and percentage dehydration by weight (F=111.32; df=1,17; p<0.0001). For every increase in dehydration of one percentage unit, in the range 5–24%, there was a 2.4 minute increase in the delay of hatching following immersion (Figure 6-29). The relationship is adequately described by:

```
time = -0.328113 + 2.429912 * weight
```

where time is in minutes and dehydration measured in terms of final egg weight expressed as a percentage of initial weight. A total of 86.75% of variation in time to pipping could be explained by level of egg dehydration. The intercept of 0.32 minutes was not significantly different from zero, suggesting that fully turgid eggs would hatch immediately on immersion.

Note that the relationship is described, its statistical significance established (p <0.0001), the strength of the regression is explicitly stated, and the adequacy of fit, assuming linearity, is given (percentage variation explained). A graphic representation of the data and regression accompanies the results summary.

Discussion

The strong linear relationship between dehydration of the eggs and delay in the time to pipping following immersion is consistent with the notion that it is oxygen deprivation that stimulates the hatchling to emerge from its torpor and break free of the egg. The greater the level of dehydration, the greater the size of the air pocket that forms between the egg shell and internal shell membranes. This air presumably sustains the partial pressure of oxygen available to the young turtle within the egg after immersion, and delays the trigger for hatching.

The very small intercept, not significantly different from zero, suggests that fully turgid eggs would hatch immediately upon immersion. This is consistent with the observations of Webb et al (1986) who report that eggs of the pig-nosed turtle hatch immediately and forcibly on immersion in water.

Dehydration of eggs occurs more frequently in nests laid higher above the water, where the sand is dryer and the temperatures higher. This study suggests that a greater stimulus will be required for these eggs than for those closer to the water, which may afford some selective advantage.

Tidy up the program by ensuring there are no elements remaining of the program that did not work. Save the program to disk within Tinn-R for future reference. You may also want to save your workspace and Rhistory, if you want to return to this example at some stage.

Exit from R and Tinn-R by choosing File->Exit from the Menu Bar.

Sources

Venables, W. & Ripley, B. (2002). Modern Applied Statistics Using S-Plus. 4th Edition. Springer-Verlag, Berlin.

Example 6-2: Mercury in Gemfish

This is a simple linear regression where a transformation is applied to rectify curvilinearity and heterogeneity of variances.

The problem

Raymond Chvojka and Denis Reid of the NSW Fisheries Research Institute collected data on mercury contamination of gemfish (*Rexea solandri*) of different lengths. Mercury accumulates in biological systems, and gemfish, being higher tropic consumers, are ideal candidates for mercury contamination. Furthermore, as fish are accumulators, we might expect larger and therefore older fish to be contaminated to a greater extent than smaller fish.

The NSW Environmental Protection Authority's safe limit for the ingestion of mercury in fish is 0.5 mg/kg. A limit on the size of gemfish for sale may need to be set. Analyse the data of Chvojka and Reid to see if there is a relationship between fish length and concentration of mercury in their tissues.

The data

Fish were sampled from commercially exploited stocks, their length measured, and tissue extracted and assayed for mercury according to standard protocols. The data are in the file gemfish.dat, with the first column containing the length data (cm) and the second column containing the mercury concentrations (mg/kg).

The analysis

Double click on the Tinn-R icon and launch R from within Tinn-R (Click in the Menu on R->Initiate/Close Rgui->Initiate preferred Rgui)

Data entry and exploratory examination

The first step in the analysis is to read the data in, and compute some basis statistics. It is often conventional to signify a "missing" data point with a period (.). The convention in **R** is "NA" and when reading in a text data file we need to warn **R** of the possibility of this using the argument `na.strings="."` within the `read.table()` function:

```
> setwd("H:\\Biometry\\data")
> gemfish <- read.table("GEMFISH.DAT", header=T,
na.strings=".")
```

Submit the above commands for execution.

The resulting **R** data frame `gemfish` should contain two variables—`lenght` and `mercury`. By simply entering the name of the data frame object at the command prompt, you can peruse the data at this point to see if it has been read as intended:

```
> str(gemfish)

'data.frame':   243 obs. of  2 variables:
 $ length : num  56 56 50 57 48 59 72.5 64.5 63 63 ...
 $ mercury: num  0.16 0.2 0.16 0.21 0.19 0.27 0.32 0.22
0.16 0.22 ...

> names(gemfish)

[1] "length"  "mercury"
```

Note there is a missing value, signified by a "." in the data file. **R** has replaced it by NA as intended.

Extra task

Try to find out which case is missing?

```
> summary(gemfish)
```

Submit the above commands for execution.

Concentrations of mercury vary from 0.08–3.46 mg/kg, with many fish having mercury concentrations in excess of the safe limit of 0.5 mg/kg (Box 6-3).

*Box 6-3. Summary statistics for mercury concentration against fish length for the gemfish (*Rexea solandri*) captured in eastern Australian waters.*

```
     length          mercury
 Min.   : 40.0   Min.   :0.0800
 1st Qu.: 64.0   1st Qu.:0.2600
 Median : 81.5   Median :0.6400
 Mean   : 78.5   Mean   :0.7592
 3rd Qu.: 93.0   3rd Qu.:1.0450
 Max.   :114.5   Max.   :3.4600
                 NA's   :1.0000
```

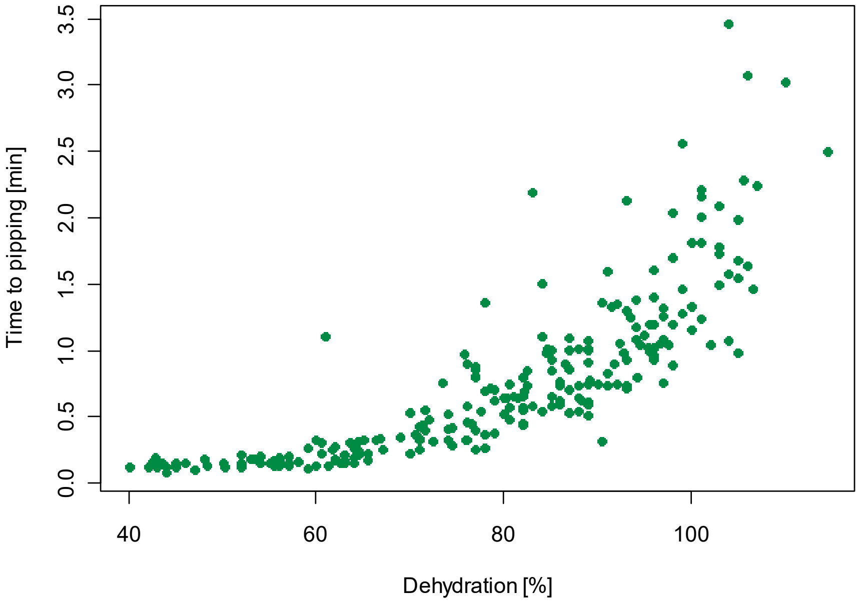
The next step in the analysis is to plot the data. Launching into a regression analysis without first plotting the data is not a good idea.

```
> plot(gemfish$length, gemfish$mercury, pch=16,
col="springgreen4", xlab="Dehydration [%]", ylab="Time to
pipping [min]")
```

Submit the above command for execution.

Again, we have added a few bells and whistles to the plot in terms of optional options on the axis statements. The plot is shown in Figure 6-27.

Figure 6-27. A plot of mercury concentration against fish length for the gemfish (Rexea solandri) captured in eastern Australian waters.



Regression analysis

Clearly a linear model is not appropriate. We can also see that the variances in mercury concentration for fish of given lengths are not equal—the larger the fish, the greater the variation in mercury concentration.

A plot of residuals brings this point home strongly (Figure 6-28). We need to run a linear regression model to generate the residuals, then plot them. The linear model function `lm()` is used here to reinforce the point that we are treating regression and ANOVA as two classes of a linear model. The `glm()` function is also available, has more diagnostic options available, and is more interactive in model building beyond simple linear regression.

```
> lm.gemfish <- lm(mercury ~ length, data=gemfish)  # Fit a
linear model
> plot(lm.gemfish, pch=16)
```

Submit the above program for execution.

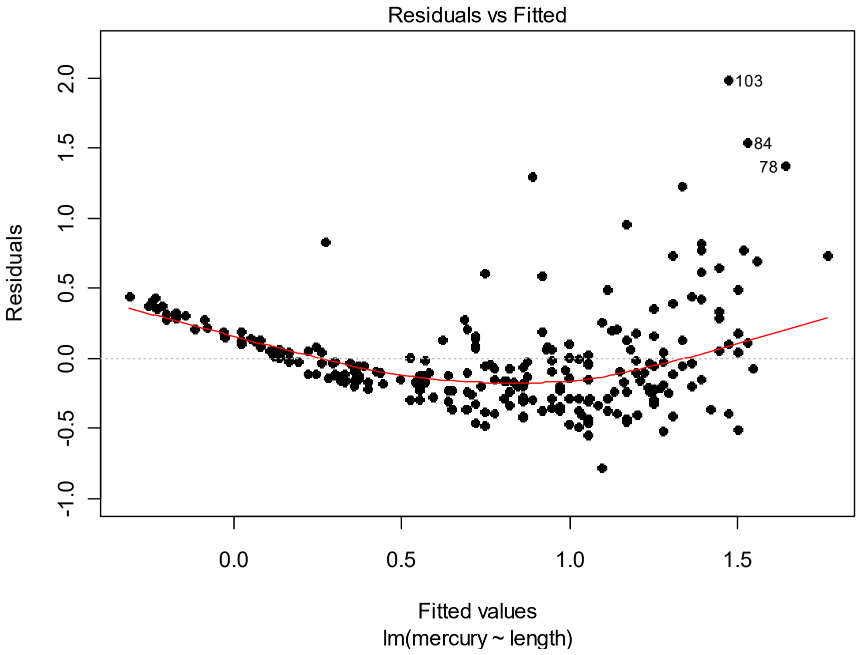


Figure 6-28. A plot of residuals for the linear regression of mercury concentration against the predicted value of fish length for the gemfish (Rexea solandri).

So we have two violations of the assumptions of simple linear regression. A transformation is needed, and several spring to mind. You can try some for yourself, but the log transformation applied to mercury concentration appears to work best.

We need to transform the data and refit the model using the transformed data. We then generate the residuals once more, and plot them. We do this by adding a new column `lg.mercury` to the `gemfish` `data.frame`.

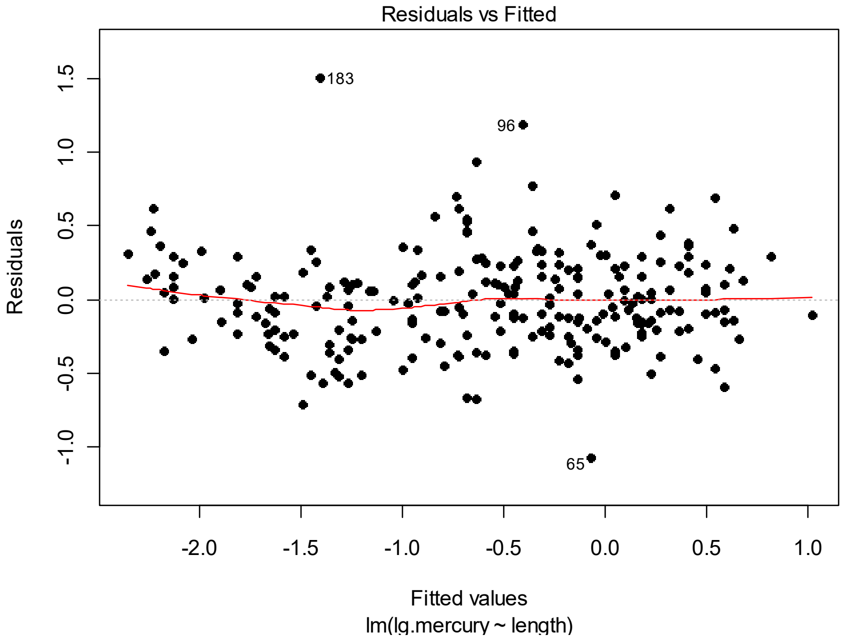
```
> gemfish$lg.mercury <- log(gemfish$mercury)
> head(gemfish)

  length mercury lg.mercury
1     56    0.16  -1.832581
2     56    0.20  -1.609438
3     50    0.16  -1.832581
4     57    0.21  -1.560648
5     48    0.19  -1.660731
6     59    0.27  -1.309333
```

Now we can fit a new model using the transformed mercury data and check the residuals.

```
> lm.log.gemfish <- lm(lg.mercury ~ length, data=gemfish)
> plot(lm.log.gemfish, pch=16)
```

Submit the above program for execution.

Figure 6-29. A plot of studentised residuals for the linear regression of logged mercury concentration against the predicted value of fish length for the gemfish (Rexea solandri).

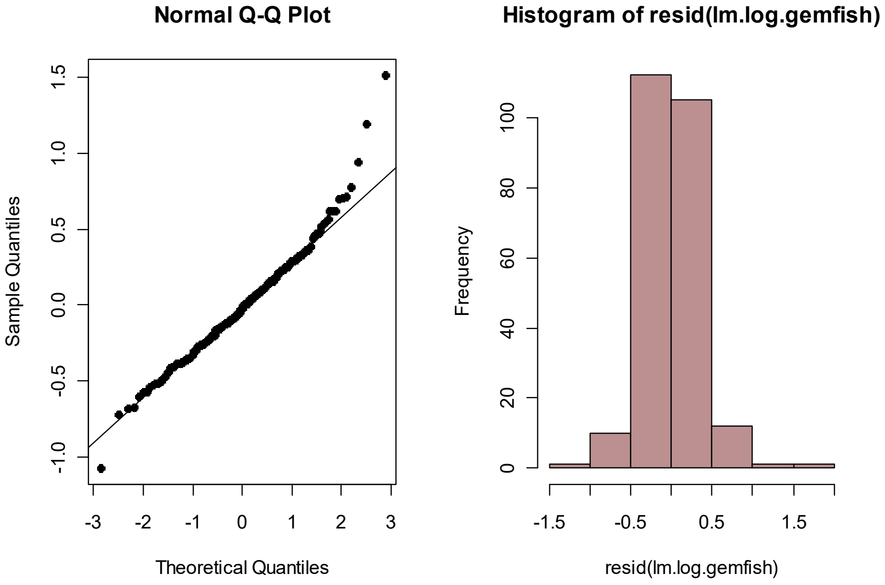


Apart from a few outliers, the plot of the residuals shows an adequate spread of the response variable across the range of the expected values (and therefore the regressor) (Figure 6-29). Having achieved linearity and homogeneity of variances, we can check the normality of the residuals (Figure 6-30).

```
> par(mfrow=c(1,2))
> qqnorm(resid(lm.log.gemfish), pch=16)
> qqline(resid(lm.log.gemfish))
> hist(resid(lm.log.gemfish), col="rosybrown")
```

Submit the above program for execution.

Figure 6-30. A normal probability plot of the residuals and a histogram of the residuals.



Apart from the nuisance outliers evident in the residual plot, the residuals appear normally distributed (refer to Module 2). With the assumptions satisfied, we can proceed to interpret the output of the regression (Box 6-4). This is done by using the `anova()` and `summary()` functions:

```
> summary(lm.log.gemfish)  #brings you the summary
statistics on the regression, R2, F statistic etc.

> anova(lm.log.gemfish) #brings you the ANOVA table for
this regression
```

Box 6-4. Output of a regression analysis of log-transformed mercury concentration against fish length for the gemfish (Rexea solandri).

```
Call:
lm(formula = lg.mercury ~ length, data = gemfish)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.07019 -0.21346 -0.01326  0.18690  1.50932 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.166968   0.095181  -43.78   <2e-16 ***
length       0.045279   0.001184   38.25   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 0.3268 on 240 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.8591,     Adjusted R-squared: 0.8585 
F-statistic:  1463 on 1 and 240 DF,  p-value: < 2.2e-16 
Analysis of Variance Table

Response: lg.mercury
           Df  Sum Sq Mean Sq F value    Pr(>F)    
length      1 156.314 156.314  1463.4 < 2.2e-16 ***
Residuals 240  25.636   0.107                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 
1 
Analysis of Variance Table

Response: lg.mercury
           Df  Sum Sq Mean Sq F value    Pr(>F)    
length      1 156.314 156.314  1463.4 < 2.2e-16 ***
Residuals 240  25.636   0.107                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 
```

The regression is highly significant (F=1463; df=1,240; p<0.0001) with a coefficient of determination of $R^2 = 0.8591$.

The equation for the regression line is:

$$Log_e(mercury) = 0.04528 \times length - 4.167$$

where `length` is in cm and `mercury` concentration is in mg/kg.

We should also plot the transformed data, with the regression line shown (Figure 6-31).

```
> plot(gemfish$length, gemfish$lg.mercury, pch=16,
col="springgreen4", xlab="Dehydration [%]", ylab="Time to
pipping [min]")
> abline(lm.log.gemfish, lwd=2, lty=2, col="red")
```

🏃 Submit the above program for execution.

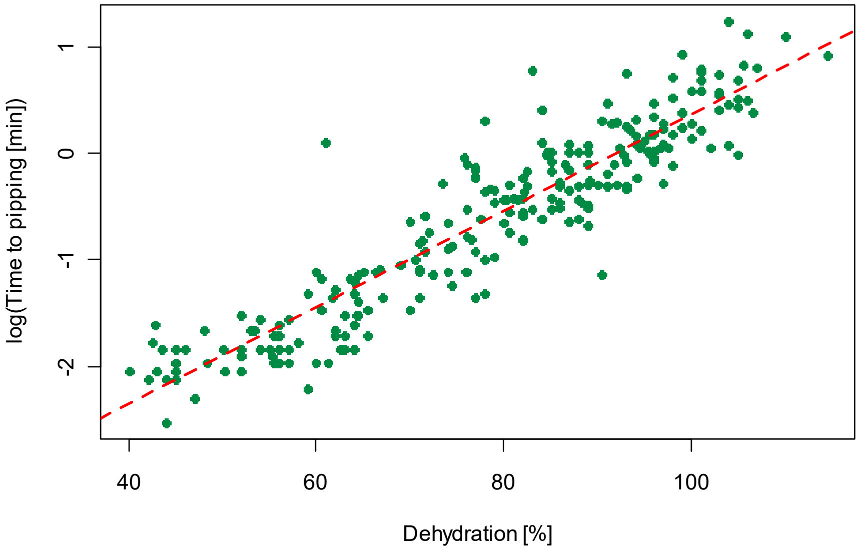


Figure 6-31. A plot of log-transformed mercury concentration against fish length for the gemfish (Rexea solandri) *captured in eastern Australian waters.*

To express the relationship in the original units of measurement, we can reverse transform the equation:

$$mercury = 0.01550 \times e^{0.04528 \times length}$$

where `length` is in cm and `mercury` concentration is in mg/kg. Note that this is the least squares solution for the log transformed values of the response variable, *mercury*. The R^2 value should only be presented with the linearised form of the relationship.

We now have a satisfactory predictive relationship. How do we use it to answer some management questions? The Environmental Protection Agency has, somewhat arbitrarily, set the safe limits for human ingestion at 0.5 mg/kg. Does this mean that we must prevent any fish with mercury levels in excess of this limit from hitting the market, or that we should ensure that the average mercury levels in the fish do not exceed this level?

There are two sets of confidence limits that have a bearing on these questions—the 99% confidence limits of the prediction of an individual Y for a given X (int="c") and the 99% confidence limits for the prediction of the mean value of Y for a given value of X (int="p"). Both are shown in Figure 6-28. The confidence limits for the mean are the narrow limits. We want to predict the confidence limits for all values of `length`, to make the plot easier we have first to sort the values and store them in a new object, `newlength`. The predict function has to be supplied with the model object, a new data.frame for the value that are used to predict the values and the type and level of the confidence interval (see ?predict.lm)

```
> newlength <- sort(gemfish$length)
> conf99 <- predict(lm.log.gemfish,new=data.frame(length
=newlength), int="c", level=0.99)
> pred99 <- predict(lm.log.gemfish,new=data.frame(length
=newlength),  int="p", level=0.99)
```

The objects we have now created, conf99 and pred99, are in fact matrices containing the fitted values of the model (column "fit"), and the lower (column "lwr") and upper (column "upr") intervals:

```
> head(conf99)

        fit       lwr       upr
1 -1.631358 -1.719266 -1.543451
2 -1.631358 -1.719266 -1.543451
3 -1.903031 -2.006036 -1.800026
4 -1.586080 -1.671598 -1.500561
5 -1.993588 -2.101856 -1.885321
6 -1.495522 -1.576402 -1.414642
```

The easiest way to plot this data is using the matlines() function, which plots the columns of a matrix. We want to plot the confidence intervals against length, so let us create a new plot and then add the new confidence limits to this plot.

```
> plot(gemfish$length, gemfish$lg.mercury, pch=16,
col="springgreen4", xlab="Dehydration [%]", ylab="log(Time
to pipping [min])")
> abline(lm.log.gemfish, lwd=2, lty=1, col="red")
> matlines(newlength, conf99, lty=2,
col=c("red","black","black"), lwd=2)
> matlines(newlength, pred99, lty=2,
col=c("red","green","green"), lwd=2)
```

Submit the above program for execution.

Now we need to find the length, where the upper limit of the confidence limits crosses the 0.5 mg/kg threshold. This involves some clever indexing operations. The basic idea is, to ask, at which position are the values in the confidence limits still below the threshold (which()) and then look in the newlength object at this position to find out the length.

Graphically we are looking for the intersection of a horizontal line at the value mercury=0.5 mg/kg. Remember we are plotting on the log scale so we need to plot the line at log(0.5) = -0.69. So let us plot this line first:

```
> abline(h=-0.69, col="orange")
```

From the plot (figure 6=32) we can see that the value for the predicted (the wider confidence interval should be something below 60). Now we check which values are below the threshold (we need to retransform the log values (using the `exp()` function) of the upper confident limit `pred99[,3]`.

```
> which(exp(pred99[,3])<0.5)
```

```
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
27 28 29 30 31 32 33 34 35 36 37 38 39 40
27 28 29 30 31 32 33 34 35 36 37 38 39 40
```

So the 92nd value is still below the threshold.

```
> newlength[40]
```

```
[1] 57
```

And this belongs to the length of of 57 cm. We can check this graphically and draw a vertical line at this value.

```
> abline(v=newlength[40], col="green", lty=2)
```

The intersection of the black dashed line and the orange threshold line, is directly at the upper confidence limits of the confidence interval of the mean mercuary value. We do the same for the other criteria.

```
> which(exp(conf99[,3])<0.5)
```

```
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52
27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52
53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78
53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78
79 80 81 82 83 84 85 86 87 88 89 90 91 92
79 80 81 82 83 84 85 86 87 88 89 90 91 92
```

```
> newlength[92]
```

```
[1] 74.5
```

```
> abline(v=newlength[92], col="black", lty=2)
```

If we adopt the first, more stringent approach, then we seek the length of fish for which we can be 99% sure that the mercury level will be equal to or less than 0.5 mg/kg. On our log scale this corresponds to $\log_e 0.5=-0.69$. From Figure 6-32, we can be 99% sure that an individual fish of length 57 cm will have a mercury concentration equal to or less than 0.5 mg/kg. Perhaps this should be our size limit?

If we adopt the second approach, on the basis that it is the average intake of mercury over a long period that matters, then we should use the confidence limits for the prediction of mean concentration. We can be 99% sure that fish of length 74.5 cm will have a mean concentration of mercury less than or equal to 0.5 mg/kg. Perhaps this should be our limit?

Clearly, the choice between the two possibilities will have major commercial implications, yet we need to protect human health.

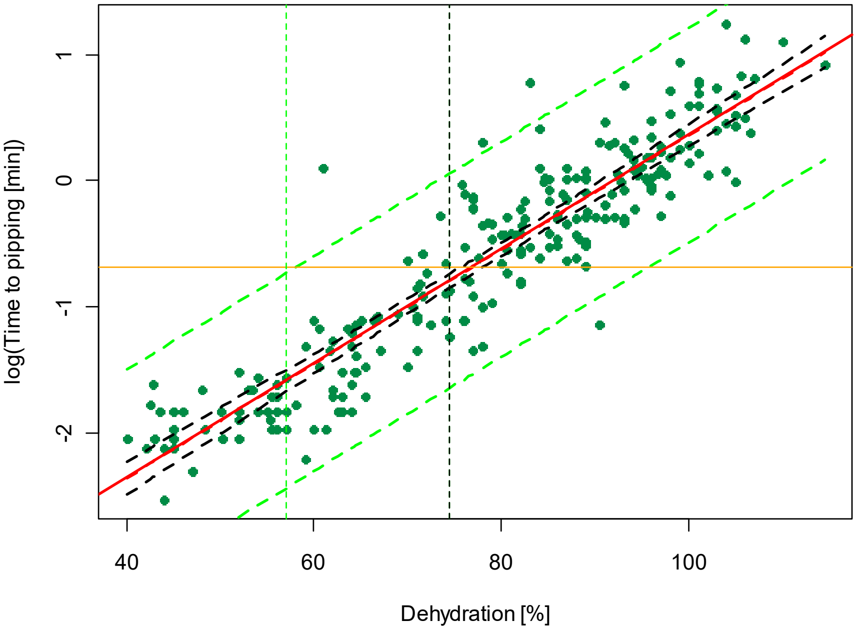


Figure 6-32. A plot of mercury concentration (log-transformed) against the length of gemfish (Rexea solandri)*. The 99% confidence limits for the prediction of the mean value of Y for a given X (narrow limits, black), and for the prediction of an individual value of Y for a given X (wide limits, green) are shown.*

Results

The following is an example of an appropriate results summary, as would appear in your report or publication.

Mercury concentration and the length of gemfish were positively related (Figure 6-27). A linear regression applied after a log transformation of mercury concentration was highly significant (F=1463.35; df=1,240; p<0.0001) with a coefficient of determination of R^2 = 0.8591. Mercury concentration could be predicted from the equation:

$$Log_e(mercury) = 0.04528 \times length - 4.167$$

where length is in cm and mercury concentration is in mg/kg (Figure 6-32).

Two approaches to setting safe size limits for gemfish were explored—the 99% confidence limits for the prediction of mercury in an individual fish suggest a limit of 57 cm; the 99% confidence limits for the prediction of mean mercury levels for a given fish length suggest a limit of 74.5 cm (Figure 6.32).

Note that the relationship is described with an equation, its statistical significance established ($p < 0.0001$), the strength of the regression is explicitly stated, and the adequacy of fit, assuming linearity, is given ($R^2 = 0.8591$). It is clear that a transformation has been applied, and the form of the transformation is evident. A graphic representation of the data and regression accompanies the results summary.

Discussion

Setting size limits for the commercial capture and sale of fish for human consumption is always going to be difficult when those fish are carrying mercury in their tissues. The commercial interests resist setting size limits unless it can be shown that not to do so would cause risk to human health. The public has a right to be protected from unacceptable health risks, and it is the government's responsibility to ensure that appropriate measures to protect public health are in place.

Providing advice from an analysis such as ours must be done in the context of the biology of the fish and the behaviour of the humans catching and eating them. How much mercury is actually consumed? This depends on the amount of gemfish eaten, the average concentration of mercury in the fish, and the level of other dietary sources of mercury. Gemfish are sold as fillets, or occasionally as cutlets, and so it would be rare for individual consumers to consistently eat large fish. The EPA set the limit at 0.5 mg/kg rather arbitrarily, without a clear documented basis addressing these issues, so the decision on the limit to apply is a matter for both the manager and the researcher.

In the context of these caveats, two approaches to setting size limits for the gemfish catch were presented—the 99% confidence limits for the prediction of mercury in an individual fish suggest a limit of 57.5 cm; the 99% confidence limits for the prediction of mean mercury levels for a given fish length suggest a limit of 75 cm (Figure 6-32). The lower limit of 57.5 cm is based on a desire to have any fish with a mercury content in excess of the safe level excluded from sale. However, if the safe levels are to be based on average consumption of mercury, then the limit of 75 cm would be appropriate. The choice will have a major impact on the commercial returns of the fishery.

The high variability in mercury concentration among fish of the same size (14% of variation unexplained) suggests that fish vary in their exposure to mercury contamination. Identifying locations of the catch, that yield fish with particularly high concentrations of mercury, excluding those locations from the fishery, and repeating this study, may be an appropriate response that meets the requirements of human health and allows fish of greater size to reach the market.

Tidy up the program by ensuring there are no elements remaining of the program that did not work. Save the program to disk within Tinn-R for future reference. You may also want to save your workspace and Rhistory, if you want to return to this example at some stage.

Exit from R and Tinn-R by choosing File->Exit from the Menu Bar.

Sources

Box & Jenkins (1976). *Time Series Analysis: Forecasting and Control,* 2nd ed, San Francisco: Holden-Day.

Neter J, Kutner, MH, Nachtsheim, CJ & Wasserman, W (1996). *Applied Linear Statistical Models*. 4th ed, Chicago: Irwin.

Sokal & Rohlf (1994). *Biometry. The Principles and Practice of Statistics in Biological Research.* 3rd ed, San Francisco: W.H. Freeman and Company.

Venables, W. & Ripley, B. (2002). Modern Applied Statistics Using S-Plus. 4th Edition. Springer-Verlag, Berlin.

Webb, G.J.W., Choquenot, D. & Whitehead, P. (1986). Nests, eggs and embryonic development of *Carettochelys insculpta* (Chelonia: Carettochelidae [sic]) from northern Australia. *Journal of Zoology* 1B:521-550.

Where have we come?

In this lesson, we have put theory into practice with a couple of fully worked examples. These examples reinforced the concepts developed earlier by putting them in context, and provided the means for introducing the concepts of confidence limits more thoroughly.

It is now time to put what you have learned into practice with some challenging exercises.

Lesson 5: Exercises

Exercise 6-1: Mercury Contamination of Gemfish

Australian dietary intakes of fish are relatively small. Nevertheless, fish is a comodity of potential public health concern as it can be contaminated with a range of environmentally persistent chemicals, including metals.

Maximum permitted concentrations are set for contaminants in food when the health of consumers cannot be safeguarded by other mechanisms. These maximum permitted concentrations are determined by the Australia New Zealand Food Authority and are set out in Standard A12 of the Food Standards Code. Any maximum permitted concentration must be considered safe at the upper end of the range of dietary intakes of the population. The maximum permitted concentration for mercury in fish is 1.0 mg/kg.

In an earlier step-through example, size limits for Gemfish were explored on the basis of a maximum permitted concentration of 0.5 mg/kg. In this exercise, you will be asked to set a size limit on the basis of the revised figure of 1.0 mg/kg set by Standard A12. You will be further asked to consider the impact of this size limit on the commercial catch of Gemfish in eastern Australian waters.

Three datasets are at your disposal. The first file (gemfish_mercury.dat) contains data on mercury concentrations in the tissue of fish of various sizes. It can be used to develop a relationship between fish length and mercury concentration. The second file (gemfish_lw.dat) contains data on length and weight for a large sample of gemfish. It can be used to develop a relationship linking fish length with weight. The third file (gemfish_size.dat) contains data on the size distribution of gemfish in the fishery for each year since 1975.

gemfish_mercury.dat

Fish length is measured as the length to caudal fork (cm) and mercury concentration is measured as mg/kg. The data are represented as two columns in the datafile, with fish length values first.

gemfish_size.dat

Fish length is measured as the length to caudal fork to the nearest whole cm below the true length. The dataset contains a column for fish length, and a series of columns containing fish counts in each size class for each year from 1975 to 1997.

gemfish_lw.dat

Fish length is again measured as the length to caudal fork to the nearest whole cm below the true length. Fish whole weight is given in kg. The sex of the fish is also provided. There are three columns in the datafile – sex, length, weight.

In this exercise, you are asked to

- Develop a relationship between mercury and fish length.
- Set a length limit on the basis of the mercury considerations.
- Develop a length-weight relationship for gemfish (male and female data combined).
- Construct a size frequency distribution for gemfish
- Estimate the magnitude of forfeited catch in tonnes if the length limit is applied.

Gemfish (*Rexea solandri*) have only recently been fished commercially, but their stocks have become rapidly depleated. Their rates of growth are exceptionally slow, a fact not appreciated at the time the fishery began. Furthermore, gemfish tend to school and are spatially clustered near ocean mounts, increasing their vulnerability to overfishing. Changes in their size distribution over time are clearly evident in the data contained in gemfish_size.dat. We require estimates of the impact of size limits on the contemporary catch (say combining the statistics for 1993-97).

Analysis -- Mercury

- Plot mercury concentration against fish length and present the graph below. Would a standard least-squares linear regression be appropriate? If not, give two reasons why not? Do not forget to plot the residuals for the raw data, and to include the plot below.
- Explore a range of transformations likely to linearise the relationship between mercury and fish length, and select the one that you regard as most appropriate – give the formula for the transformation below. Apply the transformation, and re-plot the data and residuals. Present the plots below. Are there any outliers?
- Which confidence limits are appropriate for setting a maximum allowable mercury concentration in a given fish presented to market. Justify your choice. Plot the data again, this time showing the regression line and the appropriate confidence limits. Use the graph to set a size limit appropriate to the government maximum allowable standard of 1.0 mg/kg.
- Conduct the Regression Analysis and summarise the results in the form of an ANOVA table. Present the regression equation in two forms – first as a linear equation involving the transformed

variable(s) (with an R^2 value), and second in terms of the original measurement variables.

Analysis – Length-Weight Relationship

- Plot whole body weight against fish length, with data for males and females combined, plot the residuals, and present the graphs below. Would a linear regression be appropriate? If not, give two reasons why not?
- What would you regard as the most likely form of the relationship between weight and length? Give an equation for it below. What transformation will linearize such a relationship? Apply the transformation, and re-plot the residuals and the data with the regression line included. Present the plots below. There is a strange pattern in the residuals – can you suggest a cause?
- Conduct the Regression Analysis and summarise the results in the form of an ANOVA table. Present the regression equation in two forms – first as a linear equation involving the transformed variable(s) (with an R^2 value), and second in terms of the original measurement variables.

Analysis – Size Frequency Distribution

- Plot the size frequency distribution for gemfish for the years 1993 to 1997 inclusive. Present your plot below.
- Calculate the total biomass of the fish contributing to the frequency histogram presented above and the biomass of fish that exceed your size limit. What proportion of potential gemfish biomass will be sacrificed by setting a maximum permitted mercury concentration?

Results summary

- Write a summary of the results of the entire analysis, as might be included in the results section of a report or manuscript. Refer in your summary to ANOVA tables and figures generated during the analysis as appropriate. Include in your results, only descriptions of any clear and statistically significant trends, but do not at this stage attempt to explain them.

Discussion

- Discuss the analysis in the context of the reasons for conducting the study. Refer to the document Metal Contamination of Major NSW Fish Species available for Human Consumption, which is available for download at
  http://learnonline.canberra.edu.au/file.php/5339/papers/Exercise_6-1_paper.pdf
  for background information.

Exercise 6-2: Zinc Accumulation in Cockles

Bivalves can be used as a bio-indicator because they readily accumulate trace metal pollutants through their filter-feeding or by absorption into the body tissue.   Oysters, cockles and mussels are efficient tools to monitor the water quality of a polluted estuarine environment. Bioaccumulation often depends on an size, gender or breeding condition of an organism, so these factors must be addressed when investigating patterns of bioaccumulation.

Rajani Rai undertook a study of pollution in an enclosed estuarine environment, in Lake Macquarie, New South Wales. She chose the Sydney cockle (*Anadara trapezia*) as a potential bio-indicator of cadmium, copper and zinc pollution. Her study sampled cockles across all size ranges, sexes, and seasons in order to unravel the pattern of bioaccumulation in these organisms.

Rai determined the dry mass of the individual cockles and total body zinc concentration in the laboratory. She found that the cockles were accumulators of cadmium and copper but were regulators of zinc. As part of her study, Rajani wanted to investigate the relationship between zinc and cockle mass.

The data reside in the file cockle.dat, and comprise four columns – sample label, sex (M or F), weight (g), and zinc load (μg/g). Weight is total dry weight and zinc load is total body concentration in μg/g.

- Plot zinc concentration against cockle weight and present the graph below. Would a linear regression be appropriate? If not, give two reasons why not? Do not forget to plot the residuals for the raw data, and to also include the residual plot below.
- Explore a range of transformations likely to linearise the relationship between zinc concentration and cockle weight, and select the one that you regard as most appropriate – give the formula for the transformation below. Remember, the objective of transformation is to both linearize the relationship and homogenize the variances.

  Apply the transformation, and re-plot the data and residuals. Present the plots below. Are there any outliers?
- Conduct the Regression Analysis and summarise the results in the form of an ANOVA table. Present the regression equation as a linear equation involving the transformed variable(s) (with an R2 value).
- Write a summary of the results of the entire analysis, as might be included in the results section of a report or manuscript. Refer in your summary to ANOVA tables and figures generated during the analysis as appropriate. Include in your results, only descriptions of

any clear and statistically significant trends, but do not at this stage attempt to explain them.

- Discuss the analysis in the context of the reasons for conducting the study. If you were comparing zinc concentrations in cockles from two areas, what measures would you need to take to ensure that the comparison was informative.

Exercise 6-3: Patterns of butterfly species richness

The distribution of most pollinator species is poorly documented despite the important ecosystem services they provide. Pollinator species are threatened by human-induced environmental change, and a potentially severe concern is that of climate change. There is mounting evidence for a biotic response to relatively small climate changes within this century, evidence such as the spatial shifts in the distribution of European and North American butterfly and bird species. Although much of the evidence is circumstantial, the findings are also consistent with predictions from current models of spatial patterns of species richness.

Jeremy Kerr of the University of Oxford used GIS to investigate the spatial patterns of butterfly diversity in Canada. He was particularly interested to learn if butterfly species richness was influenced by climate, and in particular average annual potential evapotranspiration (PET, measured in mm/yr). The butterflies of Canada provide a superb baseline for studying the effects of climate on contemporary patterns of species richness and comprise the only complete pollinator taxon for which this sort of analysis is possible.

The data are found in the file butterfly.dat and comprise three columns – quadrat number, potential evapotranspiration (mm/yr) and butterfly species richness (no. of species).

- Plot butterfly species richness against potential evapotranspiration and present the graph below. Would a linear regression be appropriate? If not, give two reasons why not? Do not forget to plot the residuals for the raw data, and to also include the residual plot below.
- Explore a range of transformations likely to linearise the relationship between butterfly species richness and potential evapotranspiration, and select the one that you regard as most appropriate – give the formula for the transformation below. Apply the transformation, and re-plot the data and residuals. Present the plots below. Are there any outliers?
- Conduct the Regression Analysis and summarise the results in the form of an ANOVA table. Present the regression equation in two forms – first as a linear equation involving the transformed variable(s) (with an R2 value), and second in terms of the original measurement variables.
- Write a summary of the results of the entire analysis, as might be included in the results section of a report or manuscript. Refer in your summary to ANOVA tables and figures generated during the analysis as appropriate. Include in your results, only descriptions of

any clear and statistically significant trends, but do not at this stage attempt to explain them.

- Discuss the analysis in the context of the reasons for conducting the study. Refer to the document Butterfly species richness patterns in Canada: energy, heterogeneity, and the potential consequences of climate change, which is available for download at http://www.consecol.org/vol5/iss1/art10, for background information.

Where have we come?

In this Module, you have been introduced to simple linear regression first from a classical perspective then from the perspective of an Analysis of Variance.

You should now see Analysis of Variance and Regression as two facets of a broader class of analyses, that of general linear models.

This brings us to the end of the series of modules on statistical analysis for ecology and natural resource management. I hope that you have now a solid foundation for extending your knowledge of analysis options in your study and work.

More advanced topics such as multiple regression, polynomial regression, analysis of covariance and more complex designs in ANOVA, should be within your reach.

There are a very many excellent texts in these areas, and I wish you well in your future studies.