

Chapter 1: Introduction

Q: Discuss, by using **examples**, the use of models as descriptions: (a) of everyday life, (b) in maths, and (c) in statistics.

- (a) When you think of models in everyday life, you think of the model trains or houses children play with, or the model aeroplanes people of all ages build. Seriously, models are used by architects for town planning, by medical schools (models of the brain for teaching neurology), and when testing aeroplanes in a wind tunnel. The models need only be as accurate as necessary for the task in hand. For example, a child's car model need not be very detailed but a model aeroplane in a wind tunnel needs to be as accurate as is possible in order to **predict** the behaviour of the real counterpart.
- (b) In a mathematical model, we focus on a particular aspect of a given system and attempt to find an equation or set of equations that describes this aspect. Example: equations of motion describing the motion of a projectile. The accuracy is again governed by the purpose of the model (prediction vs. description) but is additionally constrained by how the factors can be accommodated mathematically. Example: a pebble skimmed across a lake — atmospheric conditions, level of tide, presence of rocks — can these be modelled mathematically?

A central feature of mathematical models concerns the handling of imposed constraints and external influences. Such factors could be affected by chance fluctuations which could be very important in some situations e.g. testing the efficiency of a new headache tablet — people could vary in their tolerance, or the tablet could affect different severities of headache differently. “Errors” such as the fickleness of people's responses in questionnaires must be taken into account in any analysis of the resulting data.

- (c) Valid analysis of either observational or experimental data must be founded on techniques that pay due regard to possible chance fluctuations. By incorporating chance into equations, we turn mathematical models into statistical models. There are many statistical models encountered in a wide variety of practical situations.

Q: Explain the difference between (a) populations and samples, (b) variables and factors, and (c) observational and experimental data.

- (a) A population is the complete set of individuals that could be measured. Sometimes, this can be finite e.g. the total number of maths students in Bangor in 2000, and so in principle, every individual can be asked a question. But, the population is often infinite, or it is not possible to examine every individual.

In nearly all practical investigations, the data will come from a **subset** of the population, or a sample. Hopefully, the sample is representative of the whole population, and the investigator can seek to draw *inferences* about the population from the sample.

It is essential to guard against bias in the choice of samples. Random sampling is usually enough to ensure the later assumptions we make are reasonable ones. A random sample is one where every member of the population has an equal chance to be included.

- (b) Commonly, several variables are measured on each individual in an experiment. The variables of primary interest to the investigator are called **response** or **dependent** variables, and other background or supplementary variables are called **explanatory** or **independent** variables. For example, life style will be analysed when studying respiratory problems in a group of individuals.

One of the main applications of statistical models is in the study of dependence between *response* and *explanatory* variables. Mathematical equations describing relationships are formed, sometimes using dummy variables, which are used when a variable is qualitative e.g. occupation. In this example, suppose that there are only 3 occupations: x_1 , x_2 and x_3 . If a respondent has job x_1 , then the values $x_1 = 1$, $x_2 = 0$ and $x_3 = 0$ are assigned, etc. Variables that have only **two** possible values are called binary or dichotomous.

The word “factor” is used in several ways. Generally, “factor” can be used for any general explanatory variable. In a narrower sense, the word “factor” is usually synonymous with qualitative explanatory variables, whose states are then called “*levels*” of the factor.

- (c) In statistics, a survey can be used to describe any occasion on which data has been collected simply by observation. The defining characteristics are that (i) *the data collection takes place in “real” surroundings*, and (ii) *the researcher has no control over prevailing conditions*. This is the opposite of an experiment, where (i) *the data collection takes place in artificial surroundings*, and (ii) *the researcher has complete control over prevailing conditions*.

Example: in an experiment to determine whether temperature has any effect on plant height, a researcher might randomly place plants in a garden (observational) or control different temperatures in a greenhouse (experimental). One drawback of experimental data is that the results might not be extrapolated into the “real world” where we want to use them.

- Q:** Explain the difference between *linear* and *non-linear* models.

The defining feature of a linear model is that a unit change in any parameter leads to the same change in the dependent variable — whatever the values of the parameters. Suppose that we have a model $\mu = a + bx_1 + cx_2$; and suppose that b changes to $(b+1)$. Then the change in μ is $[a + (b+1)x_1 + cx_2] - [a + bx_1 + cx_2] = x_1$, and this will be the same *whatever* the values of a , b and c .

Any non-linearity of variables in a model is irrelevant — it is in terms of the parameters that the linearity property has to be tested.

Consider the model $\mu = a \exp(bx_1 + cx_2)$. When b changes to $(b+1)$, the change in μ can be simplified to $a(e^{x_1-1}) \exp(bx_1 + cx_2)$, but no further. Thus the change in μ will be different for different values of a , b and c . Hence this model is non-linear. But, taking logarithms on both sides of the model, we obtain $\log \mu = \log a + bx_1 + cx_2$. This *is* a linear model. Some non-linear models can be reparameterised like this, but others are intrinsically non-linear.

Chapter 2: Distributions and Inference

Q: What is a *random variable*? Give **examples**.

Where it is **impossible** to predict in advance of observation what the actual value of a variable will be for a specific individual, i.e. we have inherent unpredictability, we have a random variable. A complete description of a random variable is given by specifying all the conceivable values and quoting their associated probabilities in a probability distribution. Examples: height, weight, age, of a person; score obtained by throwing 2 die; distance travelled to University from home.

Q: Explain the difference between *discrete* and *continuous* variables.

A discrete variable is one in which all possible values come from either a finite or a countable finite set, so that these values can all be listed in a (possibly infinite) sequence. A continuous variable, on the other hand, is one in which all values can be any real numbers in a specified range. Such a set of values is uncountably infinite — it is not possible to list all values in a sequence.

Q: Show that $E[(Y-\mu)^2] = E[Y^2]-\mu^2$ for *random* variables.

$$E[(Y-\mu)^2] = E(Y^2-2Y\mu+\mu^2) = E(Y^2)-E(2Y\mu)+E(\mu^2) = E(Y^2)-\mu E(2Y)+\mu^2.$$

As $E(Y) = \mu$, then we *have* $E(Y^2)-\mu(2\mu)+\mu^2 = E(Y^2)-2\mu^2+\mu^2 = E(Y^2)-\mu^2$.

Q: Use the *method of moments* to estimate the parameters of a normal distribution when a sample of n values y_1, \dots, y_n has been collected.

The method of moments chooses the estimate of any general unknown parameter to be the value which makes the population mean *equal* to the sample mean. In our example, the Y_i come from the p.d.f. $f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2\}$ ($-\infty < y < \infty$). The normal distribution has mean μ and variance σ^2 , i.e. we have 2 unknowns. Here, we set $E(Y)$ equal to $\frac{1}{n} \sum_{i=1}^n Y_i$, and $E(Y^2)$ equal to $\frac{1}{n} \sum_{i=1}^n Y_i^2$. We then solve the pair of simultaneous equations for μ and σ .

So $E(Y) = \frac{1}{n} \sum_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2\}$, and $E(Y^2) = \frac{1}{n} \sum_{i=1}^n (\frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2\})^2$. Notice that the **first** equation gives the sample mean, and that the **second** equation minus the **first** equation *squared* gives the sample variance. After solving the equations, we would expect to get back the “obvious” answers, that is the parameters of a normal distribution when a sample of n values has been collected are μ and σ , i.e. $Y \sim N(\mu, \sigma^2)$.

Q: Use the method of *least squares* in the above question.

If Y_1, Y_2, \dots, Y_n are such that $Y_i = g(\beta_1, \beta_2, \dots, \beta_k) + \epsilon_i$ for $i = 1, 2, \dots, n$, where $g(\beta_1, \beta_2, \dots, \beta_k)$ is a constant function of k parameters β_i , and the ϵ_i are iid random variables each with *zero* mean and a common variance σ^2 , then the least squares estimators (lse) of the parameters are the values which minimise $V = \sum_{i=1}^n [Y_i - g(\beta_1, \beta_2, \dots, \beta_k)]^2$. Standard calculus can be employed to obtain these values. Note: Least squares estimators coincide with *maximum likelihood* ones if the distribution of the ϵ_i is normal, but not *necessarily* otherwise.

In our example, we know that the Y_i are iid normal ($N(\mu, \sigma^2)$) variables. We can equivalently write $Y_i = \mu + \epsilon_i$ for $i = 1, 2, \dots, n$, where the ϵ_i are iid $N(0, \sigma^2)$ variables. In this formulation, we can think of the ϵ_i as random variables which measure the *departures* of the Y_i from the population mean μ .

Secondly, we can see that the *log-likelihood* has the form (this will be derived in the **next** question) $l(\mu, \sigma^2) = -n \log \sigma - \frac{n}{2} \log(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2$. If σ^2 is simply regarded as an *irrelevant* constant, then maximising the log-likelihood with respect to μ is the same as minimising $\sum_{i=1}^n (y_i - \mu)^2$ with respect to μ . In other words, the m.l.e. of μ is the value which **minimises** the sum of the squared departures between the sample values and the parameter. This is how we should *proceed*.

Q: Show that the m.l.e. of the **mean** of a normal distribution is *unbiased*.

First, we shall derive the maximum likelihood estimate (m.l.e.) of some samples from a normal distribution.

Theory: Let Y_1, \dots, Y_n be a random sample from any p.d.f. $f(y; \theta)$ depending on an unknown parameter θ . The *likelihood* of the sample is the joint probability density $f(y_1, \dots, y_n; \theta)$ of the sample, treated as a function of θ . Writing the likelihood as $L(\theta)$, and because random samples are *independent*, we have $L(\theta) = \prod_{i=1}^n f(y_i; \theta)$. The *function* $\hat{\theta} = g(y_1, y_2, \dots, y_n)$ that maximises $L(\theta)$ w.r.t. θ is the maximum likelihood estimator of θ , and its **actual** value for a given sample is the *maximum likelihood estimate* (m.l.e.) of θ for that sample. **Note** that if $\hat{\theta}$ is the maximum of $L(\theta)$, then $\hat{\theta}$ is also the maximum of $\log L(\theta) = l(\theta)$ (which we shall be *working* with).

Calculation: Suppose that y_1, y_2, \dots, y_n is a random sample from a normal distribution with mean μ and variance s^2 , two parameters which we would like to estimate. Then we can say from the above that $L(\mu, \sigma^2) = \prod_{i=1}^n \left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{y_i - \mu}{\sigma}\right)^2\right\}\right)$, so that $l(\mu, s^2) = \log L(\mu, s^2)$

$$= \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 = -n \log \sigma - \frac{n}{2} \log(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2.$$

Hence $\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mu)$, and $\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (y_i - \mu)^2$. Setting these equations to zero and solving them for σ and μ yields the m.l.e.'s $\hat{\sigma}$ and $\hat{\mu}$. Thus once we *do* set them to zero, we must change the σ and μ to $\hat{\sigma}$ and $\hat{\mu}$ *respectively*.

Setting the first equation to zero (and multiplying through by $\hat{\sigma}^2$ which cannot be zero) yields $\sum_{i=1}^n (y_i - \hat{\mu}) = 0$, which gives $\hat{\mu} = \frac{1}{n} \sum y_i = \bar{y}$. Setting the **second** equation to zero (and multiplying through by $\hat{\sigma}^3$) yields $\hat{\sigma}^2 = \sum_{i=1}^n (y_i - \hat{\mu})^2 = \frac{1}{n} \sum (y_i - \bar{y})^2$. These are the m.l.e.'s of μ and σ^2 respectively. Note that the **divisor is n** in $\hat{\sigma}^2$, as opposed to the divisor (n-1) in the usual definition of sample variance.

We have shown above that the m.l.e. of the mean of a normal distribution is given by $\hat{\mu} = \frac{1}{n} \sum y_i = \bar{y}$. An estimator is said to be unbiased if $E(\hat{\theta}) = \theta$, i.e. if the *mean* of its sampling distribution is θ . Otherwise, it is said to be biased, with bias $b(\theta) = E(\hat{\theta}) - \theta$. An unbiased estimator is therefore one which gives the correct value “*on average*” over repeated sampling, which is an useful property.

For the normal distribution, $E(\hat{\mu}) = \frac{1}{n} \sum_{i=1}^n E(y_i) = \frac{1}{n} (\mu + \mu + \dots + \mu) = \frac{1}{n} (n)\mu = \mu$. Therefore, the m.l.e. of μ in a *random* sample of size n from an $N(\mu, \sigma^2)$ distribution is unbiased.

Q: Show that the m.l.e. of the **variance** in the normal distribution is *biased*.

We have shown in the *previous* question that the m.l.e. of the variance in the normal distribution is $\frac{1}{n} \sum (y_i - \bar{y})^2$. To test whether this is **biased**, we find the expectation of the expression and see if it comes out as σ^2 . Now $E(\frac{1}{n} \sum (y_i - \bar{y})^2) = E(\frac{1}{n} \sum (y_i - \mu + \mu - \bar{y})^2) =$

$$E(\frac{1}{n} (\sum (y_i - \mu)^2 - \sum (\bar{y} - \mu)^2)) = \frac{1}{n} \sum E((y_i - \mu)^2) - \frac{1}{n} \sum E(\bar{y} - \mu)^2 = \sigma^2 - \frac{1}{n} \sigma^2.$$

Therefore, the m.l.e. is biased, as we do not get σ^2 coming out of our *calculations*. Notice that if we were to use the **usual** sample variance $s = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$, then this would be an *unbiased* estimator for σ^2 . This explains the preference for the *latter* estimator in **practical** applications.

Q: State the *advantages* and *disadvantages* of the three methods of estimation.

Broadly speaking, the method of moments is the *simplest* of the three methods, works well in straightforward cases, but doesn't always produce estimators with good properties. Maximum likelihood is a good general-purpose method which produces estimators that have excellent *large-sample* properties — but sometimes poor *small-sample* properties. Least squares is the best in the context of *linear* models, and is equivalent to maximum likelihood when the data comes from the *normal* distribution.

Q: State and define **three** properties of estimators, and show that they hold for the estimate of μ in a normal distribution.

We shall look at **Bias, Mean Square Error** and **Consistency**.

(1) As stated above, an estimator is said to be unbiased if $E(\hat{\theta}) = \theta$, i.e. if the *mean* of its sampling distribution is θ ; otherwise it is biased with bias $b(\theta) = E(\hat{\theta}) - \theta$. Also shown above is that if we have an estimate for μ in a normal distribution, e.g. the m.l.e., then the estimate $\hat{\mu} = \frac{1}{n} \sum y_i = \bar{y}$ is unbiased.

- (2) An unbiased estimator merely ensures that the average of a (*large*) series of repeated estimates is correct, but says nothing about **individual** elements. As in practice we usually take only one sample, it would be nice if this particular estimate was close to θ . A good estimator is one with small MSE. If we have a choice of several estimators, the best one will be the one with the smallest MSE.

It can be shown that the MSE is identically equal to **squared bias** plus **variance**. Hence the MSE of an unbiased estimator is equal to its variance, so that the usual characterisation of a “good” estimator is an unbiased estimator with *small* variance. However, there is a result (The Cramer-Rao Lower Bound Theorem) which establishes that the *variance* of any unbiased estimator of a parameter θ cannot be less than $I^2(\theta)$, where

$$I^2(\theta) = \frac{1}{-E\left(\frac{\partial^2 l}{\partial \theta^2}\right)}. \quad (l \text{ denotes the log-likelihood of the sample providing the estimator}).$$

Therefore, no unbiased estimator can do “*better*” than one whose variance is equal to $I^2(\theta)$, and hence such an estimator is said to be fully efficient. For the *normal* distribution, we already know that \bar{Y} is an *unbiased* estimator of μ . We also know from a previous question that $\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mu)$, so that $\frac{\partial^2 l}{\partial \mu^2} = -\frac{1}{\sigma^2} \sum_{i=1}^n (1) = -\frac{n}{\sigma^2}$. Since this expression just involves the **constants** n and σ^2 , then

$$I^2(\theta) = -\frac{1}{E\left(\frac{\partial^2 l}{\partial \theta^2}\right)} = -\frac{1}{-\frac{n}{\sigma^2}} = \frac{\sigma^2}{n}. \quad \text{Thus } \bar{Y} \text{ is a } \textit{fully efficient} \text{ estimator of } \mu.$$

- (3) The preceding properties all involve the sampling distribution of an estimator $\hat{\theta}$ obtained from a sample of size n . To check that we have a sensible estimator, we need to check that $\hat{\theta}$ is increasingly *likely* to yield the right answer θ as the sample size gets bigger, and, in the limiting case, is **certain** to yield θ if the whole population is sampled.

The latter condition is satisfied if the MSE is zero, so that $\hat{\theta}$ is a consistent estimator of θ if its $\text{MSE} \rightarrow 0$ as $n \rightarrow \infty$. If $\hat{\theta}$ is *unbiased*, then it is also consistent if its variance $\rightarrow 0$ as $n \rightarrow \infty$. For our **normal** example, the previous question says that it is evident that $\text{var}(\bar{Y}) = \sigma^2/n \rightarrow 0$ as $n \rightarrow \infty$, so that \bar{Y} is a consistent estimator of μ .

Q: For the following sample of *heights* in cm of sitka spruce springs one month after plantation, calculate a confidence interval for the population mean. Sample: 9, 6, 12, 15, 13, 14, 11, 13, 12, 9, 8, 14.

It can be shown that $(\bar{Y} - t_{n-1, \alpha} \frac{s}{\sqrt{n}}, \bar{Y} + t_{n-1, \alpha} \frac{s}{\sqrt{n}})$ is a $100(1-\alpha)\%$ confidence interval for μ . Consider that in our example, we want a 95% confidence interval for the population mean. To plug in values into the above expression, we want to find the *sample mean*, \bar{Y} ; the number of samples, n , which is 12; and the *standard deviation*, s . ($s = \sqrt{\text{variance}}$). Calculations:

$$\bar{Y} = (1/12)(9+6+12+15+13+14+11+13+12+9+8+14) = (1/12)(136) = 11\frac{1}{3}.$$

$$E(Y^2) = (1/12)(9^2+6^2+12^2+\dots+14^2) = (1/12)(1626) = 135\frac{1}{2}.$$

$$\text{So Variance} = E(Y^2) - [E(Y)]^2 = 135\frac{1}{2} - (11\frac{1}{3})^2 = 135\frac{1}{2} - 128\frac{4}{9} = 7\frac{1}{18}.$$

$$\text{Therefore, the Standard Deviation is } s = \sqrt{7\frac{1}{18}}.$$

So our 95% confidence interval is $(11\frac{1}{3} \pm t_{(11, 0.05)} \frac{\sqrt{7\frac{1}{18}}}{\sqrt{11}})$. Knowing *t-distribution values*, we could **simply** the expression further.

Q: State the *assumptions* you are making in the calculations you performed in the above question.

The assumption made is that the data comes from a *normal distribution* which has mean μ and variance σ^2 , i.e. $N \sim (\mu, \sigma^2)$, a random variable. To verify this, we would plot the data, and see if the shape of the distribution followed the shape of a *normal* distribution.

Q: Test whether the population mean height grown is **10cm** for the above sample of data.

To answer this question, we shall calculate a *95% Prediction Interval* for the mean, and check to see whether 10 lies within it. If it does, then the **null hypothesis** that the mean height grown is 10cm holds; but if 10 lies outside the prediction interval, then we accept the alternate hypothesis that the mean height grown is not 10cm.

The 95% prediction interval is given by $\bar{Y} \pm t_{(n-1, 0.05)} \times \text{sd}_{\text{prediction}}$. We need to calculate the *prediction standard deviation*. To do this, consider that Prediction = estimate of μ + estimate of ϵ , both of which are subject to *error*. From earlier questions, recall that the **variance** of $\hat{\mu}$ is σ^2/n , and that the **variance** of ϵ is σ^2 . Therefore, the *estimated variance of prediction* is given by $\hat{\sigma}^2(n+1/n)$.

Using $(n-1)$, so that $\hat{\sigma}^2 = s^2$, we have estimated variance of prediction $s^2(n+1/n)$, which in *this* instance is $7\frac{1}{18}(13/12) = 1651/216$. So our 95% prediction interval is $11\frac{1}{3} \pm t_{(11, 0.05)} \times 1651/216$. Again, knowing the t-value for $t_{(11, 0.05)}$, we could go on to answer the question concerned, and state whether the **null** or the **alternate** hypothesis is the one that holds.

Chapter 3: Normal Response and Quantitative Explanatory Variables: Regression

Q: For a simple linear regression model $y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \epsilon_i$, find the *least squares estimates* of β_0 and β_1 .

We look for estimates of the two parameters that minimise the sum of the squared departures, $S = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1[x_i - \bar{x}])^2$. *Differentiating*, we obtain

$$\partial S / \partial \beta_0 = \sum 2(y_i - \beta_0 - \beta_1[x_i - \bar{x}])(-1), \text{ and } \partial S / \partial \beta_1 = \sum 2(y_i - \beta_0 - \beta_1[x_i - \bar{x}])(-[x_i - \bar{x}]).$$

Setting these *partial derivatives* to zero (for the minimum), and noting that $\sum_i [x_i - \bar{x}] = 0$, we yield the **normal** equations for the minimising values:

$n\hat{\beta}_0 = \sum y_i$, and $\sum (x_i - \bar{x})^2 \hat{\beta}_1 = \sum (x_i - \bar{x})y_i$. But $\sum_i (x_i - \bar{x})y_i = \sum_i (x_i - \bar{x})(y_i - \bar{y})$ (since $\bar{y} \sum_i (x_i - \bar{x}) = 0$), so that if we adopt the **standard** notation $S_{xx} = \sum (x_i - \bar{x})^2$, $S_{yy} = \sum (y_i - \bar{y})^2$, and $S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$, then the *normal equations* can be written as $n\hat{\beta}_0 = n\bar{y}$ and $S_{xx}\hat{\beta}_1 = S_{xy}$, and the parameter estimates follow directly from these *equations* as $\beta_0 = \bar{y}$ and $\beta_1 = S_{xy}/S_{xx}$. Differentiating S a second time with respect to each parameter shows that the estimates do indeed provide **minimums**.

Q: Show that the estimate of β_1 is *unbiased*, and derive an expression for the **variance** of the estimate.

$\hat{\beta}_1$ can be written as $(\sum [x_i - \bar{x}]y_i) / (\sum [x_i - \bar{x}]^2)$ from the direct solution of the above normal equations. This is of the form $\sum l_i y_i$, where $l_i = [x_i - \bar{x}] / (\sum [x_i - \bar{x}]^2)$. Thus $\hat{\beta}_1$ is a linear combination of *independent* normal variables (the y_i), so that we deduce that its **sampling distribution** is also normal, with mean $\sum l_i E(y_i)$ and variance $\sigma^2 (\sum l_i^2)$. Therefore, we have $E(y_i) = \beta_0 + \beta_1(x_i - \bar{x})$, so that $\sum l_i E(y_i) = \sum \left\{ \frac{[x_i - \bar{x}][\beta_0 + \beta_1(x_i - \bar{x})]}{\sum [x_i - \bar{x}]^2} \right\} = \beta_0 \frac{\sum [x_i - \bar{x}]}{\sum [x_i - \bar{x}]^2} + \beta_1 \frac{\sum [x_i - \bar{x}]^2}{\sum [x_i - \bar{x}]^2} = \beta_1$.

(The first term in the *second* expression disappears because the sum in the numerator is always **zero**, and the second fraction in this line cancels to 1). Further, we have the following: $\sum l_i^2 = \sum \{ [x_i - \bar{x}] / (\sum [x_i - \bar{x}]^2) \}^2 = (\sum [x_i - \bar{x}]^2) / (\sum [x_i - \bar{x}]^2)^2 = 1 / (\sum [x_i - \bar{x}]^2)$. Thus the sampling distribution of $\hat{\beta}_1$ is normal, with **mean** β_1 (so that $\hat{\beta}_1$ is an unbiased estimator of β_1), and **variance** σ^2 / S_{xx} . Remembering that σ^2 is estimated by the residual mean square in the analysis of variance table, and denoting this estimate by s^2 , it follows that the estimated standard error of $\hat{\beta}_1$ is $s / \sqrt{S_{xx}}$.

Q: Show that the estimate of β_0 is *unbiased*, and derive an expression for the **variance** of the estimate.

Since $\beta_0 = \bar{y}$, inferences about the *estimate* of β_0 is conducted by the usual methods on \bar{y} . Standard results show that \bar{y} has a normal distribution with mean β_0 and variance σ^2/n . From the questions on chapter 2, we therefore know that the *estimate* of β_0 is **unbiased**.

Q: Show that for the simple linear regression model with **residual** e_i , we have the following result: $E_1 = \sum e_i^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$.

In the absence of any information on X, the implicit model for the data is $y_i = \mu + \epsilon_i$ (for $i = 1, \dots, n$), in which case the *estimate* of μ is the sample mean \bar{y} . Hence $\hat{y}_i = \bar{y}$ for all i , so that $e_i = y_i - \bar{y}$, and the *sum of the squared residuals* is $E_0 = \sum_{i=1}^n (y_i - \bar{y})^2 = S_{yy}$.

If we fit the *simple regression model*, then y_i is “predicted” by the point on the regression line **corresponding** to $X = x_i$, so that $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1(x_i - \bar{x}) = \bar{y} + (S_{xy}/S_{xx})(x_i - \bar{x})$ on substituting for the *parameter estimates*. Thus $e_i = y_i - \hat{y}_i = (y_i - \bar{y}) - (S_{xy}/S_{xx})(x_i - \bar{x})$, so that the *sum of squared residuals* is $E_1 = \sum_{i=1}^n ([y_i - \bar{y}] - \frac{S_{xy}}{S_{xx}}[x_i - \bar{x}])^2$.

On **squaring** out the bracket and summing, we find that $E_1 = S_{yy} + (S_{xy}^2/S_{xx}^2) \times S_{xx} - 2 \times S_{xy} \times (S_{xy}/S_{xx})$, which reduces to $E_1 = S_{yy} - (S_{xy}^2/S_{xx})$. QED.

Q: Show that $E_0 - E_1 = \sum (\hat{y}_i - \bar{y})^2$, where $E_0 = \sum (y_i - \bar{y})^2$, and E_1 is as defined above.

We know that $E_0 = \sum_{i=1}^n (y_i - \bar{y})^2 = S_{yy}$, and that $E_1 = S_{yy} - (S_{xy}^2/S_{xx})$. Therefore, $E_0 - E_1 = (S_{xy}^2/S_{xx})$. Now in the above *question*, we had $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1(x_i - \bar{x}) = \bar{y} + (S_{xy}/S_{xx})(x_i - \bar{x})$. **Rearranging** the terms, we can obtain $(S_{xy}/S_{xx})(x_i - \bar{x}) = \hat{\beta}_0 + \hat{\beta}_1(x_i - \bar{x}) - \bar{y} = \hat{y}_i - \bar{y}$. Thus, *squaring* and *summing* over i gives us $\frac{S_{xy}^2}{S_{xx}^2} \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1[x_i - \bar{x}] - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$. But, we know that $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$, so it **follows** that $E_0 - E_1 = \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1[x_i - \bar{x}] - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$. QED.

Q: Explain in words what E_0 , E_1 and $\sum (\hat{y}_i - \bar{y})^2$ mean.

$\sum (\hat{y}_i - \bar{y})^2$ is a quantity that represents the variation accounted for **by** the relationship between Y and X; it is termed the *regression sum of squares*, and can be expressed as $E_0 - E_1$. E_1 represents the **residual** variation about the regression line — so is termed the *residual sum of squares*. E_0 is clearly the **total** sum of squares about the sample mean. The fundamental analysis of variance identity thus states that the **total** sum of squares equals the **regression** sum of squares plus the **residual** sum of squares.

Q: Perform a *full linear regression analysis* on the data in BARLEY.MTW. This is the data from an experiment at College Farm to study the effect of nitrogen fertiliser on barley yield. The first column represents the amount of fertiliser (in kg/ha), and the second column gives the yield of barley (in tons/ha).

After *importing* the data into Minitab, the following output was obtained (see the next page):

Regression Analysis

The regression equation is
 $\text{Yield} = 4.47 + 0.0158 \text{ Fert}$

Predictor	Coef	StDev	T	P
Constant	4.4675	0.1769	25.25	0.000
Fert	0.015841	0.001461	10.85	0.000

S = 0.5465 R-Sq = 80.8% R-Sq(adj) = 80.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	35.132	35.132	117.62	0.000
Residual Error	28	8.364	0.299		
Total	29	43.496			

Unusual Observations

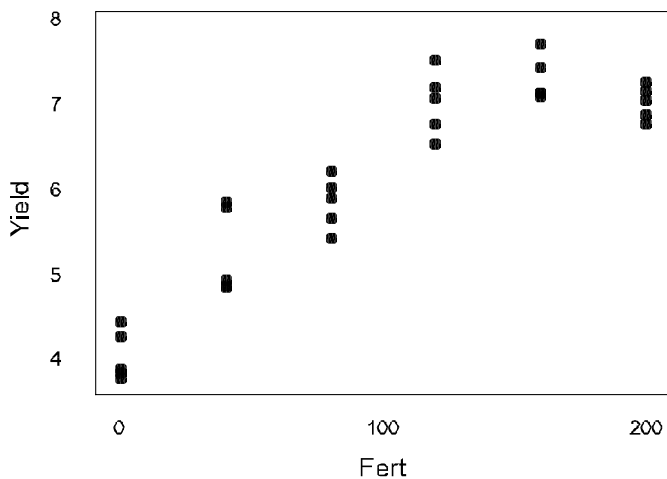
Obs	Fert	Yield	Fit	StDev Fit	Residual	St Resid
17	120	7.4950	6.3685	0.1040	1.1265	2.10R

R denotes an observation with a large standardized residual

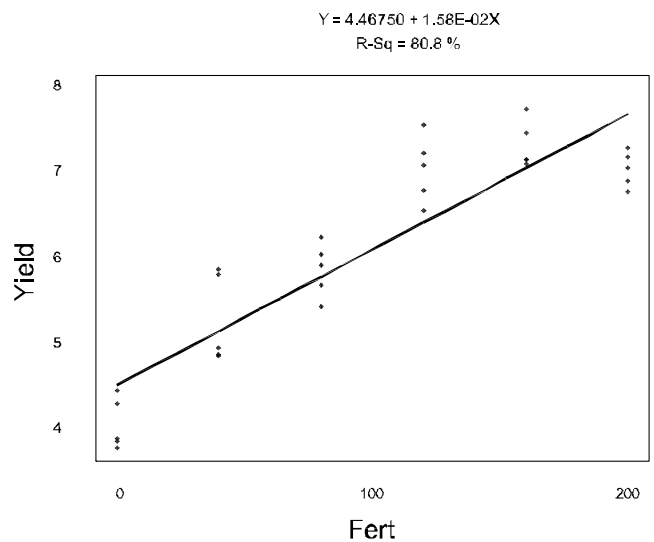
MTB >

Graphs

Data Plot

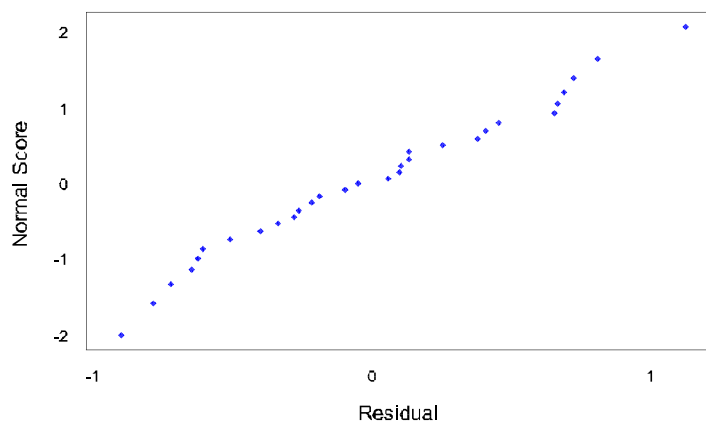


Regression Plot



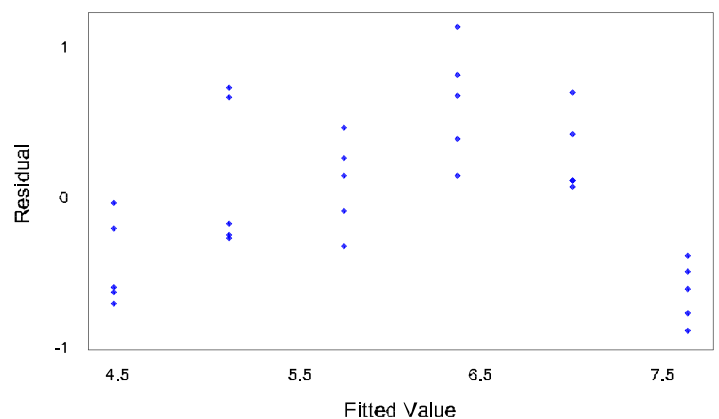
Normal Probability Plot of the Residuals

(response is Yield)



Residuals Versus the Fitted Values

(response is Yield)



The first thing we do is to plot the data, and from the graph shown, we conclude that there could be a *straight line relationship* between Fertiliser and Yield, although the graph suggests a curved relationship for higher values of fertiliser. The next thing we do is to ask for a regression analysis, complete with a normal probability plot of the residuals, and a residuals vs. fits plot.

From the regression output provided by Minitab, we see that the regression equation is $\text{Yield} = 4.47 + 0.0158 \text{ Fert}$. The p-values in the next section tells us that there is a very small probability (not zero as shown) that the null hypothesis holds, i.e. that there is **no relationship** between fertiliser and yield.

By applying a straight line relationship, Minitab tells us that **80%** of the variability is accounted for by the straight line relationship. However, there is still 20% of variability to be accounted for, and this could be reduced by looking for further (e.g. *quadratic* or *cubic*) relationships between fertiliser and yield. This is seemingly confirmed by the *residuals vs. fits* plot, which seems to have an “n” shape to the values, and the *regression plot*. (Both graphs are shown on the previous page).

On the plus side, the *normal probability plot* of the residuals looks like a straight line, so we can say that the samples were probably taken from a normal distribution.

Let us now look for another relationship to model the data. Using the *fitted line plot* command in Minitab, and selecting the quadratic option, we get the following output and graph:

Polynomial Regression

$Y = 3.94971 + 3.53E-02X - 9.71E-05X^{**2}$
R-Sq = 91.1 %

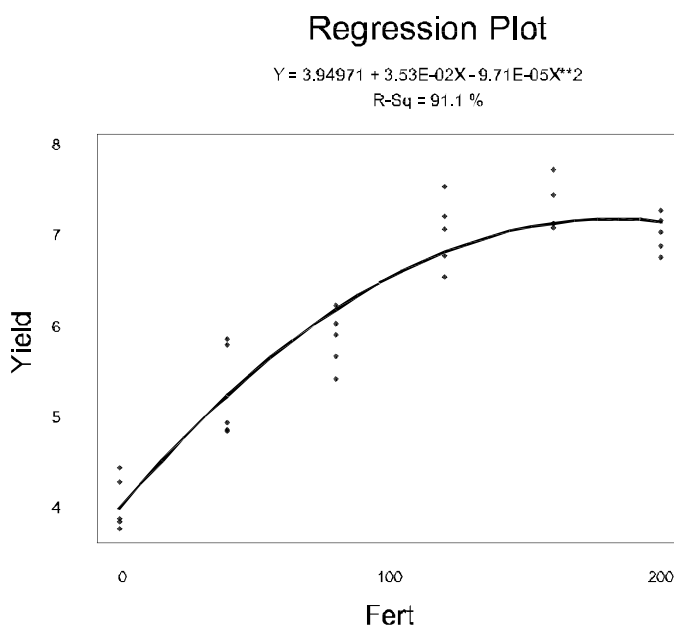
Analysis of Variance

SOURCE	DF	SS	MS	F	P
Regression	2	39.6368	19.8184	138.648	6.22E-15
Error	27	3.8594	0.1429		
Total	29	43.4962			

SOURCE	DF	Seq SS	F	P
Linear	1	35.1325	117.616	1.56E-11
Quadratic	1	4.5043	31.5118	5.89E-06

MTB >

As you can see, there is a marked increase in the variability accounted for by the line, the increase being from 80% to 91%. Conclusion: a quadratic model is best for this data. (A cubic model, when applied, only produces marginal increases in the R-Sq value).



Q: Use the *matrix* approach to find the least squares estimates for a **multiple** regression model. ($\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$).

As in the case of *simple linear regression*, we estimate the parameters β_j ($j = 0, 1, \dots, p$) by the method of least squares. Our aim is to find the vector $\boldsymbol{\beta}$ that minimises $S(\boldsymbol{\beta}) = \sum_{i=1}^n \epsilon_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$. The *usual way of proceeding* is by differentiating S with respect to each of the β_j ; setting the resulting *expressions* to zero; and solving the ensuing system of simultaneous equations.

An **alternative** route, not involving calculus, is to note that if $\boldsymbol{\beta}_0$ is any value of $\boldsymbol{\beta}$ satisfying $\mathbf{X}'\mathbf{X}\boldsymbol{\beta}_0 = \mathbf{X}'\mathbf{y}$, then $S(\boldsymbol{\beta}_0) \leq S(\boldsymbol{\beta})$ for all $\boldsymbol{\beta}$. This result can be *established* in the following steps: First, from the definition of S above, we have $S(\boldsymbol{\beta}) - S(\boldsymbol{\beta}_0) = \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + 2\boldsymbol{\beta}_0'\mathbf{X}'\mathbf{y} - \boldsymbol{\beta}_0'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}_0$.

Next, *substituting* for $\mathbf{X}'\mathbf{y}$ from the above, we obtain $S(\boldsymbol{\beta}) - S(\boldsymbol{\beta}_0) = \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\beta}_0'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}_0$. Finally, collecting terms together and tidying up, we see that $S(\boldsymbol{\beta}) - S(\boldsymbol{\beta}_0) = [\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)]'[\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)]$, which is a *sum of squares* — which must be greater than or equal to zero. This establishes the result.

If two distinct vectors $\boldsymbol{\beta}_0^{(1)}$ and $\boldsymbol{\beta}_0^{(2)}$ satisfy $\mathbf{X}'\mathbf{X}\boldsymbol{\beta}_0 = \mathbf{X}'\mathbf{y}$, then by subtraction, $\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}_0^{(1)} - \boldsymbol{\beta}_0^{(2)}) = \mathbf{0}$. *Elementary linear algebra* tells us that this can only be the case if the determinant of $\mathbf{X}'\mathbf{X}$ is zero. Thus if $\det(\mathbf{X}'\mathbf{X}) \neq 0$, then $\mathbf{X}'\mathbf{X}$ is non-singular, and $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is the (*unique*) least squares estimator of $\boldsymbol{\beta}$.

Q: Show that for the case of a *simple linear regression*, the matrix approach leads to the **same** estimates for β_0 and β_1 in terms of S_{xy} , S_{xx} , n , Σy_i , etc.

The *simple linear regression model* is a case of the general linear model with $p = 1$. We have

$$\mathbf{y} = (y_1, y_2, \dots, y_n)'; \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \dots & \dots \\ 1 & x_n - \bar{x} \end{pmatrix}; \quad \boldsymbol{\beta} = (\beta_0, \beta_1)'; \quad \text{and } \boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)'$$

Straightforward matrix transposition and multiplication yields $\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & 0 \\ 0 & \sum_j (x_j - \bar{x})^2 \end{pmatrix} = \begin{pmatrix} n & 0 \\ 0 & S_{xx} \end{pmatrix}$, from which we immediately *obtain* $(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 1/n & 0 \\ 0 & 1/S_{xx} \end{pmatrix}$. Hence $\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \sigma^2/n & 0 \\ 0 & \sigma^2/S_{xx} \end{pmatrix}$.

Therefore, we see that $\hat{\beta}_0$ and $\hat{\beta}_1$ are *uncorrelated*, and the regressors (in this case, trivially, X_0 and X_1) are orthogonal. Moreover, the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ from the **diagonal** elements of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, and these agree with the expressions derived *previously*. Finally,

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} \sum y_j \\ \sum (x_j - \bar{x})y_j \end{pmatrix} = \begin{pmatrix} \sum y_j \\ S_{xy} \end{pmatrix}, \text{ so that } \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} \frac{1}{n} \sum y_j \\ S_{xy}/S_{xx} \end{pmatrix}.$$

Again these expressions **agree** with what was obtained in previous questions.

Chapter 3: Further Questions

Q: Using the *matrix* approach to regression, state what problems may occur when fitting a polynomial model.

In multiple regression, the matrix $(\mathbf{X}'\mathbf{X})^{-1}$ plays an important part in the calculations, which break down if $\mathbf{X}'\mathbf{X}$ is singular — the inverse does not exist in this case. However, several problems can be encountered if $\mathbf{X}'\mathbf{X}$ can be inverted but is *nearly* singular, which will happen if there is an *approximate* linear relationship among some or all of the regressor variables. In this case, we say that there is multicollinearity among the regressor variables. The main problems caused by multicollinearity are as follows:

- Some or all of the regression coefficients will have large *standard errors*, so are unreliable as estimators of the model parameters.
- There is instability in the fitted model, meaning that a small perturbation to an observation, or the deletion of an observation from the data set, will produce a very **different** fitted model.
- Difficulties arise in variable selection.

It is important to recognise multicollinearity if it is present, so that *remedial* action can be taken. For example, in polynomial models, such as $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$, powering rapidly increases the *magnitudes* of values. As the order of the model increases, so will the danger of $\mathbf{X}'\mathbf{X}$ becoming ill-conditioned, and we are increasingly likely to encounter *multicollinearity* problems.

Mean-centering is a useful device in such circumstances — use $z_i = (x_i - \bar{x})$ instead of x_i , and fit a *polynomial model* in terms of the z_i — and deduce the parameters for the polynomial in x_i from this fitted model. For example, if we wish to fit $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$, we fit instead $y_i = \gamma_0 + \gamma_1 z_i + \gamma_2 z_i^2 + \epsilon_i$, where $z_i = x_i - \bar{x}$.

The **latter** model is just $y_i = \gamma_0 + \gamma_1(x_i - \bar{x}) + \gamma_2(x_i - \bar{x})^2 + \epsilon_i$ so on expanding the squared term and collecting terms together, we see that $\beta_0 = \gamma_0 - \gamma_1 \bar{x} + \gamma_2 \bar{x}^2$; $\beta_1 = \gamma_1 - 2\gamma_2 \bar{x}$; and $\beta_2 = \gamma_2$. Thus estimates $\hat{\beta}_i$ are readily deduced from the $\hat{\gamma}_i$, and the latter were obtained with greater numerical stability because many of the *correlations* between the regressors are greatly reduced by mean-centering.

Q: Perform a polynomial regression on the **barley** data.

This was done in the previous section inadvertently — a *quadratic* model was fitted to the model, which was found to represent 11% more of the variability than when applying a linear model only. (The *increase* in the R^2 value was from 80% to 91%).

Formally, we should (in Minitab) produce a **new** column containing the squared x-values, and then perform the linear regression, allowing the x and x^2 columns to be treated as “explanatory” variables. In this way, we would get the same equation as if we would just ask Minitab to apply a quadratic model with its fitted line plot command.

Q: State why standardised residuals (rather than residuals) are used to assess the regression model assumptions.

Let us look at the basic *regression* model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ has mean vector $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$. The vector of *fitted values* is $\hat{\mathbf{y}} = (\hat{y}_1, \dots, \hat{y}_n)' = \mathbf{X}\hat{\boldsymbol{\beta}}$, and the vector of residuals is $\mathbf{e} = (e_1, \dots, e_n)' = \mathbf{y} - \hat{\mathbf{y}}$. We also know that $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, and that $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, the “*hat*” matrix. It can be shown that $E(\mathbf{e}) = \mathbf{0}$, and $\text{cov}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$. If we write the $(i, j)^{\text{th}}$ element of \mathbf{H} as h_{ij} , it follows that $\text{var}(e_i) = (1 - h_{ii})\sigma^2$, and $\text{cov}(e_i, e_j) = -h_{ij}\sigma^2$.

Intuitively, looking at the residuals should give us an indication of the *quality* of fit of the regression. If all the residuals are “small”, then the fit is good; but if some or all are “large”, then the fit is suspect. Unfortunately, the raw residuals cannot be easily interpreted directly, as their scale of values will depend on the *range* and *scales* of the response and regressor variables.

An “average” variance of the residuals is given by the residual mean square, s^2 , which estimates the unknown departure variance σ^2 . It follows that a simple scaling of the residuals is to divide each by the average standard deviation, s . **However**, $\text{var}(e_i) = (1 - h_{ii})\sigma^2$ shows that the residuals do not have *constant* variance. Therefore, it is preferable to work with the **standardised residuals**, defined by $e'_i = \frac{e_i}{s\sqrt{(1 - h_{ii})}}$.

Furthermore, $\text{cov}(e_i, e_j) = -h_{ij}\sigma^2$ shows that the residuals are *not* independent. However, the covariance between any two of them will usually be low, particularly if the sample size n from which the model has been fitted is large. Finally, if we assume the normality of the ϵ_i in $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, then the fact that the residuals are *linear* functions of the y_i implies the normality of the e'_i as well.

Thus, if all the assumptions underlying the regression model are correct, we can view the e'_i as iid $N(0,1)$ random variables. The graphical inspection of the e'_i will quickly show up any violation of the model assumptions, or any model *inadequacy*, as such situations will induce deviation from iid $N(0,1)$ behaviour on the part of the **standardised** residuals.

Q: Define an *influential observation*, an *outlier*, and the *leverage* of an observation.

An influential observation is defined to be one which, for whatever reasons, causes a large change in some or all of the *estimated* regression parameters when it is omitted from the data set. Sometimes, an **outlier** is an influential observation.

Outliers are observations that are not well fitted by the assumed model, and have large residuals. A “rule of thumb” is that an observation with a standardised residual greater than 2.5 in absolute value indicates a possible outlier, and then the source of the data should be *investigated*. The observation should be omitted if there is a doubt about the accuracy of the observation.

If a large difference in the model parameter estimates is consequently obtained, then the observation is said to be influential. Influential outliers are more important than ones which do not cause much perturbation in model parameters on omission, because they raise the possibility of model instability.

Observations well separated from the others in terms of their regressor variable values will have *large* values of h_{ii} , with h_{ii} coming from the “hat” matrix, \mathbf{H} . We call h_{ii} the leverage on the i^{th} observation. As a rough guide, observations with $h_{ii} > 3(p+1)/n$ are influential, where p is the number of *regressor variables* in the model, and n is the *number of observations*.

However, the drawback with leverage as a measure of influence is that it only takes into account the data configuration regarding the **regressor** variables, so that it may miss those observations whose influence arises partly or wholly through their response variable values. Various other measures of influence have therefore been proposed, such as *Cook’s distance*.

Q: Perform a multiple regression of aerial biomass on the 5 variables *salinity, acidity, potassium, sodium* and *zinc levels*, using the data in O:\STATDATA\DATA\CAPEFEAR.MTW. Perform the regression as described below.

O:\STATDATA\DATA\CAPEFEAR.TXT

The data in O:\STATDATA\DATA\CAPEFEAR.MTW comes from Applied Regression Analysis, a research tool by J.O. Rawlings 1988, and is in the Science Library QA278.2.R38.

The data describes a piece of research to identify the important soil characteristics influencing aerial biomass production of the marsh grass *Spartina alterniflora* in the Cape Fear estuary of North Carolina.

Factors

type revegetated dead areas, short *Spartina* areas and tall *Spartina* areas
location 3 places Oak Island, Smith Island and Snows Marsh

5 samples from each of the type by location sites were taken.

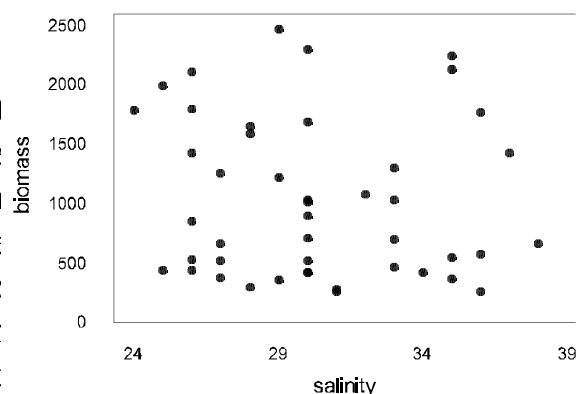
Measurements taken

salinity - 0/00
 pH - acidity as measured in water
 Kk - potassium in ppm (the name K causes Minitab problems!)
 Na - sodium in ppm
 Zn - zinc in ppm

The dependent variable is aerial biomass in /gm²

The objective is to identify which measurements both singly and together are related to biomass.

To start with, I looked to see if the variable “*salinity*” affected our dependent variable “*biomass*”. Looking at the data in the plot shown on the right, it seems that there is no significant relationship between the two variables, and this is confirmed by the Regression output on the following page, where only **1.1%** of the variability in the data is explained by the relationship between salinity and biomass. (A low F-value and a high p-value). I therefore conclude that salinity does not affect the aerial biomass of the marsh grass.



Regression Analysis

The regression equation is
biomass = 1555 - 18.3 salinity

Predictor	Coef	StDev	T	P
Constant	1554.9	820.7	1.89	0.065
salinity	-18.31	26.92	-0.68	0.500

S = 664.1 R-Sq = 1.1% R-Sq(adj) = 0.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	204048	204048	0.46	0.500
Residual Error	43	18966915	441091		
Total	44	19170963			

Unusual Observations

Obs	salinity	biomass	Fit	StDev Fit	Residual	St Resid
26	29.0	2436.0	1024.0	104.7	1412.0	2.15R
27	35.0	2216.0	914.1	161.4	1301.9	2.02R

R denotes an observation with a large standardized residual

Next, I looked to see whether **Type** or **Location** affected the biomass, either *together* or *singly*. The following is the regression output obtained.

Regression Analysis (Type & Location)

The regression equation is
biomass = - 35 + 367 type + 151 location

Predictor	Coef	StDev	T	P
Constant	-35.2	315.3	-0.11	0.912
type	366.7	107.1	3.42	0.001
location	151.3	107.1	1.41	0.165

S = 586.6 R-Sq = 24.6% R-Sq(adj) = 21.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	4720387	2360193	6.86	0.003
Residual Error	42	14450577	344061		
Total	44	19170963			

Source	DF	Seq SS
type	1	4033333
location	1	687053

Unusual Observations

Obs	type	biomass	Fit	StDev Fit	Residual	St Resid
33	1.00	1960.0	785.5	174.9	1174.5	2.10R
34	1.00	2080.0	785.5	174.9	1294.5	2.31R

R denotes an observation with a large standardized residual

Regression Analysis (Type)

The regression equation is
biomass = 267 + 367 type

Predictor	Coef	StDev	T	P
Constant	267.5	234.0	1.14	0.259
type	366.7	108.3	3.38	0.002

S = 593.3 R-Sq = 21.0% R-Sq(adj) = 19.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	4033333	4033333	11.46	0.002
Residual Error	43	15137630	352038		
Total	44	19170963			

Unusual Observations

Obs	type	biomass	Fit	StDev Fit	Residual	St Resid
33	1.00	1960.0	634.1	139.8	1325.9	2.30R
34	1.00	2080.0	634.1	139.8	1445.9	2.51R

R denotes an observation with a large standardized residual

Regression Analysis (Location)

The regression equation is
 $\text{biomass} = 698 + 151 \text{ location}$

Predictor	Coef	StDev	T	P
Constant	698.1	258.6	2.70	0.010
location	151.3	119.7	1.26	0.213

S = 655.6 R-Sq = 3.6% R-Sq(adj) = 1.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	687053	687053	1.60	0.213
Residual Error	43	18483910	429858		
Total	44	19170963			

Unusual Observations

Obs	location	biomass	Fit	StDev Fit	Residual	St Resid
26	2.00	2436.0	1000.8	97.7	1435.2	2.21R

R denotes an observation with a large standardized residual

25% of the variability is accounted for by both variables together; 21% of the variability is accounted for by the Type-Biomass relationship; and only 3.6% of the variability is accounted for by the Location-Biomass relationship. It therefore looks like Location only has a *negligible* effect on the Biomass obtained, and this variable will be henceforth *disregarded*.

As an aside, here is the **Anova** output related to the above:

One-way Analysis of Variance (Location)

Analysis of Variance for biomass

Source	DF	SS	MS	F	P
location	2	817013	408507	0.93	0.401
Error	42	18353950	436999		
Total	44	19170963			

Individual 95% CIs For Mean Based on Pooled StDev

Level	N	Mean	StDev	CI Lower	CI Upper
1	15	887.5	374.5	140.0	635.0
2	15	924.8	901.8	22.0	1647.6
3	15	1190.1	597.9	992.2	1388.0

Pooled StDev = 661.1

One-way Analysis of Variance (Type)

Analysis of Variance for biomass

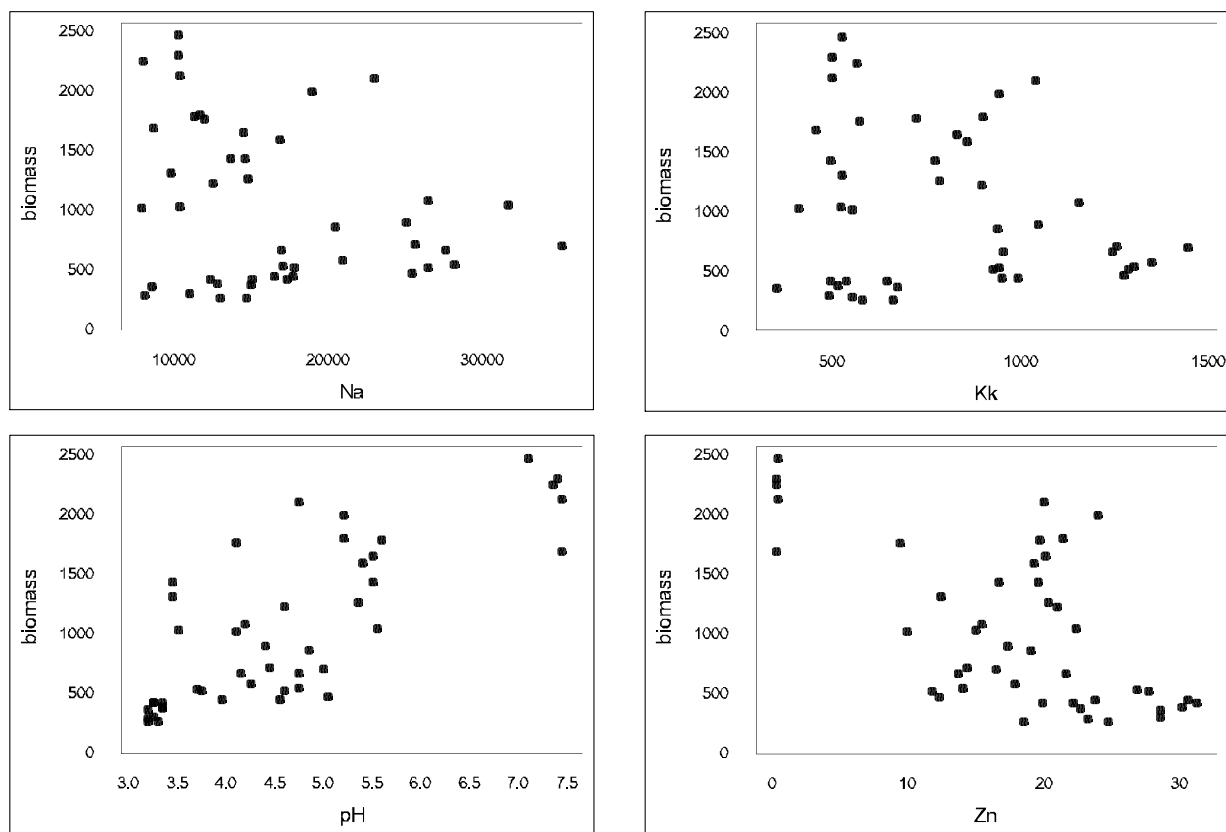
Source	DF	SS	MS	F	P
type	2	10875932	5437966	27.53	0.000
Error	42	8295031	197501		
Total	44	19170963			

Individual 95% CIs For Mean Based on Pooled StDev

Level	N	Mean	StDev	CI Lower	CI Upper
1	15	909.9	607.6	292.3	1127.5
2	15	449.3	140.2	309.1	589.5
3	15	1643.2	451.3	741.6	2544.8

Pooled StDev = 444.4

Let us now study whether any of the “**chemicals**” affect the dependent variable. To start with, let us look at the *plots* where we plot biomass against a “chemical”:



I used the `best subsets` command to regress the four above variables on biomass, as this selects automatically the best **single** “chemical”-biomass relationship; the best “two-chemicals”/biomass relationship; and so on. Here is the output obtained:

Best Subsets Regression

Response is biomass

Vars	R-Sq	Adj. R-Sq	C-p	s	p	K	N	Z
					H	k	a	n
1	59.9	59.0	6.7	422.63	X			
1	39.0	37.6	31.7	521.55				X
2	65.8	64.2	1.7	394.85	X	X		
2	64.8	63.1	3.0	401.07	X	X		
3	66.2	63.8	3.2	397.28	X	X	X	
3	66.0	63.6	3.5	398.51	X	X	X	
4	66.4	63.1	5.0	401.19	X	X	X	X

Looking at the **R²-values** in the output, we see that *66% of the variability* is accounted for by a model that takes into account all four variables — but note that we can account for 64% of the variability by using a model that takes into account two chemicals only. (The *third* row in the output). This small decrease is acceptable given the relative simplicity of the two-variable model as compared to the four-variable model.

In summary, the variables “*location*” and “*salinity*” do not affect the biomass, our dependent variable. **Type** does affect the biomass to some degree, but combining it with the two variable model obtained on the previous page produces negligible increases in R². Therefore, our model for prediction of biomass levels consists of an equation relating the biomass to the *acidity* and the *sodium*,

$$\text{Biomass} = 476 + 405 \times \text{Acidity} - 0.0233 \times \text{Sodium}.$$

Regression Analysis

The regression equation is
 $\text{biomass} = -476 + 405 \text{ pH} - 0.0233 \text{ Na}$

Predictor	Coef	StDev	T	P
Constant	-475.7	273.5	-1.74	0.089
pH	404.95	47.77	8.48	0.000
Na	-0.023326	0.008655	-2.70	0.010

S = 394.9 R-Sq = 65.8% R-Sq(adj) = 64.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	12622789	6311394	40.48	0.000
Residual Error	42	6548175	155909		
Total	44	19170963			

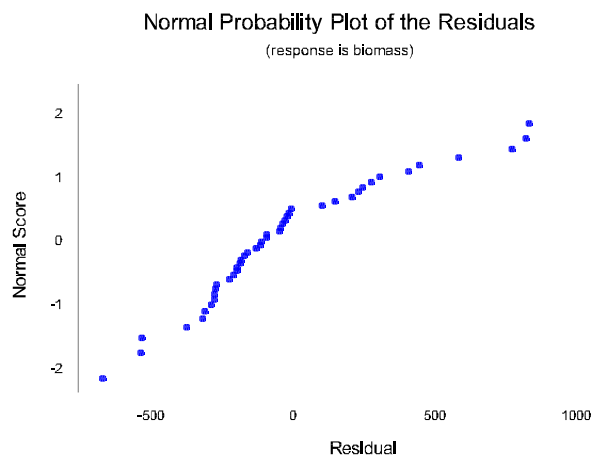
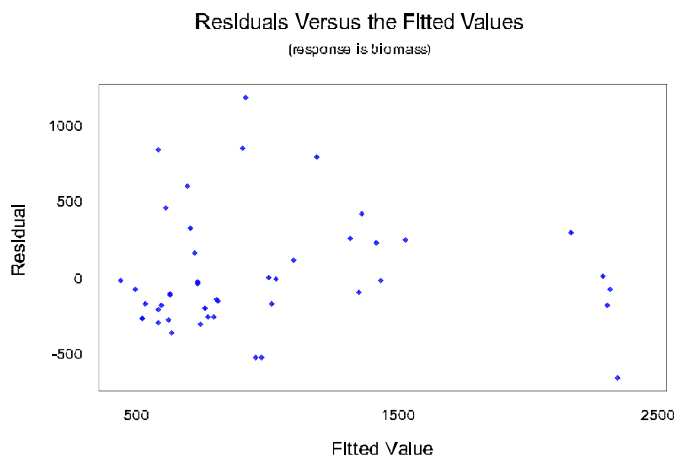
Source	DF	Seq SS
pH	1	11490388
Na	1	1132401

Unusual Observations

Obs	pH	biomass	Fit	StDev Fit	Residual	St Resid
12	3.45	1400.0	580.9	82.9	819.1	2.12R
14	4.10	1736.0	905.2	75.6	830.8	2.14R
34	4.75	2080.0	911.6	81.3	1168.4	3.02R

R denotes an observation with a large standardized residual

Here are the *usual two graphs* asked for with a regression analysis: a normal probability plot of the residuals, together with a residuals vs. fits plot:



Q: Perform a *forward selection*, a *backward elimination*, and a *stepwise regression* on the above data.

All of the above regression methods can be performed using the **Stepwise** command from the Regression section of the Stat menu. There is a useful page in the Minitab Help File clarifying how we should use the Stepwise command to perform the above methods of regression:

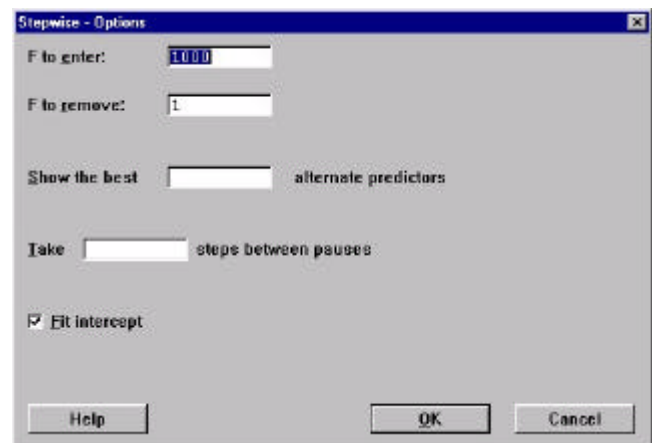
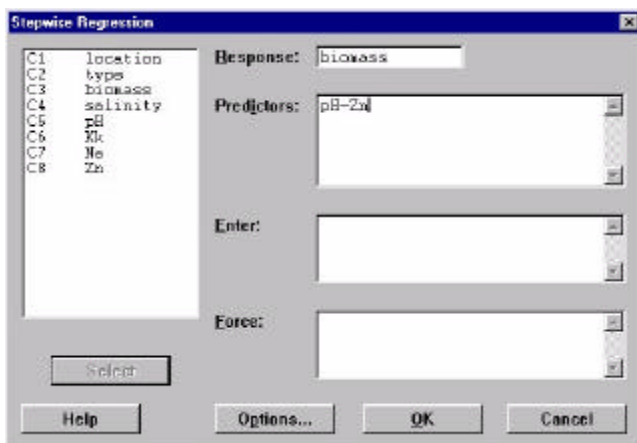
Stat > Regression > Stepwise

Stepwise regression removes and adds variables, for the purpose of identifying a useful subset of the predictors. Three commonly used procedures are provided: standard stepwise regression (adds and removes variables), forward selection (adds variables), and backwards elimination (removes variables).

1. **Stepwise:** In step one, an F-statistic for each predictor already in the model is calculated. If the F-statistic for any predictor is less than the value specified in the "F to remove" text box under <Options>, Minitab removes the predictor with the lowest F-statistic and prints output from the resulting model. In step two, Minitab calculates an F-statistic for each predictor not in the current model. If any value is greater than the value specified in the F to enter text box in the <Options> for any predictor, Minitab enters the predictor with the highest F-statistic and prints the output from the resulting model. These steps are repeated until no variables meet the criteria for addition or removal. Stepwise selection is the default. See Method Used by Stepwise for more information.
2. **Forward selection:** Adds predictors to the model as in Stepwise, but once added, a variable is never removed. The forward selection procedure ends when no additional variables have an F-value greater than F to enter. To do forward selection: Click the Options button and set "F to remove" to 0.
3. **Backwards elimination:** Begins with a model containing all possible predictors and removes them one at a time without re-entering any. Ends when no variable in the model has an F-value less than F to remove. To do backwards elimination: List all predictors in the Enter text box, click the Options button, and set "F to enter" to 10000 (a value virtually impossible to obtain).

In this question, consider that we want to find out if any of the four “*chemical*” variables are related to the **biomass** in any way. We can do this by using the three methods of regression.

- (1) **Forward Selection.** As suggested by the help file, we perform forward selection by setting the “F-to-remove” to zero in the options sub-screen. After doing this, we select the variables involved and click OK.



Stepwise Regression (Forward Selection)

```
F-to-Enter:      4.00    F-to-Remove:      0.00
Response is biomass on 4 predictors, with N = 45

      Step      1      2
Constant    -885.2  -475.7

pH          410     405
T-Value     8.02     8.48

Na          -0.0233
T-Value     -2.70

S           423     395
R-Sq       59.94    65.84
```

Using Forward Selection, the *first* iteration of the algorithm suggests an association with acidity. After the *second* iteration, the algorithm suggests an association with both acidity and sodium. The algorithm then terminates. Note that these were the variables used at the end of the previous question.

(2) Backward Elimination.

Stepwise Regression (Backward Elimination)

```
F-to-Enter:    1000.00    F-to-Remove:      4.00
Response is biomass on 4 predictors, with N = 45

No variables entered or removed
```

Using Backward Elimination, we start effectively with an equation linking the four “chemical” variables specified. Because the algorithm cannot find a relationship using three variables that has a better “fit” than the four-variable version, the algorithm terminates at the first step.

(3) Stepwise Regression (Normal).

Stepwise Regression

```
F-to-Enter:      4.00    F-to-Remove:      4.00
Response is biomass on 4 predictors, with N = 45

      Step      1      2
Constant    -885.2  -475.7

pH          410     405
T-Value     8.02     8.48

Na          -0.0233
T-Value     -2.70

S           423     395
R-Sq       59.94    65.84
```

Using Stepwise Regression, we arrive at the same relationship as what we arrived at at the end of the previous question, i.e. a relationship linking Biomass to Acidity and Sodium.

Q: Using Mallows's C_p criterion, and the minimum residual mean square criterion, which model would you choose to predict aerial biomass for the above data?

To see what we would *choose*, let us reprint the Best Subsets output obtained when performing analysis two questions ago:

Best Subsets Regression

Response is biomass

Vars	R-Sq	Adj. R-Sq	C-p	s	p	K	N	Z
					H	k	a	n
1	59.9	59.0	6.7	422.63	X			
1	39.0	37.6	31.7	521.55				X
2	65.8	64.2	1.7	394.85	X		X	
2	64.8	63.1	3.0	401.07	X	X		
3	66.2	63.8	3.2	397.28	X		X	X
3	66.0	63.6	3.5	398.51	X	X	X	
4	66.4	63.1	5.0	401.19	X	X	X	X

Using Mallows's C_p criterion, we would choose the **third** option, as this has the lowest C-p value. Using the *minimum residual mean square* criterion, we would also choose the **third** option, as this has the lowest 's' value in the table. These results agree with what we obtained previously.

Q: Perform a comparison of *regression lines* analysis on the data in O:\STATDATA\DATA\DADSON.MTW, where the first column gives the height in inches of 20 first year students at UWB in two different years as given in the third column; and the second column gives the height of each student's *father*.

After creating another column from the "year" column, changing 1969 to "0" and 1979 to "1", I regressed the heights on each other, including the adjusted year column as another *predictor*. This is the regression output, together with a plot of "dad's height" against "son's height" by year.

Regression Analysis

The regression equation is
 $son_ht = 34.2 + 0.516 \text{ dad_ht} - 0.219 \text{ year_adj}$

Predictor	Coef	StDev	T	P
Constant	34.190	8.412	4.06	0.000
dad_ht	0.5160	0.1208	4.27	0.000
year_adj	-0.2194	0.7277	-0.30	0.765

S = 2.297 R-Sq = 33.4% R-Sq(adj) = 29.8%

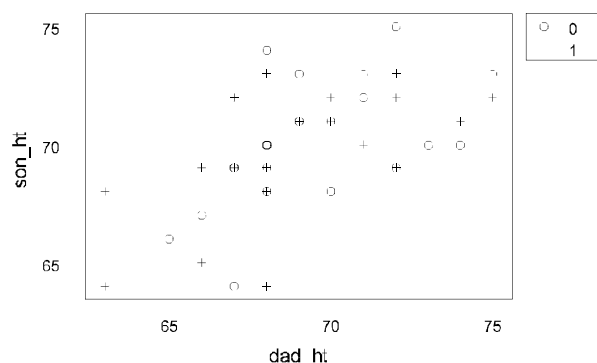
Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	97.855	48.927	9.27	0.001
Residual Error	37	195.245	5.277		
Total	39	293.100			

Source	DF	Seq SS
dad_ht	1	97.375
year_adj	1	0.480

Unusual Observations

Obs	dad_ht	son_ht	Fit	StDev Fit	Residual	St Resid
8	67.0	64.000	68.760	0.596	-4.760	-2.15R



Conclusion: For 1969, we have the relationship Son's Height = $34.2 + (0.516) \times \text{Dad's Height}$. For 1979, we have the relationship Son's Height = $(34.2 + 0.219) + (0.516) \times \text{Dad's Height}$. This seems to suggest that *on average*, heights grew by 0.219 inches in 10 years.

Chapter 4: Analysis Of Variance

Q: Show that the following equation holds: $\sum \sum (y_{ij} - y_{i..})^2 = \sum n_i (y_{i.} - y_{..})^2 + \sum \sum (y_{ij} - y_{i.})^2$.

The *left hand side* is the total sum of the responses $\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - y_{i..})^2$, and is used to derive a test for differences among groups. Adding and subtracting $y_{i.}$ within the square, and suitably bracketing the terms, the sum of squares can be written as $\sum_{i=1}^k \sum_{j=1}^{n_i} [(y_{ij} - y_{i.}) + (y_{i.} - y_{i..})]^2$. Expanding the square in terms of the *quantities* in the round brackets, we obtain

$$\begin{aligned} \sum_{i=1}^k \sum_{j=1}^{n_i} [(y_{ij} - y_{i.}) + (y_{i.} - y_{i..})]^2 &= \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - y_{i.})^2 + 2 \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - y_{i.})(y_{i.} - y_{i..}) + \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{i.} - y_{i..})^2 \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - y_{i.})^2 + 2 \sum_{i=1}^k \sum_{j=1}^{n_i} (0)(y_{i.} - y_{i..}) + \sum_{i=1}^k n_i (y_{i.} - y_{i..})^2 \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - y_{i.})^2 + \sum_{i=1}^k n_i (y_{i.} - y_{i..})^2. \end{aligned}$$

Explanation: The **middle** term disappears because $\sum_{j=1}^{n_i} (y_{ij} - y_{i.}) = 0$, and because the **final** term does not depend on j , we will have n_i copies of $\sum_{i=1}^k (y_{i.} - y_{i..})^2$ added together.

Q: Describe in words the *meaning* of each of the **terms** in the above question.

The term on the **left-hand side** is the familiar total sum of squares of the responses, denoted by S_{yy} . The first term on the right-hand side expresses the (weighted) squared differences among the group means, so is usually termed the *between-group sum of squares*, denoted by SS_T (in recognition of the fact that the groups are often “*treatment* groups”). The second term on the right-hand side can be recognised as the sum of squares *pooled within the groups*, so is called the *within-group sum of squares*, denoted by SS_E (in recognition of the fact that it is a measure of the **experimental error**). Thus the fundamental identity in one-way analysis of variance is $S_{yy} = SS_T + SS_E$.

Q: Perform a one way Anova on the barley data used in the *regression examples* above.

One-way Analysis of Variance

Analysis of Variance for C2

Source	DF	SS	MS	F	P
C1	5	40.657	8.131	68.72	0.000
Error	24	2.840	0.118		
Total	29	43.496			

Individual 95% CIs For Mean
Based on Pooled StDev

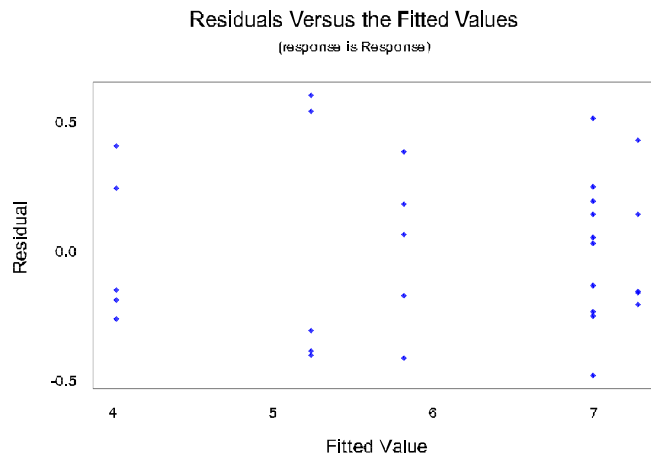
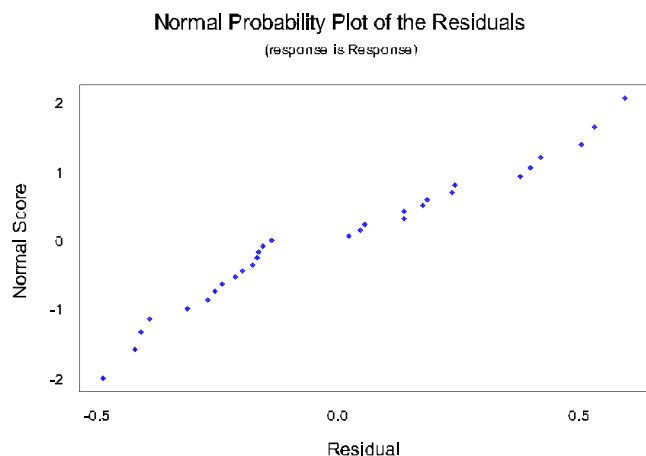
Level	N	Mean	StDev	CI
0	5	4.0150	0.2967	(- * - -)
40	5	5.2284	0.5140	(- - * -)
80	5	5.8146	0.3097	(- * - -)
120	5	6.9910	0.3837	(- * - -)
160	5	7.2712	0.2714	(- - * -)
200	5	6.9896	0.2020	(- * - -)

Pooled StDev = 0.3440

-----+-----+-----+-----
4.8 6.0 7.2

From the above, we see that there is *significant evidence* that the fertiliser levels in column 1 affect the yield levels in column 2: a high **F-value**; a low **probability** level (< 0.001); and a **small error SS** is the catalyst behind this suggestion.

To test whether the data is *normally distributed*, here are the graphs asked for with the Anova output: a **normal probability plot**, and a **residuals vs. fits plot**. As you can see, the normal plot is relatively *straight*, so it passes the “test”; and the residuals are *evenly distributed* about the *middle*, so this “test” is also passed.



Q: Decide whether a linear, quadratic, cubic or other model is the most appropriate for the barley data.

This has already been done in previous questions, and it was found that the quadratic model was the best one for the data. To recap, we used the **fitted line plot** command off the Regression menu. Specifying a linear model, the R^2 value obtained was 80.1%, which meant that 80.1% of the variability in the data was accounted for by the linear model.

Specifying a quadratic model, an R^2 value of 91.1% was obtained. Specifying a cubic model, an R^2 value of 92.2% was obtained. Due to the relatively large jump in the R^2 value between the *linear* and *quadratic* models, and the relatively small jump in the R^2 value between the *quadratic* and *cubic* models, I decided that the most appropriate model for the barley data was the quadratic model. The gain in simplicity of the model was worth it given the small jump in R^2 between the quadratic and cubic models.

This is the final model for the barley data: ($Y = \text{Yield}$, $X = \text{Fertiliser Level}$)

Polynomial Regression

$$Y = 3.94971 + 3.53E-02X - 9.71E-05X^{**2}$$

$$R\text{-Sq} = 91.1 \%$$

Q: The following table allows *linear*, *quadratic*, *cubic*, *quartic* and *quintic* orthogonal contrasts to be calculated for a factor with 6 levels. Confirm the findings for, say, the **linear** and **quadratic** effects from the Minitab output for the *barley* data.

<i>linear</i>	-5	-3	-1	1	3	5
<i>quadratic</i>	5	-1	-4	-4	-1	5
<i>cubic</i>	-5	7	4	-4	-7	5
<i>quartic</i>	1	-3	2	2	-3	1
<i>quintic</i>	-1	5	-10	10	-5	1

Quadratic Model for the Barley Data

Polynomial Regression

$$Y = 3.94971 + 3.53E-02X - 9.71E-05X^{**2}$$

$$R-Sq = 91.1 \%$$

Analysis of Variance

SOURCE	DF	SS	MS	F	P
Regression	2	39.6368	19.8184	138.648	6.22E-15
Error	27	3.8594	0.1429		
Total	29	43.4962			

SOURCE	DF	Seq SS	F	P
Linear	1	35.1325	117.616	1.56E-11
Quadratic	1	4.5043	31.5118	5.89E-06

Any *linear combination* of group totals $z_w = l_{w1}Y_{1.} + l_{w2}Y_{2.} + \dots + l_{wk}Y_{k.}$ is called a *linear comparison* or *linear contrast* (among the group totals) if the coefficients satisfy the condition $\sum_{i=1}^k n_i l_{wi} = 0$. **Two** contrasts z_1 and z_2 are said to be orthogonal if $\sum_{i=1}^k n_i l_{1i} l_{2i} = 0$.

Thus, for example, if there are three groups with equal sample sizes $n_1 = n_2 = n_3 = m$, say, such that group 1 is a “*control*”, while groups 2 and 3 correspond to two new *experimental* regimes A and B, then $z_1 = Y_{1.} - 1/2(Y_{2.} + Y_{3.})$ is a **contrast** between the control group and the two experimental regimes; $z_2 = Y_{2.} - Y_{3.}$ is a contrast between the two *experimental* regimes; and these two contrasts are orthogonal. We can partition the between-group sum of squares SS_T using the following results:

1. If $z_w = \sum_i l_{wi} Y_{i.}$ is *any* contrast among the $Y_{i.}$, and $D_w = n_1 l_{w1}^2 + n_2 l_{w2}^2 + \dots + n_k l_{wk}^2$, then the quantity z_w^2/D_w is a **component** of SS_T representing *one* degree of freedom.
2. If z_1 and z_2 are orthogonal, then z_2^2/D_2 is similarly a *one-degree-of-freedom* component of $SS_T - (z_1^2/D_1)$.
3. If SS_T has $k-1$ degrees of freedom, and if z_1, z_2, \dots, z_{k-1} are (*any*) mutually orthogonal contrasts, then $SS_T = \frac{z_1^2}{D_1} + \frac{z_2^2}{D_2} + \dots + \frac{z_{k-1}^2}{D_{k-1}}$.

In the *barley* example, we have $SS_T = 35.1325 + 4.5043$ as shown [above](#). What we want to show is that by using the *coefficients* given, and the above *formulae*, we can *calculate* the above two numbers. In the absence of knowing the **actual** data, we will estimate $Y_{i.}$ by $n(\bar{y}_i)$, where \bar{y}_i is the group mean as given by the Anova **output**.

So $z_1 = 5(4.0150)(-5) + 5(5.2284)(-3) + (5)(5.8146)(-1) + (5)(6.9910)(1) + (5)(7.2712)(3) + (5)(6.9896)(5) = 110.889$; $z_1^2 = 110.889^2 = 12296$. So $D_1 = 5[(-5)^2 + (-3)^2 + (-1)^2 + 1^2 + 3^2 + 5^2] = 5(70) = 350$; and the *Linear Component* of SS_T is therefore $12296/350 = 35.132$, which within rounding errors and estimation is *very* good.

Now $z_2 = 5[(4.0150)(5) + (5.2284)(-1) + \dots] = -43.495$; $z_2^2 = (-43.495)^2 = 1891.815$. And $D_2 = 5(5^2 + (-1)^2 + (-4)^2 + (-4)^2 + (-1)^2 + 5^2] = 84 \times 5 = 420$; so SS_T (*Quadratic*) = $1891.815/420 = 4.50$, which is again a good estimate.

Q: Perform a Two Factor Anova on the data in O:\STATDATA\DATA\CEREAL.MTW.

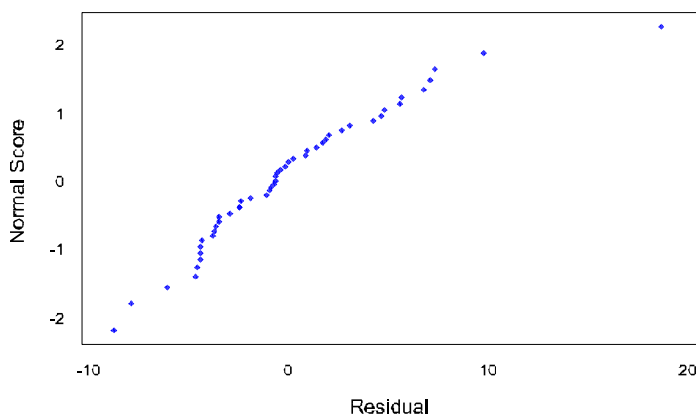
This is the **output** obtained:

Two-way Analysis of Variance

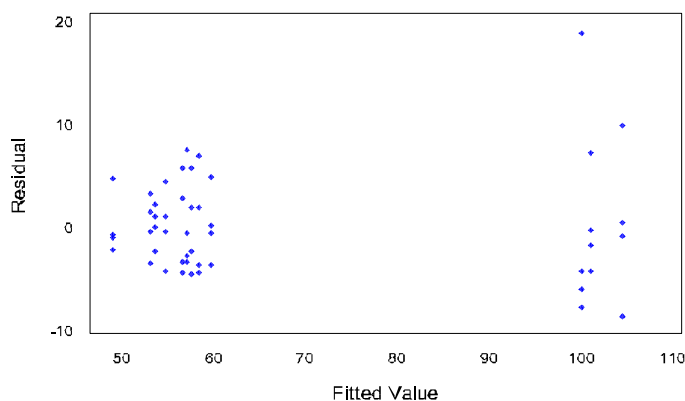
Analysis of Variance for height

Source	DF	SS	MS	F	P
cereal	3	19558.8	6519.6	205.54	0.000
depth	2	13.7	6.9	0.22	0.806
Interaction	6	120.2	20.0	0.63	0.704
Error	36	1141.9	31.7		
Total	47	20834.7			

Normal Probability Plot of the Residuals
(response is height)



Residuals Versus the Fitted Values
(response is height)



From the above, it seems that cereal *does* have an effect on the response variable “height”. However, neither the explanatory variable “depth” nor the *interaction* between cereal and depth has an effect on “height”. Further, large p-values suggest that the NH holds, so that there are no effects ($p < 0.05$). Therefore, I decided to perform a *one-way analysis of variance* using just the cereal. This is what I *obtained*:

One-way Analysis of Variance

Analysis of Variance for height

Source	DF	SS	MS	F	P
cereal	3	19558.8	6519.6	224.84	0.000
Error	44	1275.9	29.0		
Total	47	20834.7			

Individual 95% CIs For Mean
Based on Pooled StDev

Level	N	Mean	StDev	CI Lower	CI Upper
1	12	51.82	3.23	45.35	58.29
2	12	101.76	8.27	85.22	118.30
3	12	58.05	4.14	50.77	65.33
4	12	56.46	4.47	47.52	65.40

Pooled StDev = 5.38

Here, we have a larger F-value for cereal, and only a marginal increase in the Error SS, which is reduced anyway if we add in another explanatory variable. Therefore, I conclude that only **cereal** affects this data. The inclusion of depth in the model does nothing to help us predict future values.

Q: Explain the difference between **fixed** effect and **random** effect factors.

The process of *subsampling* removes the connections between the experimental units that exist in a cross-classification. For example, “*ward 1*” in the first hospital chosen in an experiment has nothing in common with “*ward 1*” in the second hospital chosen; “*leaf 1*” in the first tree chosen in an experiment has nothing to do with “*leaf 1*” in each of the **subsequent** trees chosen.

Since they are all chosen purely *randomly*, and no interest attaches itself to the particular ones actually selected, they are said to exhibit **random effects**. By contrast, all the levels of *factors*, or *blocks*, in a cross-classification, are common across all experimental units. Moreover, the actual levels used in the experiment are generally of interest in their own right, so are treated as having **fixed effects**.

Q: Explain why, before the advent of *computers*, the advice of statisticians was to collect **balanced** data.

Before the advent of electronic computers, the calculations for a multiple regression analysis were very *onerous*. Consequently, whenever possible, alternative (simple) computational methods were developed, such as Anova. Such methods almost always involved some constraints on the allowable forms of data.

We will be looking at Anovas requiring **balanced** arrangements, in which the same number of observations appear in each *block*, each *factor*, and so on. When this is so, the corresponding regression models are at least partially orthogonal, which simplifies matters considerably. Manual calculations are therefore much easier to do than if the Anova was not balanced, and we can at least get some sensible answers and results without having to use a computer.

Q: Write out the *oneway Anova* model for the barley data in a multiple regression format — i.e. what is the **design matrix X**?