31st January 2000

Introductory Lecture

Regression (quantitative from quantitative), **Analysis of Variance** (quantitative from qualitative), and the **General Linear Model** (quantitative from quantitative *and* qualitative) are Normal. Note that the Generalised Linear Model (not covered) is NOT normal.

Example: A drug company develops a drug to *lower heart rate* (H.R.). Measure the H.R., give the drug, and measure again, giving the change, (B-A). We do not have the **same** response for everybody, and the change for person i is not C units (b.p.m.) — it is C units ± variability. *Variability* is normally distributed.

Consider C - C_i, where C_i depends on the *gender/age/initial H.R.* The change in person i is C_i + $error_i$, e.g. $y_i = \mu_i + \varepsilon_i$. Populations/Samples; Variables (*quantitative*)/Factors (*qualitative*); Observational/Experimental; Linear/Non Linear. **Regression**: $y = \alpha + \beta x + \epsilon$ (α and β are parameters): a *unit* change in parameter leads to the **same** change in y for all x. $y = \alpha \exp(\beta x) + \epsilon$ (*additive* error): the change in y depends on x.

1st February 2000

Random Variables (r.v.'s)

Typically, we know the **range** of a measurement, but not an **exact** value for an individual, e.g. we know that Heart Rate > 0 but < 500. *Use a* lower case y for an *individual* measurement, e.g. $y_1, \ldots, y_n$. **Discrete**: Finite and Countably infinite. **Continuous**: any real number in the range. *Probability Distributions.* Example: a coin can be H or T, so *y* can be H or T. Pr(Y = y) = ½ (H) or ½ (T). f(y) = Pr(Y = y) is the **p.d.f.** (probability density function). $f(y) \geq 0$ and $\Sigma f(y) = 1$ (or $\int_R f(y)dy = 1$). The purpose of applied stats is to get information about a population from a sample.

Distributions. The *expected value* of a r.v. is $E(Y) = \Sigma_i y_i p_i = \mu$. ($\int y f(y) dy$). In the **table,** E(Y) = ¾. *Transform* Y to g(Y), then $E(g(Y)) = \Sigma_i g(y_i) p_i$. g(Y) could be Y^r (the r^{th} moment), or $(Y-\mu)^r$ (the r^{th} moment *about the mean*). Now $E((Y-\mu)^2) = \Sigma_i (y-\mu)^2 p_i = \sigma^2$ ($= \int (y-\mu)^2 f(y) dy$), the **variance**, V(Y). *Standard deviation* = √(variance). Properties: $E(a+cY) = a + cE(Y)$; $V(a+cY) = c^2 V(Y)$ (a and c are *constants*); $E(\Sigma a_i Y_i) = \Sigma a_i E(Y_i)$; and $V(\Sigma a_i Y_i) = \Sigma a_i^2 V(Y_i)$. $E(\bar{Y})$ is the *expected* value of **sample** means, while $V(\bar{Y})$ is the *variance* of the sample means. $\bar{Y} = (\Sigma Y_i)/n$. ($a_i = {}^1/_n$).

y_i	0	1
p_i	¼	¾

Normal Distribution

$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2\}$ (= $N(\mu, \sigma^2)$). Here, $E(Y) = \mu$ and $V(Y) = \sigma^2$. For the *standardised* normal, *where* $Z = {}^{Y-\mu}/_{\sigma}$, we have $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$ (= N(0,1) — mean 0 and s.d. 1).

2nd February 2000

3 Methods of estimation: *Moments*, *Least Squares*, *Maximum Likelihood.* We want to estimate the *population* mean. Obvious estimate: sample mean ($(\Sigma y_i)/n$). Equate the *sample* mean to E(Y), and equate $(\Sigma y_i^2)/n$ to $E(Y^2)$. In a distribution with **1** parameter, we need **1** equation. **2** parameters — need **2** equations. E(Y) & $E(Y^2)$ are *functions* of parameters.

Maximum Likelihood

Take a random *sample* from a distribution $f(Y,\theta)$. *1st observation:* $f(Y_1,\theta) \rightarrow f(y_1,\theta)$. **For** y_2, $f(Y_2,\theta) \rightarrow f(y_2,\theta)$, and so on, *until* y_n: $f(Y_n,\theta) \rightarrow f(y_n,\theta)$. Here, $f(y_1\ y_2 \dots y_n,\ \theta) = f(y_1,\theta).f(y_2,\theta)\dots f(y_n,\theta)$. **Define Likelihood** as $L(\theta) = \prod_{i=1}^{n} f(y_i,\ \theta)$. We want $\hat{\theta}$ as a *function* of $y_1 \dots y_n$; and want to *minimise* $L(\theta)$. $\hat{\theta}$ is the value of θ that gives the *highest possible probability* to the data obtained. If $\hat{\theta}$ is the maximum value of $L(\theta)$, then $\hat{\theta}$ is the maximum value of log $L(\theta) = l(\theta)$. **Standard** calculus works for "*nice*" functions. $\hat{\theta}$ is the root of $\partial l/\partial\theta = 0$, and $\partial^2 l/\partial\theta^2$ is -ve.

Take $y_1,\dots,y_n$, a *random sample* (r.s.) from $N(\mu,\sigma^2)$. $f(y,\ \mu,\ \sigma^2) = 1/\sigma\sqrt{(2\pi)} \times \exp\{-1/2\sigma^2 (y-\mu)^2\}$. $L(\theta) = \prod_{i=1}^{n}(1/\sigma\sqrt{(2\pi)})\exp\{-1/2\sigma^2\ \Sigma_i(y_i-\mu)^2\}$. And $l(\theta) = \log(1/\sigma\sqrt{(2\pi)})^n - 1/2\sigma^2\ \Sigma_{i=1}^{n}(y_i-\mu)^2 = -n\log\sigma - n/2 \log 2\pi$. Now $\partial l/\partial\mu = 1/\sigma^2\ \Sigma_i(y_i-\mu)$, and $\partial l/\partial\sigma = -n/\sigma + 1/\sigma^3\ \Sigma_i(y_i-\mu)^2$. **Further,** $\hat{\mu} = \bar{y} = (\Sigma y_i)/n$. So $\hat{\sigma}^2 = 1/n\ \Sigma_i(y_i - \hat{\mu})^2$. Check the *maximum* using $\partial^2 l/\partial\mu^2$ and $\partial^2 l/\partial\sigma^2$.

Let $Y_1\ Y_2 \dots Y_n$ **be iid** (Independent and *identically* distributed), with distribution $N(\mu,\sigma^2)$. Let $Y_i = \mu + \epsilon_i$. (ϵ_i is distributed as $N(0,\sigma^2)$). The *least squares estimate* (**lse**) is $\Sigma(y_i-\mu)^2$. (Minimise this).

<u>Bias</u>. *We have an unbiased* estimator if $E(\hat{\theta}) = \theta$ — if the **mean** of the *sampling distribution* is θ. We would like $E(\frac{\Sigma(y_i-\bar{y})^2}{n})$ to be σ^2. $E(1/n\ \Sigma(y_i-\mu+\mu-\bar{y})^2) \dots\dots E(1/n(\Sigma(y_i-\mu)^2 - \Sigma(\bar{y}-\mu)^2))$ ... $1/n \Sigma E((y-\mu)^2) - 1/n \Sigma E(\bar{y}-\mu)^2 = \sigma^2 - 1/n\,\sigma^2$.

7th February 2000

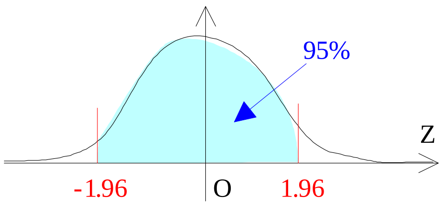
Central Limit Theorem

Let $\bar{Y} = (\Sigma_{i=1}^{n} y_i)/n$. This *converges* in probability to $N(\mu,\ \sigma^2/n)$ as $n \rightarrow \infty$. n = 5 may be all right if *symmetric & continuous;* n ~ 30 if *heavily skewed* or otherwise. (Note: the Y_i are i.i.d. as $N(\mu,\ \sigma^2)$, and the a_i are *constants*). **Let** $T = \Sigma a_i Y_i$, where $T \sim N(\mu\Sigma a_i,\ \sigma^2\Sigma a_i^2)$. Contrasts in **Anova**. **<u>Bias</u>**: *unbiased* if $E(\hat{\theta}) = \theta$. The Bias is $E(\hat{\theta})-\theta$. **<u>Mean Square Error</u>**: $E(\hat{\theta}-\theta)^2$ is the average squared *distance* of $\hat{\theta}$ from θ. We want small MSE.

<u>Cramer Rao Lower Bound</u>. The variance of any *unbiased estimator* cannot be <u>less</u> than $I^2(\theta)$, where $I^2(\theta) = -1/E(\partial^2 l/\partial\theta^2)$. If it *attains its bounds*, then it is <u>fully efficient</u>. Normal: $\partial l/\partial\mu = 1/\sigma^2\,\Sigma(y_i-\mu)$; $\partial^2 l/\partial\mu^2 = -1/\sigma^2.\Sigma 1 = -n/\sigma^2$; $I^2(\theta) = -1/E(\partial^2 l/\partial\mu^2) = -1/(-n/\sigma^2) = \sigma^2/n$. $\Sigma(y_i-\bar{y})^2/n$ is an *estimator* of σ^2. Note: the n in the *denominator* is **biased**, whereas (n-1) is **unbiased**. <u>Most</u> packages use (n-1).

<u>Consistency</u>. $\hat{\theta}$ is *consistent* if the MSE $\rightarrow 0$ as $n \rightarrow \infty$. The **larger** the sample, the **smaller** the distance of $\hat{\theta}$ from θ. $V(\bar{Y}) = \sigma^2/n \rightarrow 0$ as $n \rightarrow \infty$.

<u>Interval Estimate</u>. Let $\bar{Y} \sim N(\mu,\ \sigma^2/n)$; and let $Z = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$, *standardising*. ($Z \sim N(0,1)$). Now $Pr(-1.96 \le Z \le 1.96) = 0.95$. *95% confidence interval:* $(\bar{Y}-1.96\sigma/\sqrt{(n)},\ \bar{Y}+1.96\sigma/\sqrt{(n)})$. (Caution! We do not *know s*). *The sample statistic* $T = \frac{\bar{Y}-\mu}{s/\sqrt{n}}$ is the sample s.d. T has the **t-distribution**, with df = (n-1). And $(\bar{Y} - t_{n-1,\ \alpha} \times s/\sqrt{(n)},\ \bar{Y} + t_{n-1,\ \alpha} \times s/\sqrt{(n)})$ is a $100(1-\alpha)\%$ Confidence Interval (CI) for μ.



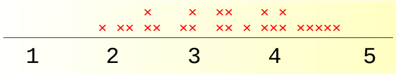
Hypothesis Testing. An investigator has a *theory*. Can it be substantiated using data? Formulate a model, with theory ≡ model parameters. We cannot obtain *absolute* evidence, only evidence comparative to some "*neutral*" position, the null hypothesis, H_0. The **theory** is the alternative hypothesis, H_a (or H_1). The NH is what we are *prepared* to believe unless there is sufficient evidence for the AH. **Specify** the NH and the AH. Choose a *test statistic* T s.t. T behaves differently under the NH and the AH. The sampling distribution is **known** for the NH.

8th February 2000

An Experiment to Analyse

24 patients with *h.l.p* (hypo lypo protenaemia) tend to have high levels of hypo proteins. (h.l.p. is associated with heart disease). We measure the plasma level of the total cholesterol, in mg/ml. We can plot **stem & leaf** plots (where the *first* value shown is 1.9) and **dot plots** of the data, as shown on the *right*. Guessing, the mean is between 3 and 3.5; and we have an idea of the standard deviation: it is approximately $^{range}/_4 \approx 0.9$. Actually, $\bar{Y} = 3.354$, $s^2 = 0.6052$, $s = 0.778$ and $n = 24$.

1	9
2	1 3 5 5 6 9
3	0 0 2 2 3 3 5 8 8 9
4	0 0 1 2 3 5 6

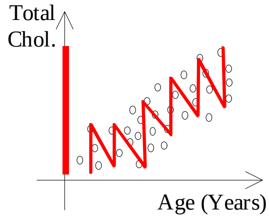


The 95% Confidence Interval for the *population mean μ* is $(\bar{Y} \pm t_{(23,\ 0.05)} \times {}^{s}/_{\sqrt{(n)}})$; $(3.354 \pm 2.069 \times ({}^{0.778}/_{\sqrt{(24)}}))$; (3.354 ± 0.329). Suppose that the doctor wishes to predict the *total cholesterol level* of the next patient admitted **with** h.l.p. Is the mean and the 95% CI related to *"predict next individual"?*

Let Y be the total cholesterol level, *a random* variable. It is unpredictable from one person to the **next**. *Assume* the Normal Distribution — the stem & leaf plot looks **roughly** normal. (We might want to look at literature to actually find that it is normal). So let $Y \sim N(\mu, \sigma^2)$; so that we get the $y_1, y_2, \ldots, y_{24}$ from $N(\mu,\sigma^2)$. Let $y_i = \mu + \epsilon_i$ *(observation = mean + departure), where* $\epsilon_i \sim N(0, \sigma^2)$. (The ϵ_i are *independent*). So $y_{25} = \mu + \epsilon_{25}$; $E(y_{25}) = E(\mu + \epsilon_{25}) = E(\mu) + E(\epsilon_{25}) = \mu + 0$. The **estimate** of μ is $\hat{\mu} = \bar{Y} = 3.354$ mg/ml.

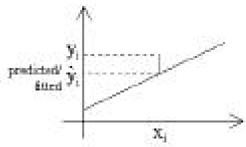
Accuracy of Prediction. Prediction is the estimate of μ + the estimate of ∈, and both are subject to *error*. The *variance* of $\hat{\mu}$ is $^{\sigma^2}/_n$ (from *earlier*). The *variance* of ∈ is σ^2. So the *variance* of the prediction is $^{\sigma^2}/_n + \sigma^2 = \sigma^2({}^{n+1}/_n)$. The estimated variance of *prediction is* $\hat{\sigma}^2({}^{25}/_{24})$. Now $\hat{\sigma}^2 = s^2$ (using *n-1*) $= 0.6052({}^{25}/_{24}) = 0.6304$. And the *s.d. of prediction* is $\sqrt{(0.6304)} = 0.794$.

95% Prediction Interval: $\bar{Y} \pm t_{(n-1,\ 0.05)} \times sd_{pred}$; $3.354 \pm 2.069 \times 0.794$. So values go from 1.711 to 4.997. **BUT**, does a person's total cholesterol level depend on something which could *account* for the variation *between* people, e.g. age as shown?



Cholesterol depends on *age* and *unexplained* variability. **data = model + residual;** $y_i = \alpha + \beta x_i + \epsilon_i$. (i = 1,...,n). After taking account age, the **large** variability in cholesterol becomes much **smaller**. Before, we had $y_i = \mu + \epsilon_i$ ($\epsilon_i \sim N(0, \sigma^2)$). Now, we have $\mu = \mu_i = \alpha + \beta x_i$. Remember to maintain *normality*, *independence* and *constant variance*.

Simple **linear regression model**: $y_i = \alpha+\beta x_i+\epsilon_i$ ($i = 1,\ldots,n$). ϵ_i is a mutually *independent* random variable, with **mean** 0 and constant **variance** σ^2. x_i is an *explanatory* variable — fixed, and measured <u>without</u> error. We do not require the **normality** of ϵ_i to get estimates of α and β. <u>Reparametrise</u>: $y_i = \beta_0+\beta_1(x_i-\bar{x})+\epsilon_i \Rightarrow \beta = \beta_1$ and $\alpha = \beta_0-\beta_1\bar{x}$. *Get estimates* of β_0 and β_1 using *Least Squares*. In the *graph*, a good line has the differences shown as **small** as possible.



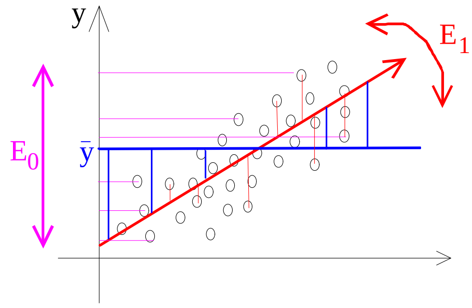
Predicted: $\hat{y}_i = \beta_0+\beta_1(x_i-\bar{x})$, and $\Sigma\epsilon_i = \Sigma(y_i-\hat{y}_i) = \Sigma(y_i-\beta_0-\beta_1(x_i-\bar{x}))$. *Choose* β_0 and β_1 so that we *minimise* $\Sigma\epsilon_i^2 = \Sigma(y_i-\beta_0-\beta_1(x_i-\bar{x}))^2 = S$. Take $\partial S/\partial\beta_0$ and $\partial S/\partial\beta_1$ and *equate to zero*. Now let $S_{xx} = \Sigma_i(x_i-\bar{x})^2$; $S_{yy} = \Sigma_i(y_i-\bar{y})^2$; and $S_{xy} = \Sigma_i (x_i-\bar{x})(y_i-\bar{y})$. Now $\hat{\beta}_0 = \bar{y}$, and $\hat{\beta}_1 = S_{xy}/S_{xx}$. Also, $\partial^2S/\partial\beta_0^2$ and $\partial^2S/\partial\beta_1^2$ are +**ve,** so that *estimates* give a <u>minimum</u> of $\Sigma\epsilon_i^2$.

14th February 2000

Assessing the Model

$\Sigma\epsilon_i^2$ should be a measure of how *good* the regression is. (**No** predictor — no 'x' — no age). Let $y_i = \mu+\epsilon_i$, with $\hat{\mu} = \bar{y}$. Now $\Sigma\epsilon_i^2 = \Sigma(y_i-\bar{y})^2 = S_{yy}$ (= E_0). Let us define a **regression** model as follows: x = age; $y_i = \beta_0+\beta_1(x_i-\bar{x})+\epsilon_i$; $\hat{y}_i = \bar{y} + (S_{xy}/S_{xx})(x_i-\bar{x})$; $e_i = y_i-\hat{y}_i = y_i-\bar{y}-(S_{xy}/S_{xx})(x_i - \bar{x})$; and $\Sigma e_i^2 = \Sigma(y_i-\bar{y}-(S_{xy}/S_{xx})(x_i-\bar{x}))^2$. *Square* and *simplify* to get $\Sigma e_i^2 = S_{yy}-(S_{xy}^2/S_{xx})$ (= E_1).

E_1 is never **larger** than E_0: S_{xy}^2 is +ve, and S_{xx} is a sum of *squares*. This suggests that including any x-variable will be **good** for prediction, as $E_1 \leq E_0$. But will x give us an **useful** prediction?



We can show that $E_0-E_1 = \Sigma_i(\hat{y}_i-\bar{y})^2$, where E_0 is the total *sum of squares* (**SS**) of y_i about $\bar{y}$, and E_1 is the *unexplained* variation around the **line**. The residual SS (*Error* SS, $\Sigma(\hat{y}_i-\bar{y})^2$) is the variation accounted for by the **line**. So $E_0 = E_1+\Sigma_i(\hat{y}_i-\bar{y})^2$; <u>**Total SS = Residual (Error) SS + Line (Regression SS).**</u>

<u>Total SS</u>: n deviations about $\bar{y}$, with df = (n-1). Variance: $\Sigma(y_i-\bar{y})^2/(n-1) = {}^{SS}/_{df}$. <u>Residual SS</u>: n deviations around $\hat{y}_i$, the line, which has *2 parameters:* $\hat{\beta}_0$ and $\hat{\beta}_1$, with df = (n-2). <u>Regression SS</u>: the effect of the line (the 2 parameters), with df = 2-1 = 1. <u>**Mean Square (MS):**</u> MS = ${}^{SS}/_{df}$. <u>**Residual MS**</u>: the average variation of the *observation* around the line, *estimating* $\hat{\sigma}^2$.

Let $y = \beta_0+\beta_1x+\epsilon$. ($\epsilon$ has *mean* 0 and *variance* σ^2). If there is **no** linear relationship between Y and X, then the <u>Regression MS</u> also estimates σ^2. If there **is** a linear relationship, then the Regression MS contains a *systematic component* as well as **random** variation, and so Regression MS >> Residual MS. So $F = {}^{\text{Regression MS}}/_{\text{Residual MS}} \sim F_{(\text{regression df, residual df})}$. (n-2, **Anova** table). **Null** Hypothesis: there is **no** linear relationship between Y and X. **Alternate** Hypothesis: there *is* a relationship between Y and X.

Assessing the Line

Goodness of fit: ${}^{\text{Regression SS}}/_{\text{Total SS}} = R^2$, the *coefficient* of *determination*.

15th February 2000

Inferences

$\hat{\beta}_1 = S_{XY}/S_{XX} = \Sigma(x_i-\bar{x})(y_i-\bar{y})/\Sigma(x_i-\bar{x})^2 = \Sigma(x_i-\bar{x})y_i/\Sigma(x_i-\bar{x})^2 - \Sigma(x_i-\bar{x})\bar{y}/\Sigma(x_i-\bar{x})^2$. Let us define $l_i = (x_i-\bar{x})/\Sigma(x_i-\bar{x})^2$, so we *have* $\hat{\beta}_1 = \Sigma l_i y_i - 0$. Now $E(\hat{\beta}_1) = E(\Sigma l_i y_i) = \Sigma l_i E(y_i)$ = (by the property of *estimators)* = $\Sigma l_i(\beta_0+\beta_1(x_i-\bar{x})) = \beta_0\Sigma l_i+\beta_1\Sigma l_i(x_i-\bar{x}) = \beta_0\Sigma(x_i-\bar{x})/\Sigma(x_i-\bar{x})^2 + \beta_1\Sigma(x_i-\bar{x})(x_i-\bar{x})/\Sigma(x_i-\bar{x})^2 = 0+\beta_1.1 = \beta_1$. And $V(\hat{\beta}_1) = V(\Sigma l_i y_i) = \sigma^2/S_{xx}$. ($\sigma^2$ estimated by the **residual MS**). A 95% CI for β_1 is $\hat{\beta}_1 \pm t_{(\text{residual df})}\times\sqrt{(\text{residual MS}/S_{xx})}$.

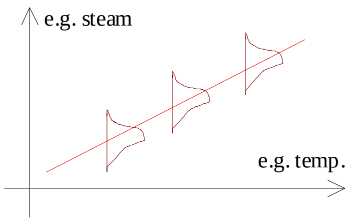
Prediction

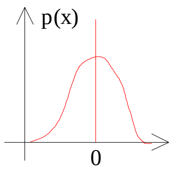
(1) *Mean* response ($\hat{\mu}_i$) for x_0. (2) *Individual* response for x_0: $\hat{\mu}_i = \hat{\beta}_0+\hat{\beta}_1(x_0-\bar{x})$. Now $V(\hat{\mu}) = V(\hat{\beta}_0+\hat{\beta}_1(x_0-\bar{x})) = V(\hat{\beta}_0)+(x_0-\bar{x})^2V(\hat{\beta}_1) = {}^{\sigma^2}/_n+(x_0-\bar{x})^2(\sigma^2/S_{xx})$. *Individual response*: we have $V(\hat{\mu}_i) = V(\hat{\beta}_0+\hat{\beta}_1(x_0-\bar{x})+\epsilon_i) = V(\hat{\beta}_0)+(x_0-\bar{x})^2V(\hat{\beta}_1)+V(\epsilon_i)$ = (*where* $V(\epsilon_i) = \sigma^2$) = $\sigma^2(1+{}^1/_n+[(x_0-\bar{x})^2/S_{xx}])$. (The 1 is "extra").

16th February 2000

There is data in **o:/statdata/G2M81**. **Trimmed Mean:** Ignore 5% of the data *at each end*. **Regression:** be sure to do a normal plot of the residuals (look for a *straight* line), and residuals versus fits (check for *curves*) graphs.

21st February 2000





Suppose you wanted to prove that **two** variables had a relationship. *Questions*: Does it follow a straight line relationship, and is there roughly equal variability along the line? Can we explain the *unexplained variability* (residuals) by a **third** variable?

Suppose that y = a+bx. Are the *residuals* $y_i-\hat{y}$ *related* to the z_i? Only now look at the relationship between x and the residuals — we want the *"extra effect"* over x. Do a **regression** of z on x. Look at the *residuals* $z_i-\hat{z}$. Then do a **regression** of $y_i-\hat{y}$ on $z_i-\hat{z}$. You *might* get e.g. $\hat{y}_i = 13-0.04x_i$; $\hat{z}_i = 22-0.8x_i$; and $y_i-\hat{y}_i = 0.2(z_i-\hat{z}_i)$. From these, we can *deduce* a new equation: $y_i = 9-0.8x_i+0.8z_i$ (3 *parameters*, so 2 d.o.f.).

We might also get $R^2_{x+z} = 0.849$ and $R^2_x = 0.714$, which means that 0.135 is *explained*. Doing an **Anova**, we might get under *Sequential SS*: X: 45.592, Z: 8.595. The X value is the SS **from** Y **on** X, or the **effect** of X ignoring Z. The Z value is the **extra** SS due to **adding** Z into a **model** containing X.

Unusual Observations: large residuals. **Standardised** residual = ${}^{\text{residual}}/_{\text{s.d. residual}}$ (look out for values > +2 or < -2). *Approximately* 95% of the standardised residuals lie **within** ±2 standard deviations, and 5% **outside** the ±2. Put observations with large influence on the *fitted* equation.

Multiple Regression

Let $y_i = \beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p x_{ip}+\in_i$. ($i = 1,\ldots,n$, and $\beta_0 = \beta_0.x_{i0}$ because $x_{i0} = 1$ *for all i*). We have **vectors y** ($n\times1$) = $(y_1,\ldots,y_n)^T$, and **β** ($n\times1$) = $(\beta_0,\beta_1,\ldots,\beta_p)^T$. We also have the *matrix* **X** as shown in yellow, and the **vector** ∈ ($n\times1$) = $(\in_1 \in_2,\ldots,\in_n)^T$. So using **y** = **Xβ**+∈, each *element* $\in_i$ of ∈ has *mean* (and *variance* s^2, and is also *independent* of all the other $\in_j$ ($i \neq j$).

$$X = \begin{pmatrix} 1 & x_{11} & x_{12}\ldots & x_{1p} \\ 1 & & & \\ \ldots & & & \\ 1 & x_{n1} & x_{n2}\ldots & x_{np} \end{pmatrix}$$

<u>Define</u> $f(y_1, y_2) = Pr(Y_1=y_1 \text{ and } Y_2=y_2)$, the *joint p.d.f.* $Pr(Y_1=y_1) = \Sigma_{y2} Pr(Y_1=y_1 + Y_2=y_2)$; $f_{Y1}(y_1) = \Sigma_{y2} f(y_1, y_2)$. *Similarly* for Y_2. **<u>Independence</u>**: if $f(y_1, y_2) = f_{Y1}(y_1).f_{Y2}(y_2)$ *leads to* $E(Y_1,Y_2) = \Sigma_{y1}\Sigma_{y2} y_1.y_2 f(y_1, y_2)$, define the **<u>covariance</u>** of Y_1 and Y_2 as $cov(Y_1, Y_2) = E(Y_1, Y_2) - E(Y_1).E(Y_2)$. **<u>Correlation</u>** of Y_1 and Y_2: $corr(Y_1, Y_2) = cov(Y_1, Y_2)/\sqrt{(V(Y_1).V(Y_2))}$. This is in the *range* -1 to 1. <u>Independence</u> ⇒ <u>uncorrelated,</u> **<u>but</u>** *uncorrelated* ⇏ *independence*.

The vector of departures/residuals is ∈ = $(\in_1,\ldots,\in_n)^T$. This is a *random* vector with <u>multivariable</u> distribution. The mean of the **random** vector is the vector of the means, $E(\in) = (0,\ldots,0)^T$, an $n\times1$ vector. cov(∈) is a covariance matrix (or a *variance-covariance* matrix). The **variances** of $\in_i$ sit along the *main diagonal*, while **covariances** of $\in_i$ and $\in_j$ are in the $(ij)^{th}$ position. But the $\in_i$ are *independent*, ∴ $E(\in_i\in_j) = E(\in_i)E(\in_j)$; ∴ $cov(\in_i\in_j) = 0$.

$$cov(\in) = \begin{pmatrix} \sigma^2 & & & 0 \\ & \sigma^2 & & \\ & & \ldots & \\ 0 & & & \sigma^2 \end{pmatrix}$$

The cov(∈) as shown on the *left* is **σ²I**, an $n\times n$ matrix. **<u>Properties of Z:</u>** (A random vector): cov(**Z**) = E(**ZZ**^{T})-E(**Z**)E(**Z**^{T}); E(**AZ**+**b**) = **Aμ**+**b**; cov(**AZ**+**b**) = **AΣA**^{T}, where E(**Z**) = **μ** (a *vector*), cov(**Z**) = Σ (not *sum*), **A** = constant <u>matrix</u>, and **b** = constant <u>vector</u>. **Regression**: **Y** = **Xβ**+∈, with E(∈) = 0, and cov(∈) = **σ²I**.

Here, E(**Y**) = E(**Xβ**+∈) = E(**Xβ**)+E(∈) = E(**Xβ**); and cov(**Y**) = cov(**Xβ**+∈) = **σ²I**. *Estimation* of **β**: <u>least squares</u>. $\Sigma_i \in_i^2 = S(\beta) = \in^T\in = (\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta) = \mathbf{y}^T\mathbf{y}-2\beta^T\mathbf{X}^T\mathbf{y}+\beta^T\mathbf{X}^T\mathbf{X}\beta$. <u>We can differentiate,</u> but it is easier if β_0 is s.t. $\mathbf{X}^T\mathbf{X}\beta_0 = \mathbf{X}^T\mathbf{y}$.

So $S(\beta)-S(\beta_0) = \beta^T\mathbf{X}^T\mathbf{X}\beta-2\beta^T\mathbf{X}^T\mathbf{y}+2\beta_0\mathbf{X}^T\mathbf{y}-\beta_0^T\mathbf{X}^T\mathbf{X}\beta_0$. *But* $\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\beta_0$, so $S(\beta)-S(\beta_0) = \beta^T\mathbf{X}^T\mathbf{X}\beta-2\beta^T\mathbf{X}^T\mathbf{X}\beta_0+2\beta_0^T\mathbf{X}^T\mathbf{X}\beta_0 = [\mathbf{X}(\beta-\beta_0)]^T[\mathbf{X}(\beta-\beta_0)]$ = *a sum of squares*, so ≥ 0. **Therefore,** $S(\beta_0)$ is a <u>minimum</u>, and we can also show that it is **unique**. Note: $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ **if** $\mathbf{X}^T\mathbf{X}$ is not *singular*. Show that $\hat{\beta}$ is *unbiased*: $E(\hat{\beta}) = E((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{X}^T.E(\mathbf{y}) = \beta$ (because $E(\mathbf{Y}) = \mathbf{X}\beta$). Now $cov(\hat{\beta}) = cov((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y})$. *Using* $cov(\mathbf{A}.\mathbf{y}) = \mathbf{A}.cov(\mathbf{y}).\mathbf{A}^T$, we have $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T.cov(\mathbf{y})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = r^2(\mathbf{X}^T\mathbf{X})^{-1}$. So $\mathbf{X}^T\mathbf{X}$ is *not* singular.

Assessing the Regression

Fitted values: $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}$. (Puts a *hat* on the **y**'s; the <u>hat matrix</u> **H**). **Residual: e** = $\mathbf{y}-\hat{\mathbf{y}} = \mathbf{y}-\mathbf{H}\mathbf{y} = (\mathbf{I}-\mathbf{H})\mathbf{y}$. Now $\Sigma(y_i-\bar{y})^2 = S_{yy}$ = *corrected SS*. The **vector** $\mathbf{y}^T\mathbf{y}-n\bar{y}^2$ is S_{yy}, so that $S_{yy} = \mathbf{y}^T\mathbf{y}-n\bar{y}^2 = [\mathbf{y}^T\mathbf{y}-\hat{\beta}^T\mathbf{X}^T\mathbf{y}]+\hat{\beta}^T\mathbf{X}^T\mathbf{y}-n\bar{y}^2$.

But $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$; *so* $\hat{\mathbf{y}}^T\hat{\mathbf{y}} = \hat{\beta}^T\mathbf{X}^T\mathbf{X}\hat{\beta}$. And $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, *therefore* $\hat{\mathbf{y}}^T\hat{\mathbf{y}} = \hat{\beta}^T\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \hat{\beta}^T\mathbf{X}^T\mathbf{y}$. The [...] $\Rightarrow \mathbf{y}^T\mathbf{y} - \hat{\beta}^T\mathbf{X}^T\mathbf{y} = \mathbf{y}^T\mathbf{y} - \hat{\mathbf{y}}^T\hat{\mathbf{y}} = \Sigma(y_i - \hat{y}_i)^2$ (residual SS). So $\hat{\beta}^T\mathbf{X}^T\mathbf{y} - n\bar{y}^2$ = *regression* SS.

Source	SS	df	MS = $^{SS}/_{df}$
Regression	$\hat{\beta}^T X^T y - n\bar{y}^2$	p	
Residual	$y^T y - \hat{\beta}^T X^T y$	n-p-1	
Total	$y^T y - n\bar{y}^2$	n-1	

Minitab

eqn. $\mathbf{y} = \mathbf{X} + \mathbf{Z}$

	coeff	s.e.	t	p
$\hat{\beta}_0$				
$\hat{\beta}_1$				
$\hat{\beta}_2$				

Anova Table

	df	SS	MS	F	p
regr					?
resid					
total					

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} c_{00} & & & \\ & c_{11} & & c_{ji} \\ & & c_{22} & \\ c_{ij} & & \cdots & c_{pp} \end{pmatrix}$$

In the *Anova* table, if '?' < 0.05, then there is a significant relationship at the 5% significance level. The **Null** Hypothesis is that $\beta_1 = \beta_2 = \ldots = \beta_p = 0$, i.e. **none** of the X_i's are related to y. The **Alternative** Hypothesis is that at least one β_j is non zero. *Earlier*, we saw that $E(\hat{\beta}) = \beta$ and $Var(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. In *particular*, $E(\hat{\beta}_j) = \beta_j$, and $Var(\hat{\beta}_j) = r^2.C_{jj}$.

$\hat{\beta}_j$-j is valid for the range 0...p. Look at the **boxed** matrix above. Assume that the departures $\in_j$ *are normally distributed*, then y_j is normal, with $\hat{\beta}_j$ *normal:* $\sim N(\beta_j, \sigma^2 C_{jj})$. *Therefore*, $\hat{\beta}_j - \beta_j/\sqrt{(\sigma^2 C_{jj})} \sim N(0,1)$, and $\hat{\beta}_j - \beta_j/\sqrt{(s^2 C_{jj})} \sim t_{n-p-1}$. *Null* Hypothesis: $\beta_j = 0$. *Alternative* Hypothesis: $\beta_j \neq 0$. This is a test of the *significance* of the x_j in the **presence** of all other the variables. If the $\hat{\beta}_j$ is not significant, then x_j does not *contribute significantly* to the variation in Y after all the other X's have been taken **into** account. X_j may have become significant if *some* of the other X's had been dropped from the model.

$\hat{\beta}$, $V(\hat{\beta})$, $V(\boldsymbol{\mu})$ and $V(\hat{\mathbf{y}})$ all need $(\mathbf{X}^T\mathbf{X})^{-1}$. This is singular if there is *linear dependence* among the **X**'s. Estimate enough **X**'s to get *non-singularity*. We have problems if it is "nearly singular", i.e. if there is approximate linear dependence among the x's. This is known as the causes of multi collinearity.

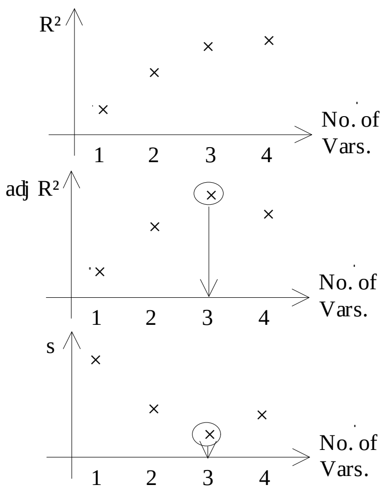
Some or all of the β's will have a large S.E.. *Instability:* a small change to an **X** value ⇒ a large change in the *fitted* model. How do we assess linear dependence among the **X**'s? R^2_j is the amount of variation in X_j explained by all the other X's. The **VIF** (the variance inflation factor) is $1/1-R^2_j$ ($R^2_j = 0.9$?). *Rule of Thumb*: Look for the VIF to be < 10. (Others **prefer** < 5).

1st March 2000

Polynomial Models

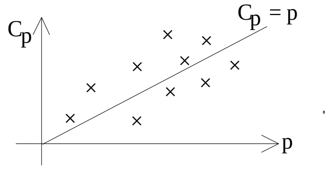
For $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots$, let $Z = x - \bar{x}$ so that $Y = \chi_0 + \chi_1 z + \chi_2 z^2 + \chi_3 z^3$.

Forward Stepwise Method



Forward Stepwise Method: Take the **best** single variable, then the variable which is next best in **conjunction** with the first; then thi one which is next best in **conjunction** with the second; etc. R^2 is the coefficient of *determination*. We want the first graph on the right tc be non-decreasing. Note that $R^2 = {}^{Regression\ SS}/_{Total\ SS} = 1 - {}^{Residual\ SS}/_{Total\ SS}$. Now adj $R^2_p = 1 - {}^{residual\ SS/(n-p)}/_{Total\ SS/(n-1)}$. (Note that the p *subscript* in adj R^2_p comes from (p-1) regressors **plus** 1 for β_0). The third graph concerns the *Residual Mean Square*, s^2. Note: Minitab gives s.

Mallow's C_p



If we have p-1 regressors (+1 for β_0), then E(Residual SS) = $(n-p)\sigma^2$. (s^2 is the *Residual MS* from the full model). So $E(R_{SS}/s^2) = {}^{(n-p)\sigma^2}/_c$ = n-p. And C_p = (Residual SS_p/s^2) + 2p - n; $E(C_p)$ = n-p+2p-n = p Choose a **model** with $C_p \approx p$.

6th March 2000

Models that Incorporate Factors

Plot data *by group*, overlaying the Graphs. Is there a suggestion of **parallel** lines? **Example**: We have cutter makes *A and B* (denoted by 1 and 2), represented by (0, 1) and (-1, 1). Regress lifetime (y) on speed (x) to get Life = 37-0.026x+15×(0, 1). So for cutter A, we have *y = 37-0.026x;* and for cutter B, we have *y = (37+15)-0.026x*.

Minitab: [CALC - MAKE INDICATOR VARIABLES]. Normal regression assumes *parallel* lines — we only get different **intercepts**. To get around this, let (empty column) = (x column) × (e.g. "medium" column). Regress on these as well, and get e.g. *y = 158+42.5x-174mi-63hi+38mi_x-22hi_x*. So for **Lo**, y = 158+452x; for **Mi,** y = (158-174)+(42.5+3.8)x; and for **Hi,** y = (158-63)+(42.5-22.6)x. Are these *parallel* lines — look at the sequential SS. Use Minitab to look up the F-values.

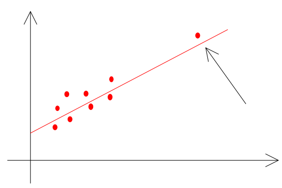
7th March 2000

We will ignore *section 3.7*, non-linear models. Consider the model $\mathbf{y} = \mathbf{X}\beta+\epsilon$. We have residuals $\mathbf{e} = \mathbf{y}-\hat{\mathbf{y}}$ (*where* $\hat{\mathbf{y}} = X\hat{\beta}$) = $(\mathbf{I}-\mathbf{H})\mathbf{y}$. (*where* $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, and $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$). Now $E(\mathbf{e}) = E[(\mathbf{I}-\mathbf{H})\mathbf{y}] = (\mathbf{I}-\mathbf{H})E(\mathbf{y}) = (\mathbf{I}-\mathbf{H})\beta = \mathbf{X}\beta-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta = \mathbf{X}\beta-\mathbf{X}\beta = \mathbf{0}$. And $\text{cov}(\mathbf{e}) = \text{cov}[(\mathbf{I}-\mathbf{H})\mathbf{y}] = (\mathbf{I}-\mathbf{H})\text{cov}(\mathbf{y})(\mathbf{I}-\mathbf{H})' = (\mathbf{I}-\mathbf{H})\sigma^2\mathbf{I}(\mathbf{I}'-\mathbf{H}') = \sigma^2(\mathbf{I}-\mathbf{H})(\mathbf{I}'-\mathbf{H}')$.

Here, $\mathbf{I}' = \mathbf{I}$, and $\mathbf{H}' = (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' = (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \mathbf{H}$. *So* $\text{cov}(\mathbf{e}) = \sigma^2(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H}) = \sigma^2(\mathbf{I}-\mathbf{H}-\mathbf{H}+\mathbf{H}.\mathbf{H})$. *As* $\mathbf{H}.\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}$, *then* $\text{cov}(\mathbf{e}) = \sigma^2(\mathbf{I}-\mathbf{H})$. The diagonal entries are $\sigma^2(1-h_{ii})$, while the off-diagonal entry is $\sigma^2 h_{ij}$. ϵ: *zero* mean; *constant* variance σ^2; and *independent*. **e**: *zero* mean; *non-constant* variance $\sigma^2(1-h_{ii})$; *not* independent.

It follows that $\text{var}(e_i) = (1-h_{ii})$, and that $\text{cov}(e_i e_j) = -h_{ij}.\sigma^2$. The **dependence** is small, especially if n is large, "*ignorable*". The standardised residuals (studentised) are e'_i, where $e'_i = \frac{e_i}{s\sqrt{(1-h_{ii})}}$, and s is our estimate of σ from **the Residual MS**. ($e'_i \sim N(0,1)$).

For a residual plot, Minitab asks you whether you want regular/standardised/deleted residuals. The default is regular — use standardised instead! (They have **constant** variance). Minitab calls a point an "outlier" if its standardised residual is > 2 or < -2. (Note: the book uses 2.5 instead). *Roughly* 99% of the standardised residuals will be within ± 2.5 s.d.'s from the mean. **Leverage**: based on h_{ii}. ($h_{ii} > {}^{3(p+1)}/_n$, where p is the number of *regressor variables*, and n is the *sample size*). For *simple* linear regression, $h_{ii} = {}^1/_n + \frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}$.



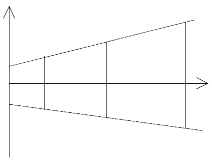
Cook's Distance

Cook's Distance is $D_k = \frac{1}{(p+1)s^2}\sum_{i=1}^{n}(\hat{y}_{i(k)} - \hat{y}_i)^2$, where $\hat{y}_i$ is the *fitted* value, and $\hat{y}_{i(k)}$ is the fitted value when observation k is **omitted**. Now $D_k = (e'_k)^2(h_{kk}) \times {}^1/_{(1-p)(1-h_{kk})}$, *where* $(e'_k)^2$ comes from the **standardised residuals;** (h_{kk}) is the **leverage;** and ${}^1/_{(1-p)(1-h_{kk})}$ is standardised. Now $D_k \sim F(0.5, p, n\text{-}p)$. If D_k is larger than this, then we have a *potential problem.*

13th March 2000

Some Tips on Anova (The Tyre Example)

Descriptive stats — wear by position. Rule of thumb — it is all right if the largest s.d. is not more than 2 × (the smallest s.d.). One-way / One-way unstacked: depends on how you enter the data: (**unstacked**) in different *columns* rather than *on top* of each other. Tip: use the **stacked** method. Position MS = n × (variance of the means). (n = 9). Error MS = mean of the 4 position variances.



Now **perform** the Anova, starting by checking the assumptions. *Equal variability:* residuals vs. fits. Look out here for the "megaphone effect" as shown. *Normal Probability plot:* make sure there is a straight line. *Dotplots:* similar to residuals vs. fits. Test: *Homogeneity of variance* — are variances the same? Use **Bartlett's** for the normal distribution, and **Levene's** for other distributions. If $p > 0.05$, then accept the NH (equal variances).

Put the *residuals* in a column, then use the descriptive statistics (a **graphical** summary) to test normality. Comparisons: which position is *different* from another? (This is a crude method).

14th March 2000

Anova

If we want, we can use *regression commands* with indicator (0-1) variables to get the Anova tables. In the General Linear Model, *Anova* and *Regression* are branches. Now $(y_{ij}-y_{..}) = (y_{ij}-y_{i.})+(y_{i.}-y_{..})$. We can **show** that $\Sigma_i\Sigma_j\,(y_{ij}-y_{..})^2 = \Sigma_i\Sigma_j\,(y_{ij}-y_{..})^2 + n\Sigma_i\,(y_{i.}-y_{..})^2$. **Residual** MS = mean of *variances*. **Treatment** MS = n×(*variance* of the *means*).

15th March 2000

(Tyre Example) Error MS = 29 for **One Way** Anova. Two Way: the Error MS reduces to 25.7. Ask for *means*, and fit the *additive* model at first. Use cross tabulation for analysis. **Interactions Plot**: Data Means for Wear. Look for *differences*. Do not use the additive model in this example. The Error MS is now 1.09. Use Interaction (*car* by *position*). Shorthand Notation in Minitab: car + position + car*position is written as car ! position (use in **Balanced** Anova).

Note: Ignore *sections 4.4 and 4.6* of the book.

20th March 2000

Contrasts

Tyre Wear example: 4 positions. 1 way Anova: does **wear** depend on **position**? Question: are (AH) / are there not (NH) differences in the mean wear in the 4 positions? We usually want to go further than this, and ask *new* questions.

	F	R	
O	1	-1	1
N	-1	1	-1
	1	-1	

C_1 compares *Front and Rear*. C_2 compares *Off and Near*. We have orthogonality if $\Sigma c_1.c_2 = 0$. There is only *one other* orthogonal contrast: C_3: 1,-1,-1,1.

Pos.	**F/R**	**N/O**	**$Y_i.$**	**C_1**	**C_2**
1	F	O	194.59	1	1
2	F	N	157.66	1	-1
3	R	O	293.2	-1	1
4	R	N	249.43	-1	-1
			894.89		

Let $l_i \to 1, -1$. So $Z_{F/R} = \Sigma l_i \times y_i$; $D_{F/R} = n(\Sigma l_i^2)$; and $^{Z^2}/_D$ = SS due to Front vs. Rear. The SS due to *position* can be **partitioned** into the SS due to *Front vs. Rear*; the SS due to *Off vs. Near*; and the SS due to *interaction*.

Paired Comparisons

In the table, the *first* column denotes position; the *second* column compares position 1 to position 2; the *third* column compares position 3 to position 4; and the *fourth* column is orthogonal to the previous constraints.

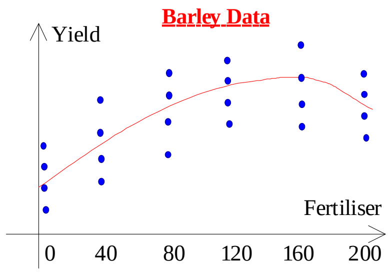
1	1	0	1
2	-1	0	1
3	0	1	-1
4	0	-1	-1

Another example: Suppose we have fertiliser levels 0, 40, 80 and 120. (A *quantitative* factor). Is there a linear or quadratic or cubic relationship between the response and a factor?

Treatment	Lin.	Quad.	Cub.
1	-3	1	-1
2	-1	-1	3
3	1	-1	-3
4	3	1	1

Regression

Is the regression **cubic**? In this model, we have *linear, quadratic, cubic and residual* terms. The **residual** term can be split up into a pure error term and a lack of fit term (which comes from replication). We "want" the lack of fit term to be *non-significant*.



Fixed and Random Effects (1-Way Anova)

In the **tyre wear** example, we have 4 positions. Also, $y_{ij} = \mu_i + \epsilon_{ij}$ (i = 1,2,3,4), and $\epsilon_{ij} \sim N(0, \sigma^2)$. It is common to *rewrite this* as $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$. ($\alpha_i$ is the effect of *position* i). For balanced problems (where we have an equal number of replicates), $\Sigma\alpha_i = 0$. The *expected mean squares* are: **Residual**: σ^2; **Treatment**: $\sigma^2 + n\Sigma\alpha_i^2$.

"A survey of children's behaviour": a teacher fills out a questionnaire on each child. For each teacher, we have a **number** of children. How different is the variation for each child? What is the mean rating over all teachers? Remember that we only have a **sample**, with $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$; $\epsilon_{ij} \sim N(0, \sigma^2)$; and $\tau_i \sim N(0, \sigma^2_\tau)$.

Expected Mean Squares: Teachers: $\sigma^2 + n.\sigma^2_\tau$; *Residual*: σ^2. (A **Variance Component Model**). Fixed: *There's a small* number of values you're interested in. Random: There is potentially a *large* number. We are "not particularly interested" in the actual ones you've got.

22nd March 2000

Balanced Anova will not work with *unbalanced* data. (Where we have unequal sample sizes). In this situation, use the General Linear Model instead, remembering to use ! to get the *interactions.*

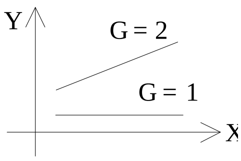
<u>Sequential SS</u>

car ignoring all others
fr accounting for car
on accounting for car and fr

<u>Adjusted SS</u>

car accounting for all others
fr accounting for all others
on accounting for all others

If we have *2 columns of data*, X and Y, then if the response is Y, the model X, and the covariant X, we have a <u>simple linear regression</u>. If we add a third column G (some group variable), then the response may be Y, the model X ! G, and the covariant X. If X*G is significant, then this implies that we **don't** have parallel lines.



Exam Paper: May 2000

Answer 3 questions out of 4 (Questions Done: 1, 2, 3)

(1) You have taken a random sample of n observations $y_1, y_2, \ldots, y_n$ from a normal distribution with mean μ and variance σ^2.

(a) Find the maximum likelihood estimators of μ and σ^2. **[12 marks]**

(b) State the expected value of the m.l.e. of σ^2. **[3 marks]**

(c) State the effect this has in common statistical usage. **[5 marks]**

(2) The multiple linear regression model can be expressed as $y = X\beta+\varepsilon$ where y is a vector of response measurements y_i, X is a matrix of row vectors of explanatory variables l, $x_{i1}, x_{i2}, \ldots, x_{ip}$, $\beta_0 \ldots. \beta_p$ is a vector of parameters and ε is a vector of random departures.

(a) Show that the least squares estimates of β are $\hat{\beta} = (X^tX)^{-1}X^ty$. **[12 marks]**

(b) Show that $\hat{\beta}$ is unbiased and that the covariance matrix of $\hat{\beta}$ is $\sigma^2(X^tX)^{-1}$. **[8 marks]**

(3) For a one factor Anova model where n observations have been randomly sampled from each of k groups, show that the total SS = $\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-y_{..})^2$ can be partitioned into two parts $\sum_{i=1}^{k} n(y_{i.}-y_{..})^2$ and $\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-y_{i.})^2$. **[8 marks]**

Describe in words the meaning of each of the two parts. **[4 marks]**

In an experiment at College farm 30 plots of land were used to study the effect of nitrogen fertiliser on barley yield. Six fertiliser levels were used (0, 40, 80, 120, 160, 200 kg/ha) and the yield of barley measured in tons/ha. The table below shows the mean and variance of the yield of barley for each fertiliser level. These were 5 replicates for each fertiliser level. The total SS is 43.50.

	0	40	80	120	160	200
Mean	4.02	5.23	5.81	6.99	7.27	6.99
Variance	0.088	0.264	0.096	0.147	0.074	0.041

Using the above information write out the analysis of variance table, state any further appropriate analyses you would perform and summarise your conclusions about the relationship between fertiliser level and barley yield. **[8 marks]**

(4) You have performed an experiment to examine the relationship between areal biomass and five chemical measurements of the soil. Show below is a table containing the regression SS for each model. Using the information that the total SS is 19 170 963 and that 45 observations were made,

(a) Write out the forward regression Anova table; **[8 marks]**

(b) Which model would you choose from the Anova table in (a) using:
(i) The residual mean square criterion. **[4 marks]**
(ii) Mallows C_p statistic. **[4 marks]**

(c) Which model would you recommend using from your answers to (a) and (b). **[4 marks]**

In the table below the following abbreviations have been used:
Salinity = S, pH = P, Potassium = K, Sodium = N and Zinc = Z.

1 Predictor	
S	204 048
P	11 490 388
K	806 574
N	1 419 069
Z	7 474 474
2 Predictors	
S P	11 567 715
S K	1 027 686
S N	1 487 655
S Z	10 594 197
P K	12 415 118
P N	12 622 789
P Z	11 661 321
K N	1 424 635
K Z	7 961 377
N Z	12 205 083
5 Predictors	
S P K N Z	12 984 915

3 Predictors	
S P K	12 503 299
S P N	12 634 567
S P Z	12 205 083
S K N	1 487 757
S K Z	11 052 633
S N Z	10 820 655
P K N	12 659 874
P K Z	12 501 663
P N Z	12 699 814
K N Z	8 244 127
4 Predictors	
S P K N	12 685 656
S P K Z	12 937 009
S P N Z	12 878 489
S K N Z	11 068 315
P K N Z	12 732 925