STATISTICAL ANALYSIS AND MODELLING

David A. Stephens

Department of Mathematics, Imperial College

d.stephens@imperial.ac.uk
stats.ma.ic.ac.uk/∼das01/EPSCourse/

11th March 2004

WEEK 2: Regression Analysis

- Linear Regression
- Generalized Regression
- Classification
- Multivariate Analysis

SECTION 1.

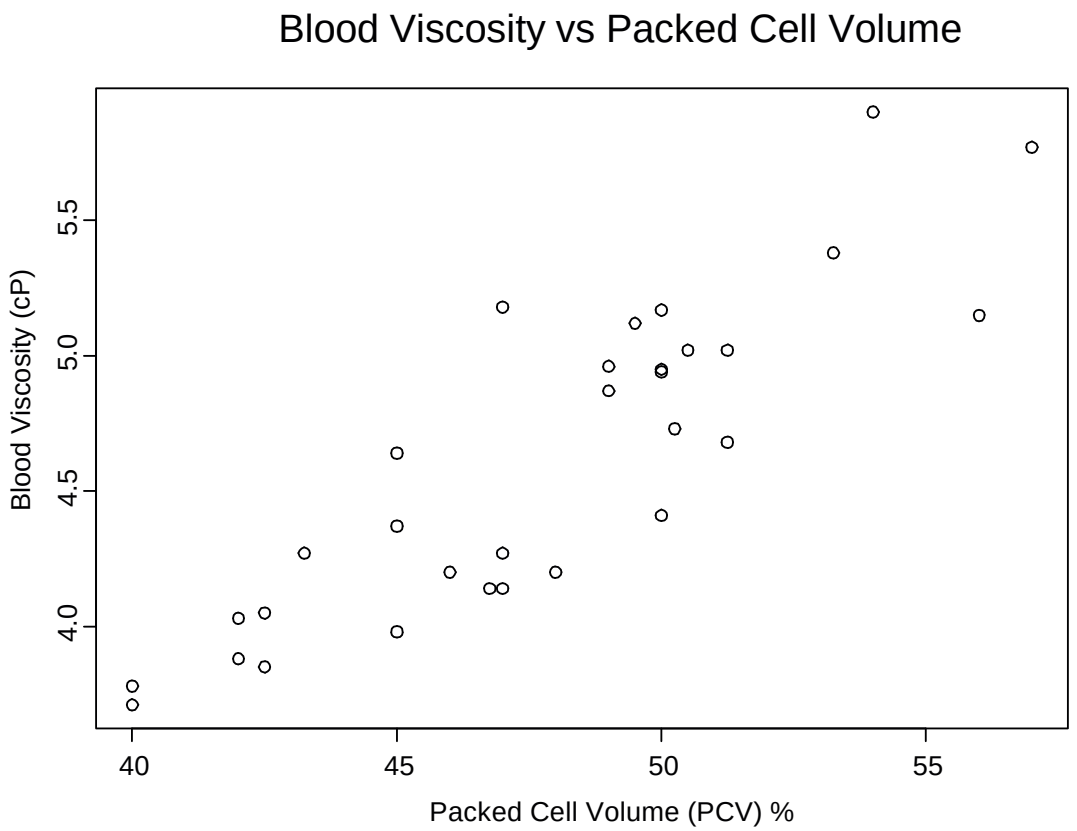
REGRESSION MODELLING

Aim: To explain the **systematic** variation of one observed variable with another in the presence of **random** variation

- two related samples (predictor-response)
- simplest case - a linear ("straight-line") relationship
- typically assume **normal random errors**
- extension to non-linear relationships
- extension to non-normal data
- lead into multivariate modelling

1.1 LINEAR REGRESSION

EXAMPLE



1.1.1 TERMINOLOGY AND NOTATION

Y is the **response** or **dependent** variable

X is the **predictor, covariate** or **independent** variable

A simple relationship between Y and X is the **linear regression model**, where

$$\mathrm{E}[Y|X = x] = \alpha + \beta x,$$

that is, conditional on $X = x$, the expected or "predicted" value of Y is given by $\alpha+\beta x$, where α and β are unknown parameters; in other words, we model the relationship between Y and X as a straight line with **intercept** α and **slope** β.

For data $\{(x_i, y_i) : i = 1, ..., n\}$, the objective is to estimate the unknown parameters α and β. A simple estimation technique, is **least-squares estimation**.

1.1.2 LEAST-SQUARES ESTIMATION

Suppose that a sample, $\{(x_i, y_i) : i = 1, ..., n\}$, is believed to follow a linear regression model, $\mathrm{E}[Y|X = x] = \alpha + \beta x$. For fixed values of α and β, let $y_i^{(P)}$ denote the expected value of Y conditional on $X = x_i$, that is

$$y_i^{(P)} = \alpha + \beta x_i$$

Now define error terms e_i, $i = 1, ..., n$ by

$$e_i = y_i - y_i^{(P)} = y_i - \alpha - \beta x_i$$

that is, e_i is the vertical discrepancy between the **observed** and **expected** values of Y.

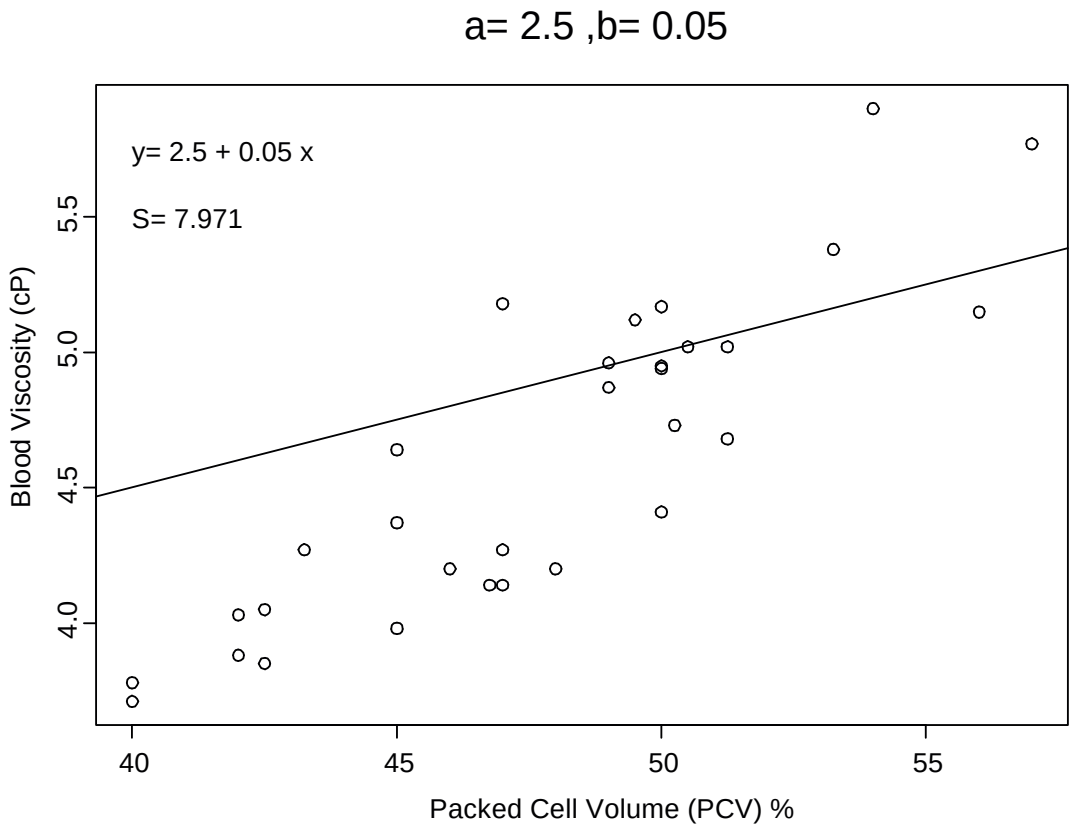
The objective in least-squares estimation is find a "line of best fit", and this is achieved by inspecting the squares of the error terms e_i, and choosing α and β such that the sum of the squared errors is **minimized**.

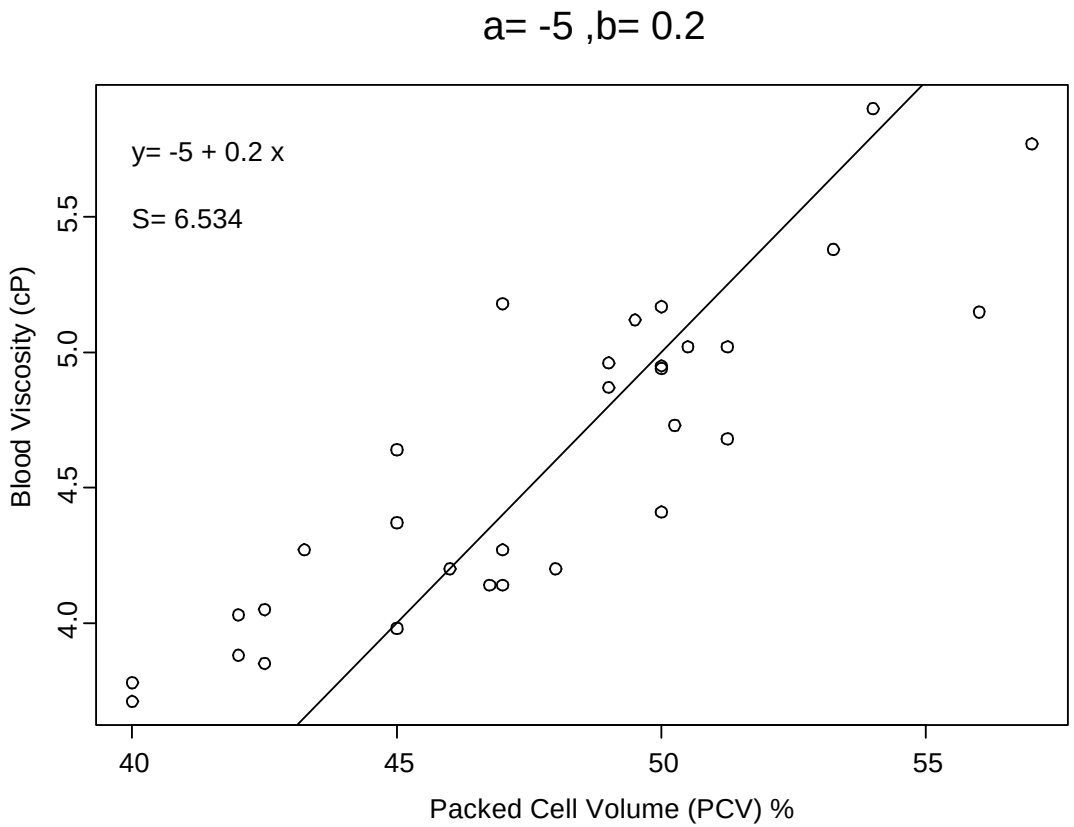
We aim to find the straight line model for which the total error is smallest.

Let $S(\alpha, \beta)$ denote the error in fitting a linear regression model with parameters α and β. Then

$$S(\alpha, \beta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - y_i^{(P)})^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Different values of α, β give different S values; we aim to choose the "best" pair of parameters


a= 2.5 ,b= 0.05
y= 2.5 + 0.05 x
S= 7.971
Blood Viscosity (cP)
4.0
4.5
5.0
5.5
40
45
50
55
Packed Cell Volume (PCV) %


a= -5 ,b= 0.2
y= -5 + 0.2 x
S= 6.534
Blood Viscosity (cP)
4.0
4.5
5.0
5.5
40
45
50
55
Packed Cell Volume (PCV) %

To calculate the least-squares estimates, we have to minimize $S(\alpha, \beta)$ as a function of α and β. This can be achieved in the usual way by taking partial derivatives with respect to the two parameters, and equating the partial derivatives to zero simultaneously.

$$(1)\frac{\partial}{\partial\alpha}\{S(\alpha,\beta)\} = -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0$$

$$(2)\frac{\partial}{\partial\beta}\{S(\alpha,\beta)\} = -2\sum_{i=1}^{n} x_i(y_i - \alpha - \beta x_i) = 0$$

Solving (1), we obtain an equation for the least-squares estimates $\widehat{\alpha}$ and $\widehat{\beta}$

$$\widehat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} y_i - \widehat{\beta}\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{y} - \widehat{\beta}\bar{x}.$$

Solving (2) in the same way, and then solving for $\widehat{\beta}$ gives

$$\widehat{\beta} = n\frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left\{\sum_{i=1}^{n} x_i\right\}^2} = \frac{nS_{xy} - S_x S_y}{nS_{xx} - \{S_x\}^2}$$

so that

$$\widehat{\alpha} = \frac{\sum_{i=1}^{n} x_i y_i - \widehat{\beta}\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i} = \bar{y} - \widehat{\beta}\bar{x}$$

where

$$S_x = \sum_{i=1}^{n} x_i \qquad S_y = \sum_{i=1}^{n} y_i \qquad S_{xx} = \sum_{i=1}^{n} x_i^2 \qquad S_{xy} = \sum_{i=1}^{n} x_i y_i$$

Therefore it is possible to produce estimates of parameters in a linear regression model using least-squares, without any specific reference to probability models. In fact, the least-squares approach is very closely related to maximum likelihood estimation for a specific probability model.

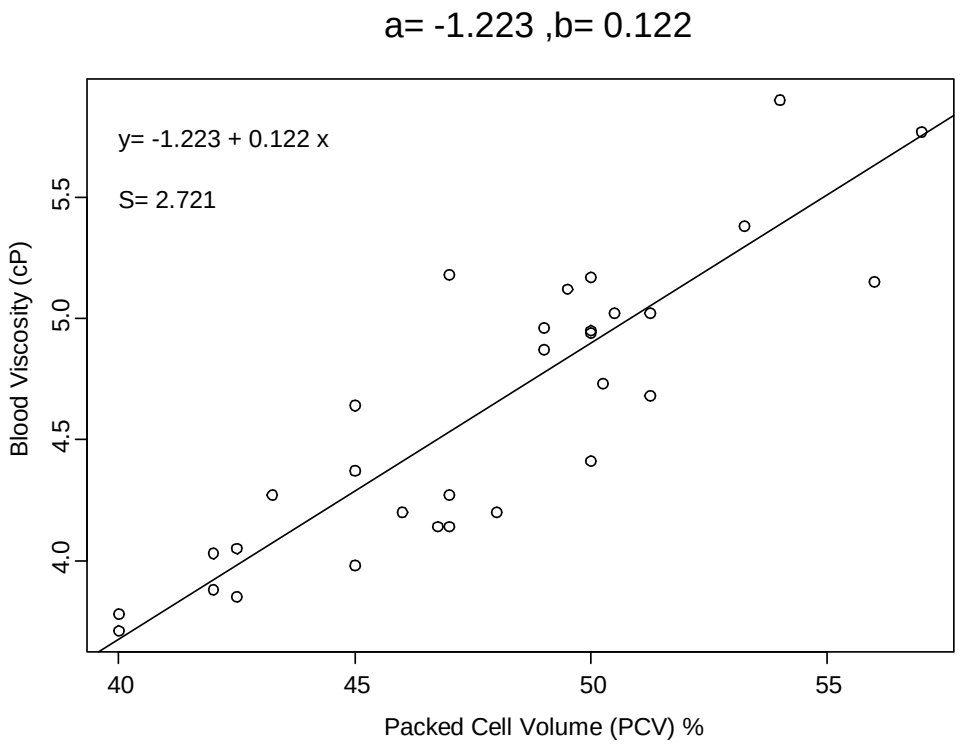
Alternative formulae: let

$$V_{xx} = S_{xx} - \frac{S_x^2}{n} \qquad V_{yy} = S_{yy} - \frac{S_y^2}{n} \qquad V_{xy} = S_{xy} - \frac{S_x S_y}{n}$$

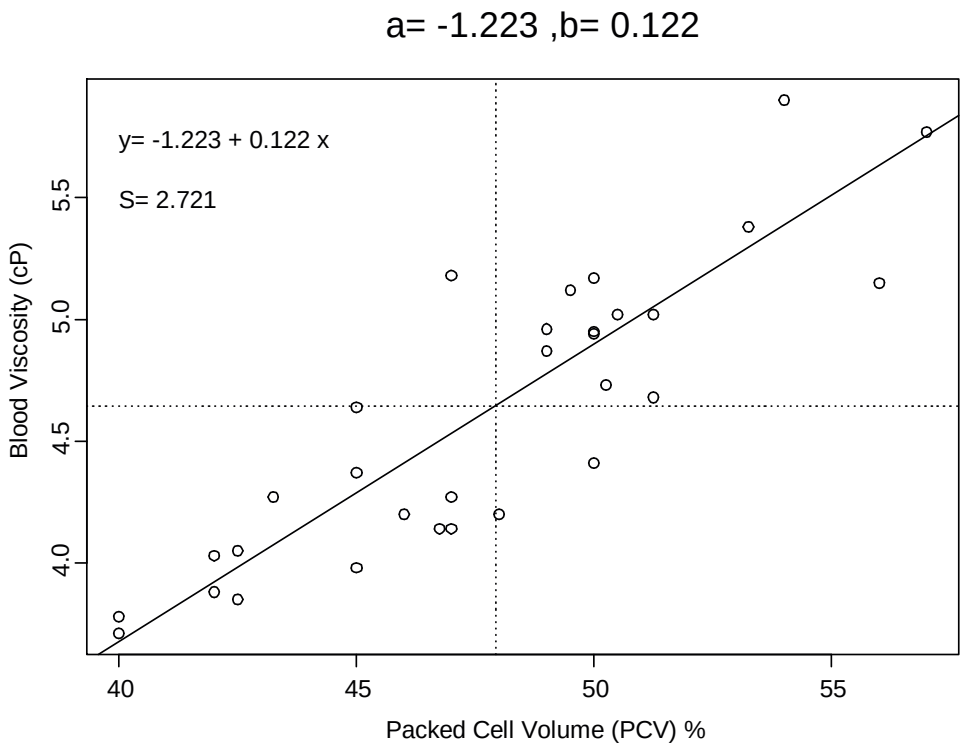
Then

$$\widehat{\beta} = \frac{V_{xy}}{V_{xx}} \qquad \widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x}$$

The line of best fit:



Note: the regression line passes through the mean value point $(\bar{x}, \bar{y})$



The least-squares estimates of α and β assuming a Normal error model are **exactly equivalent** to the maximum likelihood estimates.

1.1.3 ESTIMATES OF ERROR VARIANCE

In addition to the estimates of α and β, we can also obtain the maximum likelihood estimate of σ^2,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\alpha} - \widehat{\beta}x_i)^2 = S^2$$

Often, a **corrected** estimate, s^2, of the error variance is used, defined by

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{\alpha} - \widehat{\beta}x_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where $\hat{y}_i = \widehat{\alpha} + \widehat{\beta}x_i$ is the **fitted value** of Y at $X = x_i$.

1.1.4 RESIDUALS

Having fitted a model with parameters $\widehat{\alpha}$ and $\widehat{\beta}$, we can calculate the error in fit at each data point, or **residual**, denoted $e_i, i = 1, ..., n$, where

$$e_i = y_i - \hat{y}_i = y_i - \widehat{\alpha} - \widehat{\beta} x_i$$

The residuals can be used to assess **model fit**. By the modelling assumptions, if the model is correct, it should be that the residuals are an independent and identically distributed random normal sample, that is

$$\epsilon_i \sim N(0, \sigma^2) \Longrightarrow e_i \text{ should be an observation from } N(0, \sigma^2).$$

This indicates a standardization mechanism

$$\frac{\epsilon_i}{\sigma} \sim N(0, 1)$$

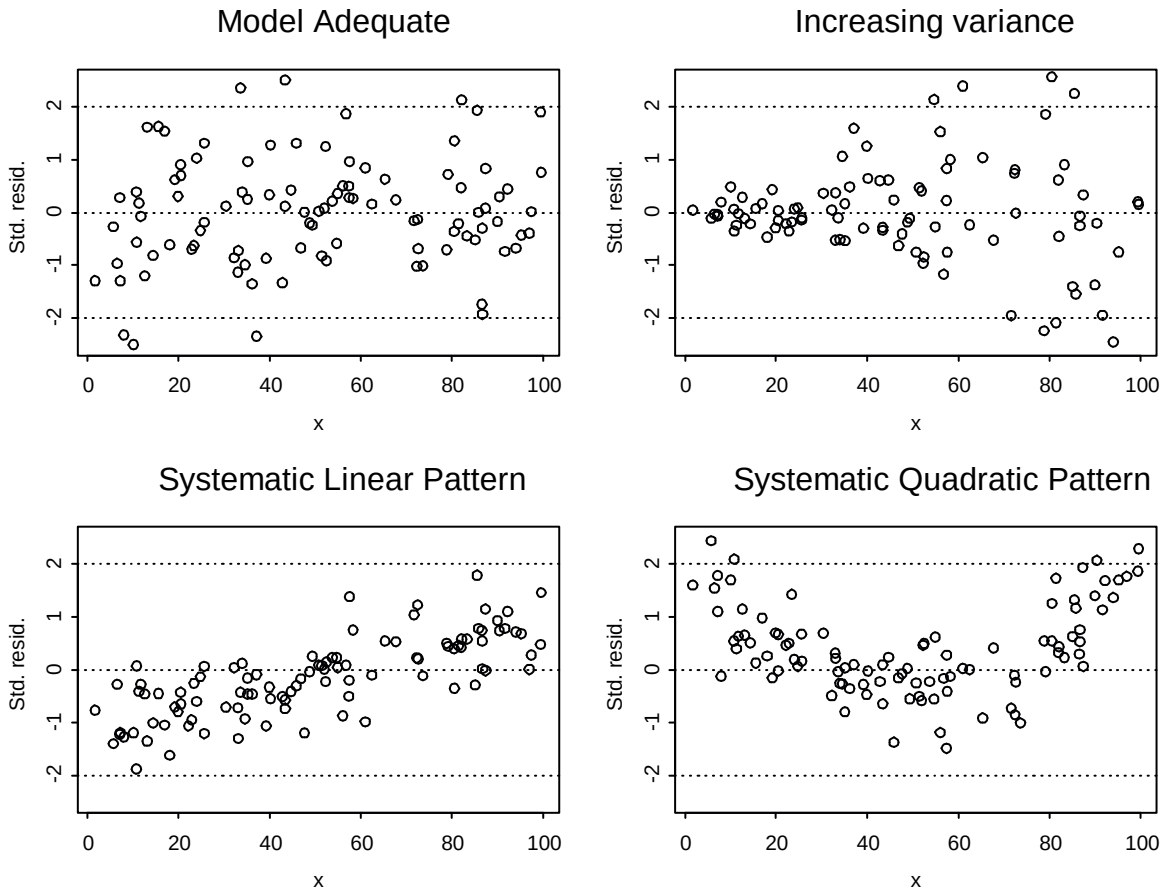
so that instead of inspecting merely residuals we inspect **standardized residual**

$$\widehat{e}_i = \frac{e_i}{s}$$

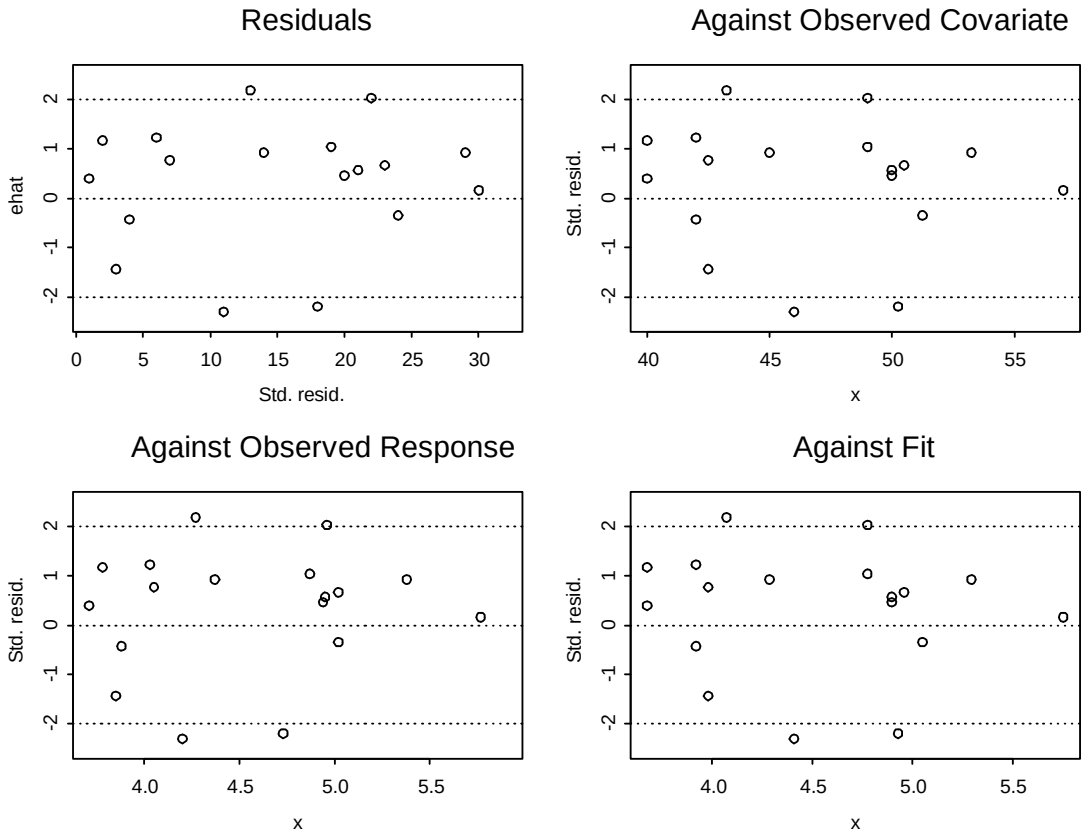
These standardized residuals should

- be internally uncorrelated
- be uncorrelated with any of the response or predictor values
- have a variance approximately 1
- lie within a band ± 2 away from zero

Any deviation from this behaviour indicates that the model is deficient in some way


Model Adequate
Increasing variance
Systematic Linear Pattern
Systematic Quadratic Pattern
Std. resid.
x
-2
-1
0
1
2
0
20
40
60
80
100

FOR BLOOD VISCOSITY DATA



1.1.5 PREDICTION FOR A NEW x VALUE

Suppose that, having fitted a model, and obtained estimates $\widehat{\alpha}$ and $\widehat{\beta}$ using maximum likelihood or least-squares, we want to predict the Y value for a new value x^* of covariate X. By considering the nature of the regression model, we obtain the predicted value y^* as

$$y^* = \widehat{\alpha} + \widehat{\beta} x^*$$

1.1.6 STANDARD ERRORS

We need to be able to understand how the estimators corresponding to $\widehat{\alpha}$ and $\widehat{\beta}$ behave, and by how much the estimate is likely to vary. This can be partially achieved by inspection of the **standard errors** of estimates, that is, the square-root of the variance in the sampling distribution of the

corresponding estimator. It can be shown thay

$$\begin{aligned} s.e.(\widehat{\alpha}) &= s\sqrt{\frac{S_{xx}}{nS_{xx} - \{S_x\}^2}} = s\sqrt{\frac{V_{xx} + \frac{S_x^2}{n}}{nV_{xx}}} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{V_{xx}}} \\ s.e.(\widehat{\beta}) &= s\sqrt{\frac{n}{nS_{xx} - \{S_x\}^2}} = s\sqrt{\frac{1}{V_{xx}}} \end{aligned}$$

where s is the square-root of the corrected estimate of the error variance. It is good statistical practice to report standard errors whenever estimates are reported. The standard error of a parameter also allows a test of the hypothesis "parameter is equal to zero". The test is carried out by calculation of the **t-statistic**, that is, the ratio of a parameter estimate to its standard error. The $t-$statistic must be compared with the 0.025 and 0.975 percentiles of a Student-t distribution with $n-2$ degrees of freedom as described below.

1.1.7 TESTS AND UNCERTAINTY INTERVALS

We may carry out hypothesis tests for the parameters in a linear regression model; as usual we need to be able to understand the sampling distributions of the corresponding estimators. In the linear regression model, the sampling distributions of the estimators of α and β have **Student-t distributions** with $n-2$ degrees of freedom, hence we use the test statistics

$$t_\alpha = \frac{\widehat{\alpha} - c}{s.e.(\widehat{\alpha})} \qquad t_\beta = \frac{\widehat{\beta} - c}{s.e.(\widehat{\beta})}$$

to test the null hypothesis that the parameter is equal to c.

Typically, we use a test at the 5 % significance level, so the appropriate critical values are the 0.025 and 0.975 quantiles of a $St(n-2)$ distribution. It is also useful to report, for each parameter, a confidence interval in which we think the **true** parameter value (that we have estimated by $\widehat{\alpha}$ or $\widehat{\beta}$) lies with high probability.

It can be shown that the 95% **confidence intervals** are given by

$$\alpha : \widehat{\alpha} \pm t_{n-2}(0.975) s.e.(\widehat{\alpha}) \qquad \beta : \widehat{\beta} \pm t_{n-2}(0.975) s.e.(\widehat{\beta})$$

where $t_{n-2}(0.975)$ is the 97.5th percentile of a Student-t distribution with $n-2$ degrees of freedom.

The confidence intervals are useful because they provide an alternative method for carrying out hypothesis tests. For example, if we want to test the hypothesis that $\alpha = c$, say, we simply note whether the 95% confidence interval contains c. If it does, the hypothesis can be accepted; if not the hypothesis should be rejected, as the confidence interval provides evidence that $\alpha \neq c$.

The prediction interval for a new covariate has two forms, depending on whether the predicted **expected** response or the predicted **observed** response is required; the two forms for a prediction at new predictor x^* are

- **EXPECTED value of** y^*

$$\widehat{\alpha} + \widehat{\beta}x^* \pm s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{V_{xx}}}$$

- **OBSERVED value of** y^*

$$\widehat{\alpha} + \widehat{\beta}x^* \pm s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{V_{xx}}}$$

Note that the latter interval is wider, as it takes into account the random observation error.

1.2 CORRELATION

The sample **correlation coefficient**, r, measures the degree of association between X and Y variables and is given by

$$r = \frac{nS_{xy} - S_x S_y}{\sqrt{(nS_{xx} - S_x^2)(nS_{yy} - S_y^2)}} = \frac{V_{xy}}{\sqrt{V_{xx} V_{yy}}}$$

and therefore is quite closely related to $\widehat{\beta}$.

We may carry out a hypothesis test to carry out whether there is significant correlation between two variables. We denote by ρ the true correlation; then, wish to test the hypothesis

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

1.2.1 THE Z-TEST FOR CORRELATION

A test of the hypothesis is given by the **Fisher z statistic**

$$z_r = \frac{\sqrt{n-3}}{2} \log\left(\frac{1+r}{1-r}\right)$$

which has a null distribution that is $N(0,1)$. Hence, if

$$|z_r| > \Phi^{-1}(0.975) = 1.96$$

then we can conclude that the true correlation ρ is significantly different from zero.

1.2.2 THE T-TEST FOR CORRELATION

An alternative test of the hypothesis is based on the test statistic

$$t_r = r\sqrt{\frac{n-2}{1-r^2}}$$

which we compare with the null distribution which is Student-t with $n-2$ degrees of freedom. If

$$|t_r| > t_{n-2}(0.975)$$

then we can conclude that the true correlation ρ is significantly different from zero.

EXAMPLE PCV/Blood Viscosity $r = 0.879$

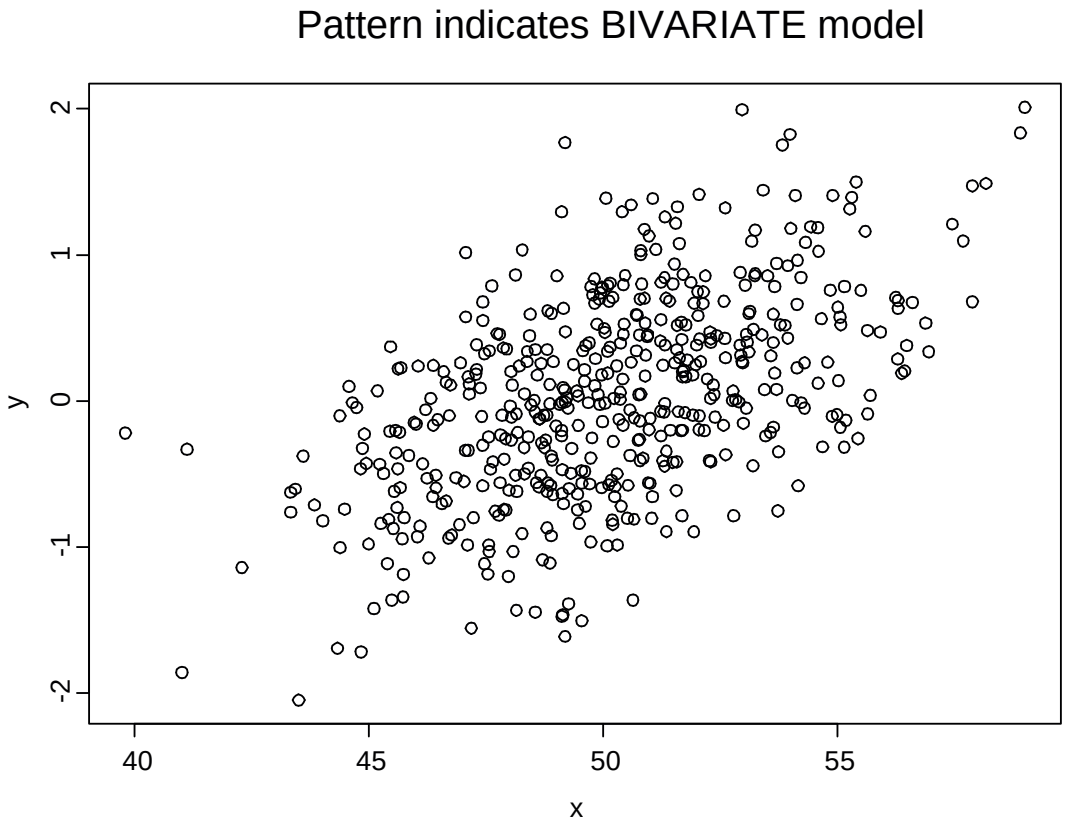
- FISHER Z TEST $z = 7.38065$ $(p = 7.875952e - 014)$
- T-TEST $t = 10.08784, (p = 1.865336e - 011)$

$\therefore$ STRONG EVIDENCE TO REJECT $\rho = 0.0$

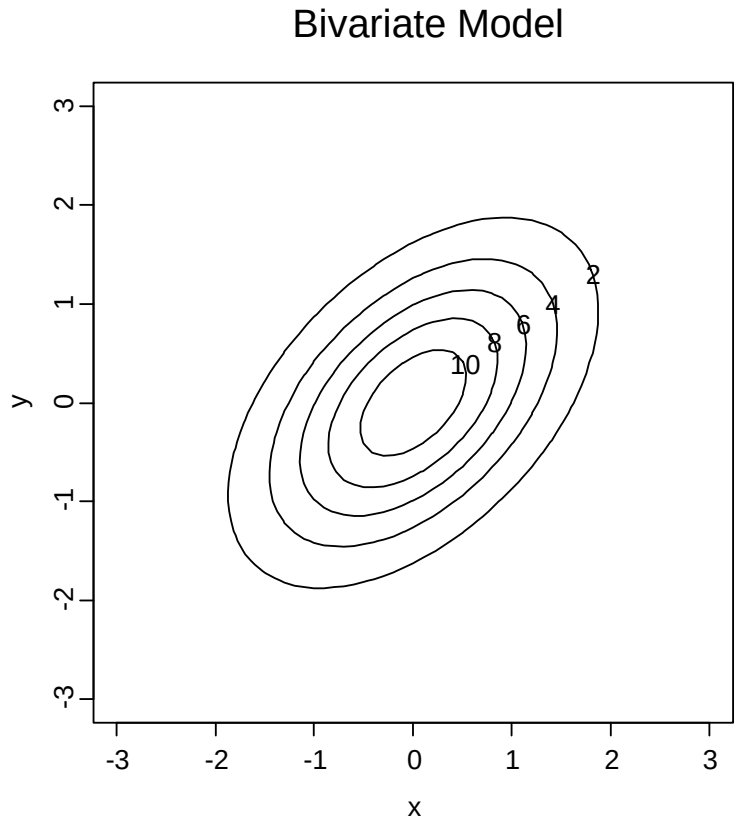
REGRESSION vs MULTIVARIATE MODELLING

Regression

Bivariate Data



Bivariate Contour



1.3 MULTIPLE LINEAR REGRESSION

The simple model above can be extended to the case where Y is modelled as a function of p covariates $X_1, ..., X_p$, that is, we have the conditional expectation of Y given by

$$\mathrm{E}[Y|X_1 = x_1, ..., X_p = x_p] = \alpha + \beta_1 x_1 + ... + \beta_p x_p$$

,so that the observation model is given by

$$Y_i|X_1 = x_{i1}, ..., X_p = x_{ip} \sim N(\alpha + \beta_1 x_{i1} + ... + \beta_p x_{ip}, \sigma^2).$$

Again, we can use maximum likelihood estimation to obtain estimates of the parameters in the model, that is, parameter vector $(\alpha, \beta_1, ..., \beta_p, \sigma^2)$, but the details are slightly more complex, as we have to solve $p+1$ equations simultaneously.

1.4 INFERENCE FOR REGRESSION

In the Normal Linear Model,

$$Y_i = \beta_0 + \beta_1 x_1 + ... + \beta_K x_{iD} + \varepsilon_i = \beta_0 + \sum_{j=1}^{K} \beta_j x_{ij} + \varepsilon_i \qquad (1)$$

where $\varepsilon_i \sim N\left(0, \sigma^2\right)$ for $i = 1, ...n$ are independent and identically distributed random error terms. Note that this implies

$$Y_i | X_i = x_i \sim N\left(\beta_0 + \sum_{j=1}^{K} \beta_j x_{ij}, \sigma^2\right) \therefore E_{f_{Y|X}}\left[Y_i | X_i = x_i\right] = \beta_0 + \sum_{j=1}^{K} \beta_j x_{ij}. \qquad (2)$$

In vector notation, (1) can be re-written $Y_i = x_i^T\beta + \varepsilon_i$, where $x_i = (1, x_{i1}, x_{i2}, ..., x_{iK})^T$, and thus, for vector $Y = (Y_1, ..., Y_n)^T$ we have

$$Y = \mathbf{X}\beta + \varepsilon$$

where $\mathbf{X}$ is a $n \times (K+1)$ matrix called the **design** matrix

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1K} \\ 1 & x_{21} & \cdots & x_{2K} \\ 1 & x_{31} & \cdots & x_{3K} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nK} \end{bmatrix}$$

and to mimic (2), $Y \sim N_n\left(\mathbf{X}\beta, \sigma^2 I_n\right)$, where I_n is the $n \times n$ identity matrix, giving a joint pdf for Y given $\mathbf{X}$ of the form

$$f_{Y|\beta,\sigma^2}(y; \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(y - \mathbf{X}\beta)^T (y - \mathbf{X}\beta)\right\} \qquad (3)$$

1.4.1 THE EXTENDED LINEAR MODEL

The formulation of the linear model above can be extended to allow for more general dependence on the predictors. Suppose that $g_1, g_2, ..., g_K$ are K (potentially non-linear) functions of the K original predictors, that is

$$g_k(x_i) = g_k\left(x_{i1}, ..., x_{iK}\right)$$

is some scalar function, for example, we could have

- $g_k(x_{i1}, ..., x_{iK}) = g_k(x_{i1}) = x_{i1}$ (the identity function)
- $g_k(x_{i1}, ..., x_{iK}) = g_k(x_{i1}) = a_k\sqrt{x_{i1}}$
- $g_k(x_{i1}, ..., x_{iK}) = g_k(x_{i1}) = a_k \log x_{i1}$
- $g_k(x_{i1}, ..., x_{iK}) = g_k(x_{i1}, x_{i2}) = a_k x_{i1} + b_k x_{i2}$

and so on. This reformulation does not effect our probabilistic definition of the model in (3); we can simply redefine design matrix $\mathbf{X}$ as

$$\mathbf{X} = \begin{bmatrix} 1 & g_1(x_1) & \cdots & g_K(x_1) \\ 1 & g_1(x_2) & \cdots & g_K(x_2) \\ 1 & g_1(x_3) & \cdots & g_K(x_3) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & g_1(x_n) & \cdots & g_K(x_n) \end{bmatrix}$$

now an $n \times (K+1)$ matrix. In the discussion below, we will regard the **transformed** variables $(g_1(X), g_2(X), ..., g_K(X))$ as the predictors and drop the dependence on the transformation functions. Hence we have

- Y as a $n \times 1$ column vector
- $\mathbf{X}$ as a $n \times (K+1)$ matrix with ith row $(1, g_1(x_i), ..., g_K(x_i))$
- β as a $(K+1) \times 1$ column vector

1.4.2 MAXIMUM LIKELIHOOD ESTIMATION

Maximum likelihood estimation for the normal linear model is straightforward. Recall that if $\theta = \left(\beta, \sigma^2\right)$ then the mle $\hat{\theta}$ is given by

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} f_{Y|\beta,\sigma^2}(y; \beta, \sigma^2) = \arg\max_{\theta \in \Theta} L\left(\beta, \sigma^2; y, x\right)$$

where parameter space $\Theta \equiv \mathbb{R}^K \times \mathbb{R}^+$. Taking logs in (3) gives

$$\log L\left(\beta, \sigma^2; y, x\right) = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\left(y - \mathbf{X}\beta\right)^T\left(y - \mathbf{X}\beta\right) \quad (4)$$

and considering the maximization for β indicates

$$\arg\max_{\beta \in \mathbb{R}^K} \log L\left(\beta, \sigma^2; y, x\right) = \arg\min_{\beta \in \mathbb{R}^K} \left(y - \mathbf{X}\beta\right)^T\left(y - \mathbf{X}\beta\right)$$

and thus,

$$S\left(\beta\right) = \left(y - \mathbf{X}\beta\right)^T\left(y - \mathbf{X}\beta\right) = y^T y - 2y^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta$$

as $\beta^T\mathbf{X}^T y = y^T\mathbf{X}\beta$.

Using vector/matrix differentiation

$$\frac{d}{d\beta}\left\{y^T\mathbf{X}\beta\right\} = y^T\mathbf{X} \qquad \frac{d}{d\beta}\left\{\beta^T\mathbf{X}^T\mathbf{X}\beta\right\} = 2\mathbf{X}^T\mathbf{X}\beta \tag{5}$$

and so if $\widehat{\beta}$ is the solution of

$$\frac{dS(\beta)}{d\beta} = -y^T\mathbf{X} + \mathbf{X}^T\mathbf{X}\beta = 0$$

then it follows that $\widehat{\beta}$ satisfies

$$\mathbf{X}^T\mathbf{X}\widehat{\beta} = \mathbf{X}^T y. \tag{6}$$

If the matrix $\mathbf{X}^T\mathbf{X}$ is non-singular, then we have the ML estimates of β as

$$\widehat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T y \tag{7}$$

and substituting back into (4) gives

$$\hat{\sigma}^2 = \frac{1}{n}\left(y - \mathbf{X}\widehat{\beta}\right)^T\left(y - \mathbf{X}\widehat{\beta}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{8}$$

where $\hat{y}_i = x_i^T\widehat{\beta}$ is the **fitted value,** and $y_i - \hat{y}_i$ is the residual. Note that $\mathbf{X}^T\mathbf{X}$ is a symmetric matrix. The expression $\left(y - \mathbf{X}\widehat{\beta}\right)^T\left(y - \mathbf{X}\widehat{\beta}\right)$ is termed the **residual sum of squares** (or **RSS**). A common **adjusted** estimate is

$$\hat{\sigma}^2_{ADJ} = \frac{1}{n-K-1}\left(y - \mathbf{X}\widehat{\beta}\right)^T\left(y - \mathbf{X}\widehat{\beta}\right) \tag{9}$$

the justification for this result depends on the sampling distribution of the estimator. It can be shown that $\hat{\sigma}^2_{ADJ}$ is unbiased for σ^2.

If $K = 1$, with identity function $g(t) = t$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

and thus

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \frac{1}{\hat{\sigma}^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix}$$

So

$$\widehat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T y$$

$$= \frac{1}{\hat{\sigma}^2}\begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix}\begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

$$= \frac{1}{\hat{\sigma}^2}\begin{bmatrix} \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i \\ n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \end{bmatrix}$$

1.4.3 PROPERTIES OF THE ML ESTIMATOR

By elementary properties of random variables, the properties of ML estimator $T = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T Y$

$$\begin{aligned} E_{Y|X,\beta,\sigma^2}[T] &= E_{Y|X,\beta,\sigma^2}\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T Y\right] = \left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right) E_{Y|X,\beta,\sigma^2}[Y] \\ &= \left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)\mathbf{X}\beta = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\beta = \beta \end{aligned}$$

so that T is unbiased for β, and

$$\begin{aligned} Var_{Y|X,\beta,\sigma^2}[T] &= Var_{Y|X,\beta,\sigma^2}\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T Y\right] \\ &= \left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right) Var_{Y|X,\beta,\sigma^2}[Y] \left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)^T \\ &= \left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)\sigma^2 I_n \left(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right) \\ &= \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}. \end{aligned}$$

Note that, in fact, given β and σ^2

$$Y \sim N_n\left(\mathbf{X}\beta, \sigma^2 I_n\right) \Longrightarrow T = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T Y \sim N_{K+1}\left(\beta, \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right). \tag{10}$$

It also follows that

$$\left(y - \mathbf{X}\beta\right)^T\left(y - \mathbf{X}\beta\right) = \left(y - \mathbf{X}\widehat{\beta}\right)^T\left(y - \mathbf{X}\widehat{\beta}\right) + \left(\widehat{\beta} - \beta\right)^T\left(\mathbf{X}^T\mathbf{X}\right)\left(\widehat{\beta} - \beta\right)$$

or

$$S\left(\beta\right) = S\left(\widehat{\beta}\right) + \left(\widehat{\beta} - \beta\right)^T\left(\mathbf{X}^T\mathbf{X}\right)\left(\widehat{\beta} - \beta\right)$$

where

$$S(\beta) = (y - \mathbf{X}\beta)^T (y - \mathbf{X}\beta) \quad \textbf{(1)}$$

$$S\left(\widehat{\beta}\right) = \left(y - \mathbf{X}\widehat{\beta}\right)^T \left(y - \mathbf{X}\widehat{\beta}\right) = (y - \hat{y})^T (y - \hat{y}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \textbf{(2)}$$

$$\left(\widehat{\beta} - \beta\right)^T (\mathbf{X}^T\mathbf{X}) \left(\widehat{\beta} - \beta\right) = \left(\mathbf{X}\beta - \mathbf{X}\widehat{\beta}\right)^T \left(\mathbf{X}\beta - \mathbf{X}\widehat{\beta}\right) \quad \textbf{(3)}$$

are the **(1) TOTAL , (2) RESIDUAL** and **(3) FITTED** sum of squares (**TSS, RSS** and **FSS).**

Therefore, by normal distribution theory, it follows that

$$\frac{S(\beta)}{\sigma^2} \sim \chi^2_n \qquad \frac{S\left(\widehat{\beta}\right)}{\sigma^2} \sim \chi^2_{n-K-1}$$

so that

$$s^2 = \frac{S\left(\widehat{\beta}\right)}{(n-K-1)} \text{ is an } \mathbf{UNBIASED} \text{ estimator of } \sigma^2$$

and the quantity

$$\frac{\widehat{\beta}-\beta}{s.e.\left(\widehat{\beta}\right)} = \frac{\widehat{\beta}-\beta}{s\sqrt{v_{ii}}} \sim Student\,(n-K-1)\,.$$

It also follows that

$$\frac{S(\beta) - S\left(\widehat{\beta}\right)}{\sigma^2} = \frac{\left(\widehat{\beta} - \beta\right)^T \left(\mathbf{X}^T\mathbf{X}\right)\left(\widehat{\beta} - \beta\right)}{\sigma^2} \sim \chi^2_{K+1}$$

so that finally

$$\frac{\left[S(\beta) - S\left(\widehat{\beta}\right)\right]/(K+1)}{S\left(\widehat{\beta}\right)/(n-K-1)} \sim Fisher\,(K+1, n-K-1)$$

It follows that in this case the ML estimator is the Minimum Variance Unbiased Estimator (MVUE) and the Best Linear Unbiased Estimator (BLUE).

1.5 ANOVA IN REGRESSION

Analysis of variance or **ANOVA** is used to display the sources of variability in a collection of data samples. The ANOVA F-test compares variability **between** samples with the variability **within** samples. In the above analysis, we have that

$$S(\beta) = S\left(\widehat{\beta}\right) + \left(\widehat{\beta} - \beta\right)^T \left(\mathbf{X}^T\mathbf{X}\right)\left(\widehat{\beta} - \beta\right)$$

or, as previously

$$TSS = RSS + FSS.$$

Now, using the distributional results above, we can construct the following **ANOVA Table** to test the hypothesis H_0 :All $\beta_k = 0$ against the general alternative that H_0 is not true.

Source	D.F.	Sum of sq.	Mean square	F
FITTED	K	FSS	$M_{FSS} = \frac{FSS}{K}$	$\frac{M_{FSS}}{M_{RSS}}$
RESIDUAL	$n-K-1$	RSS	$M_{RSS} = \frac{RSS}{(n-K-1)}$	
TOTAL	$n-1$	TSS		

This test allows a comparison of the fits of the two competing models implied by the null and alternative hypotheses.

- Under the null model, if H_0 is true, then the model has $Y_i \sim N\left(\beta_0, \sigma_0^2\right)$ for $i = 1, 2, ...n$, for some β_0 and σ_0^2 to be estimated.
- Under the alternative hypothesis, there are a total of $K+1$ β parameters to be estimated.

The **degrees of freedom** column headed (D.F.) details how many parameters are used to describe the amount of variation in the corresponding row of the table; for example, for the FIT row, D.F. equals K as there are K parameters used to extend the null model to the alternative model.

Now consider the following design; suppose that there are K possible medical treatments and you wish to test for any difference between them.

The parameter vector is $\beta = [\beta_1, \beta_2, ..., \beta_K]^T$ say, and the null hypothesis is that, for some β,

$$H_0 : \beta_1 = \beta_2 = ... = \beta_K = \beta$$

The Normal Linear Model Theory above applies, and the ANOVA test proceeds as usual.

1.6 MIXED LINEAR MODELS

The equation for response Y in terms of covariates X

$$Y = \mathbf{X}\beta + \varepsilon$$

so that

$$Y_i = x_i^T \beta + \varepsilon_i$$

indicates that the variation in Y_i is the result of a systematic component $x_i^T \beta$ plus some random variation ε. The parameters β are termed **fixed effects** parameters. An extension of this model adds a further, individual random component

$$Y_i = x_i^T \beta + Z_i + \varepsilon_i$$

where $Z_i \sim N(0, \sigma_Z^2)$ is a **random** individual specific-random variable. If multiple observations are available,

$$Y_{ij} = x_{ij}^T \beta + Z_i + \varepsilon_{ij}$$

A model that includes both fixed and random effects terms is called a **mixed effects model**.

The $\{Z_i\}$ terms are identically distributed, with one Z_i specific to each individual's observations.

It is possible to **marginalize** this model by integrating out over the unobsrerved Z.

Standard likelihood theory does not extend to this case

1.7 NON-LINEAR REGRESSION

The linear model

$$Y_i = x_i^T \beta + \varepsilon_i$$

is termed linear because the terms in the vector β appear in a linear combination. It can be exteneded to the **non-linear** case, for example

$$Y_i = g\left(x_i^T \beta\right) + \varepsilon_i$$

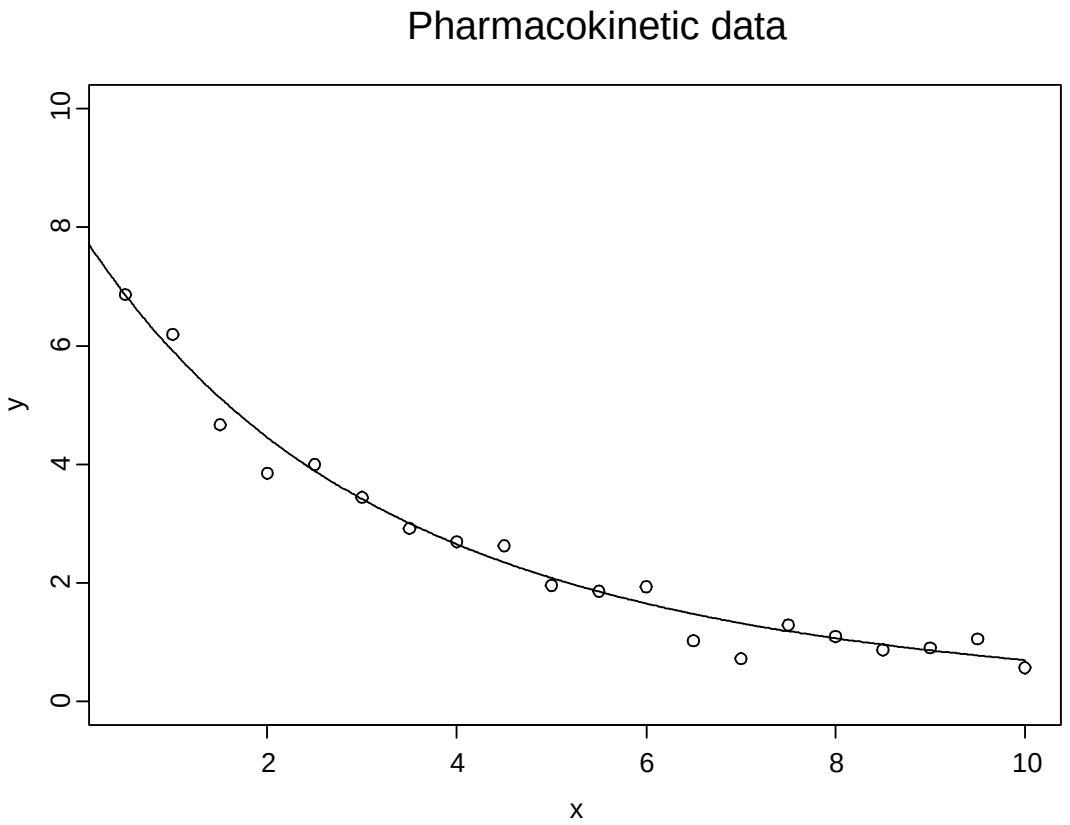
for some non-linear function g of the parameters.

Likelihood & Least Squares estimation still available.

EXAMPLE: Pharmacokinetics

$$Y_i = g\left(x_i^T \beta\right) + \varepsilon_i = \beta_{00} \exp\left\{-\beta_{01} x_i\right\} + \beta_{10} \exp\left\{-\beta_{11} x_i\right\} + \varepsilon_i$$

where $\beta_{01} < \beta_{11}$ for identifiablity.


Pharmacokinetic data
y
x
0
2
4
6
8
10
2
4
6
8
10

SECTION 2.

GENERALIZED LINEAR MODELS

The central idea of **Generalized Linear Models** (GLMs) is to extend the ideas from the normal linear model to allow the possibility of modelling non-normal data. In the GLM, we will model

$$E_{f_{Y|X}}\left[Y_i|X_i = x_i\right] = g^{-1}\left(x_i^T\beta\right)$$

where

$$x_i^T\beta = \beta_0 + \sum_{j=1}^{K}\beta_j x_{ij}.$$

for some monotonic/invertible function g; in the normal linear model, g is the **identity** function.

2.1 GLM TERMINOLOGY

There are two key terms in the model description:

- **Linear predictor**: for observed predictor $x_i = (x_{i1}, ..., x_{iK})$ and parameters $\beta = (\beta_0, \beta_1, ...\beta_K)$, the **linear predictor** is

$$\eta_i = x_i^T \beta = \beta_0 + \sum_{j=1}^{K} \beta_j x_{ij}$$

- **Link function:** a **link function** g is a function that connects the linear predictor to the expected value of the response

$$g\left(E_{f_{Y|X}}\left[Y_i | X_i = x_i\right]\right) = x_i^T \beta.$$

EXAMPLES:

- **LOGISTIC REGRESSION**: data are 0/1 (FAILURE/SUCCESS) variables, where

$$P[Y_i = 1] = \theta_i \qquad 0 < \theta_i < 1$$

(that is, the probability is "personalized" for individual i, and

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = x_i^T\beta = \beta_0 + \beta_1 x_i \Longleftrightarrow \theta_i = \frac{\exp\{\beta_0 + \beta_1 x_i\}}{1+\exp\{\beta_0 + \beta_1 x_i\}}$$

- **PROBIT REGRESSION**: data are 0/1 as above, but

$$\Phi^{-1}(\theta_i) = \beta_0 + \beta_1 x_i \Longleftrightarrow \theta_i = \Phi(\beta_0 + \beta_1 x_i)$$

where Φ is the standard normal distribution function.

- **POISSON REGRESSION :** data are counts $\{0, 1, 2, 3, ...\}$ modelled using a Poisson model

$$P[Y_i = y] = \frac{e^{-\lambda_i}\lambda_i^y}{y!} \qquad \lambda_i > 0$$

with

$$\log \lambda_i = \beta_0 + \beta_1 x_i$$

Poisson Regression is a special case of a class of regression models called **log-linear models** that are used in the analysis of contingency tables. In a two-way table, with observed cell entries n_{ij} it is common to use the model

$$P[N_{ij} = n_{ij}] = \frac{e^{-\lambda_{ij}}\lambda_{ij}^{n_{ij}}}{n_{ij}!} \qquad \log \lambda_{ij} = \beta_{ij}$$

where β_{ij} is further decomposed as required.

2.2 LINK FUNCTIONS

Link functions that are in common usage for the usual statistical models include:

- **log**
- **logistic**
- **power**
- **Box-Cox**
- **Probit**
- **Complementary log-log**
- **Log-log**

2.3 CHECKING THE FIT OF A GLM

Deviance is a way of measuring the goodness of fit of a GLM. It is based on a **Likelihood Ratio** statistic

$$D = 2\log\frac{l_S\left(\widehat{\beta}_S\right)}{l_M\left(\widehat{\beta}_M\right)} = -2\log\frac{l_M\left(\widehat{\beta}_M\right)}{l_S\left(\widehat{\beta}_S\right)}$$

where

- $\widehat{\beta}_M$ is the mle under a model, M
- $\widehat{\beta}_S$ is the mle baseline model the **saturated** model, which corresponds to the **best possible fit**
- l_M and l_S are the **likelihood** functions under the model and saturated model respectively.

We have a complete range of model fits to calibrate the fit of any individual model:

$$\text{SATURATED MODEL} \quad \rightarrow \quad \text{MODEL} \quad \rightarrow \quad \text{NULL MODEL}$$

$$\text{MOST COMPLEX} \quad \rightarrow \quad \text{LEAST COMPLEX}$$

$$\text{LOWEST DEVIANCE} \quad \rightarrow \quad \text{HIGHEST DEVIANCE}$$

Changes in deviance are assessed for significance against a **chi-squared** distribution, or an F distribution.

SECTION 3.

MULTIVARIATE MODELLING

A typical experimental design involves taking repeated measurements of the **same** feature on the **same** experimental unit.

- observations made at time $t = 0, t_1, t_2, ...$
- observations made on treatment 1,2,3,...

The resulting observed values are necessarily **correlated** through "time" in this design because of the "within individual" factors. However, we may still be interested in the impact of different **predictors/covariates/fixed effects.**

3.1 MODELLING ASSUMPTIONS

We want to assess the systematic differences in **mean** response level (between time points, between treatment groups etc.) whilst accounting for the correlation in the observed data. Let y_{ijk} be the

- k^{th} observation for the
- j^{th} experimental unit in the
- i^{th} group

for $i = 1, ..., n$, $j = 1, ..., n_i$ and $k = 1, ..., K_{ij}$. Then

$$E\left[Y_{ijk}\right] = \mu_{ijk}$$

and we attempt to model μ_{ijk}, but $\mathbf{Y}_{ij} = \left(Y_{ij1}, ...Y_{ijK_{ij}}\right)$ have some **multivariate probability distribution** due to the within individual correlation.

A common (and typical) assumption is to use a multivariate Normal distribution,

$$\mathbf{Y}_{ij} \sim N_{K_{ij}}\left(\boldsymbol{\mu}_{ij}, \Sigma_{ij}\right)$$

where

- $\boldsymbol{\mu}_{ij} = \left(\mu_{ij1}, \mu_{ij2}, ..., \mu_{ijK_{ij}}\right)^T$ is a $K_{ij} \times 1$ mean vector
- Σ_{ij} is a $K_{ij} \times K_{ij}$ variance-covariance matrix

for the $(i,j)^{th}$ unit.

If μ_{ijk} does not depend on time k, then the model simplifies to

$$\boldsymbol{\mu}_{ij} = \left(\mu_{ij}, \mu_{ij}, ..., \mu_{ij}\right)^T$$

and our interest centres on differences between the collection of μ_{ij}s.

We have encountered this situation previously when studying **ANOVA;** the distinction here is the correlation in the data.

- Classical ANOVA: **"BETWEEN SUBJECTS" -** data conditionally independent given group classification
- Here: **"WITHIN SUBJECTS" -** data correlated

The within-subjects design introduces **nuisance** parameters such as the parameters in Σ_{ij}; some simplifications, such as

$$\Sigma_{ij} = \Sigma_i \qquad \text{common covariance for group } i$$

or

$$\Sigma_{ij} = \Sigma \qquad \text{common covariance for all groups}$$

can be made (and tested).

3.2 WITHIN SUBJECTS ANOVA

The within subjects or **repeated measures** approach is used for several reasons:

- **Clinical:** some research hypotheses require repeated measures. Longitudinal research, for example, measures each sample member at each of several ages. In this case, age would be a repeated factor
- **Statistical:** in cases where there is a great deal of variation between sample members, error variance estimates from standard ANOVAs are large. Repeated measures of each sample member provides a way of accounting for this variance, thus reducing error variance.
- **Economic:** when sample members are difficult to recruit, repeated measures design are economical because each member is measured under all conditions.

Alternative Design:

Repeated measures ANOVA can also be used when sample members have been matched according to some important characteristic.

- matched sets of sample members are generated
- each set having the same number of members
- each member of a set being exposed to a different random level of a factor or set of factors.

When sample members are matched, measurements across conditions are treated like repeated measures in a repeated measures ANOVA.

EXAMPLE: Suppose that a group of depressed subjects is selected, and their levels of depression measured. Suppose then that these subjects are arranged into pairs having similar depression levels; one subject from each matching pair is then given a treatment for depression, and afterwards the level of depression of the entire sample is measured again.

ANOVA comparisons between the two groups for this final measure would be most efficient using a repeated measures ANOVA. In this case, each matched pair would be treated as a single sample member.

As with any ANOVA, within subjects, or **repeated measures** ANOVA tests the equality of means.

NOTE: We should be clear about the difference between a **repeated measures** design and a **multivariate** design .

- In the **repeated measures** design, each trial represents the measurement of the **same characteristic under a different condition**.
- In the **multivariate design**, each trial represents the measurement of a **different characteristic**. It is generally inappropriate to test for simple mean differences between measurements of different characteristics.

3.3 WORKED EXAMPLE

(taken from `http://www.utexas.edu/cc/docs/stat38.html`)

A health researcher wants to investigate the impact of dietary habits and types of exercise on individuals' **pulse rates** over time.

- To investigate these issues, a sample of individuals is collected, and grouped according to their **dietary preferences** D: (either meat eater or vegetarians)
- Each diet category is then split into three groups, with each group assigned one of three types of **exercise** E: (aerobic stair climbing, squash, and weight training).

Thus, this design has two **between-subjects** grouping factors: **dietary preference** and **exercise type**.

In addition to these between-subjects factors, a single **within-subjects** factor is to be included.

- Each subject's pulse rate will be measured at **three** exercise intervals (immediately after warm-up exercises, after jogging, and after running). Thus, **intensity of exertion** I is the within-subjects factor in this design. The order of these three measurements will be randomly assigned for each subject.

NOTE: all the factors described can be considered **fixed effects** (and not **random effects)**. The trials and groups were selected because of the research hypothesis. The levels of a random effect are chosen at random from a population of possible levels; random effects can not be appropriately analyzed with the method being described.

ID	DIET	EX.	WEIGHT	PULSE 1	PULSE 2	PULSE 3
1	Meat Eater	Stair	75.0	86	115	119
2	Meat Eater	Stair	85.6	72	117	129
3	Meat Eater	Stair	61.2	78	106	132
4	Meat Eater	Stair	72.2	68	108	129
⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	Meat Eater	Stair	62.8	87	103	125
11	Meat Eater	Squash	74.7	90	127	128
12	Meat Eater	Squash	79.0	75	123	141
⋮	⋮	⋮	⋮	⋮	⋮	⋮
30	Meat Eater	Weights	70.1	61	185	204
31	Vegetarian	Stair	80.6	86	121	125
32	Vegetarian	Stair	79.7	68	105	143
⋮	⋮	⋮	⋮	⋮	⋮	⋮

There are four questions to be addressed:

- **Within-Subjects Main Effect**
  - Does exertion intensity influence pulse rate? (Does mean pulse rate change across the trials for exertion intensity?) This is the test for a within-subjects main effect of intensity.
- **Between-Subjects Main Effects**
  - Does dietary preference influence pulse rate? (Do vegetarians have different mean pulse rates than meat eaters?) This is the test for a between-subjects main effect of dietary preference.
  - Does exercise type influence pulse rate? (Are there differences in mean pulse rates between stair climbers, squash players, and weight trainers?) This is the test for a between-subjects main effect of exercise type.

- **Between-Subjects Interaction Effect**
  - Does the influence of exercise type on pulse rate depend on dietary preference? (Does the pattern of differences between mean pulse rates for exercise-type groups change for each dietary-preference group?) This is the test for a between-subjects interaction of exercise type by dietary preference. Note that other formulations of this interaction are equivalent. This hypothesis can also be expressed as "Does the influence of dietary preference depend on exercise type?".

The **interaction hypotheses** can be interpreted as follows; we may wish to test that vegetarian squash players have lower pulse rates than all meat eaters and other vegetarians; is something unique in the combination of a vegetarian diet and squash exercise that produces an unusually low mean pulse rate.

- **Within-Subjects by Between-Subjects Interaction Effects**
  - Does the influence of diet on pulse rate depend upon exertion intensity? (Does the pattern of differences between mean pulse rates for dietary-preference groups change at each intensity trial?) This is the test for a between-subjects by within-subjects interaction of dietary preference by exertion intensity. You might suspect, for example, that the mean pulse rate of meat eaters will increase more than the mean pulse rate of vegetarians as the intensity of exercise changes.
  - Does the influence of exercise type on pulse rate depend upon exertion intensity? (Does the pattern of differences between mean pulse rates for exercise-type groups change at each intensity trial?) This is the test for a between-subjects by within-subjects interaction of exercise type by exertion intensity.

– Does the influence of dietary preference on pulse rate depend upon exercise type and exertion intensity? (Does the pattern of differences between mean pulse rates for dietary-preference groups change for some exercise-type group and for some intensity trial?) This is the test for a between-subjects by within-subjects interaction of dietary preference by exercise type by exertion intensity.

Each of these hypotheses relates to a different aspect of the model specification. Many of the statistical tests that are to be used are extensions of the ones used previously based on ANOVA considerations (in tests of equality of mean response), tests for equality of variance (Levene's Test etc.). The novelty here relates to testing hypotheses concerned with the correlation structure.

3.4 ASSESSING COVARIANCE

Several methods are used to investigate the covariance structure in a repeated measures (or general multivariate) experiment.

3.4.1 BOX'S TEST OF EQUALITY OF COVARIANCE

Box's M-test tests for the equality of covariance matrices across G multivariate subgroups (defined by the fixed effects cross-categorization), that is

$$\begin{aligned} H_0 &: \Sigma_g = \Sigma \\ H_1 &: \Sigma_{g_1} \neq \Sigma_{g_2} \qquad \text{for some pair } (g_1, g_2) \text{ of sub-populations} \end{aligned}$$

against a general alternative.

The test statistic is M defined by

$$M = (n - G) \log |\mathbf{S}| - \sum_{g=1}^{G} (n_g - 1) \log |\mathbf{S}_g|$$

where

$$\mathbf{S} = \frac{\sum_{g=1}^{G} (n_g - 1)\, \mathbf{S}_g}{n - G} \qquad \mathbf{S}_g = \frac{\sum_{i=1}^{n_g} \left(\mathbf{y}_{gi} - \overline{\mathbf{y}}_g\right) \left(\mathbf{y}_{gi} - \overline{\mathbf{y}}_g\right)^T}{n_g - 1}$$

and where $\overline{\mathbf{y}}_g$ is the (vector) sample mean, $\mathbf{S}_g$ is the sample covariance matrix for subgroup g, and $\mathbf{S}$ is the pooled sample covariance matrix, and

$$n = n_1 + ... + n_g$$

is the total sample size. The null distribution for this test is the $Fisher - F$ distribution.

3.4.2 MULTIVARIATE TESTS

Multivariate tests are joint tests of the **significance of a main effects and within subjects effects**; the tests reported in SPSS are

- Pillai's Trace
- Wilks's Lambda
- Hotelling's Trace
- Roy's Largest Root

The details of these tests can be largely ignored; the important thing to note is that they follow the usual procedure of statistical hypothesis testing; a test statistic is derived (usually based on some transformation or eigen-representation of the sample covariance matrices) and the suprisingness of the observed test statistic is assessed against some null distribution. SPSS reports the appropriate $p-$values.

Which test is preferable ?

- Schatzoff (1966)
  - Roy's largest-latent root was the most sensitive when populations differ along a single dimension, but was otherwise least sensitive.
  - Under most conditions it was a toss-up between Wilks' and Hotelling's criteria.
- Olson (1976)
  - Pillai's criteria was the most robust to violations of assumptions concerning homogeneity of the covariance matrix.
  - Under diffuse **noncentrality** the ordering was Pillai, Wilks, Hotelling and Roy.

  - Under concentrated noncentrality the ordering is Roy, Hotelling, Wilks and Pillai.

- So which is best ?

  - When sample sizes are very large the Wilks, Hotelling and Pillai become asymptotically equivalent.

(http://www.gseis.ucla.edu/courses/ed231a1/notes3/manova.html)

3.4.3 MAUCHLY'S TEST OF SPHERICITY

The correlations between the different measurement times (not only successive but rather any time) are not usually the same, which prevents the use of the usual (Fisher F) test calculated as for ANOVA. The normal F test assumes the **sphericity** of the data, which means that variance of **all mutual differences of all possible pairs** of measuring times is the **same.**

This test is equivalent to an assumption that the $(i, j)^{th}$ element of the common covariance matrix for the random errors is some constant value σ

$$[\Sigma]_{ij} = [\Sigma]_{ji} = \sigma \qquad i \neq j$$

or that the correlation is some constant ρ.

Lack of sphericity causes concern about the ANOVA F-test; this is tested with **Mauchly's test**, whose null hypothesis is sphericity.. The rejection of sphericity does not prevent the analysis of variance, but the degrees of freedom should be adjusted before the ANOVA F-test result is reported.

A correction should be applied to the within-subject (time) effects and their corresponding error. Mean-Square values change, but the value of the F test does not, however the degrees of freedom values used to calculate the $p-$value do change, and can make a large difference to the conclusions made.

If the number of measurement times is K, then the coefficient for correction of the degrees of freedom, ϵ, can take values between $1/(K-1)$ and 1, where 1 corresponds to the complete sphericity situation.

THE GREENHOUSE-GEISSER AND HUYNH-FELDT CORRECTIONS

- If the assumption of sphericity is strongly rejected, i.e. the coefficient of correction is near the lower limit $1/(K-1)$, we can use the **Greenhouse-Geisser** correction (rough limit $\epsilon < 0.75$).

- If the assumption of sphericity is broken just a little, i.e. ϵ is near one ($\epsilon > 0.75$), an adequate option is the **Huynh-Feldt** correction which is more liberal than the Greenhouse-Geisser correction (and hence more sensitive to differences).

3.5 TESTS OF WITHIN-SUBJECTS EFFECTS

Univariate tests for the fixed effects (and contrasts) can also be carried out; this respects the true repeated measures aspect of the design, as opposed to the **multivariate tests** described above. **Repeated measures** ANOVA is used.

Repeated measures ANOVA carries the standard set of assumptions associated with an ordinary analysis of variance; extended to the matrix case: multivariate normality, homogeneity of covariance matrices, and independence. Repeated measures ANOVA is robust to violations of the first two assumptions.

- Violations of the independence assumption produce a non-normal distribution of the residuals, which results in invalid F ratios.
- The most common violations of independence occur when either random selection or random assignment is not used.

In addition to these assumptions, the univariate approach to tests of the within-subject effects requires the assumption of sphericity; if the sphericity assumption is not valid, conservative correction methods (such as Greenhouse-Geisser or Huynh-Feldt) should be utilized.

When sample sizes are small, the univariate approach can be more powerful, but this is true only when the assumption of a common spherical covariance matrix has been met.

Finally, a test of **homogeneity** of variances based on Levene's procedure should also be carried out.

3.6 MULTIVARIATE RESPONSES

A **multivariate response** experiment has much of the same structure as the repeated measures experiment that is described in the previous chapter. The principal extensions are that

- a number different variables can be measured for each experimental unit, possibly at different time points.
- simplifying assumptions necessary for repeated measures ANOVA (homogeneity, sphericity) can be relaxed; this may result in a less powerful analysis, but this is unavoidable if the simplyfing assumptions are not valid.

SECTION 4.

CLASSIFICATION AND DISCRIMINATION

Classification is another special type of regression modelling that explains the observed variability in a **response** variable Y via consideration of **predictors** $X = (X_1, ..., X_K)$. The principal difference between classification and conventional regression is that the response variable is a **nominal categorical variable**, that is, for data item i

$$Y_i \in \{1, 1, 2, ...K\}$$

so that the value of Y_i is a **label** rather than a numerical value, where the label represents the **group** or **class** to which that item belongs.

We again wish to use the **predictor** information in X to allocate Y to one of the classes

Thus, there are two main goals:

- to partition the observations into two or more labelled classes. The emphasis is on **deriving a rule** that can be used to **optimally assign** a new object to the labeled classes.
  - This is the process of **CLASSIFICATION**
- to describe either graphically or algebraically, the different features of observations from several known collections. We attempt to find **discriminants** whose numerical values are such that the collections are separated as much as possible.
  - This is the process of **DISCRIMINATION**

Both are special cases of what we have previously termed **MULTIVARIATE ANALYSIS**

Typically, the exercise of classification will be **predictive**, that is,

- we have a set of data available where both the **response** and **predictor** information is known
  - these data are the **training** data
- we also have a set of data where only the **predictor** information is known, and the **response** is to be predicted
  - these data are the **test** data
- often we will carry out an exercise of **model-building** and **model-testing** on a given data set by extracting a **training set**, building a model using the training data, whilst holding back a proportion (the **test set**) for model-testing.

4.1 CLASSIFICATION FOR TWO CLASSES ($K = 2$)

Let $f_1(x)$ and $f_2(x)$ be the probability functions associated with a (vector) random variable X for two populations 1 and 2. An object with measurements x must be assigned to either class 1 or class 2. Let $\mathbb{X}$ denote the sample space. Let $\mathcal{R}_1$ be that set of x values for which we classify objects into class 1 and $\mathcal{R}_2 \equiv \mathbb{X} \backslash \mathcal{R}_1$ be the remaining x values, for which we classify objects into class 2.

The **conditional probability**, $P(2|1)$, of classifying an object into class 2 when, in fact, it is from class1 is:

$$P(2|1) = \int_{\mathcal{R}_2} f_1(x) \; dx.$$

Similarly, the conditional probability, $P(1|2)$, of classifying an object into class 1 when, in fact, it is from class 2 is:

$$P(1|2) = \int_{\mathcal{R}_1} f_2(x)\ dx$$

Let p_1 be the *prior* probability of being in class 1 and p_2 be the *prior* probability of 2, where $p_1 + p_2 = 1$. Then,

$$\begin{aligned} P(\text{Object correctly classified as class 1}) &= P(1|1)p_1 \\ P(\text{Object misclassified as class 1}) &= P(1|2)p_2 \\ P(\text{Object correctly classified as class 2}) &= P(2|2)p_2 \\ P(\text{Object misclassified as class 2}) &= P(2|1)p_1 \end{aligned}$$

Now suppose that the *costs* of misclassification of a class 2 object as a class 1 object, and vice versa are, respectively.

$$c(1|2) \qquad \text{and} \qquad c(2|1).$$

Then the expected cost of misclassification is therefore

$$c(2|1)P(2|1)p_1 + c\,(1|2)\,P(1|2)p_2.$$

The idea is to choose the regions $\mathcal{R}_1$ and $\mathcal{R}_2$ so that this expected cost is minimized. This can be achieved by comparing the predictive probability density functions at each point x

$$\mathcal{R}_1 \equiv \left\{ x : \frac{f_1\,(x)}{f_2\,(x)} \frac{p_1}{p_2} \geq \frac{c\,(1|2)}{c\,(2|1)} \right\}$$

or, equivalently

$$\mathcal{R}_2 \equiv \left\{ x : \frac{f_1\,(x)}{f_2\,(x)} \frac{p_1}{p_2} < \frac{c\,(1|2)}{c\,(2|1)} \right\}$$

or indeed minimizing the total probability of misclassification

$$p_1 \int_{\mathcal{R}_2} f_1\,(x)\;\, dx + p_2 \int_{\mathcal{R}_1} f_2\,(x)\;\, dx$$

If $p_1 = p_2$, then

$$\mathcal{R}_1 \equiv \left\{ x : \frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2)}{c(2|1)} \right\}$$

and if $c(1|2) = c(2|1)$, equivalently

$$\mathcal{R}_1 \equiv \left\{ x : \frac{f_1(x)}{f_2(x)} \geq \frac{p_2}{p_1} \right\}$$

and finally if $p_1 = p_2$ and $c(1|2) = c(2|1)$ then

$$\mathcal{R}_1 \equiv \left\{ x : \frac{f_1(x)}{f_2(x)} \geq 1 \right\} \equiv \{ x : f_1(x) \geq f_2(x) \}$$

4.2 CLASSIFICATION FOR TWO NORMAL SAMPLES

Suppose that we have two (multivariate) normal classes (in d dimensions), that is where

- **class 1**: $X \sim N_d(\mu_1, \Sigma_1)$

$$f_1(x) = \left(\frac{1}{2\pi}\right)^{d/2} \frac{1}{|\Sigma_1|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)\right\}$$

- **class 2**: $X \sim N_d(\mu_2, \Sigma_2)$

$$f_2(x) = \left(\frac{1}{2\pi}\right)^{d/2} \frac{1}{|\Sigma_2|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_2)^T \Sigma_2^{-1}(x-\mu_2)\right\}$$

We sometimes assume that $\Sigma_1 = \Sigma_2 = \Sigma$ (*homogeneity of variances).* Using the previous formula, we identify the following **classification rule**; we allocate an observation with predictor variable x_0 to class 1 if

$$(\mu_1 - \mu_2)^T \Sigma^{-1} x_0 - \frac{1}{2}(\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 + \mu_2) \geq \log\left[\frac{c(1|2)\, p_2}{c(2|1)\, p_1}\right]. \quad (11)$$

More generally, if $\Sigma_1 \neq \Sigma_2$, we allocate an observation with predictor variable x_0 to class 1 if

$$-\frac{1}{2} x_0^T \left(\Sigma_1^{-1} - \Sigma_2^{-1}\right) x_0 + \left(\mu_1^T \Sigma_1^{-1} - \mu_2^T \Sigma_2^{-1}\right) x_0 - k \geq \log\left[\frac{c(1|2)\, p_2}{c(2|1)\, p_1}\right] \quad (12)$$

where

$$k = \frac{1}{2}\log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + \frac{1}{2}\left(\mu_1^T \Sigma_1^{-1} \mu_1 - \mu_2^T \Sigma_2^{-1} \mu_2\right)$$

The parameters μ_1, μ_2 and Σ, Σ_1 and Σ_2 may be estimated from training data.

- if the covariance matrices are presumed **equal** then we have a total of

$$2d + \frac{1}{2}d\,(d+1)$$

  parameters to estimate

- if the covariance matrices are presumed **unequal** then we have a total of

$$2d + d\,(d+1)$$

  parameters to estimate

Thus with limited data in d dimensions, we may be limited in the type of analysis can be done. In fact, we may have to further restrict the type of covariance structure that we may assume; for example, we might have to restrict attention to

- **diagonal** covariance matrices ($2d$ parameters in total),
- or an assumption of **sphericity** ($2\,(d+1)$ parameters in total)

Despite their simplicity, such models often work well in practice.

4.3 DISCRIMINATION

Discriminant analysis works in a very similar fashion; from equations (11) and (12) we note that the boundary between regions $\mathcal{R}_1$ and $\mathcal{R}_2$ takes one of two forms

- **Equal covariances**: we have a **straight line/plane** defined by an equation of the form

  $$A_1 x + a_0$$

  where A_1 is a $d \times d$ matrix

- **Unequal covariances**: we have a **quadratic surface** defined by an equation of the form

  $$x^T B_2 x + B_1 x + b_0$$

  where B_1 and B_2 are $d \times d$ matrices.

4.4 ASSESSMENT OF CLASSIFICATION ACCURACY

The performance of a classification rule can be achieved in a number of ways: we can examine

- the **within-sample** classification error: the proportion of elements in the training sample that are misclassified by the rule
- the **leave-one-out** classification error: the proportion of elements in the training sample when the model is built (that is, the parameters are estimated) on a training sample that omits a single data point, and then attempts to classify that point on the trained model

- an m-**fold cross-validation** : the data are split into m subsamples of equal size, and one is selected at random to act as a **pseudo-test** sample. The remaining data are used as **training** data to build the model, and the prediction accuracy on the pseudo-test sample is computed. This procedure is repeated for all possible splits, and the prediction accuracy computed as a average of the accuracies over all of the splits.

- accuracy using **bootstrap resampling** to achieve the **cross-validation** based estimates of accuracy from above.

The theory behind the assessment of classification accuracy is complex.

4.5 ROC CURVES

Receiver Operating Characteristic (ROC) curves can also be used to compare the classification performance classifiers. We consider the results of a particular classifier for two populations, say one population with a disease, the other population without the disease. Suppose that a single characteristic, x, is to be used to classify individuals.

The classification procedures above reduce to a simple rule; we classify an individual to class 1 if

$$x < t_0$$

for some threshold t_0, and to class 2 otherwise. We then consider the following quantities:

- **Sensitivity**: probability that a test result will be positive when the disease is present (true positive rate, expressed as a percentage).
- **Specificity**: probability that a test result will be negative when the disease is not present (true negative rate, expressed as a percentage).
- **Positive likelihood ratio**: ratio between the probability of a positive test result given the presence of the disease and the probability of a positive test result given the absence of the disease

$$\frac{\text{True Positive Rate}}{\text{False Positive Rate}}$$

- **Negative likelihood ratio:** ratio between the probability of a negative test result given the presence of the disease and the probability of a negative test result given the absence of the disease

$$\frac{\text{False Negative Rate}}{\text{True Negative Rate}}$$

- **Positive predictive value**: probability that the disease is present when the test is positive (expressed as a percentage).
- **Negative predictive value:** probability that the disease is not present when the test is negative (expressed as a percentage).

		Disease Class		
		1	2	Total
Predicted	1	a	c	$a+c$
Class	2	b	d	$b+d$
	Total	$a+b$	$c+d$	$a+b+c+d$

- Sensitivity:/Specificity:

$$Sensitivity : \frac{a}{a+b} \qquad Specificity : \frac{d}{c+d}$$

- Likelihood Ratios

$$PLR = \frac{Sensitivity}{1 - Specificity} \qquad NLR = \frac{1 - Sensitivity}{Specificity}$$

- Predictive Values

$$PPV = \frac{a}{a+c} \qquad NPV = \frac{d}{b+d}$$

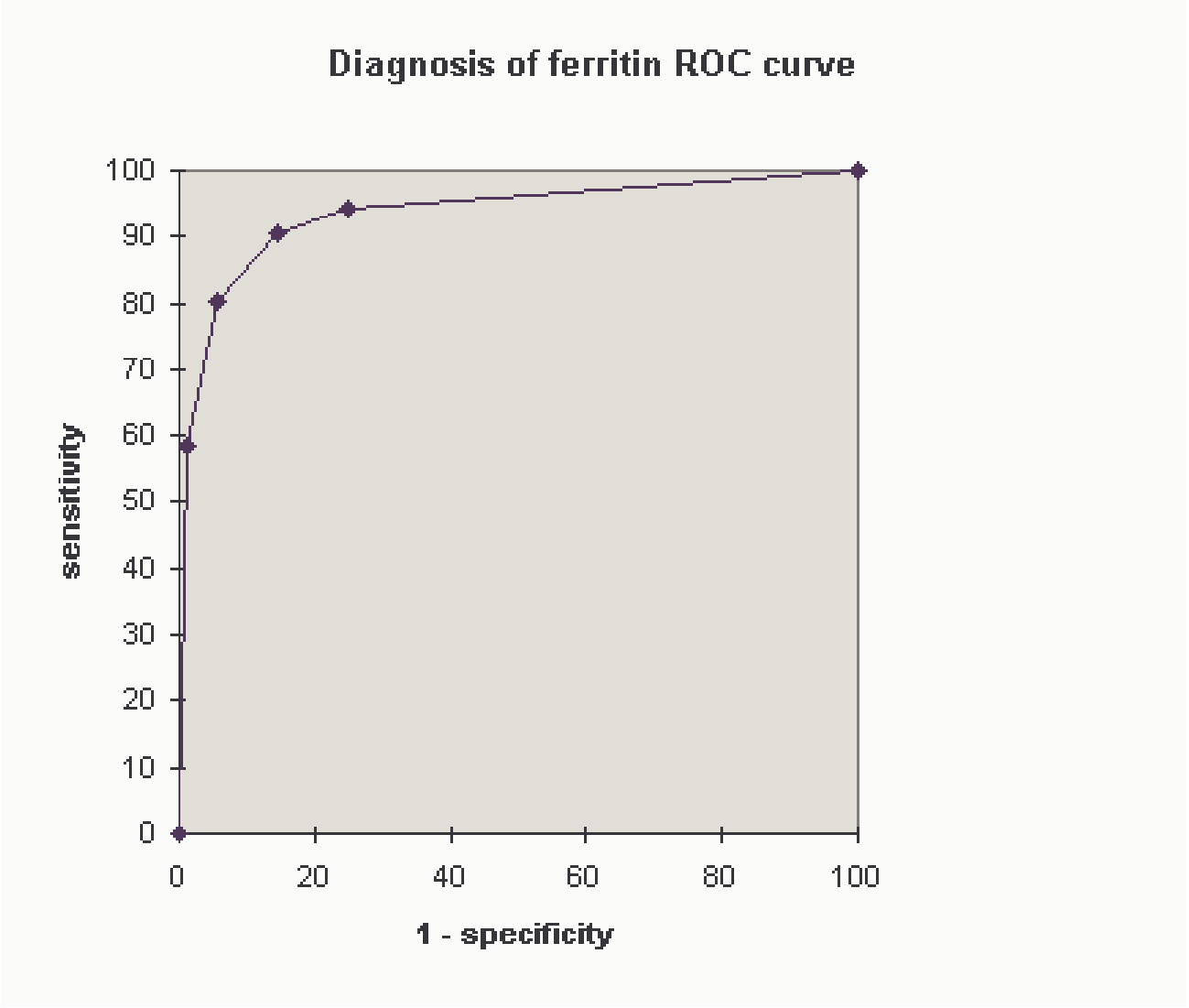
As the classifier producing the predicted class depends on the threshold t_0, we can produce a plot of how these quantities change as t_0 changes.

If we plot

$$\begin{aligned} y(t_0) &: \textit{Sensitivity} \text{ at } t_0 \\ x(t_0) &: 1 - \textit{Specificity} \text{ at } t_0 \end{aligned}$$

then we obtain an **ROC curve**;

- for a good classifier would rise steeply and then flatten off ; such a curve would have a large area underneath it on the unit square (the domain of $(x(t_0), y(t_0))$)
- for a poor classifier would be have an ROC curve near the line $y = x$.


Diagnosis of ferritin ROC curve
sensitivity
100
90
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100
1 - specificity

ROC Curve

4.6 GENERAL CLASSIFICATION SCHEMES

The general exercise of classification can be seen as a exercise in **regression modelling** for a nominal categorical variable. Previously, we studied **regression**, and more briefly **generalized linear regression**.

- For a **binary response,** or a **two-class** problem, we can use **logistic** or **binary** regression
- For a **multinomial response,** or a **multi-class** problem we can use **multinomial** regression

Because of this regression context, we can use all the previous tools for analysis in regression models that we have used previously.