USING STATISTICS IN RESEARCH

David A. Stephens

Department of Mathematics, Imperial College

d.stephens@imperial.ac.uk
stats.ma.ic.ac.uk/~das01/StatsShortCourse/

June 18, 2003

WEEK 3

REGRESSION, CORRELATION

&

RELATED METHODS

SECTION 6.

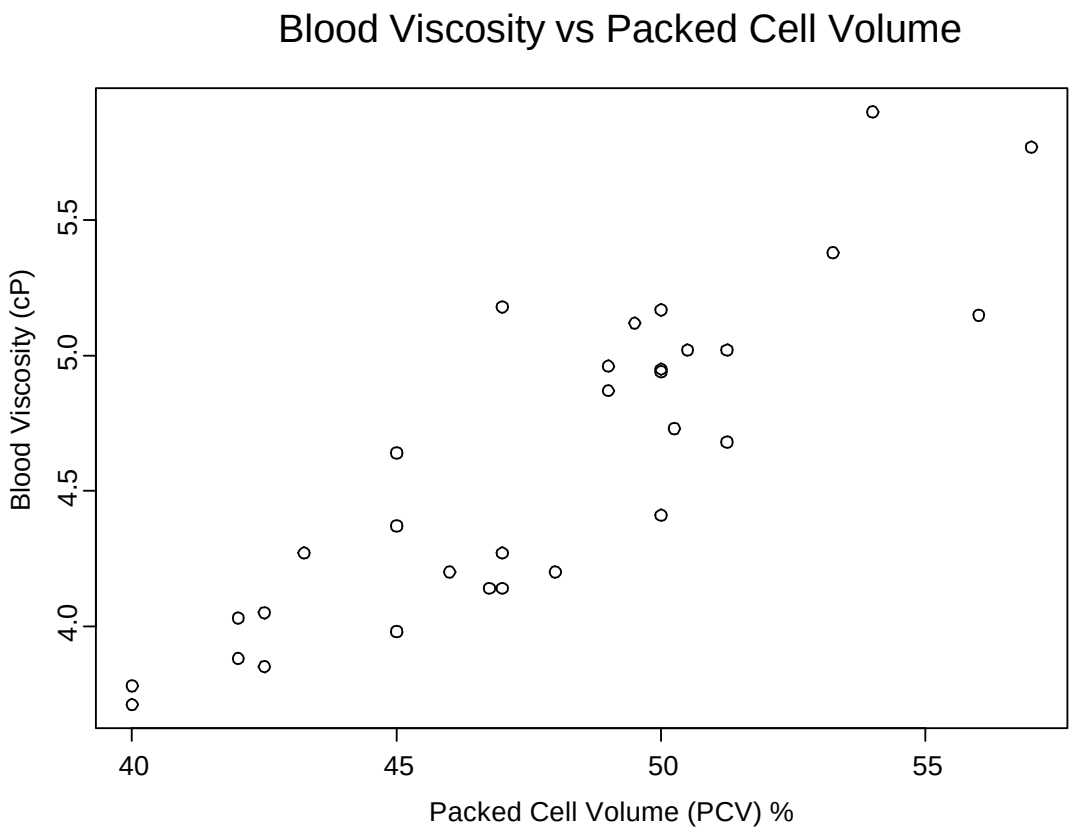
REGRESSION MODELLING

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

6.1 LINEAR REGRESSION

EXAMPLE



6.1.1 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**.

6.1.2 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

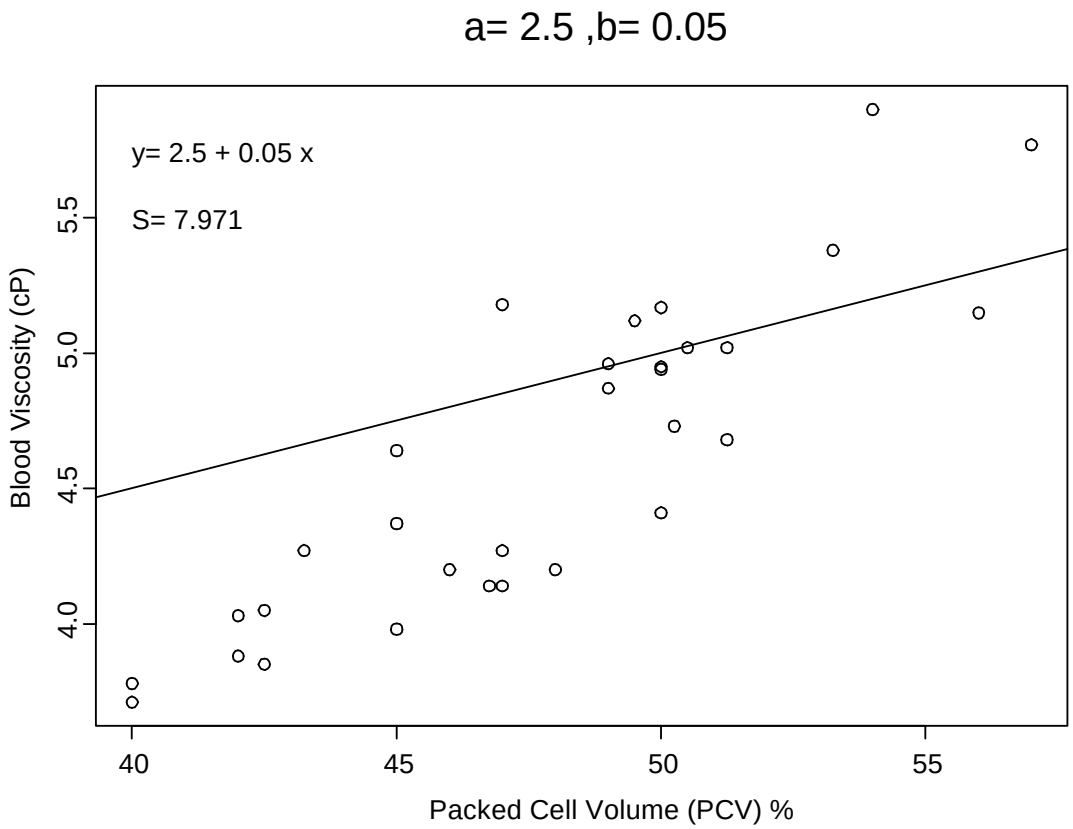
$$e_i = y_i - y_i^{(P)} = y_i - \alpha - \beta x_i$$

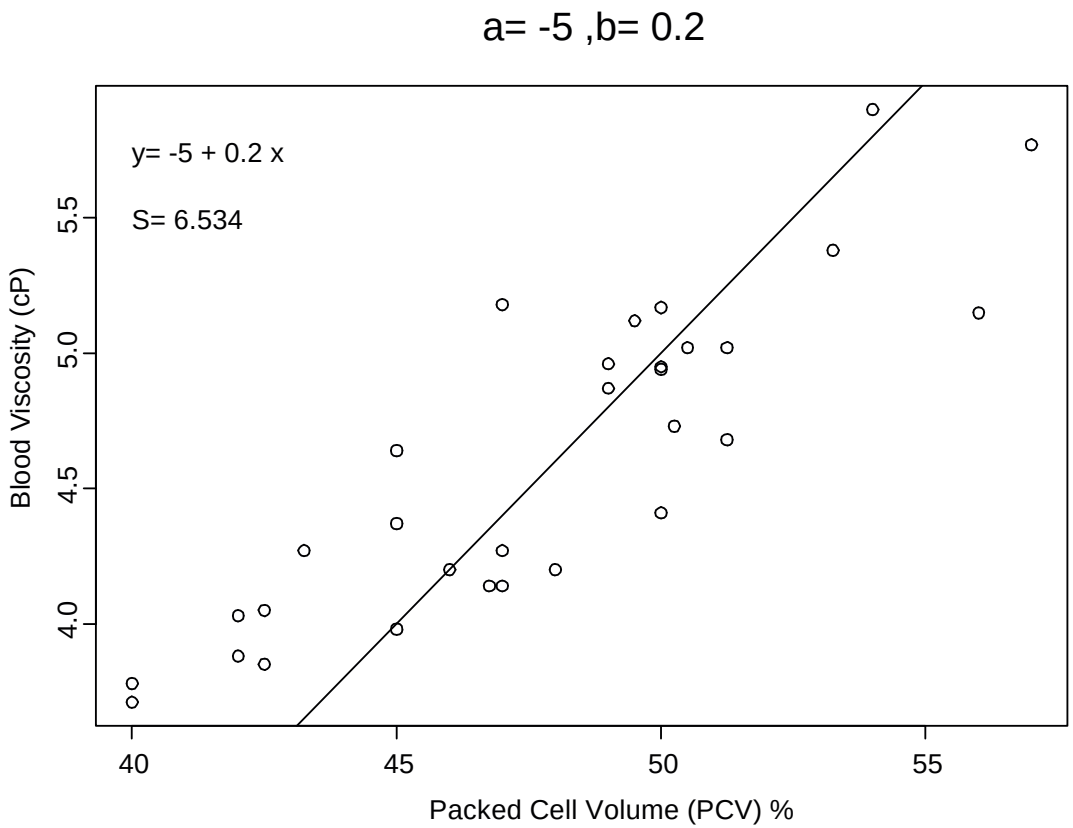
that is, e_i is the vertical discrepancy between the **observed** and **expected** values of Y. The objective in least-squares estimation is find a "line of best fit", and this is achieved by inspecting the squares of the error terms e_i, and choosing α and β such that the sum of the squared errors is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

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

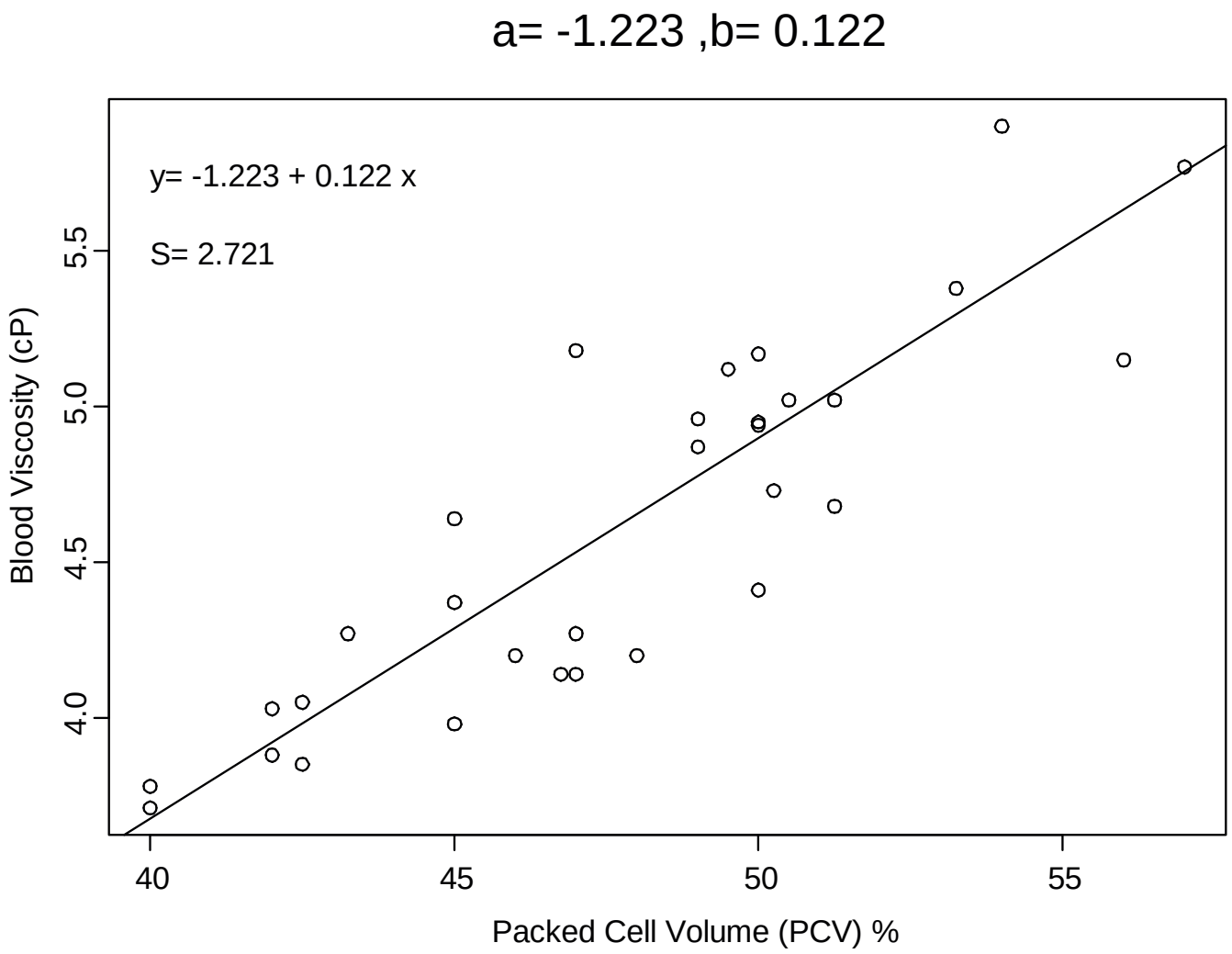
so that

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

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.


a= -1.223 ,b= 0.122
y= -1.223 + 0.122 x
S= 2.721
Blood Viscosity (cP)
4.0
4.5
5.0
5.5
40
45
50
55
Packed Cell Volume (PCV) %

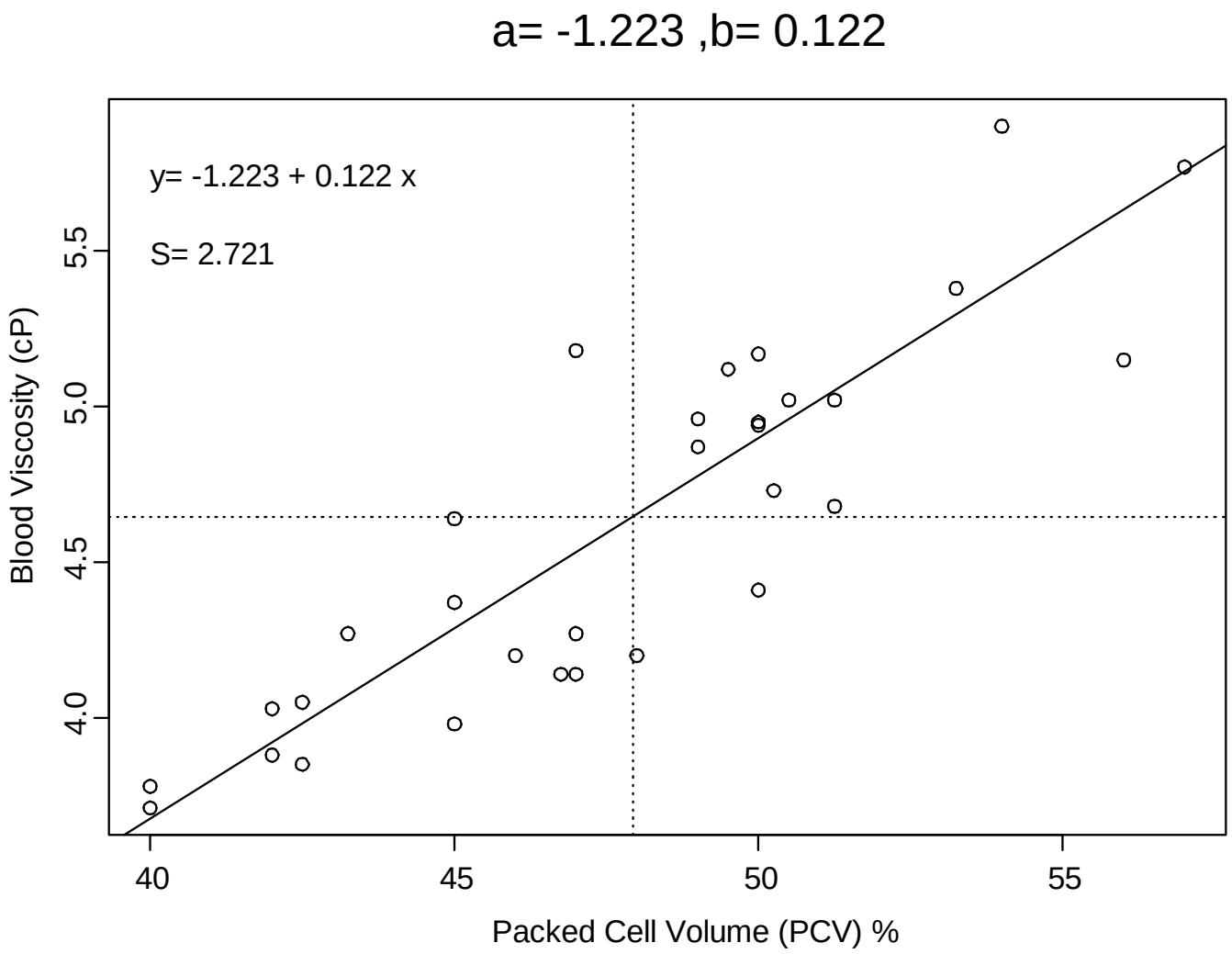
Alternative formulae: let

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

Then

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

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


a= -1.223 ,b= 0.122
y= -1.223 + 0.122 x
S= 2.721
Blood Viscosity (cP)
4.0
4.5
5.0
5.5
40
45
50
55
Packed Cell Volume (PCV) %

6.1.3 LEAST-SQUARES AS MAXIMUM LIKELIHOOD

Suppose that X and Y follow a linear regression model

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

and recall that the error terms e_i were defined

$$e_i = y_i - \alpha - \beta x_i.$$

Now, e_i is the vertical discrepancy between observed and expected behaviour, and thus e_i could be interpreted as the observed version of a **random variable**, say ϵ_i, which represents the random uncertainty involved in measuring Y for a given X.

A plausible probability model might therefore be that the random variables ϵ_i, $i = 1, ...n$, were independent and identically distributed, and

$$\epsilon_i \sim N(0, \sigma^2),$$

for some error variance parameter σ^2. Implicit in this assumption is that the distribution of the random error in measuring Y does not depend on the value of X at which the measurement is made.

This distributional assumption about the error terms leads to a probability model for the variable Y. As we can write

$$Y = \alpha + \beta X + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$, then given on $X = x_i$, we have the conditional distribution Y_i as

$$Y_i | X = x_i \sim N(\alpha + \beta x_i, \sigma^2),$$

where random variables Y_i and Y_j are **independent** (as ϵ_i and ϵ_j are independent).

On the basis of this probability model, we can derive a likelihood function, and hence derive maximum likelihood estimates. For example, we have the likelihood $L(\theta) = L(\alpha, \beta, \sigma^2)$ defined as the product of the n conditional density terms derived as the conditional density of the observed y_i given x_i,

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(y_i; x_i, \theta) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} (y_i - \alpha - \beta x_i)^2 \right\} \\ &= \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \right\} \end{aligned}$$

The maximum likelihood estimates of α and β, and error variance σ^2, are obtained as the values at which $L(\alpha, \beta, \sigma^2)$ is **maximized**. But, $L(\alpha, \beta, \sigma^2)$ is maximized when the term in the exponent, that is

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

is minimized. But this is **precisely** the least-squares criterion described above, and thus the m.l.e s of α and β assuming a Normal error model are **exactly equivalent** to the least-squares estimates.

6.1.4 ESTIMATES OF ERROR VARIANCE

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

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

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 $X = x_i$.

6.1.5 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

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

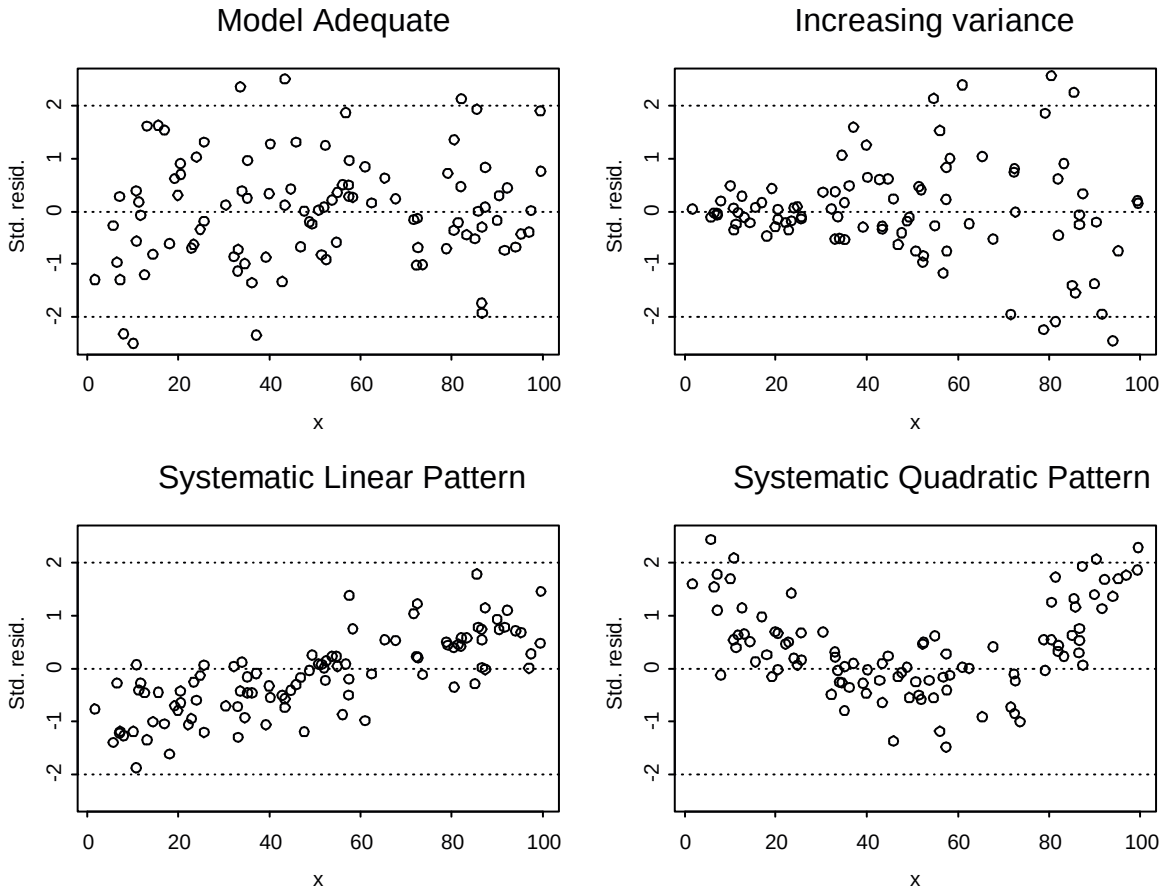
so that instead of inspecting merely residuals we inspect **standardized residual**

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

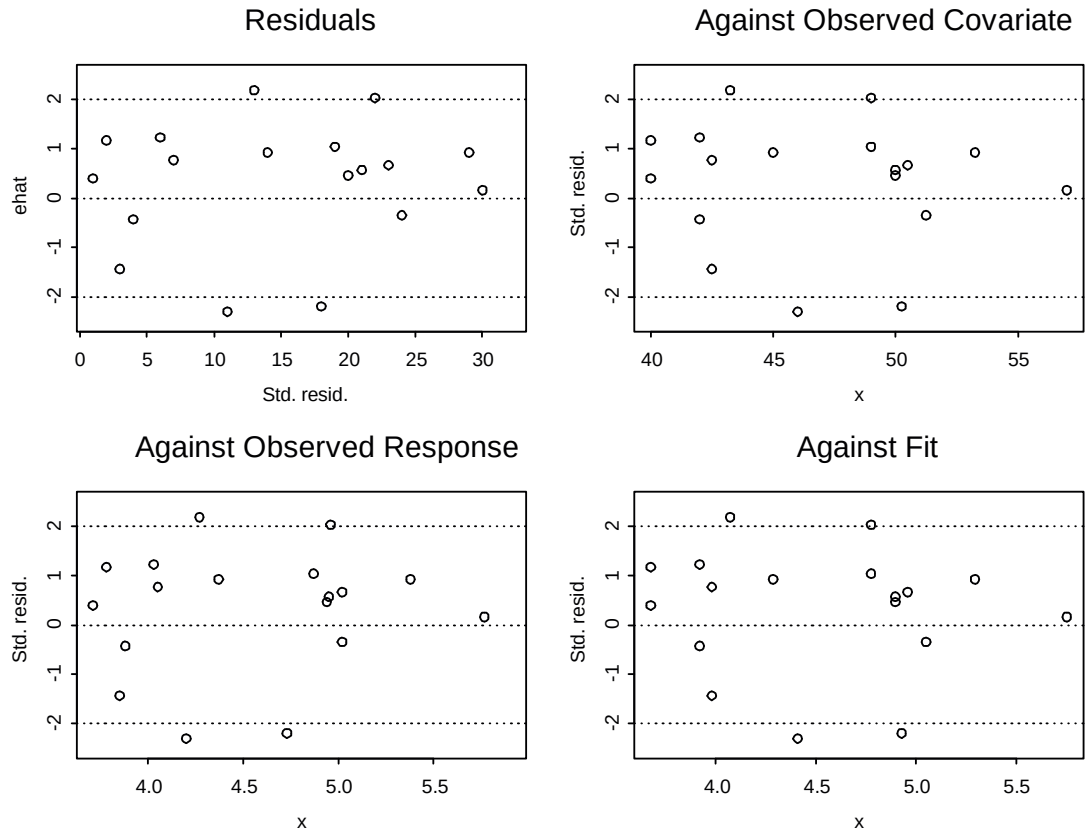
These standardized residuals should

- 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



6.1.6 PREDICTION FOR A NEW COVARIATE 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

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

6.1.7 STANDARD ERRORS OF ESTIMATORS AND T-STATISTICS

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.

6.1.8 HYPOTHESIS TESTS AND CONFIDENCE 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

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

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

6.1.9 WORKED EXAMPLE

The following data are believed to follow a linear regression model;

x	0.54	2.03	3.15	3.96	6.25	8.17
y	11.37	11.21	11.61	8.26	14.08	16.25
x	11.08	12.44	14.04	14.34	18.71	19.90
y	11.00	14.94	16.91	15.78	21.26	20.25

We want to calculate estimates of α and β from these data. First, we calculate the summary statistics;

$$S_x = \sum_{i=1}^{n} x_i = 118.63 \qquad S_y = \sum_{i=1}^{n} y_i = 172.92$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 = 1598.6 \qquad S_{xy} = \sum_{i=1}^{n} x_i y_i = 1930.9$$

with $n = 12$ which leads to parameter estimates

$$\widehat{\beta} = \frac{nS_{xy} - S_x S_y}{nS_{xx} - \{S_x\}^2} = \frac{12 \times 1930.9 - 118.63 \times 172.92}{12 \times 1598.6 - (118.63)^2} = 0.5201$$

$$\widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x} = 14.410 - 0.5201 \times 9.8842 = 9.269$$

The **corrected variance estimate**, s^2, is given by

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{\alpha} - \widehat{\beta}x_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 3.438 \qquad \Longrightarrow \qquad s = 2.332$$

The **standard errors** for the two parameters are given by

$$s.e.(\widehat{\alpha}) = s\sqrt{\frac{S_{xx}}{nS_{xx} - \{S_x\}^2}} = 1.304$$

$$s.e.(\widehat{\beta}) = s\sqrt{\frac{n}{nS_{xx} - \{S_x\}^2}} = 0.113$$

The **t-statistics** for the two parameters are given by

$$t_\alpha = \frac{\widehat{\alpha}}{s.e.(\widehat{\alpha})} = \frac{9.269}{1.304} = 7.109$$

$$t_\beta = \frac{\widehat{\beta}}{s.e.(\widehat{\beta})} = \frac{0.520}{0.113} = 4.604.$$

The 0.975 percentile of a Student-t distribution with $n - 2 = 10$ degrees of freedom is found from tables to be 2.228. Both t-statistics are more extreme than this critical value, and hence it can be concluded that both parameters are significantly different from zero.

To calculate the **confidence intervals** for the two parameters. we need to use the 0.975 percentile of a $St(10)$ distribution. >From above, we have that $St(10)(0.975) = 2.228$, and so the confidence intervals are given by

$$\alpha \quad : \quad \widehat{\alpha} \pm t_{n-2}(0.975) s.e.(\widehat{\alpha}) = 9.269 \pm 2.228 \times 1.304 = (6.364 : 12.174)$$

$$\beta \quad : \quad \widehat{\beta} \pm t_{n-2}(0.975) s.e.(\widehat{\beta}) = 0.5201 \pm 2.228 \times 0.113 = (0.268 : 0.772)$$

so that, informally, we are 95% certain that the true value of α lies in the interval $(6.724 : 12.174)$, and that the true value of β lies in the interval $(0.268 : 0.772)$. This amounts to evidence that, for example, $\alpha \neq 0$ (as the confidence interval for α does not contain 0), and evidence that $\beta \neq 1$ (as the confidence interval for β does not contain 1).

This fit leads to the following fitted values and residuals;

x	0.54	2.03	3.15	3.96	6.25	8.17
y	11.37	11.21	11.61	8.26	14.08	16.25
$\hat{y}$	9.55	10.33	11.95	12.37	12.52	13.52
e	1.82	0.88	-0.34	-4.11	1.56	2.73

x	11.08	12.44	14.04	14.34	18.71	19.90
y	11.00	14.94	16.91	15.78	21.26	20.25
$\hat{y}$	15.03	15.73	16.57	16.73	19.00	19.62
e	-4.03	-0.80	0.34	-0.95	2.26	0.63

6.2 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 to test the hypothesis

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

6.2.1 THE Z-TEST FOR CORRELATION

An alternative 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.

6.2.2 THE T-TEST FOR CORRELATION

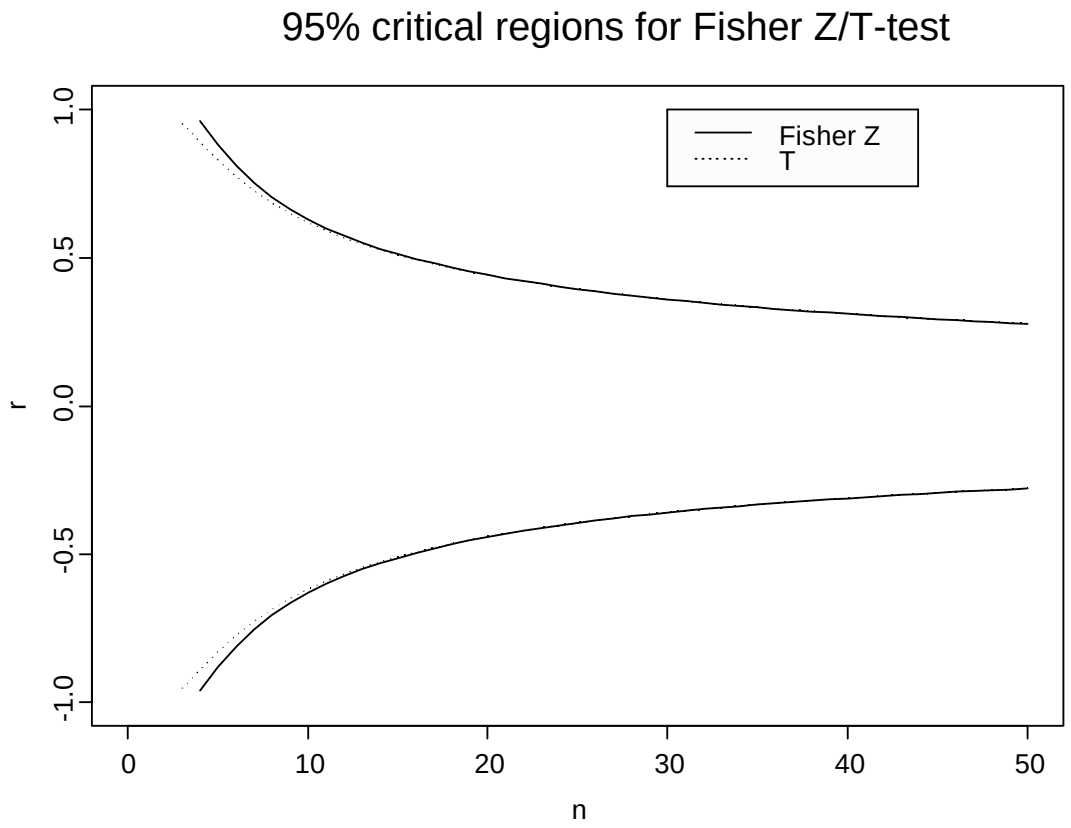
An alternative test of the hypothesis is based on the test statistic

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

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.

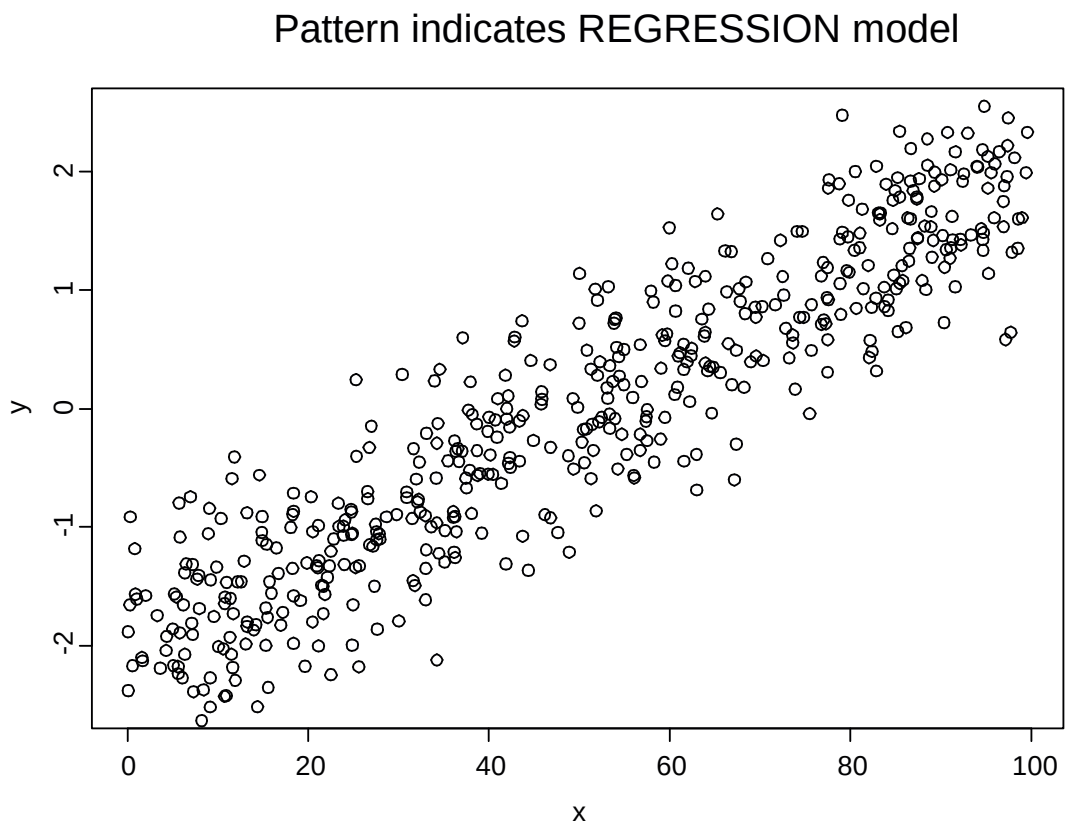

95% critical regions for Fisher Z/T-test
Fisher Z
T
r
1.0
0.5
0.0
-0.5
-1.0
0
10
20
30
40
50
n

EXAMPLE PCV/Blood Viscosity $r = 0.879$

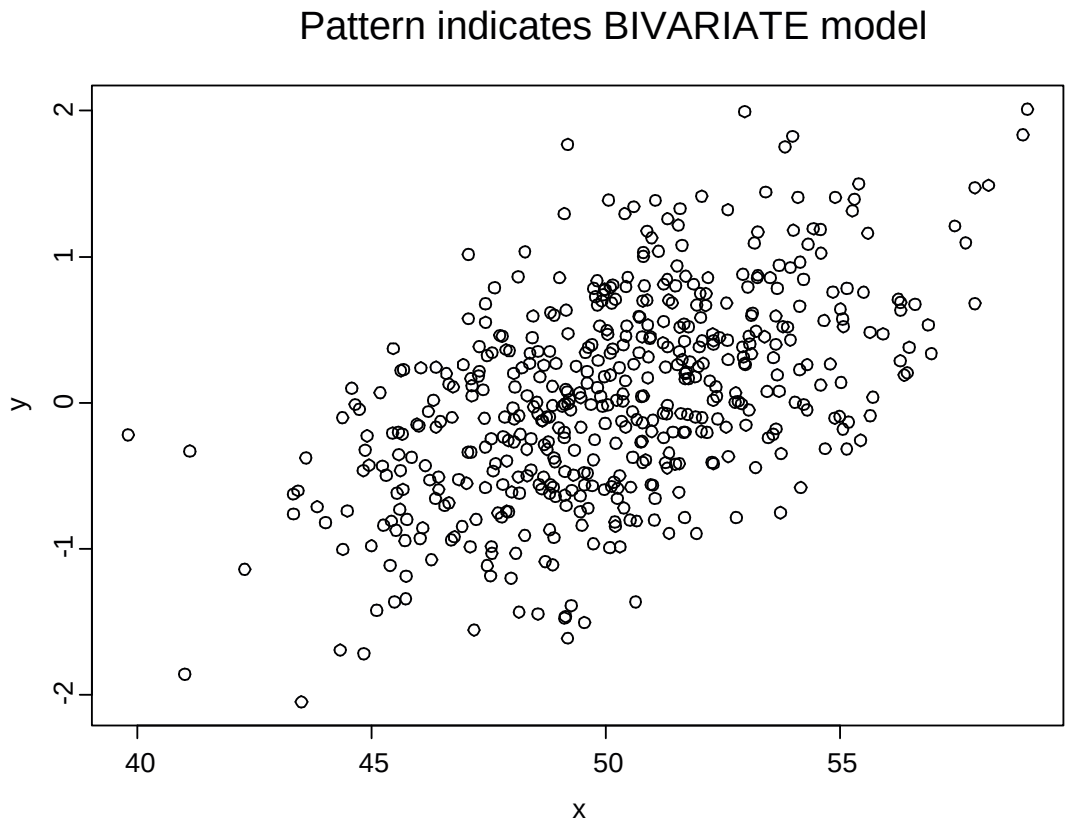
- FISHER Z TEST $z = 7.38065$ $(p = 7.875952e - 014)$
- T-TEST $t = 10.08784, (p = 1.865336e - 011)$

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

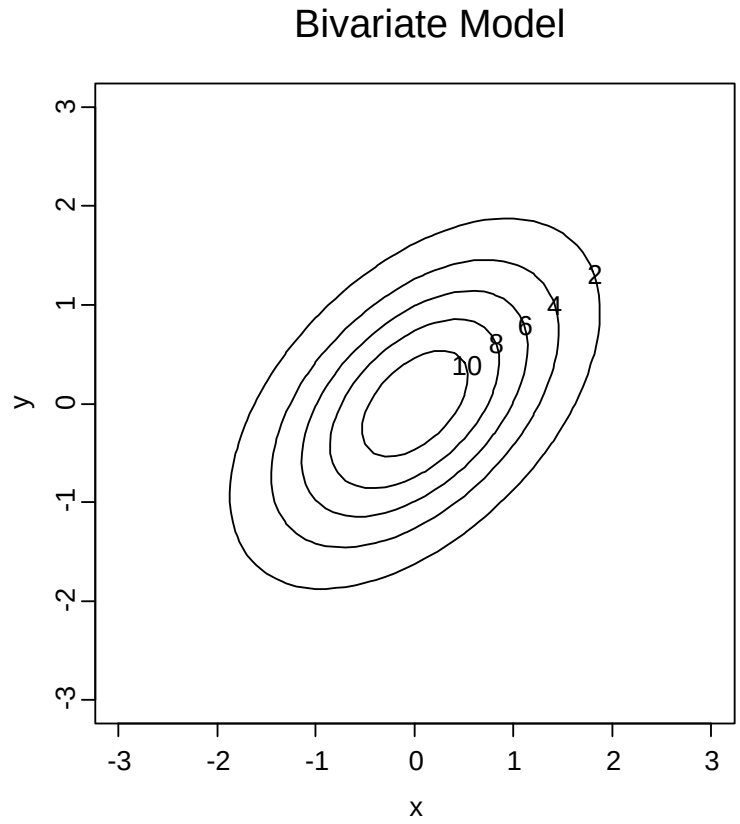
Regression


Pattern indicates REGRESSION model
y
x
-2
-1
0
1
2
0
20
40
60
80
100

Bivariate Data



Bivariate Contour



6.3 MULTIPLE LINEAR REGRESSION

In everything that is described above, we have used a model in which we predicted a response Y from a single covariate X. This simple model 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.

6.4 THE NORMAL LINEAR REGRESSION MODEL

We assume that the variables to be modelled are as follows; we will observe paired data, with response data y_i paired to predictor variables stored in vector form $x_i = (x_{i1}, ..., x_{iD})^T$, and our aim is to explain the variation in $(y_1, ..., y_n)$. We achieve this by modelling the conditional distribution of response variable Y_i given the observed value of predictor variable $X_i = x_i$. Specifically, we may write

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

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

$$Y_i | X_i = x_i \sim N\left(\beta_0 + \sum_{j=1}^{D} \beta_j x_{ij}, \sigma^2\right) \tag{2}$$

so that

$$E_{f_{Y|X}}\left[Y_i | X_i = x_i\right] = \beta_0 + \sum_{j=1}^{D} \beta_j x_{ij}.$$

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

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

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

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

and to mimic (2)

$$Y \sim N_n\left(\mathbf{X}\beta, \sigma^2 I_n\right) \tag{3}$$

where I_n is the $n \times n$ identity matrix, giving a joint pdf for Y given $\mathbf{X}$ of the form

$$f_{Y|\beta,\sigma^2}(y;\beta,\sigma^2) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\left(y-\mathbf{X}\beta\right)^T\left(y-\mathbf{X}\beta\right)\right\} \tag{4}$$

6.5 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 D original predictors, that is

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

is some scalar function, for example, we could have

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

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

6.6 FACTOR PREDICTORS: CONTRAST PARAMETERIZATIONS

The linear model formulation can be used for **categorical** predictors, or **factors**; suppose that predictor X takes K distinct **levels** $(l_1, l_2, ..., l_K)$, and that there is a different mean response for each level

$$E[Y] = \begin{cases} \beta_1 & x = c_1 \\ \beta_2 & x = c_2 \\ & \\ \beta_K & x = c_K \end{cases}$$

The parameters $(\beta_1, \beta_2, ..., \beta_K)$ can be estimated in the usual way.

Other parameterizations that will permit inferences about specific differences of interest, or **contrasts,** include

- **Deviation:** differences from overall mean level

$$\mu_0 = \frac{1}{K}(\beta_1 + \beta_2 + ... + \beta_K)$$

$$\mu_k = \left(1 - \frac{1}{K}\right)\beta_k - \frac{1}{K}\sum_{j \neq k}\beta_j \qquad k = 1, 2, ..., K-1$$

- **Simple:** differences of levels $c_1, ..., c_{K-1}$ from c_K

$$\mu_0 = \frac{1}{K}(\beta_1 + \beta_2 + ... + \beta_K)$$

$$\mu_k = \beta_k - \beta_K \qquad k = 1, 2, ..., K-1$$

(with arbitrary labelling of the levels)

- **Helmert:** differences of each level from mean of **subsequent** categories

$$\mu_0 = \frac{1}{K}(\beta_1 + \beta_2 + ... + \beta_K)$$

$$\mu_k = \beta_k - \frac{1}{K-k}\sum_{j=k+1}^{K}\beta_K \qquad k = 1, 2, ..., K-1$$

- **Difference:** differences of each level from mean of **previous** categories

$$\mu_0 = \frac{1}{K}(\beta_1 + \beta_2 + ... + \beta_K)$$

$$\mu_k = \beta_{k+1} - \frac{1}{k}\sum_{j=1}^{k}\beta_k \qquad k = 1, 2, ..., K-1$$

- **Polynomial:** for ordinal categorical variables

  CONTRAST 1 : **LINEAR** EFFECT ACROSS LEVELS
  CONTRAST 2 : **QUADRATIC** EFFECT ACROSS LEVELS
  $\vdots$

- **Repeated:** differences for **adjacent** levels

$$\mu_0 = \frac{1}{K}(\beta_1 + \beta_2 + ... + \beta_K)$$

$$\mu_k = \beta_k - \beta_{k+1} \qquad k = 1, 2, ..., K-1$$

Most of these contrast specifications can be written as linear transformations of the original parameters, that is

$$\mu = C\beta$$

for a $K \times K$ matrix C.

Often, **orthogonal** contrasts are used for ease of interpretation; for orthogonal linear contrasts

$$C^T C = I$$

where I is the $K \times K$ **identity** matrix (ones on the diagonal, zeros elsewhere).

Contrasts can be defined to examine specific effects.

6.7 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

$$S(\beta) = S\left(\hat{\beta}\right) + \left(\hat{\beta} - \beta\right)^T \left(\mathbf{X}^T\mathbf{X}\right)\left(\hat{\beta} - \beta\right) \qquad \text{or} \qquad TSS = RSS + FSS.$$

Now, using the distributional results above, we can construct the following **ANOVA Table** to test the hypothesis

$$H_0 : \beta_1 = ... = \beta_K = 0$$

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 using equation (**??**). 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,

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

Suppose that there are $n_1, ..., n_K$ observations in the K treatment groups respectively. Then the design matrix in the corresponding (full) linear

model takes the form

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \\ \mathbf{X}_K \end{bmatrix} \qquad \mathbf{X}_k = \underbrace{\begin{bmatrix} 0 & 0 & \cdots & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 1 & \cdots & 0 \end{bmatrix}}_{K \text{ columns}} \Bigg\} n_k \text{ rows}$$

that is, $\mathbf{X}_k$ is a $n_k \times K$ block matrix with only the kth column non-zero, and equal to the $n_k \times 1$ vector of 1s. Under the assumption that the observed responses are normally distributed **with common variance** σ^2 we are in the linear model framework, and all of the above likelihood and statistical theory applies.

6.8 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

6.9 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

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

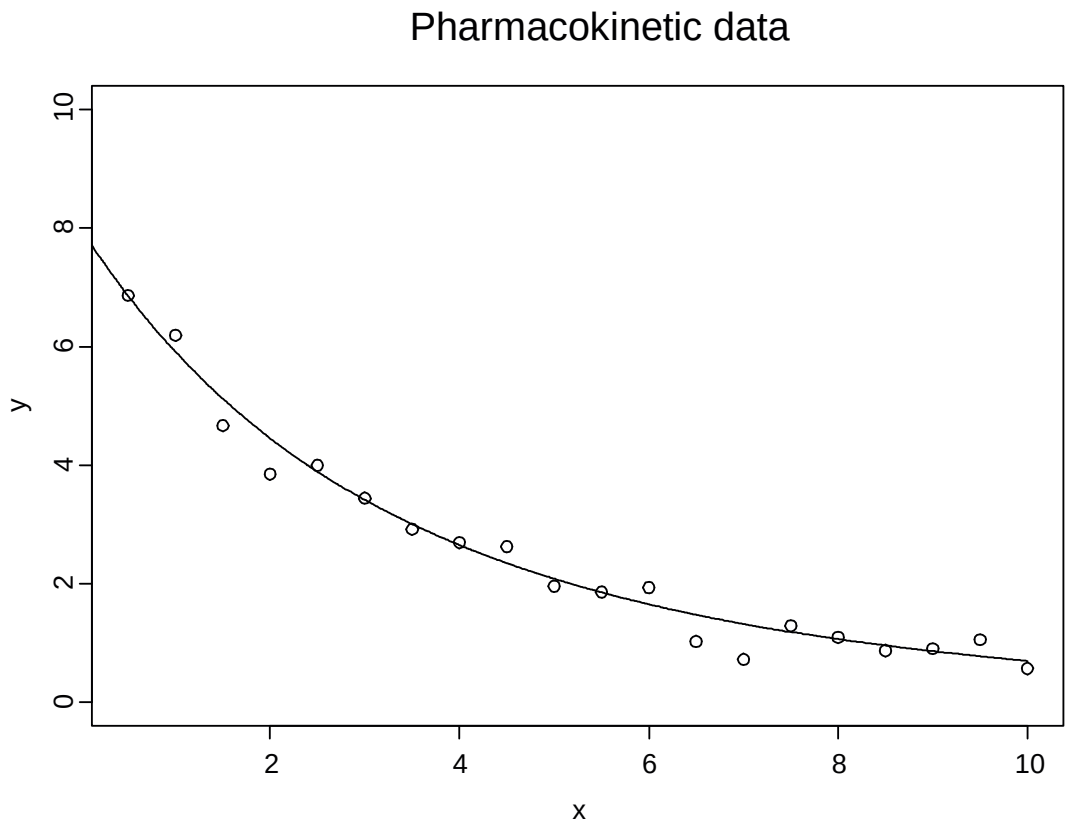
for some non-linear function g of the parameters.

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

6.10 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.

6.10.1 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

- **POISSON MODEL**

$$f_{Y|\theta,\phi}(y;\theta,\phi) = f_{Y|\lambda,\phi}(y;\lambda,\phi) = \frac{e^{-\lambda_0}\lambda^y}{y!}$$

- **BINOMIAL MODEL**

$$f_{Y|\theta,\phi}(y;\theta,\phi) = f_{Y|\theta,\phi}(y;\theta,\phi) = \binom{n}{y}\theta^y\,(1-\theta)^{n-y}$$

6.10.2 LINK FUNCTIONS

For the **Poisson** model;

- The **canonical link** is the **log** link: it is the link function that connects the naive parameter to the linear predictor

$$\lambda = \log \lambda_0 = x^T \beta$$

Here,

$$E_{f_{Y|\lambda}}[Y] = \lambda_0 = \exp\{\lambda\} \therefore g\left(E_{f_{Y|\lambda}}[Y]\right) = x^T\beta \text{ where } g(t) = \log t$$

- **Power link**

$$g(t) = t^\alpha$$

for some real parameter α

- **Box-Cox link**

$$g(t) = \frac{t^{\alpha} - 1}{\alpha}$$

for some real parameter α.

For the **Binomial** model;

- The **canonical link** is the **logit** link:

$$\theta = \log\left(\frac{\theta_0}{1-\theta_0}\right) = x^T\beta$$

Here,

$$E_{f_{M|\lambda}}[M] = \theta_0 = \frac{\exp\{\theta\}}{(1+\exp\{\theta\})} \therefore g\left(E_{f_{m|\lambda}}[M]\right) = x^T\beta$$

where $g(t) = \log\left(t/\left(1-t\right)\right)$

- **Probit link**

$$g(t) = \Phi^{-1}\left(\frac{\exp\{t\}}{1+\exp\{t\}}\right) = x^T\beta \qquad \therefore \qquad \Phi^{-1}(\theta_0) = x^T\beta$$

where Φ is the standard normal cdf.

- **Complementary log-log link**

$$g(t) = \log\{\log(1+\exp\{t\})\} = x^T\beta \qquad \therefore \qquad \log\{-\log(1-\theta_0)\} = x^T\beta$$

- **Log-log link**

$$g(t) = -\log\{-t + \log(1+\exp\{t\})\} = x^T\beta \qquad \therefore \qquad -\log\{-\log\theta_0\} =$$

6.10.3 CHECKING THE FIT OF A GLM

In the normal linear framework, the fit of a model is assessed by inspection of the magnitude of the residual sum of squares (RSS) and the fitted sum of squares (FSS) in an ANOVA F-test. For example, to test a model with K predictors plus an intercept ($K+1$ parameters) against the model with just an intercept (1 parameter), we use either a Chi-squared goodness-of-fit statistic or the F-ratio statistic, or by inspecting the error in fit as measured by the **residual,** e,

$$e = y_i - \hat{y}_i = y_i - x_i^T \hat{\beta}.$$

where $\hat{\beta}_0$ and $\hat{\beta}$ are mles computed under the 1 and $K+1$ parameter models respectively. In the GLM case, we will use similar, Likelihood Ratio (LR) statistics to perform tests.

6.10.4 DEVIANCE

In the following we use the following notation for data Y modelled via linear predictor $\eta = x^T\beta$ through canonical parameter θ and related expected value $\mu = E_{f_{Y|X,\beta}}[Y]$ with link function, g.. After the model is fitted, we have ML estimates in the linear predictor, β, for the following parameters

$$\hat{\eta} = x^T\hat{\beta} \qquad \hat{\theta} = g^{-1}\left(x^T\hat{\beta}\right) \qquad \hat{\mu} = h\left(\hat{\theta}\right)$$

We may also write $\hat{y} = \hat{\mu}$.

Deviance is a way of measuring the goodness of fit of a GLM. From a previous definition, the deviance, D, for a model M is the likelihood ratio statistic in an LR test of the model against the **saturated** model, S,

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

where

- $\hat{\beta}_M$ is the mle under model M
- $\hat{\beta}_S$ is the mle baseline model the saturated model, which corresponds to the **best possible fit,** and which occurs when

$$\hat{\mu}_i = \hat{y}_i = y_i$$

- 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:

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

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

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

6.10.5 STATISTICAL PROPERTIES OF DEVIANCE

Likelihood Ratio theory gives a means of calibrating the magnitude of the deviance; we have approximately

$$D^*\left(y;\hat{\theta},\phi\right) = \frac{D\left(y;\hat{\theta}\right)}{\phi} \sim \chi^2_{n-K-1}$$

if η has $K+1$ parameters. From this result we have two possible estimates of dispersion parameter ϕ; the **Deviance-based** estimate

$$\hat{\phi}_D = \frac{D\left(y;\hat{\theta}\right)}{n-K-1}$$

or the **Pearson-type** estimate

$$\hat{\phi}_P = \frac{1}{n-K-1}\sum_{i=1}^{n}\frac{\left(y_i-\hat{\mu}_i\right)^2}{V\left(\hat{\mu}_i\right)/w_i}$$

For choosing between two models

$$\begin{aligned} M_1 &: K_1 \text{ predictors, parameter } \theta_1 \\ M_2 &: K_2 \text{ predictors, parameter } \theta_2 \end{aligned}$$

where, without loss of generality $K_2 > K_1$, we have

$$\frac{D_{M_1}\left(y;\hat{\theta}_1\right) - D_{M_1}\left(y;\hat{\theta}_1\right)}{\phi} \sim \chi^2_{K_2-K_1}$$

and, if ϕ is not known

$$\frac{D_{M_1}\left(y;\hat{\theta}_1\right) - D_{M_1}\left(y;\hat{\theta}_1\right)}{\hat{\phi}\left(K_2-K_1\right)} \sim Fisher\left(K_2-K_1, n-K_2-1\right)$$

where $\hat{\phi}$ is either of $\hat{\phi}_D$ or $\hat{\phi}_P$.