BS1A APPLIED STATISTICS - LECTURES 1-14

GEOFF NICHOLLS

Date: 16 lectures, MT14.

1. Course

The course develops the theory of statistical methods, and introduces students to the analysis of data using a statistical package. The main topics of the MT14 course are: Practical aspects of normal linear models, Logistic regression and generalized linear models.

1.1. **Website.** Information about the course, classes and assessment etc are posted at

`http://www.stats.ox.ac.uk/~nicholls/sb1a/`

In particular Lecture notes, problem sheets, etc. are linked from that page.

2. Linear Regression

2.1. **Normal linear models.** Does y get bigger when there is more x? Normal linear models are attractive because they are simple, and in some respects easy to interpret. If, for a given observation process, they happen to give a correct or near-correct description of the distribution of the response y, and its dependence on predictive factors x, then they are hard to beat.

We choose to model a randomly variable response $Y = y$ as a linear function of explanatory variables $x_1, x_2, ..., x_p$. At the ith observation, we set the explanatory variables to values $\mathbf{x}_i = (x_{i,1}, x_{i,2}, ..., x_{i,p})$ (with $\mathbf{x}_i$ a row vector), and measure a response $Y_i = y_i$. Under a normal linear model, the expected response is given as a linear combination of the explanatory variables, weighted by parameters $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$,

$$y_i = \sum_{j=1}^{p} \beta_j x_{i,j} + \epsilon_i,$$

with $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ iid normal *errors* for $i = 1, 2, ..., n$. If $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_n)^T$ then $y_i = \mathbf{x}_i\beta + \epsilon_i$. If $y = (y_1, y_2, ..., y_n)^T$, and X is the $n \times p$ *design* matrix with rows $\mathbf{x}_i$, then our linear model has matrix form

$$y = X\beta + \epsilon.$$

We will write $X_j,\ j = 1, 2, ..., p$ for the columns of $X = (X_1, X_2, ..., X_p)$. We will use y and ϵ to denote both a column vector of responses, as above, *and* a single generic realization of the scalar response $Y = y$ etc. For example if we write down a model in terms of scalars y and ϵ,

$$y = \alpha + \gamma_1 x_1 + ... + \gamma_m x_m + \epsilon$$

omitting the $i = 1, 2, ..., n$ subscript, we have in mind $y = X\beta + \epsilon$ (now vectors) with $\beta_1 = \alpha$ and $\beta_i = \gamma_{i+1}$ and $p = m + 1$. In this example, $X_1 = 1_{n,1}$ that is, the first column if X is a column of ones and corresponds to the explanatory variable for the intercept parameter α.

The nomenclature assumes that we set the values of the explanatory variables and measure the response. When we choose the $\mathbf{x}_i$, $i = 1, 2, ..., n$ we are designing the experiment. We will see that not all designs are equally good. The subject of experimental design is one we will just touch on.

From time to time, we will assume that the columns of X are linearly independent vectors, that is, the explanatory variables are not linearly dependent. If they are, we can throw out linearly dependent columns till we have a linearly independent set; the discarded columns tell us nothing new about the measurement context. This issue comes up, for example, in Section **??**. We are, as a consequence, often assuming $p \leq n$, that is, we have more measurements than parameters.

The number one problem for interpreting linear models arises from correlation between explanatory variables. You can think of this as a kind of weak linear dependence between variables, and groups of variables.

Example 2.1. The dataset `cig` contains measurements of the carbon-monoxide (variable `CO`), tar and nicotine content and tobacco weight for $n = 25$ cigarettes. The data are plotted in Figure 1. In the normal linear model `CO` $\sim$ `1` + `Nicotine` + `Tar` + `Weight` the response is `CO` and all the other variables (including intercept) are

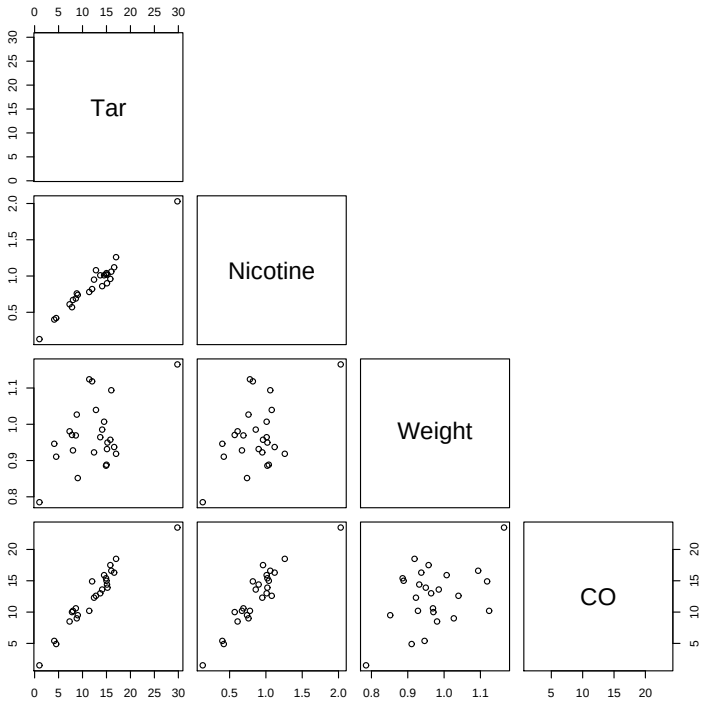

FIGURE 1. Cigarette CO data.

explanatory. This notation (which comes from R) means we are fitting the model

$$y = \beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + \epsilon$$

with $y =$ `CO` output for one cigarette, $x_1 = 1$ and x_2, x_3 and x_4 respectively the measured nicotine and tar content and weight of the cigarette.

```
> loc<-'http://www.stats.ox.ac.uk/~nicholls/bs1a/data/cigarettes.txt'
> cig<-read.table(loc,header=T) #load the data from a file
> names(cig) #inspect the data
[1] "Brand"    "Tar"      "Nicotine" "Weight"   "CO"
> dim(cig) #number of observations by number of variables
[1] 25  5
> head(cig)
          Brand  Tar Nicotine Weight   CO
1        Alpine 14.1     0.86 0.9853 13.6
2 Benson&Hedges 16.0     1.06 1.0938 16.6
3    BullDurham 29.8     2.03 1.1650 23.5
4   CamelLights  8.0     0.67 0.9280 10.2
5       Carlton  4.1     0.40 0.9462  5.4
6  Chesterfield 15.0     1.04 0.8885 15.0
> pairs(cig[,c("Tar","Nicotine","Weight","CO")])
```

The following partial R-output gives the fitted MLE parameter values for this model.

```
> cig.lm<-lm(CO~Nicotine+Tar+Weight,data=cig) #ignoring brand for now
> summary(cig.lm)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.2022     3.4618   0.925 0.365464
Nicotine     -2.6317     3.9006  -0.675 0.507234
Tar           0.9626     0.2422   3.974 0.000692 ***
Weight       -0.1305     3.8853  -0.034 0.973527
...
```

The columns give estimates ($\hat{\beta}$), standard errors (*ie* estimates of var($\hat{\beta}$)), t-values for the test that the parameter is zero (Estimate/Std. Error) and p values for those test statistics. Notice that the parameter for the variable `Nicotine` is negative. Look at the pairs plot. How does this make sense? The problem is that tar and nicotine are correlated. Tar is explaining the bulk of the variation in `CO` making the contribution from nicotine hard to interpret. Which variables are explanatory? Is there a minimal set? My guess would be that Tar gives rise to CO and Tar separately predicts Nicotine so Nicotine is only indirectly linked to CO. We will return to this sort of problem later in the course.

For the linear model theory to go through, the response must be a linear function of the parameters β. It need not be a linear function of the explanatory variables. Also, we may find that some function of the response is a linear function of the explanatory variables.

Example 2.2. Consider the `trees` data. What variables, and what functions of those variables are important on physical grounds? A lattice plot of the logged trees data is shown in Figure 2.

```
> data(trees)    #bring the data into the workspace
> names(trees)   #what are the variable names?
[1] "Girth"  "Height" "Volume"
> dim(trees)     #n=31 observations
[1] 31  3
> pairs(log(trees),main='logged trees data')   #Make the lattice plot
```

If v is the volume, and h and g are the height and girth, a natural model on physical grounds would be

$$v = \eta h^{1+\beta_2} g^{2+\beta_3} \gamma$$

with η a fixed constant and γ varying randomly about one. The idea of using a multiplying error γ here is that large volume trees have a higher volume variance than lower volume trees. We will investigate the linear model

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

with $y = \log(v/hg^2)$, $\beta_1 = \log(\eta)$, $x_2 = \log(h)$, $x_3 = \log(g)$ and $\epsilon = \log(\gamma)$. In the lattice plot of Figure 2, the logged data has little curvature, skew or uneven distribution of X-values (bunching by height or girth) so a linear model seems acceptable.

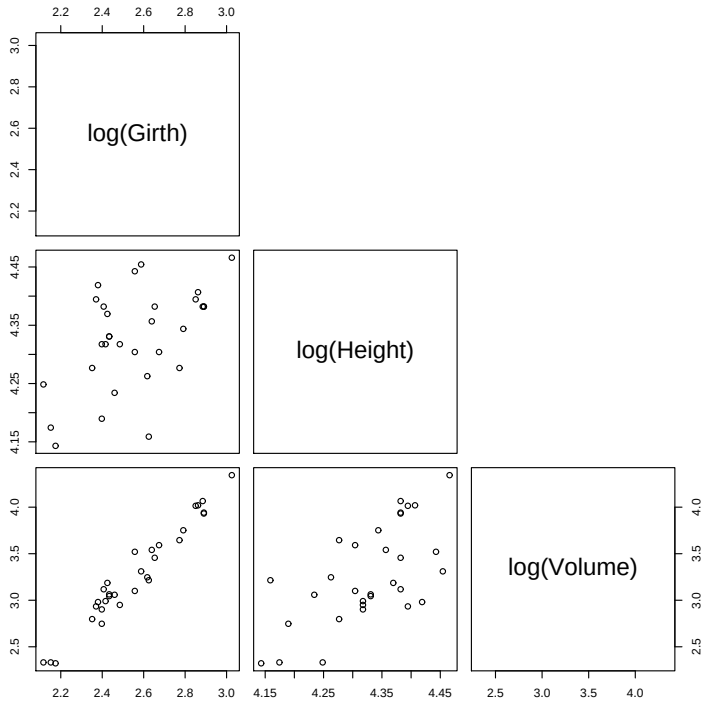

FIGURE 2. Lattice plot of the of the logged response (volume) and the logged explanatory variables (girth and height) for the 31 observations in the `trees` data.

The model we are fitting has $n = 31$ observations and $p = 3$ parameters. Fitting the model we obtain

```
> trees.lm1<-lm(log(Volume/(Height*Girth^2))~1+log(Height)+log(Girth),data=trees)
> names(trees.lm1)
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
> summary(trees.lm1)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.63162    0.79979  -8.292 5.06e-09 ***
log(Height)  0.11712    0.20444   0.573    0.571
log(Girth)  -0.01735    0.07501  -0.231    0.819
...
```

Reading off the estimated parameters, $\beta = (\log(\eta), \beta_2, \beta_3)$, so $\hat{\eta} = \exp(-6.6)$ *etc*, we arrive at the model

$$v = \exp(-6.6) h^{1.12} g^{1.98} \gamma,$$

with $\log(\gamma) \sim \mathrm{N}(0, 0.08^2)$.

2.2. **Estimators.** Given a normal linear model with data y and an $n \times p$ design X, the log-likelihood for β is

$$\ell(\beta, \sigma^2; y) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i\beta)^2.$$

Let $\mathrm{RSS}(\beta) = (y - X\beta)^T(y - X\beta)$ denote the residual sum of squares. The maximum likelihood estimator $\hat{\beta}$ for β minimises the RSS, for any fixed σ. Denote by

$$\mathrm{col}(X) = \{z \in R^n : z = X\beta, \beta \in R^p\}$$

the column span of X. We suppose to begin with that X is rank p, so $\mathrm{col}(X)$ is a p-dimensional linear subspace of R^n. Since $\hat{\beta}$ minimises the RSS, $X\hat{\beta}$ is that point $\hat{y}$ in $\mathrm{col}(X)$ lying closest to y. The point $\hat{y} = X\hat{\beta}$ therefore lies at the orthogonal projection of y into $\mathrm{col}(X)$. Since $y - \hat{y}$ is orthogonal to all vectors in $\mathrm{col}(X)$, we have the p normal equations

$$X^T(y - X\hat{\beta}) = 0$$

which fix the values of the p parameters β. Now if X has p linearly independent columns then the $p \times p$ matrix X^TX is rank p and invertible. It follows that

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$

gives the MLE in terms of the design matrix and observations. This is also the least-squares estimator for β, since it minimizes the RSS.

We will shortly derive an unbiased estimator for the error variance σ^2. However, recall that when we make likelihood ratio tests we substitute parameter MLEs into the likelihood, and it is for this reason that we will later need the MLE for σ^2 and the value of the maximized log-likelihood. Since $\hat{\beta}$ is the MLE for all σ^2, the MLE $\hat{\sigma}^2_{\mathrm{MLE}}$ for the error variance maximises

$$\ell(\hat{\beta}, \sigma^2; y) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y - X\hat{\beta})^T(y - X\hat{\beta})$$

so the MLE is

$$\hat{\sigma}^2_{\mathrm{MLE}} = \frac{\mathrm{RSS}}{n}.$$

This is a biased estimator (RSS means $\mathrm{RSS}(\hat{\beta})$ from here on). The value of the log-likelihood at the joint MLE, is

$$\ell(\hat{\beta}, \sigma^2_{\mathrm{MLE}}; y) = -\frac{n}{2}\log(\mathrm{RSS}/n) - n/2$$

since the two factors of $(y - X\hat{\beta})^T(y - X\hat{\beta})$ cancel.

2.3. **Properties of Estimators.** Let $\hat{y} = X\hat{\beta}$ give the estimated response. Define the $n \times n$ hat matrix H

$$H = X(X^TX)^{-1}X^T$$

so that $\hat{y} = Hy$. The hat matrix is a projection operator, projecting y into the column space of X, so $H = HH$ and H is symmetric. Define the vector of residuals, $e = y - \hat{y}$. The residual sum of squares, is the squared norm of the residuals, $\mathrm{RSS}(\hat{\beta}) = e^Te$. Under the normal linear model, the residuals e and the estimated response $\hat{y}$ are actually independent. We begin by showing that they are uncorrelated. [END L1 2010]

Exercise : Show that $e^Te + \hat{y}^T\hat{y} = y^Ty$ ($y, \hat{y}$ and e form a right-angle triangle).

Exercise : Show that H has p eigenvalues equal one and $n-p$ equal zero.

For generic random vectors $U = (U_1, ..., U_p)^T$ and $W = (W_1, ..., W_n)^T$ with means $\mu_U = \mathsf{E}(U)$ and $\mu_W = \mathsf{E}(W)$ denote by

$$\mathrm{cov}(U, W) = \mathsf{E}((U - \mu_U)(W - \mu_W)^T)$$

the $p \times n$ covariance matrix with entries $\mathrm{cov}(U, W)_{i,j} = \mathrm{cov}(U_i, W_j)$. Notice that $(U - \mu_U)(W - \mu_W)^T$ is an outer product.

Exercise Let C, C' be constant (*ie* non-random) p, n-component vectors. Show that $\mathrm{cov}(U + C, W + C') = \mathrm{cov}(U, W)$.

The variance matrix

$$\mathrm{var}(U) = \mathsf{E}((U - \mu_U)(U - \mu_U)^T)$$

is a symmetric $p \times p$ matrix with entries $\mathrm{var}(U)_{i,j} = \mathrm{var}(U_i, U_j)$. The variance matrix of Y is $\mathrm{var}(Y) = \mathrm{var}(X\beta + \epsilon)$, $X\beta$ is a constant and $\mathrm{var}(\epsilon) = \sigma^2 I_n$ so $\mathrm{var}(Y) = \sigma^2 I_n$.

Exercise Let L and M be matrices of suitable dimension. Show that $\mathrm{cov}(LU, MW) = L\mathrm{cov}(U, W)M^T$ and $\mathrm{var}(LU) = L\mathrm{var}(U)L^T$.

Under the normal linear model, the estimated responses $\hat{y}$ and residuals e are uncorrelated: $\mathrm{cov}(\hat{Y}, Y - \hat{Y}) = \mathrm{cov}(HY, (I_n - H)Y)$ and the RHS there is

$$H\mathrm{cov}(Y, Y)(I_n - H)^T = \sigma^2 I_n(H - HH^T)$$

which is zero, that is, $\mathrm{cov}(\hat{Y}, Y - \hat{Y}) = 0_{n,n}$ where $0_{n,n}$ is an $n \times n$ matrix of zeros. In fact, as we will see, $\hat{Y}$ and e are independent. This is sometimes asserted on the basis that $\hat{Y}$ and e are normal, with zero covariance, so they are independent. Beware: this kind of argument works when the two quantities are *jointly* normal. It is easy to see that $\hat{Y}$ and e are not jointly normal (the covariance matrix for each is singular). The conclusion is nevertheless correct in this case.

We next compute the distribution of our parameter estimate $\hat{\beta}$. The MLE $\hat{\beta} = (X^T X)^{-1} X^T Y$ is a linear combination of the normal random variables $Y = (Y_1, ..., Y_n)$. In general, if $W \sim N(\mu_W, \Sigma)$, so that W is an n-component multivariate normal (MVN) random vector (r.v.) with positive definite $n \times n$ variance matrix $\mathrm{var}(W) = \Sigma$, and L is a $p \times n$ matrix with $p \leq n$ linearly independent rows, then the p-component r.v. $LW \sim N(L\mu_W, L\Sigma L^T)$. The conditions are there to ensure $L\Sigma L^T$ is positive definite, and hence invertible.

Exercise Show that $\mathsf{E}(\hat{\beta}) = \beta$ and $\mathrm{var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$, so that

$$\hat{\beta} \sim \mathrm{N}(\beta, \sigma^2 (X^T X)^{-1}).$$

Ans: This is the generic case with $L = (X^T X)^{-1} X^T$ and $W = Y$, so $\mathrm{var}(\hat{\beta}) = (X^T X)^{-1} X^T \mathrm{var}(Y)((X^T X)^{-1} X^T)^T$, or $\mathrm{var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1}$ which is $\mathrm{var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$. The result for the MVN distribution follows from the result for LW above.

Let us now show that e and $\hat{Y}$ are independent. It follows that $\hat{\beta}$ and RSS are independent, since the estimator $\hat{\beta} = (X^T X)^{-1} X^T \hat{Y}$ and $\mathrm{RSS} = e^T e$. We use the independence properties when we construct test statistics in the next section.

Let $e_1, ..., e_p$ be a fixed orthonormal basis for the column space of X. Extend this basis to $e_1, ..., e_p, e_{p+1}, ..., e_n$, an orthonormal basis for R^n. The vectors $e_{p+1}, ..., e_n$ are orthogonal to the column vectors of X. Expand the n-component random vector Y in this basis,

$$Y = Z_1 e_1 + ... + Z_n e_n.$$

We should think of Z_i as a random function of Y, with $Z_i = e_i^T Y$. Now $He_i = e_i$ for $i = 1, 2, ..., p$ since these vectors are in col(X). On the other hand $He_i = 0$ for $i = p+1, ..., n$ (since $X^T e_i = 0$ for $i > p$). It follows that

$$\hat{Y} = Z_1 e_1 + ... + Z_p e_p,$$

since $HY = H(Z_1 e_1 + ... + Z_p e_p)$. Now $e = Y - \hat{Y}$ so

$$e = Z_{p+1} e_{p+1} + ... + Z_n e_n.$$

The weights $Z_i = e_i^T Y$ are distributed $Z_i \sim \mathrm{N}(e_i^T \mathsf{E}(Y), \sigma^2)$. For $i = p+1, p+2, ..., n$, they are mean zero, since $\mathsf{E}(Y) = X\beta$, and $e_i^T X = 0_{1,p}$. They are uncorrelated, since $\mathrm{cov}(Z_i, Z_j) = e_i^T \mathrm{cov}(Y, Y) e_j$ which is zero for $i \neq j$, since $\mathrm{var}(Y) \propto I_n$. Taking $i = j$ we have $\mathrm{var}(Z_i) = \sigma^2$. Since they are also jointly normal, they are independent.

The estimated response $\hat{Y}$ and the residuals $e = Y - \hat{Y}$ are functions of the two non-overlapping sets, $(Z_1, ..., Z_p)$ and $(Z_{p+1}, ..., Z_n)$, of mutually independent random variables, so $\hat{Y}$ and e are independent under the normal linear model.

We can now read off the distribution of RSS, and get an unbiased estimator for σ. Since $\mathrm{RSS} = e^T e$,

$$\mathrm{RSS} = Z_{p+1}^2 + ... + Z_n^2.$$

Since $Z_i/\sigma \sim \mathrm{N}(0, 1)$ for $i = p+1, p+2, ..., n$, and $\mathrm{RSS}/\sigma^2 = (Z_{p+1}/\sigma)^2 + ... + (Z_n/\sigma)^2$, with $(Z_i/\sigma)^2 \sim \chi^2(1)$ mutually independent rv each having a chi-squared distribution with one degree of freedom, it follows that

$$\mathrm{RSS}/\sigma^2 \sim \chi^2(n-p),$$

under $H0$. Now if $A \sim \chi^2(r)$ then $\mathsf{E}(A) = r$ so $\mathsf{E}(\mathrm{RSS}/\sigma^2) = n - p$ and

$$s^2 = \frac{\mathrm{RSS}}{n-p}$$

is an unbiased estimator for σ^2. It follows that $\hat{\sigma}^2_{MLE} = \frac{\mathrm{RSS}}{n}$ is biased (but it is also asymptotically unbiased, as it is a MLE).

2.4. **Tests.** We would like now to consider a collection of tests on the parameters of a normal linear regression. We would like to test for the significance of a parameter, of a group of parameters, test for parameters to be equal, or greater than one another, and for properties of linear combinations of parameters.

The test for significance of a single parameter, we know. Suppose we want to test $H0 : \beta_k = 0$ against $H1 : \beta_k \neq 0$ for some particular k from 1 to p. Under $H0$,

$$\frac{\hat{\beta}_k}{\sqrt{\sigma^2 (X^T X)^{-1}_{k,k}}} \sim N(0, 1)$$

and $\mathrm{RSS}/\sigma^2 \sim \chi^2(n-p)$, so writing $s^2 = \mathrm{RSS}/(n-p)$,

$$\begin{aligned} t &= \frac{\hat{\beta}_k}{\sqrt{\sigma^2 (X^T X)^{-1}_{k,k}}} \times \sqrt{\frac{\sigma^2(n-p)}{\mathrm{RSS}}} \\ &= \frac{\hat{\beta}_k}{s\sqrt{(X^T X)^{-1}_{k,k}}} \end{aligned}$$

is a suitably scaled ratio of independent standard normal and χ^2 random variates. Under the null, t is a realisation of a Student's-t distributed random variable T,

$$T \sim t(n-p).$$

The p-value for a two sided test is $2(1 - \Pr(T < |t|))$.

When we test for the significance of a group of parameters we use a test called an F-test. If there is just one parameter i the group, then the F-test reduces to the T-test we just described. Suppose we have a conjecture that there is no linear relation between the response y and the last k explanatory variables, $x_{p-k+1}, ..., x_p$. We want to test $H0 : \beta_{p-k+1} = 0, \beta_{p-k+2} = 0, ..., \beta_p = 0$ against $H1$: at least one of the last k parameters is non-zero. Under $H0$, with $\beta = \beta^{(0)}$ say, we are fitting the normal linear model

$$y = \sum_{i=1}^{p-k} \beta_i^{(0)} x_i + \epsilon.$$

Let $\tilde{X}$ be a matrix made up of the first $p-k$ columns of X. Under $H0$, $y = \tilde{X}\beta^{(0)} + \epsilon$ with $\beta^{(0)}$ a $(p-k)\times 1$ vector, $\beta^{(0)} = (\beta_1, ..., \beta_{p-k})$. When we fit this model we get $\hat{\beta}^{(0)} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T y$. Let $H^{(0)} = \tilde{X}(\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T$ be the hat matrix for the linear model with design matrix $\tilde{X}$. Let $\hat{Y}^{(0)} = H^{(0)}Y$ and

$$\mathrm{RSS}^{(0)} = (Y - \hat{Y}^{(0)})^T (Y - \hat{Y}^{(0)}).$$

Under $H0$, the MLE for σ^2 is

$$\hat{\sigma}^2_{MLE,0} = \frac{\mathrm{RSS}^{(0)}}{n}.$$

The dimension of parameter space under the null is $p-k+1$ (ie, $\beta_1, ..., \beta_{p-k}, \sigma^2$).

Under $H1$, with $\beta = \beta^{(1)}$, we are fitting the normal linear model

$$y = \sum_{i=1}^{p} \beta_i^{(1)} x_i + \epsilon.$$

This is the usual setup, with $Y = X\beta^{(1)} + \epsilon$, $\hat{\beta}^{(1)} = (X^T X)^{-1} X^T y$, $\hat{Y}^{(1)} = HY$,

$$\mathrm{RSS}^{(1)} = (Y - \hat{Y}^{(1)})^T (Y - \hat{Y}^{(1)}),$$

and

$$\hat{\sigma}^2_{MLE,1} = \frac{\mathrm{RSS}^{(1)}}{n}.$$

The dimension of parameter space under the alternative is $p+1$.

We can now give the Likelihood Ratio Test (LRT) statistic Λ for $H0$. Substituting the local values into the expression for $\ell(\hat{\beta}, \hat{\sigma}^2_{MLE}; y)$, which we gave at the

end of Section 2.2, we get

$$\begin{aligned}\Lambda(Y) &= -2(\ell(\hat{\beta}^{(0)}, \hat{\sigma}^2_{MLE,0}; Y) - \ell(\hat{\beta}^{(1)}, \hat{\sigma}^2_{MLE,1}; Y)) \\ &= n\log(\mathrm{RSS}^{(0)}) - n\log(\mathrm{RSS}^{(1)}).\end{aligned}$$

We reject $H0$ when the likelihood ratio statistic Λ falls in the critical region. We know from the Neyman-Pearson theorem that this region has the form $C_1 = \{y : \Lambda(y) > C\}$. Asymptotically in n, Λ has as χ^2 distribution with k degrees of freedom, and this would give an approximate test for $H0$. However, we will see that the LRT statistic is a strictly increasing function of another statistic, $F(y)$, for which we possess an exact distribution. This leads to an exact test with the same critical region as the LRT: if C' is chosen so that, under $H0$, $F(Y) > C'$ with probability $1-\alpha$, and C is chosen so that $\Lambda(Y) > C$ with probability $1-\alpha$, then $F(Y) > C'$ if and only if $\Lambda(Y) > C$ (imagine sorting the states y by $F(y)$ and by $\Lambda(y)$ - you get the same order so the threshold is set at the same states).

Consider the F-statistic,

$$F(y) = \frac{(\mathrm{RSS}^{(0)} - \mathrm{RSS}^{(1)})/k}{\mathrm{RSS}^{(1)}/(n-p)}.$$

The corresponding random variable $F(Y)$ has a $F(k, n-p)$ distribution under the null hypothesis in which the last k parameters are zero. The F-distribution is new to us. If $A \sim \chi^2(a)$ and $B \sim \chi^2(b)$ are two independent χ^2 r.v.'s with a and b degrees of freedom respectively, then the new r.v.

$$F = \frac{(A/a)}{(B/b)}$$

has a $F(a,b)$-distribution on $F > 0$. This property defines the distribution. The mean of $F \sim F(a,b)$ is $b/(b-2)$ so, under the null, the mean of $F \sim F(k, n-p)$ is $(n-p)/(n-p-2)$ for $n-p > 2$, or about 1 for $n \gg p$. The quantiles of an F-distribution with k numerator and $n-p$ denominator degrees of freedom are known. Let $F_{1-\alpha}(k, n-p)$ be the $1-\alpha$ quantile of $F(k, n-p)$. We reject $H0$ at significance level α if $F(y) > F_{1-\alpha}(k, n-p)$.

How do we know F has the properties we claim for it? First, it is easy to check that

$$\Lambda = n\log\left(1 + \frac{k}{(n-p)}F\right)$$

with $n > p > k > 0$, so Λ is a strictly increasing function of F, and this test based on F is indeed a LRT. Secondly,

$$\frac{\mathrm{RSS}^{(1)}}{\sigma^2} \sim \chi^2(n-p)$$

and

$$\frac{\mathrm{RSS}^{(0)} - \mathrm{RSS}^{(1)}}{\sigma^2} \sim \chi^2(k).$$

The former we know, and the latter we demonstrate shortly. Thirdly, $\mathrm{RSS}^{(1)}$ and $\mathrm{RSS}^{(0)} - \mathrm{RSS}^{(1)}$ are independent, again, something we will verify. Finally, since F is the ratio of suitably scaled and independent $\chi^2(n-p)$ and $\chi^2(k)$ r.v. it follows that F has an $F(k, n-p)$ distribution, by the definition of this distribution given above.

What is the intuition here? The question which a LRT answers is, is there evidence that the data fits the $H1$ model $Y = X\beta^{(1)} + \epsilon$ better than the $H0$ model $Y = \tilde{X}\beta^{(0)} + \epsilon$? Tests which look for significant changes in the RSS, such as the F-test above, are called ANOVA, short for *Analysis of Variance*. When we fit the more complex model, $H1$ above, there will be a reduction in the residual sum of squares compared to the residual sum of squares we get when we fit the simpler model $H0$. If $H0$ is good, then the fractional improvement $(\mathrm{RSS}^{(0)} - \mathrm{RSS}^{(1)})/\mathrm{RSS}^{(1)}$ in the RSS is slight. However, if we add lots of explanatory variables in $H1$, so that $\dim(\mathrm{col}(X))$ approaches n, then we will see a big drop in the RSS, towards zero. In order to account for this, the ratio $(\mathrm{RSS}^{(0)} - \mathrm{RSS}^{(1)})/\mathrm{RSS}^{(1)}$ must be weighted by the fractional change $(n-p)/k$ in the number of degrees of freedom. Now large values of F are a sign that that the added parameters in $H1$ are reducing the estimated variance by an amount which is too great to be put down to chance.

We need to verify the second and third properties above. We are interested in the distribution of F under $H0$, where $\mathsf{E}(Y) = \tilde{X}\beta^{(0)}$. We will modify the expansion

$$Y = Z_1 e_1 + ... + Z_n e_n.$$

As before, $e_1, ..., e_n$ are an orthonormal basis for R^n, and $e_1, ..., e_p$ are an orthonormal basis for $\mathrm{col}(X)$. We can choose this basis so that there is a first group $e_1, ...e_{p-k}$ of vectors spanning the first $p-k$ columns of X and a second group $e_{p-k+1}, ..., e_p$ completing the basis $e_1, ..., e_p$ for $\mathrm{col}(X)$ (and notice $e_{p-k+1}, ..., e_p$ are not a basis for $\mathrm{span}(X_{p-k+1}, ..., X_p)$ unless $\mathrm{span}(X_{p-k+1}, ..., X_p) \perp \mathrm{span}(X_1, ..., X_{p-k})$). Since $\hat{Y}^{(0)} = H^{(0)}Y$ $H^{(0)}$ projects into space spanned by the first $p-k$ columns of X, we must have $He_j = 0$ for $j > p-k$, so

$$\hat{Y}^{(0)} = Z_1 e_1 + ... + Z_{p-k} e_{p-k}$$

Similarly, $\hat{Y}^{(1)} = HY$ with $He_j = 0$ for $j > p$, so

$$\hat{Y}^{(1)} = Z_1 e_1 + ... + Z_p e_p.$$

Now $Y - \hat{Y}^{(1)} = Z_{p+1} e_{p+1} + ... + Z_n e_n$ so

$$\mathrm{RSS}^{(1)} = Z_{p+1}^2 + ... + Z_n^2.$$

Similarly, $\mathrm{RSS}^{(0)} = Z_{p-k+1}^2 + ... + Z_n^2$. It follows that

$$\mathrm{RSS}^{(0)} - \mathrm{RSS}^{(1)} = Z_{p-k+1}^2 + ... + Z_p^2.$$

The weights Z_i $i = 1, 2, ..., n$ are mutually independent, since they are jointly normal, with zero covariance, exactly as before. They are distributed as $Z_i \sim \mathrm{N}(e_i^T \mathsf{E}(Y), \sigma^2)$ with $E(Y) = [X_1, ...X_{p-k}]\beta$ so $Z_i \sim \mathrm{N}(0, \sigma^2)$ for $i = p-k+1, ..., n$ (*ie*, not just $i = p+1, ..., n$ as before).

We can see that $\mathrm{RSS}^{(0)} - \mathrm{RSS}^{(1)}$ and $\mathrm{RSS}^{(1)}$ are independent rv, since they are functions of disjoint sets of independent rv. Also, since there are k terms in the last sum above, $(\mathrm{RSS}^{(0)} - \mathrm{RSS}^{(1)})/\sigma^2$ has a $\chi^2(k)$ distribution and we are done demonstrating the second and third properties.

We often test for two parameters β_1 and β_2 to be equal, so $H0 : \beta_1 - \beta_2 = 0$. The MLE is $\hat{\beta}_1 - \hat{\beta}_2$ with variance

$$\begin{aligned} \mathrm{var}(\hat{\beta}_1 - \hat{\beta}_2) &= \mathrm{var}(\hat{\beta}_1) + \mathrm{var}(\hat{\beta}_2) - 2\mathrm{cov}(\hat{\beta}_1, \hat{\beta}_2) \\ &= \sigma^2 (X^T X)^{-1}_{1,1} + \sigma^2 (X^T X)^{-1}_{2,2} - 2\sigma^2 (X^T X)^{-1}_{1,2} \end{aligned}$$

so the test statistic is

$$\frac{\hat{\beta}_1 - \hat{\beta}_2}{s\sqrt{(X^TX)^{-1}_{1,1} + (X^TX)^{-1}_{2,2} - 2(X^TX)^{-1}_{1,2}}} \sim t(n-p)$$

This works for linear combinations of parameters. If v is $p \times 1$ and we want to test for $v^T\beta = 0$ then the MLE is $v^T\hat{\beta}$, and

$$\begin{aligned} \text{var}(v^T\hat{\beta}) &= v^T \text{var}(\hat{\beta}) v \\ &= \sigma^2 (v^T (X^TX)^{-1} v), \end{aligned}$$

so the test statistic is

$$\frac{v^T\hat{\beta}}{s\sqrt{(v^T(X^TX)^{-1}v)}} \sim t(n-p).$$

The quantities s^2 and $v\hat{\beta}$ are independent, since s^2 and $\hat{\beta}$ are independent.

Exercise verify that $v = (1, -1, 0, ..., 0)^T$ gives the test for $\hat{\beta}_1 - \hat{\beta}_2 = 0$.

There are some shortcuts. For example, in a test for $\beta_1 = \beta_2$ the reduced model, with $\beta_1' = \beta_1 = \beta_2$ is

$$y = \beta_1'(x_1 + x_2) + \beta_3 x_3 + ...$$

and the full model, $Y = X\beta + \epsilon$, can be written

$$y = \beta_1'(x_1 + x_2) + \beta_2'(x_1 - x_2) + \beta_3 x_3 + ...$$

so the test for $\beta_1 = \beta_2$ can be framed as a test $\beta_2' = 0$. We can run a T or F test to drop β_2', with design matrix $X = [X_1 + X_2, X_1 - X_2, X_3, ..., X_p]$ and parameter vector $\beta = (\beta_1', \beta_2', \beta_3, ..., \beta_p)$.

2.5. **ANOVA, and an example.** Because ANOVA tests are used so frequently, the important numbers in the test are laid out in a standard way, to facilitate reading. There is a little variation (the default R table doesn't exactly follow my rules), but just a little.

The ANOVA table sets out the numbers we need to make tests for dropping certain collections of variables. Suppose the variables $x_1, ..., x_p$ come in $m = 3$ groups (typically just 2 or 3 groups) and are ordered so that the groups are

$$\{1\}, \{2, 3...p-k\}, \{p-k+1, p-k+2, ..., p\}.$$

This splits off the variables x_{p-k+1} to x_p. This is the variable grouping relevant for the hypothesis $H0 : \beta_{p-k+1} = \beta_{p-k+2} = ... = \beta_p = 0$, which we test with an F-test.

A typical ANOVA table gives fitting information for each of the models starting from a simplest model with just intercept β_1, adding the groups one at a time, up to the model with all p variables. For the $m = 3$-group case, testing $H0 : \beta_{p-k+1} = \beta_{p-k+2} = ... = \beta_p = 0$, the model sequence is

$$\begin{aligned} y &= \beta_1 + \epsilon, \\ y &= \beta_1 + \beta_2 x_2 + ... + \beta_{p-k} x_{p-k} + \epsilon, \\ y &= \beta_1 + \beta_2 x_2 + ... + \beta_{p-k} x_{p-k} + ... + \beta_p x_p + \epsilon, \end{aligned}$$

If we order the variables in the right way, we can sometimes do model selection at a glance, as we read down the table from top to bottom. If $X_{1:i} = [X_1, X_2, ..., X_i]$,

Terms added	Degrees Freedom	Reduction in RSS	Mean Square	F statistic
$X_{2:(p-k)}$	$p-k-1$	$\text{TSS}-\text{RSS}_{1:(p-k)}$	$\frac{\text{TSS}-\text{RSS}_{1:(p-k)}}{p-k-1}$	$\frac{(\text{TSS}-\text{RSS}_{1:(p-k)})/(p-k-1)}{\text{RSS}_{1:p}/(n-p)}$
$X_{(p-k+1):p}$	k	$\text{RSS}_{1:(p-k)}-\text{RSS}_{1:p}$	$\frac{\text{RSS}_{1:(p-k)}-\text{RSS}_{1:p}}{k}$	$\frac{(\text{RSS}_{1:(p-k)}-\text{RSS}_{1:p})/k}{\text{RSS}_{1:p}/(n-p)}$
Residual	$n-p$	$\text{RSS}_{1:p}$	$\frac{\text{RSS}_{1:p}}{n-p}$	

TABLE 1. ANOVA table for the groups of variables $\{x_2,...,x_{p-k}\}$ and $\{x_{p-k+1},...,x_p\}$ added incrementally to the intercept group $\{x_1\}$. In some tables a final column giving the p-value is included.

then the design matrices build from X_1 to $X = X_{1:p}$. Let $\text{RSS}_{1:i}$ be the residual sum of squares for the fit with design matrix $X_{1:i}$. The decrease in the residual sum of squares when we add the variables $x_{i+1},...,x_{i+k}$ to a model that already has the variables $x_1, x_2, ..., x_i$ is $\text{RSS}_{1:i} - \text{RSS}_{1:(i+k)}$. The number of residual degrees of freedom in the fit for the model with design matrix $X_{1:i}$ is $n-i$ (assuming the columns of $X_{1:i}$ are linearly independent).

The layout of an ANOVA table for the three groups $\{1\}, \{2,...,p-k\}, \{p-k+1,...,p\}$ is shown in Table 1. $TSS = (y-\bar{y})^T(y-\bar{y})$ is the residual sum of squares for a model with just intercept, in other words, the total sum of squares adjusted for intercept. $\text{RSS}_{1:p}$ is the residual sum of squares for the full model.

The F-statistic in row two of Table 1 is the F-test statistic for the test to add the variables $\{x_{p-k+1},...,x_p\}$ to a model with variables $\{x_1,..,x_{p-k}\}$, which is the test we set up in Section 2.4.

The F-statistic in row one of Table 1 is an F-test statistic for the test to add the variables $\{x_2,...,x_{p-k}\}$ to a model with just x_1, the intercept variable. It might seem natural to use the divisor $\text{RSS}_{1:(p-k)}/(n-(p-k))$, for an F with $p-k-1$ numerator and $n-(p-k)$ denominator degrees of freedom. However, (i) the divisor $\text{RSS}_{1:p}/(n-p)$ is "just as good" as $\text{RSS}_{1:(p-k)}/(n-(p-k))$, since it too is independent of $\text{TSS}-\text{RSS}_{1:(p-k)}$, so we can see $(\text{TSS}-\text{RSS}_{1:(p-k)})(n-p)/\text{RSS}_{1:p)}(p-k-1)$ has an $F(p-k-1,n-p)$ distribution under the null, and (ii) it is better, as $\text{RSS}_{1:p)}/(n-p)$ is an estimate of σ^2 which is not biased if the variables $x_{p-k+1},...x_p$ added in the row below turn out to be explanatory. You might possibly add (iii) the divisor $\text{RSS}_{1:p}/(n-p)$ has a higher variance than $\text{RSS}_{1:(p-k)}/(n-(p-k))$, if variables in the rows below really were not related to the response, so in that case we would do better to drop them from the ANOVA. That is equivalent to using the $\text{RSS}_{1:(p-k)}/(n-(p-k))$ divisor. Another way to make point (iii) is that the $\text{RSS}_{1:(p-k)}/(n-(p-k))$ divisor is the one given by the LRT, where as $\text{RSS}_{1:p)}/(n-p)$ is just some statistic with a distribution we happen to know under the null. On balance item (ii) controls our choice of test statistic, so the table opts for higher variance in return for lower bias.

Table 1 is the table we might set out if we were carrying out the F-test for $H0 : \beta_{p-k+1} = \beta_p - k + 2 = ... = \beta_p = 0$ against $H1 : \beta \in R^p$, though we could omit the first row.

Other relevant test statistics can be computed from the numbers in an ANOVA table like Table 1. For example, the test for the model hypothesis $H0' : \beta_2 = ... = \beta_p = 0$ against the full model $H1 : \beta \in R^p$, is the test for no linear relation to the variables $x_2, ...x_p$. The test statistic is

$$F' = \frac{\text{TSS} - \text{RSS}_{1:p}}{p-1} \times \frac{n-p}{\text{RSS}_{1:p}},$$

since $k = p - 1$ here. The second factor is given at the bottom of the table. The first element $\text{TSS} - \text{RSS}_{1:p}$ is the sum of the terms in the third column of the table,

$$\text{TSS} - \text{RSS}_{1:p} = (\text{RSS}_{1:(p-k)} - \text{RSS}_{1:p}) + (\text{TSS} - \text{RSS}_{1:(p-k)}),$$

so we can form F' by taking appropriate sums and ratios of table elements. Note that the statistic F' replaces the widely used statistic R^2 as a measure of fit quality (or the lack of it). While R^2 runs from zero to one, we have no absolute scale for quality of fit (how close to one is acceptable), F' runs from 0 to infinity (and big F' is poor fit) and does give a direct test for significant linear dependence. Note that R outputs both F', the test statistic for no linear relation, and the p-value for this test, automatically, when we make a `summary()` of a `lm()` output. This is more useful to us than R^2, though R gives this as well. [End L3]

Again, these numbers are useful for calculating other tests. The test for no linear dependence on the variables $x_2, ..., x_p$, has F statistic

$$F = \frac{(\text{TSS} - \text{RSS}_{1:p})/(p-1)}{\text{RSS}_{1:p}/(n-p)},$$

with $(p-1)$ numerator and $(n-p)$ denominator degrees of freedom. The quantity $\text{TSS} - \text{RSS}_{1:p}$ is the sum of all the entries in the third column, bar the last,

$$\begin{aligned} \text{TSS} - \text{RSS}_{1:p} \quad = \quad & (\text{RSS}_{1:i_{m-1}} - \text{RSS}_{1:p}) + (\text{RSS}_{1:i_{m-2}} - \text{RSS}_{1:i_{m-1}}) + ... \\ & ... + (\text{RSS}_{1:i_{i_2}} - \text{RSS}_{1:i_3}) + (\text{TSS} - \text{RSS}_{1:i_2}). \end{aligned}$$

Example 2.3. Consider variable selection for the `trees` data. We looked at this in Example 2.2. When we have many variables, we can't test all possible combinations. Physical considerations (*aka* common sense) are always important, but especially so, when we are setting up the hypotheses. We considered the linear model

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_2 + \epsilon$$

with $y = \log(v/hg^2)$, $\beta_1 = \log(\eta)$, $x_2 = \log(h)$, $x_3 = \log(g)$ and $\epsilon = \log(\gamma)$ (and v, g and h the volume, girth and height).

Look at the R-output for this model in Example 2.2. The quoted *residual standard error* is $s^2 = \text{RSS}/(n-p)$: the RSS is $(Y - \hat{Y})^2$, or in R,

```
> (rss<-sum(trees.lm1$residuals^2))
[1] 0.1854634
```

so estimated error variance *aka* the Residual standard error s^2 is

```
> (sqrt(rss/28))
[1] 0.08138607
```

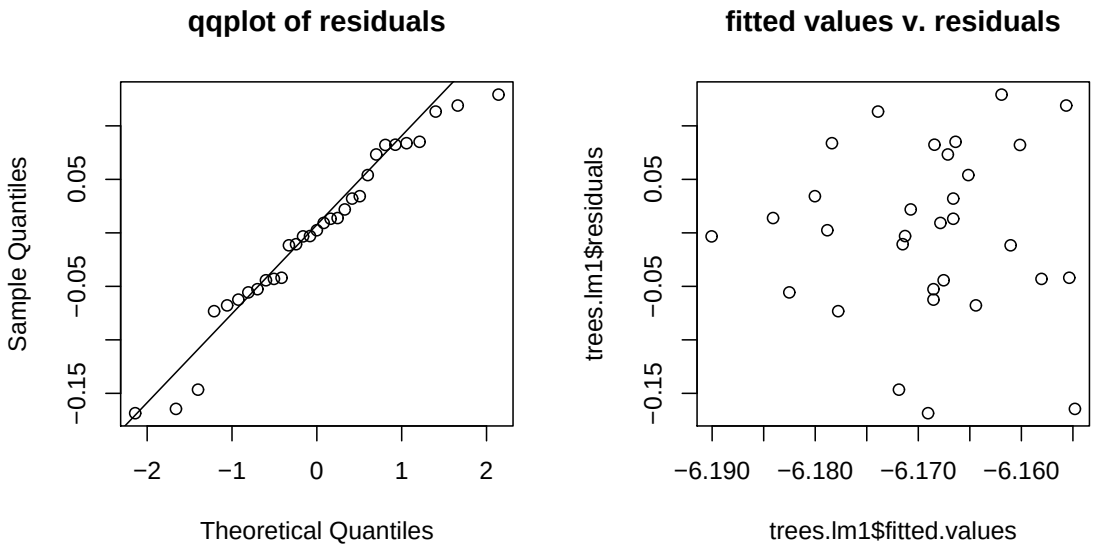

FIGURE 3. (LEFT) A qqplot of the residuals of the logged-model fit to the `trees` data looks healthy, and (RIGHT) no sign of correlation between residuals and fitted values (though a possible funnel, increasing variance with fitted value).

You might like to check the elements in the `summary(trees.lm1)` table.

Later on we look at checks on the fit. Under the model we have just estimated (which is $H1$, with the full parameter set), the residuals are given by $e = (I_n - H)\epsilon$, so that each residual e_i is normally distributed and the residuals e and fitted values $\hat{Y} = HY$ are independent. We can make a qqplot of the sample quantiles against the normal quantiles and look for normality. Also, we can plot fitted values against residuals and look for a trend. The qqplot is obviously important here, as our $\epsilon = \log(\gamma)$ is the log of a multiplicative error, so this is an area we might expect to see departures from the model. The qqplot in Figure 3 is very well behaved, except possibly in the upper tail.

We now come to the test for $\beta_2 = \beta_3 = 0$. R gives us the ANOVA tables. Here are several ways to do this. We begin by fitting the reduced, $H0$, model.

```
> trees.lm0<-lm(log(Volume/(Height*Girth^2))~1)
```

The fields of `trees.lm0` are special cases. There is just an intercept, so $\tilde{X} = 1_{n,1}$, $X^TX = n$, $(X^TX)^{-1} = 1/n$, and $(X^TX)^{-1}X^Ty = \bar{y}$.

Now we use ANOVA to see if $\beta_2 = \beta_3 = 0$. We can do this 'by hand'.

```
> (rss0<-sum(trees.lm0$residuals^2))
[1] 0.1876858
> (rss1<-sum(trees.lm1$residuals^2))
[1] 0.1854634
> k<-2
> n.minus.p<-31-3
```

```
> (F<-(rss0-rss1)*n.minus.p/(k*rss1))
[1] 0.1677617
> (p<-1-pf(F,k,n.minus.p))
[1] 0.8463989
```

The p-value, $p = 0.85$, is not significant, so the LRT supports $H0 : \beta_2 = \beta_3 = 0$ (that is, there is no evidence for dependence of y on x_2 and x_3).

We can get R to form something like a regular ANOVA table, Table **??**. The default R ANOVA adds one variable at a time, so the groups of variables are $\{x_1\}, \{x_2\}\{x_3\}$ (since F-tests for many other grouping can be computed form this 'finest resolution' table). Because the change in the number of degrees of freedom is one at each row, columns three and four of Table **??** are identical.

```
> anova(trees.lm1)
Analysis of Variance Table

Response: log(Volume/(Height * Girth^2))
            Df   Sum Sq  Mean Sq F value Pr(>F)
log(Height)  1 0.001868 0.001868  0.2820 0.5996
log(Girth)   1 0.000354 0.000354  0.0535 0.8188
Residuals   28 0.185463 0.006624
```

In this table `Sum Sq` corresponds to "Reduction in RSS" in Table **??**. Thus the residual sum of squares $\mathrm{RSS}_{1:p}$ for the full model is $\mathrm{RSS}_{1:3} = 0.185463$, and $\mathrm{RSS}_{1:p}/(n-p) = 0.185463/28 = 0.006624$. The other quantity we need, for the test $\beta_3 = \beta_2 = 0$ is $\mathrm{TSS} - \mathrm{RSS}_{1:3}$, which is

$$\mathrm{RSS}_{1:2} - \mathrm{RSS}_{1:3} + \mathrm{TSS} - \mathrm{RSS}_{1:2} = 0.000354 + 0.001868 = 0.002222.$$

So

$$\begin{aligned} F &= \frac{(\mathrm{TSS} - \mathrm{RSS}_{1:3})/2}{\mathrm{RSS}_{1:p}/(n-p)} \\ &= 0.001111/0.006624 \simeq 0.1677 \end{aligned}$$

and the F-test proceeds as before.

The ANOVA table we got from R didn't quite have the variable grouping we wanted - we got the default grouping, which just put every explanatory variable in a separate group, so we had to do some arithmetic to get our test statistic. Alternatively, we can tell R which specific models we want to compare, and get R to form a table

```
> anova(trees.lm0,trees.lm1)
Analysis of Variance Table

Model 1: log(Volume/(Height * Girth^2)) ~ 1
Model 2: log(Volume/(Height * Girth^2)) ~ 1 + log(Height) + log(Girth)
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1     30 0.187686
2     28 0.185463  2  0.002222 0.1678 0.8464
```

which we see is not a standard ANOVA table, but is nevertheless easy enough to read. We read this $\mathrm{RSS}_{1:3} = 0.185463$, $\mathrm{TSS} = 0.187686$, $\mathrm{TSS} - \mathrm{RSS}_{1:3} = 0.002222$, with $k = 2$ for the two variables we set to zero, and if $F = 28(\mathrm{TSS} - \mathrm{RSS}_{1:3})/2\mathrm{RSS}_{1:3}$ then $F = 0.1678$. Finally, if $F_{2,28} \sim F(2, 28)$ is an F-distributed

rv with 2 numerator and 28 denominator degrees of freedom, then we read off $\Pr(F_{2,28} > F) \simeq 0.85$, so the p-value shows no evidence for a departure from $H0$. This time, there was no arithmetic needed.

We could get this directly, also, from the `summary(trees.lm1)` output above. Recall that this output automatically gives the F-statistic for the reduced model with no explanatory variables except the intercept, β_1 here. That is just the model reduction we are considering, so the final line

```
F-statistic: 0.1678 on 2 and 28 DF,  p-value: 0.8464
```

gives us the same elements, $F = 0.1678$ and $\Pr(F > f) \simeq 0.85$. This part of the output would not be relevant if we were considering dropping a smaller subset of the explanatory variables.

2.6. **Categorical variables.** So far we have treated continuous explanatory variables. However, explanatory variables may be categorical. The values taken by a categorical variable are called its *levels*, and the levels may for ordered or unordered. We will discuss unordered categorical variables.

A categorical explanatory variable $x_k^{(\mathrm{cat})} \in \{1, 2, ..., c\}$ with c levels is equivalent to c binary indicator variables $g_{k,a} = \mathbb{I}_{x_k^{(\mathrm{cat})}=a}$, with $a = 1, 2, ..., c$ the level index, so the k'th response y_k has one explanatory variable $g_{k,a}$ for each level of the original categorical variable. Suppose we want to allow the response to have a mean which depends on the level of $x_k^{(\mathrm{cat})}$, and suppose that, for the k'th response y_k, there are m other explanatory variables $x_{k,1}, x_{k,2}, ..., x_{k,m}$ including an intercept, $x_{k,1} = 1$. The model

$$y_k = \alpha + \alpha_2 g_{k,2} + ... + \alpha_b g_{k,b} + \gamma_2 x_{k,2} + ... + \gamma_m x_{k,m} + \epsilon_k$$

allows the mean of y_k to vary with the level of $x_k^{(\mathrm{cat})}$. What happened to $\alpha_1 g_{k,1}$? If we include it, then our model is over-parameterized. If the level for response k is $x_k^{(\mathrm{cat})} = 1$, then $g_{k,1} = 1$ and $g_{k,2} = ... = g_{k,c} = 0$, so

$$\mathsf{E}(Y_k) = \alpha + \gamma_2 x_{k,2} + ... + \gamma_m x_m.$$

If the level is $x_k^{(\mathrm{cat})} = a$, then $g_{k,a} = 1$ and the others are zero, so

$$\mathsf{E}(Y_k) = \alpha + \alpha_a + \gamma_2 x_{k,2} + ... + \gamma_m x_m.$$

We see that α_a is the offset in the intercept of the level-a samples relative to the intercept α of the level-1 samples. We are using level 1 as the *baseline* level.

If G_a is the binary column vector $G_a = (g_{1,a}, ..., g_{n,a})^T$ for the level-a indicator, and $X_1, X_2, ..., X_m$ are column vectors for other variables, then the design matrix for the model above is $X = (X_1, G_2, ...G_c, X_2, ..., X_m)$ (so $p = m + c - 1$ here, and columns are in no particular order). The model itself is $Y = X\beta + \epsilon$ with $\beta = (\alpha, \alpha_2, ..., \alpha_c, \gamma_2, ..., \gamma_m)$. We left out G_1 when we formed the new design matrix because X has a first column of ones, corresponding to the intercept. But then $X_1 = \sum_{a=1}^{c} G_j$ since each observation must have its categorical variable in *one* of the levels $1, 2, ..., c$, and so the columns of $(X_1, G_1, G_2, ..., G_c, X_2, ...X_m)$ are not linearly independent, and again, the model with all c columns, $G_1, ..., G_c$, is over-parameterized. The variables $G_1, ..., G_c$ are sometimes called a *dummy variables* for the level and the matrix $(G_2, ..., G_c)$ is called a *contrast* matrix.

Example 2.4. The data depicted in Figure 4 shows average Oxford house prices (in thousands of pounds) for 100 months starting April 2000 ending July 2008 for Detached, Semi-Detached and Terraced houses and Flats. The website `www.home.co.uk` displays data of this kind. The Flats and Detached properties clearly have lower and higher variance respectively, than the two other classes. We will deal with them later. We start with an analysis of the two-hundred Semi-Detached and Terraced prices. Here are some rows of data

```
> ohp[1:3,]
price            type month sales
  329   Semi-Detached   100    19
  276   Semi-Detached    99    37
  300   Semi-Detached    98    45
.
.
.
> ohp[99:102,]
price            type month sales
  148   Semi-Detached     2    75
  148   Semi-Detached     1    68
  294        Terraced   100     9
  310        Terraced    99    35
.
.
.
> ohp[198:200,]
price       type month sales
  154   Terraced     3    75
  150   Terraced     2    65
  149   Terraced     1    62
```

The prices `price` are average figures for the month, in thousands of pounds, while the number of sales `sales` gives the number of individual sale prices which were averaged to form the price reported for that month. We will leave a discussion of column four for the moment. The variable `type` is now a two-level categorical variable.

Is there any difference between the price trends for the two types of houses? The lattice plot, Figure 4, supports a linear model for `price` as a function of `month`. We have omitted the two-level categorical variable `type` in Figure 4. We will begin by fitting a model with a different offset for the two levels - we assume prices grow at the same rate, but there is an offset in the price for Terraced relative to Semi-Detached properties. Let $y_k = \texttt{price[k]}$, $x_{k,1} = 1, x_{k,M} = \texttt{month[k]}$ and $g_{k,T} = \mathbb{I}_{\texttt{type[k]=Terraced}}$, so that

$$y_k \sim \alpha + \alpha_T g_{k,T} + \gamma_M x_{k,M} + \epsilon_k, \qquad \epsilon_k \sim \mathrm{N}(0, \sigma^2).$$

In this model α is the price of a Semi-Detached house in month 0, and $\alpha + \alpha_T$ is the price of a Terraced house in month 0. In R, the intercept is included by default, so `price~month+type` and `price~1+month+type` specify the same model. Also, R will automatically construct dummy variables for the levels of a categorical variable (such as `type`, which has levels `Semi-Detached` and `Terraced`).

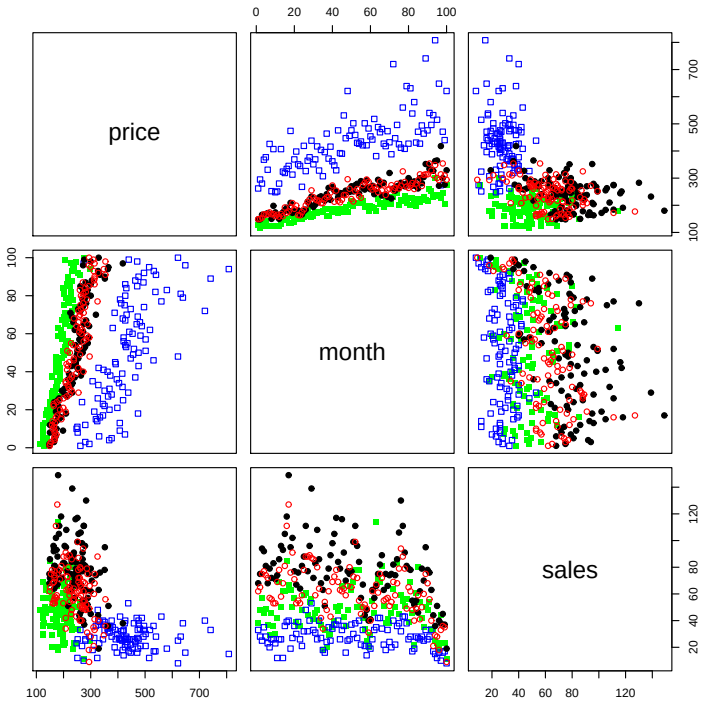

FIGURE 4. Lattice scatter plot of monthly average prices, against month and number of sales. (solid/black circle) Semi-Detached, (open/red circle) Terraced, (solid/green square) Flat, (open/blue square) Detached.

How does R code the categorical variable `type`?

```
> X<-model.matrix(price~month+type,data=ohp)
> X[1:3,]
(Intercept) month typeTerraced
          1   100            0
          1    99            0
          1    98            0
> X[99:102,]
(Intercept) month typeTerraced
          1     2            0
          1     1            0
          1   100            1
          1    99            1
> X[198:200,]
(Intercept) month typeTerraced
          1     3            1
          1     2            1
          1     1            1
```

The rightmost column of the design matrix in this R implementation is $G_T = (g_{1,T}, ..., g_{n,T})^T$. The baseline level is `Semi-Detached` and the variable mapping for the design matrix above is $(\alpha, \gamma_M, \alpha_T) = (\beta_1, \beta_2, \beta_3)$. You need to check you know which variable R is using as the baseline, though you wouldnt use `model.matrix()` to do that: the baseline level is simply the level omitted in the `summary()` output (see below). The offset in the mean for a house with `type` equal `Terraced` is β_3, the parameter for column three in `X`, the R the design matrix above.

OK, so let's fit the model `price~month+type`.

```
> ohp.lm<-lm(price~month+type,data=ohp)
> summary(ohp.lm)
...
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  158.09727    3.59152  44.020   <2e-16 ***
month          1.69887    0.05531  30.716   <2e-16 ***
typeTerraced  -2.58000    3.19306  -0.808     0.42
...
Residual standard error: 22.58 on 197 degrees of freedom
Multiple R-Squared: 0.8274,     Adjusted R-squared: 0.8256
F-statistic: 472.1 on 2 and 197 DF,  p-value: < 2.2e-16
```

We don't need an F-test to test $\alpha_T = 0$ (*ie*, $\beta_3 = 0$) here: for a single parameter, the t-test is equivalent. We have $n = 200$ data and $p = 3$. Reading off the table, $\hat{\alpha}_T = -2.58$, and the t-statistic for the test $\alpha_T = 0$,

$$\hat{\alpha}_T / s\sqrt{(X^T X)^{-1}_{33}} = -2.58/3.19$$

is equal to -0.81, with $n-p = 197$ degrees of freedom. The p-value $2(1-\Pr(T < |t|))$ is $2 * (1 - \texttt{pt}(0.81, 197)) \simeq 0.42$ (which we can read in the right column) and this shows that the Terraced/Semi-Detached distinction is not significant. The two regressions are plotted in Figure 5. There is clearly little difference. Note that this is not the same as making two regressions and using a t-test for equality of intercepts, as we are imposing (i) equal slopes, and (ii) equal error variance σ^2.

Exercise Can you see any sign of model-mispecification in Figure 5?

2.7. **Variable interactions.** We can form new explanatory variables from old by taking functions of explanatory variables. Interactions of the form $\beta_i x_i + \beta_j x_j + \beta_I x_i x_j$ are particularly common, and mean something like "variable j has more impact on the response when variable i is large" and *vis versa*. If x_i is a binary dummy variable for some level a of a categorical variable x^{cat}, then the slope with increasing x_j is β_j for observations with $x^{\text{cat}} \neq a$ (where $x_i = 0$) and the slope is $\beta_j + \beta_I$ for observations with $x^{\text{cat}} = a$ (where $x_i = 1$).

When we have interactions we often include lower order terms in the model, though we might have no 'physical' use for them. Suppose $y = \alpha + \beta x_1 x_2$ with x_1 in Celcius. If we switch to Farenheit, $x_1 = m x_1' + d$ with $m = 5/9$ and $d = 160/9$, then $y = \alpha + d\beta x_2 + m\beta x_1' x_2$. Now we have a new kind of term $d\beta x_2$. We may dislike the idea that the kinds of terms in our model (rather than just the parameter values) are dependent on the zero-location for interacting variables, and instead at least begin our modeling with $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_I x_1 x_2$. Faraway (2004) Chapter 8 page 122 has more on this.

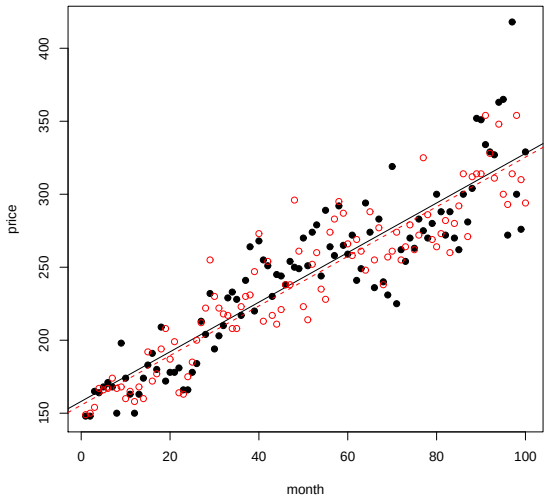

FIGURE 5. Equal-slope regression of the Oxford house-price data. (Solid line/points) Semi-Detached, (circles/dashed) Terraced.

Example 2.5. We looked in Example 2.4 at prices for Terraced and Semi-Detached houses in Oxford. Let us see if Terraced and Semi-Detached houses have grown at different rates. In order to make the whole thing a little more interesting, I will add Flats to the picture. I am ignoring the somewhat lower variance of the Flats data, and more on this anon. Have Flats increased in price at the same rate as Terraced and Semi-Detached properties? Is there any difference in the rates or intercepts for the latter house types?

With Flat providing the baseline, we fit the model

$$y_k \sim \alpha + \alpha_T g_{k,T} + \alpha_{SD} g_{k,SD} + \gamma_M x_{k,M} + \gamma_{MT} x_{k,M} g_{k,T} + \gamma_{MSD} x_{k,M} g_{k,SD} + \epsilon_k$$

with $\epsilon_k \sim \mathrm{N}(0, \sigma^2)$ and, for observation $k = 1, 2, ..., n$, we have $y_k = \texttt{price}[\texttt{k}]$, $g_{k,T}$ is the dummy indicator variable for $\texttt{type}[\texttt{k}] = \texttt{Terraced}$, $g_{k,SD}$ is the dummy indicator variable for $\texttt{type}[\texttt{k}] = \texttt{Semi}-\texttt{Detached}$ and $x_{k,M}$ is the value of $\texttt{month}[\texttt{k}]$. In month 0 (so, at the intercept), the Flat price is α, the Semi-Detached price is $\alpha + \alpha_{SD}$, and the Terraced price is $\alpha + \alpha_T$ thousands of pounds. Flat prices go up at rate γ_M, Semi-Detached up at rate $\gamma_M + \gamma_{MSD}$, and Terraced up at rate $\gamma_M + \gamma_{MT}$ thousands of pounds per month. In *R* the model `price~month*type` is expanded as the model `price~1+ month + type + month:type`, so that the `:` notation gives the product term by itself, with no lower order terms, while the `*` notation includes lower order terms by default.

```
> ohp.lm<-lm(price~month*type,data=ohp)
> (ohp.lms<-summary(ohp.lm))

Call:
lm(formula = price ~ month * type, data = ohp)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-54.522 -13.275  -1.146  10.431  93.285

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             126.50182    4.14322  30.532  < 2e-16 ***
month                     1.21838    0.07123  17.105  < 2e-16 ***
typeSemi-Detached        29.61091    5.85940   5.054 7.63e-07 ***
typeTerraced             31.00000    5.85940   5.291 2.39e-07 ***
month:typeSemi-Detached   0.51978    0.10073   5.160 4.55e-07 ***
month:typeTerraced        0.44119    0.10073   4.380 1.65e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.56 on 294 degrees of freedom
Multiple R-Squared: 0.8661,     Adjusted R-squared: 0.8638
F-statistic: 380.3 on 5 and 294 DF,  p-value: < 2.2e-16
```

The variable mapping is

$$(\alpha, \alpha_{SD}, \alpha_T, \gamma_M, \gamma_{MSD}, \gamma_{MT}) = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6).$$

We see that Flat price grows at a lower rate than Semi or Terraced (since the offsets $\alpha_{SD} \simeq 0.52, \alpha_T \simeq 0.44$ are positive, and significant). We can test for differing rates between Terraced and Semi-Detached. The test statistic is

$$\begin{aligned} T &= \frac{\hat{\beta}_5 - \hat{\beta}_6}{\sqrt{s^2(X^TX)^{-1}_{5,5} + s^2(X^TX)^{-1}_{6,6} - 2s^2(X^TX)^{-1}_{5,6}}} \\ &\simeq \frac{0.52 - 0.44}{\sqrt{0.10073^2 + 0.10073^2 - 2 \times 0.005074}} \\ &\simeq 0.78 \end{aligned}$$

Now if $T_{n-p} \sim t(n-p)$ with $n = 300$ and $p = 6$ here, then the p-value $2(1 - \Pr(T_{n-p} > T))$ is $2 * (1 - \texttt{pt}(\texttt{T}, 294)) \simeq 0.436$. Note that the matrix $(X^TX)^{-1}$ is part of the output of `summary()`, so the code I used to make this test was

```
> (beta<-ohp.lm$coefficients)
            (Intercept)                   month       typeSemi-Detached
            126.5018182               1.2183798              29.6109091
          typeTerraced month:typeSemi-Detached      month:typeTerraced
             31.0000000               0.5197840               0.4411881
> (s<-ohp.lms$sigma)
[1] 20.56092
> XTXi<-ohp.lms$cov.unscaled
> (T<-abs( (beta[5]-beta[6])/(s*sqrt(XTXi[5,5]+XTXi[6,6]-2*XTXi[5,6])) ))
               0.780243
> 2*(1-pt(T,ohp.lm$df.residual))
              0.4358756
```

We conclude that there is no evidence for different rates, when the intercepts are unequal. We can see the fitted lines in Figure 6. There is clearly little in it.

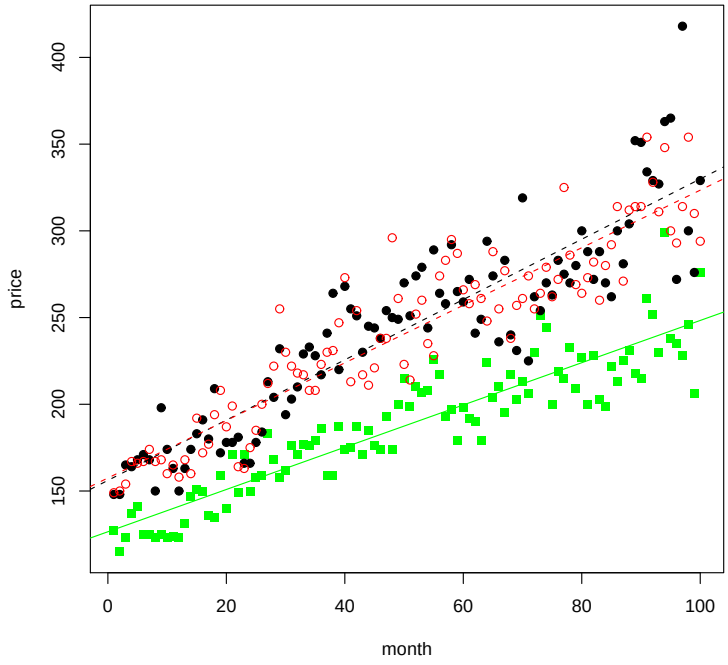

FIGURE 6. Regression of the Oxford house-price data. (upper dashed black line/black full circles) Semi-Detached, (lower dashed red line/red empty circles) Terraced, (solid green line/green squares) Flat.

We now want to make the combined test, dropping the intercept *and* slope distinction between Terraced and Semi-Detached. Since we now make no distinction between Terraced and Semi-Detached, we are effectively merging the levels `Terraced` and `Semi-Detached` within the categorical variable `type`. Let $g_{k,TSD} = g_{k,T} + g_{k,SD}$. With level `Flat` again providing the baseline, the reduced model, with $\alpha_T = \alpha_{SD}$, $\gamma_T = \gamma_{SD}$ is

$$y_k \sim \alpha + \alpha_{TSD} g_{k,TSD} + \gamma_M x_{k,M} + \gamma_{MTSD} x_{k,M} g_{k,TSD} + \epsilon_k$$

One easy way to implement this in R is to re-level the categorical variable, but otherwise proceed as before.

```
> # last two levels Semi-Detached and Terraced are over-written with 'T.or.SD'
> ohpr<-ohp
> levels(ohpr$type)<-c('Flat','T.or.SD','T.or.SD')
> # fit the reduced model
> ohpr.lm<-lm(price~month*type,data=ohpr)
> # calculate residual sums of squares for full (rss1) and reduced (rss0) models
> rss0<-sum(ohpr.lm$residuals^2)
> rss1<-sum(ohp.lm$residuals^2)
```

```
> # form the F statistic and calculate a p-value
> F<-((rss0-rss1)/2)/(rss1/294)
> (pval<-1-pf(F,2,294))
[1] 0.4983896
```

At around 0.5, the p-value for the LRT is not significant, so the test favors the reduced model. It seems that both prices and trends in prices for Terraced and Semi-Detached types are the same, as you might guess from Figure 6.

You can check, from the `summary(ohpr.lm)` output (not shown) that $\hat{\gamma}_{MTSD}$ is non-zero and positive (the p-value for $\hat{\gamma}_{MTSD} > 0$ is tiny), so Flat prices have increased at a rate which is significantly lower than the rate of increase for Terraced and Semi-Detached properties. Note that this is still not the same as making two regressions, as we are imposing equal error variance σ^2 for the response under the two types. Note also that our conclusions are based on gross data for Oxford, and the data have been slightly jittered, so local trends could differ. [End L5]

2.8. **Blocks, Treatments and Designs.** Chapters 14-16 of 'Linear Models with R' by J. Faraway covers this material at about the right level for us. See the discussion in Davison (2003) for more detail.

One important kind of categorical variable arises when the data have been gathered from b *blocks*, $y_{i,j},\ i = 1, 2, ..., b,\ j = 1, 2, ..., n_i$ corresponding to groups of subjects that are expected to have similar response to the explanatory variables. Imagine collecting observations of the body mass index (BMI) of eighty five-year-old children from different schools. In one design we measure the BMI of eight randomly selected five-year-olds in each of 10 schools. For each child we record the BMI and the school. In another design, we might do exactly the same, but fail to record the school. The first data set has a block structure, with 10 blocks. The block index is often explanatory. If, for example, there is a correlation between parent income-level, school and incidence of obesity, then 'school' will be explanatory for 'BMI'. In such cases we code the block index i as a categorical explanatory variable in the design matrix.

Besides taking subjects from distinct groups, and distinguishing group responses, we may also give subjects different treatments, and distinguish responses to different treatments. If there is a block structure to the population, with subjects in different blocks having different treatment responses, and we ignore it, then this will tend to inflate the estimated error variance s^2, and real differences in the treatment response may not be detected.

"Treatment factors are those for which we wish to determine if there is an effect. Blocking factors are those for which we believe there is an effect. We wish to prevent a presumed blocking effect from interfering with our measurement of the treatment effect", a a neat encapsulation I found in Heiberger and Holland 'Statistical Analysis and Data Display', Springer (2004).

Example 2.6. The following example is taken from Dr D. Lunn's previous lecture notes for this course. Twelve piglets were fed three different diets (A, B and C) in order to see which diet led to the greatest weight gain. The piglets came from four different litters (I,II, III and IV). We are looking for an effect (a mean shift in weight gain) due to diet, and we want to allow for variation (another mean shift) due to litter. If $y_{i,j}$ is the response in row i and column j of Table 2 then the model

	Diet		
Litter	A	B	C
I	89	68	62
II	78	59	61
III	114	85	83
IV	79	61	82

TABLE 2. Piglet diet data from D. Lunn (2007)

we want to fit is

$$\begin{aligned} y_{1,1} &= \alpha + \epsilon_{1,1} \\ y_{1,j} &= \alpha + \tau_j + \epsilon_{1,j} \\ y_{i,1} &= \alpha + \gamma_i + \epsilon_{i,1} \\ y_{i,j} &= \alpha + \gamma_i + \tau_j + \epsilon_{i,j} \end{aligned}$$

for $i = 2, 3, 4$ and $j = 2, 3$. Notice that $i = 1$ (Litter I) and $j = 1$ (Diet A) are the baseline levels for Litter and Diet. The blocks (`Litter`) and treatments (`Diet`) are categorical, and can be coded using dummy variables. Let $k = 1, 2, ..., n = 12$ run over subjects, as appears in the margin of the following output.

```
> pigs<-data.frame(c(89,78,114,79,68,59,85,61,62,61,83,82),
+                  c('I','II','III','IV'),
+                  c('A','A','A','A','B','B','B','B','C','C','C','C'))
> names(pigs)<-c('Gain','Litter','Diet')
> pigs
   Gain Litter Diet
1    89      I    A
2    78     II    A
3   114    III    A
. . .
10   61     II    C
11   83    III    C
12   82     IV    C
```

Each row corresponds to a subject (a piglet). Let $g_{k,1}, ..., g_{k,4}$ be four dummy indicator variables for the `Litter` of the subject in row k. Let $z_{k,1}, z_{k,2}, z_{k,3}$ be dummy variables for `Diet`. The model above is

$$y_k = \alpha + \gamma_2 g_{k,2} + \gamma_3 g_{k,3} + \gamma_4 g_{k,4} + \tau_2 z_{k,2} + \tau_3 z_{k,3} + \epsilon_k$$

If for $a = 1, ..., 4,\ G_a = (g_{1,a}, ..., g_{n,a})^T$ and for $a' = 1, 2, 3,\ Z'_a = (z_{1,a'}, ..., z_{n,a'})^T$, then the design is $X = (G_2, G_3, G_4, Z_2, Z_3)$.

```
> (X<-model.matrix(Gain~Litter+Diet,data=pigs))
   (Intercept) LitterII LitterIII LitterIV DietB DietC
1            1        0         0        0     0     0
2            1        1         0        0     0     0
3            1        0         1        0     0     0
. . .
10           1        1         0        0     0     1
11           1        0         1        0     0     1
```

```
12           1        0          0        1     0     1
```

In matrix notation the model we are fitting is $Y = X\beta+\epsilon$ with $\beta = (\alpha, \gamma_2, \gamma_3, \gamma_4, \tau_2, \tau_3)$. Now, lets fit the model.

```
> pigs.lm<-lm(Gain~Litter+Diet,data=pigs)
> summary(pigs.lm)

Call:
lm(formula = Gain ~ Litter + Diet, data = pigs)

Residuals:
   Min     1Q Median     3Q    Max
-8.250 -4.937 -0.375  2.938 12.750

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   86.250      5.763  14.967  5.6e-06 ***
LitterII      -7.000      6.654  -1.052  0.33332
LitterIII     21.000      6.654   3.156  0.01967 *
LitterIV       1.000      6.654   0.150  0.88547
DietB        -21.750      5.763  -3.774  0.00924 **
DietC        -18.000      5.763  -3.124  0.02049 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.15 on 6 degrees of freedom
Multiple R-Squared: 0.8569,     Adjusted R-squared: 0.7376
F-statistic: 7.184 on 5 and 6 DF,  p-value: 0.01623
```

It is clear from the parameter estimates that Litter I and Diet A have been taken as the baseline levels for the respective categorical variables.

Is `Diet` predictive for `Gain`? The ANOVA table for variable groups

$$\{1\}, \{\texttt{LitterII}, \texttt{LitterIII}, \texttt{LitterIV}\}, \{\texttt{DietB}, \texttt{DietC}\}$$

is

```
> anova(pigs.lm)
Analysis of Variance Table

Response: Gain
          Df  Sum Sq Mean Sq F value  Pr(>F)
Litter     3 1304.25  434.75  6.5458 0.02545 *
Diet       2 1081.50  540.75  8.1418 0.01952 *
Residuals  6  398.50   66.42
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

so Diet is predictive at 5% significance.

Example 2.7. What happens if the data did not record the block variable Litter? The variation within each diet is inflated by variation due to Litter. If the Litter variable is not there, this variation increases the residuals, so the estimated variance,

$s^2 = \text{RSS}/(n-p)$, is larger. This in turn pulls down the t and F test statistics, which both have RSS in the denominator, so p values tend to increase. The outcome of ignoring variation due to block variables is that the response due to treatment variables which are marginally significant can be explained away as noise, as their parameters are no longer significantly different from zero.

```
> anova(lm(Gain~Diet,data=pigs))
Analysis of Variance Table

Response: Gain
          Df  Sum Sq Mean Sq F value Pr(>F)
Diet       2 1081.50  540.75  2.8582 0.1094
Residuals  9 1702.75  189.19
```

Now the F-test for a linear relation between Gain and Diet shows Diet is no longer explanatory - the effect is lost. Notice that the error variance estimate s^2 has jumped. In the `summary()` output above, for the model $\texttt{Gain} \sim \texttt{1} + \texttt{Litter} + \texttt{Diet}$, we see $s = 8.15$ where the model $\texttt{Gain} \sim \texttt{1} + \texttt{Diet}$ with no block structure has $s = \sqrt{\text{RSS}/(n-p)} = \sqrt{1702.75/9} = 13.75$. This jump in RSS, from 398 (in the $\texttt{Gain} \sim \texttt{1} + \texttt{Litter} + \texttt{Diet}$ ANOVA) to 1702 (in the $\texttt{Gain} \sim \texttt{1} + \texttt{Diet}$ ANOVA), is due to the extra variation across litter being treated as variation across treatment.

When we set up a design, we may have some choice in the assignment of treatments to subjects. The point of the trial is to see if the response depends on the treatment. Suppose the 'treatment' levels are 'Give drug' and 'Give placebo' and the response is some index of health. If we are allowed to choose which subject gets which treatment, we could distort the trial outcome by, for example, choosing to give the drug to subjects who for some reason are more likely to get better anyway. Wonderdrug! In order to avoid all traps of this kind, we typically assign the treatments to subjects completely at random. A design with everyone in one block, and treatments assigned to subjects at random, is called *completely randomised*. If the subjects are in blocks, with m_a subjects in block a, and we apply treatment $t = 1, 2, ..., T$ to $m_{a,t}$ different subjects in block $a = 1, 2, ..., b$, then, for each treatment, we choose $m_{a,t}$ subjects independently at random and without replacement from the m_a subjects in the a'th block. If $m_{a,t} = m$, so that each treatment is applied to the same number of subjects in each block, then the design is called a *randomised complete block design*. The treatments are distributed in a balanced way through the blocks.

The piglets data is balanced, as each of the four litters contains one piglet on each of the three diets, so $b = 4, T = 3$ and $m_{a,t} = 1$ for all $a = 1, ...4$, and $t = 1, 2, 3$. A *randomised complete block design* is balanced in such a way that the block and treatment parameter estimates are independent, and so the treatments can be analyzed separately from blocks. Informally, in a completely balanced design the treatments are all tested under the same conditions, so when we compare treatments, it doesn't matter what those conditions were. You have another example of a completely balanced design in the last problem of Problem sheet 2.

Sometimes the number of subjects m_a in a block is smaller than T the number of treatments. In that case the design will be *incomplete*, since some blocks will have no instances of some treatments. An incomplete block design may still be balanced.

3. Model Checking and Model Selection

3.1. **Model Checking.** We are fitting a normal linear model $y = X\beta + \epsilon$ with $\epsilon \sim \mathrm{N}(0, I_n\sigma^2)$, and X a $n \times p$ design matrix with one column $X_i = (x_{1,i}, x_{2,i}, ..., x_{n,i})^T$ for each of the $i = 1, 2, ..., p$ explanatory variables. Under this model the vector y of n responses is a realisation of the rv $Y \sim \mathrm{N}(X\beta, I_n\sigma^2)$.

Model violations take many forms. A misspecified model is structurally inappropriate for essentially all responses. For example, the response y may not be a linear function of the linear predictors $X\beta$, and we need to transform the response. This is discussed in a section below on the Box-Cox family of transformations. We may choose to work with a misspecified model if we have reason to believe that the biases in fitted values or parameter estimates are not large enough to invalidate the inference, for parameters of interest. On the other hand the problem may lie with the data. The normal linear framework may be good for most data points, but a few of the responses may have quite different causative factors. Such data points are called outliers. We try to identify them and then typically remove them from further analysis.

We have a range of validatory checks for model mispecification and outlier detection. The most straightforward check for linearity is to plot the explanatory variables against the response, as we do in a lattice plot, such as Figure 1. Variation in the response caused by variation in other variables may obscure the linear response to any single variable. Added variable plots (Problem sheet 1 Q1) help by focusing on the relation between the response and a single variable, removing variation due to other fitted variables. These check the linearity of the running mean. We have checks also on the independence and constant variance of ϵ, the errors. We saw that the fitted values $\hat{y} = Hy$ (with hat matrix $H = X(X^TX)^{-1}X^T$) and the n-component vector of residuals $e = y - \hat{y}$ are independent under the model, so a plot of residuals against fitted values should show no correlation.

Misfit. One weakness of the residuals/fitted-values plot, as a diagnostic tool, is that the residuals may have unequal variance under the model. A large residual e_k could be a sign that the kth data point is an outlier, but it might have a large variance as a consequence of the experimental design. The individual residuals e_i $i = 1, 2, ..., n$ are all mean zero normal rv, with variance matrix

$$\begin{aligned} \mathrm{var}(e) &= \mathrm{var}((I - H)y) \\ &= \sigma^2(I - H)(I - H)^T \\ &= \sigma^2(I - H), \end{aligned}$$

since $H^2 = H = H^T$. The diagonal entries $h_{kk},\ k = 1, 2, ..., n$ of H control the variances of the components of e and $\hat{y}$. The residual variances can be unequal, since $\mathrm{var}(e_k) = \sigma^2(1 - h_{kk})$.

Exercise Show that $\mathrm{var}(\hat{y}) = \sigma^2 H$.

The standardised residuals $r_k,\ k = 1, 2, ..., n$ have zero mean and equal variance. They are obtained by scaling the residuals:

$$r_k = \frac{e_k}{s\sqrt{1 - h_{kk}}},$$

with s^2 an unbiased estimate of σ^2, under the normal linear model. They (the r_k) have approximately unit variance, and can be compared with standard normal

variables, an approximation which will be good at large $n-p$. Because s and e are correlated, we cannot easilly compute the distribution of the standardised residuals.

Exercise Show that the standardised residuals r are (like the residuals e) independent of the fitted values, $\hat{y}$.

We can make normal qqplots of standardised residuals, and plot them (the r_k) against fitted values $\hat{y}$. We often mistakenly fit a model of constant variance to data in which the variance of the response increases with the underlying mean. This model mispecification is witnessed by a trend to increasing variance in e' with $\hat{y}$.

A response y_k which generates a relatively large residual e_k need not be an outlier, since $\mathrm{var}(e_k) = \sigma^2(1-h_{kk})$, and $1-h_{kk}$ may be relatively large. We might expect the standardised residuals r to be a good basis for outlier detection, since these should have variance about one under the normal linear model. Standardised residuals exceeding two (standard deviations) are large. The problem here is that $s^2 = e^T e$, the denominator in the expression for r_k, is computed from the residuals e themselves. A response y_k, which is truly outlying, may have a large residual e_k, but this will inflate our estimate s^2 for σ, and we may end up with a moderate standardised residual r_k. A bad response with an h_{kk} value close to one has a low variance. It 'pulls' the fitted surface towards itself, so that e_k is small, and again r_k is not obviously large. What to do?

We can treat this problem using an idea related to 'cross-validation'. We remove response y_k and row k, $\mathbf{x}_k = (X_{k,1}, ..., X_{k,p})^T$ say, from the data and compare the fitted values $\hat{y}$ we got with all the data with the fitted values $\hat{y}_{-k}$ we get when $\mathbf{x}_k, y_k$ are removed. Denote by y_{-k} and X_{-k} the remaining response and design data with y_k and $\mathbf{x}_k$ removed. Let $\hat{\beta}_{-k} = (X_{-k}^T X_{-k}) X_{-k}^T y_{-k}$ give the new parameter estimates. The kth deletion, or studentised residual is

$$r'_k = \frac{y_k - \mathbf{x}_k \hat{\beta}_{-k}}{\mathrm{std.err}(y_k - \mathbf{x}_k \hat{\beta}_{-k})}.$$

It may be shown (Problem Sheet 3 Q2) that $r'_k \sim t(n-p-1)$ and r' and $\hat{y}$ are independent. *Our studentised residuals have equal variance, known distribution, and can be compared to a standard normal (using for example a qqplot). We can plot r' against $\hat{y}$. Any visible correlation is a sign of model mispecification, and data points with $|r'_k| > 2$ show misfit, and are possible outliers.*

We will now derive a useful auxiliary formula for the studentised residuals. It may be shown that

$$\hat{\beta}_{-k} = \hat{\beta} - (X^T X)^{-1} \mathbf{x}_k \frac{e_k}{1-h_{kk}}.$$

Exercise Verify this expression for $\hat{\beta}_{-k} = (X_{-k}^T X_{-k})^{-1} X_{-k}^T y_{-k}$. Quote the Woodbury formula

$$(X_{-k} X_{-k})^{-1} = (X^T X)^{-1} + (X^T X)^{-1} \mathbf{x}_k (1 - \mathbf{x}_k (X^T X)^{-1} \mathbf{x}_k) \mathbf{x}_k^T (X^T X)^{-1}$$

without proof, and use the fact that $X_{-k}^T y_{-k} = X^T y - X^T (0, ..., 0, y_k, 0, ..., 0)^T$.

Exercise Using the formula above for $\hat{\beta}_{-k}$, show that

$$y_k - \mathbf{x}_k \hat{\beta}_{-k} = \frac{e_k}{1-h_{kk}}$$

and hence show

$$\mathrm{var}(y_k - \mathbf{x}_k\hat{\beta}_{-k}) = \sigma^2/(1-h_{kk}).$$

Estimating σ^2 as s^2_{-k}, the residual standard error with the kth observation removed, we have studentised residuals

$$r'_k = \frac{e_k}{s_{-k}(1-h_{kk})^{1/2}}.$$

Exercise The residual variance is $(n-p)s^2 = y^Ty - y^TX\hat{\beta}$. The kth deletion residual-variance is $(n-p-1)s^2_{-k} = y^Ty - y_k^2 - (y^TX - y_k\mathbf{x}_k)\hat{\beta}_{-k}$. Using the formulae above, show that

$$(n-p-1)s^2_{-k} = (n-p-r_k^2)s^2$$

and hence obtain the useful form

$$r'_k = r_k\sqrt{\frac{n-p-1}{n-p-r_k^2}}.$$

This formula is important computationally, since it shows that we can compute the studentised residuals without making n linear regressions for the n deletions, just using the results of the primary regression.

3.1.1. *Leverage.* The diagonal entries $h_{kk} = H_{kk}$ of the hat matrix are called the *leverage* components. Since $h_{kk} = \mathrm{var}(y_k)/\sigma^2$ we have $0 \le h_{kk} \le 1$. Since $\mathrm{var}(e_k) = \sigma^2(1-h_{kk})$, a point with leverage h_{kk} close to one has low variance: since $\mathsf{E}(e_k) = 0$ the fitted surface $\mathbf{x}\hat{\beta}$ must be pulled close to the kth response y_k. If $\mathbf{x}_k, y_k$ is an outlier with high leverage, then predictions $\mathbf{x}\hat{\beta}$ for $\mathbf{x}$ near $\mathbf{x}_k$ will be poor.

How big is big, when it comes to leverage? The 'average' leverage for a $n \times p$ design is $\bar{h} = p/n$, as we will see shortly, so points with leverage values above $2p/n$ get special attention. Since they are having more impact on the final fit than other points, it is important that they are not outliers. The average leverage is given by the trace of the hat matrix H, so

$$\begin{aligned}\sum_{k=1}^n h_{kk} &= \mathrm{trace}(H)\\ &= \mathrm{trace}(X(X^TX)^{-1}X^T)\\ &= \mathrm{trace}(X^TX(X^TX)^{-1}),\end{aligned}$$

by the cyclic permutation property of the trace $\mathrm{trace}(ABC) = \mathrm{trace}(CAB)$. It follows that

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^n h_{kk} &= \mathrm{trace}(I_p)/n\\ &= p/n.\end{aligned}$$

Alternatively, $\mathrm{trace}(H) = p$, as H has the eigenvalue $\lambda = 1$ repeated p times, and the eigenvalue $\lambda = 0$ repeated $n-p$ times, since there are p linearly independent vectors $v = e_1, ..., e_p$ satisfying $Hv = v$ and $n-p$ vectors $u = e_{p+1}, ..., e_n$ satisfying $Hu = 0$. We saw this when we made the expansion of y in $e_1, ..., e_n$ in Section 2.

3.1.2. *Influence.* A point with high leverage need not do much damage, if it lies close to the fitted surface through other points - the surface would have gone close to it anyway. Such a point is said to have high leverage but low *influence*. Highly influential points shift the fitted surface far from where it would have lain if the influential point were not included. The *Cook's distance* for a point $\mathbf{x}_k, y_k$ is a measure of influence given by the sum of the squares of the shift in fitted values when point k is removed. Now the Cook's distance for the kth data point is defined to be

$$C_k = \frac{(\hat{y} - \hat{y}_{-k})^T(\hat{y} - \hat{y}_{-k})}{ps^2}.$$

It may be shown, using the formula in Section 3.1 above, that

$$C_k = \frac{r_k^2 h_k}{p(1 - h_{kk})}.$$

Our intuition is that high influence occurs where there is misfit and large leverage. The factor $h_{kk}/(1-h_{kk})$ rises with increasing leverage. The factor r_k^2 (standardised residual squared) is related to misfit. A rough rule of thumb is that $|r_k| > 2$ and h_{kk} are separate causes for concern, so points with Cook's distance exceeding

$$C_k \gtrsim \frac{8}{n - 2p}$$

have high influence. Such points are generally outliers, and are removed from the analysis.

3.1.3. *Model checking, example.*

Example 3.1. The `data(swiss)` dataset is described in the R documentation as follows.

The data give a standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888.

Switzerland, in 1888, was entering a period known as the Demographic Transition, its fertility was beginning to fall from the high level typical of underdeveloped countries. The data come from Mosteller, F. and Tukey, J. W. (1977) who give the original source.

The `swiss` *data frame has 47 observations on 6 variables, each of which is given as a percentage. The columns are*

Fertility	*a standardized fertility measure*
Agriculture	*% of males involved in agriculture as occupation*
Examination	*% draftees receiving highest mark on army examination*
Education	*% draftees receiving education beyond primary school*
Catholic	*% catholic (ie, not protestant)*
Infant.Mortality	*% live births who live less than 1 year*

Which variables are explanatory for Fertility?

We can map the percentages $(0, 100)$ into $(-\infty, \infty)$ using the logistic transformation $x \leftarrow log(x/(100 - x))$. Since 0 and 100 may appear, we modify this to something like $x \leftarrow log((1 + x)/(101 - x))$, checking for linearity between mapped variable and response.

```
> data(swiss)
> head(swiss) #first few rows
             Fertility Agriculture Examination Education Catholic Infant.Mortality
```

```
Courtelary        80.2        17.0          15        12     9.96              22.2
Delemont          83.1        45.1           6         9    84.84              22.2
Franches-Mnt      92.5        39.7           5         5    93.40              20.2
Moutier           85.8        36.5          12         7    33.77              20.3
Neuveville        76.9        43.5          17        15     5.16              20.6
Porrentruy        76.1        35.3           9         7    90.57              26.6
>
> sw<-swiss; #map data into R
> sw[,-1]<-log((swiss[,-1]+1)/(101-swiss[,-1]))
> n<-dim(sw)[1]; p<-dim(sw)[2]
```

Notice that each row (*ie* each data point) is named according to the Swiss province that data point measures. The transformed data are displayed in Figure 7. We will (i) fit a normal linear model, (ii) look for outliers, (iii) remove them, refit, and look again for outliers, (iv) make a model reduction to something simple which nevertheless predicts the response and finally (v) check again for outliers in the reduced model.

```
> # (i) fit a normal linear model
> sw1.lm<-lm(Fertility~Infant.Mortality+Examination+Education+Catholic+Agriculture,
+            data=sw)
> summary(sw1.lm)
...
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       76.7438    11.4611   6.696 4.44e-08 ***
Infant.Mortality  23.6269     6.9393   3.405  0.00149 **
Examination       -6.2086     3.7104  -1.673  0.10188
Education         -6.8316     2.6984  -2.532  0.01528 *
Catholic           0.8225     0.6183   1.330  0.19079
Agriculture       -1.8702     1.6896  -1.107  0.27478

Residual standard error: 8.398 on 41 degrees of freedom
F-statistic: 12.16 on 5 and 41 DF,  p-value: 2.960e-07
```

We can find the points of high influence and display them as in Figure 7. The R code to do this is available in the course website in the file L7.R. We apply the `cooks.distance()` function to the `lm()` output (and `round(...,3)` to 3dp). See the top panel of Figure 8. We see that that the data points for `V. De Geneve` and `La Vallee` have high influence. They must have high misfit (as measured by standardised residual) or high leverage, or both. The leverages are the diagonal entries in the hat matrix (rounded to 3 dp). See the centre panel of Figure 8. The data points for `V. De Geneve` and `La Vallee` have high leverage in the sense that they exceed twice the mean leverage p/n. How about misfit? The numbers are displayed in the bottom panel of Figure 8. `V. De Geneve` and `Rive Gauche` are above threshold (which equals 2) for high misfit. `Rive Droite` and `La Vallee` have highish misfit. However, `Rive Gauche` and `Rive Droite` do not have high enough leverages (0.1 and 0.178) to make them a problem. On the other hand `La Vallee` has a particularly high leverage (0.379), so its milder misfit (1.7) is still a cause for concern. There are two diagnostic functions we havnt mentioned: `rstudent()` and `fitted.values` act on the `lm()` output to give studentised residuals r' and fitted

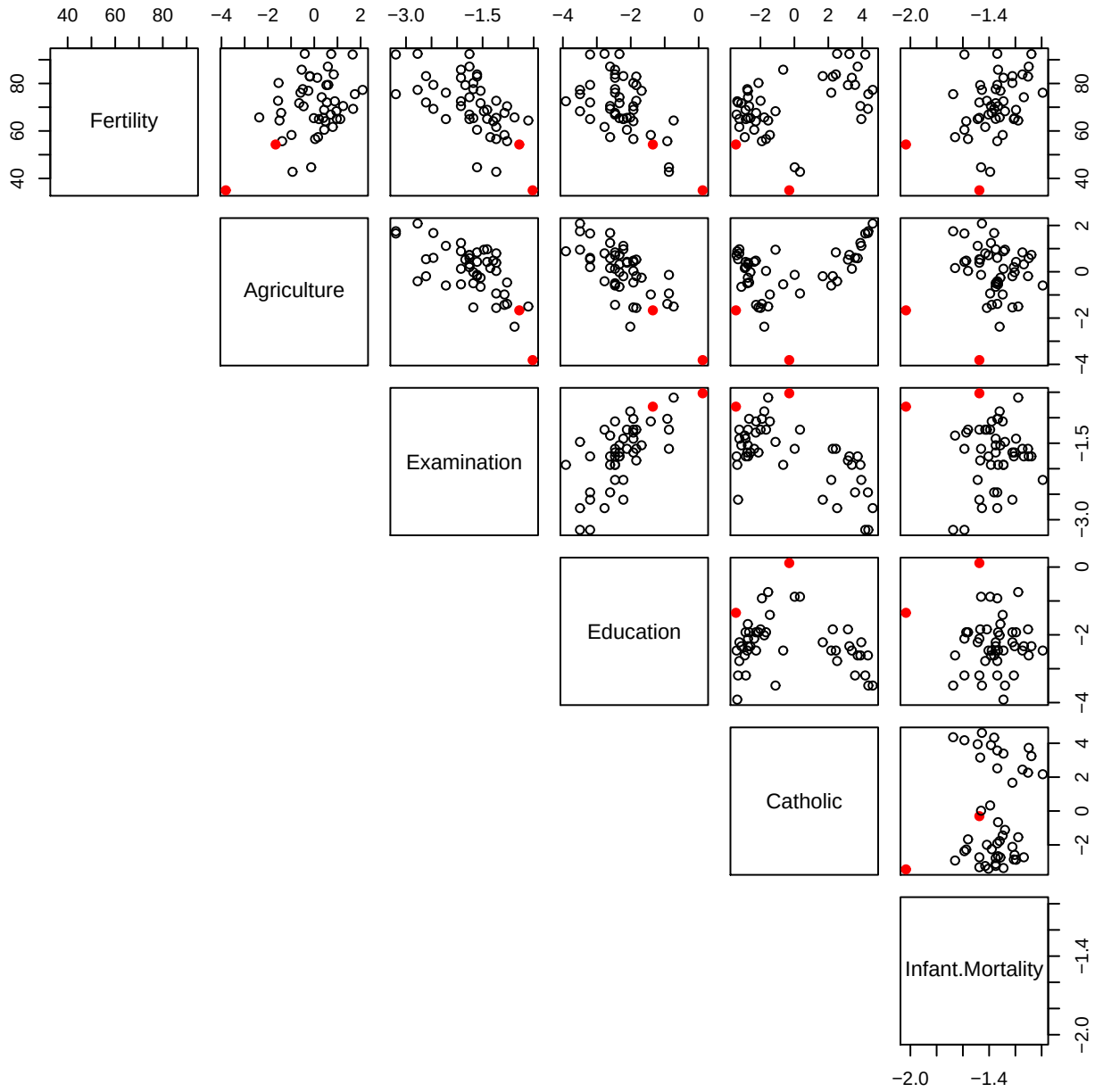

FIGURE 7. Pairs plot for Swiss fertility data. Percentage data has been mapped monotonically to the interval $(-\infty, \infty)$.

```
> #V. De Geneve and La Vallee are points of high influence
> cooks.distance(sw1.lm)
  Courtelary     Delemont Franches-Mnt      Moutier   Neuveville   Porrentruy        Broye
       0.021        0.002        0.043        0.017        0.034        0.064        0.017
       Glane      Gruyere       Sarine      Veveyse        Aigle      Aubonne     Avenches
       0.066        0.005        0.015        0.015        0.010        0.000        0.001
    Cossonay    Echallens     Grandson     Lausanne    La Vallee       Lavaux       Morges
       0.005        0.046        0.001        0.002        0.294        0.000        0.004
      Moudon        Nyone         Orbe         Oron      Payerne Paysd'enhaut       Rolle
       0.054        0.001        0.005        0.041        0.000        0.015        0.000
       Vevey      Yverdon      Conthey    Entremont       Herens     Martigwy      Monthey
       0.003        0.010        0.029        0.025        0.016        0.006        0.007
  St Maurice       Sierre         Sion       Boudry La Chauxdfnd     Le Locle    Neuchatel
       0.014        0.082        0.035        0.007        0.003        0.013        0.012
  Val de Ruz ValdeTravers V. De Geneve  Rive Droite  Rive Gauche
       0.007        0.000        0.464        0.118        0.081
> 8/(n-2*p)
[1] 0.229


> #V. De Geneve and La Vallee are points of high leverage
> hatvalues(sw1.lm)
  Courtelary     Delemont Franches-Mnt      Moutier   Neuveville   Porrentruy        Broye
       0.118        0.160        0.177        0.052        0.084        0.164        0.117
       Glane      Gruyere       Sarine      Veveyse        Aigle      Aubonne     Avenches
       0.128        0.077        0.092        0.145        0.094        0.092        0.118
    Cossonay    Echallens     Grandson     Lausanne    La Vallee       Lavaux       Morges
       0.114        0.178        0.056        0.087        0.379        0.118        0.072
      Moudon        Nyone         Orbe         Oron      Payerne Paysd'enhaut       Rolle
       0.105        0.068        0.117        0.183        0.114        0.219        0.080
       Vevey      Yverdon      Conthey    Entremont       Herens     Martigwy      Monthey
       0.052        0.069        0.223        0.109        0.138        0.117        0.102
  St Maurice       Sierre         Sion       Boudry La Chauxdfnd     Le Locle    Neuchatel
       0.105        0.191        0.091        0.054        0.213        0.073        0.140
  Val de Ruz ValdeTravers V. De Geneve  Rive Droite  Rive Gauche
       0.057        0.149        0.332        0.178        0.100
> 2*p/n
[1] 0.255


> #V. De Geneve and Rive Gauche have higher misfit than Rive Droite and La Vallee.
> rstandard(sw1.lm)
  Courtelary     Delemont Franches-Mnt      Moutier   Neuveville   Porrentruy        Broye
       0.967        0.281        1.099        1.380        1.485       -1.396        0.871
       Glane      Gruyere       Sarine      Veveyse        Aigle      Aubonne     Avenches
       1.651        0.600        0.934        0.722        0.771       -0.032        0.210
    Cossonay    Echallens     Grandson     Lausanne    La Vallee       Lavaux       Morges
      -0.469       -1.124        0.329       -0.386        1.700        0.103        0.553
      Moudon        Nyone         Orbe         Oron      Payerne Paysd'enhaut       Rolle
      -1.656       -0.324       -0.459       -1.052        0.033       -0.567       -0.048
       Vevey      Yverdon      Conthey    Entremont       Herens     Martigwy      Monthey
      -0.581       -0.882       -0.776       -1.112       -0.778       -0.522       -0.590
  St Maurice       Sierre         Sion       Boudry La Chauxdfnd     Le Locle    Neuchatel
      -0.842        1.447        1.455        0.853       -0.272        0.982        0.663
  Val de Ruz ValdeTravers V. De Geneve  Rive Droite  Rive Gauche
       0.838       -0.109       -2.365       -1.806       -2.097
```

FIGURE 8. (Top) Influence, (Centre) Leverage and (Bottom) Misfit.

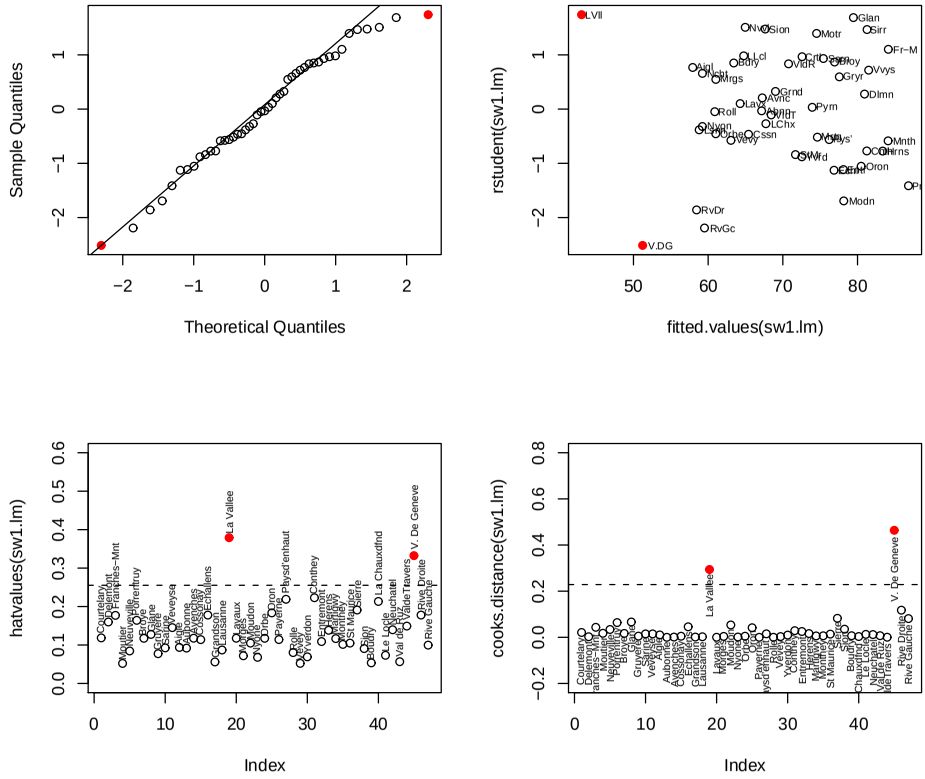

FIGURE 9. Diagnostic plots for the Swiss fertility data. The model is `Fertility` $\sim$ `Infant.Mortality` + `Examination` + `Education` + `Catholic` + `Agriculture`. The residuals displayed are all studentised. The text discusses standardised residuals in connection with the Cook's distance. Points in red, outliers by influence.

values $\hat{y}$ respectively, and were used to make the diagnostic graphs in Figure 9. The qqplot is acceptable. Notice that the outliers are in the extremes. The scatter in the plot of studentised residuals against fitted values is somewhat clustered, with some points very literally 'outlying'. These signs of correlation show signs of model violation somewhere in the model. The pairs plot itself Figure 7 seems to show acceptable linearity in at least some explanatory variables.

If, after considering background knowledge of the data gathering, and the plausibility of the candidate outlier values on physical grounds, which we don't have in this case, we decide that high influence points are indeed outliers, we may remove them, and refit.

```
> i<-cooks.distance(sw1.lm)>(8/(n-2*p))
> swr<-sw[-which(i),]
> nr<-dim(swr)[1];
```

```
> swr1.lm<-lm(Fertility~Infant.Mortality+Examination+Education+Catholic+Agriculture,
+             data=swr)
> summary(swr1.lm)
...
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       82.5002    12.3752   6.667 6.17e-08 ***
Infant.Mortality  26.9630     8.2567   3.266  0.00228 **
Examination       -6.7927     3.5219  -1.929  0.06107 .
Education         -5.9604     2.5509  -2.337  0.02469 *
Catholic           1.0270     0.6005   1.710  0.09514 .
Agriculture       -2.8355     1.7661  -1.606  0.11645

Residual standard error: 7.866 on 39 degrees of freedom
F-statistic: 10.41 on 5 and 39 DF,  p-value: 2.139e-06
```

Considering we removed just two of 47 data points, the parameters and significance levels have changed a fair bit. For example, `Examination` and `Catholic` are now marginally significant. Before we continue, lets repeat the diagnostics. Outliers can mask outliers, so we may find new points of high influence in the reduced data set. The new diagnostics are plotted in Figure 10. The qqplot was already acceptable, and is no worse. The plot of studentised residuals against fitted values is improved, with no discernable trend. No points of high influence have been exposed in the reanalysis. `La Chauxdfnd` has now a slightly elevated leverage, but its misfit is mild so its influence is not large. Conversely there are some points with studentised residuals exceeding two, however we can assume their influence remains slight.

Notice that `Examination` and `Education` have similar parameter values (around -6 to -7). Looking at the pairs plot above they are clearly correlated (as you might expect, thinking about the variables). This is an instance of variance inflation due to correlation (*ie*, near linear dependence (Section **??** and Problem Sheet 3 Q3)). If we remove one of these variables it is likely that the Std. Error of the other will drop substantially. I will go straight to the simplest model that seems to work, and make an F-test to drop `Education + Catholic + Agriculture` from the model.

```
> swr0.lm<-lm(Fertility~Infant.Mortality+Examination,data=swr)
> summary(swr0.lm)
...
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        95.133     10.871   8.751 5.15e-11 ***
Infant.Mortality   32.982      8.009   4.118 0.000175 ***
Examination       -11.811      2.100  -5.624 1.38e-06 ***

Residual standard error: 8.181 on 42 degrees of freedom
F-statistic:  21.1 on 2 and 42 DF,  p-value: 4.545e-07

> anova(swr0.lm,swr1.lm)
Analysis of Variance Table
```

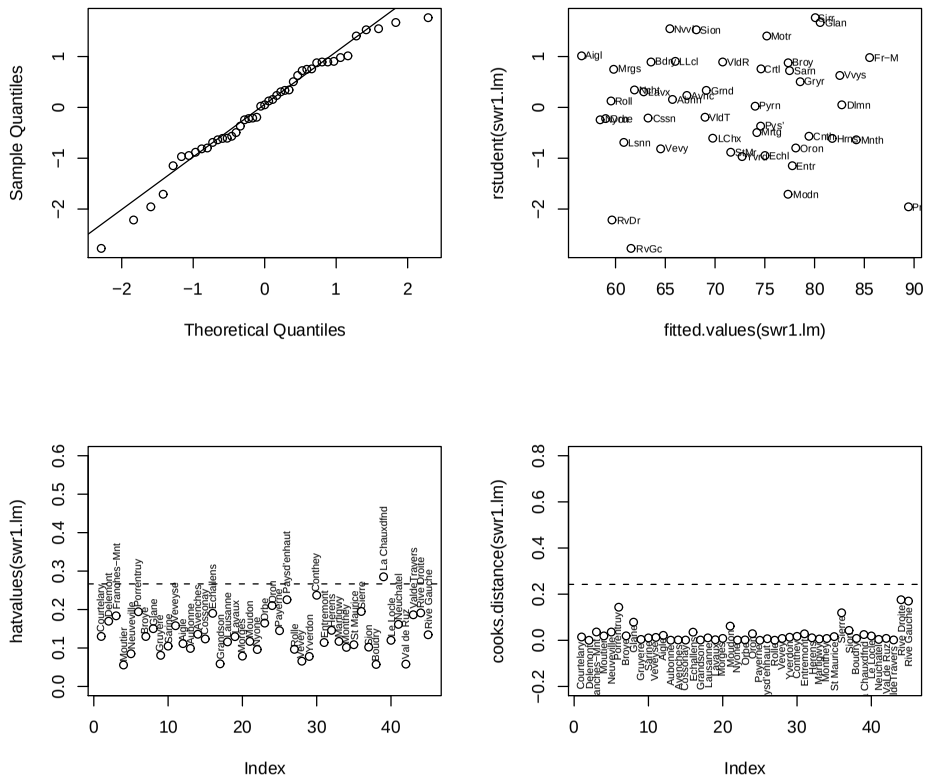

FIGURE 10. Diagnostic plots for the reduced `swiss` data (`V. De Geneve` and `La Vallee` removed) fitting the same model as Figure 9 and displaying studentised residuals.

```
Model 1: Fertility ~ Infant.Mortality + Examination
Model 2: Fertility ~ Infant.Mortality + Examination + Education + Catholic +
    Agriculture
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     42 2810.68
2     39 2413.14  3    397.54 2.1416 0.1106
```

The p-value comparing the reduced model (`Model 1.` above) to the full model (`Model 2.` above) is 0.11 so there is no evidence to support the more complex model. Notice that the p-value for `Examination` alone is now far more significant. Finally, does the reduced model have any outliers? A quick check:

```
> pr<-3 #there are p=3 parameters in the reduced model
> any(cooks.distance(swr0.lm)>(8/(nr-2*pr)))
[1] FALSE
```

shows that there are at least no points of high influence.

3.2. **Model Selection.** In some settings we aim to filter out from some original, possibly large, set of explanatory variables the ones that are explanatory for the response. Where variables are correlated, this may not be possible - which of a group of near linearly dependent variables do we drop? Physical considerations are very important. Does the model make sense? We have to go some way down the road to understanding the application domain in order to make sense of the variables and their interactions, and make sensible selections of subsets of variables.

3.2.1. *Model choice v. Exploratory data analysis.* The easiest case is where we go into the analysis with some preconceived hypothesis. The trees example, Example 2.2, worked that way. We modeled the tree as something like a cylinder, and that gave us a clear hypothesis about the relation between volume, girth and height. We fit the model and test the hypothesis.

More commonly we look at the parameters in the fitted model, observe that some are not significant, fit a range of reduced and added-variable plots, and formulate some hypothesis about the relations between variables. This leads to a test for the significance of some subset of explanatory variables which is informed by the data. We have in effect made many tests, but report just the final one. This introduces a hazard for multiple testing, which can sometimes be corrected, see Ripley (2002) section 6.10. If we don't correct, and we hardly every do, then, paraphrasing Davison (2003) section 8.7 in the subsection 'Inference after model selection', "...the only covariates for which subsequent inference using the standard confidence intervals is reliable are those for which the evidence for inclusion is overwhelming". It sometimes happens that we frame a hypothesis by looking at the data, but then realise that the hypothesis has some natural physical meaning. I can tell you that the cylinder model in the trees example (Example 2.2) was not the first model I tried for that data, but emerged in the analysis. However, the final model is so natural, that we could easily imagine a scientist going into the analysis with precisely this hypothesis (we just did).

Sometimes we use *automatic* variable selection. This means, essentially, any model selection procedure not guided by physical considerations for variable meaning. We might search over all subsets of variables for the 'best' reduced model. One choice is to look for the largest fully significant model. No variable, or set of variables, can be dropped, and if we add variables we have a model with some non-significant variables. There may be many such models. When the model space is very large we may search using stepwise methods. In *Backwards elimination* we start with the full set of variables, and successively drop the least significant, until we have a fully significant set. There is a *Forwards selection* scheme, which adds the next most significant variable. We sometimes sort the variables in an ANOVA table so that the p-values on the RHS of the table show in effect the steps of forward selection. One advantage of backwards selection, is that the initial estimator for s^2 is from the fit for all the variables, so it is not biased (upwards) by significant variables which are not included. If forwards selection started with an estimate of s^2 based on just the one or two variables in the initial model, then it might be a large overestimate, since variation in the response y due to variations in significant explanatory variables would inflate s^2, and levels of significance would suffer accordingly. This is the reason we use the same $\mathrm{RSS}_{1:p}/(n-p)$ divisor in every row of the F-coloumn of an ANOVA, rather than $\mathrm{RSS}_{1:i_k}/(n-(i_k-i_{k-1}))$.

One strategy is to search over all models and optimise some measure of the relative worth of models. The measure must somehow penalise models which are too complex or too simple, since both are poor for prediction. The AIC is just such a model choice criterion. The AIC-approach is set out in the next section.

Automatic methods are exposed to the hazard for multiple testing mentioned above. If we make alot of tests on a large number of non-significant variables we may find marginally significant variables where there are none. The approach is justified as part of exploratory data analysis. The hope is that the method will turn up some physically natural set of explanatory variables.

3.2.2. *The AIC.* One idea (there are lots, but this is core) is to consider what happens when new data Y' come along. We have parameter MLE's $\hat{\beta}(Y)$ and $\hat{\sigma}_{MLE}(Y)$ computed using the old data Y and some particular design X. If the model is a good model, then the loglikelihood $\ell(\hat{\beta}(Y), \hat{\sigma}_{MLE}(Y); Y')$ for $\hat{\beta}$ and $\hat{\sigma}_{MLE}$ in the new data should be large. This should "usually" hold, *ie*, the average over Y and Y' of $\ell(\hat{\beta}(Y), \hat{\sigma}_{MLE}(Y); Y')$ should be large. If

$$C = -2\ell(\hat{\beta}(Y), \hat{\sigma}^2_{MLE}(Y); Y')$$

then we like models that make $\mathsf{E}(C)$ small. The expectation here is over realisations of the old Y and new Y' data. Davison (2003) section 8.7.3, page 402, gives a somewhat better motivation on this crucial point, linking the analysis to ideas of cross-validation.

In an earlier version of these notes I set the following as an exercise. Here is the detail. Omit this at first reading, and jump to the definition of the AIC.

We will now derive a consistent estimator, AIC, for $\mathsf{E}(C)$. In the following, v^2 means the scalar inner product $v^T v$. We will need one auxiliary result, namely, that if $Z \sim \chi^2(\nu)$ with $\nu > 2$ then $\mathsf{E}(1/Z) = 1/(\nu - 2)$. Looking at $\ell(\hat{\beta}(Y), \hat{\sigma}_{MLE}(Y); Y)$ in Section 2.2, we see we have here

$$-2\mathsf{E}(\ell(\hat{\beta}(Y), \hat{\sigma}_{MLE}(Y); Y')) = \mathsf{E}\left(n\log(\hat{\sigma}^2_{MLE}(Y)) \ + \ \frac{(Y' - X\hat{\beta}(Y))^2}{\hat{\sigma}^2_{MLE}(Y)} \right).$$

We will deal with the second term first.

$$\begin{aligned}
\mathsf{E}\left(\frac{(Y' - X\hat{\beta}(Y))^2}{\hat{\sigma}^2_{MLE}(Y)} \right) &= \mathsf{E}\left(\frac{(Y' - X\beta + X\beta - X\hat{\beta}(Y))^2}{\hat{\sigma}^2_{MLE}(Y)} \right) \\
&= \mathsf{E}_Y\left(\frac{n\sigma^2}{\hat{\sigma}^2_{MLE}(Y)} \right) \ + \ \mathsf{E}_Y\left(\frac{(X\beta - X\hat{\beta}(Y))^2}{\hat{\sigma}^2_{MLE}(Y)} \right) + \text{constant terms},
\end{aligned}$$

since $\mathsf{E}_{Y'}((Y' - X\beta)^2) = n\sigma^2$ and

$$\mathsf{E}\left(\frac{2(Y' - X\beta)^T (X\beta - X\hat{\beta}(Y))}{\hat{\sigma}^2_{MLE}(Y)} \right) = 0$$

using the Y' expectation. Recall that $\hat{\sigma}^2_{MLE} = \mathrm{RSS}/n$. Now $X\beta - X\hat{\beta}(Y) = \mathsf{E}(\hat{Y}) - \hat{Y}$ and $\hat{Y}$ and RSS are independent, so

$$\mathsf{E}\left(\frac{(X\beta - X\hat{\beta}(Y))^2}{\hat{\sigma}^2_{MLE}} \right) = \mathsf{E}\left((\mathsf{E}(\hat{Y}) - \hat{Y})^2 \right) \mathsf{E}\left(\frac{n}{\mathrm{RSS}(Y)} \right).$$

Using $\text{var}(\hat{Y}) = \sigma^2 H$ we have

$$\begin{aligned} \mathsf{E}\left((X\beta - X\hat{\beta}(Y))^2\right) &= \sum_{k=1}^{n} \sigma^2 h_{kk} \\ &= \sigma^2 p. \end{aligned}$$

Also, RSS/σ^2 has a $\chi^2(n-p)$ distribution, so $\mathsf{E}(\sigma^2/\text{RSS}) = 1/(n-p-2)$, using our auxiliary result, so

$$\mathsf{E}\left(\frac{(Y' - X\hat{\beta}(Y))^2}{\hat{\sigma}^2_{MLE}(Y)}\right) = \frac{n^2}{n-p-2} + \frac{np}{n-p-2}.$$

We need now to deal with the first term in the expression above for $\mathsf{E}(C)$. If D is the deviance

$$D(Y) = -2\ell(\hat{\beta}(Y), \hat{\sigma}^2_{MLE}(Y); Y)$$

then

$$\mathsf{E}(D) = \mathsf{E}\left(n\log(\hat{\sigma}^2_{MLE}(Y)) + n\right)$$

from Section 2.2. We have then

$$\mathsf{E}(C) = \mathsf{E}(D) + \frac{n^2}{n-p-2} + \frac{np}{n-p-2} \ + \ \text{constant terms.}$$

Expanding the last two terms in p/n about zero, we have

$$\mathsf{E}(C) = E(D) \ + \ 2p \ + \ O(p/n) \ + \ \text{ terms not depending on } p.$$

Now $D(y)$ is an unbiased estimator for $\mathsf{E}(D)$, and $2p$ is, up to a constant term, twice the number of parameters, so the objective function we seek to minimise in our search for the optimal normal linear model is

$$AIC = -2\ell(\hat{\beta}, \hat{\sigma}^2_{MLE}; y) \ + \ 2 \times \text{ number of parameters.}$$

We derived this for normal linear models only, but it is in fact the relevant criterion for model choice in a much wider setting. In our setting

$$AIC = n\log(\text{RSS}/n) + 2p$$

where we have dropped constants, such as n, which are fixed for given data. There is another, closely related criterion, called Mallow's C_p, which is relevant for regression. It may be shown that

$$C_p = \frac{\text{RSS}}{\sigma^2} + 2p$$

is an approximation to the AIC (itself an approximation to $\mathsf{E}(C)$). We do not in general know σ^2, so it is estimated using the full model (and thereby avoiding bias from any significant variables in the full model). Mallows C_p has motivation beyond simply approximating the AIC.

Example 3.2. Returning to the trees data in Example 2.2, I mentioned in the text above that I tried a few models before settling on the (admittedly fairly obvious) product form in Example 2.2. Here is part of the model exploration I did. Let y_i be the volume, $x_{i,1}$ be the girth, and $x_{i,2}$ the height. Denote by $RSS(\alpha, \beta_1, \beta_2, \gamma)$ the residual sum of squares for the linear model

$$Y_i = \alpha + \beta_1 x_{i,1}^2 + \beta_2 x_{i,2} + \gamma x_{i,1}^2 x_{i,2} + \epsilon_i.$$

with $\epsilon \sim \mathrm{N}(0, \sigma^2)$. The following results were obtained from regression of four different models: $RSS(\alpha, \beta_1, \beta_2, \gamma) = 179.3$, $RSS(\alpha, 0, 0, \gamma) = 180.2$, $RSS(\alpha, \beta_1, \beta_2, 0) = 219.4$, $RSS(0, 0, 0, \gamma) = 181$. Carry out variable selection.

Here we have data for $n = 31$ trees, so computing the AIC's,

```
> n<-31; p<-c(4,2,3,1); RSS<-c(179.3,180.2,219.4,181.0);
> n*log(RSS/n)+2*p
[1] 62.40727 58.56248 66.66419 56.69980
```

we can make a table of AIC's by model

Model	p	RSS	AIC
$\alpha, \beta_1, \beta_2, \gamma$	4	179.3	62.4
α, γ	2	180.2	58.56
α, β_1, β_2	3	219.4	66.66
γ	1	181.0	56.70

and the winner is... the model in the last line with the least AIC, $E(Y_i) = \gamma x_{i,1}^2 x_{i,2}$. Although we did this by hand, this is in effect automatic variable selection, with no regard for the physical content of the model. The comments above about multiple testing apply here. An undesirable feature of the final model, is that we end up with the interactions without the lower order terms. The product form in the original example deals with this issue.

Note that the `step()` function in R, applied to the output of `lm()`, will carry out automatic backwards variable selection using the AIC, and report all these numbers. See th last sections of `L7.R` for another example of guided, and automatic variable selection, using the AIC, applied to the `swiss` fertility data.

3.3. **Two model revision strategies.** The following two subsections treat a couple of discrete topics concerning model revision. Suppose that, in the course of our diagnostic analysis we find that the errors are non-normal, or correlated. It may be possible to make a linear transformation of the model to get iid normal mean zero errors ϵ. We transform the data, fit, and then invert the transformation to get results for the original model. This is weighted regression, which we look at in the next part section.

We may find that the response y_i is not linearly related to the linear predictor $X\beta$. Can we find a transform which restores linearity? It would be convenient to take a family of transformations, and choose the transform which best restores linearity. This is the Box-Cox approach, presented in the second part section.

3.3.1. *Weighted Regression.* We get linear normal data with variance varying from observation to observation. If σ_i^2 is the variance for the ith observation, then

$$Y_i = X\beta + \epsilon_i \quad \text{for } i = 1, 2, ..., n$$

as before, but this time the independent errors ϵ_i are distributed $\epsilon_i \sim \mathrm{N}(0, \sigma_i^2)$. There are three cases to consider: σ_i^2 unequal and unknown, σ_i^2 unequal and known, and $\sigma_i^2 = \sigma^2/w_i$, with w_i known, but σ^2 unknown.

The first case cannot be treated without some assumption on the joint distribution of the σ_i, since we have a variance parameter for each datum, which we cannot estimate. The second case can be treated in the same framework as the third (see the 2nd exercise in Problem sheet 1 where σ^2 is known).

What is the motivation for considering the third case? If, for $i = 1, 2, ..., n$, Y_i is actually the outcome of n_i independent measurements, $Y_{i,j} \sim \mathrm{N}(x_i\beta, \sigma^2)$, then we

may pool the data, so $Y_i = n_i^{-1} \sum_j Y_{i,j}$ so that $\text{var}(Y_i) = \sigma^2/n_i$. This works for the more general case where $Y_{i,j}$ are iid for $j = 1, 2, ..., n_i$, with mean $x_i\beta$ and variance $\text{var}(Y_{i,j}) = \sigma^2$, but not normal. By the CLT, the approximation $Y_i \sim \text{N}(x_i\beta, \sigma^2/n_i)$ may be good.

We can map the weighted variance problem onto our original problem. If $W = \text{diag}(w_1, ..., w_n)$ (*ie*, an $n \times n$ matrix with square-roots of w's on the diagonal, and otherwise zero) and we define 'data' $Y' = W^{1/2}Y$ and a 'design' $X' = W^{1/2}X$ then

$$Y' = X'\beta + \epsilon'$$

with $\epsilon' \sim \text{N}(0, \sigma^2 I_n)$. We are back to our standard problem. The weighted least squares estimators for β and σ^2 are $\hat{\beta} = (X^T W X)^{-1} X^T W Y$, and $s^2 = (Y - X\hat{\beta})^T W (Y - X\hat{\beta})/(n - p)$.

Exercise suppose $Y = X\beta + \epsilon$, and the errors are correlated, with $\epsilon \sim \text{N}(0, \Sigma)$ with Σ a $n \times n$ positive definite covariance matrix. Give formulae for the weighted least squares estimators $\hat{\beta}$ and s^2 for β and σ^2. Ans: Suppose there is a transformation L such that $\text{var}(L\epsilon) = \sigma^2 I_n$ for some $\sigma^2 > 0$. Now it is the same as the last problem. Is there such a transformation?

Example 3.3. The monthly prices quoted in Example 2.4 and Example 2.5 are averages. The number of houses used to form the averages are given (as a scatterplot) in the lattice plot in Figure 4. The numbers are quite large (60 is typical) so we might expect the central limit theorem to give us averages which are approximately normally distributed, scattered about a running (straight-line regression) mean. Let $w_i =$ `sales`$[i]$. I return to the house types `Semi-Detached` and `Terraced` and the model of Example 2.4, but weighted,

$$y_k \sim \alpha + \alpha_T g_{k,T} + \gamma_M x_{k,M} + \epsilon_k, \qquad \epsilon_k \sim \text{N}(0, \sigma^2/w_i),$$

and I am looking to remove all the variance variation by house type and month by taking into account the sales weighting. We can fit this model in R using the `weights` option in the `lm()` function. We set `weights` = `sales`, and otherwise the variables are set up as before.

```
> #weighted by number of sales
> ohp.wlm<-lm(price~month+type,weights=sales,data=ohp)
> summary(ohp.wlm)
...
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  157.94603    3.29306  47.963   <2e-16 ***
month          1.70720    0.05509  30.987   <2e-16 ***
typeTerraced  -1.95419    3.06955  -0.637    0.525
...
```

In this fit `Semi-Detached` gives the baseline. So, has our weighting corrected the variance? We have been using R to plot residuals against fitted values, as a check for goodness of fit. Looking at Figure 11. We do see an improvement in the pattern of residuals, so we have a 'better' fit. The residuals and fitted values should be independent and the residuals should have homogeneous variance. There is a funnel-shape structure to the scatter plot at left in Figure 11. The scatter plot at right is somewhat more even. The weighting brings the outlier into the

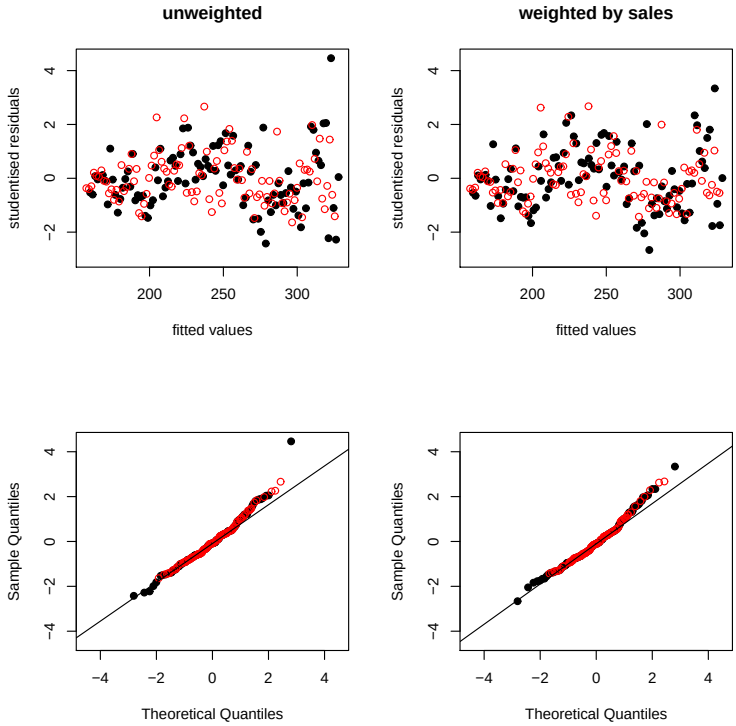

FIGURE 11. LEFT: studentised residuals against fitted values and qqplot for the unweighted fit. RIGHT the fit weighted by sales. Weighting by sales improves the variance of the terraced and semi-detached properties.

normal range (sales have fallen lately - the weighting corrects the recent part of the variance trend - but there is still some increase in variance with fitted values at small weighted fitted values - this is the trend to increased variance with price, which we havnt treated).

Note the scaling on the x and y axes at right in Figure 11. We have $\hat{Y}' = W^{1/2}X\hat{\beta}$ so the weighted residuals $e' = Y' - \hat{Y}' = W^{1/2}(Y - X\hat{\beta})$ have variance matrix $\sigma^2(I_n - H)$ with $H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$. When we make a weighted regression, $\hat{\beta}$ determines the fitted values in both weighted and unweighted coordinate systems. R returns the fitted values $\hat{Y} = X\hat{\beta}$ and residuals $e = (Y - X\hat{\beta})$ in the unweighted coordinates rather than $\hat{Y}'$ and e'. However, the standardised and studentised residuals are the same in the two coordinate systems. We plot the studentised residuals $r'(e)$ against the fitted values $\hat{y}$ since they have the same diagnostic properties (independence and unit variance) as plotting $r'(e')$ against $\hat{y}'$

3.4. **The Box-Cox family of transformations.** Suppose we have data $y, X_1, ..., X_p$ with $y_k \geq 0$. The response may not be a linear function of the linear predictor. A simple transformation is often enough to set things right. The family of transformations $y'_k = (y_k^\lambda - 1)/\lambda$ for $k = 1, 2, ..., n$ can fit in your pocket, and includes

many of the transformations modeling actually leads us to use. The idea is to find a λ value that linearises the data.

Exercise Let $g(z;\lambda) = (z^\lambda - 1)/\lambda$ for $z \geq 0$. Show that $g(z) = 1 - z^{-1}, 2z^{1/2} - 2, z-1, (z^2-1)/2$ and $g(z) \to \log(z)$ for $\lambda = -1, 1/2, 1, 2$ and as $\lambda \to 0$ respectively. Notice that, if $X_1 = 1_{n,1}$ so β_1 is the intercept, then the transform $y'_k = y_k^\lambda$ (as opposed to $y'_k = (y_k^\lambda - 1)/\lambda$) leads to MLE's for the parameters of $y' = X\beta' + \epsilon$ which differ from those above by a scale and shift, $\beta'_1 = \beta_1\lambda + 1, \beta'_i = \lambda\beta_i, i = 2, ..., p$.

If we fit $y' = X\beta + \epsilon$ for $\epsilon \sim \mathrm{N}(0, I_n\sigma^2)$, with $\mathbf{x}_k$ the kth row X, then the likelihood is

$$L(\beta, \sigma^2, \lambda; y') = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_k (y'_k - \mathbf{x}_k^T\beta)^2\right).$$

To get the likelihood for λ in terms of y, we make the change of variables $y'_k = g(y_k;\lambda)$ with Jacobian $g_z(z;\lambda) = z^{\lambda-1}$.

Exercise Make the change of variables, and calculate the log-likelihood, $\ell(\beta, \sigma^2, \lambda; y)$ in terms of y. Show that the MLE's for β and σ^2 are $\hat{\beta}' = (X^TX)^{-1}X^Ty'$ and $\hat{\sigma}^2_{MLE} = (y' - Hy')^T(y' - Hy')/n$, with $H = X(X^TX)^{-1}X^T$ and $y' = y'(y, \lambda)$ (*ie* all as usual, but the 'data' y' depends on λ). Substitute these into the likelihood, and show that the MLE for λ is the argument maximising

$$\ell(\hat{\beta}, \hat{\sigma}^2_{MLE}, \lambda; y) = -\frac{n}{2}\log(\hat{\sigma}^2_{MLE}(y;\lambda)) + (\lambda - 1)\log((y_1y_2...y_n)).$$

In order to maximise $\ell(\hat{\beta}, \hat{\sigma}^2_{MLE}, \lambda; y)$ as a function over λ, we simply evaluate it at a range of values. In the 4th problem sheet you will compute a confidence interval for λ at level α.

Having computed $\hat{\lambda}$ and a confidence interval for λ we usually fix on the nearest readily interpreted λ-value in the interval. We then recompute the fit (for $\hat{\beta}$ and s^2 etc) conditioned on this estimate.

Example 3.4. The putting data records the fraction of successful putts as a function of distance in feet. Gelman and Nolan (2001) model these data. See

`http://www.stat.berkeley.edu/users/nolan/Papers/golfnew.ps`.

We often transform proportion data as $\log(p/(1-p))$ since this (the log-odds) is the link function for a Bernoulli rv. It is a monotone map from $(0,1)$ to $(-\infty, \infty)$. In this case the odds of failure $(1-p)/p$ is the natural object (it increases with distance). A log turns out to be the wrong transform for linearity. Box-Cox gives a linear response.

```
> putts<-data.frame(2:20,
+ c(0.93,0.83,0.74,0.59,0.55,0.53,0.46,0.32,0.34,0.32,
+     0.26,0.24,0.31,0.17,0.13,0.16,0.17,0.14,0.16)
+ )
> names(putts)<-c('Dist','Prop')
> head(putts)
  Dist Prop
1    2 0.93
2    3 0.83
3    4 0.74
```

```
4     5 0.59
5     6 0.55
6     7 0.53
```

The data are plotted in the top left panel of Figure 12. The λ value was estimated

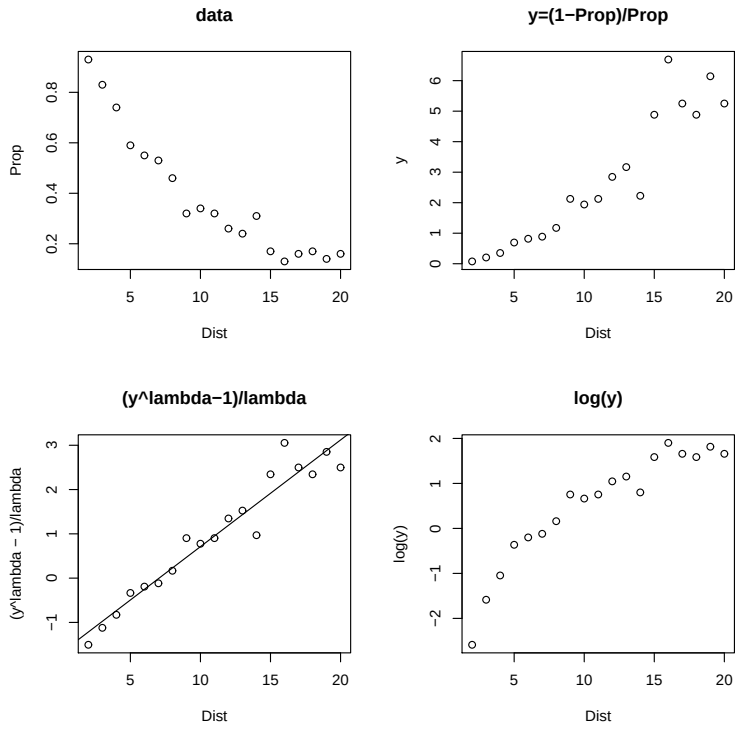

FIGURE 12. Putting, fraction successful puts as a function of putt length in feet. Box-Cox analysis at lower left, otherwise as labeled.

by maximising the likelihood, as above.

```
> y<-(1-putts$Prop)/putts$Prop
> x<-putts$Dist
> putts.bc<-boxcox(y~x)
```

This generates the graph in Figure 13, and from the graph we see that the MLE is at around 0.46 but the CI covers $\lambda = 0.5$, which is easier to deal with in later modeling. We transform the data. If distance is $x =$ `Dist` and the response is $y = (1-p)/p$ for $p =$ `Prop` then we want $\sqrt{y} = \beta_1 + \beta_2 x + \epsilon$ ($2\sqrt{y} - 2 =$ ... is not materialy different).

```
> putts.lm<-lm(sqrt(y)~x)
> summary(putts.lm)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.14342    0.09818   1.461    0.162
```

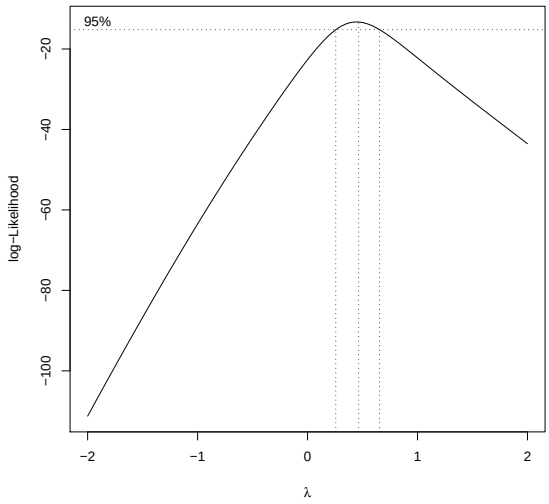

FIGURE 13. Box-Cox analysis of the putts data.

```
x            0.12293    0.00799  15.386 2.07e-11 ***
...
```

The resulting fit is plotted in the lower left panel of Figure 12. The lower right panel shows the log-transformed response, which is clearly not suitable.

Notice that β_1 (the (`Intercept`) row in the `summary`() output) is not significant. Enforcing $\beta_1 = 0$ is natural on physical grounds also, as the odds of failure should go to zero for very short puts. We conclude that the odds of putt-failure increase as the square of the distance.

4. Generalised Linear Models

We have tried just about every possible transformation of the response and explanatory variables in order to get a normal linear model. However, we encounter data for which the response is binary valued, categorical, or in some other way incorrigibly non-normal. How should we set up a model, estimate parameters and confidence intervals, and carry out variable selection? The GLM framework generalises normal linear models in a natural way. Inference for GLM's has some familiar stages.

4.1. **Exponential Families.** A probability mass function for a scalar rv y is an natural linear exponential family of order one if it can be written

$$f(y|\theta) = \exp(y\theta - \kappa(\theta) + c(y))$$

for $y \in \Omega$ the support of f. In this pmf, κ is a normalising constant, since

$$\sum_{y\in\Omega} f(y|\theta) = 1 \quad \Rightarrow \quad \exp(\kappa(\theta)) = \sum_{y\in\Omega} \exp(y\theta + c(y)).$$

The parameter space, $\theta \in \Xi$ is the set

$$\Xi = \{\theta;\ \kappa(\theta) < \infty\}$$

of values which make f a probability. The same definition applies for a probability density, with

$$\exp(\kappa(\theta)) = \int_{y\in\Omega} \exp(y\theta + c(y))\, dy.$$

The function $\kappa(\theta)$ is the cumulant generating function for a probability density proportional to $\exp(c(y))$.

For the record, the exponential family of order $q > 1$ has the form

$$f(y|\theta) = \exp(g(y)^T\theta(w) - \kappa(w) + c(y)),$$

for $y \in \Omega$ possibly a vector, $g : \Omega \to R^q$ a vector of q linearly independent functions, and $\theta(w)$ a vector of q functions of the basic parameter vector w of dimension less than or equal q. The random vector $g(y)$ is called the natural observation for the family. The parameterisation $\theta(w)$ might disallow θ that give finite κ. If the parameterisation is simply $\exp(g(y)^T\theta - \kappa(\theta) + c(y))$, for all θ that make this function a pmf (or pdf) then we have a natural exponential family. If the parameter space Ξ is open, it is regular.

There are many useful general results for inference under exponential family models. Most of the models you know fit within this framework.

Example 4.1. In the Binomial distribution Binomial(m, π) the pmf $\Pr(Y = y)$ is

$$C_y^m \pi^y (1-\pi)^{m-y} = \exp(y \log(\pi/(1-\pi)) + m \log(1-\pi) + \log(C_y^m)).$$

This is a natural exponential family of order one, with natural parameter the log odds $\theta = \log(\pi/(1-\pi))$, natural observation $y \in \Omega = \{0, 1, 2, ...\}$, $\kappa = -\log(1-\pi))$ and $c(y) = \log(C_y^m)$.

Example 4.2. In the Normal distribution $\mathrm{N}(\mu, \sigma^2)$ the pdf $f(y|\mu, \sigma^2)$ is

$$(2\pi\sigma^2)^{-1/2} \exp(-(y-\mu)^2/2\sigma^2) = \exp(y\mu/\sigma^2 - \mu^2/2\sigma^2 - (1/2)\log(2\pi\sigma^2) - y^2/2\sigma^2).$$

If σ^2 is known (so, it is not a parameter indexing members of the family) then this is a natural exponential family of order one, with natural parameter the $\theta = \mu/\sigma^2$,

natural observation $y \in R$, $\kappa = \mu^2/2\sigma^2 - (1/2)\log(2\pi\sigma^2)$ and $c(y) = -y^2/2\sigma^2$. If σ^2 is unknown we have a natural exponential family of order two, with $g(y)^T\theta = (y, -y^2/2)(\mu/\sigma^2, 1/\sigma^2)^T$ which is natural since $\theta \in R \times (0, \infty)$.

4.2. **GLM, setup.** A normal linear model has three parts: a deterministic part $\eta = X\beta$, a stochastic part $Y \sim \mathrm{N}(\mu, \sigma^2)$, and a link between the stochastic and deterministic parts, $\mu = \eta$. This kind of language is overkill for such a simple model.

The GLM has a stochastic part, $Y_i \sim f$, for $i = 1, 2, ..., n$ which we insist has the form

$$f(y_i|\theta_i, \phi) = \exp\left(\frac{y_i\theta_i - \kappa(\theta_i)}{\phi} + c(y_i; \phi)\right).$$

If the scale parameter ϕ is known, this is a natural exponential family (with natural parameter θ/ϕ).

It has a deterministic part $\eta_i = \mathbf{x}_i^T\beta$ (as before $\mathbf{x}_i$ is the $p\times 1$ vector of explanatory variables for the ith response and β is a $p\times 1$ parameter vector). The linear function $\eta_i(\beta)$ is called the linear predictor. If $\eta = (\eta_1, ..., \eta_n)^T$ and X is a $n \times p$ matrix with rows $\mathbf{x}_i^T$, then $\eta = X\beta$.

A GLM has a third part, which links the stochastic and deterministic parts. If $\mu_i = \mathsf{E}(Y_i)$, then $g(\mu_i) = \eta_i$. Here g, the link function, is a smooth invertible function of the mean.

When we find a GLM to modl given data, we make these three modeling choices. The purpose of the link function is to map the linear predictor into the scale of the response. For example, if Y_i is a binary rv, then $\mathsf{E}(Y_i)$ is in fact a probability, so we consider invertible (*ie* monotone) link functions g that map $R \to (0, 1)$.

4.2.1. *Moments.* The mean μ_i is a function of θ_i. Since

$$\int (y/\phi)^n \exp\left(y\theta_i/\phi + c(y; \phi)\right) dy = \frac{d^n}{d\theta_i^n} \exp(\kappa(\theta_i)/\phi)$$

we have

$$\begin{aligned}
\mathsf{E}(Y_i) &= \exp(-\kappa(\theta_i)/\phi) \int y \exp\left(y\theta_i + c(y; \phi)\right) dy \\
&= \exp(-\kappa(\theta_i)/\phi)\phi\frac{\kappa'(\theta_i)}{\phi} \exp(\kappa(\theta_i)/\phi) \\
&= \kappa'(\theta_i),
\end{aligned}$$

so $\mu_i = \kappa'(\theta_i)$. For the variance,

$$\begin{aligned}
\mathrm{var}(Y) &= \mathsf{E}(Y_i^2) - \mathsf{E}(Y_i)^2 \\
&= \exp(-\kappa(\theta_i)/\phi)\phi^2 \frac{d^2}{d\theta_i^2} \exp(\kappa(\theta_i)/\phi) - (\kappa')^2 \\
&= \exp(-\kappa(\theta_i)/\phi)\phi^2 \frac{d}{d\theta_i}(\kappa' \exp(\kappa(\theta_i)/\phi)/\phi) - (\kappa')^2 \\
&= \phi\kappa''(\theta_i),
\end{aligned}$$

and so $\mathrm{var}(Y) = \phi\kappa''(\theta_i)$. Since $\mathrm{var}(Y) > 0$, we have $\kappa''(\theta_i) > 0$, so $d\mu_i/d\theta_i > 0$ and μ_i is a strictly increasing function of θ_i.

We have seen that the mean of Y_i increases with θ_i. Since both $\mu_i = \kappa'(\theta_i; \phi)$ and $g(\mu_i) = \eta_i$ are monotone functions, we must have an invertible relation $\theta_i = \theta(\eta_i)$ between the original parameter θ_i and the linear predictor.

We define a variance function $V(\mu_i)$ relating the variance $\mathrm{var}(Y_i)$ and the mean, μ_i as

$$\mathrm{var}(Y_i) = \phi V(\mu_i).$$

Exercise Show that $\mathrm{var}(Y_i) = \phi\kappa''(\theta_i)$ and hence $V(\mu_i) = \kappa''(\kappa'^{-1}(\mu_i))$.

Exercise Show that $d\theta_k/d\mu_k = 1/V(\mu_k)$.

One particularly simple possibility is that $\theta_i = \eta_i$. This arises when $\kappa'^{-1}(g^{-1}(\eta_i);\phi) = \eta_i$, that is, if $g^{-1}(x) = \kappa'(x)$, corresponding to the canonical choice of link function

$$g(\mu_i) = \kappa'^{-1}(\mu_i).$$

4.3. **Inference for GLM's.**

4.3.1. *Likelihood.* One of the key ideas of modeling with GLM's is that we start with one parameter θ_i for each observation y_i $i = 1, 2, ..., n$. We have knitted them all together by modeling $\theta_i = \theta(\mathbf{x}_i\beta)$ in terms of the set of parameters, $\beta = (\beta_1, ..., \beta_p)^T$. There are now just p parameters (with $p \ll n$ often, as in Section 2) in the GLM. Since $\mathsf{E}(Y_i) = g^{-1}(\mathbf{x}_i\beta)$ is a monotone function of $\mathbf{x}_i\beta$, we are back to the situation where we have p explanatory variables $\mathbf{x}_i = (x_{i,1}, ..., x_{i,p})^T$ for the ith response and, for positive β_j, the mean response Y_i increases as $x_{i,j}$ goes up. Significant explanatory variables generate significant shift in the mean response.

Since the rv Y_i are iid, the log-likelihood for β is

$$\ell(\beta) = \sum_{i=1}^{n} \frac{y_i\theta_i - \kappa(\theta_i)}{\phi} + c(y_i;\phi),$$

with $\theta_i = \theta_i(\beta)$. The MLE for β is the solution to the p equations $d\ell/d\beta_j = 0$. There is no general convenient closed form for the MLE. However, it may be computed numerically, *via* the iteratively re-weighted least squares algorithm.

4.3.2. *Iteratively Re-weighted Lest Squares.* Suppose we want to find $z^* \in R$ such that $f(z^*) = 0$ for some continuously differentiable function $f : R \to R$. The Newton Raphson algorithm is an iteration $z_0, z_1, ...$ which, under certain conditions, converges to z^*. We start with $z = z^{(0)}$ and approximate $f(z) \simeq f(z^{(0)}) + (df/dz|_{z^{(0)}})(z - z^{(0)})$. We seek $z = z^{(1)}$ so that $f(z^{(1)}) = 0$. The local linear approximation for $z^{(1)}$ is $z^{(1)} = z^{(0)} - (df/dz|_{z^{(0)}})^{-1}f(z^{(0)})$, and we can iterate this to improve the approximation,

$$z^{(i+1)} = z^{(i)} - (df/dz|_{z^{(i)}})^{-1}f(z^{(i)}), \quad i = 0, 1, 2, ...$$

In order to solve $df/dz = 0$ numerically, we replace f by df/dz. The N.R. iteration becomes

$$z^{(i+1)} = z^{(i)} - (d^2f/dz^2|_{z^{(i)}})^{-1}df/dz(z^{(i)}), \quad i = 0, 1, 2, ...$$

This sequence may fail to converge. If f is convex then the sequence converges for some z_0 sufficiently close to z^*. The multivariate case $z = (z_1, z_2, ..., z_p)^T$ is given in terms of the Hessian of $f : R^p \to R$. Let $\partial f/\partial z^T = (\partial f/\partial z_1, \partial f/\partial z_2, ..., \partial f/\partial z_p)$

be the row vector of derivatives. Let

$$\frac{\partial^2 f}{\partial z \partial z^T} = \begin{pmatrix} \frac{\partial^2 f}{\partial z_1^2} & \frac{\partial^2 f}{\partial z_2 \partial z_1} & \cdots & \frac{\partial^2 f}{\partial z_p \partial z_1} \\ \frac{\partial^2 f}{\partial z_1 \partial z_2} & & & \frac{\partial^2 f}{\partial z_p \partial z_2} \\ \vdots & & & \vdots \\ \frac{\partial^2 f}{\partial z_1 \partial z_p} & \frac{\partial^2 f}{\partial z_2 \partial z_p} & \cdots & \frac{\partial^2 f}{\partial z_p^2} \end{pmatrix}$$

give the Hessian for $f(z)$. In order to find an extremum we seek $z^* \in R^p$ to solve the p equations $\partial f / \partial z = 0$. The local linear approximation for $\partial f / \partial z$ at $z = z^{(0)}$ is

$$\frac{\partial f}{\partial z} \simeq \left.\frac{\partial f}{\partial z}\right|_{z^{(0)}} + \left.\frac{\partial^2 f}{\partial z \partial z^T}\right|_{z^{(0)}} (z - z^{(0)}).$$

Notice that if f is a quadratic in z then "$\simeq$" is "$=$". If $\partial^2 f / \partial z \partial z^T$ is invertible, then our multivariate Newton Raphson iteration for an extremum of f is

$$z^{(i+1)} = z^{(i)} - \left(\left.\frac{\partial^2 f}{\partial z \partial z^T}\right|_{z^{(i)}} \right)^{-1} \left.\frac{\partial f}{\partial z}\right|_{z^{(i)}}.$$

This converges to z^* in a single step, if f is a quadratic function of z.

We would like to find the maximum of the log-likelihood. In our setting, with ϕ known, we have $\ell = \ell(\beta; y)$. We seek $\hat{\beta}$ to solve $\partial \ell(\hat{\beta}; y) / \partial \beta = 0$. The Newton Raphson iteration for β is

$$\beta^{(i+1)} = \beta^{(i)} - \left(\left.\frac{\partial^2 \ell}{\partial \beta \partial \beta^T}\right|_{\beta^{(i)}} \right)^{-1} \left.\frac{\partial \ell}{\partial \beta}\right|_{\beta^{(i)}}.$$

This iteration might be applied to any twice differentiable log-likelihood. It (*ie* $\beta^{(i)}$) will converge to the MLE if started sufficiently close to the MLE (since the likelihood is quadratic in the neighborhood of the MLE). Notice that, if the log-likelihood is a quadratic function function of the parameter β as is the case for regression with a normal linear model (see Section 2.1) then this iteration must converge in a single step.

The quantity

$$J(y) = -\partial^2 \ell / \partial \beta \partial \beta^T$$

is the observed information, whilst

$$I = -\mathsf{E}(\partial^2 \ell / \partial \beta \partial \beta^T)$$

is the expected information. For a GLM the expected information is often more convenient to work with. It has a special role in the inference, since the difference $\beta - \hat{\beta}$ between the true parameters vector and the MLE becomes normal,

$$(\beta - \hat{\beta}) \xrightarrow{D} \mathrm{N}(0, I^{-1}(\beta)),$$

asymptotically in n. Estimating the variance $I(\beta)$ using $I(\hat{\beta})$ or $J(y)$ are both valid, since both converge to I (see Davison (2003) Section 4.4.2). The revised iteration is

$$\beta^{(i+1)} = \beta^{(i)} + I^{-1} \left.\frac{\partial \ell}{\partial \beta}\right|_{\beta^{(i)}}.$$

For a GLM we can simplify things a bit. In that case

$$\begin{aligned}\frac{\partial \ell}{\partial \beta_i} &= \sum_k \frac{\partial \ell}{\partial \eta_k}\frac{\partial \eta_k}{\partial \beta_i} \\ &= \sum_k \frac{\partial \ell}{\partial \theta_k}\frac{d\theta_k}{d\eta_k}x_{k,i} \\ &= \sum_k \frac{y_k-\mu_k}{\phi}\frac{d\theta_k}{d\mu_k}\frac{d\mu_k}{d\eta_k}x_{k,i} \\ &= \sum_k \frac{y_k-\mu_k}{g'(\mu_k)\phi V(\mu_k)}x_{k,i}.\end{aligned}$$

since $g(\mu_k)=\eta_k$ and using the exercise above for $d\theta_k/d\mu_k$. In vector notation, with and $u=\partial\ell/\partial\eta$, so $u=(u_1,...,u_n)^T$ with $u_k=(y_k-\mu_k)/g'(\mu_k)\phi V(\mu_k)$,

$$\frac{\partial \ell}{\partial \beta} = X^T u.$$

Also,

$$\begin{aligned}\frac{\partial^2 \ell}{\partial\beta\partial\beta^T} &= \frac{\partial}{\partial\beta}\frac{\partial\ell}{\partial\eta^T}X \\ &= X^T\frac{\partial^2\ell}{\partial\eta\partial\eta^T}X,\end{aligned}$$

so the expected information is

$$I = X^T W X,$$

where $W=-\mathsf{E}(\partial^2\ell/\partial\eta\partial\eta^T)$ is a diagonal matrix. We can see it is diagonal, as u_k is a function of μ_k only, and $\mu_k=g^{-1}(\eta_k)$ is a function of η_k only. The diagonal entries are

$$\begin{aligned}W_{kk} &= -\mathsf{E}\left(\frac{\partial^2\ell}{\partial\eta_k^2}\right) \\ &= \frac{1}{g'(\mu_k)^2\phi V(\mu_k)}.\end{aligned}$$

Exercise show that $\mathsf{E}(\partial\ell/\partial\theta_k)=0$ and use it to show that

$$\mathsf{E}\left(\frac{\partial^2\ell}{\partial\eta_k^2}\right) = \mathsf{E}\left(\left[\frac{d\theta_k}{d\eta_k}\right]^2\frac{\partial^2\ell}{\partial\theta_k^2}\right).$$

Now show that

$$-\mathsf{E}\left(\frac{\partial^2\ell}{\partial\theta_k^2}\right) = \mathsf{E}\left(\left[\frac{\partial\ell}{\partial\theta_k}\right]^2\right),$$

and use this result to calculate W_{kk} as above.

Now the Newton Raphson iteration is

$$\beta^{(i+1)} = \beta^{(i)} + (X^T W^{(i)} X)^{-1} X^T u^{(i)},$$

with $W^{(i)}$ and $u^{(i)}$ evaluated at $\beta^{(i)}$. This can be written

$$\begin{aligned}\beta^{(i+1)} &= (X^TW^{(i)}X)^{-1}X^TW^{(i)}(X\beta^{(i)} + (W^{(i)})^{-1}u^{(i)}) \\ &= (X^TW^{(i)}X)^{-1}X^TW^{(i)}z^{(i)}\end{aligned}$$

where $z^{(i)} = (z_1^{(i)}, ..., z_n^{(i)})^T$ is the iterated 'data vector',

$$z_k^{(i)} = [X\beta^{(i)}]_k + g'(\mu_k^{(i)})(y_k - \mu_k^{(i)}).$$

This is iteratively reweighted least squares. If you look back at Section 3.3.1 you will see that if $y = X\beta + \epsilon$ with $\epsilon \sim \mathrm{N}(0, W^{-1})$ then $\hat{\beta} = (X^TWX)^{-1}X^TWy$. The IRLS algorithm for estimation of the MLE for a GLM via the sequence $\beta^{(i)} \to \hat{\beta}$ as $i \to \infty$ is

Start with $\mu^{(0)} = y$ so $X\beta^{(0)} = \eta^{(0)} = g(\mu^{(0)}) = g(y)$, and $z^{(0)} = g(y)$ and $W^{(0)} = \mathrm{diag}(g'(y)^2\phi V(y))^{-1}$. For $i = 0, 1, 2, ...,$

(a) set $\beta^{(i+1)} = (X^TW^{(i)}X)^{-1}X^TW^{(i)}z^{(i)}$ (*ie* $\beta^{(i+1)}$ are the MLE parameter values in the weighted regression of $z^{(i)}$ on X).

(b) $\eta^{(i+1)} = X\beta^{(i+1)}$, $\mu^{(i+1)} = g^{-1}(\eta^{(i+1)})$,

$$z^{(i+1)} = \eta^{(i+1)} + g'(\mu^{(i+1)})(y - \mu^{(i+1)})$$

and

$$W^{(i+1)} = \mathrm{diag}\left(\frac{1}{g'(\mu^{(i+1)})^2\phi V(\mu^{(i+1)})}\right).$$

Exercise Show that if the link function is the canonical link function then $u_k = (y_k - \mu_k)/\phi$, $W_{kk} = V(\mu_k)/\phi$ and

$$z_k = [X\beta]_k + \frac{(y - \mu_k)}{V(\mu_k)}.$$

Exercise Compute these quantities for the case of regression with a normal linear model with σ^2 known. Show that $\phi = \sigma^2$, $V(\mu) = 1$, $W = I_n/\sigma^2$ and $z = y$, and verify that IRLS converges in one step for regression with a normal linear model.

4.3.3. *Variance of MLEs.* Recall that if $\hat{\beta}$ is an MLE then, asymptotically in n,

$$(\beta - \hat{\beta}) \xrightarrow{D} \mathrm{N}(0, I^{-1}),$$

with I the expected information matrix, $I = -\mathsf{E}(\partial\ell/\partial\beta\partial\beta^T)$. We have seen that, for a GLM, $I = X^TWX$. This is handy, as $(X^TWX)^{-1}$ is computed in the course of the IRLS algorithm. We have variances $\mathrm{var}(\hat{\beta}_i) \simeq (X^TWX)^{-1}_{ii}$ which are good at large n, with $\hat{W} = W(\hat{\mu})$ computed in the IRLS.

Exercise Show that if the link function is the canonical link function then

$$\hat{\beta} \xrightarrow{D} \mathrm{N}(\beta, \phi(X^T\mathrm{diag}(V(\mu_1), ..., V(\mu_n))X)^{-1}).$$

Exercise Check this reduces to something familiar for the case of regression with a normal linear model.

This gives us a test for $\beta_i = 0$, the significance of a single GLM parameter, as

$$\frac{\hat{\beta}_i}{\sqrt{I_{ii}^{-1}}} \xrightarrow{D} N(0,1)$$

gives us

$$\frac{\hat{\beta}_i}{\sqrt{\hat{I}_{ii}^{-1}}} \xrightarrow{D} N(0,1)$$

at large n, with $\hat{I}_{ii}^{-1}$ estimated using $\hat{I} = I(\hat{\mu})$ or $\hat{I} = J(y)$. The `std.dev`, Z-value and p-value reported by R in the following GLM, logistic regression, are computed using the approximate standard normal distribution for $Z = \hat{\beta}_i / \sqrt{(X^T W X)_{ii}^{-1}}$ with W the final weight matrix of the converged IRLS iteration.

4.4. **Logistic regression.**

Example 4.3. *Model setup.* The Challenger data give O-ring failures as a function of temperature.

```
> #From Casella and Berger Statistical Inference (2002)
> #Challenger O-ring failures (fail=1, OK=0, temp in deg. F)
> ch.dat<-data.frame(
+ fail=c(1,1,1,1,0,0,0,0,0,0,0,0,1,1,0,0,0,1,0,0,0,0,0),
+ temp=c(53,57,58,63,66,67,67,67,68,69,70,70,70,70,72,73,75,75,76,76,78,79,81))
> head(ch.dat)
  fail temp
1    1   53
2    1   57
3    1   58
4    1   63
5    0   66
6    0   67
```

Now if $y_i =$ `fail[i]` and $\mathbf{x}_i = (x_{i,1}, x_{i,2}) = (1,$ `temp[i]`$)$ (so we have an intercept), then we model $y_i \sim$ Bernoulli(π_i) with π_i the probability for failure a function of temperature. The observation model for y_i has pmf

$$\pi_i{}^{y_i}(1-\pi_i)^{1-y_i} = \exp(y_i \log(\pi_i/(1-\pi_i)) + \log(1-\pi_i)),$$

so $\theta_i = \log(\pi_i/(1-\pi_i))$ is the natural parameter, $\phi = 1$ and $\kappa(\theta_i) = \log(1+e^{\theta_i})$. Since $\mathsf{E}(Y_i) = \pi_i$, we have $\mu_i = \pi_i$.

The linear predictor is $\eta_i = \beta_1 + \beta_2 x_{i,2}$. We need to link this to the mean. The logistic link

$$\log(\mu_i/(1-\mu_i)) = \eta_i$$

is natural here, since this is equivalent to

$$\pi_i = \frac{\exp(\beta_1 + \beta_2 x_{i,2})}{1 + \exp(\beta_1 + \beta_2 x_{i,2})}.$$

The probability for failure is a logistic function of temperature. As we vary β_1 and β_2 we move the threshold for failure to different temperatures, and change how sharply the onset occurs. We can thereby choose β_1 and β_2 to fit the binary responses (see Figure 14 below). This is *logistic* regression, modeling the probability for success for a binary response as logistic function of a linear predictor.

Since $\mu_i = \kappa'(\theta_i)$ and $\kappa'(\theta_i) = e^{\theta_i}/(1+e^{\theta_i})$, we have

$$\kappa'^{-1}(\mu_i) = \log(\mu_i/(1-\mu_i)),$$

so the link function $g(z) = \log(z/(1-z))$ that we chose when we set up the logistic regression was in fact the canonical link. The question "is temperature explanatory for O-ring failure" is answered by a test for $\beta_2 \neq 0$ significant (in fact the conjecture is that failure probability increases with decreasing temperature, so the test will check for $\beta_2 < 0$ significant).

Example 4.4. *Fitting a logistic model.* We now fit the logistic regression model we proposed in to the Challenger data. R uses IRLS to compute $\hat{\beta}$ and reports the number of (score) iterations it took to get convergence.

The log-likelihood for β in the example above, with our choice of the canonical link, is

$$\ell(\beta) = \sum_{i=1}^{n} y_i \eta_i - \log(1+e^{\eta_i})$$

with $n = 23$ and $\eta_i = \beta_1 + \beta_2 x_{i,2}$. Fitting,

```
> ch.glm<-glm(fail~temp, data=ch.dat, family=binomial())
> summary(ch.glm)

Call:
glm(formula = fail ~ temp, family = binomial(), data = ch.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0611  -0.7613  -0.3783   0.4524   2.2175

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  15.0429     7.3786   2.039   0.0415 *
temp         -0.2322     0.1082  -2.145   0.0320 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.267  on 22  degrees of freedom
Residual deviance: 20.315  on 21  degrees of freedom
AIC: 24.315

Number of Fisher Scoring iterations: 5
```

we see that the probability for failure decreases with increasing temperature. The number of scoring iterations is the number of iterations of of the IRLS algorithm we described above. We will define residual deviance shortly.

The values of the probability for failure at the observation points $\mathbf{x}_i$ are in effect the "fitted values", $\hat{\mu}_i = g^{-1}(\mathbf{x}_i\hat{\beta})$, rather like $\hat{y}_i$ in our normal linear model. Since

$\hat{\pi}(\eta) = \pi(\hat{\eta})$, we can evaluate the function

$$\hat{\pi}(t) = \frac{\exp(\hat{\beta}_1 + \hat{\beta}_2 t)}{1 + \exp(\hat{\beta}_1 + \hat{\beta}_2 t)},$$

to get the estimated probability for failure as a function of temperature. The `predict()` function takes the `glm()` output, and some $\mathbf{x}$ values and returns $\mathbf{x}^T\hat{\beta}$. With the option `type='response'` we get $g^{-1}(\mathbf{x}_i^T\hat{\beta})$, or in this case $\hat{\pi}(\mathbf{x}_i\hat{\beta})$. We can evaluate this function at other $\mathbf{x}$-values using the `newdata` option. We take a sequence of values of the explanatory variable `temp`, and plot π against temperature in Figure 14.

```
> ch.prob<-predict(ch.glm,newdata=data.frame(temp=40:90),type='response')
> plot(40:90,ch.prob,type='l'); #the solid line
```

and I omit the R for the plotting of the points, see `L10.R`.

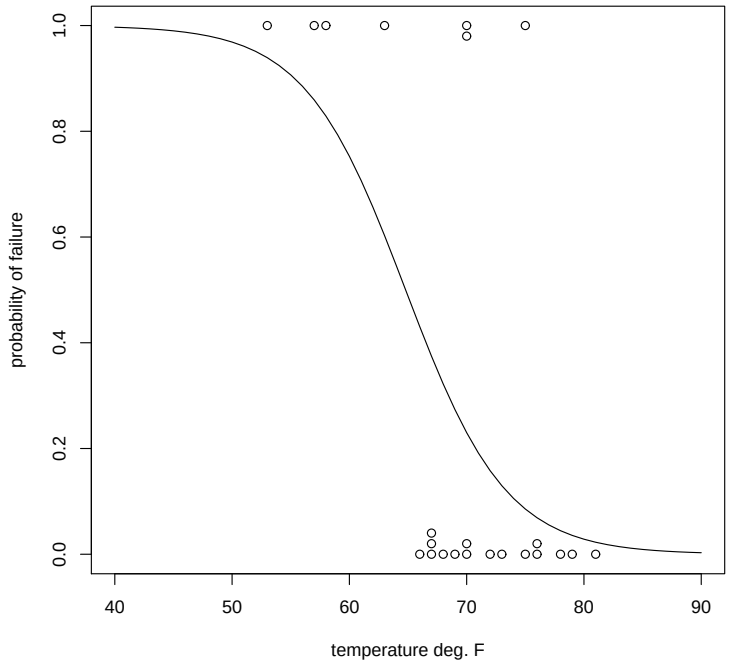

FIGURE 14. O-ring failure data (points) with fitted probability for failure (curve).

The p-value for $\beta_2 = 0$ is 0.032, supporting the view that temperature is explanatory for O-ring failure. How is this computed? The reported standard errors are $\mathrm{var}(\hat{\beta}_i) \simeq (X^T\hat{W}X)^{-1}_{ii}$ for $i = 1, 2, ..., p$. The quoted Z value is $\hat{\beta}_i/\sqrt{(X^T\hat{W}X)^{-1}_{ii}}$ which has an approximate $N(0,1)$ distribution. The p-value for the test for $\beta_2 < 0$ is 0.016.

Exercise Verify that $V(\mu_i) = \kappa''(\theta_i)/\phi = \mu_i(1-\mu_i)$, and hence

$$\hat{W} = \text{diag}(\hat{\mu}_1(1-\hat{\mu}_1), ..., \hat{\mu}_n(1-\hat{\mu}_n))$$

is the last weight matrix of the IRLS. Check the quoted std errs. Ans:

```
> # check we understand what the reported std errs are
> p<-predict(ch.glm,type='response')
> X<-model.matrix(fail~temp, data=ch.dat)
> sqrt(diag(solve(t(X)%*%diag(p*(1-p))%*%X)))
(Intercept)        temp
  7.3786364   0.1082365
```

Example 4.5. *Odds and the odds ratio.* The quantity $\pi/(1-\pi)$ is the odds of success (O-ring failure, but Bernoulli success, *ie* $Y = 1$). For a single generic binary response Y with explanatory variables $\mathbf{x}$ with $\mu = \mathsf{E}(Y)$ and $\pi = \Pr(Y = 1) = \mu$, the canonical link $g(\mu) = \eta$ for logistic regression gives $\log(\pi/(1-\pi)) = \mathbf{x}^T\beta$, and so the odds of success, $O = \pi/(1-\pi)$ say, are estimated as $\hat{O} = O(\mathbf{x}^T\hat{\beta})$, so we have $\hat{O} = \exp(\mathbf{x}^T\hat{\beta})$. This gives the interpretation of $\hat{\beta}_i$ in logistic regression: $\hat{\beta}_i$ gives the change in log odds for success (the outcome $Y = 1$) when the corresponding explanatory variable changes $x_i \to x_i + 1$. For example, for the O-ring data, if $\mathbf{x} = (x_1, x_2)^T$ and $\mathbf{x}' = (x_1, x_2 + 1)$, so the temperature goes up a degree, and $\hat{O} = O(\mathbf{x}^T\hat{\beta})$ and $\hat{O}' = O(\mathbf{x}'^T\hat{\beta})$, then

$$\frac{\hat{O}}{\hat{O}'} = \exp(\hat{\beta}_2).$$

In this example we estimated $\hat{\beta}_2 \simeq -0.2321627$ so the odds of O-ring failure go down by around $\exp(-0.2321627) \simeq 0.7928171$ for an increase in the temperature by one deg. F.

On the day of the disaster it was unusually cold, 31 deg. F. What is our predicted probability for O-ring failure that day?

```
> #launch at temp=31 deg F - estimate for the probability of failure:
> predict(ch.glm,newdata=data.frame(temp=31),type='response')
[1] 0.9996088
>
```

The estimated odds of O-ring failure are 1:1 at a temperature around 64.79465 deg F (since $\hat{O} = 1$ when $\hat{\beta}_1 + \hat{\beta}_2 x_2 = 0$ and solve for x_2). How much higher are the odds of failure at 31 def F? This must be $0.9996088/(1-0.9996088) \simeq 2555$, or $\exp(\beta_2)^{\Delta\text{temp}}$ which is $0.7928171^{(31-64.79465)}$.

Example 4.6. Binomial data in a 2 way table The data in Table 3 are taken from Dr Lunn's 2007 lecture notes, and give the number of men and women smokers and non-smokers in a particular industry. Table 3 gives smoking information for 2916 men and 2503 women. In a binomial model with p the probability for a man to be a smoker and q the probability for a woman to be a smoker, the MLE's for p and q are $\hat{p} = n_{11}/(n_{11}+n_{12}) \simeq 0.71$ and $\hat{q} = n_{21}/(n_{21}+n_{22}) \simeq 0.45$. How much higher are the odds for a man to be a smoker (smoker given man) than for a woman to be a smoker (smoker given woman)? The MLE for the odds ratio is a function of the MLEs for p and q, *ie* $[\hat{p}/(1-\hat{p})]/[\hat{q}/(1-\hat{q})] = (n_{11}/n_{12})/(n_{21}/n_{22}) \simeq 2.92$.

We will use these data to illustrate the Binomial GLM. We have two observations, $y_1 = 2059$ 'successes' from $m_1 = 2916$ trials and $y_2 = 1130$ successes from $m_2 =$

	Smoking		
Sex	Y	N	Tot
M	2059	857	2916
F	1130	1373	2503
Tot	3189	2230	

TABLE 3. Two way table for smoking and gender, giving the number of subjects in each category.

2503 trials. Let π_i give probability for probability for 'success' (here, smoker) for the ith group of trials (so $\pi_1 = p$ and $\pi_2 = q$). The likelihood for a single observation is

$$f(y_i|m_i) = \exp(y_i \log(\pi_i/(1-\pi_i)) + m_i \log(1-\pi_i) + \log(C^{m_i}_{y_i})).$$

The natural observation is y_i, the number of successes in the ith group of trials, the natural parameter is again $\theta_i = \log(\pi_i/(1-\pi_i))$. $\kappa(\theta_i) = m_i \log(1+e^{\theta_i})$, and $\phi = 1$.

We are interested in the explanatory variable 'Sex'. Let $x_{1,2} = 1$ and $x_{2,2} = 0$, so x_2 is the dummy indicator variable for the `M` level of the categorical variable `Sex`. Level `F` is the baseline. The linear predictor is $\eta_k = \beta_1 + \beta_2 x_{k,2}$ and the design matrix is

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}.$$

As for the binary data, model the probability as a logistic function of the indicator variable. Here $\pi_k = \mu_k/m_k$, so

$$\log(\mu_k/(m_k - \mu_k)) = \beta_1 + \beta_2 x_{k,2}.$$

Since $\theta_k = \eta_k$, this is again the cannonical link function, with

$$\begin{aligned} \pi_1 &= \exp(\beta_1+\beta_2)/(1+\exp(\beta_1+\beta_2)) \\ \pi_2 &= \exp(\beta_1)/(1+\exp(\beta_1)). \end{aligned}$$

We can fit these data in R using `glm()` again.

```
> # smoking data from Dunn 07
> smk<-data.frame(S=c(2059,857),NS=c(1130,1373),Sex=c('M','F'))
> # response is two columns '#successes' S and '#failures' NS
> smk
     S   NS Sex
1 2059 1130   M
2  857 1373   F
>
> #The design matrix has an intercept and an indicator for Sex==M
> model.matrix(cbind(smk$S,smk$NS)~Sex,data=smk)
  (Intercept) SexM
1           1    1
2           1    0
>
> #fit response as binomial with Sex explanatory
> smk.glm<-glm(cbind(smk$S,smk$NS)~Sex,data=smk,family=binomial)
```

```
> summary(smk.glm)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.47132    0.04353  -10.83   <2e-16 ***
SexM         1.07132    0.05715   18.75   <2e-16 ***
...
```

The estimated change in the log odds when we move from `F` to `M` (so $x_{k,2} \to x_{k,2}+1$) is $\hat{\beta}_2 = 1.07$ which is an increase by a factor of $\exp(1.07132) \simeq 2.92$.

4.5. **Model choice.** We have seen how to test for the significance of a single GLM parameter β_i, using the asymptotically standard normal distribution of $Z = \hat{\beta}_i / \sqrt{(X^T W X)^{-1}_{ii}}$.

How do we test for the significance of a group of variables? We use the likelihood ratio test. For regression with the normal linear model we managed to get an exact test, by writing the LRT statistic as a monotone function of a statistic, F, with a distribution we have exactly. For regression with GLM's, we stick to the LRT statistic, and accept a test based on the asymptotic distribution of this statistic.

There are two important models 'above' and 'below' our model. The saturated model unlinks the θ_i. They were bound by the constraint $g(\mu_i) = \mathbf{x}_i^T \beta$ on the means. In the saturated model we have one parameter θ_i for each response y_i. The MLE $\hat{\theta}_i^{(s)}$ for θ_i in the saturated model is just $\hat{\theta}_i^{(s)} = \arg\max_\theta f(y_i|\theta_i)$.

Exercise Suppose $Y_i \sim \text{Binomial}(\pi_i, m_i)$. Show that $\hat{\pi}_i^{(s)} = y_i/m_i$ so that $\hat{\theta}_i^{(s)} = \log(y_i/(m_i - y_i))$ and the log-likelihood for the saturated model is

$$\ell(\hat{\theta}_i^{(s)}; y) = \sum_i y_i \log(y_i/(m_i - y_i))) + m_i \log(1 - y_i/m_i).$$

The other model of interest is the null model, in which $\beta_2 = ... = \beta_p = 0$, with $g(\mu_i) = \beta_1$ so we just have the intercept in the linear predictor. The means are all equal so there is a single common natural parameter θ. Let $\hat{\theta}^{(0)}$ be the MLE for the natural parameter of the null model.

Exercise Show that, for the binomial model, $\hat{\pi}^{(0)} = (\sum_i y_i)/(\sum_i m_i) = \bar{y}/\bar{m}$ so that $\hat{\theta}^{(0)} = \log(\bar{y}/(\bar{m} - \bar{y}))$ and the log-likelihood for the null model is

$$\ell(\hat{\theta}^{(0)}; y) = n\bar{y} \log(\bar{y}/(\bar{m} - \bar{y}))) + n\bar{m} \log(1 - \bar{y}/\bar{m}).$$

4.5.1. *Deviance.* The *scaled* deviance $D(y)$ for our GLM is simply related to the log-likehood,

$$D(y) = -2\ell(\hat{\beta}; y) + 2\ell(\hat{\theta}^{(s)}; y).$$

This is of course the LRT statistic for a test comparing the saturated model with the GLM of interest. The deviance itself is the scaled deviance at scale parameter $\phi = 1$. Since the parameter space of $\theta^{(s)}$ includes the parameter space of $\theta(\beta)$ as a subspace, $D(y) \geq 0$. The null deviance is

$$D^{(0)} = -2\ell(\hat{\theta}^{(0)}; y) + 2\ell(\hat{\theta}^{(s)}; y).$$

4.5.2. *Goodness of fit.* Since D is the LRT statistic for a test with null parameter space of dimension p and alternative of dimension n, we expect $D(Y) \sim \chi^2(n-p)$ approximately, under the hypothesis that our GLM model includes all the factors generating variation in the response. If $D(y)$ is large on the scale of a $\chi^2(n-p)$ rv, then we question our model.

This model check is not always applicable. The problem is that $D(Y) \sim \chi^2(n-p)$ holds asymptotically, yes, but not asymptotically in n. We achieve the asymptotic distribution for a LRT statistic as our MLE's converge to their limiting values. Since, in the saturated model, we have one parameter for each observation, increasing n doesn't add to the precision of our estimates of these parameters (increasing n does increase the precision of our estimate of β, but that is just half the LRT statistic).

So, when is this distributional assumption good? Suppose we had multiple replicates for each observation $y_{i,j}$, $j = 1, ..., m_i$, for given explanatory variables $\mathbf{x}_i$, and, in the saturated model, just the one parameter $\theta_i^{(s)}$ for each batch of m_i replicates. Now $\hat{\theta}_i^{(s)} \to \theta_i^{(s)}$ as $m_i \to \infty$. This is the limit where the approximation $D(Y) \sim \chi^2(n-p)$ holds good. So, where do we have replicates?

Exercise Show that a binomial response $Y_i \sim \text{Binomial}(m_i, \pi_i)$ is m_i replicates of a Bernoulli response, and hence that $\hat{\theta}_i^{(s)}$ converges to $\log(\pi_i^{(s)}/(1-\pi_i^{(s)}))$ and $D(Y) \xrightarrow{D} \chi^2(n-p)$ as $m_i \to \infty$ for each $i = 1, 2, ..., n$.

Bernoulli data has none of this structure. In this case $D(Y) \sim \chi^2(n-p)$ will not hold and any inference that depends on this approximation (residual analysis below) is not useful.

4.5.3. *Model choice.* A test comparing two nested models (Q) with dimension q and (P) with dimension $p < q$ has LRT statistic $\Lambda = D^{(P)}(y) - D^{(Q)}(y)$, so

$$D^{(P)}(y) - D^{(Q)}(y) \ \sim \ \chi^2(q-p).$$

The test for no relation between the mean response μ_i and the explanatory variables $\eta_i = \mathbf{x}_i\beta$ in the model (Q) is the test for $\beta_2 = ... = \beta_p = 0$ (if β_1 is the intercept) has test statistic

$$D^{(0)}(y) - D^{(Q)}(y) \ \sim \ \chi^2(q-1).$$

All the remarks we made about model choice for a LM, in Section 3.2 apply equally to GLMs. The problem has a similar structure. We want to check that we have all the important variables in our model (using a Goodness of fit test, as above), and then check we have no redundant variables. The tools available which help us organise the search for a set of explanatory variables are like those for LM's. We can make forwards addition and backwards elimination, one variable at a time. We can identify a reduced model by these or physical considerations, and test to drop the corresponding complementary group of variables. We can make automatic model choice using the AIC.

Example 4.7. In the following example, quoted from Venables and Ripley (2002), based on data from Collet (1991, see V&R for reference), 12 batches of 20 tobacco budworm moths were exposed for 3 days to different dose levels of a toxin. The numbers dead or disabled were recorded. The data are displayed in Table 4. We have a categorical variable `sex` $\in \{$`M`, `F`$\}$ and a real ordinal variable `dose`, which we

	Dose					
	1	2	4	8	16	32
	Mortality					
Male	1	4	9	13	18	20
Female	0	2	6	10	12	16

TABLE 4. Tobacco budworm mortality data, from Collect 1971 *via* Venables and Ripley (2002).

will code as `ldose` $= \log_2($`dose`$)$, as suggested by the scale of the variable. We look for an effect due to `ldose` (fairly obvious!) and allow for possibly different intercepts and slopes in the linear predictors for the two sexes. Our linear predictors are, for $i = 1, 2, ...n$ with $n = 12$,

$$\eta_i = \beta_1 + \beta_2 g_{M,i} + \beta_3 x_{d,i} + \beta_4 g_{M,i} x_{d,i}$$

with $g_{M,i} = 1$ if `sex`[i] $= M$ and zero otherwise, and $x_{d,i} =$ `ldose`[i]. If y_i are the cell counts, then R will fit for the scaled response y_i/m_i (with $m_i = 20$ for $i = 1, 2, ..., n$ here). The expected response is then $\mu_i = \mathsf{E}(Y_i/m_i)$ which is $\mu_i = \pi_i$ with π_i the probability of success in the ith Binomial trial, so $y_i \sim \text{Binomial}(m_i, \pi_i)$. In the notation of Example 4.6, we have the logistic link function

$$\log(\mu_i/(1 - \mu_i)) = \eta_i$$

modeling the proportion μ_i of 'successes', which are Moth-deaths in this case.

```
> #Example from Venables and Ripley (2002) Sec 7.2
> ldose<-rep(0:5,2) #the log-dose
> numdead<-c(1,4,9,13,18,20,0,2,6,10,12,16)
> sex<-c(rep("M",6),rep("F",6))
>
> #gather data in a data-frame
> #bw$sex is categorical with levels M,F
> bw<-data.frame(numdead,numalive=20-numdead,ldose,sex)
>
> #to use glm we have some options how to enter the response
> #here we are using the [#successes,#failures] format
>
> #logistic regression
> bw.glm<-glm(cbind(numdead,numalive)~sex*ldose,data=bw,family=binomial)
> summary(bw.glm)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.9935     0.5527  -5.416 6.09e-08 ***
sexM          0.1750     0.7783   0.225    0.822
ldose         0.9060     0.1671   5.422 5.89e-08 ***
sexM:ldose    0.3529     0.2700   1.307    0.191
...
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 124.8756  on 11  degrees of freedom
```

```
Residual deviance:   4.9937  on  8  degrees of freedom AIC: 43.104

Number of Fisher Scoring iterations: 4
```

Are we missing sources of variation? The test for goodness-of-fit is $D \sim \chi^2(n-p)$ with $n = 12$ and $p = 4$. Since $m_i = 20$ is acceptably large, we can expect the asymptotics for D to hold good. Comparing $D(y) = 4.9937$ to a $\chi^2(8)$ random variable,

```
> 1-pchisq(4.9937,8)
[1] 0.7582493
```

we get a $p - value$ of around 0.76, so we see no evidence for misfit in this respect. Can we drop any terms from this model. The two variables `sexM` and `sexM:ldose` are obvious candidates. In order to get the deviance for the reduced model, we refit using the reduced model.

```
> bw.glm<-glm(cbind(numdead,numalive)~ldose,data=bw,family=binomial)
> summary(bw.glm)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.7661     0.3701  -7.473 7.82e-14 *** ldose
1.0068     0.1236   8.147 3.74e-16 ***
...
Residual deviance:  16.984  on 10  degrees of freedom AIC: 51.094
```

The change in the deviance on dropping `sex` and `sex:ldose` is $16.984 - 4.9937 = 11.9903$ with $n = 12, q = 4, p = 2$ and since $D' - D \sim \chi^2(q-p)$ under the null model without `sex` and `sex:ldose`, the test to drop the two variables has p-value

```
> 1-pchisq(11.99,2)
[1] 0.002491177
```

strong evidence for some role for at least one of these variables.

As Venables and Ripley (2202) point out, this p-value might be a bit surprising given the apparent low significance of the individual variables. As they note, the fact that the offset β_2 in the intercept is not significant is not surprising. This is the difference in mean response between sexes at `dose` $= 1$ (so `ldose` $= 0$) and there are few deaths on either side at low dose. The intercept and slope offset estimates are quite significantly (anti) correlated (inspect $I^{-1} =$ `summary(bw.glm)$cov.scaled` for detail) so we have here the kind of significance-masking due to near-linear dependence that we have discussed for linear models. If we shift the intercept to a central dose value, say `dose` $= 8$ or `ldose` $= 3$ we reduce this correlation (without changing the meaning of the other parameters). This can be thought of as a step towards orthogonalising the original variables g_M in order to reduce correlation.

```
> bw.glm<-glm(cbind(numdead,numalive)~sex*I(ldose-3),data=bw,family=binomial)
> summary(bw.glm)
...
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.2754     0.2305  -1.195  0.23215
sexM                1.2337     0.3770   3.273  0.00107 **
I(ldose - 3)        0.9060     0.1671   5.422 5.89e-08 ***
```

```
sexM:I(ldose - 3)   0.3529     0.2700   1.307  0.19117

    Null deviance: 124.8756  on 11  degrees of freedom
Residual deviance:   4.9937  on  8  degrees of freedom AIC: 43.104

Number of Fisher Scoring iterations: 4
```

We can now see that there is no evidence for different slopes for the two sexes, but the intercepts are not equal. Venables and Ripley (2002) go on to test for curvature, looking for dependence on $(\texttt{ldose} - 3)^2$.

Note that we can form an analysis of deviance table for a GLM fit. Like the default R ANOVA table, this looks at the nested sequence of models obtained by dropping the variables one at a time. The analysis of deviance table gives, in each row, the *increase* in the deviance as the model variables in the corresponding row are removed from the model. For example, here is the analysis of deviance table for the test on the original fit with intercept as $\texttt{ldose} = 0$.

```
> anova(bw.glm,test='Chisq')
Analysis of Deviance Table
          Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                        11    124.876
ldose      1  107.892        10     16.984 2.839e-25
sex        1   10.227         9      6.757     0.001
ldose:sex  1    1.763         8      4.994     0.184
```

The first row gives the null deviance, $D^{(0)}$ (called here the residual deviance). The second row gives the p-value for testing for an effect due to `ldose` in a model with just intercept. The column headed `Deviance` is what we would call the change in deviance, and the column headed `Resid. Dev` is what we would call the deviance. The levels of significance depend on the order in which the variables are added. We can see now that `sex` was signifcant from the first, but in the original `summary()` output its significance was masked by correlation with the slope offset.

4.6. **Further GLM diagnostics.** In order to identify outliers, we can make an analysis of residuals, just as for normal linear models. There are a number of different ways to define residuals for GLM's.

A poorly fitting point will make a large contribution to the deviance. Let

$$\ell_i(\theta; y_i) = \frac{y_i\theta - \kappa(\theta_i)}{\phi} + c(y_i; \phi),$$

and let

$$d_i = -2\ell_i(\hat{\beta}; y_i) + 2\ell_i(\hat{\theta}_i^{(s)}; y_i),$$

so that

$$D(y) = \sum_{i=1}^{n} d_i.$$

Now $d_i \geq 0$ and data with relatively larger d_i are data in greater conflict with the model and the rest of the data. The deviance residuals are defined to be

$$r_i = \text{sign}(y_i - \hat{\mu}_i)d_i.$$

The function $\text{sign}(y_i - \hat{\mu}_i)$ is $+1$ or -1 depending on the sign of $y_i - \hat{\mu}_i$ and shows us whether y_i is in conflict by being relatively too large or too small compared to the fitted mean.

Another way to think about residuals is to consider the misfit in the space of the linear predictor. Since g maps μ to η, we might compare $g(y)$ and $\hat{\eta}$. Expanding $g(y)$ about $y = \mu$, $g(y) \simeq g(\mu) + g'(\mu)(y - \mu)$, an object which appears in our IRLS, $z = \eta + g'(\mu)(y - \mu)$. This gives us an interpretation of z in that algorithm. The residuals $z - \hat{\eta} = g'(\hat{\mu})(y - \hat{\mu})$ are called the working residuals (in R, for example). They are exactly the residuals of the linear regression made in the IRLS. The Hat matrix for that regression is

$$H = W^{1/2} X (X^T W X)^{-1} X^T W^{1/2},$$

and the leverages h_{ii} have the same role as before in the analysis of the working residuals.

The standardised deviance residuals are

$$r'_i = \text{sign}(y_i - \hat{\mu}_i) \frac{d_i}{\sqrt{1 - h_{ii}}}.$$

It can be shown that if the original approximation $D(y) \sim \chi^2(n - p)$ is good then the r' usually have roughly unit variance and distributions close to normal. This leads to a check for misfit.

Example 4.8. For the budworm moth analysis of Example 4.7, the standardised deviance residuals are shown in Figure 15. These are given for the final model, with equal slopes for the two sexes.

```
> bw.glm<-glm(cbind(numdead,numalive)~sex+I(ldose-3),data=bw,family=binomial)
> summary(bw.glm)
...
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.10540  -0.65343  -0.02225   0.48471   1.42944

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.2805     0.2431  -1.154  0.24854
sexM           1.1007     0.3558   3.093  0.00198 **
I(ldose - 3)   1.0642     0.1311   8.119  4.7e-16 ***
...
    Null deviance: 124.876  on 11  degrees of freedom
Residual deviance:   6.757  on  9  degrees of freedom
AIC: 42.867

Number of Fisher Scoring iterations: 4
> plot(fitted.values(bw.glm),rstandard(bw.glm))
```

There is no evidence for misfit, as there are no large (compared to one) standardised deviance residuals. There is no obvious sign of a trend in the residuals. This is often a sign of a problem with the link function (see problem 2 of PS5 - for an example - try computing residuals for the log-link, and compare them with the residuals for the sqrt link).

4.7. **Dispersion and the scale parameter.** When we model using a GLM based on a one parameter family of distributions, such as Binomial and Poisson distributions, the scale parameter ϕ is equal one. When we model using other distributions

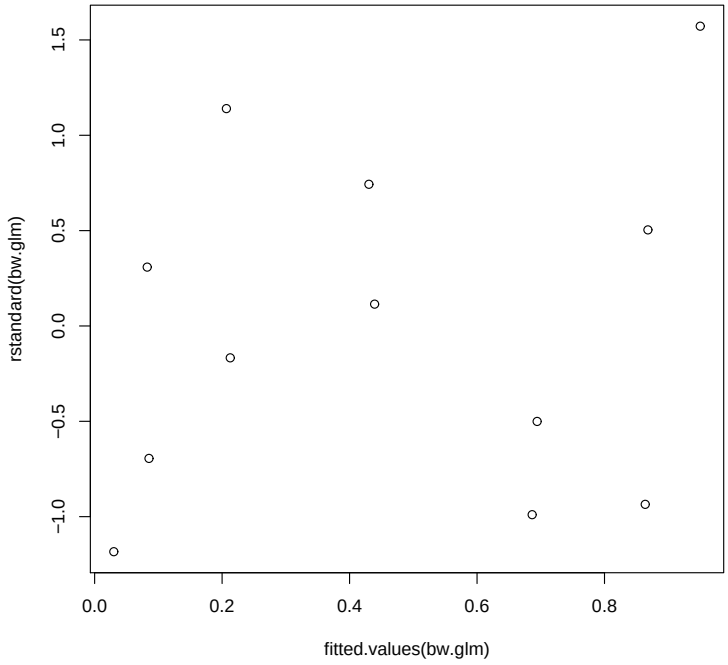

FIGURE 15. Standardised residuals for the final model for the budworm data of Example 4.7.

(Gamma, Normal *etc*) ϕ is not one, and may be unknown. In these cases we might allow for over-dispersion, and extend the model, allowing $\phi \neq 1$. We need an estimate for ϕ.

Notice that the IRLS algorithm doesnt need ϕ! It follows that, operationally, we can fit our GLM at $\phi = 1$, and then look for estimates of ϕ based on the fitted values of $\hat{\beta}$.

Exercise verify that ϕ cancels out in each step of the IRLS squares iteration.

Since $\text{var}(Y_i) = \phi V(\mu_i)$ we can form an estimate

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)},$$

with $\hat{\mu}_i = g^{-1}(\mathbf{x}_i^T \hat{\beta})$.

When we fit a simple model with $\phi = 1$, or otherwise known, and we suspect over- (or who knows? under-) dispersion, we may wish to compute $\hat{\phi}$. If this differs from one, the estimated confidence intervals for $\hat{\beta}$, which depend on ϕ, as in Section 4.3.3, will be wrong, and the estimated variance should now be multiplied by $\hat{\phi}$.

The picture cannot be so simple for Binomial and Poisson models. We can't simply scale the $y\theta - \kappa(\theta)$ term by ϕ in these models and expect the exponential family to remain the same. The requirement that the likelihood be normalised over y (ie, $\int \exp\left(y\theta/\phi + c(y;\phi)\right) dy = \exp(\kappa(\theta_i)/\phi)$) is no longer satisfied within the original family. If we define $\phi = \mathrm{var}(Y)/V(\mu)$, yielding the estimator given above, then we can get an ad-hoc adjustment for over-dispersion.

Exercise see related exercise on cloth fault data in PS6.

In either of these cases (ie when ϕ is a parameter of the likelihood, and when it is an ad-hoc scale parameter adjusting the variance of estimates), the test for model comparison based on the deviance is adjusted. If ϕ is estimated then the scaled deviance is the ratio of the deviance at $\phi = 1$ divided by $\hat{\phi}$. Since these quantities have distributions which are approximated by χ^2 distributions you will often see an F-statistic used in GLM analysis to carry out this test. [End L14]

Statistics Department
E-mail address: nicholls@stats.ox.ac.uk