

BS1A APPLIED STATISTICS - LECTURES 1-4

GEOFF NICHOLLS

Date: 16 lectures, MT10.

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 MT10 course are: Simulation based inference, Practical aspects of linear models, and Logistic regression and generalized linear models.

Reading (Michaelmas Term):

A.C. Davison, *Statistical Models*, CUP (2003)

J.J. Faraway, *Linear Models with R*, Chapman & Hall/CRC (2005)

S.M. Ross, *Simulation*, Elsevier, 4th ed., (2006)

Further Reading:

D. Lunn, Lecture notes http://www.stats.ox.ac.uk/~dlunn/BS1_05/BS1_mt05.htm

J.J. Faraway, *Extending the Linear Model with R*, Chapman & Hall/CRC (2006)

W.N. Venables and B.D. Ripley, *Modern Applied Statistics with S*, Springer (2002)

A.J. Dobson, *An Introduction to Generalized Linear Models*, Chapman and Hall (1990)

1.1. **Website.** Goto the Statistics Department website and access

http://www.stats.ox.ac.uk/current_students/bammath/course_material

Select BS1a Applied Statistics I, or straight to

<http://www.stats.ox.ac.uk/~nicholls/bs1a/>

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

1.2. **Classes.** Bookings for classes will be made in the first lecture. There will be 6 problems classes this term, each lasting for 1 hour. They will be held in weeks 3 to 8. Details of classes will be linked from the website.

1.3. **Assessment and Practicals.** There are three assessed BS1a practicals and one assessed BS1b practical. These contribute 8.5% each of a total 34% practical component in your BS1 mark. The BS1 paper in finals is a 2 hour paper worth 66% of your BS1 mark.

The practical problems for you to solve will be made available in the practical teaching labs. The deadlines for handing in your answers are as follows:

First assessed practical - 12 noon Tuesday, week 8 Michaelmas Term 2010

Second assessed practical - 12 noon, Tuesday, week 2 Hilary Term 2011

Third assessed practical - 12 noon, Tuesday, week 7 Hilary Term 2011

Fourth assessed practical - 12 noon, Friday, week 1 Trinity Term 2011

In order to solve the practical problems you need to learn R. We will have a session introducing R itself, in the lab, in week 3, and then, for each of the four assessed practicals above, a session where we show you some code relevant to the

particular practical. Charalambos Loizides is running these teaching-practicals and will provide information linked from

<http://www.stats.ox.ac.uk/~loizides/BS1/>

Each session is presented on two evenings (in the Isis room in OUCS, at 13 Banbury Road). You may choose which evening you wish to attend. The BS1 practical teaching sessions are held 5:30-7:30 Tuesday and Wednesday evenings, this term in weeks 3,5 and 8 and next term in weeks 2 and 7.

1.4. **The R package.** The R package is available free of charge on the web. Go straight to

<http://cran.r-project.org/>

and download and install R. There is a FAQ, linked from the left margin, full of useful information. Once installed, *An Introduction to R* is a good place to start. Again, from the left margin,

Manuals → An Introduction to R.

1.5. Guide.

- Read the lecture notes and interleave your own notes.
- We will begin by reviewing and extending some of the Part A material on *Linear Models*.
- Linear models lead naturally to looking at *Diagnostics*. Is the model plausible? If not, why not?
- The next topic is *Generalised Linear Models*. In a normal linear model the linear predictor $X^T\beta$ is the mean. We keep this model element, but link it to the mean of some other distribution. This greatly extends the application domain of the linear model. We introduce the theory and look at practical problems.
- The final topic is *Monte Carlo testing*. This is a framework that allows us to test relatively complex models.

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, \dots, x_p . At the i th observation, we set the explanatory variables to values $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,p})$ (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 $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T$,

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

with $\epsilon_i \sim N(0, \sigma^2)$ iid normal *errors* for $i = 1, 2, \dots, n$. If $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)^T$ then $y_i = \mathbf{x}_i \beta + \epsilon_i$. If $y = (y_1, y_2, \dots, y_n)^T$, 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 , $j = 1, 2, \dots, p$ for the columns of $X = (X_1, X_2, \dots, X_p)$. 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 ,

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

omitting the $i = 1, 2, \dots, n$ 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 , $i = 1, 2, \dots, n$ 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 ~ 1 + Nicotine + Tar + Weight` the response is `CO` and all the other variables (including intercept) are

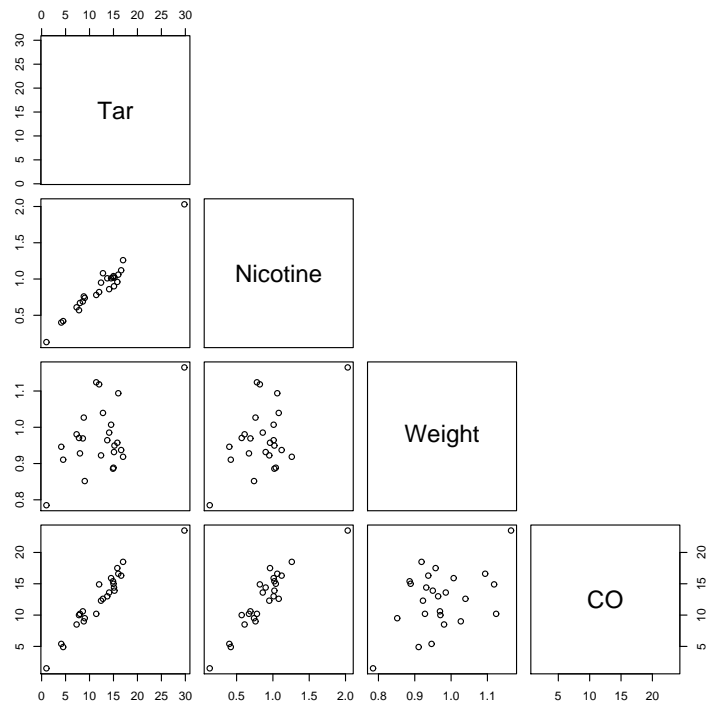


FIGURE 1. Cigarette CO data.

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 $y = \text{CO}$ 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 $\text{var}(\hat{\beta})$), 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.

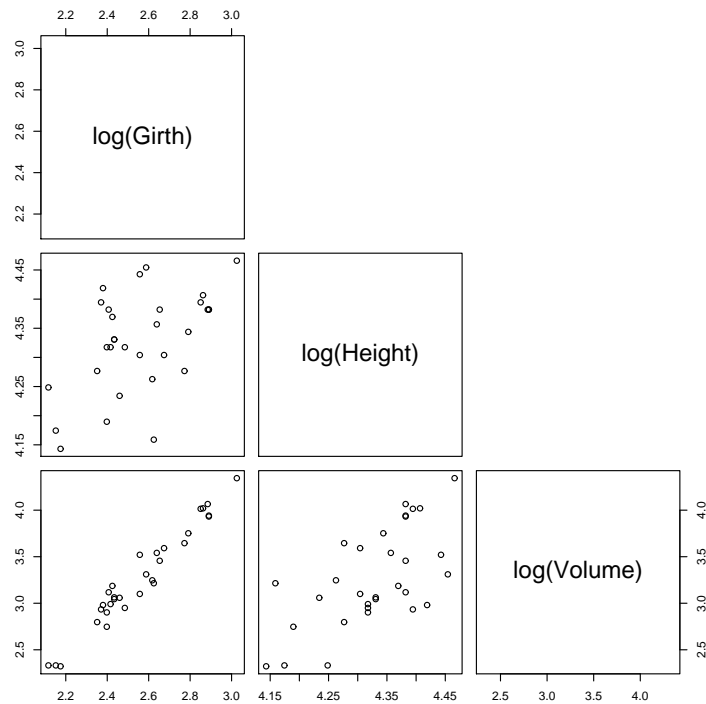


FIGURE 2. Lattice plot of the of the logged response (volume) and the logged explanatory variables (girth and height) for the 31 observations in the `trees` data.

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 $\log(\gamma) \sim N(0, 0.08^2)$.

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

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

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

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

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 $\hat{\sigma}_{\text{MLE}}^2$ 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

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

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

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

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

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

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, $\text{RSS}(\hat{\beta}) = e^T e$. 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 $e^T e + \hat{y}^T \hat{y} = y^T y$ (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 $U = (U_1, \dots, U_p)^T$ and $W = (W_1, \dots, W_n)^T$ with means $\mu_U = \mathbf{E}(U)$ and $\mu_W = \mathbf{E}(W)$ denote by

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

the $p \times n$ covariance matrix with entries $\text{cov}(U, W)_{i,j} = \text{cov}(U_i, W_j)$. 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 $\text{cov}(U + C, W + C') = \text{cov}(U, W)$.

The variance matrix

$$\text{var}(U) = \mathbf{E}((U - \mu_U)(U - \mu_U)^T)$$

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

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

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

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

which is zero, that is, $\text{cov}(\hat{Y}, Y - \hat{Y}) = 0_{n,n}$ 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 $Y = (Y_1, \dots, Y_n)$. 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 $\text{var}(W) = \Sigma$, 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 $\mathbf{E}(\hat{\beta}) = \beta$ and $\text{var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$, so that

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

Ans: This is the generic case with $L = (X^T X)^{-1} X^T$ and $W = Y$, so $\text{var}(\hat{\beta}) = (X^T X)^{-1} X^T \text{var}(Y) ((X^T X)^{-1} X^T)^T$, or $\text{var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1}$ which is $\text{var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$. 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 $\text{RSS} = e^T e$. We use the independence properties when we construct test statistics in the next section.

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

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

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

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

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

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

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

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

Since $Z_i/\sigma \sim N(0, 1)$ for $i = p+1, p+2, \dots, n$, and $\text{RSS}/\sigma^2 = (Z_{p+1}/\sigma)^2 + \dots + (Z_n/\sigma)^2$, 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

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

under H_0 . Now if $A \sim \chi^2(r)$ then $E(A) = r$ so $E(\text{RSS}/\sigma^2) = n-p$ and

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

is an unbiased estimator for σ^2 . It follows that $\hat{\sigma}_{MLE}^2 = \frac{\text{RSS}}{n}$ 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 $H_0 : \beta_k = 0$ against $H_1 : \beta_k \neq 0$ for some particular k from 1 to p . Under H_0 ,

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

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

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

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 in 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}, \dots, x_p . We want to test $H_0 : \beta_{p-k+1} = 0, \beta_{p-k+2} = 0, \dots, \beta_p = 0$ against H_1 : at least one of the last k parameters is non-zero. Under H_0 , 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 H_0 , $y = \tilde{X}\beta^{(0)} + \epsilon$ with $\beta^{(0)}$ a $(p-k) \times 1$ vector, $\beta^{(0)} = (\beta_1, \dots, \beta_{p-k})$. 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

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

Under H_0 , the MLE for σ^2 is

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

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

Under H_1 , 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 $Y = X\beta^{(1)} + \epsilon$, $\hat{\beta}^{(1)} = (X^T X)^{-1} X^T y$, $\hat{Y}^{(1)} = HY$,

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

and

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

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

We can now give the Likelihood Ratio Test (LRT) statistic Λ for H_0 . Substituting the local values into the expression for $\ell(\hat{\beta}, \hat{\sigma}_{MLE}^2; y)$, which we gave at the

end of Section 2.2, we get

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

We reject H_0 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 H_0 . 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 H_0 , $F(Y) > C'$ 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,

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

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 H_0 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,

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

and

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

The former we know, and the latter we demonstrate shortly. Thirdly, $\text{RSS}^{(1)}$ and $\text{RSS}^{(0)} - \text{RSS}^{(1)}$ 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 $(\text{RSS}^{(0)} - \text{RSS}^{(1)})/\text{RSS}^{(1)}$ in the RSS is slight. However, if we add lots of explanatory variables in $H1$, so that $\dim(\text{col}(X))$ approaches n , then we will see a big drop in the RSS, towards zero. In order to account for this, the ratio $(\text{RSS}^{(0)} - \text{RSS}^{(1)})/\text{RSS}^{(1)}$ 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 $E(Y) = \tilde{X}\beta^{(0)}$. We will modify the expansion

$$Y = Z_1e_1 + \dots + Z_ne_n.$$

As before, e_1, \dots, e_n are an orthonormal basis for R^n , and e_1, \dots, e_p are an orthonormal basis for $\text{col}(X)$. We can choose this basis so that there is a first group e_1, \dots, e_{p-k} of vectors spanning the first $p - k$ columns of X and a second group e_{p-k+1}, \dots, e_p completing the basis e_1, \dots, e_p for $\text{col}(X)$ (and notice e_{p-k+1}, \dots, e_p are not a basis for $\text{span}(X_{p-k+1}, \dots, X_p)$ unless $\text{span}(X_{p-k+1}, \dots, X_p) \perp \text{span}(X_1, \dots, X_{p-k})$). 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

$$\hat{Y}^{(0)} = Z_1e_1 + \dots + Z_{p-k}e_{p-k}$$

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

$$\hat{Y}^{(1)} = Z_1e_1 + \dots + Z_pe_p.$$

Now $Y - \hat{Y}^{(1)} = Z_{p+1}e_{p+1} + \dots + Z_ne_n$ so

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

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

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

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

We can see that $\text{RSS}^{(0)} - \text{RSS}^{(1)}$ and $\text{RSS}^{(1)}$ are independent rv, since they are functions of disjoint sets of independent rv. Also, since there are k terms in the last sum above, $(\text{RSS}^{(0)} - \text{RSS}^{(1)})/\sigma^2$ 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

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

so the test statistic is

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

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

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

so the test statistic is

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

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

Exercise verify that $v = (1, -1, 0, \dots, 0)^T$ 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

$$y = \beta'_1(x_1 + x_2) + \beta_3 x_3 + \dots$$

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

$$y = \beta'_1(x_1 + x_2) + \beta'_2(x_1 - x_2) + \beta_3 x_3 + \dots$$

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 $X = [X_1 + X_2, X_1 - X_2, X_3, \dots, X_p]$ and parameter vector $\beta = (\beta'_1, \beta'_2, \beta_3, \dots, \beta_p)$.

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, \dots, x_p come in $m = 3$ groups (typically just 2 or 3 groups) and are ordered so that the groups are

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

This splits off the variables x_{p-k+1} to x_p . This is the variable grouping relevant for the hypothesis $H_0: \beta_{p-k+1} = \beta_{p-k+2} = \dots = \beta_p = 0$, 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 $H_0: \beta_{p-k+1} = \beta_{p-k+2} = \dots = \beta_p = 0$, the model sequence is

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

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 $X_{1:i} = [X_1, X_2, \dots, X_i]$,

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

TABLE 1. ANOVA table for the groups of variables $\{x_2, \dots, x_{p-k}\}$ and $\{x_{p-k+1}, \dots, x_p\}$ 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 $RSS_{1:i}$ 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}, \dots, x_{i+k} to a model that already has the variables x_1, x_2, \dots, x_i is $RSS_{1:i} - RSS_{1:(i+k)}$. 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 $\{1\}, \{2, \dots, p - k\}, \{p - k + 1, \dots, p\}$ 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. $RSS_{1:p}$ 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 $\{x_{p-k+1}, \dots, x_p\}$ to a model with variables $\{x_1, \dots, x_{p-k}\}$, 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 $\{x_2, \dots, x_{p-k}\}$ to a model with just x_1 , the intercept variable. It might seem natural to use the divisor $RSS_{1:(p-k)}/(n - (p - k))$, for an F with $p - k - 1$ numerator and $n - (p - k)$ denominator degrees of freedom. However, (i) the divisor $RSS_{1:p}/(n - p)$ is “just as good” as $RSS_{1:(p-k)}/(n - (p - k))$, since it too is independent of $TSS - RSS_{1:(p-k)}$, so we can see $(TSS - RSS_{1:(p-k)})(n - p)/RSS_{1:p}(p - k - 1)$ has an $F(p - k - 1, n - p)$ distribution under the null, and (ii) it is better, as $RSS_{1:p}/(n - p)$ is an estimate of σ^2 which is not biased if the variables x_{p-k+1}, \dots, x_p added in the row below turn out to be explanatory. You might possibly add (iii) the divisor $RSS_{1:p}/(n - p)$ has a higher variance than $RSS_{1:(p-k)}/(n - (p - k))$, 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 $RSS_{1:(p-k)}/(n - (p - k))$ divisor. Another way to make point (iii) is that the $RSS_{1:(p-k)}/(n - (p - k))$ divisor is the one given by the LRT, where as $RSS_{1:p}/(n - p)$ 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 $H_0 : \beta_{p-k+1} = \beta_p - k + 2 = \dots = \beta_p = 0$ against $H_1 : \beta \in R^p$, 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 $H_0' : \beta_2 = \dots = \beta_p = 0$ against the full model $H_1 : \beta \in R^p$, is the test for no linear relation to the variables x_2, \dots, 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, \dots, 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,

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

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
```

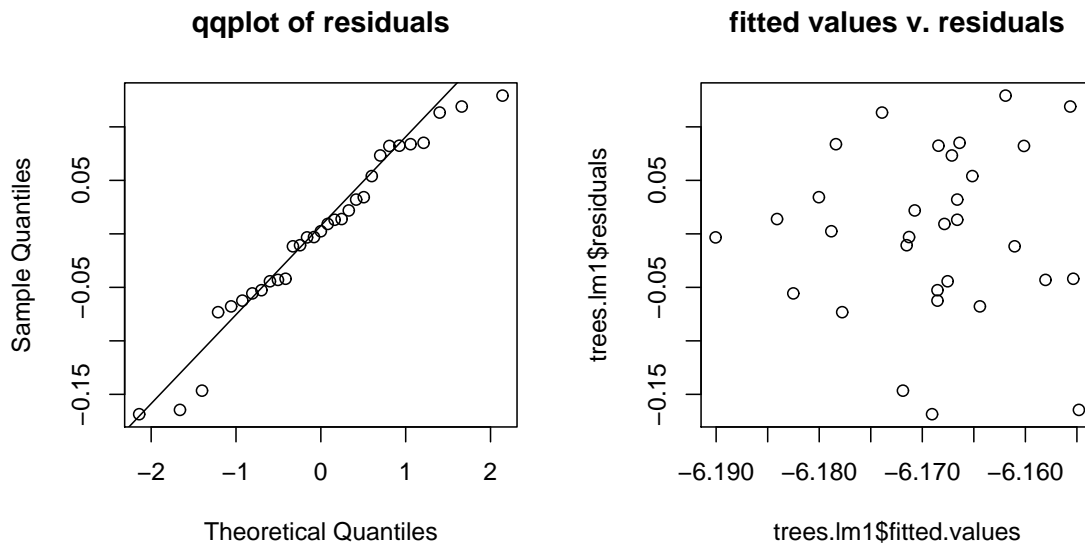


FIGURE 3. (LEFT) A qqplot of the residuals of the logged-model fit to the `trees` data looks healthy, and (RIGHT) no sign of correlation between residuals and fitted values (though a possible funnel, increasing variance with fitted value).

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 $\tilde{X} = 1_{n,1}$, $X^T X = n$, $(X^T X)^{-1} = 1/n$, and $(X^T X)^{-1} X^T y = \bar{y}$.

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 $H_0 : \beta_2 = \beta_3 = 0$ (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 from 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 $RSS_{1:p}$ for the full model is $RSS_{1:3} = 0.185463$, and $RSS_{1:p}/(n-p) = 0.185463/28 = 0.006624$. The other quantity we need, for the test $\beta_3 = \beta_2 = 0$ is $TSS - RSS_{1:3}$, which is

$$RSS_{1:2} - RSS_{1:3} + TSS - RSS_{1:2} = 0.000354 + 0.001868 = 0.002222.$$

So

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

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 $RSS_{1:3} = 0.185463$, $TSS = 0.187686$, $TSS - RSS_{1:3} = 0.002222$, with $k = 2$ for the two variables we set to zero, and if $F = 28(TSS - RSS_{1:3})/2RSS_{1:3}$ 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 H_0 . 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 be ordered or unordered. We will discuss unordered categorical variables.

A categorical explanatory variable $x_k^{(\text{cat})} \in \{1, 2, \dots, c\}$ with c levels is equivalent to c binary indicator variables $g_{k,a} = \mathbb{I}_{x_k^{(\text{cat})}=a}$, with $a = 1, 2, \dots, c$ 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 $x_k^{(\text{cat})}$, and suppose that, for the k 'th response y_k , there are m other explanatory variables $x_{k,1}, x_{k,2}, \dots, x_{k,m}$ including an intercept, $x_{k,1} = 1$. The model

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

allows the mean of y_k to vary with the level of $x_k^{(\text{cat})}$. 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 $x_k^{(\text{cat})} = 1$, then $g_{k,1} = 1$ and $g_{k,2} = \dots = g_{k,c} = 0$, so

$$\mathbb{E}(Y_k) = \alpha + \gamma_2 x_{k,2} + \dots + \gamma_m x_{k,m}.$$

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

$$\mathbb{E}(Y_k) = \alpha + \alpha_a + \gamma_2 x_{k,2} + \dots + \gamma_m x_{k,m}.$$

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 $G_a = (g_{1,a}, \dots, g_{n,a})^T$ for the level- a indicator, and X_1, X_2, \dots, X_m are column vectors for other variables, then the design matrix for the model above is $X = (X_1, G_2, \dots, G_c, X_2, \dots, X_m)$ (so $p = m + c - 1$ here, and columns are in no particular order). The model itself is $Y = X\beta + \epsilon$ with $\beta = (\alpha, \alpha_2, \dots, \alpha_c, \gamma_2, \dots, \gamma_m)$. 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 $X_1 = \sum_{a=1}^c G_a$ since each observation must have its categorical variable in *one* of the levels 1, 2, ..., c , and so the columns of $(X_1, G_1, G_2, \dots, G_c, X_2, \dots, X_m)$ are not linearly independent, and again, the model with all c columns, G_1, \dots, G_c , is over-parameterized. The variables G_1, \dots, G_c are sometimes called a *dummy variables* for the level and the matrix (G_2, \dots, 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 $y_k = \text{price}[k]$, $x_{k,1} = 1$, $x_{k,M} = \text{month}[k]$ and $g_{k,T} = \mathbb{I}_{\text{type}[k]=\text{Terraced}}$, so that

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

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`).

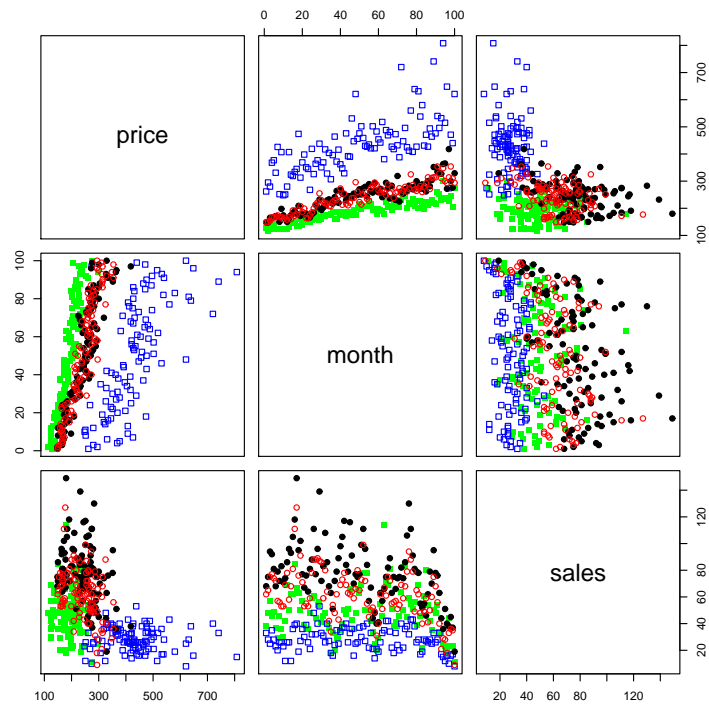


FIGURE 4. Lattice scatter plot of monthly average prices, against month and number of sales. (solid/black circle) Semi-Detached, (open/red circle) Terraced, (solid/green square) Flat, (open/blue square) Detached.

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 $G_T = (g_{1,T}, \dots, g_{n,T})^T$. 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 $2 * (1 - \text{pt}(0.81, 197)) \simeq 0.42$ (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.

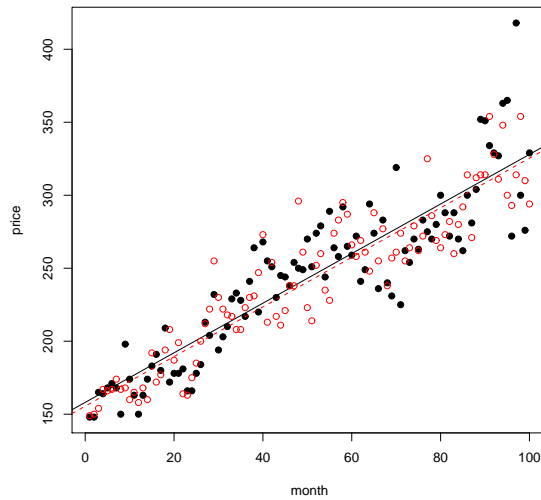


FIGURE 5. Equal-slope regression of the Oxford house-price data. (Solid line/points) Semi-Detached, (circles/dashed) Terraced.

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 $\epsilon_k \sim N(0, \sigma^2)$ and, for observation $k = 1, 2, \dots, n$, we have $y_k = \text{price}[k]$, $g_{k,T}$ is the dummy indicator variable for $\text{type}[k] = \text{Terraced}$, $g_{k,SD}$ is the dummy indicator variable for $\text{type}[k] = \text{Semi-Detached}$ and $x_{k,M}$ is the value of $\text{month}[k]$. 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

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

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 $2 * (1 - \text{pt}(T, 294)) \simeq 0.436$. Note that the matrix $(X^T X)^{-1}$ 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.

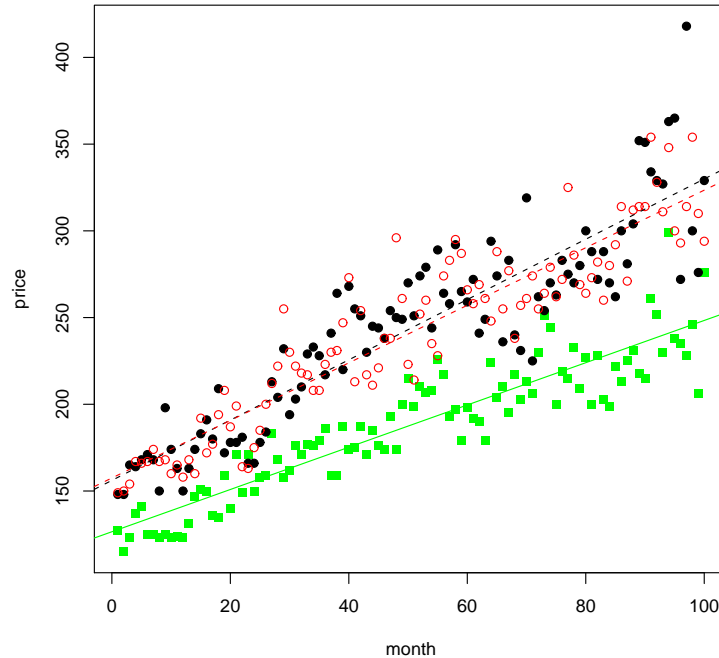


FIGURE 6. Regression of the Oxford house-price data. (upper dashed black line/black full circles) Semi-Detached, (lower dashed red line/red empty circles) Terraced, (solid green line/green squares) Flat.

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]

STATISTICS DEPARTMENT

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