M.Sc. in Applied Statistics MT2004

Formulae for Linear Models

1 Fitting Linear Models

The general setting is that we have n observations $(y, x_1, \ldots, x_p)$ of a scalar *dependent variable* or *response* y and p *carriers* or *regressors* x_i, and we seek a linear relationship of the form

$$y = b_1 x_1 + \cdots + b_p x_p + \text{ residual errors} \tag{1}$$

Note that this does not include a constant term; this can be included by defining $x_1 \equiv 1$. Care is then needed over the meaning of p, which in packages usually excludes the constant regressor. We assume[1] $p < n$.

Define a column vector $Y = (y_1, \ldots y_n)^T$, a column vector $b = (b_1, \ldots, b_p)^T$ and a $n \times p$ matrix X with ith row the vector $(x_1, \ldots, x_p)$ for observation i. Other vectors will be defined in an analogous manner (without mention). Then we have

$$Y = Xb + e \tag{2}$$

for the residual vector e. The predictor $\widehat{Y}$ of Y is given by $\widehat{Y} = Xb$.

Least squares and normal equations

The *residual sum of squares* $RSS = \sum e_i^2 = \|e\|^2$ (using the Euclidean norm on vectors). We choose b to minimize this sum of squares. We will justify this (partially) later, but it has been assumed to be reasonable for a couple of centuries, and it leads to simple computations.

The traditional approach is to write

$$\|e\|^2 = \|Y - Xb\|^2 = (Y - Xb)^T(Y - Xb) = Y^TY - 2Y^TXb + b^TX^TXb$$

which is a quadratic form in the vector b with stationary point (by partially differentiating with respect to each b_i) which solves

$$X^TXb = X^TY \tag{3}$$

the so-called **normal equations**.

[1] This is not necessary, but avoids having constantly to make exceptions.

If X has rank p, then so does X^TX (exercise or see **?**, p. 531) and so X^TX is invertible and

$$b = (X^TX)^{-1}X^TY \tag{4}$$

Let $\widehat{Y}$ denote the fitted values $Xb = X(X^TX)^{-1}X^TY = HY$ for the *hat* matrix $H = X(X^TX)^{-1}X^T$. The residuals $e = (I - H)Y$.

Approach via QR decompositions

This is the modern computational approach.

Theorem: *There is an orthogonal matrix Q such that the $n \times p$ matrix $R = QX$ is upper-triangular ($r_{ij} = 0$ for $i > j$).*

Proof: **?**, §5.2

Note: A matrix A is *orthogonal* if $A^TA = I$.

The name comes from $X = Q^TR$, and Q^T is also an orthogonal matrix. (In the notation of some references Q and Q^T are swapped.) Note that the last $(n-p)$ rows[2] of R are completely zero.

By orthogonality, the residual sum of squares $\|e\|^2$ is unchanged if we replace (Y, X, e) by $(QY, R = QX, Qe)$:

$$\begin{aligned} \|QY - Rb\|^2 &= (QY - Rb)^T(QY - Rb) \\ &= Y^TQ^TQY + X^TQ^TQX - 2Y^TQ^TQXb \\ &= Y^TY + X^TX - 2Y^TXb = \|Y - Xb\|^2 \end{aligned}$$

Thus according to the least squares principle we should choose b to minimize

$$\|e\|^2 = \|Qe\|^2 = \|QY - Rb\|^2$$

Now partition R into the first p rows U and $n - p$ zero rows, and partition QY into V and W in the same way:

$$R = \begin{bmatrix} U \\ 0 \end{bmatrix},\ QY = \begin{bmatrix} V \\ W \end{bmatrix} = \begin{bmatrix} V \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ W \end{bmatrix},\ Rb = \begin{bmatrix} Ub \\ 0 \end{bmatrix} \text{ and } QY - Rb = \begin{bmatrix} V - Ub \\ W \end{bmatrix}$$

Then

$$\|QY - Rb\|^2 = \|V - Ub\|^2 + \|W\|^2 \tag{5}$$

and since *rank*(U) = *rank*(R) = *rank*$(X) = p$ (exercise), U is invertible and we can choose b so that $Ub = V$. From (5), $\|QY - Rb\|^2 \geq \|W\|^2$, and equality is attained (uniquely) for our choice of b.

[2] we have assumed that $n - p > 0$.

Since U is also upper triangular, solving $Ub = V$ is easy:

$$\begin{aligned} b_p &= v_p/u_{pp} \\ b_{p-1} &= (v_{p-1} - u_{p-1,p}b_p)/u_{p-1,p-1} \\ \vdots & \qquad\qquad \vdots \qquad\qquad \vdots \\ b_1 &= (v_1 - u_{1,2}b_2 - \cdots - u_{1,p}b_p)/u_{11} \end{aligned} \tag{6}$$

and at each stage the right-hand side is completely known.

2 A true linear model

Thus far, we have not assumed anything about the distribution of the observations, for example independence or a normal distribution. We have just been calculating properties of fitting by least-squares.

Now suppose that the x's are fixed, but the (y_i) are random and generated by the model

$$y = \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon \tag{7}$$

where the $\epsilon \sim N(0, \sigma^2)$ and they are independent for different observations. Many of the results below are true either exactly or approximately for non-normal errors (use the Central Limit Theorem). In matrix terms $Y = X\beta + \epsilon$.

We need the concept[3] of the (co)variance matrix $\mathrm{Var}(Z)$ of a random vector. This is defined by

$$\mathrm{Var}(Z) = E\left[(Z - EZ)(Z - EZ)^T\right]$$

where EZ is the vector of means of the components. Then $\mathrm{var}(Z_i) = (\mathrm{Var}(Z))_{ii}$ and $\mathrm{cov}(Z_i, Z_j) = (\mathrm{Var}(Z))_{ij}$. For a constant matrix A, $\mathrm{Var}(AZ) = A\mathrm{Var}(Z)A^T$.

We have (by definition) $\mathrm{Var}(\epsilon) = \sigma^2 I$ for the $n \times n$ identity matrix I, and hence $\mathrm{Var}(Y) = \sigma^2 I$. Then $\mathrm{Var}(b) = (X^TX)^{-1}X^T\sigma^2 I[(X^TX)^{-1}X^T]^T = \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$.

Estimating σ^2

The residuals $e = Y - Xb = Y - HY = (I - H)Y = (I - H)(X\beta + \epsilon) = (I - H)\epsilon$. So $E\,e = (I-H)E\,\epsilon = 0$ and $E\,RSS = E\,\|e\|^2 = E\,\mathrm{trace}(ee^T) = \mathrm{trace}(I-H)E\,\epsilon\epsilon^T(I-H) = \sigma^2\mathrm{trace}(I - H)^2 = \sigma^2\mathrm{trace}(I - H) = \sigma^2(n - p)$.

Further, $RSS/\sigma^2 \sim \chi^2_{n-p}$ and b and RSS are independent.

[3] If these results are unfamiliar to you, write them out as sums over observations to see that they do encapsulate the basic formulae.

t tests and confidence intervals

We can now consider tests and confidence intervals for coefficients β_i. The statistic

$$t = \frac{b_i - \beta_i}{s\sqrt{(X^T X)^{-1}_{ii}}} \tag{8}$$

has a t_{n-p} distribution. (The numerator is normally distributed with mean zero and variance $\sigma^2(X^T X)^{-1}_{ii}$, and $(n-p)s^2/\sigma^2 = RSS/\sigma^2$.)

We can use (8) directly to test whether β_i has any specified value. It can also be used as a pivot to give a $(1-2\alpha)$ confidence interval

$$\left(b_i - t_{n-p,\alpha} s\sqrt{(X^T X)^{-1}_{ii}},\ b_i + t_{n-p,\alpha} s\sqrt{(X^T X)^{-1}_{ii}}\right) \tag{9}$$

where $t_{n-p,\alpha}$ is the $1-\alpha$ point of a t_{n-p} distribution.

F tests

The sums of squares in the lines of the basic analysis of variance table are independent, since the regression SS depends only on b, and this is independent of RSS. Suppose $\beta = 0$. Then $SS_{\text{regression}}/\sigma^2 \sim \chi^2_p$. Thus

$$F = \frac{SS_{\text{regression}}/p}{RSS/(n-p)} \sim F_{p,n-p} \tag{10}$$

provides a test for $\beta = 0$, that is for the efficacy of the whole set of regressors. Note that the divisors are the degrees of freedom, so that F is a ratio of mean squares with denominator s^2.

We can also test whether a subset of regressors suffices. Suppose $\beta_{q+1} = \cdots = \beta_p = 0$. Then

$$RSS_q - RSS_p \sim \sigma^2 \chi^2_{p-q}$$

independently of $RSS_p =$, and

$$F = \frac{(RSS_q - RSS_p)/(p-q)}{RSS_p/(n-p)} \sim F_{p-q,n-p} \tag{11}$$

This is again a ratio of mean squares.

Note that this applies to selecting any q regressors, not just the first q (as we can re-order them), and also to testing any pais of *nested* models, that is one model whose predictions from a q-dimensional subspace of the p-dimensional prediction space of the larger model.

F vs t tests

We could test $\beta_p = 0$ either via a t-test or via an F-test. Are they equivalent? Yes, $F = t^2$.

This is true of testing any one coefficient to be zero, as we can always reorder the regressors so the coefficient of interest is the last one.

Predictions

The prediction $\widehat{Y} = Xb$ has mean $XEb = X\beta$ and variance matrix $\mathrm{Var}(\widehat{Y}) = X\mathrm{Var}(b)X^T = \sigma^2 X(X^TX)^{-1}X^T = \sigma^2 H$, say

We can also consider predicting a new observation $(y, x_1, \ldots, x_p) = (y, x)$ for a *row* vector x. The prediction will be $\widehat{y} = xb$ with mean $x\beta$ and variance $\mathrm{var}(\widehat{y}) = x\mathrm{Var}(b)x^T$. Since $y = x\beta + \epsilon_0$, and ϵ_0 is independent of Y, the variance of the prediction error, $y - \widehat{y}$, is

$$\mathrm{var}(y - \widehat{y}) = \mathrm{var}(y) + \mathrm{var}(\widehat{y}) = x\mathrm{Var}(b)x^T + \sigma^2 = \sigma^2\left(1 + x(X^TX)^{-1}x^T\right) \tag{12}$$

Note that there are three variances we might want here:

(a) That of the prediction error, $\sigma^2\left(1 + x_i(X^TX)^{-1}x_i^T\right)$.

(b) The variance of the prediction itself, $\mathrm{var}(\widehat{y}) = \sigma^2 x(X^TX)^{-1}x^T$, or

(c) The variance of a residual, which will be lower than the prediction error as y_i participated in the fit, and in fact is $\mathrm{var}(e_i) = \sigma^2\left(1 - x_i(X^TX)^{-1}x_i^T\right)$ (see (17)).

Take care not to confuse them.

Likelihoods

We started by assuming fitting by least-squares as a principle. If we assume the true linear model with normal errors, we can show that the least-squares estimators coincide with the maximum-likelihood estimators.

The likelihood of Y is given by

$$\ell(\beta, \sigma^2; Y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{(y_i - x_i\beta)^2}{2\sigma^2}$$

where x_i refers to the ith row of X, as a row vector. Thus the log-likelihood is

$$L(\beta, \sigma^2; Y) = \text{const} - \|Y - X\beta\|^2/2\sigma^2 - \tfrac{n}{2}\log\sigma^2 \tag{13}$$

and this clearly maximized by $\widehat{\beta} = b$, the least-squares estimator. Then

$$L(\widehat{\beta}, \sigma^2; Y) = \text{const} - RSS/2\sigma^2 - \tfrac{n}{2}\log\sigma^2 \tag{14}$$

which is maximized by $\widehat{\sigma}^2 = RSS/n$, not by s^2. Thus

$$L(\widehat{\beta}, \widehat{\sigma}^2; Y) = \text{const} - \tfrac{n}{2}\log(RSS/n) \tag{15}$$

$$\text{or, equivalently} \quad \ell(\beta, \sigma^2; Y) \propto RSS^{-n/2} \tag{16}$$

From (16) it is clear that a likelihood ratio test of two nested hypotheses is equivalent to the ratio $RSS_p^{-n/2}/RSS_q^{-n/2}$ or to the ratio RSS_q/RSS_p of the residual sum of squares, and hence to the F-test (11). (The critical region of the LR test is 'ratio of optimised likelihood under larger model to that under smaller model is greater than k'. This is $RSS_p^{-n/2}/RSS_q^{-n/2} >$

k. Equivalently, $RSS_q/RSS_p > k_1$, or $RSS_q/RSS_p - 1 > k_1 - 1$, that is, $(RSS_q - RSS_p)/RSS_p > k_2$. But if we divide numerator and denominator by their degrees of freedom, this only changes the constant again,and now we know the distribution is F.)

The t test is (by definition) a Wald test for dropping a single regressor from the model (and it uses the estimates from the full model only).

3 Residuals and Outliers

Types of residuals

Having fitted our model, we want to check whether the fit is reasonable. We do this by looking at various types of residuals. The (ordinary) residual vector $e = Y - \widehat{Y} = Y - Xb$. This fails to be a fair measure of the errors for two reasons:

1.) The variance of e_i varies over the space of regressors, being greatest at $(\overline{x}_1, \ldots, \overline{x}_p)$. We saw

$$e = (I - H)\epsilon$$

We have

$$\begin{aligned} \mathrm{Var}(e) &= \mathrm{Var}(I - H)\epsilon = (I - H)\mathrm{Var}(\epsilon)(I - H) \\ &= (I - H)\sigma^2 I(I - H) = \sigma^2(I - H) \end{aligned}$$

(since $HH = H$) and finally

$$\mathrm{var}(e_i) = \sigma^2(1 - h_{ii}) \tag{17}$$

(For an explanation of why $1 - h_{ii}$ is greatest at the mean of the x's, see **?**, §8.2.1 or **?**, p. 308.)

The **standardized residual** is formed by normalizing to unit variance then replacing σ^2 by s^2, so

$$e_i' = \frac{e_i}{s\sqrt{1 - h_{ii}}} \tag{18}$$

Another way of looking at the reduced variance of residuals when h_{ii} is large is to say that data point i has high **leverage** and is able to pull the fitted surface toward the point. To say what we mean by 'large', note that $\sum_i h_{ii} = \mathrm{trace}\, H = \mathrm{trace}\, X(X^TX)^{-1}X^T = \mathrm{trace}\, X^TX(X^TX)^{-1} = \mathrm{trace}\, I_{p\times p} = p$, so the average leverage is p/n. We will generally take note of points with leverages more than two or three times this average.

2.) If one error is very large, the variance estimate s^2 will be too large, and this deflates the standardized residuals. Let us consider fitting the model (1) without observation i. We get a prediction $\widehat{y}_{(i)}$ of y_i. The **studentized (or deletion or jackknife) residual** is

$$e_i^* = \frac{y_i - \widehat{y}_{(i)}}{\sqrt{\mathrm{var}(y_i - \widehat{y}_{(i)})}}$$

where σ (used in the variance estimate) is replaced by its estimate in this model, $s_{(i)}$.

Fortunately, it is not necessary to re-fit the model each time deleting an observation. Let $e_{(i)} = y_i - \widehat{y}_{(i)}$. For the algebra we need the following result (the Sherman–Morrison–Woodbury formula)

$$(A - UV^T)^{-1} = A^{-1} + A^{-1}U(I - V^T A^{-1}U)^{-1}V^T A^{-1} \tag{19}$$

for a $p \times p$ matrix A and $p \times m$ matrices U and V with $m \leq p$. (Just check that this has the properties required of the inverse.)

Let X_i denote the ith row of X, and $X_{(i)}$ denote X with the ith row deleted. Then

$$X_{(i)}^T X_{(i)} = (X^T X - X_i^T X_i)$$

and

$$X_i(X^T X)^{-1} X_i^T = h_{ii}$$

and so from (19) with $A = X^T X, U = X_i^T, V = X_i$ and $m = 1$, we have

$$I - V^T A^{-1} U = I - X_i^T X^T X^{-1} X_i^T = 1 - h_{ii}$$

and

$$(X_{(i)}^T X_{(i)})^{-1} = (X^T X)^{-1} + (X^T X)^{-1} X_i^T X_i (X^T X)^{-1}/(1 - h_{ii})$$

Now

$$b_{(i)} = (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T Y_{(i)}$$

and

$$X_{(i)}^T Y_{(i)} = X^T Y - X_i^T y_i$$

Substituting for these two gives

$$\begin{aligned}
b_{(i)} &= ((X^T X)^{-1} + (X^T X)^{-1} X_i^T X_i (X^T X)^{-1}/(1 - h_{ii}))(X^T Y - X_i^T y_i) \\
&= (X^T X)^{-1} X^T Y - (X^T X)^{-1} X_i^T y_i \\
&\quad + (X^T X)^{-1} X_i^T X_i (X^T X)^{-1} X^T Y/(1 - h_{ii}) \\
&\quad - (X^T X)^{-1} X_i^T X_i (X^T X)^{-1} X_i^T Y/(1 - h_{ii}) \\
&= b - (X^T X)^{-1} X_i^T y_i + (X^T X)^{-1} X_i^T X_i b/(1 - h_{ii}) \\
&\quad - (X^T X)^{-1} X_i^T h_{ii} Y/(1 - h_{ii}) \\
&= b - (X^T X)^{-1} X_i^T y_i (1 + h_{ii}/(1 - h_{ii})) + (X^T X)^{-1} X_i^T X_i b/(1 - h_{ii}) \\
&= b - (X^T X)^{-1} X_i^T y_i/(1 - h_{ii})) + (X^T X)^{-1} X_i^T X_i b/(1 - h_{ii}) \\
&= b - (X^T X)^{-1} X_i^T (y_i - X_i b)/(1 - h_{ii}) \\
&= b - (X^T X)^{-1} X_i^T e_i/(1 - h_{ii})
\end{aligned} \tag{20}$$

Also

$$(n - p)s^2 = \|Y - Xb\|^2 = Y^T Y - b^T X^T Y$$

and after deletion

$$\begin{aligned}
(n - p - 1)s_{(i)}^2 &= Y^T Y - y_i^2 - b_{(i)}^T (X^T Y - X_i^T y_i) \\
&= Y^T Y - y_i^2 - (b^T - ((X^T X)^{-1} X_i^T e_i)^T/(1 - h_{ii}))(X^T Y - X_i^T y_i) \\
&= Y^T Y - y_i^2 - b^T X^T Y + b^T X_i^T y_i + e_i X_i (X^T X)^{-1} X^T Y/(1 - h_{ii}) \\
&\quad - e_i X_i (X^T X)^{-1} X_i^T y_i/(1 - h_{ii})
\end{aligned}$$

$$
\begin{aligned}
&= (n-p)s^2 - y_i^2 + b^T X_i^T y_i + e_i X_i b/(1-h_{ii}) - e_i h_{ii} y_i/(1-h_{ii}) \\
&= (n-p)s^2 - y_i^2 + \widehat{y}_i y_i + e_i \widehat{y}_i/(1-h_{ii}) - e_i h_{ii} y_i/(1-h_{ii}) \\
&= (n-p)s^2 - (y_i^2 - h_{ii} y_i^2 - \widehat{y}_i y_i + h_{ii}\widehat{y}_i y_i - e_i \widehat{y}_i + e_i h_{ii} y_i)/(1-h_{ii}) \\
&= (n-p)s^2 \\
&\quad -(y_i^2 - h_{ii} y_i^2 - \widehat{y}_i y_i + h_{ii}\widehat{y}_i y_i - (y_i - \widehat{y}_i)\widehat{y}_i + (y_i - \widehat{y}_i) h_{ii} y_i)/(1-h_{ii}) \\
&= (n-p)s^2 - (y_i^2 - 2\widehat{y}_i y_i + \widehat{y}_i{}^2)/(1-h_{ii}) \\
&= (n-p)s^2 - e_i^2/(1-h_{ii}) \\
&= s^2\left[(n-p) - e_i'^2\right]
\end{aligned}
$$

Now since observation i plays no part in $\widehat{y}_{(i)}$, it is independent of y_i, and $e_{(i)} = y_i - \widehat{y}_{(i)}$ has variance from (12) of

$$
\begin{aligned}
\text{var } e_{(i)} &= \sigma^2\left(1 + X_i (X_{(i)}^T X_{(i)})^{-1} X_i^T\right) \\
&= \sigma^2\left(1 + X_i\left((X^T X)^{-1} + (X^T X)^{-1} X_i^T X_i (X^T X)^{-1}/(1-h_{ii})\right) X_i^T\right) \\
&= \sigma^2\left(1 + h_{ii} + h_{ii}^2/(1-h_{ii})\right) \\
&= \sigma^2\left(1 - h_{ii} + h_{ii} - h_{ii}^2 + h_{ii}^2\right)/(1-h_{ii}) \\
&= \sigma^2/(1-h_{ii})
\end{aligned}
$$

Also,

$$
\widehat{y}_{(i)} = X_i b_{(i)} = X_i b - X_i (X^T X)^{-1} X_i^T e_i/(1-h_{ii}) = \widehat{y}_i - h_{ii} e_i/(1-h_{ii}) \tag{21}
$$

and from this

$$
e_{(i)} = y_i - \widehat{y}_{(i)} = (y_i - \widehat{y}_i) + h_{ii} e_i/(1-h_{ii}) = e_i/(1-h_{ii})
$$

Finally,

$$
e_i^* = \frac{e_{(i)}}{s_{(i)}/\sqrt{(1-h_{ii})}} = \frac{e_i}{s_{(i)}\sqrt{(1-h_{ii})}} = \frac{s e_i'}{s_{(i)}} = \frac{e_i'}{\sqrt{\frac{n-p-e_i'^2}{n-p-1}}} \tag{22}
$$

This shows explicitly that where the standardized residual e_i' is larger than one (in modulus), the studentized residual e_i^* is larger. (The denominator of (22) is less than 1.) In fact, from (22) we can deduce that the maximum value of the standardized residual is $\sqrt{n-p}$.

Cook's statistic

The studentized residuals tell us whether a point has been explained well by the model, but if it has not, they do not tell us what the size of the effect on the fitted coefficients of omitting the point might be. A badly-fitted point in the middle of the design space will have much less effect on the predictions than one at the edge of the design space.

? proposed a measure that combines both the effect of leverage and that of being badly fitted. His statistic is

$$
D_i = \frac{(b_{(i)} - b)^T X^T X (b_{(i)} - b)}{p s^2} = \frac{\|\widehat{Y}_{(i)} - \widehat{Y}\|^2}{p s^2} = \frac{(e_i')^2 h_{ii}}{p(1-h_{ii})}
$$

We can derive the last expression using (20):

$$\begin{aligned} b_{(i)} - b &= -(X^T X)^{-1} X_i^T e_i/(1-h_{ii}) \\ \widehat{Y}_{(i)} - \widehat{Y} &= Xb_{(i)} - Xb = -X(X^T X)^{-1} X_i^T e_i/(1-h_{ii}) \\ \frac{\|\widehat{Y}_{(i)} - \widehat{Y}\|^2}{s^2} &= \frac{e_i^2}{s^2(1-h_{ii})^2} X_i (X^T X)^{-1} X^T X (X^T X)^{-1} X_i^T \\ &= \frac{e_i^2 h_{ii}}{s^2(1-h_{ii})^2} = \frac{h_{ii}}{(1-h_{ii})} \frac{e_i^2}{s^2(1-h_{ii})} = \frac{(e_i')^2 h_{ii}}{(1-h_{ii})} \end{aligned}$$

using (18). Large values of D_i indicate 'influential' observations, that is those which if dropped would have a large effect on the predictions (measured by $\|\widehat{Y}_{(i)} - \widehat{Y}\|^2$) or on the simultaneous $1-\alpha$ confidence region for β

$$(\beta - b)^T X^T X (\beta - b) \leq p s^2 F_{p,n-p,\alpha}$$

Several small modifications have been proposed. One (**?**, p. 25) is to use the signed square root, taking the same sign as that of the residuals, and to drop the p. If in this we replace s by $s_{(i)}$ we get

$$\text{DFFITS}_i = \sqrt{\frac{h_{ii}}{(1-h_{ii})}} e_i^*$$

As **?**, p. 161 point out

> 'Actually, the number of measures available in the literature for identifying outliers and influential points verges on being mind-boggling.'

So use what tools your computer package makes available, and remember that deleting one or more points and re-fitting is much less onerous than when most of these measures were designed (in the days of punch cards and batch computing).