The general linear model: what you need to know

P.M.E.Altham, Statistical Laboratory, University of Cambridge.

October 14, 2004

Note: this set of basic notes contains deliberate gaps, for you to fill in.
The models

$$y_i = \alpha + \beta x_i + \epsilon_i,$$

$$y_i = \alpha + \beta x_i + \gamma {x_i}^2 + \epsilon_i$$

and

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

may all be seen as special cases of the model

$$y_i = \beta^T x_i + \epsilon_i$$

for $i = 1, ..., n$ where we assume that $(\epsilon_i, i = 1, ..., n)$ form a random sample from $N(0, \sigma^2)$. Here y_i is the 'dependent' variable, x_i is the known covariate, β the unknown parameter, of dimension say p, and ϵ_i is the unknown 'error': we assume that $(\epsilon_i, i = 1, ..., n) \sim NID(0, \sigma^2)$, ie $(\epsilon_i, i = 1, ..., n)$ form a random sample from $N(0, \sigma^2)$. The parameter σ^2 is also unknown. We rewrite this model as

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

where $\epsilon \sim N_n(0, \sigma^2 I)$ (a multivariate- normal distribution).
How do we estimate β? What is the accuracy of our estimate? How do we check the validity of our model?
The mle (maximum likelihood estimator) is say $\hat{\beta}$, which is the vector that minimises

$$\Sigma_i (y_i - \beta^T x_i)^2,$$

equivalently minimises

$$R(\beta) = (Y - X\beta)^T (Y - X\beta).$$

Take the partial derivative wrt β of the expression above, and set it to 0. Then you find that the equation for $\hat{\beta}$ is

$$2X^T Y = 2X^T X\beta.$$

Suppose that X is of full rank, p, then $X^T X$ is non-singular, and we see that $R(\beta)$ is min'd wrt β by $\hat{\beta}$, the solution of

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

We now find the distribution of $\hat{\beta}$. Write

$$\hat{\beta} = (X^T X)^{-1} X^T (X\beta + \epsilon) = \beta + (X^T X)^{-1} X^T \epsilon.$$

Now $E(\epsilon) = 0$ and $E(\epsilon\epsilon^T) = \sigma^2 I$. Hence (check), $E(\hat{\beta}) = \beta$, ie the estimator $\hat{\beta}$ is unbiased.
Further

$$\hat{\beta} - \beta = L\epsilon$$

say, where $L = (X^T X)^{-1} X^T$. Thus, since $\epsilon \sim N_n(0, \sigma^2 I)$, we see that

$$\hat\beta - \beta \sim N_p(0, \sigma^2 L L^T).$$

Furthermore, $LL^T = (X^T X)^{-1}$, CHECK THIS,
and so we see that

$$\hat\beta - \beta \sim N_p(0, \sigma^2 (X^T X)^{-1}),$$

which we rewrite as

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

If σ^2 were known, we could use this result to produce confidence intervals for say β_1, a component of β.
How do we get around the problem of σ^2 unknown?
We use the following powerful distributional result (which we do NOT PROVE here):
define

$$Q = R(\hat\beta) = (Y - X\hat\beta)^T (Y - X\hat\beta),$$

as the residual sum of squares (rss), then

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

independently of Q, and

$$Q/\sigma^2 \sim \chi^2_f$$

where $f = n - p$, thus $f =$ number of independent observations minus number of parameters fitted.
Hence (i) $E(Q) = f\sigma^2$, and so $s^2 = Q/f$ is our unbiased estimator of σ^2.
Further, (ii)

$$(\hat\beta_j - \beta_j)/\sqrt{v_{jj} s^2} \sim t_f$$

where (v_{ij}) is the matrix $(X^T X)^{-1}$ (having jth diagonal element (v_{jj})).
Observe that in general the components of $\hat\beta$ will be correlated, since $(X^T X)^{-1}$ is not necessarily a diagonal matrix.
We can always reparametrise, say from β to γ, by a non-singular linear transformation, to arrange that the components of $\hat\gamma$ are uncorrelated. In this case we say that $\gamma_1, \ldots, \gamma_p$ are orthogonal parameters.
How do we do this? We use a bit more algebra. Suppose we define $\gamma = B\beta$, for some non-singular $p \times p$ matrix B.

$$Y = X\beta + \epsilon = (XB^{-1})\gamma + \epsilon.$$

Thus

$$Y = X_1 \gamma + \epsilon,$$

say, where $X_1 = XB^{-1}$, and

$$\hat\gamma \sim N_p(\gamma, \sigma^2 (X_1^T X_1)^{-1}).$$

Hence the parameters $\gamma_1, \ldots, \gamma_p$ are orthogonal provided that we arrange that $(X_1^T X_1)$ is a diagonal matrix: without loss of generality (wlog) we can arrange that $(X_1^T X_1) = I_p$, the $p \times p$ identity matrix.
Specifically, we require $(XB^{-1})^T (XB^{-1})$ to be I_p.
This is equivalent to the problem: choose B a real $p \times p$ matrix, such that

$$X^T X = B^T B.$$

The reason that we can find such a B is that $X^T X$ is a $p \times p$ positive- definite matrix.
Important special case: orthogonal polynomials. Consider

$$y_i = \alpha + \beta x_i + \gamma x_i^2 + \epsilon_i.$$

We can rewrite this as

$$y_i = \alpha' + \beta'(x_i - \bar{x}) + \gamma'((x_i - \bar{x})^2 + b(x_i - \bar{x}) + c) + \epsilon_i,$$

with b, c such that

$$\Sigma((x_i - \bar{x})^2 + b(x_i - \bar{x}) + c) = 0$$

$$\Sigma((x_i - \bar{x})^2 + b(x_i - \bar{x}) + c)(x_i - \bar{x}) = 0.$$

What are the consequent equations for b, c?
Then the parameters α', β', γ' are mutually orthogonal, although the parameters α, β, γ were not, in general. This is used, for example, by S-Plus in the construction of *orthogonal polynomials*.

What about residuals? What about diagnostic plots? First, some more algebra. Define the *fitted values* $\hat{Y} = X\beta$, and define $\hat{\epsilon} = Y - \hat{Y}$ as the *residuals*. Then you can check that

$$\hat{Y} = X(X^TX)^{-1}X^TY = PY$$

say, where P is a *projection* matrix, ie it satisfies

$$P = P^T, PP = P,$$

thus it is symmetric and idempotent. (P is also called H, the 'hat' matrix.)
Now

$$\hat{\epsilon} = \ldots\ldots = (I - P)\epsilon$$

(CHECK THIS), and so

$$\epsilon \sim N(0, \sigma^2 I) \text{ implies } \hat{\epsilon} \sim N(0, \sigma^2(I - P)).$$

CHECK THIS.
FILL IN THE STORY BEHIND THE qqplot.

Define h_i as the ith diagonal element of the matrix P. Then $\hat{\epsilon}_i \sim N(0, (1-h_i)\sigma^2)$ and $0 \leq h_i \leq 1$ (can you prove this?). We call h_i the leverage of the point x_i.
Then $\Sigma_i h_i = trace(P)$ by definition of the function $trace()$, and, since $PP = P$, all the eigen-values of P are 0 or 1 (prove this). Hence $\Sigma_i h_i =$ sum of eigen values of $P = rank(P) = p$.
Here is a picture to illustrate a point of high leverage.

Without the special point x_i, as in the picture, we may think that there is no particular relationship between y and x.
But if we include x_i in our regression fit, then we will find that
i) the fit looks much better (R^2 is closer to 1).
ii) Since h_i is relatively large, we'll find that $h_i \simeq 1$, and since the fitted value $\hat{y}_i = \Sigma_j h_{ij} y_j$, we will find that $\hat{y}_i \simeq y_i$; the point P is exerting very high *leverage* at P, and the fit there is nearly perfect.
This could be very misleading, or it could mean that if we took more observations, then we might discover something rather interesting.
Back to residuals and fitted values. First, observe that

$$\hat{Y} = PY, \hat{\epsilon} = (I - P)\epsilon$$

imply that

$$\hat{Y} = PX\beta + P\epsilon, \hat{\epsilon} = (I - P)\epsilon$$

and hence ... $cov(\hat{\epsilon}_i, \hat{Y}_j) = 0$ for all i, j.
Typically we plot $\hat{\epsilon}_i$ against $\hat{Y}_i$ expecting to see 'no particular trend': as in the picture below:

But you may find, eg if y_i corresponds to height, weight, monetary value,... that $|\epsilon_i|$ tends to increase as $\hat{Y}_i$ increases: the residuals tend to 'fan out' as in the sketch below:

This suggests that in the model

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

we should assume that $\epsilon_i \sim N(0, \sigma_i^2)$, ie the variance is non-constant, we have *heteroscedasticity*. There is a 'fix-up' which often enables us to get over this problem: namely to work with a transformation of y_i rather than y_i itself. For example, if we take the model

$$\log(y_i) = (X\beta)_i + \epsilon_i$$

we may well find (from the diagnostic plots) that apparently

$$var(\epsilon_i) \simeq \text{constant}.$$

In this case we say that log() is a *variance-stabilising transformation.*
Motivation for variance-stabilising transformation: Lemma
Suppose the rv Y is approximately distributed as $N(\mu, \sigma^2(\mu))$.

Take any 'well-behaved' transformation $g()$.
Then, use Taylor's theorem (FILL IN) to show that, approximately:

$$g(Y) \sim N(g(\mu), \sigma^2(\mu)(g'(\mu))^2).$$

Hence, given the function $\sigma^2(\mu)$, for example

$$\sigma^2(\mu) \propto \mu^2,$$

we can choose the function $g()$ such that $\sigma^2(\mu)(g'(\mu))^2 = \text{constant}$, by solving the corresponding differential equation.
We return to the question of transforming the response variable Y when we discuss the family of Box-Cox transformations: this allows you to let mle find the best transformation.

More on testing hypotheses in the linear model

We first need to state one more theorem, which we will then use, without proof.
Let Ω be the model

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

with the usual assumption that $\epsilon \sim N(0, \sigma^2 I)$, and let X, β be partitioned so that $X\beta = X_1\beta_1 + X_2\beta_2$. Let ω be the submodel say

$$Y = X_1\beta_1 + \epsilon,$$

and let R_Ω, R_ω be the corresponding residual sums of squares. Note,

$$R_\Omega \leq R_\omega.$$

Put $Q_1 = R_\omega - R_\Omega$. Then, on Ω, $R_\Omega/\sigma^2 \sim \chi^2_f$ independently of Q_1, and $Q_1/\sigma^2 \sim$ non-central $\chi^2_{f_1}$, where $f_1 = dim(\Omega) - dim(\omega)$. Further, this second χ^2 is ordinary (ie central) χ^2 if and only if ω is true. Hence, to test ω against Ω we refer the following ratio to $F_{*,*}$.

$$\frac{Q_1/f_1}{R_\Omega/f}.$$

We will need to change our notation a little to apply this result to get the usual analysis of variance.
Consider the model

$$y_i = \mu + \beta^T x_i + \epsilon_i$$

equivalently

$$\Omega : Y = \mu 1 + X\beta + \epsilon,$$

and assume as usual that $\epsilon \sim N(0, \sigma^2 I)$.
In a general notation, we have the following **anova**, ie analysis of variance. This enables us to construct the F test of the hypothesis $\beta = 0$. Table 1 gives us the basic analysis of variance.
Here 'total' is actually the resid ss fitting the 'baseline' model

$$H_0 : Y = \mu 1 + \epsilon.$$

Thus H_0 is the hypothesis $\beta = 0$.
It can be shown that, on H_0,

$$S_\beta/\sigma^2 \sim \chi^2_p, \; R_\Omega/\sigma^2 \sim \chi^2_{n-1-p}$$

and these two are independent. From this result we can construct the $F-$test of $\beta = 0$, ie refer FILL IN FORMULA

....

Table 1: anova

due to β	S_β	p
residual	R_Ω	$n-p-1$
'total'	$\Sigma(y_i-\bar{y})^2$	$n-1$

Observe that $S_\beta/\text{'}total'$ is called the multiple correlation coefficient R^2: the closer it is to 1, the better the regression 'explains' the variation in the data y. But in fact, we are more likely to want to test whether particular components of β are zero, ie to test whether the corresponding columns of X can be 'dropped' from the model.
Suppose we now partition X, β so that

$$X\beta = X_1\beta_1 + X_2\beta_2.$$

We may wish to test the model

$$\omega_1 : Y = \mu 1 + X_1\beta_1 + \epsilon$$

against the model Ω, equivalently, to test $\omega_1 : \beta_2 = 0$, assuming that Ω is true.
Let $R_\Omega =$ resid ss fitting Ω, thus
$R_\Omega = (Y - \hat{\mu} - X\hat{\beta})^T(Y - \hat{\mu} - X\hat{\beta})$,
and let $R_{\omega_1} =$ resid ss fitting ω_1, thus

$$R_{\omega_1} \geq R_\Omega.$$

For this particular test we will use the following result: on Ω,

$$R_\Omega/\sigma^2 \sim \chi^2_{n-1-p} \text{ independently of } (R_{\omega_1} - R_\Omega)/\sigma^2 \sim \chi^2_{p_2}(c)$$

and this second χ^2 is non-central, **unless** $\beta_2 = 0$. Hence, to test $\beta_2 = 0$, we refer the following RATIO (fill in) to $F_{p_2, n-1-p}$.

and we can construct an extended aov like this
fill in big table

Observe, R_Ω is defined in terms of the fitted value $\hat{\mu}, \hat{\beta}$, as above,

and correspondingly, R_{ω_1} is defined in terms of the fitted value say μ^*, β_1^*, specifically:

.

If we now want to test, say $\beta_1 = 0$, we have to start again, in order to compute the appropriate residual ss.
In a nutshell, this is why, in general, the result of (for example)

```
anova(lm(y~ x+z))
```

will look different from the result of

```
anova(lm(y~ z+x))
```

From the point of view of the anova (ie from the point of view of hypothesis testing) the **order** in which the terms appear in the linear model is very important, unless we have a design such that the parameters corresponding to x, z are **orthogonal**. EXPAND THIS STATEMENT.

.