Theory of Linear Models

Steven Gilmour
King's College London

January – February 2020

Introduction to Statistical Modelling

This is a course on statistical modelling and especially linear models. I do not assume prior knowledge of linear models, but I take a fairly theoretical approach, so some familiarity with distribution theory, statistical inference and linear algebra is a great advantage.

A statistical model is a mathematical model, which includes a stochastic component, of a process which generates (actual or potential) data.

Note:

- There is no implication that the model is a physically correct description of the data generating mechanism. “All models are wrong, some models are useful” (G. E. P. Box).
- The aim is to represent the process which generates the actual data (including measurement error, effects of imprecise experimentation, etc.), not an idealised process which would produce “perfect” data.
- Actual data are subject to influence from more sources than we can ever hope to identify. The stochastic element of the model aims at describing all of these influences.

Simple statistical models describe a single variable, e.g. $Y_i \sim N(\mu, \sigma^2); i = 1, \ldots, n$, with all pairs of random variables (r.v.s) independent, where $y_1, \ldots, y_n$ are the heights of a sample of 6-year old boys, assumed to be realisations of r.v.s $Y_1, \ldots, Y_n$.

A typical model for a variable will include:

- an assumed family of probability distributions (e.g. the normal distribution);
- a number of unknown parameters of the distribution (e.g. μ and σ^2).

Notes:

- Usually, we will use the data to estimate the unknown parameters of the model and to draw inferences about them.
- It is also usually important to use the data to question the assumptions of the model, e.g. the assumed distribution.
- The same model can be represented by different parameterisations, e.g. σ instead of σ^2.
- It is most common to use parameters which are closely related to the central moments of the distribution, i.e. mean (location), variance and covariance (dispersion), etc.
- We might try to estimate and draw inferences about parameters relating to the central moments without assuming a particular probability distribution (so-called semi-parametric inference).

In this course, we will be considering models for datasets having more than one variable. We will be interested in relationships among the variables.

In particular, there will be one or more response variables, $Y_1, \ldots, Y_m$, variation in which are of primary interest, and one or more explanatory variables, $X_1, \ldots, X_q$, whose relationships with the response variables we must describe.

- In this course, we will mostly restrict ourselves to a single response variable, Y. In practice, multivariate responses are often analysed separately (especially if they are measuring physically different things).
- Y might be continuous, typically on $\mathbb{R}$ or $\mathbb{R}^+$, or discrete, typically on $\mathbb{Z}^+ \bigcup \{0\}$ or $\{0, 1, \ldots, m\}$.
- Y might represent the natural variable on a transformed scale, e.g. log Height.
- X_r might be quantitative, either continuous or discrete, or qualitative, either ordered, partially ordered or unordered.

An important distinction is between explanatory variables whose values are:

- controlled in an experiment;
- used as a basis for selecting these particular sampling units in a survey;
- unknown at the time the sample is selected.

This distinction determines the types of conclusions we can draw from the results of our analysis. However, it has less impact on how the analysis is carried out; in particular, we use the same methodology for estimation in each case.

Conceptually, we will always assume that our data are recorded in this form:

	Variable				
Observation	X_1	X_2	$\ldots$	X_q	Y
1	x_{11}	x_{21}	$\ldots$	x_{q1}	y_1
2	x_{12}	x_{22}	$\ldots$	x_{q2}	y_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
n	x_{1n}	x_{2n}	$\ldots$	x_{qn}	y_n

However the data were collected, it is usual to model relationships between Y and the explanatory variables by:

- assuming a family of probability distributions (e.g. the normal distribution) for Y;
- assuming that (some or all of) the unknown parameters of the distribution depend on the observed values of the explanatory variables (e.g. $\mu = \mu(x)$ and σ^2 is constant);
- assuming some functional form for the relationship between the parameters and the explanatory variables (e.g. $\mu(x) = \beta_0 + \beta_1 x$).

Again we might try to do without the distributional assumption.

The Linear Model

A model which can be written as

$$E(Y_i) = \mu_i = \sum_{s=1}^{p} \beta_s f_s(\mathbf{x}_i)$$

and $V(Y_i) = \sigma^2$, with all r.v.s uncorrelated, where $\beta_1, \ldots, \beta_p$ are unknown real-valued parameters and $f_1(\cdot), \ldots, f_p(\cdot)$ are functions of the levels of the explanatory variables, is called a linear model, or a general linear model.

If we additionally assume that Y_i has a normal distribution, it is a normal linear model.

Note:

- The model is linear *in the parameters*; $f_s(\mathbf{x})$ can be a nonlinear function of $\mathbf{x}$.
- The model *can* be written in this form, but it need not be.
- The parameters are often labelled in a different way, e.g. $\beta_0, \beta_1, \ldots, \beta_{p-1}$.
- We might broaden the assumption from the process which produced our data to the process more generally, i.e.

$$E(Y|\mathbf{X} = \mathbf{x}) = \sum_{s=1}^{p} \beta_s f_s(\mathbf{x}),$$

which allows us, for example, to predict future observations or to consider counterfactuals, i.e. data which could have been collected, but were not.

The model is usefully written in matrix form,

$$E(\mathbf{Y}) = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}; \quad V(\mathbf{Y}) = \sigma^2\mathbf{I},$$

where $\mathbf{Y}' = [Y_1, \ldots, Y_n]$, $\boldsymbol{\mu}' = [\mu_1, \ldots, \mu_n]$, $\mathbf{X}$ is an $n \times p$ matrix with ith row given by $[f_1(\mathbf{x}_i), \ldots, f_p(\mathbf{x}_i)]$ and $\boldsymbol{\beta}' = [\beta_1, \ldots, \beta_p]$.

Then to specify a particular model, we need only specify a general row i of $\mathbf{X}$ (and the labelling of the parameters in $\boldsymbol{\beta}$).

The model can also be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad E(\boldsymbol{\epsilon}) = \mathbf{0}; \; V(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I},$$

where $\boldsymbol{\epsilon}' = [\epsilon_1, \ldots, \epsilon_n]$ and the ϵ_is are known as error terms.

This form of the model is convenient for obtaining some theoretical results, but mainly it is useful for model checking.

A plot of the data $y_1, \ldots, y_n$ does not tell us anything about whether they come from a normal distribution or have constant variance, since the model states that they are from different distributions.

On the other hand, a plot of residuals (which estimate the ϵ_is) can allow such model checking, since they should all be from the same distribution.

We will now consider some commonly used examples of linear models.

The simple linear regression model:

$$\mu_i = \beta_0 + \beta_1 x_i.$$

The multiple linear regression model:

$$\mu_i = \beta_0 + \sum_{r=1}^{q} \beta_r x_{ri}.$$

The second order polynomial regression model:

$$\mu_i = \beta_0 + \sum_{r=1}^{q} \beta_r x_{ri} + \sum_{r=1}^{q} \beta_{rr} x_{ri}^2 + \sum_{r=1}^{q-1} \sum_{s=r+1}^{q} \beta_{rs} x_{ri} x_{si},$$

sometimes called the second order response surface model.

Transformed regression models, e.g.

$$\mu_i = \beta_0 + \beta_1 \log x_i$$

or

$$\mu_i = \beta_0 + \beta_1 \sqrt{x_i}.$$

A subtle point is that these are linear models, *given the transformation used*. If the transformation is itself to be estimated from the data, e.g. which power of x to use, then these models become nonlinear.

Analysis of variance models: if the explanatory variables are qualitative, we use indicator variables to represent their effects. For example, if there are m categories, we can write

$$\mu_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi},$$

where

$$x_{ri} = \begin{cases} 1 & \text{if observation } i \text{ is from category } r; \\ 0 & \text{otherwise,} \end{cases}$$

so that β_r represents the expected response from category r.

More usually, we write

$$\mu_i = \beta_0 + \beta_2 x_{2i} + \cdots + \beta_m x_{mi},$$

so that category 1 is treated as a baseline and β_r represents the difference in expected response between category r and category 1.

This parameterisation makes it simpler to include many explanatory variables, some qualitative, some quantitative, in the model.

We can also deal with the model

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}; \quad V(\mathbf{Y}) = \sigma^2\mathbf{G}, \tag{1}$$

where $\mathbf{G}$ is a *known*, nonsingular, matrix.

There exists a unique symmetric nonsingular matrix $\mathbf{G}^{1/2}$ such that $\mathbf{G} = \mathbf{G}^{1/2}\mathbf{G}^{1/2}$.

Considering $\mathbf{Z} = \mathbf{G}^{-1/2}\mathbf{Y}$, we have

$$E(\mathbf{Z}) = \mathbf{G}^{-1/2}\mathbf{X}\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\beta}; \quad V(\mathbf{Z}) = \sigma^2\mathbf{I},$$

so that model (1) is a linear model in $\mathbf{Z}$ from which we can estimate (and draw inferences about) the parameters of model (1) for $\mathbf{Y}$.

This is useful in some applications, but not universally, since in practice it is unusual to know $\mathbf{G}$.

Justifications for the Linear Model

Why should we use linear models?

If the functional form relating the parameters of Y to $\mathbf{x}$ is known and nonlinear, then we should not use linear models.

However, in many areas of scientific investigation and most areas of clinical medicine, social sciences and business, there is no known model.

Then if we locally approximate the true, but unknown, function

$$\mu(\mathbf{x}) = g(\mathbf{x}; \boldsymbol{\theta}),$$

using a Taylor series expansion, truncated after low order terms, we obtain low order polynomial regression models.

A second justification applies to almost all designed experiments.

In an experiment, a number of treatments will be compared by applying each to a number of experimental units and observing the response from each unit.

The basic model which will (almost) always be assumed is

$$y_{i(r)} = u_i + t_r,$$

where $y_{i(r)}$ is the response on unit i if treatment r is applied, u_i is the response on unit i when a treatment with zero effect is applied and t_r is the effect of treatment r.

The model can also be written

$$y_{i(r)} = \mu + t_r + e_i, \tag{2}$$

where μ is the mean of u_i across the units and e_i is the deviation of unit i from this mean, with $\sum e_i = 0$.

Note:

- this is a deterministic mathematical model, not a statistical model;
- the *only* assumption is additivity of treatment and unit effects;
- $y_{i(r)}$ might represent a transformation of the measured response to ensure additivity.

Now, randomization provides the basis for inference from the designed experiment.

We will set up the theory for a completely randomized design for n experimental units with t treatments each applied to $n_t = \frac{n}{t}$ experimental units.

Perform randomization by:

1. writing down the combinatorial design;
2. randomly allocating units to unit labels.

Over the population of possible randomizations, the deterministic model (2) becomes the stochastic model

$$Y_{i(r)} = \mu + t_r + \sum_{j=1}^{n} \delta_{ij} e_j,$$

where $\delta_{ij} = 1$ if unit j is randomized to unit label i and $\delta_{ij} = 0$ otherwise.

$\sum_{j=1}^{n} e_j = 0$ and $\sum_{j=1}^{n} e_j^2 = (n-1)\sigma^2$, where σ^2 is the (population) variance of the e_j.

Then

$$\sum_{j=1}^{n} \sum_{\substack{l=1 \\ l \neq j}}^{n} e_j e_l = -\sum_{j=1}^{n} e_j^2 = -(n-1)\sigma^2.$$

We also have

$$
\begin{aligned}
E(\delta_{ij}) &= \frac{1}{n}; \\
E(\delta_{ij}^2) = \frac{1}{n} \quad &\Rightarrow \quad Var(\delta_{ij}) = \frac{n-1}{n^2}; \\
E(\delta_{ij}\delta_{il}) = 0 \quad &\Rightarrow \quad Cov(\delta_{ij}, \delta_{il}) = -\frac{1}{n^2}; \\
E(\delta_{ij}\delta_{kj}) = 0 \quad &\Rightarrow \quad Cov(\delta_{ij}, \delta_{kj}) = -\frac{1}{n^2}; \\
E(\delta_{ij}\delta_{kl}) = \frac{1}{n(n-1)} \quad &\Rightarrow \quad Cov(\delta_{ij}, \delta_{kl}) = \frac{1}{n^2(n-1)}.
\end{aligned}
$$

Letting $\epsilon_i = \sum_{j=1}^{n} \delta_{ij} e_j$, the model is

$$Y_{i(r)} = \mu + t_r + \epsilon_i, \tag{3}$$

where the variance-covariance structure of $\boldsymbol{\epsilon}$ is known.

Hence the deterministic model (2) leads inevitably, under randomization, to a linear model.

The theory extends to designs which use more complicated forms of restricted randomization, such as block designs, row-column designs, split-plot designs, etc.

These give model (3), but with a more complicated variance-covariance structure.

A similar justification for linear models can be obtain through sampling theory for designed surveys.

Hence (in carefully planned statistical investigations) if the levels of the explanatory variables are known before the data on the response variables are collected, a linear model is *the correct model*, at least for an initial analysis of the data.

In studies in which the values of the explanatory variables are unknown before the data are collected, it is rare for the true model form to be known, so a local polynomial approximation is often appropriate.

Hence linear models have enormously wide application.

Estimation

We have the defined the general linear model, most usefully written in matrix form

$$E(\mathbf{Y}) = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}; \quad V(\mathbf{Y}) = \sigma^2\mathbf{I},$$

where $\mathbf{Y}' = [Y_1, \ldots, Y_n]$, $\boldsymbol{\mu}' = [\mu_1, \ldots, \mu_n]$, $\mathbf{X}$ is an $n \times p$ matrix with ith row given by $[f_1(\mathbf{x}_i), \ldots, f_p(\mathbf{x}_i)]$ and $\boldsymbol{\beta}' = [\beta_1, \ldots, \beta_p]$.

Now we concentrate on estimating $\boldsymbol{\beta}$. We will discuss the estimation of σ^2 later.

The idea of least squares estimation is that we find an estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ which minimises

$$S = \sum_{i=1}^{n} (Y_i - \mu_i)^2.$$

This can be written

$$S = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

and we find the minimum by differentiating S with respect to $\boldsymbol{\beta}$ and equating to zero.

Writing

$$S = \mathbf{Y}'\mathbf{Y} - 2\beta'\mathbf{X}'\mathbf{Y} + \beta'\mathbf{X}'\mathbf{X}\beta,$$

we differentiate with respect to β to get

$$\frac{dS}{d\beta} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\beta$$

and equating to $\mathbf{0}$ gives the normal equations,

$$\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{Y}. \tag{4}$$

Since $\mathbf{X}'\mathbf{y}$ is in the vector space generated by the columns of $\mathbf{X}'\mathbf{X}$, a result from linear algebra implies that the normal equations always have a real solution.

Let $\hat{\boldsymbol{\beta}}$ be any solution of equation (4). Then, for any $\boldsymbol{\beta}$,

$$\begin{aligned} S &= (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) \\ &= \{\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\}'\{\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\} \\ &= (\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})+(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\ &\geq (\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}) \end{aligned}$$

and so all solutions $\hat{\boldsymbol{\beta}}$ of the normal equations minimise S. The minimum is

$$(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}),$$

which is the residual sum of squares.

For any $p \times 1$ vector $\mathbf{c}$, $\mathbf{c}'\hat{\boldsymbol{\beta}}$ is defined to be a least squares estimator of $\mathbf{c}'\boldsymbol{\beta}$, where $\hat{\boldsymbol{\beta}}$ is any solution of equation (4).

Example (simple linear regression): If $\boldsymbol{\beta}' = [\beta_0 \ \beta_1]$ and

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix},$$

then

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

and

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \sum Y_i \\ \sum x_i Y_i \end{bmatrix}.$$

The normal equations, then, are

$$\left.\begin{aligned} \beta_0 n + \beta_1 \sum x_i &= \sum Y_i \\ \beta_0 \sum x_i + \beta_1 \sum x_i^2 &= \sum x_i Y_i \end{aligned}\right\}. \tag{5}$$

Consider the case where $x_i = x \ \forall i = 1, \ldots, n$. Then equations (5) reduce to the single equation

$$\beta_0 + \beta_1 x = \bar{Y}$$

and any $\hat{\beta}_0$ and $\hat{\beta}_1$ which satisfy this equation are least squares estimators of β_0 and β_1.

The function of parameters $\mathbf{c}'\boldsymbol{\beta}$, where $\mathbf{c}' = [1 \ x]$, has a unique least squares estimator $\mathbf{c}'\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x = \bar{Y}$.