LINEAR MODEL

ARTHUR ALLIGNOL

INTRODUCTION

INTRODUCTION

Linear modelling is used for explaining the relationship between a single quantitive variable Y and one or more explanatory variables $X_1, \ldots, X_q$

- Y is also called the *response, outcome, dependent variable*
- The X's are the *predictors, independent, explanatory* variables
  - They can be of any type (continuous, ordinal, ...)

A very general model would be

$$Y = f(X_1, \ldots, X_q) + \varepsilon,$$

with f some unknown function and ε the error. Assume a more restricted form for the model:

$$Y = \beta_0 + \beta_1 X1 + \cdots + \beta_q X_q + \varepsilon,$$

which is a linear model

INTRODUCTION

Assume y_i represents the value of the response variable on the ith individual, and that $x_{i1}, x_{i2}, \ldots, x_{iq}$ represents the individual's values on p explanatory variables, with $i = 1, \ldots, n$.

The multiple linear regression model is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq} + \varepsilon_i,$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

INTRODUCTION

Thus, the distribution of Y is also normal with expected value given by the linear combination of the explanatory variables

$$E(y|x_1, \ldots, x_q) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$$

and with variance σ^2.

The parameters of the model $\beta_k, k = 1, \ldots, q$, are known as regression coefficients with β_0 the *intercept*

INTRODUCTION

It will be convenient to collect the n responses $\mathbf{y}$, viewed as a realisation of a random vector $\mathbf{Y}$ with mean

$$\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mu}$$

and variance-covariance matrix

$$\mathrm{var}(\mathbf{Y}) = \sigma^2\mathbf{I}$$

Under the assumption of normality, $\mathbf{Y}$ has a multivariate normal distribution

$$\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$$

INTRODUCTION

The model can be rewritten in matrix form as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon,$$

or equivalently

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta},$$

- $\mathbf{X}$ a $n \times p$ matrix containing the values of the q predictors and the intercept
- $\mathbf{X}$ usually called *model matrix* or *design matrix*
- $\mathbf{X}\boldsymbol{\beta}$ is called *linear predictor*

ESTIMATION

ESTIMATION OF β

Considering i.i.d observations and Y normally distributed, the log-likelihood is

$$\log L(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum(y_i - x_i'\beta)^2/\sigma^2$$

Maximising the log-likelihood wrt β for a fixed value of σ^2 is equivalent to minimising the sum of squared differences between observed and expected values, or *residual sum of squares*

$$\mathrm{RSS}(\beta) = \sum(y_i - x_i'\beta)^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

ESTIMATION OF β

Taking derivatives of the residual sum of squares with respect to β and setting the derivative equal to zero leads to the so-called *normal equations* for the maximum-likelihood estimator $\hat{\beta}$

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\mathbf{y}$$

If the model matrix $\mathbf{X}$ has full rank, i.e., $\mathbf{X}'\mathbf{X}$ is non-singular, then the *ordinary least squares* (OLS) or maximum likelihood estimator of the linear parameters is

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

PROPERTIES OF THE ESTIMATOR β

If the model is correct

$$\mathbf{E}(\hat{\beta}) = \beta.$$

Still assuming i.i.d observations and constant variance σ^2, the variance-covariance matrix of the OLS estimator is

$$\mathbf{var}\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2$$

In large samples, $\hat{\beta} \sim \mathcal{N}(\beta, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2)$

ESTIMATION OF σ^2

Plugging-in $\hat{\beta}$ in the log-likelihood, we obtain

$$\log \mathrm{L}(\sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum(y_i - x_i'\hat{\beta})^2/\sigma^2$$

The maximum likelihood estimator is

$$\hat{\sigma}^2 = \sum(y_i - x_i'\hat{\beta})^2/n$$

Dividing by $n - p$ instead of n leads to an unbiased estimator

ILLUSTRATION: HUBBLE DATA SET

From Freedman et al. (2001)

- Relative velocity and the distance of 24 galaxies according to measurements made by the Hubble telescope
- `velocity` assessed by measuring the Doppler red shift (km/s)
- `distance` (Mega parsecs $= 3.09 \times 10^{19}$ km)

Aim is to derive the age of the universe

HYPOTHESIS TESTING

WALD TEST

Test on one particular coefficient, e.g.,

$$H_0 : \beta_j = 0$$

Under H_0, $\hat{\beta}_j$ has a distribution with mean 0 and variance given by the j-th diagonal element of $\mathrm{var}\hat{\beta}$. The test is based on the ratio

$$t = \frac{\hat{\beta}_j}{\sqrt{\mathrm{var}\hat{\beta}_j}}$$

which follows a *Student* distribution with $n - p$ degrees of freedom under the null hypothesis when σ^2 is estimated (which is the case in practice)

Confidence intervals for the coefficients can be constructed based on the Wald test, i.e., we can state with $100(1 - \alpha)$ confidence that $\hat{\beta}_j$ is between the bounds

$$\hat{\beta}_j \pm t_{1-\alpha/2;n-p} \sqrt{\mathrm{var}\hat{\beta}_j}$$

LIKELIHOOD RATIO TEST

Consider testing the joint significance of several coefficients, e.g.,

$$H_0 : \beta_2 = 0,$$

where $\beta = (\beta_1, \beta_2)$. Anagously, we partition the design matrix into $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$. The hypothesis of interest states that the response does not depend on the last p_2 predictors.

The idea of the likelihood ratio test is

1. fit two nested models: a smaller model with the first p_1 predictors in $\mathbf{X}_1$, and a larger model with all p predictors in $\mathbf{X}$
2. compare their maximized likelihoods (or log-likelihoods)

LIKELIHOOD RATIO TEST

We fit the smaller model with $\mathbf{X}_1$ as predictors. The maximised log-likelihood is

$$\max \log L(\beta_1) = c - \frac{1}{2}\mathrm{RSS}(\mathbf{X}_1)/\sigma^2,$$

with $c = -(n/2)\log(2\pi\sigma^2)$

The maximised likelihood for the model including all predictors $\mathbf{X}_1 + \mathbf{X}_2$ is

$$\max \log L(\beta_1, \beta_2) = c - \frac{1}{2}\mathrm{RSS}(\mathbf{X}_1 + \mathbf{X}_2)/\sigma^2,$$

LIKELIHOOD RATIO TEST

The likelihood ratio criterion is

$$-2\log\Lambda = \frac{\mathrm{RSS}(\mathbf{X}_1) - \mathrm{RSS}(X_1 + X_2)}{\sigma^2}$$

(where Λ is the actual ratio of likelihoods)

- The idea is to look at the reduction in the residual sum of squares when adding the predictors in $\mathbf{X}_2$ in the model
- This reduction is compared to the error variance σ^2

In practice we calculate the criterion using $\hat{\sigma}^2 = \mathrm{RSS}(\mathbf{X}_1 + X_2)/(n-p)$

- $-2\log\Lambda \sim \chi^2_{p_2}$

F-TEST

Under the assumption of normality, $-2\log\Lambda$ has an *exact* chi-squared distribution with p_2 d.f if σ^2 is known

If σ^2 is estimated, the criterion divided by p_2, i.e.,

$$F = \frac{(\mathrm{RSS}(\mathbf{X}_1) - \mathrm{RSS}(\mathbf{X}_1 + \mathbf{X}_2))/p_2}{\mathrm{RSS}(\mathbf{X}_1 + \mathbf{X}_2)/(n-p)}$$

has an exact F distribution with p_2 and $n-p$ d.f

ANOVA TABLE

Source of variation	Sum of squares	Degrees of freedoms
$\mathbf{X}_1$	$\mathrm{RSS}(\emptyset) - \mathrm{RSS}(\mathbf{X}_1)$	$p_1 - 1$
$\mathbf{X}_2$ given $\mathbf{X}_1$	$\mathrm{RSS}(\mathbf{X}_1) - \mathrm{RSS}(\mathbf{X}_1 + \mathbf{X}_2)$	p_2
Residual	$\mathrm{RSS}(\mathbf{X}_1 + \mathbf{X}_2)$	$n - p$
Total	$\mathrm{RSS}(\emptyset)$	$n - 1$

- More detailed analysis of variance tables may be obtained by introducing the predictors one at a time, while keeping track of the reduction in residual sum of squares at each step

CONTRASTS

CONTRASTS

- Linear regression can be extended to accomodate categorical variables (factors) using *dummy variables* or *contrasts*
- Below a categorical variable is represented by a dummy regressor D, coded 1 for one category, 0 for the other

$$y_i = \beta_0 + \beta X_i + \gamma D_i + \varepsilon_i$$

- This fits two regression lines with the same slope but different intercepts, i.e., γ represents the constant separation between 2 regression lines

$$y_i = \beta_0 + \beta X_i + \gamma(0) + \varepsilon_i = \beta_0 + \beta X_i + \varepsilon_i$$
$$y_i = \beta_0 + \beta X_i + \gamma(1) + \varepsilon_i = \beta_0 + \gamma + \beta X_i + \varepsilon_i$$

CONTRASTS FOR CATEGORICAL VARIABLES

- A variable with m categories has $m-1$ regressors
- As with the two category case, one of the categories is a reference group (coded 0 for all dummy regressors), e.g.,

Category	D_1	D_2
category 1	0	0
category 2	1	0
category 3	0	1

The regression model is then

$$y_i = \beta_0 + \beta X_i + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \varepsilon_i$$

CONTRASTS FOR CATEGORICAL VARIABLES

This model describes 3 parallel regression lines, which can differ in their intercepts

- Category 3: $y_i = (\beta_0 + \gamma_2) + \beta X_i + \varepsilon_i$
- Category 2: $y_i = (\beta_0 + \gamma_1) + \beta X_i + \varepsilon_i$
- Category 1: $y_i = \beta_0 + \beta X_i + \varepsilon_i$

"category 1" is coded as 0 for both dummy regressors, thus serves as *baseline* category with which the other categories are compared

These types of contrasts are called *treatment contrasts*

CHOICE OF BASELINE CATEGORY

The choice of reference category is technically irrelevant, but there are two possible consideration

- Theory may suggest we compare to a particular category, e.g., non smoker VS 10 cigarettes a day, and non-smokers VS a pack/day
- The largest category gives the smallest standard errors

In R, factors take care of the dummy coding in the design matrix. By default the reference level is the lowest level in alphabetical order.

In SAS, the `class` statement is used. Reference level is the last one.

N.B.: other types of contrasts exist. In particular, constrats can be defined in such a way as to test specific hypotheses

REGRESSION DIAGNOSTICS

REGRESSION DIAGNOSTICS

The estimation of and inference from the regression model depend on several assumptions. These should be checked using *regression diagnostics*.

The potential problems:

- **Error:** We assume $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$
- **Model:** We assume that the structural part of the model, $\mathrm{E}y = X\beta$, is correct
- **Unusual observations:** Sometimes just a few observations do not fit the model. These few observations might change the choice and fit of the model

DIAGNOSTIC PLOTS

- A plot of residuals against each explanatory variable in the model. The presence of a non-linear relationship, for example, may suggest that a higher-order term in the explanatory variable should be considered.
- A plot of residuals against fitted values. If the variance of the residuals appears to increase with predicted value, a transformation of the response variable may be in order.
- A normal probability plot of the residuals. After all the systematic variation has been removed from the data, the residuals should look like a sample from a standard normal distribution. A plot of the ordered residuals against the expected order statistics from a normal distribution provides a graphical check of this assumption.

COOK'S DISTANCE

A further diagnostic that is often very useful is an index plot of the Cook's distances for each observation. This statistic is defined as

$$D_k = \frac{1}{(p+1)\hat{\sigma}^2} \sum_{i=1}^{n} (\hat{y}_{i(k)} - y_i)^2$$

where $\hat{y}_{i(k)}$ is the fitted value of the ith observation when the kth observation is omitted from the model.

The values of D_k assess the impact of the kth observation on the estimated regression coefficients. Values of D_k greater than one are suggestive that the corresponding observation has undue influence on the estimated regression coefficients.

PARTIAL RESIDUAL PLOT

Construct the response with the predicted effect of the other X removed

$$y - \sum_{j \neq i} x_j \hat{\beta}_j = \hat{y} + \hat{\varepsilon} - \sum_{j \neq i} x_j \hat{\beta}_j = x_i \hat{\beta}_i + \hat{\varepsilon}$$

The partial residual plot is then $x_i \hat{\beta}_i + \hat{\varepsilon}$ against x_i. Additionally the `termplot` function in R centers the x_i.

- Show the relationship between a given independent variable and the response variable given that other independent variables are also in the model.

ILLUSTRATIONS

Hubble data set

A complete example: Insurance redlining

Insurance redlining refers to the practice of refusing to issue insurances to certain types of people or within some geographical area.

In the 70s, several Chicago community organisations accused insurances of redlining their neighbourhood

To assess whether redlining was taking place, the number of FAIR (Fair Access to Insurance Requirements) plan policies by zip code were collected as well as some statistics at the zip code level

- `race` racial composition in percent minority
- `fire` fires per 100 housing units
- `theft` theft per 1000 population
- `age` percent of housing units built before 1939
- `involact` new FAIR plan policies and renewals per 100 housing units
- `income` median family income in thousands of dollars

- `side` North or South side of Chicago

EXTENSIONS

INTERACTIONS

Two explanatory variables interact in determining a response variable when the partial effect of one depends on the value of the other

- If the regressions in different categories of a qualitative explanatory variable are not parallel, then the qualitative variable interacts with one or more of the quantitative variables
- The dummy-regression model can be modified to reflect interactions

The following model accomodates different intercepts and slopes for different categories (here 2)

$$y_i = \beta_0 + \beta X_i + \gamma D_i + \delta(D_i X_i) + \varepsilon_i$$

INTERACTIONS

- The interaction regressor is the product of the other two regressors.
- For category 1 (coded 0)

$$\begin{aligned} y_i &= \beta_0 + \beta X_i + \gamma(0) + \delta(0 \cdot X_i) + \varepsilon_i \\ &= \beta_0 + \beta X_i + \varepsilon_i \end{aligned}$$

- For category 2 (coded 1)

$$\begin{aligned} y_i &= \beta_0 + \beta X_i + \gamma(1) + \delta(1 \cdot X_i) + \varepsilon_i \\ &= (\beta_0 + \gamma) + (\beta + \delta) X_i + \varepsilon_i \end{aligned}$$

- β_0 and β are the intercept and slope for the regression of X among category 1
- γ gives the difference in intercept between the categories
- δ gives the difference in slopes between the groups

INTERACTIONS

Interactions between 2 factors:

The model with interactions corresponds to having different levels of the two factors.

Interactions between continuous variables:

Tricky. The interpretation is that you have a linear effect of varying one variable while keeping the other constant, but with a slope that changes as you vary the other variable.

N.B. If you include interactions in the model, it is better to also include the main effects, even if the effects are small.

ILLUSTRATION

Anorexia data set

Weight change data for young female anorexia patients following treatment.

- `Treat` with levels `Cont`: Control, `CBT`: Cognitive Behavioural treatment and `FT`: family treatment
- `Prewt` Weight of patient before study period, in lbs
- `Postwt` Weight of patient after study period, in lbs

TRANSFORMATIONS OF THE RESPONSE AND/OR PREDICTORS

Transformation of the response and/or predictors can improve the fit of the model and correct violations of model assumptions such as non-constant error variance.

We may also consider adding additional predictors that are functions of the existing predictors, quadratic terms

TRANSFORMATION OF THE RESPONSE

Suppose we want to use a log transformation of the response

$$\log y = \beta_0 + \beta_1 X + \varepsilon.$$

In the original scale of the response, the model becomes

$$y = \exp(\beta_0 + \beta_1 X)\exp(\varepsilon)$$

The error enters the model *multiplicatively*. Whether that's sensible is a trial and error thing (redo the diagnostic plots and see)

TRANSFORMATION OF THE RESPONSE

- Regression coefficients are interpreted in the transformed scale (not always easy)
- Regression coefficients for models where the transformations are different can't be compared

Use with care!

But note that a log-transformation allows for a meaningful interpretation of the regression coefficients

$$\log \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$
$$\hat{y} = e^{\hat{\beta}_0} e^{\hat{\beta}_1 X_1} \ldots e^{\hat{\beta}_p X_p}$$

An increase of one X_1 would multiply the predicted response by $e^{\hat{\beta}_1}$

TRANSFORMATION OF THE PREDICTORS — POLYNOMIALS

Includes second order and higher powers of a variable in the model along with the original linear term. That is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \varepsilon$$

- Non linear relation between y and x
- but the model is still linear

SPLINES

We consider cubic splines

- Divide the real line by an ordered set of points $\{z_i\}$ known as *knots*
- On the interval $[z_i, z_{i+1}]$ the spline is a cubic polynomial
- Impose continuity, and continuous first and second derivatives to ensure smoothness

CORRECT FOR HETEROSKEDASTICITY

- Heteroskedasticity does not bias the coefficient estimates; though not efficient
- but the standard errors are incorrect, so are the confidence intervals and p-values

We need robust estimates of the variance of the regression coefficients.

Won't be detailed here but are available in, e.g., R-package **sandwich**

ILLUSTRATIONS

The simulated data