Statistical Models in R

Some Examples

Steven Buechler

Department of Mathematics
276B Hurley Hall; 1-6233

Fall, 2007

Outline

Statistical Models

First Principles

In a couple of lectures the basic notion of a statistical model is described. Examples of anova and linear regression are given, including variable selection to find a simple but explanatory model. Emphasis is placed on *R*'s framework for statistical modeling.

General Problem

addressed by modelling

Given: a collection of variables, each variable being a vector of readings of a specific trait on the samples in an experiment.

Problem: In what way does a variable Y depend on other variables $X_1, \ldots, X_n$ in the study.

Explanation: A statistical model defines a mathematical relationship between the X_i's and Y. The model is a representation of the real Y that aims to replace it as far as possible. At least the model should capture the dependence of Y on the X_i's

General Problem

addressed by modelling

Given: a collection of variables, each variable being a vector of readings of a specific trait on the samples in an experiment.

Problem: In what way does a variable Y depend on other variables $X_1, \ldots, X_n$ in the study.

Explanation: A statistical model defines a mathematical relationship between the X_i's and Y. The model is a representation of the real Y that aims to replace it as far as possible. At least the model should capture the dependence of Y on the X_i's

General Problem

addressed by modelling

Given: a collection of variables, each variable being a vector of readings of a specific trait on the samples in an experiment.

Problem: In what way does a variable Y depend on other variables $X_1, \ldots, X_n$ in the study.

Explanation: A statistical model defines a mathematical relationship between the X_i's and Y. The model is a representation of the real Y that aims to replace it as far as possible. At least the model should capture the dependence of Y on the X_i's

The Types of Variables

in a statistical model

The response variable is the one whose content we are trying to model with other variables, called the explanatory variables.

In any given model there is one response variable (Y above) and there may be many explanatory variables (like $X_1, \ldots . X_n$).

Identify and Characterize Variables

the first step in modelling

- Which variable is the response variable;
- Which variables are the explanatory variables;
- Are the explanatory variables continuous, categorical, or a mixture of both;
- What is the nature of the response variable — is it a continuous measurement, a count, a proportion, a category, or a time-at-death?

Identify and Characterize Variables

the first step in modelling

- Which variable is the response variable;
- Which variables are the explanatory variables;
- Are the explanatory variables continuous, categorical, or a mixture of both;
- What is the nature of the response variable — is it a continuous measurement, a count, a proportion, a category, or a time-at-death?

Identify and Characterize Variables

the first step in modelling

- Which variable is the response variable;
- Which variables are the explanatory variables;
- Are the explanatory variables continuous, categorical, or a mixture of both;
- What is the nature of the response variable — is it a continuous measurement, a count, a proportion, a category, or a time-at-death?

Identify and Characterize Variables

the first step in modelling

- Which variable is the response variable;
- Which variables are the explanatory variables;
- Are the explanatory variables continuous, categorical, or a mixture of both;
- What is the nature of the response variable — is it a continuous measurement, a count, a proportion, a category, or a time-at-death?

Types of Variables Determine Type of Model

The explanatory variables

All explanatory variables continuous	Regression
All explanatory variables categorical	Analysis of variance (Anova)
Explanatory variables both continuous and categorical	Analysis of covariance (Ancova)

Types of Variables Determine Type of Model

The response variable — what kind of data is it?

Continuous	Normal Regression, Anova, Ancova
Proportion	Logistic regression
Count	Log linear models
Binary	Binary logistic analysis
Time-at-death	Survival analysis

Model Formulas

Which variables are involved?

A fundamental aspect of models is the use of model formulas to specify the variables involved in the model and the possible interactions between explanatory variables included in the model.

A model formula is input into a function that performs a linear regression or anova, for example.

While a model formula bears some resemblance to a mathematical formula, the symbols in the "equation" mean different things than in algebra.

Common Features

of model formulas

Model formulas have a format like

```
> Y ~ X1 + X2 + Z * W
```

where `Y` is the explanatory variable, $\sim$ means "is modeled as a function of" and the right hand side is an expression in the explanatory variables.

First Examples of Model Formulas

Given continuous variables `x` and `y`, the relationship of a linear regression of `y` on `x` is described as

```
> y ~ x
```

The actual linear regression is executed by

```
> fit <- lm(y ~ x)
```

First Examples of Model Formulas

If `y` is continuous and `z` is categorical we use the same model formula

```
> y ~ z
```

to express that we'll model `y` as a function of `z`, however in this case the model will be an anova, executed as

```
> fit <- aov(y ~ z)
```

Multiple Explanatory Variables

Frequently, there are multiple explanatory variables invovled in a model. The + symbol denotes inclusion of additional explanatory variables. The formula

```
> y ~ x1 + x2 + x3
```

denotes that y is modeled as a function of `x1, x2, x3`. If all of these are continuous,

```
> fit <- lm(y ~ x1 + x2 + x3)
```

executes a mutliple linear regression of `y` on `x1, x2, x3`.

Other Operators in Model Formulas

In complicated relationships we may need to include "interaction terms" as variables in the model. This is common when a model involves multiple categorical explanatory variables. A factorial anova may involve calculating means for the levels of variable `A` restricted to a level of `B`. The formula

```
> y ~ A * B
```

describes this form of model.

Just the Basics

Here, just the basic structure of modeling in *R* is given, using anova and linear regression as examples. See the Crawley book listed in the syllabus for a careful introduction to models of varying forms.

Besides giving examples of models of these simple forms, tools for assessing the quality of the models, and comparing models with different variables will be illustrated.

Outline

Approximate Y

The goal of a model is to approximate a vector Y with values calculated from the explanatory variables. Suppose the Y values are $(y_1, \ldots, y_n)$. The values calculated in the model are called the fitted values and denoted $(\hat{y}_1, \ldots, \hat{y}_n)$. (In general, a "hat" on a quantity means one approximated in a model or through sampling.)

The goodness of fit is measured with the residuals, $(r_1, \ldots, r_n)$, where $r_i = y_i - \hat{y}_i$.

Measure of Residuals

Two obtain a number that measures the overall size of the residuals we use the residual sum of squares, defined as

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

As RSS decreases $\hat{y}$ becomes a better approximation to y.

Residuals: Only Half of the Story

A good model should have predictive value in other data sets and contain only as many explanatory variables as needed for a reasonable fit.

To minimize *RSS* we can set $\hat{y}_i = y_i$, for $1 \leq i \leq n$. However, this "model" may not generalize at all to another data set. It is heavily biased to this sample.

We could set $\hat{y}_i = \bar{y} = (y_1 + \cdots + y_n)/n$, the sample mean, for all i. This has low bias in that other samples will yield about the same mean. However, it may have high variance, that is a large *RSS*.

Residuals: Only Half of the Story

A good model should have predictive value in other data sets and contain only as many explanatory variables as needed for a reasonable fit.

To minimize *RSS* we can set $\hat{y}_i = y_i$, for $1 \leq i \leq n$. However, this "model" may not generalize at all to another data set. It is heavily biased to this sample.

We could set $\hat{y}_i = \bar{y} = (y_1 + \cdots + y_n)/n$, the sample mean, for all i. This has low bias in that other samples will yield about the same mean. However, it may have high variance, that is a large *RSS*.

Residuals: Only Half of the Story

A good model should have predictive value in other data sets and contain only as many explanatory variables as needed for a reasonable fit.

To minimize *RSS* we can set $\hat{y}_i = y_i$, for $1 \leq i \leq n$. However, this "model" may not generalize at all to another data set. It is heavily biased to this sample.

We could set $\hat{y}_i = \bar{y} = (y_1 + \cdots + y_n)/n$, the sample mean, for all i. This has low bias in that other samples will yield about the same mean. However, it may have high variance, that is a large *RSS*.

Bias-Variance Trade-off

Selecting an optimal model, both in the form of the model and the parameters, is a complicated compromise between minimizing bias and variance. This is a deep and evolving subject, although it is certainly settled in linear regression and other simple models.

For these lectures make as a goal minimzing *RSS* while keeping the model as simple as possible.

Just as in hypothesis testing, there is a statistic calculable from the data and model that we use to measure part of this trade-off.

Statistics Measure the Fit

Comparing two models fit to the same data can be set up as a hypothesis testing problem. Let M_0 and M_1 denote the models. Consider as the null hypothesis "M_1 is not a significant improvement on M_0", and the alternative the negation. This hypothesis can often be formulated so that a statistic can be generated from the two models.

Model Comparison Statistics

Normally, the models are nested in that the variables in M_0 are a subset of those in M_1. The statistic often involves the RSS values for both models, adjusted by the number of parameters used. In linear regression this becomes an anova test (comparing variances).

More robust is a likelihood ratio test for nested models. When models are sufficiently specific to define a probability distribution for y, the model will report the log-likelihood, $\hat{L}$. Under some mild assumptions, $-2(\hat{L}_0 - \hat{L}_1)$ follows a chi-squared distribution with degrees of freedom = difference in number of parameters on the two models.

Comparison with Null Model

The utility of a single model M_1 is often assessed by comparing it with the null model, that reflects no dependence of y on the explanatory variables. The model formula for the null model is

```
> y ~ 1
```

signifying that we use a constant to approximate y. The natural constant is the mean of y.

Functions in R that generate models report the statistics that measure it's comparison with the null model.

Outline

Continuous $\sim$ Factors

Analysis of variance is the modeling technique used when the response variable is continuous and all of the explanatory variables are categorical; i.e., factors.

Setup: A continuous variable `Y` is modeled against a categorical variable `A`.

Anova Model Structure: On each level approximate the `Y` values by the mean for that level.

Continuous $\sim$ Factors

Analysis of variance is the modeling technique used when the response variable is continuous and all of the explanatory variables are categorical; i.e., factors.

Setup: A continuous variable `Y` is modeled against a categorical variable `A`.
Anova Model Structure: On each level approximate the `Y` values by the mean for that level.

Anova Model Structure

Suppose `Y` is the vector $(y_1, \ldots, y_m)$. In an anova model the fit value for y_i in level A_j is the mean of the y values in level A_j. So, the fit vector is a vector of level means.

The null model for an anova uses `mean(Y)` to approximate every y_i.

An anova model with two levels is basically a t test. The t test assesses whether it is statistically meaningful to consider the group means as different, or approximate both the global mean.

Setup and Assumptions

for a simple anova

Consider the continuous variable Y, partioned as y_{ij}, the j^{th} observation in factor level i. Suppose there are K levels.

Assumptions:

- The anova is balanced, meaning every level has the same number n of elements. Let $Y_i = \{ y_{ij} : 1 \leq j \leq n \}$
- Each Y_i is normally distributed.
- The variances of the Y_i's are constant.

Setup and Assumptions

for a simple anova

Consider the continuous variable Y, partioned as y_{ij}, the j^{th} observation in factor level i. Suppose there are K levels.

Assumptions:

- The anova is balanced, meaning every level has the same number n of elements. Let $Y_i = \{ y_{ij} : 1 \leq j \leq n \}$
- Each Y_i is normally distributed.
- The variances of the Y_i's are constant.

Setup and Assumptions

for a simple anova

Consider the continuous variable Y, partioned as y_{ij}, the j^{th} observation in factor level i. Suppose there are K levels.

Assumptions:

- The anova is balanced, meaning every level has the same number n of elements. Let $Y_i = \{ y_{ij} : 1 \leq j \leq n \}$
- Each Y_i is normally distributed.
- The variances of the Y_i's are constant.

Sums of Squares

measure the impact of level means

Several sums of squares help measure the overall deviation in the Y values, the deviation in the model and the RSS. Let $\bar{y}_i$ be the mean of Y_i, $\bar{\bar{y}}$ the mean of Y (all values).

$$
\begin{aligned}
SSY &= \sum_{i=1}^{K}\sum_{j=1}^{n}(y_{ij} - \bar{\bar{y}})^2 \\
SSA &= \sum_{i=1}^{K}(\bar{y}_i - \bar{\bar{y}})^2 \\
RSS = SSE &= \sum_{i=1}^{K}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2
\end{aligned}
$$

The Statistic to Assess Fit

is F

The statistic used to assess the model is calculated from SSA and SSE by adjusting for degrees of freedom. Let $MSA = SSA/(K-1)$ and $MSE = SSE/K(n-1)$. Define:

$$F = \frac{MSA}{MSE}.$$

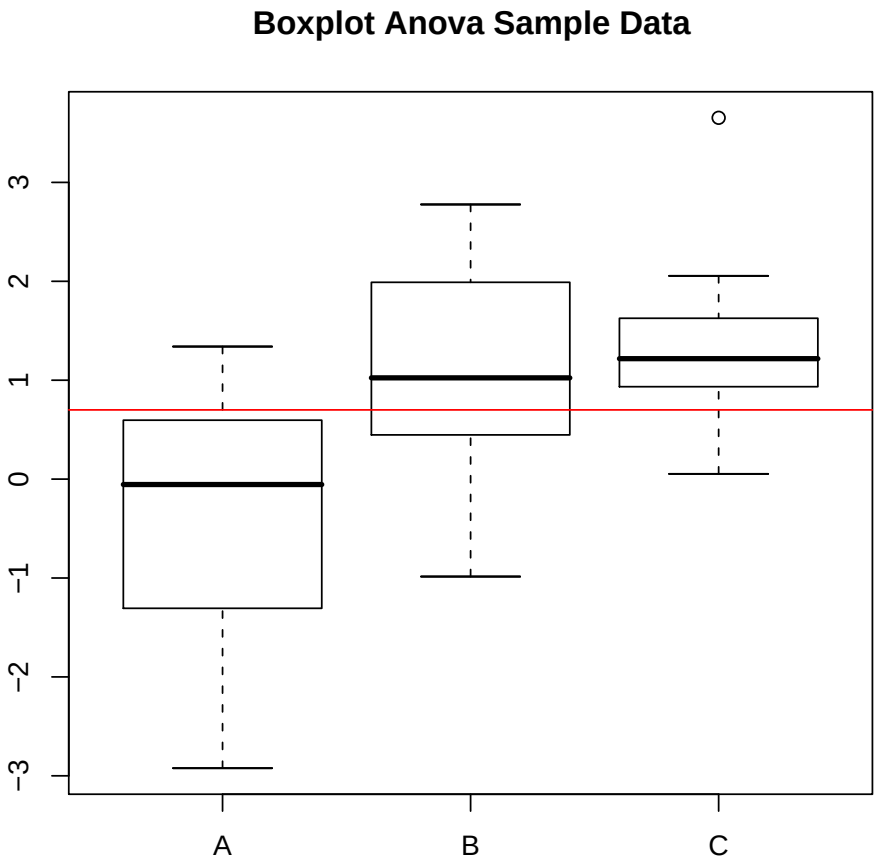
Under all of the assumptions on the data, under the null hypothesis that all of the level means are the same, F satisfies an F distribution with $K-1$ and $K(n-1)$ degrees of freedom.

Anova in R

An anova in *R* is executed with the `aov` function. The sample data have a vector `Y`, of values and an associated factor `LVS` of three levels, each containing 20 samples. As usual, it should be verified that the values on each level are normally distributed and there is a constant variance across the levels. First visualize the relationship between `Y` and `LVS` with a boxplot.

```
> plot(LVS, Y, main = "Boxplot Anova Sample Data")
> abline(h = mean(Y), col = "red")
```

Plot of Sample Data



Execute Anova and Summarize

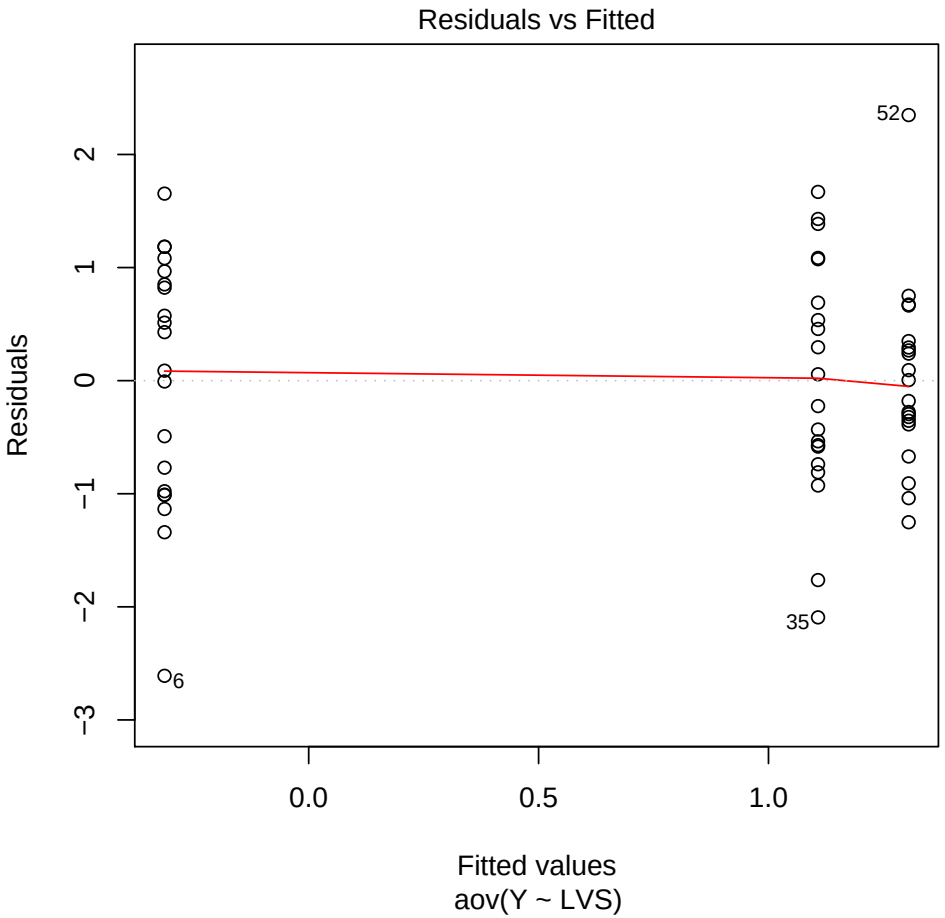
```
> aovFit1 <- aov(Y ~ LVS)
> summary(aovFit1)

             Df Sum Sq Mean Sq F value  Pr(>F)
LVS           2   31.2    15.6    15.8 3.5e-06 ***
Residuals    57   56.3     1.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclude: there is a significant difference in level means.

Plots Assessing the Anova Model

```
> plot(aovFit1)
```



Variables May be Components

Frequently, the variables on which a statistical model is generated are components in a data frame. If the data frame `dat` has components `HT` and `FLD`, an anova could be executed as

```
> aovFld <- aov(HT ~ FLD, data = dat)
```

All statistical modeling formulas have a `data` optional parameter.

Underlying Linear Model

Actually, this analysis of variance is a form of multiple linear regression, with a variable for each level in the factor. Many features of modeling are the same and *R* reflects this.

Underlying Linear Model

```
> summary.lm(aovFit1)

Call:
aov(formula = Y ~ LVS)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6095 -0.6876  0.0309  0.6773  2.3485

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.314      0.222   -1.41     0.16
LVSB           1.421      0.314    4.52  3.1e-05
LVSC           1.618      0.314    5.15  3.4e-06

(Intercept)
LVSB        ***
```

Getting (Almost) Real

The assumptions of a classical anova aren't realistic. Various refined methods handle unbalanced data (levels of different sizes), non-normally distributed data, multiple nested factors, readings within a factor collected over time (longitudinal data), and pseudo-replication.

The types of models to reference are: split-plot, random effects, nested design, mixed models.

These methods can require significant care in defining the model formula and interpreting the result.

Getting (Almost) Real

The assumptions of a classical anova aren't realistic. Various refined methods handle unbalanced data (levels of different sizes), non-normally distributed data, multiple nested factors, readings within a factor collected over time (longitudinal data), and pseudo-replication.

The types of models to reference are: split-plot, random effects, nested design, mixed models.

These methods can require significant care in defining the model formula and interpreting the result.

Getting (Almost) Real

The assumptions of a classical anova aren't realistic. Various refined methods handle unbalanced data (levels of different sizes), non-normally distributed data, multiple nested factors, readings within a factor collected over time (longitudinal data), and pseudo-replication.

The types of models to reference are: split-plot, random effects, nested design, mixed models.

These methods can require significant care in defining the model formula and interpreting the result.

Kruskal-Wallis Test

Just as an anova is a multi-level t test, the Kruskal-Wallis test is a multi-level version of the Mann-Whitney test. This is a non-parametric test that does not assume normality of the errors. It sums rank as in Mann-Whitney. For example, the Kruskal-Wallis test applied to the earlier data is executed by

```
> kTest <- kruskal.test(Y ~ LVS)
```

`kTest` will be an htest object, like that generated by `t.test`.