2021 Advanced Regression Methods for Independent Data

BIOSTAT/STAT 570

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington
jonno@uw.edu

CHAPTER 5: LINEAR MODELS

Outline

Least Squares

Parameter Interpretation and Parameterization

Assessing Modeling Assumptions

The Bias-Variance Trade-Off

Analysis of Variance

Sandwich Estimation

Discussion

Least Squares

Recap: Least Squares Estimation

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ denote the $n \times 1$ vector of responses and $\boldsymbol{x} = (x_1, \ldots, x_n)$ the $n \times (k+1)$ matrix of covariates where $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ik})$.

We assume a linear association between Y and x

$$\mathrm{E}[\boldsymbol{Y} | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k] = \boldsymbol{x}\boldsymbol{\beta},$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)$.

The least squares (LS) estimator minimizes the sum of squares:

$$\widehat{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta}} \quad (\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta})^{\mathsf{T}}(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta}),$$

to give

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{Y}.$$

Properties

Unbiasedness:

$$\begin{aligned} \mathrm{E}_Y[\widehat{\boldsymbol{\beta}}] &= \mathrm{E}[(\boldsymbol{x}^\intercal \boldsymbol{x})^{-1} \boldsymbol{x}^\intercal \boldsymbol{Y}] = (\boldsymbol{x}^\intercal \boldsymbol{x})^{-1} \boldsymbol{x}^\intercal \mathrm{E}[\boldsymbol{Y}] \\ &= (\boldsymbol{x}^\intercal \boldsymbol{x})^{-1} \boldsymbol{x}^\intercal \boldsymbol{x} \boldsymbol{\beta} = \boldsymbol{\beta}. \end{aligned}$$

This is true without saying anything about the variance-covariance structures of the error terms, or the distribution, of the observations.

This is in no know way, necessarily, getting at the "true" relationship between Y and x, we are simply estimating the linear association.

This may or may not be an appropriate summary measure.

Variance

Assume

$$\boldsymbol{y} = \mathsf{E}[Y|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k] + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\mathsf{T}$ with $\mathsf{E}[\boldsymbol{\epsilon}] = 0$ and $\text{var}(\epsilon) = \boldsymbol{I}_n\sigma^2$.

In linear models that contain an intercept the assumption of $\mathsf{E}[\boldsymbol{\epsilon}] = 0$ is not important, since if the mean of the "errors" can always be absorbed into the intercept, e.g., suppose the measurement device for y was off by a constant.

In this case,

$$\begin{aligned} \text{var}(\widehat{\boldsymbol{\beta}}) &= \text{var}\{(\boldsymbol{x}^\mathsf{T}\boldsymbol{x})^{-1}\boldsymbol{x}^\mathsf{T}\boldsymbol{Y}\} \\ &= (\boldsymbol{x}^\mathsf{T}\boldsymbol{x})^{-1}\boldsymbol{x}^\mathsf{T}\text{var}(\boldsymbol{Y})\boldsymbol{x}(\boldsymbol{x}^\mathsf{T}\boldsymbol{x})^{-1} = (\boldsymbol{x}^\mathsf{T}\boldsymbol{x})^{-1}\sigma^2. \end{aligned}$$

because $\text{var}(\boldsymbol{Y}) = \boldsymbol{I}_n\sigma^2$.

Distribution of Estimator

If, in addition, we assume $\boldsymbol{\epsilon} \sim \mathrm{N}_n(0, \boldsymbol{I}_n\sigma^2)$, then since $\widehat{\boldsymbol{\beta}}$ is a linear combination of normal random variables

$$\widehat{\boldsymbol{\beta}} \;\sim\; \mathrm{N}_{k+1}(\,\boldsymbol{\beta}, (\boldsymbol{x}^\intercal \boldsymbol{x})^{-1}\sigma^2\,).$$

Even if the errors are not normal, the above can be justified via a central limit theorem.

Specifically, if $\mathrm{E}[\boldsymbol{\epsilon}] = 0$ and $\mathrm{var}(\boldsymbol{\epsilon}) = \boldsymbol{I}_n\sigma^2$, then (under certain conditions on the information matrix),

$$(\boldsymbol{x}^\intercal \boldsymbol{x})^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \;\to_d\; \mathrm{N}_{k+1}(\,\boldsymbol{0}_n, \boldsymbol{I}_n\sigma^2\,).$$

Estimator of the Variance

All of the previous results depend on known σ^2.

An unbiased (quadratic) estimator is,

$$\widehat{\sigma}^2 = \frac{(\boldsymbol{Y} - \boldsymbol{x}\widehat{\beta})^{\intercal}(\boldsymbol{Y} - \boldsymbol{x}\widehat{\beta})}{n-k-1},$$

where we divide by $n - k - 1$ in order to correct for the fact that we have estimated $k + 1$ parameters within the residual sum of squares (so we're "closer" to the middle of the data than we should be. . .).

Under normality of errors,

$$\frac{(\boldsymbol{Y} - \boldsymbol{x}\widehat{\beta})^{\intercal}(\boldsymbol{Y} - \boldsymbol{x}\widehat{\beta})}{\sigma^2} = \frac{\text{RSS}}{\sigma^2} = \frac{(n-k-1)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-k-1}.$$

If the data are not normal, using this distribution can have terrible practical implications.

Confidence intervals and Tests

Under normality of errors we combine $\mathrm{RSS}/\sigma^2 \sim \chi^2_{n-k-1}$ with

$$\widehat{\beta} \sim \mathrm{N}_{k+1}(\,\beta, (\boldsymbol{x}^\intercal\boldsymbol{x})^{-1}\sigma^2\,).$$

to show that the distribution of $\widehat{\beta}$ is a multivariate t distribution with $n-k-1$ degrees of freedom.

This allows confidence intervals, confidence regions and tests to be derived.

If the data are not normally distributed but we have independent data with a common variance, then as $n \to \infty$ since $\widehat{\sigma}^2$ tends to σ^2, we will have asymptotic normality (Slutsky's theorem).

Using the t distribution gives wider intervals than the normal and so is conservative.

Confidence intervals and Tests

Particular results: For $j = 0, 1, \ldots, k$,

$$\frac{\widehat{\beta}_j - \beta_j}{S_j^{1/2}\widehat{\sigma}} \sim t_{n-k-1},$$

where the latter denotes the univariate t-distribution with $n - k - 1$ degrees of freedom, location β_j and scale $S_j\widehat{\sigma}^2$ and S_j is element (j, j) of $(\boldsymbol{x}^\top \boldsymbol{x})^{-1}$.

Confidence intervals and tests of $H_0 : \beta_j = c$ follow.

The latter is a partial t-test in that it is a test with $1, x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$ in the model – this is different from fitting the model

$$\mathrm{E}[Y|\boldsymbol{x}] = \beta_0 + \beta_j x_j,$$

and testing $H_0 : \beta_j = c$, unless x_j is orthogonal to other covariates.

Message: care must be taken when examining t-statistics in a multiple regression.

Confidence intervals and Tests

The test with $c = 0$ is equivalent to the partial F-statistic,

$$F = \frac{\text{FSS}(\beta_j|\beta_0, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k)/1}{\text{RSS}/(n-k-1)},$$

where

$$\text{FSS}(\beta_j|\beta_0, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k) =$$

$$\text{RSS}(\beta_0, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k) - \text{RSS}(\beta),$$

with $F = t^2$, and under H_0, $F \sim F_{1,n-k-1}$.

Consider the model,

$$\text{E}[Y|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k] = \beta_0 + \beta_1 \boldsymbol{x}_1 + \cdots + \beta_p \boldsymbol{x}_k,$$

and the hypothesis

$$H_0 : \beta_{q+1} = \cdots = \beta_k = 0,$$

for $1 \leqslant q \leqslant k$.

Confidence intervals and Tests

Under H_0, we have the partial F-statistic,

$$F = \frac{\text{FSS}(\beta_{q+1}, \ldots, \beta_k | \beta_0, \beta_1, \ldots, \beta_q)/(k-q)}{\text{RSS}/(n-k-1)} \sim F_{k-q,n-k-1},$$

partial because $\beta_0, \beta_1, \ldots, \beta_q$ are in the model.

Note that

$$\text{FSS}(\beta_{q+1}, \ldots, \beta_k | \beta_0, \beta_1, \ldots, \beta_q) \neq \text{FSS}(\beta_{q+1}, \ldots, \beta_k),$$

unless $(x_1, \ldots, x_q)$ is orthogonal to $(x_{q+1}, \ldots, x_k)$.

The table below lays out the calculations for the above partial F-test of

$$H_0 : \beta_{q+1} = \cdots = \beta_k = 0$$

in the form of an Analysis of Variance (ANOVA) table.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-statistic
x_1	$\text{FSS}(\beta_1)$	q		
x_2	$\text{FSS}(\beta_2 \| \beta_1)$	$k-q$	$\frac{\text{FSS}(\beta_2 \| \beta_1)}{k-q}$	$\frac{\text{FSS}(\beta_2 \| \beta_1)/(k-q)}{\text{RSS}/(n-k-1)}$
Error	RSS	$n-k-1$	$\frac{\text{RSS}}{n-k-1}$	
Total	CTSS	$n-1$		

In the table:

$$\boldsymbol{x}_1 = (1, x_1, \ldots, x_q), \qquad \boldsymbol{x}_2 = (x_{q+1}, \ldots, x_k),$$

$$\boldsymbol{\beta}_1 = (\beta_0, \ldots, \beta_q), \qquad \boldsymbol{\beta}_2 = (\beta_{q+1}, \ldots, \beta_k)$$

and

$$\text{CTSS} = (Y - \bar{Y})^{\intercal}(Y - \bar{Y})$$

is the corrected total sum of squares.

For a $100(1-\alpha)\%$ confidence region we can use the elliptically-shaped contours in k-dimensional space determined by

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\intercal}(\boldsymbol{x}^{\intercal}\boldsymbol{x})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = ks^2 F_{k,n-k-1}(1-\alpha).$$

Prediction

For inference concerning the expected response at a covariate $\boldsymbol{x}_0$, we define

$$\theta = \boldsymbol{x}_0\boldsymbol{\beta}.$$

Then

$$\widehat{\theta} = \boldsymbol{x}_0\widehat{\boldsymbol{\beta}}$$

and

$$\widehat{\theta} \sim \mathrm{N}\left(\theta, \boldsymbol{x}_0(\boldsymbol{x}^\top\boldsymbol{x})^{-1}\boldsymbol{x}_0^\top\sigma^2\right).$$

For prediction of an observed response at $\boldsymbol{x}_0$ we define

$$Y_0 = \boldsymbol{x}_0\boldsymbol{\beta} + \epsilon_0$$

where $\epsilon_0 \sim \mathrm{N}(0, \sigma^2)$, with estimator

$$\widehat{Y}_0 = \boldsymbol{x}_0\widehat{\boldsymbol{\beta}} + \epsilon_0.$$

Prediction

It then follows that,

$$\widehat{Y}_0 \sim N\left(y_0, [1 + \boldsymbol{x}_0(\boldsymbol{x}^\intercal\boldsymbol{x})^{-1}\boldsymbol{x}_0^\intercal]\sigma^2\right).$$

Intervals constructed in this way can be awful if the data are not normally distributed.

Alternative approaches are to:

- make another distributional assumption (a parametric approach), or
- bootstrap the residuals (a semi-parametric approach).

Random component

Consider the model,

$$y_i = \boldsymbol{x}_i \boldsymbol{\beta} + \epsilon_i.$$

Weaker assumptions about ϵ_i lead to weaker conclusions:

The Gauss-Markov Theorem tells us that if errors are uncorrelated then the least squares estimators have minimum variance in the class of linear unbiased estimators.

Least squares has smallest variance amongst all unbiased estimators, if $\epsilon_i \sim_{iid} \mathrm{N}(0, \sigma^2)$.

Intuitively: if we knew the distribution of the errors and we used this in our estimator, then we should do better than if we don't use this information.

The big questions:

- if we get the distribution wrong (model misspecification), how well does our estimator perform?
- What is the loss in efficiency if we don't pick the "right" model?

Generalized Least Squares (GLS)

If $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{var}(\boldsymbol{\epsilon}) = \boldsymbol{V}\sigma^2$, for known $\boldsymbol{V}$, then it is more appropriate to minimize

$$(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})^\intercal \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}) = \boldsymbol{y}^\intercal \boldsymbol{V}^{-1}\boldsymbol{y} - 2\boldsymbol{\beta}^\intercal \boldsymbol{x}^\intercal \boldsymbol{V}^{-1}\boldsymbol{y} + \boldsymbol{\beta}^\intercal \boldsymbol{x}^\intercal \boldsymbol{V}^{-1}\boldsymbol{x}\boldsymbol{\beta},$$

to give estimating function

$$\boldsymbol{G} = \boldsymbol{x}^\intercal \boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta}),$$

to obtain the GLS estimator

$$\widehat{\boldsymbol{\beta}}_G = (\boldsymbol{x}^\intercal \boldsymbol{V}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}^\intercal \boldsymbol{V}^{-1}\boldsymbol{Y},$$

with

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}_G) = (\boldsymbol{x}^\intercal \boldsymbol{V}^{-1}\boldsymbol{x})^{-1}\sigma^2.$$

Under the correct variance-covariance structure this is as good as it gets for unbiased linear estimators.

Parameter Interpretation and Parameterization

Parameterization

The parameterization of a model can be based on:

- Mathematical considerations.
- Parameter interpretation.
- Computational Convenience.

Example: A beta distribution can be parameterized via:

- The “powers” a and b to give density proportional to $x^{a-1}(1-x)^{b-1}$.
- The mean $m = a/(a+b)$ and scale $s = a+b$.
- Parameters on the whole real line: $\log a$ and $\log b$.

Parameter interpretation

We initially consider the situation in which we have a single covariate x in which case a simple model is

$$\mathrm{E}[Y|x] = \beta_0 + \beta_1 x.$$

The model

$$\mathrm{E}[Y|x] = \beta_0,$$

says that the expected response does not vary with x, which is a strong statement.

The model,

$$\mathrm{E}[Y] = \beta_0$$

is very different and is simply saying that there is an average response in the population – no assumptions here – this response is given by

$$\mathrm{E}[Y] = \mathrm{E}_x\{\mathrm{E}[Y|x]\}.$$

Message: care should be taken in the notation we use.

Parameter interpretation

The “reciting” interpretation of the parameters in

$$\mathrm{E}[Y|x] = \beta_0 + \beta_1 x,$$

is that β_1 represents the increase in the expected response for a unit change in x, and β_0 is the expected response at $x = 0$.

The latter expectation may make little sense however (e.g., response blood pressure, predictor height).

If we reparameterize the model as

$$\mathrm{E}[Y|x] = \beta_0^* + \beta_1(x - x^*),$$

then $\beta_0^\star$ is the expected response at $x = x^\star$ where $x^\star$ may be taken as a value that is more interesting/meaningful scientifically.

Taking $x^* = \bar{x}$ leads to computational/analytical convenience (e.g., prediction).

Key Point: The models are the same!

Parameter interpretation

In the linear model,

$$\mathrm{E}[Y|x] = \beta_0 + \beta_1 x,$$

the interpretation of β_1 here is the expected change in the response for a unit change in x, in the population that we sampled – it is an association.

If, we were to select two groups of individuals from the population, one with $x + 1$, and another with x, then β_1 would represent the difference in the expected response between these groups.

This is very different to stating that β_1 is the expected change in the expected response if we were to increase x by one unit for an individual (via an intervention).

The latter is a causal interpretation and is only valid under very strict conditions.

Parameter interpretation

To see the difference between associations and causal statements, suppose the true relationship is

$$\mathrm{E}[Y|x, Z] = \beta_0 + \beta_1 x + \beta_2 Z,$$

but Z is unobserved.

Here, β_1 is the expected change in the expected response for a unit increase in x with Z held constant.

Define a and b through,

$$\mathrm{E}[Z|x] = a + bx,$$

so that

$$\mathrm{E}[Y|x] = \mathrm{E}_{Z|x}\{\mathrm{E}[Y|x, Z]\} = \beta_0^\star + \beta_1^\star x,$$

where $\beta_0^\star = \beta_0 + \beta_2 a$ and $\beta_1^\star = \beta_1 + \beta_2 b$.

If we were to select an individual and change $x \rightarrow x + 1$ but were to leave Z unchanged, then β_1 is the change in that individual's expected response, and not $\beta_1^\star$.

Parameter interpretation

In an observational study, nature's allocation of x is not carried out at random – whether $x = 0/1$ (say) depends on the values of other variables (due, say, to common causes) and so x and Z are not independent.

In observational studies is that we are modeling associations and not causal relationships, and so great care is required in parameter interpretation.

In a designed experiment in which x is randomly assigned to each individual, via randomization for example, then a causal interpretation is valid (in expectation and assuming the experiment was successfully completed so there are no non-compliers for example).

Here $x \perp Z$ by construction.

Parameterization

Suppose we wish to examine the association between a response Y and sex x_1.

Obvious formulation of the model:

$$\mathrm{E}[Y|x] = \begin{cases} \beta_0 + \beta_1 & \text{if female } x_1 = 0, \\ \beta_0 + \beta_2 & \text{if male } x_1 = 1. \end{cases}$$

The parameters are not identifiable, however.

This non-identifiability phenomenon is known as (intrinsic) aliasing.

A solution is to place a constraint on the parameters.

(Extrinsic aliasing – anomaly of the data that makes the columns of x linearly dependent.)

Parameterization

In the popular sum-to-zero parameterization we impose the constraint $\beta_1 + \beta_2 = 0$ which gives the model

$$\mathrm{E}[Y|\boldsymbol{x}] = \begin{cases} \beta_0' - \beta_1' & \text{if female } x_1 = -1, \\ \beta_0' + \beta_1' & \text{if male } x_1 = 1. \end{cases}$$

In this case we have $\mathrm{E}[Y|\boldsymbol{x}] = \boldsymbol{x}\boldsymbol{\beta}'$ with $\boldsymbol{x} = [1 \quad -1]$ if female, and $\boldsymbol{x} = [1 \;\; 1]$ if male.

In this model β_0' is the average response, and $2\beta_1'$ is the expected response for males minus the expected response for females.

Parameterization

An alternative, the corner-point constraint, is to assign $\beta_1 = 0$ and assume a model of the form

$$\mathrm{E}[Y|\boldsymbol{x}] = \begin{cases} \beta_0^\star & \text{if female } x_1 = 0, \\ \beta_0^\star + \beta_1^\star & \text{if male } x_1 = 1. \end{cases}$$

In this case we have $\mathrm{E}[Y|\boldsymbol{x}] = \boldsymbol{x}\beta^\star$ where $\boldsymbol{x} = [1 \;\; 0]$ if female, and $\boldsymbol{x} = [1 \;\; 1]$ if male.

In this model $\beta_0^\star$ is the expected response for a female, and $\beta_1^\star$ is the change in the expected response for males, as measured against females.

PARAMETERIZATION

A final model is,

$$\mathrm{E}[Y|\boldsymbol{x}] = \begin{cases} \beta_0^\dagger & \text{if female } x_1 = 0, \\ \beta_1^\dagger & \text{if male } x_1 = 1. \end{cases}$$

In this case we have $\mathrm{E}[Y_i|x] = x\beta^\dagger$ where $\boldsymbol{x} = [1 \;\; 0]$ if female, and $\boldsymbol{x} = [0 \;\; 1]$ if male.

In this model $\beta_0^\star$ is the expected response for a female, and $\beta_1^\star$ is the expected response for a male – yay, easy!!! But it doesn't generalize to many factors.

In general, the benefits of the above alternative parameterizations should also be considered in the light of the possibility of their extension to the case of more than one factor – sum-to-zero and corner point are advantageous in this respect.

Note that inference for the each of the formulations is identical, the only thing that changes is the interpretation of the parameters.

INTERACTIONS

We discuss interactions with respect to the models:

$$\begin{aligned}
\text{Null model}: \quad \mathrm{E}[Y|x_1, x_2] &= \beta_0, \\
x_1 \text{ only}: \quad \mathrm{E}[Y|x_1, x_2] &= \beta_0 + \beta_1 x_1, \\
x_2 \text{ only}: \quad \mathrm{E}[Y|x_1, x_2] &= \beta_0 + \beta_2 x_2, \\
x_1 \text{ and } x_2 \text{ only}: \quad \mathrm{E}[Y|x_1, x_2] &= \beta_0 + \beta_1 x_1 + \beta_2 x_2, \\
\text{Interaction}: \quad \mathrm{E}[Y|x_1, x_2] &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2,
\end{aligned}$$

where $x_1 = 0, 1$ and $x_2 = 0, 1$.

Figure 1 will be used to interpret interactions.

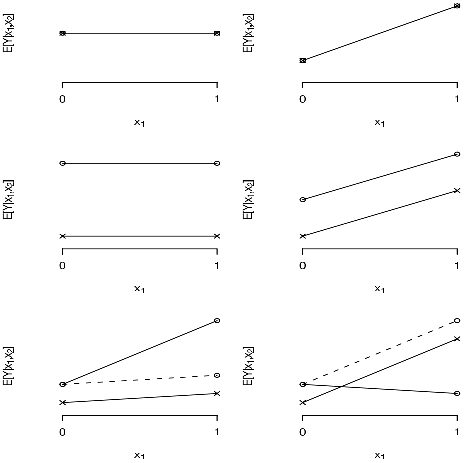

FIGURE 1: Interpretation of interactions: (a) null model, (b) x_1 main effect only, (c) x_2 main effect only, (d) x_1 and x_2 main effects only, (e) interaction, (f) interaction.

Assessing Modeling Assumptions

Assessing Modeling Assumptions

Diagnostics may be examined to access the adequacy of the assumptions underlying the analysis performed.

So far we have seen that both frequentist and Bayesian approaches to inference rely on modeling assumptions for valid inference – assumptions are for both the deterministic and stochastic components of the model.

Diagnostics should not be viewed as a way of avoiding careful initial thought about the model that is used since inference requires the model to not have been chosen on the basis of the current data set.

Assessing Modeling Assumptions

In a frequentist analysis, the operating characteristics of estimators (e.g. confidence intervals) and hypothesis tests (type I and II error rates) are based upon repeat sampling from the model.

Hence, if we follow a particular procedure (for example, examination of residual plots) to reach a certain model, this whole procedure forms the sampling distribution.

In a Bayesian analysis, the model chosen is itself a random variable and hence all inference must average across all possible models that may be examined.

Residuals

Consider the model $Y_i = \boldsymbol{x}_i\beta + \epsilon_i$ with $\dim(\beta) = k + 1$ and suppose we assume $\text{var}(\boldsymbol{\epsilon}) = \boldsymbol{I}_n\sigma^2$.

For $i = 1, \ldots, n$, the observed errors are

$$e_i = Y_i - \widehat{Y}_i, \tag{1}$$

while the true errors are

$$\epsilon_i = Y_i - \text{E}[Y_i|\boldsymbol{x}_i].$$

In residual analysis we examine the observed residuals for discrepancies from the assumed model.

Residuals

Define residuals as

$$\boldsymbol{e} = (e_1, \ldots, e_n)^\top = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{v})\boldsymbol{Y}, \tag{2}$$

where $\boldsymbol{v} = \boldsymbol{x}(\boldsymbol{x}^\top \boldsymbol{x})^{-1}\boldsymbol{x}^\top$.
$\boldsymbol{v}$ is sometimes known as the hat (or projection) matrix and is

- symmetric ($\boldsymbol{v}^\top = \boldsymbol{v}$) and
- idempotent ($\boldsymbol{v}\boldsymbol{v}^\top = \boldsymbol{v}$).

The matrix $\boldsymbol{v}$ is the linear transformation that orthogonally projects any n-vector onto the space spanned by the columns of $\boldsymbol{x}$, i.e., $\widehat{y} = \boldsymbol{v}\boldsymbol{y}$.

Residuals

We want to derive the relationship between e and ϵ so we can use the former to assess whether the assumptions that we have made on the latter are appropriate, i.e., uncorrelated, constant variance and normality (if we want to carry out prediction of an observable).

If we substitute

$$Y = \boldsymbol{x}\beta + \epsilon$$

into (2), we obtain

$$\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{v})\epsilon, \tag{3}$$

or

$$e_i = \epsilon_i - \sum_{j=1}^{n} v_{ij}\epsilon_j, \tag{4}$$

showing that the estimated residuals differ from the true residuals.

The Hat Matrix: Example

For the simple linear model $\mathrm{E}[Y_i|\boldsymbol{x}_i] = \beta_0 + \beta_1 x_i$ we have

$$v_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}.$$

We may write

$$\widehat{Y}_i = v_{ii} Y_i + \sum_{j=1, j\neq i}^{n} v_{ij} Y_j,$$

so if v_{ii} is large relative to the other elements in the i-th row of v, then the i-th fitted value will be largely determined by Y_i.

In this case the i-th observation is said to have large leverage.

Note that the leverage depends on the design matrix (i.e., $\boldsymbol{x}$) only, which distinguishes it from observations that have strong influence on the fitted values.

Since $\mathrm{var}(\widehat{Y}_i) = v_{ii}\sigma^2$, fitted values extreme in the x-space have high variance, while the corresponding residuals have low variance.

Residuals

The $\boldsymbol{e}$-vector is a function of the vector of random variables ϵ, and we have

$$\mathrm{E}[\boldsymbol{e}] = 0 \quad \text{and} \quad \mathrm{var}(\boldsymbol{e}) = (\boldsymbol{I} - \boldsymbol{v})\sigma^2.$$

In particular:

$$\mathrm{var}(e_i) = (1 - v_{ii})\sigma^2.$$

Based on these results we may define standardized residuals to be equal to:

$$e_i^* = \frac{Y_i - \widehat{Y}_i}{\widehat{\sigma}(1 - v_{ii})^{1/2}}, \qquad (5)$$

for $i = 1, \ldots, n$. These residuals are such that $\mathrm{E}[e_i^*] = 0$ and $\mathrm{var}(e_i^*) = 1$, but they are not independent since they are based on $n - (p + 1)$ independent quantities.

The constraints are given by

$$(\boldsymbol{Y} - \boldsymbol{x}\widehat{\beta})^\top \boldsymbol{x} = 0.$$

Often the $(1 - v_{ii})^{1/2}$ term in the denominator of (5) is ignored.

Residuals

Consider fitting the model without observation i to get a prediction $\widehat{y}_{(i)}$ of y_i.

Then predicted or case-deletion residuals are given by

$$e_i^\dagger = Y_i - \widehat{Y}_{(i)} = Y_i - \boldsymbol{x}_i \widehat{\boldsymbol{\beta}}_{(i)},$$

where $\widehat{\boldsymbol{\beta}}_{(i)}$ is the OLS estimator obtained from a reduced data set with case i removed.

It can be shown that

$$e_i^\dagger = e_i/(1 - v_{ii}),$$

these residuals are useful for outlier detection. Sometimes two or more points may obscure outliers (masking), so the above can be extended by deleting more than one case at a time.

Jack-knifed residuals

If one error is very large then the variance estimate $\widehat{\sigma}^2$ will be too large and will deflate all of the standardized residuals.

Jack-knifed residuals are given by

$$e_i' = \frac{Y_i - \widehat{Y}_{(i)}}{\sqrt{\mathrm{var}(Y_i - \widehat{Y}_{(i)})}},$$

which are standardized residuals with $\widehat{\sigma}$ replaced by $\widehat{\sigma}_{(i)}$.

Fortunately we don't need to refit the model each time an observation is deleted since

$$e_i' = e_i^* / \left(\frac{n-p-1-e_i^{*2}}{n-p-2} \right).$$

Scatterplots

Recall assumptions:

(A1) The error terms have constant variance.

(A2) The error terms are uncorrelated.

(A3) The estimator is normally distributed.

Deviations from the horizontal axis are easiest to detect.

Local smoothers are very useful for detecting departures from that expected.

Residuals may be plotted against fitted values $\widehat{Y}_i$ for assessment of (A1).

The same assessment may also be made by plotting squared residuals, e.g. e_i^2 versus $\widehat{Y}_i$.

Scatterplots

Plotting residuals against covariates can:

- Indicate nonlinear relationships, need for other functional forms or transformations of covariates to linearity (beware dredging), or robustify the regression (though interpretation may be more difficult).
- Indicate the possibility of interactions (use different plotting symbols for sub-populations defined by factors)
- Outliers may also be indicated – is the regression being defined by a few points, and do we believe them? Can also report conclusions only for stable regions.

To assess dependence between error terms (A3):

- Plot e_i versus spatial or temporal indicators.
- Dependence may also be detected using scatter plots of lagged residuals, e.g. plotting e_i versus e_{i-1} for $i = 2, \ldots, n$. Independent residuals should produce a plot with a random scatter of points.
- Variogram of residuals in spatial analysis.

Normal Plots

We don’t need normality of error terms but the mean is more likely to be a useful summary of differences when the data are approximately normal with constant variance.

A normal plot takes the ordered residuals and plots them versus the expected residuals (may be used for other distributions also).

Points on a straight line can indicate normal errors while systematic departures can indicate

- light-tails
- heavy-tails, and/or
- skewness.

Normal Plots

Care must be taken in interpretation since (4) may exhibit supernormality since

$$e_i = \epsilon_i - \sum_{j=1}^{n} v_{ij}\epsilon_j,$$

and so even if ϵ_i is not normal, $\sum_{j=1}^{n} v_{ij}\epsilon_j$ may tend towards normality (and dominate the first term).

This may be examined by simulating data from a standard normal, with the same design matrix (so that v is the same as in the observed data) and forming confidence envelopes.

Simulation is good for seeing what we would expect to see if the model was correct.

Residuals: Final Comments

Problem: If assumptions are found wanting and we change the model, what are the frequentist properties in terms of bias, the coverage of intervals, and the α level of tests?

Moral of story: try and think as much as possible about model choice before the data are analyzed.

Also always report the exact procedure followed, so that inferential summaries can be interpreted.

Variance Stabilization

The linear model assumes uncorrelated additive errors with constant variance and linearity of the response-covariate relationship.

If these assumptions are inadequate in a particular application we may consider transformation of the response and/or the covariates (again beware dredging).

An alternative is to assume a model with different assumptions and model on the original scale.

We first consider how the variance may be stabilized.

Variance Stabilization

Suppose we know that $\mathrm{E}[Y] = \mu_Y$ and $\mathrm{var}(Y) = h(\mu_Y)\sigma_Y^2$.

Then consider the Taylor series for a differentiable function $g(\cdot)$:

$$Z = g(Y) \approx g(\mu_Y) + (Y - \mu_Y)\frac{\partial g}{\partial y}\Big|_{\mu_Y},$$

to give

$$\mathrm{E}[Z] \approx g(\mu_Y)\ , \quad \mathrm{var}(Z) \approx h(\mu)\sigma_Y^2 \left(\frac{\partial g}{\partial y}\Big|_{\mu_Y}\right)^2 .$$

The variance-stabilizing transformation $g(y)$ is such that $\partial g/\partial y = h(y)^{-1/2}$ or

$$g(y) = \int h(y)^{-1/2}\mathrm{d}y.$$

Variance Stabilization

Transformations may also robustify the analysis to outliers (e.g. the log transform).

Examples:

(i) $h = \mu$ then $g(y) = \sqrt{y}$, (historically used for Poisson data),

(ii) $h = \mu^2$ then $g(y) = \log y$, (data with a constant coefficient of variation, alternative lognormal distribution).

(iii) $h = \mu(1 - \mu)$ then $g(y) = \arcsin\{(y)^{1/2}\}$ (historically used for binomial data).

(ii) can be useful, while (iii) is a bit ridiculous... I have used (i), since it works for $y = 0$, unlike (ii).

Transformations

Suppose we fit the model

$$\log Y = \beta_0 + \beta_1 x + \epsilon, \tag{6}$$

or equivalently

$$Y = \exp(\beta_0 + \beta_1 x + \epsilon) = \exp(\beta_0 + \beta_1 x)\delta,$$

where $\delta = \exp(\epsilon)$.

The expectation of Y depends on the expectation of ϵ.

For interpretation:

$$\frac{E[Y|x-1]}{E[Y|x-1]} = \frac{\exp(\beta_0 + \beta_1 x)}{\exp(\beta_0 + \beta_1 (x-1))} \times \frac{E[\delta(x)}{E[\delta(x)} = \exp(\beta_1) \times \frac{E[\delta(x)]}{E[\delta(x)]}$$

Transformations

For example, if $\epsilon \sim \mathrm{N}(0, \sigma^2)$ then, using the fact that δ is lognormal,

$$\mathrm{E}[Y|x] = \exp(\beta_0 + \beta_1 x + \sigma^2/2).$$

Hence,

$$\frac{\mathrm{E}[Y|x-1]}{\mathrm{E}[Y|x-1]} = \exp(\beta_1)$$

So long as $\mathrm{E}[\delta|\boldsymbol{x}]$ does not depend on $\boldsymbol{x}$ then we can interpret $\exp(\beta_1)$ as the ratio of expected responses between sub-populations with covariate values x and $x-1$.
It may be useful to think about

$$\frac{\mathrm{median}(Y|x-1)}{\mathrm{median}(Y|x-1)}$$

So if $\delta(x)$ follows a symmetric error distribution that doesn't depend on x, the ratio is $\exp(\beta_1)$.

Transformations

This model with the assumption of normal errors is useful if the standard deviation on the original scale is proportional to the mean since, evaluating the variance of a lognormal distribution

$$\text{var}(Y|\boldsymbol{x}) = \text{E}[Y|\boldsymbol{x}]^2\{\exp(\sigma^2) - 1\}.$$

If σ^2 is small then $\exp(\sigma^2) \approx 1 + \sigma^2$ and

$$\text{var}(Y|\boldsymbol{x}) = \text{E}[Y|\boldsymbol{x}]^2\sigma^2,$$

showing that for this model we have a constant coefficient of variation (CV) (which is defined to be the standard deviation divided by the mean).

Constant CV models are popular in many disciplines including chemistry and phamacokinetics.

Transformations

A model that at first sight looks very similar has

$$\mathsf{E}[Y|\boldsymbol{x}] = \exp(\beta_0 + \beta_1 x),$$

with $Y = \mathsf{E}[Y|\boldsymbol{x}] + \epsilon$.

The interpretation of $\exp(\beta_1)$ is the same as in the previous case but here we have additive errors, whereas in the previous case they were multiplicative.

For the above model we may question whether additive errors are reasonable given that the mean function is always positive.

This model is also non-linear in the parameters, whereas (6) is linear.

Transformations

Rather than transform one may consider a broader class of models, GLMs provide one class.

A huge advantage of a GLM is that the data may be modeled on their natural scale – for example, for count data we can a count model, such as binomial, Poisson or overdispersed versiosn!

The Bias-Variance Trade-Off

Bias-Variance Trade-Off

Suppose the true model is

$$Y = \boldsymbol{x}\boldsymbol{\beta} + \epsilon,$$

where $\boldsymbol{x}$ is $n \times (k+1)$, $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{var}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$.

The LS estimator

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{x}^\top \boldsymbol{x})^{-1} \boldsymbol{x}^\top \boldsymbol{Y},$$

is such that

$$\mathrm{E}[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta} \quad \text{and} \quad \mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{x}^\top \boldsymbol{x})^{-1} \sigma^2,$$

where we assume $\boldsymbol{x}^\top \boldsymbol{x}$ is of full rank.

Since $\boldsymbol{x}^\top \boldsymbol{x}$ is symmetric we can find an upper-triangular matrix $\boldsymbol{U}$ such that $(\boldsymbol{x}^\top \boldsymbol{x})^{-1} = \boldsymbol{U}\boldsymbol{U}^\top$ which leads to

$$\mathrm{var}(\widehat{\boldsymbol{\beta}})_{jj} = \sigma^2 \sum_{l=1}^{k+1} U_{jl}^2,$$

with $U_{jl} = 0$ if $j > l$.

Bias-Variance Trade-Off

Suppose we fit the model

$$\mathsf{E}[Y|\boldsymbol{x}_A] = \boldsymbol{x}_A\boldsymbol{\beta}_A^\star,$$

where $\boldsymbol{x} = (\boldsymbol{x}_A, \boldsymbol{x}_B)$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_A^\star, \boldsymbol{\beta}_B^\star)^\intercal$ and $\boldsymbol{x}_A$ is $n \times (q+1)$ with $q < k$.

Then

$$\widehat{\boldsymbol{\beta}}_A^\star = (\boldsymbol{x}_A^\intercal\boldsymbol{x}_A)^{-1}\boldsymbol{x}_A^\intercal Y,$$

and

$$\begin{aligned} \mathsf{E}[\widehat{\boldsymbol{\beta}}_A^\star] &= (\boldsymbol{x}_A^\intercal\boldsymbol{x}_A)^{-1}\boldsymbol{x}_A^\intercal(\boldsymbol{x}_A\boldsymbol{\beta}_A + \boldsymbol{x}_B\boldsymbol{\beta}_B) \\ &= \boldsymbol{\beta}_A + (\boldsymbol{x}_A^\intercal\boldsymbol{x}_A)^{-1}\boldsymbol{x}_A^\intercal\boldsymbol{x}_B\boldsymbol{\beta}_B, \end{aligned}$$

the second term is the bias arising from omission of the last $k - q$ covariates. This bias is zero if $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ are orthogonal (or $\boldsymbol{\beta}_B = 0$).

So for bias to result we need $\boldsymbol{x}_B$ to be associated with both Y and $\boldsymbol{x}_A$.

Bias-Variance Trade-Off

We write

$$(\boldsymbol{x}_A^\top \boldsymbol{x}_A)^{-1} = \boldsymbol{U}_A \boldsymbol{U}_A^\top$$

where $\boldsymbol{U}_A$ is upper-triangular and consists of the first $q+1$ rows and columns of $\boldsymbol{U}$.

We then have

$$\operatorname{var}(\widehat{\beta}_A^\star)_{jj} = \sigma^2 \sum_{l=1}^{q+1} U_{jl}^2 = \operatorname{var}(\widehat{\beta}_A^\star)_{jj} \leqslant \operatorname{var}(\widehat{\beta}_A)_{jj},$$

$j = 0, 1, \ldots, q$, with equality iff $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ are orthogonal.

From this it would appear that adding covariates will never increase precision (since there are competing explanations for the data).

It is not as straightforward as this, however, since we have assumed that σ^2 is known.

Bias-Variance Trade-Off

The other consideration is explaining variability in the data – if a covariate is strongly associated with the response then there will be a large reduction in the variance of the 'error' if we omit that covariate, as the following shows.

We have

$$\text{var}(Y|\boldsymbol{x}_A, \boldsymbol{x}_B) = \sigma^2 I$$

with both covariates and

$$\text{var}(Y|\boldsymbol{x}_A) = \sigma^2 I + \text{var}(\boldsymbol{x}_B \boldsymbol{\beta}_B | \boldsymbol{x}_A).$$

In practice we will see a decrease in $\widehat{\sigma}^2$ if we include $\boldsymbol{x}_B$.

The overall effect in terms of precision of estimation of $\widehat{\boldsymbol{\beta}}_A$ will depend on $\boldsymbol{\beta}_B$ and the degree of orthogonality.

Bias-Variance Trade-Off

Consider the model

$$Y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \beta_2(z_i - \bar{z}) + \epsilon_i,$$

compared to

$$Y_i = \beta_0^\star + \beta_1^\star(x_i - \bar{x}) + \epsilon_i.$$

Then

$$\mathrm{E}[\widehat{\beta}_0^\star] = \beta_0,$$

since the n vector of 1's is orthogonal to $\boldsymbol{x}$ and

$$\mathrm{E}[\widehat{\beta}_1^\star] = \beta_1 + \beta_2 \times S_{xz}/S_{xx} = \beta_1 + \beta_2 \times \rho_{xz}(S_{zz}/S_{xx})^{1/2} \tag{7}$$

where $S_{xx} = \sum_{\underline{i}}(x_i - \bar{x})^2$, $S_{xz} = \sum_i(x_i - \bar{x})(Z_i - \bar{Z})$, $S_{zz} = \sum_i(Z_i - \bar{Z})^2$, and $\rho_{xz} = S_{xz}/(S_{xx}S_{zz})^{1/2}$.

Equation (7) may be used to assess sensitivity to unmeasured confounding.

Bias-Variance Trade-Off

We have,

$$(\boldsymbol{x}^\top \boldsymbol{x})^{-1} = \begin{bmatrix} 1/n & 0 & 0 \\ 0 & S_{zz}/D & -S_{xz}/D \\ 0 & -S_{xz}/D & S_{xx}/D \end{bmatrix},$$

where $D = S_{xx}S_{zz} - S_{xz}^2$, giving

$$\text{var}(\widehat{\beta}_1) = \frac{\sigma^2}{S_{xx} - S_{xz}^2/S_{zz}} \geqslant \frac{\sigma^2}{S_{xx}} = \text{var}(\widehat{\beta}_1^*),$$

with equality iff $S_{xz} = 0$, i.e., x and Z are orthogonal.

Bias-Variance Trade-Off: Summary

For a linear model, the inclusion of an additional variable that is truly related to the response will reduce the bias (or at least not increase the bias), but may increase the variance of the estimator of the association with the covariate of interest.

When additional covariates are included, there are competing factors that must be considered.

The unexplained variation in the data (via $\widehat{\sigma}^2$) may be reduced but the uncertainty in which of the covariates to assign the variation in the response to is increased.

This uncertainty is reduced the greater the orthogonality between the exposure and the confounder.

Example of Bias-Variance Trade-off: prostate cancer study

	Full model			Reduced model		
Variable	Estimate	St. Er.	T score	Estimate	St. Er.	T score
lcavol	0.5870	0.0879	6.6768	0.5516	0.0747	7.3879
lweight	0.4545	0.1700	2.6731	0.5085	0.1502	3.864
age	-0.0196	0.0112	-1.7576	–	–	–
lbph	0.1071	0.0584	1.8316	–	–	–
svi	0.7662	0.2443	3.1360	0.6662	0.2098	3.17562
lcp	-0.1055	0.0910	-1.1589	–	–	–
gleason	0.0451	0.1575	0.2866	–	–	–
pgg45	0.0045	0.0044	1.0236	–	–	–
σ	0.7804	–	–	0.7168	–	–

Table 1: Parameter estimates from full model and a reduced model for the prostate cancer data.

Notice the change in the estimates in the reduced model (so potentially biased), but the reduction in standard errors.

The variance estimate $\widehat{\sigma}$, is smaller in the reduced model.

Bayesian analysis

We now discuss Bayesian analyses of these data.

With the improper prior,

$$\pi(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2},$$

inference is identical with the frequentist approach so that the estimates and standard errors in Table 1 are also posterior means and posterior standard deviations.

Turning now to an informative prior distribution on $\boldsymbol{\beta}$,

$$\pi(\boldsymbol{\beta}, \sigma^2) \propto \prod_{j=0}^{8} \pi(\beta_j) \times \sigma^{-2},$$

with $\pi(\beta_0) \propto 1$ and $\pi(\beta_j) \sim \mathrm{N}(0, v_j)$, with v_j known.

Bayesian analysis

The standard deviations for the prior, $\sqrt{v_j}$, were chosen in the following way.

For the prostate data we believe that it is unlikely that any one covariate, over it's range (which we denote $x_{\max}$), will change the log(PSA) by more than $y_{\max} = 2$ units (on the log scale).

The way we include this information in the prior is by assuming that the $\beta_j = 0 + 2\sqrt{v_j}$ point of the prior corresponds to the maximum value of β_j that we believe a priori is plausible.

Formally we have

$$2\sqrt{v_j} = \frac{y_{\max}}{x_{\max}} \quad \text{to give} \quad v_j = \frac{y_{\max}^2}{2^2 x_{\max}^2}.$$

Bayesian analysis

	Prior model				Shrinkage model			
Variable	Median	St. Dev.	95% Interval		Median	St. Dev.	95% Interval	
lcavol	0	0.197	-0.387	0.387	0.4946	0.0806	0.3335	0.6495
lweight	0	0.273	-0.536	0.536	0.3472	0.1452	0.0592	0.6267
age	0	0.027	-0.053	0.053	-0.0133	0.0103	-0.0334	0.0072
lbph	0	0.275	-0.539	0.539	0.1116	0.0564	0.0006	0.2220
svi	0	1.020	-2.000	2.000	0.7660	0.2359	0.3039	1.2260
lcp	0	0.238	-0.466	0.466	-0.0386	0.0837	-0.2009	0.1278
gleason	0	0.340	-0.667	0.667	0.0577	0.1397	-0.0045	0.0110
pgg45	0	0.010	-0.020	0.020	0.0033	0.0039	-0.0045	0.0110
$\widehat{\sigma}$	–	–	–	–	0.7137	0.0554	0.6193	0.8370

TABLE 2: Prior and posterior summaries for the Bayesian shrinkage model for the prostate cancer data.

We see that the posterior standard deviations are all smaller than the prior standard deviations, as they must be in the normal-normal linear model.

If we compare the full model estimates with Table 1 we have reduced uncertainty in the Bayesian analysis (compare the standard errors with the posterior standard deviations), and some shrinkage in the estimates: one way of addressing the bias-variance issue.

Analysis of Variance

Analysis of Variance

The name ANOVA has been given to analyses in which a univariate response is modeled as a function of factors.

The model is just a special case of a multiple linear regression model, however.

Numerous examples (and books) exist on this topic but much of the development has been carried out in the context of agricultural field trials (Fisher), genetics, animal breeding, and industrial processes.

We briefly outline the ANOVA approach to analysis distinguishing between crossed and nested (or hierarchical) designs and fixed and random effects modeling.

One-Way ANOVA

First we consider the simplest situation in which we have a single factor, this is known as the one-way classification.

Suppose a textile company weaves fabric on a number of looms and we are interested in whether a particular a looms produce fabric of the same strength.

We may then model the strength Y_{ij} of the j-th sample from loom i as

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \tag{8}$$

with $\epsilon_{ij} \sim_{i.i.d.} \mathrm{N}(0, \sigma^2)$, $i = 1, \ldots, a; j = 1, \ldots, r$.

We need a constraint to prevent aliasing, we may choose the sum-to-zero constraint $\sum_{i=1}^{a} \alpha_i = 0$ or the corner-point constraint $\alpha_1 = 0$ (any one of the effects could have been chosen).

One-Way ANOVA

This model is an example of the multiple linear regression

$$\mathrm{E}[Y|\boldsymbol{x}] = \boldsymbol{x}\beta,$$

where Y is $ar \times 1$, $\boldsymbol{\beta} = (\mu, \alpha_1, \ldots, \alpha_a)^\intercal$, $\boldsymbol{x}$ is $ar \times (a+1)$ and is given by

$$\boldsymbol{x} = \left[\begin{array}{ccccc}
1 & 1 & 0 & \ldots & 0 \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
1 & 1 & 0 & \ldots & 0 \\
1 & 0 & 1 & \ldots & 0 \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
1 & 0 & 1 & \ldots & 0 \\
\ldots & \ldots & \ldots & \ldots & \\
1 & 0 & 0 & \ldots & 1 \\
\ldots & \ldots & \ldots & \ldots & \\
1 & 0 & 0 & \ldots & 1
\end{array}\right].$$

TESTING

Suppose we are interested in whether there are differences between the strengths from different looms, i.e.

$$H_0 : \alpha_1 = \cdots = \alpha_a = 0.$$

We may carry out $a(a-1)/2$ t-tests but this would lead to serious problems of multiple testing.

For example, if $a = 5$ we have 10 tests of pairs of looms and with an individual type I error of 0.05 this gives an overall type I error of $1 - 0.95^{10} = 0.4$.

As an alternative we derive an F-test.

One-Way ANOVA

The above hypothesis may be tested using the F-statistic

$$F = \frac{\mathrm{FSS}(\alpha_1, \ldots, \alpha_a | \mu)/(a-1)}{\mathrm{RSS}/(ar-a)}$$

and $F \sim F_{a-1, ar-a}$ if H_0 is true where

$$\mathrm{FSS}(\alpha_1, \ldots, \alpha_a | \mu) = \mathrm{RSS}(\mu) - \mathrm{RSS}(\mu, \alpha_1, \ldots, \alpha_a).$$

An ANOVA table may then be constructed to lay out the calculations of the F-test.

The success of the F-test depends on the fact that we may decompose the overall sum of squares into the sum of independent χ^2 random variables (Cochran's Theorem).

One-Way ANOVA

Source of Variation	Sum of Squares	D of F	E[SS/DF]
Between Looms	$r\sum_{i=1}^{a}(\bar{Y}_{i.} - \bar{Y}_{..})^2$	$a-1$	$\sigma^2 + r\frac{\sum_{i=1}^{a}\alpha_i^2}{a-1}$
Error	$\sum_{i=1}^{a}\sum_{j=1}^{r}(Y_{ij} - \bar{Y}_{i.})^2$	$a(r-1)$	σ^2
Total	$\sum_{i=1}^{a}\sum_{j=1}^{r}(Y_{ij} - \bar{Y}_{..})^2$	$ar-1$	

Table 3: ANOVA table for test of $H_0 : \alpha_2 = \cdots = \alpha_a = 0$.

We are comparing the reduction in variation attributable to the between-loom parameters, to the residual variation in the data.

ONE-WAY ANOVA

- It is straightforward to extend this test to the case of different sample sizes within looms, i.e. r_i, $i = 1, \ldots, a$.

- Suppose we are interested in the strength of fabric. If there are loom effects then we cannot ignore them in our model (even if they are not of interest in themselves) because a model with no effects would not allow for the correlations between strengths from the same loom.

Crossed Designs

Suppose we have two factors, A and B with a and b levels respectively.

Then if each level of A is crossed with each level of B we have a factorial design.

Suppose that there are r replicates within each of the ab cells.

The interaction model is,

$$\mathsf{E}[Y_{ijk} | \boldsymbol{x}_{ijk}] = \mu + \alpha_i + \beta_j + \gamma_{ij},$$

for $i = 1, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, r$.

As written this model contains $1 + a + b + ab$ parameters while there are only ab means – hence, constraints are required.

Crossed Designs

In the corner-point parameterization these $1 + a + b$ constraints are

$$\alpha_1 = \beta_1 = \gamma_{11} = \cdots = \gamma_{1b} = \gamma_{21} = \ldots \gamma_{a1} = 0.$$

Alternatively we may adopt the sum-to-zero constraints:

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0.$$

To fit this model in `R` using MLE, we write `lm(y∼A+B+A:B)`, or `lm(y∼A*B)`.

Crossed Designs

Suppose we wish to test *a* treatments with *b* patients.

An example of a crossed design is one in which each patient receives each of the *a* treatments.

In this case we have two factors, treatments and patients and we have a crossed design because each treatment appears in each patient.

The main effects only model is

$$\mathrm{E}[Y_{ijk}|\boldsymbol{x}_{ijk}] = \mu + \alpha_i + \beta_j,$$

with $i = 1, \dots, a$ treatments, $j = 1, \dots, b$ patients.

Hence α_i is the change in the expected response for treatment *i*, and β_j is the change in the expected response for patient *j*.

The treatment effects are parameters of interest, between-patient parameters represent nuisance parameters.

Crossed Designs

	Treatment				
Subject	1	2	3	4	Mean
1	8.4	9.4	9.8	12.2	9.95
2	12.8	15.2	12.9	14.4	13.82
3	9.6	9.1	11.2	9.8	9.92
4	9.8	8.8	9.9	12.0	10.12
5	8.4	8.2	8.5	8.5	8.40
6	8.6	9.9	9.8	10.9	9.80
7	8.9	9.0	9.2	10.4	9.38
8	7.9	8.1	8.2	10.0	8.55
Mean	9.30	9.71	9.94	11.02	9.99

Table 4: Crossed design data, from Armitage and Berry (1994).

The response Y_{ij} represents the clotting time of plasma for individual i under treatment j, $i = 1, \ldots, 8$; $j = 1, \ldots, 4$.

Crossed Designs

These data also provide an example of a randomized block design in which the aim is to provide a more homogeneous experimental setting within which to compare the treatments.

Ignoring the blocking factor (patient) increases the unexplained variability and reduces efficiency.

There are no replicates within each of the 8×4 cells in Table 4 and so it is not possible to examine interactions between subjects and treatments.

Consequently we concentrate on the main effects only model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \tag{9}$$

for $i = 1, \ldots, 4; j = 1, \ldots, 8$ and with $\epsilon_{ij} \mid \sigma^2 \sim_{iid} \mathrm{N}(0, \sigma^2)$.

Here we adopt the corner-point parameterization with $\alpha_1 = 0$ and $\beta_1 = 0$.

Crossed Designs

Table 5 gives the generic ANOVA table (fit using MLE) for a two-way classification with no replicates and Table 6 gives the numerical values for the plasma data.

For these data, primary interest is in treatment effects (the α_i's) and Table 6 shows the steps to obtaining a *p*-value of 0.0026 for the null of

$$H_0 : \alpha_2 = \alpha_3 = \alpha_4 = 0$$

which, for this small sample size, points strongly towards the null being unlikely.

In passing, we note that there are large between-subject differences for these data, so that the crossed design is very efficient.

Crossed Designs

Source	Sum of Squares	DF	EMS	F Statistic
Factor A	$\mathrm{SS_A} = b\sum_{i=1}^{a}(\overline{Y}_{i.} - \overline{Y}_{..})^2$	$a - 1$	$\frac{\mathrm{SS_A}}{a-1}$	$\frac{\sigma^2 + b\sum_{i=1}^{a}\alpha_i^2}{a-1}$
Factor B	$\mathrm{SS_B} = a\sum_{j=1}^{b}(\overline{Y}_{.j} - \overline{Y}_{..})^2$	$b - 1$	$\frac{\mathrm{SS_B}}{b-1}$	$\frac{\sigma^2 + a\sum_{j=1}^{b}\beta_j^2}{b-1}$
Error	$\mathrm{SS_E} =$ $\sum_{i=1}^{a}\sum_{j=1}^{b}(Y_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}_{..})^2$	$(a-1)(b-1)$	$\frac{\mathrm{SS_E}}{(a-1)}$	σ^2
Total	$\mathrm{SS_T} = \sum_{i=1}^{a}\sum_{j=1}^{b}(Y_{ij} - \overline{Y}_{..})^2$	$ab - 1$		

Table 5: ANOVA table for the two-way crossed classification with one observation per cell; DF is shorthand for degrees of freedom and EMS for the expected mean square.

Crossed Designs

Source of Variation	Sum of Squares	DF	Mean Square	*F* Statistic
Treatment	13.0	3	4.34	6.62 (0.0026)
Subjects	79.0	7	11.3	17.2 (2.2×10^{-7})
Error	13.8	21	0.656	
Total	105.8	31		

Table 6: ANOVA table for the plasma clotting time data in Table 4; DF is shorthand for degrees of freedom. The quantity in brackets in the final column is the *p*-value.

We now examine treatment differences using Bayesian estimation.

Under the improper prior,

$$p(\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}$$

interval estimates obtained from Bayesian, likelihood and least squares analyses are identical.

Crossed Designs

Under a Bayesian approach we report the posterior distribution for each of the treatment effects.

We let $\boldsymbol{\theta} = [\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}]$ where $\boldsymbol{\alpha} = [\alpha_2, \ldots, \alpha_4]$ and $\boldsymbol{\beta} = [\beta_2, \ldots, \beta_8]$.

The joint posterior for $\boldsymbol{\theta}$ is multivariate Student's *t*, with $n - k - 1 = 32 - 11 = 21$ degrees of freedom, posterior mean $\widehat{\boldsymbol{\theta}}$ and posterior scale matrix, $(\boldsymbol{x}^\top \boldsymbol{x})^{-1}\widehat{\sigma}^2$, where $\widehat{\sigma}^2$ is the usual unbiased estimator of the residual error variance.

Crossed Designs

We assume a corner-point parameterization and take treatment 1 as the reference, and examine treatment differences with respect to this baseline group.

Figure 2 gives the posterior distributions for $\alpha_2, \alpha_3, \alpha_4$.

The posterior probabilities that the average responses under treatment 2, 3 and 4 are greater than treatment 1 are 0.16, 0.065 and 0.00017, respectively.

Consequently, we would conclude that there is strong evidence that treatment 4 differs from treatment 1, with decreasingly lesser evidence of differences between treatment 1 and treatments 3 and 2.

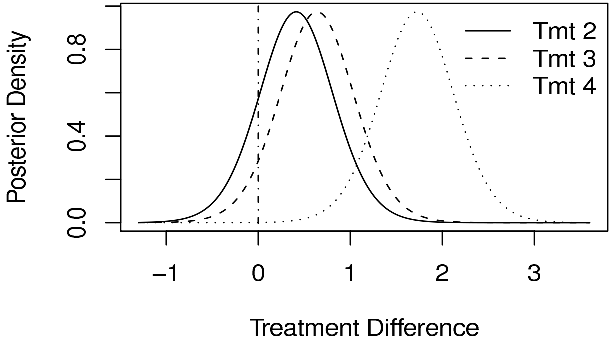

FIGURE 2: Marginal posterior distributions for the treatment contrasts, with treatment 1 as the baseline, for the plasma clotting time data in Table 4.

Crossed versus Nested Designs

The × symbols show where observations are measured.

	Treatment			
Subject	1	2	3	4
1	×	×	×	×
2	×	×	×	×
3	×	×	×	×
4	×	×	×	×
5	×	×	×	×
6	×	×	×	×
7	×	×	×	×
8	×	×	×	×

Table 7: Crossed design.

	Treatment			
Subject	1	2	3	4
1	×			
2	×			
3		×		
4		×		
5			×	
6			×	
7				×
8				×

Table 8: Nested design.

Nested Designs

For a design with two factors suppose that Y_{ijk} denotes a response at level i of factor A and level j of factor B (with replication indexed by k).

With respect to Table 7, the treatments correspond to factor A and the subjects to factor B (and there is no replication).

Then in a nested situation $j = 1$ in level 1 of factor A has no meaningful connection with $j = 1$ in level 2 of factor A.

For ML estimation in `R` this model is specified as `lm(y∼A/B)`.

Nested Designs

Suppose now that we have the previous example, but now each treatment is given to only two patients, and each patient receives just one treatment.

In this case we again have two factors, treatments and patients but the patient effects are nested within treatments.

A nested model is,

$$\mathrm{E}[Y_{ijk}|\boldsymbol{x}_{ijk}] = \mu + \alpha_i + \beta_{j(i)},$$

with $i = 1, \ldots, 4$ treatments, $j = 1, \ldots, 8$ patients, so that $\beta_{j(i)}$ represents the change in expected response for patient j within treatment i.

Random Effects Models

The above models are fixed effects ANOVA model since looms were selected to be of particular.

Suppose instead that a random sample of looms were selected (from the population of all looms in the factory – note frequentist interpretation).

In this case we may assume a random effects model in which the α_i are viewed as a random sample from a distribution.

Random Effects Models

In this case the model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

is combined with the assumption that $\alpha_i \sim \mathrm{N}(0, \sigma_\alpha^2)$, $i = 1, \dots, a$, $j = 1, \dots, r$.

This is a situation in which frequentist view unknowns as arising from a distribution – more on this in Biostat 571!

Note that we do not need a constraint in this case. No between-loom differences can then be examined via the hypothesis $H_0 : \sigma_\alpha^2 = 0$.

In this one-way classification this test turns out to be equivalent to the F-test given previously – not true for general models (which may contain both fixed and random effects, known as mixed-effects models).

The ANOVA table is very similar to that for the fixed effects model (but note the final column).

Source of Variation	Sum of Squares	D of F	E[SS/DF]
Between Looms	$r\sum_{i=1}^{a}(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2$	$a-1$	$\sigma^2+r\sigma_a^2$
Error	$\sum_{i=1}^{a}\sum_{j=1}^{r}(Y_{ij}-\bar{Y}_{i\cdot})^2$	$a(r-1)$	σ^2
Total	$\sum_{i=1}^{a}\sum_{j=1}^{r}(Y_{ij}-\bar{Y}_{\cdot\cdot})^2$	$ar-1$	

Estimation: For a likelihood approach, the marginal distribution is needed and may be derived as

$$L(\mu,\sigma^2,\sigma_a^2)=p(y_i|\mu,\sigma^2,\sigma_a^2)=\int p(y_i|\mu,\alpha_i,\sigma^2)\times p(\alpha_i|\sigma_a^2)\;d\alpha_i,$$

and results in

$$y_i|\mu,\sigma^2,\sigma_a^2\sim_{iid}\mathrm{N}(\mu 1_r,\sigma^2 I_r+\sigma_a^2 J_r),$$

where 1_r is an $r-$vector of 1's, I_r is the $r\times r$ identity matrix, and J_r is the $r\times r$ matrix of 1's.

Notes on ANOVA

Cox and Reid (2000) do not view ANOVA as only an “outgrowth” of linear models, but rather “An older and in our view more important role is in clarifying the structure of sets of data, especially relatively complicated mixtures of crossed and nested data.

This indicates what contrasts can be estimated and the relevant basis for estimating error. From this viewpoint the analysis of variance table comes first, then the linear model, not vice-versa”.

Notes on ANOVA

Randomized Block Designs: If there is another factor that is related to the response then we randomly assign the factor of interest (treatment say) within levels of this blocking factor.

The aim is to provide a more homogeneous experimental unit within which to compare the treatments.

Ignoring block increases the unexplained variability (recall mean-variance trade-off discussed earlier).

Simplest example: paired t-test where we have two observations on each individual.

With ANCOVA we have a combination of factors and quantitative covariates.

Sandwich Estimation

Sandwich Estimation

We have already examined the properties of the ordinary least squares/maximum likelihood estimator

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x})^{-1}\boldsymbol{x}\boldsymbol{Y}$$

and seen that

$$\text{var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x})^{-1}\sigma^2,$$

if we assume $\text{var}(\boldsymbol{Y} \mid \boldsymbol{x}) = \sigma^2 \boldsymbol{I}_n$.

Suppose that the correct variance model is $\text{var}(\boldsymbol{Y} \mid \boldsymbol{x}) = \sigma^2 \boldsymbol{V}$ so that the model from which the estimator was derived was incorrect.

Sandwich Estimation

The estimator is still unbiased but the appropriate variance estimator is

$$\begin{aligned} \text{var}(\widehat{\beta}) &= (\boldsymbol{x}^\intercal\boldsymbol{x})^{-1}\boldsymbol{x}^\intercal\text{var}(\boldsymbol{Y} \mid \boldsymbol{x})\boldsymbol{x}(\boldsymbol{x}^\intercal\boldsymbol{x})^{-1} \\ &= (\boldsymbol{x}^\intercal\boldsymbol{x})^{-1}\boldsymbol{x}^\intercal\boldsymbol{V}\boldsymbol{x}(\boldsymbol{x}^\intercal\boldsymbol{x})^{-1}\sigma^2, \end{aligned} \tag{10}$$

Expression (10) can also be derived directly from the estimating function

$$\boldsymbol{G}(\beta) = \boldsymbol{x}^\intercal(\boldsymbol{Y} - \boldsymbol{x}\beta),$$

since we know

$$(\boldsymbol{A}_n^{-1}\boldsymbol{B}_n\boldsymbol{A}_n^{\intercal\,-1})^{1/2}(\widehat{\beta} - \beta) \rightarrow_d \text{N}_{k+1}(\boldsymbol{0}_n, \boldsymbol{I}_n),$$

where

$$\begin{aligned} \boldsymbol{B}_n &= \text{var}(\boldsymbol{G}) = \boldsymbol{x}^\intercal\boldsymbol{V}\boldsymbol{x}\sigma^2 \\ \boldsymbol{A}_n &= \text{E}\left[\frac{\partial\boldsymbol{G}}{\partial\beta}\right] = -\boldsymbol{x}^\intercal\boldsymbol{x}, \end{aligned}$$

to give

$$\text{var}(\widehat{\beta}) = (\boldsymbol{x}^\intercal\boldsymbol{x})^{-1}\boldsymbol{x}^\intercal\boldsymbol{V}\boldsymbol{x}(\boldsymbol{x}^\intercal\boldsymbol{x})^{-1}\sigma^2.$$

Sandwich Estimation

We now describe a sandwich estimator of the variance that relaxes the constant variance assumption but assumes uncorrelated responses.

When the variance is not constant the ordinary least squares estimator is consistent (since the mean specification is correct), but the usual standard errors will be inappropriate.

Consider the estimating function

$$\boldsymbol{G}(\beta) = \boldsymbol{x}^{\mathsf{T}}(\boldsymbol{Y} - \boldsymbol{x}\beta).$$

The “bread” of the sandwich, $\boldsymbol{A}^{-1}$, remains unchanged since $\boldsymbol{A}$ does not depend on Y.

Sandwich Estimation

The "filling" becomes

$$\boldsymbol{B} = \text{var}(\boldsymbol{G}) = \boldsymbol{x}^{\top}\text{var}(\boldsymbol{Y})\boldsymbol{x} = \sum_{i=1}^{n} \sigma_i^2 \boldsymbol{x}_i^{\top} \boldsymbol{x}_i, \tag{11}$$

where $\sigma_i^2 = \text{var}(Y_i)$ and we have assumed that the data are uncorrelated.

Unfortunately σ_i^2 is unknown but various simple estimation techniques are available.

An obvious estimator stems from setting $\widehat{\sigma}_i^2 = (Y_i - \boldsymbol{x}_i\beta)^2$, to give

$$\widehat{\boldsymbol{B}}_n = \sum_{i=1}^{n} \boldsymbol{x}_i^{\top} \boldsymbol{x}_i (Y_i - \boldsymbol{x}_i\widehat{\beta})^2, \tag{12}$$

and its use provides a consistent estimator of (11). However, this variance estimator has finite sample downward bias.

Sandwich Estimation

For linear regression the MLE

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i \widehat{\boldsymbol{\beta}})^2 = \frac{1}{n} \sum_{i=1}^{n} \widehat{\sigma}_i^2,$$

is downwardly biased, with bias $-(k+1)\sigma^2/n$, which suggests using

$$\widehat{\boldsymbol{B}}_n = \frac{n}{n-k-1} \sum_{i=1}^{n} \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{x}_i (Y_i - \boldsymbol{x}_i \widehat{\boldsymbol{\beta}})^2. \tag{13}$$

Sandwich Estimation

This simple correction provides an estimator of the variance that has finite bias, since the bias in $\widehat{\sigma}^2$ changes as a function of the design points $\boldsymbol{x}_i$, but will often improve on (12).

In linear regression, if $\text{var}(Y_i) = \sigma^2$, then $\text{E}[(Y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})^2] = \sigma^2(1 - h_{ii})$ where h_{ii} is the i-th diagonal element of the hat matrix $\boldsymbol{x}(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathsf{T}}$.

Therefore, another suggested correction is

$$\widehat{\boldsymbol{B}}_n = \sum_{i=1}^{n} \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{x}_i \frac{(Y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})^2}{(1 - h_{ii})}. \tag{14}$$

For each of (12), (13) and (14) the variance of the estimator $\widehat{\boldsymbol{\beta}}$ is consistently estimated by $\widehat{\boldsymbol{A}}_n^{-1}\widehat{\boldsymbol{B}}_n\widehat{\boldsymbol{A}}_n^{-1}$.

Example: Prostate Cancer

We fit the model

$$\log \text{PSA}_i = \beta_0 + \beta_1 \log_{10}(\text{can vol}_i) + \epsilon_i \tag{15}$$

with $\epsilon_i \mid \sigma^2 \sim_{iid} \text{N}(0, \sigma^2)$.

Table 9 gives summaries of the linear association under model-based and sandwich variance estimates.

The point estimates and model-based standard error estimates arise from either ML estimation or ordinary least squares estimation of $\boldsymbol{\beta}$.

For a 10-fold increase in cancer volume (in cc), there is a $\exp(\widehat{\beta}_1) = 2.1$ increase in PSA concentration.

Example: Prostate Cancer

The sandwich estimates of the standard errors relax the constancy of variance but assume uncorrelated errors.

The standard error of the intercept is essentially unchanged under sandwich estimation, when compared to the model based version, while that for the slope is slightly increased.

The sample size of $n = 97$ is large enough to guarantee asymptotic normality of the estimator.

Parameter	Estimate	Model-based Standard Error	Sandwich Standard Error
β_0	1.51	0.122	0.123
β_1	0.719	0.0682	0.0728

Table 9: Least squares/maximum likelihood parameter estimates and model-based and sandwich estimates of the standard errors, for the prostate cancer data.

Discussion

Discussion

- OLS, Likelihood and Bayes with flat priors all inferentially agree, which is comforting.
- Always be careful to distinguish models from estimation techniques: parameterization, constraints and model checking are important for all inferential paradigms.

References

Armitage, P. and Berry, G. (1994). *Statistical Methods in Medical Research, Third Edition*. Blackwell Science, Oxford.

Cox, D. and Reid, N. (2000). *The Theory of the Design of Experiments*. Chapman and Hall/CRC, Boca Raton.