Sequential procedures for model selection
(Permanent-press fabric example)
Forward selection: start with no expl. variables
and add them one at a time.
Each step: choose the variable that most
reduces the RSS (M1)
  - compare with the previous smaller
    model (M0)
    - if \( F > F_{in} (4) \) then include, i.e. we
      have a significant reduction in RSS
      and repeat. Else stop.

Backward selection: start with all variables
and remove them one at a time.
Each step: choose the variable that least
increases the RSS (H0)
  - compare with the previous
    larger model (M1)
    - if \( F < F_{in} (4) \) then remove, i.e. we
      do not have a significant increase in RSS
      and repeat. Else stop.
9.6 Past Exam question

(Semester 1, 1999)

Demographic data are collected for 12 male patients with congestive heart failure enrolled in a study of an experimental drug.

\( y = \text{cardiac index (L/min/m}^2\text{)}, \)
\( x_1 = \text{age (years)}, \)
\( x_2 = \text{disease duration (years)}, \)
\( x_3 = \text{weight (kg)} \)

The following multiple linear models have been fitted and their residual SS shown.

<table>
<thead>
<tr>
<th>Model</th>
<th>df</th>
<th>Residual SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>11</td>
<td>2.329</td>
</tr>
<tr>
<td>( x_1 )</td>
<td>10</td>
<td>0.987</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>10</td>
<td>2.172</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>10</td>
<td>2.024</td>
</tr>
<tr>
<td>( x_1, x_2 )</td>
<td>9</td>
<td>0.798</td>
</tr>
<tr>
<td>( x_1, x_3 )</td>
<td>9</td>
<td>0.668</td>
</tr>
<tr>
<td>( x_2, x_3 )</td>
<td>9</td>
<td>1.967</td>
</tr>
<tr>
<td>( x_1, x_2, x_3 )</td>
<td>8</td>
<td>0.600</td>
</tr>
</tbody>
</table>
(a) The utility of a model is judged on whether it is significantly better than the null model. Test for the utility of the model that contains only the variables $x_1$ and $x_3$.

(b) Which model is selected by the backward elimination process?

(c) Another way to select models is to compute some measures of "goodness" of the models. Four commonly used measures are R-sq, adjusted R-sq, Mallow’s $C_p$ and $s$. Compute these for the model where the mean of $Y$ is $\alpha + \beta x_1$. ($C_p = \frac{RSS \text{ of model}}{RSS_{allv}/df_{allv}} + 2p - n$, allv denoting a model with all the variables in.)
9.6 (c) \[ H_0 : \mu_y = \alpha \]
\[ H_1 : \mu_y = \alpha + \beta_1 x_1 + \beta_3 x_3 \]
\[ F = \frac{(2.329 - 0.668)/(11 - 2)}{0.668/9} = 11.27 F_{(2, 9)}(0.99) = 8.02 \]
So \( H_1 \) is significantly better than the null model.

(b) Of the two variable models, \( x, x_3 \) has the smallest resid. SS:
\[ H_0 : \mu_y = \alpha + \beta_1 x_1 + \beta_3 x_3 \]
\[ H_1 : \mu_y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \]
\[ F = \frac{(0.669 - 0.600)/1}{0.600/9} = 0.91 \quad \text{not sign.} \quad (\leq 1) \]
\[ [F_{(1, 9)}(0.95) = 5.3] \]
So accept \( H_0 \), i.e. remove \( x_2 \).

Of the one variable models, \( x_1 \) has the smallest resid. SS:
\[ H_0 : \mu_y = \alpha + \beta_1 x_1 \]
\[ H_1 : \mu_y = \alpha + \beta_1 x_1 + \beta_3 x_3 \]
\[ F = \frac{(0.967 - 0.668)/1}{0.668/9} = 4.03 \quad \text{reject} \quad F_{(1, 9)}(0.95) = 5.12 \]
So accept \( H_0 \), i.e. remove \( x_2 \).

Compare with the null:
\[ H_0 : \mu_y = \alpha \]
\[ H_1 : \mu_y = \alpha + \beta_3 x_3 \]
\[ F = \frac{2.329 - 0.967}{0.967/10} = 14.08 \quad \text{significant} - \text{reject null} \]
\[ \therefore \text{Select model with only } x_1 \].
(c) \[ \mu_y = \alpha + \beta x_i \]

\[ R^2 = 1 - \frac{RSS}{Total SS} = 1 - \frac{0.967}{2.329} = 0.585 \]

\[ adj \ R^2 = 1 - \frac{0.967/10}{2.329/11} = 0.543 \]

\[ C_p = \frac{0.967}{0.600/8} + 2 \times 2 - 12 \uparrow \]
\[ p=2 \]

\[ s = \sqrt{0.967/10} = 0.311 \]
Chapter 10
Analysis of covariance

10.1. Introduction
10.2. Models
10.3. Extrapolation
10.4. Minitab output for *Nassarius* data
10.5. Past exam question

D+P: pages 560-562 — too little detail

O+n: Ch 16 (not 16.4) — too much detail
10.1 Introduction

So far, in all the data analysis we have done,

- Response variable - continuous, from a normal distribution.

- Explanatory variables
  - one or two or more;
  - all numerical, or all categorical

\[ \text{ANOVA} \begin{cases} \text{1-way} \\ \text{2-way} \end{cases} \]

\[ \text{Regression} \begin{cases} \text{Simple linear} \\ \text{Multiple linear} \\ \text{Polynomial} \end{cases} \]

We now consider

- explanatory variables - one numerical, and one is categorical.
Two questions requiring a similar model:

Can Chinese basketballers jump higher (on average) than Australian basketballers?

Yes, but on average they are taller.  \[ \rightarrow \text{Do taller people jump higher?} \]

Is the relationship of jumping ability to height similar for Chinese and Australian basketballers?

Yes, except the range of heights is different.

or

No, the Chinese jump higher on average with the same height.

or

No, jumping ability increases more quickly (slowly) with height for the Chinese basketballers.
Example 10.1.1 \( \textit{(Nassarius)} \)

An experiment was carried out to investigate the factors which affect the time taken for food particles to pass through the guts of \textit{Nassarius}. Two groups of food were distinguished, organic and inorganic, and various types of food particles from each group were fed to the animals. For each type of particle, the time in \underline{seconds}, and the particle size in \underline{thousandth} of a \underline{centimetre}, were recorded.

Preliminary investigation suggested that an approximately linear relationship exists between \( y = \log(\text{time}) \) and \( x = \log(\text{size}) \). The basic records were therefore transformed to a logarithmic scale and were as follows:

\begin{align*}
\text{Inorganic foods} \\
\text{log time, } y: & \quad 1.25 \quad 1.18 \quad 1.79 \quad 1.99 \quad 1.72 \quad 1.86 \quad 2.28 \quad 1.70 \quad 2.97 \quad 2.27 \quad 2.97 \\
\text{log size, } x: & \quad 1.17 \quad 1.39 \quad 1.73 \quad 1.94 \quad 2.18 \quad 2.30 \quad 2.54 \quad 2.63 \quad 3.77 \quad 3.10 \quad 3.40 \\
\text{Organic foods} \\
\text{log time, } y: & \quad 2.81 \quad 2.47 \quad 3.50 \quad 3.87 \quad 3.36 \quad 3.39 \quad 3.57 \quad 2.80 \quad 3.01 \\
\text{log size, } x: & \quad 2.48 \quad 2.31 \quad 2.55 \quad 3.49 \quad 3.45 \quad 3.53 \quad 4.07 \quad 1.91 \quad 2.23
\end{align*}

(a) What are the response variable and explanatory variable here? What type are they?

(b) What are the questions of interest?
(c) response variable: time taken - numerical.

explanatory variables:
particle size - numerical
food group - categorical

(b) Does particle size affect time taken?
Does food group affect time taken?

Is the effect of particle size dependent on food group?
(interaction?)

Response \( y \)
covariate \( x \)
dummy variable

Data entry:

<table>
<thead>
<tr>
<th>Time</th>
<th>Size</th>
<th>Food</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.25</td>
<td>1.17</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.40</td>
<td>2.97</td>
<td>1</td>
</tr>
<tr>
<td>2.81</td>
<td>2.49</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.01</td>
<td>2.23</td>
<td>2</td>
</tr>
</tbody>
</table>
The primary object of the investigation at this stage was to compare the two groups of foods and to estimate the difference between the organic and inorganic foods.

**Example 10.1.2 (Nassarius)**

A scatter plot of the data is shown below. Comment on it. (Inorganic foods is coded 1 and organic foods coded 2.)

- Larger food takes longer to pass through.
- Organic food takes longer to pass through than inorganic food.
- The relationships appear to be quite parallel, i.e., no interaction.
Example 10.2.1 (*Nassarius*)
The diagrams in Figure 10.1 suggest various possible models for the *Nassarius* data. Interpret each one and write down the mathematical specification of each model.

Figure 10.1: Possible models for *Nassarius* data
Example 10.2.1
(a) Suggest various models that could be fitted and what is the interpretation of these models.

Data: \((x_{ij}, y_{ij}); i = 1, 2, j = 1, 2, \ldots, n_i\)

- \(i = 1\) for inorganic food, \(i = 2\) for organic food.
- \(n_1 = 9, n_2 = 7, q\)

We let \(\mu(i, j) = \text{mean of } Y_{ij}\).

The following are some of the possible models.

**Model 1:** Two arbitrary straight lines, one for each group.

Inorganic:
\[
\mu(1, j) = \alpha_1 + \beta_1 x_{1j}
\]

Organic:
\[
\mu(2, j) = \alpha_2 + \beta_2 x_{2j}
\]

General (in general):
\[
\mu_{ij} = \mu + \alpha_i + \beta x_{ij} + \gamma x_{ij}
\]

where \(\alpha_1 + \alpha_2 = 0\) and \(\beta_1 + \beta_2 = 0\).

Interpretation: Both the type of food and size of food affect the passage time. The effect of size on time depends on the type of food. That is, there is interaction between type of food and size.
Model 2: Two parallel straight lines, one for each group.

Inorganic: \( \mu(1, j) = \alpha_1 + \beta x_{1j} \quad \mu(2, j) = \alpha_2 + \beta x_{2j} \)

Organic: \( \mu(1, j) = \lambda_1 + \beta x_{1j} \quad \mu(2, j) = \lambda_2 + \beta x_{2j} \)

OR (in Minitab):
\(
\mu(i, j) = \mu + \alpha_i + \beta x_{ij}, \quad \text{where } \alpha_1 + \alpha_2 = 0
\)

\( \mu(\xi, x) = \mu + \delta i + \beta x \)

Interpretation: Both the type of food and size of food affect the passage time. The effect of size on time is the same for the two type of food. That is, there is no interaction between type of food and size.

Model 3: One single straight line, for both groups of data.

For \( i = 1 \) and \( 2 \):
\( \mu(i, j) = \alpha + \beta x_{ij} \quad \mu(\xi, x) = \lambda + \beta x \)

Interpretation: The size of food affects the passage time, but the type of food does not.
**Model 4:** Two parallel horizontal lines, one for each group.

For $i = 1$ and $2$: $\mu(i, j) = \mu + \alpha_i$ where $\alpha_1 + \alpha_2 = 0$

$\mu(i, x) = \mu + \alpha_i$

$\omega(i, x) = \mu_i$

**Interpretation:** The type of food affects the passage time, but the size of food does not.

**Model 5:** One horizontal line for both groups of data.

$\mu(i, j) = \mu$

$\mu(x, x) = \mu$

**Interpretation:** Neither type of food nor size affect the passage time.

**Other models:** Depending on the data, it may be appropriate to use quadratic or cubic polynomials instead of straight lines.