15 Computational methods

Computers have freed statisticians from the grip of mathematical tractability.

15.1 Analysis of residuals from a regression

Lacking powerful computers, statisticians could once only analyse data in ways that were not computationally too difficult, or using pre-calculated tables. Modern computing power, high resolution screens and computational packages such as MINITAB and SPLUS make it easy to display and analyse data in useful ways.

In our regression models we hypothesised that errors are IID $N(0, \sigma^2)$. It is worth checking this by an analysis of the residuals. We estimate the errors by

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{a} - \hat{\beta} x_i\,.$$

It can be shown by a calculation, which we omit here, that

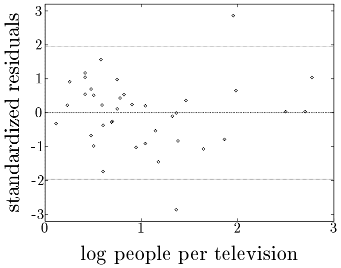
$$\mathrm{var}(Y_i - \hat{Y}_i) = \big(1 - 1/n - (x_i - \bar{x})^2/S_{xx}\big)\sigma^2.$$

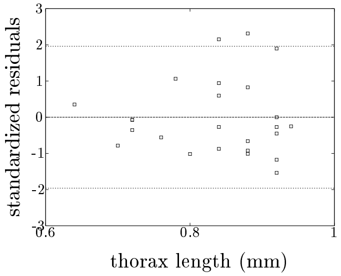
Recall that an estimate of σ^2 is obtained from the residual sum of squares, R, as $\hat{\sigma}^2 = R/(n-2)$. So we can calculate **standardized residuals**,

$$\hat{\epsilon}_{s,i} = \frac{Y_i - \hat{Y}_i}{\hat{\sigma}\sqrt{\mathrm{var}(Y_i - \hat{Y}_i)/\sigma^2}} = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - 1/n - (x_i - \bar{x})^2/S_{xx}}}\,,$$

These should be distributed approximately as $N(0,1)$.

Example 15.1 *Standardized residuals for (a) life expectancy on log people per television, and (b) fruitfly longevity on thorax length (for the 25 kept with no companions).*

(a) 

(b) 

We draw lines at ± 1.96, the values between which samples from a $N(0,1)$ will lie 95% of the time. In (a) the pattern of residuals is consistent with samples from $N(0,1)$. In (b) it looks as though the magnitude of the errors might be increasing with thorax length. This is known as 'heteroscedasticity'. Perhaps a better model would be $\epsilon_i \sim N(0, \sigma^2 x_i)$. This would suggest we try fitting, with $\eta_i \sim N(0, \sigma^2)$:

$$y_i/\sqrt{x_i} = a/\sqrt{x_i} + \beta\sqrt{x_i} + \eta_i\,.$$

15.2 Discriminant analysis

A technique which would be impossible in practice without computer assistance is the credit-scoring used by banks and others to screen potential customers.

Suppose a set of individuals $\{1, 2, \ldots, n\}$ can be divided into two disjoint sets, A and B, of sizes n_A and n_B respectively. Those in set A are known good credit risks and those in set B are known bad credit risks. For each individual we have measured p variables which we believe to be related to credit risk. These might be years at present address, annual income, age, etc. For the ith individual these are $x_{i1}, \ldots, x_{ip}$. The question is: given measurements for a new individual, say $x_{01}, \ldots, x_{0p}$, is that individual more likely to be a good or bad credit risk? Is he more similar to the people in group A or to those in group B?

One approach to this problem is to use least squares to fit a model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

where y_i is defined to be 1 or -1 as $i \in A$ or $i \in B$. Then the 'discriminant function'

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_p x_{0p}$$

is used to classify the new individual as being in group A or group B as $\hat{y}_0$ is closer to $(1/n_A)\sum_{i\in A}\hat{y}_i$ or to $(1/n_B)\sum_{i\in B}\hat{y}_i$. We do not go any further with the theory here. The point is that this is a practically important application of statistics, but a lot of calculation is required to find the discriminant function. Of course a mail order company will experiment with building its discriminant function upon different variables and doing this research is also computer-intensive.

Other uses of discriminant analysis, (and related ideas of 'cluster analysis' when there are more than two groups), include algorithms used in speech recognition and in finance to pick investments for a portfolio.

15.3 Principal components / factor analysis

Suppose we have measured a large number of variables, say $x_{i1}, \ldots, x_{ip}$, for individuals, $i = 1, \ldots, n$. Maybe these are answers to p questions on a psychological test, such as the Myers–Briggs. The question is: can we find a much smaller number of variables (factors) which explain most of the variation? In the Myers–Briggs test subjects answer a large number of questions of the sort 'when the telephone rings, are you pleased?' and the answers are converted to scores on 4 factors measuring strengths of extroversion, intuition, thinking and judging. How might these four factors have been identified from the data?

To keep things simple, we explain an approach via 'principal components analysis'. True factor analysis involves some further ideas that we skip over here. We begin by finding that linear function of the variables with the greatest variance, i.e.,

$$\text{maximize} \sum_{i=1}^{n} \left[(\beta_1 x_{i1} + \cdots + \beta_p x_{ip}) - (\beta_1 \bar{x}_1 + \cdots + \beta_p \bar{x}_p)\right]^2 \quad \text{subject to} \sum_{i=1}^{p} \beta_i^2 = 1$$

where $\bar{x}_i$ is the mean of the ith variable within the population. Equivalently,

$$\text{maximize } \beta^\top G \beta \quad \text{subject to } \beta^\top \beta = 1,$$

where G is the $p \times p$ matrix with $G_{jk} = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$. By Lagrangian methods we find that the maximum equals the largest eigenvalue of G, say λ_1, and is achieved when β is the corresponding right hand eigenvector, say $\beta^1 = (\beta_1^1, \ldots, \beta_p^1)^\top$. We call β^1 the 'first principal component'. Similarly, we can find the eigenvector β^2 of G corresponding to the second largest eigenvalue, λ_2. Continuing, we find an orthogonal set of eigenvectors $\beta^1, \ldots, \beta^m$, $m < p$, such that the proportion of variance explained, i.e.,

$$\sum_{j=1}^{m} \sum_{i=1}^{n} \left[(\beta_1^j x_{i1} + \cdots + \beta_p^j x_{ip}) - (\beta_1^j \bar{x}_1 + \cdots + \beta_p^j \bar{x}_p)\right]^2 \Bigg/ \sum_{j=1}^{p} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

is near 1. This amounts to the same thing as $\sum_{j=1}^{m} \lambda_j / \sum_{j=1}^{p} \lambda_j$; indeed the denominator above is $\text{trace}(G) = \sum_{j=1}^{p} \lambda_j$. The above ratio is also the proportion of variation explained by using least squares to fit

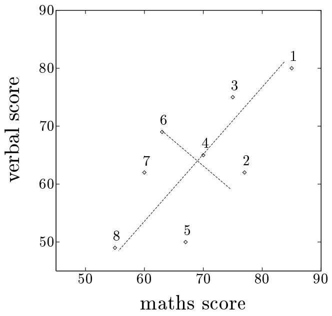
$$x_{ij} = \alpha_1^j z_{i1} + \cdots + \alpha_m^j z_{im} + \epsilon_{ij},$$

when we take $z_{ij} = \beta_1^j x_{i1} + \cdots \beta_p^j x_{ip}$. Here z_{ij} is the 'score of individual i on factor j'.

The final step is to try to give some natural interpretation to the factors, $z_1, \ldots, z_m$. For example, if we observe that the components of β^1 which are large in magnitude seem to match up with components of x which have something to do with whether or not an individual is extroverted, and other components of β^1 are near 0, then we might interpret factor 1 as an 'extroversion factor'. Then if z_{i1}, the score of individual i on this factor, is large and positive we could say that i is extroverted, and if large and negative that i is introverted.

To be fair, we should say that things are rarely so simple in practice and that many statisticians are dubious about the value of factor analysis. For one thing, the factors depend on the relative units in which the variables are measured.

Nevertheless, here is a simple illustration for $p = 2$, $m = 1$, $n = 8$. Suppose 8 students are scored on two tests, one consisting of verbal puzzles and the other of maths puzzles; the ith student scores (x_{i1}, x_{i2}). The first principal component is a line through the data which minimizes the sum of squared differences between the data points and their orthogonal projections onto this line. A reasonable name for this component might be 'IQ'. The 'IQ' of student i is $z_{i1} = \beta_1^1 x_{i1} + \beta_2^1 x_{i2}$.



student	math score	verbal score	IQ factor	mathmo factor
1	85	80	116.1	12.1
2	77	62	97.2	17.8
3	75	75	105.8	7.8
4	70	65	94.9	10.5
5	67	50	81.6	18.1
6	63	69	93.4	2.6
7	60	62	86.1	4.9
8	55	49	73.0	9.6

15.4 Bootstrap estimators

Suppose students are scored on two tests, and we wish to reduce their scores to single 'IQ' scores. Let $\mathbf{x} = (x_1, \ldots, x_8)$ be the vector of test scores, where $x_i = (x_{i1}, x_{i2})$. Define the statistic $t(\mathbf{x}) = \lambda_1(\mathbf{x}) / \sum_{j=1}^{2} \lambda_j(\mathbf{x})$, i.e., the proportion of variation that is explained by a single factor corresponding to the first principal component of $G(\mathbf{x})$. Suppose we are interested in $\theta = \mathbb{E}\, t(\mathbf{X})$, a measure of how well we can do on average when using this procedure to summarise 8 pairs of test scores in 8 single 'IQ' scores.

We can estimate θ by $\hat{\theta} = t(\mathbf{x})$. But to assess the accuracy of $\hat{\theta}$ we need to know its variance. This depends on the distribution from which our IID samples $X_1, \ldots, X_8$ have been drawn, say F. It is no surprise that there is not a nice formula for the variance of $t(\mathbf{X})$, nor that percentage points of the distribution of $t(\mathbf{X})$ have not been tabulated; (that would require some assumption about F, e.g., that it is bivariate normal).

A modern method of estimating the variance of $\hat{\theta}$ is the **bootstrap estimate**. The idea is to approximate F by the empirical distribution $\hat{F}$, a sample from which is equally likely to take any of the values $x_1, \ldots, x_8$. We take a sample of 8 pairs of tests scores from $\hat{F}$; this corresponds to randomly choosing 8 out of the set $\{x_1, \ldots, x_8\}$, *with replacement*. Perhaps we get $\mathbf{x}^* = (x_3, x_8, x_1, x_2, x_3, x_3, x_5, x_1)$. From this sample we calculate a value of the estimator, $\hat{\theta}^* = t(\mathbf{x}^*)$. We repeat this procedure B times, to get $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$. Of course we use a computer to do the random sampling, the calculation of G and of λ_1. The bootstrap estimate of the variance of $\hat{\theta} = t(\mathbf{X})$ under F is then the estimate of the variance of $t(\mathbf{X})$ under $\hat{F}$ given by

$$\hat{\sigma}_{\hat{\theta}}^2 = \frac{1}{B-1} \sum_{i=1}^{B} \left(\hat{\theta}_i^* - \frac{1}{B} \sum_{k=1}^{B} \hat{\theta}_k^* \right)^2 .$$

For the data above, $z_1 = 0.653 x_1 + 0.757 x_2$. The proportion of variation explained is $\hat{\theta} = t(\mathbf{x}) = 0.86$. A bootstrap estimate with $B = 240$ gives $\hat{\sigma}_{\hat{\theta}} = 0.094$.

Formalisation of the bootstrap method dates from 1979; the study of its use for constructing estimators, tests and confidence intervals is an active area of research.