12 The Analysis of Residuals

12.1 Errors and residuals

Recall that in the statistical model for the completely randomized one-way design,

\[ Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad i = 1, \ldots, a; j = 1, \ldots, n, \]  

(1)

the errors \( \varepsilon_{ij} \) are assumed to be NID(0, \( \sigma^2 \)). Specifically, we have the following two assumptions:

- homogeneity of variance — the error variance is the same for each treatment;
- normality — the errors are normally distributed.

The errors \( \varepsilon_{ij} \) are themselves unobservable. From the model (1),

\[ \varepsilon_{ij} = Y_{ij} - \mu - \tau_i \quad i = 1, \ldots, a; j = 1, \ldots, n, \]

but \( \mu \) and the \( \tau_i \) are unknown. However, we may replace \( \mu + \tau_i \) in the above expression by \( \hat{\mu} + \hat{\tau}_i \). This gives the residuals \( \hat{\varepsilon}_{ij} \) (as also defined in Section 10.2),

\[
\hat{\varepsilon}_{ij} = Y_{ij} - \bar{Y}_{..} - (\bar{Y}_{i.} - \bar{Y}_{..}) = Y_{ij} - \bar{Y}_{i.} \quad i = 1, \ldots, a; j = 1, \ldots, n,
\]

whose observed values, \( e_{ij} \), can be calculated from the experimental data.

The residuals are closely related to the errors but are not the same and do not have the same distributional properties. Specifically, the errors \( \varepsilon_{ij} \) are assumed to be NID(0, \( \sigma^2 \)), but the residuals \( \hat{\varepsilon}_{ij} \) are not even independently distributed of each other. They are in fact linearly dependent, since they satisfy the identities \( \hat{\varepsilon}_{i.} = 0 \) or \( e_{i.} = 0 \), \( i = 1, \ldots, a \). Furthermore, under our assumptions, the residuals are not normally distributed. Nevertheless, the residuals do have approximately the same joint distribution as the errors, and we shall use them as proxies for the errors to investigate the validity of the assumptions about the error distribution.

Recall that, from Equation (1), \( E[Y_{ij}] = \mu + \tau_i \). Replacing \( \mu + \tau_i \) by the observed versions of \( \hat{\mu} + \hat{\tau}_i \) in this expression, we obtain the fitted value, \( \hat{y}_{ij} = \hat{\mu} + \hat{\tau}_i = \bar{y}_i \). From our definition of residuals and fitted values we obtain a relationship that provides an alternative characterization of the residuals as the differences between the observed and fitted values:

\[ e_{ij} = y_{ij} - \hat{y}_{ij} \quad i = 1, \ldots, a; j = 1, \ldots, n. \]
12.2 Standardization of residuals

An unbiased estimator of the error variance $\sigma^2$ is given by $\hat{S}^2 \equiv MS_R$, and, to a first approximation, the residuals, when divided by $\hat{S}$, have the standard normal distribution. However, an improved standardization of the residuals may be used.

It may be established that

$$\text{var}(\hat{\varepsilon}_{ij}) = \left(1 - \frac{1}{n_i}\right)\sigma_i^2 \quad i = 1, \ldots, a; j = 1, \ldots, n_i.$$  

Thus an unbiased estimator of var($\hat{\varepsilon}_{ij}$) is given by $(1 - 1/n_i)\hat{S}^2$, and we define standardized residuals $D_{ij}$ by

$$D_{ij} = \frac{\hat{\varepsilon}_{ij}}{\hat{S}\sqrt{1 - 1/n_i}} \quad i = 1, \ldots, a; j = 1, \ldots, n_i.$$  

If our assumptions about the error distribution are correct, the standardized residuals are approximately NID(0,1).

The following S+ output refers to our example on tensile strength of a synthetic fibre. Recall that the object storing the results of the model fit are stored in fibre.aov. The function resid() is used to obtain the ordinary residuals $e_{ij}$ (and hence the numerator of the standardized residuals) and est.stdev is constructed so as to yield the denominator of the standardized residuals.

```r
> est.stdev <- sqrt(8.06 * (1 - 1/5))
> std.res <- resid(fibre.aov)/est.stdev
> plot(fitted(fibre.aov), std.res)
```

Another way to obtain these values is to utilize the command stdres from the S+ library MASS (which is associated with the book “Modern Applied Statistics in S”).

```r
> library(MASS)
> plot(fitted(fibre.aov), stdres(fibre.aov))
```

Various plots of the standardized residuals may be examined to check the assumptions about the error distribution. More generally, any apparently non-random pattern in the residuals or the standardized residuals may be taken to indicate inadequacy of the model in some respect.

12.3 Homogeneity of variance

One of the assumptions of the model that may be open to question is that of the homogeneity of the error variance with respect to treatments, i.e., that the variance of the errors does not depend upon the treatment. It might be conjectured that $\sigma^2$ is not constant but that it is a function of the treatment mean $\mu_i$, typically, though not necessarily, increasing with $\mu_i$. This conjecture may be investigated by plotting the observed standardized residuals $d_{ij}$ against the fitted values $\hat{y}_{ij}$ and looking to see if the dispersion of the standardized residuals appears to increase with increasing fitted value. The following plot again refers to our example.
From inspection of the plot, there is no evidence to suggest in this case that there is any tendency for the error variance to increase with treatment mean.

12.4 Normality

Assuming homogeneity of variance, the assumption that the errors are normally distributed may be checked by investigating whether the standardized residuals appear to be approximately distributed as a random sample from a standard normal distribution. This may be done in a simple way by examining the summary statistics and histogram of the standardized residuals, as in the following S+ output, which again refers to the data on tensile strength.

```r
> summary(stdres(fibre.aov))

       Min. 1st Qu.  Median        Mean 3rd Qu.       Max.
-1.496e+00 -1.024e+00  1.575e-01 -6.828e-17  5.513e-01  2.048e+00

> bpoints <- c(-1.75, -1.25, -0.75, -0.25, 0.25, 0.75, 1.25, 1.75, 2.25)
> hist(stdres(fibre.aov), breaks = bpoints)
```

To set the ‘bin’ positions of the histogram to something other than the default ones, we need to explicitly set the flag `breaks` within the `hist()` function. This can be done by creating a vector, called `bpoints` say, which contains the break points or demarcations of the endpoints of the bins, which is then assigned to `breaks`. 

---

3
If data really are a random sample from a standard normal distribution, about 95\% of the sample values should lie within the limits ±2 and it would be very rare indeed to obtain a value outside the limits ±3. If we find one or more sample values whose standardized residuals are substantially greater than 2 in modulus, and especially if they are greater than 3 in modulus, then we might regard the corresponding observations as outliers, exceptional observations requiring special investigation and, perhaps, exclusion from the analysis. A number of such large values might also indicate that the error distribution was not normal but had longer tails. There are no such exceptionally large values among our standardized residuals, the largest value in modulus being 2.048.

- Extreme outliers may possibly be due to errors in data entry — a useful check.

For a standard normal distribution, about 50\% of the observations should lie within the limits ±0.6745, so that these should be the approximate values of the sample upper and lower quartiles — although there will be great variability in these sample values for small sample sizes.

The histogram does contain some indication that the distribution of the residuals is positively skewed (skewed to the right), so perhaps there is some not very serious departure from normality in the error distribution. However, there is not nearly enough data for us to reach any firm conclusions.
12.5 Normal scores

We now present another approach to investigating whether data are from a normal distribution. If a sample, \( y_1, y_2, \ldots, y_n \), of size \( n \) has been taken then let \( y(1), y(2), \ldots, y(n) \) denote the sample values when ranked in order of increasing size, so that

\[ y(1) < y(2) < \ldots < y(n). \]

We refer to \( y(i) \), the observation of rank \( i \), as the \( i \)-th order statistic.

Let \( F_Y(\cdot) \) and \( \Phi(\cdot) \) be the cumulative distribution functions for \( N(\mu, \sigma^2) \) and \( N(0, 1) \), respectively.

Then

\[ \frac{i}{n} = P \left( Y \leq F_Y^{-1} \left( \frac{i}{n} \right) \right) \]

which implies that

\[ 1 - \frac{i}{n} = P \left( Y > F_Y^{-1} \left( \frac{i}{n} \right) \right) = P \left( \frac{Y - \mu}{\sigma} > \frac{F_Y^{-1} \left( \frac{i}{n} \right) - \mu}{\sigma} \right). \]

Since \( \frac{Y - \mu}{\sigma} \sim N(0, 1) \), it follows that

\[ \frac{F_Y^{-1} \left( \frac{i}{n} \right) - \mu}{\sigma} = \Phi^{-1} \left( \frac{i}{n} \right) \]

i.e.

\[ F_Y^{-1} \left( \frac{i}{n} \right) = \mu + \sigma \Phi^{-1} \left( \frac{i}{n} \right) \]

Now \( y(i) \) represents that point for which a proportion of \( i/n \) of the sample values lie on or below this point, for each \( i = 1, \ldots, n \). Hence, if the \( y_i, i = 1, \ldots, n \), are drawn from \( N(\mu, \sigma^2) \), then each \( y(i) \) should be a reasonable approximation to \( F_Y^{-1} \left( \frac{i}{n} \right), i = 1, \ldots, n \), and hence

\[ \Phi^{-1} \left( \frac{i}{n} \right) \approx \frac{1}{\sigma} y(i) - \frac{\mu}{\sigma}. \]

Arguably, this approximation could be further improved by replacing \( \frac{i}{n} \) by \( \frac{i - \frac{1}{2}}{n} \).

Let \( \xi_i \) denote the \( i \)-th normal score,

\[ \xi_i = \Phi^{-1} \left( \frac{i - \frac{1}{2}}{n} \right) \quad i = 1, \ldots, n. \]

Hence, our check for normality will be to plot \( \xi_i \) against \( y(i) \), for \( i = 1, \ldots, n \). Any significant departure from a linear relationship in the plot would be indicative that the original \( y_i, i = 1, \ldots, n \) are not normally distributed. In the following S+ output, standardized residuals are plotted against the normal scores using \texttt{qqnorm()} \texttt{qqline()}. To facilitate the comparison and check for normality, a line with appropriate slope and intercept is added using \texttt{qqline()}.
Since the standardized residuals are calculated to have zero mean and unit standard deviation, the plot should be approximately a straight line with unit slope, passing through the origin - `qqline()` does not necessarily produce such a line; in fact it plots a line through the first and third quartile of the data, and the corresponding quantiles of the standard normal distribution. However, this is good enough for our purposes!

If you want to actually produce a line with intercept 0 and slope 1, then use `abline(0,1)`.

The plot appears to be more or less linear, although the kink in the region of -1.2 for the standardized residuals and -1 for the quantiles indicates some departure from normality, corresponding to the small peak centred around -1.2 in the histogram of the standardized residuals.

We may, as above, just inspect the plot visually to see if it appears to be roughly linear. The data sets that we shall use are relatively small and the evidence regarding normality is often not very conclusive either way.

Certain test statistics may be calculated which essentially are measures of the linearity of the plot of the sample values against their normal scores. There are special tables of the percentage points of the sampling distributions of such statistics. One such statistic is simply the square of the correlation coefficient between the $y_{(i)}$ and the $\xi_i$. 

\[
> \text{qqnorm}(`stdres(fibre.aov))$
\]
\[
> \text{qqline}(`stdres(fibre.aov))$
\]
12.6 Appendix: transformations to stabilize variance

If we decide that the assumption of equal variances is untenable, we may carry out a transformation of the response variable to stabilize the variance, i.e., to make the variance at least approximately the same for each treatment.

We may find appropriate transformations by making use of an approximate result for the variance of a function of a random variable. Given a r.v. $Y$ with $E[Y] = \mu$ and $\text{var}(Y) = \sigma^2$, consider in general some function $g(Y)$ and its Taylor series expansion about $\mu$,

$$g(Y) = g(\mu + (Y - \mu)) = g(\mu) + g'(\mu)(Y - \mu) + \ldots$$

Taking expectations and noting that $E[Y - \mu] = 0$, to a first approximation,

$$E[g(Y)] = g(\mu).$$

To the same level of approximation,

$$[g(Y) - E(g(Y))]^2 = g'(\mu)^2(Y - \mu)^2.$$  

Taking expectations, we obtain the approximate formula:

$$\text{var}(g(Y)) = g'(\mu)^2\sigma^2. \quad (3)$$

To summarize:

<table>
<thead>
<tr>
<th>mean</th>
<th>variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>$\mu$</td>
</tr>
<tr>
<td>$g(Y)$</td>
<td>$g(\mu)$</td>
</tr>
</tbody>
</table>

We now turn to the problem of finding a transformation of the data to stabilize the variance. We adapt the earlier notation by writing $\mu$ for the treatment mean $\mu_i$ and $\sigma^2(\mu_i)$ for the treatment variance $\sigma^2_i$, where it is now explicitly assumed that the treatment variance $\sigma^2(\mu)$ is some function of the treatment mean $\mu$. Substituting into the formula of Equation (3), we obtain

$$\text{var}(g(Y)) = g'(\mu)^2\sigma^2(\mu).$$

Hence, to obtain the transformation $g$ that stabilizes the variance to a constant value, we require that

$$g'(\mu) \propto \frac{1}{\sigma(\mu)}.$$  

For example, if $\sigma^2(\mu) \propto \mu$ then

$$g'(\mu) \propto \frac{1}{\sqrt{\mu}},$$

7
and to stabilize the variance we may use the square root transformation,

\[ g(x) = \sqrt{x}. \]

If \( \sigma(\mu) \propto \mu \) then

\[ g'(\mu) \propto \frac{1}{\mu}, \]

so that we use the logarithmic transformation,

\[ g(x) = \ln x. \]

- A transformation of the data, which may have been carried out primarily to stabilize the variance, also has the effect of altering the error distribution, possibly transforming a non-normal distribution to something closer to normality.

Having decided upon a transformation, we may then transform the response variable and apply the ANOVA to the transformed data. Such transformations are readily carried out in S+ using `sqrt()` and `log()`.