1 Chapter 6: Checking assumptions

1.1 Valid results depend on model assumptions

This point is discussed at the beginnnig of Chapter 6.

1.2 Model diagnostics: residual analysis

Recall that the residuals e_{ij} in the ANOVA model $y_{ij} = \mu_i + e_{ij}$ should 1) be independent, 2) have a normal distribution, and 3) have a common variance σ^2. Of these assumptions, the most important (and hardest to check) in terms of the effect of a violation is 1). In terms of importance of a violation the next most important assumption is 3), and finally 2) . Transformations are a common classical remedy for problems with assumptions 2) and 3) . The ANOVA null hypothesis is not affected by transformation, but confidence intervals for the mean are affected, and can instead be interpreted as confidence intervals for the median on the original scale. We can estimate the e_{ij} terms from our sample via one of several types of residuals:

Name of residual	**Formula**	**SAS Proc GLM keyword**	**R function**
residual	$r_{ij} = y_{ij} - \widehat{y}_{ij} = y_{ij} - \overline{y}_{i.}$	r = or residual =	residuals()
standardized residual	$z_{ij} = r_{ij}/\sqrt{MS_E}$		
studentized residual	$s_{ij} = r_{ij}/\sqrt{MS_E(1-H_{ij})}$	student =	rstandard()
jackknife residual	$t_{ij} = r_{ij}/\sqrt{MS_{E(-ij)}(1-H_{ij})}$	rstudent =	rstudent()

where $MS_{E(-ij)}$ is MS_E computed without the jth observation in the ith group. The last three types (standardized, studentized, and jackknife) are fairly similar if the model assumptions are satisfied. A sound strategy for diagnosing the adequacy of a linear model is to obtain a variety of plots of residuals. These include histograms, normal plots, and scatter plots of the residuals versus the predicted values. A useful way to evaluate the collection of residual plots is to look at the plots for problems, and focus on characterizing the potential effect of these problems on the model. Departures from independence due to either time or spatial effects can be assessed via residual-by-time plots or variogram plots, respectively. In a residual-by-time plot, autocorrelation can be detected if successive residuals are either too close together (positive autocorrelation) or too far apart (negative autocorrelation). If there is no spatial association then the variogram should be relatively flat.

1.3 Transformations to help satisfy model assumptions

The text discusses different approaches for obtaining transformations of data to help meet model assumptions, focusing primarily on two methods, i) variance stabilizing transformations for known distributions (Table 6.3), and ii) power transformations suggested by either a regression of $\log(\overline{y})$ on $\log(s)$ or use of the Box-Cox method.

The idea behind the first power transformation method is that if $\sigma_i = \alpha\mu_i^{\beta}$, then we have the relationship $\log(\sigma_i) = \log(\alpha) + \beta\log(\mu_i)$. Thus we estimate the slope of the least-squares line predicting $\log(s)$ from $\log(\overline{y})$ as $\widehat{\beta}$. Then if we select a power transformation $x = y^{\lambda}$, with λ is chosen as $\widehat{\lambda} = 1-\widehat{\beta}$, the new variable x should have a variance that is approximately constant.The Box-Cox method uses a likelihood-based approach and is more widely applicable.