CHECKING MODEL ASSUMPTIONS (CH5)

The experimenter should carefully choose a model before collecting data.

It's also important to use the data, once it has been collected, to check the model.

We'll use *residual plots*.

Recall: The *residuals* are

$$\hat{e}_{it} = y_{it} - (\hat{\mu} + \hat{\tau}_i) = y_{it} - \overline{y}_{i\cdot}$$

Plots of residuals typically show trends more readily than plots of response values.

Comments on residuals:

- Minitab will store residuals when doing an Analysis of Variance if you check the appropriate box in the dialogue window.

  Example: Batteries

- The residuals sum to zero. This follows readily from the first step in deriving the normal equations.

  Example: Check with Battery data using Column Statistics in Minitab

- Thus the sample mean of the residuals is zero.

- Recall from the handout Estimating Parameters and Variance that ssE $= \sum_{i=1}^{t}\sum_{t=0}^{r_i}\hat{e}_{it}^{\,2}$.

- The sample variance of the residuals is thus $\sum_{i=1}^{t}\sum_{t=0}^{r_i}\hat{e}_{it}^{\,2}$ /(n-1) = ssE/(n-1).

Recall model assumptions:

The errors ε_{it} are independent random variables with distributions $N(0, \sigma^2)$.

Since the $\hat{e}_{it}$'s are estimates of the ε_{it}'s, they should also be approximately independent, with approximate $N(0, \sigma^2)$ distributions.

Note: Since the residuals sum to zero, they are *not* independent. But other evidence of lack of independence may be taken as lending doubt to the assumption that the errors ε_{it} are independent.

Many prefer to use *standardized residuals* -- residuals divided by their sample standard deviation:

$$z_{it} = \frac{\hat{e}_{it}}{\sqrt{ssE/(n-1)}}$$

Plots of standardized residuals make it a little easier to identify outliers than do plain residual plots.

As above, if the model assumptions are correct, the standardized residuals should be approximately independent (except for the fact that they sum to zero) and have approximately a N(0,1) distribution.

Note: Standardized residuals can be formed from residuals in Minitab by using the Calculate menu.

Example: Battery data

Additional note for students who have had regression: We don't typically use leverages in "standardizing" residuals for regression, even though ANOVA can be done by regression. But when we do ANOVA by regression, we get leverages that are all the same. However, there are other types of "studentized" residuals that are sometimes used with ANOVA.

Recommended Order for Checking Model Assumptions:

1. Check the form of the model.

2. Check for outliers

3. Check for independence.

4. Check for constant variance.

5. Check for normality.

Rationale for this order: The checks are ordered so that you should stop if you find a major problem. (Possibly remedies to problems will be discussed later.)

1. If you've got the wrong model, everything else will be messed up.

2. Outliers might indicate mistakes in recording data.

3. Procedures are least robust to lack of independence.

4. ANOVA is next least robust to lack of constant variance. Also, lack of independence might falsely suggest lack of constant variance, or might hide it.

5. ANOVA is fairly robust to some departure from normality.

1. *Check the form of the model.*

Plot residuals against :

- Each independent variable (treatment factor, block factor, or covariate) included in the model.
- Levels of factors not included in the model.

Any non-random pattern suggests lack of fit of the model.

Possible remedy: Transformation or more sophisticated model.

Example: Battery experiment

Plot residuals against

- Type
- Order

2. *Check for outliers.*

Plot (standardized) residuals against levels of treatment factor.

- If normality assumption is true, standardized residuals beyond ± 3 are likely outliers.
- Outliers should be investigated for possible recording errors.

  Example: Battery experiment (p. 108)

- Outliers or unusual numbers of standardized residuals beyond ±2 may indicate non-normality (in which case, transformations or non-parametric methods may be needed)
- Outliers should *not* automatically be discarded.

3. *Check for independence of error terms*.

- Do before checks for constant variance and normality.

- Plot residuals against time or spatial variables, or anything else likely to cause lack of independence. (*Always record these when doing an experiment!)*

- Data showing time or other dependence might be salvaged by analysis of covariance. (Might not be covered in this class.)

Examples: Battery against order.

4. *Check for equal variance*

- Plot residuals against fitted values. A pattern of increasing variance as mean increases is the most common (but not the only possible) sign of unequal variances.

- Also compare sample variances of residuals for each treatment:

  $s_i^2 = \frac{1}{r_i - 1}\sum_{t=1}^{r_i} \hat{e}_{it}^2$ is an unbiased estimator of the error variance σ_i^2 of the i^{th} treatment group.

  Rule of thumb (from simulation studies): If the ratio s^2_{max}/ s^2_{min} of the largest treatment variance to the smallest does not exceed 3 (some say 4), then the inference procedures for the equal variance model are still approximately valid.

  (However, a larger ratio might also occur by chance even when model assumptions are correct.)

- With unequal sample sizes, departures from equal variance can have different effects depending on the relationship between sample size and treatment variance.

- With unequal sample sizes and/or unequal variance, looking at the p-value (effectively setting a smaller alpha) may give additional information in deciding whether to reject the null hypotheses.

Methods for dealing with unequal variance:

a. Transform the response variable: Replace Y with h(Y), so that the model becomes

   - $h(Y_{it}) = \mu^* + \tau_i^* + \varepsilon_{it}^*$

   - The ε_{it}^*'s are independent random variables.

   - For each i and t, $\varepsilon_{it}^* \sim N(0\,,\,\sigma^2)$

   (More later)

b. Use a method such as Satterthwaite's (pp. 116 – 117; we won’t cover this)

   - It's only approximate and less powerful than an equal - variance model.

   - It may be useful if there is no suitable transformation available or if it is important to keep the original units. Sometimes this can be useful in getting at least something from the data after a “disaster” in running the experiment.

c. There are also “Weighted ANOVA” methods, analogous to weighted least squares in regression.

d. Use a non-parametric test such as the Kruskal-Wallis test for medians. (Not covered in this course.)

e. In some circumstances, graphical techniques may be sufficient. (e.g., if the purpose is to see if the groups are really the same in the variable of interest, and there is enough data to be confident from graphical methods that one group has a distribution vastly different from that of the others.)

f. Resampling, permutation, quantile regression, or Bayesian methods might be useful. (Not covered in this course.)

5. *Check for normality*

- Form a *normal probability plot* of residuals. (Details later)

- Approximate normality is usually good enough for inferences concerning treatment means and contrasts to be reasonably good, especially if sample sizes are large (thanks to the CLT).

- Heavy tails can be a problem -- non-parametric methods may be better in this case.

- Transformations can sometimes achieve normality -- but care is needed to get equal variance as well.

Probability Plots

Many tests or other procedures in statistics assume a certain (e.g., normal) distribution. Some procedures are *robust* (i.e., still work pretty well) to some departures from assumptions, but often not to dramatic ones.

This raises the question: How to judge whether data come from a given distribution?

Histograms *don*'t serve this purpose well -- e.g., bin sizes, samples sizes, and their interaction cause problems.

Probability plots (also known as *Q-Q plots* or *quantile plots*) are not perfect, but somewhat better. The idea:

- Order the data: $y_1 \leq y_2 \leq \ldots \leq y_n$.
- Plot vs $q_1 \leq q_2 \leq \ldots \leq q_n$, where

  q_k = the expected value (as approximated by computer) of the kth smallest member of a simple random sample of size n from the distribution of interest.

If the data come from this distribution, we expect $y_k \approx q_k$, so the graph will lie approximately along the line y = x.

(We will apply this when the y_i's are the residuals or standardized residuals.)

Variation often used to test for normality:

Take the q_k's from the *standard normal* distribution. So if the y_k's are sampled from an N(μ,σ) distribution, then the transformed data $\frac{y_k - \mu}{\sigma}$ come from a standard normal distribution, so we expect

$$\frac{y_k - \mu}{\sigma} \approx q_k$$

In other words, if the y_k's are sampled from an N(μ,σ) distribution, then

$$y_k \approx \sigma\, q_k + \mu,$$

so the graph should lie approximately on a straight line with slope and intercept σ and μ, respectively.

Note: Minitab uses a variation where the y_i's are on the horizontal axis.