Stat 3515

Lecture 4

Model Adequacy - Checking the Underlying Assumptions

The model for the one way ANOVA is given by:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

where $i=1,....,a$ and $j=1,....,n_i$.

Fixed Effects	**Random Effects**
1. τ_i is a fixed parameter.	1. τ_i are iid $N(0, \sigma_\tau^2)$
2. ε_{ij} are iid $N(0,\sigma_\varepsilon^2)$.	2. ε_{ij} are iid $N(0, \sigma_\varepsilon^2)$.
3. μ is a fixed constant	3. τ_i and ε_{ij} are independent.
	4. μ is a fixed constant

We will now proceed to examine these assumptions and present remedial measures that are used in the case these assumptions do not hold. The main diagnostic tools will be based on the *residuals*. For the one-way ANOVA model the residuals are given by:

$$e_{ij} = \hat{\varepsilon}_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \bar{Y}_{i.} .$$

If the randomization process of assigning units to treatments is done properly then the independence and identically distributed assumption is usually satisfied.

To examine the normality assumption we will employ the normal probability plot and the W test via proc univariate.

Model Diagnostics-Analysis of Residuals for ANOVA models.

Most of the regression diagnostics are based on the analysis of various types of residuals. We proceed to define these residuals for a one way ANOVA model and discuss their properties and applications.

The results remain valid for more general ANOVA models, under the classical assumptions for the random errors. The *raw residuals* or just residuals have been defined as

$$e_{ij} = Y_{ij} - \widehat{Y_{ij}} = Y_{ij} - \overline{Y_{i.}} \ .$$

1. $\sum_{i=1}^{a} \sum_{j=1}^{n_i} e_{ij} = 0.$

2. This immediately implies the raw residuals are *dependent.* If the number of observation in the data set is large this dependence can be ignored.

3. The estimate of the variance of the random errors, ε_{ij}, is based on the residuals:

$$MSE = \frac{1}{N-a} \sum_{i=1}^{a} \sum_{j=1}^{n_i} e_{ij}^2,$$

where

$$N = \sum_{i=1}^{a} n_i$$

is the number of observation in the data set. One can show that

$$E(MSE) = \sigma^2.$$

To simplify the presentation of the results we adopt the following notation:

$$S^2 = MSE.$$

4. Since the range of the raw residuals varies from one data set to another, depending on the units of the data set, it makes sense to standardize the residuals. One approach is to define the *standardized residuals*

$$z_{ij} = \frac{e_{ij}}{S}.$$

The standardized residuals have unit sample variance in the sense that

$$\frac{1}{N-a}\sum_{i=1}^{a}\sum_{j=1}^{n_i} z_{ij}^2 = \frac{1}{N-a}\sum_{i=1}^{a}\sum_{j=1}^{n_i}\frac{e_{ij}^2}{S^2} = 1.$$

5. One can show that for $i = 1, ..., a$ and $j = 1, ..., n_i$

$$E(e_{ij}) = 0$$

and

$$Var(e_{ij}) = \sigma^2(1 - h_{ij}),$$

where h_{ij} is called the *leverage* of Y_{ij} observation. The expression for h_{ij} is complex and will not be presented here. Also, one can show that the $e'_{ij}s$ are combination of the observations Y_{ij}, $i = 1, ..., a$ and $j = 1, ..., n_i$, and therefore have a normal distribution.

6. Since the $e'_{ij}s$ do not have the same population variance a better way to standardize the residuals is to consider the *studentized residuals* defined as follows:

$$r_{ij} = \frac{e_{ij}}{S\sqrt{1 - h_{ij}}}.$$

In some books these are called *internally* studentized residuals. The reason for that is that S is a function of $e'_{ij}s$ and therefore not independent of it. Studentized residuals have a mean that is close to zero and its variance can be estimated by

$$\frac{1}{N-a}\sum_{i=1}^{a}\sum_{j=1}^{n_i} r_{ij}^2,$$

which slightly greater than 1. For large N, r_{ij} have an approximate t-distribution with $N - a$ degrees of freedom. Also, $r_{ij}^2/(N - a)$ have a beta distribution with parameters $1/2$ and $(N - 2)/2$.

7. Analogously one can define the so called *jackknife residuals*:

$$r_{(-ij)} = \frac{e_{ij}}{S_{(-ij)}\sqrt{1 - h_{ij}}} = r_{ij}\sqrt{\frac{N - a - 1}{N - a - r_{ij}^2}},$$

where $S^2_{(-ij)}$ is the residual variance computed with the ith observation deleted, i.e.

$$S^2_{(-ij)} = \frac{1}{N-a-1} \sum_{u=1,u\neq i}^{a} \sum_{v=1,v\neq j}^{n_u} e^2_{uv}.$$

In some books these residuals are called *externally* studentized residuals. Jackknife residuals also have mean approximately equal to zero and its variance can be approximated by

$$\frac{1}{N-a-1} \sum_{i=1}^{a} \sum_{j=1}^{n_i} r^2_{(-ij)},$$

which is also slightly greater than 1. If the classical assumptions hold then each of the jackknife residuals have an exact t distribution with $N-a-1$ degrees of freedom.

Remarks.

1. Under normal circumstances all types of residuals will exhibit similar behavior. On the other hand, if we have unusual observations, with a high value of h_{ij}, then the r_{ij} and especially $r_{(-ij)}$ will point out this problem.

2. One can show that

$$\widehat{Y}_{ij} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} h_{ij} Y_{ij} = h_{ij} Y_{ij} + \sum_{u=1,u\neq i}^{a} \sum_{v=1,v\neq j}^{n_u} h_{uv} Y_{uv}.$$

If Y_{ij} has an unusually large value of h_{ij}, compared to the other observations, then Y_{ij} will have a greater effect on the value of $\widehat{Y}_{ij}$. In some cases an unusual observation needs to be removed. Observations with large values of h_{ij} will have a small value of $Var(e_{ij})$, and regardless of what value of Y_{ij} is, it will have a raw residual close to 0. That is the reason why it is important to examine studentized or even better the jackknifed residuals to detect effectively unusual observations, that might not belong to the data set.

Outliers.

Outliers are extreme observations that do not belong to the data set. They might cause the normality assumption for the random errors to be

not valid. Also, outliers might cause the variance of the random errors to be not constant or even violate the linearity assumption of the model. To detect outliers one examines the various types of residuals. For example, if we decide to examine the jackknifed residuals the fact that they have a t distribution with $N - a - 1$ degrees of freedom is helpful.

As a rule of thumb we will declare an observation as an outlier if its jackknife residual has an absolute value larger than 4.

Outliers detection based on the leverages has been discussed by Hoaglin and Welsch (1978). They recommend to look carefully at any observation with an leverage

$$h_{ij} > 4/N.$$

Influential Observations.

One of the most popular measures for evaluating influence of an observation is *Cook's distance.* It measures the extent of change in the estimates of the regression coefficients for the linear model, if a particular observation is deleted. Cook's distance for the ij th observation is given by

$$d_{ij} = \frac{1}{2} r_{ij}^2 \frac{h_{ij}}{1 - h_{ij}} = \frac{e_{ij}^2 h_{ij}}{2S^2(1 - h_{ij})^2}.$$

Cook and Weisberg (1982) recommend to scrutinize any observation with $d_{ij} > 1$.

Another measure to examine influence of observations is

$$DFFITS_i = \frac{\widehat{Y_{ij}} - \widehat{Y_{(-ij)}}}{S_{(-ij)}\sqrt{h_{ij}}} = \frac{S}{S_{(-ij)}}\sqrt{2d_{ij}} = r_{ij}\sqrt{\frac{h_{ij}}{1 - h_{ij}}}.$$

If $|DFFITS_i)| > 2$ the observation has to be scrutinized.