DOE (Linear Model)
– Strategy for checking experimental model assumptions
Part 1

When we discuss experiments whose data are described or analyzed by the one-way analysis of variance ANOVA model, we note that a complete statement of this model must make clear of its error assumptions. This is because any violation of the assumptions may lead to incorrect or misleading analysis results.

For example, if the assumption of error independence is violated, the simple one-way ANOVA is no long appropriate. If the assumption of error normality is violated (in other words some outliers are present), the one-way ANOVA test is less superior than a nonparametric test. Furthermore unequal population variances would be even more damaging assumption violation.

Therefore, checking the assumptions becomes an important exercise in the one-way ANOVA on experimental data.

For a completely randomized design with n specifically selected treatments (fixed effects), the model is,

$$Y_{it} = \mu + \tau_i + \varepsilon_{it}\ ; \quad \varepsilon_{it} \sim N(0, \sigma^2) \tag{1}$$

where

$\sim N(0, \sigma^2)$ denotes "has a mean 0 and variance σ^2".

ε_{ij}*'s* are mutually independent,

$t = 1, 2, \ldots, r_i$; $\quad i = 1, 2, \ldots, n$

This model implies that the response variable Y_{it} are mutually independent

and have a normal distribution with mean $\mu + \tau_i$ and variance σ^2 , that is

$$Y_{it} \sim N(\mu + \tau_i, \sigma^2).$$

Hence, when the data of the experiments have been collected, we have to study the adequacy of the experimental model by checking the model assumptions. Usually a series of pilot scale experiments is conducted to check these assumptions before the main experiments are conducted in order to save time and costs.

The general strategy for checking these assumptions is to use the following sequence of checks:

1. Check the form of the model – are the mean responses for the treatment adequately described by $E[Y_{it}] = \mu + \tau_i,\ i = 1, 2, \ldots, n$
2. Check for outliers – are there any unusual observations?
3. Check for independence – do the error variables ε_{it} appear to be independent?
4. Check for constant variance – do the error variables ε_{it} have similar variances for each treatment?
5. Check for normality – do the error variables ε_{it} appear to be a random sample from a normal distribution?

The check items (3) to (5) are distributional assumptions about the residuals or errors. Many of these checks can be done visually on the residual plots made.

Let us illustrate how these checks can be performed from the results of a series of hypothetical experiments with a fixed factor having four levels A, B, C and D and 10 replicates for each level (Table 1):

Table 1:
Results of 4-leveled experiments with 10 replicates

A	B	C	D
6.7	9.9	10.4	9.3
7.8	8.4	8.1	9.3
5.5	10.4	10.6	7.2
8.4	9.3	8.7	7.8
7.0	10.7	10.7	9.3
7.8	11.9	9.1	10.2
8.6	7.1	8.8	8.7
7.4	6.4	8.1	8.6
5.8	8.6	7.8	9.3
7.0	10.6	8.0	7.2

From here, we have found:

	A	B	C	D
Mean	7.20	9.33	9.03	8.69
Std Deviation	1.019	1.717	1.135	1.000
Variance	1.038	2.947	1.289	1.001

Now the assumptions on the model involve the error variable, $\varepsilon_{it} = Y_{it} - E[Y_{it}]$ and can be checked by the analysis of its residuals. The (it)th **residual** ε_{it} is defined as the observed value of $Y_{it} - \hat{Y}_{it}$ where $\hat{Y}_{it}$ is the least square estimator of $E[Y_{it}]$. That is,

$$\varepsilon_{it} = y_{it} - \hat{y}_{it}$$

For the one-way analysis of variance model, we have $E[Y_{it}] = \mu + \tau_i$, so the (it)th residual ε_{it} is

$$\varepsilon_{it} = y_{it} - (\mu + \tau_i) = y_{it} - \bar{y}_{it} \qquad (2)$$

That is to say the residual ε_{it} is the difference between the observed value and the mean value of the replicates, i.e. $y_{it} - \bar{y}_{it}$ for the one-way analysis of variance model. For example, when the results of Treatment A = 6.7, the residual = 6.7−7.2 = −0.50.

All the residuals of the data are calculated and tabulated in Table 2 as below:

Table 2:
Calculated residuals of experimental data

A	B	C	D
−0.50	0.57	1.37	0.61
0.60	−0.93	−0.93	0.61
−1.70	1.07	1.57	−1.49
1.20	−0.03	−0.33	−0.89
−0.20	1.37	1.67	0.61
0.60	2.57	0.07	1.51
1.40	−2.23	−0.23	0.01
0.20	−2.93	−0.93	−0.09
−1.40	−0.73	−1.23	0.61
−0.20	1.27	−1.03	−1.49

If we square each residual value in Table 2 and sum them up altogether, we get sum of squares of error (SSE) being 56.47. The mean square error (MSE) is to divide SSE by the total degrees of freedom and in this case, the degrees

of freedom = 40 −1 = 39. Hence, the standard error (SE) which is the square root of MSE is:

$$SE = \sqrt{\frac{SSE}{n-1}} = \sqrt{\frac{56.47}{39}} = 1.203$$

Now, let us standardize or normalize the residual data because standardization facilitates the identification of outliers.

If the assumptions of the model are correct, the standardized error variables ε_{it}/σ are independently distributed with a $N(0,1)$ distribution, so the observed standardized value $\varepsilon_{it}/\sigma = (y_{it} - (\mu + \tau_i))/\sigma$ would constitute independent observation from a standard normal distribution.

In short, the standardized residuals are obtained from the Table 2 by equation:

$$z_{it} = \frac{\varepsilon_{it}}{SE} \qquad (3)$$

and we obtain the following standardized residual values in Table 3:

Table 3:
Standardized residuals of the experimental data

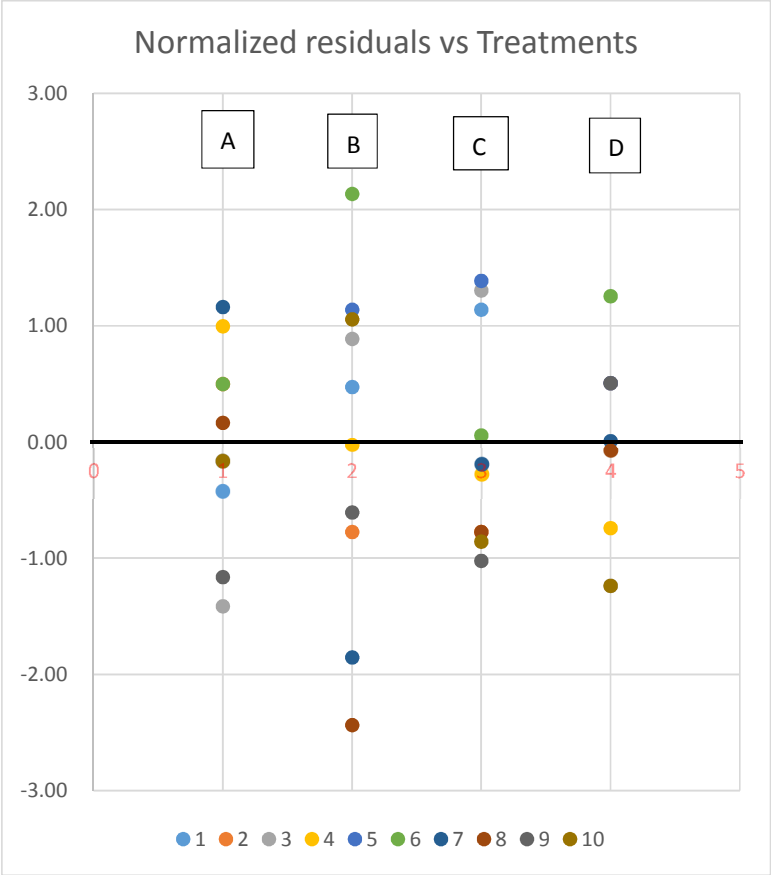
A	B	C	D
−0.42	0.47	1.14	0.51
0.50	−0.77	−0.77	0.51
−1.41	0.89	1.30	−1.24
1.00	−0.02	−0.27	−0.74
−0.17	1.14	1.39	0.51
0.50	2.14	0.06	1.25
1.16	−1.85	−0.19	0.01
0.17	−2.43	−0.77	−0.07
−1.16	−0.61	−1.02	0.51
−0.17	1.06	−0.86	−1.24

1. Checking the fit of the model

The first assumption to be checked is the assumption that the model $E[Y_{it}]$ for the mean response is correct. Our purpose of running a pilot scale experiment is to choose a model that is a reasonable description of the experimental data. If this is done, the model assumption checks for the main experiment should pose no problems. If the model for mean response does not adequately fit the data, we say the model is *lack of fit*.

In general, the fit of the model is checked by plotting the standardized residuals versus the levels of each individual independent variables (treatment factor, treatment block or covariate) included in the model. Lack of fit is indicated if the residuals exhibit a non-random pattern about zero in any such plot, being too often positive for some levels of the independent variable and too often negative for others.

We now examine if the above example data conform to $\sim N(0,1)$ by plotting the normalized residuals against the four levels (A, B, C, and D) of only treatment factor as shown in Figure 1 below:



From the above plot, it is observed that the standardized residuals of the various levels of treatment factor scattered randomly around zero without noticeable trend and there were no unusually large or small normalized observations, being $-3 < z_{it} < +3$. The lack of fit can also be checked if the normalized values were to be plotted against the levels of other treatment factors which were omitted in the experiments. If the plot were to show a pattern, it indicates that that factor should have also been included in the model as a covariate.

2. Check for outliers

If there is any suspicion on extremely large or small experimental data collected in the analysis, there are many outlier statistic tests to choose from. The popular outlier tests are Dixon's, Grubb's, etc.

However ultimately the experimenter has to decide if he wants to keep the unusual value in the analysis or to omit it altogether. Another approach is to analyze the data for its mean without the unusual value. If the conclusions of the experiments remain the same, then the outlier can safely be left in the analysis. On the other hand, if the conclusions change dramatically, the outlier is said to be influential and the experimenter has to make a professional judgment as to whether the outlying observation is likely to be experimental error or whether this unusual observation do occur from time to time.

If the experimenter decides on the former, then the analysis should be reported without the outlier. On the other hand, he can conclude that the model is not adequate to describe the experimental situation and a more complicated model would be needed.

(End of Part 1. Part 2 will discuss the checking of error independency, equal population variance and the normality assumptions.)