STAT 441/541 Statistical Methods II
Handout: Checking ANOVA Assumptions

Introduction

For any statistical method that we use, it is always important to be aware of underlying assumptions. Our text refers to these as conditions in Chapter 8.

Section 8.2 Page 405

The analysis of variance procedures are developed under the following conditions:

1. Each of the t populations has a normal distribution.
2. The variances of the t populations are equal; that is, $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2 = \sigma^2$.
3. The t sets of measurements are independent random samples from their respective populations.

Section 8.3 Page 412

We will impose the following conditions concerning the sample measurements and the population from which they are drawn:

1. The samples are independent random samples. Results from one sample in no way affect the measurements observed in another sample.
2. Each sample is selected from a normal population.
3. The mean and variance for population i are, respectively, $\mu_i \; and \; \sigma_i^2 \; (i = 1, 2, \ldots, t)$. The t variances are equal: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2 = \sigma^2$.

Combined assumptions for ANOVA:

1. The t sets of measurements are independent random samples from their respective populations. The samples are independent such that results from one sample in no way affect the measurements observed in another sample.
2. Each sample is selected from a normal population and the errors are independently normally distributed with a mean of 0 and a common variance of σ^2, $\varepsilon_{ij} \sim Normal(0, \sigma^2)$.
3. The mean and variance for population i are, respectively, $\mu_i \; and \; \sigma_i^2 \; (i = 1, 2, \ldots, t)$. The t variances are equal: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2 = \sigma^2$.

*****Since these conditions are mostly on the errors, we will use residuals to check the following:**

Errors are normally distributed

A graphical method to assess this is the Normal Q-Q plot, labeled as "C" on the next page. The guideline is that the points tend to follow a straight line. A histogram of residuals is also a graphical method that we can use. The shape should be symmetrical, unimodal, with no outliers. Since these methods are rather subjective, we will also use the Shapiro-Wilk test to determine if the errors are normally distributed. This is the same test we used for regression.

There are no outliers

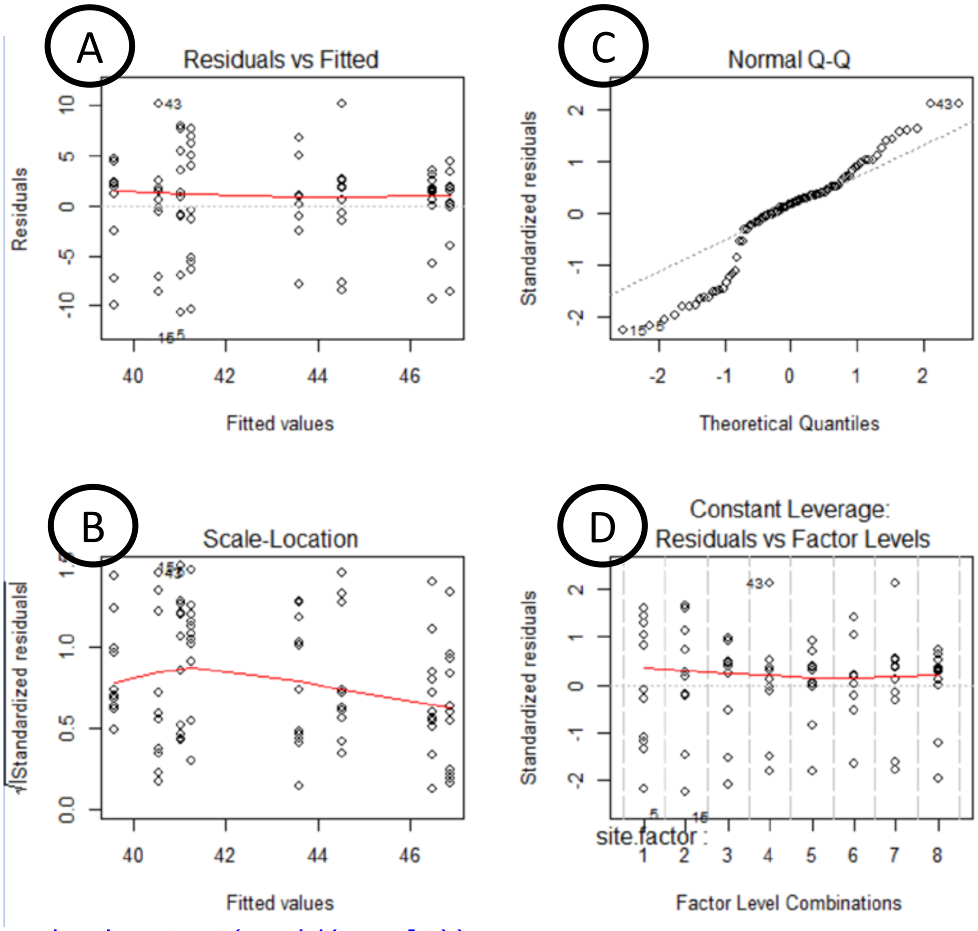
This can be assessed by looking at all the plots (A, B, C, D, histogram, boxplot) to see if any points are labeled or appear far from other points. We will consider these as potential outliers which require a thorough investigation before we can label them as outliers and remove them from the analysis.

There is a common variance

The plots labeled "A" and "B" can be used to assess a common variance. If there are any non-random patterns in the points, then this is an indication there is not a common variance. We will also use Levene's test to determine if the errors have a common variance. The null hypothesis is that there is a common variance and the alternative is that there is not a common variance. The test statistic is the F-value in the R output.

Practice:

Check each assumption for the Silt Content of Soils example:




```
> shapiro.test(resid(result))
        Shapiro-Wilk normality test
data:  resid(result)
W = 0.94572, p-value = 0.001069

> # Need R package car for Levene's Test for Homogeneity of Variances
> # Install R package car before running the following code
> library(car)
Loading required package: carData
> leveneTest(Silt~site.factor,data=dataobj)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  7  1.0702 0.3903
      80
```