Analysis of Variance
Part 4: Model Diagnostics

STAT 705: Regression and Analysis of Variance

The Basic Assumptions

- For both the cell means and the effects models, we assume that the errors
  - are independent
  - follow a normal distribution
  - have mean 0
  - have constant variance σ^2

Importance of the Assumptions

- All inference depends on the assumptions. This includes
  - p-values for all hypothesis tests
  - Levels of confidence for the confidence intervals
- We can never 'prove' the assumptions are satisfied
  - Instead we look for evidence they are violated
- Consequences of violating the assumptions
  - Incorrect p-values, resulting in incorrect conclusions
- Severity of the consequences
  - Vary from inconsequential to serious
  - Depends on which assumption is violated and the extent to which it is violated

How to Check Assumptions

- Follow the same steps we did with regression models
- Fit the model and get the residuals
- Generate diagnostics plots
  - Normal probability plot - - points should follow the line
  - Residual plot - - look for outliers and evidence of unequal variances; can also use the Brown-Forsythe test to test for equal variances

Note: For ANOVA models, the residual plot will appear 'stacked' like the scatterplot did. Unequal variances are evident when one (or more) vertical columns of points are much more widespread than the others.

Independence Assumption

- Violation of independence can ***severely*** affect conclusions
- To check this assumption, we must consider how the data were collected
- Intuitively, observations are statistically independent if one outcome does not affect another
- Observations from ***random*** samples are independent

Violations of Independence

Examples

- Repeated measures
  - Observations are taken on the *same* object over time
- Subsampling
  - The same object is measured more than once (to improve the accuracy of the measurement)
  - Groups of objects are treated as a whole rather than individually (e.g. a classroom of students all receive the same 'new' lesson plan, but test scores are measured on individual students)

Example Violation #1

Suppose that a medical researcher wishes to compare two medicines for reducing cholesterol. Each patient in the study has cholesterol measured every week. The observations over time on each patient would not be independent. For instance, a patient with higher than average cholesterol one week would likely be followed by a higher than average cholesterol level the next week. However, observations associated with one patient would be independent of observations associated with another because they are treated individually.

Example Violation #2

Suppose a researcher takes random samples of soil from a contaminated area. Each sample is subdivided into three parts and a measurement of a pesticide in the soil is made on all three parts. Because the parts are made on the same sample, the 3 parts would not be independent. These parts are called subsamples. For instance, a higher than normal reading on one subsample would likely be associated with a higher than normal reading on another. Typically, we average the subsample readings to obtain a single reading for each sample.

Example Violation #3

Suppose we are interested in how the oven temperature affects the quality of baked bread. We make 5 loaves and put them in the oven at the same time. The measurements on these loaves might not be independent. For instance, if the loaves are baked a bit longer than called for, this would affect all of the loaves in the same way, so the observations on the loaves within the oven would not be independent of one another.

Remedies for Lack of Independence

- Analyzing dependent data as if observations are independent can have **serious** consequences in terms of producing wrong p-values, etc.
- It is beyond the scope of the course to discuss methods that may be applied when the independence assumption is violated
- However, you should be aware of time-dependent data and subsampling so that you do not mistakenly apply the wrong methods to such data

Normality and/or Variance Violations

- Good news!
  - Modest deviations from normality or from the equal variance assumptions will have little effect on the p-values, etc.
- We say that ANOVA is **robust** to minor violations of these assumptions
- In some cases, extreme violations can be alleviated by transforming the response variable
  - Often, a logarithmic transformation is used

Formal Tests for Unequal Variances

- These tests are usually used only for ANOVA models (not regression), because they require replication
- There are a number of formal hypothesis tests for unequal variances, but many are very sensitive to violations of the normality assumptions
- One recommended test is Brown-Forsythe
  - It is a modification of a test developed by Levene
  - Compute the absolute value of the difference between each observation and the *median* of the treatment from which the observation came
  - Apply one-way ANOVA to these values.
  - A significant F test indicates unequal population variances

Generate Diagnostics Caffeine Data

```
symbol value=dot i=none color=purple;
proc glm data=caffeine plots=diagnostics;
class dose (ref='0');
model taps = dose / solution ;
lsmeans dose / stderr cl;
means dose / hovtest=bf;
run;
```

Generate diagnostic plots

The option `hovtest=bf` *generates the Brown-Forsythe test for equality of variances*

Brown and Forsythe's Test for Homogeneity of taps Variance
ANOVA of Absolute Deviations from Group Medians

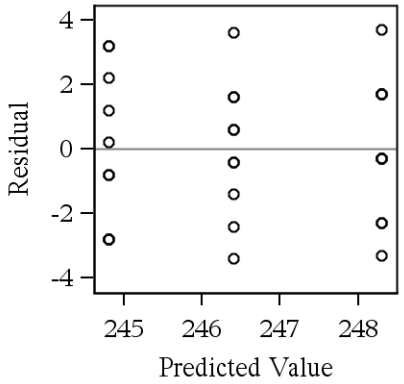
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
dose	2	0.8667	0.4333	0.28	0.7565
Error	27	41.5000	1.5370		

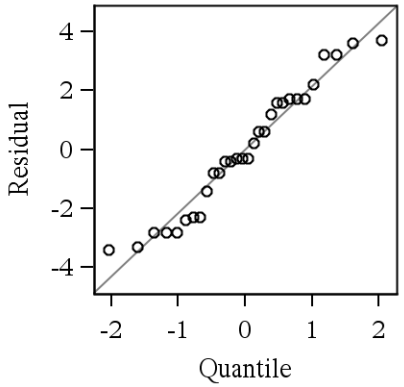
For the Brown-Forsythe test, the null hypothesis is that the variances are all equal.

We do not reject H_0 (p=0.7565).

We conclude the variances are equal.

Check Assumptions: Caffeine Data





These types of graphs should look familiar to you

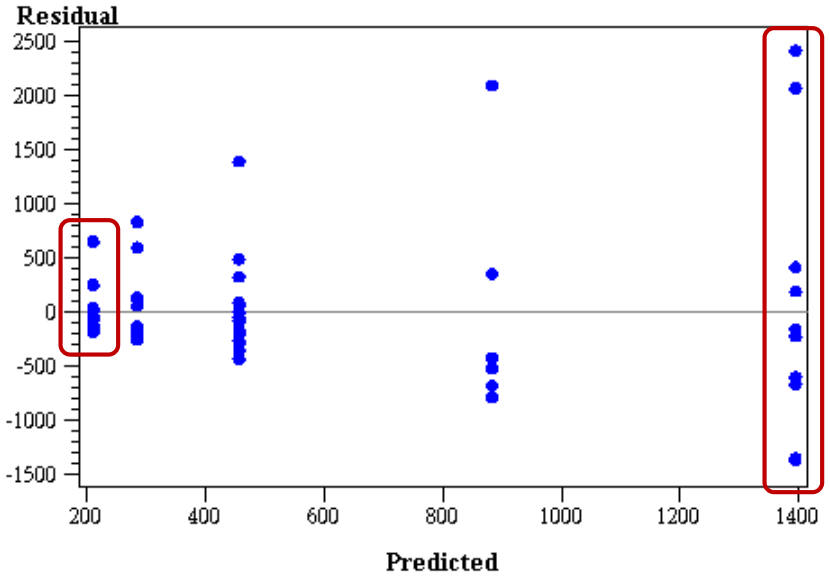
Top graph: Residual vs. fitted

- Vertical spread of the points is about the same for all three groups
- Also consider Brown-Forsythe test (previous slide, p=0.7565)
- No reason to suspect variances are unequal

Bottom graph: Normal probability plot

- No major departures from the line
- Nothing to indicate non-normality

A Different Dataset



Brown and Forsythe's Test for Homogeneity of Time Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Cancer	4	4778360	1194590	4.45	0.0033
Error	59	15829818	268302		

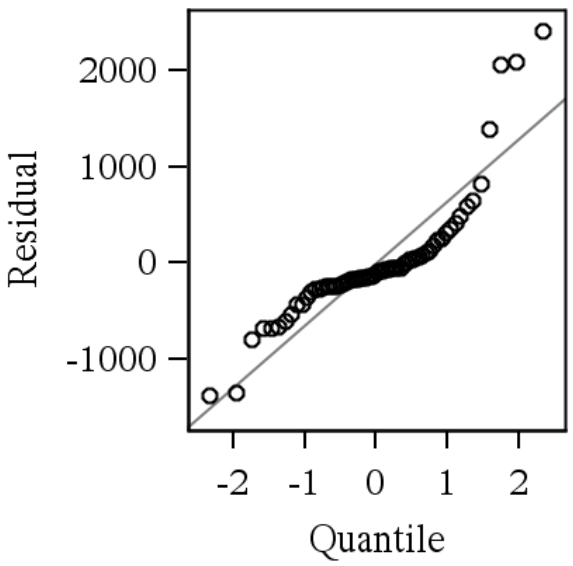
There is convincing evidence that these five groups DO have different variances.

The differences in spread are clear.

The Brown-Forsythe test has p-value 0.0033, which implies we should reject the null hypothesis that the variances are equal.

The normal probability plot is given on the next slide.

Normal Probability Plot



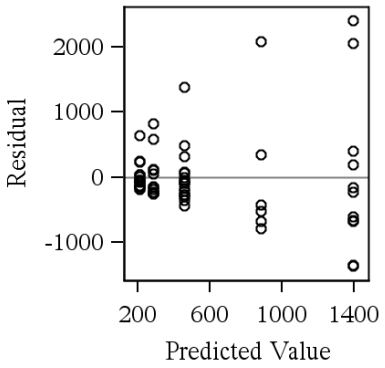
- This is for the same data as the previous slide
- There is a distinct curve, especially in the upper right
- Combine this information with the truly awful residual plot on the previous slide

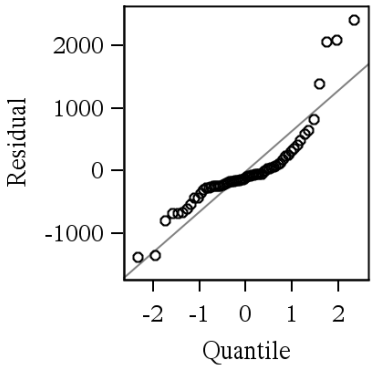
  ⇨ we need to do ***something*** to make these better

Variance-Stabilizing Transformation

(Use log(Y) instead of Y as the response variable)

Original Y





Brown-Forsythe Test

p = 0.0033

Equal variances is not likely

p = 0.6164

Equal variances seems satisfied

Normality

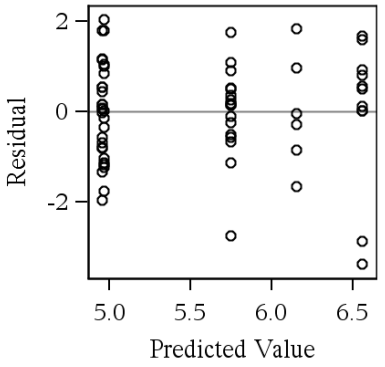
Distinct curve.

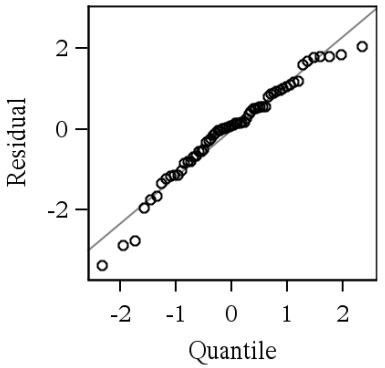
Normality may be violated

Much flatter.

Normality seems satisfied

log(Y)





Back-Transformations

- If you transform Y
  - Do ALL of the analysis using the transformed Y
  - When you are finished with the analysis BACK-TRANSFORM all estimates so the values will be in the original units
- Example: Suppose Y is measured in weeks
  - Then log Y is measured in log-weeks (and what's that?)
  - Use the exponential function to back-transform log Y to the original units
  - So if you generated a confidence interval (1.15, 2.13) for the mean of one treatment, these numbers are in log-weeks
  - The same estimate in weeks is $\left(e^{1.15}, e^{2.13}\right)$, or (3.16, 8.41) weeks

Other Transformations

- If the logarithmic transformation does not work, there are other transformations that might
- One possibility: Power transformations create a new Y by raising the old Y to some power
- The Box-Cox procedure can be used to choose an appropriate power
- This is beyond the scope of this course – but Google "Box-Cox" it if you need it!

What You Should Know

KNOW THE ASSUMPTIONS FOR AN ANOVA MODEL

- Understand the consequences for violating these assumptions
- Be able to generate and interpret diagnostic plots, and the Brown-Forsythe test
- Be able to apply a variance-stabilizing transformation (and back-transformation) when needed. (We did an example of this in Simple Linear Regression, Part 7.)