

Assumptions

Tim Hanson

Department of Statistics
University of South Carolina

January, 2017

Modified from originals by Gary W. Oehlert

ASSUME

Our models have assumptions. We assume that, e.g.

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

or, e.g.

$$y_{ij} = \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \epsilon_{ij}$$

where the ϵ_{ij} s are independent with distribution $N(0, \sigma^2)$.

If the model is correct, our inference is good and matches with randomization inference (if available).

Unfortunately, wishing doesn't make it so.

What can go wrong?

If the assumptions are not true, our inferences might not be valid, for example,

- A confidence interval might not cover with the stated error rate.
- A test with Type I error of \mathcal{E} could actually have a larger or smaller Type I error rate.
- Contrasts are estimating garbage.

This is obviously bad news and can be the source of controversy and disagreement over how the analysis was done and the validity of the results.

(But if you did a randomization, your randomization inference is still valid.)

Some procedures work reasonably well (e.g., actual interval coverage rate is near to nominal, or actual p-value is close to nominal p-value) even when some assumptions are violated.

This is called robustness of validity.

Generally these procedures work better when violations are mild and work less well as violations become more extreme.

A procedure that has robustness of validity can be inefficient, so we might not want to use it even if it is robust.

The basic assumptions are

- Independence (most important)
- Constant variance
- Normality (least important)

Many ways that data can fail to be independent; we will learn to check for one.

In this course we will not generally try to fix or accommodate dependence. We leave that for other courses: time series (STAT 520/720), multivariate analysis STAT 530/730, longitudinal STAT 771, etc. However, certain types of dependence will be useful later when we discuss blocked designs, repeated measures, etc.

Our assumptions are about the ϵ_{ij} , but we never get to see them.

What we do have are residuals.

The basic raw residual is

$$r_{ij} = y_{ij} - \text{fitted value}$$

In the separate means (oneway ANOVA) model they are

$$r_{ij} = y_{ij} - (\hat{\mu} + \hat{\alpha}_i) = y_{ij} - \bar{y}_{i\bullet}$$

Standardized residuals

The raw residual is useful for many purposes, and is often good enough in balanced designed experiments. However, we can do better.

The standardized residual (sometimes called internally Studentized) adjusts r_{ij} for its estimated standard deviation:

$$s_{ij} = \frac{r_{ij}}{\sqrt{MS_E(1 - H_{ij})}}$$

The H_{ij} value is called the leverage; it is a diagonal element of the “Hat” matrix, which is why we call it H .

Roughly speaking, the s_{ij} should look like standard normals, particularly in large samples.

Externally Studentized or “deleted” residuals

One further step is the Studentized residual (or the externally Studentized residual if you like calling standardized by internally Studentized):

$$t_{ij} = s_{ij} \sqrt{\frac{\nu - 1}{\nu - s_{ij}^2}}$$

where ν is the df in the MS_E .

If model is correct, t_{ij} follows a t distribution with $\nu - 1$ df. A t with reasonable df will look pretty much like a normal.

Formal rule for detecting outliers

Studentized residuals are especially good in looking for outliers.

A formal rule using Bonferroni for classifying an observation as “outlying” is to compare $|t_{ij}|$ to a t-distribution cutoff of $1 - \frac{\mathcal{E}}{2N}$ with freedom degrees ν ; usually $\mathcal{E} = .05$. I think this goes overboard, but some people do this.

Studentized residuals say how well the data value fits the model estimated from the rest of the data.

Formal tests for normality, constant variance, etc. exist, BUT...

- With small sample sizes, you'll never be able to reject the null that there are no problems.
- With large sample sizes, you'll constantly detect little problems that have no practical effect.

It's really all shades of gray (at least 50), and we would like to know where we are on the scale from mild issues to severe issues.

Assess assumptions qualitatively; plots help a lot here.

Our principal tools for assessing assumptions are various plots of residuals:

- Normal probability plot
- Residuals versus predicted plot
- Residuals in time order

The first two are the basic plots for assessing normality and constant variance; the last one is just one of many potential plots for assessing independence.

Normal probability plot

The NPP plots the residual against its corresponding normal score. The smallest residual plots against the smallest normal score for a sample of N ; the second smallest residual against the second smallest normal score, and so on.

Normal scores depend on N . Think about an independent sample of N standard normals. They all have mean 0, but if you just consider the smallest one, it has a negative expectation. That expectation is its normal score.

The *rankit* approximates the normal score:

$$\text{rankit}_{i,N} = \Phi^{-1} \left(\frac{i - 3/8}{n + 1/4} \right)$$

where Φ^{-1} gives normal percent points.

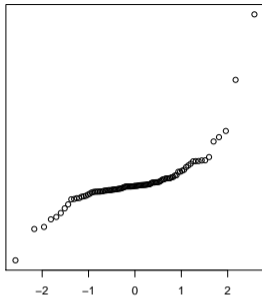
Normal probability plot

It's probably best to use the Studentized residuals, but the others also work fine in most situations.

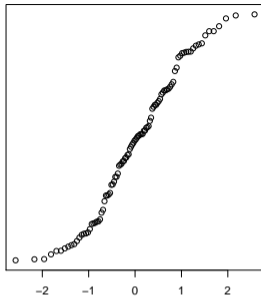
Normally distributed data (and, we hope, residuals from iid normally distributed errors) should have a roughly linear shape, although even normal data can look crooked in small samples.

You can tell the shape of the data from the shape of the plot, but you need to practice.

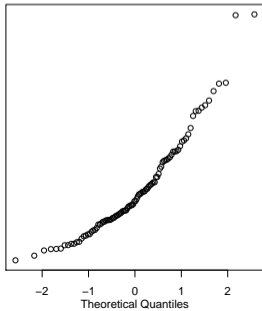
long tails



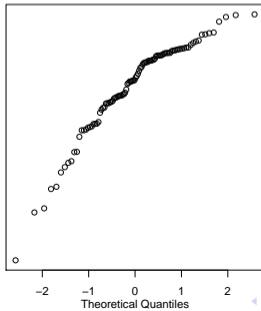
short tails



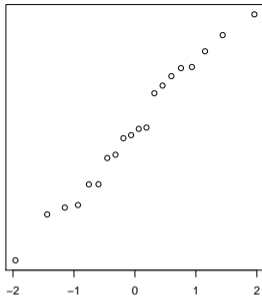
skewed right



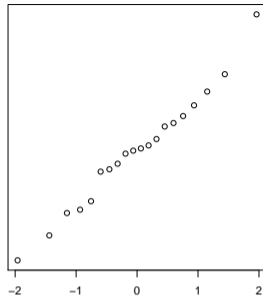
skewed left



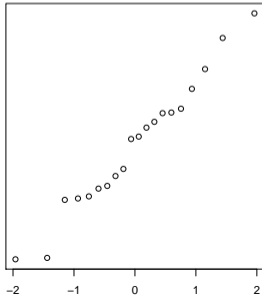
iid Normal



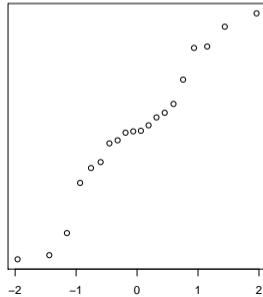
iid Normal



iid Normal



iid Normal



R example: resin lifetimes

```
resin=read.table("http://users.stat.umn.edu/~gary/book/fcdae.data/exmpl3.2",header=T)
resin # data from website slightly different than in library 'oehlert'
colnames(resin)=c("temp","logTime") # column names now match
resin[,1]=c(rep(175,8),rep(194,8),rep(213,8),rep(231,7),rep(250,6))
resin$Time=10^resin$logTime
resin$ftemp=factor(resin$temp)

f=lm(Time~ftemp,data=resin)
residuals(f) # raw
rstandard(f) # standardized
rstudent(f) # studentized
qqnorm(rstudent(f)) # normal probability plot
par(mfrow=c(2,2))
plot(f)

shapiro.test(rstudent(f)) # formal normality test of H0: residuals are normal
```

Formal detection of “outliers” ...

```
abs(rstudent(f))>qt(1-0.025/nrow(resin),f$df.residual)  
which(abs(rstudent(f))>qt(1-0.025/nrow(resin),f$df.residual))
```

Residual vs. predicted

The diagnostic plot for non-constant variance is to plot each residual against its corresponding predicted/fitted value.

We are hoping to see no pattern in the vertical dispersion.

The most common problem occurs when larger means go with larger variances. In this case we see a “right opening megaphone.”

We sometimes see the reverse, particularly when there is an upper bound on the response.

There are several variations on this, including box plots of residuals and plots of square root absolute residuals against fitted values.

R example: resin lifetimes

```
par(mfrow=c(1,1))
plot(f$fitted.values,rstandard(f),main="Std. residuals vs. predicted")
plot(f$fitted.values,sqrt(abs(rstandard(f))))
plot(rstandard(f)~resin$temp)
par(mfrow=c(2,2))
plot(f)
```

```
par(mfrow=c(1,1))
library(car)
residualPlots(f)
```

What do you think of non-constant variance?

Formal test for constant variance

If you must test for equality of variances, your best bet is Levene's test. This makes a new response as the absolute value of the deviations of the original data from the predicted value, and then does an ANOVA test for the separate means model on the absolute deviations.

There are several variations on this where you might take absolute deviations from the median of each group, or the absolute deviations to some power, etc.

There are several classical tests of equality of variance including Barlett's test and Hartley's test; they are highly sensitive to normality.

R example: resin

```
bartlett.test(rstandard(f)~resin$temp) # assumes normality within groups  
leveneTest(f) # does not assume normality but has lower power
```

There are many ways that data could fail to be independent, but we will only talk about the simplest of these: temporal dependence.

In some data sets, but not all data sets, there is a time order of some kind.

One common failure of independence is when data close in time tend to have similar ϵ_{ij} s and thus similar residuals. This is called positive temporal dependence or positive serial correlation.

The reverse can also happen (near in time tend to be unusually far apart), but it is much more rare.

The simplest diagnostic is to plot the residuals in time order and look for patterns.

Do the residuals seem to be high and low together in patches? That is positive serial correlation.

Do the residuals seem to bounce up and down very roughly and alternately? That could be negative serial correlation.

The stronger the pattern, the stronger the correlation and the greater the problem it will cause with inference.

Testing for independence

There are a couple of simple tests for serial correlation. Let r_i be one of the kinds of residuals sorted into time order.

The Durbin-Watson statistic is

$$DW = \frac{\sum_{i=1}^{n-1} (r_i - r_{i+1})^2}{\sum_{i=1}^n r_i^2}$$

Independent data tend to have DW around 2; positive correlation makes DW smaller; negative correlation makes DW bigger.

If DW gets as low as 1.5 or as high as 2.5, it's definitely time to start worrying about what is happening to the inference.

Testing for independence

There are also a whole variety of “runs” tests, variously defined. These look for things like runs of residuals that are positive (or negative), or runs of data that are increasing (or decreasing).

In any event, there are several runs tests, but they can also be used to assess temporal correlation.

Only assess temporal correlation if your data have a time order!

R example: thermocouples (Ex. 6.3, p. 121)

Christensen and Blackwood (1993) give data on five thermocouples inserted into a high-temperature furnace to ascertain their relative bias. $N = 64$ temperature readings were taken using each thermocouple, with the readings taken simultaneously from the five devices. We examine the paired $d_i = x_i - y_i$ differences between 3 and 5. The relative bias between 3 and 5 is estimated by the mean difference μ , i.e. $d_i = \mu + \epsilon_i$.

```
oven=read.table("http://users.stat.umn.edu/~gary/book/fcdae.data/exmpl6.3",header=T)
f=lm(tempdiff~1,data=oven)
plot(residuals(f))
durbinWatsonTest(f)

library(tseries)
runs.test(factor(sign(residuals(f))))
```

Independence reasonable?

Accommodating problems

There are two basic approaches to dealing with things when assumptions are not met:

- Alternate methods
- Massaging the data

Developing alternate methods is basically fulltime employment for academic statisticians.

Let's look at a few broad areas, but only talk about a couple alternatives.

Robustness is a philosophy and class of techniques that deal with long-tailed, outlier prone data.

Generalized Linear Models (GLM) is a class of techniques for using models with linear predictors but which have non-normal data including count data and various kinds of non-constant variance.

Time series is a class of statistical models for working with serial correlation (among other things).

Spatial statistics includes, among other things, the ability to fit linear models when the data are correlated in space.

Direct replacements are usually developed to solve specific narrow issues without building a whole new class of statistical models.

Two-sample direct replacement for non-constant variance

Many of you are familiar with the version of the t-test that does not use a pooled estimate of variance. Instead, it uses

$$t = \frac{\bar{y}_{i\bullet} - \bar{y}_{j\bullet}}{\sqrt{\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}}}$$

where s_i^2 and s_j^2 are the sample variances in two groups. There is a formula for approximate df, and then you compare with a t-distribution.

This is the direct replacement for ANOVA when $g = 2$ and there is non-constant variance.

In this case, the replacement is so easy and works so well that there is little reason not to use it all the time.

Non-constant variance for oneway ANOVA

The Brown-Forsythe method generalizes this to $g > 2$ groups, but even this simple problem is getting a bit messy. Let

$$d_i = s_i^2(1 - n_i/N)$$

Then the Brown Forsythe F is

$$BF = \frac{SS_{Trt}}{\sum_{i=1}^g d_i}$$

Treat this as F with $g - 1$ and ν df where

$$\nu = \frac{\sum_{i=1}^g d_i^2}{\sum_{i=1}^g d_i^2 / (n_i - 1)}$$

R example: resin

Recall that the resin data had non-constant variance; the sample variance increased with increasing lifetimes when temperatures were lower.

```
library(cfcdae) # has Brown-Forsythe test
boxplot(Time~ftemp,data=resin)
brown.forsythe.test(Time~ftemp,data=resin)
oneway.test(Time~ftemp,data=resin) # generalizes Welch
source("http://people.stat.sc.edu/hansont/stat506/gh.R")
gh(resin$Time,resin$ftemp) # assumes normality but not constant variance
library(coin) # one type of permutation test
independence_test(Time~ftemp,data=resin)
kruskal.test(logTime~temp) # Kruskal-Wallis nonparametric one-way ANOVA
library(rcompanion) # methods are holm, fdr (Benjamini & Hochberg), etc.
?pairwisePermutationTest # see options
pairwisePermutationTest(Time~ftemp,method="holm",data=resin)
```

Both Brown-Forsythe and Welch assume normality but allow non-constant variance. One approach that assumes nothing is a permutation test (as in Chapter 2) found in the `coin` package for R. This package is in fact very powerful and can be applied to multivariate responses, categorical responses, etc.

This sounds like iniquity, but it's really not that bad.

The simplest form of this practice is removing outliers and reanalyzing the data. Ideally, we would like to get the same basic inference with and without the outliers.

If the inference changes substantially, this means that it is dependent on just a handful of the data.

You can't automatically reject a data value simply because it does not fit the model you assume.

Transforming the response

Our go-to approach is usually to transform the data, that is, to re-express the data on another scale. Thus we might use

- pH instead of hydrogen ion concentration (log transformation);
- diameter of a bacterial colony rather than area (square root transformation);
- time to distance instead of rate of advance (reciprocal transformation).

In general, any monotone transformation will work, but we concentrate on power family transformations.

Power family transformations work for positive data. If you have some zeroes or negatives, you must first add a constant to all data.

So

$$y_{ij} \rightarrow y_{ij}^{\lambda}$$

Use a log transformation instead where $\lambda = 0$ would go.

A lower power tends to reduce right-skewness and reduce increasing variance.

A higher power tends to reduce left-skewness and reduce decreasing variance.

Note: if the data only range over a factor of 2 or 3, then power transformations are of limited utility. As the ratio of largest to smallest increases, power transformations can have more effect.

Serendipity. More often than we have any right to expect, transformations that make variance more constant also improve normality.

The Box-Cox procedure helps us

- Pick a reasonable range of transformation powers
- Decide whether we need a transformation: $\lambda = 1$ gives us the original response.

Try to pick a transformation power that both fixes the problems and is also interpretable.

In R, Box-Cox gives us a likelihood profile for λ as well as a 95% confidence interval.

```
library(MASS)
boxcox(Time~ftemp,data=resin)
```

Closest value is 0; so try log-transformation. I usually use natural log, not \log_{10} .

Note that this is how the data were originally given to us, log-transformed. Now we know how someone decided on this transformation!

Simpson, Olsen, and Eden (1975) provide data giving the rainfall in acre feet of 52 clouds, 26 of which were chosen at random for seeding with silver oxide. The problem is to determine if seeding has an effect and what size the effect is (if present). `trt=1` indicates not seed and `trt=2` is seeded.

```
cloud=read.table("http://users.stat.umn.edu/~gary/book/fcdae.data/exmpl6.1",header=T)
boxplot(y~trt)
library(lattice)
histogram(~y|trt)
boxplot(log(y)~trt)
```

Let's keep going...

If the null is that distributions for different treatments are the same on one scale, they will also be the same on some other scale.

We might as well use the one where our assumptions are plausible.

We can test equality of means on any scale and get proper inference.

That's the good news . . .

$$g\{E(Y)\} \neq E\{g(Y)\}$$

The bad news shows up when you want to make inference on means across scales.

Means do not transform cleanly across power transformations.

That is, you cannot exponentiate the mean of the log data to get the mean of the natural scale data.

A transformed CI for the mean of normal data is a CI for the median on the transformed scale, *not* for the mean.

Land's method helps in the specific case of logs and anti-logs, but in general you either make due with medians or work on the original scale and take your lumps on the quality of inference.

So how bad is this, really?

Skewness measures how asymmetric a distribution is.

Kurtosis measures how long-tailed (outlier prone) a distribution is.

The normal has both 0 skewness and 0 kurtosis.

Absent outliers, F-test is only slightly affected by non-normality.

F-test has reasonable robustness of validity, but it is not resistant; individual outliers can change test results.

Often check to see if inference is consistent with and without outliers.

For balanced data (all sample sizes equal),

- Skewness has little effect
- Long tails (positive kurtosis) leads to conservative tests. These tests have nominal p-values larger than they really should be, so fewer rejections than we should have.
- Short tails (negative kurtosis) leads to liberal tests. These tests have nominal p-values smaller than they really should be, so more rejections than we should have.

See book for some numerical results.

Inconsistent results for unbalanced data.

Also see discussion on how skew affects CI's, non-constant variance, temporal dependence, etc.

- Always look at your data first.
- There are nonparametric, e.g. permutation-based, ways to test $H_0 : \mu_1 = \dots = \mu_g$ and follow up w/ pairwise comparisons. These work on the original untransformed data, do not require constant variance, etc.
- A nonparametric pairwise procedure is due to Dwass, Steel, Critchlow, and Fligner. Could not get associated R package to work. Available in SAS though. Permutation-based approach seemed to work.
- If your data are normal, you may get a bit more power using traditional F-tests, pairwise comparisons based on the studentized range etc. Check residuals for constant variance, normality, etc. via plots.
- If your residuals/data are highly non-normal, try a power transformation via the Box-Cox procedure.
- If your data are normal but with non-constant variance, there are the Brown-Forsythe and Welch tests of $H_0 : \mu_1 = \dots = \mu_g$. You can follow up with the Games and Howell pairwise procedure.