Displays for Statistics 5303

Lecture 12

September 30, 2002

Christopher Bingham, Instructor

612-625-7023 (St. Paul)
612-625-1024 (Minneapolis)

Class Web Page
http://www.stat.umn.edu/~kb/classes/5303

Checking Assumptions

Checking assumptions is always in the context of some model.

The one-way ANOVA model for the CRD design is

$$y_{ij} = \mu^* + \alpha_i + \varepsilon_{ij},\ i = 1,\ldots,g, j = 1,\ldots,n_i$$

The **multiple regression model** is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i,$$
$$i = 1,\ldots,n_i;\ x_1,\ldots,x_k \text{ predictor variables}$$

The ε's - disturbances or errors - always have zero mean, that is $\mu_\varepsilon = 0$.

Both these models are of the form
y = predictable part + unpredictable part

The fact that the parts are *added* together rather than, say, multiplied, is an important feature of both models.

In both cases the predictable part is itself a *sum* of various terms.

These models have 3 assumptions about the ε's in common

- All ε's are *independent*
  For ANOVA model, this implies
  1. Different observations in the same group are independent
  2. Data from different groups are independent
- The variances of all ε's are all σ^2
  For ANOVA model, this implies that each group has the same variance
- The ε's are normally distributed

You can combine all these assumptions in one statement:
ε's are a random sample from $N(0,\sigma^2)$

The three assumptions are listed in *decreasing* order of importance.
Independence (most important)
Constant σ (next most important)
Normal distribution (least important)

Vocabulary:

Homoskedastic errors all have the same variances. This is the condition of **homoskedasticity**.

Heteroskedastic errors do *not* all have the same variances. This is the condition of **heteroskedasticity.**

In ANOVA **heteroskedasticity** means σ differs between groups, often depending on the value of μ

In ANOVA **heteroskedasticity** means σ depends on the values of the predictors. An important case is when σ depends on the mean $\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}$.

The condition that $\mu_\varepsilon = 0$ essentially means that your model for the mean of y is correct.

- In regression, dependence of y on the x's is linear and that you haven't left out any important x's.
- In ANOVA situation, you haven't left out any factors such as time of day that might affect mean of y.

The possibility of such unknown factors is one of the reasons randomization is important. If you have randomized well, on the average unknown factors have no systematic effect, although they can increase the variability.

You need the assumptions to be true, or at least true "enough", so that your statistical methods will "**work as advertised**".

- Confidence intervals have the intended coverage
- Significance tests have the intended type I error rate ε, whether comparisonwise, experimentwise, strong experimentwise, false discovery rate, … .

When sample sizes are moderately large, ANOVA methods based on means such as the F test work quite well even when the data are not normal. That is, they are **robust** against non-normality.

Inference about variances tends to be very non-robust against non-normality.

When sample sizes are close to equal, the F-test is fairly robust against heteroskedasticity, but standard errors are off.

Probably all these assumptions are never <u>exactly</u> true.

With any assumption, there are at least two issues:

- How to diagnose that the data do not satisfy the assumption (a violation)
- What to do when you find a violation

I discussed ion Lecture 7 (September 18) some ways to **diagnose lack of independence** when you collect data sequentially in time. I did not give any remedy, since that would lead us too far in the direction of time series analysis.

Proper randomization is the best protection against lack of independence, since the randomization itself induces independence.

Generally, even with dependent (not independent) errors estimates of means and regression coefficients are unbiased, so there is no systematic error.

However, the estimate of σ^2 can be very biased and hence standard errors, t-statistics and confidence intervals computed the usual way can be misleading.

I did a simulation with $g = 5$ treatments, each with $n = 4$ observations. For each serial correlation values -.8, -.6, .-4, -.2, 0, .2, .4, .6 and .8 I did an ANOVA with simulated 2500 sets of normal data with $\sigma = 1$, but with correlated errors.

-0.8	-0.6	-0.4	-0.2	0	0.2	0.4	0.6	0.8
1.18	1.16	1.12	1.06	1.01	0.90	0.77	0.59	0.35

Row 1: Serial correlation
Row 2: Average MSE
Average MSE $\neq$ 1 indicates bias.
Positive serial correlation $\Rightarrow$ serious underestimation of σ

Essentially all methods for diagnosing violation of assumptions are based on study of the observed residuals

$$r_{ij} \equiv y_{ij} - \hat{\mu}_i = r_{ij} - \overline{y_{i\bullet}} = r_{ij} - \hat{\mu}^* - \hat{\alpha}_i$$

This is because you **can't observe** the **true** residuals $\varepsilon_{ij} = y_{ij} - \mu_i = y_{ij} - \mu^* - \alpha_i$.

You should check residuals as a standard part of every analysis of a designed experiment.

Comment:
Even when the $\{\varepsilon_{ij}\}$ are independent, the $\{r_{ij}\}$ are not. For one thing, in each group $\sum_{1\leq j\leq n_i} r_{ij} = 0$, so $r_{in_i} = -\sum_{1\leq j\leq n_i-1} r_{ij}$.

In fact, correlation of two residuals in the same group is $-1/n_i$.
And, even when σ_ε is constant, $\sigma^2_{r_{ij}}$ may not be constant. In the ANOVA case, $V[r_{ij}] = ((n_i-1)/n_i)\sigma^2 = (1 - 1/n_i)\sigma^2 < \sigma^2$

- When you find non-normal errors, you often also find heteroskedastic errors.
- When you find heteroskedastic errors, you often also find non-normal errors.

Although this often happens, there are lots of exceptions. Because heteroskedasticity is more important than non-normality, it should have take priority in seeking a remedy.

The principal remedial tool available is re-expression of the response, that is analyzing some **transformation** of the response instead of the response itself.

Common transformations are $\log(y)$, $\sqrt{y}$, $y^{1/3}$, $1/\sqrt{y}$ and $1/y$.

Because $y^{-p} = 1/y^p$ reverses order (if $y_1 > y_2$, then $1/y_1^p < 1/y_2^p$, $p > 0$), Oehlert suggests using $-y^{-p}$ which preserves order. I don't see the advantage.

Remark:

$\log_{10}(y) = \log_e(y)/\log_e(10) = \log_e(y)/2.3026$

$\log_e(y) = \log_e(10)\times\log_{10}(y) = 2.3026\times\log_{10}(y)$

That is, they differ by a multiplicative constant and hence serve equally well (or badly) to correct non-normality and/or non-constant σ

This is a reflection of the following fact:

If you have two transformations

$$\tilde{y}_1 = f_1(y) \text{ and } \tilde{y}_2 = f_2(y)$$

such that

$$\tilde{y}_2 = (\tilde{y}_1 - a)/b$$

then they are completely equivalent in terms of their use to cope with violations of assumptions.

The **Box-Cox power family** of transformations for a positive response variable is

$$\tilde{y} = (y^p - 1)/p,\quad p \neq 0$$
$$\tilde{y} = \log(y),\ p = 0$$

Clearly, when $p \neq 0$, the Box-Cox transformation, is equivalent to the power transformation $\tilde{y} = y^p$, which includes $\sqrt{y}$ ($p = 1/2$) and $1/y$ ($p = -1$).

Oehlert uses a slightly different definition which matches what MacAnova macro **boxcox()** computes.

$$y \rightarrow \{(y^p - 1)/p\}/GM^{p-1},\ p \neq 0$$
$$y \rightarrow GM\times\log(y)\qquad,\ p = 0$$

where $GM = e^{\sum_i \log(y_i)/n}$ is the geometric mean. Since this is a multiple of the first definition, it is equivalent to the first definition and to y^p or $\log(y)$.

Here is a very small computation to demonstrate that for p near 0, $(y^p-1)/p$ is very close to log(y):

```
Cmd> y # short vector of positive data
(1)     0.35376     0.46584      2.1432       11.08      1.8897

Cmd> p <- .0001; hconcat(log(y),(y^p - 1)/p)
(1,1)     -1.0391     -1.0391
(2,1)    -0.76392    -0.76389
(3,1)     0.76229     0.76232
(4,1)      2.4052      2.4055
(5,1)     0.63643     0.63645
```

hconcat() binds its arguments side by side to form a matrix or table.

Here's a comparison of the simple form $(y^p - 1)/p$ and the form involving GM.

```
Cmd> p <- .5

Cmd> (y^p - 1)/p # Simple form of Box-Cox transformation
(1)    -0.81045    -0.63495     0.92793      4.6573     0.74933

Cmd> GM <- exp(sum(log(y))/5);GM # geometric mean
(1)      1.4921

Cmd> (y^p - 1)/(p*GM^(p-1)) # as defined by Oehlert
(1)    -0.98996    -0.77559      1.1335      5.6889      0.9153

Cmd> boxcox(y,p) #as computed by boxcox()
(1)    -0.98996    -0.77559      1.1335      5.6889      0.9153
```


Because residuals may have different variances, it is common to standardize them in some way.

$V[r_{ij}] = (1 - 1/n_i)\sigma^2 = (1 - H_{ij})\sigma^2$, $H_{ij} = 1/n_i$

The quantity H_{ij} is called the *leverage*.

anova() and **regress()** always compute a vector **HII**, the same length as y, which contains the leverages for each case.

Since σ^2 is estimate by MS_E, the *internally standardized* residuals are

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

These all have the same variance, which is approximately, but slightly < 1

They are called *internally* Studentized, since MS_E the estimate of variance includes a contribution from r_{ij}. If, say, r_{ij} is an outlier, it inflates MS_E, The externally studentized residuals are

$$t_{ij} = \sqrt{\{(df_{error} - 1)s_{ij} \times / \sqrt{(df_{error} - s_{ij}^2)}$$

These have the property that when all the assumptions are satisfied, t_{ij} has a t-distribution on d_{ferror} - 1 d.f.

They are called *externally* studentized since it can be shown that

$$t_{ij} = (y_{ij} - \hat{\mu}_i^{(-j)})/\sqrt{\{(1 - H_{ij})MS_E^{(-ij)}\}}$$

where $\hat{\mu}_i^{(-j)}$ is the mean of the responses in group j, *omitting* case i and $MS_E^{(-ij)}$ is the MS_E in an ANOVA of the data omitting y_{ij}. Thus the estimated standard error in the denominator is computed by data that is "external" to y_{ij}.

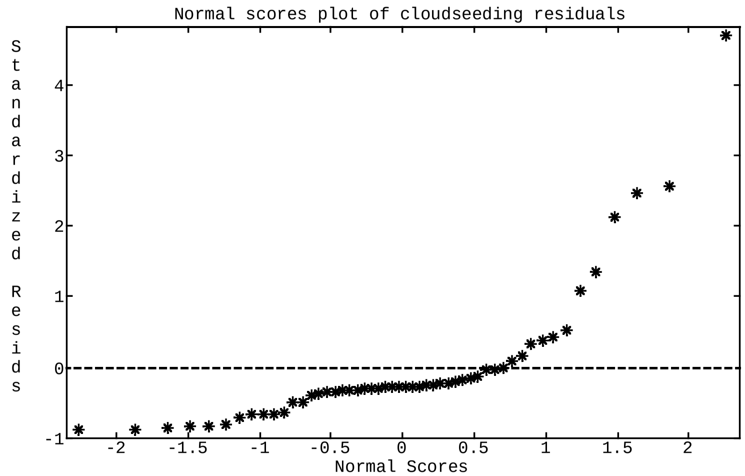
Statistics 5303 Lecture 12 September 30, 2002

Cloud seeding example from the text.

```
Cmd> data <- read("","exmpl6.1",quiet:T)
Read from file "TP1:Stat5303:Data:OeCh06.dat"

Cmd> treat <- factor(vector(data[,1]))

Cmd> rainfall <- vector(data[,2])

Cmd> list(treat,rainfall)
rainfall        REAL   52
treat           REAL   52    FACTOR with 2 levels

Cmd> n <- tabs(rainfall,treat,count:T);n
(1)         26          26

Cmd> anova("rainfall = treat")
Model used is rainfall=treat
                DF          SS           MS
CONSTANT         1  4.7831e+06  4.7831e+06
treat            1  1.0003e+06  1.0003e+06
ERROR1          50  1.2526e+07  2.5052e+05

Cmd> list(HII) # HII was computed by anova()
HII             REAL   52

Cmd> unique(HII) # all the values are the same
(1)    0.038462

Cmd> 1/n # In 1-way ANOVA, HII = 1/n[i]
(1)    0.038462    0.038462
```

You can check normality of residuals using a normal scores or rankit plot. When the residuals are normal, this plot should be close to linear.

```
Cmd> resvsrankits(title:"Normal scores plot of cloudseeding
residuals")
```



Curved in an asymmetrical way, indicating a skewed distribution of residuals.

resid() computes both s_{ij} (column 2) and t_{ij} for each residuals, plus other quantities.


```
Cmd> stuff <- resid() # must follow anova()

Cmd> stuff[run(5),] # first 5 rows
          Depvar  StdResids        HII   Cook's D    t-stats
(1)       1202.6     2.1149   0.038462   0.089458     2.1941
(2)        830.1      1.356   0.038462   0.036773     1.3677
(3)        372.4    0.42341   0.038462  0.0035855    0.41991
(4)        345.5     0.3686   0.038462  0.0027174     0.3654
(5)        321.2    0.31909   0.038462  0.0020364    0.31621
```

Column 1 is y_{ij}, followed by s_{ij}, H_{ij}, D_{ij} and t_{ij}. D_{ij} is Cook's distance, a measure how much influence the case had on the parameter estimates. It can be increased by large leverage (H_{ij}) or large t_{ij}.

```
Cmd> J <- grade(abs(stuff[,5]),down:T)
```

J now contains the case numbers of the data rearranged in order of decreasing $|t_{ij}|$.

```
Cmd>  stuff[J[run(10)],]
           Depvar   StdResids          HII    Cook's D      t-stats
(27)       2745.6      4.6936     0.038462     0.44059       6.2123
(28)       1697.8      2.5587     0.038462     0.13094       2.7171
(29)         1656      2.4735     0.038462     0.12237       2.6138
(1)        1202.6      2.1149     0.038462    0.089458       2.1941
(2)         830.1       1.356     0.038462    0.036773       1.3677
(30)          978      1.0921     0.038462    0.023855       1.0943
(52)          4.1    -0.89218     0.038462     0.01592     -0.89033
(51)          7.7    -0.88485     0.038462    0.015659     -0.88289
(50)         17.5    -0.86488     0.038462     0.01496     -0.86266
(49)         31.4    -0.83656     0.038462    0.013997     -0.83401
```

These are the rows associated with the residuals with the largest $|t_{ij}|$. The first is large and might be an outlier. You can test it by comparing it with $t_{1-(\alpha/n)/2,n-g-1}$, a Bonferronized cut point.

You Bonferronize by n because there are potentially n values of t_{ij} to test.

```
Cmd> invstu(1 - .025/52,DF[3]-1)
(1)      3.5135
```

$\max(|t_{ij}|) = 6.21 > 3.51$ confirms that case 27 may be an outlier. You could delete it, and refit, and test the new residuals.