

Week 7.1--IES 612-STA 4-573-STA 4-576.doc

IES 612/STA 4-576 Winter 2009

ANOVA MODELS - model adequacy aka RESIDUAL ANALYSIS

Numeric data - samples from "t" populations obtained

Assume $Y_{ij} \sim$ independent $N(\mu_i, \sigma^2)$ or $Y_{ij} = \mu_i + \varepsilon_{ij}$ with $\varepsilon_{ij} \sim$ independent $N(0, \sigma^2)$

$i = 1, 2, \dots, t$ (populations or treatments) and $j = 1, 2, \dots, n_i$ (observations)

Residual definition: $e_{ij} = Y_{ij} - \bar{Y}_i$. same as before residual = _____ - _____

Assumption	Checking?	Addressing?
1. Constant variance?	Plot e_{ij} vs. sample means and look for a pattern	- Transformation (e.g. log, sqrt) - Weighted Least Squares
2. Normal responses?	- Normal probability plot (normal scores vs. residual quantiles) - Histogram? Boxplot? Stemplot? - 68% of standardized residuals with -1 and +1 (95% within -2 and +2) - Shapiro-Wilk Test of Normality	- Transformation (sqrt - count responses, arcsin-sqrt - proportions, log - right skewed responses) - GLiMs (generalized linear models - binomial / Poisson responses as example)
3. Independent?	Plot residuals vs. order of observations? Often implicit part of the design	Analysis that reflects dependence? Time series/spatial data/mixed models
4. Outliers?	Large standardized residuals? ± 3 rule	- Check? - Run analysis with and without points? - Rank-based methods?

In Regression, we were concerned about the "correct form" of the regression model. Why are we **NOT** concerned about this in ANOVA?

What happens if assumptions NOT satisfied?

Assumption	Impact?
1. Constant variance?	MSE (pooled variance estimate) may be incorrect
2. Normal responses?	Stat tests may not be correct

3. Independent?	MOST important - $se(\bar{y})$'s too small
4. Outliers?	MSE inflated - CI's wider/harder to detect diffs

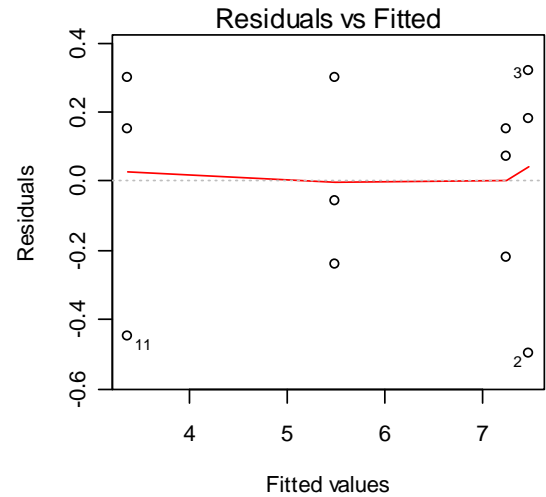
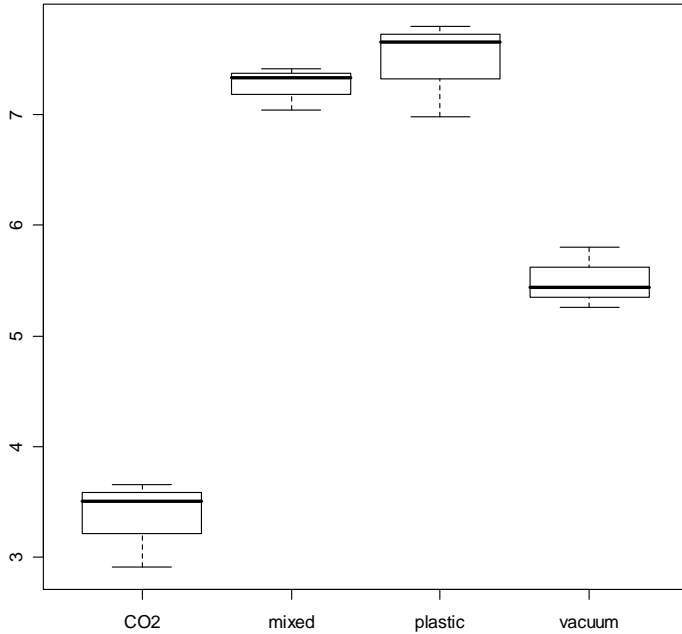
Example Bacteria growth in meat under different packaging conditions (again!)

Model: $\text{Log}(\text{Bacterial Growth})_{ij} = \mu_i + \varepsilon_{ij}$, where ε_{ij} are independent $N(0, \sigma^2)$.

Using R

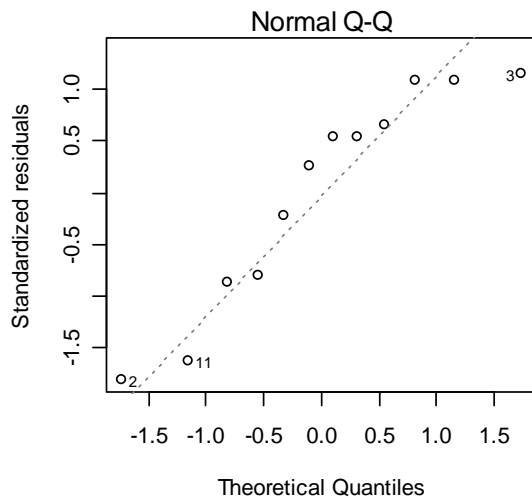
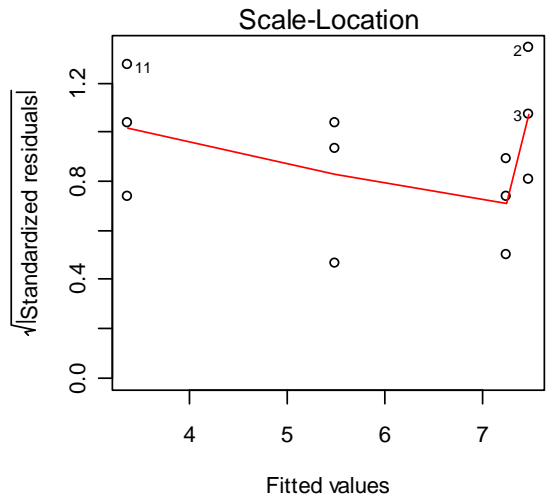
1. Constant Variance?

```
> boxplot(logCount ~ Condition)
```



2. Normality?

```
> plot(MeatANOVA)
> shapiro.test(MeatANOVA$residuals)
Shapiro-Wilk normality test
data: MeatANOVA$residuals
W = 0.8942, p-value = 0.1336
```



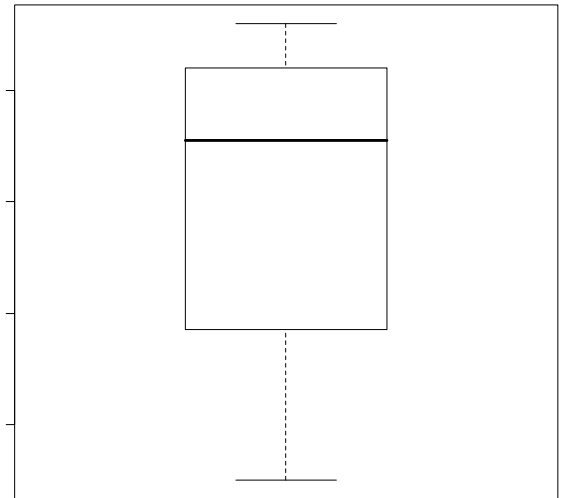
3. Independence?

Assume that correct sampling methods were used to insure independence.

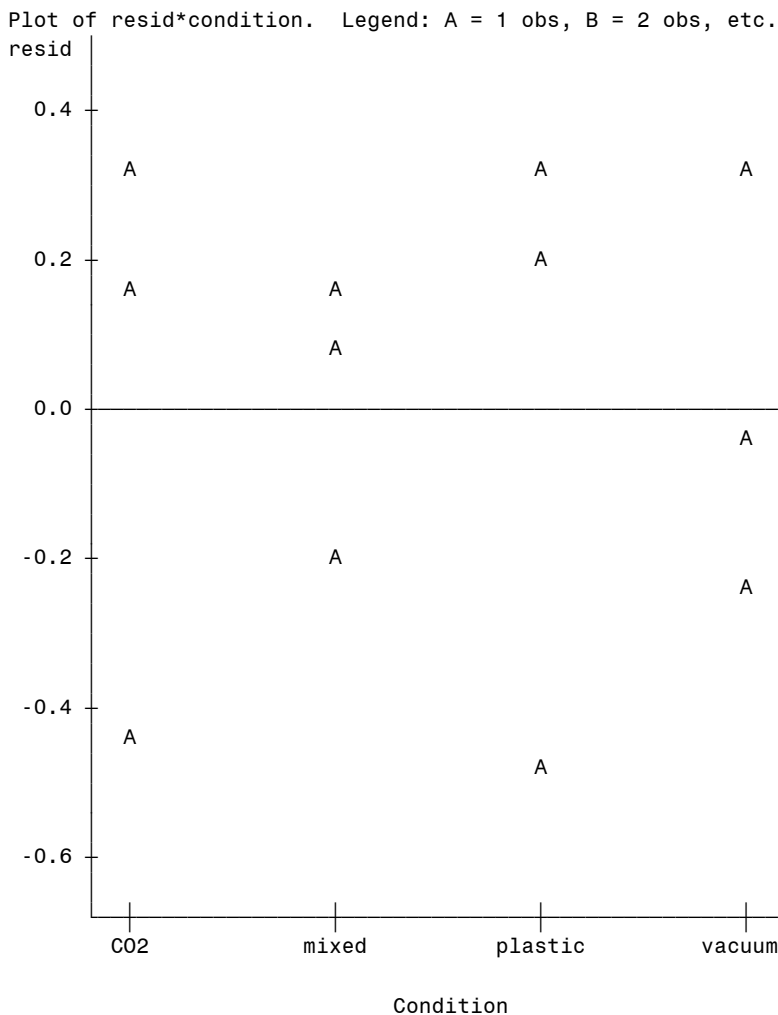
4. Outliers?

Using the above plots we note that none of the (standardized or studentized) residuals is beyond ± 3 .

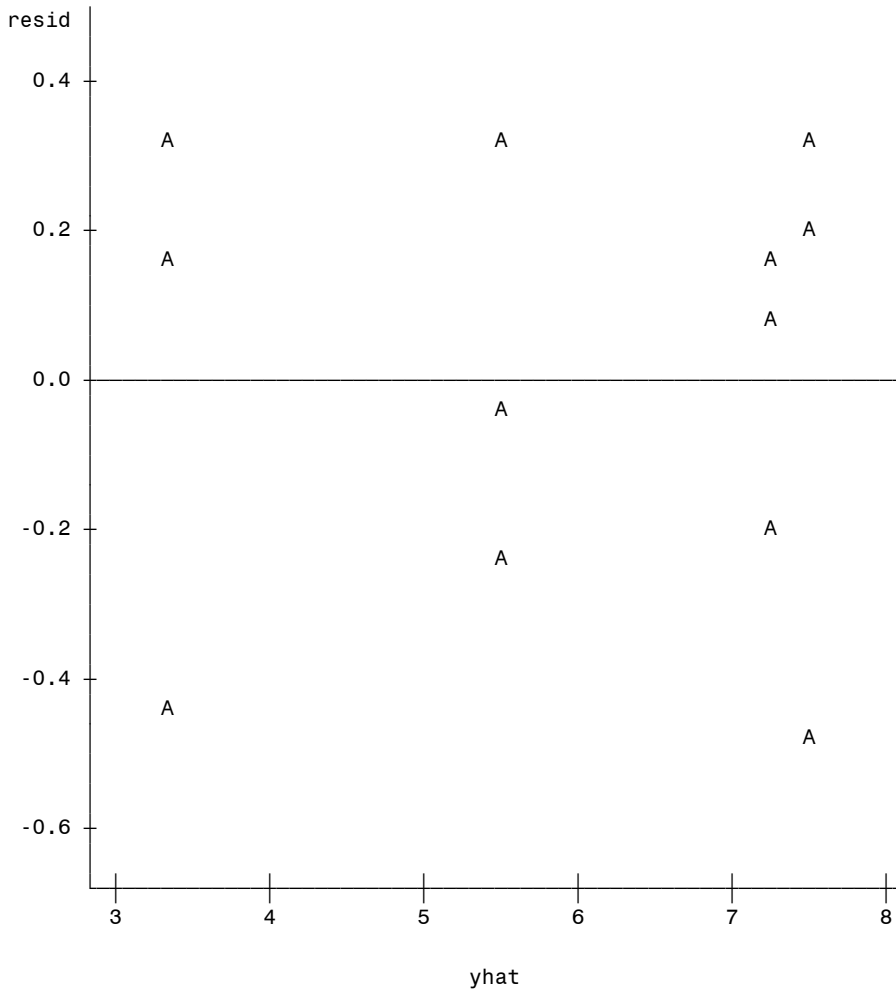
Or using a boxplot of the residuals we note that none of the outliers are flagged as being outlying using the "boxplot" rule of being beyond the "fences."



Using SAS

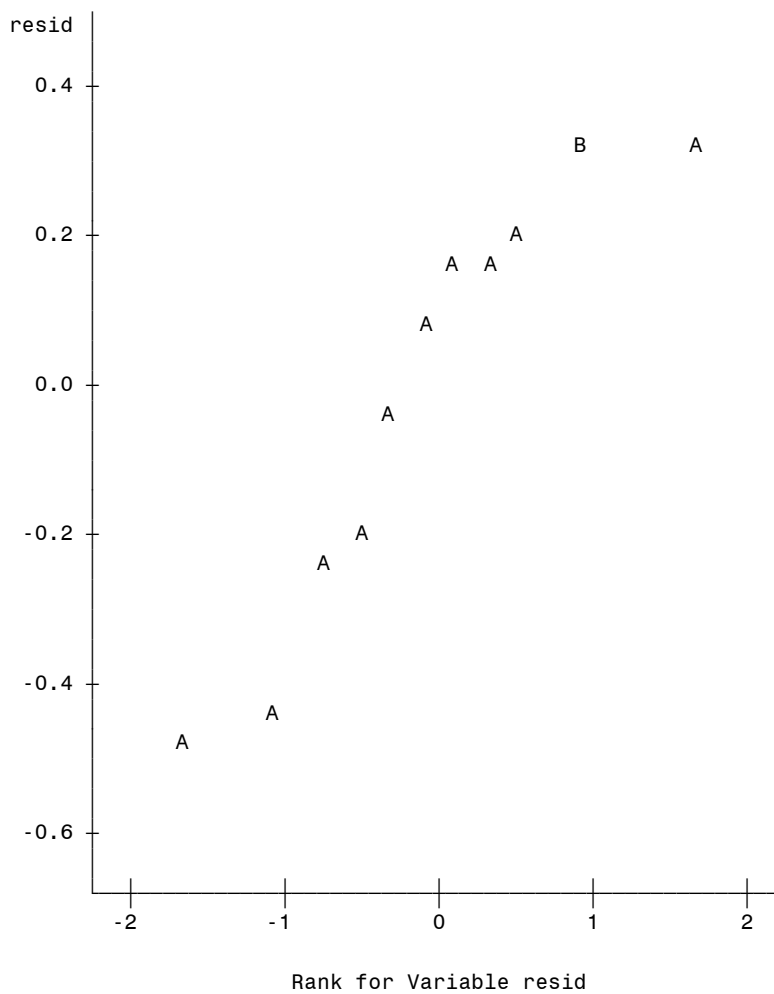


Plot of resid*yhat. Legend: A = 1 obs, B = 2 obs, etc.



- * Variances look constant [know pattern with increasing mean response]
- * None of the residuals stand out and look "outlying"

Plot of resid*nscore. Legend: A = 1 obs, B = 2 obs, etc.



* normal probability plot looks approximately linear

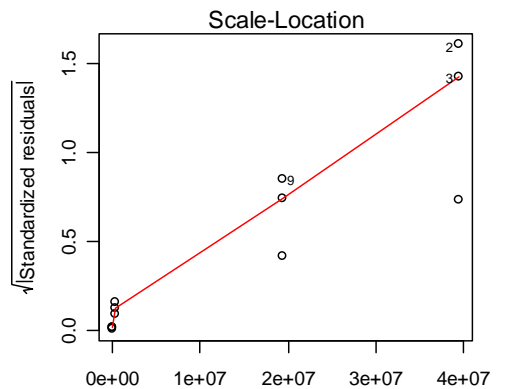
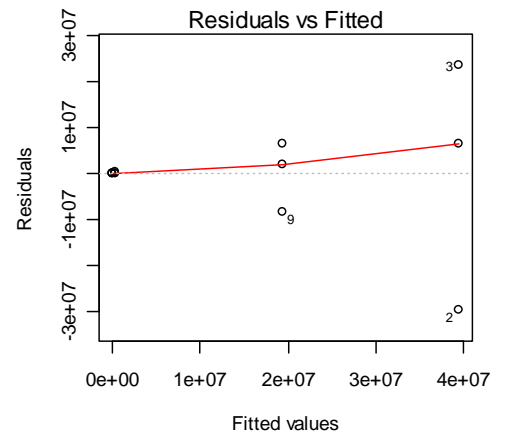
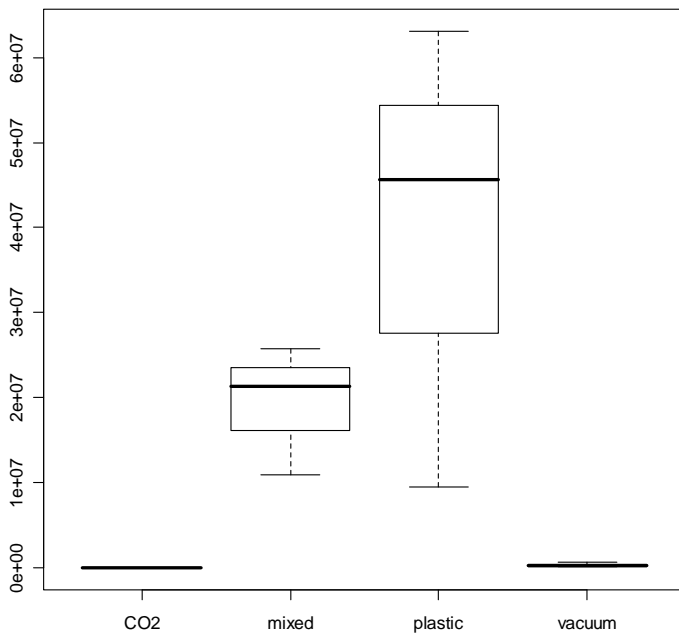
Example: How about the Meat study with using Bacterial Counts to show "When assumptions go BAD!"

Model: Bacterial Growth $_{ij} = \mu_i + \varepsilon_{ij}$, where ε_{ij} are independent $N(0, \sigma^2)$.

Using R

1. Constant Variance? Serious issues!

```
> MeatData$Count = 10^MeatData$logCount  
> boxplot(Count ~ Condition)
```

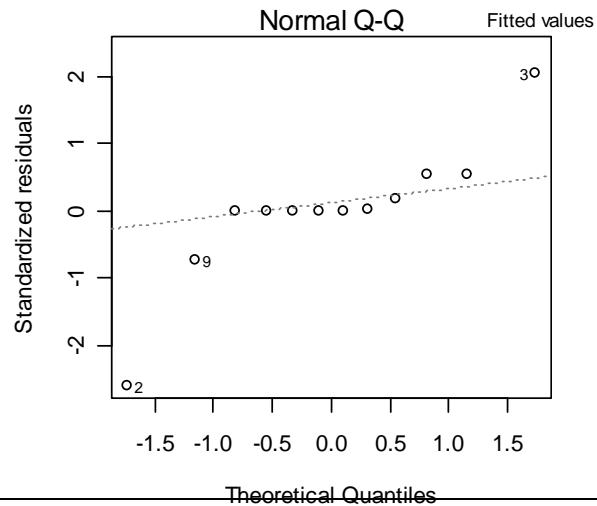


2. Normality? Serious issues!

```
> plot(MeatANOVACount)  
> shapiro.test(MeatANOVACount$residuals)
```

Shapiro-Wilk normality test

```
data: MeatANOVACount$residuals  
W = 0.8102, p-value = 0.01227
```



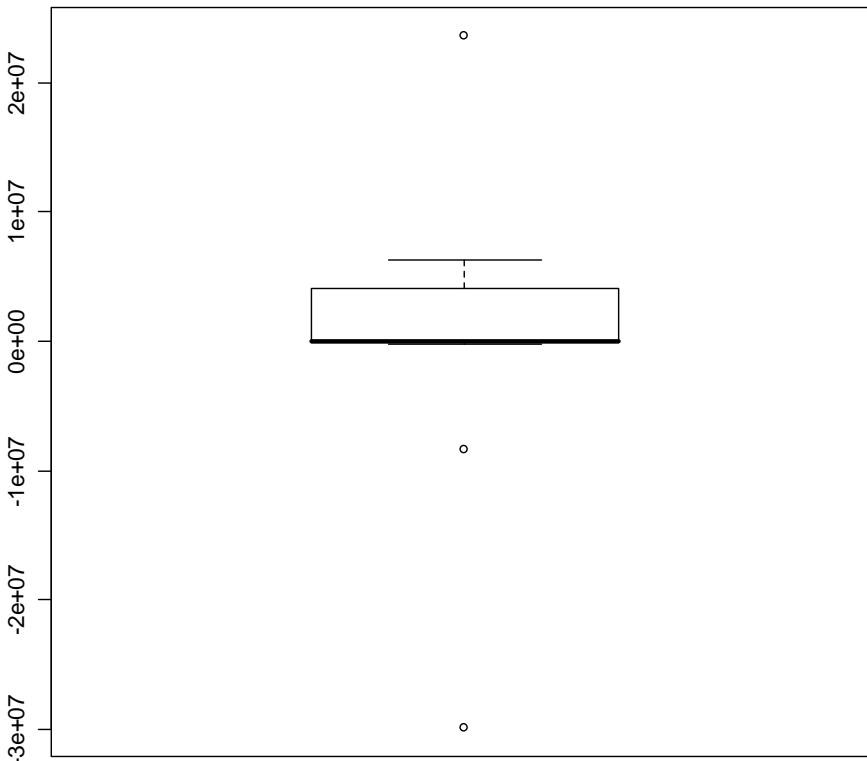
3. Independence?

Assume that correct sampling methods were used to insure independence.

4. Outliers?

Using the above plots we note that none of the residuals is beyond ± 3 .

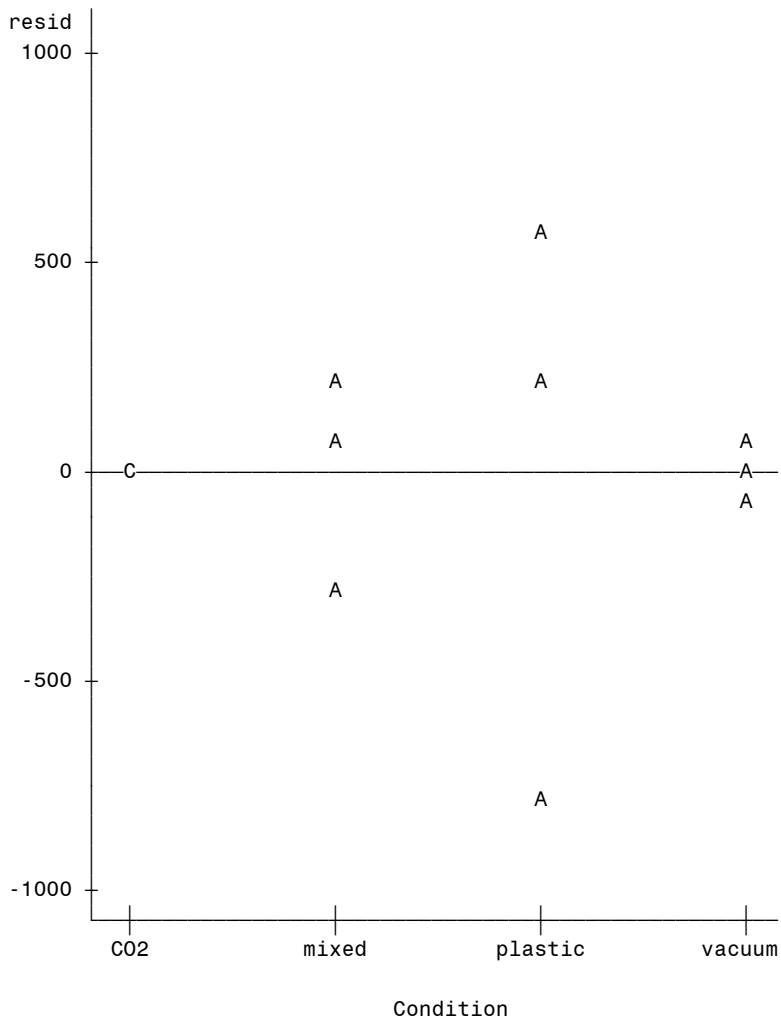
Or using a boxplot of the residuals we note that three of the outliers are flagged as being outlying using the "boxplot" rule of being beyond the "fences."



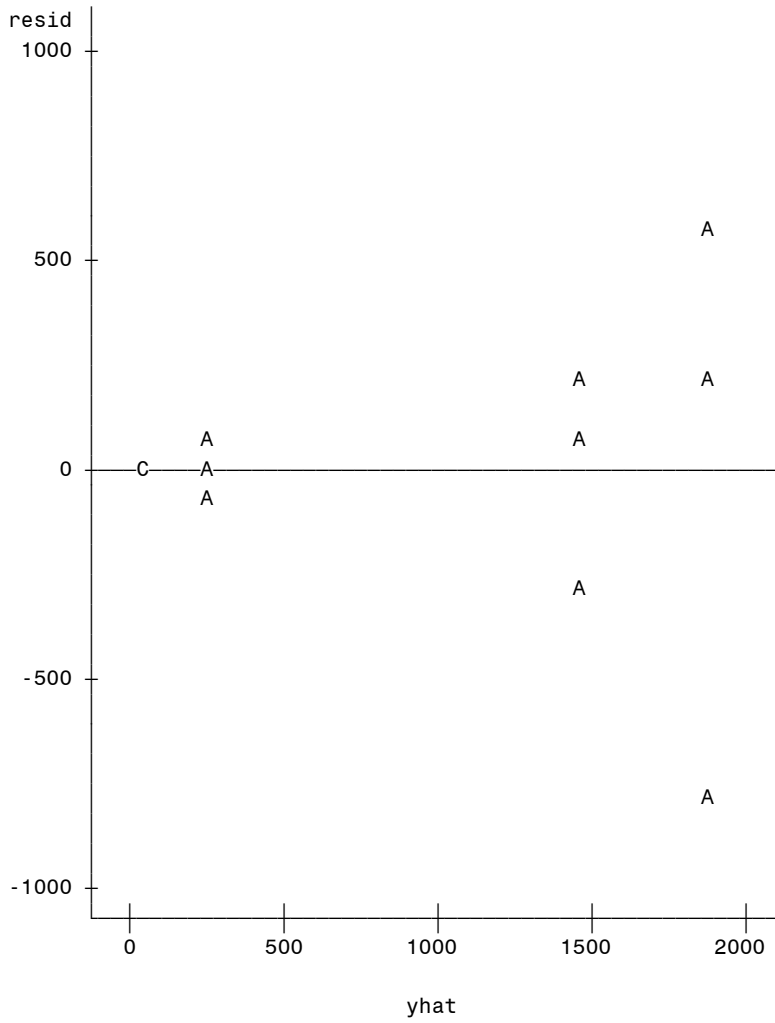
NOTE: By taking the log transformation of the Bacterial Counts, BOTH the Non Constant Variance AND the Normality problems were eliminated!

Using SAS

Plot of resid*condition. Legend: A = 1 obs, B = 2 obs, etc.

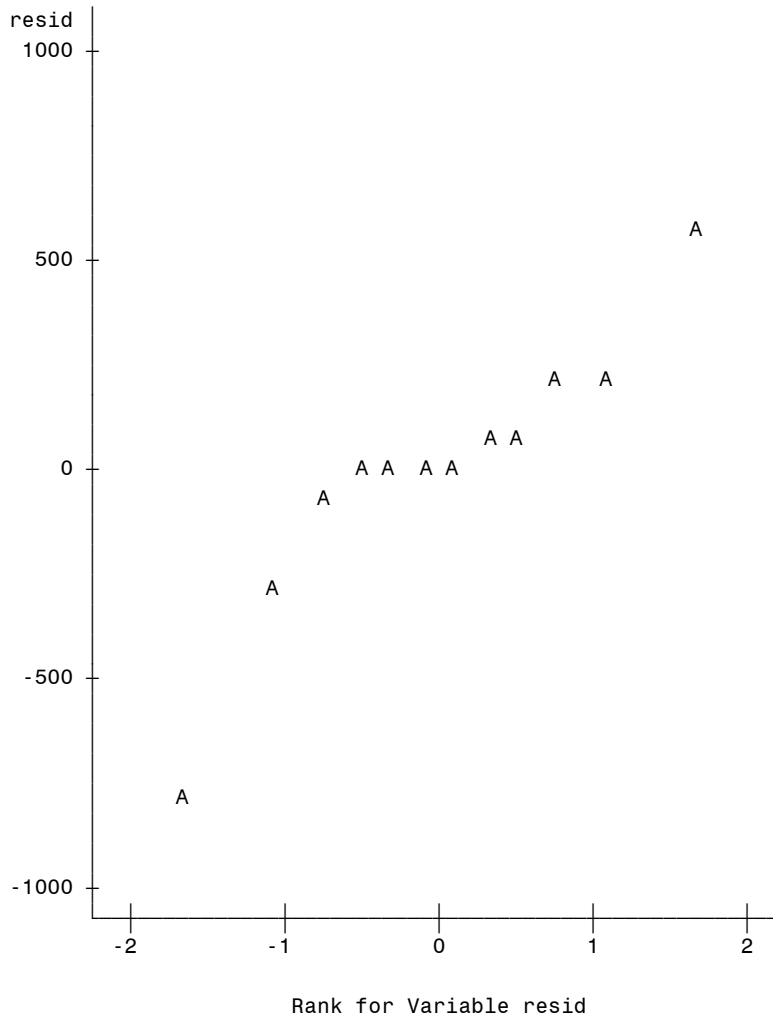


Plot of resid*yhat. Legend: A = 1 obs, B = 2 obs, etc.



* Variance obviously increases with increasing mean response

Plot of resid*nscore. Legend: A = 1 obs, B = 2 obs, etc.



* Nonlinear pattern here and nonconstant variance goes hand-in-hand

LEVINE'S TEST FOR HOMOGENEITY

Levine's test is a commonly used test to determine if the variances are constant in the popln's.

Levine's test is very simple.

1. Fit an ANOVA model and obtain the residuals about sample mean of group (or sample median of group as in Rcmdr).
2. Use the absolute values of these residuals in another ANOVA.
3. The resulting F test is Levine's Test of equality of variances.

Rcmdr: can request this directly: Statistics > Variances > Levene's test

Example: The Meat study with using logBacterial Counts

Using R

1. Constant Variance?

```
> summary(lm(abs(MeatANOVA$res) ~ MeatBacteria$Condition))
Call:
lm(formula = abs(MeatANOVA$res) ~ MeatBacteria$Condition)

Residuals:
    Min       1Q   Median       3Q      Max
-0.153333 -0.092500  0.001667  0.080000  0.166667

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.30000    0.07608   3.943  0.00428 **
MeatBacteria$Conditionmixed -0.15333    0.10760  -1.425  0.19197
MeatBacteria$Conditionplastic  0.03333    0.10760   0.310  0.76464
MeatBacteria$Conditionvacuum -0.10000    0.10760  -0.929  0.37989
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1318 on 8 degrees of freedom
Multiple R-squared:  0.3272,    Adjusted R-squared:  0.0749
F-statistic: 1.297 on 3 and 8 DF, p-value: 0.3404
```

H_0 : variances equal in all groups

H_A : not all equal

Conclusion: P-value = 0.34 implies fail to reject H_0 - insufficient evidence to conclude variances differ between the groups.

Interpretation: Can not conclude that the variances are different (at $\alpha = 0.05$ or $\alpha=0.10$ or ...)

Example: How about the Meat study with using Bacterial Counts to show "When assumptions go BAD!"

Using R

1. Constant Variance? Serious issues!

```
> summary(lm(abs(MeatCountANOVA$res) ~ MeatBacteria$Condition))  
Call:  
lm(formula = abs(MeatCountANOVA$res) ~ MeatBacteria$Condition)  
Residuals:  
    Min       1Q   Median       3Q      Max   
-13677052  -23614    1359   1272265   9967189  
Coefficients:  
                Estimate Std. Error  t value Pr(>|t|)      
(Intercept)         1374    3658626  0.000375  0.99971      
MeatBacteria$Conditionmixed  5588408    5174079   1.080  0.31159      
MeatBacteria$Conditionplastic 19933005    5174079   3.852  0.00486 **    
MeatBacteria$Conditionvacuum  177409    5174079   0.034  0.97349      
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
Residual standard error: 6337000 on 8 degrees of freedom  
Multiple R-squared:  0.711,    Adjusted R-squared:  0.6027  
F-statistic: 6.561 on 3 and 8 DF,  p-value: 0.01504
```

H_0 :

H_A :

Conclusion:

Interpretation:

NONPARAMETRIC KRUSKAL-WALLIS TEST

Suppose assumptions are not met: data not Normal, data have outliers that are REAL, and/or the variances are not constant, WHAT DO YOU DO?

CONSIDER A NONPARAMETRIC ALTERNATIVE!

H_0 : the distributions all have the same location/center

H_a : at least two distributions differ in terms of location/center

TS: Function of the RANKS of the observations

RR: Special Tables or a normal approximation for large sample sizes

Assumptions: distributions have same SHAPE and same SPREAD [so it isn't a free lunch]

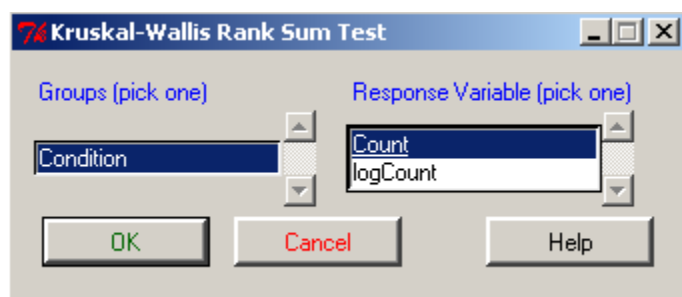
Opinion: Often a useful alternative if outliers present

Example: The Meat study with using **Bacterial Counts**

Using R

With *MeatBacteria* the active data set, using RCommander:

Statistics > Nonparametric tests > Kruskal-Wallis test ...



```
> tapply(MeatBacteria$Count, MeatBacteria$Condition, median, na.rm=TRUE)
      CO2      mixed      plastic      vacuum
3235.937 21379620.895 45708818.961 275422.870

> kruskal.test(Count ~ Condition, data=MeatBacteria)

Kruskal-Wallis rank sum test

data:  Count by Condition
Kruskal-Wallis chi-squared = 9.4615, df = 3, p-value = 0.02374
```

```

title "Nonparametric ANOVA/ Kruskal-Wallis";
title2 "Bacteria in meat data";
data meat;
  input condition $ logcount @@;
  count = exp(logcount);
  datalines;
plastic 7.66 plastic 6.98 plastic 7.80
vacuum 5.26 vacuum 5.44 vacuum 5.80
mixed 7.41 mixed 7.33 mixed 7.04
CO2 3.51 CO2 2.91 CO2 3.66
;
options ls=70 nocenter nodate formdlm="-";
proc nparlway;
title3 "response = log(count)";
class condition;
var logcount;
run;

proc nparlway;
title3 "response = count";
class condition;
var count;
run;

```

(output edited- NPAR1WAY gives lots of different nonparametric methods)

Nonparametric ANOVA/ Kruskal-Wallis 1
Bacteria in meat data
response = log(count)

The NPAR1WAY Procedure

Analysis of Variance for Variable logcount
Classified by Variable condition

condition	N	Mean
plastic	3	7.480
vacuum	3	5.500
mixed	3	7.260
CO2	3	3.360

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Among	3	32.87280	10.957600	94.5844	<.0001
Within	8	0.92680	0.115850		

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable logcount
Classified by Variable condition

condition	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
plastic	3	30.0	19.50	5.408327	10.0
vacuum	3	15.0	19.50	5.408327	5.0
mixed	3	27.0	19.50	5.408327	9.0

CO2 3 6.0 19.50 5.408327 2.0

Kruskal-Wallis Test

Chi-Square 9.4615
 DF 3
 Pr > Chi-Square 0.0237

 Nonparametric ANOVA/ Kruskal-Wallis 7
 Bacteria in meat data
 response = count

The NPAR1WAY Procedure

Analysis of Variance for Variable count
 Classified by Variable condition

condition	N	Mean
plastic	3	1879.09259
vacuum	3	251.07441
mixed	3	1439.73191
CO2	3	30.22214

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Among	3	7282652.347795	2427550.783	16.5587	0.0009
Within	8	1172820.616230	146602.577		

 Nonparametric ANOVA/ Kruskal-Wallis 8
 Bacteria in meat data
 response = count

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable count
 Classified by Variable condition

condition	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
plastic	3	30.0	19.50	5.408327	10.0
vacuum	3	15.0	19.50	5.408327	5.0
mixed	3	27.0	19.50	5.408327	9.0
CO2	3	6.0	19.50	5.408327	2.0

Kruskal-Wallis Test

Chi-Square 9.4615
 DF 3
 Pr > Chi-Square 0.0237

A SMALL DIGRESSION INTO DESIGNS AND A CRD MODEL

SOME QUICK TERMINOLOGY

RESPONSE: Continuous measurement of interest

FACTOR: Categorical variable that "defines" the populations

LEVELS: The different values that the FACTOR can take on

Also known as **TREATMENTS**.

TREATMENTS: Different levels of the factor **OR**

Combinations of several factors.

COMPLETELY RANDOMIZED DESIGN (CRD):

Treatments (or factor levels) are randomly assigned to experimental units or randomly sampling from existing populations (factor levels) that you want to compare. Only "constraint" on randomization is how many experimental units receive a particular treatment.

BALANCED DESIGN: Any design in which there are an equal number of observations in each treatment (or factor level). For the moment in CRD's, "balancing" has no effect, but with more complex designs which we will encounter later, **IT CAN AND WILL MAKE A HUGE DIFFERENCE!**

DIFFERENT CRD MODELS

Numeric data - independent Random Samples from "t" populations obtained

Assume $Y_{ij} \sim$ independent $N(\mu_i, \sigma^2)$ $i = 1, 2, \dots, t$ and $j = 1, 2, \dots, n_i$

CELL MEANS or MEANS MODEL

Defn: $Y_{ij} = \mu_i + \varepsilon_{ij}$ with $\varepsilon_{ij} \sim$ independent $N(0, \sigma^2)$ is the Cell Means or Means Model

b/c: the observations are expressed as deviations from the Means of the populations with the population means, the μ_i 's, the parameters of interest.

EFFECTS MODEL

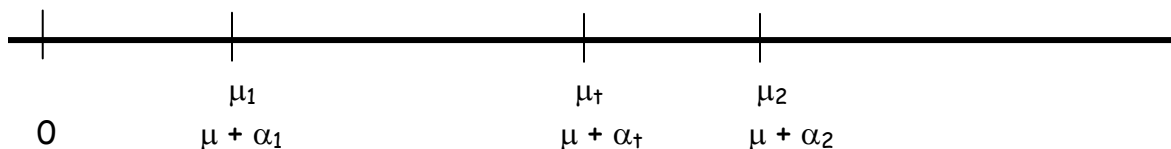
Defn: $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ with $\varepsilon_{ij} \sim$ independent $N(0, \sigma^2)$ is the Effects Model

b/c: the observations are still expressed as deviations from the Means of the populations, but now the means of the populations are represented in terms of an "overall" mean and an effect (or deviation) from this overall mean.

NOTE

1. While models appear to be different, the pertinent questions of interest (namely _____) can be investigated in EITHER MODEL!
2. While the Means Model is easier to present and understand, the Effects Model is the one that is used in software packages.

PICTORIALLY



HYPOTHESIS OF INTEREST

MEANS MODEL: $H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_t$ ie all "t" popln means equal

EFFECTS MODEL: $H_0: ?$

Note that $Y_{ij} = \mu_i + \varepsilon_{ij} = Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ implies

$$\mu_1 = \mu + \alpha_1$$

$$\mu_2 = \mu + \alpha_2$$

...

$$\mu_t = \mu + \alpha_t$$

and $\mu_1 = \mu_2 = \mu_3 = \dots = \mu_t \Rightarrow \mu + \alpha_1 = \mu + \alpha_2 = \dots = \mu + \alpha_t \Rightarrow \alpha_1 = \alpha_2 = \dots = \alpha_t$

so for EFFECTS MODEL: $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_t$

ESTIMATES OF MODEL PARAMETERS

MEANS MODEL: $Y_{ij} = \mu_i + \varepsilon_{ij}$ with $\varepsilon_{ij} \sim$ independent $N(0, \sigma^2)$

Parameters are: the _____ and _____.

Recall we estimated _____ with _____ = _____.

How would we estimate the _____ ?

Note this model has "t" unknown μ 's, but each is UNIQUELY estimated by the _____.

EFFECTS MODEL: $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ with $\varepsilon_{ij} \sim$ independent $N(0, \sigma^2)$

Parameters are: _____.

We estimate _____ also with _____ = _____.

How would we estimate _____ and the _____ ?

Note that this model has "t + 1" unknowns in μ and the α 's. To estimate the parameters requires a constraint. Typical constraints are:

1. set $\alpha_1 = 0$ or
2. set $\alpha_t = 0$ or
3. set $\sum_{i=1}^t n_i \alpha_i = 0$

Most programs (including R and SAS) use a constraint like the first or second.

NOTES:

1. To test the hypothesis of interest in ANOVA it doesn't matter which constraint is used.
2. However, estimates of the parameters of the EFFECTS MODEL WILL be different depending on which constraint is used.

Here's R and SAS output for the MeatBacteria data.

```
> summary(MeatANOVA)

Call:
lm(formula = logCount ~ Condition, data = MeatBacteria)

Residuals:
    Min       1Q   Median       3Q      Max
-0.500 -0.225  0.110  0.210  0.320

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.3600    0.1965   17.10 1.39e-07 ***
Condition[T.mixed] 3.9000    0.2779   14.03 6.45e-07 ***
Condition[T.plastic] 4.1200    0.2779   14.82 4.22e-07 ***
Condition[T.vacuum] 2.1400    0.2779    7.70 5.74e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3404 on 8 degrees of freedom
Multiple R-squared:  0.9726,    Adjusted R-squared:  0.9623
F-statistic: 94.58 on 3 and 8 DF,  p-value: 1.376e-06
```

```
proc glm;
  class condition;
  model logcount= condition/solution;
run;
```

Dependent Variable: logcount

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	32.87280000	10.95760000	94.58	<.0001
Error	8	0.92680000	0.11585000		
Corrected Total	11	33.79960000			

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	5.500000000	0.19651124	27.99	<.0001
condition C02	-2.140000000	0.27790886	-7.70	<.0001
condition mixed	1.760000000	0.27790886	6.33	0.0002
condition plastic	1.980000000	0.27790886	7.12	<.0001
condition vacuum	0.000000000	.	.	.

FITTING THE MEANS AND EFFECTS MODELS IN R AND SAS

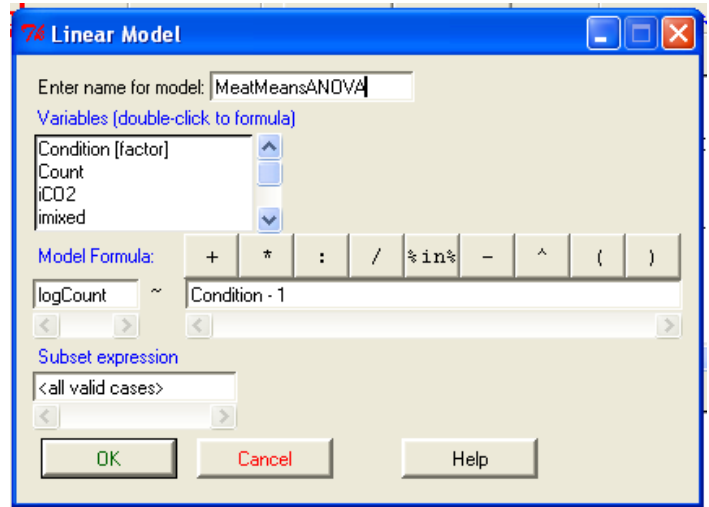
The default models for most ANOVA program/packages is the Effects Model ($Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$) as we just saw in R and SAS! However, the Means Model can easily be obtained in such programs/packages by eliminating the μ term from the model.

Using R

With *MeatBacteria* the active data set, using RCommander:

Statistics > Fit Models > Linear model ...

With the following output:



```
> MeatMeansModel <- lm(logCount ~ Condition - 1, data=MeatBacteria)
> summary(MeatMeansModel)

Call:
lm(formula = logCount ~ Condition - 1, data = MeatBacteria)

Residuals:
    Min       1Q   Median       3Q      Max
-0.500 -0.225  0.110  0.210  0.320

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
ConditionCO2      3.3600     0.1965  17.10 1.39e-07 ***
Conditionmixed    7.2600     0.1965  36.94 3.16e-10 ***
Conditionplastic  7.4800     0.1965  38.06 2.49e-10 ***
Conditionvacuum   5.5000     0.1965  27.99 2.87e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3404 on 8 degrees of freedom
Multiple R-squared:  0.9979,    Adjusted R-squared:  0.9969
F-statistic: 972.4 on 4 and 8 DF,  p-value: 8.861e-11
```

Using SAS

```
options ls=110 nocenter nodate formdlm="-";
title "Fitting a Means Model in SAS";
title2 "Bacteria in meat data";
data meat;
  input condition $ logcount @@;
  datalines;
plastic 7.66 plastic 6.98 plastic 7.80
vacuum 5.26 vacuum 5.44 vacuum 5.80
mixed 7.41 mixed 7.33 mixed 7.04
CO2 3.51 CO2 2.91 CO2 3.66
;
proc glm;
  class condition;
  model logcount = condition / noint solution;
run;
```

Fitting a Means Model in SAS
Bacteria in meat data

4

The GLM Procedure
Dependent Variable: logcount

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	450.5928000	112.6482000	972.36	<.0001
Error	8	0.9268000	0.1158500		
Uncorrected Total	12	451.5196000			

R-Square	Coeff Var	Root MSE	logcount Mean
0.972580	5.768940	0.340367	5.900000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
condition	4	450.5928000	112.6482000	972.36	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
condition	4	450.5928000	112.6482000	972.36	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > t
condition CO2	3.360000000	0.19651124	17.10	<.0001
condition mixed	7.260000000	0.19651124	36.94	<.0001
condition plastic	7.480000000	0.19651124	38.06	<.0001
condition vacuum	5.500000000	0.19651124	27.99	<.0001

Comparing both the R and SAS outputs of this Means Models, our estimates of the parameters, the μ_i 's, are nothing more than the sample averages for each sample. Recall from R these sample averages were:

```
> numSummary(logCount, groups=Condition, statistics=c("mean", "sd"))
      mean      sd n
CO2    3.36 0.3968627 3
mixed  7.26 0.1946792 3
plastic 7.48 0.4386342 3
vacuum  5.50 0.2749545 3
```

Below is the "Means Model" ANOVA output.

```
> summary(MeatMeansANOVA)

Call:
lm(formula = logCount ~ Condition - 1, data = MeatBacteria)

Residuals:
    Min       1Q   Median       3Q      Max
-0.500 -0.225  0.110  0.210  0.320

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
ConditionCO2      3.3600    0.1965   17.10 1.39e-07 ***
Conditionmixed     7.2600    0.1965   36.94 3.16e-10 ***
Conditionplastic   7.4800    0.1965   38.06 2.49e-10 ***
Conditionvacuum    5.5000    0.1965   27.99 2.87e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3404 on 8 degrees of freedom
Multiple R-squared:  0.9979,    Adjusted R-squared:  0.9969
F-statistic: 972.4 on 4 and 8 DF,  p-value: 8.861e-11
```

Below is the default "Effects Model" ANOVA output for comparison purposes.

```
> summary(MeatANOVA)

Call:
lm(formula = logCount ~ Condition, data = MeatBacteria)
---
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.3600    0.1965   17.10 1.39e-07 ***
Condition[T.mixed]  3.9000    0.2779   14.03 6.45e-07 ***
Condition[T.plastic] 4.1200    0.2779   14.82 4.22e-07 ***
Condition[T.vacuum]  2.1400    0.2779    7.70 5.74e-05 ***
---
Residual standard error: 0.3404 on 8 degrees of freedom
Multiple R-squared:  0.9726,    Adjusted R-squared:  0.9623
F-statistic: 94.58 on 3 and 8 DF,  p-value: 1.376e-06
```

Note that

1. the estimate of σ , square root mse, is the same,
2. estimates of the β 's are different (not surprising since the "models" are different!)
3. the F-test results are different
4. the R^2 value is different.