

Analysis of Covariance in Agronomy and Crop Research

Rong-Cai Yang
Alberta Agriculture and Rural Development
and
University of Alberta
Rong-cai.yang@ualberta.ca

Outline

- Overview of Analysis of Covariance (ANCOVA)
- Basic theory and principles
- Conventional uses
- Elaborated applications
- Take-home messages

- Most stats textbooks would devote one chapter to ANCOVA (e.g., Steel et al. 1997, Ch 17; Snedecor & Cochran 1980, Ch 18)
 - Milliken and Johnson (2002, Analysis of messy data. Volume 3: Analysis of covariance) devote the entire book to the subject
- Other books specifically for SAS users also have a chapter on ANCOVA (e.g., Littell et al. 2006. SAS for mixed models, 2nd ed., Ch 7)

However, ANCOVA is a more advanced topic, often appearing towards the end of books;

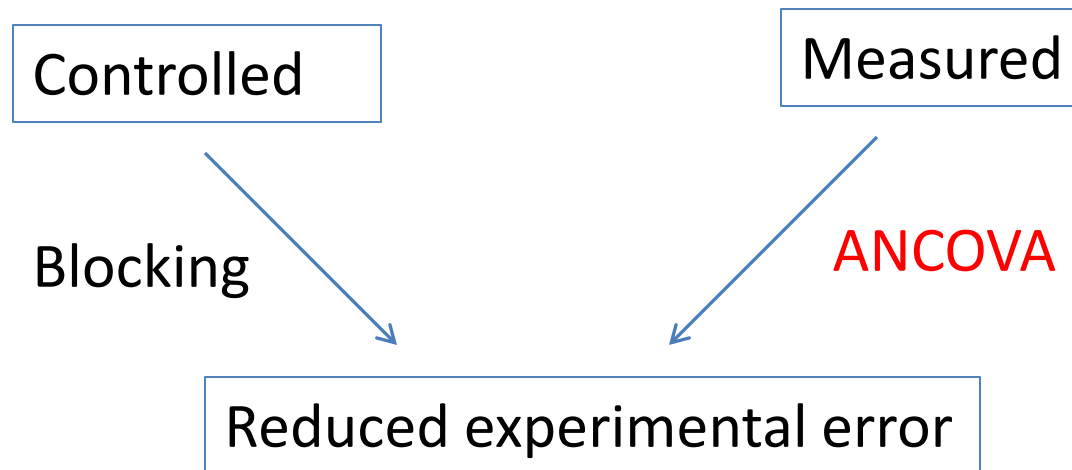
ANCOVA is taught cursorily or ignored completely in many stats classes

What is ANCOVA?

- ANCOVA is a statistical technique that combines the methods of ANOVA and regression.
- ANCOVA has two types of independent variables
 - Dummy (0-1) variables for treatment IDs
 - Continuous variables (covariates) – directly measured
- If there are only dummy variables, ANCOVA becomes ANOVA
- If there are only covariates, ANCOVA becomes regression analysis

Why ANCOVA?

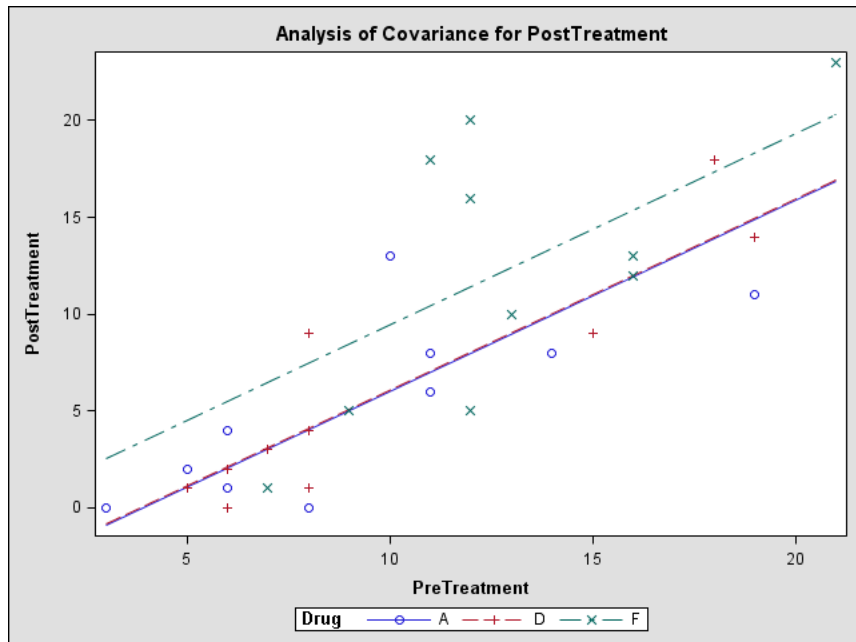
- Two kinds of nuisance factors contribute to experimental error



Choice of covariates

- Nuisance factors that can be measured but not controlled (e.g., blocking)
 - Closely related to the response variable (y)
- Remove more error variation that cannot be accounted for by blocking
 - e.g., insect movement and soil fertility gradient are not in the same direction
- Pre-study variables (i.e., measured before the start of study) to ensure that they are not influenced by the treatments being tested
 - Plot-to-plot heterogeneity (e.g., soil moisture, soil nutrients, weed population, non-uniform insect distribution)
 - Residual effects of previous trials

Statistical model



- The simplest ANCOVA model: one-way trt structure with t_i , one independent covariate x_{ij} and associated regression coefficient β
 - $y_{ij} = \beta_0 + t_i + \beta X_{ij} + \varepsilon_{ij}$
- This model represents a set of parallel lines
- The common slope of lines is β
- The intercept of the i th line is $(\beta_0 + t_i)$.
- If X s were not measured, then $\beta X_{ij} +$ could not be determined and would thus be included in the error term
 - $y_{ij} = \beta_0 + t_i + e_{ij}$

How does ANCOVA work?

- ANCOVA is essentially an ANOVA of the quantity $y_{ij} - \beta X_{ij}$. The value of slope β is chosen so that error SS of $y_{ij} - \beta X_{ij}$,

- $E_{yy} - 2\beta E_{xy} + \beta^2 E_{xx}$
is minimized.

Some re-arrangements lead to

- $E_{xx}(\beta - E_{xy}/E_{xx})^2 + E_{yy} - (E_{xy})^2/E_{xx}$

Thus, least squares estimate of the slope is:

- $\beta = E_{xy}/E_{xx}$

and the minimum error SS is:

- $E_{yy} - (E_{xy})^2/E_{xx}$

	df	SS_x	CP_{xy}	SS_y	df	$SS_{y x}$
Treatment (T)	t-1	T_{xx}	T_{xy}	T_{yy}		
Error (E)	t(r-1)	E_{xx}	E_{xy}	E_{yy}	t(r-1)-1	$SS1 = E_{yy} - (E_{xy})^2/E_{xx}$
Total (T+E)	tr-1	$T_{xx} + E_{xx}$	$T_{xy} + E_{xy}$	$T_{yy} + E_{yy}$	tr-2	$SS2 = T_{yy} + E_{yy} - (T_{xy} + E_{xy})^2 / (T_{xx} + E_{xx})$
Adj treatment					t-1	$T'_{yy} = SS2 - SS1$



T'_{yy} is the adjusted SS for treatment = SAS output

Conventional uses of ANCOVA

- Adjusted means
 - Trt means of the y variable are adjusted to a common value of covariate => equitable comparison of trt means
- Statistical control of errors
 - Variation in y due to its association with covariates is removed from the error variance => more precise estimates of trt means and more powerful test
- Testing for homogeneity of slopes for different treatment groups
 - Are regression lines parallel for different groups?
- Estimating missing values
 - Less useful now => present-day stats software can easily handle unbalanced data

Stand (x) and yield (y) (lbs field weight of ear corn) of six varieties in RCBD with four blocks (Snedecor & Cochran 1980, Table 18.5.2)

SAS code:

```
proc mixed data=sc;
class variety block;
model y=variety
      block x;
run;
```

Output from SAS PROC MIXED:

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
variety	5	14	6.64	0.0023
block	3	14	5.15	0.0132
x	1	14	76.01	<.0001

Detailed calculations

$$\beta = 917.25/113.83 = 8.06$$

	df	SS _x	CP _{xy}	SS _y	df	SS _{y x}	MS _{y x}	F	Pr>F
Block	3	21.67	8.50	436.17					
Variety	5	45.83	559.25	9490.00					
Error	15	113.83	917.25	8752.33	14	1361.26	97.23		
V+E	20	159.67	1476.50	18242.33	19	4588.50			
Adj V					5	3227.24	645.45	6.64	0.0023

Adjusting treatment means

- Single mean:

$$y_{i. (adj)} = y_{i.} - \beta(x_{i.} - x_{..})$$

with a standard error (SE)

$$SE = \{MS_{y|x} [1/r + (x_{i.} - x_{..})^2/E_{xx}]\}^{0.5}$$

For example, adj variety a,

$$y_{a. (adj)} = 173.0 - (8.06)(24.0 - 26.3) = 191.8$$

$$SE = \{97.23(1/4 + (-2.3)^2/113.8)\}^{0.5} = 5.38$$

$x_{i.}$	$y_{i.}$	$y_{i. (adj)}$	SE
24.0	173.0	191.8	5.38
25.3	182.3	191.0	5.03
26.5	194.5	193.2	4.93
28.0	232.8	219.3	5.17
27.8	201.0	189.6	5.10
26.5	215.0	213.7	4.93
$x_{..}=26.3$			

SAS PROC MIXED output with LSMEANS statement:

Effect	variety	Estimate	Standard Error	DF	t Value	Pr > t
variety	a	191.80	5.3814	14	35.64	<.0001
variety	b	190.98	5.0310	14	37.96	<.0001
variety	c	193.16	4.9328	14	39.16	<.0001
variety	d	219.32	5.1654	14	42.46	<.0001
variety	e	189.58	5.1013	14	37.16	<.0001
variety	f	213.66	4.9328	14	43.31	<.0001

Difference between adjusted treatment means

- Difference between two adjusted means:

$$Y_{i. (adj)} - Y_{j. (adj)} = y_i - y_j - \beta(x_i - x_j)$$

with a standard error (SE)

$$SE = \{MS_{y|x} [(1/r_i + 1/r_j + (x_i - x_j)^2/E_{xx})]\}^{0.5}$$

For example, diff between varieties a and b,

$$Y_{a. (adj)} - Y_{b. (adj)} = 191.8 - 191.0 = 0.8,$$

$$SE = \{97.23(2/4 + (24.0-25.3)^2/113.8)\}^{0.5} \\ = 7.07$$

x_i	y_i	$y_i (adj)$
24.0	173.0	191.8
25.3	182.3	191.0
26.5	194.5	193.2
28.0	232.8	219.3
27.8	201.0	189.6
26.5	215.0	213.7
26.3		

SAS PROC MIXED output with LSMEANS statement:

Effect	variety	_variety	Estimate	Standard Error	DF	t Value	Pr > t
variety	a	b	0.8223	7.0677	14	0.12	0.9090
variety	a	c	-1.3554	7.3455	14	-0.18	0.8562
variety	a	d	-27.5187	7.8920	14	-3.49	0.0036
variety	a	e	2.2169	7.7865	14	0.28	0.7800
variety	a	f	-21.8554	7.3455	14	-2.98	0.0100
. . .							

Relative efficiency of ANCOVA vs. ANOVA

- Error MS from ANOVA without considering covariate (x)
= $8752.33/15 = 583.49$
- Error MS from ANOVA after considering covariate (x)
= 97.23
- Effective error MS = $MS_{y|x} [1 + T_{xx}/(t-1)/E_{xx}] = 97.23 * [1 + 45.83/5/113.83] = 105.06$
- Rel Efficiency = $583.49/105.06 = 5.55$
 - ANCOVA with 10 replications gives as precise estimates as unadjusted means with 55 replications!!

	df	SS _x	CP _{xy}	SS _y	df	SS _{y x}	MS _{y x}	F	Pr>F
Block	3	21.67	8.50	436.17					
Variety	5	45.83	559.25	9490.00					
Error	15	113.83	917.25	8752.33	14	1361.26	97.23		
V+E	20	159.67	1476.50	18242.33	19	4588.50			
Adj V					5	3227.24	645.45	6.64	0.0023

Elaborated applications of ANCOVA to agronomy and crop research

- Application #1: Analysis of dosage response
- Application #2: Analysis of treatment stability across environments
- Application #3: Analysis of spatial variability

Application #1: Analysis of dosage response

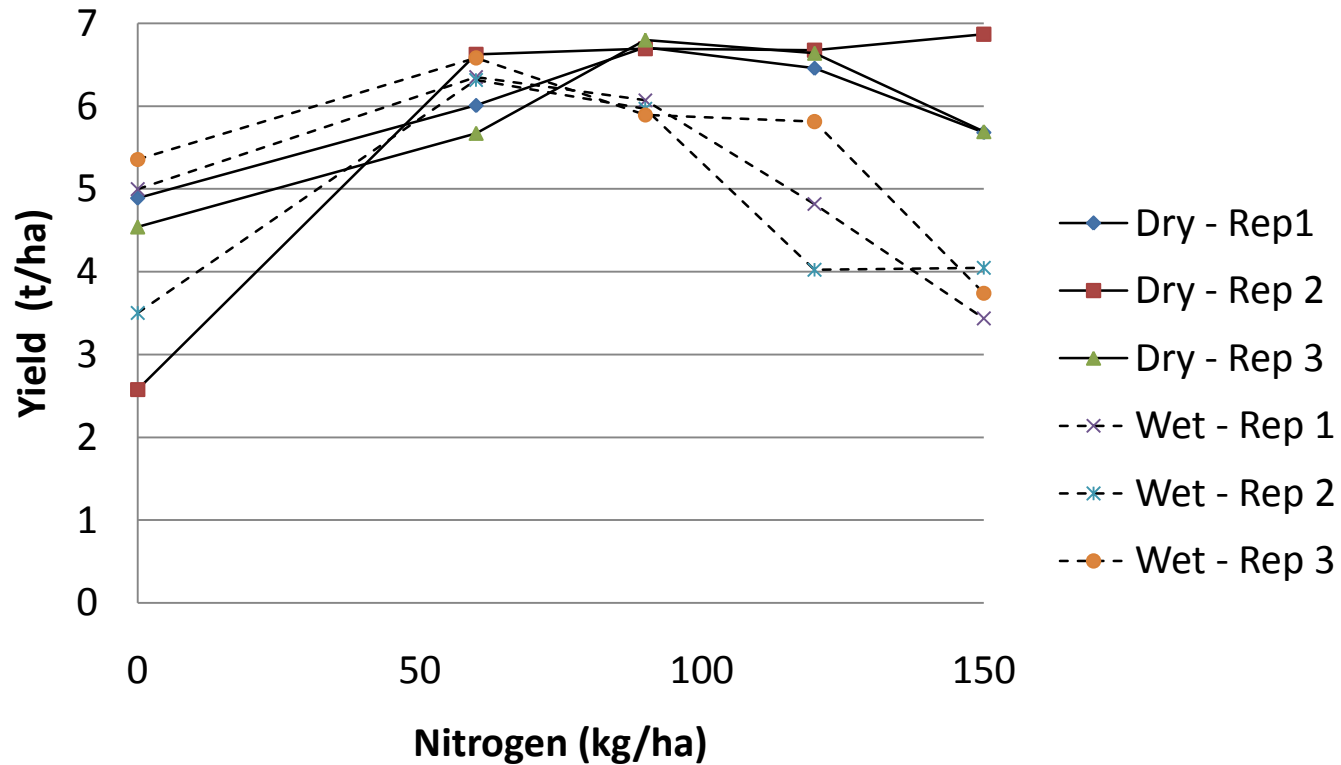
```
/*Gomez and Gomez 1984, pages 317 - 327*/  
/*A fertilizer trial with five nitrogen rates  
   (kg/ha) tested on rice yield (tonne/ha)  
   for each of two seasons, dry and wet. Each  
   trial has a RCBD with three replications.  
*/
```

```
options ls=100 ps=6000;  
data raw;  
input season $ nitrogen r1 r2 r3;  
n=nitrogen; *n is set to be a covariate;  
datalines;  
dry0      4.891    2.577    4.541  
dry60     6.009    6.625    5.672  
dry90     6.712    6.693    6.799  
dry120    6.458    6.675    6.639  
dry150    5.683    6.868    5.692  
wet0      4.999    3.503    5.356  
wet60     6.351    6.316    6.582  
wet90     6.071    5.969    5.893  
wet120    4.818    4.024    5.813  
wet150    3.436    4.047    3.740  
;run;
```

Objectives:

- 1.Examine if there is differential yield response to N rates in dry and wet seasons
- 2.Determine if it is necessary to have a separate technology recommendation for two seasons.

Since there are five N rates, a fourth-degree polynomial can be fit



Standard (old) method:

- Partition trt SS using orthogonal contrasts due to linear, quadratic and higher-order regression effects
 - $\text{Trt SS} = \text{SS}(\text{linear}) + \text{SS}(\text{quadratic}) + \dots$
- Older editions of textbooks give tables of orthogonal polynomial coefficients for balanced data with equally spaced treatment levels
 - How about unbalanced data with unequally spaced treatment levels?
 - How to estimate regression equations?
 - In a factorial experiment, are all regressions the same over all levels of the other factor (i.e., homogeneity of slopes)?

Orthogonal polynomial analysis

- Using the SAS IML ORPOL function to obtain orthogonal polynomial coefficients for five unequally spaced nitrogen levels (0, 60, 90, 120 and 150)

```
proc iml;  
levels={0 60 90 120 150};  
coef=orpol(levels`);  
print coef;  
quit;  
run;
```

Orthogonal polynomial analysis

Nitrogen level	0	60	90	120	150
Linear	-0.7278	-0.2080	0.0520	0.3119	0.5719
Quadratic	0.4907	-0.4729	-0.4595	-0.1160	0.5576
Cubic	-0.1677	0.6312	-0.2170	-0.6213	0.3748
Quartic	0.0367	-0.3671	0.7342	-0.5507	0.1468

- The coefficients are orthonormal because the squared coefficients for each contrast sum to one.
 - E.g., $(-0.7278)^2 + \dots + (0.5719)^2 = 1$.
- Orthogonal means the sum of cross-products of coefficients in any pair of rows would be zero:
 - E.g., $-0.7278 \cdot 0.4907 + \dots + 0.5719 \cdot 0.5576 = 0$

PROC GLM is used to show exact partitioning of the total trt SS into components due to linear, quadratic, cubic and quartic responses based on orthogonal polynomial coefficients

```
proc glm data=new ;  
class season nitrogen rep;  
model y=season rep(season) nitrogen season*nitrogen/ss1;  
random rep(season)/test;  
/*make sure the sum of coefficients for each contrast must be  
numerically zero!!*/  
contrast 'Linear' nitrogen -0.7278 -0.2080 0.0520 0.3119 0.5719;  
contrast 'Quadratic' nitrogen 0.4907 -0.4729 -0.4595 -0.1160 0.5577;  
contrast 'Cubic' nitrogen -0.1677 0.6312 -0.2170 -0.6213 0.3748;  
contrast 'Quartic' nitrogen 0.0368 -0.3671 0.7342 -0.5507 0.1468;  
run;
```



Partition the total TRT SS

PROC GLM outputs

Source	DF	Type I SS	Mean Square	F Value	Pr > F
season	1	4.49771520	4.49771520	10.19	0.0057
rep(season)	4	1.26136747	0.31534187	0.71	0.5943
nitrogen	4	18.75018453	4.68754613	10.62	0.0002
season*nitrogen	4	9.65721480	2.41430370	5.47	0.0057

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Linear	1	1.43575348	1.43575348	3.25	0.0902
Quadratic	1	17.17340157	17.17340157	38.90	<.0001
Cubic	1	0.09260605	0.09260605	0.21	0.6531
Quartic	1	0.04722634	0.04722634	0.11	0.7479

Use of ANCOVA for dosage response analysis

```
proc glm data=new ;  
class season nitrogen rep;  
model y=season rep(season) n n*n nitrogen season*n  
      season*n*n season*nitrogen/ss1;  
random rep(season)/test;  
run;
```

Note:

(i) Variable n is not in CLASS statement and thus is treated as a covariate; but nitrogen is a CLASS variable.

(ii) **SS1** option is used... Why not **SS3** option??

Don't use Type III SS as adjustments of linear and linear x type interaction effects for quadratic effects produce nonsense results!!!

ANOVA

Source	DF	Type I SS	Mean Square	F Value	Pr > F
season	1	4.49771520	4.49771520	10.19	0.0057
rep(season)	4	1.26136747	0.31534187	0.71	0.5943
nitrogen	4	18.75018453	4.68754613	10.62	0.0002
season*nitrogen	4	9.65721480	2.41430370	5.47	0.0057

ANCOVA

Source	DF	Type I SS	Mean Square	F Value	Pr > F
season	1	4.49771520	4.49771520	10.19	0.0057
rep(season)	4	1.26136747	0.31534187	0.71	0.5943
n	1	1.43627121	1.43627121	3.25	0.0901
n*n	1	17.17400333	17.17400333	38.90	<.0001
nitrogen	2	0.13990999	0.06995499	0.16	0.8548
n*season	1	8.81004604	8.81004604	19.95	0.0004
n*n*season	1	0.54976514	0.54976514	1.25	0.2810
season*nitrogen	2	0.29740363	0.14870181	0.34	0.7190



(i) ANCOVA provides the same partitioning of trt SS into linear, quadratic,...

(ii) ANCOVA reveals significant season*nitrogen is due to heterogeneity of linear responses between 2 seasons

PROC MIXED can be used to produce the same results, but the METHOD=TYPE1 option should be used

```
proc mixed data=new method=type1;  
class season nitrogen rep;  
model y=season n n*n nitrogen season*n season*n*n  
      season*nitrogen;  
random rep(season);  
run;
```

PROC MIXED will produce the same output as PROC GLM plus a correct F-test for season effect.

PROC MIXED makes it easier to obtain regression equations; SEs are unbiased with random blocks

```
proc mixed data=new
  method=type1;
class season nitrogen rep;
model y=season n(season)
      n*n(season) /noint
      solution;
random rep(season);
run;
```

In PROC MIXED, the RANDOM rep(season) statement and the NOINT option cause the intercepts to be estimated directly.

```
proc glm data=new;
class season nitrogen rep;
model y=season rep n(season)
      n*n(season) /ss1 solution;
random rep;
estimate 'beta_0--dry season'
intercept 3 rep 1 1 1 season 3
0/divisor=3;
estimate 'beta_0--wet season'
intercept 3 rep 1 1 1 season 0
3/divisor=3;
run;
```

In PROC GLM, the ESTIMATE statement is needed to estimate the intercepts, but SEs are underestimated.

Use of MIXED to obtain regression equations

Solution for Fixed Effects

Effect	season	Estimate	Standard Error	DF	t Value	Pr > t
season	dry	3.9825	0.3426	4	11.63	0.0003
season	wet	4.6749	0.3426	4	13.65	0.0002
n(season)	dry	0.05233	0.01005	20	5.21	<.0001
n(season)	wet	0.04772	0.01005	20	4.75	0.0001
n*n(season)	dry	-0.00025	0.000065	20	-3.93	0.0008
n*n(season)	wet	-0.00037	0.000065	20	-5.64	<.0001

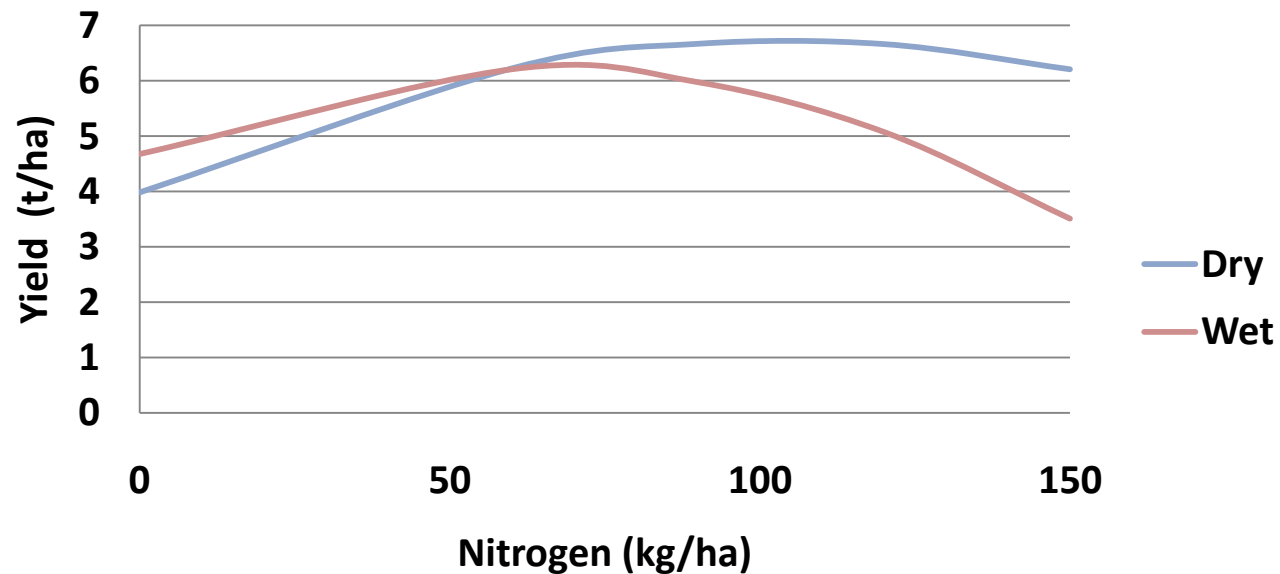
Quadratic regression equations for dry and wet seasons can be obtained as

$$\text{Dry: } Y = 3.9825 + 0.05233N - 0.00025N^2$$

$$\text{Wet: } Y = 4.6749 + 0.04772N - 0.00037N^2$$

The rate of yield increase with increase in the N rate is higher in dry season than in wet season.

Max yield reached at 102.5 kg N/ha in dry season and 65.2 kg N/ha in wet season



So there is the need for different nitrogen recommendations for dry vs. wet seasons!!

To test for season*n and season*n*n, further partitioning of error variance is needed

```
proc mixed data=new method=type1 covtest;  
class season nitrogen rep;  
model y=season n n*n nitrogen season*n season*n*n  
      season*nitrogen;  
random rep(season) n*rep(season) n*n*rep(season);  
run;
```

Use of MIXED to further divide error variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	34.16648200	2.62819092	5.95	0.0006
Error	16	7.06412187	0.44150762		
Corrected Total	29	41.23060387			

Source	DF	Sum of Squares	Mean Square
season	1	4.497715	4.497715
n	1	1.436271	1.436271
n*n	1	17.174003	17.174003
nitrogen	2	0.139910	0.069955
n*season	1	8.810046	8.810046
n*n*season	1	0.549765	0.549765
season*nitrogen	2	0.297404	0.148702
rep(season)	4	1.261367	0.315342
n*rep(season)	4	3.864242	0.966061
n*n*rep(season)	4	0.472544	0.118136
Residual	8	2.727335	0.340917

$$F = 8.810046 / 0.966061$$

$$F = 0.549765 / 0.118136$$

What have we learned from Application #1

- ANOCVA is a much easier approach to studying dosage responses than conventional methods (e.g., orthogonal polynomial coefficient analysis)
- Use of PROC MIXED ensures all tests are correct but the **METHOD=TYPE1** option (i.e. SS1) should be used for exact partitioning of the total trt SS into components due to linear, quadratic, cubic and quartic responses
 - Use of SS3 or REML would produce nonsense results because it doesn't make sense that linear effect is adjusted for quadratic or higher-order terms, etc.!!!

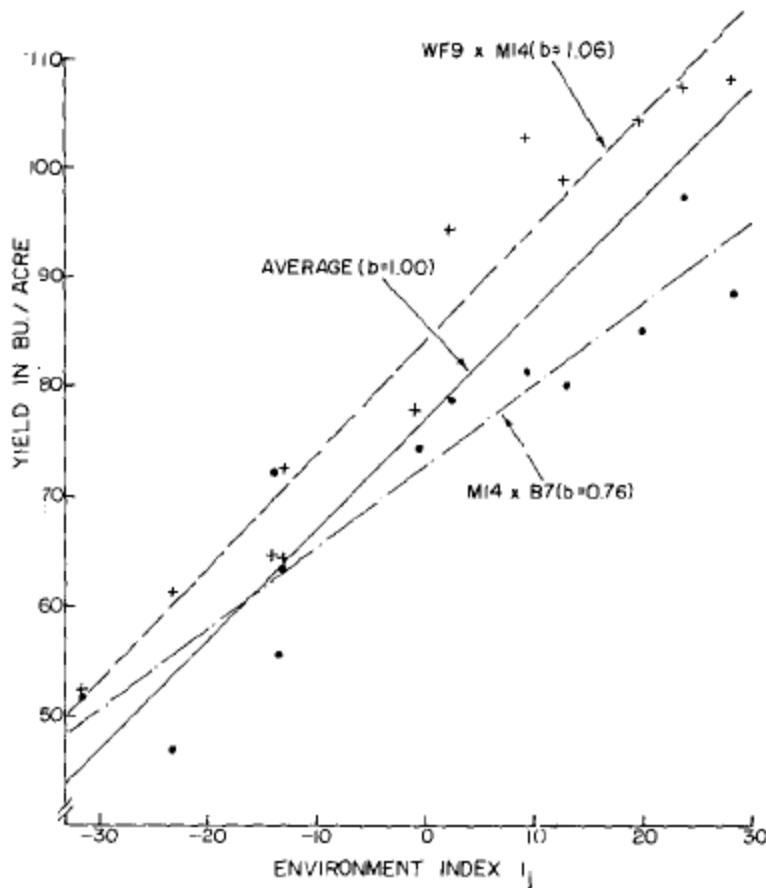
Application #2: Stability of treatments across environments

```
/*Littell et al. (2002) SAS for linear models, 4th
  edition. Pp. 420-431*/
/* A study was carried out to compare 3 treatments
  (trt) conducted at 8 locations (loc). At each
  location, a RCBD design was used, but the number of
  blocks varied: 3 blocks in locations 1-4, 6 blocks in
  locations 5 & 6, and 12 blocks in locations 7 & 8*/
```

Issues in this example:

- Preliminary analysis indicates sig trt x loc interaction
- Trt 1 is favored in 'poor' locations but trt 3 is favored in 'good' locations
- How stable are treatments over locations?

Stability analysis (Eberhart and Russell, 1966 Crop Sci 6: 36-40)



- An ANCOVA model:

$$Y_{ijk} = \mu + L_i + B(L)_{ij} + t_k + \beta_k I_i + (tL)_{ik} + e_{ijk}$$

Where I_i is a location index defined as the mean response over all obs in location i .

- A regression coefficient β_k is for linear regression on location index

Location index analysis

```
proc sort data=mloc;
  by loc;
proc means noprint data=mloc;
  by loc; var y;
  output out=env_indx mean=index;
run;
data all;
  merge mloc env_indx;
  by loc;
proc print data=all;
run;

proc mixed data=all;
  class loc blk trt;
  model y=trt trt*index/noint solution
    ddfm=satterth;
  random loc blk(loc) loc*trt;
  lsmeans trt/diff;
  contrast 'trt at mean index'
    trt 1 -1 0 trt*index 45.2 -45.2 0,
    trt 1 0 -1 trt*index 45.2 0 -45.2;
run;
```

- The PROC SORT and PROC MEANS statements generate a new data set, env_indx, which contains the means of Y by location, named INDEX;
- The MIXED program is for a mixed-model analysis where trt effects are fixed and location effects are random
- The term trt*index in the MODEL statement is to examine if there are similar responses for different treatments
- The NOINT and SOLUTION options allow easier interpretation of output

Covariance Parameter Estimates with and without INDEX as a covariate

- Without INDEX

Cov Parm	Estimate
loc	63.7513
blk(loc)	0
loc*trt	34.4302
Residual	29.1199

- With INDEX

Cov Parm	Estimate
loc	0
blk(loc)	0
loc*trt	0.8334
Residual	27.9211

A comparison between the two panels indicates that the linear regression of Y on INDEX (covariate) accounts for most of the variation among locations

Estimating and interpreting location index

Effect	trt	Estimate	Standard Error	DF	t Value	Pr > t
trt	1	12.4035	5.1377	32.7	2.41	0.0215
trt	2	17.0483	5.1377	32.7	3.32	0.0022
trt	3	-29.4519	5.1377	32.7	-5.73	<.0001
index*trt	1	0.6345	0.1128	29.2	5.62	<.0001
index*trt	2	0.6232	0.1128	29.2	5.52	<.0001
index*trt	3	1.7423	0.1128	29.2	15.44	<.0001

Least Squares Means

Effect	trt	Estimate	Standard Error	DF	t Value	Pr > t
trt	1	41.0822	0.8483	19.2	48.43	<.0001
trt	2	45.2182	0.8483	19.2	53.31	<.0001
trt	3	49.2975	0.8483	19.2	58.12	<.0001

The TRT estimates $\mu + t_k$ and the INDEX*TRT estimates β_k .

TRT + (INDEX*TRT)*(location index) is the expected trt mean at a given value of location index (e.g., at mean index = 45.2)

For trt 1, $12.4035 + (0.6345) * (45.2) = 41.4 \rightarrow$ LSmean

Estimating and interpreting location index

Effect	trt	Estimate	Standard Error	DF	t Value	Pr > t
trt	1	12.4035	5.1377	32.7	2.41	0.0215
trt	2	17.0483	5.1377	32.7	3.32	0.0022
trt	3	-29.4519	5.1377	32.7	-5.73	<.0001
index*trt	1	0.6345	0.1128	29.2	5.62	<.0001
index*trt	2	0.6232	0.1128	29.2	5.52	<.0001
index*trt	3	1.7423	0.1128	29.2	15.44	<.0001

The INDEX*TRT estimate is much larger for trt 3, but the intercept (TRT) is much smaller

Trt 3 performs worse in poor locations but better in good locations than trt 1 and 2. To verify, issue the following LSMEANS statements:

```
lsmeans trt/at index=30.9 diff;  
lsmeans trt/at means diff;  
lsmeans trt/at index=57.9 diff;
```

Trt means at poorest (index=30.9), average (index=45.2) and best (index=57.9) locations

Least Squares Means

Effect	trt	index	Estimate	Standard Error	DF	t Value	Pr > t
trt	1	30.90	32.0094	1.7935	36.6	17.85	<.0001
trt	2	30.90	36.3064	1.7935	36.6	20.24	<.0001
trt	3	30.90	24.3842	1.7935	36.6	13.60	<.0001
trt	1	45.20	41.0822	0.8483	19.2	48.43	<.0001
trt	2	45.20	45.2182	0.8483	19.2	53.31	<.0001
trt	3	45.20	49.2975	0.8483	19.2	58.12	<.0001
trt	1	57.90	49.1407	1.6936	19.6	29.02	<.0001
trt	2	57.90	53.1338	1.6936	19.6	31.37	<.0001
trt	3	57.90	71.4255	1.6936	19.6	42.17	<.0001

Differences of Least Squares Means

Effect	trt	_trt	index	Estimate	Standard Error	DF	t Value	Pr > t
trt	1	2	30.90	-4.2970	2.5363	36.6	-1.69	0.0987
trt	1	3	30.90	7.6252	2.5363	36.6	3.01	0.0048
trt	2	3	30.90	11.9221	2.5363	36.6	4.70	<.0001
trt	1	2	45.20	-4.1360	1.1996	19.2	-3.45	0.0027
trt	1	3	45.20	-8.2153	1.1996	19.2	-6.85	<.0001
trt	2	3	45.20	-4.0793	1.1996	19.2	-3.40	0.0030
trt	1	2	57.90	-3.9930	2.3951	19.6	-1.67	0.1114
trt	1	3	57.90	-22.2848	2.3951	19.6	-9.30	<.0001
trt	2	3	57.90	-18.2917	2.3951	19.6	-7.64	<.0001

What have we learned from Application #2?

- ANCOVA can be used to partition the total trt x location interaction variability into two parts, one due to the linear regression of Y on **location index** and the residual.
- If the linear regression is significant, then a focus should be on examining changes in trt responses at different locations (i.e., lowest to highest INDEX values)
 - Open question: **why do different trts perform differently over locations??**

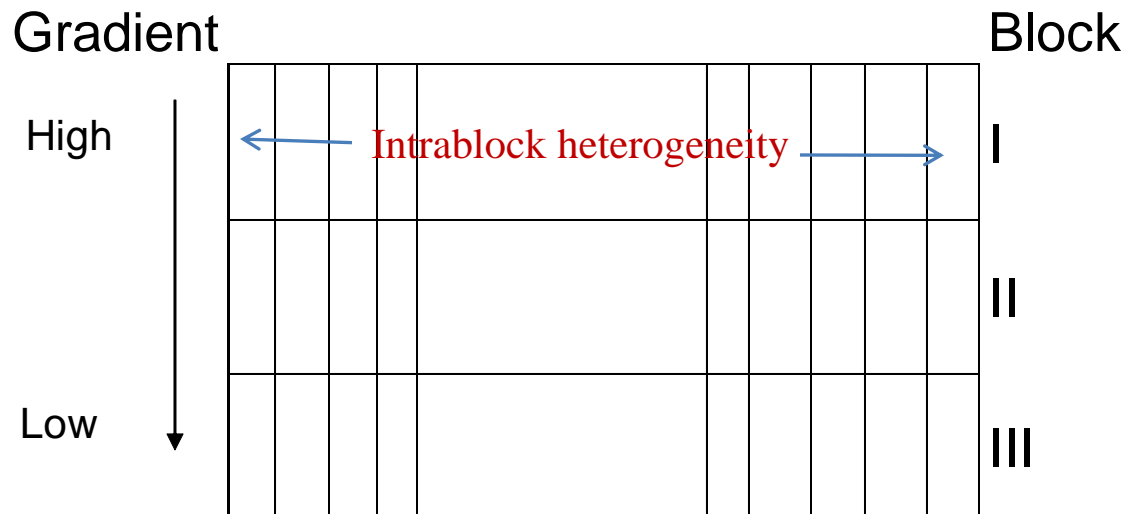
Application #3: Analysis of spatial variability

*/*Data taken from a field pea variety trial as described by Yang et al. (2004, Crop Sci. 44:49-55): Experimental design is a RCBD with 4 replications (blocks) and 28 varieties in each of the four blocks and thus a total of $4 \times 28 = 112$ plots*/*

```
data raw;
input PLOT_NR    BLOCK    ENTRY    YIELD;
datalines;
426 4    1        6419
310 3    1        6143
207 2    1        6219
121 1    1        6121
128 1    2        4934
422 4    2        5419
326 3    2        5203
215 2    2        4454
408 4    27       6482
210 2    27       5885
113 1    27       5698
323 3    27       5021
. . .
```

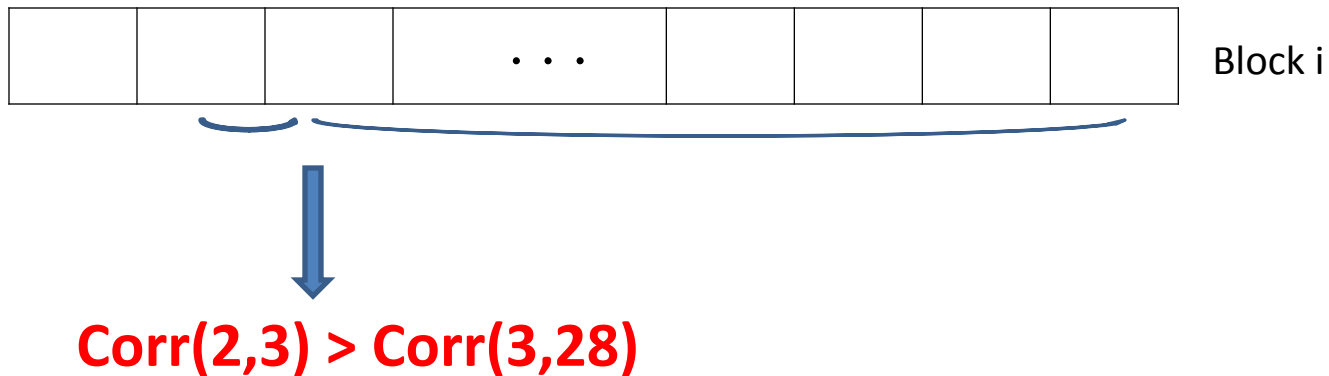
Problems with RCBD analysis

- In RCBD, proper blocking can reduce error by maximizing the difference between blocks and maintaining the plot-to-plot homogeneity within blocks,
- But blocking is ineffective if heterogeneity between plots does not follow a definite pattern (e.g., spotty soil heterogeneity; unpredictable pest incidence after blocking)
- **When block size is large [$> 8 - 12$ plots (trts) per block], intrablock heterogeneity is inevitable!**



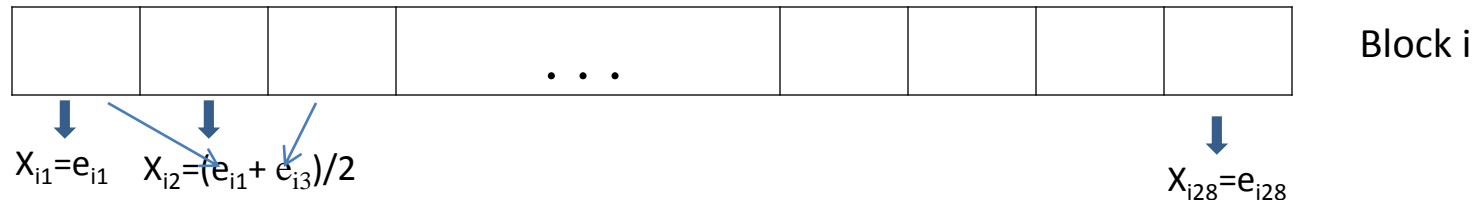
Nearest neighbor adjustment (NNA)

- To adjust a plot performance for spatial variability by using information from the immediate neighboring plots.
 - NNA would be effective if the correlation between residuals for two adjacent plots is higher than that for two plots far apart.



Statistical models for NNA

- For an observation (Y_{ij}) in block i and trt j ,
- RBCD model: $Y_{ij} = \mu + B_i + t_j + e_{ij}$
- NNA model: $Y_{ij} = \mu + B_i + t_j + \beta_i X_{ij} + \varepsilon_{ij}$
– where $X_{ij} = (e_{i,j-1} + e_{i,j+1})/2$



SAS code for RCBD vs NNA analysis

```
title "RCBD analysis";  
proc mixed  
class entry block;  
model yield= entry;  
random block;  
lsmeans entry/diff;  
ods output LSMeans=lsm_rcbd;  
run;
```

```
Title 'NNA analysis';  
proc mixed  
class entry block;  
model yield= entry x;  
random block;  
lsmeans entry/diff;  
ods output LSMeans=lsm_nna;  
run;
```

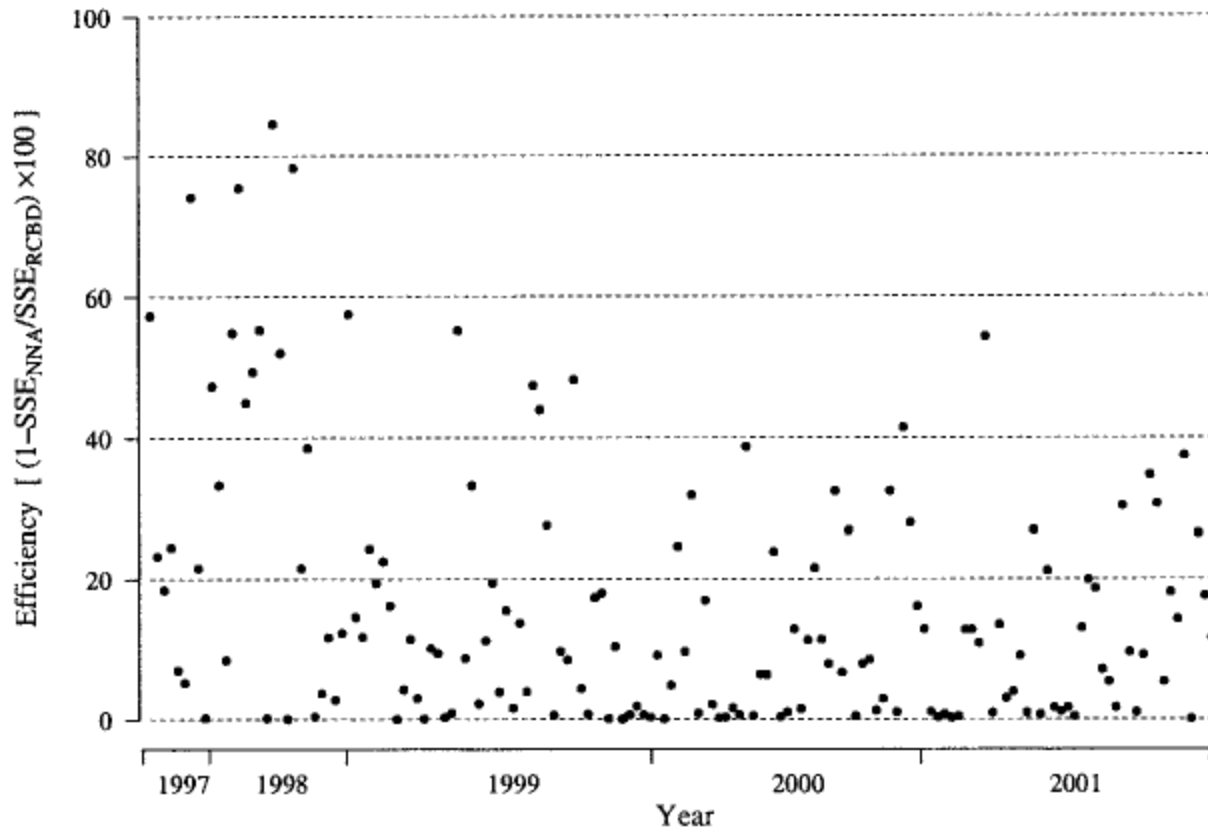
Error variance = 407639

Error variance = 224805

~45% error reduction!!

Yang et al. (2004) calculated efficiency of NNA over RCBD for 157 field trials across Alberta during 1997 to 2001.

Block sizes are larger in 1997-98 (28 – 32 varieties per block) than in 1999-2001 (12 – 22 varieties per block)... so NNA removed more error variation due to spatial heterogeneity in 1997-98 than in 1999-2001.



NNA affects variety ranking

Geno	RCBD	NNA	r_RCBD	r_NNA
1	6226	5796	2	5
2	5003	4948	21	24
3	5707	5478	8	12
4	6184	6017	3	2
5	5536	5195	11	20
6	5082	5100	20	21
7	5102	5392	19	15
8	5616	5654	10	7
9	5739	5901	6	4
10	6013	5972	4	3
11	4743	4832	26	25
12	5349	5394	15	14
13	4909	5030	24	23
14	5620	5697	9	6
15	5172	5260	18	18
16	5723	5539	7	11
17	5220	4799	17	26
18	5517	5567	12	9
19	6287	6053	1	1
20	4274	4593	27	27
21	5405	5540	13	10
22	5277	5268	16	17
23	4998	5401	22	13
24	2862	2863	28	28
25	5375	5390	14	16
26	4879	5038	25	22
27	5772	5600	5	8
28	4960	5229	23	19

What have we learned from Application #3?

- NNA is an application of ANCOVA technique using the information of neighboring plots in block designs
- NNA is able to account for plot-to-plot spatial variability within blocks, thereby further reducing experimental error
- Trt means may be ranked differently before and after adjustment for spatial variability

Take-home messages

- ANCOVA provides an easier analysis of dosage responses than conventional analyses (e.g., orthogonal polynomial coefficient analysis)
- ANCOVA can analyze stability of treatments across environments based on location index.
- ANCOVA can reduce experimental error by removing unpredictable spatial variability within blocks in designed experiments
- SAS software provides a convenient computing platform for ANCOVA