Analysis of Covariance

(ANCOVA)

Bruce A Craig

Department of Statistics
Purdue University

When to Use ANCOVA

- In experiment, there is a nuisance factor x that is
  1. Correlated with y
  2. Unaffected by treatment
- Can measure x but can't control it (otherwise block)
- Factor x then called a covariate or concomitant variable
- ANCOVA adjusts y for effect of covariate x
- Combination of regression and analysis of variance
- Without adjustment, effects of x on y
  - Will inflate σ^2
  - May alter trt mean comparisons (in extreme cases)

Examples

- **Pretest/Posttest score analysis**: The change in score y may be associated with current GPA. Also the posttest score y may be associated with the pretest score x. Analysis of covariance provides a way to "handicap" students.
- **Weight gain experiments in animals**: When comparing different feeds, the weight gain y may be associated with the dominance x of the animal. While it may be hard to control for dominance, it is not too difficult to measure.
- **Comparing competing drug products**: The effect of the drug y after two hours may be associated with the initial mental and physical shape of the subject. Variables describing mental and physical shape x at baseline may be used as covariates.

Model Description

- Consider single covariate in CRD
- Statistical model is

$$y_{ij} = \mu + \tau_i + \beta(x_{ij} - \overline{x}_{..}) + \epsilon_{ij} \quad \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots n_i \end{cases}$$

- Additional assumptions
  - x_{ij} not affected by treatment
  - x and y are linearly related
  - Constant slope across groups (can be relaxed)
  - Note: Subtracting off $\bar{x}_{..}$ is not needed (conceptual)

Estimation

- Conceptual Approach:
  - Fit one-way model ($y = \text{trt}$)
  - Fit one-way model ($x = \text{trt}$)
  - Regress residuals (residuals1 = residuals2)

    Provides estimate of slope after adjusting for trt
  - Model estimates are

$$\hat{\mu} = \overline{y}_{..}$$
$$\hat{\beta} = \sum\sum(y_{ij} - \overline{y}_{i.})(x_{ij} - \overline{x}_{i.}) / \sum\sum(x_{ij} - \overline{x}_{i.})^2$$
$$\hat{\tau}_i = \overline{y}_{i.} - \overline{y}_{..} - \hat{\beta}(\overline{x}_{i.} - \overline{x}_{..})$$

F Tests

- Test $H_0 : \tau_1 = \tau_2 = \ldots = \tau_a = 0$
  - Compare treatment means **after adjusting for differences among treatments due to differences in covariate levels**
  - Trt and covariate not orthogonal (order of fit matters)

$$F_0 = \frac{\mathrm{SS}(\mathrm{trt}|x)/a-1}{\mathrm{SS_E}/(N-a-1)}$$

- Test: $\beta = 0$
  - Sum of Squares regression (SS_x): $\hat{\beta}^2 \sum\sum(x_{ij} - \overline{x}_{i.})^2$

$$F_0 = \frac{\mathrm{SS}_x/1}{SS_{\mathrm{E}}/(N-a-1)}$$

Mean Estimates

- Adjusted treatment means
  - Estimate $\hat{\mu}_i = \hat{\mu} + \hat{\tau}_i = \overline{y}_{i.} - \hat{\beta}(\overline{x}_{i.} - \overline{x}_{..})$
  - Using the expected value of y when x is equal to the average covariate value
  - Can really use any value of x, just make sure it is reasonable for all factor levels
  - Variance: $\hat{\sigma}^2 \left(1/n + (\overline{x}_{i.} - \overline{x}_{..})^2 / \sum\sum(x_{ij} - \overline{x}_{i.})^2\right)$
- Pairwise differences
  - Estimate: $\hat{\tau}_i - \hat{\tau}_{i'} = \overline{y}_{i.} - \overline{y}_{i'.} - \hat{\beta}(\overline{x}_{i.} - \overline{x}_{i'.})$
  - Variance: $\hat{\sigma}^2 \left(2/n + (\overline{x}_{i.} - \overline{x}_{i'.})^2 / \sum\sum(x_{ij} - \overline{x}_{i.})^2\right)$

Analysis of Covariance

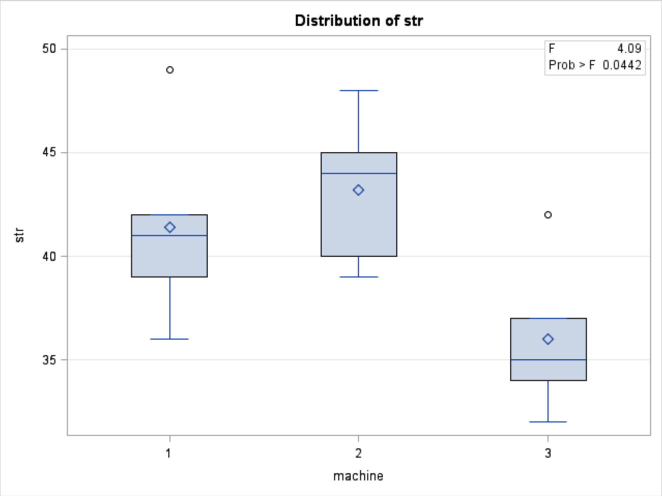
Table 15.10

- Looking at the breaking strength (in pounds) of a monofilament fiber produced by 3 different machines
- Known that strength depends on the fiber thickness
- Machines designed to keep thickness within specification limits but thickness will vary fiber to fiber
- Will consider diameter of the fiber as a covariate

SAS Code

```
data ancova;
 input machine str dia @@;
 datalines;
 1 36 20 1 41 25 1 39 24 1 42 25 1 49 32
 2 40 22 2 48 28 2 39 22 2 45 30 2 44 28
 3 35 21 3 37 23 3 42 26 3 34 21 3 32 15
 ;

symbol1 i=rl v=circle;
proc gplot;  plot str*dia=machine;

proc glm;
 class machine; model str = machine;
 lsmeans machine / adjust=tukey;

proc glm;
 class machine; model str = machine dia;
 lsmeans machine / adjust=tukey;
```

Boxplot



Distribution of str
F 4.09
Prob > F 0.0442
50
45
40
35
str
1
2
3
machine

SAS Output - No Covariate

```
                              The GLM Procedure
Dependent Variable: str
                                 Sum of
 Source                DF       Squares     Mean Square    F Value    Pr > F
 Model                  2   140.4000000      70.2000000       4.09    0.0442
 Error                 12   206.0000000      17.1666667
 Corrected Total       14   346.4000000

         R-Square     Coeff Var      Root MSE      str Mean
         0.405312      10.30664      4.143268      40.20000

 Source                DF     Type I SS     Mean Square    F Value    Pr > F
 machine                2   140.4000000      70.2000000       4.09    0.0442

Source                 DF   Type III SS     Mean Square    F Value    Pr > F
 machine                2   140.4000000      70.2000000       4.09    0.0442
```

SAS Output - No Covariate

```
                 The GLM Procedure
                  Least Squares Means
      Adjustment for Multiple Comparisons: Tukey-Kramer
                                               LSMEAN
            machine        str LSMEAN          Number
            1              41.4000000               1
            2              43.2000000               2
            3              36.0000000               3

                           str                   LSMEAN
                         LSMEAN     machine      Number
                  A        43.2     2                 2
                  A
             B    A        41.4     1                 1
             B
             B             36.0     3                 3
```

Difference Plot

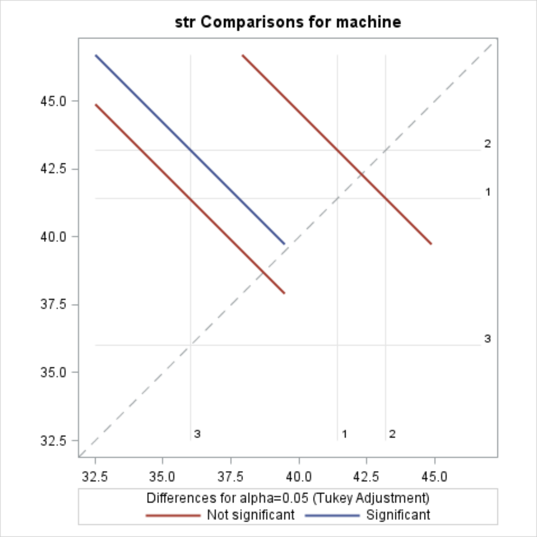
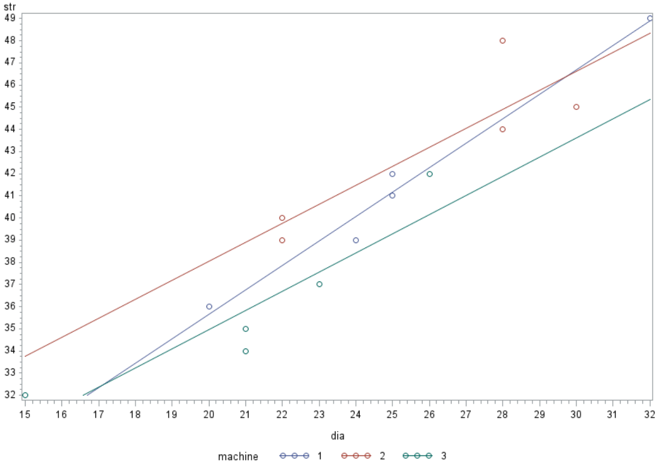


Table of Means

```
                          The MEANS Procedure
------------------------------ machine=1 ---------------------------------

 Variable    N         Mean       Std Dev       Minimum       Maximum
 str         5   41.4000000     4.8270074    36.0000000    49.0000000
 dia         5   25.2000000     4.3243497    20.0000000    32.0000000

------------------------------ machine=2 ---------------------------------

 Variable    N         Mean       Std Dev       Minimum       Maximum
 str         5   43.2000000     3.7013511    39.0000000    48.0000000
 dia         5   26.0000000     3.7416574    22.0000000    30.0000000

------------------------------ machine=3 ---------------------------------

 Variable    N         Mean       Std Dev       Minimum       Maximum
 str         5   36.0000000     3.8078866    32.0000000    42.0000000
 dia         5   21.2000000     4.0249224    15.0000000    26.0000000
```

Scatterplot



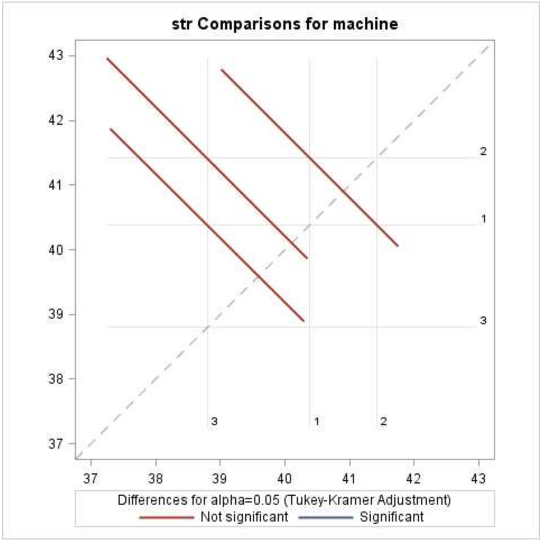
SAS Output

```
                          The GLM Procedure
Dependent Variable: str
                                   Sum of
 Source               DF          Squares     Mean Square    F Value    Pr > F
 Model                 3      318.4141104     106.1380368      41.72    <.0001
 Error                11       27.9858896       2.5441718
 Corrected Total      14      346.4000000

           R-Square     Coeff Var      Root MSE      str Mean
           0.919209      3.967776      1.595046      40.20000

 Source               DF        Type I SS     Mean Square    F Value    Pr > F
 machine               2      140.4000000      70.2000000      27.59    <.0001
 dia                   1      178.0141104     178.0141104      69.97    <.0001

 Source               DF      Type III SS     Mean Square    F Value    Pr > F
 machine               2       13.2838506       6.6419253       2.61    0.1181
 dia                   1      178.0141104     178.0141104      69.97    <.0001
```

SAS Output

```
                 The GLM Procedure
                   Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer
                                                LSMEAN
       machine       str LSMEAN                 Number
       1             40.3824131                      1
       2             41.4192229                      2
       3             38.7983640                      3

                     str                        LSMEAN
                  LSMEAN    machine             Number

       A        41.41922    2                        2
       A
       A        40.38241    1                        1
       A
       A        38.79836    3                        3

****Must use LSMEANS to get adjusted means ****
```

Difference Plot



Summary

- Positive linear association between diameter and strength. Are slopes constant? Will investigate shortly.
- Model including covariate better explains the data. Percent of explained variation jumps from 40.5% to 91.9%. MSE drops from 17.167 to 2.544.
- Because Machine 3 had narrower fibers, its adjusted mean strength is shifted upwards. Likewise Machine 2 had wider fibers so mean shifted downward
- No significant difference among the machines relies on assumption that diameter not different across machines

Nonconstant Slope in ANCOVA

- Statistical model for constant slope is

$$y_{ij} = \mu + \tau_i + \beta(x_{ij} - \overline{x}_{..}) + \epsilon_{ij} \quad \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots n_i \end{cases}$$

- Can allow for different slope by including interaction

$$y_{ij} = \mu + \tau_i + (\beta + (\beta\tau)_i)(x_{ij} - \overline{x}_{..}) + \epsilon_{ij} \quad \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots n_i \end{cases}$$

- In SAS, simply add interaction term into model
- Provides test for nonconstant slope

SAS Code

```
data ancova;
 input machine str dia @@;
 datalines;
 1 36 20 1 41 25 1 39 24 1 42 25 1 49 32
 2 40 22 2 48 28 2 39 22 2 45 30 2 44 28
 3 35 21 3 37 23 3 42 26 3 34 21 3 32 15
 ;

proc glm;
 class machine; model str = machine dia;
 lsmeans machine / adjust=tukey lines;

proc glm;
 class machine; model str = machine dia machine*dia;
 lsmeans machine / adjust=tukey lines;
run;
```

SAS Output

The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	321.1512879	64.2302576	22.90	<.0001
Error	9	25.2487121	2.8054125		
Corrected Total	14	346.4000000			

R-Square	Coeff Var	Root MSE	str Mean
0.927111	4.166509	1.674937	40.20000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
machine	2	140.4000000	70.2000000	25.02	0.0002
dia	1	178.0141104	178.0141104	63.45	<.0001
dia*machine	2	2.7371774	1.3685887	0.49	0.6293

Source	DF	Type III SS	Mean Square	F Value	Pr > F
machine	2	2.6641625	1.3320812	0.47	0.6367
dia	1	171.1192314	171.1192314	61.00	<.0001
dia*machine	2	2.7371774	1.3685887	0.49	0.6293

Regression Approach to ANCOVA

- Consider ANCOVA model with $a = 3$

$$y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \epsilon_j$$

$$\begin{aligned} j &= 1, 2, \ldots N \\ X_{1j} &= 1 \text{ if Trt 1 and } X_{1j} = -1 \text{ if Trt 3} \\ X_{2j} &= 1 \text{ if Trt 2 and } X_{2j} = -1 \text{ if Trt 3} \\ X_{3j} &= (x_j - \overline{x}_{..}) \end{aligned}$$

- Trt 1: $y_j = \beta_0 + \beta_1 + \beta_3(x_j - \overline{x}_{..}) + \epsilon_j$
- Trt 2: $y_j = \beta_0 + \beta_2 + \beta_3(x_j - \overline{x}_{..}) + \epsilon_j$
- Trt 3: $y_j = \beta_0 - \beta_1 - \beta_2 + \beta_3(x_j - \overline{x}_{..}) + \epsilon_j$
- Results in estimates
  $\hat{\mu} = \hat{\beta}_0 \quad \hat{\tau}_1 = \hat{\beta}_1 \quad \hat{\tau}_2 = \hat{\beta}_2 \quad \hat{\beta} = \hat{\beta}_3$

Analysis of Covariance

- Can incorporate covariate into any model
- For two factor model

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \beta(x_{ijk} - \overline{x}_{...}) + \epsilon_{ijk}$$

- Assume constant slope **for each *ij* combination**
- Can include interaction terms to vary slope
- Plot y vs x for each combination

Background Reading

- ANCOVA Model: Montgomery 15.3.1
- General Regression Significance Test: Montgomery 15.3.3