Analysis of Covariance

Design of Experiments - Montgomery
Section 15-3
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 (i.e., treat as a 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 $x$ may
  - Inflate $\sigma^2$
  - Alter treatment 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 $x$ may be used as covariates.
Model Description

• Consider single covariate in CRD

• Statistical model is

\[ y_{ij} = \mu + \tau_i + \beta(x_{ij} - \bar{x}.) + \epsilon_{ij} \]

\[ \begin{cases} 
  i = 1, 2, \ldots, a \\
  j = 1, 2, \ldots n_i 
\end{cases} \]

• Additional assumptions
  
  – \( x_{ij} \) not affected by treatment
  
  – \( x \) and \( y \) are linearly related
  
  – Constant slope across groups (can be relaxed)
Estimation

- Conceptual Approach:
  - Fit one-way model \( y = \text{trt} \)
  - Fit one-way model \( x = \text{trt} \)
  - Regress residuals \( \text{residuals1} = \text{residuals2} \)
    
    Provides estimate of slope after adjusting for trt
  
  - Model estimates are
    \[
    \begin{align*}
    \hat{\mu} &= \bar{y}_.
    
    \hat{\beta} &= \frac{\sum \sum (y_{ij} - \bar{y}_i)(x_{ij} - \bar{x}_i)}{\sum \sum (x_{ij} - \bar{x}_i)^2}
    
    \hat{\tau}_i &= \bar{y}_i. - \bar{y}_. - \hat{\beta}(\bar{x}_i. - \bar{x}_.)
    \end{align*}
    \]
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)
    
    \[ F_0 = \frac{\text{SS(trt}|x)/a - 1}{\text{SS}_E/(N - a - 1)} \]

- Test: $\beta = 0$
  
  - Sum of Squares regression ($\text{SS}_x$): $\hat{\beta}^2 \sum \sum (x_{ij} - \bar{x}_i)^2$
    
    \[ F_0 = \frac{\text{SS}_x/1}{\text{SS}_E/(N - a - 1)} \]
Mean Estimates

- Adjusted treatment means
  - Estimate $\hat{\mu}_i = \hat{\mu} + \hat{\tau}_i = \bar{y}_i - \hat{\beta}(\bar{x}_i - \bar{x}.)$
  - 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: $\hat{\sigma}^2 \left( \frac{1}{n} + (\bar{x}_i - \bar{x}.)^2 / \sum \sum (x_{ij} - \bar{x}_i)^2 \right)$

- Pairwise differences
  - Estimate: $\hat{\tau}_i - \hat{\tau}_i' = \bar{y}_i - \bar{y}_i' - \hat{\beta}(\bar{x}_i - \bar{x}_i')$
  - Variance: $\hat{\sigma}^2 \left( \frac{2}{n} + (\bar{x}_i - \bar{x}_i').^2 / \sum \sum (x_{ij} - \bar{x}_i)^2 \right)$
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
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

str

1 2 3

machine

ANCova
### SAS Output - No Covariate

The GLM Procedure

Dependent Variable: str

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>140.4000000</td>
<td>70.2000000</td>
<td>4.09</td>
<td>0.0442</td>
</tr>
<tr>
<td>Error</td>
<td>12</td>
<td>206.0000000</td>
<td>17.1666667</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>14</td>
<td>346.4000000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>R-Square</th>
<th>Coeff Var</th>
<th>Root MSE</th>
<th>str Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.405312</td>
<td>10.30664</td>
<td>4.143268</td>
<td>40.20000</td>
</tr>
</tbody>
</table>

Source | DF | Type I SS       | Mean Square   | F Value | Pr > F |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>machine</td>
<td>2</td>
<td>140.4000000</td>
<td>70.2000000</td>
<td>4.09</td>
<td>0.0442</td>
</tr>
</tbody>
</table>

Source | DF | Type III SS     | Mean Square   | F Value | Pr > F |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>machine</td>
<td>2</td>
<td>140.4000000</td>
<td>70.2000000</td>
<td>4.09</td>
<td>0.0442</td>
</tr>
</tbody>
</table>
### SAS Output - No Covariate

**The GLM Procedure**

**Least Squares Means**

Adjustment for Multiple Comparisons: Tukey-Kramer

<table>
<thead>
<tr>
<th>machine</th>
<th>str</th>
<th>LSMEAN</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>41.4000000</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>43.2000000</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>36.0000000</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>str</th>
<th>LSMEAN</th>
<th>machine</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>43.2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>41.4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>36.0</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
Difference Plot

str Comparisons for machine

Differences for alpha=0.05 (Tukey Adjustment)
- Red: Not significant
- Blue: Significant
# Table of Means

The MEANS Procedure

---

### machine=1

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>str</td>
<td>5</td>
<td>41.4000000</td>
<td>4.8270074</td>
<td>36.0000000</td>
<td>49.0000000</td>
</tr>
<tr>
<td>dia</td>
<td>5</td>
<td>25.2000000</td>
<td>4.3243497</td>
<td>20.0000000</td>
<td>32.0000000</td>
</tr>
</tbody>
</table>

### machine=2

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>str</td>
<td>5</td>
<td>43.2000000</td>
<td>3.7013511</td>
<td>39.0000000</td>
<td>48.0000000</td>
</tr>
<tr>
<td>dia</td>
<td>5</td>
<td>26.0000000</td>
<td>3.7416574</td>
<td>22.0000000</td>
<td>30.0000000</td>
</tr>
</tbody>
</table>

### machine=3

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>str</td>
<td>5</td>
<td>36.0000000</td>
<td>3.8078866</td>
<td>32.0000000</td>
<td>42.0000000</td>
</tr>
<tr>
<td>dia</td>
<td>5</td>
<td>21.2000000</td>
<td>4.0249224</td>
<td>15.0000000</td>
<td>26.0000000</td>
</tr>
</tbody>
</table>
**SAS Output**

The GLM Procedure

Dependent Variable: str

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>318.4141104</td>
<td>106.1380368</td>
<td>41.72</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>11</td>
<td>27.9858896</td>
<td>2.5441718</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>14</td>
<td>346.4000000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>machine</td>
<td>2</td>
<td>140.4000000</td>
<td>70.2000000</td>
<td>27.59</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>dia</td>
<td>1</td>
<td>178.0141104</td>
<td>178.0141104</td>
<td>69.97</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>machine</td>
<td>2</td>
<td>13.2838506</td>
<td>6.6419253</td>
<td>2.61</td>
<td>0.1181</td>
</tr>
<tr>
<td>dia</td>
<td>1</td>
<td>178.0141104</td>
<td>178.0141104</td>
<td>69.97</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>
## SAS Output

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

<table>
<thead>
<tr>
<th>machine</th>
<th>str LSMEAN</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40.3824131</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>41.4192229</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>38.7983640</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>str</th>
<th>LSMEAN</th>
<th>machine</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>41.41922</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>A</td>
<td>40.38241</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>38.79836</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

****Must use LSMEANS to get adjusted means ****
Difference Plot

str Comparisons for machine

Differences for alpha=0.05 (Tukey-Kramer Adjustment)
- Not significant
- Significant
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

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

• Can allow for different slope by including interaction

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

• 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

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>5</td>
<td>321.1512879</td>
<td>64.2302576</td>
<td>22.90</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>9</td>
<td>25.2487121</td>
<td>2.8054125</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>14</td>
<td>346.4000000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>machine</td>
<td>2</td>
<td>140.4000000</td>
<td>70.2000000</td>
<td>25.02</td>
<td>0.0002</td>
</tr>
<tr>
<td>dia</td>
<td>1</td>
<td>178.0141104</td>
<td>178.0141104</td>
<td>63.45</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>dia*machine</td>
<td>2</td>
<td>2.7371774</td>
<td>1.3685887</td>
<td>0.49</td>
<td>0.6293</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>machine</td>
<td>2</td>
<td>2.6641625</td>
<td>1.3320812</td>
<td>0.47</td>
<td>0.6367</td>
</tr>
<tr>
<td>dia</td>
<td>1</td>
<td>171.1192314</td>
<td>171.1192314</td>
<td>61.00</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>dia*machine</td>
<td>2</td>
<td>2.7371774</td>
<td>1.3685887</td>
<td>0.49</td>
<td>0.6293</td>
</tr>
</tbody>
</table>

R-Square: 0.927111  
Coeff Var: 4.166509  
Root MSE: 1.674937  
str Mean: 40.200000
Regression Approach to ANCOVA

- Consider ANCOVA model with \( a = 3 \)

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

\( j = 1, 2, \ldots N \)

\( X_{1j} = 1 \) if Trt 1 and \( X_{1j} = -1 \) if Trt 3

\( X_{2j} = 1 \) if Trt 2 and \( X_{2j} = -1 \) if Trt 3

\( X_{3j} = (x_j - \bar{x}_{..}) \)

- Trt 1: \( y_j = \beta_0 + \beta_1 + \beta_3(x_j - \bar{x}_{..}) + \epsilon_j \)

- Trt 2: \( y_j = \beta_0 + \beta_2 + \beta_3(x_j - \bar{x}_{..}) + \epsilon_j \)

- Trt 3: \( y_j = \beta_0 - \beta_1 - \beta_2 + \beta_3(x_j - \bar{x}_{..}) + \epsilon_j \)

- 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

\[ y_{ijk} = \mu + \tau_i + \beta_j + (\tau \beta)_{ij} + \beta(x_{ijk} - \bar{x}) + \epsilon_{ijk} \]

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