ANCOVA

Combining Quantitative and Qualitative Predictors

In an ANCOVA we try to adjust for differences in the quantitative variable.
For example, suppose that we were to compare men’s average faculty income to women’s average faculty income here at OU faculty.
- Looking for a difference involves an ANOVA
- Explaining the difference (if one is found) involves an ANCOVA

Explaining the Difference

• In trying to explain the difference between men and women, we may want to control for certain variables:
  - Experience
  - Rank
  - Performance record
  - Etc.
• That is, we would like to show that the difference is due to relevant performance criteria. If we can’t show that this is the case, then we have a serious discrimination problem.

What if?

• The ANCOVA is many ways is a what if analysis– what if men and women had the same amount of experience? Would we still see the difference in income?
• Inherent in this analysis is the possibility that the what if question is relevant. For example, it would be silly to compare basketball teams adjusting for the heights of players. This would be a meaningless comparison.

Difference Are Due to Experience

![Graph showing difference due to experience]

ANCOVA

![Graph showing ANCOVA analysis]
Difference Due to Something Besides Experience

Difference Are Due to Interaction

Situational and Individual Differences

• In the social sciences researcher use ANCOVA to adjust the results for individual differences.
• Suppose that you are looking at ethical decisions under a variety of situations (personal gain, accountability, etc.)
  – You would also like to see if certain individual difference variables (introversion, conscientiousness, cognitive style, etc.) moderate the situational results, you can adjust for these individual difference variables using an ANCOVA design.

SAS ANCOVA Setup

```
proc glm; class race;
model inc = educ race / solution;
means race / tukey;
lsmeans race / tdiff adj=tukey;
/* Note that contrast and estimate statements are based on the adjusted means */
contrast 'black vs white' race 1 0 -1;
estimate 'black vs white' race 1 0 -1; run;
proc glm data=anc; class race;
model inc= race;
estimate 'black vs white' race 1 0 -1;
contrast 'black vs white' race 1 0 -1;run;
```
### ANOVA: Unadjusted Means

<table>
<thead>
<tr>
<th>Blocks Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Income education</td>
<td>16</td>
<td>13.87</td>
<td>6.64</td>
<td>8.00</td>
<td>33.00</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>12.25</td>
<td>3.35</td>
<td>7.00</td>
<td>19.00</td>
</tr>
<tr>
<td>Hispanics</td>
<td>N</td>
<td>Mean</td>
<td>Std Dev</td>
<td>Minimum</td>
<td>Maximum</td>
</tr>
<tr>
<td>Income education</td>
<td>14</td>
<td>15.50</td>
<td>6.40</td>
<td>8.00</td>
<td>29.00</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>11.64</td>
<td>2.30</td>
<td>8.00</td>
<td>16.00</td>
</tr>
<tr>
<td>Whites</td>
<td>N</td>
<td>Mean</td>
<td>Std Dev</td>
<td>Minimum</td>
<td>Maximum</td>
</tr>
<tr>
<td>Income education</td>
<td>50</td>
<td>13.12</td>
<td>2.80</td>
<td>7.00</td>
<td>20.00</td>
</tr>
</tbody>
</table>

### ANOVA Results

(Comparing the Unadjusted Means)

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of 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>838.117</td>
<td>419.058</td>
<td>7.06</td>
<td>0.0173</td>
</tr>
<tr>
<td>Error</td>
<td>77</td>
<td>7662.370</td>
<td>98.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>79</td>
<td>8440.487</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Tukey on the Unadjusted Means

<table>
<thead>
<tr>
<th>race Comparison</th>
<th>Difference Between Means</th>
<th>Simultaneous 95% Confidence Limits</th>
<th>race</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 - 2</td>
<td>5.740</td>
<td>-1.440 - 12.920</td>
<td>***</td>
</tr>
<tr>
<td>3 - 1</td>
<td>7.365</td>
<td>0.544 - 14.188</td>
<td>***</td>
</tr>
<tr>
<td>2 - 3</td>
<td>-5.740</td>
<td>-12.920 - 1.440</td>
<td></td>
</tr>
<tr>
<td>2 - 1</td>
<td>1.625</td>
<td>-7.065 - 10.315</td>
<td></td>
</tr>
<tr>
<td>1 - 3</td>
<td>-7.365</td>
<td>-14.188 - 0.544</td>
<td>***</td>
</tr>
<tr>
<td>1 - 2</td>
<td>-1.625</td>
<td>-10.315 - 7.065</td>
<td></td>
</tr>
</tbody>
</table>

Comparisons significant at the 0.05 level are indicated by ***.

### ANCOVA Results

<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>race</td>
<td>2</td>
<td>838.117</td>
<td>419.058</td>
<td>7.06</td>
<td>0.0173</td>
</tr>
<tr>
<td>ed</td>
<td>1</td>
<td>3061.307</td>
<td>3061.307</td>
<td>51.23</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

### Adjusted Means for Income

We see that even after we adjust for education there is still a difference between the averages.

### Adjustment for Multiple Comparisons: Tukey-Kramer

Least Squares Means for Effect race

<table>
<thead>
<tr>
<th>Least Squares Means for Effect race t for H0: LSMean(i)-LSMean(j) / Pr &gt;</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>for H0: LSMean(i)=LSMean(j) / Pr &gt;</td>
<td></td>
</tr>
<tr>
<td>Dependent Variable: inc</td>
<td></td>
</tr>
<tr>
<td>coefficient</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>-1.04773</td>
</tr>
<tr>
<td>2</td>
<td>-2.43112</td>
</tr>
<tr>
<td>3</td>
<td>1.04773</td>
</tr>
</tbody>
</table>
Using the Solution option in SAS

| Parameter     | Estimate | Standard Error | t Value | Pr > |t|  |
|---------------|----------|----------------|---------|------|--------|
| Intercept     | -7.831   | B              | 4.2060  | -1.86| 0.0663 |
| Race 1 (Blacks) | -5.457   | B              | 2.2365  | -2.43| 0.0174 |
| Race 2 (Hispanics) | -2.466 | B              | 2.3816  | -1.04| 0.3036 |
| Race 3 (Whites) | 0.000    |                |        |      |        |
| sd            | 2.215    |                | 0.3095  | 7.16 | <.0001 |

Blacks

\[ y_1 = (-7.83 - 5.44) + 2.22x \]
\[ y_1 = -13.27 + 2.22x \]

Regression Equation for each Group

\[ y_1 = -13.27 + 2.22x \]
\[ y_2 = -10.3 + 2.22x \]
\[ y_3 = -7.83 + 2.22x \]

Adjusted Means

To obtain the adjusted means we use the regression equation for each group and the overall x mean

14.92 = -13.27 + 2.22(12.7) \( 14.84 \)
17.89 = -10.3 + 2.22(12.7) \( 17.81 \)
20.36 = -7.83 + 2.22(12.7) \( 20.28 \)

The Adjusted Means

SAS Type Sum of Squares for unequal n’s

<table>
<thead>
<tr>
<th>Source</th>
<th>SS I</th>
<th>SS II</th>
<th>SS III</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>SS(A) ( \mu )</td>
<td>SS(A) ( \mu, B )</td>
<td>SS(A) ( \mu, B, AB )</td>
</tr>
<tr>
<td>B</td>
<td>SS(B) ( \mu, A )</td>
<td>SS(B) ( \mu, A )</td>
<td>SS(B) ( \mu, A, AB )</td>
</tr>
<tr>
<td>A*B</td>
<td>SS(AB) ( \mu, A,B )</td>
<td>SS(AB) ( \mu, A,B )</td>
<td>SS(AB) ( \mu, A,B )</td>
</tr>
</tbody>
</table>
Unequal n’s Designs and Ancova Models

• Under the MCAR (Missing data complete at random) assumption:
  – SAS Type III Sum of Squares provides a test of the partial effects, all submodels are compared to the overall model,

\[
y_{ij} = \mu + \tau_j + \beta x_i + e_{ij} = \beta o_j + \beta x_i + e_{ij}
\]

Sequential Sum of Squares
SAS Type I

• SAS model statement: (testing the equality of slopes assumption in ancova)
  model y= trt cov trt*cov;
  SS(trt | µ)
  SS(cov | µ, trt)
  SS(trt*cov | µ, trt, cov)
  For Type I SS, the sum of all effects add up to the model SS:
  SS(trt)+SS(cov)+SS(trt*cov)+SS(error)=SS(total)
  SS’s are also independent

SAS Type II SS

• SAS model statement:
  model y= trt cov trt*cov;
  SS(trt | µ,cov)
  SS(cov | µ, trt)
  SS(trt*cov | µ, trt, cov)
  Type II SS do not necessarily add up to the model SS.
  The SS’s are not independent.

SAS Type III SS
Partial Sum of Squares

• SAS model statement:
  model y= trt cov trt*cov;
  SS(trt | µ, cov)
  SS(cov | µ, trt)
  SS(trt*cov | µ, trt, cov)
  Type III SS do not necessarily add up to the model SS.
  • The SS’s are not independent

Partial F Test
Type III SS

\[
F = \frac{(SS_p - SS_f)/1}{df_e} = \frac{(R_p^2 - R_f^2)}{(1 - R_f^2)/df_e}
\]

The reduced model is the full model minus the element being tested.

SAS ANCOVA Setup
Unequal Slopes Model

\texttt{proc glm; class race; model inc = educ race educ*race / solution; means race / tukey; lsmeans race / tdiff adj=tukey; run;}

Unequal Slopes
Comparing Means

Testing for Difference in Means for a Given Value

```sas
proc glm; class race;
  model inc = educ race educ*race / solution;
  means race / tukey;
  /* Comparison at the mean */
  lsmeans race / tdiff adj=tukey;
  /* Comparison at 10 grade level */
  lsmeans race / tdiff adj=tukey at educ=10;
  estimate 'one vs two' race 1 -1 0
    educ*race 10 -10 0;
run;
```