STAT362 Homework Assignment 13

Sharon O’Boyle

Problem 1: Problem 1, 7.11 p. 235

SAS Program

* Sharon O'Boyle;
* Stat 362;
* Homework Assignment 13;

* Problem 1, 7.11 p. 235 ;
* Topic: Analysis of Covariance;

* Create dataset;

```sas
data math;
input GROUP $ M_SCORE AGE @@;
datalines;
A 90 16 B 92 18 C 97 18 A 88 15 B 88 13 C 92 17 A 72 12 B 76 12 C 88 16 A 82 14 B 78 14 C 92 17 A 65 12 B 90 17 C 99 17 A 74 13 B 68 12 C 82 14 ;
run;
```

```sas
proc print data=math;
title 'math data';
run;
```

*** Part A - One-Way ANOVA ;

```sas
PROC ANOVA DATA = math;
class group;
MODEL m_score age = group;
means group / snk;
TITLE 'REGRESSION of m_score and age on group';
RUN;
QUIT;
```

*** Part B - look for homogeneity of score and age among the 3 groups ;

```sas
PROC GLM DATA = math;
class group;
MODEL m_score = age group age*group;
TITLE 'PROC GLM Homogeneity ';
RUN;
QUIT;
```

*** Part C - analysis of covariance;

```sas
PROC GLM DATA = math;
class group;
MODEL m_score = age group / SS3 ;
```
LSMEANS GROUP / PDIFF;
TITLE 'Analysis of Covariance';
RUN;
QUIT;

SAS Log

NOTE: Copyright (c) 2002-2010 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) Proprietary Software 9.3 (TS1M1)
Licensed to GEORGE MASON UNIVERSITY-SPA T&R, Site 70008900.
NOTE: This session is executing on the W32_7PRO platform.

NOTE: Updated analytical products:
SAS/STAT 9.3_M1, SAS/ETS 9.3_M1, SAS/OR 9.3_M1

NOTE: SAS initialization used:
   real time          26.53 seconds
   cpu time           2.74 seconds

1 2 * Sharon O'Boyle;
3 4 * Stat 362;
5 6 * Homework Assignment 13;
7 8 * Problem 1, 7.11 p. 235 ;
9 10 * Topic: Analysis of Covariance;
11 12 * Create dataset;
13 14 data math;
15 16 input GROUP $ M_SCORE AGE @@;
17 18 datalines;
19 20 ;
21 22 run;
23 24 proc print data=math;
25 26 NOTE: Writing HTML Body file: sashtml.htm
27 28 title 'math data';
29 30 run;
31
32 NOTE: There were 18 observations read from the data set WORK.MATH.
33 NOTE: PROCEDURE PRINT used (Total process time):
   real time          2.21 seconds
   cpu time           0.54 seconds
*** Part A - One-Way ANOVA 
PROC ANOVA DATA = math;
class group;
MODEL m_score age = group;
means group / snk;
TITLE 'REGRESSION of m_score and age on group';
RUN;
QUIT;

NOTE: PROCEDURE ANOVA used (Total process time):
real time 5.91 seconds
cpu time 0.78 seconds

*** Part B - look for homogeneity of score and age among the 3 groups ;
PROC GLM DATA = math;
class group;
MODEL m_score = age group age*group;
TITLE 'PROC GLM Homogeneity ';
RUN;
QUIT;

NOTE: PROCEDURE GLM used (Total process time):
real time 0.67 seconds
cpu time 0.12 seconds

*** Part C - analysis of covariance;
PROC GLM DATA = math;
class group;
MODEL m_score = age group /SS3 ;
LSMEANS GROUP / PDIFF;
TITLE 'Analysis of Covariance';
RUN;
QUIT;

NOTE: PROCEDURE GLM used (Total process time):
real time 1.10 seconds
cpu time 0.34 seconds
*** Part A - One-Way ANOVA;

REGRESSION of m_score and age on group

The ANOVA Procedure

Class Level Information
Class Levels Values
GROUP 3 A B C

Number of Observations Read 18
Number of Observations Used 18

REGRESSION of m_score and age on group

The ANOVA Procedure

Dependent Variable: M_Score

<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>558.1111111</td>
<td>279.055556</td>
<td>3.75</td>
<td>0.0479</td>
</tr>
<tr>
<td>Error</td>
<td>15</td>
<td>1116.833333</td>
<td>74.455556</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>17</td>
<td>1674.944444</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square Coeff Var Root MSE M_Score Mean
0.333212 10.26555 8.628763 84.05556

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Anova SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>GROUP</td>
<td>2</td>
<td>558.1111111</td>
<td>279.055556</td>
<td>3.75</td>
<td>0.0479</td>
</tr>
</tbody>
</table>

For M_Score (Math score) F= 3.75 and p=0.0479. So there is a significant result for math score.
Student-Newman-Keuls Test for M_SCORE

<table>
<thead>
<tr>
<th>SNK Grouping</th>
<th>Mean</th>
<th>N</th>
<th>GROUP</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>91.667</td>
<td>6</td>
<td>C</td>
</tr>
<tr>
<td>B</td>
<td>82.000</td>
<td>6</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>78.500</td>
<td>6</td>
<td>A</td>
</tr>
</tbody>
</table>

Means with the same letter are not significantly different.

SNK shows no difference in Math Score between Group A and Group B or between Group B and Group C. However there is a difference between Group A and Group C.

Dependent Variable: AGE

<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>26.33333333</td>
<td>13.16666667</td>
<td>3.52</td>
<td>0.0559</td>
</tr>
<tr>
<td>Error</td>
<td>15</td>
<td>56.16666667</td>
<td>3.74444444</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>17</td>
<td>82.50000000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square: 0.319192, Coeff Var: 13.04533, Root MSE: 1.935053, AGE Mean: 14.83333

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Anova SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>GROUP</td>
<td>2</td>
<td>26.33333333</td>
<td>13.16666667</td>
<td>3.52</td>
<td>0.0559</td>
</tr>
</tbody>
</table>

For AGE F= 3.52 and p=0.0559. So there is not a significant result for age at the .05 level (although it would be significant at the 0.1 level).

Student-Newman-Keuls Test for AGE

<table>
<thead>
<tr>
<th>SNK Grouping</th>
<th>Mean</th>
<th>N</th>
<th>GROUP</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>16.500</td>
<td>6</td>
<td>C</td>
</tr>
<tr>
<td>A</td>
<td>14.333</td>
<td>6</td>
<td>B</td>
</tr>
<tr>
<td>A</td>
<td>13.667</td>
<td>6</td>
<td>A</td>
</tr>
</tbody>
</table>

Means with the same letter are not significantly different.

SNK shows no difference in Age between between any of the Groups at .05 level.
*** Part B - Homogeneity of relationship among 3 groups; ***

<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>AGE</td>
<td>1</td>
<td>768.8994431</td>
<td>768.8994431</td>
<td>39.76</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>GROUP</td>
<td>2</td>
<td>79.2386314</td>
<td>39.6193157</td>
<td>2.05</td>
<td>0.1716</td>
</tr>
<tr>
<td>AGE*GROUP</td>
<td>2</td>
<td>77.0383234</td>
<td>38.5191617</td>
<td>1.99</td>
<td>0.1790</td>
</tr>
</tbody>
</table>

In the test for significant interaction between AGE and GROUP, the F-value=1.99 and p-value=0.1790. So there is no significant interaction and we can do the analysis of covariance.

*** Part C - Analysis of Covariance; ***

Dependent Variable: M_SCORE

<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>3</td>
<td>1365.867788</td>
<td>455.289263</td>
<td>20.62</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>14</td>
<td>309.076657</td>
<td>22.076904</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>17</td>
<td>1674.944444</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square Coeff Var Root MSE M_SCORE Mean
0.815471 5.589882   4.698607  84.05556

<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>AGE</td>
<td>1</td>
<td>807.7566766</td>
<td>807.7566766</td>
<td>36.59</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>GROUP</td>
<td>2</td>
<td>12.3240166</td>
<td>6.1620083</td>
<td>0.28</td>
<td>0.7606</td>
</tr>
</tbody>
</table>

For Age, F=36.59 and P <.0001, so the result for Age is very significant. ie Math Score is affected by age.

For Group, F=0.28 and P =0.7606, so the result for Group is not significant. ie Math Score is not affected by group.
Problem 2: Problem 2, 9.2 p. 316

SAS Program

* Problem 1 from Lesson 12, 9.1 p. 315 ;
* Topic: Multiple Regression;

* Create dataset;

data tomato;
input yield light water @@;
datalines;
12 1 1 20 2 2
9 1 1 16 2 2
8 1 1 16 2 2
13 1 2 18 3 1
15 1 2 25 3 1
14 1 2 20 3 1
16 2 1 25 3 2
14 2 1 27 3 2
12 2 1 29 3 2
;
run;

proc print data=tomato;
title 'tomato data';
run;

data tomato2;
set tomato;
light_int = light * 5;
run;

proc print data=tomato2;
title 'tomato2 data';
run;

PROC REG DATA = tomato2;
MODEL yield = light_int water;
title 'REGRESSION ON TOMATO DATA';
RUN;
QUIT;

* Problem 2 , 9.2 p. 316 ;
* Topic: Re-do 9.1 Multiple Regression using Dummy Variables;

data tomato3;
set tomato;

if light = 1 then L5=1; else L5 = 0; /* Reference- will not be used in model */
if light = 2 then L10=1; else L10 = 0;
if light = 3 then L15=1; else L15 = 0;

if water = 1 then W1=1; else W1 = 0; /* Reference- will not be used in model */
if water = 2 then W2=1; else W2 = 0;
run;

proc print data=tomato3;
title 'tomato3 data';
run;

PROC REG DATA = tomato3;
MODEL yield = L10 L15 W2;
TITLE 'REGRESSION ON TOMATO DATA using dummy variables';
RUN;
QUIT;
SAS Log

* Problem 1 from Lesson 12, 9.1 p. 315 ;
* Topic: Multiple Regression;

* Create dataset;

data tomato;
input yield light water @@;
datalines;
NOTE: SAS went to a new line when INPUT statement reached past the end of a line.
NOTE: The data set WORK.TOMATO has 18 observations and 3 variables.
NOTE: DATA statement used (Total process time):
  real time 0.00 seconds
  cpu time 0.01 seconds

RUN;
proc print data=tomato;
title 'tomato data';
run;
NOTE: There were 18 observations read from the data set WORK.TOMATO.
NOTE: PROCEDURE PRINT used (Total process time):
  real time 0.02 seconds
  cpu time 0.03 seconds

data tomato2;
set tomato;
light_int = light * 5;
run;
NOTE: There were 18 observations read from the data set WORK.TOMATO.
NOTE: The data set WORK.TOMATO2 has 18 observations and 4 variables.
NOTE: DATA statement used (Total process time):
  real time 0.02 seconds
  cpu time 0.01 seconds

proc print data=tomato2;
title 'tomato2 data';
run;
NOTE: There were 18 observations read from the data set WORK.TOMATO2.
NOTE: PROCEDURE PRINT used (Total process time):
  real time 0.02 seconds
  cpu time 0.03 seconds
PROC REG DATA = tomato2;
MODEL yield = light_int water;
TITLE 'REGRESSION ON TOMATO DATA';
RUN;
QUIT;

NOTE: PROCEDURE REG used (Total process time):
      real time           1.85 seconds
      cpu time            0.40 seconds

* Problem 2, 9.2 p. 316 ;
* Topic: Re-do 9.1 Multiple Regression using Dummy Variables;

data tomato3;
set tomato;
if light = 1 then L5 = 1; else L5 = 0; /* Reference- will not be used in model */
if light = 2 then L10 = 1; else L10 = 0;
if light = 3 then L15 = 1; else L15 = 0;
if water = 1 then W1 = 1; else W1 = 0; /* Reference- will not be used in model */
if water = 2 then W2 = 1; else W2 = 0;
run;

NOTE: There were 18 observations read from the data set WORK.TOMATO.
NOTE: The data set WORK.TOMATO3 has 18 observations and 8 variables.
NOTE: DATA statement used (Total process time):
      real time           0.01 seconds
      cpu time            0.00 seconds

proc print data=tomato3;
title 'tomato3 data';
run;

NOTE: There were 18 observations read from the data set WORK.TOMATO3.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.04 seconds
      cpu time            0.03 seconds

PROC REG DATA = tomato3;
MODEL yield = L10 L15 W2;
TITLE 'REGRESSION ON TOMATO DATA using dummy variables';
RUN;
QUIT;

NOTE: PROCEDURE REG used (Total process time):
      real time           1.39 seconds
      cpu time            0.40 seconds
SAS Output

[Note: Here is my original result from Problem 9.1]

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|--------------------|----------------|---------|------|---|
| Intercept | 1 | -1.83333            | 2.29689        | 0.80    | 0.4372 |
| light_int | 1 | 1.21667             | 0.14066        | 8.65    | <.0001 |
| water    | 1 | 4.55556             | 1.14845        | 3.97    | 0.0012 |

The original equation from Problem 9.1 is:

\[ \text{yield} = -1.83333 + (1.21667 \times \text{light\_int}) + (4.55556 \times \text{water}) \]

REGRESSION ON TOMATO DATA using dummy variables

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|--------------------|----------------|---------|------|---|
| Intercept | 1 | 9.55556            | 1.04485        | 9.15    | <.0001 |
| L10      | 1 | 3.83333            | 1.27967        | 3.00    | 0.0096 |
| L15      | 1 | 12.16667           | 1.27967        | 9.51    | <.0001 |
| W2       | 1 | 4.55556            | 1.04485        | 4.36    | 0.0007 |

New equation:

\[ \text{yield} = 9.55556 + (3.83333 \times \text{L10}) + (12.16667 \times \text{L15}) + (4.55556 \times \text{water}) \]

Note that the parameter estimate for the water variable stays the same using both methods. But the intercept and light parameters changed.