Lab Objectives
After today’s lab you should be able to:

1. Use PROC GLM to generate least squares means and differences with confidence limits, and to make pair-wise comparisons adjusted for multiple comparisons.
2. Use PROC GLM to run analysis of covariance.
3. Interpret output from PROC GLM.
4. Use PROC REG to run multiple linear regression.
5. Interpret output from PROC REG.
6. Output diagnostics (predicted values, residuals) from linear regression into a new dataset.
7. Use PROC GPLOT to create simple scatter plots and diagnostic plots for linear regression.
8. Begin to produce enhanced graphs using PROC GPLOT.

Recommended reading in Walker: Chapters 10-11
LAB EXERCISE STEPS:

Follow along with the computer in front…

1. Double-click on to open the SAS editor file “data creation code” which should be saved in your stats210 folder from last week; run the libname statement:

```
libname stats210 'C:\Documents and Settings\mitl- pc.LANE-LIB\My Documents\Stats210';
```

2. Using the Explorer Browser on the left hand side of your screen, double check that a stats210 library has been properly created, and that it contains the SAS dataset “runners”.

3. Try ANOVA for the outcome variable sumedi1 (though not perfectly normally distributed outcome…). And examine output.

```
proc anova data=stats210.runners;
   class mencat;
   model sumedi1=mencat;
run;
```

4. To figure out which groups are different after adjusting the p-value post-hoc for having done 3 pairwise comparisons (using a scheffe adjustment):

```
proc glm data= stats210.runners;
   class mencat;
   model sumedi1=mencat;
   lsmeans mencat/pdiff adjust=scheffe cl;
run;
```

Proc glm="General linear model”—more powerful than ANOVA...does ANOVA “plus”...we are actually making a linear regression model : “model sumedi1=mencat” with sumedi1 as the outcome and mencat as the categorical predictor.

“automatically adjust my p-values for all pairwise comparisons” using a scheffe adjustment…
Generates the same ANOVA table as before, plus the following:

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Scheffe

<table>
<thead>
<tr>
<th>mencat</th>
<th>SPINE LSMEAN</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16.0000000</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>20.6923077</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>7.3000000</td>
<td>3</td>
</tr>
</tbody>
</table>

mean sumedi1 score for each group

Least Squares Means for effect mencat
Pr > |t| for H0: LSMean(i)=LSMean(j)

<table>
<thead>
<tr>
<th>i/j</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.7912</td>
<td>0.4007</td>
<td>0.0064</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.7912</td>
<td>0.0064</td>
</tr>
<tr>
<td>3</td>
<td>0.4007</td>
<td></td>
<td>0.0064</td>
</tr>
</tbody>
</table>

After adjusting for multiple comparisons, only groups 2 and 3 (oligomenorrheic eumenorrheic) are significantly different at p<.05 level.

sumedi1 mencat LSMEAN 95% Confidence Limits

<table>
<thead>
<tr>
<th>mencat</th>
<th>LSMEAN</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16.0000000</td>
<td>3.948831  28.051169</td>
</tr>
<tr>
<td>2</td>
<td>20.692308</td>
<td>14.007522  27.377094</td>
</tr>
<tr>
<td>3</td>
<td>7.3000000</td>
<td>2.899535   11.700465</td>
</tr>
</tbody>
</table>

Difference Between Means
Simultaneous 95% Confidence Limits for
LSMean(i)-LSMean(j)

<table>
<thead>
<tr>
<th>i</th>
<th>j</th>
<th>Difference</th>
<th>Simultaneous 95% Confidence Limits for Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>-4.692308</td>
<td>-22.016236  12.631620</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>8.700000</td>
<td>-7.427699   24.827699</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>13.392308</td>
<td>3.331670   23.452945</td>
</tr>
</tbody>
</table>

95% confidence intervals for the mean sumedi1 scores. Note that the confidence interval for group 1 is wide, because it is a very small group; use interactive data analysisÆdistributionÆtablesÆfrequency counts to find that n=4.

95% Confidence intervals for the difference in means between group 1 and group j. Note that 2 vs. 3 does not cross 0.
5. Run the following code; note the change in p-values of differences if we hadn’t adjusted for multiple comparisons; also note that SAS gives you a warning!

```sas
proc glm data= stats210.runners;
  class mencat;
  model sumedi1=mencat;
  lsmeans mencat/pdiff cl;
run;
```

6. Controlling for confounders (ANCOVA)
Sometimes, you want to control for confounders. This requires ANCOVA *(analysis of covariance). We’ll return to this when we talk about regression. For now, use PROC GLM again and add confounders to your model.

```sas
proc glm data=stats210.runners;
  class mencat;
  model neck1=mencat pounds1 age;
  lsmeans mencat/pdiff adjust=tukey;
run;
```

**GIVES:**

The GLM Procedure

Dependent Variable: neck1

<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>4</td>
<td>0.25314374</td>
<td>0.06328593</td>
<td>5.28</td>
<td>0.0016</td>
</tr>
<tr>
<td>Error</td>
<td>42</td>
<td>0.50370337</td>
<td>0.01199294</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>46</td>
<td>0.75684711</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square  | Coeff Var | Root MSE | neck1 Mean |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.334471</td>
<td>11.90846</td>
<td>0.109512</td>
<td>0.919617</td>
</tr>
</tbody>
</table>

Overall ANOVA table. This says that at least some of the predictors in the model significantly explain differences (variation) in neck bone density (p<.0016).

Estimated standard deviation of neck bone density (average variability within groups) 

$$\sqrt{MSE} = \sqrt{0.0119929} = 0.1095$$

R-squared:

$$R^2 = \frac{SSModel}{TSS} = \frac{.25}{.75} = .33$$

An R-square of .33 means that 33% of the total variance in neck bone density is explained by age, weight, and menstrual group.

standard deviation of mean 

$$s_{100} \times 100 = .109512$$

$$s_{100} = 11.9$$

$$\bar{x}_{100} = s_{100}$$
The mean neck BMD each group would have if everyone were the same age (the average age of the whole sample) and the same weight (the average weight of the whole sample):
7. Now run the same model using PROC REG (multiple linear regression); here you must dummy code on your own. The difference is that you get out regression coefficients, but the overall ANOVA results are identical.

/**Run the same thing as above in PROC REG--do dummy coding on your own**/

data stats210.runners;
set stats210.runners;
if mencat=1 then amen=1; else amen=0;
if mencat=2 then olig=2; else olig=0;
run;

proc reg data=stats210.runners;
model neck1=amen olig pounds1 age/clm cli ;
run;

Translates to a regression model: \( \text{neck1} = \alpha + \beta_{\text{amen}}(1/0) + \beta_{\text{olig}}(1/0) + \beta_{\text{pounds}}(\text{pounds}) + \beta_{\text{age}}(\text{age}) \)

OUTPUT:

The REG Procedure
Model: MODEL1
Dependent Variable: neck1 neck1

Analysis of Variance

<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>4</td>
<td>0.25314</td>
<td>0.06329</td>
<td>5.28</td>
<td>0.0016</td>
</tr>
<tr>
<td>Error</td>
<td>42</td>
<td>0.50370</td>
<td>0.01199</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>46</td>
<td>0.75685</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 0.10951
R-Square 0.3345
Dependent Mean 0.91962
Adj R-Sq 0.2711
Coeff Var 11.90846

ANOVA table is identical to the ANOVA table from step 6, because it’s the same model!

In addition, though, PROC REG also gives you the regression coefficients="parameter estimates."

| Variable       | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------------|-------|----|--------------------|----------------|---------|------|-----|
| Intercept      |       | 1  | 0.65500            | 0.32228        | 2.03    | 0.0485 |
| amen           |       | 1  | -0.07166           | 0.06286        | -1.14   | 0.2608 |
| olig           |       | 1  | -0.04833           | 0.03670        | -1.32   | 0.1950 |
| pounds1        | pounds1 | 1  | 0.00479           | 0.00119        | 4.04    | 0.0002 |
| age            |       | 1  | -0.01639           | 0.01369        | -1.20   | 0.2382 |

FINAL MODEL:
\( \text{neck1} = .655 - .07166(1/0) - .04833(1/0) + .00479(\text{pounds}) - .01639(\text{age in years}) \)
Example: predicted neck bone density for an amenorrheic runner who is 25 years old and weighs 105 lbs:

\[ 0.655 - 0.07166(1) - 0.04833(0) + 0.00479(105) - 0.01639(25) = \frac{0.6765}{cm^2} \]

**Calculation of least squares means (compare with values in output from step 6, page 5 of this lab); the mean weight and age of the entire sample are: 20.8 years and 130.3 lbs:

“Least squares mean” for eumenorrheic group:

\[ 0.655 - 0 - 0 + 0.00479(130.3) - 0.01639(20.8) = 0.939 \frac{g}{cm^2} \]

For oligomenorrheic group:

\[ 0.655 - 0 - 0.04833(0) + 0.00479(105) - 0.01639(25) = 0.89 \frac{g}{cm^2} \]

For amenorrheic group:

\[ 0.655 - 0.07166(1) + 0 - 0.01639(25) = 0.867 \frac{g}{cm^2} \]

FROM page 5:

<table>
<thead>
<tr>
<th>Least Squares Means</th>
<th>Adjustments for Multiple Comparisons: Tukey-Kramer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.86742642</td>
</tr>
<tr>
<td>2</td>
<td>0.89075083</td>
</tr>
<tr>
<td>3</td>
<td>0.93908445</td>
</tr>
</tbody>
</table>

Note output from “CLM” and “CLI” options; each individual has a predicted neck1 value based on their weight, age, and menstrual status, and the regression equation:

<table>
<thead>
<tr>
<th>Dep Var</th>
<th>Predicted neck1</th>
<th>Std Error</th>
<th>Obs</th>
<th>Value Mean</th>
<th>Predict</th>
<th>95% CL Mean</th>
<th>95% CL Predict</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.1540</td>
<td>0.8876</td>
<td>0.0284</td>
<td>0.8303</td>
<td>0.9449</td>
<td>0.6593</td>
<td>1.1169</td>
<td>0.2664</td>
</tr>
<tr>
<td>2</td>
<td>0.8390</td>
<td>0.9849</td>
<td>0.0404</td>
<td>0.9033</td>
<td>1.0664</td>
<td>0.7493</td>
<td>1.2204</td>
<td>-0.1459</td>
</tr>
<tr>
<td>3</td>
<td>0.7870</td>
<td>0.8769</td>
<td>0.0270</td>
<td>0.8223</td>
<td>0.9314</td>
<td>0.6492</td>
<td>1.1045</td>
<td>-0.0899</td>
</tr>
</tbody>
</table>

For example, participant 1 is eumenorrheic, 21.96 years old, 123.6 pounds, and has a true neck density value of 1.154. Predicted value:

\[ 0.655 - 0.07166(0) - 0.04833(0) + 0.00479(123.6) - 0.01639(21.96) = \frac{0.8876}{cm^2} \]

Difference between observed and predicted = residual = 1.154 - 0.8876 = 0.2664
9. Type in the following code to make diagnostic plots of residuals against individual predictors. We will learn more about the features of PROC GPLOT in the coming weeks.

```sas
proc reg data=stats210.runners;
   model neck1=olig amen pounds1 age;
   output out=outdata p=predicted r=residual;
run;
```

Goptions reset=all; *resets graphing options;
sequence symbol1 value=circle color=red w=1 h=1;
   title1 'Plot of residuals against predictor oligomenorrheic';
   label olig='1 if oligomenorrheic';
   plot residual*olig /vaxis = -.5 to .5 by .1
       haxis = -1 to 2 by 1;
run; quit;

NOTE: Titles stay in effect until they are replaced by new ones or removed by entering a blank title: title1 '';

NOTE: label statements assigned to a variable within a PROC only are valid for duration of that PROC.

```sas
goptions reset=all; *resets graphing options;
symbol1 value=* color=blue w=1 h=1;
proc gplot data=outdata;
   title1 'Plot of residuals against predictor weight';
   label pounds1='Weight (pounds)';
   plot residual*pounds1 /vaxis = -.5 to .5 by .1
       haxis = 90 to 180 by 10;
run; quit;
```

DUE FROM LAB FIVE: Use right click→file→export as image to save last graph and email to me (or hand in printed copy).