

ST512 - Osborne
Lab 10

1. Refer to “beer.sas”. These are sodium contents in lager beer for six brands randomly sampled from the (large) population of beers produced in the U.S. and Canada. For each beer, $n = 8$ 12-ounce cans or bottles are sampled and measured for sodium content, y . Address each of these questions using the bits of code below:

- (a) Plot the sodium contents by brand of beer:

```
proc gplot;  
  plot y*brand;  
run;
```

- i. Which is a bigger source of variance, the random brand effect, or the bottle-to-bottle effect within brand?
 - ii. Which parameters are used to quantify these variabilities?
- (b) Obtain a one-way ANOVA table for the sodium contents.

```
proc glm;  
  class brand;  
  model y=brand;  
run;
```

- (c) Get the expected means square for the random brand effect using a random statement:

```
proc glm;  
  class brand;  
  model y=brand;  
  random brand;  
run;
```

- (d) Estimate the variance components using PROC VARCOMP

```
proc varcomp;  
  class brand;  
  model y=brand;  
run;
```

- (e) Do the estimated variance components agree with your characterization in part 1.(a)i.?
- (f) Do all of this using PROC MIXED, including confidence intervals for the model parameters μ , σ^2 and σ^2 :

```
proc mixed cl;  
  class brand;  
  model y=/cl s;  
  random brand;  
run;
```

2. Consider the percentage protein content of soybeans data listed below. This experiment involved randomly sampling 10 F_2 plants and randomization of 30 plots to seeds from the 10 F_2 plants. (3 plots per original plant.) Protein contents were measured on the subsequent offspring (F_3).

plant	plot1	plot2	plot3
1	42.4	41.0	39.6
2	28.6	36.3	42.2
3	43.2	42.1	40.2
4	40.8	41.0	38.9
5	41.0	38.3	41.1
6	39.4	39.5	37.2
7	39.6	40.4	38.9
8	38.1	38.3	37.9
9	35.9	36.1	35.6
10	39.6	39.9	39.7

The printed data in the text from which this example was taken have been known to disagree with those provided on the accompanying floppy diskette in the back of the book. The fitted values are given by the plant mean, $\hat{y}_{ij} = \bar{y}_{i+}$ and the residuals are the differences $e_{ij} = y_{ij} - \hat{y}_{ij}$.

- (a) Obtain some diagnostic plots:
- residuals against fitted values
 - residuals against fitted values using a different symbol for each plant.
 - histogram and normal plots of residuals
 - histogram and normal plots of raw protein contents
- (b) Can you find one observation that looks suspicious? Consider possible typos for this observation in the text. Try to fix them and then see if you can obtain the same ANOVA table as that given in the back of the book:

The GLM Procedure						
Class Level Information						
	Class	Levels	Values			
	plant	10	1	2	3	4 5 6 7 8 9 10

Dependent Variable: protein						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	9	74.0120000	8.2235556	4.25	0.0034	
Error	20	38.7066667	1.9353333			
Corrected Total	29	112.7186667				
	R-Square	Coeff Var	Root MSE	protein Mean		
	0.656608	3.528481	1.391163	39.42667		

Source	DF	Type I SS	Mean Square	F Value	Pr > F	
plant	9	74.0120000	8.2235556	4.25	0.0034	

(This may just be a transposition error, which happens all the time with real data. Changing 38.6 to 3.86 is more common than changing 38.6 to 28.6). To see the change I made to the data, see the file "soybeans.dat".

- (c) Use the `proc mixed` code in “soybeans.sas” to address these questions
- i. Construct an ANOVA table w/ complete with expected mean squares
 - ii. Is there evidence of variation among the offspring due to the parent (i.e. “plant”)?
 - iii. Estimate all variance components.
 - iv. Estimate the coefficient of variation for a protein content measurement.
 - v. Obtain a 95% confidence interval for the proportion of variation in protein content that can be ascribed to plant-to-plant variability (due to genetic attributes of the parent).
 - vi. Test (at level $\alpha = 0.05$) the hypothesis that the average protein content is 40%.
3. The data for a study of serum cholesterol can be found on the “datafiles” webpage under the name “cholest2.dat” and the SAS code can be found under “cholest2.sas.” The file contains serum cholesterol measurements from 8 randomly sampled healthy subjects. Two samples were made and assayed for each subject.

To better understand the notion of *intraclass* correlation or in this case, *intrasubject* correlation, obtain a plot of sample 2 measurements versus sample 1 measurements. To do this,

- (a) use PROC TRANSPOSE to create a dataset with two observations for each subject, called `sample1` and `sample2`.
 - (b) use PROC Gplot to plot `sample1` against `sample2`.
Imagine taking another serum sample from another normal subject and measuring the serum cholesterol on this sample twice. Wouldn't you want to model these two measurements as correlated?
 - (c) To get SAS to report the *intraclass* correlation coefficient, you can use the following statement
`REPEATED / SUBJECT=subject TYPE=CS rcorr;` However, note that the confidence limits on the variance component σ_T^2 for normal subject are symmetric about $\hat{\sigma}_T^2$ and are therefore different from the ones based on the Satterthwaite approximation discussed in class.
4. Fit a random effects model for the milk contamination with factors LAB and SAMPLE using “milk.sas”. (p. 196 of lecture notes) We'll discuss some of the output in lecture this week.