Analysis of Dietary Patterns over Freshman Year of College for Western Users of SAS® Software 2012
Chelsea Loftland, California Polytechnic, San Luis Obispo, CA

ABSTRACT

The FLASH at Cal Poly San Luis Obispo has been an ongoing research study aimed at discovering more about college students' health behavior and habits throughout their college career. The data for this study comes from a combination of an online survey and a physical assessment. Students join the study during their freshman year, and data are collected each fall and spring quarter through their college career.

Our current research focuses on an investigation of the progression of relationships between dietary patterns with sleep, stress, and exercise over students' freshman year. Our cohort consists of data from fall 2009 and spring 2010 quarters which represents Cal Poly students in the class of 2013.

Data management techniques consisted of routine data cleaning as well as the use of PROC SQL to combine the survey data from each quarter into one analytic data set. The data were then manipulated so that the final analytic data set had variables of interest that could be properly utilized. This required taking the multiple food consumption variables, which were grouped into seven food groups based on government health guidelines, and then transposing them to create one variable for food group and one for consumption.

Our analysis consists of a split-split plot design using PROC MIXED to create models that can be used to explain changes in consumption over the course of the year. In particular, interactions with food group and quarter are a priori specified questions of interest. Variables analyzed as potential predictors are sex, race, sleep quantity, sleep quality, international physical assessment questionnaire (IPAQ) score, perceived body mass index (BMI), and stress score. Our final analysis provides insight as to the associations behind college dietary patterns and other traits.

INTRODUCTION

This analysis is an investigation of changes in Cal Poly students' eating habits over freshman year. The motivation behind this was an interest in college students' lifestyles; college is the first time most students live on their own and it can be an important maturation period. College is stressful, exciting, liberating, and terrifying all at the same time. This distinctive life experience, along with a desire to handle big and messy data, led us to this research question.

The response variable to be analyzed is food consumption and the explanatory variables are: sex, race, quarter, food group, stress, exercise, BMI, sleep quality and quantity. These variables were chosen based on interest in how they could relate to a change in dietary patterns over the first year of college.

DATA

Cal Poly's Stride Program started the FLASH study in order to investigate college health. It is composed of an online survey and a physical assessment. Student responses are collected every fall and spring quarter, beginning in a students' freshman year and continuing throughout their college career. The goal is not only to examine aspects of students' health but to also see whether and how these change over time. This analysis uses the survey data for the class of 2013 only.

All data management and analysis was done using SAS® 9.2 software. The fall 2009 and spring 2010 survey data sets were merged using PROC SQL and kept only students who took the survey in both quarters. Removing students who were missing an entry for sex gave a total sample size of 740 observational units.

RESPONSE VARIABLE

A portion of the survey contained questions detailing consumption frequency of specific foods. These consumptions were converted into average consumption per month. The consumptions were then grouped into the seven categories in Table 1 based on government health guidelines [5].
Analysis of Dietary Patterns over Freshman Year of College

<table>
<thead>
<tr>
<th>Food Group</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grains</td>
<td>Cereals, deli sandwiches, white bread/English muffins/bagels, wheat bread/English muffins/bagels, and pasta.</td>
</tr>
<tr>
<td>Vegetables</td>
<td>Deli sandwiches, other potatoes, cooked or canned beans, lettuce or green leafy salad, and other vegetables.</td>
</tr>
<tr>
<td>Fruits</td>
<td>Fruits and 100% fruit juice.</td>
</tr>
<tr>
<td>Dairy</td>
<td>Milk</td>
</tr>
<tr>
<td>Protein</td>
<td>Deli sandwiches, cooked or canned beans, fish and seafood, red meat, chicken, peanut butter, and nuts/seeds.</td>
</tr>
<tr>
<td>Oils</td>
<td>Peanut butter, and mayonnaise or salad dressing.</td>
</tr>
<tr>
<td>Empty Calories</td>
<td>Bacon or sausage, hot dogs or other processed meat such as lunch meats, hamburgers, pizza, French fries, jelly/jam/preserves, chips, cake/muffins/pies/brownies/cookies/sweet pastries, frozen desserts, fruit-flavored drinks, regular soft drinks, and energy drinks.</td>
</tr>
</tbody>
</table>

**Table 1. Groupings of Food Consumption**

A food item could be placed in multiple categories based on the nutritional guidelines, i.e. peanut butter is grouped with both proteins and oils. However it should be noted that the food groups are created so that variables in the “empty calories” group cannot be in any other nutritional group. The food/drink variables that were not put in to any category are: coffee/specialty coffee beverages, diet soft drinks, and water because there were no dietary guidelines regarding these beverages.

**EXPLANATORY VARIABLES**

Along with sex, the following variables were of interest in exploring an association with changes in dietary patterns:

- **Race** (categorical): Due to the small percentage of non-white students at Cal Poly, race was collapsed into three categories: White/Asian, Hispanic, and other.
- **Stress Score** (quantitative): Stress score was created using questions that originated from Sheldon Cohen’s “Perceived Stress Scale” article (1983) that attempts to evaluate a person’s perceived level of stress. Answers were ranked then summed together to create the stress score variable so that a higher score indicates higher perceived level of stress. The possible range of score is 0-56.
- **Perceived Body Mass Index** (quantitative): Age, height, and weight were used to create perceived BMI for each person. According to the Center of Disease Control and Prevention, normal BMI is between 18.5 to 24.9 kg/m² [1]
- **International Physical Activity Questionnaire Score** (categorical): Using guidelines for IPAQ scoring [4], three categories were created:
  - **High**: 1+ hour per day of at least moderate-intensity activity or half an hour of vigorous-intensity activity
  - **Moderate**: Half an hour of at least moderate-intensity activity on most days
  - **High**: Not meeting any of the criteria for either of the previous categories
- **Sleep Quantity and Quality** (categorical): Sleep quality and sleep quantity were answered on a Likert scale of very bad, bad, average, good, and very good. Based on the small amount of people for some of these levels, very bad was combined with bad and very good was combined with good, resulting in three levels for both sleep variables.

**TRANSFORMATION**

This analysis is longitudinal, so only analyzing the difference in consumption between the two quarters will not suffice. FLASH is an ongoing study and one of the main goals of this project was to leave the analysis in a place where it could be continued with new data from the following quarter(s). To recognize this longitudinal nature the data were transformed. First, the data were stacked to create two observations per person—one per quarter, creating a quarter variable. Then the food consumptions were transposed using PROC TRANSPOSE to create a consumption variable and a food group variable. Any person with missing food consumption was removed.

This resulted in a new sample size of 509 students, each with measurements for all seven food groups across each quarter, giving 14 observations per student. An example of the data is given in Table 2.
### Table 2. Example of transformed data

The response variable is consumption and the main focus is how the explanatory variables of interest interact with food and quarter since that explains how that explanatory variable is related to changes in consumption over time for each food group.

To reduce computing time for some of the analyses tried, a subset of 100 students was extracted from the data. Only students with no missing values were used. This data set was used for trials of code that required long computing time and once final code was decided upon it would be run on the full data set.

### STATISTICAL ANALYSIS

To recap, we are interested in whether and how dietary patterns change over freshman year, particularly in relation to sex, race, quarter, food group, stress, exercise, BMI, sleep quality and quantity. There are many different ways to approach this, and this section will detail the methods investigated. This analysis is intriguing and complicated due to the nature of the data. Some variables, such as sex and ID, are constant for each student throughout all quarters. Other variables (stress score, perceived BMI, etc.) are recorded once per quarter. Furthermore, there are seven food groups that have consumptions recorded each quarter. The data is collected on the same students every other quarter, suggesting a repeated measures approach. Due to these different experimental units and a possible repeated nature, conducting the correct analysis is tricky.

#### HOTELLING’S PAIRED $T^2$ TEST

The first step in the analysis was to compare fall 2009 consumption to spring 2010 consumption in order to see if the consumptions significantly differ between the two quarters. A multivariate approach analyzing the vector of differences among the seven food consumption groups was preferred over a univariate paired t-test for each separate food group. This was done because consumptions of the food groups can be correlated with each other.

There are multiple ways to conduct Hotelling’s Paired $T^2$ Test. PROC IML was used at first in order to acquire results that could be compared with PROC REG and PROC GLM [4]. This was done to verify that the analysis was doing what we expected. The original data of 740 students that hadn’t been transposed was used for this test. After the $T^2$ test the transposed data is used.

The codes for the three methods are as follows:

```r
proc iml;
    start hotelling;
    mu0=(0, 0, 0, 0, 0, 0, 0);
    one=j(nrow(x),1,1);
    ident=i(nrow(x));
    ybar=x`*one/nrow(x);
    s=x`*(ident-one*one`/nrow(x))*x/(nrow(x)-1.0);
    print mu0 ybar;
    print s;
    t2=nrow(x) *(ybar-mu0) `*inv(s) *(ybar-mu0);
    f= (nrow(x)-ncol(x)) *t2/ncol(x)/(nrow(x)-1);
    df1=ncol(x);
    df2=nrow(x)-ncol(x);
    pvalue=1-probf(f,df1,df2);
    print t2 f df1 df2 pvalue;
    finish;
use hotelling;
    read all var{graindiff vegdiff fruitdiff dairydiff proteindiff oildiff emptycaldiff} into x;
run hotelling;quit;
```
proc reg data=flash.finaldataset;
  model graindiff vegdiff fruitdiff dairydiff proteindiff oildiff emptycaldiff = ;
  mtest intercept;
run;quit;

proc glm data=flash.finaldataset;
  model graindiff vegdiff fruitdiff dairydiff proteindiff oildiff emptycaldiff = ;
  manova h=intercept;
run;quit;

Hotelling's Paired $T^2$ Test yielded an overall p-value <0.0001, providing evidence that the true mean food consumptions are jointly significantly different for fall 2009 and spring 2010. The next step it to explore an association between this change in consumption and the explanatory variables of interest.

**MULTIVARIATE REPEATED MEASURES**

The first approach treats food and quarter both as repeated measures within a person and uses an unstructured covariance matrix. The intention was for this analysis to capture the idea that multiple food groups were being recorded across multiple quarters.

proc mixed data=flash._100_2;
  class Food ID Sex Quarter;
  model Consumption = Food|Quarter|Sex StressScore;
  repeated Food Quarter / type=un@un subject=ID;
run;

The repeated statement defines what has been repeated within each subject (here stated as ID). The type option is used to request the unstructured covariance matrix.

A segment of the output generated is given below.

<table>
<thead>
<tr>
<th>Effect</th>
<th>Num</th>
<th>Den</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food</td>
<td>6</td>
<td>588</td>
<td>31.96</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Quarter</td>
<td>1</td>
<td>98</td>
<td>1.26</td>
<td>0.2649</td>
</tr>
<tr>
<td>Food*Quarter</td>
<td>6</td>
<td>588</td>
<td>3.12</td>
<td>0.0051</td>
</tr>
<tr>
<td>Sex</td>
<td>1</td>
<td>98</td>
<td>0.32</td>
<td>0.5701</td>
</tr>
<tr>
<td>Food*Sex</td>
<td>6</td>
<td>588</td>
<td>3.93</td>
<td>0.0007</td>
</tr>
<tr>
<td>Sex*Quarter</td>
<td>1</td>
<td>98</td>
<td>0.05</td>
<td>0.8260</td>
</tr>
<tr>
<td>Food<em>Sex</em>Quarter</td>
<td>6</td>
<td>588</td>
<td>1.15</td>
<td>0.3290</td>
</tr>
<tr>
<td>StressScore</td>
<td>1</td>
<td>1273</td>
<td>0.42</td>
<td>0.5173</td>
</tr>
</tbody>
</table>

**Output 1. Portion of output from the multivariate repeated measures analysis**

Covariates like stress score give incorrect degrees of freedom that are too high. The 1273 suggests that stress score is being analyzed as if it had been recorded along with each food consumption rather than once per quarter. This is incorrect so we must look at the analysis a different way.

**STRIP PLOT**

The next method attempted was a strip plot analysis. This treats a person as the whole plot unit, quarter as the row, food type as the column, and food by quarter as the cell. This was considered due to the lack of randomization of food and quarter.

proc mixed data=flash._100_2 method=type3 lognote maxiter=10;
  class ID Sex Quarter Food;
  model Consumption = Sex|Quarter|Food StressScore|Food|Quarter|Sex;
  random ID(Sex) Food*ID(Sex) StressScore*Quarter*ID(Sex);
run;
The random statement is used to define the multiple error terms since variables were measured differently and therefore do not have a common experimental unit. These error terms can be found below in Table 3.

Again, adding covariates raised issues. Including stress score into the model would not produce results and removing method=type3 would run but with incorrect degrees of freedom. As seen in Output 2 the 564 degrees of freedom for quarter look too high.

<table>
<thead>
<tr>
<th>Effect</th>
<th>Num</th>
<th>Den</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>1</td>
<td>98</td>
<td>2.20</td>
<td>0.1415</td>
</tr>
<tr>
<td>Quarter</td>
<td>1</td>
<td>564</td>
<td>0.00</td>
<td>0.9886</td>
</tr>
<tr>
<td>Sex*Quarter</td>
<td>1</td>
<td>564</td>
<td>2.61</td>
<td>0.1067</td>
</tr>
<tr>
<td>Food</td>
<td>6</td>
<td>588</td>
<td>8.95</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Sex*Food</td>
<td>6</td>
<td>588</td>
<td>3.83</td>
<td>0.0009</td>
</tr>
</tbody>
</table>

Output 2. Portion of output from the strip plot analysis using the default method=REML

Since the degrees of freedom for quarter appear too high, the correct degrees of freedom for a strip plot analysis of the 100 student data set are computed and given in Table 3 with error terms bolded. This shows that quarter is being incorrectly tested using the overall error term.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>1</td>
</tr>
<tr>
<td>ID(Sex)</td>
<td>98</td>
</tr>
<tr>
<td>Quarter</td>
<td>1</td>
</tr>
<tr>
<td>Quarter * Sex</td>
<td>1</td>
</tr>
<tr>
<td>Quarter*ID(Sex)</td>
<td>94</td>
</tr>
<tr>
<td>Food</td>
<td>6</td>
</tr>
<tr>
<td>Sex * Food</td>
<td>6</td>
</tr>
<tr>
<td>Food * ID(Sex)</td>
<td>588</td>
</tr>
<tr>
<td>Quarter * Food</td>
<td>6</td>
</tr>
<tr>
<td>Quarter * Sex * Food</td>
<td>6</td>
</tr>
<tr>
<td>Error</td>
<td>564</td>
</tr>
<tr>
<td>Total</td>
<td>1399</td>
</tr>
</tbody>
</table>

Table 3. ANOVA table for strip plot analysis of 100 student data set

The ddf= option in the model statement allows us to redefine the denominator degrees of freedom. However, it changes the denominator degrees of the F critical value that the F statistic is being compared to, not the degrees of freedom used to calculate the F statistic. Therefore it is not an appropriate fix to the problem.

SPLIT PLOT WITH REPEATED MEASURES

A split plot with repeated measures design treats each quarter as the split plot factor and food as a repeated measure within each quarter. The ddfm= option in the model statement defines Kenward Roger as the method for calculating denominator degrees of freedom. This is considered to be the correct analysis, however errors occurred when trying to use the random and repeated statements together in PROC MIXED. An example is the code below, which yielded an infinite likelihood error.

```plaintext
proc mixed data=flash._100_2 maxiter=10;
  class id sex food quarter;
  model consumption = sex|food|quarter / ddfm=kr;
  random id(sex) quarter*id(sex);
  repeated food / type=un subject=quarter*id(sex);
run;
```

WARNING: Stopped because of infinite likelihood.

Through research, we found that this is an ongoing and common problem with using the REPEATED statement in PROC MIXED on large data sets that has yet to be solved.
SPLIT-SPLIT PLOT

Split-split plot is the final analysis decided upon based on the issues stated above. Due to limited time and computing resources, further investigation had to be put on hold. The most recent results from this analysis will be presented and interpreted.

Under the split-split plot design the whole plot unit is a person, the split plot unit is food*ID, and the split-split plot unit is food*quarter*ID. By analyzing the data using a split-split plot design we assume that food and quarter were both randomized. This is not true. As stated before, issues exist in trying to model this data and this is still a work in progress. The results from this analysis will be presented, but it is important to keep this violation of assumptions in mind.

```plaintext
proc mixed data=flash.narrow lognote;
   class id sex food quarter;
   model consumption = sex|food|quarter
                       stressscore
                       stressscore*food
                       / ddfm=kr solution;
   random id(sex) food*id(sex);
   lsmeans sex*food*quarter / adj=tukey;
run;
```

The random statement is again used to define the error terms from multiple experimental units. LSMEANS provides pairwise comparisons using a Tukey adjustment option. The lognote option sends notes to the log giving more specific details of what the analysis is doing as it runs.

A large amount of output is generated from the code above. The portion containing the main effect p-values is given in Output 3.

<table>
<thead>
<tr>
<th>Effect</th>
<th>Num DF</th>
<th>Den DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>1</td>
<td>513</td>
<td>2.22</td>
<td>0.1365</td>
</tr>
<tr>
<td>Food</td>
<td>6</td>
<td>5960</td>
<td>54.22</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Sex*Food</td>
<td>6</td>
<td>3079</td>
<td>9.26</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Quarter</td>
<td>1</td>
<td>3546</td>
<td>2.82</td>
<td>0.0934</td>
</tr>
<tr>
<td>Sex*Quarter</td>
<td>1</td>
<td>3546</td>
<td>0.13</td>
<td>0.7220</td>
</tr>
<tr>
<td>Food*Quarter</td>
<td>6</td>
<td>3546</td>
<td>0.90</td>
<td>0.4924</td>
</tr>
<tr>
<td>Sex<em>Food</em>Quarter</td>
<td>6</td>
<td>3546</td>
<td>2.40</td>
<td>0.0259</td>
</tr>
<tr>
<td>StressScore</td>
<td>1</td>
<td>2628</td>
<td>8.89</td>
<td>0.0029</td>
</tr>
<tr>
<td>StressScore*Food</td>
<td>6</td>
<td>6435</td>
<td>2.32</td>
<td>0.0307</td>
</tr>
</tbody>
</table>

Output 3. Portion of split-split plot output

The three way interaction of food, sex, and quarter is illustrated in Figure 1. The graph was created using JMP® 9 software.
Figure 1. Plot of consumption vs. quarter by sex and food group

Since interest is in how consumption changed over time, quarter is along the x-axis with mean consumption along the y-axis. Dashed lines are female and solid lines are male. Additionally, each food group has its own color. A flat line suggests no significant change in consumption between the quarters. The dashed red line with a negative slope from fall 2009 to spring 2010 suggests that females tended to eat less dairy in spring than in fall.

The interaction stress and food is seen in Figure 2. The graph was created using Microsoft Excel using the coefficients from the split-split plot output.

Figure 2. Plot of stress by food
In Figure 2 stress is plotted against mean consumption and each line represents a food group. Since quarter and sex were not significant, these variables were held constant at males in spring 2010. This plot suggests that food consumption tends to decrease as stress increases, with an exception of empty calories. This is interesting since it contradicts the notion of “stress eating”.

CONCLUSION

With these results, there is evidence that how consumption changed from fall 2009 to spring 2010 significantly differs between males and females and among food groups. There is also evidence that stress relates to changes in consumption over time.

Please note that there is work left to do with this analysis and should therefore be interpreted with caution. In addition, this is a voluntary response study that did not utilize random sampling. This is a limitation with the FLASH study that hinders the ability to generalize to all Cal Poly students.

Multiple analyses were tried in order to address the multiple experimental units as well as the lack of randomization of the effects. This was done by mainly utilizing the random and repeated statements in PROC MIXED. However many of these methods would not use the correct error term in the analysis or would not run.

As a work in progress, there is still much to be done. The appropriate method of analysis needs to be chosen using more powerful computing resources, and then the additional explanatory variables of interest stated earlier can be included. Another idea is to bootstrap the data in an attempt to avoid issues associated with large data sets. Once this is accomplished the analysis can continue with the data from subsequent quarters.

REFERENCES


ACKNOWLEDGMENTS

Thank you to Cal Poly’s STRIDE program for the data and to SAS for the use of SAS/STAT® 9.2 software and JMP® 9 software. Very special thanks to Dr. Karen McGaughey, Dr. Soma Roy, and Becky Ottesen for all of their guidance and support.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

    Name: Chelsea Loomis Lofland
    Address: 557 Sunnymount Ave
    City, State ZIP: Sunnyvale, CA 94087
    Work Phone: (408)242-2322
    E-mail: Chelsea.Loomis.Lofland@gmail.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.