Paper 60-2018

"Training Wheels" for learning mixed models

YuTing Tian, M.S. Statistics , WCU PA
Russ Lavery, Contractor

ABSTRACT

This paper intends to provide insights, rather than math, as part of an introduction to mixed models. It's intended for someone just starting out to learn mixed models and wants to show the link between one-way ANOVA, two-way ANOVA and mixed models.

Along the way, it is a good explanation of fixed effect ANOVA. There is a thread that links all of the sections of this paper together. That thread is the idea that error, or variability, is the inability of a process to repeat its mean.

Another deliverable of this paper is that it contains worked examples of the matrix calculations and, in the appendix, SAS code that displays the matrices that appear in typical explanations of mixed models.

INTRODUCTION

The authors acknowledge that some of the figures are rather complicated and small. It is suggested that a reader might print this paper and download the PDF. Where printed numbers are hard for a particular reader to read, the PDF can be blown up to 200% making the details of figures readable.

This paper is divided into the following _9_ sections:

1) A brief review of model history

2) One-way fixed variable ANOVA- linking the formula to SAS via hand calculation

3) Two-way Fixed ANOVA- linking the formula to SAS via hand calculation

4) One Way Fixed ANOVA using PROC ANOVA & PROC MIXED

   → Fixed ANOVA-using two different PROCS → Same Result

5) One-way Fixed ANOVA vs. Two-way Fixed ANOVA

   → 2nd x variable reduces SSE using Fixed Effect formula

6) TWO Way Fixed ANOVA – done using PROC GLIMMIX→matrix formula

7) An introduction to random variables

8) Two-way mixed model – City is Random - using PROC Mixed, PROC GliMix and Matrix Formula

9) A warning about convergence

1) A BRIEF REVIEW OF MODEL HISTORY

The chart in Figure 1 illustrates the historical relationship between the different PROCs that we will use in this paper.

PROC Reg and PROC ANOVA were programmed early in SAS's history and only do linear combinations of fixed effects. Proc Logistic came later and does a similar analysis after SAS has applied a logit function to the Y value.

These three PROCs were combined PROC GLM – a very versatile PROC that only handles fixed effects.

PROC Mixed was created to handle models with both fixed effects and random effects but did not have the ability to do transforms on the Y variable.

GlimMix united these two development streams. It can model fixed or random X variables as well as applying a variety of transforms to the Y variable.

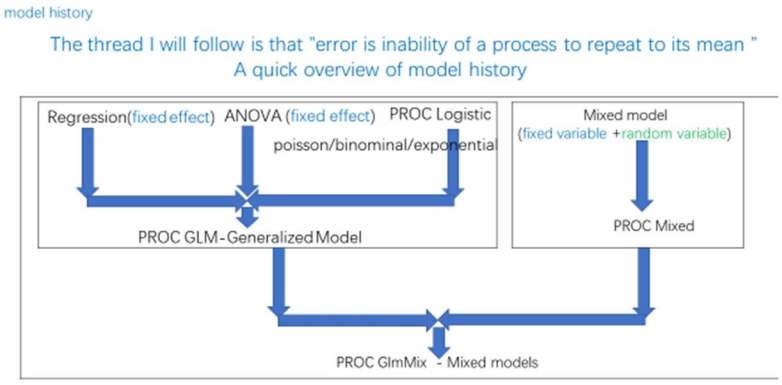

Figure 1

2) ONE WAY FIXED VARIABLE ANOVA- LINKING THE FORMULA TO SAS VIA HAND CALCULATION

The thread that the authors are going to weave through this paper is that variability, or error, is the inability of a process to repeat a mean. Summed squared total (SST), summed squared error (SSE) and summed squared model are all calculated by summing the squared distances of something (an observation or a mean) from a mean. An important thing to note is ***which*** mean is used in which calculation.

Assume we randomly assigned six stores to each of three salespeople and record the dollars sold at each store.

We are going to define, and then partition, the variability in dollars sold. Russ is the blue salesperson Ajoy is the red salesperson and Bo is the orange salesperson.

Think of the three salespeople as three processes that each have a mean and a standard deviation. The processes then generate data

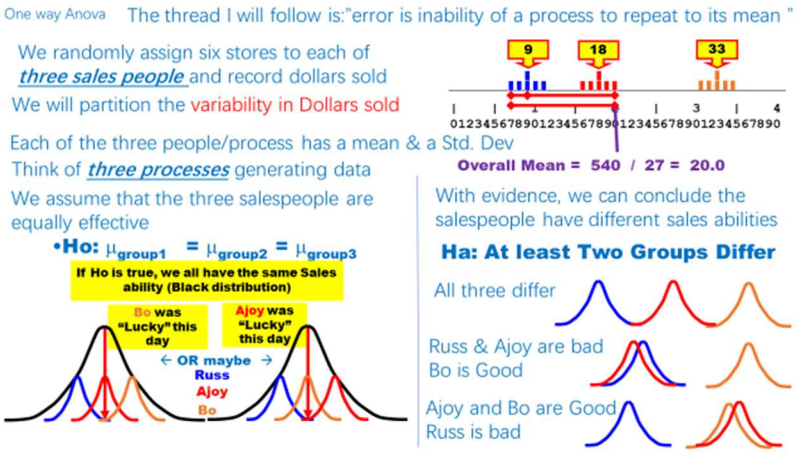

Figure 2

While the histogram of the sales for the three salespeople (see upper right of Figure 2) suggest that the salespeople have different skills, we are going to assume that all three people are equally effective.

Ho is $\mu_1 = \mu_2 = \mu_3$ where the mean of Russ is μ_1 ,the mean of Ajoy μ_2, and the mean of Bo is μ_3

Please look at the lower left corner of Figure 2 for graphical representations of Ho being true. Ho being true implies one black distribution is the "parent distributions" for the three salespeople.

Let's look at the distribution in the lower left (labeled "Bo was lucky this day)- it represents the distribution for one "contest". If Ho is true then all salespeople have the same sales ability and should have the same sales average. Their true ability is shown by the black curve and Bo, in this picture, sold more dollars, because he had a lucky day.

On any particular day, people are little bit above, or little bit below, their overall average. In the black curve on the bottom left, Bo was lucky. He sold more goods than anybody, that day, because he had a good day.

If Ho is true then the distributions labeled “Ajoy was lucky” might happen on some other day. In that distribution, even though all salespersons have the same skill, Ajoy was the lucky person. If Ho is true we expect that, every time we run this experiment, the order of the salespeople will change – just due to chance.

The right-hand side of Figure 2 shows pictures of how the null hypothesis can be true. For ANOVA problems Ha is usually stated that at least two groups differ. It could be that all three salespeople have different selling abilities (their averages differ greatly compared to their spreads). Ha could also be true if two of the groups are similar and one group is different. The picture in the bottom right-hand side of the slide shows how the situation “at least two groups differ” can occur.

Figure 3 shows how summed squared total variability (SST) is calculated. Variability is inability of a process to repeat to a mean. For fixed effects, summed squared total (SST) is calculated by taking the squared distance between an observation and **the overall mean of the data**.

Variability is the inability of a process to repeat a mean and the formula for summed squared total (SST) uses the **overall mean of the data**.

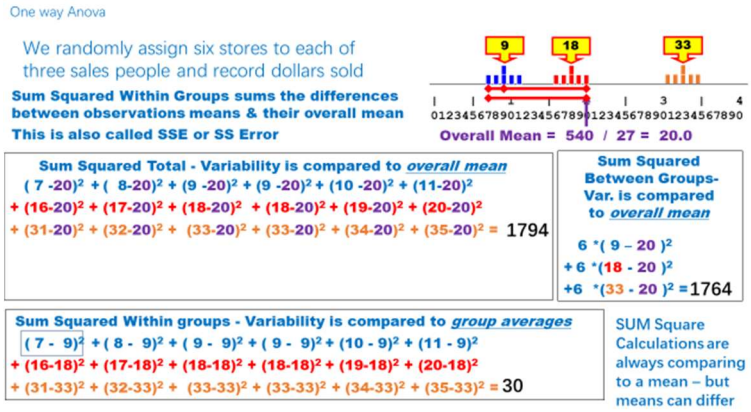

Figure 3

Summed squared error within groups (SSE) is also a variability calculation and it calculates the squared distance between an observation and the ***mean*** for its group. Calculating SSE involves calculating variability and the mean**s** that are used in this calculation are the mean**s** for the individual salespeople and the **overall mean of the data**.

SSModel, or “Summed squared variability explained by our model”, is calculated by taking the ***means*** for the salespeople minus the **overall *mean* of the data** times the number of independent observations for each of the salespeople.

So we’ve partitioned the variability (SST = SSM + SSE) by just taking some “observation/number minus some ***mean”*** squaring that and repeating for all “observations/numbers”. Summed squared calculations are always comparing something to a mean (subtracting a mean from something) but the ***means*** are different for different calculations.

Figure 4 shows a crucial bit of logic and understanding how variability relates to these PROCs.

In the picture, in the upper right-hand corner, it's pretty easy to see that Bo is a much better salesperson than the others.

This picture shows, as a reminder, the values for SST, SSModel and SSE. It also reminds us of the pictures associated with Ho and Ha. This is just one picture of how Ha can be true.

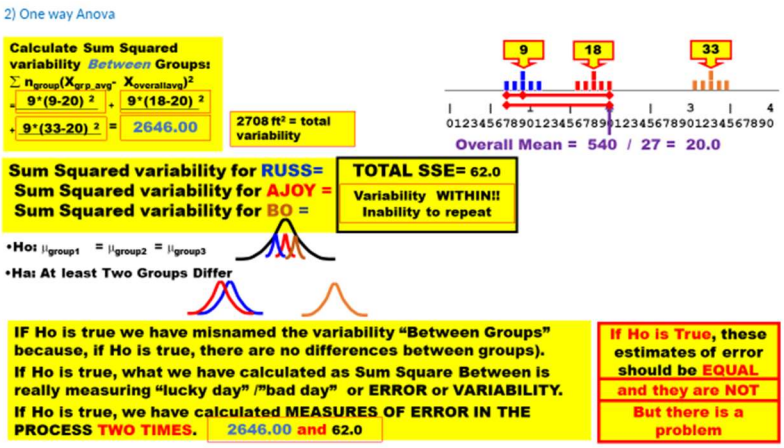

Figure 4

If Ho is true, there will be some small variation of the three salespersons' means around the overall mean.

An important twist in the ANOVA logic is that the formula we have called "summed squared variability between groups" is a just math formula. No matter what the data looks like, this formula will always give us some non-zero answer. If Ho is true or false, this formula will give a non-zero answer. If H0 is true, and the salespersons', means are far from the overall average, this number will be large. If Ho is true, and the salespersons', means are close to the overall average, this number will be small.

This idea is critical! If Ho is true we have mislabeled the variability "between groups". If Ho is true there are no differences between groups. If Ho is true, what we have calculated as SS Model is really a measurement of how lucky or unlucky the salespeople were on a particular day. How lucky or unlucky a person is, is a measure of the inability of a process to always repeat to its mean. If Ho is true, the formula (which, remember, *always* returns an answer – no matter if Ho is true or false) has calculated a measure of good day or bad day – and this is a measurement of error.

If Ho is true we've calculated measures of error in our processes two times and we've gotten two answers (2646 and 62) – we have measured the same "thing" (error) two times. Those two numbers don't look very similar – and there is an adjustment that must be made to them before we can do a real comparison of the two measurements.

When we calculate variability we sum distance between something and the average for each of the "somethings". The more observations we have in the calculation, the larger the variability is likely to be – because we are summing squared distance can never be negative. The 2646 and 62 were calculated using different numbers of observations and we must adjust for the different number of observations.

Rather than comparing summed squared variability we're going to compare the average variability per independent observation. We are going to take the two summed squared totals and divide them by their degrees of freedom.

Mean Squared Model (or MS Treatment) is: 2646 / 2 = 1323 mean squared error is 62 / 24 = 2.58

If Ho is true these numbers should be the same.

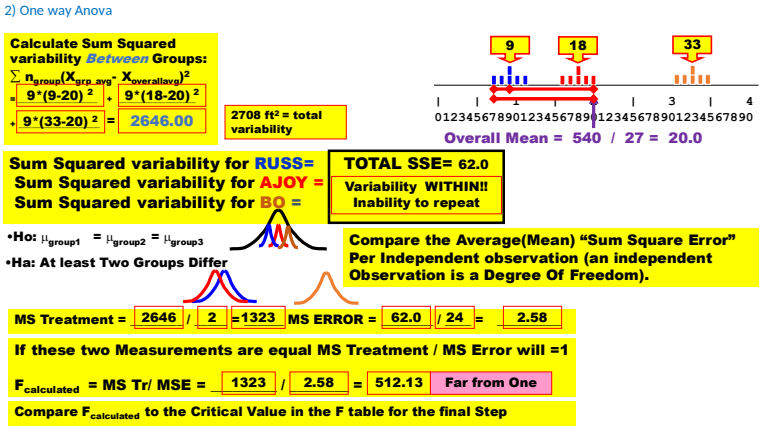

Figure 5

One natural way to compare these two measurements- of the same thing- would be to subtract them. If the numbers are the same, and we subtracted one from the other, we should expect to see a zero – if Ho is true.

Statisticians didn't do that. They decided that if they divided one number by the other number, and if Ho were true (there is no real between group affect) that ratio would be one. This is the logic of the F test.

I encourage you to look at the picture in Figure 5. The large value of SS between (Summed square Model) and the large F value are caused by the averages of the salespeople not being close to the overall average for the data.

If the averages for the salespeople were closer to the overall average, SS between (AKA SSModel) would be smaller. The farther apart salespersons averages are from the overall average, the bigger summed squared between (SSModel) will be. The bigger summed squared between is, the bigger is the calculated F.

If Ho is true, we expect the calculated F value to be approximately one. If Ho as false expect F to be large.

We hope you can link, in your mind the relationship between how far the salespersons averages are from the overall average, through the variability calculations, through to the F test.

Figure 6 shows the output from a SAS analysis of this data.

Note that the numbers that SAS produced are the same numbers we got by hand.

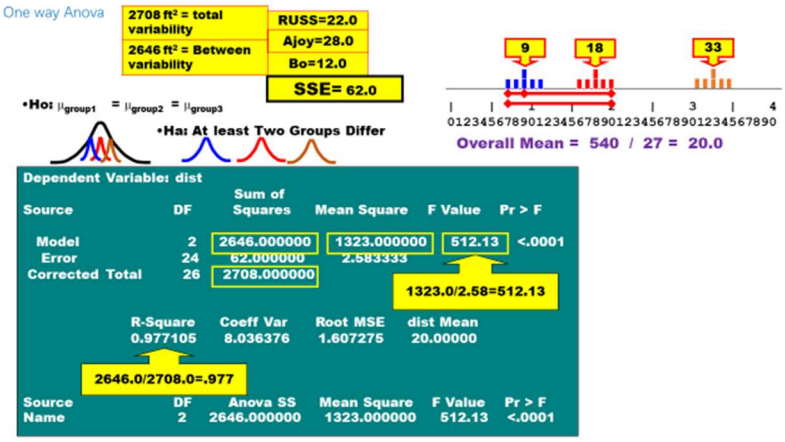

Figure 6

3) TWO-WAY FIXED ANOVA - LINKING THE FORMULA TO SAS VIA HAND CALCULATION

In the next section we will bring the thread of variability through to the analysis of a two-way fixed ANOVA.

There is a new story here. Our company not only owns some grocery stores but manufactures food products under our own brand name.

The business situation is that our company sells sausages and we have contracts to sell our sausages in just three cities.

We send salespeople into grocery stores to stand in the isle with an electric hot plate, warm up some sausages, and let people have a small sample of how good the sausages taste.

In our experiment, we have 24 different stores and 24 different demonstrators. We randomly assigned four stores in each city to a female or male food demonstrator. We have the ability to measure the sales of stores in Boston Philadelphia and Washington DC where we have contracts with distributors.

We are interested in both the effect of gender of the food demonstrator on sales and the effect of city on sales. This managerial interest causes us to code the SAS procedure in a way that it calculates a fixed effects ANOVA.

Error is the inability of a process to repeat it's mean and the graphic in Figure 7 we actually have six processes (six "cells"): female – Boston, male – Boston, female – Philadelphia, male – Philadelphia, female – DC, male – DC.

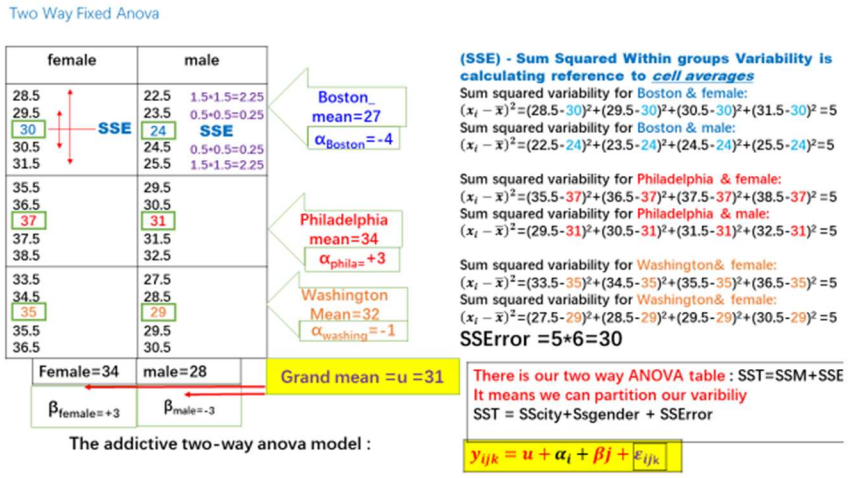

Figure 7

When we calculate SSE we are going to calculate the squared distances between an observation and the mean of its cell (See Figure 7). . This is very similar to the calculation we did in the previous section, when we calculated how far the salesperson's observations were from their individual average. No matter if Ho is true, or false, this is a measurement of error.

We will proceed in a manner that was similar to the calculation of SSM in the previous section (where we compared the averages for the salespeople to the overall average of the data).

To get SSM in this situation we will see how much the averages for female and male differ from the overall average from the data.

We will also see how the city averages differ from the overall average for the data set.

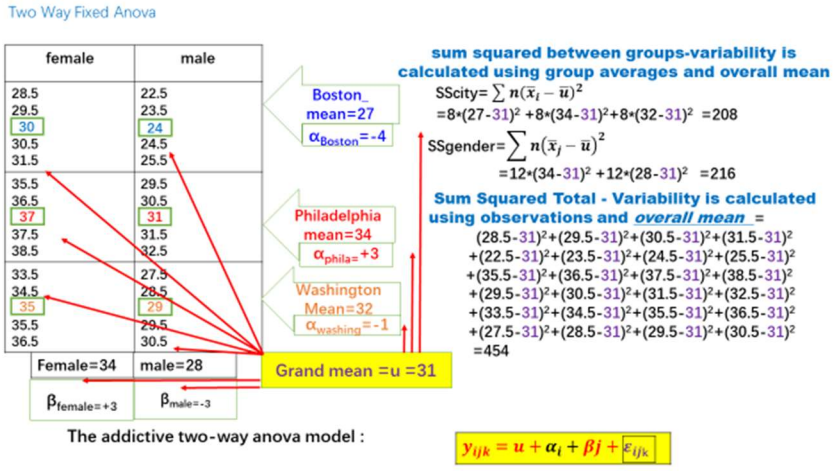

Figure 8

SST is calculated by comparing the squared distance between individual observations and the overall average for the data set.

Fixed effect variables use a calculation that's referencing the ***overall mean of the data set*** (the grand mean) as a way of assigning variability to a cause (the cause being an x variable- here gender and city) and reducing SSE. Reducing SSE makes the F test more powerful and is our goal.

Figure 9 shows that the hand calculations agree with the SAS printout.

Ho for this ANOVA is

Ho: $\mu_M = \mu_F$

& $\mu_{BOSTON} = \mu_{PHILADELPHIA} = \mu_{WASHINGTON}$

In figures 7 to 9, we have calculated summed squared variabilities, and you can calculate the mean squared variabilities by dividing the summed squared numbers by their degrees of freedom.

.

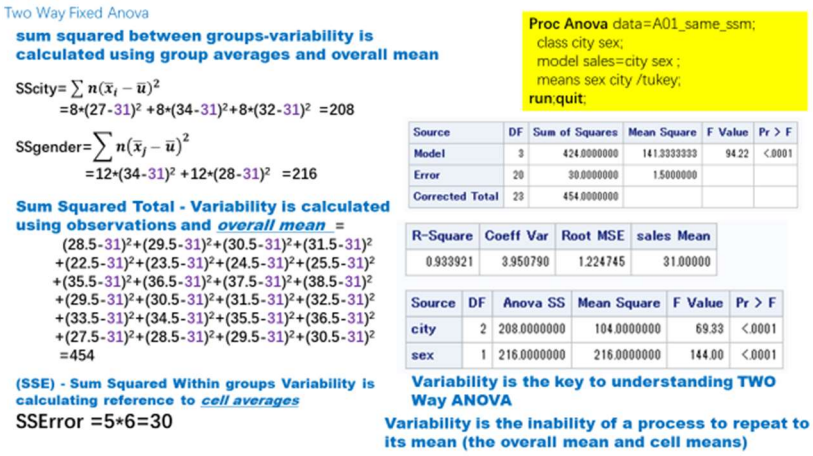

Figure 9

The F test for the model is: (MSCity + MSGender) / MSE

If neither city or gender have an effect on sales we expect the F test of the model to be one. Even if there is no true effect of city or gender we expect the city and gender averages to differ slightly from the grand mean of the data set.

If Ho is true, these differences from the overall average will be small and a measure of error. If Ho is false, these differences from the overall average will be large and due to the fact that the gender of the food demonstrator, and/or the city in which the demonstration is performed have an effect on sales.

If the F test for the model is not significant, ***an analyst must stop at that point***. An analyst is only allowed to look at the effect of city, or of gender, if the model "as a whole" is significant. A commonly used phrase is that the F tests for city and/or gender are protected by the F test for the model.

Assuming a significant model:

The F test for city is: MSCity / MSE.
The F test for gender is: MSGender / MSE.

Hopefully; a reader can link the SAS printout, through the hand calculations, to the picture of the data (Figures 7 and 8).

4) ONE WAY FIXED ANOVA USING PROC ANOVA AND PROC MIXED - FIXED ANOVA-USING TWO DIFFERENT PROCS →SAME RESULT

This section, and figure 10, illustrate a small point. PROC ANOVA is only able to analyze fixed effects (only X variables that are considered to be fixed).

PROC Mixed can analyze both fixed, and random, X variables.

This slide shows that, if you analyze the same "fixed X variable model" with the two different procedures, you get the same analysis and numbers.

However; the **tables produced are different** and the descriptions of the analysis is different. This one slide shows the output of a one way fixed ANOVA using PROC ANOVA and also PROC Mixed.

While the two PROCs produce much of the same output, elements of information, that are useful to an analyst, have different names, or are in different looking tables, in the output from the two PROCs. Since SSE and MSE are so important in judging the quality of a model, it is important to note that the MSE, which one might be used to seeing in an ANOVA table, is in, for PROC Mixed, the table titled covariance parameter estimates.

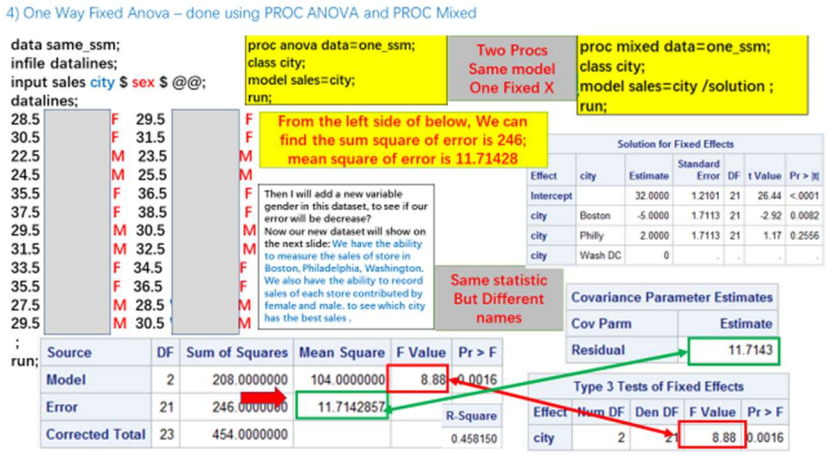

Figure 10

5) ONE-WAY FIXED ANOVA VS. TWO-WAY FIXED ANOVA 2ND X REDUCES SSE USING THE FIXED EFFECT CALCULATION

Section 5 illustrates a fundamental principle of statistical analysis. The principle is, that one tries to add X variables to a model in hopes that the new variable can help explain variability in Y. It is hoped, that a new X variable will explain some of the error in the data and reduce SSE.

This section, has two goals: a quick comparison of 1X vs 2X models and to introduce matrix notation.

Figure 11 uses PROC ANOVA to do a one X variable and then a 2X variable analysis.

In this data set, variables city and sex are both able to predict sales.

Adding the second variable, as a fixed variable this time, reduced SSE and made the F test more powerful.

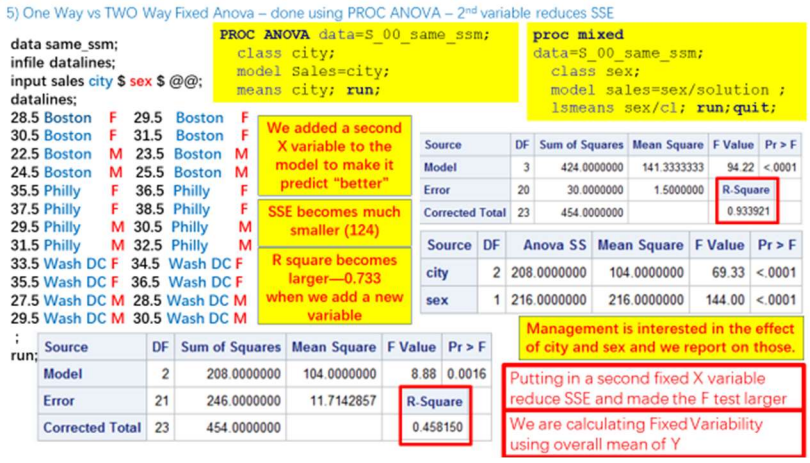

Figure 11

Since this is a fixed variable analysis, the variability explained by the variables city and sex is calculated by comparing something **to the grand mean of the data**.

Figure 12 introduces the matrix form of a one-way ANOVA.

The algebraic form is:

$\mathbf{Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}}$

The matrix form is:

$\mathbf{Y = X\beta + \varepsilon}$

The design matrix for this model is shown in the lower right-hand corner of the figure. The yellow column is an intercept column the columns for males and females are colored pink and blue.

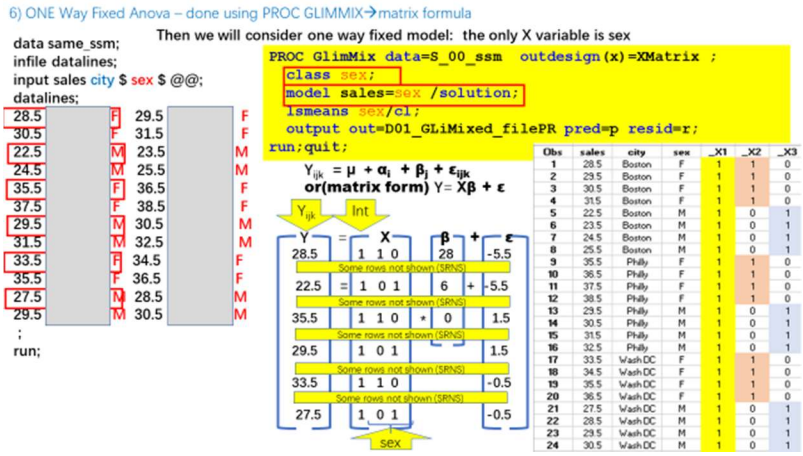

Figure 12

Figure 13 shows, in a gray box, the calculation for the Y=28.5.

This is the example of hand calculations that was mentioned in the abstract.

It's an illustration of the matrix multiplication and summation processes that are implied in the formula

$\mathbf{Y = X\beta + \varepsilon}$

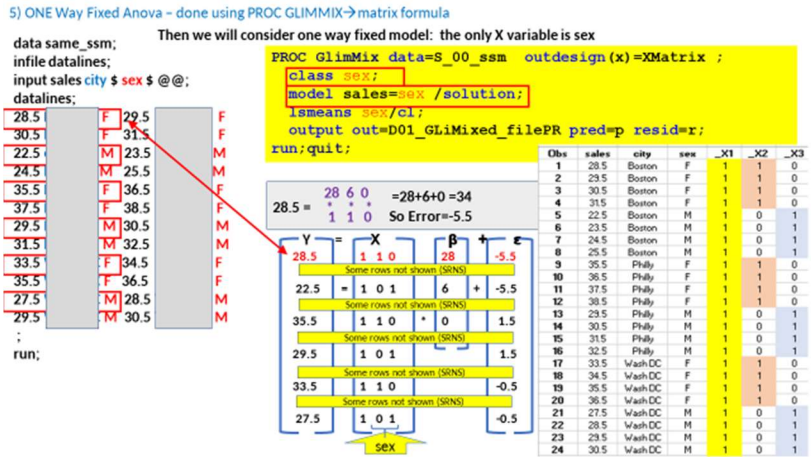

Figure 13

Figure 14 shows another example of performing the matrix calculations for a one variable fixed effect model.

The figure, in a gray box, the calculation for the Y=29.5

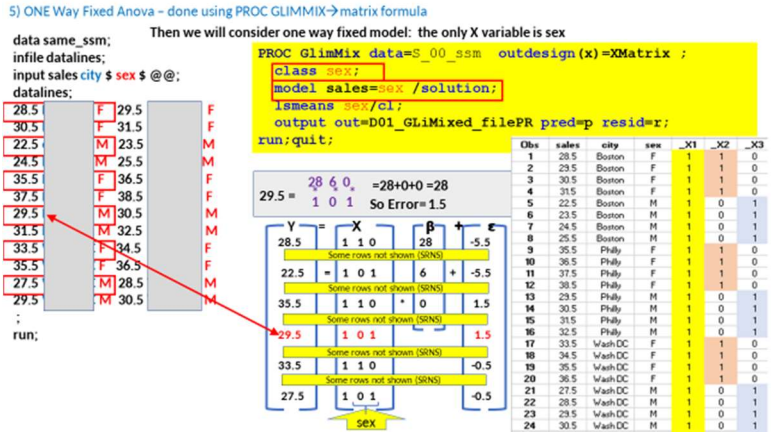

Figure 14

Figure 15 is an illustrates the design matrix for a two variable fixed effect model. This model could be done in PROC ANOVA but we used PROC GlimMix because that PROC outputs design matrices.

The source for the beta matrix is shown in the lower left-hand corner of the figure. That beta is printed by SAS.

The design matrix for this model is shown in the lower right-hand corner of the figure.

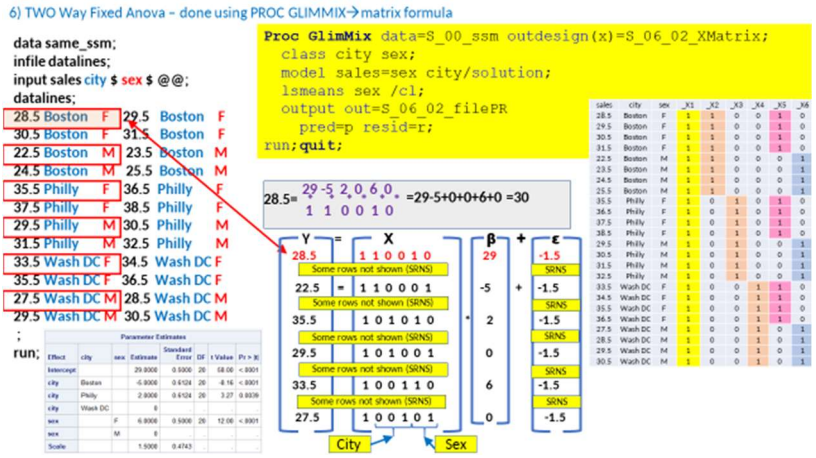

Figure 15

6) TWO WAY FIXED ANOVA – DONE USING PROC GLIMMIX→MATRIX

Figure 16 shows an example of the calculations are required for a two variable fixed effect model.

The design matrix (the X matrix has more columns)

The source for the beta vector is shown in the lower left-hand corner of the figure.

The design matrix is shown in the lower right-hand corner of the figure.

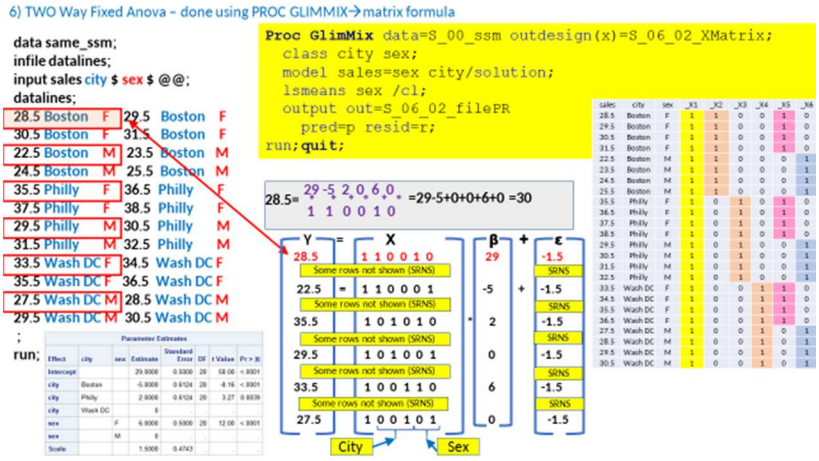

Figure 16

Figure 17 shows another calculation using the same two variable fixed effect model that was presented above.

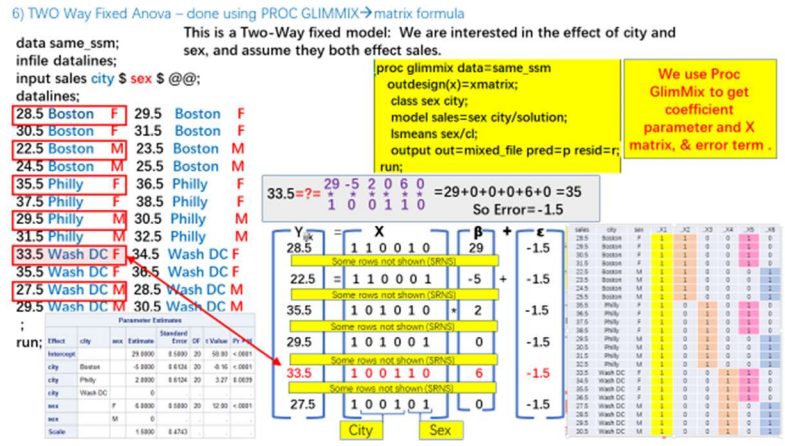

Figure 17

7) TWO WAY FIXED ANOVA – DONE USING PROC GLIMMIX→MATRIX

Practically, a variable becomes "fixed" because your code it that way in your SAS program – you type in options that cause SAS to treat a variable as fixed. You would want to code a variable as fixed for one of two reasons.

Firstly; you would want to code a variable as fixed because you are interested, managerially, in knowing about the averages of each of the levels of X. If you are interested in knowing the effect of different levels of X on the Y variable, you are thinking of that X as a fixed variable.

This means, in a very practical way, if you wanted to repeat this experiment **you would use the same levels of X** in the new experiment. This re-using of levels of x in a future experiment is a characteristic of a fixed effect. This means that if you're interested in the effects of gender on sales, in the next experiment you would use the same genders. If you are interested in the effect of some particular cities on sales, *because you had to make managerial decisions about advertising in a particular city or maybe putting another store in a city*, you would use the same cities in your next experiment.

The second reason is a variable is coded as fixed is that you have used all possible levels of the x variable in your study. In our study, our company can only sell in three cities – right now - so these three cities represent the universe of our possible stores and, therefore, cities are fixed.

It is time to consider random X variables. If we had the potential to sell product in any city in America then the three cities we picked for the experiment would be a sample for all the possible levels of the variable cities. In that case, the variable would be modeled as a random x variable (levels of x are a random selection from the universe of cities). If we were to re-run the experiment we would probably use different cities. Therefore; cities should be considered a random variable.

In a two-way fixed regression, we entered cities into our model, as Fixed, in the hopes of reducing SSE and also getting information about the averages of the cities.

However; in mixed models (a mixture of Fixed and random X) ,while we enter random variables into the model because we want to reduce SSE, the logic is different and the results are different.

If we code cities as a random variable we do not get beta values for the cities – and we don't care. By definition, you are not interested in details about particular cities if cities is a random variable.

What we do get is an estimate of the variability in sales caused by cities having different characteristics. The variability that we estimate **is not the variability caused by these three cities, but the variability that would because by the larger population of cities**.

When one author was taught about random variables the professor said "if you want to make statements about the population in general, make the variable random". However, we have learned, the statement that can be made about a random variable is not very enlightening. The only statement that one can make about the "general population of cities" is "does cities cause variability in sales or not". If different levels of cities cause variability in Y, we can use the random variable to remove variability and make the calculated F test larger. From Mixed analysis we only get information about the variability in Y caused by a random variable. We can say that city-to-city variability is small (non-significant) or large (significant)

If we make cities a random variable we are able to say whether there's **a lot of variability in cities** (the solid black line with the normal shape in figure 19) or if **there is just a small amount of variability** caused by cities (the dotted black line with the normal shape in Figure 18).

The statement that we make, about variability in cities, is not very interesting but cities *can* explain variability and reduce SSE and that is interesting.

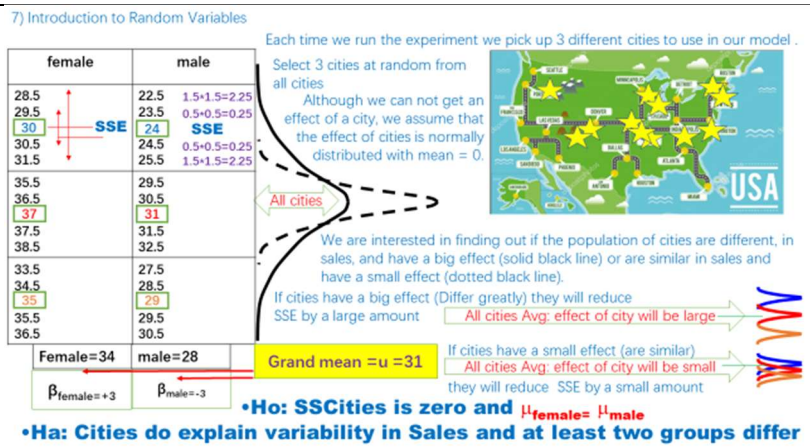

Figure 18

Making cities a random variable allows us to pull variability out of SSE and increase the power of the F test.

As a thought exercise, let's review how cities, in the story we've been using through much of this paper, **could be considered either a fixed variable or a random variable**.

Firstly, Imagine our food company only has stores in Boston Philadelphia and Washington DC – those are our three cities. Because every "level" of possible city is represented in the experiment - city is a fixed level variable. The second way a variable can be fixed is that management is specifically interested in means and standard deviations about the levels of that variable.

If, our retail chain of stores had 20 stores in many different cities and the test was done only in Boston Philadelphia and Washington, then city would be a random variable.

8)TWO WAY FIXED ANOVA – DONE USING PROC GLIMMIX→MATRIX

The common formulas for mixed models are shown in figure 20

the basic formula

$$Y = (X * \beta) + (Z * \gamma) + \varepsilon$$

is complicated enough but the other formulas are even worse – especially when you cannot see the matrices. ---→

SAS comes to our aid and can print many of the matrices used in the calculation.

The basic formula is:

$$Y = (X * \beta) + (Z * \gamma) + \varepsilon$$

Y is an N by 1 vector of observations of Y.

X is a design matrix for the fixed effects (SAS will print this)

B is the coefficients for the fixed effects (think Betas in Mult. reg.)

Z is a design matrix for the fixed effects.

γ is a vector of the effects of the levels of the random variable. It is assumed to have an expected value of 0 and a variance covariance matrix that is represented by G. PROC Mixed will print the G matrix.

ε is a vector of errors. It has an expected value =0 and a variance R

Figure 19

You can use PROC GlimMix to have SAS output the X and Z matrices.

You can use PROC mixed to get SAS to output the beta and gamma matrices (and also the G matrix).

If you have a repeated measures problem, you can use PROC GlimMix to have SAS output the R matrix.

Figure 20 shows a calculation for a mixed model.

Sex is fixed because were interested in finding out whether men or women are better food demonstrators. City is random because there are three cities selected from all the possible cities in America.

The first column in X is the intercept and it adds 28 to every row during a matrix multiplication.

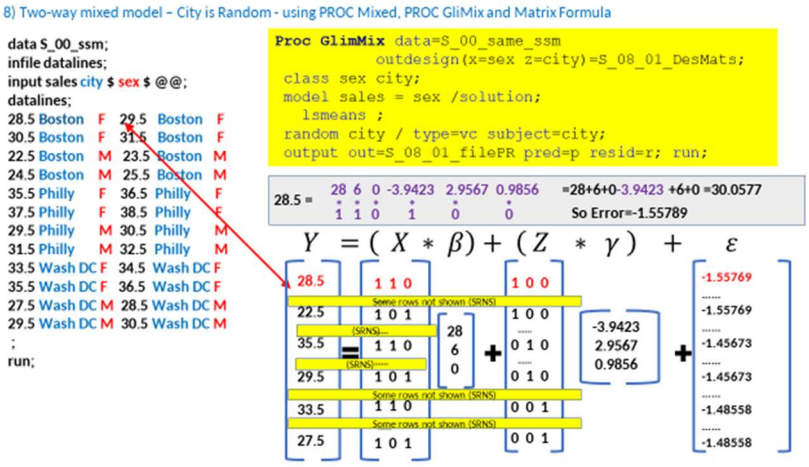

Figure 20

Male is the base case sex (if sorts higher than F). 6 is the effect of being Female. Since M is the base case, the effect of being male is 0 – the effect of male is in "part of" the intercept.

Figure 21 shows another calculation for a mixed model.

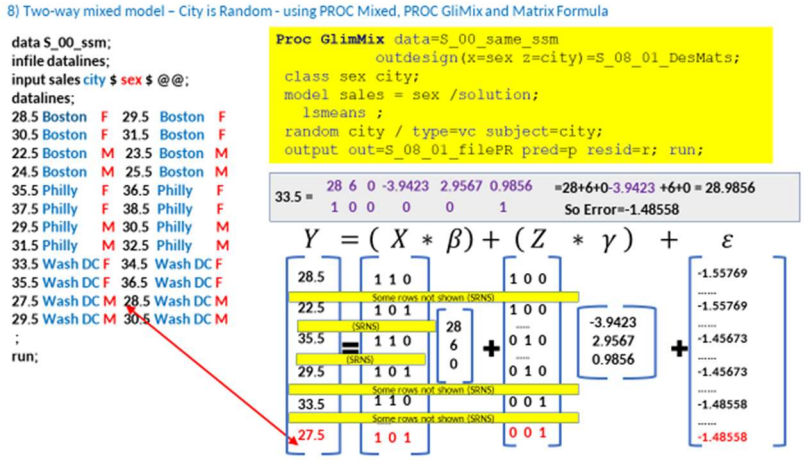

Figure 21

9) A WARNING ABOUT CONVERGENCE

There seem to be a lot of problems that occur when attempting to get answers out mixed models.

The first problem is failing to converge. They seem to be many reasons why a mixed model fails to converge and the topic is both too broad and too complicated to be treated here. We will refer you to some excellent articles in the reference section to help you start your journey of learning.

The second problem is that a message can be sent to the log that says that convergence was achieved but there's a problem with the Hessian matrix. This topic is also too broad and too complicated to be treated here – especially since there seems to be some disagreement on the web as to whether this is a serious problem and not. The authors plan to treat this "warning" about the Hessian as an error message.

Many people say that if there is a note indicating a problem with the Hessian the model SHOLD NOT BE USED. It seems a majority of people writing on the web are of the opinion that reasonable looking model results, when there is a note about the Hessian, can not be used.

There are some suggestions on what to do, when a researcher encounters one of these two problems. Some people suggest that a researcher should estimate a simpler model. The idea is that the researcher should estimate a model with fewer X variables or a simpler covariance matrix (all the elements in the covariance matrix must be estimated that estimation “soaks” up degrees of freedom).

The problem with these solutions is that the researcher ends up estimating the wrong model. There are lots of papers that discuss the effect of missing X variables on model accuracy. There are fewer papers on the effects of estimating a model with an incorrect covariance matrix – but the ones that do exist don’t offer any great encouragement.

The authors have not been able to find any process, or sequence of steps, to address these two issues, but will continue to research (Maybe model convergence with training wheels will be written)

CONCLUSION

The authors hope that this paper, by showing calculation details, and how to get SAS to print matrices that are internal to the model (see appendix) , is a help to people starting to learn Mixed models.

REFERENCES

Hamer , Robert M., Ph.D , 2000, Mixed-Up Mixed Models: Things That Look Like They Should Work But Don't, and Things That Look Like They Shouldn't Work But Do *Proceedings of the SAS Global 2000 Conference* Available at: http://www2.sas.com/proceedings/sugi25/25/aa/25p020.pdf

Grace-Martin, Karen, When the Hessian Matrix goes Wacky, The Analysis Factor,
Available at: http://www.theanalysisfactor.com/wacky-hessian-matrix/

Kiernan , Tao & Gibbs, 2012, Tips and Strategies for Mixed Modeling with SAS/STAT ® Procedures, *Proceedings of the SAS Global 2012 Conference*
Available at: http://support.sas.com/resources/papers/proceedings12/332-2012.pdf

Liu, Cao Chen and Zagar, Good RANDOM and REPEATED statements – How to Use Them to Model the Covariance Structure in Proc Mixed , 2007 , *Proceedings of the MidWest SUG 2012 Conference*
Available at: https://www.mwsug.org/proceedings/2007/stats/MWSUG-2007-S02.pdf

Kiernan , Tao & Gibbs, 2015, Advanced Techniques for Fitting Mixed Models Using SAS/STAT ® Software, *Proceedings of the SAS Global 2000 Conference*
Available at: https://support.sas.com/resources/papers/proceedings15/SAS1919-2015.pdf

Kincaid, Chuck, 2005, Guidelines for Selecting the Covariance Structure in Mixed Model Analysis *Proceedings of the 30th SAS User Group Internationalt 2005 Conference* Available at:
http://www2.sas.com/proceedings/sugi30/198-30.pdf

CONTACT INFORMATION

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

YuTing Tian, YT899963@WCUPA.EDU
Russ Lavery, contractor, Bryn Mawr PA russ.lavery@veriozon.net

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.

```
/*************************************************************
******************
*  Program name      : WUSS 2018-60
*  Project           : YTRL
*  Written by        :
*  Date of creation  :
*  Description       :
*  Macros called     :
*  Input file        :
*  Output file       :
*  Revision History  :
*  Date      Author   Description of the change
*
*************************************************************
******************/
Footnote "Prog Name: %sysget(SAS_EXECFILENAME) **By: &SYSUSERID
** Run on: %sysfunc(datetime(),datetime16.)";
/*************************************************************
**
Section _00_: data read and QC
*************************************************************
**/
ODS LISTING;
OPTIONS NOCENTER;
Title "section A: data read and qc";
Data S_00_same_ssm;
infile datalines truncover firstobs=2;
input @1 sales 4.1 @7 city $char7. @16 sex $char1. @;
output;
input @19 sales 4.1 @25 city $char7. @34 sex $char1. @;
output;
datalines;
12345678901234567890123456789012345678901234567890123456
28.5  Boston   F  29.5  Boston   F
30.5  Boston   F  31.5  Boston   F
22.5  Boston   M  23.5  Boston   M
24.5  Boston   M  25.5  Boston   M
35.5  Philly   F  36.5  Philly   F
37.5  Philly   F  38.5  Philly   F
29.5  Philly   M  30.5  Philly   M
31.5  Philly   M  32.5  Philly   M
33.5  Wash DC  F  34.5  Wash DC  F
35.5  Wash DC  F  36.5  Wash DC  F
27.5  Wash DC  M  28.5  Wash DC  M
29.5  Wash DC  M  30.5  Wash DC  M
;
run;quit;
```

```
proc summary  data= S_00_same_ssm ;
  class sex city;
  var sales;
  types () sex city sex*city;
  output out=S_01_Raw_data_QC;
run;

Proc print data=S_01_Raw_data_QC;
  where upcase(_STAT_)="MEAN";
run;

/*************************************************************
**
Section _1_: a history of the procs
*************************************************************
**/

/*************************************************************
**
Section _2_: ONE way anova using anova, glm and mixed
 show you get same result
*************************************************************
**/
data S_02_sales;
infile datalines ;
input salesperson $ Sales  @@;
datalines;
Russ 7      Russ 8      Russ 9    Russ 9      Russ 10     Russ
11
Ajoy 16     Ajoy 17     Ajoy 18   Ajoy 18     Ajoy 19     Ajoy
20
Bo   31     Bo   32     Bo   33   Bo   33     Bo   34     Bo
35
;
run;
Proc anova data=S_02_sales;
  class salesperson;
  model sales=salesperson;
run; quit;


/*************************************************************
**
Section _3_: Two Way Fixed Anova
SSE is smaller than 1 way Fixed ANOVA
```

```
**********************************************************************
**/
Proc Anova data=S_00_same_ssm;
  class city sex;
  model sales=city sex ;
  means sex city /tukey;
run;quit;

/**********************************************************************
**
Section _4_: One Way Fixed Anova - done using PROC ANOVA and
PROC Mixed
COMPARE OUTPUT
**********************************************************************
**/
PROC ANOVA data=S_00_same_ssm;
  class city;
  model Sales=city;
run;

PROC Mixed data=S_00_same_ssm;
  class city;
  model Sales=city / solution;
run;

/**********************************************************************
**
Section _5_: One Way
 done using PROC ANOVA - vs Proc Mixed and GlimMix (get design
matrices
**********************************************************************
**/
title "Section 5 One Fixed way - PROC ANOVA";
PROC ANOVA data=S_00_same_ssm;
  class city;
  model Sales=city;
  means city;
run;


title "Section 5 One Fixed way - PROC Mixed";
PROC Mixed data=S_00_same_ssm;
  class sex;
  model sales=sex/solution ;
  lsmeans sex/cl;
run;quit;
```

```
title "Section 5 One Fixed way - PROC GLIMMix to get design
matrices";
PROC GlimMix data=S_00_same_ssm  outdesign(x)=S_05_01_XDesMat ;
  class sex;
  model sales=sex /solution;
  lsmeans sex/cl;
  output out=S_05_01_filePR pred=p resid=r;
run;quit;

PROC Print data=S_05_01_XDesMat;
run;

PROC Print data=S_05_01_filePR;
run;

title "Section 5 Two Fixed way - PROC GLIMMix to get design
matrices";
PROC Mixed data=S_00_same_ssm;
  class city sex;
  model sales=sex city/solution ;
  lsmeans sex/cl;
run;quit;

PROC GlimMix data=S_00_same_ssm  outdesign(x)=S_05_01_XDesMat ;
  class sex;
  model sales=sex /solution;
  lsmeans sex/cl;
  output out=S_05_01_filePR pred=p resid=r;
run;quit;

PROC Print data=S_05_01_DesMat;
run;

PROC Print data=S_05_01_filePR;
run;


/*************************************************************
**
Section _6_:
*************************************************************
**/
/*ONE Way Fixed Anova - done using PROC GLIMMIX - matrix formula
*/
title "section 6 - ONe way fixed using glimmix";
PROC GlimMix data=S_00_same_ssm  outdesign(x)=S_06_01_XMatrix ;
  class sex;
```

```
  model sales=sex /solution;
  lsmeans sex/cl;
  output out=S_06_01_filePR pred=p resid=r;
run;quit;

title "section 6 - Two way fixed using glimmix";
/*TWO Way Fixed Anova - done using PROC GLIMMIX-matrix formula*/
Proc GlimMix data=S_00_same_ssm
  outdesign(x)=S_06_02_XMatrix ;
  class city sex;
  model sales=sex city/solution;
  lsmeans sex /cl;
  output out=S_06_02_filePR pred=p resid=r;
run;quit;

/*************************************************************
**
Section _7_: intro to mixed var - all text
*************************************************************
**/


/*************************************************************
**
Section _8_: Two-way mixed model - City is Random - using PROC
Mixed,
  PROC GliMix and Matrix Formula
*************************************************************
**/
Title "Section 8: analysis &Predicted and residuals for 2 way
MIXED (Sex=F city=R) anova ** GLMMix";
Title2 "GlimMix";
title3 "Having the design matrix allows us to easily show fixed
calculations";
/*do a fixed using glimmix*/
PROC GlimMix data=S_00_same_ssm    outdesign(x=sex
z=city)=S_08_01_DesMats;
 class sex city;
 model sales = sex /solution;
   lsmeans ;
 random city / type=vc subject=city;
 output out=S_08_01_filePR pred=p resid=r;
 run;

PROC Print data=S_08_01_DesMats;
run;
```

```
PROC Print data=S_08_01_filePR;
run;

PROC Summary  data= S_08_01_filePR ;
class sex city;
 var p r;
 types () sex city sex*city;
 output out=S_08_02_GLiMixed_filePR ;
run;

PROC Print data=S_08_02_GLiMixed_filePR;
  where upcase(_STAT_)='MEAN';
run;
```