SAS 9.4:
A Survival Guide for EPRS 8530, EPRS 8540, and EPRS 8550 Students

Edition 1

April 2019

Theresa L. Dell-Ross
Department of Educational Policy Studies
Georgia State University

T. Chris Oshima
Department of Educational Policy Studies
Georgia State University

Table of Contents

Foreword

The purpose of this guidebook is to help you successfully use SAS 9.4 to complete the course assignments for EPRS 8530: Quantitative Methods & Analysis I, EPRS 8540: Quantitative Methods & Analysis II, and EPRS 8550: Quantitative Methods & Analysis III. Specifically, this guide will provide assistance as you (1) analyze data using a variety of statistical models, (2) verify the tenability of the associated statistical assumptions, (3) read and interpret the corresponding output, and (4) report your results in an appropriate format and style.

For each method of statistical analysis, you will find the following:

1. a sample research scenario with a small dataset,
2. SAS code for executing the analysis with explanatory notes and tips, and
3. excerpts of selected output noting the results most commonly of interest to the researcher, and
4. sample summary statements of the results following American Psychological Association (APA) 6th Edition reporting guidelines.

Please note that most of the scenarios and datasets herein are intentionally small and simplistic, so that you can focus your attention on the SAS code and output presented. The code needed to run analyses will invariably differ for your own particular course assignments and research projects. *Pay careful attention to the explanatory notes provided*, as these should help you adapt the code to your specific needs.

It is our hope that you will find this guidebook helpful as you learn new statistical methods and that it also comes in handy as a reference for you later in your statistics career.

Theresa Dell-Ross
Department of Educational Policy Studies
Georgia State University

T. Chris Oshima
Department of Educational Policy Studies
Georgia State University

Acknowledgement

This guide is modeled in large part after *SPSS for Windows Versions 20.0 (for Windows 7): A Survival Guide for EPRS 8530 and EPRS 8540* (Edition 5, Version 1.0; December 2014) by Tianna C. S. Floyd, Keith D. Wright, H. A. Russell III, J. Randy Beggs, Gary L. May, and T. Chris Oshima.

The purposes for the development of *SPSS: A Survival Guide for EPRS 8530 and EPRS 8540* are the same as the purposes for the development of this guide – to help students analyze data, interpret output, and report results – except that the focus is on SPSS software. EPRS 8530 and 8540 students interested in SPSS have been successfully using Floyd et al.'s guide for several years; however, to date, a similar guide has not been made available for SAS. This guide is intended to fill in that gap, using SAS 9.4, as many statistics students are interested in this powerful programming language.

It is entirely possible that students enrolled in these courses may wish to learn both programs. The simplest way to facilitate this is to share the sample research scenarios and datasets between the two guides. **Therefore, unless otherwise noted, all of the scenarios and datasets herein are taken directly from the work of Floyd and her colleagues.** Furthermore, as the two guides present the procedures for the same statistical models, with the same ultimate objectives, the structure and contents of this guide will have much in common with its predecessor; in some cases, verbatim reproduction may be observed.

The authors of this guide express their gratitude for the hard work and dedication that went in to the development and revision of *SPSS: A Survival Guide for EPRS 8530 and EPRS 8540*. Thank you for providing a thorough, high-quality model.

About This Guidebook

This guidebook was developed specifically for students taking statistical courses in the Educational Policy Studies Department of the College of Education and Human Development at Georgia State University. Anyone conducting research analysis who is new to using SAS may find this guide helpful.

This guidebook is organized in three main sections. The first section, Getting Started, provides a brief introduction to SAS 9.4. This section includes directions for you to install SAS; write, run, and save code; check your code for execution errors; and "clean" your SAS windows. The second section, Descriptive Statistics, will provide instruction related to the generation of measures of central tendency, measures of spread, Pearson's *r* for correlation, and frequency tables and histograms. In the third section, Inferential Statistics, you will learn to run analyses ranging from a one-sample *t* test to multiple regression and model building with the incorporation of assumption-checking methods.

The code herein was written using SAS 9.4, with default installation settings, in a Windows 10 environment; it is assumed that the reader is using this version of SAS (or a more current one) in Windows 10 or higher.

It is assumed that students already possess a basic knowledge of Windows and standard computer software (e.g., word processing programs). A working knowledge of any programming language will be helpful, but is not required. As the guidebook progresses from one analysis to another, certain levels of detail are omitted due to space limitations. (For example, explanation of the assignment of raw data to variable names need not be repeated for each successive statistical analysis once it has been introduced.) **It is expected that the user will develop a *cumulative* working knowledge of SAS as the guidebook progresses.**

The SAS code presented herein is limited to the procedures required for the statistical analyses being conducted. (For example, topics such as the manipulation of character data are omitted.) Of course, there are often multiple methods for achieving the same results; however, in most cases, only one way is presented. Readers wishing to learn SAS in depth are therefore referred to SAS courses provided by the SAS Institute, online resources, or print resources.

Every effort has been made to ensure that this guidebook is free from errors. If you should find any mistakes within this guide, please contact Theresa Dell-Ross (tdellross1@student.gsu.edu) or Dr. Chris Oshima (oshima@gsu.edu) so that corrections may be made.

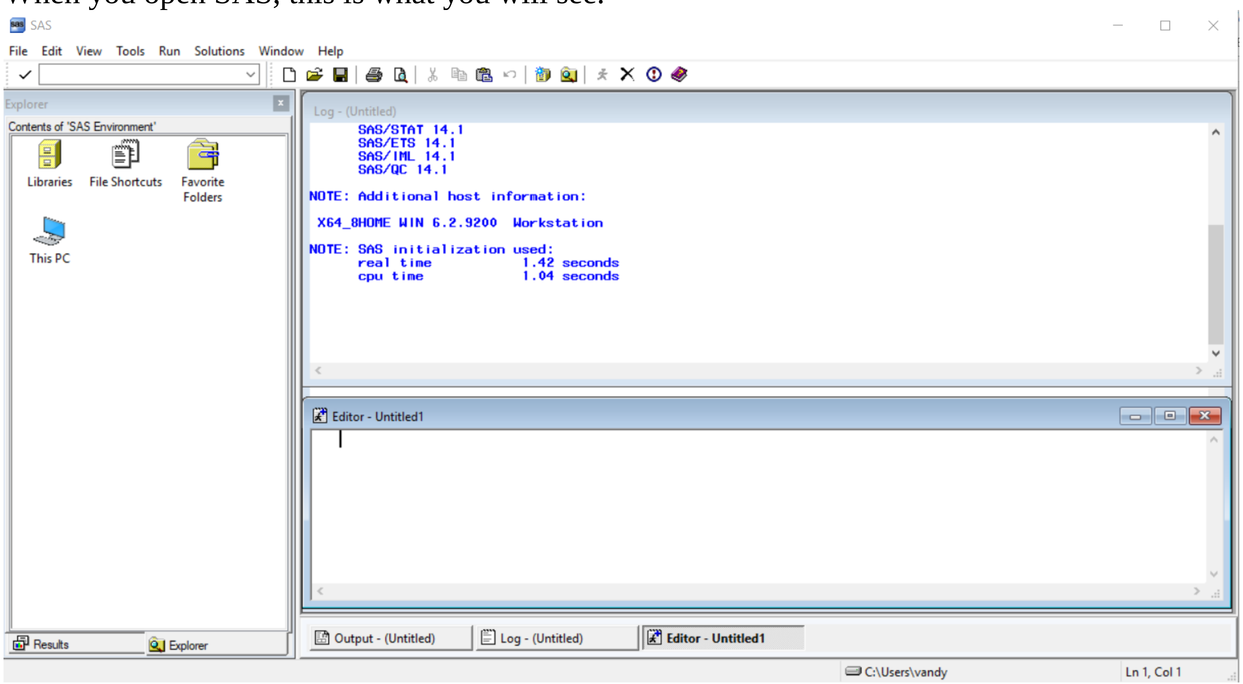
Getting Started with SAS 9.4

Installing SAS

SAS 9.4 is currently offered for free to Georgia State University students. To download and install SAS, go to the GSU Technology home page at https://technology.gsu.edu/technology-services/services-for-you/it-services-for-students/. Scroll down to the "Download Software" section, find SAS, and follow the directions.

Using SAS

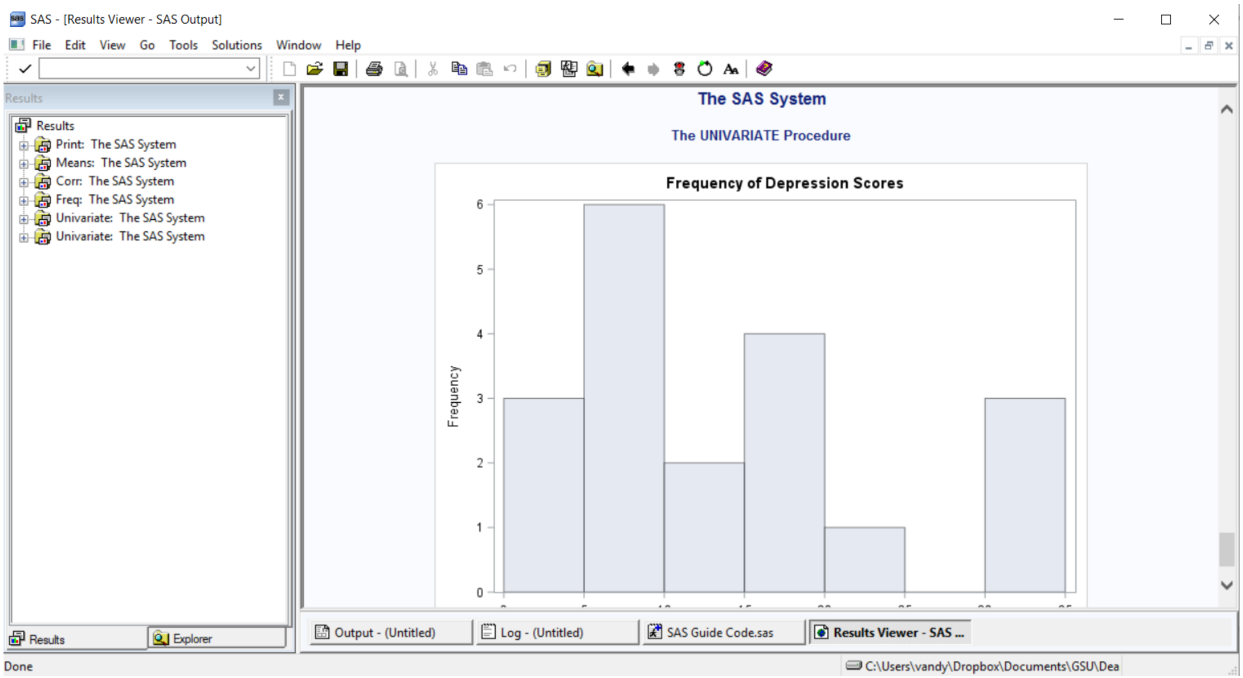
When you open SAS, this is what you will see.



The "Editor" window is where you will write and run your code. The "Log" window is where you will check to make sure that your code was executed properly. You can toggle back and forth using the Editor and Log tabs at the bottom of the SAS screen.

You may begin typing code in the Editor window. You may also open an existing code file by going to File > Open. To save your code, make sure you are in the Editor window and then go to File > Save As or File > Save. To run your code, highlight it and click on the icon of the running man.

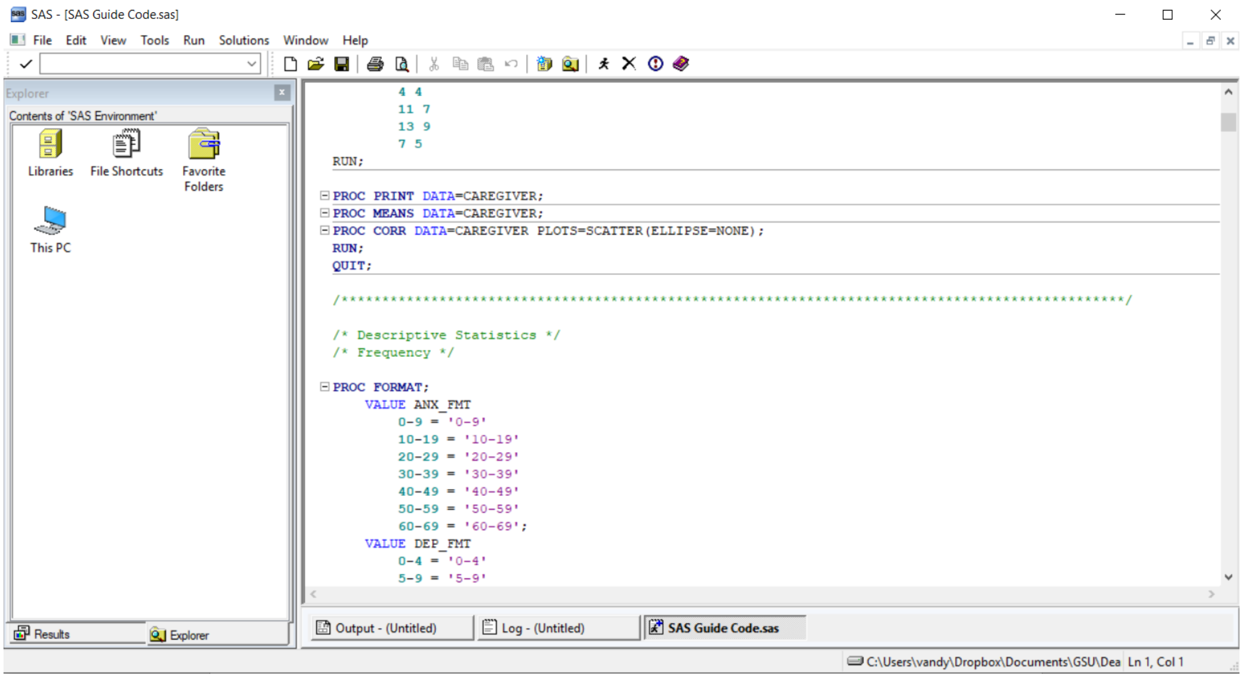
Here is what you will see once you have written and run some code.



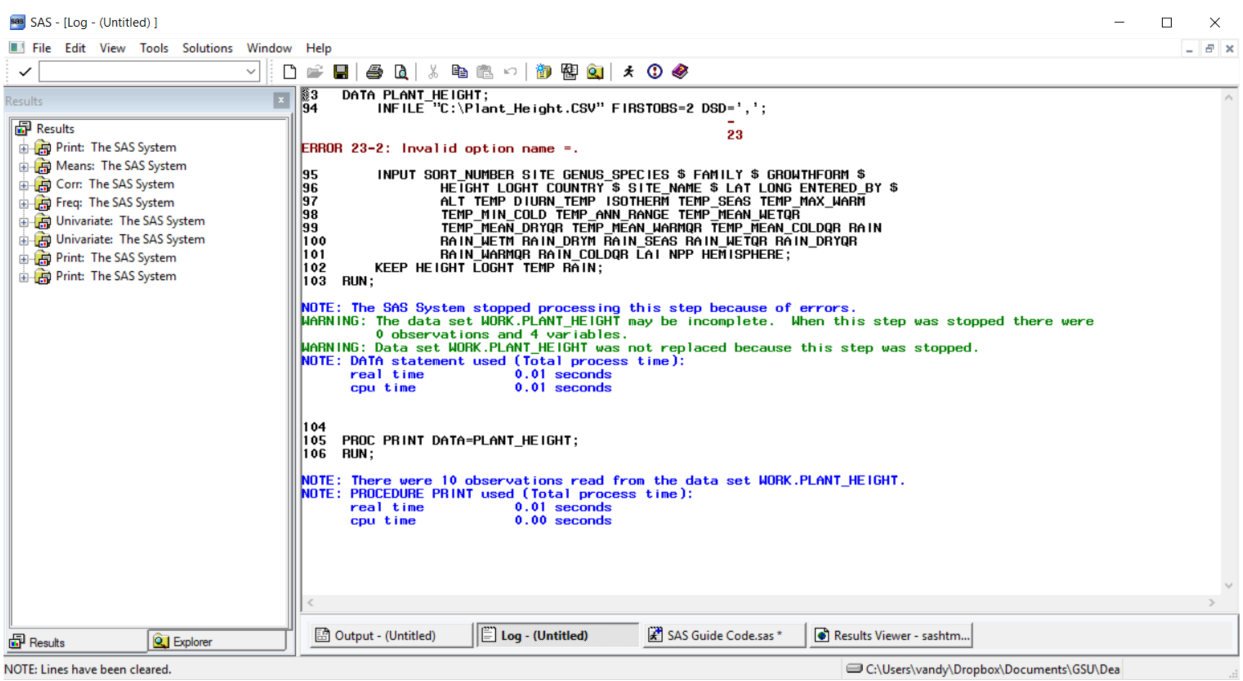
Notice that there is now a list of results in the left-hand window. There is also a "Results" window (with a Results tab at the bottom).

IT IS HIGHLY RECOMMENDED THAT YOU GET IN THE HABIT OF READING THE LOG WINDOW IMMEDIATELY AFTER YOU RUN CODE, EVERY TIME YOU RUN CODE, BEFORE YOU LOOK AT ANY OF THE RESULTS. It is possible that SAS will generate results, even if you made mistakes in your code, and you do not want to interpret or report results without first checking to make sure that the code was executed properly.

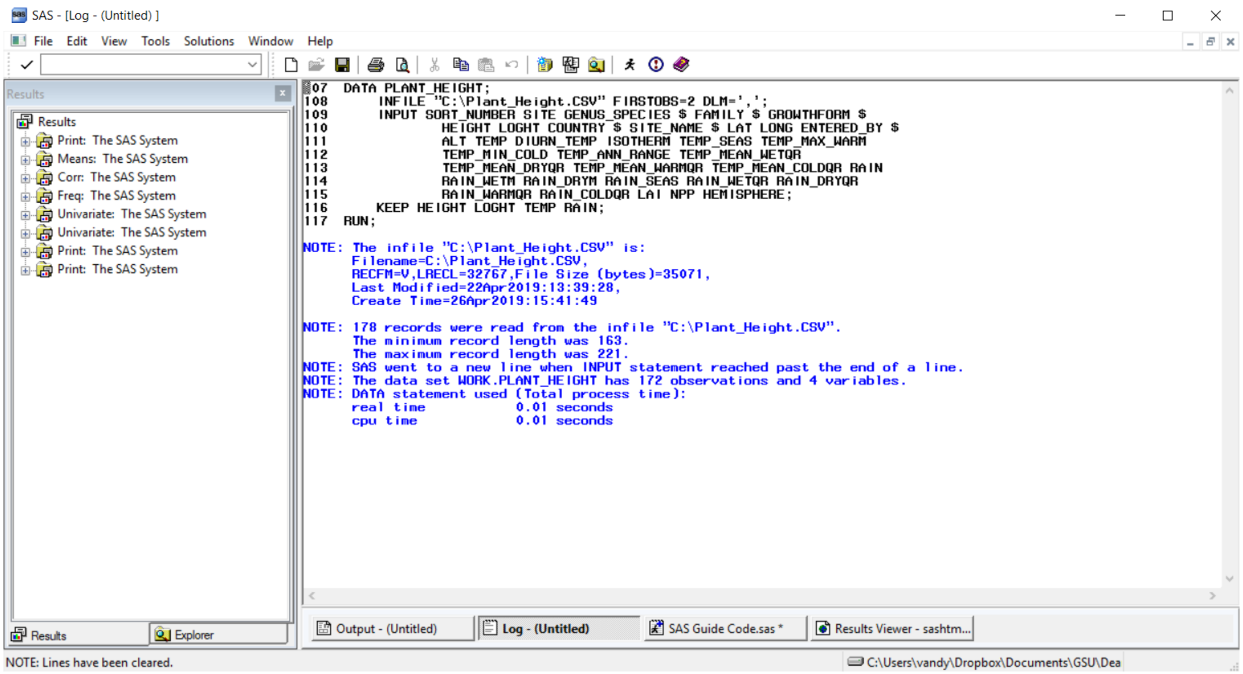
IN SAS, COLOR IS CRITICAL. When you type in your code, SAS automatically color-codes certain words. As you can see from the screenshot below, it colors some words in dark blue, light blue, green, teal, purple, and black. As you use this guide, make sure your colors match the colors you see in the code here. If your colors don't match the colors here, it is a clue that you may have made an error (e.g., forgetting a semicolon in the line above or misspelling a word).



COLOR IS ALSO CRITICAL IN THE LOG WINDOW. When you check the Log window, you always want to see black and blue. The Log window will show your code in black. If there are any errors, they will be shown in maroon (red). If there are warnings, they will be shown in green. This screenshot shows an error and two warnings:



All other notes will be shown in blue. In most cases, blue means success. However, you must be careful. Blue does not mean success 100% of the time. For example, look at the following:

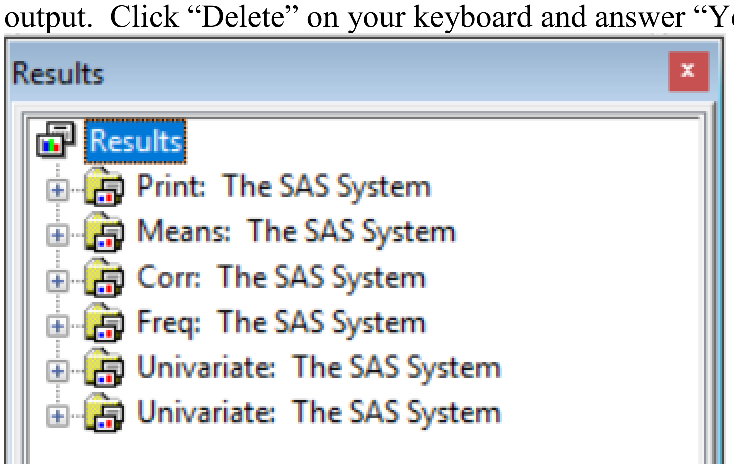


Notice that "178 records were read" but the new dataset has "172 observations." There was, in fact, an error here (having to do with missing data in the .CSV file that was being read in), but SAS did not recognize it as such. ALWAYS read the Log window before interpreting and reporting results!!

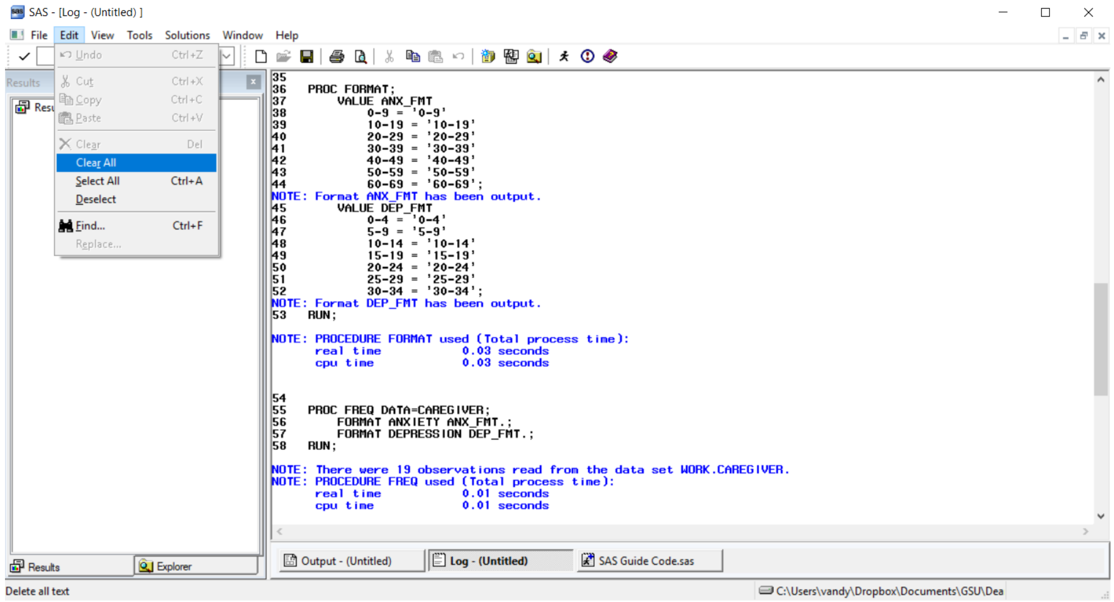
Keeping Your Windows Manageable

Sometimes you may run the same code over and over as you troubleshoot, and you may become overloaded with output in Results window and notes in the Log window. It is easy to fix this. If you want to clear everything out, simply do the following.

1. In the left-hand window, click on the word "Results" at the top of the list of all the output. Click "Delete" on your keyboard and answer "Yes."



2. Select the “Log” window tab, go to “Edit,” and select “Clear all.”



3. Continue working.

Descriptive Statistics
Research Scenario

A counselor working with a group of caregivers of patients living with a terminal illness is interested in forming a support group to share experiences and therefore reduce the sense of isolation often associated with catastrophic illness. The counselor, working with hospital staff, administers a depression and anxiety inventory to each caregiver that has volunteered for the program. The counselor feels that knowing the levels of depression and anxiety within the group will help in the design of an effective intervention program. The scores obtained from the administration of the two inventories are given below.

ANXIETY	DEPRESSION
22	16
12	8
68	33
10	6
5	5
53	24
44	18
37	17
0	2
21	14
64	31
33	17
55	30
18	13
3	3
4	4
11	7
13	9
7	5

Descriptive Statistics

SAS Code for Measures of Central Tendency, Measures of Spread, and Correlation

```
(1)   DATA CAREGIVER;
(2)         INPUT ANXIETY DEPRESSION;
(3)         LINES;
                  22 16
                  12 8
                  68 33
                  10 6
                  5 5
                  53 24
                  44 18
                  37 17
                  0 2
                  21 14
                  64 31
                  33 17
                  55 30
                  18 13
                  3 3
                  4 4
                  11 7
                  13 9
                  7 5
(4)   RUN;

(5)   PROC PRINT DATA=CAREGIVER;
(6)   PROC MEANS DATA=CAREGIVER;
(7)   PROC CORR DATA=CAREGIVER PLOTS=SCATTER(ELLIPSE=NONE);
(8)   RUN;
(9)   QUIT;
```

(1) The DATA step creates a new dataset and assigns it the specified name, CAREGIVER in this case.

(2) The INPUT statement creates the variable names and assigns the order of the variables to the new dataset. This command also assigns a variable type to each new variable. The default variable type, which applies to both variables in this case, is *numeric*. If a *character* variable is being created, put a dollar sign ($) in back of it. For example, if ANXIETY was a character variable, the code would read INPUT ANXIETY $ DEPRESSION.

(3) LINES indicates that the data are being entered manually (as opposed to read in or imported from a file). The data are listed beginning on the next line. Make sure that you enter the data in the same order that you entered the variable names in the INPUT statement (i.e. type the anxiety score first, followed by the depression score, because ANXIETY comes first and DEPRESSION comes second). *Please note that the data do NOT include a semicolon (;) at the end.*

(4) This RUN statement is optional, because SAS automatically ends the step that is currently running each time SAS encounters a new DATA or PROC step. Please note that, for some mysterious reason, this RUN statement does not turn dark blue like other RUN statements. ☺

(5) PROC PRINT is the "proc"edure that returns all of the raw data from CAREGIVER to your output window. Making use of this procedure is optional, but it's recommended when you enter data manually; it is a built-in check for data entry errors.

(6) PROC MEANS is the procedure that outputs the sample size, mean, standard deviation, minimum, and maximum for each numeric variable in CAREGIVER.

(7) PROC CORR is the procedure that runs the correlations for each pair of numeric variables in CAREGIVER, as well as the *p* value for each correlation. It also returns the same descriptive statistics as PROC MEANS (rendering that procedure redundant). In this procedure, a scatterplot is also requested, with the default prediction ellipses omitted.

(8) Each time SAS encounters a new DATA or PROC step, SAS automatically ends the step that is currently running. Because there are no more DATA or PROC steps in this code, the RUN step is needed in order to force SAS to end the PROC CORR step, effectively ending this code. The RUN step can be used at the end of any DATA or PROC step, not just the last one in the code. It is generally good practice to include RUN at the end of each DATA or PROC step.

(9) In SAS, some procedures will not end using RUN alone; a QUIT step is also required. If you forget to include the QUIT step, and SAS shows a procedure as still running at the top of the window, you can simply type "QUIT;" (without the quotation marks), then highlight and run this code; it will end any procedure still running. *It is simpler – and, therefore, recommended – to get in the habit of including a QUIT statement at the end of your code.*

Descriptive Statistics
SAS Code for Frequency

```
(10)   PROC FORMAT;
(11)         VALUE ANX_FMT
(12)               0-9 = '0-9'
                   10-19 = '10-19'
                   20-29 = '20-29'
                   30-39 = '30-39'
                   40-49 = '40-49'
                   50-59 = '50-59'
                   60-69 = '60-69';
(13)         VALUE DEP_FMT
                   0-4 = '0-4'
                   5-9 = '5-9'
                   10-14 = '10-14'
                   15-19 = '15-19'
                   20-24 = '20-24'
                   25-29 = '25-29'
                   30-34 = '30-34';
       RUN;

(14)   PROC FREQ DATA=CAREGIVER;
(15)         FORMAT ANXIETY ANX_FMT.;
(16)         FORMAT DEPRESSION DEP_FMT.;
       RUN;

(17)   PROC UNIVARIATE DATA=CAREGIVER;
(18)         VAR ANXIETY;
(19)         HISTOGRAM ANXIETY /
(20)               VSCALE=COUNT
(21)               ENDPOINTS = 0 TO 70 BY 10
(22)               ODSTITLE='Frequency of Anxiety Scores'
(23)               VAXISLABEL='Frequency';
(24)         LABEL ANXIETY='Anxiety';
       RUN;

(25)   PROC UNIVARIATE DATA=CAREGIVER;
             VAR DEPRESSION;
             HISTOGRAM DEPRESSION /
                   VSCALE=COUNT
                   ENDPOINTS = 0 TO 35 BY 5
                   ODSTITLE='Frequency of Depression Scores'
                   VAXISLABEL='Frequency';
             LABEL DEPRESSION='Depression';
(26)   RUN;
(27)   QUIT;
```

(10) PROC FORMAT is used to create guidelines for formatting a variable. In this case, the procedure will create a format that specifies the ranges, or bins, for the frequency analysis. Your bins should always include the minimum observed value for the variable, the maximum observed value, and all values in between *even if unobserved*. For example, if

your variable X has values of 1, 2, 3, 4, and 10, you need to make sure that (a) your lowest bin includes the value 1, (b) your highest bin includes the value 10, and (c) you have bins for 5, 6, 7, 8, and 9, even though X does not include these values. If you omit bins for 5-9, your X data will be misrepresented.

(11) The VALUE statement assigns a name to the format you are creating. In this case, we are creating the formatting for ANXIETY, which is reflected in the name assigned here (ANX_FMT).

(12) Beginning on this line, each bin is created and assigned a label. Values of 0-9 (on the left side of the equal sign), for example, are assigned a label of "0-9" (on the right side of the equal sign). Because the label is contained inside quotation marks, you can include any character in the label (e.g., <, >, $, letters/words, etc.). For example, you could write: 0-9 = '≤ 9'. ANX_FMT has no impact on your output yet, because it has not been applied to a particular variable.

(13) The VALUE statement assigns a name to the format you are creating. In this case, we are creating the formatting for DEPRESSION, which is reflected in the name assigned here (DEP_FMT). Following this statement, you will see the creation and labeling of the bins, as in (12).

(14) PROC FREQ generates a frequency table for each numeric variable in CAREGIVER.

(15) The FORMAT statement tells SAS that you want a frequency table, but that you need to override the default frequency table format. The default for numeric variables is to include a bin for every single observed value. As we want the frequency by ranges, this statement is used to apply the ANX_FMT range values and labels to ANXIETY for this particular analysis. (It does NOT change the actual ANXIETY data.) *Note that a period (.) follows ANX_FMT. The code will not work properly if you omit the period.*

(16) Just as in (15), the FORMAT statement overrides the default frequency table, applying the DEP_FMT values and labels to DEPRESSION. *Note that a period (.) follows DEP_FMT. The code will not work properly if you omit the period.*

(17) PROC UNIVARIATE provides descriptive statistics for each variable, and renders PROC MEANS redundant. In fact, PROC UNIVARIATE provides a greater variety of descriptive statistics than PROC MEANS, including median, mode, skewness, and kurtosis. PROC UNIVARIATE is used here to generate histograms.

(18) The VAR statement limits the analysis in the PROC UNIVARIATE step to ANXIETY (i.e. DEPRESSION is omitted).

(19) The HISTOGRAM statement creates histograms for the specified variable(s). Here, a histogram is created for ANXIETY. The forward slash (/) is used to separate the call for a histogram of ANXIETY from the options that follow.

(20) VSCALE is the first option in relation to the histogram being created. The default scale for the vertical axis is the frequency *percent* (e.g., a particular value occurs 20% of the time). VSCALE overrides this and sets the vertical axis to the frequency *count* (e.g., a particular value occurs 5 times).

(21) The ENDPOINTS option sets the bin ranges, in this case from 0 to 70 in increments of 10. (This matches ANX_FMT.)
(22) The ODSTITLE option assigns a title to the histogram.
(23) The VAXISLABEL option overrides the default vertical axis label ("Count") and assigns a different label. This is optional; you can always use the default label. As this is the last option in the code, it ends with a semicolon (;).
(24) In order to change the horizontal axis label, the LABEL statement is used. This is optional; the default is the name of the variable being represented in the histogram (ANXIETY in this case).
(25) Here, PROC UNIVARIATE is used to create a histogram for DEPRESSION. The code matches that of the PROC UNIVARIATE code for ANXIETY, except for the variable included (DEPRESSION in this case), and title, labels, and bin ranges of the histogram. (The ENDPOINTS option sets the bin ranges from 0 to 35 in increments of 5; this matches DEP_FMT.)
(26) Remember that it is generally good practice to include RUN at the end of each DATA or PROC step; in this code it is used to end PROC FORMAT, PROC FREQ, and PROC UNIVARIATE. You may decide not to use RUN in the first two instances – although this is NOT recommended – but you MUST use it at the end of the code.
(27) Remember that some procedures will not end using RUN alone; a QUIT step is also required. Therefore, you should include it at the end of every code.

Descriptive Statistics
Selected Output

It is always a good idea to verify that your data have been input/imported correctly.

Obs	ANXIETY	DEPRESSION
1	22	16
2	12	8
3	68	33
4	10	6
5	5	5
6	53	24
7	44	18
8	37	17
9	0	2
10	21	14
11	64	31
12	33	17
13	55	30
14	18	13
15	3	3
16	4	4
17	11	7
18	13	9
19	7	5

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
ANXIETY	19	25.2631579	21.9894978	0	68.0000000
DEPRESSION	19	13.7894737	9.8464824	2.0000000	33.0000000

The CORR Procedure

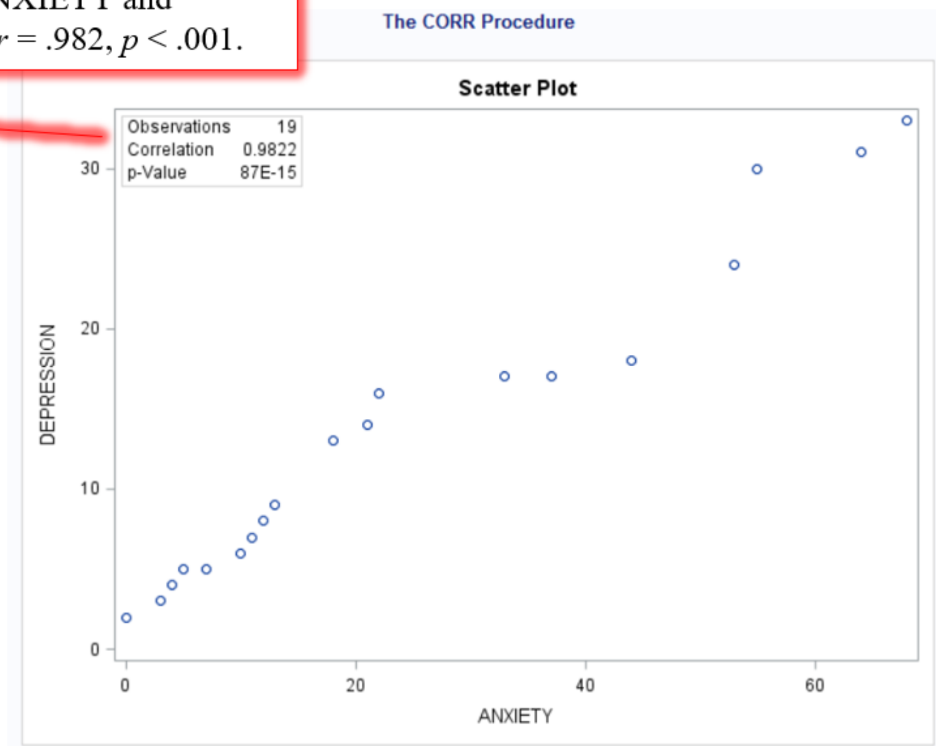
2 Variables:	ANXIETY DEPRESSION

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
ANXIETY	19	25.26316	21.98950	480.00000	0	68.00000
DEPRESSION	19	13.78947	9.84648	262.00000	2.00000	33.00000

Pearson Correlation Coefficients, N = 19 Prob > |r| under H0: Rho=0		
	ANXIETY	DEPRESSION
ANXIETY	1.00000	0.98222 <.0001
DEPRESSION	0.98222 <.0001	1.00000

There is a significant relationship between ANXIETY and DEPRESSION, $r = .982$, $p < .001$.

There is a significant relationship between ANXIETY and DEPRESSION, $r = .982$, $p < .001$.



The FREQ Procedure

ANXIETY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0-9	5	26.32	5	26.32
10-19	5	26.32	10	52.63
20-29	2	10.53	12	63.16
30-39	2	10.53	14	73.68
40-49	1	5.26	15	78.95
50-59	2	10.53	17	89.47
60-69	2	10.53	19	100.00

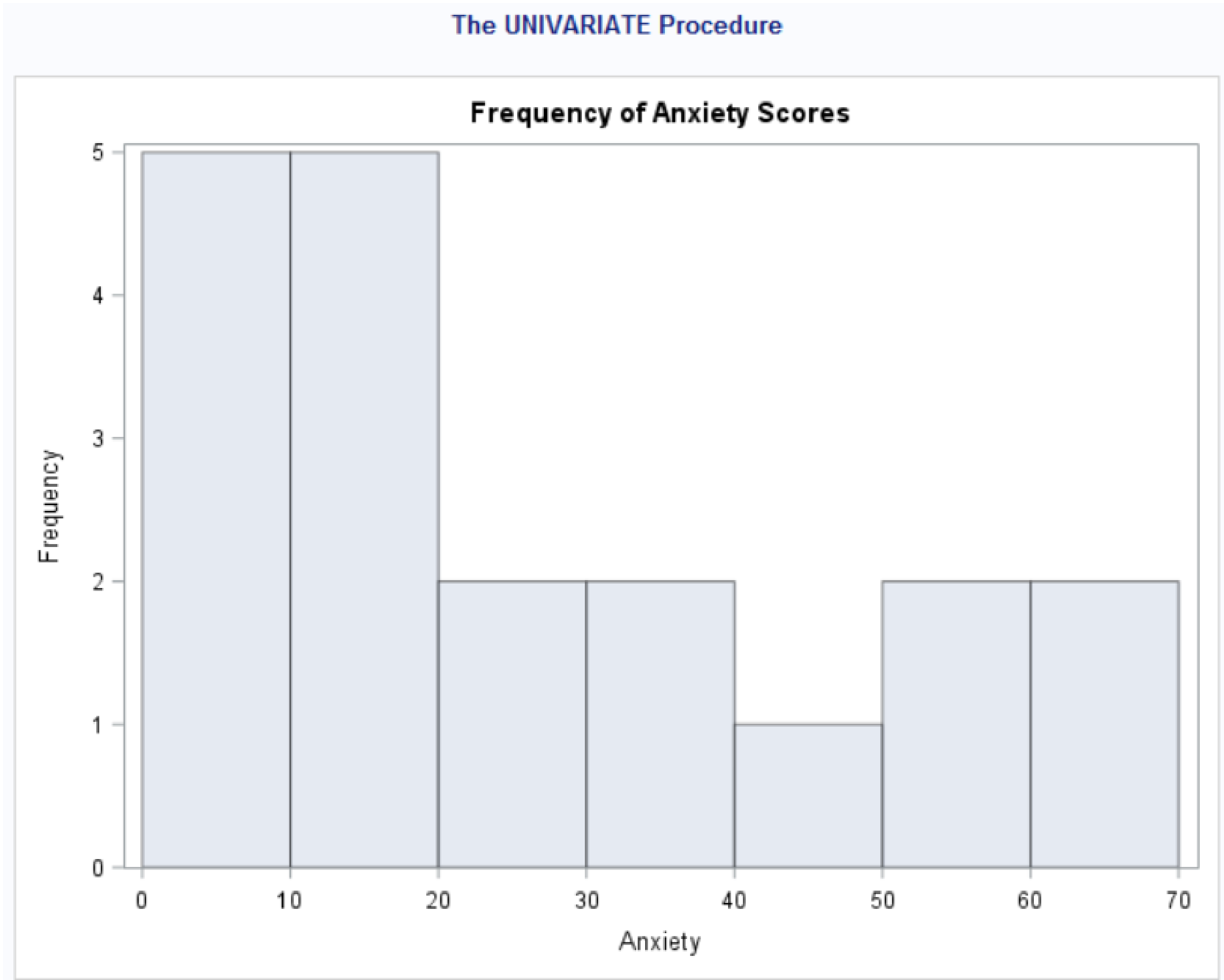
DEPRESSION	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0-4	3	15.79	3	15.79
5-9	6	31.58	9	47.37
10-14	2	10.53	11	57.89
15-19	4	21.05	15	78.95
20-24	1	5.26	16	84.21
30-34	3	15.79	19	100.00

ANXIETY

The UNIVARIATE Procedure
Variable: ANXIETY (Anxiety)

Moments			
N	19	Sum Weights	19
Mean	25.2631579	Sum Observations	480
Std Deviation	21.9894978	Variance	483.538012
Skewness	0.76009557	Kurtosis	-0.7861464
Uncorrected SS	20830	Corrected SS	8703.68421
Coeff Variation	87.041762	Std Error Mean	5.04473677

Basic Statistical Measures			
Location		Variability	
Mean	25.26316	Std Deviation	21.98950
Median	18.00000	Variance	483.53801
Mode	.	Range	68.00000
		Interquartile Range	37.00000


The UNIVARIATE Procedure
Frequency of Anxiety Scores
Frequency
Anxiety
0
1
2
3
4
5
0
10
20
30
40
50
60
70

DEPRESSION

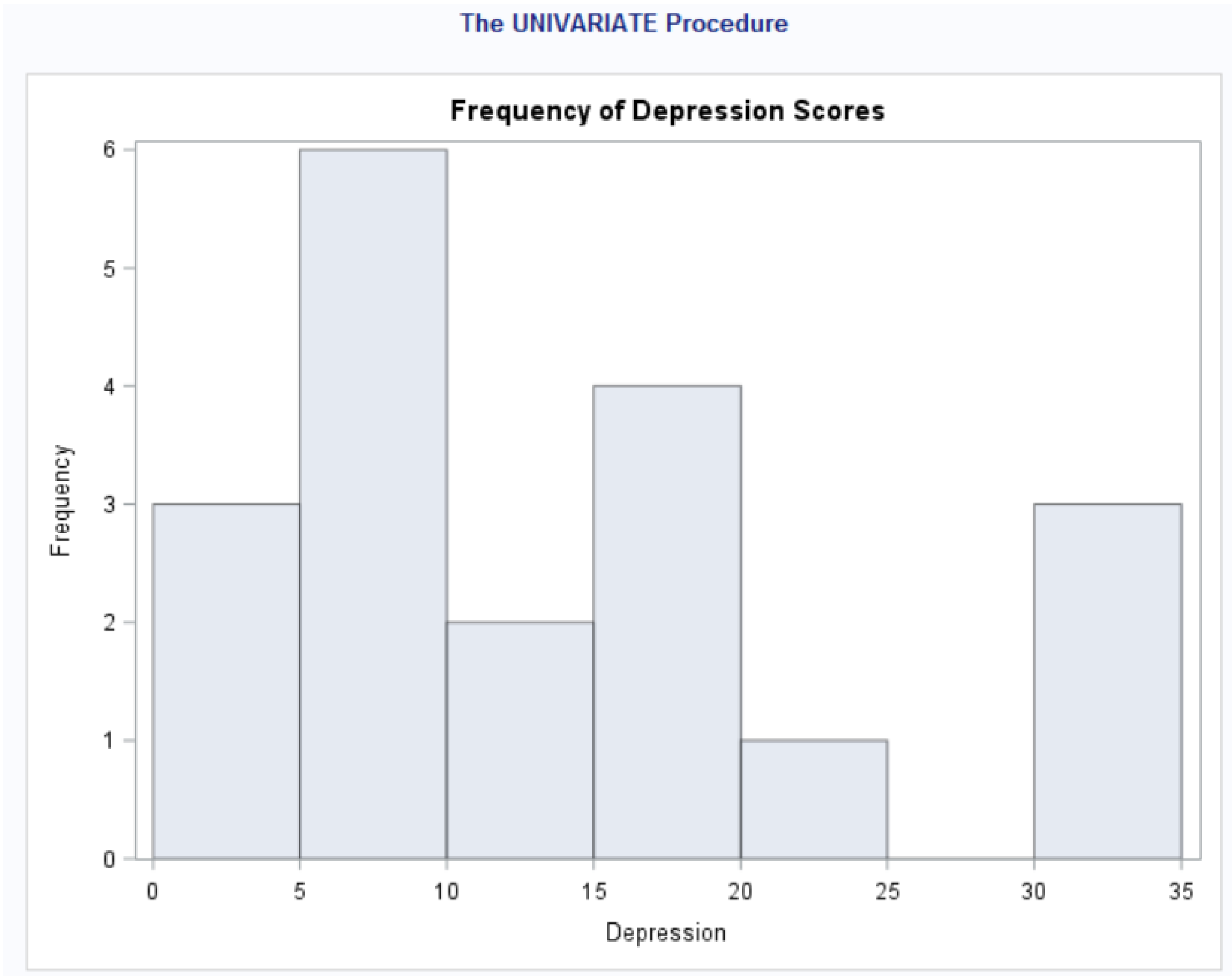
The UNIVARIATE Procedure
Variable: DEPRESSION (Depression)

Moments			
N	19	Sum Weights	19
Mean	13.7894737	Sum Observations	262
Std Deviation	9.84648244	Variance	96.9532164
Skewness	0.74032431	Kurtosis	-0.5527289
Uncorrected SS	5358	Corrected SS	1745.15789
Coeff Variation	71.4057887	Std Error Mean	2.25893799

Basic Statistical Measures			
Location		Variability	
Mean	13.78947	Std Deviation	9.84648
Median	13.00000	Variance	96.95322
Mode	5.00000	Range	31.00000
		Interquartile Range	13.00000

Special Note

Note: The mode displayed is the smallest of 2 modes with a count of 2.


The UNIVARIATE Procedure
Frequency of Depression Scores
Frequency
Depression
0
1
2
3
4
5
6
0
5
10
15
20
25
30
35

Inferential Statistics
One-Sample *t* Test
Research Scenario

Suppose that Professor Coffey learns from a national survey that the average high school student in the United States spends 6.75 hours each week exploring particular websites on the internet. The professor is interested in knowing how internet use among students at the local high school compares with this national average. Is local use more than, or less than, this average? Professor Coffey randomly selects a sample of 10 students. Each student is asked to report the number of hours he or she spends exploring these websites on the internet in a typical week during the school year. The data appear below.

Student	Number of Hours
1	6
2	9
3	12
4	3
5	11
6	10
7	18
8	9
9	13
10	8

Source of Data and Scenario: Coladarci, T., & Cobb, C. (2014). *Fundamentals of Statistical Reasoning in Education* (4th ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Inferential Statistics
One-Sample *t* Test
SAS Code

```
       DATA STUDYHOURS;
              INPUT HOURS;
              LINES;
                     6
                     9
                     12
                     3
                     11
                     10
                     18
                     9
                     13
                     8
       RUN;

       PROC PRINT DATA=STUDYHOURS;
       RUN;

       PROC UNIVARIATE DATA=STUDYHOURS;
              VAR HOURS;
              HISTOGRAM HOURS /
                     VSCALE=COUNT
                     ENDPOINTS = 0 TO 20 BY 1
                     ODSTITLE='Frequency of Study Hours'
                     VAXISLABEL='Frequency';
              LABEL HOURS='Hours';
       RUN;

(1)    PROC TTEST DATA=STUDYHOURS PLOTS(shownull)=interval H0=6.75;
(2)           VAR HOURS;
(3)           TITLE "One-Sample t Test: Study Hours";
       RUN;

(4)    TITLE;
       QUIT;
```

(1) PROC TTEST may be used for one-sample *t* tests, as well as independent and dependent *t* tests. The PLOTS(shownull) option will add a reference line for the null hypothesis value (i.e. the national average of 6.75). Setting “H0” to 6.75 will make this a one-sample *t* test.
(2) The one-sample *t* test is being conducted for the variable HOURS.
(3) An optional title is being assigned to the output. Although this is not necessary here, it becomes very helpful in more advanced analyses that contain a lot of output, which may easily become confusing. If you do not want a title, simply remove this line of code.
(4) *The title that was created in (3) will be applied to all future SAS output indefinitely*. For example, if you were to run PROC MEANS or PROC FREQ after this code, it would label the output “One-Sample t Test: Study Hours,” even though they have nothing to do with *t* tests. You must take one of three explicit actions to clear this title: (a) exit and restart

SAS, (b) assign a new title, or (c) type “TITLE;” (without the quotation marks) to reset the title to SAS defaults. Option C is being enacted here to reset default titles.

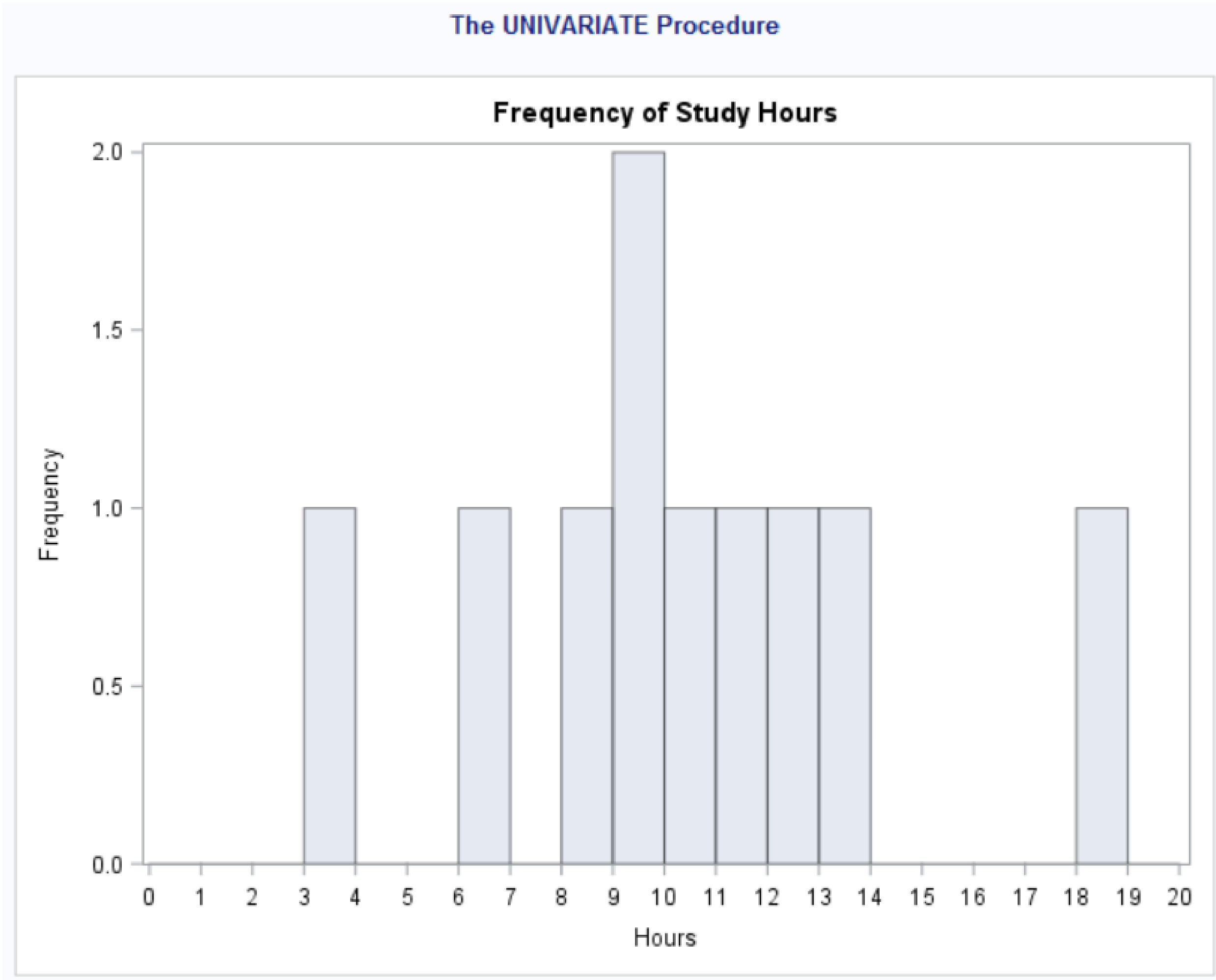
Inferential Statistics
One-Sample *t* Test
Selected Output

Obs	HOURS
1	6
2	9
3	12
4	3
5	11
6	10
7	18
8	9
9	13
10	8

The UNIVARIATE Procedure
Variable: HOURS (Hours)

Moments			
N	10	Sum Weights	10
Mean	9.9	Sum Observations	99
Std Deviation	4.06748626	Variance	16.5444444
Skewness	0.36258643	Kurtosis	1.18673013
Uncorrected SS	1129	Corrected SS	148.9
Coeff Variation	41.0857198	Std Error Mean	1.28625209

Basic Statistical Measures			
Location		Variability	
Mean	9.900000	Std Deviation	4.06749
Median	9.500000	Variance	16.54444
Mode	9.000000	Range	15.00000
		Interquartile Range	4.00000



The mean of HOURS is 9.9 hours.

One-Sample t Test: Study Hours

The TTEST Procedure

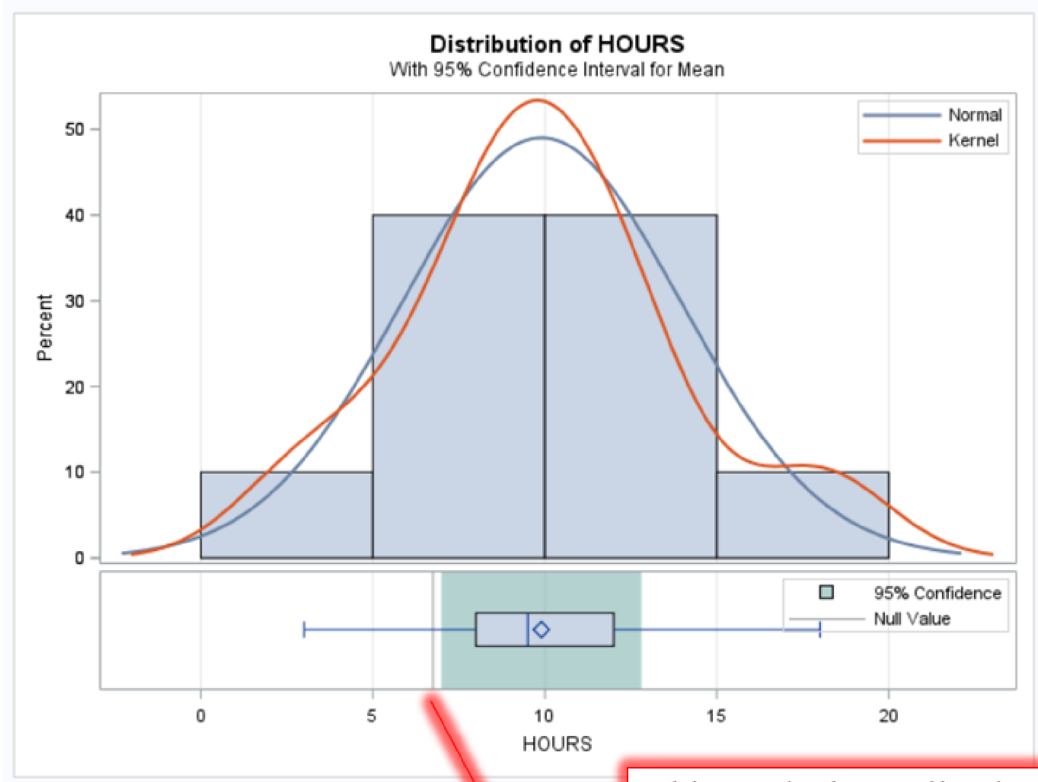
Variable: HOURS

N	Mean	Std Dev	Std Err	Minimum	Maximum
10	9.9000	4.0675	1.2863	3.0000	18.0000

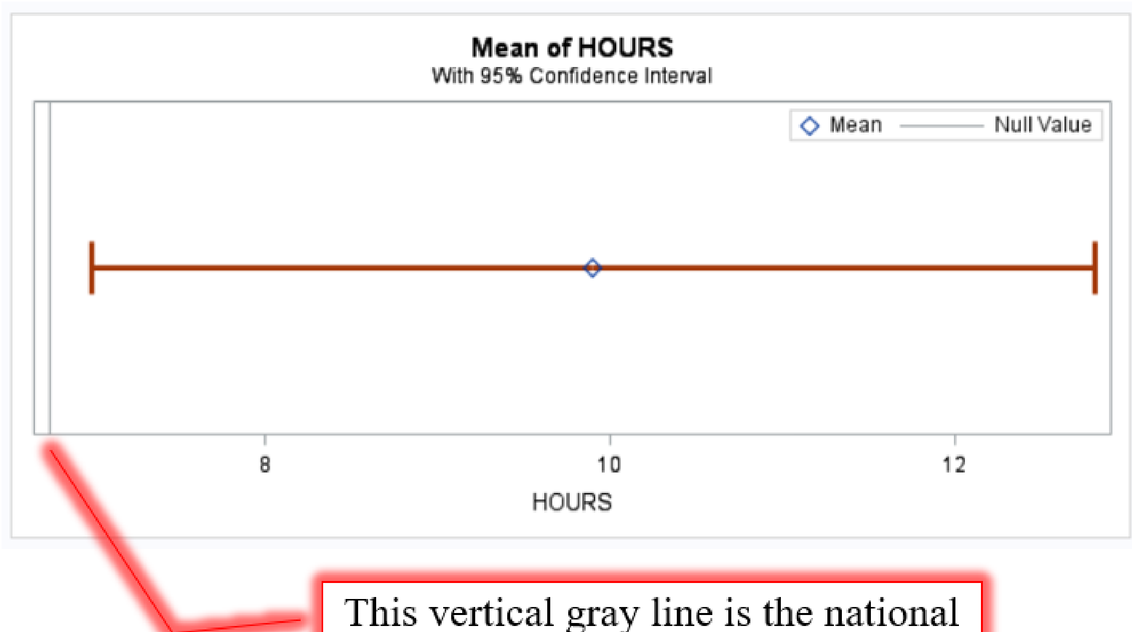
Mean	95% CL Mean		Std Dev	95% CL Std Dev	
9.9000	6.9903	12.8097	4.0675	2.7978	7.4256

DF	t Value	Pr > |t|
9	2.45	0.0368

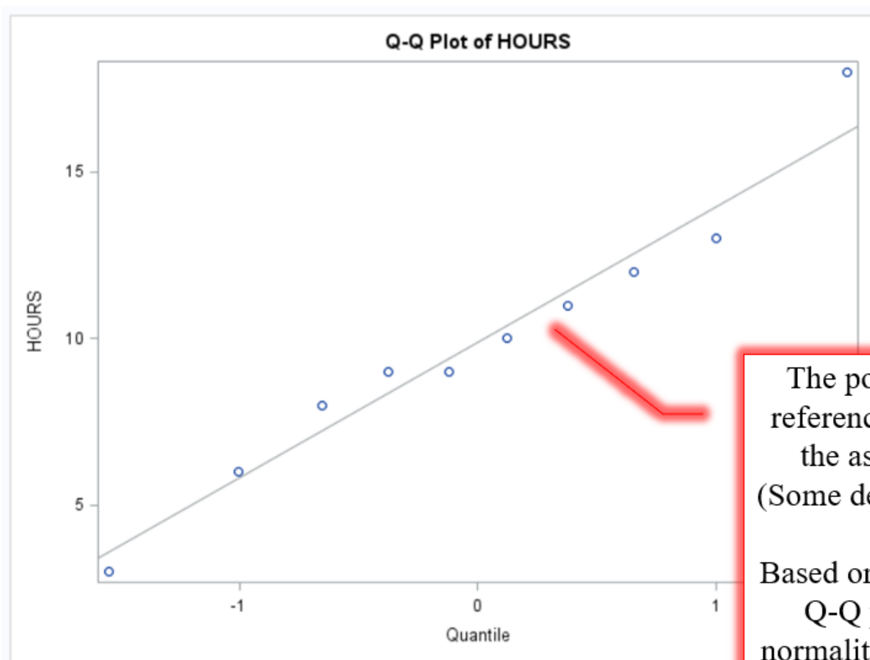
The mean of HOURS is significantly different from the national average of 6.75 hours, $t(9) = 2.45$, $p = .037$.



This vertical gray line is the national mean of 6.75 hours (i.e. the value being tested in the null hypothesis).



This vertical gray line is the national mean of 6.75 hours (i.e. the value being tested in the null hypothesis).



The points appear to "hug" the reference line, lending support to the assumption of normality. (Some deviation is to be expected.)

Based on a visual inspection of the Q-Q plot, the assumption of normality was found to be tenable.

Inferential Statistics
Independent *t* Test
Research Scenario

A training manager believes that a new interactive computer-based training package will help improve the production rate of order assemblers. She arranges for a production area of 21 experienced employees to complete the new training package over a six-week period. Another group from the production area of 23 employees received no additional training. The following are the average production rates per person per hour based on a 12-week period following the training.

Training Group (T)	Control Group (C)
26	44
45	45
60	57
73	28
45	64
51	39
63	35
46	43
69	21
51	56
55	22
58	87
61	48
54	12
64	19
56	62
59	55
35	44
48	39
45	44
59	57
--	30
--	30

Inferential Statistics
Independent *t* Test
SAS Code

```
(1)     PROC FORMAT;
(2)           VALUE $TRTMNT_FMT
(3)                 "T"="Training"
                    "C"="Control";
        RUN;

        DATA PRODUCTION;
(4)           INPUT TREATMENT $ PRODRATE @@;
(5)           FORMAT TREATMENT $TRTMNT_FMT.;
              LINES;
                    T 26 T 45 T 60 T 73 T 45
                    T 51 T 63 T 46 T 69 T 51
                    T 55 T 58 T 61 T 54 T 64
                    T 56 T 59 T 35 T 48 T 45
                    T 59 C 44 C 45 C 57 C 28
                    C 64 C 39 C 35 C 43 C 21
                    C 56 C 22 C 87 C 48 C 12
                    C 19 C 62 C 55 C 44 C 39
                    C 44 C 57 C 30 C 30
        RUN;

        PROC PRINT DATA=PRODUCTION;
        RUN;

(6)     PROC GLM DATA=PRODUCTION PLOTS=DIAGNOSTICS;
(7)           CLASS TREATMENT;
(8)           MODEL PRODRATE=TREATMENT;
(9)           MEANS TREATMENT / HOVTEST=LEVENE(TYPE=ABS);
(10)          TITLE "Independent t Test: Production Rate";
        RUN;

(11)    TITLE;
        QUIT;
```

(1) The PROC FORMAT step is used here to create labels for the TREATMENT codes.

(2) The VALUE statement assigns a name to the format you are creating, TRTMNT_FMT in this case, and matches each code with a label. The dollar sign ($) at the front of TRTMNT_FMT indicates that this format will be applied to a character variable. (If the $ is absent, SAS defaults to applying the format to a numeric variable.) *Remember to use quotation marks (“”) with character data and labels.*

(3) Beginning on this line, individual labels are assigned. For example, “T” is assigned the label “Training.” TRTMNT_FMT has no impact on your output yet, because it has not been applied to a particular variable.

(4) The INPUT statement creates a character variable TREATMENT (the $ that follows it indicates that it is a character variable) and a numeric variable PRODRATE (the lack of $

means that it defaults to numeric). TREATMENT is coded as "T" for employees who participated in the training package and "C" for those in the control group. PRODRATE is the rate of production. The @@ symbols tell SAS that there are multiple observations on a single line of data. For example, the first line of data (T 26 T 45 T 60 T 73 T 45) contains the data for five observations (i.e. employees). This is optional, but tends to save space in longer datasets.

(5) The FORMAT statement is used to apply the labels created in TRTMNT_FMT (1) to TREATMENT. *Note that a period (.) follows TRTMNT_FMT; the code will not work properly if you omit the period.*

(6) PROC GLM (General Linear Model) is used to conduct regression analysis; a *t* test is one specific type of regression analysis. In this case, an independent *t* test is being conducted on the PRODUCTION data. The diagnostic plots have also been requested here; these will be used to verify the tenability of the *t* test assumptions (i.e. that the assumptions have been met). Specifically, these plots will help when verifying that the data are normally distributed.

(7) The CLASS statement is used to identify variables with nominal data (as opposed to continuous data); in other words, use a CLASS statement to identify categorical variables. In this case, you have the categorical/grouping variable TREATMENT.

(8) The MODEL statement is written as Dependent Variable (DV) = Independent Variable (IV). *Again, the DV goes on the left of the equal sign and the IV goes on the right.*

(9) The MEANS statement requires SAS to compute the means for each group in TREATMENT. The option HOVTEST=LEVENE calls for the Levene's Test of Homogeneity of Variances, to test the assumption that the groups have equal variances. There are multiple homogeneity tests available; Levene's Test is the default. There are two methods for calculating Levene's statistic: using the absolute residuals (TYPE=ABS) or the squared residuals (TYPE=SQUARE, the default). As the users of this guide may be more familiar with SPSS or may be referencing *SPSS: A Survival Guide for EPRS 8530 and EPRS 8540*, the TYPE option is set to the method used by SPSS: ABS.

(10) An optional title is being assigned to the output. Although this is not necessary here, it becomes very helpful in more advanced analyses that contain a lot of output, which may easily become confusing. If you do not want a title, simply remove this line of code.

(11) *The title that was created in (10) will be applied to all future SAS output indefinitely.* For example, if you were to run PROC MEANS or PROC FREQ after this code, it would label the output "Independent t Test: Production Rate," even though they have nothing to do with *t* tests. You must take one of three explicit actions to clear this title: (a) exit and restart SAS, (b) assign a new title, or (c) type "TITLE;" (without the quotation marks) to reset the title to SAS defaults. Option C is being enacted here to reset default titles.

Inferential Statistics
Independent *t* Test
Selected Output

Obs	TREATMENT	PRODRATE
1	Training	26
2	Training	45
3	Training	60
4	Training	73
5	Training	45
6	Training	51
7	Training	63
8	Training	46
9	Training	69
10	Training	51
11	Training	55
12	Training	58
13	Training	61
14	Training	54
15	Training	64
16	Training	56
17	Training	59
18	Training	35
19	Training	48
20	Training	45
21	Training	59
22	Control	44
23	Control	45
24	Control	57

It is always a good idea to verify that your data have been input/imported correctly.

The GLM Procedure

Class Level Information		
Class	Levels	Values
TREATMENT	2	Control Training

Number of Observations Read	44
Number of Observations Used	44

It is always a good idea to verify that your data have been input/imported correctly.

Independent t Test: Production Rate

The GLM Procedure

Dependent Variable: PRODRATE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1286.08997	1286.08997	5.99	0.0186
Error	42	9016.45549	214.67751		
Corrected Total	43	10302.54545			

R-Square	Coeff Var	Root MSE	PRODRATE Mean
0.124832	30.64081	14.65188	47.81818

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TREATMENT	1	1286.089968	1286.089968	5.99	0.0186

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TREATMENT	1	1286.089968	1286.089968	5.99	0.0186

There was a significant difference in production rates between the training and control groups, $t(42) = 2.447$, $p = .019$.

PROC GLM reports F values, not t values. The conversion is very simple: $t^2 = F$.

For an independent t test, the degrees of freedom are equal to $n_1 + n_2 - 2$. In this case, $df = 21 + 23 - 2 = 42$.

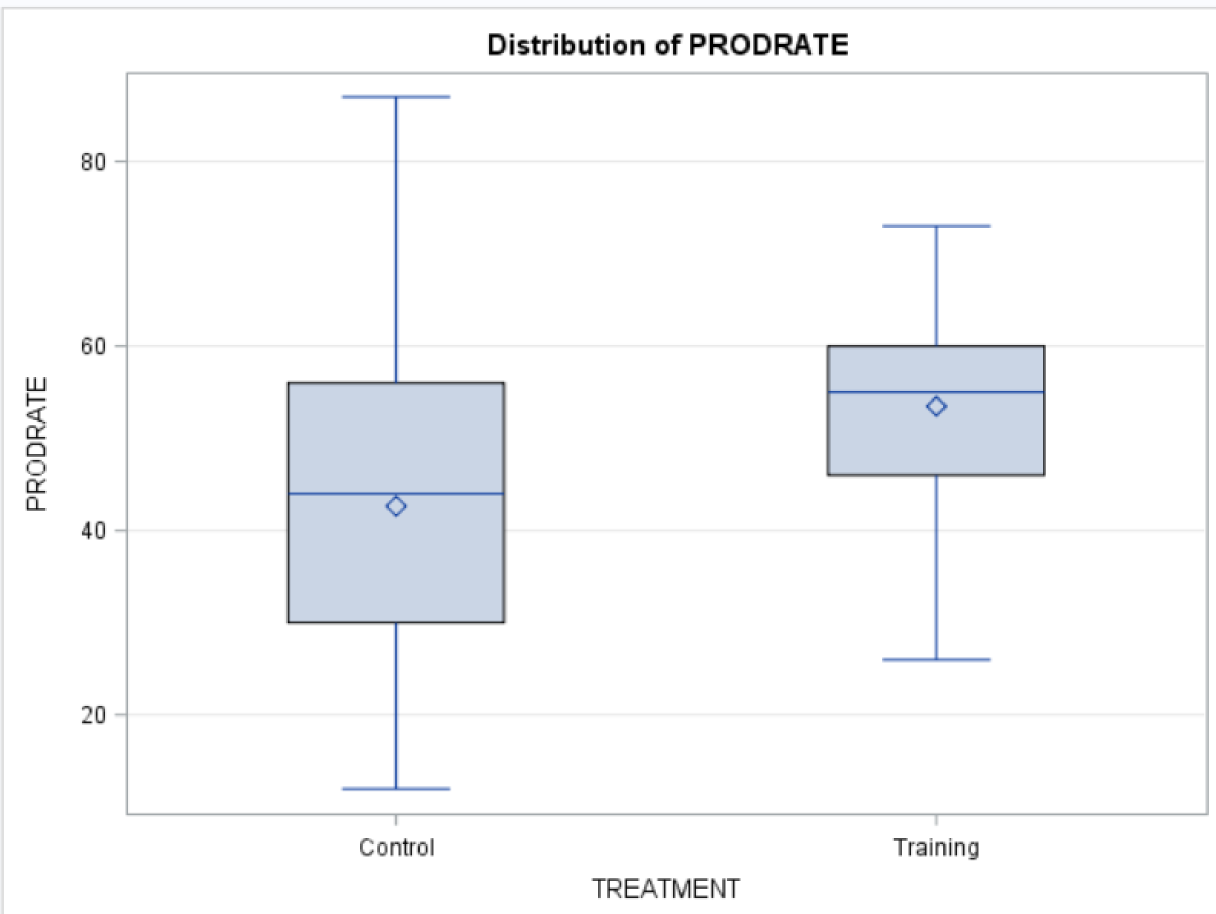
The assumption of homogeneity of variances was found to be tenable, $F(1, 42) = 2.87$, $p = .098$.

The GLM Procedure

Levene's Test for Homogeneity of PRODRATE Variance ANOVA of Absolute Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
TREATMENT	1	239.6	239.6	2.87	0.0976
Error	42	3505.1	83.4559		

Independent t Test: Production Rate

The GLM Procedure



Level of TREATMENT	N	PRODRATE Mean	PRODRATE Std Dev
Control	23	42.6521739	17.3116059
Training	21	53.4761905	11.0073568

Group means

Inferential Statistics
Dependent *t* Test
Research Scenario

A sports psychologist was interested in testing the effect of a simple relaxation technique on college basketball players' free throw shooting accuracy. Each player was asked to shoot 20 consecutive free throws and the number of successful attempts was recorded. The players were then trained to use a simple five-second relaxation technique while preparing to shoot a free throw. The players then returned to the court and shot 20 consecutive free throws again. The resulting data are given below.

Pre-Training	Post-Training
12	13
15	15
9	11
16	15
12	15
15	18
17	17
10	12
12	13
14	17

Inferential Statistics
Dependent *t* Test
SAS Code

```
      DATA THROWS;
            INPUT THROWS_PRE THROWS_POST @@;
            LINES;
                  12 13 15 15 9 11
                  16 15 12 15 15 18
                  17 17 10 12 12 13
                  14 17
      RUN;

      PROC PRINT DATA=THROWS;
      RUN;

(1)   PROC MEANS DATA=THROWS;
      RUN;

(2)   PROC TTEST DATA=THROWS PLOTS(SHOWH0)=INTERVAL;
(3)         PAIRED THROWS_PRE*THROWS_POST;
            TITLE "Dependent t Test: Free Throws";
      RUN;

      TITLE;
      QUIT;
```

(1) PROC MEANS provides the means and standard deviations for each variable in the THROWS dataset (THROWS_PRE and THROWS_POST).

(2) PROC TTEST will perform a *t* test on the THROWS data. PROC TTEST includes several useful plots in the output by default. The SHOWH0 option shows the null hypothesis (H_0) critical value in the appropriate plots; the INTERVAL option adds an additional plot that will aid in interpreting the results. [*Note: PROC TTEST can also be used with independent* t *tests. The reason that it was omitted from the previous section is because PROC TTEST does not offer Levene's Test to assess the homogeneity of variances assumption (it uses a different test). Therefore, PROC GLM, which defaults to Levene's Test, was used.*]

(3) The PAIRED statement makes this a **dependent** *t* test. The difference scores are calculated as the score on the left of the asterisk (*) minus the score on the right of the asterisk (*). In this case, the analysis is THROWS_PRE minus THROWS_POST. This may seem backwards, as the resulting difference values will be negative if the number of free throws increases after the treatment. The reason they have been placed in this order is so that THROWS_PRE is placed on the *x*-axis and THROWS_POST is placed on the *y*-axis in applicable output graphs. The negative sign on the difference values and the corresponding *t* value can be disregarded. Please note that you can reverse this order if

you prefer; the only consequence will be the reversal of the variables on the axes of the resulting plots.

Inferential Statistics
Dependent *t* Test
Selected Output

Obs	THROWS_PRE	THROWS_POST
1	12	13
2	15	15
3	9	11
4	16	15
5	12	15
6	15	18
7	17	17
8	10	12
9	12	13
10	14	17

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
THROWS_PRE	10	13.2000000	2.6161889	9.0000000	17.0000000
THROWS_POST	10	14.6000000	2.3190036	11.0000000	18.0000000

The THROWS_PRE mean was 13.2; the THROWS_POST mean was 14.6. Therefore, the mean difference is –1.4.

Dependent t Test: Free Throws

The TTEST Procedure

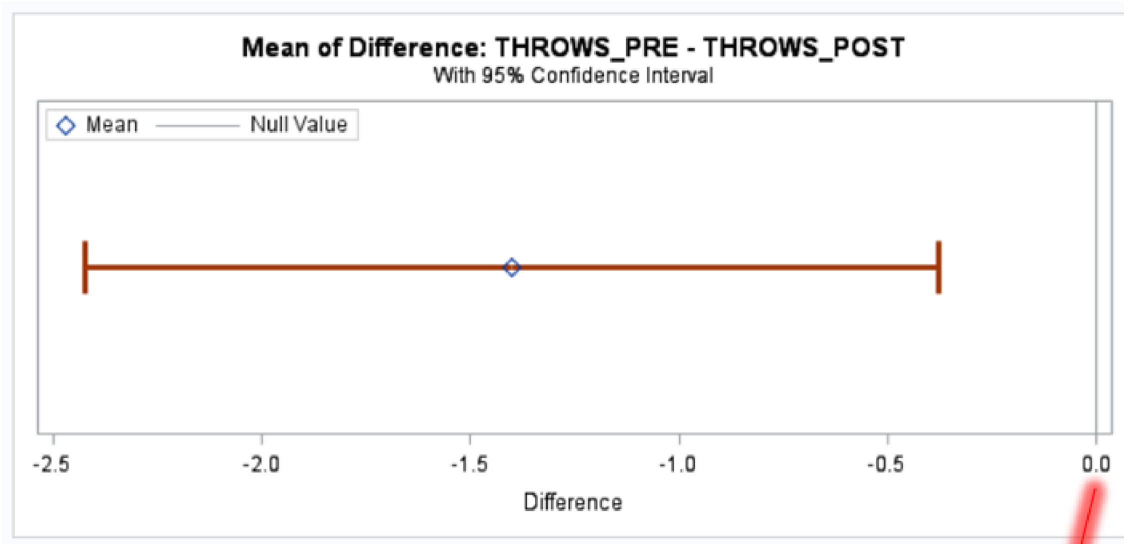
Difference: THROWS_PRE - THROWS_POST

N	Mean	Std Dev	Std Err	Minimum	Maximum
10	-1.4000	1.4298	0.4522	-3.0000	1.0000

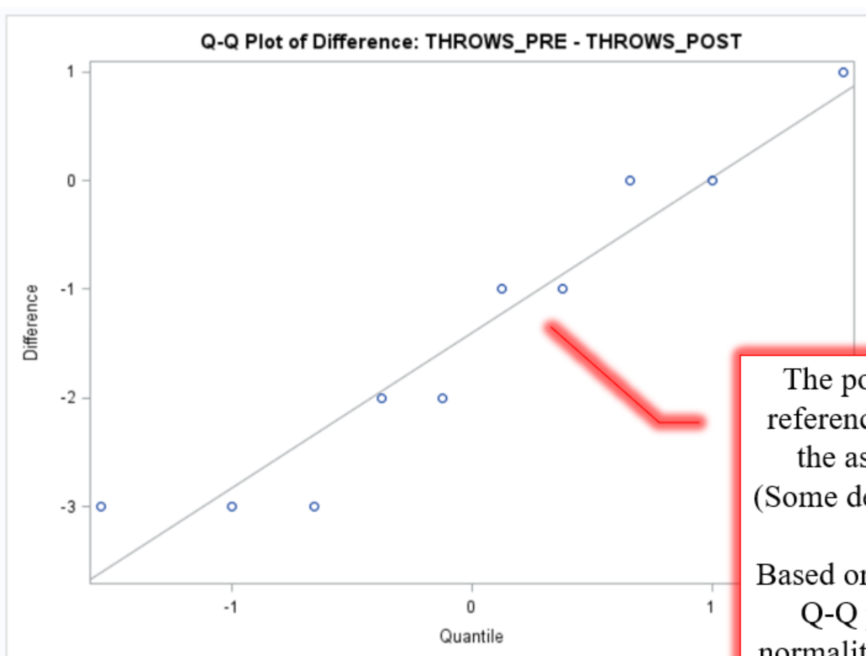
Mean	95% CL Mean		Std Dev	95% CL Std Dev	
-1.4000	-2.4228	-0.3772	1.4298	0.9835	2.6103

DF	t Value	Pr > |t|
9	-3.10	0.0128

The relaxation technique had a significant impact on free throws scored, $t(9) = -3.10$, $p = .013$.



This vertical gray line is the mean difference of zero (i.e. the value being tested in the null hypothesis).



The points appear to "hug" the reference line, lending support to the assumption of normality. (Some deviation is to be expected.)

Based on a visual inspection of the Q-Q plot, the assumption of normality of difference scores was found to be tenable.

Inferential Statistics
One-Way ANOVA
Research Scenario

A direct marketer of insurance wanted to evaluate the effect of age on the response rate to a new insurance product. Below are the response rates per 1,000 mailings by age group from 12 different metropolitan areas.

Young	Middle-Aged	Elderly
25	30	25
27	29	22
23	29	27
24	31	23
23	28	24
24	31	25
22	29	23
25	32	22
21	30	21
24	29	22
21	28	24
23	31	23

Inferential Statistics
One-Way ANOVA
SAS Code

```
      PROC FORMAT;
(1)         VALUE AGE_FMT
                  1="1. Young"
                  2="2. Middle-Aged"
                  3="3. Elderly";
      RUN;

      DATA INSURANCE;
            INPUT AGE RATE @@;
            FORMAT AGE AGE_FMT.;
            LINES;
                  1 25 1 27 1 23 1 24 1 23 1 24 1 22 1 25 1 21 1 24 1 21 1 23
                  2 30 2 29 2 29 2 31 2 28 2 31 2 29 2 32 2 30 2 29 2 28 2 31
                  3 25 3 22 3 27 3 23 3 24 3 25 3 23 3 22 3 21 3 22 3 24 3 23
      RUN;

      PROC PRINT DATA=INSURANCE;
      RUN;

      PROC GLM DATA=INSURANCE PLOTS=DIAGNOSTICS;
(2)         CLASS AGE;
(3)         MODEL RATE=AGE;
(4)         MEANS AGE / HOVTEST=LEVENE(TYPE=ABS);
(5)         LSMEANS AGE / PDIFF=ALL ADJUST=TUKEY;
            TITLE "One-Way ANOVA: Insurance Response Rate";
      RUN;

      TITLE;
      QUIT;
```

(1) When labels are applied to the values (code numbers or letters) of a variable, *the output will show these labels and their corresponding data in alphabetical order*. In this case, AGE is coded as 1 = "Young," 2 = "Middle," and 3 = "Elderly." If these labels were used in the VALUE statement (instead of the labels shown in the code above), then the output would appear in the order "Elderly," "Middle," and "Young" (i.e. alphabetical order). This may make the output confusing for the reader, who would expect the axis displaying AGE to go from youngest to oldest. The easiest way to correct this is to include the code number in the label (e.g., "1. Young"), as "1" comes first in alphabetical order, followed by "2. Middle-Aged" and "3. Elderly." Now, alphabetical order and logical order will match.

(2) The CLASS statement identifies AGE as a nominal/categorical/grouping variable.

(3) The MODEL statement is written as Dependent Variable (DV) = Independent Variable (IV).

(4) The MEANS statement requires SAS to compute the means for each group in AGE. The option HOVTEST=LEVENE calls for the Levene's Test of Homogeneity of Variances, to test the assumption that the groups have equal variances. There are multiple homogeneity

tests available; Levene's Test is the default. There are two methods for calculating Levene's statistic: using the absolute residuals (TYPE=ABS) or the squared residuals (TYPE=SQUARE, the default). As the users of this guide may be more familiar with SPSS or may be referencing *SPSS: A Survival Guide for EPRS 8530 and EPRS 8540,* the TYPE option is set to the method used by SPSS: ABS.

(5) *The LSMEANS statement shown here is only required if the one-way ANOVA is found to be significant.* The LSMEANS statement will perform Tukey's HSD post-hoc testing by AGE. The PDIFF=ALL option requests the *p* values for each of the pairwise comparisons. If post-hoc testing is not required, you may remove this line of code.

Inferential Statistics
One-Way ANOVA
Selected Output

Obs	AGE	RATE
1	1. Young	25
2	1. Young	27
3	1. Young	23
4	1. Young	24
5	1. Young	23
6	1. Young	24
7	1. Young	22
8	1. Young	25
9	1. Young	21
10	1. Young	24
11	1. Young	21
12	1. Young	23
13	2. Middle-Aged	30
14	2. Middle-Aged	29
15	2. Middle-Aged	29
16	2. Middle-Aged	31
17	2. Middle-Aged	28
18	2. Middle-Aged	31
19	2. Middle-Aged	29
20	2. Middle-Aged	32
21	2. Middle-Aged	30
22	2. Middle-Aged	29

One-Way ANOVA: Insurance Response Rate

The GLM Procedure

Dependent Variable: RATE

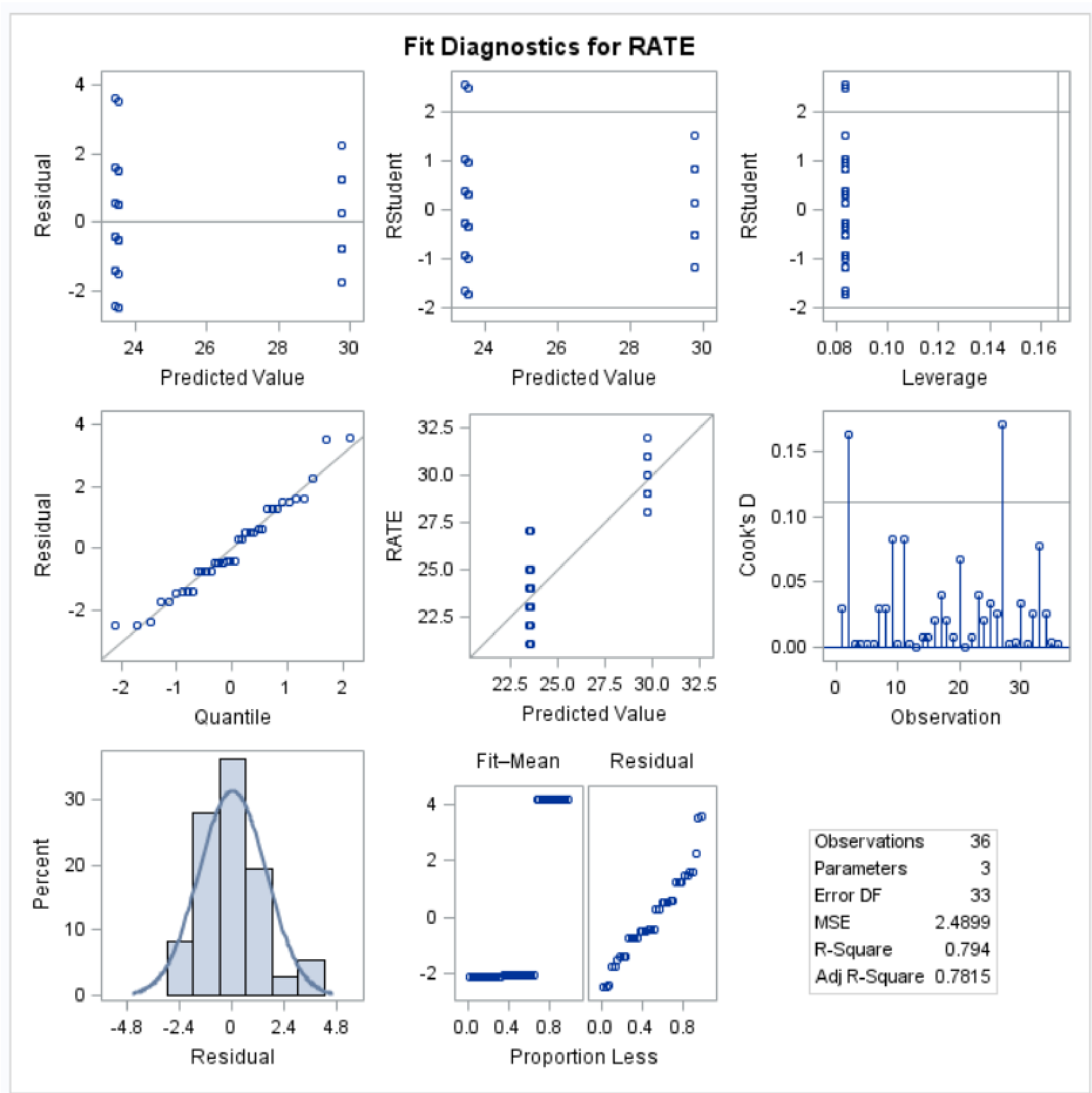
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	316.7222222	158.3611111	63.60	<.0001
Error	33	82.1666667	2.4898990		
Corrected Total	35	398.8888889			

R-Square	Coeff Var	Root MSE	RATE Mean
0.794011	6.174553	1.577941	25.55556

Source	DF	Type I SS	Mean Square	F Value	Pr > F
AGE	2	316.7222222	158.3611111	63.60	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
AGE	2	316.7222222	158.3611111	63.60	<.0001

There was a significant effect of AGE on the response rate, $F(2, 33) = 63.60$, $p < .001$.


Fit Diagnostics for RATE
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
RATE
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 36
Parameters 3
Error DF 33
MSE 2.4899
R-Square 0.794
Adj R-Square 0.7815

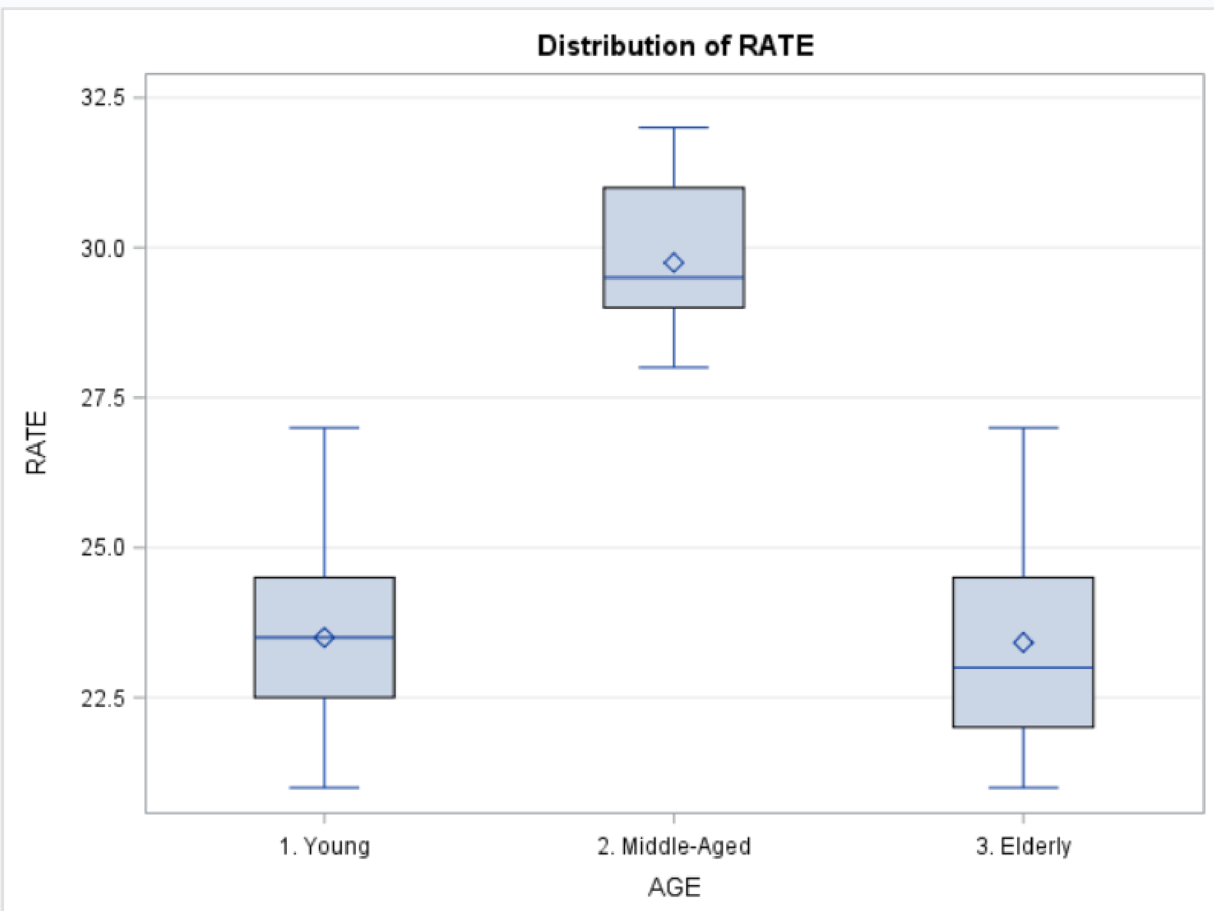
The assumption of homogeneity of variances was found to be tenable, $F(2, 33) = 0.30$, $p = .741$.

The GLM Procedure

Levene's Test for Homogeneity of RATE Variance ANOVA of Absolute Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
AGE	2	0.4738	0.2369	0.30	0.7411
Error	33	25.8588	0.7836		

One-Way ANOVA: Insurance Response Rate

The GLM Procedure



Level of AGE	N	RATE	
		Mean	Std Dev
1. Young	12	23.5000000	1.73205081
2. Middle-Aged	12	29.7500000	1.28805703
3. Elderly	12	23.4166667	1.67648622

The "LSMEAN Number" column gives the codes that are used in the Tukey pairwise comparison table below. In this case, "Young" is LSMEAN #1, "Middle-Aged" is LSMEAN #2, and "Elderly" is LSMEAN #3.

One-Way ANOVA: Insurance Response Rate

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

AGE	RATE LSMEAN	LSMEAN Number
1. Young	23.5000000	1
2. Middle-Aged	29.7500000	2
3. Elderly	23.4166667	3

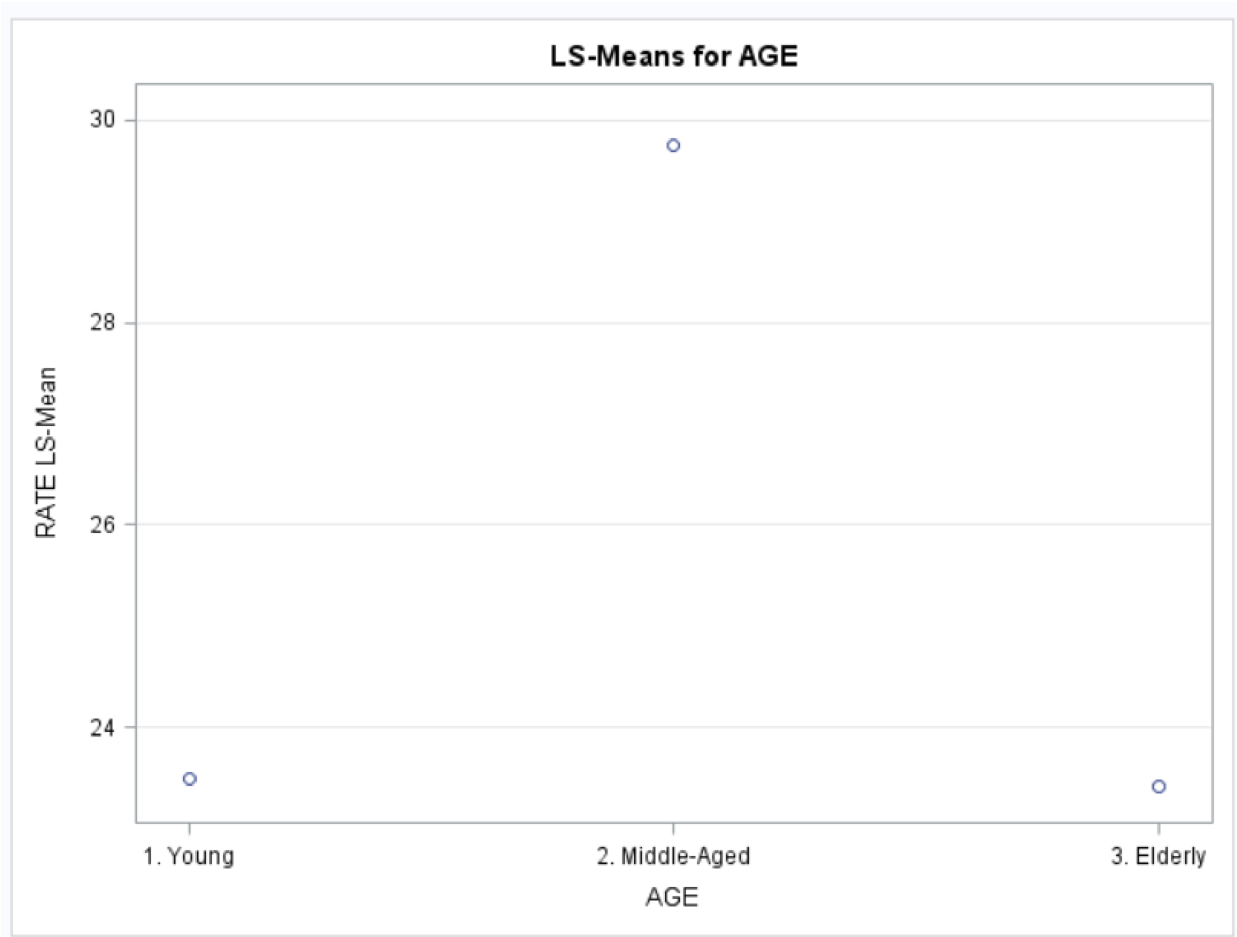
Least Squares Means for effect AGE
Pr > \|t\| for H0: LSMean(i)=LSMean(j)
Dependent Variable: RATE

i/j	1	2	3
1		<.0001	0.9908
2	<.0001		<.0001
3	0.9908	<.0001	

"Young"

"Young"

Tukey's HSD post-hoc test revealed a significant difference between response rates of young and middle-aged participants ($p < .001$) and between response rates of middle-aged and elderly participants ($p < .001$). The difference between the response rates of young and elderly participants was not significant ($p = .991$).


LS-Means for AGE
30
28
26
24
RATE LS-Mean
1. Young
2. Middle-Aged
3. Elderly
AGE

Inferential Statistics
Two-Way ANOVA with Nonsignificant Interaction
Research Scenario

A marketing manager for a supermarket chain was interested in the effect of both retail price and display location on a new promotional line of cookies. A group of 24 stores with matching store volume, layout, and customer demographics was split at random into 6 groups of 4 stores each. One group was assigned to each of the 6 combinations of retail price (regular retail vs. discounted retail) and display location (entrance aisle, cookie aisle, and checkout). Below are the average weekly unit sales by store over a 13-week period.

Price	**Display Location**		
	1. Entrance	**2. Cookies**	**3. Checkout**
1. Regular Retail	38	28	21
	31	25	32
	27	23	30
	33	20	22
2. Discounted Retail	35	22	19
	21	24	15
	39	16	25
	30	17	20

Inferential Statistics
Two-Way ANOVA with Nonsignificant Interaction
SAS Code

```
      PROC FORMAT;
            VALUE PRICE_FMT
                  1="1. Regular"
                  2="2. Discounted";
            VALUE LOC_FMT
                  1="1. Entrance"
                  2="2. Cookie Aisle"
                  3="3. Checkout";
      RUN;

      DATA NEW_COOKIES;
(1)         INPUT PRICE LOCATION SALES @@;
(2)         FORMAT PRICE PRICE_FMT. LOCATION LOC_FMT.;
            LINES;
                  1 1 38 1 1 31 1 1 27 1 1 33
                  1 2 28 1 2 25 1 2 23 1 2 20
                  1 3 21 1 3 32 1 3 30 1 3 22
                  2 1 35 2 1 21 2 1 39 2 1 30
                  2 2 22 2 2 24 2 2 16 2 2 17
                  2 3 19 2 3 15 2 3 25 2 3 20
      RUN;

      PROC PRINT DATA=NEW_COOKIES;
      RUN;

(3)   PROC SGPLOT DATA=NEW_COOKIES;
(4)         VLINE PRICE /
                  GROUP=LOCATION
                  STAT=MEAN
                  RESPONSE=SALES
                  MARKERS;
      RUN;

(5)   PROC GLM DATA=NEW_COOKIES ORDER=INTERNAL;
(6)         CLASS PRICE LOCATION;
(7)         MODEL SALES=PRICE*LOCATION;
(8)         MEANS PRICE*LOCATION / HOVTEST=LEVENE(TYPE=ABS);
            TITLE "Levene's Test of Homogeneity of Variances";
      RUN;

(9)   PROC GLM DATA=NEW_COOKIES ORDER=INTERNAL;
(10)        CLASS PRICE LOCATION;
(11)        MODEL SALES=PRICE LOCATION PRICE*LOCATION;
(12)        MEANS PRICE LOCATION PRICE*LOCATION;
(13)        LSMEANS LOCATION / PDIFF=ALL ADJUST=TUKEY;
            TITLE "Two-Way ANOVA: Cookie Sales";
      RUN;

      TITLE;
      QUIT;
```

(1) *Please take note of the order used for data entry.* PRICE is the variable initiated first, then LOCATION, then SALES. Correspondingly, the raw data are entered with PRICE first, LOCATION second, and SALES third. Therefore, the observations are entered in the following order.

PRICE = 1 and LOCATION = 1
PRICE = 1 and LOCATION = 2
PRICE = 1 and LOCATION = 3
PRICE = 2 and LOCATION = 1
PRICE = 2 and LOCATION = 2
PRICE = 2 and LOCATION = 3

(2) The FORMAT statement is used to apply the labels created in PRICE_FMT and LOC_FMT to PRICE and LOCATION, respectively. *Note that a period (.) follows PRICE_FMT* ***and*** *LOC_FMT; the code will not work properly if you omit the periods.*

(3) PROC SGPLOT creates a line graph comparing the means by PRICE and LOCATION.

(4) The VLINE statement specifies PRICE as the variable to be placed on the *x*-axis. A forward slash (/) is used prior to entering the VLINE options. The GROUP option specifies LOCATION as the variable to use for the lines; each LOCATION will be shown as a separate color-coded line. The STAT option requests the mean for each group to be the statistic that is graphed. The RESPONSE option identifies the *y*-axis variable, SALES in this case. MARKERS tells SAS to show the means for each group as data points.

(5) **This PROC GLM step is used for Levene's Test of the homogeneity of variances only. Do NOT use or report any of the other results from this step.** A quirk of SAS is that Levene's Test is only available for one-way ANOVAs. Therefore, this PROC GLM step is written such that the interaction term (PRICE*LOCATION) is the only independent variable (IV) in the analysis, making it a one-way ANOVA while retaining all six cells whose variances need to be compared. **Once again, do NOT use or report any results – other than Levene's Test – from this step. Although some of the ANOVA results will match the final (correct) results, some of them (including the Sum of Squares partitioning) will NOT.** ORDER=INTERNAL tells SAS to analyze the data in the order that it was entered in the DATA step.

(6) The CLASS statement identifies PRICE and LOCATION as the grouping variables.

(7) The MODEL statement is written as Dependent Variable (DV) = the interaction term (PRICE*LOCATION) for the two Independent Variables (IVs). Remember, this is NOT the correct model for a factorial ANOVA. This MODEL statement is required in order to run Levene's Test.

(8) The MEANS statement requires SAS to compute the means for each cell of PRICE by LOCATION. The option HOVTEST=LEVENE calls for Levene's Test of Homogeneity of Variances, to test the assumption that the cells have equal variances.

(9) **This PROC GLM step is the correct step for a Factorial ANOVA. These are the true results to use and report for this analysis.** ORDER=INTERNAL tells SAS to analyze the data in the order that it was entered in the DATA step.

(10) The CLASS statement identifies PRICE and LOCATION as the grouping variables.

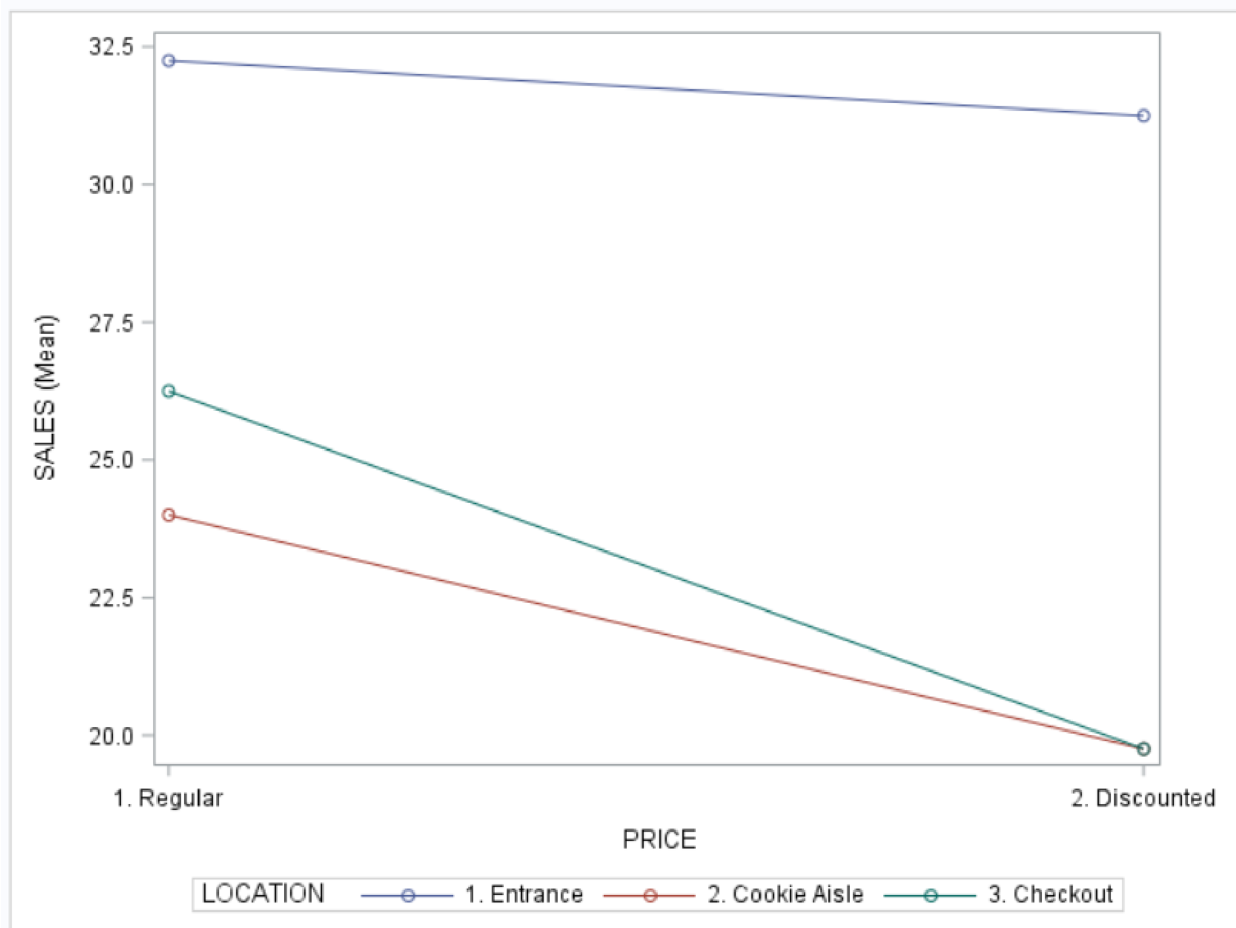
(11) This MODEL statement – which is the correct model for a factorial ANOVA – is written as Dependent Variable (DV) = Independent Variable 1 (IV1) and Independent Variable 2 (IV2) and the interaction term for the two IVs (IV1*IV2).

(12) The MEANS statement requires SAS to compute the means for (A) each group in PRICE, (B) each group in LOCATION, and (C) each cell of PRICE by LOCATION.

(13) *The LSMEANS statement shown here is only required if (A) there is a significant interaction effect* ***<u>or</u>*** *(B) there is a significant main effect.* Remember, if there is a significant interaction effect, then this effect takes priority and you need to conduct post-hoc analyses to find out which interaction (IV1*IV2) means differ significantly. If the interaction is nonsignificant, you need to examine the main effects for significance; if a main effect (or multiple main effects) are significant, conduct post-hoc analyses to find out which main effect (IV1 and/or IV2) means differ significantly. The LSMEANS statement will perform Tukey's HSD post-hoc testing for any effect you specify: interaction or main effects. *In this example, the LSMEANS statement performs Tukey's HSD by LOCATION.* The PDIFF=ALL option requests the *p* values for each of the pairwise comparisons.

Inferential Statistics
Two-Way ANOVA with Nonsignificant Interaction
Selected Output

Obs	PRICE	LOCATION	SALES
1	1. Regular	1. Entrance	38
2	1. Regular	1. Entrance	31
3	1. Regular	1. Entrance	27
4	1. Regular	1. Entrance	33
5	1. Regular	2. Cookie Aisle	28
6	1. Regular	2. Cookie Aisle	25
7	1. Regular	2. Cookie Aisle	23
8	1. Regular	2. Cookie Aisle	20
9	1. Regular	3. Checkout	21
10	1. Regular	3. Checkout	32
11	1. Regular	3. Checkout	30
12	1. Regular	3. Checkout	22
13	2. Discounted	1. Entrance	35
14	2. Discounted	1. Entrance	21
15	2. Discounted	1. Entrance	39
16	2. Discounted	1. Entrance	30
17	2. Discounted	2. Cookie Aisle	22
18	2. Discounted	2. Cookie Aisle	24
19	2. Discounted	2. Cookie Aisle	16
20	2. Discounted	2. Cookie Aisle	17
21	2. Discounted	3. Checkout	19
22	2. Discounted	3. Checkout	15
23	2. Discounted	3. Checkout	25



The assumption of homogeneity of variances was found to be tenable, $F(5, 18) = 1.12$, $p = .385$.

Levene's Test of Homogeneity of Variances

The GLM Procedure

Levene's Test for Homogeneity of SALES Variance ANOVA of Absolute Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
PRICE*LOCATION	5	32.2083	6.4417	1.12	0.3848
Error	18	103.5	5.7500		

Notes:

Use Type III Sums of Squares (Type III SS).

The *DF* associated with the IVs (PRICE, LOCATION, and PRICE*LOCATION, in the bottom table) add up to the "Model" *DF* in the ANOVA table at the top.

Two-Way ANOVA: Cookie Sales

The GLM Procedure

Dependent Variable: SALES

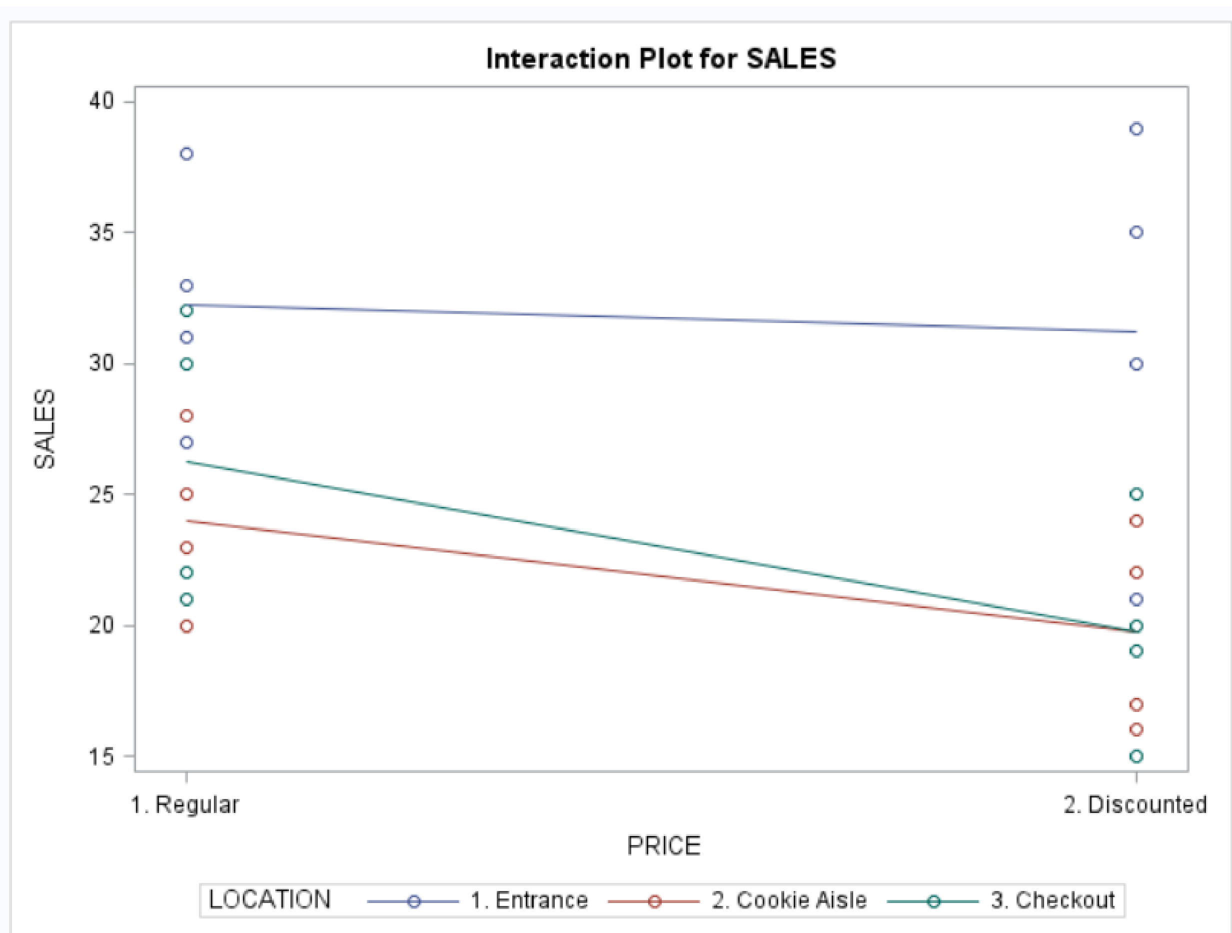
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	590.208333	118.041667	4.56	0.0073
Error	18	465.750000	25.875000		
Corrected Total	23	1055.958333			

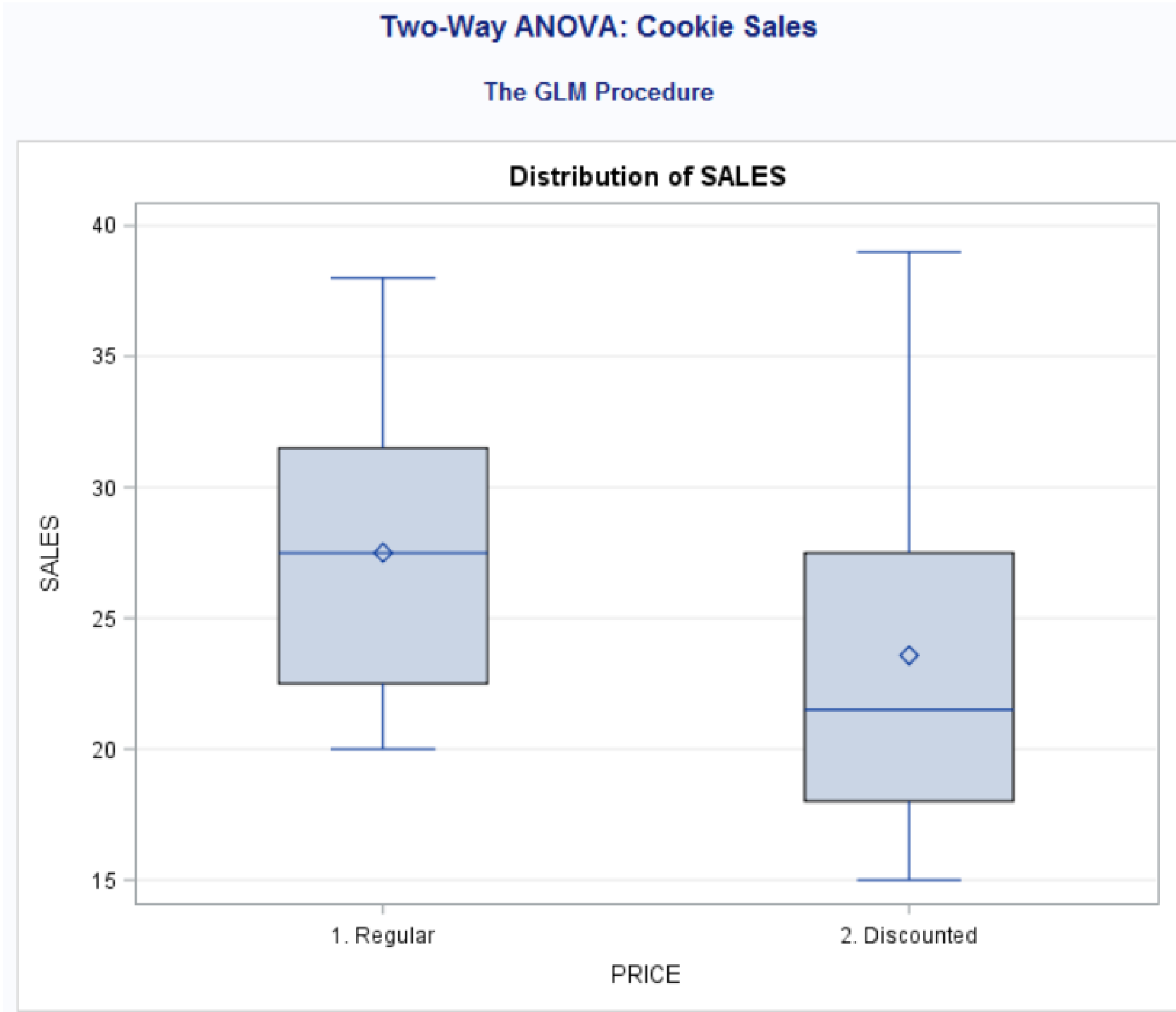
R-Square	Coeff Var	Root MSE	SALES Mean
0.558931	19.91549	5.086747	25.54167

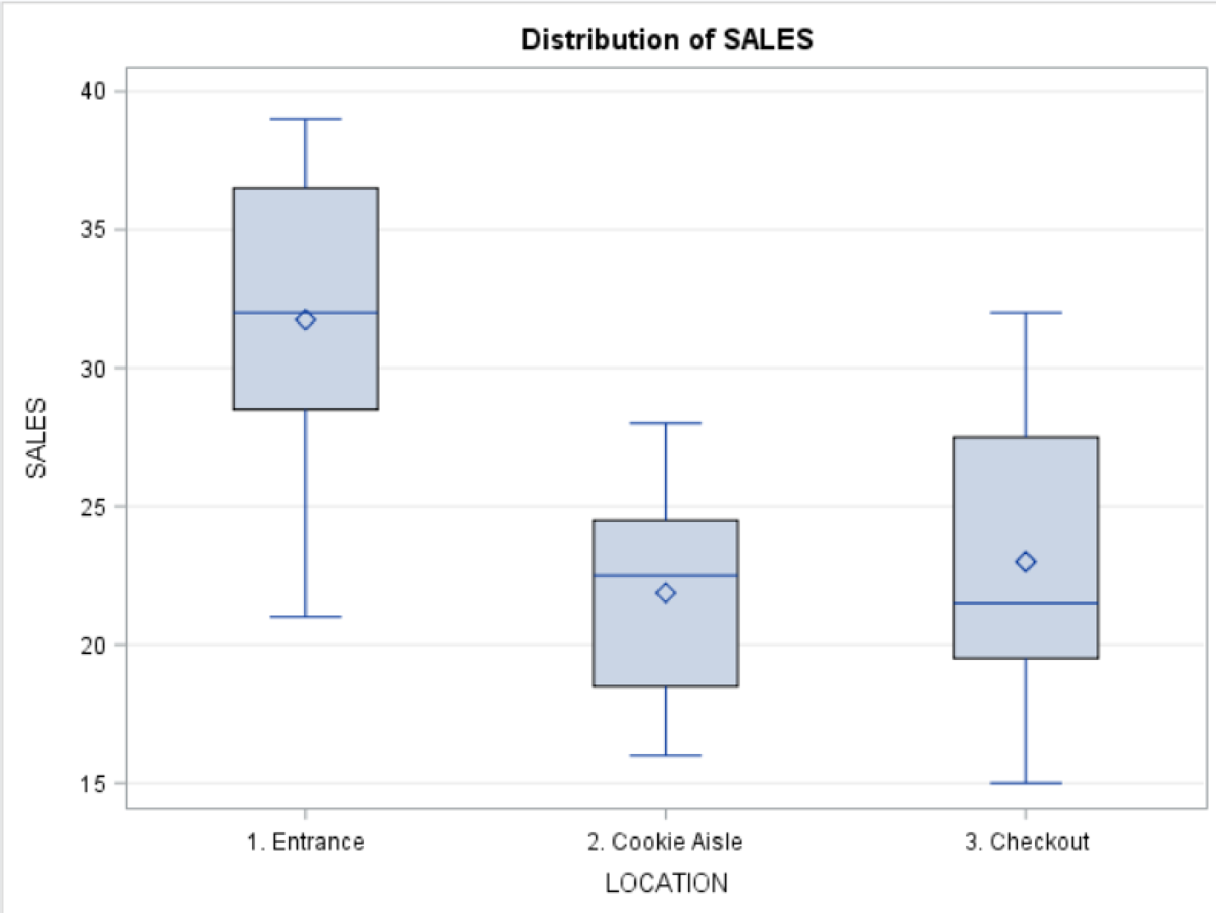
Source	DF	Type I SS	Mean Square	F Value	Pr > F
PRICE	1	92.0416667	92.0416667	3.56	0.0755
LOCATION	2	467.5833333	233.7916667	9.04	0.0019
PRICE*LOCATION	2	30.5833333	15.2916667	0.59	0.5642

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PRICE	1	92.0416667	92.0416667	3.56	0.0755
LOCATION	2	467.5833333	233.7916667	9.04	0.0019
PRICE*LOCATION	2	30.5833333	15.2916667	0.59	0.5642

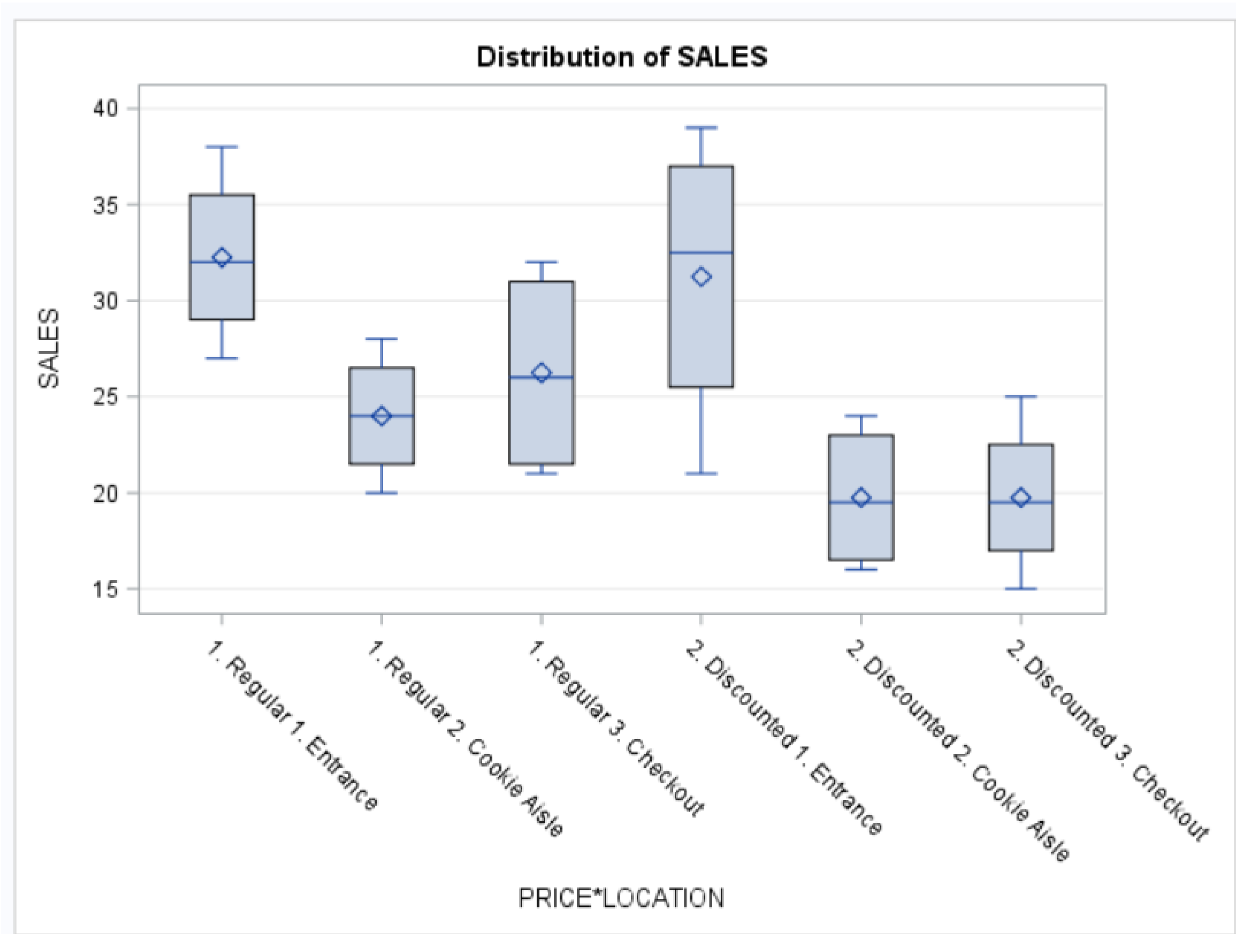
The interaction between PRICE and LOCATION was not found to be significant, $F(2, 18) = 0.59$, $p = .564$. The main effect of PRICE on cookie sales was not significant either, $F(1, 18) = 3.56$, $p = .076$. There was a significant main effect of LOCATION, $F(2, 18) = 9.04$, $p = .002$.




Two-Way ANOVA: Cookie Sales
The GLM Procedure
Distribution of SALES
40
35
30
25
20
15
SALES
1. Regular
2. Discounted
PRICE



Level of LOCATION	N	SALES Mean	Std Dev
1. Entrance	8	31.7500000	5.92211352
2. Cookie Aisle	8	21.8750000	4.05101398
3. Checkout	8	23.0000000	5.70713839



Level of PRICE	Level of LOCATION	N	SALES Mean	SALES Std Dev
1. Regular	1. Entrance	4	32.2500000	4.57347424
1. Regular	2. Cookie Aisle	4	24.0000000	3.36650165
1. Regular	3. Checkout	4	26.2500000	5.56027577
2. Discounted	1. Entrance	4	31.2500000	7.76208735
2. Discounted	2. Cookie Aisle	4	19.7500000	3.86221008
2. Discounted	3. Checkout	4	19.7500000	4.11298756

The "LSMEAN Number" column gives the codes that are used in the Tukey pairwise comparison table below. In this case, the "Entrance" location is LSMEAN #1, the "Cookie Aisle" is LSMEAN #2, and the "Checkout" is LSMEAN #3.

Two-Way ANOVA: Cookie Sales

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

LOCATION	SALES LSMEAN	LSMEAN Number
1. Entrance	31.7500000	1
2. Cookie Aisle	21.8750000	2
3. Checkout	23.0000000	3

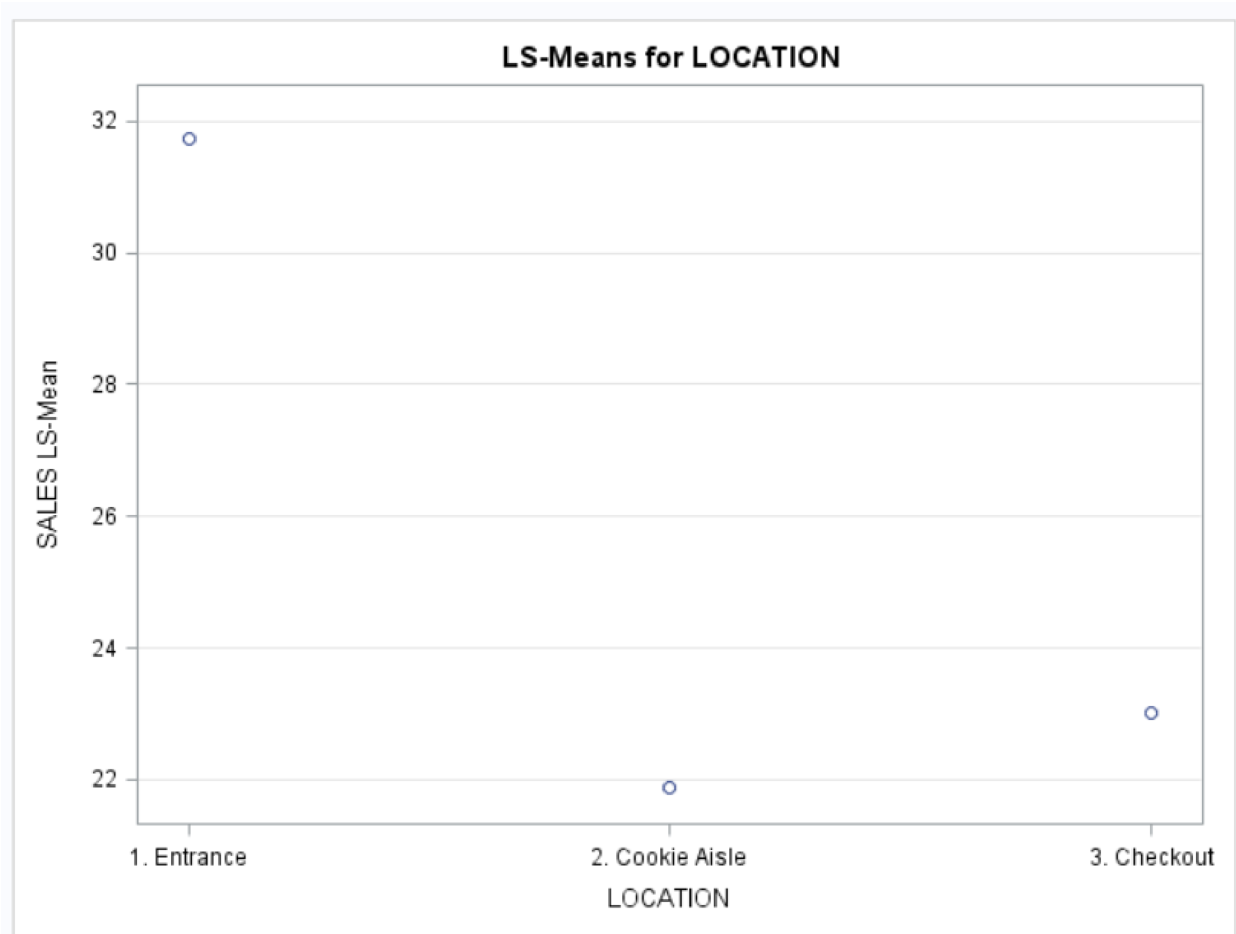
"Entrance" Location

"Entrance" Location

Least Squares Means for effect LOCATION
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: SALES

i/j	1	2	3
1		0.0030	0.0078
2	0.0030		0.8984
3	0.0078	0.8984	

Tukey's HSD post-hoc test revealed a significant difference between the entrance and the cookie aisle (p = .003) and between the entrance and the checkout (p = .008). The difference between the cookie aisle and the checkout was not significant (p = .898).


LS-Means for LOCATION
32
30
28
26
24
22
SALES LS-Mean
1. Entrance
2. Cookie Aisle
3. Checkout
LOCATION

Inferential Statistics
Two-Way ANOVA with Significant Interaction
Research Scenario

Eighteen students were randomly assigned to one of three different classroom teaching methods (online, hybrid, or in person) to learn a new math concept. At the end of 12 weeks, students were given a test to assess their understanding of the concept. A two-way ANOVA will be run to assess the difference between the teaching methods as well as between student's previous online experiences. The table below shows the scores by teaching method and previous experience. Two students did not finish the 12-week course so the distribution between the teaching methods and the previous experience was not equal.

	1. In person	**2. Hybrid**	**3. Online**
1. Experienced (Has taken an online course in the past)	57.5 80.0 62.5	72.5 75.0	77.5 90.0 82.5
2. Not Experienced (Has not taken an online class in the past)	80.0 65.0 65.0	85.0 100.0 87.5	57.5 65.0

Inferential Statistics
Two-Way ANOVA with Significant Interaction
SAS Code

```
      PROC FORMAT;
            VALUE EXPER_FMT
                  1="1. Experienced"
                  2="2. Not Experienced";
            VALUE METHOD_FMT
                  1="1. In Person"
                  2="2. Hybrid"
                  3="3. Online";
      RUN;

      DATA TEACHING_METHODS;
(1)         INPUT EXPERIENCE METHOD SCORE @@;
(2)         FORMAT EXPERIENCE EXPER_FMT. METHOD METHOD_FMT.;
            LINES;
                  1 1 57.5 1 1 80.0 1 1 62.5
                  1 2 72.5 1 2 75.0
                  1 3 77.5 1 3 90.0 1 3 82.5
                  2 1 80.0 2 1 65.0 2 1 65.0
                  2 2 85.0 2 2 100.0 2 2 87.5
                  2 3 57.5 2 3 65.0
      RUN;

      PROC PRINT DATA=TEACHING_METHODS;
      RUN;

(3)   PROC SGPLOT DATA=TEACHING_METHODS;
(4)         VLINE METHOD /
                  GROUP=EXPERIENCE
                  STAT=MEAN
                  RESPONSE=SCORE
                  MARKERS;
      RUN;

(5)   PROC GLM DATA=TEACHING_METHODS ORDER=INTERNAL;
(6)         CLASS EXPERIENCE METHOD;
(7)         MODEL SCORE=EXPERIENCE*METHOD;
(8)         MEANS EXPERIENCE*METHOD / HOVTEST=LEVENE(TYPE=ABS);
            TITLE "Levene's Test of Homogeneity of Variances";
      RUN;

(9)   PROC GLM DATA=TEACHING_METHODS ORDER=INTERNAL;
(10)        CLASS EXPERIENCE METHOD;
(11)        MODEL SCORE=EXPERIENCE METHOD EXPERIENCE*METHOD;
(12)        MEANS EXPERIENCE METHOD EXPERIENCE*METHOD;
(13)        LSMEANS EXPERIENCE*METHOD / PDIFF=ALL ADJUST=TUKEY;
            TITLE "Two-Way ANOVA: Math Teaching Methods";
      RUN;

      TITLE;
      QUIT;
```

(1) *Please take note of the order used for data entry.* EXPERIENCE is the variable initiated first, then METHOD, then SCORE. Correspondingly, the raw data are entered with EXPERIENCE first, METHOD second, and SCORE third. Therefore, the observations are entered in the following order.

EXPERIENCE = 1 and METHOD = 1
EXPERIENCE = 1 and METHOD = 2
EXPERIENCE = 1 and METHOD = 3
EXPERIENCE = 2 and METHOD = 1
EXPERIENCE = 2 and METHOD = 2
EXPERIENCE = 2 and METHOD = 3

(2) The FORMAT statement is used to apply the labels created in EXPER_FMT and METHOD _FMT to EXPERIENCE and METHOD, respectively. *Note that a period (.) follows EXPER_FMT* ***and*** *METHOD_FMT; the code will not work properly if you omit the periods.*

(3) PROC SGPLOT creates a line graph comparing the means by EXPERIENCE and METHOD.

(4) The VLINE statement specifies EXPERIENCE as the variable to be placed on the *x*-axis. A forward slash (/) is used prior to entering the VLINE options. The GROUP option specifies METHOD as the variable to use for the lines; each METHOD will be shown as a separate color-coded line. The STAT option requests the mean for each group to be the statistic that is graphed. The RESPONSE option identifies the *y*-axis variable, SALES in this case. MARKERS tells SAS to show the means for each group as data points.

(5) **This PROC GLM step is used for Levene's Test of the homogeneity of variances only. Do NOT use or report any of the other results from this step.** A quirk of SAS is that Levene's Test is only available for one-way ANOVAs. Therefore, this PROC GLM step is written such that the interaction term (EXPERIENCE*METHOD) is the only independent variable (IV) in the analysis, making it a one-way ANOVA while retaining all six cells whose variances need to be compared. **Once again, do NOT use or report any results – other than Levene's Test – from this step. Although some of the ANOVA results will match the final (correct) results, some of them (including the Sum of Squares partitioning) will NOT.** ORDER=INTERNAL tells SAS to analyze the data in the order that it was entered in the DATA step.

(6) The CLASS statement identifies EXPERIENCE and METHOD as the grouping variables.

(7) The MODEL statement is written as Dependent Variable (DV) = the interaction term (EXPERIENCE*METHOD) for the two Independent Variables (IVs). Remember, this is NOT the correct model for a factorial ANOVA. This MODEL statement is required in order to run Levene's Test.

(8) The MEANS statement requires SAS to compute the means for each cell of EXPERIENCE by METHOD. The option HOVTEST=LEVENE calls for the Levene's Test of Homogeneity of Variances, to test the assumption that the cells have equal variances.

(9) **This PROC GLM step is the correct step for a Factorial ANOVA. These are the correct results to use and report for this analysis.** ORDER=INTERNAL tells SAS to analyze the data in the order that it was entered in the DATA step.

(10) The CLASS statement identifies EXPERIENCE and METHOD as the grouping variables.

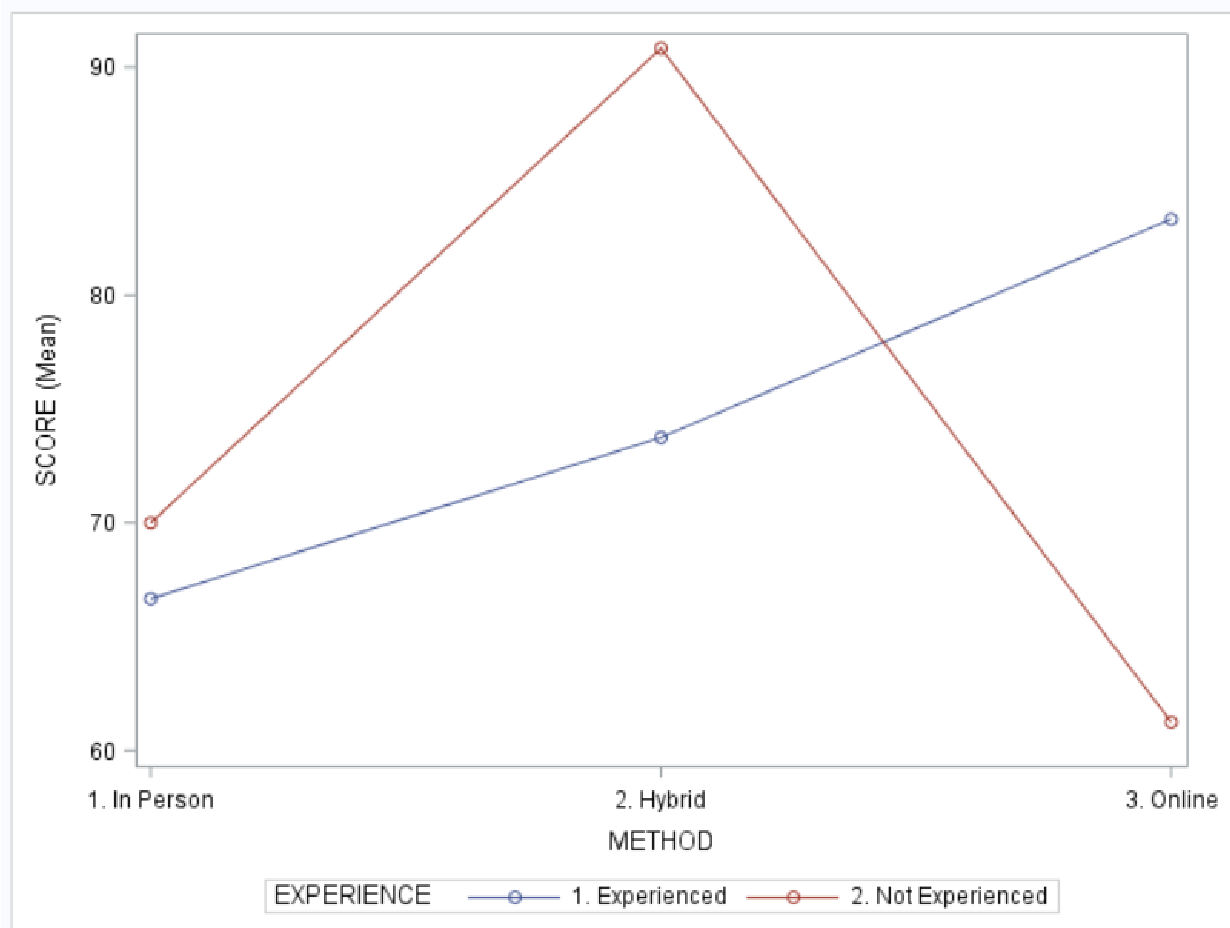
(11) This MODEL statement – which is the correct model for a factorial ANOVA – is written as Dependent Variable (DV) = Independent Variable 1 (IV1) and Independent Variable 2 (IV2) and the interaction term for the two IVs (IV1*IV2).

(12) The MEANS statement requires SAS to compute the means for (A) each group in EXPERIENCE, (B) each group in METHOD, and (C) each cell of EXPERIENCE by METHOD.

(13) *The LSMEANS statement shown here is only required if (A) there is a significant interaction effect* ***<u>or</u>*** *(B) there is a significant main effect.* Remember, if there is a significant interaction effect, then this effect takes priority and you need to conduct post-hoc analyses to find out which interaction (IV1*IV2) means differ significantly. If the interaction is nonsignificant, you need to examine the main effects for significance; if a main effect (or multiple main effects) are significant, conduct post-hoc analyses to find out which main effect (IV1 and/or IV2) means differ significantly. The LSMEANS statement will perform Tukey's HSD post-hoc testing for any effect you specify: interaction or main effects. *In this example, the LSMEANS statement performs Tukey's HSD by METHOD.* The PDIFF=ALL option requests the *p* values for each of the pairwise comparisons.

Inferential Statistics
Two-Way ANOVA with Significant Interaction
Selected Output

Obs	EXPERIENCE	METHOD	SCORE
1	1. Experienced	1. In Person	57.5
2	1. Experienced	1. In Person	80.0
3	1. Experienced	1. In Person	62.5
4	1. Experienced	2. Hybrid	72.5
5	1. Experienced	2. Hybrid	75.0
6	1. Experienced	3. Online	77.5
7	1. Experienced	3. Online	90.0
8	1. Experienced	3. Online	82.5
9	2. Not Experienced	1. In Person	80.0
10	2. Not Experienced	1. In Person	65.0
11	2. Not Experienced	1. In Person	65.0
12	2. Not Experienced	2. Hybrid	85.0
13	2. Not Experienced	2. Hybrid	100.0
14	2. Not Experienced	2. Hybrid	87.5
15	2. Not Experienced	3. Online	57.5
16	2. Not Experienced	3. Online	65.0



The assumption of homogeneity of variances was found to be tenable, $F(3, 8) = 0.84$, $p = .507$.

Levene's Test of Homogeneity of Variances

The GLM Procedure

Levene's Test for Homogeneity of SCORE Variance ANOVA of Absolute Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
EXPERIENCE*METHOD	3	30.3241	10.1080	0.84	0.5074
Error	8	95.8333	11.9792		

Level of EXPERIENCE	Level of METHOD	N	SCORE Mean	SCORE Std Dev
1. Experienced	1. In Person	3	66.6666667	11.8145391
1. Experienced	2. Hybrid	2	73.7500000	1.7677670
1. Experienced	3. Online	3	83.3333333	6.2915287
2. Not Experienced	1. In Person	3	70.0000000	8.6602540
2. Not Experienced	2. Hybrid	3	90.8333333	8.0363756
2. Not Experienced	3. Online	2	61.2500000	5.3033009

Notes:

Use Type III Sums of Squares (Type III SS).

The *DF* associated with the IVs (EXPERIENCE, METHOD, and EXPERIENCE*METHOD, in the bottom table) add up to the "Model" *DF* in the ANOVA table at the top.

Two-Way ANOVA: Math Teaching Methods

The GLM Procedure

Dependent Variable: SCORE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1624.609375	324.921875	4.86	0.0163
Error	10	668.750000	66.875000		
Corrected Total	15	2293.359375			

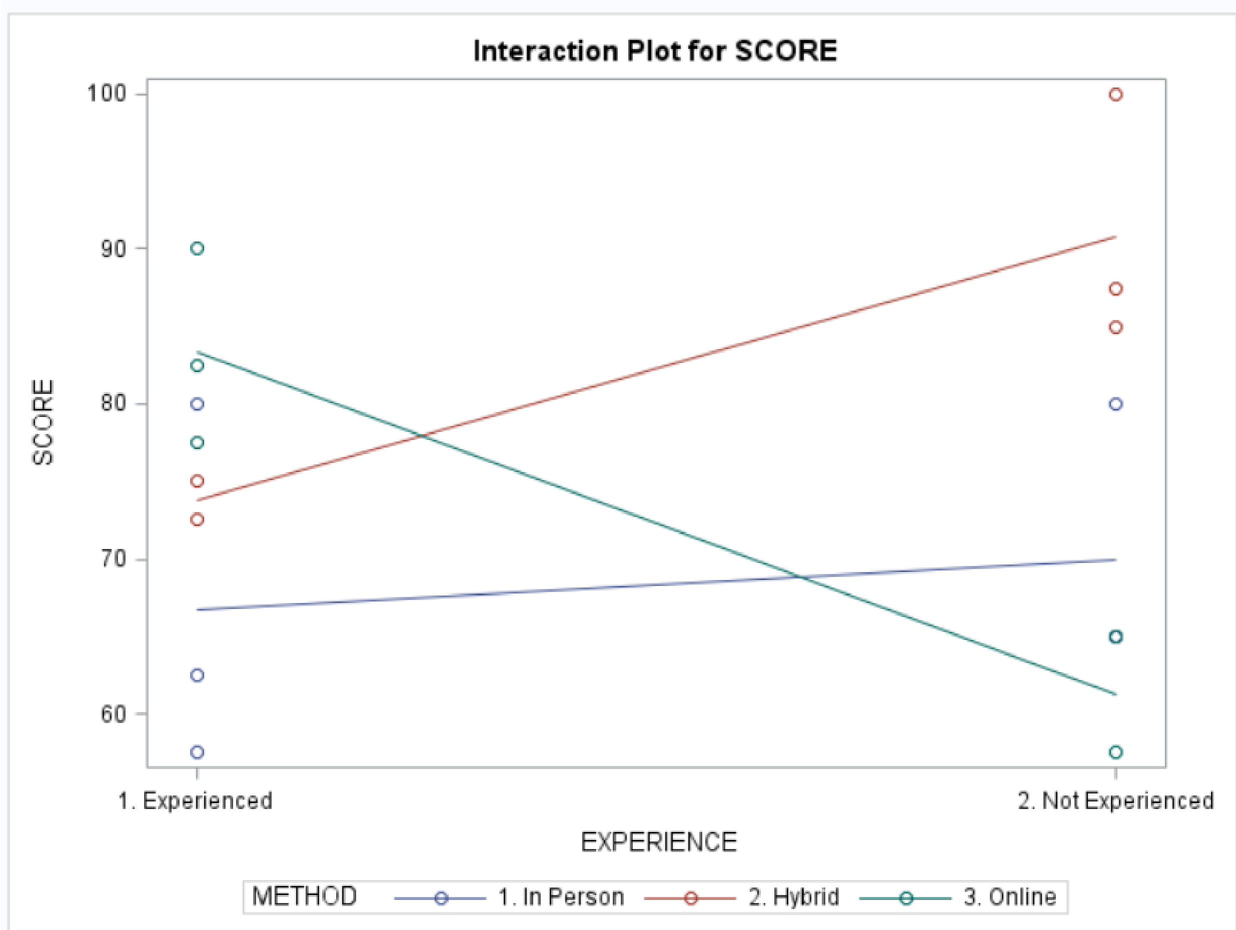
R-Square	Coeff Var	Root MSE	SCORE Mean
0.708397	10.88095	8.177714	75.15625

Source	DF	Type I SS	Mean Square	F Value	Pr > F
EXPERIENCE	1	3.5156250	3.5156250	0.05	0.8233
METHOD	2	669.2668269	334.6334135	5.00	0.0312
EXPERIENCE*METHOD	2	951.8269231	475.9134615	7.12	0.0120

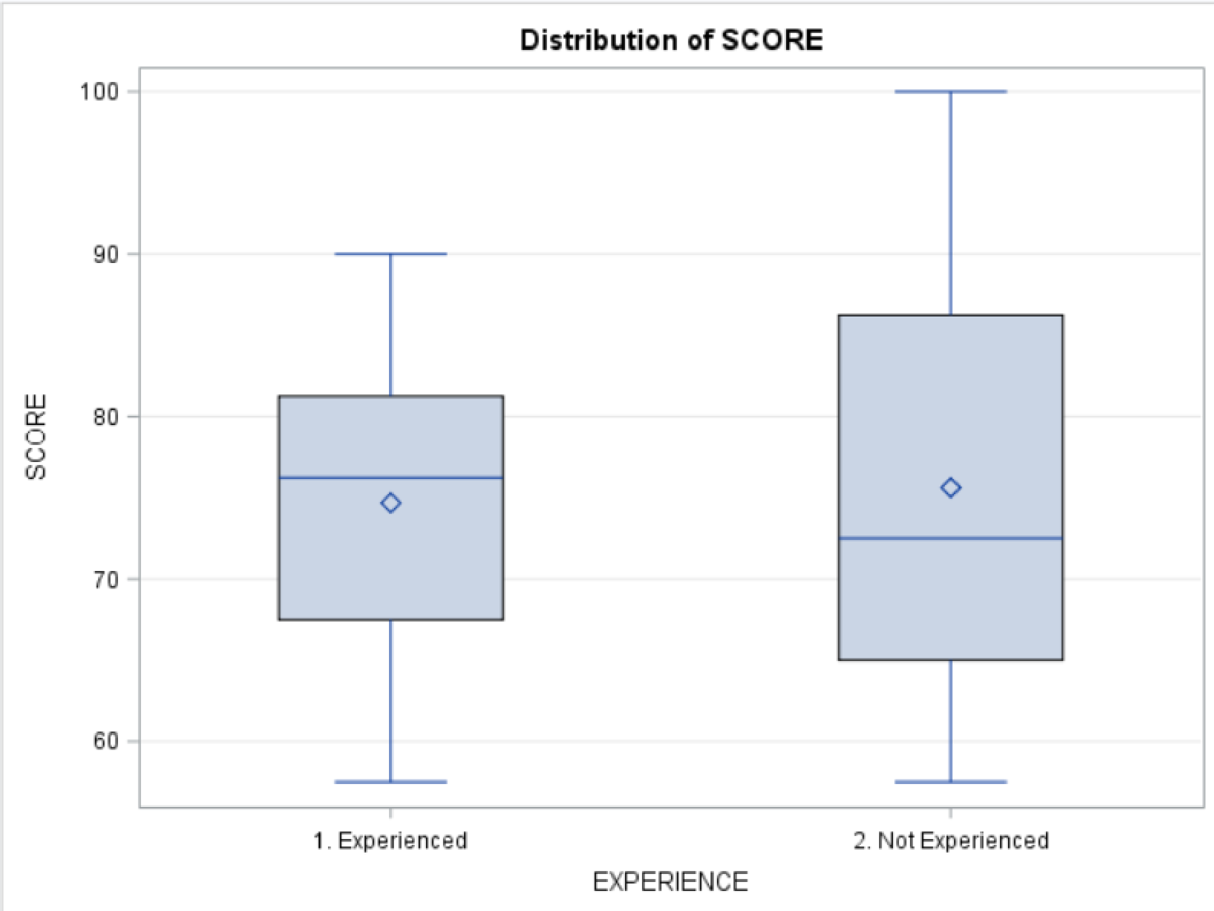
Source	DF	Type III SS	Mean Square	F Value	Pr > F
EXPERIENCE	1	1.1904762	1.1904762	0.02	0.8965
METHOD	2	536.3141026	268.1570513	4.01	0.0526
EXPERIENCE*METHOD	2	951.8269231	475.9134615	7.12	0.0120

The interaction between EXPERIENCE and METHOD was found to be significant, $F(2, 10) = 7.12$, $p = .012$.

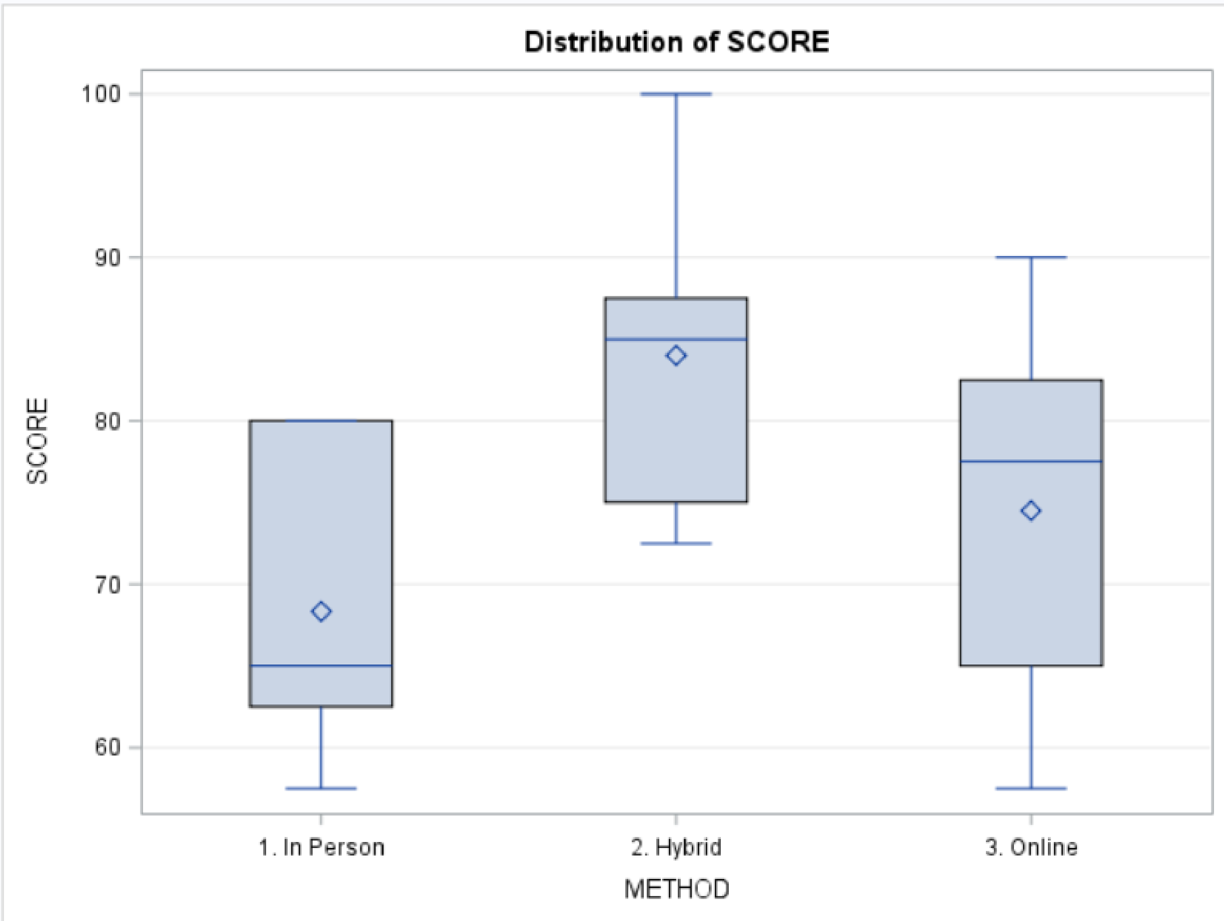
Remember: When there is a significant interaction, this takes precedence over main effects, regardless of whether or not they are significant.



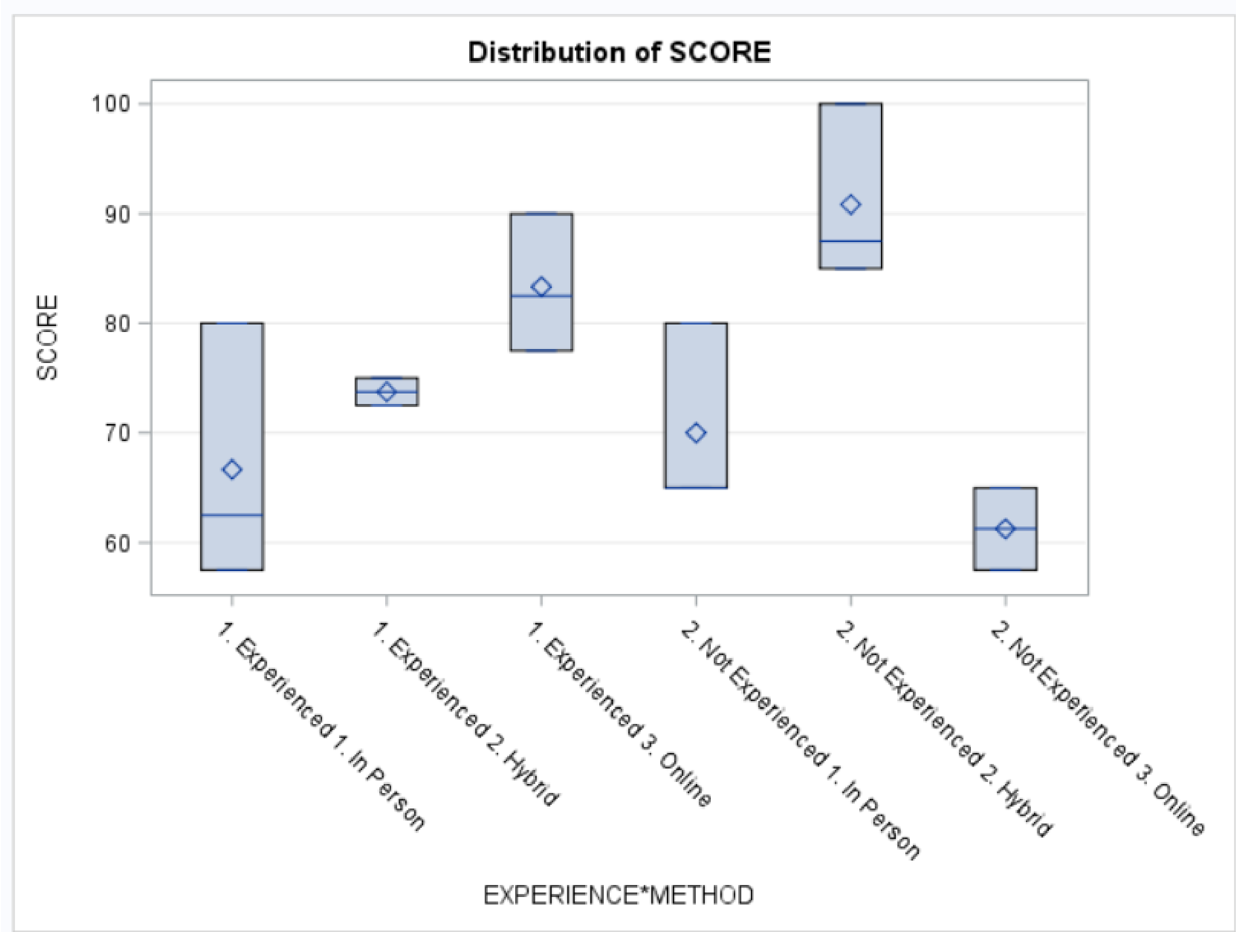
Interaction Plot for SCORE
SCORE
100
90
80
70
60
1. Experienced
2. Not Experienced
EXPERIENCE
METHOD
1. In Person
2. Hybrid
3. Online



Level of EXPERIENCE	N	SCORE	
		Mean	Std Dev
1. Experienced	8	74.6875000	10.5591446
2. Not Experienced	8	75.6250000	14.6841752



Level of METHOD	N	SCORE	
		Mean	Std Dev
1. In Person	6	68.3333333	9.4428103
2. Hybrid	5	84.0000000	10.9829413
3. Online	5	74.5000000	13.1576974



Level of EXPERIENCE	Level of METHOD	N	SCORE	
			Mean	Std Dev
1. Experienced	1. In Person	3	66.6666667	11.8145391
1. Experienced	2. Hybrid	2	73.7500000	1.7677670
1. Experienced	3. Online	3	83.3333333	6.2915287
2. Not Experienced	1. In Person	3	70.0000000	8.6602540
2. Not Experienced	2. Hybrid	3	90.8333333	8.0363756
2. Not Experienced	3. Online	2	61.2500000	5.3033009

The "LSMEAN Number" column gives the codes that are used in the Tukey pairwise comparison table below. In this case, the crossed condition "Experienced" and "In Person" is LSMEAN #1.

Two-Way ANOVA: Math Teaching Methods

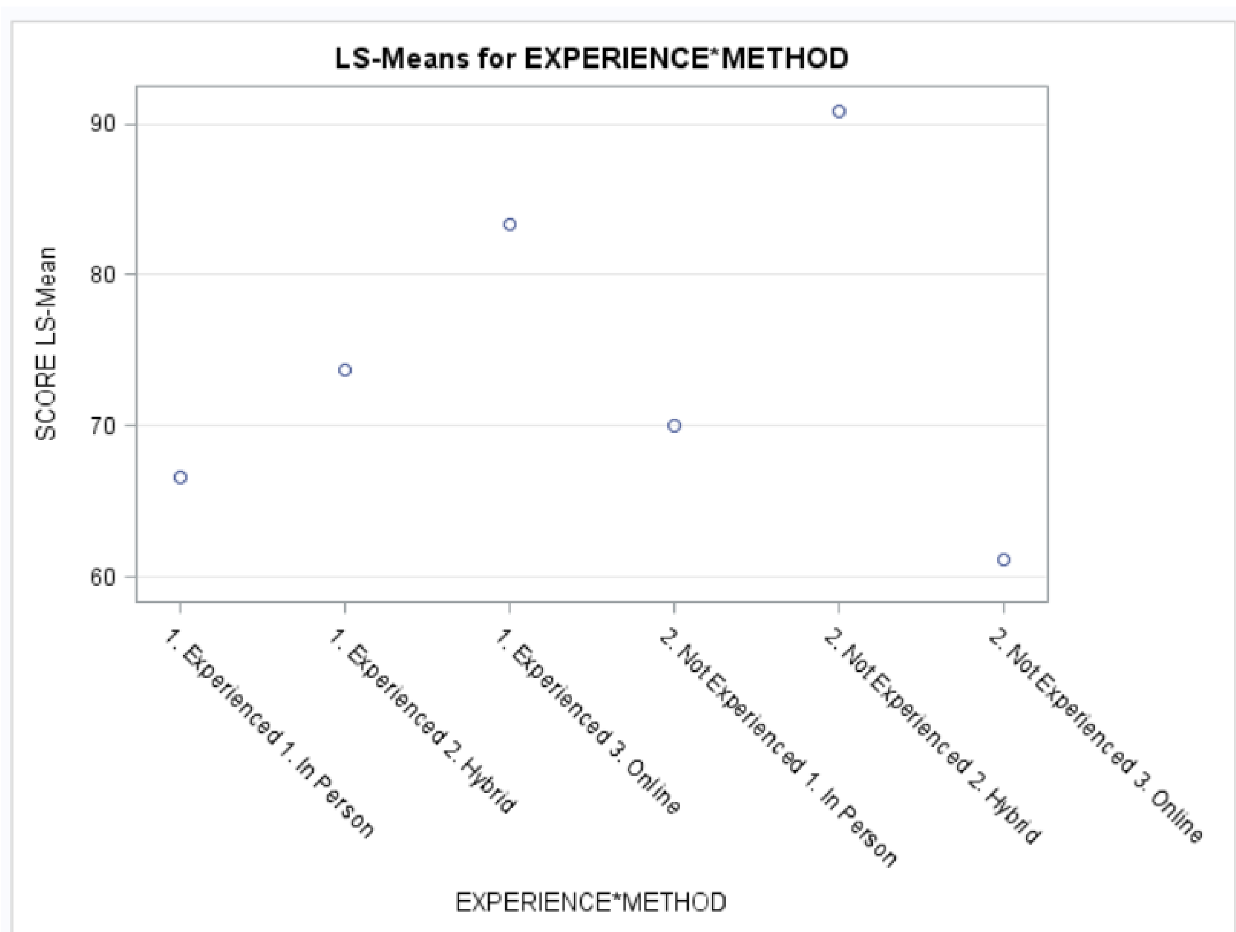
The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

EXPERIENCE	METHOD	SCORE LSMEAN	LSMEAN Number
1. Experienced	1. In Person	66.6666667	1
1. Experienced	2. Hybrid	73.7500000	2
1. Experienced	3. Online	83.3333333	3
2. Not Experienced	1. In Person	70.0000000	4
2. Not Experienced	2. Hybrid	90.8333333	5
2. Not Experienced	3. Online	61.2500000	6

Least Squares Means for effect EXPERIENCE*METHOD
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: SCORE

i/j	1	2	3	4	5	6
1		0.9241	0.2123	0.9951	0.0400	0.9740
2	0.9241		0.7874	0.9949	0.2814	0.6561
3	0.2123	0.7874		0.4056	0.8613	0.1088
4	0.9951	0.9949	0.4056		0.0854	0.8402
5	0.0400	0.2814	0.8613	0.0854		0.0238
6	0.9740	0.6561	0.1088	0.8402	0.0238	

Tukey's HSD test revealed a significant difference between students with online course experience assigned to the in-person method and students without online course experience assigned to the hybrid method (p = .040) and between students without online course experience assigned to the hybrid method and students without online course experience assigned to the online method (p = .024).



Inferential Statistics
Analysis of Covariance (ANCOVA)
Research Scenario

A personnel manager wished to evaluate the effect of positive and negative reinforcement on tardiness in a large manufacturing plant. A group of 30 chronic late arrivers were identified based on their previous tardiness records. They were then randomly assigned in equal numbers to one of three study groups: (1) positive reinforcement when on time, (2) negative reinforcement when late, and (3) no reinforcement. After a 10-week treatment period, data was collected on the number of tardies over an additional 10-week period. The data below presents the number of tardies for 10 weeks prior to the intervention and 10 weeks following termination of the treatment.

Type of Reinforcement					
1. Positive Reinforcement		**2. Negative Reinforcement**		**3. No Reinforcement**	
Pre	**Post**	**Pre**	**Post**	**Pre**	**Post**
4	2	4	3	5	5
5	3	5	5	5	4
6	1	5	3	5	4
7	4	6	6	6	8
7	3	7	6	8	8
8	5	8	7	8	9
9	3	8	6	9	7
9	5	9	8	10	10
10	6	10	6	10	8
11	5	11	4	7	10

Inferential Statistics
Analysis of Covariance (ANCOVA)
SAS Code

```
PROC FORMAT;
      VALUE REINF_FMT
            1="1. Positive"
            2="2. Negative"
            3="3. None";
RUN;

DATA TARDINESS_REINF;
      INPUT REINF PRE POST @@;
      FORMAT REINF REINF_FMT.;
      LINES;
            1     4     2     1     5     3
            1     6     1     1     7     4
            1     7     3     1     8     5
            1     9     3     1     9     5
            1     10    6     1     11    5
            2     4     3     2     5     5
            2     5     3     2     6     6
            2     7     6     2     8     7
            2     8     6     2     9     8
            2     10    6     2     11    4
            3     5     5     3     5     4
            3     5     4     3     6     8
            3     8     8     3     8     9
            3     9     7     3     10    10
            3     10    8     3     7     10
RUN;

PROC PRINT DATA=TARDINESS_REINF;
RUN;

(1)   PROC GLM DATA=TARDINESS_REINF PLOTS=DIAGNOSTICS;
            CLASS REINF;
            MODEL POST=REINF;
            MEANS REINF / HOVTEST=LEVENE(TYPE=ABS);
            TITLE "Levene's Test of Homogeneity of Variances";
      RUN;

(2)   PROC GLM DATA=TARDINESS_REINF;
            CLASS REINF;
            MODEL POST=REINF PRE REINF*PRE;
            TITLE "Test of Assumption of Homogeneity of Slopes";
      RUN;

(3)   PROC GLM DATA=TARDINESS_REINF;
            CLASS REINF;
            MODEL POST=REINF PRE;
            TITLE "ANCOVA: Employee Tardiness";
      RUN;
```

```
(4)    PROC GLM DATA=TARDINESS_REINF;
             CLASS REINF;
             MODEL POST=REINF PRE;
             LSMEANS REINF / PDIFF ADJUST=TUKEY;
             TITLE "Adjusted Means & Tukey Post-Hoc Comparisons";
       RUN;

(5)    PROC GLM DATA=TARDINESS_REINF;
             CLASS REINF;
             MODEL PRE=REINF;
             TITLE "Bryant-Paulson (BP) F Value";
       RUN;

       TITLE;
       QUIT;
```

(1) **This PROC GLM step is used for Levene's Test of the homogeneity of variances. Do NOT use or report any of the other significance test results from this step.** It will perform Levene's Test by conducting a one-way ANOVA of POST on REINF. It will also produce graphics to help determine whether the assumption of normality is tenable. Additionally, the *unadjusted* means and standard deviations will be produced by this procedure.
(2) **This PROC GLM step is used to test the homogeneity of slopes assumption. Do not use any other results from this procedure.** An ANCOVA test is produced by this PROC GLM by including an interaction term (REINF*PRE) in the model.
(3) **This PROC GLM step is the correct step for ANCOVA. These are the correct results to use and report for this analysis.** This PROC GLM is the same as that in (2), except that the interaction term is excluded. The output from this PROC GLM should be reported as the ANCOVA results.
(4) **This PROC GLM step is used to generate the *adjusted* pretest/pre-treatment means and Tukey's post-hoc comparisons.** The LSMEANS ("least squares" or "estimated marginal" means) statement requests the adjusted means, as well as the post-hoc comparisons with a Tukey alpha level adjustment for multiple significance tests.
(5) **This PROC GLM step is used to calculate the *F* value for the Bryant-Paulson (BP) Post-Hoc Procedure. Do not use any other results from this procedure.** If you want to conduct the BP Procedure, you can use PROC GLM to run a one-way ANOVA modeling PRE as the dependent variable and REINF as the independent variable. This will provide the *F* value that you will need for the manual BP calculations.

Inferential Statistics
Analysis of Covariance (ANCOVA)
Selected Output

The assumption of homogeneity of variances was found to be tenable, $F(2, 27) = 1.04$, $p = .368$.

Levene's Test of Homogeneity of Variances

The GLM Procedure

Levene's Test for Homogeneity of POST Variance ANOVA of Absolute Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
REINF	2	1.8747	0.9373	1.04	0.3684
Error	27	24.4200	0.9044		

Test of Assumption of Homogeneity of Slopes

The GLM Procedure

Dependent Variable: POST

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	106.2716858	21.2543372	9.96	<.0001
Error	24	51.1949808	2.1331242		
Corrected Total	29	157.4666667			

R-Square	Coeff Var	Root MSE	POST Mean
0.674884	26.71686	1.460522	5.466667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
REINF	2	64.86666667	32.43333333	15.20	<.0001
PRE	1	36.70054432	36.70054432	17.21	0.0004
PRE*REINF	2	4.70447484	2.35223742	1.10	0.3482

Source	DF	Type III SS	Mean Square	F Value	Pr > F
REINF	2	4.36894310	2.18447155	1.02	0.3743
PRE	1	39.30169928	39.30169928	18.42	0.0003
PRE*REINF	2	4.70447484	2.35223742	1.10	0.3482

The assumption of homogeneity of regression slopes was found tenable, as the PRE×REINF interaction was nonsignificant, ($F = 1.10$, $p = .348$).

The nonsignificant interaction term is removed from the model. The results of the new model are displayed; these are the ANCOVA results that should be reported.

ANCOVA: Employee Tardiness

The GLM Procedure

Dependent Variable: POST

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	101.5672110	33.8557370	15.75	<.0001
Error	26	55.8994557	2.1499791		
Corrected Total	29	157.4666667			

R-Square	Coeff Var	Root MSE	POST Mean
0.645008	26.82221	1.466281	5.466667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
REINF	2	64.86666667	32.43333333	15.09	<.0001
PRE	1	36.70054432	36.70054432	17.07	0.0003

Source	DF	Type III SS	Mean Square	F Value	Pr > F
REINF	2	70.45606548	35.22803274	16.39	<.0001
PRE	1	36.70054432	36.70054432	17.07	0.0003

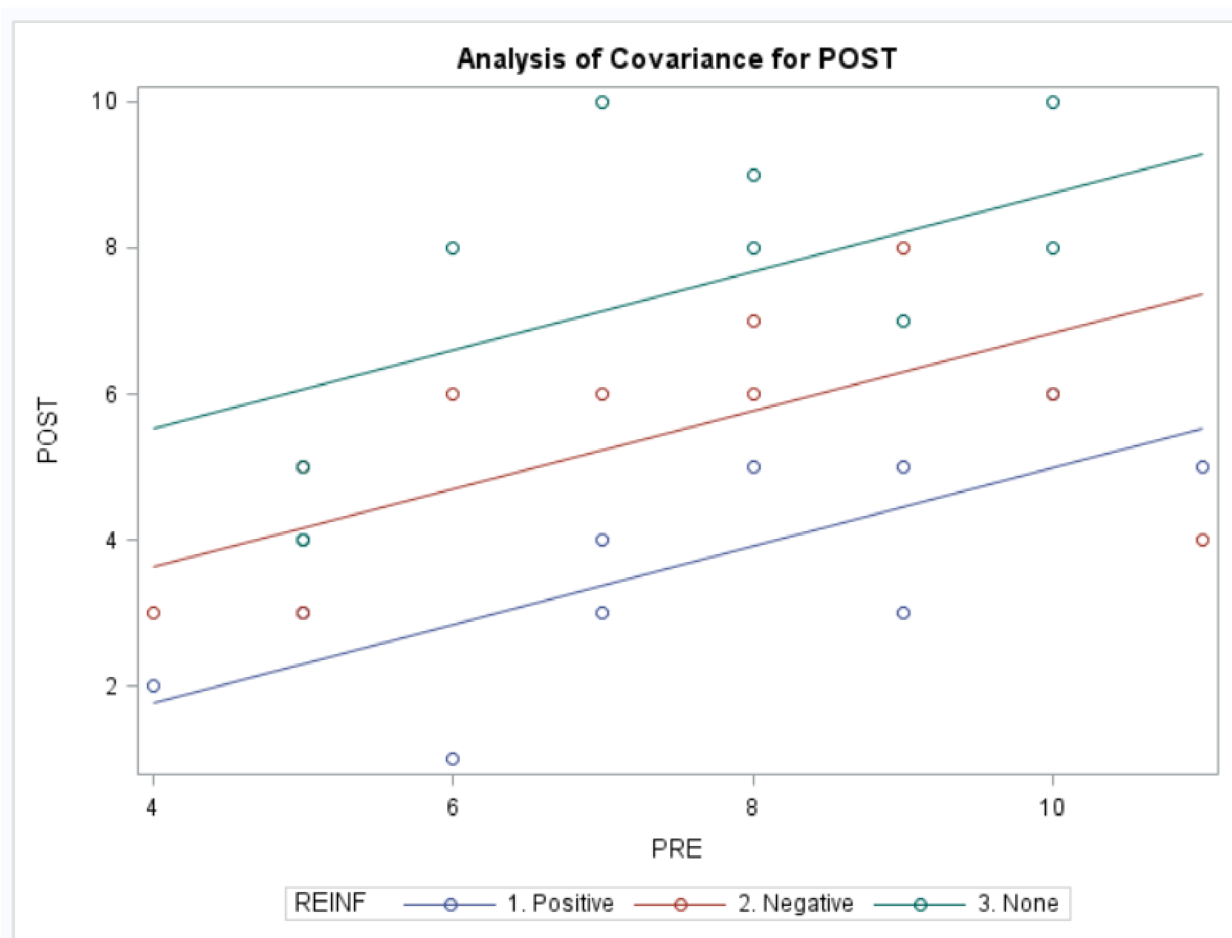
Notes:

Use Type III Sums of Squares (Type III SS).

The *DF* associated with the IVs (EXPERIENCE, METHOD, and EXPERIENCE*METHOD, in the bottom table) add up to the "Model" *DF* in the ANOVA table at the top.

There was a significant effect of REINF, $F(2, 26) = 16.39$, $p < .001$.

Notice that there was a significant effect of PRE, $F(1, 26) = 17.07$, $p < .001$. This is to be expected, because PRE is the covariate and should be highly correlated with POST.



Adjusted POSTTEST means

Adjusted Means & Tukey Post-Hoc Comparisons

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

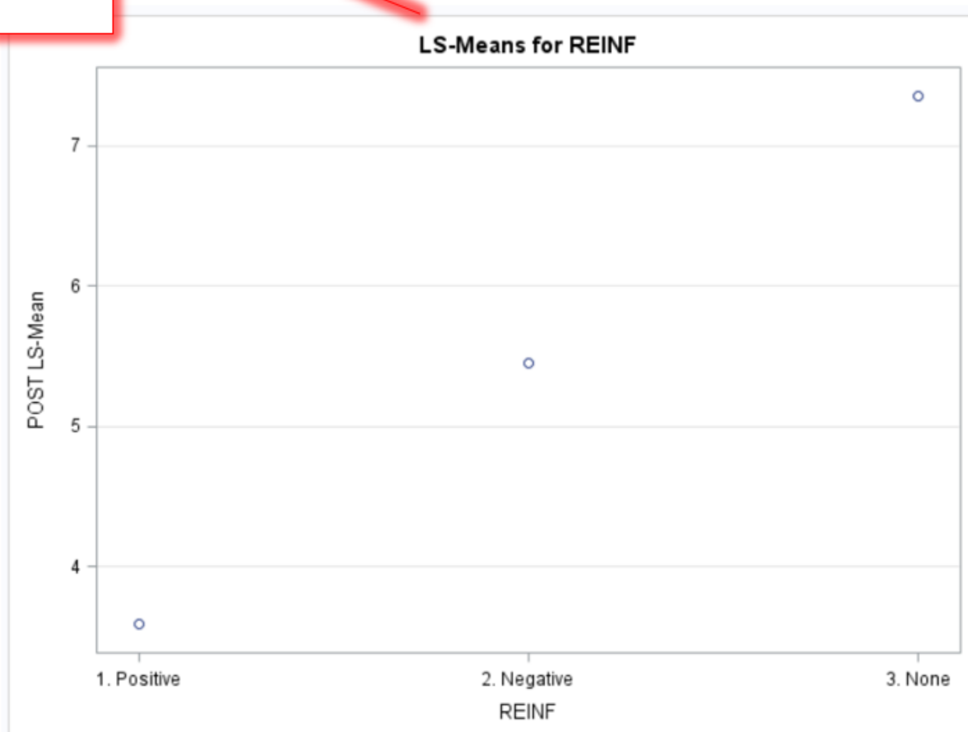
REINF	POST LSMEAN	LSMEAN Number
1. Positive	3.59315708	1
2. Negative	5.45342146	2
3. None	7.35342146	3

Least Squares Means for effect REINF
Pr > \|t\| for H0: LSMean(i)=LSMean(j)
Dependent Variable: POST

i/j	1	2	3
1		0.0232	<.0001
2	0.0232		0.0199
3	<.0001	0.0199	

Tukey's HSD post-hoc test revealed a significant difference between positive and negative reinforcement (p = .023), between positive and no reinforcement (p < .001), and between negative and no reinforcement (p = .020).

Adjusted POSTTEST means



Bryant-Paulson (BP) F Value

The GLM Procedure

Dependent Variable: PRE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.6000000	0.3000000	0.06	0.9391
Error	27	128.6000000	4.7629630		
Corrected Total	29	129.2000000			

R-Square	Coeff Var	Root MSE	PRE Mean
0.004644	29.49218	2.182421	7.400000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
REINF	2	0.60000000	0.30000000	0.06	0.9391

Source	DF	Type III SS	Mean Square	F Value	Pr > F
REINF	2	0.60000000	0.30000000	0.06	0.9391

F value needed for the manual calculation of the Bryant-Paulson post-hoc procedure

Inferential Statistics
Repeated Measures: One Within Factor Design
Research Scenario

A high school math teacher studies the impact of paper color on mathematics test scores. The hypothesis was test scores would be higher on tests taken on pastel green paper than tests taken on bright yellow paper or traditional white paper because the cool color would have a calming effect and reduce test-taking anxiety. Weekly math tests for an Algebra I class were printed in equal quantities on the three colors of paper. The order of treatment was counterbalanced such that one third of the students were randomly assigned to a different color each week over a three-week period. Below are the test scores by student by paper color.

Student	Paper Color		
	Yellow (Y)	Green (G)	White (W)
1	80	76	77
2	81	89	70
3	39	64	55
4	95	93	91
5	71	90	87
6	86	76	92
7	98	94	83
8	95	92	92
9	73	77	53
10	78	88	83
11	54	64	57
12	73	92	96
13	82	75	69
14	49	67	55
15	83	91	79
16	91	90	88
17	94	97	91
18	85	89	90
19	62	69	45

Inferential Statistics
Repeated Measures: One Within Factor Design
SAS Code

```
      DATA PAPER_COLOR;
            INPUT ID YELLOW GREEN WHITE;
            LINES;
                  1     80    76    77
                  2     81    89    70
                  3     39    64    55
                  4     95    93    91
                  5     71    90    87
                  6     86    76    92
                  7     98    94    83
                  8     95    92    92
                  9     73    77    53
                  10    78    88    83
                  11    54    64    57
                  12    73    92    96
                  13    82    75    69
                  14    49    67    55
                  15    83    91    79
                  16    91    90    88
                  17    94    97    91
                  18    85    89    90
                  19    62    69    45
      RUN;

      PROC PRINT DATA=PAPER_COLOR;
      RUN;

      PROC MEANS DATA=PAPER_COLOR;
      RUN;

(1)   PROC GLM DATA=PAPER_COLOR ORDER=INTERNAL;
(2)         MODEL YELLOW GREEN WHITE= / NOUNI;
(3)         REPEATED COLOR / PRINTE;
            TITLE "Repeated Measures ANOVA: Paper Color";
      RUN;

      TITLE;

      PROC FORMAT;
            VALUE COLOR_FMT
                  1="Yellow"
                  2="Green"
                  3="White";
      RUN;

(4)   DATA PAPER_COLOR_2;
(5)         SET PAPER_COLOR;
(6)         SCORE=YELLOW; COLOR=1; OUTPUT;
(7)         SCORE=GREEN; COLOR=2; OUTPUT;
(8)         SCORE=WHITE; COLOR=3; OUTPUT;
```

```
(9)         FORMAT COLOR COLOR_FMT.;
(10)        DROP YELLOW GREEN WHITE;
      RUN;

(11)  PROC PRINT DATA=PAPER_COLOR_2;
      RUN;

(12)  PROC SGPANEL DATA=PAPER_COLOR_2;
(13)        PANELBY COLOR / UNISCALE=ROW;
(14)        HISTOGRAM SCORE;
(15)        DENSITY SCORE;
      RUN;

(16)  PROC GLM DATA=PAPER_COLOR_2 ORDER=INTERNAL PLOTS=DIAGNOSTICS;
            CLASS ID COLOR;
(17)        MODEL SCORE=ID COLOR;
(18)        LSMEANS COLOR / PDIFF CL ADJUST=BON;
            TITLE "Bonferroni Post-Hoc Comparisons";
      RUN;

      TITLE;
      QUIT;
```

(1) **This PROC GLM step is the correct step for repeated measures ANOVA. These are the correct results to use and report for this analysis.** The output from this PROC GLM should be reported as the repeated measures ANOVA results.

(2) As always, the MODEL statement is written as DV(s) = IV(s). Repeated measures ANOVA can be approached from a univariate (single DV) perspective or a multivariate (multiple DV) perspective. *It is for this reason that you will see the repeated measures variable(s) placed in the MODEL statement to the left of the equal sign (=) where the DVs belong.* In this case, the colors (YELLOW, GREEN, and WHITE) are the *within (repeated) factors*, so they are treated as the DVs. There are no *between (grouping) factors*, so there are no IVs; there is nothing to the right of the equal sign. The NOUNI option suppresses some univariate output that you will not need.

(3) The REPEATED statement is what makes this a repeated measures analysis. For the purpose of making the output more informative, you can follow the statement REPEATED with a word to describe/name the repeated measure being analyzed; in this case, it was named COLOR. (TIME, TREATMENT, and TRIAL may be good options in other circumstances.) It is important for you to understand that this word is not a variable; it is just a label that SAS will use in the output. The PRINTE option produces supplemental output, including Mauchly's Test of Sphericity.

(4) Unfortunately, if you need Bonferroni post-hoc testing of the *within factor(s)*, you cannot get it with the previous PROC GLM code. In order to get Bonferroni results, you will first need to "reshape" your data; you will need to go from the "wide" format you began with to a "long" format. In this DATA step, a long data format is created and named PAPER_COLOR_2.

(5) The SET statement copies the data from PAPER_COLOR into PAPER_COLOR_2.

(6) Two new variables are created for the PAPER_COLOR_2 dataset: SCORE and COLOR. SAS copies each YELLOW value to SCORE and simultaneously codes that observation as a "1" for COLOR.

(7) SAS copies each GREEN value to SCORE and simultaneously codes that observation as a "2" for COLOR.
(8) SAS copies each WHITE value to SCORE and simultaneously codes that observation as a "3" for COLOR.
(9) The FORMAT statement is used to apply the labels created in COLOR_FMT to COLOR. *Note that a period (.) follows COLOR_FMT; the code will not work properly if you omit the period.*
(10) The DROP statement deletes the variables YELLOW, GREEN, and WHITE from PAPER_COLOR_2 because they are no longer needed (because COLOR was created and coded as 1, 2, or 3). *Note: These variables still exist in the original dataset PAPER_COLOR.*
(11) Again, it is always a good idea to display your new dataset and confirm there are no data entry errors.
(12) PROC SGPANEL creates a panel of graphs. This will be used to assess whether SCORE is normally distributed for each COLOR.
(13) The PANELBY statement requests that each panel represent one level of COLOR. The UNISCALE=ROW option requests that all panels have the same *x*-axis scale.
(14) The HISTOGRAM statement will produce histograms of SCORE. These histograms will be produced by COLOR per the statement in (13).
(15) The DENSITY statement will overlay a normal curve on each histogram.
(16) **This PROC GLM step is used generate the Bonferroni post-hoc comparisons. Do NOT use or report any of the other results from this step.** Some of the results from this PROC GLM will match the repeated measures ANOVA results, but some of the results are different. Only the Bonferroni post-hoc comparisons should be reported.
(17) The MODEL statement includes the ID variable as an IV. Note that the new variable SCORE is the DV and the new grouping variable COLOR is an IV.
(18) The LSMEANS ("least squares" or "estimated marginal" means) statement requests the adjusted means, as well as the Bonferroni post-hoc comparisons. The CL option requests confidence limits (confidence intervals) for the results.

Inferential Statistics
Repeated Measures: One Within Factor Design
Selected Output

Obs	ID	YELLOW	GREEN	WHITE
1	1	80	76	77
2	2	81	89	70
3	3	39	64	55
4	4	95	93	91
5	5	71	90	87
6	6	86	76	92
7	7	98	94	83
8	8	95	92	92
9	9	73	77	53
10	10	78	88	83
11	11	54	64	57
12	12	73	92	96
13	13	82	75	69
14	14	49	67	55
15	15	83	91	79
16	16	91	90	88
17	17	94	97	91
18	18	85	89	90
19	19	62	69	45

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
ID	19	10.0000000	5.6273143	1.0000000	19.0000000
YELLOW	19	77.3157895	16.4217086	39.0000000	98.0000000
GREEN	19	82.7894737	11.0784021	64.0000000	97.0000000
WHITE	19	76.4736842	16.2356549	45.0000000	96.0000000

The assumption of sphericity was found to be tenable, Mauchly's criterion ($df = 2$) = 0.937, $\chi^2 = 1.110$, $p = .574$. [Note: The "Orthogonal Components" results are reported.]

Sphericity Tests

Variables	DF	Mauchly's Criterion	Chi-Square	Pr > ChiSq
Transformed Variates	2	0.6269285	7.9376864	0.0189
Orthogonal Components	2	0.9367721	1.1103591	0.5740

This output is from the multivariate perspective, and you may disregard it.

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of no COLOR Effect
H = Type III SSCP Matrix for COLOR
E = Error SSCP Matrix

S=1 M=0 N=7.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.62033770	5.20	2	17	0.0173
Pillai's Trace	0.37966230	5.20	2	17	0.0173
Hotelling-Lawley Trace	0.61202519	5.20	2	17	0.0173
Roy's Greatest Root	0.61202519	5.20	2	17	0.0173

Repeated Measures ANOVA Results (Within Subjects Effects)

There was a significant effect of paper color on math test scores, $F(2, 36) = 4.14$, $p = .024$.

Repeated Measures ANOVA: Paper Color

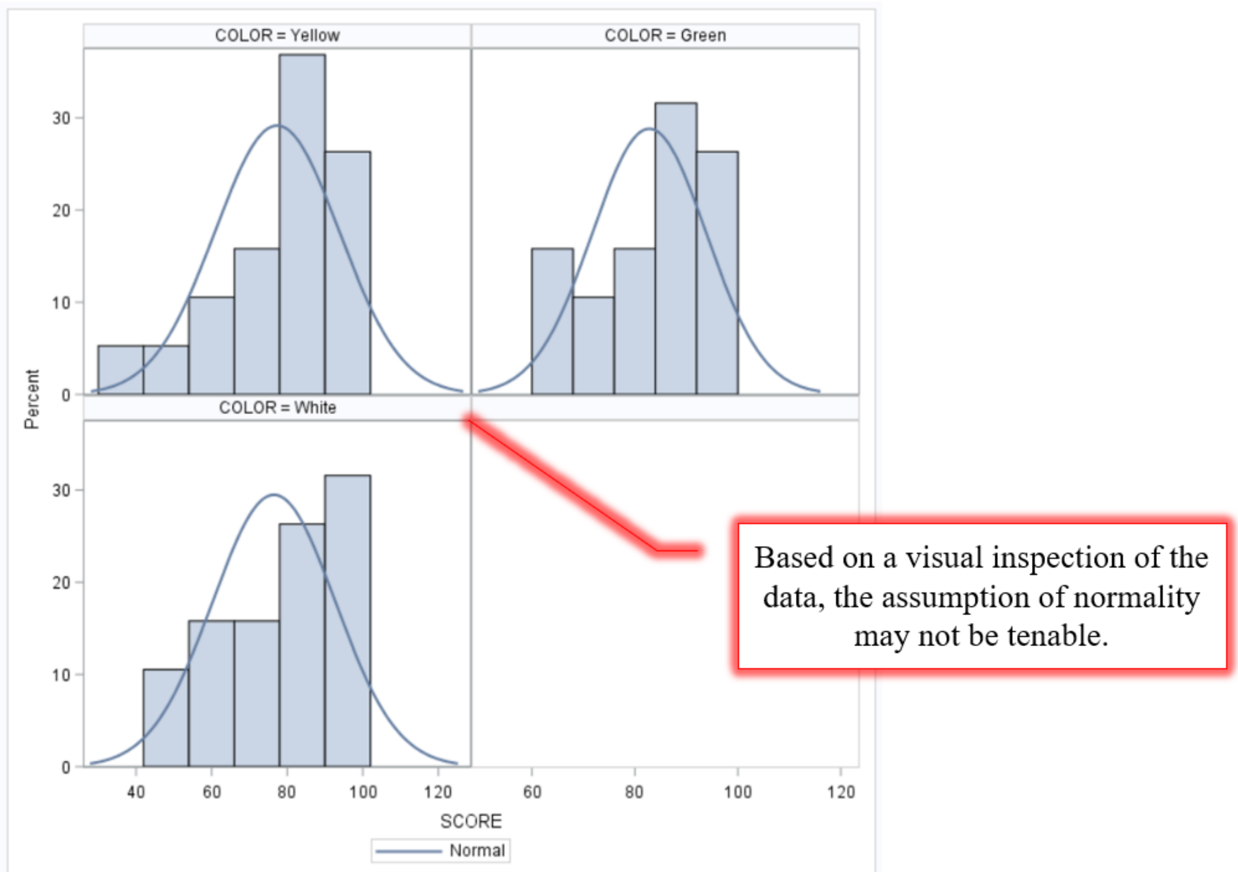
The GLM Procedure
Repeated Measures Analysis of Variance
Univariate Tests of Hypotheses for Within Subject Effects

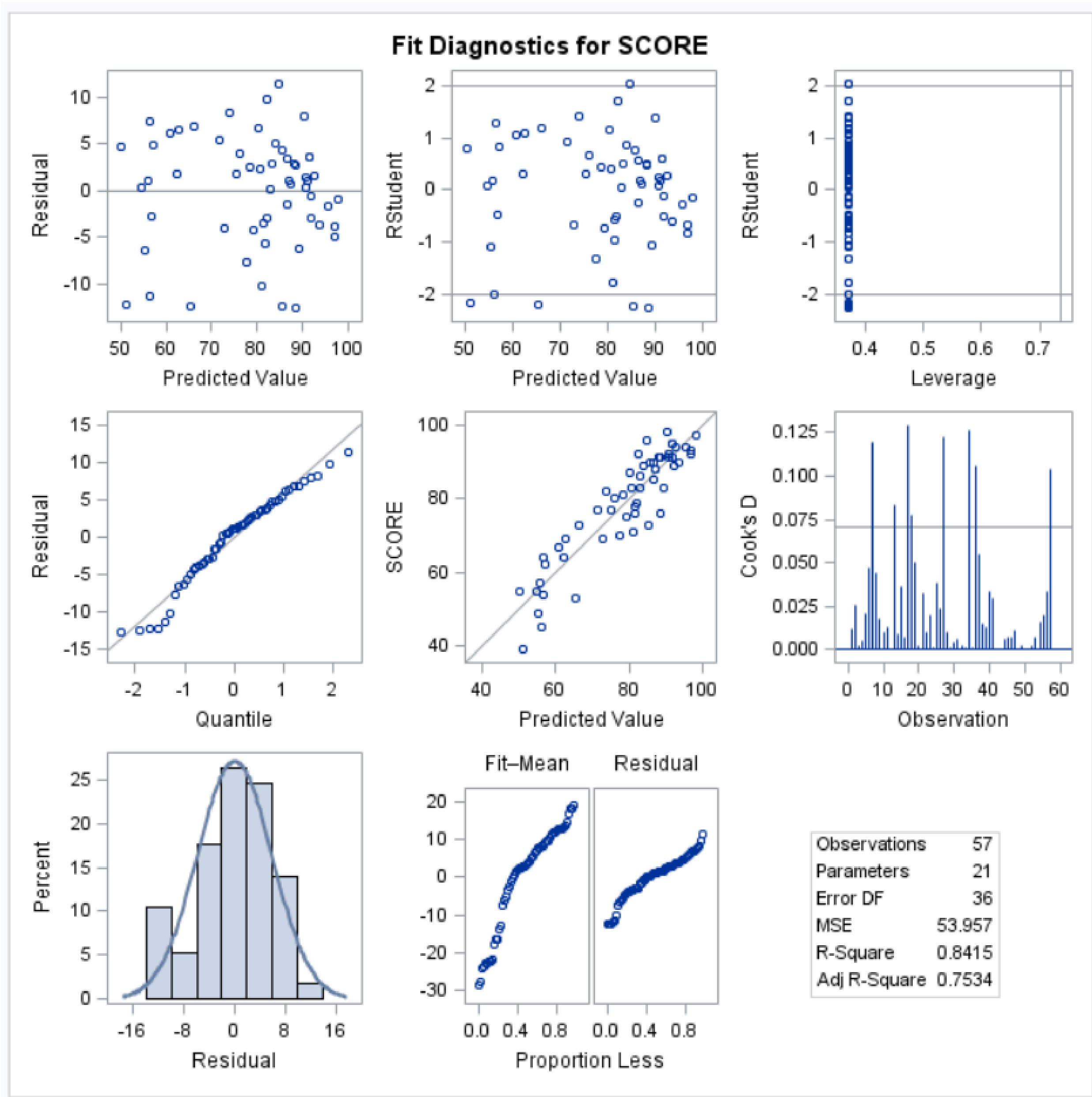
Source	DF	Type III SS	Mean Square	F Value	Pr > F	Adj Pr > F G - G	Adj Pr > F H - F
COLOR	2	446.877193	223.438596	4.14	0.0241	0.0267	0.0241
Error(COLOR)	36	1942.456140	53.957115				

Greenhouse-Geisser Epsilon	0.9405
Huynh-Feldt Epsilon	1.0466

If the assumption of sphericity is NOT tenable, the Greenhouse-Geisser (G-G) or Huynh-Feldt (H-F) the adjusted p value, along with the corresponding epsilon value, should be reported.

Obs	ID	SCORE	COLOR
1	1	80	Yellow
2	1	76	Green
3	1	77	White
4	2	81	Yellow
5	2	89	Green
6	2	70	White
7	3	39	Yellow
8	3	64	Green
9	3	55	White
10	4	95	Yellow
11	4	93	Green
12	4	91	White
13	5	71	Yellow
14	5	90	Green
15	5	87	White
16	6	86	Yellow
17	6	76	Green
18	6	92	White
19	7	98	Yellow
20	7	94	Green
21	7	83	White
22	8	95	Yellow


COLOR = Yellow
COLOR = Green
COLOR = White
Percent
0
10
20
30
40
60
80
100
120
SCORE
Normal
Based on a visual inspection of the data, the assumption of normality may not be tenable.


Fit Diagnostics for SCORE
Residual
Predicted Value
RStudent
Leverage
Quantile
SCORE
Cook's D
Observation
Percent
Fit–Mean
Residual
Proportion Less
Observations 57
Parameters 21
Error DF 36
MSE 53.957
R-Square 0.8415
Adj R-Square 0.7534

Bonferroni Post-Hoc Comparisons

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

COLOR	SCORE LSMEAN	LSMEAN Number
Yellow	77.3157895	1
Green	82.7894737	2
White	76.4736842	3

Least Squares Means for effect COLOR
Pr > \|t\| for H0: LSMean(i)=LSMean(j)
Dependent Variable: SCORE

i/j	1	2	3
1		0.0827	1.0000
2	0.0827		0.0356
3	1.0000	0.0356	

Bonferroni's post-hoc test revealed a significant difference between test scores on green and white paper (p = .036).

COLOR	SCORE LSMEAN	95% Confidence Limits	
Yellow	77.315789	73.898076	80.733503
Green	82.789474	79.371760	86.207187
White	76.473684	73.055971	79.891398

Least Squares Means for Effect COLOR

i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-5.473684	-11.458024	0.510656
1	3	0.842105	-5.142235	6.826445
2	3	6.315789	0.331450	12.300129


LS-Means for COLOR
SCORE LS-Mean
82
80
78
76
Yellow
Green
White
COLOR

Inferential Statistics
Repeated Measures: One Within Factor and One Between Factor Design
Research Scenario

A researcher wanted to investigate the effect of anxiety on math performance for fourth graders in a variety of testing time constraints. She came up with a 20-question multiplication test and gave it to twelve participants. For the first trial, she allowed participants one hour. For the second trial, participants had 45 minutes to take the test. The time for the third trial was 30 minutes. The time for the final trial was 15 minutes. Prior to giving the math test, she assessed students' test anxiety level; students with low test anxiety were in Group 1 and those with high test anxiety were in Group 2. The tests, however, were the same between groups. The data can be seen below.

Subject	Anxiety	Trial 1	Trial 2	Trial 3	Trial 4
1	1	18	14	12	6
2	1	19	12	8	4
3	1	14	10	6	2
4	1	16	12	10	4
5	1	12	8	6	2
6	1	18	10	5	1
7	2	16	10	8	4
8	2	18	8	4	1
9	2	16	12	6	2
10	2	19	16	10	8
11	2	16	14	10	9
12	2	16	12	8	8

Inferential Statistics
Repeated Measures: One Within Factor and One Between Factor Design
SAS Code

```
      PROC FORMAT;
            VALUE ANX_FMT
                  1="1. Low Anxiety"
                  2="2. High Anxiety";
      RUN;

      DATA TESTING_TIME;
            INPUT ID ANXIETY TRIAL1 TRIAL2 TRIAL3 TRIAL4;
            FORMAT ANXIETY ANX_FMT.;
            LINES;
                  1     1     18    14    12    6
                  2     1     19    12    8     4
                  3     1     14    10    6     2
                  4     1     16    12    10    4
                  5     1     12    8     6     2
                  6     1     18    10    5     1
                  7     2     16    10    8     4
                  8     2     18    8     4     1
                  9     2     16    12    6     2
                  10    2     19    16    10    8
                  11    2     16    14    10    9
                  12    2     16    12    8     8
      RUN;

      PROC PRINT DATA=TESTING_TIME;
      RUN;

      PROC MEANS DATA=TESTING_TIME;
(1)         BY ANXIETY;
      RUN;

(2)   PROC GLM DATA=TESTING_TIME;
            CLASS ANXIETY;
(3)         MODEL TRIAL1-TRIAL4=ANXIETY / NOUNI;
(4)         MEANS ANXIETY / HOVTEST=LEVENE(TYPE=ABS);
(5)         REPEATED TRIAL / PRINTE;
(6)         LSMEANS ANXIETY / PDIFF CL ADJUST=BON;
            TITLE "Repeated Measures ANOVA: Testing Time";
      RUN;

      TITLE;

(7)   DATA TESTING_TIME_2;
            SET TESTING_TIME;
            SCORE=TRIAL1; TRIAL=1; OUTPUT;
            SCORE=TRIAL2; TRIAL=2; OUTPUT;
            SCORE=TRIAL3; TRIAL=3; OUTPUT;
            SCORE=TRIAL4; TRIAL=4; OUTPUT;
            DROP TRIAL1-TRIAL4;
      RUN;
```

```
      PROC PRINT DATA=TESTING_TIME_2;
      RUN;

(8)   PROC GLM DATA=TESTING_TIME_2 ORDER=INTERNAL PLOTS=DIAGNOSTICS;
            CLASS ID ANXIETY TRIAL;
(9)         MODEL SCORE=ID ANXIETY TRIAL;
(10)        LSMEANS TRIAL / PDIFF CL ADJUST=BON;
            TITLE "Bonferroni Post-Hoc Comparisons";
      RUN;

      TITLE;
      QUIT;
```

(1) The BY statement requires PROC MEANS to calculate statistics for each level of ANXIETY; in other words, it will calculate statistics for each TRIAL*ANXIETY cell.

(2) **This PROC GLM step is the correct step for repeated measures ANOVA. These are the correct results to use and report for this analysis.** The output from this PROC GLM should be reported as the repeated measures ANOVA results.

(3) As always, the MODEL statement is written as DV(s) = IV(s). Repeated measures ANOVA can be approached from a univariate (single DV) perspective or a multivariate (multiple DV) perspective. *It is for this reason that you will see the repeated measures variable(s) placed in the MODEL statement to the left of the equal sign (=) where the DVs belong.* In this case, the trials (TRIAL1, TRIAL2, TRIAL3, and TRIAL4) are the *within (repeated) factors*, so they are treated as the DVs. You have the option of writing each of these DVs individually, but writing TRIAL1-TRIAL4 is quicker. ANXIETY is a *between (grouping) factor*, which is treated as an IV and written to the right of the equal sign. The NOUNI option suppresses some univariate output that you will not need.

(4) The MEANS statement with the option HOVTEST=LEVENE will conduct Levene's Test of Homogeneity of Variances, to test the assumption that the cells have equal variances, for TRIAL1-TRIAL4.

(5) The REPEATED statement is what makes this a repeated measures analysis. For the purpose of making the output more informative, you can follow the statement REPEATED with a word to describe/name the repeated measure being analyzed; in this case, it was named TRIAL. (TIME and TREATMENT may be good options in other circumstances.) It is important for you to understand that this word is not a variable; it is just a label that SAS will use in the output. The PRINTE option produces supplemental output, including Mauchly's Test of Sphericity.

(6) The LSMEANS ("least squares" or "estimated marginal" means) statement requests the adjusted means, as well as the Bonferroni post-hoc comparisons, for the *between factor(s)*. The CL option requests confidence limits (confidence intervals) for the results.

(7) Unfortunately, if you need Bonferroni post-hoc testing of the *within factor(s)*, you cannot get it with the previous PROC GLM code. In order to get Bonferroni results, you will first need to "reshape" your data; you will need to go from the "wide" format you began with to a "long" format. In this DATA step, a long data format is created and named TESTING_TIME_2.

(8) **This PROC GLM step is used generate the Bonferroni post-hoc comparisons. Do NOT use or report any of the other results from this step.** Some of the results from

this PROC GLM will match the repeated measures ANOVA results, but some of the results are different. Only the Bonferroni post-hoc comparisons should be reported.

(9) The MODEL statement includes the ID variable as an IV. Note that the new variable SCORE is the DV and the new grouping variable TRIAL is an IV.

(10) The LSMEANS ("least squares" or "estimated marginal" means) statement requests the adjusted means, as well as the Bonferroni post-hoc comparisons. The CL option requests confidence limits (confidence intervals) for the results.

Inferential Statistics
Repeated Measures: One Within Factor and One Between Factor Design
Selected Output

Obs	ID	ANXIETY	TRIAL1	TRIAL2	TRIAL3	TRIAL4
1	1	1. Low Anxiety	18	14	12	6
2	2	1. Low Anxiety	19	12	8	4
3	3	1. Low Anxiety	14	10	6	2
4	4	1. Low Anxiety	16	12	10	4
5	5	1. Low Anxiety	12	8	6	2
6	6	1. Low Anxiety	18	10	5	1
7	7	2. High Anxiety	16	10	8	4
8	8	2. High Anxiety	18	8	4	1
9	9	2. High Anxiety	16	12	6	2
10	10	2. High Anxiety	19	16	10	8
11	11	2. High Anxiety	16	14	10	9
12	12	2. High Anxiety	16	12	8	8

The MEANS Procedure

ANXIETY=1. Low Anxiety

Variable	N	Mean	Std Dev	Minimum	Maximum
ID	6	3.5000000	1.8708287	1.0000000	6.0000000
TRIAL1	6	16.1666667	2.7141604	12.0000000	19.0000000
TRIAL2	6	11.0000000	2.0976177	8.0000000	14.0000000
TRIAL3	6	7.8333333	2.7141604	5.0000000	12.0000000
TRIAL4	6	3.1666667	1.8348479	1.0000000	6.0000000

ANXIETY=2. High Anxiety

Variable	N	Mean	Std Dev	Minimum	Maximum
ID	6	9.5000000	1.8708287	7.0000000	12.0000000
TRIAL1	6	16.8333333	1.3291601	16.0000000	19.0000000
TRIAL2	6	12.0000000	2.8284271	8.0000000	16.0000000
TRIAL3	6	7.6666667	2.3380904	4.0000000	10.0000000
TRIAL4	6	5.3333333	3.4448028	1.0000000	9.0000000

The assumption of homogeneity of variances for the "anxiety" and "no anxiety" groups was found to be tenable for TRIAL1, $F(1, 10) = 3.31$, $p = .099$.

The assumption of homogeneity of variances for the "anxiety" and "no anxiety" groups was found to be tenable for TRIAL2, $F(1, 10) = 0.16$, $p = .701$.

The assumption of homogeneity of variances for the "anxiety" and "no anxiety" groups was found to be tenable for TRIAL3, $F(1, 10) = 0.27$, $p = .617$.

The assumption of homogeneity of variances for the "anxiety" and "no anxiety" groups was not found to be tenable for TRIAL4, $F(1, 10) = 7.79$, $p = .019$.

The GLM Procedure

Levene's Test for Homogeneity of TRIAL1 Variance
ANOVA of Absolute Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
ANXIETY	1	3.3426	3.3426	3.31	0.0988
Error	10	10.0926	1.0093		

Levene's Test for Homogeneity of TRIAL2 Variance
ANOVA of Absolute Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
ANXIETY	1	0.3333	0.3333	0.16	0.7009
Error	10	21.3333	2.1333		

Levene's Test for Homogeneity of TRIAL3 Variance
ANOVA of Absolute Deviations from Group Means

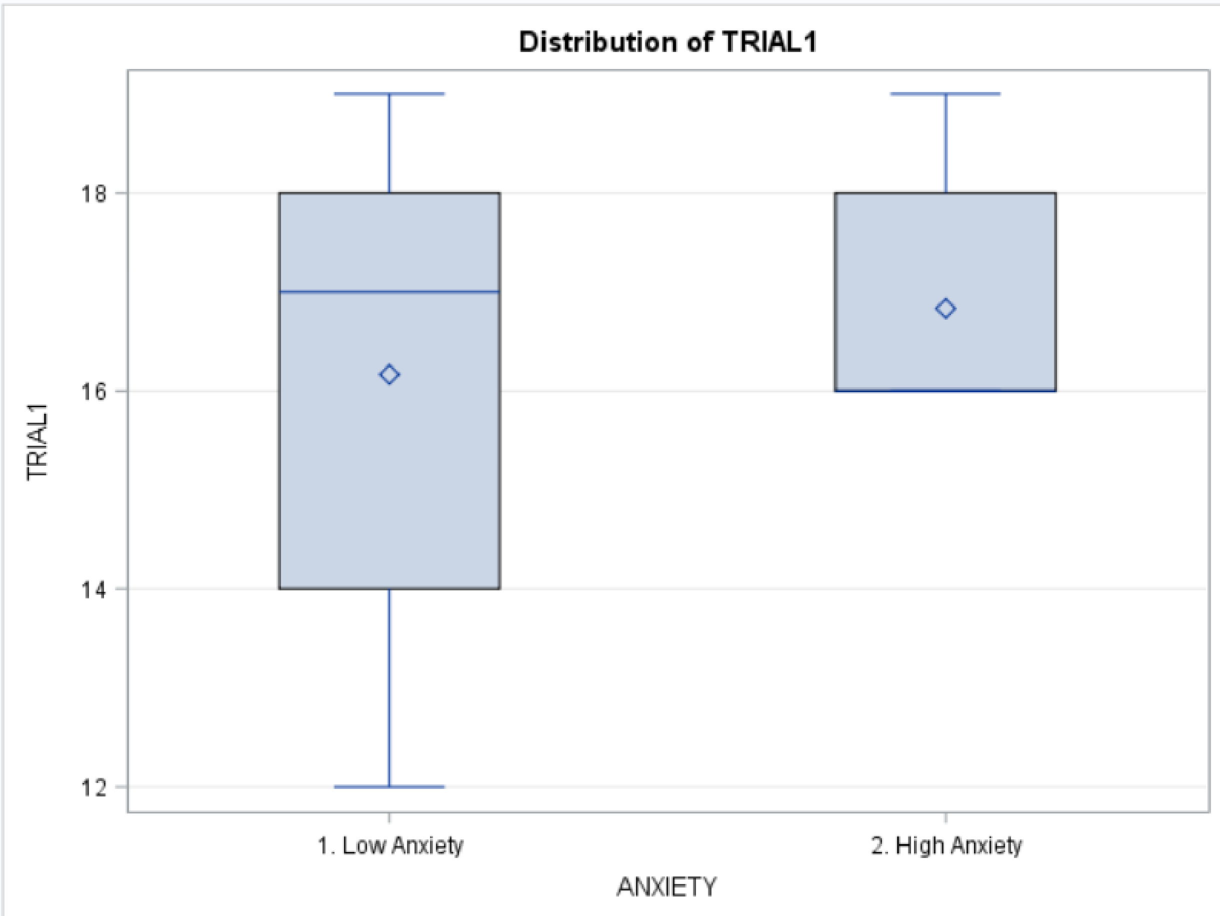
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
ANXIETY	1	0.4537	0.4537	0.27	0.6170
Error	10	17.0370	1.7037		

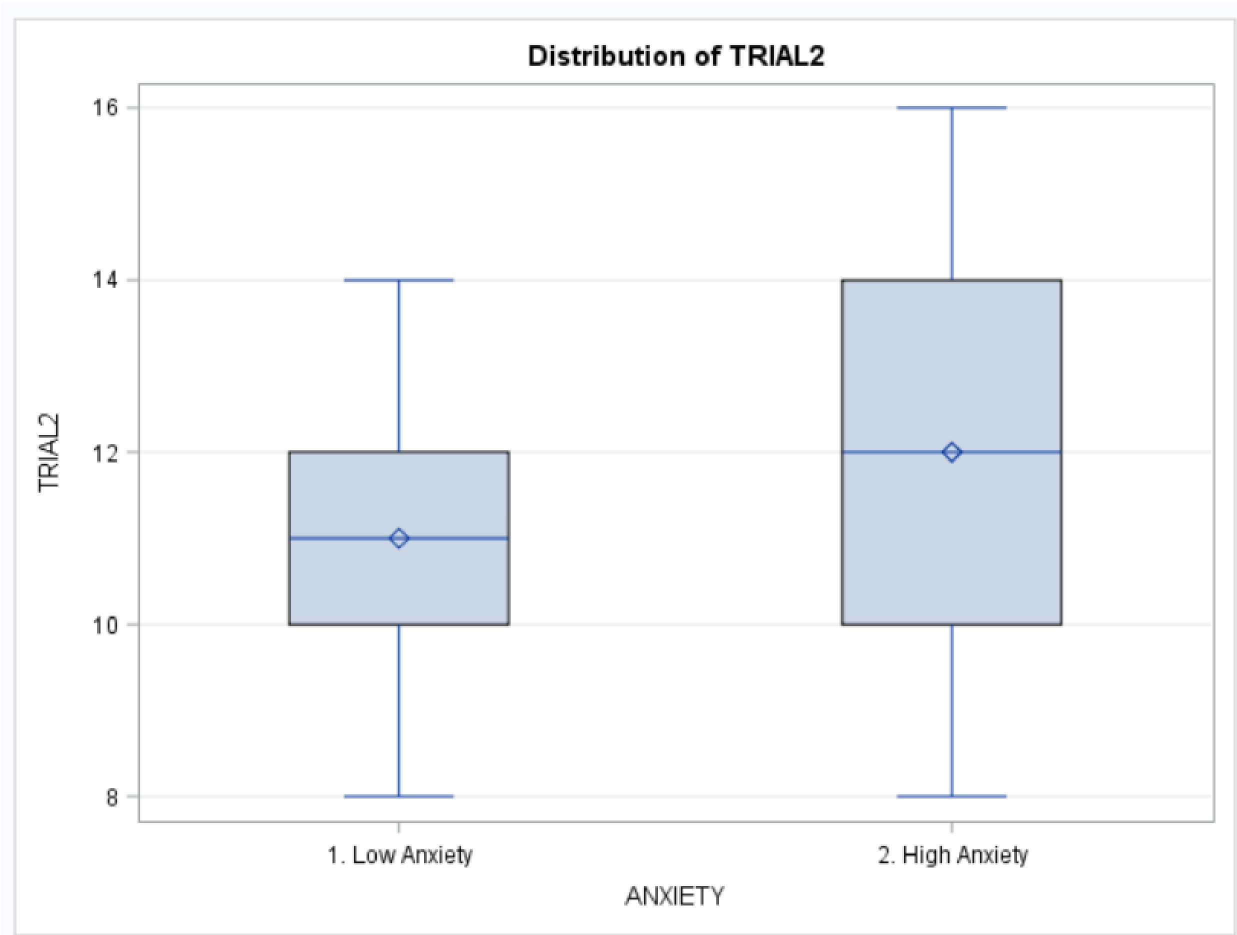
Levene's Test for Homogeneity of TRIAL4 Variance
ANOVA of Absolute Deviations from Group Means

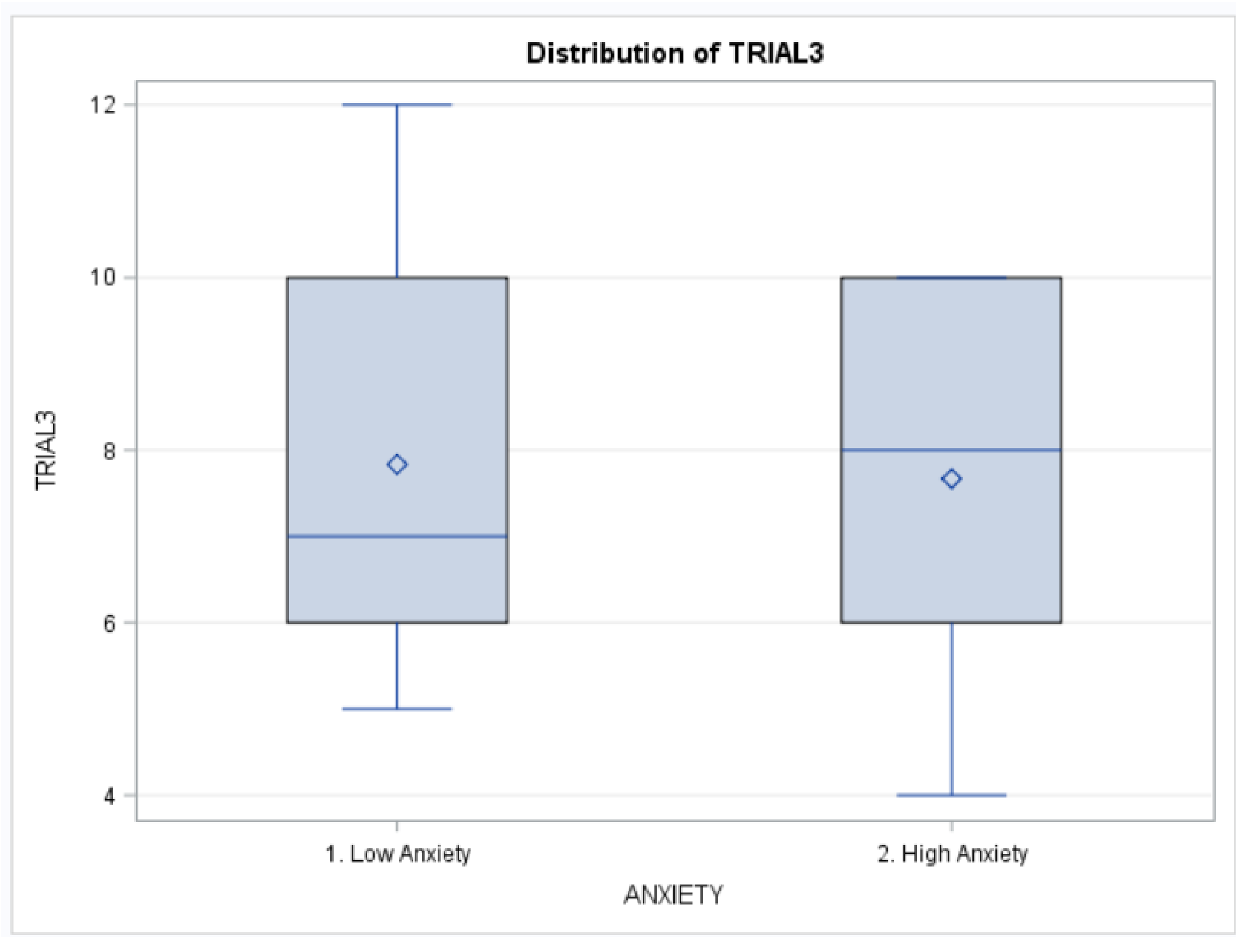
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
ANXIETY	1	6.7500	6.7500	7.79	0.0191
Error	10	8.6667	0.8667		

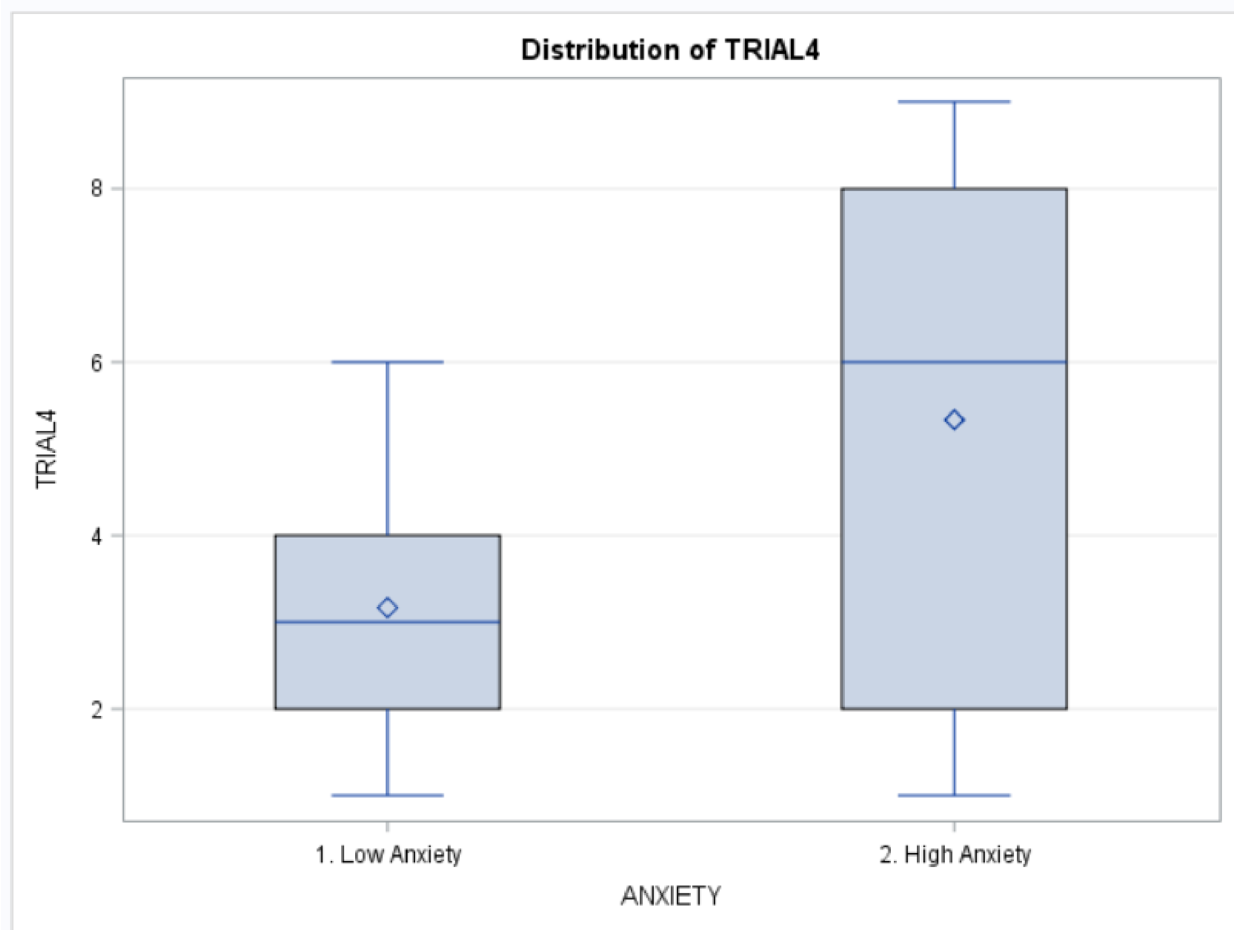
Repeated Measures ANOVA: Testing Time

The GLM Procedure




Distribution of TRIAL2
16
14
12
10
8
TRIAL2
1. Low Anxiety
2. High Anxiety
ANXIETY


Distribution of TRIAL3
12
10
8
6
4
TRIAL3
1. Low Anxiety
2. High Anxiety
ANXIETY



Level of ANXIETY	N	TRIAL1		TRIAL2		TRIAL3		TRIAL4	
		Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
1. Low Anxiety	6	16.1666667	2.71416040	11.0000000	2.09761770	7.83333333	2.71416040	3.16666667	1.83484786
2. High Anxiety	6	16.8333333	1.32916014	12.0000000	2.82842712	7.66666667	2.33809039	5.33333333	3.44480285

The assumption of sphericity was found to be tenable, Mauchly's criterion (5) = 0.283, $\chi^2 = 11.011$, $p = .051$. [Note: The "Orthogonal Components" results are reported.]

Sphericity Tests

Variables	DF	Mauchly's Criterion	Chi-Square	Pr > ChiSq
Transformed Variates	5	0.1588511	16.04704	0.0067
Orthogonal Components	5	0.282986	11.010565	0.0512

This output is from the multivariate perspective, and you may disregard it.

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of no TRIAL Effect
H = Type III SSCP Matrix for TRIAL
E = Error SSCP Matrix

S=1 M=0.5 N=3

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.03949423	64.85	3	8	<.0001
Pillai's Trace	0.96050577	64.85	3	8	<.0001
Hotelling-Lawley Trace	24.32015172	64.85	3	8	<.0001
Roy's Greatest Root	24.32015172	64.85	3	8	<.0001

This output is from the multivariate perspective, and you may disregard it.

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of no TRIAL*ANXIETY Effect
H = Type III SSCP Matrix for TRIAL*ANXIETY
E = Error SSCP Matrix

S=1 M=0.5 N=3

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.52102440	2.45	3	8	0.1381
Pillai's Trace	0.47897560	2.45	3	8	0.1381
Hotelling-Lawley Trace	0.91929590	2.45	3	8	0.1381
Roy's Greatest Root	0.91929590	2.45	3	8	0.1381

Repeated Measures ANOVA Results (Between Subjects Effects)

There was not a significant effect of ANXIETY on math test scores, $F(1, 10) = 0.59$, $p = .460$.

Repeated Measures ANOVA: Testing Time

The GLM Procedure
Repeated Measures Analysis of Variance
Tests of Hypotheses for Between Subjects Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
ANXIETY	1	10.0833333	10.0833333	0.59	0.4602
Error	10	170.9166667	17.0916667		

Repeated Measures ANOVA Results (Within Subjects Effects)

There was not a significant difference between TRIAL*ANXIETY means, $F(3, 30) = 1.09$, $p = .368$. There was a significant difference between TRIAL means, $F(3, 30) = 128.63$, $p < .001$.

Repeated Measures ANOVA: Testing Time

The GLM Procedure
Repeated Measures Analysis of Variance
Univariate Tests of Hypotheses for Within Subject Effects

						Adj Pr > F	
Source	DF	Type III SS	Mean Square	F Value	Pr > F	G - G	H-F-L
TRIAL	3	991.5000000	330.5000000	128.63	<.0001	<.0001	<.0001
TRIAL*ANXIETY	3	8.4166667	2.8055556	1.09	0.3677	0.3463	0.3529
Error(TRIAL)	30	77.0833333	2.5694444				

Greenhouse-Geisser Epsilon	0.5441
Huynh-Feldt-Lecoutre Epsilon	0.6355

If the assumption of sphericity is NOT tenable, the Greenhouse-Geisser (G-G) or Huynh-Feldt (H-F) adjusted p value, along with the associated epsilon value, should be reported.

Bonferroni post-hoc testing is not required here because ANXIETY (a) was not significant and (b) only has two levels.

Repeated Measures ANOVA: Testing Time

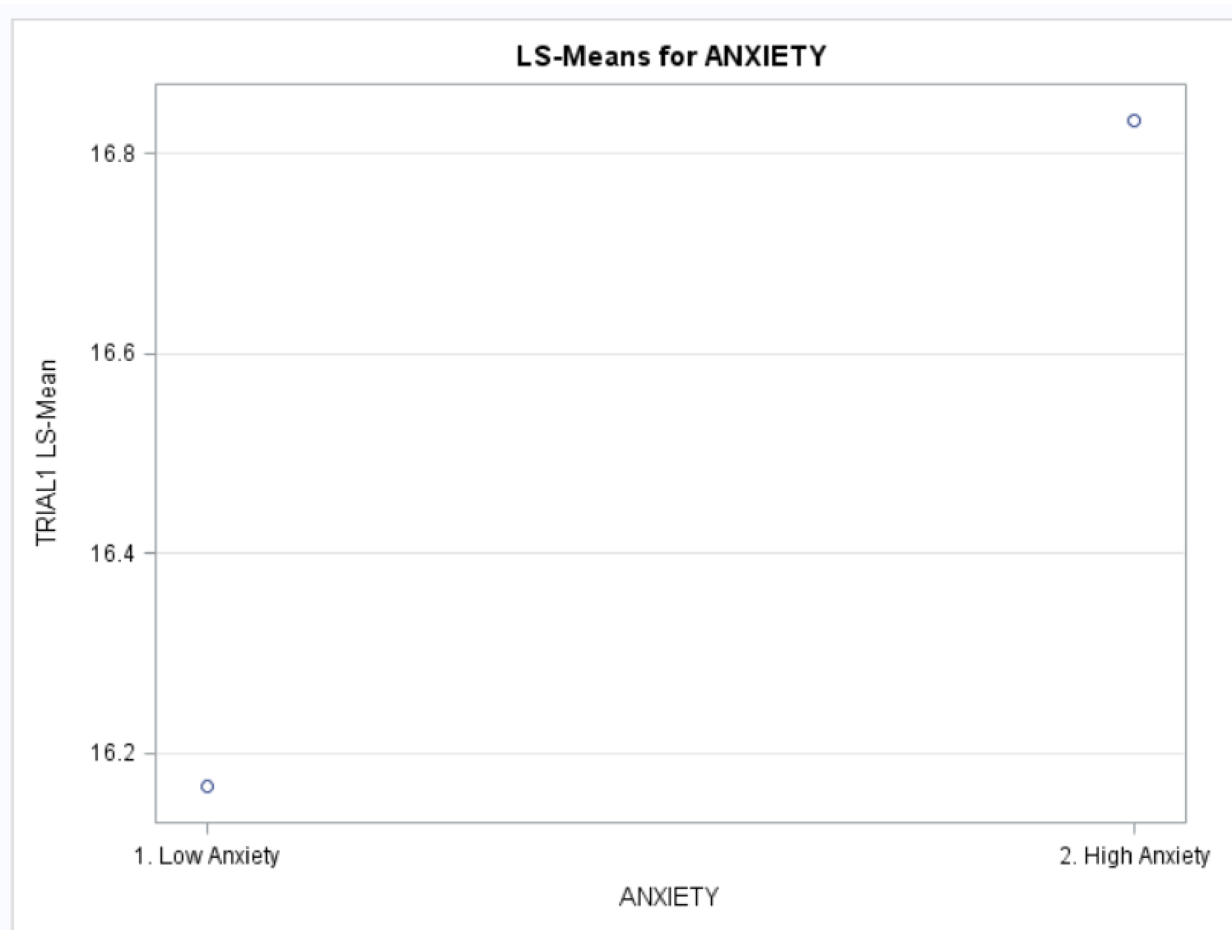
The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Bonferroni

ANXIETY	TRIAL1 LSMEAN	H0:LSMean1=LSMean2 Pr > |t|
1. Low Anxiety	16.1666667	0.6008
2. High Anxiety	16.8333333	

ANXIETY	TRIAL1 LSMEAN	95% Confidence Limits	
1. Low Anxiety	16.166667	14.222801	18.110532
2. High Anxiety	16.833333	14.889468	18.777199

Least Squares Means for Effect ANXIETY

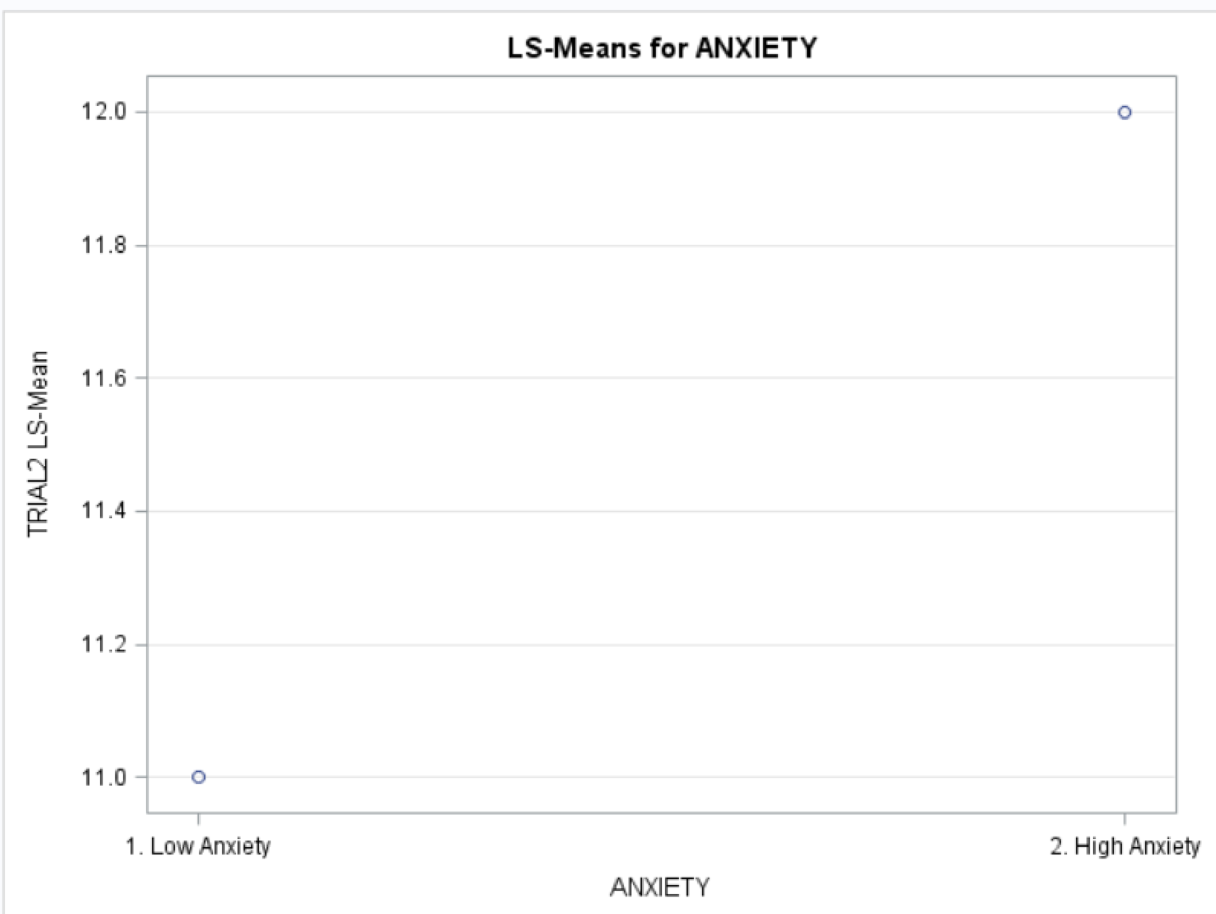
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-0.666667	-3.415708	2.082375



ANXIETY	TRIAL2 LSMEAN	H0:LSMean1=LSMean2 Pr > |t|
1. Low Anxiety	11.0000000	0.5025
2. High Anxiety	12.0000000	

ANXIETY	TRIAL2 LSMEAN	95% Confidence Limits	
1. Low Anxiety	11.000000	8.735030	13.264970
2. High Anxiety	12.000000	9.735030	14.264970

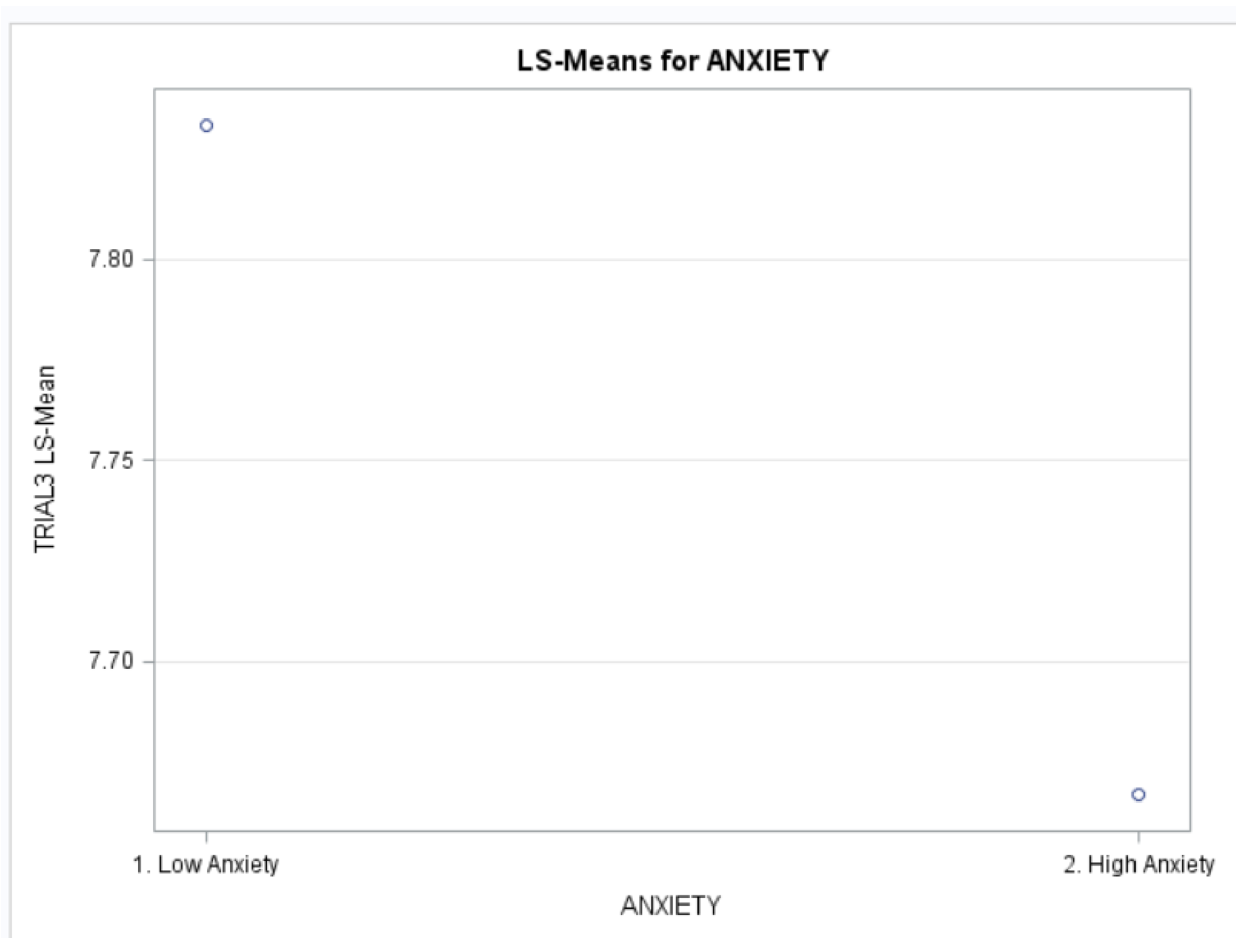
Least Squares Means for Effect ANXIETY				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-1.000000	-4.203151	2.203151



ANXIETY	TRIAL3 LSMEAN	H0:LSMean1=LSMean2 Pr > |t|
1. Low Anxiety	7.83333333	0.9115
2. High Anxiety	7.66666667	

ANXIETY	TRIAL3 LSMEAN	95% Confidence Limits	
1. Low Anxiety	7.833333	5.529127	10.137540
2. High Anxiety	7.666667	5.362460	9.970873

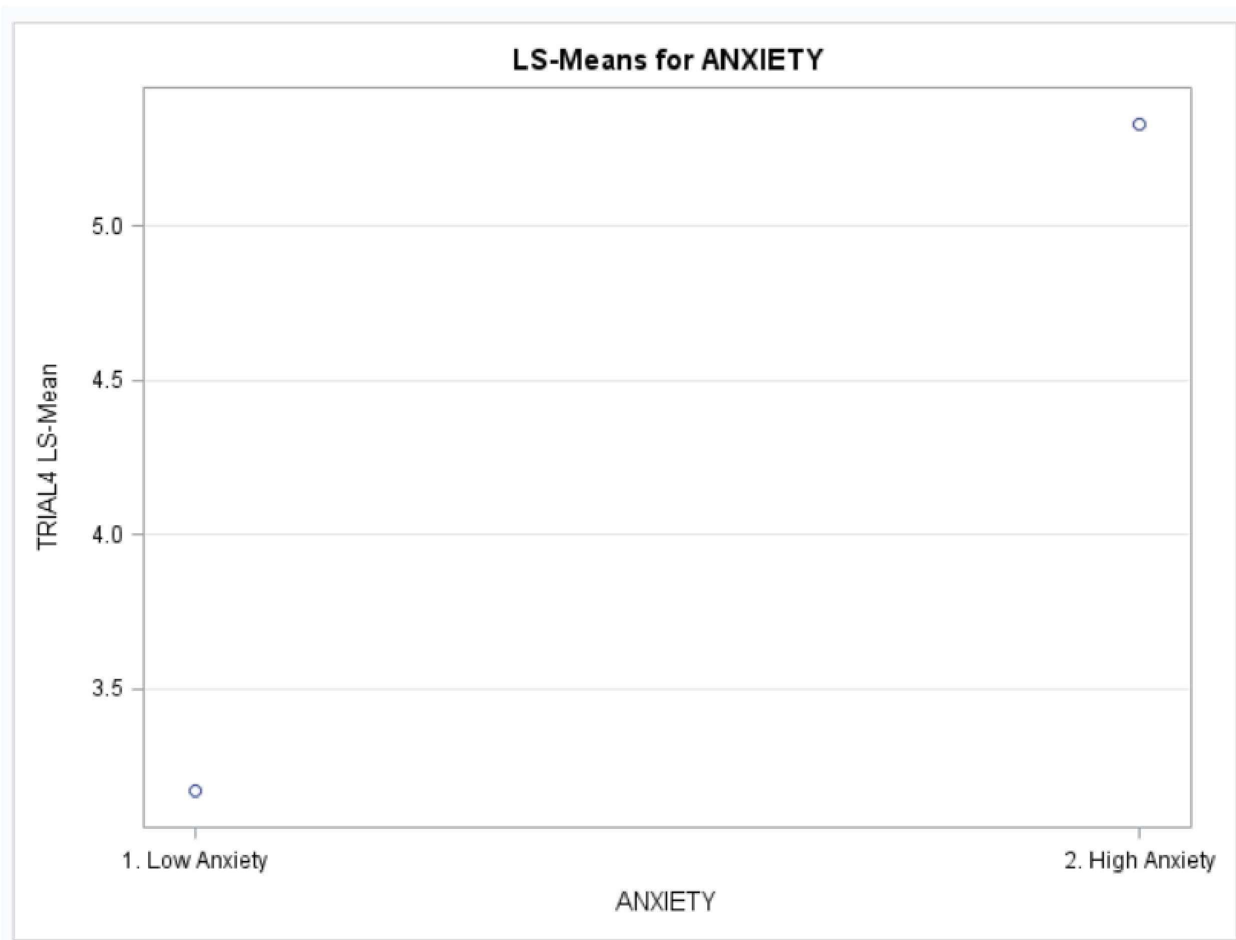
Least Squares Means for Effect ANXIETY				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	0.166667	-3.091973	3.425307



ANXIETY	TRIAL4 LSMEAN	H0:LSMean1=LSMean2 Pr > |t|
1. Low Anxiety	3.16666667	0.2038
2. High Anxiety	5.33333333	

ANXIETY	TRIAL4 LSMEAN	95% Confidence Limits	
1. Low Anxiety	3.166667	0.656231	5.677102
2. High Anxiety	5.333333	2.822898	7.843769

Least Squares Means for Effect ANXIETY				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-2.166667	-5.716959	1.383626



Bonferroni post-hoc testing revealed that all pairwise TRIAL comparisons were significant ($p < .001$).

Bonferroni Post-Hoc Comparisons

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

TRIAL	SCORE LSMEAN	LSMEAN Number
1	16.5000000	1
2	11.5000000	2
3	7.7500000	3
4	4.2500000	4

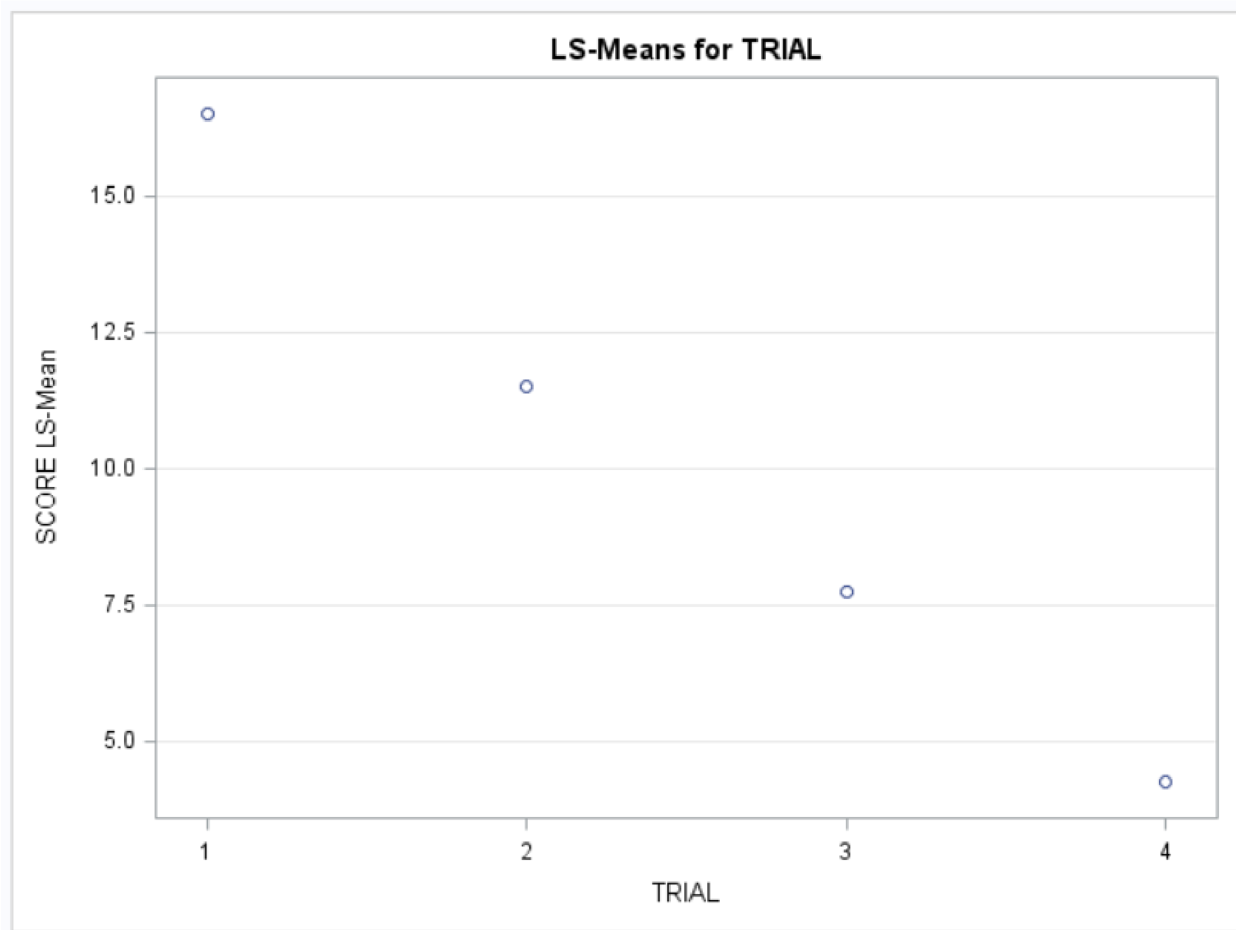
Least Squares Means for effect TRIAL
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: SCORE

i/j	1	2	3	4
1		<.0001	<.0001	<.0001
2	<.0001		<.0001	<.0001
3	<.0001	<.0001		<.0001
4	<.0001	<.0001	<.0001	

TRIAL	SCORE LSMEAN	95% Confidence Limits	
1	16.500000	15.554642	17.445358
2	11.500000	10.554642	12.445358
3	7.750000	6.804642	8.695358
4	4.250000	3.304642	5.195358

Least Squares Means for Effect TRIAL

i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	5.000000	3.155597	6.844403
1	3	8.750000	6.905597	10.594403
1	4	12.250000	10.405597	14.094403
2	3	3.750000	1.905597	5.594403
2	4	7.250000	5.405597	9.094403
3	4	3.500000	1.655597	5.344403


LS-Means for TRIAL
15.0
12.5
10.0
7.5
5.0
SCORE LS-Mean
1
2
3
4
TRIAL

Inferential Statistics
Simple Linear Regression
Research Scenario

A cereal-lover would like to lose weight and is interested to know if the amount of complex carbohydrates in his cereal significantly affects the cereal's calorie count. The complex carbohydrates are measured in grams; carbohydrates and calories values are per serving.

This analysis will use the "UScereal.csv" file that can be found as a companion with this SAS guide. If you prefer to enter the data manually, instead of reading in the data file, the raw data for the variables of interest may be found in the Appendix of this guide.

	A	B	C	D	E	F	G	H	I	J	K	L
1	CEREAL	MFR	CALORIES	PROTEIN	FAT	SODIUM	FIBER	CARBO	SUGARS	SHELF	POTASSIUM	VITAMINS
2	100% Bran	N	212.1212	12.12121	3.0303	393.9394	30.30303	15.15152	18.18182	3	848.48485	enriched
3	All-Bran	K	212.1212	12.12121	3.0303	787.8788	27.27273	21.21212	15.15152	3	969.69697	enriched
4	All-Bran with Extra Fiber	K	100	8	0	280	28	16	0	3	660	enriched
5	Apple Cinnamon Cheerios	G	146.6667	2.66667	2.66667	240	2	14	13.33333	1	93.33333	enriched
6	Apple Jacks	K	110	2	0	125	1	11	14	2	30	enriched
7	Basic 4	G	173.3333	4	2.66667	280	2.66667	24	10.66667	3	133.33333	enriched
8	Bran Chex	R	134.3284	2.98507	1.49254	298.5075	5.97015	22.38806	8.95522	1	186.56716	enriched
9	Bran Flakes	P	134.3284	4.47761	0	313.4328	7.46269	19.40299	7.46269	3	283.58209	enriched
10	Cap'n'Crunch	Q	160	1.33333	2.66667	293.3333	0	16	16	2	46.66667	enriched
11	Cheerios	G	88	4.8	1.6	232	1.6	13.6	0.8	1	84	enriched
12	Cinnamon Toast Crunch	G	160	1.33333	4	280	0	17.33333	12	2	60	enriched
13	Clusters	G	220	6	4	280	4	26	14	3	210	enriched
14	Cocoa Puffs	G	110	1	1	180	0	12	13	2	55	enriched
15	Corn Chex	R	110	2	0	280	0	22	3	1	25	enriched
16	Corn Flakes	K	100	2	0	290	1	21	2	1	35	enriched
17	Corn Pops	K	110	1	0	90	1	13	12	2	20	enriched
18	Count Chocula	G	110	1	1	180	0	12	13	2	65	enriched
19	Cracklin' Oat Bran	K	220	6	6	280	8	20	14	3	320	enriched
20	Crispix	K	110	2	0	220	1	21	3	3	30	enriched
21	Crispy Wheat & Raisins	G	133.3333	2.66667	1.33333	186.6667	2.66667	14.66667	13.33333	3	160	enriched

Source of Data: StatLib. (1993). *Serial correlation or cereal correlation??* [Data file]. Retrieved from http://lib.stat.cmu.edu/datasets/1993.expo/

Inferential Statistics
Simple Linear Regression
SAS Code

```
      DATA US_CEREAL;
(1)         INFILE "C:\UScereal.CSV" FIRSTOBS=2 DSD MISSOVER;
(2)         INPUT CEREAL $ MFR $ CALORIES PROTEIN FAT SODIUM
                  FIBER CARBO SUGARS SHELF POTASSIUM VITAMINS $;
(3)         KEEP CALORIES CARBO;
      RUN;

      PROC PRINT DATA=US_CEREAL;
      RUN;

      PROC CORR DATA=US_CEREAL PLOTS=SCATTER(ELLIPSE=NONE);
      RUN;

(4)   PROC GLM DATA=US_CEREAL PLOTS=(DIAGNOSTICS RESIDUALS);
(5)         MODEL CALORIES=CARBO / CLPARM;
            TITLE "Simple Linear Regression: US Cereal";
      RUN;

(6)   PROC REG DATA=US_CEREAL
(7)               PLOTS(LABEL)=(COOKSD RSTUDENTBYLEVERAGE DFFITS DFBETAS);
(8)         MODEL CALORIES=CARBO;
            TITLE "Simple Linear Regression: US Cereal";
      RUN;

      TITLE;
      QUIT;
```

Note: There are two primary ways of performing a simple linear regression in SAS: PROC GLM and PROC REG. Both methods return the exact same results, so you can use either procedure. However, there are some supplemental results provided by PROC REG that are not provided by PROC GLM (e.g., adjusted R-square) and vice versa, so you may want to run both simultaneously so that you have everything you need. Both procedures are modeled here.

(1) The INFILE statement is used to identify the file path and file name of the raw data file. The optional statement FIRSTOBS=2 tells SAS that the first observation is located on Row 2 (in this case, because Row 1 of the data file contains the variable names). The DSD (delimiter-separated data) option performs two functions here. First, it tells SAS that the data is separated by commas (i.e. it is a .CSV file). Second, if two consecutive commas are found, it forces SAS to treat that as missing data. If the end of an observation (row) is blank, there will not be two consecutive commas, but you still need to treat it as missing data; the MISSOVER option is used to accomplish this.

(2) In the same manner as when you manually entered data, the INPUT statement creates the variable names and assigns the order of the variables to the new dataset. This command also assigns a variable type to each new variable. The default variable type is *numeric*. If a *character* variable is being created, put a dollar sign ($) in back of it.

(3) You must read in every variable in order. However, for the purpose of this analysis, we only need to KEEP two of the variables: CALORIES and CARBO.

(4) **This is the PROC GLM method of simple linear regression.**

(5) The MODEL statement is written as Dependent Variable (DV) = Independent Variable (IV). The CLPARM option requests the confidence limits (confidence interval) for the parameter estimates.

(6) **This is the PROC REG method of simple linear regression.**

(7) Both PROC REG and PROC GLM include diagnostic plots that help determine if outliers are exerting undue influence on the regression analysis. However, only PROC REG offers the LABEL option. This option requests that the observation numbers of influential observations be included in the graphics, making it much easier to determine which observations (if any) are impacting the results.

(8) The MODEL statement is written as Dependent Variable (DV) = Independent Variable (IV).

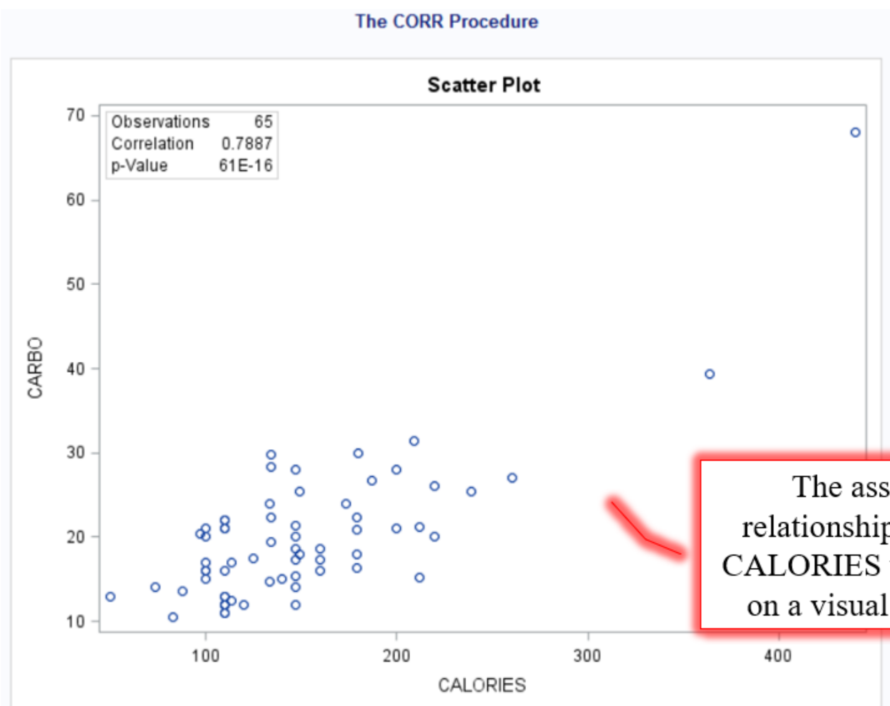
Inferential Statistics
Simple Linear Regression
Selected Output

The CORR Procedure

2 Variables:	CALORIES CARBO

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
CALORIES	65	149.40826	62.41187	9712	50.00000	440.00000
CARBO	65	19.96762	8.46847	1298	10.52632	68.00000

Pearson Correlation Coefficients, N = 65 Prob > |r| under H0: Rho=0		
	CALORIES	CARBO
CALORIES	1.00000	0.78872 <.0001
CARBO	0.78872 <.0001	1.00000



The assumption of a linear relationship between CARBO and CALORIES was found tenable based on a visual inspection of the data.

PROC GLM Output

Simple Linear Regression: US Cereal

The GLM Procedure

Dependent Variable: CALORIES

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	155082.6063	155082.6063	103.70	<.0001
Error	63	94212.8880	1495.4427		
Corrected Total	64	249295.4943			

$R^2 = .622$

R-Square	Coeff Var	Root MSE	CALORIES Mean
0.622083	25.88274	38.67095	149.4083

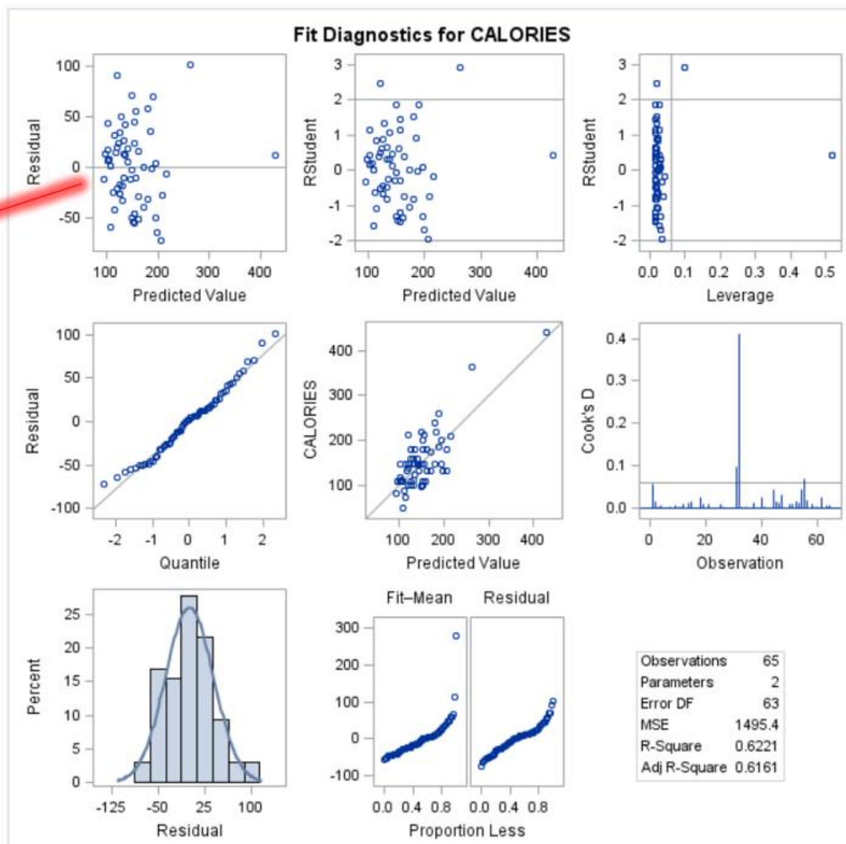
Source	DF	Type I SS	Mean Square	F Value	Pr > F
CARBO	1	155082.6063	155082.6063	103.70	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
CARBO	1	155082.6063	155082.6063	103.70	<.0001

There is a significant effect of CARBO on CALORIES, $F(1, 63) = 103.70$, $p < .001$, $R^2 = .622$.

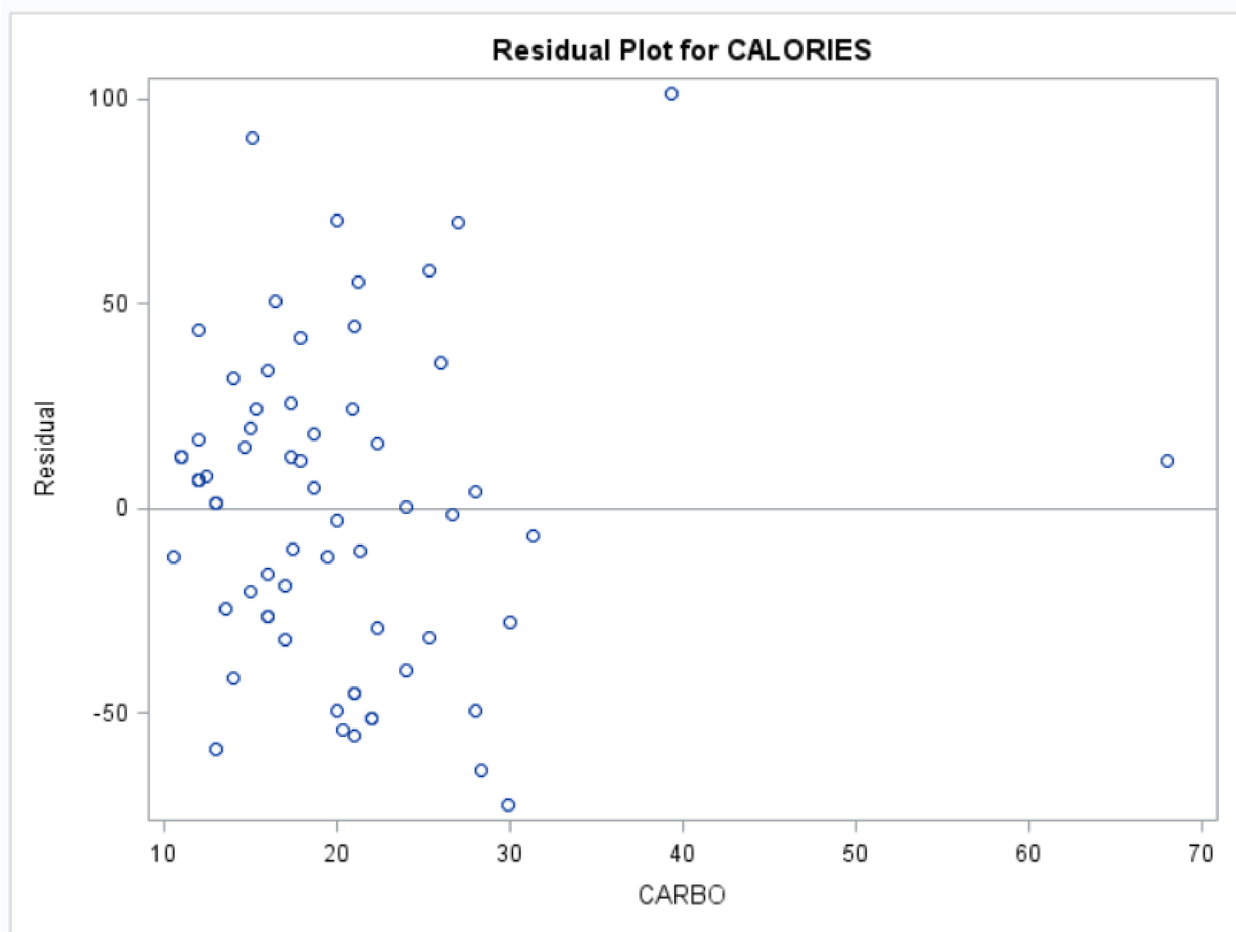
Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	33.34012848	12.36583350	2.70	0.0090	8.62898206	58.05127489
CARBO	5.81281755	0.57080797	10.18	<.0001	4.67214884	6.95348627

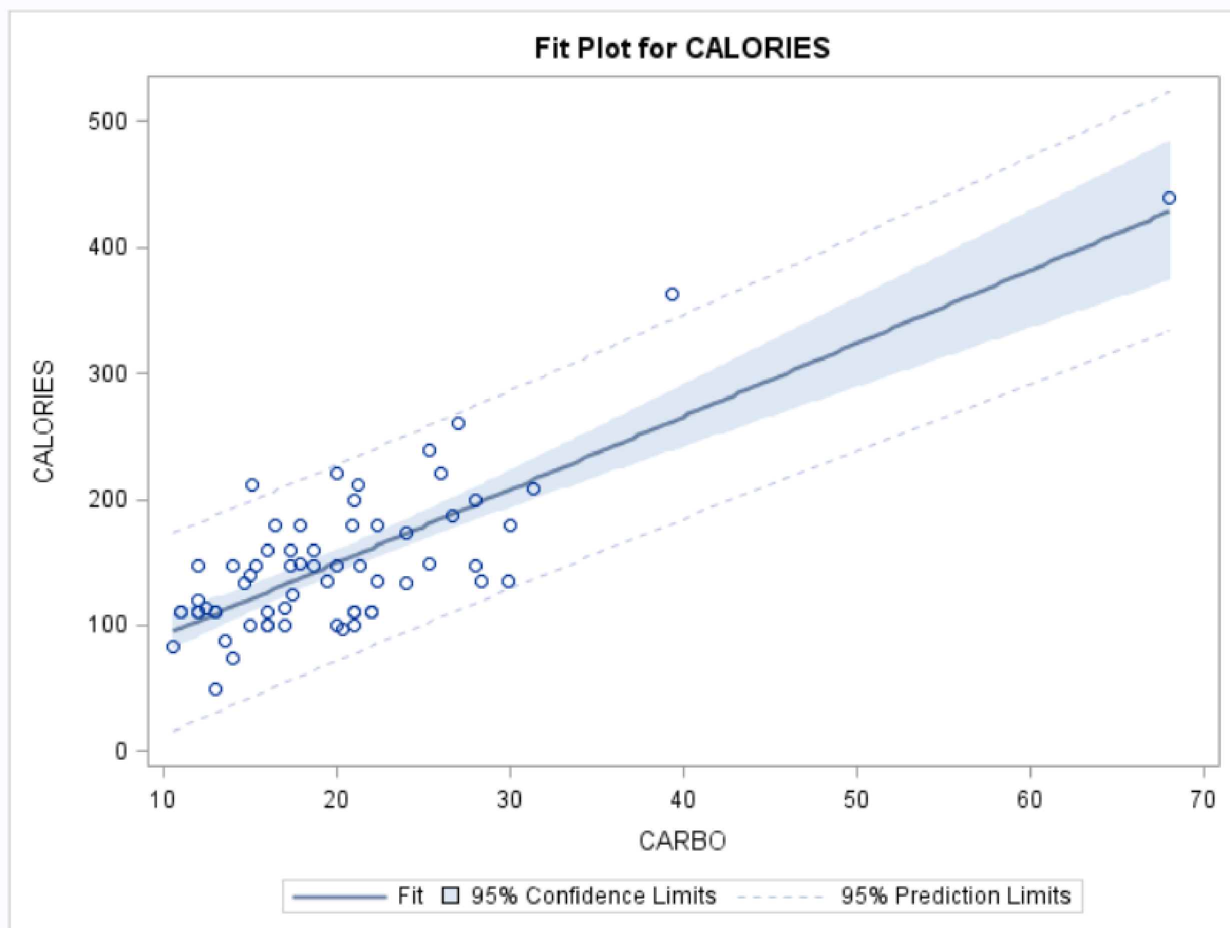
$$\widehat{\text{CALORIES}} = 33.340 + 5.813\text{CARBO}$$



The data appear to be randomly scattered about the reference line.

The assumption of homoscedasticity was found tenable based on a visual inspection of the data.


Residual Plot for CALORIES
100
50
0
-50
Residual
10
20
30
40
50
60
70
CARBO



PROC REG Output

Simple Linear Regression: US Cereal

The REG Procedure
Model: MODEL1
Dependent Variable: CALORIES

Number of Observations Read	65
Number of Observations Used	65

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	155083	155083	103.70	<.0001
Error	63	94213	1495.44267		
Corrected Total	64	249295			

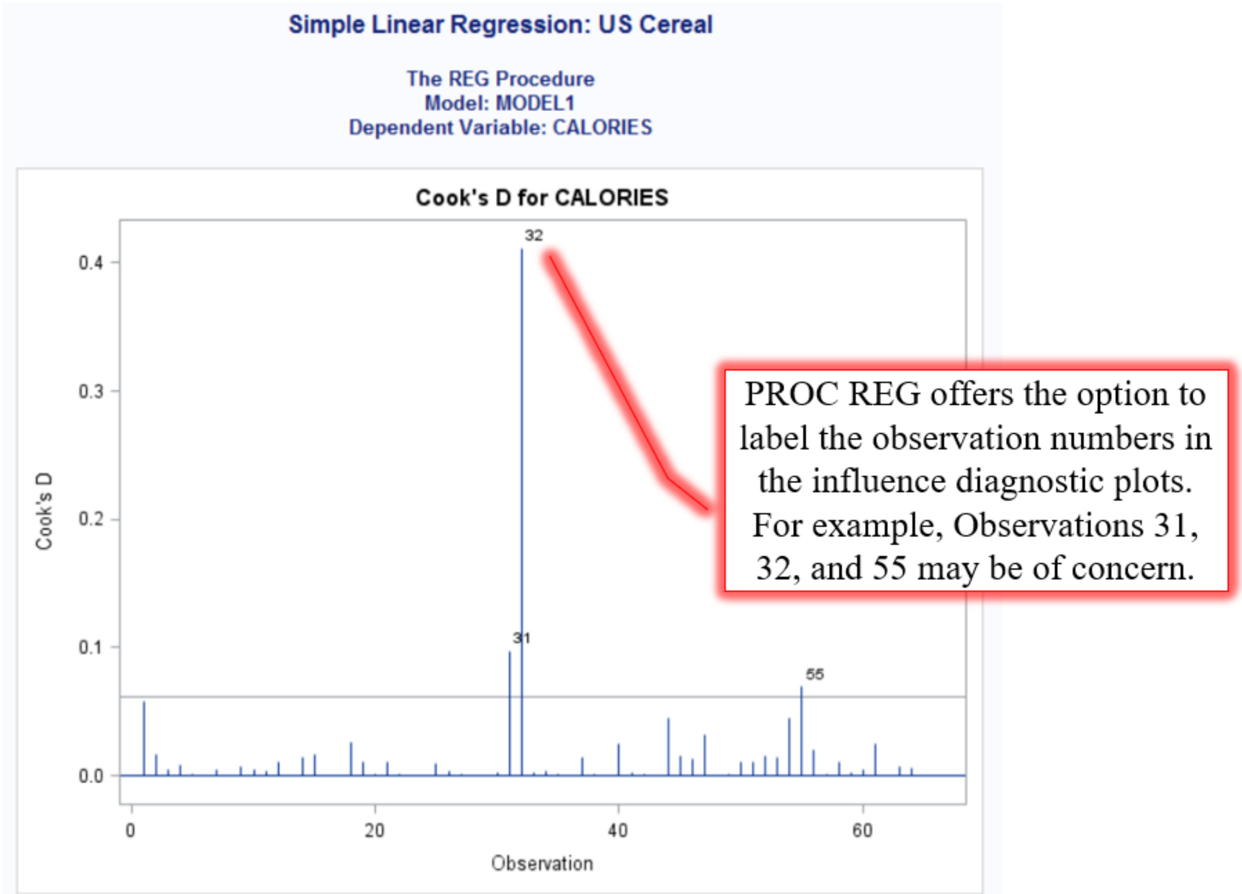
There is a significant effect of CARBO on CALORIES, $F(1, 63) = 103.70$, $p < .001$, $R^2 = .622$, adjusted $R^2 = .616$.

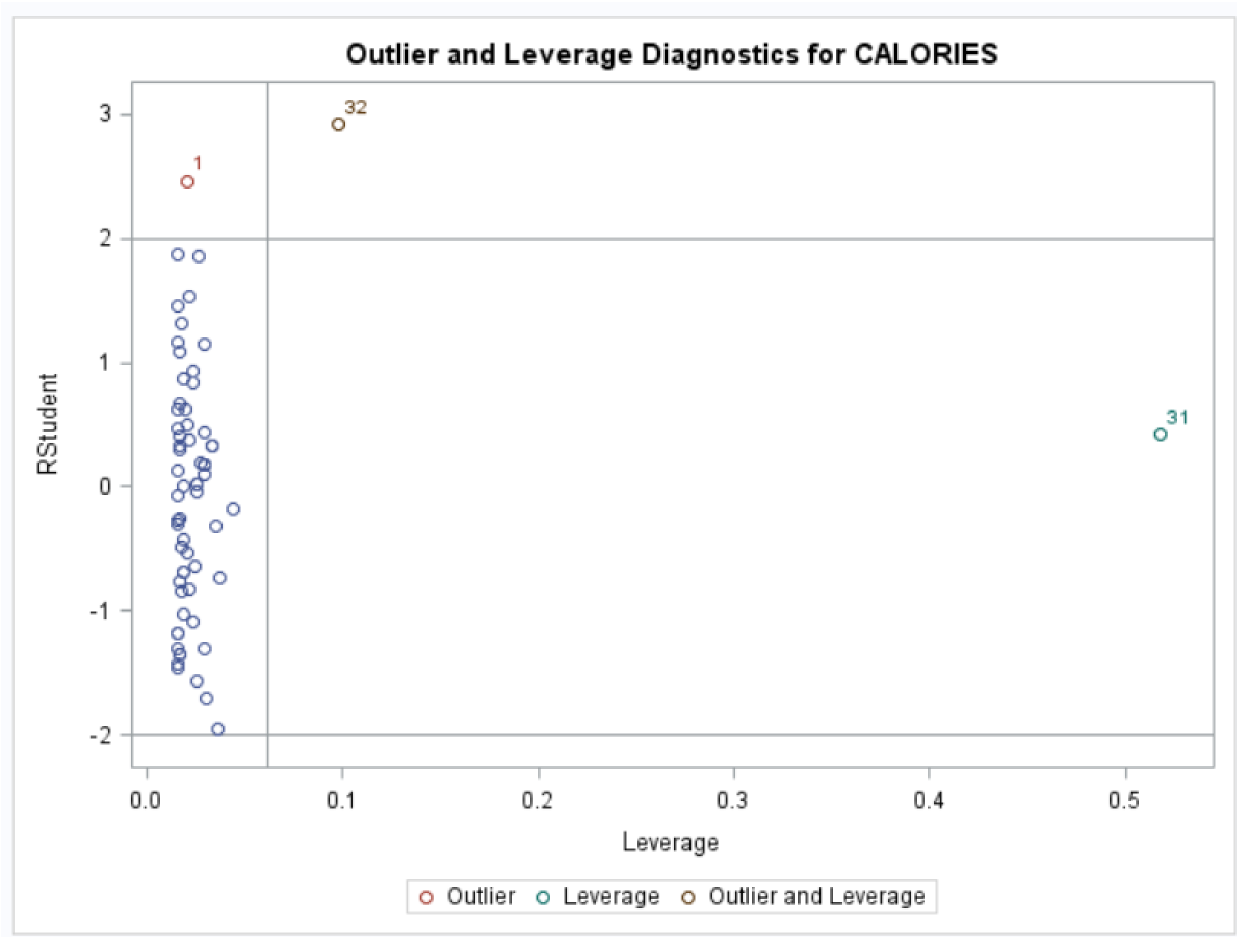
Root MSE	38.67095	R-Square	0.6221
Dependent Mean	149.40826	Adj R-Sq	0.6161
Coeff Var	25.88274		

$R^2 = .622$
Adjusted $R^2 = .616$

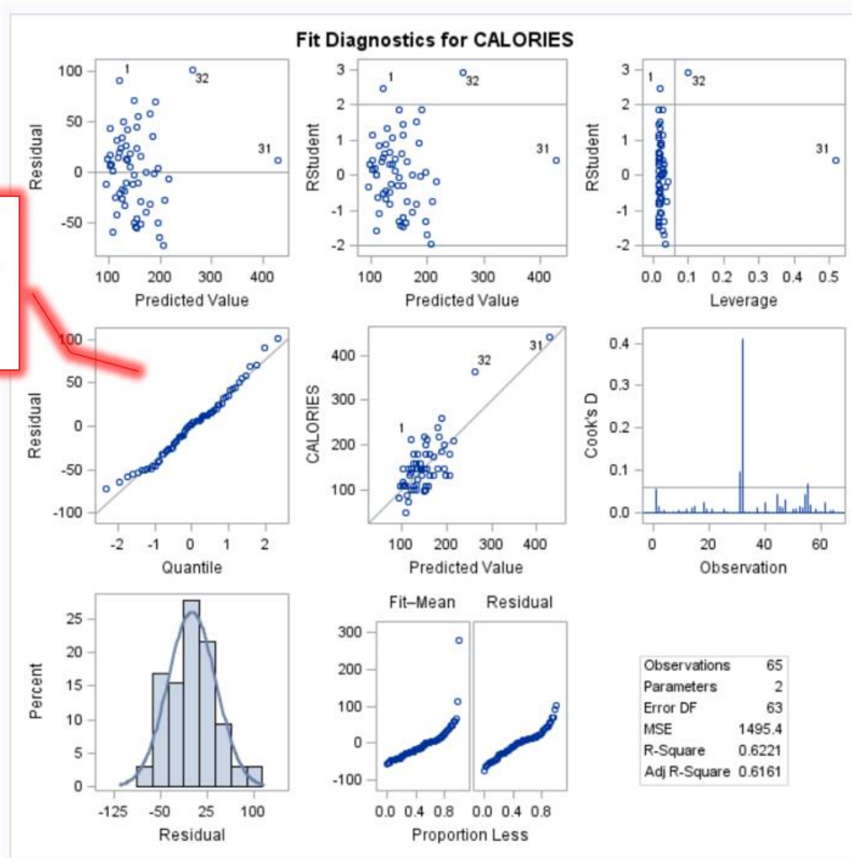
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	33.34013	12.36583	2.70	0.0090
CARBO	1	5.81282	0.57081	10.18	<.0001

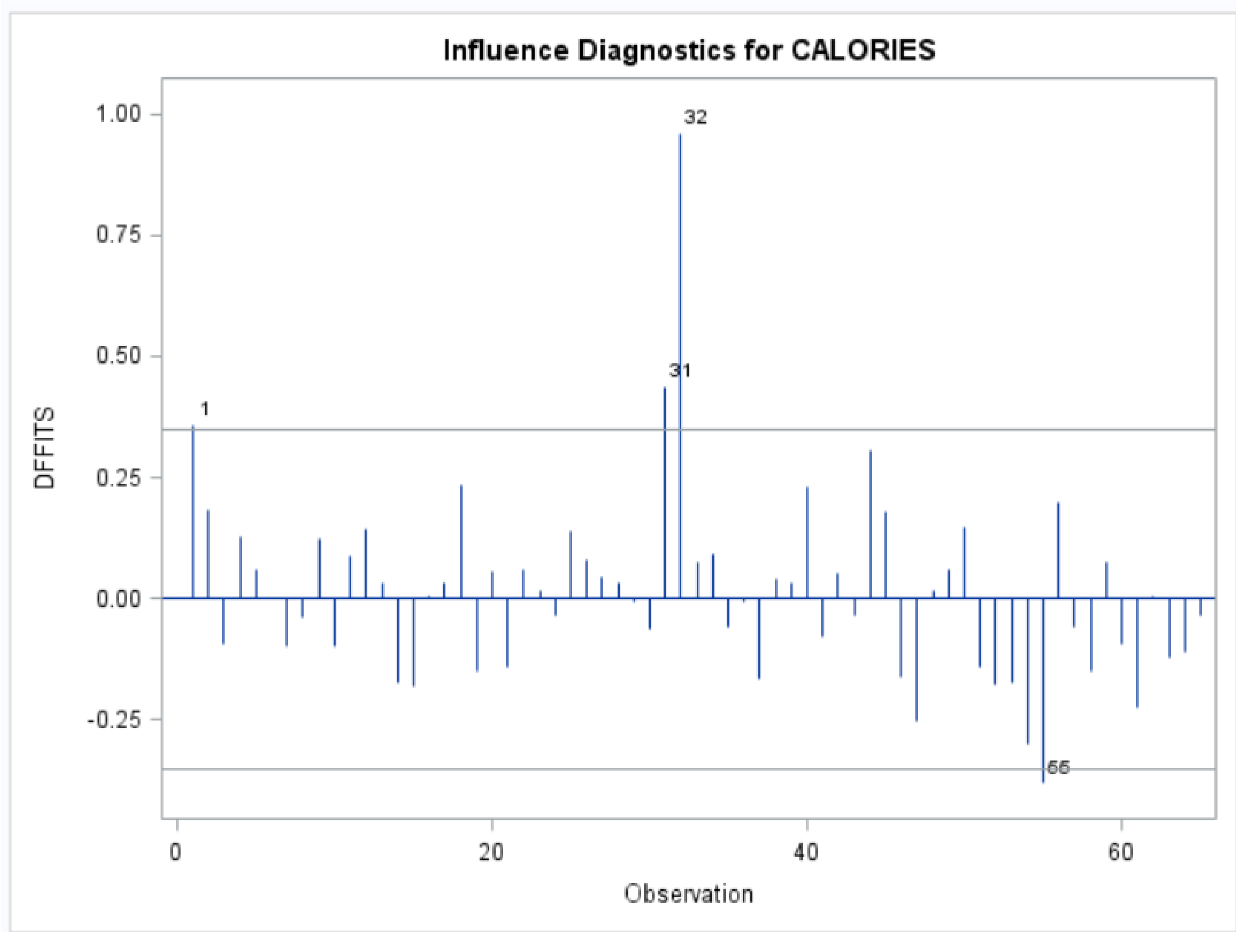
$\widehat{\text{CALORIES}} = 33.340 + 5.813\text{CARBO}$

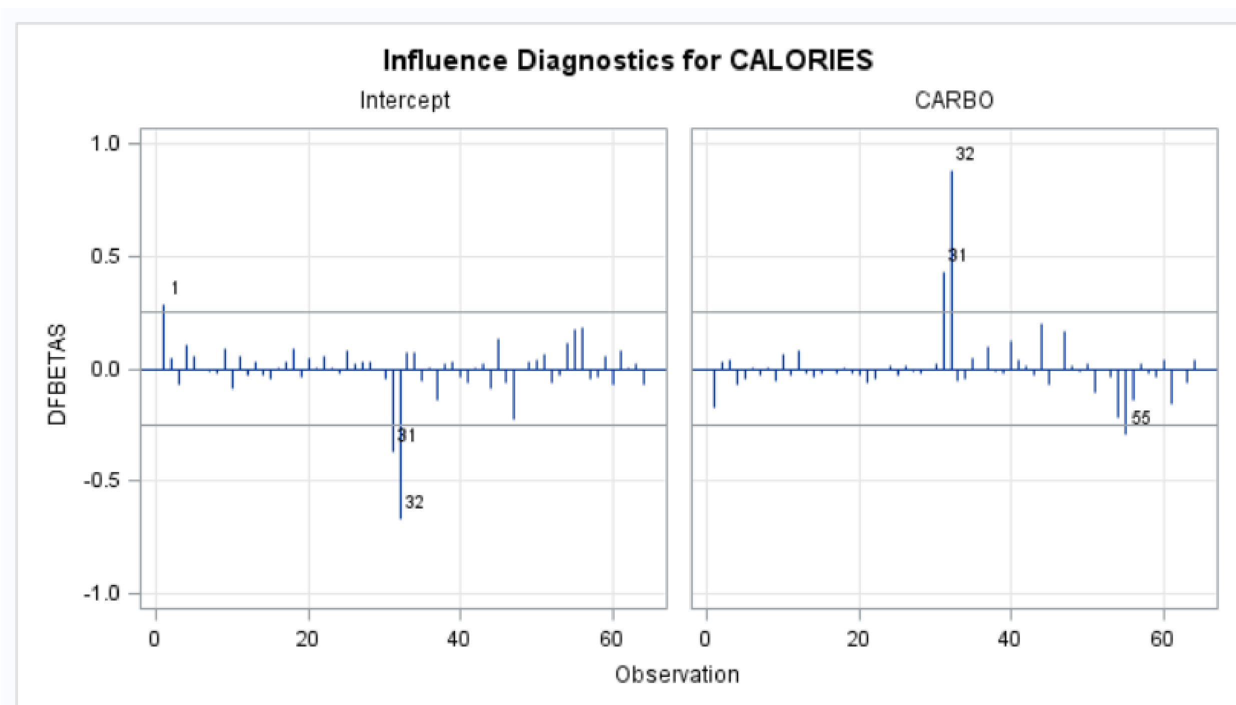

Simple Linear Regression: US Cereal
The REG Procedure
Model: MODEL1
Dependent Variable: CALORIES
Cook's D for CALORIES
32
31
55
0.4
0.3
0.2
0.1
0.0
Cook's D
0
20
40
60
Observation
PROC REG offers the option to label the observation numbers in the influence diagnostic plots. For example, Observations 31, 32, and 55 may be of concern.

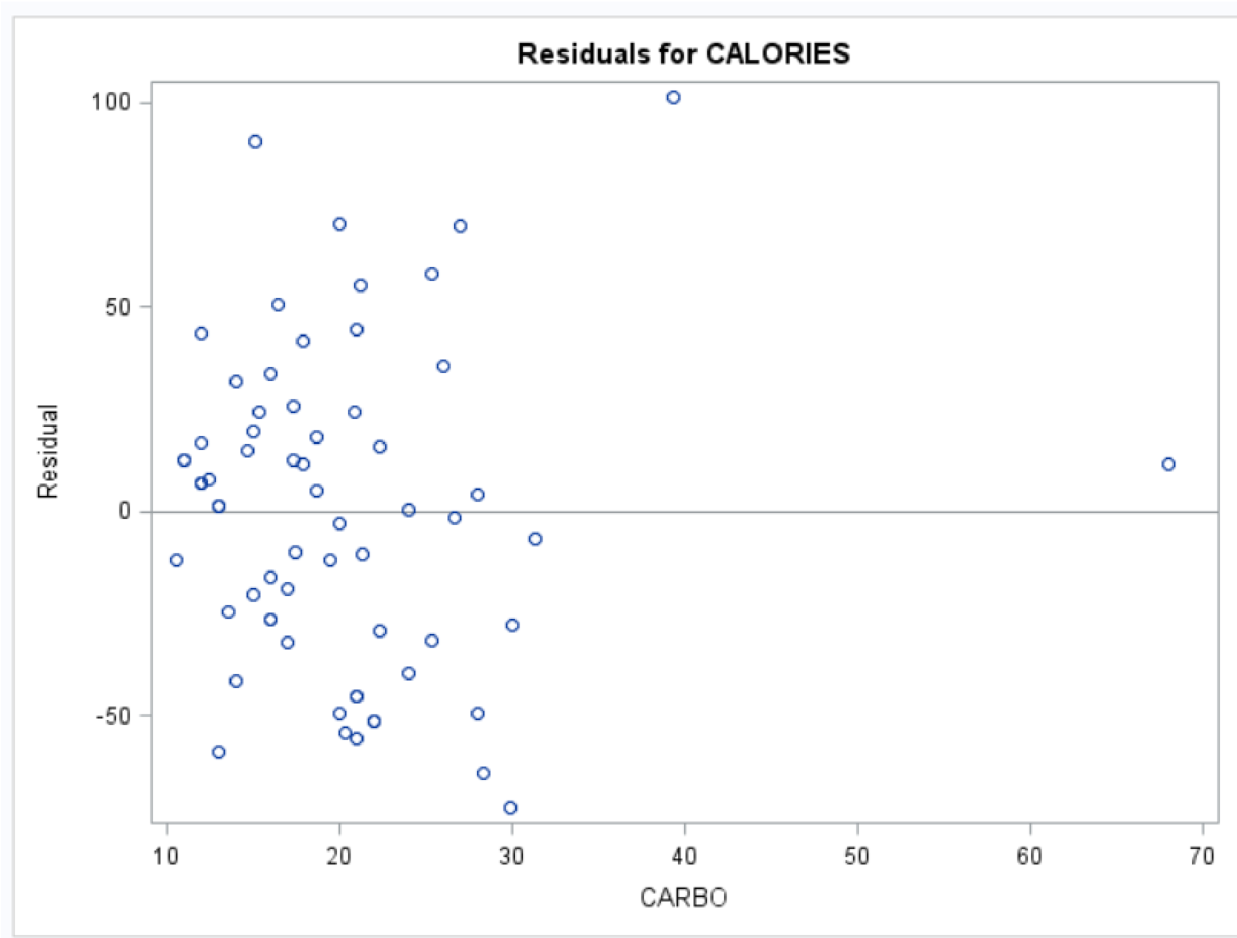

Outlier and Leverage Diagnostics for CALORIES
RStudent
Leverage
32
1
31
3
2
1
0
-1
-2
0.0
0.1
0.2
0.3
0.4
0.5
Outlier
Leverage
Outlier and Leverage

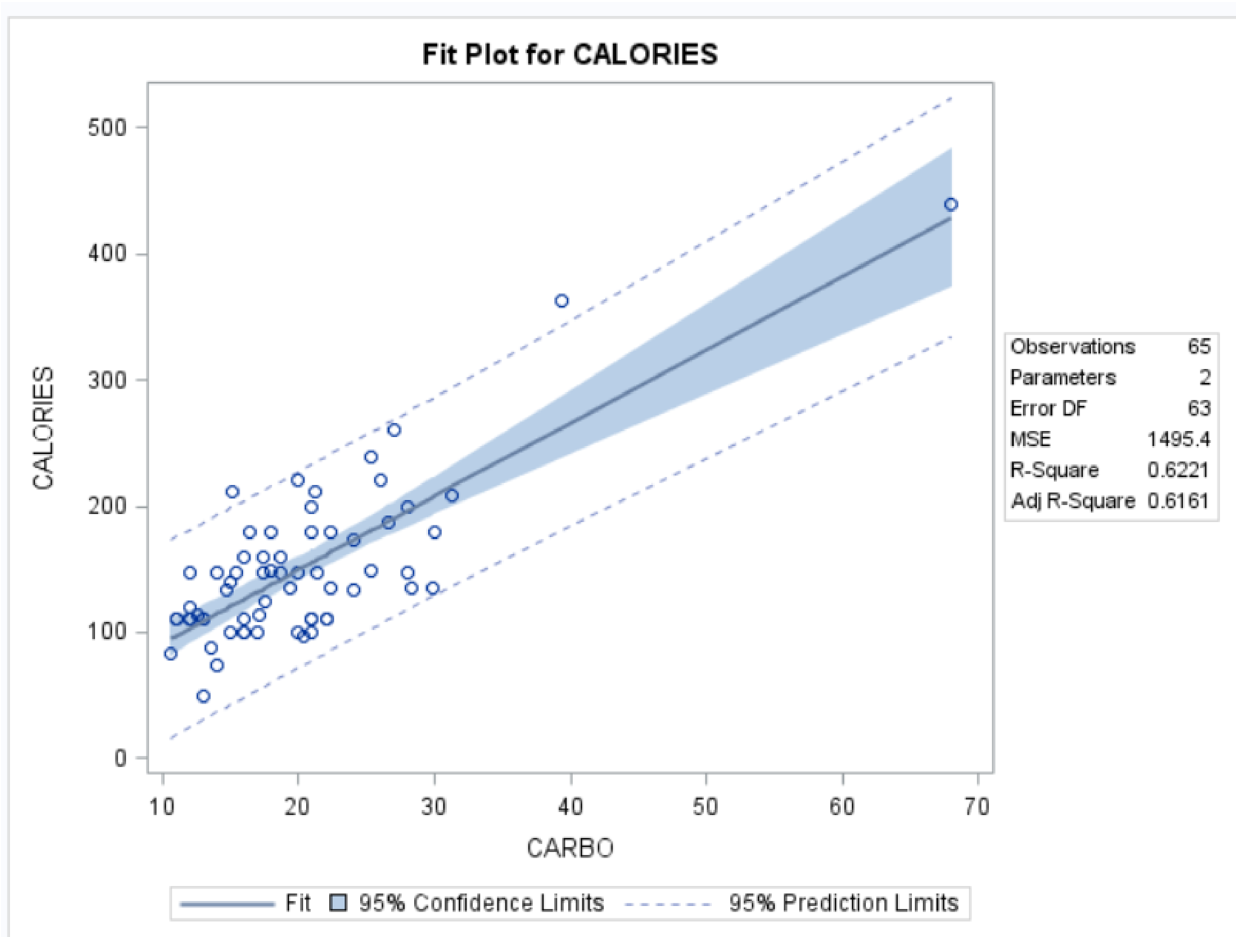
Based on a visual inspection of the data, the assumption of normality was found to be tenable.




Influence Diagnostics for CALORIES
DFFITS
1.00
0.75
0.50
0.25
0.00
-0.25
32
31
1
55
0
20
40
60
Observation


Influence Diagnostics for CALORIES
Intercept
CARBO
DFBETAS
1.0
0.5
0.0
-0.5
-1.0
1
31
32
55
0
20
40
60
Observation


Residuals for CALORIES
Residual
100
50
0
-50
10
20
30
40
50
60
70
CARBO


Fit Plot for CALORIES
CALORIES
0
100
200
300
400
500
10
20
30
40
50
60
70
CARBO
Observations 65
Parameters 2
Error DF 63
MSE 1495.4
R-Square 0.6221
Adj R-Square 0.6161
Fit
95% Confidence Limits
95% Prediction Limits

Inferential Statistics
Multiple Regression
Research Scenario

A botanist decides to test whether temperature and rainfall are good predictors of plant height.

This analysis will use the "Plant_Height.csv" file that can be found (A) as a companion with this SAS guide or (B) at the website below by scrolling down to the "Running the analysis" section and clicking on the download link. Please be aware that, if you choose to download the file from the website, you will need to complete the following "cleaning" of the dataset.

1. Remove all commas from Column I ("Site"). You will be reading this file as a .CSV (comma-separated value) file. These commas in the data would cause SAS to act as though it was reading data for the variable in Column J when it was still reading Column I.
2. Change all "NA" values to blanks or periods in Columns AG and AH. This is because Columns AG and AH are numeric data. In SAS, a period is used to represent missing numeric data. ("NA" would be considered character data.)
3. Column B is named "site" (lowercase) and Column I is named "Site" (with a capital). You do not have to change anything, but be aware that the SAS code herein is written such that the variable in Column I will be called SITE_NAME to differentiate the two variables.

If you use the "Plant_Height.csv" file that can be found as a companion with this SAS guide, this cleaning has already been completed for you.

If you prefer to enter the data manually, instead of reading in the data file, the raw data for the variables of interest may be found in the Appendix of this guide.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	sort_num	site	Genus_sp	Family	growthfor	height	loght	Country	Site	lat	long	entered.b	alt	temp	diu
2	1402	193	Acer_mac	Sapindace	Tree	28	1.447158	USA	Oregon - I	44.6	-123.334	Angela	179	10.8	
3	25246	103	Quararibe	Malvaceae	Tree	26.6	1.424882	Peru	Manu	12.183	-70.55	Angela	386	24.5	
4	11648	54	Eragrostis	Poaceae	Herb	0.3	-0.52288	Australia	Central Au	23.8	133.833	Michelle	553	20.9	
5	8168	144	Cistus_sal	Cistaceae	Shrub	1.6	0.20412	Israel	Hanadiv	32.555	34.938	Angela	115	19.9	
6	22422	178	Phlox_bifi	Polemoni	Herb	0.2	-0.69897	USA	Indiana Du	41.617	-86.95	Michelle	200	9.7	
7	15925	59	Homaliun	Salicaceae	Shrub	1.7	0.230449	New Cale	NA	21.5	165.5	Laura	95	22.6	
8	25151	27	Pultenaea	Fabaceae	Shrub	0.5	-0.30103	Australia	Kuringai C	33.65	151.2	Michelle	157	16.8	
9	26007	118	Rhizophor	Rhizophor	Tree	10	1	NA	Marshall I	9	168	Laura	2	27.7	
10	6597	154	Carya_ova	Juglandac	Tree	40	1.60206	USA	Colorado	35.8	-89.9	Angela	71	15.5	
11	16908	106	Ischaemu	Poaceae	Herb	0.5	-0.30103	Australia	Christmas	10.417	105.667	Laura	2	26.4	
12	4610	201	Betula_na	Betulacea	Shrub	0.55	-0.25964	Estonia	NA	58.5	25	Angela	28	5.4	
13	1593	86	Acmena_g	Myrtaceae	Tree	32	1.50515	Australia	Cairns - Da	16.103	145.446	Angela	263	25.2	
14	22359	69	Phaleria_i	Thymelae	Tree	5	0.69897	Fiji	Viti Levu	17.8	178	Laura	1108	19.3	

Source of Data: Letten, A. (2016). *Linear regression*. Retrieved from http://environmentalcomputing.net/linear-regression/

Inferential Statistics
Multiple Regression
SAS Code

```
     DATA PLANT_HEIGHT;
(1)       INFILE "C:\Plant_Height.CSV" FIRSTOBS=2 DSD MISSOVER;
(2)       INPUT SORT_NUMBER SITE GENUS_SPECIES $ FAMILY $ GROWTHFORM $
                HEIGHT LOGHT COUNTRY $ SITE_NAME $ LAT LONG ENTERED_BY $
                ALT TEMP DIURN_TEMP ISOTHERM TEMP_SEAS TEMP_MAX_WARM
                TEMP_MIN_COLD TEMP_ANN_RANGE TEMP_MEAN_WETQR
                TEMP_MEAN_DRYQR TEMP_MEAN_WARMQR TEMP_MEAN_COLDQR RAIN
                RAIN_WETM RAIN_DRYM RAIN_SEAS RAIN_WETQR RAIN_DRYQR
                RAIN_WARMQR RAIN_COLDQR LAI NPP HEMISPHERE;
(3)       KEEP HEIGHT LOGHT TEMP RAIN;
     RUN;

     PROC PRINT DATA=PLANT_HEIGHT;
     RUN;

     PROC CORR DATA=PLANT_HEIGHT PLOTS=SCATTER(ELLIPSE=NONE);
     RUN;

(4)  DATA PLANT_HEIGHT;
(5)       SET PLANT_HEIGHT;
(6)       DROP HEIGHT;
     RUN;

     PROC PRINT DATA=PLANT_HEIGHT;
     RUN;

     PROC CORR DATA=PLANT_HEIGHT PLOTS=SCATTER(ELLIPSE=NONE);
     RUN;

(7)  PROC GLM DATA=PLANT_HEIGHT PLOTS=(DIAGNOSTICS RESIDUALS);
(8)       MODEL LOGHT=TEMP RAIN / CLPARM;
(9)       OUTPUT OUT=INFLUENCE_RESULTS
(10)                  PREDICTED=Y_PRED
                      RESIDUAL=RESID
                      STUDENT=SRESID
                      RSTUDENT=SDRESID
                      H=HAT_H
                      COOKD=COOKS_D
                      DFFITS=DFFITS
                      LCLM=LCL_MEAN
                      UCLM=UCL_MEAN
                      LCL=LCL_INDIVID
                      UCL=UCL_INDIVID
                      PRESS=PRESS;
          TITLE "Multiple Regression: Plant Height";
     RUN;

     TITLE;

(11) PROC PRINT DATA=INFLUENCE_RESULTS;
```

```
      RUN;

(12)  PROC CORR DATA=INFLUENCE_RESULTS;
(13)        VAR TEMP RAIN RESID;
      RUN;

(14)  PROC EXPORT DATA=INFLUENCE_RESULTS
(15)        OUTFILE="C:\temp\PlantHeight_Influence.CSV"
(16)        REPLACE;
      RUN;

(17)  PROC REG DATA=PLANT_HEIGHT
(18)              PLOTS(LABEL)=(COOKSD RSTUDENTBYLEVERAGE DFFITS DFBETAS);
(19)        MODEL LOGHT=TEMP RAIN / INFLUENCE TOL VIF PARTIAL;
            TITLE "Multiple Regression: Plant Height";
      RUN;

      TITLE;

(20)  DATA PLANT_HEIGHT_2;
(21)        SET PLANT_HEIGHT;
(22)        ID = _N_;
(23)        IF ID=22 OR ID=102 OR ID=146 THEN DELETE;
      RUN;

(24)  PROC PRINT DATA=PLANT_HEIGHT_2;
      RUN;

      TITLE;
      QUIT;
```

Note 1: There are two primary ways of performing a simple linear regression in SAS: PROC GLM and PROC REG. Both methods return the exact same results, so you can use either procedure. However, there are some supplemental results provided by PROC REG that are not provided by PROC GLM (e.g., adjusted R-square) and vice versa, so you may just want to run both simultaneously so that you have everything you need. Both procedures are modeled here.

Note 2: The code for multiple regression is EXACTLY the same as the code for simple linear regression except that there is at least one more IV in the MODEL statement [Lines (8) and (19) of this example]. The multiple regression code presented here is comprehensive. It includes additional (optional) code for conducting diagnostic analyses, predictive analyses, etc., topics which you will learn throughout EPRS 8550. If you are in the first few weeks of the course and are learning multiple regression for the first time, you are welcome to use the simple linear regression code from the previous section of this SAS guide. All you need to do is make sure you include all of your IVs in the PROC GLM and PROC REG MODEL statements. ☺

(1) The INFILE statement is used to identify the file path and file name of the raw data file. The optional statement FIRSTOBS=2 tells SAS that the first observation is located on Row 2 (in this case, because Row 1 of the data file contains the variable names). The DSD (delimiter-separated data) option performs two functions here. First, it tells SAS that

the data is separated by commas (i.e. it is a .CSV file). Second, if two consecutive commas are found, it forces SAS to treat that as missing data. If the end of an observation (row) is blank, there will not be two consecutive commas, but you still need to treat it as missing data; the MISSOVER option is used to accomplish this.

(2) In the same manner as when you manually entered data, the INPUT statement creates the variable names and assigns the order of the variables to the new dataset. This command also assigns a variable type to each new variable. The default variable type is *numeric*. If a *character* variable is being created, put a dollar sign ($) in back of it.

(3) You must read in every variable in order. However, for the purpose of this analysis, we only need to KEEP four of the variables: HEIGHT, LOGHT, TEMP, and RAIN.

(4) From the output shown below, you will see that it is prudent to drop HEIGHT from the model. This DATA step is used to rewrite the PLANT_HEIGHT dataset without HEIGHT.

(5) The data from the previous PLANT_HEIGHT dataset is copied to the new PLANT_HEIGHT dataset using the SET statement.

(6) The DROP statement is used to drop HEIGHT from the dataset.

(7) **This is the PROC GLM method of simple linear regression.**

(8) The MODEL statement is written as Dependent Variable (DV) = Independent Variable (IV). The CLPARM option requests the confidence limits (confidence interval) for the parameter estimates.

(9) PROC GLM offers a wide variety of influence statistics, some of which are not available in PROC REG. However, it will not produce influence statistics in the output window. Instead, you must request that SAS create an output dataset by using the OUTPUT statement. Next, use OUT= to assign that dataset a name; in this case, the name of the new dataset is INFLUENCE_RESULTS.

(10) Beginning with Line (10), you need to identify every influence statistic you want included in the new dataset; this will be written to the left of the equal sign. Then, you need to assign each statistic the name you want it to have in your new dataset; this will be written to the right of the equal sign. For example, in Line (10), the PREDICTED statistic is being requested; it will be called Y_PRED in the new dataset. (*Note: You can use the same name on the left and right of the equal sign if you want.*) All of the influence statistics included in this code will be taught in EPRS 8550. 😊

- PREDICTED: The predicted value of Y (Y')
- RESIDUAL: The residual (Y – Y')
- STUDENT: The residual divided by its standard error (studentized residuals; SRESID)
- RSTUDENT: A studentized residual with the current observation deleted (SDRESID)
- H: Leverage
- COOKD: Cook's D
- DFFITS: Standard influence of observation on predicted value
- LCLM: Lower confidence limit for the conditional mean
- UCLM: Upper confidence limit for the conditional mean
- LCL: Lower confidence limit for an individual prediction
- UCL: Upper confidence limit for an individual prediction
- PRESS: Residual for the i^{th} observation that results from dropping it and predicting it on the basis of all other observations. This is the residual divided by $(1 - h_i)$, where h_i is the leverage.

(11) PROC PRINT is used to display the new INFLUENCE_RESULTS dataset in the output window.
(12) One assumption of multiple regression is that the IVs are independent of (i.e. not correlated to) the model residuals. PROC CORR will generate the correlations to test this assumption.
(13) The VAR statement tells SAS to generate the correlations for TEMP, RAIN, and RESID (the model residuals from the INFLUENCE_RESULTS dataset).
(14) In order to find the most influential observations in a large dataset, you may need to sort your influence diagnostics. For example, you may want to sort the studentized residuals from greatest to least. You cannot do this in the output window. You may copy and paste the data from the output window into Excel. Another option is to export the data. PROC EXPORT creates a data file that you can then open and manipulate in Excel. In this case, you want to export the INFLUENCE_RESULTS data.
(15) The OUTFILE statement assigns a file path, file name, and file extension.
(16) The REPLACE statement is optional. It tells SAS to replace (overwrite) an existing file of the same name in the same location (if there is one). *It is recommended that you include this option in case you make changes to your data or your analysis, because SAS will NOT give you an error or a warning if there is an existing file and you do not replace it.* (The log will report this as a blue note, which blends in with all of the other successful log notes. You may think that you overwrote the file, when you did not.)
(17) **This is the PROC REG method of simple linear regression.**
(18) Both PROC REG and PROC GLM include diagnostic plots that help determine if outliers are exerting undue influence on the regression analysis. However, only PROC REG offers the LABEL option. This option requests that the observation numbers of influential observations be included in the graphics, making it much easier to determine which observations (if any) are impacting the results. Notice also that PROC REG has a DFBETA plot. PROC REG offers DFBETA plots and statistics; PROC GLM does not.
(19) Obtaining influence statistics in PROC REG is much simpler than in PROC GLM; all you have to do is add the INFLUENCE option to the MODEL statement. The influence statistics will appear in the output window. If you want to sort them, you can copy and paste them into Excel. The TOL option requests tolerance statistics; the VIF option requests VIF statistics. The PARTIAL option is used to generate partial regression plots (also called “added variable plots”).
(20) As you can see from the output screenshots in the next section of this guide, there are several data points that may be considered influential. Suppose you decide that it is appropriate to remove Observations #22, 102, and 146 from the analysis. In order to analyze the data without these influential observations, it is best to create a new dataset. In this DATA step, a new dataset, PLANT_HEIGHT_2, is created.
(21) The SET statement copies the data from PLANT_HEIGHT into PLANT_HEIGHT _2.
(22) A new variable, ID, is created and each observation is assigned a number from 1 to *N*.
(23) If the data is associated with ID #22, 102, or 146, it is deleted. In other words, the influential observations are being deleted from the new dataset.
(24) In this PROC PRINT, you should now see the new dataset, without those three influential data points. Now, to conduct the analysis without the influential data points, simply re-run all of the above code using PLANT_HEIGHT_2 (instead of PLANT_HEIGHT) as your data.

Note: The decision as to whether to exclude data points must be made carefully. Such a decision should be based on a thorough knowledge of the data and theory. This code is provided so that you have the option to exclude data and rerun analyses; it is not intended as a recommendation that you do so. You must use your best judgment in that regard.

Inferential Statistics
Multiple Regression
Selected Output

The CORR Procedure

4 Variables:	HEIGHT LOGHT TEMP RAIN

Simple Statistics

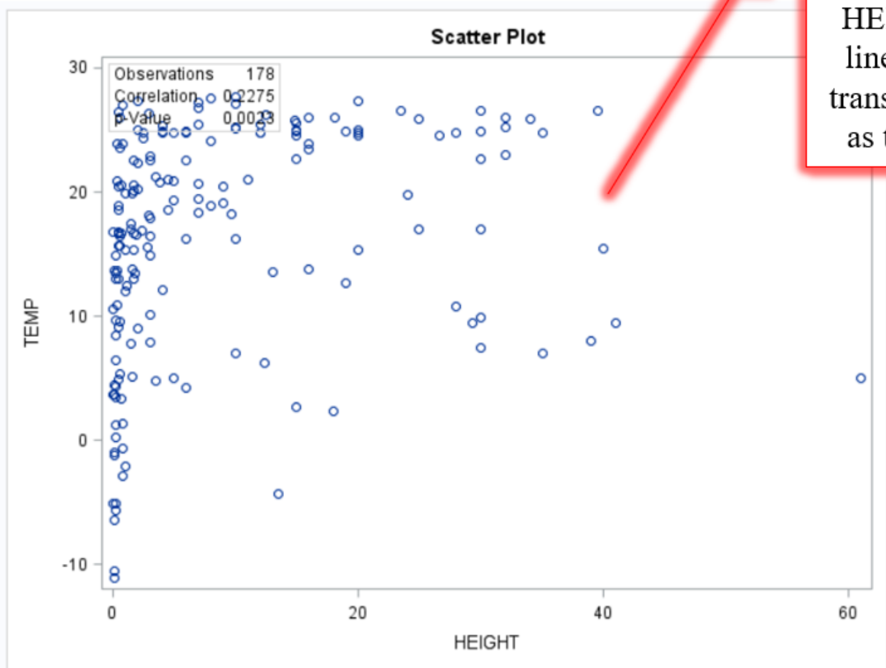
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
HEIGHT	178	8.90899	11.31967	1586	0.03220	61.00000
LOGHT	178	0.45827	0.78657	81.57152	-1.49214	1.78533
TEMP	178	16.12528	9.20288	2870	-11.10000	27.70000
RAIN	178	1344	954.78816	239257	73.00000	3991

Pearson Correlation Coefficients, N = 178
Prob > |r| under H0: Rho=0

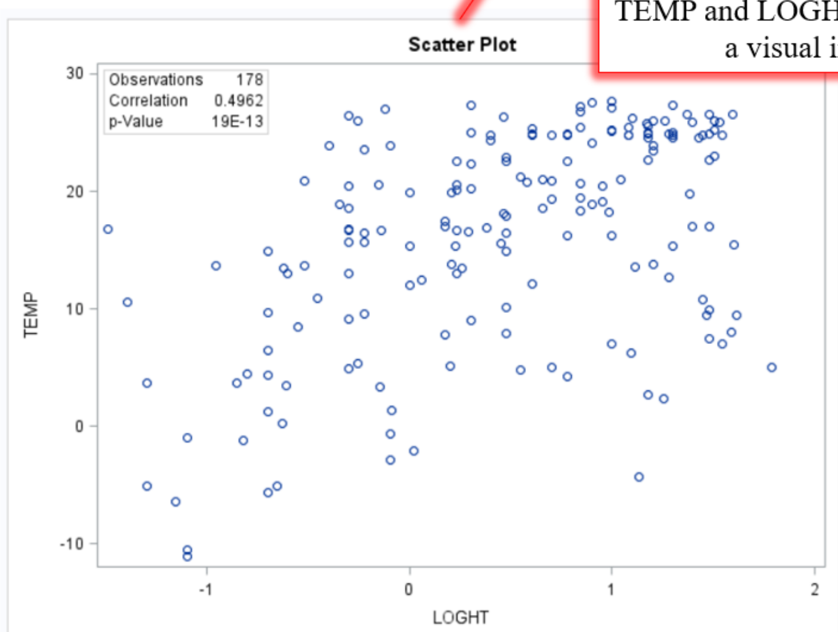
	HEIGHT	LOGHT	TEMP	RAIN
HEIGHT	1.00000	0.79848 <.0001	0.22751 0.0023	0.37415 <.0001
LOGHT	0.79848 <.0001	1.00000	0.49624 <.0001	0.48177 <.0001
TEMP	0.22751 0.0023	0.49624 <.0001	1.00000	0.55155 <.0001
RAIN	0.37415 <.0001	0.48177 <.0001	0.55155 <.0001	1.00000

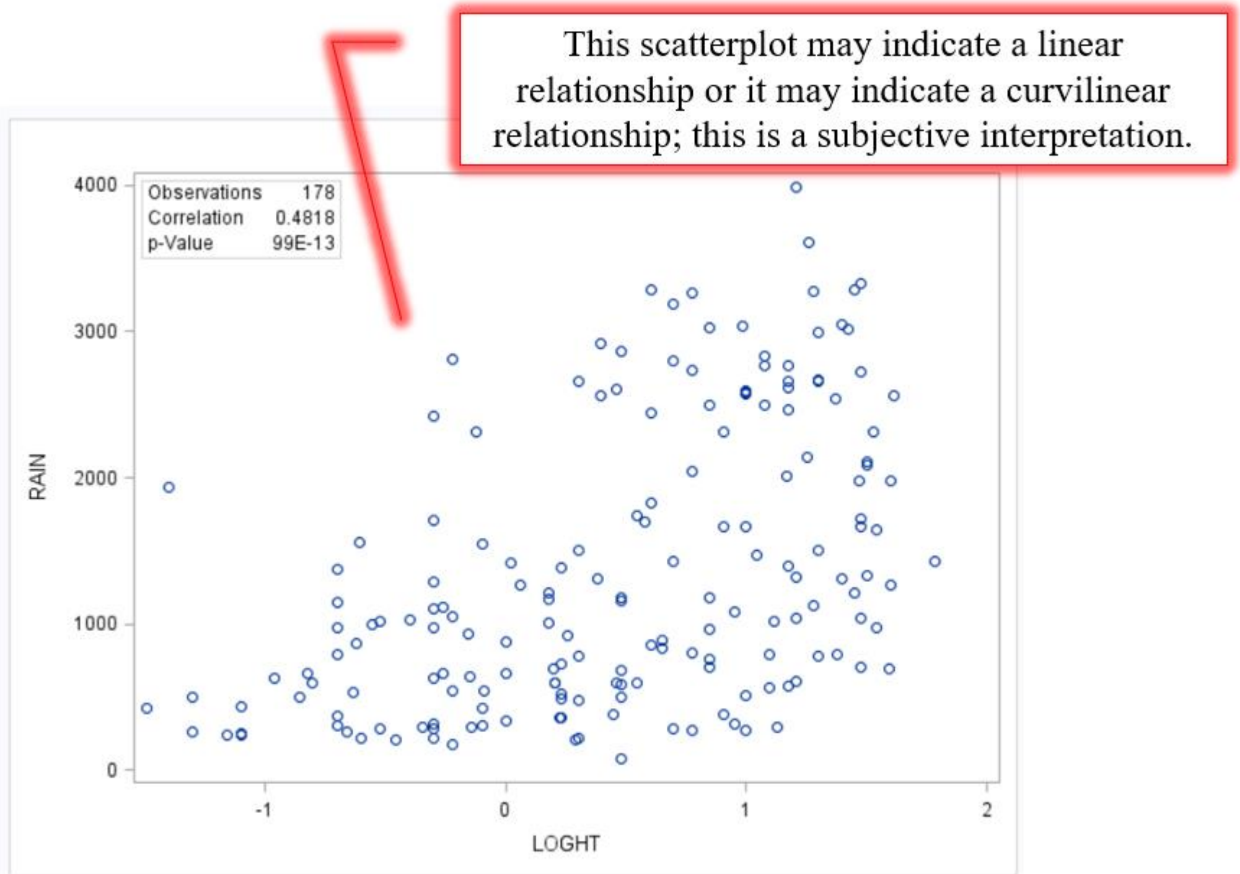
The correlation between the two IVs, TEMP and RAIN, is significant ($p < .001$); therefore, multicollinearity may be a problem in this analysis. The variance inflation factor (VIF) should be checked.

The relationship between HEIGHT and TEMP is not linear. Therefore, the log-transformed LOGHT is used as the DV in the analyses.



The assumption of a linear relationship between TEMP and LOGHT was found tenable based on a visual inspection of the data.




This scatterplot may indicate a linear relationship or it may indicate a curvilinear relationship; this is a subjective interpretation.
Observations 178
Correlation 0.4818
p-Value 99E-13
RAIN
LOGHT
4000
3000
2000
1000
0
-1
0
1
2

PROC GLM Output

The GLM Procedure

Dependent Variable: LOGHT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	33.7806895	16.8903447	39.03	<.0001
Error	175	75.7266241	0.4327236		
Corrected Total	177	109.5073136			

TEMP and RAIN account for a significant amount of the variation in LOGHT, $F(2, 175) = 39.03$, $p < .001$, $R^2 = .308$.

$R^2 = .308$

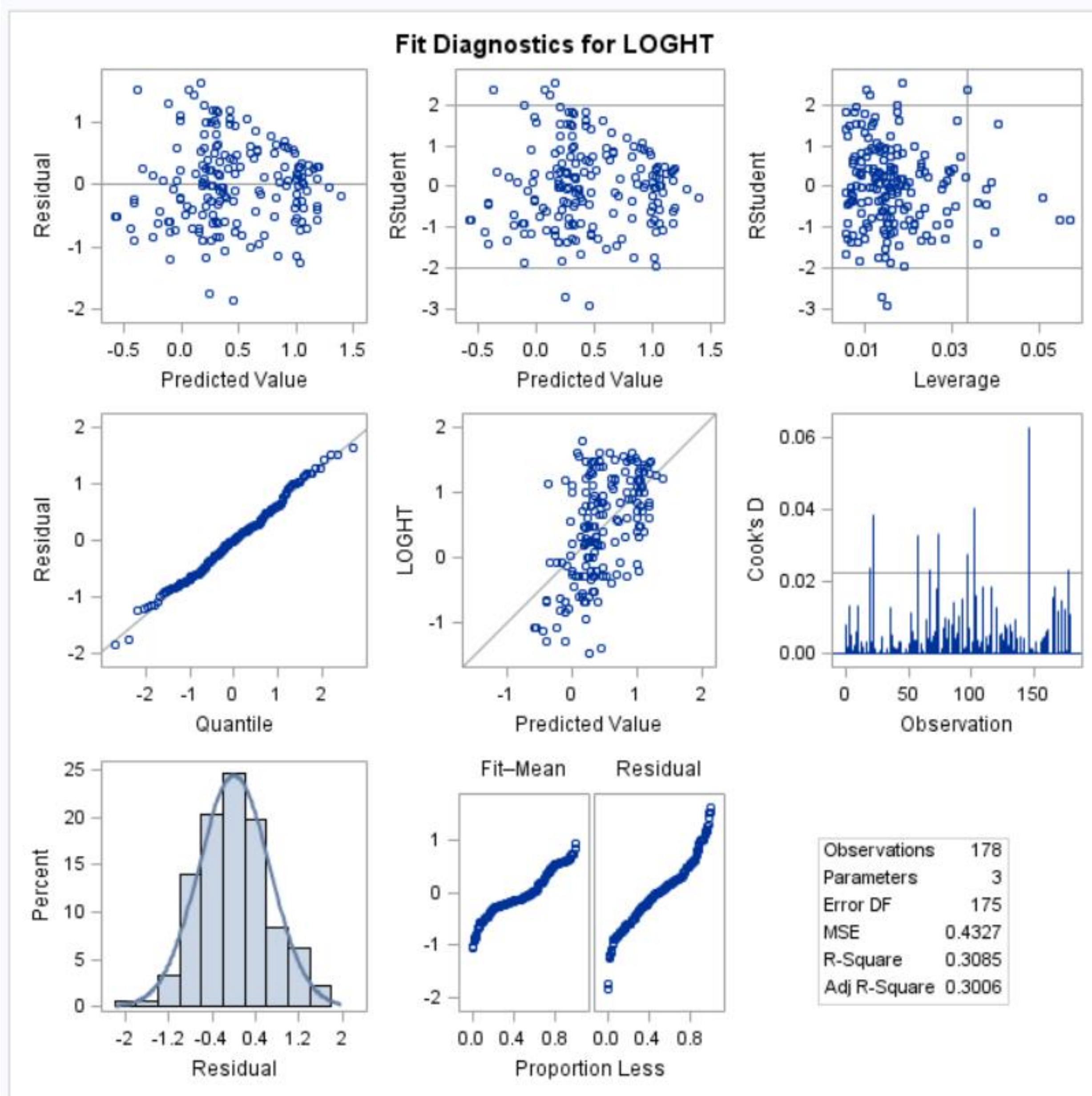
R-Square	Coeff Var	Root MSE	LOGHT Mean
0.308479	143.5446	0.657817	0.458267

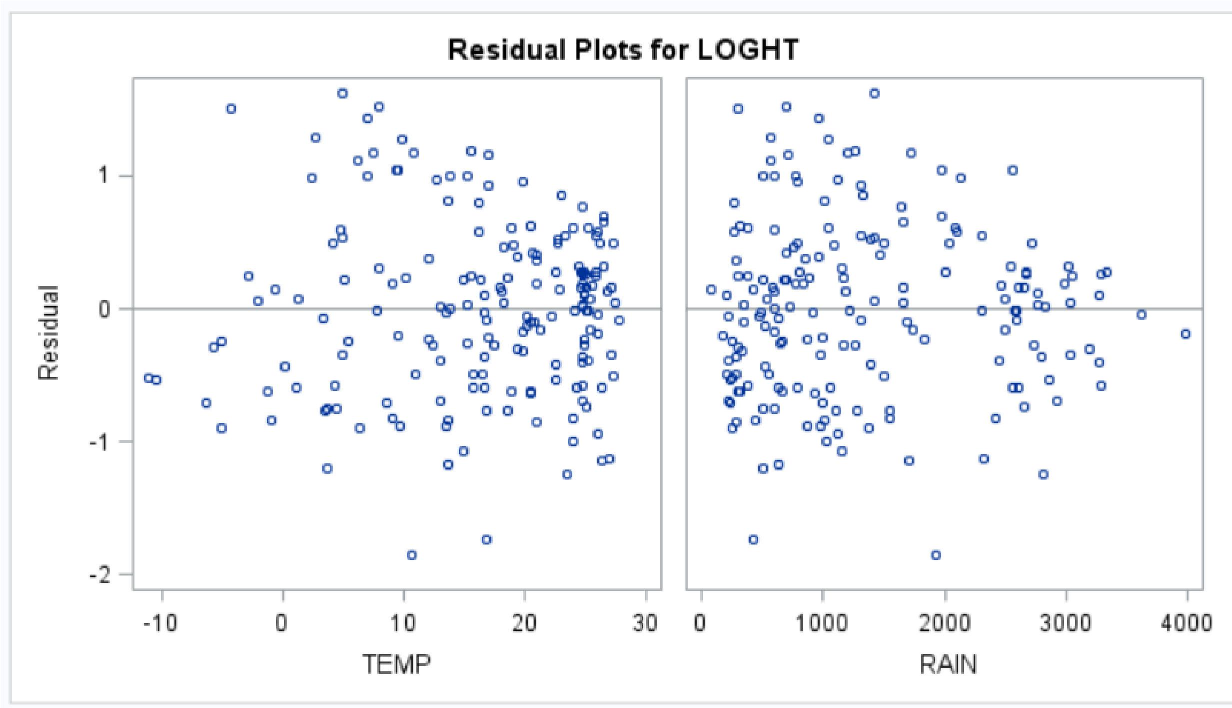
Source	DF	Type I SS	Mean Square	F Value	Pr > F
TEMP	1	26.96691279	26.96691279	62.32	<.0001
RAIN	1	6.81377671	6.81377671	15.75	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TEMP	1	8.36350580	8.36350580	19.33	<.0001
RAIN	1	6.81377671	6.81377671	15.75	0.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	-.3294816828	0.10306038	-3.20	0.0016	-.5328829265	-.1260804391
TEMP	0.0283166976	0.00644101	4.40	<.0001	0.0156046481	0.0410287471
RAIN	0.0002463537	0.00006208	3.97	0.0001	0.0001238265	0.0003688808

$$\widehat{\text{LOGHT}} = -.329 + .028\text{TEMP} + .0002\text{RAIN}$$


Fit Diagnostics for LOGHT
Residual
Predicted Value
RStudent
Leverage
Quantile
LOGHT
Cook's D
Observation
Percent
Fit–Mean
Residual
Proportion Less
Observations 178
Parameters 3
Error DF 175
MSE 0.4327
R-Square 0.3085
Adj R-Square 0.3006



Obs	LOGHT	TEMP	RAIN	Y_PRED	RESID	SRESID	SDRESID	HAT_H	COOKS_D	DFFITS	LCL_MEAN	UCL_MEAN	LCL_INDIVID	UCL_INDIVID	PRESS
1	1.44716	10.8	1208	0.27393	1.17322	1.79047	1.80193	0.007763	0.008360	0.15938	0.15955	0.38832	-1.02937	1.57724	1.18240
2	1.42488	24.5	3015	1.10703	0.31785	0.48883	0.48776	0.022945	0.001870	0.07475	0.91038	1.30369	-0.20605	2.42012	0.32531
3	-0.52288	20.9	278	0.33082	-0.85370	-1.31305	-1.31579	0.023117	0.013600	-0.20241	0.13343	0.52822	-0.98237	1.64402	-0.87390
4	0.20412	19.9	598	0.38134	-0.17722	-0.27142	-0.27070	0.014814	0.000369	-0.03319	0.22332	0.53936	-0.92652	1.68920	-0.17988
5	-0.69897	9.7	976	0.18563	-0.88460	-1.35042	-1.35362	0.008372	0.005132	-0.12438	0.06684	0.30442	-1.11807	1.48933	-0.89207
6	0.23045	22.6	1387	0.65217	-0.42172	-0.64411	-0.64303	0.009371	0.001308	-0.06254	0.52649	0.77784	-0.65218	1.95651	-0.42571
7	-0.30103	16.8	1283	0.46231	-0.76334	-1.16376	-1.16494	0.005737	0.002605	-0.08849	0.36398	0.56065	-0.83968	1.76431	-0.76775
8	1.00000	27.7	2585	1.09172	-0.09172	-0.14066	-0.14027	0.017536	0.000118	-0.01874	0.91979	1.26364	-0.21790	2.40133	-0.09335
9	1.60206	15.5	1262	0.42033	1.18173	1.80156	1.81330	0.005663	0.006162	0.13685	0.32262	0.51803	-0.88162	1.72227	1.18847
10	-0.30103	26.4	1704	0.83787	-1.13890	-1.74280	-1.75309	0.013124	0.013464	-0.20216	0.68914	0.98659	-0.46890	2.14463	-1.15404
11	-0.25964	5.4	664	-0.01299	-0.24664	-0.37747	-0.37654	0.013331	0.000642	-0.04377	-0.16289	0.13691	-1.31989	1.29391	-0.24998
12	1.50515	25.2	2087	0.89824	0.60691	0.92799	0.92762	0.011557	0.003356	0.10030	0.75867	1.03781	-0.40752	2.20400	0.61401
13	0.69897	19.3	3191	1.00315	-0.30418	-0.46974	-0.46869	0.030988	0.002352	-0.08381	0.77460	1.23169	-0.31509	2.32138	-0.31390
14	0.84510	27.2	3031	1.18743	-0.34233	-0.52668	-0.52559	0.023678	0.002242	-0.08185	0.98765	1.38721	-0.12613	2.50099	-0.35063
15	1.07918	24.8	2770	1.05517	0.02401	0.03684	0.03673	0.018333	0.000008	0.00502	0.87939	1.23096	-0.25495	2.36529	0.02446
16	0.22531	15.3	355	0.19122	0.03409	0.05218	0.05203	0.013566	0.000012	0.00610	0.04001	0.34243	-1.11583	1.49827	0.03456
17	-0.15490	20.5	926	0.47913	-0.63404	-0.96913	-0.96896	0.010875	0.003442	-0.10160	0.34375	0.61452	-0.82618	1.78445	-0.64101
18	0.60206	24.9	1831	0.82668	-0.22462	-0.34331	-0.34244	0.010756	0.000427	-0.03571	0.69203	0.96133	-0.47856	2.13192	-0.22706
19	-0.22185	23.5	2814	1.02920	-1.25105	-1.92017	-1.93517	0.019026	0.023837	-0.26950	0.85012	1.20828	-0.28137	2.33977	-1.27531
20	0.20412	13.8	598	0.20861	-0.00449	-0.00685	-0.00684	0.009327	0.000000	-0.00066	0.08323	0.33399	-1.09571	1.51292	-0.00453
21	1.50515	26.0	2110	0.92656	0.57859	0.88510	0.88455	0.012482	0.003301	0.09945	0.78151	1.07161	-0.37979	2.23291	0.58590
22	1.78533	5.0	1427	0.16365	1.62168	2.48835	2.52633	0.018485	0.038871	0.34670	-0.01287	0.34016	-1.14657	1.47387	1.65222
23	1.17026	25.8	2012	0.89675	0.27351	0.41830	0.41731	0.011978	0.000707	0.04595	0.75466	1.03884	-0.40928	2.20278	0.27682

The CORR Procedure

3 Variables:	TEMP RAIN RESID

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
TEMP	178	16.12528	9.20288	2870	-11.10000	27.70000
RAIN	178	1344	954.78816	239257	73.00000	3991
RESID	178	0	0.65409	0	-1.84556	1.62168

Pearson Correlation Coefficients, N = 178
Prob > |r| under H0: Rho=0

	TEMP	RAIN	RESID
TEMP	1.00000	0.55155 <.0001	0.00000 1.0000
RAIN	0.55155 <.0001	1.00000	0.00000 1.0000
RESID	0.00000 1.0000	0.00000 1.0000	1.00000

TEMP and the model residuals are not significantly related ($p > .999$).

RAIN and the model residuals are not significantly related ($p > .999$).

PROC REG Output

Multiple Regression: Plant Height

The REG Procedure
Model: MODEL1
Dependent Variable: LOGHT

Number of Observations Read	178
Number of Observations Used	178

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	33.78069	16.89034	39.03	<.0001
Error	175	75.72662	0.43272		
Corrected Total	177	109.50731			

TEMP and RAIN account for a significant amount of the variation in LOGHT, $F(2, 175) = 39.03$, $p < .001$, $R^2 = .309$, adjusted $R^2 = .301$.

Root MSE	0.65782	R-Square	0.3085
Dependent Mean	0.45827	Adj R-Sq	0.3006
Coeff Var	143.54456		

$R^2 = .309$
Adjusted $R^2 = .301$

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	-0.32948	0.10306	-3.20	0.0016	.	0
TEMP	1	0.02832	0.00644	4.40	<.0001	0.69580	1.43720
RAIN	1	0.00024635	0.00006208	3.97	0.0001	0.69580	1.43720

Variance inflation factor (VIF) results

A VIF < 2 indicates that multicollinearity is not a problem in this model.

$\widehat{\text{LOGHT}} = -.329 + .028\text{TEMP} + .0002\text{RAIN}$

Multiple Regression: Plant Height

The REG Procedure
Model: MODEL1
Dependent Variable: LOGHT

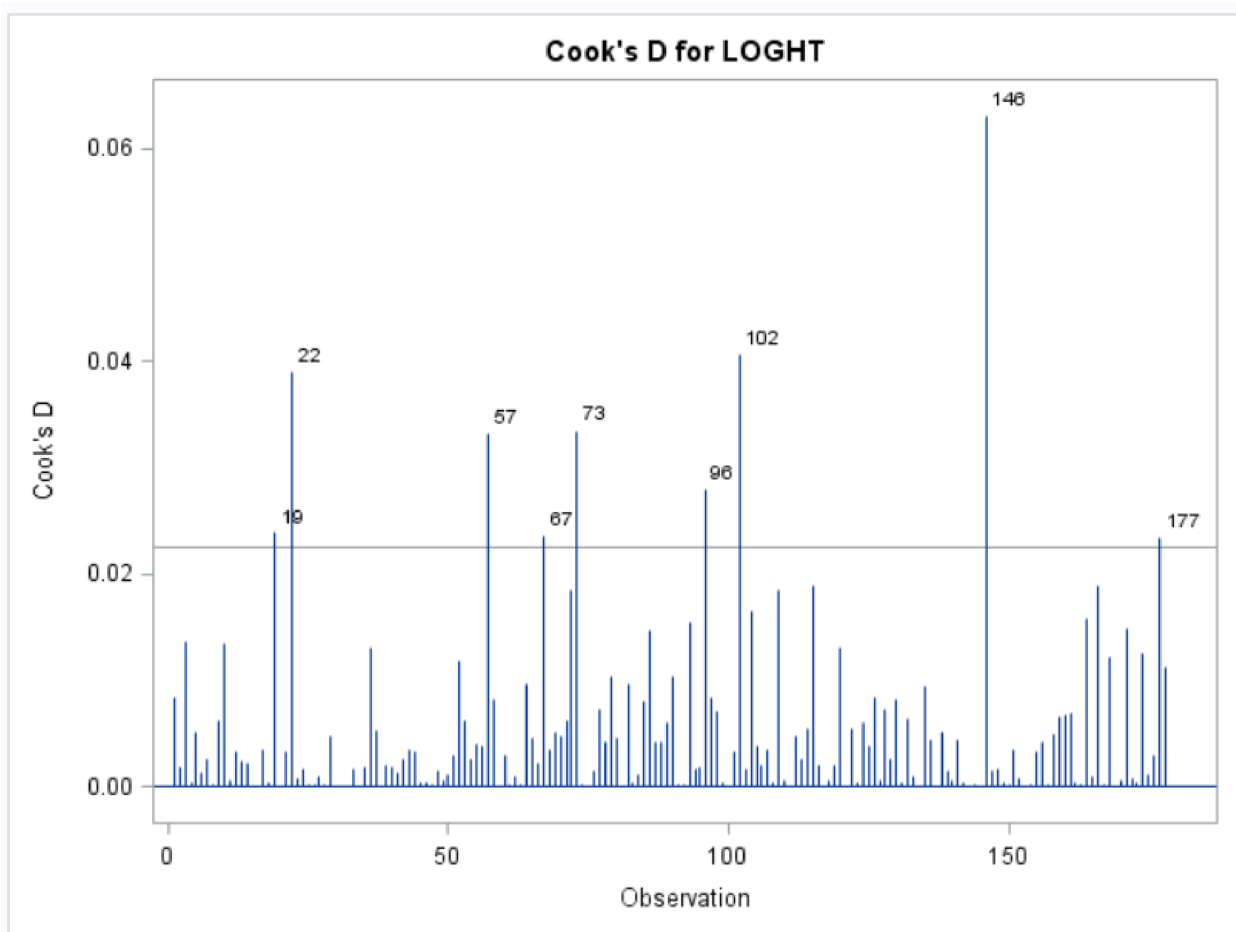
DFBETAS are available in PROC REG, but not PROC GLM.

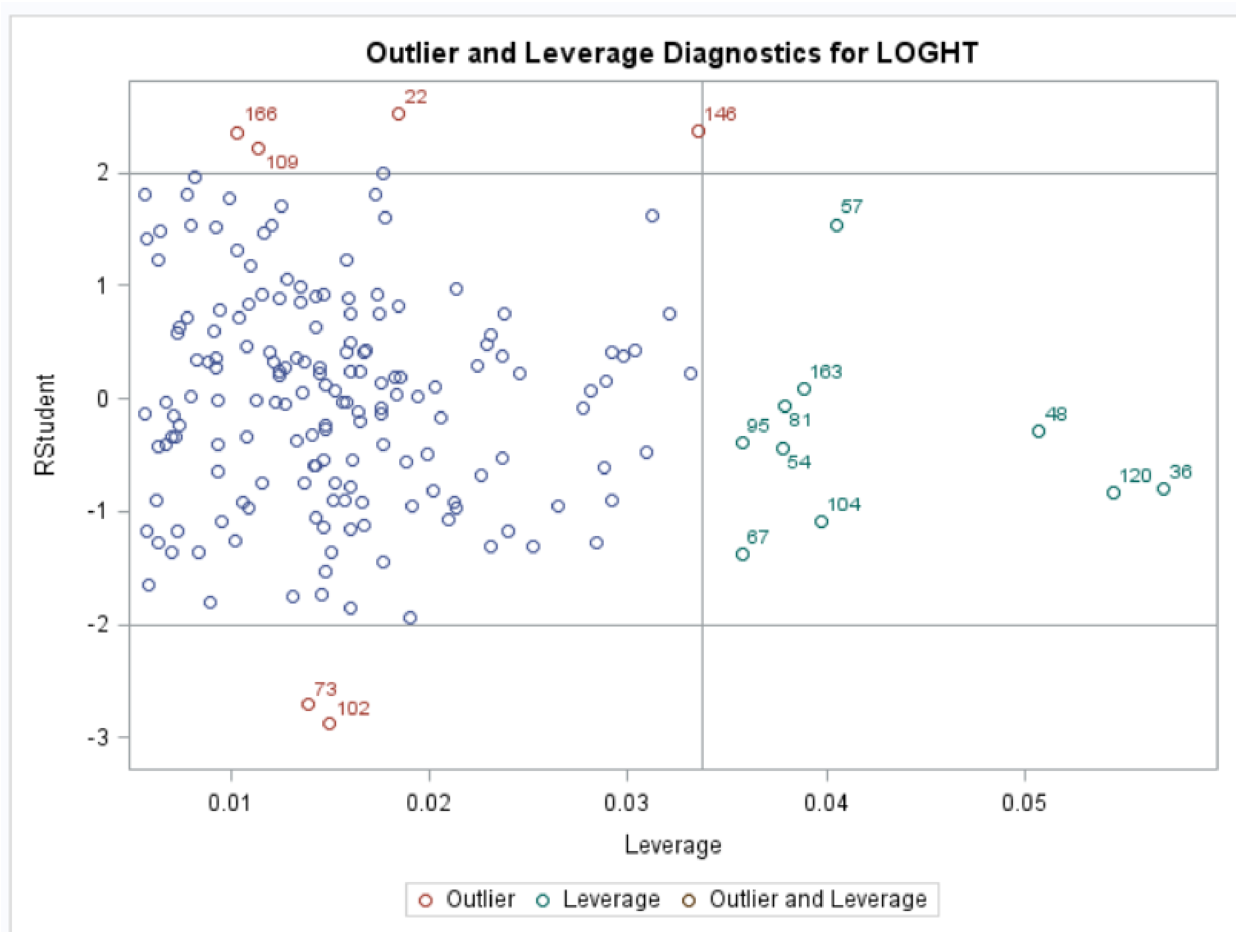
Output Statistics

Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS Intercept	DFBETAS TEMP	DFBETAS RAIN
1	1.1732	1.8019	0.0078	0.9700	0.1594	0.1237	-0.0815	0.0288
2	0.3178	0.4878	0.0229	1.0370	0.0747	-0.0248	-0.0025	0.0555
3	-0.8537	-1.3158	0.0231	1.0109	-0.2024	-0.0468	-0.1361	0.1683
4	-0.1772	-0.2707	0.0148	1.0313	-0.0332	-0.0090	-0.0207	0.0248
5	-0.8846	-1.3536	0.0084	0.9942	-0.1244	-0.1087	0.0595	0.0001
6	-0.4217	-0.6430	0.0094	1.0197	-0.0625	0.0005	-0.0395	0.0200
7	-0.7633	-1.1649	0.0057	0.9996	-0.0885	-0.0393	-0.0114	0.0110
8	-0.0917	-0.1403	0.0175	1.0351	-0.0187	0.0081	-0.0069	-0.0077
9	1.1817	1.8133	0.0057	0.9673	0.1368	0.0750	-0.0034	-0.0080
10	-1.1389	-1.7531	0.0131	0.9781	-0.2022	0.0516	-0.1445	0.0380
11	-0.2466	-0.3765	0.0133	1.0286	-0.0438	-0.0421	0.0264	0.0024
12	0.6069	0.9276	0.0116	1.0141	0.1003	-0.0297	0.0468	0.0197
13	-0.3042	-0.4687	0.0310	1.0459	-0.0838	0.0123	0.0310	-0.0748
14	-0.3423	-0.5256	0.0237	1.0371	-0.0819	0.0348	-0.0110	-0.0529
15	0.0240	0.0367	0.0183	1.0363	0.0050	-0.0017	0.0004	0.0033
16	0.0341	0.0520	0.0136	1.0313	0.0061	0.0034	0.0023	-0.0047
17	-0.6340	-0.9690	0.0109	1.0121	-0.1016	-0.0213	-0.0629	0.0615

Sum of Residuals	0
Sum of Squared Residuals	75.72662
Predicted Residual SS (PRESS)	78.17889

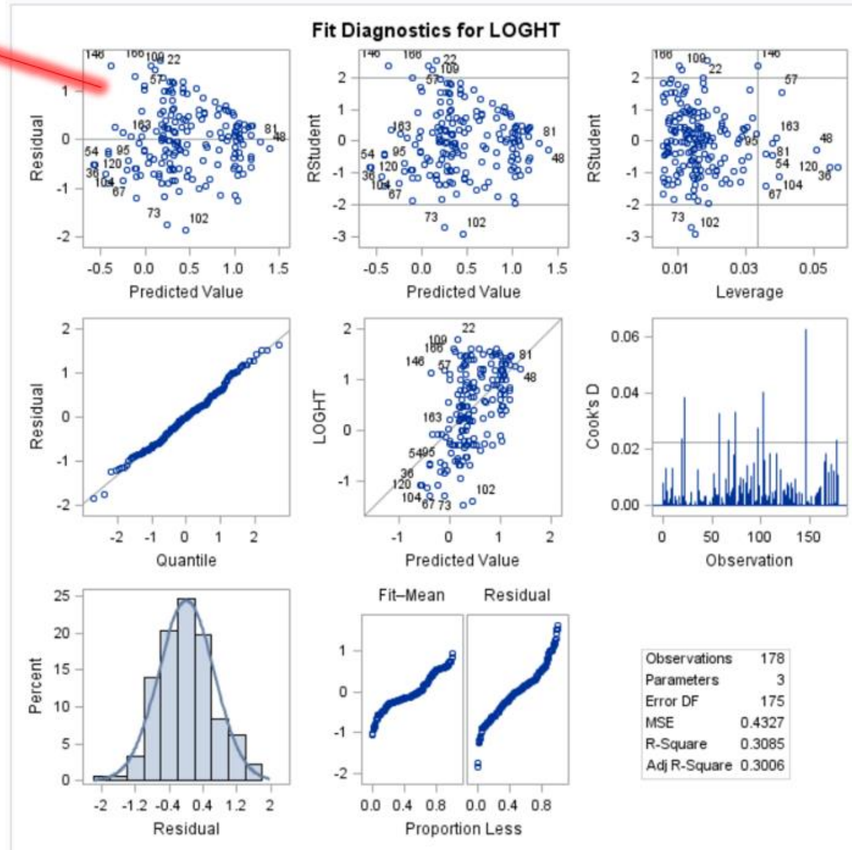
The PRESS Statistic

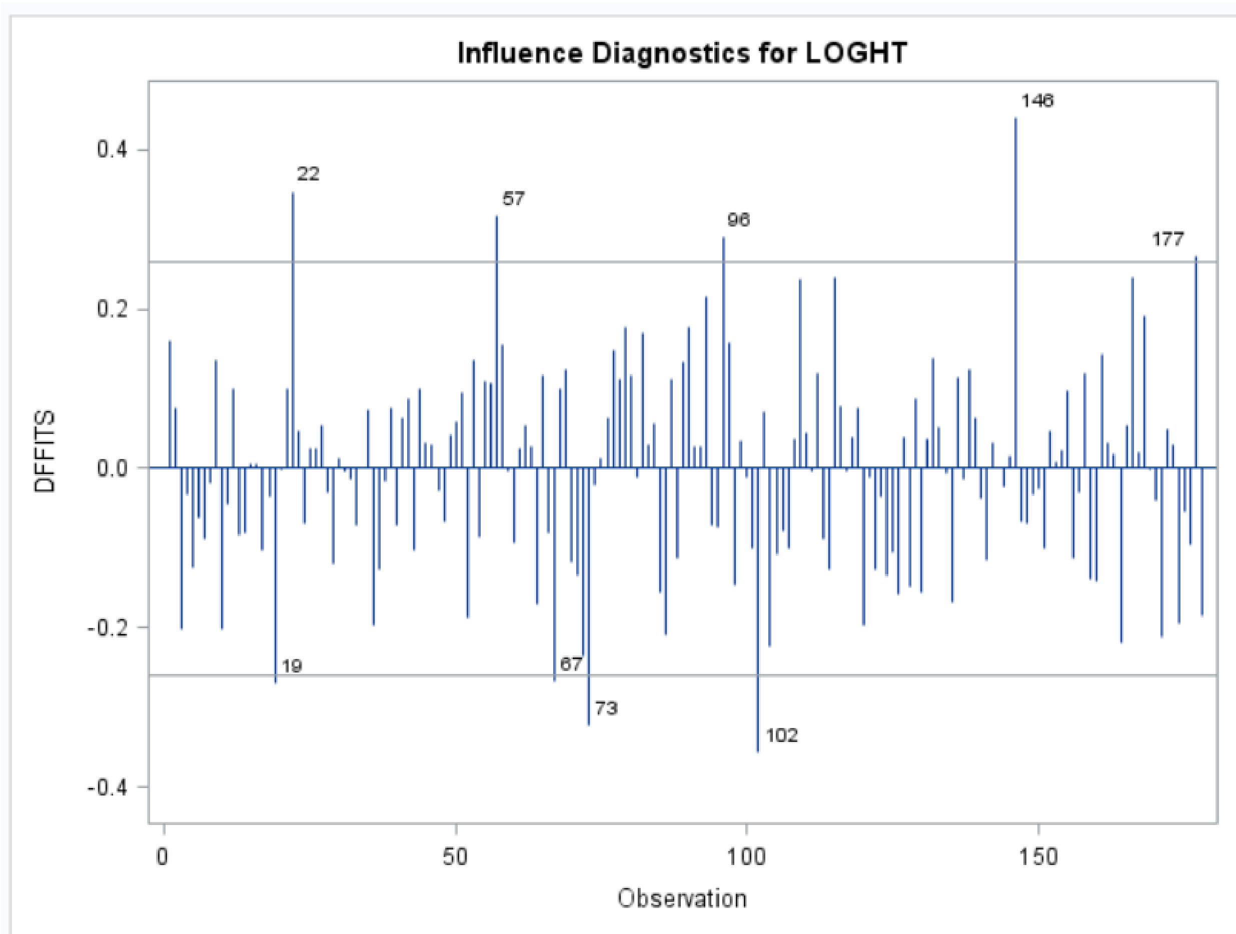

Cook's D for LOGHT
Cook's D
0.06
0.04
0.02
0.00
146
22
102
57
73
96
19
67
177
0
50
100
150
Observation

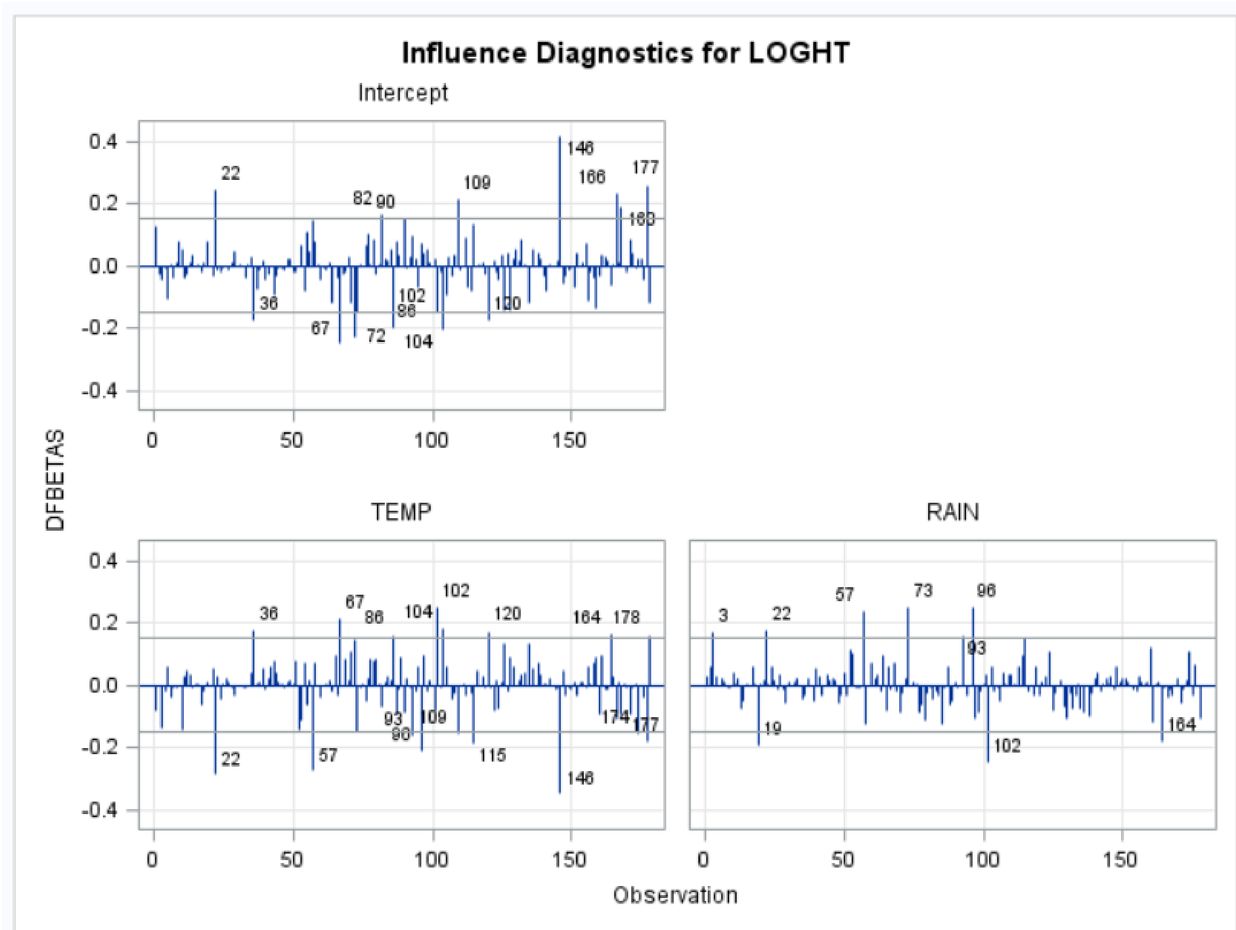

Outlier and Leverage Diagnostics for LOGHT
RStudent
Leverage
0.01
0.02
0.03
0.04
0.05
2
1
0
-1
-2
-3
166
22
146
109
57
163
95
81
54
48
120
36
104
67
73
102
Outlier
Leverage
Outlier and Leverage

The data do not appear to be randomly scattered about the reference line. Instead, it looks funnel-shaped, narrowing as the *x* value increases.

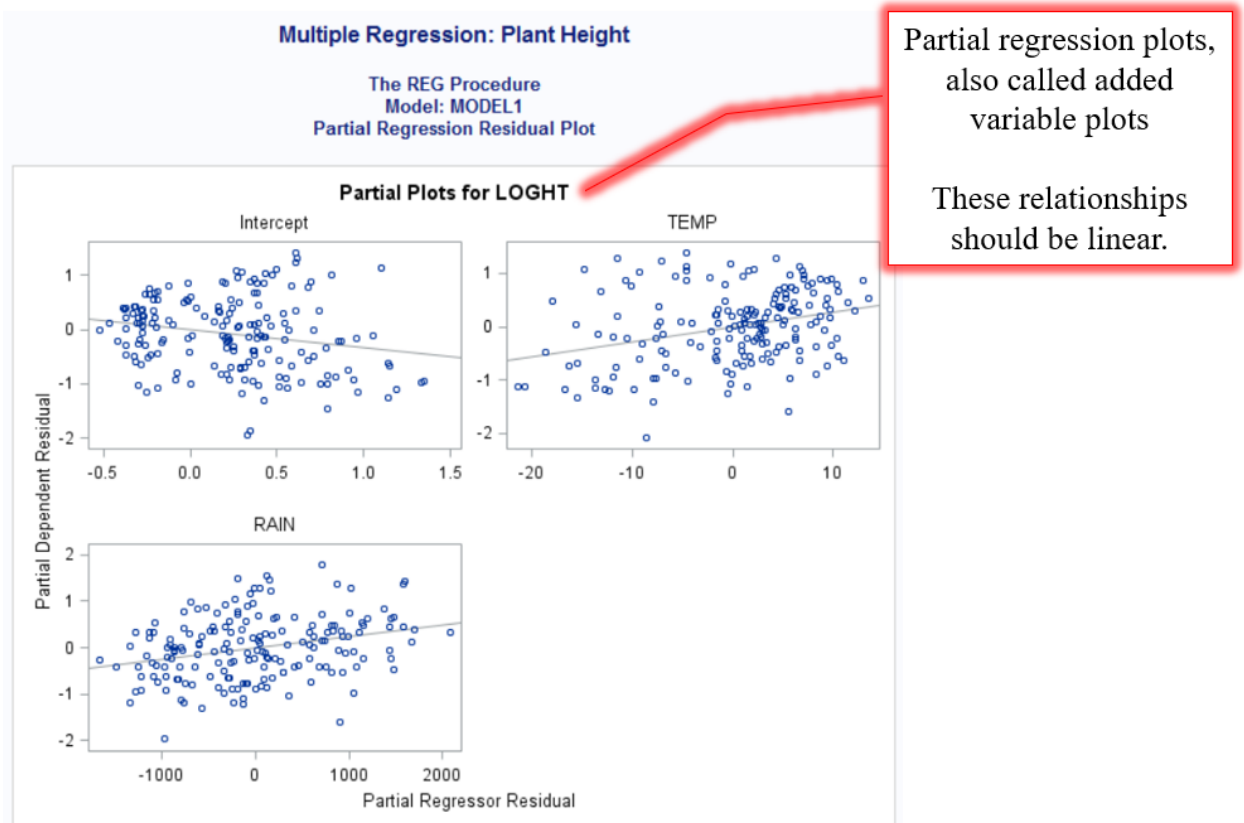
There may be a violation of the assumption of homoscedasticity.




Influence Diagnostics for LOGHT
DFFITS
Observation
0.4
0.2
0.0
-0.2
-0.4
0
50
100
150
22
57
96
146
177
19
67
73
102


Influence Diagnostics for LOGHT
Intercept
TEMP
RAIN
DFBETAS
Observation


Residual by Regressors for LOGHT
Residual
TEMP
RAIN


Multiple Regression: Plant Height
The REG Procedure
Model: MODEL1
Partial Regression Residual Plot
Partial Plots for LOGHT
Intercept
TEMP
RAIN
Partial Dependent Residual
Partial Regressor Residual
Partial regression plots, also called added variable plots
These relationships should be linear.

Obs	LOGHT	TEMP	RAIN	ID
1	1.44716	10.8	1208	1
2	1.42488	24.5	3015	2
3	-0.52288	20.9	278	3
4	0.20412	19.9	598	4
5	-0.69897	9.7	976	5
6	0.23045	22.6	1387	6
7	-0.30103	16.8	1283	7
8	1.00000	27.7	2585	8
9	1.60206	15.5	1262	9
10	-0.30103	26.4	1704	10
11	-0.25964	5.4	664	11
12	1.50515	25.2	2087	12
13	0.69897	19.3	3191	13
14	0.84510	27.2	3031	14
15	1.07918	24.8	2770	15
16	0.22531	15.3	355	16
17	-0.15490	20.5	926	17
18	0.60206	24.9	1831	18
19	-0.22185	23.5	2814	19
20	0.20412	13.8	598	20
21	1.50515	26.0	2110	21
22	1.17026	25.8	2012	23
23	0.00000	19.9	338	24
24	1.17609	24.9	2767	25

This is Plant_Height_2, the dataset from which Influential Observations 22, 102, and 146 have been excluded. Notice that ID 22, for example, is not included in this dataset.

Inferential Statistics
Model Building / Variable Selection
Research Scenario

A transportation engineer, interested in designing safer roads, would like to know the best set of predictors for vehicular accidents (rate; expressed as the accident rate per million vehicle miles). Due to the high costs of data collection, he is not interested in the full model. The predictors he is considering are:

- average daily traffic count in thousands (adt),
- truck volume as a percent of the total volume (trks),
- speed limit in 1973 (slim),
- total number of lanes of traffic (lane), and
- number of access points per mile (acpt).

This analysis will use the "Highway1.csv" file that can be found as a companion with this SAS guide. If you prefer to enter the data manually, instead of reading in the data file, the raw data for the variables of interest may be found in the Appendix of this guide.

	A	B	C	D	E	F	G	H	I	J	K	L
1	RATE	LEN	ADT	TRKS	SIGS1	SLIM	SHLD	LANE	ACPT	ITG	LWID	HTYPE
2	4.58	4.99	69	8	0.200401	55	10	8	4.6	1.2	12	FAI
3	2.86	16.11	73	8	0.062073	60	10	4	4.4	1.43	12	FAI
4	3.02	9.75	49	10	0.102564	60	10	4	4.7	1.54	12	FAI
5	2.29	10.65	61	13	0.093897	65	10	6	3.8	0.94	12	FAI
6	1.61	20.01	28	12	0.049975	70	10	4	2.2	0.65	12	FAI
7	6.87	5.97	30	6	2.007504	55	10	4	24.8	0.34	12	PA
8	3.85	8.57	46	8	0.816686	55	8	4	11	0.47	12	PA
9	6.12	5.24	25	9	0.57084	55	10	4	18.5	0.38	12	PA
10	3.29	15.79	43	12	1.453331	50	4	4	7.5	0.95	12	PA
11	5.88	8.26	23	7	1.331065	50	5	4	8.2	0.12	12	PA
12	4.2	7.03	23	6	1.992248	60	10	4	5.4	0.29	12	PA
13	4.61	13.28	20	9	1.285301	50	2	4	11.2	0.15	12	PA
14	4.8	5.4	18	14	0.745185	50	8	2	15.2	0	12	PA
15	3.85	2.96	21	8	0.337838	60	10	4	5.4	0.34	12	PA
16	2.69	11.75	27	7	0.685106	55	10	4	7.9	0.26	12	PA
17	1.99	8.86	22	9	0.112867	60	10	4	3.2	0.68	12	PA
18	2.01	9.78	19	9	0.202249	60	10	4	11	0.2	12	PA
19	4.22	5.49	9	11	0.362149	50	6	2	8.9	0.18	12	PA
20	2.76	8.63	12	8	0.115875	55	6	2	12.4	0.14	13	PA
21	2.55	20.31	12	7	1.039237	60	10	4	7.8	0.05	12	PA
22	1.89	40.09	15	13	0.144944	55	8	4	9.6	0.05	12	PA
23	2.34	11.81	8	8	0.084674	60	10	2	4.3	0	12	PA
24	2.83	11.39	5	9	0.177796	50	8	2	11.1	0	12	PA
25	1.81	22	5	15	0.045455	60	7	2	6.8	0	12	PA
26	9.23	3.58	23	6	2.78933	40	2	4	53	0.56	12	MA
27	8.6	3.23	13	6	1.239598	45	2	2	17.3	0.31	12	MA

Source of Data: Hoffstedt, C. (n.d.). [Highway accidents: Unpublished masters paper]. Unpublished raw data. Retrieved from R Package "car."

Inferential Statistics
Model Building / Variable Selection
SAS Code

```
      DATA HIGHWAY;
            INFILE "C:\Highway1.CSV" FIRSTOBS=2 DSD MISSOVER;
            INPUT RATE LEN ADT TRKS SIGS1 SLIM SHLD LANE ACPT ITG
                  LWID HTYPE $;
            KEEP RATE ADT TRKS SLIM LANE ACPT;
      RUN;

      PROC PRINT DATA=HIGHWAY;
      RUN;

(1)   PROC GLMSELECT DATA=HIGHWAY PLOTS=ALL;
(2)         MODEL RATE=ADT TRKS SLIM LANE ACPT /
(3)               SELECTION=STEPWISE
(4)               DETAILS=STEPS
(5)               SELECT=SL SLSTAY=0.05 SLENTRY=0.05;
            TITLE "Stepwise Model Selection: Highway Accident Rate";
      RUN;

      PROC GLMSELECT DATA=HIGHWAY PLOTS=ALL;
            MODEL RATE=ADT TRKS SLIM LANE ACPT /
(6)               SELECTION=FORWARD
                  DETAILS=STEPS
(7)               SELECT=SL SLENTRY=0.05;
            TITLE "Forward Model Selection: Highway Accident Rate";
      RUN;

      PROC GLMSELECT DATA=HIGHWAY PLOTS=ALL;
            MODEL RATE=ADT TRKS SLIM LANE ACPT /
(8)               SELECTION=BACKWARD
                  DETAILS=STEPS
(9)               SELECT=SL SLSTAY=0.05;
            TITLE "Backward Model Selection: Highway Accident Rate";
      RUN;

(10)  PROC REG DATA=HIGHWAY PLOTS(ONLY)=(RSQUARE ADJRSQ CP);
            MODEL RATE=ADT TRKS SLIM LANE ACPT /
(11)              SELECTION=RSQUARE
(12)              ADJRSQ CP;
            TITLE "All Possible Subsets Analysis: Highway Accident Rate";
      RUN;

      TITLE;
      QUIT;
```

(1) PROC GLMSELECT is used for model building.
(2) The MODEL statement is written as Dependent Variable (DV) = Independent Variable (IV). Be sure to include all of the IVs that are candidates for inclusion.
(3) The SELECTION option is used to identify the model building method you want to use (STEPWISE, FORWARD, or BACKWARD). Here, STEPWISE was chosen.

(4) The DETAILS=STEPS option is used if you want to see what SAS did at each step of model building (as opposed to seeing the final model only).
(5) The SELECT= option is used to identify the statistic that is used in variable selection. Options include AIC, AICC, BIC, and SBC. Here, SL is used. SL refers to “significance level.” In other words, a variable has to make a significant *p* value contribution to the model. If you choose SELECT=SL, you must also identify a significance level for a variable to qualify for entry in the model (SLENTRY=) and a significance level for a variable to qualify to stay in the model (SLSTAY=).
(6) The SELECTION option is used to identify the model building method you want to use (STEPWISE, FORWARD, or BACKWARD). Here, FORWARD was chosen.
(7) For forward selection, if you choose SELECT=SL, you must also identify the significance level for a variable to qualify for entry in the model (SLENTRY=).
(8) The SELECTION option is used to identify the model building method you want to use (STEPWISE, FORWARD, or BACKWARD). Here, BACKWARD was chosen.
(9) For backward selection, if you choose SELECT=SL, you must also identify the significance level for a variable to qualify to stay in the model (SLSTAY=).
(10) This PROC REG is used to conduct an all-possible-subsets analysis.
(11) The SELECTION=RSQUARE option tells SAS to analyze all possible subsets and rank them based on their R-square values. You may also use adjusted R-square (SELECTION=ADJRSQ) or Mallow’s Cp (SELECTION=CP).
(12) The ADJRSQ and CP options request these values to be output for each subset, even though they were not used as the ranking criteria. Remember: You may copy and paste the output into Excel if you want to sort or manipulate the results.

Inferential Statistics
Model Building / Variable Selection
Selected Output

Stepwise Model Selection: Highway Accident Rate

The GLMSELECT Procedure

Data Set	WORK.HIGHWAY
Dependent Variable	RATE
Selection Method	Stepwise
Select Criterion	Significance Level
Stop Criterion	Significance Level
Entry Significance Level (SLE)	0.05
Stay Significance Level (SLS)	0.05
Effect Hierarchy Enforced	None

Stepwise Selection

Statistical significance is being used as the criteria for selection and stopping. Other options include AIC, AICC, BIC, and SBC.

A variable must have a *p* value of .05 or less to stay in the model.

A variable must have a *p* value of .05 or less to enter the model.

Stepwise Model Selection: Highway Accident Rate

The GLMSELECT Procedure
Stepwise Selection: Step 0

Step 0: Intercept-only (null) model

Effect Entered: Intercept

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	0	0	.	.
Error	38	149.88607	3.94437	
Corrected Total	38	149.88607		

Root MSE	1.98604
Dependent Mean	3.93333
R-Square	0.0000
Adj R-Sq	0.0000
AIC	95.50624
AICC	95.83957
SBC	56.16980

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	3.933333	0.318022	12.37

Stepwise Model Selection: Highway Accident Rate

The GLMSELECT Procedure
Stepwise Selection: Step 1

Step 1: $\widehat{RATE} = ACPT$

Effect Entered: ACPT

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	1	84.76691	84.76691	48.16
Error	37	65.11915	1.75998	
Corrected Total	38	149.88607		

Root MSE	1.32664
Dependent Mean	3.93333
R-Square	0.5655
Adj R-Sq	0.5538
AIC	64.99363
AICC	65.67934
SBC	27.32075

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	1.984477	0.352115	5.64
ACPT	1	0.160281	0.023095	6.94

Entry Candidates			
Rank	Effect	Log pValue	Pr > F
1	ACPT	-17.1945	<.0001
2	SLIM	-13.2096	<.0001
3	TRKS	-7.0697	0.0009
4	LANE	-0.1719	0.8420
5	ADT	-0.1474	0.8629

Stepwise Model Selection: Highway Accident Rate

The GLMSELECT Procedure
Stepwise Selection: Step 2

Step 2:
$\widehat{RATE} = TRKS + ACPT$

Effect Entered: TRKS

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	2	94.82015	47.41008	30.99
Error	36	55.06591	1.52961	
Corrected Total	38	149.88607		

Root MSE	1.23677
Dependent Mean	3.93333
R-Square	0.6326
Adj R-Sq	0.6122
AIC	60.45380
AICC	61.63027
SBC	24.44448

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	4.429324	1.008565	4.39
TRKS	1	-0.234177	0.091344	-2.56
ACPT	1	0.138964	0.023081	6.02

Entry Candidates			
Rank	Effect	Log pValue	Pr > F
1	TRKS	-4.2214	0.0147
2	SLIM	-3.4540	0.0316
3	ADT	-1.6654	0.1891
4	LANE	-1.3978	0.2471

Stepwise Model Selection: Highway Accident Rate

The GLMSELECT Procedure
Stepwise Selection: Step 3

Step 3:
$\widehat{RATE} = TRKS + SLIM + ACPT$

Effect Entered: SLIM

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	3	101.52930	33.84310	24.50
Error	35	48.35676	1.38162	
Corrected Total	38	149.88607		

Root MSE	1.17542
Dependent Mean	3.93333
R-Square	0.6774
Adj R-Sq	0.6497
AIC	57.38673
AICC	59.20491
SBC	23.04098

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	10.206900	2.791567	3.66
TRKS	1	-0.219942	0.087053	-2.53
SLIM	1	-0.098439	0.044671	-2.20
ACPT	1	0.098149	0.028709	3.42

Entry Candidates			
Rank	Effect	Log pValue	Pr > F
1	SLIM	-3.3749	0.0342
2	ADT	-1.0350	0.3552
3	LANE	-0.6273	0.5341

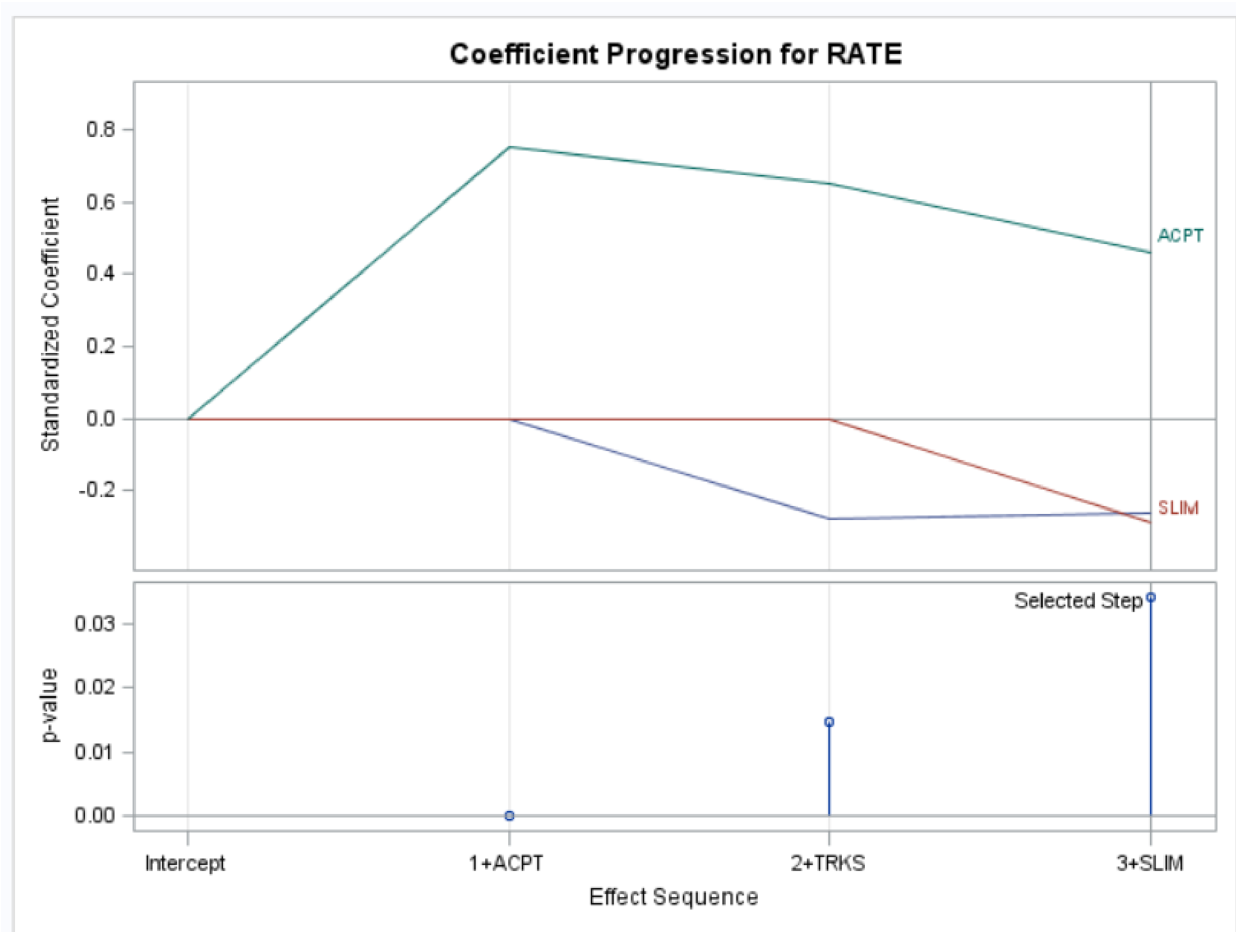
Stepwise Model Selection: Highway Accident Rate

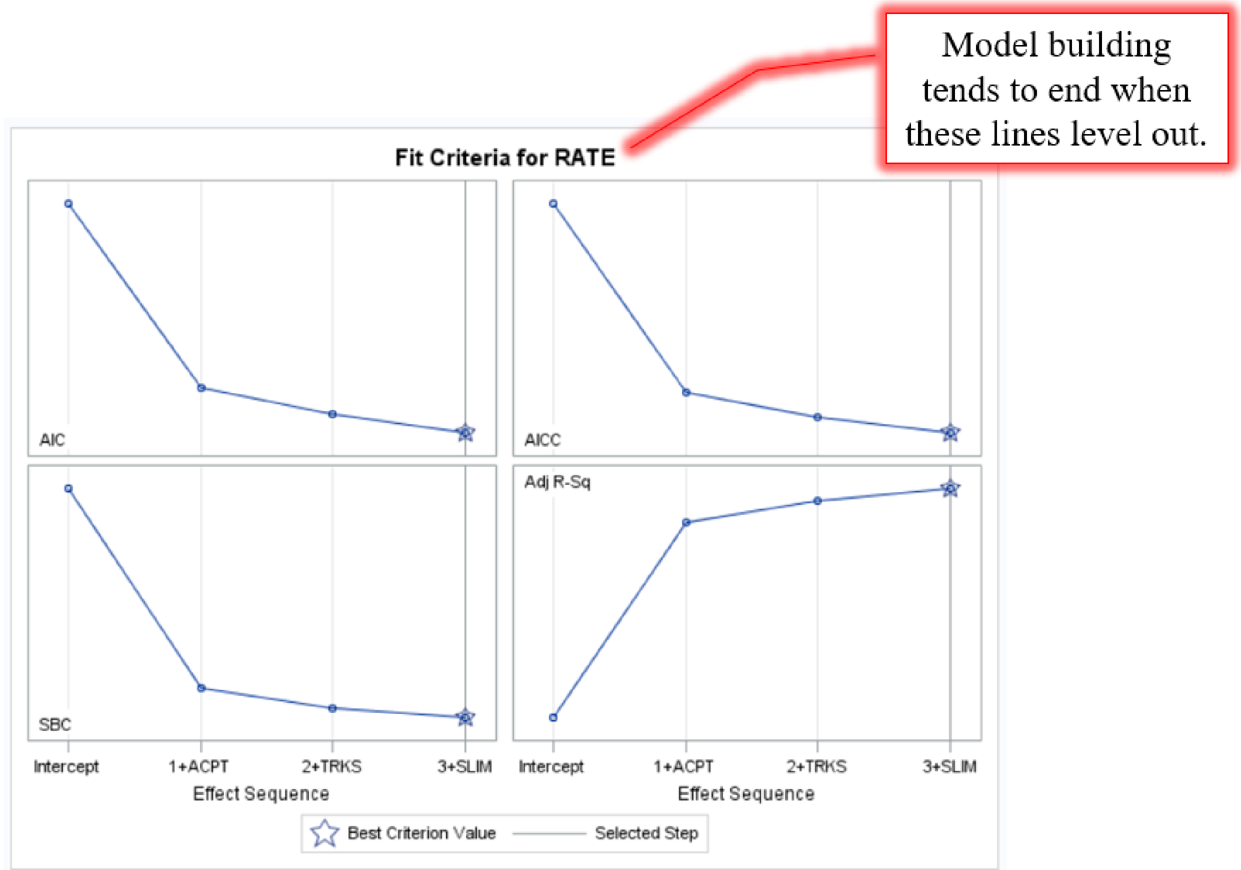
The GLMSELECT Procedure

Stepwise Selection Summary					
Step	Effect Entered	Effect Removed	Number Effects In	F Value	Pr > F
0	Intercept		1	0.00	1.0000
1	ACPT		2	48.16	<.0001
2	TRKS		3	6.57	0.0147
3	SLIM		4	4.86	0.0342

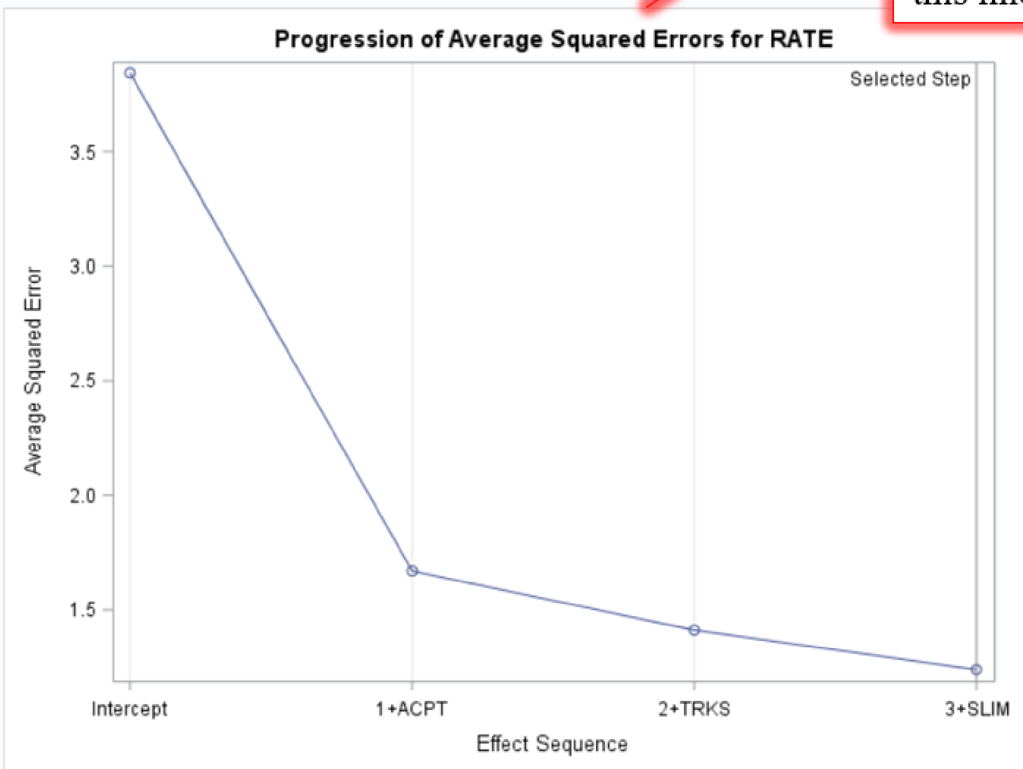
Selection stopped because the candidate for entry has SLE > 0.05 and the candidate for removal has SLS < 0.05.

Stop Details					
Candidate For	Effect	Candidate Significance		Compare Significance	
Entry	ADT	0.1897	>	0.0500	(SLE)
Removal	SLIM	0.0342	<	0.0500	(SLS)


Coefficient Progression for RATE
Standardized Coefficient
0.8
0.6
0.4
0.2
0.0
-0.2
ACPT
SLIM
p-value
0.03
0.02
0.01
0.00
Selected Step
Intercept
1+ACPT
2+TRKS
3+SLIM
Effect Sequence


Model building tends to end when these lines level out.
Fit Criteria for RATE
AIC
AICC
SBC
Adj R-Sq
Intercept
1+ACPT
2+TRKS
3+SLIM
Effect Sequence
Best Criterion Value
Selected Step

Model building tends to end when this line levels out.



Stepwise Model Selection: Highway Accident Rate

The GLMSELECT Procedure
Selected Model

Final STEPWISE results

The selected model is the model at the last step (Step 3).

Effects:	Intercept TRKS SLIM ACPT

Step 3:
$\widehat{RATE} = TRKS + SLIM + ACPT$

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	3	101.52930	33.84310	24.50
Error	35	48.35676	1.38162	
Corrected Total	38	149.88607		

Multiple regression analysis for the final model

Root MSE	1.17542
Dependent Mean	3.93333
R-Square	0.6774
Adj R-Sq	0.6497
AIC	57.38673
AICC	59.20491
SBC	23.04098

Multiple regression analysis for the final model (Cont.)

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	10.206900	2.791567	3.66
TRKS	1	-0.219942	0.087053	-2.53
SLIM	1	-0.098439	0.044671	-2.20
ACPT	1	0.098149	0.028709	3.42

$$\widehat{\text{RATE}} = 10.207 - .220\text{TRKS} - .098\text{SLIM} + .098\text{ACPT}$$

Forward Model Selection: Highway Accident Rate

The GLMSELECT Procedure

Data Set	WORK.HIGHWAY
Dependent Variable	RATE
Selection Method	Forward
Select Criterion	Significance Level
Stop Criterion	Significance Level
Entry Significance Level (SLE)	0.05
Effect Hierarchy Enforced	None

Forward Selection

Statistical significance is being used as the criteria for selection and stopping. Other options include AIC, AICC, BIC, and SBC.

A variable must have a *p* value of .05 or less to enter the model.

Number of Observations Read	39
Number of Observations Used	39

Dimensions	
Number of Effects	6
Number of Parameters	6

Forward Model Selection: Highway Accident Rate

The GLMSELECT Procedure
Forward Selection: Step 0

Step 0: Intercept-only (null) model

Effect Entered: Intercept

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	0	0	.	.
Error	38	149.88607	3.94437	
Corrected Total	38	149.88607		

Root MSE	1.98604
Dependent Mean	3.93333
R-Square	0.0000
Adj R-Sq	0.0000
AIC	95.50624
AICC	95.83957
SBC	56.16980

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	3.933333	0.318022	12.37

Forward Model Selection: Highway Accident Rate

The GLMSELECT Procedure
Forward Selection: Step 1

Step 1:
$\widehat{RATE} = ACPT$

Effect Entered: ACPT

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	1	84.76691	84.76691	48.16
Error	37	65.11915	1.75998	
Corrected Total	38	149.88607		

Root MSE	1.32664
Dependent Mean	3.93333
R-Square	0.5655
Adj R-Sq	0.5538
AIC	64.99363
AICC	65.67934
SBC	27.32075

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	1.984477	0.352115	5.64
ACPT	1	0.160281	0.023095	6.94

Entry Candidates			
Rank	Effect	Log pValue	Pr > F
1	ACPT	-17.1945	<.0001
2	SLIM	-13.2096	<.0001
3	TRKS	-7.0697	0.0009
4	LANE	-0.1719	0.8420
5	ADT	-0.1474	0.8629

Forward Model Selection: Highway Accident Rate

The GLMSELECT Procedure
Forward Selection: Step 2

Step 2:
$\widehat{RATE} = TRKS + ACPT$

Effect Entered: TRKS

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	2	94.82015	47.41008	30.99
Error	36	55.06591	1.52961	
Corrected Total	38	149.88607		

Root MSE	1.23677
Dependent Mean	3.93333
R-Square	0.6326
Adj R-Sq	0.6122
AIC	60.45380
AICC	61.63027
SBC	24.44448

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	4.429324	1.008565	4.39
TRKS	1	-0.234177	0.091344	-2.56
ACPT	1	0.138964	0.023081	6.02

Entry Candidates			
Rank	Effect	Log pValue	Pr > F
1	TRKS	-4.2214	0.0147
2	SLIM	-3.4540	0.0316
3	ADT	-1.6654	0.1891
4	LANE	-1.3978	0.2471

Forward Model Selection: Highway Accident Rate

The GLMSELECT Procedure
Forward Selection: Step 3

Step 3:
$\widehat{RATE} = TRKS + SLIM + ACPT$

Effect Entered: SLIM

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	3	101.52930	33.84310	24.50
Error	35	48.35676	1.38162	
Corrected Total	38	149.88607		

Root MSE	1.17542
Dependent Mean	3.93333
R-Square	0.6774
Adj R-Sq	0.6497
AIC	57.38673
AICC	59.20491
SBC	23.04098

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	10.206900	2.791567	3.66
TRKS	1	-0.219942	0.087053	-2.53
SLIM	1	-0.098439	0.044671	-2.20
ACPT	1	0.098149	0.028709	3.42

Entry Candidates			
Rank	Effect	Log pValue	Pr > F
1	SLIM	-3.3749	0.0342
2	ADT	-1.0350	0.3552
3	LANE	-0.6273	0.5341

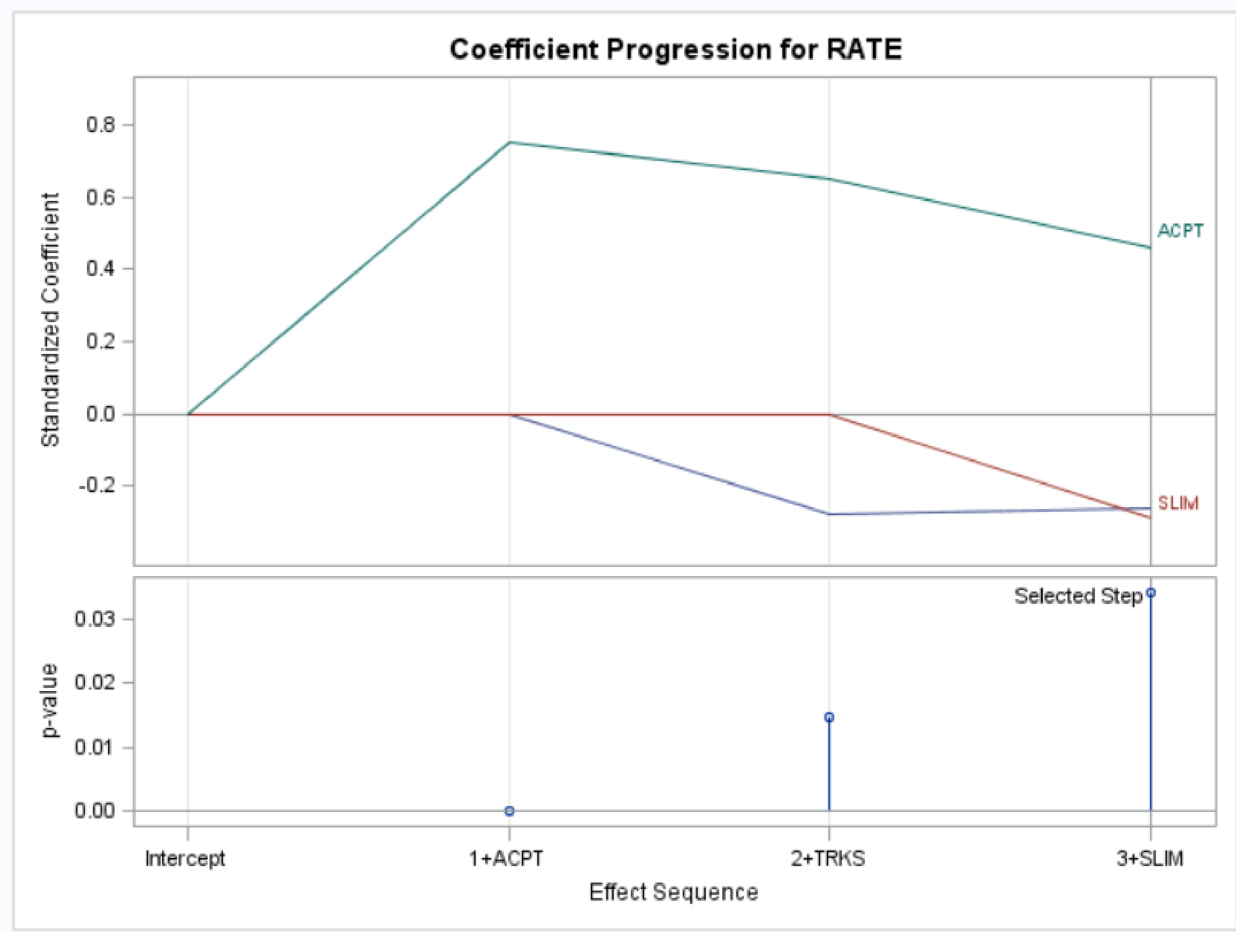
Forward Model Selection: Highway Accident Rate

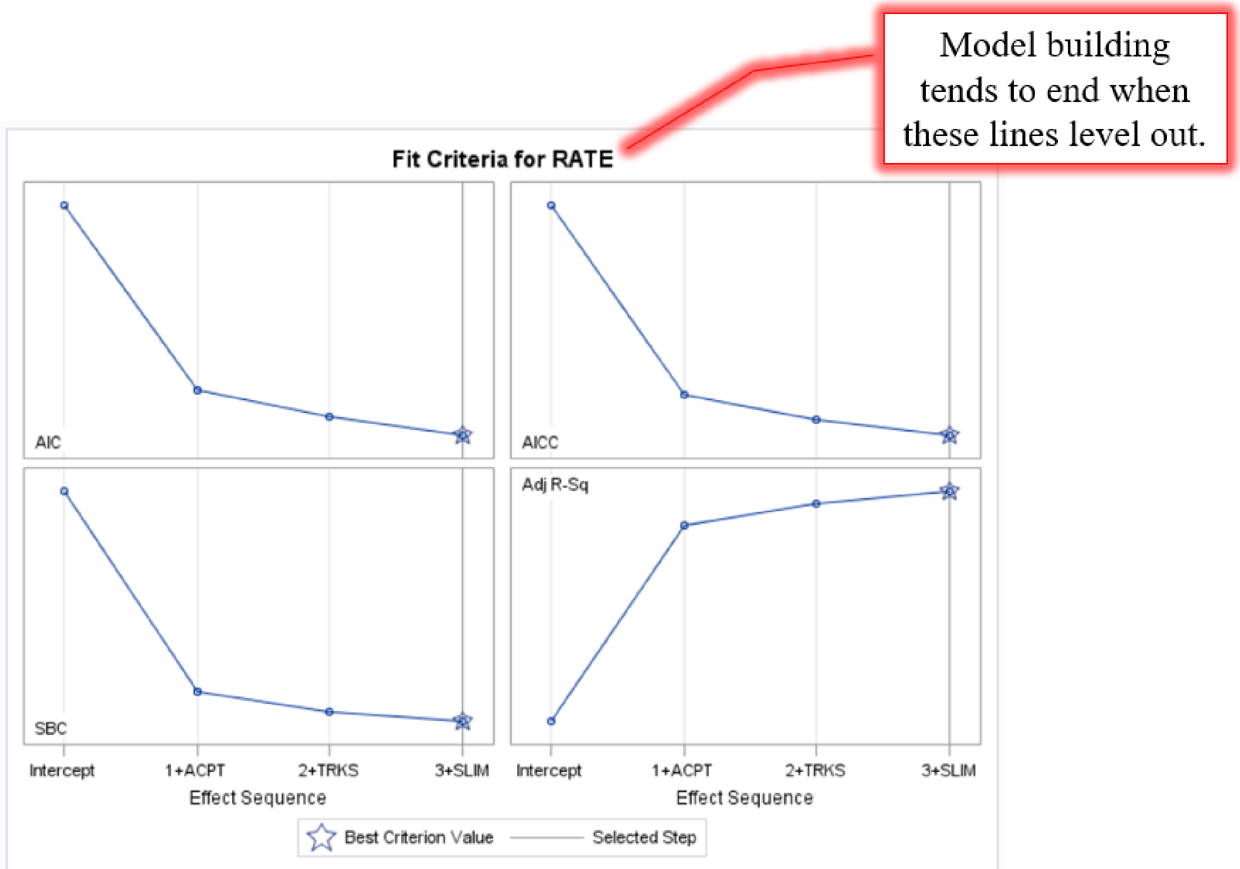
The GLMSELECT Procedure

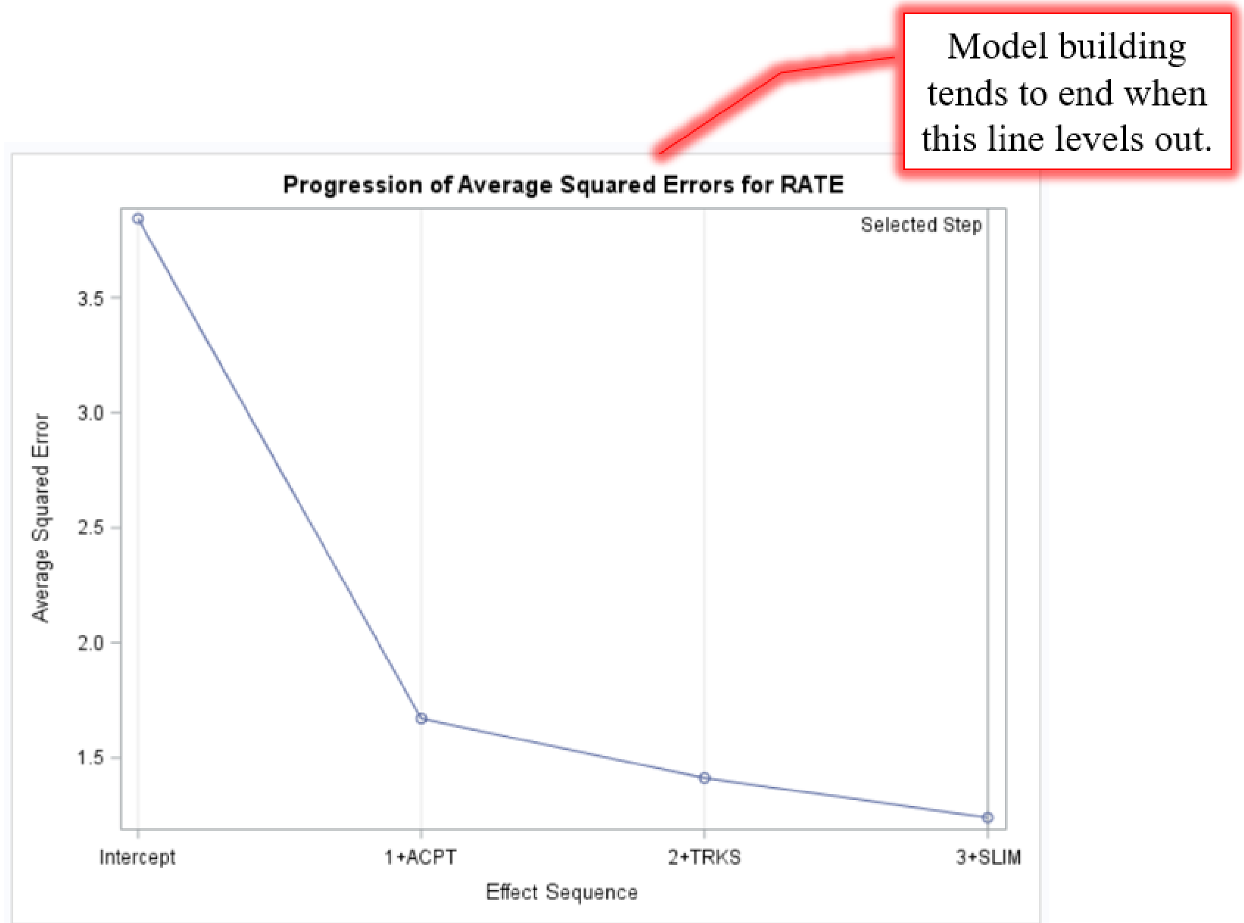
Forward Selection Summary				
Step	Effect Entered	Number Effects In	F Value	Pr > F
0	Intercept	1	0.00	1.0000
1	ACPT	2	48.16	<.0001
2	TRKS	3	6.57	0.0147
3	SLIM	4	4.86	0.0342

Selection stopped as the candidate for entry has SLE > 0.05.

Stop Details					
Candidate For	Effect	Candidate Significance		Compare Significance	
Entry	ADT	0.1897	>	0.0500	(SLE)


Coefficient Progression for RATE
Standardized Coefficient
0.8
0.6
0.4
0.2
0.0
-0.2
ACPT
SLIM
p-value
0.03
0.02
0.01
0.00
Selected Step
Intercept
1+ACPT
2+TRKS
3+SLIM
Effect Sequence


Model building tends to end when these lines level out.
Fit Criteria for RATE
AIC
AICC
SBC
Adj R-Sq
Intercept
1+ACPT
2+TRKS
3+SLIM
Effect Sequence
Best Criterion Value
Selected Step


Model building tends to end when this line levels out.
Progression of Average Squared Errors for RATE
Selected Step
Average Squared Error
3.5
3.0
2.5
2.0
1.5
Intercept
1+ACPT
2+TRKS
3+SLIM
Effect Sequence

Forward Model Selection: Highway Accident Rate

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 3).

Effects:	Intercept TRKS SLIM ACPT

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	3	101.52930	33.84310	24.50
Error	35	48.35676	1.38162	
Corrected Total	38	149.88607		

Root MSE	1.17542
Dependent Mean	3.93333
R-Square	0.6774
Adj R-Sq	0.6497
AIC	57.38673
AICC	59.20491
SBC	23.04098

Final FORWARD results

Step 3:
$\widehat{\text{RATE}} = \text{TRKS} + \text{SLIM} + \text{ACPT}$

Multiple regression analysis for the final model

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	10.206900	2.791567	3.66
TRKS	1	-0.219942	0.087053	-2.53
SLIM	1	-0.098439	0.044671	-2.20
ACPT	1	0.098149	0.028709	3.42

$$\widehat{\text{RATE}} = 10.207 - .220\text{TRKS} - .098\text{SLIM} + .098\text{ACPT}$$

Backward Model Selection: Highway Accident Rate

The GLMSELECT Procedure

Data Set	WORK.HIGHWAY
Dependent Variable	RATE
Selection Method	Backward
Select Criterion	Significance Level
Stop Criterion	Significance Level
Stay Significance Level (SLS)	0.05
Effect Hierarchy Enforced	None

Backward Selection

Statistical significance is being used as the criteria for selection and stopping. Other options include AIC, AICC, BIC, and SBC.

A variable must have a *p* value of .05 or less to stay in the model.

Number of Observations Read	39
Number of Observations Used	39

Dimensions	
Number of Effects	6
Number of Parameters	6

Backward Model Selection: Highway Accident Rate

The GLMSELECT Procedure
Backward Selection: Step 0

Step 0: Full model
$\widehat{RATE} = ADT + TRKS + SLIM + LANE + ACPT$

Full Least Squares Model

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	5	103.96214	20.79243	14.94
Error	33	45.92393	1.39163	
Corrected Total	38	149.88607		

Root MSE	1.17968
Dependent Mean	3.93333
R-Square	0.6936
Adj R-Sq	0.6472
AIC	59.37356
AICC	62.98647
SBC	28.35493

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	10.087521	2.820490	3.58
ADT	1	0.012914	0.018231	0.71
TRKS	1	-0.194036	0.090697	-2.14
SLIM	1	-0.107760	0.045732	-2.36
LANE	1	0.024717	0.253993	0.10
ACPT	1	0.103051	0.029052	3.55

Backward Model Selection: Highway Accident Rate

The GLMSELECT Procedure
Backward Selection: Step 1

Effect Removed: LANE

Step 1:
$\widehat{RATE} = ADT + TRKS + SLIM + ACPT$

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	4	103.94896	25.98724	19.23
Error	34	45.93711	1.35109	
Corrected Total	38	149.88607		

Root MSE	1.16236
Dependent Mean	3.93333
R-Square	0.6935
Adj R-Sq	0.6575
AIC	57.38475
AICC	60.00975
SBC	24.70256

Parameter Estimates				
Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**
Intercept	1	10.118501	2.761342	3.66
ADT	1	0.014338	0.010714	1.34
TRKS	1	-0.195580	0.087990	-2.22
SLIM	1	-0.107163	0.044654	-2.40
ACPT	1	0.103049	0.028625	3.60

Removal Candidates			
Rank	**Effect**	**Log pValue**	**Pr > F**
1	LANE	-0.0801	0.9231
2	ADT	-0.7263	0.4837
3	TRKS	-3.2215	0.0399
4	SLIM	-3.7072	0.0245
5	ACPT	-6.7325	0.0012

Backward Model Selection: Highway Accident Rate

The GLMSELECT Procedure
Backward Selection: Step 2

Step 2:
$\widehat{RATE} = TRKS + SLIM + ACPT$

Effect Removed: ADT

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	3	101.52930	33.84310	24.50
Error	35	48.35676	1.38162	
Corrected Total	38	149.88607		

Root MSE	1.17542
Dependent Mean	3.93333
R-Square	0.6774
Adj R-Sq	0.6497
AIC	57.38673
AICC	59.20491
SBC	23.04098

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	10.206900	2.791567	3.66
TRKS	1	-0.219942	0.087053	-2.53
SLIM	1	-0.098439	0.044671	-2.20
ACPT	1	0.098149	0.028709	3.42

Removal Candidates			
Rank	Effect	Log pValue	Pr > F
1	ADT	-1.6623	0.1897
2	TRKS	-3.4117	0.0330
3	SLIM	-3.8156	0.0220
4	ACPT	-6.9055	0.0010

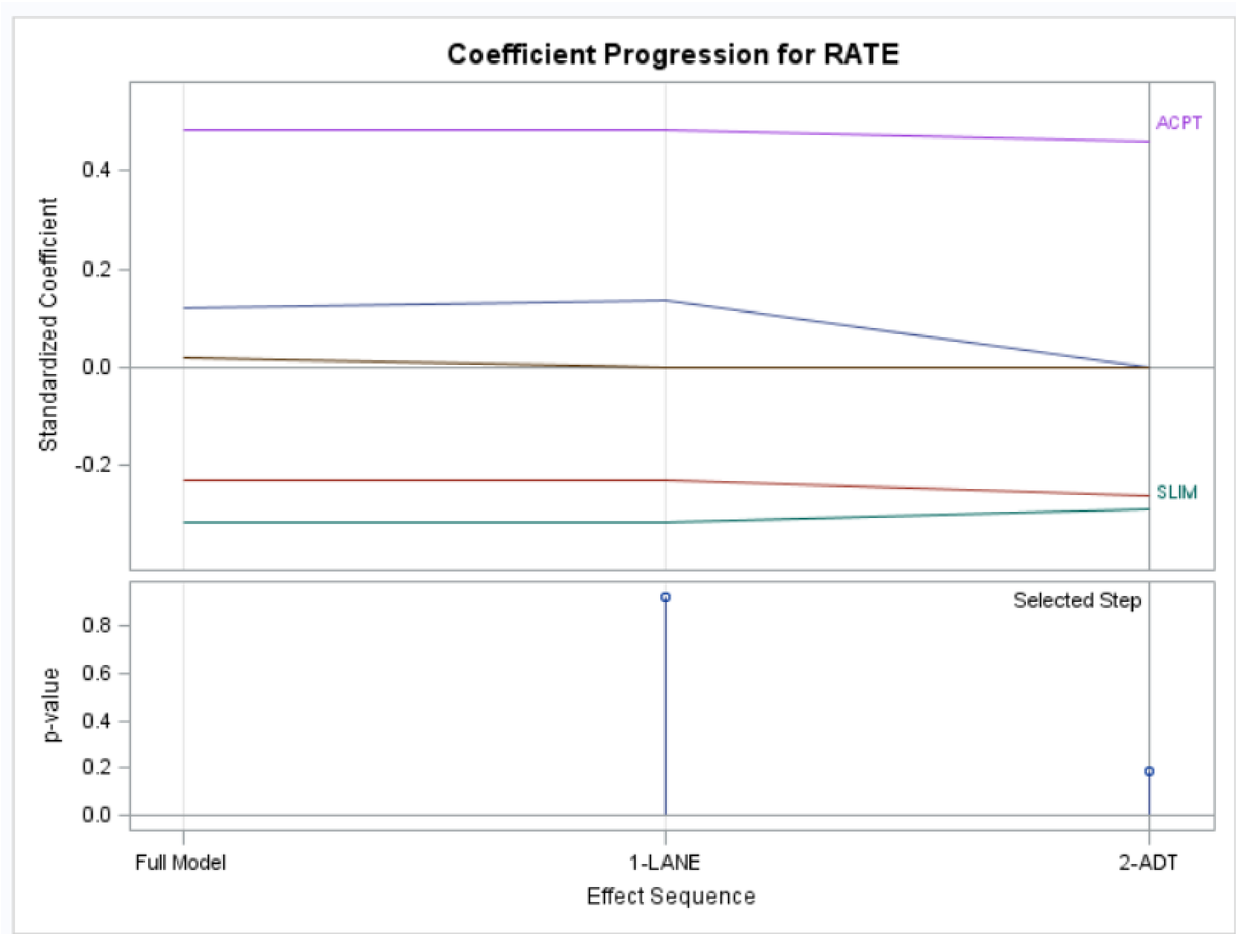
Backward Model Selection: Highway Accident Rate

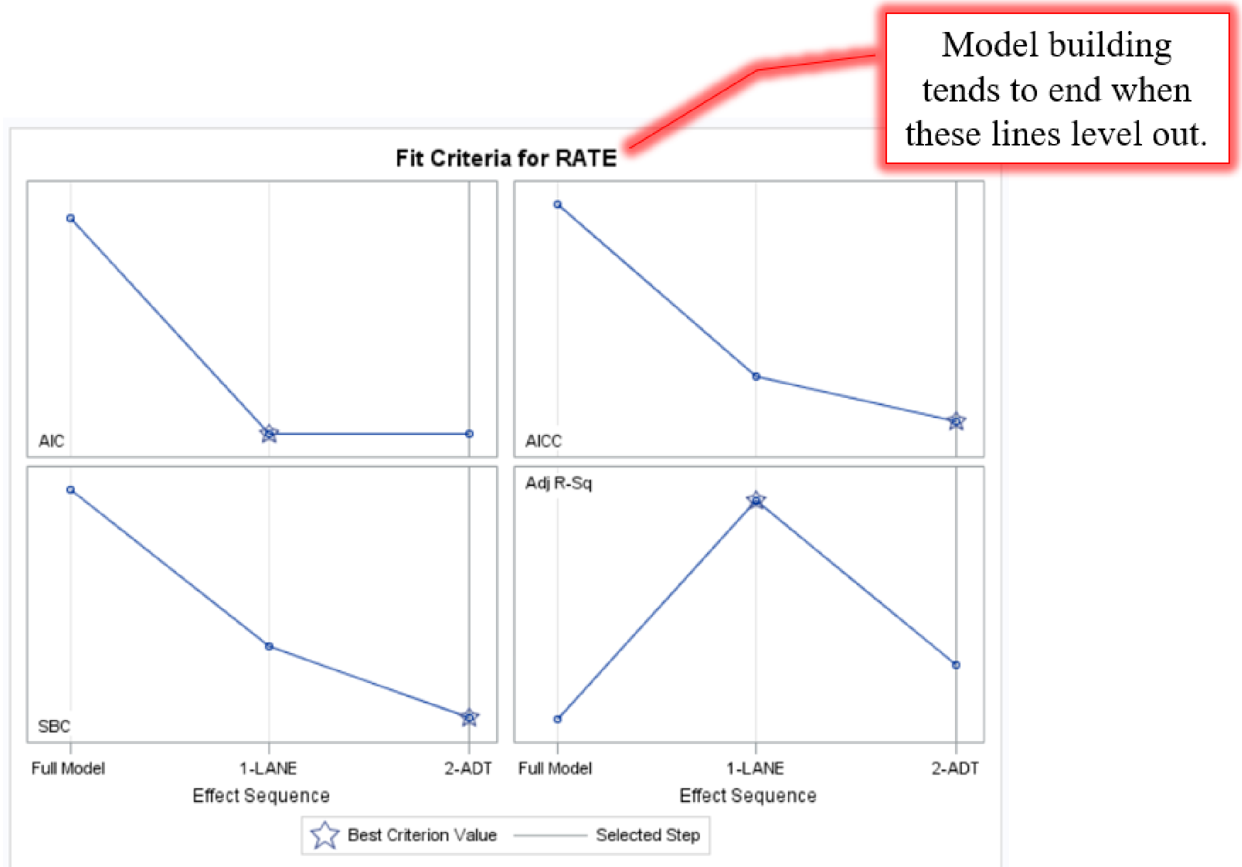
The GLMSELECT Procedure

Backward Selection Summary				
Step	Effect Removed	Number Effects In	F Value	Pr > F
0		6		
1	LANE	5	0.01	0.9231
2	ADT	4	1.79	0.1897

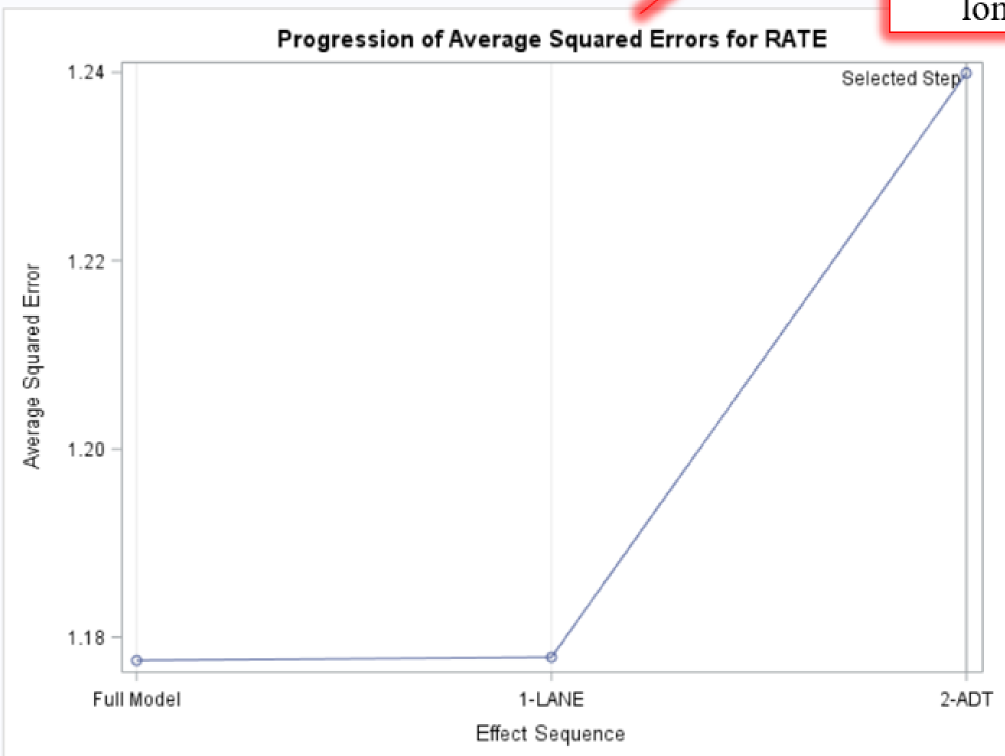
Selection stopped because the next candidate for removal has SLS < 0.05.

Stop Details					
Candidate For	Effect	Candidate Significance		Compare Significance	
Removal	SLIM	0.0342	<	0.0500	(SLS)


Coefficient Progression for RATE
Standardized Coefficient
0.4
0.2
0.0
-0.2
ACPT
SLIM
p-value
0.8
0.6
0.4
0.2
0.0
Selected Step
Full Model
1-LANE
2-ADT
Effect Sequence


Model building tends to end when these lines level out.
Fit Criteria for RATE
AIC
AICC
SBC
Adj R-Sq
Full Model
1-LANE
2-ADT
Full Model
1-LANE
2-ADT
Effect Sequence
Effect Sequence
Best Criterion Value
Selected Step

Model building tends to end when this line is no longer level.



Backward Model Selection: Highway Accident Rate

The GLMSELECT Procedure
Selected Model

Final BACKWARD results

The selected model is the model at the last step (Step 2).

Effects:	Intercept TRKS SLIM ACPT

Step 2:
$\widehat{RATE} = TRKS + SLIM + ACPT$

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	3	101.52930	33.84310	24.50
Error	35	48.35676	1.38162	
Corrected Total	38	149.88607		

Multiple regression analysis for the final model

Root MSE	1.17542
Dependent Mean	3.93333
R-Square	0.6774
Adj R-Sq	0.6497
AIC	57.38673
AICC	59.20491
SBC	23.04098

Multiple regression analysis for the final model (Cont.)

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	10.206900	2.791567	3.66
TRKS	1	-0.219942	0.087053	-2.53
SLIM	1	-0.098439	0.044671	-2.20
ACPT	1	0.098149	0.028709	3.42

$$\widehat{\text{RATE}} = 10.207 - .220\text{TRKS} - .098\text{SLIM} + .098\text{ACPT}$$

All Possible Subsets Analysis: Highway Accident Rate

The REG Procedure
Model: MODEL1
Dependent Variable: RATE

R-Square Selection Method

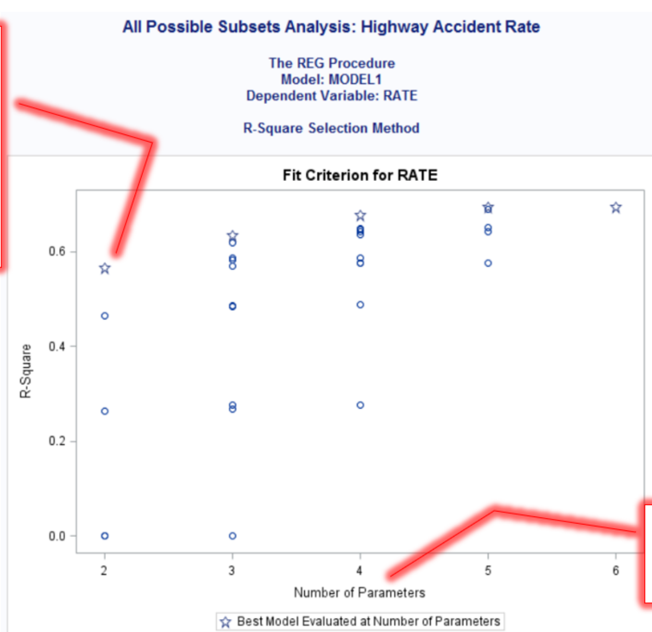
Method for determining the best subset(s)

Number of Observations Read	39
Number of Observations Used	39

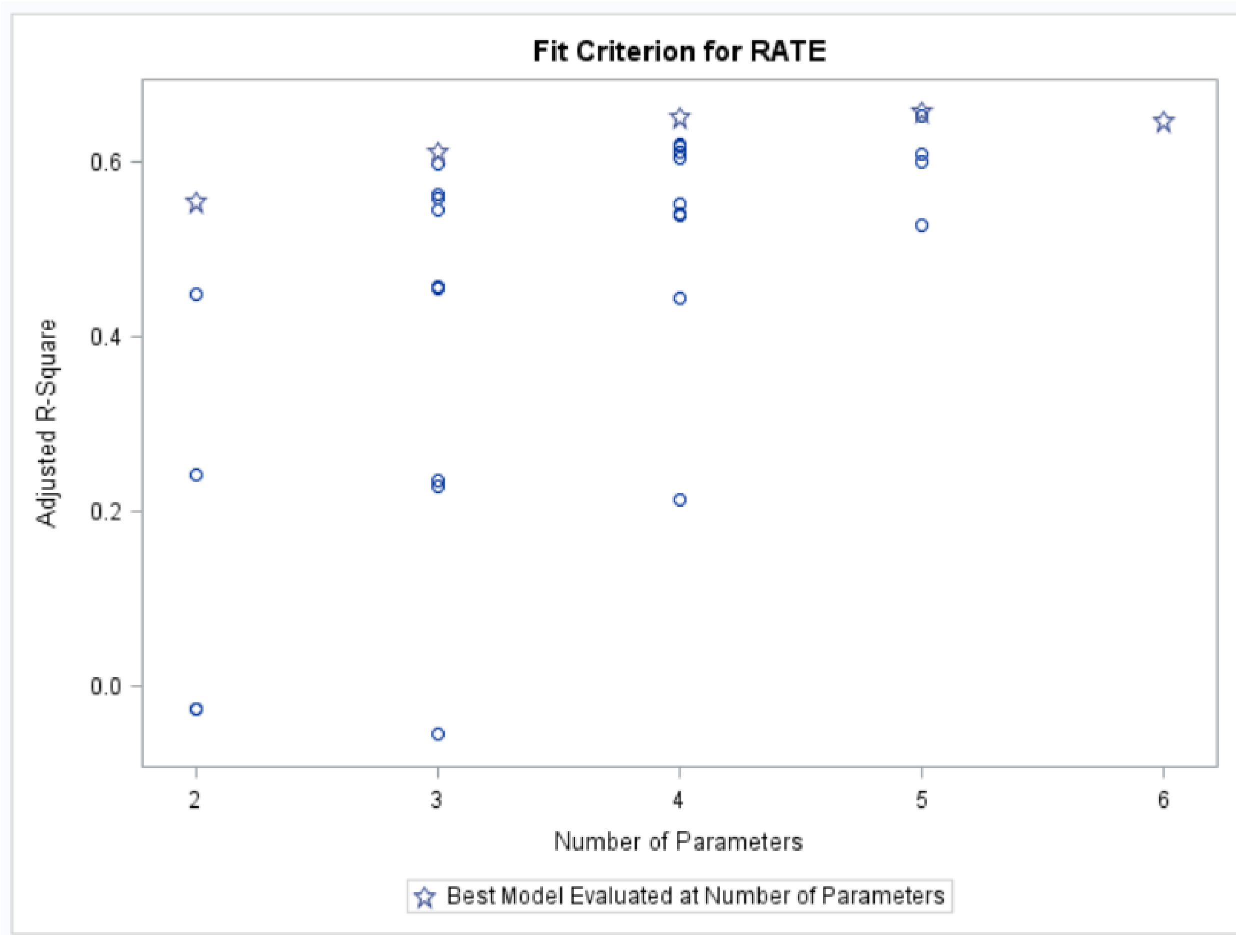
Model Index	Number in Model	R-Square	Adjusted R-Square	C(p)	Variables in Model
1	1	0.5655	0.5538	11.7933	ACPT
2	1	0.4637	0.4492	22.7581	SLIM
3	1	0.2627	0.2428	44.4132	TRKS
4	1	0.0011	-.0259	72.5879	LANE
5	1	0.0008	-.0262	72.6172	ADT
6	2	0.6326	0.6122	6.5692	TRKS ACPT
7	2	0.6185	0.5973	8.0856	SLIM ACPT
8	2	0.5861	0.5631	11.5749	ADT ACPT
9	2	0.5816	0.5584	12.0610	LANE ACPT
10	2	0.5696	0.5457	13.3518	TRKS SLIM
11	2	0.4870	0.4585	22.2503	SLIM LANE
12	2	0.4839	0.4552	22.5865	ADT SLIM
13	2	0.2754	0.2352	45.0405	TRKS LANE
14	2	0.2688	0.2282	45.7497	ADT TRKS
15	2	0.0011	-.0544	74.5873	ADT LANE
16	3	0.6774	0.6497	3.7482	TRKS SLIM ACPT
17	3	0.6490	0.6189	6.8062	ADT SLIM ACPT
18	3	0.6470	0.6167	7.0233	SLIM LANE ACPT
19	3	0.6416	0.6109	7.6011	ADT TRKS ACPT
20	3	0.6367	0.6056	8.1283	TRKS LANE ACPT
21	3	0.5864	0.5510	13.5441	ADT LANE ACPT
22	3	0.5767	0.5404	14.5913	ADT TRKS SLIM
23	3	0.5750	0.5386	14.7764	TRKS SLIM LANE

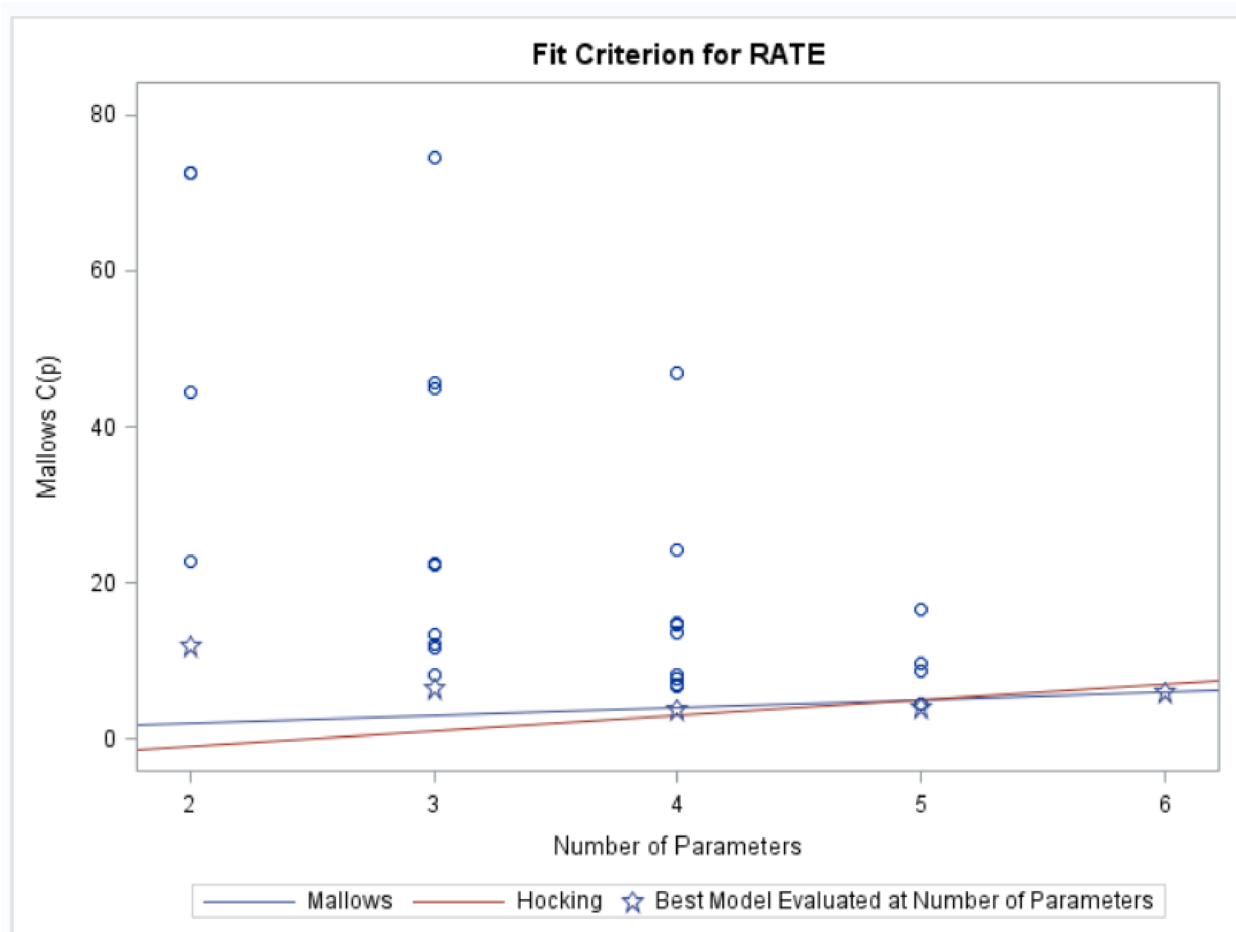
This data may be copied and pasted into Excel if you want to sort or otherwise manipulate the results.

This is the best two-parameter model (i.e. the best model with one IV). It has an R^2 value just shy of .6. From the previous table, you can see that this is the model that includes ACPT ($R^2 = .5655$).



Here, the number of parameters is equal to the number of IVs + 1 intercept.


Fit Criterion for RATE
Adjusted R-Square
0.6
0.4
0.2
0.0
2
3
4
5
6
Number of Parameters
Best Model Evaluated at Number of Parameters


Fit Criterion for RATE
Mallows C(p)
80
60
40
20
0
2
3
4
5
6
Number of Parameters
Mallows
Hocking
Best Model Evaluated at Number of Parameters

Appendix
US Cereal Data

CALORIES	CARBO
212.12121	15.15152
212.12121	21.21212
100.00000	16.00000
146.66667	14.00000
110.00000	11.00000
173.33333	24.00000
134.32836	22.38806
134.32836	19.40299
160.00000	16.00000
88.00000	13.60000
160.00000	17.33333
220.00000	26.00000
110.00000	12.00000
110.00000	22.00000
100.00000	21.00000
110.00000	13.00000
110.00000	12.00000
220.00000	20.00000
110.00000	21.00000
133.33333	14.66667
133.33333	24.00000
110.00000	11.00000
146.66667	18.66667
125.00000	17.50000
179.10448	17.91045
179.10448	20.89552
146.66667	17.33333
113.63636	12.50000
146.66667	20.00000
113.63636	17.04545
440.00000	68.00000
363.63636	39.39394
120.00000	12.00000
146.66667	15.33333
82.70677	10.52632
186.66667	26.66667
73.33333	14.00000
149.25373	17.91045
110.00000	12.00000

238.80597	25.37313
100.00000	15.00000
179.10448	22.38806
208.95522	31.34328
260.00000	27.00000
179.10448	16.41791
100.00000	20.00000
50.00000	13.00000
200.00000	28.00000
160.00000	18.66667
200.00000	21.00000
180.00000	30.00000
97.34513	20.35398
110.00000	22.00000
134.32836	28.35821
134.32836	29.85075
146.66667	12.00000
110.00000	16.00000
110.00000	21.00000
140.00000	15.00000
100.00000	16.00000
146.66667	28.00000
110.00000	13.00000
149.25373	25.37313
100.00000	17.00000
146.66667	21.33333

Appendix
Plant Height Data

HEIGHT	LOGHT	TEMP	RAIN
28.000000	1.447158	10.80	1208
26.600000	1.424882	24.50	3015
0.300000	-0.522880	20.90	278
1.600000	0.204120	19.90	598
0.200000	-0.698970	9.70	976
1.700000	0.230449	22.60	1387
0.500000	-0.301030	16.80	1283
10.000000	1.000000	27.70	2585
40.000000	1.602060	15.50	1262
0.500000	-0.301030	26.40	1704
0.550000	-0.259640	5.40	664
32.000000	1.505150	25.20	2087
5.000000	0.698970	19.30	3191
7.000000	0.845098	27.20	3031
12.000000	1.079181	24.80	2770
1.680000	0.225309	15.30	355
0.700000	-0.154900	20.50	926
4.000000	0.602060	24.90	1831
0.600000	-0.221850	23.50	2814
1.600000	0.204120	13.80	598
32.000000	1.505150	26.00	2110
61.000000	1.785330	5.00	1427
14.800000	1.170262	25.80	2012
1.000000	0.000000	19.90	338
15.000000	1.176091	24.90	2767
7.000000	0.845098	26.80	1184
20.000000	1.301030	25.00	2664
7.000000	0.845098	25.40	2494
2.900000	0.462398	26.30	2607
9.670000	0.985426	18.20	3042
8.000000	0.903090	24.10	2314
2.000000	0.301030	22.30	216
0.600000	-0.221850	15.70	1052
1.700000	0.230449	13.00	723
7.000000	0.845098	18.30	762
0.080000	-1.096910	-11.10	252
0.200000	-0.698970	14.90	1150
0.707000	-0.150580	3.40	637
7.000000	0.845098	20.70	703

0.500000	-0.301030	13.00	214
23.500000	1.371068	26.50	2542
16.000000	1.204120	23.40	1315
0.233000	-0.632640	0.20	526
34.000000	1.531479	25.90	2315
15.000000	1.176091	25.50	2462
15.000000	1.176091	24.60	2660
1.500000	0.176091	17.00	1003
16.000000	1.204120	26.00	3991
2.800000	0.447158	15.60	384
25.000000	1.397940	25.90	3048
20.000000	1.301030	27.30	1505
0.400000	-0.397940	23.90	1027
6.000000	0.778151	4.20	2043
0.200000	-0.698970	-5.70	305
3.500000	0.544068	4.80	599
25.000000	1.397940	17.00	1307
18.000000	1.255273	2.40	2142
10.000000	1.000000	16.20	276
10.000000	1.000000	25.20	2576
0.500000	-0.301030	15.70	281
19.000000	1.278754	24.90	3273
20.000000	1.301030	24.60	2674
3.000000	0.477121	17.90	583
0.110000	-0.958610	13.70	630
12.500000	1.096910	26.20	790
0.600000	-0.221850	16.50	546
0.050000	-1.301030	-5.10	257
30.000000	1.477121	22.70	2726
35.000000	1.544068	24.80	1649
3.000000	0.477121	22.50	2865
0.200000	-0.698970	1.20	788
0.050000	-1.301030	3.70	500
0.032200	-1.492140	16.80	420
3.500000	0.544068	21.20	1741
12.000000	1.079181	25.40	2494
0.800000	-0.096910	-2.90	301
16.000000	1.204120	13.80	603
16.000000	1.204120	23.90	1036
30.000000	1.477121	17.00	706
30.000000	1.477121	26.50	1661
18.100000	1.257679	26.00	3612
10.000000	1.000000	7.00	508
3.000000	0.477121	14.90	682
7.000000	0.845098	19.40	964

4.000000	0.602060	24.80	3283
0.080000	-1.096910	-1.00	436
5.000000	0.698970	5.00	1427
2.500000	0.397940	24.30	2567
32.000000	1.505150	23.00	1327
30.000000	1.477121	9.90	1037
10.000000	1.000000	27.10	1664
2.000000	0.301030	9.00	781
29.300000	1.466868	9.40	1975
4.000000	0.602060	25.30	2444
0.220000	-0.657580	-5.10	257
41.000000	1.612784	9.50	2561
24.000000	1.380211	19.80	793
2.000000	0.301030	25.00	2662
4.500000	0.653213	18.60	882
2.400000	0.380211	16.90	1313
2.000000	0.301030	27.30	1505
0.040000	-1.397940	10.60	1936
28.000000	1.447158	24.80	3283
0.070000	-1.154900	-6.40	244
0.280000	-0.552840	8.50	996
5.000000	0.698970	24.80	2803
0.500000	-0.301030	18.60	1099
0.800000	-0.096910	-0.70	422
35.000000	1.544068	7.00	972
20.000000	1.301030	24.80	2993
1.800000	0.255273	13.50	915
19.000000	1.278754	12.70	1121
0.350000	-0.455930	10.90	208
0.250000	-0.602060	13.00	214
30.000000	1.477121	7.50	1720
15.000000	1.176091	22.70	1397
10.000000	1.000000	25.10	2598
3.000000	0.477121	22.90	73
30.000000	1.477121	24.90	3329
0.080000	-1.096910	-10.50	236
2.020000	0.305351	20.20	475
0.800000	-0.096910	23.90	1545
1.150000	0.060698	12.40	1263
0.450000	-0.346790	18.90	293
6.000000	0.778151	24.80	3269
0.150000	-0.823910	-1.20	657
1.584893	0.200000	5.10	691
0.140000	-0.853870	3.70	500
5.000000	0.698970	20.90	278

2.500000	0.397940	24.80	2920
3.000000	0.477121	16.40	501
20.000000	1.301030	15.30	780
6.000000	0.778151	22.60	803
1.700000	0.230449	16.70	484
0.200000	-0.698970	6.40	1379
6.000000	0.778151	16.20	272
3.800000	0.579784	20.80	1698
8.000000	0.903090	18.90	380
9.000000	0.954243	19.10	1085
0.600000	-0.221850	9.60	174
0.239000	-0.621600	13.50	867
4.500000	0.653213	21.00	834
12.000000	1.079181	24.80	2835
1.700000	0.230449	20.50	354
0.810000	-0.091510	1.30	539
13.500000	1.130334	-4.30	296
0.500000	-0.301030	4.90	977
0.720000	-0.142670	16.70	290
1.500000	0.176091	17.50	1165
1.710000	0.232996	20.10	520
0.300000	-0.522880	13.70	1019
3.000000	0.477121	7.90	1156
8.000000	0.903090	27.50	1663
2.900000	0.462398	18.10	597
13.000000	1.113943	13.60	1016
0.200000	-0.698970	4.30	374
1.000000	0.000000	12.00	872
39.600000	1.597695	26.50	1974
0.158000	-0.801340	4.50	597
0.500000	-0.301030	20.40	310
9.000000	0.954243	20.40	310
3.000000	0.477121	10.10	1176
1.050000	0.021189	-2.10	1418
0.500000	-0.301030	9.10	2421
11.000000	1.041393	21.00	1476
39.000000	1.591065	8.00	692
1.940000	0.287802	16.60	212
12.400000	1.093422	6.20	564
1.500000	0.176091	7.80	1211
1.000000	0.000000	15.30	656
0.750000	-0.124940	27.00	2319
4.000000	0.602060	12.10	859
15.000000	1.176091	25.00	2616
0.550000	-0.259640	26.00	1117

6.000000	0.778151	24.90	2731
0.500000	-0.301030	16.70	630
15.000000	1.176091	2.70	572
0.246000	-0.609060	3.50	1555

Appendix
Highway1 Data

RATE	ADT	TRKS	SLIM	LANE	ACPT
4.58	69	8	55	8	4.6
2.86	73	8	60	4	4.4
3.02	49	10	60	4	4.7
2.29	61	13	65	6	3.8
1.61	28	12	70	4	2.2
6.87	30	6	55	4	24.8
3.85	46	8	55	4	11.0
6.12	25	9	55	4	18.5
3.29	43	12	50	4	7.5
5.88	23	7	50	4	8.2
4.20	23	6	60	4	5.4
4.61	20	9	50	4	11.2
4.80	18	14	50	2	15.2
3.85	21	8	60	4	5.4
2.69	27	7	55	4	7.9
1.99	22	9	60	4	3.2
2.01	19	9	60	4	11.0
4.22	9	11	50	2	8.9
2.76	12	8	55	2	12.4
2.55	12	7	60	4	7.8
1.89	15	13	55	4	9.6
2.34	8	8	60	2	4.3
2.83	5	9	50	2	11.1
1.81	5	15	60	2	6.8
9.23	23	6	40	4	53.0
8.60	13	6	45	2	17.3
8.21	7	8	55	2	27.3
2.93	10	10	55	2	18.0
7.48	12	7	45	2	30.2
2.57	9	8	60	2	10.3
5.77	4	8	45	2	18.2
2.90	5	10	55	2	12.3
2.97	4	13	55	2	7.1
1.84	5	12	55	2	14.0
3.78	2	10	55	2	11.3
2.76	3	8	50	2	16.3
4.27	1	11	55	2	9.6
3.05	3	11	60	2	9.0
4.12	1	10	55	2	10.4