STA 216 Handouts 2019

Prof. Adrian

Table of Contents

[Handout 1 – Basic terminology 2](#_Toc26094019)

[Handout 2 - Introduction to SAS 4](#_Toc26094020)

[Handout 3 – Descriptive statistics (in SAS) 18](#_Toc26094021)

[Handout 4 – Road Map of STA 216 and intro to two-sample model 32](#_Toc26094022)

[Handout 5 - Review of Hypothesis Testing through the two-sample t test 42](#_Toc26094023)

[Handout 6 – Confidence Intervals 50](#_Toc26094024)

[Handout 7 – Extra practice on two-sample model 57](#_Toc26094025)

[Handout 8 – Hypothesis tests errors 63](#_Toc26094026)

[Handout 9: Rank-sum test 65](#_Toc26094027)

[Handout 10 Supplement 72](#_Toc26094028)

[Handout 11 – Assumptions in SLR and how to check them 79](#_Toc26094029)

[Handout 12 – Statistical Inference for simple linear regression 83](#_Toc26094030)

[Handout 13 – Introducing Multiple Linear Regression 90](#_Toc26094031)

[Handout 14 – Separate regression lines by a categorical variable 102](#_Toc26094032)

[Handout 15 – Comparing models fit with R, root MSE, and Adj. R 112](#_Toc26094033)

[Handout 16 – One-way ANOVA 122](#_Toc26094034)

[Handout 17 – Comparing means 129](#_Toc26094035)

[Handout 18 – Two-way ANOVA (no interaction) 134](#_Toc26094036)

[Handout 19 – Two-way ANOVA with interaction 141](#_Toc26094037)

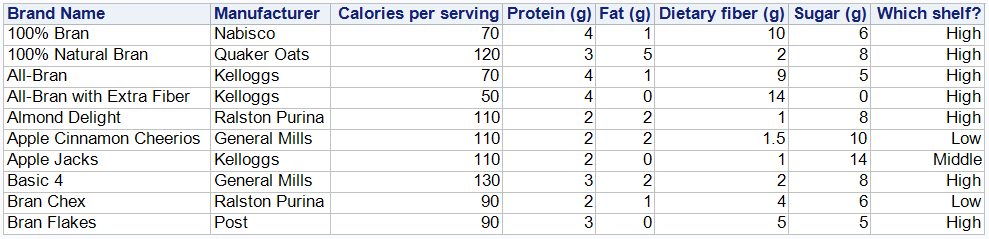
# Handout 1 – Basic terminology

STA 216

Basic vocab about datasets

* Individuals: who or what the information is about
* Variables: characteristics about the individuals that are recorded
* Observations: values of the characteristics
* In the standard format of a data table,
  + The individuals are the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ of the table
  + The variables are the \_\_\_\_\_\_\_\_\_\_\_\_\_\_ of the table
  + The observations are the \_\_\_\_\_\_\_\_\_\_\_\_\_ of the table

Example: Breakfast cereals



* Identify each of the underlined terms as an individual, variable, or observation:
  + The manufacturer of Apple Jacks is Kelloggs.
  + All-Bran with Extra Fiber has 14g of dietary fiber per serving.

Categorical/Quantitative Variables

* Categorical: assigns a category to individuals
* Quantitative: assigns a quantity to individuals
* What types are the variables above:  
  
* Generally,
  + categorical variables have observations that are letters/words
  + quantitative variables have observations that are numeric
* But there are exceptions:
  + Zip code, area code are categorical because these are not meaningful quantities

Explanatory/Response Variables

* Changes in an explanatory variable can explain changes in a response variable.
* Examples from cereal dataset:
  + Explanatory variable: sugar
  + Response variable: calories
* The explanatory variable is often labeled \_\_\_\_ and the response variable is labeled \_\_\_\_.

Example

Several groups of apple trees are sprayed with different amounts of pesticides (either high concentration, low concentration, or no pesticides). The resulting yield (in bushels of apples) of each group of apple trees is recorded.

* Individuals: groups of apples trees
* Two variables: Amount of pesticides, Yield
  + Explanatory variable: amount of pesticides
  + Response variable: Yield
* Is each variable categorical or quantitative?  
  Amount of pesticides is categorical; yield is quantitative
* Is “low concentration” an individual, variable, or observation?  
  An observation (it’s the value of the variable “amount of pesticides” for particular individuals – groups of apple trees)

# Handout 2 - Introduction to SAS

(Reference for Lab 1)

STA 216

Overview

* Comments
* Reading in data with PROC IMPORT
* PROC PRINT with VAR and WHERE statements
* DATA step and creating a new variable
* Labels for variables
* Formats for observations
* PROC SORT and BY statement

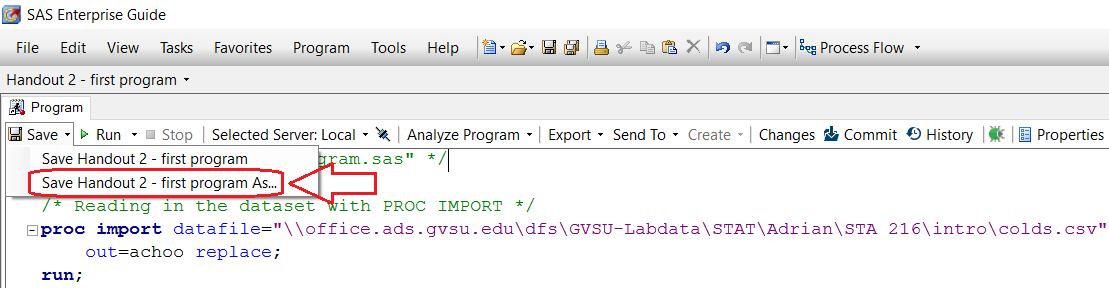
Getting Started:

* Go to our class’s Blackboard site 🡪 Handouts 🡪 Under Handout 2, click on “Handout 2 – first program.sas”
  + This will open SAS Enterprise Guide with the code displayed that is saved in the program “Handout 2 – first program.sas”
* Another way to open SAS Enterprise Guide 7.1:
  + Go to Lab Applications on Desktop 🡪 STATISTICS 🡪 SAS 9.4 🡪 SAS Enterprise Guide 7.1

Comments

* Note all the text in green on the program. These are comments.
* A comment is a piece of text that SAS ignores when it processes your program.
  + The main purpose of comments are to help explain what a part of your code does to yourself or someone else. (You can see that I’ve already added a lot of these.)
  + One way to make a comment in SAS is to put the text you want to be a comment between the start- and end-comment marks as follows: /\* This is a comment \*/
* Let’s add some comments at the top of the program: Under /\* "Handout 2 - first program.sas" \*/
  + Write a comment that says that this program is a good reference for Lab 1 (so you don’t forget that)
  + Under that, write a comment describing what your favorite thing in the world is.

Saving your program



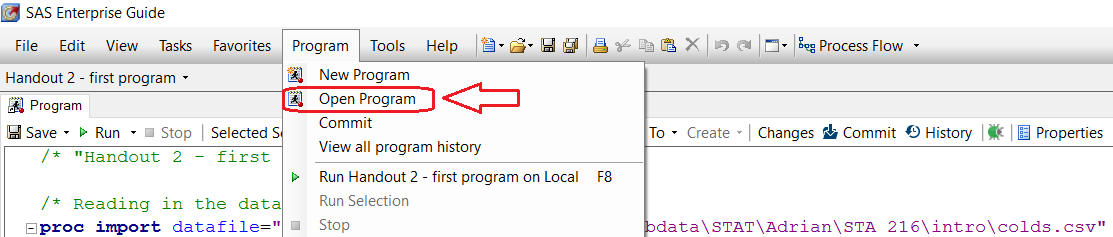
* Save it the network folder you usually use (feel free to take a minute to make a new folder for STA 216)

Reopen the program you saved

* Two ways:

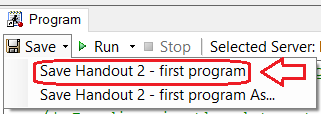
1. Navigate to the “Handout 2 – first program.sas” file using File Explorer and double-click it

2. Open SAS Enterprise Guide from Lab Applications 🡪 STATISTICS 🡪 SAS 9.4 🡪 SAS Enterprise Guide 7.1

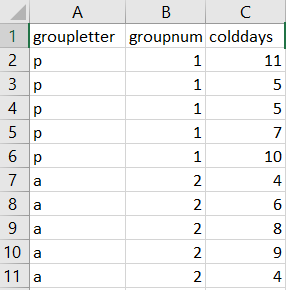
Use Program menu  


* Note: **Make sure you save and open SAS Programs and not SAS Projects!**
  + Projects are more advanced files that can incorporate several programs and data files together.
  + We won’t need them. Using them by mistake can make it hard to find saved work!
  + This is why I discourage using the File menu on the left.

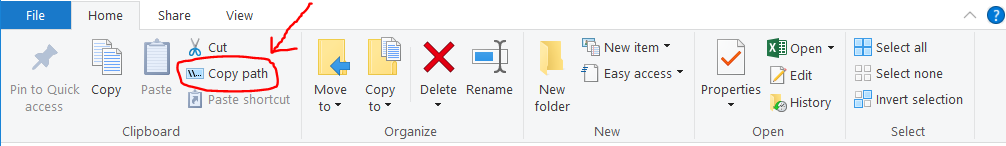
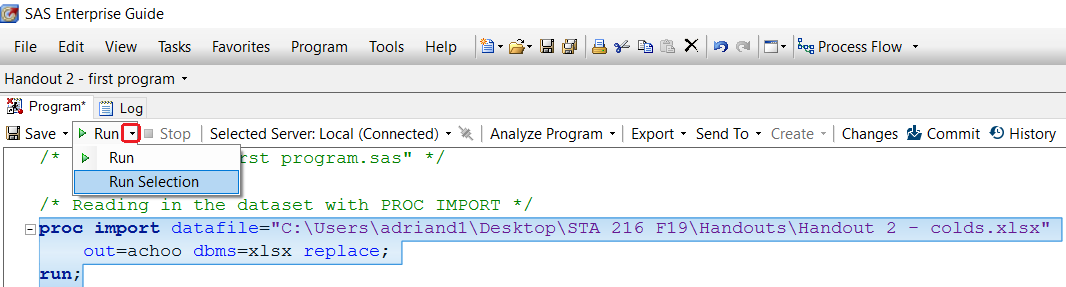
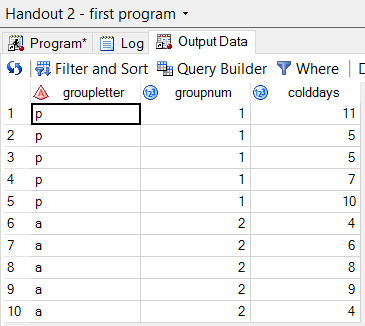
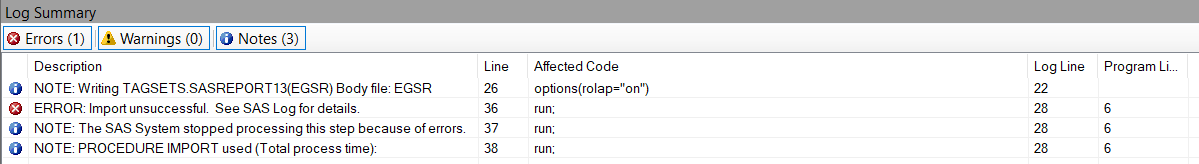
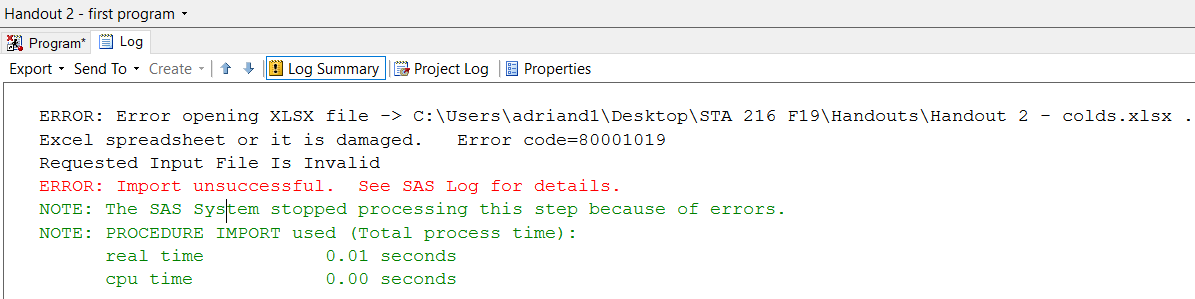
One more save/open practice

* Below your comment describing your favorite thing in the world, write a comment describing your least favorite thing in the world.
* Now that we have a name/folder established for the .sas program, save the program using
* I encourage you to do this frequently.
* Now close SAS EG and open the program again using one of the two methods of your choice.

Reading in the dataset with PROC IMPORT

* “PROC” stands for procedure and represent a group of SAS statements that do something.
* In this case, PROC IMPORT imports a dataset from an external file into SAS.
* Here, the external file is named “Handout 2 - colds.xlsx” and is on Blackboard under Handout 2.
* Save the file “Handout 2 - colds.xlsx” in the same folder as you saved the SAS program earlier.
* The dataset describes ten people with a cold, five who were given Alka-Seltzer and five who were given a placebo, and records how many days the cold lasted.
* Try opening the Excel file:  
  
* Note: **The Excel file must be closed** in order for PROC IMPORT to work.
* There are four parts of this PROC IMPORT:

|  |  |
| --- | --- |
| SAS code | Description |
| datafile="PATH\Handout 2 - colds.xlsx" | Gives the sequence of folders that contains the file |
| out=achoo | Names the SAS dataset named “achoo” |
| dbms=xlsx | Tells SAS this is an Excel file |
| replace | Replaces the current SAS dataset named “achoo” (if one exists) |

* Trick to get the path right (since it must be exactly right):
  + Use File Explorer to navigate to the “Handout 2 – colds.xlsx” file.
  + Use the “Copy Path” button:  
    
* Note:
  + You can name the SAS dataset whatever you want in the OUT= statement.
  + SAS is not case-sensitive or space sensitive, except inside quotes
  + The PROC IMPORT ends in a semicolon.
  + It is common to end each PROC with “RUN;”
* To run the PROC IMPORT, highlight the PROC IMPORT portion of the program, click on the black triangle of Run and click on “Run Selection”.  
  
  + Clicking on “Run” itself will run the entire program, which might give more output (or take longer) than you want.
* Result:  
  
* Note:
  + The SAS variable names “groupletter”, “groupnum”, and “colddays” were read in from the first line of the Excel file.
  + “groupletter” is a character variable (i.e. containing letters), as given by the icon 
  + “groupnum” and “colddays” are numeric variables (i.e. containing numbers), as given by the icon 
* Let’s try running PROC IMPORT while the Excel file is open to see what an error looks like:  
  Log Summary at bottom  
    
    
  Log window  
  

PROC PRINT:

* Highlight and run the first PROC PRINT portion of the code.
* Say “Yes”; you want to replace the results from previous run. (This pops up every run and is kind of annoying.)

|  |  |
| --- | --- |
| Code: | Output (results tab): |
| **proc** **print** data=achoo;  **run**; |  |

VAR Statement in PROC PRINT:

* allows you to only print some of the variables.
* To print out just the variables groupletter and colddays (not groupnum):

|  |  |
| --- | --- |
| Code: | Output: |
| **proc** **print** data=achoo;  var groupletter colddays;  **run**; |  |

WHERE Statement in PROC PRINT:

* allows you to only print individuals satisifying some criteria.
* In this case, let’s only print the individuals who took Alka-Seltzer

|  |  |  |
| --- | --- | --- |
| Code: | Explanation: | Output: |
| **proc** **print** data=achoo;  where groupletter = 'a';  **run**; | Only Alka-Seltzer observations are printed ->  (Note that the “a” needs to be in quotation marks for this to work because it is the value of a character variable.) |  |

More Examples of PROC PRINT with VAR and WHERE statements

|  |  |
| --- | --- |
| Code | Output |
| **proc** **print** data=achoo;  where groupnum = **1**;  var groupnum colddays;  **run**; |  |
| **proc** **print** data=achoo;  where colddays > **8**;  **run**; |  |
| **proc** **print** data=achoo;  where groupletter = 'a' and colddays <= **8**;  var groupletter colddays;  **run**; |  |

DATA steps and how to create new variables

* Suppose we want to add a variable for the number of hours the cold lasted and save the result in a new dataset.
* We could use the following DATA step (the lines 1-4 below)

|  |  |  |
| --- | --- | --- |
| **1**  **2**  **3**  **4**  **5**  **6** | **data** achoonew;  set achoo;  coldhrs = colddays \* **24**;  **run**;  **proc** **print** data=achoonew;  **run**; |  |

Explanation by line:

* Actually starts with line 2: the SET statement says to start from the dataset “achoo”
* Line 3: Defines the new variable coldhrs to be the variable colddays multiplied by 24. (+, -, \*, and / are for addition, subtraction, multiplication and division)
* Line 1: Stores the resulting dataset (the three variables from the old dataset “achoo” and the new variable coldhrs) and names it “achoonew”.
* Note: often, we will want the new dataset to replace the old one (i.e. delete the old one). In this case, the DATA and SET lines will have the same dataset name.

Labels for variables:

* SAS variable names are not allowed to have spaces and are often short for convenience.
* While this helps for coding, to make the SAS variable names more understandable, we can provide them with labels.

|  |  |  |
| --- | --- | --- |
|  | Code | Output |
| **1**  **2**  **3**  **4**  **5** | **data** achoo;  set achoo;  label groupletter = 'Drug Group Letter'  groupnum = 'Drug Group Number'  colddays = 'Number of days the cold lasted';  **proc** **print** data=achoo label;  where colddays < **10**;  **run**; |  |

Explanation by line:

* Line 2: Start with dataset achoo
* Line 3: Add labels for each variable (note that this is all one line of code – it ends at the semicolon)
* Line 1: The updated dataset achoo new contains the labels
* Line 4: the option LABEL must be included to display the labels
* Line 5: Only individuals with less than 10 cold days are printed. (Note that the variable name is used here, not its label)

Note that the variable names are replaced by the labels in the output.

Formats for observations:

* To make the observations (values of the variables) more understandable, we can use formats. For example, it would be clearer if we could see “Placebo” and “Group” on the output instead of 1 and 2.
* There are two steps to this:
  + Use PROC FORMAT to define what to replace each observation by
  + Use the FORMAT statement in a DATA step to define which variable each format applies to

1. PROC FORMAT

|  |
| --- |
| **proc** **format**;  value $chandler 'p' = 'Placebo' 'a' = 'Alka-Seltzer';  value joey **1** = 'PB' **2** = 'AS';  **run**; |

Note:

* One VALUE statement is used for the format corresponding to each variable. The format named “$chandler” corresponds with the variable groupletter and the format “joey” corresponds with the variable groupnum. (I named them after “Friends” characters to illustrate that they can be named anything you want.)
* $chandler says to show observations of “p” as “Placebo” and observations of “a” as “Alka-Seltzer”
* “$chandler” needs a dollar sign because it corresponds to a character variable; “joey” does not because it corresponds to a numeric variable

2. FORMAT statement in DATA step

|  |  |
| --- | --- |
| **1**  **2**  **3** | **data** achoo;  set achoo;  format groupletter $chandler. groupnum joey.;  **run**; |

Explanation by line:

* Line 2: Starts with the dataset achoo
* Line 3: FORMAT statement matches format $chandler with variable groupletter and format joey with groupnum (you need to follow the format names by periods here).
* Line 1: Save the dataset (with the added formats) as achoo (over the old one)

Result:

|  |  |
| --- | --- |
| **proc** **print** data=achoo label;  where groupletter = 'a';  **run**; |  |

* Note that the unformatted value of “a” is used in the WHERE statement.
* We may want to consider new labels considering the new formats.

PROC SORT

* Sorts by values of a variable – by default, from smallest to largest (or A to Z)

|  |  |
| --- | --- |
| **proc** **sort** data=achoo;  by colddays;  **proc** **print** data=achoo label;  **run**; |  |

* Or it can sort by descending values

|  |  |
| --- | --- |
| **proc** **sort** data=achoo;  by descending colddays;  **proc** **print** data=achoo label;  **run**; |  |

* Can sort by one variable first, then another

|  |  |
| --- | --- |
| **proc** **sort** data=achoo;  by groupletter colddays;  **proc** **print** data=achoo label;  **run**; |  |

Using Snipping Tool on your labs:

* On your labs, you will be required to copy and paste code and output from SAS into a Word document.
* To copy and paste output, you should use the Snipping Tool.
* I will demonstrate this.

Saving your SAS program and Exiting SAS

* Make sure to save your program before you exit.
* When you close SAS, it asks whether you want to save the project. Say no. (Remember we always deal with programs and never projects.)

# Handout 3 – Descriptive statistics (in SAS)

Guide for Lab 2

STA 216

* Descriptive statistics: numerical and graphical summaries of a dataset
* Inferential statistics:
  + Methods to use a sample to make conclusions about the larger population
  + Includes confidence intervals and hypothesis tests
  + Only makes sense to do if you have a **random** sample from the population
* We focus on descriptive statistics in this handout.
* In descriptive statistics, what numerical or graphical summaries you make depends on
  + How many variables you are summarizing: one (univariate), two (bivariate), or more than two (multivariate)
  + Type of variable: categorical or quantitative
* Here we’ll focus on univariate and bivariate descriptive statics (we’ll do multivariate in the next handout)

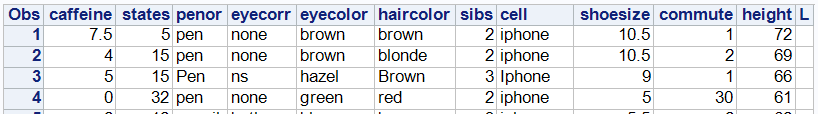
|  |  |  |
| --- | --- | --- |
| How many/type of variable | Numerical summaries | Graphical summaries |
| One categorical | Count, percent, proportion | Pie chart, bar chart |
| One quantitative | Mean, standard deviation, 5-number summary | Boxplot, histogram |
| Two categorical | Counts, percentages in two-way table | Clustered/stacked bar chart, Mosaic plot |
| Two quantitative | Correlation, equation of regression line | Scatterplot |
| Quantitative response, Categorical explanatory | Separate mean, sd, etc (same as for 1 quant) for each group defined by the categorical variable | Side-by-side boxplot or comparative histogram |

* I will show how to make each of these numerical and graphical summaries in SAS for a dataset I collected from my STA 215 students.
* Follow along with the SAS program “Handout 3 – descriptive statistics.sas”
* There is a summary table of the different SAS commands at the end of this handout.
* Note that inferential statistics (hypothesis tests or confidence intervals) aren’t appropriate for this dataset because this is not a random sample of all GVSU students or even of all STA 215 students.

Variables include:

* caffeine: How many caffeinated beverages do you drink per week?
* states: How many U.S. states have you been to?
* pencil: Overall, do you prefer to write in pen or pencil?
* eyecorr: What kind of eye correction do you have? -- near-sighted, far-sighted, both, or none
* eyecolor: What color are your eyes?
* yourhair: What color is your hair (naturally)?
* sibs: How many siblings do you have?
* cell: What cell phone brand do you have?
* shoesize: What is your shoe size? (Please subtract 1.5 for women’s shoes for equivalence)
* commute: How many miles is your commute to Grand Valley (Allendale)?
* height: What is your height in inches? (for example, 5’7” = 67 inches)

First four individuals:



Introduction

* Use PROC IMPORT to import the Excel file “sta 215 student data.xlsx” into SAS
* How to drop the variable “L”
* Adding labels
* Adding a format

One categorical variable

* Numerical summaries: count, percent, proportion
* Graphical summaries: pie chart, bar chart

**Table of frequencies (another word for counts) and percentages with PROC FREQ**

|  |  |
| --- | --- |
| **proc** **freq** data=sta215;  table eyecorr;  **run**; |  |
| **proc** **freq** data=sta215;  table eyecorr /nocum;  **run**; |  |

The option NOCUM suppresses the printing of cumulative frequencies and percents in the table. (For example, the cumulative frequency for far-sighted is 7 because it’s adding together the frequencies for both and Far-sighted.)

**Bar chart: PROC SGPLOT with VBAR**

(We won’t make pie charts in this class)

|  |  |
| --- | --- |
| **proc** **sgplot** data=sta215;  vbar eyecorr;  **run**; | img0.png |

VBAR is for vertically oriented bars. HBAR will give horizontally oriented bars.

One quantitative variable

* Numerical summaries: Mean, standard deviation, 5-number summary
* Graphical summaries: Boxplot, histogram

**PROC MEANS for mean, sd, and other numerical summaries**

|  |  |
| --- | --- |
| **proc** **means** data=sta215;  **run**; |  |

By default, for each numeric variable in the dataset, PROC MEANS reports the sample size (N), the mean, standard deviation (Std Dev), the minimum, and maximum

|  |  |
| --- | --- |
| **proc** **means** data=sta215;  var states;  **run**; |  |

The VAR statement can be used to show only the variables desired.

|  |  |
| --- | --- |
| **proc** **means** data=sta215  n mean std min q1 median q3 max;  var states;  **run**; |  |

Statistic keywords can be added to the PROC MEANS line to specify which statistics are desired. These include: max, mean, min, mode, n, range, std (standard deviation), sum, var (variance), median, p10 (10th percentile), q1 (first quartile), q3, and qrange (interquartile range)

|  |  |
| --- | --- |
| **proc** **means** data=sta215 mean std maxdec=**1**;  var states;  **run**; |  |

The MAXDEC option is used to specify how many decimal places are reported.

**Histogram: PROC SGPLOT with HISTOGRAM command**

|  |  |
| --- | --- |
| **proc** **sgplot** data=sta215;  histogram caffeine;  **run**; | img0.png |
| **proc** **sgplot** data=sta215;  histogram caffeine /scale=count;  **run**; | img1.png |

By default, the histogram has percent on the y-axis. Adding the option SCALE=count makes it so counts are on the y-axis.

**Boxplot: PROC SGPLOT with VBOX or HBOX**

|  |  |
| --- | --- |
| **proc** **sgplot** data=sta215;  hbox caffeine;  **run**; | img0.png |

Note:

* HBOX is for a horizontally oriented boxplot. VBOX is for a vertical one.
* The diamond is the mean.
* The circles are outliers; greater than Q3 + 1.5\*IQR

Two Categorical Variables

* Numerical summaries: Counts, percentages in two-way table
* Graphical summaries: Clustered/stacked bar chart, Mosaic plot

**Two-way table of counts/percentages with PROC FREQ**

|  |  |
| --- | --- |
| **proc** **freq** data=sta215;  table eyecolor \* haircolor;  **run**; |  |

Note:

* The first variable in the TABLE statement has its categories as the rows and the second variable is in the columns.
* According to the red box, 6 students (reported that they) had blue eyes and blonde hair. According to the green box, 9 students had blue eyes (out of all the hair colors). According to the purple box, 11 students had blonde hair.
* By default, each combination of two categories has 4 entries: frequency, percent, row pct, col pct. Focusing on the combination with blue eyes and blonde hair (red box):
  + The 6 for frequency is the count of students that have blue eyes and blonde hair
  + The 19.35 for percent is the percent out of all students that have blue eyes and blonde hair, which is 6 out of the total of 31 students (the 31 is from the lower right corner).
  + The 66.67 for row percent is the percent out of students in the row (or out of students with blue eyes) that have blue eyes and blonde hair, which is 6 out of 9
  + The 54.55 for column percent is the percent out of the students in the column (or out of students with blonde hair) that have blue eyes and blonde hair, which is 6 out of 11.

|  |  |
| --- | --- |
| **proc** **freq** data=sta215;  table eyecolor \* haircolor  /norow nocol nopercent;  **run**; |  |

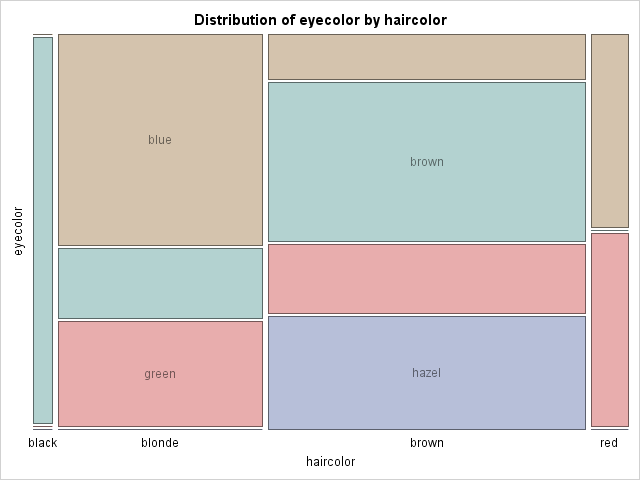
The table above only shows the frequencies because the options NOROW, NOCOL, and NOPERCENT suppress the printing of the row, column, and overall percentages, respectively.

**Stacked and clustered bar charts with PROC SGPLOT with VBAR and GROUP option**

|  |  |
| --- | --- |
| /\* Stacked bar chart \*/  **proc** **sgplot** data=sta215;  vbar haircolor /group=eyecolor;  **run**; | img0.png |
| /\* Clustered bar chart \*/  **proc** **sgplot** data=sta215;  vbar haircolor /group=eyecolor  groupdisplay=cluster;  **run**; | img1.png |

**Mosaic plot with ODS GRAPHICS and PROC FREQ**

|  |  |
| --- | --- |
| ods graphics on;  **proc** **freq** data=sta215;  table eyecolor \* haircolor /norow  nocol nopercent plots=mosaicplot;  **run**;  ods graphics off; |  |



Note:

* It is necessary to “turn on” ODS GRAPHICS to get some graphs from SAS. ODS is the output delivery system.
* A mosaic plot is like a graphical depiction of the two-way table.
  + The width of the different vertical rectangular sections for black, blonde, brown, and red hair is proportional to their counts in the dataset (the total counts 1, 11, 17, and 2 at the bottom of each column in the table.)
  + Each of the vertical sections are divided according to the counts of eye color within each hair color.

Two quantitative variables

* Numerical summaries: correlation and equation of regression line
* Graphical summary: scatterplot (with or without regression line)

**Scatterplot: PROC SGPLOT with SCATTER statement**

|  |  |
| --- | --- |
| **proc** **sgplot** data=sta215;  scatter x=shoesize y=height;  **run**; | img0.png |

Use x= for the variable to go on the x-axis; y= for the y-axis.

**Scatterplot with regression line: PROC SGPLOT with REG statement**

|  |  |
| --- | --- |
| **proc** **sgplot** data=sta215;  reg x=shoesize y=height;  **run**; | img0.png |

**Calculate correlation: PROC CORR**

|  |  |
| --- | --- |
| **proc** **corr** data=sta215;  var shoesize height;  **run**; |  |

Note:

* Correlation of 0.79432 is circled
* PROC CORR gives a lot of univariate summary statistics in the top table

**Calculate equation of regression line**: PROC REG

We will consider predicting height from shoesize so x=shoesize and y=height

|  |  |
| --- | --- |
| **proc** **reg** data=sta215;  model height = shoesize;  **run**; |  |

Note:

* The MODEL height = shoesize statement has to be given in the order “y = x”. This is important and you will get the wrong result if you reverse the variables.
* The slope and intercept of the regression line are given in the circled part of the output
* Intercept = 55.0 and slope = 1.4 so the equation of the regression line of

Quantitative response and categorical explanatory variables

* Numerical summaries: different mean, sd, etc for individuals in each group
* Graphical summaries: Side-by-side boxplot or comparative histogram

**Separate means, sds, etc. with PROC MEANS and CLASS statement**

|  |  |
| --- | --- |
| **proc** **means** data=sta215  mean std maxdec=**1**;  class penor;  var caffeine;  **run**; |  |

Note:

* The quantitative variable goes under the VAR statement
* The categorical variable goes under the CLASS statement
* Interpretation: the average caffeinated drinks per week for the 17 students that use a pen was 5.4; for the 14 students that use a pencil the average was 3.2 caffeinated drinks per week.

**Side-by-side boxplots with PROC SGPLOT, VBOX, and CATEGORY option**

|  |  |
| --- | --- |
| **proc** **sgplot** data=sta215;  vbox caffeine  /category=penor;  **run**; | img0.png |

**Comparative histograms with PROC SGPANEL**

|  |  |
| --- | --- |
| **proc** **sgpanel** data=sta215;  panelby penor;  histogram caffeine;  **run**; | img0.png |

Summary of SAS commands

* One categorical
  + Numerical summaries: PROC FREQ for table of counts and percents
  + Graphs: PROC SGPLOT with HBAR/VAR for bar charts
* One quantitative
  + Numerical summaries: PROC MEANS for mean, sd, etc.
  + Graphs: PROC SGPLOT with HBOX/VBOX for boxplot  
     with HISTOGRAM for histogram
* Two categorical
  + Numerical summaries: PROC FREQ for two-way table of counts and percents
  + Graphs: PROC SGPLOT with VBAR and GROUP for stacked bar chart  
     with VBAR, GROUP, and GROUPDISPLAY=cluster for clustered bar chart  
    PROC FREQ with ODS GRAPHICS for mosaic plot
* Two quantitative
  + Numerical summaries: PROC CORR for correlation  
     PROC REG for equation of regression line
  + Graphs: PROC SGPLOT with SCATTER for scatterplot  
     with REG for scatterplot with regression line
* Quantitative response and categorical explanatory
  + Numerical summaries: PROC MEANS with CLASS statement
  + Graphs: PROC SGPLOT with HBOX/VBOX and CATEGORY for side-by-side boxplot  
     PROC SGPANEL with PANELBY, HISTOGRAM for comparative histograms

# Handout 4 – Road Map of STA 216 and intro to two-sample model

STA 216

* Focus on the statistical model
* Definition: A statistical model describes the assumptions about the distribution of the data
* All the models in the class will follow the form

Response = Mean + Error

Y = f(X) + E

* The Mean depends on the explanatory variables X
* The Error is the part of the response variable that remains unexplained after considering X
* The model depends on the type of data (categorical or quantitative).
* Some models for bivariate (two-variable) data:

Two-sample model  
Y: quantitative  
X: categorical with 2 categories

Model: the mean of Y is determined by group membership

Simple linear regression model

Y: quantitative

X: quantitative

Model: the mean of Y is determined by a linear function of X

One-way ANOVA model

Y: quantitative

X: categorical with more than 2 categories

Model: the mean of Y is determined by group membership

* Later in the semester, models for multivariate data (more than one explanatory variable)
  + Multiple regression
  + Two-way ANOVA
* First: two-sample model

Two-sample model example

Losing weight is an important goal for many individuals. An article in the *Journal of the American Medical Association* describes a study in which researchers investigated whether financial incentives would help people lose weight more successfully. Some participants in the study were randomly assigned to a treatment group that offered financial incentives for achieving weight loss goals, while others were assigned to a treatment group that did not use financial incentives. All participants were monitored over a 4-month period and the net weight change (Beginning – End in pounds) was recorded for each individual. Note that a positive value corresponds to a weight loss and a negative value is a weight gain. The data are given below:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Control | 12.5 | 12.0 | 1.0 | -5.0 | 3.0 | -5.0 | 7.5 | -2.5 | 20.0 | -1.0 |
|  | 2.0 | 4.5 | -2.0 | -17.0 | 19.0 | -2.0 | 12.0 | 10.5 | 5.0 |  |
| Incentive | 25.5 | 24.0 | 8.0 | 15.5 | 21.0 | 4.5 | 30.0 | 7.5 | 10.0 | 18.0 |
|  | 5.0 | -0.5 | 27.0 | 6.0 | 25.5 | 21.0 | 18.5 |  |  |  |

Two variables:

* Weight change (quantitative)
* Group (categorical): control or incentive

Notation

Y = weight change (response)

i = group (i=1 for control, i=2 for incentive)

population mean weight change for group i

population standard deviation of weight change for group i

Model

* In words:
  + For individuals in the control group, the weight change follows a normal distribution with mean and standard deviation . [Or for short, .]
  + For individuals in the incentive group, the weight change follows a normal distribution with mean and standard deviation . [Or for short, .]
* In Response = Mean + Error form, the two-sample model is written as   
    
  Y = ,  
    
  where the errors .
* The model also assumes that the weight loss of the individuals are **independent**.

Four-step Process to Modeling

* **Choose** model: specify the model (like we just did)
* **Fit** the model: estimate parameters based on the data
* **Assess** how well the model describes the data.
* **Use** the model (make predictions, explain relationships, assess differences, statistical inference)

FIT for two-sample model

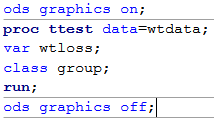
* The parameters are the population mean weight changes and the population standard deviations of weight change .
* Their estimates are the sample means and the sample standard deviations .
* We learned to calculate these with PROC MEANS.
* Here’s some code from the file “handout 4.sas”:

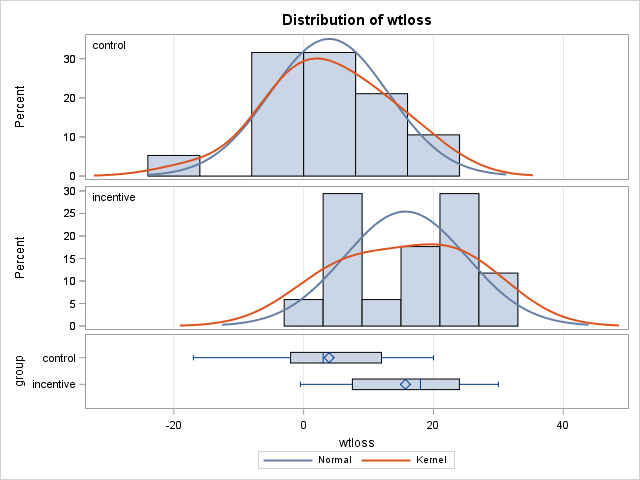
|  |  |
| --- | --- |
|  |  |

ASSESS

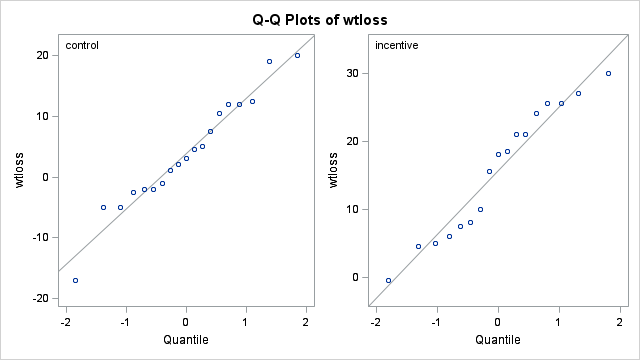
Is it reasonable to assume the weight changes in each group follow a normal distribution?

* The default ODS GRAPHICS from PROC TTEST (which we’ll use shortly to do the 2-sample t test) gives side-by-side boxplots, histograms, and normal quantile plots.





* We’re looking for the histograms to be roughly bell-shaped and the Kernel and normal density curves (two curves plotted over each histogram) to be close to each other.
* The boxplots should be roughly symmetric.



* On the normal quantile plots (Q-Q Plots), we are looking for the points to follow the line (and not have a systematic curve).

USE

Statistical inference: using a sample to make a conclusion about the population

* Should only be used if the data is a random sample from a population
  + It should not be used if the data is already the population of interest. (Then you can calculate population parameters exactly!)
  + It should not be used if the data is a sample, but not a random one. (Then the sample is likely not representative of the population.)
* Next handouts: Hypothesis tests and confidence intervals in the two-sample setting.

Handout 4 Supplement

Calculating and interpreting the standard deviation

Formula:

Steps for calculating the standard deviation (s)

* 1. Calculate the mean
  2. Compute the deviations of each observation as “observation minus mean”,
  3. Square each deviation,
  4. Sum the squared deviations,
  5. Divide the sum of squared deviations by the sample size minus 1 (n-1). This gives the variance (s2), .
  6. Take the square root of the variance to get the standard deviation.

Interpretation

* Variance is average squared distance from observations to the mean
* Standard deviation is the average distance from observations to the mean.
* For example, sd = 9.4 for the incentive group means that the average distance from the weight loss measurements to the mean weight loss is 9.4 lb.

“By-hand” calculation with SAS

Calculates standard deviation of weight loss for individuals in the incentive group.

/\* Makes a new dataset "incgp" which only contains observations in the incentive group.

The IF statement says to only include these observations \*/

**data** incgp;

set wtdata;

if group='incentive';

**run**;

**proc** **print** data=incgp;

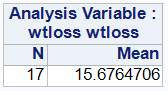
**run**;

/\* Finds sample size and mean weight loss for incentive group \*/

**proc** **means** data=incgp n mean;

var wtloss;

**run**;



/\* Calculates deviations (dev) and squared deviations (devsq) for each observation \*/

**data** incgp;

set incgp;

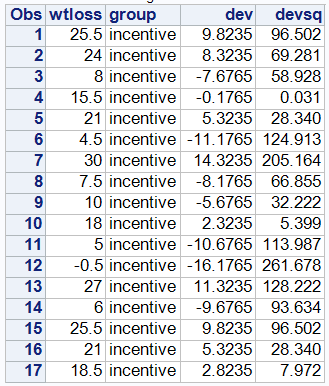
dev = wtloss - **15.67647**;

devsq = dev \* dev;

**run**;

**proc** **print** data=incgp;

**run**;

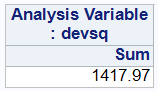


/\* Sums the squared deviations \*/

**proc** **means** data=incgp sum;

var devsq;

**run**;



/\* Uses SAS as a calculator to divide by (n-1) and take square root \*/

**data** calc;

variance = **1417.97** / **16**;

sd = sqrt(variance);

**run**;

**proc** **print** data=calc;

**run**;



# Handout 5 - Review of Hypothesis Testing through the two-sample t test

Five steps of hypothesis test

1. Write Ho and Ha in notation and define that notation in the context of the problem.
2. Calculate test statistic t.
3. Calculate p-value.
4. Decision (about Ho)
5. Conclusion (about Ha)

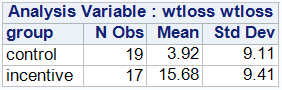
Example: Comparing weight loss among adults given incentives and not given incentives.

1. Null and alternative hypotheses

* Null hypothesis (Ho) and Alternative hypothesis (Ha)
* Both are written in terms of parameters (numbers calculated from the population)
  + For the two-sample t test, these are the population means and .
  + For our example,   
     = mean weight loss for all adults not given incentives  
     = mean weight loss for all adults given incentives
  + Doesn’t matter which is and .
* Ha says that the effect the researchers are looking for is present; Ho says that effect is absent.
* For the two-sample t test, there are three possibilities:
  + vs.
  + vs.
  + vs.
* The weight loss study includes the description “researchers investigated whether financial incentives would help people lose weight more successfully”,   
  so hypotheses are   
   vs. ,  
  where population mean weight loss over 4-month period **without** incentives  
   population mean weight loss over 4-month period **with** incentives

2. Test statistic

* Definition: a number that summarizes the distance between the observed result and what is expected to be observed if Ho is true.
* The test statistic for the two-sample t test is .
* The denominator is called the standard error, which estimates the standard deviation of if we repeated the experiment over and over.



* In this case, , , ,   
  (note that the group numbering 1 & 2 has to match the hypotheses).
* Test statistic

3. P-value

* Hypothesis testing logic (“Proof by contradiction”):
  + We want to show there is an effect (like the incentive increasing weight loss)
  + So we assume there is not an effect (i.e. assume that Ho is true)
  + And under this assumption, calculate the probability of observing the data 🡪 this is the p-value
  + If the p-value is low, it is unlikely under Ho that we observed the data. This is evidence that contradicts our assumption of Ho.
  + So we reject Ho and conclude that we have evidence of the effect (of Ha).
* So:
  + The p-value is a probability (that’s what “p” stands for).
  + **P-value: probability of observing the data if Ho is true**
* To calculate the p-value:
  + We need to think about (hypothetically) repeating the experiment over and over again.
  + Each experiment would have different weight loss values (and therefore different means and sd’s) which would make different test statistics.
  + The p-value is the probability of observing the test statistic (or one more extreme in the direction of Ha) if Ho is true.
  + If we assume Ho is true (and some other things we’ll talk about later), the test statistic follows the t distribution.
* The t distribution has degrees of freedom .
  + This formula is called the Satterthwaite approximation.
  + Calculate in two steps:
    - Calculate and .
    - Calculate

For our example,

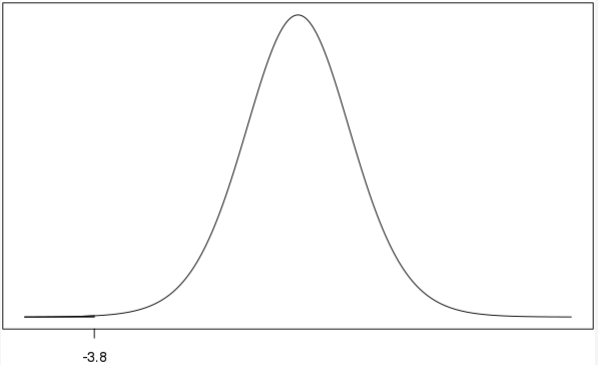
a = 9.112 / 19 = 4.3680 and b = 9.412 / 17 = 5.2087

* On your calculator, the function tcdf (2nd VARS) can be used to calculate the p-value. It has arguments:

tcdf(lower, upper, df).

* Use lower = -9999 for negative infinity and lower = 9999 for positive infinity

Picture:



LOWER = -9999, UPPER = -3.8, df = 33.28

p-value = tcdf(-9999, -3.8, 33.28) = 2.9E-4 = .00029

4. Decision (about Ho)

* By Comparing the p-value to the significance level .
* The most common value of the significance level is .05 historically, but it can be different. In this class, you can assume that on problems unless it says otherwise.
* How to make the decision:
* If p-value , decision: Reject Ho
* If p-value , decision: Fail to reject Ho

For the weight loss example, p-value = .0007 .05 🡪 Decision: Reject Ho

Reasoning about why we reject / fail to reject Ho

* When we do a hypothesis test, we assume Ho is true.
* Informally, the p-value is the probability of observing the data if Ho is true.
* So if p-value , there is a **small** chance of observing the data if Ho is true. Thus, the data leads us to reject Ho.
* On the other hand, if p-value , there is a **“not small”** chance of observing the data if Ho is true. The sample isn’t unlikely enough assuming Ho to reject Ho so our decision is to fail to reject Ho.
* Note that we don’t say that we “accept Ho” when p-value . This is because we’re not saying that have evidence in favor of Ho, just that we don’t have strong enough evidence against Ho.

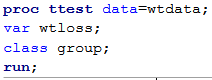
5. Conclusion

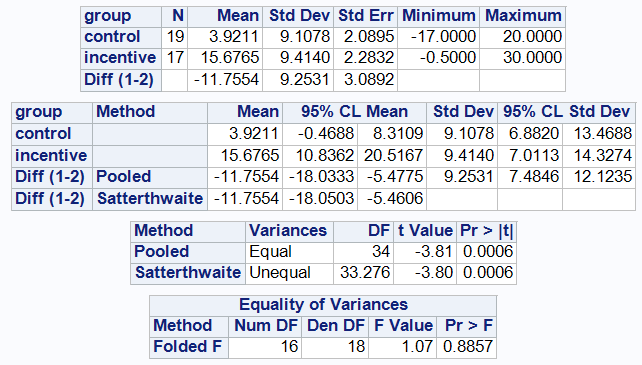
* While the decision about Ho, the conclusion is about the alternative hypothesis (Ha) and whether we have statistically significant evidence to support it.  
  + Reject Ho 🡪 conclusion: there **is** statistically significant evidence that Ha in the context of the problem.
  + Fail to reject Ho 🡪 conclusion: there **is not** statistically significant evidence that Ha in the context of the problem.
* You also need to start your conclusion with “At the \_\_\_\_\_\_\_ significance level, …”
* In writing “Ha in the context of the problem”, you might as well copy from your definitions of and in step 1.

For exercise example: At the alpha=.05 significance level, there is statistically significant evidence that the population mean weight loss after 4 months with no incentives is less than the population mean weight loss after 4 months with incentives.

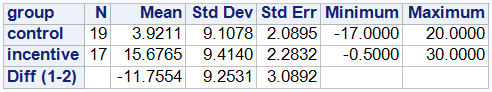
* Note: Although we make binary conclusions in this class (there is / is not statistically significant evidence), it is important to note that the amount of evidence is a continuum based on p-value.
* For example,
  + p-value = .051 represents about the same evidence as p-value = .049 (even though   
    .051 > .05 and .049 < .05)
  + p-value = .0001 represents much stronger evidence than p-value = .049 (even though both are less than .05)

2-sample t test in SAS: PROC TTEST





Top table



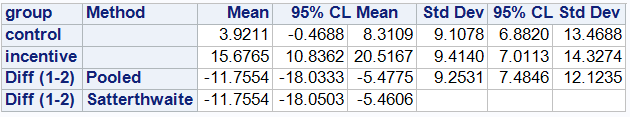
Gives

, ,

,

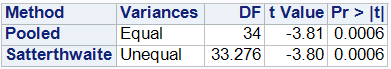
Standard error = denominator of test statistic =

Next table



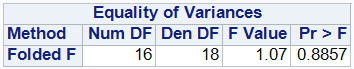
This gives confidence intervals which we haven’t covered yet.

Next table



* The row for Method: Pooled is based on the model assumption of equal population variances (i.e. ). We did not make this assumption in our two-sample model and when calculating the test statistic and p-value.
* So we will use the row for Method: Satterthwaite which is based on the model assumption of unequal population variances. In this row, we see
  + The DF = 33.28 value we calculated
  + The test statistic value t = -3.80
  + By default, SAS gives the p-value for the two-sided test. In this case, we divide this p-value by 2 to get the p-value we calculated for our one-sided test (.0006 / 2 = .0003).

Bottom Table

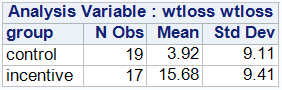


* This is a hypothesis test of Ho: vs. Ha:.
* With the p-value = .886 > .05, we fail to reject Ho 🡪 There is not statistically significant evidence that the population variances are different.
* This could be used to argue that the model assumption is reasonable.
* But there is not really much of an advantage to this assumption and we will never make it in this class.
* So effectively we won’t use this table.

# Handout 6 – Confidence Intervals

STA 216

Statistical significance does not always mean practical significance

* Recall: for weight loss data,  
  
  + Test statistic
  + P-value = tcdf(-9999, -3.8, 33.28) =.0003
* Suppose we decreased the difference between the mean by a factor of 10 and increased the sample sizes by a factor of 100. That is,
  + Made , ,
  + Made so
  + Test statistic
  + P-value = tcdf(-9999, -3.81, 3525.3) =.00007
  + Test statistic is exactly the same,
  + p-value is about half as before 🡪 our conclusion as far as having statistically significant evidence doesn’t change
* A problem with p-values and statistical significance is that they give no information about the **effect size** (in this case, the difference between the means).
  + In one case, the difference is 11.8 lb, which is practically significant (a significant size in real life).
  + In the other case, the difference is 1.2 lb, which is not practically significant (within the variation of day-to-day weight measurements).
  + But both are “statistically significant” differences.
* Confidence intervals are better than hypothesis tests because they estimate the effect size.

Confidence Intervals in two-sample case

* Range of numbers that we believe contains the difference in population means with a certain level of confidence.
  + Written as (Lower, Upper)
  + Most common level of confidence = 95%
* Formula:
* The “t-multiplier”
  + is NOT the same as the test statistic t we have been calculating so far.
  + It is the value of t distribution with so the confidence level (like 95%) is contained between and .
  + Picture:
* As the confidence level increases, the value of t\* increases.
* Some values of t\* are given on the table on the last page.

Hand calculation of CI for for weight loss example

,

t\* = 2.042 from table (df=30 on table is closest to the actual df=33.28).

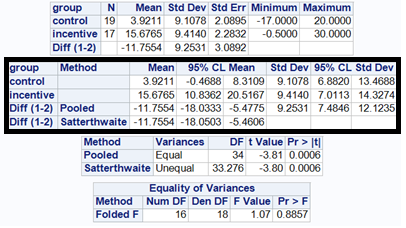
|  |
| --- |
| Accurate t\* value from SAS: |

= (-11.76 – 6.30, -11.76 + 6.30) = (-18.06, -5.46)

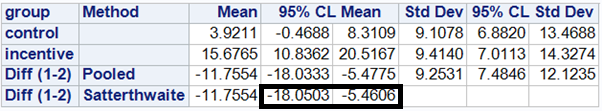
* For altered values , , ,   
  CI =   
   = (-1.79, -0.57)
* Mostly, we will use PROC TTEST to calculate confidence intervals for .

PROC TTEST OUTPUT

* The CIs are given in the second table:



* The CI for is given in the Diff(1-2), Satterthwaite row and the 95% CL Mean columns



* The ALPHA option can be used to change the confidence level.   
  Use ALPHA = 1 – confidence level.  
  For example, for a 99% confidence interval,

|  |  |
| --- | --- |
| **proc** **ttest** data=wtdata alpha=**.01**;  var wtloss;  class group;  **run**; |  |

Interpretation

* We are 95% confident that is in the range from -18.05 to -5.46 pounds.
* In the context of the problem, we are 95% confident that the population mean weight loss over a 4-month period for people given incentives is between 5.46 and 18.05 pounds greater than for people not given incentives.
* In general, 3 cases:
* Contains zero: ( - , +): we cannot be \_\_\_ % confident that and are different.
* Negative: ( - , - ): We are 95% confident that is less than by between lower and upper (use positive numbers).
* Positive: CI is ( + , + ): We are \_\_\_% confident that is greater than by between lower and upper.

What does the confidence level mean?

* When we say we are 95% confident that is in the confidence interval, it means that if we took tons of samples and made a confidence interval based on each sample, 95% of the intervals would contain the true value of .
* In other words, it’s the capture probability of the process that makes the interval.

You can do a test with a CI!

* Only for a specific test:
  + A two-sided hypothesis test
  + with significance level (for example 95% with alpha=.05)

Two cases:

* If the CI for does not contain zero 🡪 reject   
  Reasoning: Zero isn’t a plausible value for population mean difference.
* If the CI for contains zero 🡪 fail to reject .  
  Reasoning: Zero is a plausible value for the pop mean difference.

Examples for weight loss data:

* 95% CI for = (-18.05, -5.46)  
  Reject at alpha=.05 level
* 99% CI for = (-20.21, -3.30)  
  Also can reject at alpha=.01 level
* 95% CI for for data with one-tenth difference in means, 100 times sample sizes:  
  (-1.79, -0.57)  
  Reject at alpha=.05 level  
  But you can see from the CI that the actual difference in mean weight loss is much smaller.

Table for finding t\* (for confidence intervals)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Confidence Level | | | | | |
| df | 80% | 90% | 95% | 98% | 99% | 99.8% |
| 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 | 318.309 |
| 2 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 22.327 |
| 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 10.215 |
| 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173 |
| 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.893 |
| 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208 |
| 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.785 |
| 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 4.501 |
| 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.297 |
| 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.144 |
| 11 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.025 |
| 12 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.930 |
| 13 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.852 |
| 14 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.787 |
| 15 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.733 |
| 16 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.686 |
| 17 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.646 |
| 18 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.610 |
| 19 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.579 |
| 20 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.552 |
| 21 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.527 |
| 22 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.505 |
| 23 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.485 |
| 24 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.467 |
| 25 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.450 |
| 26 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.435 |
| 27 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.421 |
| 28 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.408 |
| 29 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.396 |
| 30 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.385 |
| 40 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 | 3.307 |
| 50 | 1.299 | 1.676 | 2.009 | 2.403 | 2.678 | 3.261 |
| 60 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 | 3.232 |
| 80 | 1.292 | 1.664 | 1.990 | 2.374 | 2.639 | 3.195 |
| 100 | 1.290 | 1.660 | 1.984 | 2.364 | 2.626 | 3.174 |
| (z\*) | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 3.090 |

# Handout 7 – Extra practice on two-sample model

“Proper treatment will cure a cold in seven days, but left to itself a cold will hang on for a week” ([Darrell Huff](https://www.goodreads.com/author/show/28848.Darrell_Huff), [How to Lie with Statistics](https://www.goodreads.com/work/quotes/415346)). Suppose a person working for Bayer wants to show that Alka-Seltzer is effective in reducing the length of a cold. To do so, she finds 10 people with colds and randomly assigns them to two groups of 5 people:

* Placebo group (group #1): subjects daily take a placebo tablet until cold is over
* Treatment group (group #2): subjects daily takes an Alka-Seltzer tablet until cold is over

The number of days of the cold recorded for each subject are as follows:

Placebo group: 11, 5, 5, 7, 10 Treatment group: 4, 6, 8, 9, 4

Two-sample model

, with , what is represented by

* 1. Y?
  2. ?
  3. ?
  4. ?
  5. ?
  6. What does mean? (In other words, how does it translate into words?)

What about the following SAS output that indicates that the normality assumption is reasonable for this data?

ods graphics on;

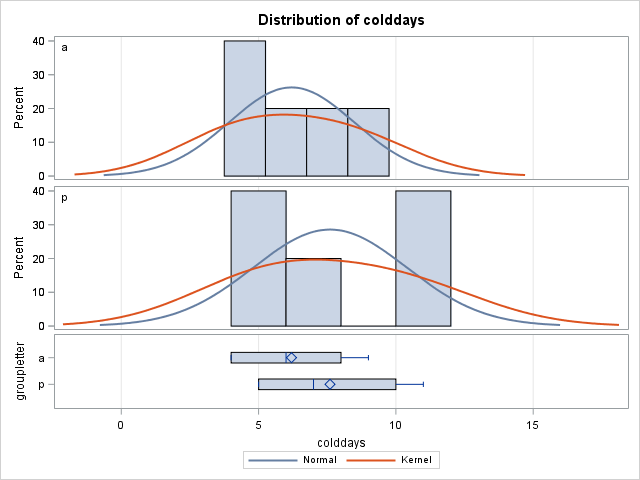
**proc** **ttest** data=achoo;

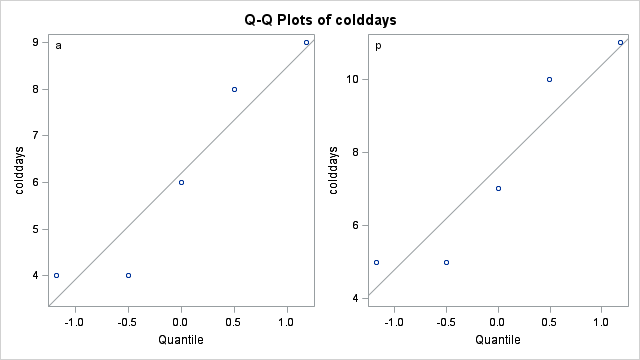
var colddays;

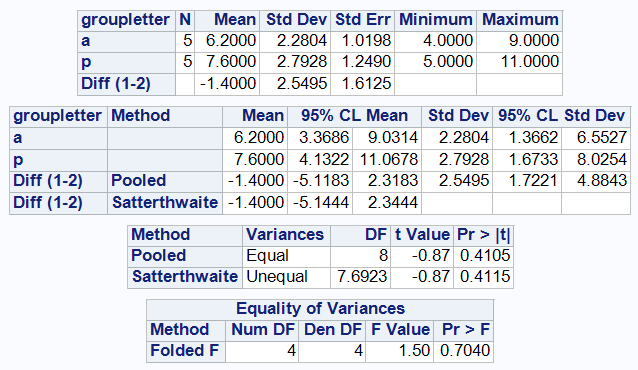
class groupletter;

**run**;

ods graphics off;





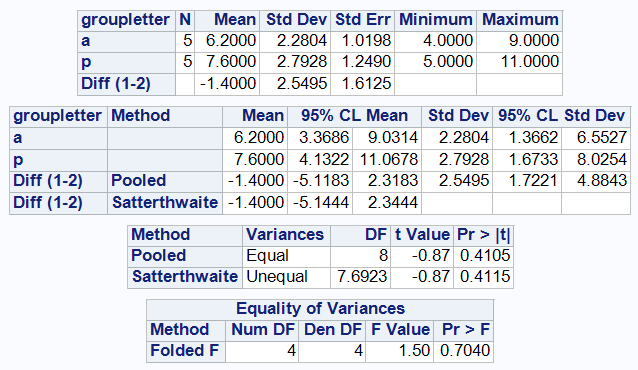


Hypothesis test

Show the five steps of a hypothesis test:

1. Write Ho and Ha in notation and define that notation in the context of the problem.
2. Calculate test statistic t.
3. Calculate p-value and draw a picture of it.
4. Decision (about Ho)
5. Conclusion (about Ha)

Confidence interval



Report the confidence interval we’re interested from the output.

What quantity is this confidence interval for?

Interpret it within the context of the problem.

What hypothesis test can the CI be used to do (what hypotheses and alpha)?

What is the result (decision and conclusion)? What about the CI tells you this?

# Handout 8 – Hypothesis tests errors

STA 216 F19

Errors in Hypothesis Tests:

* The conclusion of a hypothesis test comparing two groups is about and , the **population** means.
* Because we don’t have information from the entire population (only a sample), we don’t know for sure whether our conclusion is correct, and errors can be made.
* Actually, there are two different kinds of errors (Type 1 and 2 errors). What are they?
* Let’s consider a hypothesis test with Ho: vs. Ha:
* There are two possibilities of the (unknown) truth:
  + Ho: is true (equivalently, Ha: is false)
  + Ho: is false (equivalently, Ha: is true)
* There are two possible decisions/conclusions:
  + Reject Ho 🡪 There is statistically significant evidence that Ha is true
  + Fail to reject Ho 🡪 There is statistically significant evidence that Ha is true
* Thus, there are four possibilities overall:

|  |  |  |  |
| --- | --- | --- | --- |
|  | | The Truth! | |
| Ho is true | Ho is false |
| Decision in the  Hypothesis Test | Reject Ho | 1. Type 1 error | 2. Correct decision ☺ |
| Fail to reject Ho | 3. Correct decision ☺ | 4. Type 2 error |

* #1 is called a **Type 1 Error** or a **false positive** (“false”: because it is an error; “positive”: because the conclusion is that there is statistically significant evidence of a difference in population means).
* #4 is called a **Type 2 Error** or a **false negative** (“false”: because it is an error; “negative”: because the conclusion is not that there is not statistically significant evidence).
* Also, #2 is called a true positive; #3 is called a true negative.

Example

A good analogy for hypothesis testing is a criminal trial in the American judicial system. Because the American judicial system believes in a “presumption of innocence” (and we assume Ho is true when we do a hypothesis test), the hypotheses are:

Ho: the defendant is innocent Ha: the defendant is guilty

and the verdict in the trial represents the decision in the hypothesis test (for example, a verdict of guilty means rejecting Ho).

1. Suppose the defendant is truly innocent but they are found guilty. Is this a type 1 error, type 2 error, or the correct decision?
2. In another trial, suppose the defendant is truly guilty but they are found innocent. Is this a   
   type 1 error, type 2 error, or the correct decision?

“Controlling” errors in Hypothesis Tests:

* We don’t know in practice whether an error occurs on a single test, again because we don’t have   
  information from the population.
* So errors can’t be controlled in individual tests (this is why it is in quotes above).
* However, we can find the probability of an error. Here’s the idea: if we took samples over and over from the population and performed a test based on each sample, the probability of an error would be the proportion of the tests that result in an error.
* Important fact: This probability is controlled by a quantity we’ve already run into. **The probability of   
  a Type 1 Error is equal to the significance level (denoted by ).**
* Thus, the smaller the , the smaller the probability of Type 1 Error. If the consequences of a false positive are serious, we may want to use a smaller so there is a smaller chance is making that mistake.

Power

* Instead of talking about the Type 2 Error probability, statisticians usually refer to the **power**,

the probability of a true positive (rejecting Ho correctly, when Ho is actually false).

* As decreases, the power decreases because stronger evidence (a smaller p-value) is needed to produce a statistically significant result.
* Thus, there is a trade-off in the choice of between minimizing the chance of false positives and maximizing the power.

Connection between model assumptions and errors in hypothesis tests

* The two-sample t test (like all hypothesis tests) depends on the model assumptions being true.
* Specifically, we rely on the normality assumption when we use the t distribution to calculate the p-value.
* Fact: If normality assumption is severely violated for a dataset, the probability of a Type 1 error will not be equal to the significance level.
* In the next handout, we introduce the **rank-sum test**, which is an alternative to the two-sample t test that does not rely on the model assumption of normality.

# Handout 9: Rank-sum test

STA 216 F19

* Recall from Handout 8:
  + If model assumptions are met, Type 1 error rate = significance level alpha
  + If assumption of normality is severely violated for a dataset, the Type 1 error rate will not be equal to the significance level alpha
* Rank-sum test
  + proper names: Wilcoxon rank-sum test or Mann-Whitney test
  + alternative to the two-sample t test which does not rely on the assumption of normality
  + Type 1 error rate = significance level alpha for this test even if normality not satisfied.
* Generally, tests that do not rely on model assumptions are called **nonparametric tests**.
* STA 317 is all about nonparametric tests.

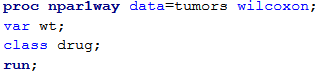
Example

During an in vivo comparison of two antitumor drugs, Collings and Hamilton (1998) collected the following data on tumor weights (in g) for two groups of mice, each exposed to a different drug. Does the sample show evidence of a difference, on average, for the population of mice, in tumor weights for mice under the standard drug and test drug?

|  |  |
| --- | --- |
| Data | Normality check plots (from PROC TTEST with ODS GRAPHICS ON) |
|  | img0.png  img1.png |

* For both drugs, the distributions of tumor weight are strongly right skewed, as shown by the histograms, boxplots, and normal quantile plots.
* We cannot assume the tumor weights come from the normal distribution.
* Thus, we should not use the two-sample t test.
* The rank-sum test is a good alternative.

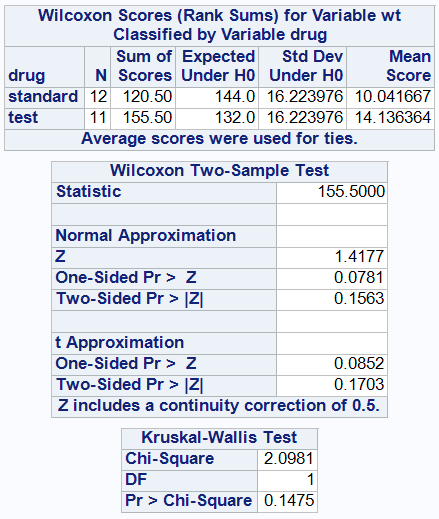
Code:

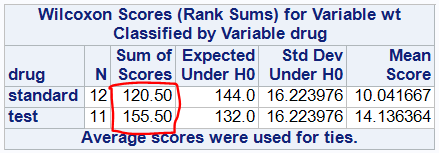


Note: In PROC NPAR1WAY

* NPAR is for nonparametric
* 1WAY is for one-way ANOVA

Output:



* The hypotheses for the rank sum test are in terms of the **medians** (not means) of the two populations.
* We will denote the population medians using the Greek letter eta, .
* Thus, the 3 options for the hypotheses are
  + Ho: vs. Ha:
  + Ho: vs. Ha:
  + Ho: vs. Ha:
* For this example, this part of the problem statement:  
    
  *Does the sample show evidence of a* ***difference****, on average, for the population of mice, in tumor weights for mice under the standard drug and test drug?*  
    
  tells us that the hypotheses are
* Ho: vs. Ha: , where
  + population median tumor weight for mice on the standard drug
  + population median tumor weight for mice on the test drug
* The test statistics are the rank sums of each group, which are 120.5 and 155.5.  
  
* Calculation:
  + Sort response variable (wt) from smallest to largest.
  + Rank each observation from 1 to N = n1 + n2 (here, 23).
  + Ties get average of ranks they are over.
  + Sum the ranks for each group.
* In this case, first using PROC SORT by wt, we can find the sum of ranks in each group as

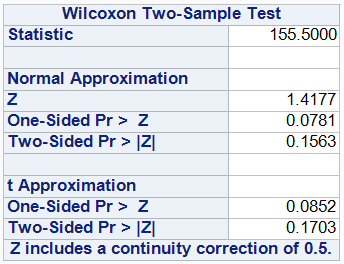
|  |  |
| --- | --- |
| Tied ranks | Ranks | Sum of ranks  Standard drug:  1+2+3+4.5+4.5+7.5+9+13.5+15+18.5+20+22  = 120.5  Test drug:  6+7.5+10+11.5+11.5+13.5+16.5+16.5+18.5+21+23  = 155.5 |

Note:

* No assumption is made about the distribution (normal or not) because only the order of the observations is used.
* This test is resistant to outliers.

P-value

* We’ll use the **Normal Approximation** section of the output to find the p-value.



Because our hypothesis test is two-sided, our p-value = 0.1563.

Decision

Fail to reject Ho because p-value > .05

Conclusion

At the alpha=.05 significance level, there is **not** statistically significant evidence that the population median tumor weight for mice on the standard drug is different than for the test drug.

# Handout 10 Supplement

Correlation:

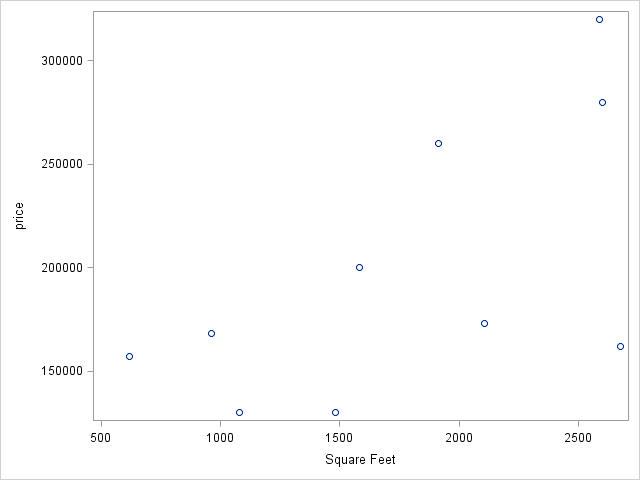
* Notation: r
* Number that is between **-1** **and 1** for any dataset.
* Measures two things about the **linear** relationship between the two variables: **direction** and **strength**.
* Direction: As the x values increase, if the y values tend to
  + Increase 🡪 **positive** direction
  + Decrease 🡪 **negative** direction
  + The direction is determined by the **sign** of the correlation.
  + It is the same of the sign of the slope of the least squares regression line.
* Strength: how closely do the points on the scatterplot follow a straight line?

|  |  |  |
| --- | --- | --- |
| Very strong (r=.99) | Moderate (r=.68) | Weak (r= -.21) |
|  |  |  |

* Strength is determined by the **magnitude** (or absolute value) of the correlation
  + Large magnitude (close to 1 or -1): strong strength
  + Small magnitude (close to 0): weak strength
  + In between: moderate strength

Example of correlation:

Square footage and price for the ten houses



|  |  |
| --- | --- |
| **proc** **corr** data=houses;  var sqft price;  **run**; |  |

Correlation: r = 0.62

Interpretation of correlation value: The linear relationship between square footage and price  
has **positive** direction with **moderate** strength.

Handout 10

STA 216 F19

Simple Linear Regression Model

* Recall from Handout 4: All the models in the class will follow the form

Response = Mean + Error

Y = f(X) + E

* For the two-sample model, this was Y = , where the errors
* For the simple linear regression model, the explanatory variable X is quantitative instead of categorical.
* Instead of having a different mean for each group i, we have a different mean of Y for each X, which we denote as .
* We assume follows a line. In other words, , where
  + is the intercept of the line
  + is the slope of the line
* Another difference is that we assume the same standard deviation for each observation (i.e. instead of and ).
* Putting this altogether, we have the simple linear regression model  
    
  .
* To summarize, there are 4 assumptions of the simple linear regression model:
  + Linearity: (the mean of Y is a linear function of X); equivalently, the mean of the errors is 0.
  + Constant variance: Each Y has the same standard deviation around
  + Normality
  + Independence
* In the model: , , and are parameters.
  + is the population intercept
  + is the population slope
  + is the population standard deviation
* These are estimated from the sample:
  + , the sample intercept, is the estimate of
  + , the sample slope, is the estimate of
  + , the sample standard deviation, is the estimate of

Example

Price and square feet for 10 houses.

* Response variable Y = price (in $1000s)
* Explanatory variable X = square feet

|  |  |
| --- | --- |
| **proc** **sgplot** data=houses;  reg x=sqft y=price;  **run**; | img0.png |

|  |  |
| --- | --- |
| (MODEL y = x) |  |

The output tells us that

* (the Root MSE, for reasons we will learn later)
* We can get the estimate of the population mean of Y at X, by plugging in for and for .
* We call this the least squares regression line .
  + We call (y-hat) the estimated mean Y or predicted Y
  + In this case, we have

Interpreting slope

* The slope of the regression is the change (increase or decrease) in when x increases by one unit.
* In this case, = predicted price in $1000s, x = square feet, slope = 0.056
* Interpretation: The predicted price of houses increases by 0.056 thousand dollars (or $56) when square feet increases by one.

Interpreting intercept

* The intercept of the regression line is the value of when x = 0.
* In this case, intercept = .
* Interpretation: The predicted price of houses with 0 square feet is 99.55 thousand dollars (or $99,550).
* Often, it does not make sense in the data context to have x = 0.

Important: The word **predicted** (or mean is fine too) must be included in these interpretations (or points will be taken off), as in this case “predicted price” to distinguish for actual price.

Interpreting standard deviation

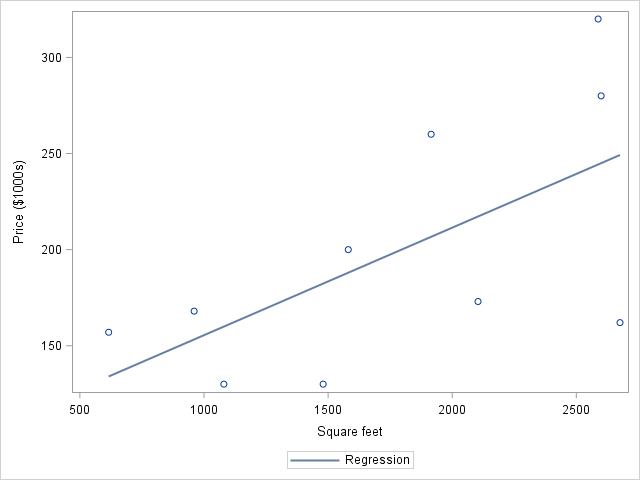
* The average distance of the actual y values from the values.
* In this case,
* Interpretation: The average difference from the actual prices to the predicted prices is 54.65 thousand dollars (or $54,650).

Calculating predicted Y for a given X

* Just plug in the X to find !
* Example: Estimate the mean price of all houses with 2500 square feet.  
  Using the regression line ,

Residual:

The formula for the residual is . It gives the difference between the actual Y and the predicted Y.

* Example: Find the residual for the house with 1915 square feet costing $260,000.  
  + The actual price is Y = 260 (in thousand dollars)
  + The predicted price is
  + The residual is
* Note: We use the notation for residuals because they are estimates of the errors
* Graphical interpretation: vertical distance from point to line.  
  For the example above,  
  

How is the least squares regression line determined?

* Out of all possible lines (with all possible intercepts and slopes), it is the “best fitting” line to the points on the scatterplot.
* “Best fitting” means that the line makes the **sum of squared residuals** as small as possible.
* This is why it’s called the **least squares** line.

Handout 10 Supplement

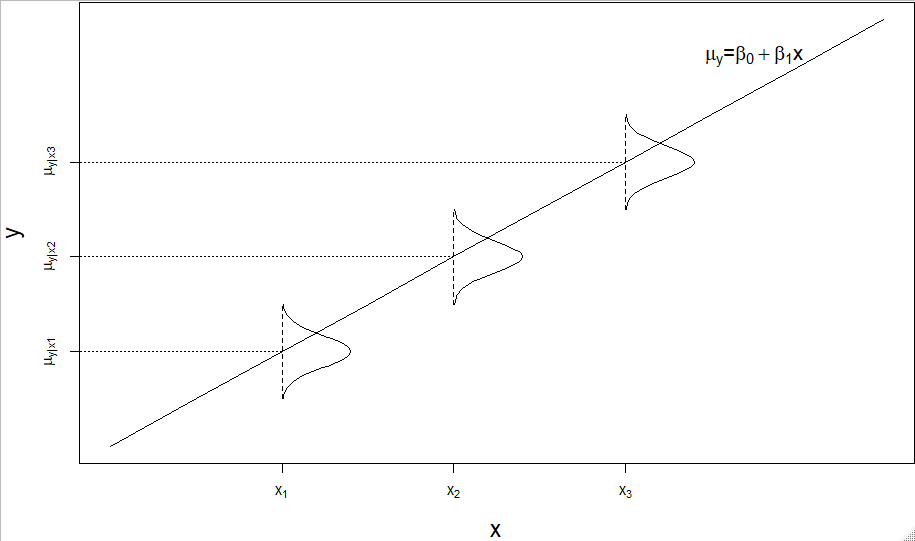
Review of correlation: Read for your own review

# Handout 11 – Assumptions in SLR and how to check them

STA 216 F19

* Simple linear regression model  
    
  .
* To summarize, there are 4 assumptions of the simple linear regression model:
  + **Linearity (or mean zero errors):** (the mean of Y is a linear function of X); equivalently, the mean of the errors E is zero.
  + **Constant variance:** The variability of the errors is the same for all values of X. This means that the spread of the points around the line remains fairly constant.
  + **Normality:** The errors follow a normal distribution
  + **Independence:** The errors are assumed to be independent of each another. Thus, one point falling above or below the line has no influence on the location of another point.

Pictorial representation



* Note that each of the assumptions above is in terms of the errors E. We don’t know the values of the errors because we don’t know and .
* But the residuals are estimates of the errors, so can be used to check the assumptions.
* Recall that the residuals and are graphically the vertical distance from a point to the line. (See example on last page of Handout 10.)
* Main point: we check the assumptions using the residuals.

Two kinds of plots of residuals to check assumptions\*

* Residual vs. predicted plot: Used to check
  + Linearity (mean zero errors)
  + Constant variance
* Normal quantile plot of residuals: used to check normality

\*We won’t concern ourselves with checking independence.

Residual vs. predicted plot

Residuals on y-axis; Predicted values on x-axis

|  |  |
| --- | --- |
| Scatterplot of y vs. x | Residuals vs. predicted plot |
| Data 1: Linearity satisfied; constant variance satisfied | |
| img0.png | img1.png |
| Data 2: Linearity violated; constant variance satisfied | |
| img0.png | img1.png |
| Data 3: Linearity satisfied; constant variance violated | |
| img0.png | img1.png |
| Data 4: Linearity violated; constant variance violated | |
| img0.png | img1.png |

Interpreting residual vs predicted plots

* To check linearity / mean zero errors:
  + on y vs. x scatterplot: points should follow a linear pattern
  + on residual vs. predicted plot: residuals should be vertically centered around zero for each predicted value
* To check constant variance:
  + on y vs. x scatterplot: points should be spread out vertically the same amount for all values of x
  + on residual vs. predicted plot: residuals should be vertically spread out the same amount for all predicted values.

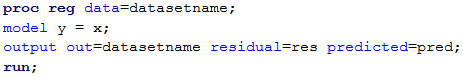
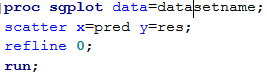
Assessing normality

Using the residuals, we use the same strategy as before.

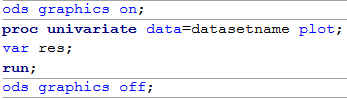
* Is the histogram roughly bell-shaped?
* Is the boxplot roughly symmetric?
* Are the points roughly following the line on the normal quantile plot?

|  |  |
| --- | --- |
| Normality satisfied | Normality violated |
| img0.png | img1.png |

Making residual vs. predicted plots in SAS

* Step 1: Use PROC REG to calculate the residuals and the predicted values.  
    
  Note: The OUTPUT statement is used to add the residuals and predicted values as new variables to the dataset named “datasetname” containing y and x. In this case, they will have names “res” and “pred”.
* Step 2: Use PROC SGPLOT with the SCATTER statement  
    
  Note that “refline 0;” adds the horizontal line at zero.

Making the histogram, boxplot, normal quantile plot combo

* Step 1: Obtain the residuals as in step 1 above.
* Step 2: Use PROC UNIVARIATE  
  

# Handout 12 – Statistical Inference for simple linear regression

STA 216 F19

Overview

* T-test and confidence interval for slope
* Analysis of Variance (ANOVA) table
* F test for slope
* R-squared and root MSE
* Confidence interval for population mean, Prediction interval for individual response

Statistical inference

* When our data is a random sample from a population, we use statistical inference (in the form of hypothesis tests and confidence intervals) to make conclusions about the population.
* For the simple linear regression model statistical inference is making conclusions about the population slope .
* Note: if , then there is no relationship between X and Y.
* For the cereal dataset, would it be useful to make a confidence interval for the slope between rating and fiber?
* We will use the dataset of the square feet and price of 10 houses, which are a random sample from a population of houses in Saratoga County, NY.

T-test for the slope

|  |  |
| --- | --- |
|  |  |

The bottom table tells us that

* Sample intercept , sample slope
* Equation of regression line:

For the hypothesis test of the slope Ho: vs. Ha: ,

* The test statistic t = 2.26, which is calculated as
* The p-value of the two-sided test is 0.0534. (Comes from t distribution with df = n – 2.)
* Thus, at the .05 level, we fail to reject Ho
* At the .05 level, there is not statistically significant evidence that the population slope is nonzero, or that the population mean price depends on the square feet.

Confidence interval of the slope

|  |  |
| --- | --- |
| **proc** **reg** data=houses;  model price = sqft /clb;  **run**; |  |

* Adding the CLB option tacks on the confidence interval to the table. The CI for the population slope is (-0.001, 0.113).
* The fact that the CI contains zero agrees with our test’s decision to fail to reject Ho: .
* Let’s calculate a 90% confidence interval instead by adding the ALPHA=.10 option

|  |  |
| --- | --- |
| **proc** **reg** data=houses alpha=**.10**;  model price = sqft /clb;  **run**; |  |

* Interpretation of the slope (): The **predicted** price of houses increases by 0.056 thousand dollars ($56) for every increase of 1 square feet.
* Interpretation of confidence interval for the population slope (.010, .102): We are 90% confident that the population mean price of houses increases by between .010 and .102 thousand dollars ($10 and $102) for every increase of 1 square feet.

Analysis of Variance (ANOVA)

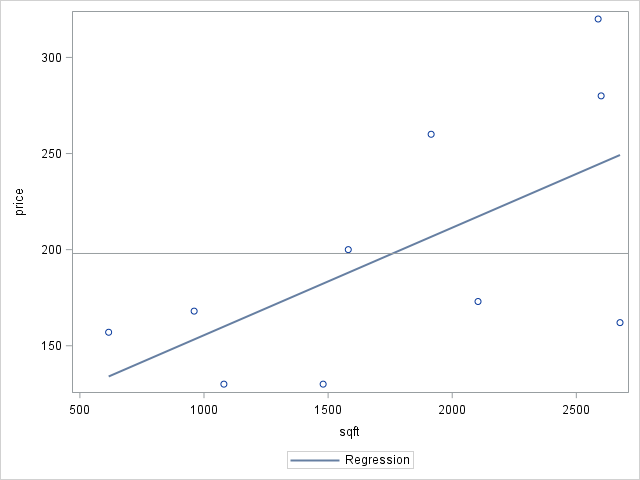
* Partitions the total variability in the response variable into two pieces:
  + Variation explained by the model (by the relationship with X)
  + Unexplained variation in the residuals
* The total variability is quantified by the sum of the squared deviations, which we call sum of squares Total, or SS(Total).   
  SS(Total) =
* The variability explained by the model is   
  SS(Model) =
* The unexplained variability is the sum of squared residuals (also called sum of squares error), with notation SS(E).  
  SS(E) =
* ANOVA sum of squares identity: SS(Total) = SS(Model) + SS(E)

Example: Calculating SS(Total), SS(Model), and SS(E)

|  |  |
| --- | --- |
|  |  |

* yhat is calculated by plugging each sqft value into the regression line .
* ybar is the mean of the 10 prices.

|  |  |
| --- | --- |
|  |  |



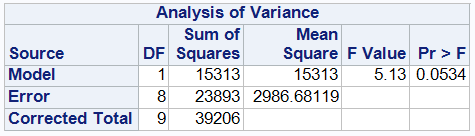
* For example, y\_ybar \*\*2 is

|  |  |
| --- | --- |
|  |  |

* SS(Model) =
* SS(E) =
* SS(Total) =

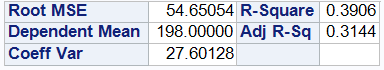
Note that these sums of squares are given in the Analysis of Variance table (top table of output)

* SS(Model) =
* SS(E) =
* SS(Total) =



ANOVA Table

* DF (degrees of freedom)
  + DF(Model) = # of explanatory variables = p
  + DF(Error) = n – p – 1
  + DF(Total) = n – 1
* Mean squares (MS) = Sum of squares DF
  + MS(Model) = SS(Model) / DF(Model) = 15,313 / 1 = 15,313
  + MS(Error) = MSE = SS(E) / DF(Error) = 23,893 / 8 = 2986.7
* F Value: This is the F statistic for the F test of the population slope Ho: vs. Ha:
  + F = MS(Model) / MSE = 15,313 / 2987.7 = 5.13
* Pr > F: P-value for the F test
  + Comes from the F distribution with numerator df = DF(Model) and   
    denominator df = DF(Error)
* Note: the p-value = .0534 for the F test for the slope is exactly the same as the t test for the slope.



Root MSE

* Recall that we found s, the estimate of the error standard deviation, as the root MSE.
* The root MSE is the square root of the MSE.
* Here, s =
* Interpretation: The average distance from actual prices to the predicted prices is 54.65 thousand dollars.

R-squared

* R-squared (the coefficient of determination) is the correlation (r) squared.
* So it takes values between 0 and 1 and measures the strength of the linear relationship.
* It is also calculated as , or
* Thus, is interpreted as the fraction of the total variation in y [SS(total)] which is explained by the model [SS(model)]. It’s often written as a percentage.
* For this example,
* Interpretation: 39.06% of total variation in price is explained by the model with square feet.

Confidence and Prediction Intervals for a simple linear regression response

* Recall that we have calculated estimated mean or predicted y values for a particular value of the explanatory variable by plugging in the x value into the regression line to find .
* What if we want an interval on this estimated mean /prediction?
* Two types of intervals
  + Confidence interval for the population mean
  + Prediction interval for an individual response y.
* For the houses data, this would be
  + A confidence interval for the population **mean** price of all houses that are (say) 2000 sqft.
  + A prediction interval for the price of an individual house that is 2000 sqft.

To get in SAS

* Step 1: Add an observation to the data with that x value.

|  |  |
| --- | --- |
|  |  |

* Step 2: Use this new dataset in a PROC REG with
  + Option CLM for confidence interval for the population mean
  + Option CLI for confidence interval for an individual y

|  |  |
| --- | --- |
| **proc** **reg** data=houses2;  model price = sqft /clm;  **run**; |  |
| **proc** **reg** data=houses2;  model price = sqft /cli;  **run**; |  |

# Handout 13 – Introducing Multiple Linear Regression

STA 216 F19

Overview

Multiple linear regression is an extension of simple linear regression

* Statement of the model
* 4 assumptions and checking them
* Scatterplot matrix
* Equation of the least squares regression line
* Calculating predicted values and residuals
* Interpreting the intercept and slope estimates
* Interpreting the root MSE and R-squared
* t tests for slopes, confidence intervals for slopes
* ANOVA table and the global F test
* Confidence intervals for the mean response and prediction intervals for an individual response

The multiple linear regression model

* Instead of having a single explanatory variable X as in simple linear regression, we use p explanatory variables , .
* The population mean of Y is assumed to be a linear combination of these variables:
* As before, the Y’s are assumed to vary from these means by error terms which are independent of each other, have constant variance, and follow a normal distribution.
* The model:  
  , where

4 assumptions implied by the model

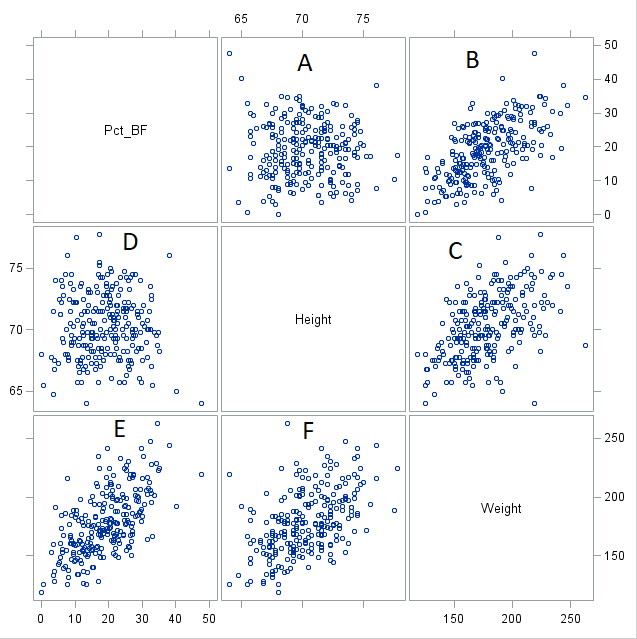
* Linearity: (the mean of Y is a linear combination of the X’s); equivalently, the mean of the errors is 0.
* Constant variance: Each Y has the same standard deviation around
* Normality
* Independence

Data Example

Random sample of 250 adult men. Goal: analyzing how x1 = height (in inches) and x2 = weight (in pounds) are related to Y = body fat percentage.

Scatterplot matrix





Shows scatterplots for different bivariate relationships between the variables

* A has y=Pct\_BF, x=Height
* B has y=Pct\_BF, x=Weight
* C has y=Height, x=Weight
* D has y=Height, x=Pct\_BF (mirror image of A)
* E has y=Weight, x=Pct\_BF (mirror image of B)
* F has y=Weight, x=Height (mirror image of C)

Note axes around the plots

* Range of Pct\_BF includes 0 to 50
* Range of Height includes 65 to 75 inches
* Range of Weight includes 150 to 250 pounds

Interpretations of scatterplots

* No correlation between Pct\_BF and Height (A)
* Positive correlation between Pct\_BF and Weight (B)
* Positive correlation between Height and Weight (C)

Correlation matrix

|  |  |
| --- | --- |
|  |  |

PROC REG output

|  |  |
| --- | --- |
|  |  |

Least squares regression line

* As before, we get the estimates of the intercept and slopes from the Parameter Estimate column of the bottom table.
* The model here is
* The estimates of the parameters are:
  + is the estimate of
  + is the estimate of
  + is the estimate of
* Thus, the equation of the least squares regression line is   
  Pct\_BF(hat) = 76.78 – 1.49\*Height + 0.263\*Weight

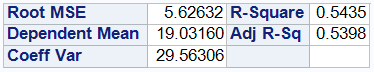
Interpretation of slopes in multiple regression

* Still means the change in predicted y for a one unit increase in x
* Additional part: other explanatory variables are **held constant**
* For example, the slope estimate of height = -1.49 is interpreted as  
  The predicted percent body fat of adult males decreases by 1.49 for every increase of one inch in height, considering adult males of the same weight.
* Upshot: If we consider guys that are the same weight, the taller guys will be skinnier (less % fat).

Predicted response values and residuals

* One male in the dataset (the first row) has Pct\_BF = 12.3, Height = 67.75 in, Weight = 154.25 lb. Let’s calculate the predicted percent body fat and the residual for this guy.
* The predicted Pct\_BF for this guy based on the regression line is  
  Pct\_BF(hat) = 76.78 – 1.49(67.75) + 0.263(154.25) = 16.6
* As before, residual = y – yhat  
  Here, Pct\_BF – Pct\_BF(hat) = 12.3 – 16.6 = -4.3

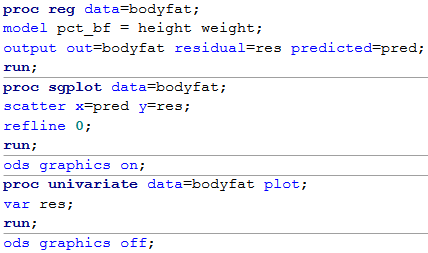
Interpreting root MSE and R-squared



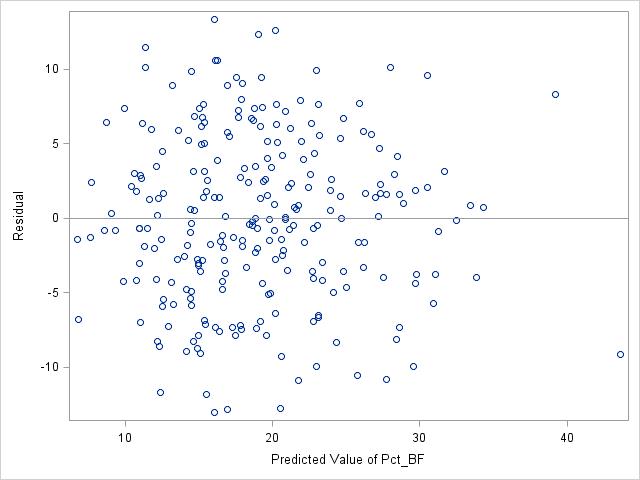
* The root MSE (or s) is still the estimate of , the standard deviation of the errors.
* Interpretation: The average distance from the actual response variables from the predicted values.
* In this context: The average distance of the actual percent body fat values from their predicted values according to the model with height and weight is 5.63 percent.
* R-squared is still the percent of the total variation in the response that is explained by the model.
* In this context: 54.35% of the total variation in percent body fat is explained by the model with height and weight.

Checking assumptions

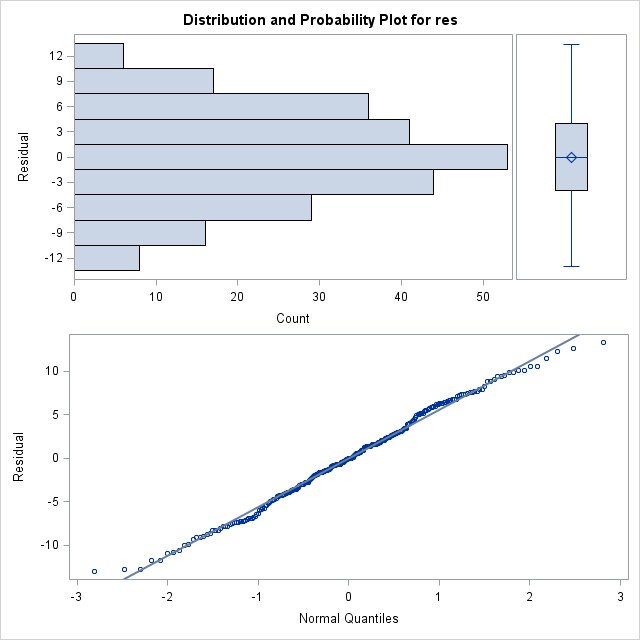
* We will check the assumptions of linearity (zero mean errors), constant variance, and normality as before.
* Linearity (zero mean errors) and constant variance use the residual vs. predicted plot
  + Linearity looks for the residuals to be centered around the zero line for all ranges of predicted values.
  + Constant variance looks for the residuals to be have the same vertical spread for all ranges of predicted values.
* Normality uses a histogram, boxplot, and normal quantile plot of the residuals.



Checking assumptions for body fat example



* Linearity (zero mean errors) is satisfied because the residuals are centered around the zero line for all ranges of predicted values
* Constant variance is satisfied because the vertical spread of the residuals is roughly the same for all ranges of predicted values

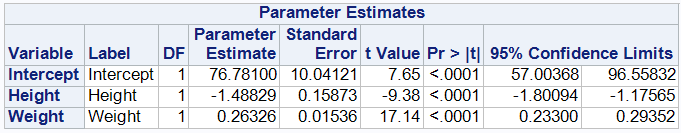


* Normality is satisfied because the histogram is bell-shaped and symmetric, the boxplot is symmetric, and the points in the normal quantile plot follow the line nicely.

Hypothesis tests and confidence intervals for slopes

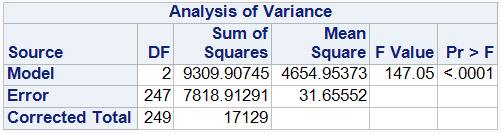
* This is a random sample from the population of all adult males, so it makes sense to do hypothesis tests and confidence intervals.
* As before, we add the CLB option to get confidence intervals.





* For the model , the hypothesis test for the slope of height has
  + hypotheses Ho: vs. Ha: ,
  + The test statistic t = -9.38, which is calculated as
  + p-value < .0001
  + Decision: reject Ho
  + Conclusion: At the .05 significance level, there is statistically significant evidence that the population slope of height is nonzero **when weight is in the model**.
* 95% confidence interval for is (-1.80, -1.18)
* Interpretation: We are 95% confident that the population mean percent body fat for adult males decreases by between 1.18 and 1.80 for every increase of one inch in height, considering males of the same weight.

ANOVA Table



* DF (degrees of freedom)
  + DF(Model) = # of explanatory variables = p
  + DF(Error) = n – p – 1
  + DF(Total) = n – 1
* Sum of squares
  + SS(Model) =
  + SS(Error) =
  + SS(Total) =
* Mean squares (MS) = Sum of squares DF
* F = MS(Model) / MS(Error)

Global F test

* F test in multiple regression has hypotheses  
  Ho: all vs. Ha: at least one of
* Here, p-value < .0001 🡪 reject Ho
* conclusion: At .05 level, there is statistically significant evidence that at least one of weight and height has a linear relationship with population mean body fat.

Confidence intervals and prediction intervals

* Confidence interval is capturing , the mean response for the population with particular values of x1, x2, …, xp
* Prediction interval is capturing Y, the response value of an individual with particular values of x1, x2, …, xp
* Example: Let’s consider a height of 6 ft (72 in) and a weight of 190 pounds

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

* 95% Confidence interval = (18.8, 20.5)  
  Interpretation: We are 95% confident that the population mean percent body fat for all 6ft tall, 190 pound adult males is between 18.8 and 20.5 percent.
* 95% Prediction interval = (8.5, 30.8)  
  Interpretation: We are 95% confident that the percent body fat for an individual 6ft tall, 190 pound adult male is between 8.5 and 30.8 percent.
* As before, prediction intervals are wider than confidence intervals because it is harder to estimate the response value for an individual than the population mean response.

# Handout 14 – Separate regression lines by a categorical variable

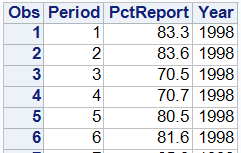
STA 216 F19

* So far we have only considered quantitative explanatory variables to model the quantitative response Y.
* What if we have both categorical and quantitative x’s?

Example

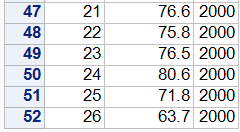
Tom Shields, jury commissioner for the Franklin County Municipal Court in Columbus, OH, is responsible for making sure that the judges have enough potential jurors to conduct jury trials. Only a small percent of the possible cases go to trial, but potential jurors must be available and ready to serve the court on short notice. Jury duty for this court is two weeks long, so Tom must bring together a new group of potential jurors 26 times a year. Random sampling methods are used to obtain a sample of registered voters in Franklin County every two weeks, and these individuals are sent a summons to appear for jury duty. One of the most difficult aspects of Tom’s job is to get those registered voters who receive a summons to actually appear at the courthouse for jury duty. The dataset contains the percentages of individuals who reported for jury duty after receiving a summons in 1998 and 2000. The reporting dates vary slightly from year to year, so the two week periods are coded sequentially from 1 to 26. A variety of methods were used after 1998 to try to increase participation rates. How successful were these methods in 2000?

Portion of the Dataset:



. . .

. . .



* Here, the response variable is PctReport and the explanatory variables are period and year.
* Period is quantitative and year is categorical.

Scatterplot of PctReport vs. Period with points grouped according to Year

|  |  |
| --- | --- |
|  | img0.png |

Adds separate regression line for each year

|  |  |
| --- | --- |
|  | img1.png |

Calculates equations of both regression lines based on considering each year as a separate dataset (using WHERE statement).

|  |  |
| --- | --- |
| **proc** **reg** data=jurors;  where year=**1998**;  model pctreport = period;  **run**; |  |
| **proc** **reg** data=jurors;  where year=**2000**;  model pctreport = period;  **run**; |  |

* The slopes of -0.67 and -0.77 are close together (which you can see because the lines are close to parallel).
* Using multiple regression, we can combine both years into a single model (and not have to separate the data).

Dummy variables

* The key in using categorical variables in a regression model is to use a **dummy variable** (also called indicator variable) that distinguishes between the groups formed by the categorical variable.
* Dummy variables take values
  + 1 if the observation belongs to the group
  + 0 if the observation does not
* In this case, a dummy variable for the year 2000 (let’s call it D2000 for short) would be defined as
  + D2000 = 0 if Year = 1998
  + D2000 = 1 if Year = 2000
* Thus, the regression model modeling PctReport using Period and Year (using D2000 instead of Year) would be

PctReport =

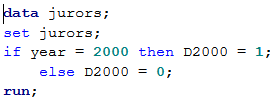
* This model contains the two separate regression lines inside it for the two different years because

|  |  |
| --- | --- |
| For Year = 1998, D2000 = 0 so the model is PctReport =   = | For Year = 2000, D2000 = 1 so the model is PctReport =   = |

* Note that these are two lines with the same slope () but the intercept is adjusted by for   
  Year = 2000.

In SAS

* Making the dummy variable D2000 with an if-then-else statement.



* Fitting the model with PROC REG

|  |  |
| --- | --- |
|  |  |

* Equation of the regression line:   
  PctReport(hat) = 77.08 – 0.717\*Period + 17.83\*D2000
* Separate regression lines for each year

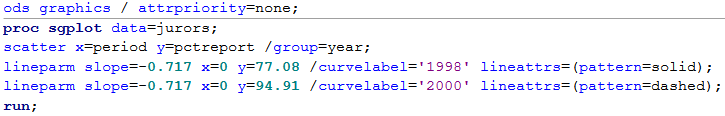
|  |  |
| --- | --- |
| Year = 1998 D2000 = 0 | Year = 2000  D2000 = 1 |
| PctReport(hat) = 77.08 – 0.717\*Period + 17.83(0)  = 77.08 – 0.717\*Period | PctReport(hat) = 77.08 – 0.717\*Period + 17.83(1)  = 94.91 – 0.717\*Period |

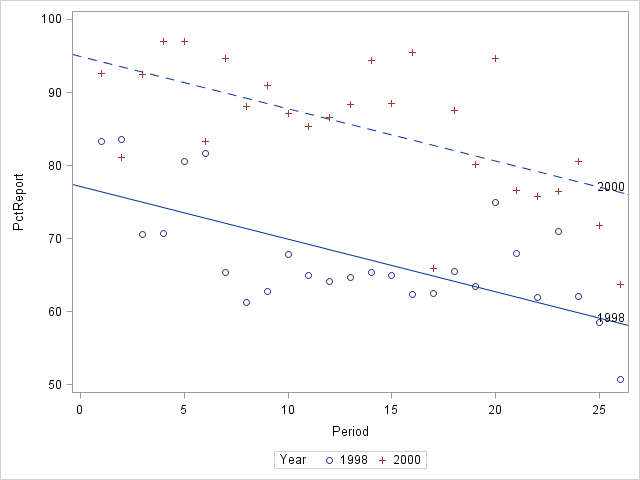
Plotting these lines

Uses the LINEPARM statement:

LINEPARM slope=b1 x=0 y=b0;

(Recall that b0 and b1 are the slope and intercept estimates, respectively.)





Interpretations of “slopes”

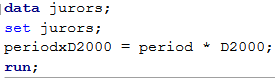
* Interpretation of slope of Period: The predicted percent of jurors reporting for duty decreases by about 0.7% per period, considering the same year.
* Interpretation of ~~slope~~ coefficient of D2000: The predicted percent of jurors reporting for duty was 17.8% higher in 2000 than in 1998, considering the same period.

Lines with different slopes

* Is this model that produces parallel (equal slope) regression lines “good enough”?
* Let’s try one that allows the slopes of the lines for 1998 and 2000 to be different and see.

Interaction term

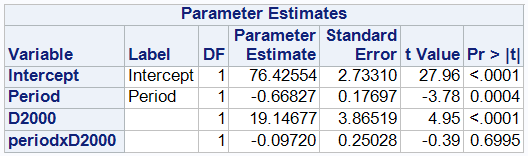
* The key is adding an interaction term obtained from multiplying explanatory variables together.
* Here, we’ll multiply together period and D2000.
* In SAS, we’ll call it periodxD2000. We make a new variable for it like:



* Adding it to the model with period and D2000 (by themselves), we have the model  
  PctReport =
* Again, we can get the separate models for 1998 and 2000 by plugging in 0 or 1 for D2000.
  + For 1998 (D2000 = 0):  
    PctReport =   
     =
  + For 2000 (D2000 = 1):  
    PctReport =   
     =
* Now, we have lines with different intercepts and slopes.
* The coefficient for this new interaction term, , shows how much the slope changes as we move from the regression line for 1998 to the line for 2000.

Fitting the model with PROC REG





* Equation of the regression line:   
  PctReport(hat) = 76.43 – 0.668\*Period + 19.15\*D2000 – 0.097\*period\*D2000
* Separate regression lines for each year

|  |  |
| --- | --- |
| Year = 1998 D2000 = 0 | PctReport(hat) = 76.43 – 0.668\*Period + 19.15\*(0) – 0.097\*period\*(0)  = 76.43 – 0.668\*Period |
| Year = 2000 D2000 = 1 | PctReport(hat) = 76.43 – 0.668\*Period + 19.15\*(1) – 0.097\*period\*(1)  = 95.58 – 0.765\*Period |

* Note that these are the same equations we got when we did the separate PROC REG’s with WHERE statements.
* So is the model that produces parallel (equal slope) regression lines “good enough”?
* Comparing the output from the parallel model (without interaction term) to nonparallel model (with interaction term)

|  |  |
| --- | --- |
| Parallel lines |  |
| Non-parallel lines |  |

* Hypothesis test for period\*D2000 term, (measures difference between slopes)  
  Ho: vs. Ho:   
  p-value = .6995 🡪 Not statistically siginificant evidence of a difference between the slopes
* Root MSE
  + 6.71 for parallel lines
  + 6.76 for nonparallel lines
  + Average distance between actual percent reporting and predicted percent reporting is less for parallel lines.
* R-squared
  + 0.7188 for parallel lines
  + 0.7197 for nonparallel lines
  + Percent of total variation in percent reporting explained by the model is almost the same for parallel lines and nonparallel lines.

# Handout 15 – Comparing models fit with R, root MSE, and Adj. R

STA 216 F19

Overview

* Comparing model fits with R-squared, root MSE, and adjusted R-squared
* Incorporating categorical explanatory variables with more than 2 categories

Comparing how well models fit with R-squared, root MSE, and adjusted R-squared

* In general, for a good fitting model, we want
  + R-squared to be **high** because it represents the total variation in y explained by the model.
  + Root MSE to be **low** because it represents the average distance between the actual y values and the predicted y values by the model.
* Recall that they are calculated as
  + , or
  + Root MSE =
* If we are deciding whether or not to include an additional explanatory variable, though, using R-squared can be misleading because **it always goes up** when another x variable is added.
* This is because, when another x is added, the SS(Model) always goes up and the SS(Error) always goes down (even if it’s just by a tiny bit).
* Also, the SS(Total) always stays the same.
* Thus, if you decide based on which R-squared is higher, you will always choose to add the x variable.
* The root MSE doesn’t have this problem because, when adding a weakly predicting x, a tiny decrease in SS(Error) can be offset by a larger decrease in (n-p-1), making the root MSE go up.
* That is, it accounts for the number of explanatory variables p.
* The adjusted R-squared also accounts for the number of explanatory variables. It’s calculated as
* It is possible for the adjusted R-squared to go down when a weakly predicting x is added.

Example from Handout 14

|  |  |  |
| --- | --- | --- |
|  | Parallel lines | Non-parallel lines |
| Root MSE | 6.71 | 6.77 |
| Adjusted R-squared | 0.7073 | 0.7022 |

Note that the root MSE and adjusted R-squared agree in calling the parallel lines model better, because it has a lower root MSE and a higher .

* In Handout 14, we learned how to incorporate a categorical x variable into a regression model.
* Our example used a variable with two categories (in other words, a binary variable).
* What about one with more than 2 categories?

Example

In Lab 5, we predicted the prices (in $1000s) of used 30 Porsches offered for sale at an internet site based on their mileages (in thousands of miles). Suppose we collected similar data for two other car models, Jaguars and BMWs.

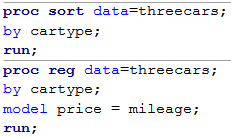
* Response variable: price
* Explanatory variables:
  + Mileage: quantitative
  + Cartype: categorical with three categories – Porsche, Jaguar, and BMW

Scatterplot with three separate regression lines

|  |  |
| --- | --- |
|  | img1.png |

Calculating separate equations of regressions with BY processing

* We could use 3 PROC REG’s with WHERE statements as on the last handout.
* If there are a lot of categories, it is more efficient to use BY processing:
  + First: use PROC SORT to sort the data BY the categorical variable.
  + Then: use PROC REG with a BY statement



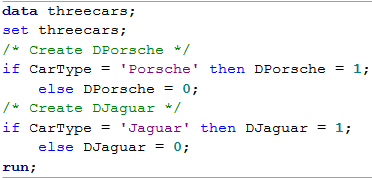
|  |  |  |
| --- | --- | --- |
|  |  |  |

Making multiple dummy variables

* In Handout 14, we incorporated a binary categorical variable into a regression model by making a dummy variable that takes values 0 or 1.
* To deal with this three-category variable, we make dummy variables for all the categories **except one**.
* Approaches that **do not** work:
  + Making a single variable with values {0, 1, 2}
  + Making dummy variables for every single category.
* The category we don’t make the dummy variable for is called the **reference category** (for reasons we’ll soon see).
* It doesn’t matter which category we make the reference category. I’ll choose to make it BMW here.
* Thus, we’ll make a dummy variables for Porsche and Jaguar.
* DPorsche
  + = 1 if car is a Porsche
  + = 0 if not
* DJaguar
  + = 1 if car is a Jaguar
  + = 0 if not

Making these variables in SAS

* Use if-then-else statements
* I also put in comments to organize my code.



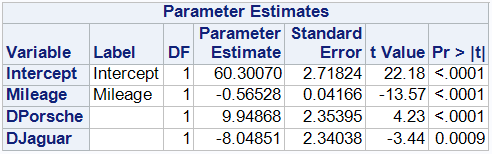
The Model contains the three regression lines

|  |  |
| --- | --- |
| BMW  (DPorsche=0, DJaguar=0) |  |
| Porsche  (DPorsche=1, DJaguar=0) |  |
| Jaguar  (DPorsche=0, DJaguar=1) |  |

* These lines all have the same slope, , but different intercepts:
  + The Porsche line is adjusted by from the BMW line
  + The Jaguar line is adjusted by from the BMW line

Fitting the model with PROC REG





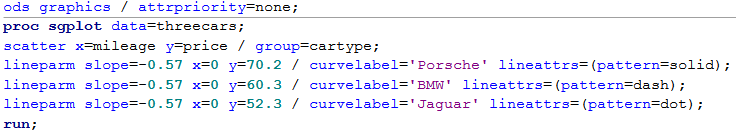
Multiple Regression Line:

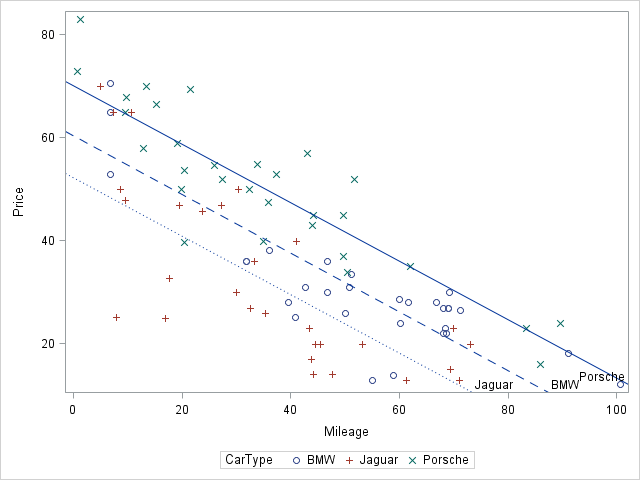
Price(hat) = 60.3 – 0.565\*Mileage + 9.9\*DPorsche – 8.0\*DJaguar

Contains the 3 regression lines:

|  |  |
| --- | --- |
| BMW  (DPorsche=0, DJaguar=0) | Price(hat) = 60.3 – 0.565\*Mileage + 9.9\*(0) – 8.0\*(0)  = 60.3 – 0.565\*Mileage |
| Porsche  (DPorsche=1, DJaguar=0) | Price(hat) = 60.3 – 0.565\*Mileage + 9.9\*(1) – 8.0\*(0)  = (60.3+9.9) – 0.565\*Mileage |
| Jaguar  (DPorsche=0, DJaguar=1) | Price(hat) = 60.3 – 0.565\*Mileage + 9.9\*(0) – 8.0\*(1)  = (60.3-8.0) – 0.565\*Mileage |

Plot of the three lines



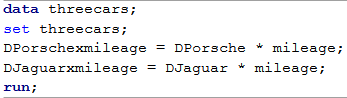


Interpretations

* Coefficient for mileage (-0.565): The predicted price decreases by 0.565 thousand dollars ($565) for every increase in 1000 miles of mileage, considering the same model of car.
* Coefficient for DPorsche (9.9): The predicted price of Porsches is 9.9 thousand dollars more than BMWs, if both cars have the same mileage.
* Coefficient for DJaguar (-8.0): The predicted price of Jaguars is 8.0 thousand dollars less than BMWs, if both cars have the same mileage.
* We see here why BMW (the category we didn’t make the dummy variable for) is called the **reference category**, because the interpretations of the coefficients DPorsche and DJaguar are both in reference (or in comparison) to BMW

Add interaction terms for nonparallel regression lines

* To allow the regression lines to have different slopes, we add the interaction terms
  + DPorsche\*Mileage
  + DJaguar\*Mileage

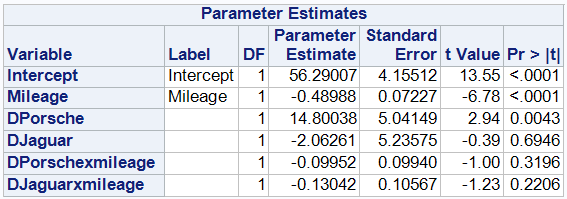


The Model contains the three regression lines

|  |  |
| --- | --- |
| BMW  (DPorsche=0, DJaguar=0) |  |
| Porsche  (DPorsche=1, DJaguar=0) |  |
| Jaguar  (DPorsche=0, DJaguar=1) |  |

* Now each type of car has a different slope, with
  + The slope of Porsche adjusted by from BMW
  + The slope of Jaguar adjusted by from BMW
* Note that the BMW is used as the “reference category” here too.





Multiple Regression Line:

Price(hat) = 56.3 – 0.49\*Mileage + 14.8\*DP – 2.1\*DJ – 0.10\*DP\*mileage –0.13\*DJ\*mileage

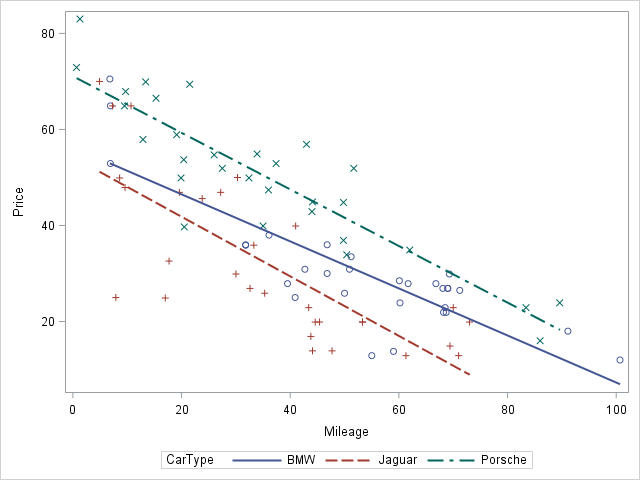
Contains the 3 regression lines:

* BMW (DPorsche=0, DJaguar=0)  
  Price(hat) = 56.3 – 0.49\*Mileage + 14.8\*(0) – 2.1\*(0) – 0.10\*(0)\*mileage –0.13\*(0)\*mileage  
   = 56.3 – 0.49\*Mileage
* Porsche (DPorsche=1, DJaguar=0)  
  Price(hat) = 56.3 – 0.49\*Mileage + 14.8\*(1) – 2.1\*(0) – 0.10\*(1)\*mileage –0.13\*(0)\*mileage  
   = (56.3+14.8) + (-0.49-0.10)\*Mileage
* Jaguar (DPorsche=0, DJaguar=1)  
  Price(hat) = 56.3 – 0.49\*Mileage + 14.8\*(0) – 2.1\*(1) – 0.10\*(0)\*mileage –0.13\*(1)\*mileage  
   = (56.3-2.1) + (-0.49-0.13)\*Mileage

Interpretations of coefficients

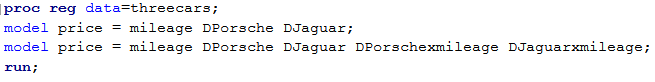
* Mileage (-0.49): this is the slope of the line for BMW only  
  For BMWs, the predicted price decreases by 0.49 thousand dollars for every increase of 1000 miles.
* DPorsche\*mileage (-0.10): the slope of the line for Porsche is 0.10 less than BMW’s line.  
  The predicted price of Porsches decreases by 0.10 thousand dollars more for every increase of 1000 miles than BMWs.
* DJaguar\*mileage (-0.13): the slope of the line for Jaguar is 0.13 less than BMW’s line.  
  The predicted price of Jaguars decreases by 0.13 thousand dollars more for every increase of 1000 miles than BMWs.

In the plot earlier, we can see that the line for BMW is less steep than the other two.



Comparing models

You can use multiple MODEL statements under the same PROC REG



|  |  |
| --- | --- |
| Parallel lines | Nonparallel lines |
|  |  |

* Note: you can keep track of which output is for which model by looking at the DF(Model), which gives the number of explanatory variables.

|  |  |  |
| --- | --- | --- |
|  | Adjusted R-squared | Root MSE |
| Parallel lines | 0.7610 | 8.62 |
| Nonparallel lines | 0.7602 | 8.64 |

* The parallel lines model is a better fit according to adjusted R-squared (higher) and root MSE (lower).

# Handout 16 – One-way ANOVA

STA 216 F19

The models we will cover in this course are:

* Two-sample model:
  + Quantitative Y, one Categorical X with 2 categories
* Simple linear regression model:
  + Quantitative Y, one Quantitative X
* Multiple Linear Regression:
  + Quantitative Y, multiple X’s (at least one quantitative)
* One-way ANOVA:
  + Quantitative Y, one Categorical X
  + Extends two-sample model to when categorical X has 3 or more categories
* Two-way ANOVA:
  + Quantitative Y, two Categorical X’s
* What about a categorical Y?
  + Logistic regression handles the case of a binary Y
  + In STA 310 (biostatistics) and STA 321 (regression)

Example

Does reproductive behavior reduce longevity in fruit flies? A study by Partridge and Farquhar (1981). It was already know that increased reproduction leads to shorter life spans for female fruit flies. But the question remained whether an increase in sexual activity would also reduce the life spans of male fruit flies. The researchers designed an experiment to answer this question. They had a total of 125 male fruit flies to use, and they randomly assigned each of the 125 to one of the following five groups:

* 8 virgins: Each male fruit fly assigned to live with 8 virgin female fruit flies.
* 1 virgin: Each male fruit fly assigned to live with 1 virgin female fruit fly.
* 8 pregnant: Each male fruit fly assigned to live 8 pregnant female fruit flies. (The theory was that pregnant female fruit flies would be receptive to sexual relations.)
* 1 pregnant: Each male fruit fly assigned to live with 1 pregnant female fruit fly
* None: Each male fruit fly was subjected to a lonely existence without females at all.

The lifespan of each male fruit fly, in days, was recorded.

|  |  |
| --- | --- |
|  | img0.png |

One-way ANalysis Of VAriance (ANOVA) model

* + Y is the response variable (longevity)
  + is the group mean, the population mean longevity in the kth group, k=1,2,…,K
  + In general, let K be the number of groups. (K=5 here).
  + E is the error term. As before, we assume
* We can write
  + is the “grand m/ean” (the average of the group means)
  + is the “group effect”

One mean or several?

* The major question we start off is: Are all the means the same or is there some difference?
* We can write this with null/alternative hypotheses as
  + Ho: vs. Ha: at least one is different
* Another way to write this:
  + Ho: vs. Ha: at least one

Model assumptions

* The 4 model assumptions are the same as they were for regression.
* They can all be write in terms of the error terms
  + The errors have mean zero
  + Constant variance: the errors have the same standard deviation in each group
  + Normality
  + Independence

Estimating the model terms

* Estimate of grand mean is , the sample mean of all the responses

|  |  |
| --- | --- |
|  |  |

* Estimate of group mean is , the sample mean of the responses in group k

|  |  |
| --- | --- |
|  |  |

* Estimate of group effect is

|  |  |
| --- | --- |
| Group | Group effect estimate |
| 1 pregnant | 64.80 – 57.44 = 7.36 |
| 1 virgin | 56.76 – 57.44 = -0.68 |
| 8 pregnant | 63.36 – 57.44 = 5.92 |
| 8 virgin | 38.72 – 57.44 = -18.72 |
| None | 63.56 – 57.44 = 6.12 |

* Estimate of error terms E is the residual
* As before, the residual (actual minus predicted),   
  where the predicted Y here is the sample group mean

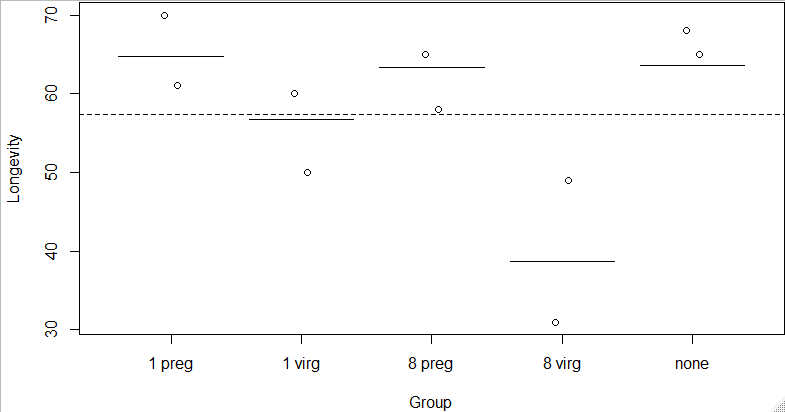
Example

One fruit fly in the 1 pregnant group lived Y = 70 days

* Because the group mean for the 1 pregnant group was 64.8,   
  his residual was 70 – 64.8 = 5.2 days.
* Note that the observation can be decomposed as   
  Y = grand mean + group effect + residual  
  70 = 57.44 + 7.36 + 5.2

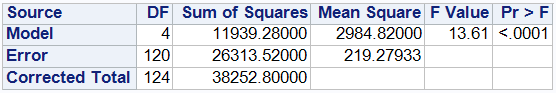
Graphical representation

Showing just two fruit flies per group (instead of 25),



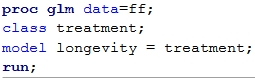
Sum of squares 🡪 ANOVA Table

* Like in regression, we have SS(total) = SS(model) + SS(error)
  + SS(total) =
  + SS(Model) =
  + SS(error) = = sum of squared residuals
* For one-way ANOVA, the SS(model) is the sum of the squared group effects



* DF (degrees of freedom)
  + DF(Model) = K - 1
  + DF(Error) = n – K
  + DF(Total) = n – 1
* Mean squares (MS) = Sum of squares DF
  + MS(Model) = SS(Model) / DF(Model) = 11,939.28 / 4 = 2984.82
  + MS(Error) = MSE = SS(E) / DF(Error) = 26313.52 / 120 = 219.28
* F Value: F = MS(Model) / MS(error) = 2984.82 / 219.28 = 13.61
* Pr > F: P-value for the F test
  + Comes from the F distribution with numerator df = DF(Model) and   
    denominator df = DF(Error)
* This test statistic and p-value are for the test of equality of means  
  Ho: vs. Ha: at least one is different
* P-value < .0001 🡪 Reject Ho
* Thus, in this case, there is statistically significant evidence that the population mean longevity of at least one of the 5 groups is different.

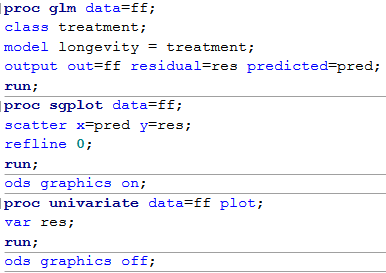
SAS code

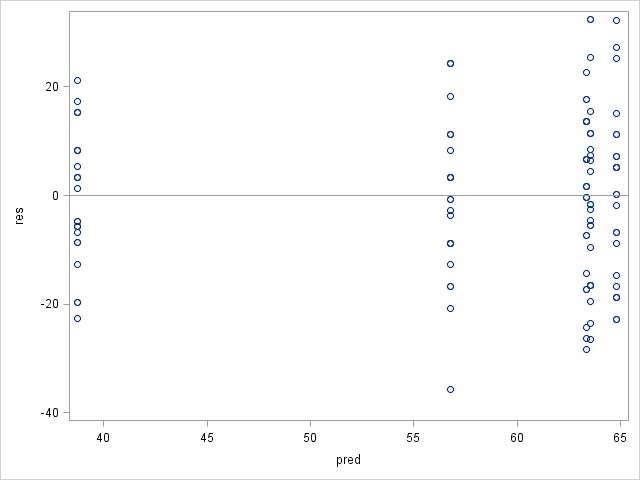


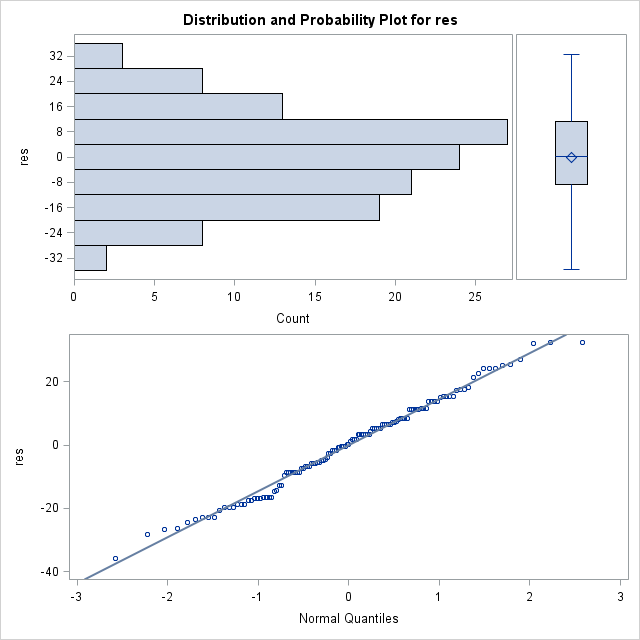
Checking assumptions

Similar process as in regression.

* Difference: we do not check zero mean error assumption because the residuals in each group will always be centered around 0.
* Constant variance and normality





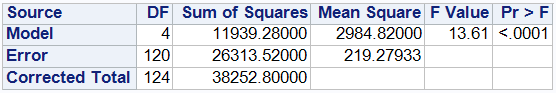


# Handout 17 – Comparing means

STA 621 F19

Continuing with fruit fly example from Handout 16

|  |  |
| --- | --- |
| img0.png |  |



Recall that the F test given in the ANOVA table is testing whether

Ho: vs. Ha: at least one is different

P-value < .0001 🡪 Reject Ho

There is statistically significant evidence that the population mean longevity of at least one of the 5 groups is different.

Follow-up question:

* Which means are different?
* Strategy: compare pairs of group means: (1P, 1V), (1P, 8P), … there are 10 pairs in all.
* We can start by comparing the sample means.
* To extend the results of the sample to the population, we can make confidence intervals for

Confidence interval for

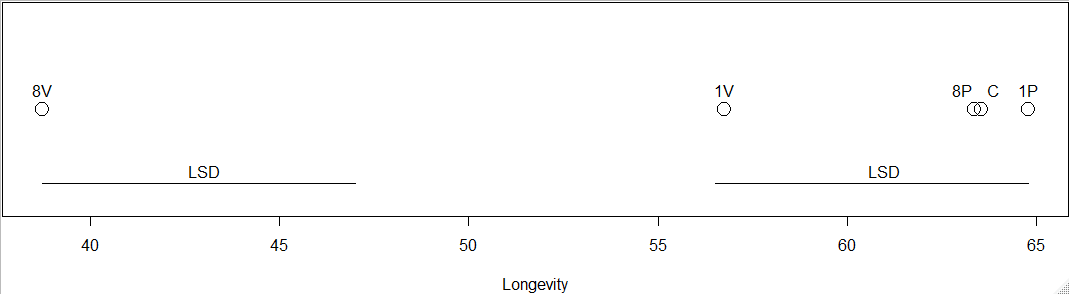
|  |  |
| --- | --- |
| 95% CI for mu[1P] – Mu[1V] | 95% CI for mu[1P] – mu[8V] |
| , ,   for df = 120 (error df) s = root MSE =  CI = = (-0.25, 16.33) | , ,   for df = 120 (error df) s = root MSE =  CI = = (17.79, 34.37) |

* 95% CI for mu[1P] – Mu[1V] = (-0.25, 16.33)  
  Interpretation: [Because the CI contains zero] We cannot be 95% confident that the population means of longevity for male fruit flies living with 1 pregnant female fruit fly and living with 1 virgin female fruit fly are different.
* 95% CI for mu[1P] – mu[8V] = (17.79, 34.37)  
  Interpretation: We are 95% confident that the population mean longevity for male fruit flies living with 1 pregnant female fruit fly is **greater than** the population mean longevity for male fruit flies living with 8 virgin female fruit flies by between 17.79 and 34.37 days.

Recall:

* If the confidence interval for a difference in means doesn’t contain zero 🡪 there is a significant difference between the means.
* The CI won’t contain zero if the (absolute value of the) difference in the means is greater than the margin of error of the interval.
* Examples:
  + CI for mu[1P] – Mu[1V] = = (-0.25, 16.33)
  + 95% CI for mu[1P] – mu[8V] = = (17.79, 34.37)
* For this reason, the margin of error is also called the **least significant difference (LSD)**
* Note: if all groups have the same sample size (called a balanced experiment), the margin of error of every the CI for every pairwise difference will be the same. 🡪 Same LSD.

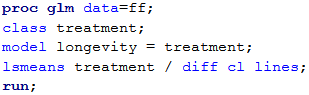
Graphical representation

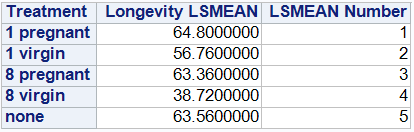


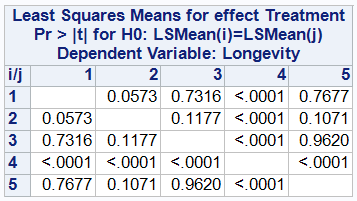
* 8V is farther than the LSD away from all the other groups
* All the other groups are within the LSD of each other.

In SAS

* Use LSMEANS statement within PROC GLM

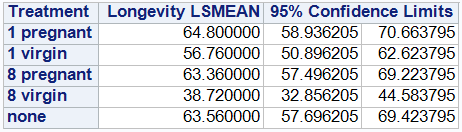


 Numbers the groups

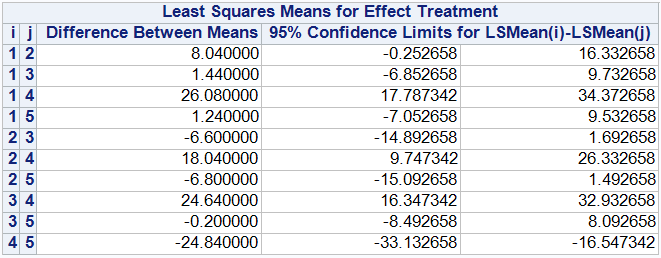


P-values for hypothesis tests of Ho: vs. Ha:

* .0573 in row 1, column 2 is for the comparison of 1P and 1V (following numbering above)
* Note that all the p-values in row/column 4 (corresponding to 8V) are <.0001, representing that 8V is significantly different from all the other groups.

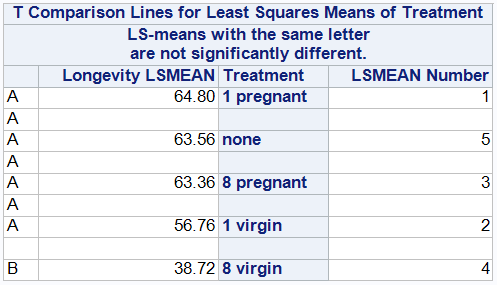


CI’s of population means



CIs of

* First row is for , which we did by hand earlier.



Note message: “LS-means with the same letter are not significantly different”. This is referring to the letters in the first column.

* 1pregnant, none, 8pregnant, and 1virgin all share “A” 🡪 they are not significantly different
* 8virgin has “B”, does not share a letter with other categories 🡪 significantly different from all other groups.

Controlling Type 1 Error rate for multiple comparisons

* Recall that is the probability of a Type 1 Error (false positive).
* When we’re doing multiple tests when we compare all the pairs of means, it increases the chances of a false positive.
* Comic: <https://xkcd.com/882/>
  + If we do 20 tests at , we expect 1 out of the 20 tests to be a false positive.
  + So find the significant result about the green jelly beans is really meaningless.
* Instead, we can focus on the controlling the family-wise error rate, the probability of getting at least 1 Type 1 Error on all the tests.
* Two methods of controlling family-wise error rate: Fisher’s LSD and Bonferroni.

Fisher’s LSD

1. Do the ANOVA F test
2. If the F test is not significant, **STOP**
3. If the F test is significant, do the pairwise comparisons using the LSD.

Bonferroni method

* Do the pairwise comparisons right away, but multiply the p-values times the number of tests.
* Suppose there are 10 tests.
  + Then the individual error rate for each test is
  + The family-wise error rate is

|  |  |
| --- | --- |
|  | Compare with unadjusted p-values |
|  |  |

# Handout 18 – Two-way ANOVA (no interaction)

STA 216 F19

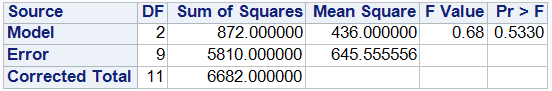
* In one-way ANOVA, we were modeling a quantitative response variable with a single categorical explanatory variable.
* Examples:
  + Handouts 16-17: Y = Longevity of fruit flies, X = cohabitants (1 pregnant, 1 virigin, etc.)
  + Lab 8: Y = Systolic blood pressure, X = Weight category (normal, overweight, or obese)  
    or X = Exercise Level (low, medium, or high)
* In two-way ANOVA, we use 2 categorical explanatory variables.
* Example: Use X1 = weight category and X2 = Exercise level in the same model.

Example

Are you affected by caffeine? What about chocolate? In their paper Scott and Chen (1994), scientists published research that compare the effects of caffeine with those of theobromine (a similar chemical found in chocolate) and with those of a placebo. Their experiment used 4 human subjects and took place over three days. Each day, each subject swallowed a tablet containing one of caffeine, theobromine, or the placebo. Two hours later, they were timed while tapping a finger in a specified manner (that they had practiced earlier, to control for learning effects). The response is the number of taps over a fixed time interval, which are shown in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Subject | Placebo | Caffeine | Theobromine | Mean |
| I | 11 | 26 | 20 | 19 |
| II | 56 | 83 | 71 | 70 |
| III | 15 | 34 | 41 | 30 |
| IV | 6 | 13 | 32 | 17 |
| Mean | 22 | 39 | 41 | 34 |

A critical feature of this study was the way the pills were assigned to subjects. Each subject was given one pill of each kind, so that over the course of the experiment, each drug was given to all four subjects. In order to protect against possible bias from carryover effects, subjects were given the drugs in an order determined by chance, with a separate randomized order for each subject.

* Response: tap rate
* Two categorical explanatory variables (also called factors)
  + drugs (factor of interest) with levels: placebo, caffeine, theobromine
  + subjects (factor creating nuisance variation)
* Note that different means of tap rate are given in the table
  + Right-most column: means by subject 🡪 subjects I and IV are slow, II is fast
  + Bottom row: means by drug 🡪 caffeine and theobromine are close together, almost double of placebo
  + Bottom-right corner: grand mean
* Results from one-way ANOVA with drug only  
  
* Not statistically significant evidence of difference between means of tap rate from different drugs
* Because variation due to subject makes the SS(error) too large. 🡪 Need to account for subjects

Two-way ANOVA model (without interaction)

* Response variable Y
* Two categorical factors
  + Factor A with K levels
  + Factor B with J levels
* The model:

where

* + : grand mean
  + : the effect of the kth level of Factor A
  + : the effect of the jth level of Factor B
  + E: error term, assume that and are independent.
* Same 4 assumptions: zero mean errors, constant variance, normality, independence
* For our example, Y = tap rate, factor A = drug (with K=3 levels), B = subject (with J=4 levels)

Estimates of parameters

* grand mean of Y
* (mean over factor A level k) – grand mean
* (mean of factor B level j) – grand mean
* Estimate of : root MSE

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Subject | Placebo | Caffeine | Theobromine | Mean |
| I | 11 | 26 | 20 | 19 |
| II | 56 | 83 | 71 | 70 |
| III | 15 | 34 | 41 | 30 |
| IV | 6 | 13 | 32 | 17 |
| Mean | 22 | 39 | 41 | 34 |

For our example,

* Alphahats (Drug effects, where 1:Placebo, 2:Caffeine, 3:Theobromine)
* Betahats (Subject effects)
* Predicted values can then be calculated as (grand mean) + (drug effect) + (subject effect)
  + For subject I on Placebo, this is = 34 + (-12) + (-15) = 7
* Residuals are as before.
  + For subject I on Placebo, Y = 11 so residual = 11 – 7 = 4

I added a row and column to the previous table for the subject and drug effects

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Subject | Placebo | Caffeine | Theobromine | Mean | Effect |
| I | 11 | 26 | 20 | 19 | -15 |
| II | 56 | 83 | 71 | 70 | 36 |
| III | 15 | 34 | 41 | 30 | -4 |
| IV | 6 | 13 | 32 | 17 | -17 |
| Mean | 22 | 39 | 41 | 34 |  |
| Effect | -12 | 5 | 7 |  |  |

Partitioning variation

(Y – grand mean) = (Drug effect) + (Subject effect) + (Residual)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Subj | Drug | Y - grand | Drug Eff | Subj Eff | Res |
| I | Pl | 11 - 34 | -12 | -15 | 4 |
| I | Caf | 26 - 34 | 5 | -15 | 2 |
| I | Th | 20 - 34 | 7 | -15 | -6 |
| II | Pl | 56 - 34 | -12 | 36 | -2 |
| II | Caf | 83 - 34 | 5 | 36 | 8 |
| II | Th | 71 – 34 | 7 | 36 | -6 |
| III | Pl | 15 – 34 | -12 | -4 | -3 |
| III | Caf | 34 – 34 | 5 | -4 | -1 |
| III | Th | 41 – 34 | 7 | -4 | 4 |
| IV | Pl | 6 – 34 | -12 | -17 | 1 |
| IV | Caf | 13 – 34 | 5 | -17 | -9 |
| IV | Th | 32 - 34 | 7 | -17 | 8 |

Sums of squares are used to quantify the amount of variation due to each source.

* SS(total) = sum of (Y – grand mean)2
* SS(drug) = sum of (Drug Eff)2
* SS(subject) = sum of (Subj Eff)2
* SS(error) = sum of (Residual)2
* Note that SS(total) = SS(drug) + SS(subject) + SS(error)

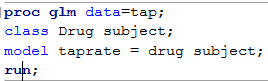
These sums of squares provide the basis for the ANOVA table:

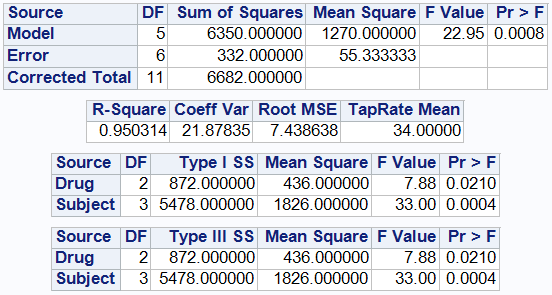
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | DF | SS | MS | F stat | P-value |
| Factor A | K-1 | SS(A) | MS(A) = SS(A) / (K-1) | MS(A) / MS(E) |  |
| Factor B | J-1 | SS(B) | MS(B) = SS(B) / (J-1) | MS(B) / MS(E) |  |
| Error | (K-1)(J-1) | SS(error) | MS(E) = SS(E) / (K-1)(J-1) |  |  |
| Total | KJ - 1 | SS(total) |  |  |  |

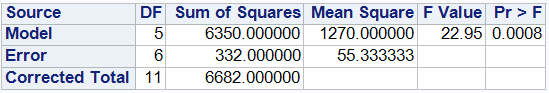
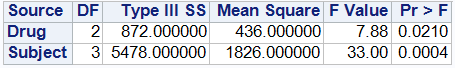
This contains two F tests:

* For factor A: vs.
* For factor B: vs.

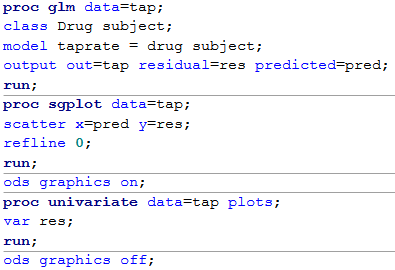
SAS code





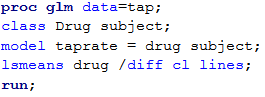
* In the ANOVA table at the top  
    
  the Source=Model row combines both the Drug and Subject effects.
* The drug and subject effects are shown separately below  
    
  (For STA 216, there’s no difference in the Type I SS and Type III SS tables.)
* Thus, the test of the drug effect has
  + F stat = MS(drug) / MS(error) = 436 / 55.33 = 7.88
  + P-value = .0210
  + At alpha = .05, there is statistically significant evidence that the population mean tap rates differ for at least one of the drugs.

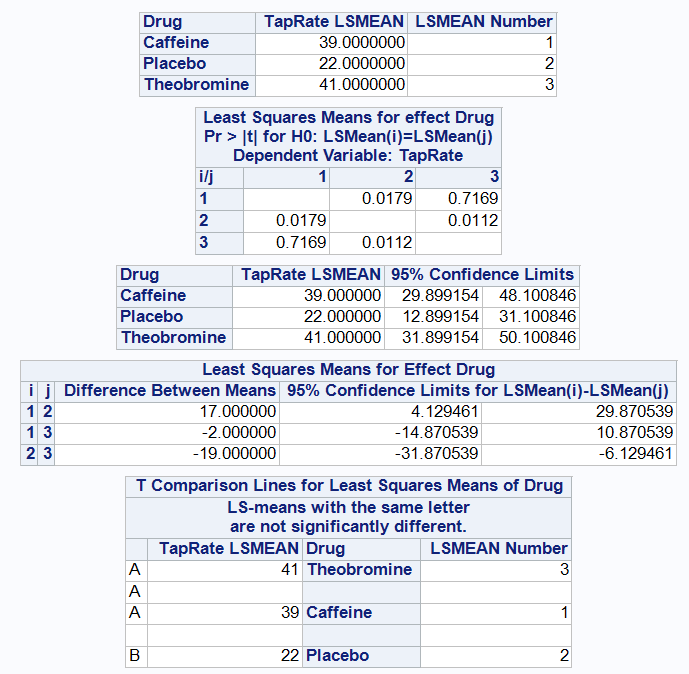
Checking assumptions



|  |  |
| --- | --- |
| img0.png | img1.png |
| Zero mean errors?  - Residuals centered around zero  Constant variance?  - Constant spread of residuals | Normality?  - Symmetric histogram / boxplot  - Points following line in QQ plot |

Pairwise comparisons of drug means





# Handout 19 – Two-way ANOVA with interaction

STA 216 F19

* What is interaction?

Example

A scientist in Iowa was interested in additives to standard pig chow that might increase the rate at which the pigs gained weight. For example, he could add antibiotics or he could add vitamin B12. He could also include both additives in a diet or leave both out (using just standard pig chow as a control). If we let Factor A be yes or no depending on whether the diet has antibiotics and let Factor B keep track of the presence or absence of vitamin B12, we can summarize the four possible diets in a two-way table. To perform the experiment, the scientist randomly assigned 12 pigs, 3 to each of the diet combinations. Their daily weight changes (in hundredths of a pound over 1.00) are summarized below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Factor B (B12) | |
| No | Yes |
| Factor A (Antiobiotics) | No | 30, 19, 8 | 26, 21, 19 |
| Yes | 5, 0, 4 | 52, 56, 54 |

Means for each of four diets

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Factor B (B12) | |
| No | Yes |
| Factor A (Antibiotics) | No | 19 | 22 |
| Yes | 3 | 54 |

Effects of B12:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Factor B (B12) | | Difference due  to B12 |
| No | Yes |
| Factor A (Antibiotics) | No | 19 | 22 | 3 |
| Yes | 3 | 54 | 51 |

Effects of Antibiotics

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Factor B (B12) | |
| No | Yes |
| Factor A (Antibiotics) | No | 19 | 22 |
| Yes | 3 | 54 |
| Difference due to Antibiotics | | -16 | 32 |

* Treatment = Combination of levels of the factors  
  In this case, there are 4 treatments (the 4 diets).
* Interaction can be thought of as a “difference of differences”. For instance,
  + The difference in the weight gain due to adding B12 to the diet is different depending on whether the pigs get antibiotics (3 vs. 51)
  + The difference in weight gain due to adding antibiotics to the diet is different depending on whether the pigs got B12 (-16 vs. 32).
* Because there is a difference between the differences, there is an interaction between antibiotics and vitamin B12 in how they affect the weight gain.

Interaction plots

* Interaction plots give visual method of looking at these differences in treatment means.
  + Y-axis: values of response
  + X-axis: Different locations for levels of one factor
  + Lines on the plot connect the means for the same level of the other factor.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Factor B (B12) | | Difference due  to B12 |
| No | Yes |
| Factor A (Antibiotics) | No | 19 | 22 | 3 |
| Yes | 3 | 54 | 51 |
| Difference due to Antibiotics | | -16 | 32 |  |

|  |  |
| --- | --- |
|  | Effects due to antibiotics  img0.png |
|  | Effects due to B12  img1.png |

* Parallel lines indicates that there is no interaction
* Non-parallel lines indicate that there is an interaction

Two-way ANOVA model with interaction

where

* : grand mean
* : the effect of the kth level of Factor A
* : the effect of the jth level of Factor B
* : interaction effect for the kth level of A with the jth level of B
* E: error term, assume that and are independent.
* Same 4 assumptions: zero mean errors, constant variance, normality, independence
* New term: introduces interaction

Estimates of parameters

* grand mean of Y
* (mean over factor A level k) – grand mean
* (mean of factor B level j) – grand mean
* (cell mean of factor A level k with factor B level j) – [
* Estimate of : root MSE

Cell means, row means, column means, grand mean

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Factor B (B12) | | Row  Mean |
| No | Yes |
| Factor A (Antibiotics) | No | 19 | 22 | 20.5 |
| Yes | 3 | 54 | 28.5 |
| Column Mean | | 11 | 38 | 24.5 |

For our example,

* Alphahats (Antibiotics effects, where 1:No, 2:Yes)
* Betahats (B12 effects, where 1:No, 2:Yes)
* Gammahats
  + (cell mean 11) – [  
     = 19 – [24.5 + (-4) + (-13.5)] = 12
  + (cell mean 12) – [  
     = 22 – [24.5 + (-4) + 13.5] = -12
* Predicted values can then be calculated as   
  (grand mean) + (antibiotic effect) + (B12 effect) + (interaction effect) = cell mean
* For pigs not on antibiotics or B12, this is  
  = 24.5 + (-4) + (-13.5) + 12 = 19
* There are residuals for each observation.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Factor B (B12) | |
| No | Yes |
| Factor A (Antiobiotics) | No | 30, 19, 8 | 26, 21, 19 |
| Yes | 5, 0, 4 | 52, 56, 54 |

Means for each of four diets

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Factor B (B12) | |
| No | Yes |
| Factor A (Antibiotics) | No | 19 | 22 |
| Yes | 3 | 54 |

Residuals

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Factor B (B12) | |
| No | Yes |
| Factor A (Antibiotics) | No | 11, 0, -11 | 4, -1, -3 |
| Yes | 2, -3, 1 | -2, 2, 0 |

Partitioning variation

(Y – grand mean) = (Antibiotic eff) + (B12 effect) + (Interaction effect) + (Residual)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Antb | B12 | Y - grand | Antb eff | B12 eff | Int eff | Res |
| No | No | 30 – 24.5 | -4 | -13.5 | 12 | 11 |
| No | No | 19 – 24.5 | -4 | -13.5 | 12 | 0 |
| No | No | 8 – 24.5 | -4 | -13.5 | 12 | -11 |
| No | Yes | 26 – 24.5 | -4 | 13.5 | -12 | 4 |
| No | Yes | 21 – 24.5 | -4 | 13.5 | -12 | -1 |
| No | Yes | 19 – 24.5 | -4 | 13.5 | -12 | -3 |
| Yes | No | 5 – 24.5 | 4 | -13.5 | -12 | 2 |
| Yes | No | 0 – 24.5 | 4 | -13.5 | -12 | -3 |
| Yes | No | 3 – 24.5 | 4 | -13.5 | -12 | 1 |
| Yes | Yes | 52 – 24.5 | 4 | 13.5 | 12 | -2 |
| Yes | Yes | 56 – 24.5 | 4 | 13.5 | 12 | 2 |
| Yes | Yes | 54 – 24.5 | 4 | 13.5 | 12 | 0 |

Sums of squares are used to quantify the amount of variation due to each source.

* SS(total) = sum of (Y – grand mean)2
* SS(Antb) = sum of (Antb Eff)2
* SS(B12) = sum of (B12 Eff)2
* SS(Int) = sum of (Int eff)2
* SS(error) = sum of (Residual)2
* Note that SS(total) = SS(Ant) + SS(B12) + SS(Int) + SS(error)

These sums of squares provide the basis for the ANOVA table:

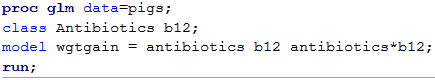
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | DF | SS | MS | F stat | P-value |
| Factor A | K-1 | SS(A) | MS(A) = SS(A) / (K-1) | MS(A) / MS(E) |  |
| Factor B | J-1 | SS(B) | MS(B) = SS(B) / (J-1) | MS(B) / MS(E) |  |
| A x B | (K-1)(J-1) | SS(AB) | MS(AB) = SS(AB) / (K-1)(J-1) | MS(AB) / MS(E) |  |
| Error | KJ(r-1) | SS(error) | MS(E) = SS(E) / [KJ(r-1)] |  |  |
| Total | n - 1 | SS(total) |  |  |  |

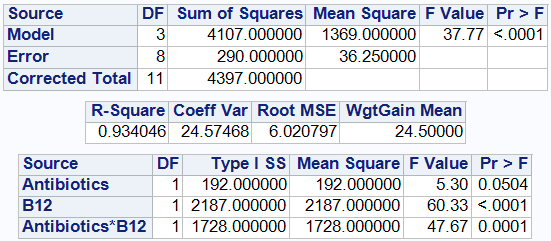
where r is the number of observations per treatment.

This contains three F tests:

* For factor A: vs.
* For factor B: vs.
* For interaction: All vs.

SAS code



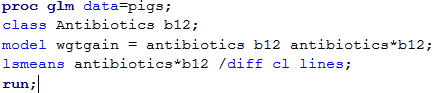


* Looking at the Antibiotics\*B12 line, there is statistically significant evidence of an interaction between antibiotics and B12.
  + F stat = MS(Antb\*B12) / MS(error) = 1728 / 36.25 = 47.67
  + P-value = .0001
* Because of the interaction, we have seen a “difference between the differences”.
* Therefore, it doesn’t make sense to speak of a single effect of each factor (called a main effect) because the effect depends on the other factor.
* NO: Average effect of adding B12 is a 27 unit increase.
  + YES: without antibiotics, B12 gives 3 unit increase  
     with antibiotics, B12 gives a 51 unit increase
* NO: Average effect of adding antibiotics is a 8 unit increase
  + YES: without B12, antibiotics give 16 unit **decrease**

With B12, antibiotics give a 32 unit increase

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Factor B (B12) | | Difference due  to B12 |
| No | Yes |
| Factor A (Antibiotics) | No | 19 | 22 | 3 |
| Yes | 3 | 54 | 51 |
| Difference due to Antibiotics | | -16 | 32 |  |

Comparisons of treatment means



|  |  |
| --- | --- |
|  |  |