Selected Statistical Methods in SAS and Output Delivery System (ODS)

STAT 3505

Week 13 (April 11, 2024)

Gunes Fleming Ph.D.

Simple Linear Regression

- The regression line that SAS calculates from the data is an estimate of a theoretical line describing the relationship between the independent variable (X) and the dependent variable (Y).
- The theoretical line is $Y = \alpha + \beta x + \varepsilon$

where α is the y-intercept, β is the slope, and ε is an error term that is normally distributed with zero mean and constant variance.

- It should be noted that β = 0 indicates that there is no linear relationship between X and Y.
- A simple linear regression analysis is used to develop an equation (a linear regression line) for predicting the dependent variable given a value (x) of the independent variable.

Simple Linear Regression

- The regression line calculated by SAS is given by $\hat{Y} = a + bx$

  where a and b are the least-squares estimates of α and β.

- The null hypothesis that there is no predictive linear relationship between the two variables is that the slope of the regression equation is zero.

- Specifically, the hypotheses are: $H_0: \beta = 0$
  $H_a: \beta \neq 0$

- A low p-value for this test (say, less than 0.05) indicates significant evidence to conclude that the slope of the line is not 0; that is, that knowledge of X would be useful in predicting Y.

Simple Linear Regression: PROC REG

Syntax

- The general syntax for PROC REG is

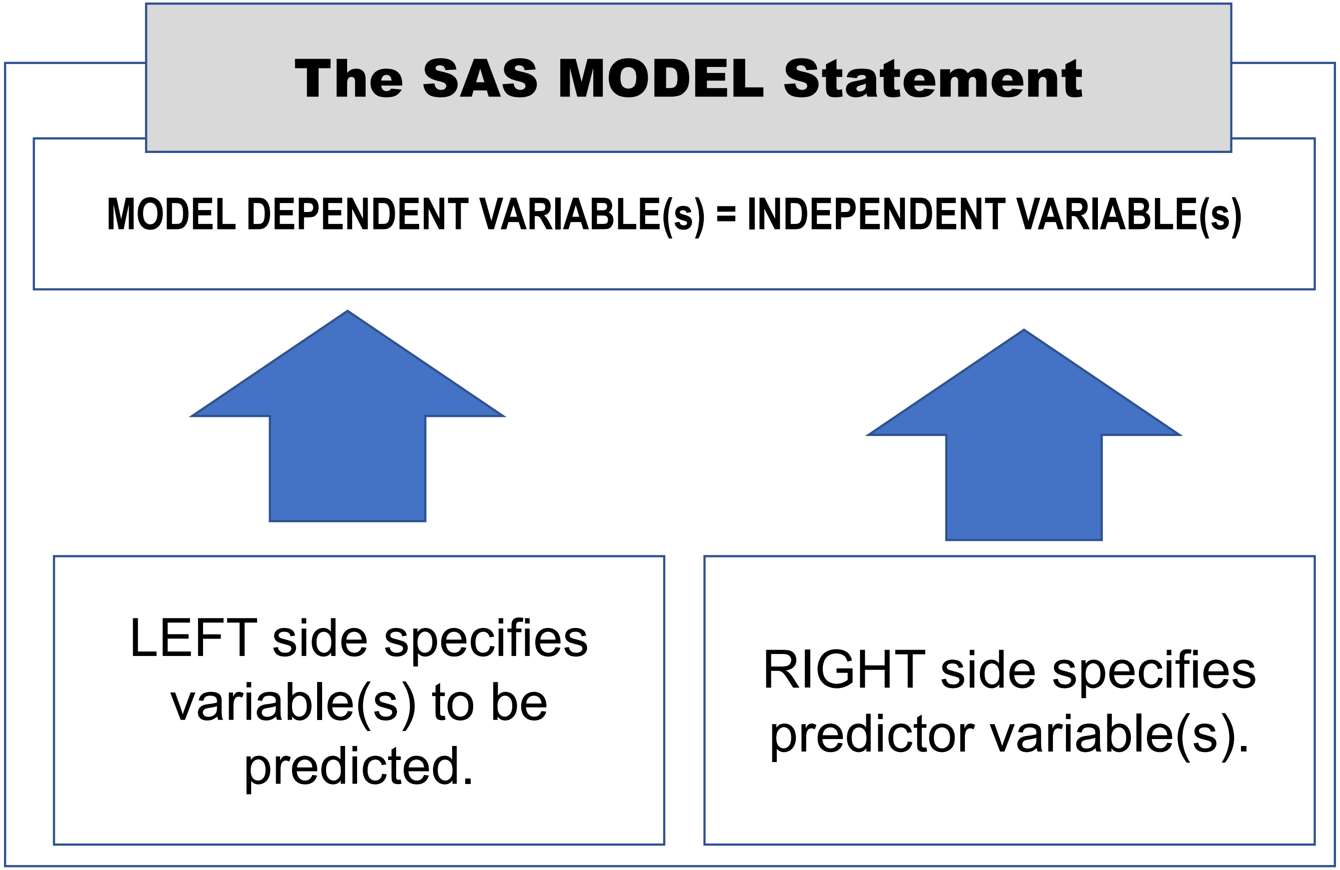
PROC REG <options>; <statements>;

- Common OPTIONS are
  - **DATA=*datsetname*** **-** Specifies dataset.
  - **SIMPLE -** Displays descriptive statistics.

Common Statements

- **MODEL** dependentvar = independentvar</options>;
- **BY groupvariable;** Produces separate regression analyses for each value of the BY variable

- Note: Several MODEL options are available, but we will defer their discussion to the section on using SAS for multiple linear regression.



The SAS MODEL Statement

MODEL DEPENDENT VARIABLE(s) = INDEPENDENT VARIABLE(s)

LEFT side specifies variable(s) to be predicted.

RIGHT side specifies predictor variable(s).

Simple Linear Regression: PROC REG

- Example: A random sample of fourteen elementary school students is selected from a school, and each student is measured on a creativity score (X) using a new testing instrument and on a task score (Y) using a standard instrument. The task score is the mean time taken to perform several hand–eye coordination tasks. Because administering the creativity test is much cheaper, the researcher wants to know if the CREATE score is a good substitute for the more expensive TASK score.

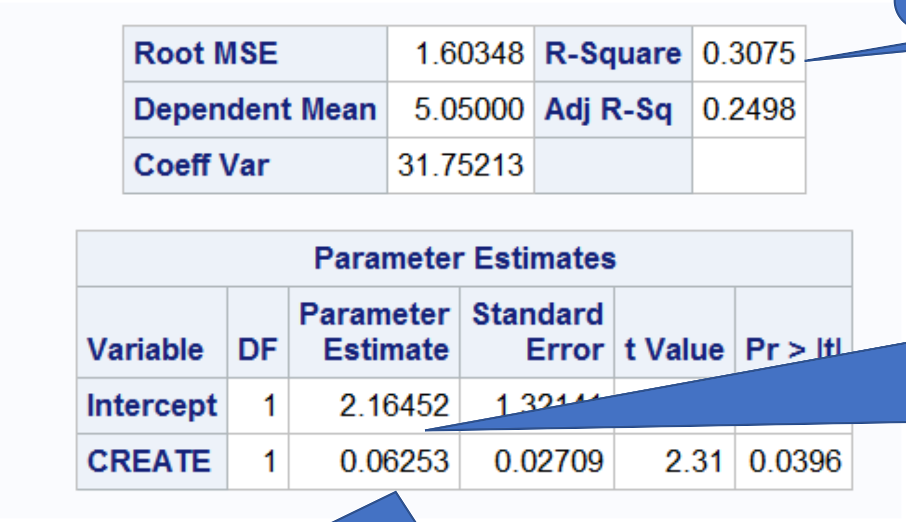
EXAMPLE

```
(data entered)
PROC REG;
MODEL TASK=CREATE;
TITLE 'Example simple linear
regression using PROC REG';
RUN;
QUIT;
```

The MODEL statement defines the linear regression equation you are calculating.

A QUIT statement is recommended for PROC REG to end the analysis.

Selected Output from PROC REG



Root MSE	1.60348	R-Square	0.3075
Dependent Mean	5.05000	Adj R-Sq	0.2498
Coeff Var	31.75213		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.16452	1.32144	[illegible]	[illegible]
CREATE	1	0.06253	0.02709	2.31	0.0396

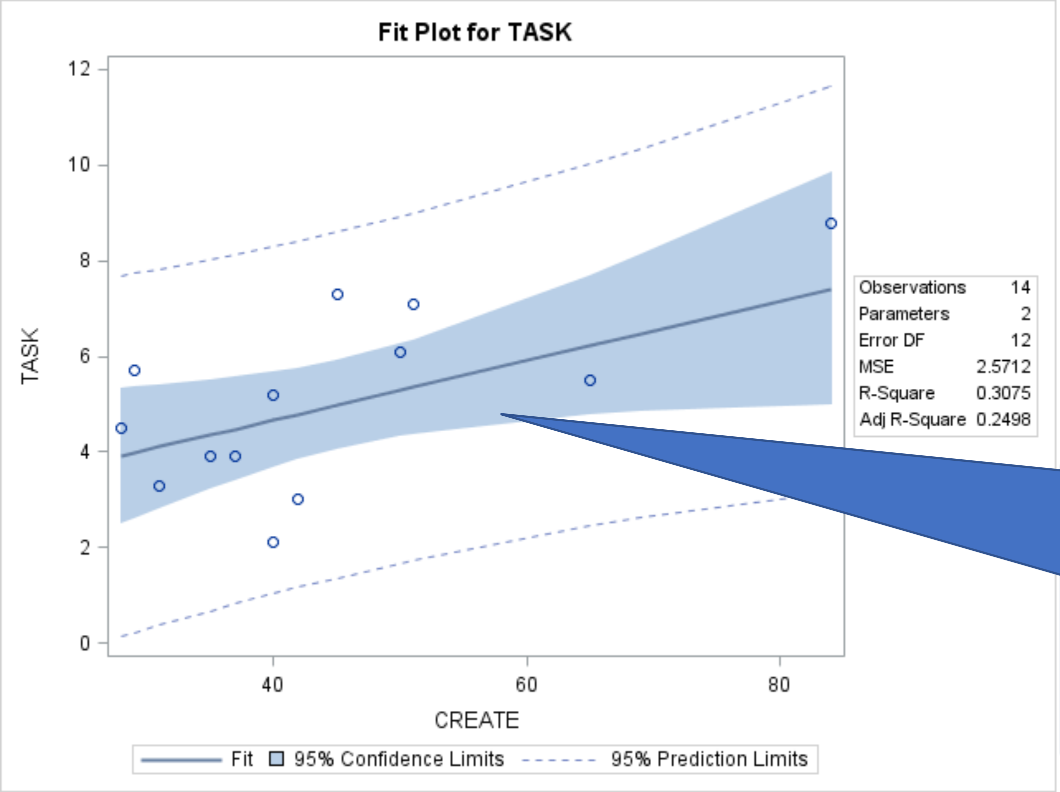
R-Squared is a measure of the strength of the association.

The regression equation from this analysis is

TASK = 2.16+0.0625*CREATE

The parameter estimates are the estimates of alpha (Intercept) and beta (slope/CREATE).

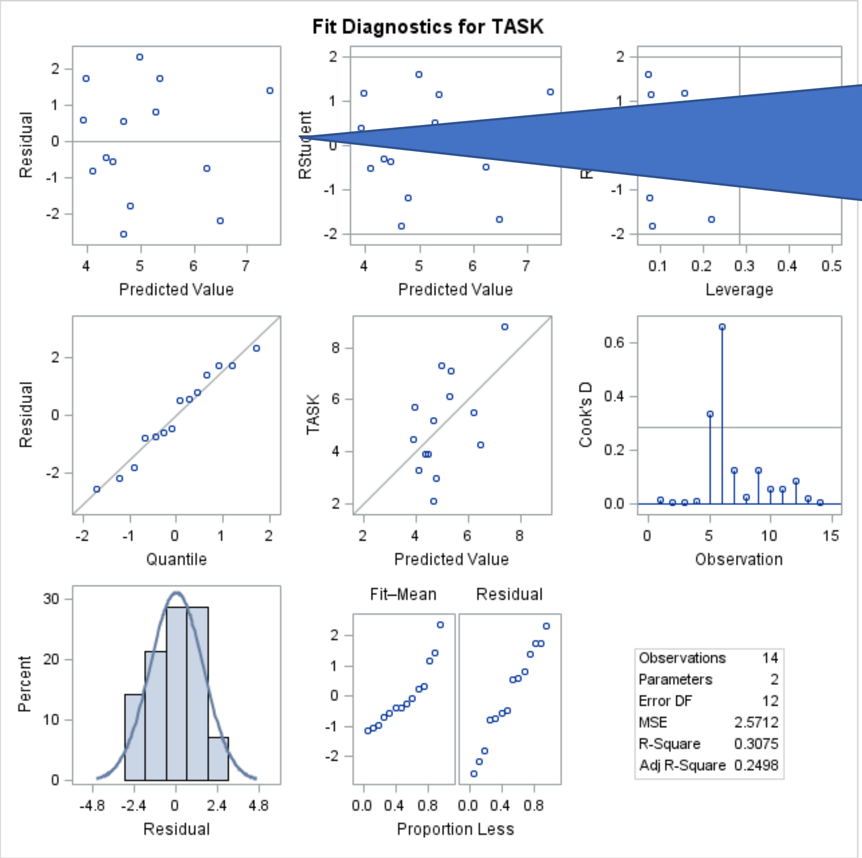
Graphical Results of Regression Analysis



Observations	14
Parameters	2
Error DF	12
MSE	2.5712
R-Square	0.3075
Adj R-Square	0.2498

The shaded area represents a 95% confidence interval for the average TASK score for a given CREATE score.

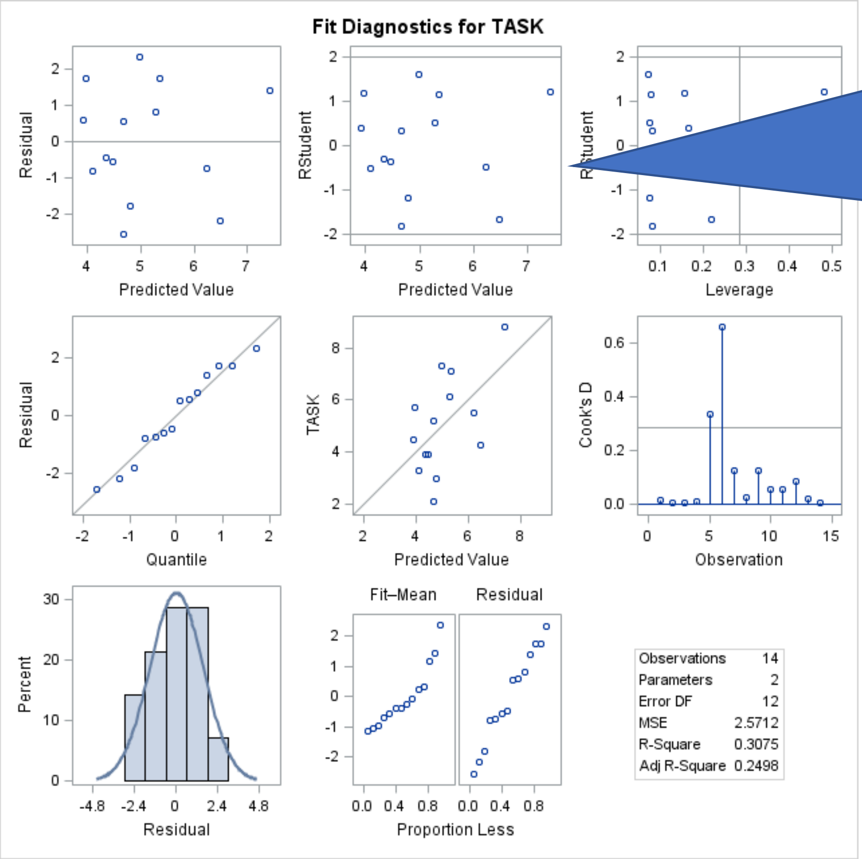
Diagnostic Plots for Linear Regression



Observations	14
Parameters	2
Error DF	12
MSE	2.5712
R-Square	0.3075
Adj R-Square	0.2498

Residual by Predicted Value plot (upper left), we want to see a random scatter of points above and below the 0 line, which is the case here. A nonrandom pattern of dots could indicate an inadequate model.

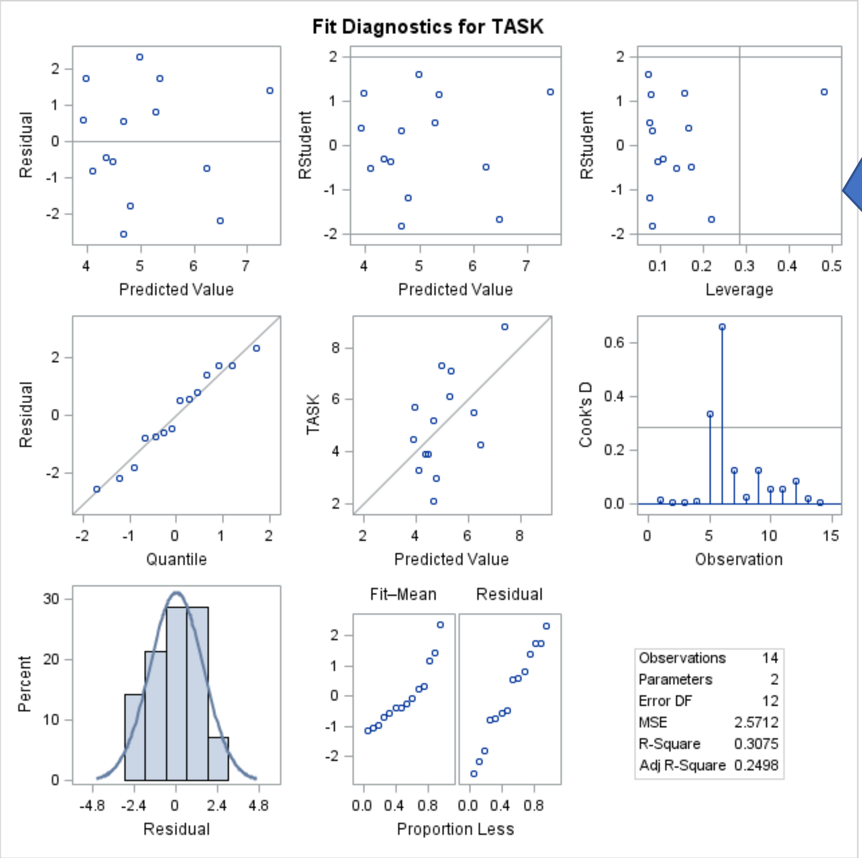
Diagnostic Plots for Linear Regression



Observations	14
Parameters	2
Error DF	12
MSE	2.5712
R-Square	0.3075
Adj R-Square	0.2498

The RStudent by Predicted Value plot indicates whether any Studentized residuals fall beyond two standard deviations, which would indicate unusual values. In this case, none fall outside the ±2 limits.

Diagnostic Plots for Linear Regression

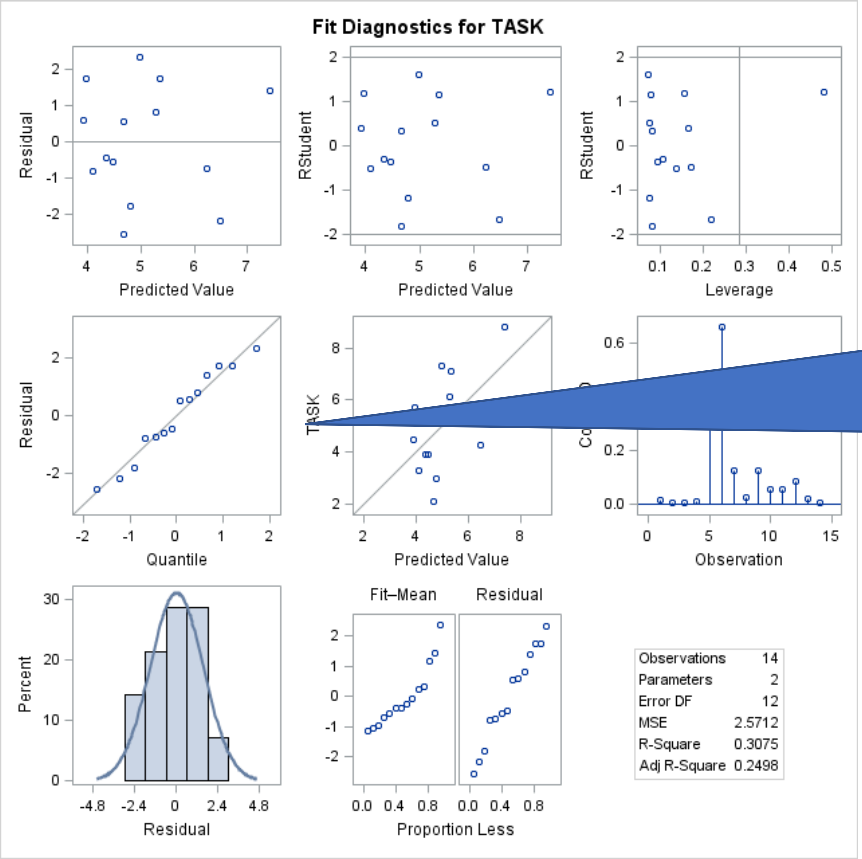


Observations	14
Parameters	2
Error DF	12
MSE	2.5712
R-Square	0.3075
Adj R-Square	0.2498

The RStudent by Leverage plot attempts to locate observations that might have unusual influence (leverage) on the calculation of the regression coefficients. In this case, there is possibly one observation that has undue influence. We'll identify *this* observation later.

Fox School of Business
TEMPLE UNIVERSITY®

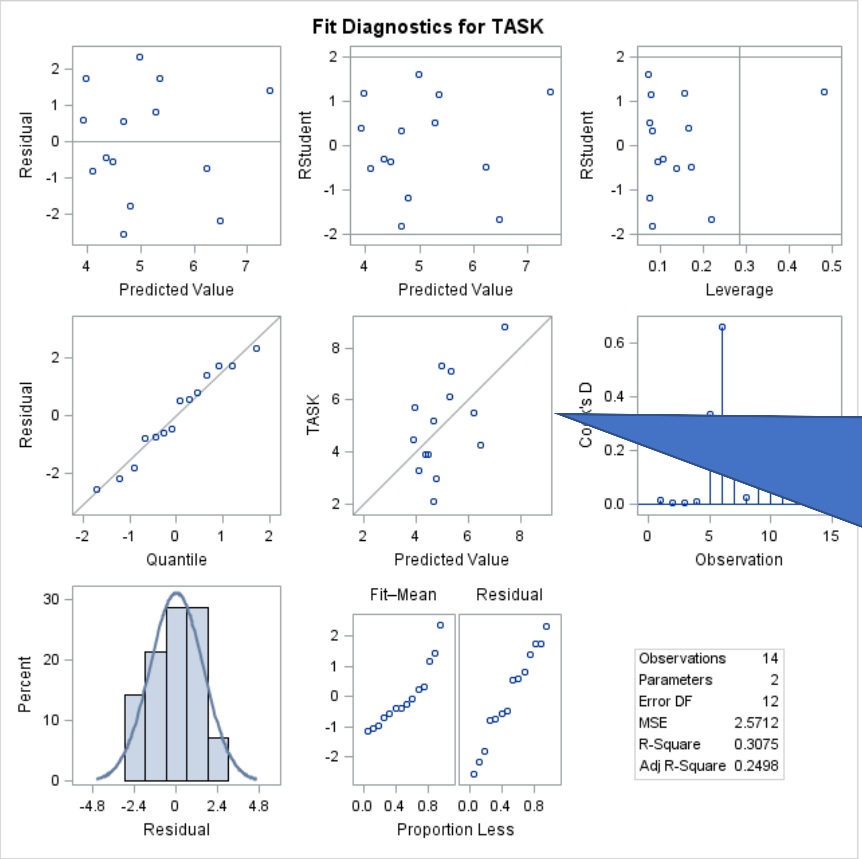
Diagnostic Plots for Linear Regression



Observations	14
Parameters	2
Error DF	12
MSE	2.5712
R-Square	0.3075
Adj R-Square	0.2498

In the Residual by Quartile plot, a tight and random scatter along the diagonal line indicates an adequate fit to the model.

Diagnostic Plots for Linear Regression

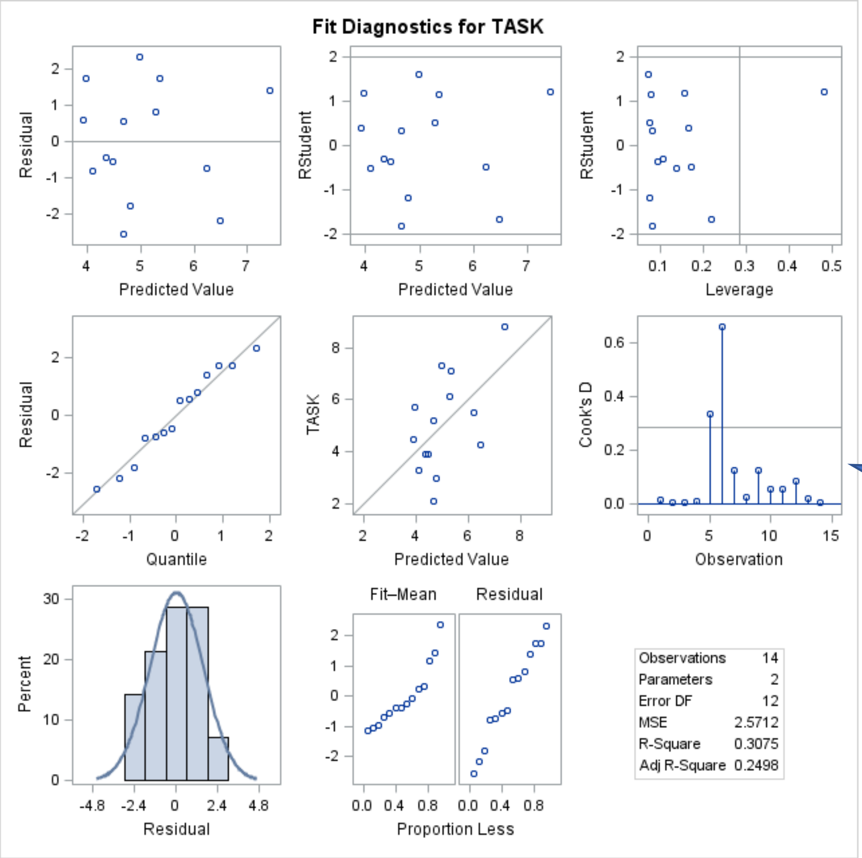


Observations	14
Parameters	2
Error DF	12
MSE	2.5712
R-Square	0.3075
Adj R-Square	0.2498

The Dependent Variable (TASK) by Predicted Value plot visualizes variability in the prediction, so if there is a pattern (e.g., variability increases as the predicted value increases) it indicates a nonconstant variance of the error.

Fox School of Business
TEMPLE UNIVERSITY®

Diagnostic Plots for Linear Regression

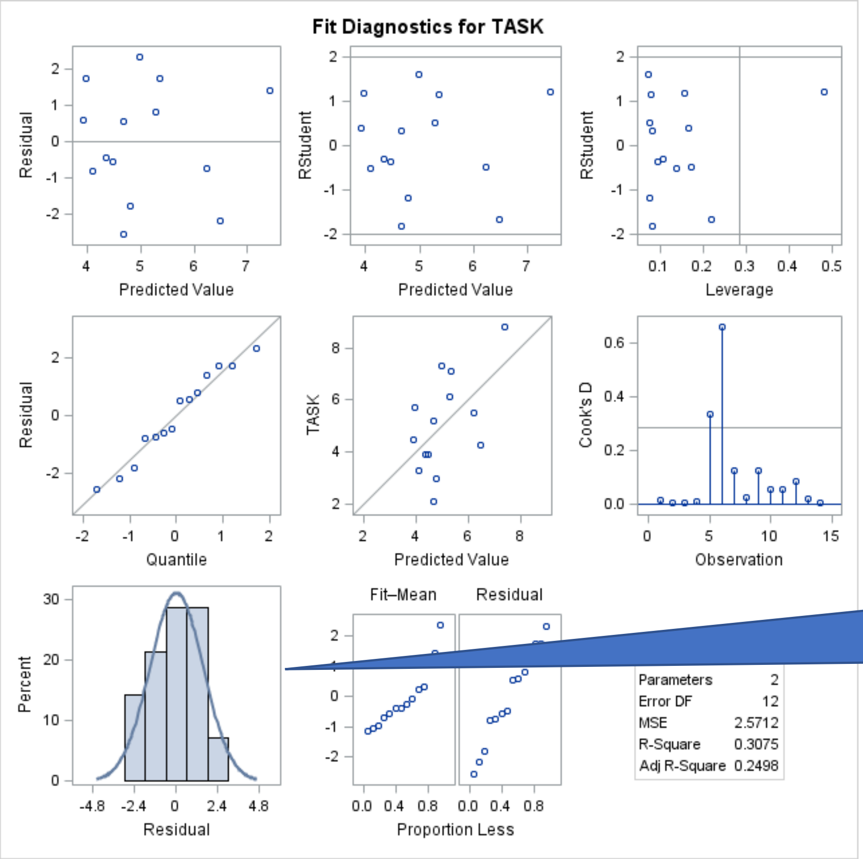


Fit Diagnostics for TASK

Observations	14
Parameters	2
Error DF	12
MSE	2.5712
R-Square	0.3075
Adj R-Square	0.2498

The Cook's *D* plot is designed to identify outliers or leverage points. In this case, it appears that observations 5 and 6 are suspect.

Diagnostic Plots for Linear Regression

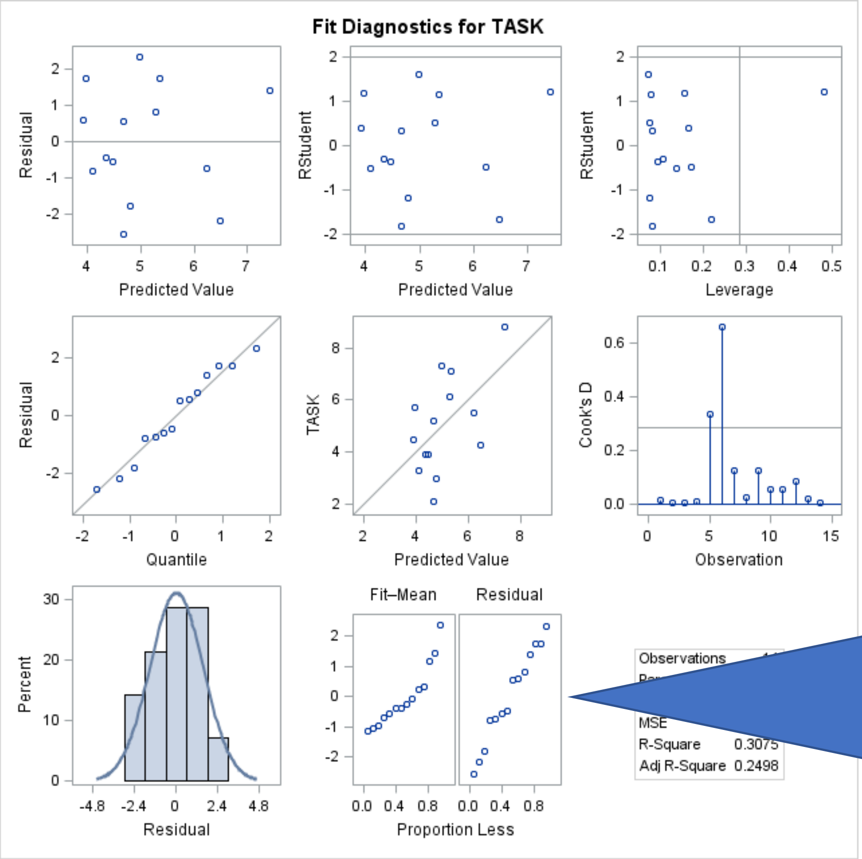


Parameters	2
Error DF	12
MSE	2.5712
R-Square	0.3075
Adj R-Square	0.2498

Residuals by Percent plot assesses the normality of the residuals.

Fox School of Business
TEMPLE UNIVERSITY®

Diagnostic Plots for Linear Regression



The Proportion Less (Spread plot) plots the proportion of the data by the rank for two or more categories. If the vertical spread (base on ranked data) is about the same, it means that there is about the same variance in both the fitted and residual values.

Fox School of Business
TEMPLE UNIVERSITY®

Predicting a New Value

- For this model, you might conclude that there is a moderate linear fit between CREATE and TASK, but it is not impressive ($R^2 = 0.3075$) or about 31% of the variation is accounted for by the regression using CREATE.
- Using the information in the regression equation, you could predict a value of TASK from CREATE=40.

```
4.67 = 2.16452 + 0.06235 * 40;
```

The value of TASK predicted.

The value of CREATE used for prediction

MULTIPLE LINEAR REGRESSION USING PROC REG

- Multiple Linear Regression (MLR) is an extension of simple linear regression. In MLR, there is a single dependent variable (Y) and more than one independent (X_i) variable.

- As with simple linear regression, the multiple regression equation calculated by SAS is a sample-based version of a theoretical equation describing the relationship between the *k* independent variables and the dependent variable *Y.*

$$Y = \boldsymbol{a} + \boldsymbol{\beta}_1 x_1 + \boldsymbol{\beta}_2 x_2 + \ldots + \boldsymbol{\beta}_k x_k + \boldsymbol{\varepsilon}$$

Hypotheses Tested

- As part of the analysis, the statistical significance of each of the coefficients is tested using a Student's *t*-test to determine if it contributes significant information to the predictor.

- These are tests of the hypotheses:

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$

- For these tests, if the p-value is low (say, <0.05), the conclusion is that the ith independent variable contributes significant information to the equation.

Using SAS PROC REG for Multiple Linear Regression

- As mentioned previously, the general syntax for PROC REG is

```
PROC REG <Options>; <Statements>;
```

Table. Additional Statement Options for the PROC REG MODEL statement (Options follow /) (Relevant to Multiple Linear Regression)

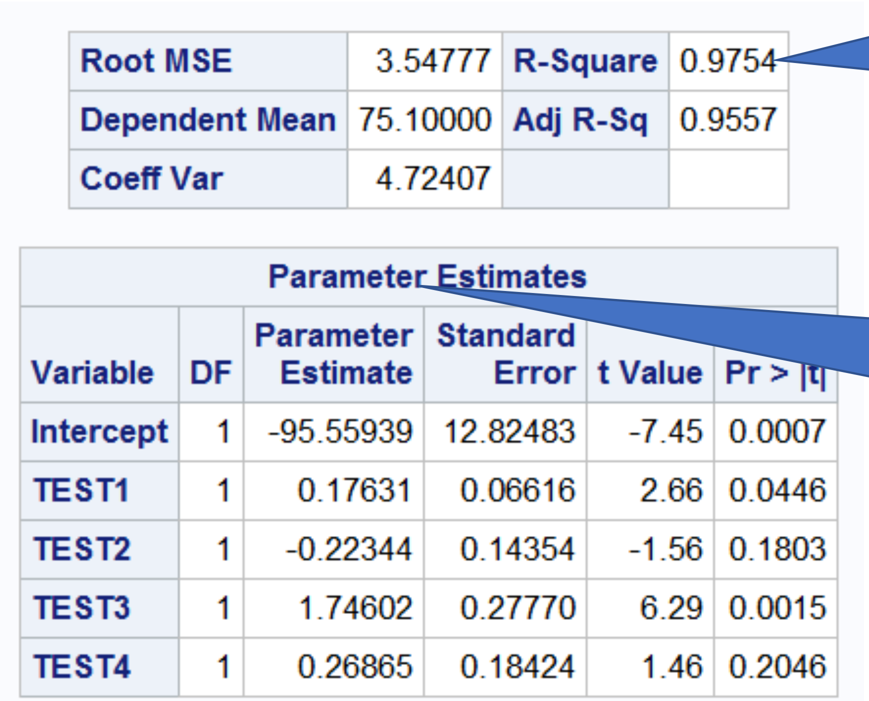
Option	**Explanation**
P	Requests a table containing predicted values from the model.
R	Requests that the residuals be analyzed.
CLM	Prints the 95 percent upper and lower confidence limits.
CLI	Requests the 95 percent upper and lower confidence limits for an individual value.
INCLUDE=k	Include the first k variables in the variable list in the model (for automated selection procedures).
SELECTION=option	Specifies automated variable selection procedure: BACKWARD, FORWARD, and STEPWISE, etc.
SLSTAY=p	Specifies the maximum p-value for a variable to stay in a model during automated model selection.
SLENTRY=p	Minimum p-value for a variable to enter a model for forward or stepwise selection.

EXAMPLE

In this model all of the predictors (independent variables) are specified

```
(enter data)
PROC REG;
MODEL JOBSCORE=TEST1 TEST2 TEST3 TEST4;
TITLE 'Job Score Analysis using PROC REG';
RUN;
QUIT;
```

Results



Root MSE	3.54777	R-Square	0.9754
Dependent Mean	75.10000	Adj R-Sq	0.9557
Coeff Var	4.72407		

Parameter Estimates

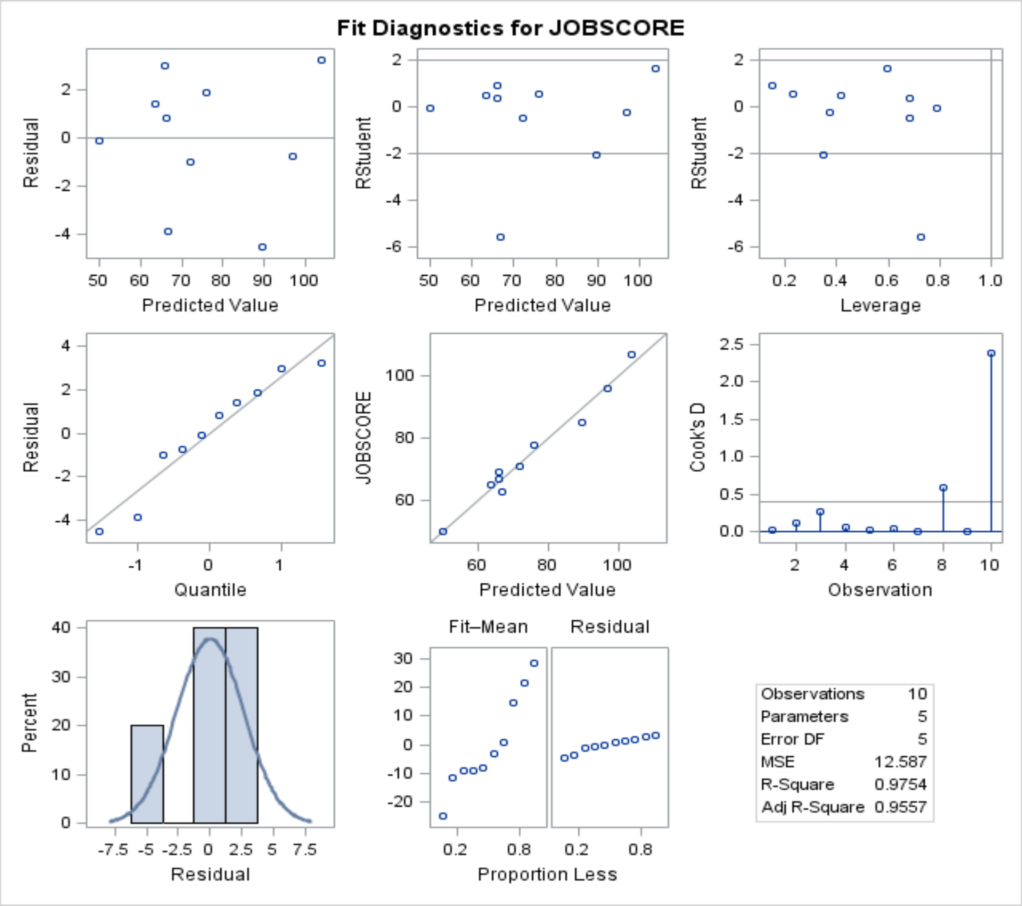
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-95.55939	12.82483	-7.45	0.0007
TEST1	1	0.17631	0.06616	2.66	0.0446
TEST2	1	-0.22344	0.14354	-1.56	0.1803
TEST3	1	1.74602	0.27770	6.29	0.0015
TEST4	1	0.26865	0.18424	1.46	0.2046

R-Square provides a measure of the strength of the prediction equation.

The Parameter Estimates are the estimates of the coefficients in the prediction equation.

Fox School of Business
TEMPLE UNIVERSITY®

Diagnostics for MLR Same as for SLR



Observations	10
Parameters	5
Error DF	5
MSE	12.587
R-Square	0.9754
Adj R-Square	0.9557

PERFORMING A ONE-SAMPLE T-TEST

- A one-sample t-test is often used to compare an observed mean with a known or "gold standard“ value.

- In general, for a one-sample t-test you obtain a random sample from some population and then compare the observed sample mean to some fixed value. The typical hypotheses for a one-sample t-test are as follows:

$H_0 : m = m_0$:The population mean is equal to a hypothesized value, m_0

$H_a : m \neq m_0$:The population mean is not equal to m_0

t-test Assumptions

- The key assumption underlying the one-sample *t-test* is that the population from which the random sample is selected is normal.

- If the data are non-normal, then nonparametric tests such as the sign test and the signed rank test are available.

- However, because of the central limit theorem, whenever the sample size is sufficiently large, the distribution of the sample mean is often approximately normal even when the population is non-normal. (See next slide)

Sample Size Recommendations

- The following are general guidelines:

- Small sample size (N < 15): You should not use the one-sample t-test if the data are clearly skewed or if outliers are present
- Moderate sample size (N> 15): The one-sample t-test can be safely used except when there are severe outliers.
- Large sample size *(N*> 40): The one-sample t-test can be safely used without regard to skewness or outliers.

Running the One-Sample t-Test in SAS

- Here are two ways to perform a one-sample t-test: First, using PROC UNIVARIATE, specify the value of m_0 for the test reported in the "Tests for Location.” For example,

```
PROC UNIVARIATE MU0=4 ;VAR LENGTH ;RUN;
```

- does a *t-test* of the null hypothesis that $m_0 = 4$.
- A second method in SAS for performing this one-sample *t-test* is to use the PROC TTEST procedure. For this procedure, the corresponding SAS code is

```
PROC TTEST H0=4;VAR LENGTH; RUN;
```

EXAMPLE

```
DATA ONESAMPLE;
INPUT LENGTH @@;
DATALINES;
4 3.95 4.01 3.95 4.00
3.98 3.97 3.97 4.01 3.98
3.99 4.01 4.02 4.02 3.98
4.01 3.99 4.03 4.00 3.99
;
Title 'Single sample t-test, using PROC UNIVARIATE';
PROC UNIVARIATE DATA=ONESAMPLE MU0=4;VAR LENGTH; RUN;
Title 'Single sample t-test using PROC TTEST';
PROC TTEST DATA=ONESAMPLE H0=4;var LENGTH;
RUN;
```

Two ways to perform the t-test – PROC UNIVARIATE and PROC TTEST.

Two Methods for a One Sample *t*-test

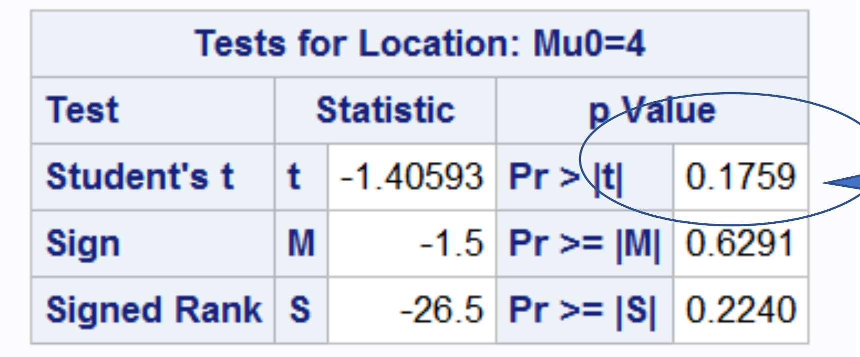
```
Title 'Using PROC UNIVARIATE';
PROC UNIVARIATE DATA=ONESAMPLE MU0=4;
    VAR LENGTH;
RUN;
```

Note indication of hypothesized value

```
Title 'Using PROC TTEST';
PROC TTEST DATA=ONESAMPLE H0=4 ;
   VAR LENGTH;
RUN;
```

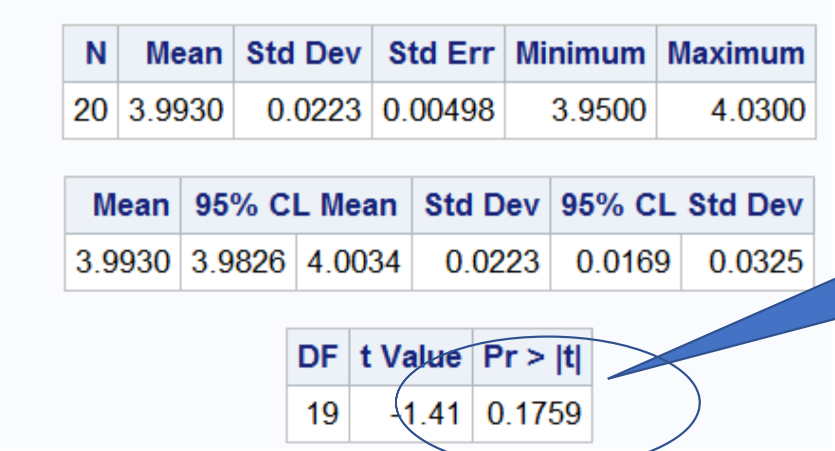
Note indication of hypothesized value

Results for a One-Sample *t*-test – Two Ways



Tests for Location: Mu0=4				
Test	**Statistic**		**p Value**	
Student's t	**t**	-1.40593	**Pr > |t|**	0.1759
Sign	**M**	-1.5	**Pr >= |M|**	0.6291
Signed Rank	**S**	-26.5	**Pr >= |S|**	0.2240

PROC UNIVARIATE RESULTS – See the line for Student’s t where p=0.1759



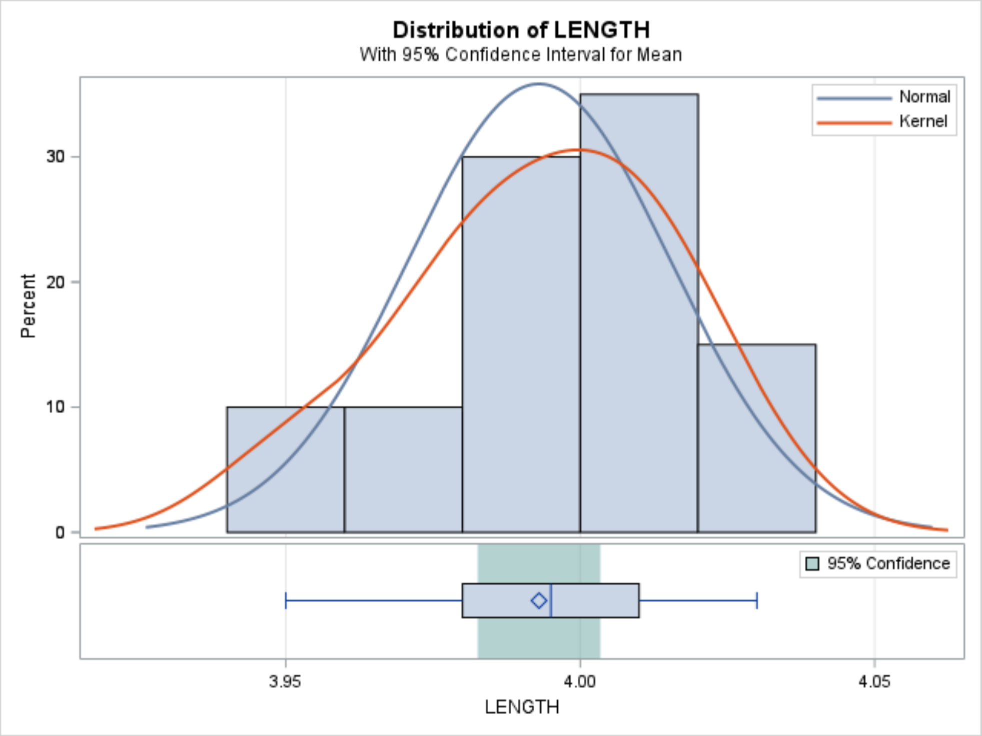
N	Mean	Std Dev	Std Err	Minimum	Maximum
20	3.9930	0.0223	0.00498	3.9500	4.0300

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
3.9930	3.9826	4.0034	0.0223	0.0169	0.0325

DF	t Value	Pr > |t|
19	-1.41	0.1759

PROC TTEST RESULTS – See where p=0.1759

Plot Created by PROC TTEST

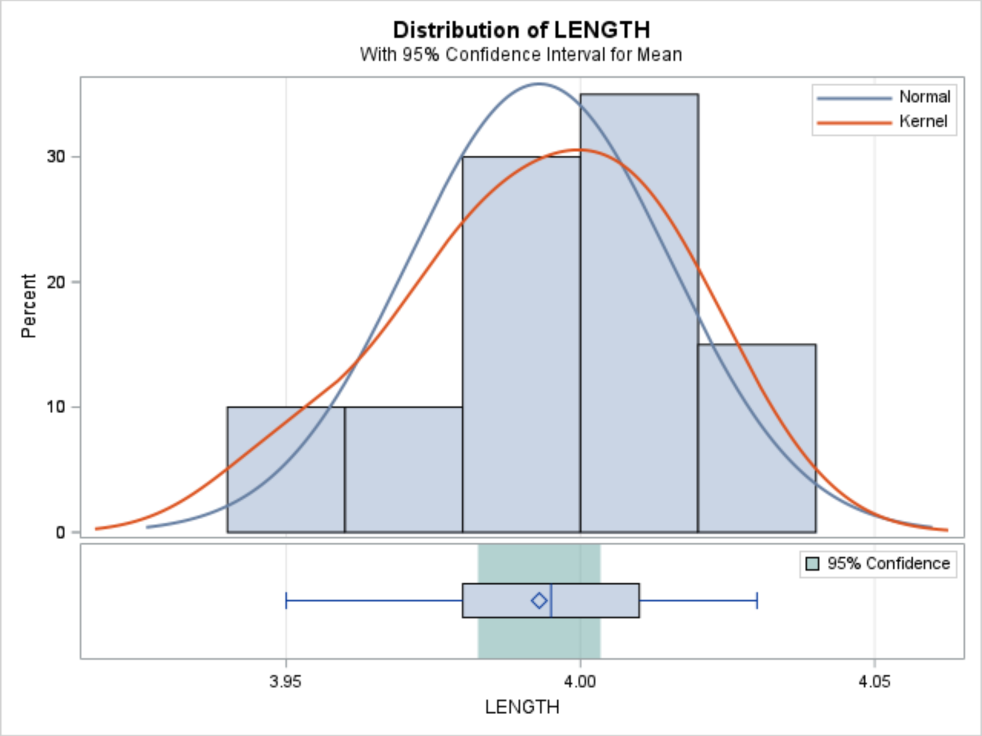


- The blue curve is a normal curve based on the mean and sd estimated from the data.

- The red curve is a kernel density estimator, which is a smoothed version of the histogram.

- If dramatic skewness were evident in the data, then the skewness would also be displayed in the kernel density estimator.

Plot Created by PROC TTEST



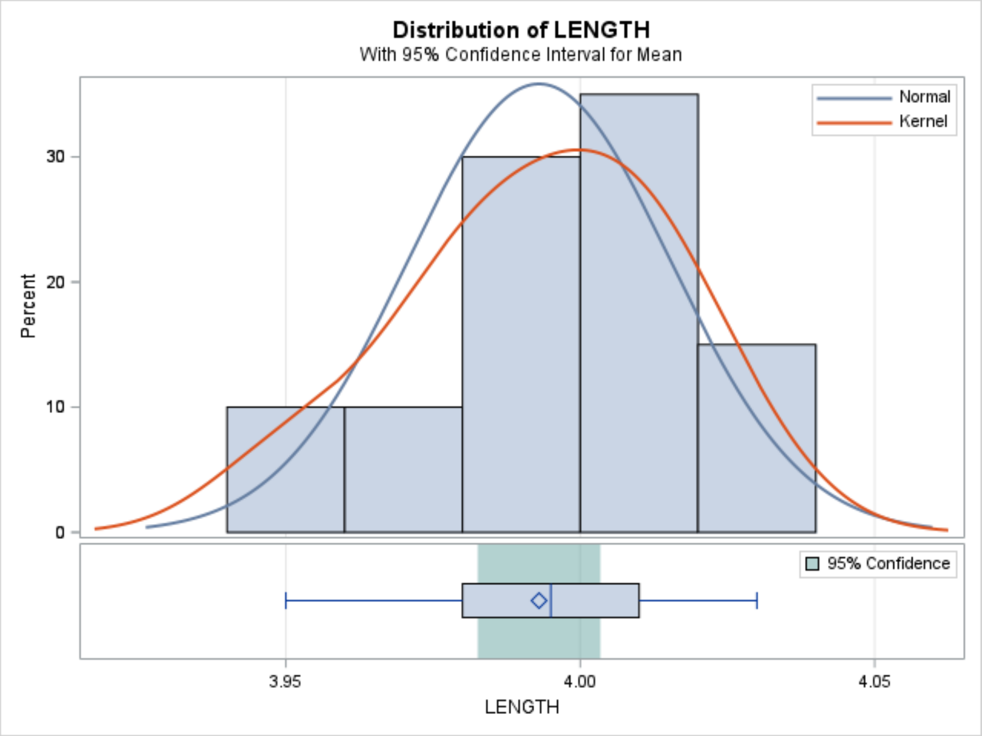
- These plots provide information about the normality assumption.

- There does not appear to be a dramatic departure from normality because the kernel density estimator is fairly bell-shaped.

- At the bottom of the plot is a boxplot.

- The boxplot is fairly symmetrical in shape with some tendency for the left whisker to be longer than the right one.

Plot Created by PROC TTEST



- Special graphics output such as the plots that follow are sometimes produced by SAS procedures

- Whenever you are in doubt about what type of graphics are available or what graphics may have been implemented in a new veSsion of SAS, it is a good idea to run your code with `ODS GRAPHICS ON/OFF`.

-
```
ODS GRAPHICS ON;
```
  before the procedure(s) and
```
ODS GRAPHICS OFF;
```

One-tailed tests

- If you are in rejecting the null hypothesis if the population mean differs from the hypothesized value in a particular direction of interest, you may want to use a one-tailed test. For example, the hypotheses if there is sufficient evidence that the mean is smaller than the hypothesized value, the hypotheses become as follows:

H_0:μ = μ0: The population mean is equal to a hypothesized value, μ_0.
H_a:μ < μ0: The population mean is less than μ

- In order to report the results of a one- tailed test you need to modify the reported p-value to fit a one-tailed test by dividing it by 2.

PERFORMING A TWO-SAMPLE *T-TEST*

- The SAS PROC TTEST procedure is used to test for the equality of means for a two-sample (independent group) t-test.
- The purpose of the two-sample t-test is to determine whether your data provide you with enough evidence to conclude that there is a difference in means.
- For a two-sample t-test you obtain independent random samples of size *N1* and *N2* from the two populations and compare the observed sample means.
- The typical hypotheses for a two-sample t-test are as follows:

$\boldsymbol{H_0{:}\mu_1 = \mu_2}$: The population means of the two groups are equal.
$\boldsymbol{H_a{:}\mu_1 \neq \mu_2}$: The population means are not equal.

Key Assumptions for a Two-Sample t-test

- Key assumptions underlying the two-sample t-test are

1. random samples are independent

2. populations are normally distributed with equal variances.

***If the data are non-normal, then nonparametric tests such as the Mann-Whitney U are available.

Guidelines regarding normality and equal variance assumptions

- **Normality:** As in the one-sample case, rules of thumb are available to help you determine whether to go ahead and trust the results of a two-sample t-test even when your data are non-normal. The sample size guidelines given earlier in this chapter for the one-sample test can be used in the two-sample case by replacing *N* with *N1* + *N2.*

- **Equal variances:** There are two t-tests reported by SAS in this setting: one based on the assumption that the variances of the two groups are equal (and thus using a pooled estimate of the common variance) and one (Satterthwaite) not making that assumption.

Running the Two-Sample *t*-Test in SAS

- Two sample t-tests can be obtained using the PROC TTEST procedure which was previously introduced in the context of a one-sample test.
- The syntax for the TTEST procedure is as follows:

```
PROC TTEST <options>;
  CLASS variable; <Statements>;
RUN;
```

Common Options for PROC TTEST

Common Options for PROC TTEST	
Option	**Explanation**
`DATA = datasetname`	Specifies which dataset to use.
`COCHRAN`	Use Cochran and Cox probability approximation for unequal variances.
`H0=n`	Specifies the hypothesized value under H_0.

Common statements for PROC TTEST

Common statements for PROC TTEST	

`CLASS variables;`	For a two-sample t-test, specify the grouping variable for the analysis.
	`PROC TTEST;`
	`CLASS GROUP;   VAR SCORE;`

`VAR variables;`	Specify observed variables for test:
	`PROC TTEST;`
	`CLASS GROUP; VAR SCORE WEIGHT HEIGHT;`

`PAIRED x*y;`	Specifies that a paired t-test is to be performed and which variables to use.
	`PROC TTEST;`
	`PAIRED BEFORE*AFTER;`

`BY, FORMAT, LABEL, WHERE`	These statements are common to most procedures, and may be used here.

EXAMPLE

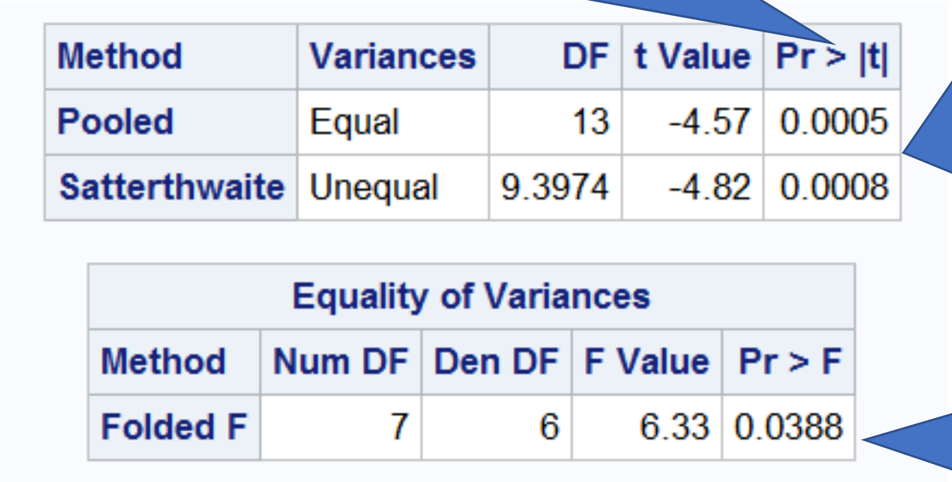
```
DATA TTEST;
INPUT BRAND $ HEIGHT;
DATALINES;
   A    20.00
   A    23.00
   A    32.00
etc
;
PROC TTEST;
     CLASS BRAND;
     VAR HEIGHT;
     Title 'Independent Group t-Test Example';
RUN;
QUIT;
```

Note that the CLASS statement identifies the “grouping” variable that specifies which groups are to be compared. In this case BRAND is A or B, the two groups to be compared.

The VAR statement identifies the observed variable.

Understanding the output

THIRD– using either t-test, the p-value is small, so reject the null hypothesis and conclude that there is a difference in means.



Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	13	-4.57	0.0005
Satterthwaite	Unequal	9.3974	-4.82	0.0008

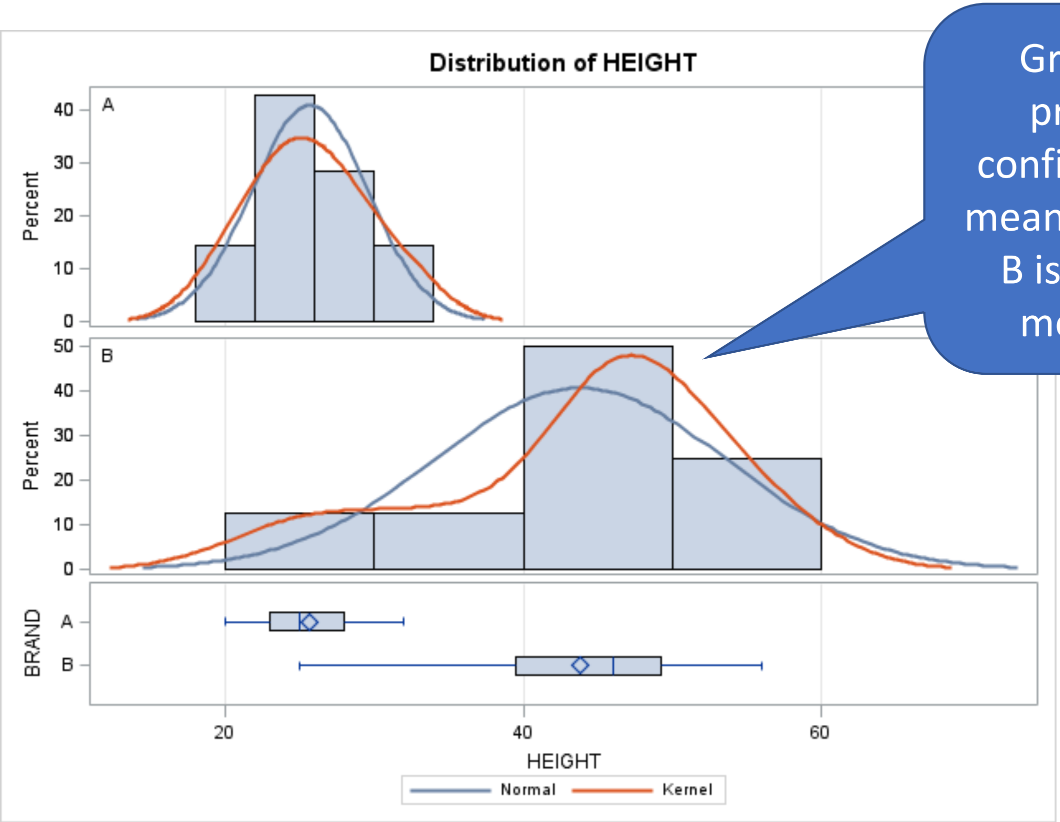
Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	7	6	6.33	0.0388

SECOND - If variances are equal, do the standard (Pooled) t-test. If variances unequal, do the unequal (Satterthwaite) version of the t-test

FIRST… use this test to determine if variances are "equal." A low p value p<.05 indicates variances are unequal.

START HERE

Graphical Results from Two Sample *t*-test



Graphical results provide a visual confirmation that the mean HEIGHT of group B is larger than the mean of group A

PERFORMING A PAIRED *T*-TEST

- To perform a paired t-test to compare two repeated measures (such as in a before–after situation) where both observations are taken from the same or matched subjects, use PROC TTEST with the PAIRED statement.

- Suppose your data contain the variables WBEFORE and WAFTER (before and after weight on a diet) for eight subjects. The hypotheses for this test are:

$\boldsymbol{H_0{:}\mu_{\text{Loss}} = 0}$**:** The population average weight loss is zero.

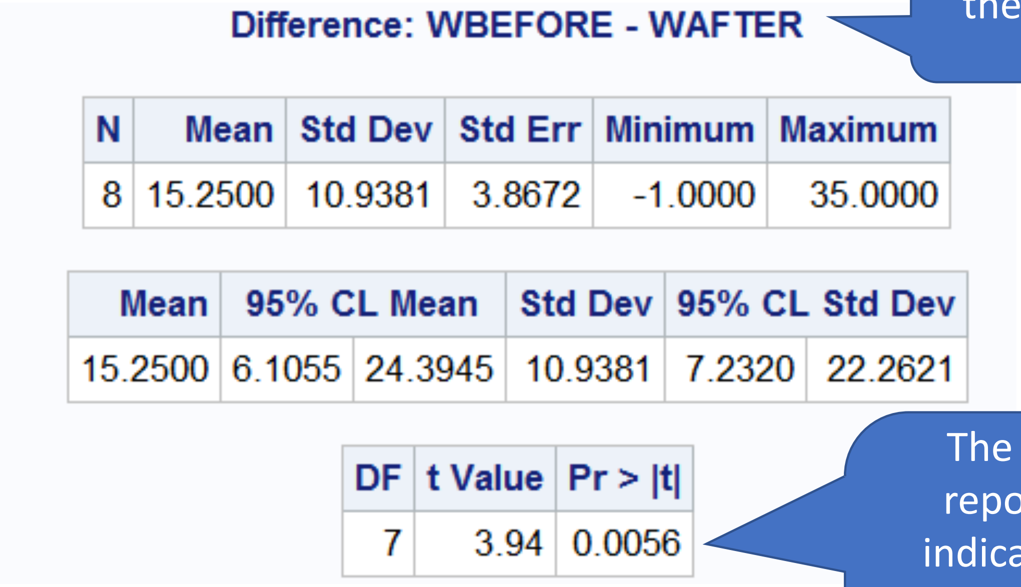
$\boldsymbol{H_a{:}\ \mu_{\text{Loss}} \neq 0}$**:** The population average weight loss is not zero.

EXAMPLE

```
PROC TTEST;
PAIRED WBEFORE*WAFTER;
TITLE 'Paired t-test Example';
RUN;
```

Note the PAIRED statement – which indicates the two paired variables WBEFORE and WAFTER.

Understanding the Paired *t*-test output



Difference: WBEFORE - WAFTER

N	Mean	Std Dev	Std Err	Minimum	Maximum
8	15.2500	10.9381	3.8672	-1.0000	35.0000

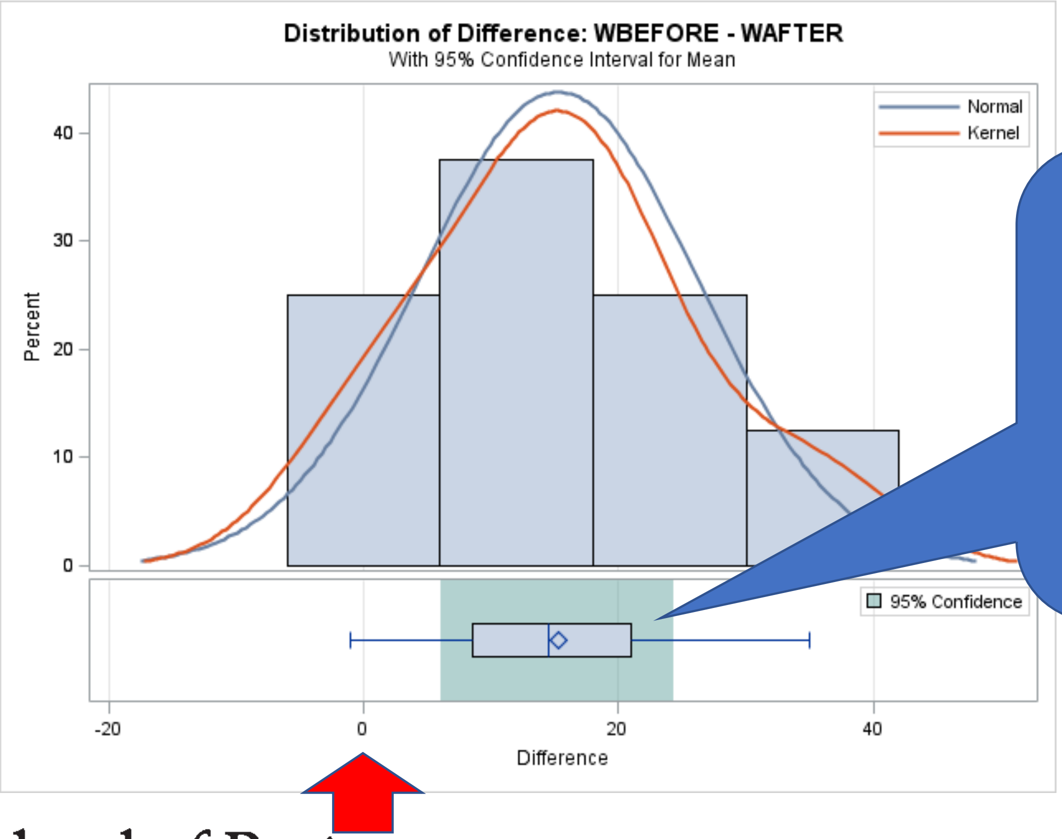
Mean	95% CL Mean		Std Dev	95% CL Std Dev	
15.2500	6.1055	24.3945	10.9381	7.2320	22.2621

DF	t Value	Pr > |t|
7	3.94	0.0056

Note the test is performed on the difference of the paired variables.

The results of the t-test reports p=0.0056 which indicates to reject the null hypothesis of no difference and conclude that there was statistically significant weight loss.

Graphical Results for Paired *t*-test



Notice that the 95% confidence interval (indicated by the green highlight) does not include 0, which is another way to conclude that the mean difference is greater than 0.

ANOVA

- To be able to compare three or more means , a one-way ANOVA with multiple comparisons can be used.

- **PROC ANOVA:** a basic procedure useful for one-way ANOVA or for multiway factorial designs with fixed factors and an equal number of observations per cell.

- **PROC GLM:** for one-way repeated measures analysis, and techniques not supported by PROC ANOVA.

COMPARING THREE OR MORE MEANS USING ONE-WAY ANALYSIS OF VARIANCE

- A one-way ANOVA is an extension of the independent group t-test where there are more than two groups.

- Assumptions for this test are similar to those for the t-test:
  - Data within groups are normally distributed with equal variances across groups.
  - Groups are from independent samples.

- The hypotheses for the comparison of independent groups are as follows (k is the number of groups):

$\boldsymbol{H_0}$**:** $\boldsymbol{\mu_1 = \mu_2 = \ldots = \mu_k}$**:** Means of all the groups are equal.

$\boldsymbol{H_a}$**:** $\boldsymbol{\mu_i \neq \mu_j}$ **for some** $\boldsymbol{i \neq j}$**:** At least two means are not equal.

Simplified Syntax for PROC ANOVA

- The syntax for the statement is as follows:

```
PROC ANOVA <Options>;
   CLASS variable;
   MODEL dependentvar = independentvars;
   MEANS independentvars / typecomparison
         <meansoptions>;
```

CLASS defines grouping variable.

The MODEL statement defines the model tested.

The MEANS statement defines post hoc multiple comparisons.

Table. Common Options for PROC ANOVA and PROC GLM for preforming a One-Way ANOVA or simple Repeated Measures

Option	**Explanation**
`DATA = dataname`	Specifies which data set to use.
`NOPRINT`	Suppresses output. This is used when you want to extract information from ANOVA results but don’t want SAS to produce output in the Results Viewer.
`OUTSTAT=dataname`	Names an output data set that saves a number of the results from the ANOVA calculation.
`PLOTS=options`	Specify PLOTS=NONE to suppress plots that are generated by default.
`ORDER=option`	Specifies order in which to display the CLASS variable (similar to what was covered in Chapter 10:Analyzing Counts and Tables.) Options are DATA, FORMATTED, FREQ, or INTERNAL.
`ALPHA=p`	Specifies alpha level for a Confidence Interval (GLM only)

(Table Continued)	
CLASS variable list;	This statement is required and specifies the grouping variable(s) for the analysis.
MODEL specification	Specifies the dependent and independent variables for the analysis. More specifically, it takes the form MODEL dependentvariable=independentvariable(s);
FREQ var	Specifies that a variable represents the count of values for an observation. Similar to the WEIGHT statement for PROC FREQ.
MEANS vars	Calculates means for dependent variables and may include comparisons.
LSMEANS vars	Calculates least square means for a dependent variable & to request comparisons. (GLM Only)
REPEATED vars	Used to specify repeated measure variables.
TEST specification	Used to specify a hypothesis test value.
CONTRAST specification	Allows you to create customized posthoc comparisons. (GLM Only)
BY, FORMAT, LABEL, WHERE	These statements are common to most procedures, and may be used here.

Using the MEANS or LSMEANS Statement

- When you perform a one-way ANOVA, typically there is a two-step procedure:

  (1) test the null hypothesis to determine whether any significant differences exist, and

  (2) if H_0 is rejected, run subsequent multiple comparison tests to determine which differences are significantly different.

- Pairwise comparison of means can be performed using one of several multiple comparison tests specified using the MEANS statement, which has the following format (where independantvar is a CLASS variable):

```
MEANS independentvar/typecomparison <meansoptions>;
```

- For PROC GLM, use the LSMEANS statement:

```
LSMEANS independentvar / typecomparison <meansoptions>;
```

Table. Common type comparison options for the PROC ANOVA or GLM MEANS Statement (Options following the slash /)

Option	Explanation
`BON`	Bonferroni t-tests of difference
`DUNCAN`	Duncan’s multiple range test
`SCHEFFE`	Scheffe multiple comparison
`SNK`	Student Newman Keuls multiple range test
`LSD`	Fisher’s Least Significant Difference
`TUKEY`	Tukey’s studentized range test
`DUNNETT (‘x’)`	Dunnett’s test—compare to a single control, where 'x' is the category value of the control group
`ALPHA=pvalue`	Specifies the significance level for comparisons (default: 0.05)
`CLDIFF`	Requests that confidence limits be included in the output.

(Table continued)	
Common type comparison options for the **PROC GLM** LSMEANS Statement (options following the slash /	
`ADJUST=option`	Specify type of multiple comparison. Examples are BON, DUNCAN, SCHFEE, SNK, LSD, DUNNETT
`PDIFF=`	Calculates p-values base (default is T). You can also specify TUKEY or DUNNETT options.

EXAMPLE

```
PROC ANOVA DATA=ACHE;
    CLASS BRAND;
    MODEL RELIEF=BRAND;
    MEANS BRAND/TUKEY;
TITLE 'COMPARE RELIEF ACROSS MEDICINES  - ANOVA
EXAMPLE';
RUN;
QUIT;
```

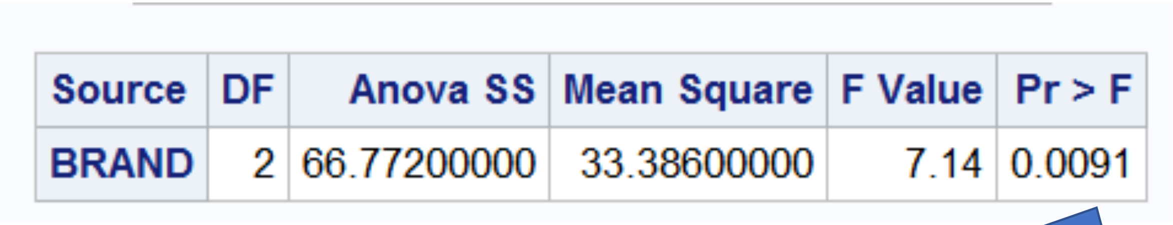
CLASS defines the grouping variable, BRAND.

The MODEL statement indicates you are wanting to test if BRAND can predict mean RELIEF.

The MEANS statement is used for a post hoc test (if Ho is rejected) to determine which means are different.

Results of a One-Way ANOVA

- The primary results for a One-Way ANOVA test are in the following table:

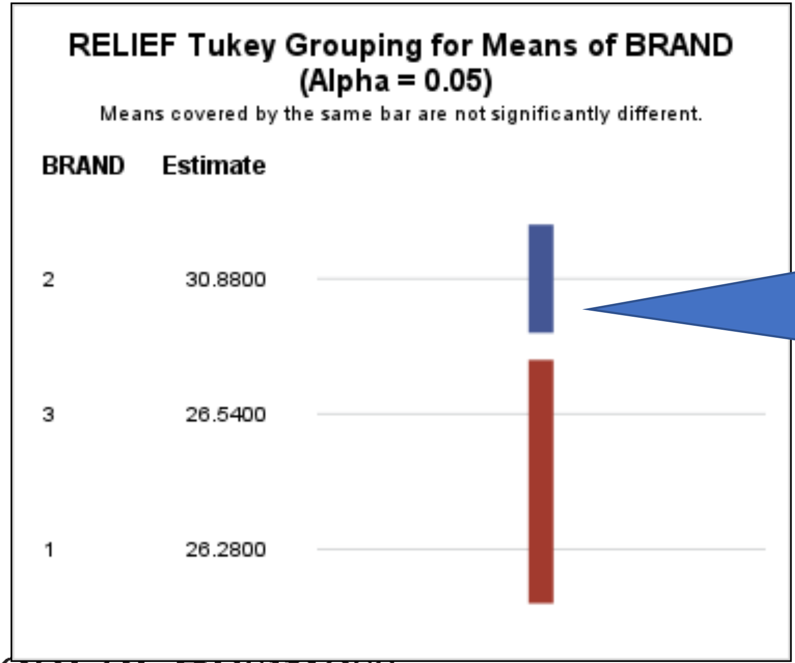


Source	DF	Anova SS	Mean Square	F Value	Pr > F
BRAND	2	66.77200000	33.38600000	7.14	0.0091

The p-value is used to decide whether or not to reject the null hypothesis. Typically, if p<0.05, you reject Ho. If you reject Ho, it indicates that some means (by group) are different, so you proceed to look at the post hoc results.

Post Hoc Multiple Comparisons test – Tukey Test

- Depending on your SAS version, you may get this description of grouping using TUKEY comparisons – both come to the same conclusion.



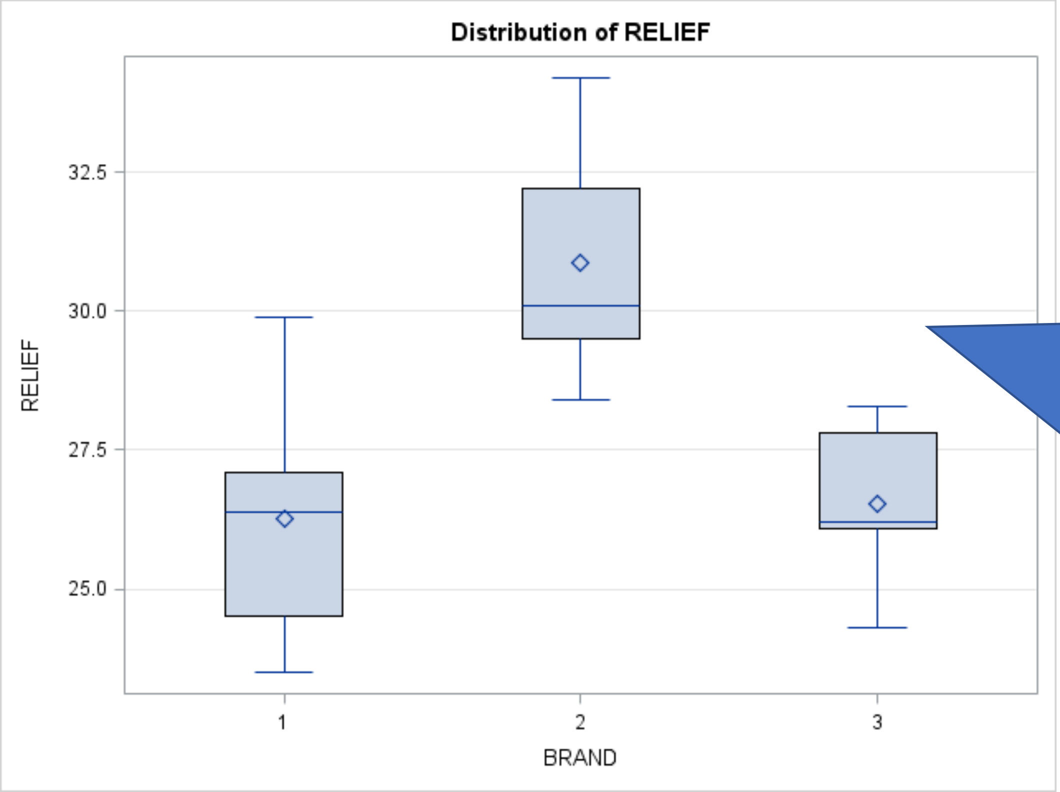
RELIEF Tukey Grouping for Means of BRAND (Alpha = 0.05)

Means covered by the same bar are not significantly different.

```
BRAND   Estimate

2       30.8800

3       26.5400

1       26.2800
```

In this graph, (at the 0.05 significance level) the mean for BRAND 2 (30.89(is different from the means of group 3 and 1 (26.54 and 26.28).

Graphical Comparison of Groups



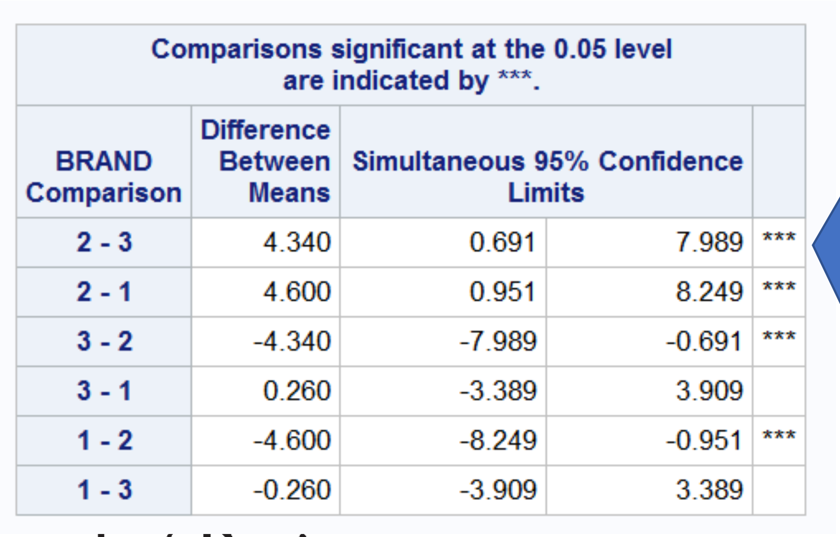
This graph reinforces the statistical results --- that groups 1 and 3 are very similar, but the mean for group 2 is larger than for either groups 1 or 2.

Multiple Comparison Test Using Confidence Limits

- Using this code for the comparison test:

```
MEANS BRAND/TUKEY CLDIFF;
```

- Results in this table



Comparisons significant at the 0.05 level are indicated by ***.

BRAND Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
2 - 3	4.340	0.691	7.989	***
2 - 1	4.600	0.951	8.249	***
3 - 2	-4.340	-7.989	-0.691	***
3 - 1	0.260	-3.389	3.909	
1 - 2	-4.600	-8.249	-0.951	***
1 - 3	-0.260	-3.909	3.389	

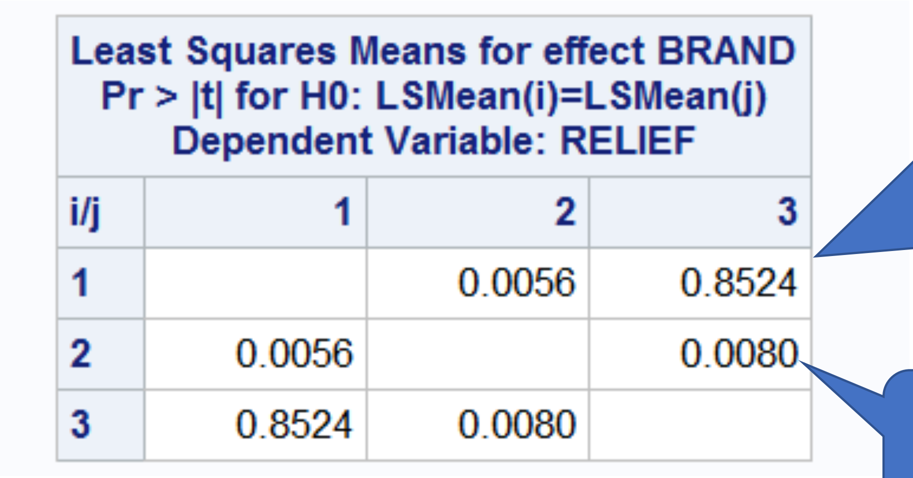
In this table, mean differences are compared. For example, the first line tests the difference between means for groups 2 minus 3 = 4.340 and reports a 95% CL of 0.691 to 7,989. Since this range **does not include** 0.0, the difference is considered statistical different at the 0.05 significance level. The *** indicates a 0.05 significant difference for that comparison

Multiple Comparisons using p-values

- Using **PROC GLM** instead of PROC ANOVA, and using this code for the comparison test:

```
LSMEANS BRAND/ PDIFF;
```

- Results in this table:



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

i/j	1	2	3
1		0.0056	0.8524
2	0.0056		0.0080
3	0.8524	0.0080	

This table reports the results of mean comparisons. For example, the comparison of mean 1 vs 3 reports a p-value of 0.8524, indicating that the difference in means is NOT statistically different.

The comparison of means 2 vs 3 is statistically different at p=0.0080.

COMPARING THREE OR MORE REPEATED MEASURES

- Repeated measures are observations taken from the same or related subjects over time or in differing circumstances.

- When there are three or more repeated measures, the corresponding analysis is a repeated measures ANOVA.

- The hypotheses being tested with repeated measures ANOVA are as follows:

  ***H_0*:** There is no difference among the group means (repeated measures).

  ***H_a* :** There is a difference among the group means.

Fox School of Business
TEMPLE UNIVERSITY®

Example Syntax for a Repeated Measures ANOVA

```
PROC GLM DATA=STUDY;
      CLASS SUBJ DRUG;
      MODEL RESULT = SUBJ DRUG;
      MEANS DRUG/DUNCAN;
      TITLE 'Repeated Measures ANOVA';
RUN;
QUIT;
```

The CLASS statement indicates grouping variables. In repeated measures, a subject variable is included.

The MODEL statement indicates that you want to predict RESULT from type of DRUG. Subject is included to account for subject differences

Example Repeated Measures Data

Each Subject received each of the 4 drugs (in random order, with a washout period between administrations.)

Subj	Drug1	Drug2	Drug3	Drug4
1	31	29	17	35
2	15	17	11	23
3	25	21	19	31
4	35	35	21	45
5	27	27	15	31

EXAMPLE

- The data for the repeated measures in not like in the table. Each line represents an observation, and each subject has 4 lines representing the 4 drugs.

```
DATA STUDY;
INPUT SUBJ DRUG RESULT;
DATALINES;
1     1     31
1     2     29
1     3     17
1     4     35
2     1     15
Etc…
```

Notice how data is set up for repeated measures – each subject has 4 records – one for each drug observation.

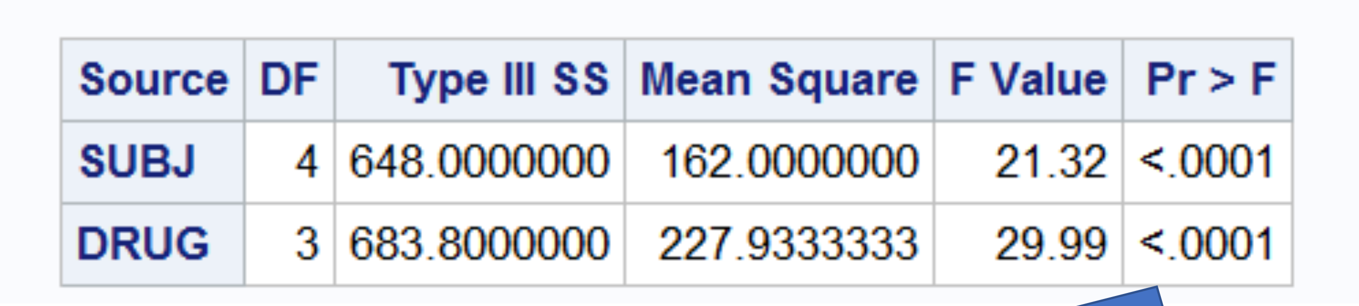
SAS Code for Repeated Measures

- This example illustrates how to compare three or more repeated measures (dependent samples) and perform pairwise comparisons using the DUNCAN procedure.

```
PROC GLM DATA=STUDY;
    CLASS SUBJ DRUG;
    MODEL RESULT= SUBJ DRUG;
    MEANS DRUG/DUNCAN;
    TITLE 'Repeated Measures ANOVA';
RUN;
```

Results from Repeated Measures ANOVA

- The results of interest are in the Type III table:

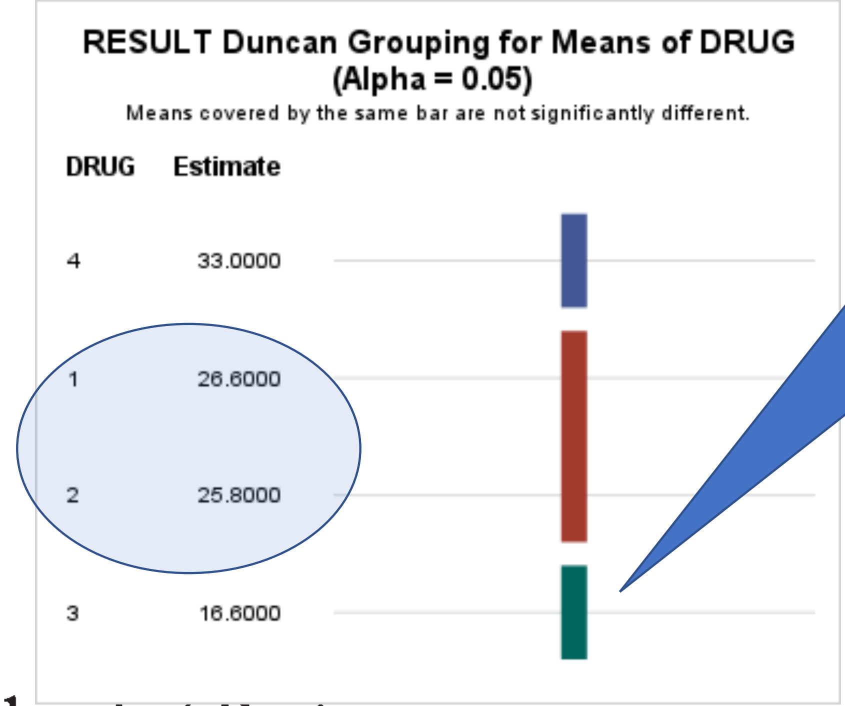


Source	DF	Type III SS	Mean Square	F Value	Pr > F
SUBJ	4	648.0000000	162.0000000	21.32	<.0001
DRUG	3	683.8000000	227.9333333	29.99	<.0001

Typically, you are not interested in the SUBJ line in this table (or p-value). The line of interest is the DRUG line, which tests the hypothesis of interest. In this case p<0.0001, which indicates a significant difference in means for the 4 Drugs. Do a post hoc test to determine which drugs are different.

Multiple Comparisons for Repeated Measures ANOVA

```
Results for MEANS DRUG/DUNCAN;
```



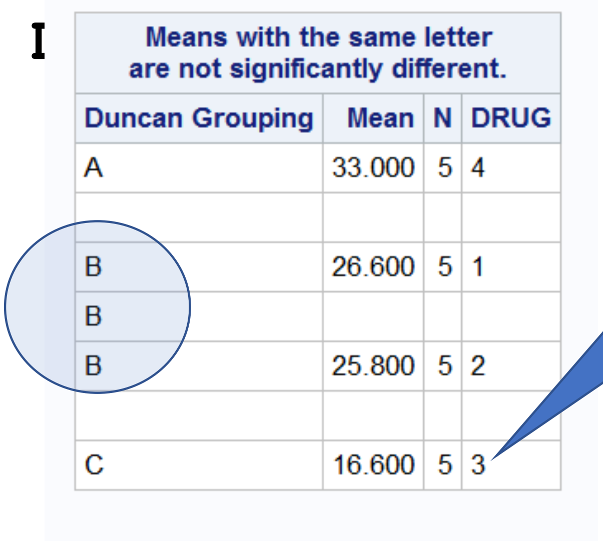
```
RESULT Duncan Grouping for Means of DRUG
             (Alpha = 0.05)
Means covered by the same bar are not significantly different.

DRUG    Estimate

4        33.0000

1        26.6000

2        25.8000

3        16.6000
```

Duncan test indicates time to relief for drug 3 (16.6) is significantly (at 0.05) lower than that for all other drugs. There's no statistical difference between drugs 2 and 1; drug 4 has the highest time to relief for all drugs tested. Thus, on this basis, drug 3 would be the preferred drug.

Some SAS Version may output this table:

- This statement provides a multiple comparison test, which is appropriate if the main hypothesis is s

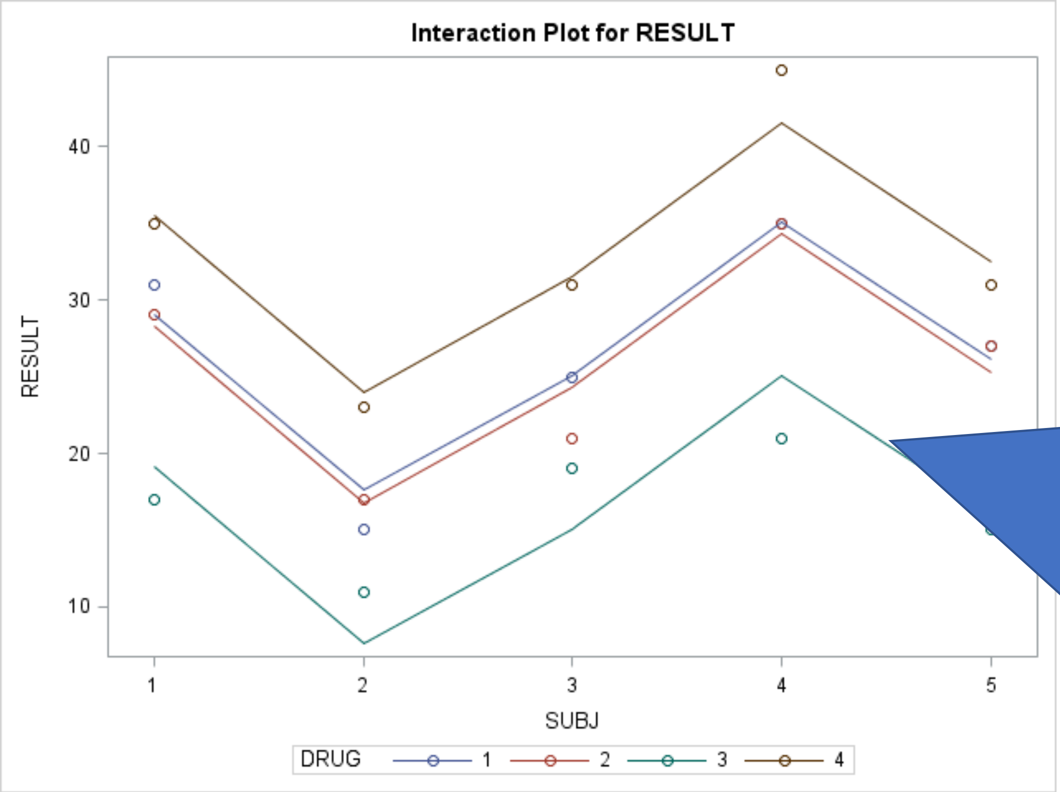
MEANS I



Means with the same letter are not significantly different.			
Duncan Grouping	Mean	N	DRUG
A	33.000	5	4
B	26.600	5	1
B			
B	25.800	5	2
C	16.600	5	3

Results indicate that there is NO DIFFERNCE in DRUGS 1 and 2 (Means of 26.6 vs 25.8). However, DRUG4 has the largest (statistically significant) mean at 33.0 and DRUG 3 has the smallest at 16.60. (Same as previous graph)

Graphical Results of a Repeated Measures ANOVA



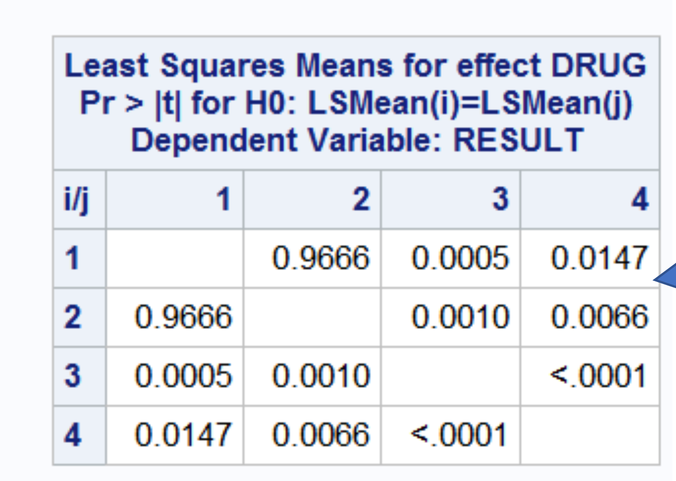
This is visual confirmation of the multiple comparisons – the line for DRUG4 is consistently higher than all the others. DRUGS 1 and 2 are too close to call different , and DRIG 3 has the smallest means.

Using LSMEANS for Comparisons (Tukey)

- Using this code:

```
LSMEANS DRUG/PDIFF ADJUST=TUKEY;
```

- You get the fol



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

i/j	1	2	3	4
1		0.9666	0.0005	0.0147
2	0.9666		0.0010	0.0066
3	0.0005	0.0010		<.0001
4	0.0147	0.0066	<.0001	

Results indicate that there is NO DIFFERNCE in DRUGS 1 and 2 (p=.97). However, the mean for DRUG4 is different than for DRUG1 (p=0.0147) and so on…

- Other common ADJUST= options are BON, DUNNETT, and SCHEFFE.

MODEL TYPES

- Typical Model Statement

```
MODEL dependentvar = independentver(s)
```

Type	Dependent	Independent
ANOVA	Group	Quantitative
Linear	Continuous	Quantitative and /or grouping
Logistic	Binary	Quantitative and /or grouping
Chi-Square	Group	Group

Binary Logistic Regression

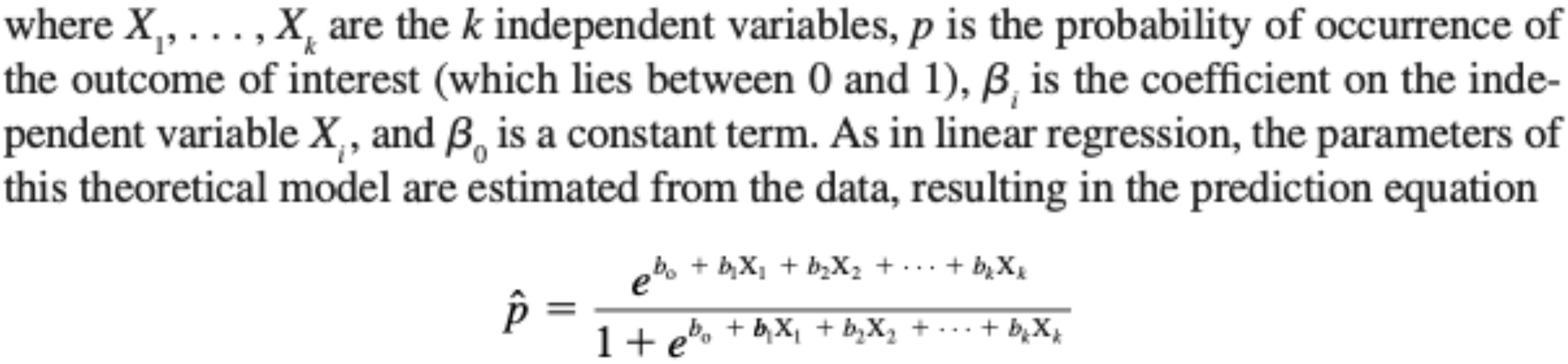
- Binary logistic regression models are based on a dependent variable that can take on only one of two values, such as presence or absence of a disease, deceased or not deceased, married or unmarried, and so on.

- In this setting, the independent (sometimes called explanatory or predictor) variables **are used for predicting the probability of occurrence of an outcome** (such as mortality).

LOGISTIC ANALYSIS BASICS: Logistic Regression Model

- The basic form of the logistic equation is

$$p = \frac{e^{\beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k}}{1 + e^{\beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k}}$$



where $X_1, \ldots, X_k$ are the k independent variables, p is the probability of occurrence of the outcome of interest (which lies between 0 and 1), β_i is the coefficient on the independent variable X_i, and β_0 is a constant term. As in linear regression, the parameters of this theoretical model are estimated from the data, resulting in the prediction equation

$$\hat{p} = \frac{e^{b_o + b_1\mathrm{X}_1 + b_2\mathrm{X}_2 + \cdots + b_k\mathrm{X}_k}}{1 + e^{b_o + b_1\mathrm{X}_1 + b_2\mathrm{X}_2 + \cdots + b_k\mathrm{X}_k}}$$

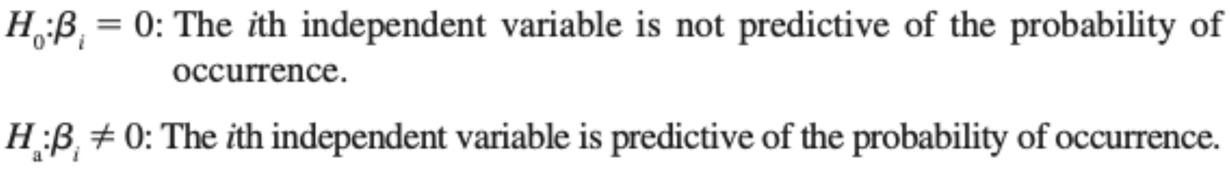
Hypotheses for the Logistic Model

- Any variable with a zero coefficient in the theoretical model is not useful in predicting the probability of occurrence.

- SAS reports tests of the **null hypothesis** that all of the β_i 's, *i* = 1, ... , *k* are zero.
- If this null hypothesis is **not rejected**, then there is **no statistical evidence that the independent variables as a group are useful in the prediction**.

- If the overall test is rejected, then we conclude that at least some of the variables are useful in the prediction. For each β_i = 1, ... , *k,* SAS reports the results of the tests.

- The hypotheses test are thus



$H_0{:}\beta_i = 0$: The *i*th independent variable is not predictive of the probability of occurrence.

$H_a{:}\beta_i \neq 0$: The *i*th independent variable is predictive of the probability of occurrence.

Understanding Odds and Odds Ratios

- Another use of the logistic model is the calculation of odds ratio (OR) for each independent variable.

- **The odds of an event measures the expected number of times an event will occur relative to the number of times it will not occur**. Thus, if the odds ratio of an event is 5, this indicates that we expect five times as many occurrences as non-occurrences. An odds of 0.2 (=1/5) would indicate that we expect five times as many non occurrences as occurrences.

PERFORMING A LOGISTIC ANALYSIS USING PROC LOGISTIC

- PROC LOGISTIC is the SAS procedure that allows you to analyze the data using a binary logistic model.
- An abbreviated syntax for this stateme

CLASS variables are categorical such as Gender “Male” and “Female” or Cancer Stage 1, 2, 3

```
PROC LOGISTIC <options>;
CLASS variables;
MODEL depend[illegible]var <(variable_options)> =
            [illegible]tions>;
```

The dependent variable is binary – that is it takes on only one of two values.

What You Are Predicting

- By default, SAS assumes that the outcome predicted (with p) in the logistic regression equation corresponds to the case in which the dependent variable is 0. (or the lowest number or alphabetic character.)

- If, for example, you have a variable such as DISEASE with DISEASE=0 indicating the disease is absent and DISEASE= 1 indicating the disease is present, then SAS will predict the probability of "disease absent" by default.

Table Common Options for PROC LOGISTIC

Option	**Explanation**
DATA = dataname	Specifies which dataset to use.
DESCENDING	Reverses the sorting order for the levels of the response variable. By default, the procedure will predict the outcome corresponding to the lower value of the dichotomous dependent variable. So, if the dependent variable takes on the values 0 and 1, then by default SAS predicts the probability that the dependent variable is 0 unless you use the DESCENDING option. (See information about the (EVENT=) option below.)
ALPHA= value	Specifies significance level for confidence limits.
NOPRINT	Suppresses output.
SIMPLE	Displays descriptive statistics.
PLOTS= option	In current versions of SAS, the Odds Ratio plots is displayed by default. Use PLOTS=NONE; to suppress this plot. PLOTS=ALL produces a number of plots include ROC, and influence diagnostics.

Common Statements for PROC LOGISTIC (Table continued)	
MODEL depvar=indvar(s);	Specifies the dependent and independent variables for the analysis. More specifically, it takes the form MODEL depvariable=indvariable(s);
CLASS variable list;	Specifies classification (either categorical character or discrete numeric) variables for the analysis. They can be numeric or character. See text for more details
ODDSRATIO ‘label’ var;	Creates a separate table with Odds Ratio Estimates and Wald Confidence Intervals. See text for more details
OUTPUT out=NAME;	Creates an output dataset with all predictors and response probabilities. For example OUTPUT OUT=MYFILE P=PRED;
BY, FORMAT, LABEL, WHERE	These statements are common to most procedures, and may be used here.

The DESCENDING Option in the MODEL Statement

- The MODEL statement specifies the dependent (outcome) variable as well as the independent variables. For example,

```
PROC LOGISTIC;
MODEL DEPVAR = INDVARl INDVAR2 etc/options;
```

- Care must be taken as to how the DEPVAR is defined.
- For example, if your dependent variable is FAIL (0 means not failed & 1 means failed), then SAS will predict FAIL=0.
- To reverse the default prediction, use the **DESCENDING** option. When that option is included in the PROC LOGISTIC statement, FAIL=1 will be modeled instead of FAIL=0. Thus:

```
PROC LOGISTIC DESCENDING;
MODEL DEPVAR = INDVARl INDVAR2 etc/options;
```

Another Way to Specify What is Predicted

- Another way to choose the value modeled is to explicitly define it in the MODEL statement. For example,

```
MODEL FAIL(EVENT='1') = independentvars;
```

- Causes SAS to use 1 as the value to model for the dependent variable FAIL.
- We recommend that you choose to use either the DESCENDING option or the EVENT= option to specify a value of the response variable to predict.

Table. Common MODEL statement options for PROC Logistic

Option	**Explanation**
`EXPB`	Displays the exponentiated values of parameter, (the odds ratios.)
`SELECTION=type`	Specifies variable selection method (examples are STEPWISE, BACKWARD, and FORWARD).
`SLENTRY=value`	Specifies significance level for entering variables. Default is 0.05.
`SLSTAY=value`	Specifies significance level for removing variables. Default is 0.05.
`LACKFIT`	Requests Hosmer-Lemershow test
`RISKLIMITS`	Requests confidence limits for odds ratios.

`CTABLE`
`PPROB=(list)`
	Requests a classification table report. PPROB specifies cutpoints to display.

`INCLUDE=n`	Includes first n independent variables in model.
`OUTROC=name`	Outputs ROC values to a dataset.

The CLASS Statement

- If a model includes independent variables that are categorical, they must be indicated in a CLASS statement.

- For example, suppose the variable CATNUM is (i.e., 1, 2, 3) and CARALPH is character (i.e., A, B, C). Your LOGISTIC code might be:

```
CLASS CATNUM CATALPH;
MODEL Y = Xl X2 . . . Xk CATNUM CATALPH;;
```

Categorical variables identified in the CLASS statement.

And those same categorical variables are used as independent variables in the model.

How Logistic Handles Categorical Variables

- When a variable is defined as a classification variable, SAS sets up a default parameterization of *N* - 1 comparisons (where *N* is the number of categories).
- The **default reference** value to which the other categories are compared is based on the **last ordered (alphabetic or numeric) value.**
- For example, if RACE categories are AA, H, C and O, then ORs are reported for AA, H, and C, **based on the reference to** O since O is the last ordered (alphabetic) value.
- Similarly, if RACE is defined using discrete numeric codes such as 1, 2, 3, 4, and 5, then the last ordered (numeric) value is 5.
- **Change the reference category by including the options (REF= "value")** after the name in the CLASS statement. For example, the statement CLASS RACE (REF= " AA") makes AA the reference value rather than "0".

Another Way to Handle Categorical Variables

- Another way to handle categorical variables with three or more categories is to recode them into a series of dichotomous variables (indicator or dummy variables).
- This may make ORs easier to interpret. For example, for RACE, create three 0/1 variables in the DATA step:

```
IF RACE="AA" then RACEA=l; ELSE RACEA= 0;
IF RACE="H" then RACEH=l; ELSE RACEH= 0;
IF RACE="C" then RACEC=l; ELSE RACEC= 0;
```

- You need one less than the number of categories. Therefore, if RACEA, RACEH, and RACEC are all 0, the race must be OTHER.

USING SIMPLE LOGISTIC ANALYSIS

- A simple logistic model is one that has only one predictor (independent) variable. This predictor variable can be either a binary or a quantitative measure.

```
PROC LOGISTIC DATA=STAT3505.ACCIDENTS DESCENDING;
      MODEL DEAD=PENETRATE / RISKLIMITS;
RUN;
```

RISKLIMITS requests ORs to be output.

DEAD (which is coded 0 and 1) is what is being modeled. The DESCENDING option tells SAS to model DEAD=1

PENETRATE is a 0, 1 dichotomous variable

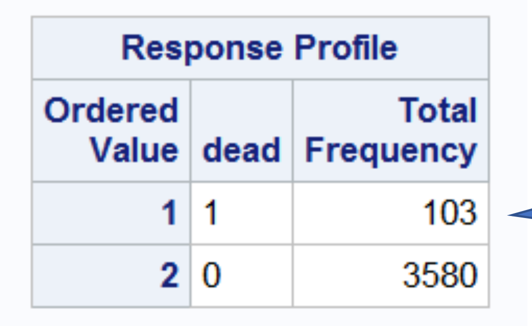
EXAMPLE

```
PROC LOGISTIC DATA=STAT3505.ACCIDENTS DESCENDING;
     MODEL DEAD=PENETRATE / RISKLIMITS;
TITLE 'Trauma Data Model Death by Penetration Wound';
RUN;
```

- In this example, the independent variable PENETRATE, which is a 0,1 variable, is used to predict death (DEAD=1), so the DESCENDING option is used.

Exercise Continued…

- Run the program. Pay special attention to this statement in the output “`Probability modeled is dead=1.`”
- It indicates what is modelled – make sure it is the output you want to model. In this case you are predicting death.
- Also note the “Response Profile”

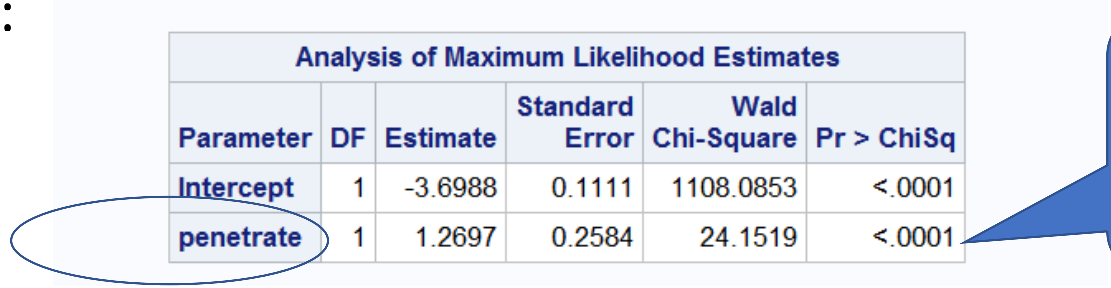


Response Profile		
Ordered Value	dead	Total Frequency
1	1	103
2	0	3580

In this case there are 103 deaths and 3580 non-deaths. IMPORTANT: Make sure these numbers are what you expect from your data.

Continued… Logistic Model Results

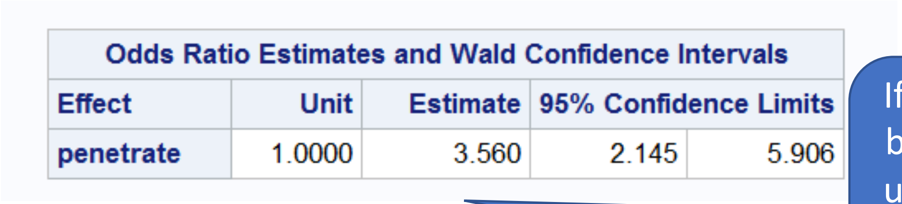
- Your primary tables of interest in the output are the estimates for the model:



Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.6988	0.1111	1108.0853	<.0001
penetrate	1	1.2697	0.2584	24.1519	<.0001

Indicates if the variable (PENETRATE) is a good predictor. Since p<0.001, we conclude that it is (when p<0.05)

- And the Odds Ratios



Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
penetrate	1.0000	3.560	2.145	5.906

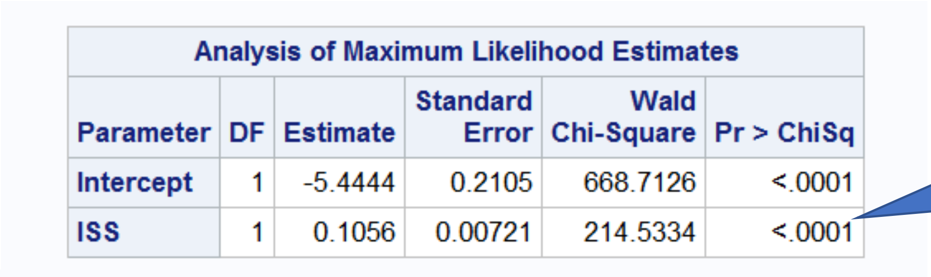
If the predictor is shown to be important, the OR gives us an idea of its strength in predicting the outcome. In this case OR=3.56

OR=3.56 indicates that the odds of a person’s dying who had a penetrating wound is 3.56 greater than that for a person who did not suffer this type of wound.

Change DEAD=PENETRATE to DEAD=ISS, Which is a Continuous Variable

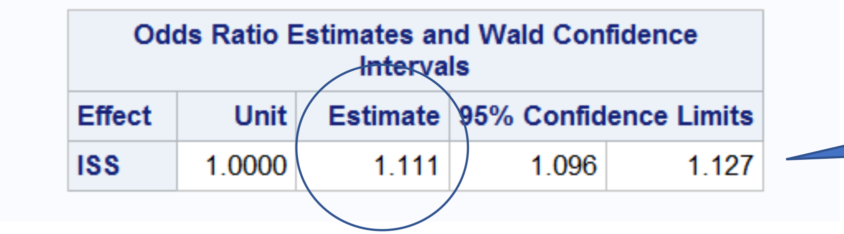
```
PROC LOGISTIC DATA=STAT3505.ACCIDENTS DESCENDING;
      MODEL DEAD=ISS / RISKLIMITS;
RUN;
```

- In this model only ISS (injury Severity Score) has changed – it is a **continuous variable** whereas PENETRATE was dichotomous.



Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.4444	0.2105	668.7126	<.0001
ISS	1	0.1056	0.00721	214.5334	<.0001

ISS is also an important predictor of death…



Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
ISS	1.0000	1.111	1.096	1.127

The odds ratio for ISS is 1.11

OR for Dichotomous vs Continuous Variables

IMPORTANT

- OR is interpreted *differently* for PENETRATE than for ISS as ISS is a quantitative measure and PENETRATE is a binary measure. ***Pay close attention to this difference.***

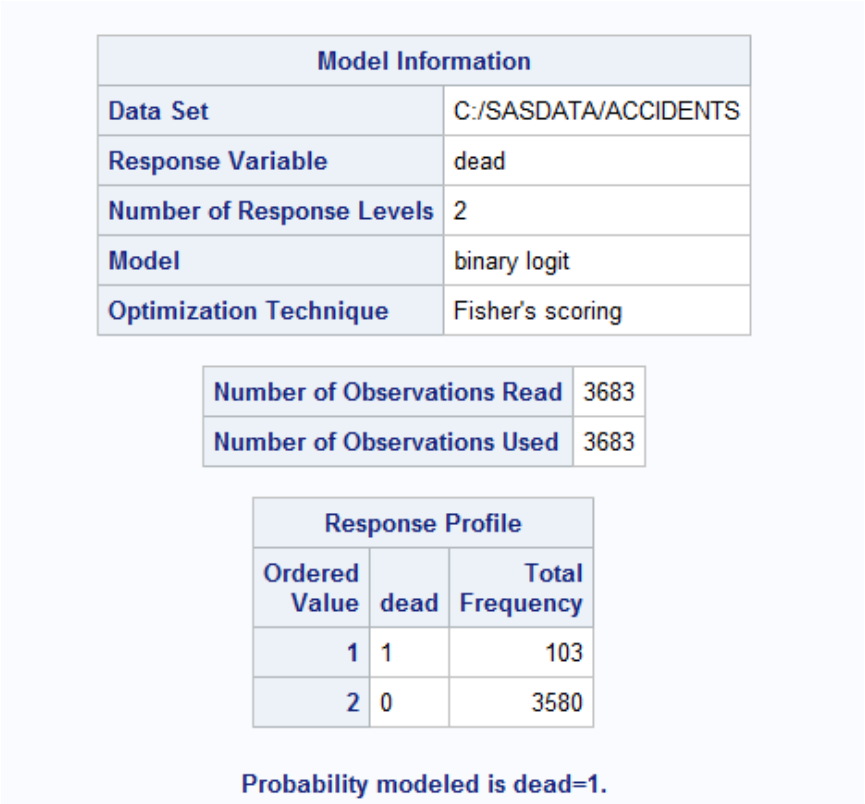
- For PENETRATE OR=3.56 indicates that the odds of a person's dying who had a penetrating wound is 3.56 times greater than that for a person who did not suffer this type of wound.

- For ISS OR=1.11 indicates that **for each unit increase** in ISS, the odds of dying increases by 1.11. (or 11%)

When OR is Less than 1

- An Odds Ratio less than 1 can also be important.
- For example, suppose a significant OR in this dataset (say AGE) was .89.
- It would be interpreted as – for each increase in AGE year, the odds of dying is LESS by about 11%.
- One way to look at it is variables with a high OR are predictive of death (the predicted outcome) and variables with an OR less than 1 are **protective** of death (the predicted outcome).

Results – First 3 Tables



Model Information	
Data Set	C:/SASDATA/ACCIDENTS
Response Variable	dead
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	3683
Number of Observations Used	3683

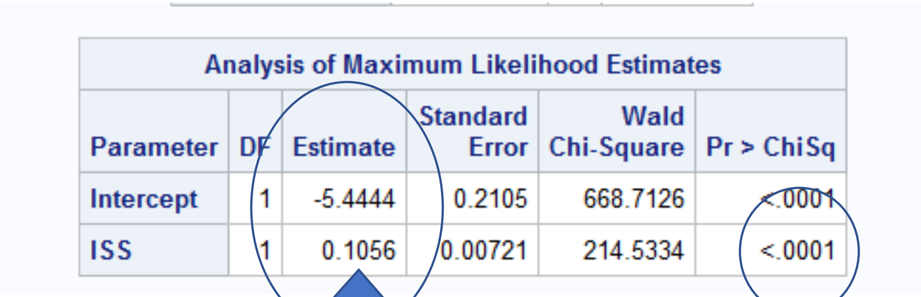
Response Profile

Ordered Value	dead	Total Frequency
1	1	103
2	0	3580

Probability modeled is dead=1.

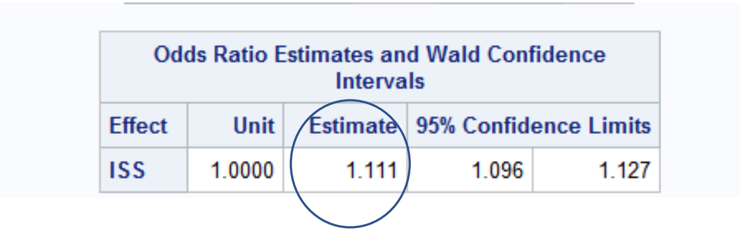
- Top table – Summary information about model (note Response Variable)
- Middle Table – Make sure observations are as expected
- Bottom Table – Make sure the key variable (In this case DEAD) has the expected number of obs.
- Note the Probability modeled is DEAD = 1

Results Continued



Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.4444	0.2105	668.7126	<.0001
ISS	1	0.1056	0.00721	214.5334	<.0001

Note these values



Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
ISS	1.0000	1.111	1.096	1.127

- Maximum Likelihood table – Note the ISS row – the p-value is p<.0001 indicating that ISS is a good predictor of DEAD. (Recall ISS is a continuous Variable)
- The ODDS RATIO estimate is 1.111

Simple Logistic Results

- The logistic equation based on estimates given in the Maximum-Likelihood Estimates tables is

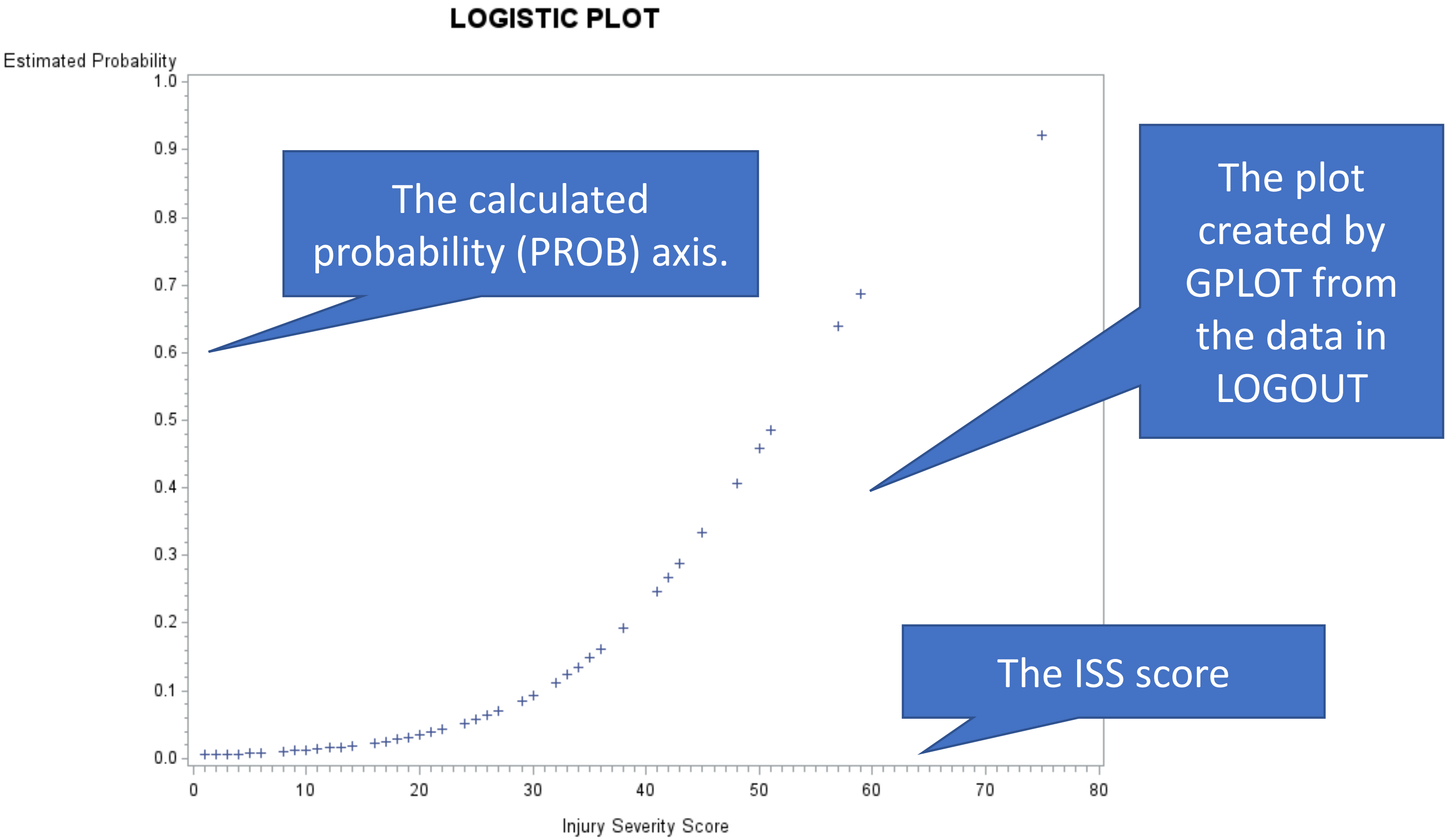
$$\hat{p} = \frac{e^{-5.444+.1056*ISS}}{1+e^{-5.444+.1056*ISS}}$$

- Where p-hat is the prediction calculated for a value of ISS.
- Because the code

```
OUTPUT OUT=LOGOUT PREDICTED=PROB;
```

was used in the program, a file named LOGOUT contains the values of p-hat (labeled PROB) for each value of ISS.

Plotting the values of PROB (Predictions)

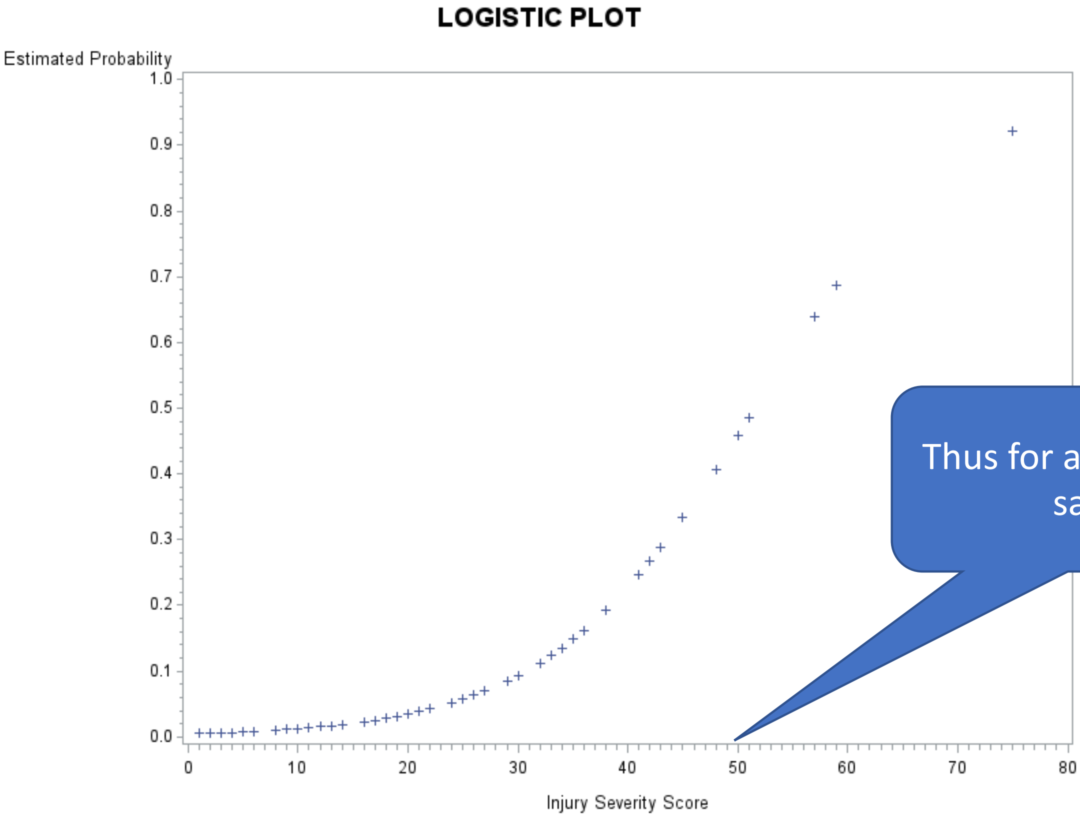


The calculated probability (PROB) axis.

The plot created by GPLOT from the data in LOGOUT

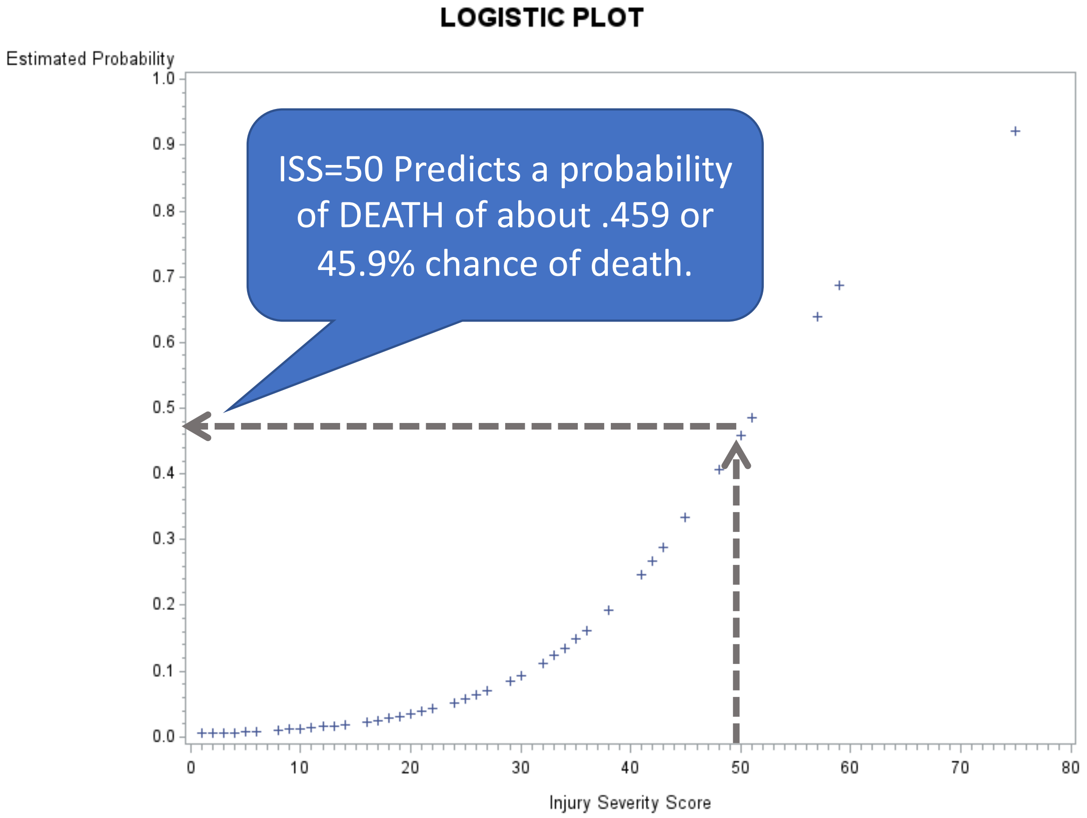
The ISS score

Plotting the values of PROB (Predictions)



Thus for a value of ISS, say 50

Plotting the values of PROB (Predictions)



ISS=50 Predicts a probability of DEATH of about .459 or 45.9% chance of death.

MULTIPLE BINARY LOGISTIC ANALYSIS

- A multiple binary logistic regression model has more than one independent variable. As such, it is analogous to a multiple regression model in the case in which the dependent variable is binary.

- It is common to have several potential predictor variables.

- One of the tasks of the investigator is to select the best set of predictors to create a parsimonious and effective prediction equation.

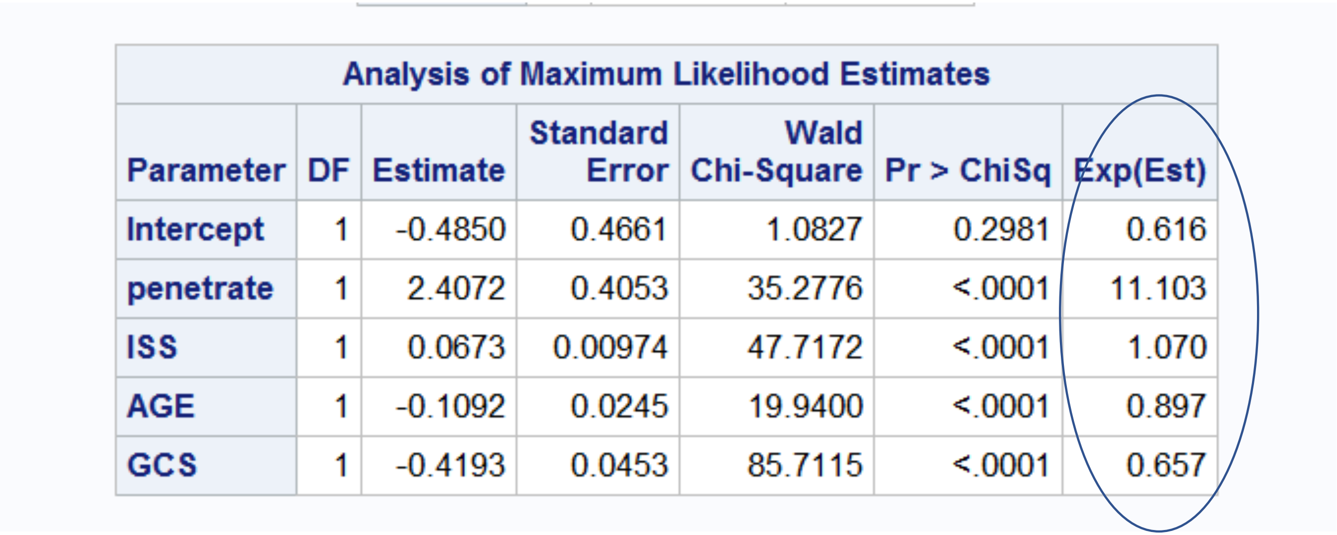
Results – NEXT STEPS

- Variables entered into the model

1. Penetrate (Forced into model by `INCLUDE=1)`
2. GCS (Next variable selected by STEPWISE)
3. ISS
4. AGE
5. Note: No (additional) effects met the 0.05 significance level for entry into the model.

The Final Model

- This table report the estimates of the parameters for the logistic model. The EXP(Est) column are the Odds Ratios.
- The Pr>ChiSq indicates the significance of each variable.
- (We are usually not interested in p for the Intercept.)

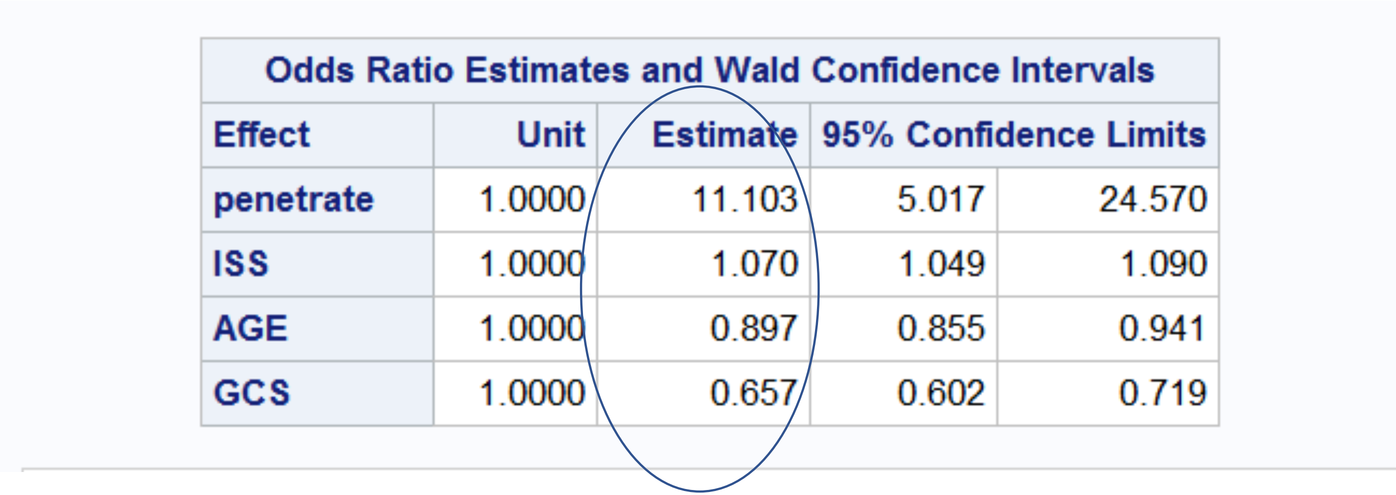


Analysis of Maximum Likelihood Estimates

			Standard	Wald		
Parameter	**DF**	**Estimate**	**Error**	**Chi-Square**	**Pr > ChiSq**	**Exp(Est)**
Intercept	1	-0.4850	0.4661	1.0827	0.2981	0.616
penetrate	1	2.4072	0.4053	35.2776	<.0001	11.103
ISS	1	0.0673	0.00974	47.7172	<.0001	1.070
AGE	1	-0.1092	0.0245	19.9400	<.0001	0.897
GCS	1	-0.4193	0.0453	85.7115	<.0001	0.657

The Odds Ratio Report

- More information about the Odds Ratios are given in this table:
- The Estimate provides a measure of the importance of the OR. *Keep in mind the different interpretations for binary and quantitative variables*.



Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
penetrate	1.0000	11.103	5.017	24.570
ISS	1.0000	1.070	1.049	1.090
AGE	1.0000	0.897	0.855	0.941
GCS	1.0000	0.657	0.602	0.719

What is ODS?

What is ODS?

- ODS, or **Output Delivery System**, is a method for controlling the output from SAS® procedures. ODS began with version 8 and continues with added enhancements in more recent versions.
- Prior to SAS 9.3, SAS output appeared in the Output Window. This output listing is like a monospaced typewriter font (with no graphics) and there are few options that allow you to control the "look" of the listing.
- Beginning with 9.3 **default** output is HTML type output.

SPECIFYING THE ODS OUTPUT FORMAT AND DESTINATION

- The SAS ODS is set up so that you "turn on" or initiate output into a designated output format. Once the output format has been initiated, SAS procedures send information to that output format. You can send output from one or more procedures to the output stream.

```
ODS OUTDESTINATION <OPTIONS>;
```

Used to tell SAS to start outputting results to a specified **output** type.

- To end the ODS output, use the **CLOSE** statement:

```
ODS OUTDESTINATION CLOSE;
```

Tells SAS that you are finished outputting results.

What do you do if your output is not showing up in the Results Viewer?

What do you do if your output is not showing up in the Results Viewer?

ODS PREFERENCES;

(Or end SAS and restart)

USING ODS TO SELECT SPECIFIC OUTPUT TABLES

- SAS procedures often output a lot of information you don't want or need. In ODS output, each part of the output is contained in a table.
- Using ODS options, you can customize **which tables you want SAS to output** to the ODS
- To include or exclude a table from the output, you first need to know the table's name. You can discover this information by using the ODS TRACE command in the following way:

```
ODS TRACE ON;
   PROC whatever;
ODS TRACE OFF;
```

Putting the TRACE code around a PROC produces information in the LOG that tells you the names of the ODS output tables.

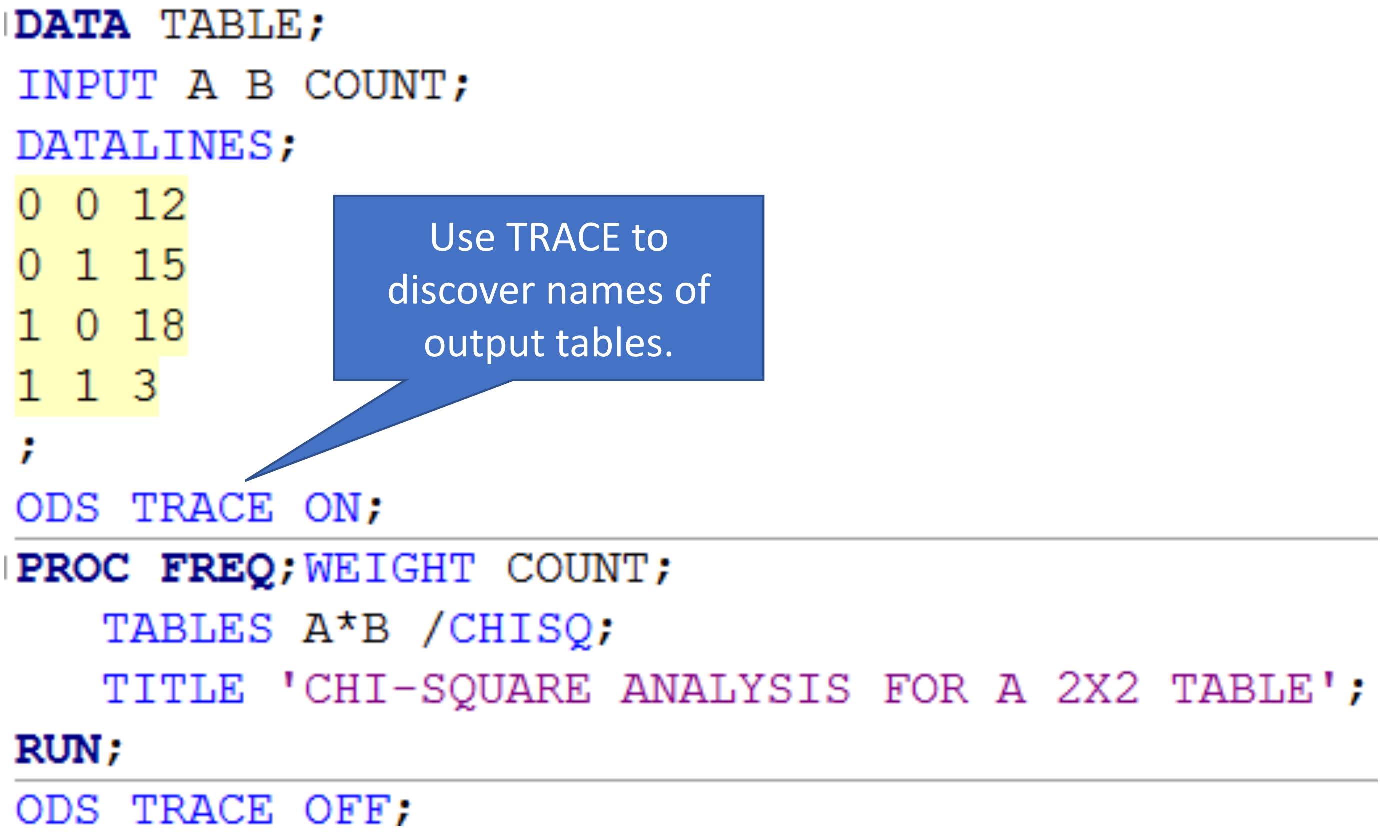
Specifying which tables to display

- Once you know the names of the tables you want to display (using TRACE), use the following code to make that request:

This ODS statement tells SAS to output only the specified table from PROC whatever

```
ODS SELECT name-of-tables-to-include;
  PROC whatever;
Etc…
```

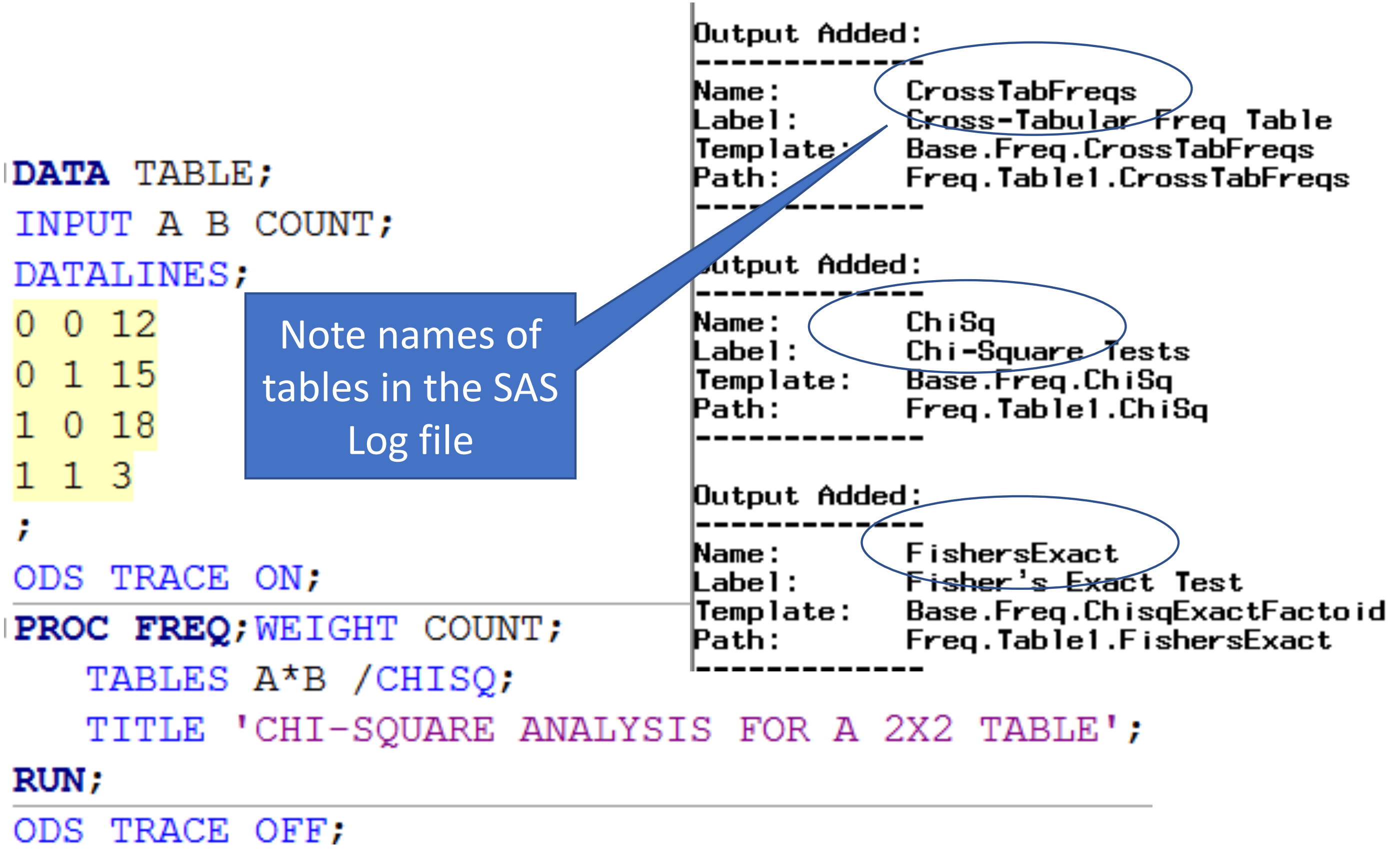
EXAMPLE:



```
DATA TABLE;
INPUT A B COUNT;
DATALINES;
0 0 12
0 1 15
1 0 18
1 1 3
;
ODS TRACE ON;
PROC FREQ;WEIGHT COUNT;
   TABLES A*B /CHISQ;
   TITLE 'CHI-SQUARE ANALYSIS FOR A 2X2 TABLE';
RUN;
ODS TRACE OFF;
```

Use TRACE to discover names of output tables.

Results of running this code:



```
DATA TABLE;
INPUT A B COUNT;
DATALINES;
0 0 12
0 1 15
1 0 18
1 1 3
;
ODS TRACE ON;
PROC FREQ;WEIGHT COUNT;
   TABLES A*B /CHISQ;
   TITLE 'CHI-SQUARE ANALYSIS FOR A 2X2 TABLE';
RUN;
ODS TRACE OFF;
```

Note names of tables in the SAS Log file

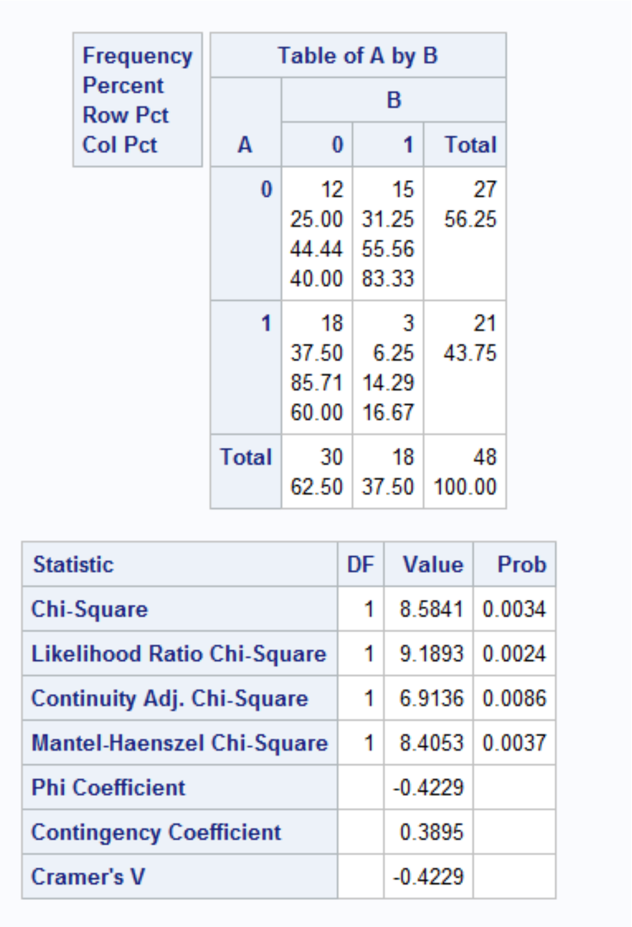
```
Output Added:
-------------
Name:       CrossTabFreqs
Label:      Cross-Tabular Freq Table
Template:   Base.Freq.CrossTabFreqs
Path:       Freq.Table1.CrossTabFreqs
-------------

Output Added:
-------------
Name:       ChiSq
Label:      Chi-Square Tests
Template:   Base.Freq.ChiSq
Path:       Freq.Table1.ChiSq
-------------

Output Added:
-------------
Name:       FishersExact
Label:      Fisher's Exact Test
Template:   Base.Freq.ChisqExactFactoid
Path:       Freq.Table1.FishersExact
-------------
```

Use that information:

- **STEP 2:** Use **SELECT** ***tablenames*** to produce output that ONLY contains the tables of interest:

```
ODS SELECT CROSSTABFREQS CHISQ;
PROC FREQ;WEIGHT COUNT;
   TABLES A*B /CHISQ;
RUN;
```

- (You can also use **ODS EXCLUDE** to exclude certain tables from output.)



Frequency Percent Row Pct Col Pct	Table of A by B		
	B		
A	0	1	Total
0	12 25.00 44.44 40.00	15 31.25 55.56 83.33	27 56.25
1	18 37.50 85.71 60.00	3 6.25 14.29 16.67	21 43.75
Total	30 62.50	18 37.50	48 100.00

Statistic	DF	Value	Prob
Chi-Square	1	8.5841	0.0034
Likelihood Ratio Chi-Square	1	9.1893	0.0024
Continuity Adj. Chi-Square	1	6.9136	0.0086
Mantel-Haenszel Chi-Square	1	8.4053	0.0037
Phi Coefficient		-0.4229	
Contingency Coefficient		0.3895	
Cramer's V		-0.4229	

CAPTURING INFORMATION FROM ODS TABLES

- Once you know the name of an output table, you can use ODS to save the table contents into a SAS data file using:

- `ODS OUTPUT NAMEOFTABLE=OUTPUTDATASET;`

You learn the name of the table using the TRACE statement

When you output a table using ODS OUTPUT, you are creating a SAS data set.

EXAMPLE

- This example shows how to capture a specific statistics from the output.

```
DATA WT;
INPUT WEIGHT @@;
DATALINES;
64 71 53 67 55 58
77 57 56 51 76 68
;
ODS TRACE ON;
PROC MEANS DATA=WT;
RUN;
ODS TRACE OFF;RUN;
QUIT;
```

- Run this program

Capturing Output Continued…

- The ODS TRACE outputs this information in Log:

```
Output Added:
-------------
Name: Summary
Label: Summary statistics
Template: base.summary
Path: Means.Summary
```

Summary is the name of the output table containing the statistics from the PROC.