Chapter 226

Analysis of Covariance (ANCOVA) with Two Groups

Introduction

This procedure performs analysis of covariance (ANCOVA) for a grouping variable with 2 groups and one covariate variable. This procedure uses multiple regression techniques to estimate model parameters and compute least squares means. This procedure also provides standard error estimates for least squares means and their differences, and computes the T-test for the difference between group means adjusted for the covariate. The procedure also provides response vs covariate by group scatter plots and residuals for checking model assumptions.

This procedure will output results for a simple two-sample equal-variance T-test if no covariate is entered and simple linear regression if no group variable is entered. This allows you to complete the ANCOVA analysis if either the group variable or covariate is determined to be non-significant. For additional options related to the T-test and simple linear regression analyses, we suggest you use the corresponding procedures in **NCSS**.

The group variable in this procedure is restricted to two groups. If you want to perform ANCOVA with a group variable that has three or more groups, use the One-Way Analysis of Covariance (ANCOVA) procedure.

This procedure cannot be used to analyze models that include more than one covariate variable or more than one group variable. If the model you want to analyze includes more than one covariate variable and/or more than one group variable, use the General Linear Models (GLM) for Fixed Factors procedure instead.

Kinds of Research Questions

A large amount of research consists of studying the influence of a set of independent variables on a response (dependent) variable. Many experiments are designed to look at the influence of a single independent variable (factor or group) while holding other factors constant. These experiments are called single-factor or single-group experiments and are analyzed with the one-way analysis of variance (ANOVA) or a two-sample T-test. Analysis of covariance (ANCOVA) is useful when you want to improve precision by removing extraneous sources of variation from your study by including a covariate.

The ANCOVA Model

The analysis of covariance uses features from both analysis of variance and multiple regression. The usual one-way classification model in analysis of variance is

$$Y_{ij} = \mu_i + e_{1ij}$$

where Y_{ij} is the j^{th} observation in the i^{th} group, μ_i represents the true mean of the i^{th} group, and e_{1ij} are the residuals or errors in the above model (usually assumed to be normally distributed). Suppose you have measured a second variable with values X_{ij} that is linearly related to Y. Further suppose that the slope of the relationship between Y and X is constant from group to group. You could then write the analysis of covariance model assuming equal slopes as

$$Y_{ij} = \mu_i + \beta\left(X_{ij} - \bar{X}_{..}\right) + e_{2ij}$$

where $\bar{X}_{..}$ represents the overall mean of X. If X and Y are closely related, you would expect that the errors, e_{2ij}, would be much smaller than the errors, e_{1ij}, giving you more precise results.

The classical analysis of covariance is useful for many reasons, but it does have the (highly) restrictive assumption that the slope is constant over all the groups. This assumption is often violated, which limits the technique's usefulness. You will want to study more about this technique in statistical texts before you use it.

If it is not reasonable to conclude that the slopes are equal, then a covariate-by-group interaction term should be included in the model.

Assumptions

The following assumptions are made when using the F-test.

1. The response variable is continuous.
2. The treatments do not affect the value of the covariate, X_{ij}.
3. The e_{2ij} follow the normal probability distribution with mean equal to zero.
4. The variances of the e_{2ij} are equal for all values of i and j.
5. The individuals are independent.

Normality of Residuals

The residuals are assumed to follow the normal probability distribution with zero mean and constant variance. This can be evaluated using a normal probability plot of the residuals. Also, normality tests are used to evaluate this assumption. The most popular of the five normality tests provided is the Shapiro-Wilk test.

Unfortunately, a breakdown in any of the other assumptions results in a departure from this assumption as well. Hence, you should investigate the other assumptions first, leaving this assumption until last.

Limitations

There are few limitations when using these tests. Sample sizes may range from a few to several hundred. If your data are discrete with at least five unique values, you can assume that you have met the continuous variable assumption. Perhaps the greatest restriction is that your data comes from a random sample of the population. If you do not have a random sample, the F-test will not work.

Representing Group Variables

Categorical group variables take on only a few unique values. For example, suppose a therapy variable has three possible values: A, B, and C. One question is how to include this variable in the regression model. At first glance, we can convert the letters to numbers by recoding A to 1, B to 2, and C to 3. Now we have numbers. Unfortunately, we will obtain completely different results if we recode A to 2, B to 3, and C to 1. Thus, a direct recode of letters to numbers will not work.

To convert a categorical variable to a form usable in regression analysis, we must create a new set of numeric variables. If a categorical variable has *k* values, *k* - 1 new binary variables must be generated.

Indicator (Binary) Variables

Indicator (dummy or binary) variables are created as follows. A *reference group* is selected. Usually, the most common value or the control is selected as the reference group. Next, a variable is generated for each of the groups other than the reference group. For example, suppose that B is selected as the reference group. An indicator variable is generated for the remaining value, A. The value of the indicator variable is one if the value of the original variable is equal to the value of interest, or zero otherwise. Here is how the original variable T and the two new indicator variable TA looks in a short example.

T	TA
A	1
A	1
B	0
B	0

The generated variable, TA, would be used as columns in the design matrix, *X*, in the model.

Representing Interactions of Numeric and Categorical Variables

When the interaction between a group variable and a covariate is to be included in the model, all proceeds as above, except that an interaction variable must be generated for each categorical variable. This can be accomplished automatically in **NCSS** based on the slopes assumption. When assuming that the slopes are unequal all applicable covariate-by-group interaction variables are automatically created.

In the following example, the interaction between the group variable *T* and the covariate variable *X* is created.

T	TA	X	XTA
A	1	1.2	1.2
A	1	1.4	1.4
B	0	2.3	0
B	0	4.7	0

When the variable *XTA* is added to the model, it will account for the interaction between *T* and *X*.

Technical Details

This section presents the technical details of the analysis method (multiple regression) using a mixture of summation and matrix notation.

The Linear Model

The linear model can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

where **Y** is a vector of N responses, **X** is an $N \times p$ design matrix, **β** is a vector of p fixed and unknown parameters, and ***e*** is a vector of N unknown, random error values. Define the following vectors and matrices:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_j \\ \vdots \\ y_N \end{bmatrix},\ \mathbf{X} = \begin{bmatrix} 1\ x_{11} \cdots x_{1p} \\ \vdots \\ 1\ x_{1j} \cdots x_{pj} \\ \vdots \\ 1\, x_{1N} \cdots x_{pN} \end{bmatrix},\ \mathbf{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_j \\ \vdots \\ e_N \end{bmatrix},\ \mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{bmatrix},\ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

X is the design matrix that includes the covariate, a binary variable formed from the group variable, and a variable resulting from the covariate-by-group interaction (if included).

Least Squares

Using this notation, the least squares estimates of the model coefficients, **b**, are found using the equation.

$$\mathbf{b} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$$

The vector of predicted values of the response variable is given by

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$$

The residuals are calculated using

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$$

Estimated Variances

An estimate of the variance of the residuals is computed using

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{N - p - 1}$$

An estimate of the variance of the model coefficients is calculated using

$$\mathbf{V}\begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix} = s^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

An estimate of the variance of the predicted mean of Y at a specific value of X, say X_0, is given by

$$s^2_{Y_m|X_0} = s^2\left(1, X_0\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}\begin{pmatrix} 1 \\ X_0 \end{pmatrix}$$

An estimate of the variance of the predicted value of *Y* for an individual for a specific value of *X*, say X_0, is given by

$$s^2_{Y_I|X_0} = s^2 + s^2_{Y_m|X_0}$$

Hypothesis Tests of the Intercept and Coefficients

Using these variance estimates and assuming the residuals are normally distributed, hypothesis tests may be constructed using the Student's *t* distribution with *N* - *p* - 1 degrees of freedom using

$$t_{b_i} = \frac{b_i - \mathrm{B}_i}{s_{b_i}}$$

Usually, the hypothesized value of B_i is zero, but this does not have to be the case.

Confidence Intervals of the Intercept and Coefficients

A $100(1-\alpha)\%$ confidence interval for the true model coefficient, β_i, is given by

$$b_i \pm (t_{1-\alpha/2,N-p-1})s_{b_i}$$

Confidence Interval of Y for Given X

A $100(1-\alpha)\%$ confidence interval for the mean of *Y* at a specific value of *X*, say X_0, is given by

$$b'X_0 \pm (t_{1-\alpha/2,N-p-1})s_{Y_m|X_0}$$

A $100(1-\alpha)\%$ prediction interval for the value of *Y* for an individual at a specific value of *X*, say X_0, is given by

$$b'X_0 \pm (t_{1-\alpha/2,N-p-1})s_{Y_I|X_0}$$

R2 (Percent of Variation Explained)

Several measures of the goodness-of-fit of the model to the data have been proposed, but by far the most popular is R^2. R^2 is the square of the correlation coefficient between *Y* and $\hat{Y}$. It is the proportion of the variation in *Y* that is accounted by the variation in the independent variables. R^2 varies between zero (no linear relationship) and one (perfect linear relationship).

R^2, officially known as the *coefficient of determination*, is defined as the sum of squares due to the linear regression model divided by the adjusted total sum of squares of *Y*. The formula for R^2 is

$$R^2 = 1 - \left(\frac{\mathbf{e'e}}{\mathbf{Y'Y} - \frac{(\mathbf{1'Y})^2}{\mathbf{1'1}}} \right)$$
$$= \frac{SS_{Model}}{SS_{Total}}$$

R^2 is probably the most popular measure of how well a model fits the data. R^2 may be defined either as a ratio or a percentage. Since we use the ratio form, its values range from zero to one. A value of R^2 near zero indicates no linear relationship, while a value near one indicates a perfect linear fit. Although popular, R^2 should not be used indiscriminately or interpreted without scatter plot support. Following are some qualifications on its interpretation:

1. *Additional independent variables*. It is possible to increase R^2 by adding more independent variables, but the additional independent variables may cause an increase in the mean square error, an unfavorable situation. This usually happens when the sample size is small.
2. *Range of the independent variables*. R^2 is influenced by the range of the independent variables. R^2 increases as the range of the *X's* increases and decreases as the range of the *X's* decreases.
3. *Slope magnitudes*. R^2 does not measure the magnitude of the slopes.
4. *Linearity*. R^2 does not measure the appropriateness of a linear model. It measures the strength of the linear component of the model. Suppose the relationship between *X* and *Y* was a perfect sphere. Although there is a perfect relationship between the variables, the R^2 value would be zero.
5. *Predictability*. A large R^2 does not necessarily mean high predictability, nor does a low R^2 necessarily mean poor predictability.
6. *Sample size*. R^2 is highly sensitive to the number of observations. The smaller the sample size, the larger its value.

Rbar² (Adjusted R²)

R^2 varies directly with *N*, the sample size. In fact, when $N = p$, $R^2 = 1$. Because R^2 is so closely tied to the sample size, an adjusted R^2 value, called $\overline{R}^2$, has been developed. $\overline{R}^2$ was developed to minimize the impact of sample size. The formula for $\overline{R}^2$ is

$$\overline{R}^2 = 1 - \frac{(N-1)(1-R^2)}{N-p-1}$$

Least Squares Means

As opposed to raw or arithmetic means which are simply averages of the grouped raw data values, least squares means are adjusted for the other terms in the model, such as covariates. In balanced designs with no covariates, the least squares group means will be equal to the raw group means. In unbalanced designs or when covariates are present, the least squares means usually are different from the raw means.

The least squares means and associated comparisons (i.e. differences) can be calculated using a linear contrast vector, $\mathbf{c_i}$. Means and differences are estimated as

$$\mathbf{c_i}'\mathbf{b},$$

with estimated standard error,

$$SE(\mathbf{c_i}'\mathbf{b}) = s\sqrt{\mathbf{c_i}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c_i}}\,.$$

where *s* is the square root of the estimated mean square error (MSE) from the model based on *v* degrees of freedom.

For an ANCOVA model with a group variable with 2 levels and a covariate, and if level 2 were the reference value, the components of the contrast vector would take the form

$$c_i = (I, \mu_1, X)$$

where *I* represents the and indicator for the intercept and *X* is the value of the covariate which the mean or difference is evaluated. The contrast vector used to estimate μ_2 would be

$$c_i = (1, 0, X)\,.$$

The contrast vector used to estimate $\mu_1 - \mu_2$ would be

$$c_i = (0,1,0)\,.$$

Confidence intervals for the estimable functions are of the form

$$\mathbf{c}_i'\mathbf{b} \pm c_\alpha SE(\mathbf{c}_i'\mathbf{b}),$$

where c_α is the critical value, usually selected to keep the experimentwise error rate equal to α for the collection of all comparisons (see Multiple Comparisons).

Two-sided significance tests for the mean and the difference (against a null value of zero) use the test statistic

$$T_i = \frac{|\mathbf{c}_i'\mathbf{b}|}{SE(\mathbf{c}_i'\mathbf{b})} \geq c_\alpha \ .$$

Data Structure

The data must be entered in a format that puts the responses in one column, the group values in a second column, and the covariate values in a third column. An example of data that might be analyzed using this procedure is shown next. The data contains a response variable, a group variable with two groups (State), and a covariate (Age).

ANCOVA2Grp dataset

State	Age	Response
Iowa	12	100
Iowa	13	102
Iowa	12	97
.	.	.
.	.	.
.	.	.
Utah	14	104
Utah	11	105
Utah	12	106
.	.	.
.	.	.
.	.	.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

These panels specify the variables used in the analysis and the model.

Variables

Response Variable(s)

Enter one or more numeric response (Y or dependent) variables. You can enter the variable names or numbers directly, or double-click in the box to display a Column Selection window that will let you select the variable(s) from a list. Each variable contains all data for all categories, one row per subject. If multiple variables are selected, each gives the values of a different response variable and a separate analysis is conducted for each.

Group Variable

The values of the group (categorical) variable indicate which group (category) the observation (row) belongs in. The variable must have only 2 unique values. Values may be text or numeric. Examples of group variables are gender, age group, or treatment.

You can enter the variable name or number directly, or double-click in the box to display a Column Selection window that will let you select a variable from a list. Only one group variable may be entered.

Reference Group

This option specifies which group will be designated as the reference group, which is used when generating internal numeric variables from categorical variables. No internal binary variable is generated for the reference group. It is commonly the baseline or control group, and the other groups are compared to it in Each vs. Reference Group comparisons. The choices are

- **First Group after Sorting – Fifth Group after Sorting**
  Use the first (through fifth) group in alpha-numeric sorted order as the reference group.
- **Last Group after Sorting**
  Use the last group after sorting as the reference group.
- **Custom**
  Use the user-entered group as the reference group.

Custom Reference Group

This group will be used as the reference or control group. If this option is left blank, then the first group after sorting will be used as the reference group.

Covariate Variable

Specify a numeric covariate variable to include in the model. This option is required for an ANCOVA model. If this variable is left blank, the model will be a one-way ANOVA model.

What is a Covariate?

We consider a variable to be a covariate if its values are numbers that are at least ordinal (i.e. “numeric”). Nominal variables are classified as categorical, even if their values are numbers.

Calculating Least Squares Means with Covariates

When a covariate is included in the model, the least squares means for groups are calculated with the covariate variable evaluated at a specific value. Often, the covariate mean is used, but the least squares means can be calculated with the covariate evaluated at any value, which can be specified below. When the model includes the covariate-by-group interaction term, you may want to calculate and compare means at various values of the covariate.

Calc Group Means at

This option allows you to specify what value(s) to use for the covariate variable when calculating the least squares means for the group variable. Often, the least squares means are calculated at the mean of the covariate. When the model includes the covariate-by-group interaction term, however, you may want to calculate and compare means at various values of the covariate.

Note: If this option is left blank, least squares means will be calculated with the covariate evaluated at its mean.

You can specify exactly what value to use for each covariate using the following syntax:

- **"Covariate Mean"**
  Least squares means are calculated with the covariate evaluated at its mean.

- **[Enter a Covariate Value]**
  If a single numeric value is entered, least squares means are calculated with the covariate evaluated at this value (for example, enter "70" to calculate least squares means with the covariate evaluated at 70).

- **[Enter a List of Covariate Values]**
  If a list separated by blanks or commas is entered, sets of least squares means are calculated with the covariate evaluated at each value (for example, enter "70 80 90" to calculate 3 sets of least squares means with the covariate evaluated at 70, then 80, and then 90).

ANCOVA Model Slopes Assumption

This section specifies the model design based on the assumption for the group regression line slopes.

Assume Slopes are

Select what assumption to use for the slopes. This will determine the type of ANCOVA model to analyze.

Possible choices are:

- **Equal (No Covariate-by-Group Interaction)**
  Slopes are assumed to be equal. The covariate variable and group variable are included in the model without the covariate-by-group interaction term.

  Note: This model should only be used if the covariate-by-group interaction term is not significant. To determine if the interaction is significant, first run the model with unequal slopes that includes the covariate-by-group interaction and review the test of the interaction term in the Analysis of Variance report.

- **Unequal (Include Covariate-by-Group Interaction)**
  The covariate variable, group variable, and the covariate-by-group interaction term are included in the model. This model is suitable when the slopes are not equal among groups.

  Note: When the model includes a significant covariate-by-group interaction, you may want to calculate and compare means at various values of the covariate and consider the results collectively. If you calculate and compare means at only one covariate value, the results may be misleading.

Reports Tab

The following options control which reports are displayed.

Alpha and Confidence Level

Tests Alpha

Alpha is the significance level used in conducting the hypothesis tests.

Confidence Level

Enter the confidence level (or confidence coefficient) for the confidence intervals reported in this procedure. Note that, unlike the value of alpha, the confidence level is entered as a percentage.

Select Reports – Summaries

Run Summary

This report summarizes the results. It presents the number of variables and rows used, basic statistical results such as R^2 and mean square error, and whether the procedure completed normally.

Descriptive Statistics

This report provides the count, arithmetic mean, standard deviation, minimum, and maximum of each variable. It is particularly useful for checking that the correct variables were used.

Select Reports – Analysis of Variance

ANOVA Table

This report provides an ANOVA table that provides tests for each individual term in the model and the model as a whole.

Select Reports – Model Coefficients

Coefficient T-Tests

This reports the estimated model coefficients, their standard errors, and significance tests.

Coefficient C.I.'s

This report provides the estimated model coefficients, their standard errors, and confidence intervals.

Select Reports – Least Squares Means

Least Squares Means

This report provides estimates of the model-adjusted least squares means, their standard errors, and confidence intervals. The confidence limits are calculated using the user-entered *Confidence Level* value.

Least Squares Means with Hypothesis Tests of H0: Mean = 0

This report provides estimates of the model-adjusted least squares means, their standard errors, and tests of H0: Mean = 0 vs. H1: Mean ≠ 0.

Select Reports – Comparison of Group Least Squares Means

Hypothesis Test and Confidence Interval Direction

Specify the direction of the T-tests and confidence intervals for the differences between group means.

Possible Choices:

- **Two-Sided**
  Hypotheses Tested: H0: Diff = 0 vs. H1: Diff ≠ 0

  Creates confidence intervals with both upper and lower limits (e.g. [5.3, 14.2]).

- **One-Sided Lower**
  Hypotheses Tested: H0: Diff ≥ 0 vs. H1: Diff < 0

  Creates confidence intervals with an upper limit only (e.g. [-Infinity, 14.2]).

- **One-Sided Upper**
  Hypotheses Tested: H0: Diff ≤ 0 vs. H1: Diff > 0

  Creates confidence intervals with a lower limit only (e.g. [5.3, Infinity]).

T-Tests for Group Least Squares Mean Differences

This report provides a T-test of the difference between group least squares means at each user-entered covariate value. Tests and confidence intervals may be either two-sided or one-sided.

Show Confidence Intervals

This report provides a confidence interval for the group least squares mean difference at each user-entered covariate value. The confidence limits are calculated using the user-entered *Confidence Level* value.

Note: If you want the confidence limits to match the hypothesis test conclusions in the reports, you should make sure that the user-entered *Confidence Level* value equals 100 × (1 minus the user-entered *Tests Alpha* value).

Select Reports – Assumptions

Residual Normality Tests

This report provides the results of several normality tests of the residuals including the Shapiro-Wilk test and the Anderson-Darling test.

Normality Test Alpha

This value specifies the significance level that must be achieved to reject the null hypothesis of residual normality. In regular hypothesis tests, common values of alpha are 0.05 and 0.01. However, most statisticians recommend that preliminary tests of assumptions use a larger alpha such as 0.10, 0.15, or 0.20.

Select Reports – Row-by-Row Lists

Show Which Rows

This option makes it possible to limit the number of rows shown in the lists. This is useful when you have a large number of rows of data.

- **Only Rows Missing Y**
  Only those rows in which the dependent variable's value is missing are displayed.

- **All Rows**
  All rows are displayed.

Residuals

Besides showing the residuals, this report also gives the predicted Y value and associated absolute percent error for each row.

Predicted Values for Means

This report the predicted values with a confidence interval for the mean of each row of the database.

Predicted Values for Individuals

This report the predicted values with a prediction interval for an individual of each row of the database.

Report Options Tab

Variable Labels

Variable Names

Specify whether to use variable names, variable labels, or both to label output reports. This option is ignored by some of the reports.

Stagger label and output if label length is ≥

When writing a row of information to a report, some variable names/labels may be too long to fit in the space allocated. If the name (or label) contains more characters than specified here, the rest of the output for that line is moved down to the next line. Most reports are designed to hold a label of up to 15 characters.

Enter ***1*** when you always want each row's output to be printed on two lines. Enter ***100*** when you want each row printed on only one line. Note that this may cause some columns to be miss-aligned.

Decimal Places

Precision

This option is used when the number of decimal places is set to ***All***. It specifies whether numbers are displayed as single (7-digit) or double (13-digit) precision numbers in the output. All calculations are performed in double precision regardless of the Precision selected here.

Single

Unformatted numbers are displayed with 7-digits

Double

Unformatted numbers are displayed with 13-digits. This option is most often used when the extremely accurate results are needed for further calculation. For example, double precision might be used when you are going to use the Multiple Regression model in a transformation.

Double Precision Format Misalignment

Double precision numbers may require more space than is available in the output columns, causing column alignment problems. The double precision option is for those instances when accuracy is more important than format alignment.

Decimal Places

Specify the number of digits after the decimal point to display on the output of values of this type. This option in no way influences the accuracy with which the calculations are done.

All

Select ***All*** to display all digits available. The number of digits displayed by this option is controlled by whether the *Precision* option is *Single* (7) or *Double* (13).

Note

This option in no way influences the accuracy with which the calculations are done.

Plots Tab

These options control the inclusion and the settings of each of the plots.

Select Plots – Response vs Covariate

Response vs Covariate by Group Scatter Plot

Check this box to display a scatter plot with the response variable on the vertical Y axis, the covariate variable on the horizontal X axis, and the group variable in the legend. By default, regression trendlines are displayed. These regression lines and associated confidence and prediction intervals correspond to the model being analyzed, whether with or without the assumption of equal slopes.

This plot shows the relationship between the response variable and the covariate variable for each level of the group variable. It allows you to visually inspect whether the assumption of equal slopes is reasonable or whether the covariate-by-group interaction should be included in the model.

This plot also allows you to study whether a linear relationship exists and whether there are outliers.

Select Plots – Means

Means Plots

Check this box to display least squares means error-bar plots. The variation lines on the plots can be either standard errors for the means (SE's) or confidence limits.

The individual means plots can be edited using the first plot format button. The combined multiple-covariate-value plots can be edited using the second plot format button.

Means plots provide a quick graphical representation of the means and their variation.

Variation Line Type

Specify what the error bars on the Means Plots represent.

Possible Choices:

- **None**
  Do not display error bars on the plots. Choose this option if you want to focus only on the means and not the variation of each mean.

- **Standard Error (SE)**
  Error bars represent the standard deviation of the least squares means.

- **Confidence Interval**
  Error bars represent confidence interval limits for the least squares means.

Select Plots – Group Comparisons

Group Comparison Plots

Check this box to display plots depicting the group comparisons. The least squares mean difference is displayed along with the confidence interval for the difference. A separate comparison and plot is generated for each user-entered covariate value, along with a combined plot.

The individual comparison plots can be edited using the first plot format button. The combined multiple-covariate-value comparison plots can be edited using the second plot format button.

Comparison plots allow you to quickly see which comparisons are significant. A reference line is displayed at Difference = 0. Any comparison for which the 100(1 - Alpha)% Confidence Interval for the difference does not include 0 is significant at level Alpha.

Select Plots – Residual Analysis Plots

Histogram

Check this box to display a histogram and/or density trace of the residuals.

The histogram lets you evaluate the assumption that the residuals are normally (bell-shaped) distributed with a mean of zero and a constant variance.

Normality

Evaluate whether the distribution of the residuals is normal (bell-shaped).

Outliers

Evaluate whether there are outliers.

Anomalies

Look for other anomalies in the residuals that should be considered.

Probability Plot

Check this box to display a normal probability plot of the residuals.

If the residuals are normally distributed, the data points of the normal probability plot will fall along a straight line through the origin with a slope of 1.0. Major deviations from this ideal picture reflect departures from normality. Stragglers at either end of the normal probability plot indicate outliers, curvature at both ends of the plot indicates long or short distributional tails, convex or concave curvature indicates a lack of symmetry, and gaps or plateaus or segmentation in the normal probability plot may require a closer examination of the data or model. Of course, use of this graphic tool with very small sample sizes is not recommended.

If the residuals are not normally distributed, then the t-tests on model coefficients, the F-tests, and any interval estimates are not valid. This is a critical assumption to check.

Residuals vs. Yhat Scatter Plot

Check this box to display a plot of the residuals versus the predicted values (Yhat).

This plot should always be examined. The preferred pattern to look for is a point cloud or a horizontal band. A wedge or bowtie pattern is an indicator of non-constant variance, a violation of a critical model assumption. A sloping or curved band signifies inadequate specification of the model. A sloping band with increasing or decreasing variability suggests non-constant variance and inadequate specification of the model.

Residuals vs. X Scatter Plots

Check this box to display a plot of the residuals versus each X variable. (Remember that the group variable is converted to a set of binary numeric variables.) The preferred pattern is a rectangular shape or point cloud which indicates that there is no relationship between this X and the response, Y, that is not accounted for by the term in the model. Other, non-random, patterns may require redefining the model.

Storage Tab

These options let you specify if, and where on the dataset, various statistics are stored.

Warning: Any data already in these variables are replaced by the new data. Be careful not to specify columns that contain important data.

Data Storage Options

Storage Option

This option controls whether the values indicated below are stored on the dataset when the procedure is run.

- **Do not store data**
  No data are stored even if they are checked.

- **Store in empty columns only**
  The values are stored in empty columns only. Columns containing data are not used for data storage, so no data can be lost.

- **Store in designated columns**
  Beginning at the *First Storage Variable*, the values are stored in this column and those to the right. If a column contains data, the data are replaced by the storage values. Care must be used with this option because it cannot be undone.

Store First Item In

The first item is stored in this column. Each additional item that is checked is stored in the columns immediately to the right of this variable.

Leave this value blank if you want the data storage to begin in the first blank column on the right-hand side of the data.

Warning: any existing data in these columns is automatically replaced, so be careful.

Data Storage Options – Select Items to Store

Predicted Y ... Upper C.L. Individual

Indicate whether to store these row-by-row values, beginning at the column indicated by the *Store First Variable In* option.

Example 1 – ANCOVA Model Assuming Unequal Slopes (with Covariate-by-Group Interaction)

This section presents an example of how to run an analysis of the data presented above. These data are contained in the ANCOVA2Grp dataset. In this example, the responses from two states, Iowa and Utah, are compared with an adjustment for the age of the respondent. The two groups will be compared at three different covariate values.

This example will run all reports and plots so that they may be documented.

You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu of the Analysis of Covariance (ANCOVA) with Two Groups window.

1 Open the ANCOVA2Grp dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **ANCOVA2Grp.NCSS**.
- Click **Open**.

2 Open the Analysis of Covariance (ANCOVA) with Two Groups window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Analysis of Covariance (ANCOVA) with Two Groups** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables and the model.

- Select the **Variables tab**.
- Double-click in the **Response Variable(s)** box. This will bring up the variable selection window.
- Select **Response** from the list of variables and then click **Ok**.
- Double-click in the **Group Variable** box. This will bring up the variable selection window.
- Select **State** from the list of variables and then click **Ok**.
- Double-click in the **Covariate Variable** box. This will bring up the variable selection window.
- Select **Age** from the list of variables and then click **Ok**.
- Set **Calc Group Means at** to **12 Mean 18**.
- Under **Model**, leave **Assume Slopes are** set to **Unequal (Include Covariate-by-Group Interaction)**.

4 Specify the reports.

- Select the **Reports tab**.
- Leave all default reports checked.
- **Check all unchecked checkboxes** to output all available reports.

5 Specify the plots.

- Select the **Plots tab**.
- Leave all default plots checked.
- **Check all unchecked checkboxes** to output all available plots.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Run Summary

Response Variable	Response
Group Variable	State
Reference Group	"Iowa"
Covariate Variable	Age
Slopes Assumed to be	Unequal
Model	Age + State + Age*State

Parameter	Value	Rows	Value
R^2	0.3947	Rows Processed	20
Adj R^2	0.2812	Rows Filtered Out	0
Coefficient of Variation	0.0266	Rows with Response Missing	0
Mean Square Error	7.445498	Rows with Group or Covariate Missing	0
Square Root of MSE	2.728644	Rows Used in Estimation	20
Ave Abs Pct Error	1.711	Completion Status	Normal Completion
Error Degrees of Freedom	16		

This report summarizes the results. It presents the variables used, the model, the number of rows used, and basic summary results.

R^2

R^2 , officially known as the *coefficient of determination*, is defined as

$$R^2 = \frac{SS_{Model}}{SS_{Total(Adjusted)}}$$

R^2 is probably the most popular measure of how well a model fits the data. R^2 may be defined either as a ratio or a percentage. Since we use the ratio form, its values range from zero to one. A value of R^2 near zero indicates no linear relationship, while a value near one indicates a perfect linear fit. Although popular, R^2 should not be used indiscriminately or interpreted without scatter plot support. Following are some qualifications on its interpretation:

1. *Additional independent variables*. It is possible to increase R^2 by adding more independent variables, but the additional independent variables may cause an increase in the mean square error, an unfavorable situation. This usually happens when the sample size is small.
2. *Range of the independent variables*. R^2 is influenced by the range of the independent variables. R^2 increases as the range of the *X's* increases and decreases as the range of the *X's* decreases.
3. *Slope magnitudes*. R^2 does not measure the magnitude of the slopes.
4. *Linearity*. R^2 does not measure the appropriateness of a linear model. It measures the strength of the linear component of the model. Suppose the relationship between *X* and *Y* was a perfect sphere. Although there is a perfect relationship between the variables, the R^2 value would be zero.
5. *Predictability*. A large R^2 does not necessarily mean high predictability, nor does a low R^2 necessarily mean poor predictability.
6. *Sample size*. R^2 is highly sensitive to the number of observations. The smaller the sample size, the larger its value.

Adjusted R^2

This is an adjusted version of R^2. The adjustment seeks to remove the distortion due to a small sample size. The formula for adjusted R^2 is

$$\overline{R}^2 = 1 - \frac{(N-1)\left(1-R^2\right)}{N-p-1}$$

Coefficient of Variation

The coefficient of variation is a relative measure of dispersion, computed by dividing root mean square error by the mean of the response variable. By itself, it has little value, but it can be useful in comparative studies.

$$CV = \frac{\sqrt{MSE}}{\bar{y}}$$

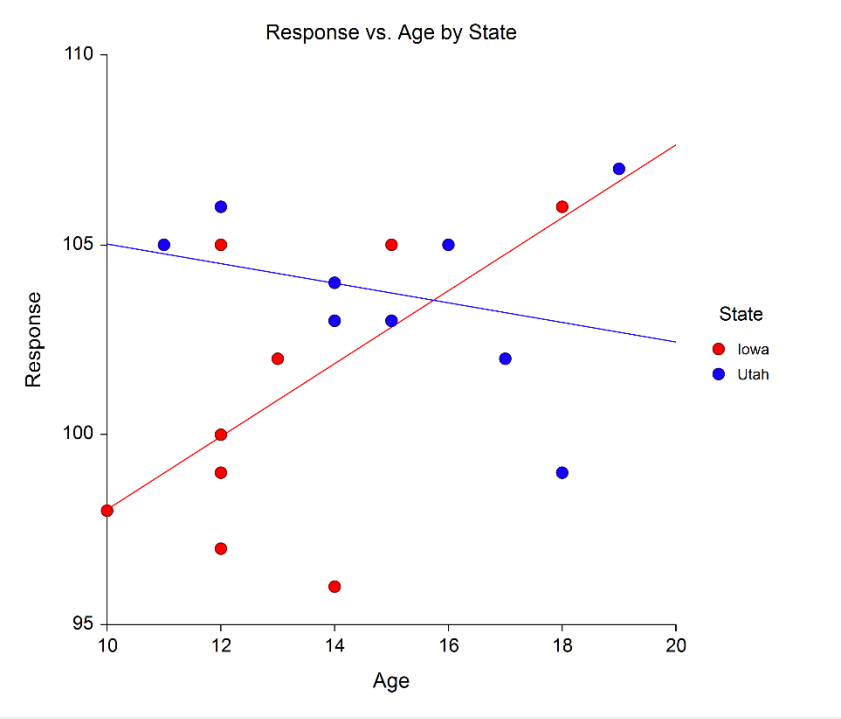
Ave Abs Pct Error

This is the average of the absolute percent errors. It is another measure of the goodness of fit of the linear model to the data. It is calculated using the formula

$$AAPE = \frac{100\sum_{j=1}^{N}\left|\frac{y_j - \hat{y}_j}{y_j}\right|}{N}$$

Note that when the response variable is zero, its predicted value is used in the denominator.

Response vs Covariate by Group Scatter Plot



This is a scatter plot with the response variable on the Y-axis, the covariate variable, Age, on the X-axis, and the group variable, State, in the legend. The slopes appear to be quite different, one positive and the other negative.

Descriptive Statistics

Variable	Count	Mean	Standard Deviation	Minimum	Maximum
Age	20	14.15	2.497894	10	19
(State="Utah")	20	0.5	0.5129892	0	1
Age*(State="Utah")	20	7.55	7.937088	0	19
Response	20	102.4	3.218368	96	107

For each variable, the count, arithmetic mean, standard deviation, minimum, and maximum are computed. This report is particularly useful for checking that the correct variables were selected. Recall that the group variable with two levels is represents by one binary indicator variable. The reference group is not listed.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-Ratio	P-Value	Significant at 5%?
Model	3	77.67203	25.89068	3.477	0.0408	Yes
Age	1	12.10835	12.10835	1.626	0.2204	No
State	1	44.73146	44.73146	6.008	0.0261	Yes
Age*State	1	36.56446	36.56446	4.911	0.0415	Yes
Error	16	119.128	7.445498			
Total(Adjusted)	19	196.8	10.35789			

An analysis of variance (ANOVA) table summarizes the information related to the variation in data. The Age*State interaction is significant, indicating that the slopes are not equal. This being the case, you may want to perform comparisons at several covariate values.

Source

This represents a partition of the variation in *Y*.

DF

The degrees of freedom are the number of dimensions associated with this term. Note that each observation can be interpreted as a dimension in *n*-dimensional space. The degrees of freedom for the intercept, model, error, and adjusted total are 1, *p*, *n*-*p*-1, and *n*-1, respectively.

Sum of Squares

These are the sums of squares associated with the corresponding sources of variation. Note that these values are in terms of the dependent variable. The formulas for each are

$$SS_{Model} = \Sigma\left(\hat{y}_j - \bar{y}\right)^2$$

$$SS_{Error} = \Sigma\left(y_j - \hat{y}_j\right)^2$$

$$SS_{Total} = \Sigma\left(y_j - \bar{y}\right)^2$$

Mean Square

The mean square is the sum of squares divided by the degrees of freedom. This mean square is an estimated variance. For example, the mean square error is the estimated variance of the residuals.

F-Ratio

This is the *F*-statistic for testing the null hypothesis that all $\beta_j = 0$. This *F*-statistic has *p* degrees of freedom for the numerator variance and *n*-*p*-1 degrees of freedom for the denominator variance.

P-Value

This is the *p*-value for the above *F-test*. The *p*-value is the probability that the test statistic will take on a value at least as extreme as the observed value, if the null hypothesis is true. If the *p*-value is less than α, say 0.05, the null hypothesis is rejected. If the *p*-value is greater than α, then the null hypothesis is accepted.

Significant at [5%]?

This is the decision based on the *p*-value and the user-entered Tests Alpha value. The default is Tests Alpha = 0.05.

Model Coefficient T-Tests

Independent Variable	Model Coefficient b(i)	Standard Error Sb(i)	T-Statistic to Test H0: β(i)=0	P-Value	Reject H0 at 5%?
Intercept	88.44495	5.52261	16.015	0.0000	Yes
Age	0.9587156	0.4132412	2.320	0.0339	Yes
(State="Utah")	19.1561	7.815332	2.451	0.0261	Yes
Age*(State="Utah")	-1.217064	0.5492	-2.216	0.0415	Yes

This section reports the values and significance tests of the model coefficients.

Independent Variable

The names of the independent variables are listed here. The intercept is the value of the *Y* intercept.

Note that the name may become very long, especially for interaction terms. These long names may misalign the report. You can force the rest of the items to be printed on the next line by using the Stagger label … option on the Report Options tab. This should create a better-looking report when the names are extra-long.

Model Coefficient b(i)

The coefficients are the least squares estimates of the parameters. The value indicates how much change in *Y* occurs for a one-unit change in a particular *X* when the remaining *X's* are held constant. These coefficients are often called partial-regression coefficients since the effect of the other *X's* is removed. These coefficients are the values of $b_0, b_1, \cdots, b_p$.

Standard Error Sb(i)

The standard error of the coefficient, s_{b_j}, is the standard deviation of the estimate. It is used in hypothesis tests and confidence limits.

T-Statistic to Test H0: β(i)=0

This is the t-test value for testing the hypothesis that $\beta_j = 0$ versus the alternative that $\beta_j \neq 0$ after removing the influence of all other *X*'s. This *t*-value has *n*-*p*-1 degrees of freedom.

To test for a value other than zero, use the formula below. There is an easier way to test hypothesized values using confidence limits. See the discussion below under Confidence Limits. The formula for the *t*-test is

$$t_j = \frac{b_j - \beta_j^*}{s_{b_j}}$$

P-Value

This is the *p*-value for the significance test of the coefficient. The *p*-value is the probability that this *t*-statistic will take on a value at least as extreme as the observed value, assuming that the null hypothesis is true (i.e., the coefficient estimate is equal to zero). If the *p*-value is less than alpha, say 0.05, the null hypothesis of equality is rejected. This *p*-value is for a two-tail test.

Reject H0 at [5%]?

This is the decision based on the *p*-value and the user-entered Tests Alpha value. The default is Tests Alpha = 0.05.

Model Coefficient Confidence Intervals

Independent Variable	Model Coefficient b(i)	Standard Error Sb(i)	Lower 95% Conf. Limit of β(i)	Upper 95% Conf. Limit of β(i)
Intercept	88.44495	5.52261	76.73754	100.1524
Age	0.9587156	0.4132412	0.08268333	1.834748
(State="Utah")	19.1561	7.815332	2.588337	35.72387
Age*(State="Utah")	-1.217064	0.5492	-2.381316	-0.05281156

Note: The T-Value used to calculate these confidence limits was 2.120.

This section reports the values and confidence intervals of the model coefficients.

Independent Variable

The names of the independent variables are listed here. The intercept is the value of the *Y* intercept.

Note that the name may become very long, especially for interaction terms. These long names may misalign the report. You can force the rest of the items to be printed on the next line by using the Stagger label … option on the Report Options tab. This should create a better-looking report when the names are extra-long.

Model Coefficient

The coefficients are the least squares estimates of the parameters. The value indicates how much change in *Y* occurs for a one-unit change in a particular *X* when the remaining *X's* are held constant. These coefficients are often called partial-regression coefficients since the effect of the other *X's* is removed. These coefficients are the values of $b_0, b_1, \cdots, b_p$.

Standard Error

The standard error of the coefficient, s_{b_j}, is the standard deviation of the estimate. It is used in hypothesis tests and confidence limits.

Lower and Upper 95% Conf. Limit of β(i)

These are the lower and upper values of a $100(1-\alpha)\%$ interval estimate for β_j based on a *t*-distribution with *n*-*p*-1 degrees of freedom. This interval estimate assumes that the residuals for the regression model are normally distributed.

These confidence limits may be used for significance testing values of β_j other than zero. If a specific value is not within this interval, it is significantly different from that value. Note that these confidence limits are set up as if you are interested in each regression coefficient separately.

The formulas for the lower and upper confidence limits are:

$$b_j \pm t_{1-\alpha/2,n-p-1}\ s_{b_j}$$

Note: The T-Value …

This is the value of $t_{1-\alpha/2,n-p-1}$ used to construct the confidence limits.

Least Squares Means

Error Degrees of Freedom (DF): 16
Means Calculated at: Covariate Values 1 to 3 (See below)

Name	Count	Least Squares Mean	Standard Error	Lower 95% Conf. Limit for Mean	Upper 95% Conf. Limit for Mean
Covariate Value 1: Age = 12					
Intercept					
All	20	102.2252	0.8649418	100.3916	104.0588
State					
Iowa	10	99.94954	0.9952167	97.83978	102.0593
Utah	10	104.5009	1.414935	101.5014	107.5004
Covariate Value 2: Age = 14.15 (Mean)					
Intercept					
All	20	102.9781	0.6635722	101.5714	104.3848
State					
Iowa	10	102.0108	0.9479811	100.0012	104.0204
Utah	10	103.9454	0.9287863	101.9765	105.9144
Covariate Value 3: Age = 18					
Intercept					
All	20	104.3263	1.277113	101.619	107.0337
State					
Iowa	10	105.7018	2.163112	101.1162	110.2874
Utah	10	102.9508	1.358314	100.0713	105.8303

Note: When the model includes a significant covariate-by-group interaction, you may want to calculate and compare means at various values of the covariate and consider the results collectively. If you calculate and compare means at only one covariate value, the results may be misleading

This section reports the least squares means and associated confidence intervals. In this example, the least squares means are calculated at Age = 12, 14.15 (the mean), and 18. The results are based on *n*-*p*-1 = 16 degrees of freedom for error.

Name

The name of the group variable and its individual group names are listed here. The intercept is the value of the *Y* intercept.

Note that the name may become very long, especially for interaction terms. These long names may misalign the report. You can force the rest of the items to be printed on the next line by using the Stagger label … option on the Report Options tab. This should create a better-looking report when the names are extra-long.

Count

This column specifies the number of observations in each group.

Least Squares Mean

This is the least squares mean estimate, $\hat{\mu}_j$. The least squares means are adjusted based on the model. In balanced designs with no covariates, the least squares group means will be equal to the raw group means. In unbalanced designs or when covariates are present, the least squares means usually are different from the raw means.

Standard Error

The standard error of the mean, $\mathrm{SE}(\hat{\mu}_j)$, is the standard deviation of the estimate. It is used in hypothesis tests and confidence limits.

Lower and Upper 95% Conf. Limits for Mean

These are the lower and upper values of a $100(1-\alpha)\%$ interval estimate for the mean, μ_j, based on a *t*-distribution with *n*-*p*-1 degrees of freedom.

The formulas for the lower and upper confidence limits are:

$$\hat{\mu}_j \pm t_{1-\frac{\alpha}{2},n-p-1} \times \mathrm{SE}(\hat{\mu}_j)$$

Least Squares Means with Hypothesis Tests of H0: Mean = 0

Error Degrees of Freedom (DF): 16
Means Calculated at: Covariate Values 1 to 3 (See below)
Hypotheses Tested: H0: Mean = 0 vs. H1: Mean ≠ 0

Name	Count	Least Squares Mean	Standard Error	T-Statistic to Test H0: Mean=0	P-Value	Reject H0 at 5%?
Covariate Value 1: Age = 12						
Intercept						
All	20	102.2252	0.8649418	118.187	0.0000	Yes
State						
Iowa	10	99.94954	0.9952167	100.430	0.0000	Yes
Utah	10	104.5009	1.414935	73.856	0.0000	Yes
Covariate Value 2: Age = 14.15 (Mean)						
Intercept						
All	20	102.9781	0.6635722	155.187	0.0000	Yes
State						
Iowa	10	102.0108	0.9479811	107.608	0.0000	Yes
Utah	10	103.9454	0.9287863	111.915	0.0000	Yes
Covariate Value 3: Age = 18						
Intercept						
All	20	104.3263	1.277113	81.689	0.0000	Yes
State						
Iowa	10	105.7018	2.163112	48.866	0.0000	Yes
Utah	10	102.9508	1.358314	75.793	0.0000	Yes

Note: When the model includes a significant covariate-by-group interaction, you may want to calculate and compare means at various values of the covariate and consider the results collectively. If you calculate and compare means at only one covariate value, the results may be misleading.

This section reports the least squares means and associated hypothesis tests. In this example, the least squares means are calculated at Age = 12, 14.15 (the mean), and 18. The results are based on *n*-*p*-1 = 16 degrees of freedom for error.

Name

The name of the group variable and its individual group names are listed here. The intercept is the value of the *Y* intercept.

Note that the name may become very long, especially for interaction terms. These long names may misalign the report. You can force the rest of the items to be printed on the next line by using the Stagger label … option on the Report Options tab. This should create a better-looking report when the names are extra-long.

Count

This column specifies the number of observations in each group.

Least Squares Mean

This is the least squares mean estimate, $\hat{\mu}_j$. The least squares means are adjusted based on the model. In balanced designs with no covariates, the least squares group means will be equal to the raw group means. In unbalanced designs or when covariates are present, the least squares means usually are different from the raw means.

Standard Error

The standard error of the mean, $\text{SE}(\hat{\mu}_j)$, is the standard deviation of the estimate. It is used in hypothesis tests and confidence limits.

T-Statistic to Test H0: Mean=0

This is the t-test value for testing the hypothesis that the mean is equal to 0 versus the alternative that it is not equal to 0. This *t*-value has *n*-*p*-1 degrees of freedom and is calculated as

$$t_j = \frac{\hat{\mu}_j}{\text{SE}(\hat{\mu}_j)}$$

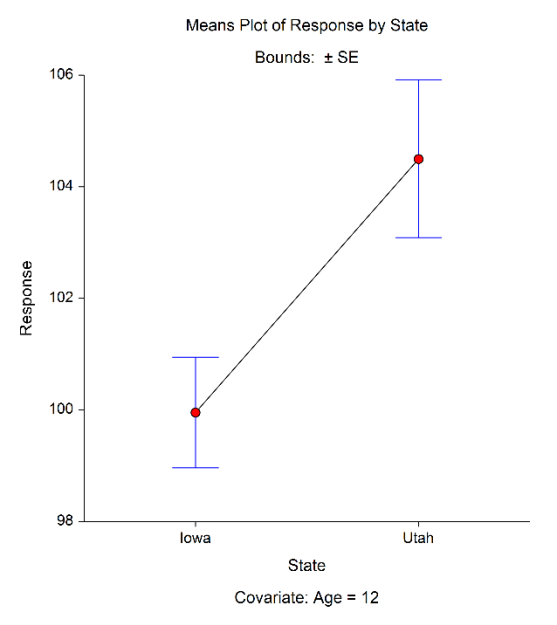
P-Value

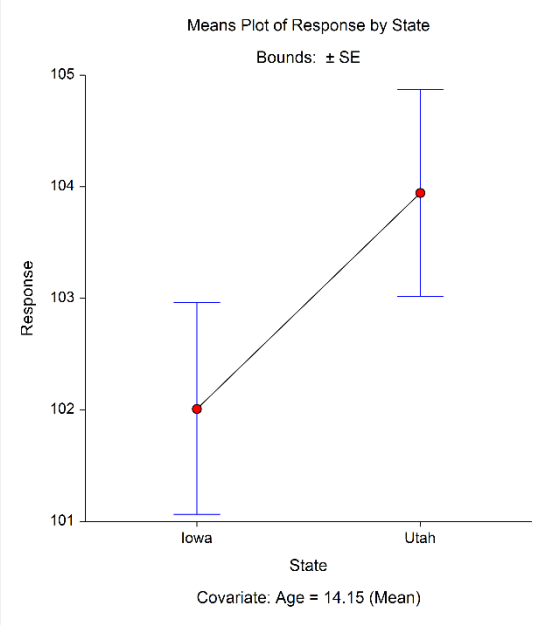
This is the *p*-value for the significance test of the mean. The *p*-value is the probability that this *t*-statistic will take on a value at least as extreme as the observed value, if the null hypothesis is true (i.e., the mean estimate is equal to zero). If the *p*-value is less than alpha, say 0.05, the null hypothesis of equality is rejected. This *p*-value is for a two-tail test.

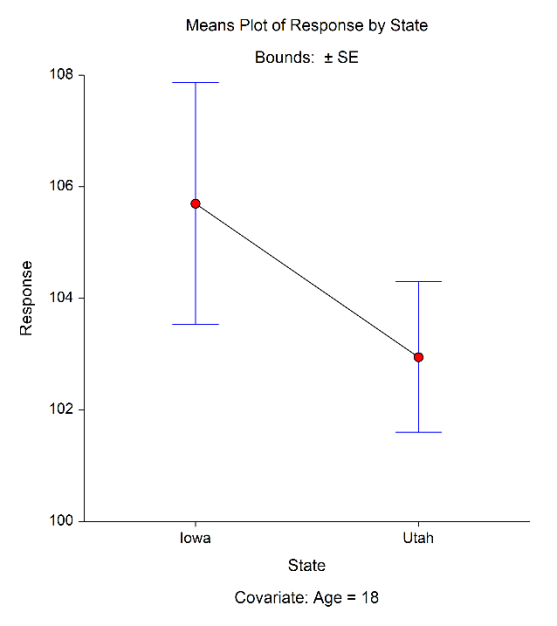
Reject H0 at [5%]?

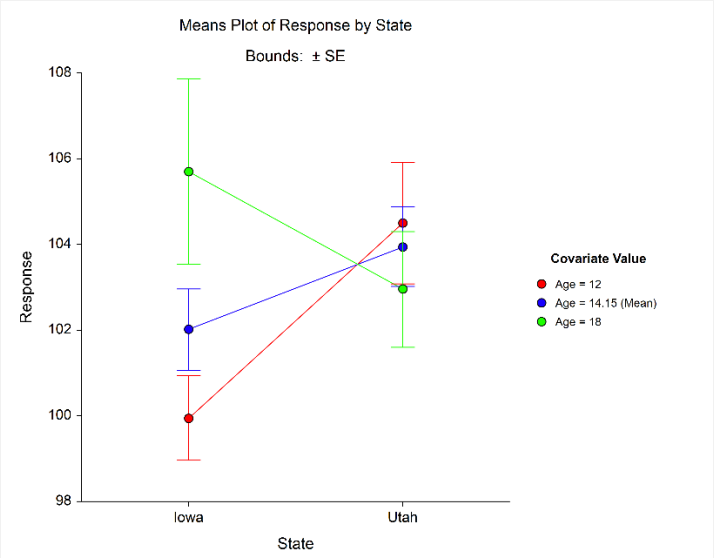
This is the decision based on the *p*-value and the user-entered Tests Alpha value. The default is Tests Alpha = 0.05.

Means Plots









The means plots displays the least squares means along with user-selected variability lines, in this case ±SE. Individual plots and a combined plot are created for the various covariate values evaluated. Notice how the means change quite drastically depending on the covariate value used for the evaluation. This happens because the covariate-by-group interaction term is significant.

T-Tests for Group Least Squares Mean Differences

Error Degrees of Freedom (DF): 16
Means Calculated at: Covariate Values 1 to 3 (See below)
Hypotheses Tested: H0: Diff = 0 vs. H1: Diff ≠ 0

Comparison	Least Squares Mean Difference	Standard Error	T-Statistic to Test H0: Diff=0	P-Value	Reject H0 at 5%?
Covariate Value 1: Age = 12					
State					
Utah - Iowa	4.551337	1.729884	2.631	0.0182	Yes
Covariate Value 2: Age = 14.15 (Mean)					
State					
Utah - Iowa	1.934651	1.327144	1.458	0.1643	No
Covariate Value 3: Age = 18					
State					
Utah - Iowa	-2.751044	2.554226	-1.077	0.2974	No

Note: When the model includes a significant covariate-by-group interaction, you may want to calculate and compare means at various values of the covariate and consider the results collectively. If you calculate and compare means at only one covariate value, the results may be misleading.

This section reports the least squares mean difference and associated hypothesis test for each covariate value. In this example, the least squares means are calculated at Age = 12, 14.15 (the mean), and 18. These tests indicate that the difference is only significant for Age = 12. The results are based on n-p-1 = 16 degrees of freedom for error.

Confidence Intervals for Group Least Squares Mean Differences

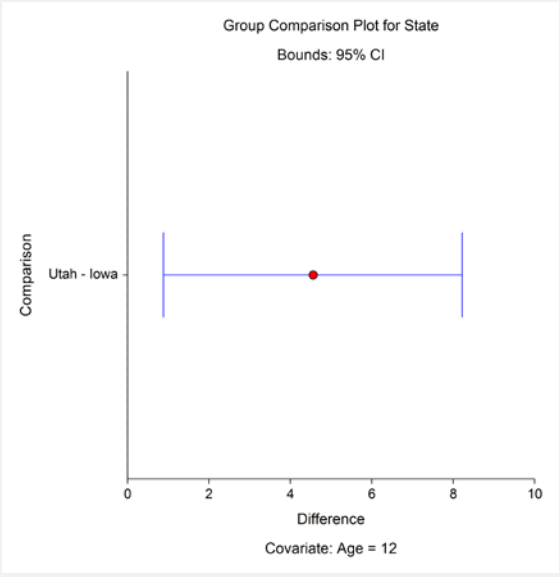
Error Degrees of Freedom (DF): 16
Means Calculated at: Covariate Values 1 to 3 (See below)

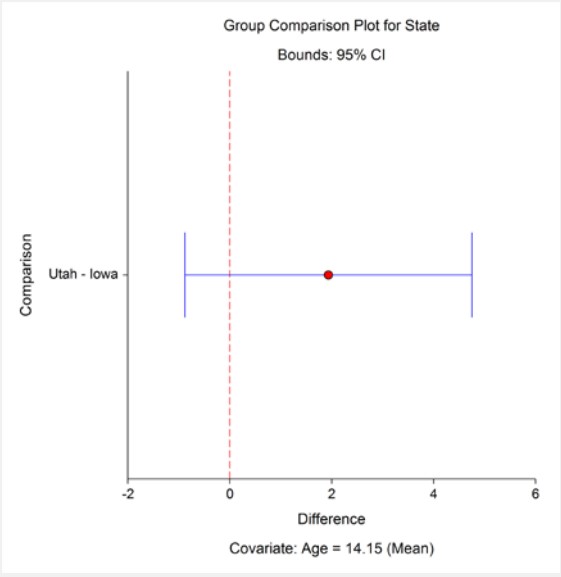
Comparison	Least Squares Mean Difference	Standard Error	Lower 95% Conf. Limit for Difference	Upper 95% Conf. Limit for Difference
Covariate Value 1: Age = 12				
State				
Utah - Iowa	4.551337	1.729884	0.8841482	8.218527
Covariate Value 2: Age = 14.15 (Mean)				
State				
Utah - Iowa	1.934651	1.327144	-0.8787697	4.748071
Covariate Value 3: Age = 18				
State				
Utah - Iowa	-2.751044	2.554226	-8.165761	2.663673

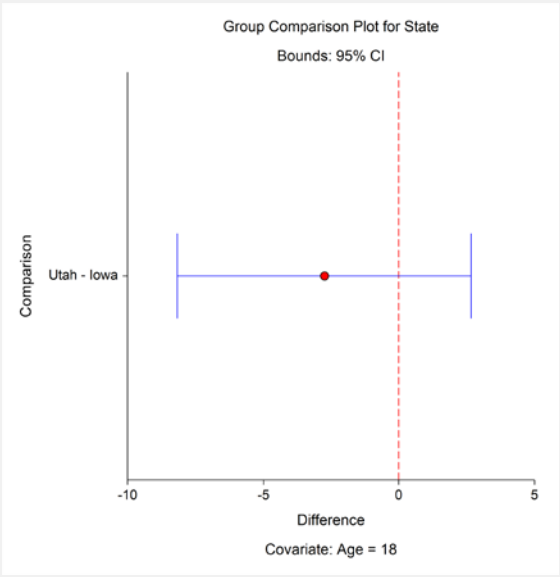
Note: When the model includes a significant covariate-by-group interaction, you may want to calculate and compare means at various values of the covariate and consider the results collectively. If you calculate and compare means at only one covariate value, the results may be misleading.

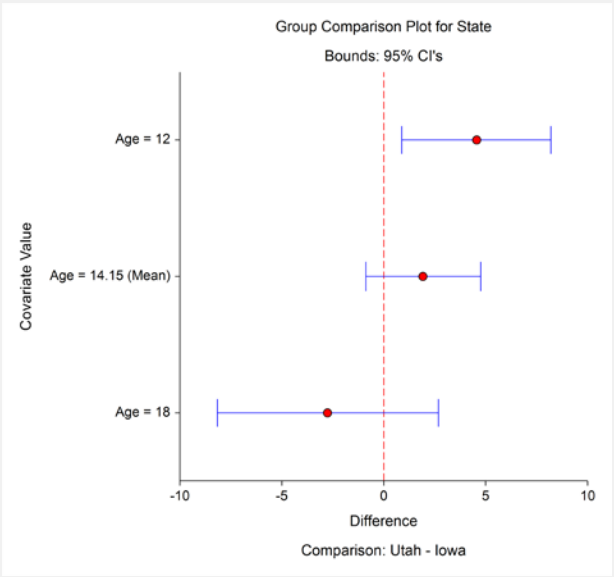
This section reports the least squares mean difference and associated confidence interval for each covariate value. In this example, the least squares means are calculated at Age = 12, 14.15 (the mean), and 18. These intervals indicate that the difference is only significant for Age = 12. The results are based on n-p-1 = 16 degrees of freedom for error.

Group Comparison Plots









These comparison plots display the mean differences along with 95% confidence intervals. Comparisons for which the interval does not contain zero are significant (i.e. for Age = 12).

Residual Normality Assumption Tests

Test Name	Test Statistic	P-Value	Reject Residual Normality at 20%?
Shapiro-Wilk	0.956	0.4700	No
Anderson-Darling	0.486	0.2264	No
D'Agostino Skewness	-0.499	0.6180	No
D'Agostino Kurtosis	1.205	0.2282	No
D'Agostino Omnibus	1.701	0.4272	No

This report gives the results of applying several normality tests to the residuals. The Shapiro-Wilk test is probably the most popular, so it is given first. These tests are discussed in detail in the Normality Tests section of the Descriptive Statistics procedure.

Graphic Residual Analysis

The residuals can be graphically analyzed in numerous ways. You should examine all of the basic residual graphs: the histogram, the density trace, the normal probability plot, the scatter plot of the residuals versus the predicted value of the dependent variable, and the scatter plot of the residuals versus each of the independent variables.

For the basic scatter plots of residuals versus either the predicted values of *Y* or the independent variables, Hoaglin (1983) explains that there are several patterns to look for. You should note that these patterns are very difficult, if not impossible, to recognize for small data sets.

Point Cloud

A point cloud, basically in the shape of a rectangle or a horizontal band, would indicate no relationship between the residuals and the variable plotted against them. This is the preferred condition.

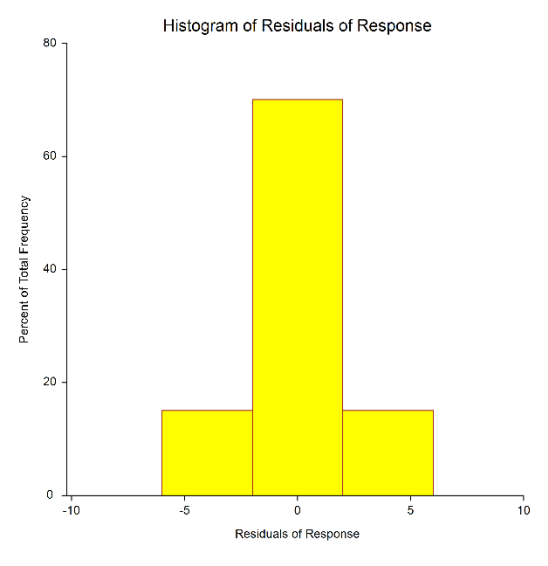
Wedge

An increasing or decreasing wedge would be evidence that there is increasing or decreasing (non-constant) variation. A transformation of *Y* may correct the problem.

Bowtie

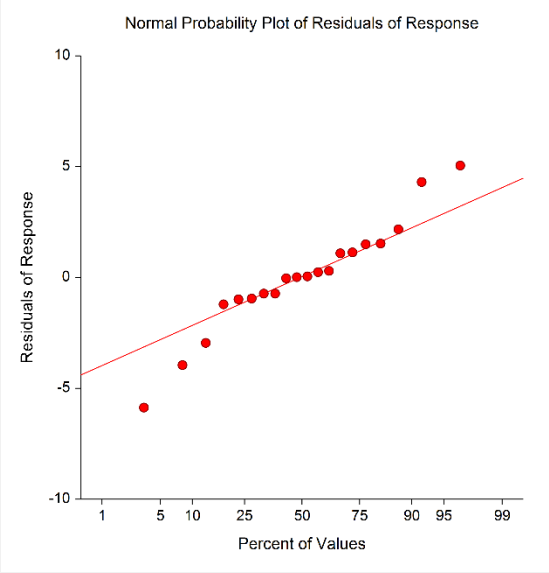
This is similar to the wedge above in that the residual plot shows a decreasing wedge in one direction while simultaneously having an increasing wedge in the other direction. A transformation of *Y* may correct the problem.

Histogram of Residuals



The purpose of the histogram and density trace of the residuals is to evaluate whether they are normally distributed. Unless you have a large sample size, it is best not to rely on the histogram for visually evaluating normality of the residuals. The better choice would be the normal probability plot.

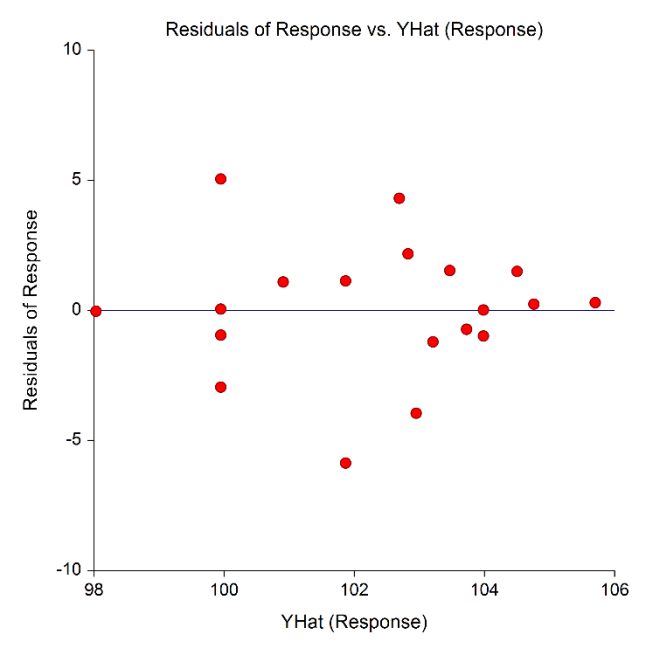
Probability Plot of Residuals



If the residuals are normally distributed, the data points of the normal probability plot will fall along a straight line through the origin with a slope of 1.0. Major deviations from this ideal picture reflect departures from normality. Stragglers at either end of the normal probability plot indicate outliers, curvature at both ends of the plot indicates long or short distributional tails, convex or concave curvature indicates a lack of symmetry, and gaps or plateaus or segmentation in the normal probability plot may require a closer examination of the data or model. Of course, use of this graphic tool with very small sample sizes is not recommended.

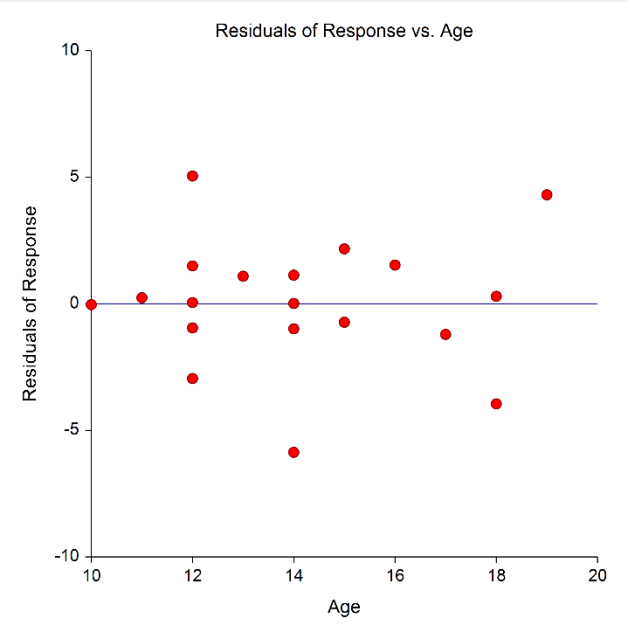
If the residuals are not normally distributed, then the t-tests on regression coefficients, the F-tests, and any interval estimates are not valid. This is a critical assumption to check.

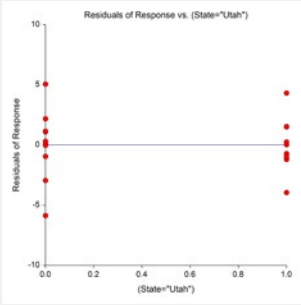
Residuals vs Yhat (Predicted) Plot

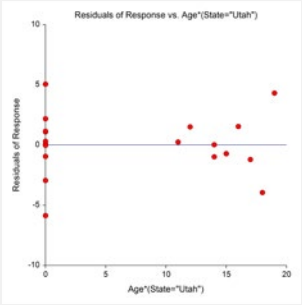


This plot should always be examined. The preferred pattern to look for is a point cloud or a horizontal band. A wedge or bowtie pattern is an indicator of non-constant variance, a violation of a critical assumption. The sloping or curved band signifies inadequate specification of the model. The sloping band with increasing or decreasing variability suggests non-constant variance and inadequate specification of the model.

Residuals vs X Plots







These are scatter plots of the residuals versus each independent variable. Again, the preferred pattern is a rectangular shape or point cloud. Any other nonrandom pattern may require a redefining of the model.

Residuals List Report

Row	Actual Test	Predicted Test	Residual	Absolute Percent Error	Sqrt(MSE) Without This Row
1	100	99.94954	0.05045871	0.050	2.818097
2	102	100.9083	1.091743	1.070	2.802407
3	97	99.94954	-2.949541	3.041	2.696829
4	96	101.867	-5.866972	6.111	2.312976
5	105	102.8257	2.174312	2.071	2.749573
6	106	105.7018	0.2981651	0.281	2.8153
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

This section reports on the sample residuals, or e_i's.

Actual

This is the actual value of *Y*.

Predicted

The predicted value of *Y* using the values of the independent variables given on this row.

Residual

This is the error in the predicted value. It is equal to the *Actual* minus the *Predicted.*

Absolute Percent Error

This is percentage that the absolute value of the *Residual* is of the *Actual* value. Scrutinize rows with the large percent errors.

Sqrt(MSE) Without This Row

This is the value of the square root of the mean square error that is obtained if this row is deleted. A perusal of this statistic for all observations will highlight observations that have an inflationary impact on mean square error and could be outliers.

Predicted Values with Confidence Limits of Means

Row	Actual Test	Predicted Test	Standard Error of Predicted	Lower 95% Conf. Limit of Mean	Upper 95% Conf. Limit of Mean
1	100	99.94954	0.9952167	97.83978	102.0593
2	102	100.9083	0.8668221	99.07068	102.7458
3	97	99.94954	0.9952167	97.83978	102.0593
4	96	101.867	0.9240355	99.9081	103.8258
5	105	102.8257	1.139227	100.4106	105.2407
6	106	105.7018	2.163112	101.1162	110.2874
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

Confidence intervals for the mean response of *Y* given specific levels for the group and covariate variables are provided here. It is important to note that violations of any assumptions will invalidate these interval estimates.

Actual

This is the actual value of *Y*.

Predicted

The predicted value of *Y*. It is predicted using the values of the group and covariate variables for this row. If the input data had both group and covariate values but no value for *Y*, the predicted value is still provided.

Standard Error of Predicted

This is the standard error of the mean response for the specified values of the group and covariate variables. Note that this value is not constant for all variable values. In fact, it is a minimum at the average value of each group and covariate variable.

Lower 95% C.L. of Mean

This is the lower limit of a 95% confidence interval estimate of the mean of *Y* for this observation.

Upper 95% C.L. of Mean

This is the upper limit of a 95% confidence interval estimate of the mean of *Y* for this observation.

Predicted Values with Prediction Limits of Individuals

Row	Actual Test	Predicted Test	Standard Error of Predicted	Lower 95% Pred. Limit of Individual	Upper 95% Pred. Limit of Individual
1	100	99.94954	2.904471	93.79234	106.1067
2	102	100.9083	2.863019	94.83893	106.9776
3	97	99.94954	2.904471	93.79234	106.1067
4	96	101.867	2.880857	95.75983	107.9741
5	105	102.8257	2.956913	96.55731	109.0941
6	106	105.7018	3.482033	98.32026	113.0834
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

A prediction interval for the individual response of *Y* given specific values of the group and covariate variables is provided here for each row.

Actual

This is the actual value of *Y*.

Predicted

The predicted value of *Y*. It is predicted using the values of the group and covariate variables for this row. If the input data had both group and covariate values but no value for *Y*, the predicted value is still provided.

Standard Error of Predicted

This is the standard error of the mean response for the specified values of the group and covariate variables. Note that this value is not constant for all variable values. In fact, it is a minimum at the average value of the group and covariate variable.

Lower 95% Pred. Limit of Individual

This is the lower limit of a 95% prediction interval of the individual value of *Y* for this observation.

Upper 95% Pred. Limit of Individual

This is the upper limit of a 95% prediction interval of the individual value of *Y* for this observation.

Example 2 – ANCOVA Model Assuming Equal Slopes (No Covariate-by-Group Interaction)

In this example, the responses from two states, Iowa and Utah, are compared with an adjustment for the age of the respondent. This section presents an example of how to run an analysis on the same data assuming equal slopes. (Note: You would not normally analyze data assuming equal slopes if the interaction is found to be significant. This is done here for demonstration purposes only.)

You may follow along here by making the appropriate entries or load the completed template **Example 2** by clicking on Open Example Template from the File menu of the Analysis of Covariance (ANCOVA) with Two Groups window.

1 Open the ANCOVA2Grp dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **ANCOVA2Grp.NCSS**.
- Click **Open**.

2 Open the Analysis of Covariance (ANCOVA) with Two Groups window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Analysis of Covariance (ANCOVA) with Two Groups** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables and the model.

- Select the **Variables tab**.
- Double-click in the **Response Variable(s)** box. This will bring up the variable selection window.
- Select **Response** from the list of variables and then click **Ok**.
- Double-click in the **Group Variable** box. This will bring up the variable selection window.
- Select **State** from the list of variables and then click **Ok**.
- Double-click in the **Covariate Variable** box. This will bring up the variable selection window.
- Select **Age** from the list of variables and then click **Ok**.
- Leave **Calc Group Means at** set to **Covariate Mean**.
- Under **Model**, set **Assume Slopes are** to **Equal (No Covariate-by-Group Interaction)**.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Output

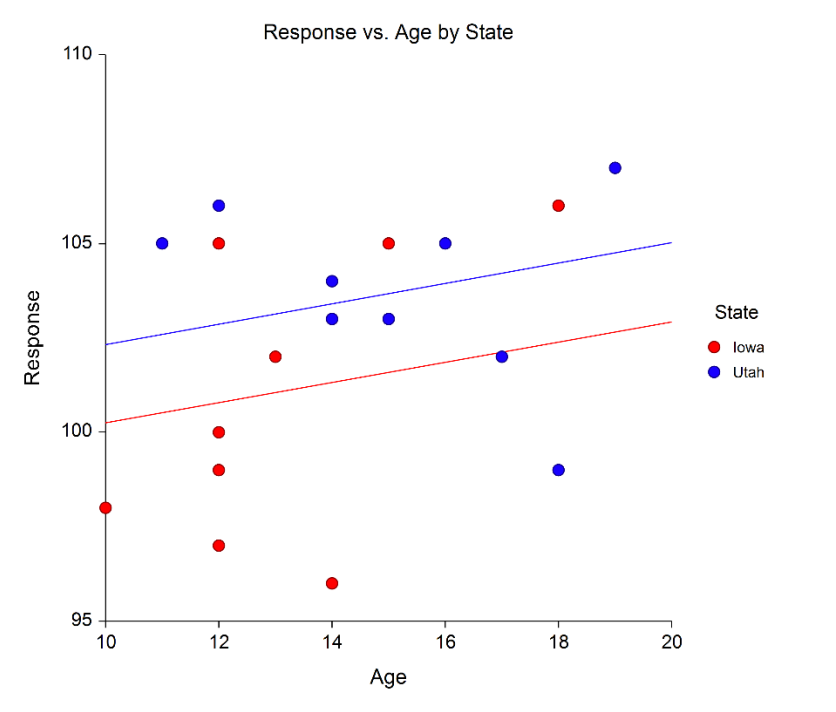
Run Summary

Response Variable	Response
Group Variable	State
Reference Group	"Iowa"
Covariate Variable	Age
Slopes Assumed to be	Equal
Model	Age + State

Parameter	Value	Rows	Value
R^2	0.2089	Rows Processed	20
Adj R^2	0.1158	Rows Filtered Out	0
Coefficient of Variation	0.0296	Rows with Response Missing	0
Mean Square Error	9.158379	Rows with Group or Covariate Missing	0
Square Root of MSE	3.026281	Rows Used in Estimation	20
Ave Abs Pct Error	2.293	Completion Status	Normal Completion
Error Degrees of Freedom	17		

Analysis of Covariance (ANCOVA) with Two Groups

Response vs Covariate by Group Scatter Plot



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-Ratio	P-Value	Significant at 5%?
Model	2	41.10756	20.55378	2.244	0.1365	No
Age	1	7.307562	7.307562	0.798	0.3842	No
State	1	18.47374	18.47374	2.017	0.1736	No
Error	17	155.6924	9.158379			
Total(Adjusted)	19	196.8	10.35789			

Model Coefficient T-Tests

Independent Variable	Model Coefficient b(i)	Standard Error Sb(i)	T-Statistic to Test H0: β(i)=0	P-Value	Reject H0 at 5%?
Intercept	97.5406	4.098049	23.802	0.0000	Yes
Age	0.2696517	0.3018744	0.893	0.3842	No
(State="Utah")	2.087662	1.469914	1.420	0.1736	No

Least Squares Means

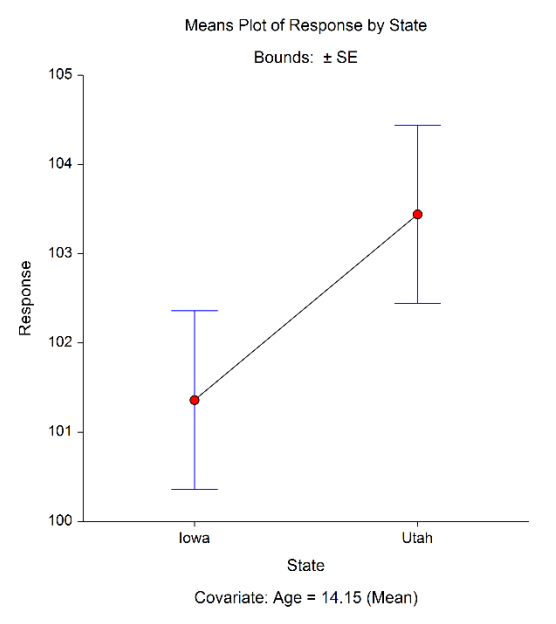
Error Degrees of Freedom (DF): 17
Means Calculated at: Age = 14.15 (Mean)

Name	Count	Least Squares Mean	Standard Error	Lower 95% Conf. Limit for Mean	Upper 95% Conf. Limit for Mean
Intercept					
All	20	102.4	0.6766971	100.9723	103.8277
State					
Iowa	10	101.3562	0.9990401	99.24838	103.464
Utah	10	103.4438	0.9990401	101.336	105.5516

Note: These results assume that the slopes for both groups are equal (i.e. the covariate-by-group interaction is not significant). To check this assumption, run the model with unequal slopes that includes the covariate-by-group interaction and review the test of the interaction term in the Analysis of Variance report.

Analysis of Covariance (ANCOVA) with Two Groups

Means Plots



T-Tests for Group Least Squares Mean Differences

Error Degrees of Freedom (DF): 17
Means Calculated at: Age = 14.15 (Mean)
Hypotheses Tested: H0: Diff = 0 vs. H1: Diff ≠ 0

Comparison	Least Squares Mean Difference	Standard Error	T-Statistic to Test H0: Diff=0	P-Value	Reject H0 at 5%?
State					
Utah - Iowa	2.087662	1.469914	1.420	0.1736	No

Note: These results assume that the slopes for both groups are equal (i.e. the covariate-by-group interaction is not significant). To check this assumption, run the model with unequal slopes that includes the covariate-by-group interaction and review the test of the interaction term in the Analysis of Variance report.

Confidence Intervals for Group Least Squares Mean Differences

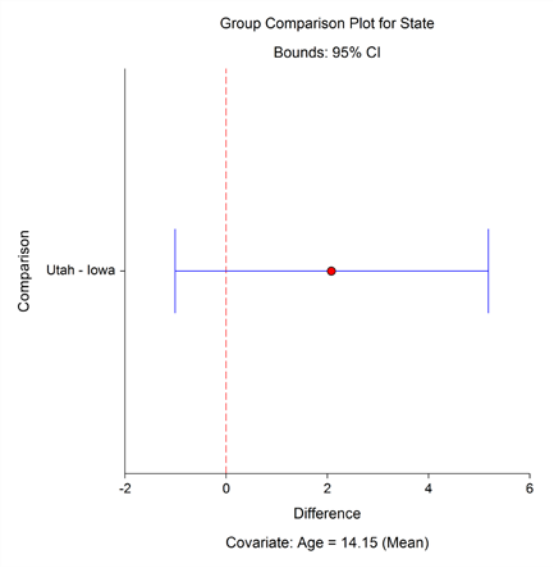
Error Degrees of Freedom (DF): 17
Means Calculated at: Age = 14.15 (Mean)

Comparison	Least Squares Mean Difference	Standard Error	Lower 95% Conf. Limit for Difference	Upper 95% Conf. Limit for Difference
State				
Utah - Iowa	2.087662	1.469914	-1.013587	5.18891

Note: These results assume that the slopes for both groups are equal (i.e. the covariate-by-group interaction is not significant). To check this assumption, run the model with unequal slopes that includes the covariate-by-group interaction and review the test of the interaction term in the Analysis of Variance report.

Analysis of Covariance (ANCOVA) with Two Groups

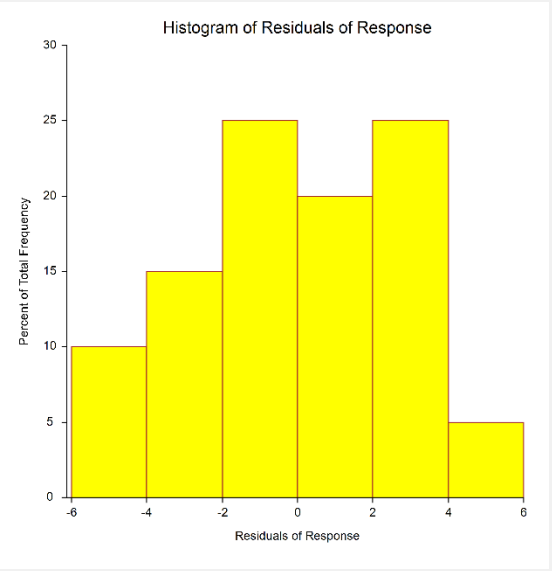
Group Comparison Plots

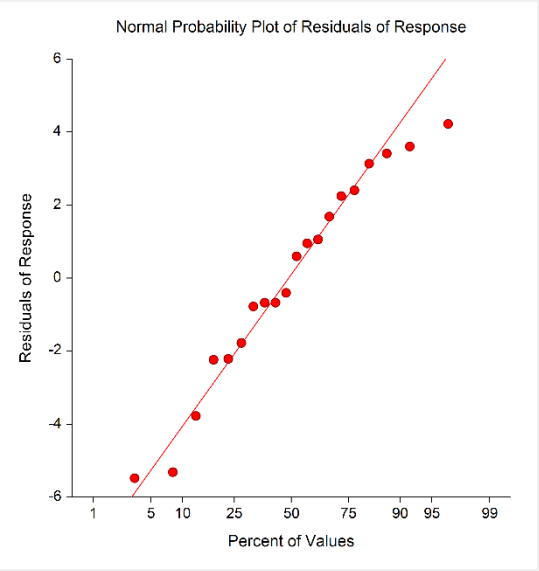


Residual Normality Assumption Tests

Test Name	Test Statistic	P-Value	Reject Residual Normality at 20%?
Shapiro-Wilk	0.955	0.4523	No
Anderson-Darling	0.277	0.6560	No
D'Agostino Skewness	-0.878	0.3797	No
D'Agostino Kurtosis	-0.509	0.6104	No
D'Agostino Omnibus	1.031	0.5971	No

Residual Analysis Plots





The scatter plot shows the two regression lines with equal slopes. The lines don't appear to fit the data very well as is expected when the interaction is significant. Neither State nor Age is significant in these results.

Example 3 – ANCOVA Model with One-Sided Multiple Comparison Tests and Confidence Intervals

In this example, we'll run the same analysis as in Example 1, except that we will perform one-sided tests and create one-sided confidence intervals for the comparisons. Only reports that are different from Example 1 will be shown.

You may follow along here by making the appropriate entries or load the completed template **Example 3** by clicking on Open Example Template from the File menu of the Analysis of Covariance (ANCOVA) with Two Groups window.

1 Open the ANCOVA2Grp dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **ANCOVA2Grp.NCSS**.
- Click **Open**.

2 Open the Analysis of Covariance (ANCOVA) with Two Groups window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Analysis of Covariance (ANCOVA) with Two Groups** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables and the model.

- Select the **Variables tab**.
- Double-click in the **Response Variable(s)** box. This will bring up the variable selection window.
- Select **Response** from the list of variables and then click **Ok**.
- Double-click in the **Group Variable** box. This will bring up the variable selection window.
- Select **State** from the list of variables and then click **Ok**.
- Double-click in the **Covariate Variable** box. This will bring up the variable selection window.
- Select **Age** from the list of variables and then click **Ok**.
- Set **Calc Group Means at** to **12 Mean 18**.
- Under **Model**, leave **Assume Slopes are** set to **Unequal (Include Covariate-by-Group Interaction)**.

4 Specify the reports.

- Select the **Reports tab**.
- Set **Hypothesis Test and Confidence Interval Direction** to **One-Sided Upper**.
- Check only **T-Tests for Group Least Squares Mean Differences** and **Show Confidence Intervals**.

5 Specify the plots.

- Select the **Plots tab**.
- Check only **Group Comparison Plots**.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Output

T-Tests for Group Least Squares Mean Differences

Error Degrees of Freedom (DF): 16
Means Calculated at: Covariate Values 1 to 3 (See below)
Hypotheses Tested: H0: Diff ≤ 0 vs. H1: Diff > 0

Comparison	Least Squares Mean Difference	Standard Error	T-Statistic to Test H0: Diff=0	P-Value	Reject H0 at 5%?
Covariate Value 1: Age = 12					
State					
Utah - Iowa	4.551337	1.729884	2.631	0.0091	Yes
Covariate Value 2: Age = 14.15 (Mean)					
State					
Utah - Iowa	1.934651	1.327144	1.458	0.0821	No
Covariate Value 3: Age = 18					
State					
Utah - Iowa	-2.751044	2.554226	-1.077	0.8513	No

Note: When the model includes a significant covariate-by-group interaction, you may want to calculate and compare means at various values of the covariate and consider the results collectively. If you calculate and compare means at only one covariate value, the results may be misleading.

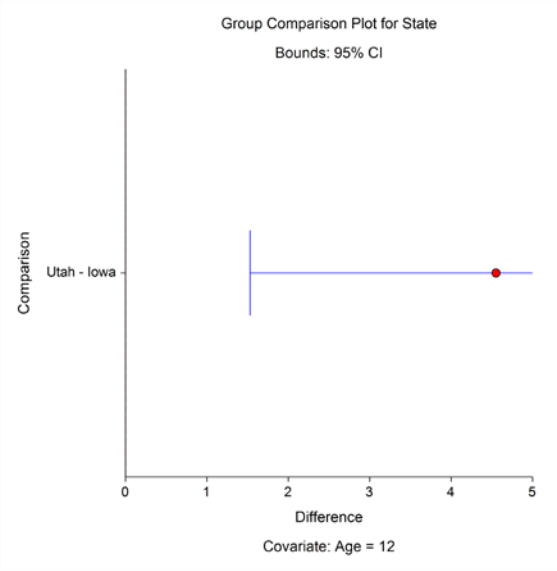
Confidence Intervals for Group Least Squares Mean Differences

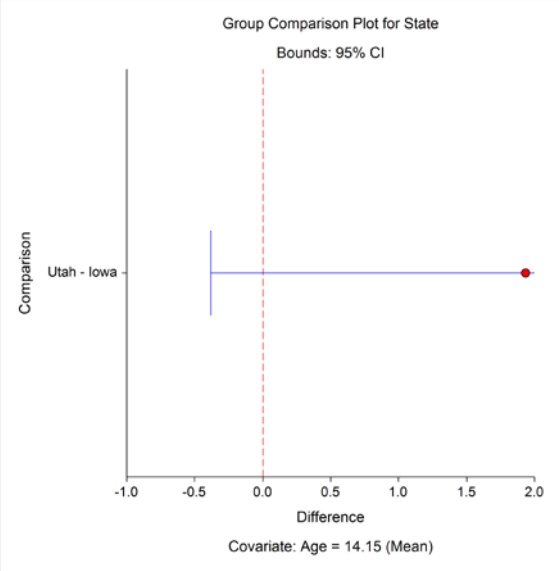
Error Degrees of Freedom (DF): 16
Means Calculated at: Covariate Values 1 to 3 (See below)

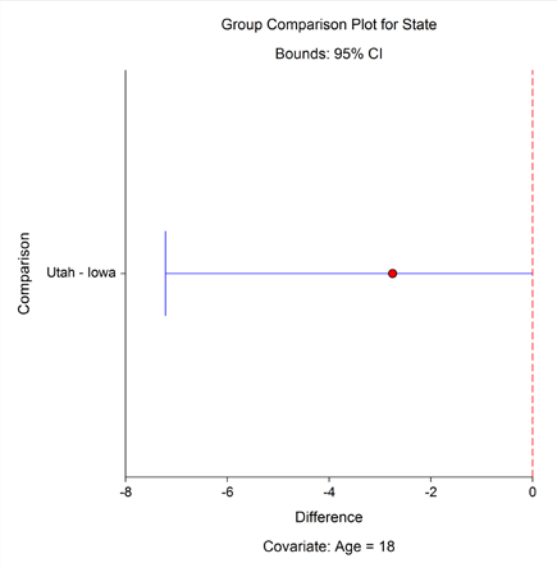
Comparison	Least Squares Mean Difference	Standard Error	Lower 95% Conf. Limit for Difference	Upper 95% Conf. Limit for Difference
Covariate Value 1: Age = 12				
State				
Utah - Iowa	4.551337	1.729884	1.531162	Infinity
Covariate Value 2: Age = 14.15 (Mean)				
State				
Utah - Iowa	1.934651	1.327144	-0.382389	Infinity
Covariate Value 3: Age = 18				
State				
Utah - Iowa	-2.751044	2.554226	-7.210425	Infinity

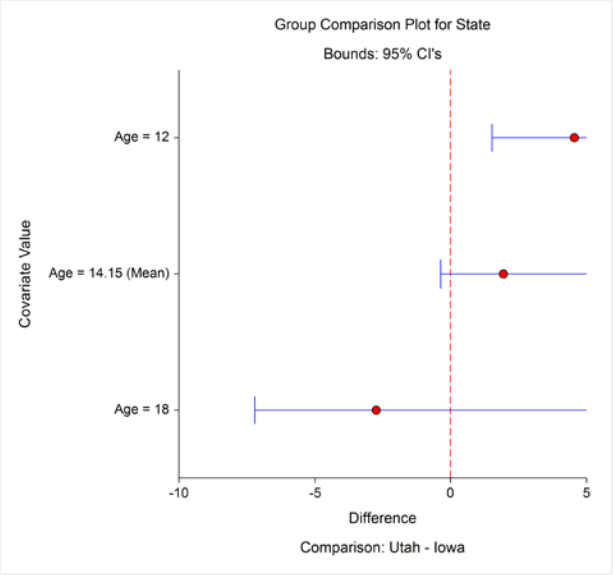
Note: When the model includes a significant covariate-by-group interaction, you may want to calculate and compare means at various values of the covariate and consider the results collectively. If you calculate and compare means at only one covariate value, the results may be misleading.

Group Comparison Plots









These results indicate that for one-sided tests and confidence intervals, only the comparison with Age = 12 is significant.

Example 4 – Two-Sample Equal-Variance T-Test

If you run an ANCOVA analysis with two groups and find the covariate to be non-significant, you may want to remove the covariate from the analysis and run a simple two-sample equal-variance T-test. This example will show you how to perform a two-sample T-test using this procedure.

Note: The two-sample T-test options are limited in this procedure. For additional options specifically related to the two-sample T-test scenario, we suggest you use the Two-Sample T-Test procedure in **NCSS** instead.

You may follow along here by making the appropriate entries or load the completed template **Example 4** by clicking on Open Example Template from the File menu of the Analysis of Covariance (ANCOVA) with Two Groups window.

1 Open the ANCOVA2Grp dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **ANCOVA2Grp.NCSS**.
- Click **Open**.

2 Open the Analysis of Covariance (ANCOVA) with Two Groups window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Analysis of Covariance (ANCOVA) with Two Groups** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables and the model.

- Select the **Variables tab**.
- Double-click in the **Response Variable(s)** box. This will bring up the variable selection window.
- Select **Response** from the list of variables and then click **Ok**.
- Double-click in the **Group Variable** box. This will bring up the variable selection window.
- Select **State** from the list of variables and then click **Ok**.
- Leave the **Covariate Variable** box **empty**.

4 Specify the reports.

- Select the **Reports tab**.
- Check only **Run Summary**, **T-Tests for Group Least Squares Mean Differences**, and **Show Confidence Intervals**.

5 Specify the plots.

- Select the **Plots tab**.
- Check only **Group Comparison Plots**.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Output

Response Variable	**Response**
Group Variable	State
Reference Group	"Iowa"
Covariate Variable	[None]
Slopes Assumed to be	Unequal
Model	State

Parameter	**Value**	**Rows**	**Value**
R^2	0.1717	Rows Processed	20
Adj R^2	0.1257	Rows Filtered Out	0
Coefficient of Variation	0.0294	Rows with Response Missing	0
Mean Square Error	9.055555	Rows with Group or Covariate Missing	0
Square Root of MSE	3.009245	Rows Used in Estimation	20
Ave Abs Pct Error	2.359	Completion Status	Normal Completion
Error Degrees of Freedom	18		

T-Tests for Group Least Squares Mean Differences

Error Degrees of Freedom (DF): 18
Hypotheses Tested: H0: Diff = 0 vs. H1: Diff ≠ 0

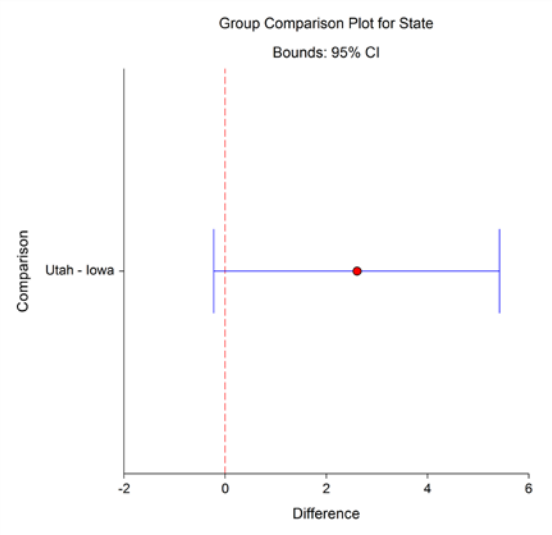
Comparison	Least Squares Mean Difference	Standard Error	T-Statistic to Test H0: Diff=0	P-Value	Reject H0 at 5%?
State					
Utah - Iowa	2.6	1.345775	1.932	0.0693	No

Confidence Intervals for Group Least Squares Mean Differences

Error Degrees of Freedom (DF): 18

Comparison	Least Squares Mean Difference	Standard Error	Lower 95% Conf. Limit for Difference	Upper 95% Conf. Limit for Difference
State				
Utah - Iowa	2.6	1.345775	-0.227369	5.427369

Group Comparison Plots



These results are equivalent to those that would be obtained for the equal-variance T-test in the Two-Sample T-Test procedure in **NCSS**.