Chapter 15
Analysis of Variance

Chapter Table of Contents

ASSIGNING MEASUREMENT LEVELS .................................. 238
CREATING THE ANALYSIS OF VARIANCE .......................... 240
  Model Information .................................................. 242
  Summary of Fit ...................................................... 242
  Analysis of Variance ................................................. 243
  Type III Tests ....................................................... 243
  Parameter Estimates ............................................... 244
  Residuals-by-Predicted Plot ...................................... 245
EXAMINING THE MEANS .................................................. 246
REFERENCES .............................................................. 250
Chapter 15. Analysis of Variance
Chapter 15
Analysis of Variance

In this chapter, you consider analyses that use least-squares methods to fit the general linear model. Such analyses include regression, analysis of variance, and analysis of covariance. You can choose Analyze:Fit (Y X) to carry out an analysis of variance. You can use box plots to examine individual group means.

Figure 15.1. Fit Window with Analysis of Variance Results
Assigning Measurement Levels

The DRUG data set contains results of an experiment carried out to evaluate the effect of four drugs (DRUG) with three experimentally induced diseases (DISEASE). Each drug by disease combination was applied to six randomly selected dogs. The response variable is the increase in systolic blood pressure (CHANG_BP) due to the drug treatment. DRUG and DISEASE are classification or class variables; that is, variables that identify distinct levels or groups. DRUG contains four levels or classes and DISEASE contains three.

Open the DRUG data set.

Figure 15.2. Data Window

A variable’s measurement level determines the way it is treated in analyses. In the data window, measurement levels appear above the variable names, in the upper right portion of the column header. SAS/INSIGHT software supports two measurement levels: interval (Int) and nominal (Nom).

Interval variables contain values that vary across a continuous range. In this data set, the change in blood pressure (CHANG_BP) is an interval variable.

Nominal variables contain a discrete set of values. In this data set, both DRUG and DISEASE contain a discrete set of values. However, since these are numeric variables, by default they have interval measurement levels (Int).

You need to assign both these variables the nominal measurement level (Nom) in order to treat them as classification variables. To do so, use the data measurement level pop-up menu.
Click on the Int measurement level indicator for the variable DRUG. This displays a pop-up menu.

![Figure 15.3. Measurement Levels Menu](image)

The radio mark beside Interval shows the current measurement level. Because DRUG is a numeric variable, it can use either an interval or a nominal measurement level.

Choose Nominal in the pop-up menu to change DRUG’s measurement level.

Repeat these steps to change the measurement level for DISEASE.

Check the measurement levels for DRUG and DISEASE in the data window. Both have Nom measurement levels.

![Figure 15.4. Data with Nominal Variables DRUG and DISEASE](image)
Creating the Analysis of Variance

Consider the two-way analysis of variance model Kutner (1974) proposed for these data:

$$\text{CHANG – BP}_{ijk} = \mu + \gamma_i + \tau_j + (\gamma \tau)_{ij} + \epsilon_{ijk}$$

where $\mu$ is the overall mean effect, $\gamma_i$ is the effect of the $i$th level of DRUG, $\tau_j$ is the effect of the $j$th level of DISEASE, $(\gamma \tau)_{ij}$ is the joint effect of the $i$th level of DRUG with the $j$th level of DISEASE, and $\epsilon_{ijk}$ is the random error term for the $k$th observation in the $i$th level of DRUG and $j$th level of DISEASE. The $\epsilon_{ijk}$’s are assumed to be normally distributed and uncorrelated and to have mean 0 and common variance $\sigma^2$.

The effects for DRUG and DISEASE are often referred to as the main effects in the model and the DRUG*DISEASE effect as an interaction effect. The interaction effect enables you to determine whether the level of DRUG affects the change in blood pressure differently for different levels of DISEASE.

To begin the analysis of variance, follow these steps.

1. Choose Analyze:Fit (Y X).
2. Select CHANG_BP in the variables list on the left, then click the Y button. CHANG_BP appears in the Y variables list and is now defined as the response variable.
3. Select DRUG and DISEASE, then click the Expand button. Your variables dialog should now appear, as shown in Figure 15.5.

Figure 15.5. Fit Variables Dialog with Variable Roles Assigned
The **Expand** button provides a convenient way to specify interactions of any order. The degree of expansion is controlled by the value below the **Expand** button. The order 2 is the default, so clicking **Expand** constructs all possible effects from the selected variables up to second-order effects. This adds **DRUG**, **DISEASE**, and **DRUG*DISEASE** to the effects list.

† **Note:** You could have added the same effects by using the **X** and **Cross** buttons, but the **Expand** button is faster. There is also a **Nest** button for specifying nested effects. For more information on the effects buttons, see Chapter 39, “Fit Analyses.”

⇒ **Click the OK button.**

A fit window appears, as shown in Figure 15.6.

You can control which tables and graphs the fit window contains by clicking the **Output** button in the fit variables dialog or by choosing from the **Tables** and **Graphs** menus. By default, the fit window contains tables for model specification, **Nominal Variable Information**, **Parameter Information**, **Model Equation**, **Summary of Fit**, **Analysis of Variance**, **Type III Tests**, and **Parameter Estimates**, as well as a residual-by-predicted plot.

![Figure 15.6. Fit Window - Model Information](image)
Model Information

The first four tables in the fit analysis contain model information. The first table displays the model specification, the response distribution, and the link function. The Nominal Variable Information table shows the levels of the nominal variables. The levels are determined from the formatted values of the nominal variables.

An additional Parameter Information table shows the variable indices for the parameters in the model equation, the $X'X$ matrix, the estimated covariance matrix, and the estimated correlation matrix. The Model Equation table gives the fitted equation for the model.

Summary of Fit

The Summary of Fit table, as shown in Figure 15.7, contains summary statistics. The Mean of Response 19.1667 is the overall mean of CHANG\_BP. The Root MSE 9.7886 is the square root of the mean square error given in the Analysis of Variance table. Root MSE is an estimate of $\sigma$ in the preceding analysis of variance model.

The R-Square value is 0.5250, which means that 52% of the variation in CHANG\_BP is explained by the fitted model. Adj R-Sq is an alternative to R-Square, adjusted for the number of parameters in the model.

Figure 15.7.  Fit Window - Summary of Fit
Analysis of Variance

The **Analysis of Variance** table summarizes the information related to the sources of variation in the data. **Sum of Squares** measures variation present in the data. It is calculated by summing squared deviations. There are three sources of variation: **Model**, **Error**, and **C Total**. The **Model** row in the table corresponds to the variation among class means. The **Error** row corresponds to \( \epsilon \) in the model and represents variation within class means. **C Total** is the total sum of squares corrected for the mean, and it is the sum of **Model** and **Error**. Degrees of Freedom, **DF**, are associated with each sum of squares and are related in the same way. **Mean Square** is the **Sum of Squares** divided by its associated **DF** (Moore and McCabe 1989, p.685).

If the data are normally distributed, the ratio of the **Mean Square** for the **Model** to the **Mean Square** for **Error** is an **F statistic**. This **F** statistic tests the null hypothesis that all the class means are the same against the alternative hypothesis that the means are not all equal. Think of the ratio as a comparison of the variation among class means to variation within class means. The larger the ratio, the more evidence that the means are not the same. The computed **F** statistic (labeled **F Stat**) is **6.0276**. You can use the **p-value** (labeled **Pr > F**) to determine whether to reject the null hypothesis. The **p-value**, also referred to as the **probability value** or **observed significance level**, is the probability of obtaining (by chance alone) an **F** statistic greater than the computed **F** statistic when the null hypothesis is true. The smaller the **p-value**, the stronger the evidence against the null hypothesis.

In this example, the **p-value** is so small that you can clearly reject the null hypothesis and conclude that at least one of the class means is different. At this point, you have demonstrated statistical significance but cannot make statements about which class means are different.

Type III Tests

The **Type III Tests** table is a further breakdown of the variation due to **MODEL**. The **Sum of Squares** and **DF** for **Model** are broken down into terms corresponding to the main effect for **DRUG**, the main effect for **DISEASE**, and the interaction effect for **DRUG*DISEASE**. The sum of squares for each term represents the variation among the means for the different levels of the factors.

The **Type III Tests** table presents the Type III sums of squares associated with the effects in the model. The Type III sum of squares for a particular effect is the amount of variation in the response due to that effect after correcting for all other terms in the model. Type III sums of squares, therefore, do not depend on the order in which the effects are specified in the model. Refer to the chapter on “The Four Types of Estimable Functions,” in the SAS/STAT User’s Guide for a complete discussion of Type I–IV sums of squares.
Chapter 15. Analysis of Variance

F tests are formed from this table in the same fashion that was explained previously in the section “Analysis of Variance.” In this case, there are three null hypotheses being tested: class means are all the same for the main effect DRUG, the main effect DISEASE, and the interaction effect DRUG*DISEASE. Begin by examining the test for the interaction effect since a strong interaction makes the interpretation of main effects difficult if not impossible. The computed F statistic is 1.7406 with a p-value of 0.1271. This gives little evidence for an interaction effect. Now examine the main effects. The computed F statistic for DRUG is 15.8053 with a p-value less than or equal to 0.0001. The computed F statistic for DISEASE is 4.2220 with a p-value of 0.0193. While both effects are significant, the DRUG effect appears to be stronger.

Now you have more information about which means are significantly different. The results of the F test in the Analysis of Variance table indicated only that at least one of the class means is different from the others. Now you know that the difference in means can be associated with the different levels of the main effects, DRUG and DISEASE.

Parameter Estimates

Parameter estimates resulting from analysis of variance models where the effects are all classification variables are different from those observed in a regression model. They represent a non-unique solution to the normal equations, and thus the individual elements in the table are not as easily interpretable as they are in multiple regression. For a complete discussion of parameter estimates involving classification variables, refer to the chapter “Details of the Linear Model: Understanding GLM Concepts,” in SAS System for Linear Models, Third Edition.
Residuals-by-Predicted Plot

It is appropriate to examine the residuals from the fitted model for analysis of variance just as you did with the multiple regression model you fit in Chapter 14, “Multiple Regression.” The residuals-by-predicted graph illustrated in Figure 15.8, along with several other diagnostic plots, are available for examining residuals. Since this topic is discussed in Chapter 14, residual plots are not examined here.

Figure 15.8.  Fit Window - Parameter Estimates
Examining the Means

Before you can interpret the results for the significant main effects you observed in the Type III Tests table, you need to examine the means for the different levels of these effects. Box plots are an excellent tool for displaying means because means and standard deviations for each level of a variable can be placed side-by-side for easy comparison.

Follow these steps to add box plots for each level of DRUG to the Fit(Y X) window.

1. Select an area for the box plot.
   Drag the cursor until you have a rectangle of suitable size.

2. Choose Analyze:Box Plot/Mosaic Plot ( Y ).
3. Select CHANG_BP in the list at the left, then click the Y button.
   This assigns the Y role to this variable.
4. Select DRUG in the list at the left, then click the X button.
   This assigns the X role to this variable and requests a separate box plot for each level of DRUG. Your variables dialog should now appear, as shown in Figure 15.10.
Click the Output button.

The output options dialog shown in Figure 15.11 appears on your display. In this dialog, you can specify options to determine the output produced by the box plot.

Click on Means.

Means displays mean diamonds for all boxes. The central line in the mean diamond marks the mean; the size of the mean diamond is two standard deviations, one above and one below the mean.

Click OK in both dialogs to create the Box Plots.
Chapter 15. Analysis of Variance

Examine the box plot representing the four levels of DRUG. Recall that the central line in each mean diamond marks the mean while the height of the mean diamond shows one standard deviation on either side of the mean. The box and whiskers display percentiles for the data. (See Chapter 4, “Exploring Data in One Dimension,” for a complete description of the parts of the box plot.)

Follow these steps to hide the display of box and whiskers in order to display the means and standard deviations better.

⇒ Click on Observations in the box plot pop-up menu.
This toggles the display of observations and thus turns off the display of the box, whiskers, and individual observations in the box plot.

⇒ Click on Values in the box plot pop-up menu.
This toggles the display of values of the mean for each box plot.
Figure 15.14. Mean Diamonds for DRUG

The largest effect noted in these plots is that drugs 1 and 2 have a higher average increase in systolic blood pressure than drugs 3 and 4 (averaged over all three levels of DISEASE). This difference resulted in the significant main effect for DRUG that was observed in the Type III Tests table.

Repeat the preceding steps and display box plots for the levels of DISEASE.

Figure 15.15. Mean Diamonds for DISEASE

The differences between the three DISEASE levels are not as pronounced as those observed for DRUG. Disease 3 is associated with a lower average increase in systolic blood pressure than the other two diseases (averaged over all four levels of DRUG).
The smaller $p$-value observed for the DRUG main effect is more evidence that the mean differences for DISEASE are not as pronounced as those for DRUG.

This example illustrates one way to use Analyze:Fit to fit the general linear model. Turn to the next chapter to see how to fit the generalized linear model.

Related Reading: Box Plots, Chapter 33.

Related Reading: Linear Models, Chapter 39.

References
