

Chapter 10

Analysis of Variance

Chapter Table of Contents

Introduction	205
One-Way Analysis of Variance	209
Nonparametric One-Way Analysis of Variance	215
Factorial Analysis of Variance	219
Linear Models	224
References	231

Chapter 10

Analysis of Variance

Introduction

Analysis of variance is a technique for exploring the variation of a continuous response variable (dependent variable). The response variable is measured at different levels of one or more classification variables (independent variables). The variation in the response due to the classification variables is computed, and you can test this variation against the residual error to determine the significance of the classification effects.

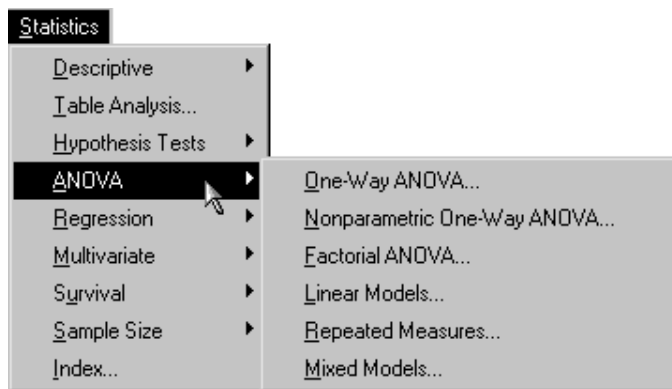


Figure 10.1. Analysis of Variance Menu

The Analyst Application provides several types of analyses of variance (ANOVA). The One-Way ANOVA task compares the means of the response variable over the groups defined by a single classification variable. See the section “One-Way Analysis of Variance” beginning on page 209 for more information.

The Nonparametric One-Way ANOVA task performs tests for location and scale differences over the groups defined by a single classification variable. Eight nonparametric tests are offered. See the section “Nonparametric One-Way Analysis of Variance” beginning on page 215 for more information.

The Factorial ANOVA task compares the means of the response variable over the groups defined by one or more classification variables. This type of analysis is useful when you have multiple ways of classifying the response values. See the “Factorial Analysis of Variance” section beginning on page 219 for more information.

The Linear Models task enables you to compare means and explain variation when you have a model that includes classification variables, quantitative variables, or both (such as in an analysis of covariance). See the “Linear Models” section beginning on page 224 for more information.

You can use the Repeated Measures task when you have multiple measurements of the response variable for the same experimental unit over different times or conditions or when the response values are assumed to be correlated within certain groups. For detailed information, see Chapter 16, “Repeated Measures.”

The Mixed Models task enables you to fit basic mixed models. A mixed model is a linear model that contains both fixed effects and random effects. For detailed information, see Chapter 15, “Mixed Models.”

The examples in this chapter demonstrate how you can use the Analyst Application to perform one-way and factorial ANOVA as well as to fit the linear model.

The Air Quality Data Set

The data set used in the following examples contains measurements on air quality recorded in an industrial valley. The measurements are taken hourly for a period of one week.

The first variable in the data set `Air` is a SAS datetime variable (`datetime`) that contains the date and the time of day on which the observation was taken. The data set contains two additional time-related variables related to `datetime` that record the day of the week (`day`) and the hour of the day (`hour`).

The variables measuring air quality are `co` (carbon monoxide), `o3` (ozone), `so4` (sulfate), `no` (nitrous oxide), and `dust` (particulates). The final variable provided is `wind`, which gives the wind speed in knots.

Open the Air Data Set

The data are provided in the Analyst Sample Library. To access this Analyst sample data set, follow these steps:

1. Select **Tools** → **Sample Data** . . .
2. Select `Air`.
3. Click **OK** to create the sample data set in your `Sasuser` directory.
4. Select **File** → **Open By SAS Name** . . .
5. Select `Sasuser` from the list of **Libraries**.
6. Select `Air` from the list of members.
7. Click **OK** to bring the `Air` data set into the data table.

Create a New Variable

To perform the analyses in the following examples, you need to create a new variable to represent the factory workshift periods. The new character variable, `shift`, recodes the variable `hour` into three factory workshift periods. For information on recoding ranges and computing variables, see the section “Recoding Ranges” on page 36 in Chapter 2.

Figure 10.2 displays the Recoding Ranges Information dialog. Enter the information to create the new variable as shown in Figure 10.2.

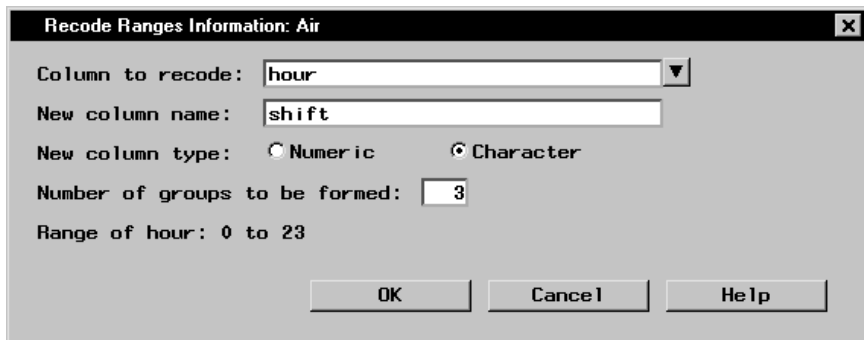


Figure 10.2. Recoding Ranges Information: Defining the New Variable

Click **OK** to display the Recoding Ranges dialog (Figure 10.3). To define the values for the new variable, **shift**, enter the values as shown in Figure 10.3.

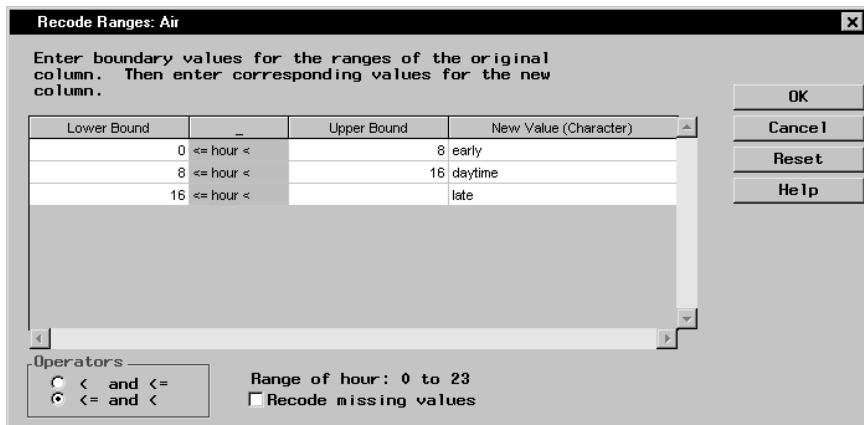


Figure 10.3. Recoding Ranges: Defining the Values for the New Variable

The values of the new variable **shift** are as follows: ‘early’ corresponds to the hours between 0 and 8 (from midnight until 8 a.m.), ‘daytime’ corresponds to the hours between 8 and 16 (from 8 a.m. until 4 p.m.), and ‘late’ corresponds to the hours greater than or equal to 16 (from 4 p.m. to midnight).

One-Way Analysis of Variance

The One-Way ANOVA task enables you to perform an analysis of variance when you have a continuous dependent variable and a single classification variable.

For example, consider the data set on air quality (Air), described in the preceding section. Suppose you want to compare the ozone level corresponding to each of the three factory workshift periods.

Request the One-Way ANOVA Task

To request the one-way ANOVA task, follow these steps:

1. Select **Statistics** → **ANOVA** → **One-Way ANOVA** . . .
2. Select o3 as the dependent variable.
3. Select shift as the independent variable.

Figure 10.4 defines the one-way ANOVA model.

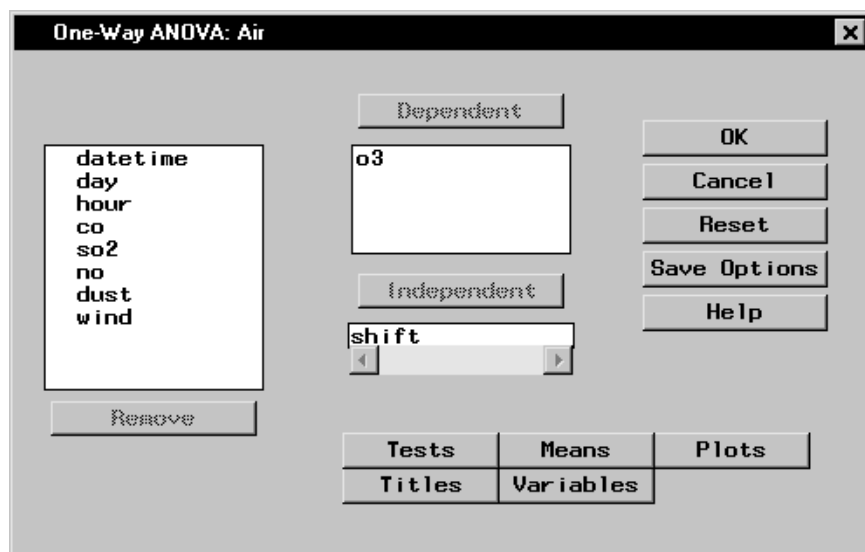


Figure 10.4. One-Way ANOVA Dialog

Request a Means Comparison Test

The analysis of variance performed in the One-Way ANOVA task indicates whether the means of the groups are different; it does not indicate which particular means are different. To generate more detailed information about the differences between the means, follow these steps:

1. Click on the **Means** button in the main dialog. The resulting window displays the **Comparisons** tab.
2. Click on the arrow adjacent to the **Comparison method** list.
3. Select **Tukey's HSD**.
4. Highlight the variable **shift** in the **Main Effects:** box.
5. Click on the **Add** button.

You can click on the arrow next to **Significance level:** to select a significance level, or you can type in the desired value.

6. Click **OK**.

Figure 10.5 specifies Tukey's studentized range (HSD) means comparison test at the 0.05 significance level.

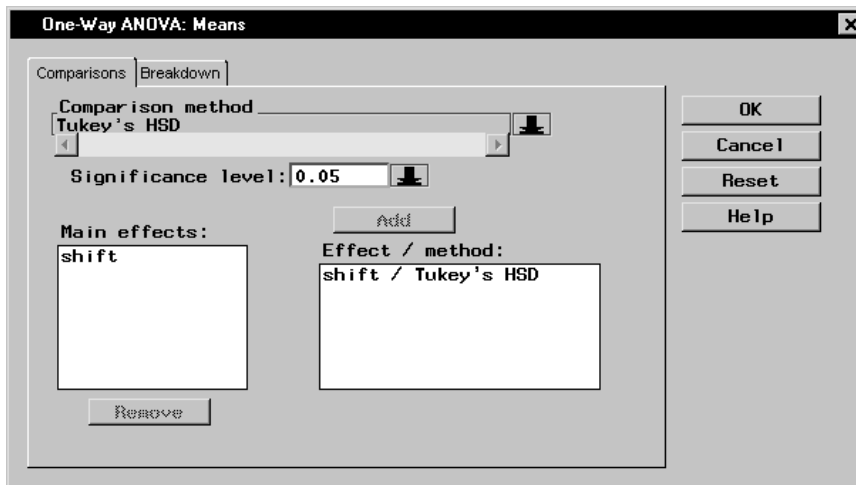


Figure 10.5. One-Way ANOVA: Means Dialog

Request a Box-and-Whisker Plot

To request a box-and-whisker plot in addition to the analysis, follow these steps:

1. Click on the **Plots** button in the main dialog.
2. Select **Box-&-whisker plot**.
3. Click **OK**.

Figure 10.6 displays the Plots dialog with the **Box-&-whisker plot** selected.

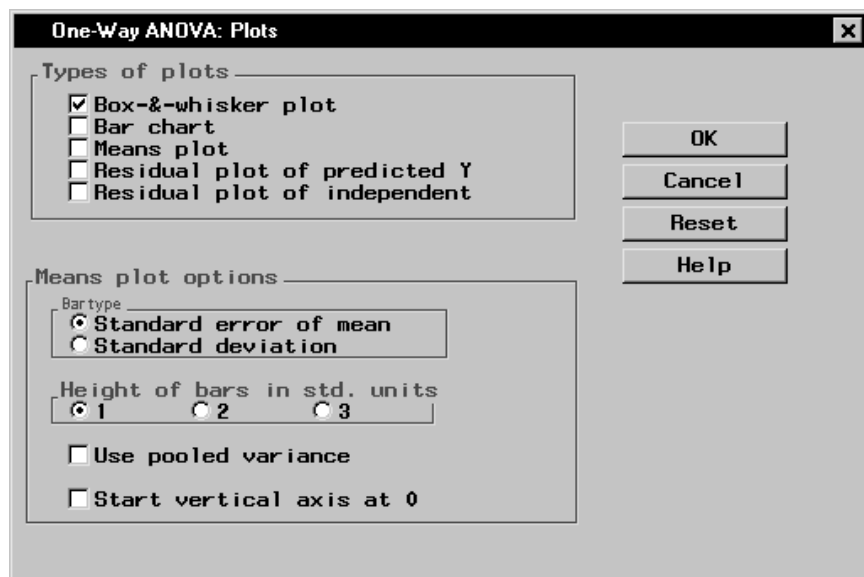


Figure 10.6. One-Way ANOVA: Plots Dialog

Click **OK** in the One-Way ANOVA dialog to perform the analysis.

Review the Results

This analysis tests whether the independent variable (shift) is a significant factor in accounting for the variation in ozone levels. Figure 10.7 displays the analysis of variance table, with an F statistic of 31.93 and an associated p -value that is less than 0.0001. The small p -value indicates that the model explains a highly significant proportion of the variation present in the dependent variable.

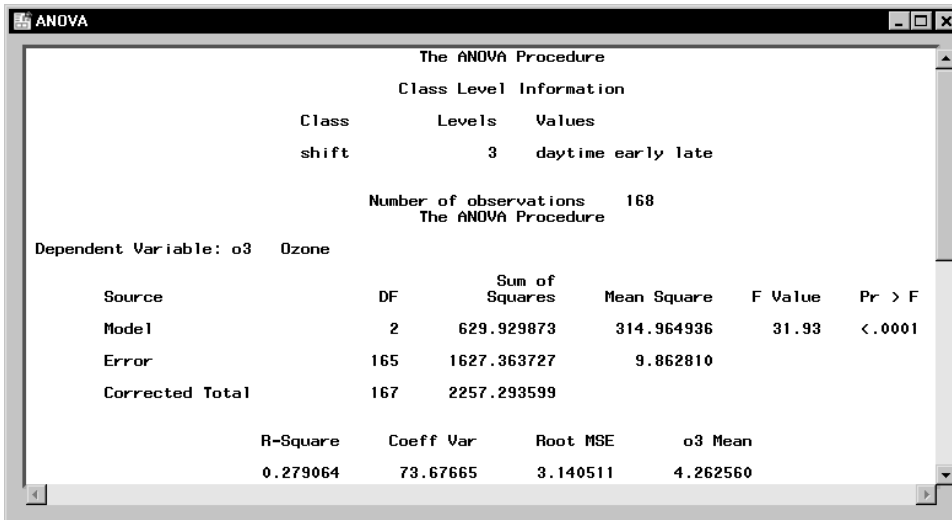


Figure 10.7. One-Way ANOVA: Analysis Results

The R-square value, which follows the ANOVA table in Figure 10.7, represents the proportion of variability accounted for by the independent variable. Approximately 28% of the variability in the ozone level can be accounted for by differences between shifts.

Information detailing which particular means are different is available in the multiple comparison test, as displayed in Figure 10.8. The means comparison output provides the alpha value, error degrees of freedom, and error mean square.

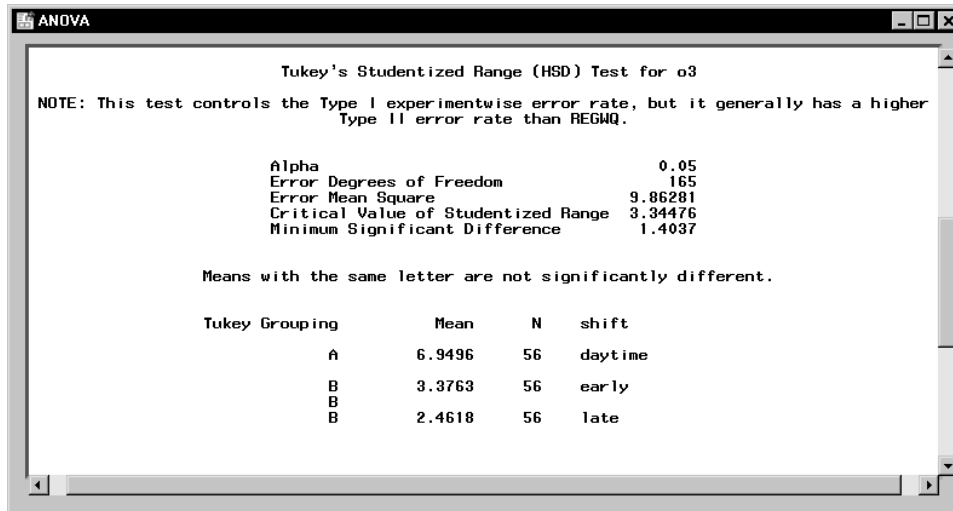


Figure 10.8. One-Way ANOVA: Multiple Comparisons Results

In the “Tukey Grouping” table, means with the same letter are not significantly different. The analysis shows that the daytime shift is associated with ozone levels that are significantly different from the other two shifts. The early and late shifts cannot be statistically distinguished on the basis of mean ozone level.

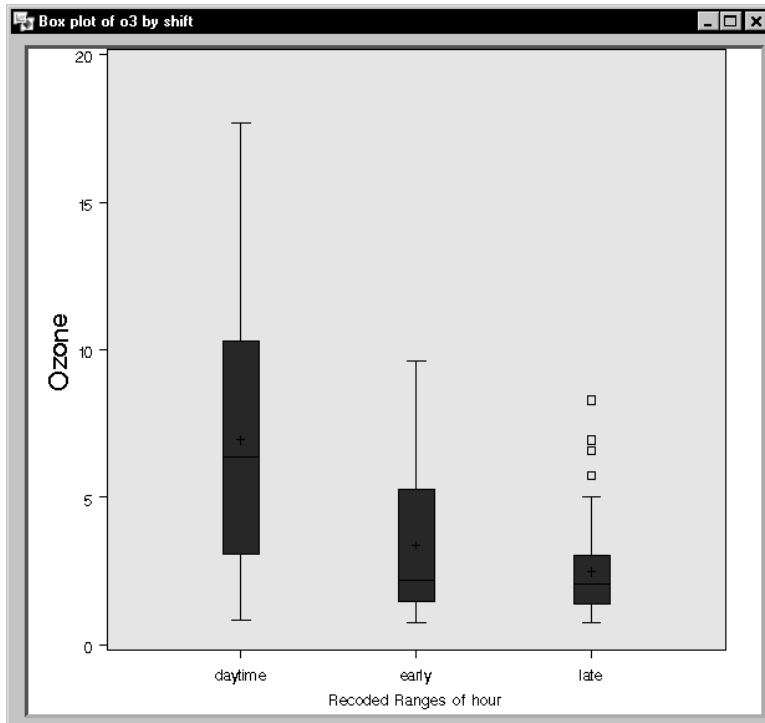


Figure 10.9. One-Way ANOVA: Box-and-Whisker Plot

The box-and-whisker plot displayed in Figure 10.9 provides a graphical view of the multiple comparison results. The variance among the ozone levels may be unequal: subsequent analyses may include a test for homogeneity of variance or a transformation of the response variable, `o3`.

Nonparametric One-Way Analysis of Variance

In statistical inference, or hypothesis testing, the traditional tests are called parametric tests because they depend on the specification of a probability distribution (such as the normal) except for a set of free parameters. Parametric tests are said to depend on distributional assumptions. Nonparametric tests, on the other hand, do not require distributional assumptions. Even if the data are distributed normally, nonparametric methods are often almost as powerful as parametric methods.

The Nonparametric One-Way ANOVA task enables you to perform nonparametric tests for location and scale when you have a continuous dependent variable and a single independent classification variable. You can perform a nonparametric one-way ANOVA using Wilcoxon (Kruskal-Wallis), median, Van der Waerden, and Savage scores. In addition, you can test for scale differences across levels of the independent variable using Ansari-Bradley, Siegal-Tukey, Klotz, and Mood scores. The Nonparametric One-Way ANOVA task provides asymptotic and exact p -values for all tests for location and scale.

For example, consider the air quality data set (Air), described in the section “The Air Quality Data Set” on page 206. Suppose that you want to perform a nonparametric one-way ANOVA and also test for scale differences for ozone levels across shift periods.

Request the Nonparametric One-Way ANOVA

To request a nonparametric one-way ANOVA, follow these steps:

1. Select **Statistics** → **ANOVA** → **Nonparametric One-Way ANOVA** . . .
2. Select **o3** as the dependent variable.
3. Select **shift** as the independent variable.

Figure 10.10 defines the nonparametric one-way ANOVA model.

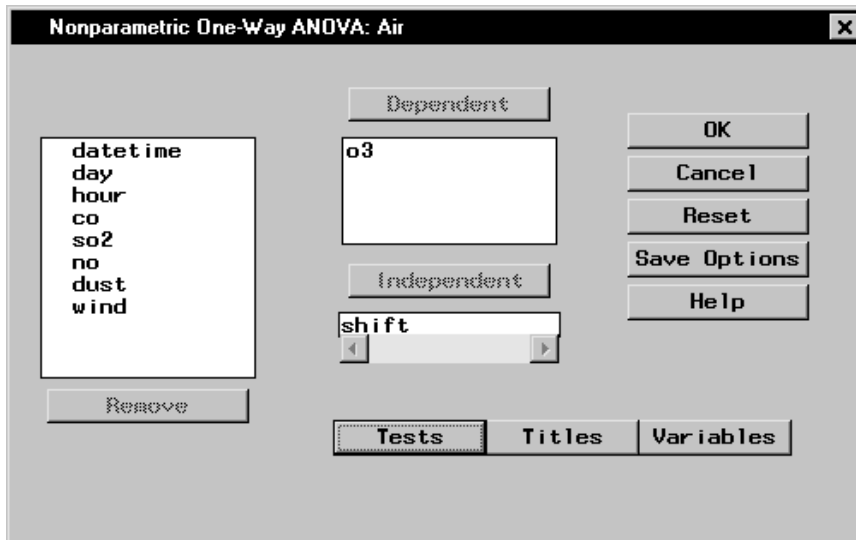


Figure 10.10. Nonparametric One-Way ANOVA: Main Dialog

Request Nonparametric Tests

You can use a nonparametric test for location to determine whether the air quality is the same at different times of the day. The Kruskal-Wallis test is a commonly used nonparametric technique for testing location differences and is produced using Wilcoxon scores.

The box-and-whisker plot in Figure 10.9 indicates that ozone levels may be more variable during the daytime shift than during the early shift or at night. You can use the Ansari-Bradley test to test for scale differences across shifts.

To request the Kruskal-Wallis and Ansari-Bradley tests, follow these steps:

1. Click on the **Tests** button in the main dialog.
2. Select **Wilcoxon (Kruskal-Wallis test)** in the **Location test scores**.
3. Select **Ansari-Bradley** in the **Dispersion test scores** box.

Figure 10.11 displays the Tests dialog with the **Wilcoxon (Kruskal-Wallis)** and **Ansari-Bradley** tests selected.

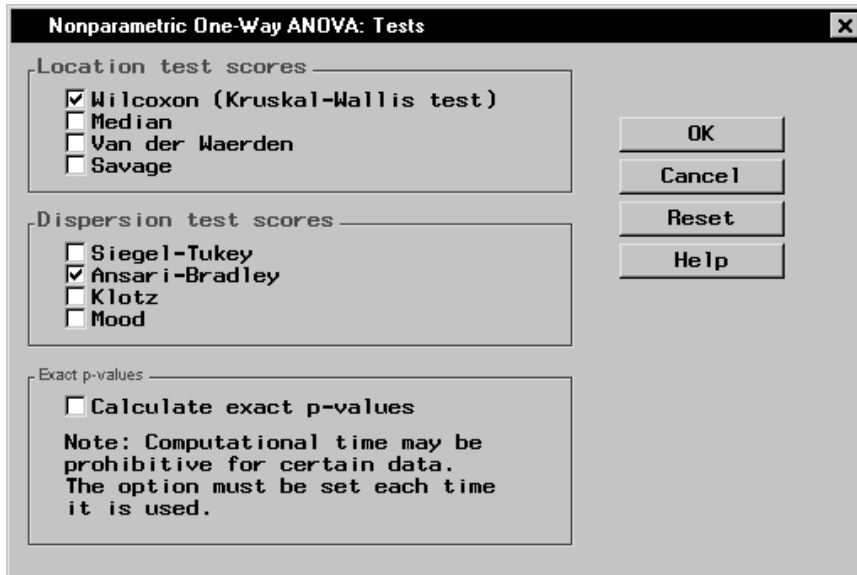


Figure 10.11. Nonparametric One-Way ANOVA: Tests Dialog

Click **OK** in the Nonparametric One-Way ANOVA dialog to perform the analysis.

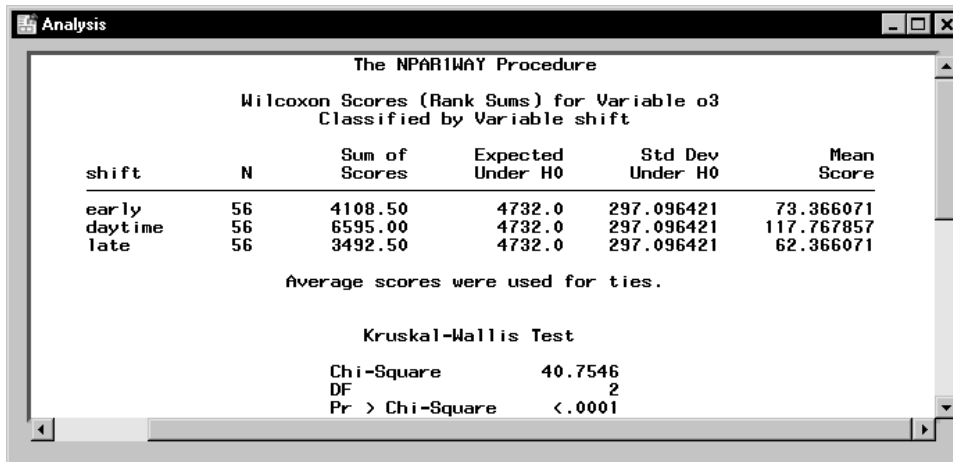


Figure 10.12. Nonparametric One-Way ANOVA: Kruskal-Wallis Test Results

Figure 10.12 displays the Wilcoxon scores and Kruskal-Wallis test results. The table labeled “Wilcoxon Scores (Rank Sums) for Variable o3” contains the sum of the rank scores, expected sum, and mean score for each shift. The daytime shift has a mean score of 117.77, which is higher than the mean scores of both the early and late shift. The “Kruskal-Wallis Test” table displays the results of the Kruskal-Wallis test. The test statistic of 40.75 indicates that there is a significant difference in ozone levels across shift times (the p -value is less than 0.0001).

shift	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
early	56	2345.250	2380.0	148.379376	41.879464
daytime	56	2089.500	2380.0	148.379376	37.312500
late	56	2705.250	2380.0	148.379376	48.308036

Average scores were used for ties.

Ansari-Bradley One-Way Analysis

Chi-Square	5.7952
DF	2
Pr > Chi-Square	0.0552

Figure 10.13. Nonparametric One-Way ANOVA: Ansari-Bradley Test Results

Figure 10.13 displays the results of the Ansari-Bradley test. The Ansari-Bradley test chi-square has the value of 5.80 with 2 degrees of freedom, which is not significant at the $\alpha = 0.05$ level. Since the p -value is just slightly higher than 0.05, there is moderate evidence of scale differences across shift times.

Factorial Analysis of Variance

The Factorial ANOVA task enables you to perform an analysis of variance when you have multiple classification variables.

For example, consider the data set on air quality (Air), described in the section “The Air Quality Data Set” on page 206. Suppose you want to compare ozone levels for each day of the week and for each factory workshift. You can define a factorial model that includes the two classification variables, `day` and `shift`.

In this example, a factorial model is specified, and a plot of the two-way effects is requested.

Request the Analysis

To request a factorial analysis of variance, follow these steps:

1. Click on **Statistics** → **ANOVA** → **Factorial ANOVA . . .**
2. Select `o3` as the dependent variable.
3. Select `shift` and `day` as the independent variables.

The resulting Factorial ANOVA dialog is displayed in Figure 10.14.

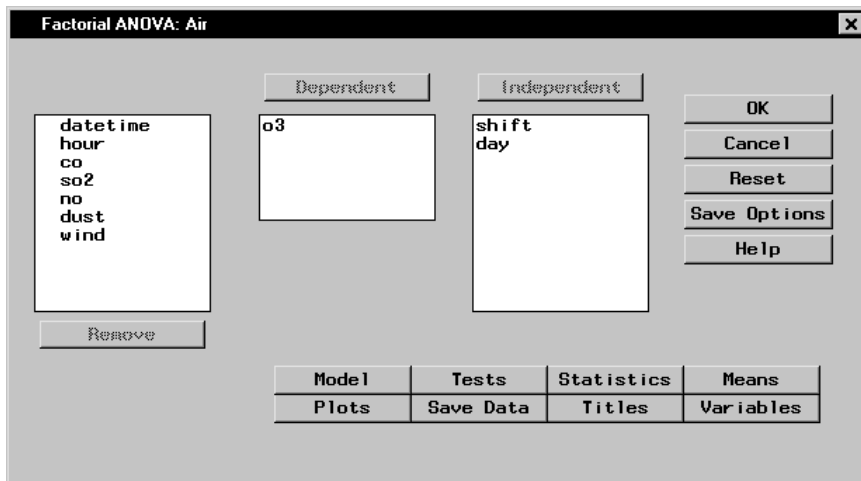


Figure 10.14. Factorial ANOVA Dialog

The default ANOVA model includes only the main effects (that is, the terms representing **shift** and **day**). To include an interaction term, or to specify other options for your analysis, you can use the dialogs available in the Factorial ANOVA task.

Specify the Model

To specify a factorial model, follow these steps:

1. Click on the **Model** button in the main dialog.
2. Highlight the variables **shift** and **day** in the resulting dialog.
3. Click on the **Factorial** button.
4. Click **OK**.

Figure 10.15 displays the Model dialog with the terms **shift**, **day**, and the interaction term **shift*day** selected as effects in the model.

Note that you can build specific models with the **Add**, **Cross**, and **Factorial** buttons, or you can select a model by clicking on the **Standard Models** button and making a selection from the drop-down list. From this list, you can request that your model include main effects only, effects up to two-way interactions, or effects up to three-way interactions.

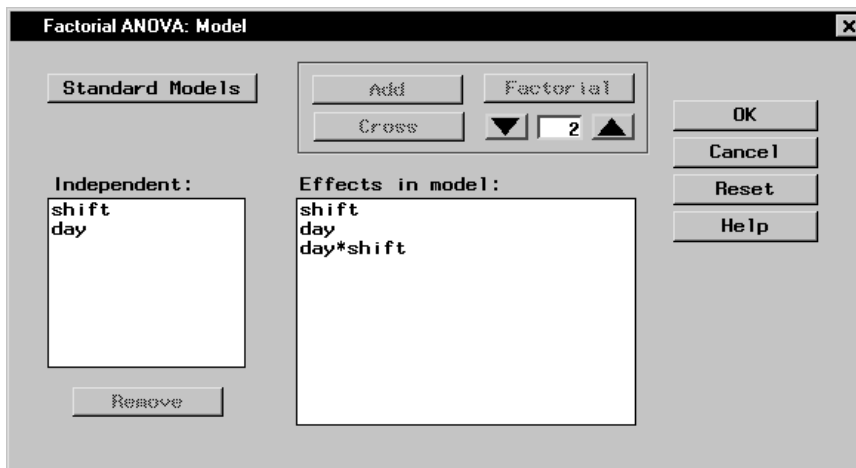


Figure 10.15. Factorial ANOVA: Model Dialog

Request a Means Plot

A means plot displays a symbol for the observed or predicted means at each level of a specified variable, with vertical bars extending for a specified number of standard errors. The means for each level of an effect are joined with line segments. To request a plot of the dependent means, follow these steps:

1. Click on the **Plots** button in the main dialog. The resulting window displays the **Means** tab.
2. Select **Plot dependent means for two-way effects**.

You can choose to plot either the observed or predicted means of the dependent variable. Additionally, you can choose whether the vertical bars should represent one, two, or three standard errors.

3. Click **OK**.

Figure 10.16 requests a plot of the observed dependent means for the two-way effects.

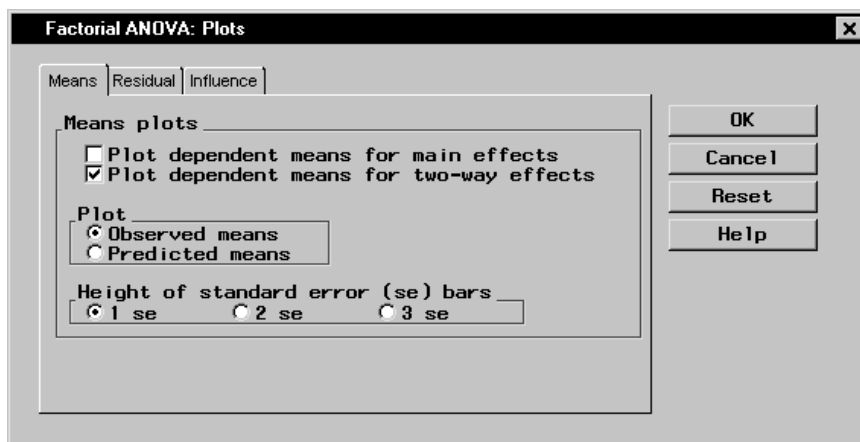


Figure 10.16. Factorial ANOVA: Plots Dialog

Click **OK** in the main dialog to perform the analysis.

Review the Results

Figure 10.17 displays information on the levels of the two classification variables, shift and day, followed by the ANOVA table. The model sum of squares is partitioned into the separate contributions of the individual model effects, and F tests are provided for each effect.

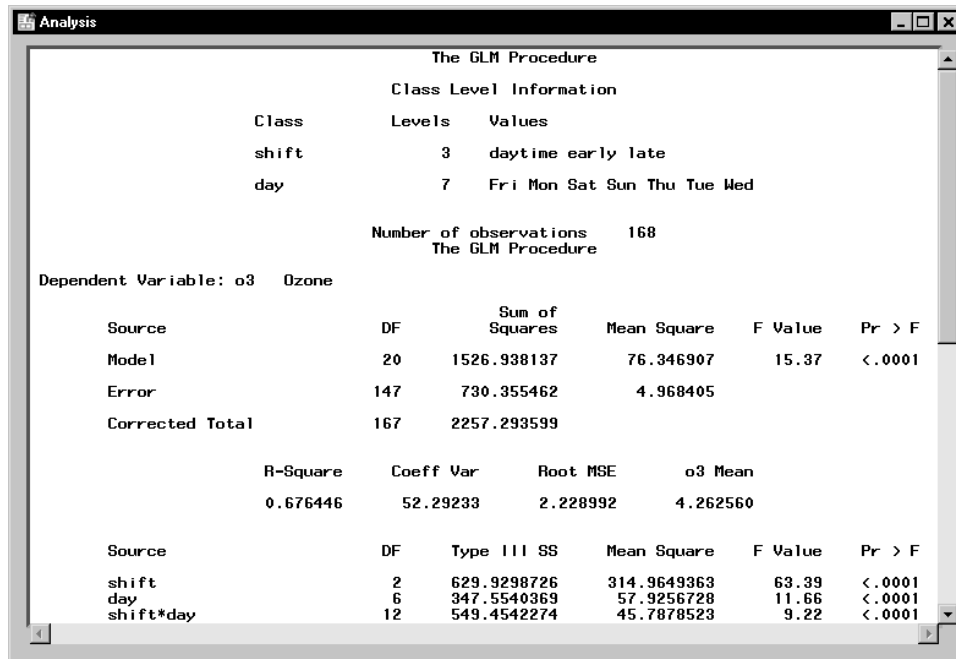


Figure 10.17. Factorial ANOVA: Analysis Results

The F statistic of 15.37 indicates that the model as a whole is highly significant (the p -value is less than 0.0001). Additionally, the R-square value of 0.6764 means that about 68% of the variation of ozone can be accounted for by the factorial model.

The table at the bottom of Figure 10.17 displays the significance test for each term of the model. The main effects and the interaction term are each significant at the $\alpha = 0.05$ level (that is, each p -value is much less than 0.05).

In Figure 10.18, the three curves display ozone concentration across days of the week. Each curve represents the relationship for one of the three factory workshift periods.

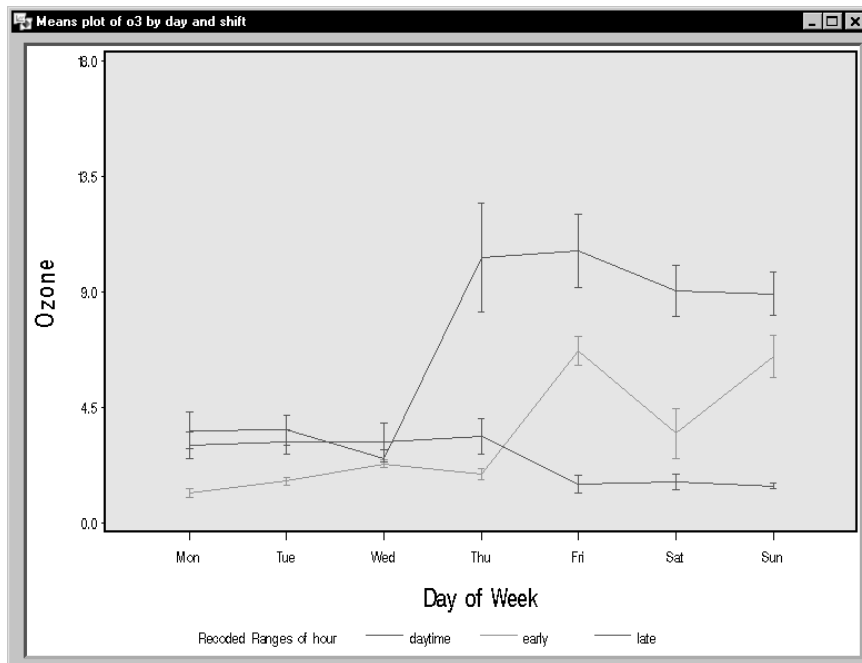


Figure 10.18. Factorial ANOVA: Means Plot

The means plot indicates an inverse relationship between the daytime and late shifts. The ozone levels during the daytime shift rise dramatically on Thursday and remain high throughout the weekend. Ozone levels for the late shift, on the other hand, start to decrease after Thursday and remain low throughout the weekend.

Linear Models

The Linear Models task enables you to perform an analysis of variance when you have a continuous dependent variable with classification variables, quantitative variables, or both.

The data set *Air*, described in the section “The Air Quality Data Set” on page 206, includes quantitative measures; for example, the variable *wind* represents wind speed, in knots. Suppose that you want to model ozone levels using the variables *day* (day of week), *shift* (factory workshift period), and *wind* (wind speed). Suppose that you also want your model to include the interaction between the variables *day* and *shift*. That is, you want to perform a simple two-way analysis of covariance with unequal slopes.

The following example fits this linear model and additionally requests a retrospective power analysis and a plot of the observed values versus the predicted values.

Request the Linear Models Analysis

To request the linear models analysis, follow these steps:

1. Select **Statistics** → **ANOVA** → **Linear Models** . . .
2. Select *o3* as the dependent variable.
3. Select *shift* and *day* as the class variables.
4. Select *wind* as the quantitative variable.

Figure 10.19 displays the Linear Models dialog.

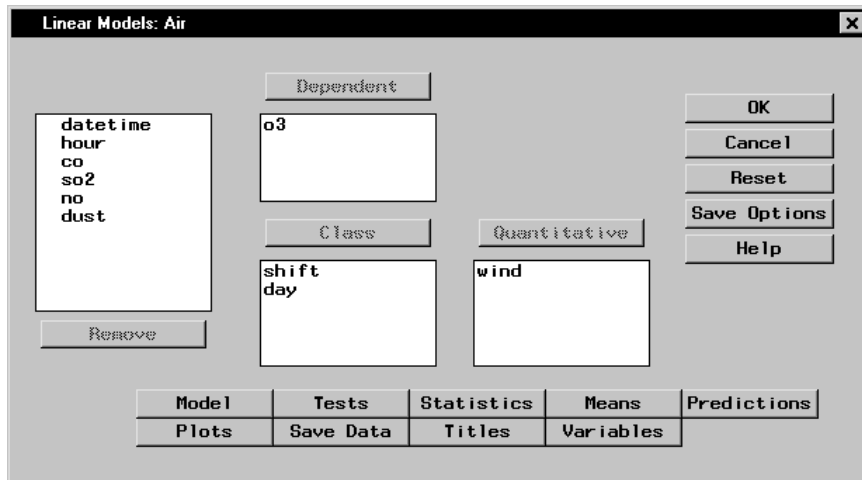


Figure 10.19. Linear Models Dialog

By default, the linear model analysis includes only the main effects specified in the main dialog: no interaction term is included.

Specifying an Interaction Term in the Model

To include the interaction term `shift*day` in your model, follow these steps:

1. Click on the **Model** button in the main dialog.
2. Highlight the variables `shift` and `day`.
3. Click on the **Cross** button.
4. Click **OK**.

Note that you can build specific models with the **Add**, **Cross**, and **Factorial** buttons, or you can select a model by clicking on the **Standard Models** button and making a selection from the pop-up list.

Figure 10.20 displays the Model dialog with the terms `shift` and `day` and the interaction term `shift*day` selected as effects in the model.

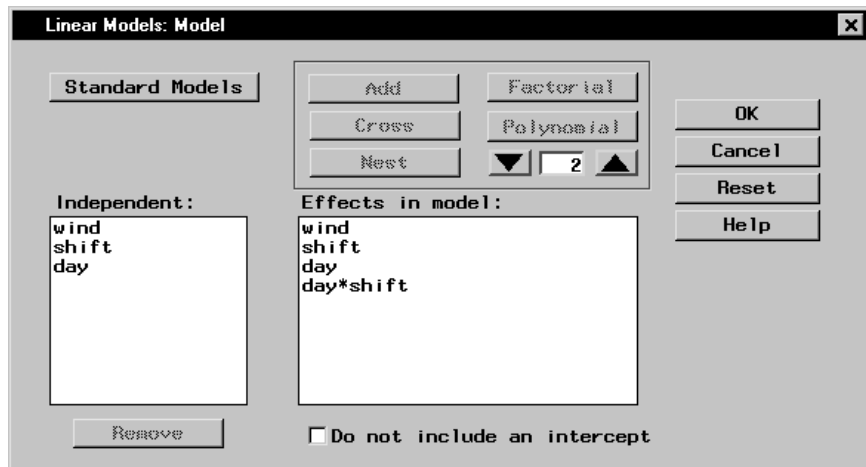


Figure 10.20. Linear Models: Model Dialog

Request a Power Analysis

The power of a test is the probability of correctly rejecting the null hypothesis of no difference. It depends on the sample size as well as the precise difference specified in the alternative hypothesis. Ideally, you consider power before gathering data to ensure that you gather enough data to detect a difference. However, once you have gathered your data, you can perform a retrospective power analysis in order to determine how much data is needed to detect the observed difference. To perform a retrospective power analysis with the Analyst Application, follow these steps:

1. Click on the **Tests** button in the main dialog.
2. Click on the **Power Analysis** tab.
3. Select **Perform power analysis**.

To request power calculations for tests performed at several α values, you can enter the values, separated by a space, in the box labeled **Alphas**. You can request power analysis for additional sample sizes in the **Sample sizes** box. You can enter one or more specific values for the sample sizes, or you can specify a series of sample sizes in the boxes labeled **From:**, **To:**, and **By:**.

4. Click **OK**.

Figure 10.21 displays the **Power Analysis** tab, which requests a retrospective power analysis with an alpha, or significance level, of 0.05.

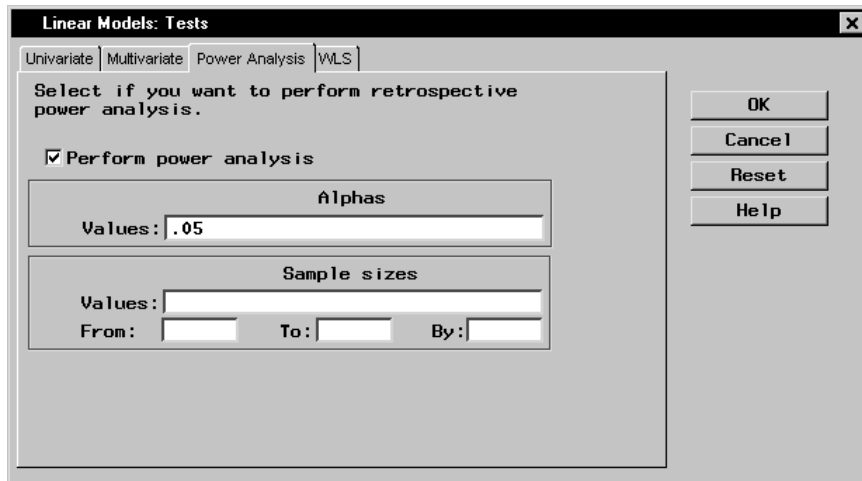


Figure 10.21. Linear Models: Tests Dialog

Request a Scatter Plot

To request a scatter plot of the predicted values versus the observed values, follow these steps:

1. Click on the **Plots** button in the main dialog.
2. Click on the **Predicted** tab.
3. Select **Plot observed vs predicted**.
4. Click **OK**.

Figure 10.22 displays the **Predicted** tab in the Plots dialog.

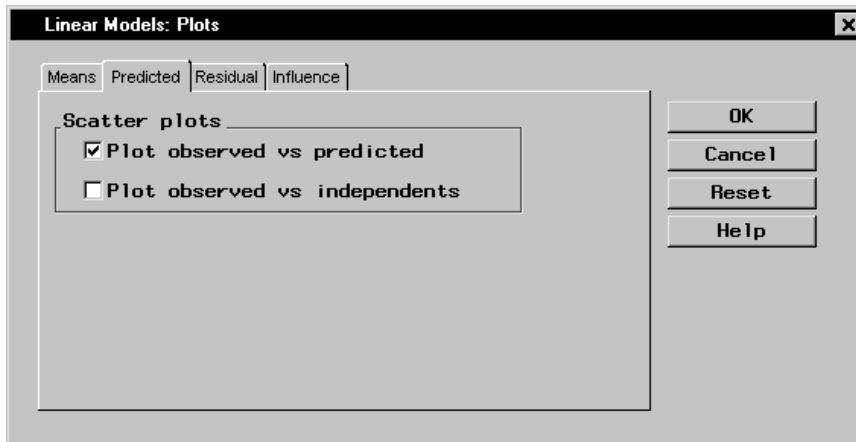


Figure 10.22. Linear Models: Plots Dialog

Click **OK** in the Linear Models dialog to perform the analysis.

Review the Results

The output of the analysis includes information about the levels of the independent variables, followed by the ANOVA table.

Figure 10.23 displays the analysis of variance table, with an F statistic of 19.44 and an associated p -value less than 0.0001. A p -value this small indicates that the model explains a highly significant proportion of the variation in the dependent variable.

The R-square value represents the proportion of variability accounted for by the independent variables. In this analysis, about 74% of the variation of the ozone level can be accounted for by the model (that is, by mean differences in **day** and **shift**, in conjunction with a linear dependence on wind speed).

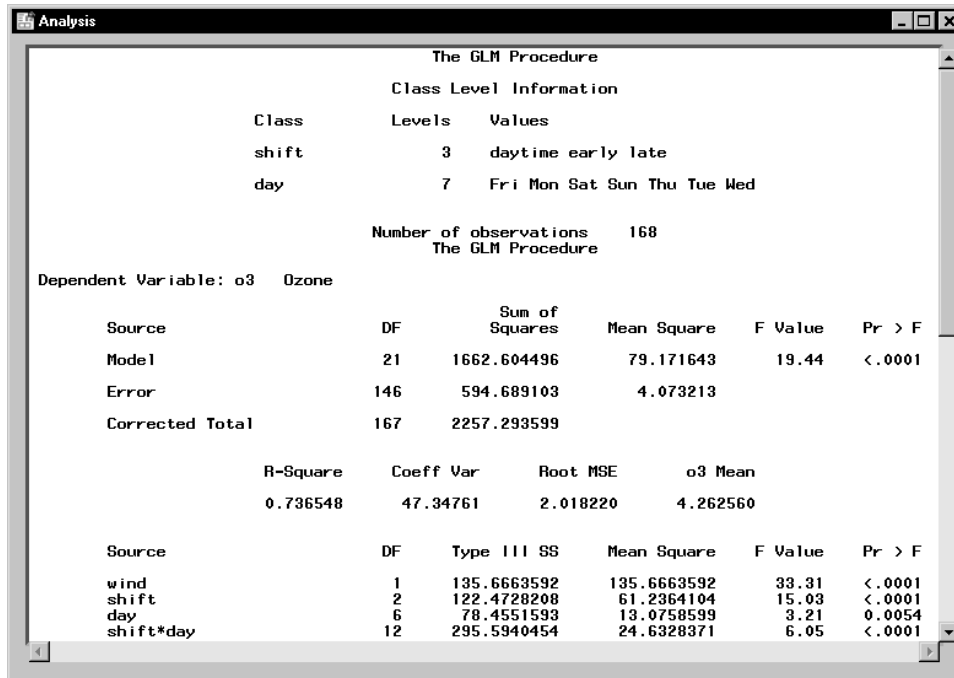
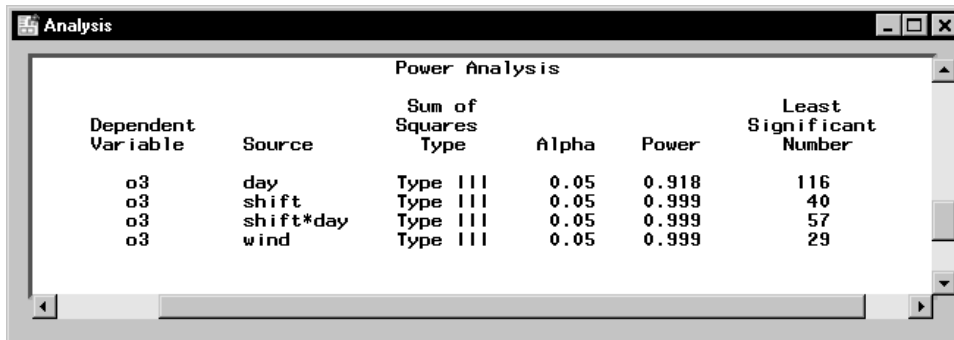


Figure 10.23. Linear Models: ANOVA Results

The last table displayed in Figure 10.23 partitions the model sum of squares into the separate contribution for each model effect and tests for the significance of each effect. The main effects and the interaction term are significant at the $\alpha = 0.05$ level (that is, each p -value is less than 0.05).

Figure 10.24 displays the retrospective power analysis. The observed power is given for each effect in the linear model.



Dependent Variable	Source	Sum of Squares Type	Alpha	Power	Least Significant Number
o3	day	Type III	0.05	0.918	116
o3	shift	Type III	0.05	0.999	40
o3	shift*day	Type III	0.05	0.999	57
o3	wind	Type III	0.05	0.999	29

Figure 10.24. Linear Models: Power Analysis

The column labeled Least Significant Number in Figure 10.24 displays the smallest number of observations required to determine that the effect is significant at the given α value.

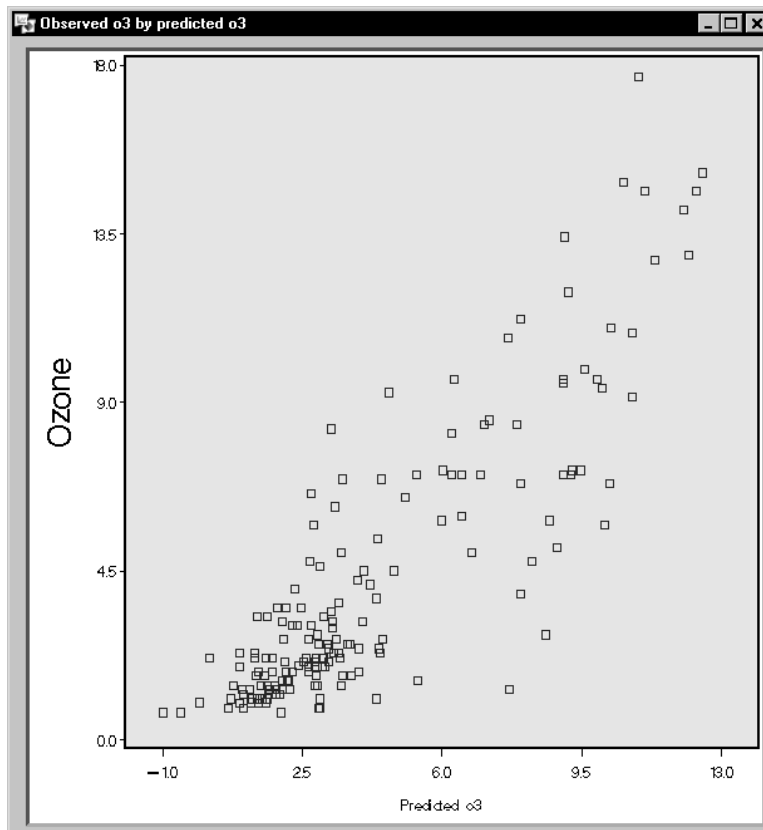


Figure 10.25. Linear Models: Observed Ozone Levels versus Predicted Values

Figure 10.25 displays the plot of the observed values versus the predicted values from the model. If the model predicts the observed values perfectly, the points on the plot fall on a straight line with a slope of 1. This plot indicates reasonable prediction.

References

- SAS Institute Inc. (1999), *SAS/STAT User's Guide, Version 7-1*, Cary, NC: SAS Institute Inc.
- Littell, Ramon C., Freund, Rudolf J., and Spector, Philip C. (1991), *SAS System for Linear Models, Third Edition* by Ramon C. Littell, Rudolf J. Freund, and Philip C. Spector

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *The Analyst Application, First Edition*, Cary, NC: SAS Institute Inc., 1999. 476 pp.

The Analyst Application, First Edition

Copyright © 1999 SAS Institute Inc., Cary, NC, USA.

ISBN 1-58025-446-2

All rights reserved. Printed in the United States of America. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, by any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute, Inc.

U.S. Government Restricted Rights Notice. Use, duplication, or disclosure of the software by the government is subject to restrictions as set forth in FAR 52.227-19 Commercial Computer Software-Restricted Rights (June 1987).

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999

SAS[®] and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. [®] indicates USA registration.

IBM[®], ACF/VTAM[®], AIX[®], APPN[®], MVS/ESA[®], OS/2[®], OS/390[®], VM/ESA[®], and VTAM[®] are registered trademarks or trademarks of International Business Machines Corporation. [®] indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

The Institute is a private company devoted to the support and further development of its software and related services.