

Chapter 11

Regression

Chapter Table of Contents

Introduction	231
Simple Linear Regression	232
Multiple Linear Regression	238
Logistic Regression	247
References	254

Chapter 11

Regression

Introduction

Regression techniques enable you to investigate the relationship between a dependent variable (also called a *response* variable) and one or more explanatory variables (also called *predictor*, or *independent*, variables). In linear regression, the dependent variable is modeled as a linear function of the quantitative independent variables. For example, you can write the simple linear regression equation as

$$Y = b_0 + b_1X$$

where Y represents the single dependent variable, X is the explanatory variable, and b_0 and b_1 are regression coefficients.

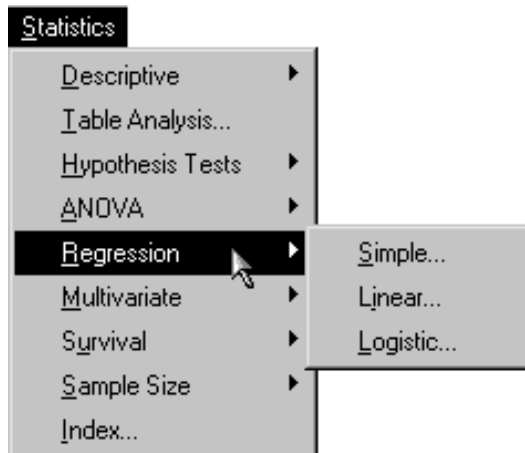


Figure 11.1. Regression Menu

The Analyst Application enables you to perform simple linear regression, multiple linear regression and logistic regression. In the Simple linear regression task, you model your dependent variable using a single explanatory variable. In the Linear regression task, you model your dependent variable using one or more explanatory variables. In the Logistic regression task, the dependent variable is discrete, and you model the variable using one or more explanatory variables.

The examples in this chapter demonstrate how you can use the Analyst Application to perform simple linear regression, multiple linear regression, and logistic regression.

Simple Linear Regression

In simple linear regression, there is a single quantitative independent variable. Suppose, for example, that you want to determine whether a linear relationship exists between the asking price for a house and its area in square feet. The area of the house is the quantitative independent variable, and the asking price for the house is the dependent variable.

The data set analyzed in this example is called **Houses**, and it contains the characteristics of fifteen houses for sale. The data set contains the following variables.

style	style category (ranch, split-level, condominium, or two-story)
sqfeet	area in square feet
bedrooms	number of bedrooms
baths	number of bathrooms
street	name of the street on which the house is located
price	asking price for the house

The task includes performing a simple regression analysis to predict the variable **price** from the explanatory variable, **sqfeet**.

Open the Houses Data Set

The data are provided in the Analyst Sample Library. To open the Houses data set, follow these steps:

1. Select **Tools** → **Sample Data** . . .
2. Select Houses.
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 Houses from the list of members.
7. Click **OK** to bring the Houses data set into the data table.

Request the Simple Regression Analysis

To request the simple regression analysis, follow these steps:

1. Select **Statistics** → **Regression** → **Simple** . . .
2. Select price from the candidate list as the Dependent variable.
3. Select sqfeet from the candidate list as the Explanatory variable.

Figure 11.2 displays the resulting dialog.

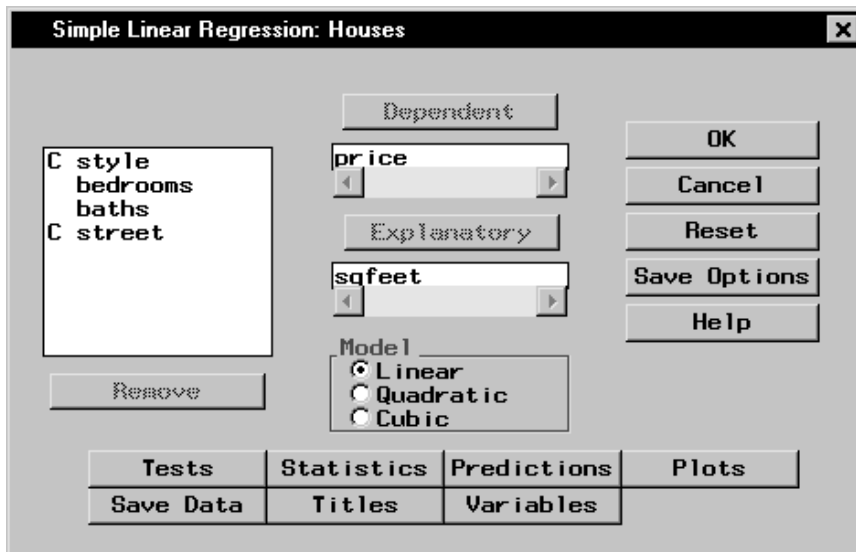


Figure 11.2. Simple Linear Regression Dialog

The model defined in this analysis is

$$\text{price} = b_0 + b_1 \text{sqfeet}$$

If you select **Quadratic** or **Cubic** in the **Model** box, the respective model is

$$\text{price} = b_0 + b_1 \text{sqfeet} + b_2 \text{sqfeet}^2$$

or

$$\text{price} = b_0 + b_1 \text{sqfeet} + b_2 \text{sqfeet}^2 + b_3 \text{sqfeet}^3$$

The default analysis fits the simple regression model.

Request a Scatter Plot of the Data

To request a plot of the observed values versus the independent values, follow these steps.

1. Click on the **Plots** button.

2. Select **Plot observed vs independent**.

You can add 95% confidence limits for the mean of the independent variable by selecting **Confidence limits**, or you can produce 95% prediction limits for individual predictions.

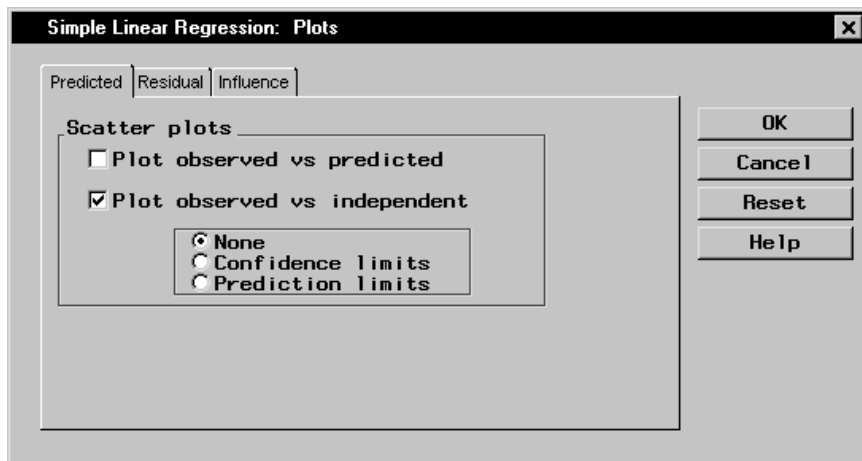
3. Click **OK**.

Figure 11.3. Simple Linear Regression: Plots Dialog

Click **OK** in the Simple Linear Regression dialog to perform the analysis.

Review the Results

The results are displayed in Figure 11.4. The ANOVA table is displayed in the results, followed by the table of parameter estimates. The least squares fit is

$$\text{price} = -14982 + 67.52 \times \text{sqfeet}$$

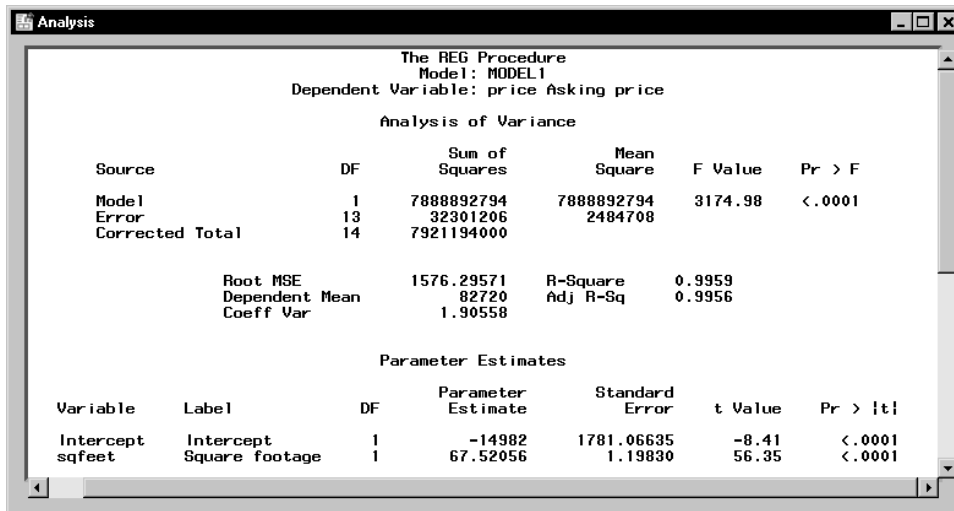


Figure 11.4. Simple Linear Regression: Results

The small p -values listed in the Pr > |t| column indicate that both parameter estimates are significantly different from zero.

The plot of the observed and independent variables is displayed in Figure 11.5. The plot includes the fitted regression line.

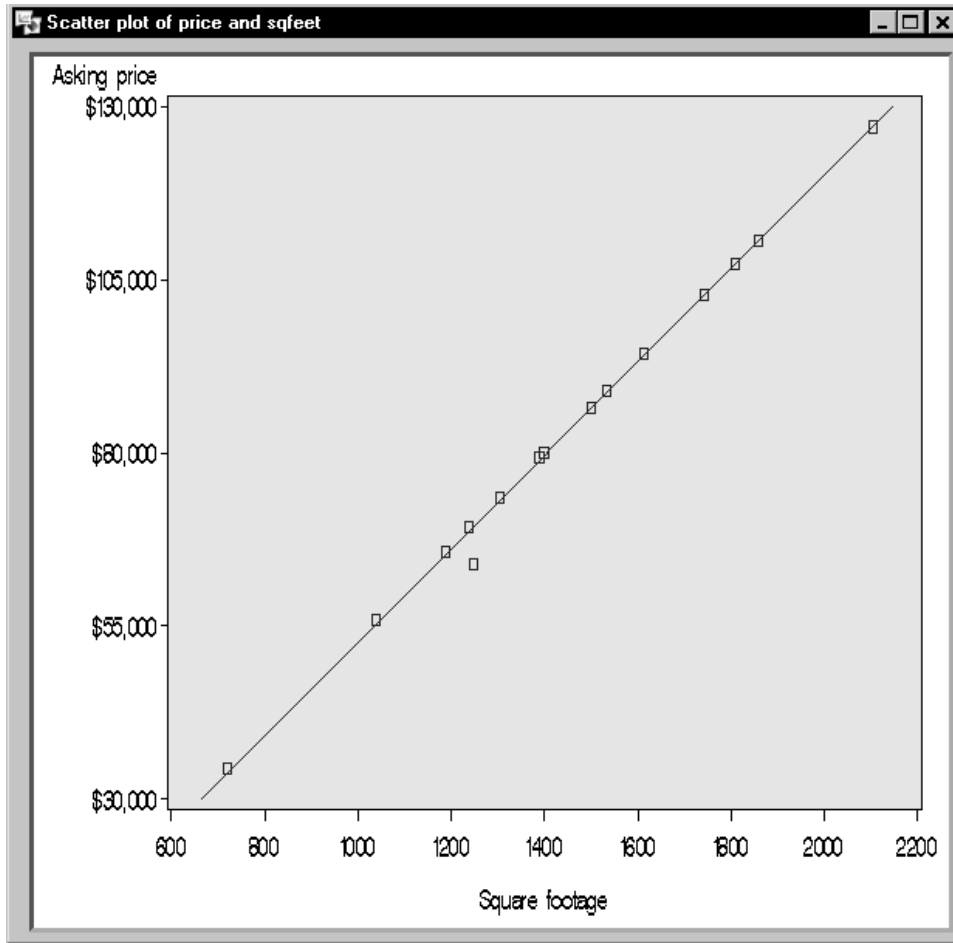


Figure 11.5. Simple Linear Regression: Scatter Plot with Regression Line

Multiple Linear Regression

You perform a multiple linear regression analysis when you have more than one explanatory variable for consideration in your model. You can write the multiple linear regression equation for a model with p explanatory variables as

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_pX_p$$

where Y is the response, or dependent, variable, the X s represent the p explanatory variables, and the b s are the regression coefficients.

For example, suppose that you would like to model a person's aerobic fitness as measured by the ability to consume oxygen. The data set analyzed in this example is named **Fitness**, and it contains measurements made on three groups of men involved in a physical fitness course at North Carolina State University. See "Computing Correlations" in Chapter 7, "Descriptive Statistics," for a complete description of the variables in the Fitness data set.

The goal of the study is to predict fitness as measured by oxygen consumption. Thus, the dependent variable for the analysis is the variable **oxygen**. You can choose any of the other quantitative variables (**age**, **weight**, **runtime**, **rstpulse**, **runpulse**, and **maxpulse**) as your explanatory variables.

Suppose that previous studies indicate that oxygen consumption is dependent upon the subject's age, the time it takes to run 1.5 miles, and the heart rate while running. Thus, in order to predict oxygen consumption, you estimate the parameters in the following multiple linear regression equation:

$$\text{oxygen} = b_0 + b_1 \text{age} + b_2 \text{runtime} + b_3 \text{runpulse}$$

This task includes performing a linear regression analysis to predict the variable **oxygen** from the explanatory variables **age**, **runtime**, and **runpulse**. Additionally, the task requests confidence intervals

for the estimates, a collinearity analysis, and a scatter plot of the residuals.

Open the Fitness Data Set

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

1. Select **Tools** → **Sample Data** . . .
2. Select **Fitness**.
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 **Fitness** from the list of members.
7. Click **OK** to bring the **Fitness** data set into the data table.

Request the Linear Regression Analysis

To specify the analysis, follow these steps:

1. Select **Statistics** → **Regression** → **Linear** . . .
2. Select the variable **oxygen** from the candidate list as the dependent variable.
3. Select the variables **age**, **runtime**, and **runpulse** as the explanatory variables.

Figure 11.6 displays the resulting Linear Regression task.

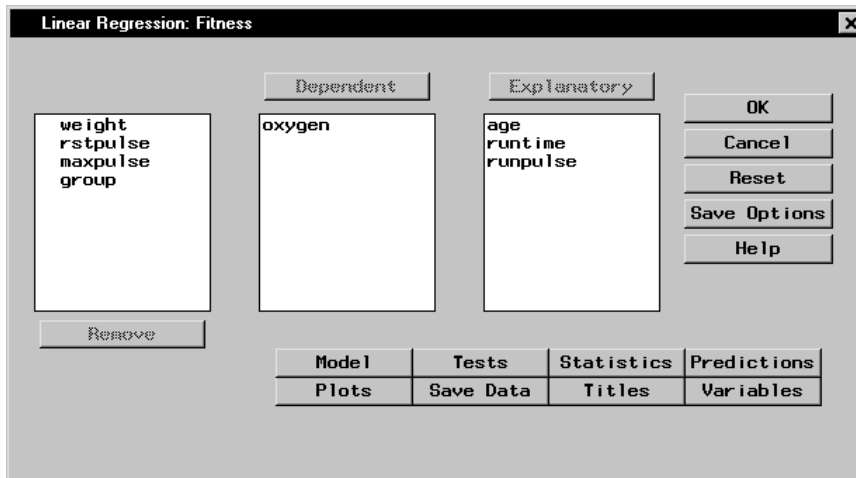


Figure 11.6. Linear Regression Dialog

The default analysis fits the linear regression model.

Request Additional Statistics

You can request several additional statistics for your analysis in the Statistics dialog.

To request that confidence limits be computed, follow these steps:

1. Click on the **Statistics** button.
2. In the **Statistics** tab, select **Confidence limits for estimates**.

Figure 11.7 displays the **Statistics** tab in the Statistics dialog.

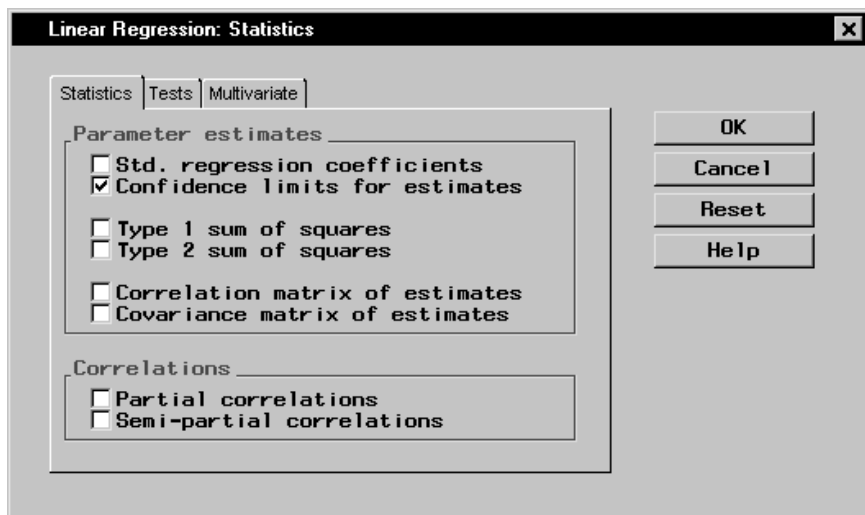


Figure 11.7. Linear Regression: Statistics Dialog, Statistics Tab

To request a collinearity analysis, follow these steps:

1. Click on the **Tests** tab in the Statistics dialog.
2. Select **Collinearity analysis**.
3. Click **OK**.

The dialog in Figure 11.8 requests a collinearity analysis in order to assess dependencies among the explanatory variables.

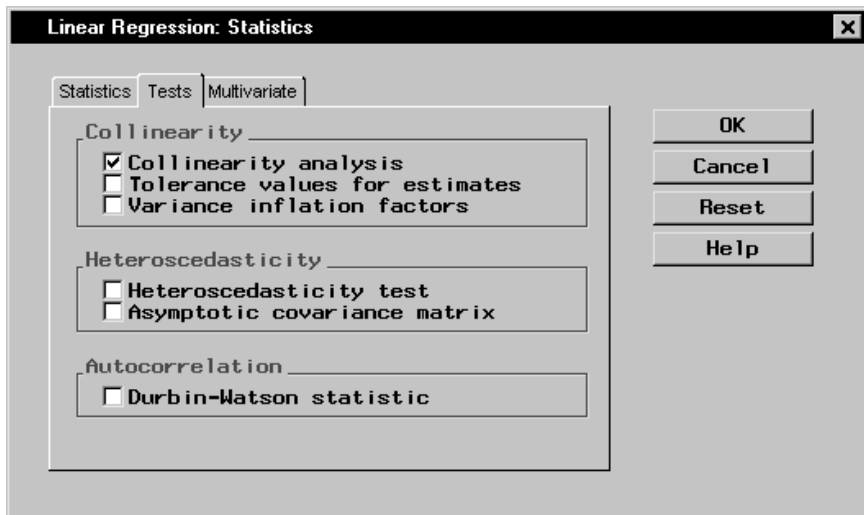


Figure 11.8. Linear Regression: Statistics Dialog, Tests Tab

Request a Scatter Plot of the Residuals

To request a plot of the studentized residuals versus the predicted values, follow these steps:

1. In the Linear Regression main dialog, click on the **Plots** button.
2. Click on the **Residual** tab.
3. Select **Plot residuals vs variables**.
4. In the box labeled **Residuals**, check the selection **Studentized**.
5. In the box labeled **Variables**, check the selection **Predicted Y**.
6. Click **OK**.

Figure 11.9 displays the **Residual** tab.

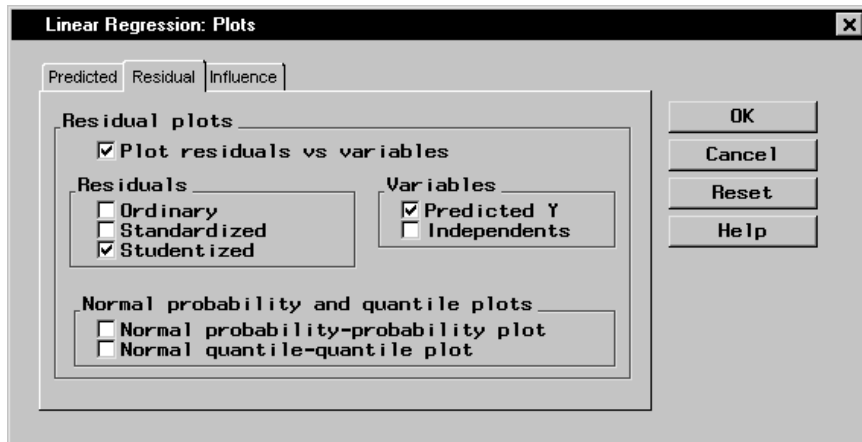


Figure 11.9. Linear Regression: Plots Dialog, Residual Tab

An ordinary residual is the difference between the observed response and the predicted value for that response. The standardized residual is the ratio of the residual to its standard error; that is, it is the ordinary residual divided by its standard error. The studentized residual is the standardized residual calculated with the current observation deleted from the analysis.

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

Review the Results

Figure 11.10 displays the analysis of variance table and the parameter estimates.

The REG Procedure
Model: MODEL1
Dependent Variable: oxygen Oxygen consumption

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	690.55086	230.18362	38.64	<.0001
Error	27	160.83069	5.95669		
Corrected Total	30	851.38154			

Root MSE 2.44063 R-Square 0.8111
Dependent Mean 47.37581 Adj R-Sq 0.7901
Coeff Var 5.15165

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	111.71806	10.23509	10.92	<.0001
age	Age in years	1	-0.25640	0.09623	-2.66	0.0129
runtime	Min. to run 1.5 miles	1	-2.82538	0.35828	-7.89	<.0001
runpulse	Heart rate while running	1	-0.13091	0.05059	-2.59	0.0154

Figure 11.10. Linear Regression: ANOVA Table and Parameter Estimates

In the analysis of variance table displayed in Figure 11.10, the F value of 38.64 (with an associated p -value that is less than 0.0001) indicates a significant relationship between the dependent variable, oxygen, and at least one of the explanatory variables. The R-square value indicates that the model accounts for 81% of the variation in oxygen consumption.

The “Parameter Estimates” table lists the degrees of freedom, the parameter estimates, and the standard error of the estimates. The final two columns of the table provide the calculated t values and associated probabilities (p -values) of obtaining a larger absolute t value. Each p -value is less than 0.05; thus, all parameter estimates are significant at the 5% level. The fitted equation for this model is as follows:

$$\text{oxygen} = 111.718 - 0.256 \times \text{age} - 2.825 \times \text{runtime} - 0.131 \times \text{runpulse}$$

Figure 11.11 displays the confidence limits for the parameter estimates and the table of collinearity diagnostics.

The screenshot shows a SAS Analysis window with two tables. The first table, 'Parameter Estimates', shows the coefficients and 95% confidence limits for the intercept and three predictors: age, runtime, and runpulse. The second table, 'Collinearity Diagnostics', shows the eigenvalues, condition indices, and the proportion of variation accounted for by each parameter in the model.

		Parameter Estimates			
Variable	Label	DF	95% Confidence Limits		
Intercept	Intercept	1	90.71740	132.71873	
age	Age in years	1	-0.45384	-0.05895	
runtime	Min. to run 1.5 miles	1	-3.56051	-2.09025	
runpulse	Heart rate while running	1	-0.23471	-0.02711	

Collinearity Diagnostics						
Number	Eigenvalue	Condition Index	-----Proportion of Variation-----			
			Intercept	age	runtime	runpulse
1	3.97790	1.00000	0.00011565	0.00056585	0.00082368	0.00016363
2	0.01183	18.33958	0.00296	0.38305	0.49678	0.00697
3	0.00919	20.80033	0.03198	0.19423	0.42448	0.09749
4	0.00108	60.60078	0.96495	0.42215	0.07792	0.89538

Figure 11.11. Linear Regression: Confidence Limits and Collinearity Analysis

The collinearity diagnostics table displays the eigenvalues, the condition index, and the corresponding proportion of variation accounted for in each estimate. Generally, when the condition index is around 10, there are weak dependencies among the regression estimates. When the index is larger than 100, the estimates may have a large amount of numerical error. The diagnostics displayed in Figure 11.11, though indicating unfavorable dependencies among the estimates, are not so excessive as to dismiss the model.

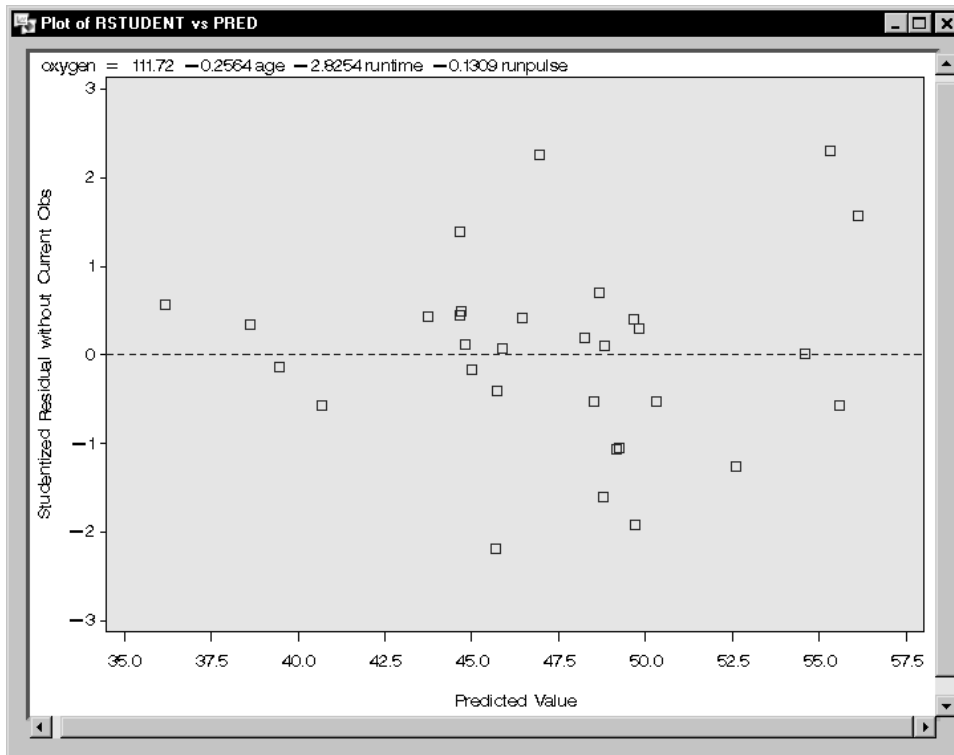


Figure 11.12. Linear Regression: Plot of Studentized Residuals versus Predicted Values

The plot of the studentized residuals versus the predicted values is displayed in Figure 11.12. When a model provides a good fit and does not violate any model assumptions, this type of residual plot exhibits no marked pattern or trend. Figure 11.12 exhibits no such trend, indicating an adequate fit.

Logistic Regression

Logistic regression enables you to investigate the relationship between a categorical outcome and a set of explanatory variables. The outcome, or response, can be dichotomous (yes, no) or ordinal (low, medium, high). When you have a dichotomous response, you are performing standard logistic regression. When you are modeling an ordinal response, you are fitting a proportional odds model.

You can express the logistic model for describing the variation among probabilities $\{\theta_h\}$ as

$$\theta_h = \{1 + \exp[-\alpha - \sum_{k=1}^t \beta_k x_{hk}]\}^{-1}$$

where α is the intercept parameter, β is a vector of t regression parameters, and \mathbf{x}'_h is a row vector of explanatory variables corresponding to the h th subpopulation.

You can show that the odds of success for the h th group are

$$\frac{\theta_h}{1 - \theta_h} = \exp\{\alpha + \sum_{k=1}^t \beta_k x_{hk}\}$$

By taking logs on both sides, you obtain a linear model for the *logit*:

$$\log\left\{\frac{\theta_h}{1 - \theta_h}\right\} = \alpha + \sum_{k=1}^t \beta_k x_{hk}$$

This is the log odds of success to failure for the h th subpopulation. A nice property of the logistic model is that all possible values of $(\alpha + \mathbf{x}'_h\beta)$ in $(-\infty, \infty)$ map into $(0, 1)$ for θ_h . Note that $\exp\{\beta_k\}$ are the odds ratios. Maximum likelihood methods are used to estimate α and β .

In a study on the presence of coronary artery disease, walk-in patients at a clinic were examined for symptoms of coronary artery disease. Investigators also administered an ECG. Interest lies in determining whether there is a relationship between presence or absence of coronary artery disease and ECG score and gender of patient. Logistic regression is the appropriate tool for such an investigation.

The data set analyzed in this example is called **Coronary2**. It contains the following variables:

sex	sex (m or f)
ecg	ST segment depression (low, medium, or high)
age	patient age
ca	disease (yes or no)

The task includes performing a logistic analysis to determine an appropriate model.

Open the Coronary2 Data Set

The data are provided in the Analyst Sample Library. To open the **Coronary2** data set, follow these steps:

1. Select **Tools** → **Sample Data . . .**
2. Select **Coronary2**.
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 **Coronary2** from the list of members.
7. Click **OK** to bring the **Coronary2** data set into the data table.

Request the Logistic Regression Analysis

To request the logistic regression analysis, follow these steps:

1. Select **Statistics** → **Regression** → **Logistic** . . .
2. Ensure that **Single trial** is selected as the **Dependent type**.
3. Select **ca** from the candidate list as the dependent variable.
4. Select **ecg** and **sex** from the candidate list as the class variables.
5. Select **age** from the candidate list as the quantitative variable.
6. Select **yes** from the drop-down list for **Model Pr{ }**:

Note that **Model Pr{ }**: determines which value of the dependent variable the model is based on; usually, the value representing an event (such as yes or success) is chosen.

Figure 11.13 displays the resulting dialog.

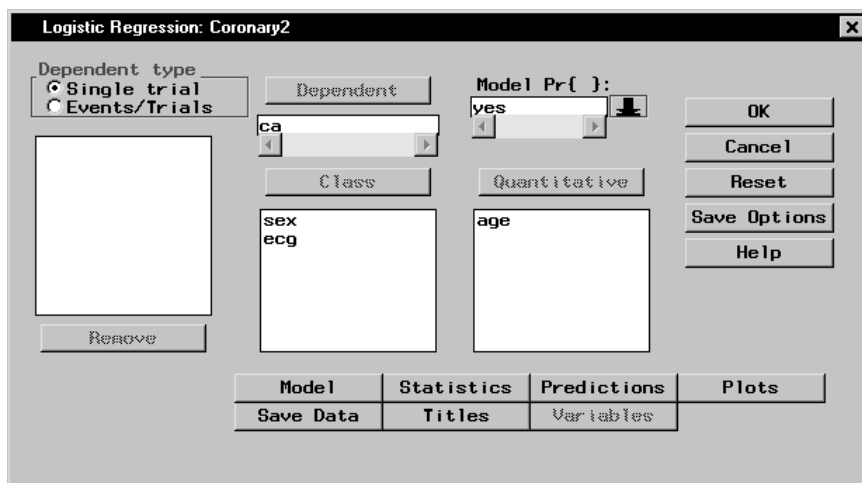


Figure 11.13. Logistic Regression Dialog

Specify the Model

By default, a main effects model is fit. To define a different model, with terms such as interactions, or to specify various model selection methods, such as forward selection or backward elimination, use the Model dialog.

To specify a forward selection model with main effects and their interactions, follow these steps:

1. Click on the **Model** button in the main dialog.
2. Highlight the variables **age**, **ecg**, and **sex** in the **Explanatory:** list of the model dialog.
3. Click on the **Factorial** button to specify main effects and their interactions.

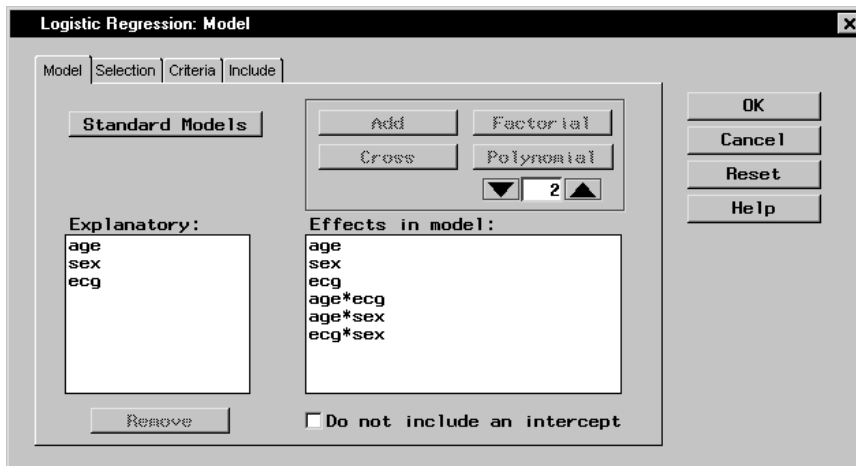


Figure 11.14. Logistic Regression: Model Dialog, Model Tab

Figure 11.14 displays the Model dialog with the terms **age**, **ecg**, **sex**, and their interactions 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 pop-up list. From this list, you can request that your model include main effects only or effects up to two-way interactions.

Now, to specify your model-building technique, follow these steps:

1. Click on the **Selection** tab.
2. Select **Forward selection**. The forward selection technique starts with a default model and adds significant variables to the model according to the specified criteria.
3. To specify which variables to include in every model, click on the **Include** tab, and select the variables **age**, **ecg**, and **sex**.
4. Click **OK**.

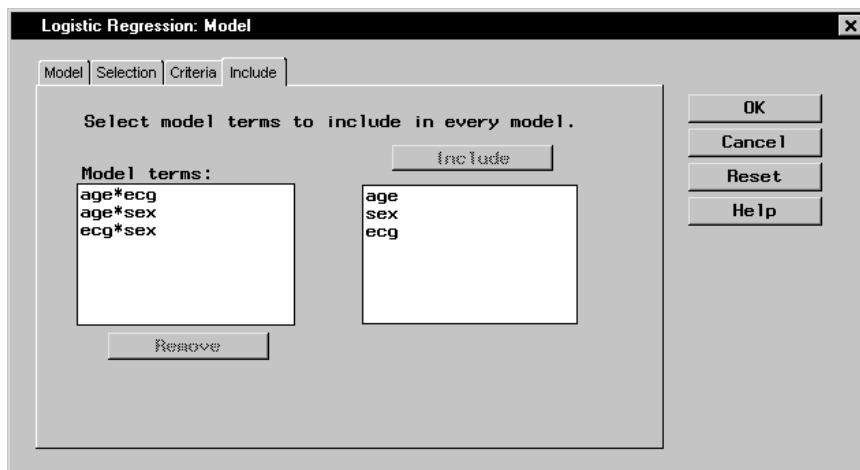


Figure 11.15. Logistic Regression: Model Dialog, Include Tab

Figure 11.15 displays the **Include** tab with the terms **age**, **ecg**, and **sex** selected as model terms to be included in every model.

When you have completed your selections, click **OK** in the main dialog to produce your analysis.

Review the Results

Figure 11.16 displays the “Testing Global Null Hypothesis: BETA = 0” table, which lists statistics that test whether the parameters are collectively equal to zero. This is similar to the overall F statistic in a regression model.

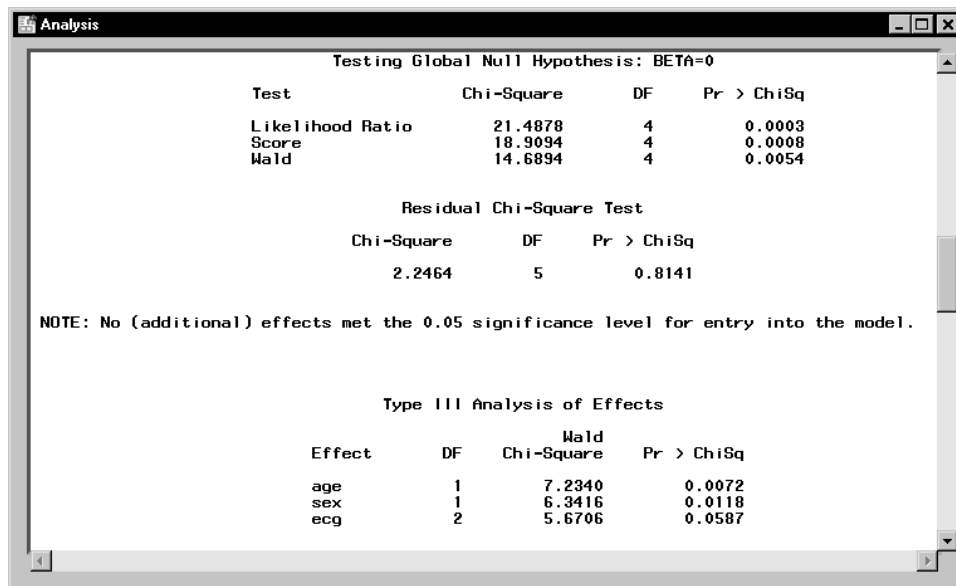


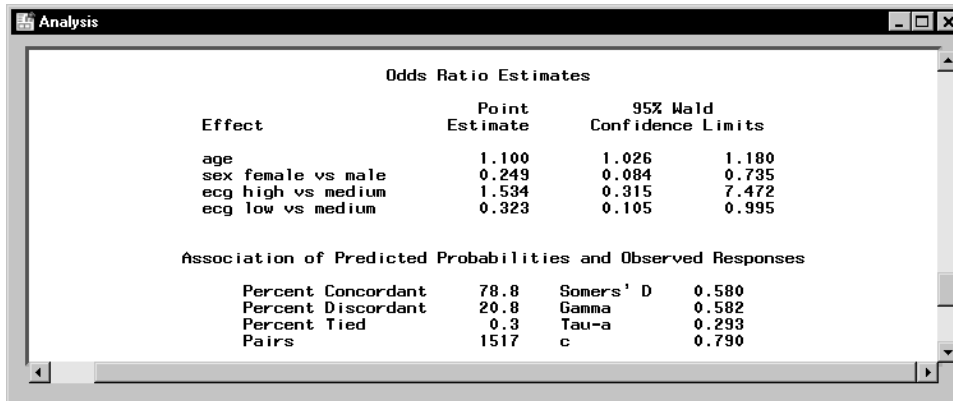
Figure 11.16. Logistic Regression: Analysis Results

When the explanatory variables in a logistic regression are relatively small in number and are qualitative, you can request a goodness-of-fit test. However, when you also have quantitative variables, the sample size requirements for these tests are not met. An alternative strategy for testing goodness of fit in this case is to examine the residual score statistic. This criterion is based on the relationship of the residuals of the model with other potential explanatory variables. If an association exists, then the additional explanatory variable should also be included in the model. This test is distributed as chi-square, with degrees of freedom equal to the difference in the number of parameters in the original model and the number of parameters in the expanded model.

The residual score statistic is displayed in Figure 11.16 as the “Residual Chi-Square Test” table. Since the difference in the number of parameters for the expanded model and the original model is $9 - 4 = 5$, the score statistic has 5 degrees of freedom. Since the value of the statistic is 2.24 and the p -value is 0.81, the main ef-

fects model fits adequately and no additional interactions need to be added.

The “Type III Tests of Effects” table provides Wald chi-square statistics that indicate that both **age** and **sex** are clearly significant at the $\alpha = 0.05$ level of significance. The **ecg** variable approaches significance, with the Wald statistic of 5.67 and $p = 0.059$. Although you may want to delete the **ecg** variable because it does not meet the $\alpha = 0.05$ significance criteria, there may be reasons for keeping it.



The screenshot shows a window titled "Analysis" with two tables. The first table, "Odds Ratio Estimates", lists effects and their point estimates and 95% Wald confidence limits. The second table, "Association of Predicted Probabilities and Observed Responses", lists various association statistics.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
age	1.100	1.026	1.180
sex female vs male	0.249	0.084	0.735
ecg high vs medium	1.534	0.315	7.472
ecg low vs medium	0.323	0.105	0.995

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	78.8	Somers' D	0.580
Percent Discordant	20.8	Gamma	0.582
Percent Tied	0.3	Tau-a	0.293
Pairs	1517	c	0.790

Figure 11.17. Logistic Regression: Analysis Results

Figure 11.17 displays odds ratio estimates and statistics describing the association of predicted probabilities and observed responses. The value of 1.10 for **age** is the extent to which the odds of coronary heart disease increase each year. The odds ratio for **sex**, 0.249, is the odds for females relative to males adjusted for **age** and **ecg**. Thus, the odds of coronary heart diseases for females are approximately one-fourth that of males.

References

- Freund, Rudolf J. and Littell, Ramon C. (1991), *SAS System for Regression, Second Edition*, Cary, NC: SAS Institute Inc.
- SAS Institute Inc. (1999), *SAS/STAT User's Guide, Version 7-1*, Cary, NC: SAS Institute Inc.
- Stokes, Maura E., Davis, Charles S., and Koch, Gary G. (1995), *Categorical Data Analysis Using the SAS System*, Cary, NC: SAS Institute Inc.

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.