Learning to Interpret Residual Plots in SAS for Regression Diagnostics

Authored by

Mohammed looti

October 27, 2025

Residual plots are fundamental diagnostic tools in **regression analysis**, offering crucial insights into the validity of a statistical model's underlying assumptions. They provide a visual assessment of whether the **residuals**, which represent the errors in prediction, are normally distributed and whether they exhibit **homoscedasticity** (constant variance).

The primary purpose of examining a residual plot is to diagnose potential problems such as non-linearity, heteroscedasticity, or outliers that could invalidate the inferences drawn from the regression model. A well-behaved residual plot indicates that the model assumptions are likely met, leading to more reliable parameter estimates and predictions.

Understanding Residual Plots in Regression Analysis

In any **linear regression model**, the goal is to predict a dependent variable based on one or more independent variables. The difference between the observed value and the value predicted by the model for each data point is known as a **residual**. These residuals are essentially the "leftover" variation after the model has accounted for the relationships between the variables.

A key assumption in ordinary least squares (OLS) linear regression is that these residuals should be randomly distributed with a mean of zero, constant variance, and follow a normal distribution. Violations of these assumptions can lead to biased standard errors, incorrect p-values, and ultimately, misleading conclusions about the relationships between variables.

A residual plot typically displays the residuals on the y-axis against the **predicted values** (or sometimes the independent variable) on the x-axis. By visualizing this relationship, we can quickly identify patterns that suggest problems with the model, such as a funnel shape indicating heteroscedasticity or a curved pattern suggesting a non-linear relationship that the model has not captured.

Basic Syntax for Generating a Residual Plot in SAS

SAS provides powerful procedures for fitting regression models and generating diagnostic plots. The `PROC REG` procedure is the go-to tool for performing regression analysis. Within `PROC REG`, the `PLOT` statement is used to request various diagnostic graphs, including the residual plot.

To create a residual plot showing residuals versus predicted values, you use `plot residual. * predicted.;`. Additionally, the `SYMBOL` statement allows for customization of the plot points, which can enhance readability and aesthetic appeal.

Below is the basic syntax for fitting a regression model and producing a residual plot in SAS. The `symbol value = circle;` statement, for instance, specifies that the points in the residual plot should

be displayed as circles, overriding the default plus sign.

symbol value = circle;

proc reg data=my_data;
model y = x;
plot residual. * predicted.;
run;

Step-by-Step Example: Creating a Residual Plot

Let's walk through a practical example to demonstrate how to generate a residual plot in SAS. We will begin by creating a sample dataset that includes an independent variable `x` and a dependent variable `y`. This dataset will serve as the foundation for our **simple linear regression model**.

The following SAS code creates a dataset named `my_data` using the `DATA` and `INPUT` statements, followed by the `DATALINES` to input the observations. After the data is created, `PROC PRINT` is used to display the dataset, allowing us to review the raw data before proceeding with the regression analysis.

/*create dataset*/
data my_data;
input x y;
datalines;
8 41
12 42
12 39
13 37
14 35
16 39
17 45
22 46
24 39
26 49
29 55
30 57
;
run;

/*view dataset*/

proc print data=my_data;

Obs	x	y
1	8	41
2	12	42
3	12	39
4	13	37
5	14	35
6	16	39
7	17	45
8	22	46
9	24	39
10	26	49
11	29	55
12	30	57

Executing the Regression Model and Plotting Residuals

With our dataset prepared, we can now proceed to fit the **simple linear regression model** and generate the diagnostic residual plot. The `PROC REG` procedure is invoked, specifying `my_data` as the input dataset. The `MODEL` statement defines the relationship, with `y` as the dependent variable and `x` as the independent variable.

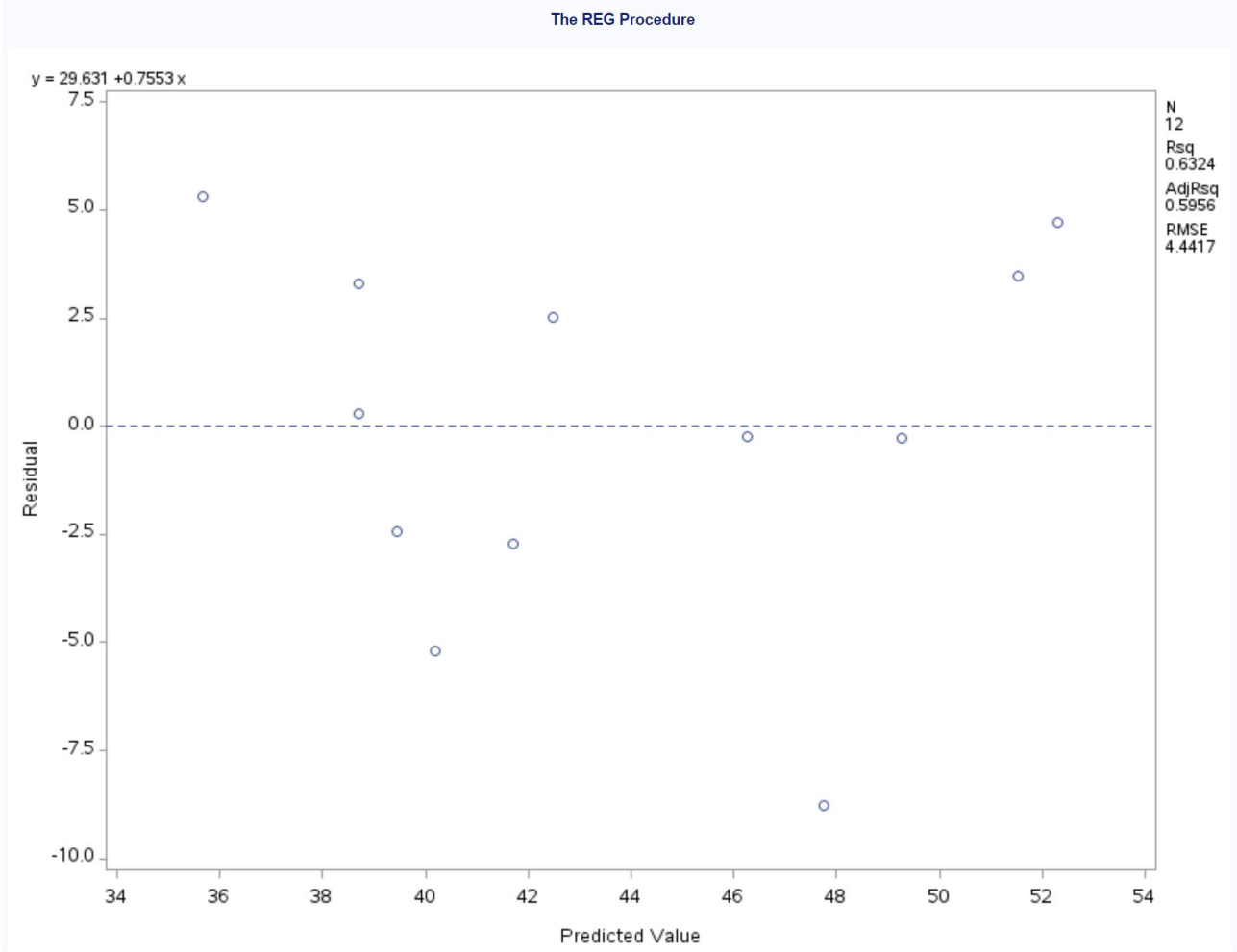
Crucially, the `PLOT residual. * predicted.;` statement is included to request the residual plot. This plot is essential for visually inspecting the model's assumptions regarding the **residuals**. The `SYMBOL value = circle;` line ensures that the data points on the plot are represented by circles, which can improve clarity compared to the default symbols.

Upon executing this syntax, SAS will produce a comprehensive output that includes the regression results, statistical tests, and, most importantly for our current objective, the residual plot. This plot is typically displayed at the bottom of the `PROC REG` output, providing an immediate visual assessment of the model's fit.

```
/*fit simple linear regression model and create residual plot*/
symbol value = circle;
proc reg data=my_data;
model y = x;
plot residual. * predicted.;
```

run;

The residual plot will be displayed at the bottom of the output:



Interpreting the Residual Plot Output

Interpreting the generated **residual plot** is a critical step in validating your **regression model**. In this plot, the x-axis represents the **predicted values** of the dependent variable (`y`), while the y-axis displays the **residuals** (the difference between observed and predicted values).

A healthy residual plot should show the residuals randomly scattered around the horizontal line at zero, with no discernible pattern. This indicates that the model has captured most of the systematic information in the data, and the remaining errors are random noise. If the residuals are randomly scattered about the value zero with no clear pattern of increasing or decreasing variance, the assumption of **homoscedasticity** (constant variance of errors) is met.

Conversely, if patterns emerge, they signal violations of regression assumptions. For instance, a "fan" or "funnel" shape suggests heteroscedasticity, where the variance of residuals changes across the range of predicted values. A curved pattern indicates that the linear model might not be appropriate, and a non-linear relationship could be present. Observing clusters or outliers can also point to influential data points or issues with the model's specification. In our example, the random scatter suggests a good model fit concerning the residual assumptions.

Analyzing Additional Regression Output Metrics

Beyond the visual insights provided by the residual plot, the `PROC REG` output also includes several important numerical metrics that quantify the model's performance and fit. These metrics are displayed alongside the plot and provide further evidence for evaluating the regression model.

Along the top of the plot, you can typically find the fitted regression equation. This equation explicitly states the relationship derived by the model, showing the intercept and the coefficient for each independent variable. For our example, this would be `Y = Intercept + Slope * X`.

On the right side of the plot, several key statistics are presented, offering a concise summary of the model's overall effectiveness:

N: This represents the **Total number of observations** used in fitting the model. In our case, N = 12.
Rsq: This is the **R-squared** value, also known as the coefficient of determination. It indicates the proportion of the variance in the dependent variable that is predictable from the independent variable(s). An R-squared of 0.6324 means that approximately 63.24% of the variability in `y` can be explained by `x`.
AdjRsq: This is the **Adjusted R-squared**. It is a modified version of R-squared that accounts for the number of predictors in the model. It is particularly useful when comparing models with different numbers of independent variables, as it penalizes for adding unnecessary predictors. An adjusted R-squared of 0.5956 suggests a slightly more conservative estimate of the explained variance, which is often preferred.
RMSE: This stands for the **Root Mean Squared Error**. It is a measure of the average magnitude of the errors. It tells us how concentrated the data is around the line of best fit. A smaller RMSE indicates a better fit. In this example, an RMSE of 4.4417 suggests that, on average, the observed values deviate from the predicted values by approximately 4.44 units.

Conclusion and Further Learning

In summary, creating and interpreting a **residual plot** in SAS is an indispensable step in conducting thorough **regression analysis**. It provides a visual confirmation of critical model assumptions, helping to ensure the reliability and validity of your statistical inferences. By following

the simple `PROC REG` and `PLOT` statements, you can quickly generate these diagnostic plots and assess the fit of your model.

A clear, randomly scattered residual plot, as demonstrated in our example, indicates a well-specified model where the assumptions of linearity and **homoscedasticity** are likely met. Conversely, patterns in the plot should prompt further investigation and potential model adjustments. Combining this visual assessment with an examination of key numerical metrics provides a robust evaluation of your regression model.

To deepen your understanding of SAS and statistical modeling, consider exploring other advanced functionalities and diagnostic tools. Continuous learning and practice are key to mastering data analysis with SAS.

Additional Resources

The following tutorials explain how to perform other common tasks in SAS: