Understanding Generalized Linear Model (GLM) Output in R: A Step-by-Step Guide

Authored by

Mohammed looti

November 1, 2025

Understanding the Generalized Linear Model (GLM) in R

The R statistical environment provides the powerful **glm()** function, which is the foundational tool used to fit generalized linear models. Unlike standard linear regression, GLMs allow the response variable to have an error distribution model other than a normal distribution, making them essential for analyzing counts, proportions, or binary outcomes.

A clear understanding of the **glm()** function's syntax is the first step toward effective statistical modeling in R. The function requires several key arguments to define the structure of the model and the underlying distribution of the data.

This function utilizes the following core syntax structure:

glm(formula, family=gaussian, data, ...)

The parameters outlined above specify the mathematical relationship and the dataset being analyzed:

formula: Defines the specific structure of the linear predictor (e.g., response variable `y` modeled by predictors `x1` and `x2` is written as `y ~ x1 + x2`).
family: Specifies the statistical family used to link the mean of the response variable to the predictors. The default is **gaussian** (equivalent to standard linear regression), but crucial alternatives include **binomial** (for binary data), **Gamma**, and **poisson** (for count data), among others.
data: The name of the data frame containing the variables referenced in the formula.

In practical data analysis, this function is most frequently employed to fit logistic regression models by specifying the **binomial** family, which is suitable for predicting the probability of a binary event. The subsequent sections will demonstrate how to interpret the complex output generated by R after fitting such a model.

Setting Up the Logistic Regression Example in R

To illustrate the interpretation of **glm()** output, we will construct a logistic regression model using the well-known built-in **mtcars** dataset in R, which contains various specifications and performance metrics for 32 automobiles. Our objective is to model a binary outcome based on continuous predictors.

Specifically, we will examine how engine displacement (**disp**) and gross horsepower (**hp**) predict the probability that a car has an automatic transmission (`am = 0`) versus a manual transmission (`am = 1`). To begin, it is good practice to inspect the structure of the data:

#view first six rows of *mtcars* dataset
head(mtcars)

mpg cyl disp hp drat wt qsec vs am gear carb
Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4
Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4
Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1
Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1
Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2
Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1

Our response variable, **am**, is binary, necessitating the use of the **family=binomial** argument within the **glm()** function. We are modeling the probability that a given car is coded as 1 (manual transmission) based on the size of its engine (disp) and its power (hp).

The following R code executes the model fit and generates the summary output that we need to interpret:

#fit logistic regression model
model <- glm(am ~ disp + hp, data=mtcars, family=binomial)

#view model summary
summary(model)

Call:
glm(formula = am ~ disp + hp, family = binomial, data = mtcars)

Deviance Residuals:
Min 1Q Median 3Q Max
-1.9665 -0.3090 -0.0017 0.3934 1.3682

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.40342 1.36757 1.026 0.3048
disp -0.09518 0.04800 -1.983 0.0474 *
hp 0.12170 0.06777 1.796 0.0725 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 43.230 on 31 degrees of freedom

Residual deviance: 16.713 on 29 degrees of freedom
AIC: 22.713

Number of Fisher Scoring iterations: 8

Examining the R Output: The Call and Residuals

Before diving into the core statistical tables, it is important to briefly review the initial lines of the summary output. The **Call** section simply restates the function and arguments used to fit the model, which is useful for verifying that the intended formula and statistical family were correctly applied. In this case, we confirm that **am** was modeled using **disp** and **hp**, assuming a **binomial** distribution.

Immediately following the Call, the **Deviance Residuals** provide a measure of how well the model fits each individual observation. Unlike standard least-squares residuals, deviance residuals are based on the contribution of each observation to the total deviance statistic. They should ideally be centered around zero and show a symmetric distribution, suggesting that the model assumptions hold true.

A glance at the summary statistics for the residuals--Min, 1Q, Median, 3Q, and Max--suggests a relatively centered distribution, as the Median value is very close to zero (-0.0017). Extreme values, such as the minimum (-1.9665) and maximum (1.3682), help identify potential outliers or observations that the model struggled to predict accurately. However, the true test of model fit lies in the coefficients and the overall deviance statistics.

Interpreting Model Coefficients and P-Values

The **Coefficients** table is the heart of the GLM output, providing the estimated effect of each predictor on the response variable. For a logistic regression, the **coefficient estimate** indicates the average change in the log odds of the response variable (in our case, `am = 1`, or manual transmission) associated with a one-unit increase in the respective predictor variable, holding all other predictors constant.

Considering the predictor **disp** (displacement), the coefficient estimate is -0.09518. This negative value signifies that a one-unit increase in engine displacement is associated with an average change of -0.09518 in the log odds of having a manual transmission. In practical terms, higher engine displacement is associated with a lower likelihood of the car having a manual transmission, which is often characteristic of larger, less sporty vehicles. Conversely, the **hp** (horsepower) coefficient of 0.12170 suggests that higher horsepower is positively associated with the likelihood of having a manual transmission, aligning with the common design of performance vehicles.

To determine the statistical significance of these effects, we examine the Z-value and the associated p-value, denoted as **Pr(>|z|)**. The **z value** is calculated by dividing the Estimate by the Standard Error (e.g., for **disp**, Z = -0.09518 / 0.04800 ≈ -1.983). This Z-score represents the test statistic for the null hypothesis that the true coefficient is zero (meaning the predictor has no effect).

The **p-value** tells us the probability of observing a Z-score as extreme as the one calculated, assuming the null hypothesis is true. If the p-value is below a chosen significance level (alpha, typically 0.05), we reject the null hypothesis, concluding that the predictor is statistically significant. For example, the p-value for **disp** is 0.0474. Since 0.0474 is less than the standard 0.05 threshold, we conclude that engine displacement is a statistically significant predictor of transmission type in this model. The **hp** variable, with a p-value of 0.0725, is marginally significant (often denoted by a dot '.') but does not meet the strict 0.05 criterion. Analysts often use significance levels of 0.01, 0.05, or 0.10, depending on the field and desired rigor.

Assessing Model Fit: Null and Residual Deviance

For generalized linear models, the concept of **deviance** replaces the sum of squared errors used in ordinary least squares regression. Deviance essentially measures the disagreement between the current model and a perfectly saturated model (one that perfectly predicts the data). Lower deviance values indicate a better fit to the observed data.

The **null deviance** shown in the output represents the deviance of the simplest possible model--one that includes only an intercept term and no predictor variables. This serves as a baseline for comparison, indicating how much variation is present in the response variable before any predictors are introduced. The null deviance is calculated with $N-1$ degrees of freedom, where N is the number of observations.

The **residual deviance**, in contrast, shows the deviance of the specific model we have fitted, which includes our p predictor variables (**disp** and **hp**). The residual deviance is calculated with $N-p-1$ degrees of freedom. The goal of fitting the model is to significantly reduce the deviance from the null model. A lower residual deviance relative to the null deviance signifies that the included predictors successfully explain a meaningful amount of the variability in the response.

To formally test whether our model is "useful"--meaning it performs significantly better than the intercept-only null model--we calculate the Chi-Square statistic (X^2), which follows a Chi-Square distribution:

X2 = Null deviance - Residual deviance

This statistic has degrees of freedom equal to the difference in the degrees of freedom between

the two models (which is equal to p, the number of predictors added).

Applying this to our example:

Null deviance: 43.230 with df = 31
Residual deviance: 16.713 with df = 29

The calculation yields:

X2 = 43.230 - 16.713
X2 = 26.517

Since we have $p = 2$ predictor variables (disp and hp), the degrees of freedom for this test is 2. By comparing this X^2 value (26.517) with the Chi-Square distribution, we find that the associated p-value is extremely small (approximately 0.000002). Because this p-value is far less than 0.05, we confidently conclude that the model containing **disp** and **hp** is significantly better at predicting transmission type than a model relying solely on the intercept.

Using the Akaike Information Criterion (AIC) for Comparison

The Akaike information criterion (**AIC**) is a key metric used for selecting the best model among a set of competing models. It balances the model's goodness of fit (how well it explains the data) against the model's complexity (the number of parameters used). The principle is simple: when comparing several regression models, the model with the **lowest AIC value** is generally preferred as it is considered the best trade-off between bias and variance.

The AIC is mathematically derived using the following formula:

AIC = 2K - 2*ln*(L)

Where the components are defined as:

K: Represents the number of parameters in the model. In our example (intercept + disp + hp), K=3.
***ln*(L)**: Represents the log-likelihood of the model. The log-likelihood is a measure of how probable the observed data is, given the fitted model parameters.

It is important to note that the absolute value of the AIC (22.713 in our output) is not interpretable on its own. Its utility is strictly comparative. If an analyst were to fit an alternative model--say, predicting **am** only using **disp**--they would calculate a new AIC. If the new AIC were lower than 22.713, that simpler model would be deemed superior according to the AIC criterion.

Summary and Additional Resources

Interpreting the output of the **glm()** function requires careful attention to both the coefficient table, which details the relationship between individual predictors and the response (via the log odds), and the deviance statistics, which quantify the overall explanatory power of the model. We determined that both engine displacement (disp) and horsepower (hp) contribute significantly to the predictive power of the model, which is highly useful compared to a null model.

Mastering the interpretation of generalized linear models is essential for advanced statistical analysis in R. For those seeking further practical guidance on implementation and common issues, the following resources are recommended.

The following tutorials provide additional information on how to use the **glm()** function in R:

The following tutorials explain how to handle common errors when using the **glm()** function: