How to Run and Interpret Hierarchical Regression in Stata

Authored by

stats writer

December 28, 2025

Hierarchical regression in **Stata** is a specialized statistical technique used extensively in fields like psychology and economics to analyze the cumulative effects of multiple predictor variables on a continuous or categorical dependent variable. Unlike standard multiple regression, this procedure involves the careful construction of regression models in a sequence of steps, or blocks. The analysis starts with a base model (Step 1) and systematically adds sets of additional **independent variables** in subsequent steps. This structured approach allows the researcher to gain critical insight into the unique contribution of each block of predictors to the overall model fit, beyond the variables already included. Furthermore, it provides a robust opportunity to assess the overall predictive accuracy and identify potential interactions or suppressions among the variables introduced sequentially.

Theoretical Foundation: Why Use Staged Modeling?

Hierarchical regression (sometimes referred to as sequential regression) is fundamentally a method for comparing nested **linear regression models**. The underlying purpose is hypothesis testing: determining whether adding a new set of predictors significantly improves the explanation of variance in the outcome variable, above and beyond the existing predictors. This methodology is particularly powerful when theory dictates the order in which variables should be entered into the analysis, allowing for precise evaluation of theoretical constructs and specific hypotheses regarding predictor importance.

The core principle involves first fitting a basic linear regression model, typically containing demographic or control variables. Subsequently, a second regression model is fit, incorporating one or more additional explanatory variables (the variables of primary interest). The improvement is quantified by comparing the coefficient of determination, commonly known as the R-squared, between the two models. If the change in the R-squared (often denoted as ΔR^2) in the second model is statistically significantly higher than that of the previous model, it provides strong evidence that the newly added variables offer substantial unique explanatory power.

This iterative process is repeated, fitting additional regression models by adding further blocks of explanatory variables and statistically testing whether these newer models offer a substantial improvement over the prior, simpler models. This rigorous comparison relies on the statistical test of the change in **R-squared** (ΔR^2), which follows an F-test distribution. Understanding the significance of the change in **R-squared** is paramount to drawing methodologically sound conclusions about variable importance.

Prerequisites and Loading the Example Dataset in Stata

To effectively illustrate the application of hierarchical regression, we will utilize a widely available, built-in dataset in **Stata** known as *auto*. This dataset contains comprehensive information on various attributes of 74 different automobiles, making it an ideal candidate for modeling price based on physical and mechanical specifications. To load this dataset and prepare our environment, execute the following command in the **Stata** Command box:

sysuse auto

Once the dataset is active, it is standard practice to quickly summarize the variables and observations available. This helps confirm the data structure, variable types, and missing values before starting the demanding modeling process. Use the following command to generate a descriptive summary:

summarize

```
. sysuse auto
(1978 Automobile Data)

. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
make	0				
price	74	6165.257	2949.496	3291	15906
mpg	74	21.2973	5.785503	12	41
rep78	69	3.405797	.9899323	1	5
headroom	74	2.993243	.8459948	1.5	5
trunk	74	13.75676	4.277404	5	23
weight	74	3019.459	777.1936	1760	4840
length	74	187.9324	22.26634	142	233
turn	74	39.64865	4.399354	31	51
displacement	74	197.2973	91.83722	79	425
gear_ratio	74	3.014865	.4562871	2.19	3.89
foreign	74	.2972973	.4601885	0	1

The output of this command confirms the data structure, showing details about the number of observations (74 cars) and key descriptive statistics for the 12 available variables. We can proceed knowing we have a clean dataset containing variables such as **price**, **mpg**, **weight**, and **gear ratio**, which will be central to our analysis.

Defining the Research Models for Comparison

For this demonstration, we will define three sequential linear regression models where the dependent variable is the **price** of the vehicle. Our primary goal is to determine if the additional predictors introduced in Model 2 and Model 3 significantly enhance the prediction of car price over their predecessors, justifying their complexity. We will systematically test the influence of **mpg** (miles per gallon), **weight** (car weight), and **gear ratio**.

The three structured models are formulated as follows:

Model 1 (Base Model): price = intercept + mpg

Model 2 (Adding Weight): price = intercept + mpg + weight

Model 3 (Adding Gear Ratio): price = intercept + mpg + weight + gear ratio

The hierarchical regression approach will allow us to specifically isolate and test the incremental contribution of **weight** (by comparing Model 2 vs. Model 1) and subsequently, the incremental contribution of **gear ratio** (by comparing Model 3 vs. Model 2). This ensures that we are testing the unique explanatory power of each variable block beyond what has already been accounted for.

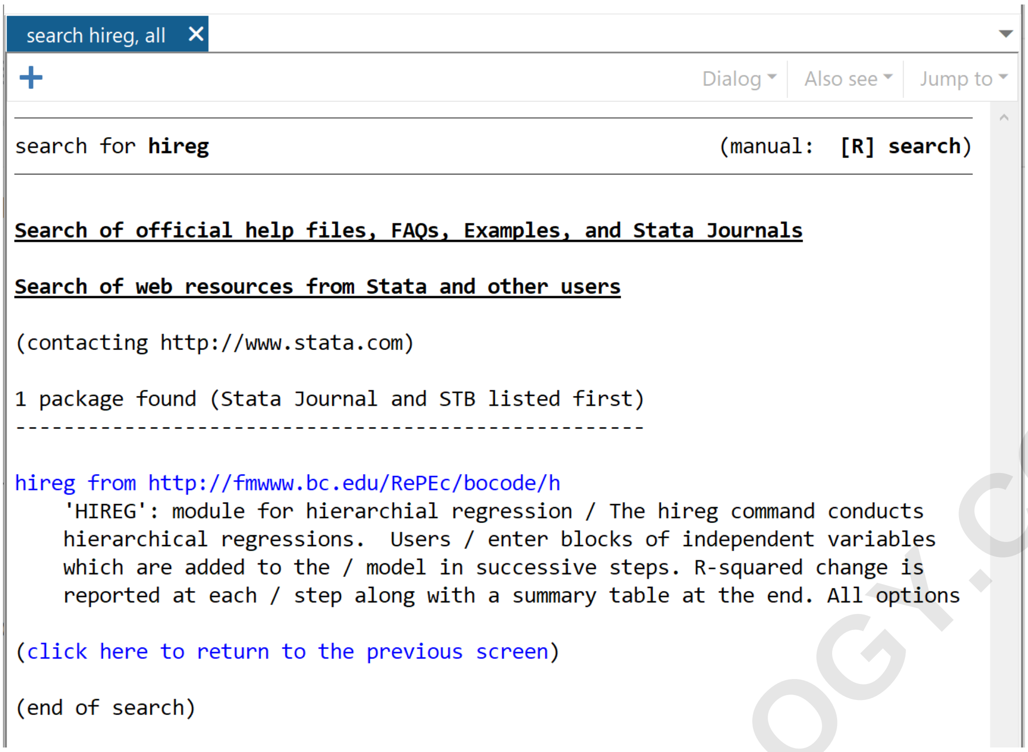
Installing the Essential `hireg` Package

To perform hierarchical regression efficiently and correctly in **Stata**, we must utilize a user-contributed external package known as **hireg**. Unlike built-in commands like `regress`, **hireg** is specifically designed to handle the sequential comparison of nested models, automatically calculating and displaying the critical statistical information, particularly the incremental F-test for the change in **R-squared** (ΔR^2).

Since this package is not part of the standard installation, we must first locate and install it. To begin this process, enter the following command into the **Stata** Command box:

findit hireg

Executing the **findit** command will open a separate viewer window providing links to the package's location, typically the Statistical Software Components (SSC) archive maintained by users.



In the resulting window, look for the link that specifies **click here to install** the package. Click this link to initiate the automated download, compilation, and installation of the **hireg** package onto your system.

```
net describe hireg, from(http://f...
Dialog   Also see   Jump to

package hireg from http://fmwww.bc.edu/RePEc/bocode/h

TITLE
      'HIREG': module for hierarchial regression

DESCRIPTION/AUTHOR(S)

      The hireg command conducts hierarchical regressions.  Users
      enter blocks of independent variables which are added to the
      model in  successive steps. R-squared change is reported at each
      step along with a summary  table at the end. All options
      available on the regress command may also be used with hireg.

      KW: regression
      KW: hierarchical regression

      Requires: Stata version 8.0


      Author: Paul H. Bern, Princeton University
      Support: email  pbern@princeton.edu

      Distribution-Date: 20050723

INSTALLATION FILES                                 (click here to install)
      hireg.ado
      hireg.hlp

(click here to return to the previous screen)
```


The installation should complete within a few seconds. Once **Stata** confirms the successful installation, we are prepared to execute the specialized hierarchical regression command using the defined model structure.

Executing the Hierarchical Analysis Command

The syntax of the **hireg** command is constructed to reflect the sequential nature of the analysis. It requires the dependent variable first, followed by the independent variables grouped in parentheses, where each group represents a step or block of variables added to the previous model.

To execute the three-step hierarchical regression defined in our structure (price predicted by mpg, then weight, then gear ratio), use the following precise command:

hireg price (mpg) (weight) (gear_ratio)

This single command instructs **Stata** to perform three distinct regression analyses and the necessary comparisons:

Step 1: Runs Model 1 using **price** as the response variable, predicted only by the variable **mpg**.

Step 2: Runs Model 2, adding the variable **weight** to the predictors from Step 1, and then statistically compares Model 2 to Model 1.

Step 3: Runs Model 3, adding the variable **gear_ratio** to the predictors from Step 2, and then statistically compares Model 3 to Model 2.

Interpreting Model 1 and Model 2 Results

The output generated by the **hireg** command systematically displays the results for each stage. We begin by analyzing the results of the base model, Model 1.

```
Model 1:
   Variables in Model:
   Adding             : mpg

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(1, 72)        =     20.26
       Model |  139449474         1  139449474     Prob > F        =    0.0000
    Residual |  495615923        72  6883554.48    R-squared       =    0.2196
-------------+----------------------------------   Adj R-squared   =    0.2087
       Total |  635065396        73  8699525.97    Root MSE        =    2623.7

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |  -238.8943   53.07669    -4.50   0.000    -344.7008   -133.0879
       _cons |   11253.06   1170.813     9.61   0.000     8919.088    13587.03
------------------------------------------------------------------------------
```


For Model 1, the overall coefficient of determination, the **R-squared**, is reported as **0.2196**. This means that **mpg** alone accounts for 21.96% of the variance observed in car price. The model's overall statistical significance is confirmed by the **Prob > F** value, which is **0.0000**. Since this p-value is highly significant (well below α = 0.05), we confirm that Model 1 provides a predictive relationship.

Moving to Model 2, we evaluate the impact of including the variable **weight** alongside **mpg**:

```
Model 2:
   Variables in Model: mpg
   Adding            : weight

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(2, 71)        =     14.74
       Model |  186321280         2  93160639.9   Prob > F        =    0.0000
    Residual |  448744116        71  6320339.67   R-squared       =    0.2934
-------------+----------------------------------   Adj R-squared   =    0.2735
       Total |  635065396        73  8699525.97   Root MSE        =      2514

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |  -49.51222   86.15604    -0.57   0.567    -221.3025     122.278
      weight |   1.746559   .6413538     2.72   0.008      .467736    3.025382
       _cons |   1946.069    3597.05     0.54   0.590    -5226.245    9118.382
------------------------------------------------------------------------------
R-Square Diff. Model 2 - Model 1 = 0.074   F(1,71) =   7.416  p = 0.008
```

The overall **R-squared** for Model 2 increases to **0.2934**. This numerical improvement suggests that adding **weight** explains an additional portion of the variance. To confirm whether this added explanatory power is statistically robust, we must examine the incremental F-test results comparing Model 2 to Model 1, which are conveniently presented at the end of the Model 2 output:

The **R-squared difference** (ΔR^2) between the two models is **0.074**, indicating a 7.4% increase in explained variance.

The corresponding **F-statistic for the difference** is **7.416**.

The associated **p-value of the F-statistic** is **0.008**.

Given that the p-value (0.008) is substantially less than the standard 0.05 significance level, we confidently conclude that the inclusion of **weight** in Model 2 resulted in a **statistically significant improvement** in the predictive power compared to Model 1.

Interpreting the Final Model Comparison and Conclusion

The final step involves Model 3, where **gear_ratio** is introduced as the last predictor, testing its unique contribution beyond **mpg** and **weight**.

```
Model 3:
   Variables in Model: mpg  weight
   Adding            : gear_ratio

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(3, 70)        =     10.73
       Model |  200028459         3  66676153.1   Prob > F        =    0.0000
    Residual |  435036937        70  6214813.38   R-squared       =    0.3150
-------------+----------------------------------   Adj R-squared   =    0.2856
       Total |  635065396        73  8699525.97   Root MSE        =      2493

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |  -50.60979   85.43696    -0.59   0.556    -221.0084    119.7889
      weight |   2.390466   .7697099     3.11   0.003     .8553283    3.925604
  gear_ratio |   1459.319   982.6304     1.49   0.142    -500.4757    3419.113
       _cons |  -4374.457   5552.978    -0.79   0.433    -15449.52    6700.609
------------------------------------------------------------------------------
R-Square Diff. Model 3 - Model 2 = 0.022   F(1,70) =  2.206  p = 0.142
```

The overall **R-squared** for Model 3 marginally increases again to **0.3150**. Although this is the highest R-squared value numerically, we must strictly rely on the incremental F-test to determine if the variable **gear_ratio** offers unique, statistically significant predictive utility. The results comparing Model 3 to Model 2 are:

The **R-squared difference** (ΔR^2) between Model 3 and Model 2 is **0.022**.

The calculated **F-statistic for the difference** is **2.206**.

The corresponding **p-value of the F-statistic** is **0.142**.

Crucially, because the resulting p-value (0.142) is greater than the alpha level of 0.05, we must conclude that there is insufficient statistical evidence to assert that Model 3 offers a significant improvement in variance explanation over the prior Model 2. The variable **gear_ratio** does not uniquely contribute significantly to predicting car price once both fuel efficiency and weight are already accounted for.

The **Stata** output concludes with a highly useful summary table that synthesizes the critical results across all sequential steps:

Model	R2	F(df)	p	R2 change	F(df) change	p
1:	**0.220**	**20.258(1,72)**	**0.000**			
2:	**0.293**	**14.740(2,71)**	**0.000**	**0.074**	**7.416(1,71)**	**0.008**
3:	**0.315**	**10.729(3,70)**	**0.000**	**0.022**	**2.206(1,70)**	**0.142**

In summary, our Stata hierarchical regression analysis confirms that Model 2 provided a demonstrably significant statistical improvement over Model 1 ($p = 0.008$), validating the inclusion of car **weight**. Conversely, the addition of **gear ratio** in Model 3 did not justify the increased model complexity ($p = 0.142$). Therefore, Model 2 is the most parsimonious and effective model among the tested sequence for predicting car price.