Learning Hierarchical Regression Analysis Using Stata: A Comprehensive Tutorial

Authored by

Mohammed looti

November 9, 2025

Defining Hierarchical Regression and Its Theoretical Foundation

Hierarchical regression is a specialized and rigorous statistical methodology employed primarily within the framework of linear models. Its primary purpose is to systematically compare a nested series of models, allowing researchers to determine the unique explanatory power of sequentially added sets of variables. It is vital to understand that this technique is conceptually distinct from methods like multilevel modeling, focusing instead on the theory-driven organization of predictor variables.

The defining characteristic of hierarchical regression is that the researcher, not the statistical software, dictates the precise order in which variable blocks are entered into the analysis. This decision must be grounded in strong theoretical reasoning, established empirical evidence from previous studies, or a clear hypothesis regarding causal priority. By imposing this strict, predefined structure, the technique facilitates a highly structured evaluation of the incremental contribution provided by specific variable groupings, enabling researchers to isolate effects with precision.

The fundamental objective of this stepwise, theory-driven process is to rigorously test whether the inclusion of a new set of predictors--referred to as a "block"--yields a statistically significant improvement in the model's overall predictive capability, above and beyond the variables already accounted for in the preceding stages. This is an indispensable tool in disciplines like psychology, sociology, and economics, where investigators frequently aim to demonstrate that novel factors explain variance beyond the influence of known, established baseline determinants. The sequential addition of variables ensures that the incremental effect of each block is clearly isolated, thereby generating highly interpretable results concerning the model's overall sufficiency, relevance, and adherence to the principle of parsimony.

Statistical Pillars of Model Comparison: R-squared and the F-Test

The rigorous comparison of these nested models relies on accurately assessing alterations in the model's goodness-of-fit. The quintessential measure of fit within this context is the R-squared (R^2), which provides a quantitative estimate of the proportion of variance in the response variable that is successfully accounted for by the explanatory variables currently included in the model. Naturally, as additional predictors are incorporated into the equation, the R^2 value will typically increase. The core statistical inquiry in hierarchical regression, however, centers on whether this increase--known as the Delta R-squared (ΔR^2)--is genuinely substantial and statistically significant.

To effectively evaluate the significance of the ΔR^2, a specific model comparison F-test is employed. This F-statistic performs a hypothesis test where the null hypothesis states that the newly introduced set of predictors contributes absolutely nothing meaningful to the accurate prediction of the outcome variable. If the calculated F-statistic is sufficiently high, it will correspond to a very small p-value (conventionally set below the alpha level of 0.05).

A calculated p-value below this threshold mandates the rejection of the null hypothesis. Rejecting the null hypothesis confirms that the incremental change in R^2 is statistically significant, decisively proving that the newly added block of variables offers a meaningful and essential improvement to the model's predictive capability. Adherence to this strict statistical criterion is crucial, ensuring that researchers avoid the trap of inflating model complexity by indiscriminately adding predictors that offer negligible explanatory power.

Preparing the Data in Stata for Sequential Modeling

To demonstrate the practical application of hierarchical regression, we will employ the powerful statistical software package Stata. For this illustrative guide, we utilize Stata's accessible built-in sample dataset, auto, which contains comprehensive information on various attributes of 74 distinct automobiles. The essential first step in any quantitative analysis is successfully loading the required data into the active Stata session. This is achieved using the straightforward command below:

sysuse auto

After the dataset is loaded, standard practice dictates a preliminary summary of the variables. This step is crucial for gaining an immediate understanding of the data structure, confirming variable types, assessing the sample size, and obtaining an initial overview of the data distribution. This summary confirms that the variables designated for the regression are correctly formatted and ready for use.

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 summary confirms a sample size of 74 observations across 12 different variables. For our hierarchical regression demonstration, we will hypothesize that a car's price is influenced sequentially by its fuel efficiency, weight, and gear ratio. We aim to test the unique, incremental contribution of each of these factors. This theoretical framework leads us to define three distinct, nested linear regression models:

Model 1 (The Baseline): We aim to predict the outcome variable, **price**, using only the car's fuel efficiency, represented by the variable **mpg**.

price = intercept + mpg

Model 2 (The First Increment): We introduce the car's physical dimension, **weight**, to the predictors established in Model 1. This tests if size adds value beyond fuel efficiency.

price = intercept + mpg + weight

Model 3 (The Full Model): We subsequently introduce the mechanical specification, **gear ratio**, to the predictors already present in Model 2, evaluating its unique explanatory power.

price = intercept + mpg + weight + gear ratio

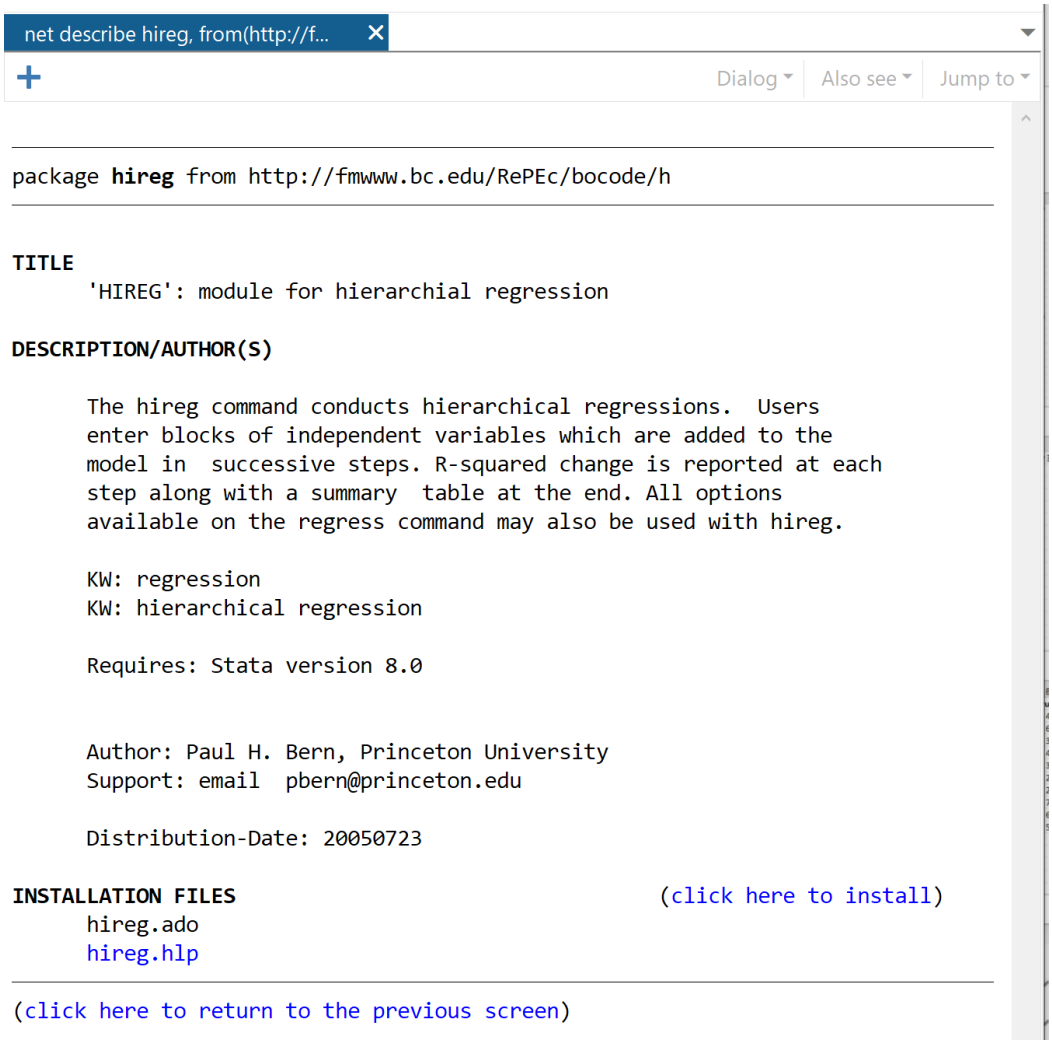
Implementing the Analysis Using the hireg User Command

It is important to note that standard versions of Stata do not include a native command specifically designed to present the output of hierarchical regression in the required comparative format,

displaying the change statistics (Delta R-squared and the associated F-test). Therefore, to execute this sophisticated analysis efficiently and obtain the essential comparison metrics, we must install a user-written package. The universally recommended community package for this specific task is **hireg**. To initiate the installation, Stata users must first utilize the `findit` command, which searches Stata's extensive online resources for the requested program:

findit hireg

Within the resulting Stata Viewer window, users should locate and click the link prominently labeled "click here to install" to seamlessly integrate the **hireg** package into their local Stata environment. This installation procedure is typically completed rapidly and securely.



Once the **hireg** package is successfully installed, executing the complex hierarchical regression becomes remarkably simple. The syntax is intuitive: it requires the dependent variable to be specified first, followed by the blocks of independent variables, each enclosed within its own set of parentheses. Each parenthetical group represents a block of variables that will be added sequentially to the model, in the order specified by the researcher's theory.

hireg price (mpg) (weight) (gear_ratio)

This single command elegantly instructs Stata to run the three theoretically defined nested models. It treats **price** as the common response variable across all stages, dictating the following exact steps: First, the baseline model includes only **mpg**. Second, the subsequent model adds **weight** to **mpg**, thereby assessing the unique incremental contribution of vehicle weight. Finally, the third model adds **gear_ratio** to the existing set of predictors (mpg and weight), evaluating the final incremental contribution.

Interpreting Model Results: Evaluating Incremental Contribution

The comprehensive output generated by the `hireg` command meticulously presents the results for each stage of the regression analysis sequentially, starting with the establishment of the baseline model.

Model 1:
Variables in Model:
Adding : **mpg**

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

Model 1, which incorporates only **mpg**, serves as the critical foundation against which all subsequent models will be compared. Reviewing the output, we observe that the R-squared value is calculated at **0.2196**. This finding indicates that approximately 22% of the observed variance in a car's price can be successfully explained by its miles per gallon rating alone. Furthermore, the overall test of the model's utility, summarized by the Prob > F, is reported as **0.0000**. Since this p-value is significantly smaller than the conventional alpha threshold of 0.05, we confidently conclude that Model 1 is statistically significant, confirming that mpg is a useful, non-trivial predictor of price.

The second stage of the hierarchical regression introduces **weight** into the model, systematically building upon the predictors already incorporated in Model 1. This step is designed to rigorously test the hypothesis that vehicle weight explains a significant and unique portion of the remaining variance in price that was not accounted for exclusively by fuel efficiency.

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

Source	SS	df	MS		
Model	**186321280**	**2**	**93160639.9**	Number of obs	= **74**
Residual	**448744116**	**71**	**6320339.67**	F(2, 71)	= **14.74**
				Prob > F	= **0.0000**
				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 R-squared for Model 2 increases noticeably to **0.2934**, a value numerically superior to the R-squared of 0.2196 achieved in Model 1. However, the most critical data for hierarchical analysis are the change statistics, presented at the bottom of the output, which directly compare Model 2 to Model 1: The R-squared difference (ΔR^2) is **0.074**; the F-statistic for this difference is **7.416**; and the corresponding p-value is **0.008**. Because the p-value (0.008) is substantially less than the 0.05 significance level, we confidently reject the null hypothesis. This rejection signifies that the addition of the **weight** variable results in a statistically significant improvement in the model's overall predictive fit. We conclude that weight is a crucial factor in determining price, even after rigorously controlling for the influence of fuel efficiency. Consequently, Model 2 is deemed statistically superior to Model 1.

The final stage of the analysis introduces the third and final predictor block, **gear_ratio**. The purpose here is to definitively determine if this mechanical specification variable provides any further incremental explanatory power beyond that already captured by the combination of **mpg** and **weight** (Model 2).

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

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

Model 3 reports a total R-squared of **0.3150**. While this value is numerically higher than the 0.2934 reported for Model 2, the increase appears marginal. We must consult the change statistics reported at the output's base to ascertain if this modest improvement is statistically meaningful: The R-squared difference (ΔR^2) is **0.022**; the F-statistic is **2.206**; and the corresponding p-value is **0.142**. In this decisive instance, the p-value (0.142) is greater than our established significance level of 0.05. Therefore, we fail to reject the null hypothesis. This crucial finding indicates that the inclusion of the **gear_ratio** does not provide a statistically significant improvement in the model's ability to predict car price over and above the predictors already established in the superior Model 2 (mpg and weight).

Conclusion: Selecting the Optimal Predictive Model

The final summary section, conveniently provided by the `hireg` command, consolidates the findings from all stages, offering a concise and powerful overview of the progression and significance of each block addition:

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

Based on the objective statistical results derived from the hierarchical analysis, the most parsimonious, statistically effective, and theoretically defensible model is **Model 2**. This model demonstrated a significant and meaningful improvement in predictive power over the baseline Model 1. Crucially, the subsequent inclusion of the final predictor block in Model 3 did not yield any further statistically significant gain in predictive accuracy. Therefore, the optimal predictive model for car price in this specific dataset includes **mpg** (fuel efficiency) and **weight** as the essential primary drivers. This conclusion successfully fulfills the goal of hierarchical regression: identifying the minimum set of variables necessary to achieve maximum explanatory power based on a specified theoretical order.