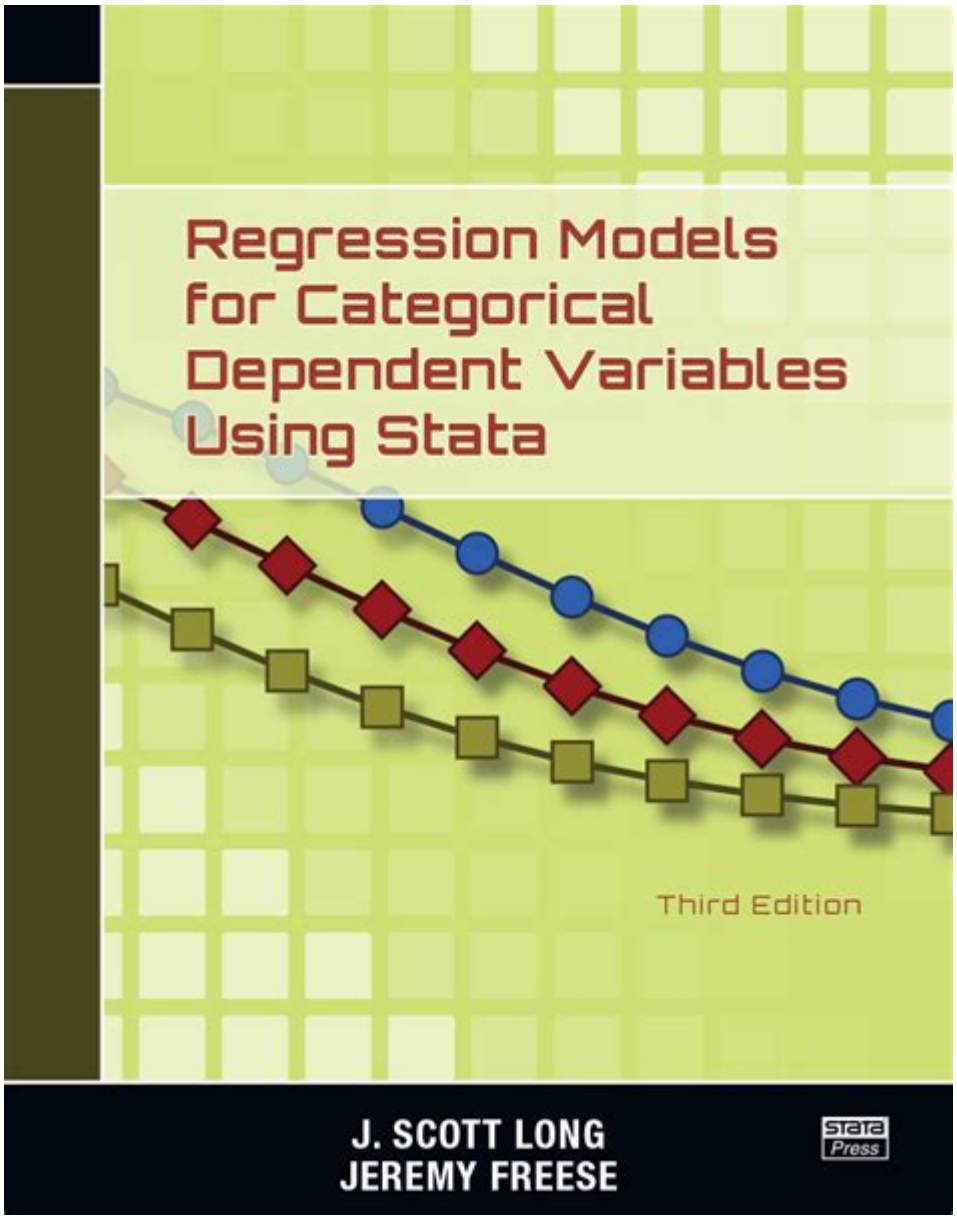
Regression Models For Categorical Dependent Variables Using Stata



Regression models for categorical dependent variables using Stata are essential tools in statistical analysis, particularly when the outcome variable is categorical rather than continuous. This article explores various regression techniques suitable for handling categorical dependent variables in Stata, a powerful statistical software widely used in research. We will discuss the types of categorical dependent variables, appropriate regression models, and the implementation of these models in Stata, along with practical examples and interpretation of results.

Understanding Categorical Dependent Variables

Categorical dependent variables are outcomes that fall into distinct categories or groups, rather than being measured on a continuous scale. These variables can be:

- **Nominal:** Categories without any intrinsic order (e.g., gender, race).
- **Ordinal:** Categories with a clear order but not evenly spaced (e.g., satisfaction ratings: low, medium, high).

The choice of regression model depends on the nature of the categorical variable and the research question.

Types of Regression Models for Categorical Dependent Variables

When working with categorical dependent variables, the following regression models are commonly used:

1. Logistic Regression

Logistic regression is used when the dependent variable is binary (two categories). For example, predicting whether a patient has a disease (yes/no) based on various independent variables.

2. Multinomial Logistic Regression

When the dependent variable has more than two categories without a natural order, multinomial logistic regression is appropriate. For instance, predicting the type of transport used (car, bus, bicycle) based on demographic factors.

3. Ordinal Logistic Regression

If the categorical dependent variable is ordinal, ordinal logistic regression (also known as proportional odds model) is the suitable choice. An example could be predicting customer satisfaction levels (satisfied, neutral, dissatisfied) based on service quality.

4. Poisson and Negative Binomial Regression

These models are used when the dependent variable is a count of occurrences (though

technically not categorical, they are often included in discussions of categorical outcomes). They can be applied in cases like predicting the number of customer complaints received.

Implementing Regression Models in Stata

Stata provides a user-friendly interface and powerful commands for implementing regression models for categorical dependent variables. Below, we outline the steps to perform these analyses.

Logistic Regression in Stata

To perform logistic regression in Stata, the command is structured as follows:

```stata
logit dependent_variable independent_variables
```

For example, to analyze the effect of age and income on the likelihood of having a health condition:

```stata
logit health_condition age income
```

To interpret the results, focus on the coefficients (log-odds) and their significance levels. The odds ratio can be obtained using the `or` option:

```stata
logit health_condition age income, or
```

Multinomial Logistic Regression in Stata

For multinomial logistic regression, use the `mlogit` command:

```stata
mlogit dependent_variable independent_variables
```

For instance, to model transport choice based on age and income:

```stata
mlogit transport_choice age income
```

Again, you can interpret the output by examining the relative risk ratios using the `rr` option:

```stata
mlogit transport_choice age income, rr
```

Ordinal Logistic Regression in Stata

To conduct ordinal logistic regression, Stata provides the `ologit` command:

```stata
ologit dependent_variable independent_variables
```

For example, predicting customer satisfaction based on service quality:

```stata
ologit satisfaction service_quality
```

The output will provide thresholds and coefficients, which can be interpreted to understand the odds of being in a higher satisfaction category.

Poisson and Negative Binomial Regression in Stata

For count data, you can use the `poisson` command:

```stata
poisson count_variable independent_variables
```

If overdispersion is present, the `nbreg` command for negative binomial regression can be used instead:

```stata
nbreg count_variable independent_variables
```

Interpreting the Results

Interpreting the results from regression analyses involving categorical dependent variables requires understanding the output provided by Stata.

1. Coefficients and Odds Ratios

In logistic and multinomial logistic regression, coefficients represent the log-odds of the dependent variable for a one-unit increase in the independent variable. Exponentiating these coefficients gives the odds ratios, which are easier to interpret.

2. Pseudo R-squared

Unlike linear regression, categorical regression models do not have a traditional R-squared value. Instead, Stata provides several pseudo R-squared values (e.g., McFadden's R-squared) that indicate the model's goodness-of-fit.

3. Predicted Probabilities

You can calculate predicted probabilities for categorical outcomes using the `predict` command after fitting a model. For example:

```stata
predict prob_health_condition if e(sample)
```

This command generates predicted probabilities for the health condition based on the independent variables.

Practical Example: Analyzing Survey Data

Let’s consider a practical example of analyzing survey data where we want to predict whether respondents are likely to recommend a service (yes/no) based on their age, income, and satisfaction level.

1. Load the Data:

```stata
use survey_data.dta, clear
```

2. Logistic Regression Model:

```stata
logit recommend age income satisfaction
```

3. Output Interpretation:

- Look at the coefficients to understand the relationship between predictors and the likelihood of recommending the service.
- Use `logit recommend age income satisfaction, or` to obtain odds ratios for easier interpretation.

4. Predicted Probabilities:

```stata
predict prob_recommend if e(sample)
```

You can then analyze how different age and income levels affect the probability of recommending the service.

Conclusion

In summary, regression models for categorical dependent variables using Stata provide powerful tools for analyzing relationships in data where outcomes are not continuous. By leveraging logistic, multinomial, ordinal, and count regression techniques, researchers can uncover insights that help inform decision-making and policy. Mastery of these models and their interpretation is essential for anyone working with categorical data, making Stata an invaluable resource in the toolkit of modern data analysis.

Frequently Asked Questions

What are regression models for categorical dependent variables in Stata?

Regression models for categorical dependent variables in Stata include logistic regression, multinomial logistic regression, and ordinal logistic regression. These models are designed to handle outcomes that are categorical, allowing researchers to analyze the relationship between independent variables and categories of a dependent variable.

How do you perform a logistic regression in Stata?

To perform a logistic regression in Stata, you can use the 'logit' or 'logistic' command. For example, you would use 'logit dependent_variable independent_variable1 independent_variable2' to estimate the model, where 'dependent_variable' is your binary outcome.

What is the difference between logistic and multinomial logistic regression in Stata?

Logistic regression is used for binary outcomes, while multinomial logistic regression is used for outcomes with more than two categories. In Stata, you can perform multinomial logistic regression using the 'mlogit' command, which allows you to model the probability

of each category relative to a baseline category.

How can you interpret coefficients from a logistic regression model in Stata?

In a logistic regression model, the coefficients represent the change in the log odds of the dependent variable occurring for a one-unit increase in the independent variable. You can also exponentiate the coefficients using the 'exp()' function to interpret them as odds ratios.

What command would you use for ordinal regression in Stata?

For ordinal regression in Stata, you can use the 'oprobit' or 'ologit' command. For example, you would use 'ologit ordinal_dependent_variable independent_variable1 independent_variable2' for an ordered outcome.

How do you assess the fit of a categorical regression model in Stata?

To assess the fit of a categorical regression model in Stata, you can look at statistics such as the pseudo R-squared, likelihood ratio tests, and goodness-of-fit tests. For logistic models, you can also use the 'estat gof' command to perform a goodness-of-fit test after running the regression.

Regression Models For Categorical Dependent Variables Using Stata

Revealing the driving factors of urban wetland park cooling effects ...
Feb 15, 2025 · In contrast, our study, which employed random forest regression and the SHAP algorithm, offers a deeper ...

Regression Analysis - an overview | ScienceDirect Topics
Regression analysis is a statistical method for analyzing a relationship between two or more variables in such a manner that one variable ...

Focal and efficient IOU loss for accurate bounding box regression
Sep 28, 2022 · In object detection, bounding box regression (BBR) is a crucial step that determines the object localization ...

Flood shocks, heterogeneous risk exposure, and housing market dyna...

This paper analyzes the economic consequences of flood shocks on housing markets in China. By combining detailed …

线性回归中的R，R平方和调整后的R平方有什么区别？ - 知乎
线性回归中的r你指的是相关系数吧，就是用来描述两个变量的线性相关程度。r绝对值越大表示2个变量间的线性相关程度越高。线性回归中的 R^2 是决定系数，表 …

Revealing the driving factors of urban wetland park cooling effects ...
Feb 15, 2025 · In contrast, our study, which employed random forest regression and the SHAP algorithm, offers a deeper understanding of the complex interactions between landscape …

Regression Analysis - an overview | ScienceDirect Topics
Regression analysis is a statistical method for analyzing a relationship between two or more variables in such a manner that one variable can be predicted or explained by using information …

Focal and efficient IOU loss for accurate bounding box regression
Sep 28, 2022 · In object detection, bounding box regression (BBR) is a crucial step that determines the object localization performance. However, we find that most p…

Flood shocks, heterogeneous risk exposure, and housing market ...
This paper analyzes the economic consequences of flood shocks on housing markets in China. By combining detailed housing transaction records with gran…

线性回归中的**R**，**R**平方和调整后的**R**平方有什么区别？ - 知乎
线性回归中的r你指的是相关系数吧，就是用来描述两个变量的线性相关程度。r绝对值越大表示2个变量间的线性相关程度越高。线性回归中的 R^2 是决定系数，表示自变量（可能有多个）对因变量的解 …

如何简单明了地理解分位数回归？ - 知乎
如何了地理解分位数回归？分位数回归 不同于传统的均值回归模型，分位数回归模型可以研究多个分位点上的回归关系，并且可以查看X对Y的影响趋势及变化情况，而且分位数回归对于异常值、因变量 …

Statistical inference for smoothed quantile regression with ...
May 1, 2025 · In this paper, we tackle the problem of conducting valid statistical inference for quantile regression with streaming data. The main difficulties are …

Multiple Linear Regression Model for Improved Project Cost ...
Jan 1, 2022 · Multiple linear regression analysis is performed to evaluate the number of regressors, the priority of the candidate EVM variables into the regression model, and to assess the …

Regression Modeling Strategies - ScienceDirect
Jun 1, 2011 · Multivariable regression models are widely used in health science research, mainly for two purposes: prediction and effect estimation. Various strateg…

鲁棒回归 Robust Regression 到底是什么？ - 知乎
）和一般回归分析方法相比，鲁棒回归（Robust Regression）不容易受 离群值（outlier）的影响。一些常见的鲁棒回归算法包括：Theil-Sen回归、Huber回归、RANSAC。1、应用情形 数据挖掘， …

Explore regression models for categorical dependent variables using Stata. Unlock insights and enhance your analysis skills. Learn more in our comprehensive guide!

Back to Home