Comparing Statistical Models

Andy Grogan-Kaylor

19 Oct 2021 11:43:01

# Introduction

In this example, we explore the predictors of the *count of Adverse Childhood Experiences* (ACES) that children experience. Using the *general linear model* framework, we could conceivably compare different statistical models on several grounds.

1. Theoretical plausibility
2. Functional form of the dependent variable
3. Functional form of the entire model
4. Statistical criteria of fit.

Frequently, there is no one correct way to analyze data, and different statistical approaches need to be weighed on multiple criteria to ascertain which approach(es) is / are appropriate.

# Theoretical and Functional Concerns

| Statistical Model | Stata Command | Theoretical Rationale | Functional Form of Dependent Variable | Functional Form of Entire Model |
| --- | --- | --- | --- | --- |
| OLS | regress | Continuous dependent variable |  | y is a linear function of the x’s |
| Logistic Regression | logit | Binary dependent variable |  | is a linear function of x’s |
| Ordinal logistic regression | ologit | Ordered dependent variable where distance between categories does not matter |  | is a linear function of x’s |
| Multinomial Logistic Regression | mlogit | Dependent variable with multiple unordered categories |  | is a linear function of x’s |
| Poisson Regression | poisson | Dependent variable representing a count |  | is a linear function of x’s |
| Negative Binomial Regression | nbreg | Dependent variable representing a count |  | is a linear function of x’s |

# Assessing Model Fit

## Get Data And Create Count of ACEs

. clear all

. use "NSCH\_ACES.dta", clear

. egen acecount = anycount(ace\*R), values(1) // generate count of ACES

## Describe The Data

. describe acecount sc\_sex sc\_race\_r higrade  
  
Variable Storage Display Value  
 name type format label Variable label  
───────────────────────────────────────────────────────────────────────────────────────────────────  
acecount byte %8.0g ace1R ace3R ace4R ace5R ace6R ace7R ace8R ace9R  
 ace10R == 1  
sc\_sex byte %30.0g sc\_sex\_lab  
 Sex of Selected Child  
sc\_race\_r byte %48.0g sc\_race\_r\_lab  
 Race of Selected Child, Detailed  
higrade byte %61.0g higrade\_lab  
 Highest Level of Education among Reported Adults

## Explore Some Models

Only some of the above listed models are relevant. We use quietly to suppress model output at this stage.

. quietly: regress acecount sc\_sex i.sc\_race\_r i.higrade // OLS

. estimates store OLS

. quietly: ologit acecount sc\_sex i.sc\_race\_r i.higrade // ordinal logit

. estimates store ORDINAL

. quietly: poisson acecount sc\_sex i.sc\_race\_r i.higrade // Poisson

. estimates store POISSON

. quietly: nbreg acecount sc\_sex i.sc\_race\_r i.higrade // Negative Binomial

. estimates store NBREG

## Compare The Models Including Fit Measures

. estimates table OLS ORDINAL POISSON NBREG, var(20) star stats(N ll aic bic) equations(1)  
  
─────────────────────┬────────────────────────────────────────────────────────────────  
 Variable │ OLS ORDINAL POISSON NBREG   
─────────────────────┼────────────────────────────────────────────────────────────────  
#1 │  
 sc\_sex │ -.01358634 -.02856135 -.01282301 -.0127557   
 │  
 sc\_race\_r │  
Black or African .. │ .32583464\*\*\* .47967243\*\*\* .26627607\*\*\* .28235733\*\*\*   
American Indian o.. │ .88542522\*\*\* .88482406\*\*\* .59710627\*\*\* .62278046\*\*\*   
 Asian alone │ -.46503425\*\*\* -.76002818\*\*\* -.62438214\*\*\* -.62012779\*\*\*   
Native Hawaiian a.. │ .2516065 .35416681 .20674094\* .21879323   
Some Other Race a.. │ .07433855 .14197623\* .06755212\* .08062919   
 Two or More Races │ .33035205\*\*\* .39265187\*\*\* .28181254\*\*\* .28198179\*\*\*   
 │  
 higrade │  
High school (inc..) │ .10021068 .17111252\* .06324858\* .06584405   
More than high sc.. │ -.45113751\*\*\* -.62649139\*\*\* -.37861085\*\*\* -.38098265\*\*\*   
 │  
 \_cons │ 1.411494\*\*\* .33994246\*\*\* .33915207\*\*\*   
─────────────────────┼────────────────────────────────────────────────────────────────  
cut1 │  
 \_cons │ -.78624597\*\*\*   
─────────────────────┼────────────────────────────────────────────────────────────────  
cut2 │  
 \_cons │ .65037457\*\*\*   
─────────────────────┼────────────────────────────────────────────────────────────────  
cut3 │  
 \_cons │ 1.5299647\*\*\*   
─────────────────────┼────────────────────────────────────────────────────────────────  
cut4 │  
 \_cons │ 2.2019291\*\*\*   
─────────────────────┼────────────────────────────────────────────────────────────────  
cut5 │  
 \_cons │ 2.8850071\*\*\*   
─────────────────────┼────────────────────────────────────────────────────────────────  
cut6 │  
 \_cons │ 3.6106908\*\*\*   
─────────────────────┼────────────────────────────────────────────────────────────────  
cut7 │  
 \_cons │ 4.4853373\*\*\*   
─────────────────────┼────────────────────────────────────────────────────────────────  
cut8 │  
 \_cons │ 5.9106719\*\*\*   
─────────────────────┼────────────────────────────────────────────────────────────────  
cut9 │  
 \_cons │ 7.5036903\*\*\*   
─────────────────────┼────────────────────────────────────────────────────────────────  
lnalpha │  
 \_cons │ -.54430672\*\*\*   
─────────────────────┼────────────────────────────────────────────────────────────────  
Statistics │   
 N │ 30530 30530 30530 30530   
 ll │ -52340.464 -42451.588 -44758.999 -42775.864   
 aic │ 104700.93 84939.175 89537.999 85573.728   
 bic │ 104784.19 85089.052 89621.263 85665.319   
─────────────────────┴────────────────────────────────────────────────────────────────  
 Legend: \* p<0.05; \*\* p<0.01; \*\*\* p<0.001

In terms of *log-likelihood* a higher value indicates a better fit. We can also use the *Akaike Information Criterion* (AIC) and the *Bayesian Information Criterion* (BIC) to compare models. For AIC and BIC, lower values indicate a better fit.

Thus, on strictly statistical grounds, the *ordinal* model would appear to provide the best fit, followed by the *negative binomial* model, the *Poisson* model, and the *OLS* model. However, we should note that the differences in fit between the *ordinal*, *negative binomial* and *Poisson* models are not exceptionally large. We would also worry that any differences in fit that we do see might be due to overfitting in this particular sample, or to capitalizing upon chance.

Lastly, we’d worry that the ordinal model might not satisfy the *proportional hazards* assumption, and should examine this with a brant test.

We need to balance these differences in fit against the fact that theoretically, a count data model seems more appropriate.

In this case, we would most likely choose to proceed with a count regression model.