Limited Dependent Variable Models

Jerome Dumortier

23 March 2023

Overview

Packages and Files

Required packages:

- AER
- censReg
- foreign
- MASS
- pscl
- truncreg

Required files:

```
data("NMES1988",package="AER")
```

Topics Covered

Regression models in which the dependent variable is somehow limited:

- Truncated data: Values above and/or below particular points are not reported
- Censored data: Values above and/or below particular points are reported at those points
- Count data: Discrete, integer count value
- Survival/duration data: Time to a certain event

Truncation

Overview

Concept

- Value above and/or below a certain point are not part of the data

Examples

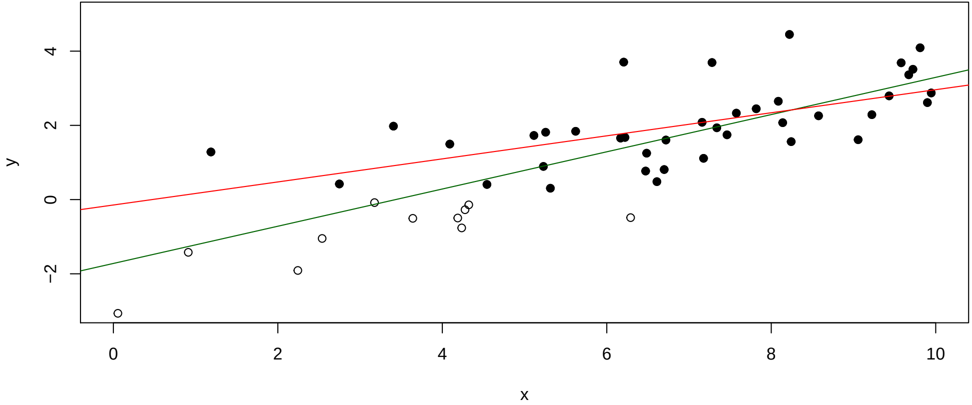
- Low income household studies
- On-site visitation data (unobserved non-visitors)
- Employment data on hours worked (excludes unemployed)

Simulated data

- "True" Coefficients: $\beta_0 = -2$ and $\beta_1 = 0.5$
- Values $y < 0$ are not reported in the data

Next slide: The green regression line is "correct" whereas the "red" is the line obtained from a regression model which ignores the truncation.

Graphical Illustration



Setup for truncation Data

```
truncation1     = truncation[c("y_real","x")]
truncation2     = subset(truncation,y_obs>0,select=c("y_obs","x"))
bhat_real       = lm(y_real~x,data=truncation1)
bhat_truncated = lm(y_obs~x,data=truncation2)
```

Required package to estimate a truncated model

- truncreg

Additional variable output *sigma*:

- Related to the truncated normal distribution

Results: Complete Data

```
##
## Call:
## lm(formula = y_real ~ x, data = truncation1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9198 -0.6360 -0.1532  0.3463  2.4094
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.72133    0.36765  -4.682 2.36e-05 ***
## x            0.50153    0.05501   9.116 4.78e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9821 on 48 degrees of freedom
## Multiple R-squared:  0.6339, Adjusted R-squared:  0.6263
## F-statistic: 83.11 on 1 and 48 DF,  p-value: 4.783e-12
```

Results: Truncated Data with Regular OLS

```
##
## Call:
## lm(formula = y_obs ~ x, data = truncation2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4247 -0.5086 -0.1122  0.3488  2.0413
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.14606    0.48297  -0.302    0.764
## x            0.31074    0.06605   4.705  3.5e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8543 on 37 degrees of freedom
## Multiple R-squared:  0.3743, Adjusted R-squared:  0.3574
## F-statistic: 22.13 on 1 and 37 DF,  p-value: 3.499e-05
```

Results: Correcting for Truncation

```
##
## Call:
## truncreg(formula = y_obs ~ x, data = truncation2, point = 0,
##     direction = "left")
##
## BFGS maximization method
## 22 iterations, 0h:0m:0s
## g'(-H)^-1g = 7.28E-09
##
##
##
## Coefficients :
##             Estimate Std. Error t-value  Pr(>|t|)
## (Intercept) -0.716370   0.668897 -1.0710    0.2842
## x            0.378806   0.086811  4.3636 1.279e-05 ***
## sigma        0.889160   0.119117  7.4646 8.349e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -46.315 on 3 Df
```

Achievement Scores: Data Load and Description

Loading the data using the package foreign

```
url         = "https://stats.idre.ucla.edu/stat/data/truncreg.dta"
achievement = read.dta(url)
```

Description of the data from UCLA Source:

"A study of students in a special GATE (gifted and talented education) program wishes to model achievement as a function of language skills and the type of program in which the student is currently enrolled. A major concern is that students are required to have a minimum achievement score of 40 to enter the special program. Thus, the sample is truncated at an achievement score of 40."

Achievement Scores: Regular OLS Estimation

```
##
## Call:
## lm(formula = achiv ~ langscore + prog, data = achievement)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.9413  -5.7033  -0.8462   5.2205  21.3010
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.63965    3.70639   7.457 4.01e-12 ***
## langscore    0.46319    0.06792   6.820 1.45e-10 ***
## progacademic  2.97343    1.44889   2.052   0.0416 *
## progvocation -0.52118    1.72739  -0.302   0.7632
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.534 on 174 degrees of freedom
## Multiple R-squared:  0.3054, Adjusted R-squared:  0.2934
## F-statistic:  25.5 on 3 and 174 DF,  p-value: 1.01e-13
```

Achievement Scores: Truncated Model

```
##
## Call:
## truncreg(formula = achiv ~ langscore + prog, data = achievement,
##     point = 40, direction = "left")
##
## BFGS maximization method
## 57 iterations, 0h:0m:0s
## g'(-H)^-1g = 2.5E-05
##
##
##
## Coefficients :
##              Estimate Std. Error t-value  Pr(>|t|)
## (Intercept)  11.29942    6.77173  1.6686   0.09519 .
## langscore     0.71267    0.11446  6.2264 4.773e-10 ***
## progacademic  4.06267    2.05432  1.9776   0.04797 *
## progvocation -1.14422    2.66958 -0.4286   0.66821
## sigma         8.75368    0.66647 13.1343 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -591.31 on 5 Df
```

Censoring

Overview

Concept

- Value above and/or below a certain point are not part of the data

Examples

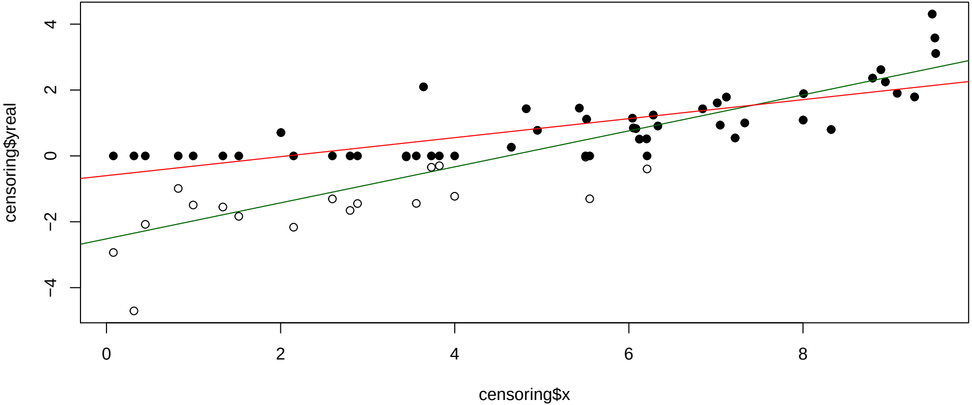
- Capacity constrained data, e.g., class enrollments or ticket sales
- Hours worked (or leisure demand), which is essentially capacity constrained
- Commodity purchases (non-negative)

Simulated data

- "True" Coefficients: $\beta_0 = -2$ and $\beta_1 = 0.5$
- Values $y < 0$ are reported at 0

R package censReg to reduce bias

Graphical Illustration



Results: Full Data

```
##
## Call:
## lm(formula = yreal ~ x, data = censoring)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.36075 -0.52032  0.04652  0.40126  2.62549
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.51759    0.28153  -8.943 8.62e-12 ***
## x            0.54628    0.04704  11.612 1.52e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9074 on 48 degrees of freedom
## Multiple R-squared:  0.7375, Adjusted R-squared:  0.732
## F-statistic: 134.8 on 1 and 48 DF,  p-value: 1.522e-15
```

Results: Censored Data with Regular OLS

```
##
## Call:
## lm(formula = y ~ x, data = censoring)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.19126 -0.47822 -0.03578  0.35424  2.17113
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5992     0.2124  -2.821  0.00695 **
## x             0.2884     0.0355   8.123 1.43e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6847 on 48 degrees of freedom
## Multiple R-squared:  0.5789, Adjusted R-squared:  0.5701
## F-statistic: 65.99 on 1 and 48 DF,  p-value: 1.434e-10
```

Estimation of a Censored Model

```
##
## Call:
## censReg(formula = y ~ x, data = censoring)
##
## Observations:
##          Total  Left-censored     Uncensored Right-censored
##             50             19             31              0
##
## Coefficients:
##             Estimate Std. error t value  Pr(> t)
## (Intercept) -2.10846    0.44764  -4.710 2.48e-06 ***
## x            0.48898    0.06582   7.429 1.09e-13 ***
## logSigma    -0.18013    0.12993  -1.386    0.166
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Newton-Raphson maximisation, 6 iterations
## Return code 1: gradient close to zero (gradtol)
## Log-likelihood: -46.05041 on 3 Df
```

Count Models

Overview

Dependent variable

- Discrete, integer count data

Examples

- What are the number of arrests for a person?
- What determines the number of credit cards a person owns?

Three count data models

1. Poisson regression
2. Quasi-Poisson Regression Model
3. Negative Binomial Regression Model

Choice criteria: Presence or absence of overdispersion

- Overdispersion Variance of the dependent variable is larger than its mean.
- Poisson model is not suitable for overdispersion

Packages

The main package used is pscl. There is also an additional resource with more theoretical details on the topic: Regression Models for Count Data in R. A more up-to-date version of the document may be found with the pscl package documentation.

Poisson Regression Model

Recall Poisson distribution:

$$Pr(Y = k) = \frac{e^{-\lambda} \cdot \lambda^k}{k!}$$

Equidispersion as key characteristics:

- Mean and variance equal to λ, i.e., $E(Y) = \lambda$ and $Var(Y) = \lambda$
- Poisson regression: $\lambda = exp(\beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k)$.

NHTS Example: Number of Vehicles (`hhpub`)

Data source

- 2017 National Household Travel Survey
- Survey quantifying trip and travel habits across the United States
- Example use: Quantifying intra-day electricity demand from electric vehicles

Outcome of interest

- Number of vehicles based on household income, home ownership, and urban/rural household location

Data preparation

- Elimination of missing and unknown data value
- Conversion of income to 1,000 dollars

Data Preparation

```
hhpubdata           = subset(hhpub,HHFAMINC %in% c(1:11) &
                                   HOMEOWN %in% c(1,2) &
                                   URBRUR %in% c(1,2) &
                                   HHVEHCNT %in% c(0:12))
HHFAMINC            = c(1:11)
INCOME              = c(10,12.5,20,30,42.5,57.5,82.5,112.5,137.5,
                        175,200)
INCOME              = data.frame(HHFAMINC,INCOME)
hhpubdata           = merge(hhpubdata,INCOME)
hhpubdata$RURAL     = hhpubdata$URBRUR-1
hhpubdata$RENT      = hhpubdata$HOMEOWN-1
```

Poisson Model Execution

Preliminary step: Calculation of mean and variance of dependent variable

```
mean(hhpubdata$HHVEHCNT)
```

```
## [1] 1.981142
```

```
var(hhpubdata$HHVEHCNT)
```

```
## [1] 1.386027
```

Similar values and thus, Poisson regression model as an appropriate first step.

```
bhat_pois = glm(HHVEHCNT~INCOME+RENT+RURAL,
                data=hhpubdata,family=poisson)
```

Results

```
##
## Call:
## glm(formula = HHVEHCNT ~ INCOME + RENT + RURAL, family = poisson,
##     data = hhpubdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6889  -0.5568  -0.1558   0.3590   5.5063
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.654e-01  4.292e-03  108.43   <2e-16 ***
## INCOME       2.986e-03  3.601e-05   82.93   <2e-16 ***
## RENT        -3.733e-01  5.797e-03  -64.39   <2e-16 ***
## RURAL        2.224e-01  4.616e-03   48.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 86505  on 124400  degrees of freedom
## Residual deviance: 68533  on 124397  degrees of freedom
## AIC: 370161
##
## Number of Fisher Scoring iterations: 5
```

Interpretation

Sign of coefficients as an indication of the direction of influence on the outcome variable, i.e., the number of cars.

- Association of higher income and rural living with a higher number of car
- Association of renting with lower number of vehicles.
- Possible correlation between income and renting

General coefficient interpretation using $\exp(\beta)$, i.e., every unit increase in X has a multiplicative effect of $\exp(\beta)$ on the mean of Y, i.e., λ:

- $\beta = 0 \Rightarrow \exp(\beta) = 1$: Y and X are not related.
- $\beta > 0 \Rightarrow \exp(\beta) > 1$: Expected count $E(y)$ is $\exp(\beta)$ times larger than when $X = 0$
- $\beta < 0 \Rightarrow \exp(\beta) < 1$: Expected count $E(y)$ is $\exp(\beta)$ times smaller than when $X = 0$

Testing for Overdispersion I

Function `overdispersion()` from the package AER:

- Tests the null hypothesis of equidispersion (i.e., assuming no overdispersion)

Executed after the main regression using `glm(...,family=poisson)`

Testing for Overdispersion II

```
dispersiontest(bhat_pois)
```

```
##
##  Overdispersion test
##
## data:  bhat_pois
## z = -115.75, p-value = 1
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##  0.5670593
```

Given the p-value, the null hypothesis cannot be rejected. If the data suggests overdispersion, two alternative regression models can be used: (1) Quasi-Poisson and (2) Negative Binomial.

Quasi-Poisson Regression Model

Dataset `blm` from article Black Lives Matter: Evidence that Police-Caused Deaths Predict Protest Activity.

- Dependent variable: Total number of protests in a city
- Note that the paper includes a significant number of supplementary materials which allows for the replication of the results and much more.

First step: Calculation of mean and variance of the variable *totalprotests*:

```
mean(blm$totprotests)
```

```
## [1] 0.4959529
```

```
var(blm$totprotests)
```

```
## [1] 6.35326
```

Presence of Overdispersion

The variance is significantly higher than the mean which suggests overdispersion. In a first step, a regular Poisson model is estimated.

```
eq1        = "totprotests~log(pop)+log(popdensity)+percentblack+
             blackpovertyrate+I(blackpovertyrate^2)+
             percentbachelor+collegeenrollpc+demshare"
bhat1      = glm(eq1,data=blm,family=poisson)
```

Estimation Results

```
##
## Call:
## glm(formula = eq1, family = poisson, data = blm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.6571  -0.5238  -0.3008  -0.1632   6.5795
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -2.001e+01  6.327e-01 -31.625  < 2e-16 ***
## log(pop)               1.129e+00  4.007e-02  28.170  < 2e-16 ***
## log(popdensity)       -1.831e-01  8.654e-02  -2.116   0.0343 *
## percentblack           1.697e-02  3.104e-03   5.467 4.59e-08 ***
## blackpovertyrate       1.461e-01  2.636e-02   5.541 3.02e-08 ***
## I(blackpovertyrate^2) -1.552e-03  3.985e-04  -3.895 9.82e-05 ***
## percentbachelor        3.893e-02  3.918e-03   9.935  < 2e-16 ***
## collegeenrollpc        9.305e-03  2.377e-03   3.914 9.06e-05 ***
## demshare               4.301e-02  5.293e-03   8.126 4.43e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3204.6  on 1225  degrees of freedom
## Residual deviance:  787.4  on 1217  degrees of freedom
##   (133 observations deleted due to missingness)
## AIC: 1242.9
##
## Number of Fisher Scoring iterations: 6
```

Testing for Overdispersion

```
##
##  Overdispersion test
##
## data:  bhat1
## z = 1.4052, p-value = 0.07998
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   2.212733
```

Null hypothesis rejected at 10% but not 5% significance level. The Quasi-Poisson Regression Model handles overdispersion by adjusting standard errors but leaving the coefficient estimates the same.

Estimation Results: Quasipoisson

```
##
## Call:
## glm(formula = eq1, family = quasipoisson, data = blm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.6571  -0.5238  -0.3008  -0.1632   6.5795
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -2.001e+01  9.841e-01 -20.332  < 2e-16 ***
## log(pop)               1.129e+00  6.232e-02  18.111  < 2e-16 ***
## log(popdensity)       -1.831e-01  1.346e-01  -1.360 0.173942
## percentblack           1.697e-02  4.828e-03   3.515 0.000457 ***
## blackpovertyrate       1.461e-01  4.100e-02   3.562 0.000382 ***
## I(blackpovertyrate^2) -1.552e-03  6.198e-04  -2.504 0.012403 *
## percentbachelor        3.893e-02  6.094e-03   6.387 2.40e-10 ***
## collegeenrollpc        9.305e-03  3.697e-03   2.517 0.011975 *
## demshare               4.301e-02  8.233e-03   5.225 2.05e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 2.419275)
##
##     Null deviance: 3204.6  on 1225  degrees of freedom
## Residual deviance:  787.4  on 1217  degrees of freedom
##   (133 observations deleted due to missingness)
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

Negative Binomial Regression Model

The Negative Binomial Regression Model can be used in the presence of count data and overdispersion. Below, the results from the article Black Lives Matter: Evidence that Police-Caused Deaths Predict Protest Activity are recreated using the negative binomial models presented in the paper.

Three models:

1. Resource mobilization and opportunity structure
2. Adding black death
3. Adding all police-caused deaths instead (victims of any race)

BLM Model I

```
bhat3 = glm.nb(eq1,data=blm,link=log)
```

BLM Model II

```
bhat4       = glm.nb(totprotests~log(pop)+log(popdensity)+
                      percentblack+blackpovertyrate+
                      I(blackpovertyrate^2)+percentbachelor+
                      collegeenrollpc+demshare+
                      deathsblackpc,data=blm,link=log)
```

BLM Model III

```
bhat5      = glm.nb(totprotests~log(pop)+log(popdensity)+
                    percentblack+blackpovertyrate+
                    I(blackpovertyrate^2)+percentbachelor+
                    collegeenrollpc+demshare+
                    deathspc,data=blm,link=log)
```