Intermediate Econometrics

12/12/2025 - Vincenzo Gioia

Count data

Introduction

- **Cases where the value of the response is a count, that is a non-negative integer**
- Examples: annual doctor visits, number of cigarettes smoked daily, number of trips taken by members of households
- Drawbacks of fitting a linear model:
  - the integer nature of the response is not taken into account
  - the fitted model can predict negative values for some values of the covariates
  - if the distribution of the response is asymmetric, the logarithm transformation can't be used in the common case where the response is 0 for a subset of observations

Count data

The Poisson distribution

- Count data can be considered as coming from a random variable that follows a Poisson distribution, which has a unique parameter, this parameter being the mean and the variance of the series

$$Y_1, \ldots, Y_n \overset{i.i.d}{\sim} Poisson(\lambda), \quad \lambda > 0$$

$$P(Y_i = y) = \frac{e^{-\lambda}\lambda^y}{y!}, \quad y \in \{0, 1, 2, \ldots\}$$

$$\mathbf{E}(Y) = \lambda \quad \mathbf{V}(Y) = \lambda$$

- The ML estimator of this parameter is the sample mean $\hat{\lambda}_M V = \bar{Y}$
- Count data often take small values
- While comparing real count data to Poisson distribution, one often faces two problems:
  - **overdispersion**, which means that the variance is much greater than the mean
  - **excess of zero**, which means that the mean probability of a zero value computed using the Poisson distribution is often much less than the observed share of zero in the sample

Count data

Overdispersion and excess of zero

- **Overdispersion** is important because we consider that the Poisson parameter is the same for every observation:
  - Adding covariates will give specific values of the Poisson parameter for every observations and will therefore reduce the (conditional) variance
  - Anyway, the overdispersion problem is often still present even after introducing covariates, because of unobserved heterogeneity
- The **excess of zero** problem is difficult to distinguish from overdispersion, as zero is an extreme value of the distribution and therefore, an excess of zero leads to overdispersion
  - The excess of zero can in part be explained by the fact that 0 is a special value that may not be correctly explained by the same process that explains strictly positive values

Econometric example

Number of trips

- Source: Terza (1998)
- Outcome: number of trips taken the day before the interview
- Let's start by exploring the dataset and only considering the unconditional distribution (computing the first two moments and percentage of zero)

```
 1 library(micsr)
 2 library(micsr.data)
 3 load("trips.rda")
 4 trips <- as.data.frame(trips)
 5 str(trips)
```

```
'data.frame':   577 obs. of  11 variables:
 $ trips   : num  2 3 0 0 1 1 2 2 1 0 ...
 $ car     : num  1 0 0 0 1 1 0 1 0 0 ...
 $ workschl: num  0.5 0.667 0 0 1 ...
 $ size    : num  6 2 2 4 5 1 2 4 6 4 ...
 $ dist    : num  0.595 0.595 0.595 0.595 0.595 ...
 $ smsa    : num  1 1 1 1 1 1 1 1 1 1 ...
 $ fulltime: num  1 2 0 0 1 1 2 2 0 0 ...
 $ adults  : num  4 2 2 4 1 1 2 2 2 1 ...
 $ distnod : num  1 1 1 1 1 1 1 4.5 1 1 ...
```

```
$ realinc : num  0.7 1.8 0.44 0.44 0.7 1.8 1.4 1.4 0.44 0.44 ...
$ weekend : num  1 0 0 0 1 0 1 1 0 0 ...
```

Econometric example

Number of trips

- This data exhibits important overdispersion, as the variance is about 5 times larger than the mean
- The predicted probability of 0 for the Poisson model is 0.01056, which is much lower than the actual percentage of 0 which is 18.5%
- There is therefore a large excess of 0, which also can be associated with overdispersion as 0 is a value far from the mean of 4.551

```
1 mean(trips$trips)
```

```
[1] 4.551127
```

```
1 var(trips$trips)
```

```
[1] 24.35545
```

```
1 mean(trips$trips==0)
```

```
[1] 0.1854419
```

```
1 dpois(0, 4.551)
```

```
[1] 0.01055664
```

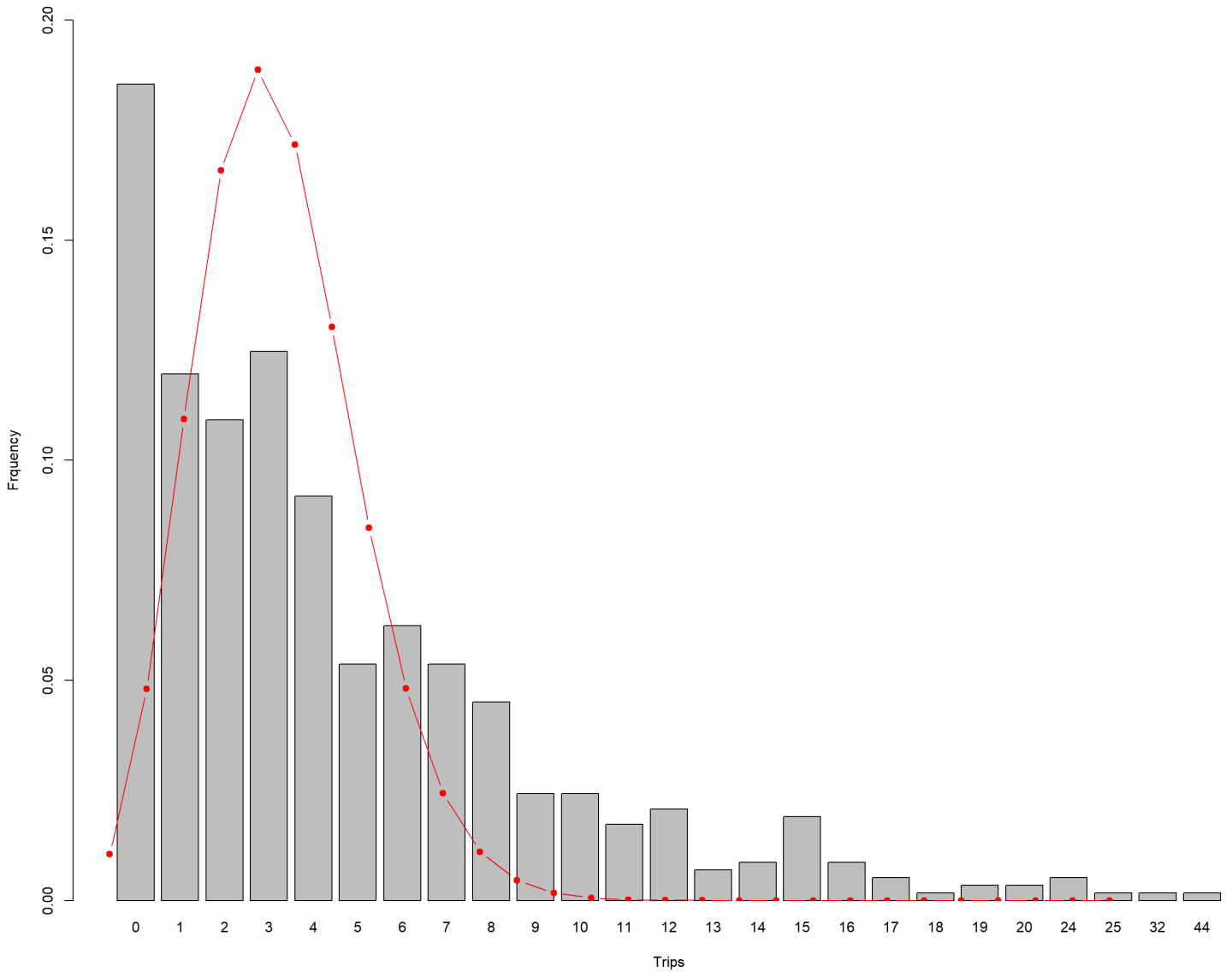
Econometric example

Number of trips

- Representation of the frequency of the data and probability mass of the fitted model
- For example, for $Y = 0$, the relative frequency is 0.18 and, as seen previously, the fitted probability is 0.01056 which implies
- The frequencies are clearly over-estimated for values of the response close to the mean (3, 4, 5, 6) and under-estimated for extreme values
- This is particularly the case for 0, which indicates that a specific problem of excess of zero may be present

```r
freq_rel <- prop.table(table(trips$trips))
barplot(freq_rel,  xlab = "Trips", ylab = "Frquency",  ylim = c(0,0.20))
lambda <- mean(trips$trips)
pmf <- dpois(0:27, lambda)
lines(0:27, pmf, type = "b", pch = 19, col = "red")
```

Econometric example



The Poisson model

Theoretical framework

- The model relies on the hypothesis that the response is a Poisson variable and therefore that the probability of a given value is

$$P(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!}, \quad y \in \{0, 1, 2, \ldots\}$$

having considered $\mu = \lambda$

- The Poisson distribution is defined by a unique parameter , which is the mean and the variance of the distribution. This is a striking feature of the Poisson distribution which differs for example from the normal distribution which has two separate position and dispersion parameters
- The Poisson probability can also be written as:

$$\mathrm{P}(Y = y, \mu) = e^{y\ln\mu - \mu - \ln y!}$$

- So, the Poisson model belongs to the generalized linear model family, with:
  - $\theta = \ln\mu$
  - $b(\theta) = \mu = e^{\theta}$

The Poisson model

Theoretical framework

- The link defines the relation between the linear predictor η and the parameter of the distribution μ: $\eta = g(\mu)$. The canonical link is, for the Poisson model, the log link: $\eta = \ln\mu$
- For a given set of covariates, the Poisson parameter for the i-th observation is $\mu_i = g^{-1}(\eta_i)$, with $\eta_i = x_i^\top \beta$ the linear predictor

$$\mu_i = \mathrm{E}(Y_i \mid x_i) = e^{x_i^\top \beta}$$

and the log-likelihood function is:

$$\ln L(\beta) = \sum_{i=1}^{n} \ln \frac{e^{-e^{x_i^\top \beta}} e^{y_i x_i^\top \beta}}{y_i!} = \sum_{i=1}^{n} -e^{x_i^\top \beta} + y_i x_i^\top \beta - \ln y_i!$$

- We can implement the log-likelihood and optimize it by using optimizer, or implementing the Newton-Raphson algorithm requiring the first and second order derivative (the latter is useful for obtaining the standard errors of the parameter estimator)

Econometric example

Number of trips

- The covariates are:
  - the share of trips for work or school (`workschl`)
  - the number of individuals in the household (`size`)
  - the distance to the central business district (`dist`)
  - a dummy (`smsa`) for large urban area
  - the number of full-time workers in a household (`fulltime`)
  - the distance from home to the nearest transit node (`distnode`)
  - household income divided by the median income of the census tract (`realinc`)
  - a dummy if the survey period is Saturday or Sunday (`weekend`)
  - a dummy for owning at least one car (`car`)

Econometric example

Number of trips

- The Poisson model is fitted using the `glm` function and setting the `family` argument to `poisson`
- For the sake of comparison, we also estimate a linear model using `lm`:

```
footrips <- trips ~ workschl + size + dist + smsa + fulltime + distnod +
  realinc + weekend + car
pois_trips <- glm(footrips, data = trips, family = poisson)
lm_trips <- lm(footrips, data = trips)
```

Econometric example

Coefficients' interpretation

- An increase of 1 of `realinc` means that household's income increases by an amount equal to the local median income
- The linear models indicates that the number of trips then increases by 0.133 trips
- The Poisson model indicates:
  - that the relative increase of trips dy/y is equal to 0.019, i.e., 2%
  - At the sample mean, it corresponds to an increase of $0.019 \times 4.55 = 0.086$ trips

```
1 rbind(lm = coef(lm_trips), poisson = coef(pois_trips))
```

```
        (Intercept)   workschl      size         dist        smsa  fulltime
lm       -0.8799070 -2.1504174 0.8284770 -0.012944001 -0.04841164 1.3330729
poisson  -0.5702721 -0.4556281 0.1667228 -0.002199494 -0.03093073 0.2482265
            distnod    realinc     weekend      car
lm       0.028952450 0.13339483 -0.47650465 2.219090
poisson  0.004878584 0.01881775 -0.07382439 1.413325
```

Econometric example

Coefficients' interpretation

- If the interview concerns a weekend day, the OLS coefficient is -0.48, which indicates that the number of trips taken during the weekend are about one-half less, compared to a weekday
- For such a dummy variable, denoting β_w and $\tilde{x}$ a given vector of covariates such that $x_w = 0$, $\tilde{x}_i^\top \beta$ is the linear predictor for a weekday and $\tilde{x}_i^\top \beta + \beta_w$ the linear predictor for a week-end day
- Therefore, the relative difference τ of the number of trips between a weekend day and a weekday is

$$\tau = \frac{P(Y = k \,|\, weekend, rest) - P(Y = k \,|\, weekday, rest)}{P(Y = k \,|\, weekday, rest)}$$

$$\tau = \frac{e^{\tilde{x}_i^\top \beta + \beta_w} - e^{\tilde{x}_i^\top \beta}}{e^{\tilde{x}_i^\top \beta}} = e^{\beta_w} - 1 = e^{-0.0738} - 1 = -0.071 = -7.1\%$$

- The previous expression indicates also that $\beta_w = \ln(1 + \tau) \approx \tau$. For small values of β_w, the coefficient can be interpreted as the relative difference of the response
- In terms of absolute value, at the sample mean, the absolute difference is $-0.071 \times 4.55 = -0.323$

	(Intercept)	workschl	size	dist	smsa	fulltime
lm	-0.8799070	-2.1504174	0.8284770	-0.012944001	-0.04841164	1.3330729
poisson	-0.5702721	-0.4556281	0.1667228	-0.002199494	-0.03093073	0.2482265

	distnod	realinc	weekend	car
lm	0.028952450	0.13339483	-0.47650465	2.219090
poisson	0.004878584	0.01881775	-0.07382439	1.413325

Econometric example

Variance of the estimator

- The matrix of the second derivatives is:

$$\frac{\partial \ln L}{\partial \beta \partial \beta^{\top}} = -\sum_{i=1}^{n} e^{x_i^{\top}\beta} x_i x_i^{\top} = -X^{\top} diag(e^{x_1^{\top}\beta}, \ldots, e^{x_n^{\top}\beta}) X$$

- The information matrix is the opposite of the hessian (as it doesn't depend on y, it is equal to its expected value)
- The estimated variance of the estimator is therefore:

$$\hat{\mathrm{V}}(\hat{\beta}) = \left(\sum_{i=1}^{n} e^{x_i^{\top}\hat{\beta}} x_i x_i^{\top} \right)^{-1} = \left(\sum_{i=1}^{n} \hat{\mu}_i x_i x_i^{\top} \right)^{-1}$$

- This estimator is consistent only if the distribution of the response is Poisson, and therefore if the conditional variance equals the conditional mean

```r
X <- model.matrix(footrips, data = trips)
H <- t(X)%*%diag(predict(pois_trips, type = "response"))%*%X
round(rbind(sqrt(diag(solve(H))), summary(pois_trips)$coefficients[,2]),4)
```

```
     (Intercept) workschl   size   dist  smsa fulltime distnod realinc weekend
[1,]      0.1252   0.0699 0.0122 0.0015 0.044   0.0263  0.0012   0.006  0.0488
[2,]      0.1252   0.0699 0.0122 0.0015 0.044   0.0263  0.0012   0.006  0.0488
       car
[1,] 0.123
[2,] 0.123
```

Econometric example

Variance of the estimator

- A more general estimator is based on the sandwich formula:

$$\left(\sum_{i=1}^{n} \hat{\mu}_i x_i x_i^\top\right)^{-1} \left(\sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 x_i x_i^\top\right) \left(\sum_{i=1}^{n} \hat{\mu}_i x_i x_i^\top\right)^{-1}$$

- One alternative is to assume that the variance is proportional to the mean: $\mathrm{V}(Y|x) = \phi \mathrm{E}(Y|x)$. ϕ can then be consistently estimated by:

$$\hat{\phi} = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\mu_i}$$

which leads to the third estimator of the covariance matrix:

$$\hat{\mathrm{V}}(\hat{\beta}) = \hat{\phi} \left(\sum_{i=1}^{n} \hat{\mu}_i x_i x_i^\top\right)^{-1}$$

- The "quasi-Poisson" model is obtained by using the log link and the equation above for the variance. It

Econometric example

Variance of the estimator

- The quasi-poisson model can be fitted using `glm` by setting the `family` argument to `quasipoisson`
- The coefficient estimates are the same of those obtained via Poisson

```
1 qpois_trips <- glm(pois_trips, data = trips, family = quasipoisson)
2 round(rbind(summary(qpois_trips)$coefficients[,1], summary(pois_trips)$coefficients[
```

```
     (Intercept) workschl size dist  smsa fulltime distnod realinc weekend  car
[1,]       -0.57    -0.46 0.17    0 -0.03     0.25       0    0.02   -0.07 1.41
[2,]       -0.57    -0.46 0.17    0 -0.03     0.25       0    0.02   -0.07 1.41
```

Econometric example

Variance of the estimator

- Compared to the Poisson model, the standard deviations of the quasi-Poisson model are inflated by a factor which is the square root of the estimate of the variance-mean ratio
- This factor can be estimated using

$$\hat{\mathrm{V}}(\hat{\beta}) = \hat{\phi}\left(\sum_{i=1}^{n} \hat{\mu}_i x_i x_i^{\mathsf{T}}\right)^{-1}$$

where

$$\hat{\phi} = \frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\mu_i}$$

which makes use of the residuals of the Poisson regression:

- The response residual is $y_i - \hat{\mu}_i$
- The Pearson residual are obtained by dividing the response residual by the standard deviation, which is $\sqrt{\hat{\mu}_n}$

Econometric example

Sandwich estimator

- The sandwich estimator can be constructed by extracting from the fitted Poisson model the model matrix (X), the fitted values ($\hat{\mu}_i$) and the response residuals ($y_i - \hat{\mu}_i$)

$$\left(\sum_{i=1}^{n} \hat{\mu}_i x_i x_i\right)^{-1} \left(\sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 x_i x_i^{\mathsf{T}}\right) \left(\sum_{i=1}^{n} \hat{\mu}_i x_i x_i\right)^{-1}$$

```
mu <- predict(pois_trips, type = "response")
e <- residuals(pois_trips, type = "response")
B <- t(X)%*%diag(mu)%*%X
M <- t(X)%*%diag(e^2)%*%X
sand_vcov <- solve(B) %*% M %*% solve(B)
round(rbind(sqrt(diag(sand_vcov)),
            sqrt(diag(sandwich::vcovHC(pois_trips, type = "HC")))),2)
```

```
     (Intercept) workschl size dist smsa fulltime distnod realinc weekend  car
[1,]        0.19     0.12 0.03    0 0.09     0.04       0    0.02     0.1 0.17
[2,]        0.19     0.12 0.03    0 0.09     0.04       0    0.02     0.1 0.17
```

Econometric example

Overdispersion

- A more general expression for the variance is:

$$\sigma_i^2 = \mu_i + \alpha\mu_i^p$$

- We will focus on $p = 1$ which implies that the variance is proportional to the expectation, $\sigma_i^2 = (1 + \alpha)\mu_i$,
- Tests for overdispersion can be done using one of the three test principles: Wald, likelihood ratio and score tests:
  - The first two require the estimation of a more general model that doesn't impose the equality between the conditional mean and the conditional variance
  - The score test requires only the estimation of the constrained (Poisson) model: we consider an auxiliary regression.^[Score tests that are not obtained from auxiliary regressions approach
- The hypothesis of equidispersion (i.e., $\alpha = 0$) can be tested by regressing $\frac{(y_n - \hat{\mu}_n)^2 - y_n}{\hat{\mu}_n}$ on a constant and comparing the Student statistic to the relevant critical value
- In both regressions the p-value is very close to zero so that the equidispersion hypothesis is strongly rejected

```r
y <- trips$trips
e_pears <- residuals(pois_trips, type = "pearson")
resp_areg <- e_pears ^ 2 - y / mu
summary(lm(resp_areg ~ 1))$coefficients
```

```
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 2.314143  0.4060474 5.699194 1.923041e-08
```

```r
summary(lm(resp_areg ~ mu - 1))$coefficients
```

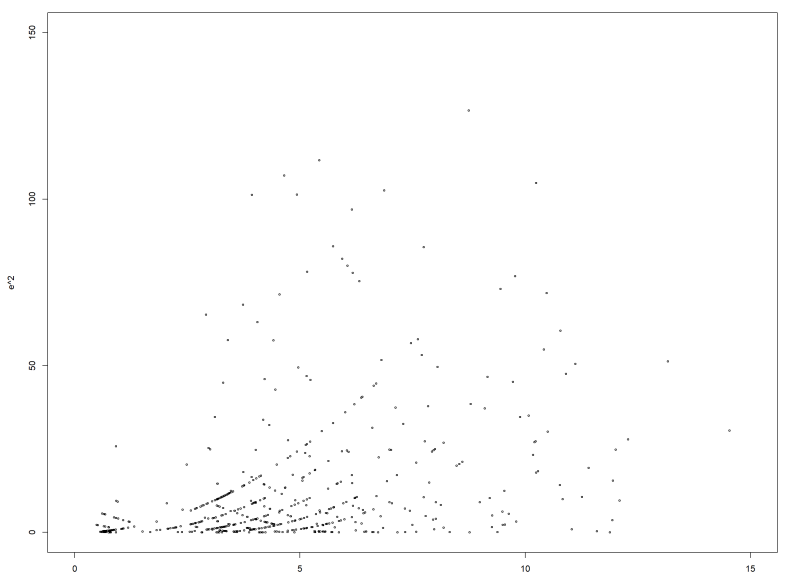
```
    Estimate Std. Error  t value     Pr(>|t|)
mu 0.4670548 0.07646609 6.107998 1.857244e-09
```

Econometric example

Overdispersion

- The overdispersion phenomenon can also be represented by plotting the square of the residuals against the fitted values
- The cloud of points forms a sort of ‘fan’ opening upward meaning that the variance is greater than mean

```
1  plot(mu, e^2, ylim = c(0,150), xlim = c(0,15), cex = 0.5)
```



Econometric example

Overdispersion

- We can handle with overdispersion using a different model: using the negative binomial distribution to describe the distribution of Y conditional to x
- Under this model

$$\mathrm{E}(Y_i \mid x) = \mu_i$$

$$\mathrm{V}(Y_i \mid x) = \mu_i(1 + \mu_i/\delta)$$

- Model fitting can be carried out using the *glm.nb* function

```
1  nb <- MASS::glm.nb(footrips, data = trips)
```

Econometric example

```
1 summary(nb)
```

```
Call:
MASS::glm.nb(formula = footrips, data = trips, init.theta = 2.062238082,
    link = log)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.627807   0.160152  -3.920 8.85e-05 ***
workschl    -0.365094   0.124781  -2.926  0.00343 **
size         0.175526   0.024111   7.280 3.34e-13 ***
dist        -0.001862   0.002473  -0.753  0.45161
smsa        -0.029920   0.081317  -0.368  0.71292
fulltime     0.318461   0.052469   6.069 1.28e-09 ***
distnod      0.005334   0.002432   2.193  0.02832 *
realinc      0.020090   0.013160   1.527  0.12684
weekend     -0.017786   0.088888  -0.200  0.84141
car          1.303900   0.154795   8.423  < 2e-16 ***
```

Econometric example

Overdispersion and excess of zero

- In the previous model output summary the value of θ corresponds to $1/\delta$ and provide support for having a overdispersion problem
- The standard errors are much higher compared to the Poisson model, but remember that, when overdispersion is present, the standard errors of the Poisson model are downward biased
- Finally, the AIC clearly shows that the binomial negative model outperform the poisson one

```
1 rbind(AIC(pois_trips), AIC(nb))
```

```
         [,1]
[1,] 3340.880
[2,] 2779.623
```

Econometric example

Excess of zero

- One of the problem `trips` data set is that the Poisson model is unable to model correctly the probability of $Y = 0$
- Different phenomena can explain zero values:
  - some people may never (or almost never) take a trip out of the house because, for example, of a bad physical condition,
  - some people may take a trip infrequently (for example, twice a week) so that a 0 value may be reported if the survey concerns a Tuesday as the individual has taken a trip on Monday and Thursday
- A viable option to consider (there exists also other alternatives) is the **Hurdle model**
  - a binomial model that explains the fact that the outcome is zero or positive
  - a model that explains the level of the outcome when it is positive
- The Hurdle model is general, meaning that it can be used for several type of outcomes
- In our case we are modelling $P(Y = 0)$ and the $P(Y = k \mid Y > 0)$ using a Poisson or Negative Binomial
- We can fit these models using the **countreg** R package. However, it is on the CRAN and must be install via github

```
#remotes::install_github("https://github.com/r-forge/countreg", subdir="pkg")

hdl_pois <- countreg::hurdle(trips ~ workschl + size + dist + smsa +
```

```
                                fulltime + distnod + realinc + weekend +
                                car, dist = "poisson",
                              zero.dist = "poisson", data = trips)
hdl_nb <- countreg::hurdle(trips ~ workschl + size + dist + smsa +
                                fulltime + distnod + realinc + weekend +
                                car, dist = "negbin",
                              zero.dist = "negbin", data = trips)
```

Econometric example

Model comparison

- It seems that the Hurdle model is responsible for an improved model fit and the preferred distribution choice is the Negative binomial

```
cbind(AIC(pois_trips), AIC(nb), AIC(hdl_pois), AIC(hdl_nb))
```

```
        [,1]     [,2]     [,3]     [,4]
[1,] 3340.88 2779.623 3008.926 2648.908
```

Econometric example

```
1 summary(hdl_nb)
```

```
Call:
countreg::hurdle(formula = trips ~ workschl + size + dist + smsa + fulltime +
    distnod + realinc + weekend + car, data = trips, dist = "negbin",
    zero.dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.3639 -0.6402 -0.3162  0.4879  6.0698

Count model coefficients (truncated negbin with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.065113   0.219796   0.296    0.767
workschl    -0.959514   0.139582  -6.874 6.23e-12 ***
size         0.163293   0.024158   6.759 1.39e-11 ***
dist        -0.002158   0.002435  -0.886    0.376
smsa        -0.134180   0.082164  -1.633    0.102
```

Endogeneity and selection

IV estimator

- Denoting W the matrix of instruments, the IV estimator is

$$\hat{\beta} = (X^\top P_W X)^{-1}(X^\top P_W y)$$

- While the GMM estimator is

$$\hat{\beta} = \left(X^\top W \hat{S}^{-1} W^\top X\right)^{-1} X^\top W \hat{S}^{-1} W^\top y$$

with

  - $$\hat{S} = \sum_i \hat{\epsilon}_n^2 w_i w_i^\top$$
  - $\hat{\epsilon}_i$ being the residual of a consistent estimation

Econometric Example

Cigarette Smoking behaviour

- Source: Mullahy (1997)
- Estimate a demand function for cigarettes which depends on the stock of smoking habits
- This variable is quite similar to a lagged dependent variable and is likely to be endogenous as the unobservable determinants of current smoking behavior should be correlated with the unobservable determinants of past smoking behavior
- The data set, called `cigmales`, contains observations of 6160 males in 1979 and 1980 from the smoking supplement to the 1979 National Health Interview Survey
- The response `cigarettes` is the number of cigarettes smoked daily

```
1 load("cigmales.rda")
2 str(cigmales)
```

```
Classes 'tbl_df', 'tbl' and 'data.frame':   6160 obs. of  11 variables:
 $ cigarettes: num  0 0 0 0 20 4 9 0 30 0 ...
  ..- attr(*, "format.stata")= chr "%9.0g"
 $ habit     : num  0 0 0 0 199 ...
  ..- attr(*, "format.stata")= chr "%9.0g"
 $ price     : num  59.2 58.3 58.2 60.5 58.5 ...
  ..- attr(*, "format.stata")= chr "%9.0g"
```

```
$ restaurant: num  0 0 0 0 0 0 0 0 0 0 ...
 ..- attr(*, "format.stata")= chr "%9.0g"
$ income    : num  8.5 20 6.5 6.5 8.5 12.5 3.5 12.5 12.5 8.5 ...
 ..- attr(*, "format.stata")= chr "%9.0g"
$ age       : num  21 17 17 71 68 28 61 67 66 29 ...
 ..- attr(*, "format.stata")= chr "%9.0g"
$ educ      : num  12 10 10 10 0 15 13.5 10 2.5 12 ...
 ..- attr(*, "format.stata")= chr "%9.0g"
$ famsize   : num  2 1 8 2 2 3 2 3 4 4 ...
 ..- attr(*, "format.stata")= chr "%9.0g"
```

Econometric Example

Cigarette Smoking behaviour

- The covariates are:
  - the habit "stock" `habit`
  - the current state-level average per-pack price of cigarettes `price`
  - a dummy indicating whether there is in the state of residence a restriction on smoking in restaurants `restaurant`
  - the age `age`
  - the number of years of schooling `educ` and their squares
  - the number of family members `famsize`
  - a dummy `race` which indicates whether the individual is white or not
- The external instruments are:
  - cubic terms in `age` and `educ` and their interaction
  - the one-year lagged price of a pack of cigarettes `lagprice`
  - the number of years the state's restaurant smoking restrictions had been in place

Econometric Example

Cigarette Smoking behaviour

- Let’s start to padd the dataset with additional variables (square and cubic terms and interaction age and education)

```
cigmales$age2 <- cigmales$age^2
cigmales$age3 <- cigmales$age^3
cigmales$educ2 <- cigmales$educ^2
cigmales$educ3 <- cigmales$educ^3
cigmales$educage <- cigmales$educ*cigmales$age
```

Econometric Example

Cigarette Smoking behaviour

- The starting point is a basic count model, i.e., a Poisson model with a log-link and robust standard errors
- The IV and the GMM estimators are estimated using the `micsr::expreg` function
- Its main argument is a two-part formula, the first part indicating the covariates and the second part the instruments
- The `method` argument can be set to `"iv"` or `"gmm"` to estimate respectively the instrumental variable and the general method of moments estimators

Econometric Example

```
pois_cig <- glm(cigarettes ~ habit + price + restaurant + income +
                  age + age2 + educ + educ2 + famsize + race,
                data = cigmales, family = quasipoisson)
iv_cig <- expreg(cigarettes ~ habit + price + restaurant + income +
                   age + age2 + educ + educ2 + famsize + race | + price + restaurant
                   age + age2 + age3 + educ + educ2 + educ3 + famsize + race  +  edu
                   lagprice + reslgth,
                 data = cigmales, method = "iv")
gmm_cig <- expreg(cigarettes ~ habit + price + restaurant + income +
                   age + age2 + educ + educ2 + famsize + race | + price + restaurant
                   age + age2 + age3 + educ + educ2 + educ3 + famsize + race  +  edu
                   lagprice + reslgth,
                 data = cigmales, method = "gmm")
```

Econometric Example

Cigarette Smoking behaviour

- For the two flavors of instrumental variable estimators, the coefficient of habit is about half of the one of the Poisson model, indicating that this covariate is positively correlated with the unobserved determinants of smoking

```
cbind(summary(pois_cig)$coefficients[,1],
      summary(iv_cig)$coefficients[,1],
      summary(gmm_cig)$coefficients[,1])
```

```
                    [,1]          [,2]          [,3]
(Intercept)  2.1544259032  0.415306390  0.414421953
habit        0.0054828117  0.003059946  0.003152737
price       -0.0094233402 -0.010550689 -0.008872943
restaurant  -0.0469020436 -0.043135676 -0.061945453
income      -0.0027728438 -0.007599625 -0.006445546
age          0.0087146732  0.099289139  0.092800292
age2        -0.0003108835 -0.001279918 -0.001209366
educ         0.0353404081  0.129910115  0.140816054
educ2       -0.0030505499 -0.008772633 -0.009280623
famsize     -0.0081011513 -0.008456225 -0.011935944
racewhite   -0.0511107480 -0.031066395 -0.090221359
```

Econometric Example

Cigarette Smoking behaviour

- Used in IV and GMM estimation when there are more instruments than endogenous variables
- Null hypothesis: all instruments are valid (they affect the dependent variable only through the endogenous regressors and are uncorrelated with the structural errors).
- Test statistic is based on the sample moments of the instruments
- Low p-value then instruments fail and the model can be misspecifid or the instruments are invalid
- The Sargan test suggest exogeneity of the instruments hypothesis is not rejected

```
1 sargan(gmm_cig)
```

```
    Sargan Test

data:  cigmales
chisq = 7.4694, df = 4, p-value = 0.1131
alternative hypothesis: the moment conditions are not valid
```