Illustration 1: Health Care Expenses and Consultations

Consider the file "CameronTrivedi2010-ch18-health.dta".

1. Present summary statistics for the variable med, both including and excluding null health care expenses.

2. Using only observations from year 1 and considering med as dependent variable and $lcoins, ndisease, female, age, lfam$ and $child$ as explanatory variables, estimate the following models:
2.1. Exponential, based on the Poisson function.
2.2. Exponential, based on the Poisson function and considering only the observations for which med is positive.
2.3. Log-linear, considering only the observations for which med is positive.
2.4. Log-linear, adding 1 to all values of med.

3. Again, consider only observations from year 1. The dependent variable is now the number of medical consultations (mdu).
3.1. Present summary statistics and a table of absolute and relative frequencies for the variable mdu.
3.2. Considering the same explanatory variables as before, estimate:
3.2.1. The Poisson regression model, by maximum likelihood.
3.2.2. The Poisson regression model, by quasi-maximum likelihood.
3.2.3. The Negative Binomial 1 regression model, by maximum likelihood.
3.2.4. The Negative Binomial 2 regression model, by maximum likelihood.
3.2.5. What can be concluded from the two overdispersion tests carried out?
3.3. Consider an individual with the following characteristics: 50 years old, male, family size of 3, no chronic disease. Using the Poisson model estimated before, fill in the table below for the following co-insurance rates: 0%, 50% e 100%.

$coins$:	0	50	100
$E(mdu\|\ldots)$			
$Pr(mdu = 0\|\ldots)$			
$Pr(mdu = 1\|\ldots)$			
$Pr(mdu \geq 2\|\ldots)$			

4. Consider the full sample (all years).
4.1. Check if the panel is balanced or not.
4.2. To explain the number of medical consultations, estimate the following panel data Poisson models:
4.2.1. Pooled.
4.2.2. Random effects.
4.2.3. Fixed effects.
4.3. Test whether the effects are random or fixed.

Illustration 2: Determinants of Firm Debt

Consider the file "CentralBalancos-BP.dta". Our aim is explaning SME's long-term debt (LEV_LT1). Use the following explanatory variables: $SIZE2, COLLAT2, PROF1, GROWTH2$ and AGE.

1. Describe, using summary statistics, SME's capital structure.

2. Find the determinants of long-term debt considering the pooled fractional logit model

3. Considering a firm with *SIZE2* = 13.54, *COLLAT2* = 0.41, *PROF1* = 0.07, *GROWTH2* = 15.03 and *AGE* = 19, predict:
3.1. The proportion of long-term debt issued by the firm.
3.2. The probability of raising debt.
3.3. The proportion of long-term debt issued by the firm conditional on being already using it.

4. Consider a logit and a probit model. Test the validity of the respective functional forms. Independently of the test result, compute the partial effects in the framework of the two models in both the version evaluated at the mean of the explanatory variables and obtained as the mean of the individual partial effects.

5. Using the exponential transformation and a fixed effects Poisson model, estimate a fractional logit model. Compare the results with those of the linear transformation.

Illustration 3: Determinants of Firm Debt (revisited)

Consider the file "CentralBalancos-BP.dta". Our aim is explaning SME's long-term debt (LEV_LT1). Use the following explanatory variables: $SIZE2, COLLAT2, PROF1, GROWTH2$ and AGE.

1. Find the determinants of long-term debt considering the following pooled models:
1.1. Two-part model based on a probit model for the first part and a logit model for the second.
1.2. Tobit model.

2. For each of the previous models, and considering a firm with *SIZE2* = 13.54, *COLLAT2* = 0.41, *PROF1* = 0.07, *GROWTH2* = 15.03 and *AGE* = 19, predict:
2.1. The proportion of long-term debt issued by the firm.
2.2. The probability of raising debt.
2.3. The proportion of long-term debt issued by the firm conditional on being already using it.

Illustration 4: Budget share on tobacco

Consider the file “Tobacco.dta”. The aim is replicating some results of illustration 7.5.4 of Veerbeck (). Ignore the fractional nature of the dependent variable, the budget share on tobacco, designated as share1. The explanatory variables are age, measured in intervals of 10 years, ranging from 0 for age<30 to 4 for age>=60, nadults, number of adults in the household, nkids, number of children aged more the 2 years, nkids2, number of children with age equal or less than 2, lnx, ln of total household expenditure, agelnx, age*lnx, and nadlnx, nadults*lnx. For the first step of the model selection approach, use in addiction, bluecol and whitecol, dummy variables for blue and white collar workers, respectively

1. Consider a tobit model.

2. Consider a two-part model where the first and the second parts are described by a probit and a linear model, respectively.

3. Consider Heckman’s two-step estimator.

Illustration 5: Missing data on LFS / UK

Consider the dataset lfs-22.dat. These data were collected from the *Labour Force Survey* (LFS), applied in the UK, and refer to the period of June-August 1999; for details see Skinner, Stuttard, Durrant & Jenkins (2002), Durrant & Skinner (2006) and Ramalho & Smith (2013). These data have been used to analyse the consequences of the introduction of the minimum wage of 3.6£ per hour in the UK. The dependent variable is wage by hour (*hrrate*) and the explanatory variables are:

- *empmon*: number of months employed
- *part*: =1 if working in part-time
- *socd1234*: =1 if manager, independent professional,, or professor
- *married*: =1 if married
- *hqd6*: =1 if no professional qualification
- *lppltw*: =1 if payment is received by periods shorter than a week
- *size25*: =1 if job place has more than 25 workers

The papers mentioned previously deal with a specific problem of the data: the overrepresentation of low-paid workers in the sample.

Consider a binary version of log(*hrrate*) defined as y=1 for log(*hrrate*)>log(3.6).

1. Estimate both a logit and a probit model, assuming that the sampling is random.

2. Take into account the overrepresentation of low-payed workers.

2.1. Comment on the previous results.
2.2. Stuttard, Durrant e Jenkins (2002) proposed an estimate for the proportion of workers with a wage equal or lower than 3.6£ of 5.5%, while the corresponding proportion in the sample is 11.18%. Use this information to construct weights in order to estimate a probit model by GMM which accounts for the response-based nature of the data.

References:

Durrant, G.B. and Skinner, C. (2006), "Using Data Augmentation to Correct for Non-Ignorable Non-Response When Surrogate Data are Available: An Application to the Distribution of Hourly Pay", *Journal of the Royal Statistical Society, Series A*, 169, 605-623.

Ramalho, E.A. e Smith, R.J. (2013), "Discrete Choice Nonresponse", *Review of Economic Studies*, 80(1), 343-364.

Skinner, C., Stuttard, N., Durrant, G.B. e Jenkins, J. (2002), "The Measurement of Low Pay in the UK Labour Force Survey", *Oxford Bulletin of Economics and Statistics*, 64, 653-676.

Illustration 6: Unemployment duration

Consider the file "CameronTrivedi2005-ch17-unemployement.dta". The aim is explaining unemployment duration, measured in number of two weeks intervals (*spell*), as a function of:

- ui (UI) = 1 if filed UI claim
- reprate (RR) = eligible replacement rate
- disrate (DR) = eligible disregard rate
- tenure (TENURE) = years tenure in lost job
- logwage (LOGWAGE) = log weekly earnings in lost job (1985$)
- other variables listed in McCall (1986) table 2 p.657

the database contains also information on

- CENSOR1 = 1 if re-employed at full-time job
- CENSOR2 = 1 if re-employed at part-time job
- CENSOR3 = 1 if re-employed but left job: pt-ft status unknown
- CENSOR4 = 1 if still jobless

The duration is considered complete when the individual is re-employeed at a full-time job.

1. Present summary statistics for variable $spell$ as well for as the other variables listed previously. Quantify the percentage of censored observations.

2. Estimate and represent in a figure the survival function. Provide the representation of the survival function for those who filled or not a UI claim.

3. Consider the additional covariates:

. gen RRUI = RR*UI
. gen DRUI = DR*UI
. gen LOGWAGE = logwage

Estimate the exponential, Weibull and Gompertz model. Present the coefficients in a table and compare results.

4. Consider the Cox proportional hazard model. Compare the results

5. Perform a simple specification check based on figures for generalized residuals of exponential and Weibull models

6. Incorporate heterogeneity in the analysis. Consider the exponential-gamma and the Weibull inverse Gaussian models. Check if the corresponding generalized residuals are close to a 45º line.

Illustration 7: Medical expenditures

Consider the file "CameronTrivedi2010-ch7-health.dta". The aim is modelling medical expenditure of individuals aged 65 or more, using data from the Medical expenditure panel survey, U.S. The approach of Cameron & Trivedi (2010), chapter 7 will be followed. Total medical expenditure is defined as *totexp*. This illustration will consider as dependent variable the ln of *totexp, ltotexp*. totexp, together with the explanatory variables, is described below:

```
. describe ltotexp suppins totchr age female white

              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------
-------------------------------------------
ltotexp         float   %9.0g                 ln(totexp) if totexp > 0
suppins         float   %9.0g                =1 if has supp priv insurance
totchr          double  %12.0g                # of chronic problems
age             double  %12.0g                Age
female          double  %12.0g                =1 if female
white           double  %12.0g                =1 if white
```

1. Present summary statistics for the variables mentioned previously. For *ltotexp* present also results on major percentiles and illustrate

2. Consider the LAD estimator and obtain the partial effects of the explanatory variables on the conditional median of *totexp* (take into account the retransformation required).

3. Compare the results of the conditional mean estimator with those of quantiles 0.25, 0.50 and 0.75. For the conditional median, consider also the individual significance that results from the use of a robust covariance matrix, that relaxes the identical distribution of errors assumption.

4. Test the presence of heteroskedasticity.

5. Test the equality of *suppins* and *totchr* coefficients, across the quantiles in analysis.

6. Illustrate the coefficients of both OLS and QR with the respective confidence interval.

Illustration 8: Number of doctor visits

Consider the file "CameronTrivedi2010-ch7-DRvisits.dta". The aim is modelling the number of doctor visits (docvis) by the Medicare system by elderly people in 2003. The explanatory variables are:

- private = 1 if having private insurance that supplements Medicare
- totchr = # of chronic conditions
- age = age in years
- female = 1 if female
- white = 1 if white

1. Present summary statistics for the variables mentioned previously. For *docvis* present also results on major percentiles and illustrate.

2. Produce the continuous version of the count response variable using the usual form of jittering. Illustrate and compare with the original variable.

3. Consider a negative binomial of type II for the conditional mean. Estimate the average marginal effects (AME).

4. Consider QR for counts. Estimate the model for the conditional median using 500 replications for the uniform distribution used in the jittering. Estimate also the AME and compare with those over the conditional mean.

5. Consider QR for counts at the .25, 0.50 and .75 quantiles and compare AME of *private* and *totchar* and the respective variability.