Illustration 1: Determinants of Firm Debt

Consider the file "CentralBalancos-BP.dta", which comprises accounting data for Portuguese firms. The aim is to explain the proportion of debt in the firm's capital structure.

1. Describe the variables in the file.

2. Present summary statistics for each variable.

3. Present a table of absolute and relative frequencies by size (micro, small, medium and large firms).

4. Repeat question 3, but also by year.

5. For the three possible dependent variables in a regression analysis of debt determinants, calculate, for each size-based group, their means.

From now on, consider only data for year 1999 and assume that the aim is to explain the proportion of long-term debt in firm's capital structure.

6. Consider the following linear regression model:

$$LEV\_LT1 = \beta_0 + \beta_1 SIZE2 + \beta_2 COLLAT2 + \beta_3 PROF1 + \beta_4 GROWTH2 + \beta_5 AGE + \beta_6 SE + \beta_7 MedE + \beta_8 LE + u$$

6.1. Estimate the proposed model.
6.2. Comment on the individual significance of the model variables, interpret the effects on LEV_LT1 and present theoretical arguments to justify them.

7. Propose and estimate models that allow testing the following hypotheses:
7.1. "The effects on LEV_LT1 of changes in the variable $PROF1$ are not uniform across different size-based groups." – assume that size-based groups have no other influence on LEV_LT1.
7.2. "When only the groups of SME (micro, small and medium enterprises) and large firms are considered, all model parameters differ significantly between the two groups."

8. For the model estimated in question 6, test:
8.1. The assumed model functional form.
8.2. For heteroskedasticity.
8.3. Re-estimate the model using heteroskedasticity-robust and bootstrap estimators of the parameter standard errors. For the latter estimator consider 500 bootstrap samples.

Illustration 2: Estimating the Returns to Schooling

Consider the file "Verbeek2008-ch5-schooling.dta", which comprises data on 3010 North-American men aged between 24-34 years old. Our aim is to estimate the returns to schooling, i.e. how wages change as schooling increase one year.

1. Describe the variables in the file.

2. Present summary statistics for each variable.

3. Consider the following linear regression model:

$$Log(Wage) = \beta_0 + \beta_1 Schooling + \beta_2 Exper + \beta_3 Exper^2 + \beta_4 SMSA + \beta_5 South + u$$

3.1. Estimate the model by OLS.
3.2. Comment on the returns to schooling and on the effects on wages of the other explanatory variables.
3.3. Test the statistical significance of the previous effects (for experience, consider its sample average).

4. Consider again the previous model, but assume that the variable $Schooling$ may be endogenous.
4.1. Test whether $NearCollege$ is a valid instrument for $Schooling$.
4.2. Estimate the model by two-stage least squares using $NearCollege$ as instrument for $Schooling$.
4.3. Compare the returns to schooling obtained by the two estimation methods.
4.4. Test whether $Schooling$ is indeed an endogenous explanatory variable. Implement the test in two alternative ways: standard and robust to heteroskedasticity.

5. Repeat questions 4.2. and 4.4. but now using GMM as estimation method.

6. Consider now $NearCollege$, $DadCollege$ and $MomCollege$ as instruments for $Schooling$.
6.1. Estimate the model by GMM using simultaneously the three instruments and compare with previous results.
6.2. Test whether $Schooling$ is an endogenous explanatory variable.
6.3. Test whether $NearCollege$, $DadCollege$ and $MomCollege$ are valid instruments for $Schooling$.

Illustration 3: Explaining Individual Wages

Consider the file "Verbeek2008-ch10-wages.dta", which comprises a sample of 545 full-time working males who completed their schooling by 1980 and were then followed over the period 1980-1987. Our aim is to test whether collective bargaining is an important determinant of wages.

1. Describe the variables in the file.

2. Present summary statistics for the variables *Wage*, *Schooling*, *Exper*, *Union*, *South* and *Public*.

3. Present a table of relative frequencies for the variables *Union*, *South* and *Public*.

4. Consider the following linear regression model:

$$Log(Wage_{it}) = \beta_0 + \beta_1 Schooling_i + \beta_2 Exper_{it} + \beta_3 Exper_{it}^2 + \beta_4 Union_{it} + \beta_5 South_{it} + \beta_6 Public_{it} + \alpha_i + u_{it}$$

4.1. Present a table with the parameter estimates, and corresponding standard errors, produced by the following methods: pooled OLS, between, random effects, fixed effects and LSDV. Present another table with only the parameter estimates and using ***, ** and * to denote which estimates are significant at the 1%, 5% and 10% levels.
4.2. Test whether the effects are random or fixed.
4.3. Estimate the model using first-differences.
4.4. Add a full set of temporal dummies and their interaction with the variable $Union$ to the model and estimate it by the random effects method. Did workers benefit more from collective bargaining in 1980 or 1987?

5. Assume that $Union$ is contemporaneously related with the error term. Estimate the original model by the fixed effects method using the closest four internal instruments in temporal terms and assuming that, apart from the current period:
5.1. $Union$ is weakly exogenous.
5.2. $Union$ is strictly exogenous.

Illustration 4: Explaining Capital Structure

Consider the file "Verbeek2008-ch10-capitalstructure.dta", which covers the years 1987 to 2001 and comprises 5449 North-American firms. Our aim is testing whether the Trade-Off theory provides a plausible explanation for firms' capital structure.

1. Describe the variables in the file.

2. Consider the following linear regression model:

$$MDR_{it} = \beta_0 + \gamma MDR_{i,t-1} + \beta_1 ebit\_ta_{it} + \beta_2 mb_{it} + \beta_3 dep\_ta_{it} + \beta_4 lnta_{it} + \beta_5 fa\_ta_{it} + \beta_6 rd\_dum_{it} + \beta_7 rd\_ta_{it} + \beta_8 indmedian_{it} + \beta_9 rated_{it} + \alpha_i + u_{it}$$

Present a table with the parameter estimates produced by the following methods: pooled OLS, random effects and fixed effects. Use ***, ** and * to denote which are significant at the 1%, 5% and 10% levels.

3. Estimate the model using the following methods:
3.1. Anderson-Hsiao, using $\Delta MDR_{i,t-2}$ as instrument for $\Delta MDR_{i,t-1}$.
3.2. Arellano-Bond, using all available instruments for $\Delta MDR_{i,t-1}$.
3.3. Arellano-Bond, using a maximum of two lags as instruments for $\Delta MDR_{i,t-1}$.
3.4. Blundell-Bond, using all available instruments for $\Delta MDR_{i,t-1}$.

4. For the model estimated in 3.2, test:
4.1. For autocorrelation.
4.2. Instrument validity, using Sargan test.
4.3. The Trade-Off theory.

Illustration 5: Modelling the Choice Between Two Brands

Consider the file "FransesPaap2001-ch4-brands.dta", which comprises data on the choice between 2 tomato ketchup brands: Heinz and Hunts. Our aim is to evaluate whether the promotional activities developed by both brands have any impact on the probability of consumers choosing one instead of the other.

1. Describe the variables in the file.

2. Present summary statistics for the variables $Heinz, Hunts,\ Dhei,\ Fhei, DFhei,\ Dhun,\ Fhun,\ DFhun,\ Phei$ and $Phun$.

3. Consider the following linear regression model:

$$Pr(Heinz = 1| \dots) = G\left[\beta_0 + \beta_1 Dhei + \beta_2 Fhei + \beta_3 DFhei + \beta_4 Dhun + \beta_5 Fhun + \beta_6 DFhun + \beta_7 log\left(\frac{Phei}{Phun}\right)\right]$$

Present a table with the parameter estimates produced by the following models: logit, probit and cloglog. Use ***, ** and * to denote which are significant at the 1%, 5% and 10% levels.

4. Use the RESET test to assess the models (Wald version; use a single power of the fitted values).

5. Consider only the model(s) which the RESET test suggested being appropriate:
5.1. Apply again the RESET test, but using an LR version (use again a single power of the fitted values).
5.2. Calculate the percentage of correct predictions.

6. Consider only the probit model:
6.1. Complete the following table (the values of Phei e Phun correspond to their sample means):

	I	II	III	IV
Dhei	0	0	0	0
Fhei	0	0	0	0
DFhei	0	1	0	1
Dhun	0	0	0	0
Fhun	0	0	0	0
DFhun	0	0	1	1
Phei (*100)	3.48	3.48	3.48	3.48
Phun (*100)	3.36	3.36	3.36	3.36
$Pr(Heinz = 1\| \dots)$				

6.2. Calculate the mean of the partial effects estimated for each individual in the sample.
6.3. Calculate the partial effects of $DFhun$ for a case where there are no promotional activities and prices are identical for both brands.

6.4. Plot the estimated values for $Pr(Heinz = 1| \dots)$ as a function of the variable $log\left(\frac{Phei}{Phun}\right)$. Consider for the latter variable values in the interval [-0.7;0.7] (at most, one price is twice the other) and compare the following three cases:

$$(Dhei, Fhei, DFhei, Dhun, Fhun, DFhun) = (0,0,0,0,0,0)\ vs.\ (0,0,1,0,0,0)\ vs.\ (0,0,0,0,0,1)$$

Illustration 6: Health Care Expenses and Consultations

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

1. Consider the following binary choice regression model:

$$Pr(dmdu = 1| \dots) = G(\beta_0 + \beta_1 lcoins + \beta_2 ndisease + \beta_3 female + \beta_4 age + \beta_5 lfam + \beta_6 child + \alpha_i)$$

1.1. Describe the variables in the file.
1.2. Present summary statistics for each variable.
1.3. Present a table of relative frequencies for the dependent variable.

2. Present a table with the parameter estimates produced by the following variants of the logit model: pooled, random effects and fixed effects. Use ***, ** and * to denote which are significant at the 1%, 5% and 10% levels.

Illustration 7: Explaining Firm's Credit Ratings

Consider the file "Verbeek2008-ch7-credit.dta", which comprises credit ratings assigned by Standard and Poor's to North-American firms. To simplify the analysis, the ratings were combined in seven categories, ranging from 1 (D - lowest rating) to 7 (AAA - highest rating). It was also considered an aggregation in only two categories: 'investment grade' - rating 4 (BBB) or superior; and 'speculative grade': rating 3 (BB) or inferior.

1. Describe the variables in the file.

2. Present summary statistics for each variable.

3. Estimate the following models:
3.1. Binary logit model for explaining the probability of a firm being classified as 'investment grade'.
3.2. Ordered logit model for explaining the assigned rating.

4. Calculate the probability of a firm (use sample means for the variables) getting the classification of 'investment grade' according to both models.

Illustration 8: Travel Mode Choice

Consider the file "Greene2003-ch21-travelmode.dta", which comprises data concerning travel mode choice for travel between Sydney and Melbourne: air, train, bus or car.

1. Describe the variables in the file.

2. Present summary statistics, by travel mode, for the variables *Mode*, *Ttme*, *GC* and *Hinc*, considering:
2.1. The full sample.
2.2. Only the observations relative to the chosen travel mode.

3. Estimate the following multinomial logit models, considering *air* as the base choice:
3.1. $Pr(Y_i = m| \dots) = G(\beta_{0m} + \beta_{1m} Hinc_i)$, where:

$$Y_i = \begin{cases} 0 \text{ if air} \\ 1 \text{ if train} \\ 2 \text{ if bus} \\ 3 \text{ if } car \end{cases}$$

3.2. $Pr(mode_i = 1| \dots) = G(\beta_{0m} + \beta_{1m} Hinc_i + \beta_2 Ttme_{im} + \beta_3 GC_{im})$.

Illustration 9: Health Care Expenses and Consultations (revisited)

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 10: 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. Describe, using summary statistics, SME's capital structure.

2. Find the determinants of long-term debt considering the following pooled models:
2.1. Fractional logit model.
2.2. Two-part model based on a probit model for the first part and a logit model for the second.
2.3. Tobit model.

3. 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:
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. Using the exponential transformation and a fixed effects Poisson model, estimate a fractional logit model.