Illustration 1: Determinants of Firm Debt

Consider the file “CentralBalancos-BP.dta”, which comprises accounting data for Portuguese firms; for additional details see Ramalho, E.A., J.J.S. Ramalho, J.M.R. Murteira (2011), "Alternative estimating and testing empirical strategies for fractional regression models", *Journal of Economic Surveys*, 25(1), 19-68 (DOI: 10.1111/j.14676419.2009.00602.x). The aim is explaining the proportion of debt in the firm’s capital structure. There are three possible measures for the proportion of debt.

1. Describe the variables in the file.

2. Present summary statistics for each variable. For the case of LEV_ST And LEV_LT, detail also by quantile.

3. Present the histogram of both LEV_ST and LEV_LT.

4. Present the boxplots of both LEV_ST and LEV_LT by size (micro, small, medium and large firms).

5. Present a table of absolute and relative frequencies by size.

6. Repeat question **5.**, but also by year.

7. Consider LEV_ST, LEV_LT, and LEV and calculate, for each size-based group:
7.1. Their means.
7.2. Their means, conditional on the use of debt.

8. Obtain the correlations between LEV_LT, COLLAT, SIZE, PROF, and GROWTH.

9. Test if the mean of LEV_LT equals 0.075 using both a bilateral and an unilateral test.

10. Test if the mean of LEV_LT differs for large firms relative to the remaining firms.

11. Test, using the ANOVA approach, whether the mean of LEV_LT differs significantly across the size groups.

12. Repeat question 11 using a non-parametric approach.

Illustration 2: College students

Consider the file "m255.dta" with information on 1428 responses of college students (complete data on 1365 observations) concerning a survey on the faculty performance; see the variable description on the .dta file. We will use item13 through item24 in our analysis.

1. Present summary statistics for each variable.

2. Present correlations among the variables.

3. Check if the correlations are statistically relevant.

4. Apply a Factorial Analysis (method: **p**rincipal-**c**omponent **f**actors).

5. Obtain a scree-plot and graph the loadings.

6. Apply a rotation varimax method.

7. Generate Factor1 and Factor2.

8. Apply a Factorial Analysis (method: maximum likelihood).

Illustration 3: Determinants of Firm Debt (cont.)

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. Consider the following linear regression model:

$$LEV\_LT = \beta_0 + \beta_1 SIZE + \beta_2 COLLAT + \beta_3 PROF + \beta_4 GROWTH + \beta_5 AGE + u$$

1. Estimate the proposed model.

2. Interpret the effects on LEV_LT of changes in the explanatory variables and present theoretical arguments to justify them.

3. Comment on both the individual and joint significance of the model variables.

4. Test also the joint significance of $PROF$ and $GROWTH$

5. Consider now an augmented version of the first model, that includes the size dummy variables. Interpret the partial effect of each of the added regressors and test whether they are jointly significant.

6. Propose and estimate models that allow testing the following hypotheses:
6.1. "The effects on LEV_LT of changes in the variable $PROF$ are not uniform across different size-based groups." – assume that size-based groups have no other influence on LEV_LT.
6.2. "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."

7. For the model estimated in question 5, test:
7.1. The assumed model functional form.
7.2. For heteroskedasticity.

8. Re-estimate the model in question 5 in a robust to heteroscedasticity version.

Illustration 4: 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*, *Black*, *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:

$$Log(Wage_{it}) = \beta_0 + \beta_1 Schooling_i + \beta_2 Exper_{it} + \beta_3 Exper_{it}^2 + \beta_4 Black_i + \beta_5 Union_{it} + \beta_6 South_{it} + \beta_7 Public_{it} + \alpha_i + u_{it}$$

4.1. Present a table with the parameter estimates, and corresponding standard errors, produced by the following methods: pooled OLS, random effects, fixed effects. Present also a similar table where the individually significant covariates are identified by stars.
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 $Black$ to the model and estimate it by the random effects method.

Illustration 5: 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 $MDR_{i,t-2}$ 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 6: 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 $nz, Hunts$, $Dhei$, $Fhei$, $DFhei$, $Dhun$, $Fhun$, $DFhun$, $Phei$ and $Phun$.

3. Consider the following 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 for the probit model.

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 $log\left(\frac{Phei}{Phu}\right)$. 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 7: 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 share2. 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. Describe the variable of interest, using summary statistics. In particular, present the percentage of 0’s.

2. Consider a tobit model

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