KAN-CEAPV2505U
Econometric Analysis of Firm Data

July 30, 2025

Copenhagen Business School
Department of Economics
Ralf A. Wilke
Mette Franck
2025/2026, Autumn semester

Problem Set 6: Limited Dependent Variable Models

1. There has been much interest in whether the presence of 401(k) pensions plans, available to many workers in the U.S., increase net savings. The data set 401KSUBS.dta contains information on net financial assets ($nettfa$), family income (inc), a binary variable for eligibility in a 401(k) plan ($e401k$), and several other variables.

   (a) What fraction of the families in the sample are eligible for participation in a 401(k) pension plan?

   (b) Estimate a linear probability model explaining 401(k) eligibility in terms of income, age, and gender. Include income and age in quadratic form, and report the results in the usual form.

(c) Would you say 401(k) eligibility is independent of income and age? What about gender? Explain.

(d) Obtain the fitted values from the linear probability model estimated in part (b). Are any fitted values negative or greater than one?

(e) Using the fitted values $\widehat{e410k}_i$ from part (d), define $\widetilde{e401k}_i = 1$ if $\widehat{e401k}_i \geq 0.5$ and $\widetilde{e401k}_i = 0$ if $\widehat{e401k}_i < 0.5$. Out of 9,275 families, how many are predicted to be eligible for a 401(k) plan?

(f) For 5,638 families not eligible for a 401(k), what percentage of these are predicted not to have a 401(k), using the predictor $\widetilde{e401k}_i$? For the 3,637 families eligible for a 401(k) plan, what percentage are predicted to have one?

(g) The overall percent correctly predicted is about 65%. Do you think this is a complete description of how well the model does, given your answers in part (f)?

2. Use GROGGER.dta to answer this question.

(a) Define a binary variable, say *arr86*, equal to unity if a man was arrested at least once during 1986, and zero otherwise. Estimate an LPM relating *arr86* to $pcnv$, $avgsen$, $tottime$, $ptime86$, $inc86$, $black$, $hispan$, and $born60$. Report the usual and heteroskedasticity-robust standard errors. What is the estimated effect on the probability of arrest if $pcnv$ goes from 0.25 to 0.75?

(b) Test the joint significance of $avgsen$ and $tottime$, using a nonrobust and robust test.

(c) Now estimate the model by probit. At the average values of $avgsen$, $tottime$, $inc86$, and $ptime86$ in the sample, and with $black = 1$, $hispan = 0$, and $born60 = 1$, what is the estimated effect on the probability of arrest if $pcnv$ goes from 0.25 to 0.75? Compare this result to the answer in part (a).

(d) For the probit model estimated in part (c), obtain the percent correctly predicted. What is the percent correctly predicted when $arr86 = 0$? When $arr86 = 1$? What do you make of these findings?

(e) In the probit model, add the terms $pcnv^2$, $ptime86^2$, and $inc86^2$ to the model. Are these individually or jointly significant? Describe the estimated relationship between the probability of arrest and $pcnv$. In particular, at what point does the probability of conviction have a negative effect on probability of arrest?

3. Use the data in PENSION.dta for this exercise.

   (a) Estimate a linear model for *pctstck*, where the explanatory variables are *choice*, *age*, *educ*, $female$, *black*, *married*, $finc25$,...,$finc101$, *wealth*89, and $prftshr$ with normal and with heteroscedasticity robust standard errors. The sample contains separate observations for some husband-wife pairs. Compute standard errors of the estimates from the model that account for the cluster correlation within family. (These should also be heteroskedasticity-robust.) Do the standard errors differ from the usual OLS or heteroskedasticity robust standard errors?

   (b) Estimate the model from part (a) with ordered probit. Estimate $E(pctstck|x)$ for a single, nonblack female with 12 years of education who is 60 years old. Assume she has net worth (in 1989) equal to \$150,000 and earns \$45,000 a year, her plan is not profit sharing and $choice = 0$. Compare this with the estimate of $E(pctsck|x)$ from the linear model.

   (c) If you want to choose between the linear model and ordered probit model based on how well each estimates $E(y|x)$, how would you proceed?

   (d) Define a variable $invest = 0$ if $pcstck = 0$, $invest = 1$ if $pcstck = 50$, $invest = 2$ if $pcstck = 100$. Estimate the ordered probit model of part (b) but with *invest* as the dependent variable. Are there any interesting quantities that would differ between using *pcstck* and *invest* as the dependent variables?

These problems are partly taken from the Wooldridge (2020) textbook.