Regresszió, gépi tanulás és oksági elemzés oktatása a tankönyv segítségével

Elek Péter

HUN-REN KRTK KTI és BCE

HUN-REN KRTK KTI – műhely
2024. október 10.

Vélemények

"Az ökonometria és az adattudomány gyönyörű integrációja, amelyben az adatgyűjtéstől és a feltáró elemzéstől egyenes út visz a hagyományos regressziós modellezésig, majd az előrejelzésig és az oksági modellezésig."

- *David Card* Nobel-emlékdíjas professzor, University of California

"Ez nem egy ökonometria, hanem egy adatelemzési tankönyv. Méghozzá egy rendkívül szokatlan darab: egyszerű nyelven íródott, egyszerűsített jelöléseken alapul, és tele van esettanulmányokkal."

- *Beata Javorcik* professzor, University of Oxford

A tankönyv szerkezete

1. Data exploration [statistics] (magyarul is)
2. Regression analysis (magyarul is)
3. Prediction [including machine learning]
4. Causal analysis

Mindegyik részen belül hat fejezet (kb. egy hét kurzus / fejezet)

Használat közgazdasági alap- és mesterszakokon a BCE-n

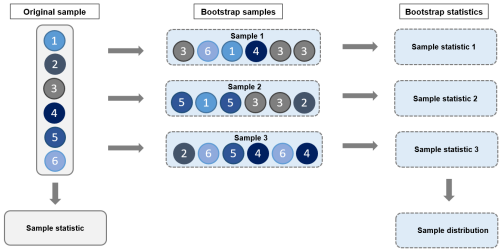
- Alkalmazott közgazdaságtan alapszak:
  - Data exploration + Regression analysis részek
- Közgazdasági elemző mesterszak:
  - Econometrics tárgy: Data exploration + Regression analysis részek
  - Causal analysis tárgy
  - Machine learning in economics tárgy: Prediction rész

Első rész: Data exploration

1. Az adatok eredete [adatgyűjtés és adatminőség]
2. Az adatok előkészítése az elemzésre
3. Feltáró adatelemzés
4. Összehasonlítás és korreláció
5. Hogyan lehet általánosítani az adatokból? [mintavétel, becslés, külső érvényesség]
6. Hipotézisek tesztelése

Első rész jellemzői

- Hangsúly az adatelőkészítésen, feltáró elemzésen és ábrázoláson
- Valószínűségszámítás és következtetéselmélet intuitív módon
- Főleg nagymintás módszerek (kisminták csak említve)
- Viszont: bootstrap, több hipotézis együttes tesztelése, p-hacking is szerepel

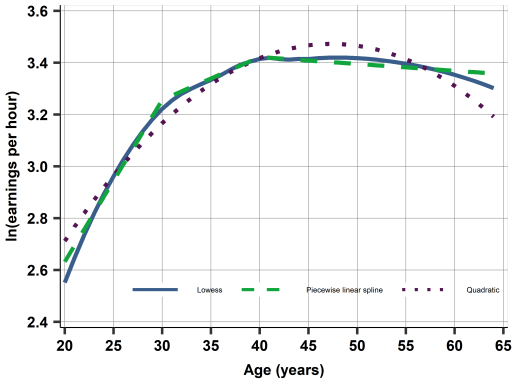


Második rész: Regression analysis

1. Egyváltozós regresszió
2. Bonyolult mintázatok és rendetlen adatok [függvényformák, kiugró értékek]
3. A regresszió eredményeinek általánosítása
4. Többváltozós lineáris regresszió
5. Valószínűségek modellezése
6. Regresszió idősoros adatokkal

Második rész jellemzői

- Szokásosnál nagyobb hangsúly a függvényformák (paraméteres és nemparaméteres) modellezésén és a bináris függő változós modellek értékelésén
- Főleg nagymintás módszerek (kisminták függelékben)
- Idősoros regresszió alapjai egy fejezetben



Harmadik rész: Prediction

1. A framework for prediction [methods of model selection]
2. Model building for prediction [including LASSO]
3. Regression trees
4. Random forests and boosting
5. Probability prediction and classification
6. Forecasting from time series data

Harmadik rész jellemzői

- Gépi tanulás (LASSO, véletlen erdők, boosting, klasszifikációs módszerek) közgazdászok számára emészthető tárgyalása
- Idősoros modellek (ARIMA, VAR)
- Deep learning (egyelőre) hiányzik

Model	RMSE
Linear regression (OLS)	48.1
Linear regression (LASSO)	46.8
Regression Tree (CART)	50.4
Random forest (basic tuning)	44.5
Random forest (atotuned)	44.7
GBM (basic tuning)	44.6
GBM (broad tuning)	44.4

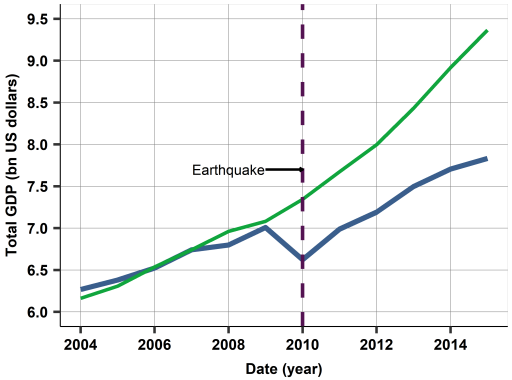
Note: 5-fold cross-validation for all models.

Negyedik rész: Causal analysis

1. A framework for causal analysis
2. Designing and analyzing experiments
3. Regression and matching with observational data [with basic IV and RDD]
4. Difference-in-differences
5. Methods for panel data
6. Appropriate control groups for panel data [with event studies and synthetic control]

Negyedik rész jellemzői

- Fontosabb oksági elemzési módszerek (párosítás, panel, diff-in-diff, eseményelemzés, szintetikus kontroll stb.) felhasználóbarát tárgyalása
- Kitekintés: Local average treatment effect stb.



Total GDP in Haiti and synthetic Haiti

Fejezetek szerkezete – egy példával (11. Valószínűségek modellezése)

- Motiváló példák a fejezet elején és esettanulmányok a fejezet során:
  - Dohányzás és egészségromlás összefüggése egyéni szintű SHARE adatok alapján
  - Jól kalibráltak-e az ausztrál időjárás-előrejelzések?
- Elméleti összefüggések
  - Tanulási eredmények nem technikai megfogalmazása a fejezet elején
  - Főbb összefüggések kiemelve boxokban
  - Részletes kifejtés
- Itt: lineáris valószínűségi modell, logit és probit tárgyalása
  - Értelmezés, marginális hatások, előrejelzés, illeszkedésvizsgálat, kalibrálás
  - Később (17. fejezetben): klasszifikáció gépi tanulási módszerekkel (is) stb.

Esettanulmány a SHARE adatok alapján

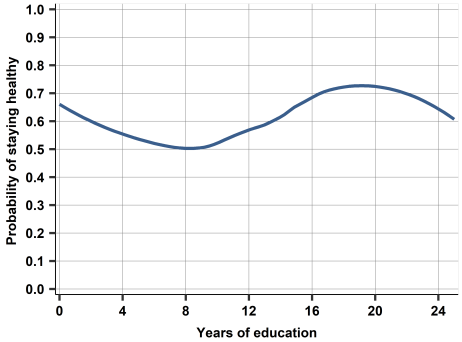
CH11A Does smoking pose a health risk?

Are smokers less likely to remain healthy than non-smokers? How about former smokers who quit?

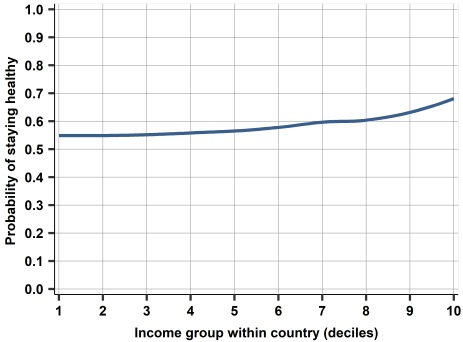
This case study uses the `share-health` data from the SHARE survey (Survey for Health, Aging and Retirement in Europe). We focus on people who were 50 to 60 years old and said to be in good health in 2011. We look at how they rated their health in 2015 and see who remained healthy ahd who changed their answer to not healthy. This case study illustrates probability models. It shows how to estimate and interpret the results of a **linear probability model** and the uses of **logit** and **probit** models. It compares the linear probability estimates to the estimated **marginal differences** from logit and probit. Finally, it illustrates when and how the different models may result in different **predicted probabilities** and how to compare their fit using **Brier-score** and other measures of fit.

Code: **Stata** or **R** or **Python** or ALL. **Data**: share-health. **Graphs**: .png or .eps

Esettanulmány: függvényforma-választás



Staying healthy and years of education



Staying healthy and income group

Esettanulmány: eredmények

Dep.var.: stays healthy	(1) LPM	(2) logit coeffs	(3) logit marginals	(4) probit coeffs	(5) probit marginals
Current smoker	-0.061*	-0.284**	-0.061**	-0.171*	-0.060*
	(0.024)	(0.109)	(0.023)	(0.066)	(0.023)
Ever smoked	0.015	0.078	0.017	0.044	0.016
	(0.020)	(0.092)	(0.020)	(0.056)	(0.020)
Female	0.033	0.161*	0.034*	0.097	0.034
	(0.018)	(0.082)	(0.018)	(0.050)	(0.018)
Years of education (if < 8)	-0.001	-0.003	-0.001	-0.002	-0.001
	(0.007)	(0.033)	(0.007)	(0.020)	(0.007)
Years of education (if $>= 8$ and < 18)	0.017**	0.079**	0.017**	0.048**	0.017**
	(0.003)	(0.016)	(0.003)	(0.010)	(0.003)
Years of education (if $>= 18$)	-0.010	-0.046	-0.010	-0.029	-0.010
	(0.012)	(0.055)	(0.012)	(0.033)	(0.012)
Income group	0.008*	0.036*	0.008*	0.022*	0.008*
	(0.003)	(0.015)	(0.003)	(0.009)	(0.003)
Exercises regularly	0.053**	0.255**	0.055**	0.151**	0.053**
	(0.017)	(0.079)	(0.017)	(0.048)	(0.017)
Age, BMI, Country	YES	YES	YES	YES	YES
Observations	3,109	3,109	3,109	3,109	3,109

Fejezetek szerkezete – egy példával (11. Valószínűségek modellezése)

- Technikai(bb) függelékek a fejezet végén
  - Itt: szaturált modellek, maximum likelihood, marginális hatások részletesen
- Nagyszámú gyakorló feladat
- Nagyszámú adatalapú feladat (főleg az esettanulmányok adatai alapján)

Technikai függelék: "Under the hood"

11.U2 Under the Hood: Maximum Likelihood Estimation and Search Algorithms

Recall that we (or, more accurately, statistical software packages) calculate the coefficients of linear regressions using the method of OLS (ordinary least squares). When applying OLS, we obtain coefficients that lead to the smallest possible (squared) deviations of the predicted values from the actual values. OLS gives formulae for the estimated coefficients, into which we can plug the values of variables in the dataset.

The estimated coefficients of logit and probit models are obtained using a different method, called **maximum likelihood estimation**. We can think of maximum likelihood estimation as another way to get coefficients that give the best fit to the data. It is a more widely applicable method than OLS, but it requires more assumptions to work.

Maximum likelihood estimation starts with assuming a theoretical distribution for the dependent variable (y) conditional on the explanatory variables (x_1, x_2, …). That theoretical distribution is called the likelihood function. It is a conditional distribution so it tells how likely certain values of y are if we observe specific values of x_1, x_2, … Hence the name likelihood. The principle of maximum likelihood estimation is that it produces coefficient values that make the theoretical distribution the most likely distribution for the observed data.

The likelihood is first specified for each observation separately, and then for the entire dataset. Consider a logit model with two explanatory variables, x_1 and x_2. The likelihood of observing a particular value y_i for observations i with particular values x_{1i} and x_{2i} is the probability that we observe that particular y value conditional on observing those particular x values. It is $P[y = 1 \mid x_{1i}, x_{2i}] = \Lambda(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$ if y_i happens to be one; it is $P[y = 0 \mid x_{1i}, x_{2i}] = 1 - P[y = 1 \mid x_{1i}, x_{2i}] = 1 - \Lambda(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$ if y_i happens to be zero.

To write down the likelihood for an observation in a single formula, we can apply a little trick that combines these two probabilities. This trick makes use of the fact that y is a binary variable with values 0 or 1, and thus $1 - y = 1$ when $y = 0$. Written this way, the likelihood of observing a particular value of y_i is

$$\ell_i = P[y = 1 \mid x_{1i}, x_{2i}]^{y_i} \left(1 - P[y = 1 \mid x_{1i}, x_{2i}]\right)^{(1-y_i)} \tag{11.15}$$

Adatalapú feladatok

Data Exercises

Easier and/or shorter exercises are denoted by [*]; harder and/or longer exercises are denoted by [**].

1. Use the `share-health` dataset and pick two countries. Carry out an appropriate analysis to answer the question of our case study in the two countries separately. Compare and discuss your results. [*]

2. Use the `share-health` dataset and pick two countries. Carry out an appropriate analysis to see how the probability of staying healthy is related to gender in the two countries separately. You may or may not want to condition on other covariates; argue for your choices. Compare and discuss your results. [*]

3. Use the `share-health` dataset to examine how exercising is related to income and family factors. In the data, the variable "br015" denotes the frequency of sport and exercise activities.

325

Using this categorical variable, create a binary one using the threshold of your choice, and argue for your choice. Estimate a probability model of your choice with age, income, and family circumstances as explanatory variables. Argue for your choices and discuss your results. [*]

4. Use the `hotels-europe` dataset and pick a city. Use hotel rating to create a binary variable: highly_rated=1 if rating ≤ 4, 0 otherwise. Examine how high rating is related to the other hotel features in the data. Estimate linear probability, logit, and probit models with distance and stars as explanatory variables. Compare coefficients, marginal differences, and predicted probabilities, and discuss your results. [*]

5. Use the `wms-management-survey` dataset and pick a country. Choose one from the 18 score variables on the quality of various aspects of management and create a binary variable for values 4 or greater to denote high management quality in that area. Are firms that export a larger share of their products more likely to have higher quality management in the area you picked? Estimate a linear probability model, a logit and a probit model, and compare their results. Do the same after

Köszönöm a figyelmet!