Data Science Lab - 3

Academic year 2019-2020

Lecturer Falco J. Bargagli-Stoffi

IMT School for Advanced Studies Lucca & KU Leuven

Linear regression

- Linear regression model of the form:

$$y = \beta_0 + x^T \beta + \epsilon \tag{1}$$

where $x \in \mathbb{R}^p$ is an input vector of predictors, $\beta_0 \in \mathbb{R}$ is a constant, $\beta \in \mathbb{R}^p$ is a parameter vector, $\epsilon \in \mathbb{R}$ is an error term, and $y \in \mathbb{R}$ is the output of the model

- Both x and y are measured, whereas β_0, β and ϵ are unknown

- To estimate β_0 and β, one uses a finite set of input-output pairs $z_i = (x_i, y_i)$, $i = 1, \ldots, N$, generated according to the model (1), i.e., each one is of the form

$$y_i = \beta_0 + x_i^T \beta + \epsilon_i \tag{2}$$

- Often, one makes additional assumptions on the error terms ϵ_i:
  - they are independent,
  - they are identically distributed according to a zero-mean Gaussian distribution: $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Estimating the parameters

- The most common ways to estimate β_0 and β are:

    1. Through the Ordinary Least Squares (OLS) method (squared-error loss minimization):

       $$\mathrm{minimize}_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p} \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - x_i^T \beta_i \right)^2 \tag{3}$$

    2. Through the Maximum Likelihood Estimation (MLE):

       $$\mathrm{maximize}_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p} \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} exp \left\{ - \frac{\left(y_i - \beta_0 - x_i^T \beta \right)^2}{2\sigma^2} \right\} \tag{4}$$

    3. Through a GMM estimator arising from the moment condition:

       $$\mathbb{E}\left[x_i (y_i - \beta_0 - x_i^t \beta) \right] = 0 \tag{5}$$

Drawbacks of OLS

- The OLS method has several drawbacks, especially in the case of regression problems with a large number p of predictors:
  - if $p > N$, then problem (3) admits an infinite number of optimal solutions: hence, it is obviously not possible to obtain a good estimate of β
  - usually, an optimal solution β° to problem (3) is not sparse: i.e., all or nearly all its components are different from zero
  - a sparse optimal solution would have better interpretability: i.e., it would highlight which predictors are important to estimate the output

The LASSO estimator

- To solve the problems described in the previous slide, the LASSO estimator has been proposed by Robert Tibshirani in 1996
- Its optimization problem is similar to the one associated with the OLS method
- Additionally, it includes a constraint on the l_1-norm of the parameter vector, which often enforces sparsity of the optimal solution
- LASSO: Least Absolute Shrinkage and Selection Operator
- The term LASSO is inspired also by the “lasso”: a long rope with a noose at one end, used to catch horses and cattle

The LASSO estimator: constrained optimization problem

- For $t > 0$, the constrained optimization problem solved by the LASSO is

$$\text{minimize}_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p} \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - x_i^T \beta \right)^2 \quad \text{subject to } \|\beta\|_1 \leq t \tag{6}$$

where $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ (here, the β_0 term is not included)

- The constraint $\|\beta\|_1 \leq t$ is a “budget constraint”, which limits how well one can fit the data.
- The parameter t has to be tuned using an external procedure (e.g., cross-validation)
- To give a-priori the same importance to each predictor, it is common to standardize them, so that each predictor has unit variance

Logit-LASSO

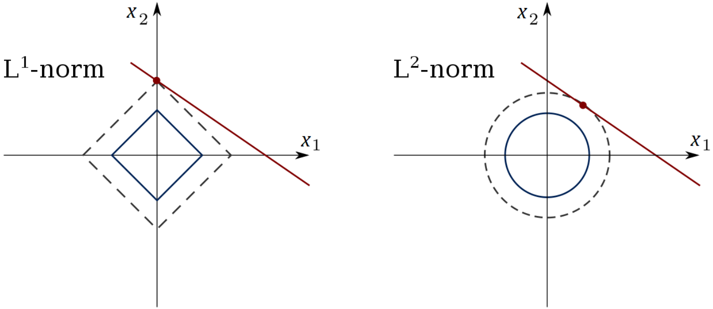
- LASSO is optimized to deal with continuous and discrete variables
- What if you have binary outcome?
- Solution: LOGIT-LASSO

$$\text{minimize}_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p} \quad \tfrac{1}{2N} \textstyle\sum_{i=1}^{N} \left(y_i(\beta_0 + x_i^T \beta) - log(1 + e^{(\beta_0 + x_i^T \beta)}) \right)^2 \\ \text{subject to } \|\beta\|_1 \leq t \tag{7}$$

Ridge regression

- An alternative to the LASSO is ridge regression; in this case, the l_1-norm constraint is replaced by an l_2-norm constraint:

$$\text{minimize}_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p} \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - x_i^T \beta\right)^2 \qquad \text{subject to } \|\beta\|_2 \leq t \tag{8}$$



Comparison between LASSO and ridge regression

- The LASSO has the advantage over ridge regression that it usually provides a sparse optimal solution, i.e., one with only a few components different from zero
- This is due to geometric reasons: more precisely, to the different shapes of the sets $\|\beta\|_1 \leq t$ and $\|\beta\|_2 \leq t$
- For the 2-dimensional case ($p = 2$), an optimal solution is obtained the first time a level curve of the quadratic objective function intersects the set $\|\beta\|_1 \leq t$ (LASSO) or $\|\beta\|_2 \leq t$ (ridge regression)
- In the first case, such an intersection is more likely to occur at a vertex, where one of the components is zero

Model selection and Post-LASSO

- LASSO is also an important tool to perform model selection (i.e., the problem of selecting a "good" subset of the set of p available predictors)
- Compared to other methods, it incorporates model selection inside its optimization problem
    - the opposite approach is to consider different (either "nested" or "non-nested") models with different parameterizations, then to select the "best among the best" such models through an external procedure, such as the classical Akaike's Information Criterion (AIC), which penalizes - the performance being the same - models with large "model complexity" with respect to models with small "model complexity"
    - in the LASSO, the role of "model complexity" is played by the l_1-norm regularization term
- After selecting the predictors through the LASSO, it is common to apply the OLS method to the subset of predictors just selected (Post-LASSO)

Numerical comparison among OLS, LASSO, and Post-LASSO

- Comparison of various regression methods to model the conditional expectation of the log-wage y given the education level z, using an overcomplete dictionary of polynomial approximating functions $x^{(j)}(z)$ depending on the number of years of schooling (example taken from (Belloni and Chernozhukov, 2011)):
- Conventional method: OLS
- s=number of predictors in the selected model (p for the conventional method, much smaller than p for the LASSO and the Post LASSO)
- The conclusion is that - the number of selected regressors being the same - LASSO can perform much better than OLS, and Post LASSO even better!

Stability Selection

- LASSO and RIDGE are potentially unstable to variability in the underlying training sample
- Stability selection (Meinshausen and Bühlmann, 2010) provides an algorithm for performing model selection while controlling the number of false discoveries
- Two main advantages over competing approaches:
  1. It works in the high-dimensional data setting ($p \gg n$)
  2. It provides control on the family-wise error rate in the finite sample setting, which is more practical than an asymptotic guarantee.

Stability Selection: the Algorithm in a Nutshell

- **Input**: dataset $z = z_1, ..., z_n$ and a regularization parameter λ and returns a selection set S^λ

  1. Define a candidate set of regularization parameters Λ and a subsample number k
  2. For each value of $\lambda \in \Lambda$, do:
     - Subsample from z to generate a smaller dataset of size $n/2$, given by $z_{(b)}$
     - Run the selection algorithm on $z_{(b)}$ with parameter λ to obtain a selection set $\hat{S}^\lambda_{(b)}$
  3. Given the selection sets from each subsample, calculate the empirical selection probability for each model component:

$$\hat{\pi}^\lambda_k = \mathbb{P}\{k \in \hat{S}^\lambda\} = \frac{1}{B}\sum_{b=1}^{B}\mathbb{I}\{k \in \hat{S}^\lambda_{(b)}\} \tag{9}$$

- **Output**: given the selection probabilities for each component and for each value of λ, construct the stable set according to:

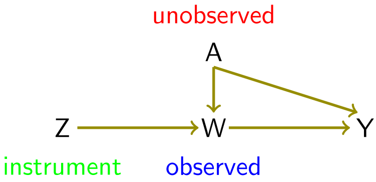
$$\hat{S}^{stable} = \{k : \max_{\lambda\in\Lambda}\hat{\pi}^\lambda_k \geq \pi_{thr}\} \tag{10}$$

Application

Using LASSO to select instruments

Instrumental Variable intuition

- Instrumental variable regression is a very powerful tool in causal inference since it gives the researcher the possibility to solve at the same time 3 issues:
  1. Omitted Variable Bias
  2. Simultaneous equation models (aka reverse causality)
  3. Measurement error
- The idea behind IV



unobserved

A

Z

W

Y

instrument

observed

- IV regression can isolate the causal effect of W by means of an instrument Z, which detect movement in W uncorrelated to D

Formalization and definitions

- $Y_i = \beta^T X_i + \rho s + \eta_i$ where $\eta_i = A_i^T \gamma + \epsilon_i$ and $\mathbb{E}[\eta_i] = A_i^T \gamma$
- $s_i = X_i^T \pi_{10} + \pi_{11} Z_i + \epsilon_{1i}$ is the First Stage
- $Y_i = X_i^T \pi_{20} + \pi_{21} Z_i + \epsilon_{2i}$ is the Reduced Form
- s_i and Y_i are the endogenous variables
- X_i and Z_i are the exogenous variables (X_i are the exogenous covariates)
- From the first stage and the reduced form we have:

$$\rho = \frac{\pi_{21}}{\pi_{11}} = \frac{Cov(Y_i, \tilde{z}_i)}{Cov(s_i, \tilde{z}_i)}$$

Angrist and Krueger (1991): on economic return to education

- Most states want student to enter school in the calendar year in which they turn 6
- Group A: children born in the 4th quarter enter school shortly before they turn 6
- Group B: children born in the 1st quarter enter school at around age 6.5
- Law requires students to remain in school until their 16th birthday
- Therefore, A and B will be in different grades, or have a different length of schooling, when the turn 16

First Stage

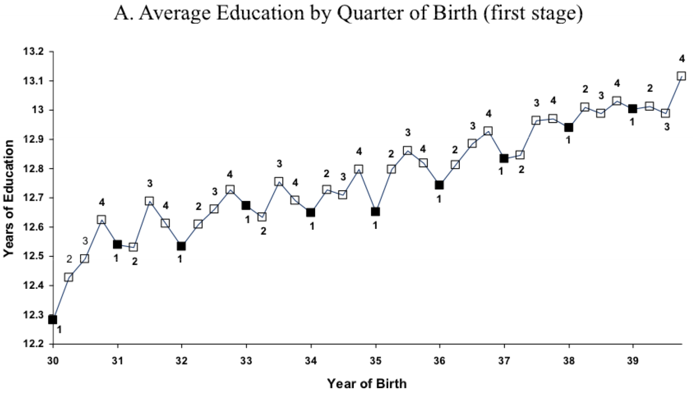
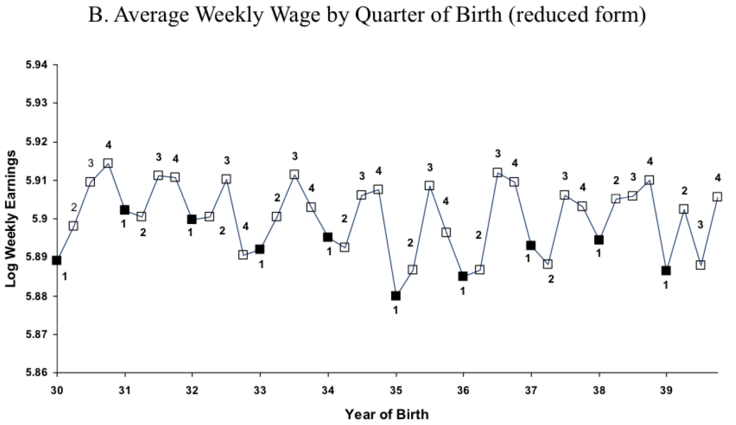


Figure: Average education by quarter of birth. Men born earlier in the calendar year tend to have lower average schooling levels

Reduced Form



B. Average Weekly Wage by Quarter of Birth (reduced form)

Figure: Average earning by quarter of birth. Man born in earlier quarters earn less, on average, than those born later in the year

Two-Stage Least Squares

- You can obtain the reduced form by substituting the 1st stage into the causal relationship

$$\begin{aligned} Y_i &= \alpha^T X_i + \rho\big[X_i \pi_{10} + \pi_{11} z_i + \epsilon_{1i}\big] + \eta_i \\ &= X_i^T\big[\alpha + \rho\pi_{10}\big] + \rho\pi_{11} z_i + \big[\rho\epsilon_{1i} + \eta_i\big] \\ &= X_i^T \pi_{20} + \pi_{21} z_i + \epsilon_{2i} \end{aligned}$$

- Note that this shows again $\rho = \frac{\pi_{21}}{\pi_{11}}$
- As we usually work with samples, we compute
  - $\hat{s}_i = X_i^T \hat{\pi}_{10} + \hat{\pi}_{11} z_i$: first-stage fitted values
  - $Y_i = \alpha^T X_i + \rho\hat{s}_i + \big[\eta_i + \rho(s_i - \hat{s}_i)\big]$: second-stage equation
- The resulting estimator is consistent for ρ because both X_i and $\hat{s}_i$ are uncorrelated with η_i as well as with $(s_i - \hat{s}_i)$
- Intuition: 2SLS retains only the variation in s_i that is generated bt exogenous quasi-experimental variation

Multiple Instrument Case

- Say that we have three instruments z_{1i}, z_{2i}, z_{3i} e.g., dummies for quarter of birth
- The 1st stage is then $s_i = X_i^T \pi_{10} + \pi_{11} z_{1i} + \pi_{12} z_{2i} + \pi_{31} z_{3i} + \epsilon_{1i}$
- The 2nd stage is the same, though $\hat{s}_i$ is now from the above
- Angrist and Krueger (1991) also include interaction terms

$$\begin{aligned} s_i &= X_i^T \pi_{10} + \pi_{11} z_{1i} + \pi_{12} z_{2i} + \pi_{13} z_{3i} \\ &+ \sum_j \big(B_{ij} z_{1i}\big) k_{1j} + \sum_j \big(B_{ij} z_{2i}\big) k_{2j} + \sum_j \big(B_{ij} z_{3i}\big) k_{3j} + \epsilon_{1i} \end{aligned}$$

where B_{ij} is a dummy for year of birth, for $j = 1931 - 39$

Using multiple instruments

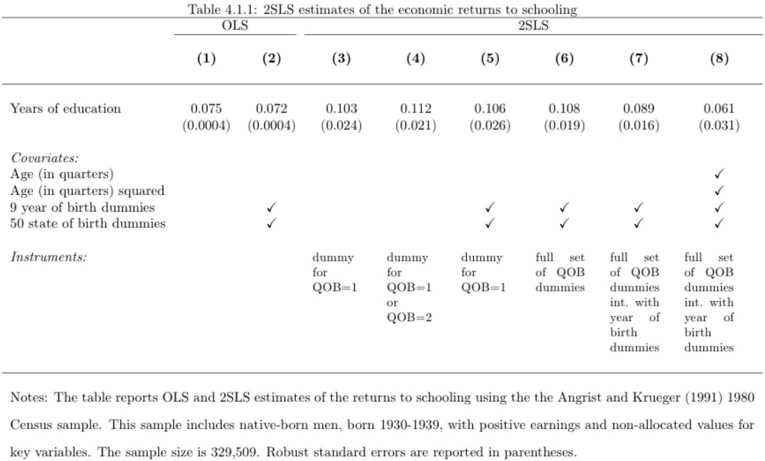


Table 4.1.1: 2SLS estimates of the economic returns to schooling

	OLS		2SLS					
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Years of education	0.075 (0.0004)	0.072 (0.0004)	0.103 (0.024)	0.112 (0.021)	0.106 (0.026)	0.108 (0.019)	0.089 (0.016)	0.061 (0.031)
Covariates:								
Age (in quarters)								✓
Age (in quarters) squared								✓
9 year of birth dummies		✓			✓	✓	✓	✓
50 state of birth dummies		✓			✓	✓	✓	✓
Instruments:			dummy for QOB=1	dummy for QOB=1 or QOB=2	dummy for QOB=1	full set of QOB dummies	full set of QOB dummies int. with year of birth dummies	full set of QOB dummies int. with year of birth dummies

Notes: The table reports OLS and 2SLS estimates of the returns to schooling using the the Angrist and Krueger (1991) 1980 Census sample. This sample includes native-born men, born 1930-1939, with positive earnings and non-allocated values for key variables. The sample size is 329,509. Robust standard errors are reported in parentheses.

Problems with (multiple) instruments

- Bound, Jaeger, and Baker (1995) note, that a possible problem with IV is caused by the selection of *weak* instruments
- Namely, weak instruments are poor predictors of the endogenous question predictor in the first-stage equation
- Weak instrument lead to bias and very large variance in the IV causal estimands
- Moreover,the weak instrument bias tends to get worse as we add more (weak instruments)
- In other words, the bias gets worse when there are many over-identifying restrictions (many instruments compared to endogenous regressors)

Selection problem

We need to select, among the different possible instruments, the ones that have the higher *predictive power* in the First Stage regression

Possible solution: LASSO regression

- Example taken from (Belloni and Chernozhukov, 2011) using data from (Angrist and Krueger, 1991)
- Model of the form

$$\begin{aligned} y_i &= \theta_0 + x_i\theta_1 + c_i^T\gamma + u_i \,, \quad \mathbb{E}\left\{u_i|c_i, z_i\right\} = 0 & (11) \\ x_i &= z_i^T\beta + c_i^T\delta + v_i \,, \quad \mathbb{E}\left\{v_i|c_i, z_i\right\} = 0 & (12) \end{aligned}$$

  where, for each person i, y_i indicates wage, x_i denotes education, c_i indicates a vector of control variables, and z_i denotes a vector of instrumental variables that affect education but do not directly affect the wage
- u_i and v_i are error terms
- In the specific problem, x and u are correlated, hence the OLS estimate of θ_1 from equation (11) - which does not use the vector of instrumental variables - is biased

Application of the LASSO to instrumental variable selection

- The vector of instrumental variables can be used to obtain an unbiased estimate of θ_1, e.g., through the following two-stage regression procedure:
    - first stage: regression of the x_i's from equation (12), using the c_i's and the z_i's
    - second stage: regression of the y_i's from equation (11), using the c_i's and the estimates of the x_i's obtained in the first stage
- In this context, the LASSO can be used to do instrumental variable selection, possibly improving the estimate of θ_1 (see the numerical results in (Belloni and Chernozhukov, 2011))
- A similar application of LASSO - this time in control variable selection - is done in (Belloni et al., 2014), using data from (Acemoglu et al. 2001), to do control variable selection when estimating the effect of institutions on output, using mortality rates for early European settlers as an instrument for institution quality

Bibliography

- Alexandre Belloni and Victor Chernozhukov, “High Dimensional Sparse Econometric Models: An Introduction,” in Inverse Problems and High-Dimensional Estimation, vol. 203 of the series Lecture Notes in Statistics, pp. 121-156, 2011

- Trevor Hastie, Robert Tibshirani, and Martin Wainwright, “Statistical Learning with Sparsity: The Lasso and Generalizations,” CRC Press, 2015 (Chapters 2 and 11)

- Ryan Tibshirani and Larry Wasserman, “Sparsity and the LASSO,” Lecture notes for the course “Statistical Machine Learning,” Carnegie Mellon University, Spring 2015

- Joshua D. Angrist and Alan B. Krueger, “Does Compulsory School Attendance Affect Schooling and Earnings?,” Quarterly Journal of Economics, vol. 106, pp. 979-1014, 1991

- Robert Tibshirani, “Regression Shrinkage and Selection via the Lasso,” Journal of the Royal Statistical Society, Series B, volume 58, pp. 267-288, 1996