Welcome

Course: High-Dimensional Econometrics

Instructor: Dr David T. Jacho-Chavez

Start Date: Monday, January 6th, 2025 **End Date:** Friday, January 10th, 2025

Server: https://server.davidjachochavez.org

GitHub Repository: https://github.com/DTJCML/HD-Metrics

Date		Schedule
Monday, January 6th		02:00 pm - 06:00 pm
Tuesday, January 7th		08:00 am - 11:45 am | 02:15 pm - 06:00 pm
Wednesday, January 8th		08:00 am - 11:45 am | 02:15 pm - 06:00 pm
Thursday, January 9th		08:00 am - 11:45 am | 02:15 pm - 06:00 pm
Friday, January 10th		08:00 am - 11:45 am | 02:15 pm - 06:00 pm

Disclaimer

All the materials in this course are taken from these sources and the original authors hold *all* text, figures, and computer code copyrights:

1. Jeffrey M. Wooldridge's '**Introductory Econometrics: A Modern Approach**' textbook.
2. Bruce Hansen's '**Econometrics**' online textbook.
3. Sergio Correia's **various** Stata packages.
4. Fernando Rios-Avila's **various** Stata packages.
5. Asjad Naqvi's **stata-schemepack** package.
6. Mingze Gao's **specurve** package.
7. Various chapters of '**An Introduction to Statistical Learning: With Applications in R**' book by Gareth James, Daniela Witten, Trevor Hastie and Rob Tibshirani.
8. David Drukker and Di Lu's '**An introduction to the lasso in Stata**.'
9. StataCorp. 2021. '**Stata Statistical Software: Release 17**.' College Station, TX: StataCorp LLC.

Exam

Date: Sunday, January 12th, 2025
Time: 09:00 am - 09:00 pm [12 hours]
Format: Electronic - Jupyter Notebook - 100%

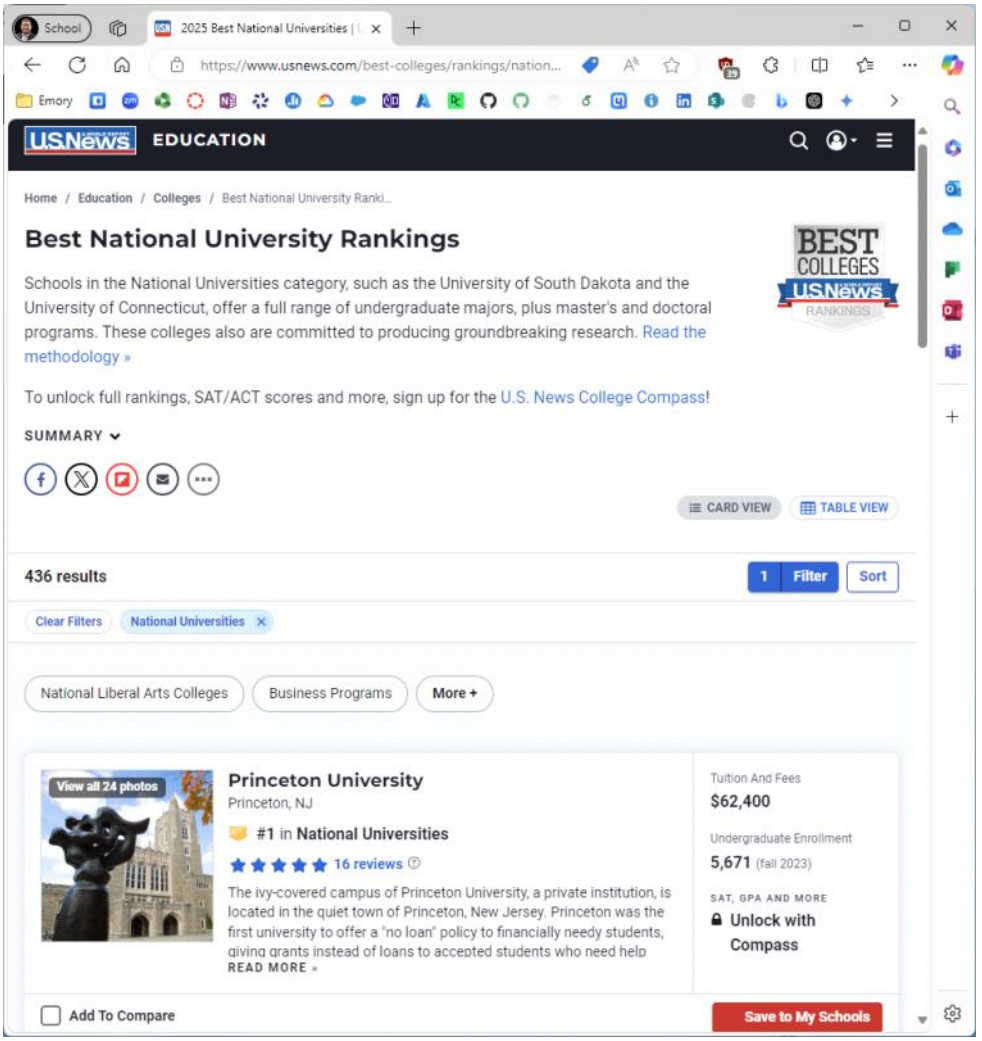
- Part 1: Theory - 60% [32% corresponds to questions in Exercises]
- Part 2: Practice - 40%

Final Score Calculation: (Exam Score/100) × 18 + (Data Camp points)
Guidelines:

- You can use your own notes, books, the internet, artificial intelligence, etc.
- The top-2 final scores among those who work individually will earn my recommendation letter for graduate school for up to 24 months, i.e., January 2027.
- Submit your `student_code.ipynb` to Manuel Vasquez [email: manuel.vasquez@bcrp.gob.pe] by 09:00 pm on Sunday, January 12th, 2025.

Emory University

Source: [US News - 2025 Best National Universities]



1. Princeton University
2. MIT
3. Harvard University
4. Stanford University
5. Yale University
6. California Institute of Technology
7. Duke University
8. Johns Hopkins University
9. Northwestern University
10. University of Pennsylvania
11. Cornell University
12. University of Chicago
13. Brown University
14. Columbia University
15. Dartmouth College
16. UCLA
17. Berkeley
18. Rice University
19. University of Notre Dame
20. Vanderbilt University
21. Carnegie Mellon University
22. University of Michigan
23. Washington University in St. Louis
24. **Emory University**
25. Georgetown University
26. University of Virginia
27. University of North Carolina - Chapel Hill
28. University of Southern California
29. University of California, San Diego
30. New York University

LinkedIn

EMORY
Department of Economics

Follow us on LinkedIn

emory-university-department-of-economics

Regression

1. Find $\mathbb{E}\big[\mathbb{E}\big[\mathbb{E}\big[Y \mid X_1, X_2, X_3\big] \mid X_1, X_2\big] \mid X_1\big]$
2. Suppose that the random variables Y and X only take the values 0 and 1 , and have the following joint probability distribution

	$X = 0$	$X = 1$
$Y = 0$	.1	.2
$Y = 1$	.4	.3

   Find $\mathbb{E}[Y \mid X]$, $\mathbb{E}\big[Y^2 \mid X\big]$ and var$[Y \mid X]$ for $X = 0$ and $X = 1$.

3. True or False?

   (a) If $Y = X\beta + e, X \in \mathbb{R}$, and $\mathbb{E}[e \mid X] = 0$, then $\mathbb{E}\big[X^2 e\big] = 0$.

   (b) If $Y = X\beta + e, X \in \mathbb{R}$, and $\mathbb{E}[Xe] = 0$, then $\mathbb{E}\big[X^2 e\big] = 0$.

   (c) If $Y = X'\beta + e$ and $\mathbb{E}[e \mid X] = 0$, then e is independent of X.

   (d) If $Y = X'\beta + e$ and $\mathbb{E}[Xe] = 0$, then $\mathbb{E}[e \mid X] = 0$.

   (e) If $Y = X'\beta + e$, $\mathbb{E}[e \mid X] = 0$, and $\mathbb{E}\big[e^2 \mid X\big] = \sigma^2$, then e is independent of X.

4. Take the homoskedastic model

$$\begin{aligned} Y &= X_1'\beta_1 + X_2'\beta_2 + e \\ \mathbb{E}[e \mid X_1, X_2] &= 0 \\ \mathbb{E}\big[e^2 \mid X_1, X_2\big] &= \sigma^2 \\ \mathbb{E}[X_2 \mid X_1] &= \Gamma X_1 \end{aligned}.$$

   Assume $\Gamma \neq 0$. Suppose the parameter β_1 is of interest. We know that the exclusion of X_2 creates omited variable bias in the projection coefficient on X_2. It also changes the equation error. Our question is: what is the effect on the homoskedasticity property of the induced equation error? Does the exclusion of X_2 induce heteroskedasticity or not? Be specific.

5. Let $\hat{\beta}_n = \big(\boldsymbol{X}_n'\boldsymbol{X}_n\big)^{-1}\boldsymbol{X}_n'\boldsymbol{Y}_n$ denote the OLS estimate when $\boldsymbol{Y}_n$ is $n \times 1$ and $\boldsymbol{X}_n$ is $n \times k$. A new observation $\big(Y_{n+1}, X_{n+1}\big)$ becomes available. Prove that the OLS estimate computed using this additional observation is

$$\hat{\beta}_{n+1} = \hat{\beta}_n + \frac{1}{1 + X_{n+1}'\big(\boldsymbol{X}_n'\boldsymbol{X}_n\big)^{-1}X_{n+1}}\big(\boldsymbol{X}_n'\boldsymbol{X}_n\big)^{-1}X_{n+1}\big(Y_{n+1} - X_{n+1}'\hat{\beta}_n\big).$$

6. Prove that R^2 is the square of the sample correlation between $\boldsymbol{Y}$ and $\widehat{\boldsymbol{Y}}$.
7. For the intercept-only model $Y_i = \beta + e_i$, show that the leave-one-out prediction error is

$$\tilde{e}_i = \left(\frac{n}{n-1}\right)\big(Y_i - \bar{Y}\big).$$

8. The observations are $\big(Y_i, X_{1i}, X_{2i}\big), i = 1, \ldots, n$. You estimate two least squares regressions.

$$\begin{aligned} Y_i &= X_{1i}'\tilde{\beta}_1 + \tilde{e}_i \\ Y_i &= X_{1i}'\hat{\beta}_1 + X_{2i}'\hat{\beta}_2 + \hat{e}_i' \end{aligned}$$

   and calculate the residual variance estimates

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \tilde{e}_i^2$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i^2 \quad .$$

Show that for any $w \in (0,1)$, there is a constant $a \in (0,1)$ such that

$$\frac{1}{n}\sum_{i=1}^{n} \left(w\hat{e}_i + (1-w)\tilde{e}_i\right)^2 = (1-a)\hat{\sigma}^2 + a\tilde{\sigma}^2.$$

Find the constant a.

9. Consider a regression of a $n \times 1$ vector of responses $\mathbf{Y}$ on a $n \times k$ matrix of explanatory variables $\mathbf{X}$. Let $\mathbf{M} = \mathbf{I}_n - \mathbf{P}$, where $\mathbf{P}$ is the projection matrix $\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$, and $\mathbf{I}_n$ is the identity matrix of order n. Suppose one adds a new explanatory variable, $\mathbf{Z}$, to the regression, so there are now $(k+1)$ regressors. Show that the new residual sum of squares is given by $\mathbf{Y}'\mathbf{M}\mathbf{Y} - b^2\mathbf{Z}'\mathbf{M}\mathbf{Z}$ where $b = \mathbf{Z}'\mathbf{M}\mathbf{Y}/\mathbf{Z}'\mathbf{M}\mathbf{Z}$ is the OLS coefficient on $\mathbf{Z}$.

Regression (Cont.)

1. For some integer k, set $\mu_k = \mathbb{E}[Y^k]$.

   (a) Construct an estimator $\hat{\mu}_k$ for μ_k.

   (b) Show that $\hat{\mu}_k$ is unbiased for μ_k.

   (c) Calculate the variance of $\hat{\mu}_k$, say $\text{var}[\hat{\mu}_k]$. What assumption is needed for $\text{var}[\hat{\mu}_k]$ to be finite?

   (d) Propose an estimator of $\text{var}[\hat{\mu}_k]$.

2. Let $\mu = \mathbb{E}[Y]$, $\sigma^2 = \mathbb{E}\left[(Y-\mu)^2\right]$ and $\mu_3 = \mathbb{E}\left[(Y-\mu)^3\right]$ and consider the sample mean $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Find $\mathbb{E}\left[(\bar{Y}-\mu)^3\right]$ as a function of μ, σ^2, μ_3 and n.

3. Now assume that

$$\sqrt{n}\begin{pmatrix}\hat{\theta}_1 - \theta_1 \\ \hat{\theta}_2 - \theta_2\end{pmatrix} \xrightarrow[d]{} \text{N}(0, \Sigma)$$

   where $\| \Sigma \| < \infty$, then find the asymptotic distribution of the following statistics

   (a) $\hat{\theta}_1\hat{\theta}_2$
   (b) $\exp(\hat{\theta}_1 + \hat{\theta}_2)$
   (c) If $\theta_2 \neq 0$, $\hat{\theta}_1/\hat{\theta}_2^2$
   (d) $\hat{\theta}_1^3 + \hat{\theta}_1\hat{\theta}_2^2$

4. Take a regression model $Y = X\beta + e$ with $\mathbb{E}[e \mid X] = 0$ and i.i.d. observations (Y_i, X_i) and scalar X. The parameter of interest is $\theta = \beta^2$. Consider the OLS estimators $\hat{\beta}$ and $\hat{\theta} = \hat{\beta}^2$.

   (a) Find $\mathbb{E}[\hat{\theta} \mid \boldsymbol{X}]$ using our knowledge of $\mathbb{E}[\hat{\beta} \mid \boldsymbol{X}]$ and $V_{\hat{\beta}} = \text{var}[\hat{\beta} \mid \boldsymbol{X}]$. Is $\hat{\theta}$ biased for θ ?

   (b) Suggest an (approximate) biased-corrected estimator $\hat{\theta}^*$ using an estimator $\hat{V}_{\hat{\beta}}$ for $V_{\hat{\beta}}$.

   (c) For $\hat{\theta}^*$ to be potentially unbiased, which estimator of $V_{\hat{\beta}}$ is most appropriate?

   Under which conditions is $\hat{\theta}^*$ unbiased?

5. Consider an i.i.d. sample $\{Y_i, X_i\} i = 1, \ldots, n$ where X_i is $k \times 1$. Assume the linear conditional expectation model $Y = X'\beta + e$ with $\mathbb{E}[e \mid X] = 0$. Assume that $n^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_k$ (orthonormal regressors). Consider the OLS estimator $\hat{\beta}$.

   (a) Find $\boldsymbol{V}_{\hat{\beta}} = \text{var}[\hat{\beta}]$

   (b) In general, are $\hat{\beta}_j$ and $\hat{\beta}_\ell$ for $j \neq \ell$ correlated or uncorrelated?

   (c) Find a sufficient condition so that $\hat{\beta}_j$ and $\hat{\beta}_\ell$ for $j \neq \ell$ are uncorrelated.

6. Take the model in vector notation

$$\begin{aligned}\boldsymbol{Y} &= \boldsymbol{X}\beta + \boldsymbol{e} \\ \mathbb{E}[\boldsymbol{e} \mid \boldsymbol{X}] &= 0 \\ \mathbb{E}[\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}] &= \Omega\end{aligned}.$$

   Assume for simplicity that Ω is known. Consider the OLS and GLS estimators $\hat{\beta} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Y}\right)$ and $\tilde{\beta} = \left(\boldsymbol{X}'\Omega^{-1}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\Omega^{-1}\boldsymbol{Y}\right)$. Compute the (conditional) covariance between $\hat{\beta}$ and $\tilde{\beta}$:

$$\mathbb{E}\left[(\hat{\beta} - \beta)(\tilde{\beta} - \beta)' \mid \boldsymbol{X}\right].$$

   Find the (conditional) covariance matrix for $\hat{\beta} - \tilde{\beta}$:

$$\mathbb{E}\left[(\hat{\beta} - \tilde{\beta})(\hat{\beta} - \beta)' \mid \boldsymbol{X}\right].$$

7. The model is $Y = X\beta + e$ with $\mathbb{E}[e \mid X] = 0$ and $X \in \mathbb{R}$. Consider the two estimators

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$$
$$\tilde{\beta} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{X_i}.$$

   (a) Under the stated assumptions are both estimators consistent for β ?

   (b) Are there conditions under which either estimator is efficient?

8. Take the linear model $Y = X\beta + e$ with $\mathbb{E}[e \mid X] = 0$ and $X_i \in \mathbb{R}$. Consider the estimator

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i^3 Y_i}{\sum_{i=1}^{n} X_i^4}.$$

   Find the asymptotic distribution of $\sqrt{n}(\hat{\beta} - \beta)$ as $n \to \infty$.

9. The model is $y = X'\beta + e$ with $\mathbb{E}[e \mid X] = 0$. An econometrician is worried about the impact of some unusually large values of the regressors. The model is thus estimated on the subsample for which $|X_i| \le c$ for some fixed c. Let $\tilde{\beta}$ denote the OLS estimator on this subsample. It equals

$$\tilde{\beta} = \left(\sum_{i=1}^{n} X_i X_i' \mathbb{1}\{|X_i| \le c\}\right)^{-1} \left(\sum_{i=1}^{n} X_i Y_i \mathbb{1}\{|X_i| \le c\}\right).$$

   (a) Show that $\tilde{\beta} \xrightarrow[p]{} \beta$.

   (b) Find the asymptotic distribution of $\sqrt{n}(\tilde{\beta} - \beta)$.

Binary Response

1. For a binary response y, let $\bar{y}$ be the proportion of ones in the sample (which is equal to the sample average of the y_i). Let $\hat{q}_0$ be the percent correctly predicted for the outcome $y = 0$ and let $\hat{q}_1$ be the percent correctly predicted for the outcome $y = 1$. If $\hat{p}$ is the overall percent correctly predicted, show that $\hat{p}$ is a weighted average of $\hat{q}_0$ and $\hat{q}_1$:

$$\hat{p} = (1 - \bar{y})\hat{q}_0 + \bar{y}\hat{q}_1.$$

   In a sample of 300 , suppose that $\bar{y} = .70$, so that there are 210 outcomes with $y_i = 1$ and 90 with $y_i = 0$. Suppose that the percent correctly predicted when $y = 0$ is 80 , and the percent correctly predicted when $y = 1$ is 40 . Find the overall percent correctly predicted.

2. Emily estimates a probit regression setting her dependent variable to equal *$Y = 1$* for a purchase and *$Y = 0$* for no purchase. Using the same data and regressors, Jacob estimates a probit regression setting the dependent variable to equal *$Y = 1$* if there is no purchase and *$Y = 0$* for a purchase. What is the difference in their estimated slope coefficients?
3. Jackson estimates a logit regression where the primary regressor is measured in dollars. Julie estimates a logit regression with the same sample and dependent variable, but measures the primary regressor in thousands of dollars. What is the difference in the estimated slope coefficients?
4. For the logistic distribution $\Lambda(x) = (1 + \exp(-x))^{-1}$ verify that
   (a) $\frac{d}{dx}\Lambda(x) = \Lambda(x)(1 - \Lambda(x))$.
   (b) $h_{\text{logit}}(x) = \frac{d}{dx}\log\Lambda(x) = 1 - \Lambda(x)$.
   (c) $H_{\text{logit}}(x) = -\frac{d^2}{dx^2}\log\Lambda(x) = \Lambda(x)(1 - \Lambda(x))$.
   (d) $\left|H_{\text{logit}}(x)\right| \leq 1$.
5. For the normal distribution $\Phi(x)$ verify that
   (a) $h_{\text{probit}}(x) = \frac{d}{dx}\log\Phi(x) = \lambda(x)$ where $\lambda(x) = \phi(x)/\Phi(x)$.
   (b) $H_{\text{probit}}(x) = -\frac{d^2}{dx^2}\log\Phi(x) = \lambda(x)(x + \lambda(x))$.

Multinomial Response

1.

a. For estimating the mean of a nonnegative random variable y using a random sample $\{y_i\}_{i=1}^n$, the Poisson quasi-log likelihood for a random draw is $\gamma_i(\mu) = y_i \log(\mu) - \mu, \mu > 0$ (where terms not depending on μ have been dropped). Letting $\mu_0 \equiv \mathbb{E}(y_i)$, we have $\mathbb{E}[\ell_i(\mu)] = \mu_0 \log(\mu) - \mu$. Show that this function is uniquely maximized at $\mu = \mu_o$.

b. The gamma (exponential) quasi-log likelihood is $\ell_i(\mu) = -y_i/\mu - \log(\mu), \mu > 0$. Show that $\mathbb{E}[\ell_i(\mu)]$ is uniquely maximized at $\mu = \mu_o$.

2. In a balanced panel data setting, i.e. $\left\{\{y_{it}, \mathbf{x}_{it}\}_{t=1}^T\right\}_{i=1}^n$, consider an unobserved effects model for count data with exponential regression function

$$\mathbb{E}(y_{it} \mid \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, c_i) = c_i \exp(\mathbf{x}_{ii}\boldsymbol{\beta}).$$

If $\mathbb{E}(c_i \mid \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}) = \exp(\alpha + \bar{\mathbf{x}}_i \gamma)$, find $\mathbb{E}(y_{it} \mid \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$ where $\bar{\mathbf{x}}_i = T^{-1}(\mathbf{x}_{i1} + \cdots + \mathbf{x}_{iT})$.

3. Let patents be the number of patents applied for by a firm during a given year. Assume that the conditional expectation of patents given sales and RD is

$$\mathbb{E}(\ patents \mid sales\ , RD) = \exp[\beta_0 + \beta_1 \log(\ sales\ ) + \beta_2 RD + \beta_3 RD^2],$$

where $sales$ is annual firm sales and RD is total spending on research and development over the past 10 years.

a. How would you estimate the β_j ? Justify your answer by discussing the nature of patents.

b. How do you interpret β_1 ?

c. Find the partial effect of RD on $\mathbb{E}(\ patents\ \mid\ sales, RD)$.

High-Dimensional Models

1. Take the model $Y = X'\beta + e$ with $\mathbb{E}[Xe] = 0$. Define the ridge regression estimator

$$\hat{\beta} = \left(\sum_{i=1}^{n} X_i X_i' + \lambda \boldsymbol{I}_k\right)^{-1} \left(\sum_{i=1}^{n} X_i Y_i\right),$$

   here $\lambda > 0$ is a fixed constant.
   (a) Find $\mathbb{E}[\hat{\beta} \mid \boldsymbol{X}]$.
   (b) Is $\hat{\beta}$ biased for β ?
   (c) Find the probability limit of $\hat{\beta}$ as $n \to \infty$.
   (d) Is $\hat{\beta}$ consistent for β ?
   (e) Derive var$\left[\hat{\beta} \mid \boldsymbol{X}\right]$.
2. Show that the ridge regression estimator can be computed as least squares applied to an augmented data set. Take the original data ($\boldsymbol{Y}$, $\boldsymbol{X}$). Add p 0's to $\boldsymbol{Y}$ and p rows of $\sqrt{\lambda}\boldsymbol{I}_p$ to $\boldsymbol{X}$, apply least squares, and show that this equals $\hat{\beta}_{\text{ridge}}$.
3. Which estimator produces a higher regression R^2, least squares or ridge regression?

Quick Reminders

The **proportionate change** in v moving from v_0 to v_1 is

$$\frac{v_1 - v_0}{v_0} = \Delta v / v_0,$$

assuming of course that $v_0 \neq 0$. The **percentage change** in v moving from v_0 to v_1 is simply 100 times the proportionate change:

$$\%\Delta v = 100(\Delta v / v_0),$$

the notation "$\%\Delta v$" is read as "the percentage change in v."
Let v_0 and v_1 be positive values, then, it can be shown that

$$\log(v_1) - \log(v_0) \approx \frac{v_1 - v_0}{v_0} = \Delta v / v_0,$$

for small changes in v. Therefore by writing $\Delta \log(v) = \log(v_1) - \log(v_0)$, then

$$100 \cdot \Delta \log(v) \approx \%\Delta v.$$

Definition (Elasticity)
The **elasticity** of y with respect to x is defined as

$$\frac{\Delta y}{\Delta x} \cdot \frac{x}{y} = \frac{\%\Delta y}{\%\Delta x}.$$

The elasticity of y with respect to x is the percentage change in y when x increases by 1%.

Consider the relationship

$$\log(y) = \beta_0 + \beta_1 \log(x),$$

Then

$$\beta_1 = \frac{\Delta \log(y)}{\Delta \log(x)} = \frac{100 \times \Delta \log(y)}{100 \times \Delta \log(x)} \approx \frac{\%\Delta y}{\%\Delta x}.$$

The slope parameter β_1 is the elasticity of y with respect to x [assuming that $x, y > 0$].

Constant Elasticity Demand Function:

If q is quantity demanded and p is price and these variables are related by

$$\log(q) = 4.7 - 1.25 \log(p),$$

then the *price elasticity of demand* is -1.25. Roughly, a 1% increase in price leads to a 1.25% fall in the quantity demanded.

Suppose that $y > 0$ and

$$\log(y) = \beta_0 + \beta_1 x,$$

Then $\Delta \log(y) = \beta_1 \Delta x$, so

$$100 \times \Delta \log(y) = (100 \times \beta_1)\Delta x$$

$$\%\Delta y \approx (100 \times \beta_1)\Delta x$$

$$\frac{\%\Delta y}{\Delta x} \approx (100 \times \beta_1)$$

The slope parameter β_1 is the semi-elasticity of y with respect to x [assuming that $y > 0$].

Logarithmic Wage Equation:

Suppose that hourly wage and years of education are related by

$$log(wage) = 2.78 + .094\ educ.$$

Then,

$$\%\Delta wage \approx 100(.094)\Delta educ = 9.4\Delta educ.$$

It follows that one more year of education increases hourly wage by about 9.4%.

Suppose that $x > 0$ and

$$y = \beta_0 + \beta_1 \log(x),$$

Then $\Delta y = \beta_1 \Delta \log(x)$, so

$\Delta y = (\beta_1/100)\ [100 \times \Delta \log(x)]$

$\Delta y \approx (\beta_1/100)\ \%\Delta x$

💡 $(\beta_1/100)$ is the unit change in y when x increases by 1%. [assuming that $x > 0$].

Labor Supply Equation:

Assume that the labor supply of a worker can be described by

$$hours = 33 + 45.1 log(wage),$$

where $wage$ is hourly wage and $hours$ is hours worked per week. Then,

$$\Delta hours \approx \left(\frac{45.1}{100}\right)(\%\Delta\ wage\ ) = .451\%\Delta\ wage.$$

In other words, a 1 % increase in wage increases the weekly hours worked by about .45, or slightly less than one-half hour.

👉 If the wage increases by 10%, then $\Delta hours = .451(10) = 4.51$, or about four and one-half hours.

Suppose that $1_A(x)$ represents the indicator function that equals one if $x \in A$ and zero otherwise, and consider the linear relationship

$$y = \beta_0 + \beta_1 1_A(x),$$

Then for pairs $(y_2, x \in A)$ and $(y_1, x \notin A)$, we have $\Delta y = \beta_1$.

💡 β_1 is the unit change in y when x belongs to A.

Logarithmic Wage Equation:

Suppose that hourly wage and the indicator for male are related by

$$log(wage) = 2.78 + .094\ male.$$

Then,

$$\%\Delta wage \approx 100(.094) = 9.4.$$

It follows that men earn roughly 9.4% more than women.

Law of Iterated Expectations

Theorem: Simple Law of Iterated Expectations If $\mathbb{E}|Y| < \infty$ then for any random vector X,

$\mathbb{E}\big(\mathbb{E}(\,Y \mid X\,)\big) = \mathbb{E}(Y)$

⚠ The average of the conditional averages is the unconditional average.

Theorem: Law of Iterated Expectations

If $\mathbb{E}|Y| < \infty$ then for any random vectors X_1 and X_2,

$\mathbb{E}[\,\mathbb{E}[\,Y \mid X_1, X_2\,] \mid X_1\,] = \mathbb{E}[\,Y \mid X_1\,]$

⚠ "The smaller information set wins."

Theorem: Conditioning Theorem

If $\mathbb{E}|Y| < \infty$ then

$\mathbb{E}\big[\,g(X)Y \mid X\,\big] = g(X)\mathbb{E}[\,Y \mid X\,]$

If in addition $\mathbb{E}|g(X)| < \infty$ then

$\mathbb{E}\big[g(X)Y\big] = \mathbb{E}\big[g(X)\mathbb{E}[\,Y \mid X\,]\big]$

⚠ A property of conditional expectations is that when you condition on a random vector X you can effectively treat it as if it is constant.

Linear Projection Model

Linear Projection Model

$$Y = X'\beta + e$$
$$\mathbb{E}[Xe] = 0$$
$$\beta = (\mathbb{E}[XX'])^{-1}\mathbb{E}[XY]$$

β is known as the *linear projection coefficient*.

Recall:

Linear CEF Model

$$Y = X'\beta + e$$
$$\mathbb{E}[e \mid X] = 0$$

Homoskedastic Linear CEF Model

$$Y = X'\beta + e$$
$$\mathbb{E}[e \mid X] = 0$$
$$\mathbb{E}[e^2 \mid X] = \sigma^2$$

Invertibility & Identification

⚠Identification means ***uniqueness***!

Invertibility and Identification

The linear projection coefficient $\beta = (\mathbb{E}[XX'])^{-1}\mathbb{E}[XY]$ exists and is unique as long as the $k \times k$ matrix $\boldsymbol{Q}_{XX} = \mathbb{E}[XX']$ is invertible. The matrix $\boldsymbol{Q}_{XX}$ is often called the design matrix as in experimental settings the researcher is able to control $\boldsymbol{Q}_{XX}$ by manipulating the distribution of the regressors X. Observe that for any non-zero $\alpha \in \mathbb{R}^k$

$$\alpha'\boldsymbol{Q}_{XX}\alpha = \mathbb{E}[\alpha'XX'\alpha] = \mathbb{E}\left[(\alpha'X)^2\right] \geq 0$$

so $\boldsymbol{Q}_{XX}$ by construction is positive semi-definite, conventionally written as $\boldsymbol{Q}_{XX} \geq 0$. The assumption that it is positive definite means that this is a strict inequality, $\mathbb{E}\left[(\alpha'X)^2\right] > 0$. This is conventionally written as $\boldsymbol{Q}_{XX} > 0$. This condition means that there is no non-zero vector α such that $\alpha'X = 0$ identically. Positive definite matrices are invertible. Thus when $\boldsymbol{Q}_{XX} > 0$ then $\beta = (\mathbb{E}[XX'])^{-1}\mathbb{E}[XY]$ exists and is uniquely defined. In other words, if we can exclude the possibility that a linear function of X is degenerate, then β is uniquely defined.

We have shown that the linear projection coefficient β is identified (uniquely determined) under Assumption 1. The key is invertibility of $\boldsymbol{Q}_{XX}$. Otherwise, there is no unique solution to the equation

$$\boldsymbol{Q}_{XX}\beta = \boldsymbol{Q}_{XY}$$

when $\boldsymbol{Q}_{XX}$ is not invertible there are multiple solutions. In this case the coefficient β is not identified as it does not have a unique value.

💡 Although $\boldsymbol{Q}_{XX}\beta = \boldsymbol{Q}_{XY}$ has multiple solutions, it turns out that one of these solutions denoted as $\boldsymbol{Q}_{XX}^{-}\boldsymbol{Q}_{XY}$ [$\boldsymbol{A}^{-}$ denoting the Moore-Penrose Inverse] is such that $\left\|\boldsymbol{Q}_{XX}^{-}\boldsymbol{Q}_{XY}\right\|_2 \leq \|b\|_2$ for *all* solutions b, i.e., minimum norm solution to a linear system.

Least Squares

The best linear predictor of Y given X for a pair of random variables $(Y, X) \in \mathbb{R} \times \mathbb{R}^k$, is summarized in the linear projection model. We are now interested in *estimating* the parameters of this model, in particular the projection coefficient

$$\beta = \left(\mathbb{E}(XX')\right)^{-1}\mathbb{E}(XY).$$

We can estimate β from observational data which includes joint measurements on the variables (Y, X).

Assumption: The variables $\{(Y_1, X_1), \dots, (Y_i, X_i), \dots, (Y_n, X_n)\}$ are identically distributed; they are draws from a common distribution F.

The linear projection model applies to the random observations (Y_i, X_i). We can write the model as

$$Y_i = X_i'\beta + e_i.$$

where the *linear projection coefficient* β is defined as

$$\beta = \underset{b \in \mathbb{R}^k}{\operatorname{argmin}}\, S(b),$$

the minimizer of the expected squared error

$$S(\beta) = \mathbb{E}\left(\left(Y_i - X_i'\beta\right)^2\right),$$

and has the explicit solution

$$\beta = \left(\mathbb{E}\left(X_i X_i'\right)\right)^{-1}\mathbb{E}(X_i Y_i).$$

Moment of functions of random variables: $\mathbb{E}[h(X, Y; \beta)]$.
$h(X, Y; \beta) = \left(Y - X'\beta\right)^2$,
$h(X, Y; \beta) = XX'$,
$h(X, Y; \beta) = XY$.

Recall that for any random variable, W, its *Cumulative Distribution Function*, $F(t)$, is well-defined as $F(t) = \Pr\{W \le t\}$. Then, if $\mathbb{E}|h(w)| < \infty$, one has

$$\mathbb{E}[h(W)] = \int h(W)dF(W).$$

Based on a sample $\{W_i; i = 1, \dots, n\}$, we can estimate this unknown quantity by replacing the unknown CDF with the **empirical distribution function**, $\widehat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(W_i \le t)$, then a natural 'plug in' estimator of $\mathbb{E}[h(W)]$ is

$$\widehat{\mathbb{E}}[h(W)] = \int h(W)d\widehat{F}(W).$$

It turns out that through the application of various measurement theory concepts one has

$$\widehat{\mathbb{E}}[h(W)] = \frac{1}{n}\sum_{i=1}^{n} h(W_i).$$

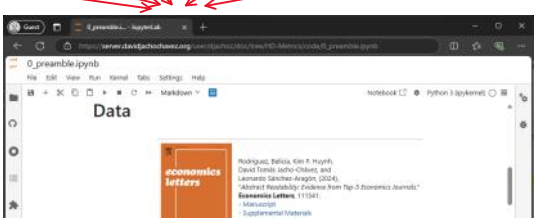
American Economic Review

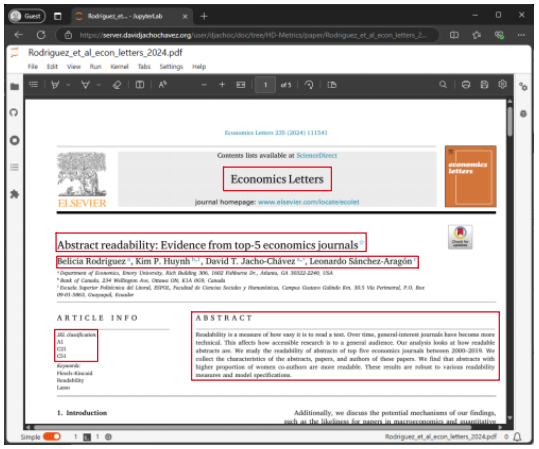
Econometrica

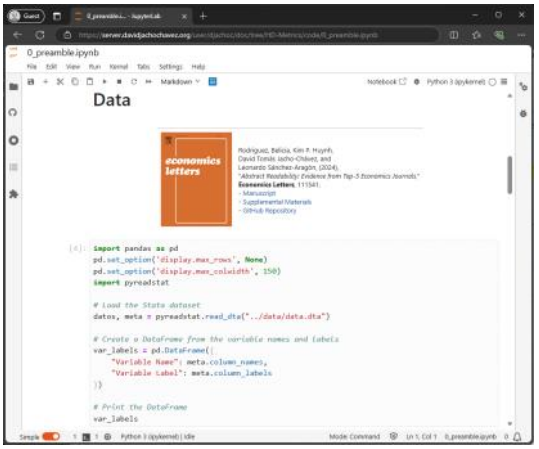
Journal of Political Economy

The Quarterly Journal of Economics

Review of Economic Studies

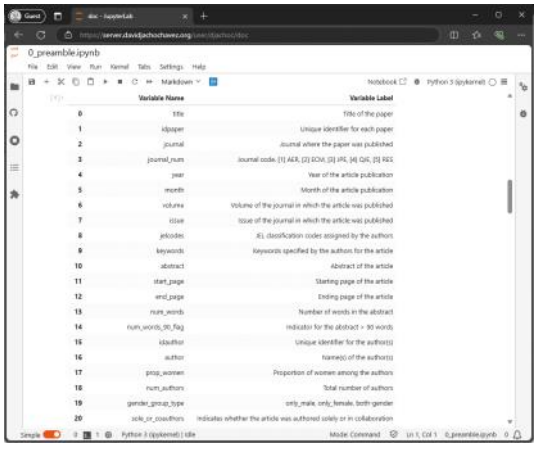


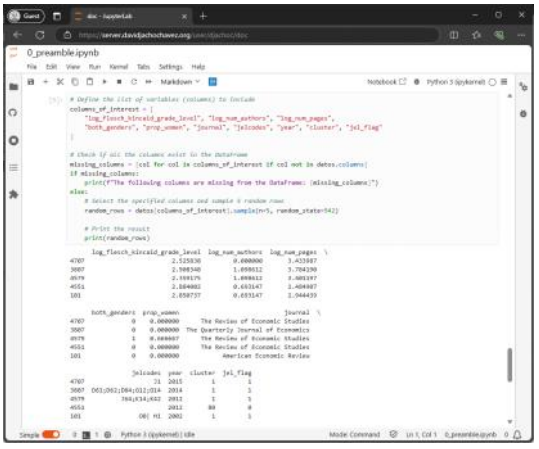




"The 'JEL' classification system originated with the *Journal of Economic Literature* and is a standard method of classifying scholarly literature in the field of economics. It is used in many of the AEA's published research materials."

Source: [American Economic Association: JEL Guide]





X = $[X_1', X_2', D', G', d]'$

X_{4707}= [0,3.433987,0,0,i.ECM,i.J1,i.2015]

G_{4707}= i.1

d_{4707}= 1

Y	=	`log_flesch_kinacid_grade_level`
X_1	=	`prop_women`
X_2	=	[`log_num_authors`, `log_num_page`, `both_genders`, 1]'
D	=	[i.`journal`, i.`jelcodes`, i.`year`]'
G	=	i.`cluster`
d	=	`jel_flag`

i.`journal`: corresponds to 4 indicators for ECM, JPE, QJE, RES

i.`jelcodes`: corresponds to 19 indicators for all JEL codes except 'D'

i.`year`: corresponds to 19 indicators for years 2001, 2002,…, 2019

i.`cluster`: corresponds to 214 indicators for clusters 2, 3,…, 215, except cluster '1'

`jel_flag`: indicator of whether the article includes a JEL classification

Base category: AER, 'D' JEL (Microeconomics), year 2000, for which the JEL was imputed

Least Squares Estimator

For given b, the expected squared error is $\mathbb{E}[(Y_i - X_i'b)^2]$. The *moment estimator* of $S(b)$ is the sample average

$$\begin{aligned}\hat{S}(b) &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'b)^2 \\ &= \frac{1}{n}\mathrm{SSE}(b)\end{aligned},$$

Sum-of-Squared-Errors function.

⚠ The estimator is also commonly referred to as the **ordinary least-squares** (**OLS**) estimator. The following notation is also used:

- $\hat{\beta}_{\text{ols}}$
- $\hat{\beta}_n$

The "^" (caret) symbol over the parameter β indicates that $\hat{\beta}$ is a sample estimate of β. Obviously

- $\hat{\beta}$ is a random variable (since it depends on the sample n).
- β is not random but the finite-dimensional vector of parameter of interest.
- $\hat{\beta} \neq \beta$.

Definition: The **least squares estimator** is

$$\widehat{\beta} = \underset{b\in\mathbb{R}^k}{\operatorname{argmin}}\, \widehat{S}(b),$$

where $\hat{S}(b) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'b)^2$.

Solving for Least Squares

The sum of squared errors can be written as

$$\mathrm{SSE}(b) = \sum_{i=1}^{n} Y_i^2 - 2b'\sum_{i=1}^{n} X_iY_i + b'\sum_{i=1}^{n} X_iX_i'\,b$$

This is a **Quadratic Function** in $\boldsymbol{b}$.

The FOC are

$$0 = \frac{\partial}{\partial b}\mathrm{SSE}(b)\Big|_{b=\widehat{\beta}} = -2\sum_{i=1}^{n} X_iY_i + 2\sum_{i=1}^{n} X_iX_i'\,\hat{\beta}$$

🔔 Notice that $\frac{\partial^2}{\partial b\,\partial b'}\mathrm{SSE}(b)\Big|_{b=\widehat{\beta}} = 2\sum_{i=1}^{n} X_iX_i' \geq 0.$

The solution for $\hat{\beta}$ may be found by solving the system of k equations

$$\sum_{i=1}^{n} X_iX_i'\,\hat{\beta} = \sum_{i=1}^{n} X_iY_i.$$

⚠ If you ever need to *numerically* calculate the solution to a system of equations, you **should** use this equation. In Python you should use the numpy.linalg.solve().

An explicit formula for the least-squares estimator is then

$$\hat{\beta} = \left(\sum_{i=1}^{n} X_iX_i'\right)^{-1}\left(\sum_{i=1}^{n} X_iY_i\right).$$

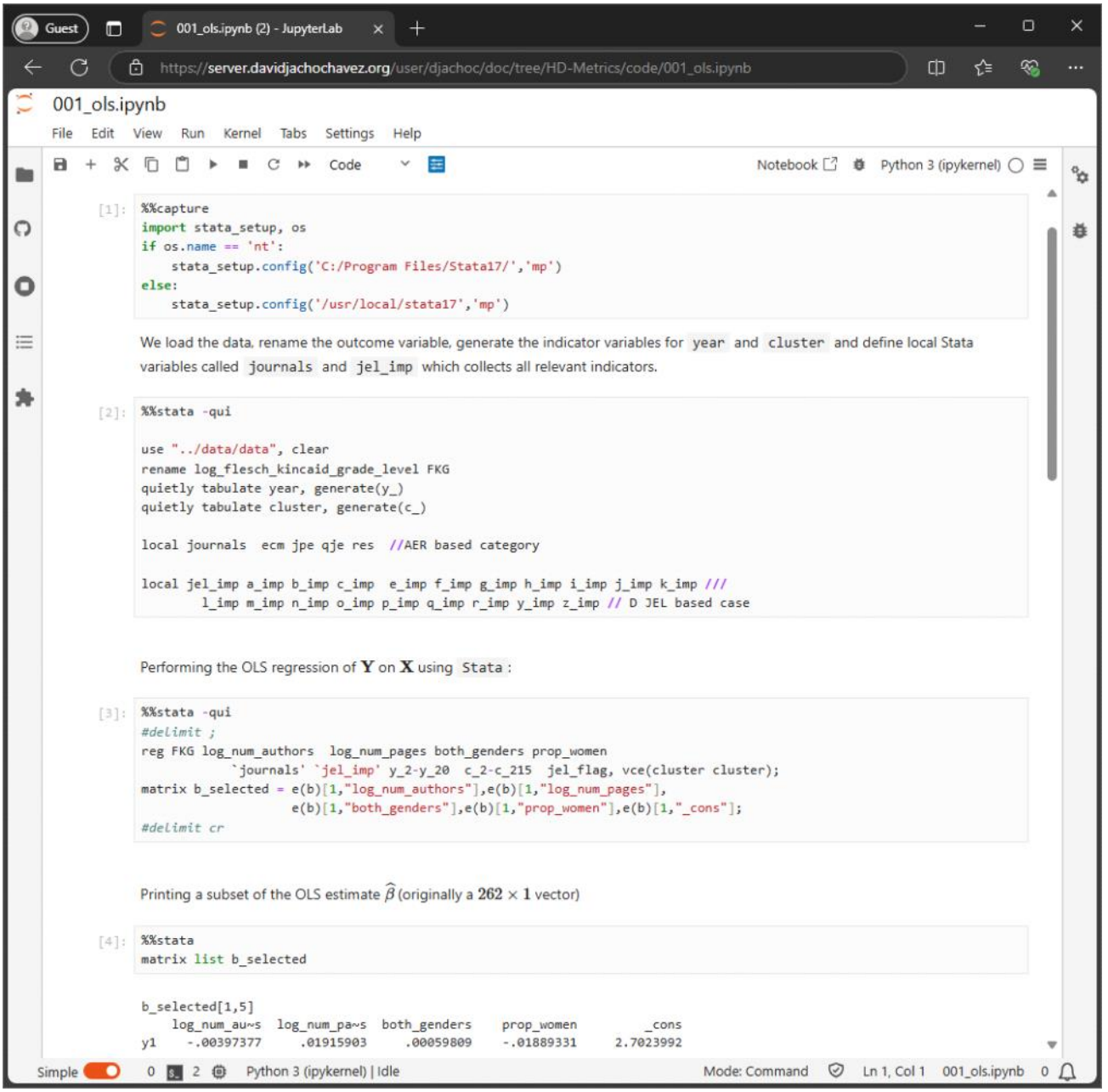
Least Squares Estimator as a *Moment* Estimator

Population		Sample

$\beta = \left(\mathbb{E}\left(X_i X_i'\right)\right)^{-1} \mathbb{E}\left(X_i Y_i\right)$		$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right)$ $= \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} X_i Y_i\right)$

Theorem: If $\sum_{i=1}^{n} X_i X_i' > 0$, the least squares estimator equals

$$\hat{\beta} = \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} X_i Y_i\right).$$



Least Squares Residuals

Fitted Value	+	Residual	=	Outcome
$\hat{Y}_i = X_i'\hat{\beta}$		$\hat{e}_i = Y_i - \hat{Y}_i = Y_i - X_i'\hat{\beta}$		$Y_i = \hat{Y}_i + \hat{e}_i = X_i'\hat{\beta} + \hat{e}_i$

$$\begin{aligned}\sum_{i=1}^{n} X_i\hat{e}_i &= \sum_{i=1}^{n} X_i\left(Y_i - X_i'\hat{\beta}\right)\\ &= \sum_{i=1}^{n} X_iY_i - \sum_{i=1}^{n} X_iX_i'\hat{\beta}\\ &= \sum_{i=1}^{n} X_iY_i - \sum_{i=1}^{n} X_iX_i'\left(\sum_{i=1}^{n} X_iX_i'\right)^{-1}\left(\sum_{i=1}^{n} X_iY_i\right)\\ &= \sum_{i=1}^{n} X_iY_i - \sum_{i=1}^{n} X_iY_i\\ &= \mathbf{0}\end{aligned}$$

⚠ When X contains a constant, then $\frac{1}{n}\sum_{i=1}^{n}\hat{e}_i = 0$.

Model in Matrix Notation

We can stack these n equations together as

$$Y_1 = X_1'\beta + e_1$$
$$Y_2 = X_2'\beta + e_2$$
$$\vdots$$
$$Y_n = X_n'\beta + e_n$$

Now define

$$\boldsymbol{Y} = \underset{n\times 1}{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}, \boldsymbol{X} = \underset{n\times k}{\begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix}}, \boldsymbol{e} = \underset{n\times 1}{\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}}$$

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{e}$$

$$\sum_{i=1}^{n} X_i X_i' = \boldsymbol{X}'\boldsymbol{X}$$
$$\sum_{i=1}^{n} X_i Y_i = \boldsymbol{X}'\boldsymbol{Y}$$

Least Squares Estimator: $\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y})$

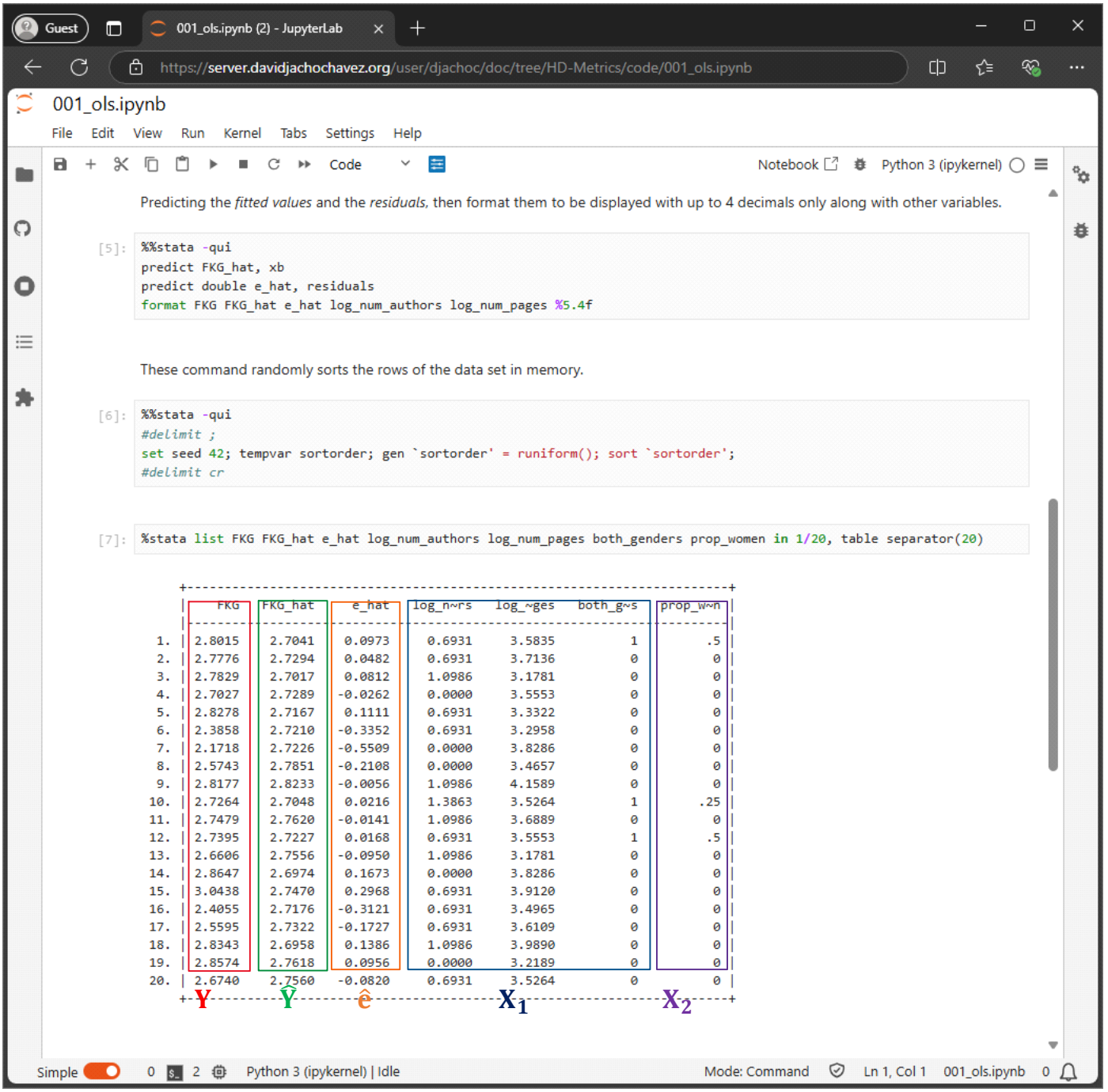
Theorem: If $\sum_{i=1}^{n} X_i X_i' > 0$, the least squares estimator equals

$$\hat{\beta} = \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\sum_{i=1}^{n} X_i Y_i\right)$$

Fitted Values: $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\beta}$

Residuals: $\hat{\boldsymbol{e}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\beta}$

Sum-of-Squared Errors: $SSE(b) = (\boldsymbol{Y} - \boldsymbol{X}b)'(\boldsymbol{Y} - \boldsymbol{X}b)$


Guest
001_ols.ipynb (2) - JupyterLab
https://server.davidjachochavez.org/user/djachoc/doc/tree/HD-Metrics/code/001_ols.ipynb
001_ols.ipynb
File Edit View Run Kernel Tabs Settings Help
Code
Notebook
Python 3 (ipykernel)
Predicting the fitted values and the residuals, then format them to be displayed with up to 4 decimals only along with other variables.
[5]:
%%stata -qui
predict FKG_hat, xb
predict double e_hat, residuals
format FKG FKG_hat e_hat log_num_authors log_num_pages %5.4f
These command randomly sorts the rows of the data set in memory.
[6]:
%%stata -qui
#delimit ;
set seed 42; tempvar sortorder; gen `sortorder' = runiform(); sort `sortorder';
#delimit cr
[7]:
%stata list FKG FKG_hat e_hat log_num_authors log_num_pages both_genders prop_women in 1/20, table separator(20)
FKG FKG_hat e_hat log_n~rs log_~ges both_g~s prop_w~n
1. 2.8015 2.7041 0.0973 0.6931 3.5835 1 .5
2. 2.7776 2.7294 0.0482 0.6931 3.7136 0 0
3. 2.7829 2.7017 0.0812 1.0986 3.1781 0 0
4. 2.7027 2.7289 -0.0262 0.0000 3.5553 0 0
5. 2.8278 2.7167 0.1111 0.6931 3.3322 0 0
6. 2.3858 2.7210 -0.3352 0.6931 3.2958 0 0
7. 2.1718 2.7226 -0.5509 0.0000 3.8286 0 0
8. 2.5743 2.7851 -0.2108 0.0000 3.4657 0 0
9. 2.8177 2.8233 -0.0056 1.0986 4.1589 0 0
10. 2.7264 2.7048 0.0216 1.3863 3.5264 1 .25
11. 2.7479 2.7620 -0.0141 1.0986 3.6889 0 0
12. 2.7395 2.7227 0.0168 0.6931 3.5553 1 .5
13. 2.6606 2.7556 -0.0950 1.0986 3.1781 0 0
14. 2.8647 2.6974 0.1673 0.0000 3.8286 0 0
15. 3.0438 2.7470 0.2968 0.6931 3.9120 0 0
16. 2.4055 2.7176 -0.3121 0.6931 3.4965 0 0
17. 2.5595 2.7322 -0.1727 0.6931 3.6109 0 0
18. 2.8343 2.6958 0.1386 1.0986 3.9890 0 0
19. 2.8574 2.7618 0.0956 0.0000 3.2189 0 0
20. 2.6740 2.7560 -0.0820 0.6931 3.5264 0 0
Y
Ŷ
ê
X1
X2
Simple
0 2
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 1
001_ols.ipynb
0

Projection Matrix

Define the matrix

$$\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'.$$

Observe that

$$\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{X}.$$

This is a property of a **projection matrix**.

Cases:

- The matrix $\boldsymbol{P}$ creates the fitted values in a least-squares regression:

$$\boldsymbol{P}\boldsymbol{Y} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{X}\hat{\beta} = \hat{\boldsymbol{Y}}.$$

Because of this, the $\boldsymbol{P}$ is sometimes called the "hat matrix"

- A special example of a projection matrix occurs when $\boldsymbol{X} = \mathbf{1}_n$ is a n-vector of ones. Then

$$\boldsymbol{P} = \mathbf{1}_n\left(\mathbf{1}_n'\mathbf{1}_n\right)^{-1}\mathbf{1}_n' = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'.$$

Note that in this case

$$\begin{aligned}\boldsymbol{P}\boldsymbol{Y} &= \mathbf{1}_n\left(\mathbf{1}_n'\mathbf{1}_n\right)^{-1}\mathbf{1}_n'\boldsymbol{Y}\\ &= \mathbf{1}_n\bar{Y}\end{aligned}.$$

An n-vector whose elements are the sample mean $\bar{Y}$ of Y_i.

Theorem: The projection matrix $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ for any $n \times k$ matrix $\boldsymbol{X}$ with $n \geq k$ has the following algebraic properties

1. $\boldsymbol{P}$ is symmetric ($\boldsymbol{P}' = \boldsymbol{P}$).
2. $\boldsymbol{P}$ is idempotent ($\boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}$).
3. $\text{tr}\ \boldsymbol{P} = k$.
4. The eigenvalues of $\boldsymbol{P}$ are 1 and 0. There are k eigenvalues equalling 1 and $n - k$ equalling 0.
5. $\text{rank}(\boldsymbol{P}) = k$.

$$\begin{aligned}\boldsymbol{P}' &= \left(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)'\\ &= (\boldsymbol{X}')'\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\right)'(\boldsymbol{X})'\\ &= \boldsymbol{X}\left((\boldsymbol{X}'\boldsymbol{X})'\right)^{-1}\boldsymbol{X}'\\ &= \boldsymbol{X}\left((\boldsymbol{X})'(\boldsymbol{X}')'\right)^{-1}\boldsymbol{X}'\\ &= \boldsymbol{P}\end{aligned}$$

$$\begin{aligned}\boldsymbol{P}\boldsymbol{P} &= \boldsymbol{P}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\\ &= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\\ &= \boldsymbol{P}\end{aligned}$$

$$\begin{aligned}\text{tr}\ \boldsymbol{P} &= \text{tr}\left(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)\\ &= \text{tr}\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\right)\\ &= \text{tr}\left(\boldsymbol{I}_k\right)\\ &= k\end{aligned}$$

Orthogonal Projection

Define

$$\begin{aligned}\boldsymbol{M} &= \boldsymbol{I}_n - \boldsymbol{P} \\ &= \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'.\end{aligned}$$

Note

$$\boldsymbol{M}\boldsymbol{X} = \big(\boldsymbol{I}_n - \boldsymbol{P}\big)\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{P}\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X} = 0.$$

Thus $\boldsymbol{M}$ and $\boldsymbol{X}$ are orthogonal. We call $\boldsymbol{M}$ an **orthogonal projection matrix**.

Cases:

- $\boldsymbol{M}$ creates least-squares residuals:

$$\boldsymbol{M}\boldsymbol{Y} = \boldsymbol{Y} - \boldsymbol{P}\boldsymbol{Y} = \boldsymbol{Y} - \boldsymbol{X}\hat{\beta} = \hat{e}.$$

- A special example of an orthogonal projection matrix occurs when $\boldsymbol{X} = \mathbf{1}_n$ is a n-vector of ones. Then

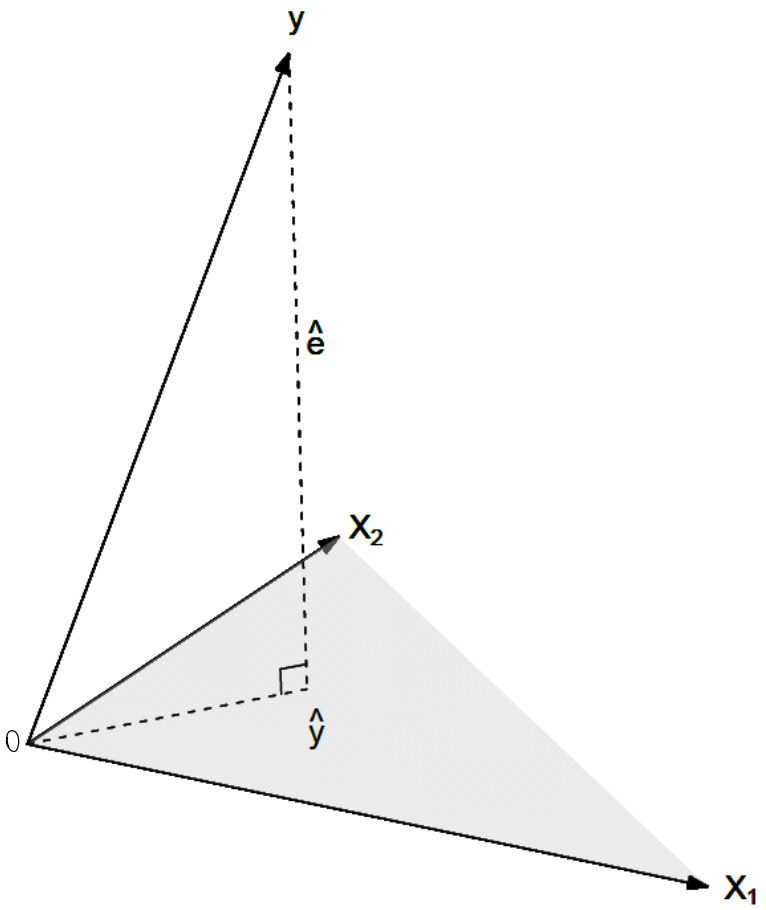
$$\boldsymbol{M} = \boldsymbol{I}_n - \mathbf{1}_n\big(\mathbf{1}_n'\mathbf{1}_n\big)^{-1}\mathbf{1}_n'.$$

Note that in this case

$$\boldsymbol{M}\boldsymbol{Y} = \boldsymbol{Y} - \mathbf{1}_n\bar{Y}.$$

$\boldsymbol{M}$ creates demeaned values.

The orthogonal projection matrix $\boldsymbol{M}$ has similar properties with $\boldsymbol{P}$, including that $\boldsymbol{M}$ is symmetric ($\boldsymbol{M}' = \boldsymbol{M}$) and idempotent ($\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$). We can calculate $\operatorname{tr}\boldsymbol{M} = n - k$, and therefore the rank of $\boldsymbol{M}$ is $n - k$.


y
ê
X2
ŷ
0
X1

Estimation of Error Variance

Substituting $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{e}$ into $\hat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{Y}$ and using $\boldsymbol{M}\boldsymbol{X} = 0$ we find

$$\hat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{Y} = \boldsymbol{M}(\boldsymbol{X}\beta + \boldsymbol{e}) = \boldsymbol{M}\boldsymbol{e},$$

which is free of dependence on the regression coefficient β.

The error variance $\sigma^2 = \mathbb{E}(e_i^2)$ is a moment, so a natural estimator is a *moment estimator*.

moment estimator	Infeasible		Feasible
Using summations	$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$		$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i^2$
Using matrix notation	$\tilde{\sigma}^2 = n^{-1}\boldsymbol{e}'\boldsymbol{e}$		$\hat{\sigma}^2 = n^{-1}\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$

Relationship between Infeasible vs Feasible

$$\begin{aligned}\hat{\sigma}^2 &= n^{-1}\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} \\ &= n^{-1}\boldsymbol{Y}'\boldsymbol{M}\boldsymbol{M}\boldsymbol{Y} \\ &= n^{-1}\boldsymbol{Y}'\boldsymbol{M}\boldsymbol{Y} \\ &= n^{-1}\boldsymbol{e}'\boldsymbol{M}\boldsymbol{e}\end{aligned}$$

An interesting implication is that

$$\begin{aligned}\tilde{\sigma}^2 - \hat{\sigma}^2 &= n^{-1}\boldsymbol{e}'\boldsymbol{e} - n^{-1}\boldsymbol{e}'\boldsymbol{M}\boldsymbol{e} \\ &= n^{-1}\boldsymbol{e}'\boldsymbol{P}\boldsymbol{e} \\ &\geq 0\end{aligned}$$

⚠ The final inequality holds because $\boldsymbol{P}$ is positive semi-definite and $\boldsymbol{e}'\boldsymbol{P}\boldsymbol{e}$ is a quadratic form.

Analysis of Variance

$\boldsymbol{Y} = \boldsymbol{PY} + \boldsymbol{MY} = \hat{\boldsymbol{Y}} + \hat{\boldsymbol{e}}$

This decomposition is **orthogonal**, that is

$\boldsymbol{TSS} = \boldsymbol{ESS} + \boldsymbol{RSS}$

$\hat{\boldsymbol{Y}}'\hat{\boldsymbol{e}} = (\boldsymbol{PY})'(\boldsymbol{MY}) = \boldsymbol{Y}'\boldsymbol{PMY} = 0$

It follows that

$$\boldsymbol{Y}'\boldsymbol{Y} = \hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}} + 2\hat{\boldsymbol{Y}}'\hat{\boldsymbol{e}} + \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} = \hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}} + \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$$

$$\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} \hat{Y}_i^2 + \sum_{i=1}^{n} \hat{e}_i^2$$

$$\boldsymbol{Y} - \boldsymbol{1}_n\bar{Y} = \hat{\boldsymbol{Y}} - \boldsymbol{1}_n\bar{Y} + \hat{\boldsymbol{e}}$$

$$\left(\hat{\boldsymbol{Y}} - \boldsymbol{1}_n\bar{Y}\right)'\hat{\boldsymbol{e}} = \hat{\boldsymbol{Y}}'\hat{\boldsymbol{e}} - \bar{Y}\boldsymbol{1}_n'\hat{\boldsymbol{e}} = 0$$

$$\left(\boldsymbol{Y} - \boldsymbol{1}_n\bar{Y}\right)'\left(\boldsymbol{Y} - \boldsymbol{1}_n\bar{Y}\right) = \left(\hat{\boldsymbol{Y}} - \boldsymbol{1}_n\bar{Y}\right)'\left(\hat{\boldsymbol{Y}} - \boldsymbol{1}_n\bar{Y}\right) + \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$$

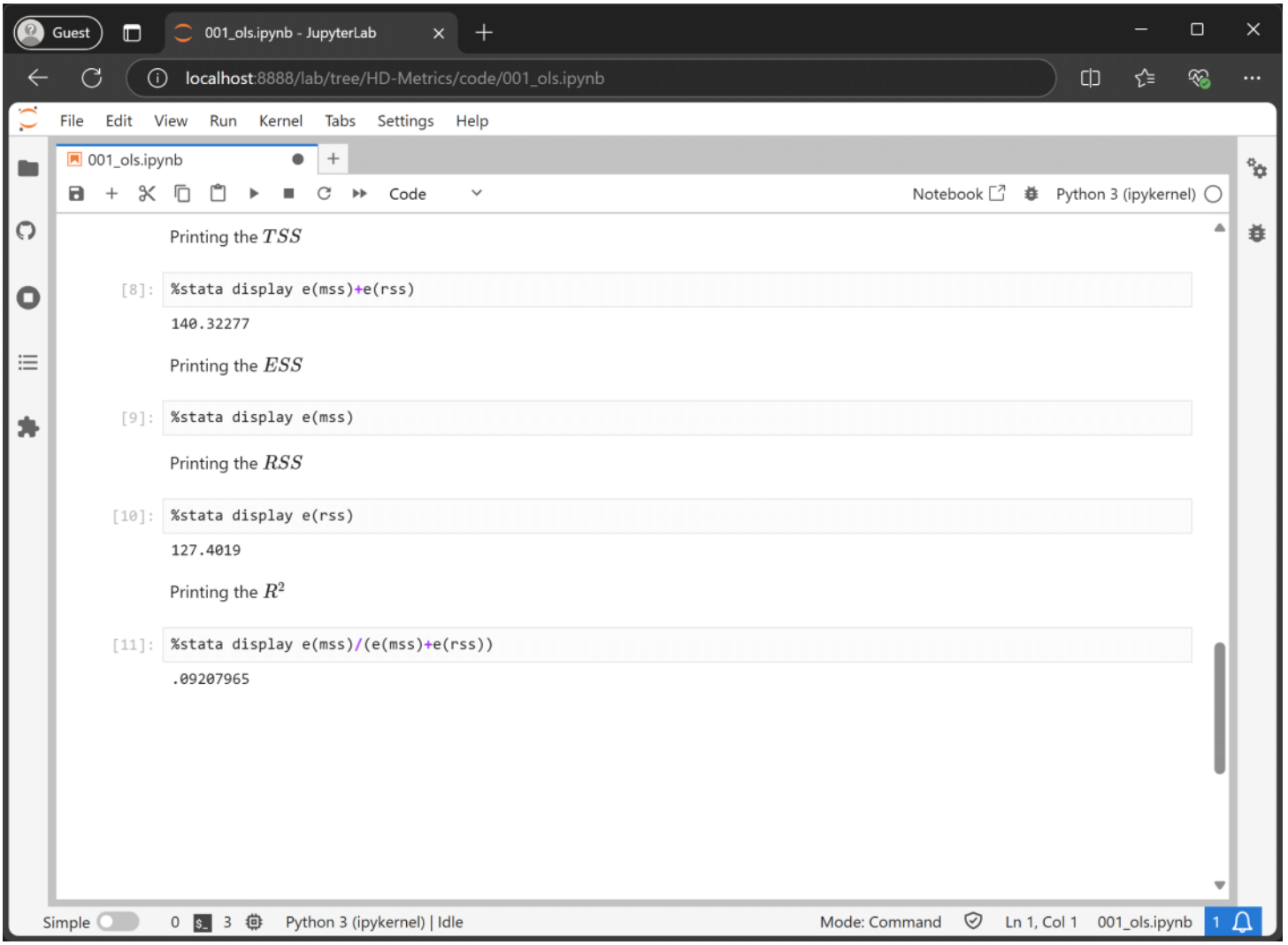
$$\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n} \hat{e}_i^2$$

Analysis-of-variance formula

A commonly reported statistic is the **coefficient of determination** or **R-squared**

$$R^2 = \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2} = 1 - \frac{\sum_{i=1}^{n} \hat{e}_i^2}{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}$$

It is often described as the fraction of the sample variance of Y_i which is explained by the least-squares fit.


Guest
001_ols.ipynb - JupyterLab
localhost:8888/lab/tree/HD-Metrics/code/001_ols.ipynb
File Edit View Run Kernel Tabs Settings Help
001_ols.ipynb
Code
Notebook
Python 3 (ipykernel)
Printing the TSS
[8]: %stata display e(mss)+e(rss)
140.32277
Printing the ESS
[9]: %stata display e(mss)
Printing the RSS
[10]: %stata display e(rss)
127.4019
Printing the R^2
[11]: %stata display e(mss)/(e(mss)+e(rss))
.09207965
Simple
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 1
001_ols.ipynb

Regression Components

Partition

$$\boldsymbol{X} = [\boldsymbol{X}_1 \boldsymbol{X}_2],$$

And

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

Then the regression model can be rewritten as

$$\boldsymbol{Y} = \boldsymbol{X}_1 \beta_1 + \boldsymbol{X}_2 \beta_2 + \boldsymbol{e}.$$

The OLS estimator of $\beta = (\beta_1', \beta_2')'$ is obtained by regression of $\boldsymbol{Y}$ on $\boldsymbol{X} = [\boldsymbol{X}_1 \boldsymbol{X}_2]$ and can be written as

$$\boldsymbol{Y} = \boldsymbol{X}\hat{\beta} + \hat{\boldsymbol{e}} = \boldsymbol{X}_1 \hat{\beta}_1 + \boldsymbol{X}_2 \hat{\beta}_2 + \hat{\boldsymbol{e}}.$$

We are interested in algebraic expressions for $\hat{\beta}_1$ and $\hat{\beta}_2$. The least-squares estimator by definition is found by the joint minimization

$$(\hat{\beta}_1, \hat{\beta}_2) = \underset{b_1, b_2}{\operatorname{argmin}} \operatorname{SSE}(b_1, b_2),$$

where

$$\operatorname{SSE}(b_1, b_2) = (\boldsymbol{Y} - \boldsymbol{X}_1 b_1 - \boldsymbol{X}_2 b_2)'(\boldsymbol{Y} - \boldsymbol{X}_1 b_1 - \boldsymbol{X}_2 b_2).$$

An equivalent expression for $\hat{\beta}_1$ can be obtained by *concentration*.

The inner expression $\min_{b_2} \operatorname{SSE}(b_1, b_2)$ minimizes over b_2 while holding b_1 fixed.

$$\hat{\beta}_1 = \underset{b_1}{\operatorname{argmin}} \left(\min_{b_2} \operatorname{SSE}(b_1, b_2) \right)$$

This is simply the least squares regression of $\boldsymbol{Y} - \boldsymbol{X}_1 b_1$ on $\boldsymbol{X}_2$. This has solution

$$\underset{b_2}{\operatorname{argmin}} \operatorname{SSE}(b_1, b_2) = (\boldsymbol{X}_2' \boldsymbol{X}_2)^{-1} \left(\boldsymbol{X}_2' (\boldsymbol{Y} - \boldsymbol{X}_1 b_1) \right)$$

Residuals:

$$\boldsymbol{Y} - \boldsymbol{X}_1 b_1 - \boldsymbol{X}_2 (\boldsymbol{X}_2' \boldsymbol{X}_2)^{-1} \left(\boldsymbol{X}_2' (\boldsymbol{Y} - \boldsymbol{X}_1 b_1) \right) = (\boldsymbol{M}_2 \boldsymbol{Y} - \boldsymbol{M}_2 \boldsymbol{X}_1 b_1)$$
$$= \boldsymbol{M}_2 (\boldsymbol{Y} - \boldsymbol{X}_1 b_1)$$

The orthogonal projection matrix for $\boldsymbol{X}_2$:

$\boldsymbol{M}_2 = \boldsymbol{I}_n - \boldsymbol{X}_2 (\boldsymbol{X}_2' \boldsymbol{X}_2)^{-1} \boldsymbol{X}_2'$ (Idempotent)

$$\min_{b_2} \operatorname{SSE}(b_1, b_2) = (\boldsymbol{Y} - \boldsymbol{X}_1 b_1)' \boldsymbol{M}_2 \boldsymbol{M}_2 (\boldsymbol{Y} - \boldsymbol{X}_1 b_1)$$
$$= (\boldsymbol{Y} - \boldsymbol{X}_1 b_1)' \boldsymbol{M}_2 (\boldsymbol{Y} - \boldsymbol{X}_1 b_1)$$

$$\hat{\beta}_1 = \underset{b_1}{\operatorname{argmin}} (\boldsymbol{Y} - \boldsymbol{X}_1 b_1)' \boldsymbol{M}_2 (\boldsymbol{Y} - \boldsymbol{X}_1 b_1)$$
$$= \left(\boldsymbol{X}_1' \boldsymbol{M}_2 \boldsymbol{X}_1\right)^{-1} \left(\boldsymbol{X}_1' \boldsymbol{M}_2 \boldsymbol{Y}\right)$$

The orthogonal projection matrix for $\boldsymbol{X}_1$:

$\boldsymbol{M}_1 = \boldsymbol{I}_n - \boldsymbol{X}_1 \left(\boldsymbol{X}_1' \boldsymbol{X}_1\right)^{-1} \boldsymbol{X}_1'$ (Idempotent)

By similar arguments we can find

$$\hat{\beta}_2 = \left(\boldsymbol{X}_2' \boldsymbol{M}_1 \boldsymbol{X}_2\right)^{-1} \left(\boldsymbol{X}_2' \boldsymbol{M}_1 \boldsymbol{Y}\right)$$

Theorem: The least squares estimator $(\hat{\beta}_1, \hat{\beta}_2)$ for $\boldsymbol{Y} = \boldsymbol{X}\hat{\beta} + \hat{\boldsymbol{e}} = \boldsymbol{X}_1\hat{\beta}_1 + \boldsymbol{X}_2\hat{\beta}_2 + \hat{\boldsymbol{e}}$ has the algebraic solution

$$\hat{\beta}_1 = \left(\boldsymbol{X}_1' \boldsymbol{M}_2 \boldsymbol{X}_1\right)^{-1} \left(\boldsymbol{X}_1' \boldsymbol{M}_2 \boldsymbol{Y}\right)$$

$$\hat{\beta}_2 = \left(\boldsymbol{X}_2' \boldsymbol{M}_1 \boldsymbol{X}_2\right)^{-1} \left(\boldsymbol{X}_2' \boldsymbol{M}_1 \boldsymbol{Y}\right)$$

where $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are defined above.

$$\hat{\beta}_2 = \left(\boldsymbol{X}_2' \boldsymbol{M}_1 \boldsymbol{X}_2\right)^{-1} \left(\boldsymbol{X}_2' \boldsymbol{M}_1 \boldsymbol{Y}\right)$$
$$= \left(\boldsymbol{X}_2' \boldsymbol{M}_1 \boldsymbol{M}_1 \boldsymbol{X}_2\right)^{-1} \left(\boldsymbol{X}_2' \boldsymbol{M}_1 \boldsymbol{M}_1 \boldsymbol{Y}\right)$$
$$= \left(\tilde{\boldsymbol{X}}_2' \tilde{\boldsymbol{X}}_2\right)^{-1} \left(\tilde{\boldsymbol{X}}_2' \tilde{\boldsymbol{e}}_1\right)$$
$$\tilde{\boldsymbol{X}}_2 = \boldsymbol{M}_1 \boldsymbol{X}_2$$
$$\tilde{\boldsymbol{e}}_1 = \boldsymbol{M}_1 \boldsymbol{Y}$$

⚠ Thus the coefficient estimate $\hat{\beta}_2$ is algebraically equal to the least-squares regression of $\tilde{\boldsymbol{e}}_1$ on $\tilde{\boldsymbol{X}}_2$.

In panel data, the FWL theorem is used to greatly improve computational time.

Theorem: Frisch-Waugh-Lovell (FWL)

In the model $\boldsymbol{Y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \boldsymbol{e}$, the OLS estimator of β_2 and the OLS residuals $\hat{\boldsymbol{e}}$ may be computed by either the OLS regression $\boldsymbol{Y} = \boldsymbol{X}\hat{\beta} + \hat{\boldsymbol{e}} = \boldsymbol{X}_1\hat{\beta}_1 + \boldsymbol{X}_2\hat{\beta}_2 + \hat{\boldsymbol{e}}$ or via the following algorithm:

1. Regress $\boldsymbol{Y}$ on $\boldsymbol{X}_1$, obtain residuals $\tilde{\boldsymbol{e}}_1$;
2. Regress $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$, obtain residuals $\tilde{\boldsymbol{X}}_2$;
3. Regress $\tilde{\boldsymbol{e}}_1$ on $\tilde{\boldsymbol{X}}_2$, obtain OLS estimates $\hat{\beta}_2$ and residuals $\hat{\boldsymbol{e}}$.

Leverage Values & LOO Regression

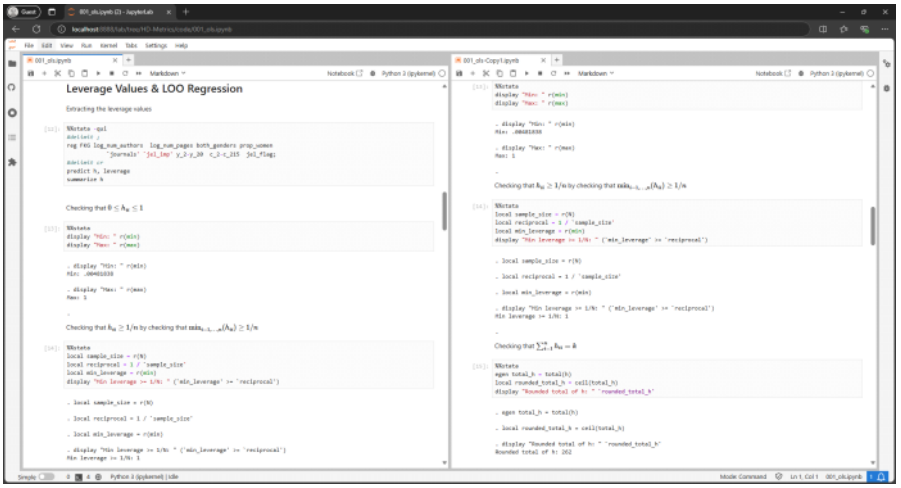
The leverage values for the regressor matrix $\boldsymbol{X}$ are the diagonal elements of the projection matrix $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. There are n leverage values, and are typically written as h_{ii} for $i = 1, \dots, n$, i.e.,

$$h_{ii} = X_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}X_i \qquad \boldsymbol{P} = \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix}(\boldsymbol{X}'\boldsymbol{X})^{-1}(X_1 X_2 \cdots X_n)$$

⚠ The leverage value h_{ii} is a normalized length of the observed regressor vector x_i.

Theorem:

1. $0 \leq h_{ii} \leq 1$.
2. $h_{ii} \geq 1/n$ if X includes an intercept.
3. $\sum_{i=1}^{n} h_{ii} = k$.



Leave-One-Out Regression

Specifically, the leave-one-out least-squares estimator of the regression coefficient β is the least squares estimator constructed using the full sample excluding a single observation i . This can be written

$$\begin{aligned}\hat{\beta}_{(-i)} &= \left(\sum_{j \neq i} X_j X_j'\right)^{-1}\left(\sum_{j \neq i} X_j Y_j\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X} - X_i X_i'\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Y} - X_i Y_i\right) \\ &= \left(\boldsymbol{X}_{(-i)}'\boldsymbol{X}_{(-i)}\right)^{-1}\boldsymbol{X}_{(-i)}'\boldsymbol{Y}_{(-i)}\end{aligned}$$

- $\boldsymbol{X}_{(-i)}$ and $\boldsymbol{Y}_{(-i)}$ are the data matrices omitting the ith row.
- There is a leave-one-out (LOO) estimator for each observation, $i = 1, \dots, n$, so we have n such estimators.

LOO predicted value:		$\tilde{Y}_i = X_i'\hat{\beta}_{(-i)}$
LOO residual, prediction error, or prediction residual:		$\tilde{e}_i = Y_i - \tilde{Y}_i$

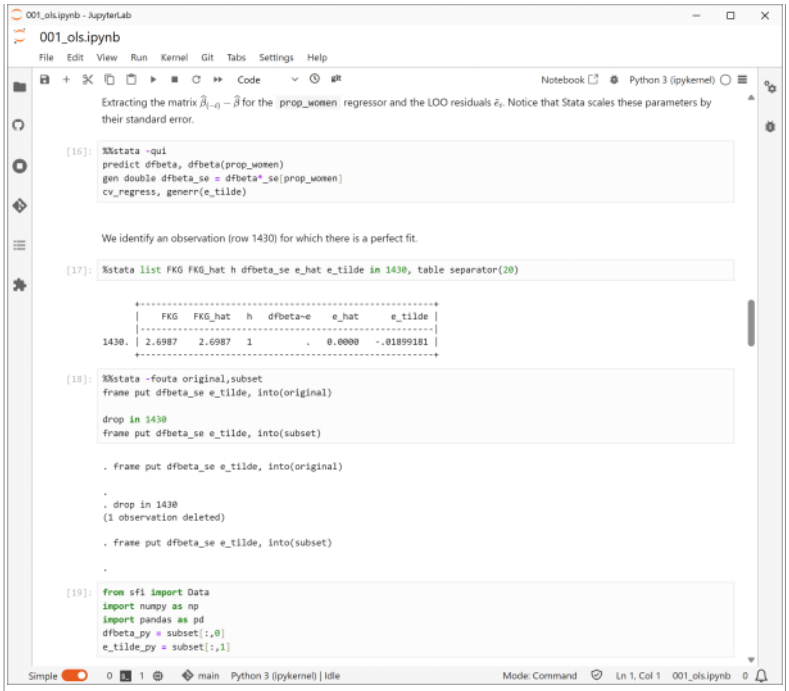
Theorem: The leave-one-out least-squares estimator and prediction error can be calculated as

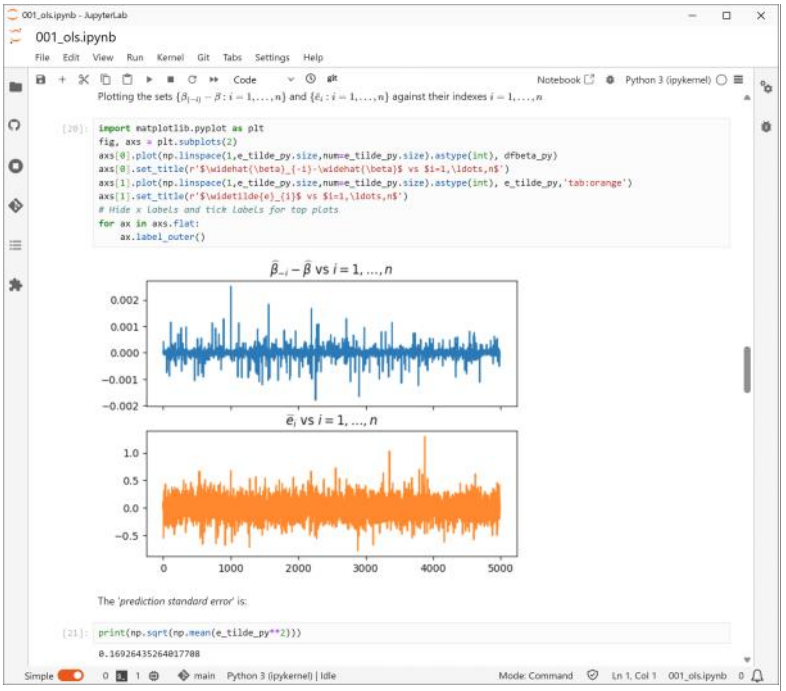
$$\hat{\beta}_{(-i)} = \hat{\beta} - (\boldsymbol{X}'\boldsymbol{X})^{-1}X_i\tilde{e}_i.$$

and

$$\tilde{e}_i = \left(1 - h_{ii}\right)^{-1}\hat{e}_i.$$

where h_{ii} are the leverage values.





Define

$$\boldsymbol{M}^* = \left(\boldsymbol{I}_n - \text{diag}\{h_{11}, \ldots, h_{nn}\}\right)^{-1} = \text{diag}\left\{\left(1 - h_{11}\right)^{-1}, \ldots, \left(1 - h_{nn}\right)^{-1}\right\} \text{ then } \tilde{\boldsymbol{e}} = \boldsymbol{M}^* \hat{\boldsymbol{e}}$$

Then a natural estimator of out-of-sample mean squared error is

$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \tilde{e}_i^2 = \frac{1}{n}\sum_{i=1}^n \left(1 - h_{ii}\right)^{-2} \hat{e}_i^2$ (**sample mean squared prediction error**). Its root $\tilde{\sigma} = \sqrt{\tilde{\sigma}^2}$ is known as the **prediction standard error**.

OLS: Mean

Assumption 2: linear Regression Model

The observations (Y_i, X_i) satisfy the linear regression equation

$$Y_i = X_i'\beta + e_i,$$
$$\mathbb{E}(e_i \mid X_i) = 0.$$

The variables have finite second moments

$$\mathbb{E}\left(Y_i^2\right) < \infty,$$
$$\mathbb{E}||X_i||^2 < \infty,$$

and an invertible design matrix

$$\boldsymbol{Q}_{XX} = \mathbb{E}(X_i X_i') > 0.$$

Assumption 3: Homoskedastic linear Regression Model

In addition to Assumption 2 ,

$$\mathbb{E}\left(e_i^2 \mid X_i\right) = \sigma^2(X_i) = \sigma^2,$$

is independent of X_i.

- We will consider both the general case of **heteroskedastic regression**, where the conditional variance $\mathbb{E}(e_i^2 \mid X_i) = \sigma^2(X_i) = \sigma_i^2$

Mean of Least-Squares Estimator

$$\mathbb{E}(Y_i \mid \boldsymbol{X}) = \mathbb{E}(Y_i \mid X_i) = X_i'\beta$$

- The first equality states that the conditional expectation of Y_i given $\{X_1, \dots, X_n\}$ only depends on X_i , since the observations are independent across i. The second equality is the assumption of a linear conditional mean.

(Conditional) Unbiasedness

$$\begin{aligned}
\mathbb{E}(\hat{\beta} \mid \boldsymbol{X}) &= \mathbb{E}\left(\left(\sum_{i=1}^n X_i X_i'\right)^{-1}\left(\sum_{i=1}^n X_i Y_i\right) \mid \boldsymbol{X}\right) \\
&= \left(\sum_{i=1}^n X_i X_i'\right)^{-1} \mathbb{E}\left(\sum_{i=1}^n X_i Y_i \mid \boldsymbol{X}\right) \\
&= \left(\sum_{i=1}^n X_i X_i'\right)^{-1} \sum_{i=1}^n \mathbb{E}(X_i Y_i \mid \boldsymbol{X}) \\
&= \left(\sum_{i=1}^n X_i X_i'\right)^{-1} \sum_{i=1}^n X_i \mathbb{E}(Y_i \mid \boldsymbol{X}) \\
&= \left(\sum_{i=1}^n X_i X_i'\right)^{-1} \sum_{i=1}^n X_i X_i'\beta \\
&= \beta
\end{aligned}$$

Theorem: Mean of Least-Squares Estimator

In the linear regression model (Assumption 2) and i.i.d. sampling (Assumption 1)

$$\mathbb{E}(\hat{\beta} \mid \boldsymbol{X}) = \beta.$$

Matrix Notation:

$$\begin{aligned}
\mathbb{E}(\hat{\beta} \mid \boldsymbol{X}) &= \mathbb{E}\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \mid \boldsymbol{X}\right) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{E}(\boldsymbol{Y} \mid \boldsymbol{X}) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\beta \\
&= \beta
\end{aligned}$$

$$\mathbb{E}(\boldsymbol{Y} \mid \boldsymbol{X}) = \begin{pmatrix} \vdots \\ \mathbb{E}(Y_i \mid \boldsymbol{X}) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ X_i'\beta \\ \vdots \end{pmatrix} = \boldsymbol{X}\beta$$

$$\mathbb{E}(\boldsymbol{e} \mid \boldsymbol{X}) = \begin{pmatrix} \vdots \\ \mathbb{E}(e_i \mid \boldsymbol{X}) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \mathbb{E}(e_i \mid X_i) \\ \vdots \end{pmatrix} = 0$$

OLS: Variance

The objective of this section is to derived the **(conditional)** **covariance matrix** of the regression coefficient estimates, i.e.,

$$\boldsymbol{V}_{\hat{\beta}} \overset{\text{def}}{=} \text{var}(\hat{\beta} \mid \boldsymbol{X}).$$

For any pair (Z, X) the **conditional** **covariance matrix** is

$$\text{var}(Z \mid X) = \mathbb{E}\big((Z - \mathbb{E}(Z \mid X))(Z - \mathbb{E}(Z \mid X))' \mid X\big)$$

For any $r \times 1$ *random* vector Z the **covariance matrix** is

$$\begin{aligned}\text{var}(Z) &= \mathbb{E}\big((Z - \mathbb{E}(Z))(Z - \mathbb{E}(Z))'\big) \\ &= \mathbb{E}(ZZ') - (\mathbb{E}(Z))(\mathbb{E}(Z))'\end{aligned}$$

Firstly, notice that

$$\begin{aligned}\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\big(\boldsymbol{X}'(\boldsymbol{X}\beta + \boldsymbol{e})\big) \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{e}) \\ &= \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}\end{aligned}$$

$$\hat{\beta} - \beta = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}$$

$$\hat{\beta} - \mathbb{E}(\hat{\beta} \mid \boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}$$

$$\boldsymbol{V}_{\hat{\beta}} = \text{var}(\hat{\beta} \mid \boldsymbol{X}) = \mathbb{E}\big((\hat{\beta} - \mathbb{E}(\hat{\beta} \mid \boldsymbol{X}))(\hat{\beta} - \mathbb{E}(\hat{\beta} \mid \boldsymbol{X}))' \mid \boldsymbol{X}\big) = \mathbb{E}\big((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}\boldsymbol{e}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \mid \boldsymbol{X}\big) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{E}(\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

$$\boldsymbol{V}_{\hat{\beta}} = \text{var}(\hat{\beta} \mid \boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

The ith diagonal element of $\boldsymbol{D}$ is

$$\mathbb{E}(e_i^2 \mid \boldsymbol{X}) = \mathbb{E}(e_i^2 \mid X_i) = \sigma_i^2$$

while the ijth off-diagonal element of $\boldsymbol{D}$ is

$$\mathbb{E}(e_i e_j \mid \boldsymbol{X}) = \mathbb{E}(e_i \mid X_i)\mathbb{E}(e_j \mid X_j) = 0$$

Therefore

$$\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X} = \sum_{i=1}^{n} X_i X_i' \sigma_i^2$$

$$\boldsymbol{D} = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

Theorem: Variance of Least-Squares Estimator In the linear regression model (Assumption 2) and i.i.d. sampling (Assumption 1)

$$\begin{aligned}\boldsymbol{V}_{\hat{\beta}} &= \text{var}(\hat{\beta} \mid \boldsymbol{X}) \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1}\end{aligned} \quad ,$$

where $\boldsymbol{D}$ is defined above.

In the homoskedastic linear regression model (Assumption 3) and i.i.d. sampling (Assumption 1)

$$\boldsymbol{V}_{\hat{\beta}} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \quad .$$

Residuals

	Residuals		Prediction Errors
Definition:	$\hat{e}_i = Y_i - X_i'\hat{\beta}$		$\tilde{e}_i = Y_i - X_i'\hat{\beta}_{(-i)}$ $\tilde{e}_i = (1-h_{ii})^{-1}\hat{e}_i$
Matrix Notation:	$\hat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{e}$ $\boldsymbol{M} = \boldsymbol{I_n} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$		$\tilde{\boldsymbol{e}} = \boldsymbol{M}^*\hat{\boldsymbol{e}} = \boldsymbol{M}^*\boldsymbol{M}\boldsymbol{e}$ $\boldsymbol{M}^*$ is a diagonal matrix with ith diagonal element $(1-h_{ii})^{-1}$
(Conditional) Mean:	$\mathbb{E}(\hat{\boldsymbol{e}} \mid \boldsymbol{X}) = \mathbb{E}(\boldsymbol{M}\boldsymbol{e} \mid \boldsymbol{X}) = \boldsymbol{M}\mathbb{E}(\boldsymbol{e} \mid \boldsymbol{X}) = \boldsymbol{0}$		$\mathbb{E}(\tilde{\boldsymbol{e}} \mid \boldsymbol{X}) = \boldsymbol{M}^*\boldsymbol{M}\mathbb{E}(\boldsymbol{e} \mid \boldsymbol{X}) = \boldsymbol{0}$
(Conditional) Variance:	$\mathrm{var}(\hat{\boldsymbol{e}} \mid \boldsymbol{X}) = \mathrm{var}(\boldsymbol{M}\boldsymbol{e} \mid \boldsymbol{X}) = \boldsymbol{M}\mathrm{var}(\boldsymbol{e} \mid \boldsymbol{X})\boldsymbol{M} = \boldsymbol{M}\boldsymbol{D}\boldsymbol{M}$		$\mathrm{var}(\tilde{\boldsymbol{e}} \mid \boldsymbol{X}) = \boldsymbol{M}^*\boldsymbol{M}\mathrm{var}(\boldsymbol{e} \mid \boldsymbol{X})\boldsymbol{M}\boldsymbol{M}^* = \boldsymbol{M}^*\boldsymbol{M}\boldsymbol{D}\boldsymbol{M}\boldsymbol{M}^*$

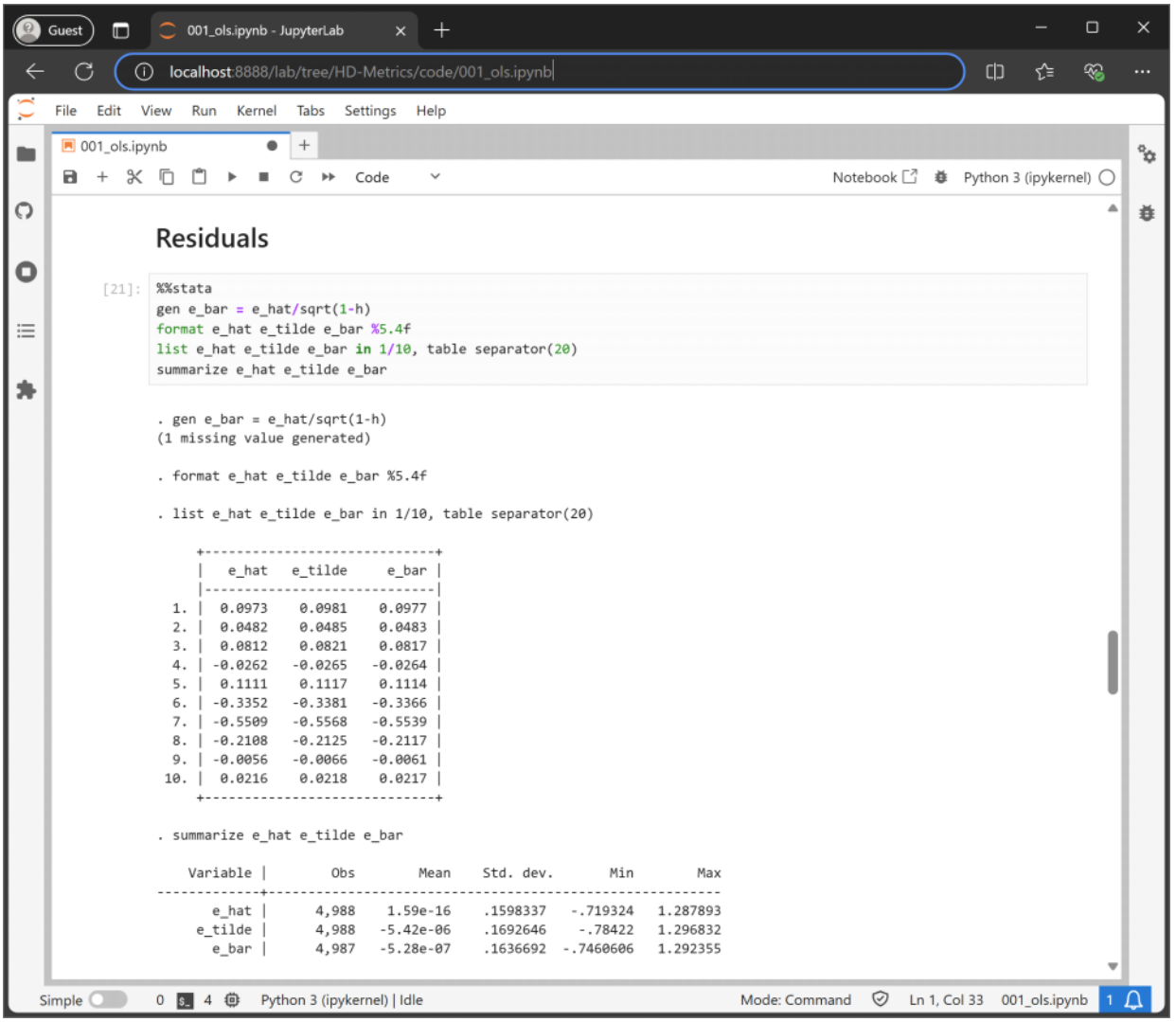
Homoskedasticity ($\boldsymbol{D} = \boldsymbol{I}_n\sigma^2$)

	Residuals		Prediction Errors
(Conditional) Variance:	$\mathrm{var}(\hat{\boldsymbol{e}} \mid \boldsymbol{X}) = \boldsymbol{M}\sigma^2$		$\mathrm{var}(\tilde{\boldsymbol{e}} \mid \boldsymbol{X}) = \boldsymbol{M}^*\boldsymbol{M}\boldsymbol{M}^*\sigma^2$
ith Observation:	$\mathrm{var}(\hat{e}_i \mid \boldsymbol{X}) = \mathbb{E}(\hat{e}_i^2 \mid \boldsymbol{X}) = (1-h_{ii})\sigma^2$		$\mathrm{var}(\tilde{e}_i \mid \boldsymbol{X}) = \mathbb{E}(\tilde{e}_i^2 \mid \boldsymbol{X}) = (1-h_{ii})^{-1}(1-h_{ii})(1-h_{ii})^{-1}\sigma^2 = (1-h_{ii})^{-1}\sigma^2$

When the true errors, $\boldsymbol{e}$, are *homoskedastic*, then a residual with constant conditional variance can be obtained by rescaling. The **standardized residuals** are

	Standardized Residuals
Definition:	$\bar{e}_i = (1-h_{ii})^{-1/2}\hat{e}_i$
Matrix Notation:	$\bar{\boldsymbol{e}} = (\bar{e}_1, \ldots, \bar{e}_n)' = \boldsymbol{M}^{*1/2}\boldsymbol{M}\boldsymbol{e}$
(Conditional) Variance:	$\mathrm{var}(\bar{\boldsymbol{e}} \mid \boldsymbol{X}) = \boldsymbol{M}^{*1/2}\boldsymbol{M}\boldsymbol{M}^{*1/2}\sigma^2$

$\mathrm{var}(\bar{e}_i \mid \boldsymbol{X}) = \mathbb{E}(\bar{e}_i^2 \mid \boldsymbol{X}) = \sigma^2$


Guest
001_ols.ipynb - JupyterLab
localhost:8888/lab/tree/HD-Metrics/code/001_ols.ipynb
File Edit View Run Kernel Tabs Settings Help
001_ols.ipynb
Code
Notebook
Python 3 (ipykernel)
Residuals
[21]:
%%stata
gen e_bar = e_hat/sqrt(1-h)
format e_hat e_tilde e_bar %5.4f
list e_hat e_tilde e_bar in 1/10, table separator(20)
summarize e_hat e_tilde e_bar
. gen e_bar = e_hat/sqrt(1-h)
(1 missing value generated)
. format e_hat e_tilde e_bar %5.4f
. list e_hat e_tilde e_bar in 1/10, table separator(20)
+-----------------------------+
e_hat e_tilde e_bar
1. | 0.0973 0.0981 0.0977 |
2. | 0.0482 0.0485 0.0483 |
3. | 0.0812 0.0821 0.0817 |
4. | -0.0262 -0.0265 -0.0264 |
5. | 0.1111 0.1117 0.1114 |
6. | -0.3352 -0.3381 -0.3366 |
7. | -0.5509 -0.5568 -0.5539 |
8. | -0.2108 -0.2125 -0.2117 |
9. | -0.0056 -0.0066 -0.0061 |
10. | 0.0216 0.0218 0.0217 |
+-----------------------------+
. summarize e_hat e_tilde e_bar
Variable | Obs Mean Std. dev. Min Max
e_hat | 4,988 1.59e-16 .1598337 -.719324 1.287893
e_tilde | 4,988 -5.42e-06 .1692646 -.78422 1.296832
e_bar | 4,987 -5.28e-07 .1636692 -.7460606 1.292355
Simple
0 4
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 33
001_ols.ipynb
1

Covariance Matrix Estimation

1. Homoskedastic Case
2. Heteroskedastic Case

Homoskedastic Case

Under homoskedasticity, the covariance matrix takes the relatively simple form

$$\boldsymbol{V}_{\widehat{\beta}}^{0} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2$$

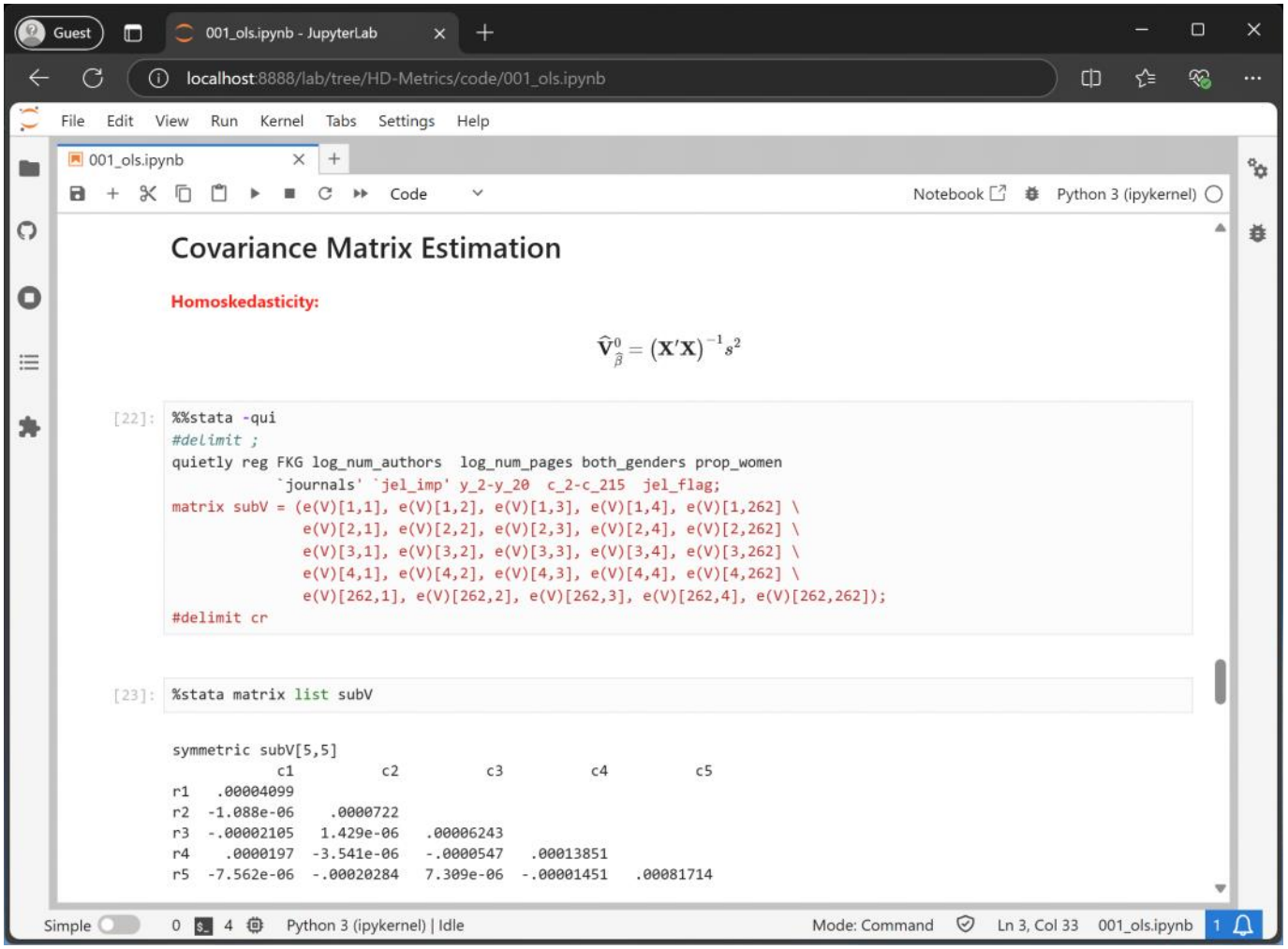
Using the 'plug-in' principle, the classic covariance matrix estimator in the homoskedastic case is

$$\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{0} = (\boldsymbol{X}'\boldsymbol{X})^{-1}s^2$$

Reminder:

Object Definition:	$\boldsymbol{V}_{\widehat{\beta}}^{0} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2$
Estimator:	$\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{0} = (\boldsymbol{X}'\boldsymbol{X})^{-1}s^2$
(Conditional) Mean Of the Estimator:	$\mathbb{E}\left(\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{0} \mid \boldsymbol{X}\right) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\mathbb{E}(s^2 \mid \boldsymbol{X})$ $= (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2$ $= \boldsymbol{V}_{\widehat{\beta}}^{0}$

✔ $\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{0}$ is conditionally unbiased for $\boldsymbol{V}_{\widehat{\beta}}^{0}$ under the assumption of homoskedasticity.


Guest
001_ols.ipynb - JupyterLab
localhost:8888/lab/tree/HD-Metrics/code/001_ols.ipynb
File Edit View Run Kernel Tabs Settings Help
001_ols.ipynb
Code
Notebook
Python 3 (ipykernel)
Covariance Matrix Estimation
Homoskedasticity:
$$\widehat{\mathbf{V}}^0_{\hat{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} s^2$$
[22]:
%%stata -qui
#delimit ;
quietly reg FKG log_num_authors log_num_pages both_genders prop_women
`journals' `jel_imp' y_2-y_20 c_2-c_215 jel_flag;
matrix subV = (e(V)[1,1], e(V)[1,2], e(V)[1,3], e(V)[1,4], e(V)[1,262] \
e(V)[2,1], e(V)[2,2], e(V)[2,3], e(V)[2,4], e(V)[2,262] \
e(V)[3,1], e(V)[3,2], e(V)[3,3], e(V)[3,4], e(V)[3,262] \
e(V)[4,1], e(V)[4,2], e(V)[4,3], e(V)[4,4], e(V)[4,262] \
e(V)[262,1], e(V)[262,2], e(V)[262,3], e(V)[262,4], e(V)[262,262]);
#delimit cr
[23]:
%stata matrix list subV
symmetric subV[5,5]
c1 c2 c3 c4 c5
r1 .00004099
r2 -1.088e-06 .0000722
r3 -.00002105 1.429e-06 .00006243
r4 .0000197 -3.541e-06 -.0000547 .00013851
r5 -7.562e-06 -.00020284 7.309e-06 -.00001451 .00081714
Simple
0 4
Python 3 (ipykernel) | Idle
Mode: Command
Ln 3, Col 33
001_ols.ipynb
1

Heteroskedastic Case

Under heteroskedasticity, the general form for the covariance matrix is

$$\boldsymbol{V}_{\hat{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Where

$$\begin{aligned}\boldsymbol{D} &= \operatorname{diag}(\sigma_1^2, \dots, \sigma_n^2) \\ &= \mathbb{E}(\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}) \\ &= \mathbb{E}(\tilde{\boldsymbol{D}} \mid \boldsymbol{X})\end{aligned}$$

$\tilde{\boldsymbol{D}} = \operatorname{diag}(e_1^2, \dots, e_n^2)$ is a conditionally unbiased estimator for $\boldsymbol{D}$.

Therefore, if the squared errors e_i^2 were observable, we could construct and unbiased 'ideal' estimator for $\boldsymbol{V}_{\hat{\beta}}$ as

Reminder:

Object Definition:	$\boldsymbol{V}_{\hat{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1}$
(ideal) Estimator:	$\tilde{\boldsymbol{V}}_{\hat{\beta}}^{\text{ideal}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\tilde{\boldsymbol{D}}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{i=1}^{n} X_i X_i' e_i^2\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}$
(Conditional) Mean Of the Estimator:	$\mathbb{E}\left(\tilde{\boldsymbol{V}}_{\hat{\beta}}^{\text{ideal}} \mid \boldsymbol{X}\right) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{i=1}^{n} X_i X_i' \mathbb{E}(e_i^2 \mid \boldsymbol{X})\right)(\boldsymbol{X}'\boldsymbol{X})^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{i=1}^{n} X_i X_i' \sigma_i^2\right)(\boldsymbol{X}'\boldsymbol{X})^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1} = \boldsymbol{V}_{\hat{\beta}}$

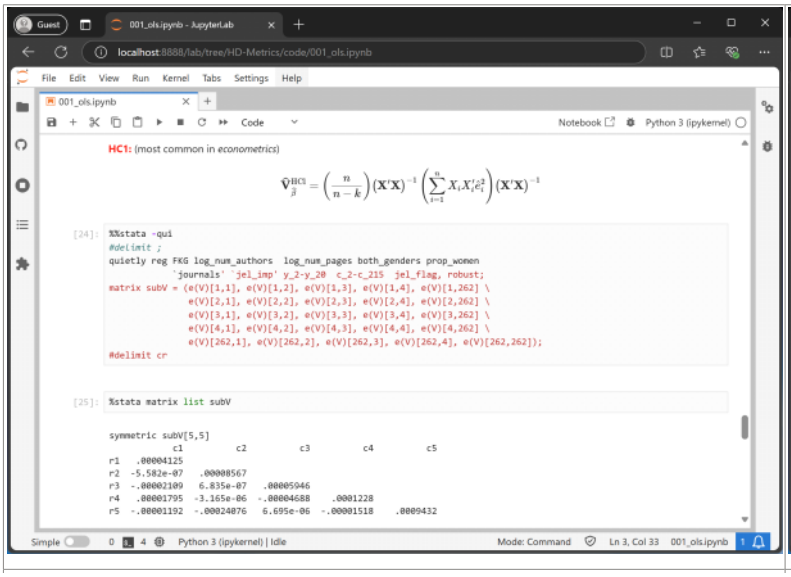
⛔ Unfortunately, this 'ideal' estimator (which is conditional unbiased) is infeasible because we do not observe $e_1^2, \dots, e_n^2$.

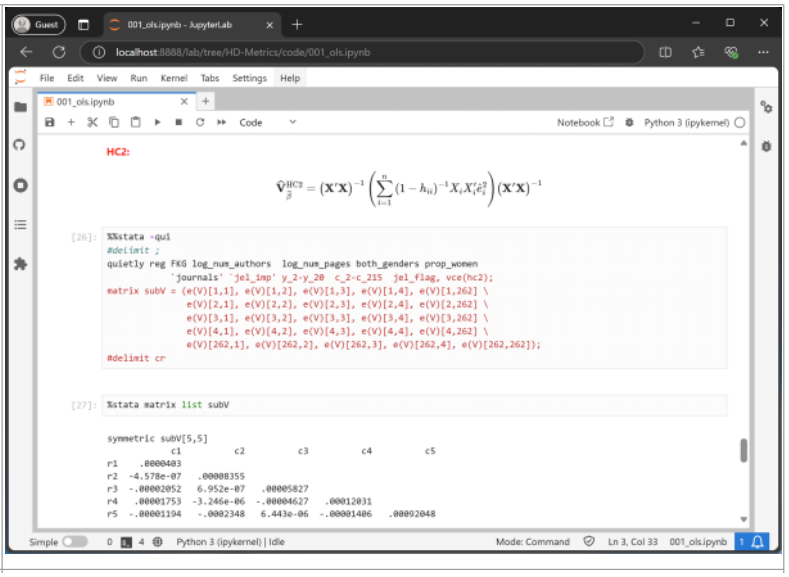
Feasible Heteroskedasticity-Consistent (HC) Estimators

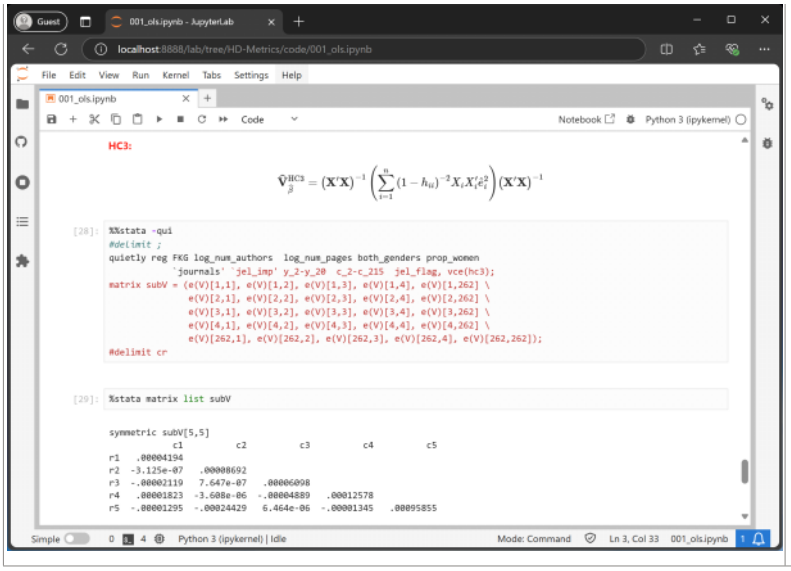
HC0:		$\hat{\boldsymbol{V}}_{\hat{\beta}}^{\text{HC0}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{i=1}^{n} X_i X_i' \hat{e}_i^2\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}$
HC1:		$\hat{\boldsymbol{V}}_{\hat{\beta}}^{\text{HC1}} = \left(\frac{n}{n-k}\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{i=1}^{n} X_i X_i' \hat{e}_i^2\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}$
HC2:		$\hat{\boldsymbol{V}}_{\hat{\beta}}^{\text{HC2}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{i=1}^{n} X_i X_i' \tilde{e}_i^2\right)(\boldsymbol{X}'\boldsymbol{X})^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{i=1}^{n} (1-h_{ii})^{-1} X_i X_i' \hat{e}_i^2\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}$
HC3:		$\hat{\boldsymbol{V}}_{\hat{\beta}}^{\text{HC3}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{i=1}^{n} X_i X_i' \tilde{e}_i^2\right)(\boldsymbol{X}'\boldsymbol{X})^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{i=1}^{n} (1-h_{ii})^{-2} X_i X_i' \hat{e}_i^2\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}$

⚠ The default *unchallenged* choice in applied work.

Since $(1-h_{ii})^{-2} > (1-h_{ii})^{-1} > 1$ it is straightforward to show that $\hat{\boldsymbol{V}}_{\hat{\beta}}^{\text{HC0}} < \hat{\boldsymbol{V}}_{\hat{\beta}}^{\text{HC2}} < \hat{\boldsymbol{V}}_{\hat{\beta}}^{\text{HC3}}$







Standard Error

Definition: A standard error $s(\hat{\beta})$ for a real-valued estimator $\hat{\beta}$ is an estimator of the standard deviation of the distribution of $\hat{\beta}$.

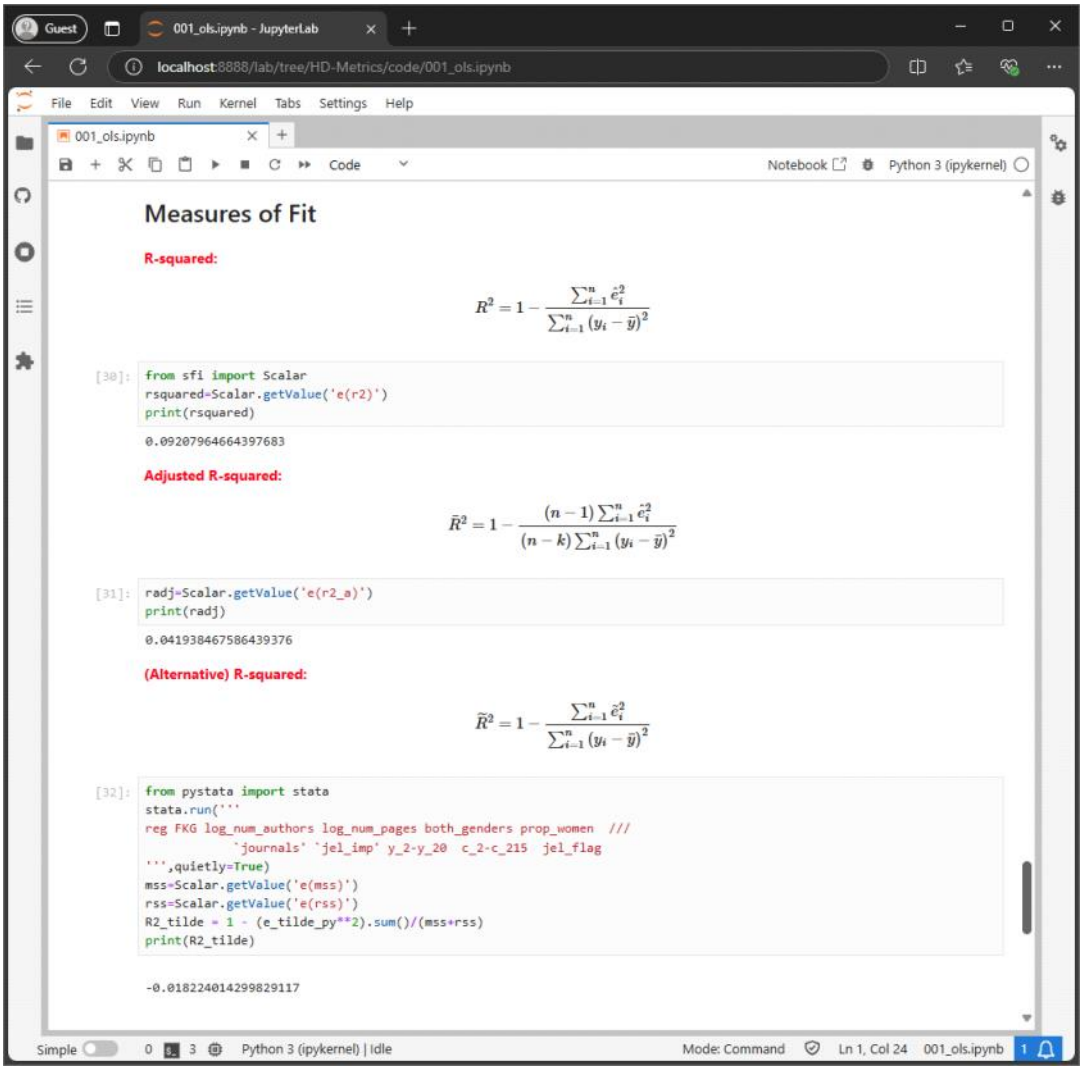
When β is a vector with estimator $\hat{\beta}$ and covariance matrix estimator $\widehat{\boldsymbol{V}}_{\hat{\beta}}^{?}$, standard errors for individual elements are the square roots of the diagonal elements of $\widehat{\boldsymbol{V}}_{\hat{\beta}}^{?}$. That is,

$$s(\hat{\beta}_j) = \sqrt{\widehat{\boldsymbol{V}}_{\hat{\beta}_j}} = \sqrt{\left[\widehat{\boldsymbol{V}}_{\hat{\beta}}^{?}\right]_{jj}}$$

Where ? ∈ {**0**, **HC0**, **HC1**, **HC2**, **HC3**}

Measures of Fit

	Definition		Comments
R-Squared (R^2):	$R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{e}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$		R^2 can be viewed as an estimator of the population parameter $\rho^2 = \frac{\text{var}(X_i'\beta)}{\text{var}(y_i)} = 1 - \frac{\sigma^2}{\sigma_y^2}$ ⚠ It cannot decrease as k increases.
(Adjusted) R-Squared ($\bar{R}^2$):	$\bar{R}^2 = 1 - \frac{(n-1)\sum_{i=1}^{n} \hat{e}_i^2}{(n-k)\sum_{i=1}^{n} (y_i - \bar{y})^2}$		It is called 'adjusted' because it corrects for the terms' degrees-of-freedom. ⚠$\bar{R}^2$ is an inappropriate choice for model selection (it tends to select models with too many parameters). However it is commonly reported in applied work.
(Alternative) R-Squared ($\tilde{R}^2$):	$\tilde{R}^2 = 1 - \frac{\sum_{i=1}^{n} \tilde{e}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$		✓ Models with high $\tilde{R}^2$ are better models in terms of expected out of sample squared error, and therefore suitable for model selection.



Cluster Sampling

⚠ A difficulty with applying the classical regression framework is that articles may share the same co-author across journals and time. Readability of different articles may be affected by writing style of shared co-authors, all of which imply dependence. These concerns, however, do not suggest that readability will be correlated across set of articles that do not share a co-author, so it seems reasonable to model readability across these sets as mutually independent.

In clustering contexts it is convenient to double index the observations as (Y_{ig}, X_{ig}) where $g = 1, \dots, G$ indexes the cluster and $i = 1, \dots, n_g$ indexes the individual within the gth cluster. The number of observations per cluster n_g may vary across clusters. The number of clusters is G. The total number of observations is $n = \sum_{g=1}^{G} n_g$. In article readability example, the number of clusters (cluster) in the estimation sample is $G = 215$ and the total number of observations is $n = 4{,}988$.

While it is typical to write the observations using the double index notation (Y_{ig}, X_{ig}), it is also useful to use cluster-level notation. Let $\boldsymbol{Y}_g = (Y_{1g}, \dots, Y_{n_g g})'$ and $\boldsymbol{X}_g = (X_{1g}, \dots, X_{n_g g})'$ denote the $n_g \times 1$ vector of

dependent variables and $n_g \times k$ matrix of regressors for the gth cluster. A linear regression model can be written for the individual observations as

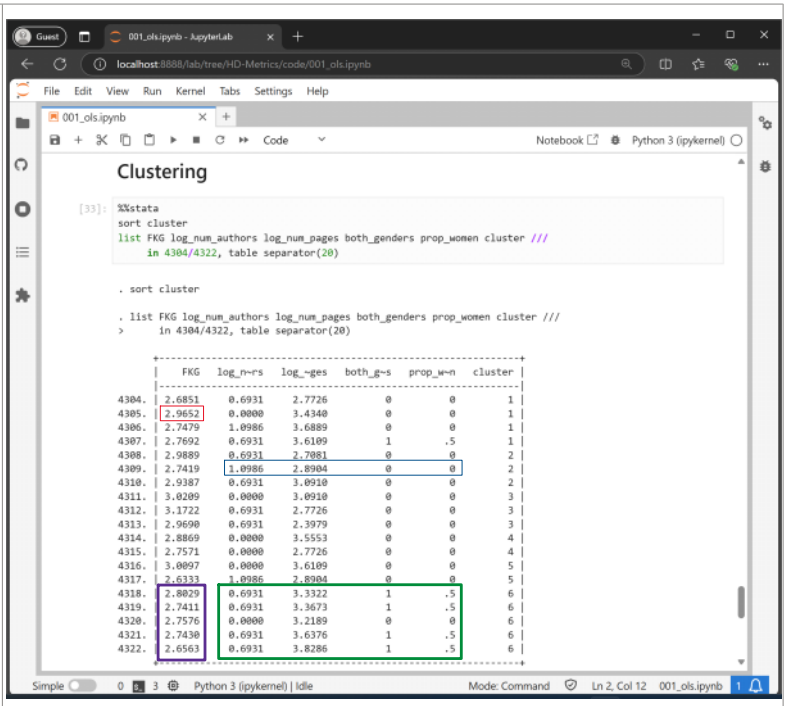
$Y_{ig} = X_{ig}'\beta + e_{ig}$

and using cluster notation as

$\boldsymbol{Y}_g = \boldsymbol{X}_g\beta + \boldsymbol{e}_g$

where $\boldsymbol{e}_g = (e_{1g}, \dots, e_{n_g g})'$ is a $n_g \times 1$ error vector. We can also stack the observations into full sample matrices and write the model as

$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{e}$



Least Squares Estimation

$$\hat{\beta} = \left(\sum_{g=1}^{G}\sum_{i=1}^{n_g} X_{ig}X_{ig}'\right)^{-1}\left(\sum_{g=1}^{G}\sum_{i=1}^{n_g} X_{ig}Y_{ig}\right)$$
$$= \left(\sum_{g=1}^{G} \boldsymbol{X}_g'\boldsymbol{X}_g\right)^{-1}\left(\sum_{g=1}^{G} \boldsymbol{X}_g'\boldsymbol{Y}_g\right)$$
$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y})$$

Individual Level:	$\hat{e}_{ig} = Y_{ig} - X_{ig}'\hat{\beta}$
Group Level:	$\hat{\boldsymbol{e}}_g = \boldsymbol{Y}_g - \boldsymbol{X}_g\hat{\beta}$

The model is a linear regression under the assumption

$\mathbb{E}(\boldsymbol{e}_g \mid \boldsymbol{X}_g) = 0$

This is the same as assuming that the individual errors are conditionally mean zero

$\mathbb{E}(e_{ig} \mid \boldsymbol{X}_g) = 0$

$$\hat{\beta} - \beta = \left(\sum_{g=1}^{G} \boldsymbol{X}_g'\boldsymbol{X}_g\right)^{-1}\left(\sum_{g=1}^{G} \boldsymbol{X}_g'\boldsymbol{e}_g\right)$$

Assumption: The clusters $(\boldsymbol{Y}_g, \boldsymbol{X}_g)$ are mutually independent across clusters g.

Theorem: In the clustered linear regression model $\mathbb{E}(\hat{\beta} \mid \boldsymbol{X}) = \beta$

(Conditional) Variance

Since the observations are *independent across clusters*

$$\text{var}\left(\left(\sum_{g=1}^{G} \boldsymbol{X}_g'\boldsymbol{e}_g\right) \mid \boldsymbol{X}\right) = \sum_{g=1}^{G} \text{var}\left(\boldsymbol{X}_g'\boldsymbol{e}_g \mid \boldsymbol{X}_g\right)$$
$$= \sum_{g=1}^{G} \boldsymbol{X}_g'\mathbb{E}\left(\boldsymbol{e}_g\boldsymbol{e}_g' \mid \boldsymbol{X}_g\right)\boldsymbol{X}_g$$
$$= \sum_{g=1}^{G} \boldsymbol{X}_g'\boldsymbol{\Sigma}_g\boldsymbol{X}_g$$
$$\overset{def}{=} \boldsymbol{\Omega}_n$$

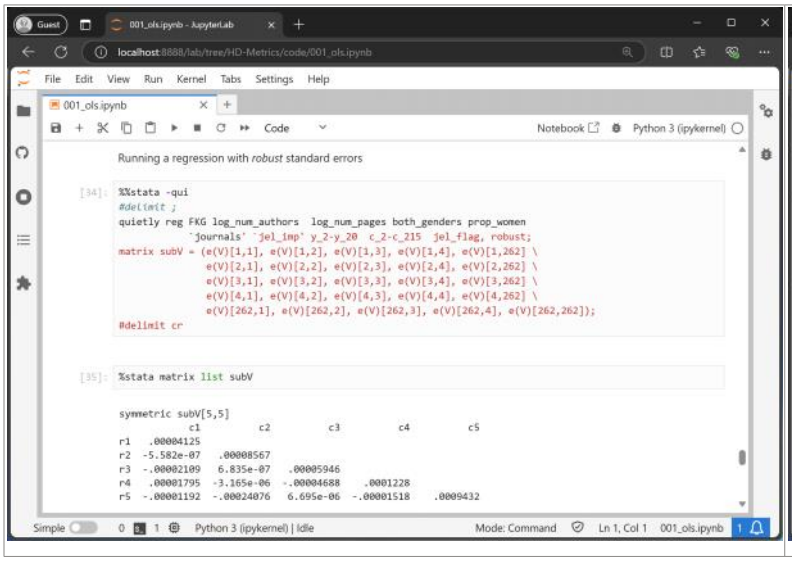
$$\boldsymbol{V}_{\hat{\beta}} = \text{var}(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\Omega}_n(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

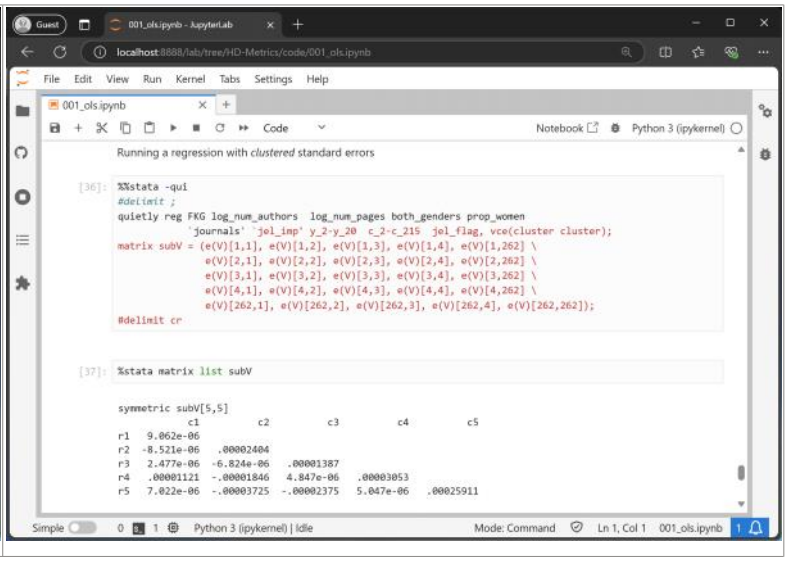
Estimator:	$\widehat{\boldsymbol{V}}_{\hat{\beta}} = a_n(\boldsymbol{X}'\boldsymbol{X})^{-1}\widehat{\boldsymbol{\Omega}}_n(\boldsymbol{X}'\boldsymbol{X})^{-1}$	

$$\widehat{\boldsymbol{\Omega}}_n = \sum_{g=1}^{G} \boldsymbol{X}_g' \hat{\boldsymbol{e}}_g \hat{\boldsymbol{e}}_g' \boldsymbol{X}_g$$
$$= \sum_{g=1}^{G} \sum_{i=1}^{n_g} \sum_{\ell=1}^{n_g} X_{ig} X_{\ell g}' \hat{e}_{ig} \hat{e}_{\ell g}$$
$$= \sum_{g=1}^{G} \left(\sum_{i=1}^{n_g} X_{ig} \hat{e}_{ig} \right) \left(\sum_{\ell=1}^{n_g} X_{\ell g} \hat{e}_{\ell g} \right)'$$

The factor $G/(G-1)$ was derived by Chris Hansen (2007) in the context of equal-sized clusters to improve performance when the number of clusters G is small. The factor $(n-1)/(n-k)$ is an ad hoc generalization which nests the adjustment used in HC1, since when $G = n$ we have the simplification $a_n = n/(n-k)$.

$$a_n = \left(\frac{n-1}{n-k} \right) \left(\frac{G}{G-1} \right)$$





A Review of Large Sample Asymptotics

Modes of Convergence

Definition: Convergence in Probability

A random vector $Z_n \in \mathbb{R}^k$ **converges in probability** to Z as $n \to \infty$, denoted $Z_n \to_p Z$ or alternatively $\text{plim}_{n\to\infty} Z_n = Z$, if for all $\delta > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left[\|Z_n - Z\| \leq \delta\right] = 1.$$

We call Z the **probability limit** (or **plim**) of Z_n.

This definition treats random variables and random vectors simultaneously using the vector norm.

It is useful to know that for a random vector, the definition holds if and only if each element in the vector converges in probability to its limit.

Definition: Convergence in Distribution

Let Z_n be a random vector with distribution $F_n(u) = \mathbb{P}\left[Z_n \leq u\right]$. We say that Z_n **converges in distribution** to Z as $n \to \infty$, denoted $Z_n \to_d Z$ if for all u at which $F(u) = \mathbb{P}[Z \leq u]$ is continuous, $F_n(u) \to F(u)$ as $n \to \infty$.

We refer to Z and its distribution $F(u)$ as the asymptotic distribution, large sample distribution, or limit distribution of Z_n.

Weak Law of Large Numbers

Theorem: Weak Law of Large Numbers (WLLN)

If $Y_i \in \mathbb{R}^k$ are i.i.d. and $\mathbb{E} \parallel Y \parallel < \infty$, then as $n \to \infty$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow[p]{} \mathbb{E}[Y].$$

The WLLN shows that the sample mean $\bar{Y}$ converges in probability to the true population expectation $\mathbb{E}[Y]$. The result applies to any transformation of a random vector with a finitemean.

Theorem:

If $Y_i \in \mathbb{R}^k$ are i.i.d., $h(y)\colon \mathbb{R}^k \to \mathbb{R}^q$ and $\mathbb{E} \parallel h(Y) \parallel < \infty$, then as $n \to \infty$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} h(Y_i) \xrightarrow[p]{} \mu = \mathbb{E}[h(Y)].$$

Lemma: For $Y \in \mathbb{R}^m$, $\mathbb{E} \parallel Y \parallel < \infty$ *if and only* if $\mathbb{E}|Y_j| < \infty$ for $j = 1, \ldots, m$.

Definition: An estimator $\hat{\theta}$ of θ is **consistent** if $\hat{\theta} \xrightarrow[p]{} \theta$ as $n \to \infty$.

An estimator which converges in probability to the population value is called **consistent**.

Central Limit Theorem

Theorem: Multivariate Lindeberg-Lévy Central limit Theorem (CLT)

If $Y_i \in \mathbb{R}^k$ are i.i.d. and $\mathbb{E} \parallel Y \parallel^2 < \infty$, then as $n \to \infty$

$$\sqrt{n}(\bar{Y} - \mu) \underset{d}{\longrightarrow} \mathrm{N}(0, \boldsymbol{V}),$$

where $\mu = \mathbb{E}[Y]$ and $\boldsymbol{V} = \mathbb{E}\left[(Y - \mu)(Y - \mu)'\right]$.

The central limit theorem shows that the distribution of the sample mean is approximately normal in large samples. For some applications it may be useful to notice that the Theorem does not impose any restrictions on $\boldsymbol{V}$ other than that the elements are finite.

Continuous Mapping Theorem

✓Continuous functions are limit-preserving.
✓There are two forms of the continuous mapping theorem, for convergence in probability and convergence in distribution.

Theorem: Continuous Mapping Theorem

Let $Z_n \in \mathbb{R}^k$ and $g: \mathbb{R}^k \to \mathbb{R}^q$. If $Z_n \xrightarrow[p]{} c$ as $n \to \infty$ and $g(u)$ is continuous at c then

$$g(Z_n) \xrightarrow[p]{} g(c) \text{ as } n \to \infty.$$

Theorem: Continuous Mapping Theorem

If $Z_n \xrightarrow[d]{} Z$ as $n \to \infty$ and $g: \mathbb{R}^m \to \mathbb{R}^k$ has the set of discontinuity points D_g such that $\mathbb{P}[Z \in D_g] = 0$ then

$$g(Z_n) \xrightarrow[d]{} g(Z) \text{ as } n \to \infty.$$

Theorem: Slutsky's Theorem

If $Z_n \xrightarrow[d]{} Z$ and $c_n \xrightarrow[p]{} c$ as $n \to \infty$, then

$$Z_n + c_n \xrightarrow[d]{} Z + c$$
$$Z_n c_n \xrightarrow[d]{} Zc$$
$$\frac{Z_n}{c_n} \xrightarrow[d]{} \frac{Z}{c}, \quad c \neq 0$$

Delta Method

Theorem: Delta Method

Let $\mu \in \mathbb{R}^k$ and $g: \mathbb{R}^k \to \mathbb{R}^q$. If $\sqrt{n}(\hat{\mu} - \mu) \longrightarrow \xi$, where $g(u)$ is continuously differentiable in a neighborhood of μ, then as $n \to \infty$

$$\sqrt{n}(g(\hat{\mu}) - g(\mu)) \underset{d}{\longrightarrow} \boldsymbol{G}'\xi,$$

where $\boldsymbol{G}(u) = \frac{\partial}{\partial u} g(u)'$ and $\boldsymbol{G} = \boldsymbol{G}(\mu)$.

In particular, if $\xi \sim \mathrm{N}(0, \boldsymbol{V})$ then as $n \to \infty$

$$\sqrt{n}(g(\hat{\mu}) - g(\mu)) \underset{d}{\longrightarrow} \mathrm{N}(0, \boldsymbol{G}'\boldsymbol{V}\boldsymbol{G}).$$

Differentiable functions of asymptotically normal random estimators are asymptotically normal.

Example:

Suppose that $\hat{\mu} \equiv [\hat{\mu}_{n,1}, \hat{\mu}_{n,2}]'$ is a random 2×1 vector that satifies $\sqrt{n}(\hat{\mu}_n - \mu) \underset{d}{\longrightarrow} N(0, V)$, where $\mu \equiv [1 \quad 2]$, and $V \equiv \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$ and we want to derive the asymptotic distribution of the sequence $\{\hat{\mu}_{n,1}^2 + \hat{\mu}_{n,2}^3\}$. Obviously, $g(\hat{\mu}) = \hat{\mu}_{n,1}^2 + \hat{\mu}_{n,2}^3$, and $g(\mu) = 1^2 + 2^3 = 9$, and $\boldsymbol{G}(u) = \frac{\partial}{\partial u} g(u)' = \left[\frac{\partial}{\partial u_1} g(u) \quad \frac{\partial}{\partial u_2} g(u)\right]' = [2u_1 \quad 3u_2^2]'$, so $\boldsymbol{G} = \boldsymbol{G}(\mu) = [2 \cdot 1 \quad 3 \cdot 2^2]'$ so $\boldsymbol{G}'\boldsymbol{V}\boldsymbol{G} = [2 \quad 12] \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 12 \end{bmatrix} = 200$, and we conclude that $\sqrt{n}(\hat{\mu}_{n,1}^2 + \hat{\mu}_{n,2}^3 - 9) \underset{d}{\longrightarrow} N(0, 200)$.

Note: See other exercises here.

Smooth Function Model

The smooth function model is $\theta = g(\mu)$ where $\mu = \mathbb{E}[h(Y)]$ and $g(\mu)$ is smooth in a suitable sense (see below).

$$\theta = g\big(\mathbb{E}[h(Y)]\big).$$

⚠ The parameter of interest θ is a smooth function of a population mean. It is not a population moment so it does not have a direct moment estimator. Instead, we use the 'plug-in' principle and replace μ by its natural moment estimator, i.e.,

$$\hat{\mu} = n^{-1}\sum_{i=1}^{n} h(Y_i),$$
$$\hat{\theta} = g\left(n^{-1}\sum_{i=1}^{n} h(Y_i)\right).$$

Theorem: If $Y_i \in \mathbb{R}^m$ are i.i.d., $h(u)\colon \mathbb{R}^m \to \mathbb{R}^k$, $\mathbb{E} \parallel h(Y) \parallel < \infty$ and $g(u)\colon \mathbb{R}^k \to \mathbb{R}^q$ is continuous at μ, then $\hat{\theta} \underset{p}{\longrightarrow} \theta$ as $n \to \infty$.

Theorem: If $Y_i \in \mathbb{R}^m$ are i.i.d., $h(u)\colon \mathbb{R}^m \to \mathbb{R}^k$, $\mathbb{E} \parallel h(Y) \parallel^2 < \infty$, $g(u)\colon \mathbb{R}^k \to \mathbb{R}^q$, and $\boldsymbol{G}(u) = \frac{\partial}{\partial u} g(u)'$ is continuous in a neighborhood of μ then as $n \to \infty$

$$\sqrt{n}(\hat{\theta} - \theta) \underset{d}{\longrightarrow} \mathrm{N}(0, \boldsymbol{V}_\theta).$$

where $\boldsymbol{V}_\theta =$

Consistency requires that $h(Y)$ has a finite expectation; asymptotic normality requires that $h(Y)$ has a finite variance.

Consistency requires that $g(u)$ be continuous; asymptotic normality requires that $g(u)$ is continuously differentiable.

Example:
Let $\{Y_1, \ldots, Y_n\}$ be a random sample taken from the population distribution $F(x) = \Pr\{Y \le x\}$. Recall that the **Empirical CDF** is defined as $\hat{F}(x) = n^{-1}\sum_{i=1}^{n} 1\{Y_i \le x\}$. Using this notation, we simply set

$g(u) = u$

$h(x) = 1\{Y \le x\}$

$\mu = \mathbb{E}[h(x)] = \mathbb{E}[1\{Y \le x\}] = \int 1\{Y \le x\} dF(Y) = F(x)$

$\hat{\mu} = n^{-1}\sum_{i=1}^{n} 1\{Y_i \le x\}.$

Clearly $\mathbb{E}|1\{Y \le x\}| \le 1 < \infty$, so we can conclude by the Theorem that $\hat{\mu} \underset{p}{\longrightarrow} \mu$ or equivalently that $\hat{F}(x) \underset{p}{\longrightarrow} F(x)$.
The Empirical CDF is a consistent estimator of the unknown CDF.

Similarly, by the second Theorem we have that

$\sqrt{n}[\hat{F}(x) - F(x)] \underset{d}{\longrightarrow} N(0, F(x)[1 - F(x)])$ as $n \to \infty$.

Stochastic Order Symbols

Simple symbols for random variables and vectors which converge in probability to zero or are stochastically bounded.

"small oh-P-one"		"big oh-P-one"
Let Z_n and $a_n, n = 1,2, \ldots$ be sequences of random variables and constants respectively. The notation $Z_n = o_p(1)$ ("small oh-P-one") means that $Z_n \underset{p}{\longrightarrow} 0$ as $n \to \infty$. We also write $Z_n = o_p(a_n)$ if $a_n^{-1} Z_n = o_p(1)$.		Similarly, the notation $Z_n = O_p(1)$ ("big oh-P-one") means that Z_n is *bounded in probability*. Precisely, for any $\epsilon > 0$ there is a constant $M_\epsilon < \infty$ such that $\limsup_{n\to\infty} \mathbb{P}\left[\left|Z_n\right| > M_\epsilon\right] \leq \epsilon.$ Furthermore, we write $Z_n = O_p(a_n)$ if $a_n^{-1} Z_n = O_p(1)$.

- $O_p(1)$ is weaker than $o_p(1)$ in the sense that $Z_n = o_p(1)$ implies $Z_n = O_p(1)$ but not the reverse.
- However, if $Z_n = O_p(a_n)$ then $Z_n = o_p(b_n)$ for any b_n such that $a_n / b_n \to 0$.

Algebra of Stochastic Orders

$$
\begin{aligned}
o_p(1) + o_p(1) &= o_p(1) \\
o_p(1) + O_p(1) &= O_p(1) \\
O_p(1) + O_p(1) &= O_p(1) \\
o_p(1) o_p(1) &= o_p(1) \\
o_p(1) O_p(1) &= o_p(1) \\
O_p(1) O_p(1) &= O_p(1)
\end{aligned}
$$

- If $X_n = O_p(f_n)$, and $Y_n = O_p(g_n)$ then
  $X_n Y_n = O_p(f_n g_n)$,
  $X_n + Y_n = O_p\left(\max(f_n, g_n)\right)$.
- If $X_n = O_p(f_n)$, and $Y_n = o_p(g_n)$, then $X_n Y_n = o_p(f_n g_n)$.

Example (Sample Mean): Let $\{Z_i\}_{i=1}^n$ be a random sample from a distribution such $\mathbb{E}[Z] = 0$ and $\text{var}[Z] = 1$ and let $X_n = n^{-1} \sum_{i=1}^n Z_i$:

$$X_n = \frac{1}{\sqrt{n}} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i \right) = \frac{1}{\sqrt{n}} A_n,$$

$\mathbb{E}[A_n] = 0$, and $\text{var}[A_n] = 1$, it then follows by the Chebyshev's inequality, i.e. $\forall \delta > 0$

$$
\begin{aligned}
\mathbb{P}\{|A_n| \geq \delta\} &\leq \mathbb{E}\left(|A_n|^2\right) / \delta^2 \\
&\leq \delta^{-2} \\
\mathbb{P}\{|A_n| \geq M_\epsilon\} &\leq \epsilon
\end{aligned}
\quad ,
$$

where $M_\epsilon \equiv \delta$, and $\epsilon \equiv \delta^{-2}$. Therefore $A_n = O_p(1)$ and

$$X_n = \frac{1}{\sqrt{n}} O_p(1) = O\left(n^{-1/2}\right) O_p(1) = O_p\left(n^{-1/2}\right).$$

Consistency: OLS Estimator

Theorem: Consistency of the OLS

Under Assumption 1, $\widehat{\boldsymbol{Q}}_{XX} \xrightarrow[p]{} \boldsymbol{Q}_{XX}$, $\widehat{\boldsymbol{Q}}_{XY} \xrightarrow[p]{} \boldsymbol{Q}_{XY}$, $\widehat{\boldsymbol{Q}}_{XX}^{-1} \xrightarrow[p]{} \boldsymbol{Q}_{XX}^{-1}$, $\widehat{\boldsymbol{Q}}_{Xe} \xrightarrow[p]{} \mathbf{0}$ and $\hat{\beta} \xrightarrow[p]{} \beta$ as $n \to \infty$.

Proof: Recall that $\hat{\beta} - \beta = \widehat{\boldsymbol{Q}}_{XX}^{-1}\widehat{\boldsymbol{Q}}_{Xe}$ where $\widehat{\boldsymbol{Q}}_{Xe} = \frac{1}{n}\sum_{i=1}^{n} X_i e_i$. The WLLN and the fact that β is the linear projection coefficient imply $\widehat{\boldsymbol{Q}}_{Xe} \xrightarrow[p]{} \mathbb{E}[Xe] = \mathbf{0}$. Therefore $\hat{\beta} - \beta = \widehat{\boldsymbol{Q}}_{XX}^{-1}\widehat{\boldsymbol{Q}}_{Xe} \xrightarrow[p]{} \boldsymbol{Q}_{XX}^{-1} \times \mathbf{0} = \mathbf{0}$.

⚠ The consistency result can be written as $\hat{\beta} = \beta + o_p(1)$

Assumption 1

1. The variables $(Y_i, X_i), i = 1, \dots, n$, are i.i.d.

2. $\mathbb{E}[Y^2] < \infty$.

These two assumptions make sure that Y and X have finite means, variances, and covariances.

3. $\mathbb{E} \parallel X \parallel^2 < \infty$

4. $\boldsymbol{Q}_{XX} = \mathbb{E}[XX']$ is positive definite.

This implies that the columns of $\boldsymbol{Q}_{XX}$ are linearly independent and that it is invertible.

Asymptotic Normality

Assumption 2

1. The variables $(Y_i, X_i), i = 1, \dots, n$, are i.i.d.
2. $\mathbb{E}[Y^4] < \infty$.
3. $\mathbb{E} \parallel X \parallel^4 < \infty$
4. $\boldsymbol{Q}_{XX} = \mathbb{E}[XX']$ is positive definite.

Expectation Inequality: For any random vector $Y \in \mathbb{R}^m$ with $\mathbb{E} \parallel Y \parallel < \infty$ then

$\parallel \mathbb{E}[Y] \parallel \leq \mathbb{E} \parallel Y \parallel$

Matrix norm:

$$\parallel A \parallel_F = \parallel \text{vec}(A) \parallel = (\text{tr}(A'A))^{1/2} = \left(\sum_{i=1}^{m} \sum_{j=1}^{k} a_{ij}^2 \right)^{1/2}$$

For any $m \times 1$ vector $\boldsymbol{a}$, one has $\parallel \boldsymbol{a}\boldsymbol{a}' \parallel_F = \parallel \boldsymbol{a} \parallel^2$

Firstly

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i e_i \right)$$

The product $X_i e_i$ is i.i.d. (since the observations (Y_i, X_i) are i.i.d.) and mean zero (since $\mathbb{E}[Xe] = 0$).

Theorem:
Under Assumption 2, $\Omega < \infty$, and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i e_i \xrightarrow[d]{} N(0, \Omega)$$

as $n \to \infty$.

In order to show that the elements of Ω are finite we only need to show that $\parallel \Omega \parallel < \infty$. By the expectation inequality, the cauchy-schwarz inequality, and Assumption 2

$\parallel \Omega \parallel \leq \mathbb{E}\|XX'e^2\| = \mathbb{E}[\parallel X \parallel^2 e^2] \leq (\mathbb{E} \parallel X \parallel^4)^{\frac{1}{2}} (\mathbb{E}[e^4])^{\frac{1}{2}} < \infty$. Therefore, the finiteness of the covariance matrix means that the CLT applies.

Therefore

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow[d]{} \boldsymbol{Q}_{XX}^{-1} \times N(0, \Omega) = N(0, \boldsymbol{Q}_{XX}^{-1} \Omega \boldsymbol{Q}_{XX}^{-1})$$

Linear combinations of normal vectors are normal

Theorem:
Under Assumption 2, as $n \to \infty$

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow[d]{} N\left(0, \boldsymbol{V}_\beta\right)$$

where $\boldsymbol{Q}_{XX} = \mathbb{E}[XX'], \Omega = \mathbb{E}[XX'e^2]$ and

$\boldsymbol{V}_\beta = \boldsymbol{Q}_{XX}^{-1} \Omega \boldsymbol{Q}_{XX}^{-1}$.

- $\hat{\beta} = \beta + O_p(n^{-1/2})$
- $\boldsymbol{V}_\beta$ is referred as the asymptotic covariance matrix of $\hat{\beta}$.
- Recall the *finite-sample* conditional variance in the CEF model

$$\boldsymbol{V}_{\hat{\beta}} = \text{var}[\hat{\beta} \mid \boldsymbol{X}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

$$n\boldsymbol{V}_{\hat{\beta}} = \left(\frac{1}{n} \boldsymbol{X}'\boldsymbol{X} \right)^{-1} \left(\frac{1}{n} \boldsymbol{X}'\boldsymbol{D}\boldsymbol{X} \right) \left(\frac{1}{n} \boldsymbol{X}'\boldsymbol{X} \right)^{-1}$$

As $n \to \infty, n\boldsymbol{V}_{\hat{\beta}} \xrightarrow[p]{} \boldsymbol{V}_\beta$

$$\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X} = \sum_{i=1}^{n} X_i X_i' \sigma_i^2$$

Heteroskedastic Covariance Matrix Estimation

The moment estimator for Ω is

$$\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \hat{e}_i^2$$

Leading to the plug-in covariance matrix estimator

$$\widehat{\boldsymbol{V}}_{\beta}^{\mathrm{HC0}} = \widehat{\boldsymbol{Q}}_{XX}^{-1}\widehat{\Omega}\widehat{\boldsymbol{Q}}_{XX}^{-1}$$

Remember the HCO covariance matrix estimator:

$$\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC0}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{i=1}^{n} X_i X_i' \hat{e}_i^2\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

! Notice that $\widehat{\boldsymbol{V}}_{\beta}^{\mathrm{HC0}} = n\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC0}}$

Theorem: Under Assumption 2, as $n \to \infty$, $\widehat{\Omega} \xrightarrow[p]{} \Omega$ and $\widehat{\boldsymbol{V}}_{\beta}^{\mathrm{HC0}} \xrightarrow[p]{} \boldsymbol{V}_{\beta}$.

Proof: Observe that

$$\begin{aligned}\widehat{\Omega} &= \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \hat{e}_i^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i X_i' e_i^2 + \frac{1}{n}\sum_{i=1}^{n} X_i X_i'\left(\hat{e}_i^2 - e_i^2\right)\end{aligned}$$

$$\hat{e}_i^2 = e_i^2 - 2e_i X_i'(\hat{\beta} - \beta) + (\hat{\beta} - \beta)' X_i X_i' (\hat{\beta} - \beta)$$

By the WLLN

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' e_i^2 \xrightarrow[p]{} \mathbb{E}\left[XX'e^2\right] = \Omega$$

We simply need to show that $\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\left(\hat{e}_i^2 - e_i^2\right) \xrightarrow[p]{} 0$

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\left(\hat{e}_i^2 - e_i^2\right)\right\| &\le \frac{1}{n}\sum_{i=1}^{n} \| X_i X_i'\left(\hat{e}_i^2 - e_i^2\right)\| \\ &= \frac{1}{n}\sum_{i=1}^{n} \|X_i\|^2 \left|\hat{e}_i^2 - e_i^2\right|\end{aligned}$$

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\left(\hat{e}_i^2 - e_i^2\right)\right\| &\le 2\left(\frac{1}{n}\sum_{i=1}^{n} \|X_i\|^3 |e_i|\right)\| \hat{\beta} - \beta \| + \left(\frac{1}{n}\sum_{i=1}^{n} \|X_i\|^4\right)\| \hat{\beta} - \beta \|^2 \\ &= o_p(1)\end{aligned}$$

Triangle Inequality: $\left|\sum_{j=1}^{m} x_j\right| \le \sum_{j=1}^{m} |x_j|$

Triangle Inequality (matrices): $\| \boldsymbol{A} + \boldsymbol{B} \| \le \| \boldsymbol{A} \| + \| \boldsymbol{B} \|$

Cauchy-Schwarz Inequality: $|\boldsymbol{a}'\boldsymbol{b}| \le \| \boldsymbol{a} \| \| \boldsymbol{b} \|$

Holder Inequality: $\mathbb{E}\|X'Y\| \le (\mathbb{E}\| X \|^p)^{1/p}(\mathbb{E}\| Y \|^q)^{1/q}$ for $p > 1$, $q > 1$, and $\frac{1}{p} + \frac{1}{q} = 1$

Matrix norm:

$$\begin{aligned}\| A \|_F &= \| \mathrm{vec}(A) \| \\ &= (\mathrm{tr}(A'A))^{1/2} \\ &= \left(\sum_{i=1}^{m} \sum_{j=1}^{k} a_{ij}^2\right)^{1/2}\end{aligned}$$

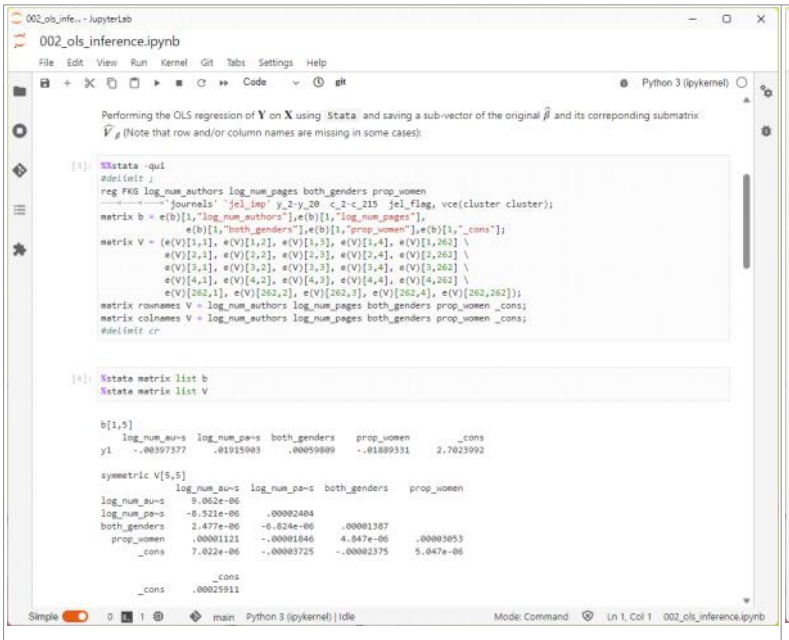
For any $m \times 1$ vector $\boldsymbol{a}$, one has $\|\boldsymbol{a}\boldsymbol{a}'\|_F = \| \boldsymbol{a} \|^2$

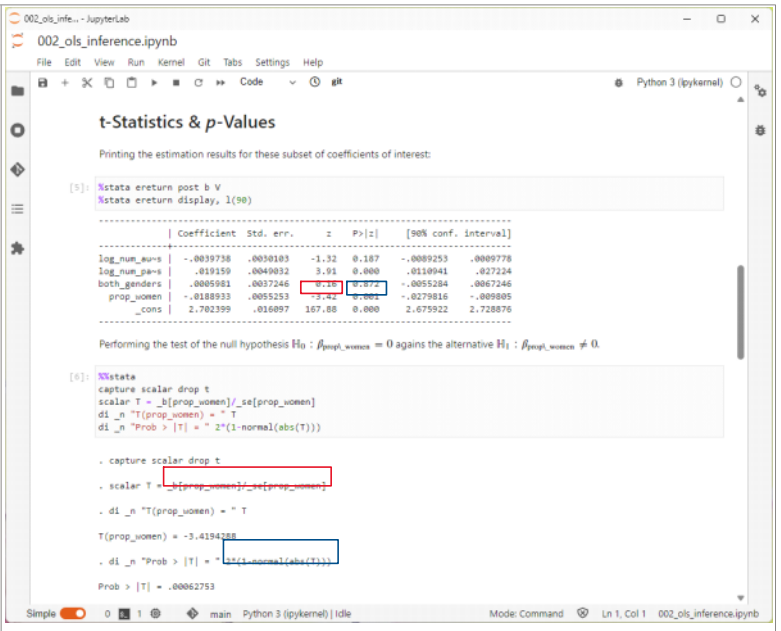
$$\begin{aligned}\left|\hat{e}_i^2 - e_i^2\right| &\le 2\left|e_i X_i'(\hat{\beta} - \beta)\right| + (\hat{\beta} - \beta)' X_i X_i' (\hat{\beta} - \beta) \\ &= 2\left|e_i\right|\left|X_i'(\hat{\beta} - \beta)\right| + \left|(\hat{\beta} - \beta)' X_i\right|^2 \\ &\le 2\left|e_i\right|\|X_i\| \| \hat{\beta} - \beta \| + \|X_i\|^2 \| \hat{\beta} - \beta \|^2\end{aligned}$$

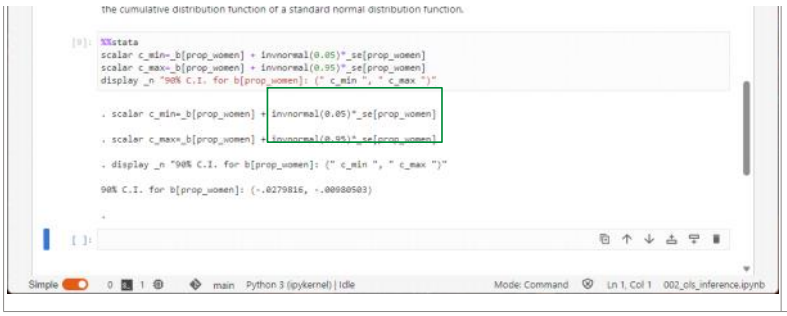
$$\begin{aligned}\mathbb{E}\left[\| X \|^3 |e|\right] &\le \left(\mathbb{E}\left[\left(\| X \|^3\right)^{4/3}\right]\right)^{3/4}\left(\mathbb{E}\left[e^4\right]\right)^{1/4} \\ &= \left(\mathbb{E}\| X \|^4\right)^{3/4}\left(\mathbb{E}\left[e^4\right]\right)^{1/4} < \infty\end{aligned}$$

S.E., t-Statistic, C.I., *p*-values

Standard Error		*t*-Statistics		Confidence Intervals
$s(\hat{\beta}) = \sqrt{n^{-1}\hat{V}_{\hat{\beta}}}$		$T(\beta) = \frac{\hat{\beta} - \beta}{s(\hat{\beta})}$		$\hat{C} = [\hat{\beta} - c \times s(\hat{\beta}), \hat{\beta} + c \times s(\hat{\beta})]$
⚠ Because this is based on *asymptotic* theory we call this *asymptotic* standard errors.		Recall that $\sqrt{n}(\hat{\beta} - \beta) \underset{d}{\to} N\left(0, V_\beta\right)$ and $\hat{V}_{\hat{\beta}} \underset{p}{\to} V_\beta$, thus $T(\beta) = \frac{\hat{\beta} - \beta}{s(\hat{\beta})} = \frac{\sqrt{n}(\hat{\beta} - \beta)}{\sqrt{\hat{V}_\beta}} \underset{d}{\to} \frac{N\left(0, V_\beta\right)}{\sqrt{V_\beta}} = Z \sim N(0,1)$		💡 For $c = 1.965$ we have $\mathbb{P}[\beta \in \hat{C}] \to 0.95$.
		p-value		
		We are interested in testing $\mathbb{H}_0: \beta = 0$ against $\mathbb{H}_1: \beta \neq 0$. Then under $\mathbb{H}_0$ one has $T(0) \equiv \hat{\beta}/s(\hat{\beta}) \underset{d}{\to} N(0,1)$. Then the test taking the form "Reject $\mathbb{H}_0$ if $\lvert T(0)\rvert > c$" has *p*-value defined as $p = 2(1 - \Phi(\lvert T(0)\rvert))$. A test with significance level α can be restated as "Reject $\mathbb{H}_0$ if $p < \alpha$".		






the cumulative distribution function of a standard normal distribution function.
[9]:
%%stata
scalar c_min=_b[prop_women] + invnormal(0.05)*_se[prop_women]
scalar c_max=_b[prop_women] + invnormal(0.95)*_se[prop_women]
display _n "90% C.I. for b[prop_women]: (" c_min ", " c_max ")"
. scalar c_min=_b[prop_women] + invnormal(0.05)*_se[prop_women]
. scalar c_max=_b[prop_women] + invnormal(0.95)*_se[prop_women]
. display _n "90% C.I. for b[prop_women]: (" c_min ", " c_max ")"
90% C.I. for b[prop_women]: (-.0279816, -.00980503)
.
[ ]:
Simple
0
1
main
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 1
002_ols_inference.ipynb

Motivation

(Practical) Implications:

As explained earlier, the information contain in Journals, JEL codes, Years, Cluster, and JEL flag will be included in the model in the form of indicator variables (dummy variables):

Journals: corresponds to 4 indicators for ECM, JPE, QJE, RES

JEL Codes: corresponds to 19 indicators for all JEL codes except 'D'

Years: corresponds to 19 indicators for years 2001, 2002,..., 2019

Cluster: corresponds to 214 indicators for clusters 2, 3,..., 215, except cluster '1'

JEL flag: indicator of whether the article includes a JEL classification

⚠ This implies that we will be estimating 257 parameters multiplying these indicator variables, i.e., 257 *fixed effects*.

Our empirical strategy is based on the following ***baseline*** specification for article a :

$$\log(\text{F} - \text{K grade})_a = \beta_1 \times \log(\text{Number authors})_a + \beta_2 \times \log(\text{Number pages})_a + \beta_3 \times \text{Both genders}_a + \theta \times \text{Proportion Women}_a + \text{cons} + u_a$$

- Some journals (`Journals`) are more likely to publish more technical articles than others, and more technical articles are on average less readable.
- Research in certain fields (`JEL codes`) require large groups of authors.
- Editors change over time (`Years`) and so journal taste for some type of articles over others.
- Articles' readability is affected by co-authors (`Cluster`).
- The method to impute JEL codes (`JEL flag`) use characteristics such as co-author information.

⚠ This implies that

$$\mathbb{E}[u_a | \log(\text{Number authors})_a, \log(\text{Number pages})_a, \text{Both genders}_a, \text{Proportion Women}_a] \neq 0$$

Because Journals, JEL codes, Years, Cluster, and JEL flag are likely correlated with $\log(\text{Number authors})_a$, $\log(\text{Number pages})_a$, Both genders_a, $\text{Proportion Women}_a$, but they have been ommitted from the specification, i.e., ommitted variable bias.

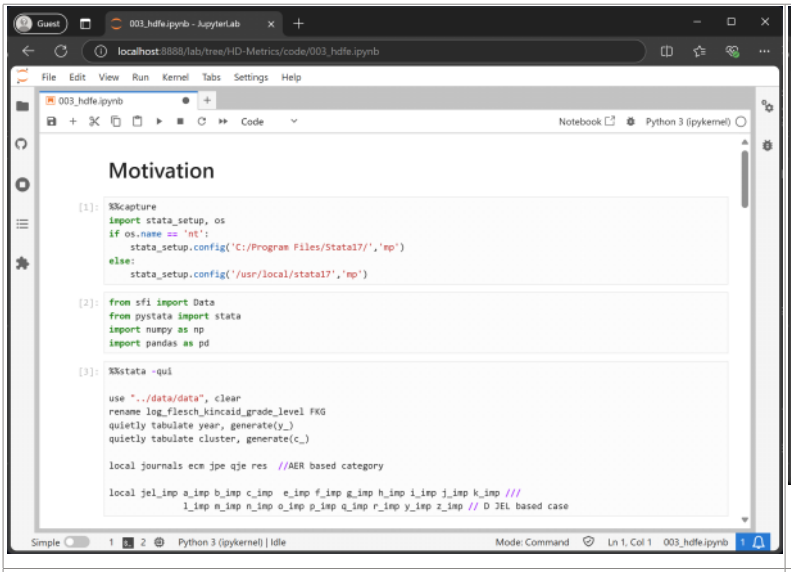
(Theoretical) Solution:

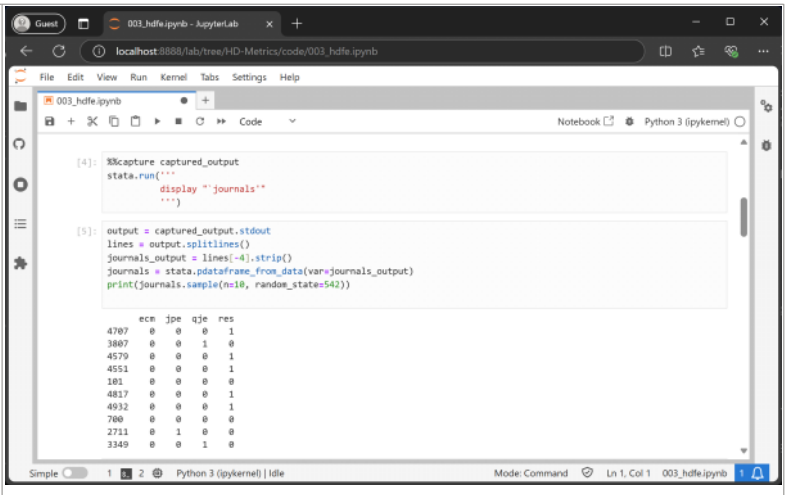
$$\log(\text{F} - \text{K grade})_a = \beta_1 \times \log(\text{Number authors})_a + \beta_2 \times \log(\text{Number pages})_a + \beta_3 \times \text{Both genders}_a + \theta \times \text{Proportion Women}_a + \text{Journals} + \text{JEL codes} + \text{Cluster} + \text{Years} + \text{JEL flag} + \text{cons} + e_a$$

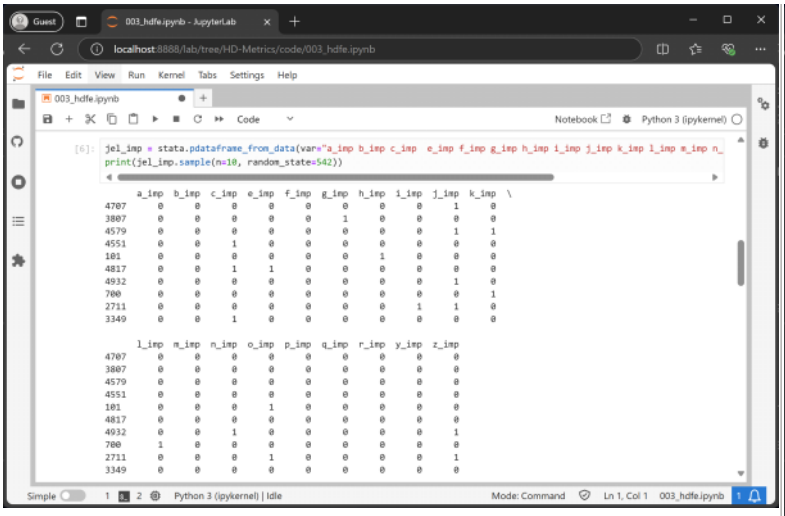
where

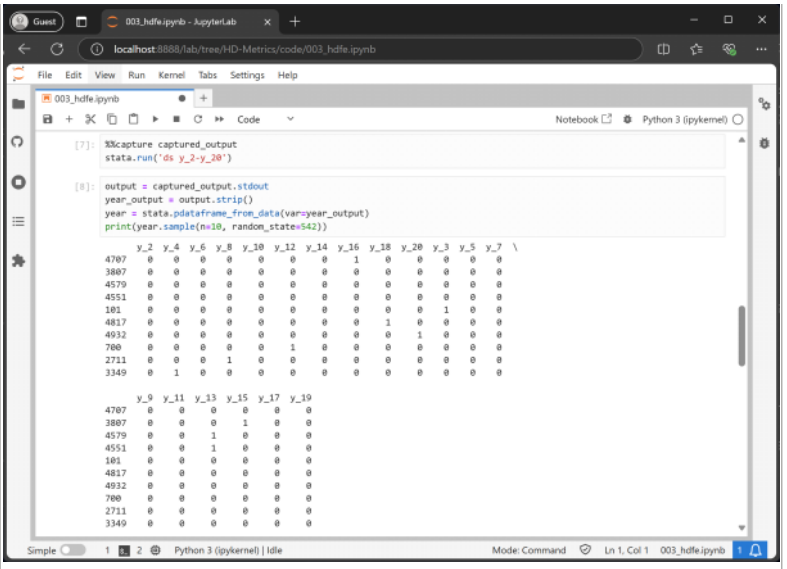
$$\mathbb{E}[e_a | \log(\text{Number authors})_a, \log(\text{Number pages})_a, \text{Both genders}_a, \text{Proportion Women}_a, \text{Journals}, \text{JEL Codes}, \text{Cluster}, \text{Years}, \text{JEL flag}] = 0$$

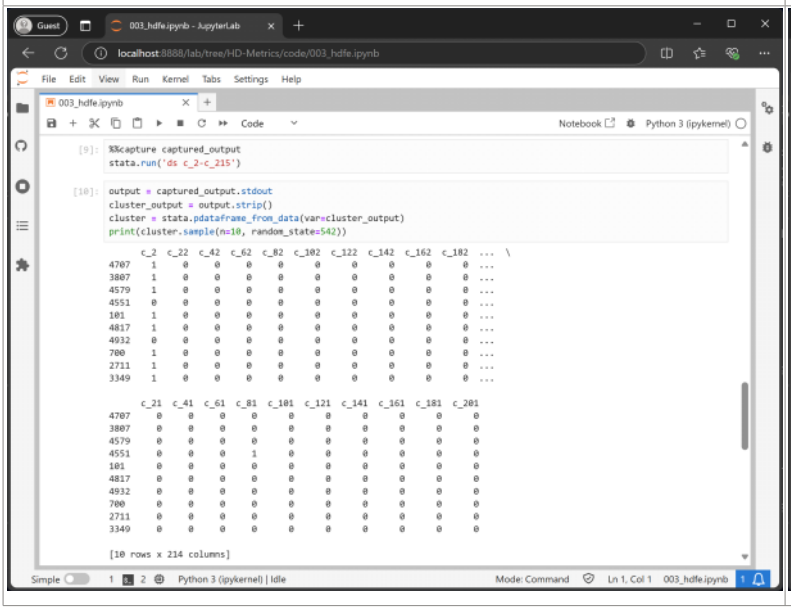
💡 Notice that *all* variables we are conditioning on are allowed to be correlated among themselves.

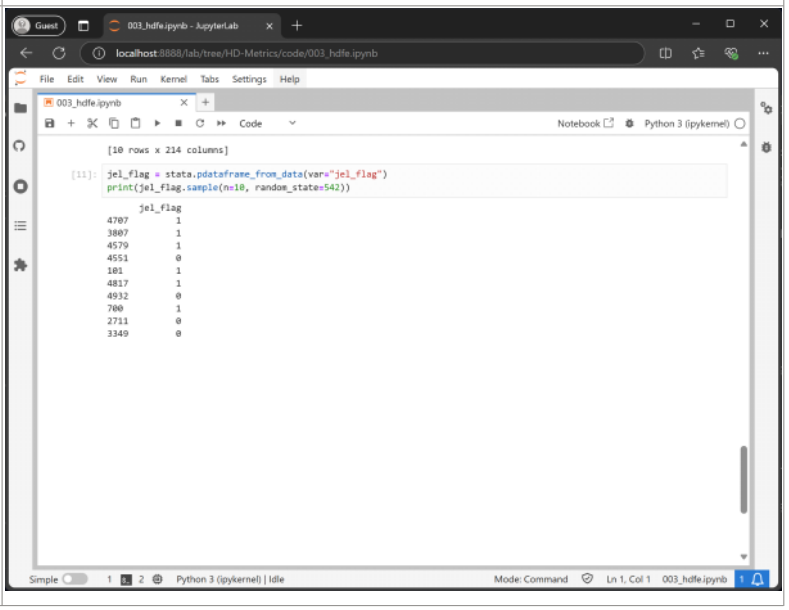





003_hdfe.ipynb - JupyterLab
localhost:8888/lab/tree/HD-Metrics/code/003_hdfe.ipynb
File Edit View Run Kernel Tabs Settings Help
003_hdfe.ipynb
Code
Notebook
Python 3 (ipykernel)
[6]:
jel_imp = stata.pdataframe_from_data(var="a_imp b_imp c_imp e_imp f_imp g_imp h_imp i_imp j_imp k_imp l_imp m_imp n_
print(jel_imp.sample(n=10, random_state=542))
a_imp b_imp c_imp e_imp f_imp g_imp h_imp i_imp j_imp k_imp \
4707 0 0 0 0 0 0 0 0 1 0
3807 0 0 0 0 0 1 0 0 0 0
4579 0 0 0 0 0 0 0 0 1 1
4551 0 0 1 0 0 0 0 0 0 0
101 0 0 0 0 0 0 1 0 0 0
4817 0 0 1 1 0 0 0 0 0 0
4932 0 0 0 0 0 0 0 0 1 0
700 0 0 0 0 0 0 0 0 0 1
2711 0 0 0 0 0 0 1 1 0
3349 0 0 1 0 0 0 0 0 0 0
l_imp m_imp n_imp o_imp p_imp q_imp r_imp y_imp z_imp
4707 0 0 0 0 0 0 0 0 0
3807 0 0 0 0 0 0 0 0 0
4579 0 0 0 0 0 0 0 0 0
4551 0 0 0 0 0 0 0 0 0
101 0 0 0 1 0 0 0 0 0
4817 0 0 0 0 0 0 0 0 0
4932 0 0 1 0 0 0 0 0 1
700 1 0 0 0 0 0 0 0 0
2711 0 0 0 1 0 0 0 0 1
3349 0 0 0 0 0 0 0 0 0
Simple
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 1
003_hdfe.ipynb


003_hdfe.ipynb - JupyterLab
localhost:8888/lab/tree/HD-Metrics/code/003_hdfe.ipynb
File Edit View Run Kernel Tabs Settings Help
003_hdfe.ipynb
Code
Notebook
Python 3 (ipykernel)
[7]:
%%capture captured_output
stata.run('ds y_2-y_20')
[8]:
output = captured_output.stdout
year_output = output.strip()
year = stata.pdataframe_from_data(var=year_output)
print(year.sample(n=10, random_state=542))
y_2 y_4 y_6 y_8 y_10 y_12 y_14 y_16 y_18 y_20 y_3 y_5 y_7 \
4707 0 0 0 0 0 0 0 1 0 0 0 0 0
3807 0 0 0 0 0 0 0 0 0 0 0 0 0
4579 0 0 0 0 0 0 0 0 0 0 0 0 0
4551 0 0 0 0 0 0 0 0 0 0 0 0 0
101 0 0 0 0 0 0 0 0 0 0 0 1 0
4817 0 0 0 0 0 0 0 0 1 0 0 0 0
4932 0 0 0 0 0 0 0 0 0 1 0 0 0
700 0 0 0 0 0 0 1 0 0 0 0 0 0
2711 0 0 0 1 0 0 0 0 0 0 0 0 0
3349 0 1 0 0 0 0 0 0 0 0 0 0 0
y_9 y_11 y_13 y_15 y_17 y_19
4707 0 0 0 0 0 0
3807 0 0 0 1 0 0
4579 0 0 1 0 0 0
4551 0 0 1 0 0 0
101 0 0 0 0 0 0
4817 0 0 0 0 0 0
4932 0 0 0 0 0 0
700 0 0 0 0 0 0
2711 0 0 0 0 0 0
3349 0 0 0 0 0 0
Simple
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 1
003_hdfe.ipynb


003_hdfe.ipynb - JupyterLab
localhost:8888/lab/tree/HD-Metrics/code/003_hdfe.ipynb
File Edit View Run Kernel Tabs Settings Help
003_hdfe.ipynb
Code
Notebook
Python 3 (ipykernel)
[9]:
%%capture captured_output
stata.run('ds c_2-c_215')
[10]:
output = captured_output.stdout
cluster_output = output.strip()
cluster = stata.pdataframe_from_data(var=cluster_output)
print(cluster.sample(n=10, random_state=542))
c_2 c_22 c_42 c_62 c_82 c_102 c_122 c_142 c_162 c_182 ... \
4707 1 0 0 0 0 0 0 0 0 0 ...
3807 1 0 0 0 0 0 0 0 0 0 ...
4579 1 0 0 0 0 0 0 0 0 0 ...
4551 0 0 0 0 0 0 0 0 0 0 ...
101 1 0 0 0 0 0 0 0 0 0 ...
4817 1 0 0 0 0 0 0 0 0 0 ...
4932 0 0 0 0 0 0 0 0 0 0 ...
700 1 0 0 0 0 0 0 0 0 0 ...
2711 1 0 0 0 0 0 0 0 0 0 ...
3349 1 0 0 0 0 0 0 0 0 0 ...
c_21 c_41 c_61 c_81 c_101 c_121 c_141 c_161 c_181 c_201
4707 0 0 0 0 0 0 0 0 0 0
3807 0 0 0 0 0 0 0 0 0 0
4579 0 0 0 0 0 0 0 0 0 0
4551 0 0 0 1 0 0 0 0 0 0
101 0 0 0 0 0 0 0 0 0 0
4817 0 0 0 0 0 0 0 0 0 0
4932 0 0 0 0 0 0 0 0 0 0
700 0 0 0 0 0 0 0 0 0 0
2711 0 0 0 0 0 0 0 0 0 0
3349 0 0 0 0 0 0 0 0 0 0
[10 rows x 214 columns]
Simple
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 1
003_hdfe.ipynb


003_hdfe.ipynb - JupyterLab
localhost:8888/lab/tree/HD-Metrics/code/003_hdfe.ipynb
File Edit View Run Kernel Tabs Settings Help
003_hdfe.ipynb
Code
Notebook
Python 3 (ipykernel)
[10 rows x 214 columns]
[11]:
jel_flag = stata.pdataframe_from_data(var="jel_flag")
print(jel_flag.sample(n=10, random_state=542))
jel_flag
4707 1
3807 1
4579 1
4551 0
101 1
4817 1
4932 0
700 1
2711 0
3349 0
Simple
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 1
003_hdfe.ipynb

High Dimensional Fixed Effects

We want to compute the least squares estimates $\hat{\boldsymbol{\beta}}$ of

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\alpha} + e.$$

- $\mathbf{D} = [\mathbf{D}_1 \quad \mathbf{D}_2 \quad \cdots \quad \mathbf{D}_F]$ consists of F indicator matrices
- If $F = 1$, this collapses to a standard fixed effect regression (`xtreg, areg`)
- Can't use dummies because $[\mathbf{D}_2 \cdots \mathbf{D}_F]$ is *too* large

Solution Strategy:

1. Compute the residuals of $\mathbf{y}$ and $\mathbf{X}$ against $\mathbf{D}$:

$$\tilde{\mathbf{Y}} = \mathbf{M}_\mathbf{D}\mathbf{Y}$$
$$\tilde{\mathbf{X}} = \mathbf{M}_\mathbf{D}\mathbf{X}$$

2. Apply the Frisch-Waugh-Lovell Theorem:

$$\hat{\boldsymbol{\beta}} = \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{Y}}.$$

Thus, we can just focus on one variable at a time: $\hat{\mathbf{Y}}$

To obtain $\hat{\mathbf{Y}} = \mathbf{M}_\mathbf{D}\mathbf{Y}$, find an $\hat{\boldsymbol{\alpha}}$ that satisfies the normal equations

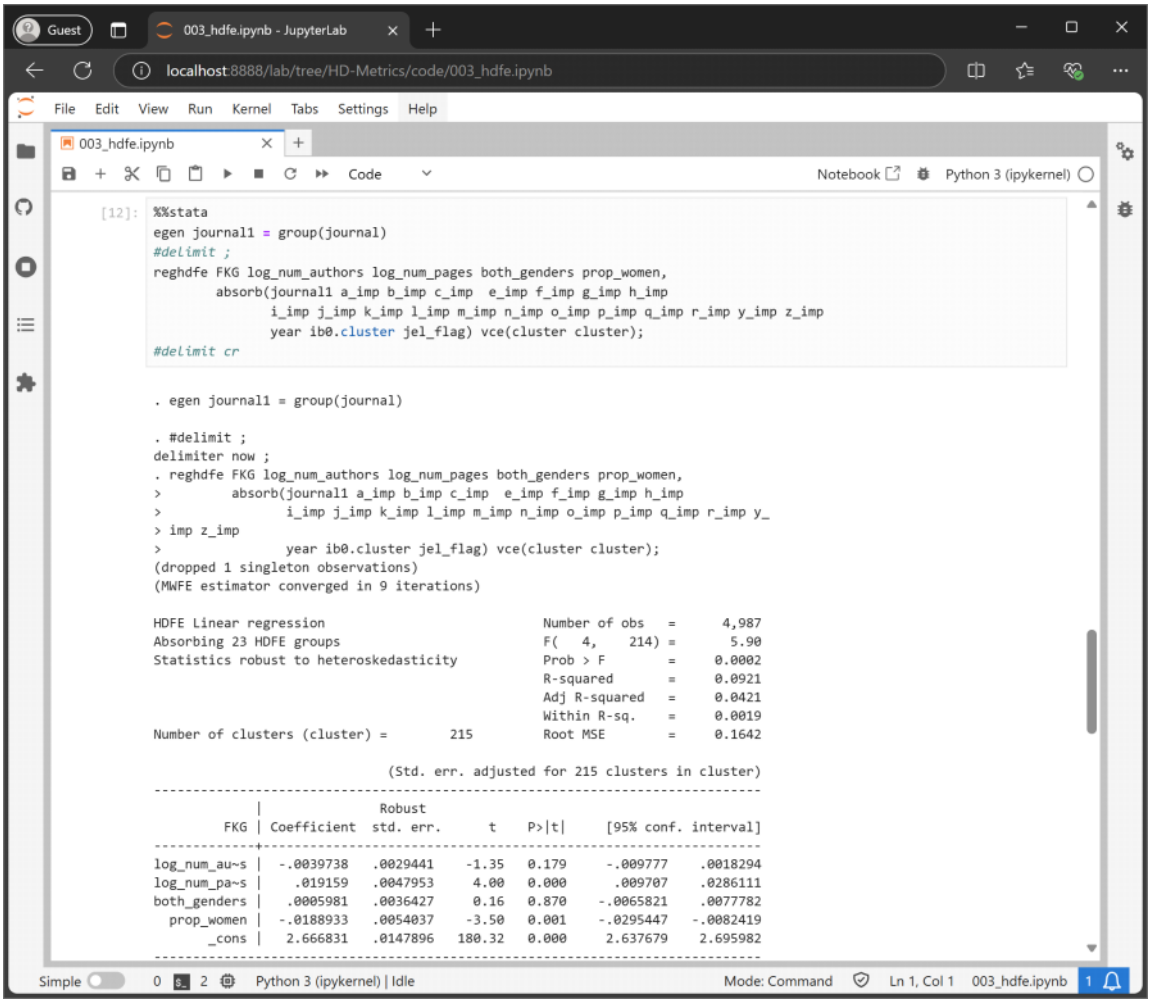
$$\mathbf{D}'\mathbf{e} = 0,\ \mathbf{e} \overset{\text{def}}{=} \mathbf{Y} - \mathbf{D}\hat{\boldsymbol{\alpha}}.$$

In plain English:

For every level g of every fixed effect f the mean of the residuals must be zero:

$$\bar{e}_i = 0\ ,\ i \in \mathcal{J}(f, g).$$

Note: We don't care if $\hat{\boldsymbol{\alpha}}$ is unique.

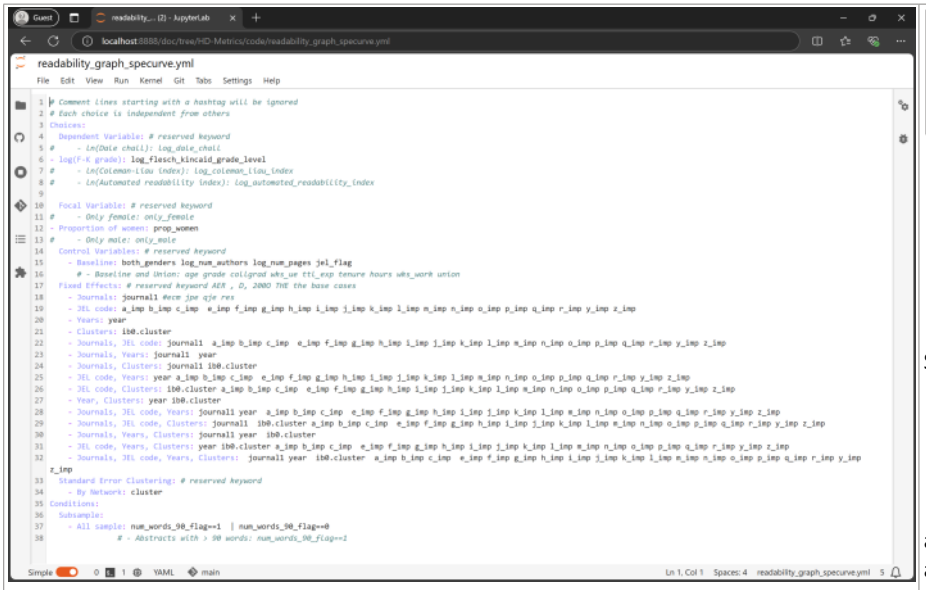


Specification Curve Analysis

Recall that we want to compute the least squares estimates $\hat{\boldsymbol{\beta}}$ of

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\alpha} + e,$$

where $\mathbf{D} = [\mathbf{D}_1 \quad \mathbf{D}_2 \quad \cdots \quad \mathbf{D}_F]$ consists of F indicator matrices.



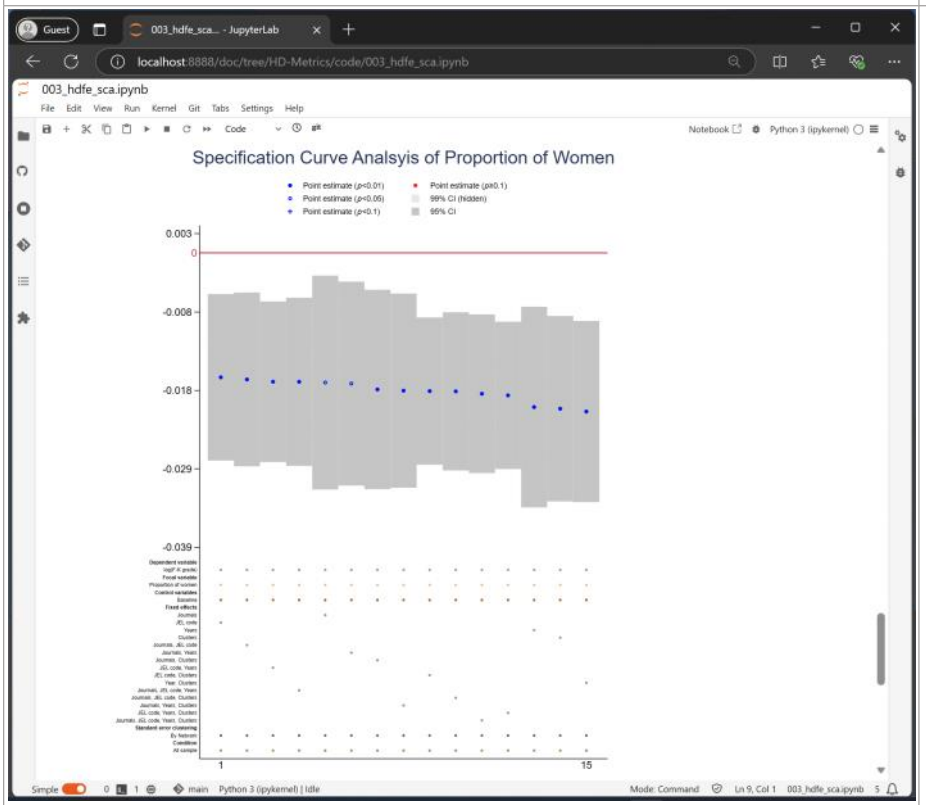
Let's rewrite this model as $\mathbf{Y} = X_1\beta_1 + \mathbf{X_2}\boldsymbol{\beta_2} + \mathbf{D_1}\boldsymbol{\alpha_1} + \cdots + \mathbf{D}_F\boldsymbol{\alpha}_F + e$, and we are only interested in the effect of the *focal* variable, X_1, and how robust this effect is to the inclusion or not of $\mathbf{D_1}$, $\mathbf{D_2}$,....$\mathbf{D}_F$ individually or grouped.

- X_1 is called the focal variable and kept in the model always, e.g., `prop_women`.
- $\mathbf{X_2}$ are called control variables and kept in the model always, e.g., `log_num_authors`, `log_num_pages`, `both_genders`, `jel_flag`, and the `_const`.

Specification Curve Analysis entails setting:

- $\boldsymbol{\alpha_l} = \mathbf{0}$ for $l = 1, \cdots, F$ one at the time.
- $\boldsymbol{\alpha_l} = \mathbf{0}$, $\boldsymbol{\alpha_m} = \mathbf{0}$ for $l, m = 1, \cdots, F; l \neq m$.
- $\boldsymbol{\alpha_l} = \mathbf{0}$, $\boldsymbol{\alpha_m} = \mathbf{0}$, $\boldsymbol{\alpha_n} = \mathbf{0}$ for $l, m, n = 1, \cdots, F; l \neq m \neq n$.
- ...

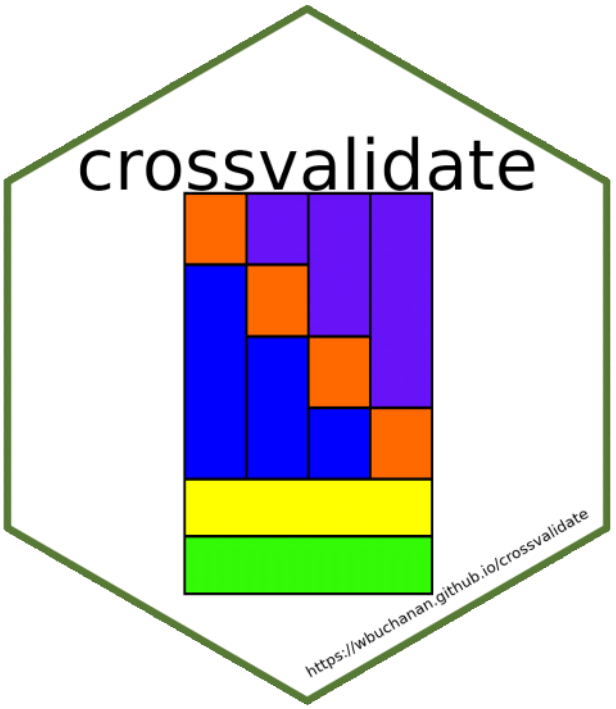
and see how the magnitude and statistical significance of $\hat{\beta}_1$ is affected by these different specifications.



Resampling Methods

There are two most commonly used *resampling methods*:

- ✔ *Cross-validation,*
- ✔ *Bootstrap*.



Source: https://github.com/wbuchanan/crossvalidate
Capabilities: https://411steven.github.io/stataConference2024/

Package is still under development.

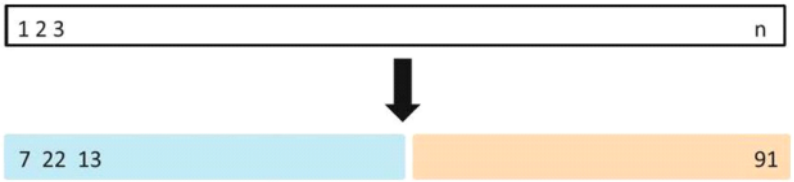
Cross-Validation

These methods are used to do two things

1. *Model Assessment* - the process of evaluating a model's performance.
2. *Model Selection* - the process of selecting the proper level of flexibility for a model.

Since models are 'trained' using training data sets, they are by construction suitable to fit data in these training data sets only (they were fit by minimizing in-sample fitted errors). Since the validation set was not used to fit the model, these set of observations can be used to assess the performance of the model and therefore will allow us to do model selection.

Validation Set Approach



A schematic display of the validation set approach. A set of n observations are randomly split into a training set (shown in blue, containing observations 7, 22, and 13, among others) and a validation set (shown in beige, and containing observation 91, among others).

☞ Usually $m/n \approx 0.8$ or 80% and $(n-m)/n \approx 0.2$ or 20%.

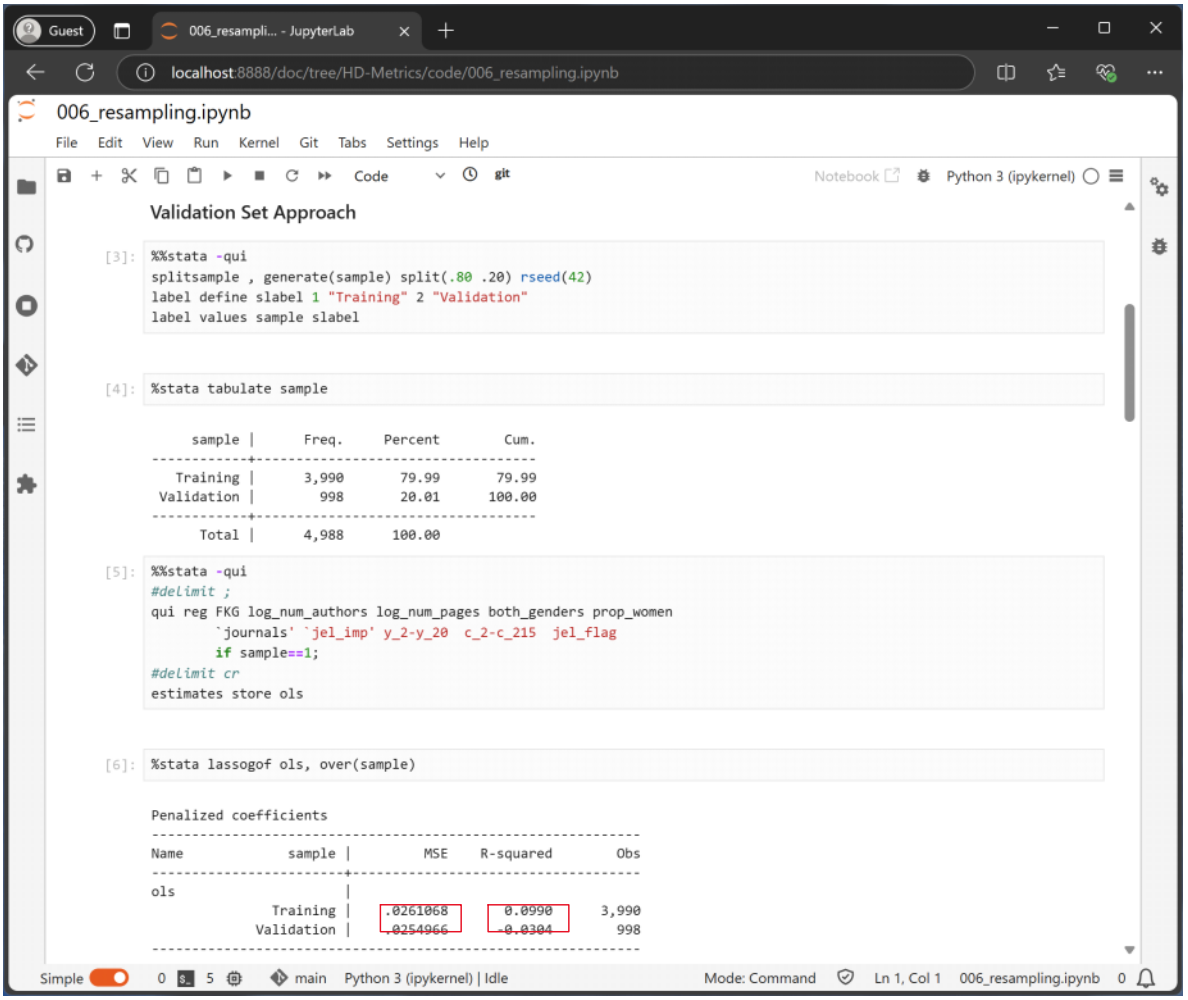
Imagine our sample $\{Y_1, Y_2, Y_3, \dots, Y_m, Y_{m+1}, \dots, Y_n\} = \{7,22,13, \cdots, 91\}$ is such that we also observe a sample of $k \times 1$ vector of realized covariates $\{X_1, X_2, X_3, \dots, X_m, X_{m+1}, \dots, X_n\}$ so that

Training Set:	$\{(Y_i, X_i)\}_{i=1}^{m}$	Using these m observations, calculate the OLS estimator, $\hat{\beta}$, using this sample.
Validation Set:	$\{(Y_j, X_j)\}_{j=m+1}^{n}$	Using these $n-m$ observations, calculate predictions $\{X'_{m+1}\hat{\beta}, \dots, X'_n\hat{\beta}\} = \{\hat{Y}_{m+1}, \dots, \hat{Y}_n\}$

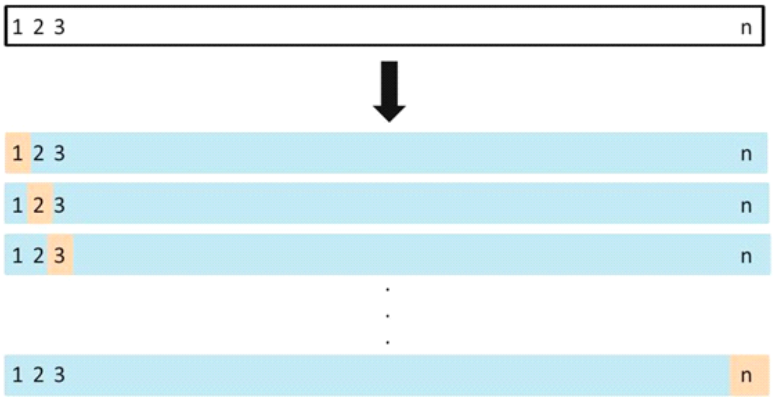
Two common out-of-sample goodness-of-fit measures:

Mean Squared Errors:	$MSE = \frac{1}{n-m}\sum_{j=m+1}^{n}\left(Y_j - \hat{Y}_j\right)^2$
R-squared:	$R^2 = 1 - \sum_{j=m+1}^{n}\left(Y_j - \hat{Y}_j\right)^2 \Big/ \sum_{j=m+1}^{n}\left(Y_j - \bar{Y}\right)^2$

The baseline is the outcome mean of your training data set, i.e., $\bar{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i$


Guest
006_resampli... - JupyterLab
localhost:8888/doc/tree/HD-Metrics/code/006_resampling.ipynb
006_resampling.ipynb
File Edit View Run Kernel Git Tabs Settings Help
Code
Notebook
Python 3 (ipykernel)
Validation Set Approach
[3]: %%stata -qui
splitsample , generate(sample) split(.80 .20) rseed(42)
label define slabel 1 "Training" 2 "Validation"
label values sample slabel
[4]: %stata tabulate sample
sample | Freq. Percent Cum.
Training | 3,990 79.99 79.99
Validation | 998 20.01 100.00
Total | 4,988 100.00
[5]: %%stata -qui
#delimit ;
qui reg FKG log_num_authors log_num_pages both_genders prop_women
`journals' `jel_imp' y_2-y_20 c_2-c_215 jel_flag
if sample==1;
#delimit cr
estimates store ols
[6]: %stata lassogof ols, over(sample)
Penalized coefficients
Name sample | MSE R-squared Obs
ols |
Training | .0261068 0.0990 3,990
Validation | .0254966 -0.0304 998
Simple 0 5 main Python 3 (ipykernel) | Idle
Mode: Command Ln 1, Col 1 006_resampling.ipynb 0

Leave-One-Out Cross-Validation



The statistical learning method is fit on the $n - 1$ training observations, and a prediction $\hat{Y}_1$ is made for the excluded observation, using its value X_1.

☑ Since (X_1, Y_1) was not used in the fitting process, $\text{MSE}_1 = (Y_1 - \hat{Y}_1)^2$ provides an approximately unbiased estimate for the test error.

☑ We can repeat the procedure by selecting (X_2, Y_2) for the validation data, training the statistical learning procedure on the $n - 1$ observations $\{(X_1, Y_1), (X_3, Y_3), \dots, (X_n, Y_n)\}$, and computing $\text{MSE}_2 = (Y_2 - \hat{Y}_2)^2$.

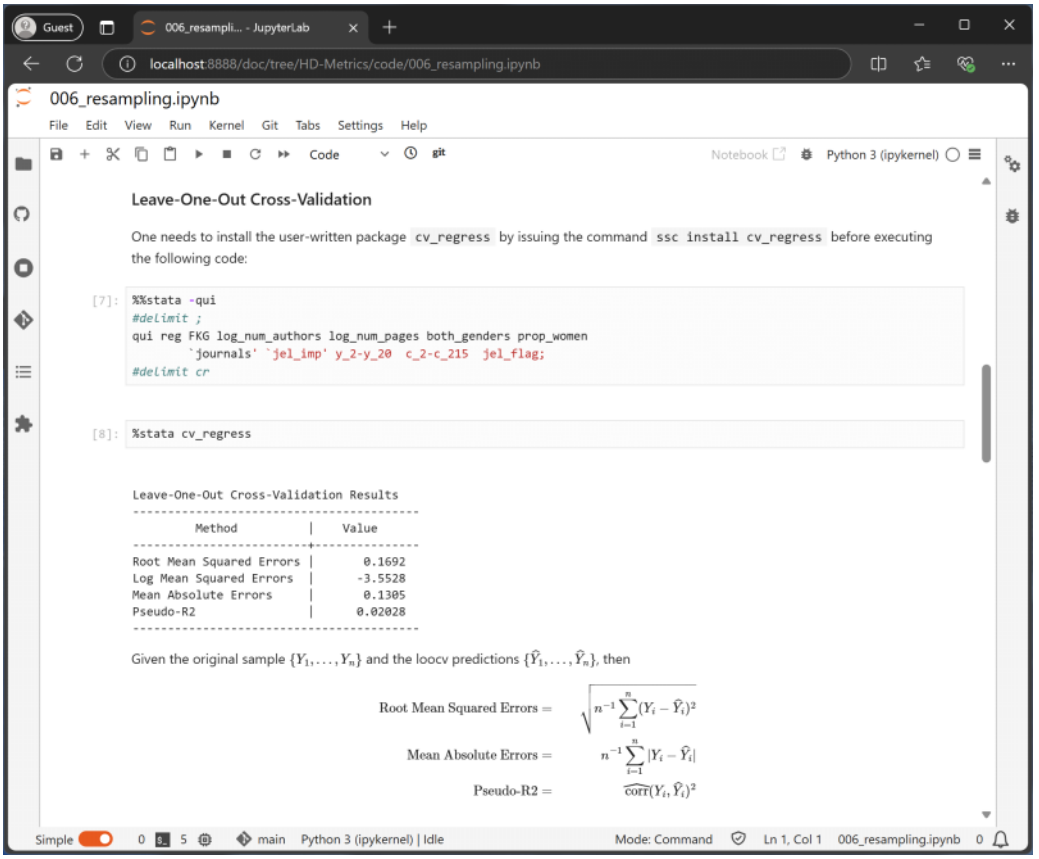
Leave-one-out cross-validation (LOOCV) is closely related to the validation set approach

Like the validation set approach, LOOCV involves splitting the set of observations into two parts.

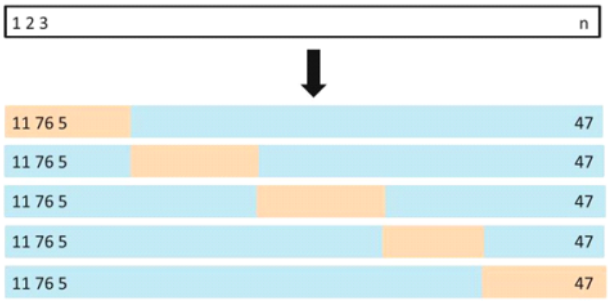
However, instead of creating two subsets of comparable size, a single observation (X_1, Y_1) is used for the validation set, and the remaining observations $\{(X_2, Y_2), \dots, (X_n, Y_n)\}$ make up the training set.

☑ Repeating this approach n times produces n squared errors, $\text{MSE}_1, \dots, \text{MSE}_n$. The LOOCV estimate for the test MSE is the average of these n test error estimates:

$$\text{CV}_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \text{MSE}_i.$$



k-Fold Cross-Validation

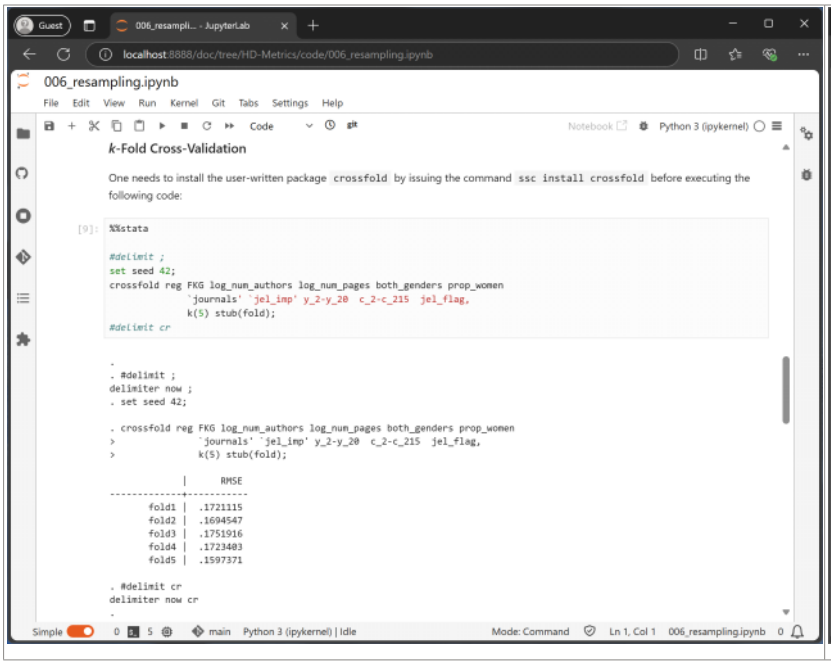


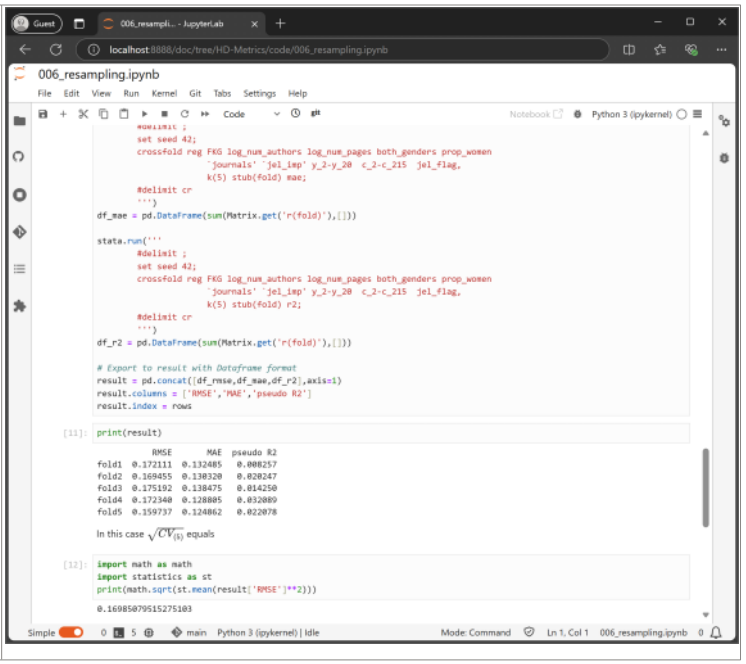
A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.

An alternative to LOOCV is k-fold CV.

This approach involves randomly dividing the set of observations into k groups, or folds, of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining $k - 1$ folds. The mean squared error, MSE M_1, is then computed on the observations in the held-out fold. This procedure is repeated k times; each time, a different group of observations is treated as a validation set. This process results in k estimates of the test error, $\text{MSE}_1, \text{MSE}_2, \ldots, \text{MSE}_k$. The k-fold CV estimate is computed by averaging these values,

$$\text{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \text{MSE}_i.$$





p-norms

We will be making extensive use of the 1-norm and 2-norm, so it is useful to review the definition of the general p-norm.

For a vector $a = (a_1, \dots, a_k)'$ the p-norm ($p \geq 1$) is

$$\| a \|_p = \left(\sum_{j=1}^{k} |a_j|^p \right)^{1/p}.$$

Important special cases include:

1-norm	*2*-norm	*sup*-norm
$\| a \|_1 = \sum_{j=1}^{k} \lvert a_j \rvert.$	$\| a \|_2 = \left(\sum_{j=1}^{k} a_j^2 \right)^{1/2}.$	$\| a \|_\infty = \max_{1 \leq j \leq k} \lvert a_j \rvert.$

We also define the "0-norm"

$$\| a \|_0 = \sum_{j=1}^{k} \mathbb{1}\{a_j \neq 0\}.$$

☞ The number of non-zero elements.

⚠ This is only heuristically labeled as a "norm."

Useful results

✔ The p-norm satisfies the following additivity property. If $a = (a_0, a_1)$ then

$$\| a \|_p^p = \| a_0 \|_p^p + \| a_1 \|_p^p.$$

✔ The Hölder inequality for $1/p + 1/q = 1$ is

$$|a'b| \leq \| a \|_p \| b \|_q.$$

✔ The case $p = 1$ and $q = \infty$ is

$$|a'b| \leq \| a \|_1 \| b \|_\infty.$$

✔ The Minkowski inequality for $p \geq 1$ is

$$\| a + b \|_p \leq \| a \|_p + \| b \|_p.$$

✔ The p-norms for $p \geq 1$ satisfy norm monotonicity, i.e.,

$$\| a \|_1 \geq \| a \|_2 \geq \| a \|_\infty.$$

✔ Applying Hölder's, we also have the inequality

$$\| a \|_1 = \sum_{j=1}^{k} |a_j| \mathbb{1}\{a_j \neq 0\} \leq \| a \|_2 \| a \|_0^{1/2}.$$

Ridge Regression

Hoerl and Kennard (1970) proposed the Ridge Regression estimator

$$\hat{\beta} = \left(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

where $\lambda > 0$ is called the ridge parameter (tuning parameter).

👉 This estimator has the property that it is well-defined and does not suffer from multicollinearity.
👉 This even holds if $p > n$! the ridge regression estimator is well-defined even when the number of regressors exceeds the sample size.

Why it works?
use the spectral decomposition to write $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{H}'\boldsymbol{D}\boldsymbol{H}$ where $\boldsymbol{H}$ is orthonormal and $\boldsymbol{D} = \text{diag}\{r_1, \dots, r_p\}$ is a diagonal matrix with the eigenvalues r_j of $\boldsymbol{X}'\boldsymbol{X}$ on the diagonal. Set $\Lambda = \lambda \boldsymbol{I}_p$. We can write

$$\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_p = \boldsymbol{H}'\boldsymbol{D}\boldsymbol{H} + \lambda \boldsymbol{H}'\boldsymbol{H} = \boldsymbol{H}'(\boldsymbol{D} + \Lambda)\boldsymbol{H},$$

which has strictly positive eigenvalues $r_j + \lambda > 0$. Thus all eigenvalues are bounded away from zero so $\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_p$ is full rank and well-conditioned.

Definition

Definition: Penalized Least Squares

Consider the sum of squared errors penalized by the squared 2-norm of the coefficient vector

$SSE_2(b, \lambda) = (\boldsymbol{Y} - \boldsymbol{X}b)'(\boldsymbol{Y} - \boldsymbol{X}b) + \lambda b'b = \| \boldsymbol{Y} - \boldsymbol{X}b \|_2^2 + \lambda \| b \|_2^2.$

The minimizer of $SSE_2(b, \lambda)$ is a regularized least squares estimator.

The first order condition for minimization of $SSE_2(b, \lambda)$ over β is

$$-2\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}b) + 2\lambda b = 0.$$

The solution is $\hat{\beta}_{\text{ridge}}$.

You specify the ridge parameter λ

Definition: Constrained Minimization

This problem is

$$\min_{b'b \le \tau} (\boldsymbol{Y} - \boldsymbol{X}b)'(\boldsymbol{Y} - \boldsymbol{X}b).$$

for some $\tau > 0$. To see the connection, the Lagrangian for the constrained problem is

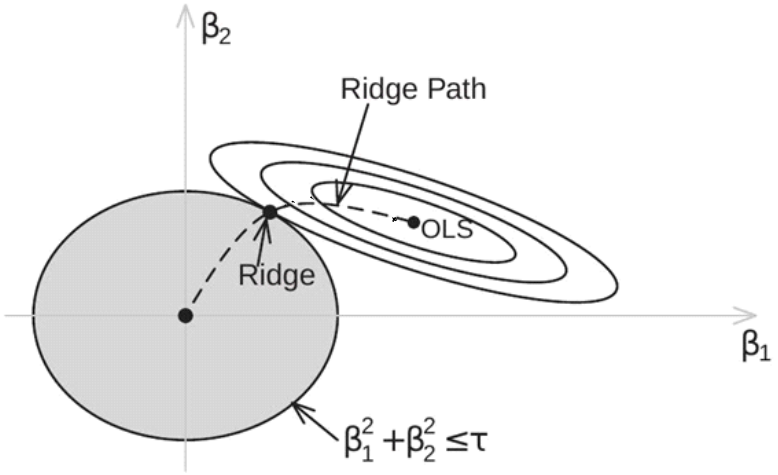
$$\min_b (\boldsymbol{Y} - \boldsymbol{X}b)'(\boldsymbol{Y} - \boldsymbol{X}b) + \lambda(b'b - \tau),$$

where λ is a Lagrange multiplier. The first order condition over b is

$-2\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}b) + 2\lambda b = 0.$

The solution is $\hat{\beta}_{\text{ridge}}$.

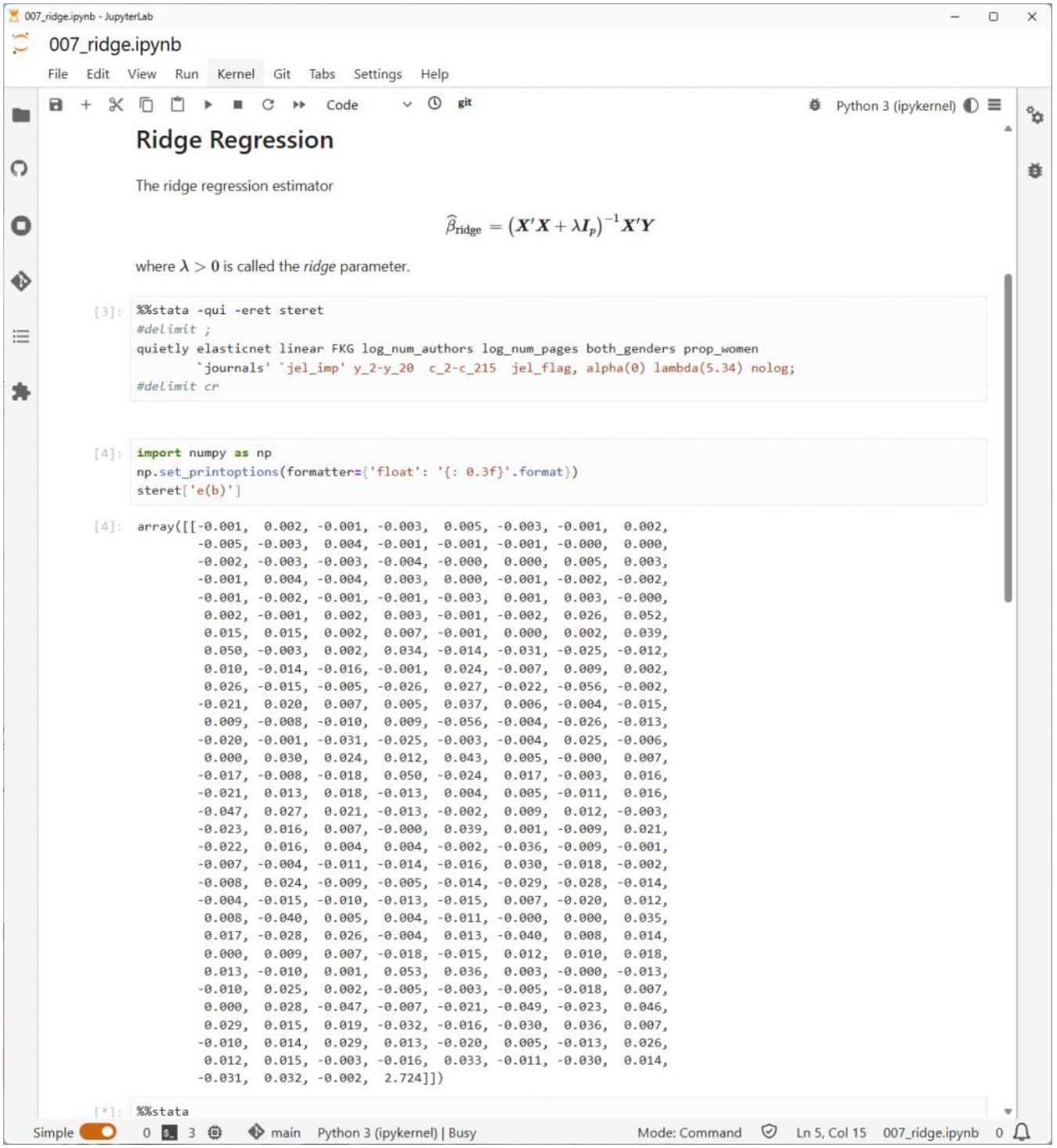
You specify the parameter τ



They are connected because the values of λ and τ satisfy the relationship $\hat{\beta}_{\text{ridge}}{}'\hat{\beta}_{\text{ridge}} - \tau = 0$

$$\boldsymbol{Y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}'\boldsymbol{Y} = \tau.$$

To find λ given τ it is sufficient to (numerically) solve this equation.


007_ridge.ipynb - JupyterLab
007_ridge.ipynb
File Edit View Run Kernel Git Tabs Settings Help
Code
git
Python 3 (ipykernel)
Ridge Regression
The ridge regression estimator
$$\widehat{\beta}_{\text{ridge}} = \left(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}$$
where $\boldsymbol{\lambda} > \mathbf{0}$ is called the *ridge* parameter.
[3]:
%%stata -qui -eret steret
#delimit ;
quietly elasticnet linear FKG log_num_authors log_num_pages both_genders prop_women
`journals' `jel_imp' y_2-y_20 c_2-c_215 jel_flag, alpha(0) lambda(5.34) nolog;
#delimit cr
[4]:
import numpy as np
np.set_printoptions(formatter={'float': '{: 0.3f}'.format})
steret['e(b)']
[4]:
array([[-0.001, 0.002, -0.001, -0.003, 0.005, -0.003, -0.001, 0.002,
-0.005, -0.003, 0.004, -0.001, -0.001, -0.001, -0.000, 0.000,
-0.002, -0.003, -0.003, -0.004, -0.000, 0.000, 0.005, 0.003,
-0.001, 0.004, -0.004, 0.003, 0.000, -0.001, -0.002, -0.002,
-0.001, -0.002, -0.001, -0.001, -0.003, 0.001, 0.003, -0.000,
0.002, -0.001, 0.002, 0.003, -0.001, -0.002, 0.026, 0.052,
0.015, 0.015, 0.002, 0.007, -0.001, 0.000, 0.002, 0.039,
0.050, -0.003, 0.002, 0.034, -0.014, -0.031, -0.025, -0.012,
0.010, -0.014, -0.016, -0.001, 0.024, -0.007, 0.009, 0.002,
0.026, -0.015, -0.005, -0.026, 0.027, -0.022, -0.056, -0.002,
-0.021, 0.020, 0.007, 0.005, 0.037, 0.006, -0.004, -0.015,
0.009, -0.008, -0.010, 0.009, -0.056, -0.004, -0.026, -0.013,
-0.020, -0.001, -0.031, -0.025, -0.003, -0.004, 0.025, -0.006,
0.000, 0.030, 0.024, 0.012, 0.043, 0.005, -0.000, 0.007,
-0.017, -0.008, -0.018, 0.050, -0.024, 0.017, -0.003, 0.016,
-0.021, 0.013, 0.018, -0.013, 0.004, 0.005, -0.011, 0.016,
-0.047, 0.027, 0.021, -0.013, -0.002, 0.009, 0.012, -0.003,
-0.023, 0.016, 0.007, -0.000, 0.039, 0.001, -0.009, 0.021,
-0.022, 0.016, 0.004, 0.004, -0.002, -0.036, -0.009, -0.001,
-0.007, -0.004, -0.011, -0.014, -0.016, 0.030, -0.018, -0.002,
-0.008, 0.024, -0.009, -0.005, -0.014, -0.029, -0.028, -0.014,
-0.004, -0.015, -0.010, -0.013, -0.015, 0.007, -0.020, 0.012,
0.008, -0.040, 0.005, 0.004, -0.011, -0.000, 0.000, 0.035,
0.017, -0.028, 0.026, -0.004, 0.013, -0.040, 0.008, 0.014,
0.000, 0.009, 0.007, -0.018, -0.015, 0.012, 0.010, 0.018,
0.013, -0.010, 0.001, 0.053, 0.036, 0.003, -0.000, -0.013,
-0.010, 0.025, 0.002, -0.005, -0.003, -0.005, -0.018, 0.007,
0.000, 0.028, -0.047, -0.007, -0.021, -0.049, -0.023, 0.046,
0.029, 0.015, 0.019, -0.032, -0.016, -0.030, 0.036, 0.007,
-0.010, 0.014, 0.029, 0.013, -0.020, 0.005, -0.013, 0.026,
0.012, 0.015, -0.003, -0.016, 0.033, -0.011, -0.030, 0.014,
-0.031, 0.032, -0.002, 2.724]])
[*]:
%%stata
Simple
0 3
main
Python 3 (ipykernel) | Busy
Mode: Command
Ln 5, Col 15
007_ridge.ipynb
0

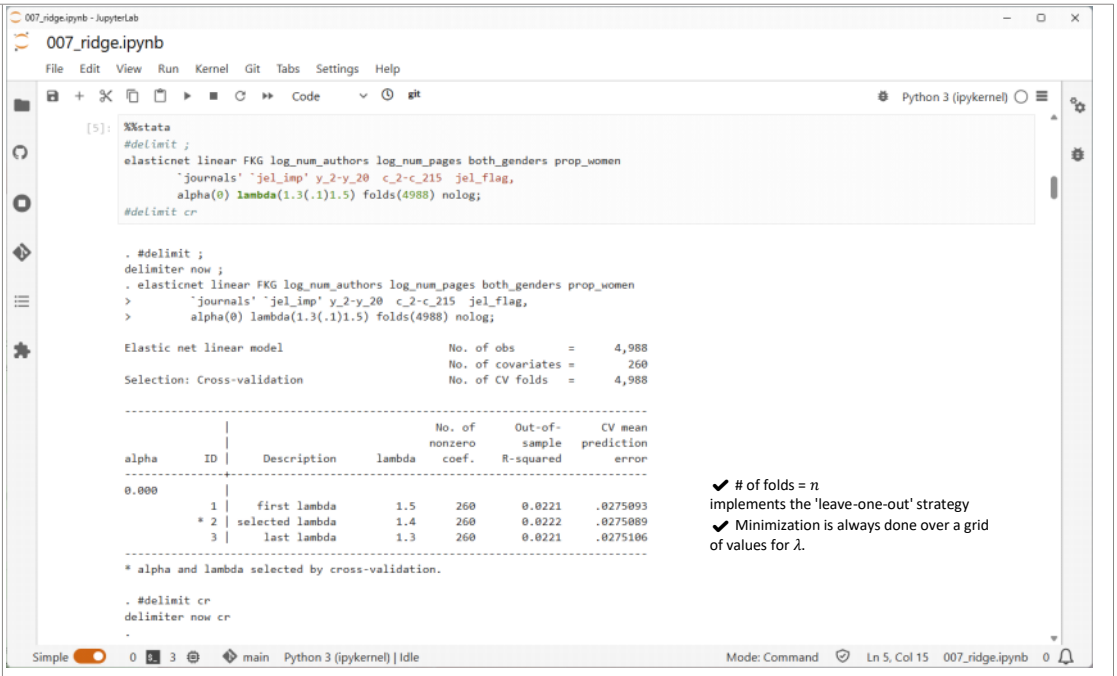
Cross Validation

The most popular method to select the ridge parameter λ is cross validation. The leave-one-out ridge regression estimator, prediction errors, and CV criterion are

$$\begin{aligned}\hat{\beta}_{-i}(\lambda) &= \left(\sum_{j\neq i} X_j X_j' + \Lambda\right)^{-1}\left(\sum_{j\neq i} X_j Y_i\right)\\ \tilde{e}_i(\lambda) &= Y_i - X_i'\hat{\beta}_{-i}(\lambda)\\ \mathrm{CV}(\lambda) &= \sum_{i=1}^{n} \tilde{e}_i(\lambda)^2\end{aligned}.$$

The CV-selected ridge parameter λ_{cv} minimizes $\mathrm{CV}(\lambda)$.

The cross-validation ridge estimator is calculated using λ_{cv}.



✔ # of folds = n implements the 'leave-one-out' strategy

✔ Minimization is always done over a grid of values for λ.

Statistical Properties

Take the linear regression model

$$\begin{array}{ll} Y & = X'\beta + e \\ \mathbb{E}[e \mid X] & = 0 \end{array}.$$

The bias of the ridge estimator with fixed λ is

$$\text{bias}\left[\hat{\beta}_{\text{ridge}} \mid \boldsymbol{X}\right] = -\lambda\left(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1}\beta .$$

Under random sampling its covariance matrix is

$$\text{var}\left[\hat{\beta}_{\text{ridge}} \mid \boldsymbol{X}\right] = \left(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1}.$$

where $\boldsymbol{D} = \text{diag}\{\sigma^2(X_1), \ldots, \sigma^2(X_n)\}$ and $\sigma^2(X) = \mathbb{E}[e^2 \mid X]$.

We can measure estimation efficiency by the mean squared error (MSE) matrix

$$\text{mse}[\hat{\beta}_{\text{ridge}} \mid \boldsymbol{X}] = \mathbb{E}\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)' \mid \boldsymbol{X}\right].$$

Define $\underline{\sigma}^2 = \min_{x \in \mathcal{X}} \sigma^2(x)$ where $\mathcal{X}$ is the support of X.

Theorem

In the linear regression model, if $0 < \lambda < 2\underline{\sigma}^2/\beta'\beta$,

$$\text{mse}\left[\hat{\beta}_{\text{ridge}} \mid \boldsymbol{X}\right] < \text{mse}\left[\hat{\beta} \mid \boldsymbol{X}\right].$$

- Theorem shows that the ridge estimator dominates the least squares estimator, if λ satisfies a specific range of values. This holds regardless of the dimension of β.
- Since the upper bound $2\underline{\sigma}^2/\beta'\beta$ is unknown, however, it is unclear if feasible ridge regression dominates least squares.
- The upper bound does not give practical guidance for selection of λ.

Variance Estimation

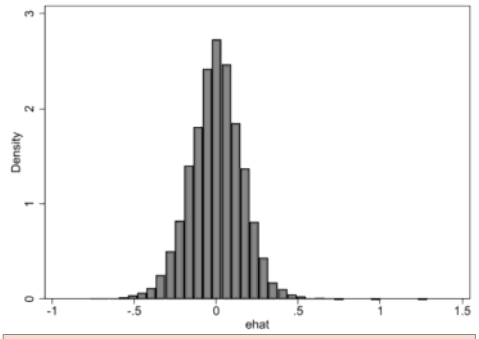
It is straightforward to construct estimators of $V_{\hat{\beta}} = \text{var}\left[\hat{\beta} \mid \boldsymbol{X}\right]$. I suggest the HC3

$$\tilde{V}_{\hat{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1}\left(\sum_{i=1}^{n} X_i X_i' \tilde{e}_i(\lambda)^2\right)\left(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1}$$

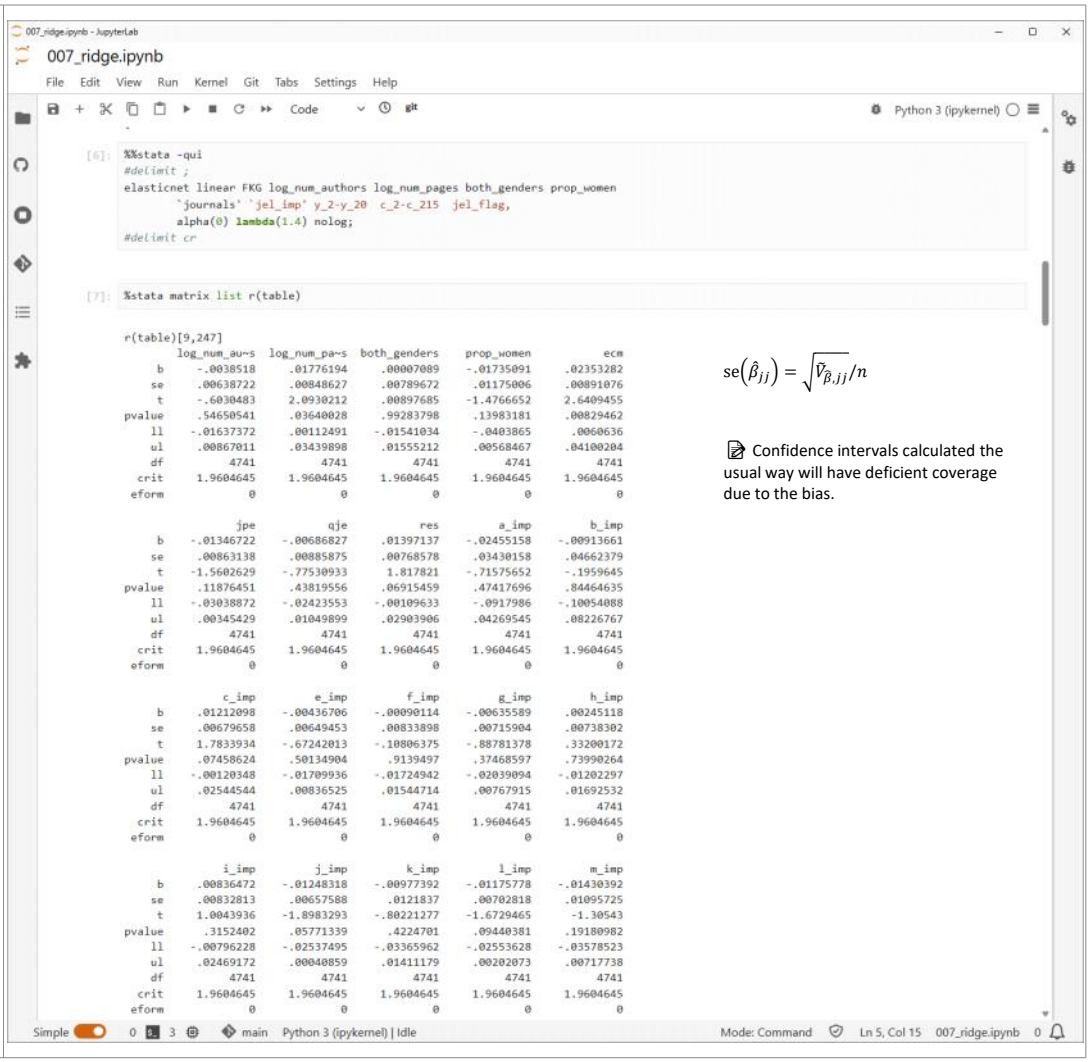
where $\tilde{e}_i(\lambda)$ are the
leave-one-out ridge regression prediction errors

$$\tilde{e}_i(\lambda) = \left(1 - X_i'\left(\boldsymbol{X}'\boldsymbol{X} + \Lambda\right)^{-1}X_i\right)^{-1}\hat{e}_i(\lambda).$$

where $\hat{e}_i(\lambda) = Y_i - X_i'\hat{\beta}_{\text{ridge}}(\lambda)$ are the ridge regression residuals.



⚠ The estimators and confidence intervals are valid for the pseudo-true projections, e.g. $\beta^* = \left(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\beta$, not the coefficients β themselves.



$$\text{se}\left(\hat{\beta}_{jj}\right) = \sqrt{\tilde{V}_{\hat{\beta},jj}/n}$$

Confidence intervals calculated the usual way will have deficient coverage due to the bias.

Lasso

Tibshirani (1996) proposed the Lasso estimator as the minimizer

$$\hat{\beta} = \underset{b}{\operatorname{argmin}}\ \mathrm{SSE}_1(b, \lambda).$$

where the least squares criterion with a 1-norm penalty is

$$\mathrm{SSE}_1(b, \lambda) = (\boldsymbol{Y} - \boldsymbol{X}b)'(\boldsymbol{Y} - \boldsymbol{X}b) + \lambda \sum_{j=1}^{p} \left|b_j\right| = \| \boldsymbol{Y} - \boldsymbol{X}b \|_2^2 + \lambda \| b \|_1.$$

👉 An important property is that when $\lambda > 0$ the Lasso estimator is well defined even if $p > n$.

Definition

Definition: Constrained Minimization

The minimization problem

$$\hat{\beta} = \underset{\|b\|_1 \le \tau}{\operatorname{argmin}} \, \mathrm{SSE}_1(b).$$

To see that the two problems are the same observe that the constrained minimization problem has the Lagrangian

$$\min_b (\boldsymbol{Y} - \boldsymbol{X}b)'(\boldsymbol{Y} - \boldsymbol{X}b) + \lambda \left(\sum_{j=1}^{p} |b_j| - \tau \right),$$

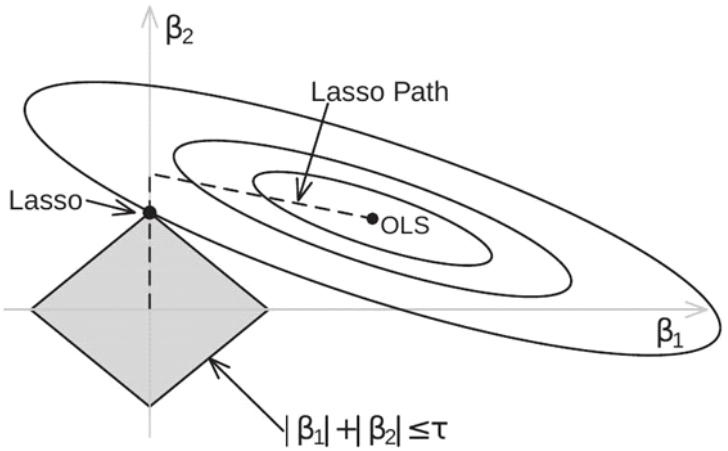
which has 'first order' conditions

$$-2\boldsymbol{X}_j'(\boldsymbol{Y} - \boldsymbol{X}b) + \lambda \mathrm{sgn}(b_j) = 0.$$

This is the same as those for minimization of the penalized criterion. Thus the solutions are identical.

The sign function of a real number x is a piecewise function which is defined as follows:

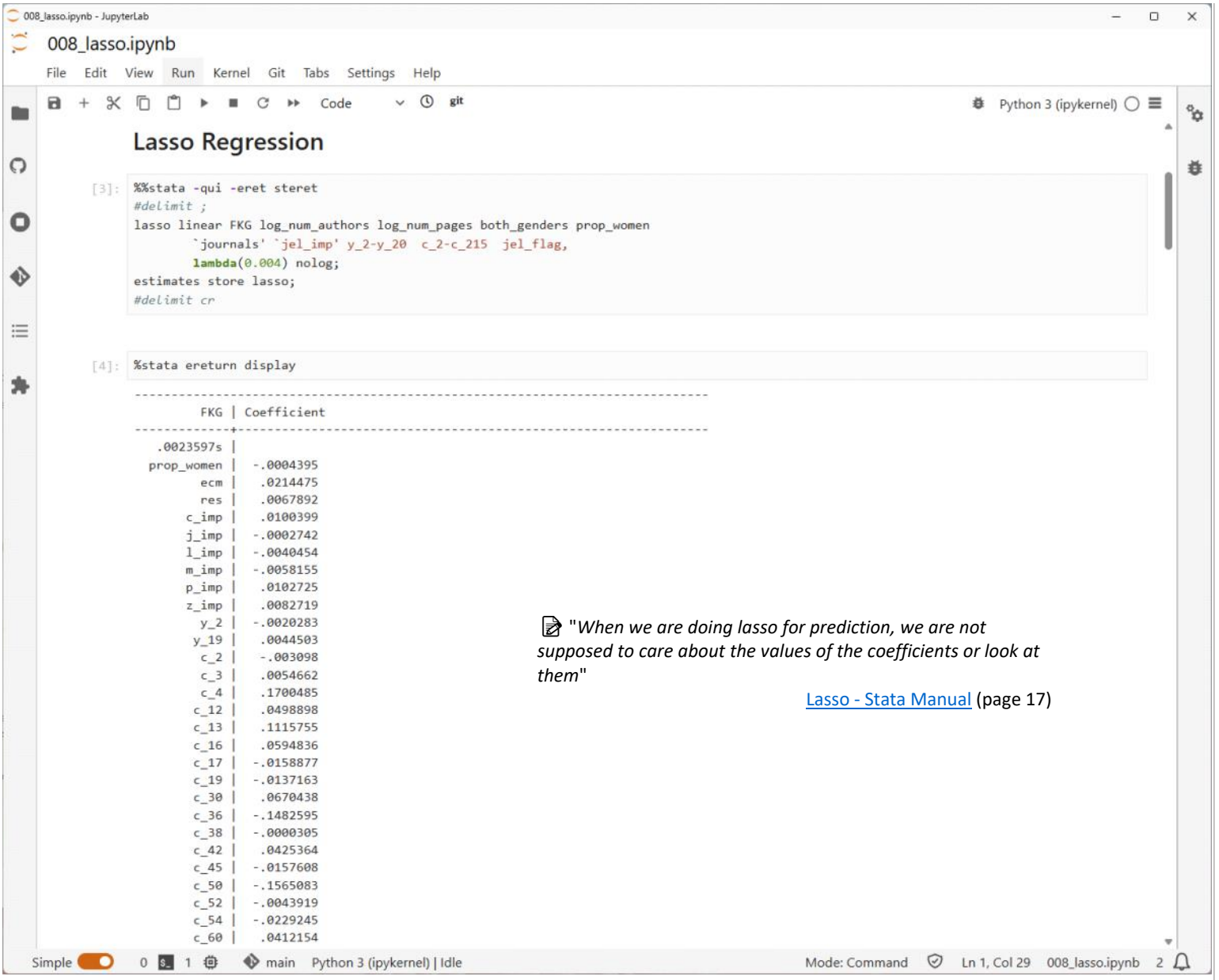
$$\operatorname{sgn} x := \begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0. \\ 1 & \text{if } x > 0 \end{cases}$$



The Lasso path is drawn with the dashed line. This is the sequence of solutions obtained as the constraint set is varied. The solution path has the property that it is a straight line from the least squares estimator to the y-axis (in this example), at which point β_2 is set to zero, and then the solution path follows the y-axis to the origin. In general, the solution path is linear on segments until a coefficient hits zero, at which point that coefficient is eliminated. In this particular example the solution path shows β_2 increasing while β_1 decreases. Thus while Lasso is a shrinkage estimator it does not shrink individual coefficients monotonically.

The Lasso is not invariant to the scaling of the regressors. If you rescale a regressor then the penalty has a different meaning. Consequently, it is important to scale the regressors appropriately before applying Lasso. It is conventional to scale all the variables to have mean zero and unit variance.

Lasso is also not invariant to rotations of the regressors. For example, Lasso on $(\boldsymbol{X}_1, \boldsymbol{X}_2)$ is not the same as Lasso on $(\boldsymbol{X}_1 - \boldsymbol{X}_2, \boldsymbol{X}_2)$ despite having identical least squares solutions. This is troubling as typically there is no default specification.



"*When we are doing lasso for prediction, we are not supposed to care about the values of the coefficients or look at them*"

Lasso - Stata Manual (page 17)

Group Lasso

Take the model

$$Y = X_1'\beta_1 + \cdots + X_G'\beta_G + e,$$

and minimize the SSE$\equiv \| \boldsymbol{Y} - \boldsymbol{X}b \|_2^2$ subject to the penalty

$$\lambda_1\sqrt{b_1'b_1} + \cdots + \lambda_G\sqrt{b_G'b_G}.$$

The penalized criterion function is

$$\text{SSE}_3(b, \lambda_1, \ldots, \lambda_G) = (\boldsymbol{Y} - \boldsymbol{X}b)'(\boldsymbol{Y} - \boldsymbol{X}b) + \sum_{g=1}^{G} \lambda_g \sqrt{b_g'b_g} = \| \boldsymbol{Y} - \boldsymbol{X}b \|_2^2 + \sum_{g=1}^{G} \lambda_g \| b_g \|_2.$$

- One important special case is $\lambda_1 = 0$, thus one group of coefficients are not penalized.
- With $G = 2$ this partitions the coefficients into two groups: penalized and non-penalized.

Relation with the Lasso

Set $\lambda_1 = \cdots = \lambda_G \equiv \lambda$, then

$$\text{SSE}_3(b, \lambda) = \| \boldsymbol{Y} - \boldsymbol{X}b \|_2^2 + \lambda \sum_{g=1}^{G} \| b_g \|_2.$$

When $G = p$, one has that the Group Lasso becomes Lasso.

Imagine we 'required' that X_1 must be in the model, but wish to perform model selection among the remaining regressors $X_2, \ldots, X_p$, then we would minimize

$$(\boldsymbol{Y} - \boldsymbol{X}b)'(\boldsymbol{Y} - \boldsymbol{X}b) + \lambda \sum_{j=2}^{p} |b_j| = \| \boldsymbol{Y} - \boldsymbol{X}b \|_2^2 + \lambda \sum_{j=2}^{p} \| b_j \|_2.$$

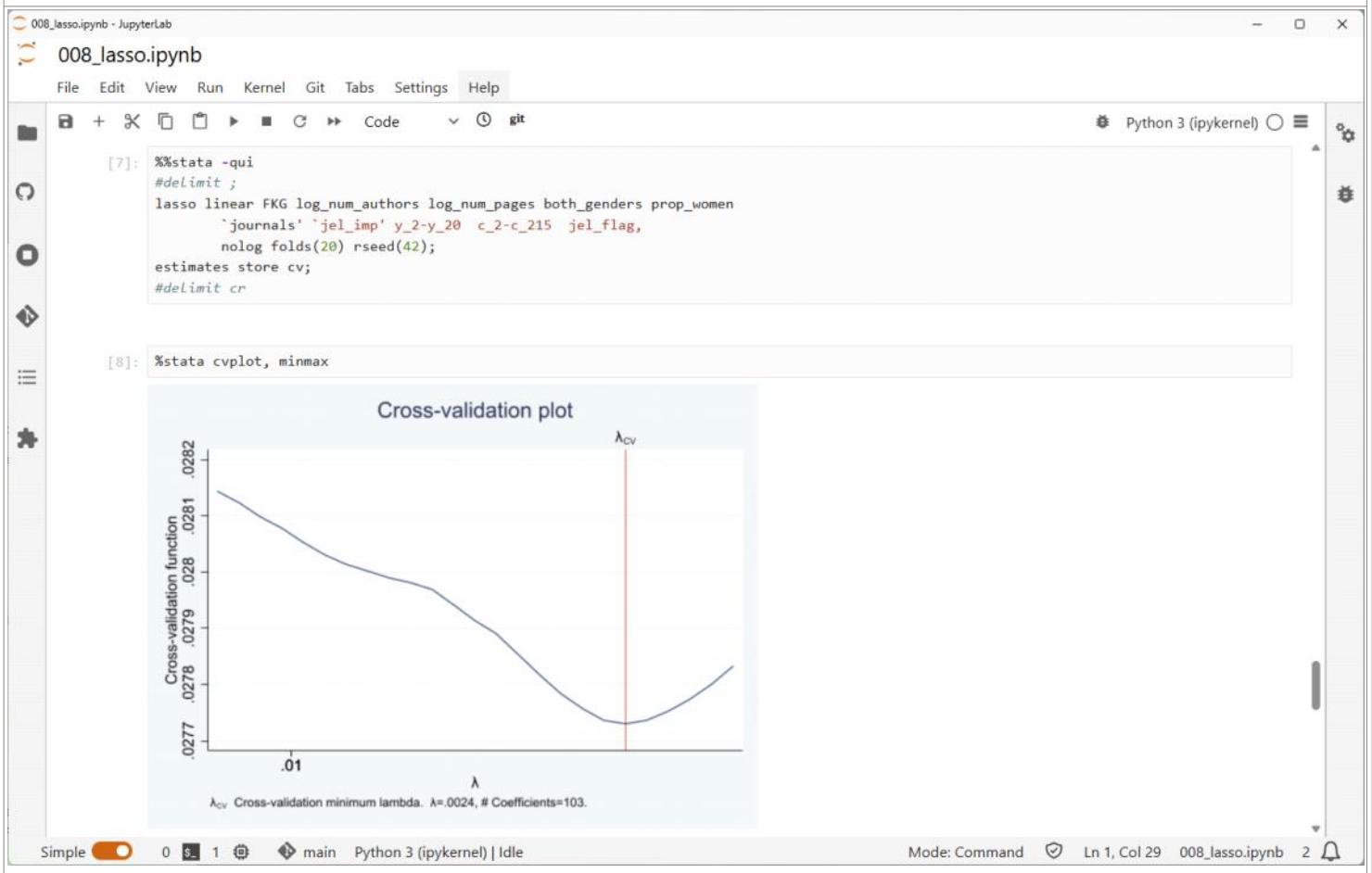
Cross Validation

👉 Critically important for Lasso estimation is the penalty λ.

- For λ close to zero the estimates are close to least squares.
- As λ increases the number of selected variables falls.
- 💡 Picking λ induces a trade-off between complexity and parsimony.

Asymptotic consistency of CV selection for Lasso estimation has been demonstrated by Chetverikov, Liao, and Chernozhukov (2021).

K-fold cross validation



"1se" Rule

Another popular choice is called the "1se" rule, which is the λ which yields the most parsimonious model for λ values within one standard error of the minimum. The idea is to select a model similar but more parsimonious than the CV-minimizing choice.

Large Sample Asymptotics

The model is the high-dimensional projection framework:

$$\begin{matrix} Y & = X'\beta + e \\ \mathbb{E}[Xe] & = 0 \end{matrix},$$

where X is $p \times 1$ with $p >> n$.

Sparsity:
The true coefficient vector β is assumed to be sparse in the sense that only a subset of the elements of β are non-zero.

For some λ let $\hat{\beta}$ be the Lasso estimator which minimizes $\text{SSE}_1(\beta, \lambda)$. Define the scaled design matrix $\boldsymbol{Q}_n = n^{-1}\boldsymbol{X}'\boldsymbol{X}$ and the **regression fit**

$$\left(\hat{\beta}_{\text{Lasso}} - \beta\right)' \boldsymbol{Q}_n \left(\hat{\beta}_{\text{Lasso}} - \beta\right) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i'\left(\hat{\beta}_{\text{Lasso}} - \beta\right)\right)^2.$$

When $p > n$ the matrix $\boldsymbol{Q}_n$ is singular.
The theory, however, requires that it not be "too singular".

- What is required is non-singularity of all sub-matrices of $\boldsymbol{Q}_n$ corresponding to the non-zero coefficients.
- Not "too many" of the zero coefficients.

If A is a $m \times n$ matrix, and B is a $n \times k$ matrix, then $\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B))$

Partition $\beta = (\beta_0, \beta_1)$ where the elements of β_0 are all 0 and the elements of β_1 are non-zero. (This partition is a theoretical device and unknown to the econometrician.)

Let $b = (b_0, b_1) \in \mathbb{R}^p$ be partitioned conformably. Define the cone $B = \left\{b \in \mathbb{R}^p : \|b_0\|_1 \leq 3\|b_1\|_1\right\}$. This is the set of vectors b such that the sub-vector b_0 is not "too large" relative to the sub-vector b_1.

Restricted Eigenvalue Condition (REC)

With probability approaching 1 as $n \to \infty$

$$\min_{b \in B} \frac{b'\boldsymbol{Q}_n b}{b'b} \geq c^2 > 0.$$

Notice that if the minimum is taken without restriction, i.e., over $\mathbb{R}^p$, it equals the smallest eigenvalue of $\boldsymbol{Q}_n$. Thus when $p < n$ a sufficient condition for the REC is $\lambda_{\min}(\boldsymbol{Q}_n) \geq c^2 > 0$. Instead, the minimum is calculated only over B. In this sense this calculation is similar to a "restricted eigenvalue" which is the source of its name.

Theorem: Suppose the high-dimensional projection framework holds with $p > 1$ and the REC Assumption holds. Assume that each regressor has been standardized so that $n^{-1}\boldsymbol{X}_j'\boldsymbol{X}_j = 1$ before applying the Lasso. Suppose $e \mid X \sim \text{N}(0, \sigma^2(X))$ where $\sigma^2(x) \leq \bar{\sigma}^2 < \infty$. For some C sufficiently large set

$$\lambda = C\sqrt{n \log p}.$$

Then there is $D < \infty$ such that with probability arbitrarily close to 1 ,

- These rates depend on the number of non-zero coefficients $\| \beta \|_0$, the number of variables p, and the sample size n.
- Suppose that $\| \beta \|_0$ is fixed. Then the bounds are $o(1)$ if $\log p = o(n)$. This shows that Lasso estimation is consistent even for an exponentially large number of variables.
- The rates, however, allow the number of non-zero coefficients $\| \beta \|_0$ to increase with n at the cost of slowing the allowable rate of increase of p.

$$\left(\widehat{\beta} - \beta\right) \boldsymbol{Q}_n \left(\widehat{\beta} - \beta\right) \leq D \parallel \beta \parallel_0 \frac{\log p}{n},$$

$$\left\|\widehat{\beta} - \beta\right\|_1 \leq D \parallel \beta \parallel_0 \sqrt{\frac{\log p}{n}}$$

and

$$\left\|\widehat{\beta} - \beta\right\|_2 \leq D \parallel \beta \parallel_0^{1/2} \sqrt{\frac{\log p}{n}}.$$

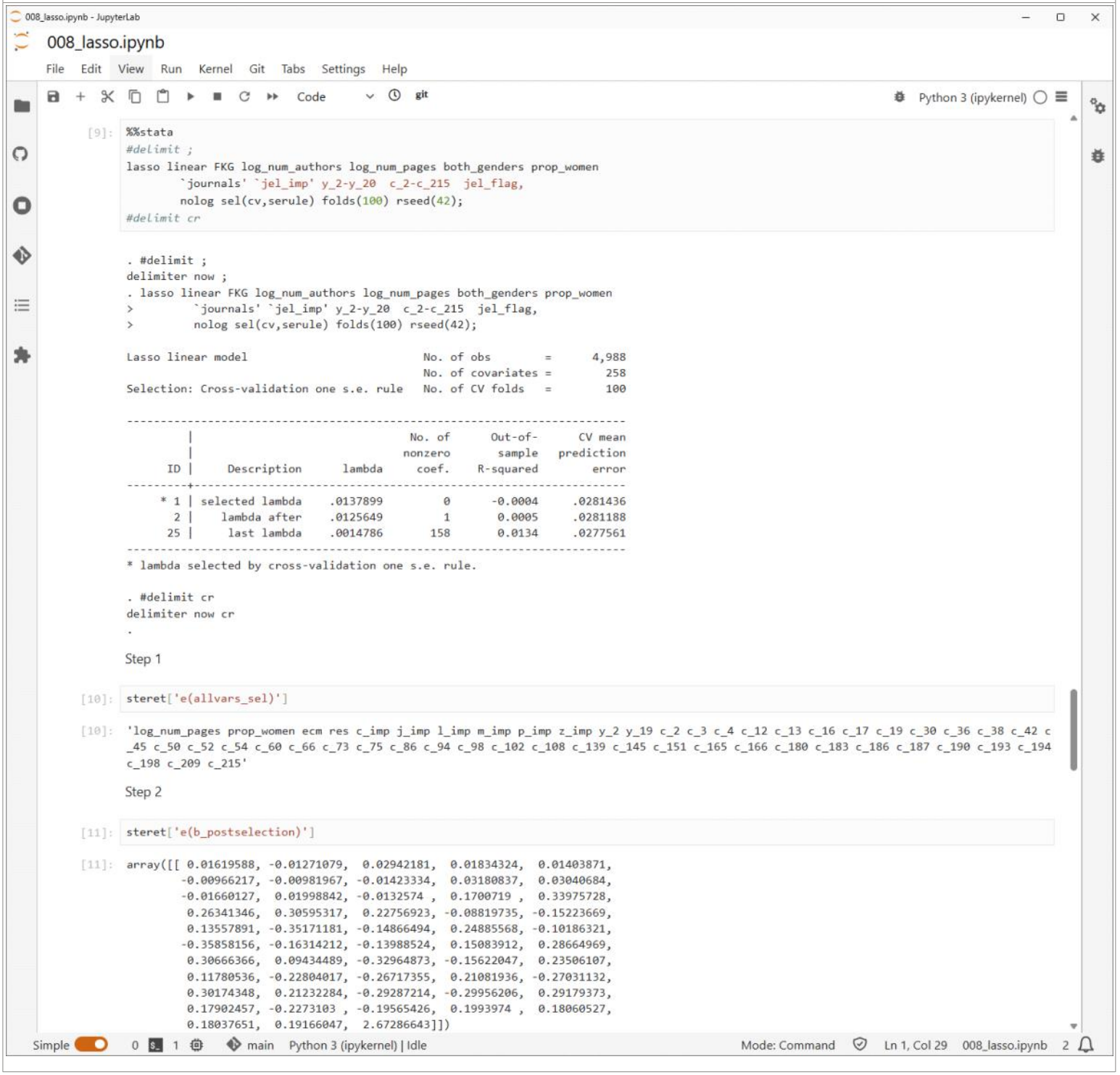
Postlasso

The Lasso estimator $\hat{\beta}$ simultaneously selects variables and shrinks coefficients. ➡ Shrinkage introduces bias into estimation. ➡ This bias can be reduced by applying least squares after Lasso selection. ⬅ This is known as the Post-Lasso estimator.

The procedure takes two steps:

(1) Estimate the model $Y = X'\beta + e$ by Lasso. Let X_S denote the variables in X which have non-zero coefficients in $\hat{\beta}$. Let β_S denote the corresponding coefficients in β.

(2) the coefficient β_S is estimated by least squares, thus $\hat{\beta}_S = \left(\boldsymbol{X}_S'\boldsymbol{X}_S\right)^{-1}\left(\boldsymbol{X}_S'\boldsymbol{Y}\right)$. This is the Post-Lasso least squares estimator.



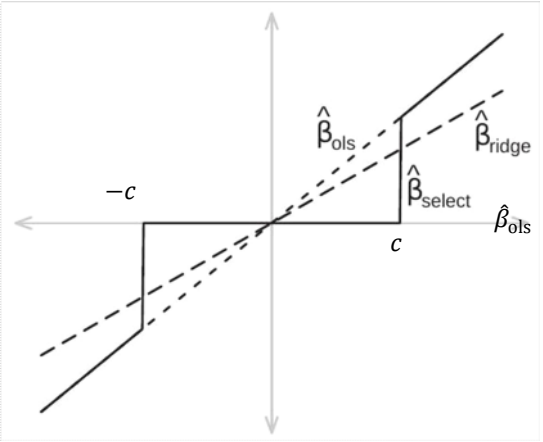
Lasso | Ridge | Selection

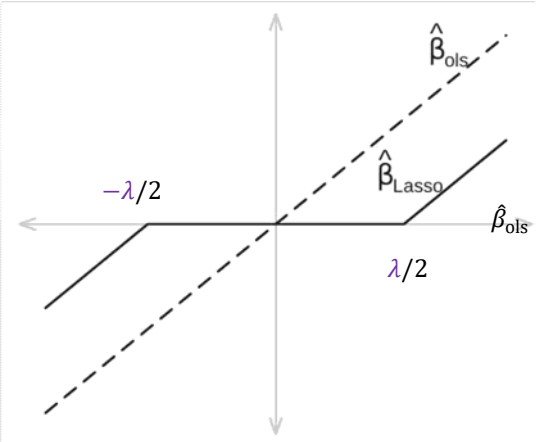
☞ One case where we can explicitly calculate the Lasso estimates is when the regressors are orthogonal, e.g., $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_p$. Then the first order condition for minimization simplifies to

$$-2\left(\hat{\beta}_{\text{ols},j} - \hat{\beta}_{\text{Lasso},j}\right) + \lambda \times \text{sgn}\left(\hat{\beta}_{\text{Lasso},j}\right) = 0,$$

which has the explicit solution

Estimator		Remarks
$\hat{\beta}_{\text{Lasso},j} = \begin{cases} \hat{\beta}_{\text{ols},j} - \lambda/2 & \hat{\beta}_{\text{ols},j} > \lambda/2 \\ 0 & \left\lvert\hat{\beta}_{\text{ols},j}\right\rvert \le \lambda/2. \\ \hat{\beta}_{\text{ols},j} + \lambda/2 & \hat{\beta}_{\text{ols},j} < -\lambda/2 \end{cases}$		• This shows that the Lasso estimate is a continuous transformation of the least squares estimate. • For small values of the least squares estimate the Lasso estimate is set to zero. • For all other values the Lasso estimate moves the least squares estimate towards zero by $\lambda/2$.
$\hat{\beta}_{\text{ridge}} = (1+\lambda)^{-1}\hat{\beta}_{\text{ols}}$		• It shrinks the OLS coefficient towards zero by a common factor.
$\hat{\beta}_{\text{select}} = \mathbb{1}\left\{\left\lvert\hat{\beta}_{\text{ols }j}\right\rvert > c\right\}\hat{\beta}_{\text{ols},j}$		A selection estimator • For simplicity consider selection based on a homoskedastic t-test with $\hat{\sigma}^2 = 1$ and critical value c.





- The Lasso and ridge estimators are continuous functions while the selection estimator is a discontinuous function.
- The Lasso ("soft") and selection ("hard")estimators are thresholding functions, meaning that the function equals zero for a region about the origin.
- Thresholding estimators are selection estimators because they equal zero when the least squares estimator is sufficiently small.

Hard thresholding rules tend to have high variance due to the discontinuous transformation. Consequently, we expect the Lasso to have reduced variance relative to selection estimators.

Elastic Net

The difference between Lasso and ridge regression is that the Lasso uses the 1-norm penalty while ridge uses the 2-norm penalty.

💡 Since both procedures have advantages it seems reasonable that further improvements may be obtained by a compromise.

Zou and Hastie (2005) showed that taking a weighted average of the penalties we obtain the Elastic Net criterion

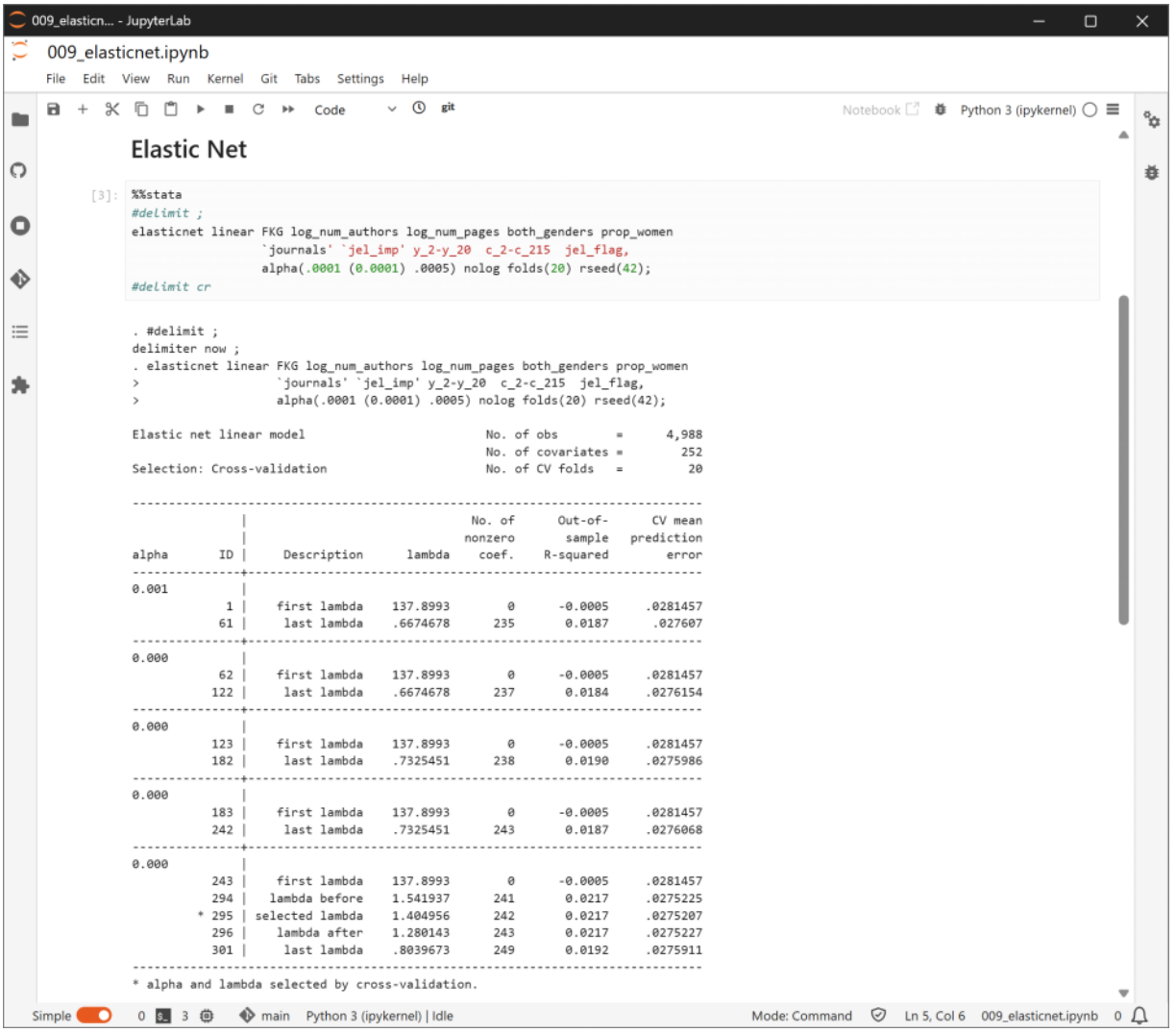
$$\text{SSE}(b, \lambda, \alpha) = (\boldsymbol{Y} - \boldsymbol{X}b)'(\boldsymbol{Y} - \boldsymbol{X}b) + \lambda\big((1-\alpha) \parallel b \parallel_2^2 + \alpha \parallel b \parallel_1\big),$$

with weight $0 \leq \alpha \leq 1$.

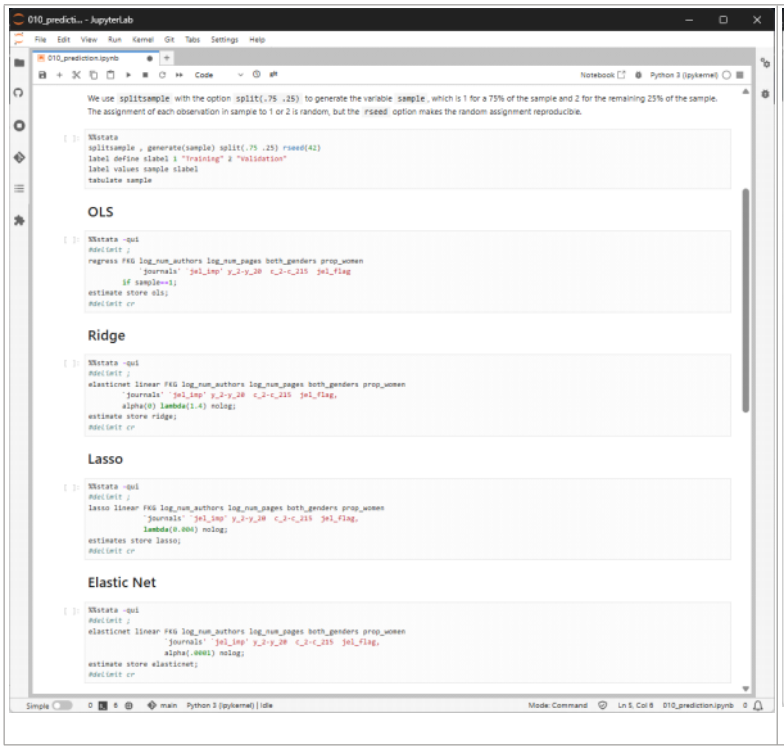
✔ This includes Lasso ($\alpha = 1$) and

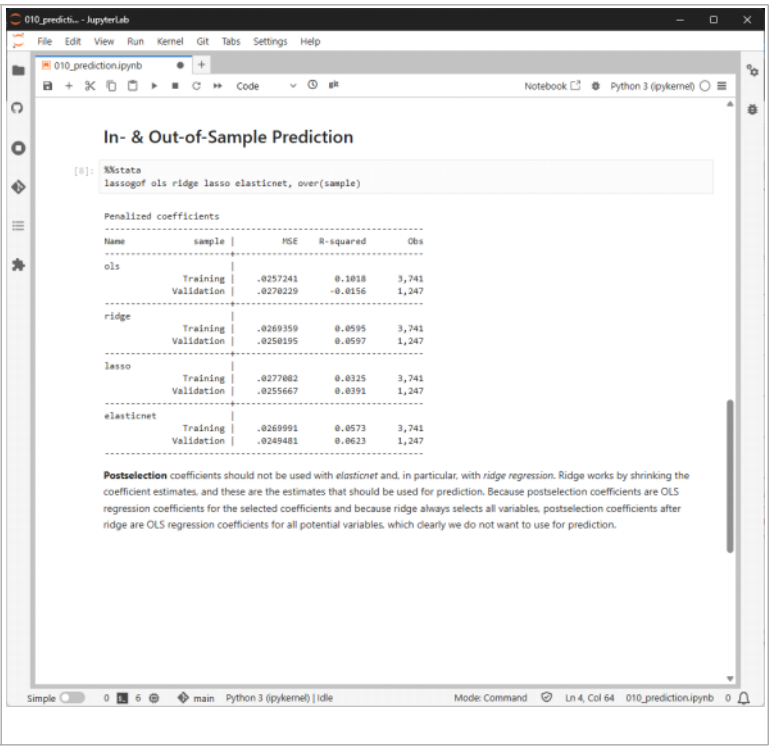
✔ ridge regression ($\alpha = 0$) as special cases.

Typically the parameters (α, λ) are selected by joint minimization of the k-fold cross-validation criterion. Since the elastic net penalty is linear-quadratic the solution is computationally similar to Lasso.



Prediction


We use splitsample with the option split(.75 .25) to generate the variable sample, which is 1 for a 75% of the sample and 2 for the remaining 25% of the sample.
The assignment of each observation in sample to 1 or 2 is random, but the rseed option makes the random assignment reproducible.
%%stata
splitsample , generate(sample) split(.75 .25) rseed(42)
label define slabel 1 "Training" 2 "Validation"
label values sample slabel
tabulate sample
OLS
%%stata -qui
#delimit ;
regress FKG log_num_authors log_num_pages both_genders prop_women
`journals' `jel_imp' y_2-y_20 c_2-c_215 jel_flag
if sample==1;
estimate store ols;
#delimit cr
Ridge
%%stata -qui
#delimit ;
elasticnet linear FKG log_num_authors log_num_pages both_genders prop_women
`journals' `jel_imp' y_2-y_20 c_2-c_215 jel_flag,
alpha(0) lambda(1.4) nolog;
estimate store ridge;
#delimit cr
Lasso
%%stata -qui
#delimit ;
lasso linear FKG log_num_authors log_num_pages both_genders prop_women
`journals' `jel_imp' y_2-y_20 c_2-c_215 jel_flag,
lambda(0.004) nolog;
estimates store lasso;
#delimit cr
Elastic Net
%%stata -qui
#delimit ;
elasticnet linear FKG log_num_authors log_num_pages both_genders prop_women
`journals' `jel_imp' y_2-y_20 c_2-c_215 jel_flag,
alpha(.0001) nolog;
estimate store elasticnet;
#delimit cr


In- & Out-of-Sample Prediction
%%stata
lassogof ols ridge lasso elasticnet, over(sample)
Penalized coefficients
Name sample MSE R-squared Obs
ols
Training .0257241 0.1018 3,741
Validation .0270229 -0.0156 1,247
ridge
Training .0269359 0.0595 3,741
Validation .0250195 0.0597 1,247
lasso
Training .0277082 0.0325 3,741
Validation .0255667 0.0391 1,247
elasticnet
Training .0269991 0.0573 3,741
Validation .0249481 0.0623 1,247
Postselection coefficients should not be used with elasticnet and, in particular, with ridge regression. Ridge works by shrinking the coefficient estimates, and these are the estimates that should be used for prediction. Because postselection coefficients are OLS regression coefficients for the selected coefficients and because ridge always selects all variables, postselection coefficients after ridge are OLS regression coefficients for all potential variables, which clearly we do not want to use for prediction.

Inference

Consider the linear model

$$\begin{aligned} Y &= D\theta + X'\beta + e \\ \mathbb{E}[e \mid D, X] &= 0 \end{aligned},$$

The goal is inference on θ.

where Y and D are scalar and X is $p \times 1$.

- The variable D is the main focus of the regression;
- the variable X are controls.

Suppose you estimate model by group post-Lasso, only penalizing β. This performs selection on the variables X, resulting in a least squares regression of Y on D and the selected variables in X.

😔 The coverage probabilities for θ are downward biased, and the distortions are serious.

😵 The distortions are primarily affected by (and increasing in) the correlation between D and X.

Belloni, Chernozhukov, and Hansen (2014b) deduce that improved coverage accuracy can be achieved if the variable X is included in the regression whenever X and D are correlated.

Double Selection Lasso

Belloni, Chernozhukov, and Hansen (2014b) proposed to perform what they call ***double-selection***. We start by specifying an auxiliary equation for D :

$$\begin{array}{ll} D & = X'\gamma + V \\ \mathbb{E}[V \mid X] & = 0. \end{array}.$$

Substituting this into our structural model, we obtain a reduced form for Y :

$$\begin{array}{ll} Y & = X'\eta + U \\ \mathbb{E}[U \mid X] & = 0 \end{array}.$$

The proposed double-selection algorithm applies model selection (e.g., Lasso selection) separately to each equations, takes the union of the sele cted regressors, and then estimates structural equation by least squares using the selected regressors. This method ensures that a variable X is included if it is relevant for the structural regression or if it is correlated with D.

The ***double-selection*** estimator as recommended by Belloni, Chernozhukov, and Hansen (2014b) is:

1 Estimate γ by Lasso. Let X_1 be the selected variables from X.

2 Estimate η by Lasso. Let X_2 be the selected variables from X.

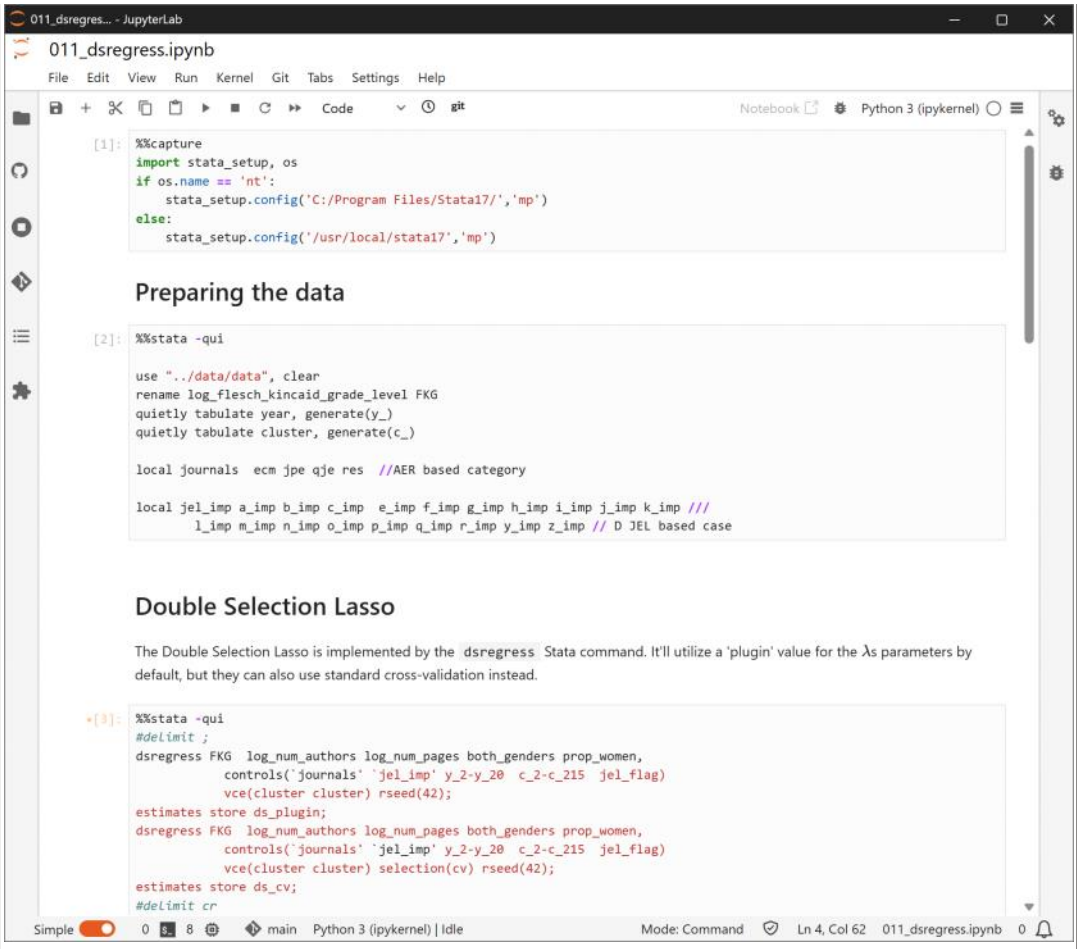
3 Let $\tilde{X} = X_1 \cup X_2$ be the union of the variables in X_1 and X_2.

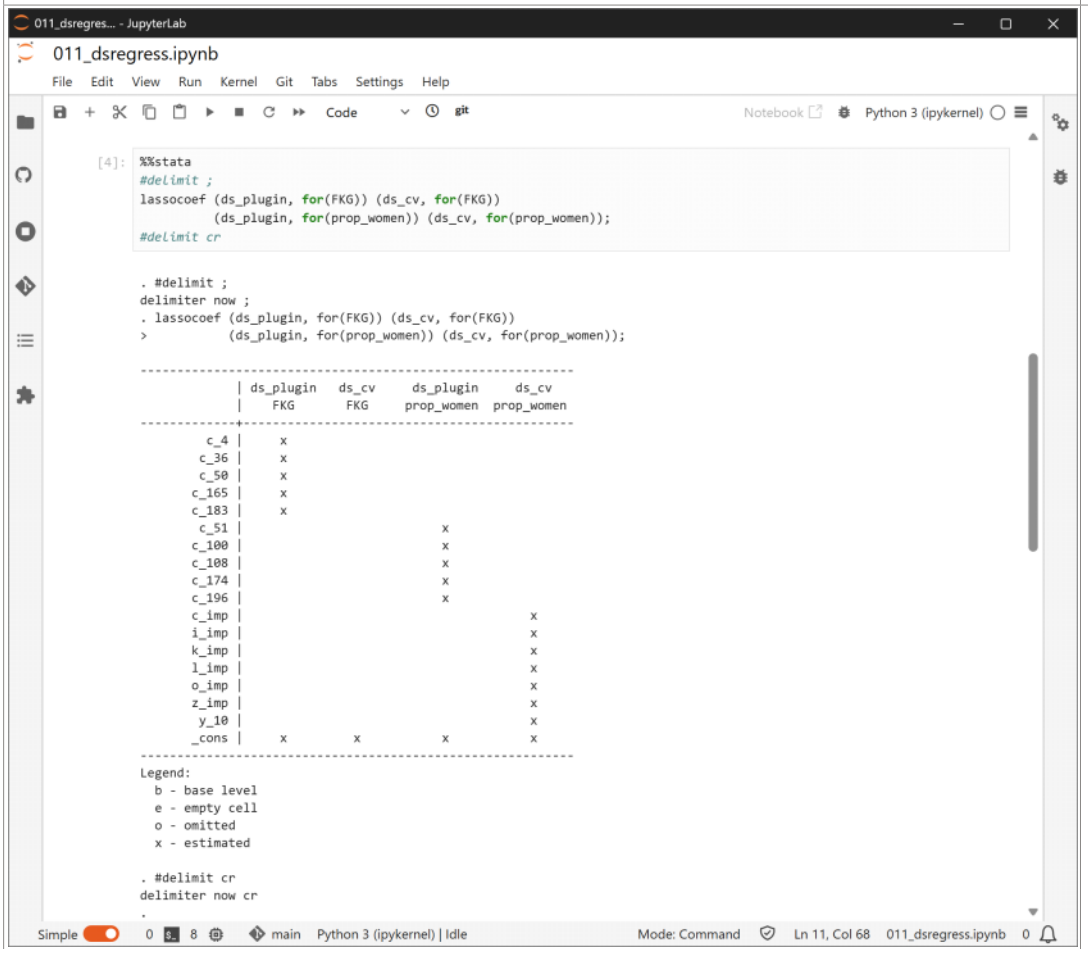
4 Regress Y on $(D, \tilde{X})$ by OLS to obtain the double-selection coefficient estimate $\hat{\theta}_{\text{Ds}}$.

5 Calculate a conventional (heteroskedastic) standard error for $\hat{\theta}_{DS}$.

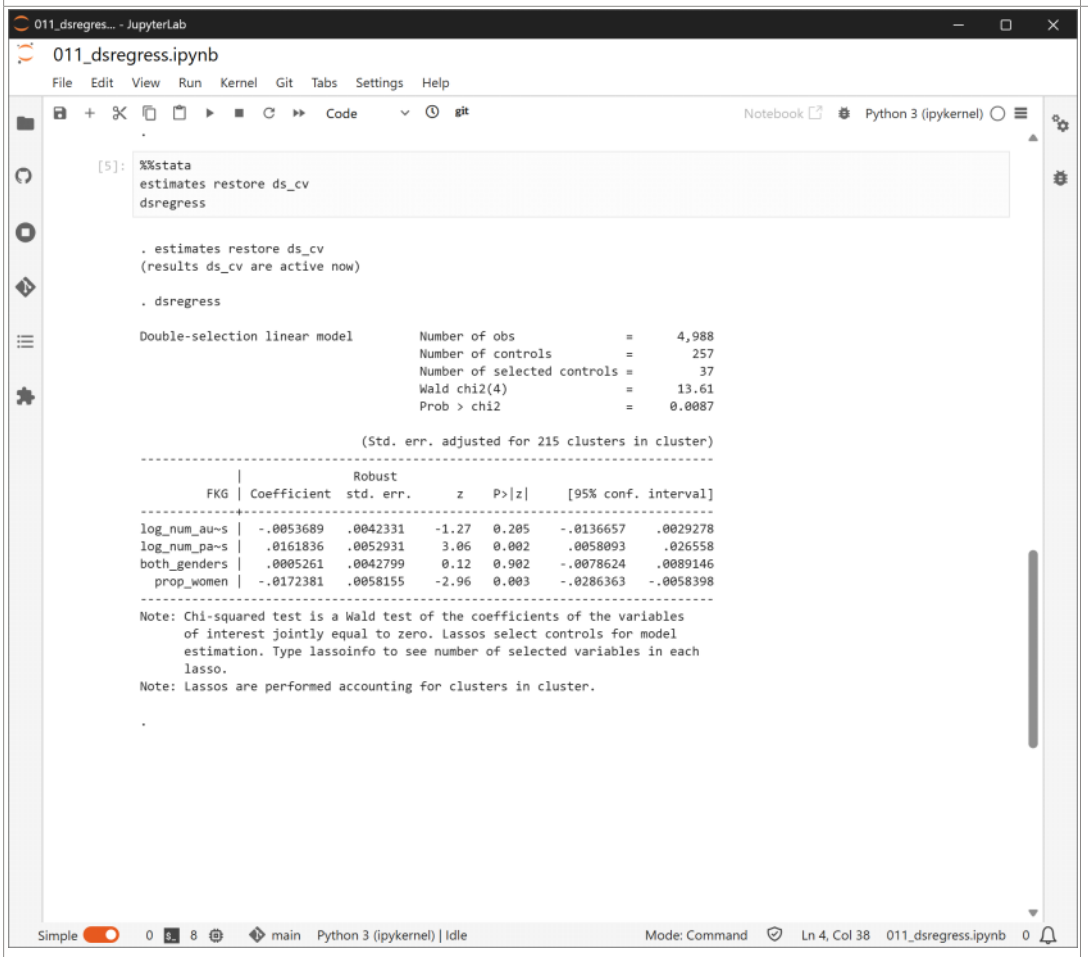
$$\eta = \beta + \gamma\theta,$$
$$U = e + V\theta$$

Provided the models in steps 1 and 2 satisfy an approximate sparsity structure (so that the regressions are well approximated by a finite set of regressors) then the double-selection estimator $\hat{\theta}_{\text{DS}}$ and its t-ratio are asymptotically normal so conventional inference methods are valid.		The essential idea is that because $\tilde{X}$ includes the variables in X_2, the estimator $\hat{\theta}_{\text{DS}}$ is asymptotically equivalent to the regression where D is replaced with the error V. Since V is uncorrelated with the regressors X the estimator and t-ratio satisfy the conventional non-selection asymptotic distribution.





The first two columns of x's show which controls were selected from the lassos for the dependent variable, `log_flesch_kincaid_grade_level`—the first column for the plugin method and the second for cross-validation. The third and fourth columns of x's show which controls were selected by the lassos for the covariate of interest, `prop_women`.



We interpret the coefficient estimates just as we would for a standard linear regression.

⚠ What we lose with the inferential lasso estimators is the ability to interpret any other coefficients.

Our point estimate for the effect of the proportion of women co-authors on abstract readability is -0.017, meaning that going from no women co-authors (`prop_women`=0) to all women co-authors (`prop_women`=1) will decrease the number of schooling needed to read the abstract by 1.724%. This value is statistically different from 0 well beyond the 5% level, in fact, beyond the 0.1% level. Our 95% confidence interval is -0.0286 to -0.0059.

Post-Regularization Lasso

[Partialing-Out] Lasso

Chernozhukov, Hansen, and Spindler (2015) proposed to transform the structural equation to eliminate the *high-dimensional* component. Take the expected value of $Y = D\theta + X'\beta + e$ conditionally on X, and subtract from each side. This leads to the equation

$$Y - \mathbb{E}[Y \mid X] = (D - \mathbb{E}[D \mid X])\theta + e.$$

Notice that this eliminates the regressor X and the high-dimensional coefficient β!

The authors specify $\mathbb{E}[Y \mid X]$ and $\mathbb{E}[D \mid X]$ as linear functions of X. Substituting these expressions we obtain

$$Y - X'\eta = (D - X'\gamma)\theta + e.$$

If η and γ were known the coefficient θ could be estimated by least squares. However, as η and γ are unknown they need to be estimated. Chernozhukov, Hansen, and Spindler (2015) recommend estimation by Lasso or post-Lasso, separately for Y and D as follows:

1. Estimate a regression of Y on X by Lasso or post-Lasso (Stata) with Lasso parameter λ_1. Let $\hat{\eta}$ be the coefficient estimator and $\hat{U}_i = Y_i - X_i'\hat{\eta}$ the residual.
2. Estimate a regression of D on X by Lasso or post-Lasso (Stata) with Lasso parameter λ_2. Let $\hat{\gamma}$ be the coefficient estimator and $\hat{V}_i = D_i - X_i'\hat{\gamma}$ the residual.
3. Let $\hat{\theta}_{\text{PR}}$ be the OLS coefficient from the regression of $\hat{U}$ on $\hat{V}$.
4. Calculate a conventional (heteroskedastic) standard error for $\hat{\theta}_{\text{PR}}$.

Theorem

Suppose

$$\begin{aligned} D &= X'\gamma + V \\ \mathbb{E}[V \mid X] &= 0. \end{aligned}$$

$$\begin{aligned} Y &= X'\eta + U \\ \mathbb{E}[U \mid X] &= 0 \end{aligned}$$

Define the scaled design matrix $\boldsymbol{Q}_n = n^{-1}\boldsymbol{X}'\boldsymbol{X}$, with probability approaching 1 as $n \to \infty$

$$\min_{b \in B} \frac{b'\boldsymbol{Q}_n b}{b'b} \geq c^2 > 0.$$

Assume that each regressor has been standardized so that $n^{-1}\boldsymbol{X}_j'\boldsymbol{X}_j = 1$. Suppose $e \mid X \sim \mathrm{N}(0, \sigma_e^2(X))$ and $V \mid X \sim \mathrm{N}(0, \sigma_V^2(X))$ where $\sigma_e^2(x) \leq \bar{\sigma}_e^2 < \infty$ and $\sigma_V^2(x) \leq \bar{\sigma}_V^2 < \infty$. For some C_1 and C_2 sufficiently large the Lasso parameters satisfy $\lambda_1 = C_1\sqrt{n \log p}$ and $\lambda_2 = C_2\sqrt{n \log p}$. Assume $p \to \infty$ and

$$\left(\| \beta \|_0 + \| \gamma \|_0\right) \frac{\log p}{\sqrt{n}} = o(1).$$

Then

$$\sqrt{n}\left(\hat{\theta}_{\text{PR}} - \theta\right) \xrightarrow[d]{} \mathrm{N}\left(0, \frac{\mathbb{E}[V^2 e^2]}{(\mathbb{E}[V^2])^2}\right).$$

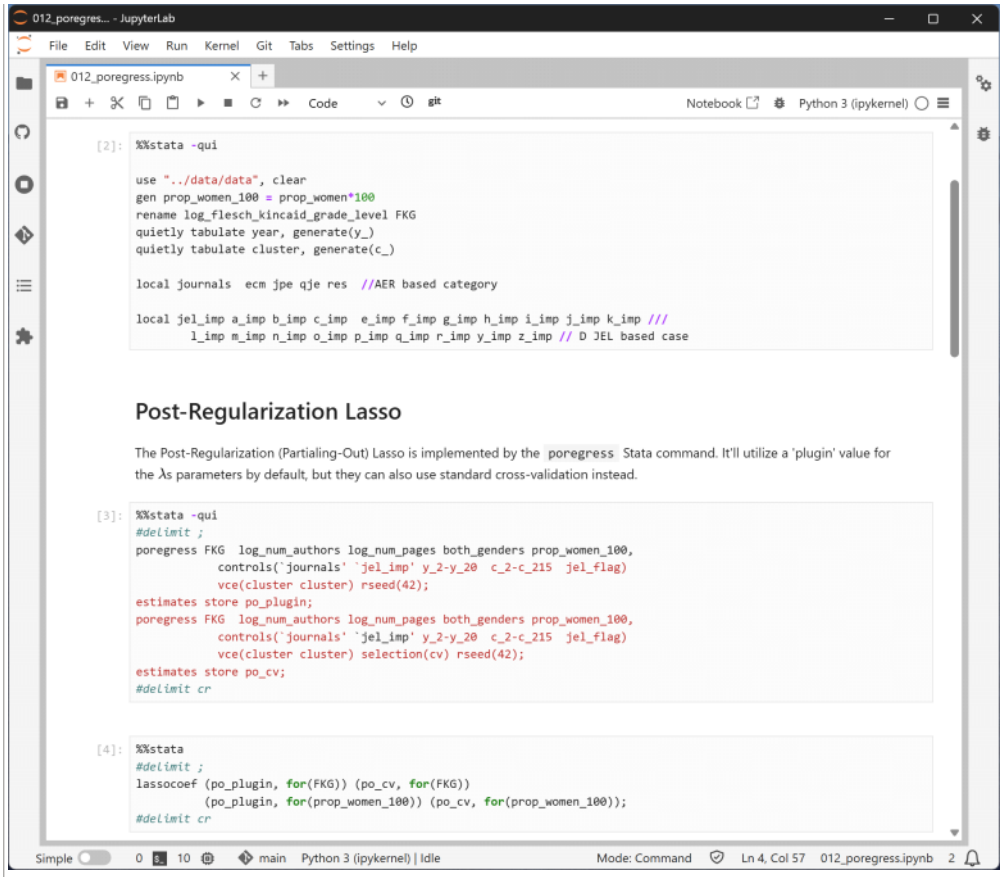
- The original result does not need the assumption of normal errors.
- Theorem shows that the post-regularization (partialing-out) Lasso estimator has a conventional asymptotic distribution, allowing conventional inference for the coefficient θ, i.e., the standard variance estimator for $\hat{\theta}_{\text{PR}}$ is consistent for the asymptotic variance.

The advantage of the post-regularization estimator $\hat{\theta}_{\text{PR}}$ over the double-selection estimator $\hat{\theta}_{\text{DS}}$ is efficiency.

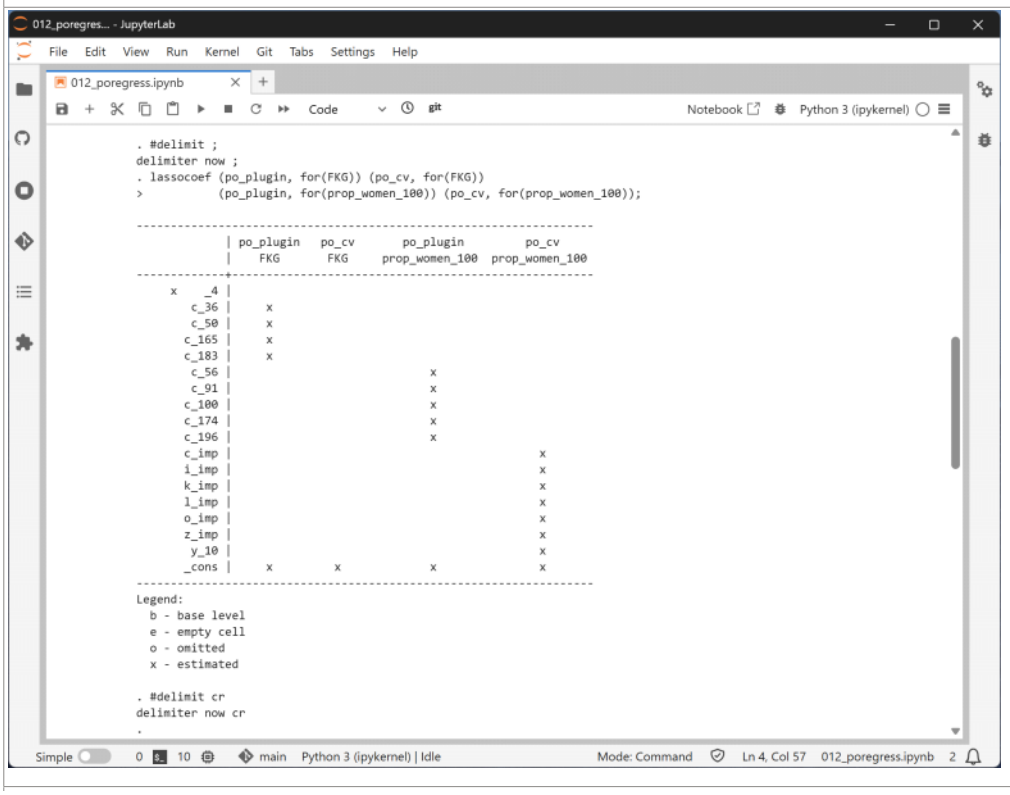
The post-regularization estimator uses only the relevant components of X to separately demean Y and D, leading to greater parsimony. Different components of X may be relevant to D and Y.

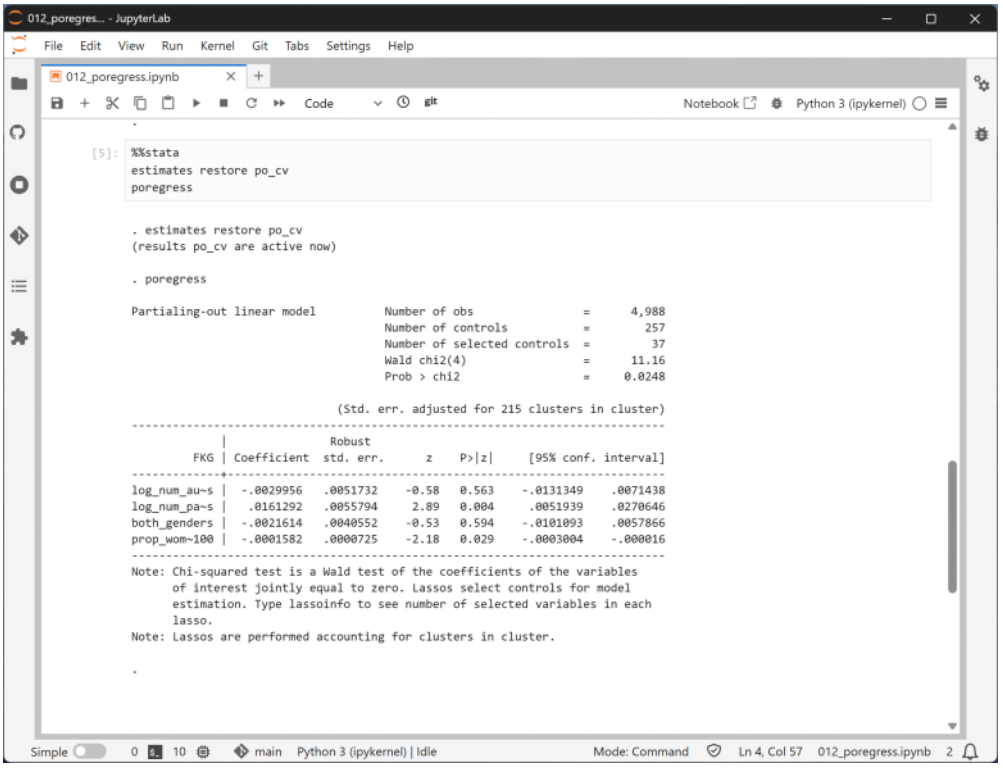
The double-selection estimator uses the union of the two regressor sets for estimation of θ, leading to a less parsimonious specification.

As a consequence, an advantage of the double-selection estimator is reduced bias and robustness.



We multiply the variable of interest `prop_women` by 100 so a 'unit' change is a 1% change.





✔ These use post-lasso estimation to construct the residuals $\hat{U}$ and $\hat{V}$.

Since the dependent variable, `log_flesch_kincaid_grade_level` is measured in logarithm, the point estimate -0.0001582 is the elasticity of the Flesch-kincaid Grade index with respect to the proportion of women co-authors. This is statistically significant at 10% & 5%.

Double/Debiased Machine Learning

Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) proposed the Double/Debiased machine learning (DML) estimator.

As presented previously, the post-regularization estimator first estimates the coefficients γ and η in the models

$$\begin{array}{ll} D & = X'\gamma + V \\ \mathbb{E}[V \mid X] & = 0. \end{array}.$$

$$\begin{array}{ll} Y & = X'\eta + U \\ \mathbb{E}[U \mid X] & = 0 \end{array}.$$

and then estimates the coefficient θ.

The DML estimator performs these estimation steps using separate samples based on K-fold partitioning. The estimation algorithm is as follows:

1 Randomly partition the sample into K independent folds A_k, $k = 1, \ldots, K$, of roughly equal size n/K.

2 Write the data matrices for each fold as $(\boldsymbol{Y}_k, \boldsymbol{D}_k, \boldsymbol{X}_k)$.

3 For $k = 1, \ldots, K$

(a) Use all observations except for fold k to estimate the coefficients γ and η by Lasso or post-Lasso (Stata). Write these leave-*fold*-out estimators as $\hat{\gamma}_{-k}$ and $\hat{\eta}_{-k}$.

(b) Set $\widehat{\boldsymbol{V}}_k = \boldsymbol{D}_k - \boldsymbol{X}_k\hat{\gamma}_{-k}$ and $\widehat{\boldsymbol{U}}_k = \boldsymbol{Y}_k - \boldsymbol{X}_k\hat{\eta}_{-k}$. These are the estimated values of V and U for observations in the k^{th} fold using the leave-*fold*-out estimators.

4 Set $\hat{\theta}_{\text{DML}} = \left(\sum_{k=1}^{K}\widehat{\boldsymbol{V}}_k'\widehat{\boldsymbol{V}}_k\right)^{-1}\left(\sum_{k=1}^{K}\widehat{\boldsymbol{V}}_k'\widehat{\boldsymbol{U}}_k\right)$. Equivalently, stack $\widehat{\boldsymbol{V}}_k$ and $\widehat{\boldsymbol{U}}_k$ into $n \times 1$ vectors $\widehat{\boldsymbol{V}}$ and $\widehat{\boldsymbol{U}}$ and set $\hat{\theta}_{\text{DML}} = \left(\widehat{\boldsymbol{V}}'\widehat{\boldsymbol{V}}\right)^{-1}\left(\widehat{\boldsymbol{V}}'\widehat{\boldsymbol{U}}\right)$.

5 Construct a conventional (heteroskedastic) standard error for $\hat{\theta}_{\text{DML}}$.

The authors label this estimator the "DML2" estimator.

An alternative they label "DML1" is $\hat{\theta}_{\text{DML1}} = K^{-1}\sum_{k=1}^{K}\left(\widehat{\boldsymbol{V}}_k'\widehat{\boldsymbol{V}}_k\right)^{-1}\left(\widehat{\boldsymbol{V}}_k'\widehat{\boldsymbol{U}}_k\right)$.

Although $\hat{\theta}_{\text{DML1}}$ is asymptotically equivalent to $\hat{\theta}_{\text{DML}}$, the latter is preferred.

Advantage:

- The advantage of the DML estimator over the post-regularization estimator is that the sample splitting eliminates the dependence between the two estimation steps, thereby reducing post-model-selection bias.

Theorem

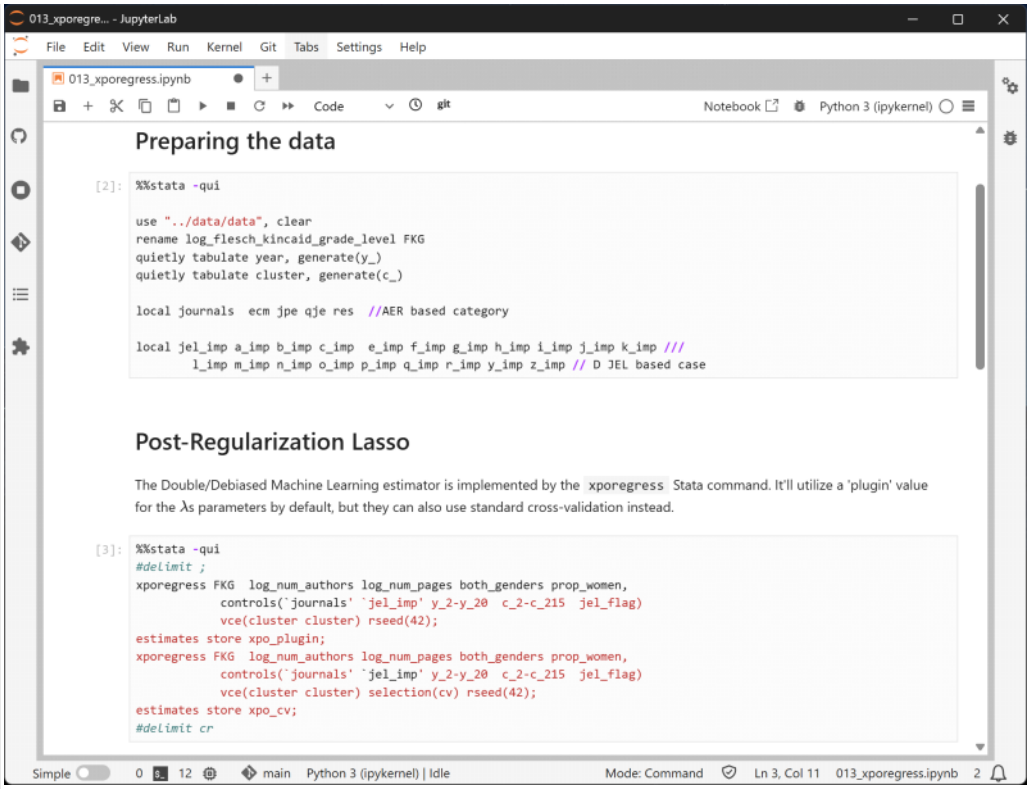
Under the assumptions of the previous Theorem,

$$\sqrt{n}\left(\hat{\theta}_{\text{DML}} - \theta\right) \xrightarrow[d]{} \text{N}\left(0, \frac{\mathbb{E}[V^2e^2]}{(\mathbb{E}[V^2])^2}\right).$$

Furthermore, the standard variance estimator for $\hat{\theta}_{\text{DML}}$ is consistent for the asymptotic variance.

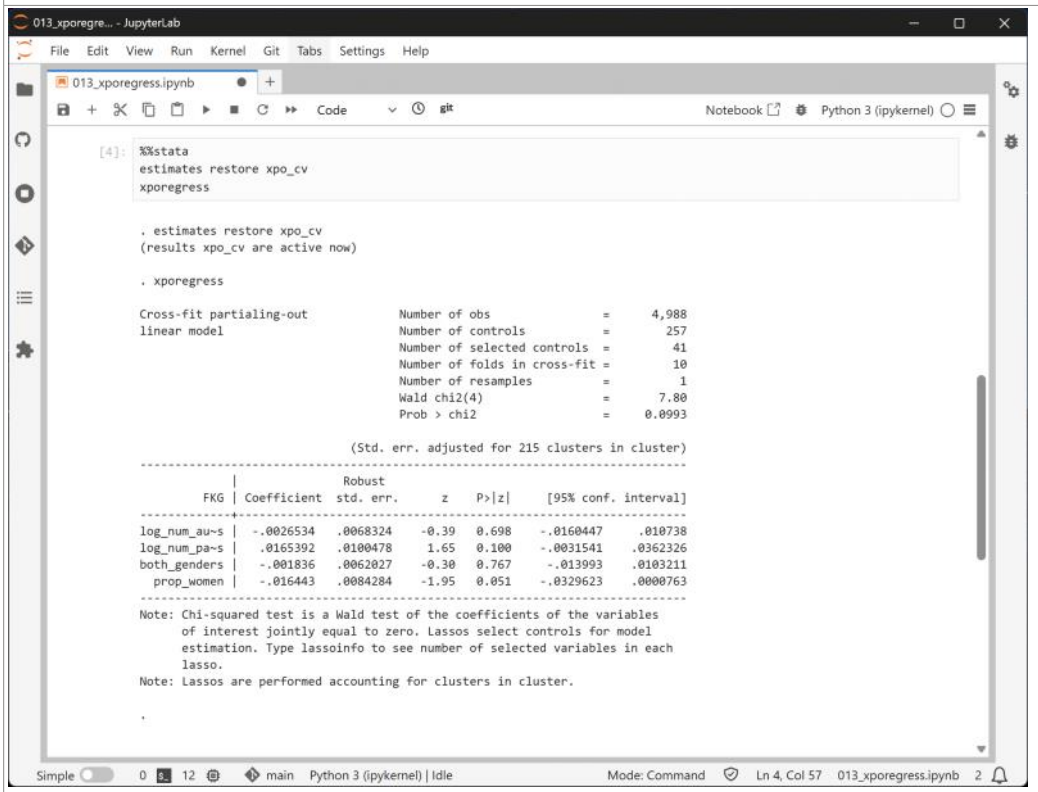
Disadvantages:

- First, the estimator is random due to the sample splitting. Two researchers with the same data set but making different random splits will obtain two distinct estimators. This randomness can be reduced by using a larger value of K, but this increases computation cost.
- Another disadvantage of sample-splitting is that estimation of γ and η is performed using smaller samples which reduces estimation efficiency, though this effect is minor if $K \geq 10$. Regardless, these considerations suggest that DML may be most appropriate for settings with large n and $K \geq 10$.



`xporegress` fits 10 lassos for the dependent variable and 10 more lassos for each covariate of interest! That is the default; you can request more. Or you can request fewer, but that is not recommended.

So, `xporegress` is orders of magnitude slower than `poregress` and `dsregress`. And it has orders of magnitude more lassos to explore.



Our point estimate for the effect of the proportion of women co-authors on abstract readability is -0.0164, meaning that going from no women co-authors (`prop_women`=0) to all women co-authors (`prop_women`=1) will decrease the number of schooling needed to read the abstract by 1.644%.

What to Use?

👉 It is recommended you should start with the Double/Debiased Machine Learning estimator (xporegress) first.

Why?

📝 It is safer if you think that the process that generated your data has lots of covariates relative to your sample size.
📝 Similarly, it is also safer if you want to explore lots of potential controls.
- The number of potential controls is not as problematic as the number of true covariates because it is the natural log of the potential control that counts. For example, needing 10 additional true covariates is the same as requesting just over 22,000 potential controls! The jargon term for this is *sparsity*.
- xporegress has a weaker sparsity requirement than do Post-Regularization Lasso (poregress) and Double Selection Lasso (dsregress).

🤔 However …

If your model is weakly identified by the data (dropping one observation or regressor would change estimates dramatically, etc.), then the Double Selection Lasso (dsregress) can be more stable than the others because it uses a union of all the selected controls from all the lassos for all of its computations after selection. ⚠ … more stable doesn't mean better !

Lasso | Logit & Poisson

Recall that the Lasso estimator for the *linear* model is defined as the minimizer

$$\hat{\beta} = \underset{b}{\operatorname{argmin}} \, \mathrm{SSE}_1(b, \lambda).$$

where the least squares criterion with a 1-norm penalty is

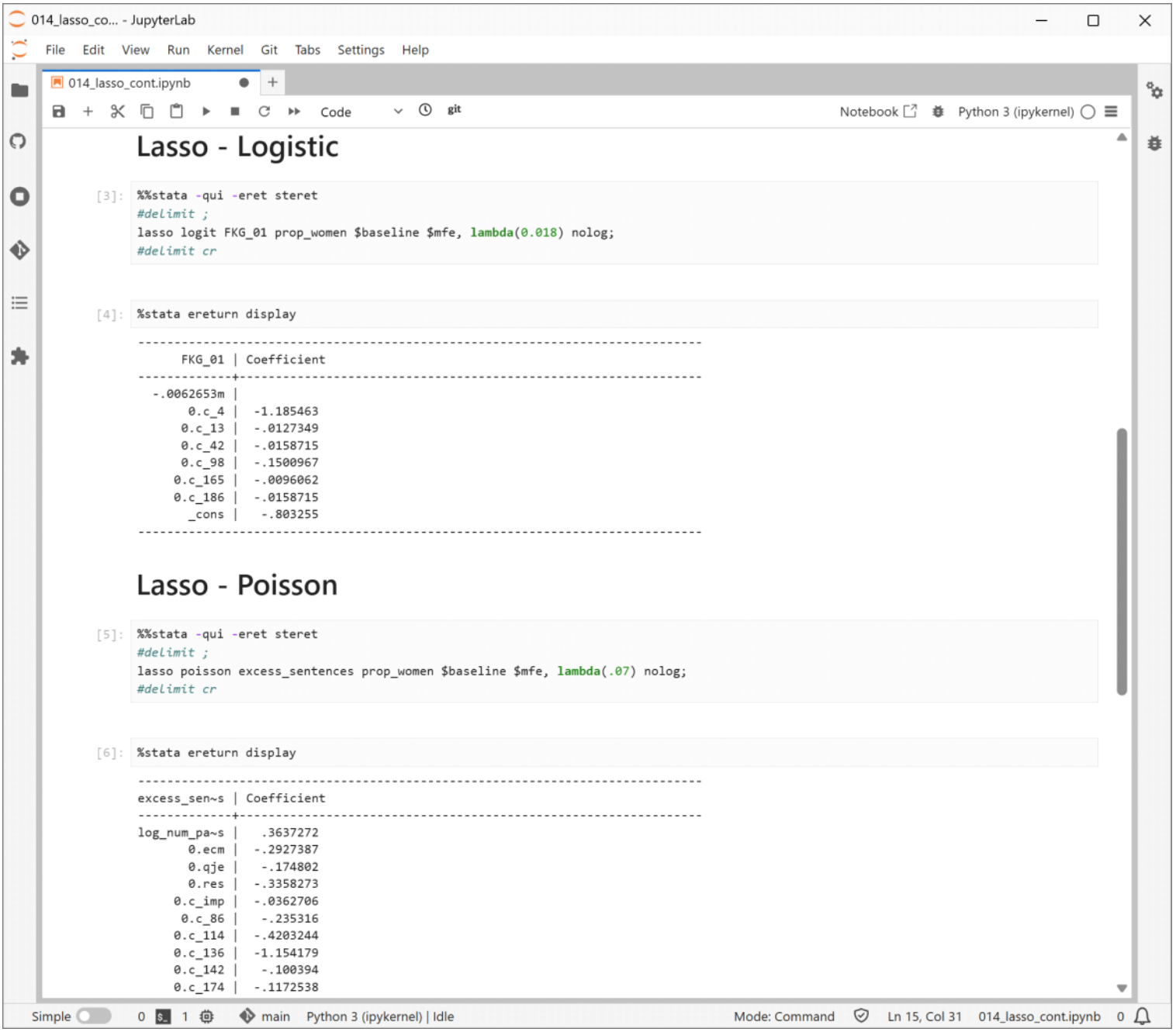
$$\mathrm{SSE}_1(b, \lambda) = (\boldsymbol{Y} - \boldsymbol{X}b)'(\boldsymbol{Y} - \boldsymbol{X}b) + \lambda \sum_{j=1}^{p} \left| b_j \right| = \| \boldsymbol{Y} - \boldsymbol{X}b \|_2^2 + \lambda \| b \|_1.$$

Equivalentely, we could write

$$\hat{\beta} = \underset{b}{\operatorname{argmin}} \, Q_n(b) + \lambda \| b \|_1.$$

where $Q_n(b) \equiv \sum_{i=1}^{n} \left(Y_i - X_i' b \right)^2$ when the model is *linear*.

Model	$Q_n(b)$.
Linear	$\sum_{i=1}^{n} \left(Y_i - X_i' b \right)^2$
Logit	$\sum_{i=1}^{n} -Y_i \left(X_i' b \right) + \log\{1 + \exp(X_i' b)\}$
Probit	$\sum_{i=1}^{n} -Y_i \times \log\{\Phi\left(X_i' b \right)\} - \left(1 - Y_i \right) \log\{1 - \Phi(X_i' b)\}$
Poisson	$\sum_{i=1}^{n} -Y_i \left(X_i' b \right) + \exp\{X_i' b\}$


014_lasso_co... - JupyterLab
File Edit View Run Kernel Git Tabs Settings Help
014_lasso_cont.ipynb
Code
git
Notebook
Python 3 (ipykernel)
Lasso - Logistic
[3]:
%%stata -qui -eret steret
#delimit ;
lasso logit FKG_01 prop_women $baseline $mfe, lambda(0.018) nolog;
#delimit cr
[4]:
%stata ereturn display
FKG_01 | Coefficient
-.0062653m |
0.c_4 | -1.185463
0.c_13 | -.0127349
0.c_42 | -.0158715
0.c_98 | -.1500967
0.c_165 | -.0096062
0.c_186 | -.0158715
_cons | -.803255
Lasso - Poisson
[5]:
%%stata -qui -eret steret
#delimit ;
lasso poisson excess_sentences prop_women $baseline $mfe, lambda(.07) nolog;
#delimit cr
[6]:
%stata ereturn display
excess_sen~s | Coefficient
log_num_pa~s | .3637272
0.ecm | -.2927387
0.qje | -.174802
0.res | -.3358273
0.c_imp | -.0362706
0.c_86 | -.235316
0.c_114 | -.4203244
0.c_136 | -1.154179
0.c_142 | -.100394
0.c_174 | -.1172538
Simple
main
Python 3 (ipykernel) | Idle
Mode: Command
Ln 15, Col 31
014_lasso_cont.ipynb

Elastic Net | Logit & Poisson

Recall that the Elastic Net estimator for the *linear* model is defined as the minimizer of

$$\text{SSE}(b,\lambda,\alpha) = (\boldsymbol{Y}-\boldsymbol{X}b)'(\boldsymbol{Y}-\boldsymbol{X}b) + \lambda\big((1-\alpha)\parallel b \parallel_2^2 + \alpha \parallel b \parallel_1\big),$$

with weight $0 \le \alpha \le 1$.

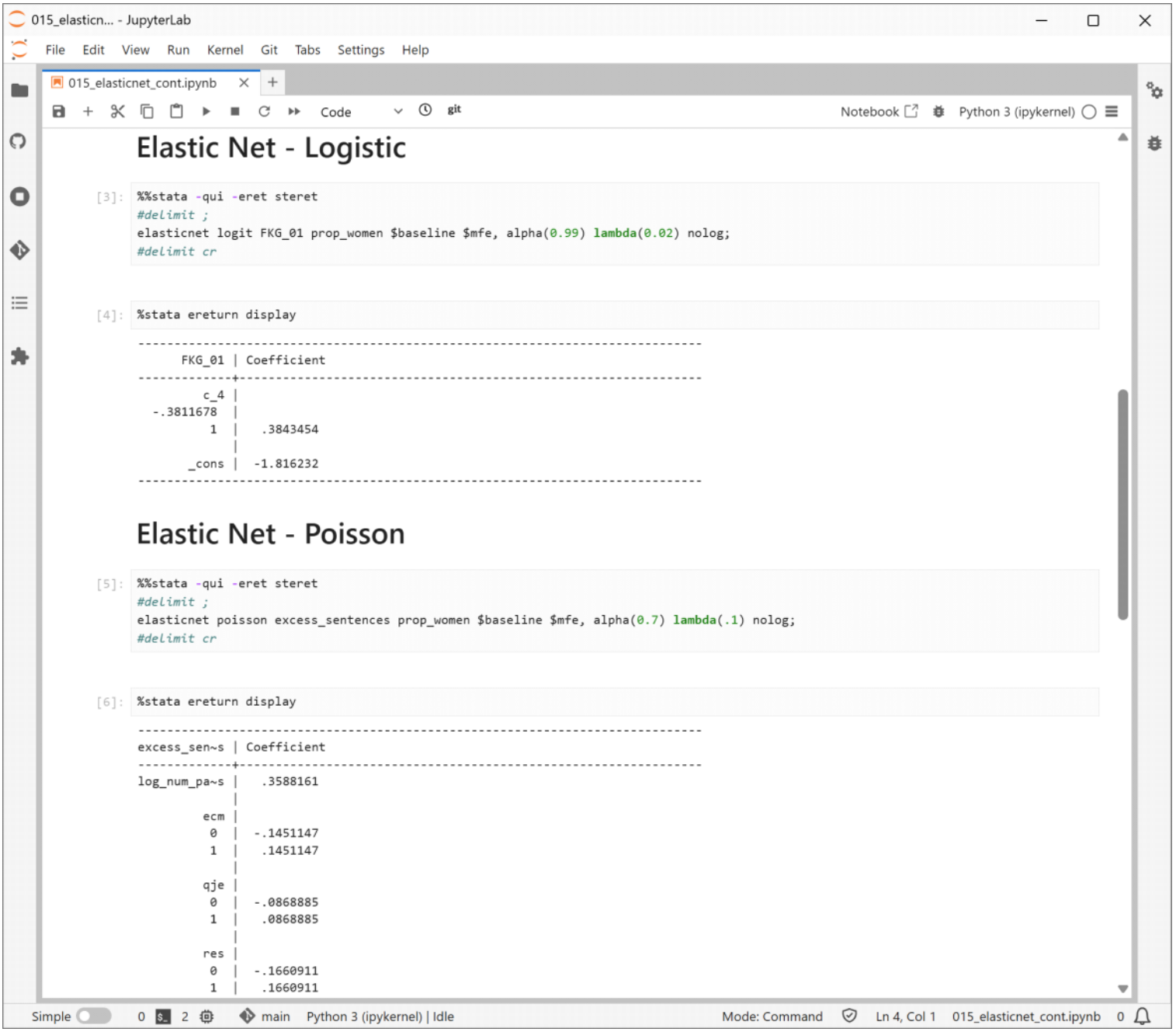
- ✔ This includes Lasso (α = 1) and
- ✔ ridge regression (α = 0) as special cases.

Equivalentely, we could write

$$\hat{\beta} = \underset{b}{\text{argmin}}\ Q_n(b) + \lambda\big((1-\alpha)\parallel b \parallel_2^2 + \alpha \parallel b \parallel_1\big).$$

where $Q_n(b) \equiv \sum_{i=1}^n \big(Y_i - X_i'b\big)^2$ when the model is *linear*.

Model	$Q_n(b)$.
Linear	$\sum_{i=1}^n \big(Y_i - X_i'b\big)^2$
Logit	$\sum_{i=1}^n -Y_i\big(X_i'b\big) + \log\{1+\exp(X_i'b)\}$
Probit	$\sum_{i=1}^n -Y_i \times \log\{\Phi\big(X_i'b\big)\} - \big(1-Y_i\big)\log\{1-\Phi(X_i'b)\}$
Poisson	$\sum_{i=1}^n -Y_i\big(X_i'b\big) + \exp\{X_i'b\}$


015_elasticn... - JupyterLab
File Edit View Run Kernel Git Tabs Settings Help
015_elasticnet_cont.ipynb
Code
git
Notebook
Python 3 (ipykernel)
Elastic Net - Logistic
[3]:
%%stata -qui -eret steret
#delimit ;
elasticnet logit FKG_01 prop_women $baseline $mfe, alpha(0.99) lambda(0.02) nolog;
#delimit cr
[4]:
%stata ereturn display
FKG_01 | Coefficient
c_4 |
-.3811678 |
1 | .3843454
_cons | -1.816232
Elastic Net - Poisson
[5]:
%%stata -qui -eret steret
#delimit ;
elasticnet poisson excess_sentences prop_women $baseline $mfe, alpha(0.7) lambda(.1) nolog;
#delimit cr
[6]:
%stata ereturn display
excess_sen~s | Coefficient
log_num_pa~s | .3588161
ecm |
0 | -.1451147
1 | .1451147
qje |
0 | -.0868885
1 | .0868885
res |
0 | -.1660911
1 | .1660911
Simple 0 2 main Python 3 (ipykernel) | Idle Mode: Command Ln 4, Col 1 015_elasticnet_cont.ipynb 0

Binary Response

When Y is **binary**, meaning that it takes two values. Without loss of generality these are taken as zero and one, thus Y has support $\{0,1\}$. In econometrics we typically call this class of models **binary choice**.

Examples:	
✔ Purchase of a single item ✔ Market entry ✔ Approval of an application ✔ Participation in a program	👉 The dependent variable may be recorded as Yes/No, True/False, or 1/-1, *but* can always be written as 1/0.

Binary Choice Models

Let (Y, X') be random with $Y \in \{0,1\}$ and $X \in \mathbb{R}^k$. The response probability of Y with respect to X is

$P(x) = \Pr[Y = 1 \mid X = x] = \mathbb{E}[Y \mid X = x]$.

The response probability completely describes the conditional distribution of $Y|X$. The marginal effect (X continuous) is

$$\frac{\partial}{\partial x} P(x) = \frac{\partial}{\partial x} \Pr[Y = 1 \mid X = x] = \frac{\partial}{\partial x} \mathbb{E}[Y \mid X = x].$$

This is the regression derivative.

⚠ Linear Regression ⚠

The variables satisfy the regression framework

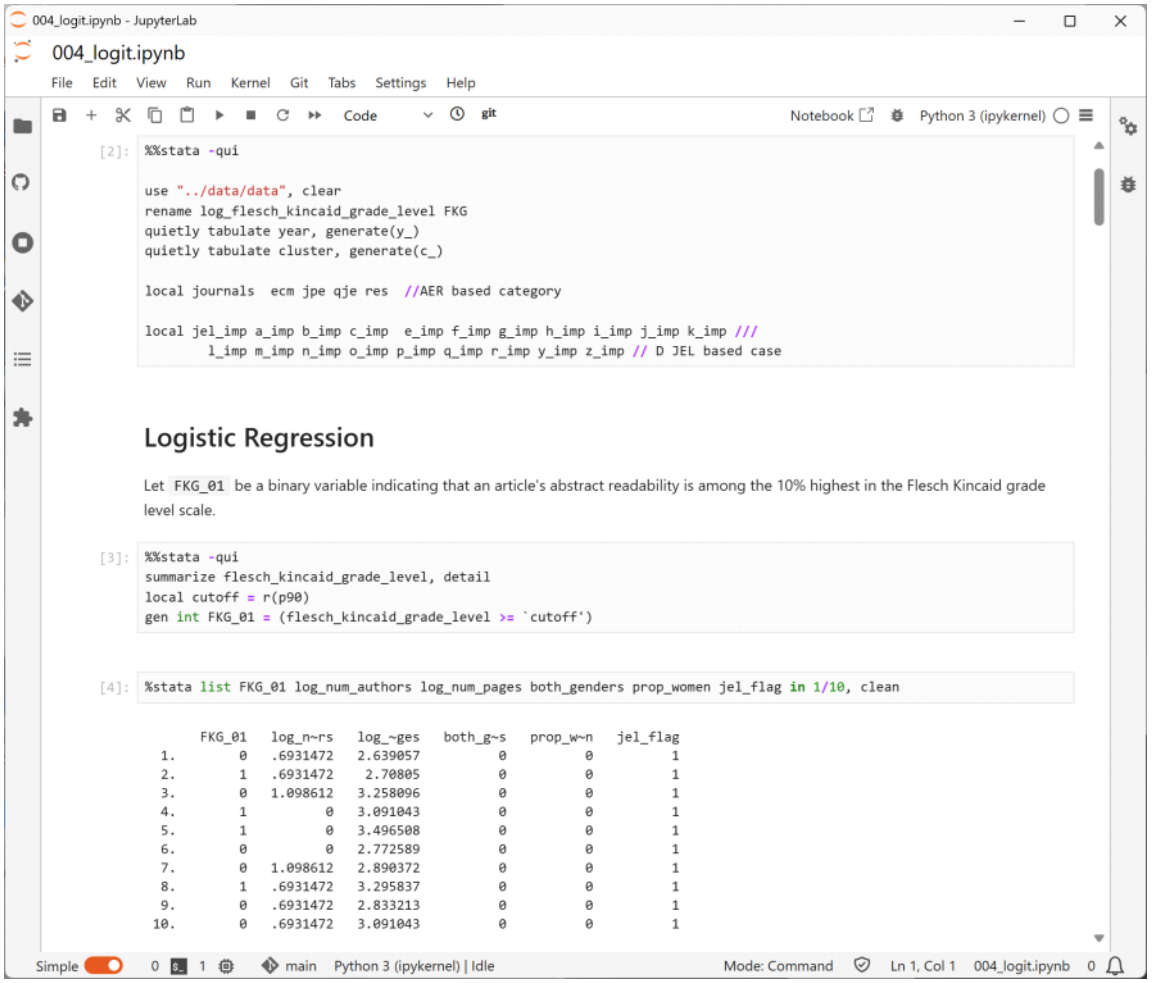
$$\begin{aligned} Y &= P(X) + e \\ \mathbb{E}[e \mid X] &= 0 \end{aligned}.$$

The error e is not "*classical*." It has the two-point conditional distribution

$$e = \begin{cases} 1 - P(X), & \text{with probability } P(X) \\ -P(X), & \text{with probability } 1 - P(X). \end{cases}$$

It is also highly heteroskedastic with conditional variance

$$\operatorname{var}[e \mid X] = P(X)(1 - P(X)).$$


004_logit.ipynb - JupyterLab
004_logit.ipynb
File Edit View Run Kernel Git Tabs Settings Help
Code
Notebook
Python 3 (ipykernel)
[2]:
%%stata -qui
use "../data/data", clear
rename log_flesch_kincaid_grade_level FKG
quietly tabulate year, generate(y_)
quietly tabulate cluster, generate(c_)
local journals ecm jpe qje res //AER based category
local jel_imp a_imp b_imp c_imp e_imp f_imp g_imp h_imp i_imp j_imp k_imp ///
l_imp m_imp n_imp o_imp p_imp q_imp r_imp y_imp z_imp // D JEL based case
Logistic Regression
Let FKG_01 be a binary variable indicating that an article's abstract readability is among the 10% highest in the Flesch Kincaid grade level scale.
[3]:
%%stata -qui
summarize flesch_kincaid_grade_level, detail
local cutoff = r(p90)
gen int FKG_01 = (flesch_kincaid_grade_level >= `cutoff')
[4]:
%stata list FKG_01 log_num_authors log_num_pages both_genders prop_women jel_flag in 1/10, clean
FKG_01 log_n~rs log_~ges both_g~s prop_w~n jel_flag
1. 0 .6931472 2.639057 0 0 1
2. 1 .6931472 2.70805 0 0 1
3. 0 1.098612 3.258096 0 0 1
4. 1 0 3.091043 0 0 1
5. 1 0 3.496508 0 0 1
6. 0 0 2.772589 0 0 1
7. 0 1.098612 2.890372 0 0 1
8. 1 .6931472 3.295837 0 0 1
9. 0 .6931472 2.833213 0 0 1
10. 0 .6931472 3.091043 0 0 1
Simple
main
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 1
004_logit.ipynb

Models

Linear Probability Model: $P(x) = x'\beta$ where $\beta \in \mathbb{R}^k$ is a coefficient vector.

Advantages:	Disadvantage:
• The coefficients β equal the marginal effects (when X does not include nonlinear transformations) • Linearity means that estimation is simple as least squares can be used to estimate the coefficients.	• It does not respect the [0,1] boundary. Fitted and predicted values from estimated linear probability models frequently violate these boundaries producing nonsense results.

Index Model: $P(x) = G(x'\beta)$ where $G(u)$ is a *link* function and β is a coefficient vector.

This framework is also called a **single index** model where $x'\beta$ is a **linear index** function.

In binary choice models, $G(u)$ is a **distribution function** which respects the probability bounds $0 \leq G(u) \leq 1$.

$G(\cdot)$ is *symmetric*, i.e., $G(-u) = 1 - G(u)$.

$\mathrm{g}(u) = \frac{\partial}{\partial u} G(u)$ denote the density function of $G(u)$.

Probit Model: $P(x) = \Phi(x'\beta)$ where $\Phi(u)$ is the standard normal distribution function.	**Logit Model**: $P(x) = \Lambda(x'\beta)$ where $\Lambda(u) = (1 + \exp(-u))^{-1}$ is the logistic distribution function.

Latent Variable Interpretation

An index model can be interpreted as a latent variable model. Consider

$$Y^* = X'\beta + e$$
$$e \sim G(e)$$
$$Y = \mathbb{1}\{Y^* > 0\} = \begin{cases} 1 & \text{if } Y^* > 0 \\ 0 & \text{otherwise} \end{cases}.$$

The event $Y = 1$ is the same as $Y^* > 0$, which is the same as

$$X'\beta + e > 0.$$

This means that the response probability is

$$P(X) = \Pr[e > -X'\beta | X] = 1 - G(-X'\beta) = G(X'\beta).$$

You may have noticed that we have discussed cases where the error e is either *standard* normal or *standard* logistic, that is, their scale is fixed.

This is because the scale of the error distribution is not *identified*.

To see this, suppose that $e = \sigma\varepsilon$ where ε has a distribution $G^*(u)$ with unit variance. Then the response probability is

$$\Pr[Y = 1 \mid X = x] = \Pr[\sigma\varepsilon > x'\beta | X = x] = G^*\left(\frac{x'\beta}{\sigma}\right) = G^*(x'\beta^*),$$

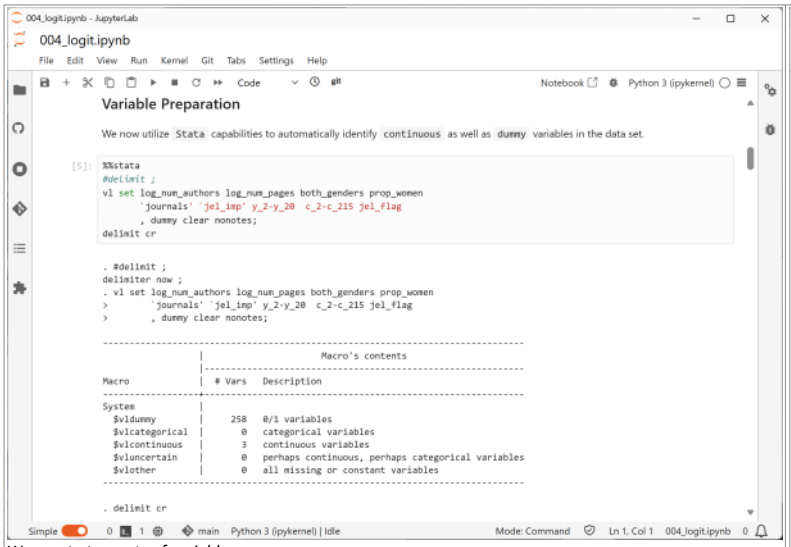
where $\beta^* = \beta/\sigma$. This is an index model with coefficient β^*.

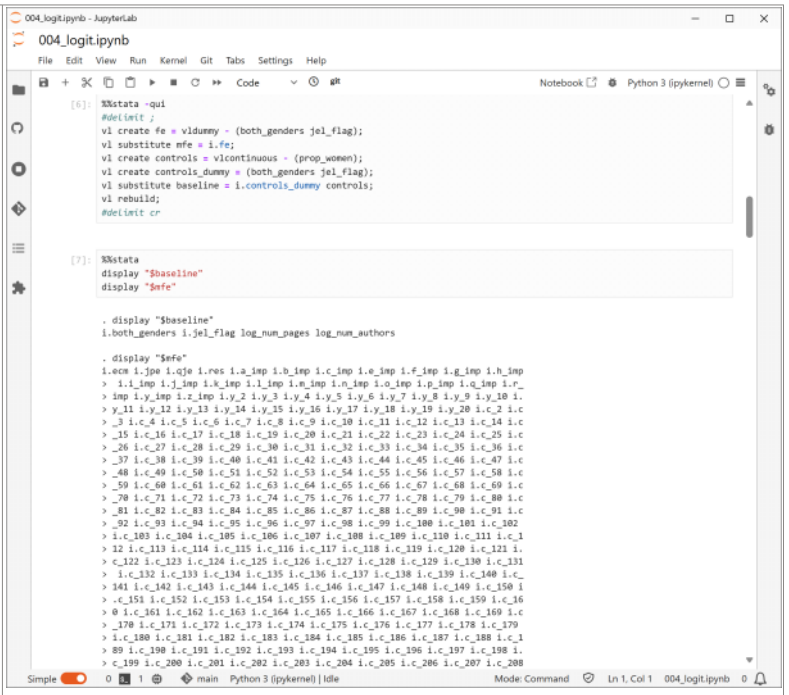
This means that β and σ are not separately identified.

While the coefficient β is not identified the following parameters are identified:

1. Scaled coefficients: $\beta^* = \beta/\sigma$.
2. Ratios of coefficients: $\beta_1/\beta_2 = \beta_1^*/\beta_2^*$.
3. Marginal effects: $\frac{\partial}{\partial x}P(x) = \frac{\beta}{\sigma}g\left(\frac{x'\beta}{\sigma}\right) = \beta^* g(x'\beta^*)$.

These only depend on β^* so are identified.





We create two sets of variables:

- baseline: `log_num_pages`, `log_num_authors`, `both_genders` (dummy)
- `mde`: Set of dummies for Journals, JEL Codes, Cluster, and Years.

Estimation - Likelihood

Recall that if Y is Bernoulli, such that $\Pr[Y=1]=p$ and $\Pr[Y=0]=1-p$, then Y has the probability mass function

$$\pi(y)=p^y(1-p)^{1-y},\ y\in\{0,1\}.$$

In the index model $\Pr[Y=1\mid X]=G(X'\beta)$, Y is conditionally Bernoulli, so its conditional probability mass function is

$$\pi(Y\mid X)=G(X'\beta)^Y\left(1-G(X'\beta)\right)^{1-Y}=G(X'\beta)^Y G(-X'\beta)^{1-Y}=G(Z'\beta),$$

where

$$Z=\begin{cases}X & ;Y=1\\ -X & ;Y=0\end{cases}.$$

Taking natural logarithm and summing across observations we obtain:

		Probit	Logit
Population:	$\ell(\beta)=\mathbb{E}[\log G(Z'\beta)]$	$\ell^{\text{probit}}(\beta)=\mathbb{E}[\log\Phi(Z'\beta)]$	$\ell^{\text{logit}}(\beta)=\mathbb{E}[\log\Lambda(Z'\beta)]$
log-likelihood function:	$\ell_n(\beta)=n^{-1}\sum_{i=1}^n \log G(Z_i'\beta)$	$\ell_n^{\text{probit}}(\beta)=n^{-1}\sum_{i=1}^n \log\Phi(Z_i'\beta)$	$\ell_n^{\text{logit}}(\beta)=n^{-1}\sum_{i=1}^n \log\Lambda(Z_i'\beta)$
First derivative:	$h(x)=\frac{d}{dx}\log G(x)$ $S_n(\beta)=\frac{\partial}{\partial\beta}\ell_n(\beta)$ $=n^{-1}\sum_{i=1}^n Z_i h(Z_i'\beta)$	$h_{\text{probit}}(x)=\frac{\phi(x)}{\Phi(x)}\overset{\text{def}}{=}\lambda(x)$ Inverse Mills Ratio	$h_{\text{logit}}(x)=1-\Lambda(x)$
Second derivative:	$H(x)=-\frac{d^2}{dx^2}\log G(x)$ $\mathcal{H}_n(\beta)=-\frac{\partial^2}{\partial\beta\,\partial\beta'}\ell_n(\beta)$ $=n^{-1}\sum_{i=1}^n X_iX_i'H(Z_i'\beta)$	$H_{\text{probit}}(x)=\lambda(x)(x+\lambda(x))$	$H_{\text{logit}}(x)=\Lambda(x)(1-\Lambda(x))$

The MLE is the value which maximizes $\ell_n(\beta)$. We write this as

$$\hat\beta^{\text{probit}}=\underset{\beta}{\text{argmax}}\,\ell_n^{\text{probit}}(\beta)$$
$$\hat\beta^{\text{logit}}=\underset{\beta}{\text{argmax}}\,\ell_n^{\text{logit}}(\beta).$$

The *pseudo* true values are defined as

$$\beta^{\text{probit}}=\underset{\beta}{\text{argmax}}\,\ell^{\text{probit}}(\beta)$$
$$\beta^{\text{logit}}=\underset{\beta}{\text{argmax}}\,\ell^{\text{logit}}(\beta).$$

Estimation - Asymptotic Distribution

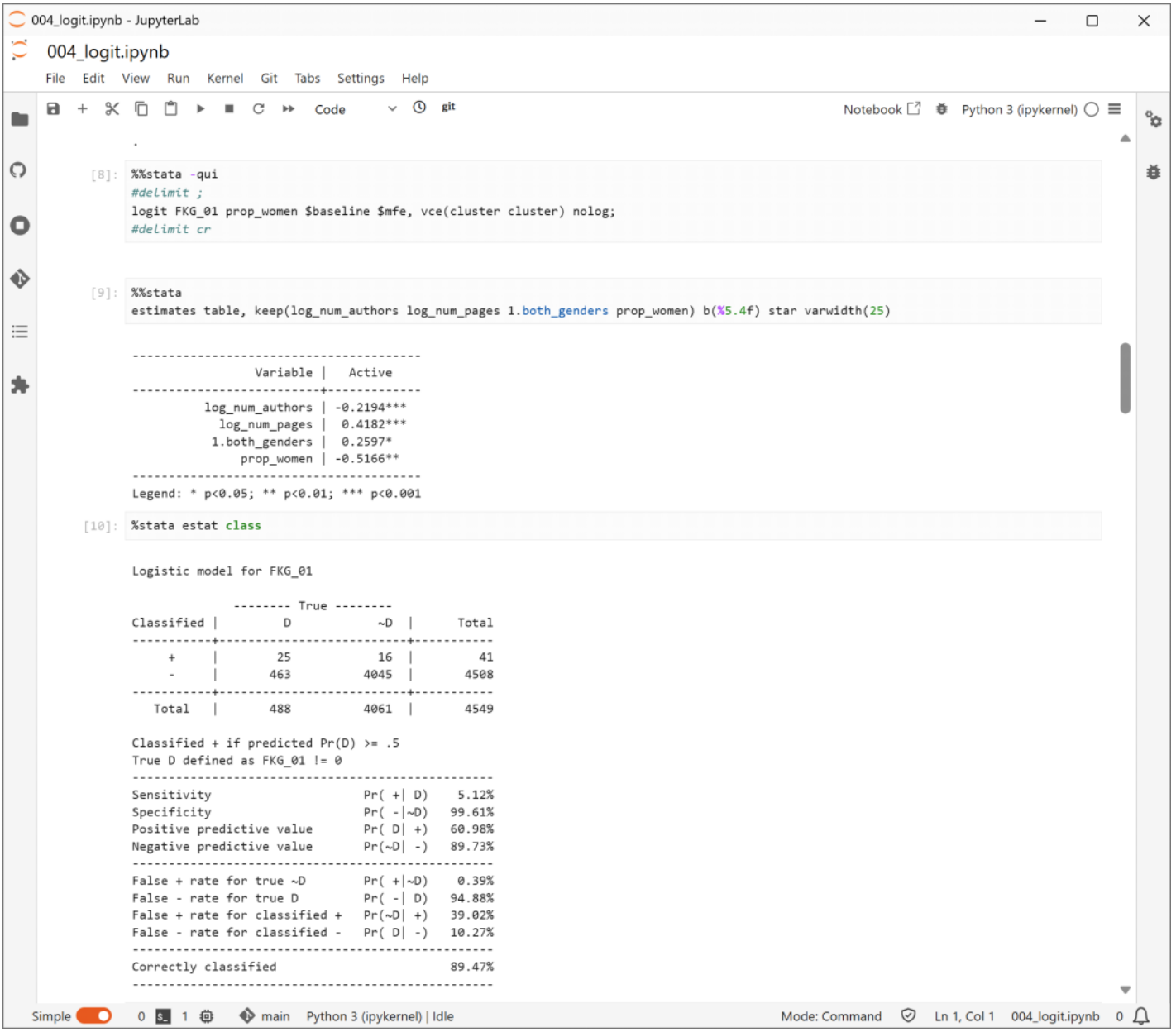
Let B be the parameter space for β

Probit		Logit
Theorem (Consistency): If $(1)(Y_i, X_i)$ are i.i.d.; (2) $\mathbb{E} \parallel X \parallel^2 < \infty$; (3) $\boldsymbol{Q}_{\mathbf{probit}} > 0$; and $(4) B$ is compact; then $\hat{\beta}^{\text{probit}} \underset{p}{\longrightarrow} \beta^{\text{probit}}$ as $n \to \infty$.		**Theorem (Consistency):** If $(1)(Y_i, X_i)$ are i.i.d.; (2) $\mathbb{E} \parallel X \parallel < \infty$; (3) $\boldsymbol{Q}_{\mathbf{logit}} > 0$; and (4) B is compact; then $\hat{\beta}^{\text{logit}} \underset{p}{\longrightarrow} \beta^{\text{logit}}$ as $n \to \infty$.
$\boldsymbol{Q}_{\mathbf{probit}} \overset{\text{def}}{=} \mathbb{E}\left[XX' H_{\text{probit}}(X'\beta^{\text{probit}})\right] > 0,$		$\boldsymbol{Q}_{\mathbf{logit}} \overset{\text{def}}{=} \mathbb{E}\left[XX' \Lambda(X'\beta^{\text{logit}})\left(1 - \Lambda(X'\beta^{\text{logit}})\right)\right] > 0,$
$\Omega_{\text{probit}} = \mathbb{E}\left[XX' \lambda(X'\beta^{\text{probit}})^2\right],$		$\Omega_{\text{logit}} = \mathbb{E}\left[XX'\left(Y - \Lambda(X'\beta^{\text{logit}})\right)^2\right].$
Theorem (Asy. Normality): If the conditions of consistency hold plus $\mathbb{E} \parallel X \parallel^4 < \infty$ and β^{probit} is in the interior of B; then as $n \to \infty$ $\sqrt{n}\left(\hat{\beta}^{\text{probit}} - \beta^{\text{probit}}\right) \underset{d}{\longrightarrow} \mathrm{N}(0, \boldsymbol{V}),$ where $\boldsymbol{V}_{\text{probit}} = \boldsymbol{Q}_{\mathbf{probit}}^{-1} \Omega_{\text{probit}} \boldsymbol{Q}_{\mathbf{probit}}^{-1}$.		**Theorem (Asy. Normality):** If the conditions of consistency hold plus $\mathbb{E} \parallel X \parallel^4 < \infty$ and β^{logit} is in the interior of B; then as $n \to \infty$ $\sqrt{n}\left(\hat{\beta}^{\text{logit}} - \beta^{\text{logit}}\right) \underset{d}{\longrightarrow} \mathrm{N}(0, V),$ where $\boldsymbol{V}_{\text{logit}} = \boldsymbol{Q}_{\mathbf{logit}}^{-1} \Omega_{\text{logit}} \boldsymbol{Q}_{\mathbf{logit}}^{-1}$.

Standard Errors

The asymptotic standard errors of $\hat{\beta}^{\text{probit}}$ and $\hat{\beta}^{\text{logit}}$ are defined as the squared-root of the main diagonal element of $n^{-1}\widehat{\boldsymbol{V}}_{\text{probit}}$ and $n^{-1}\widehat{\boldsymbol{V}}_{\text{logit}}$, where

$\widehat{\boldsymbol{V}}_{\text{probit}} = \widehat{\boldsymbol{Q}}_{\mathbf{probit}}^{-1} \widehat{\Omega}_{\text{probit}} \widehat{\boldsymbol{Q}}_{\mathbf{probit}}^{-1},$		$\widehat{\boldsymbol{V}}_{\text{logit}} = \widehat{\boldsymbol{Q}}_{\mathbf{logit}}^{-1} \widehat{\Omega}_{\text{logit}} \widehat{\boldsymbol{Q}}_{\mathbf{logit}}^{-1},$
$\widehat{\boldsymbol{Q}}_{\mathbf{probit}} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \hat{\lambda}_i(\hat{\mu}_i + \hat{\lambda}_i)$, where $\hat{\mu}_i = X_i'\hat{\beta}^{\text{probit}}$, $\hat{\lambda}_i = \lambda(\hat{\mu}_i),$		$\widehat{\boldsymbol{Q}}_{\text{logit}} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \widehat{\Lambda}_i(1 - \widehat{\Lambda}_i)$, where $\widehat{\Lambda}_i = \Lambda\left(X_i'\hat{\beta}^{\text{logit}}\right),$
$\widehat{\Omega}_{\text{probit}} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \hat{\lambda}_i^2.$		$\widehat{\Omega}_{\text{logit}} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i' (Y_i - \widehat{\Lambda}_i)^2.$

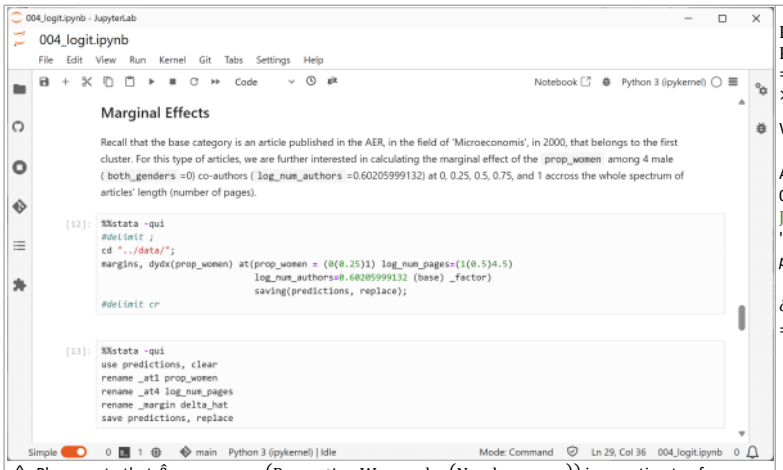

004_logit.ipynb - JupyterLab
004_logit.ipynb
File Edit View Run Kernel Git Tabs Settings Help
Code
git
Notebook
Python 3 (ipykernel)
[8]:
%%stata -qui
#delimit ;
logit FKG_01 prop_women $baseline $mfe, vce(cluster cluster) nolog;
#delimit cr
[9]:
%%stata
estimates table, keep(log_num_authors log_num_pages 1.both_genders prop_women) b(%5.4f) star varwidth(25)
Variable | Active
log_num_authors | -0.2194***
log_num_pages | 0.4182***
1.both_genders | 0.2597*
prop_women | -0.5166**
Legend: * p<0.05; ** p<0.01; *** p<0.001
[10]:
%stata estat class
Logistic model for FKG_01
True
Classified | D ~D | Total
+ | 25 16 | 41
- | 463 4045 | 4508
Total | 488 4061 | 4549
Classified + if predicted Pr(D) >= .5
True D defined as FKG_01 != 0
Sensitivity Pr(+| D) 5.12%
Specificity Pr(-|~D) 99.61%
Positive predictive value Pr(D| +) 60.98%
Negative predictive value Pr(~D| -) 89.73%
False + rate for true ~D Pr(+|~D) 0.39%
False - rate for true D Pr(-| D) 94.88%
False + rate for classified + Pr(~D| +) 39.02%
False - rate for classified - Pr(D| -) 10.27%
Correctly classified 89.47%
Simple
main
Python 3 (ipykernel) | Idle
Mode: Command
Ln 1, Col 1
004_logit.ipynb

Marginal Effects

Recall that $\Pr[Y = 1 \mid X = x] = G(x'\beta) = G(\beta_1 x_1 + \cdots + \beta_j x_j + \cdots + \beta_{k-1} x_{k-1} + \beta_k)$

	x_j is continuous	x_j is binary
Marginal Effects:	$\delta_j(x) = \beta_j g(x'\beta)$	$\delta_j(x) = G\left(\beta_1 x_1 + \cdots + \beta_j \cdot 1 + \cdots + \beta_{k-1} x_{k-1} + \beta_k\right) - G(\beta_1 x_1 + \cdots + \beta_j \cdot 0 + \cdots + \beta_{k-1} x_{k-1} + \beta_k)$
Average Marginal Effects	$\text{AME} = \mathbb{E}[\delta_j(X)] = \beta_j \mathbb{E}[g(X'\beta)]$	$\text{AME} = \mathbb{E}[\delta_j(X)]$

Estimator: $\widehat{\text{AME}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\delta}_j(X_i)$, where $\hat{\delta}_j(X_i) = \hat{\beta}_j g(X_i'\hat{\beta})$ or $\hat{\delta}_j(X_i) = G\left(\hat{\beta}_1 X_{1,i} + \cdots + \hat{\beta}_j \cdot 1 + \cdots + \hat{\beta}_{k-1} X_{k-1,i} + \hat{\beta}_k\right) - G(\hat{\beta}_1 X_{1,i} + \cdots + \hat{\beta}_j \cdot 0 + \cdots + \hat{\beta}_{k-1} X_{k-1,i} + \hat{\beta}_k)$.



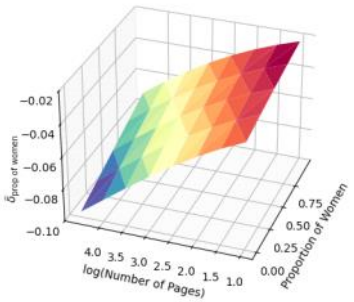
⚠ Please note that $\hat{\delta}_{prop\ of\ women}(\text{Proportion Women}, \log(\text{Number pages}))$ is an *estimate* of $\delta_{prop\ of\ women}(\text{Proportion Women}, \log(\text{Number pages}))$ and therefore one also needs to report $s.e.(\hat{\delta}_{prop\ of\ women}(\text{Proportion Women}, \log(\text{Number pages}))$ which are calculated via the *Delta Method*.

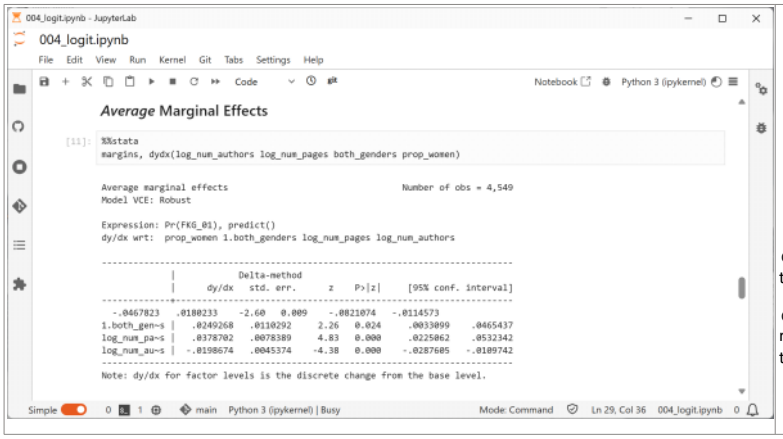
$\Pr[\text{FKG\_01}_a | \log(\text{Number authors})_a, \log(\text{Number pages})_a, \text{Both genders}_a, \text{Proportion Women}_a, \text{Journals}, \text{JEL Codes}, \text{Cluster}, \text{Years}, \text{JEL flag}]$
$= G(\beta_1 \times \log(\text{Number authors})_a + \beta_2 \times \log(\text{Number pages})_a + \beta_3 \times \text{Both genders}_a + \theta \times \text{Proportion Women}_a + \text{Journals} + \text{JEL codes} + \text{Cluster} + \text{Years} + \text{JEL flag} + \text{cons})$

We are interested in the marginal effect of education on the probability of participating in the labor force.

After a Logit regression, one obtains $[\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\theta}, \ldots, \hat{\beta}_{262}]$, we proceed to set $\log(\text{Number authors}) = 0.60205999132$ (Number authors = 4), Both genders $= 0$, and all the sets of dummies in Journals, JEL Codes, Cluster, Years, and JEL flag equal to 0, i.e., an article published in 2000, in the field of 'Microeconomics,' without author-provided JEL codes, published in the AER. Then the *marginal effect* of *prop_women* is

$\hat{\delta}_{prop\ of\ women}(\text{Proportion Women}, \log(\text{Number pages}))$
$= \hat{\theta} g(\hat{\beta}_1 \times 0.60205999132 + \hat{\beta}_2 \log(\text{Number pages}) + \hat{\theta} \times \text{Proportion Women} + \hat{\beta}_{262}) + \hat{\theta}$





☞ On average if the proportion of women authors increases by 1, the probability of it to be very difficult to read decreases by -.047.

☞ On average the probability for an article to be very difficult to read when it is written by mixed-gender co-authors is 0.025 higher than single-gender authors holding everything else constant.

Multinomial Response

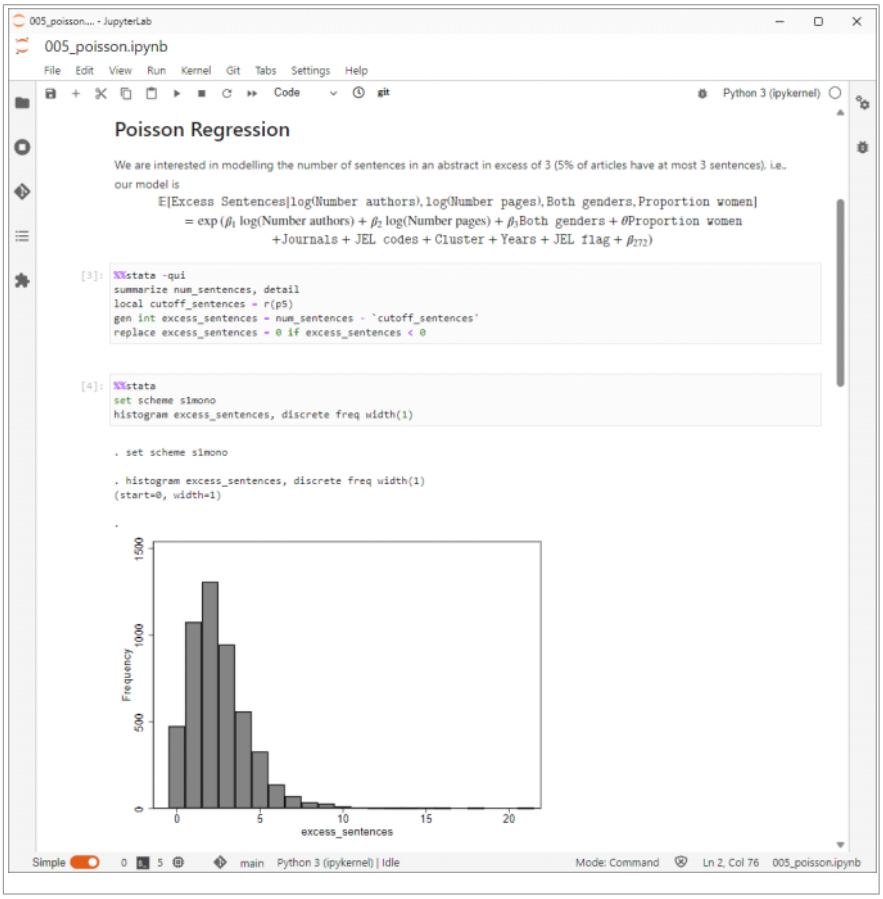
A **multinomial** random variable Y takes values in a finite set, typically written as $Y \in \{0,1,2, \ldots, J\}$. We typically describe the pair (Y, X') as **multinomial response** when Y is multinomial and $X \in \mathbb{R}^k$ are regressors. The conditional distribution of Y given X is summarized by the response probability

$$P_j(x) = \mathrm{Pr}[Y = j \mid X = x].$$

Poisson Regression

Count data refers to situations where the dependent variable is the number of "events" recorded as positive integers $Y \in \{0,1,2,\ldots\}$

Examples:	A count data model specifies the response probabilities $P_j(X) = \Pr[Y = j \mid X]$ for $j = 0,1,2,\ldots$, with the property $\sum_{j=0}^{\infty} P_j(X) = 1$.
✔ Number of doctor visits ✔ Number of accidents ✔ Number of patent registrations ✔ Number of absences	



Poisson Regression:

$P_j(X) = \dfrac{\exp(-\lambda(X))\lambda(X)^j}{j!}$ $\lambda(X) = \exp(X'\beta)$	Recall that the Poisson distribution has the property that its mean and variance equal the Poisson parameter λ. Thus (conditionally on X) $\mathbb{E}[Y \mid X] = \exp(X'\beta)$ $\text{var}[Y \mid X] = \exp(X'\beta)$.

Estimator:		The MLE $\hat{\beta}$ is the value β which maximizes $\ell_n(\beta)$.
Log-likelihood Function:		$\ell_n(\beta) = \sum_{i=1}^{n} \log P_{Y_i}(X_i \mid \beta) \overset{\text{iid}}{=} \sum_{i=1}^{n} \left(-\exp(X_i'\beta) + Y_i X_i'\beta - \log(Y_i!)\right)$

1st & 2nd Derivatives:	$\frac{\partial}{\partial\beta}\ell_n(\beta) = \sum_{i=1}^{n} X_i\left(Y_i - \exp\left(X_i'\beta\right)\right)$ $\frac{\partial^2}{\partial\beta\,\partial\beta'}\ell_n(\beta) = -\sum_{i=1}^{n} X_i X_i' \exp\left(X_i'\beta\right)$

Coefficients Interpretation (*Continuous Controls*)

Let us rewrite

$$\mathbb{E}[\,Y \mid X\,] = \exp\left(X'\beta\right) = \exp(\beta_1 X_1 + \cdots + \beta_{k-1}X_{k-1} + \beta_k),$$

then by taking logs one has

$$\log\left\{\mathbb{E}\left[\,Y \mid X_1, \ldots, X_{k-1}\,\right]\right\} = \beta_1 X_1 + \cdots + \beta_{k-1}X_{k-1} + \beta_k.$$

Let $X_j \in \mathbb{R}$, then

$$\log\left\{\mathbb{E}\left[\,Y \mid X_1, \ldots, X_j + \Delta X_j, \ldots, X_{k-1}\,\right]\right\} - \log\left\{\mathbb{E}\left[\,Y \mid X_1, \ldots, X_j, \ldots, X_{k-1}\,\right]\right\} = \beta_j \Delta X_j,$$

$$\%\Delta\mathbb{E}\left[\,Y \mid X_1, \ldots, X_{k-1}\,\right] \approx (100\beta_j)\Delta X_j.$$

$100\beta_j$ is roughly the percentage change in $\mathbb{E}[\,Y \mid X\,]$, given a one-unit increase in X_j.

If $X_j = \log(Z_j)$, for some variable $Z_j > 0$, then $\Delta X_j = \log\left(Z_{j;1}\right) - \log\left(Z_{j;0}\right) = \Delta\log(Z_j)$, so $\%\Delta\mathbb{E}\left[\,Y \mid X_1, \ldots, X_{k-1}\,\right] \approx \left(100\beta_j\right)\Delta X_j = \beta_j\left(100 \cdot \Delta\log\left(Z_j\right)\right) \approx \beta_j \%\Delta Z_j$ and therefore

$$\beta_j \approx \%\Delta\mathbb{E}\left[\,Y \mid X_1, \ldots, \log(Z_j), \ldots, X_{k-1}\,\right] / \%\Delta Z_j.$$

β_j is roughly the *elasticity* of the expected value of Y with respect to Z_j.

The **proportionate change** in v moving from v_0 to v_1 is

$$\frac{v_1 - v_0}{v_0} = \Delta v / v_0,$$

assuming of course that $v_0 \neq 0$. The **percentage change** in v moving from v_0 to v_1 is simply 100 times the proportionate change:

$$\%\Delta v = 100(\Delta v / v_0),$$

the notation "$\%\Delta v$" is read as "the percentage change in v." Let v_0 and v_1 be positive values, then, it can be shown that

$$\log(v_1) - \log(v_0) \approx \frac{v_1 - v_0}{v_0} = \Delta v / v_0,$$

for small changes in v. Therefore by writing $\Delta\log(v) = \log(v_1) - \log(v_0)$, then

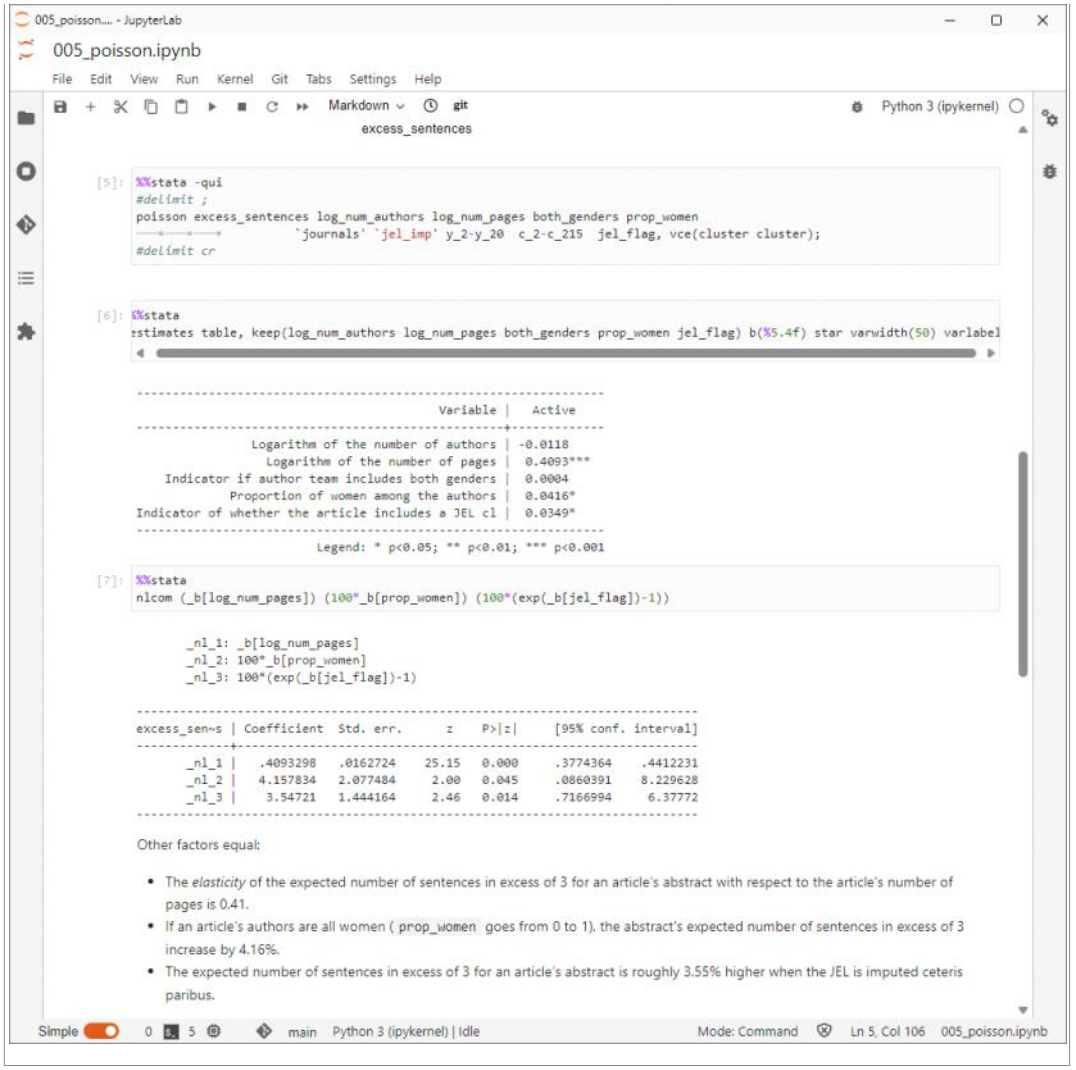
$$100 \cdot \Delta\log(v) \approx \%\Delta v.$$

Coefficient Interpretation (*Discrete Controls*)

Let $X_j \in \{0,1\}$, then the proportionate change in the expected value of Y from $X_j = 0$ to $X_j = 1$ is

$$\frac{\mathbb{E}\left[\,Y \mid X_1, \ldots, X_j = 1, \ldots, X_{k-1}\right]}{\mathbb{E}\left[\,Y \mid X_1, \ldots, X_j = 0, \ldots, X_{k-1}\right]} - 1 = \frac{\exp\left(\beta_1 X_1 + \cdots + \beta_j \cdot 1 + \cdots + \beta_{k-1}X_{k-1} + \beta_k\right)}{\exp\left(\beta_1 X_1 + \cdots + \beta_j \cdot 0 + \cdots + \beta_{k-1}X_{k-1} + \beta_k\right)} - 1 = \exp\left(\beta_j\right) - 1$$

$100 \cdot [\exp\left(\beta_j\right) - 1]$ is the percentage change in $\mathbb{E}[\,Y \mid X\,]$ from $X_j = 0$ to $X_j = 1$.


005_poisson.... - JupyterLab
005_poisson.ipynb
File Edit View Run Kernel Git Tabs Settings Help
Markdown
git
Python 3 (ipykernel)
excess_sentences
[5]:
%%stata -qui
#delimit ;
poisson excess_sentences log_num_authors log_num_pages both_genders prop_women
`journals' `jel_imp' y_2-y_20 c_2-c_215 jel_flag, vce(cluster cluster);
#delimit cr
[6]:
%%stata
estimates table, keep(log_num_authors log_num_pages both_genders prop_women jel_flag) b(%5.4f) star varwidth(50) varlabel
Variable | Active
Logarithm of the number of authors | -0.0118
Logarithm of the number of pages | 0.4093***
Indicator if author team includes both genders | 0.0004
Proportion of women among the authors | 0.0416*
Indicator of whether the article includes a JEL cl | 0.0349*
Legend: * p<0.05; ** p<0.01; *** p<0.001
[7]:
%%stata
nlcom (_b[log_num_pages]) (100*_b[prop_women]) (100*(exp(_b[jel_flag])-1))
_nl_1: _b[log_num_pages]
_nl_2: 100*_b[prop_women]
_nl_3: 100*(exp(_b[jel_flag])-1)
excess_sen~s | Coefficient Std. err. z P>|z| [95% conf. interval]
_nl_1 | .4093298 .0162724 25.15 0.000 .3774364 .4412231
_nl_2 | 4.157834 2.077484 2.00 0.045 .0860391 8.229628
_nl_3 | 3.54721 1.444164 2.46 0.014 .7166994 6.37772
Other factors equal:
The elasticity of the expected number of sentences in excess of 3 for an article's abstract with respect to the article's number of pages is 0.41.
If an article's authors are all women (prop_women goes from 0 to 1), the abstract's expected number of sentences in excess of 3 increase by 4.16%.
The expected number of sentences in excess of 3 for an article's abstract is roughly 3.55% higher when the JEL is imputed ceteris paribus.
Simple
0 5
main
Python 3 (ipykernel) | Idle
Mode: Command
Ln 5, Col 106
005_poisson.ipynb

HDFE & SCA

We are interested in estimating $\beta \equiv [\theta, \beta_2']'$ in the model

$$\mathbb{E}[Y|X_1, X_2, D_1, D_2, \cdots, D_F] = \exp(\theta X_1 + X_2'\beta_2 + D_1'\alpha_1 + D_2'\alpha_2 + \cdots + D_F'\alpha_F),$$

where $[D_1', D_2', \cdots, D_F']$ represents a row of $\mathbf{D} = [\mathbf{D}_1 \quad \mathbf{D}_2 \quad \cdots \quad \mathbf{D}_F]$ which consists of F indicator matrices.

Manuscript: https://arxiv.org/abs/1903.01690
Source Code: https://github.com/sergiocorreia/ppmlhdfe

arXiv:1903.01690v3 [econ.EM] 2 Aug 2019

PPMLHDFE: Fast Poisson Estimation with High-Dimensional Fixed Effects

Sergio Correia[1], Paulo Guimarães[2], and Tom Zylkin[3]

[1]Federal Reserve Board, sergio.a.correia@frb.gov
[2]Banco de Portugal, pfguimaraes@bportugal.pt
[3]University of Richmond, tzylkin@richmond.edu

August 5, 2019

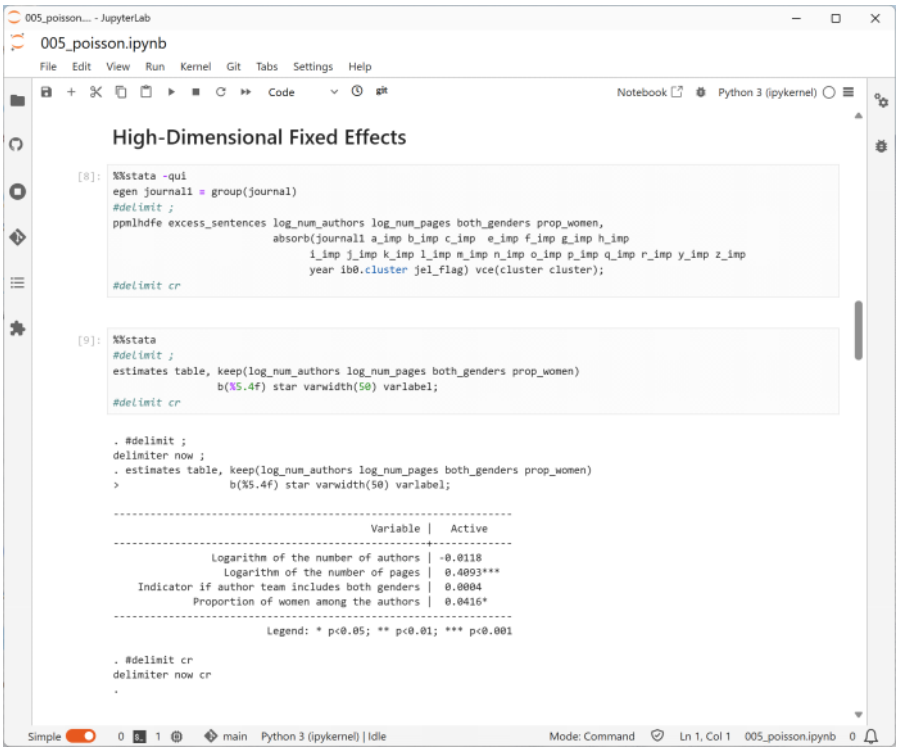
Abstract

In this paper we present **ppmlhdfe**, a new Stata command for estimation of (pseudo) Poisson regression models with multiple high-dimensional fixed effects (HDFE). Estimation is implemented using a modified version of the iteratively reweighted least-squares (IRLS) algorithm that allows for fast estimation in the presence of HDFE. Because the code is built around the **reghdfe** package, it has similar syntax, supports many of the same functionalities, and benefits from **reghdfe**'s fast convergence properties for computing high-dimensional least squares problems. Performance is further enhanced by some new techniques we introduce for accelerating HDFE-IRLS estimation specifically. **ppmlhdfe** also implements a novel and more robust approach to check for the existence of (pseudo) maximum likelihood estimates.

Keywords: **ppmlhdfe**, **reghdfe**, Poisson regression, high-dimensional fixed-effects

1 Introduction

Poisson regression is now well established as the standard approach to model count data. However, it is also gaining popularity as a viable alternative for estimation of multiplicative models where the dependent variable is nonnegative. Commonly, these models are estimated by linear regression applied to a log-transformed dependent variable. But, as with ordinary least squares (OLS), the only assumption required for consistency of the Poisson regression estimator is the correct specification of the conditional mean of the dependent variable (Gourieroux et al., 1984). In this setting, Poisson regression becomes Poisson pseudo maximum likelihood (PPML) regression. Gourieroux et al.'s results greatly extend the realm of application of Poisson regression because there is no



005_poisson.... - JupyterLab

005_poisson.ipynb

File Edit View Run Kernel Git Tabs Settings Help

Code Notebook Python 3 (ipykernel)

Specification Curve Analysis

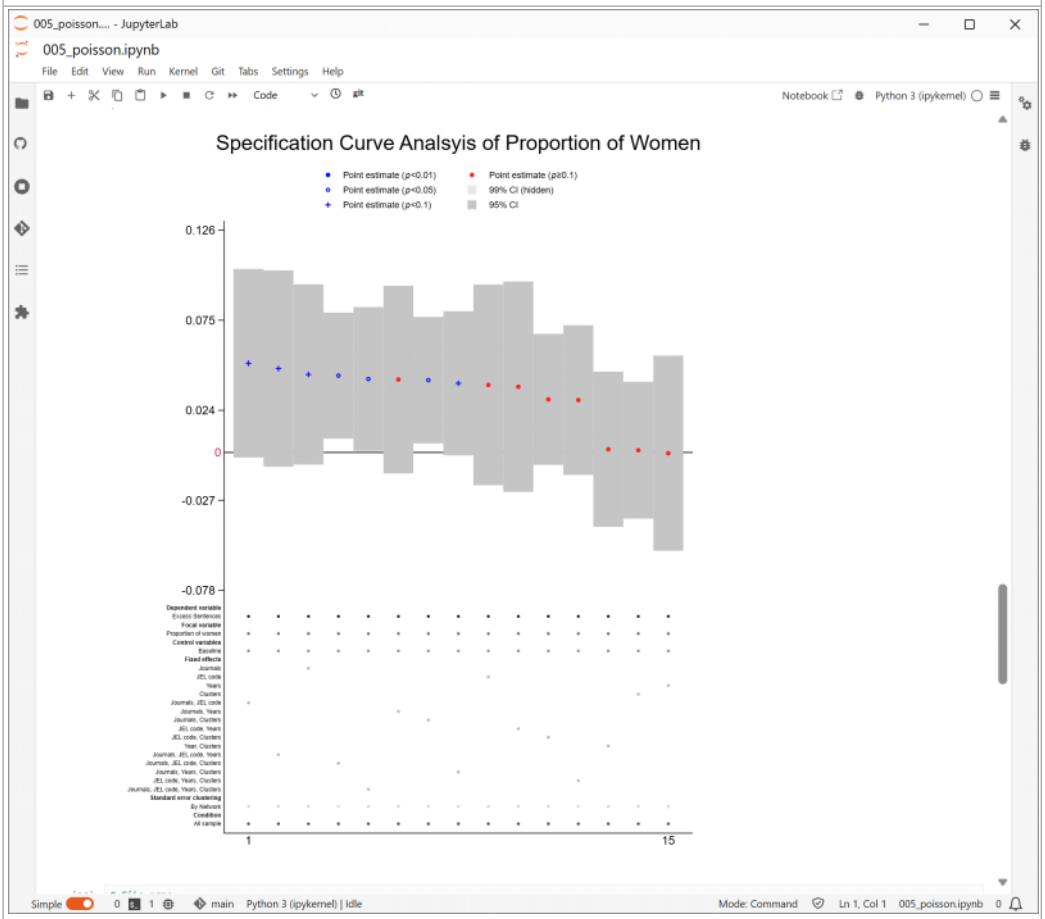
Clearing the data set in memory, reloading, and creating the `journal1` variable

[10]:
```
%%stata -qui

clear all
use "../data/data", clear
egen journal1 = group(journal)

summarize num_sentences, detail
local cutoff_sentences = r(p5)
gen int excess_sentences = num_sentences - `cutoff_sentences'
replace excess_sentences = 0 if excess_sentences < 0
```

Commenting out the `- log(F-K grade): log_flesch_kincaid_grade_level` line in the `.yml` file.

[11]:
```
import yaml

# File name
file_path = '../code/readability_graph_specurve.yml'

# Read, modify, and save the file
with open(file_path, 'r') as file:
    lines = file.readlines()

# Modify the target line
modified_lines = []
for line in lines:
    if line.strip() == "- log(F-K grade): log_flesch_kincaid_grade_level":
        modified_lines.append(f"#     {line}")  # Comment out the line
    else:
        modified_lines.append(line)

# Write the modified content back to the file
with open(file_path, 'w') as file:
    file.writelines(modified_lines)

print("Line commented successfully!")
```

Simple 0 1 main Python 3 (ipykernel) | Idle Mode: Command Ln 1, Col 1 005_poisson.ipynb 0



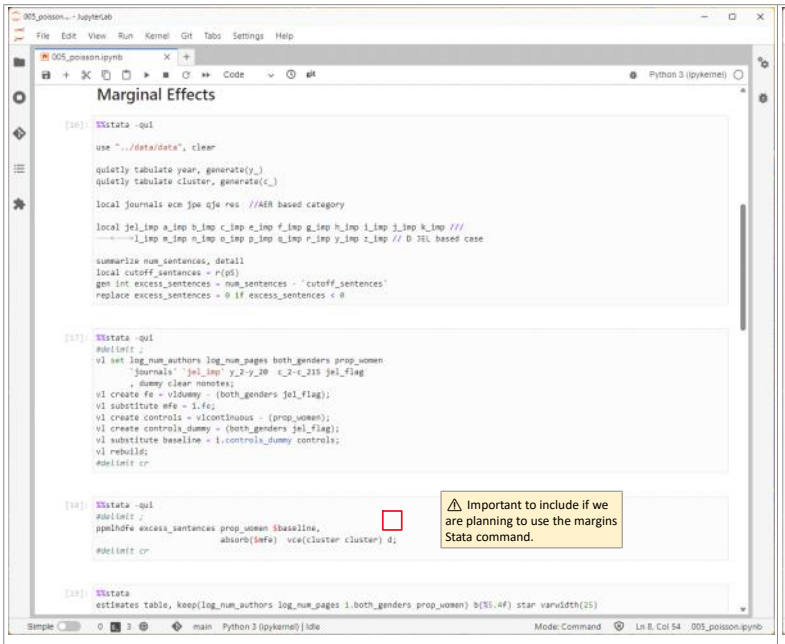
⚠ Notice that 8 out of 15 specifications yield insignificant (at 10%) effect of the variable of interest, `prop_women`, on the article's number of sentences in the abstract in excess of 3.

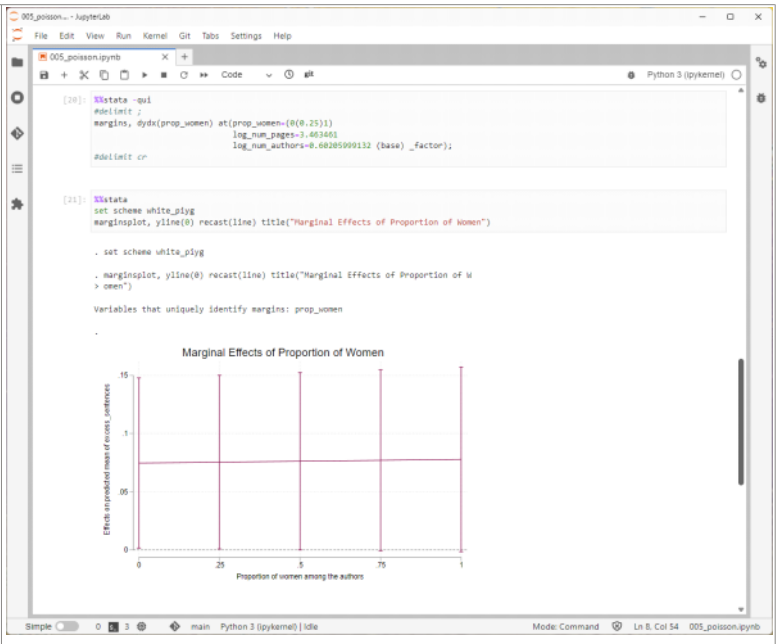
Marginal Effects

Recall that $\mathbb{E}[\,Y \mid X = x\,] = \exp(x'\beta) = \exp\,(\beta_1 x_1 + \cdots + \beta_j x_j + \cdots + \beta_{k-1} x_{k-1} + \beta_k)$

	x_j is continuous	x_j is binary
Marginal Effects:	$\delta_j(x) = \beta_j \exp(x'\beta)$	$\delta_j(x) = \exp\left(\beta_1 x_1 + \cdots + \beta_j \cdot 1 + \cdots + \beta_{k-1} x_{k-1} + \beta_k\right) - \exp(\beta_1 x_1 + \cdots + \beta_j \cdot 0 + \cdots + \beta_{k-1} x_{k-1} + \beta_k)$
Average Marginal Effects	$\text{AME} = \mathbb{E}[\delta_j(X)] = \beta_j \mathbb{E}\left[\exp\left(X'\beta\right)\right]$	$\text{AME} = \mathbb{E}[\delta_j(X)]$

Estimator: $\widehat{\text{AME}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\delta}_j(X_i)$, where $\hat{\delta}_j(X_i) = \hat{\beta}_j \exp(X_i'\hat{\beta}\,)$ or $\hat{\delta}_j(X_i) = \exp\left(\hat{\beta}_1 X_{1,i} + \cdots + \hat{\beta}_j \cdot 1 + \cdots + \hat{\beta}_{k-1} X_{k-1,i} + \hat{\beta}_k\right) - G(\hat{\beta}_1 X_{1,i} + \cdots + \hat{\beta}_j \cdot 0 + \cdots + \hat{\beta}_{k-1} X_{k-1,i} + \hat{\beta}_k)$.





For a 32-pages (=exp(3.463461)) AER article published written by 2 same-gender co-authors (=exp(0.602059999132)) in the year 2000 without author-provided JEL code that belongs to cluster 1, the expected number of sentences in the abstract in excess of 3 will increase by roughly .75 if the proportion of women changes from zero to 1.