Causal Machine Learning (II): Post-Double Selection

Prof. Tzu-Ting Yang
楊子霆

Institute of Economics, Academia Sinica
中央研究院經濟研究所

March 26, 2026

Problem of High-Dimensional Data

From Regression to High-Dimensional Settings

- **Recall from Part I:** Under the CIA, regression controls for observed confounders X to identify the causal effect of treatment D

$$Y_i = \alpha D_i + X_i\beta + \epsilon_i$$

- **Challenge:** To satisfy CIA, we may need to include **many** covariates X
  - Interaction terms, polynomials, and many control variables
  - The number of covariates p may be large relative to sample size n
- **Part II:** Post-Double Selection (PDS) uses LASSO to select the most relevant controls from high-dimensional data while preserving valid causal inference

Problem of High-Dimensional Data

- Consider linear regression model with p potential covariates where p is too large.

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i^j + \epsilon_i, \quad i = 1, \ldots, n$$

  - Y_i is observed outcome for individual i
  - X_i^j is observed covariate j for individual i
  - n is sample size
  - p is the number of covariates

- **Problem:** p could be much larger than n

Why We Need High-Dimensional Data?

- Why we need so many covariates?
  - Including more covariates can reduce **omitted variable bias (selection bias)**
  - Linear regression may predict well if include many relevant covariates
  - The number of covariates increases if we account for non-linearity or interaction effects
- **Example:**
  - Cross-country regressions, where we have only small number of countries, but thousands of macro variables.

Problem of High-Dimensional Data

Example 1: Test the Convergence Hypothesis

Sala-i-Martin, Xavier (1997), "**I Just Ran Two Million Regressions**", American Economic Review

- The author tries to examine the hypothesis of growth convergence and the determinants of economic growth
  - Whether poorer economies' per capita incomes will tend to grow at faster rates than richer economies
- He finds that a substantial number of variables can be found to be strongly related to growth
- Citations: 4,026 times

Problem of High-Dimensional Data

Example 1: Test the Convergence Hypothesis

- Examine the relation between GDP growth rate and initial per capita GDP:

$$\underbrace{\text{GrowthRate}}_{Y_i} = \beta_0 + \underbrace{\alpha}_{\text{ATE}} \underbrace{\log(\text{GDP})}_{D_i} + \sum_{j=1}^{p} \beta_j X_i^j + \epsilon_i$$

- Control a lot of covariates X_i^j
- In their data, they have $p = 60$ covariates, $n = 90$ observations
  - Need to do variable selection
- Test the convergence hypothesis: $\alpha < 0$
  - Poor countries catch up with richer countries, conditional on similar characteristics (e.g. institutions, human capital etc.)
  - Prediction from the classical Solow growth model.

Problem of High-Dimensional Data

Example 2: Effect of Abortion on Crime

John J. Donohue and Steven D. Levitt (2001), "**The Impact of Legalized Abortion on Crime,**" *Quarterly Journal of Economics*

- Examine the causal effect of legalized abortion on crime rate
  - **Mechanism:** unwanted children are at elevated risk of criminal behavior; abortion legalization reduced the number of such births
- Crime began to fall roughly 18 years after abortion legalization
- States with high abortion rates in the 1970s–80s experienced greater crime reductions in the 1990s
- **Highly controversial**: many debates on identification and confounders

Problem of High-Dimensional Data

Example 2: Effect of Abortion on Crime

- **Goal:** estimate causal effect of d_{it} (abortion rate) on y_{it} (crime rate)
- **Problem:** abortion rates are not randomly assigned
- **Key concern:** many confounders correlated with both abortion and crime rates
  - States differ in demographics, income, policing, social programs, etc.
  - Crime rates evolve differently across states for many reasons
  - Example confounder: high school dropout rate — correlated with both abortion access and crime
- **Solution:** control for $p = 284$ potential covariates but $n = 600$ (50 states $\times$ 12 years)
  - p is large relative to $n \Rightarrow$ high-dimensional problem

Problem of High-Dimensional Data

Example 2: Effect of Abortion on Crime

- Donohue and Levitt (2001) baseline model:

$$y_{it} = \alpha_0 d_{it} + \sum_{j=1}^{p} \beta_j x_{it}^j + \gamma_t + \delta_i + \epsilon_{it}$$

  - $n = 600$: 50 states $\times$ 12 years
  - y_{it}: crime rate (violent, property, or murder per 1,000 people)
  - d_{it}: "effective" abortion rate
  - $p = 284$ **potential covariates** x_{it}^j: lagged prisoners, lagged police $\times t$, income, beer consumption, initial abortion rate, etc.
  - γ_t: time fixed effects; δ_i: state fixed effects

Post-Double Selection

- Suppose we want to estimate the causal effect of treatment D_i on outcome Y_i
- Control p potential covariates where p is too large

$$Y_i = \beta_0 + \alpha D_i + \sum_{j=1}^{p} \beta_j X_i^j + \epsilon_i, \quad i = 1, \ldots, n$$

- **Traditional Variable Selection:**
  1. Drop all X_i^j that are **statistically insignificant** (e.g., using t-tests)
  2. Run OLS of Y_i on D_i and selected covariates X_i^j
- **Does not work** because fails to eliminate **omitted variable bias**

Review: Omitted Variable Bias

- Suppose the true model is:

$$Y_i = \delta + \alpha D_i + \beta X_i + \epsilon_i$$

- X_i is the observed characteristics (e.g. family income)
- But we estimate this model:

$$Y_i = \delta + \alpha D_i + u_i$$

- where $u_i = \beta X_i + \epsilon_i$

Review: Omitted Variable Bias

- OVB formula:

$$\hat{\alpha} \xrightarrow{p} \alpha + \frac{Cov(u_i, D_i)}{V(D_i)}$$
$$= \alpha + \beta \frac{Cov(X_i, D_i)}{V(D_i)}$$

- The difference between estimated treatment effect $\hat{\alpha}$ and true effect α depends on two components:

1. β: The effect of omitted variable X_i on outcome variable Y_i
2. $\frac{Cov(X_i, D_i)}{V(D_i)}$: The relationship between omitted variable X_i and treatment variable D_i

Post-Double Selection Method

Intuition

- Based on OVB formula, Belloni, Chernozhukov, Fernandez-Val and Hansen (2013) propose **Post-Double Selection (PDS)** approach:
  1. Use ML methods (LASSO) to select covariates X_i^j that can predict Y_i.
  2. Use ML methods (LASSO) to select covariates X_i^j that can predict D_i
  3. Run OLS of Y_i on D_i and the union of covariates selected in steps 1 and 2
- The additional selection step 2 can help eliminate the **omitted variable bias**

Post-Double Selection Method

Formal Illustration

- Suppose the true model is:

$$Y_i = \beta_0 + \alpha D_i + \sum_{j=1}^{p} \beta_j X_i^j + \epsilon_i$$

Post-Double Selection Method

Formal Illustration

- **Step 1 of PDS:** Use ML methods (LASSO) to select covariates X_i^j that can predict Y_i
  - Denote the set of LASSO-selected covariates by A

$$Y_i = \delta_0 + \sum_{j=1}^{p} \delta_j X_i^j + \zeta_i$$

- **Step 2 of PDS:** Use ML methods (LASSO) to select covariates X_i^j that can predict D_i
  - Denote the set of LASSO-selected covariates by B

$$D_i = \gamma_0 + \sum_{j=1}^{p} \gamma_j X_i^j + \varepsilon_i$$

Post-Double Selection Method

Formal Illustration

- **Step 3 of PDS:** Estimate the following OLS regression:

$$Y_i = \pi_0 + \alpha D_i + \sum_{j=1}^{g} \pi_j U_i^j + \upsilon_i$$

  - Let U_i^j denote the **union of covariates in A and B**

- The PDS estimator of treatment effect α is the coefficient on D in the above OLS regression

Post-Double Selection Method

Literature

Belloni, Chernozhukov, Fernandez-Val and Hansen (2013) ""**Inference on Treatment Effects after Selection amongst High-Dimensional Controls**", Review of Economic Studies

Belloni, Chernozhukov, Fernandez-Val and Hansen (2014) "**High-dimensi Methods and Inference on Structural and Treatment Effects**", Journal of Economic Perspectives

- For more details of PDS method, please read above papers

LASSO Estimation

Shrinkage Methods

Main Idea

- Consider linear regression model with p potential covariates where p is too large.

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i^j + \epsilon_i, \quad i = 1, \ldots, n$$

- Shrinkage estimators minimize **sum of squares error (SSE) with a penalty for model size**
  - This shrinks some of unimportant parameter estimates towards zero

Shrinkage Methods

Main Idea

- Depending on algorithm of penalty for model size, there are two popular methods:
  - Ridge regression
  - **Least Absolute Shrinkage and Selection Operator (LASSO)** regression
- The key assumption is **approximate sparsity**:
  - Some of the β_j coefficients are well-approximated by zero, and the approximation error is sufficiently 'small'

LASSO regression

Least Absolute Shrinkage and Selection Operator

- We can estimate the following LASSO regression:

$$y_i = \sum_{j=1}^{p} \beta_j x_i^j + \epsilon_i, \quad i = 1, \ldots, n$$

  - Subject to $\sum_{j=1}^{p} |\beta_j| \leq s$
  - s is a small number

LASSO regression

Least Absolute Shrinkage and Selection Operator

- The LASSO estimator a set of $\hat{\beta}_j$ solves the following optimization problem:

$$(\hat{\beta}_1, \hat{\beta}_2, ...\hat{\beta}_p) = \min_{\beta_j} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_i^j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

  - where $\lambda \geq 0$ is a **tuning parameter**
  - This is equivalent to minimizing SSE subject to $\sum_{j=1}^{p} |\beta_j| \leq s$ for some $s \geq 0$
  - s and λ are in one-to-one correspondence
- The LASSO estimator:
  - Minimizes the **sum of squared errors (SSE)** with a penalty for model complexity
  - Can set some coefficients exactly to zero, effectively selecting a subset of variables

Adjust Scale of Covariates

- However, the magnitude of parameter estimates β_j are related to **scale of covariates**
- Make parameter estimates β_j invariant to the scale of covariates
- We need to **standardize** the covariates X and outcome variable Y by dividing their standard deviations σ_{X^j} and σ_Y
  - The covariates we use in regression will be be transformed to:
    - ★ $x_i^j = \dfrac{X_i^j - \bar{X}^j}{\sigma_{X^j}}$
  - The outcome variable will be transformed to:
    - ★ $y_i = \dfrac{Y_i - \bar{Y}}{\sigma_Y}$

LASSO Regression

- There is no closed-form solution for the LASSO estimator
- **Intuition:**
  - Including more regressors incurs a cost: $\lambda \sum_{j=1}^{p} |\beta_j|$
  - Variables that contribute little to fit are penalized out: LASSO sets $\hat{\beta}_j = 0$ for some j, performing **variable selection**
- **Sparsity assumption:** LASSO works best when only a few $\beta_j \neq 0$ and most $\beta_j = 0$
- **Choosing λ = choosing which variables to include**
  - Large λ: heavier penalty $\Rightarrow$ fewer variables selected
  - Small λ: lighter penalty $\Rightarrow$ more variables retained
  - Extremes: $\lambda \to \infty$ sets all $\hat{\beta}_j = 0$; $\lambda \to 0$ recovers OLS criteria

How to select λ

Overview

- The **tuning parameter** λ controls the **strength of penalty** and determines a set of covariates
  - Each tuning parameter value λ corresponds to a fitted regression model

- **Example:** given $p = 3$ covariates, different λ values yield different models:

λ	Selected covariates	Fitted model
small	x^1, x^2, x^3	$\hat{y}_i = \hat{\beta}_1 x_i^1 + \hat{\beta}_2 x_i^2 + \hat{\beta}_3 x_i^3$
medium	x^1, x^3	$\hat{y}_i = \hat{\beta}_1 x_i^1 + \hat{\beta}_3 x_i^3$
large	x^1	$\hat{y}_i = \hat{\beta}_1 x_i^1$
$\to \infty$	(none)	$\hat{y}_i = 0$

- The shrinkage methods allow us to **simplify the model selection problem to a one-dimensional problem**

Three Ways to Choose λ

Overview

1. **Cross-validation approach:**
   - Data-driven: choose λ that minimizes **out-of-sample prediction error** (typically via k-fold CV)
   - Simple and widely used, but computationally costly and does not guarantee inferential properties for causal estimation
2. **Rigorous (theory-based) approach:**
   - Theory-driven: Belloni, Chernozhukov, and Hansen (2014, *ReStud*) derive the optimal $\lambda \propto \sqrt{n \log p}$
   - More data ($\uparrow n$) and more variables ($\uparrow p$) $\Rightarrow$ stronger penalty needed to avoid false selection
   - Provides formal statistical guarantees in high dimensions; used in the **Post-Double Selection** method

3 **Information criteria approach:**

- Choose λ that minimizes an information criterion, e.g., BIC:

$$\text{BIC}(\lambda) = \log\left(\frac{\text{SSE}(\lambda)}{n}\right) + \frac{\log n}{n} \cdot \hat{s}(\lambda)$$

- $\hat{s}(\lambda)$: number of selected variables (non-zero $\hat{\beta}_j$) under penalty λ
- **Intuition:** trade-off between fit and sparsity
  - ★ Large $\lambda \Rightarrow$ small $\hat{s}(\lambda)$ but large SSE
  - ★ Small $\lambda \Rightarrow$ small SSE but large $\hat{s}(\lambda)$
  - ★ BIC selects λ that balances the two

Cross-validation approach

Cross-validation approach

Overview

- Cross-validation is a simple, intuitive way to evaluate model fit (choose λ) based on **prediction error**
  - Single-split validation, K-fold cross validation, leave-one-out cross validation
  - Divide sample into:
    - ★ Training data: estimation
    - ★ Validation data: evaluate prediction
  - Computationally more expensive

Cross-validation approach

Overview

- Consider two types of data sets
  1. **Training data**
     - ⋆ Used to estimate a regression model
     - ⋆ Get estimated coefficients $\hat{\beta}_j$
  2. **Validation data**
     - ⋆ Additional data used to determine how good is the regression model fit
     - ⋆ A test observation (x_i^{j+}, y_i^+) is a previously unseen observation

Cross-validation approach

Training Data

- Use **training data** to estimate the following regression using LASSO:

$$y_i = \sum_{j=1}^{p} \beta_j x_i^j + \epsilon_i, \quad i = 1, \ldots, n$$

- Choose $(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)$ that minimize **SSE with a penalty for model size**:

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_i^j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

  - where $\lambda \geq 0$ is a **tuning parameter**

- Predict y_i using the estimated model $\hat{f}_\lambda(\cdot)$:

$$\hat{y}_i = \hat{f}_\lambda(x_i^j) = \sum_{j=1}^{p} \hat{\beta}_j x_i^j$$

Cross-validation approach

Validation Data

- Use the estimated model $\hat{f}_\lambda(\cdot)$ to predict unseen observations in the **validation data**
- Evaluate prediction accuracy via **SSE** or **MSE**:

$$\text{SSE} = \sum_{i=1}^{n} \left(y_i^+ - \hat{f}_\lambda(x_i^{j+}) \right)^2$$

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i^+ - \hat{f}_\lambda(x_i^{j+}) \right)^2$$

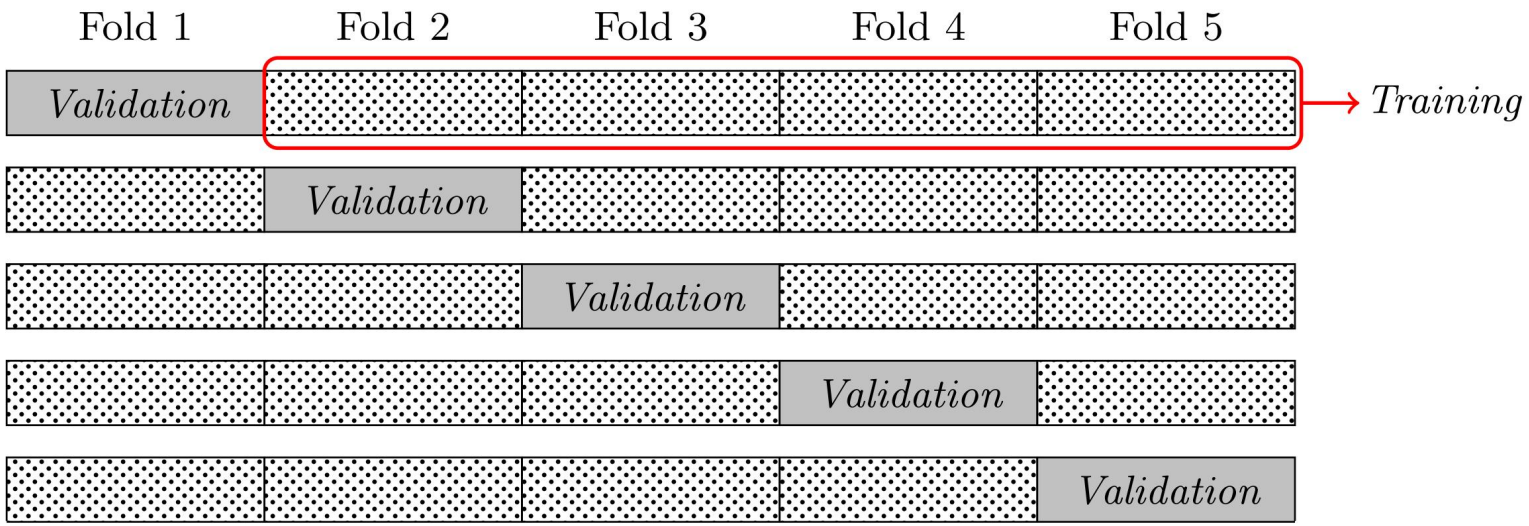
  - (x_i^{j+}, y_i^+): unseen observations in **validation data**
  - $\hat{f}_\lambda(\cdot)$: estimated model from **training data**

- **Key idea:** a model that fits training data well may not predict validation data well $\Rightarrow$ **overfitting**

K-fold Cross Validation

Step 1

1. Randomly divide the data set $\{1, \ldots, n\}$ into K groups $F_1, \ldots, F_K$ of roughly equal size
   - Commonly we split the data into 5 or 10 groups/folds ($K = 5$ or $K = 10$)
- Consider training on $(x_i^j, y_i), i \notin F_k$, and validating on $(x_i^{j+}, y_i^+), i \in F_k$



K-fold Cross Validation

Step 2

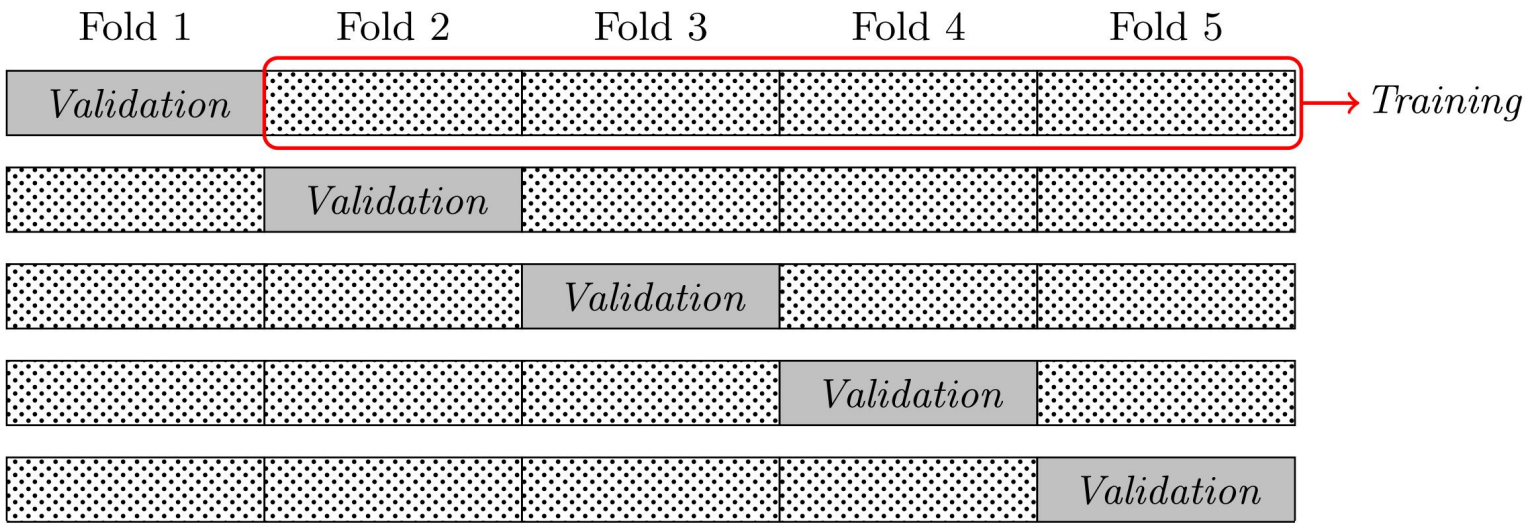
2 For each value of the tuning parameter $\lambda \in \{\lambda_1, \ldots, \lambda_m\}$, we do the following steps:

- Note that each λ has the corresponding regression model so we have m regression models in this case

K-fold Cross Validation

Step 2-1

2-1 Compute the regression estimates $\widehat{f}_\lambda^{-k}$ using the **training data**

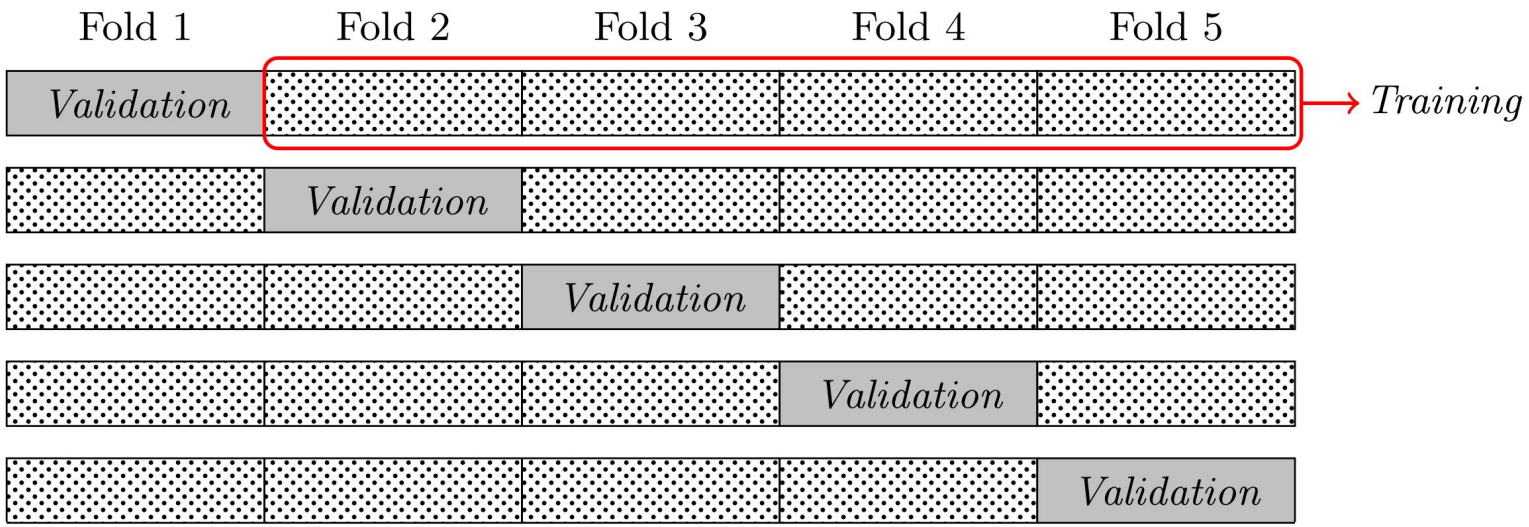


K-fold Cross Validation

Step 2-2

2-2 Compute the prediction error $e_k(\lambda)$ on the each **validation data**:

$$e_k(\lambda) = \sum_{i \in F_k} (y_i^+ - \hat{f}_\lambda^{-k}(x_i^{j+}))^2$$

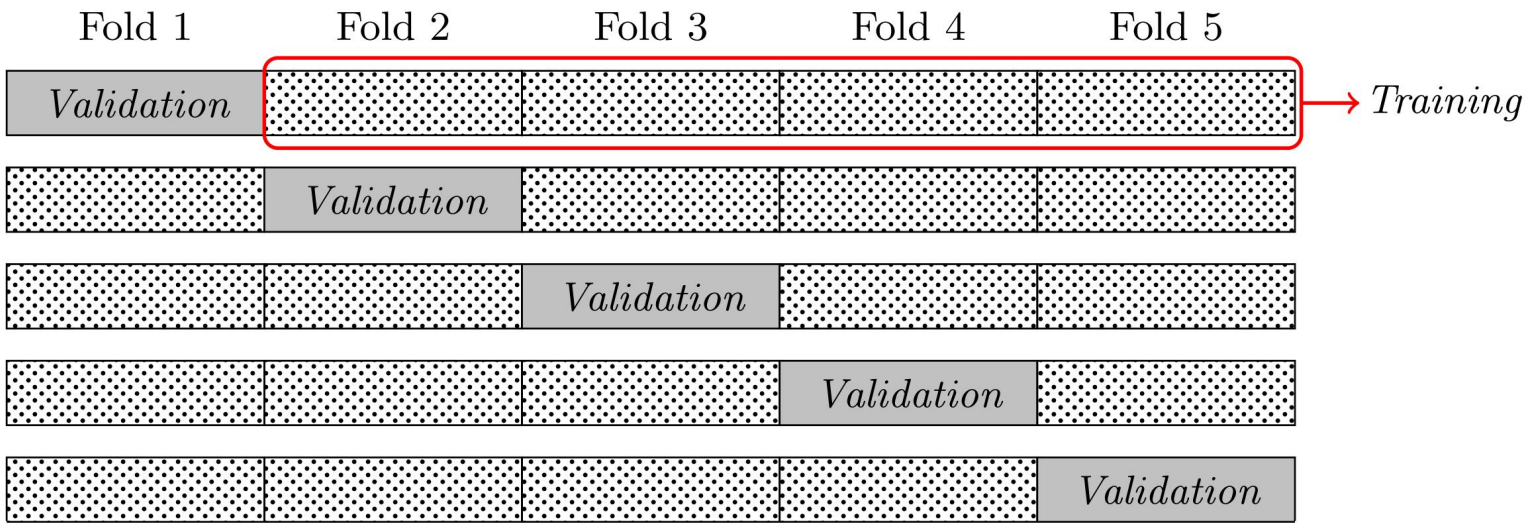


K-fold Cross Validation

Step 2-3

2-3 Then, compute the **average prediction error** over all **validation data sets**

$$CV(\lambda) = \frac{1}{n}\sum_{k=1}^{K} e_k(\lambda) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in F_k}(y_i^+ - \hat{f}_\lambda^{-k}(x_i^{j+}))^2$$



K-fold Cross Validation

Step 3: Choose $\hat{\lambda}$

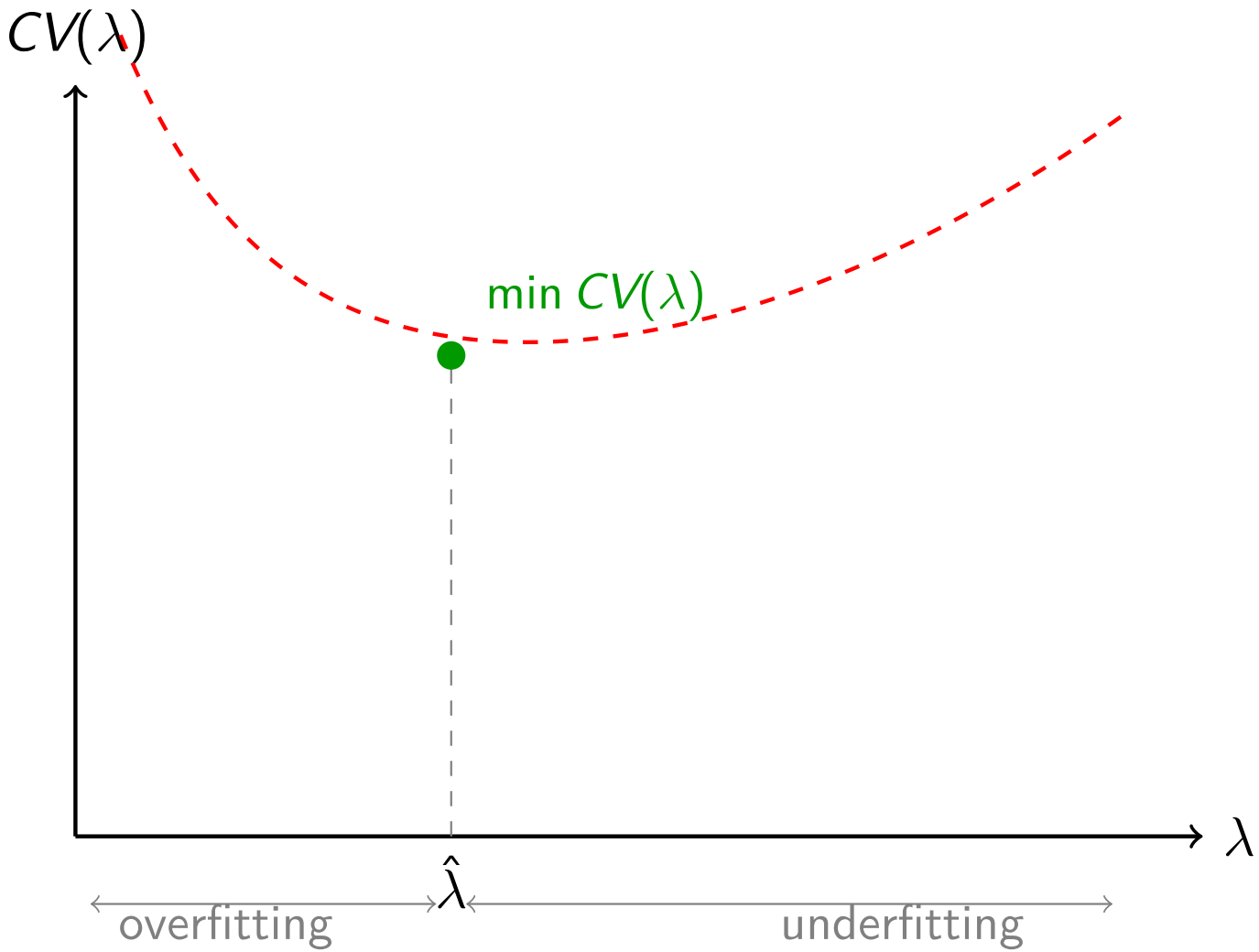
3 Choose λ that minimizes the **cross-validation error curve** $CV(\lambda)$:

$$\hat{\lambda} = \arg \min_{\lambda \in \{\lambda_1, \ldots, \lambda_m\}} CV(\lambda)$$

- $CV(\lambda)$ is a U-shaped curve
- When $K = n$: **leave-one-out cross-validation**

K-fold Cross Validation

CV Error Curve



STATA Example

STATA Example: Test the Convergence Hypothesis

Overview

Sala-i-Martin, Xavier (1997), "**I Just Ran Two Million Regressions**", *American Economic Review* (cited 2,931 times)

- Same setup as Example 1: test $\alpha < 0$ (convergence hypothesis) in

$$\underbrace{\text{GrowthRate}}_{Y_i} = \beta_0 + \underbrace{\alpha}_{\text{ATE}} \underbrace{\log(\text{GDP})}_{D_i} + \sum_{j=1}^{p} \beta_j X_i^j + \epsilon_i$$

- Data: $p = 60$ covariates, $n = 90$ observations $\Rightarrow$ need **variable selection**
- Key finding: a large number of variables are robustly related to growth

STATA Example: Test the Convergence Hypothesis

Data and Code

- See **ML.do**
- Use growth.dta

STATA Example: Test the Convergence Hypothesis

Control All Covariates

```
1 reg Outcome gdpsh465 bmp1l- tot1,r
```

- Control all 60 covariates
- This could result in imprecise estimates of coefficients due to overfitting

STATA Example: Test the Convergence Hypothesis

Control All Covariates

```
. reg Outcome gdpsh465 bmp1l- tot1,r

Linear regression                               Number of obs     =         90
                                                F(61, 28)         =       8.96
                                                Prob > F          =     0.0000
                                                R-squared         =     0.8871
                                                Root MSE          =     .03074
```

Outcome	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
gdpsh465	-.009378	.0324211	-0.29	0.775	-.0757896	.0570336
bmp1l	-.0688627	.0408531	-1.69	0.103	-.1525464	.014821
freeop	.080069	.2318888	0.35	0.732	-.3949337	.5550716
freetar	-.4889626	.4124222	-1.19	0.246	-1.333771	.355846
h65	-2.362099	.7377739	-3.20	0.003	-3.87336	-.8508374
hm65	.7071434	.518609	1.36	0.184	-.355179	1.769466
hf65	1.693448	.4580274	3.70	0.001	.7552219	2.631675
p65	.2655267	.1645915	1.61	0.118	-.0716236	.602677
pm65	.1369526	.1310772	1.04	0.305	-.1315469	.4054521
pf65	-.3312669	.1843816	-1.80	0.083	-.7089556	.0464217
s65	.0390793	.1772324	0.22	0.827	-.3239648	.4021234
sm65	-.0306685	.1230053	-0.25	0.805	-.2826334	.2212964
sf65	-.1799173	.1014187	-1.77	0.087	-.387664	.0278294
fert65	.0068808	.0289444	0.24	0.814	-.0524091	.0661708
mort65	-.2334545	.856995	-0.27	0.787	-1.988929	1.52202

STATA Example: Test the Convergence Hypothesis

Control All Covariates

- The initial per capita GDP has little impact on economic growth rate
- Does NOT support prediction from the classical Solow growth model

Review: Post-Double Selection Method

Formal Illustration

- Suppose the true model is:

$$\underbrace{\text{GrowthRate}}_{Y_i} = \beta_0 + \underbrace{\alpha}_{\text{ATE}} \underbrace{\log(\text{GDP})}_{D_i} + \sum_{j=1}^{p} \beta_j X_i^j + \epsilon_i$$

Review: Post-Double Selection Method

Formal Illustration

- **Step 1:** Use ML methods (LASSO) to select covariates X_i^j that can predict Y_i
  - Denote the set of LASSO-selected covariates by A

$$\underbrace{\text{GrowthRate}}_{Y_i} = \delta_0 + \sum_{j=1}^{p} \delta_j X_i^j + \zeta_i$$

- **Step 2:** Use ML methods (LASSO) to select covariates X_i^j that can predict D_i
  - Denote the set of LASSO-selected covariates by B

$$\underbrace{\log(\text{GDP})}_{D_i} = \gamma_0 + \sum_{j=1}^{p} \gamma_j X_i^j + \varepsilon_i$$

Review: Post-Double Selection Method

Formal Illustration

- **Step 3:** Estimate the following OLS regression:

$$\underbrace{\text{GrowthRate}}_{Y_i} = \pi_0 + \underbrace{\alpha}_{\text{ATE}} D_i + \sum_{j=1}^{g} \pi_j U_i^j + \upsilon_i$$

  - Let U_i^j denote the **union of covariates in A and B**

- The PDS estimator of treatment effect α is the coefficient on D in the above OLS regression
- The additional selection step 2 can help eliminate the **omitted variable bias**

Review: Three Ways to Choose λ

Overview

1. **Cross-validation approach:** It is a data-driven approach. Choose **tuning parameter** λ that minimize **prediction error**
   - STATA command: **cvlasso**
2. **Rigorous approach:** It is a theory-driven approach. Belloni et al. (2012, Econometrica) develop theory and feasible algorithms for the optimal λ.
   - STATA command: **rlasso**
3. **Information criteria approach:** Select the value of λ that minimizes information criterion (AIC, AICc, BIC or EBIC).
   - STATA command: **lasso2**

- To use the above commands, you need to install **lassopack** package
  - STATA command: **ssc install lassopack**

STATA Example: Test the Convergence Hypothesis

Use Post-Double Selection Method

```
1 pdslasso Outcome gdpsh465 (bmp1l- tot1),rob
```

- **pdslasso**: Implement double selection method with LASSO to select covariates
  - Select covariates from $bmp1l - tot1$ to predict outcome Y_i and treatment variable D_i

STATA Example: Test the Convergence Hypothesis

Use Post-Double Selection Method

```
. pdslasso Outcome gdpsh465 (bmp1l- tot1),rob
1.  (PDS/CHS) Selecting HD controls for dep var Outcome...
Selected: bmp1l
2.  (PDS/CHS) Selecting HD controls for exog regressor gdpsh465...
Selected: freetar hm65 sf65 lifee065 humanf65 pop6565


Estimation results:

Specification:
Regularization method:                lasso
Penalty loadings:                     heteroskedastic
Number of observations:               90
Exogenous (1):                        gdpsh465
High-dim controls (60):               bmp1l freeop freetar h65 hm65 hf65 p65 pm65 pf65 s65
                                      sm65 sf65 fert65 mort65 lifee065 gpop1 fert1 mort1
                                      invsh41 geetot1 geerec1 gde1 govwb1 govsh41 gvxdxe41
                                      high65 highm65 highf65 highc65 highcm65 highcf65 human65
                                      humanm65 humanf65 hyr65 hyrm65 hyrf65 no65 nom65 nof65
                                      pinstab1 pop65 worker65 pop1565 pop6565 sec65 secm65
                                      secf65 secc65 seccm65 seccf65 syr65 syrm65 syrf65
                                      teapri65 teasec65 ex1 im1 xr65 tot1
Selected controls (7):                bmp1l freetar hm65 sf65 lifee065 humanf65 pop6565
Unpenalized controls (1):             _cons

Structural equation:

OLS using CHS lasso-orthogonalized vars
```

Outcome	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
gdpsh465	-.0278804	.0164206	-1.70	0.090	-.0600642	.0043035

STATA Example: Test the Convergence Hypothesis

Use Post-Double Selection Method

OLS with PDS-selected variables and full regressor set

Outcome	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
gdpsh465	-.0500059	.0150732	-3.32	0.001	-.0795488	-.0204629
bmp1l	-.0782423	.0157799	-4.96	0.000	-.1091703	-.0473144
freetar	-.5746764	.2538159	-2.26	0.024	-1.072146	-.0772064
hm65	.0511529	.0538366	0.95	0.342	-.0543649	.1566707
sf65	-.0470218	.0487002	-0.97	0.334	-.1424725	.0484288
lifee065	.2122794	.054255	3.91	0.000	.1059415	.3186173
humanf65	-.000376	.0035354	-0.11	0.915	-.0073052	.0065531
pop6565	.1343893	.2301626	0.58	0.559	-.3167211	.5854996
_cons	-.4064513	.1830995	-2.22	0.026	-.7653198	-.0475828

Standard errors and test statistics valid for the following variables only:
gdpsh465

STATA Example: Test the Convergence Hypothesis

Use Post-Double Selection Method

- Instead of 60 variables, this method only select 7 variables that are correlated with outcome Y_i or treatment variable D_i
  - Higher initial per capita GDP could lead to lower GDP growth rate in the later years
  - Poor countries do catch up with richer countries
- Support prediction from the classical Solow growth model

STATA Example: Test the Convergence Hypothesis

Use Post-Double Selection Method

```
** step 1:
rlasso Outcome bmp1l- tot1,rob

** step 2:
rlasso gdpsh465 bmp1l- tot1,rob

** step 3:
reg Outcome gdpsh465 bmp1l freetar hm65 sf65 lifee065
    humanf65 pop6565,r
```

- You can implement PDS method by yourself
- PDS uses rigorous approach to select optimal λ.
  - STATA command: **rlasso**

STATA Example: Test the Convergence Hypothesis

Use Post-Double Selection Method

```
pdslasso Outcome gdpsh465 (bmp1l- tot1), rob pnotpen(ex1 im1
    )
```

- **pnotpen(ex1 im1)**: Specifies variables that are automatically included in the model without going through the selection process
  - Variables in **pnotpen()** are not penalized and will always be included in the final model, regardless of their statistical significance

STATA Example: Visualize Data

Scatter Plot

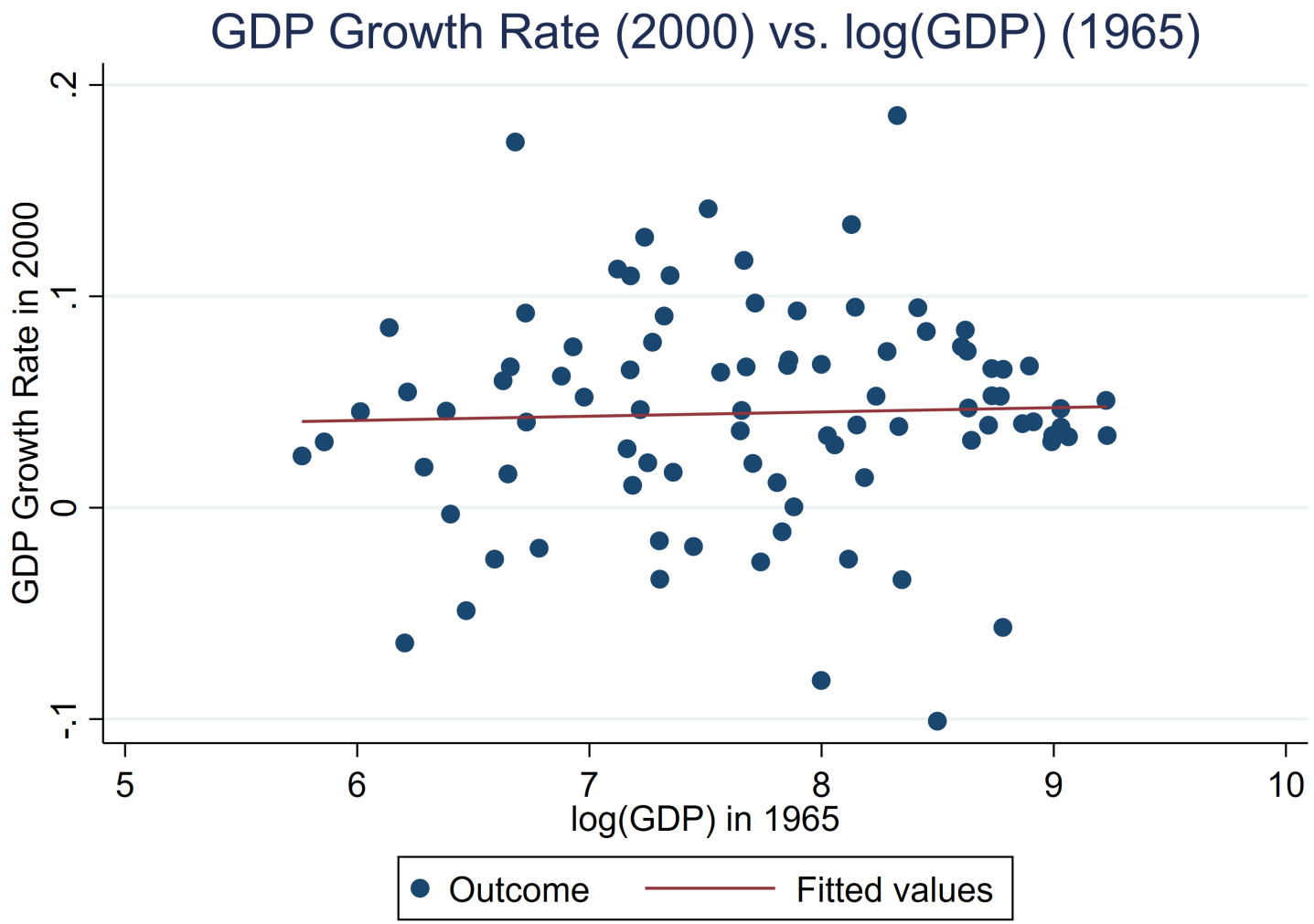
```
graph twoway ///
(scatter Outcome gdpsh465) ///
(lfit Outcome gdpsh465), ///
title("GDP Growth Rate (2000) vs. log(GDP) (1965)") ///
xtitle("log(GDP) in 1965") ///
ytitle("GDP Growth Rate in 2000") ///
graphregion(color(white)) xlabel(5(1)10) ylabel(-0.1(0.1)
    0.2)

graph export "$pic\gdp_scatter.png", replace width(3000)
```

- **graph twoway scatter**
  - Generates a scatter plot with ‘Outcome‘ as the y-axis and ‘gdpsh465‘ as the x-axis.
- **graphregion(color(white))**
  - Sets the background color of the graph to white for a clean look.
- **graph export**
  - Exports the graph as a PNG file named

STATA Example: Visualize Data

Scatter Plot



STATA Example: Visualize Data

Scatter Plot: residuals

```
1 * Regress Outcome on control variables and get residuals
2 regress Outcome bmp1l freetar hm65 sf65 lifee065 humanf65
      pop6565
3 predict res_Outcome, residuals
4
5 * Regress gdpsh465 on control variables and get residuals
6 regress gdpsh465 bmp1l freetar hm65 sf65 lifee065 humanf65
      pop6565
7 predict res_gdpsh465, residuals
```

STATA Example: Visualize Data

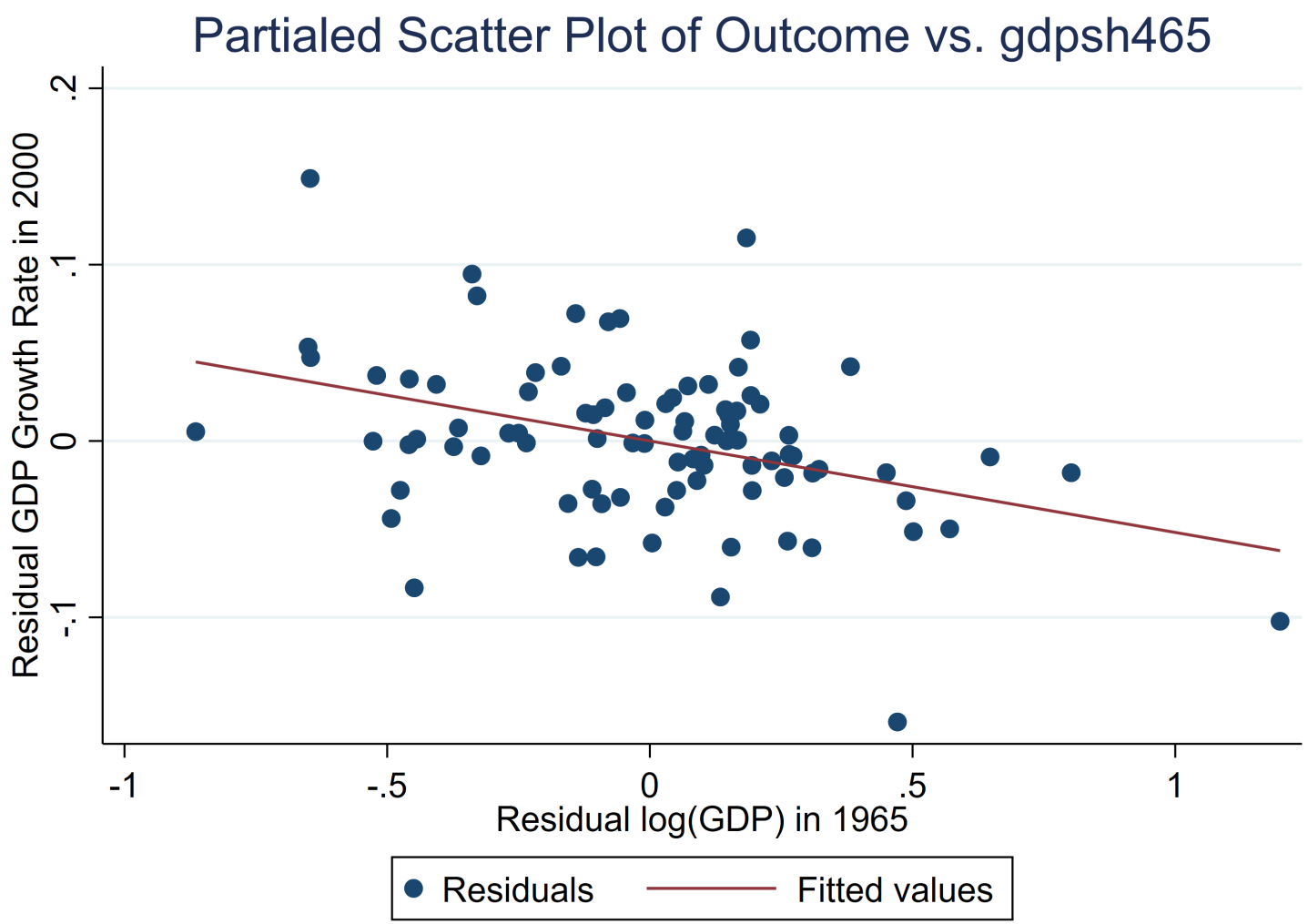
Scatter Plot: residuals

```
* Scatter plot of the residuals
graph twoway ///
(scatter res_Outcome res_gdpsh465) ///
(lfit res_Outcome res_gdpsh465), ///
title("Partialed Scatter Plot of Outcome vs. gdpsh465") ///
xtitle("Residual log(GDP) in 1965") ///
ytitle("Residual GDP Growth Rate in 2000") ///
graphregion(color(white)) xlabel(-1(0.5)1) ylabel(-0.1(0.1)
    0.2)

graph export "$pic\gdp_scatter_r.png", replace width(3000)
```

STATA Example: Visualize Data

Scatter Plot: residuals



R Example

R Example: Test the Convergence Hypothesis

Data and Code

- See **ML.R**
- Use growth.dta
- Install the following packages:
  - haven
  - hdm

R Example: Test the Convergence Hypothesis

Install Packages

```
1 install.packages("hdm")
2
3 library(hdm)
```

- Install and load package **hdm**
  - This package can help you implement post-double selection method

R Example: Test the Convergence Hypothesis

Read Data

```
GrowthData <- read_dta(paste0(rawdata, "/growth.dta"))

# Checking dimensions of the dataset
dim(GrowthData)

varnames = colnames(GrowthData)
```

- **dim(GrowthData)**: Displays the dimensions of the loaded data, ensuring it is loaded correctly.
- **varnames**: Stores the names of all variables in the dataset for reference.

R Example: Test the Convergence Hypothesis

Create Sample for Analysis

```
y = GrowthData[, 1, drop = F]
d = GrowthData[, 2, drop = F]
X = as.matrix(GrowthData)[, -c(1, 2)]
varnames = colnames(GrowthData)
```

- Load dataset and define outcome y, treatment variable d, and covariates x
  - **y**: Dependent variable, representing the first column of the dataset.
  - **d**: Treatment or key independent variable, taken from the second column.
  - **X**: Matrix of covariates, includes all columns except the first two.

R Example: Test the Convergence Hypothesis

PDS Estimation

```
1 doublesel.effect = rlassoEffect(x = X, y = y, d = d, method
      = "double selection")
2
3 summary(doublesel.effect)
```

- Implement double selection estimation and report the treatment effect

R Example: Visualize Data

Scatter Plot

```
ggplot(GrowthData, aes(x = gdpsh465, y = Outcome)) +
geom_point() +
geom_smooth(method = "lm", se = FALSE, color = "blue") +
labs(title = "GDP Growth Rate (2000) vs. log(GDP) (1965)",
x = "log(GDP) in 1965",
y = "GDP Growth Rate in 2000") +
theme(plot.background = element_rect(fill = "white")) +
scale_x_continuous(breaks = seq(5, 10, by = 1)) +
scale_y_continuous(breaks = seq(-0.1, 0.2, by = 0.1))
```

- **ggplot()**
  - Initializes the plot with ‘gdpsh465‘ on the x-axis and ‘Outcome‘ on the y-axis.
- **geom_point()**
  - Adds scatter points to the plot.

R Example: Visualize Data

Scatter Plot

```
ggplot(GrowthData, aes(x = gdpsh465, y = Outcome)) +
geom_point() +
geom_smooth(method = "lm", se = FALSE, color = "blue") +
labs(title = "GDP Growth Rate (2000) vs. log(GDP) (1965)",
x = "log(GDP) in 1965",
y = "GDP Growth Rate in 2000") +
theme(plot.background = element_rect(fill = "white")) +
scale_x_continuous(breaks = seq(5, 10, by = 1)) +
scale_y_continuous(breaks = seq(-0.1, 0.2, by = 0.1))
```

- **geom_smooth(method = "lm")**
  - Adds a linear regression line without confidence interval shading.
- **theme(plot.background = ...)**
  - Sets the background color of the plot to white.
- **scale_x_continuous() & scale_y_continuous()**
  - Controls the axis ticks to match specific intervals.

R Example: Visualize Data

Scatter Plot

```
ggsave(filename = file.path(pic, "gdp_scatter.png"), width =
    10, dpi = 300)
```

- **ggsave()**
  - Saves the plot as a PNG image named `gdp_scatter.png`.

R Example: Visualize Data

Residuals from Regressions

```
# Step 1: Regress Outcome on controls, extract residuals
model_y <- lm(Outcome ~ bmp1l + freetar + hm65 + sf65 +
    lifee065 + humanf65 + pop6565, data = GrowthData)
data$res_Outcome <- resid(model_y)

# Step 2: Regress gdpsh465 on controls, extract residuals
model_x <- lm(gdpsh465 ~ bmp1l + freetar + hm65 + sf65 +
    lifee065 + humanf65 + pop6565, data = GrowthData)
data$res_gdpsh465 <- resid(model_x)
```

R Example: Visualize Data

Scatter Plot: residuals

```
ggplot(GrowthData, aes(x = res_gdpsh465, y = res_Outcome)) +
geom_point() +
geom_smooth(method = "lm", se = FALSE, color = "blue") +
labs(title = "Partialed Scatter Plot of Outcome vs. gdpsh465
    ",
x = "Residual log(GDP) in 1965",
y = "Residual GDP Growth Rate in 2000") +
theme(plot.background = element_rect(fill = "white")) +
scale_x_continuous(breaks = seq(-1, 1, by = 0.5)) +
scale_y_continuous(breaks = seq(-0.1, 0.2, by = 0.1))

ggsave(filename = file.path(pic, "gdp_scatter_r.png"), width
     = 10, dpi = 300)
```

Recommended Resources for Self-Learning

- Machine Learning & Causal Inference: A Short Course
  - https://www.gsb.stanford.edu/faculty-research/centers-initi
    sil/research/methods/ai-machine-learning/short-course

Recommended Resources for Self-Learning

- NBER Summer Institute 2013: Econometric Methods for High-Dimensiona Data
  - https://www.nber.org/econometrics_minicourse_2013/
- One free textbook: An Introduction to Statistical Learning
  - Website:
    - ★ http://faculty.marshall.usc.edu/gareth-james/
  - Video lectures:
    - ★ https://www.youtube.com/playlist?list=PLOg0ngHtcqbPTlZzRHA2oc qZ5V

Recommended Resources for Self-Learning

- the Stata Lasso Page
  - https://statalasso.github.io/