Supervised Learning, Part 1: Regression

Elliott Ash

Max Planck Summer School 2017

Different Methods for Different Goals

- **Supervised: Pursuing a known goal – prediction or classification.**
- Unsupervised: Unknown goal, let the computer summarize the data.

Approximating $Y = f(X)$

- We want to predict a real-valued outcome Y given X, that is, constructing an approximation of the function $f(X)$.
  - With high-dimensionality and multi-collinearity, normal regression methods do not work.
- Supervised learning:
  - regularized regression
  - random forests
  - cross-validation

Approximating $Y = f(X)$

- We want to predict a real-valued outcome Y given X, that is, constructing an approximation of the function $f(X)$.
  - With high-dimensionality and multi-collinearity, normal regression methods do not work.
- Supervised learning:
  - regularized regression
  - random forests
  - cross-validation

Outline

OLS Regression

- Consider the linear model

$$Y_i = X_i'\beta + \varepsilon_i$$

  where Y_i and all elements of X_i have been de-meaned and standardized to $s.d. = 1$.

- OLS assumptions:
  - X_i uncorrelated with ε_i
    - Let's just assume this for now; will come back later.
  - Columns of X_i are not highly collinear.
    - In the case of word/n-gram frequency data, this is a bad assumption.

OLS Regression

- Consider the linear model

$$Y_i = X_i'\beta + \varepsilon_i$$

  where Y_i and all elements of X_i have been de-meaned and standardized to $s.d. = 1$.

- OLS assumptions:
  - X_i uncorrelated with ε_i
    - Let's just assume this for now; will come back later.
  - Columns of X_i are not highly collinear.
    - In the case of word/n-gram frequency data, this is a bad assumption.

OLS Regression

- Consider the linear model

$$Y_i = X_i'\beta + \varepsilon_i$$

  where Y_i and all elements of X_i have been de-meaned and standardized to $s.d. = 1$.

- OLS assumptions:
  - X_i uncorrelated with ε_i
    - Let's just assume this for now; will come back later.
  - Columns of X_i are not highly collinear.
    - In the case of word/n-gram frequency data, this is a bad assumption.

OLS Regression

- Consider the linear model

$$Y_i = X_i'\beta + \varepsilon_i$$

  where Y_i and all elements of X_i have been de-meaned and standardized to $s.d. = 1$.

- OLS assumptions:
  - X_i uncorrelated with ε_i
    - Let's just assume this for now; will come back later.
  - Columns of X_i are not highly collinear.
    - In the case of word/n-gram frequency data, this is a bad assumption.

OLS Regression

- Consider the linear model

$$Y_i = X_i'\beta + \varepsilon_i$$

  where Y_i and all elements of X_i have been de-meaned and standardized to $s.d. = 1$.

- OLS assumptions:
  - X_i uncorrelated with ε_i
    - Let's just assume this for now; will come back later.
  - Columns of X_i are not highly collinear.
    - In the case of word/n-gram frequency data, this is a bad assumption.

Univariate OLS Regressions to Rank Predictive Features

- Consider the univariate regression

$$Y_i = \beta_w x_i^w + \varepsilon_i$$

  for text feature w (e.g., relative word or n-gram frequency).

- Can be estimated with OLS
  - Can add fixed effects, or even better: residualize Y and X on fixed effects before running any regressions.
  - Robust or clustered standard errors is optional, if the goal is just to rank predictors or filter out noise features.

Univariate OLS Regressions to Rank Predictive Features

- Consider the univariate regression

$$Y_i = \beta_w x_i^w + \varepsilon_i$$

  for text feature w (e.g., relative word or n-gram frequency).

- Can be estimated with OLS
  - Can add fixed effects, or even better: residualize Y and X on fixed effects before running any regressions.
  - Robust or clustered standard errors is optional, if the goal is just to rank predictors or filter out noise features.

Univariate OLS Regressions to Rank Predictive Features

- Consider the univariate regression

$$Y_i = \beta_w x_i^w + \varepsilon_i$$

  for text feature w (e.g., relative word or n-gram frequency).

- Can be estimated with OLS
  - Can add fixed effects, or even better: residualize Y and X on fixed effects before running any regressions.
  - Robust or clustered standard errors is optional, if the goal is just to rank predictors or filter out noise features.

Univariate OLS Regressions to Rank Predictive Features

- Consider the univariate regression

$$Y_i = \beta_w x_i^w + \varepsilon_i$$

  for text feature w (e.g., relative word or n-gram frequency).

- Can be estimated with OLS
  - Can add fixed effects, or even better: residualize Y and X on fixed effects before running any regressions.
  - Robust or clustered standard errors is optional, if the goal is just to rank predictors or filter out noise features.

OLS Regression in Python statsmodels

- One could write a DO file to run these regressions in Stata.
  - But the loops and data saving would be tricky with so many feature variables.
- Easier to do in R or Python (statsmodels package)
  - Loop through features
  - run the regression
  - save t-statistics and coefficients in a list
  - [demo_code.py]

OLS Regression in Python statsmodels

- One could write a DO file to run these regressions in Stata.
  - But the loops and data saving would be tricky with so many feature variables.
- Easier to do in R or Python (statsmodels package)
  - Loop through features
  - run the regression
  - save t-statistics and coefficients in a list
  - [demo_code.py]

Gentzkow and Shapiro (2010)

- Gentzkow and Shapiro (*Econometrica* 2010) introduced quantitative text analysis to economics.
- Approach:
  - Collect speeches from U.S. Congressional Record for 2005.
  - Select 1000 n-grams that are predictive of Republican or Democrat speaker
- For each phrase w, regress $Y_i = \beta_w x_i^w + \varepsilon_i$, where Y_i is political party of speaker i and x_i^w is relative frequency of phrase w.

Gentzkow and Shapiro (2010)

- Gentzkow and Shapiro (*Econometrica* 2010) introduced quantitative text analysis to economics.
- Approach:
  - Collect speeches from U.S. Congressional Record for 2005.
  - Select 1000 n-grams that are predictive of Republican or Democrat speaker
- For each phrase w, regress $Y_i = \beta_w x_i^w + \varepsilon_i$, where Y_i is political party of speaker i and x_i^w is relative frequency of phrase w.

Gentzkow and Shapiro (2010)

- Gentzkow and Shapiro (*Econometrica* 2010) introduced quantitative text analysis to economics.
- Approach:
  - Collect speeches from U.S. Congressional Record for 2005.
  - Select 1000 n-grams that are predictive of Republican or Democrat speaker
- For each phrase w, regress $Y_i = \beta_w x_i^w + \varepsilon_i$, where Y_i is political party of speaker i and x_i^w is relative frequency of phrase w.

Gentzkow and Shapiro (2010)

- Gentzkow and Shapiro (*Econometrica* 2010) introduced quantitative text analysis to economics.
- Approach:
  - Collect speeches from U.S. Congressional Record for 2005.
  - Select 1000 n-grams that are predictive of Republican or Democrat speaker
- For each phrase w, regress $Y_i = \beta_w x_i^w + \varepsilon_i$, where Y_i is political party of speaker i and x_i^w is relative frequency of phrase w.

Gentzkow and Shapiro (2010) (2)

- Then form text-predicted ideology for newspapers by summing the prediction from each univariate regression:

$$\hat{y}_p = \sum_{w=1}^{1000} \hat{\beta}_w x_i^w$$

- This assumes that the effects of each x^w on y are independent of each other.
- The measure is then used to explore slant in newspapers.
  - They find that newspapers respond to consumer (rather than owner) political preferences.

Gentzkow and Shapiro (2010) (2)

- Then form text-predicted ideology for newspapers by summing the prediction from each univariate regression:

$$\hat{y}_p = \sum_{w=1}^{1000} \hat{\beta}_w x_i^w$$

- This assumes that the effects of each x^w on y are independent of each other.
- The measure is then used to explore slant in newspapers.
  - They find that newspapers respond to consumer (rather than owner) political preferences.

Gentzkow and Shapiro (2010) (2)

- Then form text-predicted ideology for newspapers by summing the prediction from each univariate regression:

$$\hat{y}_p = \sum_{w=1}^{1000} \hat{\beta}_w x_i^w$$

- This assumes that the effects of each x^w on y are independent of each other.
- The measure is then used to explore slant in newspapers.
  - They find that newspapers respond to consumer (rather than owner) political preferences.

Ash, Morelli, and Van Weelden (2017)

- Approach:
  - Adopt the measure from Gentzkow and Shapiro to analyze divisiveness/polarization in Congress.
- Results:
  - Senators use more divisive language when they are up for election.
  - House members respond to greater news coverage with more divisive language.
- Interpretation:
  - Electoral incentives and transparency are important contributors to polarization of U.S. politics.

Ash, Morelli, and Van Weelden (2017)

- Approach:
  - Adopt the measure from Gentzkow and Shapiro to analyze divisiveness/polarization in Congress.
- Results:
  - Senators use more divisive language when they are up for election.
  - House members respond to greater news coverage with more divisive language.
- Interpretation:
  - Electoral incentives and transparency are important contributors to polarization of U.S. politics.

Ash, Morelli, and Van Weelden (2017)

- Approach:
  - Adopt the measure from Gentzkow and Shapiro to analyze divisiveness/polarization in Congress.
- Results:
  - Senators use more divisive language when they are up for election.
  - House members respond to greater news coverage with more divisive language.
- Interpretation:
  - Electoral incentives and transparency are important contributors to polarization of U.S. politics.

Ash, Morelli, and Van Weelden (2017)

- Approach:
  - Adopt the measure from Gentzkow and Shapiro to analyze divisiveness/polarization in Congress.
- Results:
  - Senators use more divisive language when they are up for election.
  - House members respond to greater news coverage with more divisive language.
- Interpretation:
  - Electoral incentives and transparency are important contributors to polarization of U.S. politics.

Outline

1. OLS Regression Baseline
2. Regression Models
   - Principal Components and PLS
   - Regularized Linear Regression
   - Ensemble Methods: Random Forests and XGBoost
   - Structural Topic Model
3. Political Economy of Tax Code and Tax Revenues
   - Data Construction
   - Predicting Tax Revenues with Tax Code Text
   - Political Party Control and Tax Policy

Overview

- This section enumerates a set of machine learning models for prediction of a real-valued outcome with high-dimensional X.

Train/Test Split

- The models are evaluated using cross-validation and out-of-sample fit:
  - the model fit in a held out test sample – correlation between true Y and model-predicted $\hat{Y}$
- [demo_code.py]

Outline

Principal Component Regression

- The "classic" way to deal with high-dimensionality is principal components regression.
  - Take the first few principal components of X and use those as predictors
  - Popular in macroeconomics and finance.
- How does it work?
  - Constructs the best linear combination of predictors to explain variance in the data set.

Principal Component Regression

- The "classic" way to deal with high-dimensionality is principal components regression.
  - Take the first few principal components of X and use those as predictors
  - Popular in macroeconomics and finance.
- How does it work?
  - Constructs the best linear combination of predictors to explain variance in the data set.

Principal Component Regression

- The "classic" way to deal with high-dimensionality is principal components regression.
  - Take the first few principal components of X and use those as predictors
  - Popular in macroeconomics and finance.
- How does it work?
  - Constructs the best linear combination of predictors to explain variance in the data set.

Pros and Cons of PCA

- Advantages:
  - components are orthogonal by construction
  - good performance on many tasks in practice
- Disadvantages
  - lose (potentially a lot of) predictive information from X
  - Coefficients are not easily interpretable.
- [demo_code.py]

Pros and Cons of PCA

- Advantages:
  - components are orthogonal by construction
  - good performance on many tasks in practice
- Disadvantages
  - lose (potentially a lot of) predictive information from X
  - Coefficients are not easily interpretable.
- [demo_code.py]

Pros and Cons of PCA

- Advantages:
  - components are orthogonal by construction
  - good performance on many tasks in practice
- Disadvantages
  - lose (potentially a lot of) predictive information from X
  - Coefficients are not easily interpretable.
- [demo_code.py]

Pros and Cons of PCA

- Advantages:
  - components are orthogonal by construction
  - good performance on many tasks in practice
- Disadvantages
  - lose (potentially a lot of) predictive information from X
  - Coefficients are not easily interpretable.
- [demo_code.py]

Partial Least Squares

- PLS is related to PCA; high-dimensional data projected down to lower-dimensional space (orthogonoalized components) while retaining as much information as possible (Chun and Keles, 2010).
- Rather than maximizing the explained variance in X, PLS constructs components to maximize predictiveness for an outcome variable (Y).
- An interesting feature of PLS is that it is generalizable to a multi-dimensional real-valued outcome.
- [demo_code.py]

Partial Least Squares

- PLS is related to PCA; high-dimensional data projected down to lower-dimensional space (orthogonoalized components) while retaining as much information as possible (Chun and Keles, 2010).
- Rather than maximizing the explained variance in X, PLS constructs components to maximize predictiveness for an outcome variable (Y).
- An interesting feature of PLS is that it is generalizable to a multi-dimensional real-valued outcome.
- [demo_code.py]

Partial Least Squares

- PLS is related to PCA; high-dimensional data projected down to lower-dimensional space (orthogonoalized components) while retaining as much information as possible (Chun and Keles, 2010).
- Rather than maximizing the explained variance in X, PLS constructs components to maximize predictiveness for an outcome variable (Y).
- An interesting feature of PLS is that it is generalizable to a multi-dimensional real-valued outcome.
- [demo_code.py]

Outline

Lasso, Ridge, and Elastic Net

- Lasso and ridge regression are tools for dealing with large feature sets where:
  - models have multicollinearity that causes bias
  - models tend to overfit
  - models are computationally costly to fit

L1 and L2 Penalties

- Lasso uses L1 Penalty:
  - penalizes coefficients by absolute value of magnitude
  - minimize squared error, plus sum of absolute value of coefficients.
- Ridge uses L2 Penalty:
  - penalizes coefficients by square of magnitude.
  - minimize squared error, plus sum of squared coefficients.
- Elastic Net uses both.

L1 and L2 Penalties

- Lasso uses L1 Penalty:
  - penalizes coefficients by absolute value of magnitude
  - minimize squared error, plus sum of absolute value of coefficients.
- Ridge uses L2 Penalty:
  - penalizes coefficients by square of magnitude.
  - minimize squared error, plus sum of squared coefficients.
- Elastic Net uses both.

L1 and L2 Penalties

- Lasso uses L1 Penalty:
  - penalizes coefficients by absolute value of magnitude
  - minimize squared error, plus sum of absolute value of coefficients.
- Ridge uses L2 Penalty:
  - penalizes coefficients by square of magnitude.
  - minimize squared error, plus sum of squared coefficients.
- Elastic Net uses both.

Regularized Linear Equation

- OLS model:

$$Y_i = X_i'\beta + \varepsilon_i$$

- Elastic Net Model:

$$Y_i = X_i'\beta + \varepsilon_i + \lambda_1 \sum_k |\beta_k| + \lambda_2 \sum \beta_k^2$$

  - λ_1, L1 penalty parameter (Lasso)
  - λ_2, L2 penalty parameter (Ridge)

Regularized Linear Equation

- OLS model:

$$Y_i = X_i'\beta + \varepsilon_i$$

- Elastic Net Model:

$$Y_i = X_i'\beta + \varepsilon_i + \lambda_1 \sum_k |\beta_k| + \lambda_2 \sum \beta_k^2$$

  - λ_1, L1 penalty parameter (Lasso)
  - λ_2, L2 penalty parameter (Ridge)

How to set λ_1 and λ_2

- Belloni et al (Econometrica 2012) provide results for setting λ_1 to ensure consistent estimates in post-Lasso under sparsity.
- But usually you would just use grid search to maximize cross-fit.

How to set λ_1 and λ_2

- Belloni et al (Econometrica 2012) provide results for setting λ_1 to ensure consistent estimates in post-Lasso under sparsity.
- But usually you would just use grid search to maximize cross-fit.

Practicalities

- Have to standardize predictors (std. dev. $= 1$) so coefficients are penalized symmetrically.
- [demo_code.py]

Outline

Random Forests

- Random Forest Regression Model is a generalization of decision trees to a continuous real-valued outcome.
- Good prediction performance – due to out-of-sample validation being included in the training process.
- Also, interpretable because includes a feature importance ranking.
- [demo_code.py]

XGBoost: Boosted Trees

- An even newer model is XGBoost, which has proved very effective, especially in classification, with minimal tuning.
- [demo_code.py]

Outline

Structural Topic Model = LDA + Metadata

- STM provides two ways to include contextual information:
  - Topic prevalence can vary by metadata
    - e.g. Republicans talk about military issues more then Democrats
  - Topic content can vary by metadata
    - e.g. Republicans talk about military issues differently from Democrats.
- Including context improves the model:
  - may provide accurate estimation (but I haven't seen evidence of this)
  - better qualitative interpretability

Structural Topic Model = LDA + Metadata

- STM provides two ways to include contextual information:
  - Topic prevalence can vary by metadata
    - e.g. Republicans talk about military issues more then Democrats
  - Topic content can vary by metadata
    - e.g. Republicans talk about military issues differently from Democrats.
- Including context improves the model:
  - may provide accurate estimation (but I haven't seen evidence of this)
  - better qualitative interpretability

Structural Topic Model = LDA + Metadata

- STM provides two ways to include contextual information:
  - Topic prevalence can vary by metadata
    - e.g. Republicans talk about military issues more then Democrats
  - Topic content can vary by metadata
    - e.g. Republicans talk about military issues differently from Democrats.
- Including context improves the model:
  - may provide accurate estimation (but I haven't seen evidence of this)
  - better qualitative interpretability

Structural Topic Model = LDA + Metadata

- STM provides two ways to include contextual information:
  - Topic prevalence can vary by metadata
    - e.g. Republicans talk about military issues more then Democrats
  - Topic content can vary by metadata
    - e.g. Republicans talk about military issues differently from Democrats.
- Including context improves the model:
  - may provide accurate estimation (but I haven't seen evidence of this)
  - better qualitative interpretability

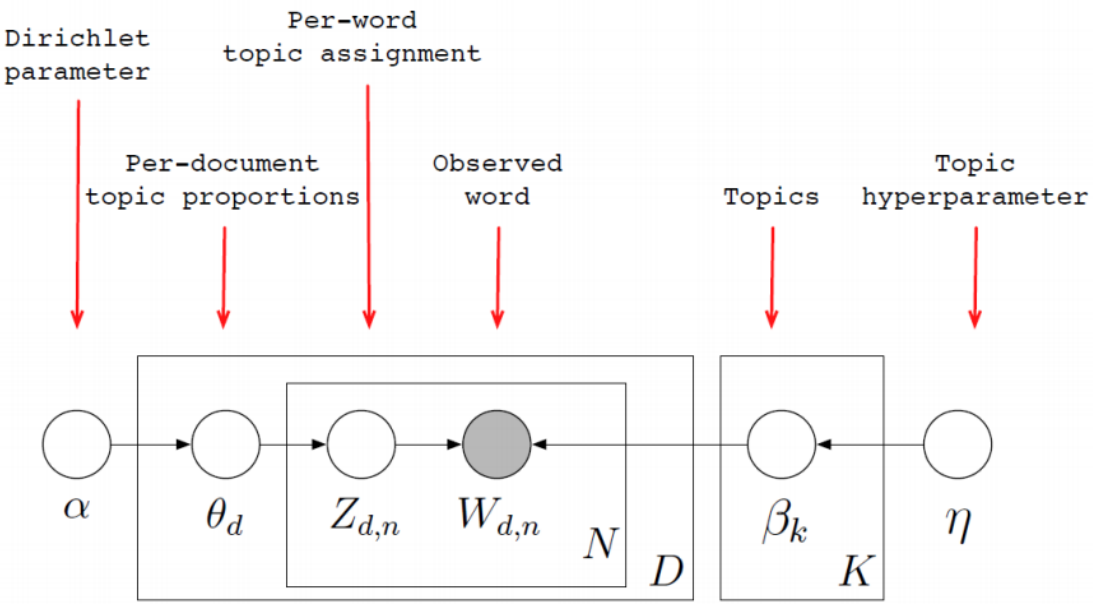
Structural Topic Model = LDA + Metadata

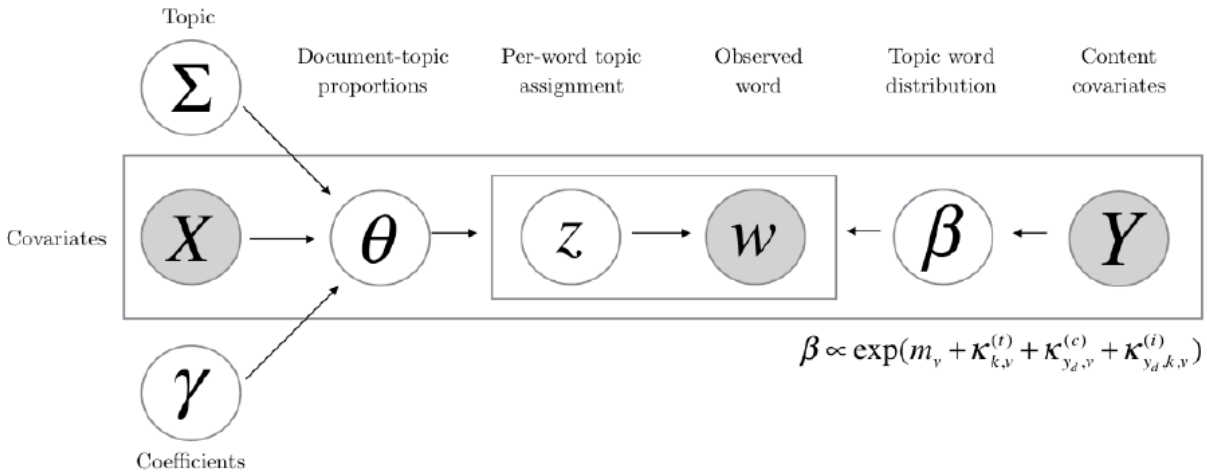
- STM provides two ways to include contextual information:
  - Topic prevalence can vary by metadata
    - e.g. Republicans talk about military issues more then Democrats
  - Topic content can vary by metadata
    - e.g. Republicans talk about military issues differently from Democrats.
- Including context improves the model:
  - may provide accurate estimation (but I haven't seen evidence of this)
  - better qualitative interpretability

Structural Topic Model = LDA + Metadata

- STM provides two ways to include contextual information:
  - Topic prevalence can vary by metadata
    - e.g. Republicans talk about military issues more then Democrats
  - Topic content can vary by metadata
    - e.g. Republicans talk about military issues differently from Democrats.
- Including context improves the model:
  - may provide accurate estimation (but I haven't seen evidence of this)
  - better qualitative interpretability

LDA vs. STM – Illustration





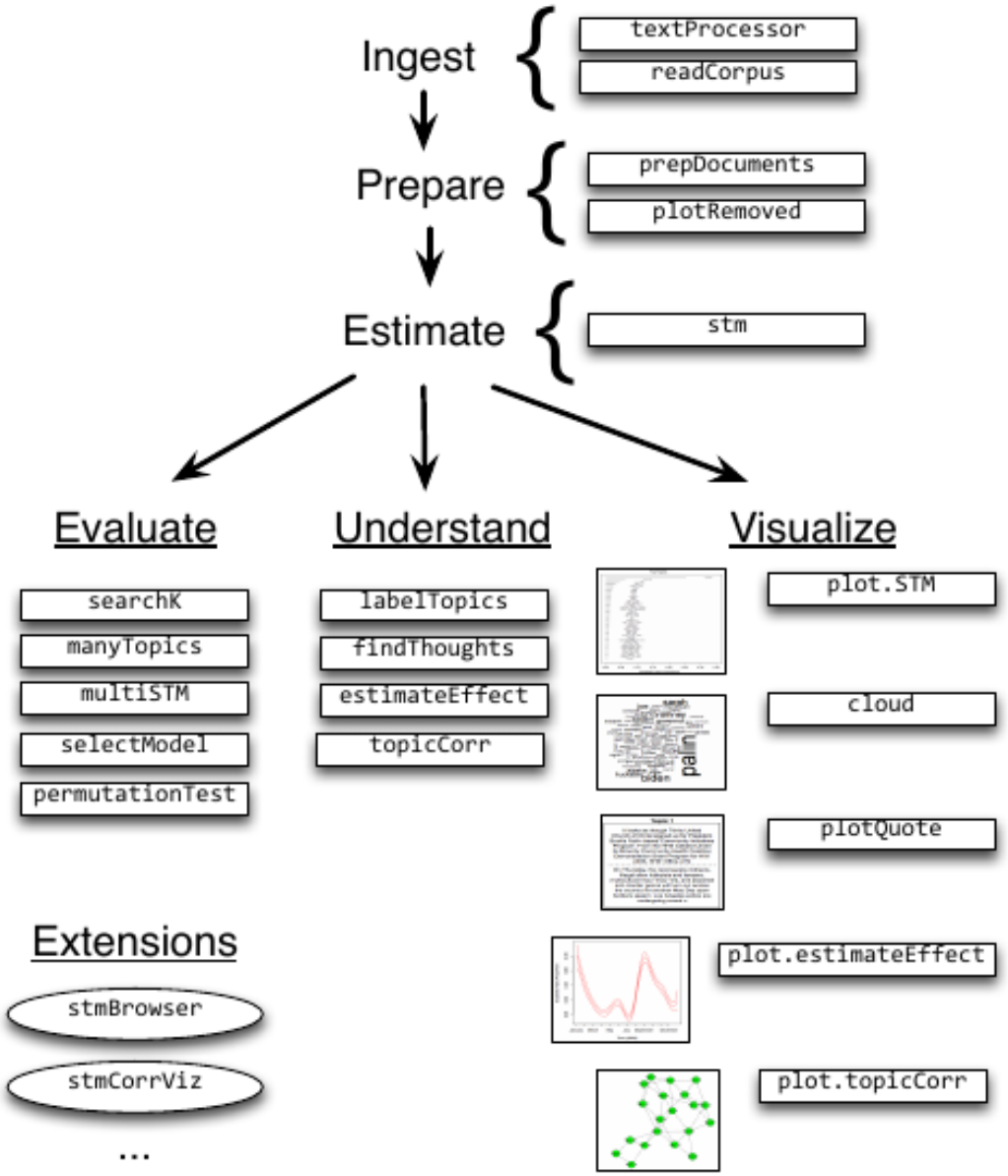
stm Package in R

- Complete workflow: raw texts $\rightarrow$ figures
- Simple regression style syntax using formulas

```
mod.out <- stm(documents,vocab, K=10, prevalence=
~paper + s(time), data=metadata, init.type="Spectral")
```

- many functions for summarization, visualization and checking
- Complete vignette online with examples

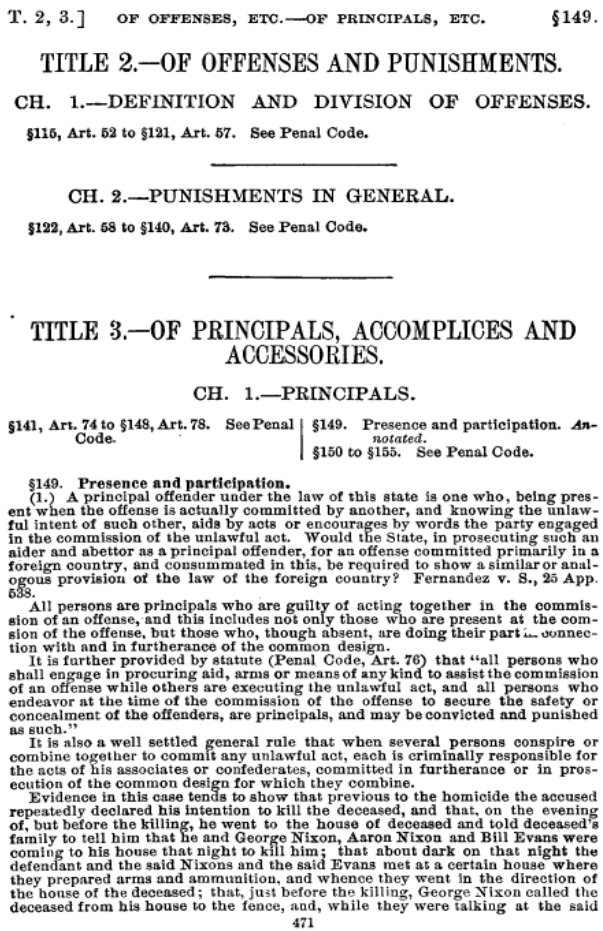
stm has great functions/features



Outline

Outline

Raw Text Data



T. 2, 3.] OF OFFENSES, ETC.—OF PRINCIPALS, ETC. §149.

TITLE 2.—OF OFFENSES AND PUNISHMENTS.

CH. 1.—DEFINITION AND DIVISION OF OFFENSES.

§115, Art. 52 to §121, Art. 57. See Penal Code.

CH. 2.—PUNISHMENTS IN GENERAL.

§122, Art. 58 to §140, Art. 73. See Penal Code.

TITLE 3.—OF PRINCIPALS, ACCOMPLICES AND ACCESSORIES.

CH. 1.—PRINCIPALS.

§141, Art. 74 to §148, Art. 78. See Penal Code.

§149. Presence and participation. *Annotated.*

§150 to §155. See Penal Code.

§149. **Presence and participation.**

(1.) A principal offender under the law of this state is one who, being present when the offense is actually committed by another, and knowing the unlawful intent of such other, aids by acts or encourages by words the party engaged in the commission of the unlawful act. Would the State, in prosecuting such an aider and abettor as a principal offender, for an offense committed primarily in a foreign country, and consummated in this, be required to show a similar or analogous provision of the law of the foreign country? Fernandez v. S., 25 App. 538.

All persons are principals who are guilty of acting together in the commission of an offense, and this includes not only those who are present at the comsion of the offense, but those who, though absent, are doing their part i... connection with and in furtherance of the common design.

It is further provided by statute (Penal Code, Art. 76) that "all persons who shall engage in procuring aid, arms or means of any kind to assist the commission of an offense while others are executing the unlawful act, and all persons who endeavor at the time of the commission of the offense to secure the safety or concealment of the offenders, are principals, and may be convicted and punished as such."

It is also a well settled general rule that when several persons conspire or combine together to commit any unlawful act, each is criminally responsible for the acts of his associates or confederates, committed in furtherance or in prosecution of the common design for which they combine.

Evidence in this case tends to show that previous to the homicide the accused repeatedly declared his intention to kill the deceased, and that, on the evening of, but before the killing, he went to the house of deceased and told deceased's family to tell him that he and George Nixon, Aaron Nixon and Bill Evans were coming to his house that night to kill him; that about dark on that night the defendant and the said Nixons and the said Evans met at a certain house where they prepared arms and ammunition, and whence they went in the direction of the house of the deceased; that, just before the killing, George Nixon called the deceased from his house to the fence, and, while they were talking at the said

471

T. 2, 3.] OF OFFENSES, ETC.-OF PRINCIPALS, ETC.

TITLE 2.-OF OFFENSES AND PUNISHMENTS.
CH. 1.-DEFINITION AND DIVISION OF OFFENSES.
§115, Art. 52 to §121, Art. 57. See Penal Code.
CH. 2.-PUNISHMENTS IN GENERAL.
§122, Art. 58 to §140, Art. 73. See Penal Code.
TITLE 3.-OF PRINCIPALS, ACCOMPLICES AND ACCESSORIES.
CH. 1.-PRINCIPALS.
§141, Art. 74 to §148, Art. 78. See Penal Code. §149. Presence and participation. Annotated.
1 §150 to §155. See Penal Code.
§149. Presence and participation.
(1.) A principal offender under the law of this state is one who, being present when the offense is actually committed by another, and knowing the unlawful intent of such other, aids by acts or encourages by words the party engaged in the commission of the unlawful act. Would the State, in prosecuting such an aider and abettor as a principal offender, for an offense committed primarily in a foreign country, and consummated in this, be required to show a similar or analogous provision of the law of the foreign country? Fernandez v. S., 25 App.
All persons are principals who are guilty of acting together in the commission of an offense, and this includes not only those who are present at the comsion of the offense, but those who, though absent, are doing their part L. connection with and in furtherance of the common design.
It is further provided by statute (Penal Code, Art. 76) that "all persons who shall engage in procuring aid, arms or means of any kind to assist the commission of an offense while others are executing the unlawful act, and all persons who endeavor at the time of the commission of the offense to secure the safety or concealment of the offenders, are principals, and may be convicted and punished as such."
It is also a well settled general rule that when several persons conspire or combine together to commit any unlawful act, each is criminally responsible for the acts of his associates or confederates, committed in furtherance or in prosecution of the common design for which they combine.
Evidence in this case tends to show that previous to the homicide the accused repeatedly declared his intention to kill the deceased, and that, on the evening of, but before the killing, he went to the house of deceased and told deceased's family to tell him that he and George Nixon, Aaron Nixon and Bill Evans were coming to his house that night to kill him; that about dark on that night the defendant and the said Nixons and the said Evans met at a certain house where they prepared arms and ammunition, and whence they went in the direction of the house of the deceased; that, just before the killing, George Nixon called the deceased from his house to the fence, and, while they were talking at the said
471

- Full text of U.S. state session laws: all statutes enacted by state legislatures.
- I segmented text into individual bills, acts, and resolutions (samples checked by RA's); 1.56 million statutes for the years 1963 through 2010.

Construction of Text Features

"`Eligible individuals must pay sales and use tax on foreign purchases.`"

- "Content" Phrases:
  - Stemmed noun and verb phrases, using parts-of-speech sequences based on Denny et al. (2015), extended for purposes of legal language:

    "`elig_individu must_pay sale_and_use_tax foreign_purchas`"

- "Style" N-grams:
  - Construct N-grams from sequences of function words, part-of-speech tags, and punctuation.
    - $N = 1$: `A, N, must, V, A, and, A, N, on, A, N, .`
    - $N = 2$: `A_N, N_must, must_V, V_A, A_and, and_A, A_N, N_on, on_A, A_, N_.` (etc.)

Construction of Text Features

"`Eligible individuals must pay sales and use tax on foreign purchases.`"

- "Content" Phrases:
  - Stemmed noun and verb phrases, using parts-of-speech sequences based on Denny et al. (2015), extended for purposes of legal language:

"`elig_individu must_pay sale_and_use_tax foreign_purchas`"

- "Style" N-grams:
  - Construct N-grams from sequences of function words, part-of-speech tags, and punctuation.
  - $N = 1$: `A, N, must, V, A, and, A, N, on, A, N, .`
  - $N = 2$: `A_N, N_must, must_V, V_A, A_and, and_A, A_N, N_on, on_A, A_, N_.` (etc.)

Construction of Text Features

"`Eligible individuals must pay sales and use tax on foreign purchases.`"

- "Content" Phrases:
  - Stemmed noun and verb phrases, using parts-of-speech sequences based on Denny et al. (2015), extended for purposes of legal language:

"`elig_individu must_pay sale_and_use_tax foreign_purchas`"

- "Style" N-grams:
  - Construct N-grams from sequences of function words, part-of-speech tags, and punctuation.
  - $N = 1$: `A, N, must, V, A, and, A, N, on, A, N, .`
  - $N = 2$: `A_N, N_must, must_V, V_A, A_and, and_A, A_N, N_on, on_A, A_, N_.` (etc.)

Construction of Text Features

"`Eligible individuals must pay sales and use tax on foreign purchases.`"

- "Content" Phrases:
  - Stemmed noun and verb phrases, using parts-of-speech sequences based on Denny et al. (2015), extended for purposes of legal language:

"`elig_individu must_pay sale_and_use_tax foreign_purchas`"

- "Style" N-grams:
  - Construct N-grams from sequences of function words, part-of-speech tags, and punctuation.
    - $N = 1$: `A, N, must, V, A, and, A, N, on, A, N, .`
    - $N = 2$: `A_N, N_must, must_V, V_A, A_and, and_A, A_N, N_on, on_A, A_, N_.` (etc.)

Construction of Text Features

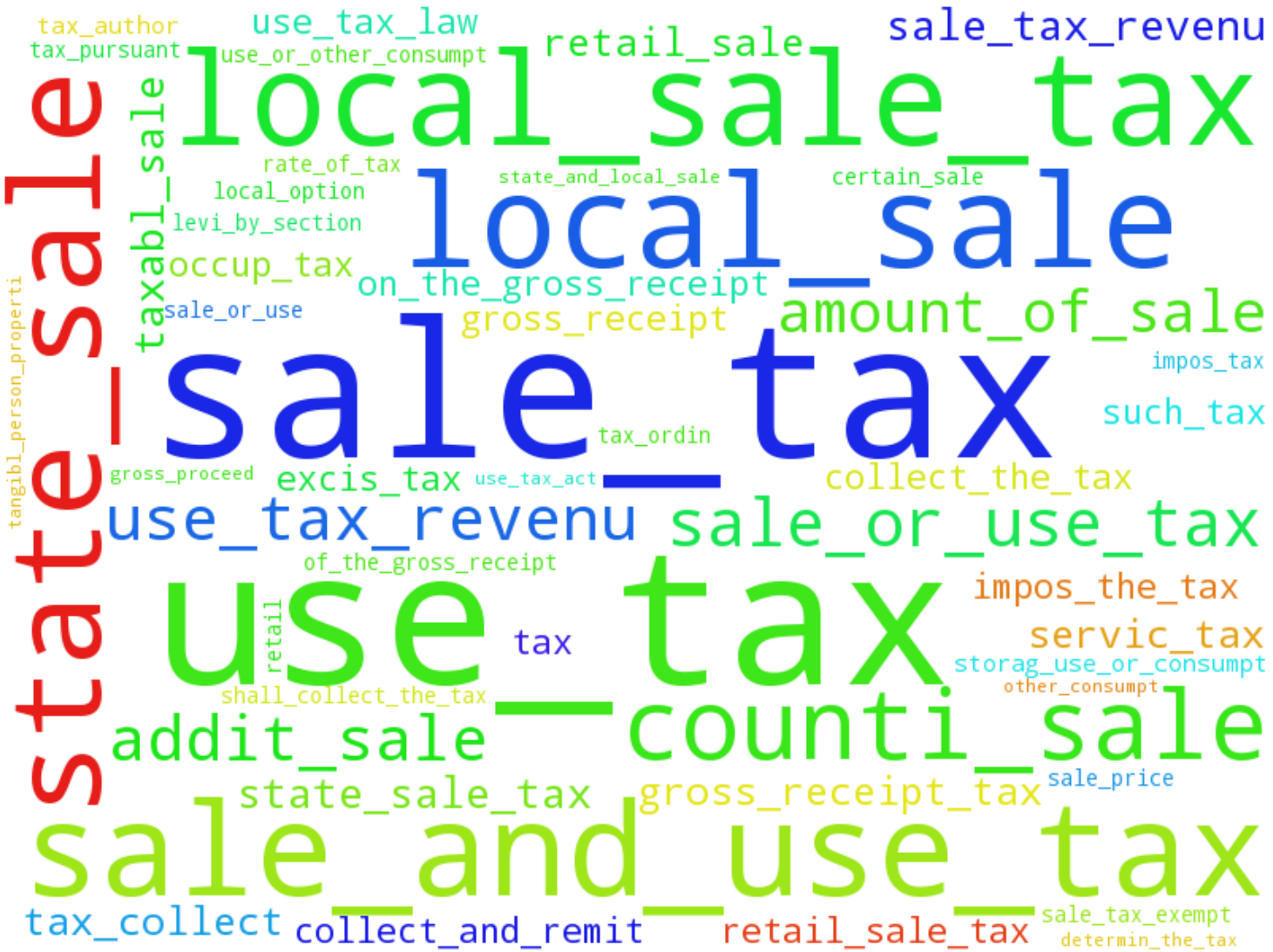
"`Eligible individuals must pay sales and use tax on foreign purchases.`"

- "Content" Phrases:
  - Stemmed noun and verb phrases, using parts-of-speech sequences based on Denny et al. (2015), extended for purposes of legal language:

"`elig_individu must_pay sale_and_use_tax foreign_purchas`"

- "Style" N-grams:
  - Construct N-grams from sequences of function words, part-of-speech tags, and punctuation.
  - $N = 1$: `A, N, must, V, A, and, A, N, on, A, N, .`
  - $N = 2$: `A_N, N_must, must_V, V_A, A_and, and_A, A_N, N_on, on_A, A_, N_.` (etc.)

Extract Tax Law Language using Word2Vec



- A statute that is geometrically close to "sales tax" in Word2Vec space is topically related to sales tax.

Classifying Statutes by Relation to Tax Law

- Each statute k gets a weighting $S(k,r) \in [-1,1]$, the cosine similarity to $r \in \{$"personal income tax", "sales tax"$\}$.
- Text feature variable x_{st}^{ir}:
  - Relative frequency of feature i, state s, time t
  - In statutes related to source $r \in \{$income tax, sales tax$\}$.
- Residualized on a state-rate fixed effect and party-year fixed effect.

Outline

Partial Least Squares

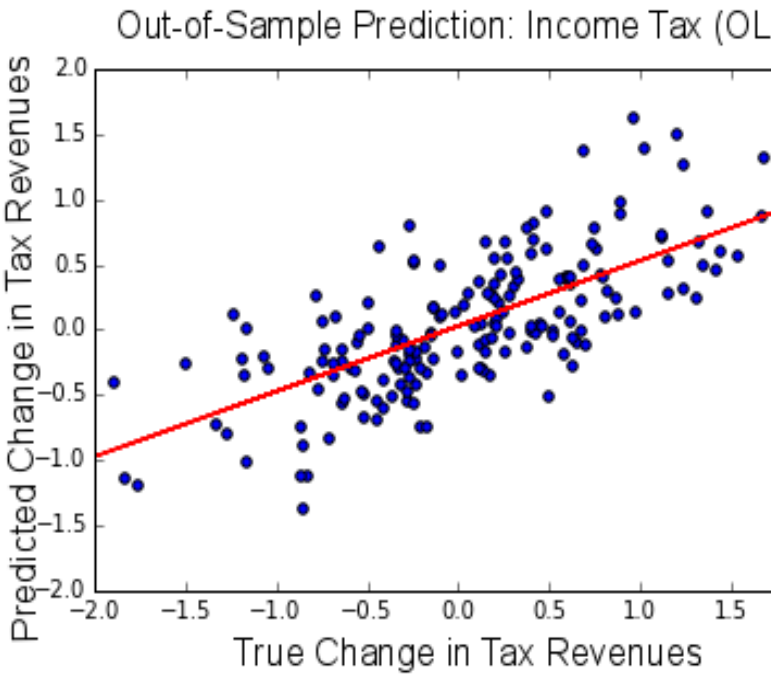
- Need to form predictions of revenue changes based on tax code changes with high-dimensional multicollinear data.

$$y_{st} = \mathbf{x}'_{st}\beta_r + \varepsilon_{st}$$

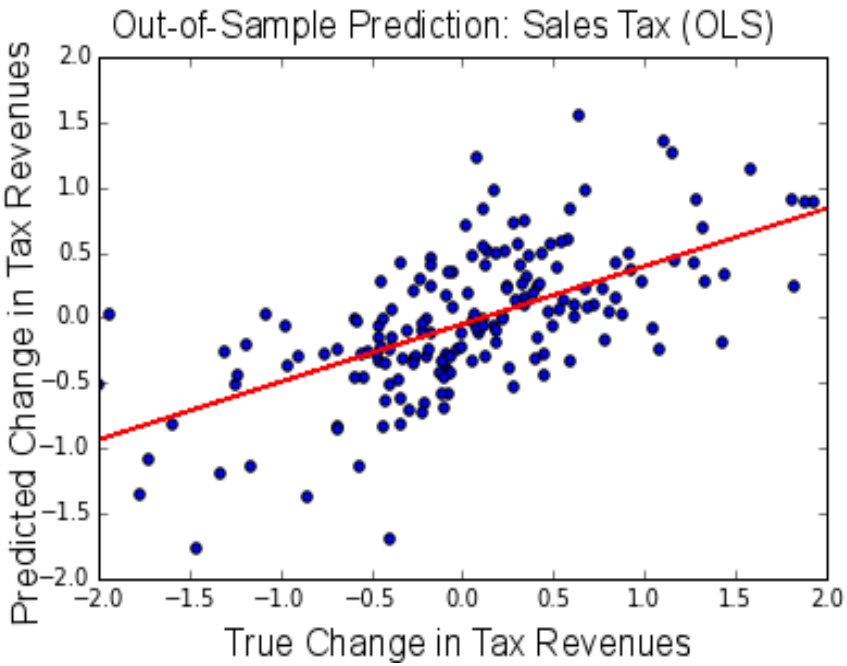
- Solution: Partial Least Squares regression (PLS)

Out-of-sample PLS predictions of tax revenue changes

Income Tax



Sales Tax



Weak predictors filtered out; 80% training, 20% testing sample.

- Predicted change in revenue (vertical axis), plotted against true change in revenue (horizontal axis).
- Correlations between truth and prediction: 0.89 and 0.84.

PLS Comments

- This method also obtains good out-of-sample predictiveness for corporate income tax and estate tax.
- The classification of statutes using Word2Vec matters; statutes related to "sales tax" cannot predict personal income tax changes nearly as well, and vice versa (about 30% worse out-of-sample correlation).
- The style n-grams (rather than content phrases) also predict quite well.
- Random forest regression also does well, but not as well as PLS.

PLS Comments

- This method also obtains good out-of-sample predictiveness for corporate income tax and estate tax.
- The classification of statutes using Word2Vec matters; statutes related to "sales tax" cannot predict personal income tax changes nearly as well, and vice versa (about 30% worse out-of-sample correlation).

- The style n-grams (rather than content phrases) also predict quite well.
- Random forest regression also does well, but not as well as PLS.

Outline

State politics data

- Democrat and Republican power shares:
  - lower house seat shares
  - upper house seat shares
  - governor vote shares
- Used in many previous papers on state politics and state finances (e.g. Besley and Case 2003, Reed 2006, Leigh 2008).

Differences-in-Differences Approach

- Given outcome variable y_{st} (tax rates and tax revenues) for state s at year t, estimate

$$y_{st} = \alpha_{st} + \delta D_{st} + f(d_{st}) + \varepsilon_{st}$$

- α_{st}: state and time fixed effects, state time trends
- $D_{st} \in \{0,1,2,3\}$, the number of state government bodies (lower house, upper house, and governor) controlled by Democrats, with 0.5 assigned for tied legislatures.
- $f(d_{st})$, polynomials in power shares for each government body (seat shares for legislatures, vote shares for governor), separately for below and above the cutoffs.
- Cluster standard errors by state (Bertrand et al. 2004).

Differences-in-Differences Approach

- Given outcome variable y_{st} (tax rates and tax revenues) for state s at year t, estimate

$$y_{st} = \alpha_{st} + \delta D_{st} + f(d_{st}) + \varepsilon_{st}$$

- α_{st}: state and time fixed effects, state time trends
- $D_{st} \in \{0,1,2,3\}$, the number of state government bodies (lower house, upper house, and governor) controlled by Democrats, with 0.5 assigned for tied legislatures.
- $f(d_{st})$, polynomials in power shares for each government body (seat shares for legislatures, vote shares for governor), separately for below and above the cutoffs.
- Cluster standard errors by state (Bertrand et al. 2004).

Differences-in-Differences Approach

- Given outcome variable y_{st} (tax rates and tax revenues) for state s at year t, estimate

$$y_{st} = \alpha_{st} + \delta D_{st} + f(d_{st}) + \varepsilon_{st}$$

- α_{st}: state and time fixed effects, state time trends
- $D_{st} \in \{0,1,2,3\}$, the number of state government bodies (lower house, upper house, and governor) controlled by Democrats, with 0.5 assigned for tied legislatures.
- $f(d_{st})$, polynomials in power shares for each government body (seat shares for legislatures, vote shares for governor), separately for below and above the cutoffs.
- Cluster standard errors by state (Bertrand et al. 2004).

Differences-in-Differences Approach

- Given outcome variable y_{st} (tax rates and tax revenues) for state s at year t, estimate

$$y_{st} = \alpha_{st} + \delta D_{st} + f(d_{st}) + \varepsilon_{st}$$

- α_{st}: state and time fixed effects, state time trends
- $D_{st} \in \{0,1,2,3\}$, the number of state government bodies (lower house, upper house, and governor) controlled by Democrats, with 0.5 assigned for tied legislatures.
- $f(d_{st})$, polynomials in power shares for each government body (seat shares for legislatures, vote shares for governor), separately for below and above the cutoffs.
- Cluster standard errors by state (Bertrand et al. 2004).

Party control has larger effect on revenue than on rates

	(1) Marginal Tax Rate	(2) Tax Revenue
Effect of Democrat Power		
Income Tax	0.0384	0.0460
	(0.0782)	(0.0811)
[% change]	**[3.1 %]**	**[7.4%]**
Sales Tax	-0.0766	-0.176
	(0.0644)	(0.114)
[%. change]	**[-3.9 %]**	**[-21.8 %]**
N	3091	3091
FE's and Trends	Yes	Yes

Observation is a state-source-session. Regressions include linear polynomials in the forcing variables for both houses and governor, separately for values above and below the cutoffs. Outcome variables are standardized. Standard errors in parentheses, clustered by state.

Regression Model for Tax Code Effect

- Define $\tilde{g}_{st}$, the predicted change in tax revenue for state s, time t, due to tax code changes, using regularized 2SLS estimates.

- Regress

$$\tilde{g}_{st} = \alpha_{st} + \phi D_{st} + f(d_{st}) + \varepsilon_{st}$$

to obtain the diffs-in-diffs effect of Democrat control, $\hat{\phi}$, on the predicted tax revenue change from the effective tax code.

  - $\tilde{g}_{st}$ is standardized: $\hat{\phi}$ can be interpreted as the predicted standard-deviations change in revenue due to tax code changes associated with Democrat control of an additional wing of state government.

Regression Model for Tax Code Effect

- Define $\tilde{g}_{st}$, the predicted change in tax revenue for state s, time t, due to tax code changes, using regularized 2SLS estimates.

- Regress

$$\tilde{g}_{st} = \alpha_{st} + \phi D_{st} + f(d_{st}) + \varepsilon_{st}$$

to obtain the diffs-in-diffs effect of Democrat control, $\hat{\phi}$, on the predicted tax revenue change from the effective tax code.

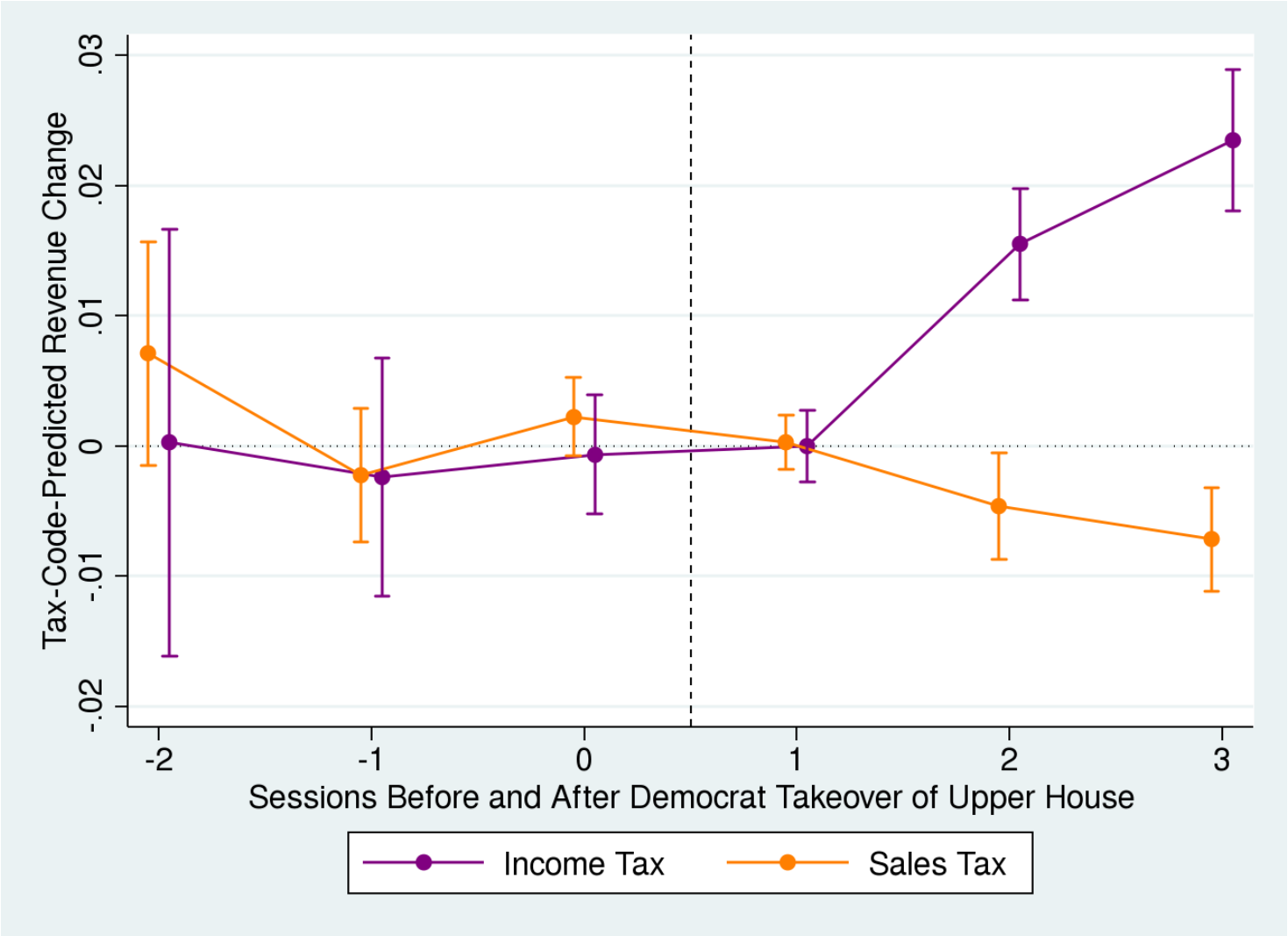
  - $\tilde{g}_{st}$ is standardized: $\hat{\phi}$ can be interpreted as the predicted standard-deviations change in revenue due to tax code changes associated with Democrat control of an additional wing of state government.

Effect of party control on text-predicted tax revenue

	Effect on $\tilde{g}$			
	(1)	(2)	(3)	(4)
Income Tax				
Democrat Power	0.0992**	0.144**	0.138**	0.145**
	(0.0337)	(0.0478)	(0.0458)	(0.0418)
Sales Tax				
Democrat Power	-0.0324	-0.0677*	-0.0829*	-0.0780*
	(0.0254)	(0.0311)	(0.0326)	(0.0310)
FE's/Trends	X	X	X	X
Forcing Var Polys		X	X	X
Lagged Covariates			X	X
Lagged Dep. Var.				X

Democrat Power is number of government bodies controlled by Democrats.
$N = 3{,}588$ observations, state-source-session. Outcome variables are standardized.
Standard errors in parentheses, clustered by state. * p<0.05, ** p<0.01.

Effect of Democrat Takeover on Tax Code Language



Event study graphs for change in text-predicted revenue before and after Democratic takeover of upper house of legislature. The vertical axis is the metric for state-predicted revenue $\tilde{g}$, as described in the text. The horizontal axis is years before and after a change in political control. Republican takeovers are also included, with the sign of the outcome variable reversed.