Change in Local Healthy Food Retail Environment through Interactions in Population, Race and Nativity

David Wutchiett, Tanya Kaufman, Daniel Sheehan, Kathryn Neckerman, Kayip Kwan, Andrew Rundle, Stephen Mooney, Jeff Goldsmith, and Gina Lovasi

Population Association of America
Annual Meeting 2015

Healthy Food Environment Matters

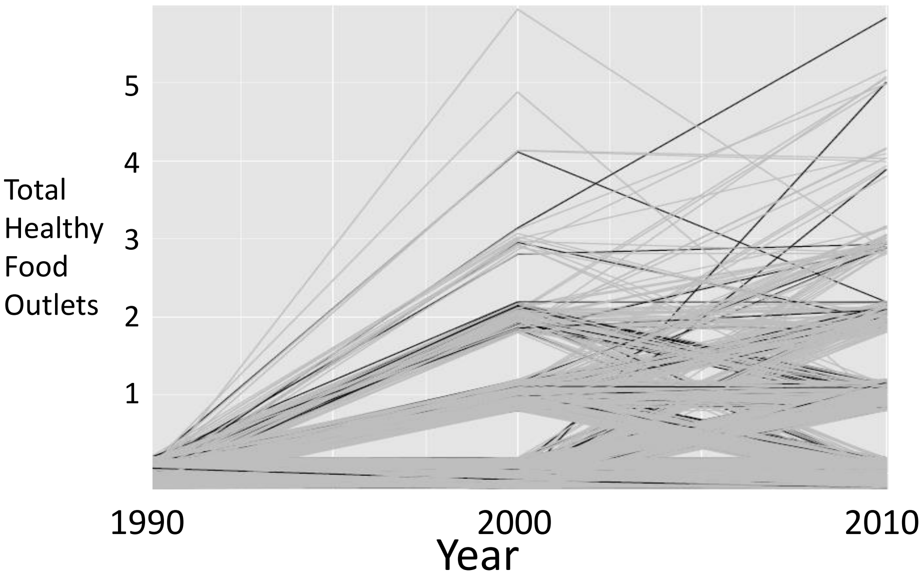
- Local characteristics and demographics are associated with presence of healthy food retail outlets. Foreign born populations have been associated with greater commercial resources including grocery stores.

Healthy food outlets = {large supermarkets, fruit & vegetable markets, natural food markets & nut stores, fish markets}

A Goal of Explaining Local Change

- How are local characteristics linked to change over time?
- Gain insight into processes leading to disparities.

Healthy Food Outlet Count across Decades for Tracts without Outlets in 1990



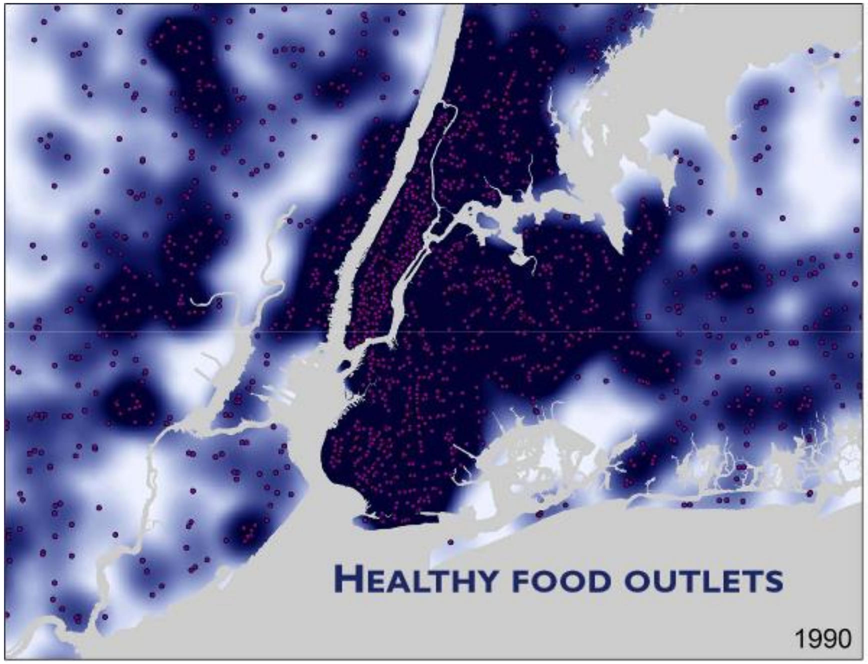
Business Data

National Establishment Time Series (NETS)

- Longitudinal ‘census of U.S. businesses’
- Geocoded to addresses and zip code
- 8 digit SIC code classifications
- 21 years of data
- 23 counties in NYC metropolitan area

Dependent Variable:

- Indicator for whether the was an increase in healthy food outlets.



Prepared by Danny Sheehan

Population Data and Tracts

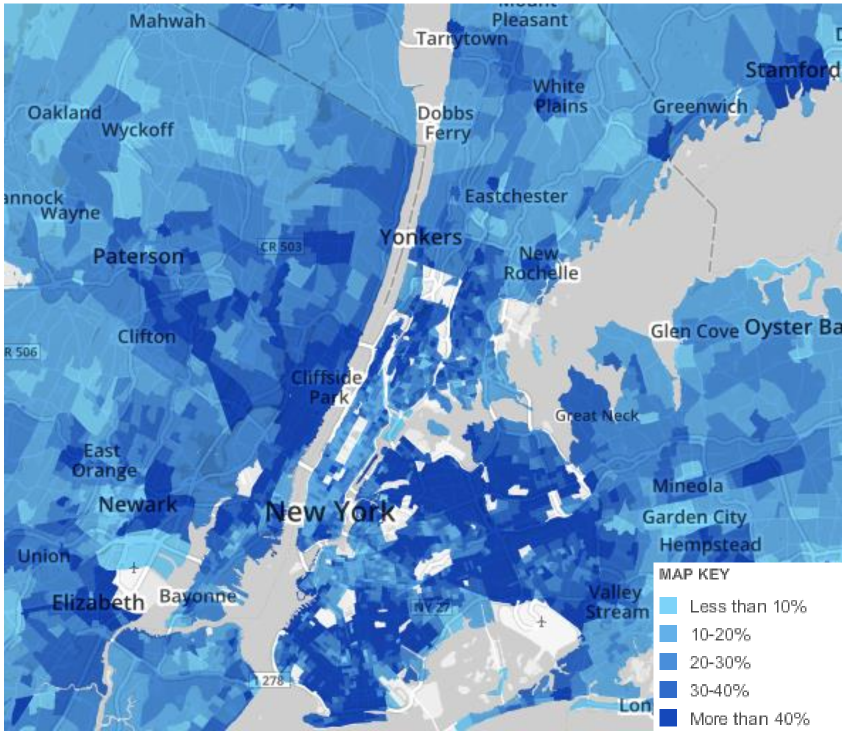
Population:

- U.S. Census
  - Decennial Census (1990, 2000, 2010)
  - American Community Survey (2007-2010)
- Geographic size
- Geographic adjacency

Explanatory Variables:

- Median Income
- % Poverty
- % Foreign Born
- % Non-Hispanic Black
- % Hispanic
- % Asian
- Total Population
- Census Tract Geographic Size

% Foreign Born Population



Derived Variables

Change in Tract Characteristics

$$\nabla\%ForeignBorn_{i,t_n} = \%ForeignBorn_{t_n} - \%ForeignBorn_{t_{n-1}}$$

Adjacent Tract Characteristic

$$\%ForeignBorn\_Adjacent_{i,t_n} = \frac{\sum_j^{N(i)} \%Poverty_{j,t_n} * Population_{j,t_n}}{\sum_j^{N(i)} Population_{j,t_n}}$$

Adjacent Tract **Change**

$$\nabla\%ForeignBorn\_Adjacent_{i,t_n} = \frac{\sum_j^{N(i)} \nabla\%\ Poverty_{j,t_n} * Population_{j,t_n}}{\sum_j^{N(i)} Population_{j,t_n}}$$

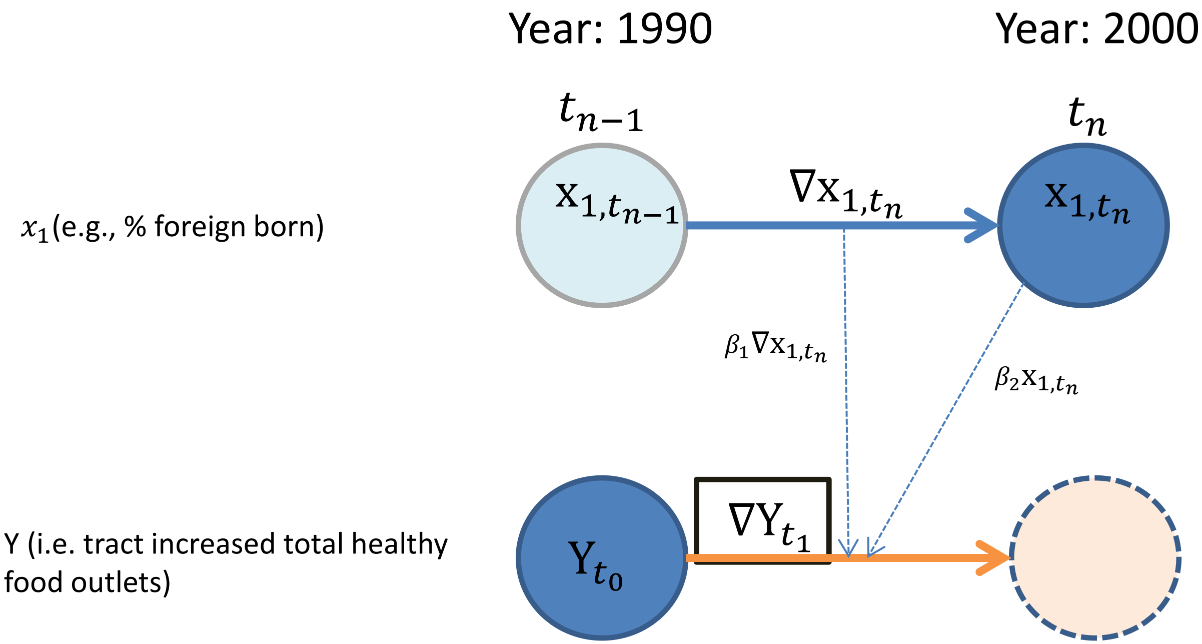
www.armellecaron.fr/works/les-villes-rangees/

Conceptualizing the Model

$$
\begin{aligned}
\mathit{Change\ in\ HealthyFoodOutlets} = {} & \%\mathit{ForeignBorn} + \\
& \%\mathit{ForeignBorn\_Change} + \\
& \%\mathit{ForeignBorn\_Adjacent} + \\
& \%\mathit{ForeignBorn\_Adjacent\_Change} + \ldots
\end{aligned}
$$

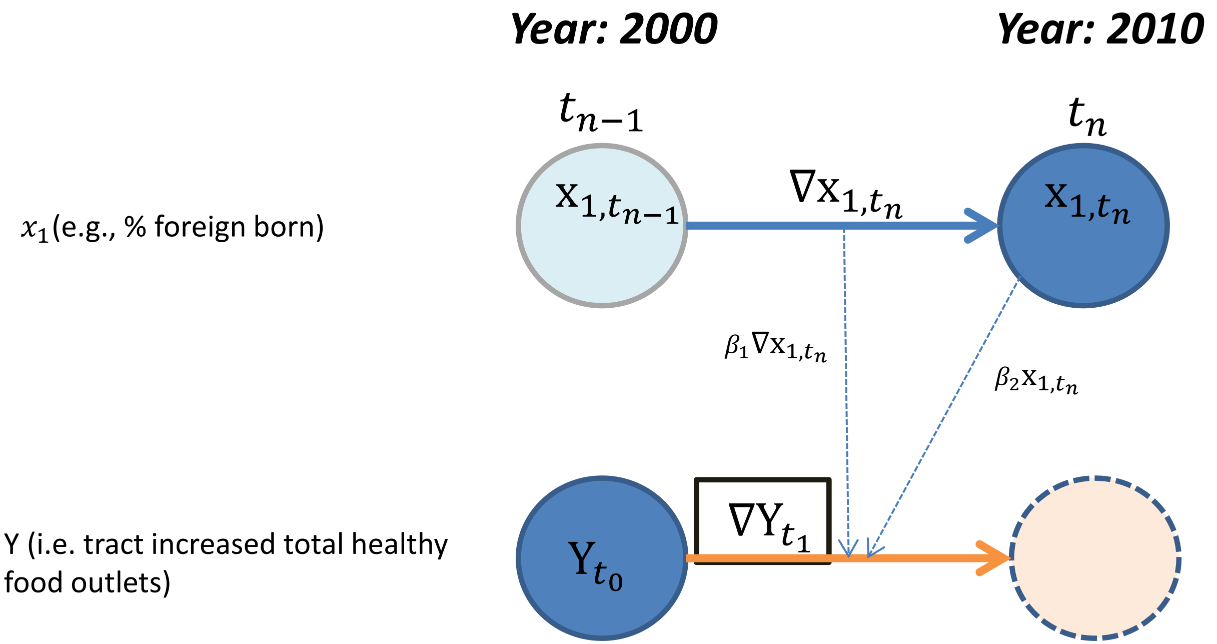
Conceptualizing the Model

$$\begin{aligned} \mathit{Change\ in\ HealthyFoodOutlets} = {} & \%ForeignBorn + \\ & \%ForeignBorn\_Change + \\ & \%ForeignBorn\_Adjacent + \\ & \%ForeignBorn\_Adjacent\_Change + \ldots \end{aligned}$$



Conceptualizing the Model

Change in HealthyFoodOutlets = *%ForeignBorn* +
%ForeignBorn_Change +
%ForeignBorn_Adjacent +
%ForeignBorn_Adjacent_Change + …



Year: 2000

Year: 2010

t_{n-1}

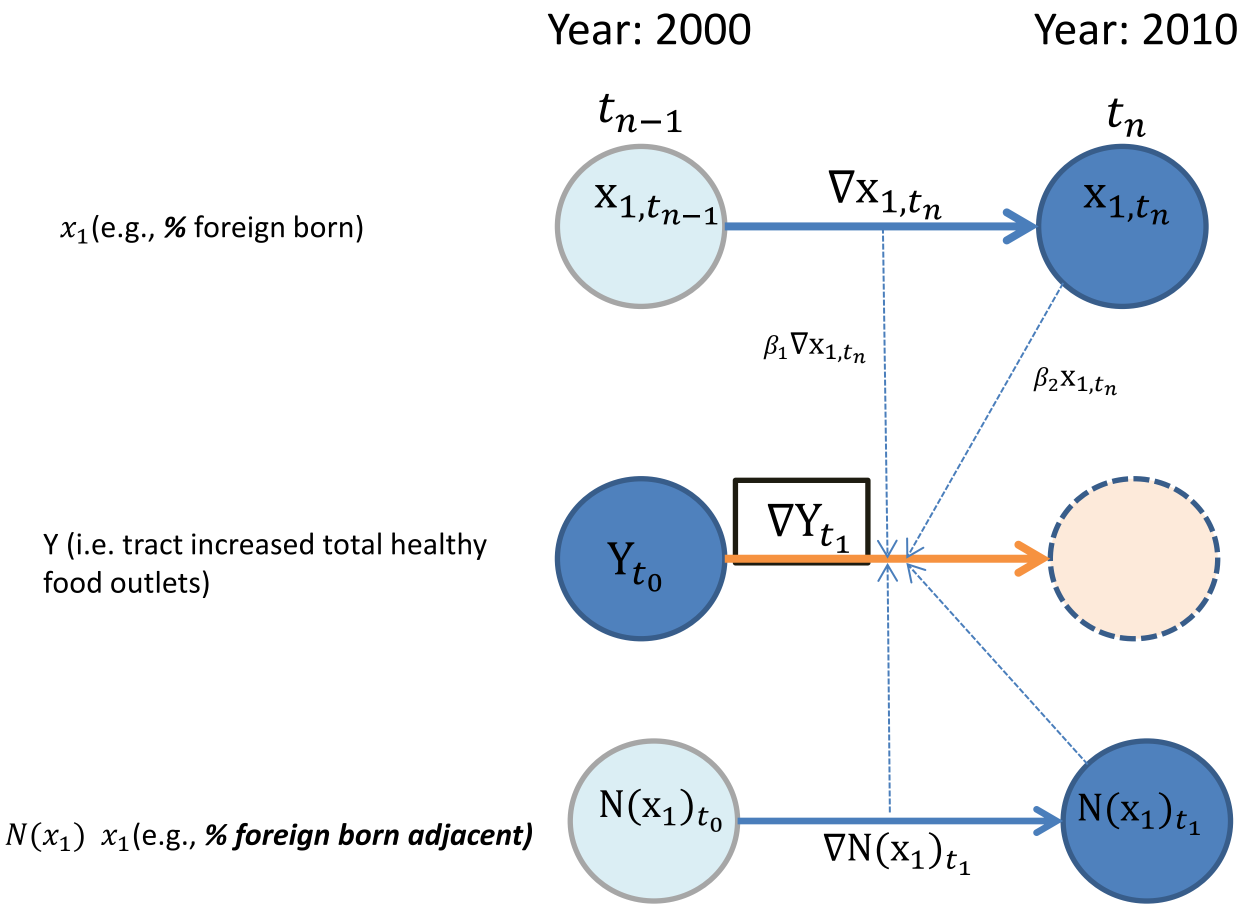
t_n

x_1 (e.g., % foreign born)

Y (i.e. tract increased total healthy food outlets)

$\mathrm{x}_{1,t_{n-1}}$
$\nabla \mathrm{x}_{1,t_n}$
x_{1,t_n}
$\beta_1 \nabla \mathrm{x}_{1,t_n}$
$\beta_2 \mathrm{x}_{1,t_n}$
Y_{t_0}
$\nabla \mathrm{Y}_{t_1}$

Conceptualizing the Model



Year: 2000

Year: 2010

t_{n-1}

t_n

x_1(e.g., ***%*** foreign born)

$\mathrm{x}_{1,t_{n-1}}$

$\nabla \mathrm{x}_{1,t_n}$

x_{1,t_n}

$\beta_1 \nabla \mathrm{x}_{1,t_n}$

$\beta_2 \mathrm{x}_{1,t_n}$

Y (i.e. tract increased total healthy food outlets)

Y_{t_0}

$\nabla \mathrm{Y}_{t_1}$

$N(x_1)$ x_1(e.g., ***% foreign born adjacent)***

$\mathrm{N}(\mathrm{x}_1)_{t_0}$

$\nabla \mathrm{N}(\mathrm{x}_1)_{t_1}$

$\mathrm{N}(\mathrm{x}_1)_{t_1}$

Expansions & Considerations

Interactions in Explanatory Variables:

38 main effects
703 interactions
9056 observations.

Risks:

Expansions & Considerations

Interactions in Explanatory Variables:

38 main effects
703 interactions
9056 observations.

Risks:

With many parameters there is risk of **overfitting relationships**

Expansions & Considerations

Interactions in Explanatory Variables:
38 main effects
703 interactions
9056 observations.

Risks:
With many parameters there is risk of **overfitting relationships**

Models suffer in terms of **interpretability**

Expansions & Considerations

Interactions in Explanatory Variables:
38 main effects
703 interactions
9056 observations.

Risks:
With many parameters there is risk of **overfitting relationships**

Models suffer in terms of **interpretability**

With overfitting, model **generalizability** to new data declines

Expansions & Considerations

Interactions in Explanatory Variables:

38 main effects
703 interactions
9056 observations.

Risks:

With many parameters there is risk of **overfitting relationships**

Models suffer in terms of **interpretability**

With overfitting, model **generalizability** to new data declines

Approaches:

Model Validation, **Resampling** and **Shrinkage/Regularization**

Expansions & Considerations

Interactions in Explanatory Variables:
38 main effects
703 interactions
9056 observations.

Risks:
With many parameters there is risk of **overfitting relationships**

Models suffer in terms of **interpretability**

With overfitting, model **generalizability** to new data declines

Approaches:
Model Validation, **Resampling** and **Shrinkage/Regularization**

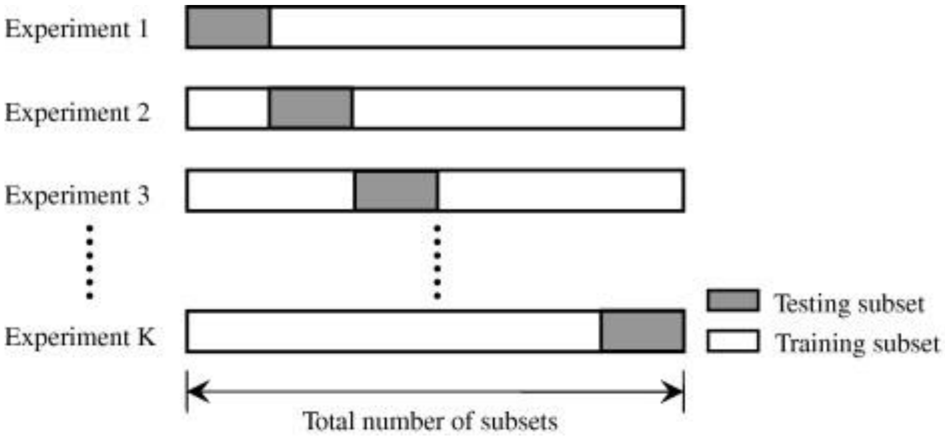
through **Cross-Validation, Bootstrapping, Lasso Shrinkage,** and **Model Averaging**

Generalization & Model Validation

Concerns regarding: **generalizability**, **interpretability**, and risk of **overfitting** relationships

K-fold cross-validation (CV)

- Divide into k subsets
- Exclude one subset as testing set; remaining k-1 subsets are combined to be the training set
- Fit model to training set, assess model on testing set

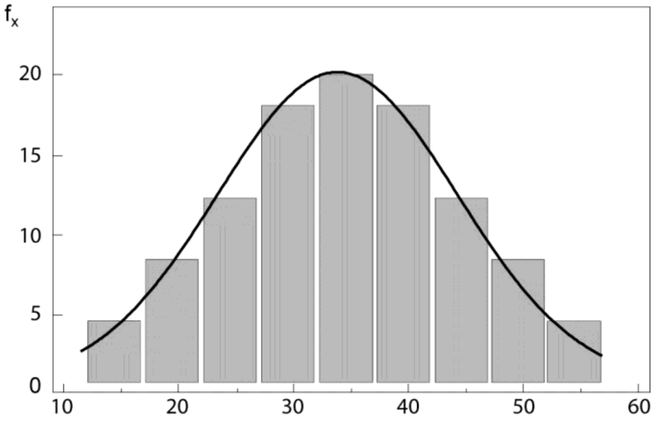


Generalization & Model Validation

Concerns regarding: **generalizability**, **interpretability**, and risk of **overfitting** relationships

Non-parametric bootstrap

- Resample from empirical distribution with replacement.



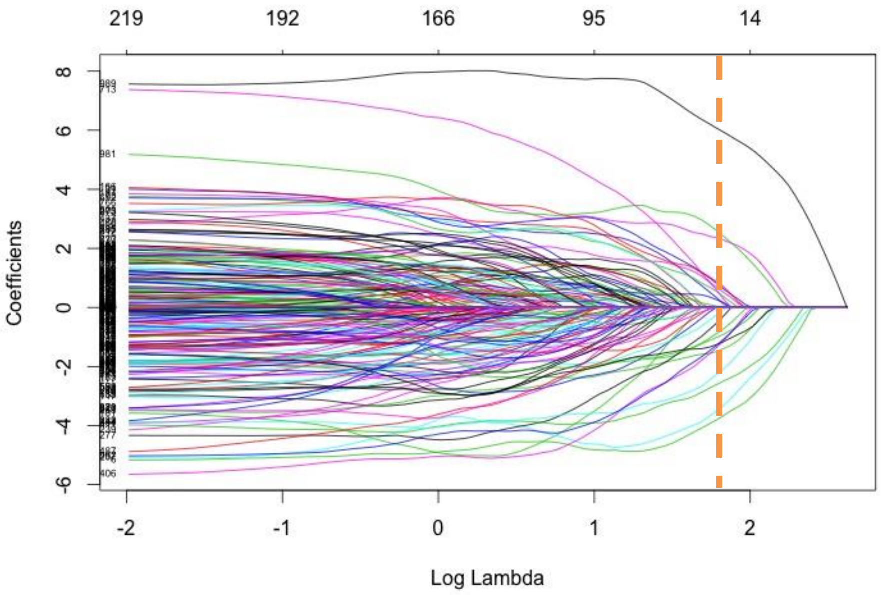
Generalization & Model Validation

Concerns regarding: **generalizability**, **interpretability**, and risk of **overfitting** relationships

Regularization/Shrinkage

- Lasso penalization on Linear Regression
- $\hat{\beta}^{lasso} = argmin_{\beta}\left\{\frac{1}{2}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|\right\}$

Lasso Regularization Path



Generalization & Model Validation

Concerns regarding: **generalizability**, **interpretability**, and risk of **overfitting** relationships

Model Averaging:

- Bagging (Bootstrap Aggregation)

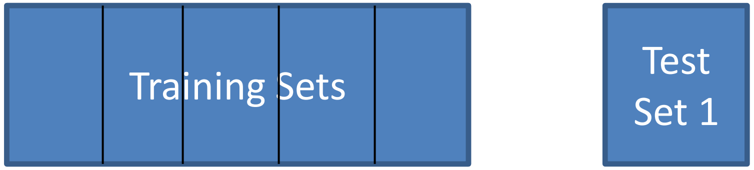
$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{*b}(x)$$

- Bayesian Model Averaging

$$\hat{\theta}_{\mathrm{BMA}} = \sum_{k=1}^{K} \hat{\theta}_k p\left(M_k \,|\, \mathbf{Z}\right)$$

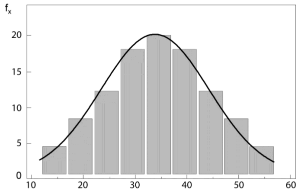
Generalization & Model Validation

1: Partition (k-fold CV)

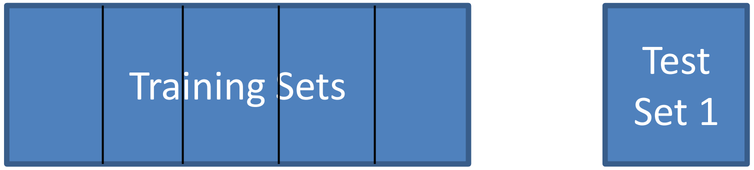


Generalization & Model Validation

2: Resample



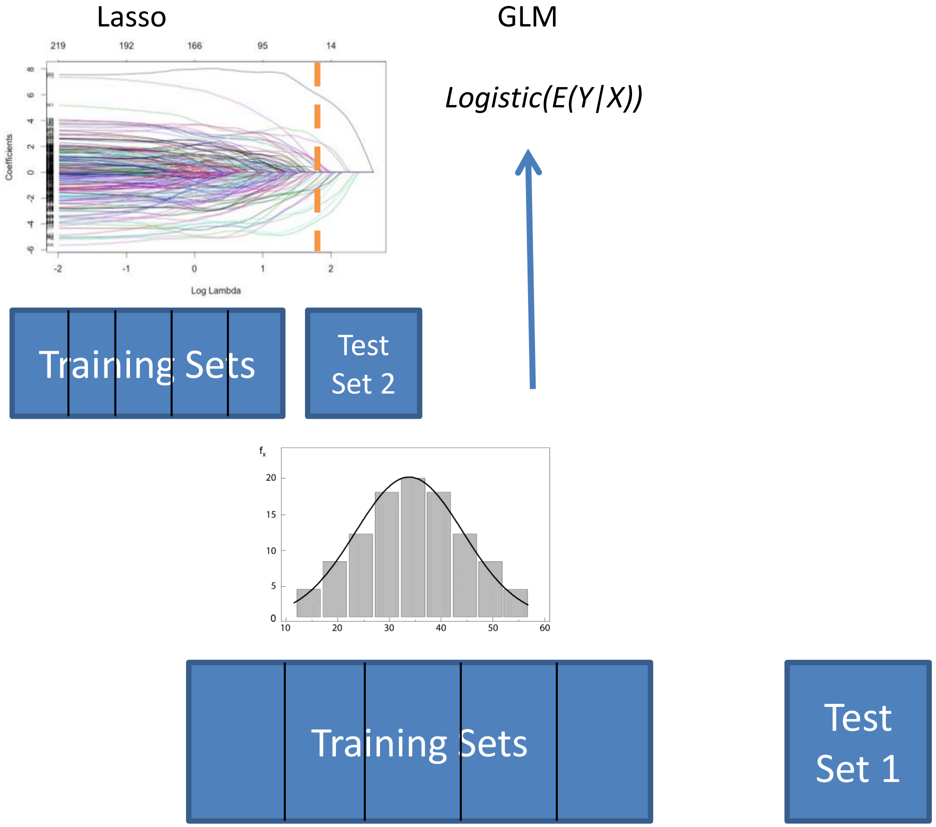
1: Partition (k-fold CV)



Training Sets

Test Set 1

Generalization & Model Validation



GLM

Logistic(E(Y|X))

Training Sets

Test Set 2

2: Resample

1: Partition (k-fold CV)

Training Sets

Test Set 1

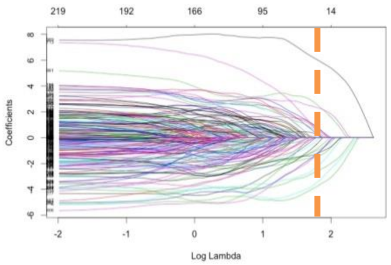
Generalization & Model Validation

4: Model averaging

$F_M(X): f(x_1),f(x_2),\ldots$

3: Fit and Evaluate Models

Lasso



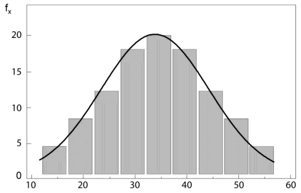
GLM

Logistic(E(Y|X))

Training Sets

Test Set 2

2: Resample



1: Partition (k-fold CV)

Training Sets

Test Set 1

Generalization & Model Validation

5: Evaluate prediction with resampled test set

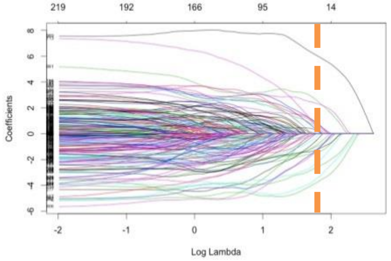
$Y = F_M(X)$ → $\{y,x\}$

4: Model averaging

$F_M(X)$: $f(x_1),f(x_2),\ldots$

3: Fit and Evaluate Models

Lasso

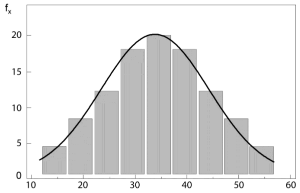


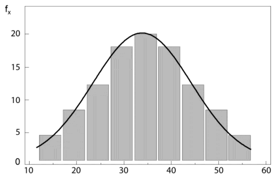
GLM

Logistic(E(Y|X))

Training Sets | Test Set 2

2: Resample





1: Partition (k-fold CV)

Training Sets | Test Set 1

Prediction

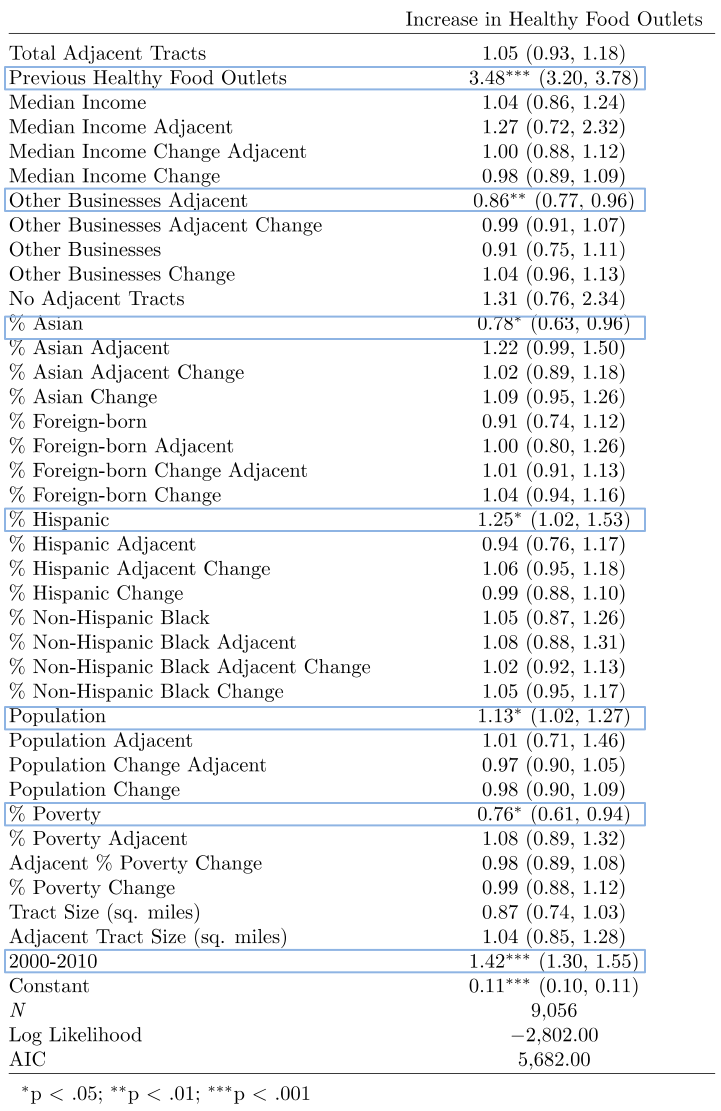
Results need correction

Misclassification

	GLM			Lasso (CV-min)			Lasso (1 Standard Error of CV-min)		
	Sample Fit	Bag	BMA	Sample Fit	Bag	BMA	Sample Fit	Bag	BMA
Training Set CV error	13.30%	13.29%	12.45%	13.25%	13.21%	12.39%	13.68%	13.63%	12.97%
Bootstrap Test Set Error	13.28%	13.66%	13.27%	13.28%	13.66%	13.23%	13.33%	13.40%	13.37%
Full Sample	13.22%	13.68%	13.21%	13.25%	13.68%	13.18%	13.30%	13.47%	13.29%

GLM Associations

Logistic Regression Results: OR(95% CI)

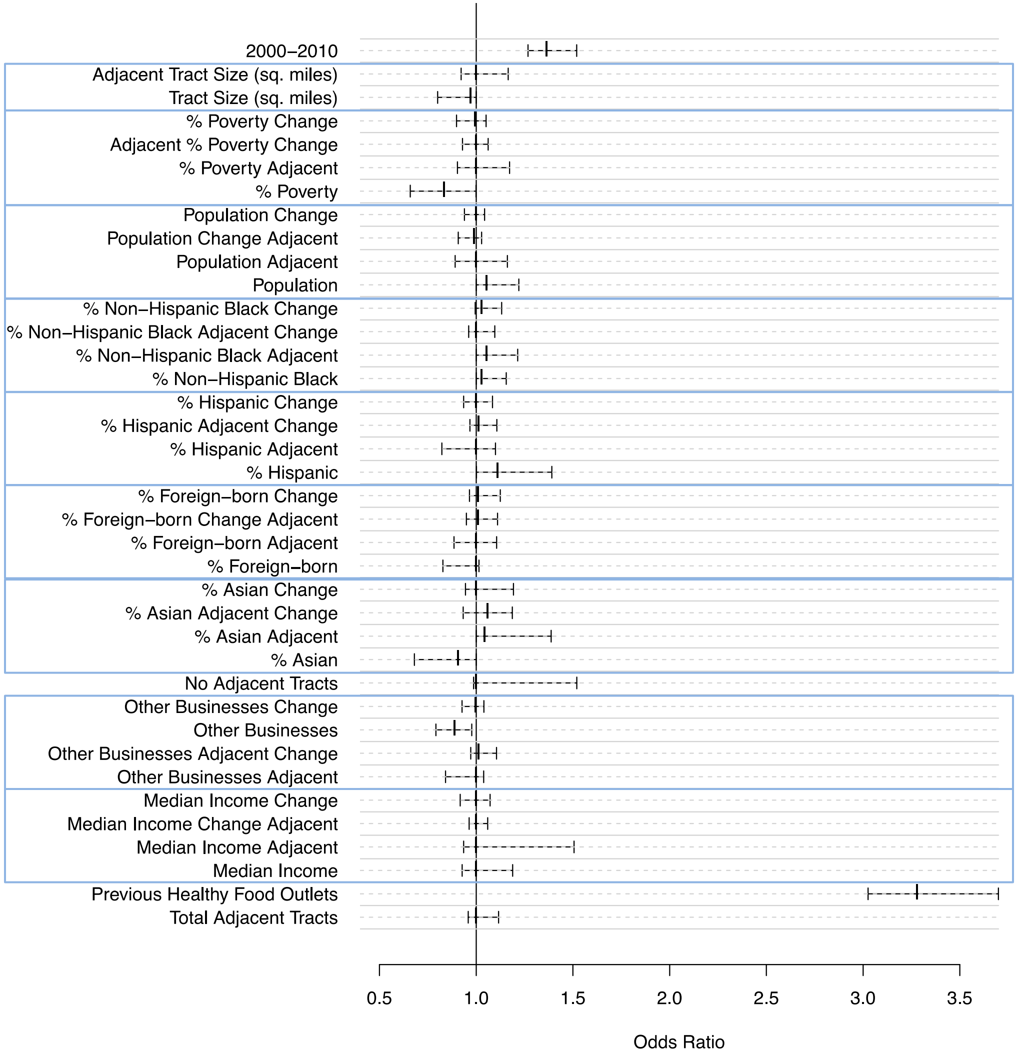


	Increase in Healthy Food Outlets
Total Adjacent Tracts	1.05 (0.93, 1.18)
Previous Healthy Food Outlets	3.48*** (3.20, 3.78)
Median Income	1.04 (0.86, 1.24)
Median Income Adjacent	1.27 (0.72, 2.32)
Median Income Change Adjacent	1.00 (0.88, 1.12)
Median Income Change	0.98 (0.89, 1.09)
Other Businesses Adjacent	0.86** (0.77, 0.96)
Other Businesses Adjacent Change	0.99 (0.91, 1.07)
Other Businesses	0.91 (0.75, 1.11)
Other Businesses Change	1.04 (0.96, 1.13)
No Adjacent Tracts	1.31 (0.76, 2.34)
% Asian	0.78* (0.63, 0.96)
% Asian Adjacent	1.22 (0.99, 1.50)
% Asian Adjacent Change	1.02 (0.89, 1.18)
% Asian Change	1.09 (0.95, 1.26)
% Foreign-born	0.91 (0.74, 1.12)
% Foreign-born Adjacent	1.00 (0.80, 1.26)
% Foreign-born Change Adjacent	1.01 (0.91, 1.13)
% Foreign-born Change	1.04 (0.94, 1.16)
% Hispanic	1.25* (1.02, 1.53)
% Hispanic Adjacent	0.94 (0.76, 1.17)
% Hispanic Adjacent Change	1.06 (0.95, 1.18)
% Hispanic Change	0.99 (0.88, 1.10)
% Non-Hispanic Black	1.05 (0.87, 1.26)
% Non-Hispanic Black Adjacent	1.08 (0.88, 1.31)
% Non-Hispanic Black Adjacent Change	1.02 (0.92, 1.13)
% Non-Hispanic Black Change	1.05 (0.95, 1.17)
Population	1.13* (1.02, 1.27)
Population Adjacent	1.01 (0.71, 1.46)
Population Change Adjacent	0.97 (0.90, 1.05)
Population Change	0.98 (0.90, 1.09)
% Poverty	0.76* (0.61, 0.94)
% Poverty Adjacent	1.08 (0.89, 1.32)
Adjacent % Poverty Change	0.98 (0.89, 1.08)
% Poverty Change	0.99 (0.88, 1.12)
Tract Size (sq. miles)	0.87 (0.74, 1.03)
Adjacent Tract Size (sq. miles)	1.04 (0.85, 1.28)
2000-2010	1.42*** (1.30, 1.55)
Constant	0.11*** (0.10, 0.11)
N	9,056
Log Likelihood	−2,802.00
AIC	5,682.00

*p < .05; **p < .01; ***p < .001

Add: Models with adjacent and time change as separate; correct results

Lasso Associations



same framework – correct the results

Conclusions

- Model validation is an important aspect of assessing generalizability.
- Selection of models fit to bootstrap samples using likelihood-based criteria and regularization may improve out-of-sample prediction.

Future Directions

- Time intervals
- Generalizability across localities
- Alternate Outcomes