Model-based matching for causal inference in observational studies

Kellie Ottoboni
with Philip B. Stark

Department of Statistics, UC Berkeley
Berkeley Institute for Data Science

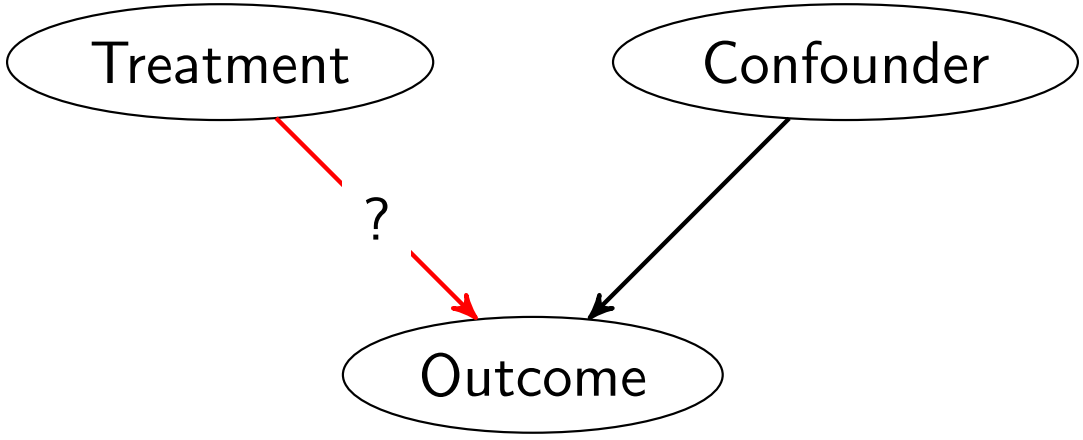
March 15, 2016

Outline

Observational Studies vs Experiments

TO DO: Check that bibliography entries look okay

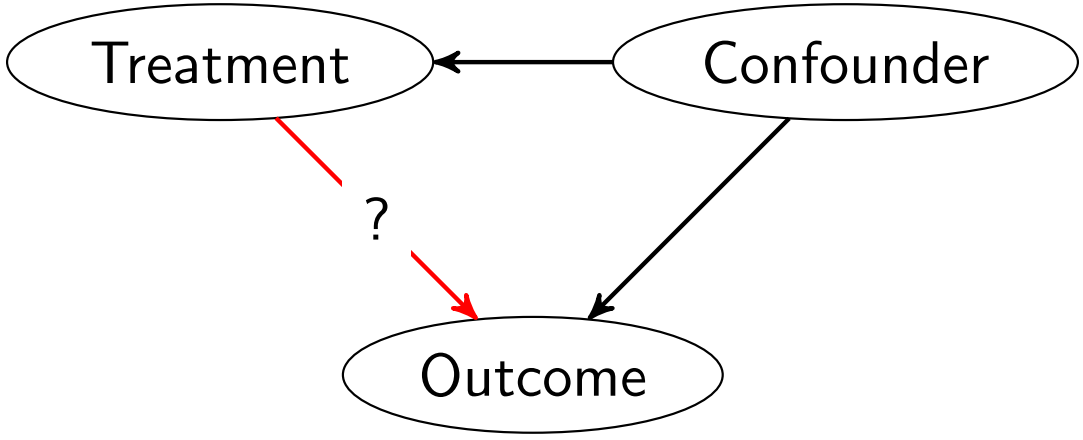
- **Problem:** Estimate the causal effect of a treatment on outcome of interest
- In randomized experiments, treatment is assigned to individuals at random.
- In observational studies, the way individuals select into treatment groups is unknown.



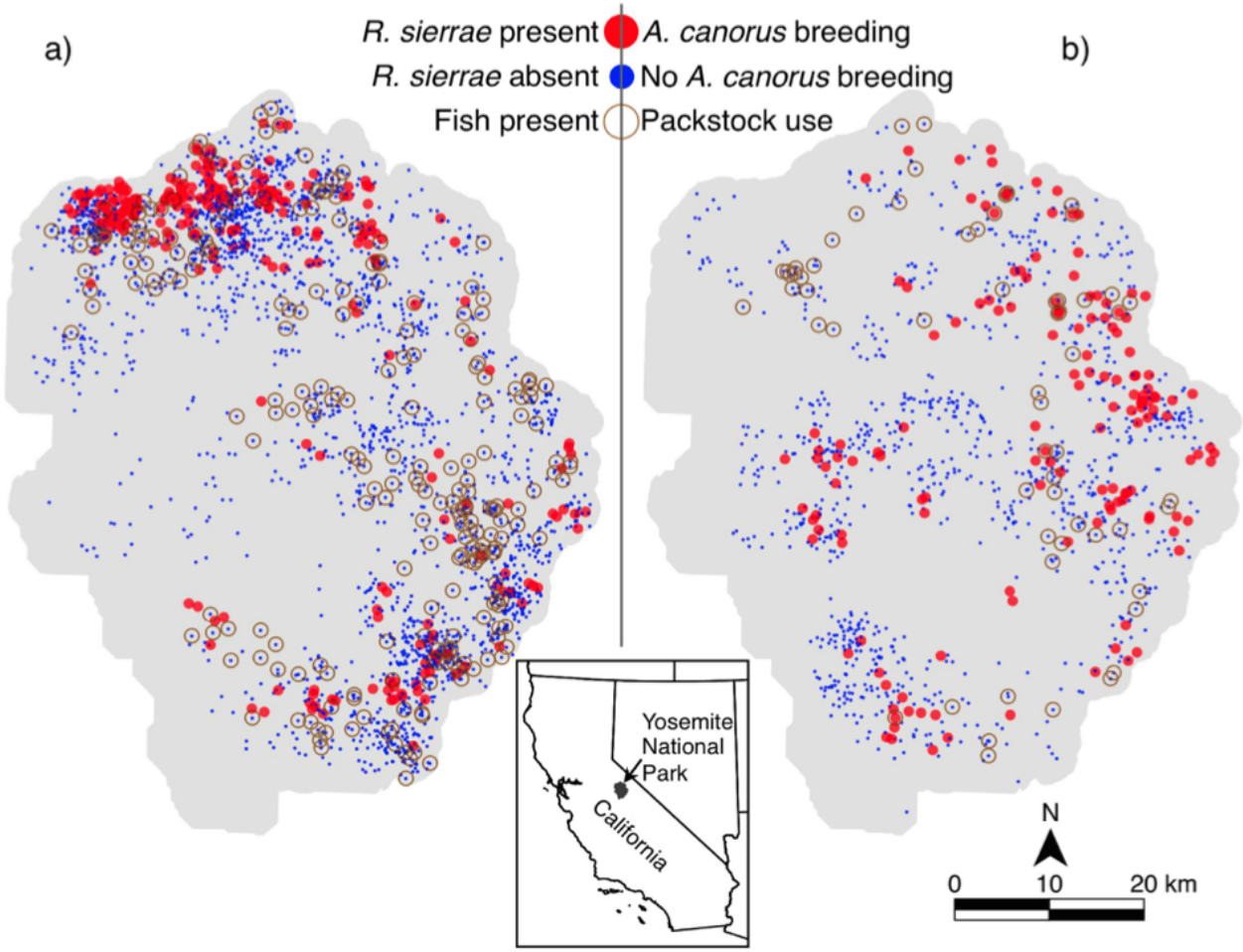
Observational Studies vs Experiments

TO DO: Check that bibliography entries look okay

- **Problem:** Estimate the causal effect of a treatment on outcome of interest
- In randomized experiments, treatment is assigned to individuals at random.
- In observational studies, the way individuals select into treatment groups is unknown.



Motivating Example: Toads and Packstock



J. R. Matchett, Philip B. Stark, Steven M. Ostoja, Roland A. Knapp, Heather C. McKenny, Matthew L. Brooks, William T. Langford, Lucas N. Joppa, and Eric L. Berlow. Detecting the influence of rare stressors on rare species in Yosemite National Park using a novel stratified permutation test. Scientific Reports, 5: 10702, June 2015.

Motivating Example: Toads and Packstock

- The response is rare (few meadows have toads).
- The treatment is rare (few meadows are used by packstock).
- Randomized experiment is impossible, and toad/packstock presence is not random across meadows.
- We're interested in detecting any effect, no matter how small. If treatment effect varies across meadows, then averages might not be informative.

Neyman-Rubin Causal Model

- Population of $i = 1, \ldots, N$ individuals. Each individual has two **potential outcomes**.
- $Y_i(1)$ is individual i's outcome if he receives treatment
- $Y_i(0)$ is individual i's outcome if he is in the control group
- The treatment effect for individual i is $\tau_i = Y_i(1) - Y_i(0)$

$Y_1(1)$	$Y_1(0)$

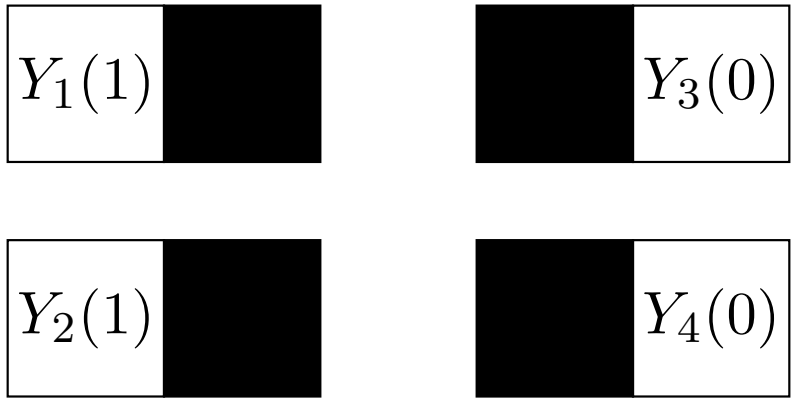
$Y_3(1)$	$Y_3(0)$

$Y_2(1)$	$Y_2(0)$

$Y_4(1)$	$Y_4(0)$

Fundamental Problem of Causal Inference [Holland, 1986]

- We may never observe both $Y_i(1)$ and $Y_i(0)$
- T_i is a treatment indicator: 1 if i is treated, 0 if i is control
- The observed outcome for individual i is $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$



TO DO: FIX BOX POSITION

Estimands

- Average treatment effect
$$\mathbb{E}(Y_i(1) - Y_i(0))$$
- Average treatment effect on the treated
$$\mathbb{E}(Y_i(1) - Y_i(0) \mid T_i = 1)$$
- Conditional average treatment effect
$$\mathbb{E}(Y_i(1) - Y_i(0) \mid X_i)$$
- If treatment effect varies by covariates X, then averages might not be informative

Goal

Goal: test the **strong null hypothesis** of no treatment effect whatsoever.

$$H_0 : Y_i(1) = Y_i(0) \text{ for all } i$$
$$H_1 : Y_i(1) \neq Y_i(0) \text{ for some } i$$

We'd like our test to have power to detect

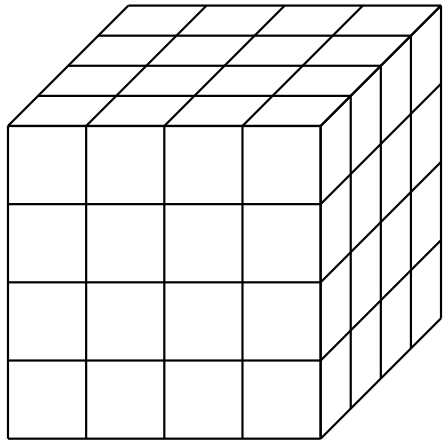
- non-constant effects
- non-linear effects
- effects with non-constant sign

Outline

Matching

Individuals with similar covariates should have similar outcomes, but for the treatment.

- **Ideal:** group individuals by X_i to estimate subgroup treatment effects and then average over subgroups
- **Reality:** many covariates, perhaps continuous, make it difficult to stratify

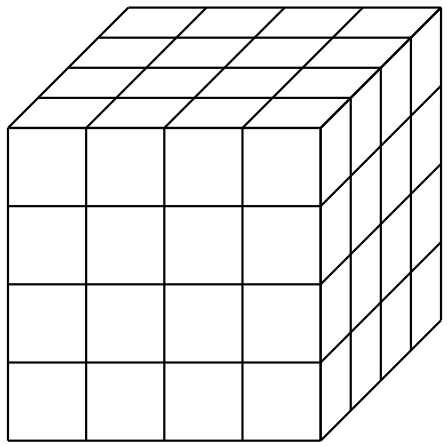


Aside... the curse of dimensionality

- If d covariates are split into k bins, we have d^k groups.
- To guarantee that we have at least one treated and one control in each group with 95% probability, we need

$$n \geq \frac{2\log(1-(0.95)^{1/k^{d+1}})}{\log(\frac{k^d-1}{k^d})}$$

- If $d = 5$ and $k = 2$, $n \geq 225$.
- If $d = 10$ and $k = 2$, $n \geq 10,844$.



Matching

- **Solution:** use a one-dimensional score to match or group individuals

Propensity score matching

- The **propensity score** is an individual's probability of being assigned treatment, conditional on their covariates

$$p(x) = \mathbb{P}(T = 1 \mid X = x)$$

- The propensity score is a balancing score: $X \perp\!\!\!\perp T \mid p(X)$
- For individuals with the same propensity score, treatment assignment is as if random

Propensity score matching

Theorem (Rosenbaum and Rubin [1983])

If treatment assignment is independent of potential outcomes given X*,*

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid X$$

and if every unit has a chance of receiving treatment,

$$0 < p(X) < 1 \text{ for all } X$$

then $(Y(1), Y(0)) \perp\!\!\!\perp T \mid p(X)$*.*

In particular, treated units can serve as the counterfactual for controls with the same $p(X)$

$$\mathbb{E}(Y(t) \mid T = 1, p(X)) = \mathbb{E}(Y(t) \mid T = 0, p(X)) \text{ for } t = 0, 1$$

Propensity score matching

This result identifies the average treatment effect in terms of quantities we can estimate:

$$\begin{aligned}
\mathbb{E}(Y(1) - Y(0)) &= \mathbb{E}_{p(x)}\left[\mathbb{E}(Y(1) - Y(0) \mid p(x))\right] \\
&= \mathbb{E}_{p(x)}\left[\mathbb{E}(Y(1) \mid p(x)) - \mathbb{E}(Y(0) \mid p(x))\right] \\
&= \mathbb{E}_{p(x)}\left[\mathbb{E}(Y \mid p(x), T = 1) - \mathbb{E}(Y \mid p(x), T = 0)\right]
\end{aligned}$$

Propensity score matching

$p(x)$ is usually unknown and estimated by $\hat{p}(x)$ using logistic or probit regressions

- Assumes a simple functional form for relationship between covariates and treatment
- Assumes that probability of treatment takes same form for all individuals
- May actually worsen balance if estimated incorrectly [Diamond and Sekhon, 2012]

Matching complicates inference

- Standard errors are difficult to compute for matching estimators [Abadie and Imbens, 2006, 2008]
- Rarely used in hypothesis testing procedures
- There's no "optimal" way to match [Austin, 2014]

Model-based Matching

Idea: Instead of modeling the propensity score, model the outcome

Computing $\hat{Y}$, the "best" prediction of the outcome based on all covariates except for the treatment, buys us two things:

- $\hat{Y}$ is a score on which to stratify observations
- Using residuals $Y - \hat{Y}$ improves precision by removing variation due to X [Rosenbaum, 2002]

Model-based Matching

Suppose that outcomes have the form

$$Y_i(t) = f(t, X_i) + \varepsilon_i$$

for $i = 1, \ldots, N$ and $t = 0, 1$. Let X_i be fixed and suppose that the ε_i are IID with $\mathbb{E}(\varepsilon_i) = 0$.

Under the strong null hypothesis, $f(0, X_i) = f(1, X_i)$ for each i.

Thus, our best guess of Y_i needn't involve the treatment:

$$\hat{Y}_i = \hat{f}(X_i)$$

Model-based Matching

Stratify or match units on their $\hat{Y}_i = \hat{f}(X_i)$.

- Let $S_i = j$ if unit i is in stratum j, where $j \in \{1, \ldots, J\}$. (For now, don't worry about how to select J strata.)
- **Under the null**, we expect units in the same strata to have the similar responses.
- **Under the alternative**, the treatment adds additional information about the responses beyond $\hat{f}$.

  The residuals will capture some of the effect of treatment:

  $$Y_i - \hat{Y}_i \not\perp\!\!\!\perp T_i$$

Permutation tests

We will use the average difference in means across strata as our test statistic:

$$\tau(Y,T) = \frac{N_j}{N}\sum_{j=1}^{J}\left|\frac{n_j}{N_j}\sum_{\substack{i:S_i=j\\T_i=1}}\left(Y_i-\hat{Y}_i\right) - \frac{N_j-n_j}{N_j}\sum_{\substack{i:S_i=j\\T_i=0}}\left(Y_i-\hat{Y}_i\right)\right|$$

NB: we can use any other test statistic that measures association between $Y_i - \hat{Y}_i$ and T_i, e.g. correlation

Permutation tests

Basic idea: If, under the null hypothesis, the probability distribution of the data is invariant under permutation of treatment assignments, then once we observe the actual data, we know other possible data sets that are equally likely.

Suppose that the jth stratum contains N_j units, n_j of which are treated. Then there are

$$\prod_{j=1}^{J} \binom{N_j}{n_j}$$

equally likely assignments to treatment.

Permutation tests

We approximate the null distribution using this invariance principle.

- Within strata, permute treatment assignments to obtain new treatment vector T_1^*.
- Compute the test statistic $\tau(Y, T_1^*)$.
- Repeat a large number B times to get a distribution $\tau(Y, T_1^*), \ldots, \tau(Y, T_B^*)$.
- The p-value of the test is

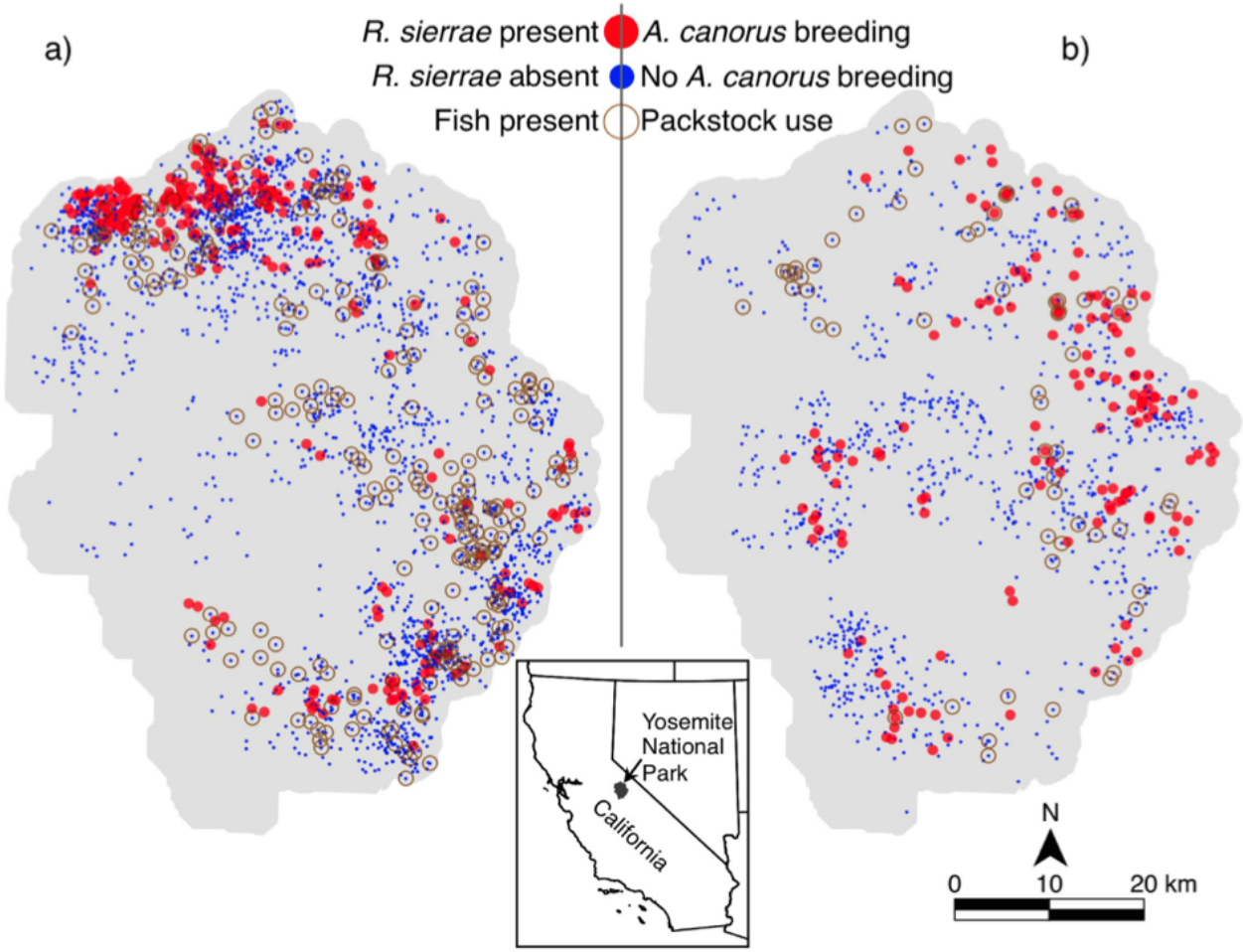
$$p = \mathbb{P}(\tau(Y, T) \geq \tau(Y, t)) \approx \frac{\sum_{i=1}^{B} \mathbb{I}(\tau(Y, T_b^*) \geq \tau(Y, T))}{B}$$

Permutation tests

- Randomization inference doesn't require distributional assumptions about ε_i or Y_i.
- Stratification allows effects to vary (in magnitude, in sign, etc.) across strata.

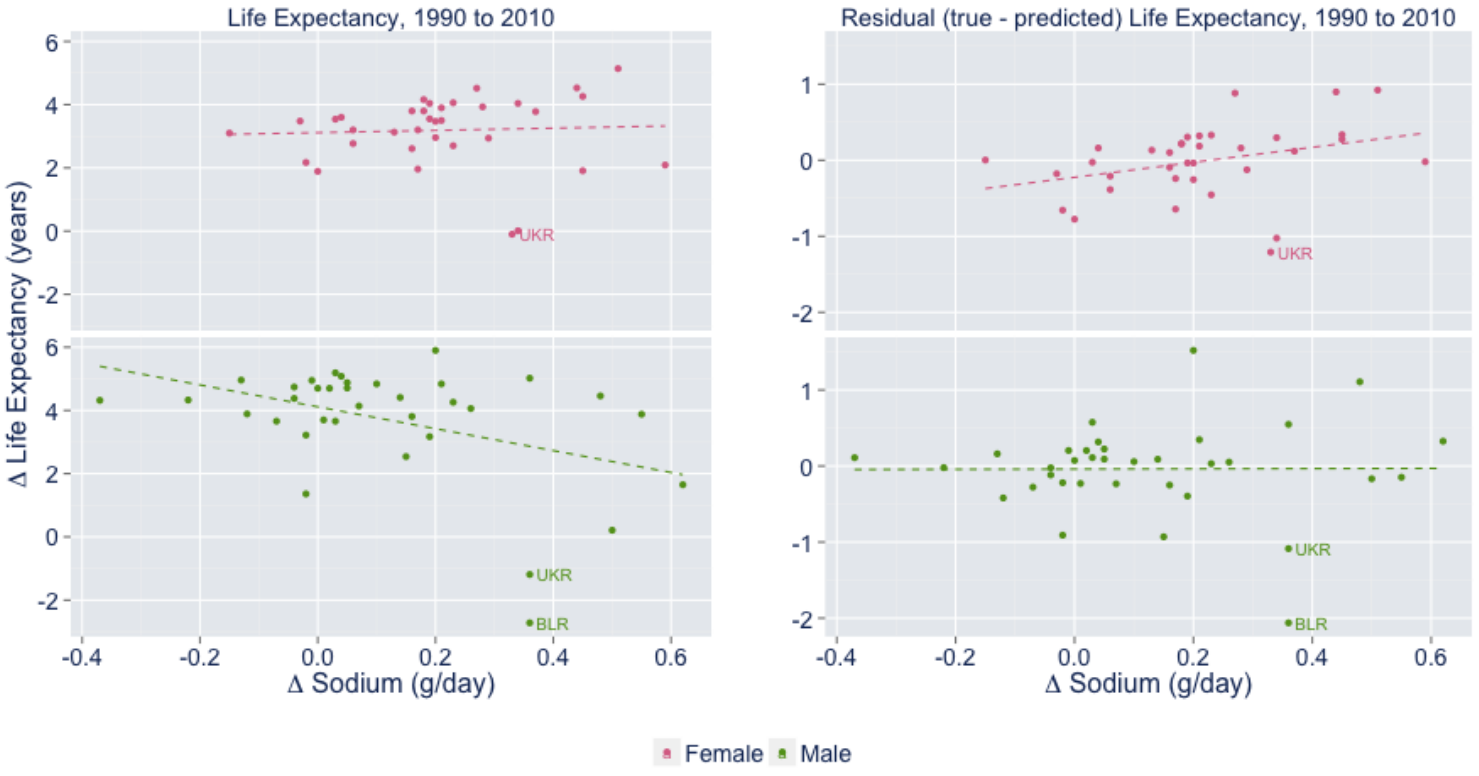
Outline

Toads and Packstock



J. R. Matchett, Philip B. Stark, Steven M. Ostoja, Roland A. Knapp, Heather C. McKenny, Matthew L. Brooks, William T. Langford, Lucas N. Joppa, and Eric L. Berlow. Detecting the influence of rare stressors on rare species in Yosemite National Park using a novel stratified permutation test. Scientific Reports, 5: 10702, June 2015.

Salt


Life Expectancy, 1990 to 2010
Residual (true - predicted) Life Expectancy, 1990 to 2010
Δ Life Expectancy (years)
Δ Sodium (g/day)
Δ Sodium (g/day)
UKR
UKR
BLR
UKR
UKR
BLR
Female
Male

Outline

Future Directions

- Do different test statistics give greater power when the treatment effect is nonlinear?
- What is the optimal way to stratify?
- How to quantify uncertainty – standard errors and confidence intervals?

References

Alberto Abadie and Guido W. Imbens. Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica*, 74(1):235–267, January 2006. ISSN 1468-0262. doi: 10.1111/j.1468-0262.2006.00655.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0262.2006.00655.x/abstract.

Alberto Abadie and Guido W. Imbens. On the Failure of the Bootstrap for Matching Estimators. *Econometrica*, 76 (6):1537–1557, November 2008. ISSN 1468-0262. doi: 10.3982/ECTA6474. URL http://onlinelibrary.wiley.com/doi/10.3982/ECTA6474/abstract.

Peter C. Austin. A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*, 33(6): 1057–1069, March 2014. ISSN 1097-0258. doi: 10.1002/sim.6004. URL http://onlinelibrary.wiley.com/doi/10.1002/sim.6004/abstract.

Alexis Diamond and Jasjeet S. Sekhon. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economics and Statistics*, 95(3): 932–945, October 2012. ISSN 0034-6535. doi: 10.1162/REST_a_00318. URL http://dx.doi.org/10.1162/REST_a_00318.

Paul W. Holland. Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396): 945–960, 1986. ISSN 0162-1459. doi: 10.2307/2289064. URL http://www.jstor.org/stable/2289064.

J. R. Matchett, Philip B. Stark, Steven M. Ostoja, Roland A. Knapp, Heather C. McKenny, Matthew L. Brooks, William T. Langford, Lucas N. Joppa, and Eric L. Berlow. Detecting the influence of rare stressors on rare species in Yosemite National Park using a novel stratified permutation test. *Scientific Reports*, 5:10702, June 2015. ISSN 2045-2322. doi: 10.1038/srep10702. URL http://www.nature.com/articles/srep10702.

Paul R. Rosenbaum. Covariance Adjustment in Randomized Experiments and Observational Studies. *Statistical Science*, 17(3):286–327, August 2002. ISSN 0883-4237, 2168-8745. doi: 10.1214/ss/1042727942. URL http://projecteuclid.org/euclid.ss/1042727942.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, April 1983. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/70.1.41. URL http://biomet.oxfordjournals.org/content/70/1/41.