Government 2000 Lecture Notes[*]

Jasjeet S. Sekhon

Center for Basic Research in the Social Sciences
Department of Government
Harvard University

http://jsekhon.fas.harvard.edu/
jasjeet_sekhon@harvard.edu

Introduction

We are interested in estimation techniques which make the fewest and weakest assumptions realistically possible.

OLS and the usual Maximum Likelihood and Bayesian methods are surprisingly brittle. Small deviations from their assumptions are often catastrophic.

We mainly focus on two issue:

- causal inference
- robust estimation

Mill's Methods of Inductive Inference

John Stuart Mill (in his *A System of Logic*) devised a set of five methods (or canons) by means of which to analyze and interpret our observations for the purpose of drawing conclusions about the causal relationships they exhibit. These methods have been used by generations of social science researchers.

Method of Agreement: “If two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree is the cause (or effect) of the given phenomenon.”

Method of Difference: “If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former; the circumstance in which alone the two instances differ is the effect, or the cause, or an indispensable part of the cause, of the phenomenon.”

These methods have been used by a vast number of researchers, including such famous ones as Durkheim and Weber. All such work as serious problems.

Here are some examples:

- The Protestant Ethic
- Deficits and interest rates
- Gun control
- The list goes on, and on....

Mill himself realized many of these problems:

> “Nothing can be more ludicrous than the sort of parodies on experimental reasoning which one is accustomed to meet with, not in popular discussion only, but in grave treatises, when the affairs of nations are the theme. “How,” it is asked, “can an institution be bad, when the country has prospered under it?” “How can such or such causes have contributed to the prosperity of one country, when another has prospered without them?” Whoever makes use of an argument of this kind, not intending to deceive, should be sent back to learn the elements of some one of the more easy physical sciences.”

The Solution: Statistical Inference

Let's look at an example Mill himself brought up:

> "In England, westerly winds blow during about twice as great a portion of the year as easterly. If, therefore, it rains only twice as often with a westerly as with an easterly wind, we have no reason to infer that any law of nature is concerned in the coincidence. If it rains more than twice as often, we may be sure that some law is concerned; either there is some cause in nature which, in this climate, tends to produce both rain and a westerly wind, or a westerly wind has itself some tendency to produce rain."

Conditional Probability

$$\begin{aligned} \text{H}: &\text{P}(\text{rain}|\text{westerly wind},\ \Omega) > \\ &\text{P}(\text{rain}|\textbf{not}\ \text{westerly wind},\ \Omega), \end{aligned}$$

where Ω is a set of background conditions we consider necessary for a valid comparison.

A lot is hidden in the Ω. This is where matching comes in.

Robust Estimation: What Are Some Good Properties Estimators Should Have?

You already know of some good properties:

- unbiasedness
- consistency
- Small RMSE

Following Huber 1981 (5–17), here are some more desirable properties estimators should have:

- have reasonably good efficiency when the model assumed for the data is correct;
- small deviations from the model assumptions (which may mean large deviations in a small fraction of the data) impair the model's performance only slightly;
- somewhat larger deviations from the model should not cause a catastrophe.

That’s Nice, But What Does that Mean?

More formal descriptions of desirable properties will be offered next class. But first, here are a two examples:

- Problem: heteroscedasticity

  Answer: Huber-White heteroscedasticity consistent standard errors, or Bootstrap standard errors. Note how different Huber-White standard errors are from Beck-Katz standard errors. Ignore this if you haven’t heard of Beck-Katz.

- Problem: a reasonable amount of measurement error in some observations in one RHS variable which renders all coefficients inconsistent.

  Answer: least medium of squares (LMS) or a number of better alternatives.

What did you say about measurement error?

Classical Assumptions

- A1. $Y_t = X_t'\alpha + \epsilon_t, \;\; t = 1, 2, 3, \cdots, n,$
  where Y_t, ϵ_t are scalar and X_t, α are $k \times 1$ vectors.

- A2. X_t is nonstochastic, $t = 1, 2, 3, \cdots, n$. X_t is a fixed constant, there is no disturbance associated with X_t. X_t may always be moved outside the expectation.

- A3. The $k \times k$ matrix $X_t X_t'$ is non-singular for every $n \geq k$.

- A4. $E[\epsilon_t] = 0$, $t = 1, 2, 3, \cdots, n$. Since X_t is assumed to be nonstochastic (A2), (A4) implies that $E[X_t \epsilon_t] = 0$. (A4) always holds if there is an intercept.

- A5. $E\left[\epsilon\epsilon'\right] = \sigma^2 I$, where ϵ is an $n \times 1$ matrix and I is an $n \times n$ identity matrix.
  - ⋆ It follows that $E[\epsilon_t^2] = \sigma^2$, $t = 1, 2, 3, \cdots, n$. But the $4^t h$ moment may vary with t. (A5) is, therefore, weaker than the iid assumption.
  - ⋆ $cov(\epsilon_t \epsilon_\tau) = 0$ for all $t \neq \tau$. iid implies these two items, but they do not imply iid.
  - ⋆ (A5) is the homoscedasticity assumption. It is similar to $E(\epsilon_t^2 | X_t) = \sigma^2$.

Correct Specification Assumption

The correct specification assumption implies that $E(\epsilon_t|X_t) = 0$. Why?

Because we are modeling the conditional mean.

$$
\begin{aligned}
Y_t &= E(Y_t|X_t) + \epsilon_t \\
\text{Then} & \\
\epsilon_t &= Y_t - E(Y_t|X_t) \\
\text{and} & \\
E(\epsilon_t|X_t) &= E\left[(Y_t - E(Y_t|X_t)|X_t\right] \\
&= E(Y_t|X_t) - E\left[E(Y_t|X_t)|X_t\right] \\
&= E(Y_t|X_t) - E(Y_t|X_t) \\
&= 0
\end{aligned}
$$

Remarks:

- The regression function $E(Y_t|X_t)$ is used to predict Y_t from knowledge of X_t.

- The term ϵ_t is called the “regression disturbance.” The fact $E(\epsilon_t|X_t) = 0$ implies that ϵ_t contains no systematic information of X_t in predicted Y_t. In other words, all information of X_t that is useful to predict Y_t has been summarized by $E(Y_t|X_t)$.

- The assumption that $E(\epsilon|X) = 0$ is crucial. If $E(\epsilon|X) \neq 0$, $\hat{\beta}$ is biased.

- Situations in which $E(\epsilon|X) \neq 0$ can arise easily. For example, X_t may contain errors of measurement.

Measurement Error

Suppose the model is

$$Y_t = W_t \beta + V_t, \quad E(W_t' V_t) = 0, \tag{1}$$

where (to make life interesting) W_t is a $1 \times k$ vector and β is $k \times 1$, Y_t, V_t are scalars.

We measure W_t subject to errors η_t as $X_t = W_t + \eta_t$,

We grant some very generous assumptions:

- the measurement error has no bias
- the underlying variable, W_t, is uncorrelated with the disturbance
- the measurement error is uncorrelated with our underlying variable
- the measurement error is uncorrelated with our disturbance

$$E(W_t' \eta_t) = 0,$$
$$E(\eta_t' \eta_t) \neq 0,$$
$$E(\eta_t' V_t) = 0.$$

Then

$$\begin{aligned} Y_t &= X_t\beta + V_t - \eta_t\beta \qquad (2) \\ &= X_t\beta + \epsilon_t. \qquad (3) \end{aligned}$$

where $\epsilon_t = V_t - \eta_t\beta$.

We have $E(X_t'\epsilon_t) = E[(W_t' + \eta_t')(V_t - \eta_t\beta)] = E(\eta_t'\eta_t)\beta \neq 0$.

Now $E(\epsilon|X) = 0$ implies that for all t, $E(X_t'\epsilon_t) = 0$, since $E(X_t'\epsilon_t) = E[E(X_t'\epsilon_t|X)] = E[X_t'E(\epsilon_t|X)] = 0$.

Hence, $E(X_t'\epsilon_t) \neq 0$ implies $E(\epsilon|X) \neq 0$. The OLS estimator will not be, in general, unbiased in the presence of measurement error.

Achen (1986) has a great paper on measurement error (and the proxy variable problem) with lots of examples and detail.

But People Believe OLS is Better Than This

Some, particularly comparative politics scholars, have often stated that measurement error only biases the variable with the error to zero.

What's some intuition for why this is wrong?

Nonorthogonal X variables!

Don't you wish you had complete experiments!

But even most experiments in science (e.g., medicine) are only quasi-experiments.

Some Examples:

Using lagged differences:
http://jsekhon.fas.harvard.edu/gov2000/R/difference1.R.

The following two examples are based on the following two articles which are available via JSTOR:

- Nathaniel Beck, Jonathan Katz, Michael Alvarez, Geoffrey Garrett and Peter Lange. "Government Partisanship, Labor Organization and Macroeconomic Performance: A Corrigendum." American Political Science Review 87(4) December 1993: 945-948.

- Michael Alvarez, Geoffrey Garrett and Peter Lange. "Government Partisanship, Labor Organization and Macroeconomic Performance, 1967-1984." American Political Science Review 85(3) June 1991: 539-556.

Replication of Table 1 for Beck et al. (1993) column 1:
http://jsekhon.fas.harvard.edu/gov2000/R/agl1.R,
http://jsekhon.fas.harvard.edu/gov2000/R/agl1.Rout.

Measurement error simulations based on Beck et al. (1993):
http://jsekhon.fas.harvard.edu/gov2000/R/MeasurementError1.R,
http://jsekhon.fas.harvard.edu/gov2000/R/MeasurementError1.Rout.

Orthogonal vs. Independent

Recall that X is called statistically independent of T if and only if $P(X|T) = P(X)$. If X and T are normally distributed and then if X and T are uncorrelated, they are also independent and vice-a-versa.

The word Orthogonal is more ambagious because it has various definitions:

- A square matrix, A, is an orthogonal matrix if $AA' = I$ so $A^{-1} = A'$
- In elementary geometry, orthogonal is the same as perpendicular. Two lines or curves are orthogonal if they are perpendicular at their point of intersection.
- two elements x and y of an inner product space are called orthogonal if the inner product of x and y is 0.
- Orthogonal can also mean statistical independence.

What Do I Need for Correct Specification?

The correct specification assumption, $E(\epsilon|X) = 0$ is only directly a statement about the first two moments. But you really need more unless you are on the natural scale of the variables.

Strictly, in order to stop uncorrelated measurement error in X_1 influencing the estimate of X_2 in OLS, one only needs that X_1 and X_2 be uncorrelated.

Post-Treatment Bias

The lesson of the lagged differences example is to not ask too much from regression. See: http://jsekhon.fas.harvard.edu/gov2000/R/difference1.R.

Many lessons follow from this. One of the most important is the issue of post-treatment bias. Post-treatment variables are those which are a consequence of the treatment of interest. Hence, they will be correlated with the treatment unless their values are the same under treatment and control.

Standard Errors and Correct Coverage

The measurement error simulation present results for correct coverage. See: `http://jsekhon.fas.harvard.edu/gov2000/R/MeasurementError1.Rout`.

Confidence intervals are said to be correct if for a given nominal level, say 95%, they contain the true parameter estimate, β, 95% of the time in repeated samples.

Recall that the nominal level for a hypothesis test or confidence interval sets its size. The size of a test is the probability that the test statistic will reject the null hypothesis when it is true—i.e., $P(H_0^R|H_0^T)$. The size of a test is also called its significance level.

We perform tests in the hope that they will reject the null hypothesis when it is false. Accordingly, the power of a test is of great interest. The power of a test statistic T is the probability that T will reject the null hypothesis when the latter is not true—i.e., $P(H_0^R|H_0^F)$.

The power of a consistent test increases with the sample size. As $n \to \infty$, power goes to 1. But the size of a test does not change when n increases.

There are two types of errors:

1. Type I rejecting a null hypothesis when it is true—$\mathrm{P}(\mathrm{H}_0^R | \mathrm{H}_0^T)$.

2. Type II accepting the null when it is false—$\mathrm{P}(\mathrm{H}_0^A | \mathrm{H}_0^F)$

Loss Function

How well the model $f(X_t)$ will explain Y_t is described by a criterion function. In general, there exists a discrepancy between $f(X_t)$ and Y_t. When $f(X_t) \neq Y_t$, a "loss" will occur. This is defined as the loss function.

A loss function $l(Y_t, f(X_t))$ is a real-valued function that describes how well the model $f(X_t)$ can explain Y_t. One form of such a loss function is:

$$l(Y_t, f(X_t)) = (Y_t - f(X_t))^p, \tag{4}$$

where $0 < p < \infty$, is a loss function.

The least square predictor is the loss function where $p = 2$. *This is an arbitrary choice.*

A Variety of Loss Functions

1. least squares (LS): $\min \Sigma_i^N \left(Y_i - \hat{Y}_i\right)^2$

2. least quads: $\min \Sigma_i^N \left(Y_i - \hat{Y}_i\right)^4$

3. least median of squares (LMS): $\min \mathrm{median} \left(Y_i - \hat{Y}_i\right)^2$

4. least absolute deviations (LAD): $\min \Sigma_i^N |Y_i - \hat{Y}_i|$

Breakdown Points of these Loss Functions

1. least squares (LS): $1/N \rightarrow 0$
2. least quads: $1/N \rightarrow 0$
3. least median of squares (LMS): $(N/2)/N \rightarrow .5$
4. least absolute value deviations (LAD): $1/N \rightarrow 0$

Recall that the breakdown point is the minimum proportion of the data which is arbitrarily contaminated so that a parameter can be made to go to infinity. That is, the minimum proportion of the data which has to be contaminated for the parameter to take on an arbitrary value.

Loss Function and MSE

The expected loss is defined as $E[l(Y_t, f(X_t))]$, where E is taken over P. When $l(Y_t, f(X_t)) = (Y_t - f(X_t))^2$, the expected loss is mean square error (MSE).

Theorem:

$$\mathrm{MSE}(f) = E\left[(Y_t - E(Y_t|X_t))^2\right] + E\left[E(Y_t|X_t) - f(X_t)\right]^2 \tag{5}$$

$$= V(\epsilon_t) + E\left[E(Y_t|X_t) - f(X_t)\right]^2, \tag{6}$$

where $\epsilon_t = Y_t - E(Y_t|X_t)$. ▮

There are two kinds of loss: the first is determined by nature, and is unavoidable. The second term comes from the specification error made by the analyst.

Back to Basics

All of the assumptions which are referred to are listed on the assumptions slide.

The existence of the least squares estimator is guaranteed by A1-A3. These assumptions guarantee that $\widehat{\beta}$ exists and that it is unique.

The unbiasedness of OLS is guaranteed by assumptions 1-4. So, $E(\widehat{\beta}) = \beta$.

For hypothesis testing, we need all of the assumptions: A1-A5. These allow us to assume that as $N \rightarrow \infty$, $\widehat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$.

For efficiency we require assumptions A1-A4 plus a modified version of A5. We need to assume that $\epsilon \sim N(0, \sigma^2 I)$, $\sigma^2 < \infty$.

Unbiasedness

Suppose (A1)-(A4) hold. Then $E(\hat{\beta}) = \beta$—i.e., $\hat{\beta}$ is an unbiased estimator of β.

Proof:

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

$$= (\sum_{t=1}^{n} X_tX_t')^{-1} \sum_{t=1}^{n} X_tY_t$$

By (A1)

$$= (\sum_{t=1}^{n} X_tX_t')^{-1} \sum_{t=1}^{n} X_t(X_t'\beta + \epsilon_t)$$

$$= (\sum_{t=1}^{n} X_tX_t')^{-1}(\sum_{t=1}^{n} X_tX_t')\beta + (\sum_{t=1}^{n} X_tX_t')^{-1} \sum_{t=1}^{n} X_t\epsilon_t$$

$$= \beta + (\sum_{t=1}^{n} X_tX_t')^{-1} \sum_{t=1}^{n} X_t\epsilon_t$$

$$\hat{\beta} = \beta + (\sum_{t=1}^{n} X_t X_t')^{-1} \sum_{t=1}^{n} X_t \epsilon_t$$

Given (A2) and (A4)

$$E\left[(\sum_{t=1}^{n} X_t X_t')^{-1} \sum_{t=1}^{n} X_t \epsilon_t\right] = (\sum_{t=1}^{n} X_t X_t')^{-1} \sum_{t=1}^{n} X_t E\left[\epsilon_t\right] = 0$$

$$E(\hat{\beta}) = \beta$$

Unbiasness may be considered a necessary condition for a good estimator, but is certainly isn't sufficient. Moreover, it is also doubtful whether it is even necessary. For example, in general, Maximum Likelihood Estimators are not unbiased—e.g., Logit, Probit, Tobit.

Variance-Covariance Matrix

Suppose Assumptions (A1)-(A5) hold. Then, the variance-covariance matrix of $\widehat{\beta}$ is:

$$E\left[(\widehat{\beta}-\beta)(\widehat{\beta}-\beta)'\right] = \sigma^2(X'X)^{-1} \tag{7}$$

Proof:

$$\widehat{\beta} = \beta + \left(\sum_{t=1}^{n} X_t X_t'\right)^{-1} \sum_{t=1}^{n} X_t \epsilon_t$$

$$\begin{aligned}
\widehat{\beta}-\beta &= (X'X)^{-1}X'\epsilon \\
(\widehat{\beta}-\beta)(\widehat{\beta}-\beta)' &= (X'X)^{-1}X'\epsilon\left[(X'X)^{-1}X'\epsilon\right]' \\
&= (X'X)^{-1}X'\epsilon\left[\epsilon'X(X'X)^{-1}\right] \\
&= (X'X)^{-1}X'(\epsilon\epsilon')X(X'X)^{-1} \\
E(\widehat{\beta}-\beta)(\widehat{\beta}-\beta)' &= (X'X)^{-1}X'E(\epsilon\epsilon')X(X'X)^{-1} \\
&= (X'X)^{-1}X'\sigma^2 I X(X'X)^{-1} \\
\text{move } \sigma^2 \text{ and drop } I &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}
\end{aligned}$$

The sigma-squared Matrix

Note that $\sigma^2 I$ is an $n \times n$ matrix which look like:

$$\begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & \sigma^2 \end{bmatrix}$$

Estimating Sigma-squared

Recall that:

Let $\hat{\epsilon}_i = Y_i - \hat{Y}_i$, then

$$\hat{\sigma}^2 = \frac{\sum \hat{\epsilon}_i^2}{n - k}, \tag{8}$$

where k is the number of parameters.

Heteroscedasticity

Assumption A5 on the assumptions slide makes the homogenous or constant variance assumption. This assumption is clearly wrong in many cases, this is particular true with cross national data like that of Alvarez, Garrett and Lange.

We can either use an estimator which directly models the heteroscedasticity and weights each observations appropriately—e.g, Weighted Least Squares (WLS)—or we can use robust standard errors.

First let's see what goes wrong.

Heteroscedasticity: OLS unbiased?

The OLS estimator is unbiased (and consistent) when there is heteroscedasticity, but it is **not** efficient. We say that $\hat{\alpha}$ is an efficient unbiased estimator if for a given sample size the variance of $\hat{\alpha}$ is smaller than (or equal to) the variance of any other unbiased estimator.

Why is OLS still unbiased? Recall that when we proved that OLS is unbiased, the variance terms played no part in the proof—i.e., Assumption A5 played no part.

Heteroscedasticity: OLS SEs?

It can be shown that the Generalized Least Squares Estimator is also unbiased, but (weakly) more efficient than OLS when there is heteroscedasticity:

$$\widehat{\beta}_{OLS} = (X'X)^{-1}X'Y \tag{9}$$

$$\widehat{\beta}_{GLS} = (X'\Omega X)^{-1}X'\Omega Y, \tag{10}$$

where Ω is a (positive definite) $n \times n$ symmetric matrix. For heteroscedasticity, only the diagonal of this matrix contains non-zero elements.

When there is heteroscedasticity, the variance estimate of $\beta\widehat{_{OLS}}$ will no longer have the **correct coverage**.

Generalized Least Squares

Suppose the classical assumptions 1–4 hold but replace 5 with $5'$:

$$\epsilon \sim N(0, \Omega), \ \Omega < \infty.$$

Then the existence and unbiasedness results hold as before, but the result required by standard errors is replaced by

$$\widehat{\beta} \sim N(\beta, (X'X)^{-1}X'\Omega X(X'X)^{-1}),$$

and the efficiency result does not hold, that is, $\widehat{\beta}$ is no longer the best unbiased estimator.

GLS: Why the Problem for OLS?

Recall this line from our derivation of the Variance-Covariance Matrix for OLS:

$$E(\hat{\beta} - \beta)(\hat{\beta} - \beta)' = (X'X)^{-1}X'E(\epsilon\epsilon')X(X'X)^{-1}$$

The problem is that unlike before, $E(\epsilon\epsilon') \neq \sigma^2 I$. Now, $E(\epsilon\epsilon') = \Omega$. So the equation does not simplify and

$$\hat{\beta} \sim N(\beta, (X'X)^{-1}X'\Omega X(X'X)^{-1}).$$

Working with Omega

As long as Ω is known, the presence of serial correlation or heteroscedasticity does not render us incapable of testing hypotheses or constructing confidence intervals. However, if Ω is unknown (apart from a factor of proportionality), testing hypotheses and constructing confidence intervals is no longer a simple matter. One might be able to construct tests based on estimates of Ω, but the resulting statistics may have very complicated distributions.

If Ω is known, efficiency can be regained by applying OLS to a linear transformation of the original model, i.e.,

$$C^{-1}Y = C^{-1}X\beta + C^{-1}\epsilon \tag{11}$$

$$\text{or}$$

$$Y^{*} = X^{*}\beta + \epsilon^{*}, \tag{12}$$

where $Y^{*} = C^{-1}Y$, $X^{*} = C^{-1}X$, $\epsilon^{*} = C^{-1}\epsilon$ and C is a nonsingular factorization of Ω such that $CC' = \Omega$ and $C^{-1}\Omega C^{-1\prime} = I$.

This transformation ensures that
$E(\epsilon^{*}\epsilon^{*\prime}) = E(C^{-1}\epsilon\epsilon' C^{-1\prime}) = C^{-1}E(\epsilon\epsilon')C^{-1\prime} = C^{-1}\Omega C^{-1\prime} = I$.

The least squares estimator for the transformed model is:

$$\begin{aligned}\beta^* &= (X^{*\prime}X^*)^{-1}X^{*\prime}Y^* \\ &= (X'C^{-1\prime}C^{-1}X)^{-1}X'C^{-1\prime}C^{-1}Y \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y.\end{aligned}$$

The estimators β^* is called the **generalized least squares** (GLS) estimator.

If Ω is known, we obtain efficiency by transforming the model "back" to a form in which OLS gives the efficient estimator. However, if Ω is unknown, this transformation is not immediately available. It might be possible to estimate Ω, say be $\hat{\Omega}$, but $\hat{\Omega}$ is then random and so is the factorization $\hat{C}$ no longer applies.

Huber-White Heteroscedasticity-Consistent Covariance Matrix Estimator

The key to obtaining Heteroscedasticity-Consistent Covariance Matrix Estimator (HCCME) is to recognize that we do *not* have to estimate Ω consistently. That would indeed be an impossible task, since Ω as n diagonal elements to estimate. Recall that when there is heteroscedasticity

$$var[\beta_{OLS}] = (X'X)^{-1}X'\Omega X(X'X)^{-1}. \tag{13}$$

$$\tag{14}$$

The only tricky thing is to estimate the second factor—$X'\Omega X$. Huber-White showed that this second term can be estimated inconsistently by

$$X'\hat{\Omega}X, \tag{15}$$

The simplest version of $\hat{\Omega}$ has n^{th} diagonal elements equal to $\hat{u}_i^2$, the n^{th} squared least squares residual.

Huber-White Heteroskedastic Omega

$$\begin{bmatrix} \widehat{\epsilon_1}^2 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \widehat{\epsilon_2}^2 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \widehat{\epsilon_{N-1}}^2 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & \widehat{\epsilon_N}^2 \end{bmatrix}$$

Panel Corrected Standard Errors

You can now read the Beck, Katz, Alvarez, Garrett and Lange (1993) paper which describes Panel Correct Standard Errors. We will cover that next time.

We will also talk in detail about the bootstrap.

The Bootstrap

Bootstrap methods are a good way of obtaining confidence intervals and other statistical estimates which require generally weaker assumptions than the usual analytical approaches.

The bootstrap is also useful when we don't know of an analytical solution. We lack such solutions when:

1. we are using complicated statistics for which a sampling distribution is difficult to solve, such as two step estimators
2. we don't want to make population assumptions

Percentile Bootstrap Intervals

The basic setup is as follows:

- We have a sample of size n from a given from a probability distribution F

$$F \rightarrow (x_1, x_2, x_3, \cdots, x_n)$$

- The empirical distribution function $\hat{F}$ is defined to be the discrete distribution that puts equal probability, $\frac{1}{n}$, on each value x_i.

- A set X which is made up of various x_i has a probability assigned by $\hat{F}$:

$$\mathrm{pr}\hat{\mathrm{o}}\mathrm{b}\{X\} = \#\{x_i \in X\}/n$$

- We can make program via the plug-in principle. The plug-in estimate of a parameter $\theta = t(F)$ is $\hat{\theta} = t(\hat{F})$

Bootstrap Intervals Algorithm

- For a dataset (denoted "full sample") iterate the following B times where B is a large number such as 1,000.

  1. Generate a random sample of size n from $x_1, \cdots, x_n$ with replacement. This is called a bootstrap sample.
  2. calculate the statistic of interest using this bootstrap sample, denoted $\hat{\theta}_b$.

- The confidence interval for $\hat{\theta}$ (our full sample estimate) is obtained by taking the quantiles from the sampling distribution in our bootstrap samples: $\hat{\theta}_b$.

- For example, for 95% CI, calculate the 2.5% and 97.5% quantiles of the distribution of $\hat{\theta}_b$.

- An estimate for the full sample quantity is provided by:

$$\hat{\theta}^* = \sum_{b=1}^{B} \hat{\theta}_b / B$$

  This is an alternative to whatever we usually do in the full sample, whether it be the usual least squares estimates, the mean, median or whatever.

Bootstrapping Regression Coefficients

- For a dataset (denoted “full sample”) iterate the following B times.

- Generate a random sample of size n from $\{Y, X\}$, with replacement. Denote this sample of data S^b.

- Estimate your model using the dataset S^b. This results in a vector of bootstrap sample estimates of β denoted β_b.

$$Y^b = \hat{\beta}^b X^b + \hat{\epsilon}^b$$

- Create CIs and SEs in the usual bootstrap fashion based on the distribution $\hat{\beta}^b$.

Comments on Non-Parametric Bootstrap

The algorithms on the previous two slides are non-parametric bootstraps. And:

- We are treating the sample as a population and sampling from it.

- does makes that assumption that the observations x_i are independent. It does not assume that they are homoscedastic.

- can be applied to any function of the data which is smooth (this is a technical assumption) such as least squares regression.

- the resulting estimates, point estimates, CIs and standard errors have $\frac{1}{\sqrt{n}}$ asymptotics just like the usual normal theory estimates. There are better bootstraps with, for example, $\frac{1}{n}$ asymptotics. Examples of such bootstraps are: calibrated bootstrap, percentile-t bootstrap and the bias corrected, BC_a, bootstrap.

Distributions of Difficult to Handle Quantities

The bootstrap provides an easy way to to obtain the sampling distribution of quantities for which it is often not known or difficult to obtain sampling distributions. For example:

- CIs for the median or other quantiles
- CIs and SEs for LMS coefficients
- In a regression, is $\widehat{\beta}_1 > \widehat{\beta}_2$
- In a regression, is $\dfrac{\widehat{\beta}_1}{\widehat{\beta}_2} < \dfrac{\widehat{\beta}_3}{\widehat{\beta}_4}$
- Two stage regression estimation. Say a logit in the first stage and a linear model in the second.
- Goodness-of-fit for two model which may not be nested.

Bootstrapping Hypothesis Testing Algorithm

- For a dataset (denoted “full sample”) iterate the following B times.

1. Generate a random sample of size n from $x_1, \cdots, x_n$ with replacement. This is called a bootstrap sample.

2. calculate the test statistic of interest using this bootstrap sample, denoted $\hat{\theta}_b$.

- The bootstrap distribution of the test statistic is

$$\widehat{\mathrm{prob}}(H_o) = \#\left\{\hat{\theta}_b \geq \hat{\theta}\right\}/B,$$

where $\hat{\theta}$ is the full sample value of the statistic.

Parametric Bootstrap

The previous bootstraps assume that the observations are independent. But this may not be the case. Also we may want to assume that X is fixed. In these cases, one may estimate a parametric bootstrap if it can be assumed that the residuals of a parametric model are independent even if the original data is not. This often occurs with time-series models.

Parametric Bootstrap Algorithm

- Estimate the model in the full-sample:

$$Y = \hat{\beta} X + \hat{\epsilon}$$

- For a dataset (denoted "full sample") iterate the following B times.

- Generate a random sample of size n from $\hat{\epsilon}$, with replacement. Denote this $\hat{\epsilon}^b$. Create the bootstrap sample by adding the new $\hat{\epsilon}^b$ to the full sample X and $\hat{\beta}$ (there is NO estimation in this step, just addition):

$$Y^b = \hat{\beta} X + \hat{\epsilon}^b$$

- Estimate your model using the full sample X but the bootstrap sample Y^b. This results in a vector of bootstrap estimates of β estimates denoted β_b.

$$Y^b = \hat{\beta}^b X + \hat{\epsilon}^b$$

- Create CIs and SEs in the usual bootstrap fashion based on the distribution $\hat{\beta}^b$.

Bootstrap Readings

For bootstrap algorithms see Venables and Ripley (2002). Especially, p.133-138 (section 5.7) and p.163-165 (section 6.6). The **R** "boot" command is of particular interest. ▮

For a statistical discussion see Chapter 16 (p.493ff) of Fox (2002). ▮

For additional readings see:

- Efron, Bradley and Tibshirani, Robert J. 1994. *An Introduction to the Bootstrap.* Chapman & Hall. ISBN: 0412042312.

- Davison, A. C. and Hinkely, D. V. 1997. *Bootstrap Methods and their Applications.* Cambridge University Press.

Two Nice Bootstrap Properties

Bootstrap CI estimates are:

1. Range Preserving. The confidence interval cannot extend beyond the range which estimates of θ range in bootstrap samples. For example, if we are bootstrapping estimates of a proportion, our confidence interval cannot be less than 0 or greater than 1. The usual normal theory confidence intervals are not range persevering. ▮

2. Take into account more general uncertainty. Many estimators, such as LMS, have the usual statistical uncertainty (due to sampling) but also have uncertainty because the algorithm used to obtain the estimates is stochastic or otherwise suboptimal. Bootstrap CIs will take this uncertainty into account. See the **R** help on the lqs function for more information about how the LMS estimates are obtained.

Causal Inference with Observational Data: An Example

- Survey research is strikingly uniform regarding the ignorance of the public (e.g., Berelson, Lazarsfeld, and McPhee 1954; Campbell, Converse, Miller, and Stokes 1960; Zaller 1992).

- Although the fact of public ignorance has not been forcefully challenged, the meaning of this observation has been (Sniderman 1993).

- Voters could use information such as polls, interest group endorsements and partisan labels to vote like their better informed compatriots (e.g., Lupia 2004; McKelvey and Ordeshook 1985a,b, 1986).

- Many claim voters use cues to preform information arbitrage, but there is almost no data analysis supporting this assertion. And there is analysis challenging it (e.g., Bartels 1996).

Where We are Going: Advanced Democracies

In democracies which have all of the institutions thought to be important for providing reliable electoral cues (e.g., U.S., Canada, U.K.):

- voters who learn information vote no differently on election day than voters who do not
- before election day, voters who learn information do differ in their vote intentions
- in the U.S., this effect is present in September but is insignificant in November
- with cross-sectional data analysis no information effects are found: the uninformed vote like the informed

Where We are Going: New Democracies

In new electoral democracies (e.g., Mexico, Russia, Serbia):

- voters who learn information vote differently on election day

- with cross-sectional data analysis information effects are also found: the uninformed vote differently than the informed

Hypothesis

Developed democratic institutions are required to preform the information arbitrage.

- Institutions such as:
  - interest groups
  - parties
  - opinion polls
  - media
- and voters who know how to use these cues (Lupia 2004)

Summary of Countries Analyzed

	Advanced Democracy	New Democracy
No Information Effects	Canada 79,80,84,88,93,97 U.K. 79,83,87,92,97 **U.S.** **1968–2000** Germany fmr. FRG 94	
Information Effects		**Mexico** **(00)** Russia (95/96) Serbia (90,92) Germany fmr. GDR 94

Where We are Not Going

- I do not argue that voters in advanced democracies make choices which are in some general sense optimal or are the same choices they would make if they were "perfectly" informed.

- However, I do argue that although voters in advanced democracies may make mistakes, these mistakes are made by the well and poorly informed alike—by the attentive and the inattentive.

Methodological Issues

- Goal: to make causal inferences using survey data
  The Rubin-Holland-Neyman causal model is used.

- Matching estimators ensure that we compare apples to apples and do not combine disparate observations.

- My **R** package called "Matching" implements rigorous matching tests and principled standard errors.

- Bootstrap does not work with Matching estimators. Why?

- Imputation is used to correct panel attrition based on the additive nonignorable model (Rubin 1976; Hausman and Wise 1979) which we will discuss late in the course.

Cross-Sectional Data

- For causal inference, it is more dubious to rely on cross-sectional data.

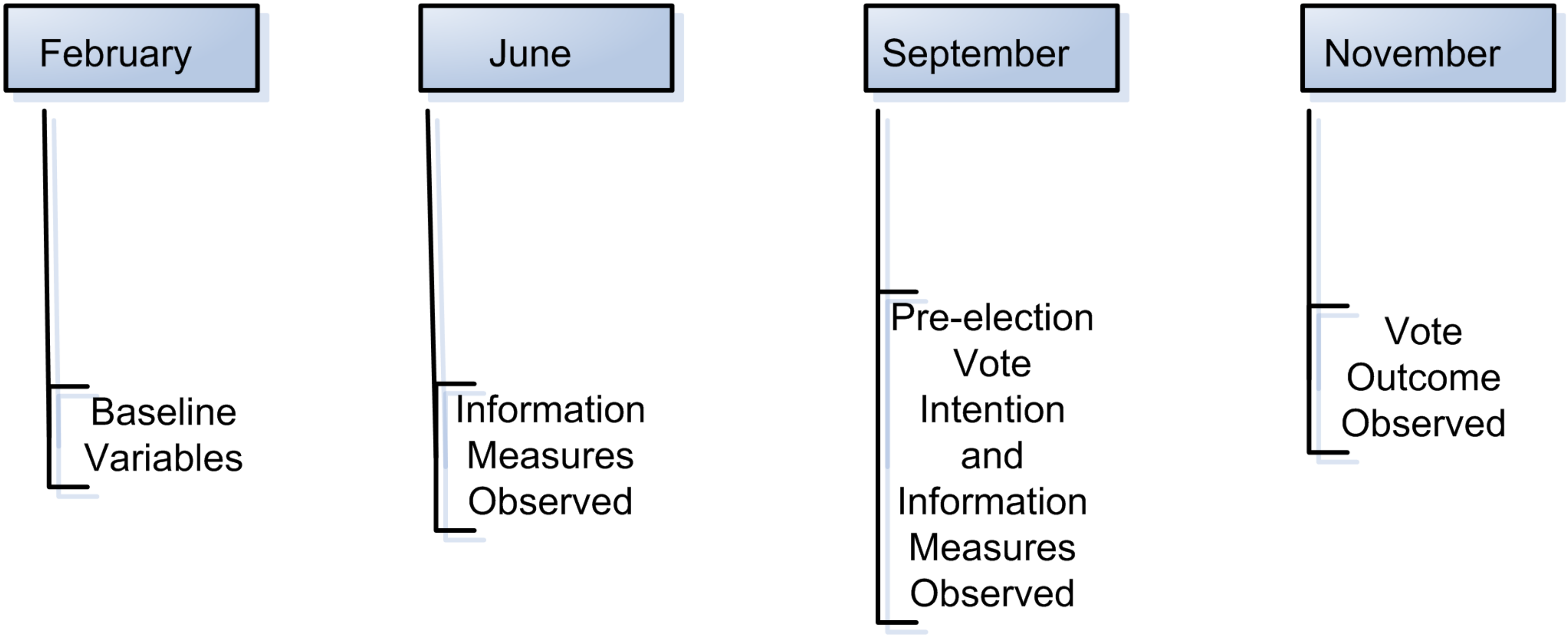
- Most of the measures known to be correlated with vote choice are endogenous.

- If you condition on too much, you have post-treatment bias.

- If you condition only on items which you are sure are pre-treatment, you are only left with sociodemographic variables.

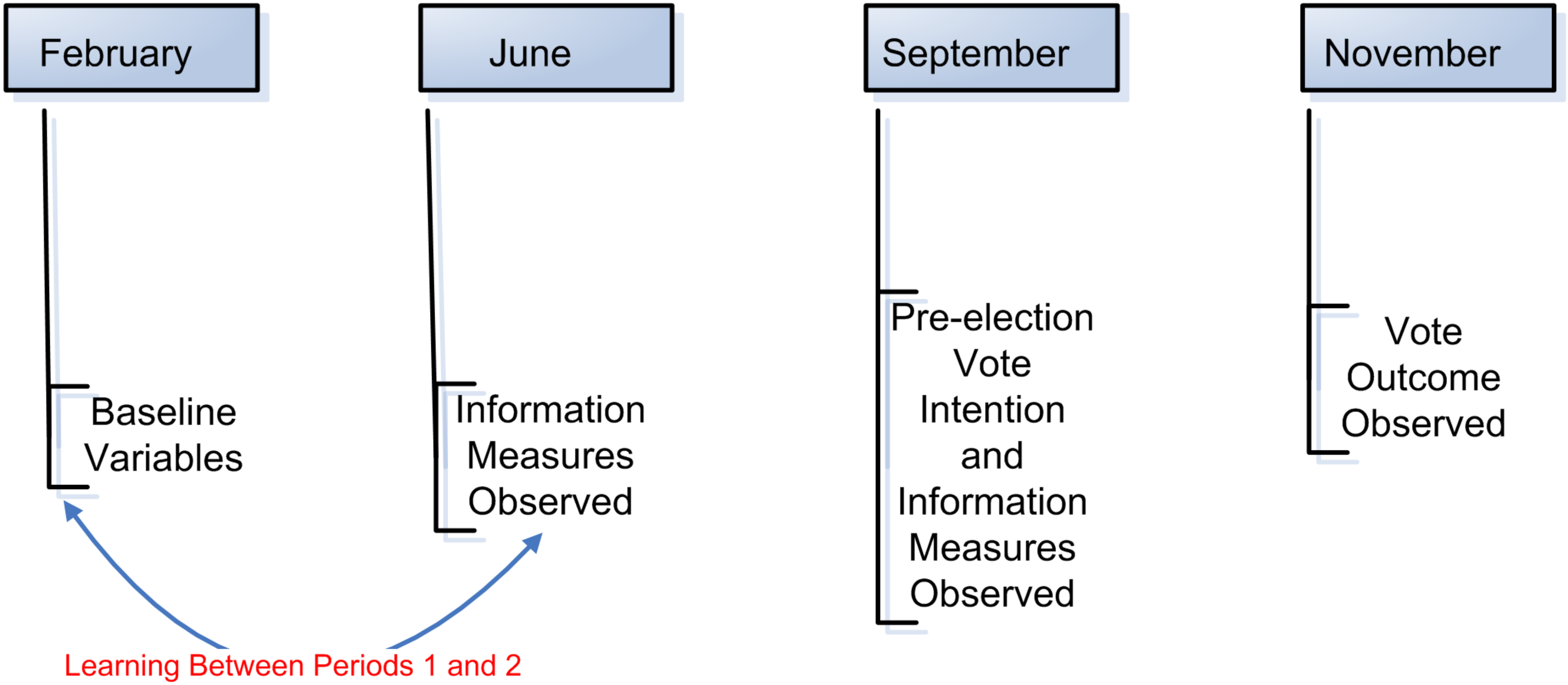
- Previous results, such as Bartels (1996), were driven by comparing apples and oranges—i.e., by untested functional form assumptions.

Panel Data from the 1980 U.S. NES


February
Baseline Variables
June
Information Measures Observed
September
Pre-election Vote Intention and Information Measures Observed
November
Vote Outcome Observed

- $t = 1$ is the first panel wave: January 22–February 25.
- $t = 2$ is the second wave: June 4–July 13.
- $t = 3$ is the third wave: September 2–October 1.
- $t = 4$ is the fourth wave: November, post-election.

Panel Data from the 1980 U.S. NES



February

Baseline Variables

June

Information Measures Observed

September

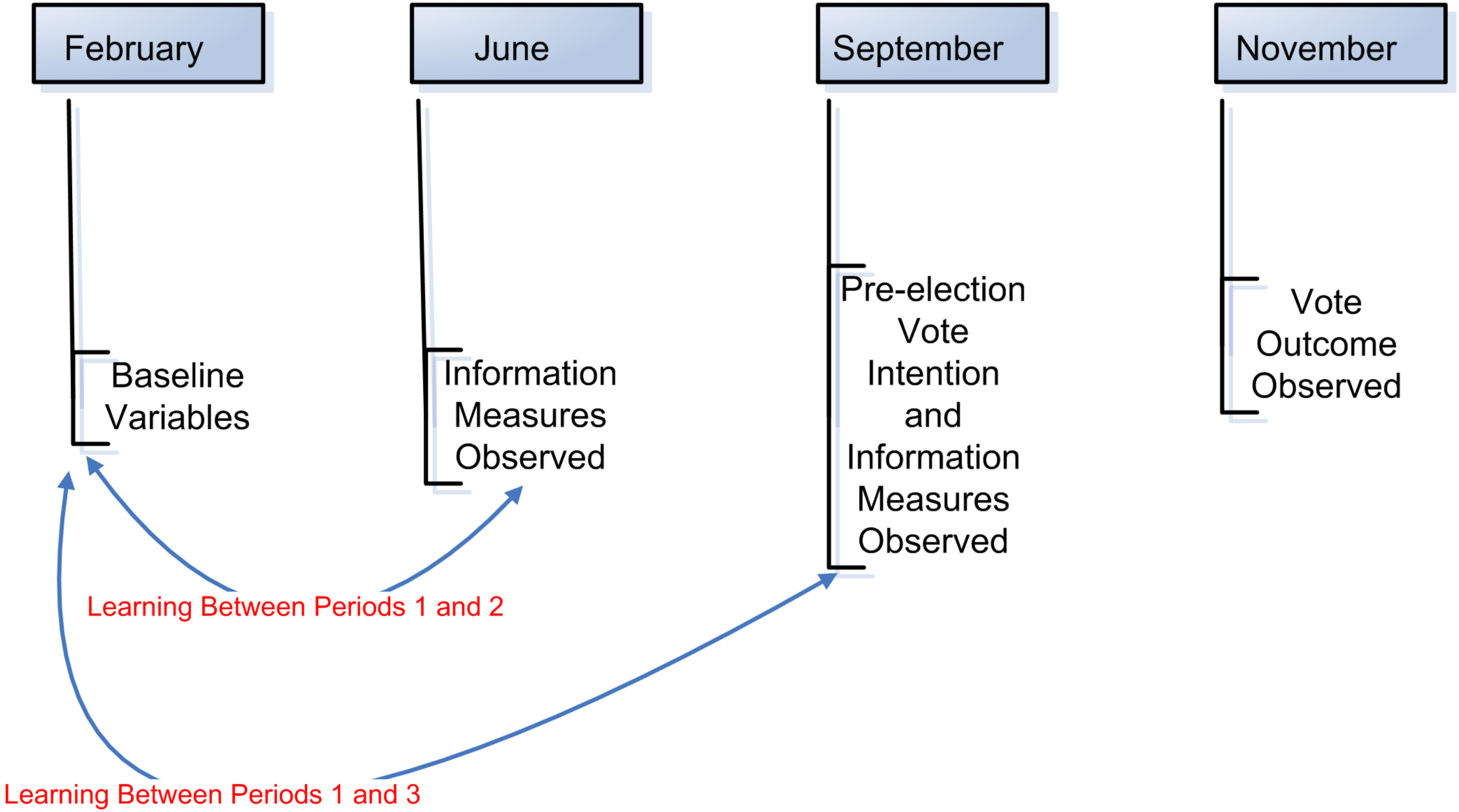
Pre-election Vote Intention and Information Measures Observed

November

Vote Outcome Observed

Learning Between Periods 1 and 2

Panel Data from the 1980 U.S. NES



February

Baseline Variables

June

Information Measures Observed

September

Pre-election Vote Intention and Information Measures Observed

November

Vote Outcome Observed

Learning Between Periods 1 and 2

Learning Between Periods 1 and 3

Fundamental Problem of Causal Inference

- Fundamental problem: not observing all of the potential outcomes or counterfactuals

- Let Y_{i1} denote i's vote intention when voter i learns during the campaign (i.e., is in the treatment regime).

- Let Y_{i0} denote i's vote intention when voter i does not learn during the campaign (i.e., is in the control regime).

- Let T_i be a treatment indicator: 1 when i is in the treatment regime and 0 otherwise.

- The observed outcome for observation i is
  $Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}$.

- The treatment effect for i is
  $\tau_i = Y_{i1} - Y_{i0}$.

Experimental Data

- If assignment to treatment is randomized, the inference problem is straightforward because the two groups are from the same population: $\{Y_{i1}, Y_{i0} \perp T_i\}$.

- Observations in the treatment and control groups are not exactly alike, but they are comparable—i.e., they are exchangeable.

- The Average Treatment Affect (ATE) can be estimated simply:

$$\begin{aligned} \tau &= \text{Mean Outcome for the treated} - \\ &\quad \text{Mean Outcome for the control} \end{aligned}$$

More formally:

$$\begin{aligned} \tau &= E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 0) \\ &= E(Y_i|T_i = 1) - E(Y_i|T_i = 0) \end{aligned}$$

Observational Data

- With observational data, the treatment and control groups are not drawn from the same population.

- Progress can be made if we assume that the two groups are comparable once we condition on observable covariates denoted by X_i.

- This is the conditional independence assumption:

$$\{Y_{i1}, Y_{i0} \perp T_i | X_i\},$$

the reasonableness of this assumption depends on the quality of the X variables.

ATT

- With observational data, the treatment and control groups are not drawn from the same population.

- Thus, we often want to estimate the average treatment effect for the treated (ATT):

$$\tau|(T = 1) \quad = \quad E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 1)$$

- We need the assumptions of unconfoundedness and overlap which together are called *strong ignorability* (Rosenbaum and Rubin 1983):
  - unconfoundedness: the conditional independence assumption stated above
  - overlap: that the X covariates for treated and control overlap—i.e., that their distributions have the same support.

ATT II

- For ATT these can be weakened to mean independence and a subset of overlap (Heckman, Ichimura, Smith, and Todd 1998):

  1. $E[Y(w)|T, X] = E[Y(w)|X]$, for $w = 0, 1$
  2. the support of X for the treated is a subset of the support of X for the control observations.

- Then, we obtain the usual exchangeability results:

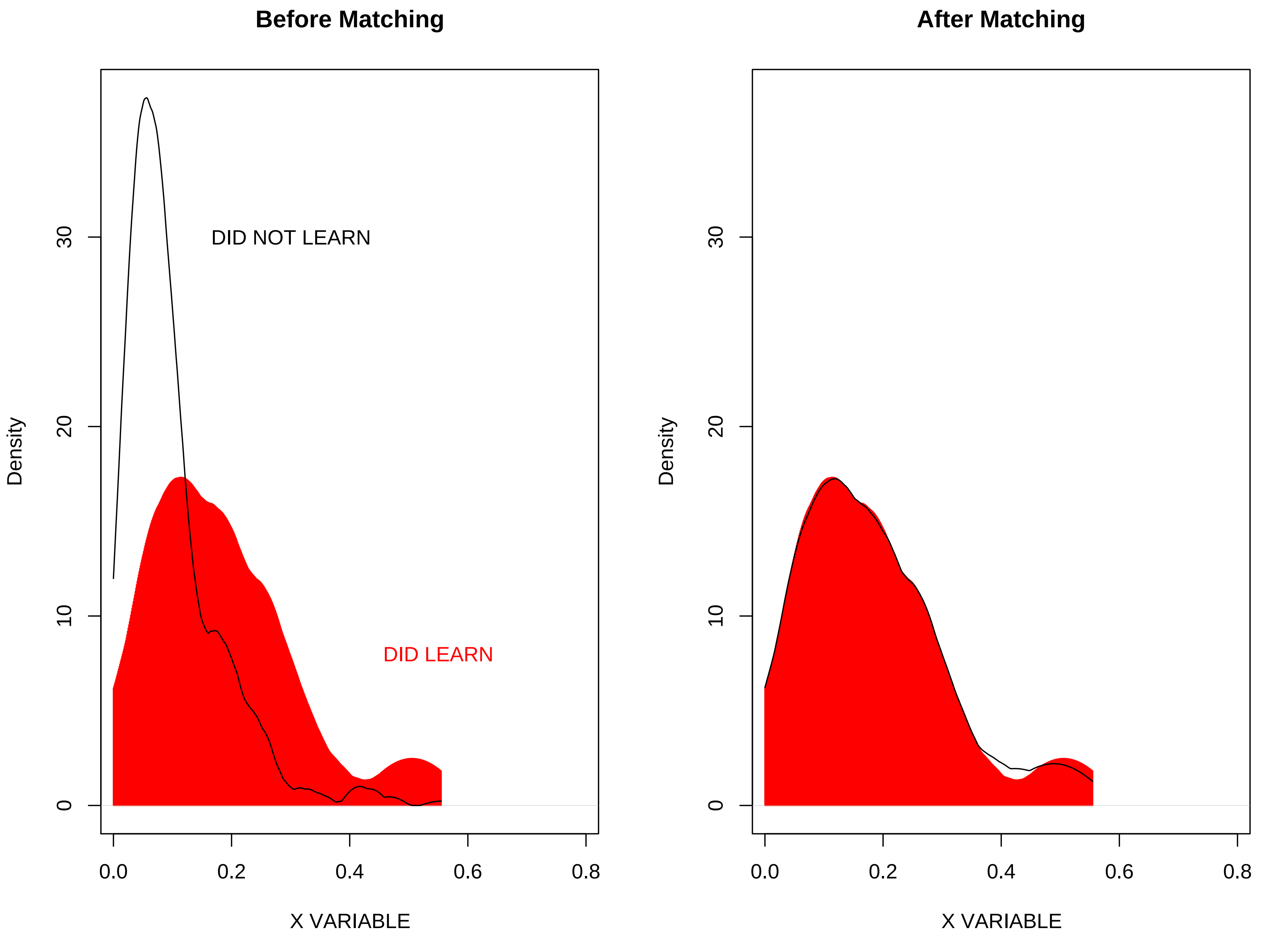
$$\begin{aligned} E(Y_{ij}|X_i, T_i = 1) &= E(Y_{ij}|X_i, T_i = 0) \\ &= E(Y_{ij}|X_i, T_i = j) \end{aligned}$$

- ATT can then be estimated by

$$\begin{aligned} \tau|(T = 1) &= E\{E(Y_i|X_i, T_i = 1) - \\ &\quad E(Y_i|X_i, T_i = 0) \mid T_i = 1\} \end{aligned}$$

where the outer expectation is taken over the distribution of $X_i|(T_i = 1)$.

Observational Data: Matching



Matching

- The nonparametric way to condition on X is to exactly match on the covariates. This is an old approach going back to at least Fechner (1966 [1860]), the father of psychophysics.

- This approach fails in finite samples if the dimensionality of X is large.

- If X consists of more than one continuous variable, multivariate matching is inefficient: matching estimators with a fixed number of matches do not reach the semi-parametric efficiency bound (Abadie and Imbens 2004).

- An alternative way to condition on X is to match on the probability of being assigned to treatment—i.e., the propensity score. The propensity score is just one dimensional.

- More formally the propensity score (also called the balancing score) is:

$$\Pr(T_i = 1|X_i) = E(T_i|X_i)$$

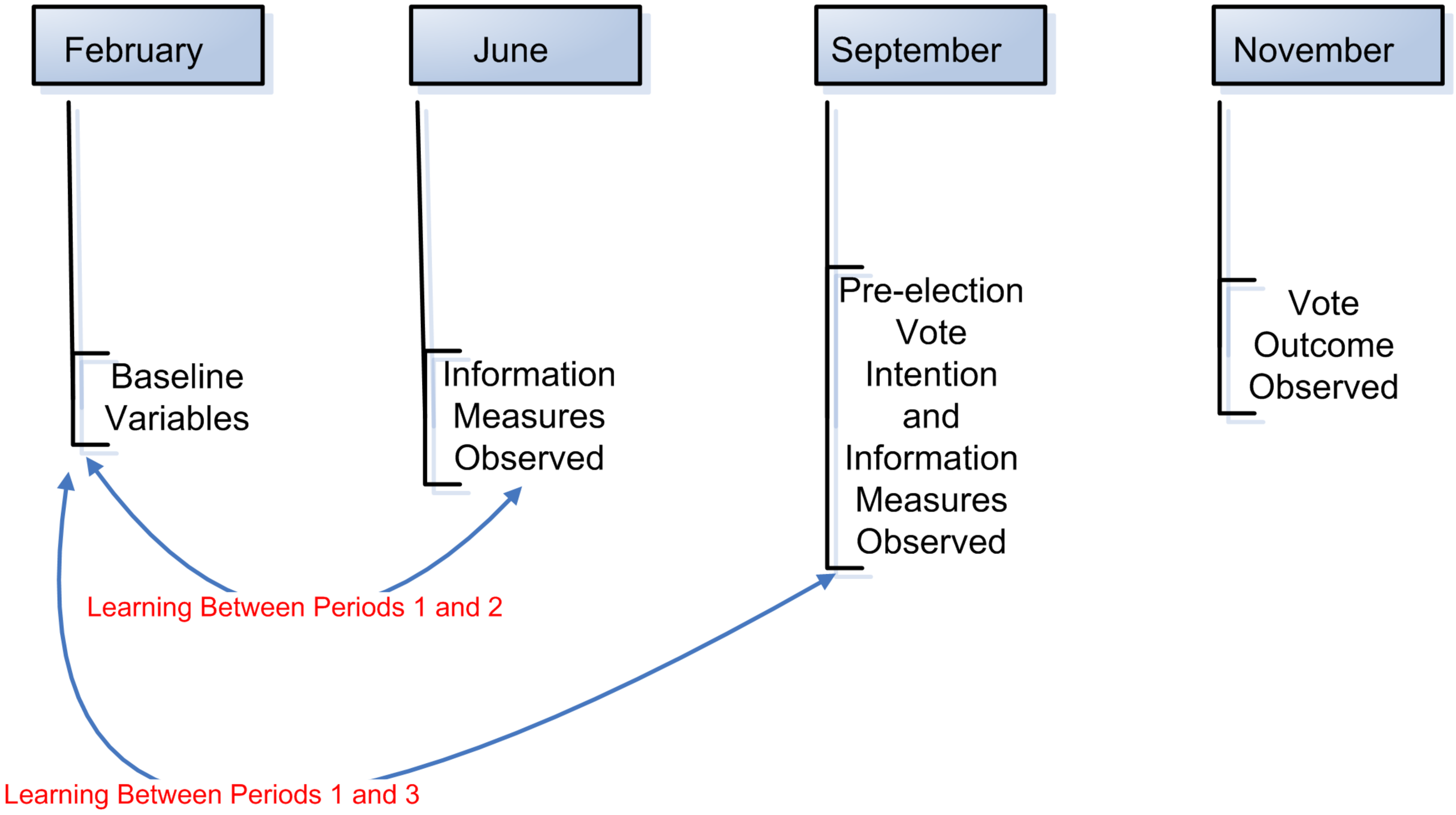
We will learn how to estimate this next class (e.g., logistic regression).

Measuring Political Information

- There is little agreement on how best to measure political information.

- I measure information in many different ways:
  - interview ratings
  - the Zaller information measure (interviewer ratings plus factual general political questions)
  - the ability to spatially locate the candidates in issue space
  - the ability to accurately place candidates in issue space
  - the ability to accurately recall specific campaign events

The United States: 1980 NES



February

June

September

November

Baseline Variables

Information Measures Observed

Pre-election Vote Intention and Information Measures Observed

Vote Outcome Observed

Learning Between Periods 1 and 2

Learning Between Periods 1 and 3

Learning During the Campaign

Issue/panel	Reagan Placement Proportion	Carter Placement Proportion	C $<$ R All	C $<$ R Of Placers
Defense $:				
p1 (Feb)	0.598	0.829	0.351	0.578
p2 (June)	0.715	0.831	0.496	0.690
p3 (Sept)	0.802	0.862	0.638	0.794
Lib-Con:				
p1	0.615	0.664	0.409	0.650
p2	0.707	0.709	0.527	0.735
p3	0.691	0.698	0.531	0.765

Learning During the Campaign

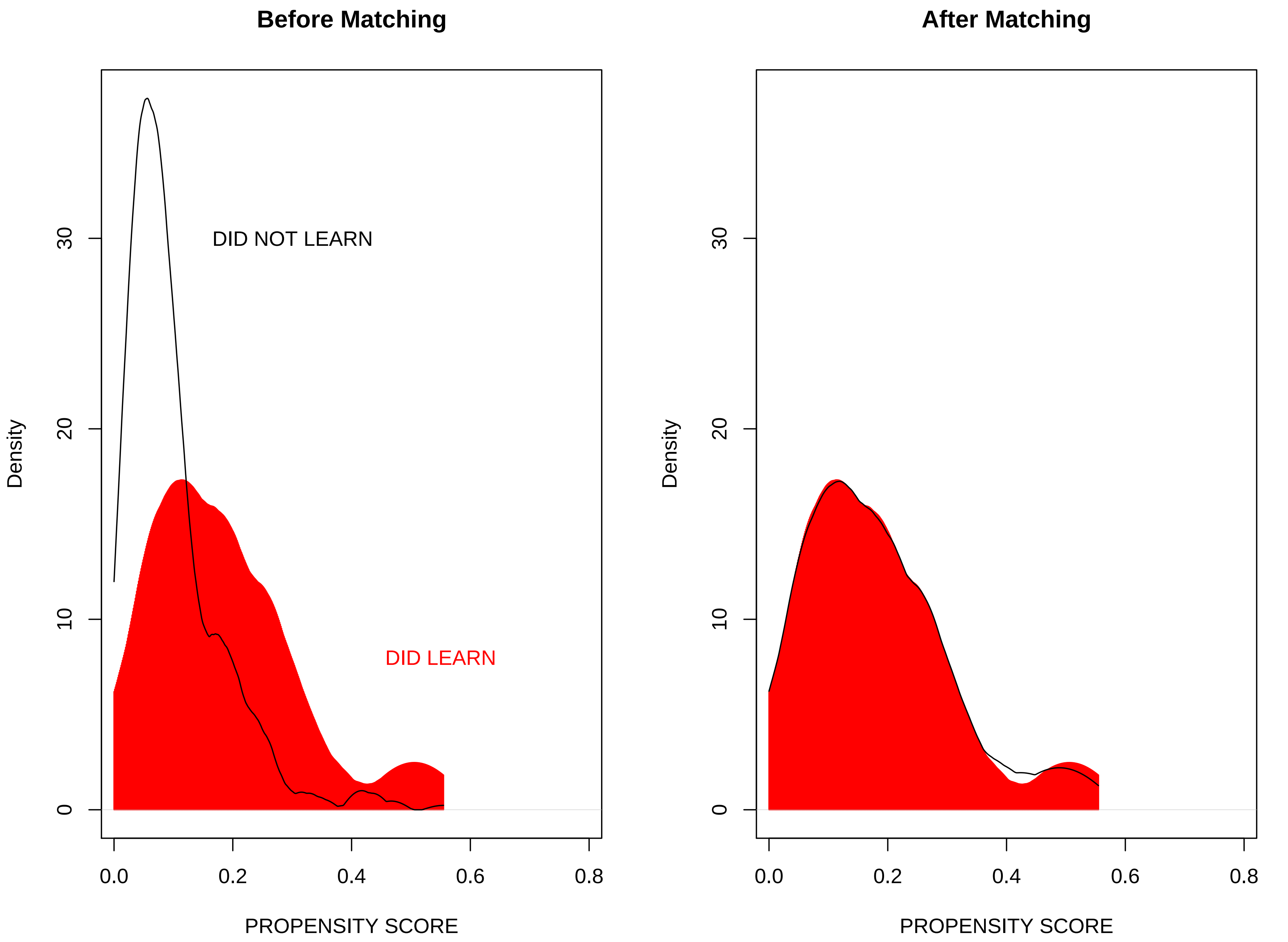
	Reagan	Carter	C < R	C < R
Issue/panel	Placement	Placement	All	Of Placers
	Proportion	Proportion		
Gov. Services $:				
p1	0.605	0.789	0.591	0.667
p2	0.687	0.804	0.567	0.652
p3	0.747	0.802	0.560	0.672
“Get Along” w/Russia:				
p1	0.658	0.885	0.380	0.576
p2	0.707	0.851	0.444	0.626
p3	0.742	0.833	0.502	0.673

Treatment Definitions: Issue Scales

- Treatment effects have been estimated for five such measures separately where treatment has been defined in many ways including:
  - T = Changing from being unable to place Reagan to being able to place him.
  - T = Changing from being unable to place Carter to the left of Reagan to being able to do so.

- Results presented for T = Changing from being unable to place Reagan on five issue questions to being able to do so.

Densities of Propensity Scores



Outcomes

The results of one representative outcome are presented:

- $Y =$ Vote intention for the Republican Party

Estimated Effect of Changing Information Level

	Changes from Feb to June		Changes from Feb to Sep
Outcome	September	November	November
Republican Vote	0.180	0.0488	−0.0384
	(0.0856)	(0.0837)	(0.0903)

SUTVA

1. We require that “the [potential outcome] observation on one unit should be unaffected by the particular assignment of treatments to the other units” (Cox 1958, §2.4). This is often called the stable unit treatment value assumption (SUTVA) (Rubin 1978).

2. SUTVA implies that Y_{i1} and Y_{i0} (the potential outcomes for person i) in no way depend on the treatment status of any other person in the dataset.

3. SUTVA is not just statistical independence between units!

4. Causal inference relies on a counterfactual of interest (Sekhon 2004), and the one which is most obviously relevant for political information is “how would Jane have voted if she were better informed?”.

5. There are other theoretically interesting counterfactuals which, because of SUTVA, I do not know how to empirically answer such as “who would have won the last election if *everyone* were well informed?”.

Properties

1. The modeling portion of the estimator is limited to the model of $p(X_i)$. Estimation of this model requires no knowledge of the outcome.

2. Unlike in the regression case, there is a clear standard for choosing an optimal model; it is the model which balances the covariates, X.

3. The key assumption required is that no variable has been left unobserved which is correlated with treatment assignment *and* with the outcome. This is called the selection on observables assumption.

4. No exclusion assumptions are being made as is necessary with instrumental variable approaches and related structural approaches such as Heckman selection models (Heckman 1979).

5. No functional form is implied for the relationship between treatment and outcome. No homogeneous causal effect assumption has been made.

6. Since we are interested in a lower dimensional representation of X_i, in particular $p(X_i)$, we do not need to estimate consistently any of the individual parameters in our propensity model—they don't even need to be identified.

Conclusion of the Example

- Although voters in advanced democracies may make mistakes, these mistakes are made by the well and poorly informed alike—by the attentive and the inattentive.

- Civic competence

- Hidden bias: results are resilient

- Combining observations

- The stable unit treatment value assumption (SUTVA)

- Deliberative polls which find significant information effects are about a very different counterfactual than the one explored here (Fishkin 1997; Luskin, Fishkin, and Jowell 2002).

Generalized Linear Models

Generalized linear models (GLMs) extend linear models to accommodate both non-normal distributions and transformations to linearity. GLMs allow unified treatment of statistical methodology for several important classes of models.

This class of models can be described by two components: stochastic and systematic.

- Stochastic component:

  1. Normal distribution for LS: $\epsilon_i \sim N(0, \sigma^2)$

  2. For logistic regression:

$$Y_{\text{Bern}}(y_i|\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} \quad \left\{ \begin{matrix} \pi_i & \text{for } y = 1 \\ 1-\pi_i & \text{for } y = 0 \end{matrix} \right\}$$

- Systematic Component:

  1. For LS: $E(Y_i|X_i) = X\beta$
  2. For logistic regression:

$$\Pr(Y_i = 1|X_i) \equiv E(Y_i) \equiv \pi_i = \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}$$

Defining Assumptions

A generalized linear model may be described by the following assumptions:

- There is a response Y observed independently at fixed values of stimulus variables $X_1, \cdots, X_k$.

- The stimulus variables may only influence the distribution of Y through a single linear function called the *linear predictor* $\eta = \beta_1 X_1 + \cdots + \beta_k X_k$.

- The distribution of Y has density of the form

$$f(Y_i; \theta_i, \phi) = \exp[A_i\{Y_i\theta_i - \gamma(\theta_i)\}/\phi + \tau(Y_i, \phi/A_i)], \tag{16}$$

where ϕ is a *scale parameter* (possibly known). A_i is a *known* prior weight and parameter θ_i depends upon the linear predictor.

This is the exponential family of distributions. Most of the distributions you know are part of this family including the normal, binomial and Poisson.

Models of the Mean

The mean $\mu = E(Y)$ is a smooth invertible function of the linear predictor:

$$\mu = m(\eta), \tag{17}$$

$$\eta = m^{-1}(\mu) = l(\mu), \tag{18}$$

where the inverse function $l(\cdot)$ is called the *link function*

For LS we use the identity link (also called the canonical link): $l(\mu) = \mu$. Thus,

$$X_i\beta = \eta_i = m(\mu_i) = \mu_i = E(Y_i)$$

For logistic regression we use the logit link: $l(\mu) = \log\left(\frac{\pi}{1-\pi}\right)$. Thus,

$$X_i\beta = \eta_i = m(\mu_i) = \frac{\exp(\eta_i)}{1+\exp(\eta_i)} = \mu_i = E(Y_i) = \Pr(Y_i = 1|X_i),$$

note that $\frac{\exp(\eta_i)}{1+\exp(\eta_i)}$ is the inverse logit link.

GLM: Gaussian

GLMs allow unified treatment of statistical methodology for several important classes of models. We here consider a couple of examples starting with the normal distribution.

The probability distribution function in the form one usually sees the distribution is:

$$f(y;\mu) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right],$$

where μ is the parameter of interest and σ^2 is regarded, in this setting, as a nuisance parameter.

The following is the canonical form of the distribution

$\theta = \mu$, $\gamma(\theta) = \theta^2/2$ and $\phi = \sigma^2$ so we can write:

$$\log f(y) = \frac{1}{\phi}\{y\mu - \frac{1}{2}\mu^2 - \frac{1}{2}y^2\} - \frac{1}{2}\log(2\pi\phi).$$

Notes on the Gaussian

If ϕ, the variance or what is called the dispersion in the GLM framework, were known the distribution of y would be a one-parameter canonical exponential family.

An unknown ϕ is handled as a nuisance parameter by moment methods. For example, note how $\hat{\sigma}^2$ is estimated in the least squares case.

We can do this because θ and ϕ are orthogonal parameters. This means that we can estimate θ and then conditioning on this estimate, calculate ϕ. We don't have to jointly estimate both.

GLM: Poisson

For a Poisson distribution with mean μ we have

$$\log f(y) = y \log(\mu) - \mu - \log(y!), \tag{19}$$

where $\theta = \log(\mu)$, $\phi = 1$, and $\phi(\theta) = \mu = e^{\theta}$.

GLM: Binomial

For a binomial distribution with fixed number of trials a and parameter p we take the response to be $y = s/a$ where s is the number of "successes". The density is

$$\log f(y) = a\left[y \log \frac{p}{1-p} + \log(1-p)\right] + \log \binom{a}{ay} \tag{20}$$

where we take $A_i = a_i$, $\phi = 1$, θ to be the logit transform of p and $\gamma(\theta) = -\log(1-p) = \log(1+e^\theta)$.

The functions supplied with **R** and **S** for handling generalized linear modeling distributions include `gaussian`, `binomial`, `poisson`, `inverse.gaussian` and `Gamma`.

Logistic Regression

Let Y^* be a continuous unobserved variable such as the propensity to vote for a particular political party, say the Republican Party, or the propensity to be assigned to treatment.

Assume that Y_i^* has some distribution whose mean is μ_i. And that Y_i^* and Y_j^* are independent for all $i \neq j$, conditional on X.

We observe Y_i such that:

$$Y_i = \left\{ \begin{array}{lcl} 1 \text{ if } Y_i^* & \geq & \tau \text{ if } i \text{ is treated or votes Republican} \\ 0 \text{ if } Y_i^* & < & \tau \text{ if } i \text{ is NOT treated or does NOT vote Republican} \end{array} \right\}$$

Since Y^* is unobserved we usually define the threshold, τ, to be zero.

Note that if Y_i^* is observed and that it is normally distributed conditional on X, we have LS regression.

If only Y_i is observed and that Y^* is distributed standardized logistic (which is very similar to the normal distribution), we obtain the logistic regression model.

Note that our guess of Y^* is $\hat{\mu}_i$–or, equivalently, $\hat{\eta}_i$!

Logistic Regression Coefficients

The estimated coefficients, $\widehat{\beta}$, in a logistic regression can be interpreted in a manner very similar to that the coefficients in LS. But there are important differences which complicate the issue.

We should interpret β as the regression coefficients of Y^* on X. So, $\widehat{\beta}_1$ is what happens to Y_i^* when X_1 goes up by one unit when all of the other explanatory variables remain the same.

The interpretation is complicated by the fact that the link function is nonlinear. So the effect of a one unit change in X_1 on the predicted probability is different if the other explanatory variables are held at one given constant value versus another. Why is this?

Logistic Regression Assumptions and Properties

The assumptions of logistic regression are very similar to those of least squares—see the assumptions slide. But there are some important difference. The most important is that in order to obtain consistency, logistic regression requires the homoscedasticity assumption while LS does not.

Another important difference is that logistic regression is NOT unbiased. But it is consistent.

The Propensity Score

Recall the propensity score. The propensity score allows us to match on X without having to match all of the k-variables in X. We may simply match on the probability of being assigned to treatment—i.e., the propensity score. The propensity score is just one dimensional.

More formally the propensity score (also called the balancing score) is:

$$\Pr(T_i = 1 | X_i) = E(T_i | X_i)$$

We we may estimate the propensity score using logistic regression where the treatment indicator is the dependent variable.

We then match on either the predicted probabilities $\hat{\pi}_i$ or the linear predictor $e\hat{t}a_i$. It is in fact preferable to match on the linear predictor because the probabilities are compressed.

For matching, we do not care about the parameters, $\hat{\beta}$. We only care about getting good estimates of $\hat{\pi}_i$.

Correct Specification of the Propensity Score?

It is unclear how one should correctly specify the propensity score. The good news is that unlike the usual circumstances, in the case of matching, the propensity score is serving a very simple purpose: matching on the propensity score should induce balance on the baseline covariates. This can be directly tested by a variety of tests such as difference of means. But difference of means are not by themselves sufficient.

Logistic Regression Loss Function and Estimation

The loss function for logistic regression is:

$$\sum_{i=1}^{n} - \left(Y_i \log(\pi_i) + (1 + Y_i) \log(1 - \pi_i)\right),$$

recall that $\pi_i = \dfrac{\exp(X_i \beta)}{1 + \exp(X_i \beta)}$.

The reasons for this loss function will be made clear when you take a maximum likelihood course. As will how this loss function is usually minimized—iterated weighted least squares.

See Venables and Ripley (2002) for more estimation details.

Obtaining Balance by Matching I

The nonparametric way to condition on X is to exactly match on the covariates.

Theory of Matching

Assumption 1: differences between groups are explained by observables

$$Y_0 \perp T | X \tag{A1}$$

Assumption 2: there exists common support for individuals in both groups

$$0 < P(T = 1 | X = x) < 1 \quad \forall x \in X \tag{A2}$$

Obtaining Balance by Matching II

Practice of Matching

Step 1: for each individual i in the treated group (T=1) find one or more individuals i' from the untreated group (T=0) who is equivalent (or proximate) in terms of X.

(a) If more than one observation matches, take average

(b) If no exact matches exist, either drop case (because of non-common support) or take closest / next best match(es)

Step 2:

$$\bar{\tau} = \frac{1}{N}\sum(Y_{1i}|X) - \frac{1}{N'}\sum(Y_{0i'}|X)$$

A simple example

i	T_i	X_i	Y_i	$\mathcal{J}$	$Y_i(0)$	$\hat{Y}_i(1)$	$K_M(i)$
1	0	2	7	$\{5\}$	7	8	3
2	0	4	8	$\{4,6\}$	8	7.5	1
3	0	5	6	$\{4,6\}$	6	7.5	1

i	T_i	X_i	Y_i	$\mathcal{J}$	$\hat{Y}_i(0)$	$Y_i(1)$	$K_M(i)$
4	1	3	9	$\{1,2\}$	7.5	9	2
5	1	2	8	$\{1\}$	7	8	1
6	1	3	6	$\{1,2\}$	7.5	6	1
7	1	1	5	$\{1\}$	7	5	0

where observed outcome is Y, covariate to match on is X, $\mathcal{J}$ indexes matches, $K_M(i)$ is number of times used as a match divided by $\#\mathcal{J}$;
$\hat{Y}_i = 1/\#\mathcal{J}(i) \sum_{l \in \mathcal{J}(i)} Y_l$.

Simple estimate of ATT (Abadie and Imbens Jan 2004):

$$\begin{aligned} \bar{\tau} &= \frac{1}{N}\sum(Y_{1i}|X) - \frac{1}{N'}\sum(\hat{Y}_{0i'}|X) = \frac{1}{N}\sum(T_i + (1-T_i)K_M(i))Y_i \\ &= .1428 \end{aligned}$$

Alternative test statistics

Treatment may be multiplicative, rather than constant additive. So the ratio of treated/non-treated should be constant: e.g., $Y_1 = \alpha Y_0$

$$\log(Y_1) - \log(Y_0) = \log(Y_1/Y_0) = \log(\alpha)$$

so,

$$\bar{\tau}_2 = \frac{1}{N}\sum \log(Y_{1i}|X) - \frac{1}{N'}\sum \log(Y_{0i'}|X)$$

Robust way to discern differences (e.g., if you fear outliers distorting means) transform to ranks:

Let R_i be the rank of the i observation, pooling over treated and non-treated.

$$\bar{\tau}_3 = \frac{1}{N}\sum (R_i|X) - \frac{1}{N'}\sum (R_{i'}|X)$$

Comments

If you think the variances are changed by treatment, but not the location, then test differences in variances.

The point: your test statistic should be motivated by your theory of what is changing when manipulation / treatment occurs!

One could in theory estimate any model one wishes on the matched data. Such as regression, logistic regression etc.

Obtaining Balance on Many Dimensions

Theory of Propensity Scores

PROBLEM: Simple matching approach fails in finite samples if the dimensionality of X is large.

An alternative way to condition on X is to match on the probability of being assigned to treatment—i.e., the propensity score. The propensity score is just one dimensional.

Defn: Propensity score is the probability that a person is treated/manipulated.

$$b(x) = P(T = 1|X = x)$$

Rubin and Rosenbaum show that we can reduce matching to this single dimension, by showing that the assumptions of matching (A1 and A2) lead to,

$$Y_0 \perp T | P(X)$$

Propensity Score Matching

Matching on b(X) similar to matching on X:

Step 1: for each Y_{1i} find Y_{0j} which has similar (or proximate) b(X).

One-one matching:
– Nearest-neighbor: choose closest control on $P(X_i)$
– Caliper: same, but drop cases which are too far

Many-one: use weighted average of neighbors near $P(X_i)$

$$Y_{0i} = \sum_j W_{ij} Y_j$$

E.g., using a weighted method where

$$W \propto K(\frac{P_i - P_j}{h})$$

Step 2: same, take average over matched values of Y_{1i} and Y_{0i}

Florida 2004 Example

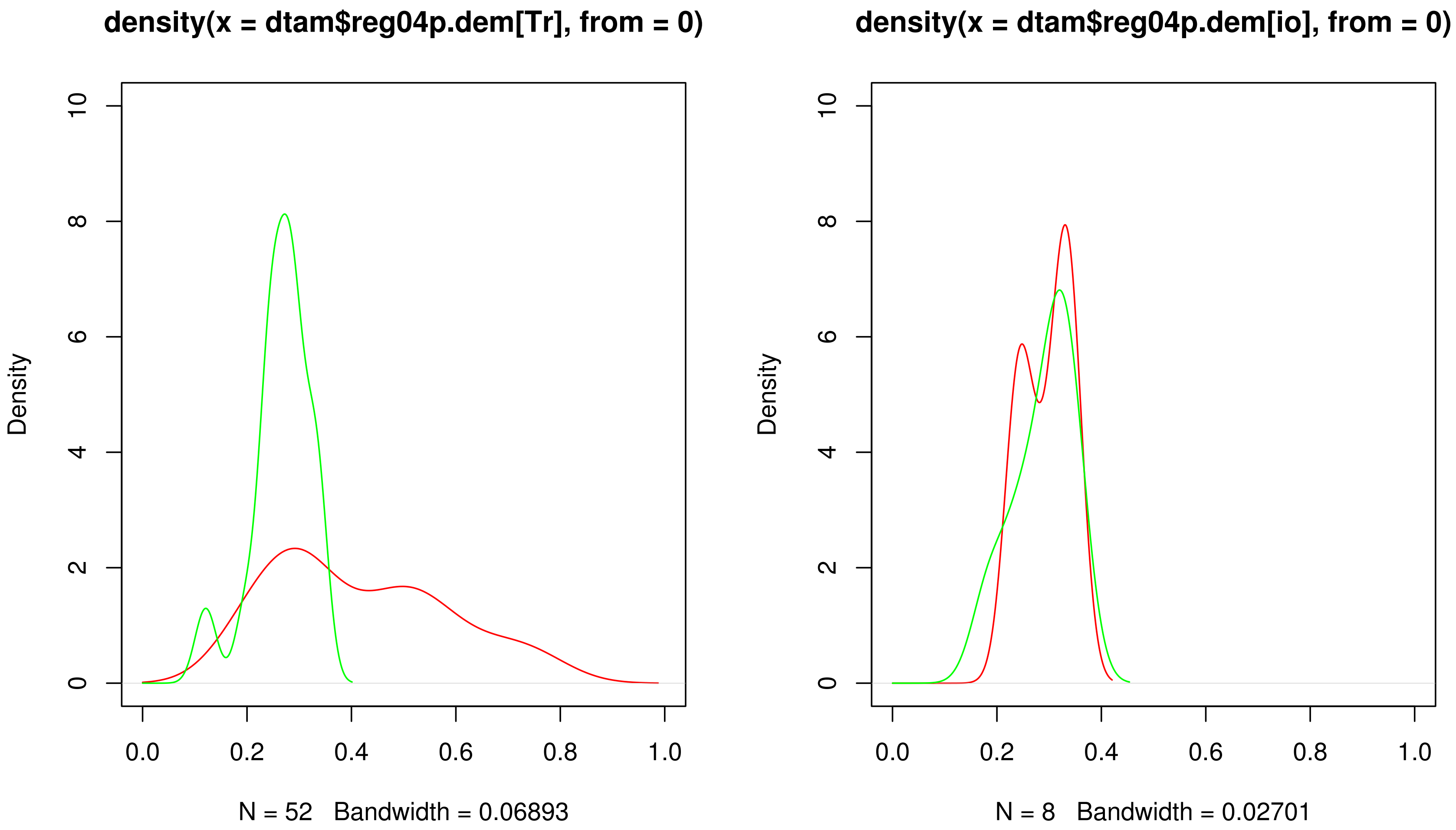
See http://jsekhon.fas.harvard.edu/papers/SekhonOpticalMatch.pdf for details

```
      Optical             DRE
[1,] "Bay"       "0.28" "Sumter"       "0.37"
[2,] "Citrus"    "0.43" "Pasco"        "0.45"
[3,] "Hernando"  "0.47" "Pasco"        "0.45"
[4,] "Manatee"   "0.43" "Lee"          "0.4"
[5,] "Marion"    "0.41" "Sumter"       "0.37"
[6,] "Orange"    "0.5"  "Hillsborough" "0.47"
[7,] "St. Johns" "0.31" "Martin"       "0.42"
[8,] "Walton"    "0.26" "Nassau"       "0.26"

Average Treatment Effect for the Treated
Estimate 0.00540
SE 0.0211
p-value 0.798

## balance on all other covariates, example:
                  optical dre    pvalue (t-test)
pre :  Dem Reg     .551    .361   .000
post:  Dem Reg     .369    .365   .556
```

Dem Registration Pre/Post Matching



Densities of democratic registration proportions by county. Green lines are DREs and red lines optical counties.

Mahalanobis Distance

- The most common method of multivariate matching is based on the Mahalanobis distance. The Mahalanobis distance measure between any two row vectors is defined as:

  $$\mathrm{md}(\mathrm{t},\mathrm{c}) \quad = \left\{(\mathrm{t}-\mathrm{c})'S^{-1}(\mathrm{t}-\mathrm{c})\right\}^{\frac{1}{2}}$$

  where t and c are the treatment and control values of the baseline covariates X for two different observations and S is the sample covariance matrix of X.

- Mahalanobis distance is an appropriate distance measure if each covariate has an elliptic distribution whose shape is common between treatment and control groups (Mitchell and Krzanowski 1985, 1989).

- Mahalanobis distance and propensity score estimation have excellent theoretical properties when the covariates have ellipsoidal distributions—e.g., such as normal and t. This is because, in such cases these matching methods have the Equal Percent Bias Reduction (EPBR) property.

Equal Percent Bias Reduction (EPBR)

- If X are distributed with ellipsoidal distributions, then the EPBR property holds for affinely invariant matching methods (e.g., propensity score, Mahalanobis Distance) (Rubin and Thomas 1992).

- An affinely invariant matching method is a matching method which produces the same matches if the covariates X are affinely transformed. An affine transformation is any transformation that preserves collinearity (i.e., all points lying on a line initially still lie on a line after transformation) and ratios of distances (e.g., the midpoint of a line segment remains the midpoint after transformation).

- Let Z be the expected value of X in the matched control group. Then we say that a matching procedure is EPBR if

  $$E(X|T=1) - Z \quad = \gamma \{E(X|T=1) - E(X|T=0)\}$$

  for a scalar $0 \leq \gamma \leq 1$.

- We say that a matching method is EPBR for X because the percent reduction in the mean biases for each of the matching variables is the same.

When is EPBR a Good Property?

- In general, if a matching method is not EPBR, than the bias for a particular linear function of X is increased.

- We may not want EPBR if we have some specific knowledge that one covariate is more important than another. For example,

$$Y = \alpha T + 2\ln(X_1^4) + 2X_2$$

  in this case we should be generally more concerned with X_1 than X_2.

- In finite samples, Mahalanobis distance and propensity score matching will not be optimal because X will not be ellipsoidally distributed in a finite sample even if that is its true distribution.

Genetic Matching (GenMatch)

- A more general way to measure distance is defined by:

$$d(t,c) \quad = \left\{ (t-c)' \left(S^{-1/2}\right)' W S^{-1/2} (t-c) \right\}^{\frac{1}{2}}$$

  where W is a $k \times k$ positive definite weight matrix and $S^{1/2}$ is the Cholesky decomposition of S which is the variance-covariance matrix of X.

- All elements of W are zero except down the main diagonal. The main diagonal consists of k parameters which must be chosen.

- This leaves the problem of choosing the free elements of W. For identification, there are only $k-1$ free parameters.

- See the paper entitled "Genetic Matching for Estimating Causal Effects: A New Method of Achieving Balance in Observational Studies" (Diamond and Sekhon 2005) for additional details: `http://jsekhon.fas.harvard.edu/papers/GenMatch.pdf`.

Parameterization

- GenMatch uses the propensity score if it is known or if it can be estimated.

- The propensity score is estimated and its linear predictor, $\hat{\mu}$, is matched upon along with the covariates X once they have been adjusted so as to be uncorrelated with the linear predictor.

- It is useful to combine the propensity score with Mahalanobis distance matching because propensity score matching is good at minimizing the discrepancy along the propensity score and Mahalanobis distance is good at minimizing the distance between individual coordinates of X (orthogonal to the propensity score) (Rosenbaum and Rubin 1985).

Optimization

- The algorithm attempts to minimize the largest discrepancy at every step.

- For a given set of matches resulting from a given W, the loss is defined as the minimum $\mathfrak{p}$-value observed across a series of balance tests.

- By default, tests are conducted for all univariate baseline covariates, as well as their first-order interactions and quadratic terms.

- The analyst may add tests of any function of X desired, including additional nonlinear functions and higher order interactions.

Tests

- The tests conducted are t-tests for the difference of means and nonparametric bootstrap Kolmogorov-Smirnov distributional tests.

- It is important the maximum discrepancy be small. p-values conventionally understood to signal balance (e.g., 0.10) are often too low to produce reliable estimates.

- After optimization, the p-values from these balance tests cannot be interpreted as true probabilities because of standard pre-test problems, but they remain useful measures of balance.

- The optimization problem described above is difficult and irregular, and we utilize an evolutionary algorithm developed by Mebane and Sekhon (1998) called GENOUD. The program is described in Sekhon and Mebane (1998) and available at: `http://jsekhon.fas.harvard.edu/rgenoud/`

Monte Carlo Experiments

- Two Monte Carlos are presented.

- MC 1: the experimental conditions satisfy assumptions for EPBR

  1. X covariates are distributed multivariate normal
  2. propensity score is correctly specified

- MC 2: the assumptions required for EPBR are **not** satisfied:

  1. X covariates are discrete and others are skewed and have point masses: they have the same distributions as the covariates of the Lalonde (1986) data.
  2. propensity score is incorrectly specified
  3. mapping from X to Y is nonlinear

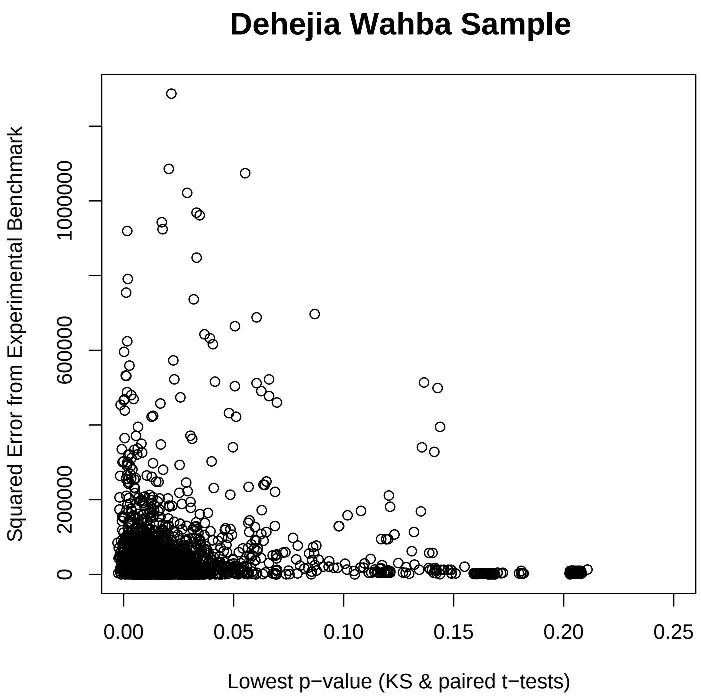
Experimental Condition 1: Multivariate Normal Distribution of Covariates

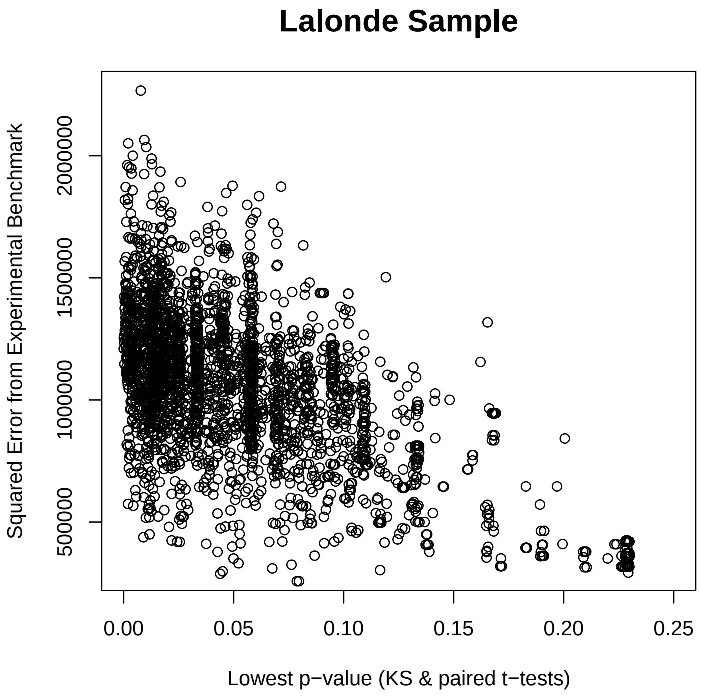
Estimator	Bias	RMSE	$\frac{\text{Bias}}{\text{Bias GM}}$	$\frac{\text{MSE}}{\text{MSE GM}}$
Raw	-604	.686	24.6	27.4
Mahalanobis (MH)	-8.63	.173	3.50	1.75
Pscore	-2.45	.210	.993	2.57
Pscore + MH	-5.96	.160	2.41	1.49
GenMatch	-2.47	.130		

Experimental Condition 2: Distribution of Lalonde Covariates

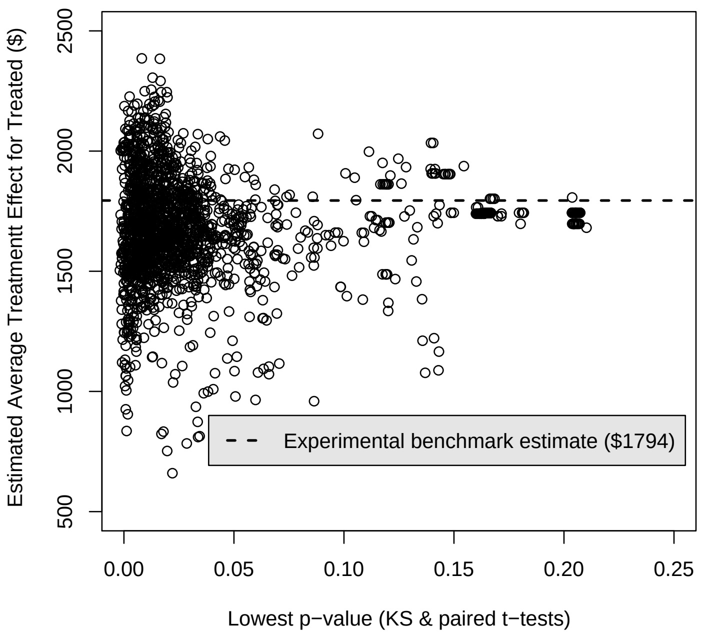
Estimator	Bias	RMSE	$\frac{\text{Bias}}{\text{Bias GM}}$	$\frac{\text{MSE}}{\text{MSE GM}}$
Raw	-485	1611	19.0	18.2
Mahalanobis (MH)	717	959	28.0	6.45
Pscore	-512	1294	20.0	11.7
Pscore + MH	-428	743	16.8	3.87
GenMatch	-25.6	378		

Squared Error Declines with Balance Achieved

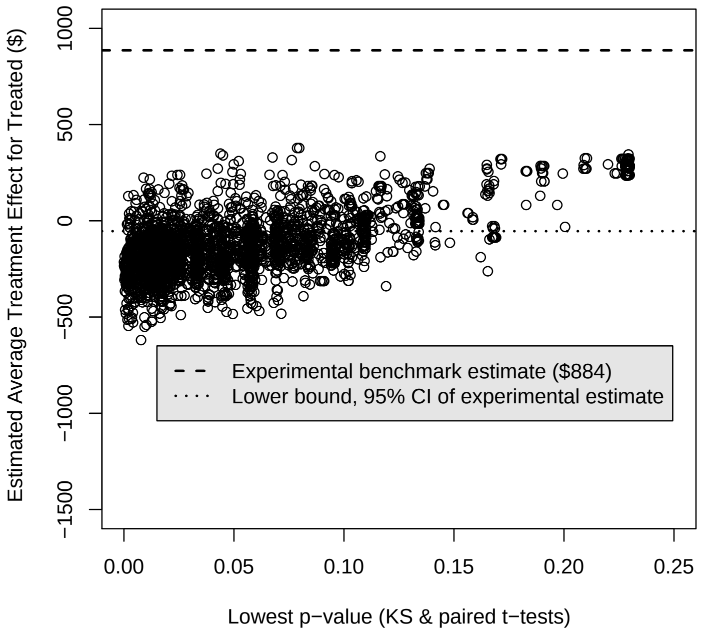




Reliable Estimates Require High Degree of Balance



Dehejia Wahba Sample



Lalonde Sample

Miscellaneous Material and Two Things You Should Remember

Outline:

1. ATE, ATT, ATC and experimental estimates
2. Presenting results
3. propensity score matching parameterization and balance tests (particularly the Kolmogorov-Smirnov Test).
4. Two Things You Should Remember

ATE, ATT, ATC

ATE (Average Treatment Effect), ATT (Average Treatment effect for the Treated) and ATC (Average Treatment effect for the Controls) are all simple to estimate when one has a valid experiment. Why is this?

Recall that with experimental data, the treatment and control groups are drawn from the same population. With observational data they (in general) are not. What about experiments will small samples?

ATT, ATT, ATC 2

Note from the ATT slide that ATT is:

$$\tau|(T=1) \quad = \quad E(Y_{i1}|T_i=1) - E(Y_{i0}|T_i=1)$$

ATT can then be estimated by

$$\begin{aligned}\tau|(T=1) \quad &= \quad E\{E(Y_i|X_i, T_i=1) - \\ &\qquad E(Y_i|X_i, T_i=0) \mid T_i=1\}\end{aligned}$$

where the outer expectation is taken over the distribution of $X_i|(T_i=1)$.

The definition of ATC is parallel to this.

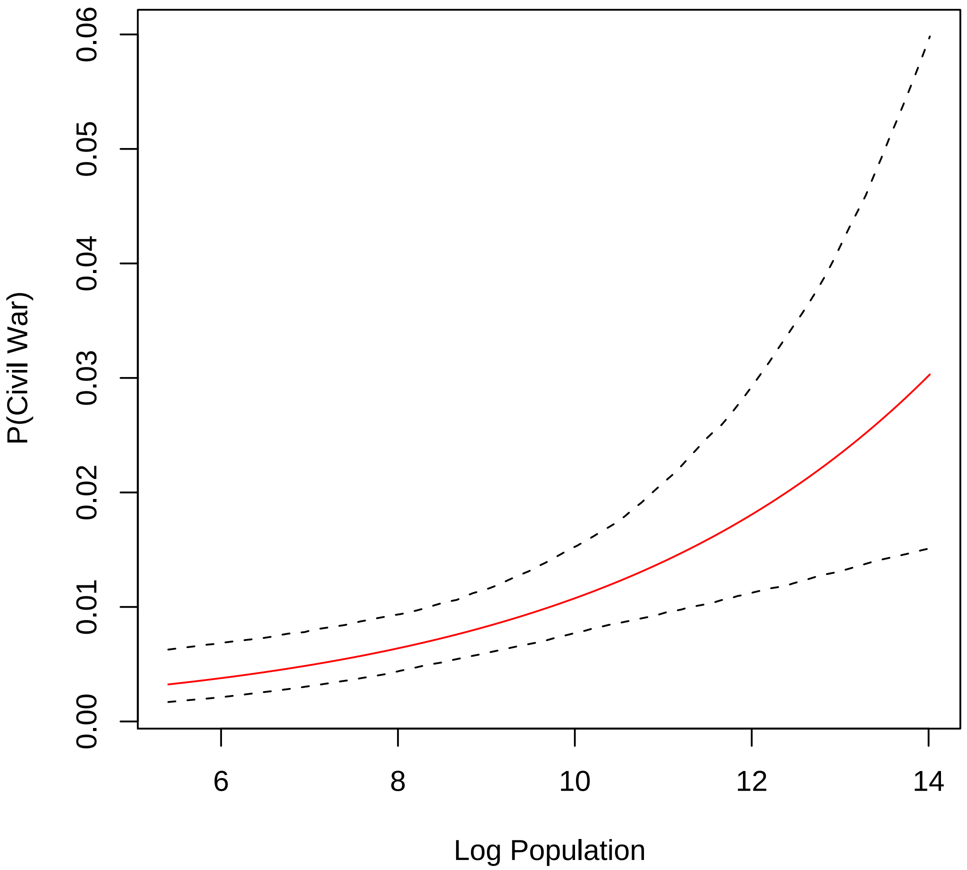
Presenting Results

Here are a number of things to remember when presenting results:

- Graphs are much better than tables
- Logit coefficients are not giving you the partial effect because the change in $\hat{Y}$ (the predicted probability) in a logistic model depends on the values of the other coefficients. This is because of the nonlinearity in the logit link. So, like in the home work. Graph the effect.

  For example, the coefficient for the effect of log population on the probability of civil war is 0.26294 in Model 1 of Fearon and Laitin. The standard error is 0.07271. What does this mean? A figure communicates much more information. Let's plot the effect along with 95% confidence bounds.

Fearon Laitin Model 1. Effect of Log Population on Civil War



More on Presentation

The confidence intervals were added to the previous figure using a parametric bootstrap.

For more useful tips on presentation see:

King, Gary; Michael Tomz; and Jason Wittenberg. “Making the Most of Statistical Analyses: Improving Interpretation and Presentation,” American Journal of Political Science, Vol. 44, No. 2 (April, 2000): 341-355. http://gking.harvard.edu/files/making.pdf

Propensity Score Parameterization

- If you do not use GenMatch, the best way to do matching is to use both the propensity score (pscore) and Mahalanobis Distance even if the pscore model isn't very good. See the GenMatch paper for details.

- There are a variety of ways of combining pscore and Mahalanobis Distance, the easiest it to add the pscore is another variable you match on.

- This is inefficient because we are using the same information many times: the information about some of the X covariates is in the pscore we have estimated. So, we want to adjust the X variables so they do not have redundant information.

- The propensity score is estimated and its linear predictor, $\hat{\mu}$, is matched upon along with the covariates X once they have been adjusted so as to be uncorrelated with the linear predictor.

- Adjustment is accomplished by regressing each covariate on the estimated linear predictor:

$$X_k = \hat{\alpha} + \hat{\mu} + \hat{\epsilon}_k$$

where k indexes the covariate number. By construction, $\mathrm{cor}(\hat{\epsilon}_k, \hat{\mu}) = 0$.

Balance Tests and the Propensity Score

Any given propensity score model may not be very good at balancing the X covariates. After matching on the pscore we need to test if the underlying X covariates have actually been balanced. You need to test more than just balance of the pscore.

The `MatchBalance()` function in Matching provides a number of tests, the two main ones being the t-test and the Kolmogorov-Smirnov Test—the KS test.

The KS test tries to determine if two distributions differ significantly. The KS-test has the advantage of making no assumption about the distribution of data—i.e., it is distribution free and non-parametric. However, because of this generality, other tests, such as the t-test, are more sensitive to certain differences—e.g., mean differences.

KS Test Details and Example

A great description of the standard KS test of offered at this webpage: `http://www.physics.csbsju.edu/stats/KS-test.html`.

But the standard KS test does not provide the correct p-value for noncontinous variables. But the bootstrap KS does provide the correct p-values. And additional problem arises if we want to test if distribution from estimated propensity scores differ. We then need to do another bootstrap to take into account the distributions of the parameters in the propensity model.

If You Only Remember Two Things:

- The definition of a causal effect which is based on the potential outcomes framework. This helps answer the questions I asked on the first day of class.

- Questions such as: “why do you believe in the effectiveness of chemotherapy for some obscure disease you have never heard of rather than key social science findings such as that government deficits *ceteris paribus* increase interest rates.”

- The other thing you should remember is that matching allows one to correct for confounding based on observables and to correct for it in a way which is testable. The specification tests for the usual methods are of limited help because they are not testing the biggest assumptions that is made: correct specification (usually linearity).

EXTRA MATERIAL

The material which follows is extra and requires knowledge of calculus.

[extra material] Derivation of Least-Squares Parameter Estimates

Our goal is to minimize $\sum_i^n (Y_i - \hat{Y}_i)^2$, where $\hat{Y}_i = \alpha + \beta X_i$ is the fitted value of Y_i corresponding to a particular observation X_i.

We minimize the expression by taking the partial derivatives with respect to α and β, setting each equal to 0, and solving the resulting pair of simultaneous equations:

$$\nabla_\alpha \sum_i^n (Y_i - \alpha - \beta X_i)^2 = -2 \sum_i^n (Y_i - \alpha - \beta X_i) \tag{21}$$

$$\nabla_\beta \sum_i^n (Y_i - \alpha - \beta X_i)^2 = -2 \sum_i^n X_i (Y_i - \alpha - \beta X_i) \tag{22}$$

Equating these two derivatives to zero and dividing by -2, we obtain:

$$\sum_i^n (Y_i - \alpha - \beta X_i) = 0 \tag{23}$$

$$\sum_i^n X_i (Y_i - \alpha - \beta X_i) = 0 \tag{24}$$

We may now rewrite these two equations to obtain what are called the **normal equations**:

$$\sum_{i}^{n} Y_i = an + \beta \sum_{i}^{n} X_i \tag{25}$$

$$\sum_{i}^{n} X_i Y_i = a \sum_{i}^{n} X_i + \beta \sum_{i}^{n} X_i^2 \tag{26}$$

We can solve for α and β simultaneously by multiplying Equation 25 by $\sum_i^n X_i$ and multiplying Equation 26 by n:

$$\sum_{i}^{n} X_i \sum_{i}^{n} Y_i = an \sum_{i}^{n} X_i + \beta (\sum_{i}^{n} X_i)^2 \tag{27}$$

$$n \sum_{i}^{n} X_i Y_i = an \sum_{i}^{n} X_i + \beta n \sum_{i}^{n} X_i^2 \tag{28}$$

Subtracting Equation 27 from Equation 28, we obtain

$$N \sum X_i Y_i - \sum_{i}^{n} X_i Y_i = \beta \left[n \sum_{i}^{n} X_i^2 - (\sum_{i}^{n} X_i)^2 \right] \tag{29}$$

It follows that:

$$\beta \frac{n \sum_i^n X_i Y_i - \sum_i^n X_i \sum_i^n}{n \sum_i^n X_i^2 - (\sum_i^n X_i)^2} \tag{30}$$

Given our solution of β, we may obtain our solution for α from Equation 25

$$\alpha = \frac{\sum_i^n Y_i}{n} - \beta \frac{\sum_i^n X_i}{n} \tag{31}$$

[extra material] Matrix Form Derivation

$$\mathfrak{min}(\beta \in \Re^k) E\left[Y_t - X_t'\beta\right]^2, \tag{32}$$

where X_t is a $k \times 1$ random vector, Y_t is a scalar random variable and β is $k \times 1$ vector.

First Order Condition (F.O.C.)

Let's set the gradient to zero:

$$\nabla_\beta E\left[(Y_t - X_t'\hat{\beta})^2\right] = 0 \tag{33}$$

We can solve for the gradient by interchanging the expectation and derivative operators and using the chain rule.

$$\begin{aligned}
\nabla_\beta E\left[(Y_t - X_t'\widehat{\beta})^2\right] &= E\left[\nabla_\beta (Y_t - X_t'\widehat{\beta})^2\right] && (34)\\
&= E\left[2(Y_t - X_t'\widehat{\beta})\nabla_\beta (Y_t - X_t'\widehat{\beta})'\right] && (35)\\
&= 2E\left[(Y_t - X_t'\widehat{\beta})\nabla_\beta (-X_t'\widehat{\beta})'\right] && (36)\\
&= -2E\left[(Y_t - X_t'\widehat{\beta})X_t'\right]. && (37)
\end{aligned}$$

Note that $\nabla_\beta(-X_t'\beta) = -X_t$.

Recall the Chain Rule

Let us be interested in:

$$
\begin{aligned}
h(x) &= f\left(g(x)\right) && (38)\\
\text{then,} & && (39)\\
\nabla_h(x) &= \nabla_f\left(g(x)\right)\nabla_g(x) && (40)\\
\text{Here is an example} & && (41)\\
h(x) &= (x^2+1)^3 && (42)\\
\text{Note that} & && (43)\\
f(x) &= x^3 && (44)\\
g(x) &= x^2+1 && (45)\\
\nabla_f(x) &= 3x^2 && (46)\\
\nabla_g(x) &= 2x && (47)\\
\text{Hence} & && (48)\\
\nabla_f\left(g(x)\right) &= 3(x^2+1)^2 && (49)\\
\nabla_h\left(x\right) &= 3(x^2+1)^2(2x) && (50)\\
& && (51)
\end{aligned}
$$

The FOC implies:

$$\mathsf{E}\left[\mathsf{X_t}(\mathsf{Y_t} - \mathsf{X_t'}\widehat{\beta})\right] = 0 \tag{52}$$

where $\widehat{\beta}$ is the value at the optimal point

$$\mathsf{E}(\mathsf{X_t Y_t}) - \mathsf{E}\left[\mathsf{X_t X_t'}\widehat{\beta}\right] = 0 \tag{53}$$

Since $\widehat{\beta}$ is considered to be a constant vector parameter,

we can move it outside of the expectation

$$\mathsf{E}(\mathsf{X_t Y_t}) = \mathsf{E}(\mathsf{X_t X_t'})\widehat{\beta} \tag{54}$$

Let us premultiply by $[\mathsf{E}(\mathsf{X_t X_t'})]^{-1}$

$$[\mathsf{E}(\mathsf{X_t X_t'})]^{-1}\,\mathsf{E}(\mathsf{X_t Y_t}) = \widehat{\beta} \tag{55}$$

We shall ignore the *Second Order Condition* (SOC).

Remarks:

1. The condition $\mathrm{E}(Y_t^2) < \infty$ ensures the existence of $\mathrm{E}(Y_t|X_t)$ and $\mathrm{E}\left[(Y_t - X_t')^2\right]$

2. Non-singularity of $\mathrm{E}(X_t X_t')$ ensures that

   (a) $\widehat{\beta}$ is unique
   (b) there is a global minimum

3. It **must** be remembered that, in general,
   $\mathrm{E}(Y_t|X_t) \neq X_t'\widehat{\beta}$

References

Abadie, Alberto and Guido Imbens. 2004. "Large Sample Properties of Matching Estimators for Average Treatment Effects." Working Paper.

Bartels, Larry M. 1996. "Uninformed Votes: Information Effects in Presidential Elections." *American Journal of Political Science* 40 (1): 194–230.

Berelson, Bernard R., Paul F. Lazarsfeld, and William N. McPhee. 1954. *Voting: A Study of Opinion Formation in a Presidential Campaign*. Chicago: University of Chicago Press.

Campbell, Angus, Philip E. Converse, Warren E. Miller, and Donald E. Stokes. 1960. *The American Voter*. New York: John Wiley & Sons.

Cox, David R. 1958. *Planning of Experiments*. New York: Wiley.

Diamond, Alexis and Jasjeet S. Sekhon. 2005. "Genetic Matching for Estimating Causal Effects: A New Method of Achieving Balance in Observational Studies." http://jsekhon.fas.harvard.edu/papers/GenMatch.pdf.

Fechner, Gustav Theodor. 1966 [1860]. *Elements of psychophysics, Vol 1.*. New York: Rinehart & Winston. Translated by Helmut E. Adler and edited by D.H. Howes and E.G. Boring.

Fishkin, James S. 1997. *The Voice of the People: Public Opinion and Democracy*. New Haven, CT: Yale University Press 2nd edition.

Hausman, Jerry A. and David A. Wise. 1979. “Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment.” *Econometrica* 47: 455–473.

Heckman, James J. 1979. “Sample Selection Bias as a Specification Error.” *Econometrica* 47: 153–161.

Heckman, James J., Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. “Characterizing Selection Bias Using Experimental Data.” *Econometrica* 66 (5): 1017–1098.

Lalonde, Robert. 1986. “Evaluating the Econometric Evaluations of Training Programs with Experimental Data.” *American Economic Review* 76 (September): 604–20.

Lupia, Arthur. 2004. “Questioning Our Competence: Tasks, Institutions, and the Limited Practical Relevance of Common Political Knowledge Measures.” Working Paper.

Luskin, Robert C., James S. Fishkin, and Roger Jowell. 2002. “Considered Opinions: Deliberative Polling in Britain.” *British Journal of Political Science* 32: 455–487.

McKelvey, Richard D. and Peter C. Ordeshook. 1985a. "Elections with Limited Information: A Fulfilled Expectations Model Using Contemporaneous Poll and Endorsement Data as Information Sources." *Journal of Economic Theory* 36: 55–85.

McKelvey, Richard D. and Peter C. Ordeshook. 1985b. "Sequential Elections with Limited Information." *American Journal of Political Science* 29 (3): 480–512.

McKelvey, Richard D. and Peter C. Ordeshook. 1986. "Information, Electoral Equilibria, and the Democratic Ideal." *Journal of Politics* 48 (4): 909–937.

Mebane, Walter R. Jr. and Jasjeet S. Sekhon. 1998. "GENetic Optimization Using Derivatives (GENOUD)." Software Package. http://jsekhon.fas.harvard.edu/rgenoud/.

Mitchell, Ann F. S. and Wojtek J. Krzanowski. 1985. "The Mahalanobis Distance and Elliptic Distributions." *Biometrika* 72 (2): 464–467.

Mitchell, Ann F. S. and Wojtek J. Krzanowski. 1989. "Amendments and Corrections: The Mahalanobis Distance and Elliptic Distributions." *Biometrika* 76 (2): 407.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

Rosenbaum, Paul R. and Donald B. Rubin. 1985. "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score." *The American Statistician* 39 (1): 33–38.

Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika* 63: 581–592.

Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *Annals of Statistics* 6 (1): 34–58.

Rubin, Donald B. and Neal Thomas. 1992. "Affinely Invariant Matching Methods with Ellipsoidal Distributions." *Annals of Statistics* 20 (2): 1079–1093.

Sekhon, Jasjeet S. 2004. "Quality Meets Quantity: Case Studies, Conditional Probability and Counterfactuals." *Perspectives on Politics* 2 (2): 281–293.

Sekhon, Jasjeet Singh and Walter R. Mebane, Jr. 1998. "Genetic Optimization Using Derivatives: Theory and Application to Nonlinear Models." *Political Analysis* 7: 189–203.

Sniderman, Paul M. 1993. "The New Look in Public Opinion Research." In Ada W. Finifter, editor, *Political Science: The State of the Discipline II* Washington, DC: American Political Science Association.

Zaller, John R. 1992. *The Nature and Origins of Mass Opinion*. New York: Cambridge University Press.