

Causal Inference

We've used *prediction* as a basis for model building:

choose a fitting routine to do the best job forecasting y at new \mathbf{x} drawn from the same distribution as data sample \mathbf{X} .

This exactly the question tackled by CV and AICc.

Today, we'll try to estimate the effect of a special covariate:

Treatment ' d ', which we can change *independently* from \mathbf{x} .

That is: we want to know the causal *treatment effect* (TE).

For example,

- ▶ $d = 1$ if you get the drug, $d = 0$ for placebo (control).
- ▶ d is some *policy tool*, like interest rates or product price.

Our treatment effect (TE) model looks like

$$\mathbb{E}[y|d, \mathbf{x}] = \alpha + d\gamma + \mathbf{x}'\beta$$

and we'll want to interpret *treatment effect* γ causally.

A coefficient is **structural** or **causal** if it is a *real-world* effect.

$\Rightarrow \gamma$ must represent change in y when d moves *independent* of any other influencers (both in \mathbf{x} or those we've omitted).

In contrast, a **reduced form** model is a simple representation for the effect of \mathbf{x} on y , without worrying about causal mechanisms.

e.g., in reduced form, β_j is the effect of x_j , but this could be due to x_j 's correlation with other variables that actually cause y .

Randomized Control Trials: **A/B experiments**

We want to know the effect on y of change in d independent of change in \mathbf{x} .

A *completely randomized design* draws two random samples of units, then applies $d = 0$ to one sample and $d = 1$ to the other.

For example, you randomize your website visitors into groups 'A' (control) and 'B' (treatment). Those in A see the current website, while those in B see a new layout.

Say y (spend or clicks) is the response of interest.

TE is treatment mean minus control mean: $\hat{\gamma} = \bar{y}_B - \bar{y}_A$

$\text{se}(\hat{\gamma}) = \sqrt{SST_A/n_A^2 + SST_B/n_B^2}$ and this is just stats 101.

◆ RCT and Sequential Design

RCT (A/B) experiments are hugely popular and hugely useful.

If you have the ability to randomize, it is very tough to find a TE estimate that is much better than $\bar{y}_B - \bar{y}_A$ from an experiment. Indeed: beware of those who claim otherwise.

However, it's century-old tech and sometimes we can do better. This is especially true if:

- ▶ You view TE $\gamma(\mathbf{x})$ as a function of covariates.
- ▶ You have many treatments to choose amongst.

In each case, if you are accumulating data over time you can use existing observations to guide where you sample next.

This is called **active learning** (AL).

◆ Active Learning

Say you have $j = 1 \dots J$ different 'ad campaigns' to show.

As each user i comes, you can show them one ad: $d_i = j$.

Your goal is to maximize ad-clicks.

Say $\mathbf{c}_n = [c_{n1} \dots c_{nJ}]$ are the clicks on each ad after n users, and $\mathbf{s}_n = [s_{n1} \dots s_{nJ}]$ are times each ad has been shown.

To find the best ad as quickly as possible you can sample *click probabilities* for each ad as $q_{nj} \sim \text{Beta}(c_{nj}, s_{nj} - c_{nj})$.

Beta has mean c_{nj}/s_{nj} here, with variance that shrinks with s_{nj}

The result is that you try (explore) all the ads, but show the ones that seem to be working more than the others.

Run `mab.R` to see this in action.

◆ Bandits and more

The prev slide's setup is called a **multi-armed bandit**.

The learning algorithm is called **thompson sampling**.

Active learning is a big area. It gets tricky once covariates are involved (e.g., click probs are functions of user attributes.)

With the trend of *site personalization*, more and more datasets for online behaviour are coming from some AL scheme.

AL is just one example of how the data you'll have to analyze can be much more complex than that from an A/B experiment.

In many cases, you don't get to experiment at all!

To figure out causal effects in more complex setups, we need to go back in time and take a look at some linear models.

Blocked-Design Experiment

Ideas born in agriculture: “does this fertilizer work?”
TE can get swamped by variation in growing conditions.

Pick *blocks* of nearby fields and split them.

1:

$d = 0$	$d = 1$
---------	---------

 2:

$d = 0$
$d = 1$

 3:

$d = 1$	$d = 0$
---------	---------

 4:

$d = 1$
$d = 0$

Response y_{kd} is some measure of yield (e.g., kg of rice).

The estimated *treatment effect* is $\hat{\gamma} = \frac{1}{4} \sum_{k=1}^4 (y_{k1} - y_{k0})$

Blocked-Design Experiment

Re-write the TE model as a regression

$$\mathbb{E}[y|d, k] = \alpha_k + d\gamma$$

where α_k is the intercept for block k .

$\hat{\gamma}$ for this regression is $\frac{1}{4} \sum_{k=1}^4 (y_{k1} - y_{k0})$, as on prev slide.

Here, we interpret γ as a *causal effect* (not just correlation) because treatment $d = 0/1$ is independent of the covariates. (in this case, 'covariates' = block membership factor variables).




We know this because that's how the experiment was designed.

Big lesson: we can use regression to analyze experiments!

SEM Experiment example

What is the effect of *paid search advertising*?


Or, if we turned it off and went organic, what would happen?




[Web](#) [Shopping](#) [Images](#) [Maps](#) [Videos](#) [More ▾](#) [Search tools](#)

About 83,000,000 results (0.29 seconds)


Shop for toddler shoes on Google




Stride Rite
Crawl Bonnie...
\$17.99
[Diapers.com](#)




Nike Kids Free
Run 5.0 (TDV...
\$24.99
[6pm.com](#)



Vans Authentic
Sneaker Pre...
\$28.00
[Shoes.com](#)
Special offer



OshKosh
B'gosh Orbit...
\$20.00
[Famous Foot...](#)
23% price drop



Saucony
Crossfire Tod...
\$32.00
[Shoes.com](#)

Ads ⓘ
Toddler Shoes - JCPenney®
[www.jcpenny.com/ToddlerShoes](#) ▾
Save on **Toddler Shoes** at JCPenney®.
JCPenney Offers Free Ship Over \$99.

Toddler's Shoes at Sears®
[www.sears.com/Free-Store-Pickup](#) ▾
4.5 ★★★★★ rating for sears.com
Save Big on **Shoes** and Get them
Today with Free Store Pickup!
📍 6153 S Western Ave, Chicago, IL
(773) 918-1400

Toddler Shoes
[www.eastbay.com/](#) ▾
Shop the Official Eastbay Store for
Brand Name Footwear, Apparel & More

Ecco Shoes Toddler
[www.shoebuy.com/](#) ▾
4.6 ★★★★★ rating for shoebuy.com
Free Shipping & Free Returns.
No Tax and a 100% price guarantee!

Toddler Squeaky Shoes
[www.squeakyshoestore.com/](#) ▾

Toddler Shoes
Ad [www.zappos.com/Toddler-Boots](#) ▾
Shop The Latest **Toddler Shoes** Enjoy Free Shipping & Free Returns!
Zappos.com has 53,726 followers on Google+

Stride Rite® for Toddlers
Ad [www.striderite.com/](#) ▾
Buy One-Get One 50% Off **Sneakers**. Free Shipping on Orders Over \$75.
Toddler Shoes - Sale & Clearance - Buy One, Get One 50% Off - Kids Shoes
📍 7501 W. CERMAK RD. RM.12, NO. RIVERSIDE, IL - (708) 442-7079

Toddler Shoes | Shipped Free at Zappos - Zappos.com
[www.zappos.com/toddler-shoes](#) ▾ [Zappos](#) ▾
6222 items - Free shipping BOTH ways on **toddler shoes**, from our vast selection of
child. Fast delivery and 24/7/365 real person service with a smile. Click or call

11

Correcting for **incomplete** randomization

eBay did a big experiment to test paid search (aka, SEM)

Blake + Nosko + Tadelis: Paid Search Effectiveness.

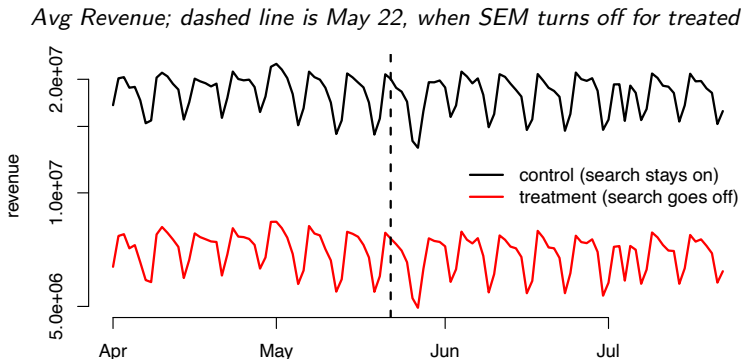
They *turned off* paid search (stopped bidding on any AdWords) for 65 of the 210 'Designated Market Areas' (DMA) in the US.

(Google guesses the DMA for a browser.)

In 2012, eBay recorded revenue in all DMAs for ≈ 8 weeks before and after turning off SEM for the treated 65 on May 22.

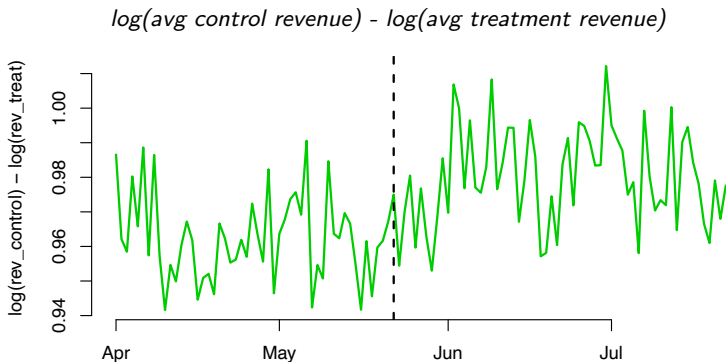
Data are in `paidsearch.csv`, and code is `paidsearch.R`. Note that this is not the real data; it's been scaled and shifted.

Problem: the treated DMAs are not a random sample.
They avoided, e.g., the largest markets.



If you just look at $\bar{y}_B - \bar{y}_A$, you see a big difference even before SEM turns off (i.e., before B is *treated*). This can't be causal!

If SEM works, then the revenue difference should be *larger after search is turned off for treatment DMAs* (after May 22). We'll look at differences in $\log(\text{revenue})$.



Maybe there is some increase. Is it real?

This is a more complicated blocking structure than before.
But we can still write out a similar regression model.

Say you have DMA i at time $t \in 0, 1$, before or after May 22, and $d_i = 1$ if DMA i is in treatment group, 0 otherwise.

Response y_{it} is the log average revenue for i during t .

Our *blocks* are the DMAs. These are like the rice fields. But instead of being divided randomly, they always split on May 22.

To tell the effect of SEM apart from that of 'time', we ask if d_i changes the effect of t : does $y_{i1} - y_{i0}$ depend upon d_i ?

Writing this out as a regression:

$$\mathbb{E}[y_{it}] = \alpha_i + t\beta_t + \gamma d_i t$$

the TE is an interaction term!

We can run the regression in R:

```
> semreg <- glm(y ~ dma + d*t-d, data=semavg)
> summary(semreg)$coef["d:t",]
      Estimate   Std. Error    t value    Pr(>|t|)
-0.006586852  0.005571899 -1.18215571  0.238493640
```

The direction is right, but it is nowhere close to significant.
See end of paidsearch.R for more intuition behind the model.

- ★ Unlike most cases, this p-value answers exactly our question:
Do we need $\gamma \neq 0$ given all other variables in the model?
- ★ Assumption for causation: *outside influences on revenue post May 22 affect treatment and control equally (i.e., are in β_t).*

NB: this is a version of what economists call 'diff in diff'.

From Experiments to **Observational Studies**

We've defined the treatment effect as change in y when d moves **independent** of all other influencers. This means that the TE represents what will happen if we move d ourselves.

This is easy to measure in a fully randomized experiment, because d is independent by design: we sample it randomly.

Under partial randomization, things remains straightforward. We know exactly which variables were not randomized over (e.g., time for SEM) and can *control* for them in regression.

Confounders and Control

The idea behind causal inference is to remove from $\hat{\gamma}$ the effect of any other influences that are correlated with d .

These influences are called 'controls' or 'confounders'.

They are variables whose effect can be confused with that of d .

Remove confounders from $\hat{\gamma}$ by *including them in regression*.

We say then that we have 'controlled for' confounding effects.

Or: "removed the effect of \mathbf{x} ", "partial effect of d given \mathbf{x} ".

For example, in the SEM study time t was correlated with treatment d because we didn't randomize treatment periods. We *controlled* for time by including it in the regression.

Experimental study: you are able to randomize treatment.

Observational study: you just observe what nature provides.

Causal TE inference requires you to control for (include in your model) all influences on y over which you have not *randomized* (i.e., those which are correlated with d).

Without an experiment, we haven't randomized over anything: with **observational** data, you need to control for 'everything'!

This is the toughest game in statistics.

In a very real sense, it is actually impossible.

But we'll take a look at how to do our best.

How does controlling for confounders work?

With \mathbf{x} in the regression model, inference for γ is measured from the effect of *the bit of \mathbf{d} that is not predictable by \mathbf{x}* .

e.g., say $d(\mathbf{x}) = \mathbf{x}'\boldsymbol{\tau} + \nu$, where ν is random noise (residual).

$$\begin{aligned}\text{Then: } \mathbb{E}[y|\mathbf{x}, d] &= d\gamma + \mathbf{x}'\boldsymbol{\beta} \\ &= (\mathbf{x}'\boldsymbol{\tau} + \nu)\gamma + \mathbf{x}'\boldsymbol{\beta} \\ &= \nu\gamma + \mathbf{x}'(\gamma\boldsymbol{\tau} + \boldsymbol{\beta}) \approx \nu\hat{\gamma} + \mathbf{x}'\hat{\boldsymbol{\beta}}\end{aligned}$$

So $\hat{\gamma}$ is *identified* as the effect of ν , the independent part of d .

This type of controlling is simple with low-D \mathbf{x} : just fit the MLE regression and your standard errors on $\hat{\gamma}$ should be correct.

Freakonomics: Abortion and Crime



Donahoe and Levitt (DL) argue a controversial thesis:

easier access to abortion
causes decreased crime.

Proposed mechanism holds that birth is postponed until the mother is more ready.

They assume stable family \Rightarrow better upbringing \Rightarrow less crime.

There's obviously no experiment here.

How have they controlled for confounders?

Crime ~ Abortion regression

The treatment variable d is by-state abortion rate,
and for response we look at $y = \text{murder rate}$.

DL *control* for bunch of state-specific *confounders*: income,
poverty, child tax credits, weapons laws, beer consumption...

They also include state effects (factor ' s ') and a time trend
(numeric ' t ') to control for missed confounders.

```
> orig = glm(y ~ d + t + s + ., data=controls)
> summary(orig)$coef['d',]
```

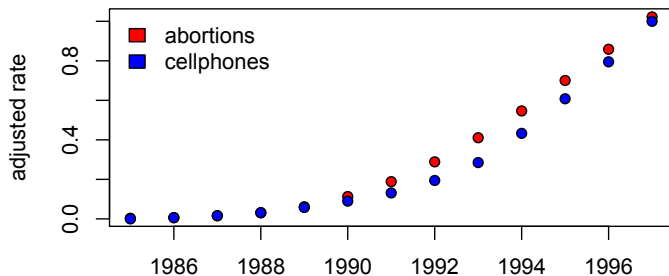
Estimate	Std. Error	t value	Pr(> t)
-2.098119e-01	4.109177e-02	-5.105936e+00	4.505925e-07

Abortion has a very significant effect! Skeptical? You should be.

Alternative story: **Cellphones and Murder**

Technology has contributed to lower murder rates, and we'll add cellphone subscribers as a variable to control for tech progress.

e.g., Cellphones lead to faster ambulances, our jobs are gentile so we're less rough, more communication increases empathy, or we don't interact in-person because we don't have to.



Abortion and cellphones move together...

```
> tech = glm(y ~ phone + t + s + ., data=controls)
> summary(tech)$coef[c('phone'),]
      Estimate      Std. Error      t value      Pr(>|t|)
-3.662545e-01  6.779352e-02 -5.402500e+00  9.699480e-08
```

Cellphones have an even more significant effect!

It took me about 10 minutes to dream up a causal variable and grab data off of wikipedia.

What is happening is that murder decreased quadratically, and we have no controls that also moved this way.

How can we be sure the abortion effect is not just a stand-in for another cause that changed quadratically over the years?

To *control* for such confounders, we just add t^2 to the model.

We should also allow confounder effects to interact with each other (e.g., different gun effect for high beer) and with time.

```
> interact <- glm(y ~ d + (s + .^2)*(t+t2), data=cntrl)s)
> summary(interact))$coef['d',]
      Estimate Std. Error    t value    Pr(>|t|)
0.8261849    0.7367764    1.1213508    0.2629207
```

Significance disappears.

This happens because we've added so many variables that there is not enough data to say anything is significant.

```
> dim(model.matrix(y ~ d + (s + .^2)*(t+t2), data=cntrl)s))
[1] 624 280
```

The authors can't be expected to fully control for every crazy story.

Multicollinearity and the lasso

MLE treatment effect estimation fails if you have too many controls. *But can't we just throw everything in a lasso?*

Not exactly.

Even if all possible influencers are in \mathbf{x} , the lasso won't always choose the right ones to remove confounding effects.

In our earlier treatment effect equations, we could have also written $\mathbf{x} = \varphi d + v$, so that \mathbf{x} is now a function of d .

$$\text{Then: } \mathbb{E}[y|\mathbf{x}, d] = d\gamma + \mathbf{x}'\beta = d(\gamma + \varphi'\beta) + v\beta$$

Since the lasso makes you pay a price for every extra nonzero coefficient, it'll choose to just collapse the effects of \mathbf{x} (β) into $\hat{\gamma}$ unless v has a big enough effect to warrant the extra cost.

Treatment Effects with High Dimensional Controls

We want to use our model selection tools to help estimate γ in $\mathbb{E}[y|\mathbf{x}, d] = d\gamma + \mathbf{x}'\beta$ when \mathbf{x} is high dimensional.

But we need to avoid confusing $\hat{\gamma}$ with β .

We need to do variable selection in a way that still allows us to control for confounding variables.

It is all about prediction: we want to forecast y for new random \mathbf{x} but where d *changes independently from \mathbf{x}* .

Causal Lasso

We have $d = d(\mathbf{x}) + \nu$, and we want the effect of ν .

So estimate $\hat{d}(\mathbf{x})$ directly and include it in the regression!

Any left over effect of d will be attributable to $d - \hat{d}(\mathbf{x}) = \nu$.

$$\mathbb{E}[y|\mathbf{x}] = (\hat{d}(\mathbf{x}) + \nu)\gamma + \hat{d}(\mathbf{x})\delta + \mathbf{x}'\beta = \nu\gamma + \hat{d}(\mathbf{x})(\gamma + \delta) + \mathbf{x}'\beta$$

Controlling for $\hat{d}(\mathbf{x})$ in regression is equivalent to estimating $\hat{\gamma}$ as the effect of ν : *the independent part of d* .

A Treatment Effects Lasso

Two stages:

1. Estimate $\hat{d}(\mathbf{x})$ with lasso regression of d on \mathbf{x} .
2. Do a lasso of y on $[d, \hat{d}(\mathbf{x}), \mathbf{x}]$, with $\hat{d}(\mathbf{x})$ *unpenalized*.

Including \hat{d} unpenalized in [2] ensures that confounder effects on d have been removed: thus $\hat{\gamma}$ measures the effect of ν .

In [2], we can apply our usual AICc lasso to see what else in \mathbf{x} effects y and, most importantly, if $\hat{\gamma} \neq 0$.

We've replaced causal estimation with two prediction problems. And prediction is something we're really good at, even in HD.

In the end, we're asking: is ν *useful* for predicting y ?

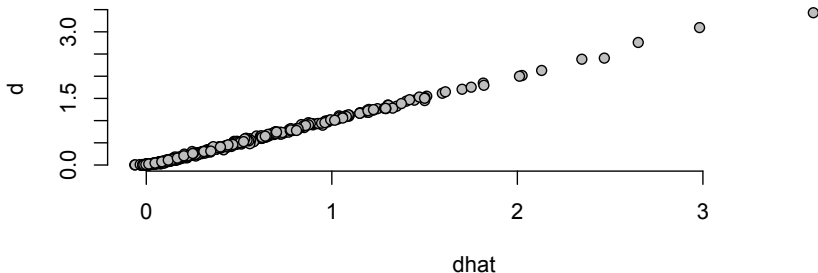
Back to abortion

If you just run a straight lasso onto d and x ,
AICc selects a significant negative abortion effect.

```
> naive <- gamlr(cBind(d,x),y)
> coef(naive)["d",]
[1] -0.09005649
```

But AICc lasso selects d as highly predicted by \mathbf{x} .

```
treat <- gamlr(x,d,lambda.min.ratio=1e-4)
dhat <- predict(treat, x, type="response")
```



In Sample R^2 is $>99\%$.

So there's almost no *independent* movement of abortion rates to measure as effecting crime (it's not much of an experiment).

Sure enough, if you include d_{hat} in lasso regression then AICc says there is no residual effect for d .

```
## free=2 here leaves dhat unpenalized  
> causal <- gamlr(cBind(d,dhat,x),y,free=2)  
> coef(causal)["d",]  
[1] 0
```

Summary: causation via two prediction models.

- ▶ Fit \hat{d} : your best predictor for d from \mathbf{x} .
- ▶ Find the best predictor for y from d and \mathbf{x} ,
after influence of \hat{d} is removed (i.e., predict y from ν and \mathbf{x}).

Then $\hat{\gamma}$ predicts what will happen if we change d independently.

Observational Study Wrap-Up

Science is hard. Keep theorizing, but hit ideas with data.

We can't say that abortion *does not* lower crime.

We just have nothing that looks like an experiment here.

And an experiment (or something that looks like one)
is what you need to estimate TEs.

Our double-lasso is one [good] way to sort out causation.

But this is a huge area, and there are *many* strategies: matching,
instrumental variables, double robust, post lasso...

Always ask yourself:

- ▶ How well would my model predict if I change d arbitrarily?
- ▶ How am I replicating what I'd get from a real experiment?