Regression Lectures 8-10

Alex Stephenson

9-13-2021

Announcements

There **is** a checkpoint due this week. All of the answers for it can be found in the Mixtape Chapter 2 Reading.

There **is not** a Weekly Practice due this week

PS1 **is** due this week

My OH this week and going forward are from 15:10-17:00 in SSB 394. Making an appointment at a different time can be done via email.

What did we do last time?

Covered Blocking and Clustering

What are we doing today?

Regression

What's the big deal about Regression?

Regression is a computational device for the estimation of an effect.

With appropriate assumptions, this estimate has a causal interpretation.

The second fact (often misused) makes regression the most common estimator into social science research

Regression has close links to other strategies for controlling for confounders

Conditional Expectation Function

The CEF is a function that characterizes all possible values of $E[Y|X = x]$

Due to the Law of Iterated Expectations $E[Y] = E[E[Y|X]]$ the unconditional expectation can be expressed as a weighted average of conditional expectations.

The CEF is the Best Predictor of $E[Y|X]$

Best Linear Prediction

For random variables X,Y, assuming $V[X] > 0$ the Best linear predictor of Y given X is

$$g(X) = \alpha + \beta X$$

where:

$$\alpha = E[Y] - \frac{Cov[X,Y]}{V[X]}E[X]$$

$$\beta = \frac{Cov[X,Y]}{V[X]}$$

The BLP is also the Best Linear Approximation of the CEF. When the CEF is linear then BLP is the CEF

Implications of Independence of Random Variables

If X and Y are independent:

1. $E[Y|X] = E[Y]$
2. $V[Y|X] = V[Y]$
3. The BLP of Y|X = E[Y]

Regression is a plug-in Estimator for the BLP

The BLP is $E[Y|X] = \alpha + \beta X$. We can rewrite the Greek terms to be:

$$\alpha = E[Y] - \frac{E[XY] - E[X]E[Y]}{E[X^2] - E[X]^2} E[X]$$

$$\beta = \frac{E[XY] - E[X]E[Y]}{E[X^2] - E[X]^2}$$

Regression is a plug-in Estimator for the BLP

We can *estimate* the BLP using plug-in estimation with sample data

$$\hat{\alpha} = \bar{Y} - \frac{\bar{XY} - \bar{X}\bar{Y}}{\bar{X^2} - \bar{X}^2}\bar{X}$$

$$\hat{\beta} = \frac{\bar{XY} - \bar{X}\bar{Y}}{\bar{X^2} - \bar{X}^2}$$

Both of these estimators are consistent estimators for the parameters of interest

Regression and Potential Outcomes

Consider the PO Model $Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$

$D_i \in \{0, 1\}$

Assume constant treatment effects. We can rewrite as

$$Y_i = \alpha + \beta D_i + \epsilon$$

Where: $\alpha = E[Y_{0i}]$, $\beta = Y_{1i} - Y_{0i}$, and $\epsilon = Y_{0i} - E[Y_{0i}]$

Regression and Potential Outcomes

Consider the conditional expectations at each value of D

$$E[Y_i|D_i = 1] = \alpha + \beta + E[\epsilon|D_i = 1]$$

$$E[Y_i|D_i = 0] = \alpha + E[\epsilon|D_i = 0]$$

Which implies

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = \beta + E[\epsilon|D_i = 1] - E[\epsilon|D_i = 0]$$

where:

$$E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]$$

Summary of Regression and the CEF

1. If the CEF is linear, then our best estimate of the BLP will be the CEF. Regression serves as a plug in estimator for the BLP
2. If the CEF is nonlinear, then our best linear approximation is the BLP. Regression serves as a plug in estimator for the BLP.
3. If we have a constant ATE (and in fact if we do not but can model it), regression provides an unbiased estimator of the treatment effect.

Regression in an Experimental Context

Suppose we now have a pre-treatment covariate X. We can write our model as:

$$Y_i = Y_i(1)D_i + (1 - D_i)Y_i(0)$$
$$Y_i - \alpha + \beta D_i + \gamma X_i + (\epsilon_i - \gamma X_i)$$

Here the error term is the last term in parenthesis. Note that we are adding and subtracting γX_i in this equation.

OLS Regression Estimator

For iid random vectors $(Y_i, \mathbf{X}_i)$ the OLS regression estimator is the function:

$$\hat{g}(\mathbf{X}) = \beta X$$

s.t.

$$\hat{\beta} = argmin \frac{1}{n} \sum_{i}^{n} (Y_i - (b_0 + \ldots + b_k X_{ki}))^2$$

OLS Estimator in Matrix Form (A Preview)

Commit the following to memory:

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

This is the solution to the least squares problem estimated with OLS. A requirement is that $(X'X)$ is invertible, which occurs when that matrix has full rank.

Lecture 9

What did we do last time?

Discussed the justification for using regression

What are we doing today?

Reminder that Checkpoint 4 and PS1 are due this Friday at 10am

OLS in Matrix Form

Standard Errors for inference

OLS Assumptions

In order for OLS to be considered an estimation of causal effects, we require the following assumptions.

1. The Population Regression Function is linear in parameters or approximated as such.
2. We have a random sample of data, or a sample that can be interpreted as such.
3. The variance of X exists and is not infinite.
4. The Zero conditional mean assumption

Examples of OLS Regression

Shoub *et al.* (2021) "Do Female Officers Police Differently? Evidence from Traffic Stops"

This paper is interested in the question of whether there is an effect of a police officer being a woman on traffic stop outcomes.

The authors collect data on traffic stops from the Charlotte Police Department and the Florida Highway Patrol

Examples of OLS Regression

TABLE 2 OLS Regressions Explaining Searches Following a Traffic Stop

	CPD	FHP
Female officer	−0.026*	−0.004*
	(0.002)	(<0.001)
Intercept	0.086*	0.026*
	(0.004)	(0.001)
Controls	Yes	Yes
Year fixed effects	Yes	Yes
Division fixed effects	Yes	No
County fixed effects	No	Yes
R^2	0.071	0.009
Adjusted R^2	0.071	0.009
N	150,547	2,712,478

Note: Each observation is an individual traffic stop.
*$p < .05$.

In general:

The primary coefficient of interest should be listed first.

Control variables should not be included unless directly relevant.

The number of observations should be included.

Standard errors of the estimates should be included.

Conditional Independence Assumption

In words, conditional on observed characteristics selection bias disappears.

$$\{Y_{0i}, Y_{1i}\} \perp D_i | X_i$$

This holds whenever D_i is randomly assigned conditional on X_i

Often referenced as "Selection on Observables"

Matrix Algebra Review

A vector

```
vector <- c(1,2,3,4)
```

A matrix

```
mat <- matrix(1:9, nrow = 3, ncol = 3)
```

Matrix Multiplication

```
mat2 <- matrix(10:18, nrow = 3, ncol = 3)
mat %*% mat2
```

Matrix Algebra Review

Matrix Transpose

```
t(mat)
```

```
##      [,1] [,2] [,3]
## [1,]    1    2    3
## [2,]    4    5    6
## [3,]    7    8    9
```

Inverting a Matrix

```
vec1 <- sample(c(rep(0,50),
                 rep(1,50)),
               100,
               replace = F)
vec2 <- runif(100, 0,1)
m <- cbind(vec1, vec2)
solve(t(m)%*%m)
```

```
##              vec1        vec2
## vec1  0.03182076 -0.02315038
## vec2 -0.02315038  0.04533887
```

Regression with Matrix Algebra

As mentioned last time the canonical closed form OLS solution is

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

as long as $X'X$ is invertible. Mechanically, the following are true:

1. $\bar{e} = 0$
2. $\overline{eX_k} = 0, \forall k \in \{1, \dots, K\}$
3. $\overline{eX_k} - \bar{e}\overline{X_k} = 0, \forall k \in \{1, \dots, K\}$

The Model Matrix

For i.i.d random vectors $(Y_i, \mathbf{X}_i), i \in \{1, \ldots, n\}$ The design matrix is

$$X = \begin{pmatrix} 1 & X_{11} & \ldots & X_{k1} \\ 1 & X_{12} & \ldots & X_{k2} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & X_{in} & \ldots & X_{kn} \end{pmatrix}$$

Regression with Matrix Algebra

```
# Beta estimates B = (X'X)^-1 X'Y
beta_est <- function(data, x, y){
  d <- data[complete.cases(data[, c(y,x)]),]

  y <- as.matrix(d[[y]])
  x <- cbind(rep(1, nrow(d)),as.matrix(d[[x]]))
  A <- solve(t(x)%*%x)

  # Beta coefficient estimates
  b <- A %*% t(x) %*% y

  return(b)
}
```

Robust Standard Errors for OLS

1. Heteroskedastic errors are the norm and should be assumed for any application
2. Without homoskedasticity OLS no longer is BLUE, but this isn't a big problem
3. We can have a valid estimator of the variance of our OLS estimators that deals with heteroskedasticity of any form.
4. For each coefficient it is possible to derive an estimate of the standard error $\sqrt{\hat{V}[\hat{\beta}_k]}$

Robust Standard Errors

Assume that $\epsilon = Y - X\beta$:

Decompose $\hat{\beta}$ as follows

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

Robust Standard Errors

The robust sampling variance estimator for $\hat{\beta}$ is:

$$\hat{V}[\hat{\beta}] = (X'X)^{-1}X'diag(e_1^2, \ldots e_n^2)X(X'X)^{-1}$$

The middle part is the n by n matrix whose i^{th} diagonal element is e_i^2 and whose off-diagonal elements are all zero.

Here $\hat{V}[\hat{\beta}]$ is the covariance matrix where the variances are on the diagonal and the covariances are on the off diagonals.

Robust Standard Errors and Causal Inference

First the takeaway, then the math:

In the Neyman model, the value of the error term depends on the realization of treatment status.

Since the value of the error term depends on the realization of treatment status, the variance of the error term depends on the realization of treatment status.

The implication is that standard errors are never going to be homoskedastic, and so we should always presume standard errors to heteroskedastic

Robust Standard Errors and Causal Inference

Now the math.

Let $\beta = \overline{Y}_1 - \overline{Y}_0$ be the ATE and $\alpha = \overline{Y}_0$ be the average PO under control for all units

$$
\begin{aligned}
Y_i &= Y_i(1)D_i + (1 - D_i)Y_i(0) \\
Y_i &= Y_i(1)D_i + (1 - D_i)Y_i(0) + \overline{Y}_0 + (\overline{Y}_1 - \overline{Y}_0) - (\overline{Y}_0 + (\overline{Y}_1 - \overline{Y}_0)) \\
Y_i &= \bar{Y}_0 + (\bar{Y}_1 - \bar{Y}_0)D_i + Y_i(0)(1 - D_i) + Y_i(1)D_i - \bar{Y}_0 - (\bar{Y}_1 - \bar{Y}_0)D_i \\
Y_i &= \bar{Y}_0 + (\bar{Y}_1 - \bar{Y}_0)D_i + [Y_i(0) - \bar{Y}_0 + ((Y_i(1) - \bar{Y}_1) - (Y_i(0) - \bar{Y}_0))D_i] \\
Y_i &= \alpha + \beta D_i + \epsilon_i
\end{aligned}
$$

where $\epsilon = Y_i(0) - \bar{Y}_0 + ((Y_i(1) - \bar{Y}_1) - (Y_i(0) - \bar{Y}_0))D_i$

Robust Standard Errors in R

```
rse <- function(data, x, y, b){
  # Get residuals
  e <- y - x %*% b
  # degrees of freedom
  df <- nrow(x) - ncol(x)-2
  u2 <- e^2
  # Sandwich estimator
  XDX <- 0
  for(i in 1:nrow(x)){
    XDX <- XDX + u2[i] * x[i,]%*%t(x[i,])
  }
  var_cov_m <- solve(t(x)%*%x)
  return(list(sqrt(diag(varcov_m))),df)
}
```

Regression in R in action

```
lm_byhand <- function(data,x,y){
  betas <- beta_est(data,x,y)
  beta_se <- rse(data, x, y, betas)
  return(tibble(
    estimate = betas,
    std.error = beta_se[1],
    t_stat = estimate/std.error,
    p_v = round(2 *pt(abs(t_val), beta_se[2], lower = F),4)
  ))
}
```

Regression in R in action

ID	x	y0	y1	z	y
001	0.9148060	1.2367313	1.5867313	1	1.5867313
002	0.9370754	0.1532365	0.5032365	1	0.5032365
003	0.2861395	1.8618671	2.2118671	1	2.2118671
004	0.8304476	1.4733469	1.8233469	1	1.8233469

	Estimate	SE	t-value	p-val	conf.low	conf.high	df
intercept	-0.1831761	0.1687923	-1.085216	0.28052	-0.5181820	0.1518299	97
z	0.2056310	0.1828161	1.124797	0.26345	-0.1572083	0.5684703	97
x	1.4386941	0.2816971	5.107238	0.00000	0.8796033	1.9977849	97

Regression and Effective Samples

Regression is a way to get a weighted average.

As a consequence: regression gives each data point a weight towards the total weighted average.

OLS induces a weighting scheme that implies that contributions from units in the sample are used differently. These weights are completely characterized by the regressors.

More weight goes to units whose treatment values are not well explained by the covariates. They measure the contribution of a unit's effect on the construction of $\hat{\beta}$

Regression and Effective Samples

Suppose we estimate a regression $Y_i = \alpha + \beta D_i + \epsilon_i$.

We might be worried about omitted variable bias to so we attempt to control for a potential confounder. $Y_i = \alpha + \beta D_i + \gamma X_i + u_i$

Fitting this model via OLS generates a weighting scheme

$$\hat{\beta} \xrightarrow{p} \frac{E[w_i \tau_i]}{E[w_i]}$$

where the weights $w_i = (D_i - E[D_i|X_i])^2$

Regression and Effective Samples

Multiple regression weights characterize the effective sample of a causal effect estimate. In situations where a given X_i confounds our relationship between treatment and outcomes, the effective sample is the most important thing to learn.

The effective sample is the original sample reweighted by the multiple regression weights.

Learning the effective sample is crucial for understanding to what type of population our effect estimates apply, especially in situations of heterogeneity.

A weight of zero implies that covariates completely explain treatment condition

Regression and Effective Samples

What happens when we actually have random assignment of D_i?

Fortunately, in this situation the weights cancel out and our regression coefficient is what we think it is, the unbiased estimate of the ATE

$$\frac{\sum_i^n \tilde{D}_i^2 \tau_i}{\sum_i^n \tilde{D}_i^2} \xrightarrow{p} \frac{E[w_i]E[\tau_i]}{E[w_i]} = E[\tau_i] = \bar{\tau}$$

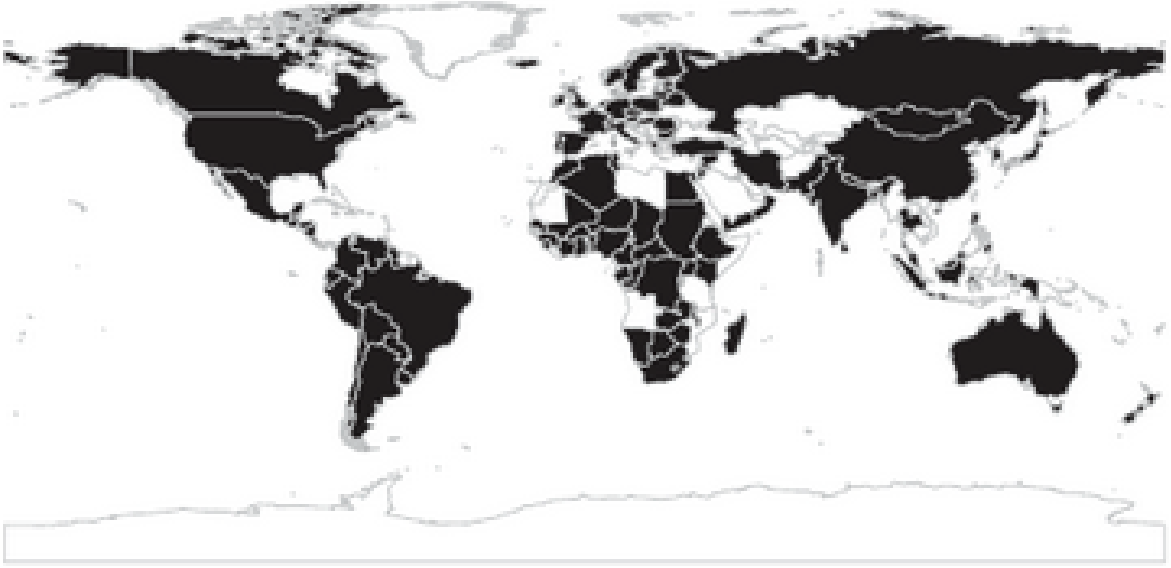
Regression and Effective Samples

We can calculate summary statistics about the nominal sample by using the weights. For example, we can estimate the mean of a covariate Z_i in the effective sample with

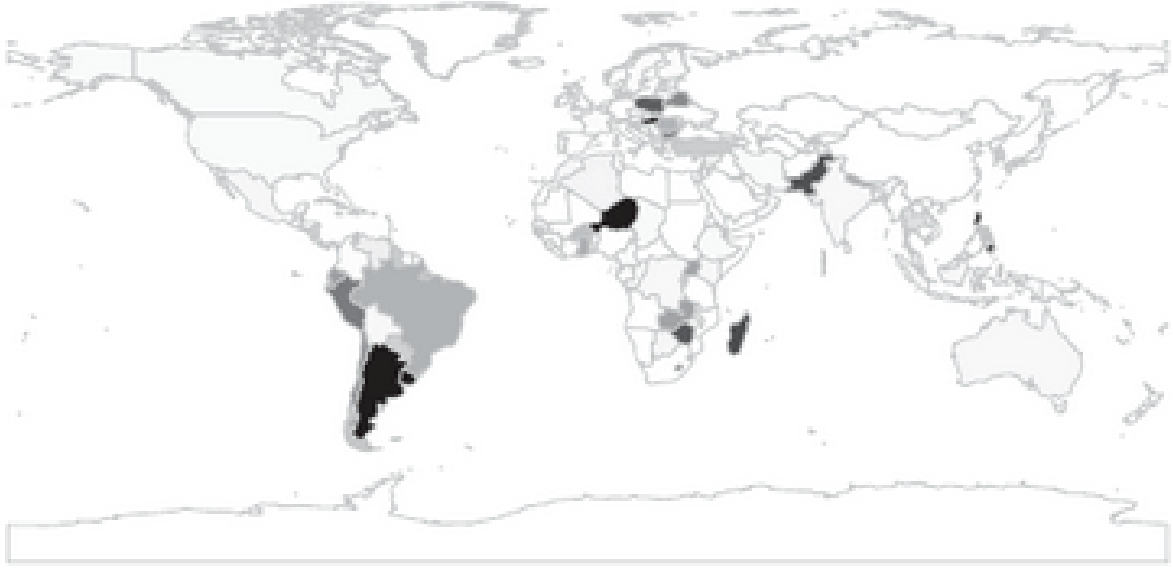
$$\hat{\mu}_{Z_i} = \text{-} \xrightarrow{p} \frac{E[w_i Z_i]}{E[w_i]} = \mu_{Zi}$$

For example, suppose we are interested in the effect of IMF agreements on foreign direct investment inflows. Z_i could be an indicator for "country in Europe" to get share of units in the effective sample that are in Europe.

Nominal Sample



Effective Sample



Saturated Models

Saturated regression models are regression models with discrete explanatory variables where the model includes a separate parameter for all possible values take on by explanatory variables

Generically

$$Y_i = \alpha + \beta_i D_{ij} + \epsilon_i$$

where $D_{ji} = 1[S_i = j]$ is a dummy variable indicating the treatment level and β_j is said to be the j^{th} level treatment effect.

Saturated models perfectly fit the CEF because the CEF is a linear function of the dummy regressors used to saturate them.

Main Effects

With two or more explanatory variables a model is saturated by including dummies for both variables, their products and a constant.

The coefficient on the dummy is the main effect. The product is the interaction term

Main Effects

Assume d_{1i} indicates treatment and x_{1i} indicates an additional variable. The CEF takes on four values

$$
\begin{aligned}
E[Y_i|d_{1i}=0, x_{1i}=0] \\
E[Y_i|d_{1i}=1, x_{1i}=0] \\
E[Y_i|d_{1i}=0, x_{1i}=1] \\
E[Y_i|d_{1i}=1, x_{1i}=1]
\end{aligned}
$$

Main Effects

We can label these with parameters

$$
\begin{aligned}
E[Y_i|d_{1i}=0, x_{1i}=0] &= \alpha \\
E[Y_i|d_{1i}=1, x_{1i}=0] &= \alpha + \beta_1 \\
E[Y_i|d_{1i}=0, x_{1i}=1] &= \alpha + \gamma \\
E[Y_i|d_{1i}=1, x_{1i}=1] &= \alpha + \beta_1 + \gamma + \delta_1
\end{aligned}
$$

As one equation in saturated regression form

$$E[Y_i|d_{1i}, x_{1i}] = \alpha + \beta_1 d_{1i} + \gamma x_{1i} + \delta_1 (d_{1i} x_{1i}) + \epsilon_i$$

Next Time

Regression and Bad Controls

Regression and Heterogeneous Effects

What we covered last time?

OLS in Matrix Form

Standard Errors for inference

What are we doing today?

Checkpoint 4 and PS1 are due today

Regression and Bad Controls

Regression with a binary outcome variable

Omitted Variable Bias

OVB formula describes the relationship between regression estimates in models with different sets of controls

Suppose we have the following regression specifications

$$Y_i = \alpha_s + \beta_s D_i + u_i$$
$$Y_i = \alpha_L + \beta_L D_i + \gamma_L X_i + v_i$$

Imagine the latter regression is "truth". What is the implication of running the short regression?

Omitted Variables Bias

Suppose we leave out X_i. Then the coefficient in the short regression is:

$$\beta_s = \frac{Cov[Y_i, D_i]}{V[D_i]} = \beta_L + \gamma' \delta_{X_s}$$

Here δ_{X_s} is the vector of coefficient from regressing the elements of X_i on D_i

Implications of the Omitted Variables Bias

1. The long and short regression will give the same results if the omitted and included variables are uncorrelated. (Can you think of a situation where this holds mechanically?)

2. The short regression is biased when an omitted variable is correlated with D_i. More generally, the error term is correlated with a variable on the RHS.

Bad Controls

What happens when we have a bad control.

Consider a PO model

$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$$
$$X_i = D_i X_{1i} + (1 - D_i) X_{0i}$$

Functionally we have two different dependent variables Y_i, X_i in this example

Due to independence

$$E[Y_{1i} - Y_{0i}] = E[Y_i | D_i = 1] - E[Y_i | D_i = 0]$$
$$E[X_{1i} - X_{0i}] = E[X_i | D_i = 1] - E[X_i | D_i = 0]$$

Bad Controls

Consider now the difference of means estimated by a regression on Y_i on D_i conditional on X_i

$$E[Y_i|D_i = 1, X_i = 1] - E[Y_i|X_i = 1, D_i = 0] =$$
$$E[Y_{1i}|X_{1i} = 1, D_i = 1] - E[Y_i|X_{0i} = 1, D_i = 0]$$

By joint independence of $\{Y_{ji}, X_{ji}\}$ and D_i

$$E[Y_{1i} - Y0i|X_{1i} = 1] + (E[Y_{0i}|X_{1i} = 1] - E[Y_{0i}|X_{1i} = 0])$$

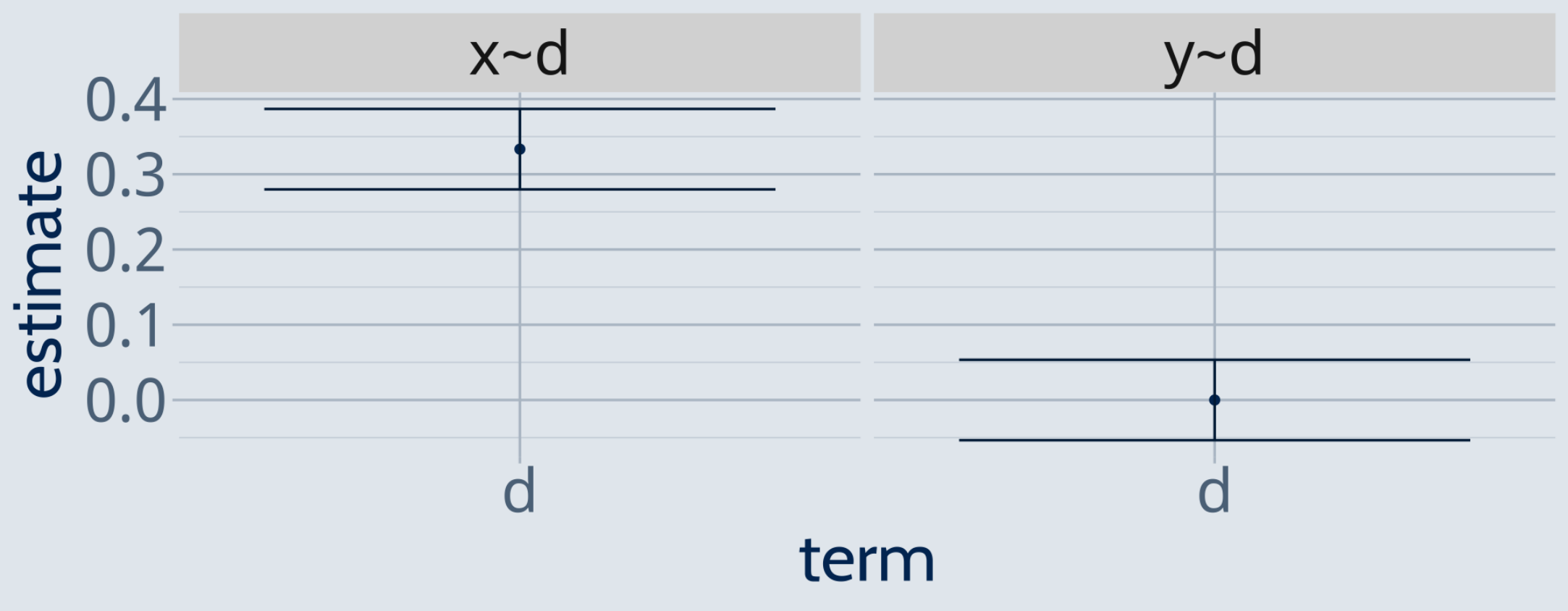
Bad Controls Example 1

Consider the following example. Here by construction there is no treatment effect of d on y. There is an effect of d on x.

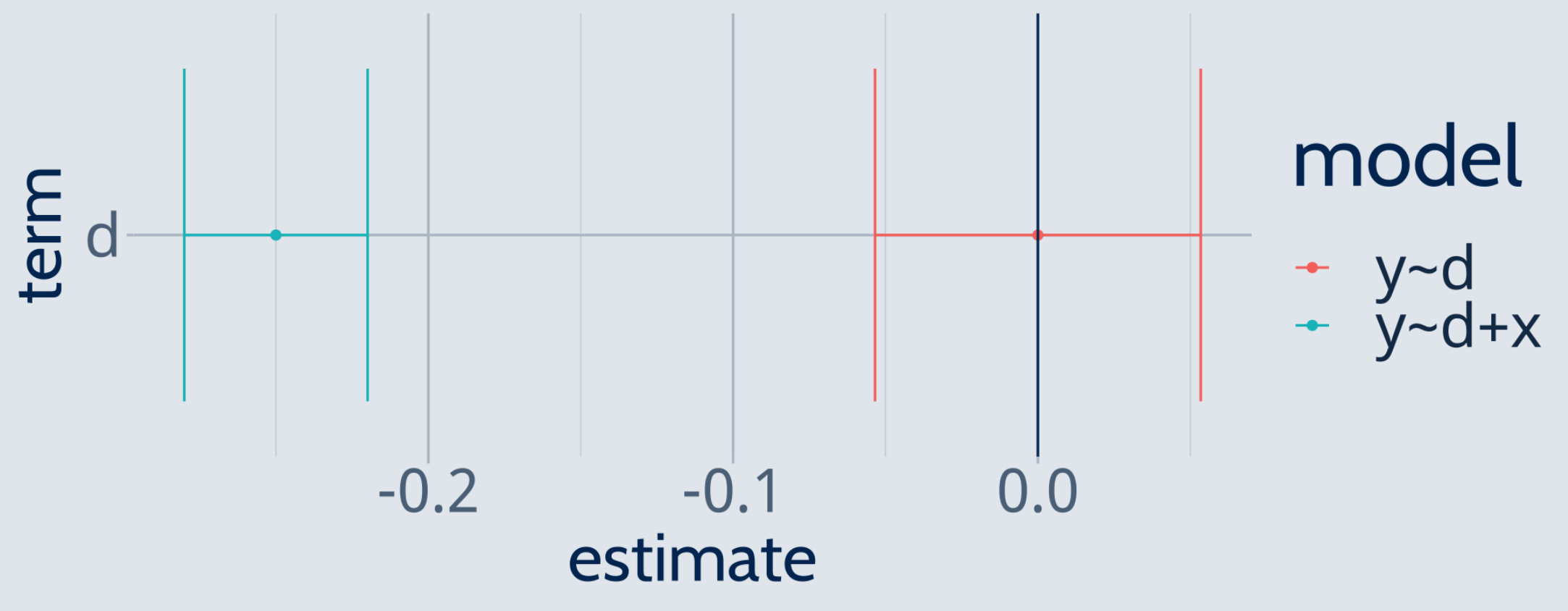
Both y and x can be thought of as different independent variables

```
# Bad Controls Example
stacked <- tibble(
  y = c(rep(0, 200), rep(1,200),
        rep(1, 200), rep(0, 200),
        rep(1, 200), rep(1, 200)),
  d = c(rep(0, 600), rep(1, 600)),
  x = c(rep(0,400), rep(1,200),
        rep(0, 200), rep(1, 400))
)
```

Bad Controls Example 1



Bad Controls Example 1



Bad Controls Example 2: Knox, Lowe, Mummolo (2020)

There is much concern (correctly) over racial bias in policing in the US.

Most research uses large administrative data sets on police-civilian interactions

Normally the ATE of interest is the effect of race on the outcome of a stop, e.g. whether a search is conducted.

Question for the room: What happens if stop decisions are influenced by the perceived race of a civilian?

Bad Controls Example 2

Administrative records do not have data on all civilians that could be stopped by police, but only those who have been stopped by the police

We actually hope this is discriminatory in the sense that stopping people at random would be a problem. The worry is that police stop individuals for reasons that are not justified (racially discriminatory)

If perceived race affects whether officers choose to stop an individuals then analyzing administrative records now amounts to conditioning on a variable affected by an individual's race

Regression with Binary Outcome variables

Our theoretical estimand is often binary (survive/not survive, did vote/did not vote, won/lost)

OLS outcome predictions are not limited to $[0, 1]$. Since it is impossible to have more than 1 or less than 0 as an outcome, OLS could make an impossible prediction. Is this a problem for regression?

No.

Regression with Binary Outcome variables

Consider our model $Y_i = \alpha + \beta D_i + \epsilon$

We have shown that the ATE is equal to β directly expressed in terms of probability. OLS always provides an unbiased estimate of the casual effect of D_i on Y_i provided our assumptions hold

In this context, our regression is often referred to as the Linear Probability Model

Regression with Binary Outcome variables

The benefits of just running the usual regression:

1. The parameter is directly interpretable
2. Interaction terms make sense and are directly interpretable
3. LPMs do better if we include "fixed effects" which practically occur a lot as a design choice
4. The LPM and most standard non-linear models give the same general answer anyway