Gov 2002: 7. Regression and Causality

Matthew Blackwell

October 15, 2015

Agnostic Regression

Regression and Causality

Regression with Heterogeneous Treatment Effects

Where are we? Where are we going?

- Last few weeks: using matching, weighting for estimating causal effects.
- This week: how to use regression to estimate causal effects.
- Regression is so widely used, it's good to know what it's actually estimating!
- Goal: salvage regression from the ashes of 1980's textbooks!
- Next week: panel data!

Reminder Email me and Stephen a half-page description of your proposed research project.

1/ Agnostic Regression

Regression as parametric modeling

- Gauss-Markov assumptions:
  - linearity, i.i.d. sample, full rank X_i, zero conditional mean error, homoskedasticity.
- $\rightsquigarrow$ OLS is BLUE, plus normality of the errors and we get small sample SEs.
- What is the basic approach here? It is a model for the conditional distribution of Y_i given X_i:

$$[Y_i | X_i] \sim N(X_i'\beta, \sigma^2)$$

- MLE from this model is the usual OLS estimator, $\hat{\beta}_{\text{OLS}}$:

$$\hat{\beta}_{\text{OLS}} = \left[\sum_{i=1}^{N} X_i X_i'\right]^{-1} \sum_{i=1}^{N} X_i Y_i$$

Agnostic views on regression

$$[Y_i|X_i] \sim N(X_i'\beta, \sigma^2)$$

- Strong distributional assumption on Y_i.
- Properties like BLUE or MLE properties depend on these assumptions holding.
- Alternative: take an **agnostic** view on regression.
  - Use OLS without believing these assumptions.
- Lose the distributional assumptions, focus on the conditional expectation function (CEF):

$$\mu(x) = \mathbb{E}[Y_i|X_i = x] = \sum_y y \cdot \mathbb{P}[Y_i = y|X_i = x]$$

Justifying linear regression

- Define linear regression:

$$\beta = \arg\min_b \mathbb{E}[(Y_i - X_i'b)^2]$$

- The solution to this is the following:

$$\beta = \mathbb{E}[X_i X_i']^{-1}\mathbb{E}[X_i Y_i]$$

- Note that the is the **population** coefficient vector, not the estimator yet.

Regression anatomy

- Consider simple linear regression:

$$(\alpha, \beta) = \underset{a,b}{\arg\min}\, \mathbb{E}\left[(Y_i - a - bX_i)^2\right]$$

- In this case, we can write the population/true slope β as:

$$\beta = \mathbb{E}[X_i X_i']^{-1}\mathbb{E}[X_i Y_i] = \frac{\text{Cov}(Y_i, X_i)}{\mathbb{V}[X_i]}$$

- With more covariates, β is more complicated, but we can still write it like this.
- Let $\tilde{X}_{ki}$ be the residual from a regression of X_{ki} on all the other independent variables. Then, β_k, the coefficient for X_{ki} is:

$$\beta_k = \frac{\text{Cov}(Y_i, \tilde{X}_{ki})}{V(\tilde{X}_{ki})}$$

Justification 1: Linear CEFs

- Justification 1: if the CEF is linear, the population regression function is it. That is, if $E[Y_i|X_i] = X_i'b$, then $b = \beta$.
- When would we expect the CEF to be linear? Two cases.
    1. Outcome and covariates are **multivariate normal**.
    2. Linear regression model is **saturated**.
- A model is **saturated** if there are as many parameters as there are possible combination of the X_i variables.

Saturated model example

- Two binary variables, X_{1i} for incumbency status and X_{2i} for party of the candidate.
- Four possible values of X_i, four possible values of $\mu(X_i)$:

$$E[Y_i|X_{1i} = 0, X_{2i} = 0] = \alpha$$
$$E[Y_i|X_{1i} = 1, X_{2i} = 0] = \alpha + \beta$$
$$E[Y_i|X_{1i} = 0, X_{2i} = 1] = \alpha + \gamma$$
$$E[Y_i|X_{1i} = 1, X_{2i} = 1] = \alpha + \beta + \gamma + \delta$$

- We can write the CEF as follows:

$$E[Y_i|X_{1i}, X_{2i}] = \alpha + \beta X_{1i} + \gamma X_{2i} + \delta(X_{1i}X_{2i})$$

Saturated models example

$$E[Y_i|X_{1i}, X_{2i}] = \alpha + \beta X_{1i} + \gamma X_{2i} + \delta(X_{1i}X_{2i})$$

- Basically, each value of $\mu(X_i)$ is being estimated separately.
  - $\rightsquigarrow$ within-strata estimation.
  - No borrowing of information from across values of X_i.
- Requires a set of dummies for each categorical variable plus **all interactions**.
- Or, a series of dummies for each unique combination of X_i.
- This makes **linearity hold mechanically** and so linearity is not an assumption.
  - Just a fact about saturated CEFs.
  - $\rightsquigarrow$ saturated models for limited dependent variables = A-OK!

Saturated model example

- Washington (AER) data on the effects of daughters.
- We'll look at the relationship between voting and number of kids (causal?).

```
girls <- foreign::read.dta("girls.dta")
head(girls[, c("name", "totchi", "aauw")])
```

```
##                    name totchi aauw
## 1   ABERCROMBIE, NEIL      0  100
## 2   ACKERMAN, GARY L.      3   88
## 3 ADERHOLT, ROBERT B.      0    0
## 4    ALLEN, THOMAS H.      2  100
## 5  ANDREWS, ROBERT E.      2  100
## 6        ARCHER, W.R.      7    0
```

Linear model

```r
summary(lm(aauw ~ totchi, data = girls))
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    61.31       1.81   33.81   <2e-16 ***
## totchi         -5.33       0.62   -8.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42 on 1733 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.0408, Adjusted R-squared:  0.0403
## F-statistic: 73.8 on 1 and 1733 DF,  p-value: <2e-16
```

Saturated model

```
summary(lm(aauw ~ as.factor(totchi), data = girls))
```

```
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)            56.41       2.76   20.42  < 2e-16 ***
## as.factor(totchi)1      5.45       4.11    1.33   0.1851
## as.factor(totchi)2     -3.80       3.27   -1.16   0.2454
## as.factor(totchi)3    -13.65       3.45   -3.95  8.1e-05 ***
## as.factor(totchi)4    -19.31       4.01   -4.82  1.6e-06 ***
## as.factor(totchi)5    -15.46       4.85   -3.19   0.0015 **
## as.factor(totchi)6    -33.59      10.42   -3.22   0.0013 **
## as.factor(totchi)7    -17.13      11.41   -1.50   0.1336
## as.factor(totchi)8    -55.33      12.28   -4.51  7.0e-06 ***
## as.factor(totchi)9    -50.41      24.08   -2.09   0.0364 *
## as.factor(totchi)10   -53.41      20.90   -2.56   0.0107 *
## as.factor(totchi)12   -56.41      41.53   -1.36   0.1745
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41 on 1723 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.0506, Adjusted R-squared:  0.0446
## F-statistic: 8.36 on 11 and 1723 DF,  p-value: 1.84e-14
```

Saturated model minus the constant

```r
summary(lm(aauw ~ as.factor(totchi) - 1, data = girls))
```

```
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## as.factor(totchi)0      56.41       2.76   20.42   <2e-16 ***
## as.factor(totchi)1      61.86       3.05   20.31   <2e-16 ***
## as.factor(totchi)2      52.62       1.75   30.13   <2e-16 ***
## as.factor(totchi)3      42.76       2.07   20.62   <2e-16 ***
## as.factor(totchi)4      37.11       2.90   12.79   <2e-16 ***
## as.factor(totchi)5      40.95       3.99   10.27   <2e-16 ***
## as.factor(totchi)6      22.82      10.05    2.27   0.0233 *
## as.factor(totchi)7      39.29      11.07    3.55   0.0004 ***
## as.factor(totchi)8       1.08      11.96    0.09   0.9278
## as.factor(totchi)9       6.00      23.92    0.25   0.8020
## as.factor(totchi)10      3.00      20.72    0.14   0.8849
## as.factor(totchi)12      0.00      41.43    0.00   1.0000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41 on 1723 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.587,  Adjusted R-squared:  0.584
## F-statistic:  204 on 12 and 1723 DF,  p-value: <2e-16
```

Compare to within-strata means

- The saturated model makes no assumptions about the between-strata relationships.
- Just calculates within-strata means:

```
c1 <- coef(lm(aauw ~ as.factor(totchi) - 1, data = girls))
c2 <- with(girls, tapply(aauw, totchi, mean, na.rm = TRUE))
rbind(c1, c2)
```

```
##     0  1  2  3  4  5  6  7   8 9 10 12
## c1 56 62 53 43 37 41 23 39 1.1 6  3  0
## c2 56 62 53 43 37 41 23 39 1.1 6  3  0
```

Other justifications for OLS

- **Justification 2**: $X_i'\beta$ is the best linear predictor (in a mean-squared error sense) of Y_i.
  - Why? $\beta = \arg\min_b \mathbb{E}[(Y_i - X_i'b)^2]$
- **Justification 3**: $X_i'\beta$ provides the minimum mean squared error linear approxmiation to $E[Y_i|X_i]$.
- Even if the CEF is not linear, a linear regression provides the best linear approximation to that CEF.
- Don't need to believe the assumptions (linearity) in order to use regression as a good approximation to the CEF.
- **Warning** if the CEF is very nonlinear then this approximation could be terrible!!

The error terms

- Let's define the error term: $e_i \equiv Y_i - X_i'\beta$ so that:

$$Y_i = X_i'\beta + [Y_i - X_i'\beta] = X_i'\beta + e_i$$

- Note the residual e_i is uncorrelated with X_i:

$$\begin{aligned}
\mathbb{E}[X_i e_i] &= \mathbb{E}[X_i(Y_i - X_i'\beta)] \\
&= \mathbb{E}[X_i Y_i] - \mathbb{E}[X_i X_i'\beta] \\
&= \mathbb{E}[X_i Y_i] - \mathbb{E}\left[X_i X_i' \mathbb{E}[X_i X_i']^{-1} \mathbb{E}[X_i Y_i]\right] \\
&= \mathbb{E}[X_i Y_i] - \mathbb{E}[X_i X_i'] \mathbb{E}[X_i X_i']^{-1} \mathbb{E}[X_i Y_i] \\
&= \mathbb{E}[X_i Y_i] - \mathbb{E}[X_i Y_i] = 0
\end{aligned}$$

- No assumptions on the linearity of $\mathbb{E}[Y_i|X_i]$.

OLS estimator

- We know the population value of β is:

$$\beta = \mathbb{E}[X_i X_i']^{-1}\mathbb{E}[X_i Y_i]$$

- How do we get an estimator of this?
- **Plug-in principle** $\rightsquigarrow$ replace population expectation with sample versions:

$$\hat{\beta} = \left[\frac{1}{N}\sum_i X_i X_i'\right]^{-1}\frac{1}{N}\sum_i X_i Y_i$$

- If you work through the matrix algebra, this turns out to be:

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

Asymptotic OLS inference

- With this representation in hand, we can write the OLS estimator as follows:

$$\hat{\beta} = \beta + \left[\sum_i X_i X_i'\right]^{-1} \sum_i X_i e_i$$

- Core idea: $\sum_i X_i e_i$ is the sum of r.v.s so the CLT applies.
- That, plus some simple asymptotic theory allows us to say:

$$\sqrt{N}(\hat{\beta} - \beta) \rightsquigarrow N(0, \Omega)$$

- Converges in distribution to a Normal distribution with mean vector 0 and covariance matrix, Ω:

$$\Omega = \mathbb{E}[X_i X_i']^{-1} \mathbb{E}[X_i X_i' e_i^2] \mathbb{E}[X_i X_i']^{-1}.$$

- No linearity assumption needed!

Estimating the variance

- In large samples then:

$$\sqrt{N}(\hat{\beta} - \beta) \sim N(0, \Omega)$$

- How to estimate Ω? **Plug-in principle** again!

$$\widehat{\Omega} = \left[\sum_i X_i X_i'\right]^{-1} \left[\sum_i X_i X_i' \hat{e}_i^2\right] \left[\sum_i X_i X_i'\right]^{-1}.$$

- Replace e_i with its emprical counterpart (residuals) $\hat{e}_i = Y_i - X_i'\hat{\beta}$.
- Replace the population moments of X_i with their sample counterparts.
- The square root of the diagonals of this covariance matrix are the "robust" or Huber-White standard errors that Stata commonly report.

Heteroskedasticity

- No assumptions of homoskedasticity.
- Heteroskedaticity will definitely occur when:
  - CEF is linear, but the $\sigma^2(x) = \mathbb{V}[Y_i|X_i = x]$ is not constant in x.
  - $E[Y_i|X_i]$ is not linear, but we use the linear regression to approxmiate it.

2/ Regression and Causality

Regression and causality

- Most econometrics textbooks: regression defined without respect to causality.
- But then when is $\hat{\beta}$ "biased"? The above derivations work for some $\mathbb{E}[Y_i|X_i]$.
- The question, then, is when does knowing the CEF tell us something about causality?
- MHE argues that a regression is causal when the CEF it approximates is causal. Identification is king.
- We will show that under certain conditions, a regression of the outcome on the treatment and the covariates can recover *a* causal parameter, but perhaps not the one in which we are interested.

Review

- Quick reminder: we have potential outcomes, $Y_i(1)$ and $Y_i(0)$, and two parameters, the ATE and ATT:

$$\tau = E[Y_i(1) - Y_i(0)],$$
$$\tau_{\text{ATT}} = E[Y_i(1) - Y_i(0)|D_i = 1].$$

- We have shown in past weeks that these effects are identified when ignorability holds. MHE calls this the conditional independence assumption (CIA).

Linear constant effects model, binary treatment

- Experiment: with a simple experiment, we can rewrite the consistency assumption to be a regression formula:

$$
\begin{aligned}
Y_i &= D_i Y_i(1) + (1 - D_i) Y_i(0) \\
&= Y_i(0) + (Y_i(1) - Y_i(0)) D_i \\
&= \mathbb{E}[Y_i(0)] + \tau D_i + (Y_i(0) - \mathbb{E}[Y_i(0)]) \\
&= \mu^0 + \tau D_i + v_i^0
\end{aligned}
$$

- Note that if ignorability holds (as in an experiment) for $Y_i(0)$, then it will also hold for v_i^0, since $\mathbb{E}[Y_i(0)]$ is constant. Thus, this satifies the usual assumptions for regression.

Now with covariates

- Now assume no unmeasured confounders: $Y_i(d) \perp\!\!\!\perp D_i | X_i$.
- We will assume a linear model for the potential outcomes:

$$Y_i(d) = \alpha + \tau \cdot d + \eta_i$$

- Remember that linearity isn't an assumption if D_i is binary
- Effect of D_i is constant here, the η_i are the only source of individual variation and we have $E[\eta_i] = 0$.
- Consistency assumption allows us to write this as:

$$Y_i = \alpha + \tau D_i + \eta_i.$$

Covariates in the error

- Let's assume that η_i is linear in X_i: $\eta_i = X_i'\gamma + \nu_i$
- New error is uncorrelated with X_i: $\mathbb{E}[\nu_i|X_i] = 0$.
- This is an assumption! Might be false!
- Plug into the above:

$$\begin{aligned}\mathbb{E}[Y_i(d)|X_i] = E[Y_i|D_i, X_i] &= \alpha + \tau D_i + E[\eta_i|X_i] \\ &= \alpha + \tau D_i + X_i'\gamma + E[\nu_i|X_i] \\ &= \alpha + \tau D_i + X_i'\gamma\end{aligned}$$

Summing up regression with constant effects

- Reviewing the assumptions we've used:
  - no unmeasured confounders
  - constant treatment effects
  - linearity of the treatment/covariates
- Under these, we can run the following regression to estimate the ATE, τ:
$$Y_i = \alpha + \tau D_i + X_i'\gamma + \nu_i$$
- Works with continuous or ordinal D_i if linearity in the effect of these variables is truly linear.

OLS constant effects simulation

Model with linear covariates, constant 0 effect of treatment:

```
library(mvtnorm)
n <- 100
p <- 4
X <- rmvnorm(n = 100, mean = rep(0, p))
gamma <- c(27.4, 13.7, 13.7, 13.7)
y <- 210 + X %*% c(gamma) + rnorm(n)

alpha <- c(-1, 0.5, -0.5, -0.1)
dprobs <- boot::inv.logit(X %*% alpha)
d <- rbinom(n, size = 1, prob = dprobs)
```

OLS with no covariates

```r
summary(lm(y ~ d))
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   220.36       4.58   48.13  < 2e-16 ***
## d             -28.69       6.54   -4.39 0.000029 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33 on 98 degrees of freedom
## Multiple R-squared:  0.164,  Adjusted R-squared:  0.156
## F-statistic: 19.2 on 1 and 98 DF,  p-value: 0.000029
```

OLS with covariates

```r
summary(lm(y ~ d + X))
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  209.952      0.159  1320.7   <2e-16 ***
## d             -0.128      0.253    -0.5     0.62
## X1            27.368      0.126   217.5   <2e-16 ***
## X2            13.677      0.114   120.1   <2e-16 ***
## X3            13.673      0.130   105.1   <2e-16 ***
## X4            13.570      0.106   128.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1 on 94 degrees of freedom
## Multiple R-squared:  0.999,  Adjusted R-squared:  0.999
## F-statistic: 2.44e+04 on 5 and 94 DF,  p-value: <2e-16
```

What happens with nonlinearity

Suppose we can only observe the following covariates:

```
z1 <- exp(X[, 1]/2)
z2 <- X[, 2]/(1 + exp(X[, 1])) + 10
z3 <- (X[, 1] * X[, 3]/25 + 0.6)^3
z4 <- (X[, 2] + X[, 4] + 20)^2
```

Implies that Y_i and D_i are functions of $\log(Z_{i1})$, Z_{i2}, $Z_{i1}^2 Z_{i2}$, $1/\log(Z_{i1})$, $Z_{i3}/\log(Z_{i1})$, and $X_{i4}^{1/2}$.

Regression is a **nonlinear** function of the observed covariates.

When linearity goes wrong

```r
summary(lm(y ~ d + z1 + z2 + z3 + z4))
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.8728    30.0799    0.73    0.469
## d            -6.9292     3.3854   -2.05    0.043 *
## z1           36.3110     2.6970   13.46   <2e-16 ***
## z2           -2.9033     3.6619   -0.79    0.430
## z3           86.2022    43.1030    2.00    0.048 *
## z4            0.4021     0.0329   12.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14 on 94 degrees of freedom
## Multiple R-squared:  0.855,  Adjusted R-squared:  0.847
## F-statistic:  111 on 5 and 94 DF,  p-value: <2e-16
```

3/ Regression with Heterogeneous Treatment Effects

Heterogeneous effects, binary treatment

- Completely randomized experiment:

$$\begin{aligned} Y_i &= D_i Y_i(1) + (1 - D_i) Y_i(0) \\ &= Y_i(0) + (Y_i(1) - Y_i(0)) D_i \\ &= \mu_0 + \tau_i D_i + (Y_i(0) - \mu_0) \\ &= \mu_0 + \tau D_i + (Y_i(0) - \mu_0) + (\tau_i - \tau) \cdot D_i \\ &= \mu_0 + \tau D_i + \varepsilon_i \end{aligned}$$

- Error term now includes two components:
    1. "Baseline" variation in the outcome: $(Y_i(0) - \mu_0)$
    2. Variation in the treatment effect, $(\tau_i - \tau)$
- Easy to verify that under experiment, $\mathbb{E}[\varepsilon_i | D_i] = 0$
- Thus, OLS estimates the ATE with no covariates.

Adding covariates

- What happens with no unmeasured confounders? Need to condition on X_i now.
- Remember identification of the ATE/ATT using iterated expectations.
- ATE is the weighted sum of CATEs:

$$\tau = \sum_x \tau(x) \Pr[X_i = x]$$

- ATE/ATT are weighted averages of CATEs.
- What about the regression estimand, τ_R? How does it related to the ATE/ATT?

Heterogeneous effects and regression

- Let's investigate this under a saturated regression model:

$$Y_i = \sum_x B_{xi}\alpha_x + \tau_R D_i + e_i.$$

- Use a dummy variable for each unique combination of X_i: $B_{xi} = \mathbb{I}(X_i = x)$
- Linear in X_i by construction!

Investigating the regression coefficient

- How can we investigate τ_R? Well, we can rely on the regression anatomy:

$$\tau_R = \frac{\text{Cov}(Y_i, D_i - E[D_i|X_i])}{\mathbb{V}(D_i - E[D_i|X_i])}$$

- $D_i - \mathbb{E}[D_i|X_i]$ is the residual from a regression of D_i on the full set of dummies.
- With a little work we can show:

$$\tau_R = \frac{\mathbb{E}\left[\tau(X_i)(D_i - \mathbb{E}[D_i|X_i])^2\right]}{\mathbb{E}[(D_i - E[D_i|X_i])^2]} = \frac{\mathbb{E}[\tau(X_i)\sigma_d^2(X_i)]}{\mathbb{E}[\sigma_d^2(X_i)]}$$

- $\sigma_d^2(x) = \mathbb{V}[D_i|X_i = x]$ is the conditional variance of treatment assignment.

ATE versus OLS

$$\tau_R = \mathbb{E}[\tau(X_i)W_i] = \sum_x \tau(x) \frac{\sigma_d^2(x)}{\mathbb{E}[\sigma_d^2(X_i)]} \mathbb{P}[X_i = x]$$

- Compare to the ATE:

$$\tau = \mathbb{E}[\tau(X_i)] = \sum_x \tau(x) \mathbb{P}[X_i = x]$$

- Both weight strata relative to their size $(\mathbb{P}[X_i = x])$
- OLS weights strata higher if the treatment variance in those strata $(\sigma_d^2(x))$ is higher in those strata relative to the average variance across strata $(\mathbb{E}[\sigma_d^2(X_i)])$.
- The ATE weights only by their size.

Regression weighting

$$W_i = \frac{\sigma_d^2(X_i)}{\mathbb{E}[\sigma_d^2(X_i)]}$$

- Why does OLS weight like this?
- OLS is a **minimum-variance estimator** $\rightsquigarrow$ more weight to more precise within-strata estimates.
- Within-strata estimates are most precise when the treatment is evenly spread and thus has the highest variance.
- If D_i is binary, then we know the conditional variance will be:

$$\begin{aligned}\sigma_d^2(x) &= \mathbb{P}[D_i = 1|X_i = x]\,(1 - \mathbb{P}[D_i = 1|X_i = x]) \\ &= e(x)\,(1 - e(x))\end{aligned}$$

- Maximum variance with $\mathbb{P}[D_i = 1|X_i = x] = 1/2$.

OLS weighting example

- Binary covariate:

$$
\begin{aligned}
\mathbb{P}[X_i = 1] &= 0.75 & \mathbb{P}[X_i = 0] &= 0.25 \\
\mathbb{P}[D_i = 1|X_i = 1] &= 0.9 & \mathbb{P}[D_i = 1|X_i = 0] &= 0.5 \\
\sigma_d^2(1) &= 0.09 & \sigma_d^2(0) &= 0.25 \\
\tau(1) &= 1 & \tau(0) &= -1
\end{aligned}
$$

- Implies the ATE is $\tau = 0.5$
- Average conditional variance: $\mathbb{E}[\sigma_d^2(X_i)] = 0.13$
- $\rightsquigarrow$ weights for $X_i = 1$ are: $0.09/0.13 = 0.692$, for $X_i = 0$: $0.25/0.13 = 1.92$.

$$
\begin{aligned}
\tau_R &= \mathbb{E}[\tau(X_i)W_i] \\
&= \tau(1)W(1)\mathbb{P}[X_i = 1] + \tau(0)W(0)\mathbb{P}[X_i = 0] \\
&= 1 \times 0.692 \times 0.75 + -1 \times 1.92 \times 0.25 \\
&= 0.039
\end{aligned}
$$

When will OLS estimate the ATE?

- When does $\tau = \tau_R$?
- Constant treatment effects: $\tau(x) = \tau = \tau_R$
- Constant probability of treatment: $e(x) = \mathbb{P}[D_i = 1|X_i = x] = e$.
  - Implies that the OLS weights are 1.
- Incorrect linearity assumption in X_i will lead to more bias.

Other ways to use regression

- What's the path forward?
  - Accept the bias (might be relatively small with saturated models)
  - Use a different regression approach
- Let $\mu_d(x) = \mathbb{E}[Y_i(d)|X_i = x]$ be the CEF for the potential outcome under $D_i = d$.
- By consistency and n.u.c., we have $\mu_d(x) = \mathbb{E}[Y_i|D_i = d, X_i = x]$.
- Estimate a regression of Y_i on X_i among the $D_i = d$ group.
- Then, $\hat{\mu}_d(x)$ is just a predicted value from the regression for $X_i = x$.
- How can we use this?

Imputation estimators

- Impute the treated potential outcomes with $\widehat{Y}_i(1) = \hat{\mu}_1(X_i)$!
- Impute the control potential outcomes with $\widehat{Y}_i(0) = \hat{\mu}_0(X_i)$!
- Procedure:
    - Regress Y_i on X_i in the treated group and get predicted values for all units (treated or control).
    - Regress Y_i on X_i in the control group and get predicted values for all units (treated or control).
    - Take the average difference between these predicted values.
- More mathematically, look like this:

$$\tau_{imp} = \frac{1}{N} \sum_i \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$$

- Sometimes called an **imputation estimator**.

Simple imputation estimator

- Use `predict()` from the within-group models on the data from the entire sample.
- Useful trick: use a model on the entire data and `model.frame()` to get the right design matrix:

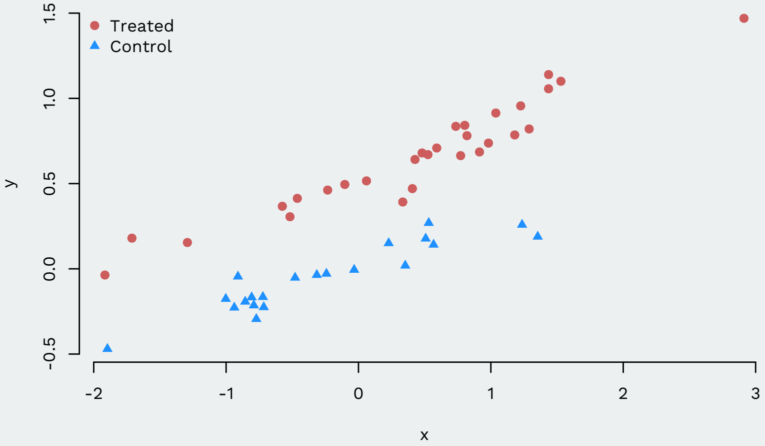
```
## heterogeneous effects
y.het <- ifelse(d == 1, y + rnorm(n, 0, 5), y)

mod <- lm(y.het ~ d + X)
mod1 <- lm(y.het ~ X, subset = d == 1)
mod0 <- lm(y.het ~ X, subset = d == 0)
y1.imps <- predict(mod1, model.frame(mod))
y0.imps <- predict(mod0, model.frame(mod))
mean(y1.imps - y0.imps)
```

```
## [1] 0.61
```

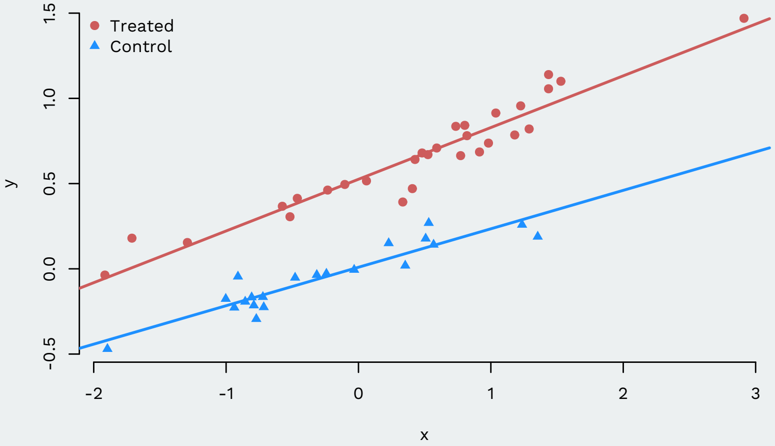
Notes on imputation estimators

- If $\hat{\mu}_d(x)$ are consistent estimators, then τ_{imp} is consistent for the ATE.
- Why don't people use this?
  - Most people don't know the results we've been talking about.
  - Harder to implement than vanilla OLS.
- Can use linear regression to estimate $\hat{\mu}_d(x) = x'\beta_d$
- Recent trend is to estimate $\hat{\mu}_d(x)$ via non-parametric methods such as:
  - Kernel regression, local linear regression, regression trees, etc
  - Easiest is generalized additive models (GAMs)

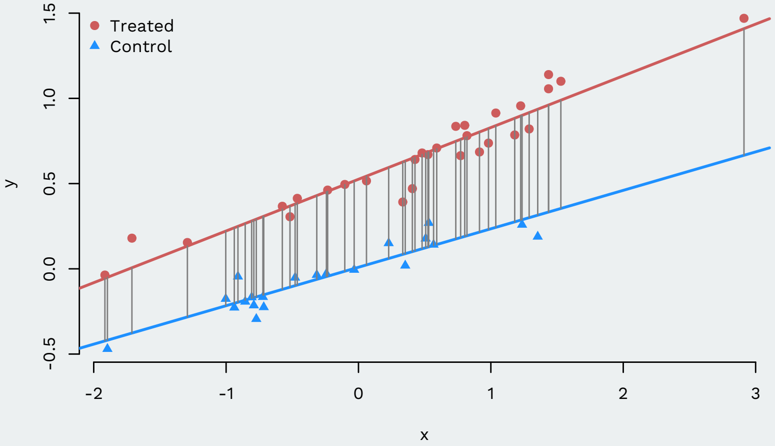
Imputation estimator visualization



Imputation estimator visualization

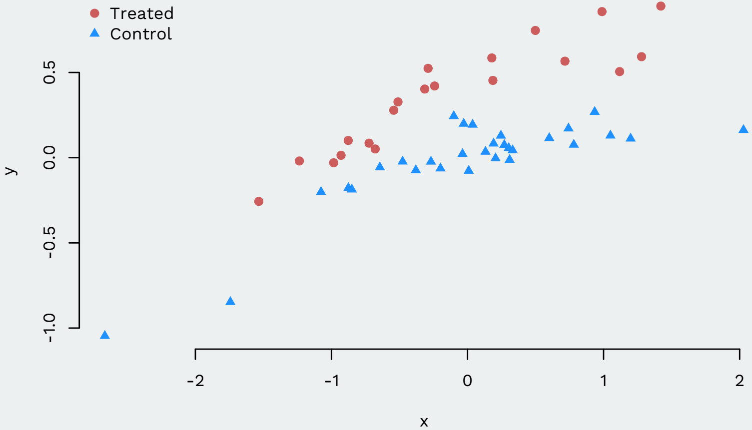


Imputation estimator visualization



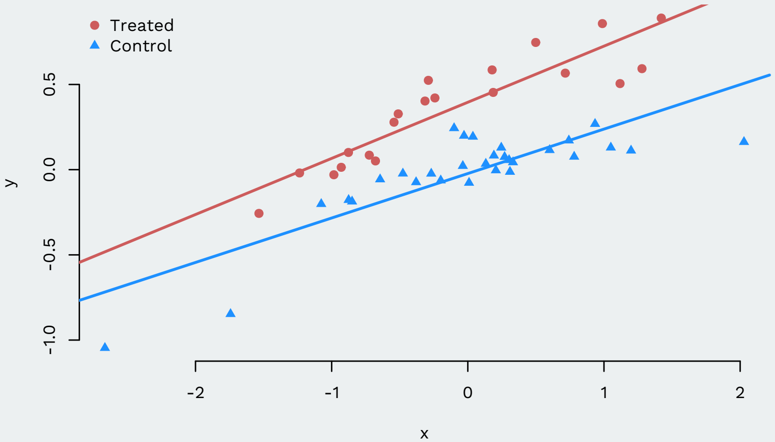
Nonlinear relationships

- Same idea but with nonlinear relationship between Y_i and X_i:



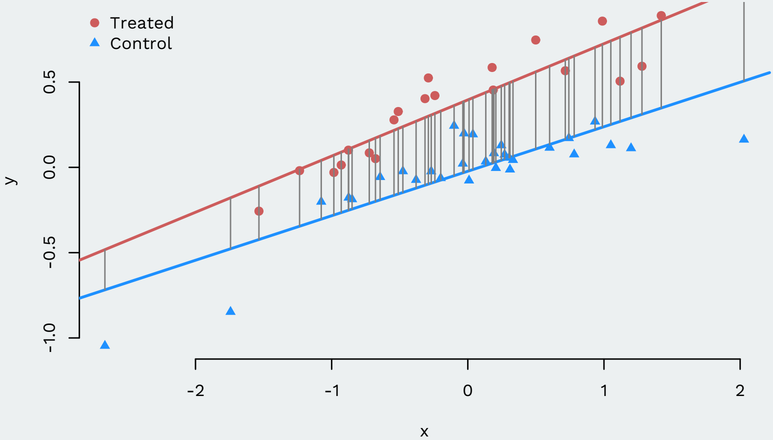
Nonlinear relationships

- Same idea but with nonlinear relationship between Y_i and X_i:



Nonlinear relationships

- Same idea but with nonlinear relationship between Y_i and X_i:



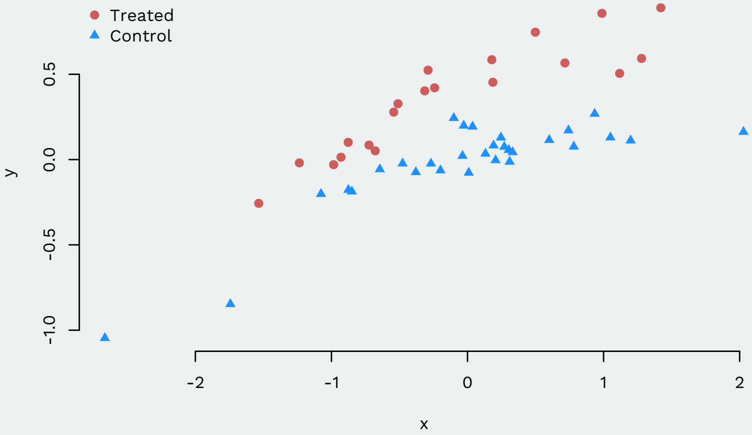
Using semiparametric regression

- Here, CEFs are nonlinear, but we don't know their form.
- We can use GAMs from the `mgcv` package to for flexible estimate:

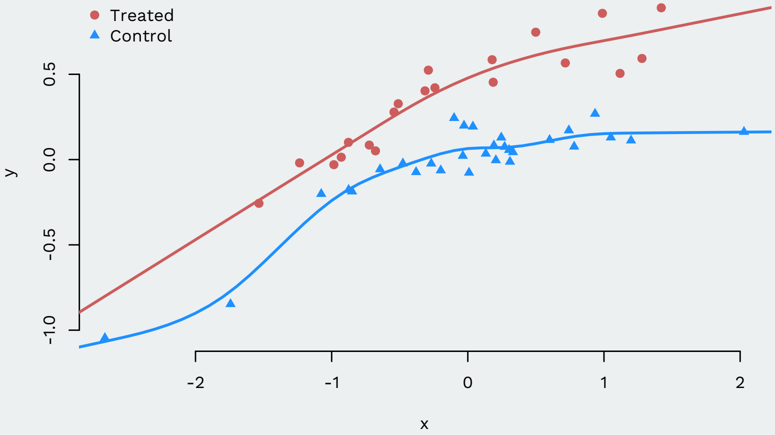
```
library(mgcv)
mod0 <- gam(y ~ s(x), subset = d == 0)
summary(mod0)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## y ~ s(x)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.0225     0.0154   -1.46     0.16
##
## Approximate significance of smooth terms:
##       edf Ref.df    F p-value
## s(x) 6.03   7.08 41.3  <2e-16 ***
## ---
```

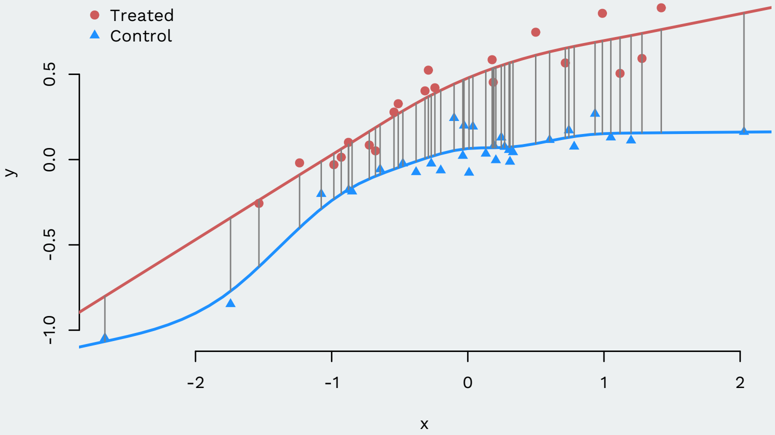
Using GAMs



Using GAMs



Using GAMs



Limited dependent variables

- Usual advice: model the data from first principles:
  - Logit/probit for binary, Poisson for counts, etc.
- OLS is a-ok with limited DVs when:
  - Binary treatment and no covariates (just diff-in-means)
  - Binary treatment, discrete covariates, and saturated models (stratified diff-in-means)
- Imposing a model on LDVs in this case imposes a distributional assumption which could be wrong!
- Even in unsaturated models, the marginal effect from OLS often decent compared to nonlinear models.
  - Could go wrong in small samples
  - If using nonlinear models, always get effects on the scale of the outcome.