Week 4: Selection on Observables II Regression
PUBL0050 Causal Inference

Julia de Romémont

Term 2 2023-24

UCL Departement of Political Science

Motivation

You have all seen tables like this

TABLE 1 Religion as a Predictor of Nazi Vote Shares, November 1932

	NSDAP Vote Share						
Percent Catholic	−.250	−.237	−.243	−.245	−.269	−.287	
	(.019)	(.016)	(.018)	(.020)	(.030)	(.025)	
Percent Nonreligious	−1.316	−.977	−.855	−.823	−.717	−.648	
	(.195)	(.150)	(.150)	(.144)	(.147)	(.113)	
Demographics							
Percent Female		.447	1.143	1.180	1.771	.650	1.216
		(.503)	(.566)	(.537)	(.546)	(.443)	(.679)
Urban County		−1.482	.424	−.191	−.800	−.140	−4.395
		(1.012)	(1.217)	(1.179)	(1.237)	(1.083)	(1.660)
Log Population		−1.750	−1.183	−.852	−.1.113	−.682	−2.028
		(.423)	(.393)	(.488)	(.489)	(.386)	(.470)
Economic Conditions							
Unemployment Rate, White-Collar Workers		.337	.379	.402	.415	.240	.870
		(.147)	(.150)	(.139)	(.147)	(.098)	(.136)
Unemployment Rate, Blue-Collar Workers		−.023	.028	−.063	−.085	−.204	−.270
		(.059)	(.069)	(.098)	(.084)	(.074)	(.099)
Unemployment Rate, Domestic Servants		.044	−.004	.114	.082	.078	.150
		(.122)	(.124)	(.107)	(.095)	(.054)	(.120)
Female Labor Force Participation Rate		.157	.057	.004	−.010	.025	.206
		(.065)	(.114)	(.118)	(.109)	(.069)	(.140)
Sectoral Composition of Workforce (in %)							
Manufacturing and Artisanry			−.117	−.064	−.109	−.022	−.066
			(.069)	(.128)	(.105)	(.063)	(.103)
Trade and Commerce			−.222	−.295	−.412	−.110	−.397
			(.077)	(.136)	(.142)	(.131)	(.145)
Services			.008	−.396	−.470	−.154	−.873
			(.075)	(.135)	(.124)	(.123)	(.242)
Domestic Labor			−.041	−.690	−.900	−2.105	−3.366
			(.304)	(2.161)	(1.655)	(1.676)	(2.558)
Occupational Composition (in %)							
White-Collar Workers				.107	.215	−.055	.518
				(.213)	(.220)	(.177)	(.274)
Civil Servants				.615	.861	.413	1.714
				(.245)	(.256)	(.203)	(.319)
Blue-Collar Workers				−.108	−.090	−.203	.345
				(.171)	(.134)	(.103)	(.194)
Domestic Servants				.838	.709	2.036	4.271
				(2.341)	(1.909)	(1.795)	(2.882)
Self-Employed				.125	.096	−.066	1.148
				(.324)	(.290)	(.211)	(.419)
Constant	39.256	23.208	−9.159	−12.285	91.185		−58.810
	(1.632)	(24.611)	(25.578)	(24.974)	(100.52)		(142.40)
Geographical Controls	No	No	No	No	Yes	Yes	Yes
Electoral District Fixed Effects	No	No	No	No	No	Yes	No
R-Squared	.584	.628	.644	.655	.672	.820	.400
Number of Observations	982	982	982	982	982	982	982

Note: Entries are coefficients and standard errors from estimating Equation (1) by weighted least squares. The dependent variable is a county's NSDAP vote share in the November elections of 1932. Weights correspond to the number of eligible voters in a given county. Heteroskedasticity robust standard errors are clustered by electoral district and reported in parentheses. The omitted category in Sectoral Composition of Workforce is Agriculture, and that in Occupational Composition is Helping Family Members. The set of Geographical Controls includes all geographical covariates listed in Table A.1 in the SI. In addition to the variables shown in the table, indicator variables for missing values on each covariate are also included in the regressions. See Appendix H in the SI for the precise definition and source of each variable.

But what **causal** quantity (if any) does β actually measure?

- τ_{ATE}?
- τ_{ATT}?
- τ_{ATC}?
- Something else?

Regression is a very widely used tool in the social sciences, and so it would be good to know what it is actually estimating!

(Re-)Running example

Do UN interventions Cause Peace?

Gilligan and Sergenti (2008) investigate whether UN peacekeeping operations have a causal effect on building sustainable peace after civil wars. They study 87 post-Cold-War conflicts, and evaluate whether peace lasts longer after conflict in 19 situations in which the UN had a peacekeeping mission compared to 68 situations where it did not.

- Y_i: Peace duration (measured in months)
- D_i: 1 if the UN intervened post-conflict, 0 otherwise
- X_{1i}: Region of conflict (categorical)
- X_{2i}: Democracy in the past (binary, based on polity)
- X_{3i}: Ethnic Fractionalization (continuous)

This Lecture

Regression and Causal Effects

Omitted Variable Bias

Non-linearity and Model Dependence

The Curse of Dimensionality

Non-standard Standard Errors

Regression and Causal Effects

Identification Assumption

1. *Potential outcomes independent of D_i given X_i:* $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i | X_i$ *("selection on observables" or "conditional independence assumption")*
2. $0 < \Pr(D = 1|X) < 1$ *for all* X *(common support)*

Identification Result

Given selection on observables we have

$$
\begin{aligned}
E[Y_{1i} - Y_{0i}|X_i] &= E[Y_{1i} - Y_{0i}|X_i, D_i = 1] \qquad \textbf{\textit{(CIA)}} \\
&= E[Y_{1i}|X_i, D_i = 1] - E[Y_{0i}|X_i, D_i = 1] \\
&= E[Y_{1i}|X_i, D_i = 1] - E[Y_{0i}|X_i, D_i = 0] \qquad \textbf{\textit{(CIA)}} \\
&= E[Y_i|X_i, D_i = 1] - E[Y_i|X_i, D_i = 0]
\end{aligned}
$$

*Implies that for any specific value for X_i, i.e. x_i, we can define the **conditional** average treatment effect (δ_x):*

$$\delta_x \equiv E[Y_i|X_i = x, D_i = 1] - E[Y_i|X_i = x, D_i = 0]$$

Identification under Selection on Observables

Identification Assumption

1. *Potential outcomes independent of D_i given X_i: $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i | X_i$ ("selection on observables" or "conditional independence assumption")*
2. $0 < \Pr(D = 1|X) < 1$ *for all X (common support)*

Identification Result

Therefore, under the common support condition and with a discrete X_i, we can calculate average effects of D_i on Y_i by taking weighted averages of δ_x:

$$
\begin{aligned}
\hat{\tau}_{ATE} &= \sum_x \delta_x P(X_i = x) \\
\hat{\tau}_{ATT} &= \sum_x \delta_x P(X_i = x | D_i = 1) \\
\hat{\tau}_{ATC} &= \sum_x \delta_x P(X_i = x | D_i = 0)
\end{aligned}
$$

i.e. where the weights are the distribution of X_i in the population ($\hat{\tau}_{ATE}$), treatment group ($\hat{\tau}_{ATT}$), and control group ($\hat{\tau}_{ATC}$).

Identification under Selection on Observables

This identification assumption and result is common to all the methods we will studied last week and will this week.

- Subclassification **(last week)**
- Matching **(last week)**
- Regression **(this week)**

These differ in

a. how we condition on X_i and
b. how we weight δ_x.

Regression, randomization, and τ_{ATE}

When we studied randomized experiments, we showed that the (bi-variate) linear regression model when D is binary

$$E[Y_i] = \alpha + \beta D_i$$

gives coefficient estimates under randomization equal to:

$$\begin{aligned} E[Y_i|D=0] &= E[Y_{0i}] = \alpha \\ E[Y_i|D=1] &= E[Y_{1i}] = \alpha + \beta \end{aligned}$$

and so:

$$\begin{aligned} E[Y_{1i}] - E[Y_{0i}] &= E[Y_i|D=1] - E[Y_i|D=0] \\ &= (\alpha + \beta) - (\alpha) \\ &= \beta \end{aligned}$$

$\Rightarrow$ Under randomisation: $\hat{\beta} = \tau_{ATE}$

Regression and "control"

- The typical introduction to regression views motivates it as a way to 'control' for potential confounding variables.
- What do we do when we include a control in regression?
  - We "hold it constant" while evaluating the relationship between D and Y
- This is also what we do in both subclassification and (exact) matching:
  1. fix the matches/subclasses
  2. calculate the mean difference for each match/class
  3. average the differences
- Regression essentially does the same thing, but does it in a single step, and the type of averaging is different

What have we been estimating all this time?

- Imagine we want to estimate the causal effect of D_i on Y_i, and we believe that D_i is independent of Y_{0i}, Y_{1i} conditional on X_i
- Does the following regression equation identify the causal effect? In other words, what does β_1 estimate?

$$Y_i = \alpha + \beta_1 D_i + \beta_2 X_i + \epsilon_i \qquad \textbf{(Long regression)}$$

- Textbook definition:

$$\beta_1 = \frac{Cov(Y_i, \tilde{D}_i)}{Var(\tilde{D}_i)}$$

where $\tilde{D}_i$ are the **residuals** from a regression of D_i on X_i

This means that estimating β_1 in the long regression is equivalent to:

1. Regressing D_i on X_i:

$$D_i = \pi_0 + \pi_1 X_i + e_i \qquad \textbf{(Treatment regression)}$$

2. Calculating the residuals from that regression:

$$\tilde{D}_i = D_i - (\hat{\pi}_0 + \hat{\pi}_1 X_i) \qquad \textbf{(Residual calculation)}$$

3. Regressing Y_i on those residuals (and nothing else):

$$Y_i = \alpha^* + \beta^* \tilde{D}_i + \epsilon^* \qquad \textbf{(Residual regression)}$$

$$\rightarrow \beta^* = \beta_1$$

Let's check using the peace data from last week:

```
# Long regression
coef(lm(dur ~ UN + ethfrac, data= peace))[2]
```

```
##       UN
## 35.24401
```

```
# Regression anatomy
## 1.
treatment_regression <- lm(UN ~ ethfrac, data = peace)
## 2.
treatment_residuals <- resid(treatment_regression)
## 3.
outcome_regression <- lm(dur ~ treatment_residuals, data = peace)

coef(outcome_regression)[2]
```

```
## treatment_residuals
##            35.24401
```

What does this tell us?

1. β_1 measures the relationship between Y_i and the part of D_i that is "not explained' ' by X_i **(i.e. the residuals)**
2. The part of D_i that is "not explained" by X_i is assumed to be independent of potential outcomes i.e. it is "as good as" random **(clear link to CIA)**

$\Rightarrow$ $\boxed{\text{Does this mean that } \beta_{\text{OLS}} \text{ is estimating } \tau_{\text{ATE}}\text{?}}$

Does β_{OLS} estimate τ_{ATE}?

Maybe, but maybe not...

1. We still have to believe that the treatment is as good as randomly assigned, conditional on covariates
   - **Conditional independence assumption**
   - i.e. all the discussion about confounding and post-treatment bias from last week applies
2. Even if we believe CIA holds, regression does something funny with the weighting step...

Regression as a weighted matching estimator

Selection on observables estimators

Last week we showed that both (exact) matching and subclassification calculate the ATE by taking weighted averages of δ_x:

$$\tau_{\text{ATE}} = \sum_x P(X_i = x)\delta_x$$

i.e. where the weights are the distribution of X_i in the population (τ_{ATE})

It can be shown (MHE, pp. 74 - 76) that the estimates for β from an OLS regression of Y on D and X have a similar form:

$$\beta_{\text{OLS}} = \sum_x \frac{Var[D_i = 1|X_i = x]P(X_i = x)}{\sum_x Var[D_i = 1|X_i = x]P(X_i = x)}\delta_x$$

Regression as a weighted matching estimator

$$\beta_{\text{OLS}} = \sum_x \frac{Var[D_i = 1|X_i = x]P(X_i = x)}{\sum_x Var[D_i = 1|X_i = x]P(X_i = x)} \delta_x$$

Therefore, regression implicitly gives higher weight to:

- Subclasses with more units (higher marginal probability, i.e. $P(X_i = x)$)
- Subclasses where the variance of the treatment is higher (i.e. $Var[D_i = 1|X_i = x]$)
- For a binary treatment, this will be subclasses with more equal numbers of treatment/control units

$\Rightarrow$ β_1 is an estimator "for the ATE but with supplemental conditional variance weighting." (Morgan and Winship, Ch 6)

Example (by hand)

$$
\begin{aligned}
w_{\text{ATE}} &\equiv P(X_i = x) \\
w_{\text{OLS}} &\equiv \frac{Var[D_i = 1|X_i = x]P(X_i = x)}{\sum_x Var[D_i = 1|X_i = x]P(X_i = x)}
\end{aligned}
$$

Region	Dem	N	δ_x	w_{ATE}	w_{OLS}	$\delta_x \cdot w_{ATE}$	$\delta_x \cdot w_{OLS}$	Diff
E. Eur	0	(5,5)	-29.2	0.14	0.2	-3.95	-5.85	1.9
E. Eur	1	(2,5)	66.8	0.09	0.11	6.32	7.65	-1.33
L. Am	0	(2,1)	144	0.04	0.05	5.84	7.69	-1.85
L. Am	1	(2,7)	5.1	0.12	0.12	0.62	0.64	-0.02
SS. Afr	0	(4,17)	27.9	0.28	0.26	7.92	7.24	0.68
SS. Afr	1	(2,12)	49	0.19	0.14	9.27	6.73	2.54
MeNa	0	(1,7)	123.7	0.11	0.07	13.37	8.67	4.7
MeNa	1	(1,1)	132	0.03	0.04	3.57	5.29	-1.72

$$
\begin{aligned}
\tau_{\text{ATE}} &= 42.96 \\
\beta_{\text{OLS}} &= 38.06
\end{aligned}
$$

So OLS gives something a bit like the ATE, but not quite...

Regression vs Matching vs Subclassification

- In practice, regression will often give very similar estimates of the ATE to matching and subclassification.
- The UN peace example, controlling for region and democratic history:

	ATE	ATT	ATC
Sub-classification	42.96	39.53	44.14
Matching (exact)	39.45	33.05	41.66
Regression	38.07		

When will β_{OLS} be an unbiased estimator for τ_{ATE}?

1. When $P(D_i = 1|X_i = x) = P(D_i = 1) \forall x$
   - treatment probability is the same for everyone
   - conditional variance is the same for everyone
   - **(e.g. in an experiment)**
2. When $\delta_x = \tau_{\text{ATE}} \forall x$
   - treatment effects are the same for each subclass
   - **(i.e. effects are homogenous)**

Note: We are still assuming conditional independence holds!

Omitted Variable Bias

Omitted variable bias

- The more typical implicit link made between regression and causality is via the idea of omitted variables.
- Consider two regression models:

$$\begin{aligned} \text{Long:} \quad Y_i &= \alpha^l + \beta_1^l X_{1i} + \beta_2^l X_{2i} + \epsilon^l \\ \text{Short:} \quad Y_i &= \alpha^s + \beta_1^s X_{1i} + \epsilon^s \end{aligned}$$

- We also have an 'auxiliary' regression:

$$\text{Auxiliary:} \quad X_{2i} = \pi_0 + \pi_1 X_{1i} + \eta_i$$

Omitted variable bias

= The bias that results from failing to control for X_{2i} in the short regression.

β^s vs β^l

- If we ignored X_{2i} and just estimated the short regression, what does β_1^s identify?
- With a little bit of work **(Mastering 'Metrics, p. 93)** we can see:

$$\beta_1^s = \beta_1^l + \beta_2^l \pi_1$$

  - $\beta_1^l \rightarrow$ the coefficient of X_{1i} on Y_i in the long regression
  - $\beta_2^l \rightarrow$ the coefficient of X_{2i} on Y_i in the long regression
  - $\pi_1 \rightarrow$ the coefficient of X_{1i} on X_{2i} in the 'auxiliary' regression

"Short equals long plus the effect of omitted times the regression of omitted on included." (MHE, p. 60)

$$\text{OVB:} \qquad \beta_1^s - \beta_1^l = \beta_2^l \pi_1$$

β^s vs β^l in practice

Is it true that "short equals long plus the effect of omitted times the regression of omitted on included"?

```
long <- lm(dur ~ UN + ethfrac,peace)
short <- lm(dur ~ UN,peace)
aux <- lm(ethfrac ~ UN,peace)

# Effect of UN in short regression
coef(short)[2]
```

```
##       UN
## 40.80263
```

```
# Long + effect of omitted*reg of omitted on included
coef(long)[2] + coef(long)[3]*coef(aux)[2]
```

```
##       UN
## 40.80263
```

Direction of OVB

$$\text{OVB:} \qquad \beta_1^s - \beta_1^l = \beta_2^l \pi_1$$

- The difficulty is that we rarely know the values for either β_2^l or π_1 and so we can't isolate β_1^l.
- The formula does however help to describe the **possible direction of bias**:

	$\beta_2^l < 0$	$\beta_2^l > 0$
$\pi_1 < 0$	$OVB > 0$	$OVB < 0$
$\pi_1 > 0$	$OVB < 0$	$OVB > 0$

Note also that:

- If $\pi_1 = 0$, then $OVB = 0$, and
- If $\beta_2^l = 0$, then $OVB = 0$

Lessons from the OVB formula

Omitting X **causes bias** in our estimate of ATE if and only if **both** the following hold

1. X is related to the treatment, conditional on other covariates
   - e.g. $\pi_1 \neq 0$
   - $\rightarrow$ no need to control for covariates when D is randomly assigned
2. X is related to the outcome, conditional on other covariates
   - e.g. $\beta_2^l \neq 0$
   - $\rightarrow$ no need to control for covariates that don't effect Y

Should I use regression?

Yes.

- Computational simplicity
- Many forms, very flexible
- Easy statistical inference
- *Easy to include continuous treatments*

No.

- Not as clearly linked to the CIA
- Do we care about the conditional-variance ATE?

My view: Yes, most of the time. The flexibility and simplicity of inference are very helpful, and it's close enough to what we care about that it's fine to use, most of the time.

Non-linearity and Model Dependence

Let's return to our civil war example. We would like to know the effect of UN interventions on peace duration:

- Y_i: Peace duration (`dur`, measured in months)
- D_i: UN intervention (`UN`, binary)

A colleague suggests that an important confounder might be the duration of the prior civil war (X_i: `lwdurat`), because:

- $\pi_1 \neq 0 \rightarrow$ the length of the previous war is probably associated with whether the UN intervenes
- $\beta_2^l \neq 0 \rightarrow$ the length of the previous war is probably associated with how long peace lasts

War duration and peace duration

- Is civil war duration **linearly** associated with peace duration?
- We might not want to use a straight line to model this relationship!

One way of modelling non-linear relationships is to include polynomial functions of explanatory variables in our model:

Polynomial models

Polynomial models take the following form:

$$\begin{array}{rrcl} \text{Linear:} & E[Y_i] & = & \alpha + \beta_1 X_1 \\ \text{Quadratic:} & E[Y_i] & = & \alpha + \beta_1 X_1 + \beta_2 X_1^2 \\ \text{Cubic:} & E[Y_i] & = & \alpha + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 \end{array}$$

Where X_1^2 is just $X_1 * X_1$ and X_1^3 is just $X_1 * X_1 * X_1$, and so on.

In theory we can keep adding polynomial terms to make our model more flexible, but it gets harder to interpret!

Polynomial functions of explanatory variables

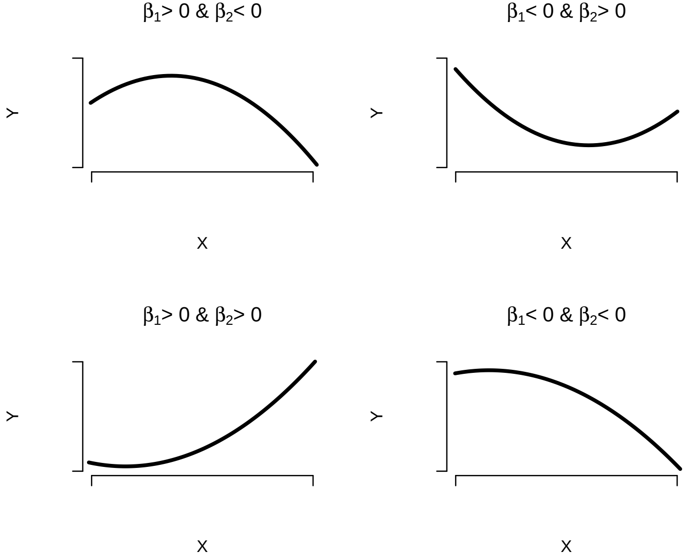
Why do polynomial terms allow for non-linear relationships?

- When we include a quadratic term in the model, we are essentially including an interaction term
  - i.e. the interaction between X_1 and itself (because $X_1 \cdot X_1 = X_1^2$)
  - This implies that the association between X_1 and Y will depend on the specific value of X_1 where we evaluate the relationship
  - $\rightarrow$ the effect of a one-unit change in X_1 will depend on the value of X_1 we are changing

Interpreting polynomial coefficients is somewhat difficult:

- It is no longer possible to hold constant all other variables
  - i.e. If you increase X_1 by one-unit, then you also increase X_1^2
  - We can still say something by looking at the sign of β_{X^2}
- In general it is much more straightforward to produce fitted value plots to describe the relationship between X and Y

For the model $E[Y] = \alpha + \beta_1 X_1 + \beta_2 X_1^2$:



Polynomial regression in R

In R we can include polynomial transformation of our X variables directly into the model formula:

```
linear_model <- lm(dur ~ lwdurat, data = peace)
quadratic_model <- lm(dur ~ lwdurat + I(lwdurat^2),
                      data = peace)
```

Polynomial regression output

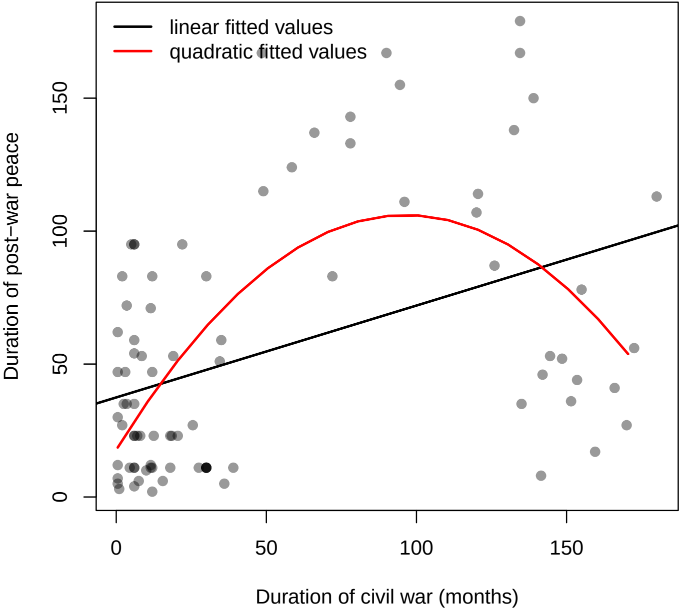
Regression output

	(1)	(2)
$lwdurat$	0.35^{***} (0.08)	1.83^{***} (0.33)
$lwdurat^2$		-0.01^{***} (0.002)
Intercept	37.44^{***} (6.33)	17.69^{**} (7.08)
Observations	87	87
R^2	0.17	0.34

The coefficients are hard to interpret, but we can see that the quadratic term is significant. What does this mean?

- The null hypothesis is that the relationship between X and Y is linear
- We can reject this null: there is evidence of non-linearity here
- We should also note that the model fit improves

Polynomial regression visualization

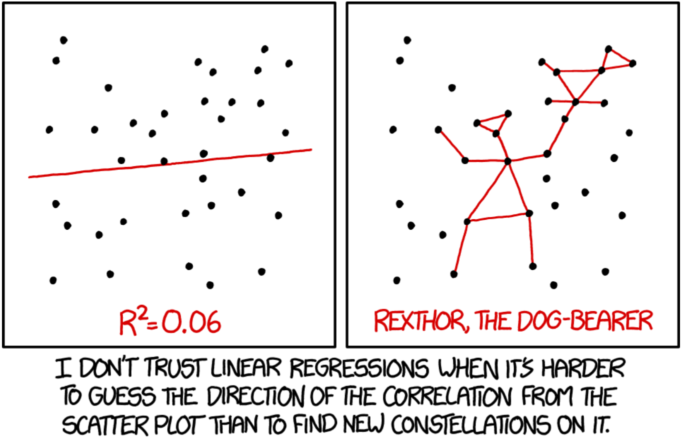


Why should we care about non-linearity?

There are 2 broad motivations for thinking about non-linearity:

1. Not all relationships are linear!
   - Regression is a model that "by default" estimates linear relationships
   - Sometimes, like here, linearity is not a good approximation of the true relationship.
   - In these cases, we may want to specify a more flexible model to capture more of reality
2. Mis-specifying a the non-linear relationship between a control variable and the outcome can lead to biased treatment effects
   - This is known as **model dependence**

Why should we care about non-linearity?



Non-linearity and model dependence

When using regression as a tool to estimate treatment effects, we also therefore need to decide *how* to control for confounders:

```
linear_control <- lm(dur ~ UN + lwdurat, data = peace)
quadratic_control <- lm(dur ~ UN + lwdurat + I(lwdurat^2),
                        data = peace)
```

	(1)	(2)
UN	37.46^{***} (10.83)	24.76^{**} (10.59)
$lwdurat$	0.33^{***} (0.08)	1.59^{***} (0.34)
$lwdurat^2$		-0.01^{***} (0.002)
Intercept	29.83^{***} (6.35)	15.86^{**} (6.94)
Observations	87	87
R^2	0.27	0.38

Note: The UN coefficient is very different in the two models!

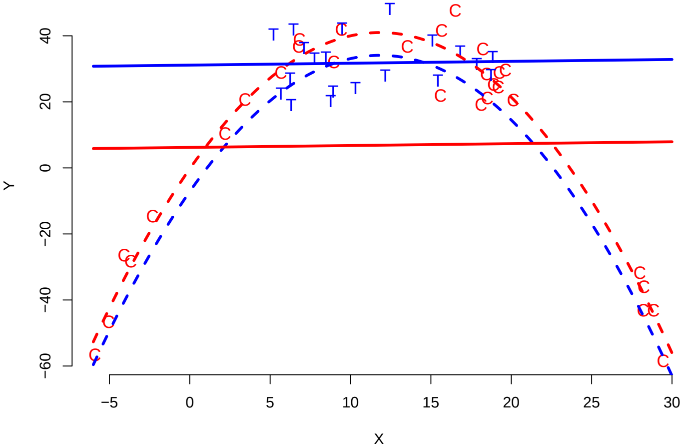
Definition

Model dependence exists when our estimates depend on specific modeling assumptions and where different specifications can yield very different causal inferences.

What can we do about this problem?

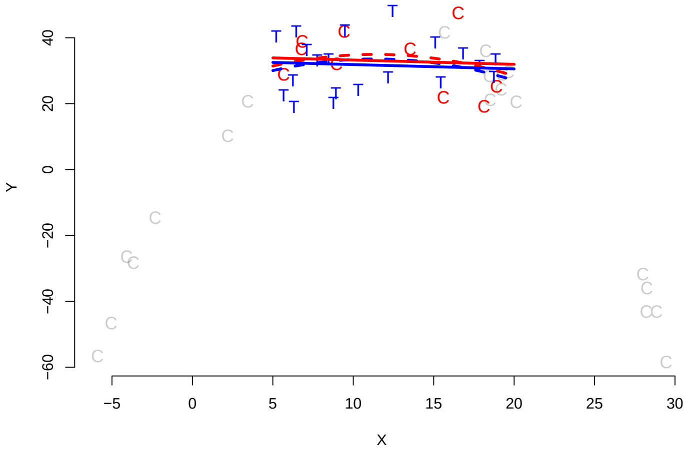
- One common approach is to use **matching** as a preprocessing tool to reduce model dependence (see Ho et. al, 2007).
- This is exactly what we have been doing with the `MatchIt` package!

Regression and model dependence



- Linear control for X: $\hat{\beta}_1 > 0$
- Quadratic control for X: $\hat{\beta}_1 < 0$

Regression, *matching* and model dependence



- Post-matching: $\hat{\beta}_{\text{linear}} \approx \hat{\beta}_{\text{quadratic}} \approx 0$

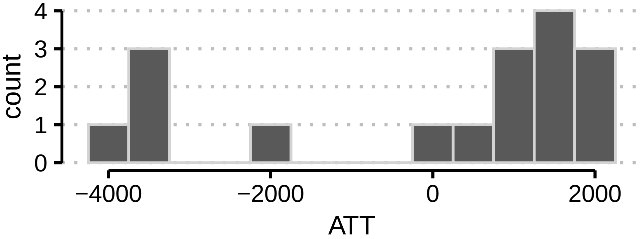
Using matching to ensure that common support holds can make our parametric estimates less model dependent.

Implications:

- Estimates will be less sensitive to small changes in modelling choices that are particularly common in regression analysis.
- We will frequently lack common support for **both** treatment and control observations, which we then discard.
- This has consequences for the interpretation of estimated treatment effects.
  - Our estimates will be $\hat{\tau}_{ATE}$ or $\hat{\tau}_{ATT}$ or $\hat{\tau}_{ATC}$ only *for those units for which the common support assumption holds*

Model dependence in matching

- Matching is not free from model dependence either!
- We saw last week how small matching decisions made quite a bit of difference
- The results from week's seminar question 2.3 illustrate this too:
  - Almost all included `age`, `married`, `black`, `hisp`
  - Then either one or two of `no degree`, `educcat`, or `educ`
  - The big difference is due to `re74` and `re75`
  - Most did 1:1, with replacement and mahalanobis distance - but these choices made less difference than the earnings variables



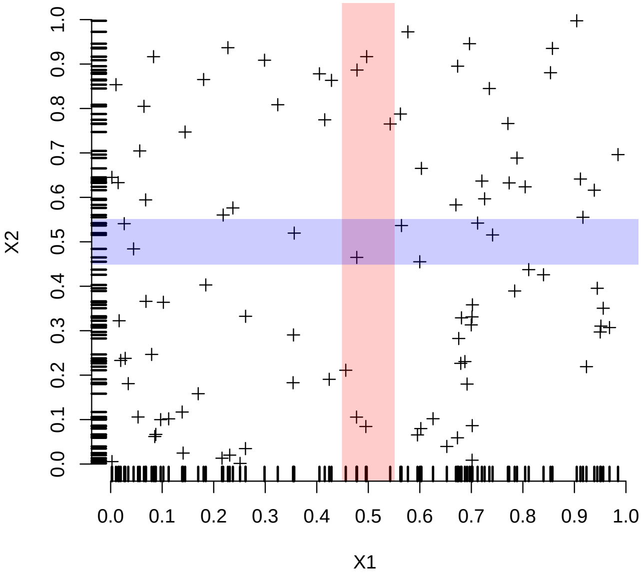
The Curse of Dimensionality

- Matching, regression, and subclassification all rely on the idea that we can make comparisons between treatment and control units that have otherwise similar X variables.
- Often, we will be able to come up with many possible confounding factors that we might want to condition upon.

Problem: The curse of dimensionality

The total quantity of data 'near' any given point in X falls off very quickly when the dimensionality of X increases.

Curse of dimensionality



Curse of dimensionality

- The average distance from the nearest observation increases very fast as we add explanatory variables
  - I.e. the data becomes 'sparse' in X
- Increasing the sample size helps, but not much!
  - To maintain the same average distance to nearest observations when going from 1 to 2 explanatory variables often requires many **thousands** more observations
  - To get the same average distance to the nearest observation that is acheived for 1 explanatory variable with 32 observations requires over 1000 observations with 2 explanatory variables
- **Implication:** Adding more covariates may make matches more "appropriate", but also makes them far harder to make.

Implications of the curse of dimensionality

- **Matching:**
  - Exact matching: very few exact matches
  - Nearest neighbour: if more distant matches are less reliable, adding X's might make 'nearest' matches poor control choices
- **Subclassification:**
  - Many empty cells, or cells with only treatment/control units
- **Regression:**
  - More reliance on model, and thus increased threat of model dependent results

Generally: As dimensionality increases, restricting to observations with reasonable matches will lead to unrepresentative estimates

Non-standard Standard Errors

Errors and residuals, re-cap

Recall the linear regression model:

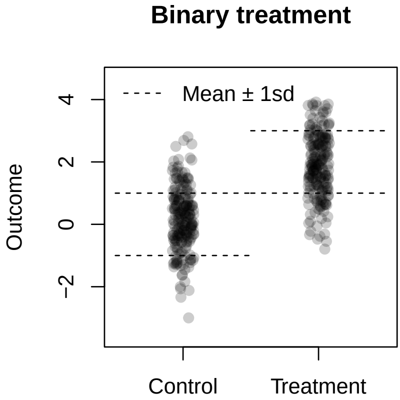
$$Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i$$

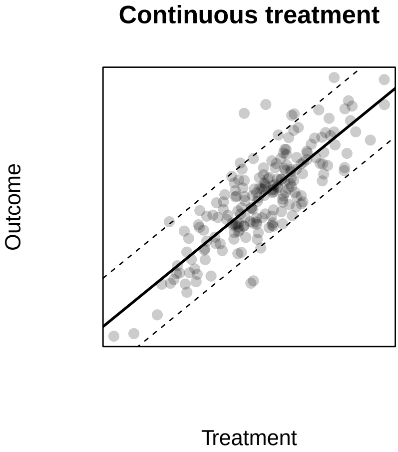
Most regression software by default makes two restrictive assumptions about the error term, ϵ:

1. Errors are independent and identically distributed for each observation **(iid)**
2. Errors have equal variance for all values of X **(homoskedasticity)**

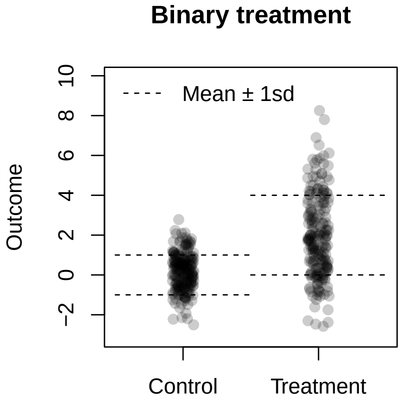
When either of these things fail, our standard errors will be wrong. Here we will discuss two potential failures of these assumptions.

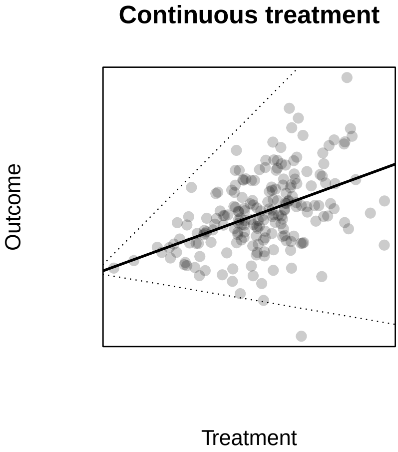
Homoskedastic errors





Heteroskedastic errors





The good news:

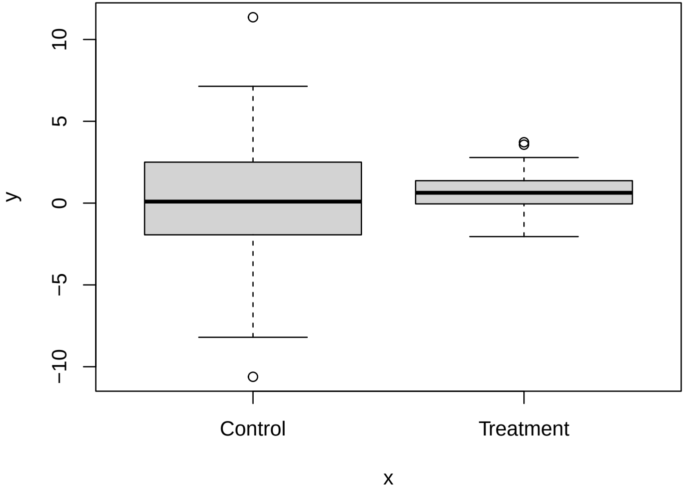
- Whether the errors are homoskedastic or heteroskedastic, $\hat{\beta}$ is both unbiased and consistent

The bad news:

- If the homoskedasticity assumption is violated:
  - t-statistics do not have a standard normal distribution
  - Conventional standard errors will be too small
  - Hypothesis tests will reject the null hypothesis too often
  - Confidence intervals will be too narrow
- Heteroskedasticity can lead to standard errors that are too small *or* too large.
  - But we generally care less about *over*estimating the standard error.

Heteroskedasticity in t-tests and regression

Let's consider an experiment where the variance of the outcome is different in the treatment and control groups:



Heteroskedasticity in t-tests and regression

By default, the `lm()` function in R assumes homoskedastic errors:

```
ols_mod <- lm(y ~ x)
summary(ols_mod)
```

```
...
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.2072     0.1913   1.083    0.279
## x             0.4462     0.2209   2.021    0.044 *
...
```

But the `t.test()` function does not:

```
t.test(y[x==1],  y[x==0])
```

```
...
## data:  y[x == 1] and y[x == 0]
## t = 1.2896, df = 104.82, p-value = 0.2
## alternative hypothesis: true difference in means is not equal to 0
...
```

Note: The regression provides the wrong conclusion!

Heteroskedasticity correction in R

```r
library(lmtest)
library(estimatr)

# Use lm_robsust from the estimatr package to tell R to calculate
# heteroskedasticity-robust SEs ("HC3")
robust_ols_mod <- lm_robust(y ~ x, se_type = "HC3")
coeftest(robust_ols_mod)
```

```
...
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.20715    0.34281  0.6043   0.5460
## x            0.44624    0.34774  1.2833   0.2001
...
```

Note: The standard error, t-statistic and p-value are now correct.

Clustered data

Another way in which normal standard errors can be wrong is when we have **clustered** data.

Examples:

- An experiment where *villages* are selected into treatment/control but the outcome is measured at the *household* level
- An observational study where we care about the effects of *class size* but we measure *individual student* outcomes

Key: Always ask yourself: at what level was the treatment assigned?

Clustering example

The STAR Experiment

The STAR project was a randomized experiment designed to test the causal effects of class sizes on learning. **Classes** in Tennessee schools were randomly assigned either to regular sized classes (22-25 students, the **control group**) or to smaller classes (15-17 students, the **treatment group**). We observe student outcomes at the **individual level**.

- Y (Dependent variable): grade of **student** on a standardised math test (0 to 100)
- X (Independent variable): size of **class** (`TRUE` = student in small class, `FALSE` = student in regular sized class)

Why does clustering affect standard errors?

Imagine we have a regression like:

$$Y_{i(g)} = \alpha + \beta_1 X_g + \epsilon_{i(g)}$$

where X_g is a covariate that only varies at the group level.

- Normal standard errors are calculated assuming that errors (ϵ) are uncorrelated across units
- This is clearly not the case here!
  - Students in the same class will have similar grades because of other factors (teacher quality; time of day; etc)
- When errors are correlated within groups, the normal standard errors will be too small
- This will be particularly bad when the number of groups is small

STAR example in R

Let's run the STAR regression:

```
linear_model <- lm(grade ~ small_class, data = star)
summary(linear_model)
```

```
...
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  45.8255     0.1869  245.20  < 2e-16 ***
## small_class   2.2234     0.3410    6.52 7.61e-11 ***
...
```

- The regression *coefficient* is an **unbiased** estimate of the ATE of small classes. Why? Randomisation!
- But, although the *errors* are almost surely correlated within class, we are treating them as independent, meaning that they are likely too small.

Cluster-robust standard errors in R

One solution to this problem is to use cluster-robust standard errors.

```r
linear_model_cl <- lm_robust(grade ~ small_class, data = star,
                             clusters = schidkn,
                             se_type = "CR2") # the default
coeftest(linear_model_cl)
```

```
...
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 45.82546    0.71378 64.2011 < 2.2e-16 ***
## small_class  2.22339    0.61902  3.5918 0.0003311 ***
...
```

Note: Here, the cluster-robust errors are twice as large as the regular standard errors, but the conclusion remains the same. This will not always be the case...

Cluster-robust standard errors

- Normal SEs are partly determined by the sample size, N
  - As $N \uparrow$, $\widehat{SE(\beta)} \downarrow$
- Clustered SEs are more sensitive to the number of clusters, G, than they are to N.
  - As $G \uparrow$, $\widehat{SE(\beta)}_{\text{Clustered}} \downarrow$

Implications:

1. Collecting more data only helps if you are collecting from new groups
2. $\widehat{SE(\beta)}_{\text{Clustered}}$ will perform poorly when the number of clusters is small (< 30).

Conclusion

- Matching or Regression (or Subclassification)?
  - **My view:** Differences between estimation strategies are far less important than the data you have collected
- Simply, for selection on observables to hold, you need good observables!
- Don't spend ages trying to persuade us that your new matching estimator is really great. Instead:
  - Think hard about which variables are crucial to condition upon
  - Collect better data relevant to these confounders
  - Find settings where there are **very** good covariates **(RDD is going to be an example of this)**
  - Find setting where confounders are less important **(experiments, natural experiments, diff-in-diff etc)**