Gov 50: 14. Regression and Causality (II)

Matthew Blackwell

Harvard University

Fall 2018

1. Today's agenda

2. Heterogeneous treatment effects

3. Non-linear relationships

4. Causality and regression wrap up

1/ Today's agenda

Where are we? Where are going?

- Last couple of lectures:
  - Learning about how to use regression to predict and estimate causal effects.
- Today:
  - More interaction terms and heterogeneous treatment effects.
  - Modeling non-linear relationships.
- HW3 due tonight.

2/ Heterogeneous treatment effects

Social pressure experiment

- We'll look at the Michigan experiment that was trying to see if social pressure affects turnout.
- Load the data and create an age variable:

```
social <- read.csv("data/social.csv")
social$age <- 2006 - social$yearofbirth
summary(social$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    20.0    41.0    50.0    49.8    59.0   106.0
```

```
social.neighbors <- subset(social,
                           neighbors == 1 | control == 1)
```

Heterogeneous effects

- Last time:
  - How does the effect of the Neighbors mailer vary by previous voter versus non-voters?
  - Used an interaction term to assess **effect heterogeneity** between groups.
- What if we want to know how the effect of the Neighbors mailer varies by age?
  - Not just two groups, but a continuum of possible age values.
- Remarkably, the same **interaction term** will work here too!

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_1 \text{age}_i + \widehat{\beta}_2 \text{neighbors}_i + \widehat{\beta}_3 (\text{age}_i \times \text{neighbors}_i)$$

Predicted values from non-interacted model

- Let $X_i = \text{age}_i$ and $Z_i = \text{neighbors}_i$:

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i$$

	Control ($Z_i = 0$)	Neighbors ($Z_i = 1$)
25 year-old ($X_i = 25$)	$\widehat{\alpha} + \widehat{\beta}_1 \cdot 25$	$\widehat{\alpha} + \widehat{\beta}_1 \cdot 25 + \widehat{\beta}_2$
26 year-old($X_i = 26$)	$\widehat{\alpha} + \widehat{\beta}_1 \cdot 26$	$\widehat{\alpha} + \widehat{\beta}_1 \cdot 26 + \widehat{\beta}_2$

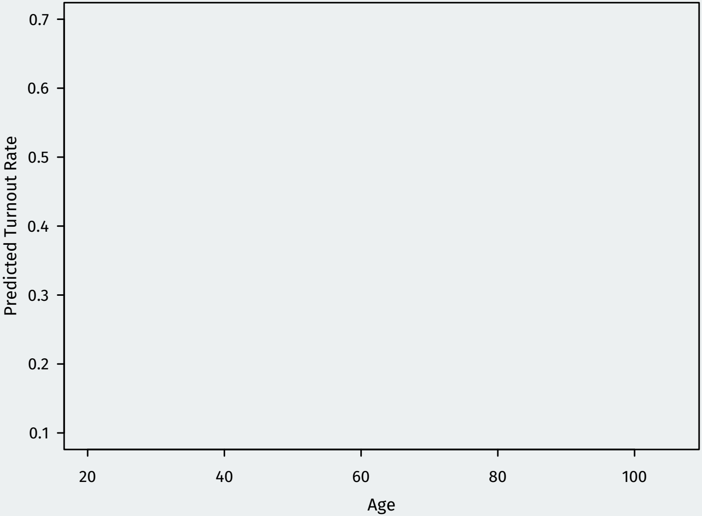
- Effect of Neighbors for a 25 year-old:

$$(\widehat{\alpha} + \widehat{\beta}_1 25 + \widehat{\beta}_2) - (\widehat{\alpha} + \widehat{\beta}_1 25) = \widehat{\beta}_2$$

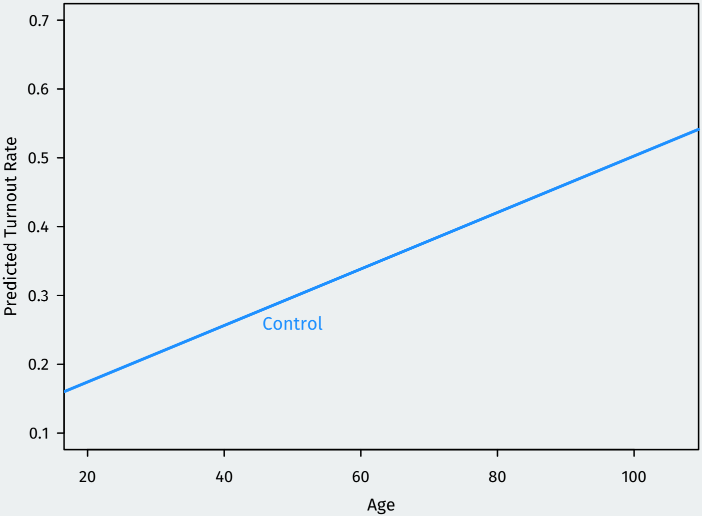
- Effect of Neighbors for a 26 year-old:

$$(\widehat{\alpha} + \widehat{\beta}_1 26 + \widehat{\beta}_2) - (\widehat{\alpha} + \widehat{\beta}_1 26) = \widehat{\beta}_2$$

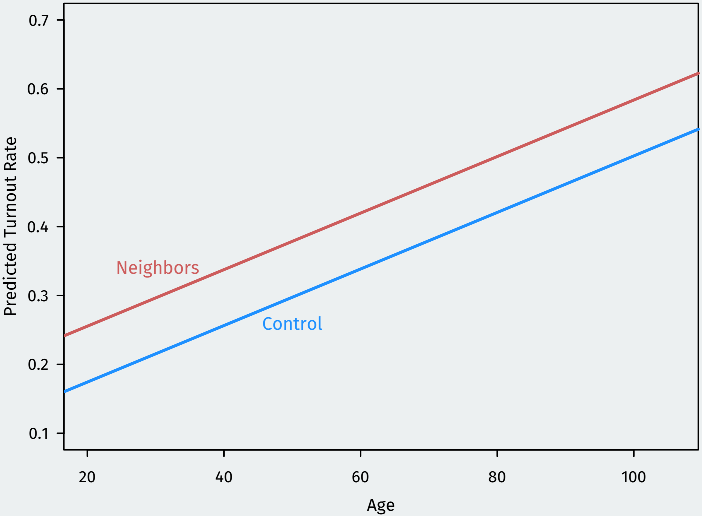
Visualizing the regression



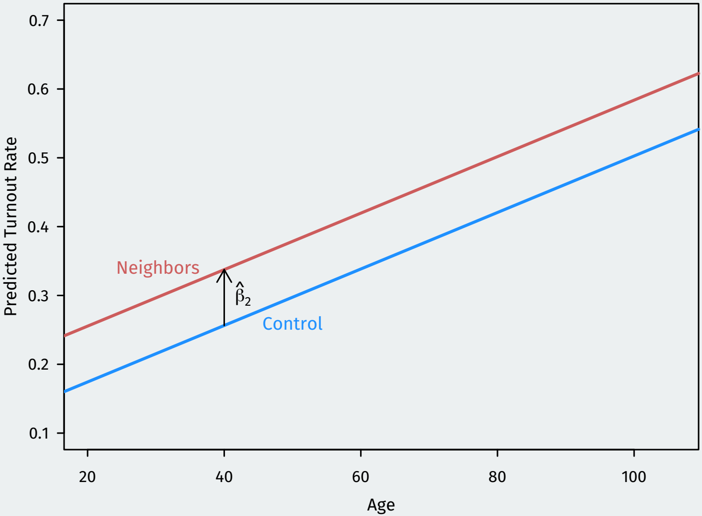
Visualizing the regression



Visualizing the regression



Visualizing the regression



Predicted values from interacted model

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

	Control ($Z_i = 0$)	Neighbors ($Z_i = 1$)
25 year-old ($X_i = 25$)	$\widehat{\alpha} + \widehat{\beta}_1 \cdot 25$	$\widehat{\alpha} + \widehat{\beta}_1 \cdot 25 + \widehat{\beta}_2 + \widehat{\beta}_3 \cdot 25$
26 year-old($X_i = 26$)	$\widehat{\alpha} + \widehat{\beta}_1 \cdot 26$	$\widehat{\alpha} + \widehat{\beta}_1 \cdot 26 + \widehat{\beta}_2 + \widehat{\beta}_3 \cdot 26$

- Effect of Neighbors for a 25 year-old:

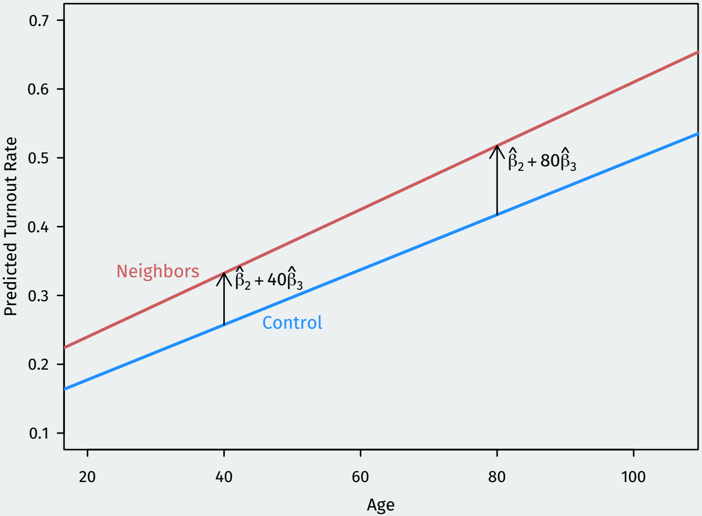
$$(\widehat{\alpha} + \widehat{\beta}_1 25 + \widehat{\beta}_2 + \widehat{\beta}_3 \cdot 25) - (\widehat{\alpha} + \widehat{\beta}_1 25) = \widehat{\beta}_2 + \widehat{\beta}_3 \cdot 25$$

- Effect of Neighbors for a 26 year-old:

$$(\widehat{\alpha} + \widehat{\beta}_1 26 + \widehat{\beta}_2 + \widehat{\beta}_3 \cdot 26) - (\widehat{\alpha} + \widehat{\beta}_1 26) = \widehat{\beta}_2 + \widehat{\beta}_3 \cdot 26$$

- Effect of Neighbors for a x year-old: $\widehat{\beta}_2 + \widehat{\beta}_3 \cdot x$

Visualizing the interaction



Interpreting coefficients

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_1 \text{age}_i + \widehat{\beta}_2 \text{neighbors}_i + \widehat{\beta}_3 (\text{age}_i \times \text{neighbors}_i)$$

- $\widehat{\alpha}$: average turnout for 0 year-olds in the control group.
- $\widehat{\beta}_1$: slope of regression line for age in the control group.
- $\widehat{\beta}_2$: average effect of Neighbors mailer for 0 year-olds.
- $\widehat{\beta}_3$: change in the **effect** of the Neighbors mailer for a 1-year increase in age.
  - Effect for x year-olds: $\widehat{\beta}_2 + \widehat{\beta}_3 \cdot x$
  - Effect for $(x + 1)$ year-olds: $\widehat{\beta}_2 + \widehat{\beta}_3 \cdot (x + 1)$
  - Change in effect: $\widehat{\beta}_3$

Interactions in R

- You can use the `:` way to create interaction terms like last time:

```
int.fit <- lm(primary2006 ~ age + neighbors + age:neighbors,
              data = social.neighbors)
coef(int.fit)
```

```
##   (Intercept)           age     neighbors
##      0.097473      0.003998      0.049829
## age:neighbors
##      0.000628
```

- Or you can use the `var1 * var2` shortcut, which will add both variable and their interaction:

```
int.fit2 <- lm(primary2006 ~ age * neighbors, data = social.neighbors)
coef(int.fit2)
```

```
##   (Intercept)           age     neighbors
##      0.097473      0.003998      0.049829
## age:neighbors
##      0.000628
```

General interpretation of interactions

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_1 X_i + \widehat{\beta}_2 Z_i + \widehat{\beta}_3 X_i Z_i$$

- $\widehat{\alpha}$: average outcome when X_i and Z_i are 0.
- $\widehat{\beta}_1$: average change in Y_i of a one-unit change in X_i when $Z_i = 0$
- $\widehat{\beta}_2$: average change in Y_i of a one-unit change in Z_i when $X_i = 0$
- $\widehat{\beta}_3$ has two equivalent interpretations:
  - Change in the effect/slope of X_i for a one-unit change in Z_i
  - Change in the effect/slope of Z_i for a one-unit change in X_i
- These hold no matter what types of variables they are!

3/ Non-linear relationships

Linear regression are linear

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_1 X_i$$

- Standard linear regression can only pick up **linear** relationships.
- What if the relationship between X_i and Y_i is non-linear?

Adding a squared term

- If we want to allow for non-linearity in age, we can add a squared term to the regression model:

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_1 \text{age}_i + \widehat{\beta}_2 \left(\text{age}_i^2\right)$$

- We are now fitting a **parabola** to the data.
- In R, we can add a squared term, but we need to wrap it in `I()`:

```
fit.sq <- lm(primary2006 ~ age + I(age^2), data = social)
coef(fit.sq)
```

```
## (Intercept)          age     I(age^2)
##  -0.0816804   0.0122736   -0.0000808
```

- $\widehat{\beta}_2$: how the effect of age increases as age increases.

Predicted values from lm()

- We can get predicted values out of R using the `predict()` function:

```
predict(fit.sq, newdata = list(age = c(20, 21, 22)))
```

```
##     1     2     3
## 0.131 0.140 0.149
```

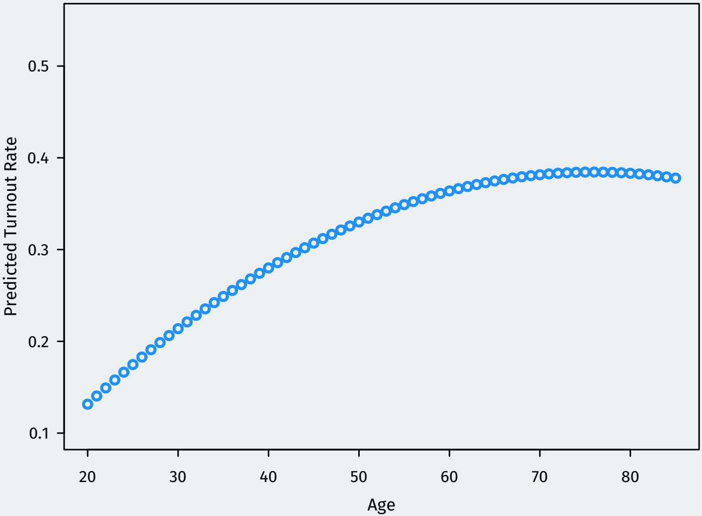
- Create a vector of ages to predict and save predictions:

```
age.vals <- 20:85
age.preds <- predict(fit.sq, newdata = list(age = age.vals))
```

- Plot the predictions:

```
plot(x = age.vals, y = age.preds, ylim = c(0.1, 0.55),
     xlab = "Age", ylab = "Predicted Turnout Rate",
     col = "dodgerblue", lwd = 2)
```

Plotting predicted values

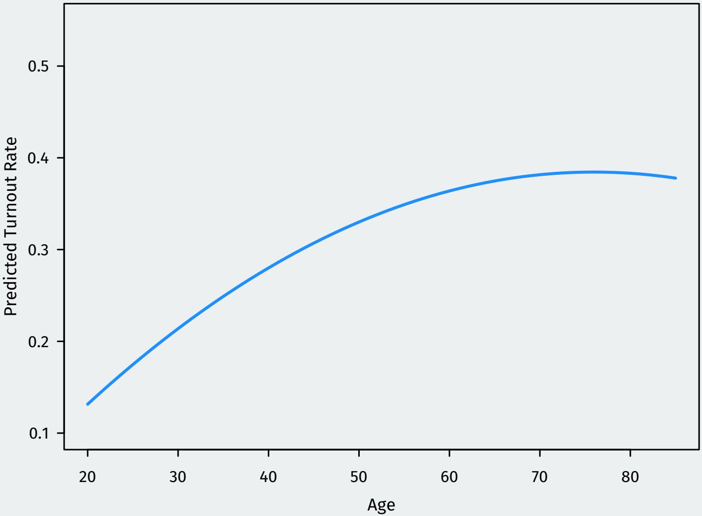


Plotting lines instead of points

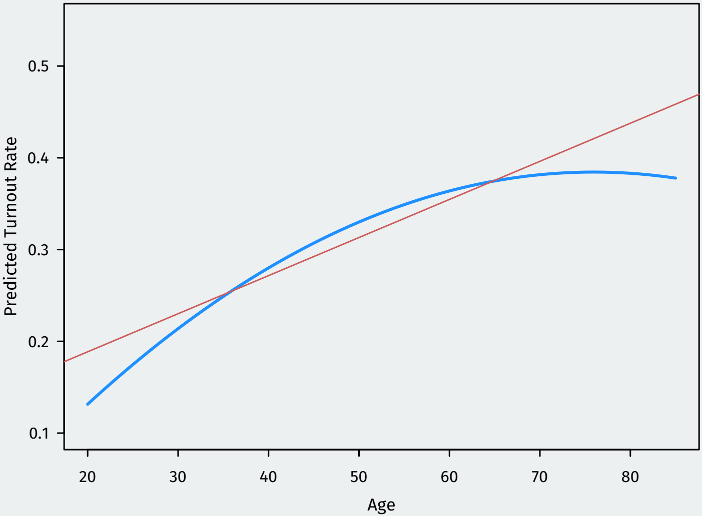
- If you want to connect the dots in your scatterplot, you can use the `type = "l"` ("line" type):

```
plot(x = age.vals, y = age.preds, ylim = c(0.1, 0.55),
     xlab = "Age", ylab = "Predicted Turnout Rate",
     col = "dodgerblue", lwd = 2, type = "l")
```

Plotting predicted values



Comparing to linear fit



Diagnosing nonlinearity

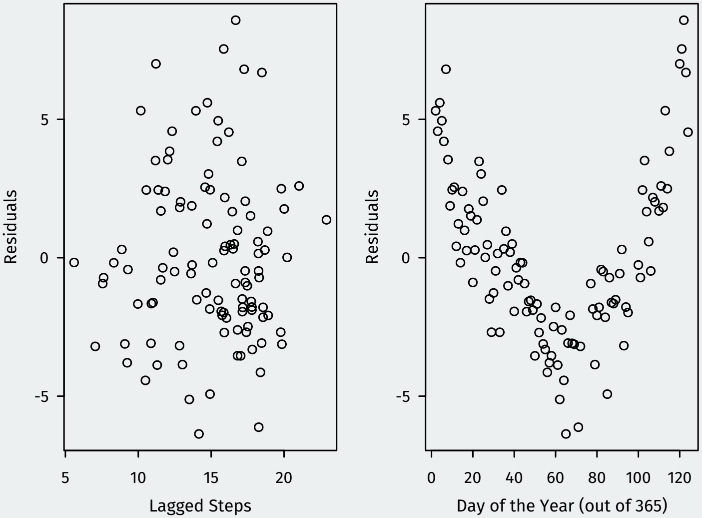
- Diagnosing nonlinearity can be easy with a single variable: just plot the scatterplot.
- With multiple variables, harder to diagnose.
- One useful tool: plotting residuals on y-axis versus variables with suspected nonlinearities on the x-axis.
- Example: my weight again

```
health <- read.csv("data/health2017.csv")
w.fit <- lm(weight ~ steps.lag + dayofyear, data = health)
```

Residual plot

```
plot(health$steps.lag, residuals(w.fit),
     xlab = "Lagged Steps", ylab = "Residuals")
plot(health$dayofyear, residuals(w.fit),
     xlab = "Day of the Year (out of 365)", ylab = "Residuals")
```

Residual plot



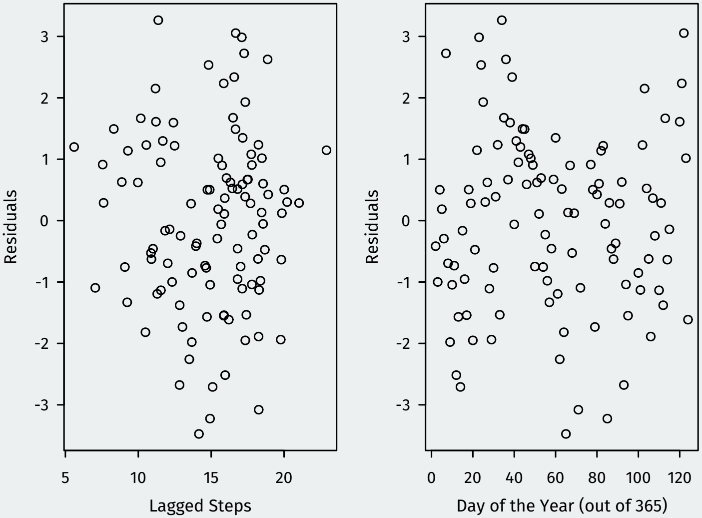
Add a squared term for a better fit

```r
w.fit.sq <- lm(weight ~ steps.lag + dayofyear + I(dayofyear^2),
               data = health)
coef(w.fit.sq)
```

```
##    (Intercept)      steps.lag      dayofyear
##       177.4679         0.0521        -0.4439
## I(dayofyear^2)
##         0.0024
```

```r
plot(health$steps.lag, residuals(w.fit.sq),
     xlab = "Lagged Steps", ylab = "Residuals")
plot(health$dayofyear, residuals(w.fit.sq),
     xlab = "Day of the Year (out of 365)", ylab = "Residuals")
```

Residual plot, redux



4/ Causality and regression wrap up

Regression and causality

- When can we interpret a regression coefficient causally?
- Randomized control trial:
  - Coefficient on binary treatment is estimate of the SATE
  - True even if we add other independent variables.
  - Other independent variables **not causal**
- Observational studies:
  - Can only interpret coefficients as causal effect **if we have controlled for all confounders** as additional independent variables.
  - Confounders: other variables that cause both treatment and outcome.
  - Before/after and diff-in-diff designs can be implemented with regression, too.

On deck

- Everything up to this point: getting estimates.
- How much uncertainty should we have about our estimates?
  - Could we have seen this regression coefficient by chance alone?
- Next part of class: quantifying uncertainty.
  - First stop: probability, the mathematical language of uncertainty.