Interactions and nonlinear relationships

Seung-Ho An, University of Arizona

Agenda

Varying effects by groups

Interactions with continous variables

Nonlinear relationships

1. Varying effects by groups

Heterogeneous treatment effects

- **Heterogeneous treatment effects:** effect varies across groups

- **Heterogeneous treatment effects:** effect varies across groups
  - Average effect of a drug is 0, but $+$ for `men` and $-$ for `women`

- **Heterogeneous treatment effects:** effect varies across groups
  - Average effect of a drug is 0, but $+$ for `men` and $-$ for `women`
  - Important questions for determining who should receive treatment

- **Heterogeneous treatment effects:** effect varies across groups
  - Average effect of a drug is 0, but $+$ for `men` and $-$ for `women`
  - Important questions for determining who should receive treatment
- Social pressure experiment:
  - primary2004 whether the person voted in 2004, before the experiment.

Heterogeneous treatment effects

- **Heterogeneous treatment effects:** effect varies across groups
  - Average effect of a drug is 0, but + for `men` and − for `women`
  - Important questions for determining who should receive treatment
- Social pressure experiment:
  - primary2004 whether the person voted in 2004, before the experiment.
  - Do 2004 voters react differently to social pressure mailer than nonvoters?

- **Heterogeneous treatment effects:** effect varies across groups
  - Average effect of a drug is 0, but + for `men` and − for `women`
  - Important questions for determining who should receive treatment
- Social pressure experiment:
  - primary2004 whether the person voted in 2004, before the experiment.
  - Do 2004 voters react differently to social pressure mailer than nonvoters?
- Two approaches:
  - Difference in effects between groups (subsetting approach)

Heterogeneous treatment effects

- **Heterogeneous treatment effects:** effect varies across groups
  - Average effect of a drug is 0, but $+$ for `men` and $-$ for `women`
  - Important questions for determining who should receive treatment
- Social pressure experiment:
  - primary2004 whether the person voted in 2004, before the experiment.
  - Do 2004 voters react differently to social pressure mailer than nonvoters?
- Two approaches:
  - Difference in effects between groups (subsetting approach)
  - Interaction terms in regression

```r
social <-read.csv("data/social.csv")

social |>
  filter(messages %in% c("Neighbors", "Control")) |>
  group_by(messages, primary2004) |>
  summarize(avg_vote = mean(primary2006)) |>
  pivot_wider(
    names_from = messages,
    values_from = avg_vote
  ) |>
  mutate(diff_exp_vote_2004 = `Neighbors` - `Control`) |>
  pull(diff_exp_vote_2004)
```

```
## [1] 0.06929617 0.09652525
```

ATE for the nonvoters:

Subset approach

```r
social <-read.csv("data/social.csv")

social |>
  filter(messages %in% c("Neighbors", "Control")) |>
  group_by(messages, primary2004) |>
  summarize(avg_vote = mean(primary2006)) |>
  pivot_wider(
    names_from = messages,
    values_from = avg_vote
  ) |>
  mutate(diff_exp_vote_2004 = `Neighbors` - `Control`) |>
  pull(diff_exp_vote_2004)
```

```
## [1] 0.06929617 0.09652525
```

ATE for the nonvoters: 0.0693

ATE for the voters:

```
social <-read.csv("data/social.csv")

social |>
  filter(messages %in% c("Neighbors", "Control")) |>
  group_by(messages, primary2004) |>
  summarize(avg_vote = mean(primary2006)) |>
  pivot_wider(
    names_from = messages,
    values_from = avg_vote
  ) |>
  mutate(diff_exp_vote_2004 = `Neighbors` - `Control`) |>
  pull(diff_exp_vote_2004)
```

```
## [1] 0.06929617 0.09652525
```

ATE for the nonvoters: 0.0693

ATE for the voters: 0.0965

How much does the esimated treatment effect differ between groups?

```
diff_exp_vote<- social |>
  filter(messages %in% c("Neighbors", "Control")) |>
  group_by(messages, primary2004) |>
  summarize(avg_vote = mean(primary2006)) |>
  pivot_wider(
    names_from = messages,
    values_from = avg_vote
  ) |>
  mutate(diff_exp_vote_2004 = `Neighbors` - `Control`) |>
  pull(diff_exp_vote_2004)

  diff(diff_exp_vote)[1]
```

```
## [1] 0.02722908
```

- Any easier way to allow for different effects of treatment by groups?

- Can allow for different effects of a variable with an **interaction term**

$$\text{turnout}_i = \alpha + \beta_1 \text{primary2004}_i + \beta_2 \text{neighbors}_i + \beta_3(\text{primary2004}_i \times \text{neighbors}_i) + \epsilon_i$$

- Can allow for different effects of a variable with an **interaction term**

$$\text{turnout}_i = \alpha + \beta_1 \text{primary2004}_i + \beta_2 \text{neighbors}_i + \beta_3(\text{primary2004}_i \times \text{neighbors}_i) + \epsilon_i$$

- Primary 2004 variable multiplied by the neighbors variable

- Can allow for different effects of a variable with an **interaction term**

$$\text{turnout}_i = \alpha + \beta_1 \text{primary2004}_i + \beta_2 \text{neighbors}_i + \beta_3 (\text{primary2004}_i \times \text{neighbors}_i) + \epsilon_i$$

- Primary 2004 variable multiplied by the neighbors variable
  - Equal to 1 if voted in 2004 (primary2004 $==$ 1) and received neighbors mailer (neighbors $==$ 1)

- Can allow for different effects of a variable with an **interaction term**

$$\text{turnout}_i = \alpha + \beta_1 \text{primary2004}_i + \beta_2 \text{neighbors}_i + \beta_3(\text{primary2004}_i \times \text{neighbors}_i) + \epsilon_i$$

- Primary 2004 variable multiplied by the neighbors variable
    - Equal to 1 if voted in 2004 (primary2004 $==$ 1) and received neighbors mailer (neighbors $==$ 1)
- Easiest to understand by investigating predicted values

Predicted values from non-interacted model

- Let $X_i = \textit{primary}2004_i$ and $Z_i = \textit{neighbors}_i$:

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

	Control ($Z_i = 0$)	Neighbors($Z_i = 1$)
no vote ($X_i = 0$)		

Predicted values from non-interacted model

- Let $X_i = \textit{primary}2004_i$ and $Z_i = \textit{neighbors}_i$:

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

	Control ($Z_i = 0$)	Neighbors($Z_i = 1$)
no vote ($X_i = 0$)	$\widehat{\alpha}$	

Predicted values from non-interacted model

- Let $X_i = primary2004_i$ and $Z_i = neighbors_i$:

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

	Control ($Z_i = 0$)	Neighbors($Z_i = 1$)
no vote ($X_i = 0$)	$\widehat{\alpha}$	$\widehat{\alpha} + \widehat{\beta}_2$
vote ($X_i = 1$)		

- Let $X_i = \textit{primary}2004_i$ and $Z_i = \textit{neighbors}_i$:

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

	Control ($Z_i = 0$)	Neighbors($Z_i = 1$)
no vote ($X_i = 0$)	$\widehat{\alpha}$	$\widehat{\alpha} + \widehat{\beta}_2$
vote ($X_i = 1$)	$\widehat{\alpha} + \widehat{\beta}_1$	

Predicted values from non-interacted model

- Let $X_i = primary2004_i$ and $Z_i = neighbors_i$:

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

	Control ($Z_i = 0$)	Neighbors($Z_i = 1$)
no vote ($X_i = 0$)	$\widehat{\alpha}$	$\widehat{\alpha} + \widehat{\beta}_2$
vote ($X_i = 1$)	$\widehat{\alpha} + \widehat{\beta}_1$	$\widehat{\alpha} + \widehat{\beta}_1 + \widehat{\beta}_2$

- Effect of neighbors for non-voters:

Predicted values from non-interacted model

- Let $X_i = \textit{primary}2004_i$ and $Z_i = \textit{neighbors}_i$:

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

	Control ($Z_i = 0$)	Neighbors($Z_i = 1$)
no vote ($X_i = 0$)	$\widehat{\alpha}$	$\widehat{\alpha} + \widehat{\beta}_2$
vote ($X_i = 1$)	$\widehat{\alpha} + \widehat{\beta}_1$	$\widehat{\alpha} + \widehat{\beta}_1 + \widehat{\beta}_2$

- Effect of neighbors for non-voters: $(\widehat{\alpha} + \widehat{\beta}_2) - (\hat{\alpha}) = \hat{\beta}_2$
- Effect of neighbors for voters:

Predicted values from non-interacted model

- Let $X_i = \mathit{primary2004}_i$ and $Z_i = \mathit{neighbors}_i$:

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

	Control ($Z_i = 0$)	Neighbors($Z_i = 1$)
no vote ($X_i = 0$)	$\widehat{\alpha}$	$\widehat{\alpha} + \widehat{\beta}_2$
vote ($X_i = 1$)	$\widehat{\alpha} + \widehat{\beta}_1$	$\widehat{\alpha} + \widehat{\beta}_1 + \widehat{\beta}_2$

- Effect of neighbors for non-voters: $(\widehat{\alpha} + \widehat{\beta}_2) - (\hat{\alpha}) = \hat{\beta}_2$
- Effect of neighbors for voters:$(\widehat{\alpha} + \widehat{\beta}_1 + \widehat{\beta}_2) - (\widehat{\alpha} + \widehat{\beta}_1) = \widehat{\beta}_2$

Now for the 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($Z_i = 1$)
no vote ($X_i = 0$)		

Now for the 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($Z_i = 1$)
no vote ($X_i = 0$)	$\widehat{\alpha}$	

Now for the 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($Z_i = 1$)
no vote ($X_i = 0$)	$\widehat{\alpha}$	$\widehat{\alpha} + \widehat{\beta}_2$
vote ($X_i = 1$)		

Now for the 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($Z_i = 1$)
no vote ($X_i = 0$)	$\widehat{\alpha}$	$\widehat{\alpha} + \widehat{\beta}_2$
vote ($X_i = 1$)	$\widehat{\alpha} + \widehat{\beta}_1$	

Now for the 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($Z_i = 1$)
no vote ($X_i = 0$)	$\widehat{\alpha}$	$\widehat{\alpha} + \widehat{\beta}_2$
vote ($X_i = 1$)	$\widehat{\alpha} + \widehat{\beta}_1$	$\widehat{\alpha} + \widehat{\beta}_1 + \widehat{\beta}_2 + \widehat{\beta}_3$

- Effect of neighbors for non-voters:

Now for the 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($Z_i = 1$)
no vote ($X_i = 0$)	$\widehat{\alpha}$	$\widehat{\alpha} + \widehat{\beta}_2$
vote ($X_i = 1$)	$\widehat{\alpha} + \widehat{\beta}_1$	$\widehat{\alpha} + \widehat{\beta}_1 + \widehat{\beta}_2 + \widehat{\beta}_3$

- Effect of neighbors for non-voters: $(\widehat{\alpha} + \widehat{\beta}_2) - (\hat{\alpha}) = \hat{\beta}_2$
- Effect of neighbors for voters:

Now for the 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($Z_i = 1$)
no vote ($X_i = 0$)	$\widehat{\alpha}$	$\widehat{\alpha} + \widehat{\beta}_2$
vote ($X_i = 1$)	$\widehat{\alpha} + \widehat{\beta}_1$	$\widehat{\alpha} + \widehat{\beta}_1 + \widehat{\beta}_2 + \widehat{\beta}_3$

- Effect of neighbors for non-voters: $(\widehat{\alpha} + \widehat{\beta}_2) - (\hat{\alpha}) = \hat{\beta}_2$
- Effect of neighbors for voters: $(\widehat{\alpha} + \widehat{\beta}_1 + \widehat{\beta}_2 + \widehat{\beta}_3) - (\widehat{\alpha} + \widehat{\beta}_1) = \widehat{\beta}_2 + \widehat{\beta}_3$

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_1 \text{primary2004}_i + \widehat{\beta}_2 \text{neighbors}_i + \widehat{\beta}_3 (\text{primary2004}_i \times \text{neighbors}_i)$$

	Control group	Neighbors group
2004 primary non-voter	$\widehat{\alpha}$	$\widehat{\alpha} + \widehat{\beta}_2$
2004 primary voter	$\widehat{\alpha} + \widehat{\beta}_1$	$\widehat{\alpha} + \widehat{\beta}_1 + \widehat{\beta}_2 + \widehat{\beta}_3$

$\widehat{\alpha}$: turnout rate for 2004 nonvoters in control group

$$\widehat{Y}_i = \widehat{\alpha}+\widehat{\beta}_1\text{primary2004}_i+\widehat{\beta}_2\text{neighbors}_i+\widehat{\beta}_3(\text{primary2004}_i\times\text{neighbors}_i)$$

	Control group	Neighbors group
2004 primary non-voter	$\widehat{\alpha}$	$\widehat{\alpha}+\widehat{\beta}_2$
2004 primary voter	$\widehat{\alpha}+\widehat{\beta}_1$	$\widehat{\alpha}+\widehat{\beta}_1+\widehat{\beta}_2+\widehat{\beta}_3$

$\widehat{\alpha}$: turnout rate for 2004 nonvoters in control group

$\widehat{\beta}_1$: avg difference in turnout between 2004 voters and nonvoters

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_1 \text{primary2004}_i + \widehat{\beta}_2 \text{neighbors}_i + \widehat{\beta}_3 (\text{primary2004}_i \times \text{neighbors}_i)$$

	Control group	Neighbors group
2004 primary non-voter	$\widehat{\alpha}$	$\widehat{\alpha} + \widehat{\beta}_2$
2004 primary voter	$\widehat{\alpha} + \widehat{\beta}_1$	$\widehat{\alpha} + \widehat{\beta}_1 + \widehat{\beta}_2 + \widehat{\beta}_3$

$\widehat{\alpha}$: turnout rate for 2004 nonvoters in control group

$\widehat{\beta}_1$: avg difference in turnout between 2004 voters and nonvoters

$\widehat{\beta}_2$: effect of neighbors for 2004 nonvoters

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_1 \text{primary2004}_i + \widehat{\beta}_2 \text{neighbors}_i + \widehat{\beta}_3 (\text{primary2004}_i \times \text{neighbors}_i)$$

	Control group	Neighbors group
2004 primary non-voter	$\widehat{\alpha}$	$\widehat{\alpha} + \widehat{\beta}_2$
2004 primary voter	$\widehat{\alpha} + \widehat{\beta}_1$	$\widehat{\alpha} + \widehat{\beta}_1 + \widehat{\beta}_2 + \widehat{\beta}_3$

$\widehat{\alpha}$: turnout rate for 2004 nonvoters in control group

$\widehat{\beta}_1$: avg difference in turnout between 2004 voters and nonvoters

$\widehat{\beta}_2$: effect of neighbors for 2004 nonvoters

$\widehat{\beta}_3$: difference in the effect of neighbors mailer between 2004 voters & nonvoters

- You can include an interaction with var1:var2:

- You can include an interaction with var1:var2:

```
social.neighbor <- social |>
  filter(messages %in% c("Neighbors", "Control"))

fit <- lm(primary2006 ~ primary2004 + messages +
            primary2004:messages, data = social.neighbor)
coef(fit)
```

```
##                  (Intercept)                  primary2004
##                   0.23710990                   0.14869507
##             messagesNeighbors primary2004:messagesNeighbors
##                   0.06929617                   0.02722908
```

- You can include an interaction with var1:var2:

```
social.neighbor <- social |>
  filter(messages %in% c("Neighbors", "Control"))

fit <- lm(primary2006 ~ primary2004 + messages +
            primary2004:messages, data = social.neighbor)
coef(fit)
```

```
##                   (Intercept)                   primary2004
##                    0.23710990                    0.14869507
##             messagesNeighbors primary2004:messagesNeighbors
##                    0.06929617                    0.02722908
```

- Compare coefficients to subset approach

ATE for nonvoters and voters:

```
## [1] 0.06929617 0.09652525
```

Difference in effects

```
## [1] 0.02722908
```

2. Interactions with Continuous Variables

We will keep using the social pressure experiment with two conditions (only) (neighbors and control)

We will keep using the social pressure experiment with two conditions (only) (neighbors and control)

Let's first create an age variable

```
social <- social |>
  mutate(age = 2006-yearofbirth)

social.neighbor <- social |>
  filter(messages %in% c("Neighbors", "Control"))
```

- In the previous example:
  - Effect of the neighbors mailer differ for previous vs. nonvoters?

- In the previous example:
  - Effect of the neighbors mailer differ for previous vs. nonvoters?
  - Used an interaction term to assess **effect heterogeneity** between groups

- In the previous example:
  - Effect of the neighbors mailer differ for previous vs. nonvoters?
  - Used an interaction term to assess **effect heterogeneity** between groups
- How does the effect of the Neighbors mailer varies by age?

Heterogeneous effects

- In the previous example:
    - Effect of the neighbors mailer differ for previous vs. nonvoters?
    - Used an interaction term to assess **effect heterogeneity** between groups
- How does the effect of the Neighbors mailer varies by age?
    - Not just two groups, but a continuum of possible age values

- In the previous example:
  - Effect of the neighbors mailer differ for previous vs. nonvoters?
  - Used an interaction term to assess **effect heterogeneity** between groups
- How does 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!

- In the previous example:
  - Effect of the neighbors mailer differ for previous vs. nonvoters?
  - Used an interaction term to assess **effect heterogeneity** between groups
- How does 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!

$$Y_i = \alpha + \beta_1 \text{age}_i + \beta_2 \text{neighbors}_i + \beta_3 (\text{age}_i \times \text{neighbors}_i) + \epsilon_i$$

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

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 group	Neighbors group
25yr old ($X_i = 25$)		

- Let X_i = age$_i$ and Z_i = neighbors$_i$:

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

	Control group	Neighbors group
25yr old ($X_i = 25$)	$\widehat{\alpha} + \widehat{\beta}_1 25$	

Predicted values from non-interacted model

- Let X_i = age$_i$ and Z_i = neighbors$_i$:

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

	Control group	Neighbors group
25yr old ($X_i = 25$)	$\widehat{\alpha} + \widehat{\beta}_1 25$	$\widehat{\alpha} + \widehat{\beta}_1 25 + \widehat{\beta}_2$
26yr old ($X_i = 26$)		

Predicted values from non-interacted model

- Let X_i = age$_i$ and Z_i = neighbors$_i$:

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

	Control group	Neighbors group
25yr old ($X_i = 25$)	$\widehat{\alpha} + \widehat{\beta}_1 25$	$\widehat{\alpha} + \widehat{\beta}_1 25 + \widehat{\beta}_2$
26yr old ($X_i = 26$)	$\widehat{\alpha} + \widehat{\beta}_1 26$	

- Let X_i = age$_i$ and Z_i = neighbors$_i$:

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

	Control group	Neighbors group
25yr old ($X_i = 25$)	$\widehat{\alpha} + \widehat{\beta}_1 25$	$\widehat{\alpha} + \widehat{\beta}_1 25 + \widehat{\beta}_2$
26yr old ($X_i = 26$)	$\widehat{\alpha} + \widehat{\beta}_1 26$	$\widehat{\alpha} + \widehat{\beta}_1 26 + \widehat{\beta}_2$

Effect of neighbors for a 25 year-old:

- 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 group	Neighbors group
25yr old ($X_i = 25$)	$\widehat{\alpha} + \widehat{\beta}_1 25$	$\widehat{\alpha} + \widehat{\beta}_1 25 + \widehat{\beta}_2$
26yr old ($X_i = 26$)	$\widehat{\alpha} + \widehat{\beta}_1 26$	$\widehat{\alpha} + \widehat{\beta}_1 26 + \widehat{\beta}_2$

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:

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

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

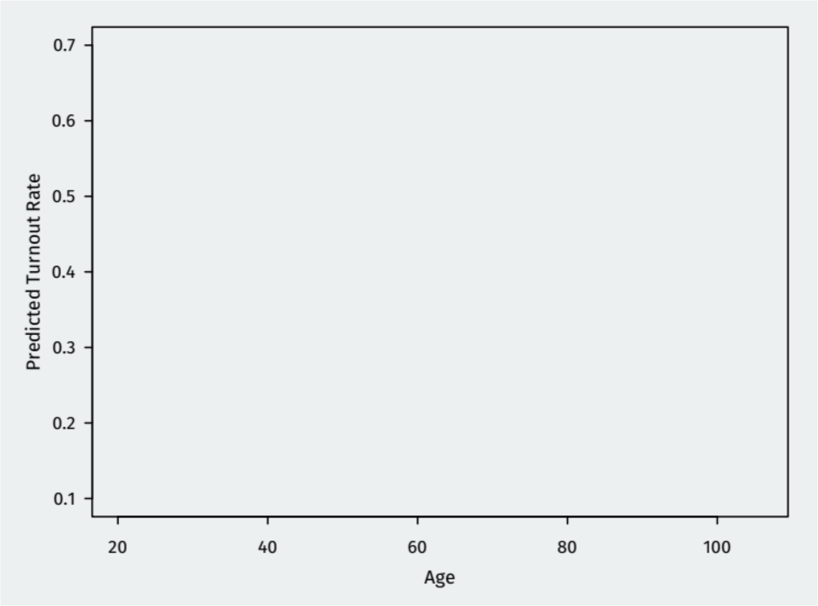
	Control group	Neighbors group
25yr old ($X_i = 25$)	$\widehat{\alpha} + \widehat{\beta}_1 25$	$\widehat{\alpha} + \widehat{\beta}_1 25 + \widehat{\beta}_2$
26yr old ($X_i = 26$)	$\widehat{\alpha} + \widehat{\beta}_1 26$	$\widehat{\alpha} + \widehat{\beta}_1 26 + \widehat{\beta}_2$

Effect of neighbors for a 25 year-old:

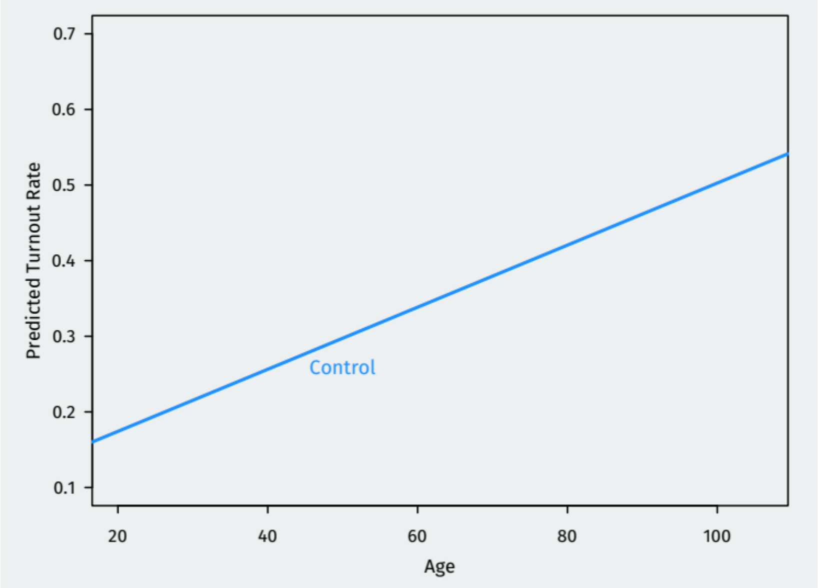
$$(\widehat{\alpha} + \widehat{\beta}_1 25 + \widehat{\beta}_2) - (\widehat{\alpha} + \widehat{\beta}_1 25) = \widehat{\beta}_2$$

Effect of neighbors for a 26 year-old:

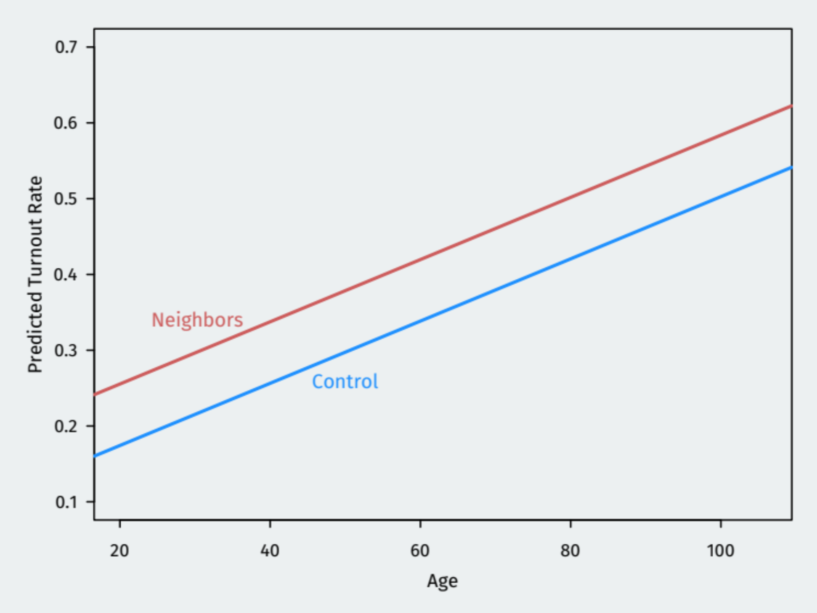
$$(\widehat{\alpha} + \widehat{\beta}_1 26 + \widehat{\beta}_2) - (\widehat{\alpha} + \widehat{\beta}_1 26) = \widehat{\beta}_2$$


0.7
0.6
0.5
0.4
0.3
0.2
0.1
Predicted Turnout Rate
20
40
60
80
100
Age

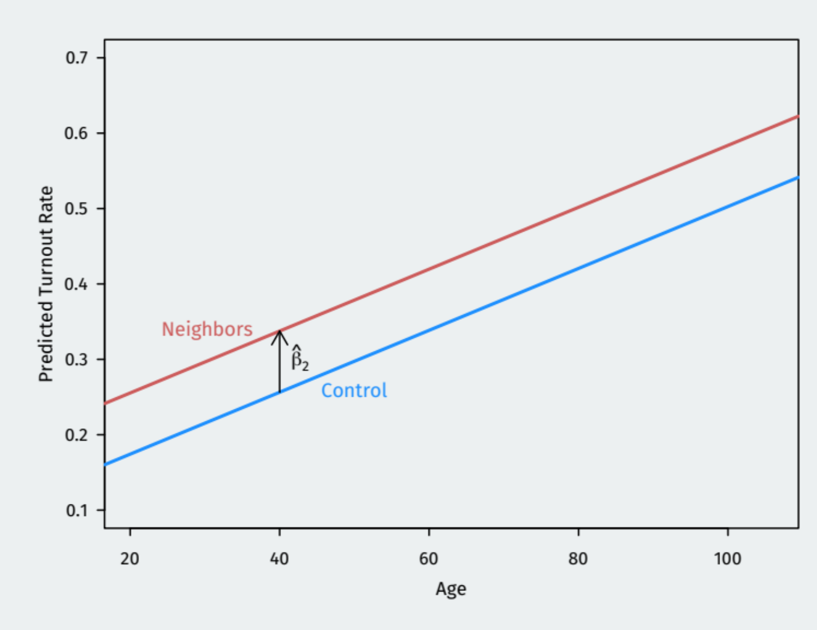
Visualizing the regression



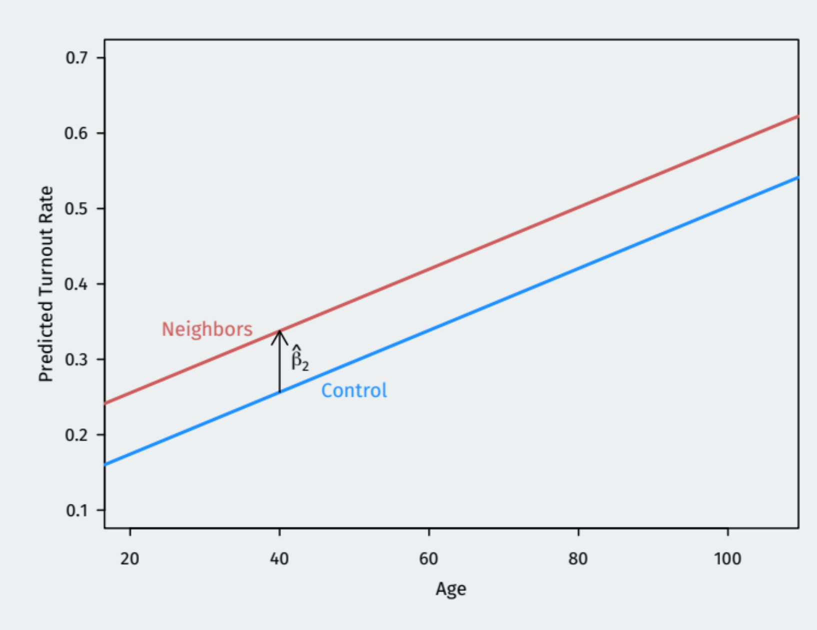
Visualizing the regression



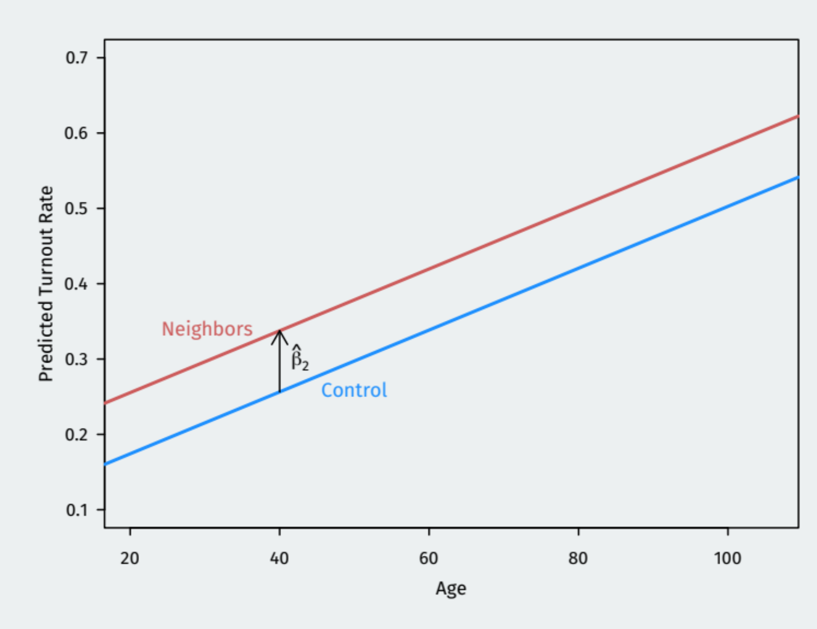
Visualizing the regression

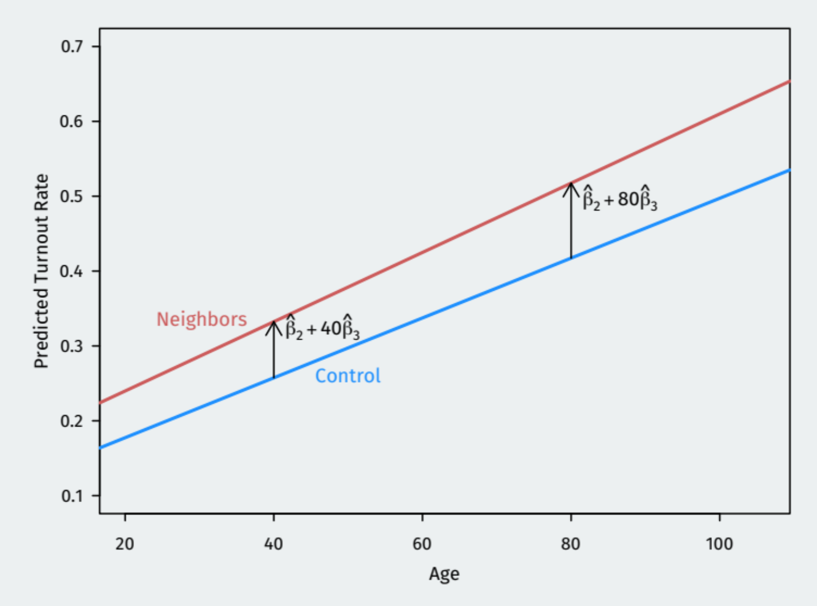


Visualizing the regression



Visualizing the regression




Predicted Turnout Rate
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Neighbors
Control
$\hat{\beta}_2 + 40\hat{\beta}_3$
$\hat{\beta}_2 + 80\hat{\beta}_3$
20
40
60
80
100
Age

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

- α: average turnout for 0 year-olds in the control group

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

- α: average turnout for 0 year-olds in the control group
- $\widehat{\beta}_1$: slope of regression line for age in the control group

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

- α: 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

Interpreting coefficients

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

- α: 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

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

- α: 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-old: $\widehat{\beta}_2 + \widehat{\beta}_3 x$

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

- α: 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-old: $\widehat{\beta}_2 + \widehat{\beta}_3 x$
    - Effect for $(x+1)$ year-olds: $\widehat{\beta}_2 + \widehat{\beta}_3 (x+1)$

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

- α: 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-old: $\widehat{\beta}_2 + \widehat{\beta}_3 x$
    - Effect for $(x+1)$ year-olds: $\widehat{\beta}_2 + \widehat{\beta}_3(x+1)$
    - Change in effect: $\widehat{\beta}_3$

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

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

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

```
##           (Intercept)                    age      messagesNeighbors
##          0.0974732574           0.0039982107           0.0498294321
## age:messagesNeighbors
##          0.0006283079
```

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

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

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

```
##           (Intercept)                    age      messagesNeighbors
##          0.0974732574           0.0039982107           0.0498294321
## age:messagesNeighbors
##          0.0006283079
```

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

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

```
##           (Intercept)                    age      messagesNeighbors
##          0.0974732574           0.0039982107           0.0498294321
## age:messagesNeighbors
##          0.0006283079
```

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

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

$\widehat{\alpha}$: 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 $Z_i = 0$

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

$\widehat{\alpha}$: 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 $X_i = 0$

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

$\widehat{\alpha}$: 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:

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

$\widehat{\alpha}$: 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

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

$\widehat{\alpha}$: 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

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. Nonlinear relationship

- 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:

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

```r
social.neighbor <- social |>
  filter(messages %in% c("Neighbors", "Control"))
```

Linear regression are linear

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

- Standard linear regression can only pick up **linear** relationships

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

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

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

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

Adding a squared term

- To allow for nonlinearity in age, add a squared term to the model:

- To allow for nonlinearity in age, add a squared term to the model:

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_1 age_i + \widehat{\beta}_2 (age_i^2)$$

- We are now fitting a **parabola** to the data

- To allow for nonlinearity in age, add a squared term to the model:

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_1 age_i + \widehat{\beta}_2 (age_i^2)$$

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

- To allow for nonlinearity in age, add a squared term to the model:

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_1 age_i + \widehat{\beta}_2 (age_i^2)$$

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

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

```
##   (Intercept)           age      I(age^2)
## -8.168043e-02  1.227357e-02 -8.078954e-05
```

- $\widehat{\beta}_2$: 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
## 0.1314752 0.1404364 0.1492361
```

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
## 0.1314752 0.1404364 0.1492361
```

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

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

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
## 0.1314752 0.1404364 0.1492361
```

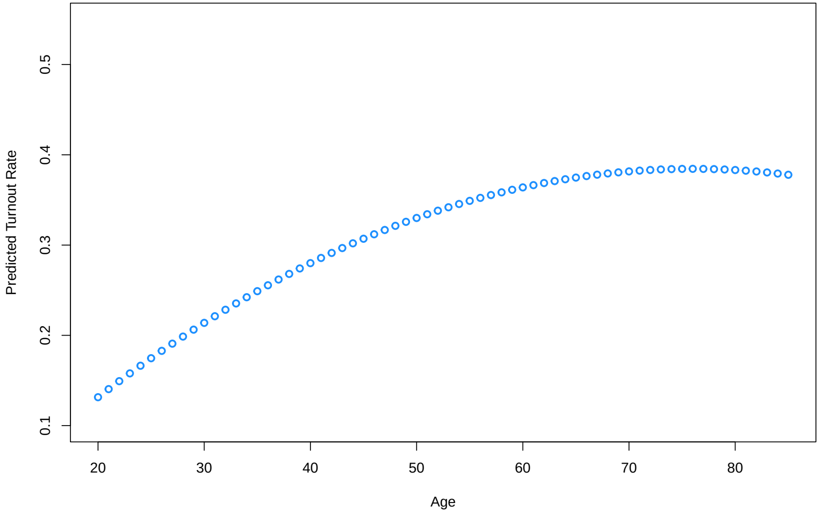
- Create a vector of ages to predict and save predictions:

```
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

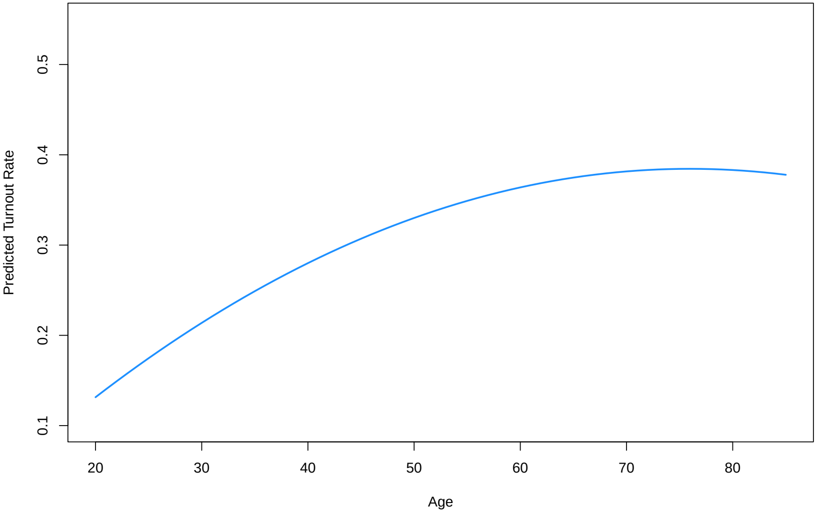


Plotting lines instead of points

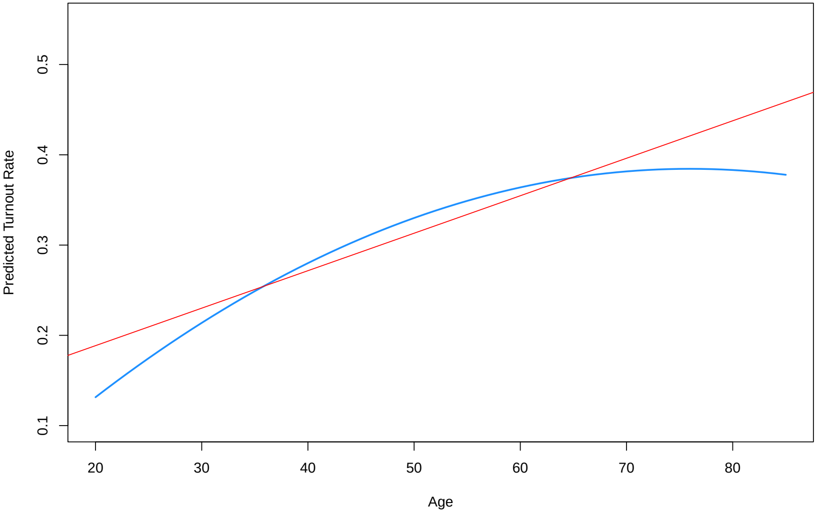
- If you want to connect the dots in your scatterplot, you can use the type = "l" ("line" type)

```
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



- One independent variable: just look at a scatterplot

- One independent variable: just look at a scatterplot
- With multiple independent variables, harder to diagnose

- One independent variable: just look at a scatterplot
- With multiple independent variables, harder to diagnose
- One useful tool: scatterplot of residuals versus independent variables

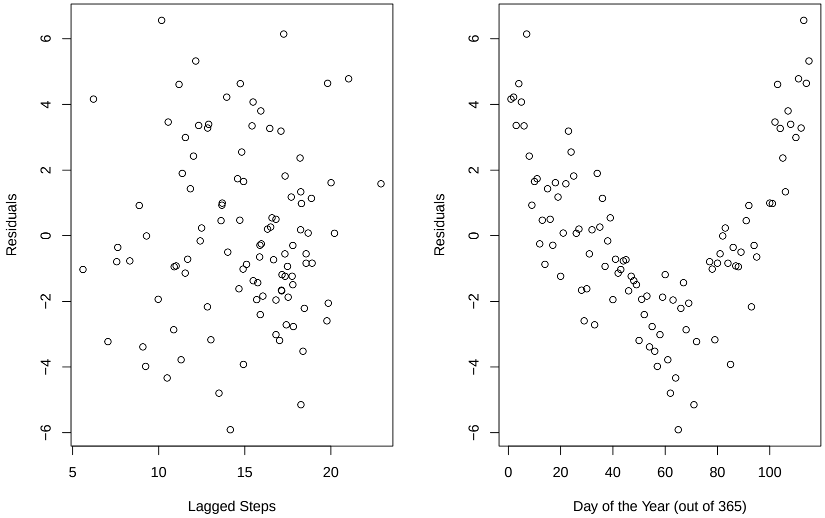
- One independent variable: just look at a scatterplot
- With multiple independent variables, harder to diagnose
- One useful tool: scatterplot of residuals versus independent variables
- Example: health data (weight and step)

```
library(TPDdata)
library(lubridate)

health <- health |>
    mutate(year=year(date)) |>
    mutate(dayofyear = yday(date)) |>
    filter(year==2017) |>
    filter(dayofyear < 120)

health <- drop_na(health)
w.fit <-lm(weight ~ steps_lag  + date, data = health)
```

```
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")
```

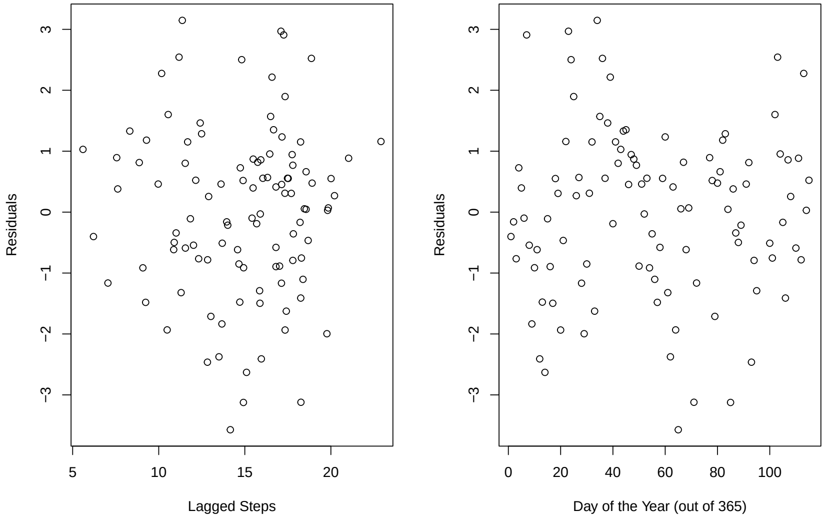

Residuals
6
4
2
0
−2
−4
−6
5
10
15
20
Lagged Steps
Residuals
6
4
2
0
−2
−4
−6
0
20
40
60
80
100
Day of the Year (out of 365)

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 I(dayofyear^2)
##  177.221763772    0.048823891   -0.425866681    0.002210719
```

```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")
```


Residuals
3
2
1
0
−1
−2
−3
5
10
15
20
Lagged Steps
Residuals
3
2
1
0
−1
−2
−3
0
20
40
60
80
100
Day of the Year (out of 365)

Class exercises

- We are going to cover some tools for exploring bivariate relationships
- We'll use the data from the Brookckman & Kalla (2016) transphobia study
- Basic summary of experiment:
  - Randomly assigned door-to-door canvassers to two conditions
  - Conditions: perspective-taking script (treatment) or recycling script (placebo)
  - Follow up surveys at 3 days, 3 weeks, 6 weeks, and 3 months

```r
library(tidyverse)
phobia<-read_csv("data/transphobia_all.csv")
phobia <- drop_na(phobia)
phobia <- phobia |>
  mutate(treat_ind = ifelse(treat_ind == "Treat.", 1, 0))
```

Variable	Description
wave	Baseline, 3 days, 21 days, 42 days, and 90 days
age	Age of the respondent in years
female	1=respondent marked “Female” on voter registration, 0 otherwise
voted_gen_14	1 if respondent voted in the 2014 general election
voted_gen_12	1 if respondent voted in the 2012 general election
treat_ind	1 if respondent was assigned to treatment, 0 for control
racename	character name of racial identity indicated on voter file
democrat	1 if respondent is a registered Democrat
therm_trans	0-100 feeling therm. about transgender people
therm_obama	0-100 feeling therm. about Barack Obama

Q 1

Run a regression of thermometer scores for transgender people in wave 1 (3 days) on the treatment indicator (`treat_ind`), the indicator for if the respondent is a Democrat (`democrat`), and the interaction between the two variables

Interpret each of the coefficients in terms of the effects of the intervention

Run a regression of thermometer scores for transgender people in wave 1 (3 days) on the treatment indicator (treat_ind), the indicator fo rif the respondent is a women (female), and the interaction between the two variables.

Interpret each of the coefficients in terms of the effects of the intervention.

Q3

Run a regression of thermometer scores for transgender people in wave 1 on the treatment indicator, age, and the interaction between the two.

What is the estimated effect for a 25 year old? For a 50 year old?

Q4

Run a regression of baseline Obama thermometer scores (therm_obama) on age and the square of age to assess the nonlinear relationship between them.

Calculate predicted values from the model for ages 18 to 90 and plot these as a line