Regression Part III: Interactions & Nonlinearities

PSC4375: Week 8 & 9

Prof. Weldzius

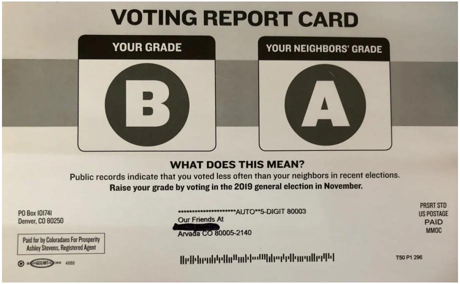
Villanova University

Slides Updated: 2025-03-18

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:
  - Subsets, subsets, subsets.
  - Interaction terms in regression.

Subset approach

- Easy way to estimate heterogeneous effects: our old friend, `filter()`, `group_by()`, and `summarize()`. Woo!
    - First, get the data

```
data(social, package="qss")
```

Subset approach

- Now, estimate the ATE for the **voters**:

```
VotersATE <- social %>%
  filter(primary2004 == 1,
         messages %in% c("Control","Neighbors")) %>%
  group_by(messages) %>%
  summarize(primary2006_mean = mean(primary2006)) %>%
  pivot_wider(names_from = "messages",
              values_from = "primary2006_mean") %>%
  mutate(ate_v = Neighbors - Control) %>%
  select(ate_v)
VotersATE
```

```
## # A tibble: 1 x 1
##    ate_v
##    <dbl>
## 1 0.0965
```

Filter approach

- Now, estimate the ATE for the **nonvoters**:

```r
NonvotersATE <- social %>%
  filter(primary2004 == 0,
         messages %in% c("Control","Neighbors")) %>%
  group_by(messages) %>%
  summarize(primary2006_mean = mean(primary2006)) %>%
  pivot_wider(names_from = "messages",
              values_from = "primary2006_mean") %>%
  mutate(ate_nv = Neighbors - Control) %>%
  select(ate_nv)
NonvotersATE
```

```
## # A tibble: 1 x 1
##   ate_nv
##    <dbl>
## 1 0.0693
```

Difference in effects

- How much does the estimated treatment effect differ between groups?

```
VotersATE$ate_v - NonvotersATE$ate_nv
```

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

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

Interaction terms

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

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

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

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

	Control ($Z_i = 0$)	Neighbors ($Z_i = 1$)
non-voter ($X_i = 0$)	$\hat{\alpha}$	$\hat{\alpha} + \hat{\beta}_2$
voter ($X_i = 1$)	$\hat{\alpha} + \hat{\beta}_1$	$\hat{\alpha} + \hat{\beta}_1 + \hat{\beta}_2$

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

Predicted from interacted model

- Now for the interacted model:

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

	Control ($Z_i = 0$)	Neighbors ($Z_i = 1$)
non-voter ($X_i = 0$)	$\hat{\alpha}$	$\hat{\alpha} + \hat{\beta}_2$
voter ($X_i = 1$)	$\hat{\alpha} + \hat{\beta}_1$	$\hat{\alpha} + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$

Interpreting coefficients

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

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

- $\hat{\alpha}$: turnout rate for 2004 nonvoters in control group.
- $\hat{\beta}_1$: avg difference in turnout between 2004 voters and nonvoters.
- $\hat{\beta}_2$: effect of neighbors for 2004 nonvoters.
- $\hat{\beta}_3$: difference in the effect of neighbors mailer between 2004 voters and nonvoters.

Interactions in R

- You can include an interaction with var1:var2:

```
social.neighbor <- social %>%
  filter(messages %in% c("Neighbors","Control")) %>%
  mutate(neighbors = ifelse(messages=="Neighbors",1,0))
```

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

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

Interactions in R

```
coef(fit)
```

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

- Compare coefficients to earlier approach:

```
NonvotersATE$ate_nv
```

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

```
VotersATE$ate_v - NonvotersATE$ate_nv
```

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

Interactions with Continuous Variables

- Create an age variable for the Michigan **social pressure get-out-the-vote experiment**:

```
social.neighbor <- social.neighbor %>%
  mutate(age = 2006 - yearofbirth)
summary(social.neighbor$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20.00   41.00   50.00   49.82   59.00  106.00
```

Hetergeneous effects

- From before:
  - Effect of the Neighbors mailer differ from previous voters vs. nonvoters?
  - Used an interaction term to assess **effect heterogeneity** between groups
- How does the effect of the Neighbors mailer vary by age?
  - Not just two groups, but a continuum of possible age values
- Remarkable, the same **interaction term** will work here too!

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

Predicted values from non-interacted model

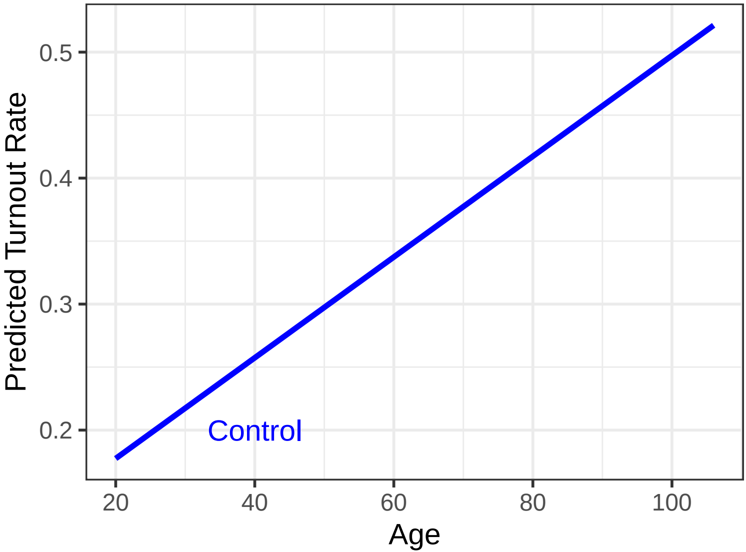
- Let $X_i = \text{age}_i$ and $Z_i = \text{neighbors}_i$

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

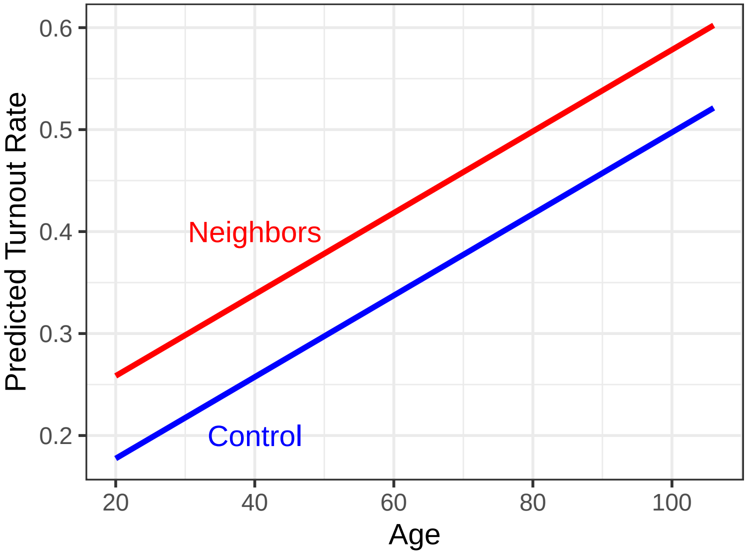
	Control ($Z_i = 0$)	Neighbors ($Z_i = 1$)
25 year-old ($X_i = 25$)	$\hat{\alpha} + \hat{\beta}_1 25$	$\hat{\alpha} + \hat{\beta}_1 25 + \hat{\beta}_2$
26 year-old ($X_i = 26$)	$\hat{\alpha} + \hat{\beta}_1 26$	$\hat{\alpha} + \hat{\beta}_1 26 + \hat{\beta}_2$

- Effect of Neighbors for a 25 year-old: $(\hat{\alpha} + \hat{\beta}_1 25 + \hat{\beta}_2) - (\hat{\alpha} + \hat{\beta}_1 25) = \hat{\beta}_2$
- Effect of Neighbors for a 26 year-old: $(\hat{\alpha} + \hat{\beta}_1 26 + \hat{\beta}_2) - (\hat{\alpha} + \hat{\beta}_1 26) = \hat{\beta}_2$

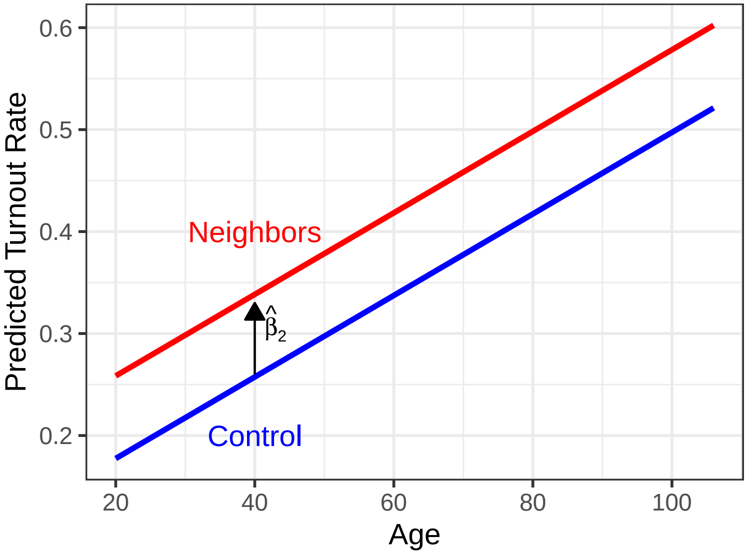
Visualizing the regression



Visualizing the regression



Visualizing the regression



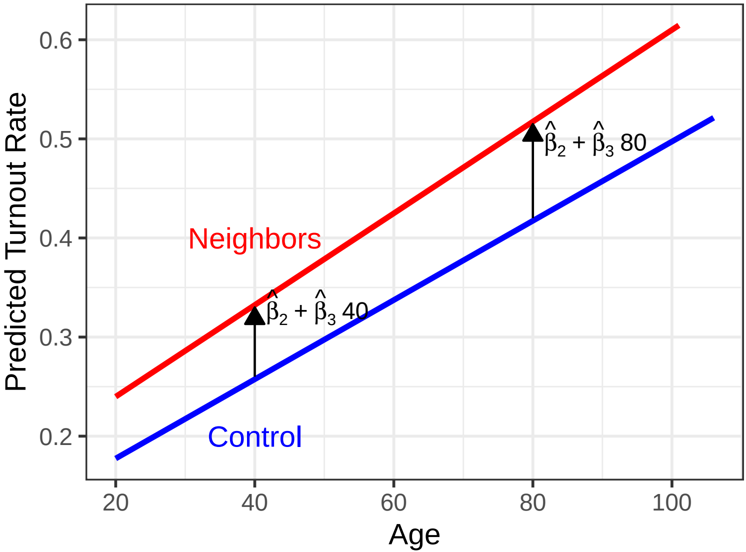
Predicted values from interacted model

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

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

- Effect of Neighbors for a 25 year-old:
  $(\hat{\alpha} + \hat{\beta}_1 25 + \hat{\beta}_2 + \hat{\beta}_3 25) - (\hat{\alpha} + \hat{\beta}_1 25) = \hat{\beta}_2 + \hat{\beta}_3 25)$
- Effect of Neighbors for a 26 year-old:
  $(\hat{\alpha} + \hat{\beta}_1 26 + \hat{\beta}_2 + \hat{\beta}_3 26) - (\hat{\alpha} + \hat{\beta}_1 26) = \hat{\beta}_2 + \hat{\beta}_3 26)$
- Effect of Neighbors for a x year-old: $\hat{\beta}_2 + \hat{\beta}_3 x$

Visualizing the interaction



Interpretting coefficients

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

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

Interactions in R

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

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

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

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

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

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

General interpretation of interactions

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

- $\hat{\alpha}$: average turnout when X_i and Z_i are 0.
  - $\hat{\beta}_1$: average change in Y_i of a one-unit change in X_i when $Z_i = 0$.
  - $\hat{\beta}_2$: average change in Y_i of a one-unit change in Z_i when $X_i = 0$.
  - $\hat{\beta}_3$: 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

Nonlinear relationships

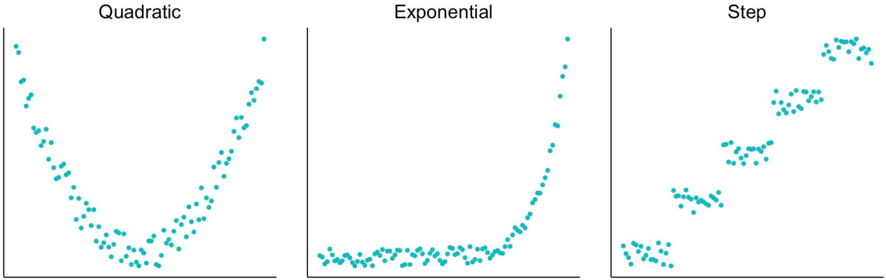


Figure 1: Types of Non-linear Relationships

Linear regression are linear

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

- 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

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 \text{age}_i + \hat{\beta}_2 \text{age}_i^2$$

- 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), data = social.neigh
coef(fit.sq)
```

```
##  (Intercept)          age     I(age^2)
## -0.080067046  0.012154358 -0.000079999
```

- $\hat{\beta}_2$: 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
## 0.1310205 0.1398949 0.1486093
```

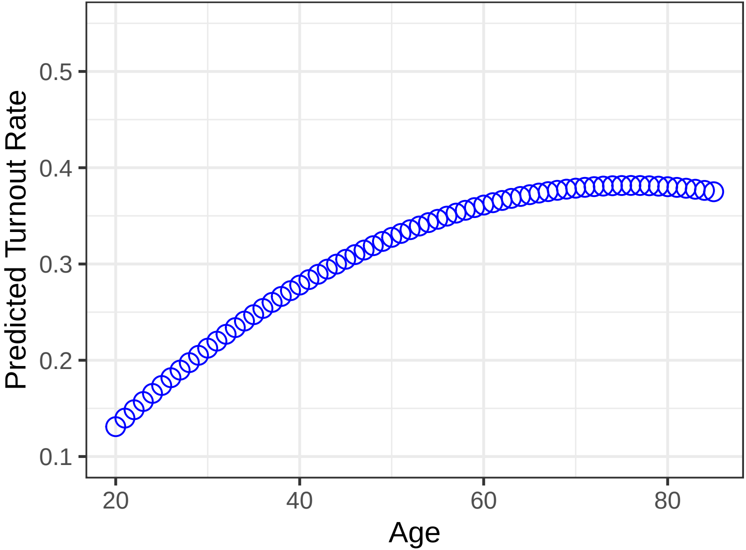
- Create a vector of ages to predict and save predictions:

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

- Plot the predictions:

```
ggplot(age.plot,aes(x = age.vals, y = age.preds)) +
  geom_point(color = "blue", size = 3, shape = 1) + ylim(0.1, 0.55)
  labs(x = "Age", y = "Predicted Turnout Rate") + theme_bw()
```

Plotting predicted values



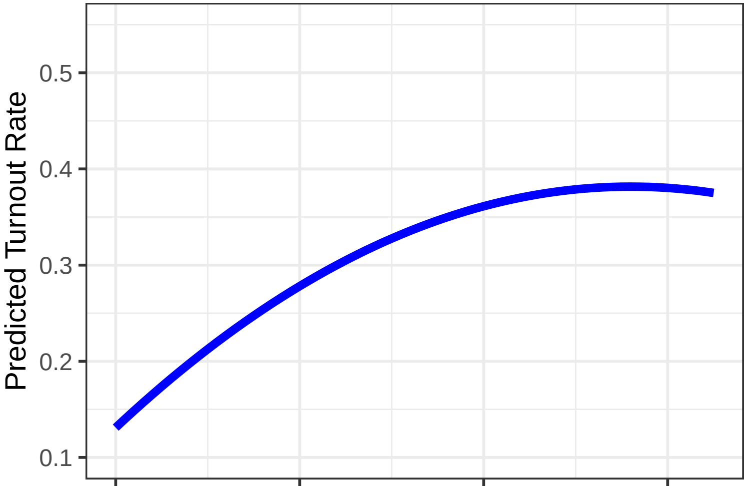
Plotting lines instead of points:

- If you want to connect the dots in your scatterplot, you can use `geom_line()`:

```
ggplot(age.plot, aes(x = age.vals, y = age.preds)) +
  geom_line(color = "blue", size = 1.5) +
  ylim(0.1, 0.55) +
  labs(x = "Age", y = "Predicted Turnout Rate") +
  theme_bw()
```

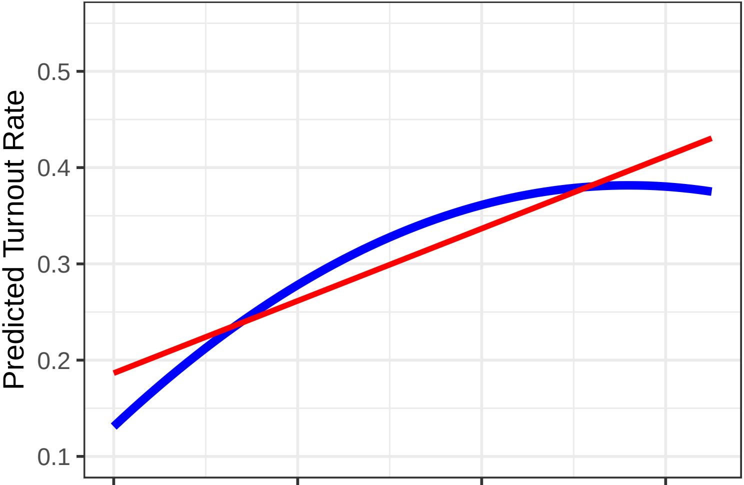
Plotting predicted values

- If you want to connect the dots in your scatterplot, you can use `geom_line()`:



Comparing to linear fit

- If you want to connect the dots in your scatterplot, you can use `geom_line()`:



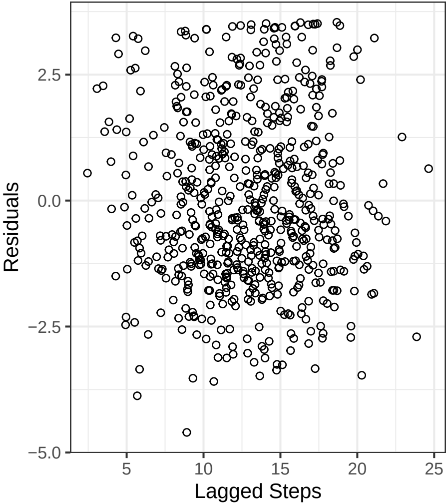
Diagnosing nonlinearity

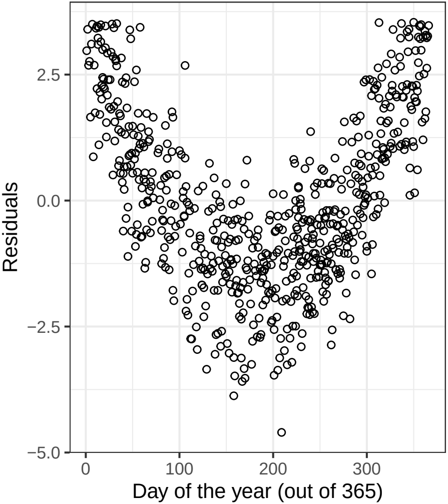
- One independent variable: just look at a scatterplot.
- With multiple independent variables, harder to diagnose.
- One useful tool: scatterplot of residuals versus independent variables.
- Example: let's talk about walking and health

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

w.fit <- lm(weight ~  steps.lag + dayofyear, data = health)
```

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 I(dayofyear^2)
##  1.749194e+02 -2.509427e-03 -5.288116e-02   1.439635e-04
```

Residual plot, redux