Linear Regression

(Estimating Linear Relationships)

Understanding Political Numbers

March 4, 2019

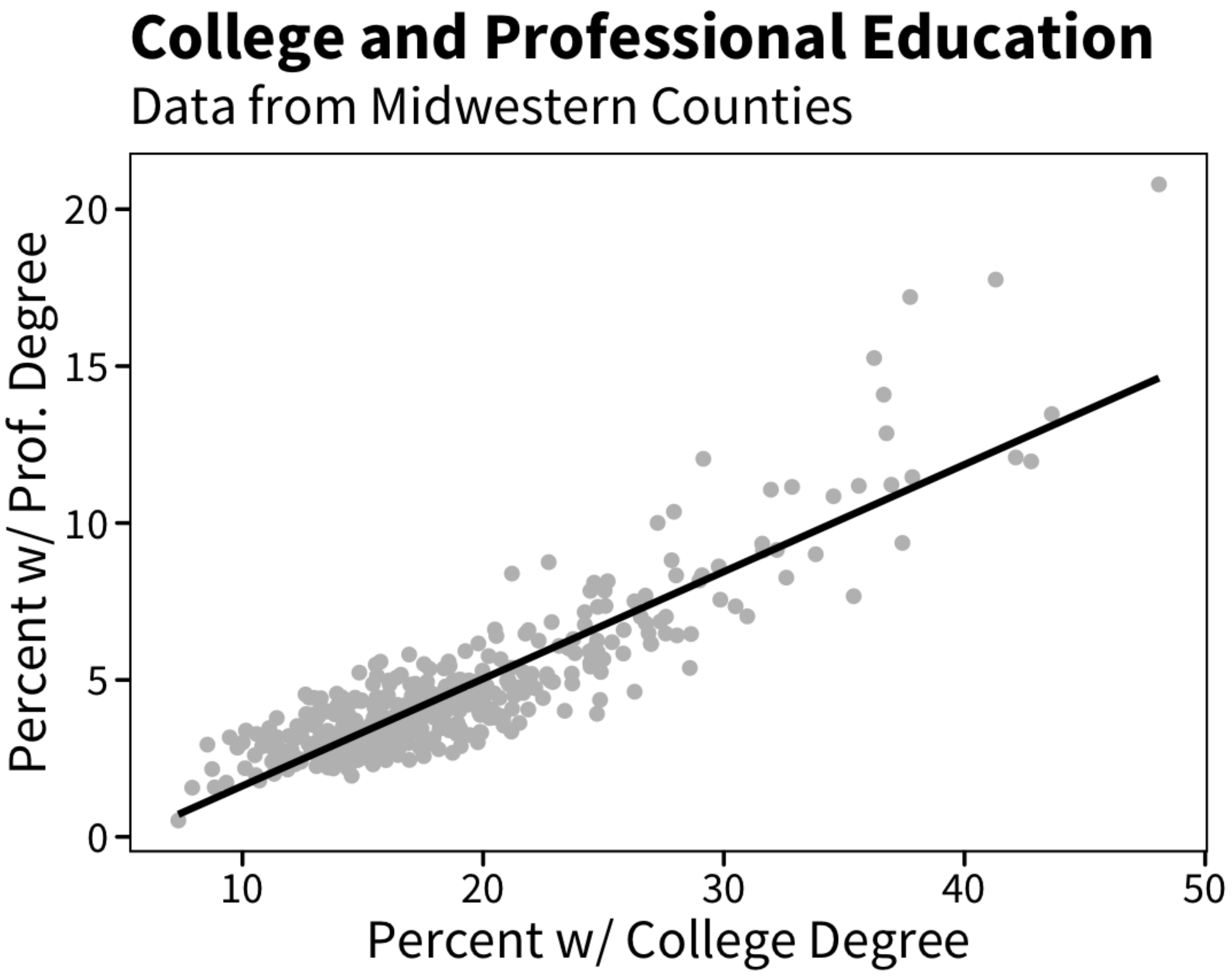
Agenda

Admin stuff

- Research Question due Monday March 11
- Exercise 2 due Wednesday March 13

Exercise 2 tips

Linear regression



Exercise 2

Follow link to data

Politics

Ho

ELECTION 2011 | RESULTS

Updated vote tallies for state Supreme Court

The following table compares vote tallies for the state Supreme Court race between incumbent David Prosser and challenger JoAnne Kloppe Associated Press totals were collected on April 6. The updated totals numbers, certified by the county boards of canvassers, were obtained Accountability Board's website. A PDF of the board's tally can be found here.

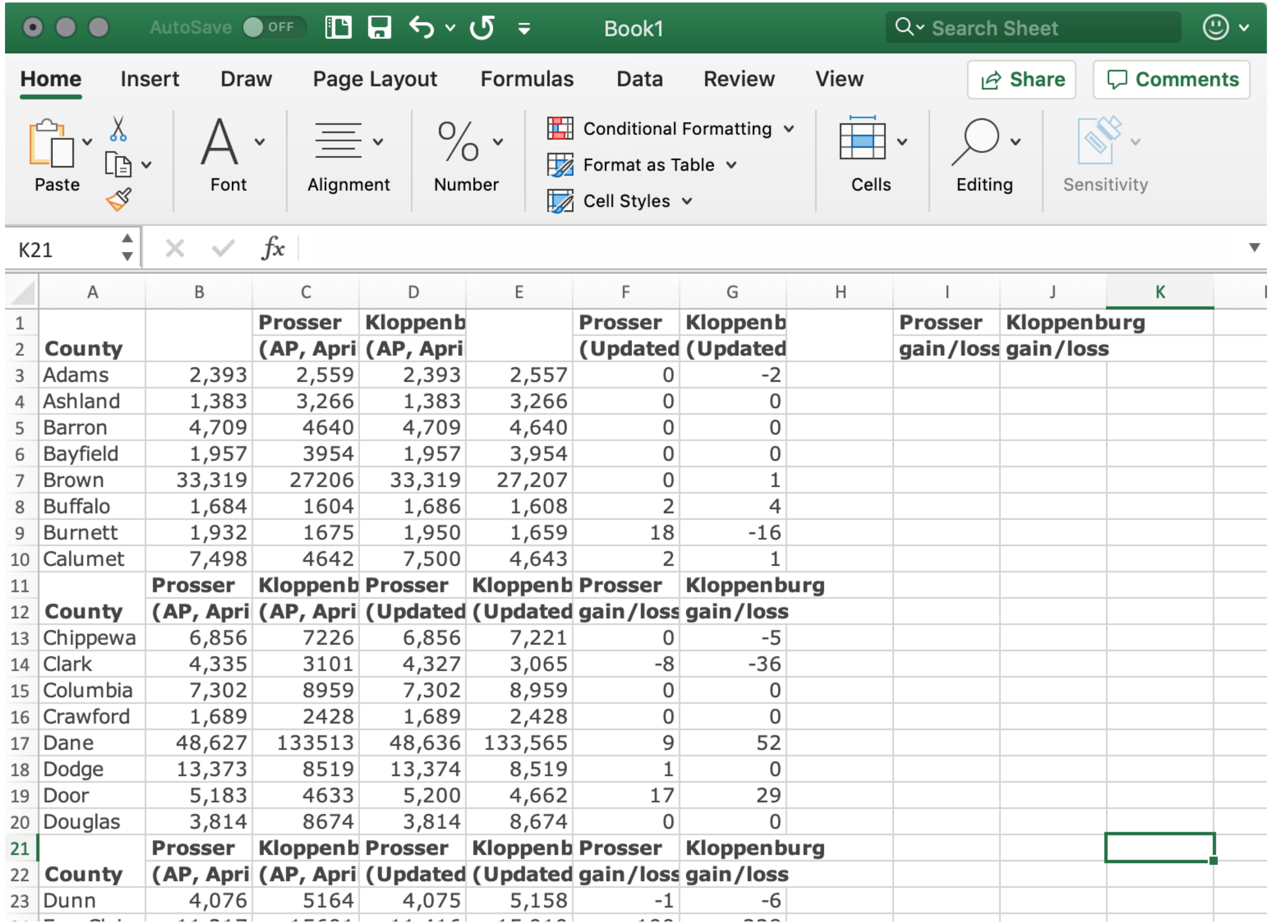
More: Results | Election section

County	Prosser (AP, April 6)	Kloppenburg (AP, April 6)	Prosser (Updated)	Kloppenburg (Updated)	Prosser gain/loss	Kloppenburg gain/loss
Adams	2,393	2,559	2,393	2,557	0	-2
Ashland	1,383	3,266	1,383	3,266	0	0
Barron	4,709	4640	4,709	4,640	0	0
Bayfield	1,957	3954	1,957	3,954	0	0
Brown	33,319	27206	33,319	27,207	0	1
Buffalo	1,684	1604	1,686	1,608	2	4
Burnett	1,932	1675	1,950	1,659	18	-16
Calumet	7,498	4642	7,500	4,643	2	1
County	**Prosser (AP, April 6)**	**Kloppenburg (AP, April 6)**	**Prosser (Updated)**	**Kloppenburg (Updated)**	**Prosser gain/loss**	**Kloppenburg gain/loss**
Chippewa	6,856	7226	6,856	7,221	0	-5
Clark	4,335	3101	4,327	3,065	-8	-36
Columbia	7,302	8959	7,302	8,959	0	0
Crawford	1,689	2428	1,689	2,428	0	0
Dane	48,627	133513	48,636	133,565	9	52
Dodge	13,373	8519	13,374	8,519	1	0
Door	5,183	4633	5,200	4,662	17	29
Douglas	3,814	8674	3,814	8,674	0	0

Paste in Excel/Numbers and CLEAN

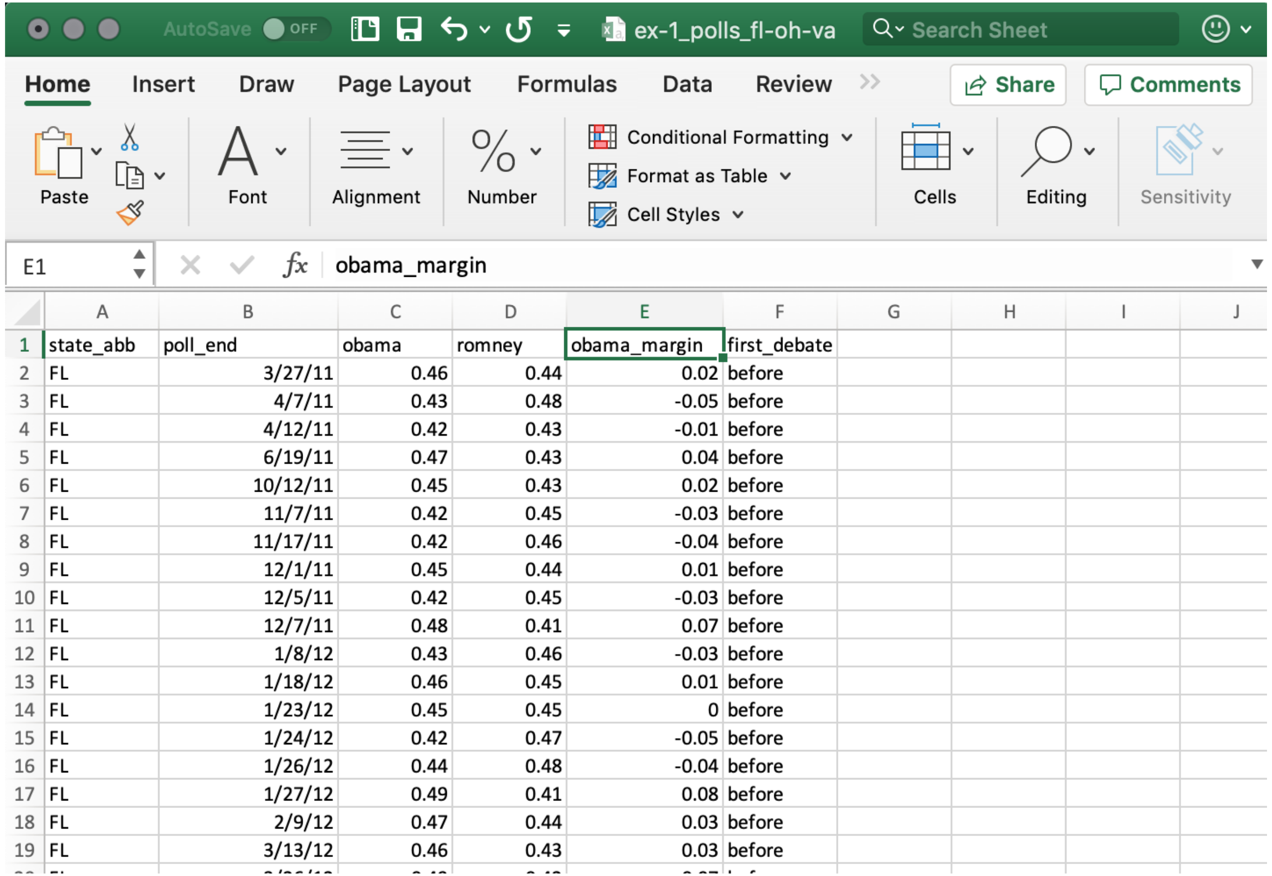
Beware:

- merged rows
- Unneeded rows
- special characters in variable names (use _)
- save as CSV



	A	B	C	D	E	F	G	H	I	J	K
1			Prosser	Kloppenb		Prosser	Kloppenb		Prosser	Kloppenburg	
2	County		(AP, Apri	(AP, Apri		(Updated	(Updated		gain/loss	gain/loss	
3	Adams	2,393	2,559	2,393	2,557	0	-2				
4	Ashland	1,383	3,266	1,383	3,266	0	0				
5	Barron	4,709	4640	4,709	4,640	0	0				
6	Bayfield	1,957	3954	1,957	3,954	0	0				
7	Brown	33,319	27206	33,319	27,207	0	1				
8	Buffalo	1,684	1604	1,686	1,608	2	4				
9	Burnett	1,932	1675	1,950	1,659	18	-16				
10	Calumet	7,498	4642	7,500	4,643	2	1				
11		Prosser	Kloppenb	Prosser	Kloppenb	Prosser	Kloppenburg				
12	County	(AP, Apri	(AP, Apri	(Updated	(Updated	gain/loss	gain/loss				
13	Chippewa	6,856	7226	6,856	7,221	0	-5				
14	Clark	4,335	3101	4,327	3,065	-8	-36				
15	Columbia	7,302	8959	7,302	8,959	0	0				
16	Crawford	1,689	2428	1,689	2,428	0	0				
17	Dane	48,627	133513	48,636	133,565	9	52				
18	Dodge	13,373	8519	13,374	8,519	1	0				
19	Door	5,183	4633	5,200	4,662	17	29				
20	Douglas	3,814	8674	3,814	8,674	0	0				
21		Prosser	Kloppenb	Prosser	Kloppenb	Prosser	Kloppenburg				
22	County	(AP, Apri	(AP, Apri	(Updated	(Updated	gain/loss	gain/loss				
23	Dunn	4,076	5164	4,075	5,158	-1	-6				

Take cues from Ex 1 data



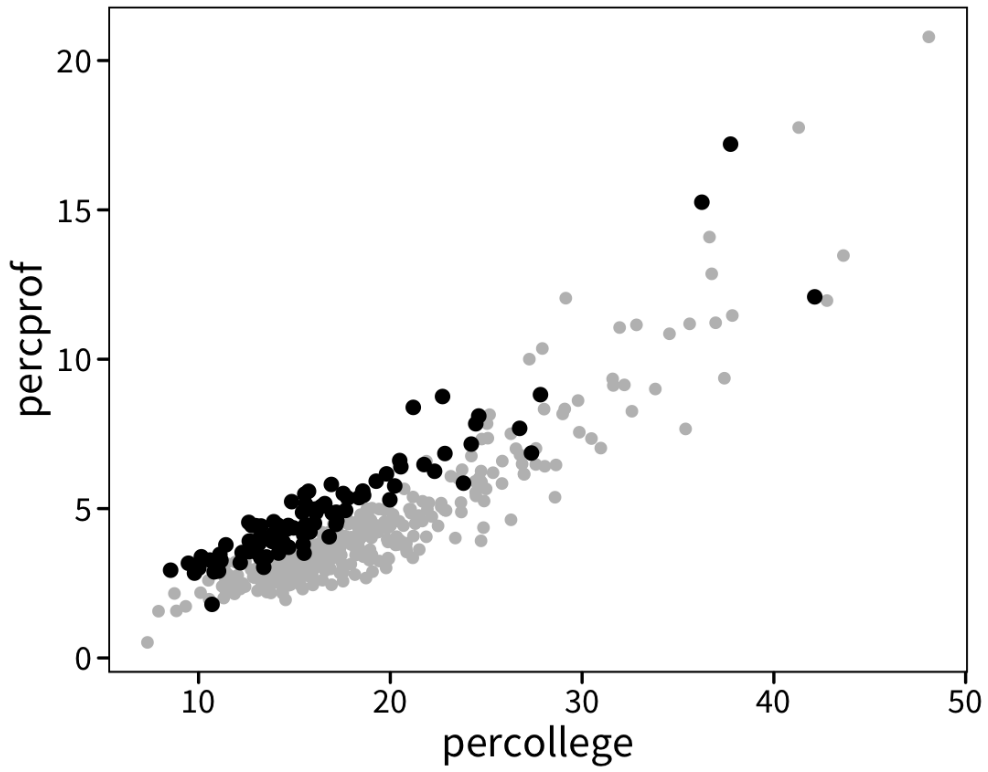
	A	B	C	D	E	F
1	state_abb	poll_end	obama	romney	obama_margin	first_debate
2	FL	3/27/11	0.46	0.44	0.02	before
3	FL	4/7/11	0.43	0.48	-0.05	before
4	FL	4/12/11	0.42	0.43	-0.01	before
5	FL	6/19/11	0.47	0.43	0.04	before
6	FL	10/12/11	0.45	0.43	0.02	before
7	FL	11/7/11	0.42	0.45	-0.03	before
8	FL	11/17/11	0.42	0.46	-0.04	before
9	FL	12/1/11	0.45	0.44	0.01	before
10	FL	12/5/11	0.42	0.45	-0.03	before
11	FL	12/7/11	0.48	0.41	0.07	before
12	FL	1/8/12	0.43	0.46	-0.03	before
13	FL	1/18/12	0.46	0.45	0.01	before
14	FL	1/23/12	0.45	0.45	0	before
15	FL	1/24/12	0.42	0.47	-0.05	before
16	FL	1/26/12	0.44	0.48	-0.04	before
17	FL	1/27/12	0.49	0.41	0.08	before
18	FL	2/9/12	0.47	0.44	0.03	before
19	FL	3/13/12	0.46	0.43	0.03	before

Helpful code tricks:

Specify data in `geom_*` function

```
library("tidyverse")

ggplot(data = midwest,
       aes(x = percollege, y = percprof)) +
  geom_point(color = "gray") +
  geom_point(
    data = filter(midwest, state == "IN"),
    color = "black",
    size = 2
  )
```

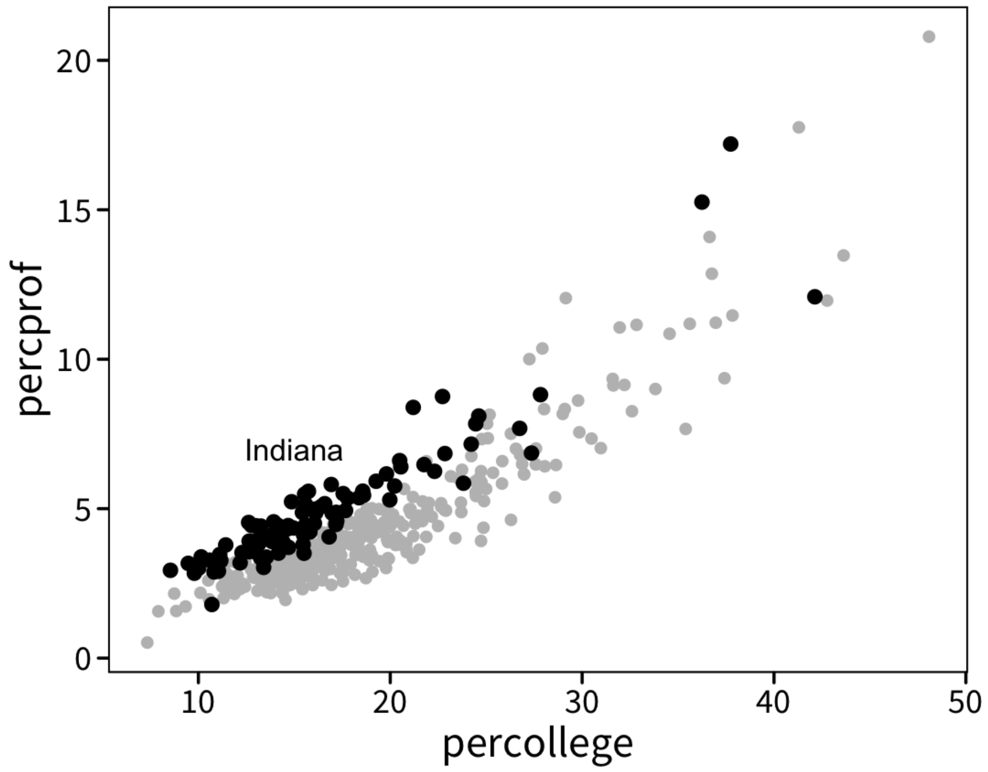
Geoms inherit data and aesthetics from `ggplot()` by default



Helpful code tricks:

Add specific annotations

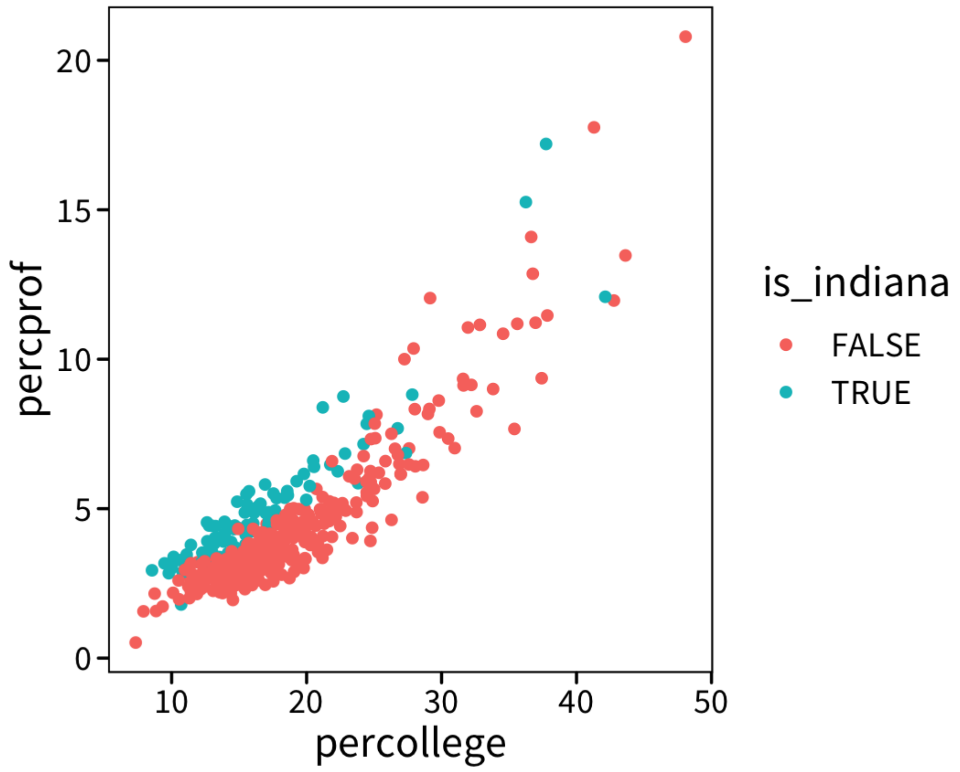
```
ggplot(data = midwest,
       aes(x = percollege, y = percprof)) +
  geom_point(color = "gray") +
  geom_point(
    data = filter(midwest, state == "IN"),
    color = "black",
    size = 2
  )  +
  annotate(geom = "text",
           x = 15, y = 7, label = "Indiana")
```



Helpful code tricks:

Create an "identifying" variable

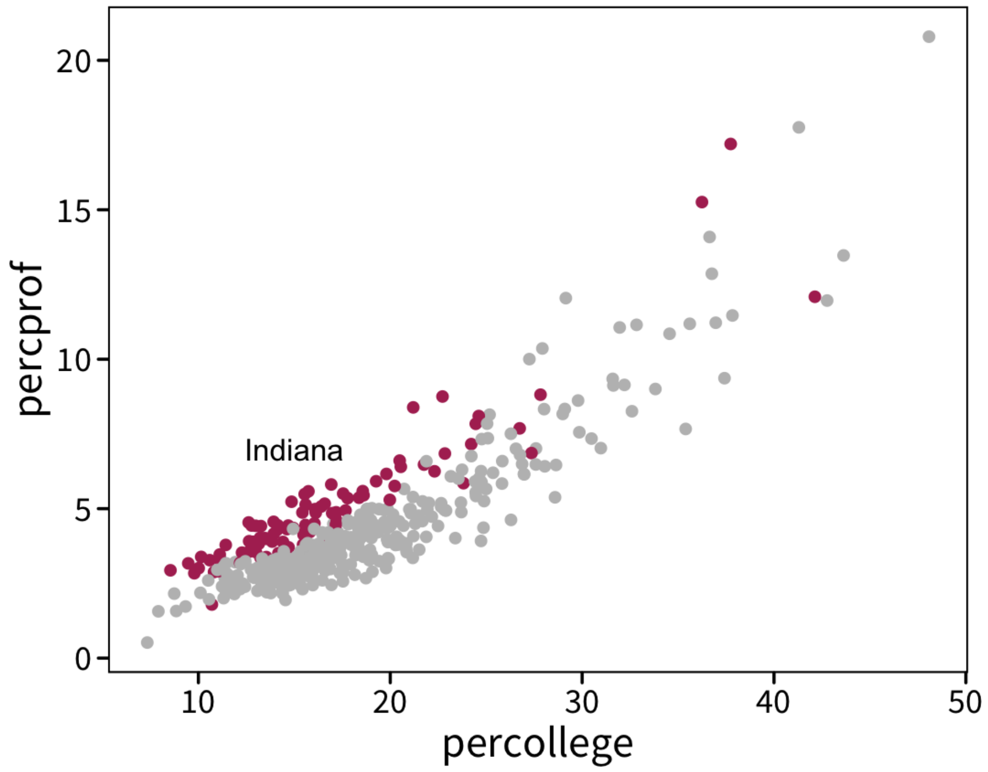
```
# logical statements are TRUE or FALSE
midwest$state == "IN"

# new logical variable
midwest2 <- midwest %>%
  mutate(is_indiana = (state == "IN"))

# map color to is_indiana
ggplot(data = midwest2,
       aes(x = percollege, y = percprof)) +
  geom_point(aes(color = is_indiana))
```



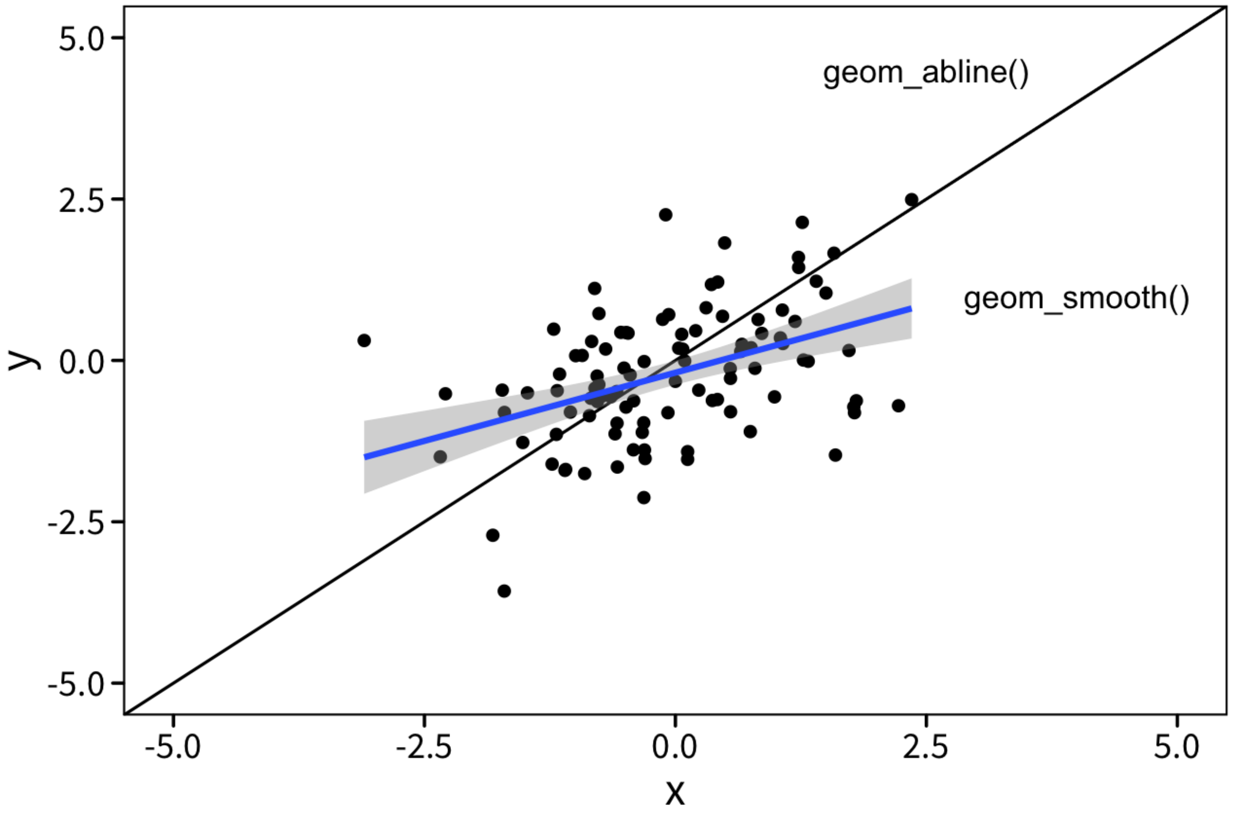
Helpful code tricks:

Hide legend

```
# also: create identifier within aes() ?
ggplot(data = midwest,
       aes(x = percollege, y = percprof)) +
  geom_point(
    aes(color = (state == "IN")),
    show.legend = FALSE,
  ) +
  annotate(geom = "text",
           x = 15, y = 7, label = "Indiana") +
  # customize colors, scale_aes_*()
  scale_color_manual(
    values = c("TRUE" = "maroon",
               "FALSE" = "gray")
  )
```



One last thing: 45° line at $y = x$

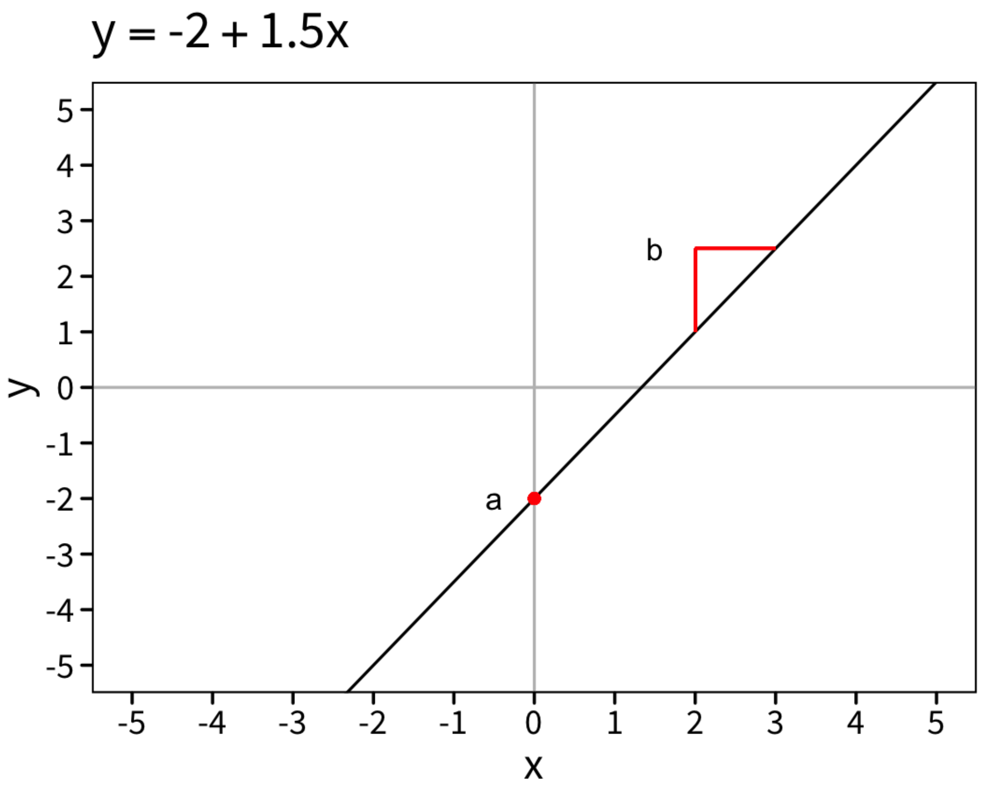


Linear Regression

Lines

A line is $y = a + bx$

- x and y: data
- a and b: parameters / coefficients
  - a is constant/y-intercept
  - b is slope



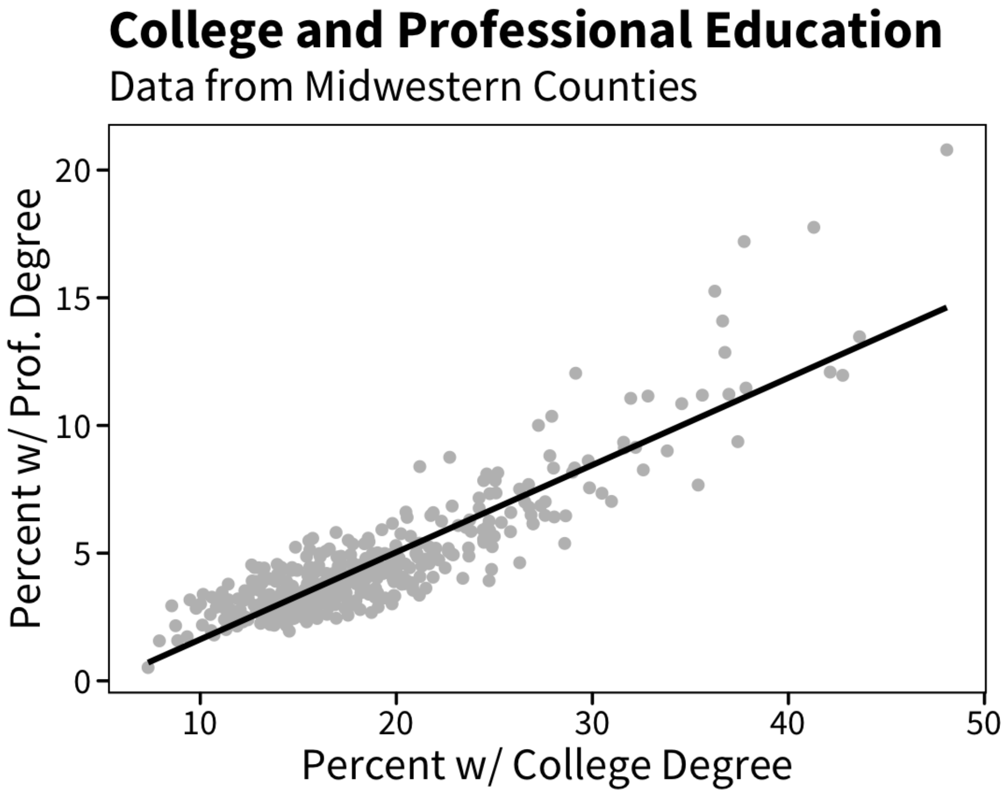
The Linear "Model"

Model: mathematical/statistical assumptions about your data

- "I think a line is a good way to summarize the relationship between X and Y"
- $\mathrm{E}[Y \mid X] \neq \mathrm{E}[Y]$
- The *conditional mean* of Y

Estimating (or "fitting") a model

- Intercept or slope are unknown
- Estimated (imperfectly) from data



"All models are wrong; some are useful"

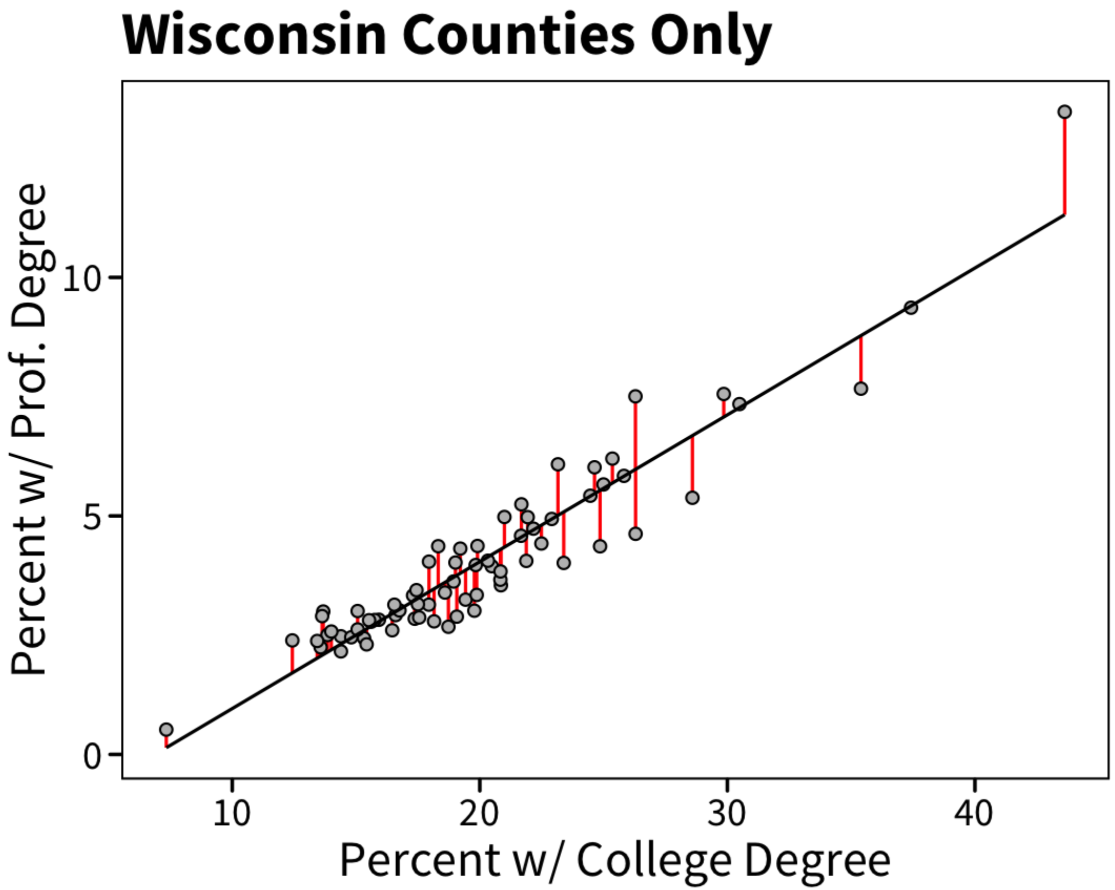
— George Box (?)

"Line of Best Fit"

Line is prediction for y, using knowledge of x

- actual y: points (data)
- prediction line: $\hat{y} = a + bx$
- residual error (in red): $e = y - \hat{y}$

"Give me the line (that is, a and b values) that result in lowest error"



An Equation for y_i

The y value for observation i

An Equation for y_i

The y value for observation i

- x_i (we know their x value)

An Equation for y_i

The y value for observation i

- x_i (we know their x value)
- $\hat{y}_i = a + bx_i$ (predicted y is on the line)

An Equation for y_i

The y value for observation i

- x_i (we know their x value)
- $\hat{y}_i = a + bx_i$ (predicted y is on the line)
- $y_i = \hat{y}_i + e_i$ (actual y is predicted + error)

An Equation for y_i

The y value for observation i

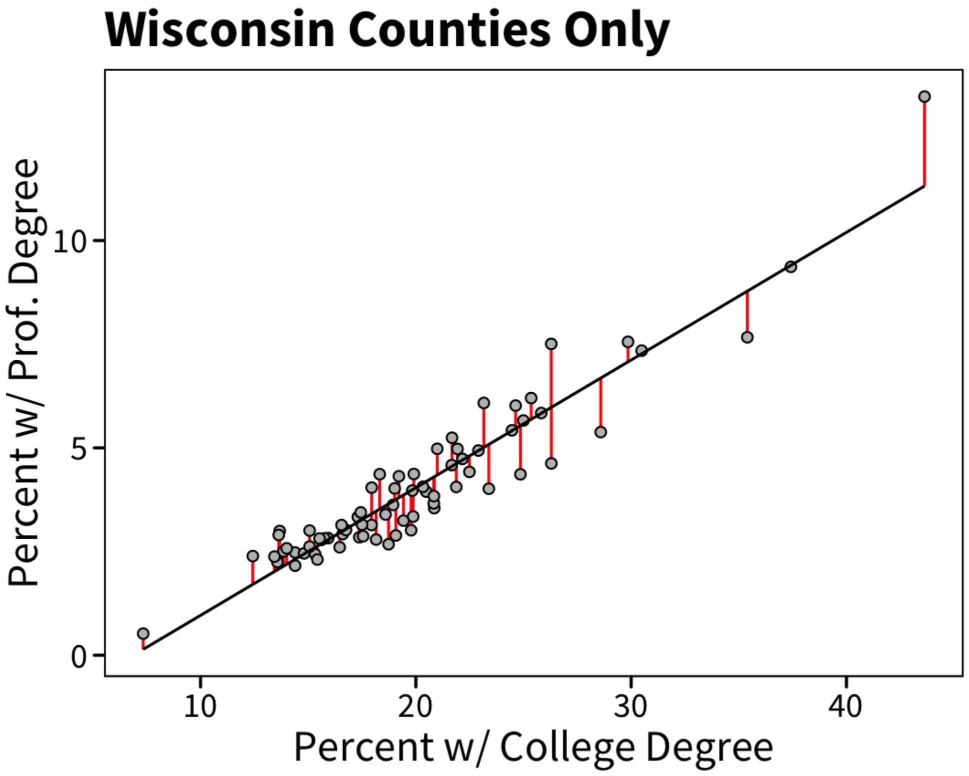
- x_i (we know their x value)
- $\hat{y}_i = a + bx_i$ (predicted y is on the line)
- $y_i = \hat{y}_i + e_i$ (actual y is predicted + error)
- $y_i = a + bx_i + e_i$

An Equation for y_i

The y value for observation i

- x_i (we know their x value)
- $\hat{y}_i = a + bx_i$ (predicted y is on the line)
- $y_i = \hat{y}_i + e_i$ (actual y is predicted + error)
- $y_i = a + bx_i + e_i$

Systematic and random components



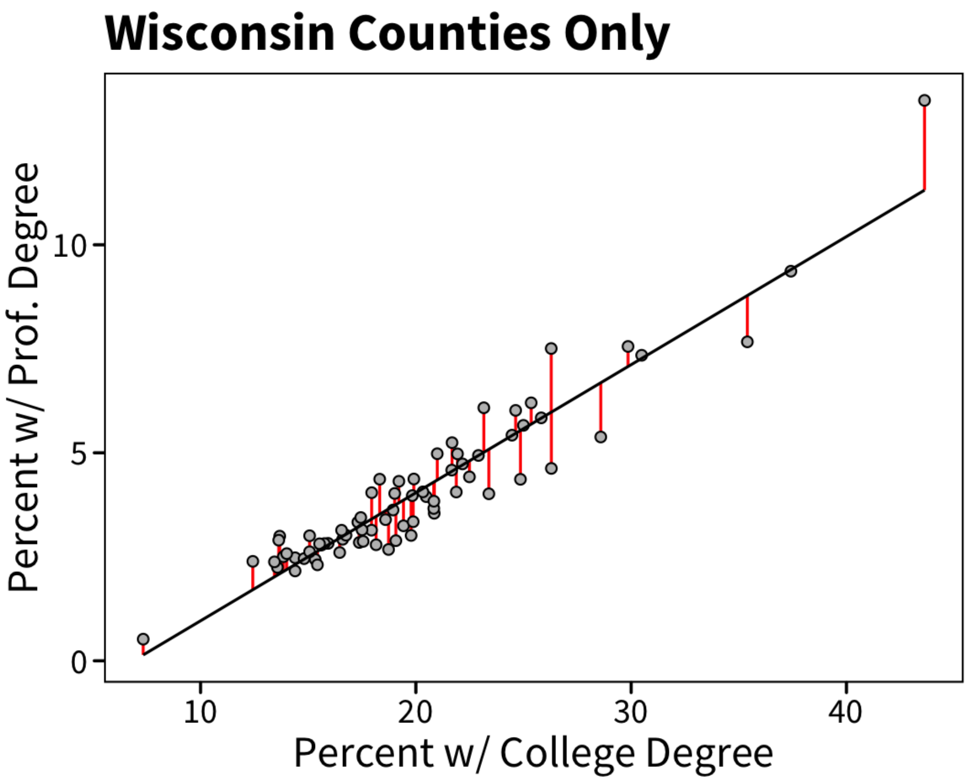
The "Ordinary Least Squares" (OLS) algorithm

Find a and b that minimize the total amount of error

Starting point: $y_i = a + bx_i + e_i$

Total error: $\sum_{i=1}^{N} e_i$

Problem?



The "Ordinary Least Squares" (OLS) algorithm

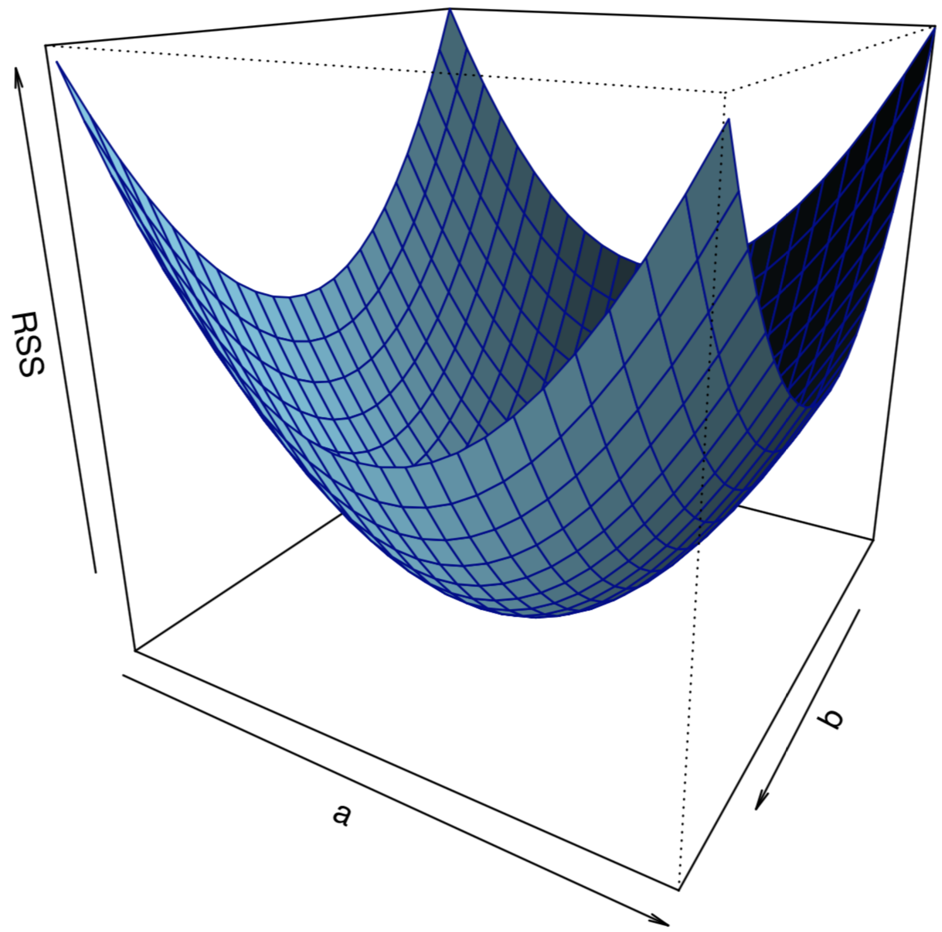
Find a and b that minimize the total amount of **squared** error

Starting point: $y_i = a + bx_i + e_i$

Total **squared** error: $\sum_{i=1}^{N} e_i^2$

$$
\begin{aligned}
y_i &= a + bx_i + e_i \\
e_i &= y_i - (a + bx_i) \\
e_i^2 &= (y_i - (a + bx_i))^2 \\
\sum_i e_i^2 &= \sum_i (y_i - (a + bx_i))^2
\end{aligned}
$$

Then minimize along a and b (calculus!)



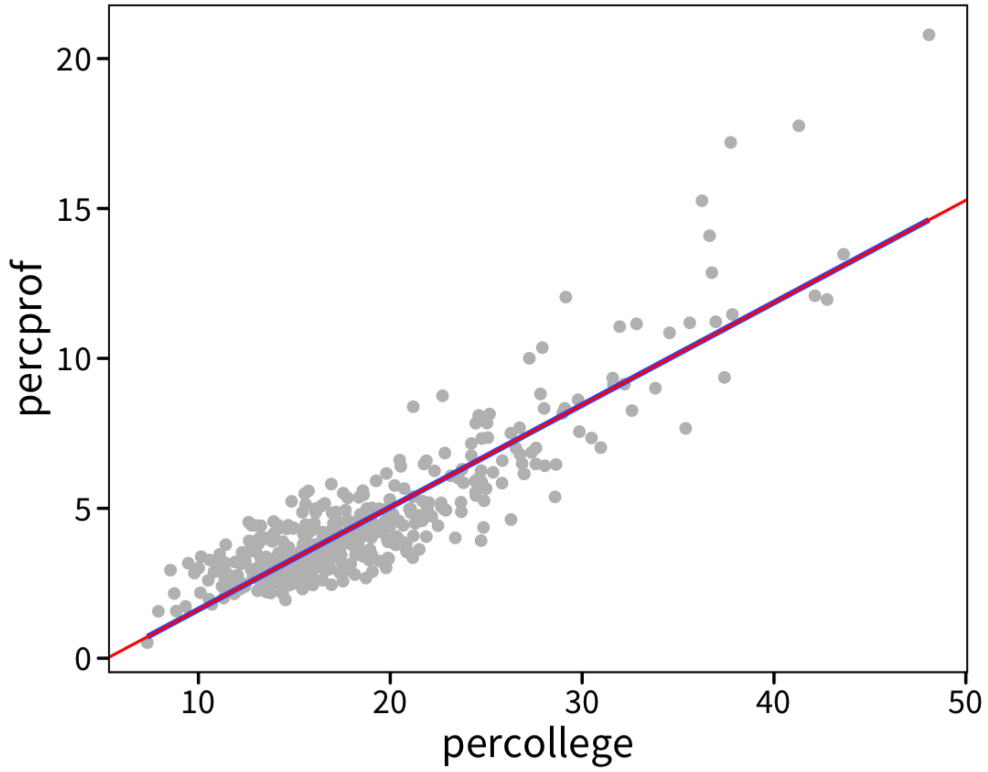
In R

```r
# y ~ x
linreg <- lm(percprof ~ percollege,
             data = midwest)
linreg
```

```
##
## Call:
## lm(formula = percprof ~ percollege, data = midwest)
##
## Coefficients:
## (Intercept)   percollege
##     -1.7899       0.3413
```

```r
ggplot(midwest,
       aes(percollege, y = percprof)) +
  geom_point(color = "gray") +
  geom_smooth(method = "lm", se = FALSE) +
  geom_abline(intercept = coef(linreg)[1],
              slope = coef(linreg)[2],
              color = "red")
```

Estimated equation is $\hat{y}_i = -1.79 + 0.34x_i$



Interpretation

```
##
## Call:
## lm(formula = percprof ~ percollege, data = midwest)
##
## Coefficients:
## (Intercept)   percollege
##     -1.7899       0.3413
```

I predict `percprof` is -1.79 when `percollege` is 0

I predict `percprof` increases by 0.34 when `percollege` increases by 1

Not necessary *cause and effect*, just a predicted change

Interpretation

```
##
## Call:
## lm(formula = percprof ~ percollege, data = midwest)
##
## Coefficients:
## (Intercept)   percollege
##     -1.7899       0.3413
```

I predict `percprof` is -1.79 when `percollege` is 0

I predict `percprof` increases by 0.34 when `percollege` increases by 1

Not necessary *cause and effect*, just a predicted change

Limitations

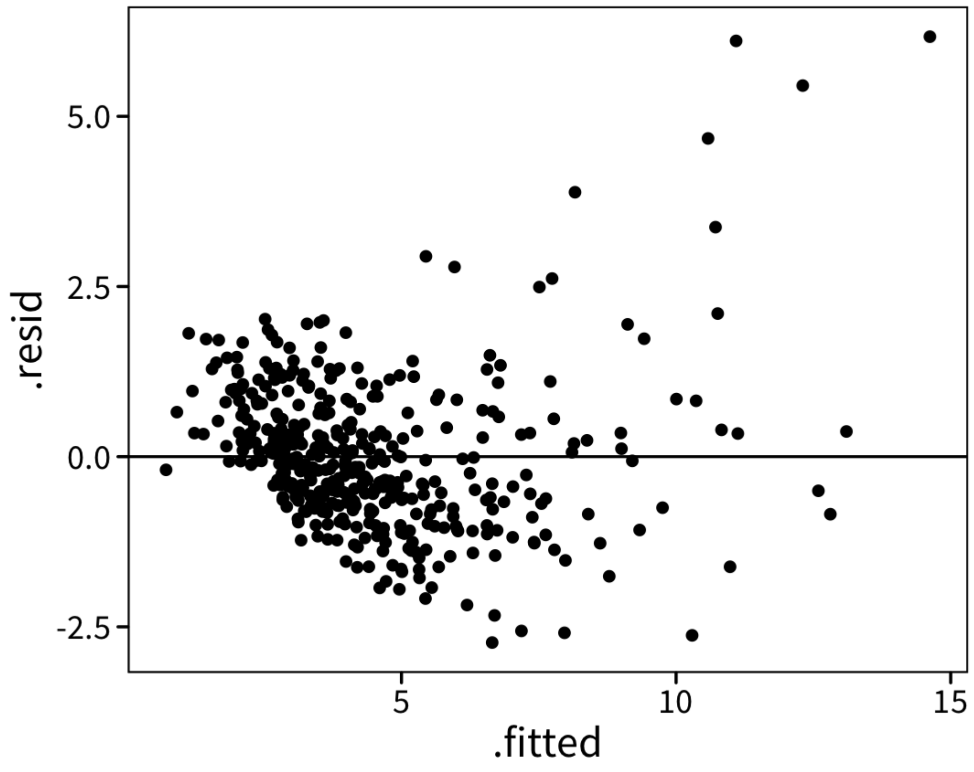
- These are just predictions (average vs. individual, population vs. sample!)
- Extrapolating beyond data = danger
- Nonsense predictions?
- Understand your assumptions ("linear enough?")
- Are there other methods out there?

Check your model

Linear models assume that errors are independent and uncorrelated

- Plot residuals vs predicted values
- Should see no pattern
- Patterns indicate something systematically wrong w/ model

```
# contains handy modeling tools
# install.packages("broom")

# use augment() from broom package
preds <- broom::augment(linreg)

ggplot(preds, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



Looking forward

On Wednesday:

- Statistical significance: How confident are we that this relationship is *real* or *just random*?

In section:

- Practicing `lm` and interpreting linear models

Through the week:

- research question meetings

Next week:

- Monday: Multiple regression (**Research questions due**)
- Wednesday: Gathering and cleaning data (**Ex 2 due**)