POLI210: Political Science Research Methods

Lecture 12.2: Linear regression

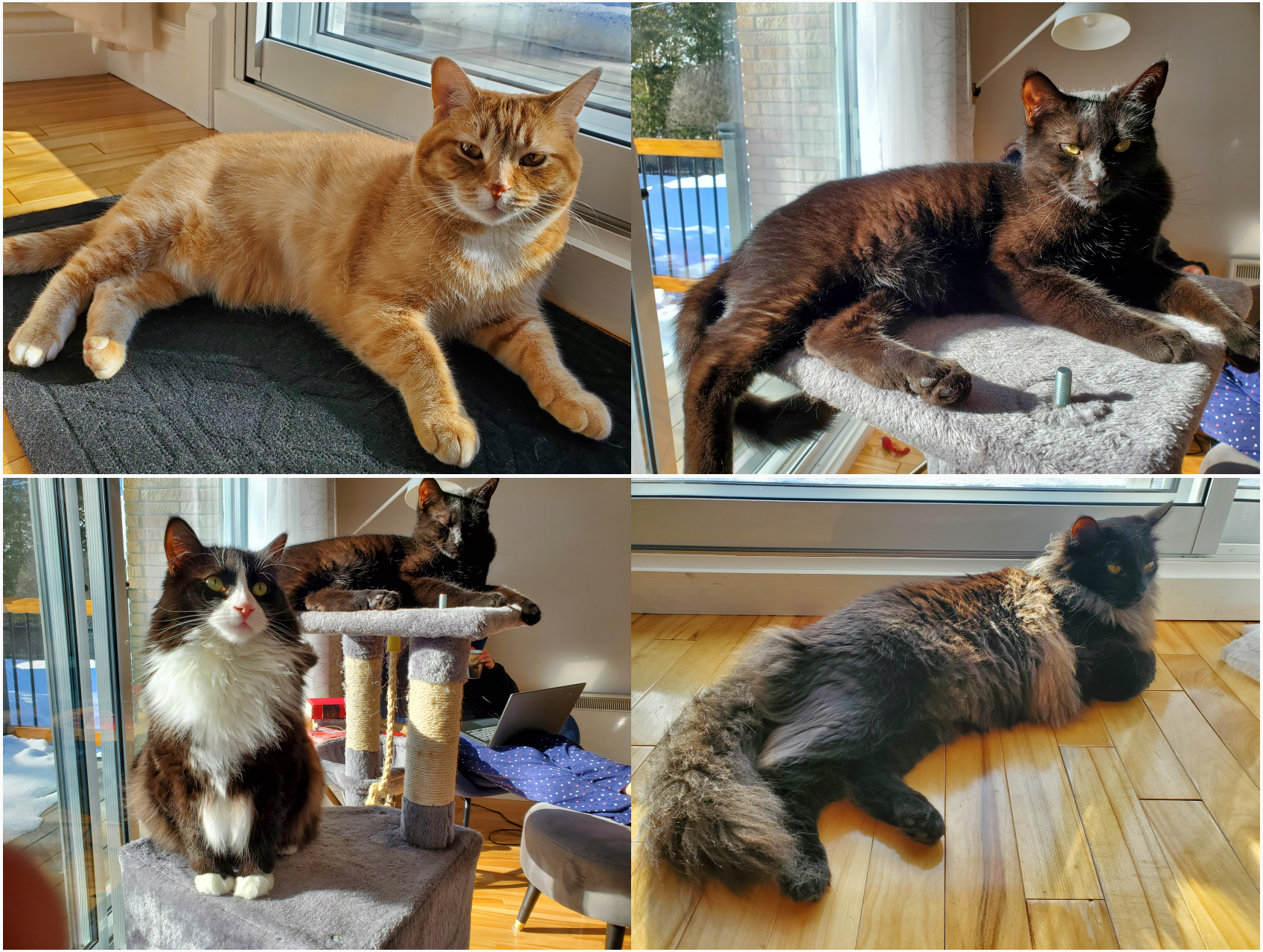
Olivier Bergeron-Boutin

November 23rd, 2021

Boring admin stuff

- More appointments with me available
- Lots of tutoring sessions
- I know there's a lot going on
    - I'm offering as much help as I can – use it!

Cats

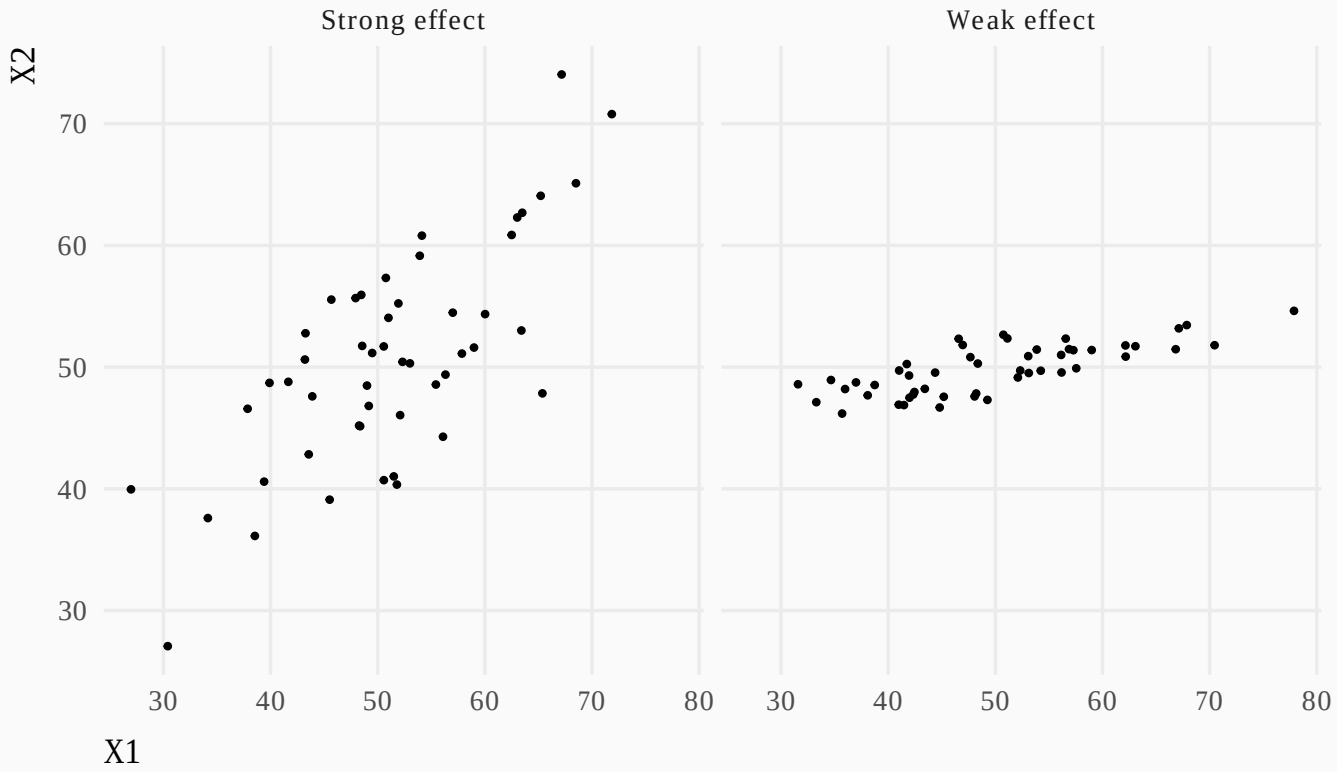


The limitations of correlation coefficients

Two limitations:

- Does not give an estimate of the **magnitude** of the effect
  - If X increases by one unit, by how much can I expect Y to change?
- Does not allow us to "control" for other variables
  - By "controlling" for confounders, we will be able to make more plausible claims about causality

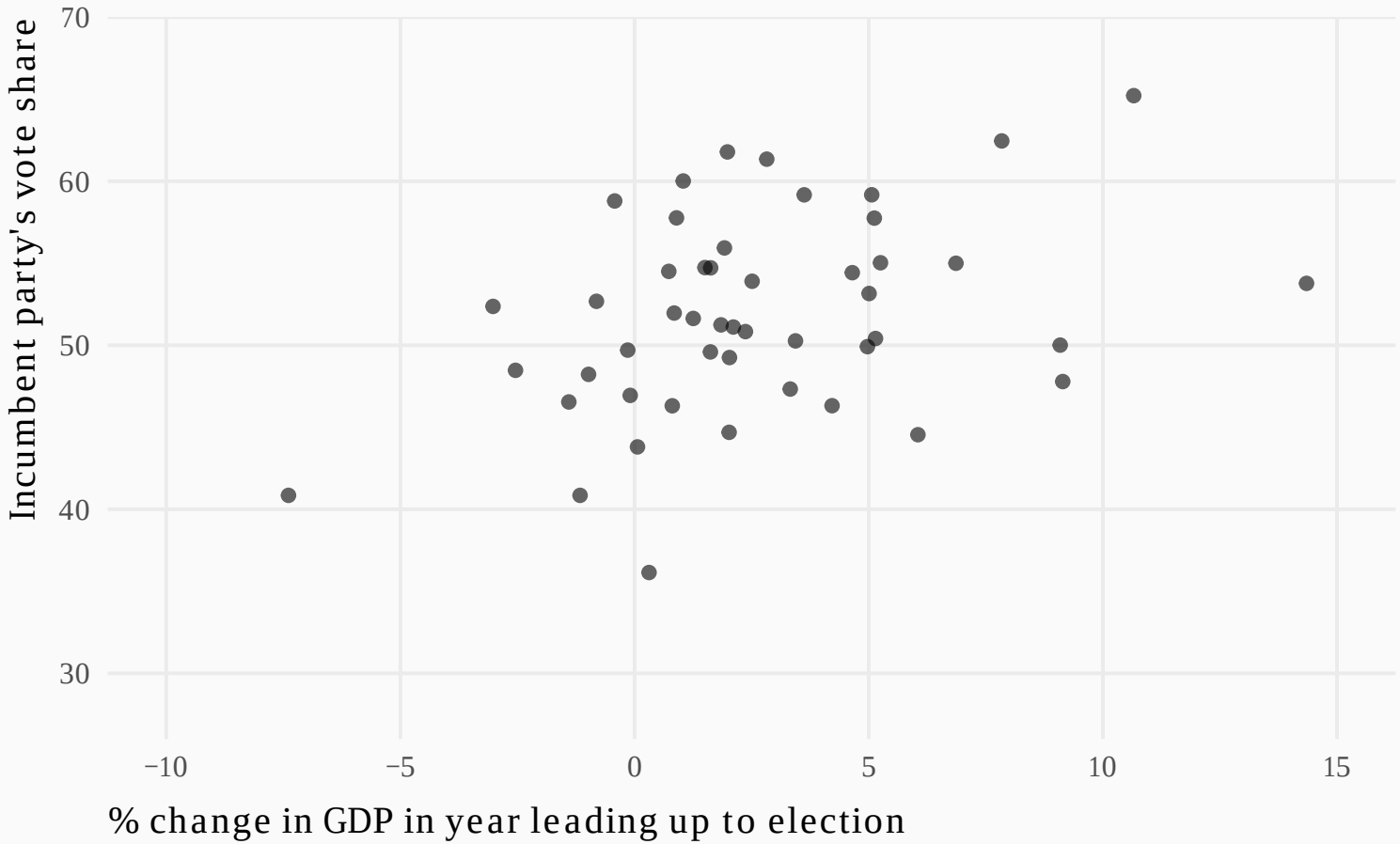
Correlation does not indicate magnitude of the effect




```
## [1] 0.7425742
```

```
## [1] 0.7616742
```

What we want to do



Our objective: draw a line through the points that best represents the relationship

Back to middleschool

We can represent lines in a graph using the following equation:

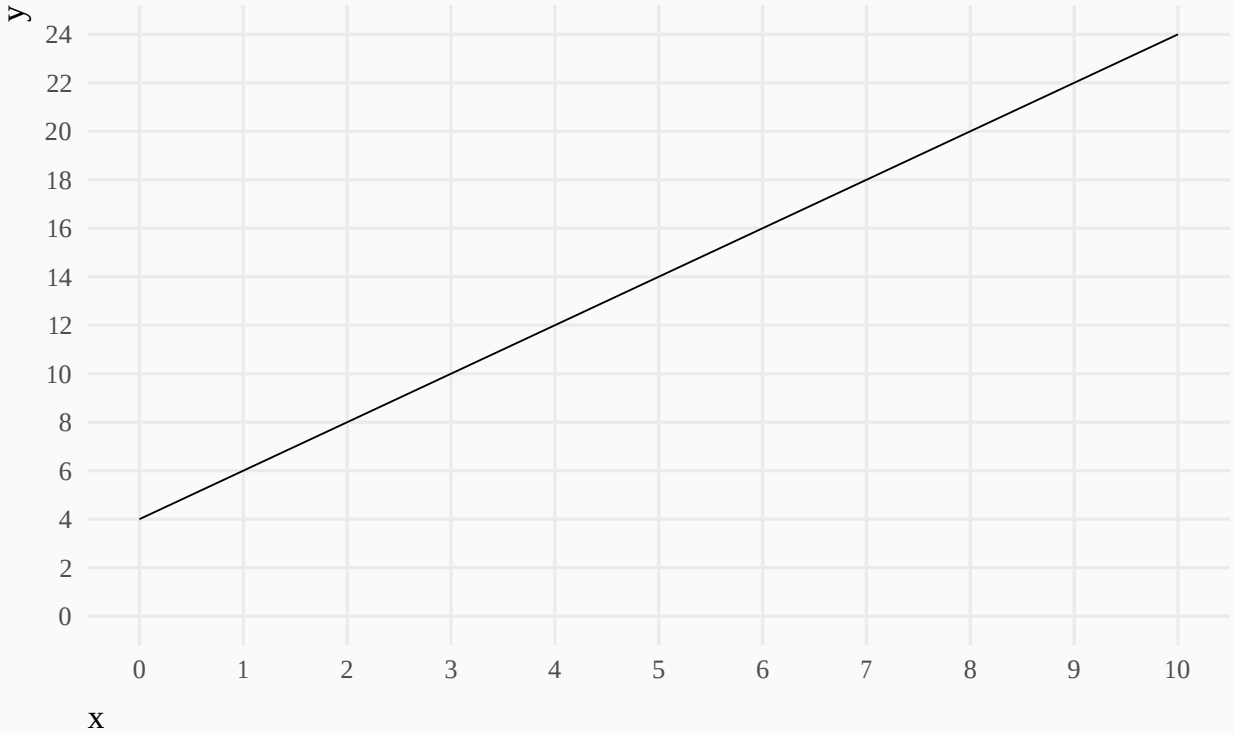
$f(x) = ax + b$

- $f(x)$: the value of y; it's determined by the right-hand side of the equation
- ax: some constant multiplied by x
  - a is the slope of my line
- b: the intercept

If I'm given the values a, x, and b, I can find the value of y

A linear function

Let's consider a simple function $f(x) = 2x + 4$



b = 0, because y is equal to 0 when x is equal to 0

a = 2, because for each increase of 1 unit in x, y increases by 2 units

Regression notation

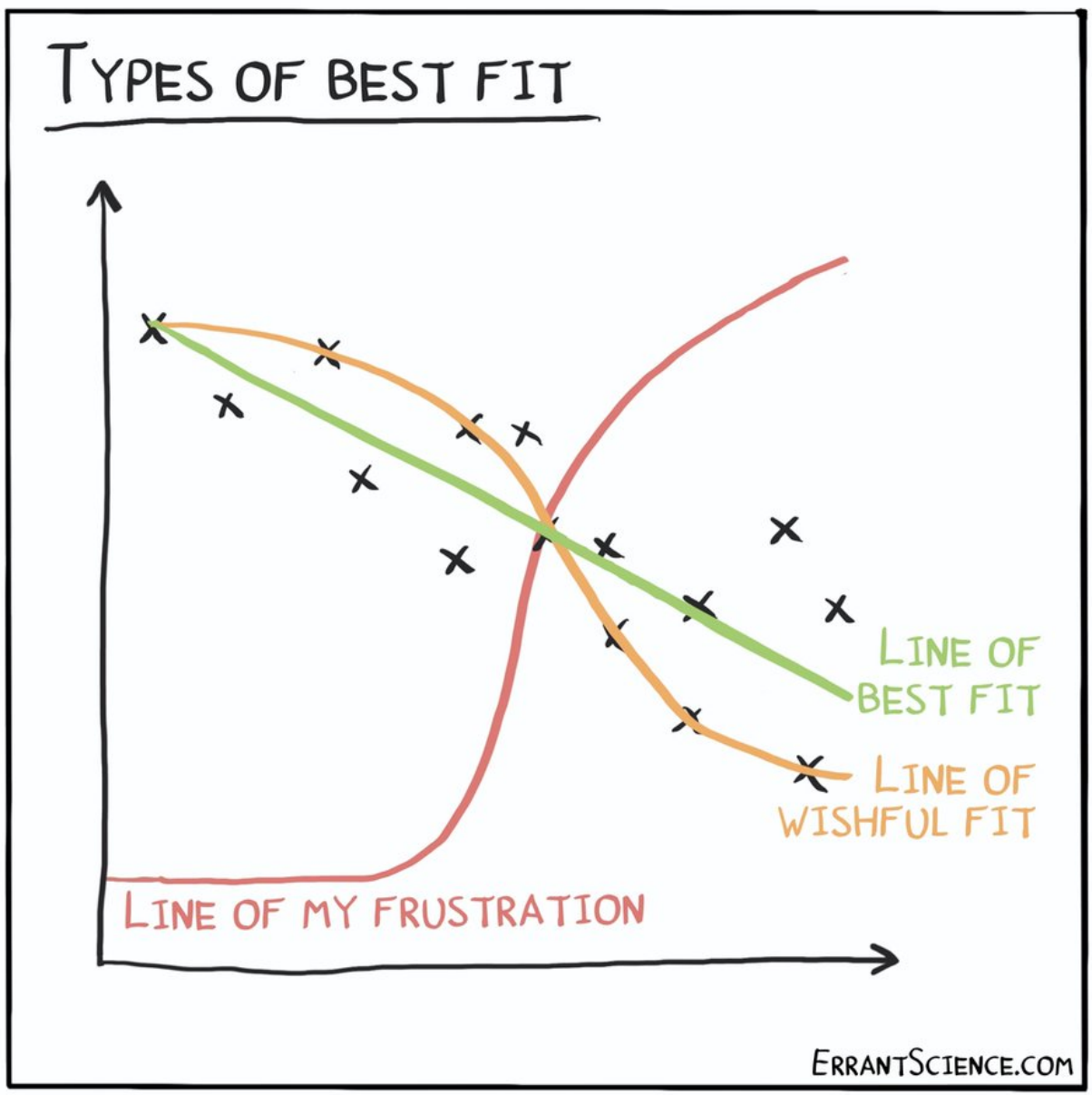
What we'll be doing: fit a line through the points

- We will want to find a rule that allows us to choose the best line
- This is the "line of best fit"

The line of best fit is generally expressed in the following way:

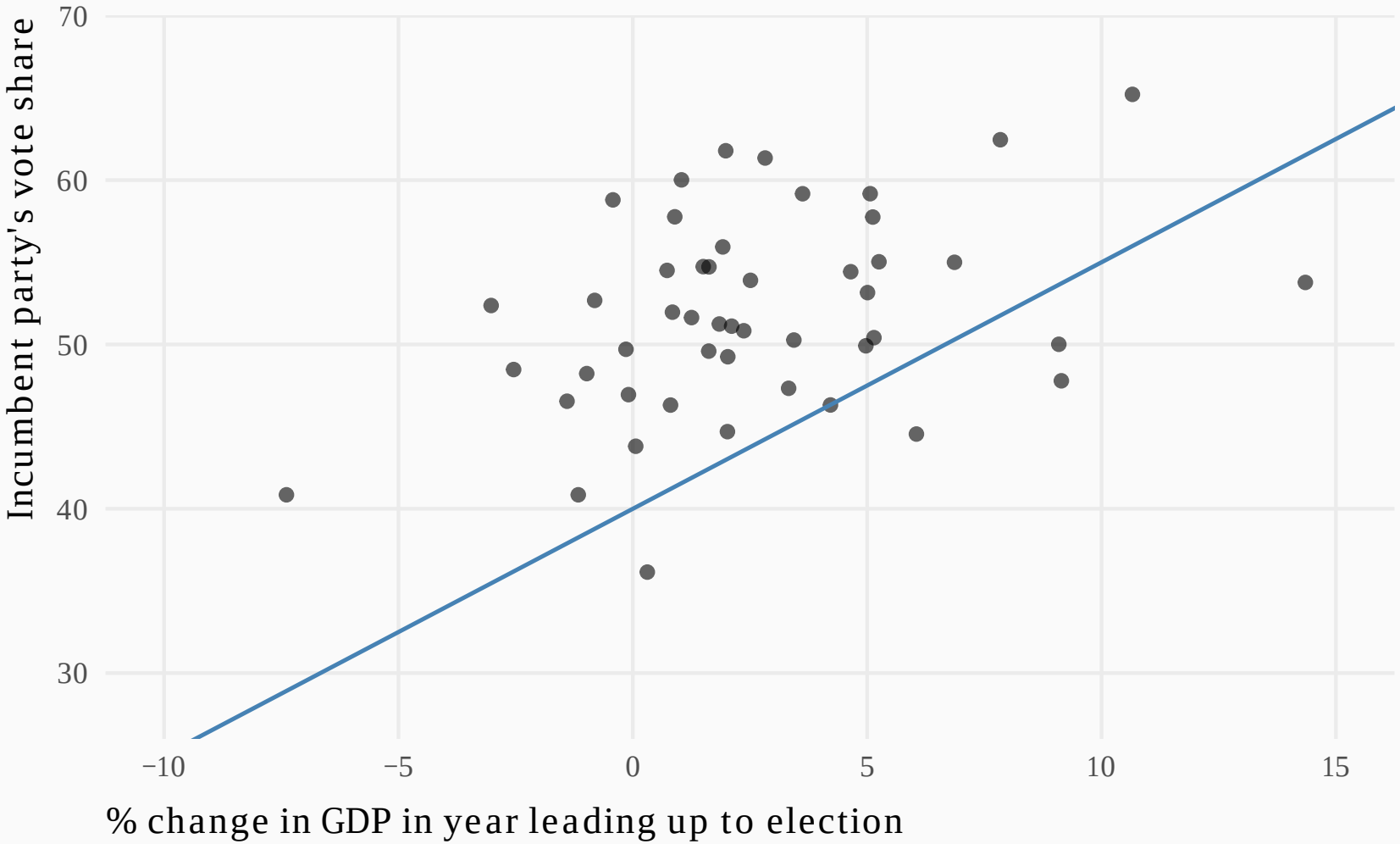
$$Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i$$

Line of best fit or...?



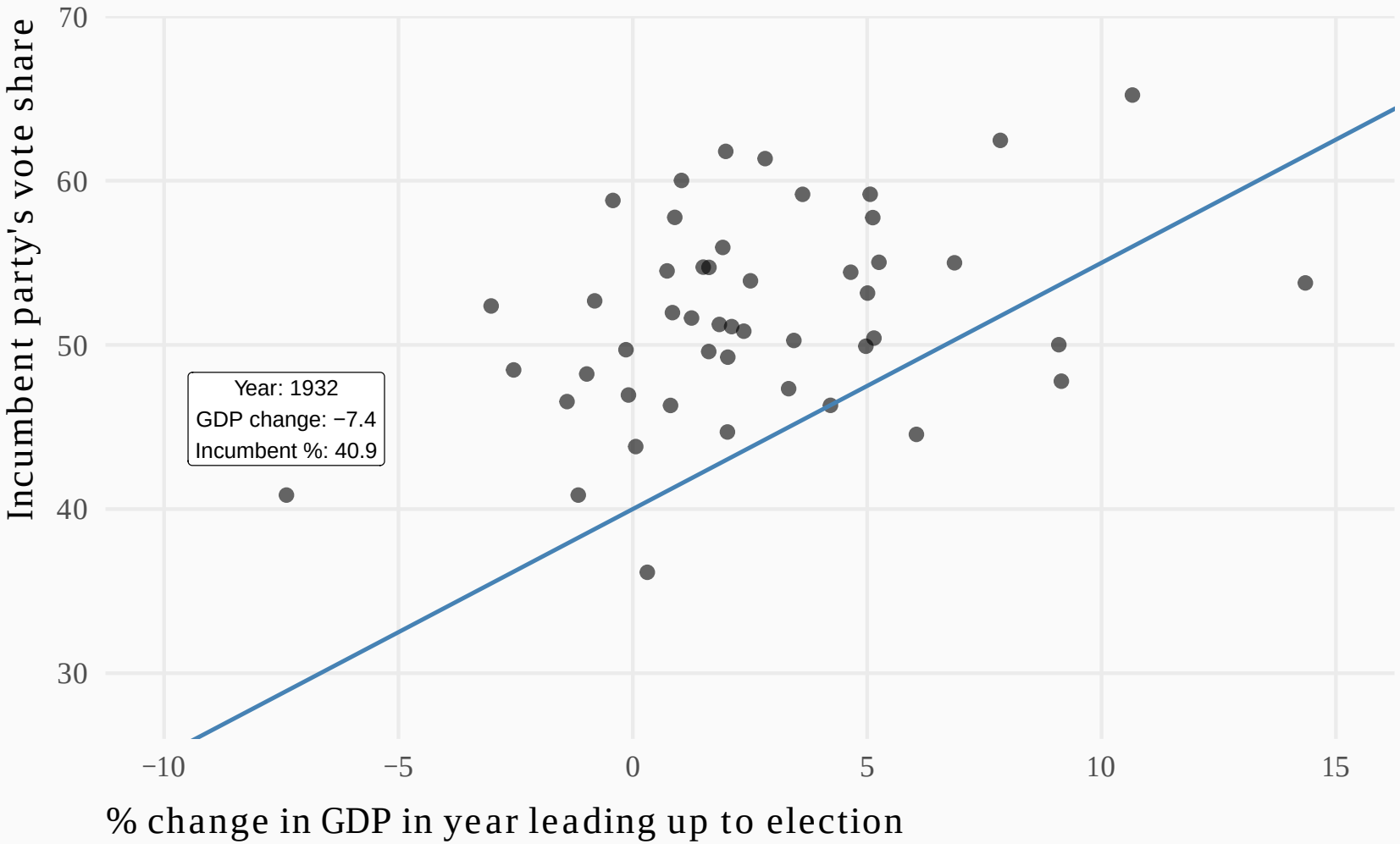
Our first attempt

$$\text{VoteShare}_i = \beta_0 + \beta_1 \text{Growth}_i + \epsilon_i$$



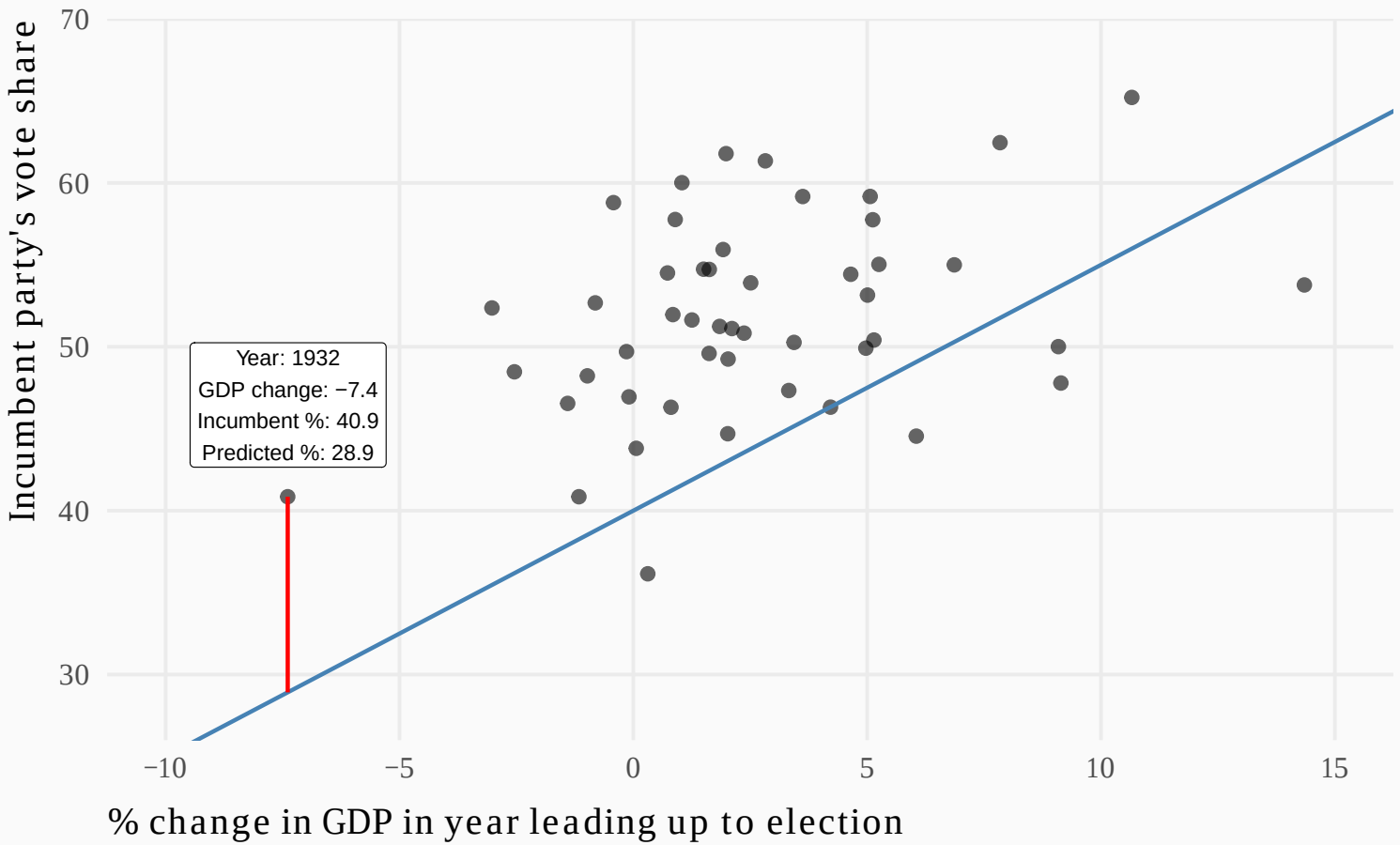
Here, I arbitrarily chose a line: $f(x) = 1.5 * GDP + 40$

Our first attempt



Let's focus on a single point: the 1932 election

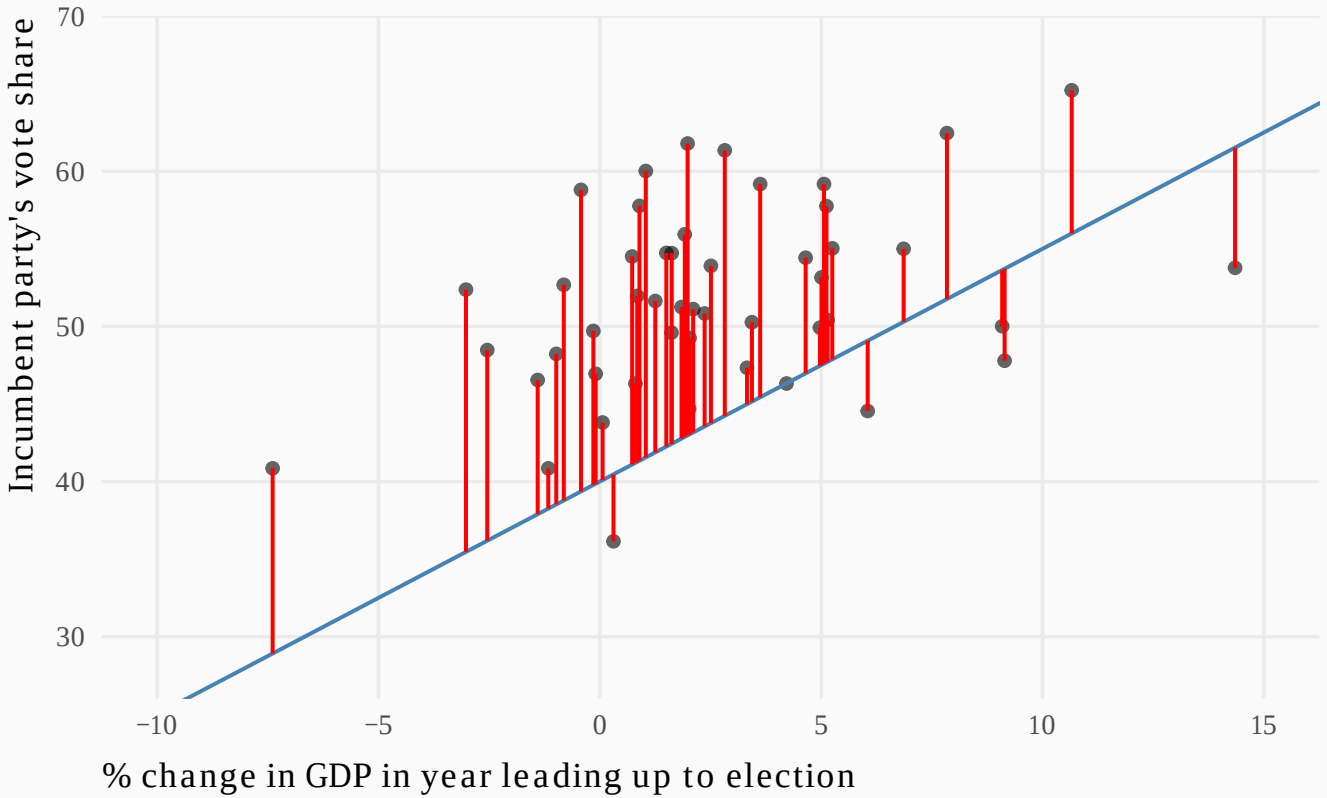
Our first attempt: residual for the 1932 observation



Residual: the difference between the actual outcome and our model's prediction of the outcome

- $\epsilon_i = y_i - \hat{y}_i = 40.9 - 28.9 = 12.0$

Our first attempt: all residuals



- We can compute the residual for each observation
- Why not try to minimize the sum of residuals?
- Some are positive, some are negative; they will cancel out
- Instead, we want to choose a line that minimizes the **sum of squared errors**

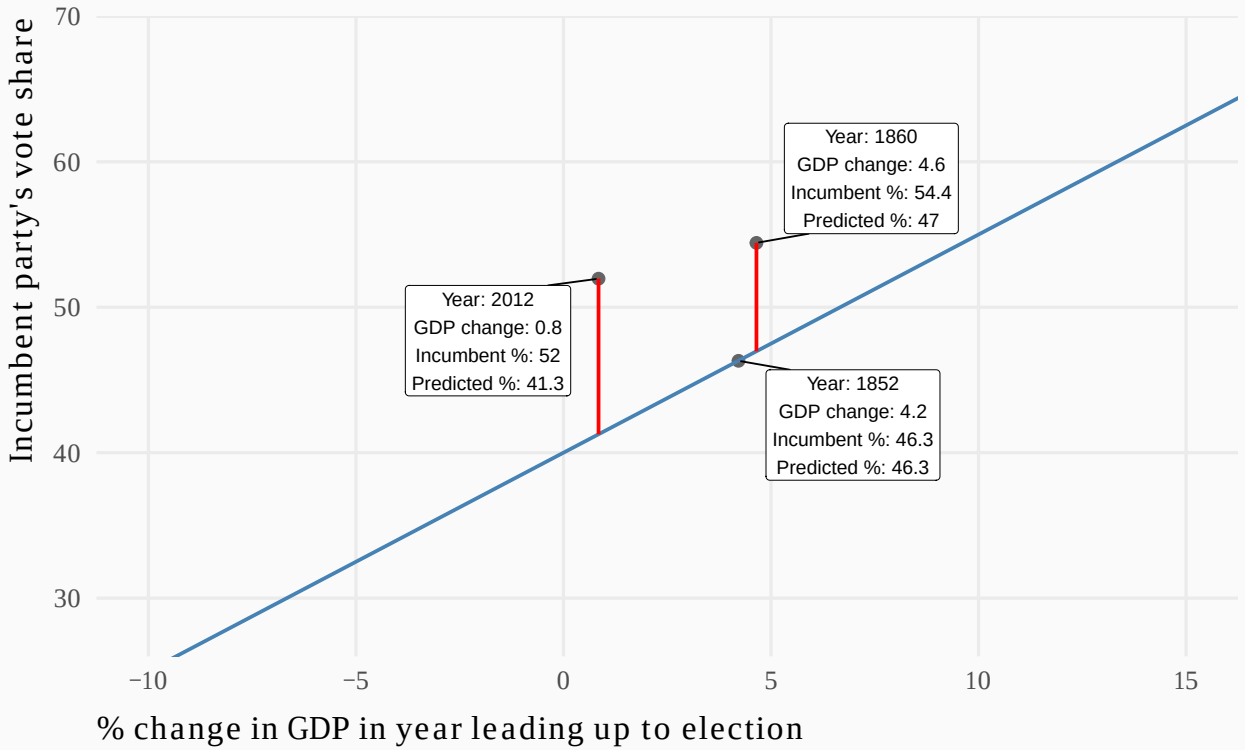
Sum of squared errors

Sum of squared errors (SSE): $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- With $n = 3$: $(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + (y_3 - \hat{y}_3)^2$

Our first attempt: why is it wrong?

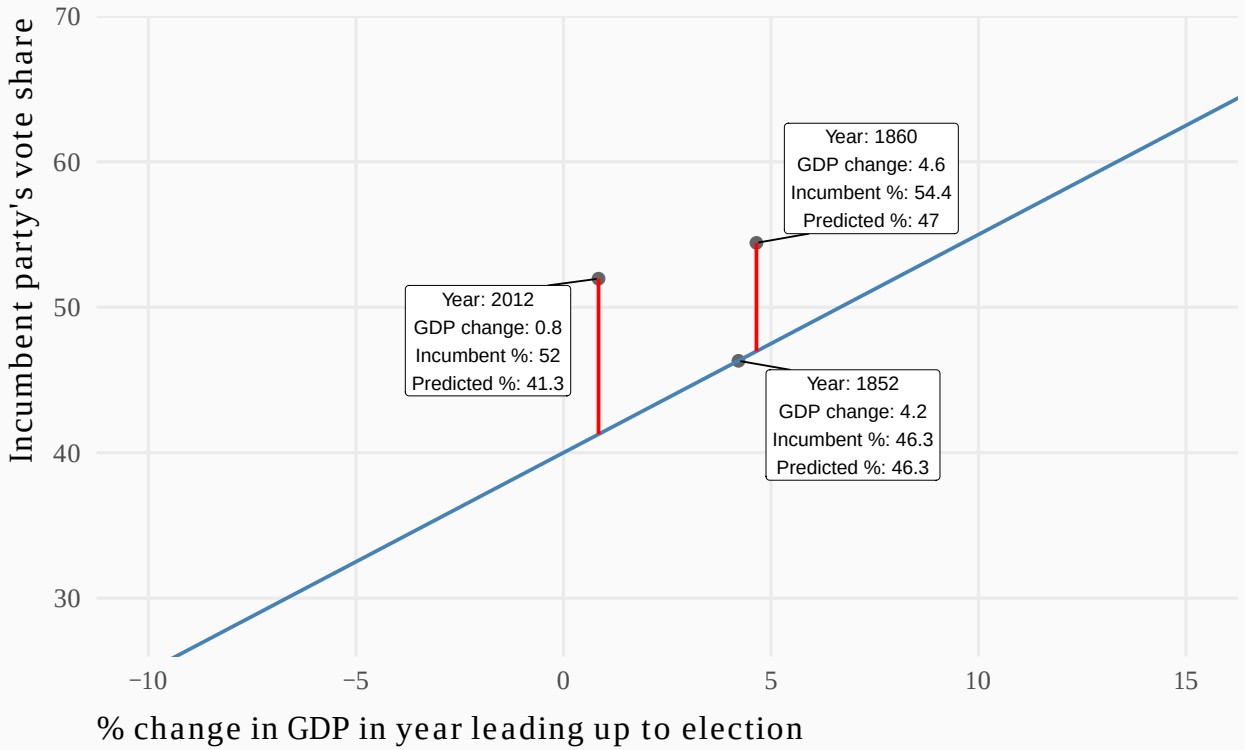
Let's select just 3 observations to simplify the task



- y_i: $52.0, 46.3, 54.4$

Our first attempt: why is it wrong?

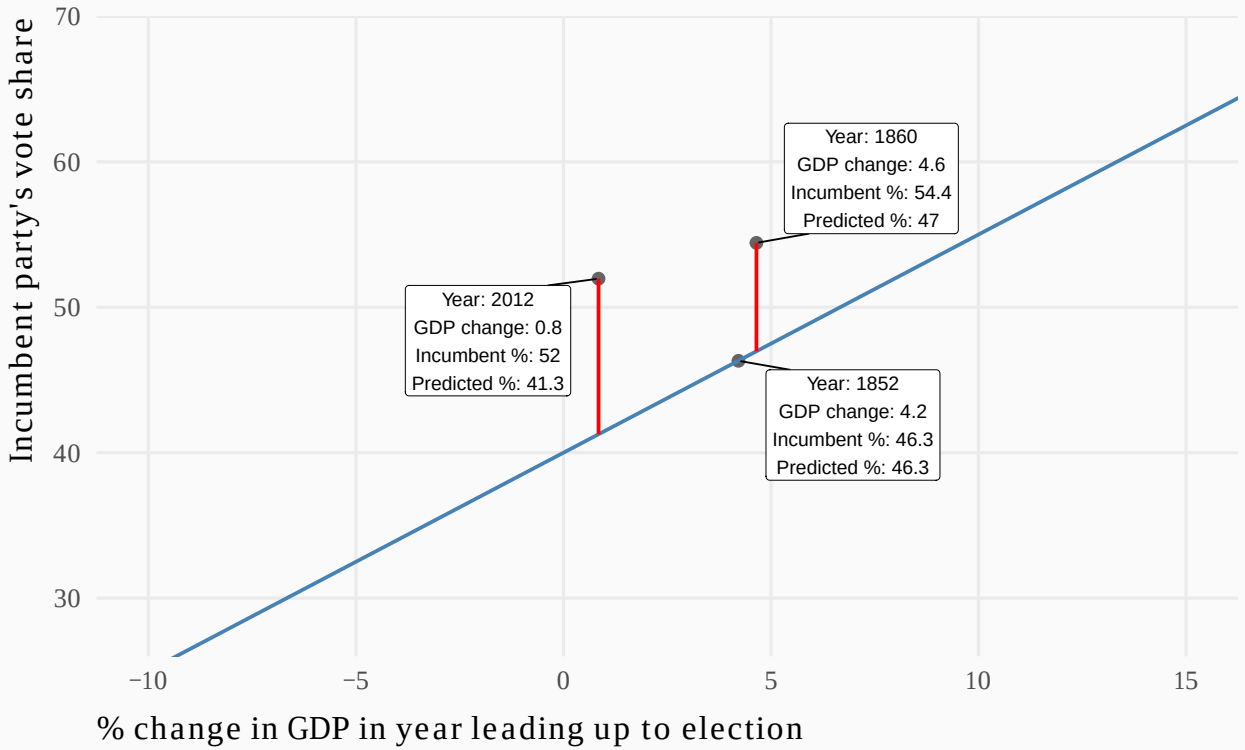
Let's select just 3 observations to simplify the task



- y_i: $52.0, 46.3, 54.4$
- $\hat{y}_i$: $41.3, 46.3, 47.0$

Our first attempt: why is it wrong?

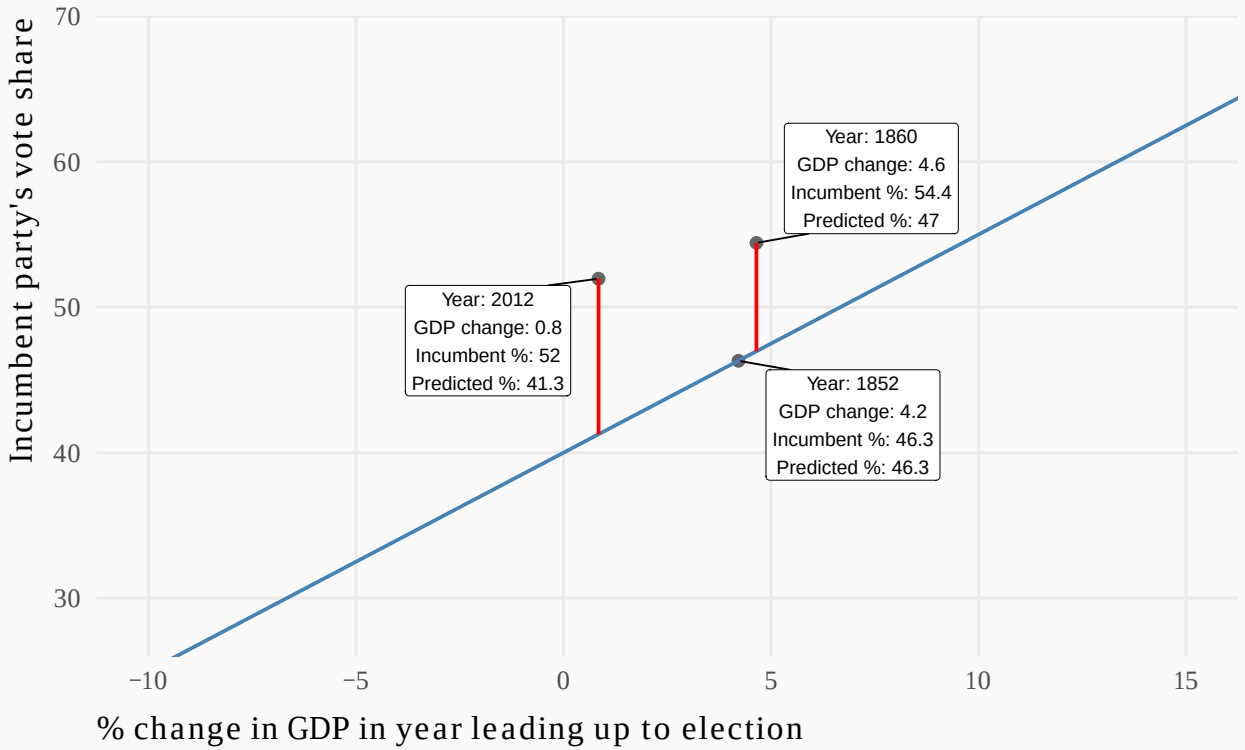
Let's select just 3 observations to simplify the task



- y_i: $52.0, 46.3, 54.4$
- $\hat{y}_i$: $41.3, 46.3, 47.0$
- ϵ_i: $10.7, 00.0, 07.4$

Our first attempt: why is it wrong?

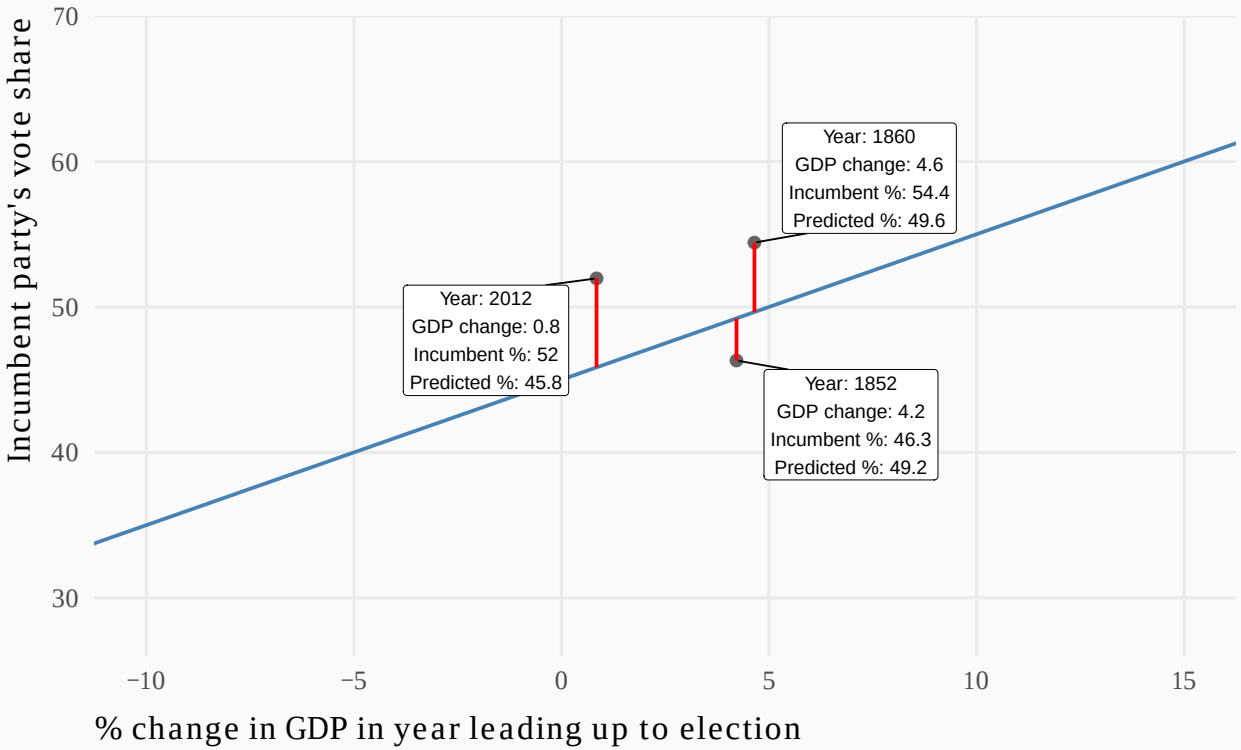
Let's select just 3 observations to simplify the task



- y_i: $52.0, 46.3, 54.4$
- $\hat{y}_i$: $41.3, 46.3, 47.0$
- ϵ_i: $10.7, 00.0, 07.4$
- SSE: $10.7^2 + 0^2 + 7.4^2 = 169.5$

Our first attempt: why is it wrong?

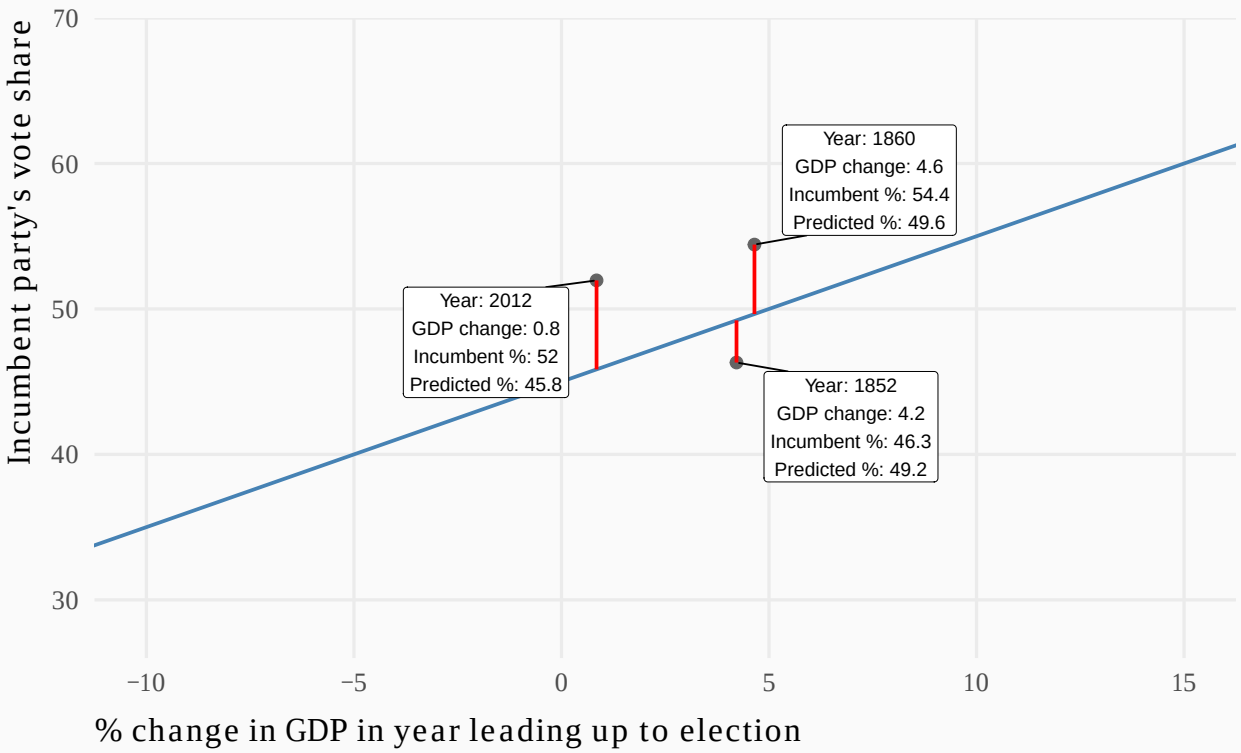
Let's instead use $\beta_0 = 45$ and $\beta_1 = 1$



- y_i: $52.0, 46.3, 54.4$

Our first attempt: why is it wrong?

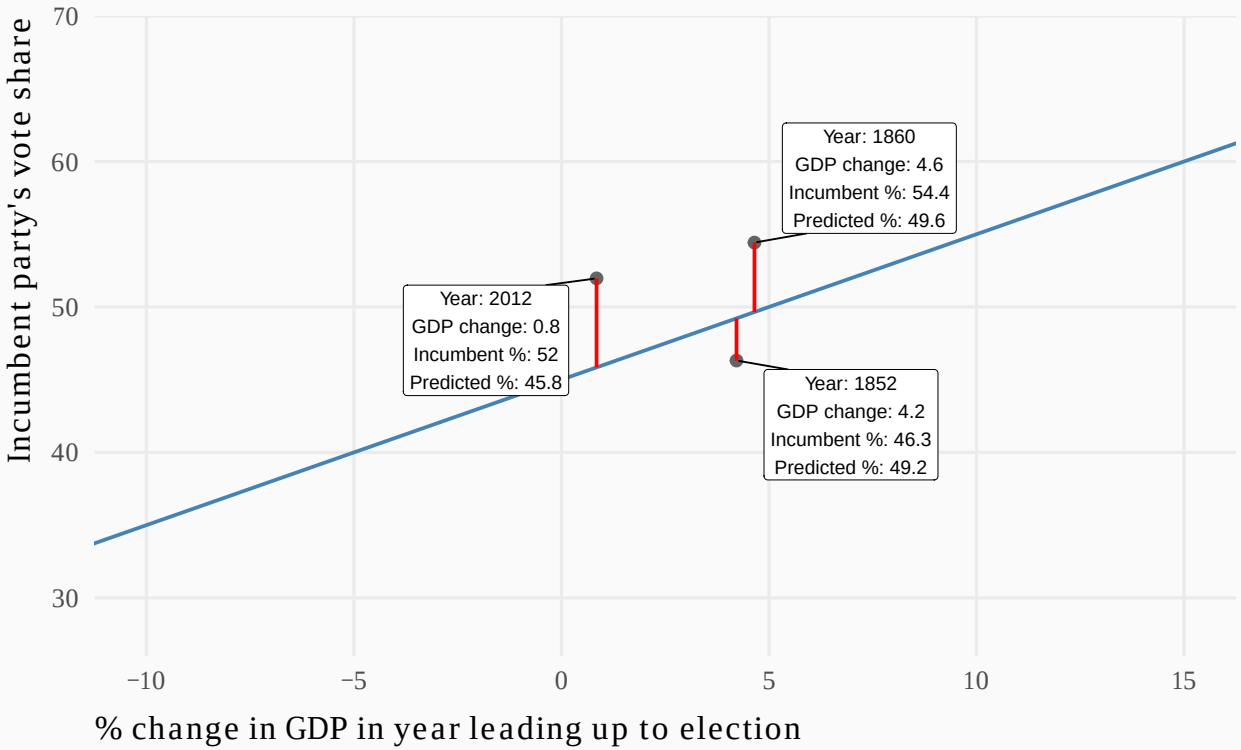
Let's instead use $\beta_0 = 45$ and $\beta_1 = 1$



- y_i: $52.0, 46.3, 54.4$
- $\hat{y}_i$: $45.8, 49.2, 49.6$

Our first attempt: why is it wrong?

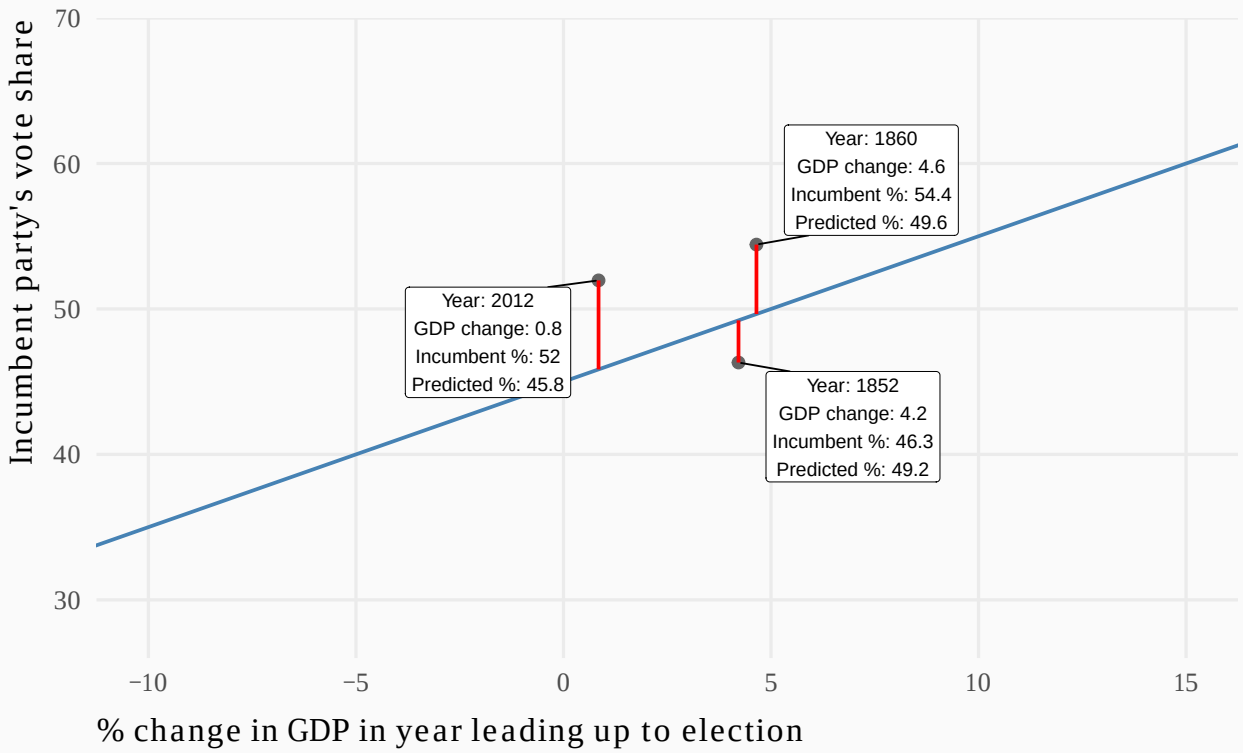
Let's instead use $\beta_0 = 45$ and $\beta_1 = 1$



- y_i: $52.0, 46.3, 54.4$
- $\hat{y}_i$: $45.8, 49.2, 49.6$
- ϵ_i: $6.2, -2.9, 4.8$

Our first attempt: why is it wrong?

Let's instead use $\beta_0 = 45$ and $\beta_1 = 1$



- y_i: $52.0, 46.3, 54.4$
- $\hat{y}_i$: $45.8, 49.2, 49.6$
- ϵ_i: $6.2, -2.9, 4.8$
- SSE: $6.2^2 + -2.9^2 + 4.8^2 = 69.89$

Running our regression

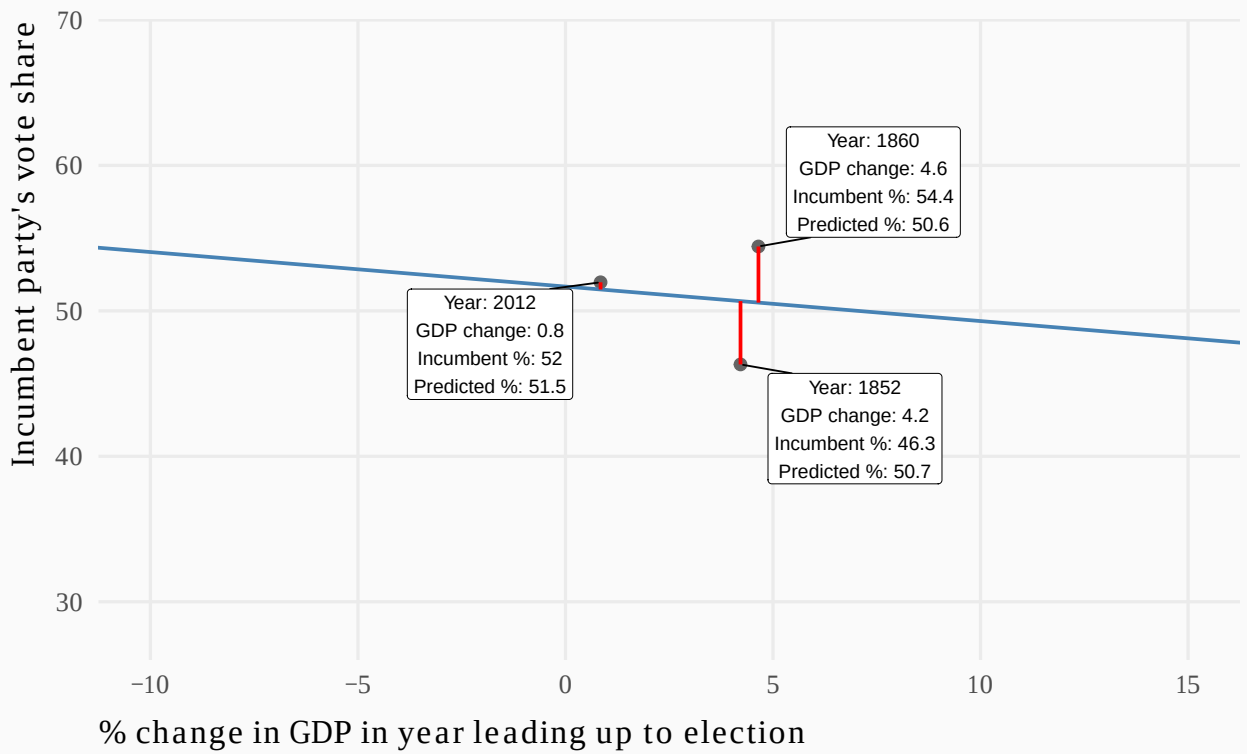
Of course, we don't have to do this by hand

- The command to run a linear regression in R is `lm()`
- Two main arguments:
  - formula, of format `y ~ x`
  - data

```
lm(partyincshr ~ gdpchangeyr3,
   data = subset(economy, year %in% c(1852, 1860, 2012)))
```

```
##
## Call:
## lm(formula = partyincshr ~ gdpchangeyr3, data = subset(economy,
##     year %in% c(1852, 1860, 2012)))
##
## Coefficients:
##  (Intercept)  gdpchangeyr3
##      51.6700       -0.2373
```

Visualizing the correct regression line

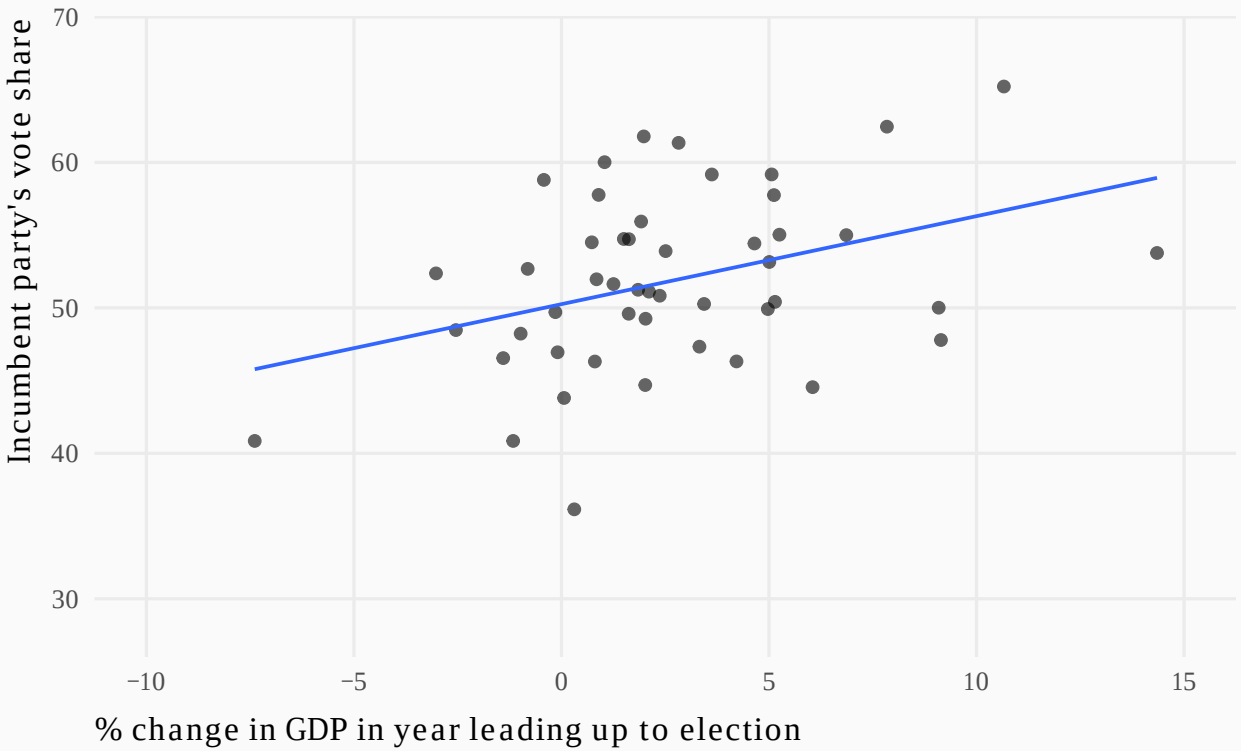


i	y_i	$\hat{y}_i$	ϵ_i	ϵ_i^2
1	46.32	50.67	-4.35	18.92
2	54.42	50.57	3.85	14.82
3	51.96	51.47	0.49	0.24

Sum of Squared Errors:

$$18.92 + 14.82 + 0.24 = 33.98$$

Back to our full data



Linear regression with our full data

```
lm(formula = partyincshr ~ gdpchangeyr3,
   data = economy)
```

```
##
## Call:
## lm(formula = partyincshr ~ gdpchangeyr3, data = economy)
##
## Coefficients:
##  (Intercept)  gdpchangeyr3
##      50.2541        0.6051
```

This is okay...but there's not a lot of information!

Linear regression with our full data

```
lm(formula = partyincshr ~ gdpchangeyr3, data = economy) %>% summary()
```

```
##
## Call:
## lm(formula = partyincshr ~ gdpchangeyr3, data = economy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2925  -3.6163  -0.1858   3.8433  10.3324
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50.2541     0.9992  50.293  < 2e-16 ***
## gdpchangeyr3   0.6051     0.2196   2.755  0.00837 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.653 on 46 degrees of freedom
##   (183 observations deleted due to missingness)
## Multiple R-squared:  0.1417, Adjusted R-squared:  0.123
## F-statistic: 7.592 on 1 and 46 DF,  p-value: 0.008372
```

Interpreting our results

```
##
## Call:
## lm(formula = partyincshr ~ gdpchangeyr3, data = economy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2925  -3.6163  -0.1858   3.8433  10.3324
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50.2541     0.9992  50.293  < 2e-16 ***
## gdpchangeyr3   0.6051     0.2196   2.755  0.00837 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.653 on 46 degrees of freedom
##   (183 observations deleted due to missingness)
## Multiple R-squared:  0.1417, Adjusted R-squared:  0.123
## F-statistic: 7.592 on 1 and 46 DF,  p-value: 0.008372
```

Interpreting our results

```
##
## Call:
## lm(formula = partyincshr ~ gdpchangeyr3, data = economy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2925  -3.6163  -0.1858   3.8433  10.3324
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50.2541     0.9992  50.293  < 2e-16 ***
## gdpchangeyr3   0.6051     0.2196   2.755  0.00837 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.653 on 46 degrees of freedom
##   (183 observations deleted due to missingness)
## Multiple R-squared:  0.1417, Adjusted R-squared:  0.123
## F-statistic: 7.592 on 1 and 46 DF,  p-value: 0.008372
```

Interpreting our results

```
##
## Call:
## lm(formula = partyincshr ~ gdpchangeyr3, data = economy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2925  -3.6163  -0.1858   3.8433  10.3324
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50.2541     0.9992  50.293  < 2e-16 ***
## gdpchangeyr3   0.6051     0.2196   2.755  0.00837 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.653 on 46 degrees of freedom
##   (183 observations deleted due to missingness)
## Multiple R-squared:  0.1417, Adjusted R-squared:  0.123
## F-statistic: 7.592 on 1 and 46 DF,  p-value: 0.008372
```

Interpreting our results

```
##
## Call:
## lm(formula = partyincshr ~ gdpchangeyr3, data = economy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2925  -3.6163  -0.1858   3.8433  10.3324
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50.2541     0.9992  50.293  < 2e-16 ***
## gdpchangeyr3   0.6051     0.2196   2.755  0.00837 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.653 on 46 degrees of freedom
##   (183 observations deleted due to missingness)
## Multiple R-squared:  0.1417, Adjusted R-squared:  0.123
## F-statistic: 7.592 on 1 and 46 DF,  p-value: 0.008372
```

Interpreting our results

```
##
## Call:
## lm(formula = partyincshr ~ gdpchangeyr3, data = economy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2925  -3.6163  -0.1858   3.8433  10.3324
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50.2541     0.9992  50.293  < 2e-16 ***
## gdpchangeyr3   0.6051     0.2196   2.755  0.00837 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.653 on 46 degrees of freedom
##   (183 observations deleted due to missingness)
## Multiple R-squared:  0.1417, Adjusted R-squared:  0.123
## F-statistic: 7.592 on 1 and 46 DF,  p-value: 0.008372
```

Interpreting our results

```
##
## Call:
## lm(formula = partyincshr ~ gdpchangeyr3, data = economy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2925  -3.6163  -0.1858   3.8433  10.3324
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50.2541     0.9992  50.293  < 2e-16 ***
## gdpchangeyr3   0.6051     0.2196   2.755  0.00837 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.653 on 46 degrees of freedom
##   (183 observations deleted due to missingness)
## Multiple R-squared:  0.1417, Adjusted R-squared:  0.123
## F-statistic: 7.592 on 1 and 46 DF,  p-value: 0.008372
```

Interpreting our results

```
##
## Call:
## lm(formula = partyincshr ~ gdpchangeyr3, data = economy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2925  -3.6163  -0.1858   3.8433  10.3324
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50.2541     0.9992  50.293  < 2e-16 ***
## gdpchangeyr3   0.6051     0.2196   2.755  0.00837 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.653 on 46 degrees of freedom
##   (183 observations deleted due to missingness)
## Multiple R-squared:  0.1417, Adjusted R-squared:  0.123
## F-statistic: 7.592 on 1 and 46 DF,  p-value: 0.008372
```

Interpreting our results

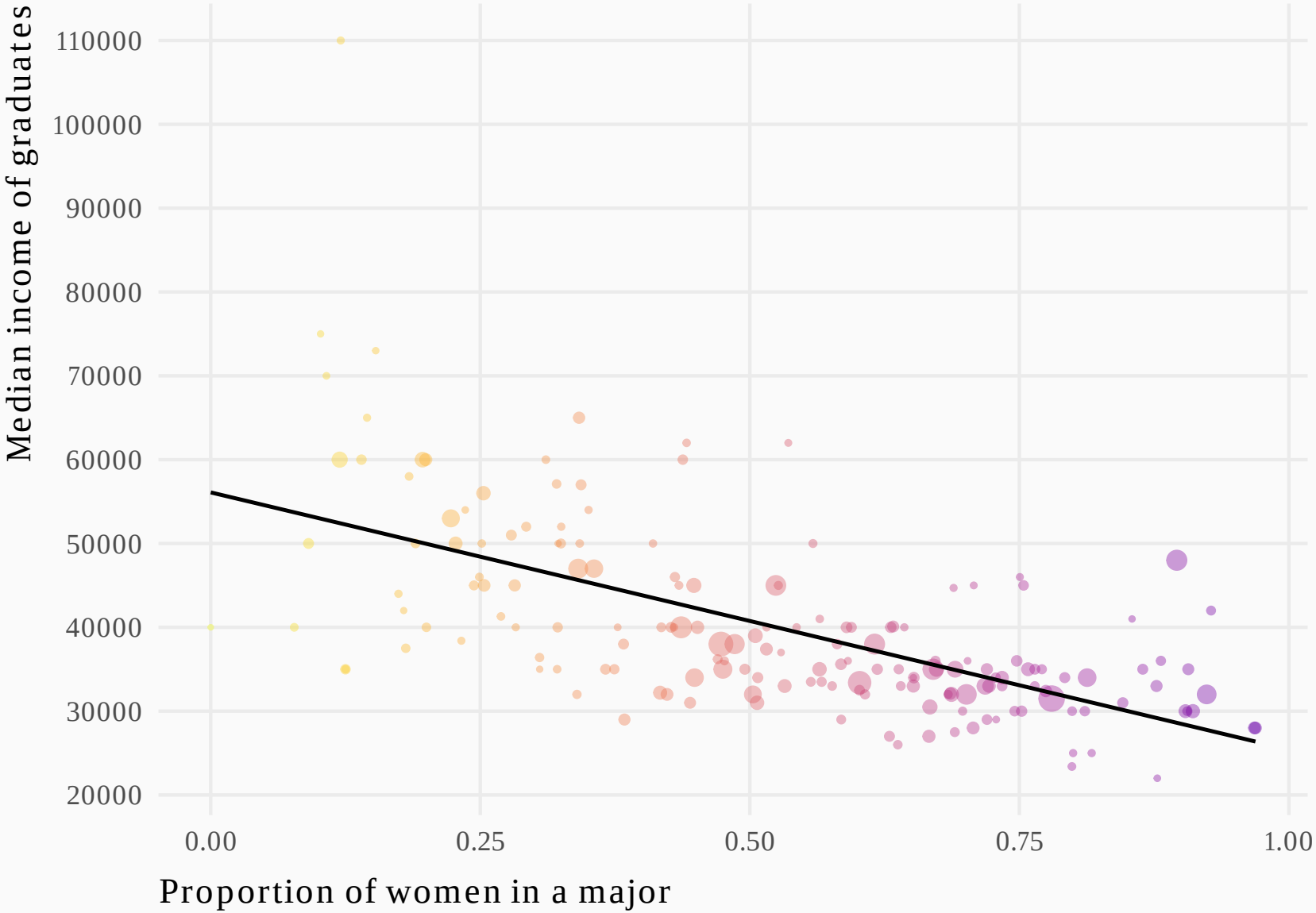
```
##
## Call:
## lm(formula = partyincshr ~ gdpchangeyr3, data = economy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2925  -3.6163  -0.1858   3.8433  10.3324
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50.2541     0.9992  50.293  < 2e-16 ***
## gdpchangeyr3   0.6051     0.2196   2.755  0.00837 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.653 on 46 degrees of freedom
##   (183 observations deleted due to missingness)
## Multiple R-squared:  0.1417, Adjusted R-squared:  0.123
## F-statistic: 7.592 on 1 and 46 DF,  p-value: 0.008372
```

How results generally appear in published work

	Model 1
(Intercept)	50.254***
	(0.999)
GDP change (year 3)	0.605**
	(0.220)
Num.Obs.	48
R2	0.142
R2 Adj.	0.123

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Predicting income



A linear regression model predicting income

	Model 1
(Intercept)	56093.305***
	(1705.115)
Proportion of women	-30669.943***
	(2987.010)
Num.Obs.	172
R2	0.383
R2 Adj.	0.379

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Why more covariates?