Introduction to Regression

Jacob M. Montgomery

2018

Regression models

Syllabus Going Forward

- 10/25: Iterated Weighted Least Squares/Logit
- 10/30: GLM: A More Formal Statement
- 11/1: MLE for counts (Lim)/ Ordered (Dom)
- 11/6: Choices (Ryden) / Spatial Models (Daniel)/PROJECT TOPICS DUE
- 11/8: Easy Bayesian (Stan)
- 11/13: Bayesian Regression/Ridge
- 11/15: Multi-level models
- 11/20: Mixture models: Bayes/EM (Joshua)
- 11/27: Variable Selection: Lasso (Jaerin) / BMA
- 11/29: Non-parametric regression: Smoothing (?) / GP
- 12/4: Random Forests /GBM (Amanda) / BART
- 12/6: Selection (Zoe) / Survival Models (Ben)

Bootstrapping the curve

Where are we?

- So far we have focused on how to estimate some parameter (or vector of parameters) θ that relate to a random variable x or random variables $\mathbf{x}$
- In the second half of the class, we want to turn to the task of asking how some variable (x) or set of variables ($\mathbf{x}$) are related to some response variable y.

The general problem

In the broadest possible terms, we want to estimate a regression function $r(x)$ which is:

$$r(x) = E(Y|X = x) = \int yf(y|x)dy$$

From our most agnostic viewpoint we want to take all of our data and assume it comes from a joint distribution

$$(Y_1, X_1), \ldots, (Y_n, X_n) \sim F_{X,Y}$$

From our most agnostic viewpoint we want to take all of our data and assume it comes from a joint distribution

$$(Y_1, X_1), \ldots, (Y_n, X_n) \sim F_{X,Y}$$

As before, the easiest way to do this is by making parametric assumptions.

From our most agnostic viewpoint we want to take all of our data and assume it comes from a joint distribution

$$(Y_1, X_1), \ldots, (Y_n, X_n) \sim F_{X,Y}$$

As before, the easiest way to do this is by making parametric assumptions. Non-parametric approaches will be discussed later in the class.

Simple linear regression model

- The simplest version is to assume

$$r(x) = \beta_0 + \beta_1 x$$

- Making the further assumption that $V(\epsilon|X = x) = \sigma^2$ does not depend on x, we can get the usual linear regression model

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

where $E(\epsilon_i|X_i) = 0$ and $V(\epsilon_i|X_i) = \sigma^2$

- The unknown parameters in the model are β_0, β_1 and the variance term σ^2.
- The fitted values $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$ and the residuals are $\hat{\epsilon}_i = Y_i - \hat{Y}_i$
- The residual sum of squares (RSS) is then

$$\sum_{i=1}^{n} \hat{\epsilon}_i^2$$

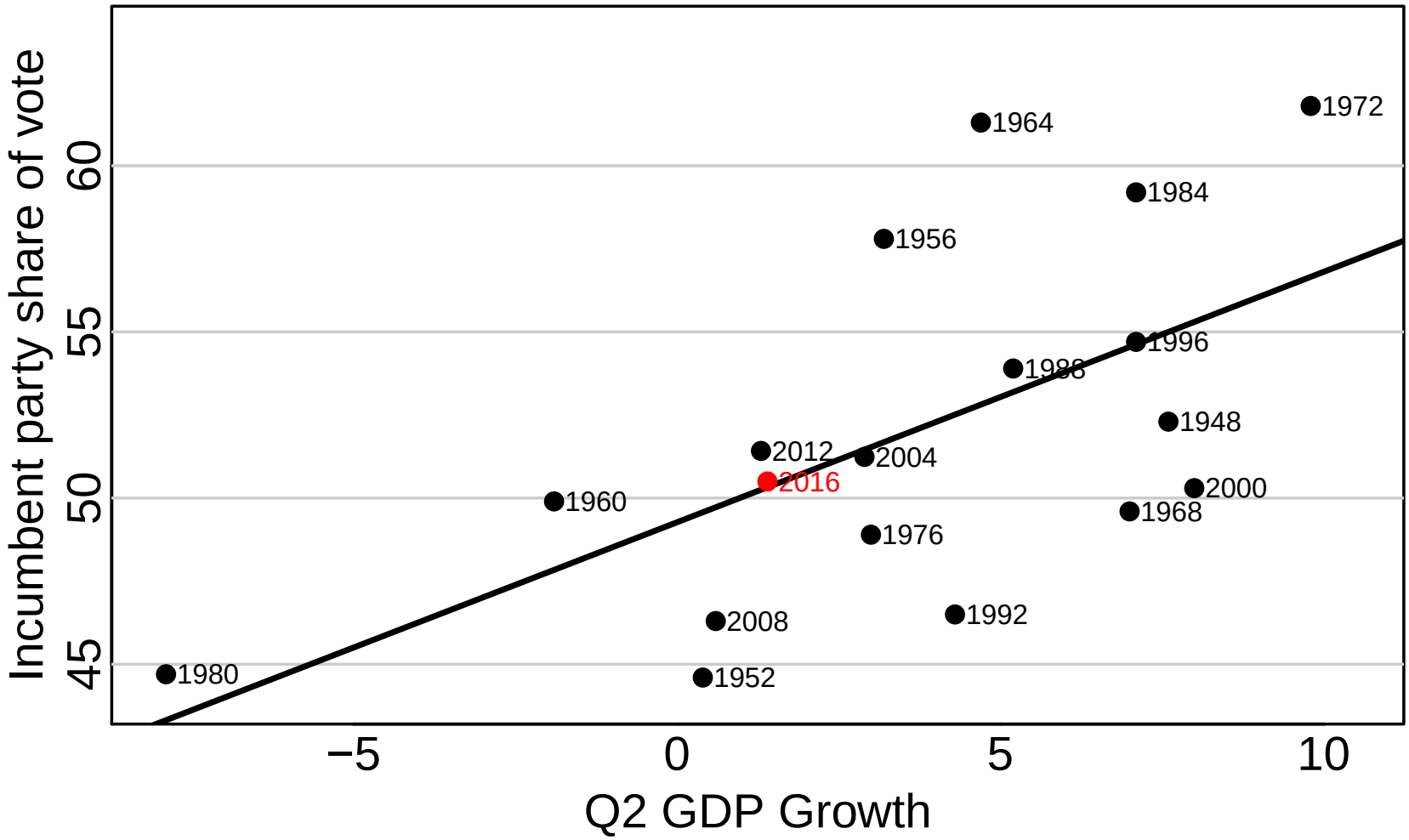
- The **least squares estimates** are the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes the RSS.

Regression: The big picture

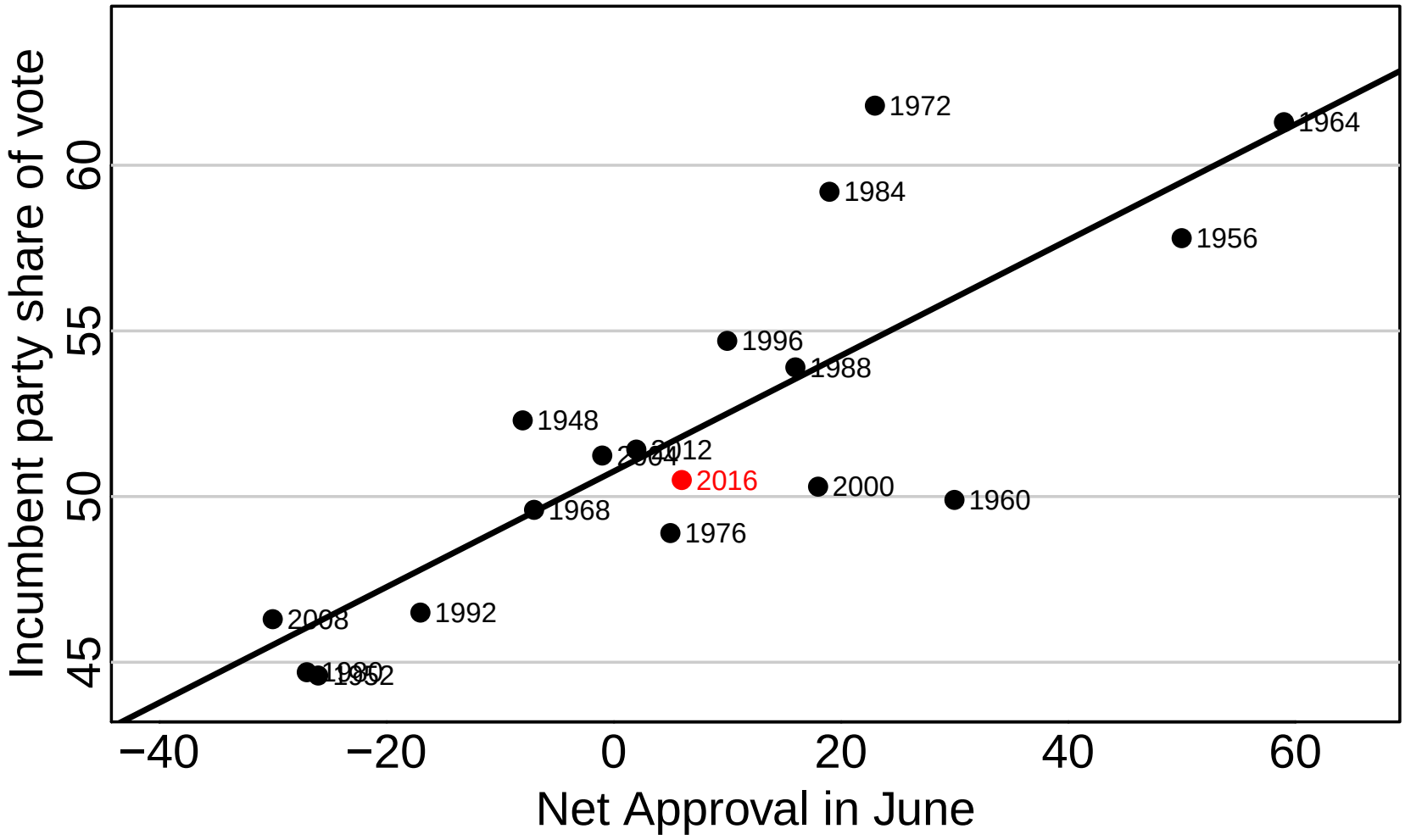
What we want to do is the following:

- Assume we have two variables where the "outcome" is interval(ish)
- Is there an "association"" between them?
- Is it statistically significant?
- Estimate "expected values" for an outcome variable given a set of covariates

Example: Presidential election and GDP growth from 1952-2016

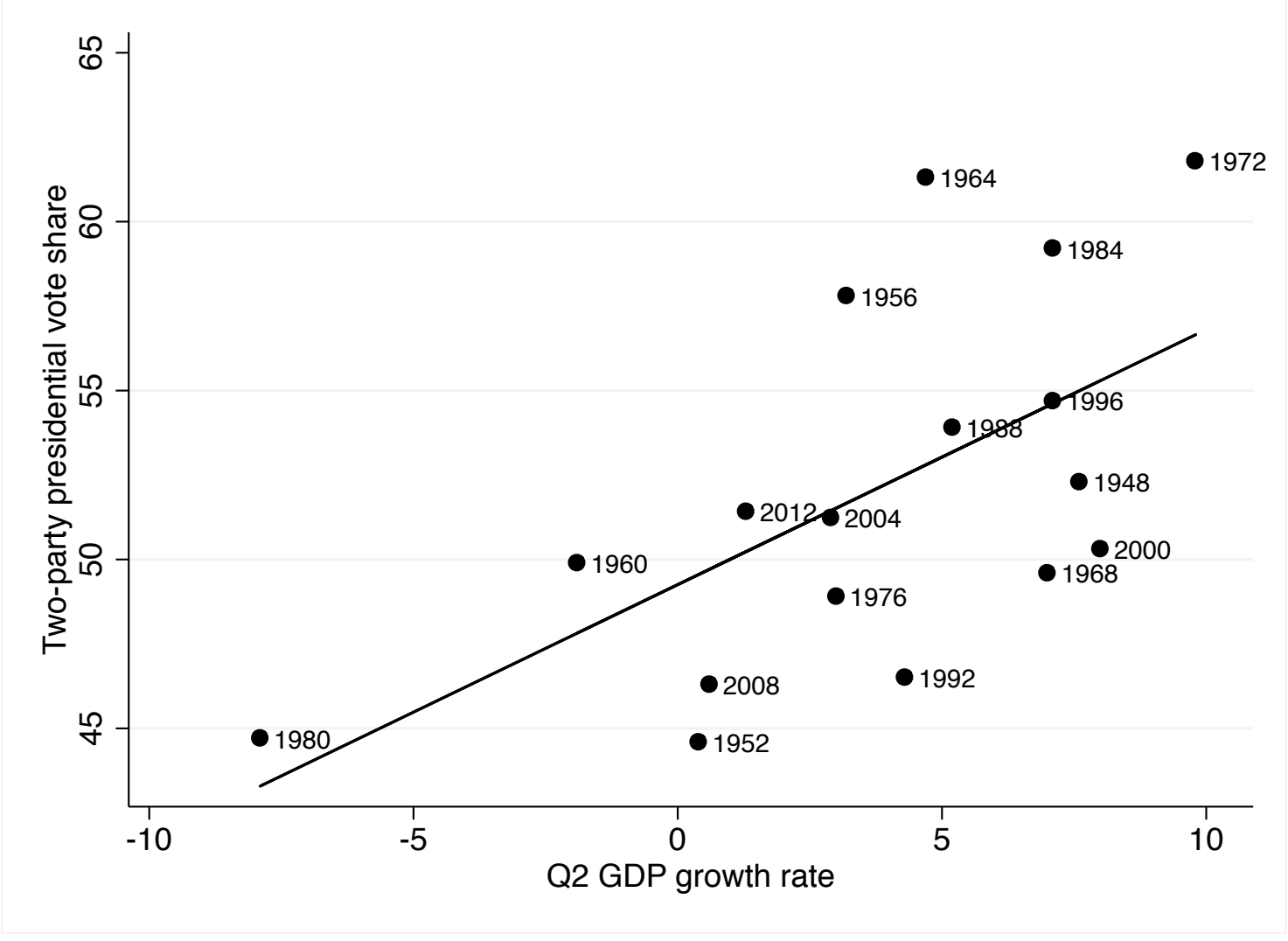


Example: Presidential election and GDP growth from 1952-2016



Our best guess for the best line leads to inference and prediction

Incumbent party vote $= 49.3 + 0.75$ Q2 GDP



Making this more formal

Let our data be the dyads (Y_i, X_i), $i = 1, \ldots, n$.

Making this more formal

Let our data be the dyads (Y_i, X_i), $i = 1, \ldots, n$.

We are going to assume that there is a linear relationship between the variables:

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

Making this more formal

Let our data be the dyads (Y_i, X_i), $i = 1, \ldots, n$.

We are going to assume that there is a linear relationship between the variables:

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

However, we also know that there is error, so

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

$$\epsilon_i \sim N(0, \sigma^2)$$

Making this more formal

Let our data be the dyads (Y_i, X_i), $i = 1, \ldots, n$.

We are going to assume that there is a linear relationship between the variables:

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

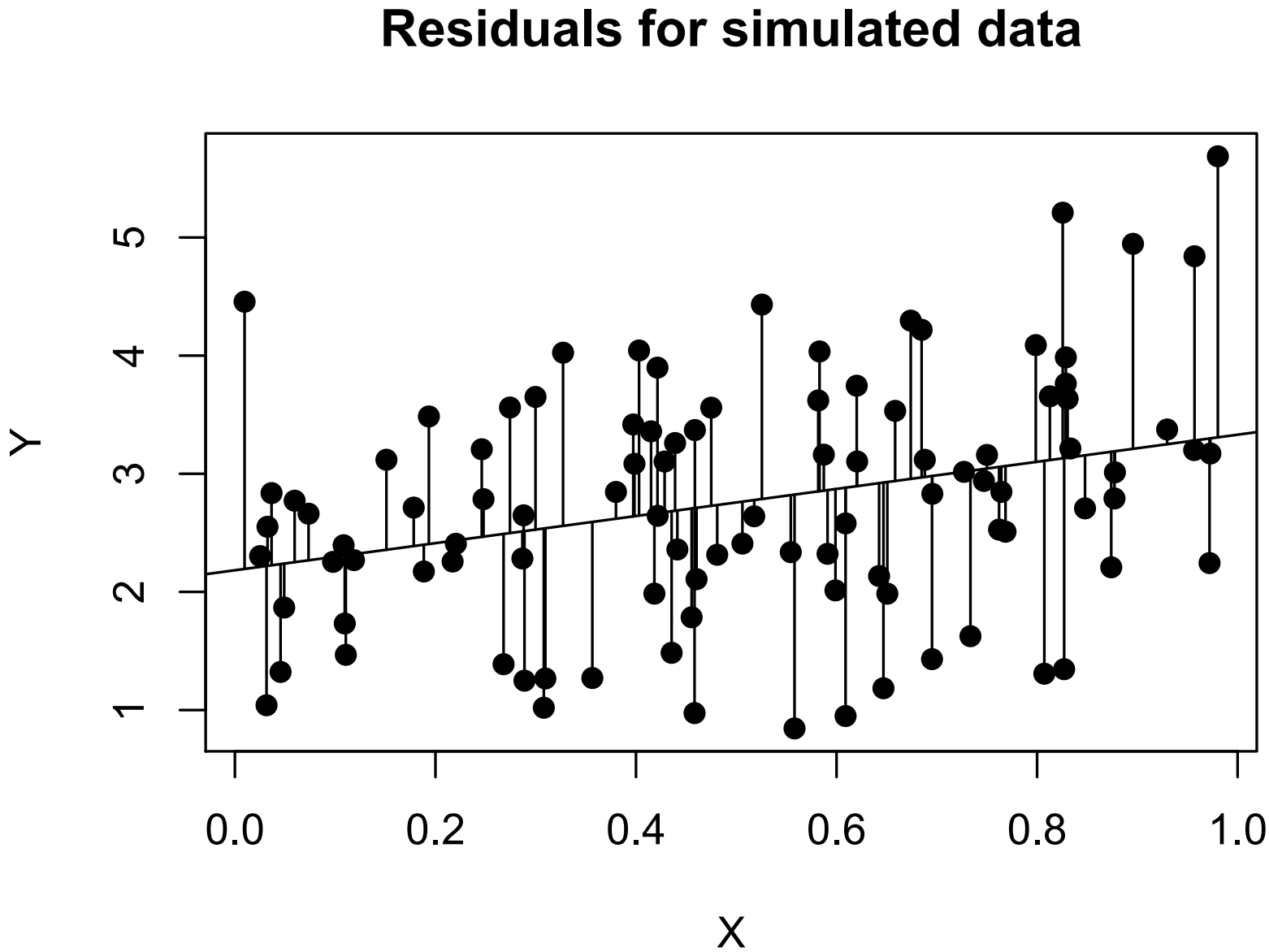
However, we also know that there is error, so

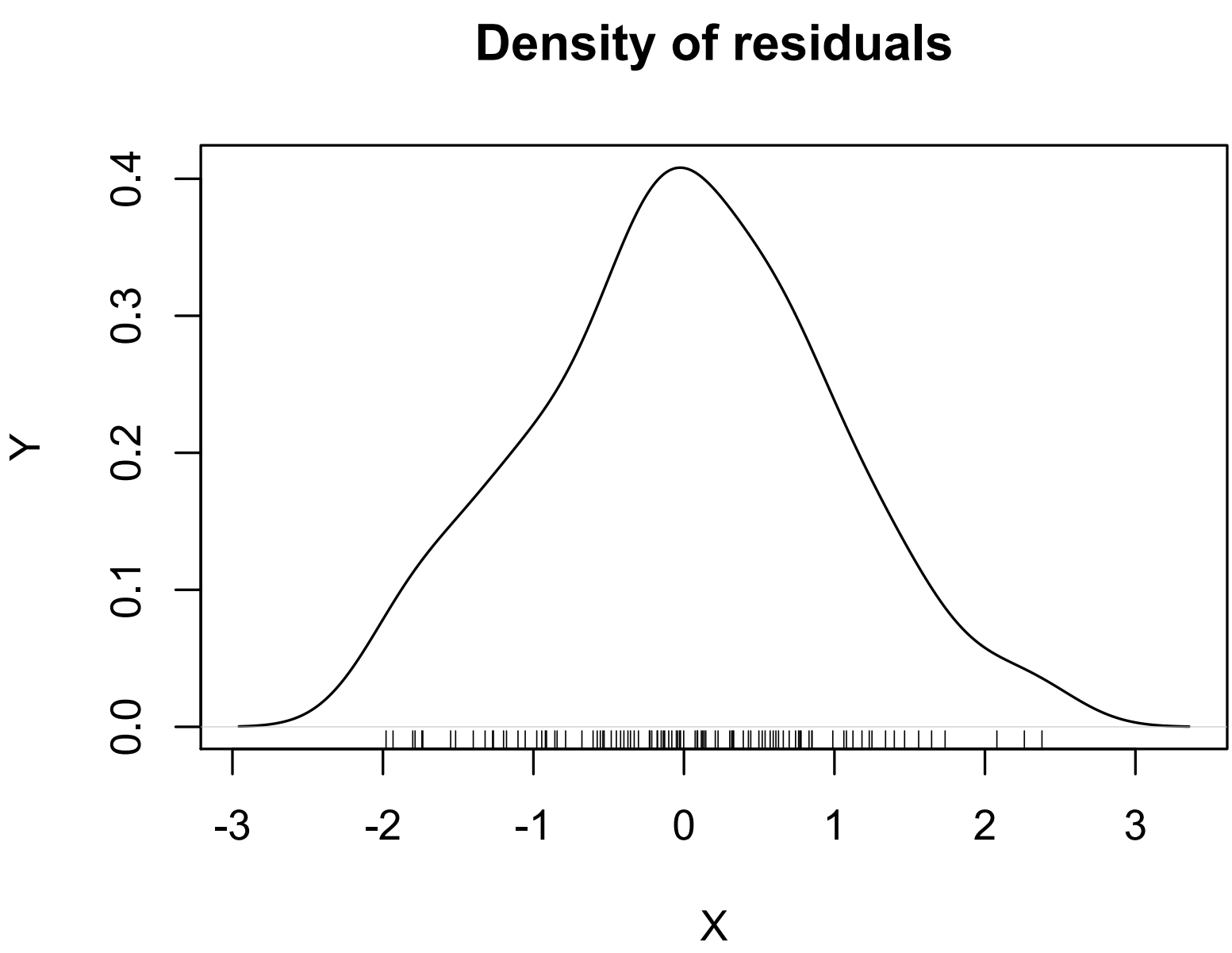
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

This is equivalent to writing:

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

Visualizing perfectly normal errors




Density of residuals
Y
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
X

- We have reduced all of the data to a simplified model
- We have three parameter $(\beta_0, \beta_1, \sigma^2)$ that we now need to estimate using our data.
- Once we have our parameter estimates, we want to then make inferences.
    - Just like before, we will set up hypotheses
    - Just like before, we will summarize how well the data supports these hypotheses

- We have reduced all of the data to a simplified model
- We have three parameter $(\beta_0, \beta_1, \sigma^2)$ that we now need to estimate using our data.
- Once we have our parameter estimates, we want to then make inferences.
  - Just like before, we will set up hypotheses
  - Just like before, we will summarize how well the data supports these hypotheses

But before we can do anything else, we need to make estimates:

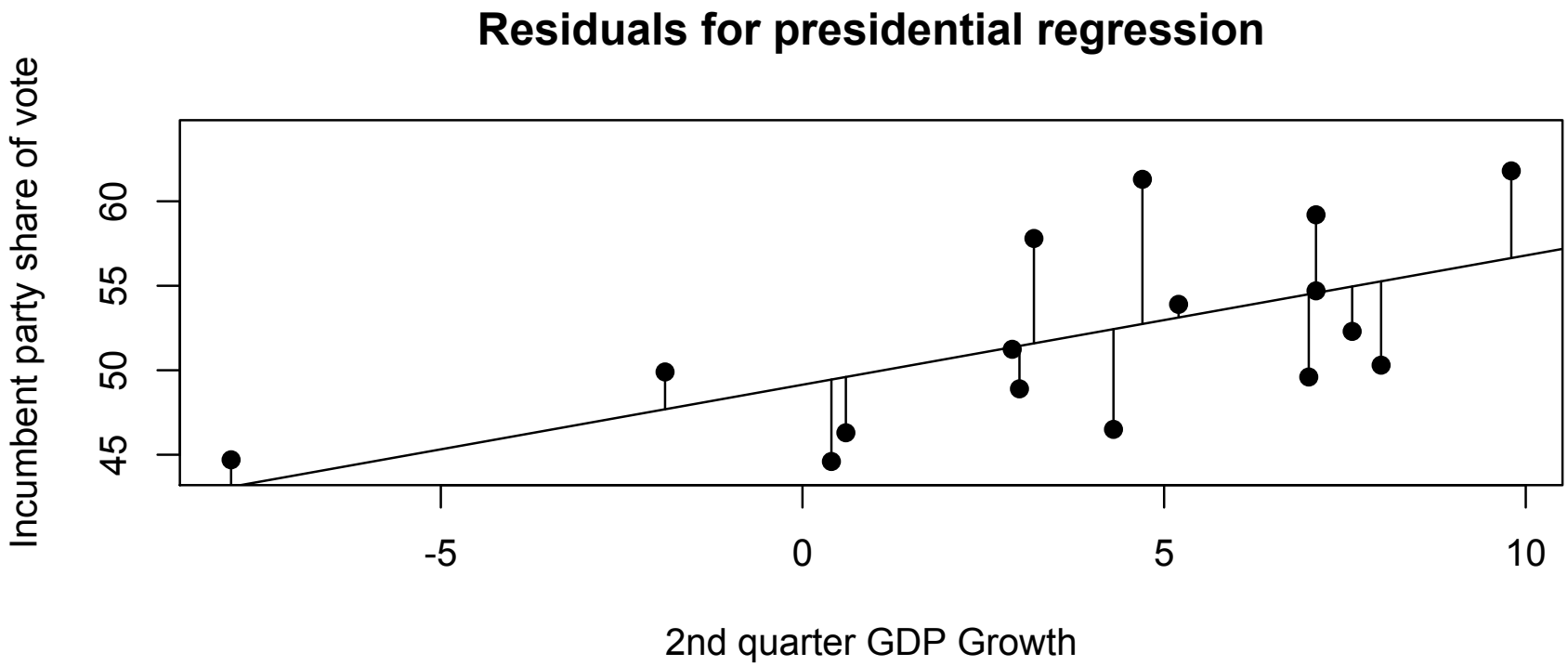
$$\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$$

- We are going to choose parameters that minimize error
- We define the observed residual for observation i as e_i.
- This is just the difference between our estimate for the value of Y_i given X_i and what was actually observed.

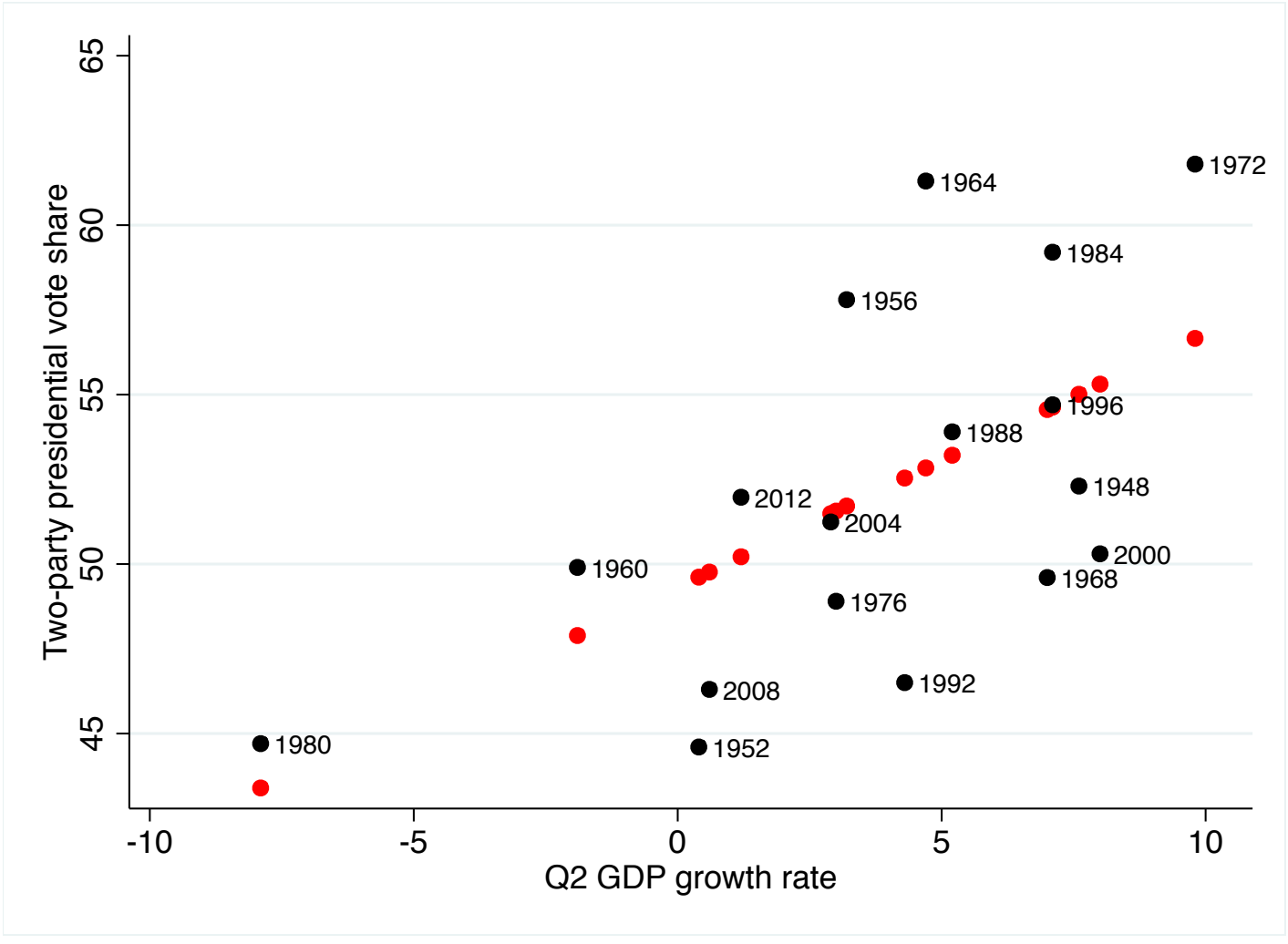
$$e_i = (Y_i - \hat{Y}_i) = (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)$$

- We are going to choose parameters that minimize error
- We define the observed residual for observation i as e_i.
- This is just the difference between our estimate for the value of Y_i given X_i and what was actually observed.

$$e_i = (Y_i - \hat{Y}_i) = (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)$$



Another look at residuals



Estimators for β_0 and β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \Big((X_i - \bar{X})(Y_i - \bar{Y}) \Big)}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Estimators for β_0 and β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \Big((X_i - \bar{X})(Y_i - \bar{Y}) \Big)}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Estimators for β_0 and β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \Big((X_i - \bar{X})(Y_i - \bar{Y}) \Big)}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Both of these are functions of the data.

Estimators for β_0 and β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \Big((X_i - \bar{X})(Y_i - \bar{Y}) \Big)}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Both of these are functions of the data. For our presidential election data:

1. $\hat{\beta}_0 = 49.3$
2. $\hat{\beta}_1 = 0.75$

Estimators for β_0 and β_1

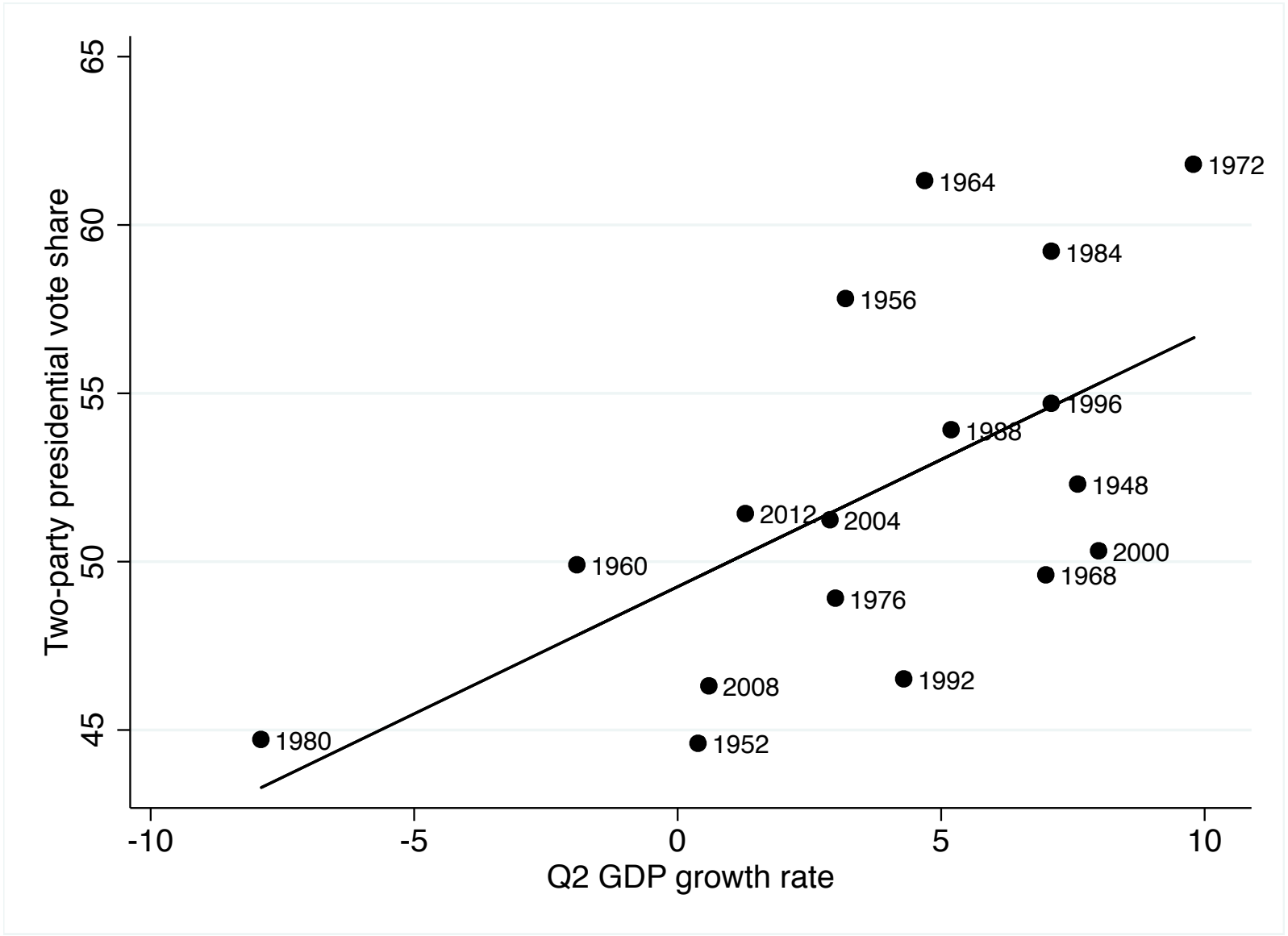
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \Big((X_i - \bar{X})(Y_i - \bar{Y}) \Big)}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Both of these are functions of the data. For our presidential election data:

1. $\hat{\beta}_0 = 49.3$
2. $\hat{\beta}_1 = 0.75$

What does that mean?


Two-party presidential vote share
65
60
55
50
45
-10
-5
0
5
10
Q2 GDP growth rate
1972
1964
1984
1956
1996
1988
1948
2012
2004
2000
1960
1968
1976
1992
2008
1980
1952

Y_i	X_i
3.8	3.5
3.0	3.3
3.5	4.0
2.8	2.3
2.4	1.8
2.7	2.7

Find $\hat{\beta}_0$ and $\hat{\beta}_1$.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \Big((X_i - \bar{X})(Y_i - \bar{Y}) \Big)}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Y_i	X_i	$(Y_i - \bar{Y})$	$(X_i - \bar{X})$	$(Y_i - \bar{Y})(X_i - \bar{X})$
3.8	3.5	0.767	0.567	0.434889
3.0	3.3	-0.033	0.367	-0.012111
3.5	4.0	0.467	1.067	0.498289
2.8	2.3	-0.233	-0.633	0.147489
2.4	1.8	-0.633	-1.133	0.717189
2.7	2.7	-0.333	-0.233	0.077589
				$\sum(Y_i - \bar{Y})(X_i - \bar{X})$

$\sum Y_i = 18.2 \quad \sum X_i = 17.6$ $= 1.863$

$\bar{Y} = 3.033 \quad \bar{X} = 2.933$

Y_i	X_i	$(Y_i - \bar{Y})$	$(X_i - \bar{X})$	$(Y_i - \bar{Y})(X_i - \bar{X})$
3.8	3.5	0.767	0.567	0.434889
3.0	3.3	-0.033	0.367	-0.012111
3.5	4.0	0.467	1.067	0.498289
2.8	2.3	-0.233	-0.633	0.147489
2.4	1.8	-0.633	-1.133	0.717189
2.7	2.7	-0.333	-0.233	0.077589
				$\sum(Y_i - \bar{Y})(X_i - \bar{X})$

$\sum Y_i = 18.2 \quad \sum X_i = 17.6$ $= 1.863$

$\bar{Y} = 3.033 \quad \bar{X} = 2.933$

$$\hat{\beta}_1 = \frac{1.863}{3.33} = .559$$

Y_i	X_i	$(Y_i - \bar{Y})$	$(X_i - \bar{X})$	$(Y_i - \bar{Y})(X_i - \bar{X})$
3.8	3.5	0.767	0.567	0.434889
3.0	3.3	-0.033	0.367	-0.012111
3.5	4.0	0.467	1.067	0.498289
2.8	2.3	-0.233	-0.633	0.147489
2.4	1.8	-0.633	-1.133	0.717189
2.7	2.7	-0.333	-0.233	0.077589
				$\sum(Y_i - \bar{Y})(X_i - \bar{X})$

$\sum Y_i = 18.2 \quad \sum X_i = 17.6$ $= 1.863$

$\bar{Y} = 3.033 \quad \bar{X} = 2.933$

$$\hat{\beta}_1 = \frac{1.863}{3.33} = .559$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Y_i	X_i	$(Y_i - \bar{Y})$	$(X_i - \bar{X})$	$(Y_i - \bar{Y})(X_i - \bar{X})$
3.8	3.5	0.767	0.567	0.434889
3.0	3.3	-0.033	0.367	-0.012111
3.5	4.0	0.467	1.067	0.498289
2.8	2.3	-0.233	-0.633	0.147489
2.4	1.8	-0.633	-1.133	0.717189
2.7	2.7	-0.333	-0.233	0.077589
				$\sum(Y_i - \bar{Y})(X_i - \bar{X})$

$\sum Y_i = 18.2 \quad \sum X_i = 17.6$ $= 1.863$

$\bar{Y} = 3.033 \quad \bar{X} = 2.933$

$$\hat{\beta}_1 = \frac{1.863}{3.33} = .559$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 3.033 - .587(2.933) = 1.394$$

Now the variance

The unbiased estimate is:

$$\hat{\sigma}^2 = \left(\frac{1}{n-2}\right) \sum_{i=1}^{n} \hat{\epsilon}^2$$

MLE estimation of a regression model

1. Show that maximizing the conditional log-likelihood is the same as the same as minimizing the RSS.
2. Find the MLE for σ^2.

Thinking through this even more

- Given the following equation:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \Big((X_i - \bar{X})(Y_i - \bar{Y}) \Big)}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

- And mean centering $\bar{y} = 0$
- We get that $b = \frac{\sum_{i=1}^{n} x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} y_i$
- Remembering that x here is constant, find the variance of b.

Thinking through this even more

- Given the following equation:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \Big((X_i - \bar{X})(Y_i - \bar{Y}) \Big)}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

- And mean centering $\bar{y} = 0$
- We get that $b = \frac{\sum_{i=1}^{n} x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} y_i$
- Remembering that x here is constant, find the variance of b.
- What is the standard error?
- Is it consistent?

Thinking through this even more

- Given the following equation:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \Big((X_i - \bar{X})(Y_i - \bar{Y}) \Big)}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

- And mean centering $\bar{y} = 0$
- We get that $b = \frac{\sum_{i=1}^{n} x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} y_i$
- Remembering that x here is constant, find the variance of b.
- What is the standard error?
- Is it consistent?
- Where did we invoke the normality assumption? And what does it mean?

$$se(\beta_1) = \frac{\hat{\sigma}}{\sqrt{(x_i - \bar{x})^2}}$$

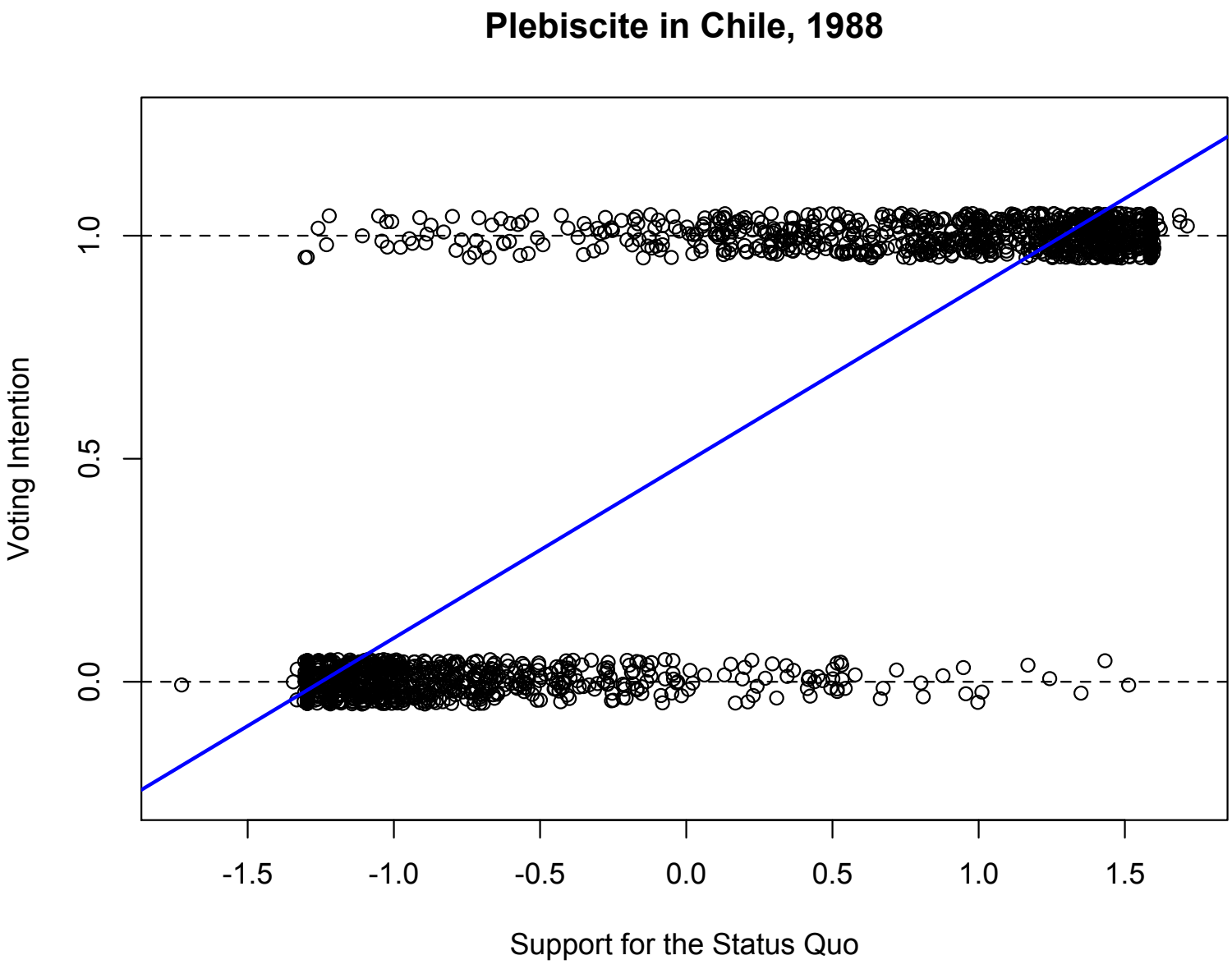
Binary outcomes

- This is perhaps one of the most common modeling set-ups in many areas of research. Supreme court votes, war/not war, turnout, etc.

Binary outcomes

- This is perhaps one of the most common modeling set-ups in many areas of research. Supreme court votes, war/not war, turnout, etc.
- We will fit a logistic regression, sometimes called "logit." Another option is called "probit" that is very similar in interpretation but estimated very differently.

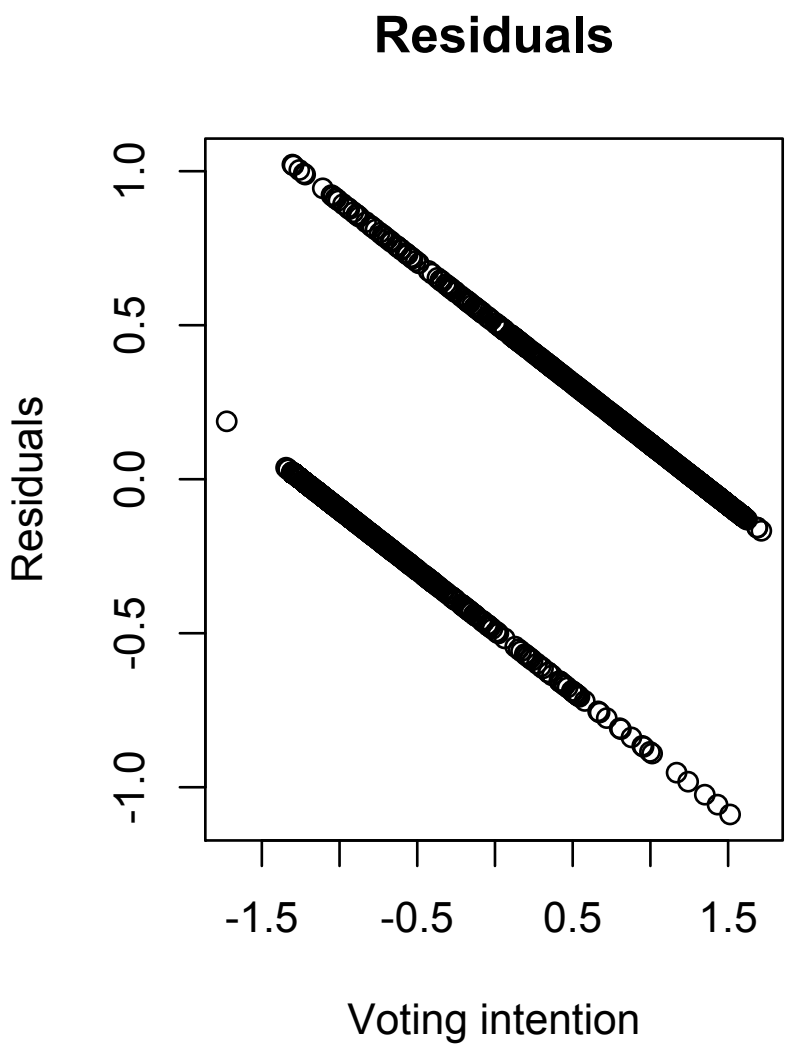
Linear probability model

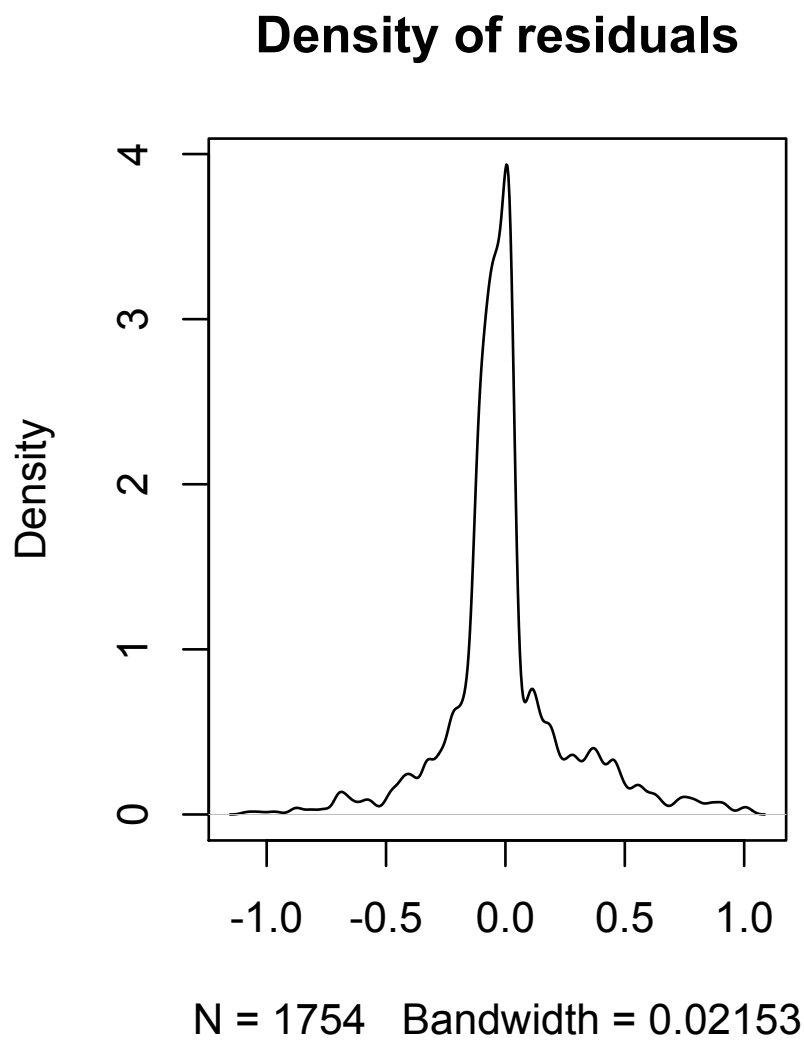


Problems with linear probability models

- Violates regression assumptions. Residuals are heteroscedastic and non-normal.
- It can also lead to nonsensical predictions.
- Interpret coefficients on the "predicted probability scale."

Residuals of linear probability model





The logit model

- Instead of fitting a line through the data, we are going to fit a curve.
- The curve will be constructed so the maximum value is 1 and the minimum value is 0.
- It is a function that maps the real numbers into probabilities.
- Although the outcome is transformed, it is still *linear*
- $\mu_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i}$

- Let $\Lambda(.)$ denote the cumulative distribution of the logistic distribution.
- We now assume:

$$Pr(Y_i = 1) = \Lambda(\mu_i) = \frac{exp(\mu_i)}{1 + exp(\mu_i)}$$

- Let $\Lambda(.)$ denote the cumulative distribution of the logistic distribution.
- We now assume:

$$Pr(Y_i = 1) = \Lambda(\mu_i) = \frac{exp(\mu_i)}{1 + exp(\mu_i)}$$

- Recall that we have already specified:

$$\mu_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i}$$

- Note that we are no longer directly modeling Y_i, we are modeling the underlying probability.

- Let $\Lambda(.)$ denote the cumulative distribution of the logistic distribution.
- We now assume:

$$Pr(Y_i = 1) = \Lambda(\mu_i) = \frac{exp(\mu_i)}{1 + exp(\mu_i)}$$

- Recall that we have already specified:

$$\mu_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i}$$

- Note that we are no longer directly modeling Y_i, we are modeling the underlying probability.
- What happens when μ_i gets big?

- Let $\Lambda(.)$ denote the cumulative distribution of the logistic distribution.
- We now assume:

$$Pr(Y_i = 1) = \Lambda(\mu_i) = \frac{exp(\mu_i)}{1 + exp(\mu_i)}$$

- Recall that we have already specified:

$$\mu_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i}$$

- Note that we are no longer directly modeling Y_i, we are modeling the underlying probability.
- What happens when μ_i gets big? Gets small?

- Interpretation of this very difficult, especially if we include many variables.
- A one unit increase in X yields differing changes in the probability of $(Y = 1)$ *depending on the value of* X.

Voting turnout	
Intercept	0.2153
	(0.0996)
Status Quo	3.2055
	(0.1431)
n=1754	

	Voting turnout
Intercept	0.2153
	(0.0996)
Status Quo	3.2055
	(0.1431)
n=1754	

- $Pr(Y_i = 1|X_i = -1.00) =$

Voting turnout	
Intercept	0.2153
	(0.0996)
Status Quo	3.2055
	(0.1431)
n=1754	

- $Pr(Y_i = 1|X_i = -1.00) = \frac{exp(.2153+(-1)*(3.2055))}{1+exp(.2153+(-1)*(3.2055))}$

Voting turnout	
Intercept	0.2153
	(0.0996)
Status Quo	3.2055
	(0.1431)
n=1754	

- $Pr(Y_i = 1|X_i = -1.00) = \frac{exp(.2153+(-1)*(3.2055))}{1+exp(.2153+(-1)*(3.2055))} = 0.0479$

Voting turnout	
Intercept	0.2153
	(0.0996)
Status Quo	3.2055
	(0.1431)
n=1754	

- $Pr(Y_i = 1|X_i = -1.00) = \frac{exp(.2153+(-1)*(3.2055))}{1+exp(.2153+(-1)*(3.2055))} = 0.0479$
- $Pr(Y_i = 1|X_i = -0.75) =$

Voting turnout	
Intercept	0.2153
	(0.0996)
Status Quo	3.2055
	(0.1431)
n=1754	

- $Pr(Y_i = 1|X_i = -1.00) = \frac{exp(.2153+(-1)*(3.2055))}{1+exp(.2153+(-1)*(3.2055))} = 0.0479$
- $Pr(Y_i = 1|X_i = -0.75) = \frac{exp(.2153+(-0.75)*(3.2055))}{1+exp(.2153+(-0.75)*(3.2055))} =$

Voting turnout	
Intercept	0.2153
	(0.0996)
Status Quo	3.2055
	(0.1431)
n=1754	

- $Pr(Y_i = 1|X_i = -1.00) = \frac{exp(.2153+(-1)*(3.2055))}{1+exp(.2153+(-1)*(3.2055))} = 0.0479$
- $Pr(Y_i = 1|X_i = -0.75) = \frac{exp(.2153+(-0.75)*(3.2055))}{1+exp(.2153+(-0.75)*(3.2055))} = 0.1008$

Voting turnout

Intercept	0.2153
	(0.0996)
Status Quo	3.2055
	(0.1431)
n=1754	

- $Pr(Y_i = 1|X_i = -1.00) = \frac{exp(.2153+(-1)*(3.2055))}{1+exp(.2153+(-1)*(3.2055))} = 0.0479$
- $Pr(Y_i = 1|X_i = -0.75) = \frac{exp(.2153+(-0.75)*(3.2055))}{1+exp(.2153+(-0.75)*(3.2055))} = 0.1008$
- $Pr(Y_i = 1|X_i = -0.25)$, $Pr(Y_i = 1|X_i = 0.00)$?

Voting turnout

Intercept	0.2153
	(0.0996)
Status Quo	3.2055
	(0.1431)
n=1754	

- $Pr(Y_i = 1|X_i = -1.00) = \frac{exp(.2153+(-1)*(3.2055))}{1+exp(.2153+(-1)*(3.2055))} = 0.0479$
- $Pr(Y_i = 1|X_i = -0.75) = \frac{exp(.2153+(-0.75)*(3.2055))}{1+exp(.2153+(-0.75)*(3.2055))} = 0.1008$
- $Pr(Y_i = 1|X_i = -0.25)$, $Pr(Y_i = 1|X_i = 0.00)$?
- 0.3675, 0.5536
- Why are the changes so different?

A study of graduate admissions produced the following prediction equation:

$$P(Y = 1) = \Lambda(-3.450 + 0.002x_1 + 0.780x_2 - 0.560x_3),$$

where $\Lambda(.)$ is the cumulative distribution of the (standard) logistic distribution.

- Y=1 if the student was admitted to graduate school
- x_1 is the student's GRE score
- x_2 is the student's G.P.A
- x_3 is a measure of the quality of the student's undergraduate school (the variable takes values 1, 2, 3, or 4)

Calculate the difference in the estimated probabilities of admission for a student with a G.P.A. of 3.2 and undergraduate ranking of 4 with a GRE score of 700 and the same student with a GRE score of 800.

Moving forward

- Next class: Estimating a logit via iterated weighted least squares. Practicum on estimation and interpretation.
- Class after: A more formal statement on GLM models