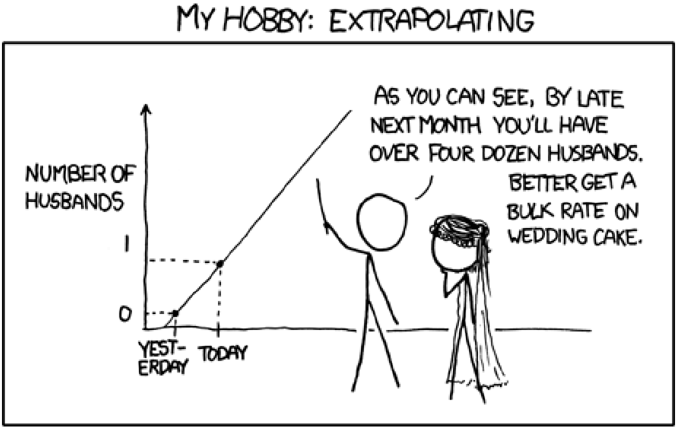
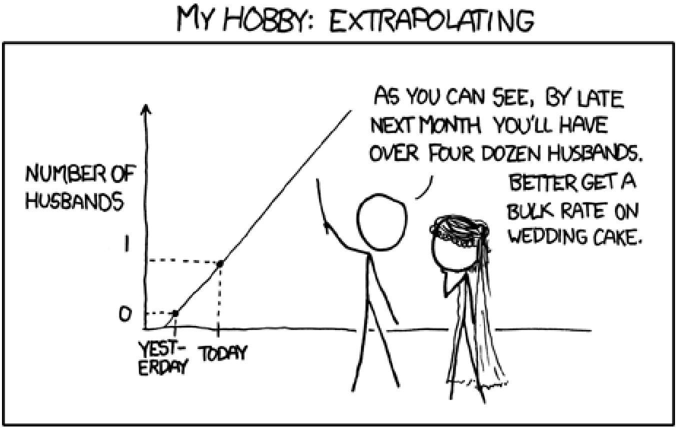
Correlation and bivariate linear regression

Prof. Jacob M. Montgomery

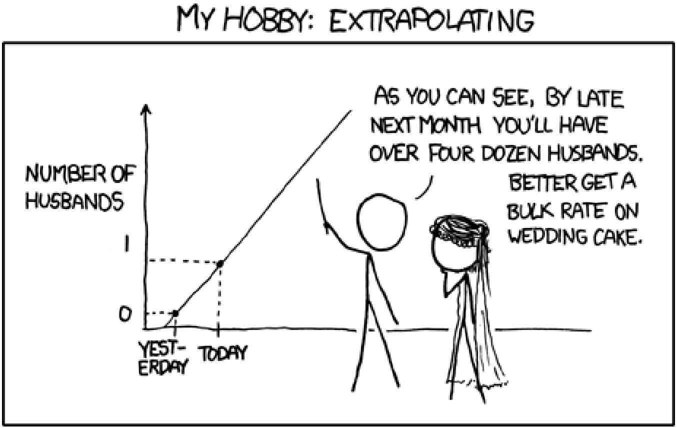
Quantitative Political Methodology (L32 363)

October 30, 2017

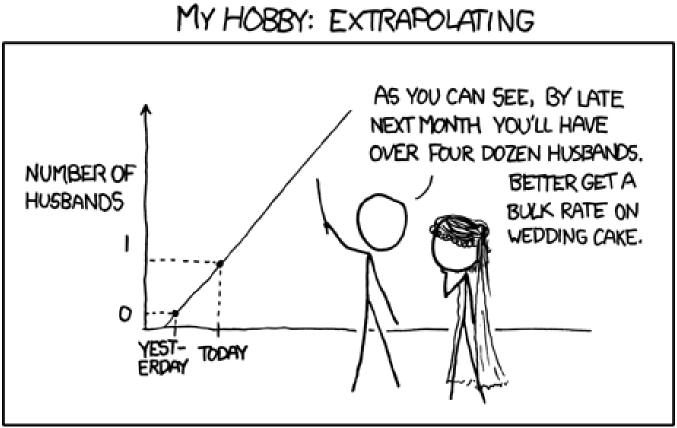

MY HOBBY: EXTRAPOLATING
NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS. BETTER GET A BULK RATE ON WEDDING CAKE.



- Scatterplots



- Scatterplots
- Correlation



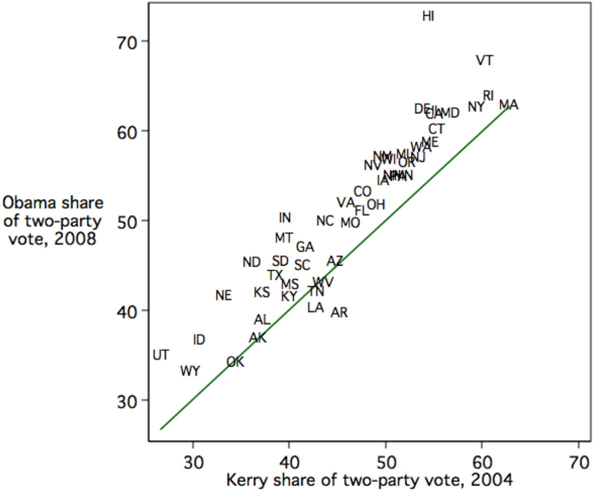
- Scatterplots
- Correlation
- Drawing the "best" line through data

Scatterplots

What are we looking for?

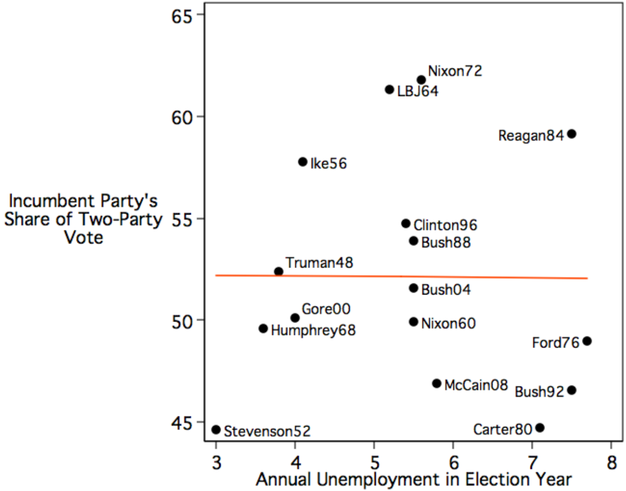
- Form/pattern
- Direction
- Strength
- Outliers

Scatterplots



(Masket 2008)

Scatterplots



(Masket 2011)

How do we quantify this?

Correlation! But first...

Standardizing variables

$$\frac{x - \bar{x}}{s}$$

Example: Populations of New England states

	x	$\frac{x-\bar{x}}{s}$
CT	3.5m	0.48
ME	1.3m	-0.47
MA	6.6m	1.83
NH	1.3m	-0.47
RI	1.0m	-0.59
VT	0.6m	-0.78

$\bar{x} = 2.40 \qquad s = 2.29$

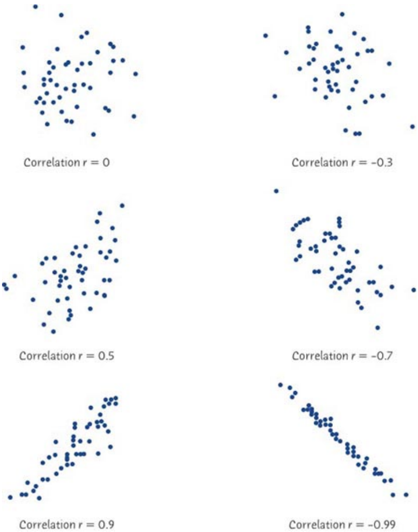
Correlation coefficient

Computation: Average of the products of the standardized values

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

What does a positive correlation mean? Negative?

Correlation visualized



Correlation $r = 0$

Correlation $r = -0.3$

Correlation $r = 0.5$

Correlation $r = -0.7$

Correlation $r = 0.9$

Correlation $r = -0.99$

Correlation visualized

http://guessthecorrelation.com/

Facts about correlation

- Linear only
- **Not** causal
- Unit-free
- $-1 \leq r \leq 1$
- Sensitive to outliers

Regression: The big picture

What we want to do is the following:

- Assume we have two variables where the "outcome" is interval(ish)

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?

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 (next class)?

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 (next class)?
- Estimate “expected values” for an outcome variable given a set of covariates

Some preliminaries

Y = Response variable/ Dependent variable/
Outcome variable/Explained variable/ Left-hand side

Some preliminaries

$$\begin{aligned}
Y = &\quad \text{Response variable/ Dependent variable/} \\
&\quad \text{Outcome variable/Explained variable/ Left-hand side} \\
X = &\quad \text{Explanatory variable/ Independent variable/} \\
&\quad \text{Treatment Variable/ Right-hand side}
\end{aligned}$$

How might Y and X be related?

Some preliminaries

Y = Response variable/ Dependent variable/
Outcome variable/Explained variable/ Left-hand side
X = Explanatory variable/ Independent variable/
Treatment Variable/ Right-hand side

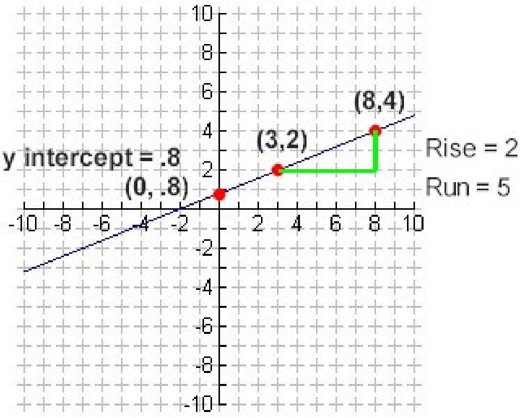
How might Y and X be related? A line of course!

Linear Model

$$Y = \alpha + \beta X$$

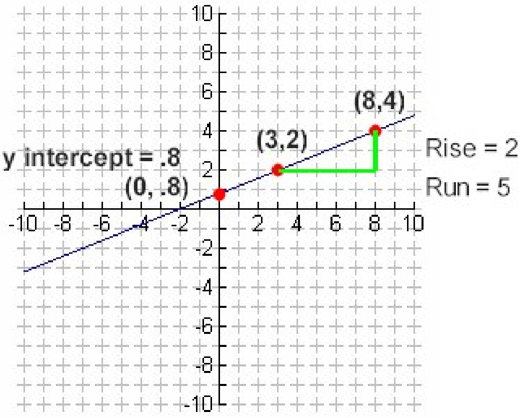
Here α is the Y-intercept and β is the slope of the line.

Reviewing the components of a line



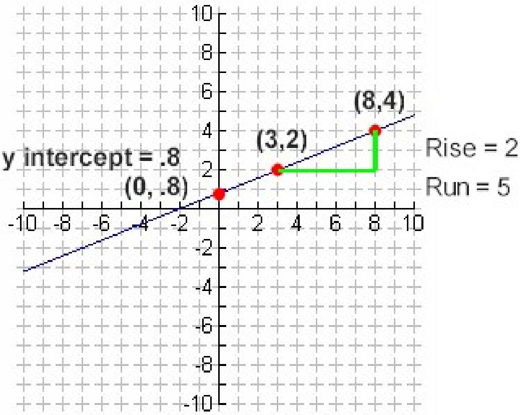
- $Y = 0.8 + \frac{2}{5}X$

Reviewing the components of a line



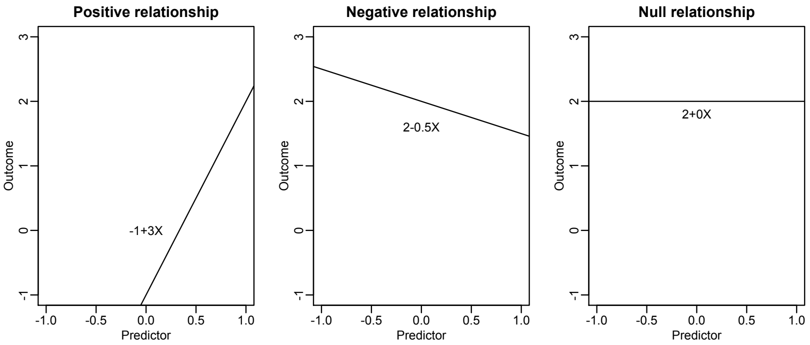
- $Y = 0.8 + \frac{2}{5}X$
- $\alpha = 0.8 =$ the value of Y when X is Zero.

Reviewing the components of a line



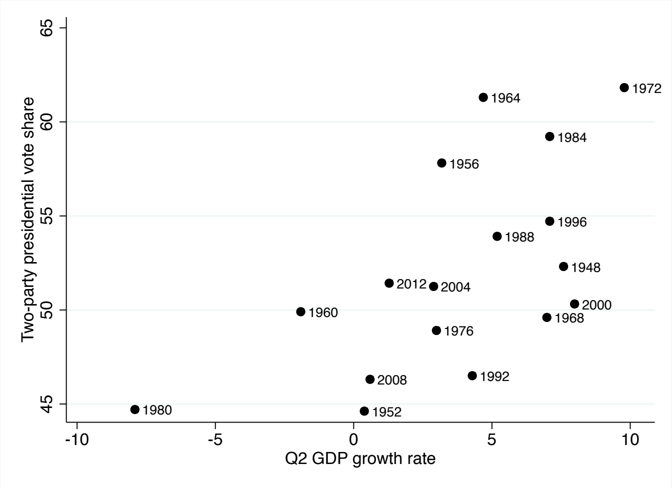
- $Y = 0.8 + \frac{2}{5}X$
- $\alpha = 0.8 =$ the value of Y when X is Zero.
- $\beta = 0.4 =$ the increase in Y associated with a one unit increase in X.

Interpretations



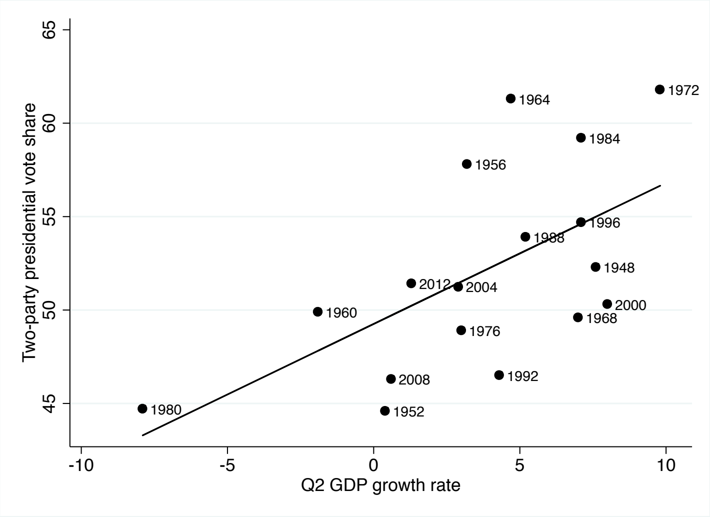
- α: The expected value of Y when X is zero
- $\beta > 0$: Positive relationship between X and Y
- $\beta < 0$: Negative relationship between X and Y
- $\beta = 0$: Null relationship between X and Y

Presidential elections and GDP growth from 1952-2012



Our best guess for the “best” line

Incumbent party vote = 49.3 + 0.75 Q2 GDP



Thinking formally about drawing the "best" line

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

Thinking formally about drawing the "best" line

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:

$$E(Y_i) = \alpha + \beta X_i$$

Thinking formally about drawing the "best" line

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:

$$E(Y_i) = \alpha + \beta X_i$$

However, we also know that there is error, so

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

Thinking formally about drawing the "best" line

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:

$$E(Y_i) = \alpha + \beta X_i$$

However, we also know that there is error, so

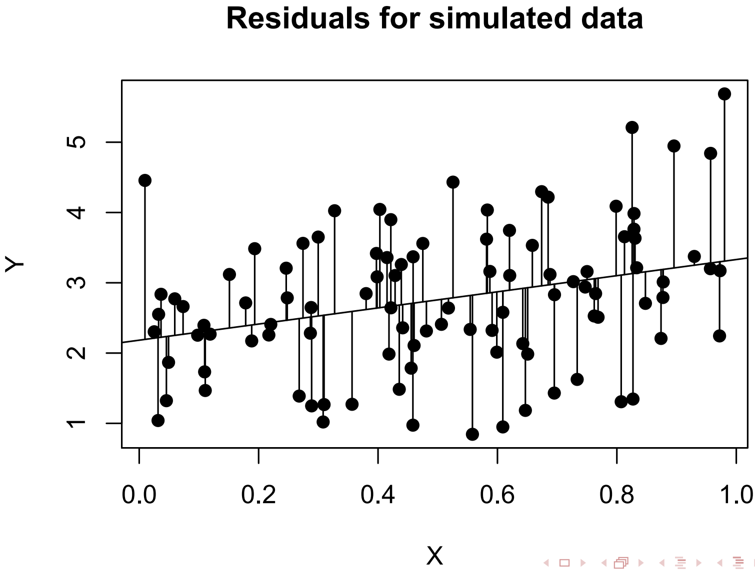
$$Y_i = \alpha + \beta X_i + \epsilon_i$$

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

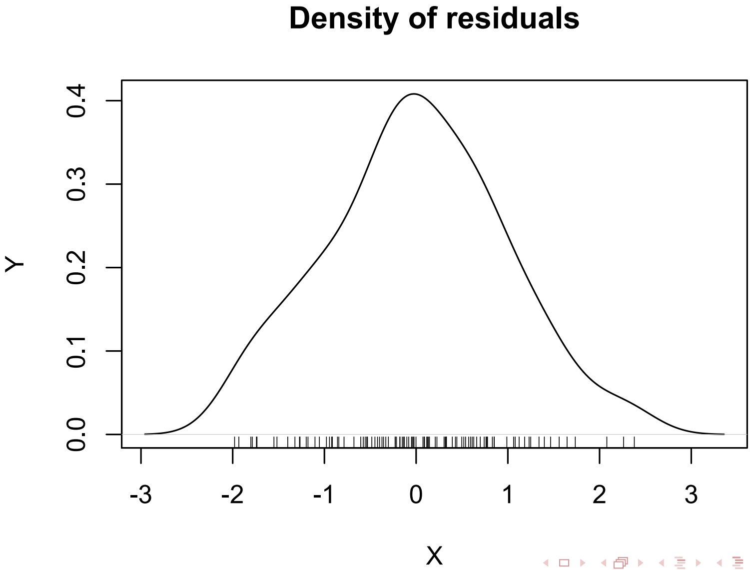
This is equivalent to writing:

$$Y_i \sim N(\alpha + \beta X_i, \sigma^2)$$

Visualizing perfectly normal errors



Visualizing perfectly normal errors



Implications

- We have reduced all of the data to a simplified model

Implications

- We have reduced all of the data to a simplified model
- We have three parameter $(\alpha, \beta, \sigma^2)$ that we now need to estimate using our data.

Implications

- We have reduced all of the data to a simplified model
- We have three parameter $(\alpha, \beta, \sigma^2)$ that we now need to estimate using our data.
- Once we have our parameter estimates, we want to then make inferences.

Implications

- We have reduced all of the data to a simplified model
- We have three parameter $(\alpha, \beta, \sigma^2)$ 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

Implications

- We have reduced all of the data to a simplified model
- We have three parameter $(\alpha, \beta, \sigma^2)$ 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

Implications

- We have reduced all of the data to a simplified model
- We have three parameter $(\alpha, \beta, \sigma^2)$ 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

Implications

- We have reduced all of the data to a simplified model
- We have three parameter $(\alpha, \beta, \sigma^2)$ 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:

$$\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2$$

Choose parameters that minimize error: Define error

Let's define the observed residual for observation i as e_i. This is just the difference between our "best guess" for the value of Y_i given X_i and what was actually observed.

Residuals

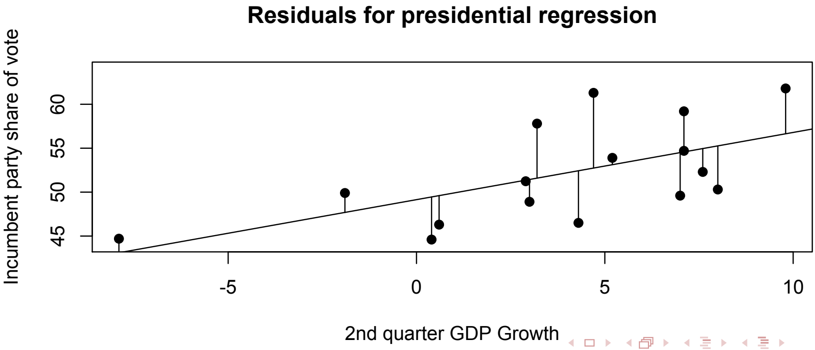
$$e_i = (Y_i - \hat{Y}_i) = (Y_i - \hat{\alpha} - \hat{\beta} X_i)$$

Choose parameters that minimize error: Define error

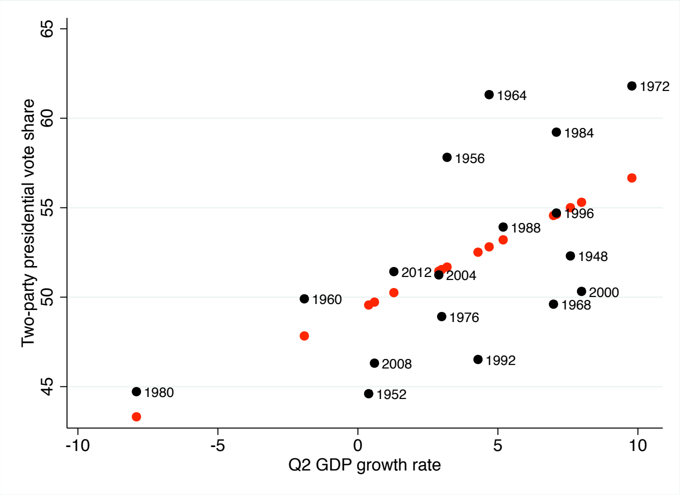
Let's define the observed residual for observation i as e_i. This is just the difference between our "best guess" for the value of Y_i given X_i and what was actually observed.

Residuals

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



Another look at residuals

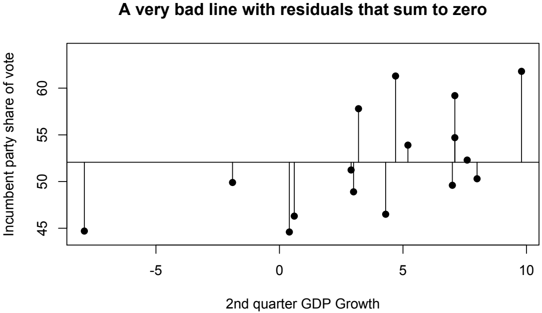


Back to drawing the "best" lines through data

The intuition here, is that the "best" line is the one that is going to reduce the amount of error.

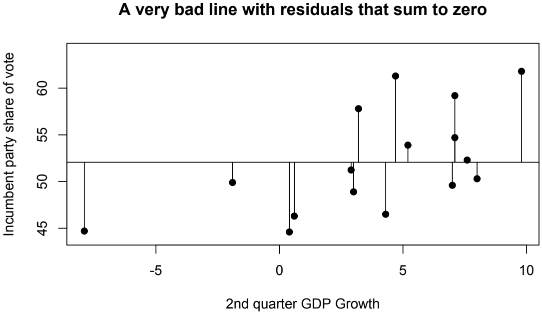
Back to drawing the "best" lines through data

The intuition here, is that the "best" line is the one that is going to reduce the amount of error. We might think to just look at the sum of the errors $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)$, but that is not actually a very good criterion.



Back to drawing the "best" lines through data

The intuition here, is that the "best" line is the one that is going to reduce the amount of error. We might think to just look at the sum of the errors $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)$, but that is not actually a very good criterion.



A long time ago, statisticians converged on the "Least Squares Criterion." We want to reduced the sum of `squared` error.

Defining "best" as minimizing SSE

For many good statistical reasons, we are going to say that any line that reduces the "Sum of Squared Error" is equivalent to having the "best" line. (Defined as most efficient unbiased estimator)

Defining "best" as minimizing SSE

For many good statistical reasons, we are going to say that any line that reduces the "Sum of Squared Error" is equivalent to having the "best" line. (Defined as most efficient unbiased estimator)

Sum of Squared Error

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Defining "best" as minimizing SSE

For many good statistical reasons, we are going to say that any line that reduces the "Sum of Squared Error" is equivalent to having the "best" line. (Defined as most efficient unbiased estimator)

Sum of Squared Error

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}X_i)^2$$

Defining “best” as minimizing SSE

For many good statistical reasons, we are going to say that any line that reduces the “Sum of Squared Error” is equivalent to having the “best” line. (Defined as most efficient unbiased estimator)

Sum of Squared Error

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}X_i)^2$$

We are going to minimize SSE with respect to $\hat{\alpha}$ and $\hat{\beta}$ (Calculus). With these parameters, we will be able to draw the “best” lines.

Estimators for α and β

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

Estimators for α and β

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

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

Estimators for α and β

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

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

Both of these are functions of the data.

Estimators for α and β

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

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

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

- $\hat{\alpha} = 49.3$
- $\hat{\beta} = 0.75$

Estimators for α and β

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

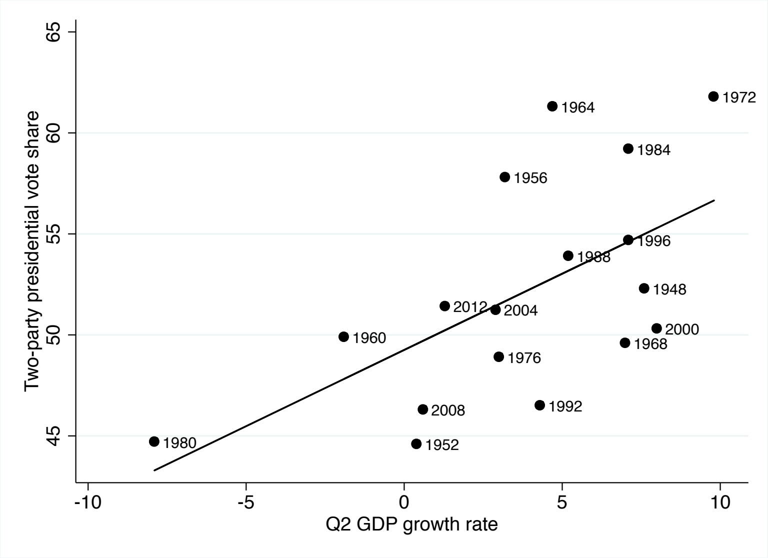
$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

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

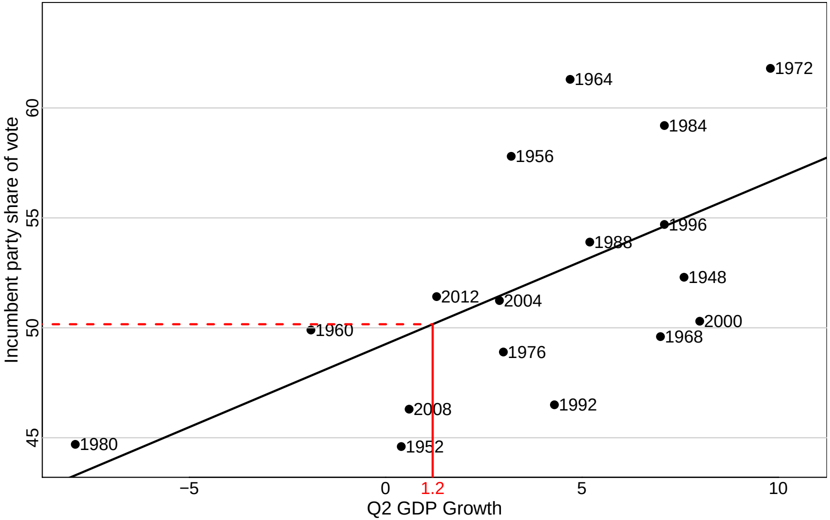
- $\hat{\alpha} = 49.3$
- $\hat{\beta} = 0.75$

What does that mean?

Visualizing bivariate regression



Who will win the election?



Example

X_i	Y_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 $\hat{\alpha}$ and $\hat{\beta}$.

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

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

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

$\sum Y_i = 18.2$ $\quad \sum X_i = 17.6$ $\quad \sum(Y_i - \bar{Y})(X_i - \bar{X}) = 1.863$

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

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
$\sum Y_i = 18.2$ $\bar{Y} = 3.033$	$\sum X_i = 17.6$ $\bar{X} = 2.933$			$\sum(Y_i - \bar{Y})(X_i - \bar{X}) = 1.863$

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

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
$\sum Y_i = 18.2$ $\bar{Y} = 3.033$	$\sum X_i = 17.6$ $\bar{X} = 2.933$			$\sum (Y_i - \bar{Y})(X_i - \bar{X})$ $= 1.863$

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

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

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

$\sum Y_i = 18.2$ $\quad \sum X_i = 17.6$ $\quad \sum(Y_i - \bar{Y})(X_i - \bar{X}) = 1.863$

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

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

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