Week 3
Contingency Tables & Regression Assumptions

Applied Statistical Analysis/Quantitative Methods I

Jeffrey Ziegler, PhD

Assistant Professor in Political Science & Data Science
Trinity College Dublin

Fall 2024

Roadmap for today

Last class we learned:

- What is a hypothesis test?
  - We will test hypotheses about population from sample
- Relationship between CI and null hypothesis testing
- Five steps of hypothesis testing
- Types of errors
- Discussion of 1-sided/2-sided tests

Sometimes, our specific aim is to understand if X causes Y

- Compare two independent samples
- So, today...

Transition: From causation to association

This class:

- Estimate if two variables are dependent
  - Chi-squared test of independence
  - Standardized residuals
- Estimating correlation
  - Does variation in one explain variation in another
- Bivariate regression
  - Assumptions
  - Estimation (i.e. drawing the "best" line through data)

We have 2 categorical variables, any relation?

- If we have three samples, our data might look like this

Variable 1 (Outcome or response)	Variable 2 (Sample or grouping)
1	1
2	0
3	1
5	2
3	2
2	0
4	0
⋮	⋮

Cross-tabs: The basics

Assume we have two variables that are nominal

- Party-ID and racial/ethnicity
- **WARNING:** Calculations are different when variables are ordinal ... especially if BOTH are ordinal (see end of chapter 8 in A&F)

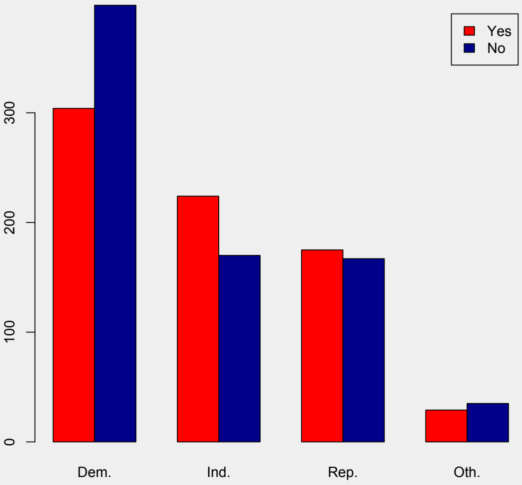
We will use a contingency table, which is usually (at least by me) referred to as a "cross-tabulation"

Example: Preferences over abortion policy

"Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if … the family has a very low income and cannot afford any more children?"

	Yes	No	Total
Democrats	304	398	702
Independents	224	170	394
Republicans	175	167	342
Other	29	35	64
Total	732	770	1502

Plot: Data comes from 1972 GSS



Not an experiment! What do we do?

We want conditional distribution of outcome

How do we get that?

- What is the distribution of the outcome variable conditioned on the independent/grouping variable? Proportions?!

	Yes	No
Democrats	0.43	0.57
Independents	0.57	0.43
Republicans	0.51	0.49
Other	0.45	0.55

Note that the rows total to 100% because the rows indicate the independent variable

- We might have the columns add up to 100% if that was our explanatory variable

***Statistical independence**: Two variables are statistically independent if the conditional distributions of the **population** are identical across categories*[1]

Note: Events can be *logically* or *physically independent* but still *statistically dependent*.

- Let A = scored above 90% on entrance exam,
- and B = attends Trinity

These are logically independent (neither one implies the other), but statistically quite dependent, because $Pr(A|B) > Pr(A)$

[1] Since $Pr(A \cap B) = Pr(A|B)Pr(B)$, if A is independent of B, then $Pr(A \cap B) = Pr(A)Pr(B)$. If this holds, then B is also independent of A since $Pr(B|A) = \frac{Pr(A \cap B)}{Pr(A)} = \frac{Pr(A)Pr(B)}{Pr(A)} = Pr(B)$

Chi-square test: The intuition

H_0: The variables are statistically independent

H_a: The variables are statistically dependent

We are going to calculate a test-statistic (the χ^2 statistic) that is distributed according to the χ^2 distribution

$f_{observed} = f_o =$ observed frequency = the raw count (NOT THE %)

$f_{expected} = f_e =$ what we would expect for independent samples

$$= \frac{\text{Row total}}{\text{Grand total}} \times \text{Column total}$$

If H_0 is true, then we would expect $f_{observed} = f_{expected}$

Chi-square statistic: The calculations

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Example: Chi-square test

	Yes	No	Total
Democrats	f_o=304	f_o=398	702
Independents	f_o=224	f_o=170	394
Republicans	f_o=175	f_o=167	342
Other	f_o=29	f_o=35	64
Total	732	770	1502

Chi-square test: Example

$$f_{1e} = \frac{\text{row total}}{\text{grand total}} * \text{column total} \quad (1)$$

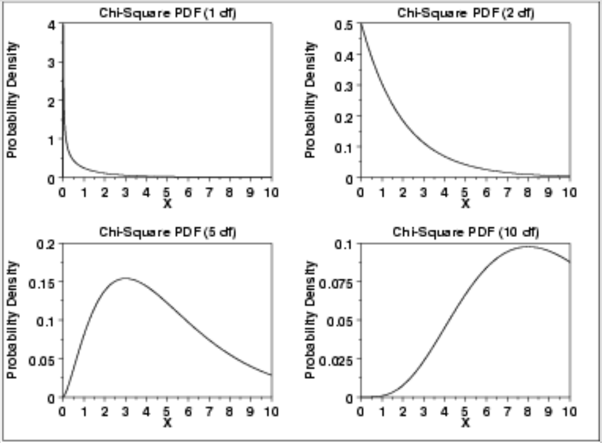
$$= \frac{702}{1502} * 732 = 342.12 \quad (2)$$

	Yes	No	Total
Democrats	f_o=304	f_o=398	702
	$\mathbf{f_e = 342.12}$	$f_e = 359.88$	
Independents	f_o=224	f_o=170	394
	$f_e = 192.12$	$f_e = 201.98$	
Republicans	f_o=175	f_o=167	342
	$f_e = 166.67$	$f_e = 175.33$	
Other	f_o=29	f_o=35	64
	$f_e = 31.19$	$f_e = 32.81$	
Total	732	770	1502

Now we want to calculate the χ^2 statistic

$$
\begin{aligned}
\chi^2 &= \sum \frac{(f_o - f_e)^2}{f_e} && (3)\\
&= \frac{(304 - 342.12)^2}{342.12} + \frac{(398 - 359.88)^2}{359.88} + \ldots + && (4)\\
&\approx 19.79 && (5)
\end{aligned}
$$

Is the χ^2 statistic "big enough?"

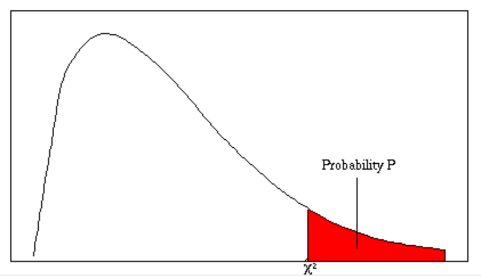


Calculating p-values for Chi-squared tests

- We are going to need to calculate the degrees of freedom
- This is skewed right and strictly positive
- Always use the upper-tail (no $\times$ 2)

Calculating p-values for Chi-squared tests

- Frequency ≥ 5 for all cells
- $df = (rows - 1)(columns - 1)$
- In R: `pchisq(`χ^2`, df = (rows-1)(columns-1), lower.tail=FALSE)`



Example: Preferences over abortion policy

What is df? $df = 3$

How should we get our p-value?

$$\begin{aligned} \text{p-value} &= \texttt{pchisq(19.79, df=3, lower.tail=F)} \quad (6) \\ &= 0.00019 \quad (7) \end{aligned}$$

Standardized residuals

Now, we have evidence that the two variables are not independent

- Where does the deviation from independence take place?
- Why did we reject the null?
- What does it mean?

Standardized residuals

How far away is each observed value from "expectation"

We need to find the **adjusted residual**:

$$z = \frac{f_{observed} - f_{expected}}{se} = \frac{f_{observe} - f_{expected}}{\sqrt{f_e(1 - \text{row prop.})(1 - \text{column prop.})}}$$

- The denominator is the standard error of the quantity $f_o - f_e$ under the null hypothesis

Example: Calculating standardized residuals

	Yes	No	Total
Democrats	f_o=304	f_o=398	702
	$f_e = 342.12$	$f_e = 359.88$	
Independents	f_o=224	f_o=170	394
	$f_e = 192.12$	$f_e = 201.98$	
Republicans	f_o=175	f_o=167	342
	$f_e = 166.67$	$f_e = 175.33$	
Other	f_o=29	f_o=35	64
	$f_e = 31.19$	$f_e = 32.81$	
Total	732	770	1502

$$z_{11} = \frac{304 - 342.12}{\sqrt{342.12(1 - \frac{702}{1502})(1 - \frac{732}{1502})}} \approx -2.395$$

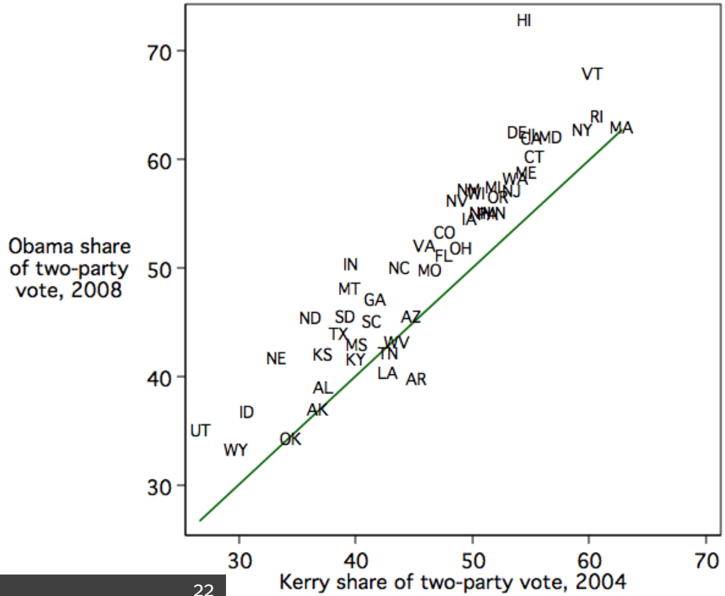
From independence to association

We should always visually inspect data using... scatterplots

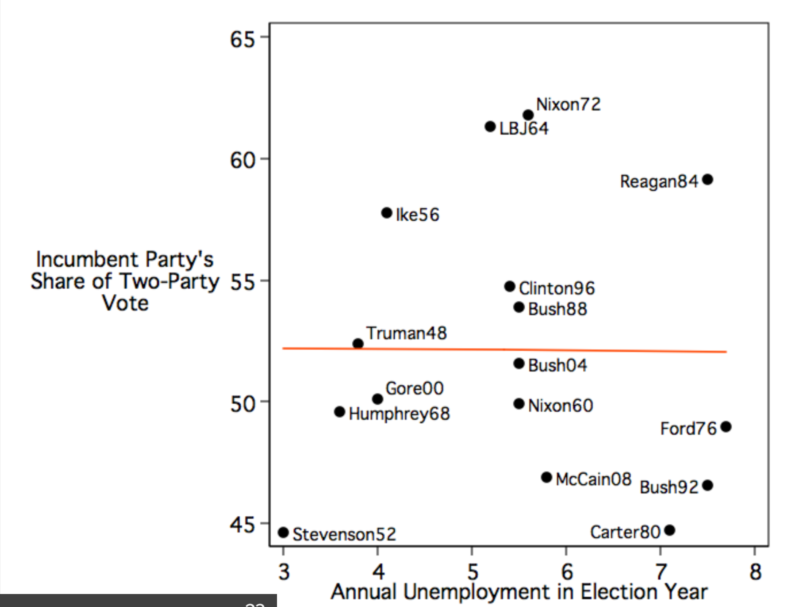
What are we looking for?

- Form/pattern
- Direction
- Strength
- Outliers

SCATTERPLOTS (MASKET 2008)



Scatterplots (Masket 2011)



How do we quantify this?

Correlation! But, first how do we assess variation in two variables...

Step 1: Standardizing variation in variables

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

Example: Populations of European countries

	x	$\frac{x-\bar{x}}{s}$
AT	8.96mil	?
BE	11.69mil	?
EL	10.34mil	?
IE	5.05mil	?
MT	0.54mil	?
PL	41.03mil	?

Standardizing variables: Create data in R

```r
# create vector
x <- c(8.96, 11.69, 10.34, 5.05, 0.54, 41.03)
```

Standardizing variables: Mean and SD in R

```
# create vector
x <- c(8.96, 11.69, 10.34, 5.05, 0.54, 41.03)
# get mean and sd
c(round(mean(x), 2), round(sd(x), 2))
```

```
[1] 12.94 14.35
```

$$\bar{x} = 12.94 \quad s = 14.35$$

STANDARDIZING VARIABLES: FILL IN TABLE USING R

	x	$\frac{x-\bar{x}}{s}$
AT	8.96mil	?
BE	11.69mil	?
EL	10.34mil	?
IE	5.05mil	?
MT	0.54mil	?
PL	41.03mil	?

Remember the formula

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

STANDARDIZING VARIABLES

```
# create standardized distance for each observation
standardized.x <- (x - mean(x))/sd(x)
# view vector of standardized values
round(standardized.x, 2)
```

```
[1] -0.28 -0.09 -0.18 -0.55 -0.86 1.96
```

	x	$\frac{x-\bar{x}}{s}$
AT	8.96mil	-0.28
BE	11.69mil	-0.09
EL	10.34mil	-0.18
IE	5.05mil	-0.55
MT	0.54mil	-0.86
PL	41.03mil	1.96

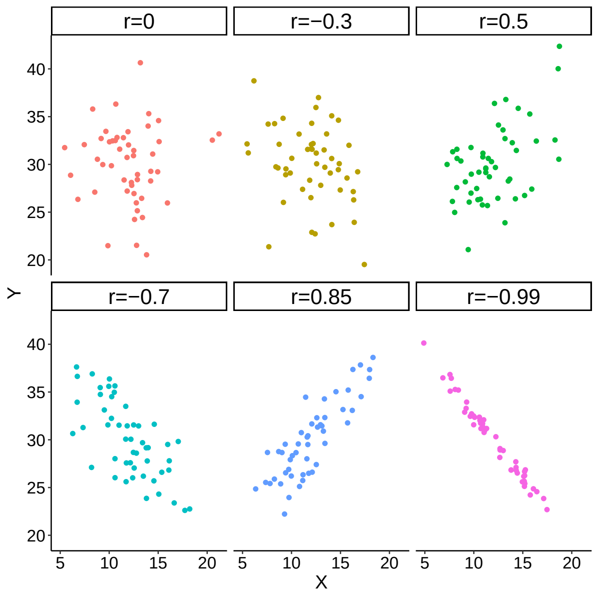
Step 2: Correlation coefficient

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



Facts about correlation

- Linear only
- **Not** causal
- Unit-free
- $-1 \leq r \leq 1$ (Remember that r is the correlation coefficient)
- Sensitive to outliers

From correlation to Regression: The big picture

Assume we have 2 variables

We need to:

- Ask: Is there an "association" between them?
- Is it statistically significant (we'll discuss this next class)?
- Estimate "expected values" for an outcome variable given a set of covariates

Starting off: Some preliminaries

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

We want to know: How might Y and X be related?

- We can visualize the relationship with a line!

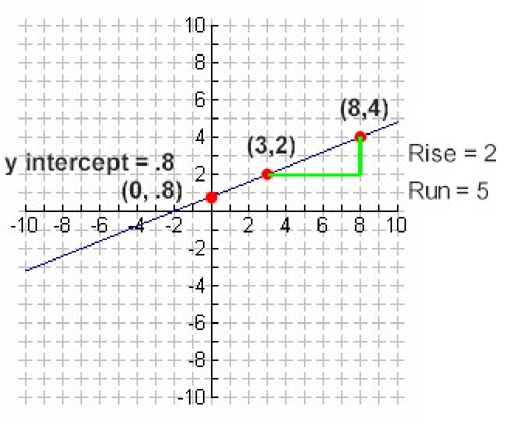
Linear Model

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

$\alpha =$ Y-intercept and $\beta =$ slope of the line

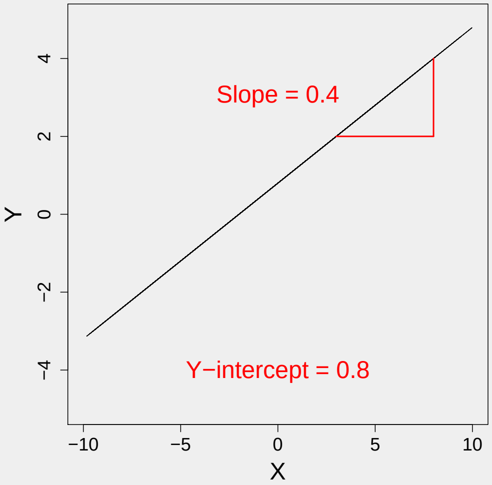
$Y =$ outcome vector and $X =$ vector of predictor

How do we write the regression line?



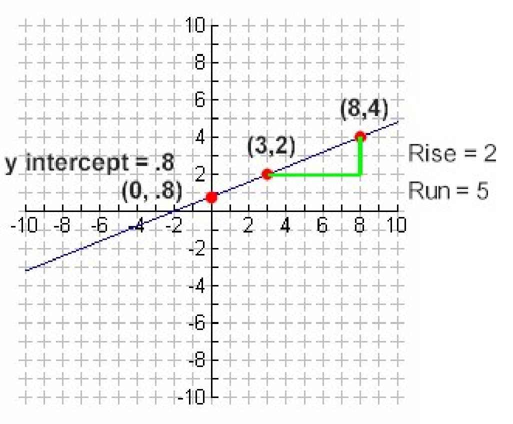
- $Y = 0.8 + \frac{2}{5}X$
- Interpret α: = 0.8 = value of Y when $X = 0$
- Interpret β: = 0.4 = a 1 unit $\uparrow$ in X is associated with a 0.4 unit $\uparrow$ in Y

Do it ourselves: Components of a line in R



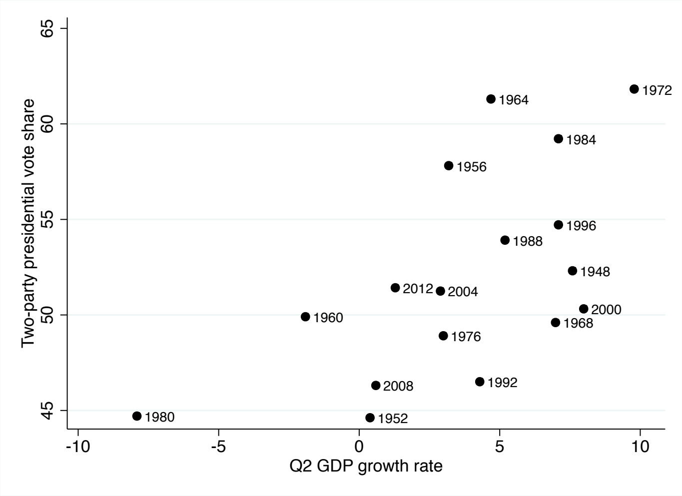

```r
# create and view components of regression lines
# we have two variables X and Y
X <- runif(100, -10, 10)
Y <- 0.8 + (2/5)*X
# plot relationship (with a line)
plot(X, Y, type="l", xlim = c(-10, 10), ylim = c(-5,5))
```

Interpreting a regression line



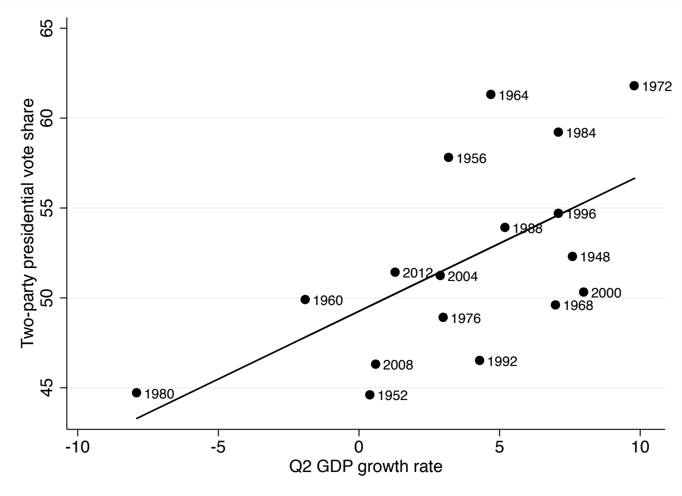
- α: E[Y] when X is zero
- $\beta > 0$: Positive relationship between X and Y
- $\beta < 0$: Negative relationship between X and Y
- $\beta = 0$: Null relationship between X and Y

Ex: Pres. elections & GDP Change from 1952-2012



Is there an association?

Our best guess for the "best" line



Incumbent party vote = 39.3 + 0.75 Q_2 GDP

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

Assumption #3: There's a linear relationship between the variables:

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

However, we know there is error, so

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

Assumption #4: $\epsilon_i \sim N(0, \sigma^2)$

This is equivalent to:

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

Reminder: Assumptions for linear regression

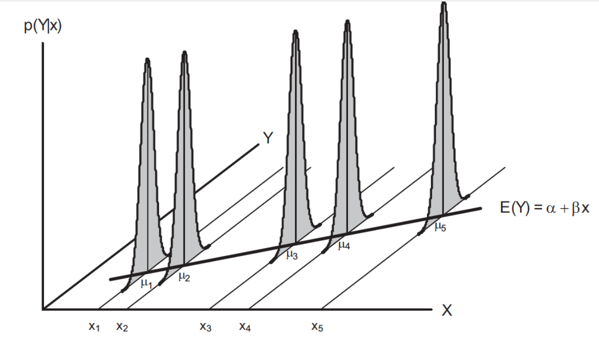
(1) Randomized data generation

(2) Independent observations

New: Assumptions for linear regression

(3) Linearity: Population means of y at different values of x have a straight-line relationship with x, i.e. $\mu_{y|x} = \beta_0 + \beta_1 x$

(4) Normality and Constant Variance: Population values of y at each value of x follow a **normal** distribution, with the **same** standard deviation σ at each x value (constant variance in y for all x)



Visualizing perfectly normal errors

First, create data:

```
# create data
# (1) draw N=100 from uniform distribution
# w/ min=0 and max=1
X <- runif(100, 0, 1)
# (2) draw corresponding outcome for N=100
# w/ mean=0 and sd=1
# embed correlation w/ X so we know what
# the relationship should be
Y <- 2 + X*1.5 + rnorm(100, 0, 1)
# run bivariate regression
temp.model <- lm(Y~X)
```

Visualizing perfectly normal errors

Notice we have an imperfect model of "true" relationship between Y and X:

Table: Estimated relationship between Y and X

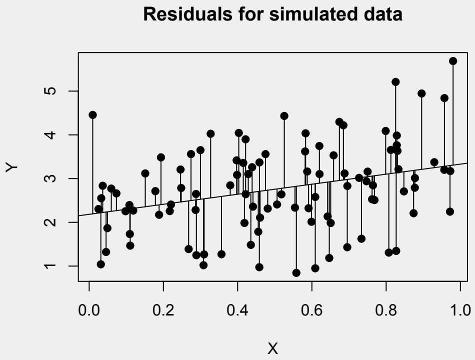
	Outcome = Y
X	1.72^{***}
	(0.34)
Intercept	2.02^{***}
	(0.19)

Notes: OLS regression coefficients shown with standard errors in parentheses.
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. In all models, N=100.

Visualizing perfectly normal errors

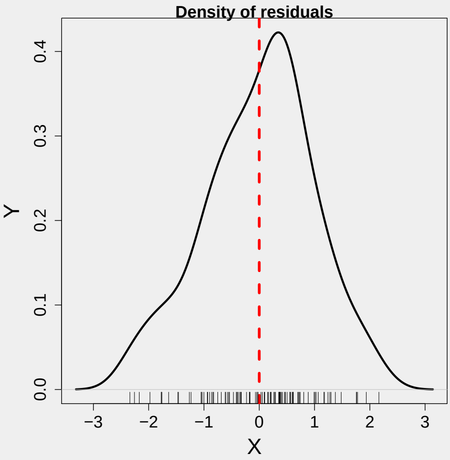
Plot residuals:

```
# plot residuals
plot(X, Y, pch=19, main
    ="Residuals for
    simulated data")
# display estimated
    regression line
abline(temp.model)
# show how far each
    prediction is from
# the estimated
    regression line
preds <- predict(temp.
    model)
segments(X,Y,X,preds)
```



Visualizing perfectly normal errors

Check that residuals are zero in expectation ($\epsilon_i \sim N(0, \sigma^2 = 1)$):




```r
plot(density(Y-preds), main="Density of residuals",
     ylab="Y", xlab="X", cex.axis=1.5, cex.lab=2, cex.main=1.5, lwd=3)
```

Implications of what we've done so far

- We have reduced all of data to a simplified model
- We have three parameter (α, β, σ^2) 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 choose estimates:

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

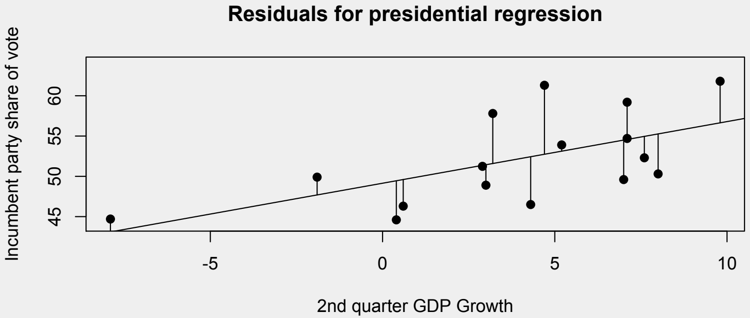
Pick parameters to minimize error: Define ϵ

Let's define observed residual for observation i as e_i (our "error")

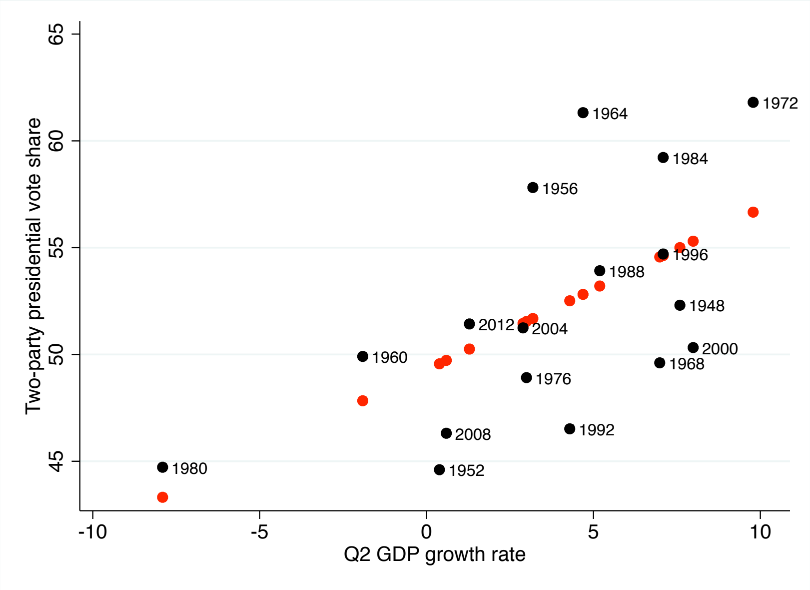
- Difference between our "best guess" for value of Y_i given X_i & what was actually observed (similar to χ^2 test)

Residuals

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



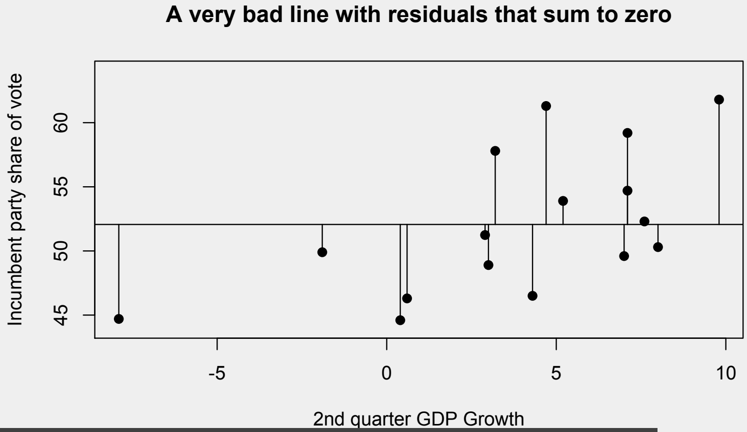
Another look at residuals from our example



Again, think back to drawing the "best" line

Intuition: "Best" line is the one that reduces the greatest error

- We might think to first check sum of the errors $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)$, but that's not actually a good criterion, why? (errors equal out)



Defining "best" as minimizing SSE

For many good statistical reasons, we'll say any line that reduces the **"Sum of Squared Error"** is equivalent to having the "best" line (defined as *"most efficient unbiased estimator"*)

- For right now, trust me!

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 $\hat{\beta}$ (using calculus in the background!)

- With these parameters, we can draw "best" lines

Estimators for α and β

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

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

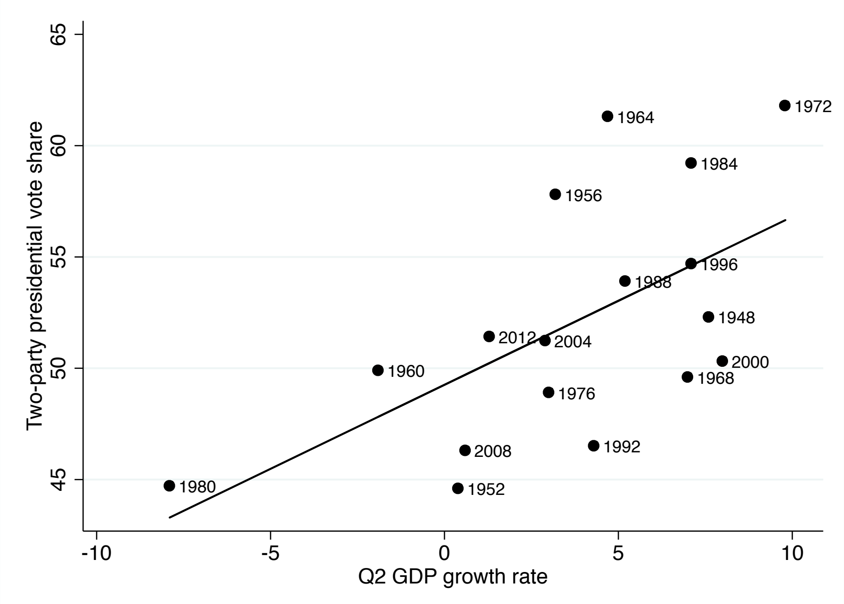
Both of these are functions of the data

For our presidential election data:

- $\hat{\alpha} = 39.3$
- $\hat{\beta} = 0.75$

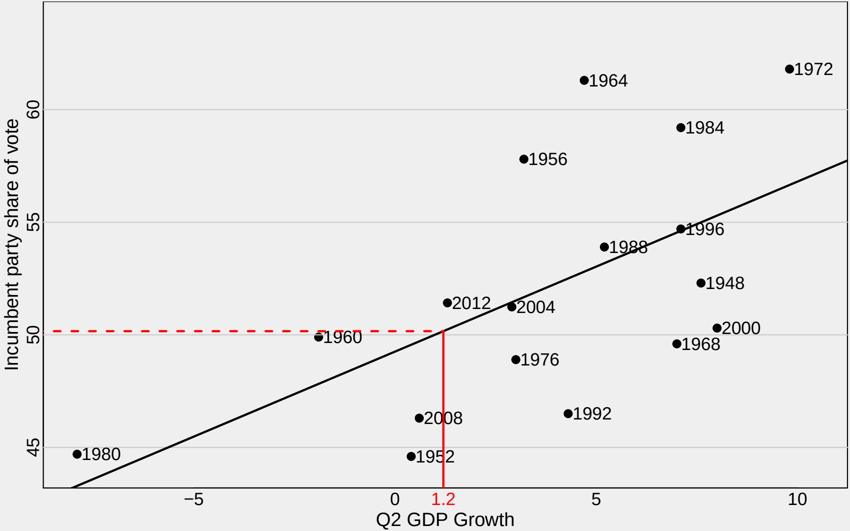
How do we interpret these estimates?

Example: Interpreting estimates



Who will win the election?

If Q_2 GDP $\Delta = 1.2$ ...



Example: Bivariate regression "by hand"

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 $\hat{\beta}$

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

For each observation...

You need to compute each component:

Y_i	X_i	$(Y_i - \bar{Y})$	$(X_i - \bar{X})$	$(Y_i - \bar{Y})(X_i - \bar{X})$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Calculate each component "by hand" in R

(1) Create data in R:

```
# bivariate regression by hand
regressMat <- as.data.frame(matrix(c(3.8, 3.5, 3.0, 3.3,
    3.5, 4.0, 2.8, 2.3, 2.4, 1.8, 2.7, 2.7), nrow=6, byrow
    = T))
colnames(regressMat) <- c("X", "Y")
```

"By hand": (2) Calculate sums and means

$\bar{Y} = 2.93$

```
mean(regressMat$Y)
```

$\bar{X} = 3.03$

```
mean(regressMat$X)
```

$\sum Y_i = 17.6$

```
sum(regressMat$Y)
```

$\sum X_i = 18.2$

```
sum(regressMat$X)
```

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

```
sum((regressMat$Y -
       mean(regressMat$Y))
    *(regressMat$X-
        mean(regressMat$X)))
```

$\sum(X_i - \bar{X})^2 = 1.373$

```
sum((regressMat$X-
       mean(regressMat$X))^2)
```

Calculate each component "by hand" in R

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	0.435
3.0	3.3	-0.033	0.367	-0.012
3.5	4.0	0.467	1.067	0.498
2.8	2.3	-0.233	-0.633	0.147
2.4	1.8	-0.633	-1.133	0.717
2.7	2.7	-0.333	-0.233	0.078

$$\hat{\beta} = \frac{1.863}{1.373} = 1.357$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 2.933 - 1.357(3.033) = -1.183$$

Check with lm( ) in R

```
lm(Y~X, data=regressMat)
```

Table: Estimated relationship between Y and X

	Outcome = Y
X	1.36^{*}
	(0.38)
Intercept	-1.18
	(1.18)

Notes: OLS regression coefficients shown with standard errors in parentheses. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

We'll talk about standard errors next week!

Wrap-up

Today we...

1. Estimate if two variables are dependent
   - Chi-squared test of independence
   - Standardized residuals
2. Correlations
3. Simple linear regression:
   - Assumptions
   - Estimation

Class business

- Read the required and suggested online materials
- Work on Problem set # 1!
- These slides are available on the course website