Lecture 16: Multiple Regression

Jacob M. Montgomery

Quantitative Political Methodology

Multiple regression

Roadmap

- **Before**: Regression with one explanatory variable
- **Today** we will learn how to:
  - Draw the best (hyper)plane through the data
  - Interpret multivariate regression results

Class business

- PS is due on Wed.
- Take notes on this one

A big day

- Introducing multivariate regression
  - An example (time for change model)
  - (Hyper)planes in (hyper)space
  - Specifying and estimating the regression model

A big day

- Introducing multivariate regression
  - An example (time for change model)
  - (Hyper)planes in (hyper)space
  - Specifying and estimating the regression model

- Three ways to think about regression
  - Hyperplanes
  - Lines within "groups"
  - Added variable plots

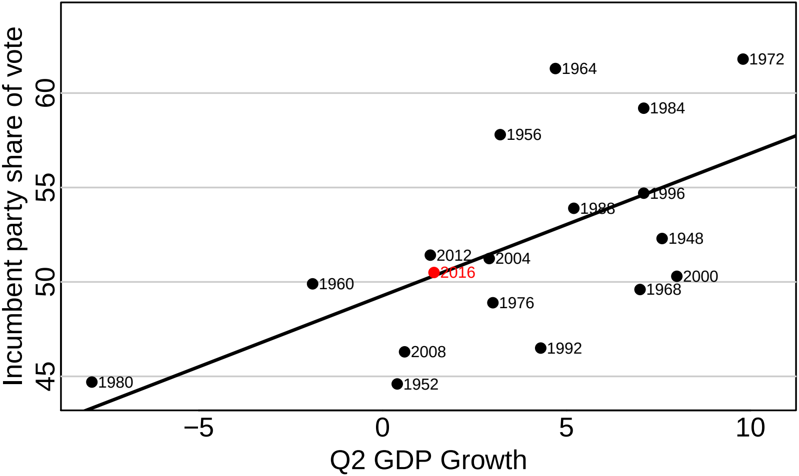
A big day

- Introducing multivariate regression
  - An example (time for change model)
  - (Hyper)planes in (hyper)space
  - Specifying and estimating the regression model
- Three ways to think about regression
  - Hyperplanes
  - Lines within “groups”
  - Added variable plots
- Inference for multivariate regression

A big day

- Introducing multivariate regression
  - An example (time for change model)
  - (Hyper)planes in (hyper)space
  - Specifying and estimating the regression model
- Three ways to think about regression
  - Hyperplanes
  - Lines within "groups"
  - Added variable plots
- Inference for multivariate regression
- A brief word on Simpson's paradox

So far we have looked at data like this



But what if it is time for change?

Success of Incumbent Party Candidate in Presidential Elections by Type of Election, 1948-2016

Results	First-Term	Second- or Later
Won	8	2
Lost	1	8
Average vote	55.3	49.3

Accounting for time in office

Estimate a more complex equation:

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

`where:`

- μ_y is mean presidential vote share
- β_0 is the y-intercept ("constant")
- β_1 is the slope ("coefficient") for Q2 GDP growth
- x_1 is Q2 GDP growth in the election year
- β_2 is the slope ("coefficient") for TFC ("time for a change")
- x_2 is an indicator ("dummy") variable for TFC (1=first term; 0=second term or later)

Equation for the graph:

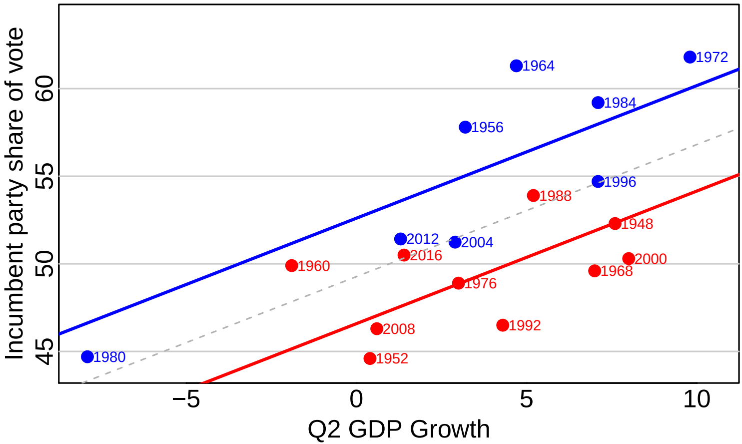
$$\text{Vote share} = 46.59 + 0.76 \times \text{Q2 GDP} + 6.02 \times \text{FirstTermInc}$$

or

$$\text{Vote share}_{\text{TFC}} = 46.59 + 0.76 \times \text{Q2 GDP}$$

$$\text{Vote share}_{\text{Not TFC}} = 52.61 + 0.76 \times \text{Q2 GDP}$$

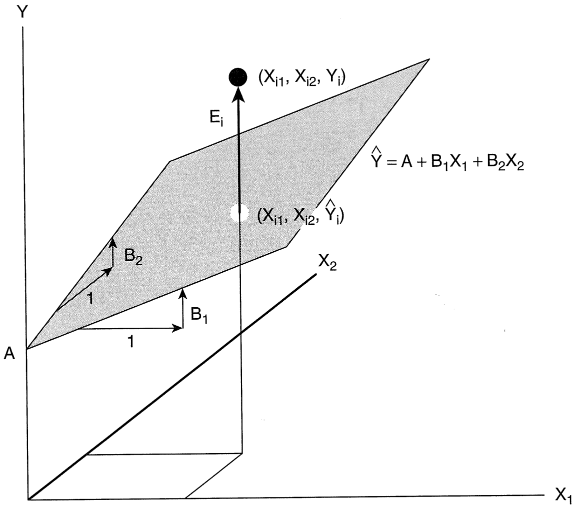
Multivariate regression



A big day

- Introducing multivariate regression
  - An example (time for change model)
  - **(Hyper)planes in (hyper)space**
  - Specifying and estimating the regression model
- Three ways to think about regression
  - Hyperplanes
  - Lines within "groups"
  - Added variable plots
- Inference for multivariate regression
- A brief word on Simpson's paradox

Multivariate regression



Beyond two dimensions

Incumbent party vote share

	Model 1	Model 2
Intercept	49.27	49.35
	(1.35)	(4.51)
2nd Qtr GDP	0.754	0.451
	(0.248)	(0.161)
June Polling		0.147
		(0.085)
Multiple R-Squared	0.366	0.781

Standard errors are in parentheses. N=18.

Beyond two dimensions

Incumbent party vote share

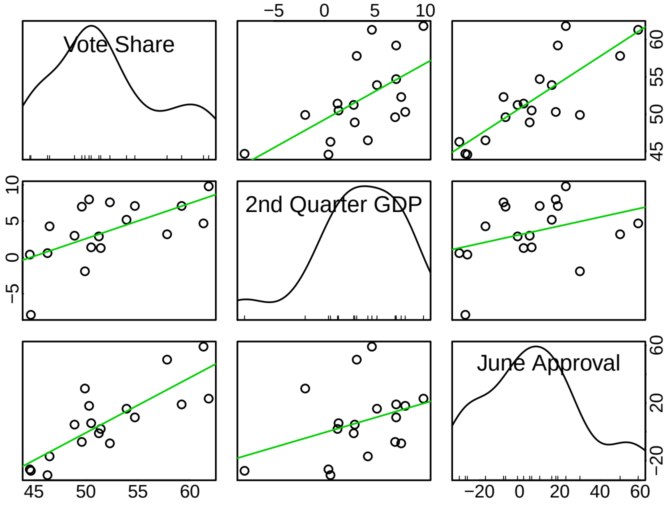
	Model 1	Model 2
Intercept	49.27	49.35
	(1.35)	(4.51)
2nd Qtr GDP	0.754	0.451
	(0.248)	(0.161)
June Polling		0.147
		(0.085)
Multiple R-Squared	0.366	0.781

Standard errors are in parentheses. N=18.

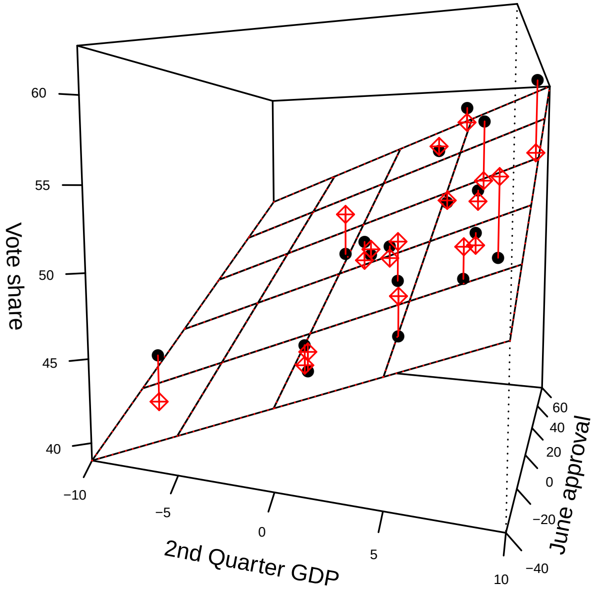
Two questions to try to understand:

- What do the coefficients (and standard errors) mean?
- Why did the “2nd Quarter GDP” coefficient change?

Now we need to think about data like this



Or even better . . . . this



A big day

- Introducing multivariate regression
  - An example (time for change model)
  - (Hyper)planes in (hyper)space
  - **Specifying and estimating the regression model**
- Three ways to think about regression
  - Hyperplanes
  - Lines within "groups"
  - Added variable plots
- Inference for multivariate regression
- A brief word on Simpson's paradox

To draw the "best" line we wanted to minimize error

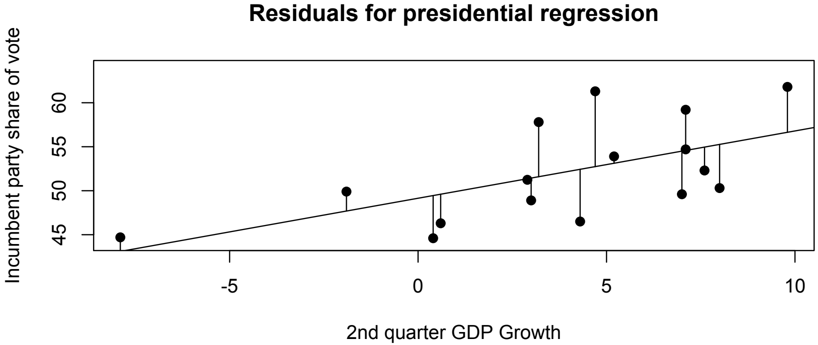
Residuals:

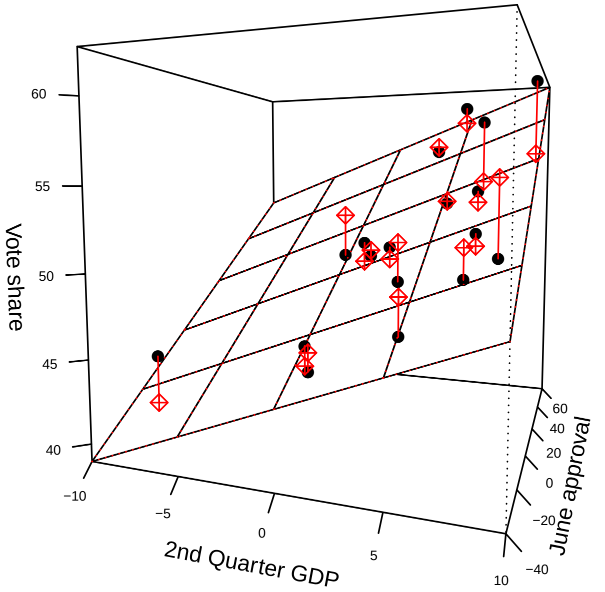
$$e_i = (Y_i - \hat{Y}_i) = (Y_i - \hat{\alpha} - \hat{\beta}X_i)$$

To draw the "best" line we wanted to minimize error

Residuals:

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





Vote share
60
55
50
45
40
2nd Quarter GDP
−10
−5
0
5
10
June approval
60
40
20
0
−20
−40

Multidimensional "linear" models

On **average**, we are hypothesizing that the **world** looks like this:

$$E(Y) = \alpha + \beta_1 X_1 + \ldots + \beta_k X_k$$

Multidimensional “linear” models

On **average**, we are hypothesizing that the **world** looks like this:

$$E(Y) = \alpha + \beta_1 X_1 + \ldots + \beta_k X_k$$

Overall, we think that the **data** looks like this

$$Y_i = \alpha + \beta_1 X_{1,i} + \ldots + \beta_k X_{k,i} + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

Just like before, we need to decide on a rule to choose the best estimates:

$$\hat{\alpha}, \hat{\sigma}^2, \hat{\beta}_1, \hat{\beta}_2, \ldots$$

Residuals, SSE, and $\hat{\sigma}^2$

- Residuals

$$e_i = (Y_i - \hat{Y}_i) = (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \ldots - \hat{\beta}_k X_{ki})$$

- Sum of squared error

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

Residuals, SSE, and $\hat{\sigma}^2$

- Residuals

$$e_i = (Y_i - \hat{Y}_i) = (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \ldots - \hat{\beta}_k X_{ki})$$

- Sum of squared error

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \ldots - \hat{\beta}_k X_{ki})^2$$

- Conditional Variance: Estimate of variance around hyperplane in population

$$\hat{\sigma}^2 = \frac{SSE}{n-(k+1)} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-(k+1)} \Rightarrow \hat{\sigma} = \sqrt{\frac{SSE}{n-(k+1)}} = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n-(k+1)}}$$

A big day

- Introducing multivariate regression
  - An example (time for change model)
  - (Hyper)planes in (hyper)space
  - Specifying and estimating the regression model
- **Three ways to think about regression**
  - Hyperplanes
  - Lines within “groups”
  - Added variable plots
- Inference for multivariate regression

A big day

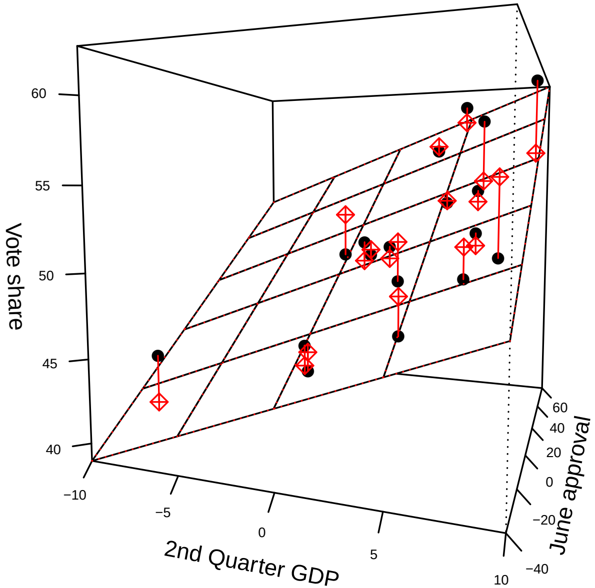
- Introducing multivariate regression
  - An example (time for change model)
  - (Hyper)planes in (hyper)space
  - Specifying and estimating the regression model
- **Three ways to think about regression**
  - Hyperplanes
  - Lines within "groups"
  - Added variable plots
- Inference for multivariate regression
- A brief word on Simpson's paradox

Incumbent party vote share

	Model 1	Model 2
Intercept	49.27	49.35
	(1.35)	(4.51)
2nd Qtr GDP	0.754	0.451
	(0.248)	(0.161)
June Polling		0.147
		(0.085)
Multiple R-Squared	0.366	0.781

Standard errors are in parentheses. N=18.

Thinking about regression 1: Planes



Thinking about regression 2: Lines within groups

Let X_i take on a value of only 0 or 1.

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

Thinking about regression 2: Lines within groups

Let X_i take on a value of only 0 or 1.

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

$$E(Y_i | X_i = 0) = \alpha$$
$$E(Y_i | X_i = 1) = \alpha + \beta$$

Thinking about regression 2: Lines within groups

Let X_i take on a value of only 0 or 1.

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

$$E(Y_i | X_i = 0) = \alpha$$
$$E(Y_i | X_i = 1) = \alpha + \beta$$

(This provides the same inference as a t-test)

Let's look at this with nominal data

$$X_1 = \{\text{Blue}, \text{Not blue}\}, X_2 = \{\text{Brown}, \text{Not brown}\}$$

Let's look at this with nominal data

$$X_1 = \{\text{Blue}, \text{Not blue}\}, X_2 = \{\text{Brown}, \text{Not brown}\}$$

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

Let's look at this with nominal data

$$X_1 = \{\text{Blue}, \text{Not blue}\}, X_2 = \{\text{Brown}, \text{Not brown}\}$$

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

$$E(Y_i | Blue) = \alpha + \beta_1$$

Let's look at this with nominal data

$$X_1 = \{\text{Blue}, \text{Not blue}\}, X_2 = \{\text{Brown}, \text{Not brown}\}$$

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

$$E(Y_i|Blue) = \alpha + \beta_1$$

$$E(Y_i|Brown) = \alpha + \beta_2$$

Let's look at this with nominal data

$$X_1 = \{\text{Blue}, \text{Not blue}\}, X_2 = \{\text{Brown}, \text{Not brown}\}$$

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

$$E(Y_i | Blue) = \alpha + \beta_1$$

$$E(Y_i | Brown) = \alpha + \beta_2$$

$$E(Y_i | Green) = \alpha$$

Let's look at this with nominal data

$$X_1 = \{\text{Blue}, \text{Not blue}\}, X_2 = \{\text{Brown}, \text{Not brown}\}$$

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

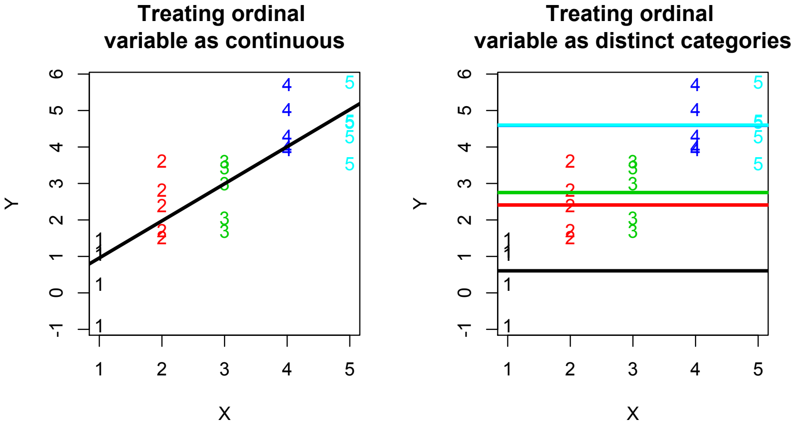
$$E(Y_i | Blue) = \alpha + \beta_1$$

$$E(Y_i | Brown) = \alpha + \beta_2$$

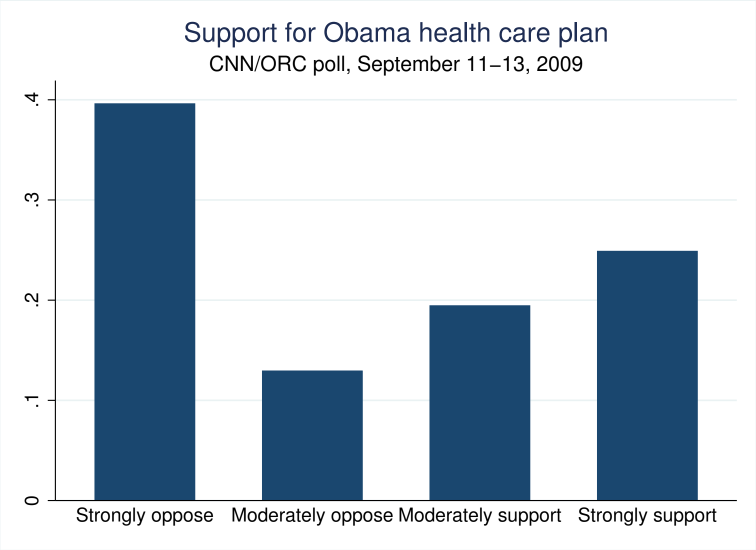
$$E(Y_i | Green) = \alpha$$

Ordinal data

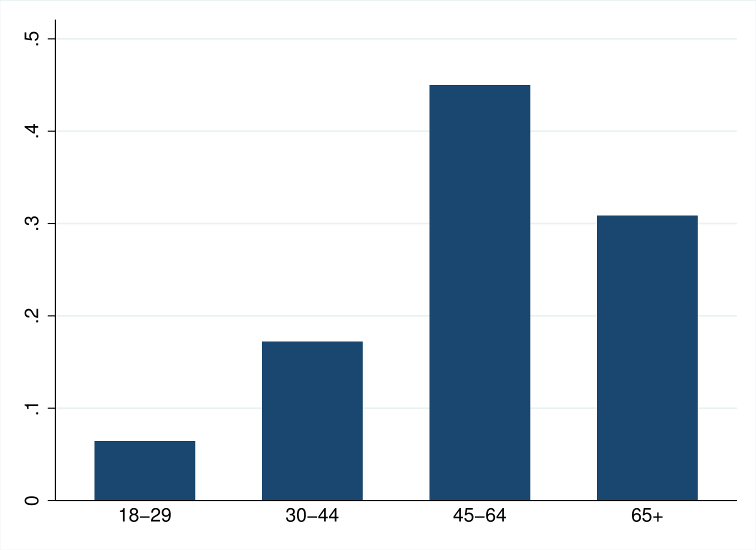
$$X = \{1, 2, 3, 4, 5\}$$



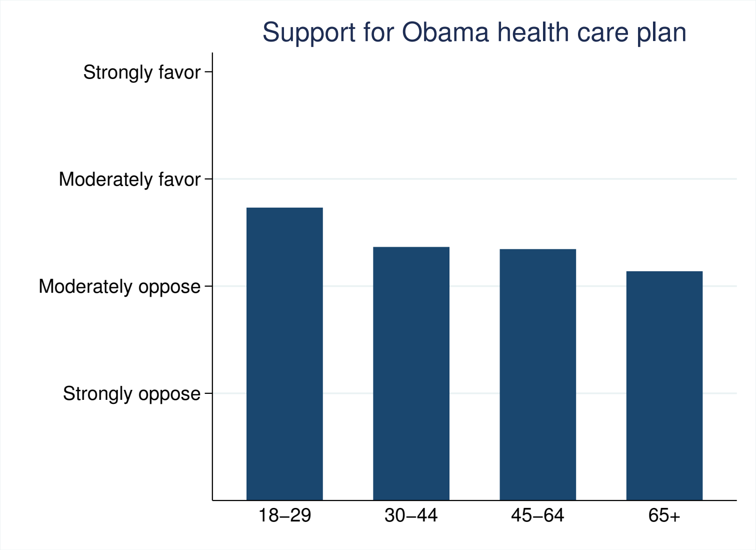
Example: 2009 health care poll



Example: 2009 health care poll



Example: 2009 health care poll



Dummy variable regression

- What is the association between age and support for HCR controlling for party?
- Goal: Recode age variable (18-29=1, 30-44=2, 45-64=3, 65+=4) into dummy variables

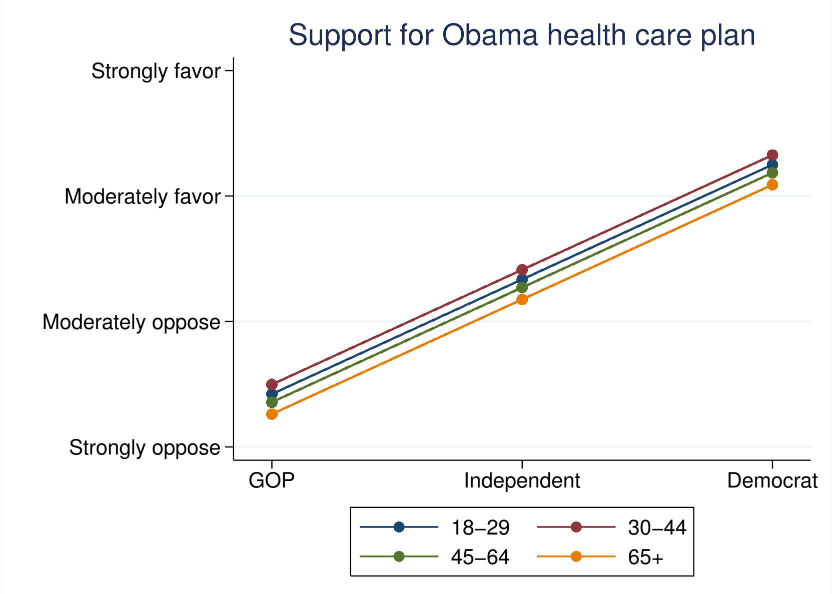
Equation:

$$\text{HCRS} = \beta_0 + \beta_1 \text{ Party} + \beta_2 \text{ Age 30-44} + \beta_3 \text{ 45-64} + \beta_4 \text{ 65+}$$

Dummy variable results

Variable	
Constant	1.421
	(0.116)
Party	0.914
	(0.031)
Age 30-44	0.77
	(0.13)
Age 45-64	-0.65
	(0.117)
Age 65 +	-0.16
	(0.121)
N = 981	
$R^2 = 0.4799$	

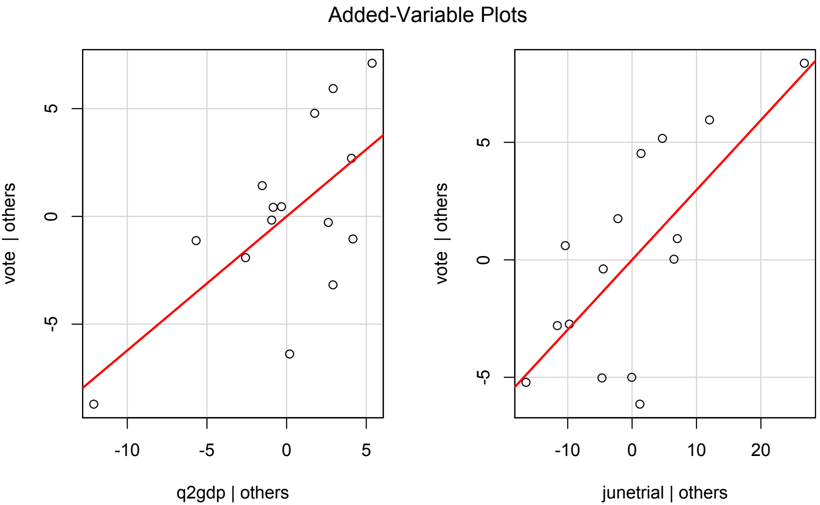
Dummy variable regression



Things to note

- For k levels of your categorical variable, you need to create $k-1$ dummy variables.
- The choice of baseline is arbitrary, but you need to know which is the baseline category in order to interpret the results correctly
- All effects are relative to the baseline category
- If you don't include them as separate dummies, you are assuming that the intercepts are equidistant and ordered.

Thinking about regression #3: Added variable plots



Thinking about X's from this point of view

- You are going to be doing this for the homework
- The slope of these lines corresponds to β estimates in the table.

Thinking about X's from this point of view

- You are going to be doing this for the homework
- The slope of these lines corresponds to β estimates in the table.
- Adding additional controls positively correlated to both the vote-share and GDP growth will lower the β for GDP growth.

Thinking about X's from this point of view

- You are going to be doing this for the homework
- The slope of these lines corresponds to β estimates in the table.
- Adding additional controls positively correlated to both the vote-share and GDP growth will lower the β for GDP growth.
- If we have many variables that are highly co-linear, often called multicollinearity, it will make coefficients smaller (and therefore less likely to be significant).
- It is difficult to decide on the "right" variables, but DO NOT use stepwise methods.

Thinking about X's from this point of view

- You are going to be doing this for the homework
- The slope of these lines corresponds to β estimates in the table.
- Adding additional controls positively correlated to both the vote-share and GDP growth will lower the β for GDP growth.
- If we have many variables that are highly co-linear, often called multicollinearity, it will make coefficients smaller (and therefore less likely to be significant).
- It is difficult to decide on the "right" variables, but DO NOT use stepwise methods.
- When in doubt, use theory.

A big day

- Introducing multivariate regression
  - An example (time for change model)
  - (Hyper)planes in (hyper)space
  - Specifying and estimating the regression model
- Three ways to think about regression
  - Hyperplanes
  - Lines within "groups"
  - Added variable plots
- **Inference for multivariate regression**
- A brief word on Simpson's paradox

Inference in regression coefficients

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

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

Inference in regression coefficients

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

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

- β_1: The coefficient for X_1.
- Interpretation: A one unit increase X_1 leads to a β_1 increase in Y *controlling for* the independent effect of X_2.

Inference in regression coefficients

$$Y_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

- β_1: The coefficient for X_1.
- Interpretation: A one unit increase X_1 leads to a β_1 increase in Y *controlling for* the independent effect of X_2.

Inference in regression coefficients

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

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

- β_1: The coefficient for X_1.
- Interpretation: A one unit increase X_1 leads to a β_1 increase in Y *controlling for* the independent effect of X_2.

We want to test whether X_1 has any effect on Y independent of X_2

$$\frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{df=n-(k+1)}$$

Inference in regression coefficients

$$Y_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

- β_1: The coefficient for X_1.
- Interpretation: A one unit increase X_1 leads to a β_1 increase in Y *controlling for* the independent effect of X_2.

We want to test whether X_1 has any effect on Y independent of X_2

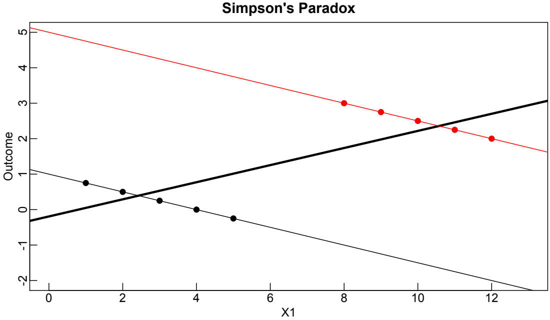
$$\frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{df=n-(k+1)}$$

- Just read these values off of the tables
- But watch your degrees of freedom.

A big day

- Introducing multivariate regression
  - An example (time for change model)
  - (Hyper)planes in (hyper)space
  - Specifying and estimating the regression model
- Three ways to think about regression
  - Hyperplanes
  - Lines within "groups"
  - Added variable plots
- Inference for multivariate regression
- **A brief word on Simpson's paradox**

Controlling for a variable can change the sign



$$E(Y) = 1 - 0.25X_1 + 2X_2$$

- Relationship between X_1 and Y is the same across groups.
- We can solve: $X_2 = 0$ for black observations, $X_2 = 2$ for red.

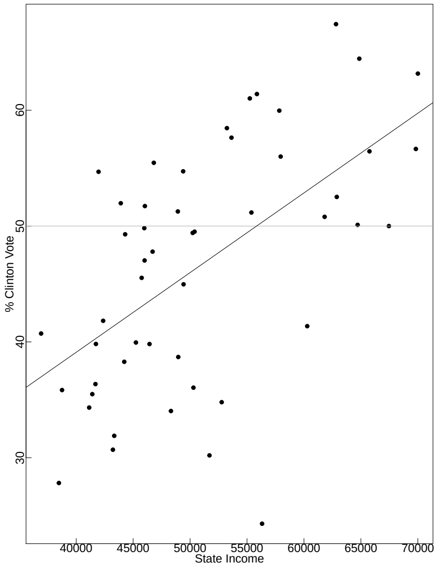
Applied example: Income in presidential voting

income

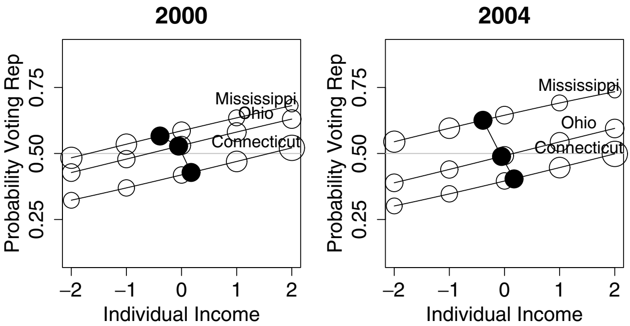
	clinton	trump	other/no answer
under $30,000 **17%**	**53%**	41%	6%
$30k-$49,999 **19%**	**51%**	42%	7%
$50k-$99,999 **31%**	46%	**50%**	4%
$100k-$199,999 **24%**	47%	**48%**	5%
$200k-$249,999 **4%**	48%	**49%**	3%
$250,000 or more **6%**	46%	**48%**	6%

24537 respondents

Applied example: Income in presidential voting



Applied example



Gelman et al. (2007):

- Rich states more likely to vote D (solid circles)
- Rich *within* states more likely to vote GOP (open circles)