Multivariate Regression

Prof. Jacob M. Montgomery

Quantitative Political Methodology (L32 363)

November 14, 2016
Class business

- PS is due Wed.
Class business

- PS is due Wed.
- You should take notes on this one.
Overview: 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
Motivating example: Presidential elections

But what if it’s “time for a change”? 

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

<table>
<thead>
<tr>
<th>Results</th>
<th>First-Term</th>
<th>Second- or Later</th>
</tr>
</thead>
<tbody>
<tr>
<td>Won</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Lost</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Average vote</td>
<td>55.3</td>
<td>49.3</td>
</tr>
</tbody>
</table>
Accounting for time in office

Estimate a more complex equation:

\[ \mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \]

where:

- \( \mu_y \) is mean presidential vote share
- \( \beta_0 \) is the y-intercept ("constant")
- \( \beta_1 \) is the slope ("coefficient") for Q2 GDP growth
- \( x_1 \) is Q2 GDP growth in the election year
- \( \beta_2 \) 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

\[ \text{Vote share}_{TFC} = 46.59 + 0.76 \times \text{Q2 GDP} \]
\[ \text{Vote share}_{\text{Not TFC}} = 52.61 + 0.76 \times \text{Q2 GDP} \]
Multivariate regression

Lecture 20 (QPM 2016)
Overview: 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

\[ \hat{Y} = A + B_1 X_1 + B_2 X_2 \]
**Beyond two dimensions**

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intercept</strong></td>
<td>49.27</td>
<td>49.35</td>
</tr>
<tr>
<td></td>
<td>(1.35)</td>
<td>(4.51)</td>
</tr>
<tr>
<td><strong>2nd Qtr GDP</strong></td>
<td>0.754</td>
<td>0.451</td>
</tr>
<tr>
<td></td>
<td>(0.248)</td>
<td>(0.161)</td>
</tr>
<tr>
<td><strong>June Polling</strong></td>
<td>0.147</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.085)</td>
<td></td>
</tr>
<tr>
<td><strong>Multiple R-Squared</strong></td>
<td>0.366</td>
<td>0.781</td>
</tr>
</tbody>
</table>

*Standard errors are in parentheses. N=18.*

Two questions to try to understand:

1. What do the coefficients (and standard errors) mean?
2. Why did the "2nd Quarter GDP" coefficient change?
## Beyond two dimensions

**Incumbent party vote share**

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>49.27</td>
<td>49.35</td>
</tr>
<tr>
<td></td>
<td>(1.35)</td>
<td>(4.51)</td>
</tr>
<tr>
<td>2nd Qtr GDP</td>
<td>0.754</td>
<td>0.451</td>
</tr>
<tr>
<td></td>
<td>(0.248)</td>
<td>(0.161)</td>
</tr>
<tr>
<td>June Polling</td>
<td>0.147</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.085)</td>
<td></td>
</tr>
<tr>
<td>Multiple R-Squared</td>
<td>0.366</td>
<td>0.781</td>
</tr>
</tbody>
</table>

*Standard errors are in parentheses. N=18.*

Two questions to try to understand:

- What do the coefficients (and standard errors) mean?
### Beyond two dimensions

**Incumbent party vote share**

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>49.27</td>
<td>49.35</td>
</tr>
<tr>
<td></td>
<td>(1.35)</td>
<td>(4.51)</td>
</tr>
<tr>
<td>2nd Qtr GDP</td>
<td>0.754</td>
<td>0.451</td>
</tr>
<tr>
<td></td>
<td>(0.248)</td>
<td>(0.161)</td>
</tr>
<tr>
<td>June Polling</td>
<td>0.147</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.085)</td>
<td></td>
</tr>
<tr>
<td>Multiple R-Squared</td>
<td>0.366</td>
<td>0.781</td>
</tr>
</tbody>
</table>

*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
Which can also be thought of like this
Overview: 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

\[ e_i = (Y_i - \hat{Y}_i) = (Y_i - \hat{\alpha} - \hat{\beta}X_i) \]
Residuals for multiple regression

```
2nd Quarter GDP

June approval

Vote share

Lecture 20 (QPM 2016)  Multivariate Regression  November 14, 2016  16 / 44
```
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:

\[ Y_i = \alpha + \beta_{1,i} X_{1,i} + \ldots + \beta_{k,i} X_{k,i} + \epsilon_i \]

\[ \epsilon_i \sim N(0, \sigma^2) \]

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

$$e_i = (Y_i - \hat{Y}_i) = (Y_i - \hat{\alpha} - \hat{\beta}_1X_{1i} - \ldots - \hat{\beta}_kX_{ki})$$
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**

$$e_i = (Y_i - \hat{Y}_i) = (Y_i - \hat{\alpha} - \hat{\beta}_1X_{1i} - \ldots - \hat{\beta}_kX_{ki})$$

**Sum of Squared Error**

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}_1X_{1i} - \ldots - \hat{\beta}_kX_{ki})^2$$
**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

\[ \hat{\sigma}^2 = \frac{SSE}{n-(k+1)} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-(k+1)} \Rightarrow \hat{\sigma} = \sqrt{\frac{SSE}{n-(k+1)}} = \sqrt{\frac{\sum(Y_i - \hat{Y}_i)^2}{n-(k+1)}} \]
Coefficient of multiple determination

Same as before:

- $TSS = \sum (Y_i - \bar{Y})^2$
Coefficient of multiple determination

Same as before:

- \( TSS = \sum (Y_i - \bar{Y})^2 \)
- \( SSE = \sum (Y_i - \hat{Y}_i)^2 \)
Coefficient of multiple determination

Same as before:

- $TSS = \sum (Y_i - \bar{Y})^2$
- $SSE = \sum (Y_i - \hat{Y}_i)^2$
Coefficient of multiple determination

Same as before:

- $TSS = \sum (Y_i - \bar{Y})^2$
- $SSE = \sum (Y_i - \hat{Y}_i)^2$

**R-squared**

$$R^2 = \frac{\text{Explained Variance}}{\text{Total Variance}} = \frac{\text{Total Variance} - \text{Unexplained Variance}}{\text{Total Variance}}$$
Coefficient of multiple determination

Same as before:

- $TSS = \sum (Y_i - \bar{Y})^2$
- $SSE = \sum (Y_i - \hat{Y}_i)^2$

**R-squared**

$$R^2 = \frac{\text{Explained Variance}}{\text{Total Variance}} = \frac{\text{Total Variance} - \text{Unexplained Variance}}{\text{Total Variance}}$$

$$= \frac{TSS - SSE}{TSS}$$
Overview: 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
### Beyond two dimensions

#### Incumbent party vote share

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>49.27</td>
<td>49.35</td>
</tr>
<tr>
<td></td>
<td>(1.35)</td>
<td>(4.51)</td>
</tr>
<tr>
<td>2nd Qtr GDP</td>
<td>0.754</td>
<td>0.451</td>
</tr>
<tr>
<td></td>
<td>(0.248)</td>
<td>(0.161)</td>
</tr>
<tr>
<td>June Polling</td>
<td>0.147</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.085)</td>
<td></td>
</tr>
<tr>
<td>Multiple R-Squared</td>
<td>0.366</td>
<td>0.781</td>
</tr>
</tbody>
</table>

*Standard errors are in parentheses. N=18.*
Conceptualization #1: Planes

![Graph showing the relationship between 2nd Quarter GDP, vote share, and June approval.](image-url)
Conceptualization #2: Lines within groups

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

$$Y_i = \alpha + \beta_1 X_i + \epsilon_i$$
Let $X_i$ take on a value of only 0 or 1.

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

$$E(Y_i|X_i = 0) = \alpha$$

$$E(Y_i|X_i = 1) = \alpha + \beta$$
Conceptualization #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)
Nominal data

\[ X_1 = \{ \text{Blue, Not blue} \}, \quad X_2 = \{ \text{Brown, Not brown} \} \]
Nominal data

\[ X_1 = \{\text{Blue, Not blue}\}, \quad X_2 = \{\text{Brown, Not brown}\} \]

\[ Y_i = \alpha + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i \]
Nominal data

\[ X_1 = \{\text{Blue, Not blue}\}, \quad X_2 = \{\text{Brown, Not brown}\} \]

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

\[ E(Y_i|\text{Blue}) = \alpha + \beta_1 \]
Nominal data

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

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

\[ E(Y_i|\text{Blue}) = \alpha + \beta_1 \]

\[ E(Y_i|\text{Brown}) = \alpha + \beta_2 \]
Nominal data

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

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

\[ E(Y_i | \text{Blue}) = \alpha + \beta_1 \]

\[ E(Y_i | \text{Brown}) = \alpha + \beta_2 \]

\[ E(Y_i | \text{Green}) = \alpha \]
Nominal data

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

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

\[ E(Y_i | \text{Blue}) = \alpha + \beta_1 \]

\[ E(Y_i | \text{Brown}) = \alpha + \beta_2 \]

\[ E(Y_i | \text{Green}) = \alpha \]
Ordinal data

\[ X = \{1, 2, 3, 4, 5\} \]

Treating ordinal variable as continuous

Treating ordinal variable as distinct categories
Example: 2009 health care poll

Support for Obama health care plan
CNN/ORC poll, September 11–13, 2009

Strongly oppose | Moderately oppose | Moderately support | Strongly support
--- | --- | --- | ---
0.4 | 0.2 | |
Example: 2009 health care poll
Example: 2009 health care poll

![Bar chart showing support for Obama health care plan by age groups](chart.png)
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:
\[
\text{HCR support} = \beta_0 + \beta_1 \text{ Party} + \beta_2 \text{ Age 30-44} + \beta_3 \text{ 45-64} + \beta_4 \text{ 65+}
\]
### Dummy variable regression results

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1.421</td>
<td>0.116</td>
</tr>
<tr>
<td>Party</td>
<td>0.914</td>
<td>0.031</td>
</tr>
<tr>
<td>Age 30-44</td>
<td>0.77</td>
<td>0.13</td>
</tr>
<tr>
<td>Age 45-64</td>
<td>-0.65</td>
<td>0.117</td>
</tr>
<tr>
<td>Age 65 +</td>
<td>-0.16</td>
<td>0.121</td>
</tr>
<tr>
<td>N</td>
<td>981</td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.4799</td>
<td></td>
</tr>
</tbody>
</table>

Lecture 20 (QPM 2016)
Multivariate Regression
November 14, 2016 30 / 44
Dummy variable regression

Support for Obama health care plan

- Strongly favor
- Moderately favor
- Moderately oppose
- Strongly oppose

- GOP
- Independent
- Democrat

- 18–29
- 30–44
- 45–64
- 65+

Lecture 20 (QPM 2016)
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.
Conceptualization #3: Added variable plots

Added-Variable Plots

```
vote | others
-10  -5  0   5
-5   0   5

q2gdp | others
-10  -5  0   5
-5   0   5

vote | others
-10  -5  0   10   20
-5   0   5
```

```
junetrial | others
-10  -5  0   10   20
-5   0   5
```
Thinking about X’s from this point of view

- You are going to be doing this for the homework.
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 \( \beta \) 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 $\beta$ estimates in the table.
- Adding additional controls positively correlated to both the vote-share and GDP growth will lower the $\beta$ 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 $\beta$ estimates in the table.
- Adding additional controls positively correlated to both the vote-share and GDP growth will lower the $\beta$ 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).
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 $\beta$ estimates in the table.
- Adding additional controls positively correlated to both the vote-share and GDP growth will lower the $\beta$ 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 $\beta$ estimates in the table.
- Adding additional controls positively correlated to both the vote-share and GDP growth will lower the $\beta$ 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.
Overview: 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

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

\( \beta_1 \): The coefficient for \( X_1 \).
Interpretation: A one unit increase \( X_1 \) leads to a \( \beta_1 \) increase in \( Y \) 
controlling for the independent effect of \( X_2 \).
Inference in regression coefficients

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

\[ \epsilon_i \sim N(0, \sigma^2) \]

\( \beta_1 \): The coefficient for \( X_1 \).
Interpretation: A one unit increase \( X_1 \) leads to a \( \beta_1 \) increase in \( Y \) \textit{controlling for} the independent effect of \( X_2 \).
Inference in regression coefficients

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

\[ \epsilon_i \sim N(0, \sigma^2) \]

\( \beta_1 \): The coefficient for \( X_1 \).

Interpretation: A one unit increase \( X_1 \) leads to a \( \beta_1 \) 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

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

\( \beta_1 \): The coefficient for \( X_1 \).

Interpretation: A one unit increase in \( X_1 \) leads to a \( \beta_1 \) 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
Inference in regression coefficients

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

\[ \epsilon_i \sim N(0, \sigma^2) \]

\( \beta_1 \): The coefficient for \( X_1 \).

Interpretation: A one unit increase \( X_1 \) leads to a \( \beta_1 \) 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.
Inference for the entire model

Before we said that $R^2$ is intuitively $R^2 = \frac{\text{Explained Variance}}{\text{Total Variance}}$. It makes sense then that $(1 - R^2)$ is the percent of variance we haven’t explained.
Inference for the entire model

Before we said that $R^2$ is intuitively $R^2 = \frac{\text{Explained Variance}}{\text{Total Variance}}$. It makes sense then that $(1 - R^2)$ is the percent of variance we haven’t explained. It turns out that:

\[
F = \frac{\frac{R^2}{k}}{\frac{(1 - R^2)}{[n - (k + 1)]}}
\]

Here $k$ is the number of covariates (gdp, incumbent, etc.), and $n$ is the number of observations. This will be distributed according to the F-distribution with $df_1 = k$, and $df_2 = n - (k + 1)$. 
Flashback: Interpreting the F-test

Is our model any good?

- This is a *formalized* way of asking whether our model is any good.
Flashback: Interpreting the F-test

Is our model any good?

- This is a *formalized* way of asking whether our model is any good.
- We compare the amount of variance explained by the regression to the amount unexplained.
Flashback: Interpreting the F-test

Is our model any good?

- This is a *formalized* way of asking whether our model is any good.
- We compare the amount of variance explained by the regression to the amount unexplained.
  - $H_0 : Y_i = \alpha + \epsilon_i$
Flashback: Interpreting the F-test

Is our model any good?

- This is a *formalized* way of asking whether our model is any good.
- We compare the amount of variance explained by the regression to the amount unexplained.
  - $H_0 : Y_i = \alpha + \epsilon_i$
  - $H_1 : Y_i = \alpha + X_{i1}\beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3 + \ldots + \epsilon_i$
Overview: 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

Simpson's Paradox

\[ E(Y) = 1 - 0.25X_1 + 2X_2 \]

- Relationship between \( X_1 \) and \( Y \) is the same across groups.
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

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

24537 respondents
Applied example: Income in presidential voting

![Graph showing the relationship between state income and Clinton vote percentage.](image-url)
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)
### Incumbent party vote share

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>49.27</td>
<td>49.35</td>
</tr>
<tr>
<td></td>
<td>(1.35)</td>
<td>(4.51)</td>
</tr>
<tr>
<td>2nd Qtr GDP</td>
<td>0.754</td>
<td>0.451</td>
</tr>
<tr>
<td></td>
<td>(0.248)</td>
<td>(0.161)</td>
</tr>
<tr>
<td>June Polling</td>
<td>0.147</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.085)</td>
<td></td>
</tr>
<tr>
<td>Multiple R-Squared</td>
<td>0.366</td>
<td>0.781</td>
</tr>
</tbody>
</table>

*Standard errors are in parentheses. N=18.*
Get in the game

<table>
<thead>
<tr>
<th>Incumbent party vote share</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>2nd Qtr GDP</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>June Polling</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Multiple R-Squared</td>
</tr>
</tbody>
</table>

Standard errors are in parentheses. N=18.

What is the predicted value for 2016? (2nd Quarter GDP = 1.4, June Approval=6)

Interpret Model 2

This should include a discussion of both substantive and statistical significance.