CSSS 569 • Visualizing Data

VISUAL DISPLAYS IN THE SOCIAL SCIENCES

Christopher Adolph

Department of Political Science
and
Center for Statistics and the Social Sciences
University of Washington, Seattle
Good visuals help social science researchers uncover patterns and relationships we’d otherwise miss.

Ever more sophisticated statistical models cry out for clear, easy-to-understand visual representations of model findings.

Casual observation suggests good visuals have a big impact on audiences for papers and job talks.

Puzzle: Social scientists seldom put as much care into designing visual displays as they devote to crafting effective prose.
### Plan of the Course

<table>
<thead>
<tr>
<th>Part I</th>
<th>Weeks</th>
<th>Principles of Effective Information Visualization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Cognitive Science and Visualization</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Graphical Programming in ( \mathbb{R} )</td>
</tr>
</tbody>
</table>

**Ideas**

<table>
<thead>
<tr>
<th>Part II</th>
<th>Weeks</th>
<th>Exploratory Data Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Visualizing Model Inference</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Visualizing Model Robustness and Interactions</td>
</tr>
</tbody>
</table>

**Tools**

<table>
<thead>
<tr>
<th>Part III</th>
<th>Weeks</th>
<th>Interactive Graphics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Tools for Scientific Writing and Presentations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Final Presentations</td>
</tr>
</tbody>
</table>
Three Examples of Visual Data Analysis

Success! Stopping infectious disease

Failure. The Challenger disaster

Confusion? Deciphering models of monetary policymaking

As we go, consider three uses of visuals:

- to explore data
- to understand model implications
- to test model fit
John Snow Saves London

Cholera outbreaks were common in 19th century London; 10,000s of deaths

Contemporary theories:

1. Cholera caused by “miasma” in the air coming from swamps…
2. Or a “poison” slowly losing strength as it passed from victim to victim…
3. London doctor John Snow thought contaminated water the cause

Outbreak in 1854: 500 deaths in 10 days in Soho

Snow collects real-time data; has Broad Street water pump handle removed

Did he stop the epidemic? And prove disease can be spread by germs?
How might a newspaper visual “analyze” John Snow’s intervention?

Overwhelming tendency to view time series data this way

A first step, but doesn’t help us make inferences about the data

The mortality data aren’t being compared to any other variables: time series plots aren’t models of the data generating process

Source: Tufte, Visual Explanations
Snow’s spatial analysis

In 1854, London water was provided by competing private firms
Each had its own network and reputation for cleanliness
Residents typically walked to the nearest street pump for water
Snow recorded the location of each death in real time
Placed these spatial data on a map along with the water pumps

Was one company’s network – or even just one pump – contaminated with cholera “germs”?
Deaths are concentrated around Broad Street pump, not other pumps.

Was it the source of the epidemic?

Or could this evidence be consistent with a different story?

How might the map be misleading? What does the map hide?

What other evidence would you like to have?
Fact: For any spot \( x \) on the map, there is a closest pump \( A \)

**Definition:** The set of all points \( x \) closest to pump \( A \) is the Voronoi cell of pump \( A \).
Fact: For any spot $x$ on the map, there is a closest pump $A$.

Definition: The set of all points $x$ closest to pump $A$ is the Voronoi cell of pump $A$. 
Fact: For any spot $x$ on the map, there is a closest pump $A$.

Definition: The set of all points $x$ closest to pump $A$ is the *Voronoi cell* of pump $A$.

Modeling Assumptions:
- Some (not all) pumps are contaminated.
- People use the closest pump.
Fact: For any spot \( x \) on the map, there is a closest pump \( A \)

Definition: The set of all points \( x \) closest to pump \( A \) is the Voronoi cell of pump \( A \)

Modeling Assumptions:
Some (not all) pumps are contaminated

People use the closest pump

Model prediction: Pattern of deaths should match Voronoi cell boundaries
Chris Adolph (University of Washington)
Problems?

Distance in a city isn’t really Euclidian – the built environment lengthens some paths.
Problems?

Distance in a city isn’t really Euclidian – the built environment lengthens some paths.

What about outliers? Can our theory be right if some cases lie outside Voronoi cell of Broad St. Pump?
Problems?

Distance in a city isn’t really Euclidian – the built environment lengthens some paths.

What about outliers? Can our theory be right if some cases lie outside Voronoi cell of Broad St. Pump?

Outliers could point to missing variables or simple randomness.
What explains outliers in this map?

Three cases:

1. A prison (work house) with its own well.
2. A brewery with its own water source. Saved by the beer.
3. Some distant deaths attributed to preference for Broad St. water.
John Snow stops the Cholera epidemic

Snow used his data and map to convince officials to remove the handle from the Broad Street pump.

Credited with stopping the outbreak & providing 1st experimental evidence for germs

Some questions to consider later:

1. Did the Broad Street Pump really cause the cholera outbreak?
2. Did removing the handle stop it?
3. How could Snow’s map be improved as a visual display of scientific information (VDSI)?

Steven Johnson (The Ghost Map, Riverhead Press, 2007) notes forensic evidence supporting Snow – final case occurred next to Broad St. Pump but didn’t spread
One way to think about data graphics is as a set of tools:

Maps, scatterplots, time series plots, histograms...

Another view: visuals can explore relationships and tell stories.
One way to think about data graphics is as a set of tools:

Maps, scatterplots, time series plots, histograms…

Another view: visuals can explore relationships and tell stories

Keys to visualizing relationships in data

1. consider a model or models linking different variables
Models, Maps, and Data Visualization

One way to think about data graphics is as a set of tools:

Maps, scatterplots, time series plots, histograms...

Another view: visuals can explore relationships and tell stories

Keys to visualizing relationships in data

1. consider a model or models linking different variables
2. find ways to usefully juxtapose variables within a visual display
Models, Maps, and Data Visualization

One way to think about data graphics is as a set of tools:
Maps, scatterplots, time series plots, histograms…

Another view: visuals can explore relationships and tell stories

Keys to visualizing relationships in data

1. consider a model or models linking different variables
2. find ways to usefully juxtapose variables within a visual display
3. be creative: move beyond rote applications of the most obvious tool
One way to think about data graphics is as a set of tools:

Maps, scatterplots, time series plots, histograms…

Another view: visuals can explore relationships and tell stories

Keys to visualizing relationships in data

1. consider a model or models linking different variables
2. find ways to usefully juxtapose variables within a visual display
3. be creative: move beyond rote applications of the most obvious tool
4. take care with details: annotations, use of color, scales, and so on
In 1986, the Challenger space shuttle exploded moments after liftoff.

The decision to launch is one of the most scrutinized in history.

Failure of O-rings in the solid-fuel rocket boosters blamed for explosion.

Could this failure have been foreseen?
The Challenger launch decision

<table>
<thead>
<tr>
<th>Flt Number</th>
<th>Temp (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>70</td>
</tr>
<tr>
<td>41b</td>
<td>57</td>
</tr>
<tr>
<td>41c</td>
<td>63</td>
</tr>
<tr>
<td>41d</td>
<td>70</td>
</tr>
<tr>
<td>51c</td>
<td>53</td>
</tr>
<tr>
<td>61a</td>
<td>79</td>
</tr>
<tr>
<td>61c</td>
<td>58</td>
</tr>
</tbody>
</table>

Engineers who made this table worried about launching below 53 degrees (Why?)

Data on O-ring failures at different launch temperatures, provided to NASA by Morton-Thiokol hours before launch
The Challenger launch decision

<table>
<thead>
<tr>
<th>Flt Number</th>
<th>Temp (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>70</td>
</tr>
<tr>
<td>41b</td>
<td>57</td>
</tr>
<tr>
<td>41c</td>
<td>63</td>
</tr>
<tr>
<td>41d</td>
<td>70</td>
</tr>
<tr>
<td>51c</td>
<td>53</td>
</tr>
<tr>
<td>61a</td>
<td>79</td>
</tr>
<tr>
<td>61c</td>
<td>58</td>
</tr>
</tbody>
</table>

Engineers who made this table worried about launching below 53 degrees (Why?)

Physical problem: O-ring would erode or “blow-by” in cold temperatures

Data on O-ring failures at different launch temperatures, provided to NASA by Morton-Thiokol hours before launch.
The Challenger launch decision

<table>
<thead>
<tr>
<th>Flt Number</th>
<th>Temp (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>70</td>
</tr>
<tr>
<td>41b</td>
<td>57</td>
</tr>
<tr>
<td>41c</td>
<td>63</td>
</tr>
<tr>
<td>41d</td>
<td>70</td>
</tr>
<tr>
<td>51c</td>
<td>53</td>
</tr>
<tr>
<td>61a</td>
<td>79</td>
</tr>
<tr>
<td>61c</td>
<td>58</td>
</tr>
</tbody>
</table>

Data on O-ring failures at different launch temperatures, provided to NASA by Morton-Thiokol hours before launch.

Engineers who made this table worried about launching below 53 degrees (Why?)

Physical problem: O-ring would erode or “blow-by” in cold temperatures

Failed to convince administrators of danger

Counter-argument:
“damages at low and high temps”

Are there problems with this presentation? With the use of data?
### Flights with O-ring damage

<table>
<thead>
<tr>
<th>Flt Number</th>
<th>Temp (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>70</td>
</tr>
<tr>
<td>41b</td>
<td>57</td>
</tr>
<tr>
<td>41c</td>
<td>63</td>
</tr>
<tr>
<td>41d</td>
<td>70</td>
</tr>
<tr>
<td>51c</td>
<td>53</td>
</tr>
<tr>
<td>61a</td>
<td>79</td>
</tr>
<tr>
<td>61c</td>
<td>58</td>
</tr>
</tbody>
</table>

Data on O-ring failures at different launch temperatures, provided to NASA by Morton-Thiokol hours before launch.

Are there problems with this presentation? With the use of data?
The Challenger launch decision

Flights with O-ring damage

<table>
<thead>
<tr>
<th>Flt Number</th>
<th>Temp (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>70</td>
</tr>
<tr>
<td>41b</td>
<td>57</td>
</tr>
<tr>
<td>41c</td>
<td>63</td>
</tr>
<tr>
<td>41d</td>
<td>70</td>
</tr>
<tr>
<td>51c</td>
<td>53</td>
</tr>
<tr>
<td>61a</td>
<td>79</td>
</tr>
<tr>
<td>61c</td>
<td>58</td>
</tr>
</tbody>
</table>

Data on O-ring failures at different launch temperatures, provided to NASA by Morton-Thiokol hours before launch

Are there problems with this presentation? With the use of data?

Selection on the dependent variable

Why sort by launch number?
The Challenger launch decision

<table>
<thead>
<tr>
<th>Damage?</th>
<th>Temp (F)</th>
<th>Damage?</th>
<th>Temp (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>53</td>
<td>Yes</td>
<td>70</td>
</tr>
<tr>
<td>Yes</td>
<td>57</td>
<td>No</td>
<td>70</td>
</tr>
<tr>
<td>Yes</td>
<td>58</td>
<td>No</td>
<td>70</td>
</tr>
<tr>
<td>Yes</td>
<td>63</td>
<td>No</td>
<td>72</td>
</tr>
<tr>
<td>No</td>
<td>66</td>
<td>No</td>
<td>73</td>
</tr>
<tr>
<td>No</td>
<td>67</td>
<td>No</td>
<td>75</td>
</tr>
<tr>
<td>No</td>
<td>67</td>
<td>No</td>
<td>76</td>
</tr>
<tr>
<td>No</td>
<td>68</td>
<td>No</td>
<td>78</td>
</tr>
<tr>
<td>No</td>
<td>69</td>
<td>Yes</td>
<td>79</td>
</tr>
<tr>
<td>Yes</td>
<td>70</td>
<td>No</td>
<td>81</td>
</tr>
</tbody>
</table>

After fixing these two problems, the evidence begins to speak for itself. What if Morton-Thiokol engineers had made this table before the launch?
Why didn’t NASA make the right decision?

Many answers in the literature:
bureaucratic politics; group think; bounded rationality, etc.

But Edward Tufte thinks it may have been a matter of presentation & modeling:

- Never made the right tables or graphics
- Selected only failure data
- Never considered a simple statistical model
The Challenger launch decision

What Morton-Thiokol presented months after the disaster
What Morton-Thiokol presented months after the disaster

A marvel of poor design – obscures the data, makes analysis harder
The Challenger launch decision

What Morton-Thiokol presented months after the disaster

A marvel of poor design – obscures the data, makes analysis harder

Can methods commonly used in social science do better?
What was the forecast temperature for launch?
26 to 29 degrees Fahrenheit (−2 to −3 degrees C)!

The shuttle was launched in unprecedented cold
Imagine you are the analyst making the launch recommendation. You’ve made the scatterplot above. What would you add to it? Put another way, what do you is the first question you expect to hear?
"What’s the chance of failure at 26 degrees?"
The scatterplot suggests the answer is “high,” but that’s vague.
But what if the next launch is at 58 degrees? Or 67 degrees?
We need a probability model and a way to convey that model to the public.
Let’s estimate a simple logit model of damage as a function of temperature:

\[ Pr(Damage|Temp) = \logit^{-1}(\hat{\beta}_0 + \hat{\beta}_1 Temp) \]

\[ \logit^{-1}(\hat{\beta}_0 + \hat{\beta}_1 Temp) \]

\[ R \] gives us this lovely logit output…
Let’s estimate a simple logit model of damage as a function of temperature:

\[
Pr(Damage|Temp) = \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 \text{Temp})
\]

R gives us this lovely logit output…

<table>
<thead>
<tr>
<th>Variable</th>
<th>est.</th>
<th>s.e.</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature (F)</td>
<td>-0.18</td>
<td>0.09</td>
<td>0.047</td>
</tr>
<tr>
<td>Constant</td>
<td>11.9</td>
<td>6.34</td>
<td>0.062</td>
</tr>
<tr>
<td>N</td>
<td>22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>log-likelihood</td>
<td>-10.9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Let’s estimate a simple logit model of damage as a function of temperature:

$$Pr(Damage | Temp) = \logit^{-1}(\hat{\beta}_0 + \hat{\beta}_1 Temp)$$

R gives us this lovely logit output…

<table>
<thead>
<tr>
<th>Variable</th>
<th>est.</th>
<th>s.e.</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature (F)</td>
<td>-0.18</td>
<td>0.09</td>
<td>0.047</td>
</tr>
<tr>
<td>Constant</td>
<td>11.9</td>
<td>6.34</td>
<td>0.062</td>
</tr>
<tr>
<td>N</td>
<td>22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>log-likelihood</td>
<td>-10.9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

…which most social scientists read as “a significant negative relationship b/w temperature and probability of damage”

…but that’s pretty vague too

Is there a more persuasive/clear/useful way to present these results?
A picture of the logistic regression shows model predictions and uncertainty.
A picture of the logistic regression shows model predictions and uncertainty.
A picture of the logistic regression shows model predictions and uncertainty.
…and gives a more precise sense of how reckless it was to launch at 29 F
When possible, it’s good to show the data giving rise to the model.
Remembering that the **Failures** are only meaningful compared to **Successes**
Looking only at the data we might think launches < 66 F are “certain failures”

This inference is based on an unstated (and flawed) model
The estimated logit model should give us pause

There is a significant risk of failure across the board
What is an acceptable risk of O-ring failure?

Was the shuttle safe at any temperature?
In a hearing, Richard Feynmann dramatically showed O-rings lose resilience when cold by dropping one in his ice water.

Experiment cut through weeks of technical gibberish concealing flaws in the O-ring

But it shouldn’t have taken a Nobel laureate:

a scientist with a year of statistical training could’ve used the launch record to reach the same conclusion

And it would take no more than a single graphic to show the result

Visualizing relationships is critical at every stage of statistical analysis
The Challenger launch decision

Lessons for social scientists:

Even relatively simple models and data are easier to understand with visuals

Tables can hide strong correlations

Imagine what might be hiding in datasets with dozens of variables?

Or in models with complex functional forms?

Visuals help make discussion more substantive

See the size of the effect, not just the sign

Make relative judgments of the importance of covariates

Make measured assessments of uncertainty – not just “accept/reject”
Coefficients are not enough

Some limits of typical presentations of statistical results:

- Everything written in terms of arcane intermediate quantities (for most people, this includes logit coefficients)
- Little effort to transform results to the scale of the quantities of interest → really want the conditional expectation, $E(y|x)$
- Little effort to make informative statements about estimation uncertainty → really want to know how uncertain is $E(y|x)$
- Little visualization at all, or graphs with low data-ink ratios
American Interest Rate Policy

Example from my own work on central banking (Bankers, Bureaucrats, and Central Bank Politics, Cambridge U.P., 2013, Ch. 4)

Federal Reserve Open Market Committee (FOMC) sets interest rates
10×/year

Members of the FOMC vote on the Chair’s proposed interest rate

Dissenting voters signal whether they would like a higher or lower rate

Dissents are rare but may be symptomatic of how the actual rate gets chosen
American Interest Rate Policy

Example from my own work on central banking (Bankers, Bureaucrats, and Central Bank Politics, Cambridge U.P., 2013, Ch. 4)

Federal Reserve Open Market Committee (FOMC) sets interest rates 10×/year

Members of the FOMC vote on the Chair’s proposed interest rate

Dissenting voters signal whether they would like a higher or lower rate

Dissents are rare but may be symptomatic of how the actual rate gets chosen

Many factors could influence interest rate votes:

Individual
- Career background
- Appointing party
- Interactions of above

Economy
- Expected inflation
- Expected unemployment

Politics
- Election cycles
American Interest Rate Policy

My main concern is the individual determinants, especially career background

I measure career background as a composite variable

Fractions of career spent in each of 5 categories:

- Financial Sector: FinExp
- Treasury Department: FMExp
- Federal Reserve: CBExp
- Other Government: GovExp
- Academic Economics: EcoExp

These 5 categories plus an (omitted) “Other” must sum to 1.0
American Interest Rate Policy

<table>
<thead>
<tr>
<th>Initial Composition</th>
<th>Hypothetical New Composition</th>
</tr>
</thead>
<tbody>
<tr>
<td>FinExp 0.1</td>
<td>FinExp 0.250</td>
</tr>
<tr>
<td>GovExp 0.3</td>
<td>= 0.15</td>
</tr>
<tr>
<td>FMExp 0.1</td>
<td></td>
</tr>
<tr>
<td>CBExp 0.2</td>
<td></td>
</tr>
<tr>
<td>EcoExp 0.3</td>
<td></td>
</tr>
<tr>
<td>Sum 1.0</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Because of the composition constraint, to consider the effects of a change in one category, we must adjust the other categories simultaneously.

What happens if I increase FinExp by 0.15, but keep all other components the same?
What happens if I increase FinExp by 0.15, but keep all other components the same?

Note – this is close to what I assume when I interpret the $\beta$ for a component as the “effect” of raising that component.
American Interest Rate Policy

<table>
<thead>
<tr>
<th></th>
<th>Initial Composition</th>
<th>Hypothetical New Composition</th>
</tr>
</thead>
<tbody>
<tr>
<td>FinExp</td>
<td>0.1</td>
<td>ΔFinExp</td>
</tr>
<tr>
<td>GovExp</td>
<td>0.3</td>
<td>=0.15</td>
</tr>
<tr>
<td>FMExp</td>
<td>0.1</td>
<td>→</td>
</tr>
<tr>
<td>CBExp</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>EcoExp</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td>1.0</td>
<td></td>
</tr>
</tbody>
</table>

Increasing one component without lowering the combined total of the other components by the same amount leads to a logical fallacy – a career that has 115% total experience!
### American Interest Rate Policy

<table>
<thead>
<tr>
<th></th>
<th>Initial Composition</th>
<th>Hypothetical New Composition</th>
</tr>
</thead>
<tbody>
<tr>
<td>FinExp</td>
<td>0.1</td>
<td>ΔFinExp 0.250</td>
</tr>
<tr>
<td>GovExp</td>
<td>0.3</td>
<td>=0.15 0.300</td>
</tr>
<tr>
<td>FMExp</td>
<td>0.1</td>
<td>→ 0.100</td>
</tr>
<tr>
<td>CBExp</td>
<td>0.2</td>
<td>→ 0.200</td>
</tr>
<tr>
<td>EcoExp</td>
<td>0.3</td>
<td>→ 0.150</td>
</tr>
<tr>
<td>Sum</td>
<td>1.0</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Alternatively, if we left out a category (say, EcoExp) as a “reference,” we would be implicitly assuming that category alone shrinks to accommodate the increase in FinExp. But that blends the effects of FinExp and EcoExp – so that in our model, the choice of reference category is no longer harmless!
And what if EcoExp (still the reference category) starts out smaller than 0.15?

Then our counterfactual would create negative career components!

<table>
<thead>
<tr>
<th></th>
<th>Initial Composition</th>
<th>Hypothetical New Composition</th>
</tr>
</thead>
<tbody>
<tr>
<td>FinExp</td>
<td>0.1</td>
<td>ΔFinExp</td>
</tr>
<tr>
<td>GovExp</td>
<td>0.3</td>
<td>=0.15</td>
</tr>
<tr>
<td>FMExp</td>
<td>0.1</td>
<td>→</td>
</tr>
<tr>
<td>CBExp</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>EcoExp</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td>1.0</td>
<td></td>
</tr>
</tbody>
</table>
American Interest Rate Policy

<table>
<thead>
<tr>
<th>Initial Composition</th>
<th>Hypothetical New Composition</th>
</tr>
</thead>
<tbody>
<tr>
<td>FinExp 0.1</td>
<td>ΔFinExp 0.250</td>
</tr>
<tr>
<td>GovExp 0.3</td>
<td>= 0.15</td>
</tr>
<tr>
<td>FMExp 0.1</td>
<td></td>
</tr>
<tr>
<td>CBExp 0.2</td>
<td></td>
</tr>
<tr>
<td>EcoExp 0.3</td>
<td></td>
</tr>
<tr>
<td>Sum 1.0</td>
<td>1.000</td>
</tr>
</tbody>
</table>

When covariates form a composition, we have two problems:

1. to avoid blending effects across components
2. to avoid impossible counterfactuals

I recommend *ratio-preserving counterfactuals*, which uniquely solve both problems.
### American Interest Rate Policy

<table>
<thead>
<tr>
<th>Initial Composition</th>
<th>Hypothetical New Composition</th>
</tr>
</thead>
<tbody>
<tr>
<td>FinExp 0.1</td>
<td>ΔFinExp 0.250</td>
</tr>
<tr>
<td>GovExp 0.3</td>
<td>=0.15 0.250</td>
</tr>
<tr>
<td>FMEExp 0.1</td>
<td>→ 0.083</td>
</tr>
<tr>
<td>CBExp 0.2</td>
<td>→ 0.167</td>
</tr>
<tr>
<td>EcoExp 0.3</td>
<td>→ 0.250</td>
</tr>
<tr>
<td>Sum 1.0</td>
<td>1.000</td>
</tr>
</tbody>
</table>

The transformations above uniquely preserve the ratios among all categories (except FinExp, of course)

Note that now, the effect of a change in one category works through all the $\beta$s for the composition
American Interest Rate Policy

We’ll fit an ordered probit model to the interest rate data:

\[
\begin{align*}
\Pr(Y_i = \text{ease}|\hat{\beta}, \hat{\tau}) &= \Phi \left(0|X_i\hat{\beta}, 1\right) \\
\Pr(Y_i = \text{assent}|\hat{\beta}, \hat{\tau}) &= \Phi \left(\hat{\tau}|X_i\hat{\beta}, 1\right) - \Phi \left(0|X_i\hat{\beta}, 1\right) \\
\Pr(Y_i = \text{tighten}|\hat{\beta}, \hat{\tau}) &= 1 - \Phi \left(\hat{\tau}|X_i\hat{\beta}, 1\right)
\end{align*}
\]

where \(\Phi\) represents the Normal CDF and \(\tau\) is a cutpoint

(don’t worry if this model is unfamiliar; suffice it to say we have a nonlinear model and not just linear regression)
Running the model yields the following estimates:

<table>
<thead>
<tr>
<th>EVs</th>
<th>param.</th>
<th>s.e.</th>
<th>EVs</th>
<th>param.</th>
<th>s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FinExp</td>
<td>−0.021</td>
<td>(0.146)</td>
<td>E(Inflation)</td>
<td>0.019</td>
<td>(0.015)</td>
</tr>
<tr>
<td>GovExp</td>
<td>−0.753</td>
<td>(0.188)</td>
<td>E(Unemployment)</td>
<td>−0.035</td>
<td>(0.022)</td>
</tr>
<tr>
<td>FMExp</td>
<td>−1.039</td>
<td>(0.324)</td>
<td>In-Party, election year</td>
<td>−0.182</td>
<td>(0.103)</td>
</tr>
<tr>
<td>CBExp</td>
<td>−0.142</td>
<td>(0.141)</td>
<td>Republican</td>
<td>−0.485</td>
<td>(0.102)</td>
</tr>
<tr>
<td>EcoExp × Repub</td>
<td>0.934</td>
<td>(0.281)</td>
<td>Constant</td>
<td>2.490</td>
<td>(0.148)</td>
</tr>
<tr>
<td>EcoExp × Dem</td>
<td>−0.826</td>
<td>(0.202)</td>
<td>Cutpoint (τ)</td>
<td>3.745</td>
<td>(0.067)</td>
</tr>
</tbody>
</table>

| N                  | 2957   |       | ln likelihood                 | −871.68|       |

**Table 1: Problematic presentation: FOMC member dissenting votes—Ordered probit parameters.** Estimated ordered probit parameters, with standard errors in parentheses, from the regression of a \( j = 3 \) category variable on a set of explanatory variables (EVs). Although such nonlinear models are often summarized by tables like this one, especially in the social sciences, it is difficult to discern the effects of the EVs listed at right on the probability of each of the \( j \) outcomes. Because the career variables XXXExp are logically constrained to a unit sum, even some of the signs are misleading. The usual quantities of interest for an ordered probit model are not the parameters (\( \beta \) and \( \tau \)), but estimates of \( \Pr(y_j|\mathbf{x}_c, \beta, \tau) \) for hypothetical levels of the EVs \( \mathbf{x}_c \), which I plot in Figure 1.
### American Interest Rate Policy

**Response variable: FOMC Votes (1 = ease, 2 = accept, 3 = tighten)**

<table>
<thead>
<tr>
<th>EVs</th>
<th>param.</th>
<th>s.e.</th>
<th>EVs</th>
<th>param.</th>
<th>s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FinExp</td>
<td>-0.021</td>
<td>(0.146)</td>
<td>E(Inflation)</td>
<td>0.019</td>
<td>(0.015)</td>
</tr>
<tr>
<td>GovExp</td>
<td>-0.753</td>
<td>(0.188)</td>
<td>E(Unemployment)</td>
<td>-0.035</td>
<td>(0.022)</td>
</tr>
<tr>
<td>FMExp</td>
<td>-1.039</td>
<td>(0.324)</td>
<td>In-Party, election year</td>
<td>-0.182</td>
<td>(0.103)</td>
</tr>
<tr>
<td>CBExp</td>
<td>-0.142</td>
<td>(0.141)</td>
<td>Republican</td>
<td>-0.485</td>
<td>(0.102)</td>
</tr>
<tr>
<td>EcoExp × Repub</td>
<td>0.934</td>
<td>(0.281)</td>
<td>Constant</td>
<td>2.490</td>
<td>(0.148)</td>
</tr>
<tr>
<td>EcoExp × Dem</td>
<td>-0.826</td>
<td>(0.202)</td>
<td>Cutpoint (τ)</td>
<td>3.745</td>
<td>(0.067)</td>
</tr>
</tbody>
</table>

| N                    | 2957   | ln likelihood | -871.68 |

**How do we interpret these results?**
American Interest Rate Policy

Because the model is non-linear, interpreting coefficients as slopes ($\partial y / \partial \beta$) is grossly misleading

Moreover, the compositional variables are tricky:
If one goes up, the others must go down, to keep the sum = 1

Finally, we can’t interpret interactive coefficients separately

<table>
<thead>
<tr>
<th>EVs</th>
<th>param.</th>
<th>s.e.</th>
<th>EVs</th>
<th>param.</th>
<th>s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FinExp</td>
<td>-0.021</td>
<td>(0.146)</td>
<td>E(Inflation)</td>
<td>0.019</td>
<td>(0.015)</td>
</tr>
<tr>
<td>GovExp</td>
<td>-0.753</td>
<td>(0.188)</td>
<td>E(Unemployment)</td>
<td>-0.035</td>
<td>(0.022)</td>
</tr>
<tr>
<td>FMExp</td>
<td>-1.039</td>
<td>(0.324)</td>
<td>In-Party, election year</td>
<td>-0.182</td>
<td>(0.103)</td>
</tr>
<tr>
<td>CBExp</td>
<td>-0.142</td>
<td>(0.141)</td>
<td>Republican</td>
<td>-0.485</td>
<td>(0.102)</td>
</tr>
<tr>
<td>EcoExp $\times$ Repub</td>
<td>0.934</td>
<td>(0.281)</td>
<td>Constant</td>
<td>2.490</td>
<td>(0.148)</td>
</tr>
<tr>
<td>EcoExp $\times$ Dem</td>
<td>-0.826</td>
<td>(0.202)</td>
<td>Cutpoint ($\tau$)</td>
<td>3.745</td>
<td>(0.067)</td>
</tr>
</tbody>
</table>

N 2957 ln likelihood $-871.68$
Looking at this table, two obvious questions arise:

What is the effect of each covariate on the probability of each kind of vote?

What are the confidence intervals or standard errors for those effects?
Cruel to leave this to the reader: it’s a lot of work to figure out.

The table above, though conventional, is an intermediate step.

Publishing the table alone is like stopping where Morton-Thiokol did, with pages of technical gibberish – the answers are there, but buried
As the researcher, I should calculate the effects and uncertainty and present them in a readable way. A single graphic achieves both goals.
American Interest Rate Policy

My final graphic will involve small multiples, but explanation should start with a single example.

“The average central banker dissents in favor of tighter interest rates 4% of the time. In contrast, former treasury officials in the FOMC dissent 0.6% of the time, with a 95% CI from 0.05% to 2%.”
American Interest Rate Policy

“Other former bureaucrats issue hawkish dissents 1% of the time [95% CI: 0.5 to 2.0], all else equal.”
Now that readers understand how to read an individual result, they are ready to explore the graphic on their own.

I can highlight broad trends, then summarize the key findings.

But starting by explaining a single instance is critical for effectively using small multiples.
American Interest Rate Policy

Response to an Increase in...

FMExp
GovExp
EcoExp × Dem
Republican
In-Party & Election
E(Unemployment)
E(Inflation)
CBExp
FinExp
EcoExp × Repub

Probability of hawkish dissent

Probability of dovish dissent

Change in \( P(\text{hawkish dissent}) \)

Change in \( P(\text{dovish dissent}) \)
No matter how complex the model, you can always summarize relationships among variables with pictures.

Well designed VDSIs make complex models (linear or nonlinear) transparent.

For example, for a regression-like model, you might calculate $E(y|x_c, \hat{\beta})$ for interesting cases $x_c$, and then plot many such quantities for comparison.
Coming later in the course...

No matter how complex the model, you can always summarize relationships among variables with pictures.

Well designed VDSIs make complex models (linear or nonlinear) transparent.

For example, for a regression-like model, you might calculate $E(y|x_c, \hat{\beta})$ for interesting cases $x_c$, and then plot many such quantities for comparison.

This works no matter how complex your model is –

Any intelligent non-specialist should be able to understand your (simple/fancy/Bayesian/dynamic/hierarchical/nonlinear/interactive) model.

If they can’t, you’re not finished writing it up, and may be missing some implications yourself!
Scope of the Course

Because visual displays can be woven throughout all empirical science, it may sound like this course covers all of applied statistics.

Course goal: complement your other statistical training.

Start by defining Visual Displays of Scientific Information & their uses.
What is a VDSI?

Almost any representation of information is a VDSI – not just graphics:

- A plot
- A table
- A confection of plots and/or tables
- A schematic
- An equation
- A paragraph
- A movie
- An interactive display
When do we use VDSIs?

VDSIs are woven through the practice of quantitative methods:

- Exploring data
- Interpreting models
- Checking model assumptions & fit
- Persuading an audience
- Making a result memorable
How do VDSIs convey information?

VDSIs can present massive amounts of data for different ends:

- for lookup
- for posterity
- for gestalt impressions
- for exploration
- for rigorous comparison

The appropriate visuals vary by task
Who uses the VDSIs the researcher designs?

- The researcher herself
- The expert reader
- Decision makers
- The general public

Different VDSIs may be best suited for each audience
So how do I choose?

*Some VDSIs are generical well suited to some tasks*

Tables are usually good for lookup, bad for gestalt impressions

For fun, type `?pie` in R.

Designing good visuals is more than “Pie charts bad; Dot charts good”

Some VDSIs will be more powerful than others for a particular purpose

Be creative – different visuals can solve the same problem in usefully different ways
So how do I choose?

Some VDSIs are generical well suited to some tasks

Tables are usually good for lookup, bad for gestalt impressions

Some VDSIs are inherently powerful

Scatterplots show relations between two continuous variables richly and simply, and will never be bettered
So how do I choose?

Some VDSIs are generical well suited to some tasks

Tables are usually good for lookup, bad for gestalt impressions

Some VDSIs are inherently powerful

Scatterplots show relations between two continuous variables richly and simply, and will never be bettered

Some VDSIs are inherently inferior

Pie charts are inefficient, awkward, and prone to misinterpretation

For fun, type `?pie` in R.
So how do I choose?

Some VDSIs are generical well suited to some tasks

Tables are usually good for lookup, bad for gestalt impressions

Some VDSIs are inherently powerful

Scatterplots show relations between two continuous variables richly and simply, and will never be bettered

Some VDSIs are inherently inferior

Pie charts are inefficient, awkward, and prone to misinterpretation

For fun, type \texttt{?pie} in \texttt{R}.

Designing good visuals is more than “Pie charts bad; Dot charts good”

Some VDSIs will be more powerful than others for a particular purpose

Be creative – different visuals can solve the same problem in usefully different ways