Statistical Modeling: Intro and Applications (or: What else is there?)

Intermediate Social Statistics

Week 8 (7 March 2017)

Andy Eggers

We've seen:

- Regression (OLS)
- RCTs
- Matching
- Instrumental variables
- RDD
- Diff-in-diff/panel

You also saw:

- Logistic regression

We’ve seen:

- Regression (OLS)
- RCTs
- Matching
- Instrumental variables
- RDD
- Diff-in-diff/panel

What else do we need?

You also saw:

- Logistic regression

Conventional approach: tour of Generalized Linear Models (GLMs) via "range matching"

Conventional approach: tour of Generalized Linear Models (GLMs) via "range matching"

If your dependent variable is …	…you need this model.	See this Stata command.
Continuous (and unbounded)	OLS	regress
Binary (e.g. join WTO or not)	Logit Probit	logit probit
A count (e.g. 0, 1, 10 wars)	Poisson Negative binomial	poisson nbreg
Ordered categories (e.g. "opposed", "neutral", in favor")	Ordinal logit Ordinal probit	ologit oprobit
Non-ordered categories (e.g. Tory, Labour, Lib Dem; Christian, Muslim, Jewish, atheist)	Multinomial logit, conditional logit	mlogit clogit
A measure of survival or duration (e.g. cabinet or war duration)	Survival or hazard model	stcox

Conventional approach: tour of Generalized Linear Models (GLMs) via "range matching"

If your dependent variable is …	…you need this model.	See this Stata command.
Continuous (and unbounded)	OLS	regress
Binary (e.g. join WTO or not)	Logit Probit	logit probit
A count (e.g. 0, 1, 10 wars)	Poisson Negative binomial	poisson nbreg
Ordered categories (e.g. "opposed", "neutral", in favor")	Ordinal logit Ordinal probit	ologit oprobit
Non-ordered categories (e.g. Tory, Labour, Lib Dem; Christian, Muslim, Jewish, atheist)	Multinomial logit, conditional logit	mlogit clogit
A measure of survival or duration (e.g. cabinet or war duration)	Survival or hazard model	stcox

See glm (generalized linear model) package for many of these.

Generalized linear models

Linear regression model:

$$E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Binary logistic models:

$$\log\left[\frac{P(Y=1)}{P(Y=0)}\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Multinomial logistic models:

$$\log\left[\frac{P(Y=j)}{P(Y=0)}\right] = \alpha_j + \beta_{j1} X_1 + \beta_{j2} X_2 + \cdots + \beta_{jk} X_k$$

Ordinal logistic models:

$$\log\left[\frac{P(Y \geq j)}{P(Y < j)}\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Count models:

$$\log\left[E(Y)\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Generalized linear models

Gailmard p. 146: "invertible function of the model parameter is expressed as a linear function of the covariate(s)"

Linear regression model:

$$\mathrm{E}(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Binary logistic models:

$$\log\left[\frac{P(Y=1)}{P(Y=0)}\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Multinomial logistic models:

$$\log\left[\frac{P(Y=j)}{P(Y=0)}\right] = \alpha_j + \beta_{j1} X_1 + \beta_{j2} X_2 + \cdots + \beta_{jk} X_k$$

Ordinal logistic models:

$$\log\left[\frac{P(Y \geq j)}{P(Y < j)}\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Count models:

$$\log[\mathrm{E}(Y)] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

What you need to know about GLMs

What you need to know about GLMs

- Syntax (trivial)

  Stata: `[model name] [outcome] [covariates], [options]`

What you need to know about GLMs

- Syntax (trivial)

  Stata: `[model name] [outcome] [covariates], [options]`
- Interpretation (not trivial)

What you need to know about GLMs

- Syntax (trivial)

  Stata: `[model name] [outcome] [covariates], [options]`
- Interpretation (not trivial)

Think about *what your model is supposed to help you understand* **(quantities of interest)**.

What you need to know about GLMs

- Syntax (trivial)

  Stata: `[model name] [outcome] [covariates], [options]`
- Interpretation (not trivial)

Think about *what your model is supposed to help you understand* **(quantities of interest)**.

Especially with GLMs, this is usually not (quite) a regression coefficient.

What you need to know about GLMs

- Syntax (trivial)

  Stata: `[model name] [outcome] [covariates], [options]`
- Interpretation (not trivial)

Think about *what your model is supposed to help you understand* **(quantities of interest)**.

Especially with GLMs, this is usually not (quite) a regression coefficient.

See lab.

What you need to know about GLMs

- Syntax (trivial)

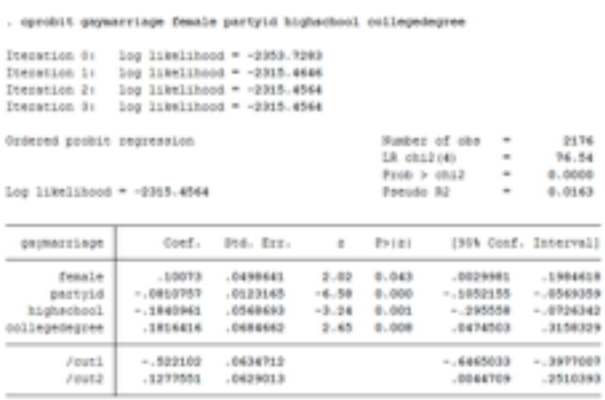
  Stata: `[model name] [outcome] [covariates], [options]`

- Interpretation (not trivial)

Think about *what your model is supposed to help you understand* **(quantities of interest)**.

Especially with GLMs, this is usually not (quite) a regression coefficient.

See lab.



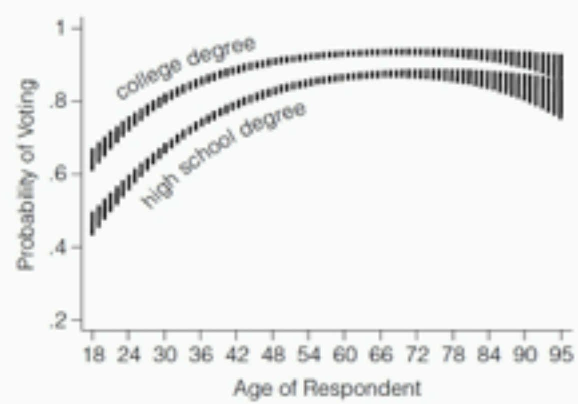
```
. oprobit gaymarriage female partyid highschool collegedegree

Iteration 0:   log likelihood = -2353.7283
Iteration 1:   log likelihood = -2315.4646
Iteration 2:   log likelihood = -2315.4564
Iteration 3:   log likelihood = -2315.4564

Ordered probit regression                         Number of obs   =       2176
                                                  LR chi2(4)      =      76.54
                                                  Prob > chi2     =     0.0000
Log likelihood = -2315.4564                       Pseudo R2       =     0.0163

------------------------------------------------------------------------------
 gaymarriage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |     .10073   .0498641     2.02   0.043     .0029981    .1984618
     partyid |  -.0810757   .0123165    -6.58   0.000    -.1052155   -.0569359
  highschool |  -.1840961   .0568693    -3.24   0.001     -.295558   -.0726342
collegedegree|   .1816416   .0684662     2.65   0.008     .0474503    .3158329
-------------+----------------------------------------------------------------
       /cut1 |   -.522102   .0634712                     -.6465033   -.3977007
       /cut2 |   .1277551   .0629013                      .0044709    .2510393
------------------------------------------------------------------------------
```

FIGURE 1 Probability of Voting by Age



Vertical bars indicate 99-percent confidence intervals

This week

This week

- **Lab**: intuition and practice with GLMs in Stata

This week

- **Lab**: intuition and practice with GLMs in Stata
- **This lecture:**

This week

- **Lab**: intuition and practice with GLMs in Stata
- **This lecture:**
  - Why I think OLS is enough for estimating treatment effects (and many other tasks)

This week

- **Lab**: intuition and practice with GLMs in Stata
- **This lecture:**
  - Why I think OLS is enough for estimating treatment effects (and many other tasks)
  - When statistical modeling might be more useful

This week

- **Lab**: intuition and practice with GLMs in Stata
- **This lecture:**
  - Why I think OLS is enough for estimating treatment effects (and many other tasks)
  - When statistical modeling might be more useful
  - Introduction to statistical models based on MLE

Ordinal probit application: Hainmueller and Hiscox 2010

Two economic explanations for (variation in) anti-immigrant sentiment:

- **Labor market competition** → natives should oppose immigrants with skill levels similar to their own
- **Fiscal burden** → rich natives should be more opposed to low-skilled immigrants than poor natives (especially where immigrants use a lot of public services)

(Random whether respondent gets A or B)

Hainmueller and Hiscox ask a sample of US respondents either

A. Do you agree or disagree that the US should allow more **highly skilled immigrants** from other countries to come and live here?

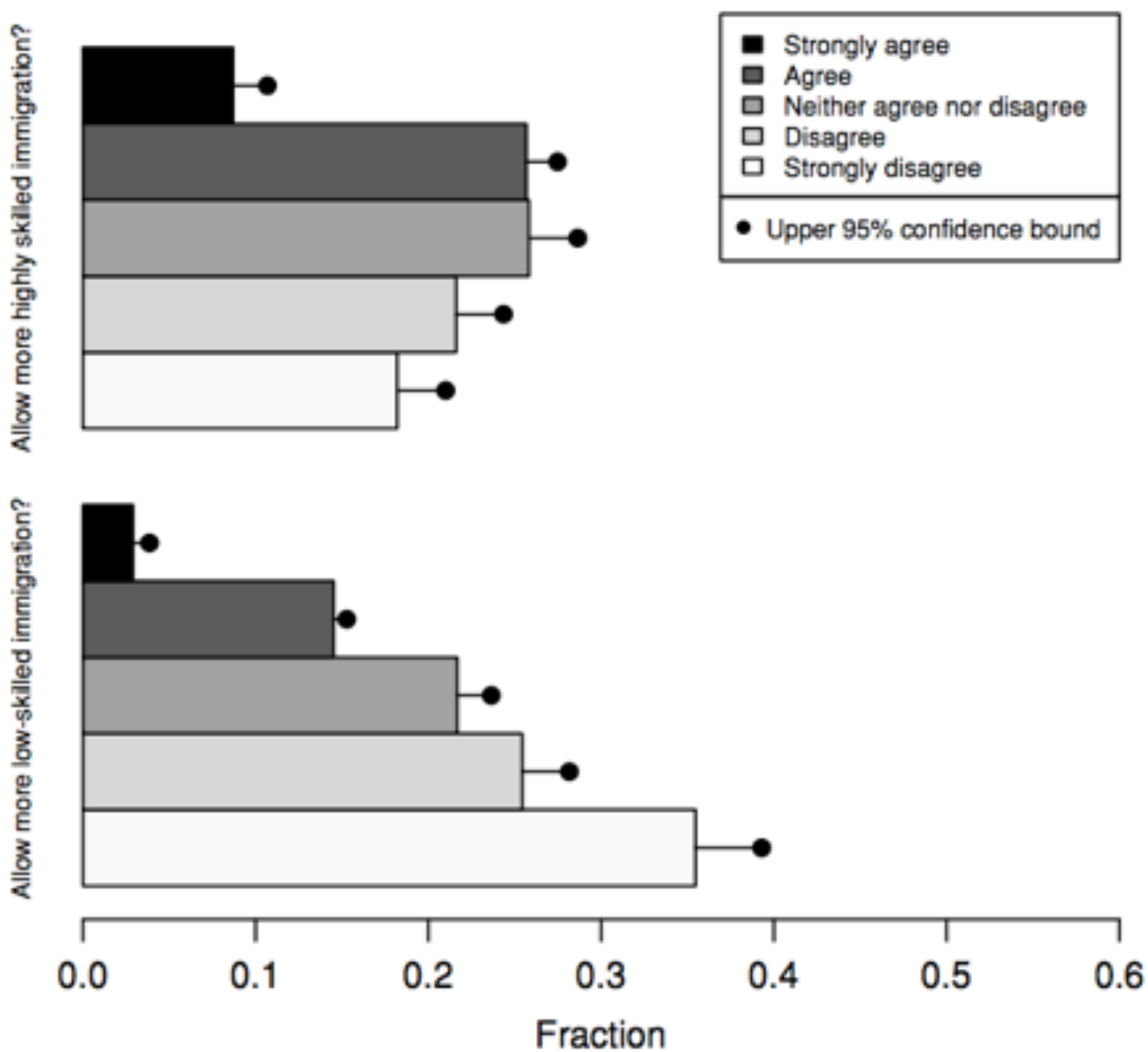
B. Do you agree or disagree that the US should allow more **low-skilled immigrants** from other countries to come and live here?

Hainmueller and Hiscox (2010): why reviewers asked for ordinal probit

Question A (allow more high-skilled immigration?)

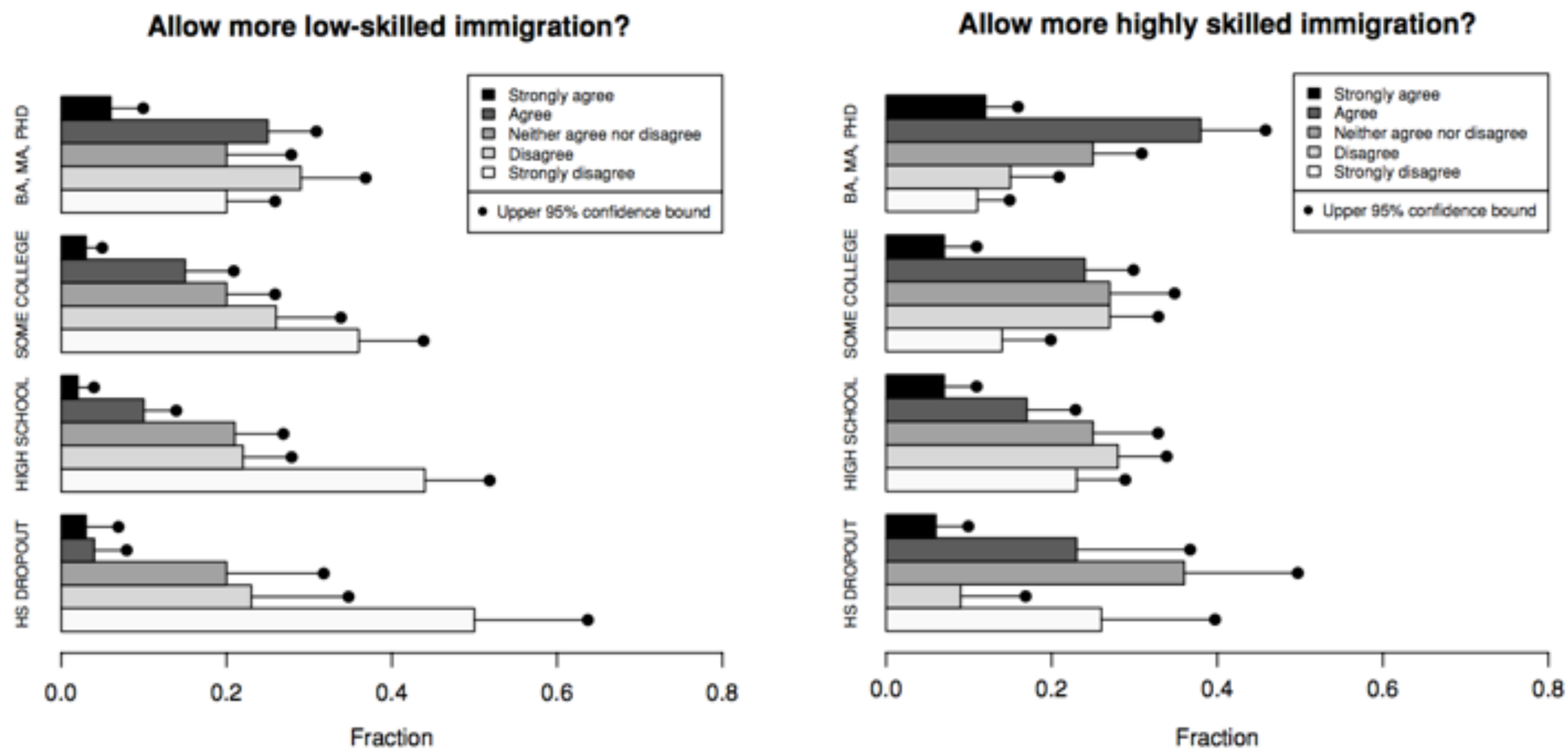
Question B (allow more low-skilled immigration?)

FIGURE 2. Support for Highly Skilled and Low-skilled Immigration



Hainmueller and Hiscox (2010): why reviewers asked for ordinal probit (cont'd)

FIGURE 3. Support for Highly Skilled and Low-skilled Immigration by Respondents' Skill Level



Ordered probit

Motivations:

- **Predict** ordered outcome Y
- Characterize the determinants of a **latent variable** Y* (e.g. support for immigration) underlying ordered outcome Y

1. Strongly disagree 2. Disagree 3. Neither agree nor disagree 4. Agree 5. Strongly agree

Ordered probit: theory

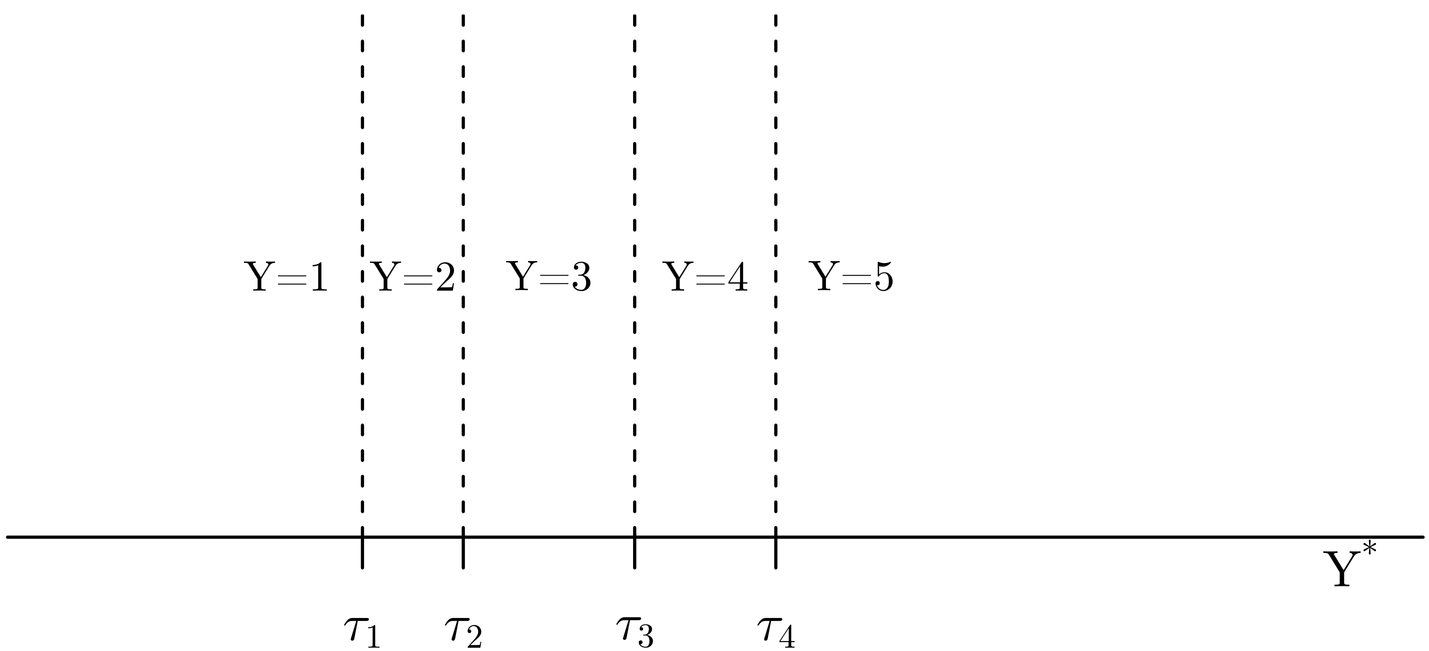
Suppose we observed Y* (support for immigration), which in conjunction with cutpoints τ_1, τ_2 etc perfectly predicts the response given:

$$Y = \begin{cases} 1, & \text{if } Y^* \leq \tau_1 \\ 2, & \text{if } Y^* \in (\tau_1, \tau_2] \\ 3, & \text{if } Y^* \in (\tau_2, \tau_3] \\ 4, & \text{if } Y^* \in (\tau_3, \tau_4] \\ 5, & \text{if } Y^* > \tau_4 \end{cases}$$

Ordered probit: theory

Suppose we observed Y* (support for immigration), which in conjunction with cutpoints τ_1, τ_2 etc perfectly predicts the response given:

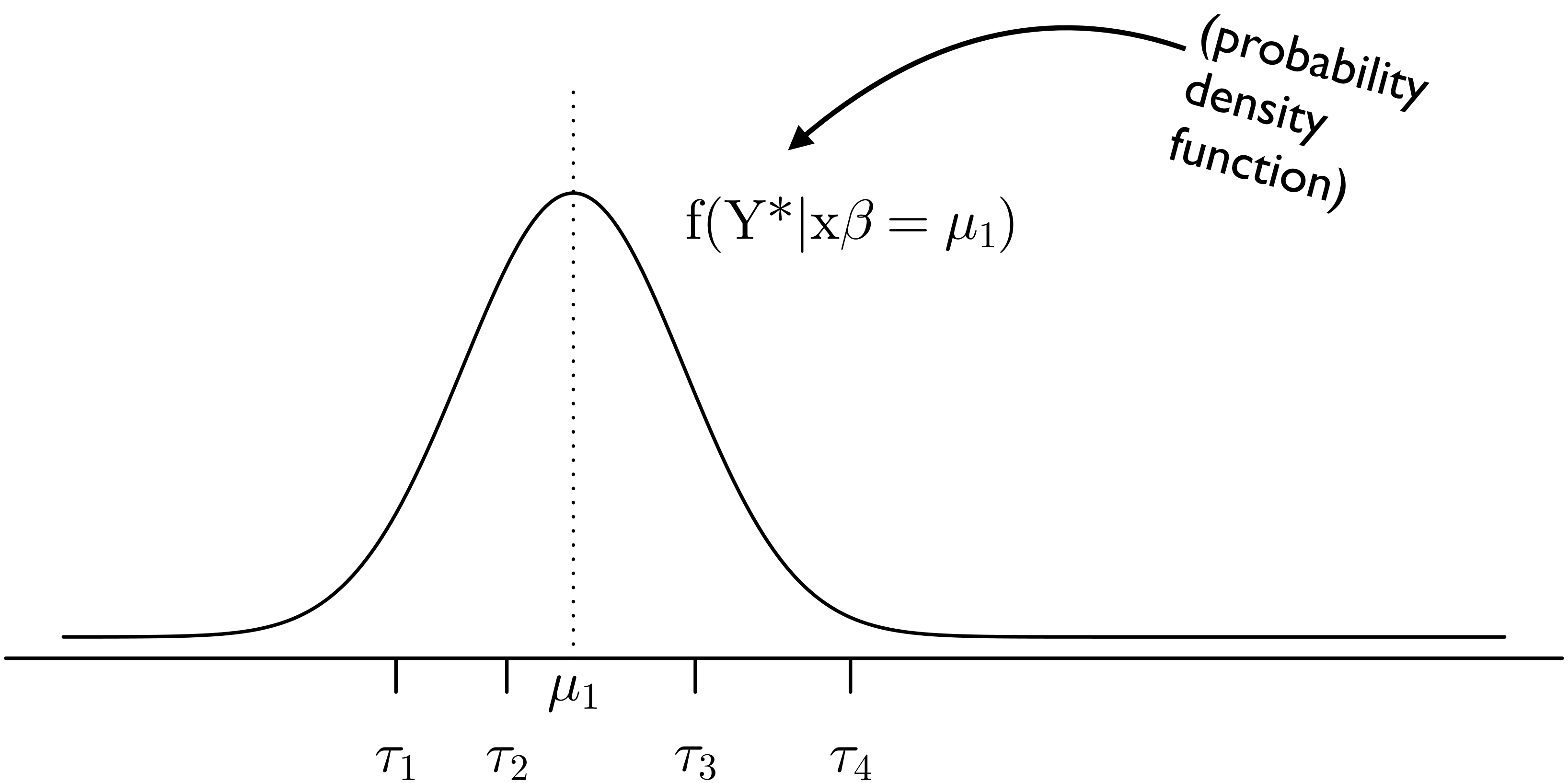
$$Y = \begin{cases} 1, & \text{if } Y^* \leq \tau_1 \\ 2, & \text{if } Y^* \in (\tau_1, \tau_2] \\ 3, & \text{if } Y^* \in (\tau_2, \tau_3] \\ 4, & \text{if } Y^* \in (\tau_3, \tau_4] \\ 5, & \text{if } Y^* > \tau_4 \end{cases}$$



Ordered probit: theory (continued)

We don't observe Y*, but we postulate that it is a linear function of covariates, plus random error (standard normal):

$$\begin{aligned} Y^* &= x\beta + \epsilon \\ \epsilon &\sim N(0,1) \end{aligned}$$

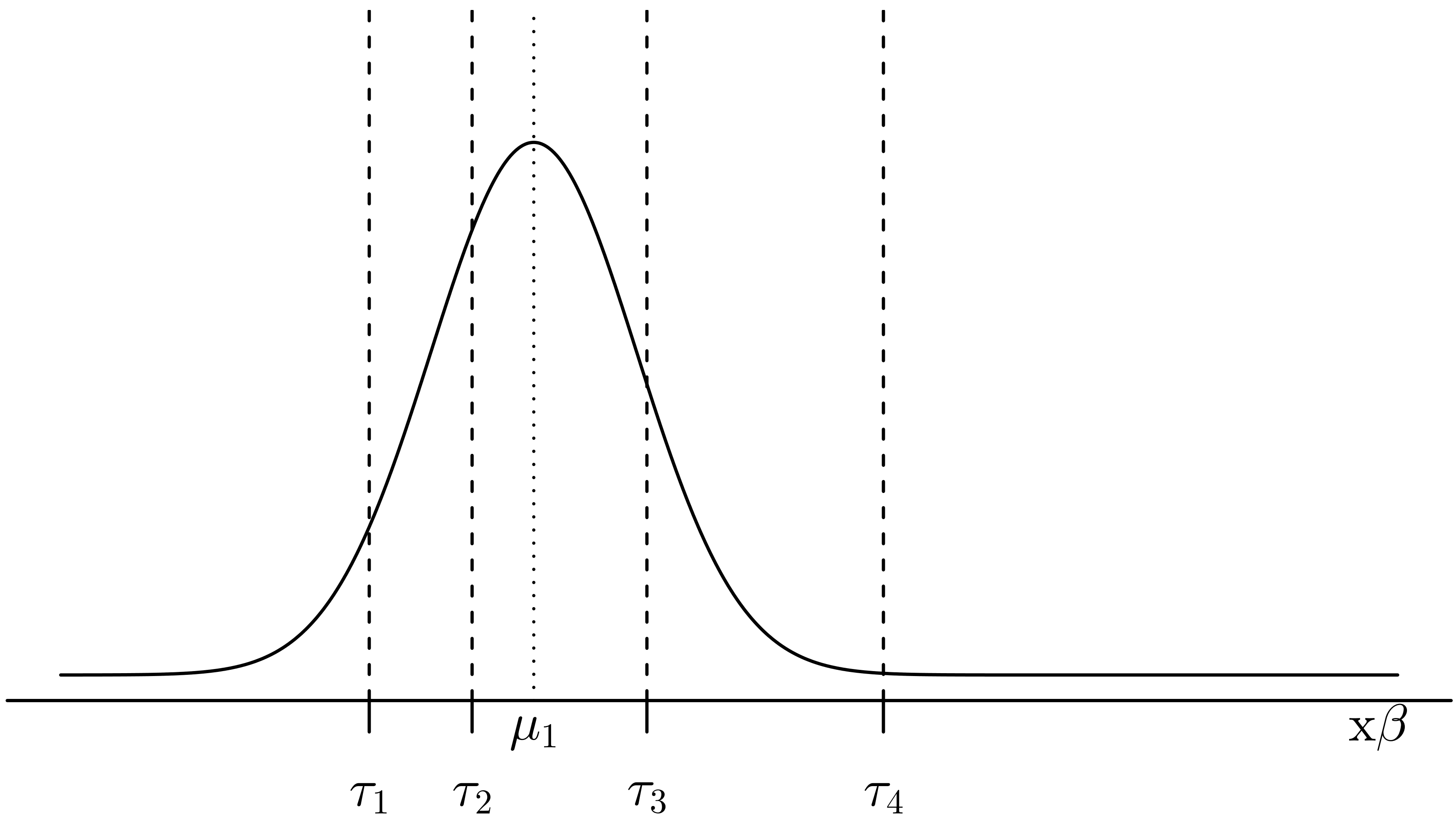


Ordered probit: visualization

That implies that given $\tau_1, \tau_2, \tau_3, \tau_4$ and $\mu_i = x_i\beta$ we know the probability of each outcome:

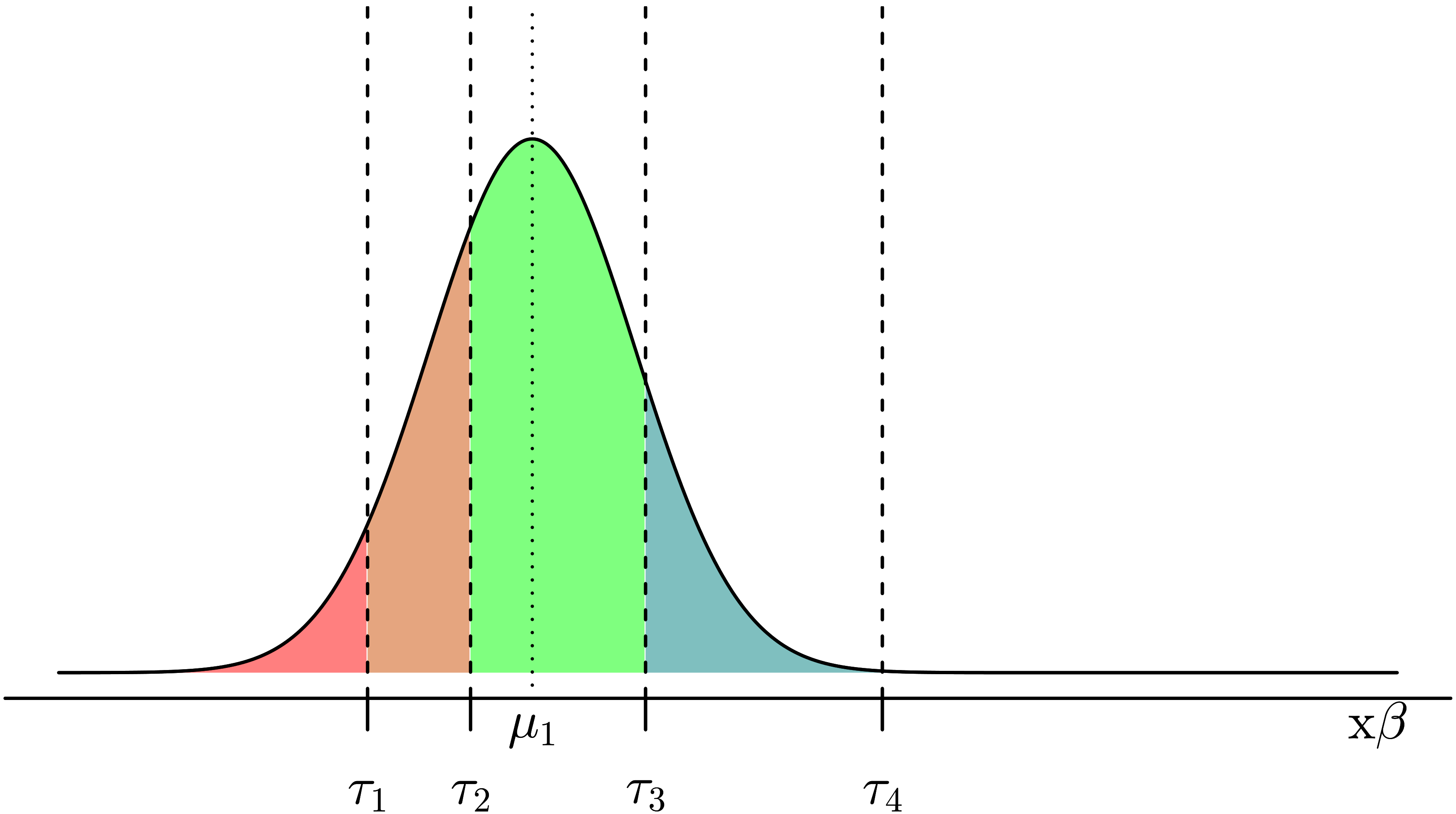
Ordered probit: visualization

That implies that given $\tau_1, \tau_2, \tau_3, \tau_4$ and $\mu_i = x_i\beta$ we know the probability of each outcome:

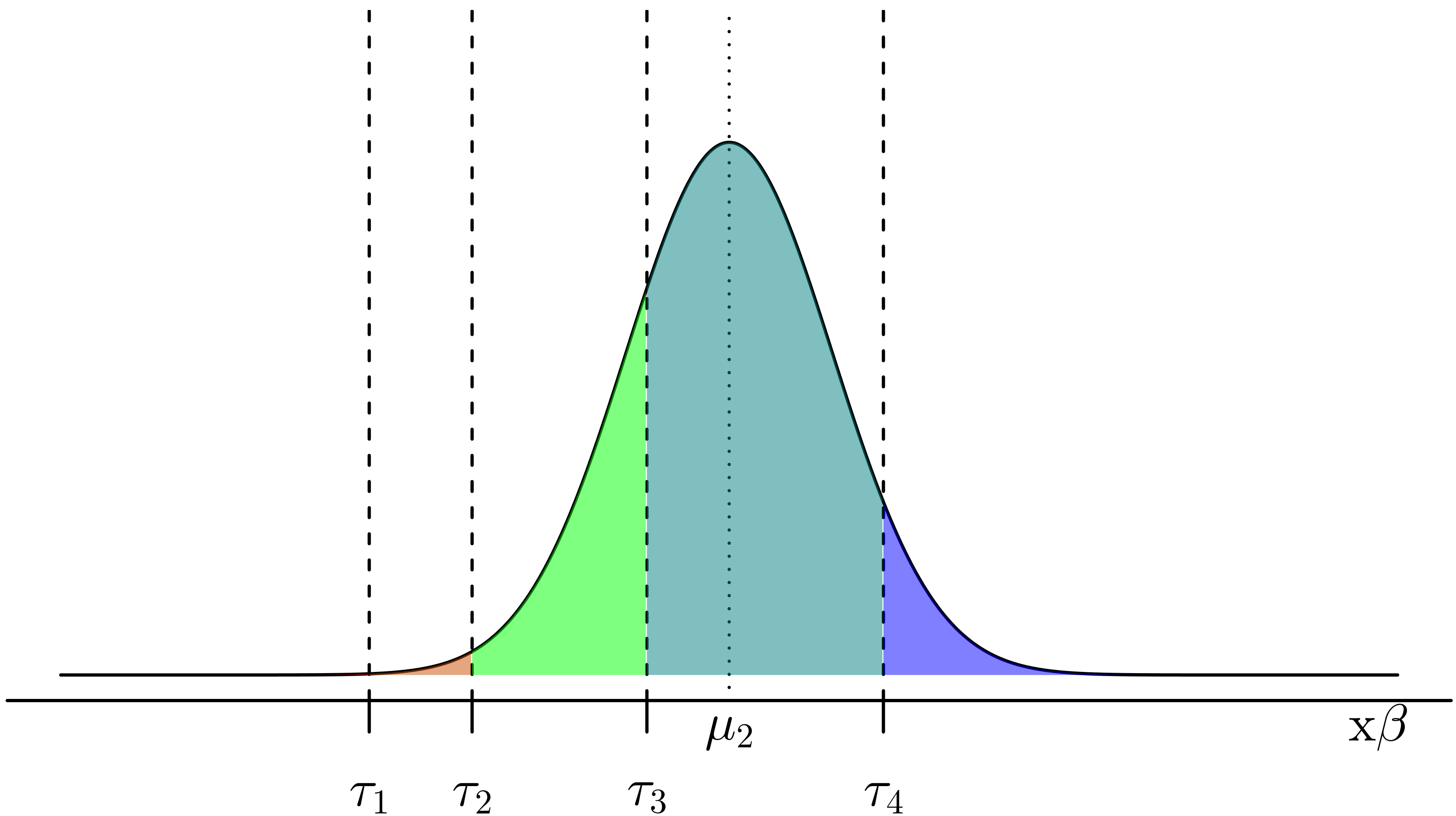


Ordered probit: visualization

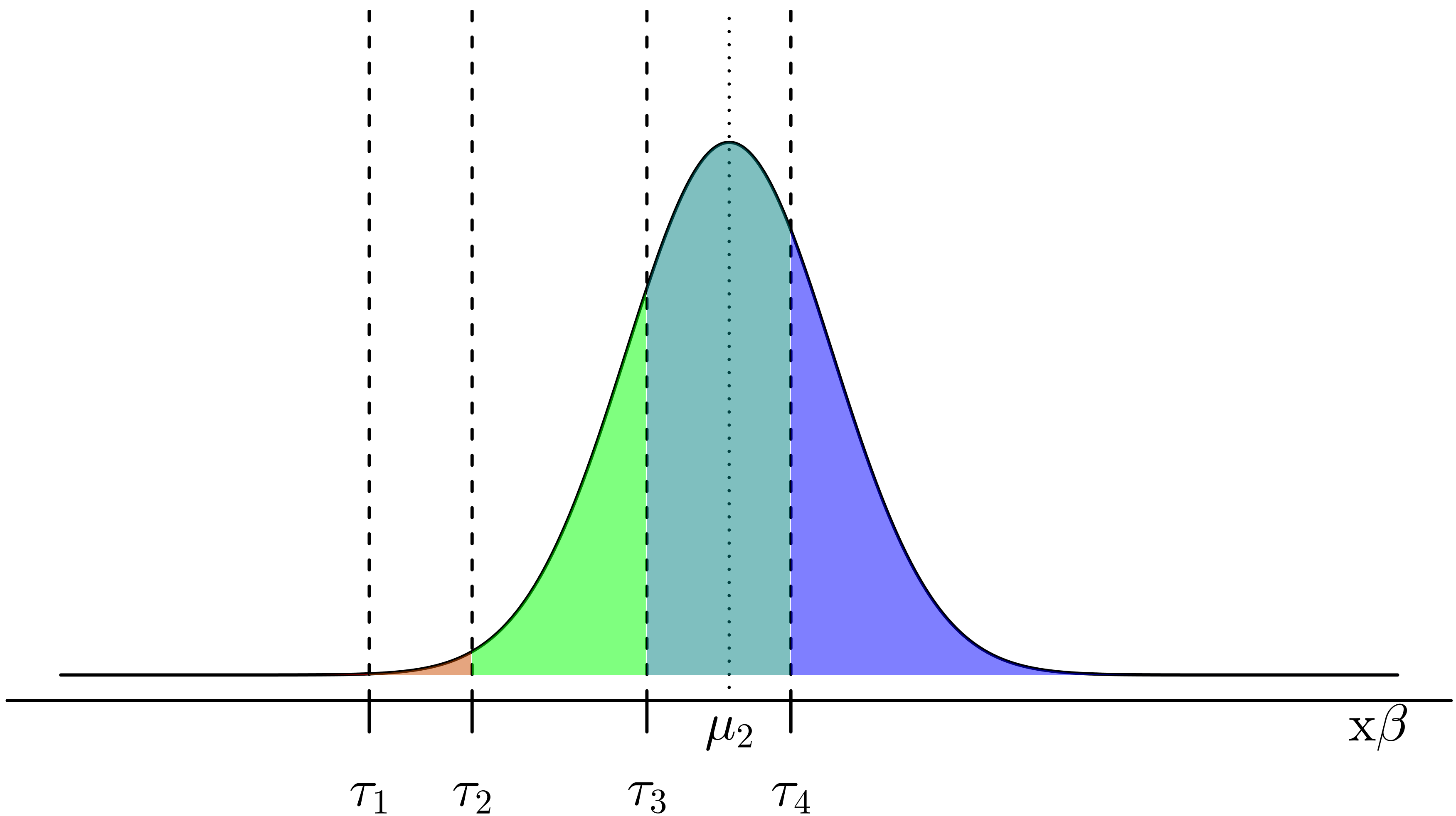
That implies that given $\tau_1, \tau_2, \tau_3, \tau_4$ and $\mu_i = x_i\beta$ we know the probability of each outcome:



Ordered probit: visualization (2)

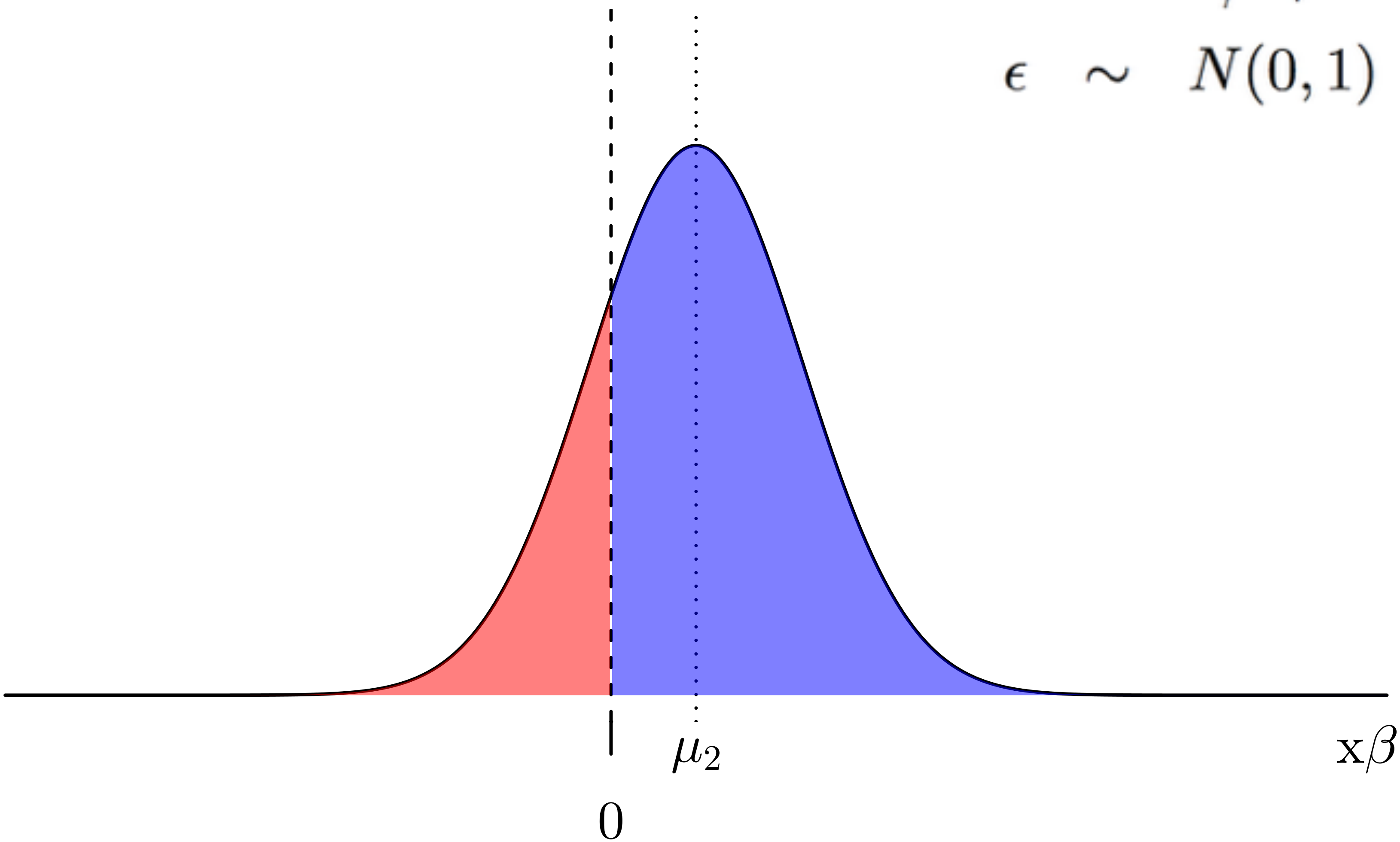


Ordered probit: visualization (3)



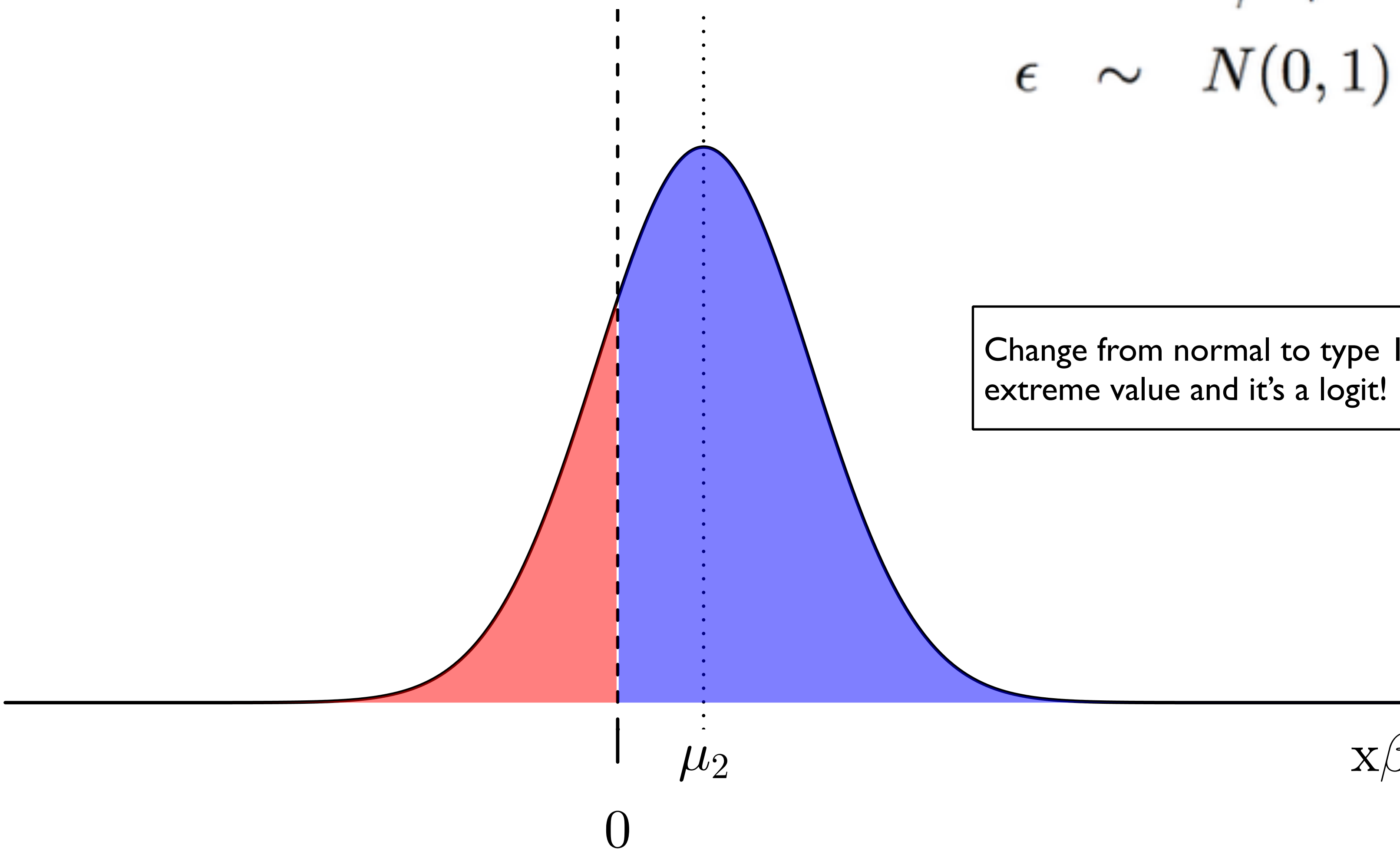
Binary probit: a special case with single threshold at 0

$$\begin{aligned} Y^* &= x\beta + \epsilon \\ \epsilon &\sim N(0,1) \end{aligned}$$



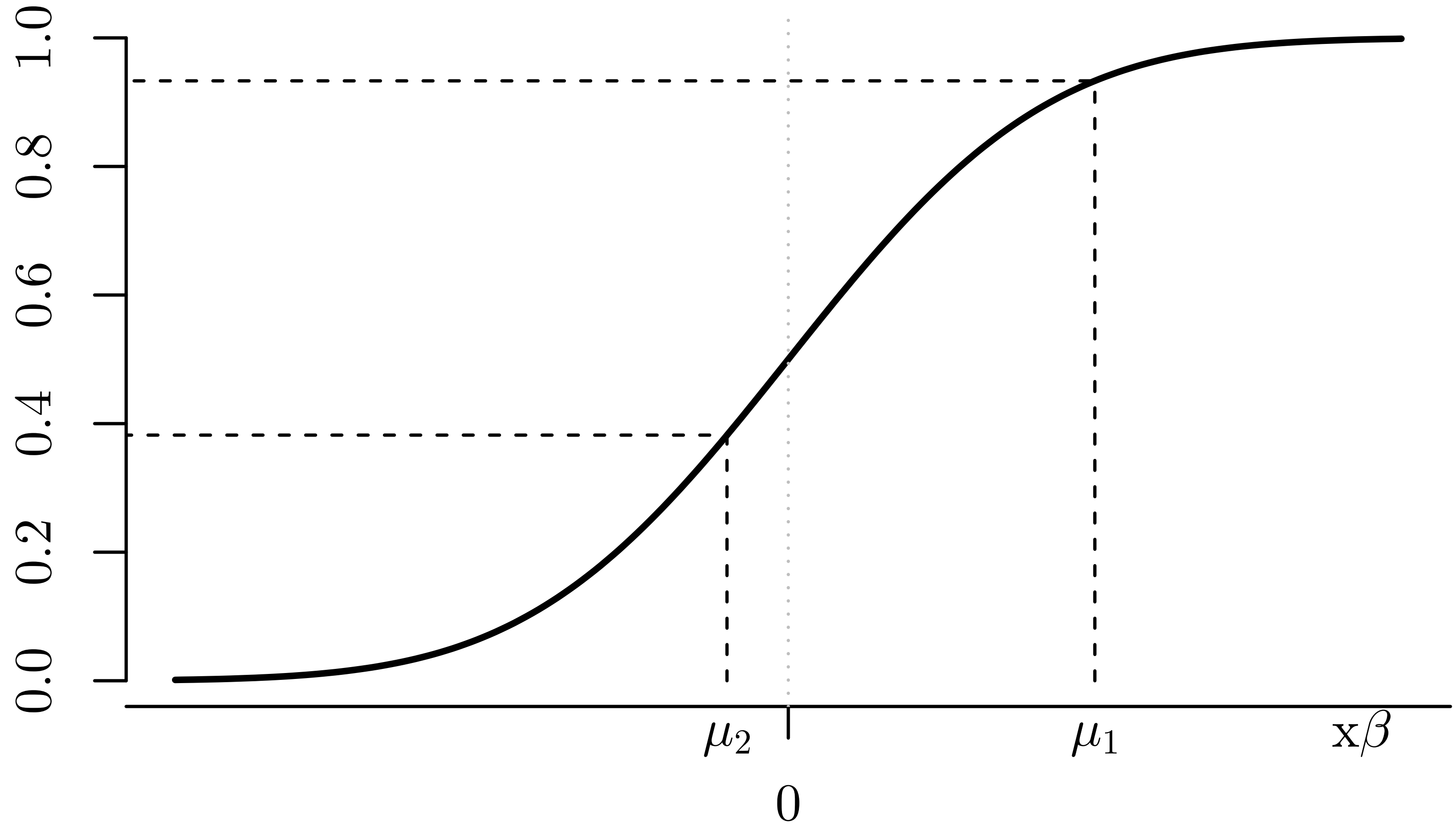
Binary probit: a special case with single threshold at 0

$$Y^* = x\beta + \epsilon$$
$$\epsilon \sim N(0,1)$$



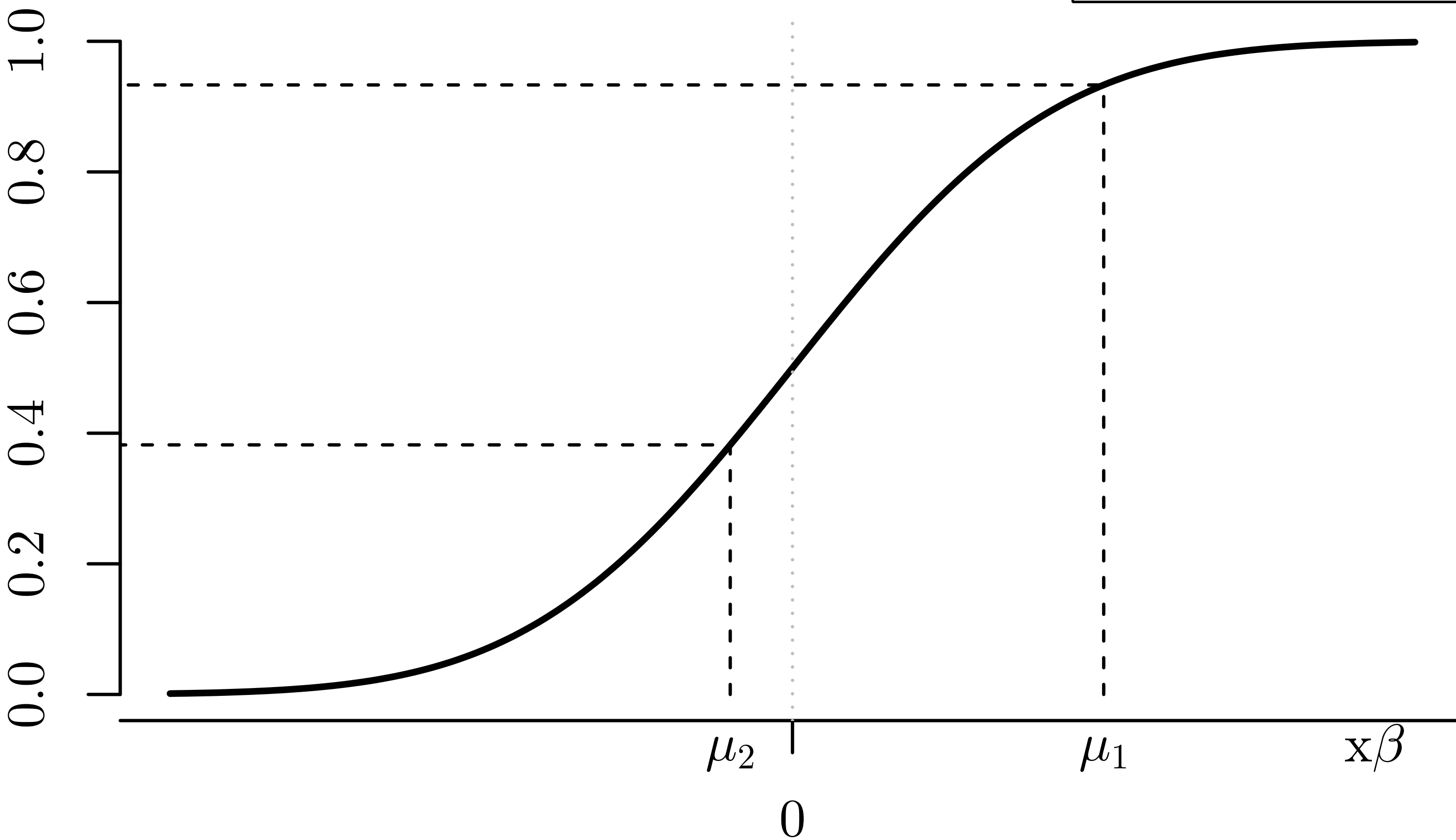
Change from normal to type I extreme value and it's a logit!

Binary probit: a special case with single threshold at 0



Binary probit: a special case with single threshold at 0

Change from normal CDF to type I extreme value and it's a logit!



Back to Hainmueller and Hiscox

To explicitly test the labor market competition argument, we estimate the systematic component of the ordered probit model with the specification.

$$\mu_i = \alpha + \gamma \text{HSKFRAME}_i + \delta(\text{HSKFRAME}_i \cdot \text{EDUCATION}_i) + \theta \text{EDUCATION}_i + Z_i \psi,$$

where the parameter γ is the lower-order term on the treatment indicator that identifies the premium that natives attach to highly skilled immigrants relative to low-skilled immigrants. The parameter δ captures how the premium for highly skilled immigration varies conditional on the skill level of the respondent.

$(\mu_i = Y^* = x_i\beta)$

Z_i contains controls: 7 age bracket dummies, gender dummy, 4 race dummies

"Notice that because the randomization orthogonalized HSKFRAME with respect to Z, the exact covariate choice does not affect the results of the main coefficients of interest." p.70

Ordered probit: estimation

Ordered probit: estimation

How do we estimate β and $\tau_1, \tau_2, \tau_3, \tau_4$?

Ordered probit: estimation

How do we estimate β and $\tau_1, \tau_2, \tau_3, \tau_4$?

Stata: `oprobit depvar [indepvars] [weight] [, options]`

Ordered probit: estimation

How do we estimate β and $\tau_1, \tau_2, \tau_3, \tau_4$?

Stata: `oprobit depvar [indepvars] [weight] [, options]`

```
. oprobit sh_both hskframe ppeducat hskeduc xx* [pweight=weight1]

Iteration 0:   log pseudolikelihood = -2418.2933
Iteration 1:   log pseudolikelihood = -2306.2688
Iteration 2:   log pseudolikelihood = -2306.1887
Iteration 3:   log pseudolikelihood = -2306.1887

Ordered probit regression                       Number of obs     =      1,589
                                                Wald chi2(8)      =     158.52
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -2306.1887               Pseudo R2         =     0.0464
```

sh_both	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
hskframe	.7261249	.2025688	3.58	0.000	.3290974	1.123152
ppeducat	.2683796	.0484328	5.54	0.000	.1734531	.3633061
hskeduc	-.0653202	.0667142	-0.98	0.328	-.1960777	.0654373
xxfemale	-.1771998	.0644352	-2.75	0.006	-.3034904	-.0509092
xxppagecat	-.0110243	.0196088	-0.56	0.574	-.0494569	.0274083
xxWhite	-.374742	.0990717	-3.78	0.000	-.5689189	-.1805651
xxBlack	-.4720909	.1352577	-3.49	0.000	-.7371911	-.2069907
xxHispanic	.0627729	.2058409	0.30	0.760	-.3406679	.4662136
/cut1	-.114744	.1910944			-.4892822	.2597941
/cut2	.5613041	.1905945			.1877457	.9348625
/cut3	1.254911	.1907666			.8810152	1.628807
/cut4	2.258038	.2003352			1.865388	2.650688

Hainmueller and Hiscox: ordered probit results

TABLE 1. Individual Support for Highly Skilled and Low-skilled Immigration—Test of the Labor Market Competition Model

	In Favor of: High Skilled Immigration	In Favor of: Low-skilled Immigration	In Favor of: Immigration				
	(1)	(2)	(3)	(4)	(5)	(6) labor force in	(7) labor force out
Dependent Variable							
EDUCATION	0.21 (0.05)	0.27 (0.05)		0.27 (0.05)		0.33 (0.06)	0.19 (0.07)
HSKFRAME			0.54 (0.07)	0.73 (0.20)	0.56 (0.12)	0.73 (0.28)	0.64 (0.29)
HSKFRAME·EDUCATION				−0.07 (0.07)		−0.08 (0.09)	0.00 (0.11)
HS DROPOUT					−0.41 (0.18)		
HSKFRAME·HS DROPOUT					0.24 (0.25)		
HIGH SCHOOL					−0.16 (0.12)		
HSKFRAME·HIGH SCHOOL					−0.05 (0.17)		
BA DEGREE					0.41 (0.12)		
HSKFRAME·BA DEGREE					−0.08 (0.16)		
(*N*)	798	791	1589	1589	1589	946	643
Covariates	x	x	x	x	x	x	x

Order Probit Coefficients shown with standard errors in parentheses. All models include a set of the covariates age, gender, and race (coefficients not shown here). The reference category for the set of education dummies is SOME COLLEGE (respondents with some college education).

Hainmueller and Hiscox: why ordered probit?

Conventional view: "Your outcome is an ordered categorical variable, so you must estimate an ordered probit model! (Although I don't remember exactly why.)"

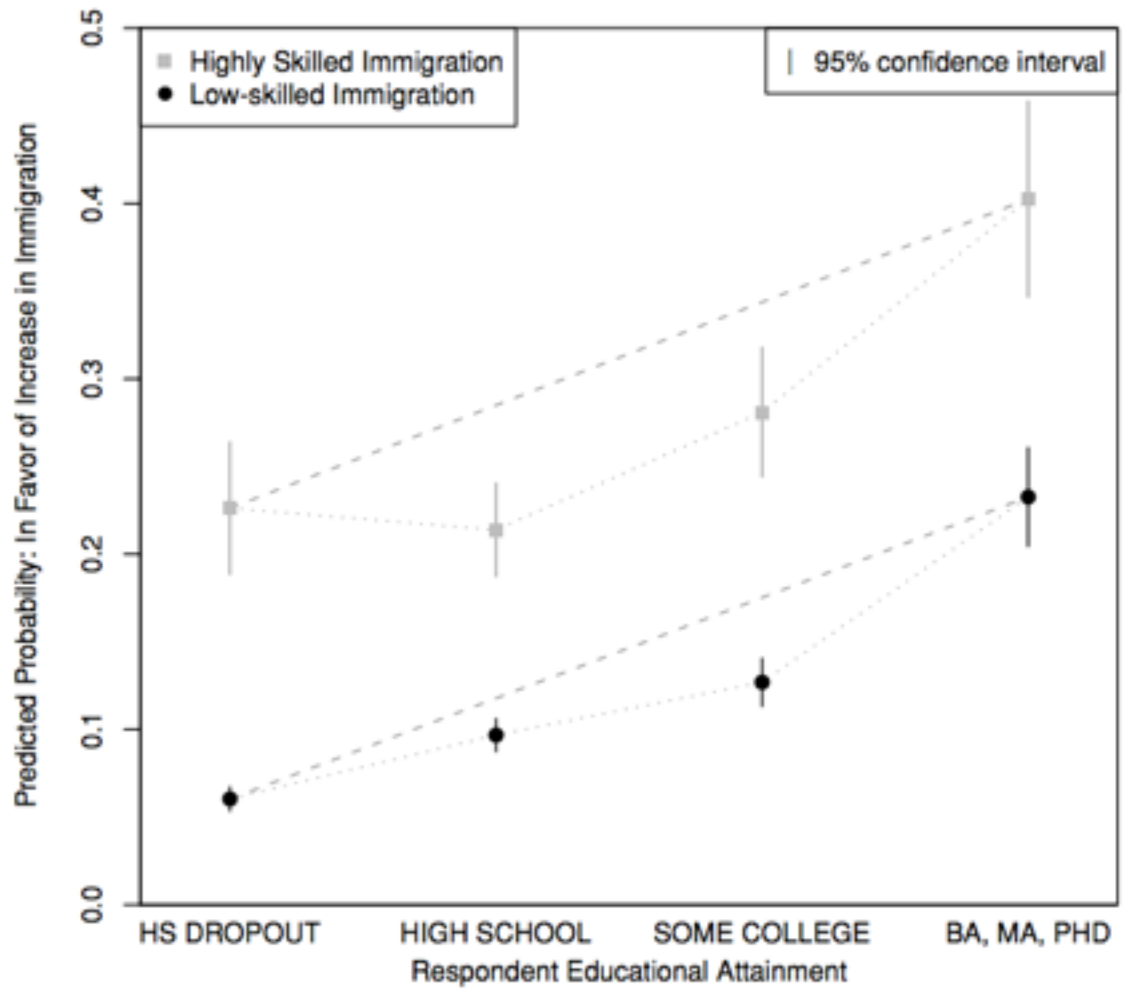
But the authors don't use the model for prediction (e.g. *estimated proportion of respondents answering category 4 given treatment status, education, gender*.)

They report the coefficients (and not the cutoffs!), and move on to logit for a different outcome: **support more immigration**.

Hainmueller and Hiscox: logit results

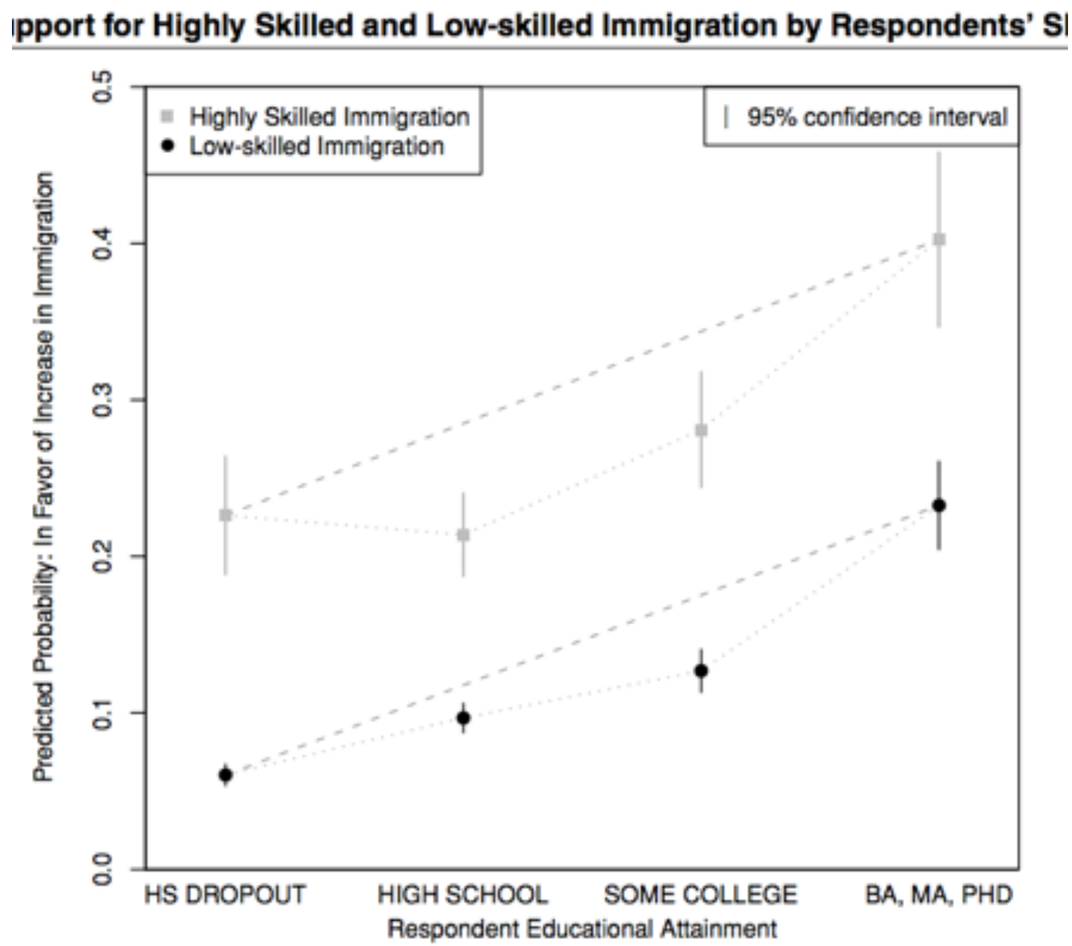
To give some sense of the substantive magnitudes involved, we simulate the predicted probability of supporting an increase in immigration (answers "somewhat agree" and "strongly agree" that the U.S. should allow more immigration) for the median respondent (a white woman aged 45) for all four skill levels and both immigration types based on the least restrictive model (model five in Table 1).

FIGURE 4. Support for Highly Skilled and Low-skilled Immigration by Respondents' Skill Level



Why logit?

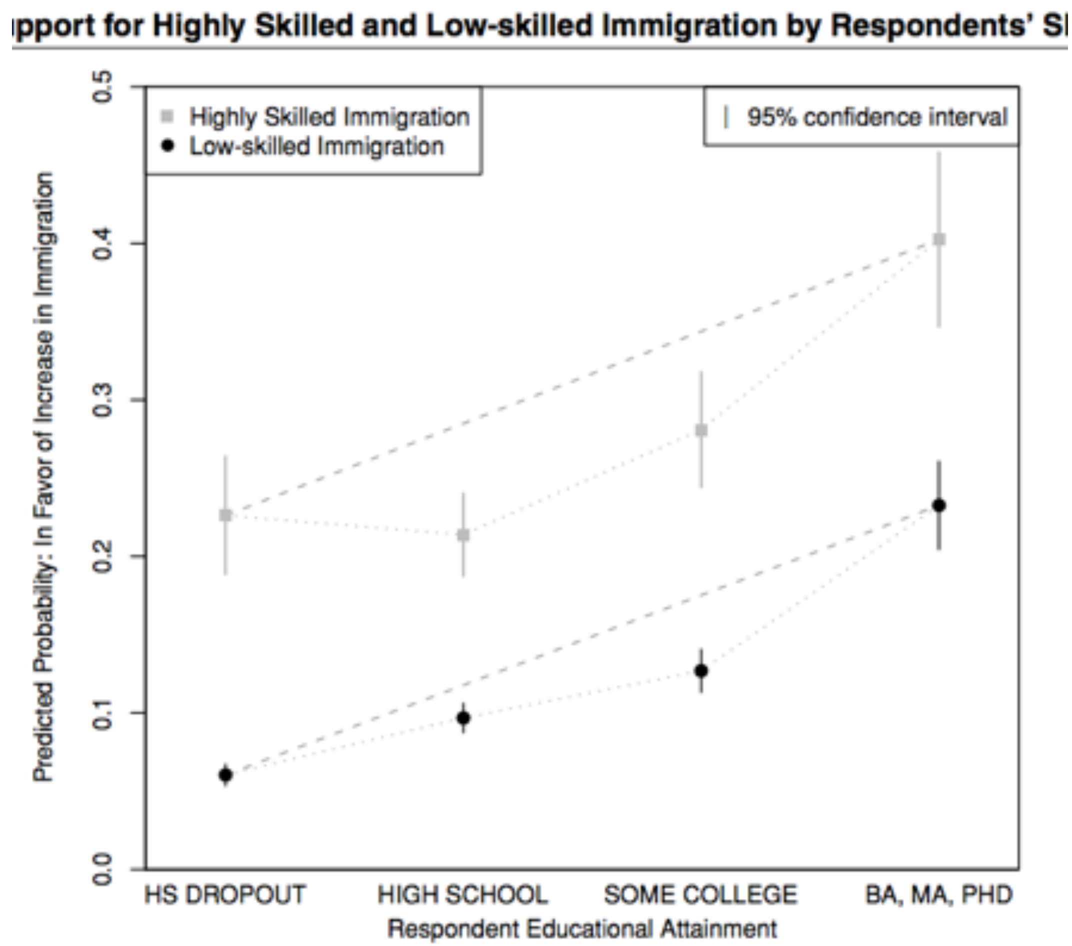
Conventional view: "Outcome is a binary variable, so you must use logit! (Although I don't remember exactly why.)"



Why logit?

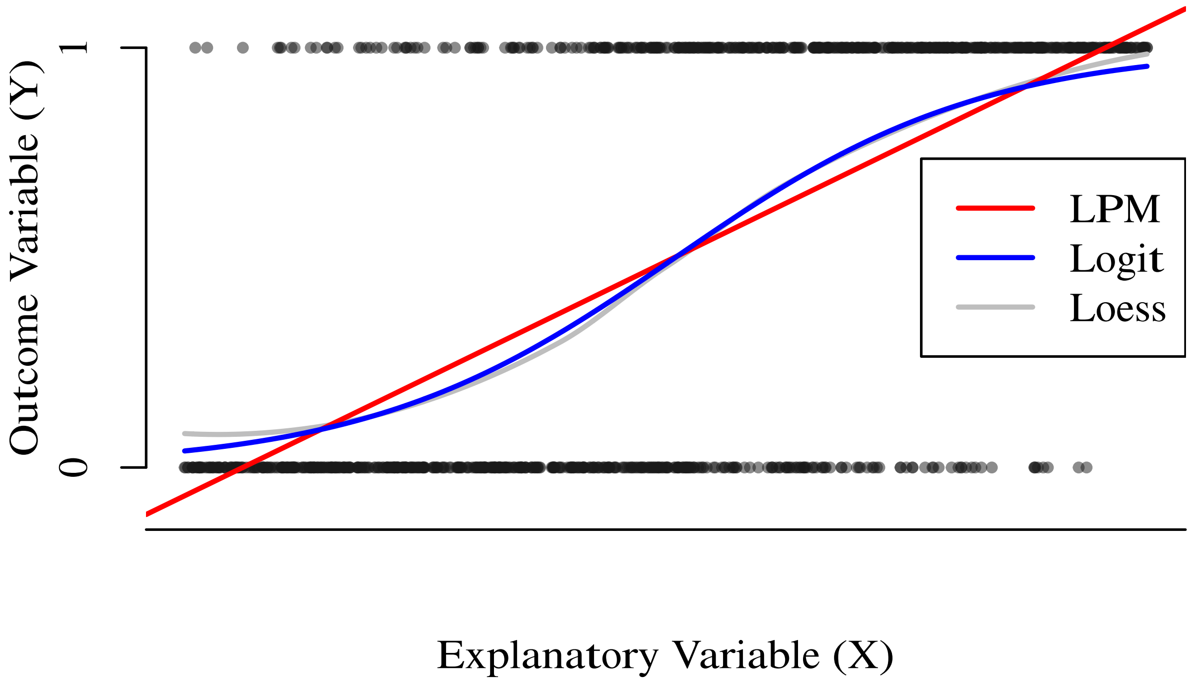
Conventional view: "Outcome is a binary variable, so you must use logit! (Although I don't remember exactly why.)"

Why not estimate a **linear probability model** (LPM) — i.e. OLS despite binary outcome?

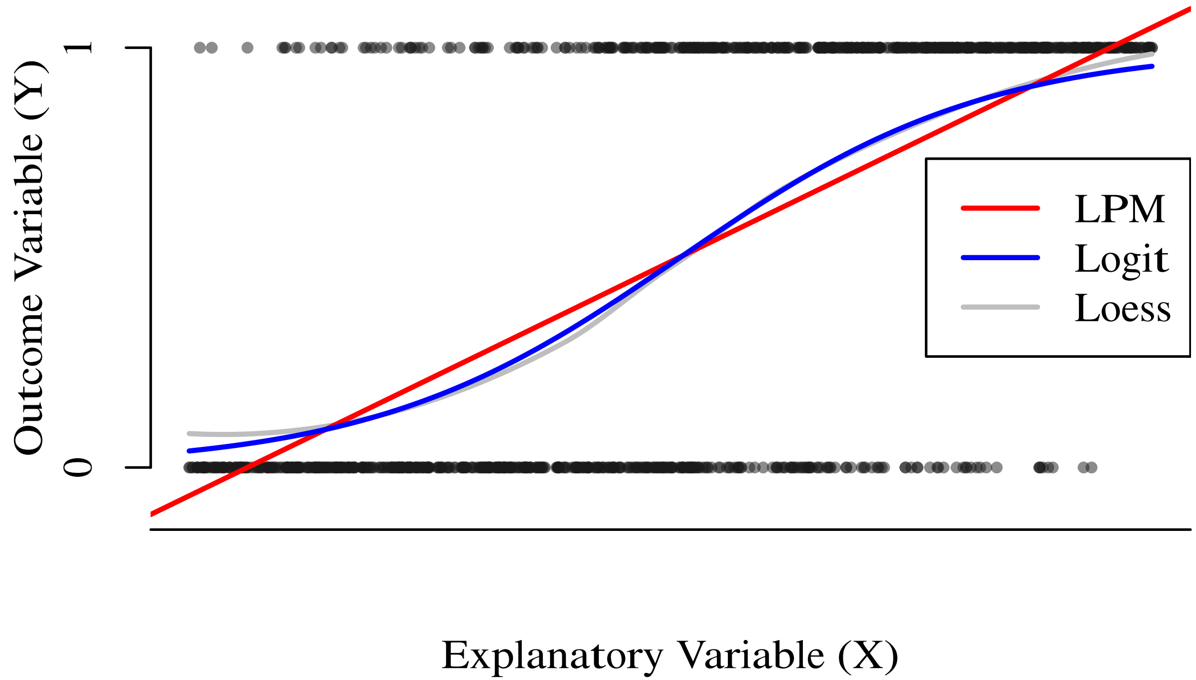


$$\text{SUPPORT}_i = \alpha + \gamma\text{HSKFRAME}_i + \delta\text{HSKFRAME}_i \times \text{EDUCATION}_i + \theta\text{EDUCATION}_i + Z_i\psi$$

The usual case against the linear probability model (LPM)



The usual case against the linear probability model (LPM)



- *Predictions outside the range of dependent variable*
- *Heteroskedasticity (violates OLS assumption)*
- *Non-normal errors (violates OLS assumption*)*
- *Unrealistic for probability to be linear in X*

The defense of the LPM: responses to critiques

The defense of the LPM: responses to critiques

- *Predictions outside the range of dependent variable*

The defense of the LPM: responses to critiques

- *Predictions outside the range of dependent variable*
  - Is prediction (for outliers) the goal?

The defense of the LPM: responses to critiques

- *Predictions outside the range of dependent variable*
  - Is prediction (for outliers) the goal?
- *Heteroskedasticity (violates OLS assumption)*

The defense of the LPM: responses to critiques

- *Predictions outside the range of dependent variable*
  - Is prediction (for outliers) the goal?
- *Heteroskedasticity (violates OLS assumption)*
  - See Huber-White standard errors, other corrections for heteroskedasticity (`robust` option in Stata)

The defense of the LPM: responses to critiques

- *Predictions outside the range of dependent variable*
  - Is prediction (for outliers) the goal?
- *Heteroskedasticity (violates OLS assumption)*
  - See Huber-White standard errors, other corrections for heteroskedasticity (`robust` option in Stata)
- *Non-normal errors (violates OLS assumption)*

The defense of the LPM: responses to critiques

- *Predictions outside the range of dependent variable*
  - Is prediction (for outliers) the goal?
- *Heteroskedasticity (violates OLS assumption)*
  - See Huber-White standard errors, other corrections for heteroskedasticity (`robust` option in Stata)
- *Non-normal errors (violates OLS assumption)*
  - That assumption is necessary for inference (i.e. valid standard errors) in small samples, but not asymptotically (see MHE section 3.1), and not for approximating the CEF

The defense of the LPM: responses to critiques

- *Predictions outside the range of dependent variable*
  - Is prediction (for outliers) the goal?
- *Heteroskedasticity (violates OLS assumption)*
  - See Huber-White standard errors, other corrections for heteroskedasticity (`robust` option in Stata)
- *Non-normal errors (violates OLS assumption)*
  - That assumption is necessary for inference (i.e. valid standard errors) in small samples, but not asymptotically (see MHE section 3.1), and not for approximating the CEF
- *Unrealistic for probability to be linear in X*

The defense of the LPM: responses to critiques

- *Predictions outside the range of dependent variable*
  - Is prediction (for outliers) the goal?
- *Heteroskedasticity (violates OLS assumption)*
  - See Huber-White standard errors, other corrections for heteroskedasticity (`robust` option in Stata)
- *Non-normal errors (violates OLS assumption)*
  - That assumption is necessary for inference (i.e. valid standard errors) in small samples, but not asymptotically (see MHE section 3.1), and not for approximating the CEF
- *Unrealistic for probability to be linear in X*
  - Yes, especially when probabilities are near 1 or 0 **(ceiling and floor effects)**; but is probit the right form?

The defense of the LPM: continued

The defense of the LPM: continued

- **Advantages of LPM**:

The defense of the LPM: continued

- **Advantages of LPM**:
  - Ease of interpretation. *Is that just because you don't understand log odds?*

The defense of the LPM: continued

- **Advantages of LPM**:
  - Ease of interpretation. *Is that just because you don't understand log odds?*
  - Best linear approximation to the CEF.

The defense of the LPM: continued

- **Advantages of LPM**:
  - Ease of interpretation. *Is that just because you don't understand log odds?*
  - Best linear approximation to the CEF.
- **Disadvantage of logit/probit**:

The defense of the LPM: continued

- **Advantages of LPM**:
  - Ease of interpretation. *Is that just because you don't understand log odds?*
  - Best linear approximation to the CEF.
- **Disadvantage of logit/probit**:
  - Doesn't directly give the Average Treatment Effect

The defense of the LPM: continued

- **Advantages of LPM**:
  - Ease of interpretation. *Is that just because you don't understand log odds?*
  - Best linear approximation to the CEF.
- **Disadvantage of logit/probit**:
  - Doesn't directly give the Average Treatment Effect
  - Can convert logit/probit estimates to something equivalent, and in simulations that is the same as the LPM estimate

The defense of the LPM: continued

- **Advantages of LPM**:
  - Ease of interpretation. *Is that just because you don't understand log odds?*
  - Best linear approximation to the CEF.
- **Disadvantage of logit/probit**:
  - Doesn't directly give the Average Treatment Effect
  - Can convert logit/probit estimates to something equivalent, and in simulations that is the same as the LPM estimate
  - Other quantities of interest are sensitive to omitted variables — even *variables uncorrelated with treatment* (Carina Mood, Eur. Soc. Rev. 2010)

The defense of the LPM: continued

- **Advantages of LPM**:
  - Ease of interpretation. *Is that just because you don't understand log odds?*
  - Best linear approximation to the CEF.
- **Disadvantage of logit/probit**:
  - Doesn't directly give the Average Treatment Effect
  - Can convert logit/probit estimates to something equivalent, and in simulations that is the same as the LPM estimate
  - Other quantities of interest are sensitive to omitted variables — even *variables uncorrelated with treatment* (Carina Mood, Eur. Soc. Rev. 2010)
- When interest is in coefficient on binary variable (e.g. treatment),

The defense of the LPM: continued

- **Advantages of LPM**:
  - Ease of interpretation. *Is that just because you don't understand log odds?*
  - Best linear approximation to the CEF.
- **Disadvantage of logit/probit**:
  - Doesn't directly give the Average Treatment Effect
  - Can convert logit/probit estimates to something equivalent, and in simulations that is the same as the LPM estimate
  - Other quantities of interest are sensitive to omitted variables — even *variables uncorrelated with treatment* (Carina Mood, Eur. Soc. Rev. 2010)
- When interest is in coefficient on binary variable (e.g. treatment),
  - CEF is linear with respect to variable of interest

The defense of the LPM: continued

- **Advantages of LPM**:
  - Ease of interpretation. *Is that just because you don't understand log odds?*
  - Best linear approximation to the CEF.
- **Disadvantage of logit/probit**:
  - Doesn't directly give the Average Treatment Effect
  - Can convert logit/probit estimates to something equivalent, and in simulations that is the same as the LPM estimate
  - Other quantities of interest are sensitive to omitted variables — even *variables uncorrelated with treatment* (Carina Mood, Eur. Soc. Rev. 2010)
- When interest is in coefficient on binary variable (e.g. treatment),
  - CEF is linear with respect to variable of interest
  - Logit vs LPM matters only if particular kind of covariate imbalance

The defense of the LPM: continued

Gailmard pp 171-2

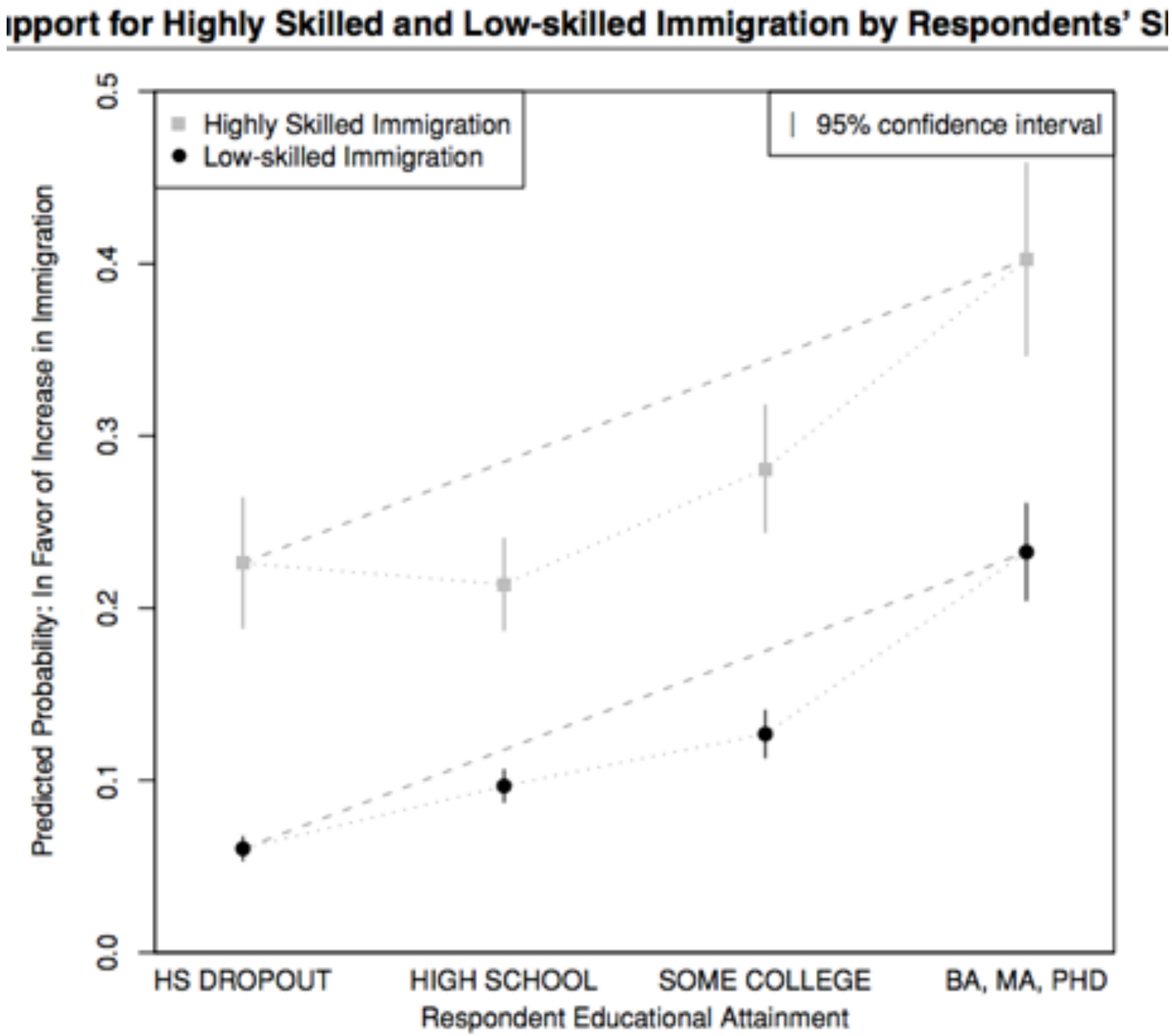
"If the CEF is linear, as it is for a saturated model, [OLS] gives the CEF.… If the CEF is non-linear, [OLS] approximates the CEF. Usually it does it pretty well. Obviously, the LPM won't give the true marginal effects from the right nonlinear model. But then, the same is true for the 'wrong' nonlinear model! The fact that we have a probit, a logit, and the LPM [shows] that we don't know what the 'right' model is. Hence, there is a lot to be said for sticking to a linear regression function as compared to a fairly arbitrary choice of a non-linear one! Nonlinearity per se is a red herring."

Steve Pischke

from MHE blog http://www.mostlyharmlesseconometrics.com/2012/07/probit-better-than-lpm/

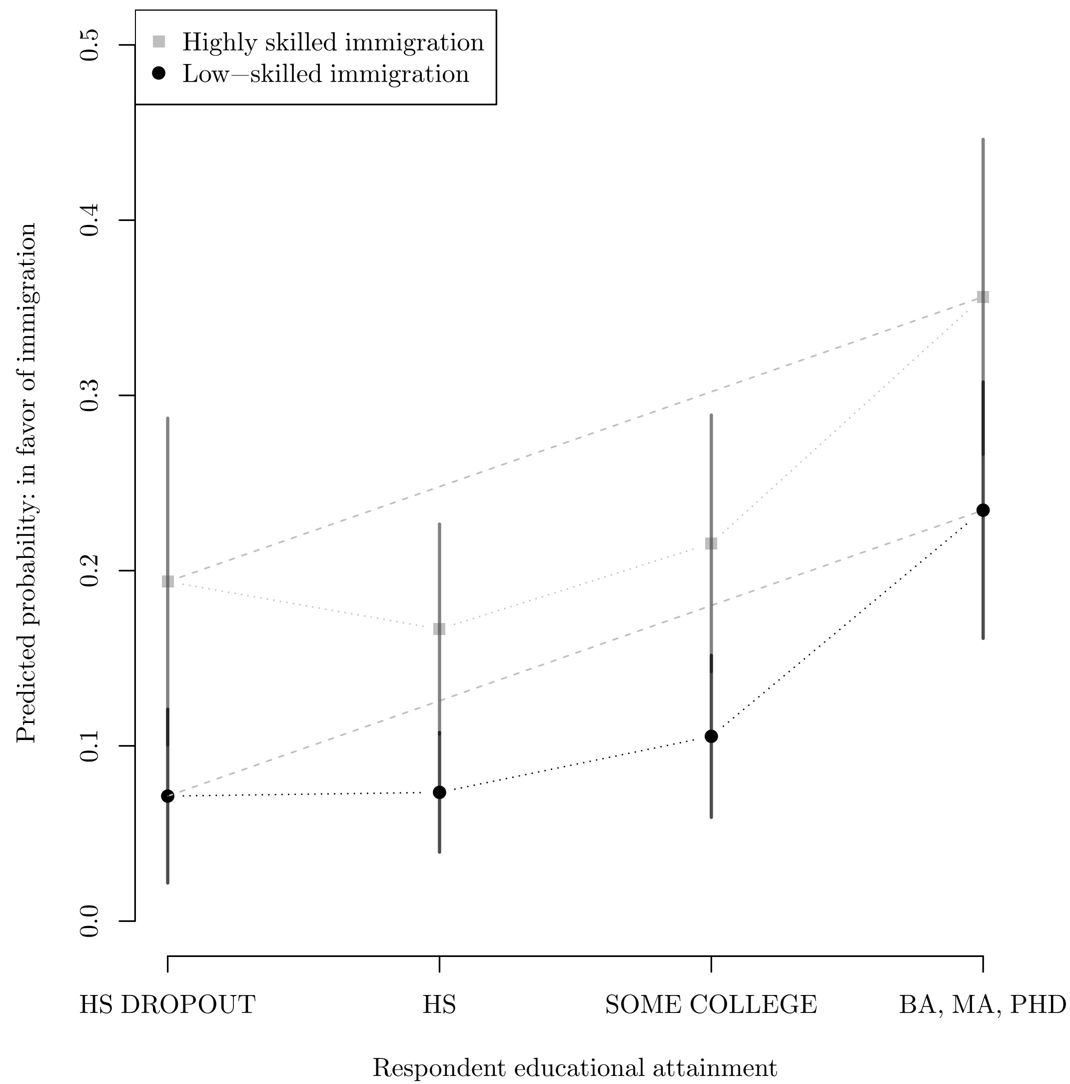
The defense of the LPM: continued

Original Figure 4 (based on logit)



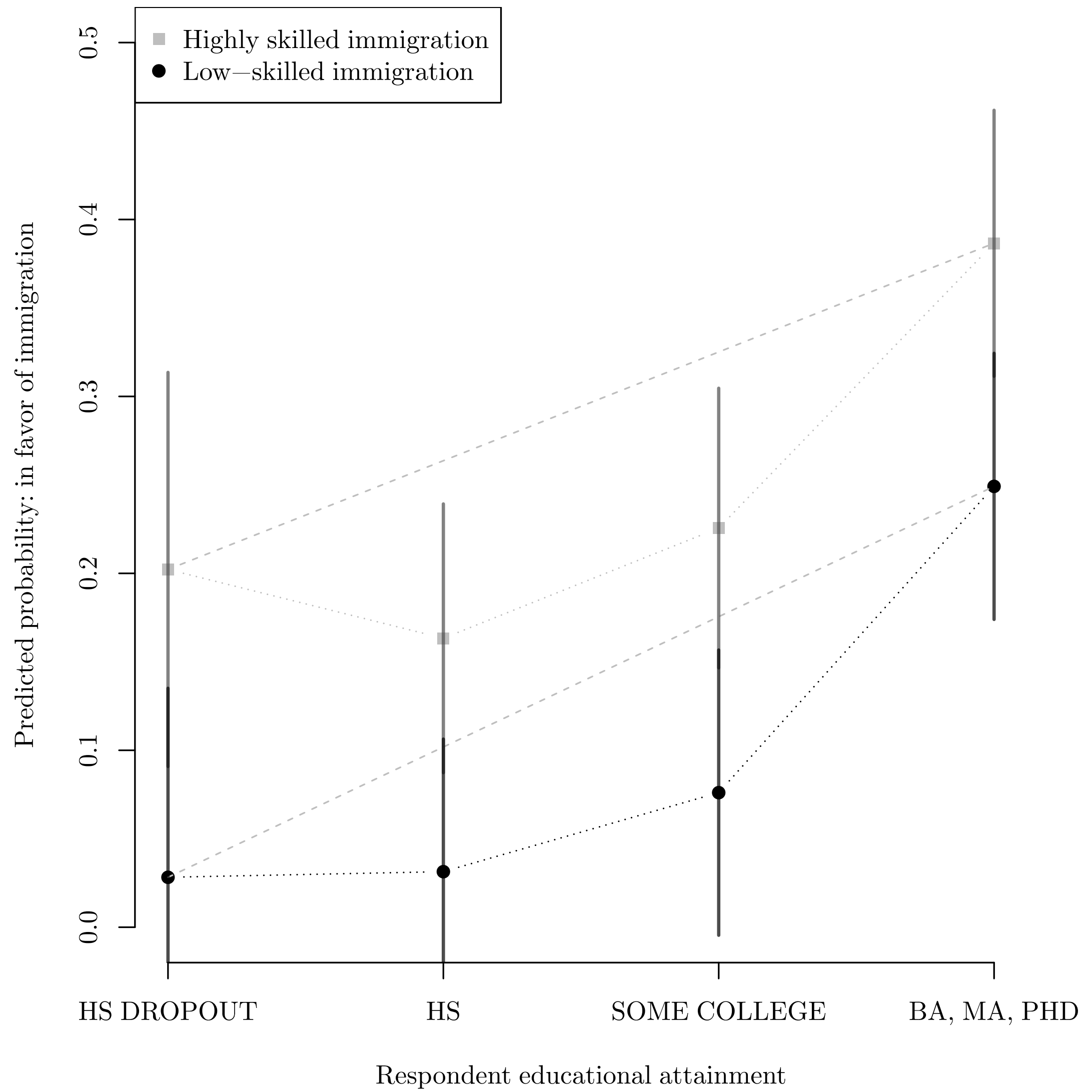
The defense of the LPM: continued

My Figure 4 (based on logit)



The defense of the LPM: continued

My Figure 4 (based on LPM)



So why learn anything other than OLS?

So why learn anything other than OLS?

There is a lot you could study: standard GLM models, other MLE estimation models, IRT measurement models, neural nets, random forests, Bayesian approaches (e.g. topic models for text) . . . .

So why learn anything other than OLS?

There is a lot you could study: standard GLM models, other MLE estimation models, IRT measurement models, neural nets, random forests, Bayesian approaches (e.g. topic models for text) . . . .

My view: for estimation of **treatment effects**, you need to focus on research design, data collection, and presentation. Usually, statistical modeling beyond OLS (with standard error corrections) is a distraction.

So why learn anything other than OLS?

There is a lot you could study: standard GLM models, other MLE estimation models, IRT measurement models, neural nets, random forests, Bayesian approaches (e.g. topic models for text) . . . .

My view: for estimation of **treatment effects**, you need to focus on research design, data collection, and presentation. Usually, statistical modeling beyond OLS (with standard error corrections) is a distraction.

But going beyond OLS can be useful for other goals:

So why learn anything other than OLS?

There is a lot you could study: standard GLM models, other MLE estimation models, IRT measurement models, neural nets, random forests, Bayesian approaches (e.g. topic models for text) . . . .

My view: for estimation of **treatment effects**, you need to focus on research design, data collection, and presentation. Usually, statistical modeling beyond OLS (with standard error corrections) is a distraction.

But going beyond OLS can be useful for other goals:

- Prediction (e.g. probability of regime breakdown)

So why learn anything other than OLS?

There is a lot you could study: standard GLM models, other MLE estimation models, IRT measurement models, neural nets, random forests, Bayesian approaches (e.g. topic models for text) . . . .

My view: for estimation of **treatment effects**, you need to focus on research design, data collection, and presentation. Usually, statistical modeling beyond OLS (with standard error corrections) is a distraction.

But going beyond OLS can be useful for other goals:

- Prediction (e.g. probability of regime breakdown)
- Measurement (e.g. of regime types, pooling information from multiple expert surveys)

So why learn anything other than OLS?

There is a lot you could study: standard GLM models, other MLE estimation models, IRT measurement models, neural nets, random forests, Bayesian approaches (e.g. topic models for text) . . . .

My view: for estimation of **treatment effects**, you need to focus on research design, data collection, and presentation. Usually, statistical modeling beyond OLS (with standard error corrections) is a distraction.

But going beyond OLS can be useful for other goals:

- Prediction (e.g. probability of regime breakdown)
- Measurement (e.g. of regime types, pooling information from multiple expert surveys)
- Description of relationships (e.g. regime types and development)

(Will statistical modeling help with **explanation**?)

So why learn anything other than OLS?

There is a lot you could study: standard GLM models, other MLE estimation models, IRT measurement models, neural nets, random forests, Bayesian approaches (e.g. topic models for text) . . . .

My view: for estimation of **treatment effects**, you need to focus on research design, data collection, and presentation. Usually, statistical modeling beyond OLS (with standard error corrections) is a distraction.

But going beyond OLS can be useful for other goals:

- Prediction (e.g. probability of regime breakdown)
- Measurement (e.g. of regime types, pooling information from multiple expert surveys)
- Description of relationships (e.g. regime types and development)

(Will statistical modeling help with **explanation**?)

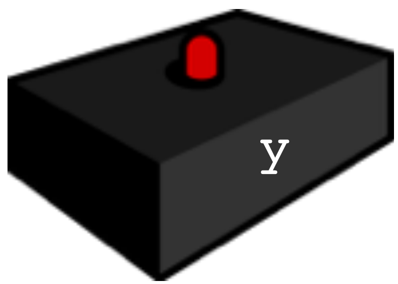
So let's get a taste of statistical modeling more generally.

What is a statistical model?

A statistical model describes how a dependent variable (Y) is thought to have been generated.

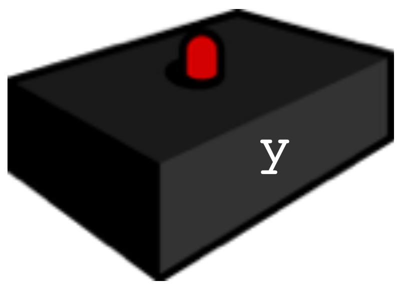
What is a statistical model?

A statistical model describes how a dependent variable (Y) is thought to have been generated.



What is a statistical model?

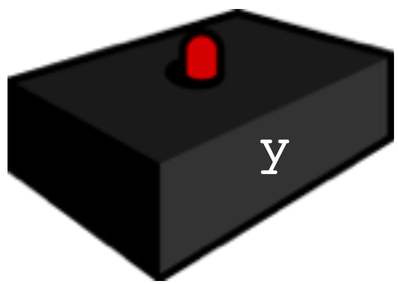
A statistical model describes how a dependent variable (Y) is thought to have been generated.

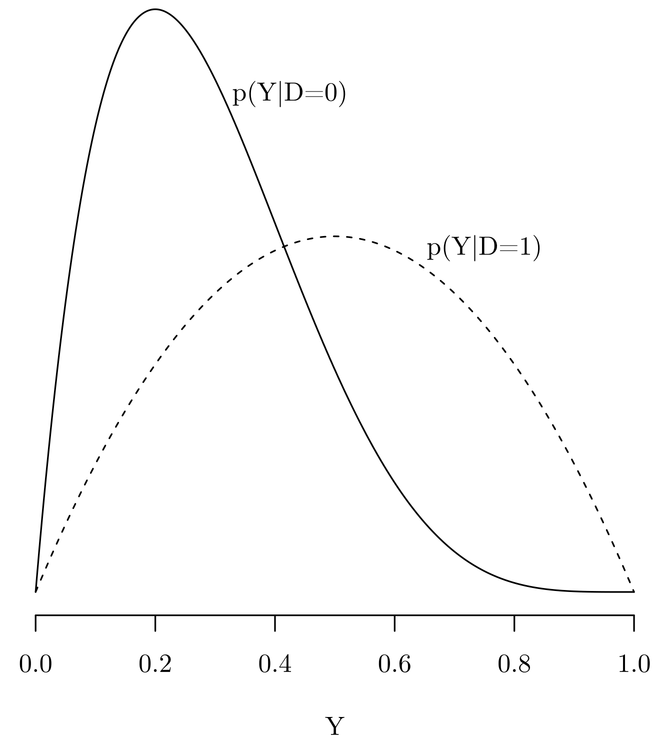


More formally, a statistical model describes a **set of probability distributions** for a random variable (Y).

What is a statistical model?

A statistical model describes how a dependent variable (Y) is thought to have been generated.





More formally, a statistical model describes a **set of probability distributions** for a random variable (Y).

In any interesting statistical model, **different units have different distributions**, depending on the features of the unit (e.g. exposure to treatment vs. control, values of covariates).

Random variables and probability distributions

Random variables and probability distributions

Gailmard 4.3

A **random variable** Y takes one of multiple possible (numerical) values depending on the outcome of an "experiment".

Random variables and probability distributions

Gailmard 4.3

A **random variable** Y takes one of multiple possible (numerical) values depending on the outcome of an "experiment".

Conventional notation: Y is the RV; y is a particular value.

Random variables and probability distributions

Gailmard 4.3

A **random variable** Y takes one of multiple possible (numerical) values depending on the outcome of an "experiment".

Conventional notation: Y is the RV; y is a particular value.

Gailmard 4.4

The probability distribution of a random variable Y can be summarized by

Random variables and probability distributions

Gailmard 4.3

A **random variable** Y takes one of multiple possible (numerical) values depending on the outcome of an "experiment".

Conventional notation: Y is the RV; y is a particular value.

Gailmard 4.4

The probability distribution of a random variable Y can be summarized by

- a **cumulative distribution function** (CDF) gives $\Pr(Y \leq y)$

Random variables and probability distributions

Gailmard 4.3

A **random variable** Y takes one of multiple possible (numerical) values depending on the outcome of an "experiment".

Conventional notation: Y is the RV; y is a particular value.

Gailmard 4.4

The probability distribution of a random variable Y can be summarized by

- a **cumulative distribution function** (CDF) gives $\Pr(Y \leq y)$
- (if discrete) a **probability mass function** (PMF) gives $\Pr(Y=y)$

Random variables and probability distributions

Gailmard 4.3

A **random variable** Y takes one of multiple possible (numerical) values depending on the outcome of an "experiment".

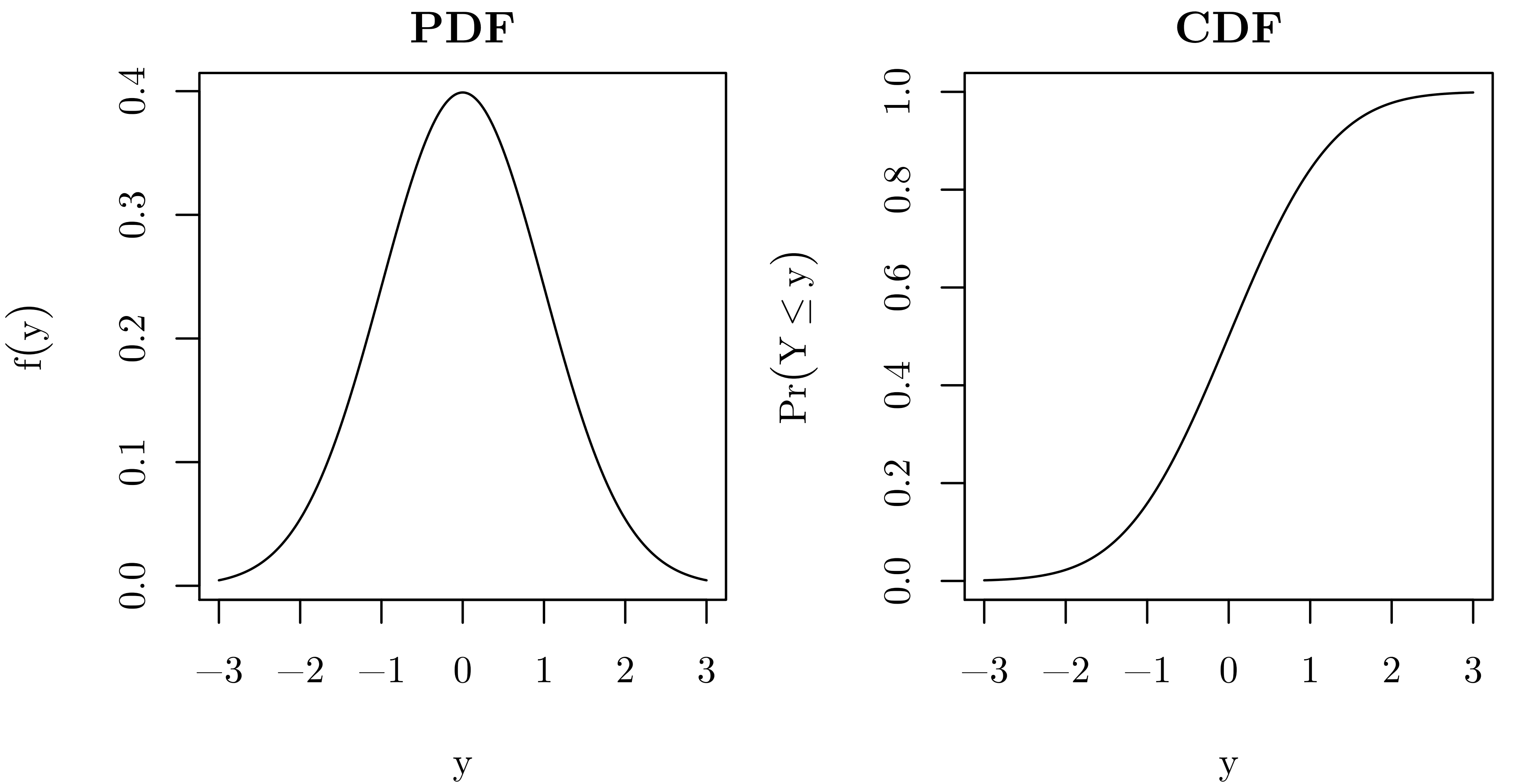
Conventional notation: Y is the RV; y is a particular value.

Gailmard 4.4

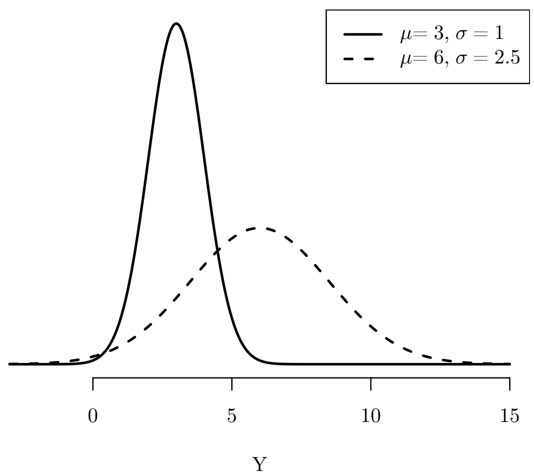
The probability distribution of a random variable Y can be summarized by

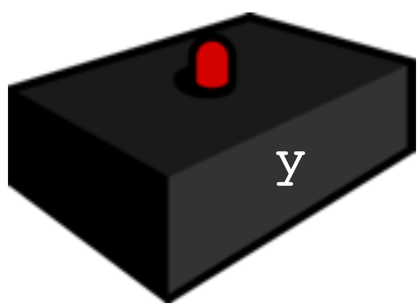
- a **cumulative distribution function** (CDF) gives $\Pr(Y \leq y)$
- (if discrete) a **probability mass function** (PMF) gives $\Pr(Y=y)$
- (if continuous) a **probability density function** (PDF) gives the derivative of the CDF at y

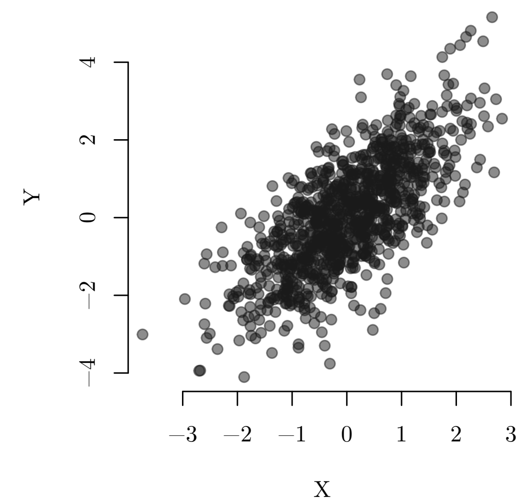
Normal PDF and CDF



How probability works

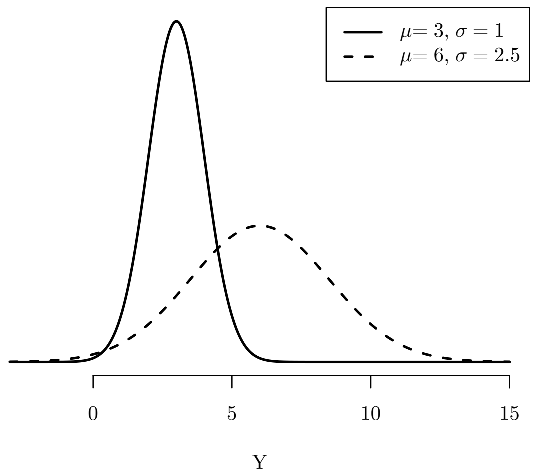


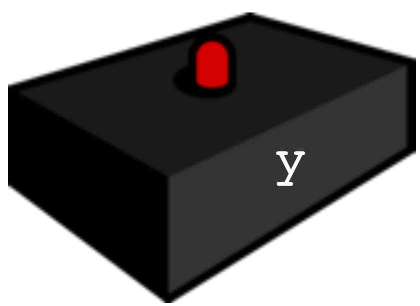


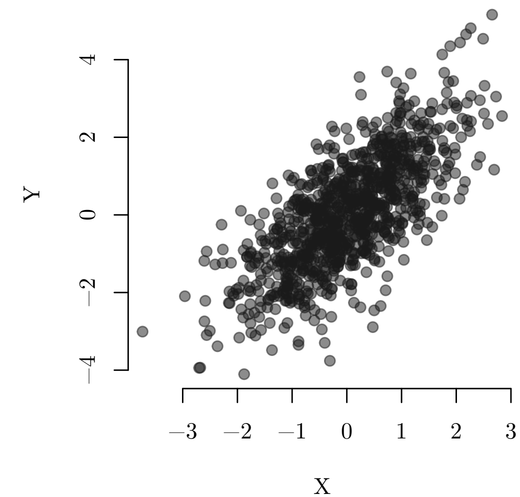


How probability works

Given a set of probability distributions…

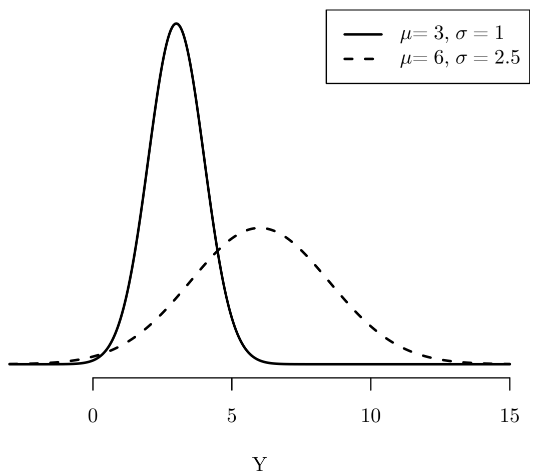




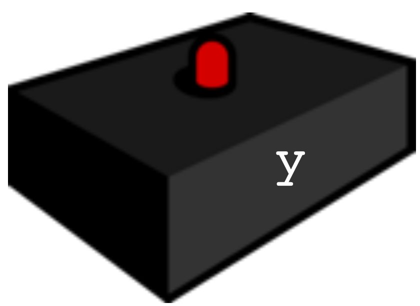


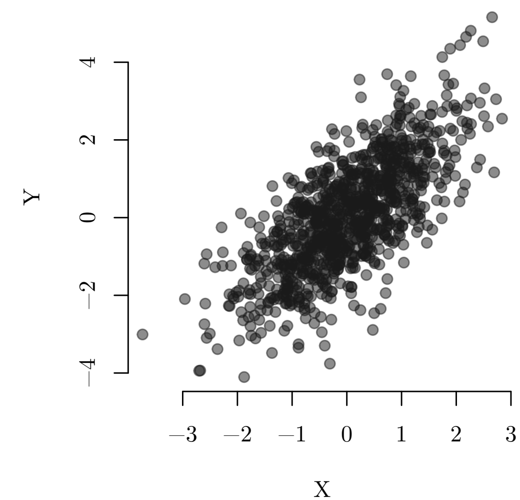
How probability works

Given a set of probability distributions…



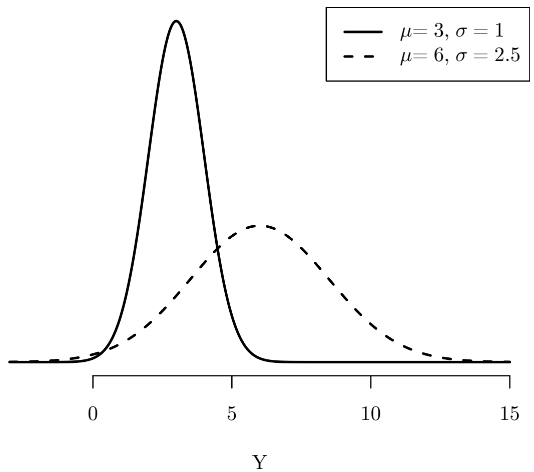
…that characterize a **data generating process** (DGP)…



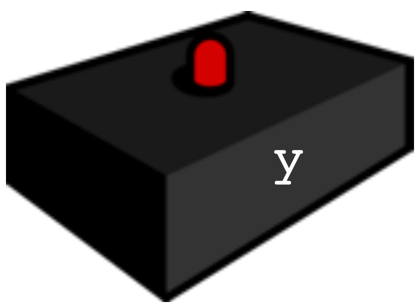


How probability works

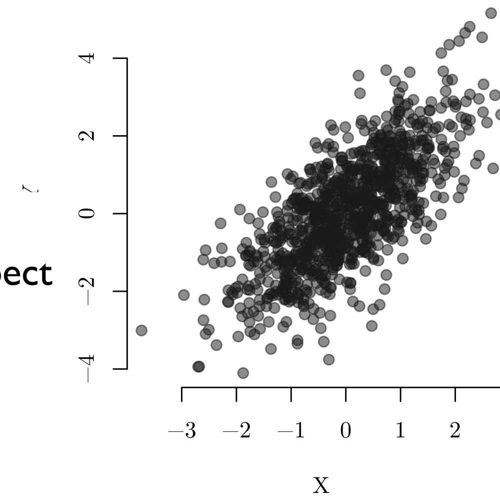
Given a set of probability distributions…



…that characterize a **data generating process** (DGP)…



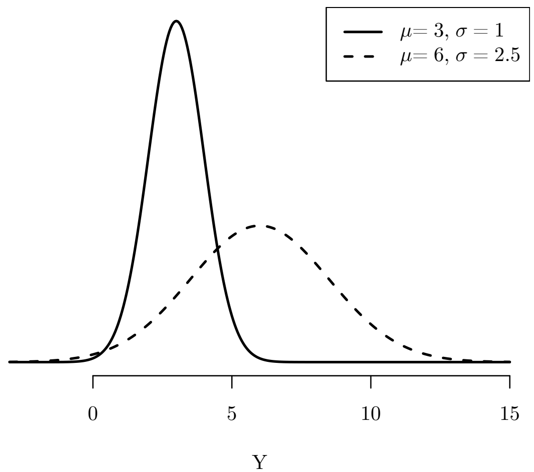
…what data should we expect to observe?



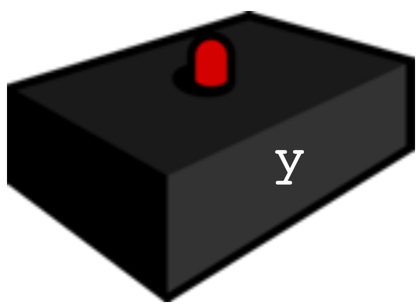
How probability works

Gailmard 4

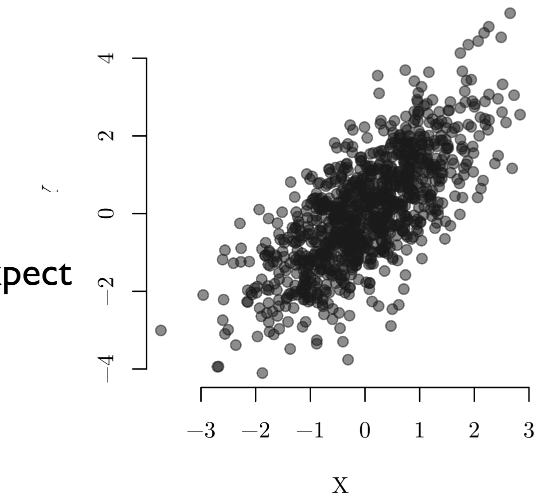
Given a set of probability distributions…



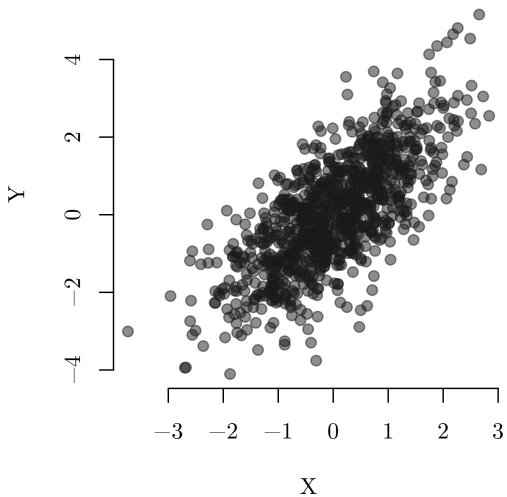
…that characterize a **data generating process** (DGP)…

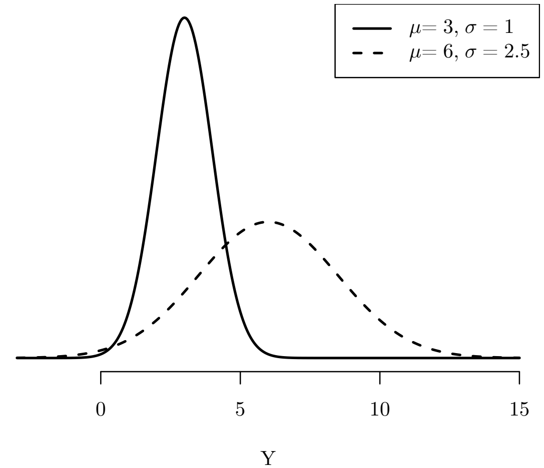


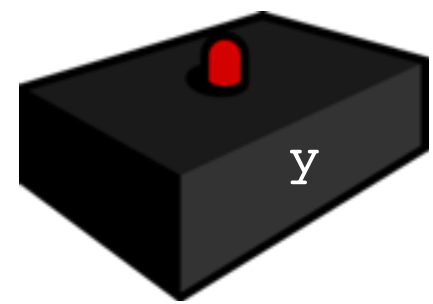
…what data should we expect to observe?



How (classical) statistics works

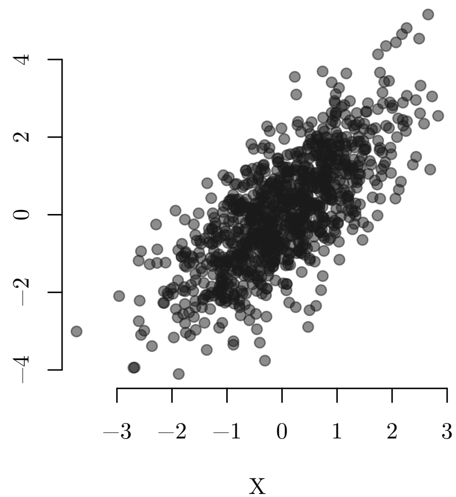


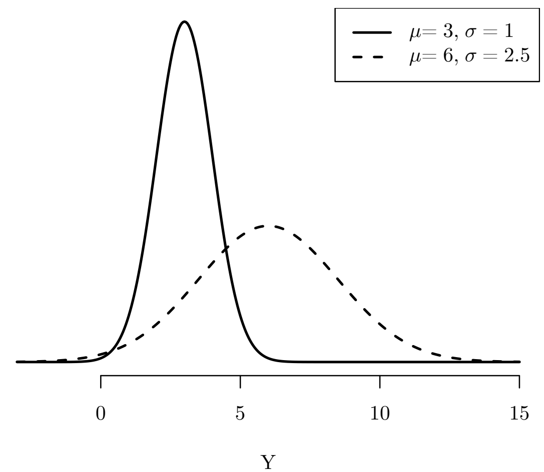


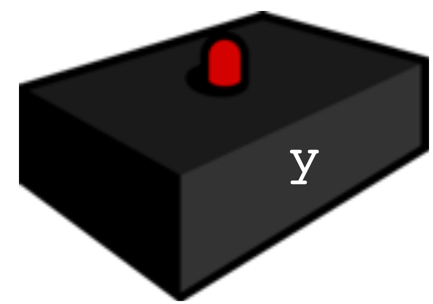


How (classical) statistics works

Given the data we observe…

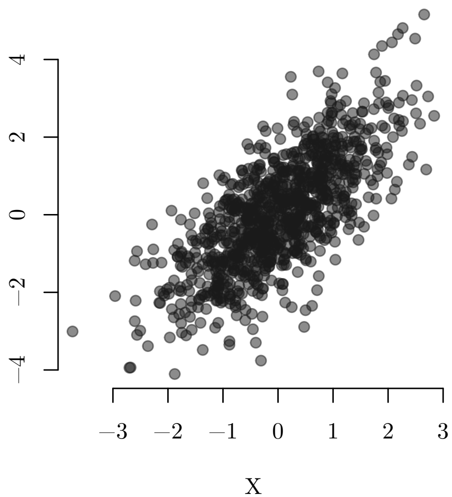




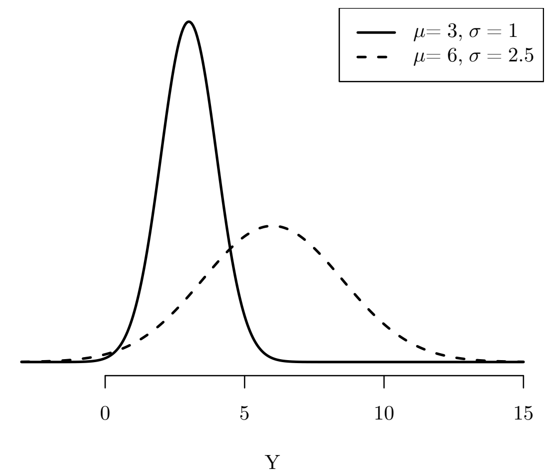


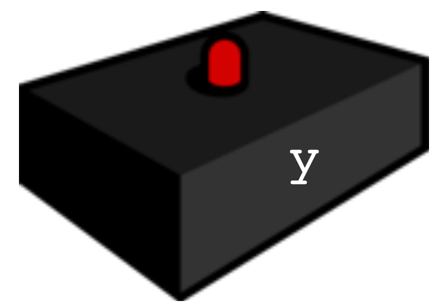
How (classical) statistics works

Given the data we observe…



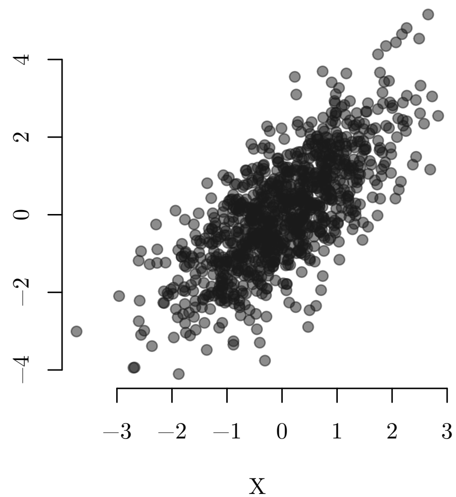
…what set of probability distributions…



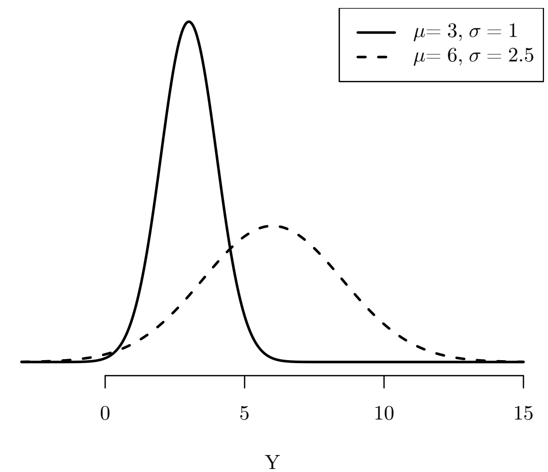


How (classical) statistics works

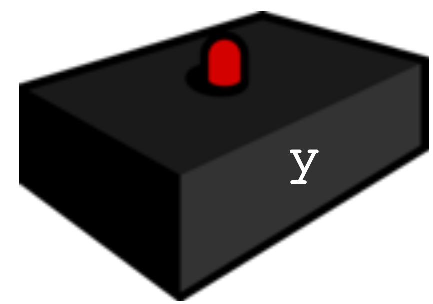
Given the data we observe…



…what set of probability distributions…



characterize the DGP?



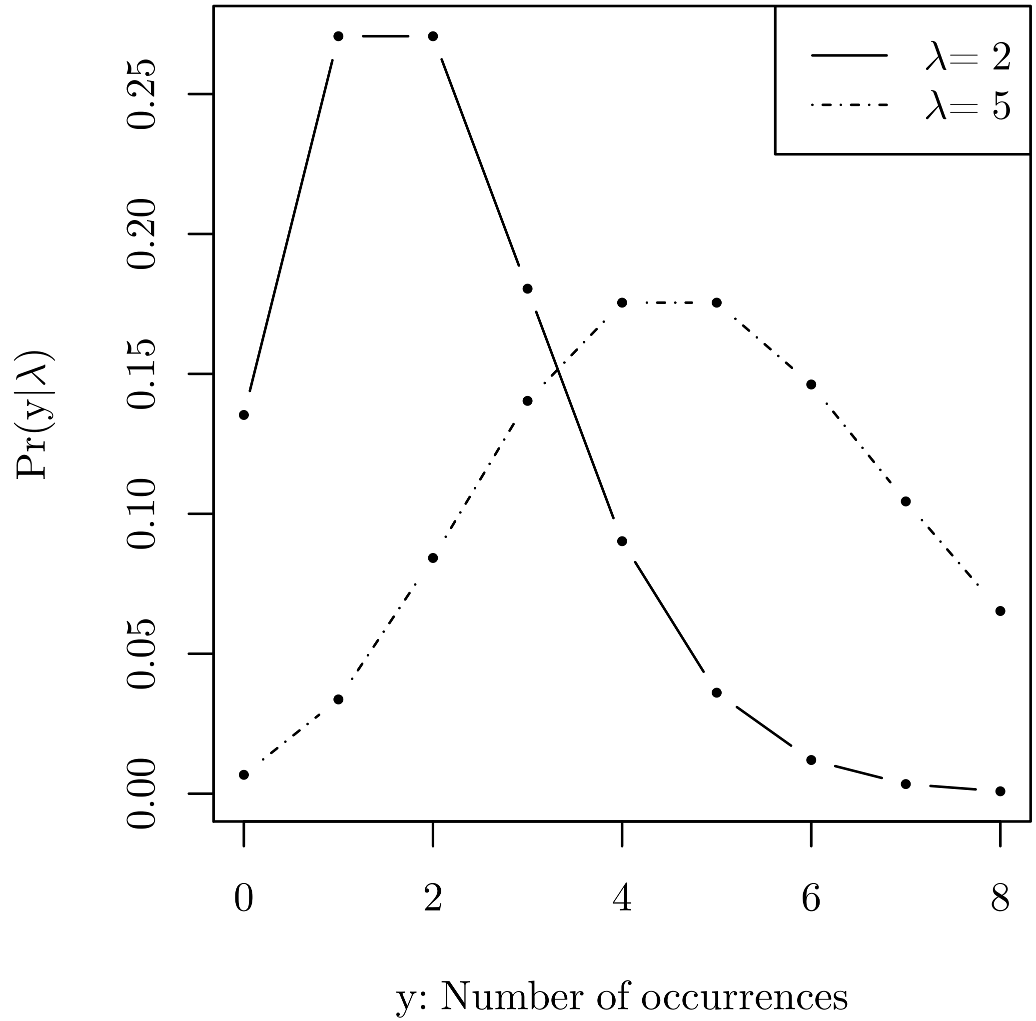
Poisson PMF

Poisson PMF

$$\Pr(Y = y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

Poisson PMF

$$\Pr(Y = y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

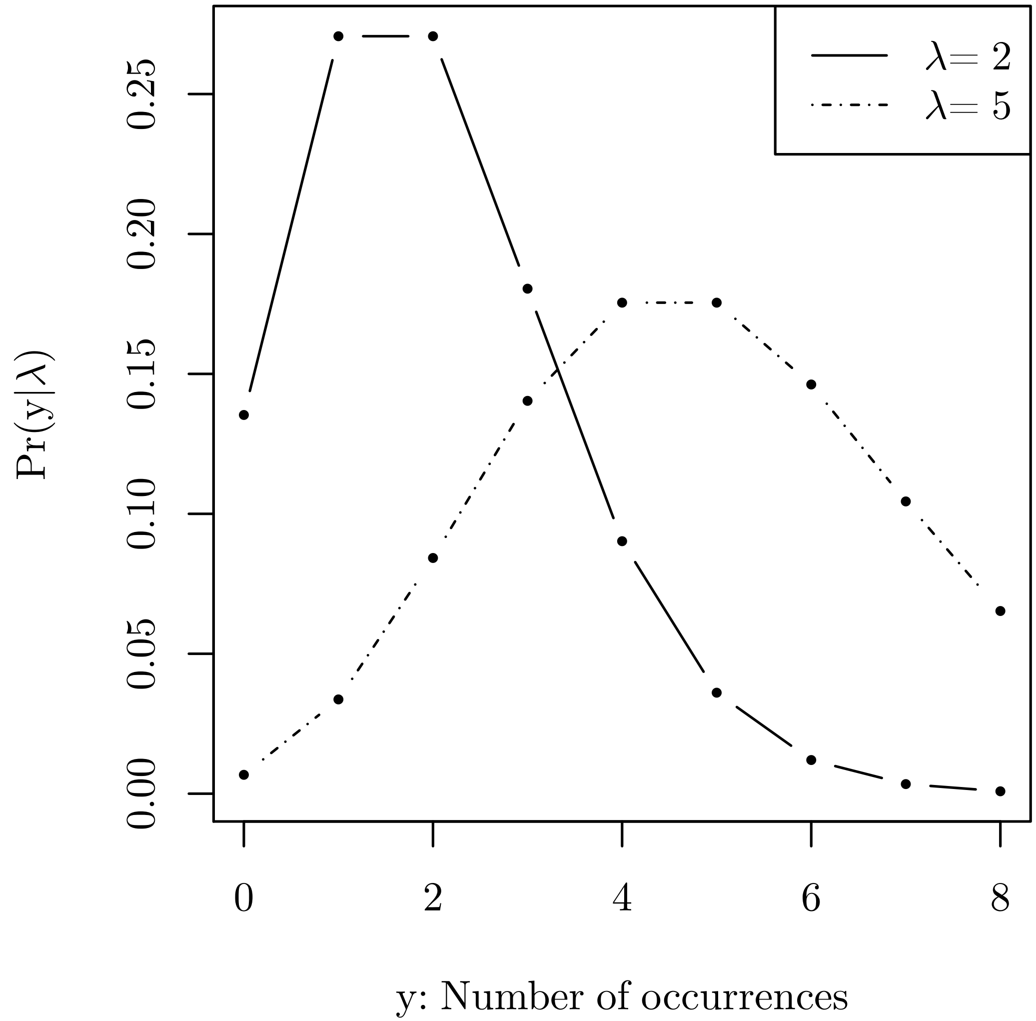


Poisson PMF

$$\Pr(Y = y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

Characterizes count of events (e.g. false convictions, horse kicks) observed in a fixed interval when

- events are independent
- rate of occurrence (probability per unit time) is constant (λ)



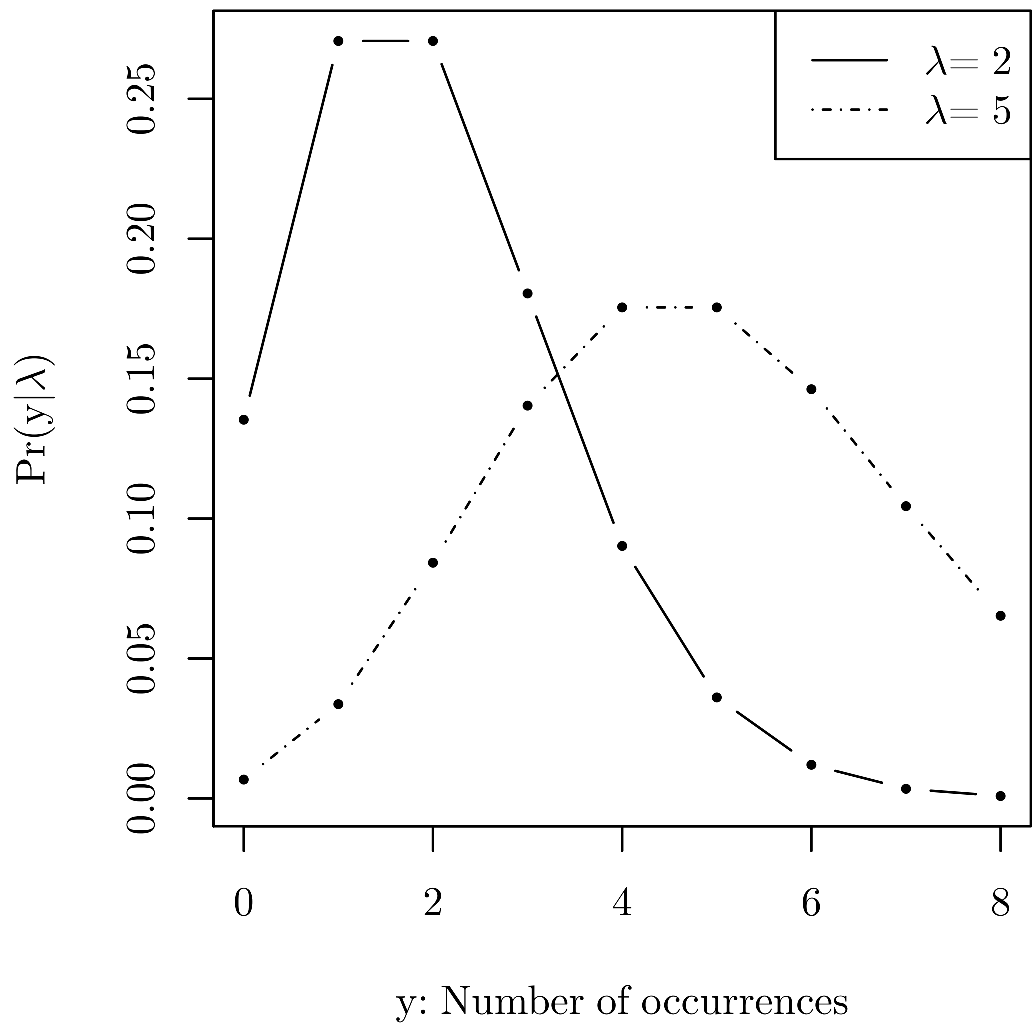
Poisson PMF

Gailmard 6.3

$$\Pr(Y = y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

Characterizes count of events (e.g. false convictions, horse kicks) observed in a fixed interval when

- events are independent
- rate of occurrence (probability per unit time) is constant (λ)



Single count

Single count

Suppose we view the number of students sitting in row 3 as a Poisson random variable.
(Reasonable?)

Single count

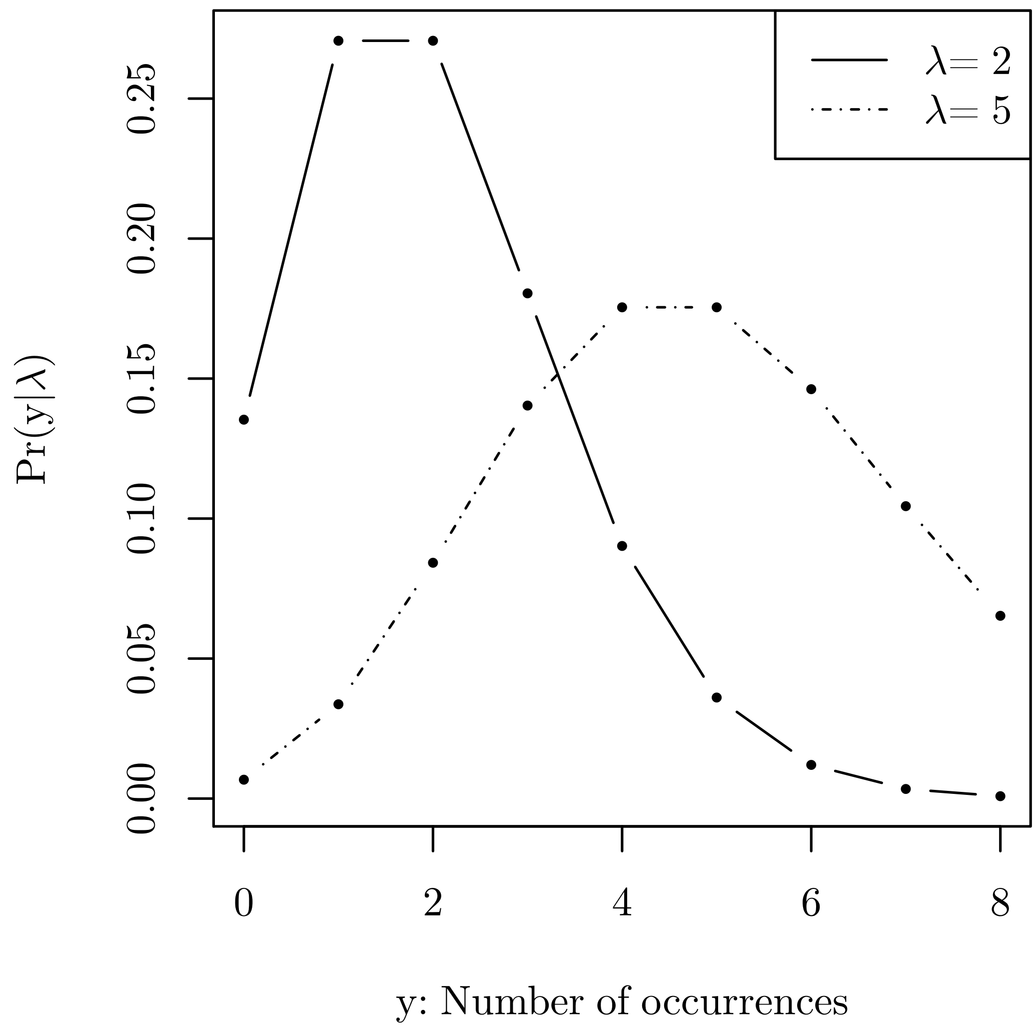
Suppose we view the number of students sitting in row 3 as a Poisson random variable.
(Reasonable?)

1) If $\lambda = 2$, how likely is the observed outcome?
2) If $\lambda = 5$, how likely is the observed outcome?

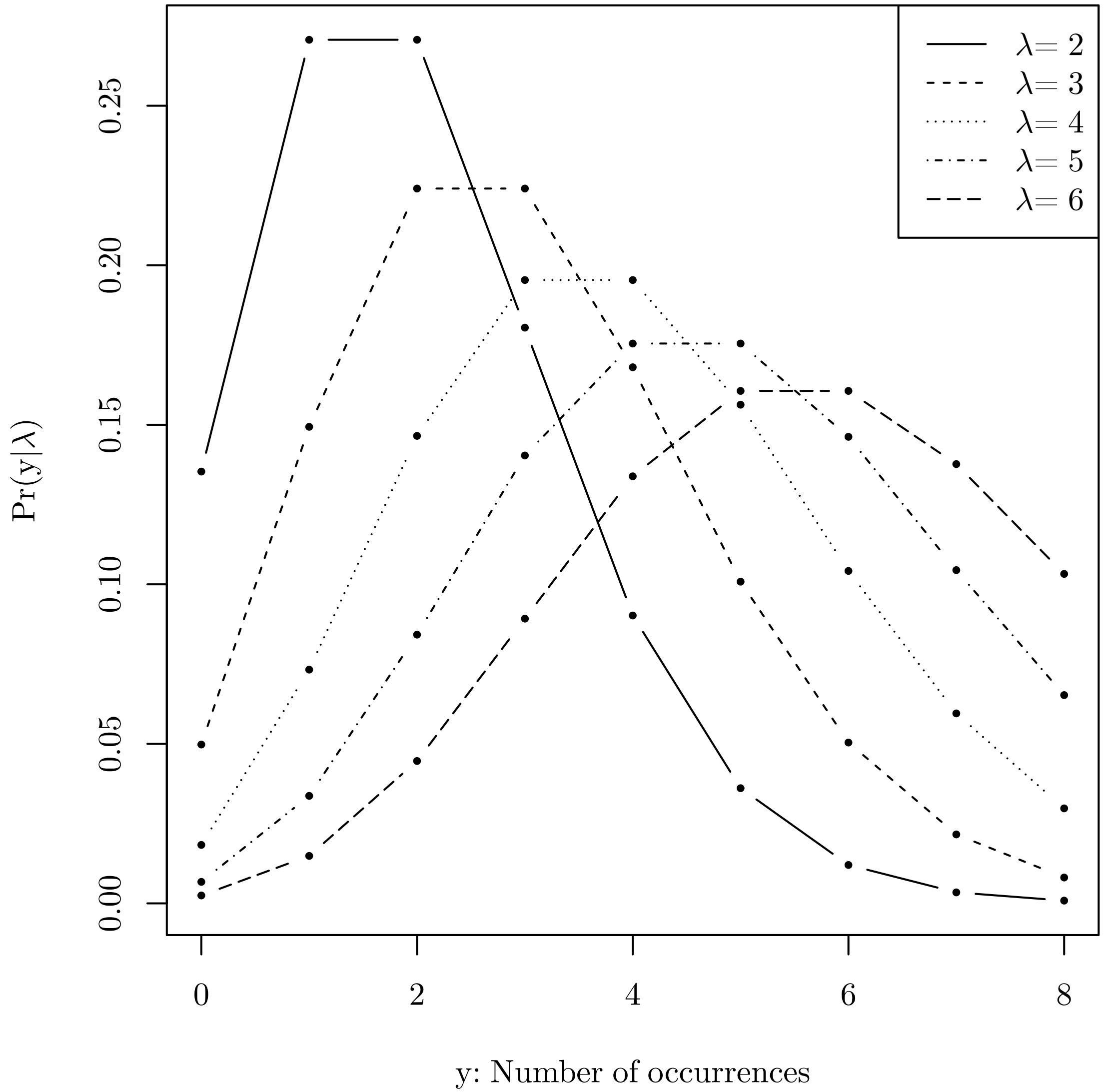
Single count

Suppose we view the number of students sitting in row 3 as a Poisson random variable.
(Reasonable?)

1) If λ = 2, how likely is the observed outcome?
2) If λ = 5, how likely is the observed outcome?



Single count (2)

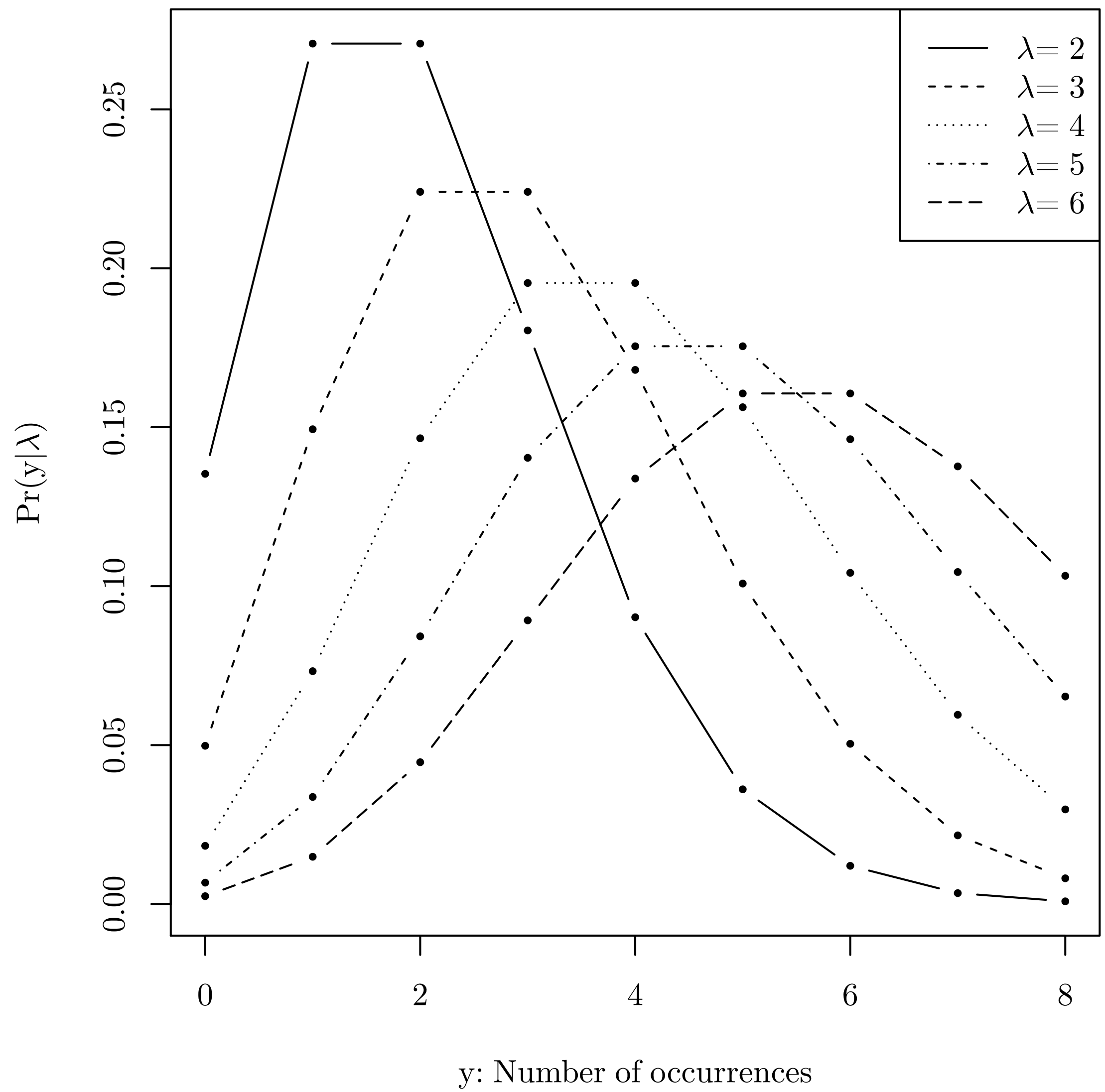


Single count (2)

Suppose we view the number of students sitting in row 3 as a Poisson random variable.

For what value of λ is the observed outcome most likely?

This is the most basic illustration of Maximum Likelihood Estimation (MLE) for λ.



Joint & conditional probability and independence

Joint & conditional probability and independence

For two events E and F, the probability of both events happening is written

$$\mathrm{P}(E, F) \qquad \text{or} \qquad P(E \cap F)$$

joint probability

Joint & conditional probability and independence

For two events E and F, the probability of both events happening is written

$$\mathrm{P}(E, F) \qquad \text{or} \qquad P(E \cap F)$$

joint probability

The probability of E happening given F is written

$$\mathrm{P}(E|F)$$

conditional probability

Joint & conditional probability and independence

For two events E and F, the probability of both events happening is written

$$\mathrm{P}(E, F) \qquad \text{or} \qquad P(E \cap F)$$

joint probability

The probability of E happening given F is written

$$\mathrm{P}(E|F)$$

conditional probability

If E and F are independent,

$$\mathrm{P}(E|F) = \mathrm{P}(E)$$

Joint & conditional probability and independence

For two events E and F, the probability of both events happening is written

$$P(E, F) \qquad \text{or} \qquad P(E \cap F)$$

joint probability

The probability of E happening given F is written

$$P(E|F)$$

conditional probability

If E and F are independent,

$$P(E|F) = P(E)$$

and:

$$P(E, F) = P(E) \times P(F)$$

Vector of counts

Vector of counts

Suppose we view the number of students sitting **in each row** as an independent Poisson random variable. (Reasonable?)

Vector of counts

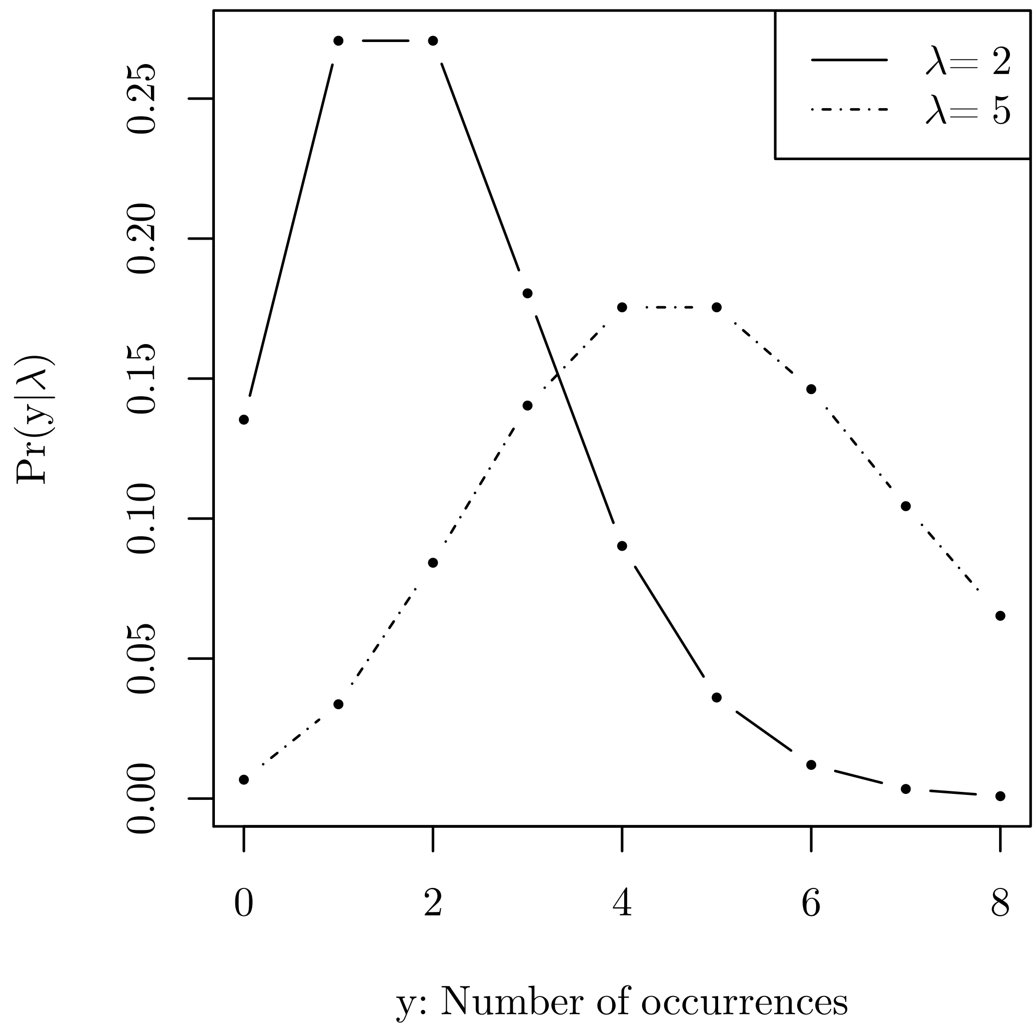
Suppose we view the number of students sitting **in each row** as an independent Poisson random variable. (Reasonable?)

1) If $\lambda = 2$, how likely is the observed outcome for rows 3-6?
2) If $\lambda = 5$, how likely is the observed outcome for rows 3-6?

Vector of counts

Suppose we view the number of students sitting **in each row** as an independent Poisson random variable. (Reasonable?)

1) If λ = 2, how likely is the observed outcome for rows 3-6?
2) If λ = 5, how likely is the observed outcome for rows 3-6?



Vector of counts

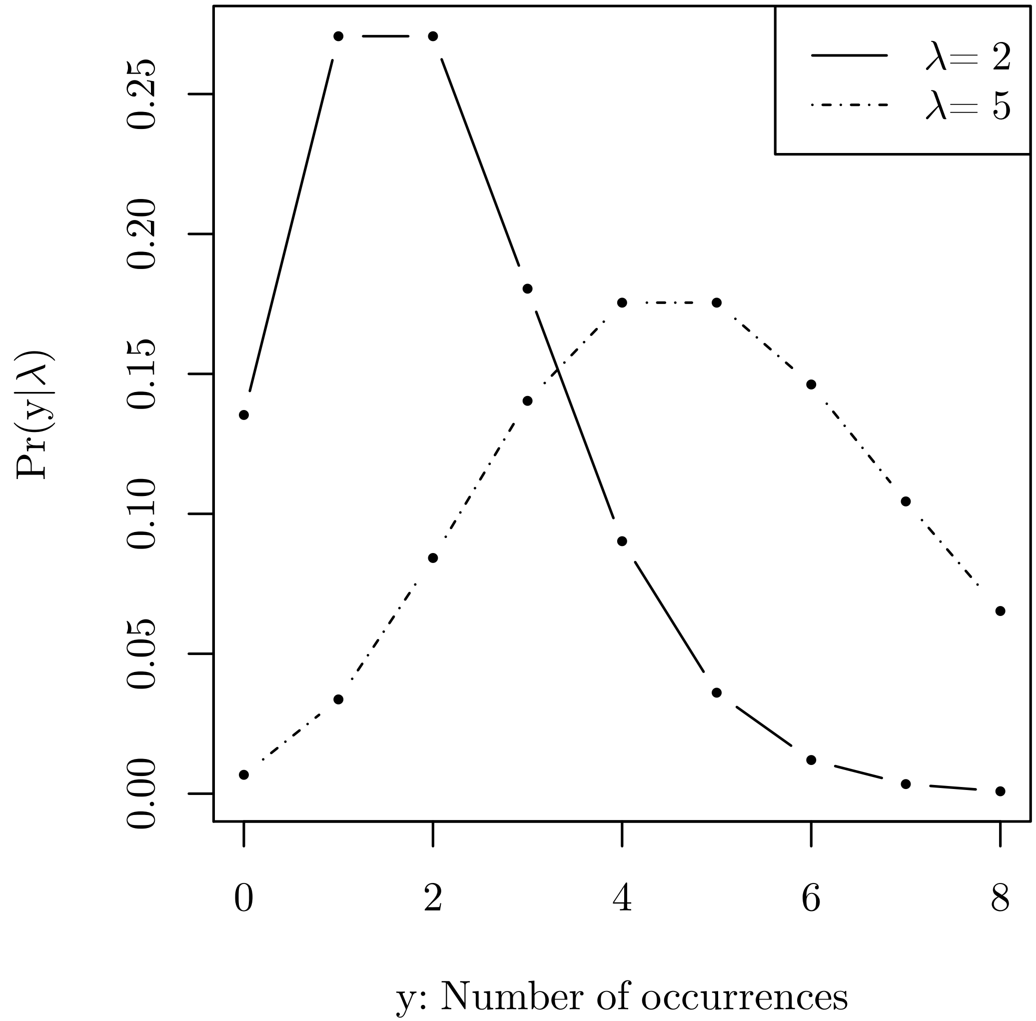
Suppose we have 5, 2, 7, 4 students in these rows.

	λ = 2	λ = 5
5	0.04	0.18
2	0.27	0.08
7	0.003	0.10
4	0.10	0.18
Likelihood = column product	3.03/1M	270.84/1M

Vector of counts

Suppose we have 5, 2, 7, 4 students in these rows.

	λ = 2	λ = 5
5	0.04	0.18
2	0.27	0.08
7	0.003	0.10
4	0.10	0.18
Likelihood = column product	3.03/1M	270.84/1M



Vector of counts (continued)

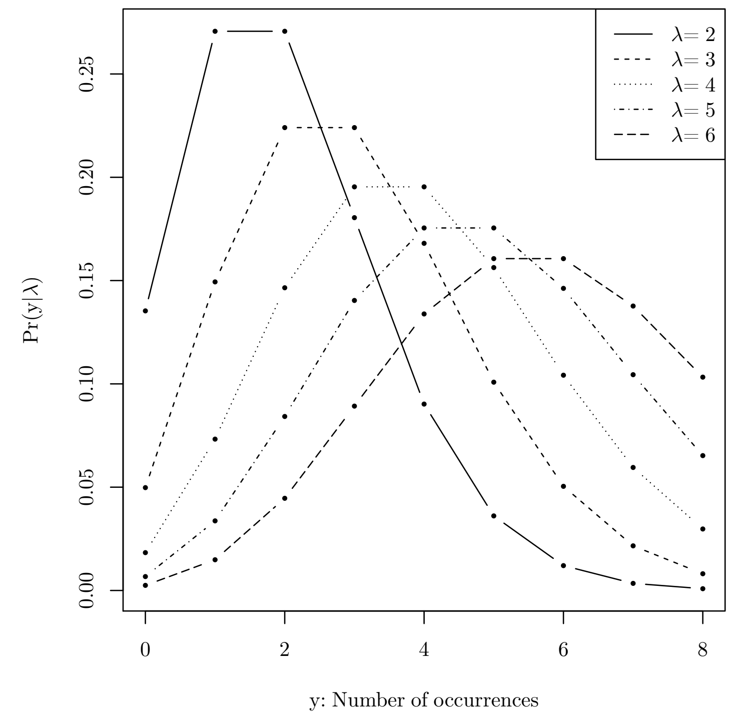
We can of course try this for more values of λ:

	λ = 2	λ = 3	λ = 4	λ = 5	λ = 6
5	0.04	0.10	0.16	0.18	0.16
2	0.27	0.22	0.15	0.08	0.04
7	0.003	0.02	0.06	0.10	0.14
4	0.10	0.17	0.20	0.18	0.13
Likelihood = column product (× 1M)	3.03	82.00	266.39	270.84	132.07

Vector of counts (continued)

We can of course try this for more values of λ:

	λ = 2	λ = 3	λ = 4	λ = 5	λ = 6
5	0.04	0.10	0.16	0.18	0.16
2	0.27	0.22	0.15	0.08	0.04
7	0.003	0.02	0.06	0.10	0.14
4	0.10	0.17	0.20	0.18	0.13
Likelihood = column product (× 1M)	3.03	82.00	266.39	270.84	132.07



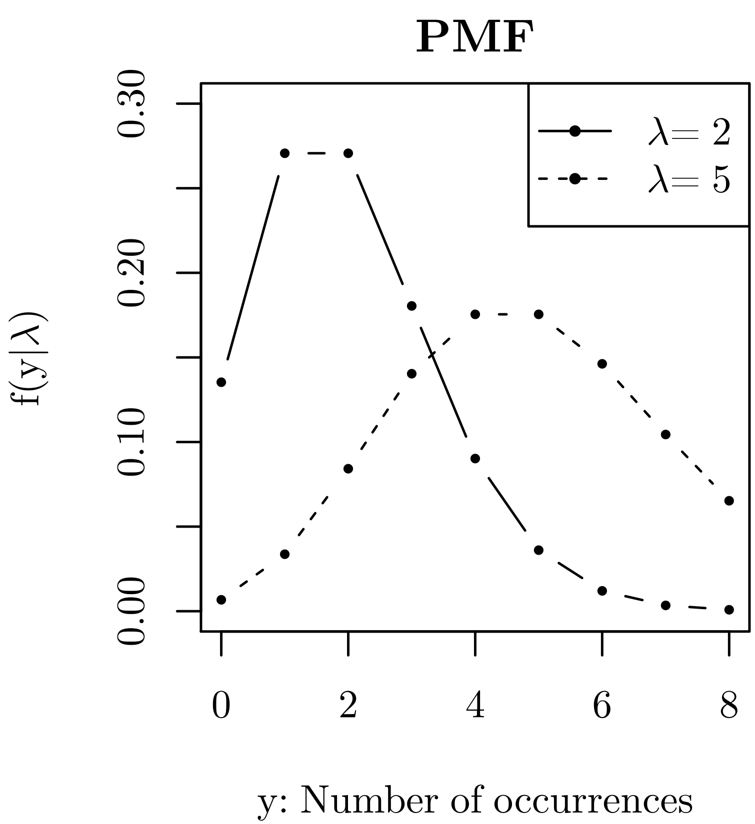
PMF vs Likelihood

PMF vs Likelihood

The pmf can be written $f(y|\lambda)$: a function of y (the observed data) whose shape depends on λ (the parameter).

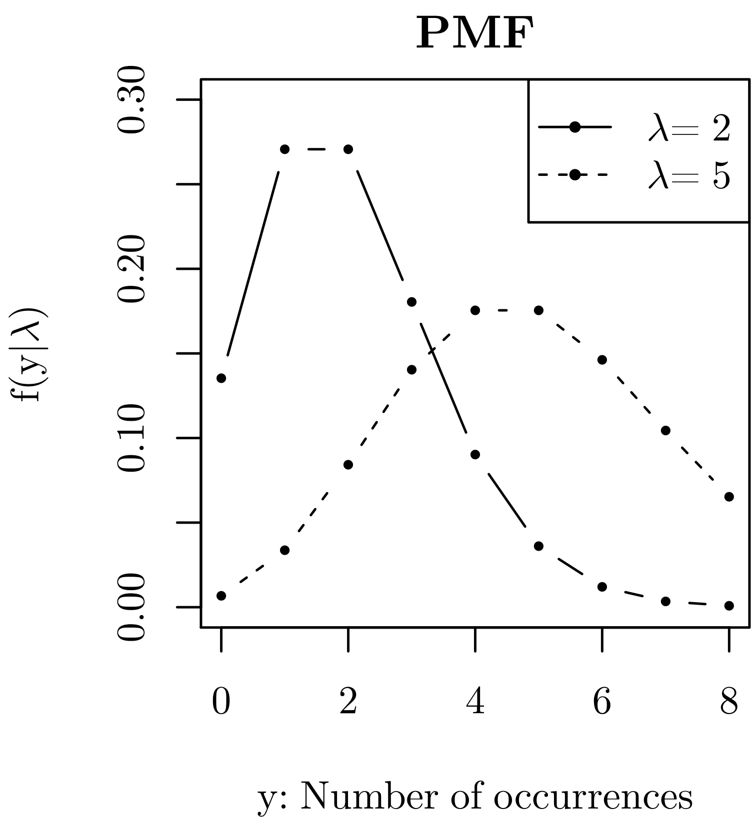
PMF vs Likelihood

The pmf can be written f(y|λ): a function of y (the observed data) whose shape depends on λ (the parameter).



PMF vs Likelihood

The pmf can be written f(y|λ): a function of y (the observed data) whose shape depends on λ (the parameter).

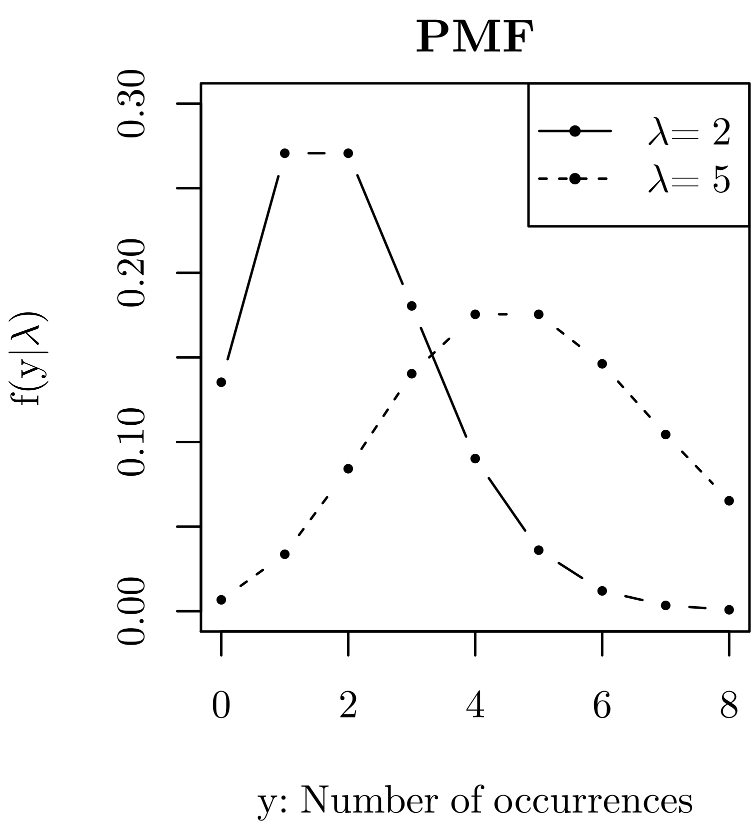


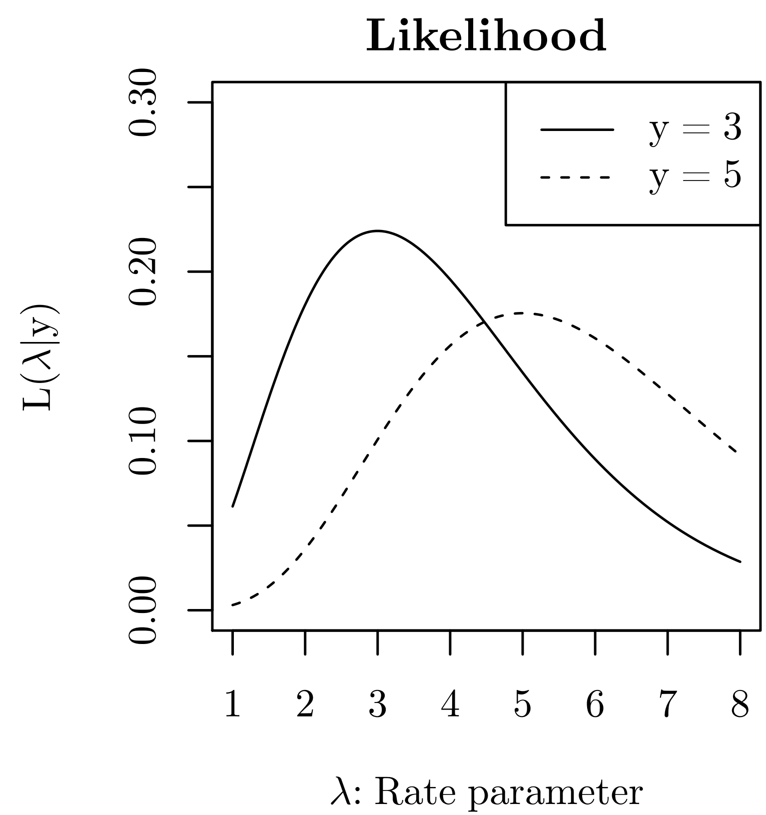
But from many of these pmfs we can derive L(λ|y): a function of λ (the parameter) whose shape depends on y (the observed data).

PMF vs Likelihood

The pmf can be written f(y|λ): a function of y (the observed data) whose shape depends on λ (the parameter).

But from many of these pmfs we can derive L(λ|y): a function of λ (the parameter) whose shape depends on y (the observed data).





Maximum likelihood

Maximum likelihood

Using θ to refer to parameters, consider:

$$\hat{\theta}(\mathbf{y}) = \mathrm{argmax}_{\theta} L(\theta|\mathbf{y})$$

Maximum likelihood

Using θ to refer to parameters, consider:

$$\hat{\theta}(\mathbf{y}) = \mathrm{argmax}_\theta L(\theta|\mathbf{y})$$

The **maximum likelihood estimate** (MLE) is the θ that makes the observed data (y) most likely.

Maximum likelihood

Using θ to refer to parameters, consider:

$$\hat{\theta}(\mathbf{y}) = \text{argmax}_{\theta} L(\theta|\mathbf{y})$$

The **maximum likelihood estimate** (MLE) is the θ that makes the observed data (y) most likely.

A general approach to statistical modeling:

- write down f(y|θ) (pdf/pmf: probability of outcomes conditional on parameters), which is also L(θ|y)
- observe data (y: actual outcomes)
- find parameters that maximize L(θ|y): the MLE!

Maximum likelihood (common notation)

$$\begin{aligned}\mathcal{L}(\theta|\mathbf{Y}) &= f(y_1, y_2, \ldots, y_n|\theta) \\ &= f(y_1|\theta)f(y_2|\theta)\ldots f(y_n|\theta) \\ &= \prod_{i=1}^{n} f(y_i|\theta)\end{aligned}$$

Maximum likelihood (common notation)

$$\begin{aligned}\mathcal{L}(\theta|\mathbf{Y}) &= f(y_1, y_2, \ldots, y_n|\theta) \\ &= f(y_1|\theta)f(y_2|\theta)\ldots f(y_n|\theta) \\ &= \prod_{i=1}^{n} f(y_i|\theta)\end{aligned}$$

iid assumption

Vector of counts with a covariate

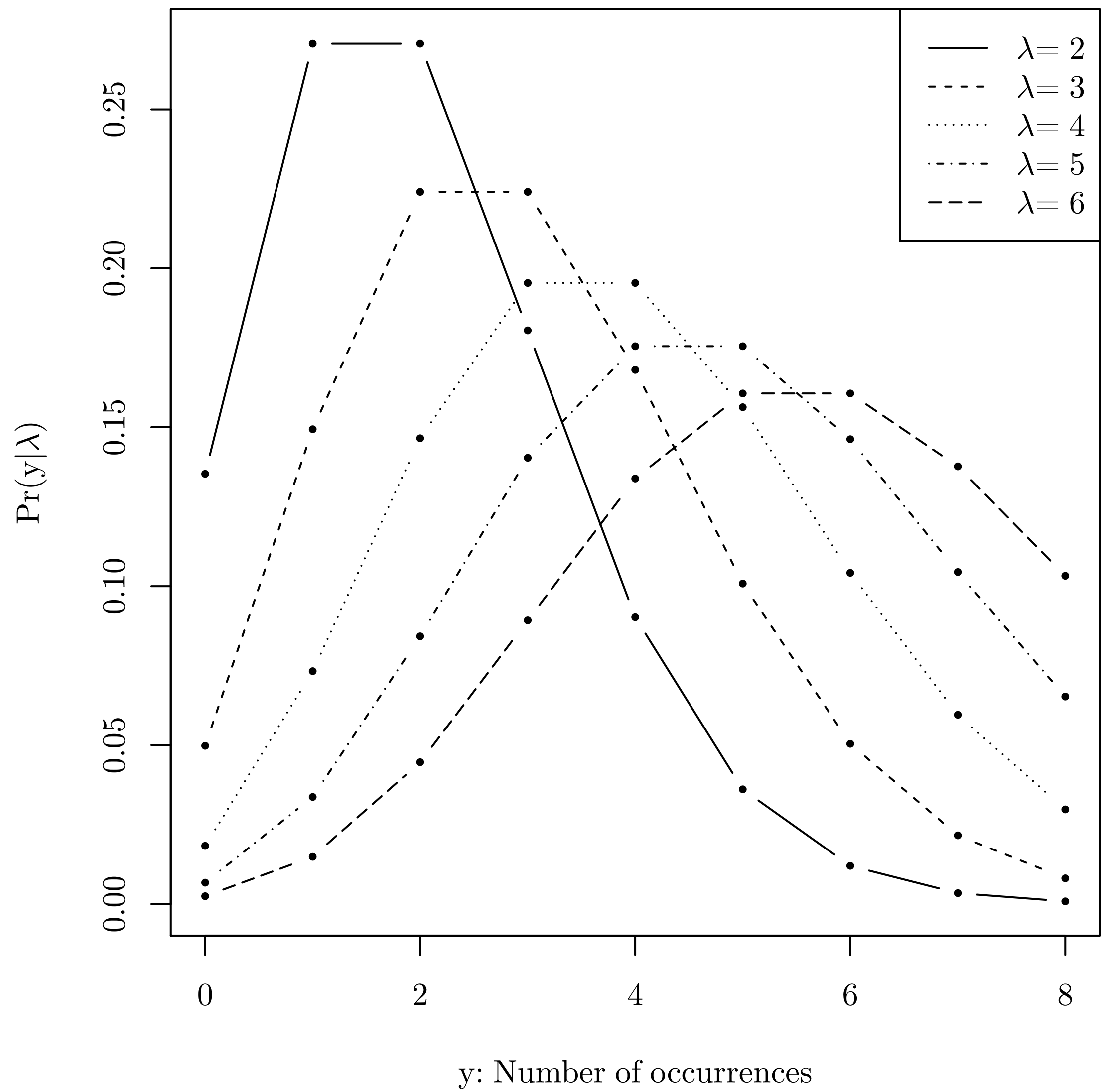
Vector of counts with a covariate

Suppose we view the number of students sitting **in each row** as an independent Poisson random variable, with $\lambda = x_i$, where x_i is the number of the row.

Vector of counts with a covariate

Suppose we view the number of students sitting **in each row** as an independent Poisson random variable, with λ = x_i, where x_i is the number of the row.

How likely is the observed outcome for rows 3-6?



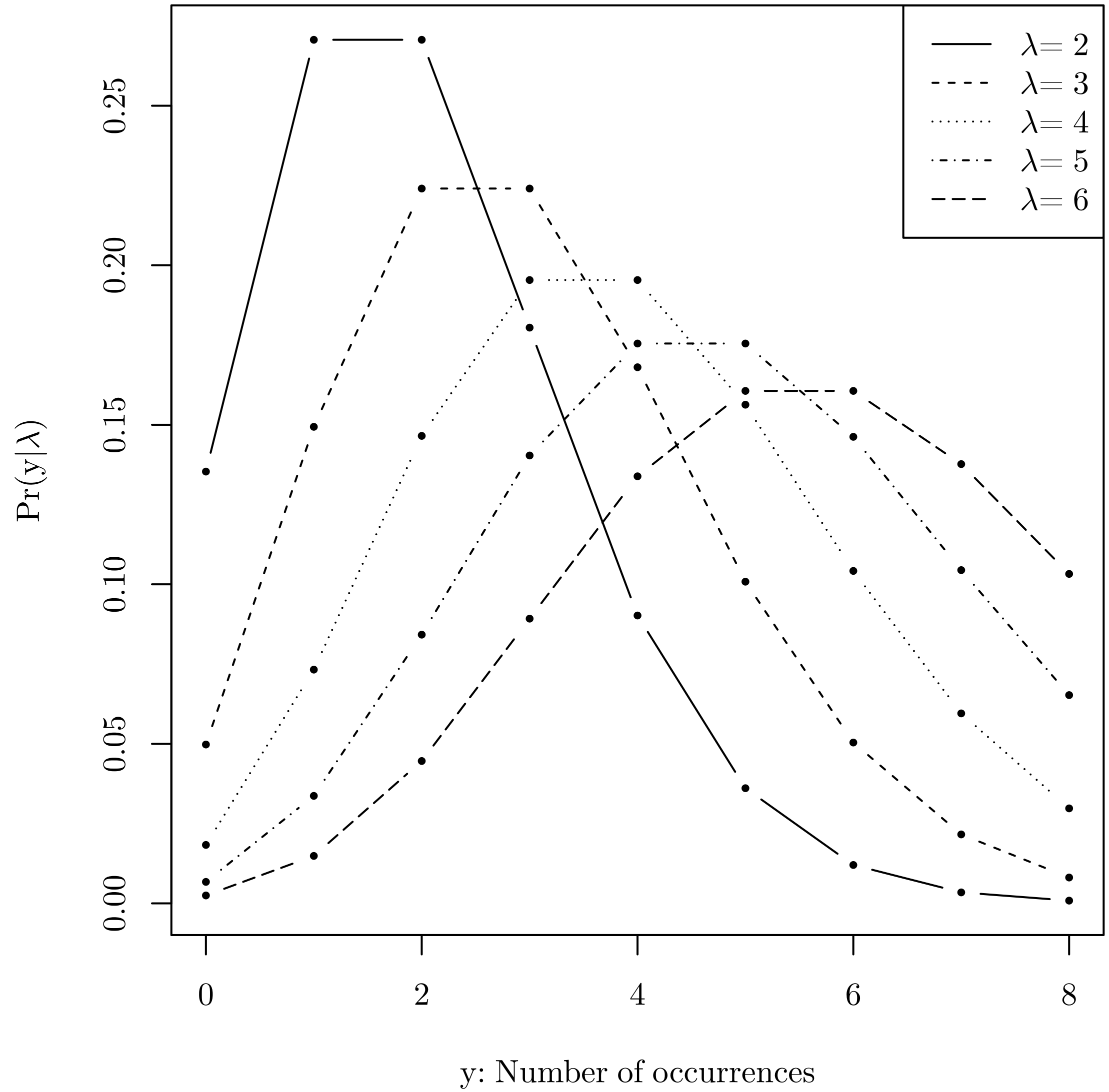
Vector of counts with a covariate

Vector of counts with a covariate

Suppose we observe
5, 2, 7, and 4 students.

Vector of counts with a covariate

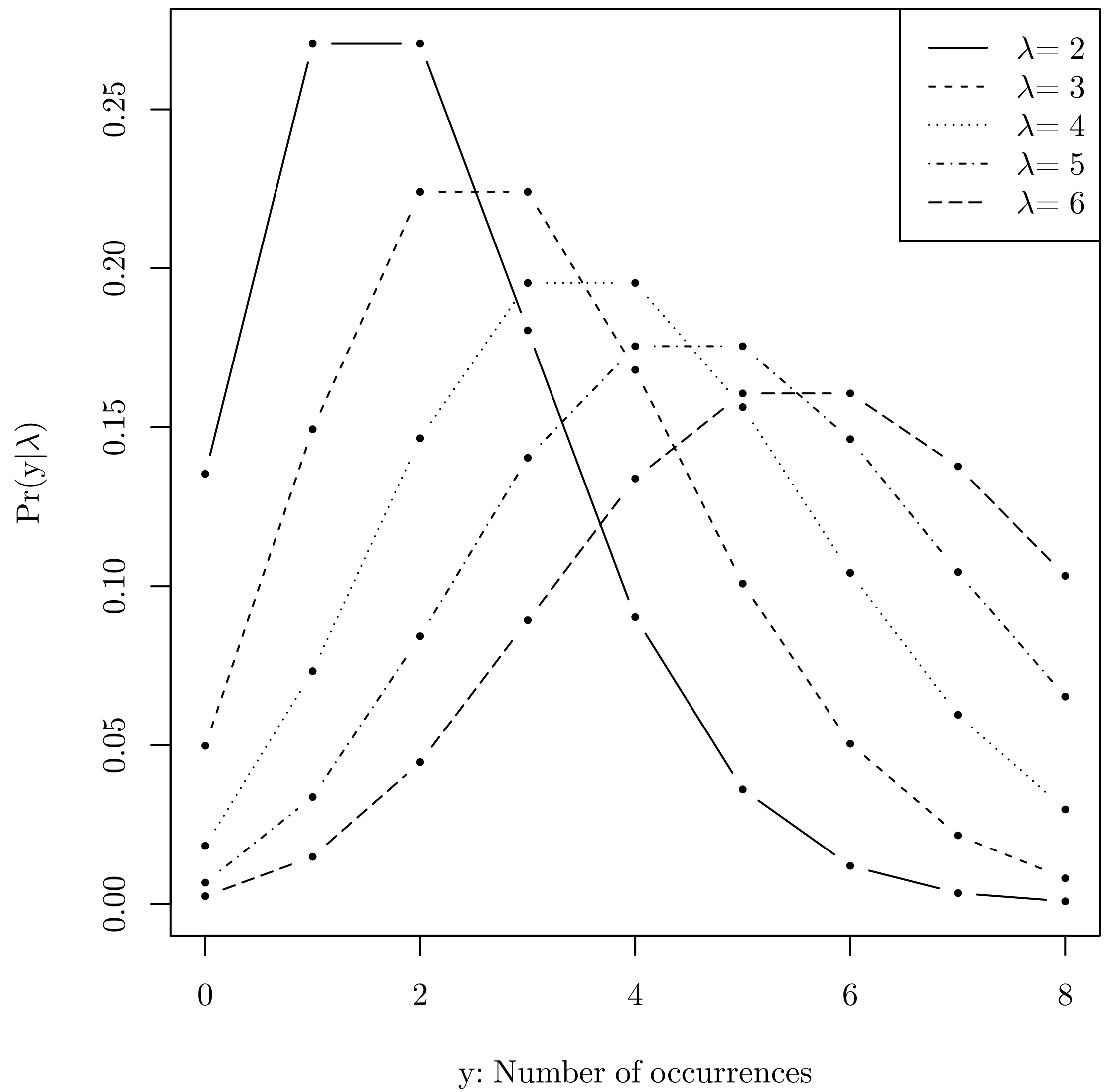
Suppose we observe 5, 2, 7, and 4 students.



Vector of counts with a covariate

Suppose we observe 5, 2, 7, and 4 students.

row	# students	λ = row
3	5	0.1
4	2	0.15
5	7	0.10
6	4	0.13
	Likelihood = column product (× 1M)	206.52

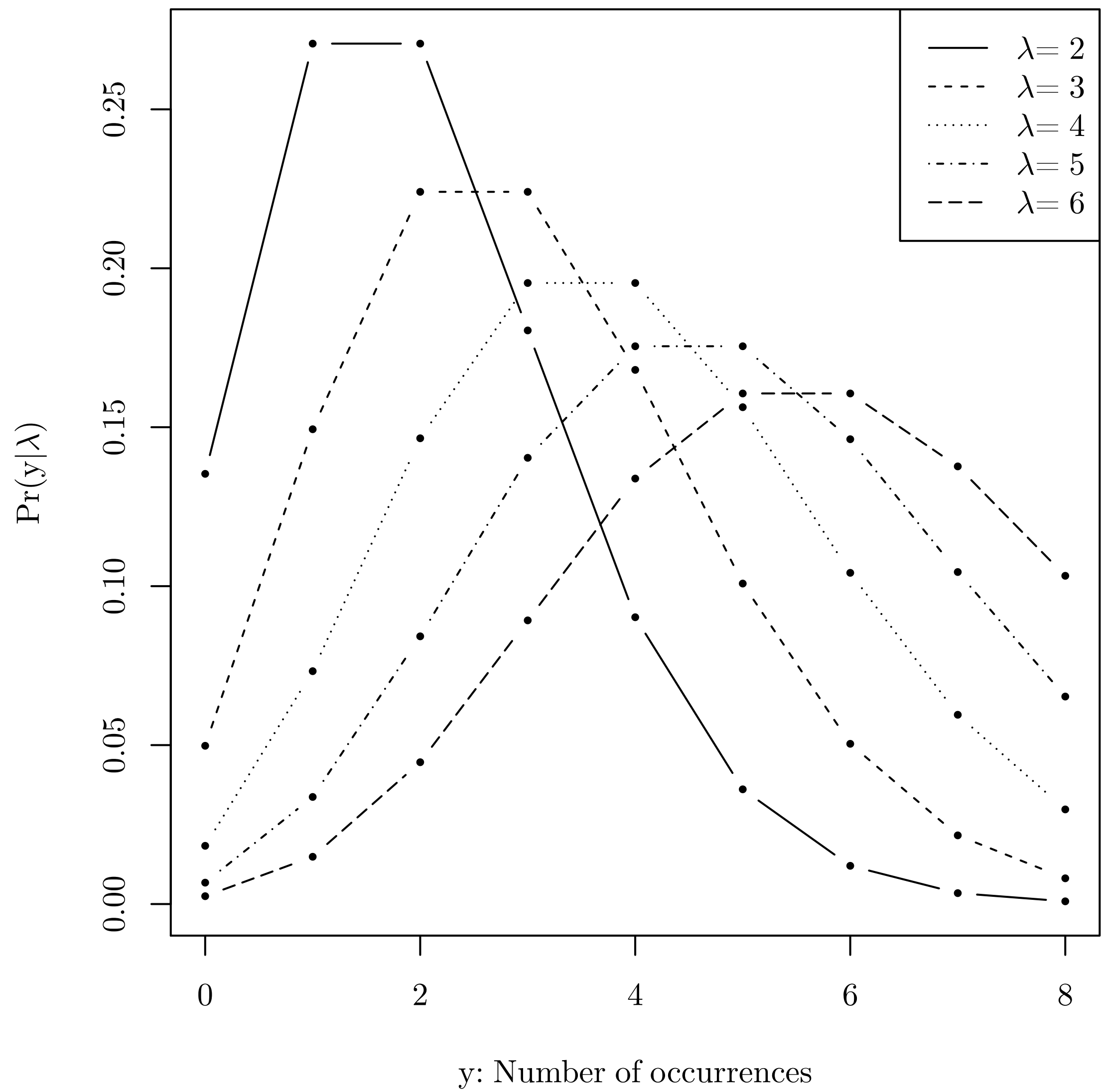


Vector of counts with a covariate

Suppose we observe 5, 2, 7, and 4 students.

row	# students	λ = row
3	5	0.1
4	2	0.15
5	7	0.10
6	4	0.13
	Likelihood = column product (× 1M)	206.52

Now suppose $\lambda = \beta_0 + \beta_1 \times \text{row}$, and find β_0, β_1 that maximize the likelihood.



Maximum likelihood (common notation)

$$\begin{aligned}
\mathcal{L}(\theta|\mathbf{Y}) &= f(y_1, y_2, \ldots, y_n|\theta) \\
&= f(y_1|\theta) f(y_2|\theta) \ldots f(y_n|\theta) \\
&= \prod_{i=1}^{n} f(y_i|\theta)
\end{aligned}$$

Maximum likelihood (common notation)

$$\begin{aligned}\mathcal{L}(\theta|\mathbf{Y}) &= f(y_1, y_2, \ldots, y_n|\theta) \\ &= f(y_1|\theta)f(y_2|\theta)\ldots f(y_n|\theta) \\ &= \prod_{i=1}^{n} f(y_i|\theta)\end{aligned}$$

iid assumption

Maximum likelihood (common notation)

$$
\begin{aligned}
\mathcal{L}(\theta|\mathbf{Y}) &= f(y_1, y_2, \ldots, y_n|\theta) \\
&= f(y_1|\theta)f(y_2|\theta)\ldots f(y_n|\theta) \\
&= \prod_{i=1}^{n} f(y_i|\theta)
\end{aligned}
$$

iid assumption

The likelihood function for the last MLE problem you just solved was:

Maximum likelihood (common notation)

$$
\begin{aligned}
\mathcal{L}(\theta|\mathbf{Y}) &= f(y_1, y_2, \ldots, y_n|\theta) \\
&= f(y_1|\theta) f(y_2|\theta) \ldots f(y_n|\theta) \\
&= \prod_{i=1}^{n} f(y_i|\theta)
\end{aligned}
$$

iid assumption

The likelihood function for the last MLE problem you just solved was:

$$
\begin{aligned}
\mathcal{L}(\theta|\mathbf{y}) &= f(y_3, y_4, y_5, y_6|\theta) \\
&= f(y_3|\theta) f(y_4|\theta) f(y_5|\theta) f(y_6|\theta) \\
&= \prod_{i=3}^{6} f(y_i|\theta) \\
&= \prod_{i=3}^{6} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} = \prod_{i=3}^{6} \frac{(\beta_0 + x_i\beta_1)^{y_i} e^{-\beta_0 - x_i\beta_1}}{y_i!}
\end{aligned}
$$

How statistical models look in research papers

A Statistical Method for Empirical Testing of Competing Theories

Kosuke Imai Princeton University
Dustin Tingley Harvard University

the model specified in equation (1) yields the following observed-data likelihood function where the latent variable Z_i has been integrated out,

$$L_{obs}(\Theta, \Pi | \{X_i, Y_i\}_{i=1}^{N}) = \prod_{i=1}^{N} \left\{ \sum_{m=1}^{M} \pi_m f_m(Y_i | X_i, \theta_m) \right\}. \quad (2)$$

Comparing Interest Group Scores across Time and Chambers: Adjusted ADA Scores for the U.S. Congress

TIM GROSECLOSE *Stanford University*
STEVEN D. LEVITT *University of Chicago*
and JAMES M. SNYDER, JR. *Massachusetts*

Given this representation, we can estimate a_t^c's, b_t^c's, and x_i's by maximizing the following likelihood function:

$$L(\bar{a}, \bar{b}, \bar{x}, \sigma; \bar{y}) = \prod_{t \in T} \prod_{c \in \{H,S\}} \prod_{i \in I_t^c} \phi\left(\frac{y_{it} - a_t^c - b_t^c x_i}{\sigma}\right) \frac{1}{\sigma},$$

How statistical models look in research papers

Ideology and Interests in the Political Marketplace

Adam Bonica Stanford University

Assuming independence across candidates and contributors, the log-likelihood to be maximized is,

$$LL(Y|\lambda,\sigma) = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{t=1}^{T}\sum_{g=0}^{1}(1-d_{ijt_g})\, ln\,(NB \times (y_{ijt_g}|\lambda_{ijt_g},\sigma_{it_g})) + (d_{ijt_g})\; ln\left(1-\sum_{k=0}^{9} NB(k|\lambda_{ijt_g},\sigma_{it_g})\right) \quad (3.3)$$

where Y is an $n \times m$ matrix of observed contribution counts with y_{ijt_g} being the contribution amount of PAC i to candidate j in period t_g.

How statistical models look in research papers

How to Analyze Political Attention with Minimal Assumptions and Costs

Kevin M. Quinn University of California, Berkeley
Burt L. Monroe The Pennsylvania State University
Michael Colaresi Michigan State University
Michael H. Crespin University of Georgia
Dragomir R. Radev University of Michigan

As will become apparent later, it will be useful to write this sampling density in terms of latent data $\mathbf{z}_1, \ldots, \mathbf{z}_D$. Here $\mathbf{z}_d$ is a K-vector with element z_{dk} equal to 1 if document d was generated from topic k and 0 otherwise. If we could observe $\mathbf{z}_1, \ldots, \mathbf{z}_D$ we could write the sampling density above as

$$p(\mathbf{Y}, \mathbf{Z} \mid \boldsymbol{\pi}, \boldsymbol{\theta}) \propto \prod_{d=1}^{D} \prod_{k=1}^{K} \left(\pi_{s(d)k} \prod_{w=1}^{W} \theta_{kw}^{y_{dw}} \right)^{z_{dk}}.$$

Surveying a suite of algorithms that offer a solution to managing large document archives.

BY DAVID M. BLEI

Probabilistic Topic Models

With this notation, the generative process for LDA corresponds to the following joint distribution of the hidden and observed variables,

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right). \quad (1)$$

Scaling text: wordfish

Scaling text: wordfish

Instead of counts of students in rows, let's model something more interesting: the number of times a political party mentions a given word in its manifesto.

Scaling text: wordfish

Instead of counts of students in rows, let's model something more interesting: the number of times a political party mentions a given word in its manifesto.

Suppose rate of use of word j by party i in year t is modeled as

$$\lambda_{ijt} = e^{\alpha_{it} + \psi_j + \beta_j \omega_{it}}$$

Scaling text: wordfish

Instead of counts of students in rows, let's model something more interesting: the number of times a political party mentions a given word in its manifesto.

Suppose rate of use of word j by party i in year t is modeled as

$$\lambda_{ijt} = e^{\alpha_{it} + \psi_j + \beta_j \omega_{it}}$$

where

- α_{it} is party-year fixed effect
- ψ_j is word fixed effect
- β_j is word weight, i.e. discrimination parameter
- ω_{it} is party i's position in year t

Scaling text: wordfish

Instead of counts of students in rows, let's model something more interesting: the number of times a political party mentions a given word in its manifesto.

Suppose rate of use of word j by party i in year t is modeled as

$$\lambda_{ijt} = e^{\alpha_{it} + \psi_j + \beta_j \omega_{it}}$$

where

- α_{it} is party-year fixed effect
- ψ_j is word fixed effect
- β_j is word weight, i.e. discrimination parameter
- ω_{it} is party i's position in year t

and we assume word use is iid (conditional on λ_{ijt}).

Can we estimate α_{it}, ψ_j, β_j, and ω_{it} with MLE? OLS?

Scaling text: wordfish (2)

Scaling text: wordfish (2)

Consider this model for the rate λ for party i using word j at time t:

$$\lambda_{ijt} = e^{\alpha_{it} + \psi_j + \beta_j \omega_{it}}$$

Scaling text: wordfish (2)

Consider this model for the rate λ for party i using word j at time t:

$$\lambda_{ijt} = e^{\alpha_{it} + \psi_j + \beta_j \omega_{it}}$$

where

- α_{it} is party-year fixed effect
- ψ_j is word fixed effect
- β_j is word weight, i.e. discrimination parameter
- ω_{it} is party i's position in year t

Scaling text: wordfish (2)

Consider this model for the rate λ for party i using word j at time t:

$$\lambda_{ijt} = e^{\alpha_{it} + \psi_j + \beta_j \omega_{it}}$$

where

- α_{it} is party-year fixed effect
- ψ_j is word fixed effect
- β_j is word weight, i.e. discrimination parameter
- ω_{it} is party i's position in year t

Consider the words "and" and "deficit".

Q: What values of ψ_j and β_j would you expect for these words?

Scaling text: wordfish (2)

Consider this model for the rate λ for party i using word j at time t:

$$\lambda_{ijt} = e^{\alpha_{it} + \psi_j + \beta_j \omega_{it}}$$

where

- α_{it} is party-year fixed effect
- ψ_j is word fixed effect
- β_j is word weight, i.e. discrimination parameter
- ω_{it} is party i's position in year t

Consider the words "and" and "deficit".

Q: What values of ψ_j and β_j would you expect for these words?

A: For the word "and":

Scaling text: wordfish (2)

Consider this model for the rate λ for party i using word j at time t:

$$\lambda_{ijt} = e^{\alpha_{it} + \psi_j + \beta_j \omega_{it}}$$

where

- α_{it} is party-year fixed effect
- ψ_j is word fixed effect
- β_j is word weight, i.e. discrimination parameter
- ω_{it} is party i's position in year t

Consider the words "and" and "deficit".

Q: What values of ψ_j and β_j would you expect for these words?

A: For the word "and":

- high ψ_j, because it is a common word

Scaling text: wordfish (2)

Consider this model for the rate λ for party i using word j at time t:

$$\lambda_{ijt} = e^{\alpha_{it} + \psi_j + \beta_j \omega_{it}}$$

where

- α_{it} is party-year fixed effect
- ψ_j is word fixed effect
- β_j is word weight, i.e. discrimination parameter
- ω_{it} is party i's position in year t

Consider the words "and" and "deficit".

Q: What values of ψ_j and β_j would you expect for these words?

A: For the word "and":

- high ψ_j, because it is a common word
- small (in magnitude) β_j because its frequency is not likely to differ between parties

Scaling text: wordfish (2)

Consider this model for the rate λ for party i using word j at time t:

$$\lambda_{ijt} = e^{\alpha_{it} + \psi_j + \beta_j \omega_{it}}$$

where

- α_{it} is party-year fixed effect
- ψ_j is word fixed effect
- β_j is word weight, i.e. discrimination parameter
- ω_{it} is party i's position in year t

Consider the words "and" and "deficit".

Q: What values of ψ_j and β_j would you expect for these words?

A: For the word "and":

- high ψ_j, because it is a common word
- small (in magnitude) β_j because its frequency is not likely to differ between parties

For the word "deficit":

Scaling text: wordfish (2)

Consider this model for the rate λ for party i using word j at time t:

$$\lambda_{ijt} = e^{\alpha_{it} + \psi_j + \beta_j \omega_{it}}$$

where

- α_{it} is party-year fixed effect
- ψ_j is word fixed effect
- β_j is word weight, i.e. discrimination parameter
- ω_{it} is party i's position in year t

Consider the words "and" and "deficit".

Q: What values of ψ_j and β_j would you expect for these words?

A: For the word "and":

- high ψ_j, because it is a common word
- small (in magnitude) β_j because its frequency is not likely to differ between parties

For the word "deficit":

- lower ψ_j

Scaling text: wordfish (2)

Consider this model for the rate λ for party i using word j at time t:

$$\lambda_{ijt} = e^{\alpha_{it} + \psi_j + \beta_j \omega_{it}}$$

where

- α_{it} is party-year fixed effect
- ψ_j is word fixed effect
- β_j is word weight, i.e. discrimination parameter
- ω_{it} is party i's position in year t

Consider the words "and" and "deficit".

Q: What values of ψ_j and β_j would you expect for these words?

A: For the word "and":

- high ψ_j, because it is a common word
- small (in magnitude) β_j because its frequency is not likely to differ between parties

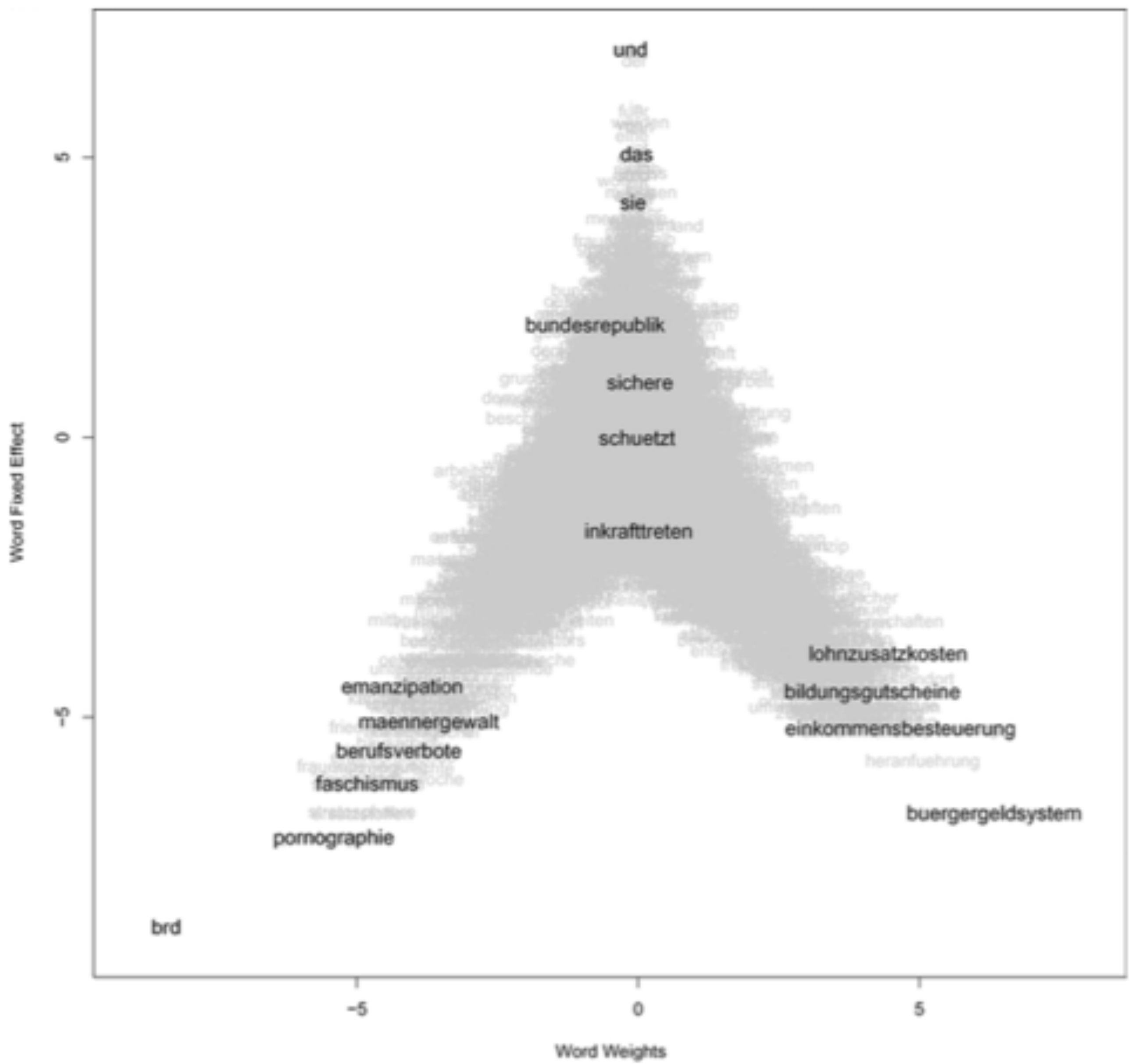
For the word "deficit":

- lower ψ_j
- larger (in magnitude) β_j; for example, if the right talks about "deficits" more frequently and party positions are oriented so that right is positive, β_j should be large and positive.

Eiffel Tower of words

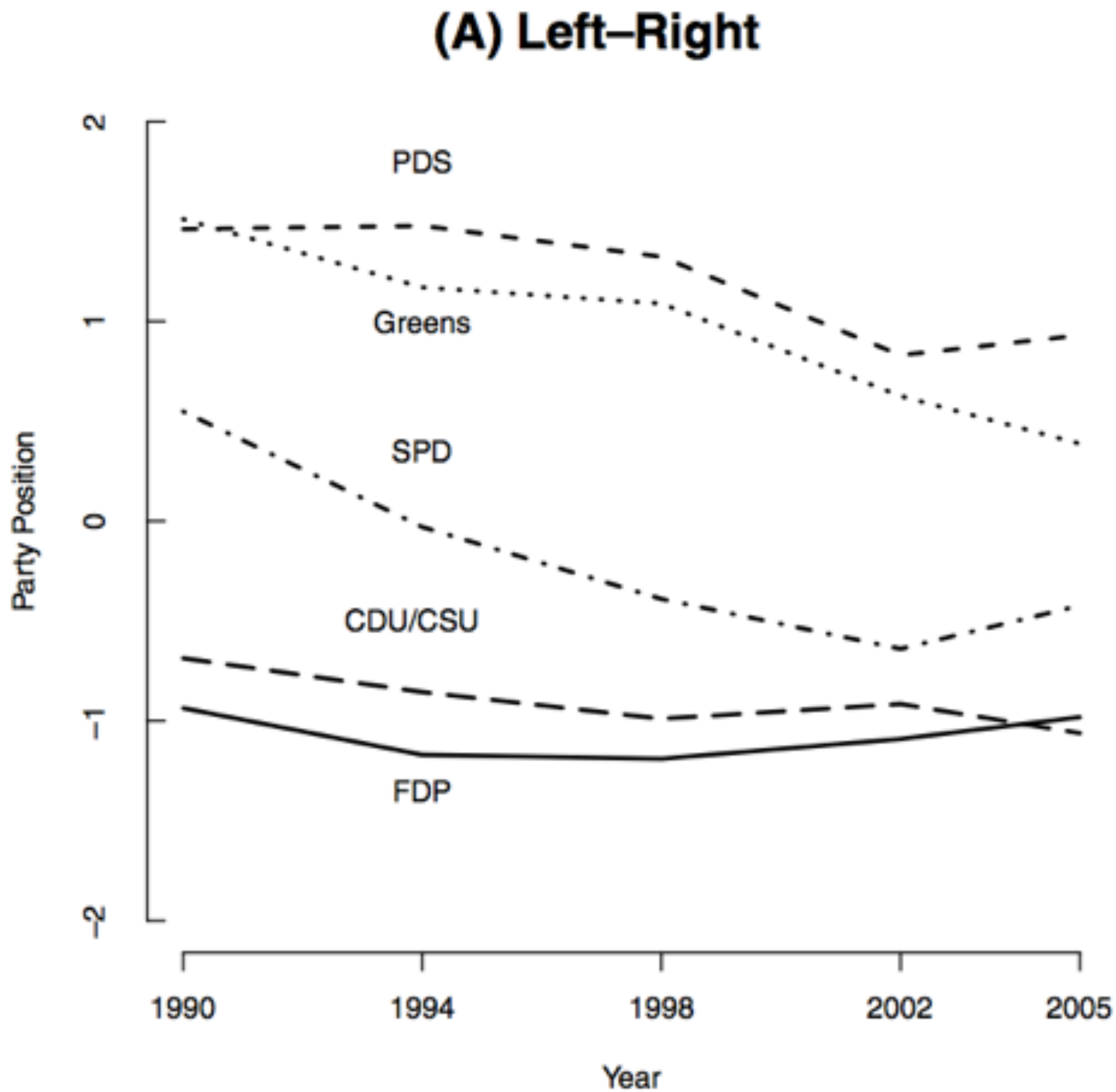
Slapin and Proksch, 2008

FIGURE 2 Word Weights vs. Word Fixed Effects. Left-Right Dimension, Germany 1990–2005 (Translations given in text)



Estimated party positions in Germany

Slapin and Proksch, 2008



How does OLS fit in?

How does OLS fit in?

OLS has attractive predictive/descriptive features *independent of a statistical model.*

How does OLS fit in?

OLS has attractive predictive/descriptive features *independent of a statistical model.*

Most importantly, the solution to

How does OLS fit in?

OLS has attractive predictive/descriptive features *independent of a statistical model*.

Most importantly, the solution to

$$\text{argmin}_{\alpha,\beta} \sum_{i=1}^{n} \Big(y_i - \alpha - \beta x_i \Big)^2$$

How does OLS fit in?

OLS has attractive predictive/descriptive features *independent of a statistical model.*

Most importantly, the solution to

$$\text{argmin}_{\alpha,\beta} \sum_{i=1}^{n} \Big(y_i - \alpha - \beta x_i \Big)^2$$

will give the best (minimum mean squared error) linear approximation of Y|X and E[Y|X] (the CEF) **regardless of linearity of CEF, distribution of errors.*** (Inference also works asymptotically.)

*This is main message of MHE 3.1 and Gailmard 132-135; see also Gailmard 314 ff.

How does OLS fit in?

OLS has attractive predictive/descriptive features *independent of a statistical model.*

Most importantly, the solution to

$$\text{argmin}_{\alpha,\beta} \sum_{i=1}^{n} \Big(y_i - \alpha - \beta x_i \Big)^2$$

will give the best (minimum mean squared error) linear approximation of Y|X and E[Y|X] (the CEF) **regardless of linearity of CEF, distribution of errors**.* (Inference also works asymptotically.)

But the OLS coefficients are also the MLE in a statistical model where $Y \sim N(\alpha + \beta X, \sigma^2)$ (i.e. mean of $\alpha + \beta X$, normal error with variance σ^2).

*This is main message of MHE 3.1 and Gailmard 132-135; see also Gailmard 314 ff.

General advice

General advice

- Keep it simple

General advice

- Keep it simple
- Keep it linked to an interesting research question

General advice

- Keep it simple
- Keep it linked to an interesting research question
- Keep it visual: before (and after) running a model, look at the data!

General advice

- Keep it simple
- Keep it linked to an interesting research question
- Keep it visual: before (and after) running a model, look at the data!
- Learn a little about a lot of techniques (so you can recognize when you need to know more) and a lot about something

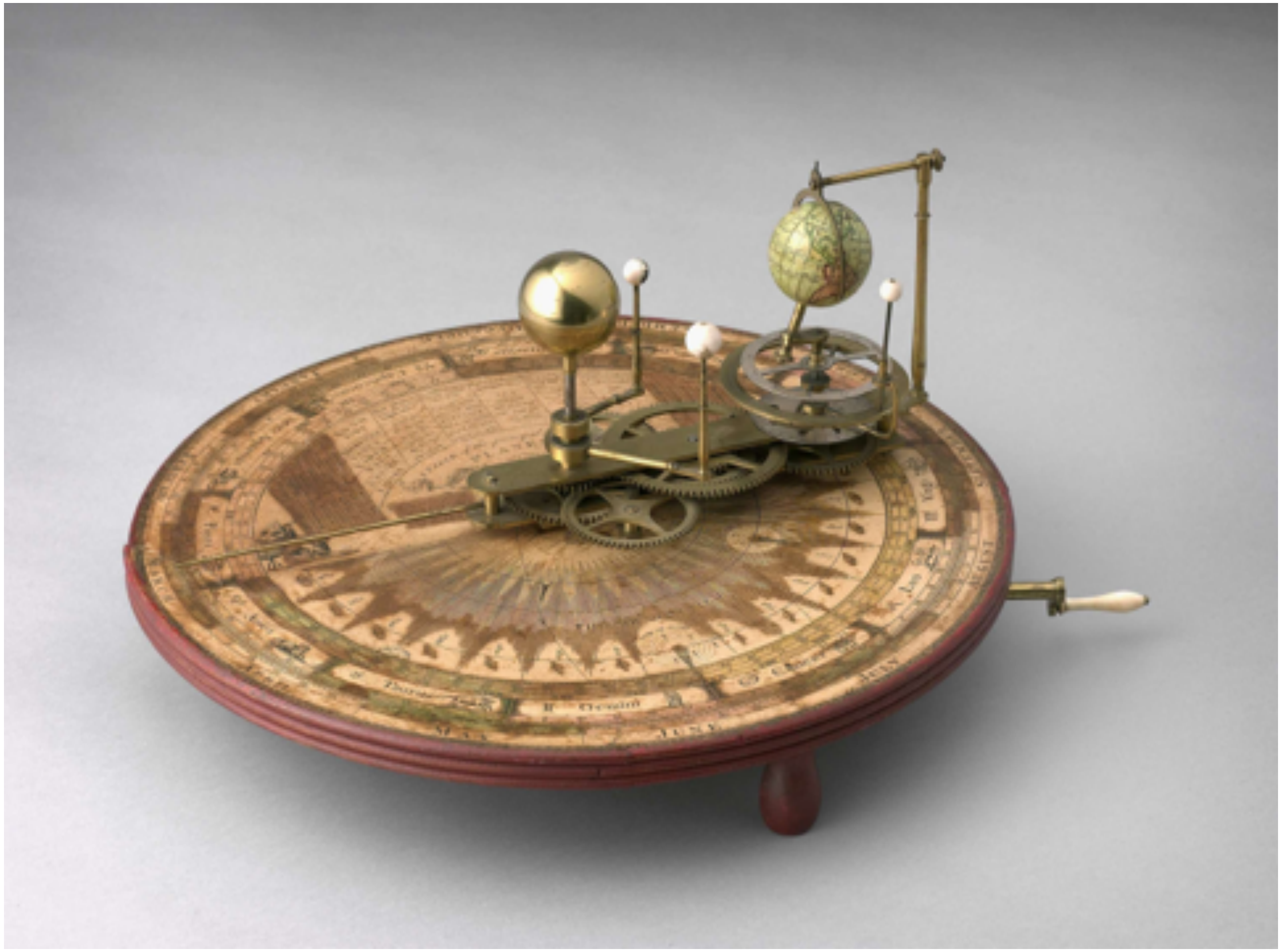
General advice

- Keep it simple
- Keep it linked to an interesting research question
- Keep it visual: before (and after) running a model, look at the data!
- Learn a little about a lot of techniques (so you can recognize when you need to know more) and a lot about something
- Get good at Stata, R, or both

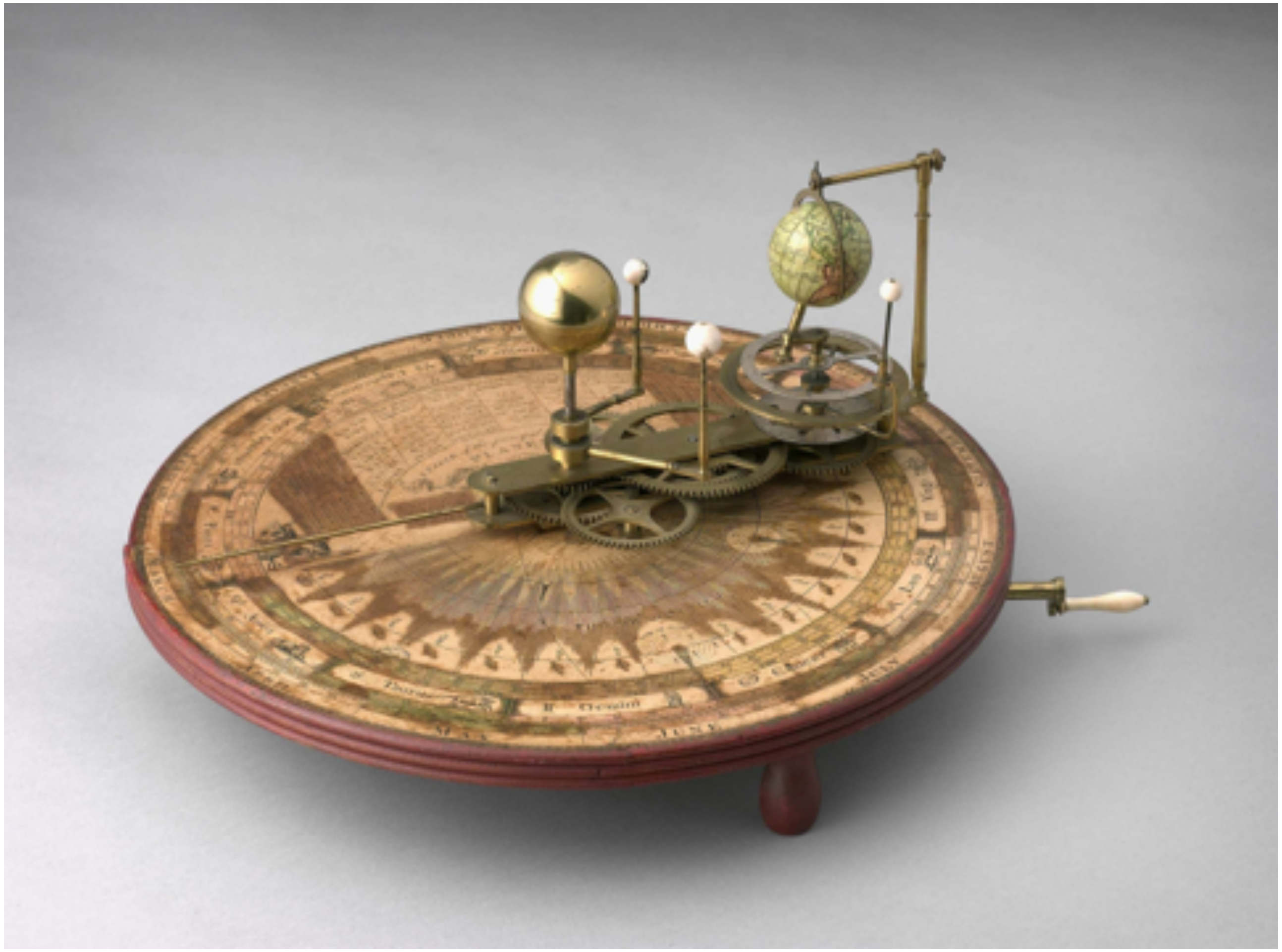
General advice

- Keep it simple
- Keep it linked to an interesting research question
- Keep it visual: before (and after) running a model, look at the data!
- Learn a little about a lot of techniques (so you can recognize when you need to know more) and a lot about something
- Get good at Stata, R, or both
- There are many ways to contribute. Choose some combination of:
  - better data
  - better design (e.g. causal inference)
  - better measurement
  - better theory

  Often one of these makes possible another.



Jones's "New Portable Orrery" (1794)



Jones's "New Portable Orrery" (1794)

"All models are wrong, but some are useful." George Box