POLS/CSSS 510:
Maximum Likelihood Methods for the Social Sciences

Modeling Nominal Outcomes

Christopher Adolph

Department of Political Science
and
Center for Statistics and the Social Sciences
University of Washington, Seattle

Categorical Variables

Theme of class: analyzing data which do not fit assumptions of the normal model

First considered binary data

Then ordered data

But what about truly categorical data?

How do we analyze data when we can't make any scaling assumptions at all?

When data are just "names" (hence the data are "nominal")?

Who did you vote for?	{Labor, Conservative, SNP, UKIP, . . . }
Which product did you purchase?	{Coke, Pepsi, Dr. Pepper }
What kind of job do you have?	{Manual labor, skilled labor, managerial, . . . }
Where to locate a plant?	{Seattle, Tokyo, London }
What is the ethnicity of a new hire?	{White, Black, Asian, . . . }
and so on . . .	

Analyzing "What", "Which", "Who", "Where" questions,
rather than "how much", or "how many"

Nominal Regression Models

We can build models for nominal data on top of our work on logit, probit, etc.

Potentially more complicated than models so far

Some models for nominal data require massive computing power
(but massive computing power is ubiquitous – even in your pocket. . .)

We will start with a "simple" model, then explore more flexible models

For nominal models, need more elaborate notation than used in earlier weeks

Notation for Nominal Regression Models

Welcome to subscript hell

Observations: $1, \ldots i, \ldots N$

Unordered Categories: $1, \ldots j, \ldots M$

Category 1 will be the "reference category",
so we will often speak of the remaining categories $2, \ldots M$

Covariates: $x_1, \ldots x_k, \ldots x_P$

Parameters: There may be a parameter for each combination of a non-reference category and a covariate, plus an intercept for non-reference category

That is, the systematic component for the jth category of the ith observation is

$$\mu_{ij} = \beta_{j0} + \sum_{k=1}^{P} \beta_{jk} \times x_{ik}$$

Note the lack of a j subscript on x . . . this will change later

Notation for Nominal Regression Models

The systematic component for the jth category of the ith observation is

$$\mu_{ij} = \beta_{j0} + \sum_{k=1}^{P} \beta_{jk} \times x_{ik}$$

or, in matrix form,

$$\boldsymbol{\mu}_j = \mathbf{X}\boldsymbol{\beta}_j$$

where

$\boldsymbol{\mu}_j$	is an $N \times 1$ vector of systematic components
$\boldsymbol{\beta}_j$	is a $(P+1) \times 1$ vector of parameters (there's a $\boldsymbol{\beta}_j$ for each category j of Y)
$\mathbf{X}$	is an $N \times (P+1)$ matrix of covariates (the same used regardless of the category j)

Notation for Nominal Regression Models

We could also put the parameters in one big matrix, like this:

$$\boldsymbol{\beta} = \beta_{jk} = \begin{pmatrix} \beta_{10} & \beta_{11} & \cdots & \beta_{1k} & \cdots & \beta_{1P} \\ \beta_{20} & \beta_{21} & \cdots & \beta_{2k} & \cdots & \beta_{2P} \\ \vdots & & \ddots & & & \vdots \\ \beta_{j0} & & & \beta_{jk} & & \beta_{jP} \\ \vdots & & & & \ddots & \vdots \\ \beta_{M0} & \beta_{M1} & \cdots & \beta_{Mk} & \cdots & \beta_{MP} \end{pmatrix}$$

Notation for Nominal Regression Models

Each row of the matrix contains the complete set of parameters for a category:

$$\boldsymbol{\beta} = \beta_{jk} = \begin{pmatrix} \beta_{10} & \beta_{11} & \cdots & \beta_{1k} & \cdots & \beta_{1P} \\ \beta_{20} & \beta_{21} & \cdots & \beta_{2k} & \cdots & \beta_{2P} \\ \vdots & & \ddots & & & \vdots \\ \beta_{j0} & & & \beta_{jk} & & \beta_{jP} \\ \vdots & & & & \ddots & \vdots \\ \beta_{M0} & \beta_{M1} & \cdots & \beta_{Mk} & \cdots & \beta_{MP} \end{pmatrix}$$

Notation for Nominal Regression Models

And each column corresponds to a set of parameters for the same covariate:

$$\boldsymbol{\beta} = \beta_{jk} = \begin{pmatrix} \beta_{10} & \beta_{11} & \cdots & \beta_{1k} & \cdots & \beta_{1P} \\ \beta_{20} & \beta_{21} & \cdots & \beta_{2k} & \cdots & \beta_{2P} \\ \vdots & & \ddots & & & \vdots \\ \beta_{j0} & & & \beta_{jk} & & \beta_{jP} \\ \vdots & & & & \ddots & \vdots \\ \beta_{M0} & \beta_{M1} & \cdots & \beta_{Mk} & \cdots & \beta_{MP} \end{pmatrix}$$

Notation for Nominal Regression Models

For identification, we set one row of coefficients to 0:

$$\boldsymbol{\beta} = \beta_{jk} = \begin{pmatrix} 0 & 0 & \dots & 0 & \dots & 0 \\ \beta_{20} & \beta_{21} & \dots & \beta_{2k} & \dots & \beta_{2P} \\ \vdots & & \ddots & & & \vdots \\ \beta_{j0} & & & \beta_{jk} & & \beta_{jP} \\ \vdots & & & & \ddots & \vdots \\ \beta_{M0} & \beta_{M1} & \dots & \beta_{Mk} & \dots & \beta_{MP} \end{pmatrix}$$

That is, we treat one category (here, "1") as the *reference category*

That leaves $(M-1) \times (P+1)$ parameters to estimate

Notation for Nominal Regression Models

For comparison:

in ordinary logit, we estimate $P + 1$ parameters

in ordered probit, we estimate $P + M - 1$ parameters

in multinomial logit, we estimate $(M - 1) \times (P + 1)$ parameters

$\rightarrow$ MNL gobbles up degrees of freedom fast

$\rightarrow$ Compared to ordinary logit, no big deal (we have more data)

$\rightarrow$ Compared to ordered probit, or linear regression, MNL is "expensive"

If the assumptions of ordered probit fit your data, use it

But the assumptions don't always fit . . .

Likelihood for Multinomial Logit (MNL)

As usual, we start with a probability model *What probability do we need?*

We need to calculate $\Pr(y_i = j|\mathbf{x}_i, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_M)$,
the probability that y_i falls in category j

For a start, we need each $\Pr(\cdot)$ to be positive, so we might try

$$\Pr(y_i = j|x_i, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_M) \propto \exp(\mathbf{x}_i\boldsymbol{\beta}_j)$$

But we also need the probabilities to sum to 1 across all the categories j
for a given observation i. . .

The inverse-logit usefully keeps each $\Pr(\cdot)$ positive, and their sum $= 1$:

$$\Pr(y_i = j|x_i, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_M) = \frac{\exp(\mathbf{x}_i\boldsymbol{\beta}_j)}{\sum_{\ell=1}^{M} \exp(\mathbf{x}_i\boldsymbol{\beta}_\ell)}$$

However, to ensure the $\boldsymbol{\beta}$'s are identified, we will need to assume, say, $\beta_{1k} = 0 \quad \forall k$.

Likelihood for Multinomial Logit (MNL)

Thus we have for $j = 1$

$$\begin{aligned}\Pr(y_i = 1|\mathbf{x}_i, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_M) &= \frac{\exp(\mathbf{x}_i \times 0)}{\exp(\mathbf{x}_i \times 0) + \sum_{\ell=2}^{M} \exp(\mathbf{x}_i \boldsymbol{\beta}_\ell)} \\ &= \frac{1}{1 + \sum_{\ell=2}^{M} \exp(\mathbf{x}_i \boldsymbol{\beta}_\ell)}\end{aligned}$$

and for $j > 2$

$$\Pr(y_i = j|x_i, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_M) = \frac{\exp(\mathbf{x}_i \boldsymbol{\beta}_j)}{1 + \sum_{\ell=2}^{M} \exp(\mathbf{x}_i \boldsymbol{\beta}_\ell)}$$

Likelihood for Multinomial Logit (MNL)

By definition, the likelihood is proportional to the probability, p_{ij},
of observing the value of y that is ultimately observed
(the probabilities for all unobserved categories are irrelevant to the likelihood)

$$\mathcal{L}(\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_M | \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{N} \prod_{j=1}^{M} {p_{ij}}^{y_{ij}}$$

Substituting for p_{ij}, we have

$$\mathcal{L}(\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_M | \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{N} \prod_{j=1}^{M} \left[\frac{\exp(\mathbf{x}_i \boldsymbol{\beta}_j)}{1 + \sum_{\ell=2}^{M} \exp(\mathbf{x}_i \boldsymbol{\beta}_\ell)} \right]^{y_{ij}}$$

Taking logs, we end up with

$$\log \mathcal{L}(\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_M | \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} \log \frac{\exp(\mathbf{x}_i \boldsymbol{\beta}_j)}{1 + \sum_{\ell=2}^{M} \exp(\mathbf{x}_i \boldsymbol{\beta}_\ell)}$$

which we can maximize with `optim()`

Calculating Expected Values in MNL

After estimating an MNL model, calculating expected probabilities is straightforward.

For a given counterfactual level of the explanatory variables, $\mathbf{x}_c$,
and a three category multinomial logit, we have

$$\Pr(y=1|\mathbf{x}_c,\hat{\boldsymbol{\beta}}_2,\hat{\boldsymbol{\beta}}_3) = \frac{1}{1+\exp(\mathbf{x}_c\hat{\boldsymbol{\beta}}_2)+\exp(\mathbf{x}_c\hat{\boldsymbol{\beta}}_3)}$$

$$\Pr(y=2|\mathbf{x}_c,\hat{\boldsymbol{\beta}}_2,\hat{\boldsymbol{\beta}}_3) = \frac{\exp(\mathbf{x}_c\hat{\boldsymbol{\beta}}_2)}{1+\exp(\mathbf{x}_c\hat{\boldsymbol{\beta}}_2)+\exp(\mathbf{x}_c\hat{\boldsymbol{\beta}}_3)}$$

$$\Pr(y=3|\mathbf{x}_c,\hat{\boldsymbol{\beta}}_2,\hat{\boldsymbol{\beta}}_3) = \frac{\exp(\mathbf{x}_c\hat{\boldsymbol{\beta}}_3)}{1+\exp(\mathbf{x}_c\hat{\boldsymbol{\beta}}_2)+\exp(\mathbf{x}_c\hat{\boldsymbol{\beta}}_3)}$$

Simulating from the MNL is also simple:

Just draw the $\hat{\boldsymbol{\beta}}$'s from the multivariate normal,

then plug them into the above equations to get a matrix of simulates

Intepreting MNL Coefficients Directly

So what do the $\boldsymbol{\beta}$'s in a multinomial logit *mean*?

Suppose we write the odds of category m against category n:

$$\begin{aligned}
\frac{\Pr(y=m|\mathbf{x}_c,\boldsymbol{\beta}_2,\ldots,\boldsymbol{\beta}_M)}{\Pr(y=n|\mathbf{x}_c,\boldsymbol{\beta}_2,\ldots,\boldsymbol{\beta}_M)} &= \frac{\frac{\exp(\mathbf{x}_c\boldsymbol{\beta}_m)}{\sum_{\ell=1}^{M}\exp(\mathbf{x}_c\boldsymbol{\beta}_\ell)}}{\frac{\exp(\mathbf{x}_c\boldsymbol{\beta}_n)}{\sum_{\ell=1}^{M}\exp(\mathbf{x}_c\boldsymbol{\beta}_\ell)}} \\
&= \frac{\exp(\mathbf{x}_c\boldsymbol{\beta}_m)}{\exp(\mathbf{x}_c\boldsymbol{\beta}_n)} \\
&= \exp\left(\mathbf{x}_c(\boldsymbol{\beta}_m-\boldsymbol{\beta}_n)\right) \\
\log\frac{\Pr(y=m|\mathbf{x}_c,\boldsymbol{\beta}_2,\ldots,\boldsymbol{\beta}_M)}{\Pr(y=n|\mathbf{x}_c,\boldsymbol{\beta}_2,\ldots,\boldsymbol{\beta}_M)} &= \mathbf{x}_c(\boldsymbol{\beta}_m-\boldsymbol{\beta}_n)
\end{aligned}$$

just as for binary logit, we find that MNL is linear in the logit of y

Intepreting MNL Coefficients Directly

$$\log \frac{\Pr(y = m|\mathbf{x}_c, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_M)}{\Pr(y = n|\mathbf{x}_c, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_M)} = \mathbf{x}_c(\boldsymbol{\beta}_m - \boldsymbol{\beta}_n)$$

In words, if the kth covariate x_k increases by 1,
then the log of the odds of category m versus n increases by the difference of their coefficients, $\beta_{mk} - \beta_{nk}$

If category n is the reference category, $\beta_{nk} = 0$ by assumption,
and this reduces to β_{mk}.

Notice something peculiar:

- If we want to calculate the effect of x_k on the shift between two categories, m and n, the other categories are *irrelevant*

- This implies that the relative probability of m and n should remain the *same* even if a close substitute to m (but not to n) is added to the choice set:
  the Independence of Irrelevant Alternatives (IIA) assumption.

An Example: Chomp!

Agresti (2002) offers the following example of nominal data

Alligators in a certain Florida lake were studied, and the following data collected:

Principal Food	1 = Invertebrates, 2 = Fish, 3 = "Other"
Size of alligator	in meters
Sex of alligator	male or female

The question is how alligator size and sex influences food choice

We fit the model in `R` using MNL and get . . .

An Example: Chomp!

Agresti (2002) offers the following example of nominal data

Alligators in a certain Florida lake were studied, and the following data collected:

Principal Food	1 = Invertebrates, 2 = Fish, 3 = "Other" !!!
Size of alligator	in meters
Sex of alligator	male or female

The question is how alligator size and sex influences food choice

We fit the model in `R` using MNL and get . . .

An Example: Chomp!

Agresti (2002) offers the following example of nominal data

Alligators in a certain Florida lake were studied, and the following data collected:

Principal Food	1 = Invertebrates, 2 = Fish, 3 = "Other" *. . . Floridians?*
Size of alligator	in meters
Sex of alligator	male or female

The question is how alligator size and sex influences food choice

We fit the model in `R` using MNL and get . . .

	$\log\left(\frac{\Pr(\text{invertebrate})}{\Pr(\text{fish})}\right)$	$\log\left(\frac{\Pr(\text{“other”})}{\Pr(\text{fish})}\right)$
Size	-2.526	0.138
	(0.848)	(0.518)
Female	-0.790	0.382
	(0.712)	(0.908)
Intercept	4.897	-1.947
	(1.706)	(1.531)
log likelihood	-48.3	
AIC	108.6	
% correctly classified (in-sample)	62.7% (null=52.5%)	
% correctly classified (LOO-CV)	55.9% (null=52.5%)	
in-sample concordance	0.71 (null=0.50)	
LOO-CV concordance	0.59 (null=0.50)	
N	59	

Can you directly interpretation the size coefficients?

	$\log\left(\frac{\Pr(\text{invertebrate})}{\Pr(\text{fish})}\right)$	$\log\left(\frac{\Pr(\text{“other”})}{\Pr(\text{fish})}\right)$
Size	-2.526	0.138
	(0.848)	(0.518)
Female	-0.790	0.382
	(0.712)	(0.908)
Intercept	4.897	-1.947
	(1.706)	(1.531)
log likelihood	-48.3	
AIC	108.6	
% correctly classified (in-sample)	62.7% (null=52.5%)	
% correctly classified (LOO-CV)	55.9% (null=52.5%)	
in-sample concordance	0.71 (null=0.50)	
LOO-CV concordance	0.59 (null=0.50)	
N	59	

1. How does size influence the choice of an invertebrate diet vs. a fish diet?

A 1 meter increase in length makes the *odds* that an alligator will eat invertebrates rather than fish $\exp(-2.526 - 0) = 0.080$ times smaller

	$\log\left(\frac{\Pr(\text{invertebrate})}{\Pr(\text{fish})}\right)$	$\log\left(\frac{\Pr(\text{“other”})}{\Pr(\text{fish})}\right)$
Size	-2.526	0.138
	(0.848)	(0.518)
Female	-0.790	0.382
	(0.712)	(0.908)
Intercept	4.897	-1.947
	(1.706)	(1.531)
log likelihood	-48.3	
AIC	108.6	
% correctly classified (in-sample)	62.7% (null=52.5%)	
% correctly classified (LOO-CV)	55.9% (null=52.5%)	
in-sample concordance	0.71 (null=0.50)	
LOO-CV concordance	0.59 (null=0.50)	
N	59	

2. How does size influence the choice of an invertebrate diet vs. an “other” diet?

A 1 meter increase in length makes the *odds* that an alligator will eat invertebrates rather than “other” food $\exp(-2.526 - 0.138) = 0.070$ times smaller

	$\log\left(\frac{\Pr(\text{invertebrate})}{\Pr(\text{fish})}\right)$	$\log\left(\frac{\Pr(\text{“other”})}{\Pr(\text{fish})}\right)$
Size	-2.526	0.138
	(0.848)	(0.518)
Female	-0.790	0.382
	(0.712)	(0.908)
Intercept	4.897	-1.947
	(1.706)	(1.531)
log likelihood	-48.3	
AIC	108.6	
% correctly classified (in-sample)	62.7% (null=52.5%)	
% correctly classified (LOO-CV)	55.9% (null=52.5%)	
in-sample concordance	0.71 (null=0.50)	
LOO-CV concordance	0.59 (null=0.50)	
N	59	

3. How does size influence the choice of a fish diet vs. an “other” diet?

A 1 meter increase in length makes the *odds* that an alligator will eat fish rather than “other” food $\exp(0 - 0.138) = 0.871$ times smaller

	$\log\left(\frac{\Pr(\text{invertebrate})}{\Pr(\text{fish})}\right)$	$\log\left(\frac{\Pr(\text{“other”})}{\Pr(\text{fish})}\right)$
Size	-2.526	0.138
	(0.848)	(0.518)
Female	-0.790	0.382
	(0.712)	(0.908)
Intercept	4.897	-1.947
	(1.706)	(1.531)
log likelihood	-48.3	
AIC	108.6	
% correctly classified (in-sample)	62.7% (null=52.5%)	
% correctly classified (LOO-CV)	55.9% (null=52.5%)	
in-sample concordance	0.71 (null=0.50)	
LOO-CV concordance	0.59 (null=0.50)	
N	59	

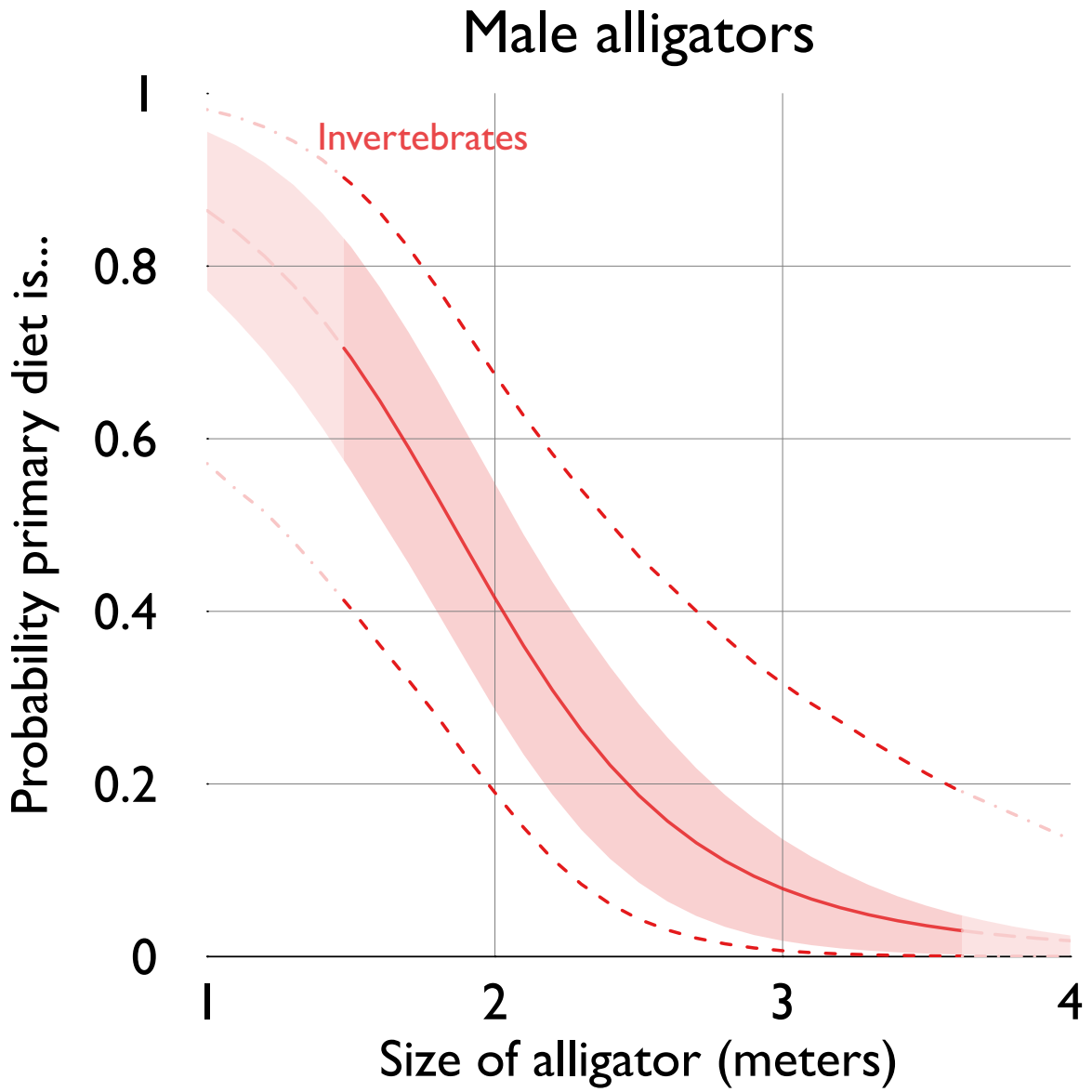
Big gators are more likely to eat mostly fish and. . . “other” food, relative to invertebrates

Really big gators still more likely to favor “other” food, even relative to fish

	$\log\left(\frac{\Pr(\text{invertebrate})}{\Pr(\text{fish})}\right)$	$\log\left(\frac{\Pr(\text{“other”})}{\Pr(\text{fish})}\right)$
Size	-2.526	0.138
	(0.848)	(0.518)
Female	-0.790	0.382
	(0.712)	(0.908)
Intercept	4.897	-1.947
	(1.706)	(1.531)
log likelihood	-48.3	
AIC	108.6	
% correctly classified (in-sample)	62.7% (null=52.5%)	
% correctly classified (LOO-CV)	55.9% (null=52.5%)	
in-sample concordance	0.71 (null=0.50)	
LOO-CV concordance	0.59 (null=0.50)	
N	59	

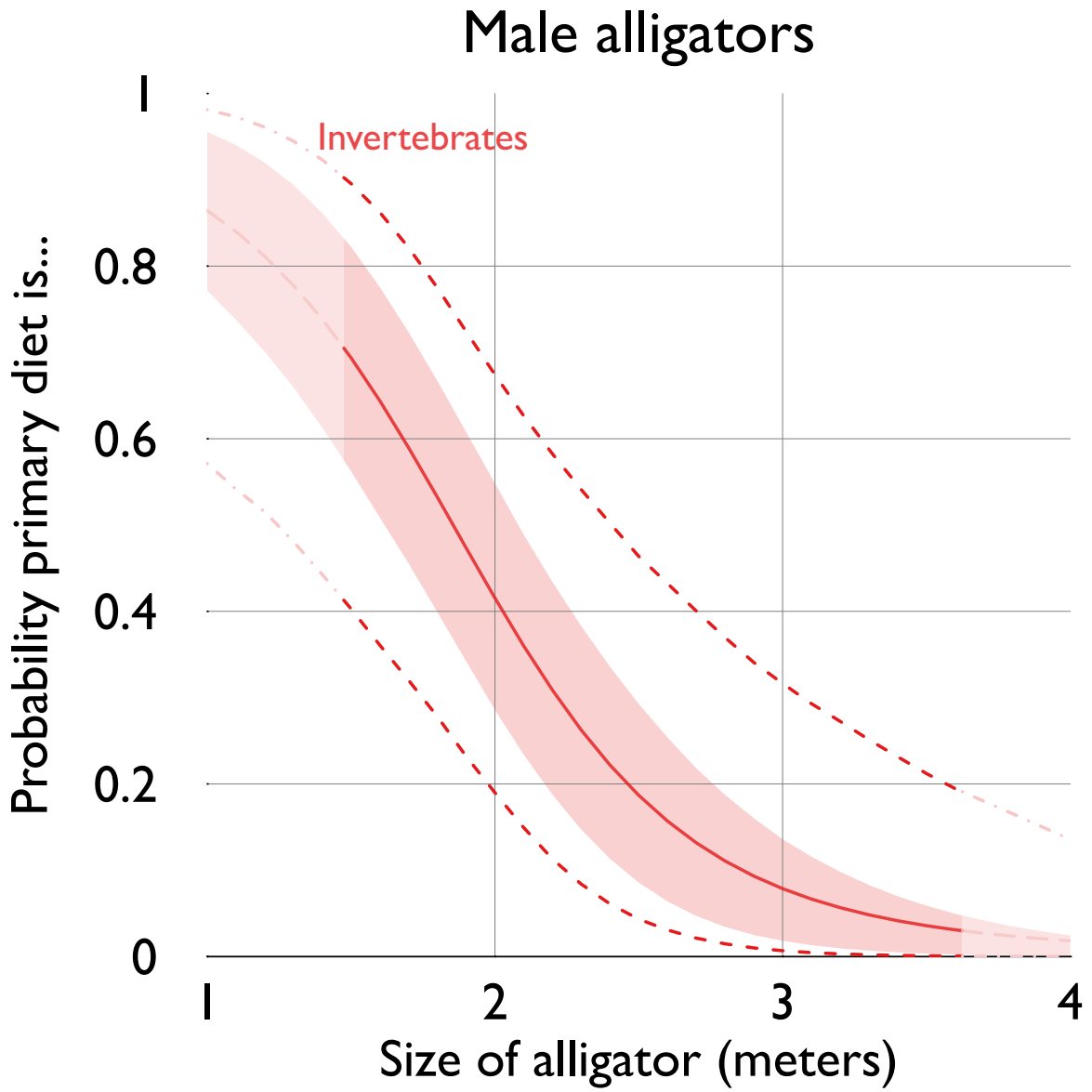
Although odds ratios are invariant to other covariate levels,
they are uninterpretable for most people

There has to be a better way . . .



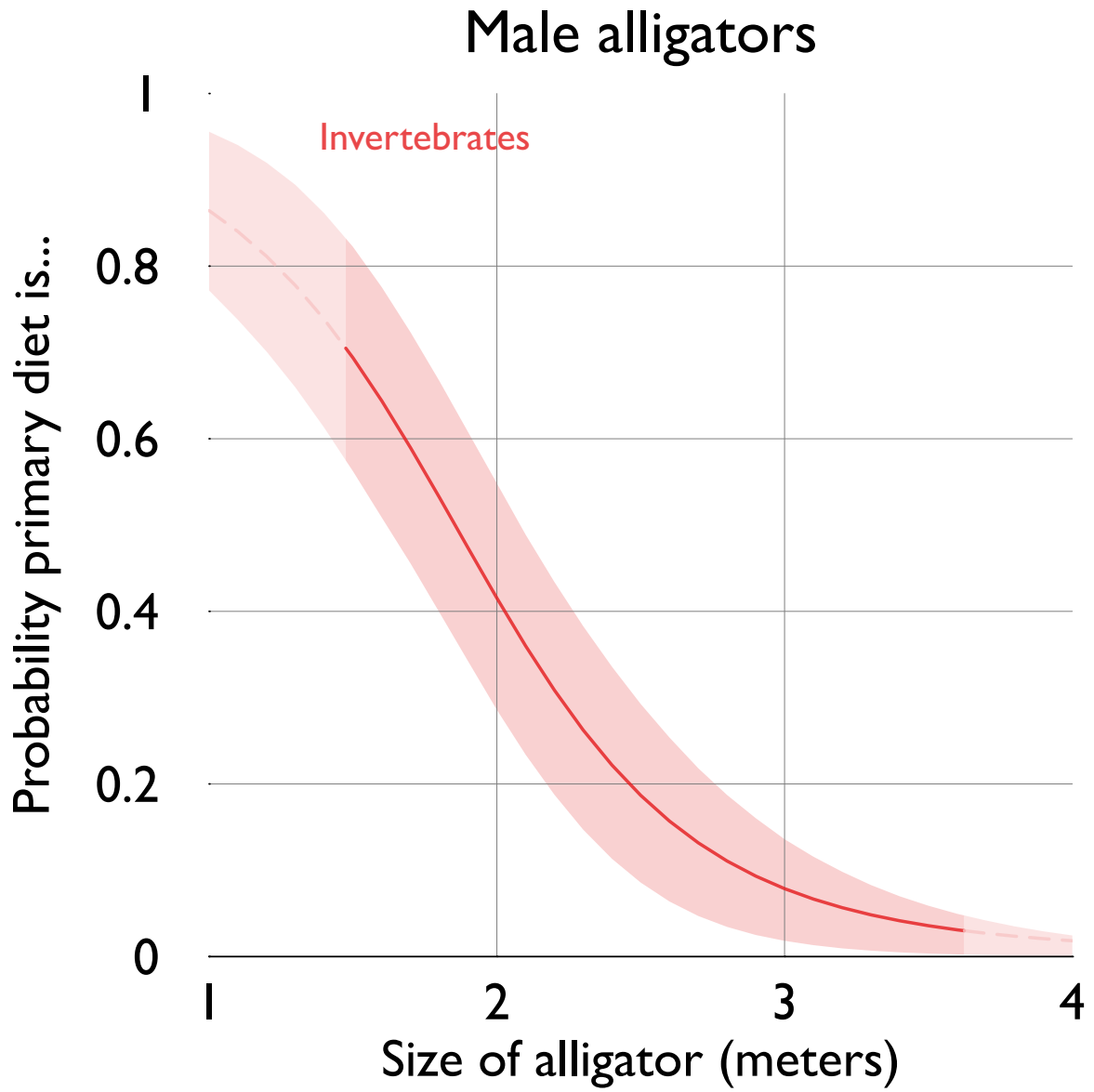
Let's simulate the predicted probability a male gator eats invertebrates, by size

Above are the predicted probabilities with 68% and 95% CIs



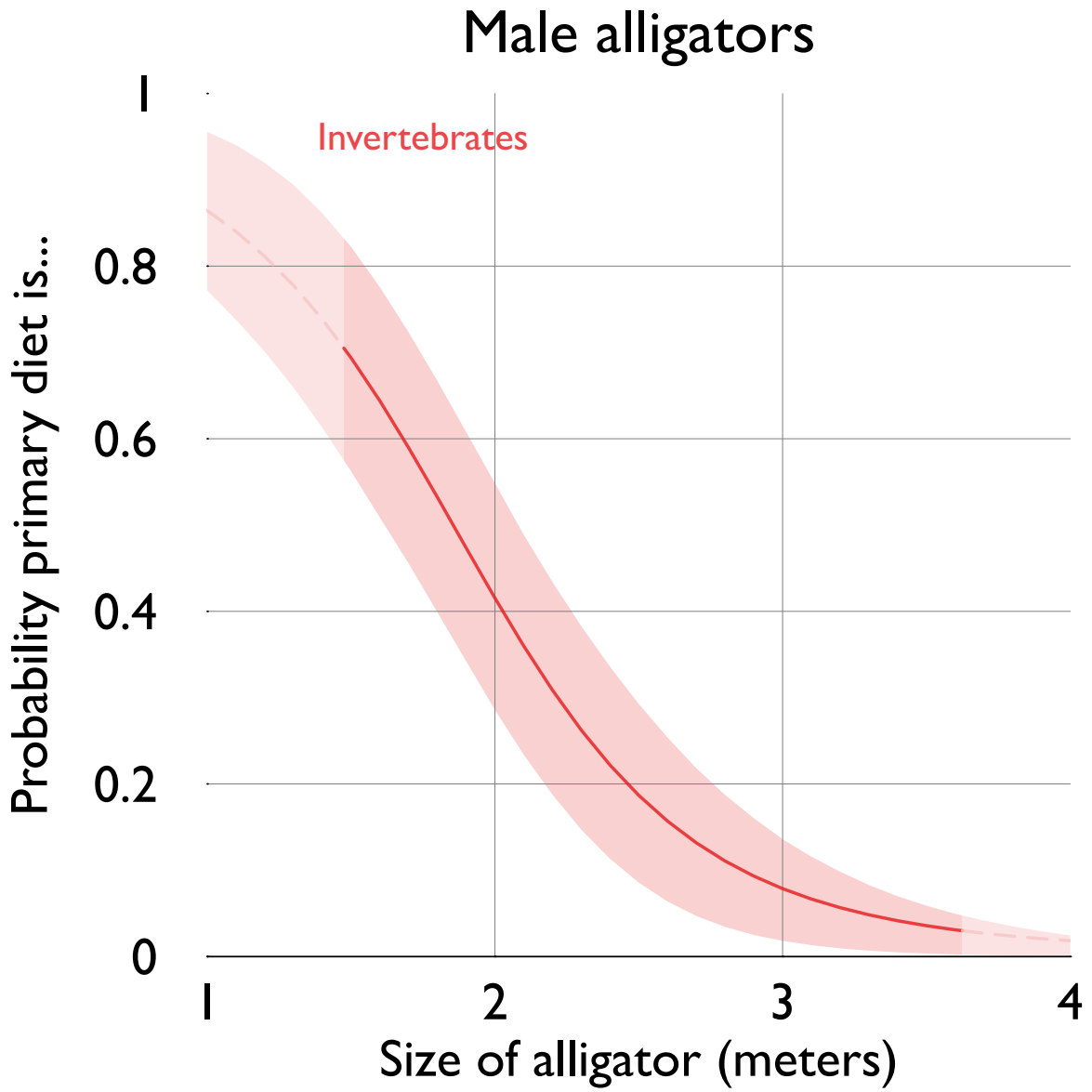
68% (95%) CIs are analogous to ± 1 (± 2) SE bars

Let's focus on 68% CIs for now. . .



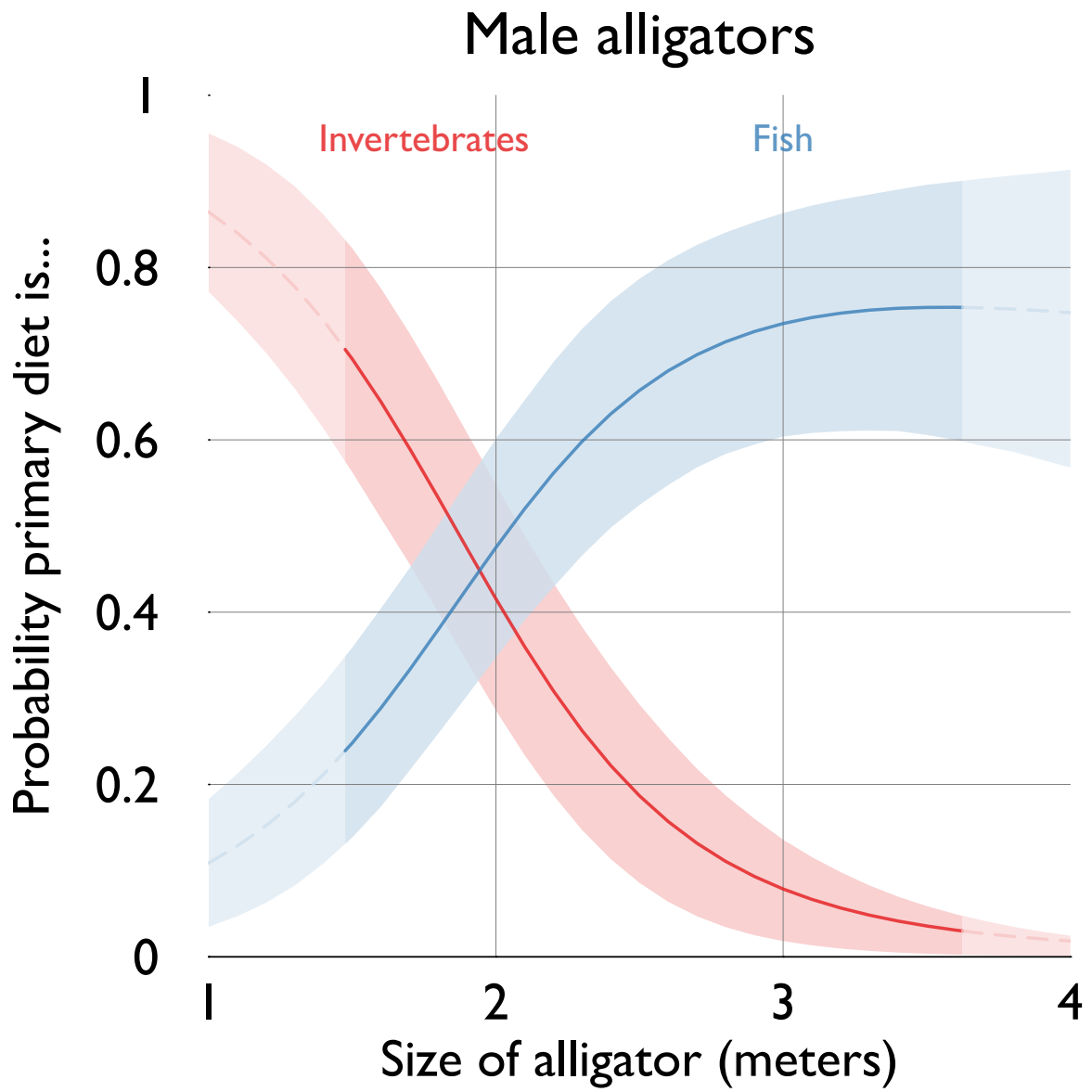
The lighter regions and dashed central line indicate *extrapolation*

The central, observed range of male gator size is emphasized



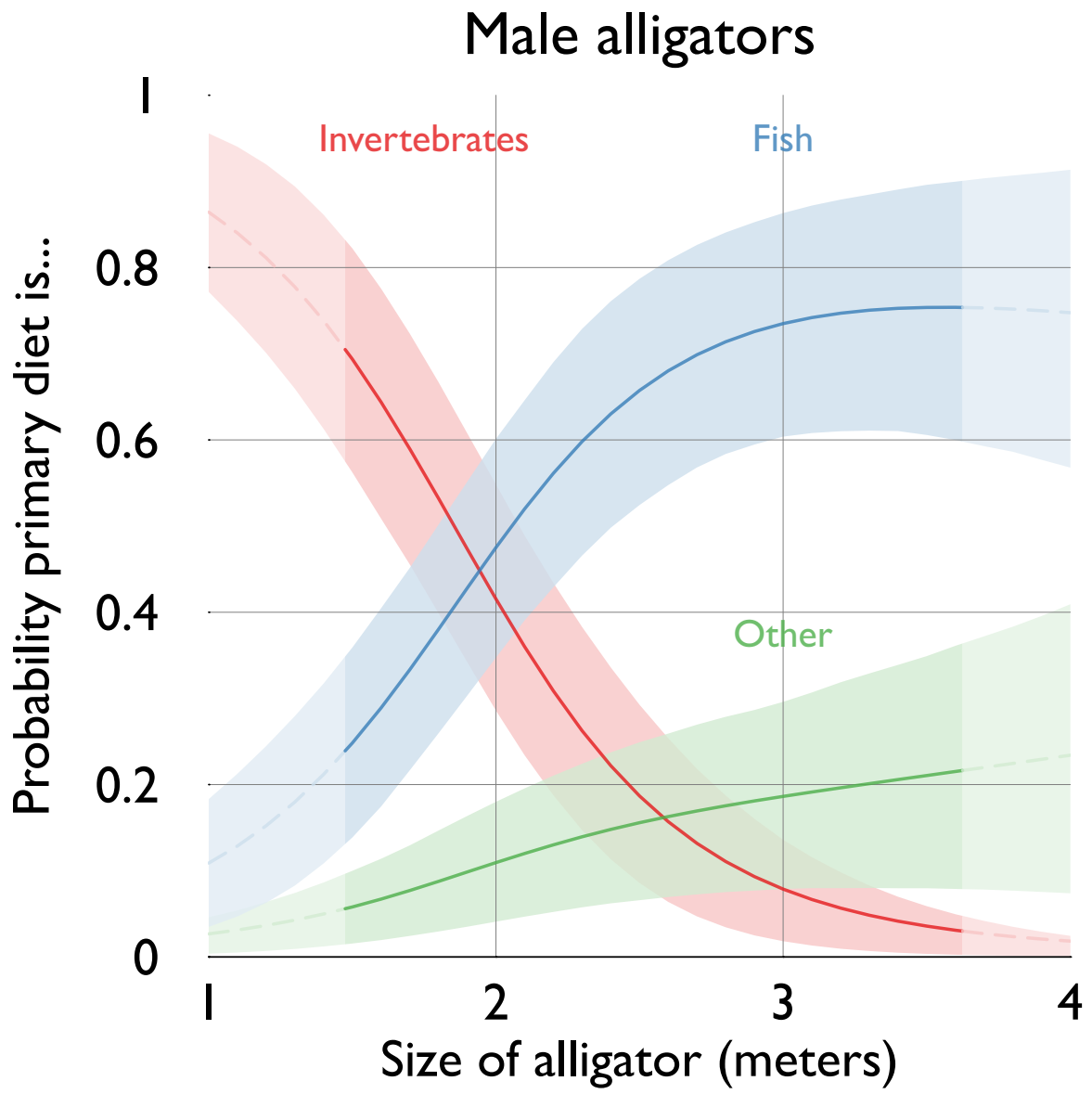
We have three categories to predict, and they can behave quite differently

Let's plot them together

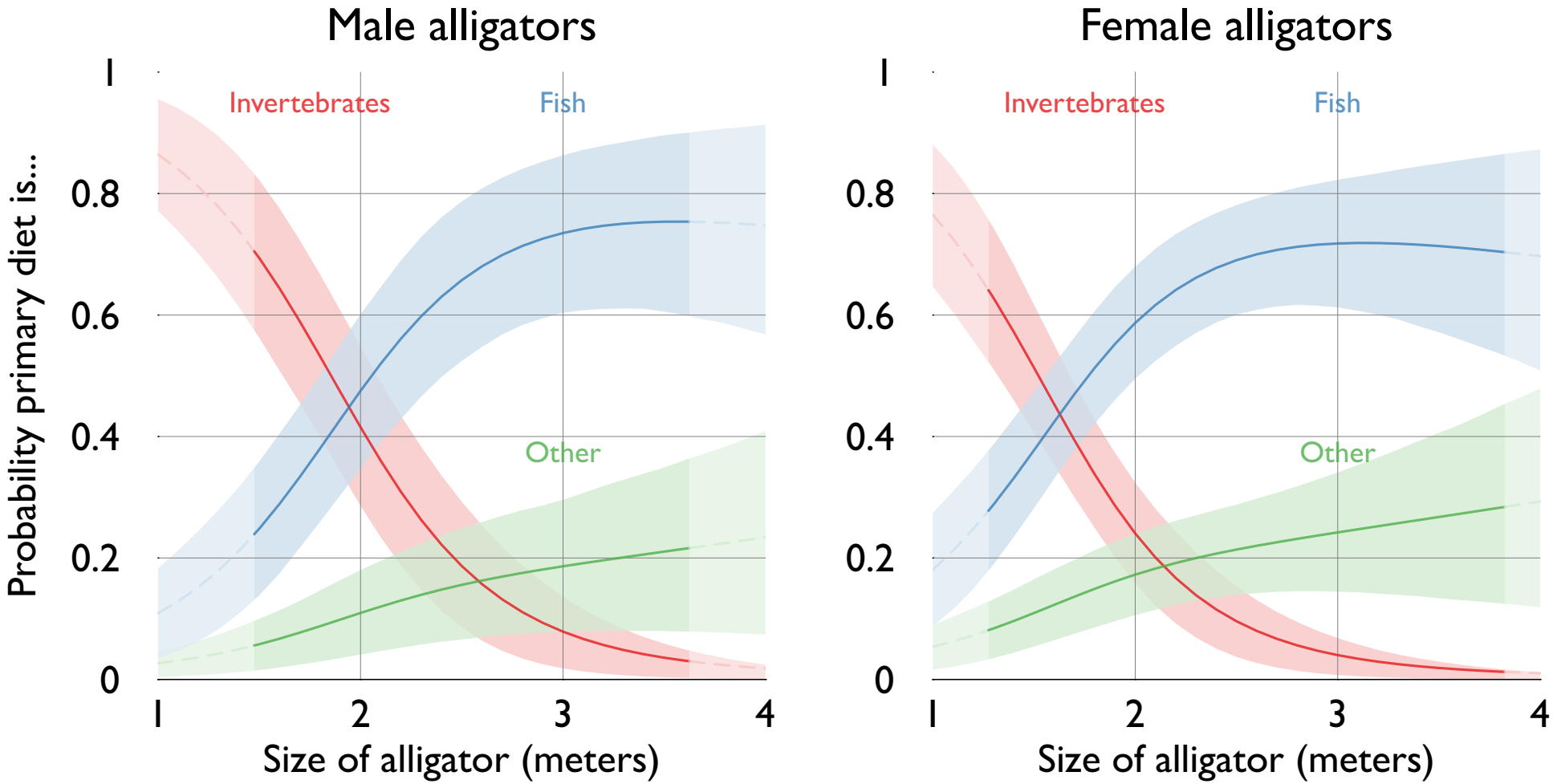


As gators get bigger, they're less likely to focus on invertebrates

and more likely to focus on fish



But as gators get really big, a minority shift to "other" food

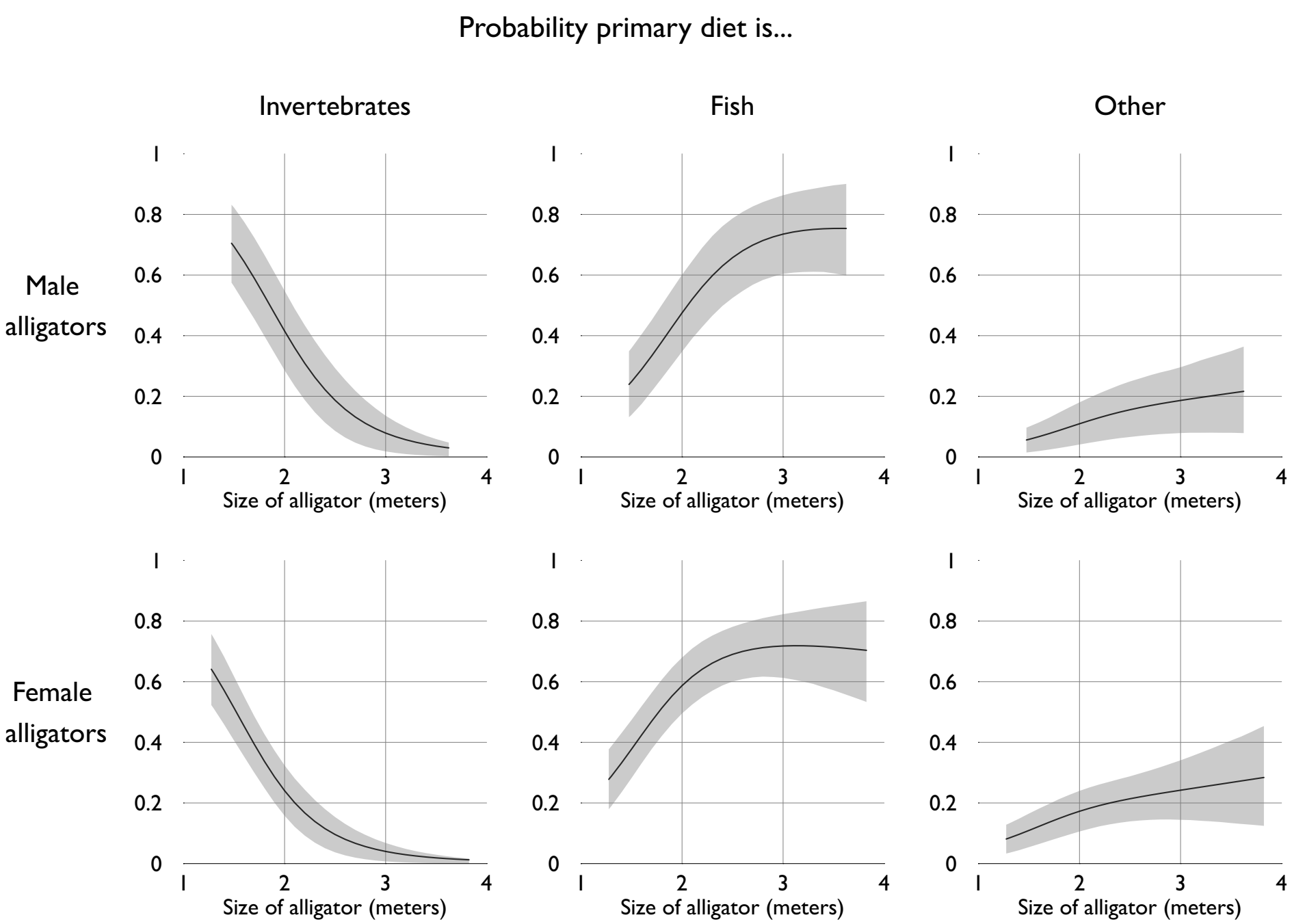


Don't forget female gators: a pair of plots now shows the full response surface

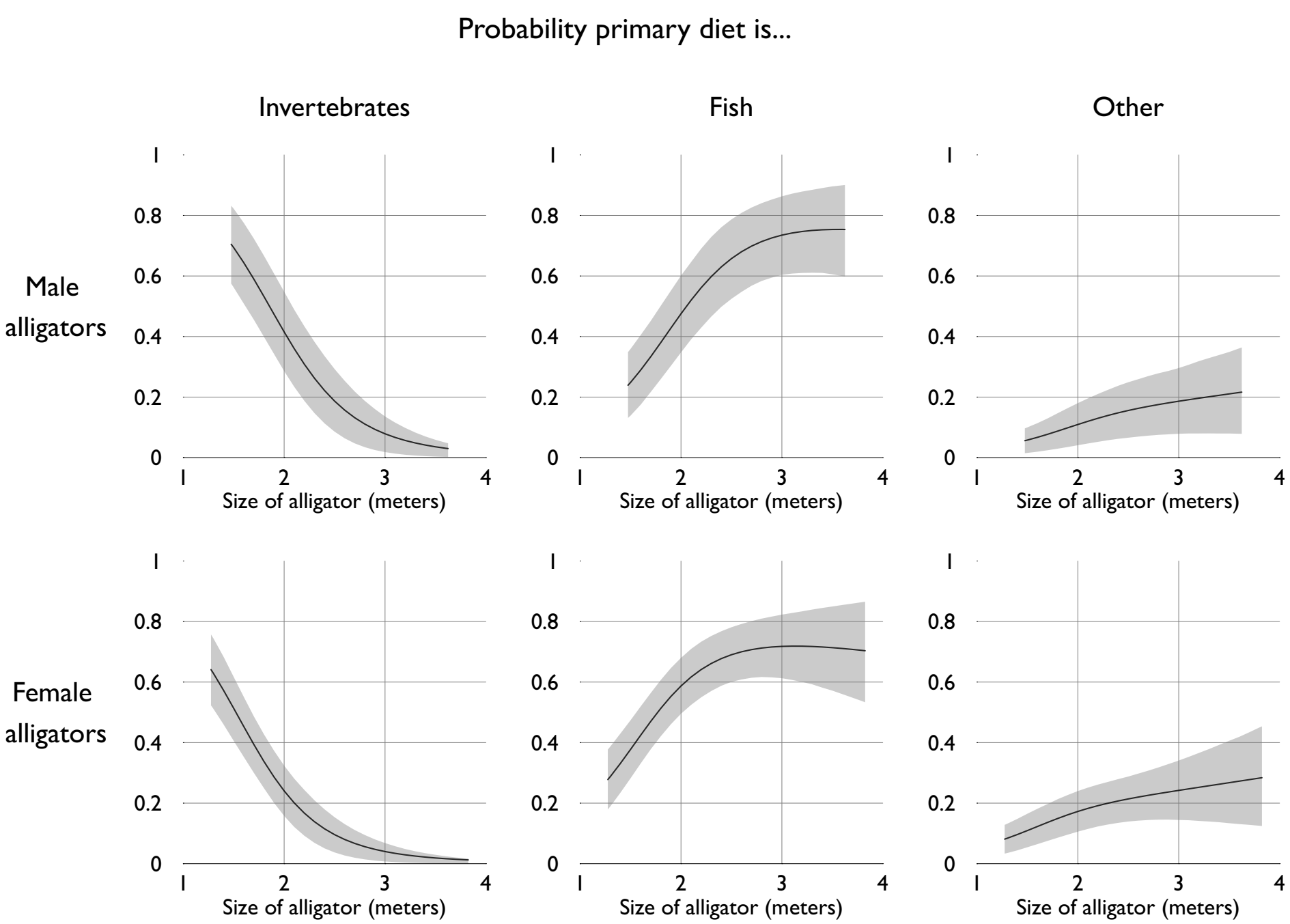
Possible here because specification so simple – usually need to be more selective

Graphs also usually get messy when you overlap >2 traces – we got lucky here

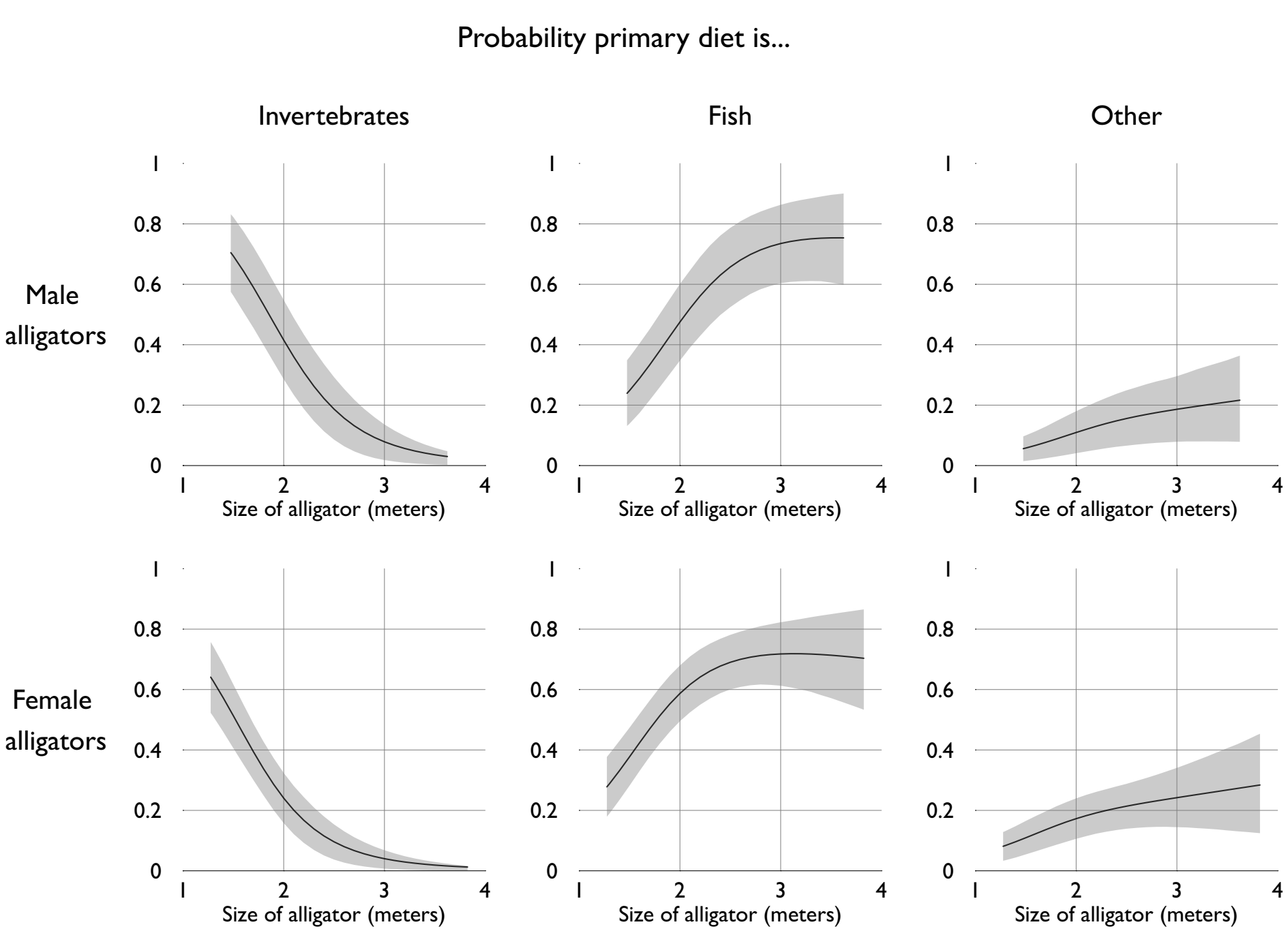
What should you do if the different categories have move overlapping probabilities?



Usually, the clearest presentation has at most one or two traces per plot

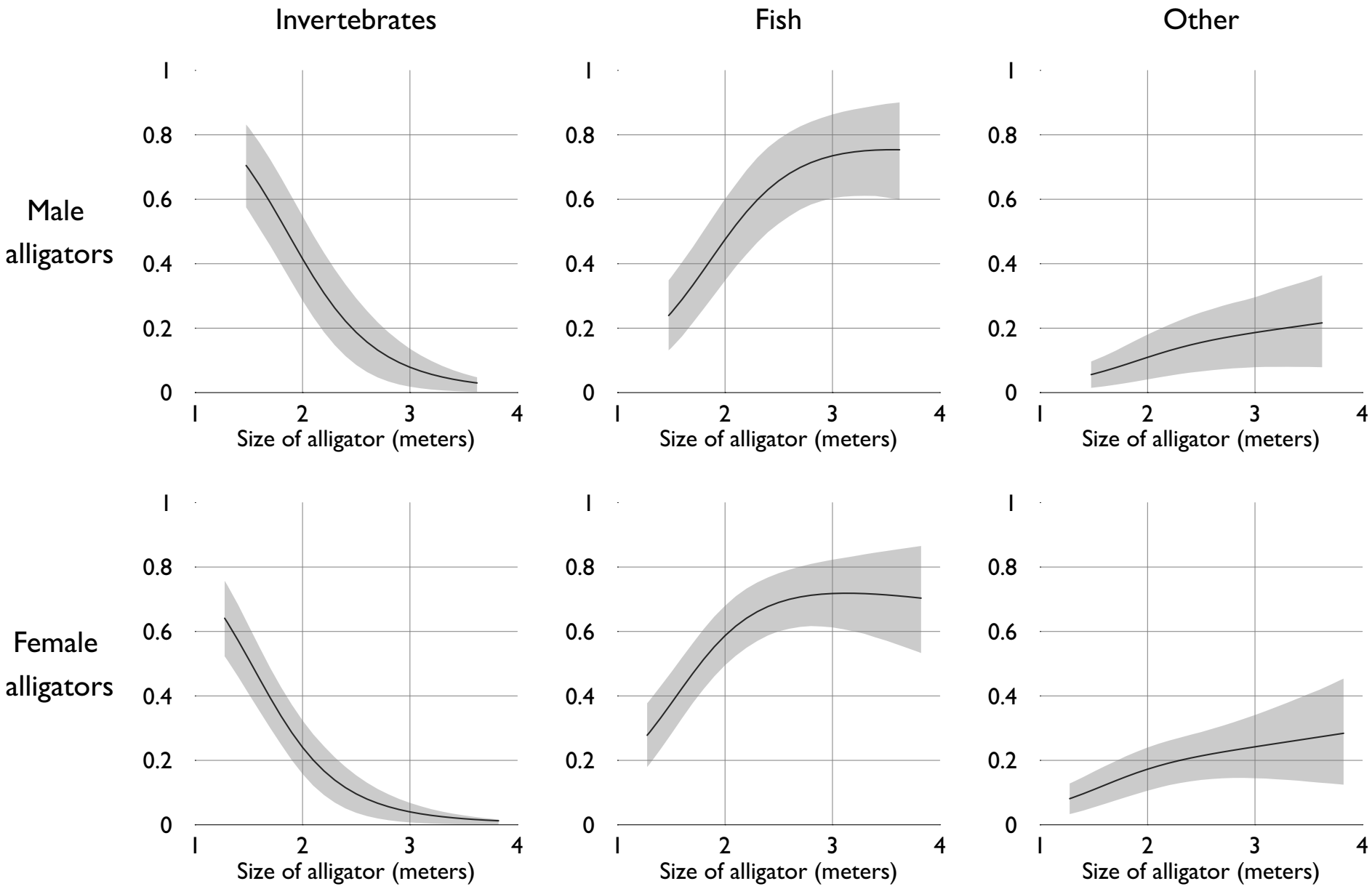


Just `tile` the plots for your scenarios and categories



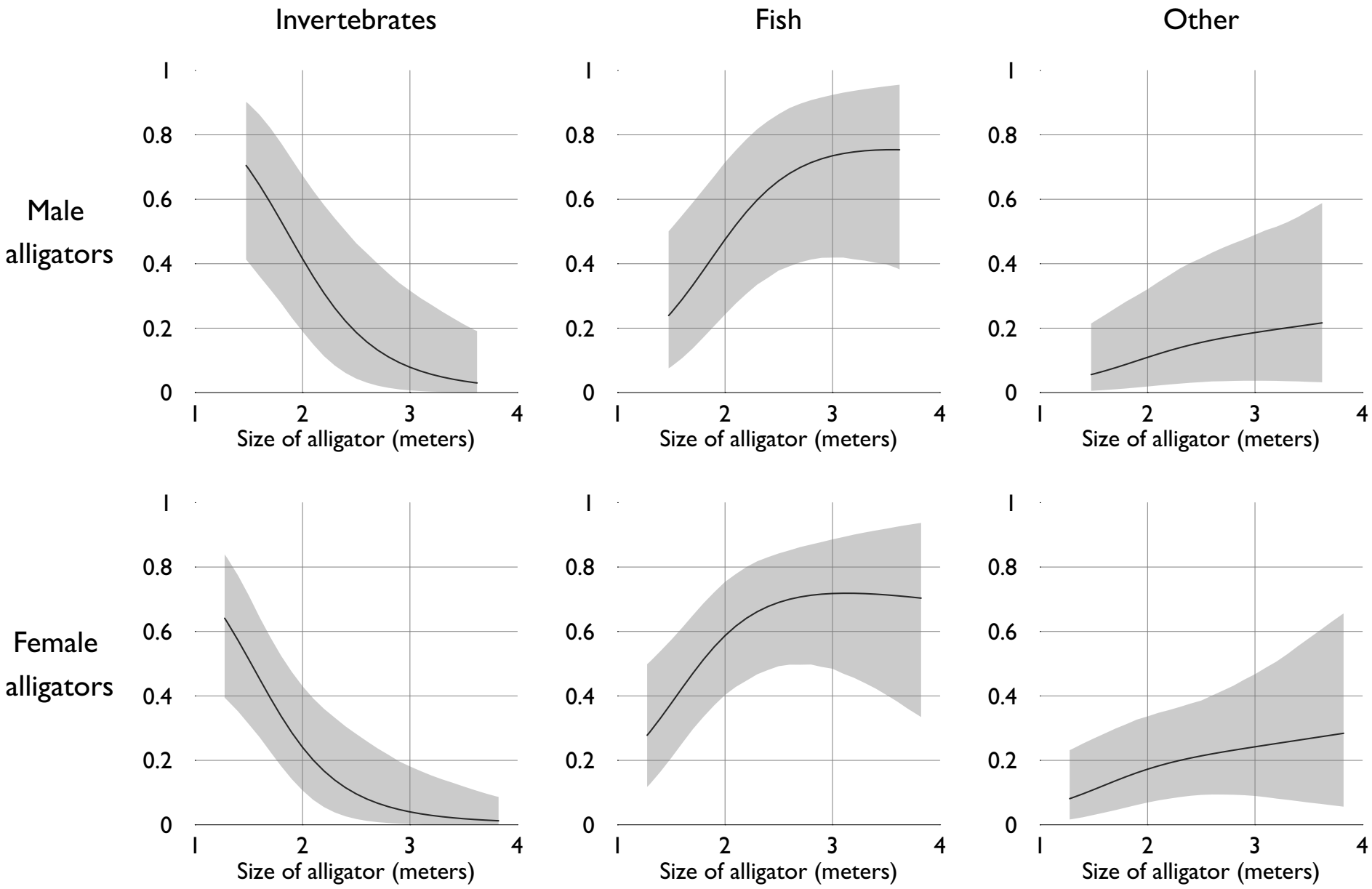
Make smaller plots simpler: remove extrapolated regions & now-unnecessary color

Probability primary diet is...

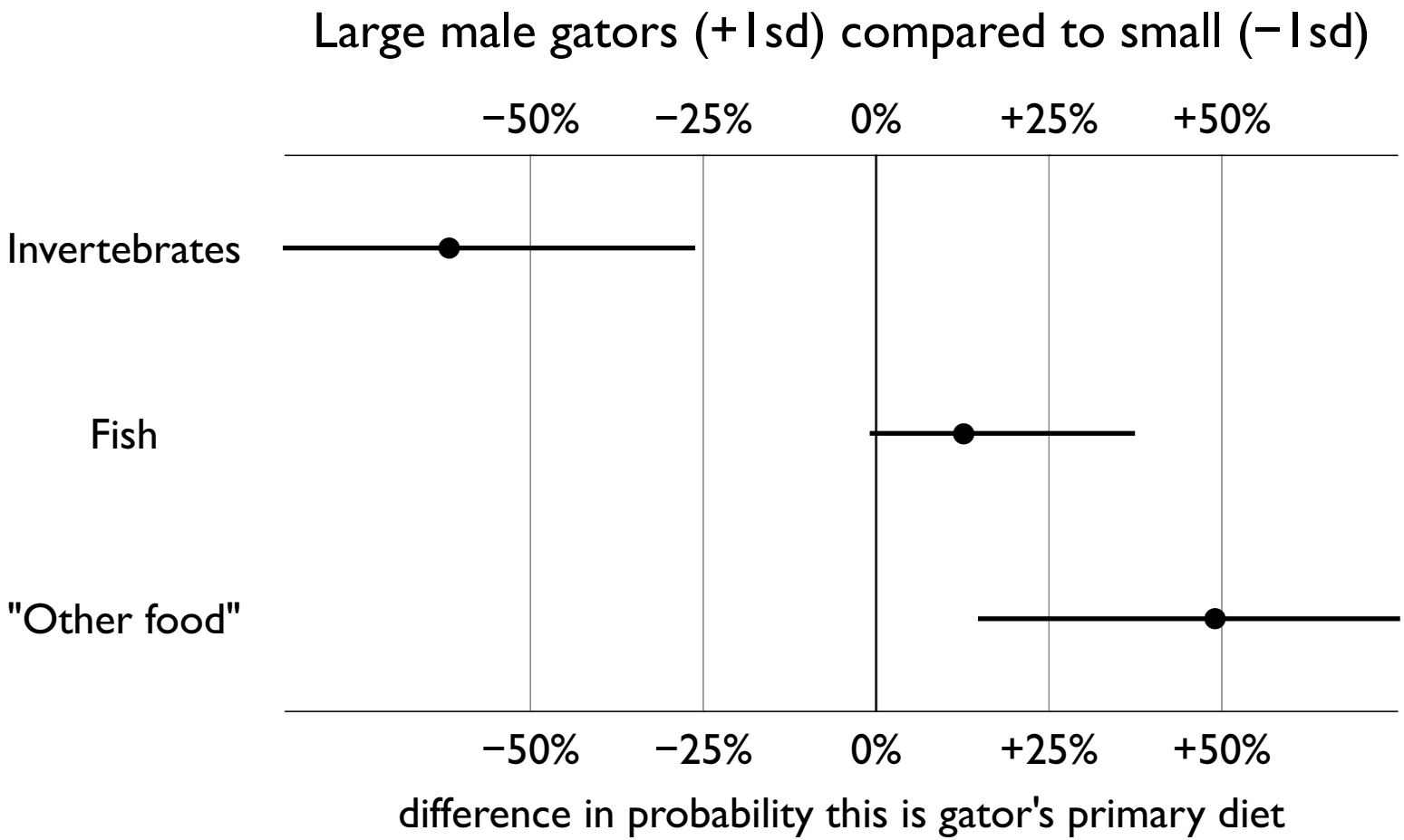


With no overlaps, we can show 95% CIs now without confusion

Probability primary diet is...



A good reminder that model is very uncertain – let's use 95% CI from here on

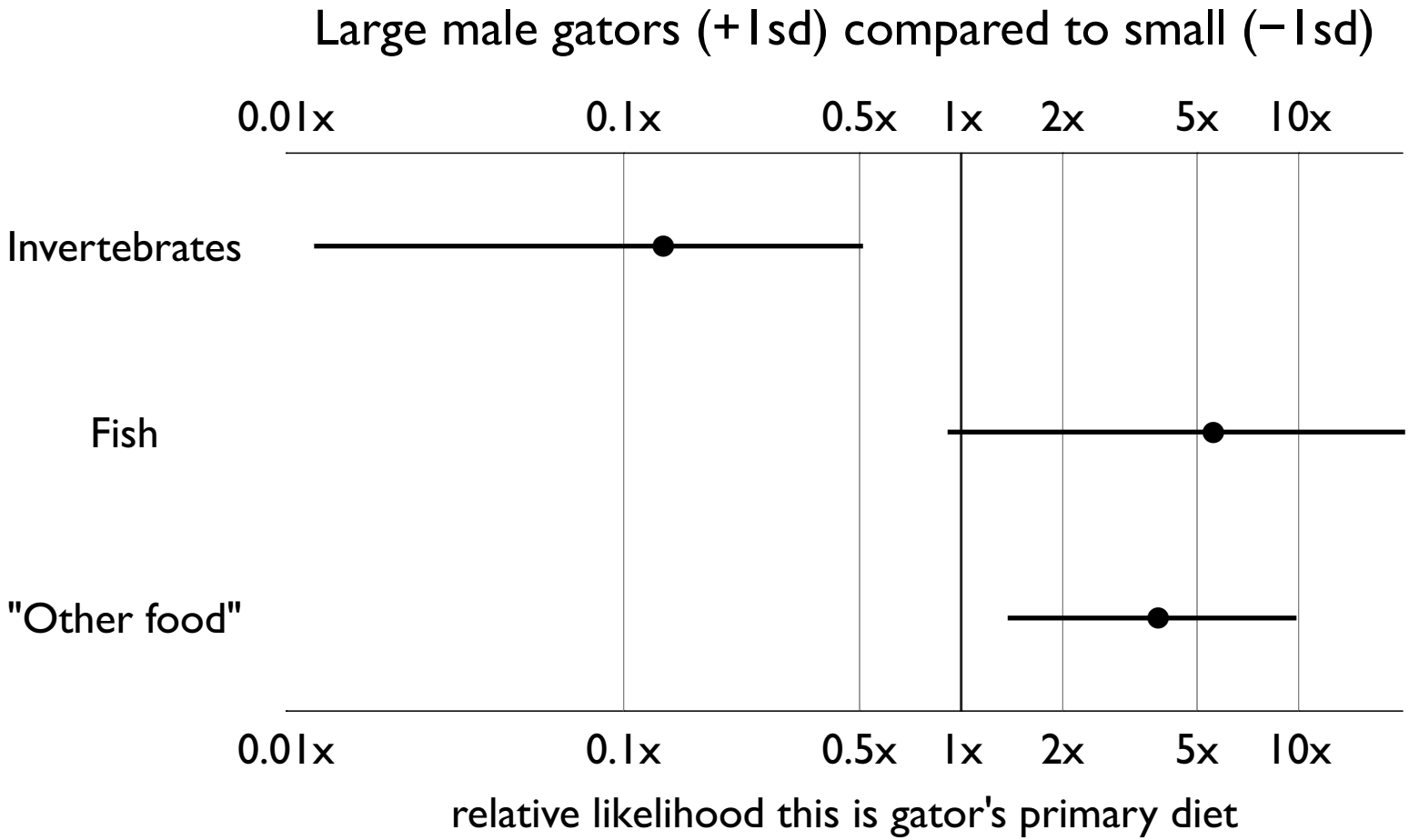


We don't *have* to show curves to visualize logit

We could just pick interesting scenarios:
like the difference between small and large male gators

Above are *first differences* between small and large males, *by category of food*

95% CI are shown as horizontal bars

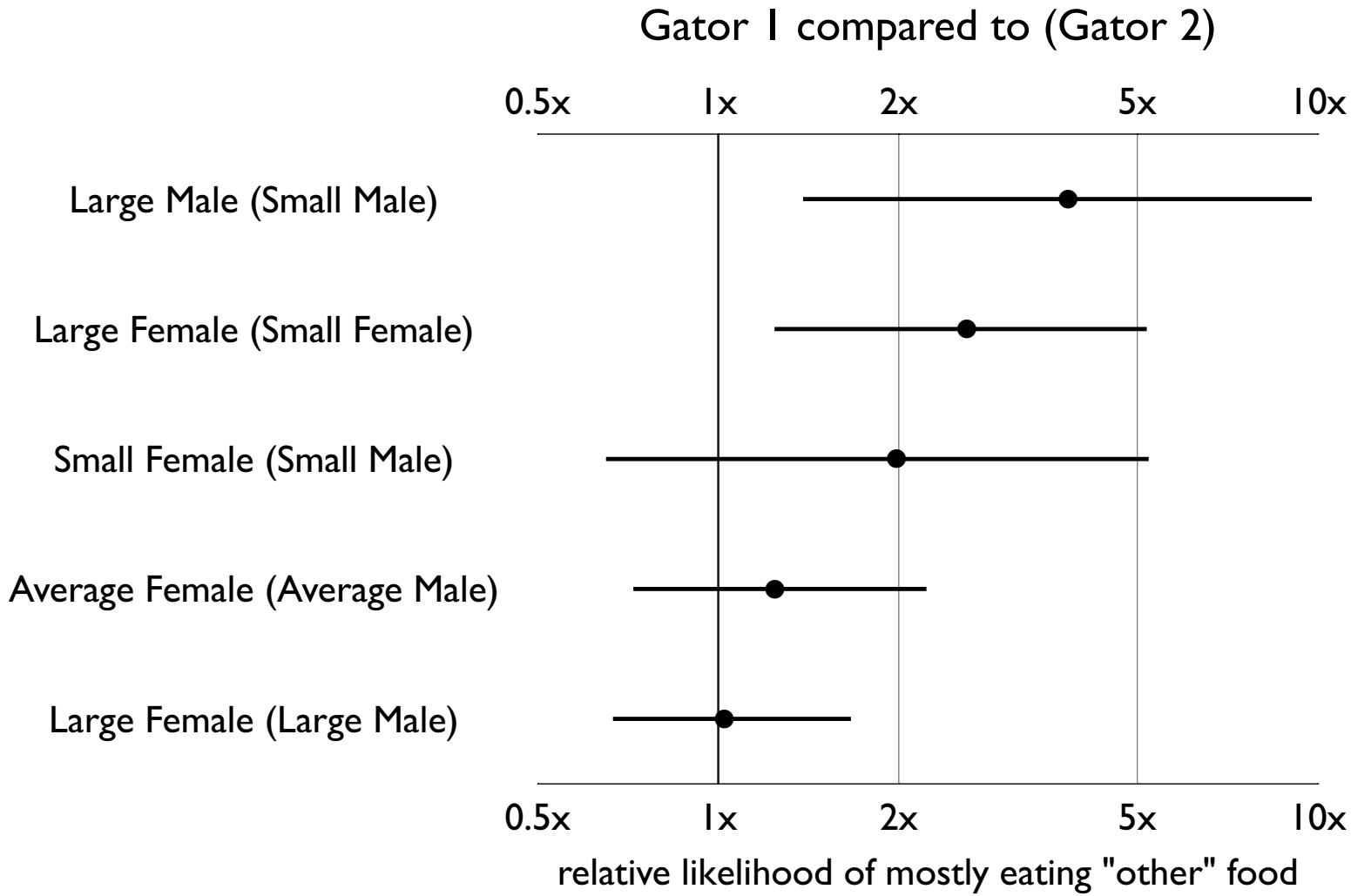


Relative risks are another way to show differences for categorical outcomes

It helps to log scale the axes for relative risks:

Here I've also relabeled them to emphasize the broad range of results & CIs

So far, we've only considered one scenario – how would we show more than one?



Let's focus on just one category of the outcome – the ominous "other food"

Above: relative risks of "other" dietary choices for 5 before-and-after scenarios

What can we say about the relationships among gator's food, size, and sex?

	$\log\left(\frac{\Pr(\text{invertebrate})}{\Pr(\text{fish})}\right)$	$\log\left(\frac{\Pr(\text{“other”})}{\Pr(\text{fish})}\right)$
Size	-2.526	0.138
	(0.848)	(0.518)
Female	-0.790	0.382
	(0.712)	(0.908)
Intercept	4.897	-1.947
	(1.706)	(1.531)
log likelihood	-48.3	
AIC	108.6	
% correctly classified (in-sample)	62.7% (null=52.5%)	
% correctly classified (LOO-CV)	55.9% (null=52.5%)	
in-sample concordance	0.71 (null=0.50)	
LOO-CV concordance	0.59 (null=0.50)	
N	59	

How well does the model fit?

Not all that well, especially in cross-validation: need a bigger sample!

"Equivalence" of MNL and Binary Logit

Before class today, how might you have tackled nominal data?

Perhaps equation-by-equation logit

Pick two categories, run a logit of the data that fall into one or the other . . .

(e.g., restrict attention to Fish and Invertebrate feeders only)

. . . then repeat with another pair of categories . . .

(next run a logit for Invertebrate and Other feeders only)

. . . and so on until the combinations are exhausted.

(finally, run a logit for Fish and Other feeders only)

"Equivalence" of MNL and Binary Logit

Drawbacks of equation-by-equation logit:

1. Time consuming, and produces a large pile of parameters

2. More seriously, each regression uses a different subset of observations:

   - Inefficient
   - Complicates significance tests

MNL is merely a more efficient version of equation-by-equation binary logit.

The same quantities are being estimated.

I.e., the parameters in MNL depend only on binary comparisons.

Limits of Multinomial Logit: IIA

In one sense a very general, or flexible, model:

No order imposed on y at all

What are the advantages of MNL vs LS for continuous data? disadvantages?

In several other senses somewhat rigid:

Covariates are fixed regardless of category (election example)

"Independence of Irrelevant Alternatives" (IIA) assumed

Next up:

Blue Bus, Red Bus conditional logit multinomial *probit*

Notation Review

Consider yourselves advanced to notation purgatory.
Just as bad as notation hell, but it won't last too much longer.

Observations: $1, \ldots, i, \ldots, N$

Unordered Categories: $1, \ldots, j, \ldots, M$

Category 1 will be the "reference category",
so we will often speak of the remaining categories $2, \ldots, M$

Covariates: $x_1, \ldots, x_k, \ldots, x_P$

Limits of Multinomial Logit

For many purposes, MNL is a limited model of nominal data.

Several restrictive assumptions with awkward names:

1. **No choice-specific variables:** $\mathbf{x}_k$ is fixed for all categories j; there's no x_{jk}'s

2. **Independence of irrelevant alternatives:** $\Pr(Y_i = m|\mathbf{x}_i, \boldsymbol{\beta}_m)/\Pr(Y_i = n|\mathbf{x}_i, \boldsymbol{\beta}_n)$ is unaffected by changes in the categories besides m and n

3. **Invariant proportion of substitution:** Given three categories, m, n, and o, and a change a covariate x_{mk}, the proportion of substitution from category n to m, relative to substitution from o to m, is insensitive to k

Goals:

- Develop an intuitive (plain English) understanding of these assumptions
- Understand options for coping with them

Limits of MNL: Unconditional Covariates

Assumption 1: $\mathbf{x}_k$ is fixed for all j

Covariates are individual- (or "chooser"-) specific,
Not choice specific.

Example:

Will a person order salmon sashimi, unagi, or California rolls?

One covariate might be x_{ik}, squeamishness about raw fish (individual-specific).

Another might be z_j, the price of each option (choice specific)

Now suppose our data include choices made in the following cities:

Seattle, Houston, Des Moines, Tokyo
Suggests a new variable:

Local price of each item, z_{ij}

$\rightarrow$ Both choice *and* chooser specific!

Limits of MNL: Unconditional Covariates

In MNL we had the following matrices of parameters & covariates

$$\boldsymbol{\beta} = \beta_{jk} = \begin{pmatrix} 0 & 0 & \dots & 0 & \dots & 0 \\ \beta_{20} & \beta_{21} & \dots & \beta_{2k} & \dots & \beta_{2P} \\ \vdots & & \ddots & & & \vdots \\ \beta_{j0} & & & \beta_{jk} & & \beta_{jP} \\ \vdots & & & & \ddots & \vdots \\ \beta_{M0} & \beta_{M1} & \dots & \beta_{Mk} & \dots & \beta_{MP} \end{pmatrix}$$

$$\mathbf{X} = x_{ik} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} & \dots & x_{1P} \\ 1 & x_{21} & \dots & x_{2k} & \dots & x_{2P} \\ \vdots & & \ddots & & & \vdots \\ 1 & & & x_{ik} & & x_{iP} \\ \vdots & & & & \ddots & \vdots \\ 1 & x_{N1} & \dots & x_{Nk} & \dots & x_{NP} \end{pmatrix}$$

And obtained

$$\boldsymbol{\mu}_j = \mathbf{X}\boldsymbol{\beta}_j$$

Adding choice specific variables: Conditional Logit

Now suppose instead we have a covariate $\mathbf{Z}_k$. In our sushi example, this is the price an individual faces for each kind of sushi.

$$\mathbf{Z}_k = z_{ij} = \begin{pmatrix} z_{11} & \cdots & z_{1j} & \cdots & z_{1M} \\ z_{21} & \cdots & z_{2j} & \cdots & z_{2M} \\ \vdots & \ddots & & & \vdots \\ z_{i1} & & z_{ij} & & z_{jM} \\ \vdots & & & \ddots & \vdots \\ z_{N1} & \cdots & z_{Nj} & \cdots & z_{NM} \end{pmatrix}$$

This matrix can viewed as a single covariate, or as the interaction of a covariate with each outcome.

Adding choice specific variables: Conditional Logit

Adding $\mathbf{Z}_k$ to the model makes this a "conditional" logit model.

The trick to conditional logit is that we assume a single parameter, γ_k links z_{ij} to μ_{ij}, regardless of the category.

That is, if $\mathbf{Z}_k$ is the price of sushi, the γ_k might measure our willingness to give up dollars for sushi.

Hence, for a model with only choice & chooser specific variables, we have

$$\mu_{ij} = \mathbf{z}_{ij}\boldsymbol{\gamma}$$

where $\boldsymbol{\gamma}$ and $\mathbf{z}_{ij}$ are k-vectors.

Adding choice specific variables: Conditional Logit

The probability model for conditional logit is analogous to MNL:

$$\Pr(y_i = j|\mathbf{z}_{ij}, \boldsymbol{\gamma}) = \frac{\exp(\mathbf{z}_{ij}\boldsymbol{\gamma})}{\sum_{\ell=1}^{M} \exp(\mathbf{z}_{ij}\boldsymbol{\gamma})}$$

but notice there is no warrant for an excluded category (we will want to have a constant for each category, though). The likelihood should also look familiar:

$$\log \mathcal{L}(\boldsymbol{\gamma}|\mathbf{y}, \mathbf{Z}) = \sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} \log \frac{\exp(\mathbf{z}_{ij}\boldsymbol{\gamma})}{\sum_{\ell=1}^{M} \exp(\mathbf{z}_{ij}\boldsymbol{\gamma})}$$

A more general form of conditional logit allows the choice to depend on both choice and individual specific variables; i.e.,

$$\mu_{ij} = \mathbf{z}_{ij}\boldsymbol{\gamma} + \mathbf{x}_i\boldsymbol{\beta}_j$$

which leads to an obvious generalization of the above probability and likelihood equations.

Adding choice specific variables: Conditional Logit

A subtle point:

We assume all the $\mathbf{z}_{ij}$'s for some covariate k are on the same scale. If this assumption holds, we can estimate fewer parameters.

An alternative is to replace the scalar z_{ijk} with the vector

$$\{z_{ijk} \times (y_{ij} = 2), \quad z_{ijk} \times (y_{ij} = 3), \quad \ldots, \quad z_{ijk} \times (y_{ij} = j)\}$$

include these variables in an MNL, which adds $J - 1$ parameters, instead of just 1.

Caution on nomenclature:

The term conditional logit is used to describe many different models. They are all nested in the model discussed here, but your statistical package may use conditional logit to describe a less general model.

R does not yet (?) have a multinomial conditional logit procedure capable of handling $\mathbf{Z}$ and $\mathbf{X}$ simultaneously.

But, using the techniques we've learned so far, I wrote a procedure that works for k covariates $\mathbf{X}$ and 1 covariate $\mathbf{z}$ which you could further generalize if you need it . . .

Today's (Fake) Example: Voting Choice

Because of the complexity and touchiness of today's models, it will help to have an example where we know the true parameters.

Today's (Fake) Example: Voting Choice

Hence, I generated the following FAKE data on voting behavior.
I'm making all this up.

Vote	Choice among a Liberal, a Conservative, and a Populist
Rural	Inverse population density for each voter's locale; suppose more rural voters really like populists ($\beta = 1.5$), kind of like conservatives ($\beta = 1$), and dislike liberals (the omitted category; $\beta = 0$)
Religion	Voter religiosity; More religious voters like conservatives ($\beta = 1$), are tepid about populists ($\beta = 0.5$), and shun liberals (the omitted category, $\beta = 0$)
Distance	The (absolute) ideological distance between the voter and the candidate. Voter ideology is distributed standard normal. Candidate ideology: Liberal = -1, Populist = 0.5, Conservative = 1 $\gamma = -1$; voters are less likely to vote for distant candidates

I generate Vote from three latent variables which are distributed multivariate normal. (Don't worry about the vc matrix $\mathbf{\Sigma}$ for now). Whichever latent variable is largest is the person's vote choice.

Today's (Fake) Example: Voting Choice

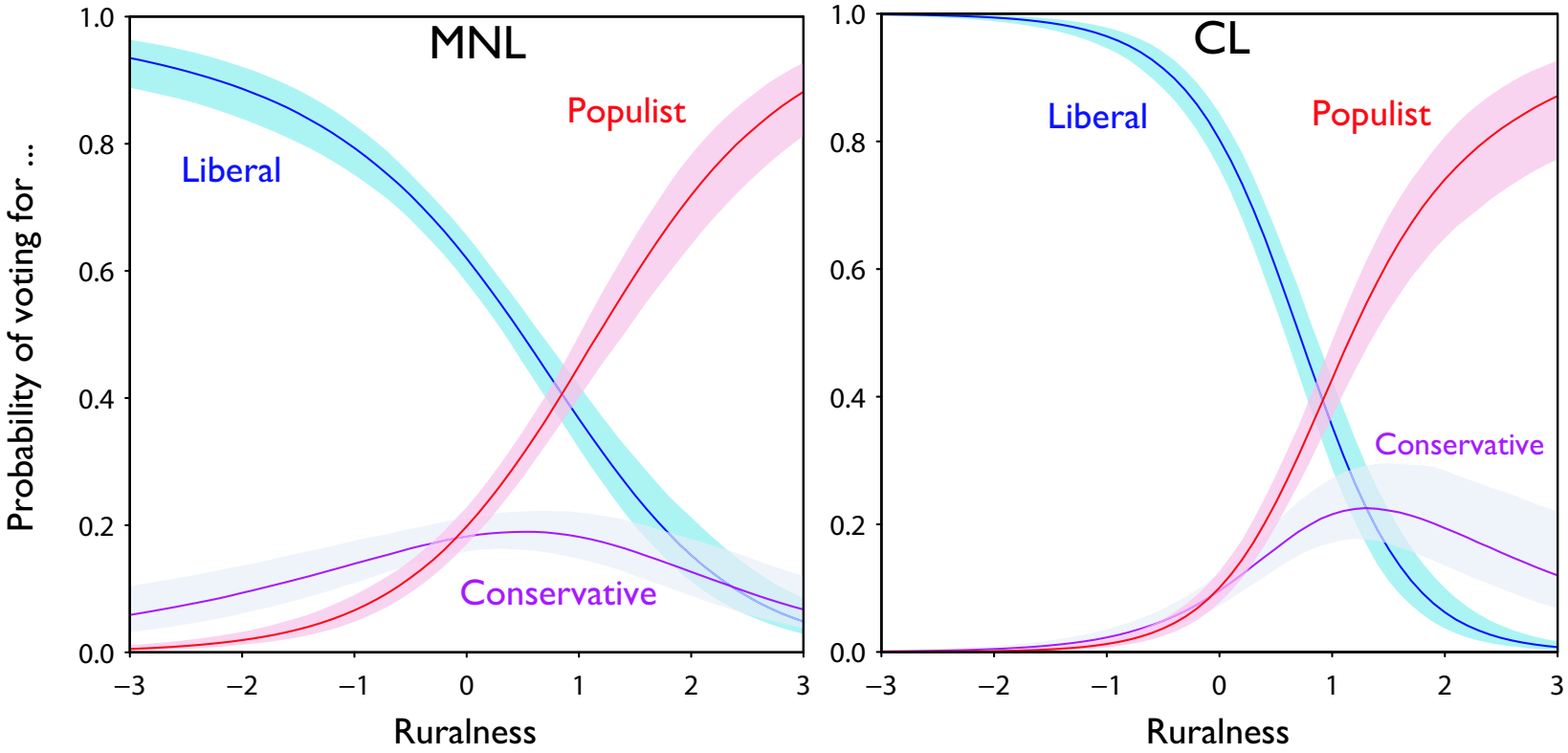
Covariates	Truth (on probit scale)	Logit coefficients MNL	CL
Constant (Cons vs. Lib)	-0.25	-1.15	-2.08
Constant (Pop vs. Lib)	-0.50	0.43	-2.05
Ruralness (Cons vs. Lib)	1.00	0.75	1.79
Ruralness (Pop vs. Lib)	1.50	-1.20	2.38
Religion (Con vs. Lib)	1.00	1.39	1.80
Religion (Pop vs Lib)	0.50	0.15	0.94
Ideological distance	-1.00		-2.42
N		1000	1000
$\log \mathcal{L}$		-824.2	-482.0

The estimated parameters are not transparently comparable to the true values (sorry). However, recalling that logit parameters tend to be about 1.6x as big as probit paramters, the CL estimates look pretty good, while the MNL estimates are lousy.

Notice the large difference in the maximum likelihood.

Today's (Fake) Example: Voting Choice

A more readable presentation of the results is to plot the expected values against size.

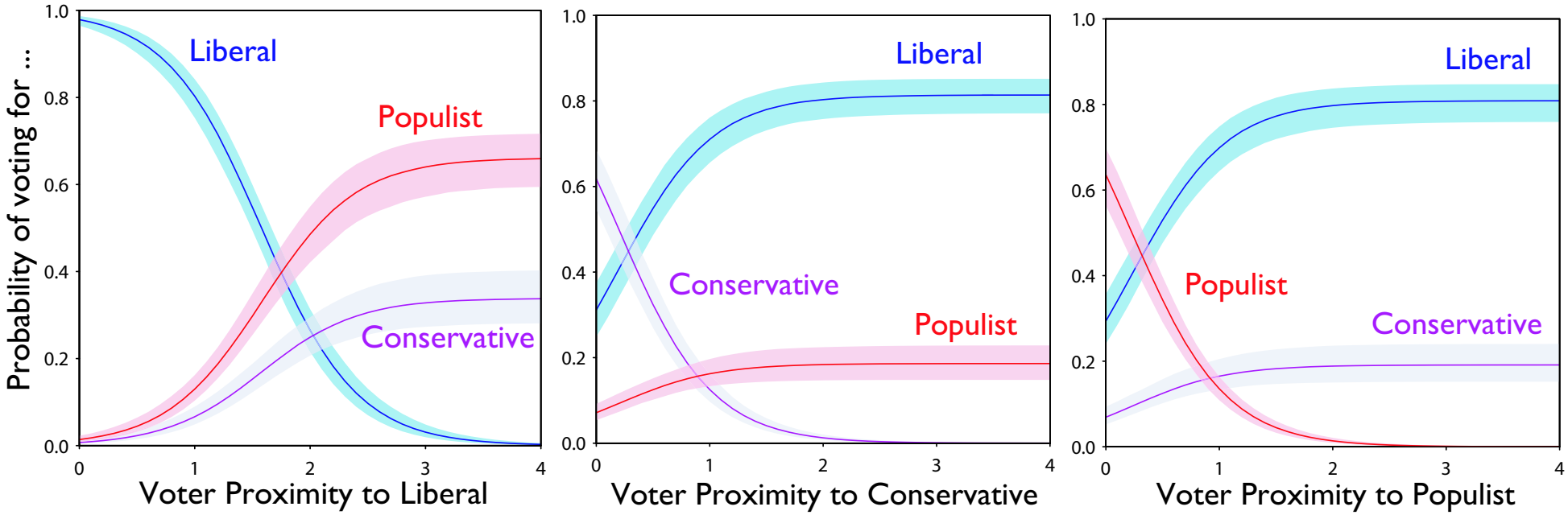


MNL estimates are on the left; CL estimates on the right. All other variables (including, for CL, the ideological distances) are held at their mean values.

Notice that the results are quite dependent on the model/omission of ideological distance.

Today's (Fake) Example: Voting Choice

Now vary voter-candidate distance, holding other voter characteristics constant



Need three separate sets of counterfactuals: each varying the distance to a different candidate.

Bottom line: If you have choice-specific covariates, CL is a practical alternative to MNL that is as easy to estimate, and nearly as "easy" to interpret.

Limits of MNL and conditional logit: IIA

The famous "blue bus"/"red bus" paradox illustrates a limitation of MNL and CL.

In MNL and CL, the odds of selecting option A versus B do not depend on the presence or characteristics of option C.

Hence, if transportation options are initially red bus, car, train, and each is equally probable (1/3), then the model predicts that adding a blue bus will lower each category to probability 1/4. A coat of paint increases bus ridership by 17 percent!

For a model containing all possible categories, IIA is innocuous.

For a model of classification, IIA may be plausible.

For a model of choice, and especially one with choice-specific variables, IIA is very difficult to swallow. (Any counterfactual choice of $\mathbf{Z}_k$ should violate it).

IIA violations are a big concern in the study of voting and transportation –
as you might guess from the examples used here

Latent Variables, again

To relax the IIA property, we return to the latent variables framework.

Assume there is a latent variable y_{ij}^* for every observed y_{ij}.

Further suppose that the latent variables are (potentially) jointly distributed.

Finally, suppose $y_{ij}^* > y_{i\ell}^* \quad \forall j \neq \ell \qquad \Rightarrow y_{ij} = 1, \quad y_{i\ell} = 0.$

We could interpret the MNL or CL model in this way by assuming iid extreme value distributions for y_{ij}^*.

But the (multivariate) normal turns out to be a much more flexible choice.

The reason is that the MVN allows us to estimate or assume non-zero covariances among the latent variables.

This breaks the IIA restriction – now a new category can change the ratio of probabilities of two other categories with respect to each other.

Multinomial Probit

As in multinomial logit, we will restrict the $\boldsymbol{\beta}$ parameters of the first category to 0

Why?

As usual, we don't know the true location or scale of the latent variables

That's okay: we only need to know which is largest

For convenience,
and to avoid the impossible problem of estimating the true location of the latent variables,
we set one category to 0 to anchor the scale

We do this by setting $\boldsymbol{\beta}_1 = 0$, which implies $\mathbb{E}(y_{i1}^*) = 0$

And by setting $\text{var}(y_{i1}^*) = 0$ also, so $y_{i1}^* = 0 \quad \forall i$

This implies that to "get out of category 1," one of the other $y_{i\ell}^*$, $\ell \neq 1$ must be > 0

Multinomial Probit

But what makes multinomial probit special is that the latent variables are assumed to be distributed jointly with covariance matrix $\mathbf{\Sigma}$

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix}$$

Identification of the elements of $\mathbf{\Sigma} = \sigma_{mn}$ is hard.

Multinomial Probit

But what makes multinomial probit special is that the latent variables are assumed to be distributed jointly with covariance matrix $\mathbf{\Sigma}$.

$$\mathbf{\Sigma} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \sigma_{22} & \sigma_{23} \\ 0 & \sigma_{32} & \sigma_{33} \end{pmatrix}$$

Identification of the elements of $\mathbf{\Sigma} = \sigma_{mn}$ is hard.

We will assume the first row and column of $\mathbf{\Sigma}$ are all 0.

Multinomial Probit

But what makes multinomial probit special is that the latent variables are assumed to be distributed jointly with covariance matrix $\mathbf{\Sigma}$.

$$\mathbf{\Sigma} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & \sigma_{23} \\ 0 & \sigma_{32} & \sigma_{33} \end{pmatrix}$$

Identification of the elements of $\mathbf{\Sigma} = \sigma_{mn}$ is hard.

We will assume the first row and column of $\mathbf{\Sigma}$ are all 0 – this is part of treating category 1 as the reference category

We need one more restriction for identification; conventionally, we set $\sigma_{22} = 1$

Why? Only the non-reference latent variables are allowed to vary,
but we still don't have true scales for them

To make relative comparisons among the latent variables, we need to anchor one scale by choosing its variance

Multinomial Probit

But what makes multinomial probit special is that the latent variables are assumed to be distributed jointly with covariance matrix $\mathbf{\Sigma}$

$$\mathbf{\Sigma} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & \sigma_{32} \\ 0 & \sigma_{32} & \sigma_{33} \end{pmatrix}$$

Identification of the elements of $\mathbf{\Sigma} = \sigma_{mn}$ is hard

We will assume the first row and column of $\mathbf{\Sigma}$ are all 0

We need one more restriction for identification; conventionally, we set $\sigma_{22} = 1$

In the 3 category MNP, this leaves 2 covariances to estimate (σ_{32} and σ_{33})

As J increases, the number of free σ's grows rapidly

In real data, the likelihood over these σ's may be (almost) flat,
and additional restrictions may be necessary to get usable estimates

Multinomial Probit: Likelihood

But identification is the lesser difficulty with MNP.
The likelihood for MNP, as for MNL, takes the form:

$$\mathcal{L}(\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_M, \boldsymbol{\Sigma} | \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{N} \prod_{j=1}^{M} {p_{ij}}^{y_{ij}}$$

But p_{ij} now contains potentially high order integrals than cannot be solved analytically, and which remain difficult to solve computationally. For example, in a three category MNP, the probability that $Y_i = 1$ is:

$$p_{i1} = \int_{-\infty}^{\frac{\mu_{i1}-\mu_{i2}}{\sqrt{\sigma_{11}+\sigma_{22}-2\sigma_{12}}}} \int_{-\infty}^{\frac{\mu_{i1}-\mu_{i3}}{\sqrt{\sigma_{11}+\sigma_{33}-2\sigma_{13}}}} \Phi\left(\text{vec}[\varepsilon_{i2}-\varepsilon_{i1}, \varepsilon_{i3}-\varepsilon_{i1}], \text{cov}[\varepsilon_{i2}-\varepsilon_{i1}, \varepsilon_{i3}-\varepsilon_{i1}]\right)$$

This likelihood is extremely difficult to maximize for "large" J (like 4 or 5).

Serious processing power and clever techniques help cope.
For example, Imai and van Dyk (2005) offer a Bayesian MCMC approach to MNP that may dominate past estimation strategies

Still isn't easy; plus there is the added necessity of assessing MCMC convergence . . .

Multinomial Probit: Example

We return to our FAKE voting example,
though we will now drop the choice specific variable to make things simpler.

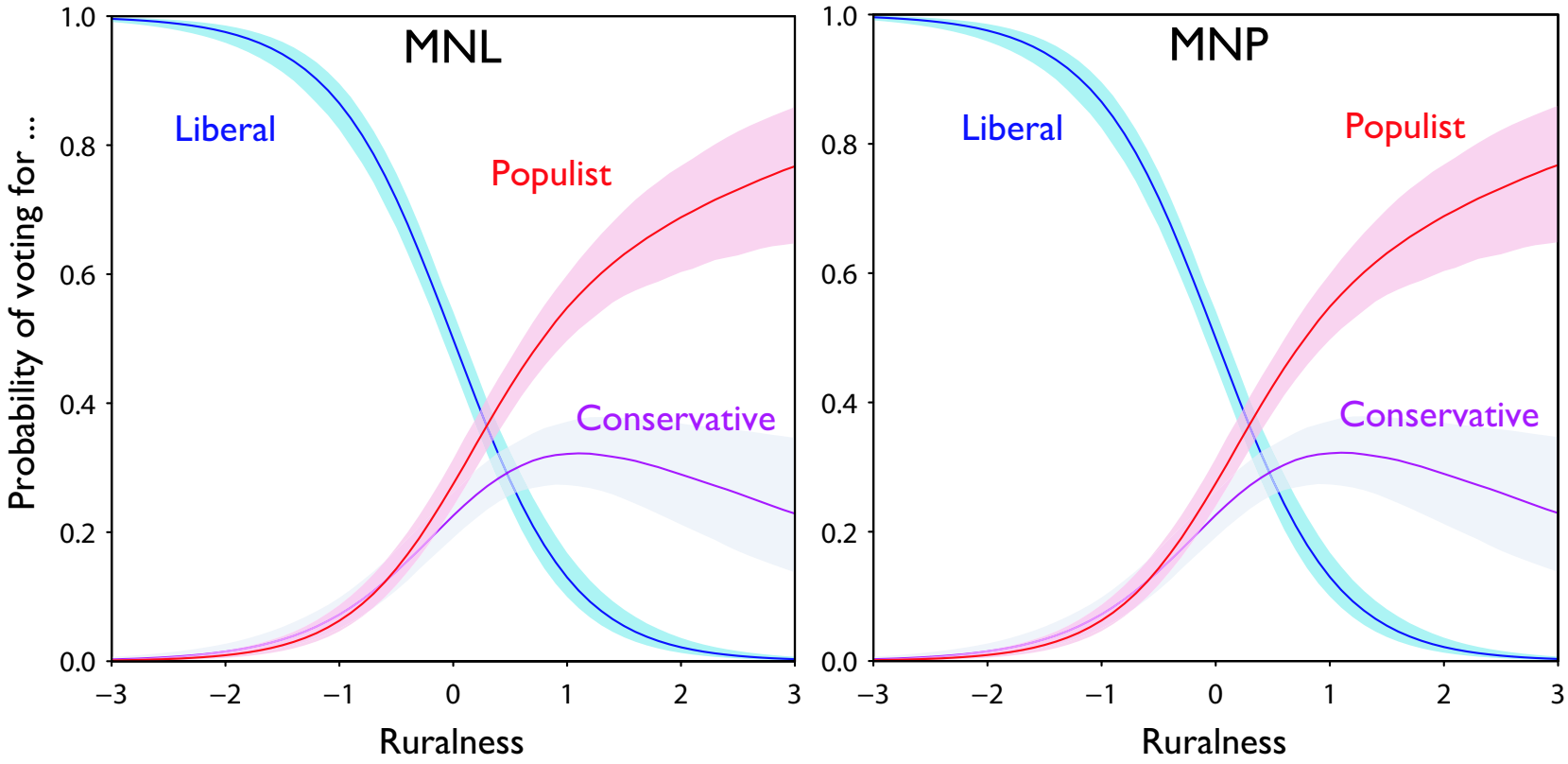
Covariates	Truth (on probit scale)	Multinomial Probit mean	95% credibility
Constant (Cons vs. Lib)	-0.25	-0.17	[-0.38, 0.05]
Constant (Pop vs. Lib)	-0.50	-0.45	[-0.81, -0.20]
Ruralness (Cons vs. Lib)	1.00	1.05	[0.82, 1.27]
Ruralness (Pop vs. Lib)	1.50	1.56	[1.12, 1.98]
Religion (Con vs. Lib)	1.00	0.95	[0.80, 1.11]
Religion (Pop vs Lib)	0.50	0.54	[0.33, 0.76]
$\sigma_{\text{Cons,Pop}}$	0.50	0.82	[0.26, 1.32]
$\sigma_{\text{Pop,Pop}}$	2.00	2.00	[0.85, 3.51]

Estimated by Imai & van Dyk (2005) data augmentation MNP procedure

Warning: This was not a typical run . . .

Multinomial Probit: Example

As with every other model in this class, we can calculate EVs and simulate confidence intervals . . .



[Why are these the same? No $\mathbf{z}_k$ is the likely reason. But the non-zero covariance estimate tells us that things would change if we added more categories]

Limits of MNL, MNP, etc.: the IPS property

Consider the following example from Steenburgh (2004).
Imagine you are looking for a new laptop, and can choose among three alternatives.

Choice	Weight	Speed
Laptop A	3 lb.	2.0 GHz
Laptop B	5 lb.	2.7 GHz
Laptop C	7 lb.	3.4 GHz

Limits of MNL, MNP, etc.: the IPS property

Let's make the example (a little) more concrete:

Choice	Weight	Speed	Ex ante Pr()
Laptop A	3 lb.	2.0 GHz	0.30
Laptop B	5 lb.	2.7 GHz	0.40
Laptop C	7 lb.	3.4 GHz	0.30

Imagine a small (read, infinitesimal) improvement in the speed of B.

This will increase the probability of buying Laptop B.

It will reduce the probability of buying Laptops A and C.

The reduction in the latter probabilities can be stated formally

$$\frac{\partial \Pr(\mathbf{y} = m)/\partial \mathbf{x}_{nk}}{\partial \Pr(\mathbf{y} = n)/\partial \mathbf{x}_{nk}} = \psi_{m,n}$$

where Invariant Proportion of Substitution (IPS) implies $\psi_{m,n}$ is a constant for all k

Limits of MNL, MNP, etc.: the IPS property

Choice	Weight	Speed	Ex ante Pr()
Laptop A	3 lb.	2.0 GHz	0.30
Laptop B	5 lb.	2.7 GHz	0.40
Laptop C	7 lb.	3.4 GHz	0.30

IPS means that if a small increase in speed increases the probability of purchasing B by 10 "units" (where units are small), while laptop A decreases by 4 units and C by 6 units,

then

if a small change decrease in weight increases the probability of purchasing B by 10 units, IPS models constrain decrease in A to be 4 units, and the decrease in C to be 6 units.

Limits of MNL, MNP, etc.: the IPS property

Choice	Weight	Speed	Ex ante Pr()
Laptop A	3 lb.	2.0 GHz	0.30
Laptop B	5 lb.	2.7 GHz	0.40
Laptop C	7 lb.	3.4 GHz	0.30

But wait! Shouldn't a drop in B's weight hurt A more than C?

People who compromised on speed to get a light laptop may think "the weight benefit is getting small; I'm going for the bigger processor."

But many C buyers probably went large for the processor, and will ignore fluctuations in the weight of smaller laptops, so long as they are slower.

Limits of MNL, MNP, etc.: the IPS property

Choice	Weight	Speed	Ex ante Pr()
Laptop A	3 lb.	2.0 GHz	0.30
Laptop B	5 lb.	2.7 GHz	0.40
Laptop C	7 lb.	3.4 GHz	0.30

A similar logic led us to assume that a speed increase in B would hurt the speed leader, C, more than it would hurt A.

But in the models we have studied—MNL, CL, MNP—we can't have it both ways.

IPS prevents a model from estimating theoretically reasonable asymetries in substitution.

Limits of MNL, MNP, etc.: the IPS property

IPS is a new idea (surprisingly?)

Few existing models generalize out of it

One that does is the "mother logit" (not my name)

The "mother logit" includes all (possible) categories; moreover, any variable enetering any equation enters the every other categories' equation.

This "complete" specification obviates all concerns discussed today

But you will probably never see this method used ever. . .

So which method do I use for nominal data?

As usual, a model isn't right or wrong, but more or less useful.

If your data are non-ordered, you will probably want a multinomial model, or your results won't make much sense

But sometimes an ordering is arguable for a particular instance.

Candidate for multinomial logit:
"What is your favorite color?"

Candidate for ordered probit:
"What is the shortest wavelength of color a species can see?"

This is up to the modeler, and her substantive knowledge to decide

So which method do I use for nominal data?

Once we've decided to use a model for nominal data, we need to make more decisions:

1. Do my data depend on choice specific variables?
Notice the word *choice*

2. Would some of my categories shrink proportionally more than others is a particular alternative were added
Notice the hypothetical addition of a new *option*

3. Would some changes in choice m make n relative less appealing compared to some other changes in m?
Notice the focus on categories substituted *from*, rather than *to*

So which method do I use for nominal data?

1. Do my data depend on choice specific variables?

2. Would some of my categories shrink proportionally more than others is a particular alternative were added

3. Would some changes in choice m make n relative less appealing compared to some other changes in m?

If the answer to all three is "no", then MNL is a good option.

Another way of putting this:
The (relatively) restrictive assumptions of MNL may be appropriate;
if so, why choose a more complicated model?

This is especially likely in classification problems, but not so tenable in choice problems.

So which method do I use for nominal data?

1. Do my data depend on choice-specific variables?

2. Would some of my categories shrink proportionally more than others is a particular alternative were added

3. Would some changes in choice m make n relative less appealing compared to some other changes in m?

If the answers are "yes, no, no", you have two simple options:

1. Include charactistics of every choice in every equation (as in the "mother" logit). Very demanding specification

2. Impose some structure:
Assume the choice-specific variables operate on a common metric,
and estimate a single parameter for each: conditional logit.

So which method do I use for nominal data?

1. Do my data depend on choice-specific variables?

2. Would some of my categories shrink proportionally more than others is a particular alternative were added

3. Would some changes in choice m make n relative less appealing compared to some other changes in m?

If your answers are "yes, yes, no", the question arises: Could there be an omitted option?

If every conceivable option is in the model, Red Bus/Blue Bus paradoxes won't arise (no one can add a Blue Bus!) In this case, logit is fine.

But new choices might emerge
(e.g., new candidates will join a race, or new products will enter a market)
$\rightarrow$ in this case, MNP will provide useful information about which options are close substitutes, helping understand likely responses to new options

A warning: in some sense, saying CL is adequate is
tantamount to saying unobserved values of $\mathbf{Z}$ are nonsensical and/or
the covariances in $\mathbf{\Sigma}$, or some analogous matrix, are 0.

So which method do I use for nominal data?

1. Do my data depend on choice-specific variables?

2. Would some of my categories shrink proportionally more than others is a particular alternative were added

3. Would some changes in choice m make n relative less appealing compared to some other changes in m?

Worried about all three? You may need a very flexible model, approaching the "mother logit" specification.

Clever specification could deal with limited violations of IPS (the property violated in 3.)

The IPS property is a new idea. . . so maybe we'll get (or make) some new methods to deal with it.

Censored data: Tobit

What do we mean by censored data?

Recall the latent variable justification for the probit model.

Suppose the latent variable y^* is partially observed,
so that above the cutpoint, we have y^*.

If the data are below the cutpoint, all we know is that fact.

These are censored data (contrast with truncated and latent)

Examples of censored data:

Survey questions on income that end with "100k or above"

Desired contribution to an 401(k) above the cap

Demand for tickets to (sometimes) sold-out concerts

Note that we are in much better shape than in regular probit:
we can identify the scale and location of the latent variable

Censored data: Tobit

How do we analyze censored data?

1. We could use linear regression on the whole dataset
   $\rightarrow$ this turns out to be inconsistent
   i.e., filling in values for the censored data isn't kosher

2. We could use linear regression on the fully observed data
   $\rightarrow$ this introduces sample selection bias
   i.e., the $\mathbf{x}$s and error terms are now correlated

3. We could model the censoring explicitly, using a probit style model
   $\rightarrow$ under certain assumptions, this is unbiased and efficient
   if we had thought of this ourselves, we'd be part-way to a Nobel. . .

Censored data: Tobit

Tobin's Probit = Tobit

Assume the censored variable is normally distributed

$$y_i^* = f_{\mathcal{N}}\left(\mu_i, \sigma^2\right)$$

and assume a deterministic observation (or censoring) mechanism

$$y_i = \begin{cases} c & \text{if } -\infty < y_i^* \leq \tau \\ y_i^* & \text{if } \tau < y_i^* \leq \infty \end{cases}$$

Note that c is an mostly-arbitrary constant, i.e., the label we attach to censored cases; τ is the threshold below which we always have censoring.

We could flip this around to have censoring above τ.

We could have censoring in multiple regions
(e.g., above and below, as in "two-limit" tobit)

Overall, this is very like probit below τ, and very like linear regression above τ

Censored data: Tobit

The tobit likelihood is formed in two parts:

Part 1: Observed cases.

The probability for observed cases is simply the normal pmf, hence

$$\mathcal{L}_1(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}, \tau) = \prod_{y_i > \tau} f_{\mathcal{N}}\left(y_i^* | \mu_i, \sigma^2\right)$$

Part 2: Unobserved cases.

The probability for an unobserved case is just like probit

$$\begin{aligned} \Pr(y_i \leq \tau) &= \int_{-\infty}^{\tau} f_{\mathcal{N}}\left(y_i^* | \mu_i, \sigma^2\right) dy^* \\ &= F_{\mathcal{N}}(\tau | \mu_i, \sigma^2) \end{aligned}$$

so

$$\mathcal{L}_2(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}, \tau) = \prod_{y_i \leq \tau} F_{\mathcal{N}}(\tau | \mu_i, \sigma^2)$$

Censored data: Tobit

The full likelihood is just the product of the likelihoods for the censored and uncensored portions of the data

$$\mathcal{L}(\boldsymbol{\mu}, \sigma^2 | \mathbf{y}, \tau) = \prod_{y_i > \tau} f_{\mathcal{N}}\left(y_i^* | \mu_i, \sigma^2\right) \prod_{y_i \leq \tau} F_{\mathcal{N}}(\tau | \mu_i, \sigma^2)$$

Tobit can be estimated by ML as usual, using `optim()`.

Survival analysis packages somtimes have tobit or "censored regression"

Tobit is also available in the `VGAM` package using:

```
vglm(..., family=tobit(Upper=K))
```

where `K` is the censoring limit. You could also/instead have `Lower=M`.

Finally, very flexible censoring (with many limits) is available in the `censReg` package

Censored data: Tobit

Interpretation:

Parameters can be interpreted as linear regression coefficients (Yay!)

Could also calculated fitted values for unobserved cases—nifty

Extensions:

Tobit is the tip of the iceberg

Beyond are methods for truncated data, and for stochastically censored data

The latter are sample selection models and widely used in econometrics

Rare events data: relogit

(The following is based on King and Zheng (2001, *Political Analysis*)

Often logit is used to study rare events:

- Whether an individual randomly selected from the population has AIDS
- Whether a given dyad of countries are at war

It turns out that logit is biased to the extent $\bar{y} << 0.5$

Assume that 1s indicate the rare event, and 0s the common alternative.

We are interested in unbiased estimation of a parameter β

Rare events data: relogit

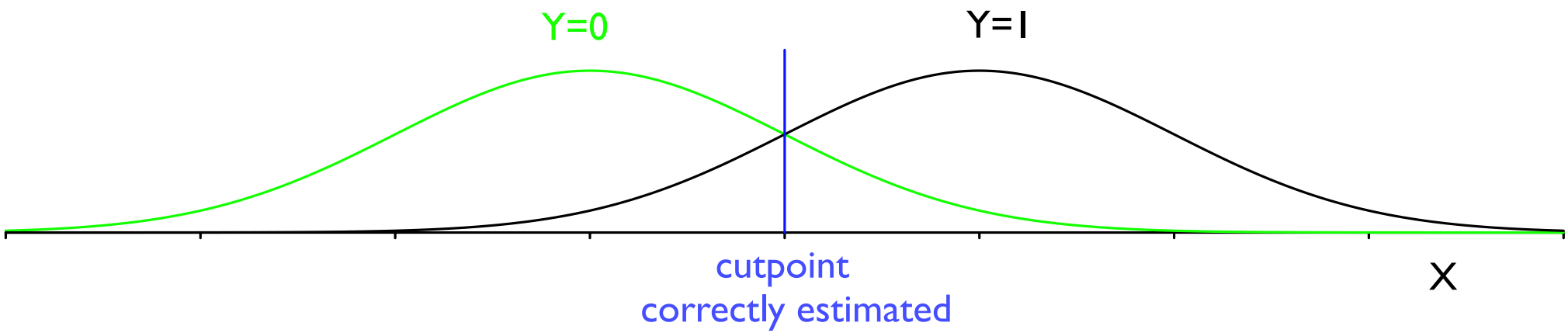
Suppose that x indeed helps classify events as 1s or 0s.

In our war example, x might be whether two countries share a border.

This indeed makes war more likely,
but most neighboring countries are nevertheless at peace most of the time

Rare events data: relogit

Suppose that if we had an infinite amount of data, we would uncover the following conditional distributions



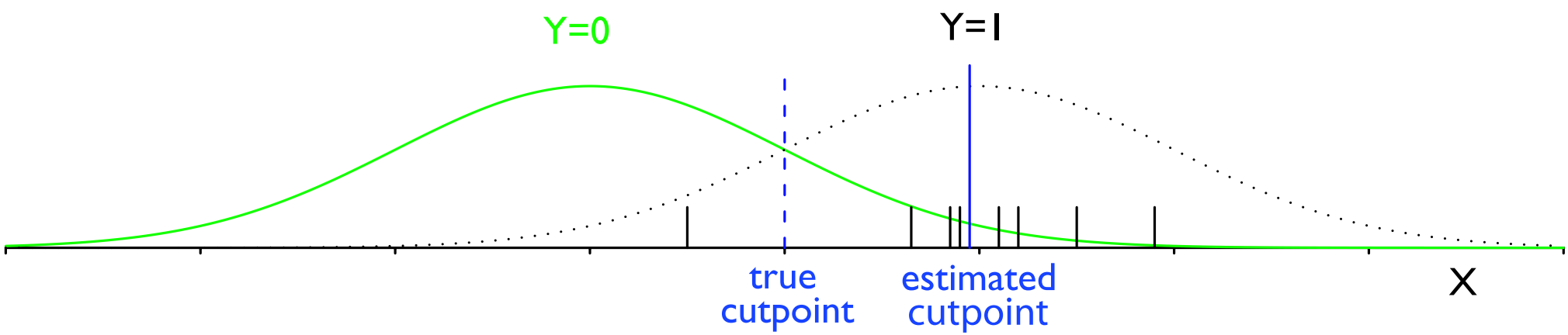
In this case, we would correctly set place the cutpoint between 0s and 1s at the indicated location

This placement of this cutpoint is related to the estimate $\hat{\beta}$, and is made to minimize incorrect classifications.

With an infinite amount of data, the consistency of logit (as an MLE) ensures $\hat{\beta} = \beta$

Rare events data: relogit

Now suppose we have a few thousand observations, of which only a handful are 1s. The distribution of 0s remains clear, but the distribution of 1s is sketchy



Logit will once again choose a cut point to min(incorrect classifications).

But with few 1s, it will place the cutpoint too far to the right, because the true cutpoint would misclassify many 0s.

(Recall ROC curves—the small sample logit is doing well at the lower part of the ROC, but very badly on the top part)

Rare events data: relogit

King and Zheng offer several pieces of advice:

1. Data collection: Collect every 1 you can; randomly sample an equal number of 0s; correct for sampling bias.

2. Estimation: when analyzing data with $\bar{y} \neq 0.5$, apply a bias correction for rare-events.

I leave the details of these bias corrections to interested readers

Implementation:

rare-events logit is part of the `R` package `Zelig`, which also contains functions for simulating quantities of interest eeriely similar to those we've been producing . . .