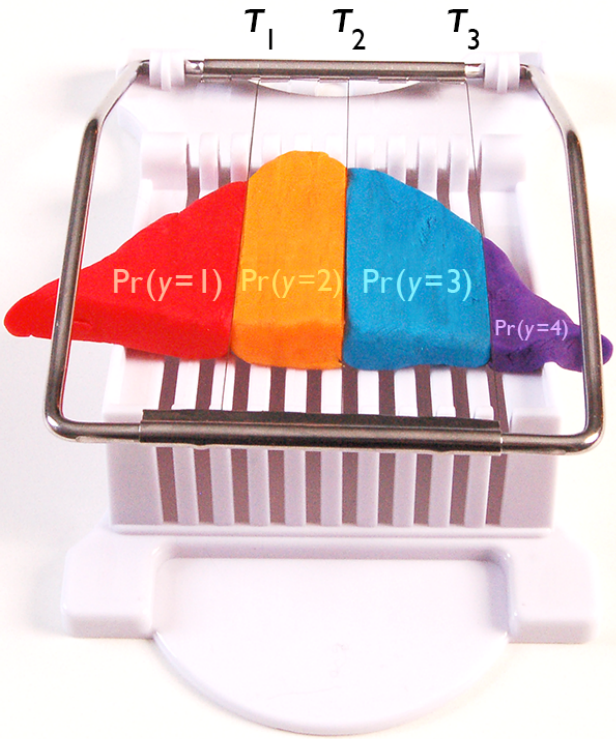
Maximum Likelihood Methods
for the Social Sciences
POLS 510 · CSSS 510

Models of Ordered Data

Christopher Adolph

Political Science *and* CSSS
University of Washington, Seattle



Ordered Discrete Variables

So far, we've discussed binary variables: $y \in \{0, 1\}$.

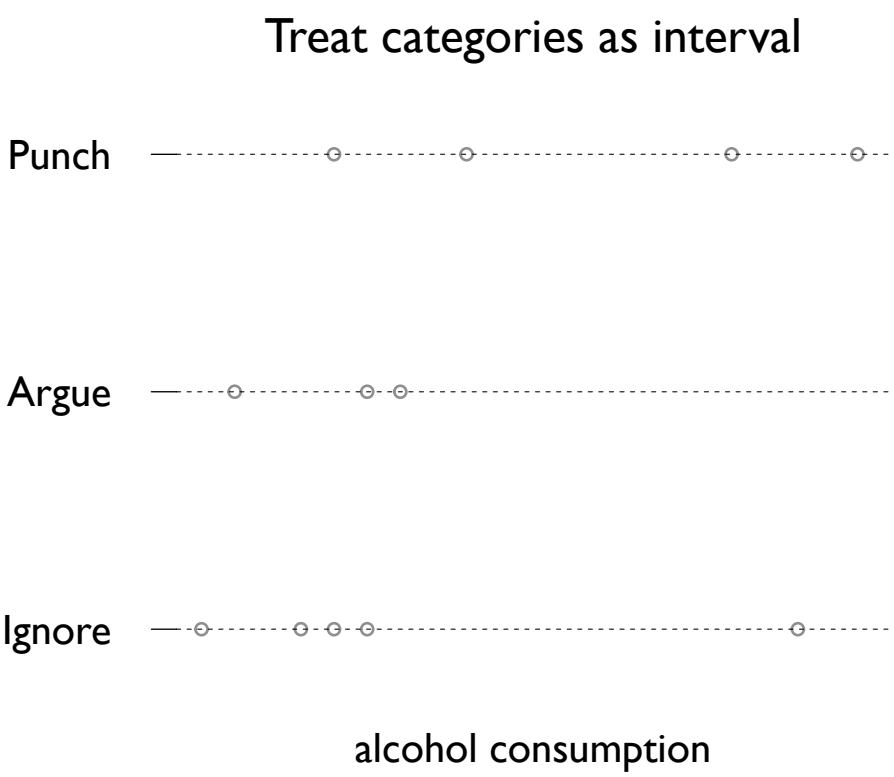
Another way to think of a binary variable is as a set of ordered values, where the size of the "gap" between values is unknown: $y \in \{\text{lowest}, \text{highest}\}$

A straightforward generalization adds more levels, again with unknown intervals: $y \in \{\text{lowest}, \text{second lowest}, \ldots \text{second highest}, \text{highest}\}$

For convenience, we will say $y \in \{1, 2, \ldots, m-1, m\}$ but understand the gaps between these variables may be of different (but unknown) sizes.

Ordered discrete data are *very* common:

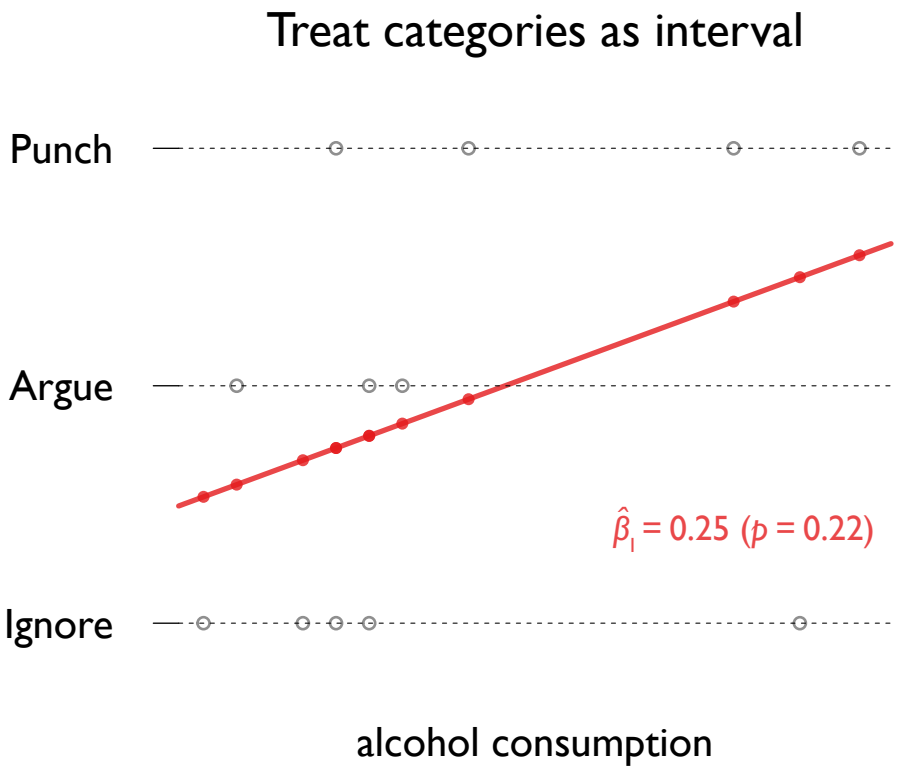
Likert scales	(Disagree strongly, disagree, somewhat disagree, . . .)
Policy options	(Privatize social security, partially privatize, leave unchanged)
Survey questions	(Some high school, HS grad, some college, college grad, . . .)
Ranks in hierarchies	(Staff, manager, upper management, executive)
Health	(Healthy, Sick, Dead)
and so on	. . .



Thought experiment: We observe a variable measuring the violence of a subject's responses to provocations made by a stranger

The variable is coded in three ordered categories:

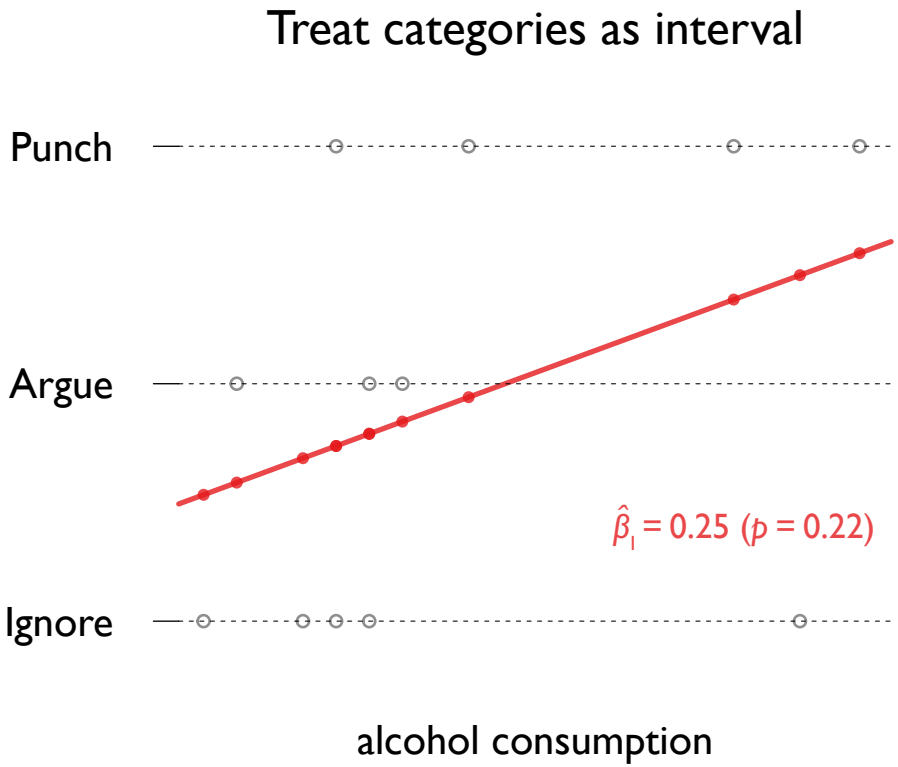
1. did the subject *ignore* the provocation?
2. did the subject verbally *argue* with the stranger?
3. did the subject physically *punch* the stranger?

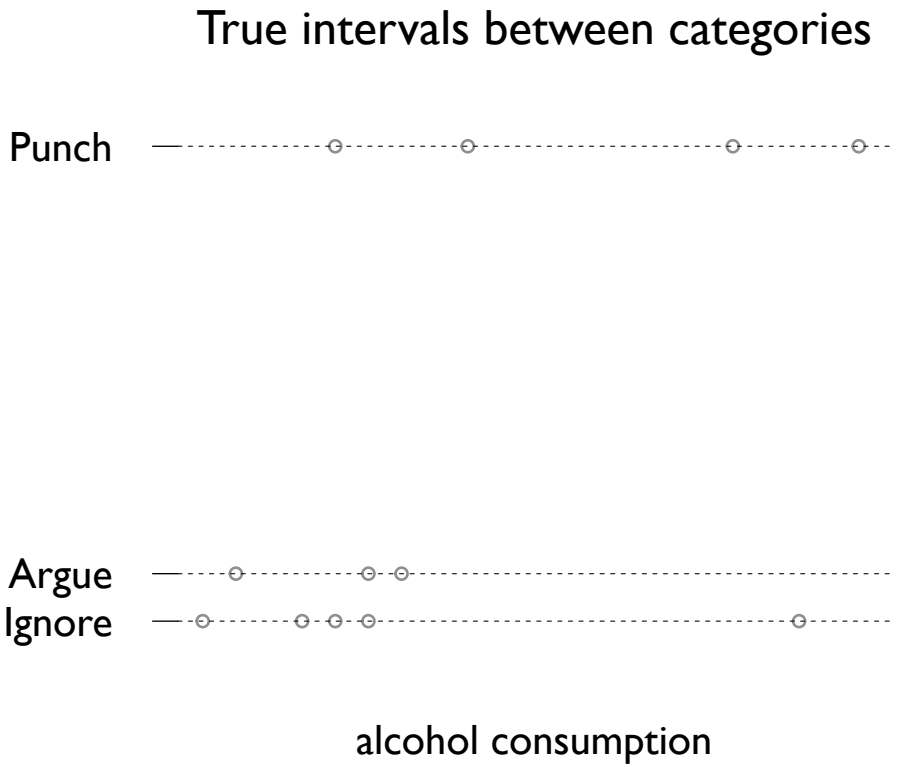


We're interested in the relationship between violence of response
and prior alcohol consumption (perhaps the data come from altercations in bars)

The plot shows what happens when
we treat the ordered categories of *violence* as an interval measure

Do we trust this model's estimate of the slope or predictions of $\hat{y}$*?*



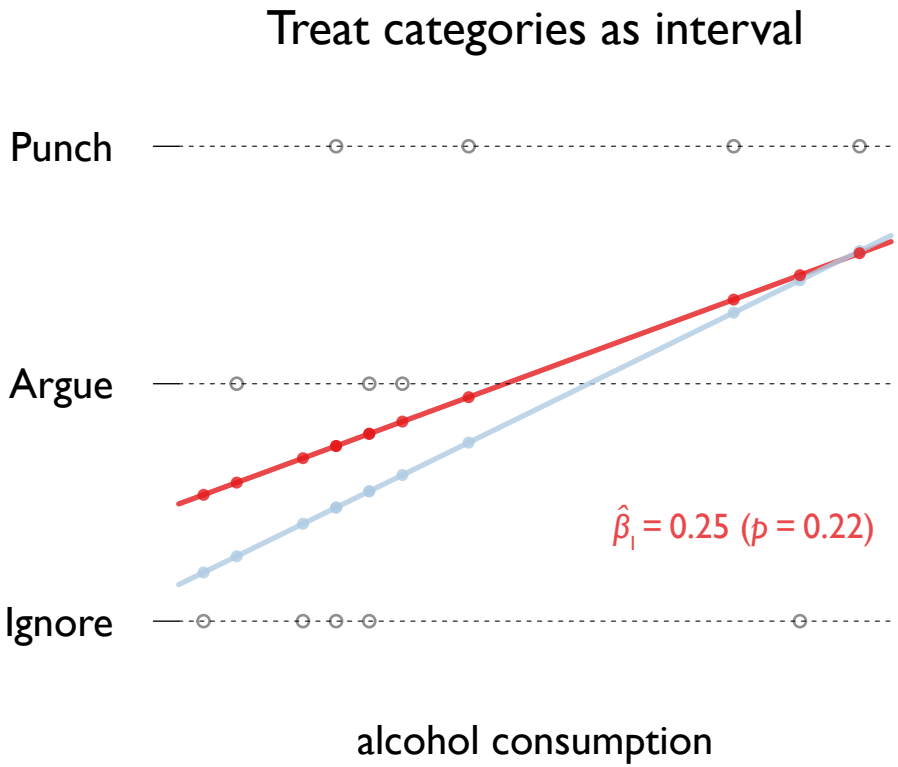


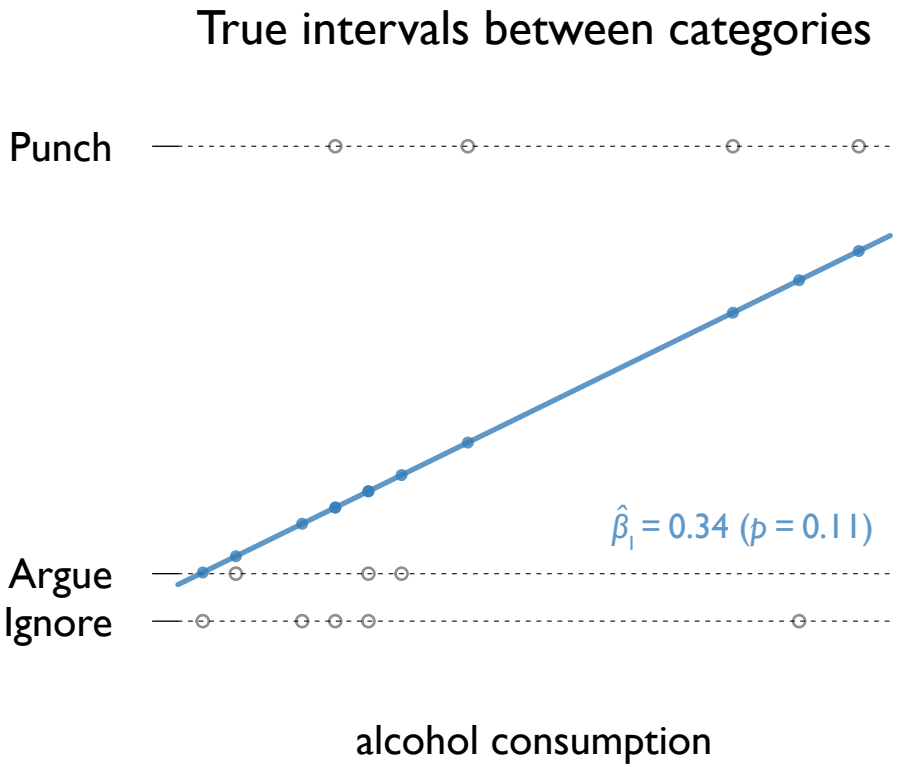
The model at left rests on very shaky assumptions, especially if the number of choices is small

Violence $\in \{$*ignore*, *argue*, *punch*$\}$ is not equally spaced on most interpretations

Most would say *argue* is close to *ignore* than it is to *punch*

Applying a model that treats them as such makes a strong & incorrect assumption



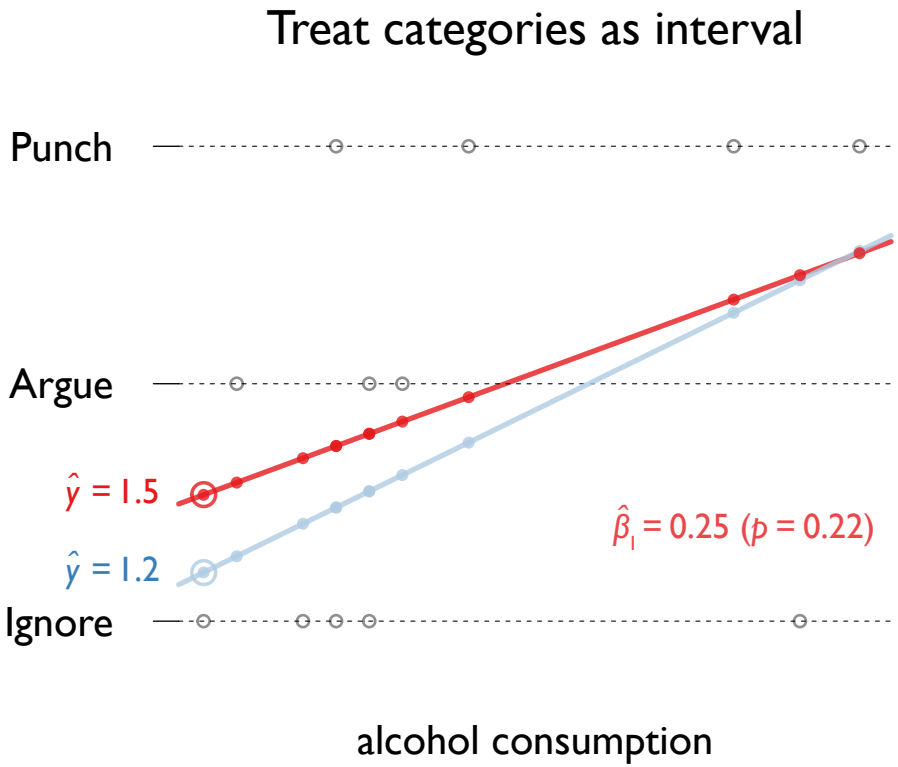


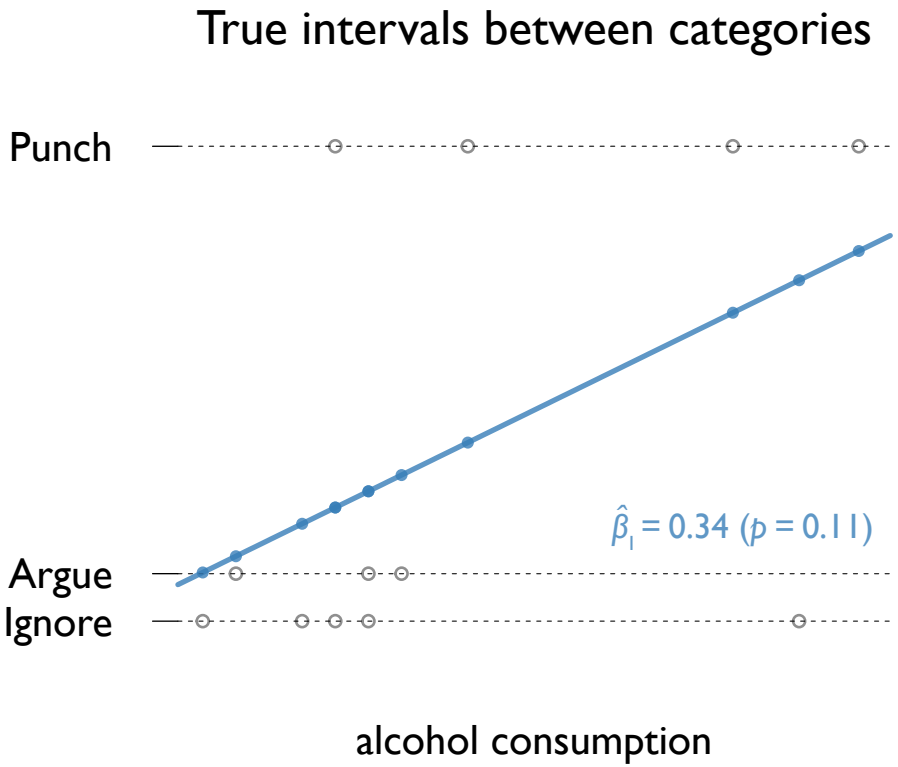
Suppose the intervals at the right are "correct"

Linear regression in this case recovers a much larger $\hat{\beta}$ –
nearly 50% larger than under the incorrect assumption of equal spacing

Linear regression is *biased* with three or more categories (it wasn't with just two)

In small samples, this bias could lead to Type II errors –
failure to reject an incorrect null of no relationship

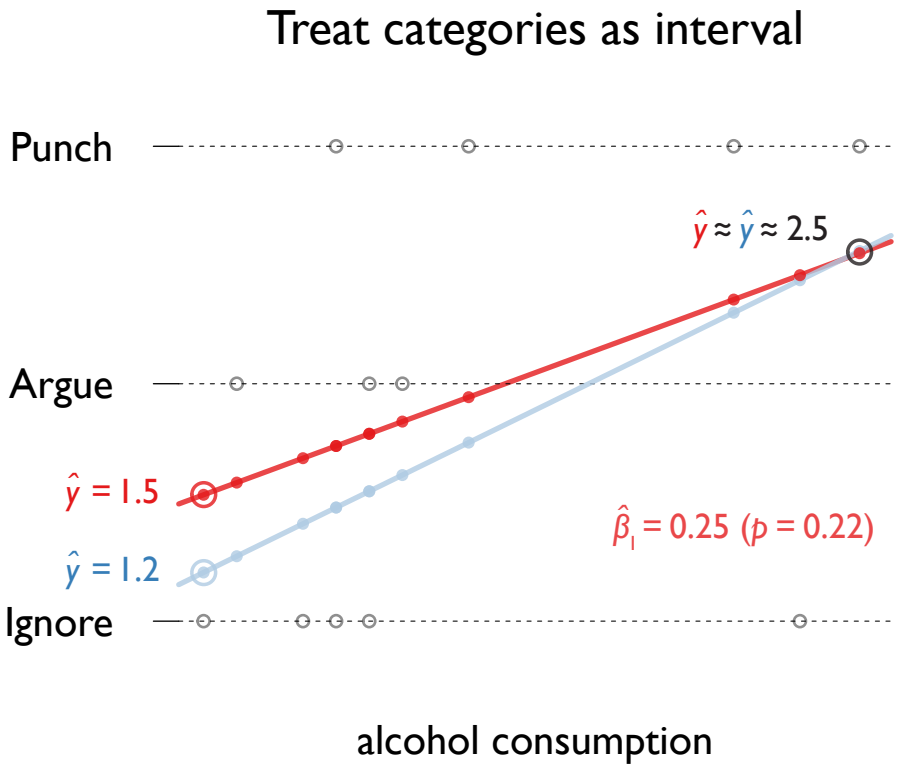


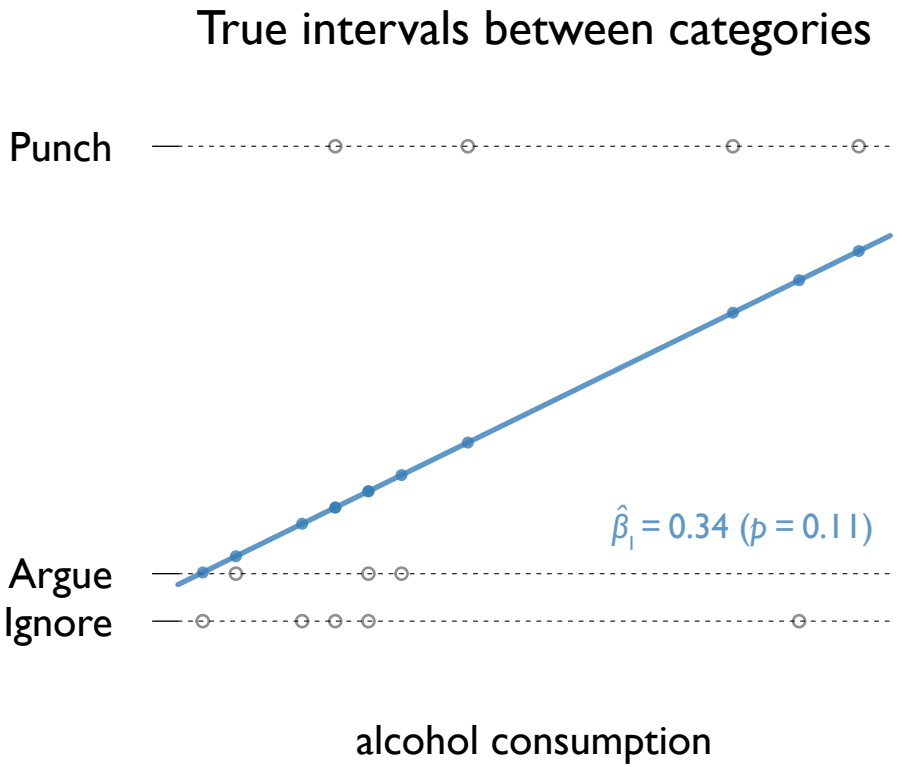


Assuming equal spacing also leads to large errors when predicting $\hat{y}$, the level of violence

Here we overestimate the risk of violence for low levels of alcohol

What does it mean to say $\hat{y} = 1.5$ *or* $\hat{y} = 1.2$*?*

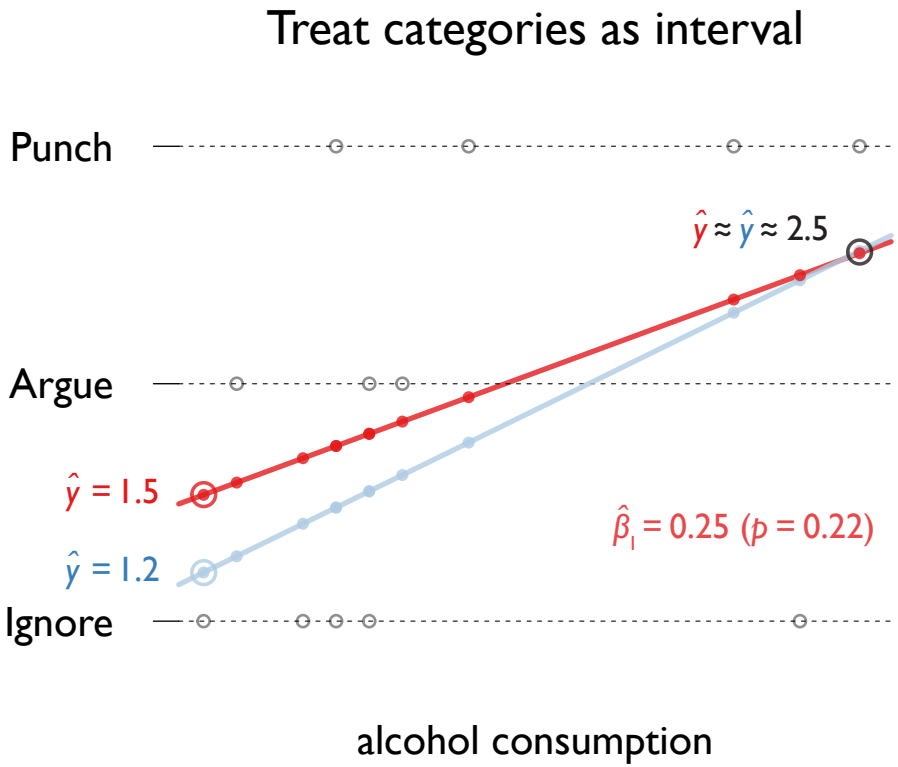


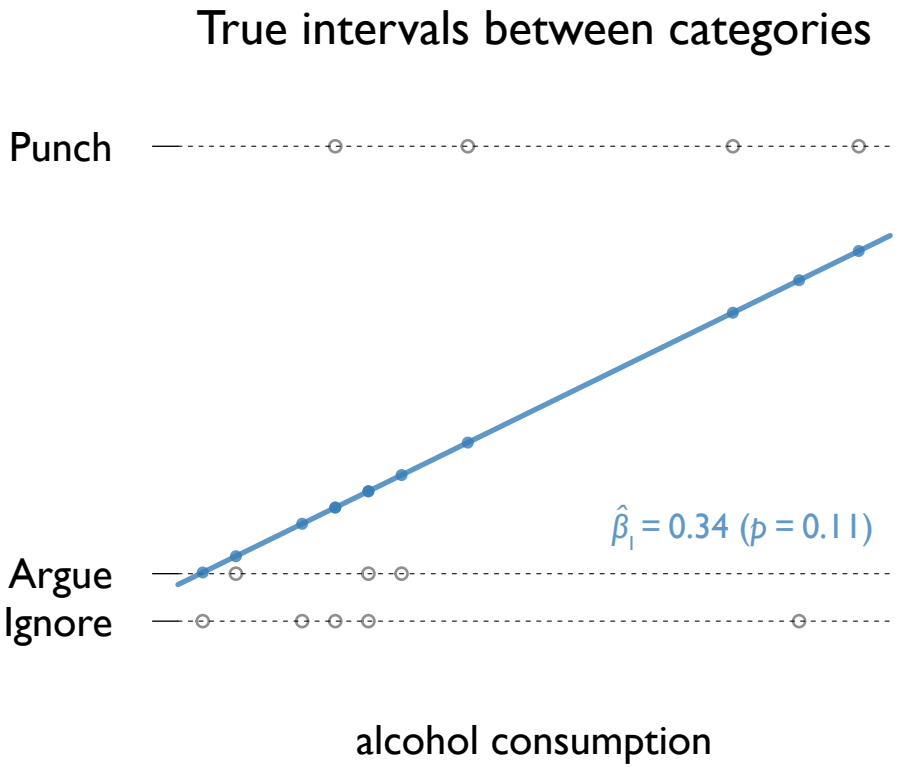


The models make similar predictions of violence for high alcohol consumption

Not much help: without knowing the true intervals,
we can't know where (if anywhere) $\hat{y}$ is approximately unbiased

But what does $\hat{y} = 2.5$ even mean? Does it mean anything at all?



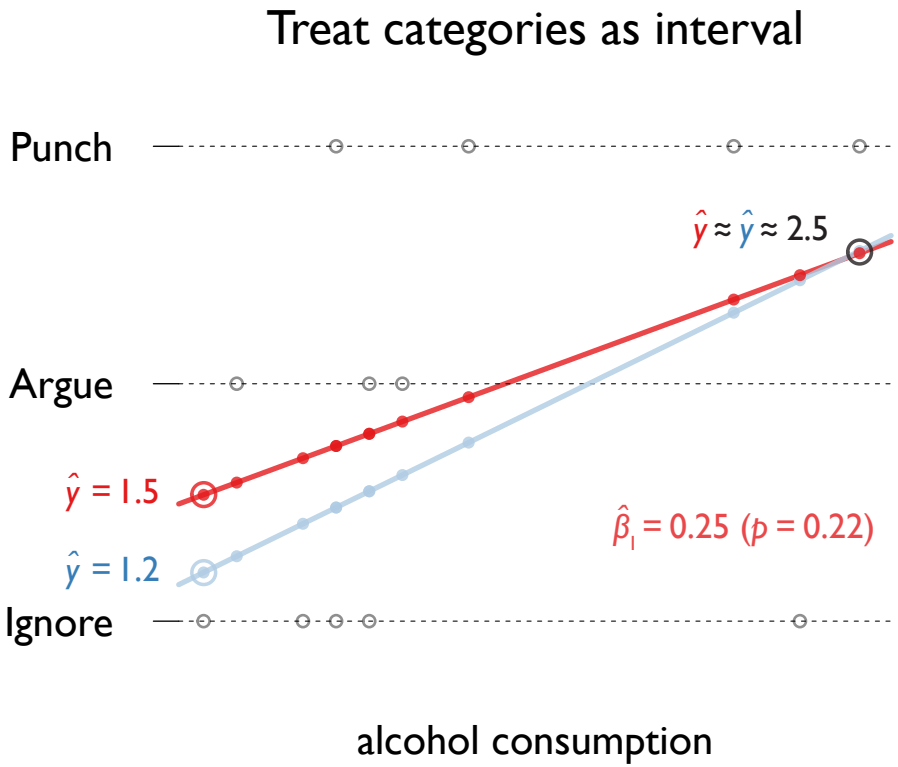


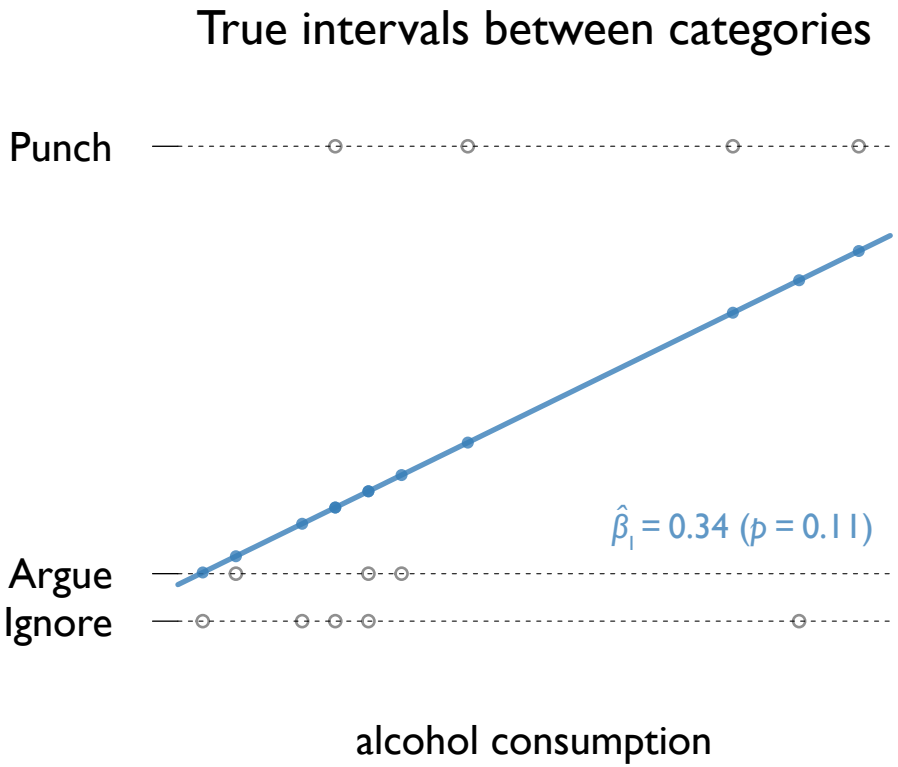
Problems with linear regression on ordered categories:

(1) Same problems as linear regression with a binary outcome:
inefficient & predictions can fall outside the range of y

(2) Treating 3+ categories as interval data creates measurement error:
potentially large bias in $\hat{\beta}$, $\hat{y}$, and p-values

(3) Data generating process completely wrong: $\hat{\beta}$ shifts units of what, exactly?
$\Rightarrow \hat{y}$ cannot be interpreted at all (certainly not as a probability)

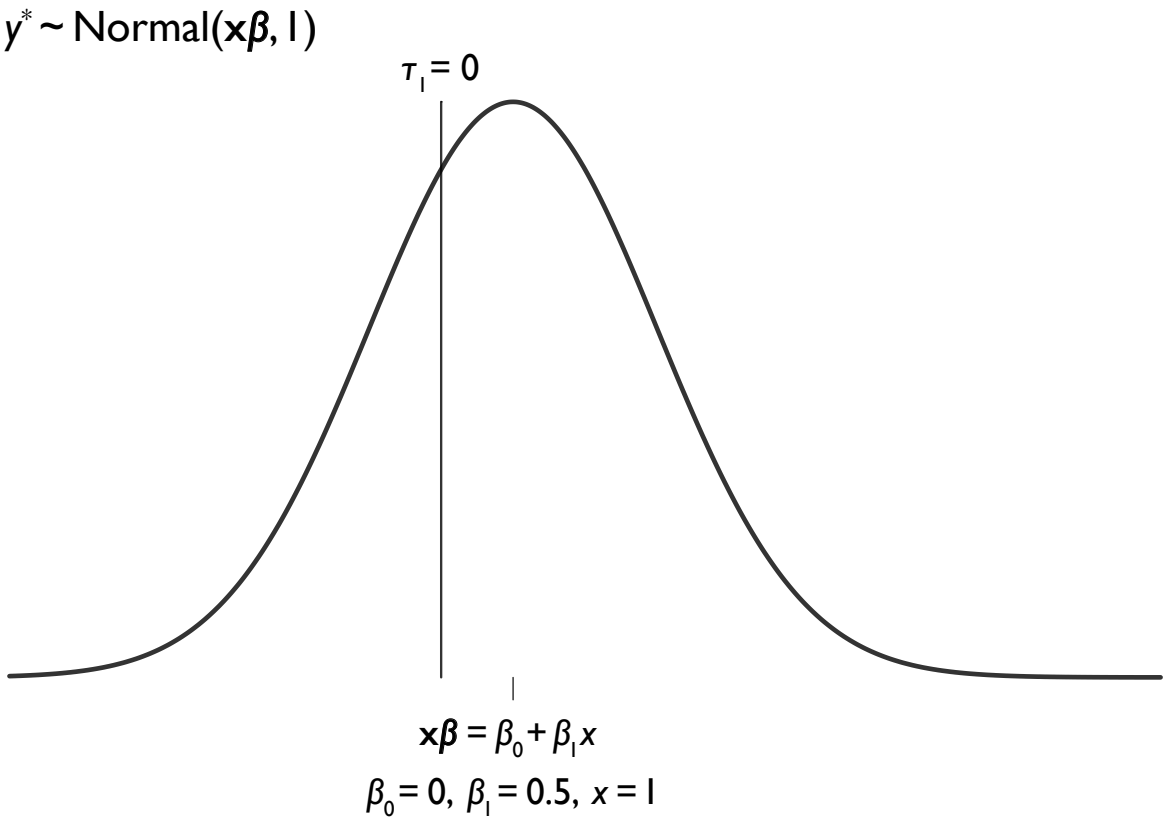


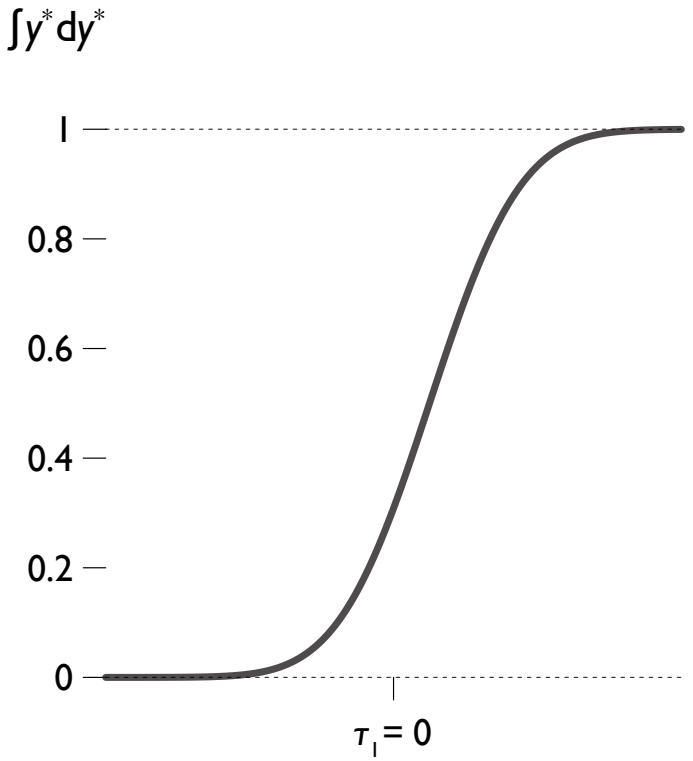


Saying a set of discrete, ordered choices are not "equally spaced" implicitly appeals to their values on a continuous latent variable

Let's make explicit the relationship between the ordered discrete variable y and the latent continuous variable y^*

Start with a review of latent variables in the binary context



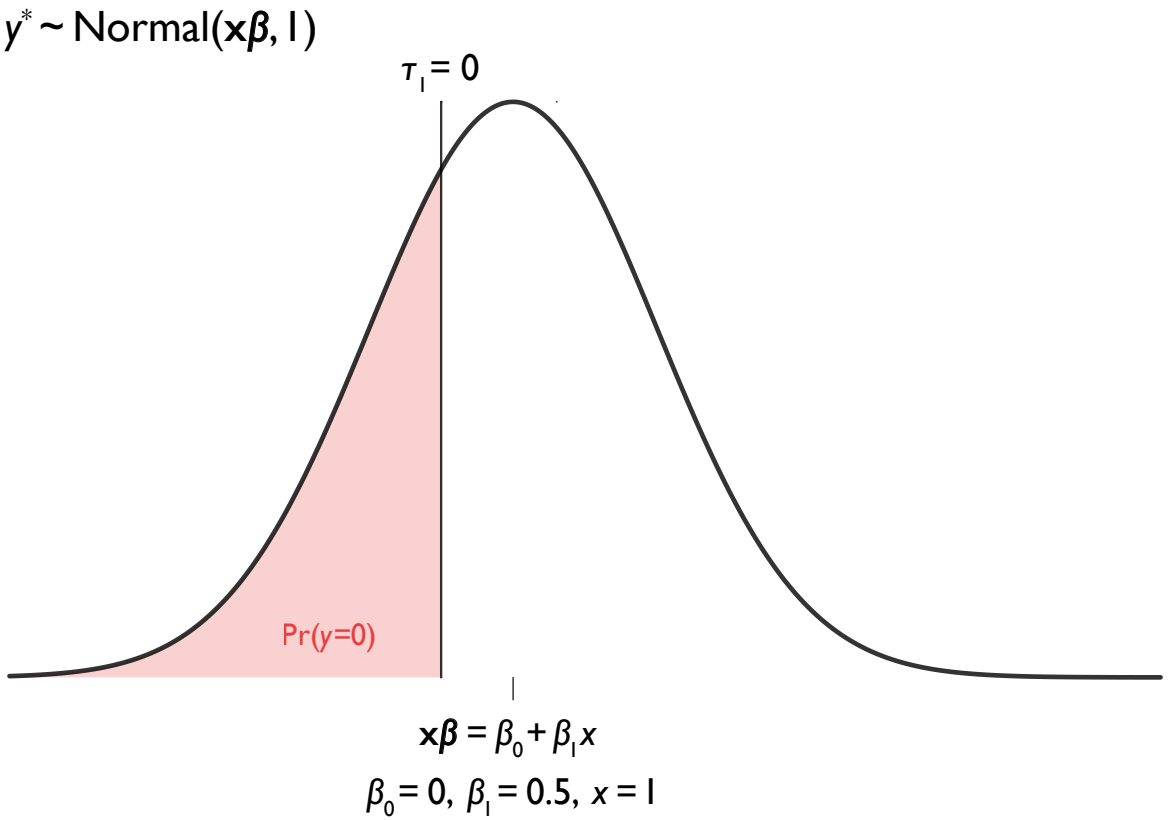


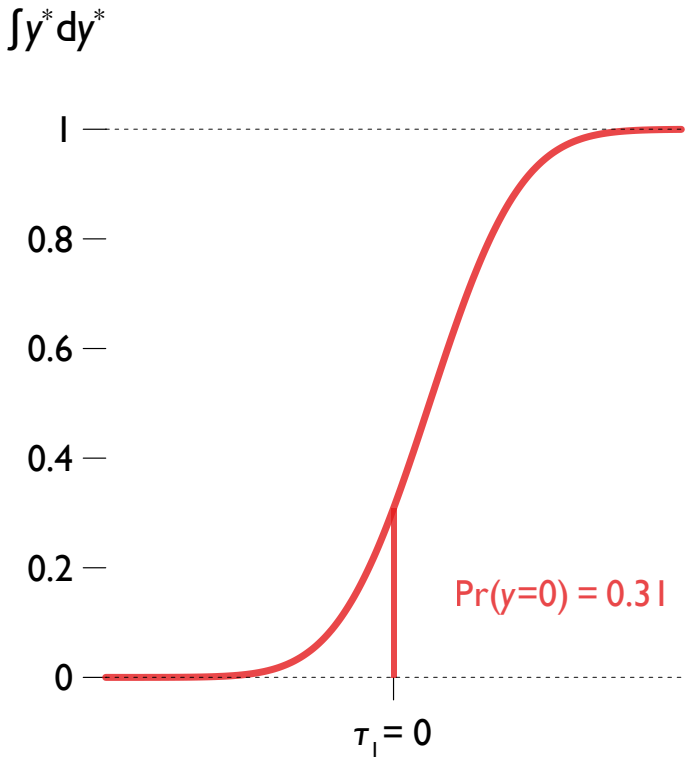
Recall: a model of binary y_i can be motivated by a latent variable y_i^* such that

$$y_i = \begin{cases} 0 & \text{if } -\infty < y_i^* \leq \tau_1 \\ 1 & \text{if } \tau_1 < y_i^* \leq \infty \end{cases}$$

b/c y^* is unobserved, we assumed $\tau_1 = 0$ for identification & w.o. loss of generality

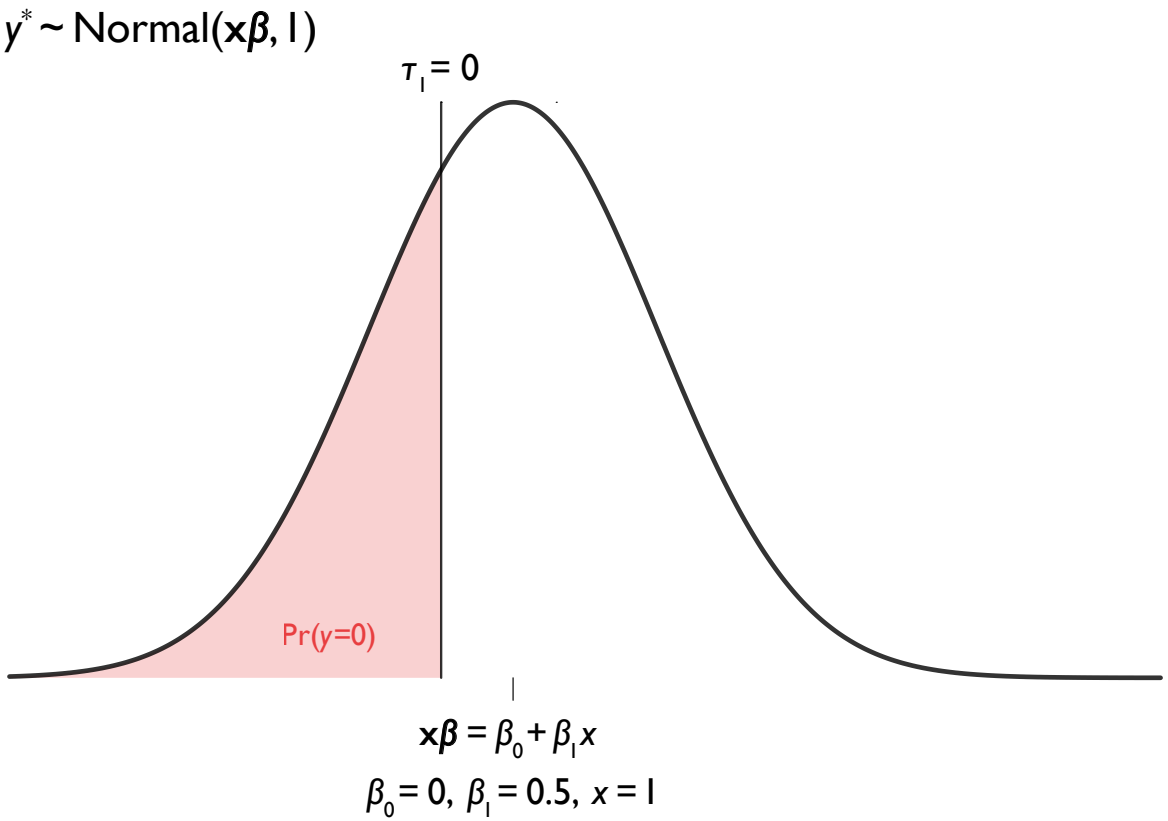
We noted that if $y_i^* \sim \text{Normal}(\mathbf{x}_i\boldsymbol{\beta}, \sigma_2 = 1)$, this produces the probit model for y_i

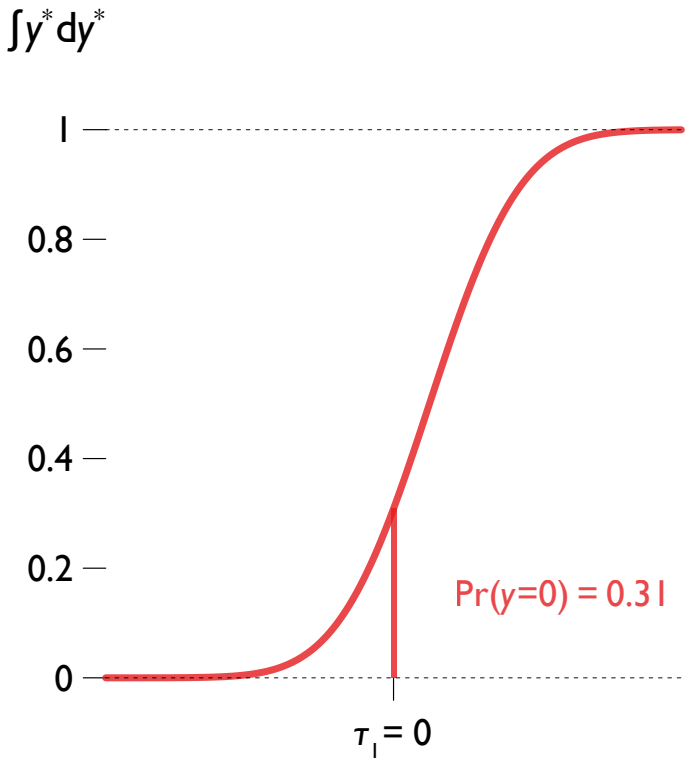




In linear regression an expected value is a *real number* $\hat{y}$ on the scale of y

If we could observe y_i^*, then its expected value would simply be $\mathbf{x}_i\hat{\boldsymbol{\beta}}$, as shown on the pdf at left

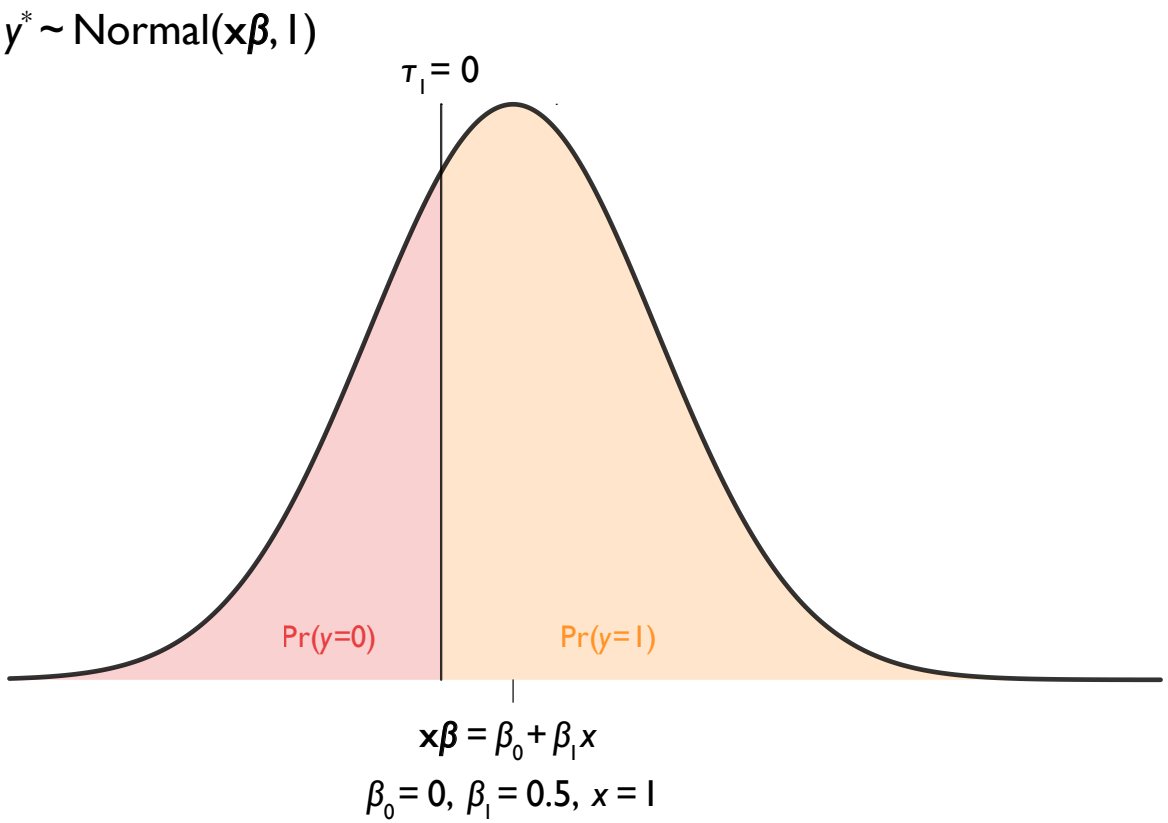


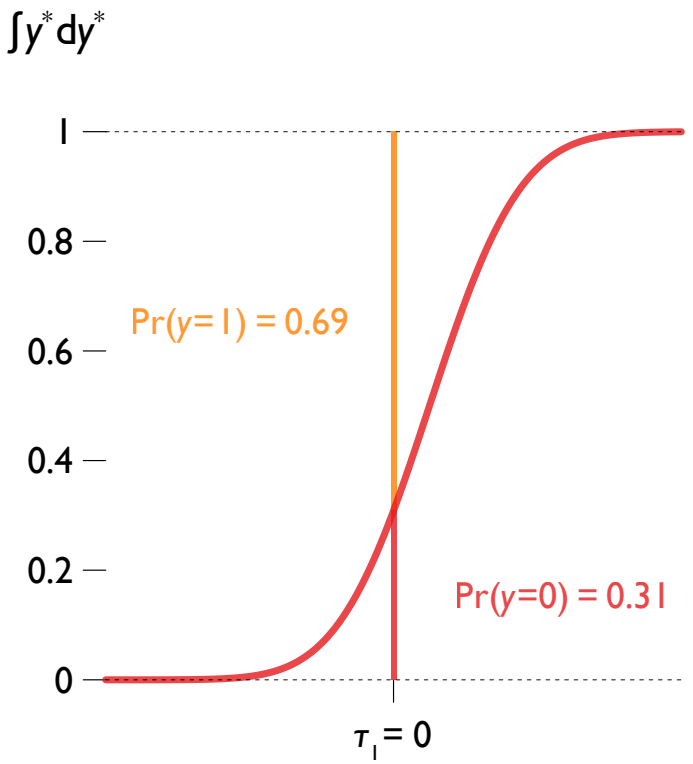


In models of binary outcomes an expected value is a *probability* $\hat{\pi}$, computed from a cdf – here, the Normal cdf

Really, it's two probabilities, one for each category of y_i:

$$\Pr(y_i = 0) = \int_{-\infty}^{\tau_1 = 0} \text{Normal}(\mathbf{x}_i \hat{\boldsymbol{\beta}}, 1) \mathrm{d}y_i^*$$

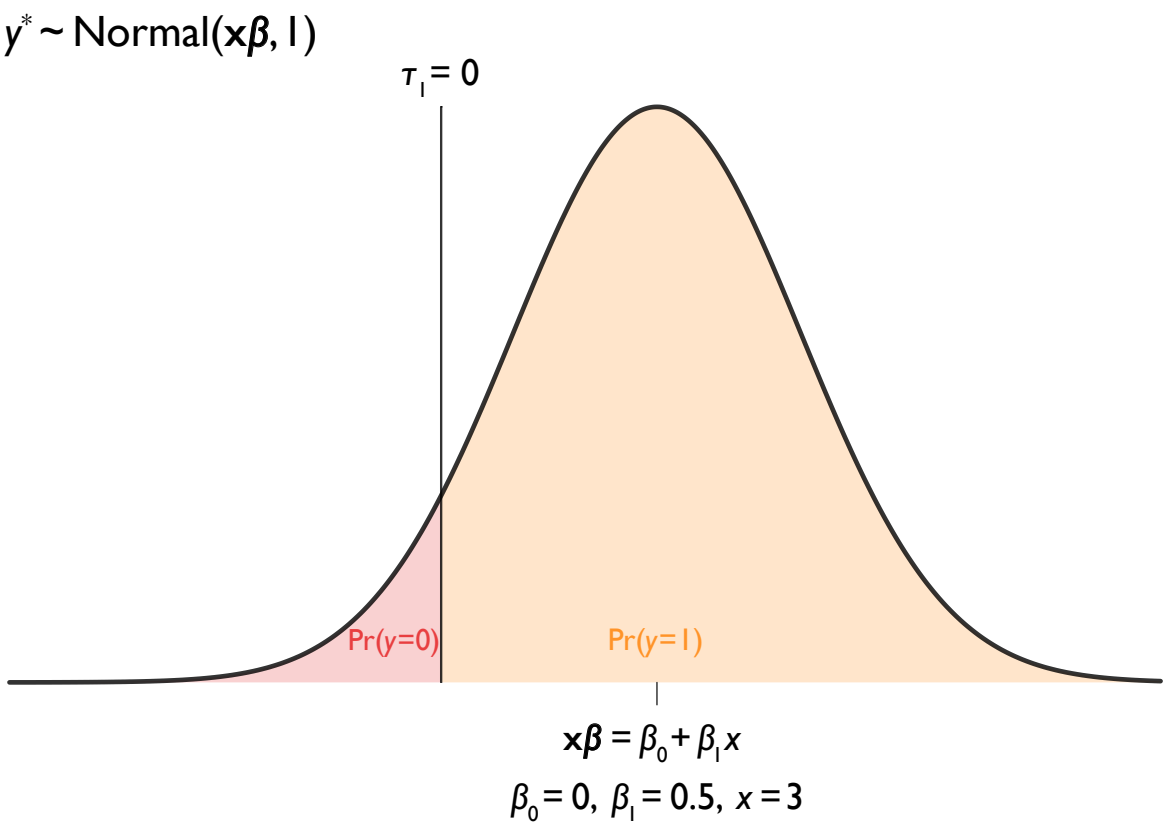


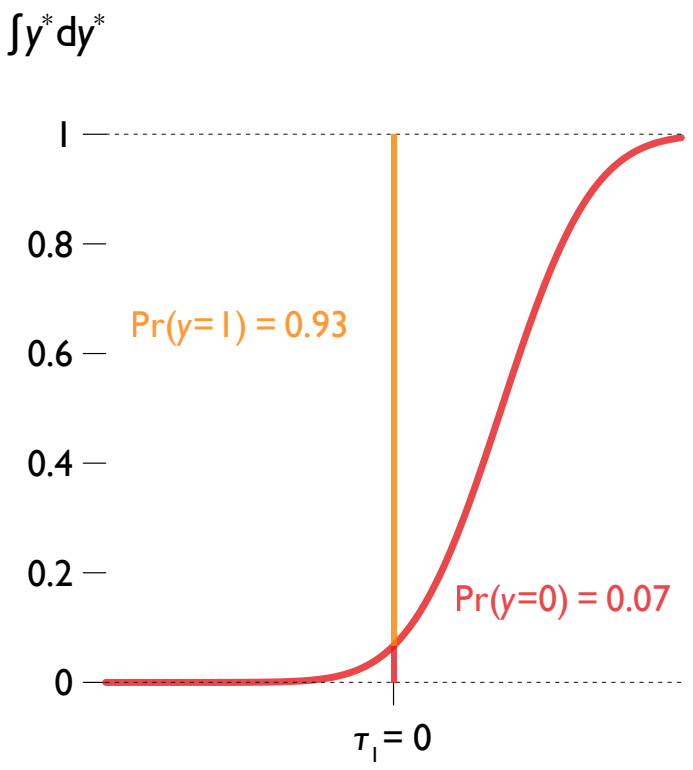


In models of binary outcomes an expected value is a *probability* $\hat{\pi}$, computed from a cdf – here, the Normal cdf

Really, it's two probabilities, one for each category of y_i:

$$\Pr(y_i = 0) = \int_{-\infty}^{\tau_1 = 0} \text{Normal}(\mathbf{x}_i \hat{\boldsymbol{\beta}}, 1) \mathrm{d}y_i^* \qquad \Pr(y_i = 1) = 1 - \Pr(y_i = 0)$$



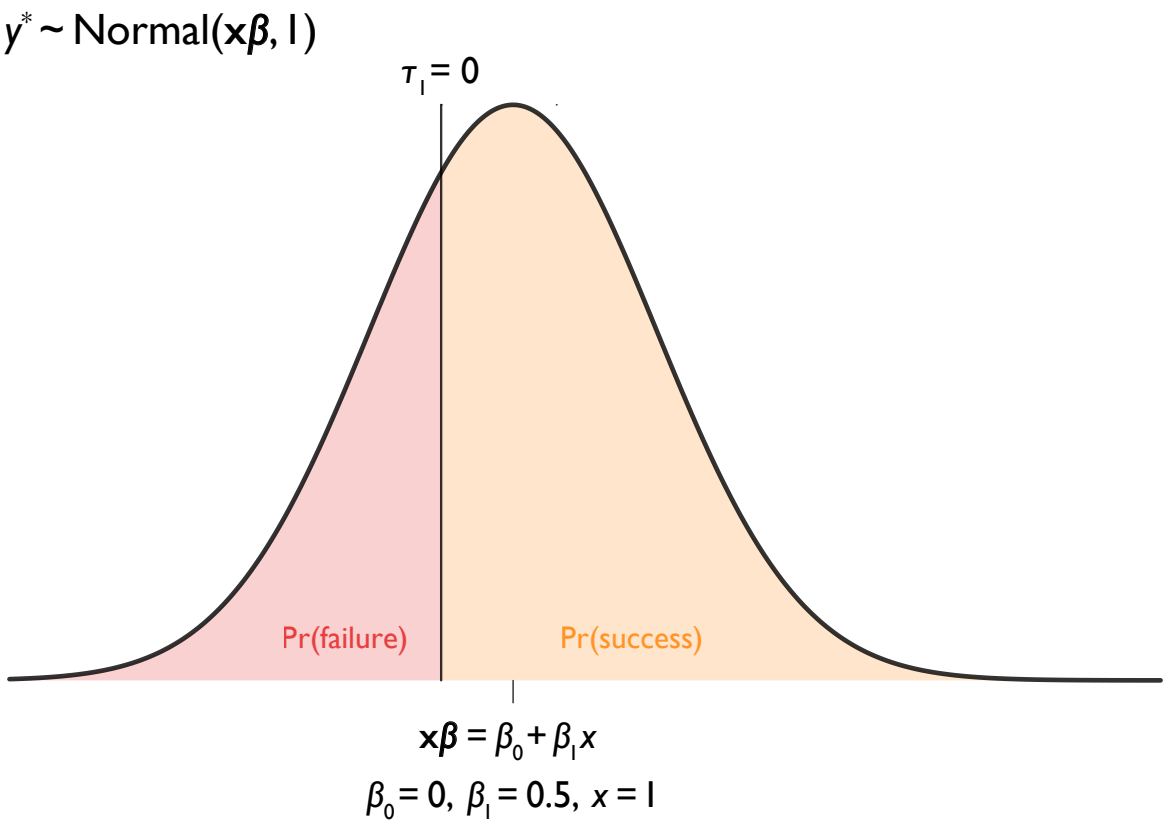


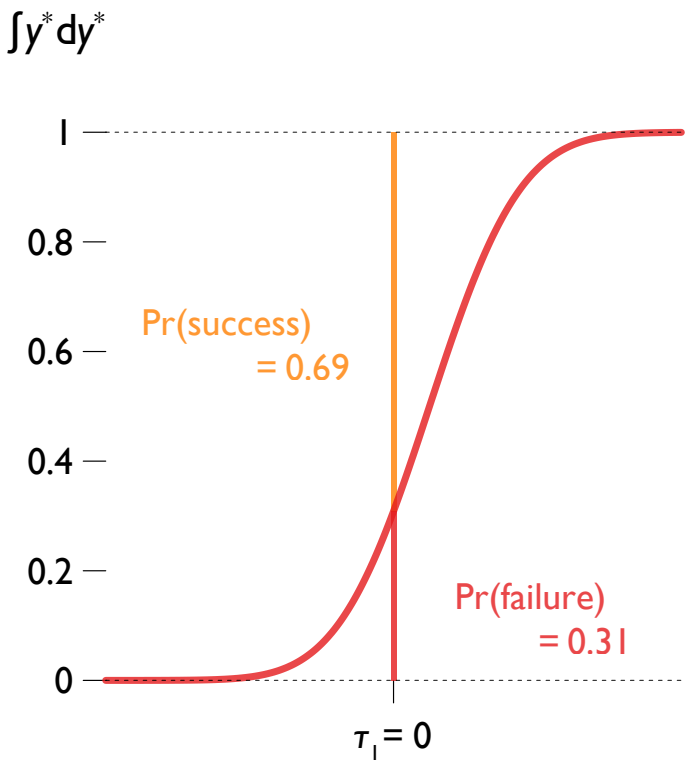
We learned these probabilities shift as $\mathbf{x}_i\boldsymbol{\beta}$ does

In our running example, we have $\mathbf{x}_i\boldsymbol{\beta} \quad = \beta_0 + \beta_1 x_i \quad = 0 + 0.5 x_i$

An increase in x_1 from 1 to 3 moves the mean of the latent variable from 0.5 to 1.5

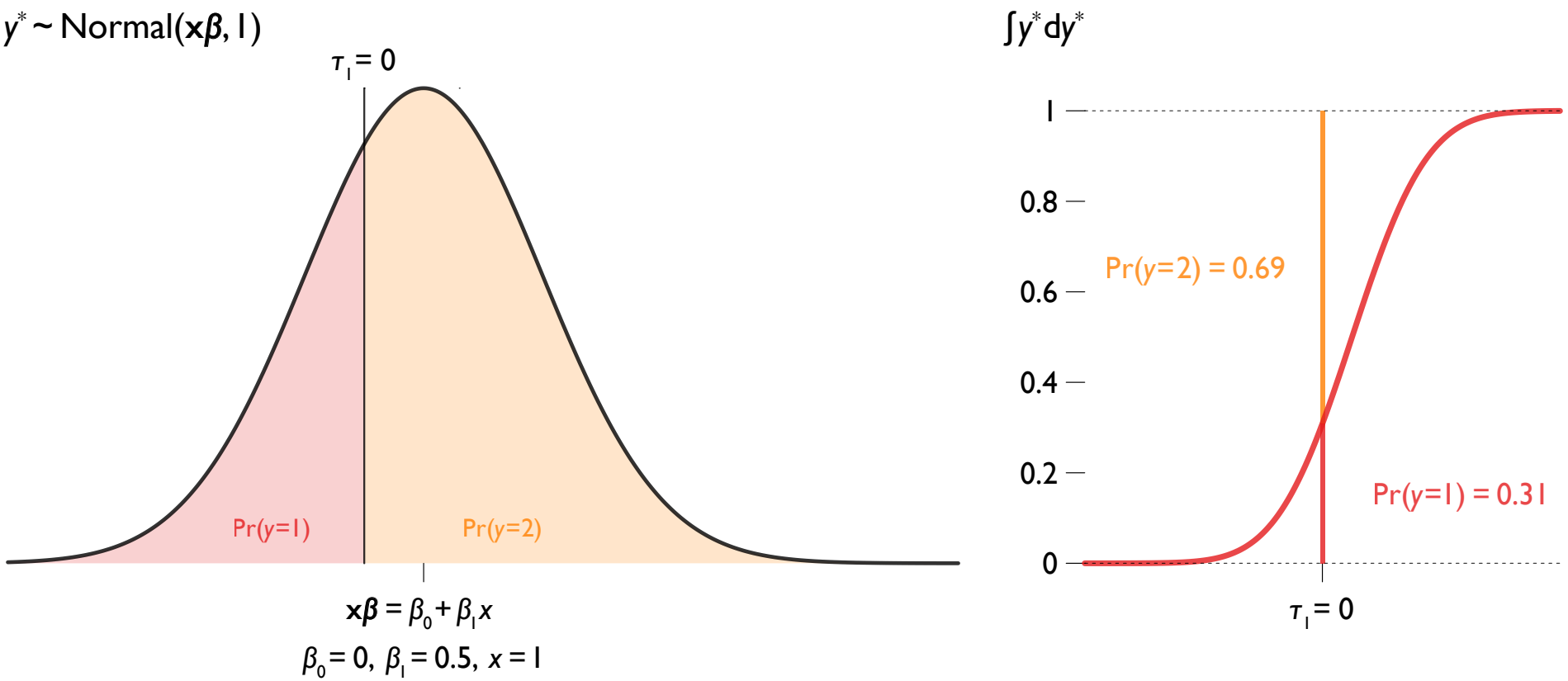
Shifting this much mass of the latent variable across the cutpoint $\tau_1 = 0$ raises the probability of category 1 from 69% to 93%, as shown by the cdf





We also learned that for binary outcomes the names of the outcomes are immaterial

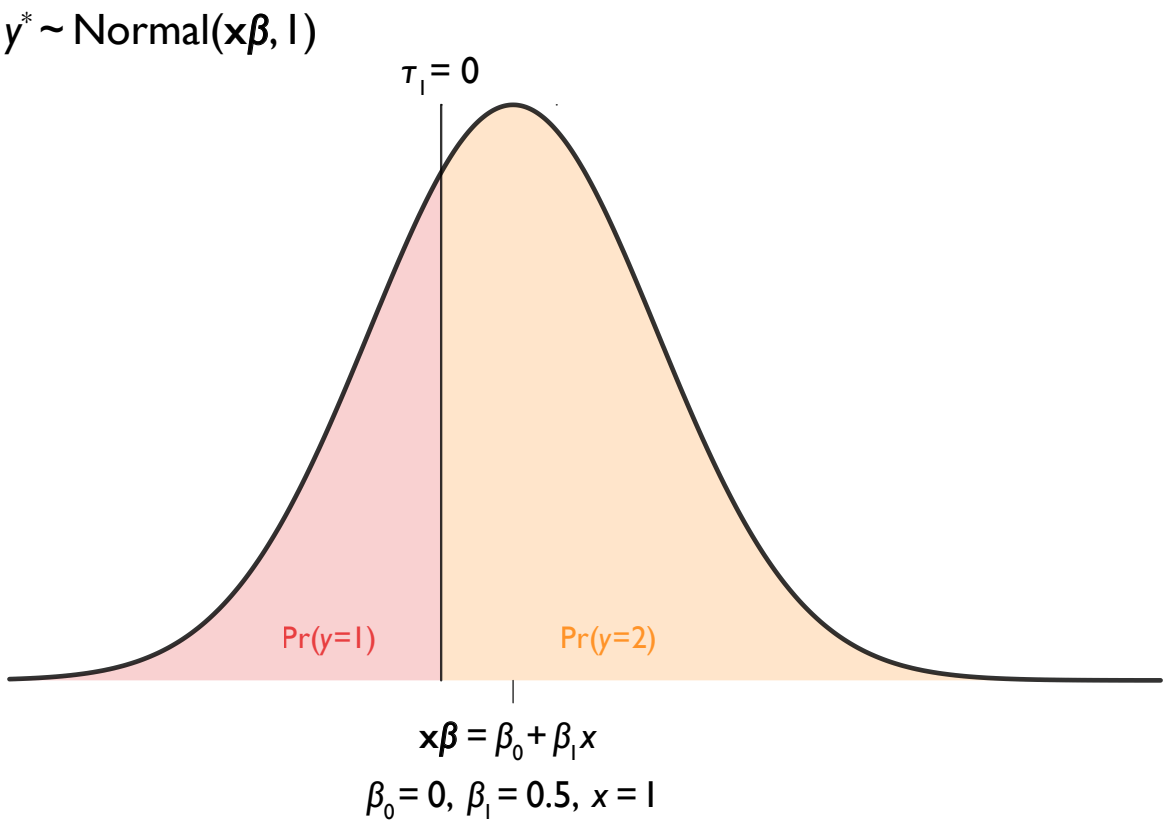
Instead of 0 and 1, we could call them "failure" and "success"

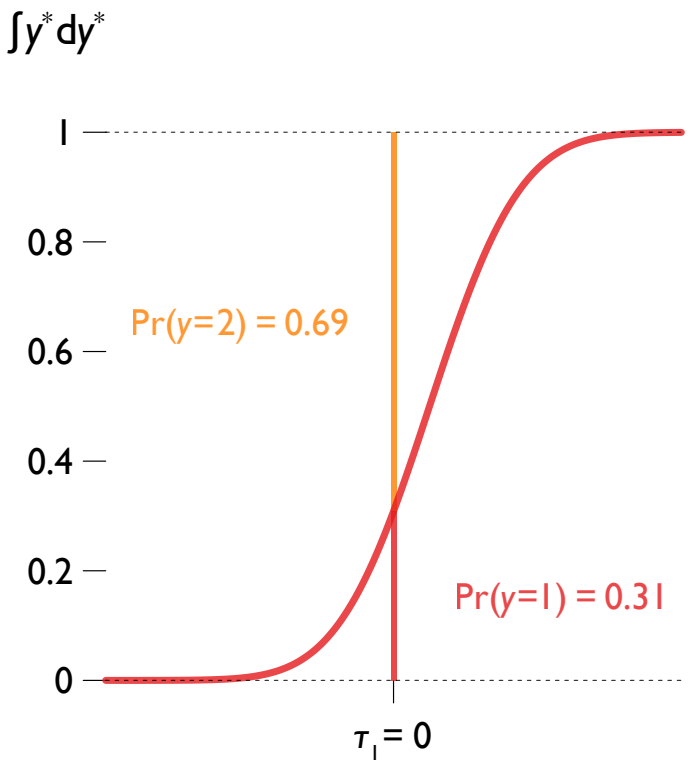


Or simply "1" and "2" for category 1 and category 2

Note that here "1" and "2" are just names –
we don't propose to treat them as numbers that can be added or multiplied

They are simply ordered categories

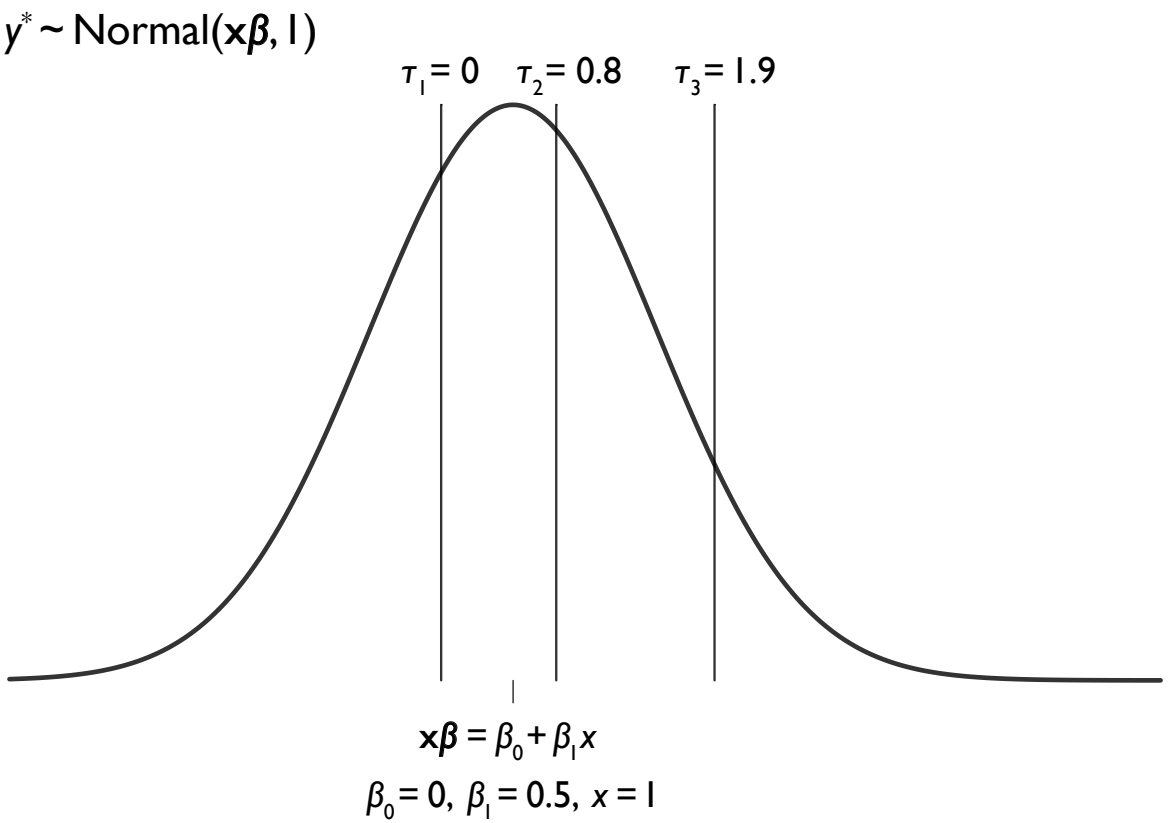


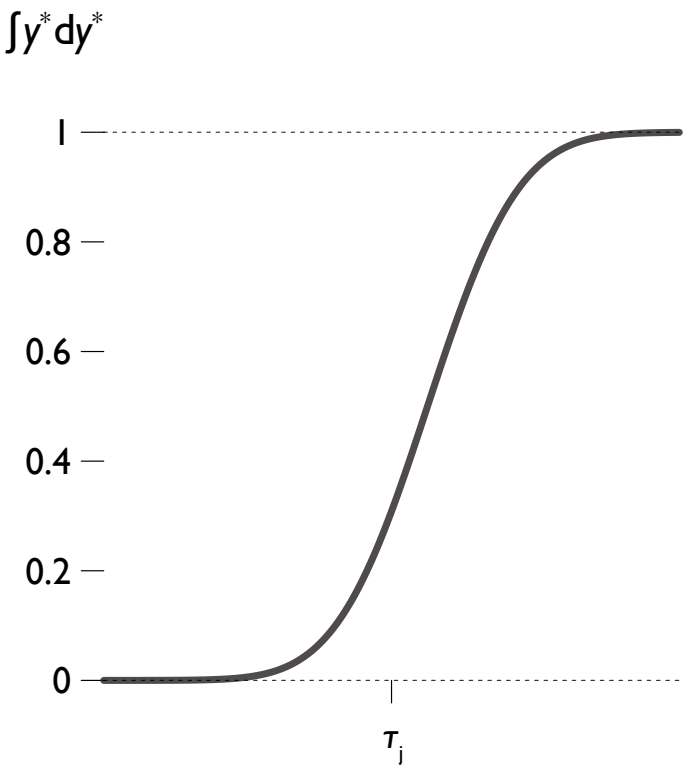


Rewrite the relationship between the binary variable $y_i \in \{1, 2\}$ and the latent y_i^*:

$$y_i = \begin{cases} 1 & \text{if } \tau_0 < y_i^* \leq \tau_1 \\ 2 & \text{if } \tau_1 < y_i^* \leq \tau_2 \end{cases}$$

$\tau_0 = -\infty$ and $\tau_2 = \infty$, so there are $m + 1 = 3$ $\boldsymbol{\tau}$'s
for $m = 2$ categories of y

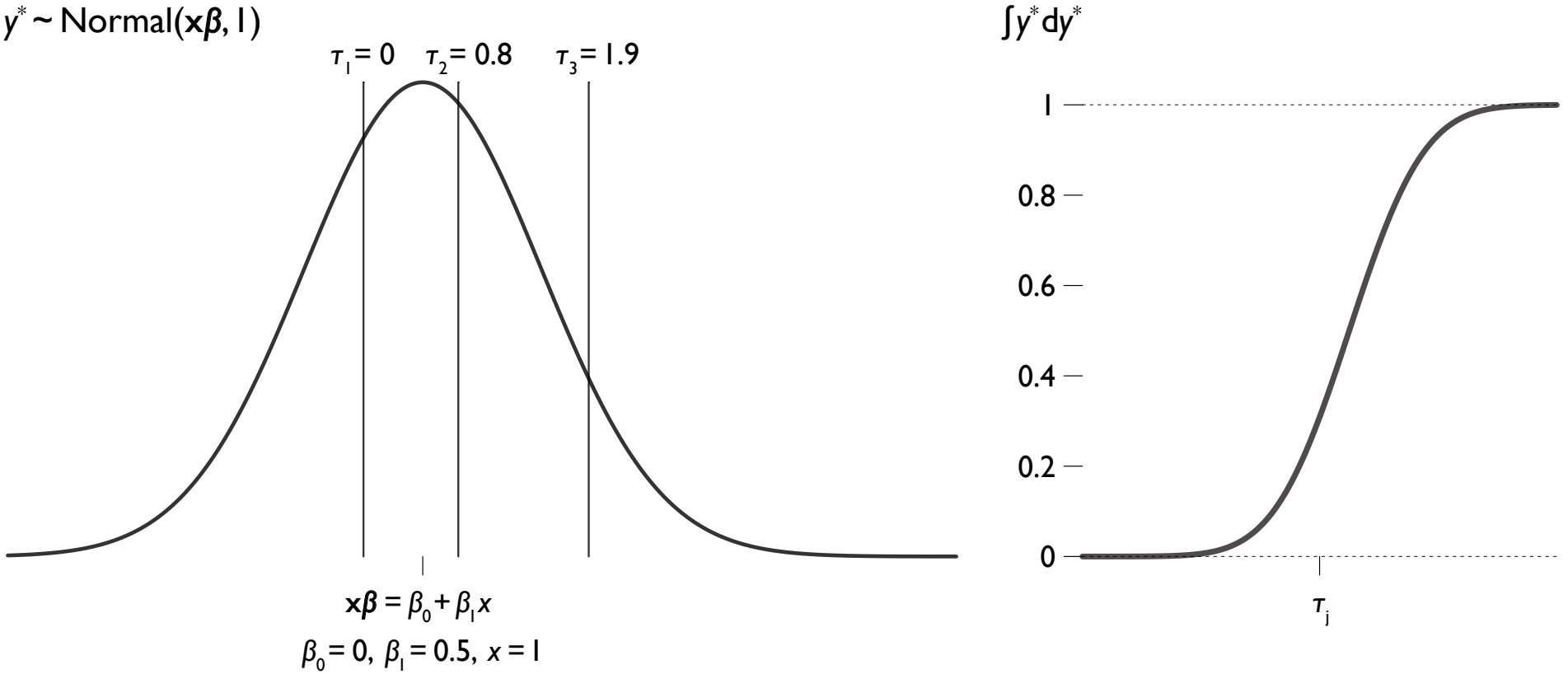




How might we add categories to this model? What if we added more *cutpoints*?

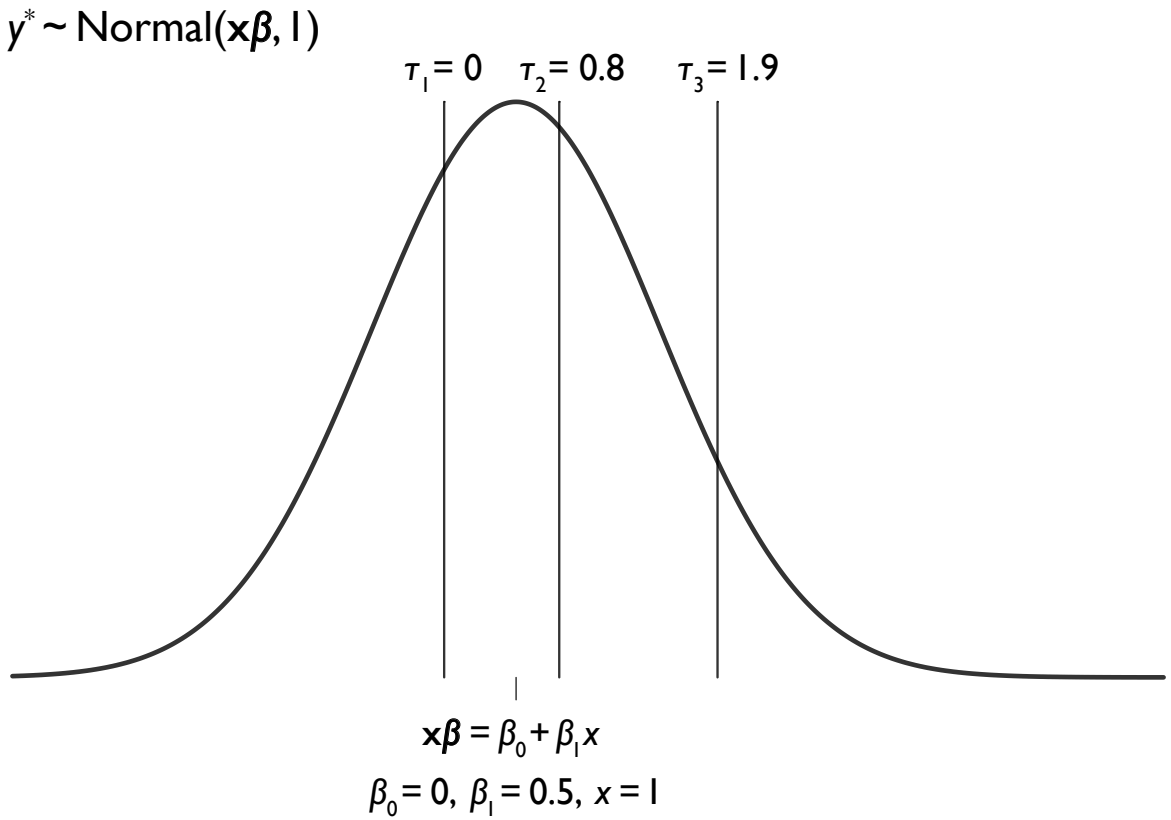
Suppose the outcome could be any of four ordered categories, $y_i \in \{1, 2, 3, 4\}$

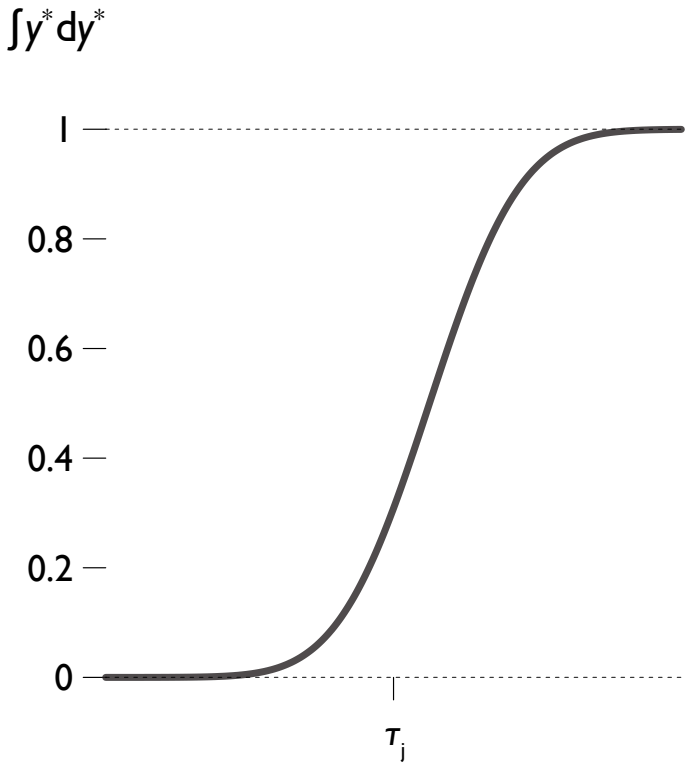
Let's add two more cutpoints τ_2 and τ_3 to mark off the new categories



Now we generalize our latent variable setup to accommodate m categories:

$$y_i = \begin{cases} 1 & \text{if } \tau_0 < y_i^* \leq \tau_1 \\ 2 & \text{if } \tau_1 < y_i^* \leq \tau_2 \\ 3 & \text{if } \tau_2 < y_i^* \leq \tau_3 \\ \dots & \\ m & \text{if } \tau_{m-1} < y_i^* \leq \tau_m \end{cases}$$



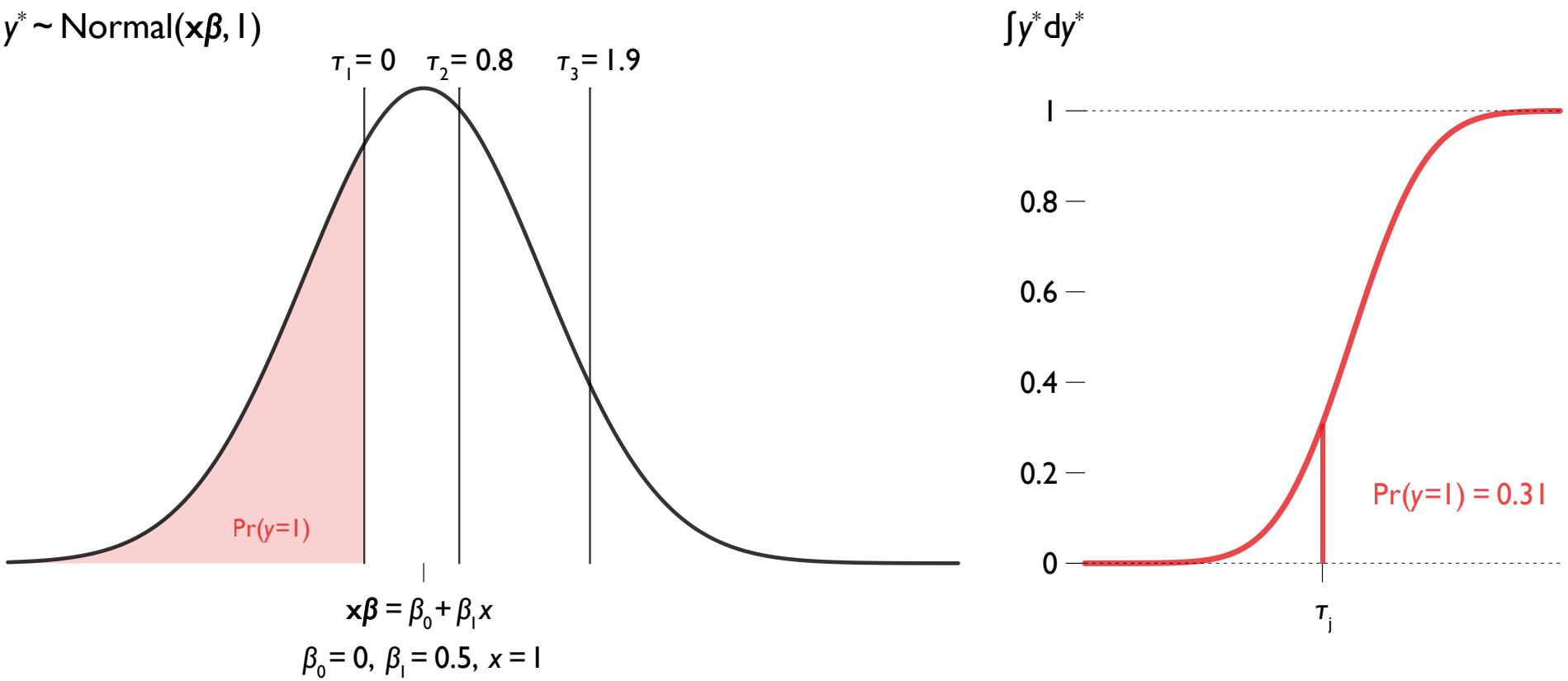


More compactly: $y_i = j \quad \text{if } \tau_{j-1} < y_i^* \leq \tau_j$

As in the binary case, we assume $\tau_0 = -\infty$, $\tau_1 = 0$, and $\tau_m = \infty$ for identification

That leaves $m - 2$ of the τ's as free parameters – here, these are τ_2 and τ_3

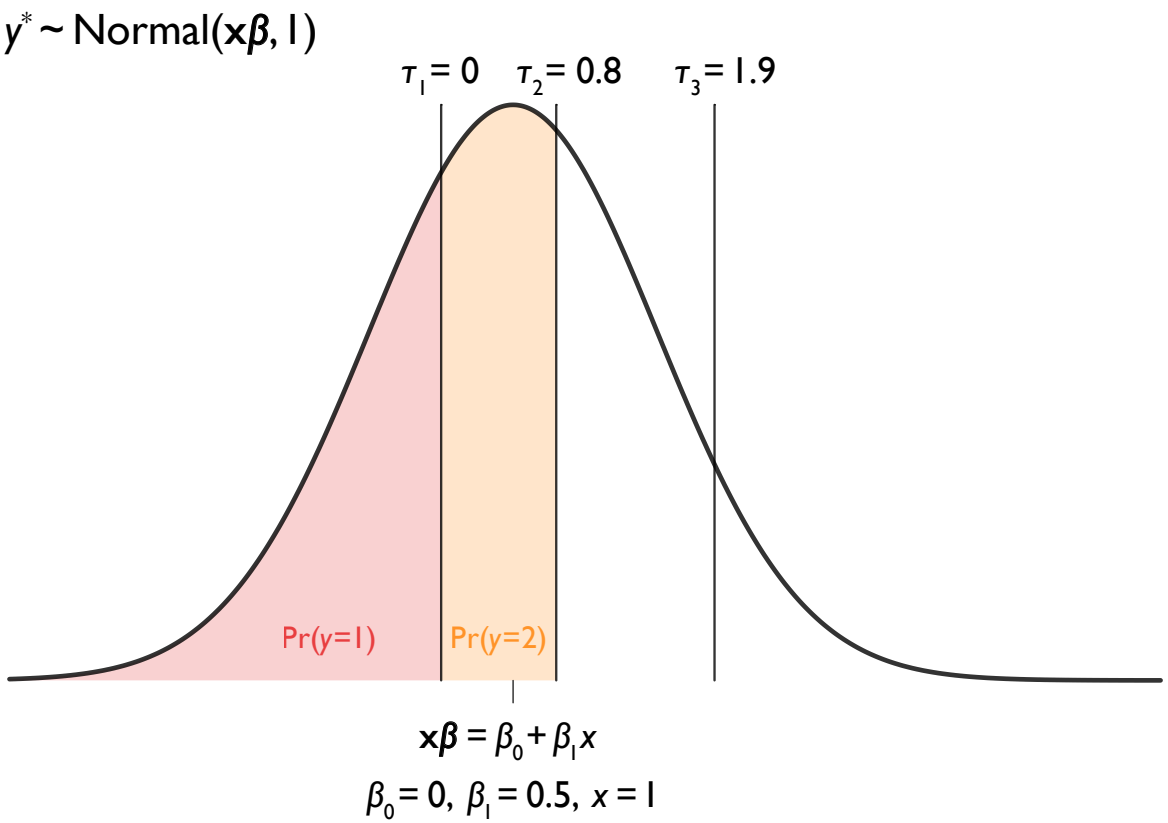
If we knew these τ's, how would we calculate the probability y_i falls in each category?

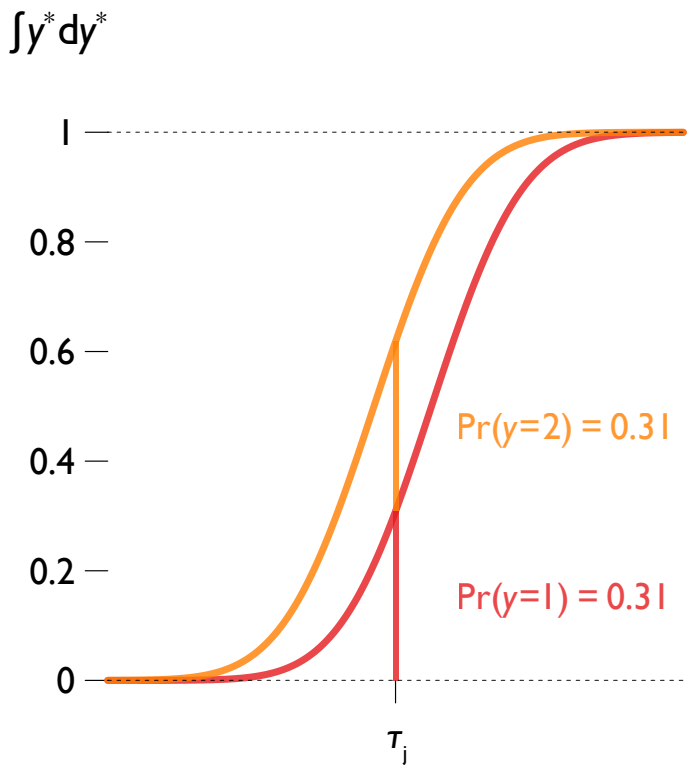


Calculating the probability of the *first* category is the same as in the binary case:

$$\Pr(y_i = 1) = \int_{-\infty}^{\tau_1 = 0} \text{Normal}(\mathbf{x}_i \hat{\boldsymbol{\beta}}, 1) \text{d}y_i^*$$

This is the probability the latent variable y_i^* falls between $\tau_0 = -\infty$ and $\tau_1 = 0$

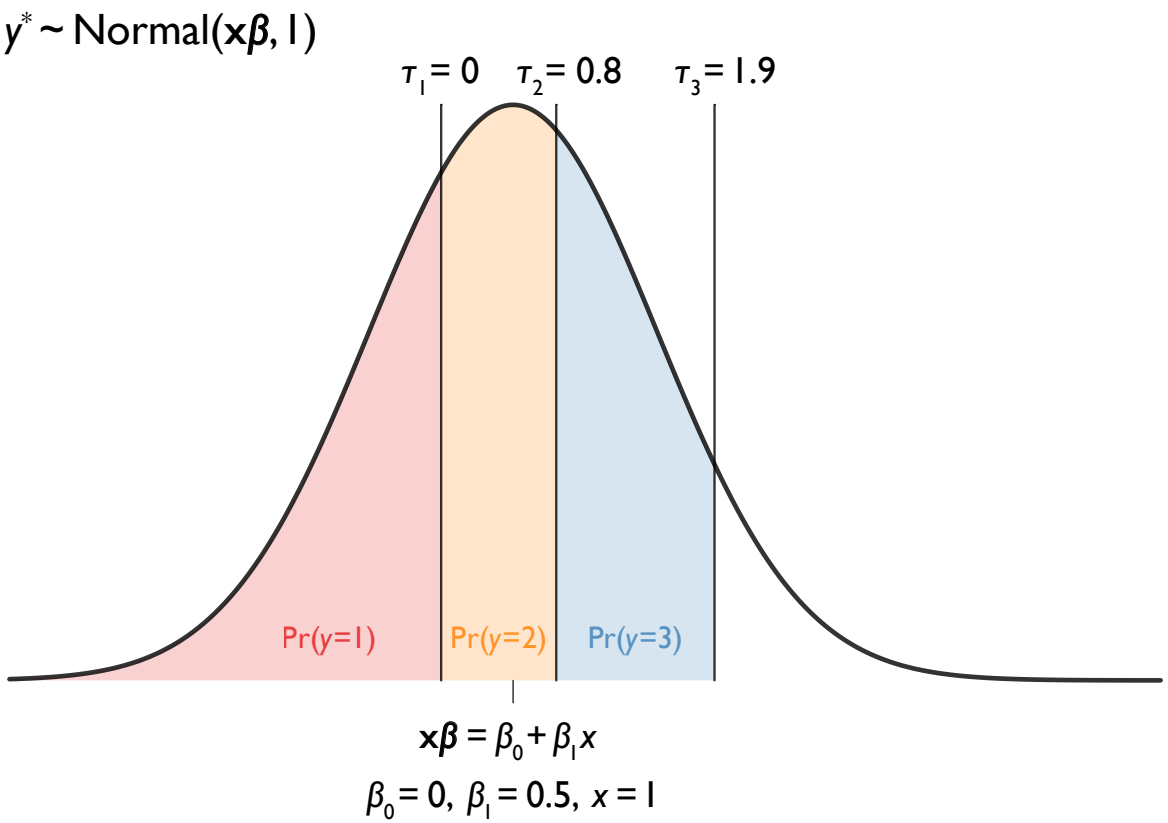


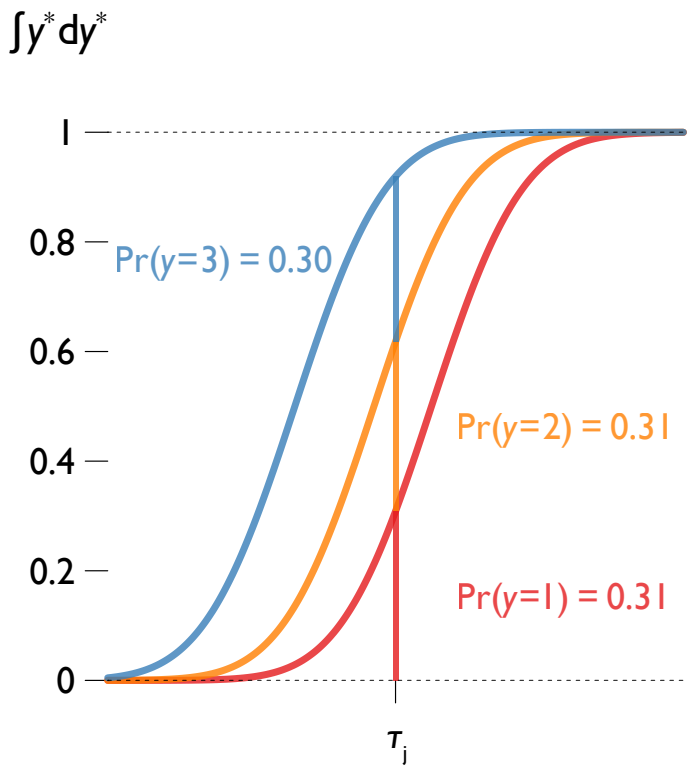


Calculating the probability of the *second* category follows the same logic – compute the chance y_i^* falls between the appropriate cutpoints

$$\Pr(y_i = 2) = \int_{\tau_1=0}^{\tau_2} \text{Normal}(\mathbf{x}_i\hat{\boldsymbol{\beta}}, 1)\text{d}y_i^*$$

But τ_2 can be any quantity between 0 and τ_3 – it is not fixed

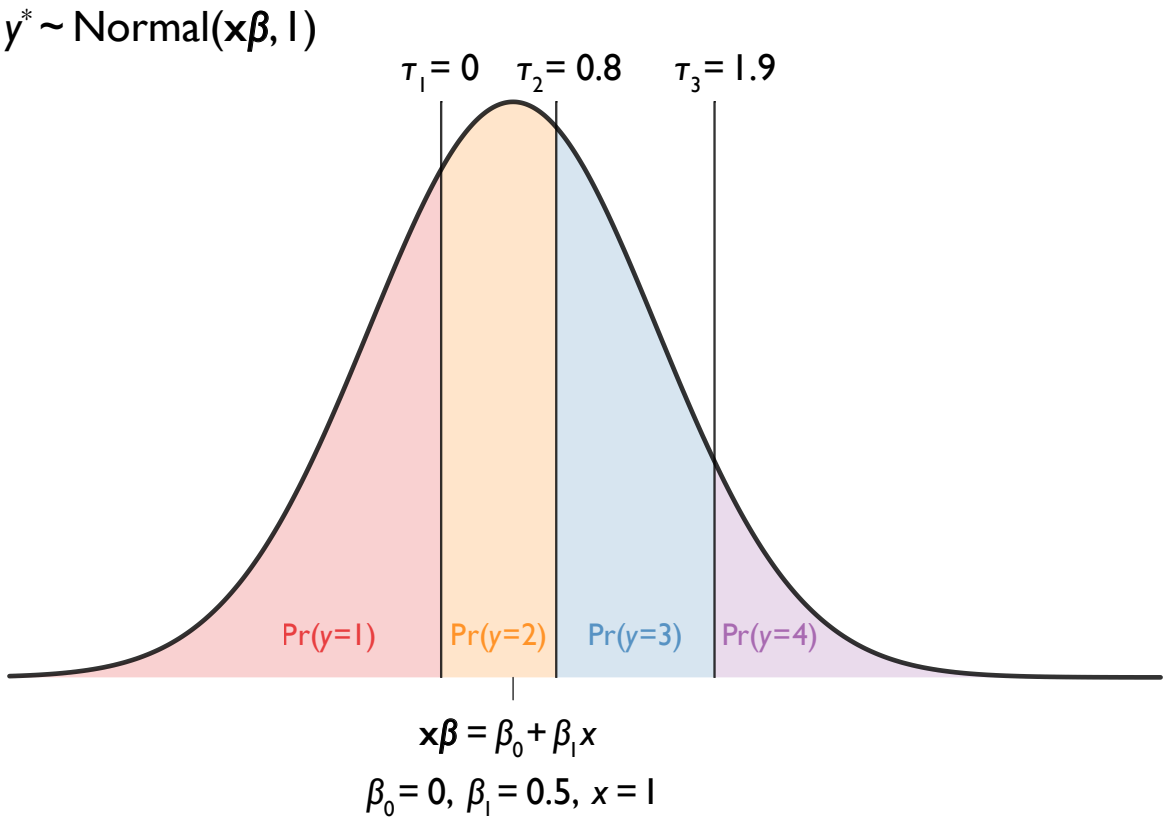


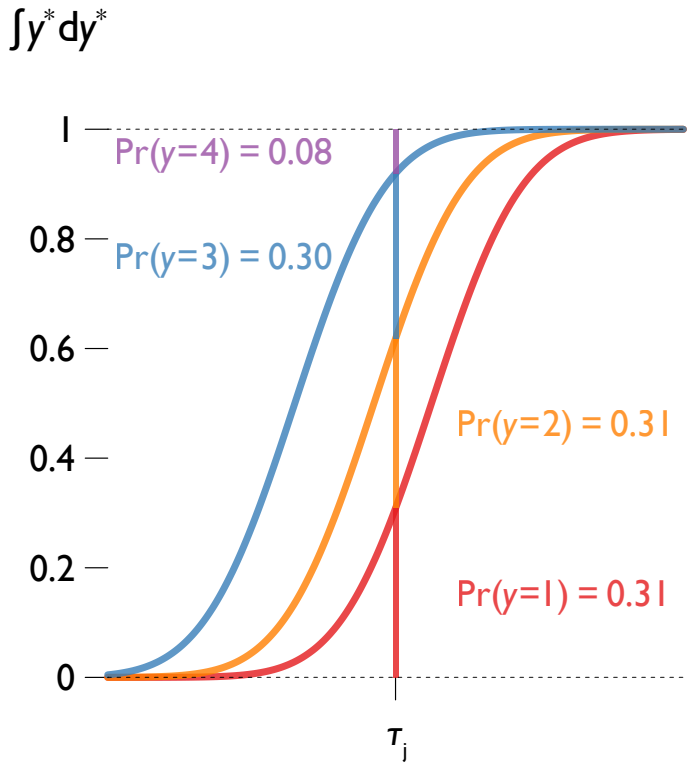


To calculate the probability of the *third* category,
insert the appropriate cutpoints into the Normal CDF:

$$\Pr(y_i = 3) = \int_{\tau_2}^{\tau_3} \text{Normal}(\mathbf{x}_i\hat{\boldsymbol{\beta}}, 1)\text{d}y_i^*$$

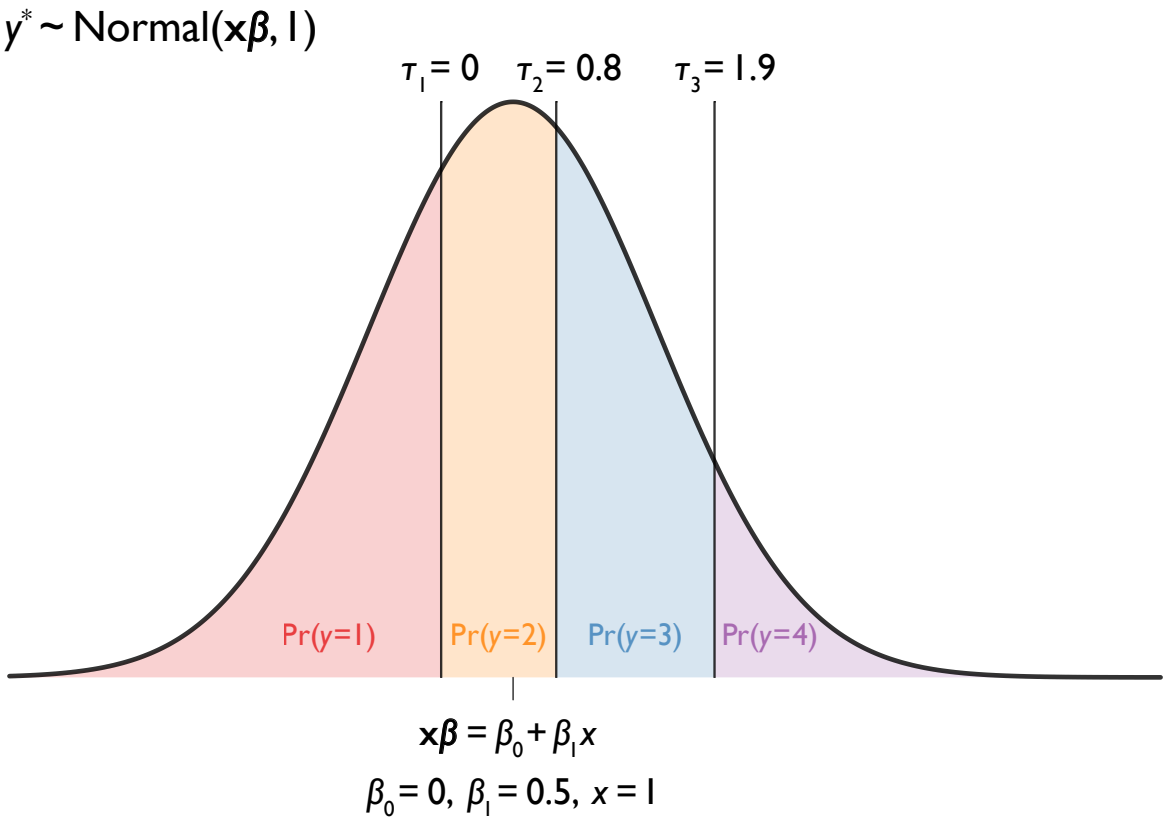
Because τ_2 and τ_3 are free parameters like $\boldsymbol{\beta}$, we'll need to estimate them

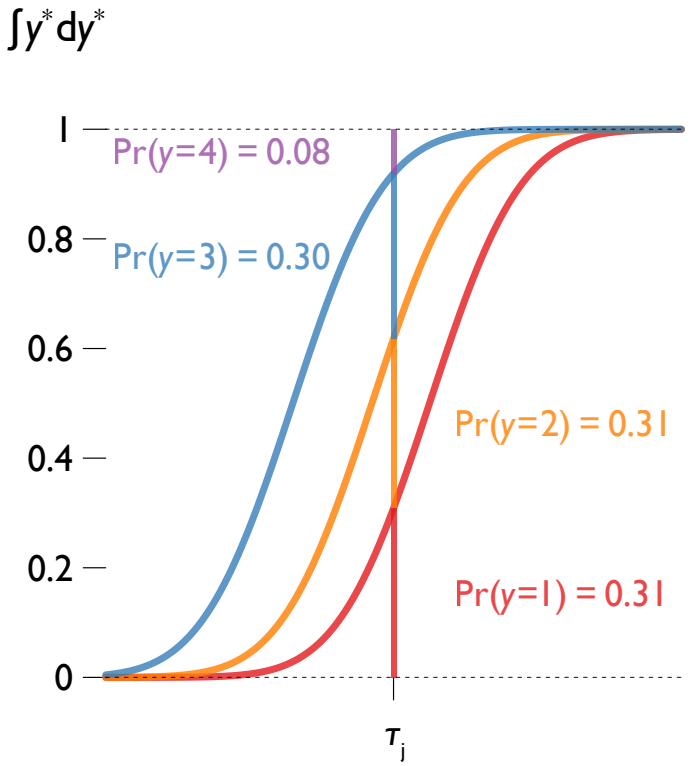




To calculate the probability of the *final* category,
we could insert the appropriate cutpoints into the Normal CDF:

$$\Pr(y_i = 4) = \int_{\tau_3}^{\infty} \text{Normal}(\mathbf{x}_i\hat{\boldsymbol{\beta}}, 1)\text{d}y_i^*$$





There's also a shortcut for the last category:
subtract the probability of the other categories from 1

$$\Pr(y_i = 4) = 1 - \sum_{j=1}^{m-1} \Pr(y_i = j)$$

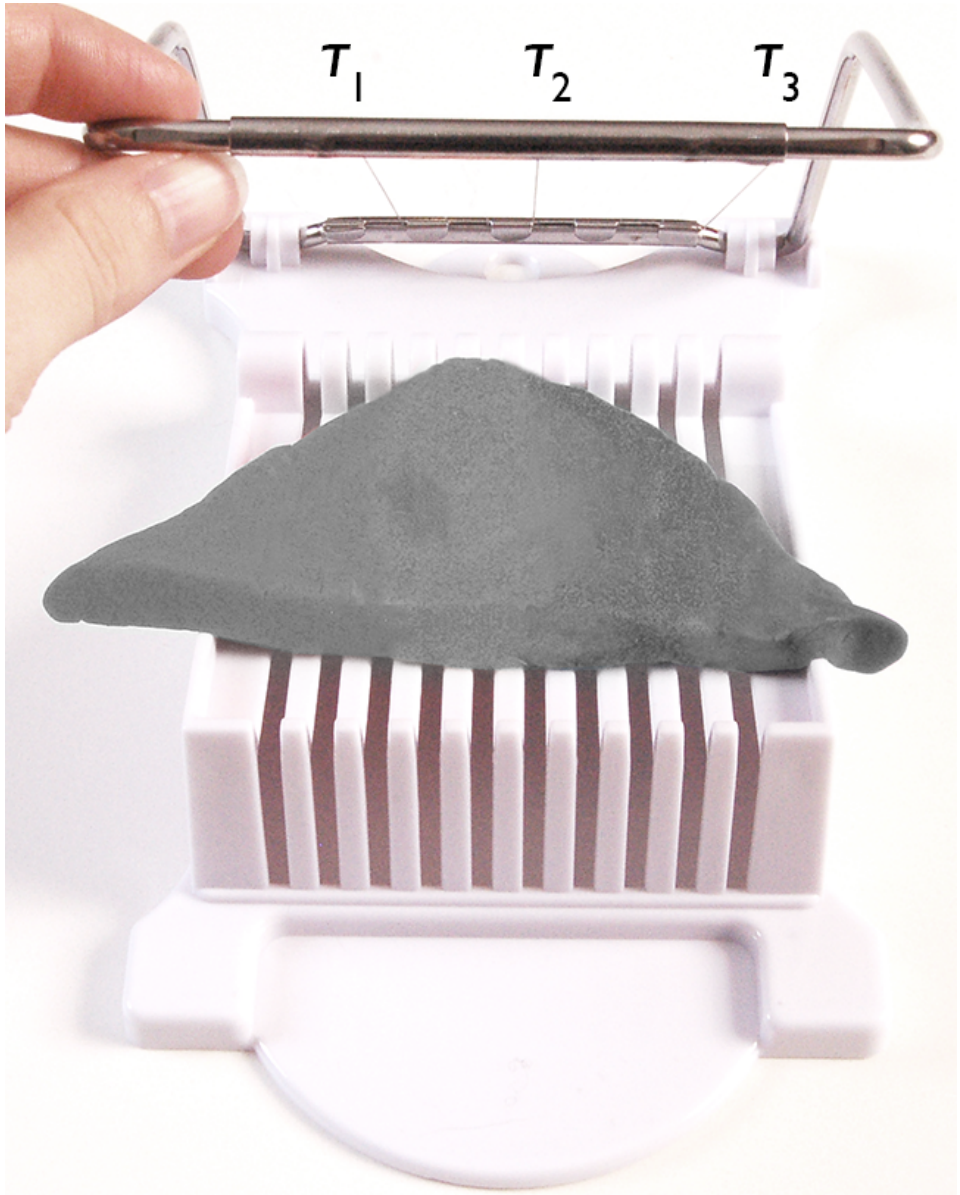
One way to solve ordered probit:
calculus

Another way:
a scale, an egg slicer, and some clay

Shape 100 grams of clay into a Normal distribution

Imagine drawing a speck of clay at random from this shape

Specks drawn from further right represent a greater chance of high categories of y

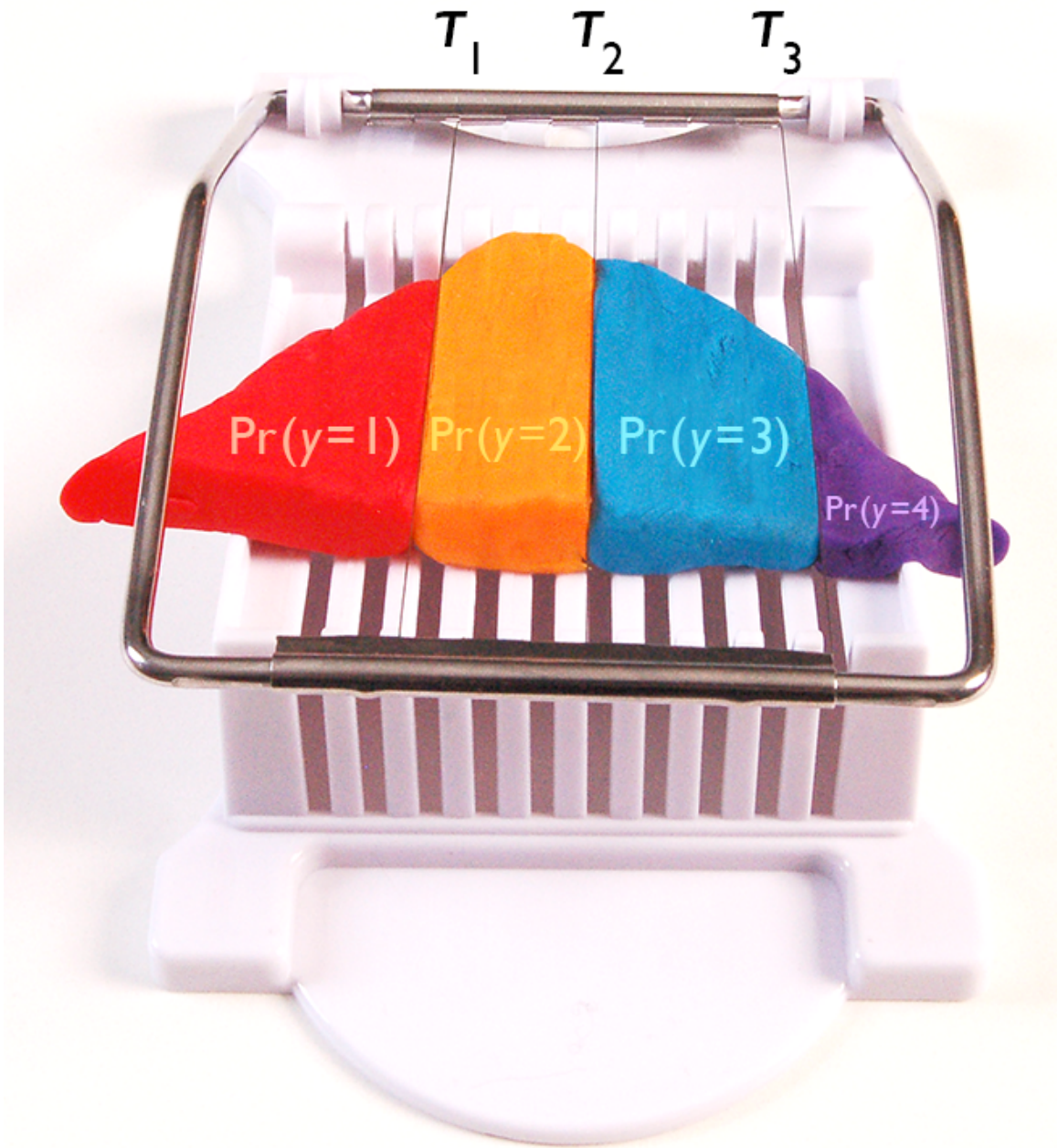


Load the clay in the egg slicer

Where do we to position the clay?
Its mean should be centered over $\mathbf{x}_i\boldsymbol{\beta}$

Let's assume this is the middle of the slicer for now

The slicer should have one wire for each (non-infinite) cutpoint, spaced apart as indicated by
τ_1, τ_2, τ_3, . . .



Slicing the clay produces four separate pieces

Each of these represents the relative weight of probability that a randomly drawn speck comes from a given category

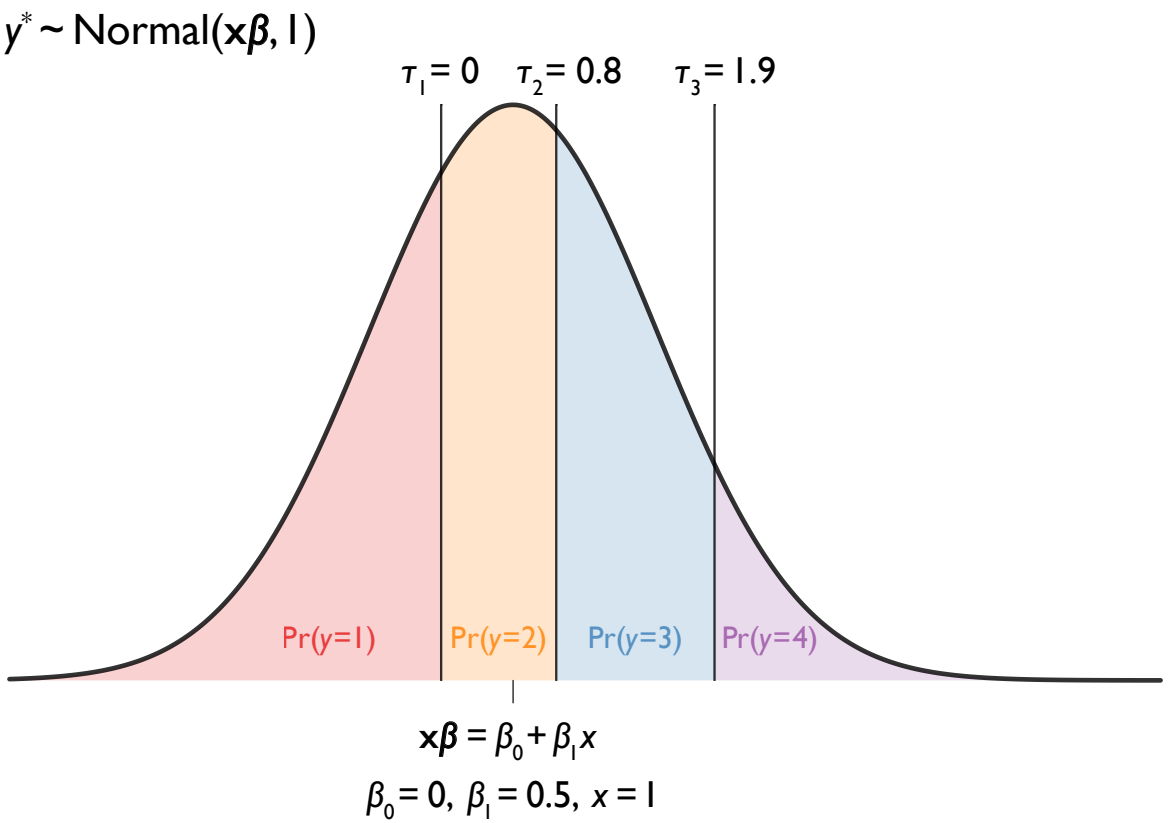
Before we used calculus to compute the probability of a draw coming from each piece. . .

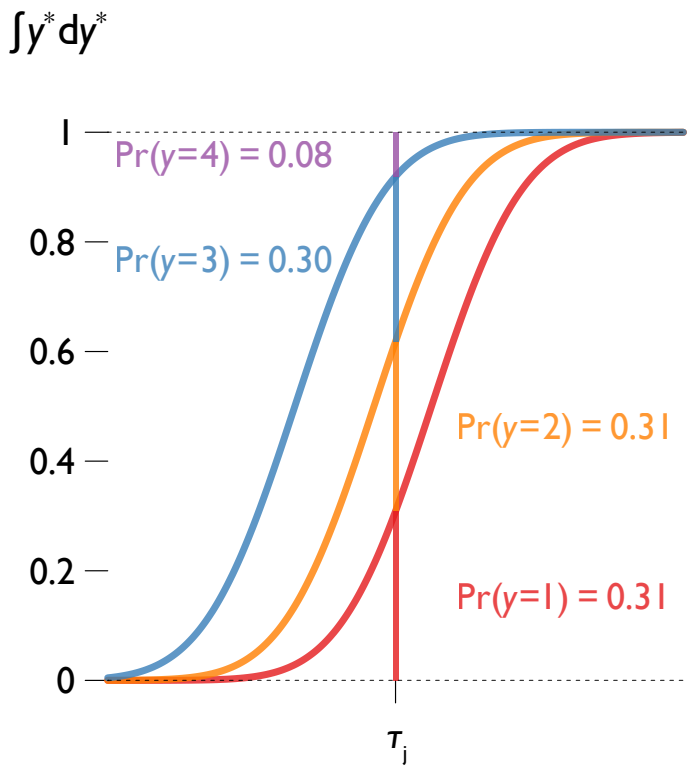
Assuming a constant density of clay, we could instead just weigh each piece

Because they sum to 100 grams, each gram indicates one percent of probability

The weights of these four categories match what we obtained from computing the Normal CDF for each piece of the latent variable

An egg slicer can do ordered probit!
But `R` is faster and cleaner, so. . .



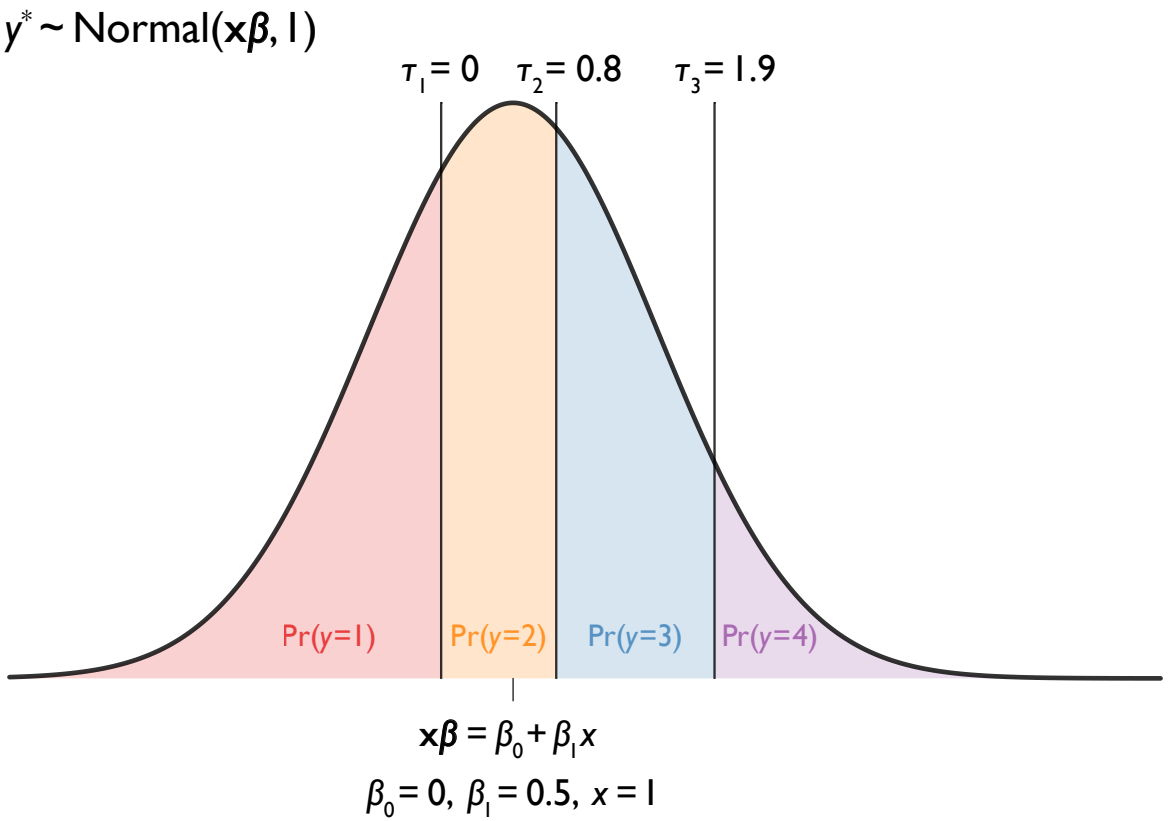


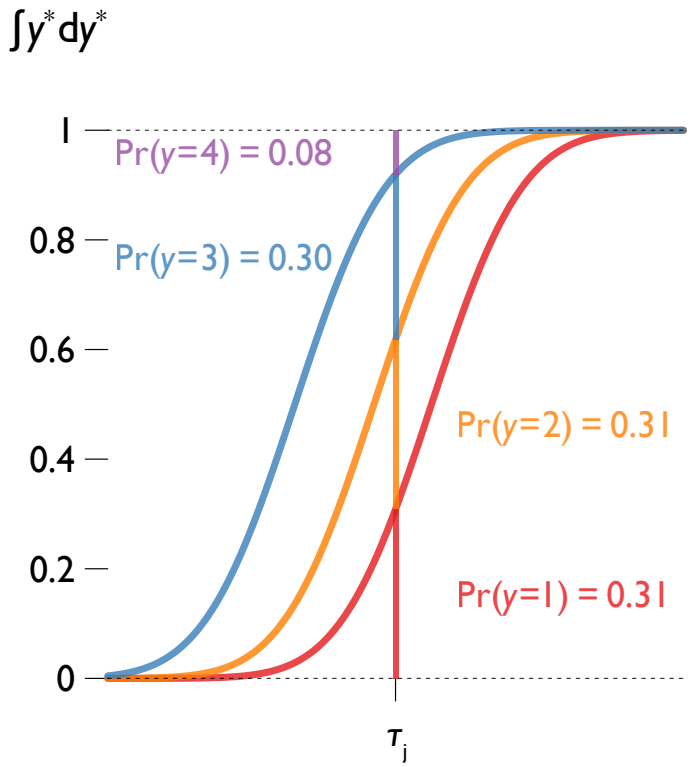
Notice the CDF curves on the right are all parallel

They are the same CDF shifted left or right by the τ's

This is the key identifying assumption of the model:
proportional (or parallel) probits for each category

$\rightarrow$ only need to estimate one probit (or logit) regardless of the number of categories

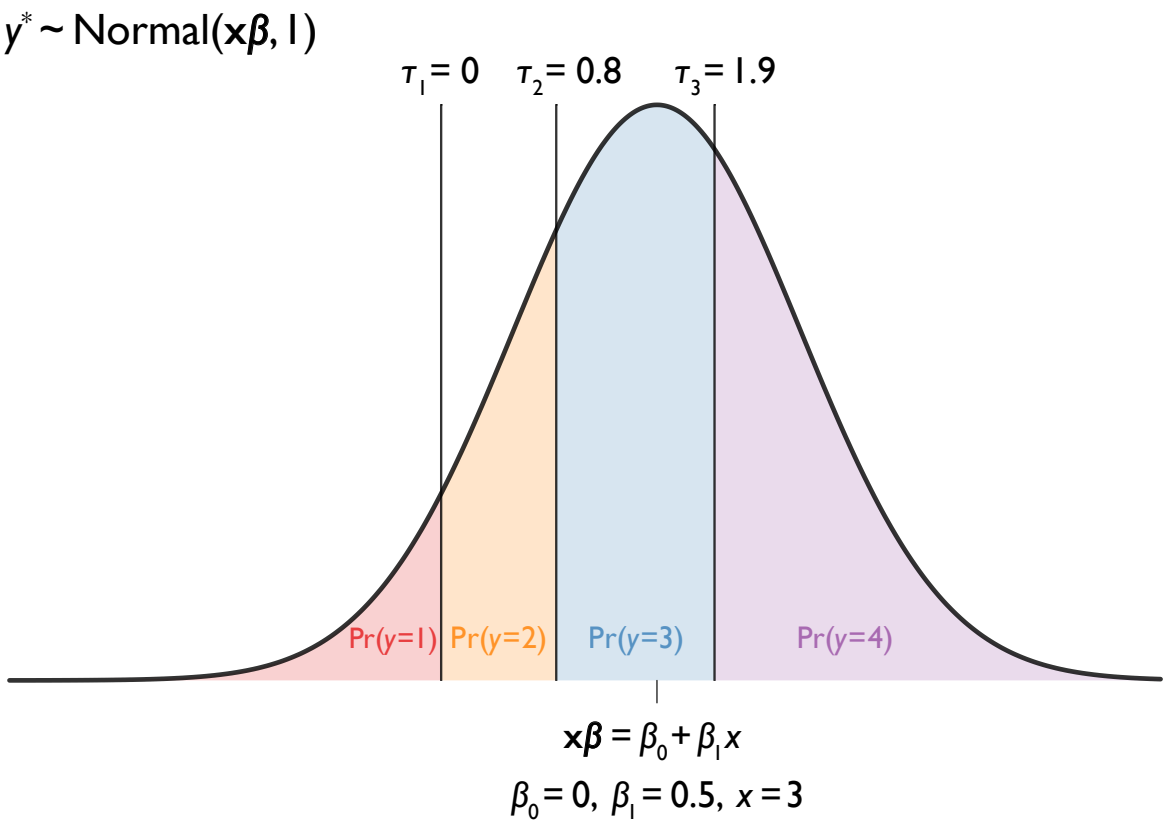


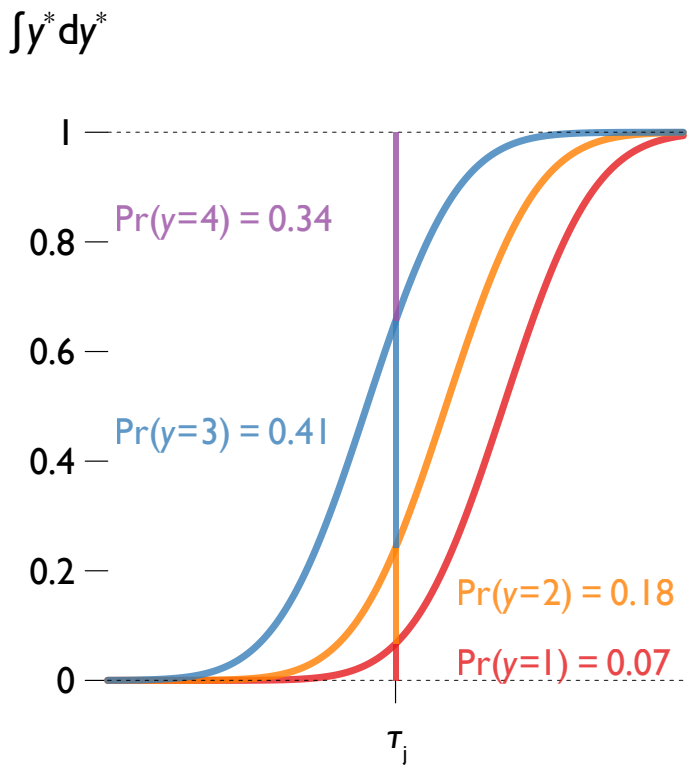


Just as in the binary case, the covariates $\mathbf{x}_i$ influence the probabilities of the categories by shifting the latent variable through the cutpoints

Suppose we increase x_1 from 1 to 3

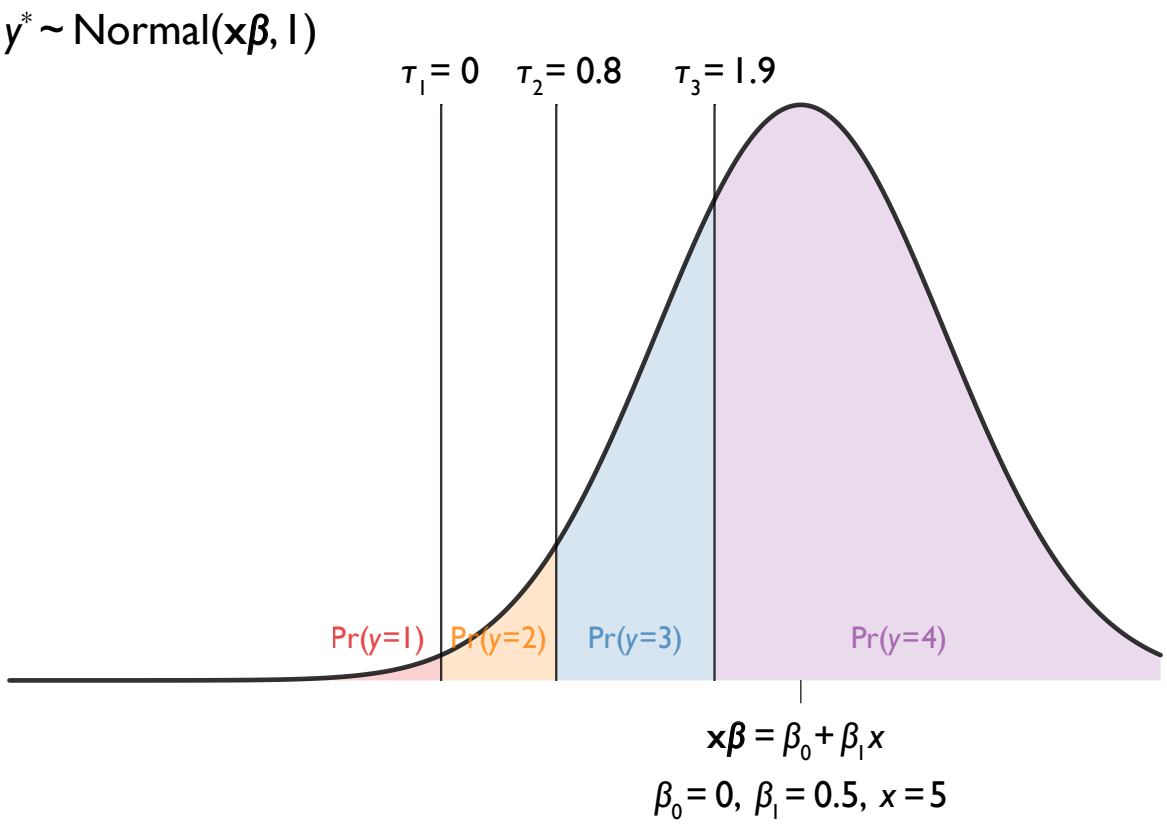
As before, this raises $\mathbf{x}_i\boldsymbol{\beta}$ by 1.5. . .

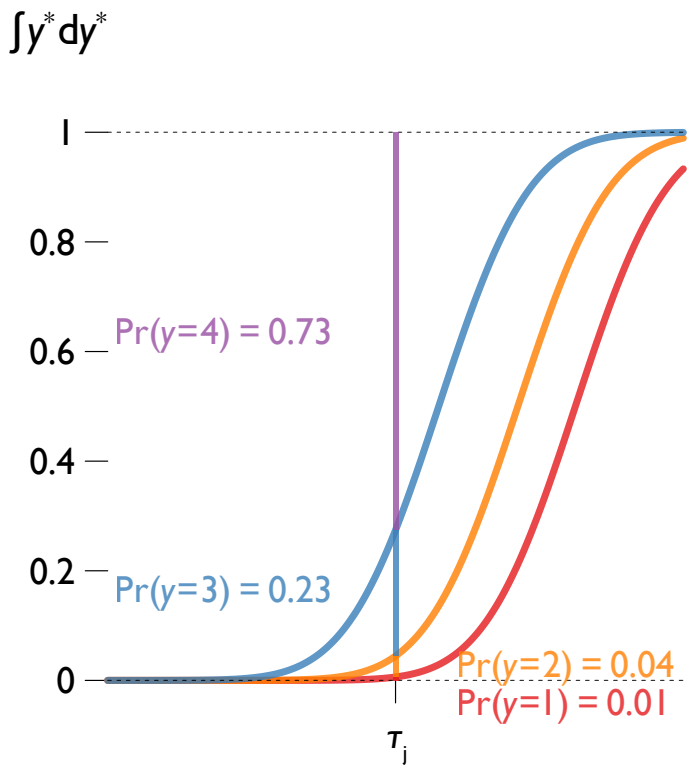




And shifts the latent variable pdf up, so more of its mass lies between high cutpoints

There is a corresponding shift in the CDFs,
so that the probability of categories 3 and 4 rise (and 1 and 2 shrink)

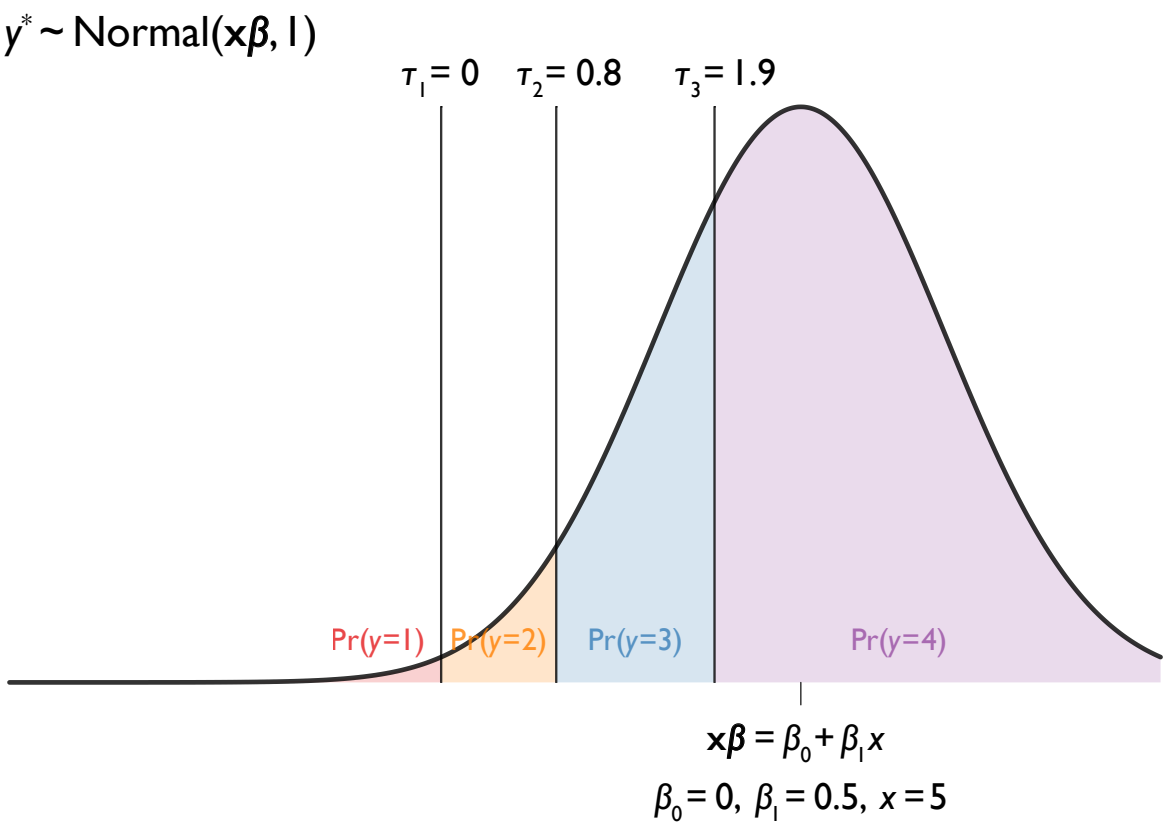


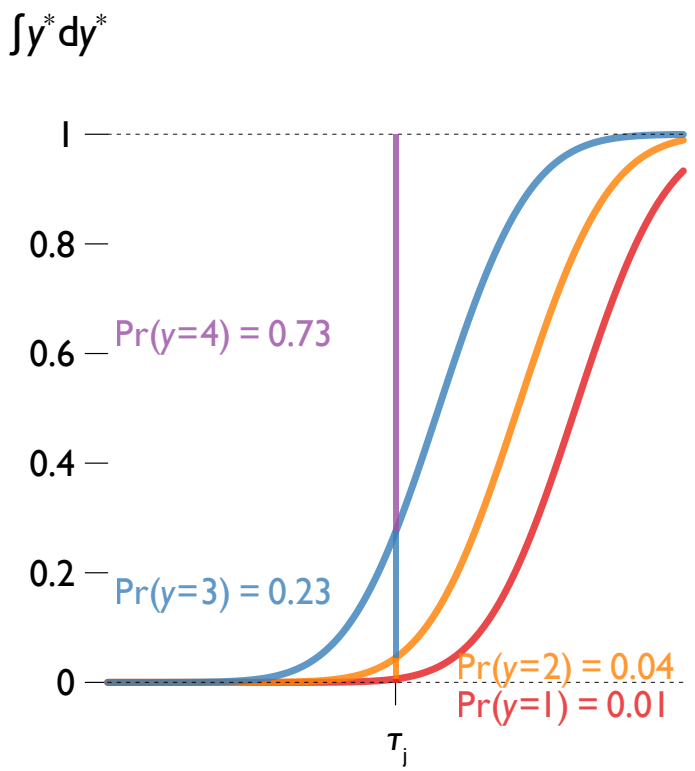


What if we raise x_1 even more, to 3, so that $\mathbf{x}_i\boldsymbol{\beta}$ is 2.5?

Now the pdf mass is largely concentrated between the two highest cutpoints

The probability of category 4 rises at the expense of all other categories



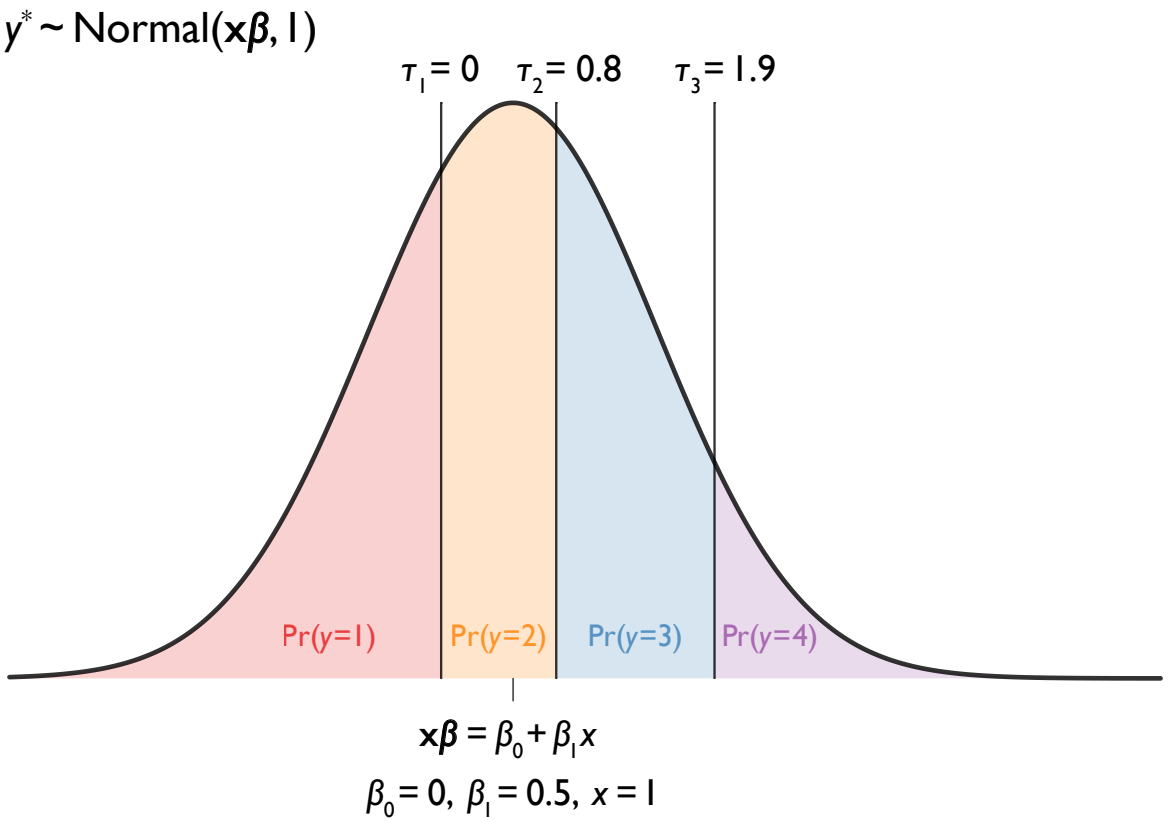


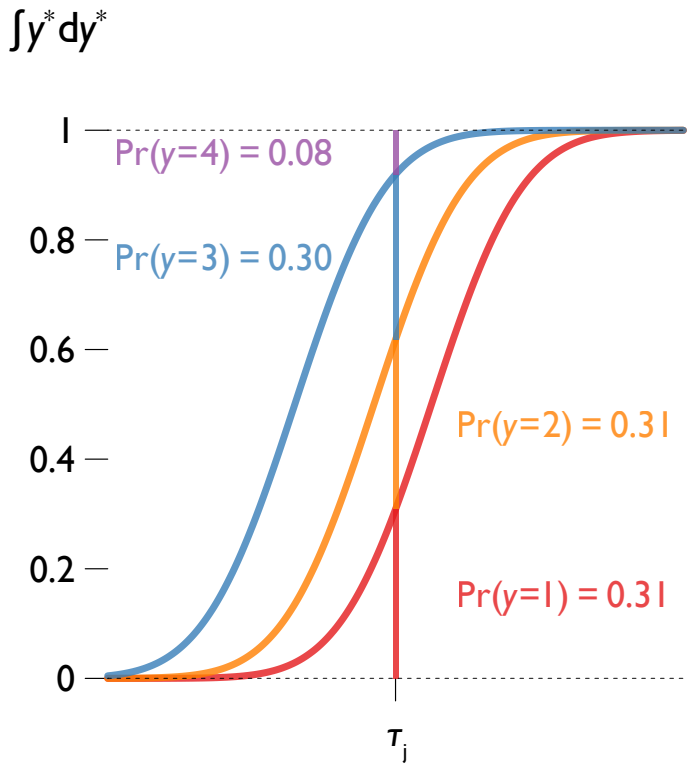
As $\mathbf{x}_i\boldsymbol{\beta}$ increases, $\Pr(y = j)$ for intermediate j first rises, then falls

If $\mathbf{x}_i\boldsymbol{\beta}$ is large enough, almost all the probability falls in the most extreme category

In that case, further increases in $\mathbf{x}_i\boldsymbol{\beta}$ will have little effect

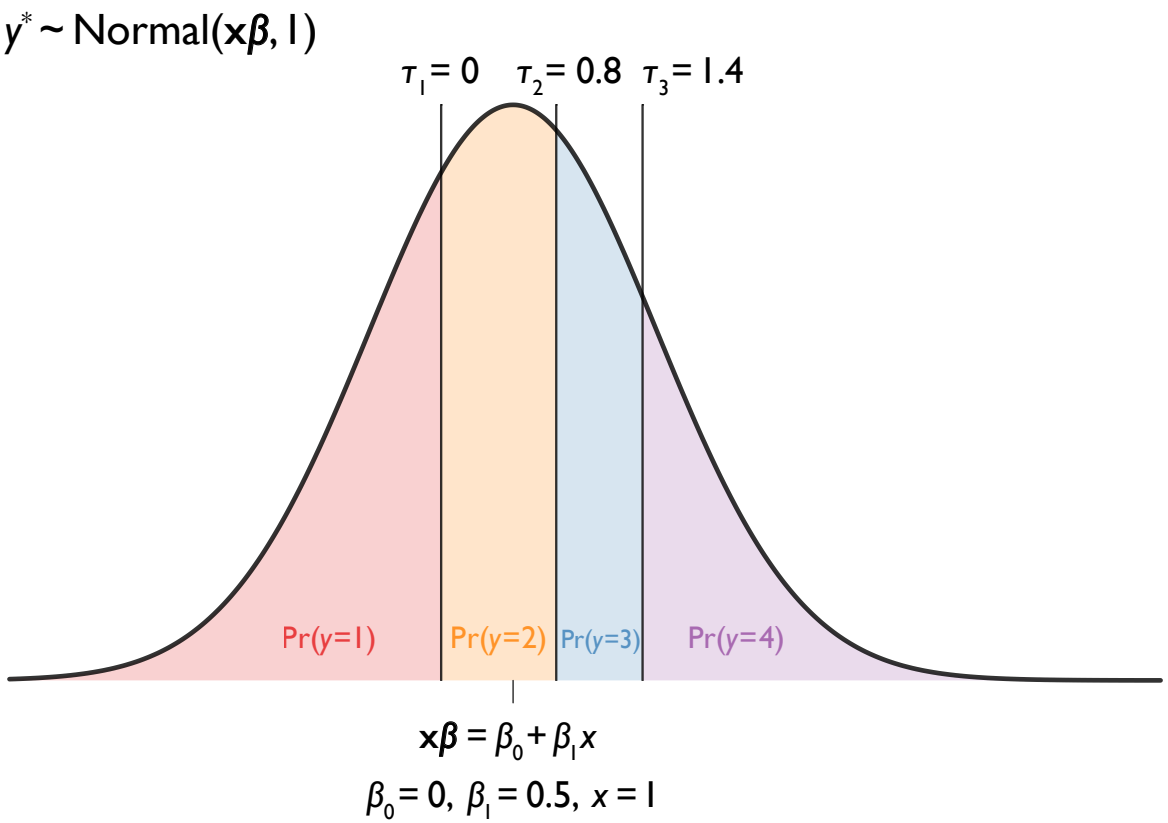
This makes the model nonlinear – like binary probit, the biggest effects of covariates are when the probabilities of the categories start out balanced

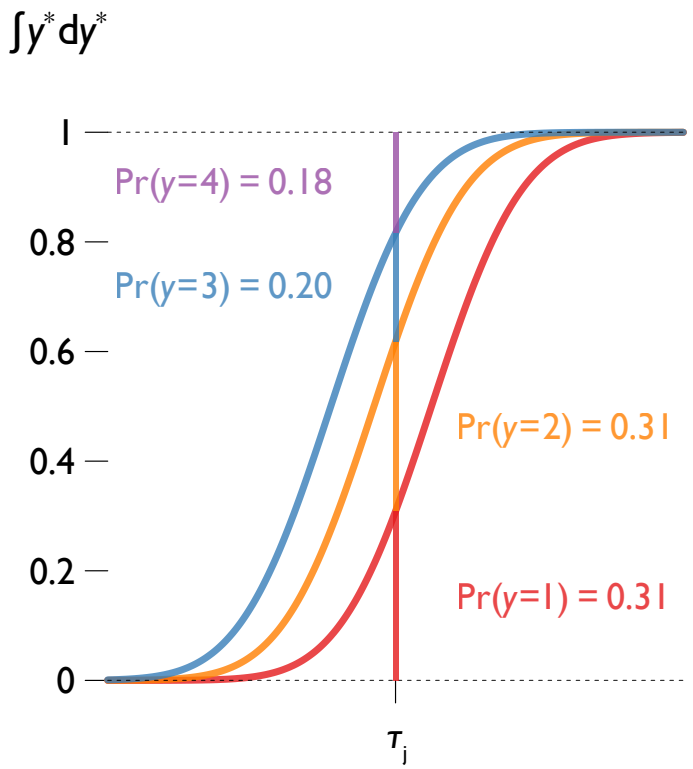




Returning to the original level of x_1,
what happens if we instead shift either of the moveable cutpoints τ_2 and τ_3?

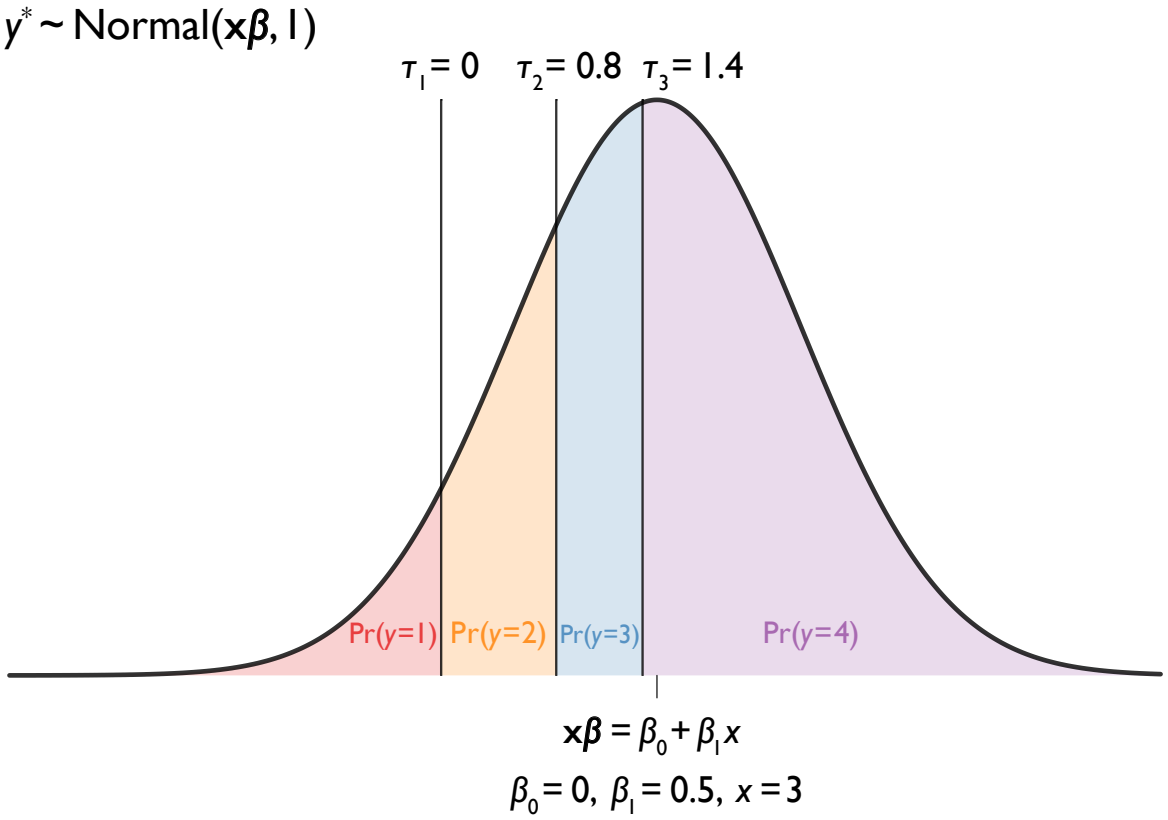
Let's move τ_3 and find out. . .

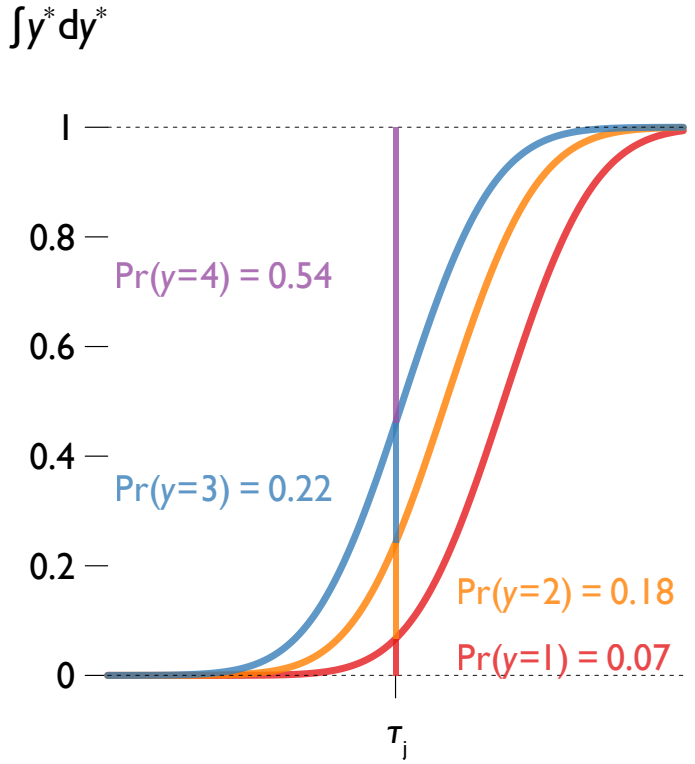




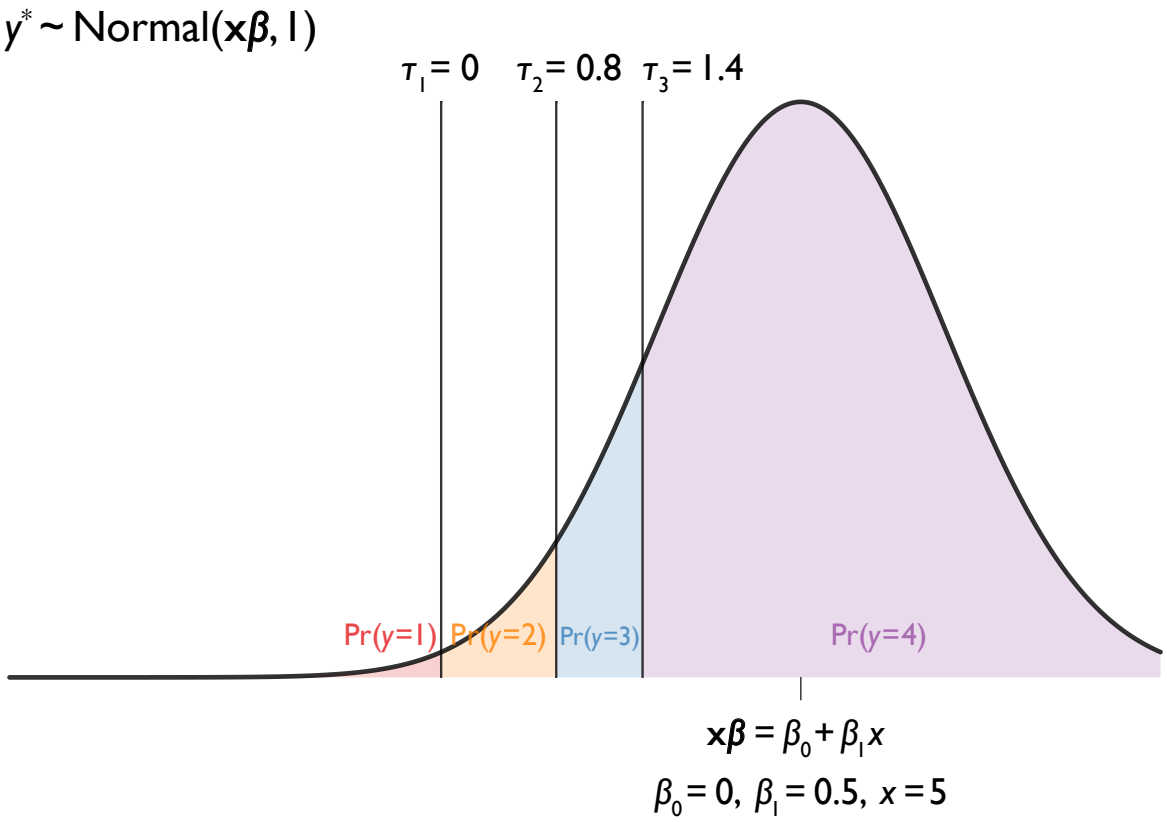
Suppose τ_3 is 1.4 instead of 1.9

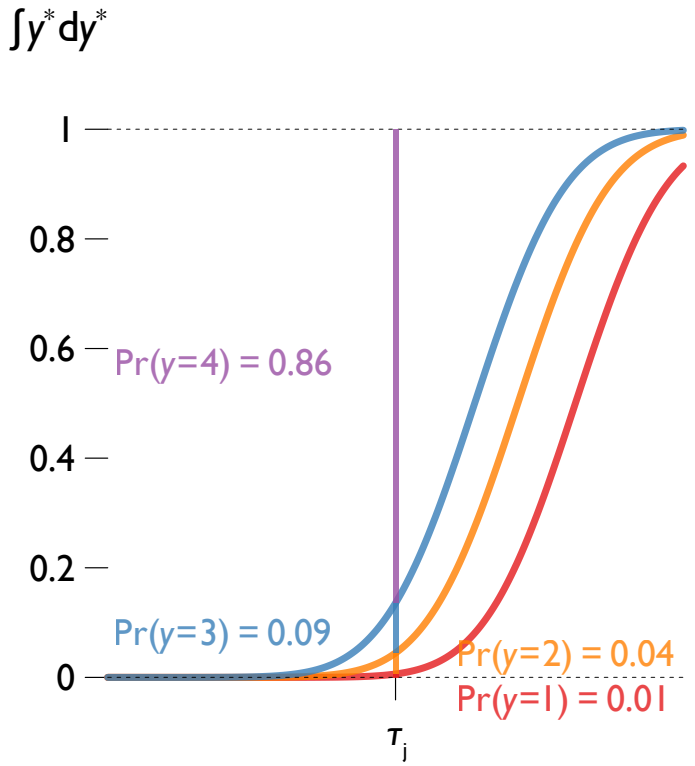
Because the cutpoints for category 3 are closer together,
the probability of this category shrinks and the probability of 4 rises





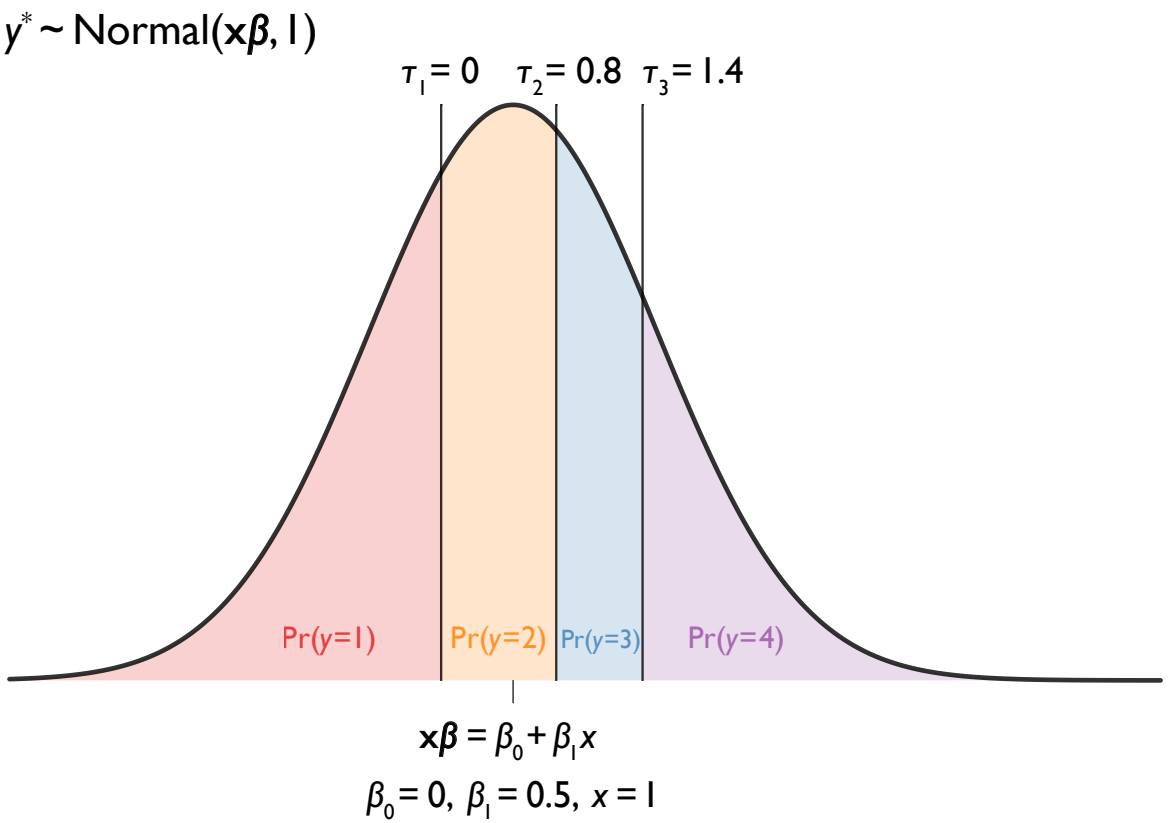
With the cutpoints closer together,
a shift in $\mathbf{x}_i\boldsymbol{\beta}$ more rapidly shifts probability to the highest category

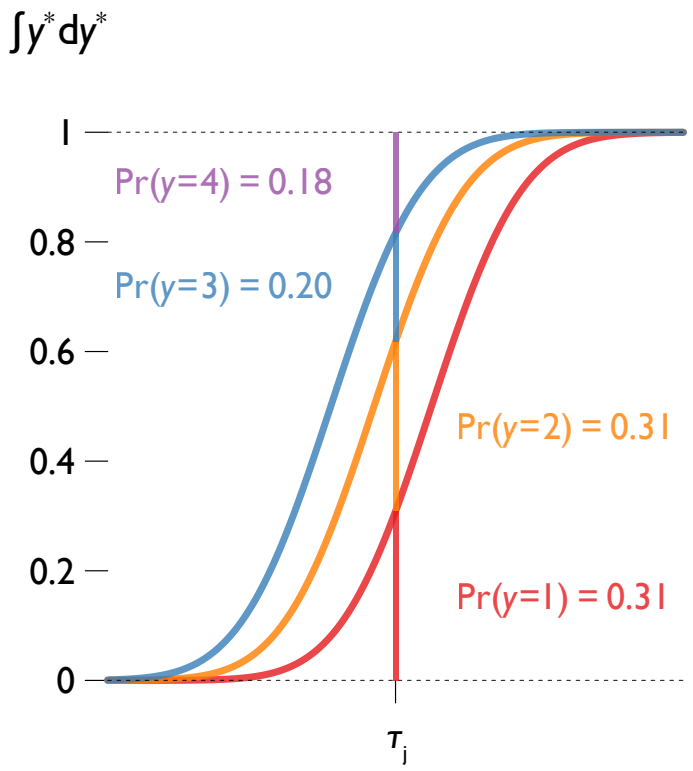




With the closer cutpoints,
$x_1 = 5$ produces a huge probability for $y = 4$
and little chance of other categories

From what you have seen so far, can you interpret β_1 directly?

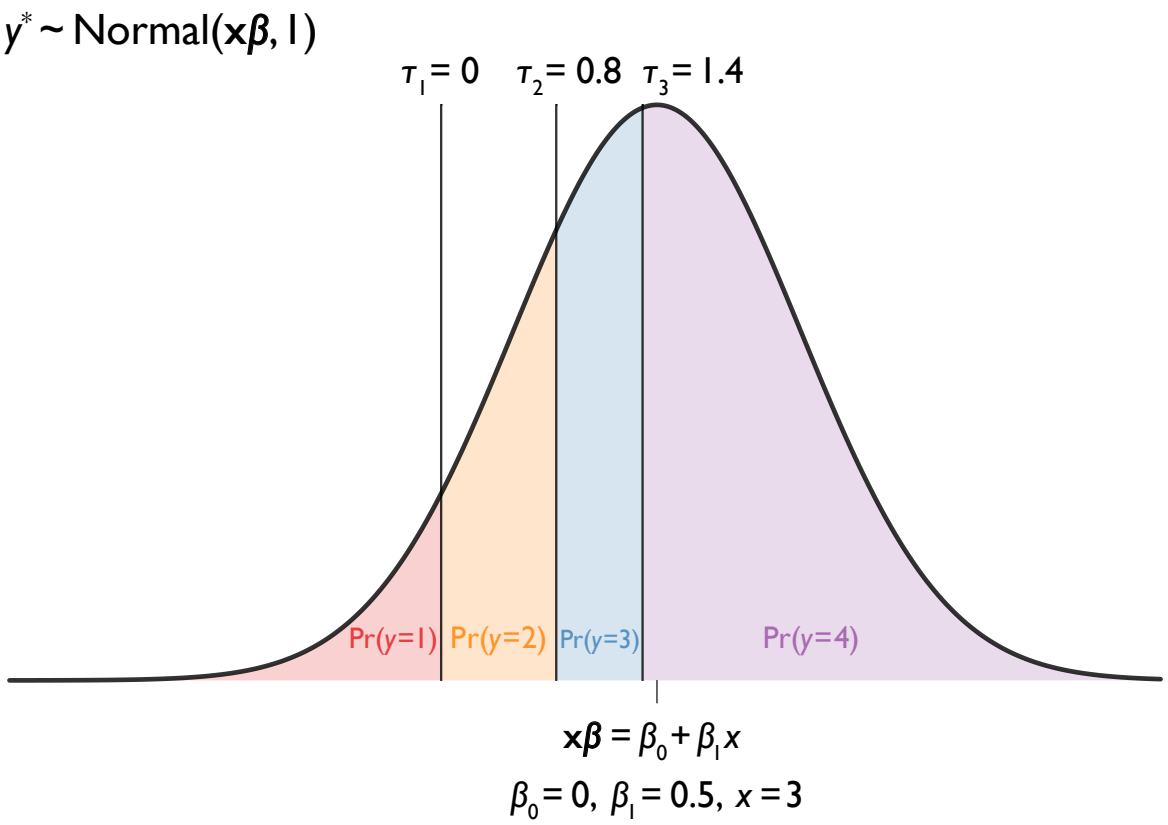


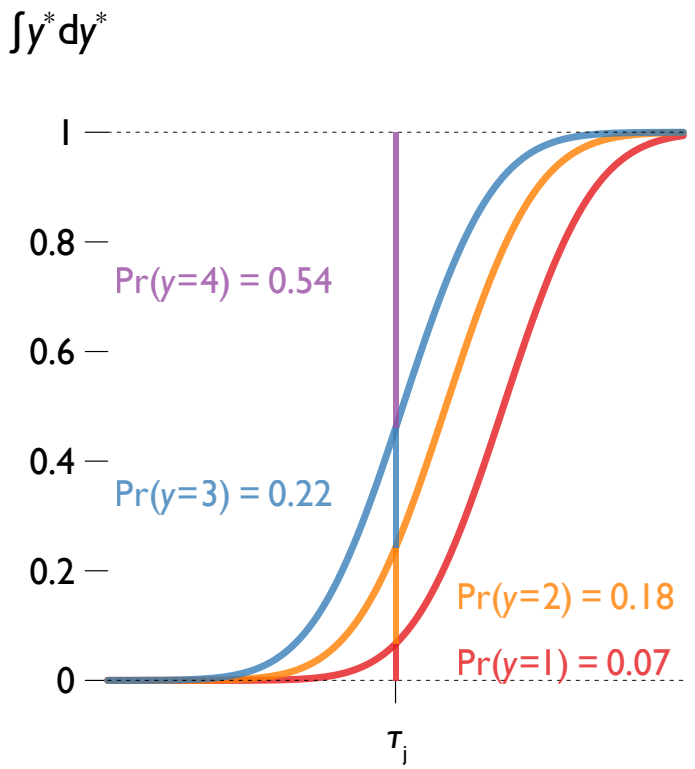


To a limited extent, ordered probit β's are interpretable, *if* we also know the τ's

Suppose that initially, x_1 is 1, so that in our model y^* is 0.5

This puts the mean of the latent variable pdf between $\tau_1 = 0$ and $\tau_2 = 0.8$, so in some sense $y_i = 2$ is the median expected category

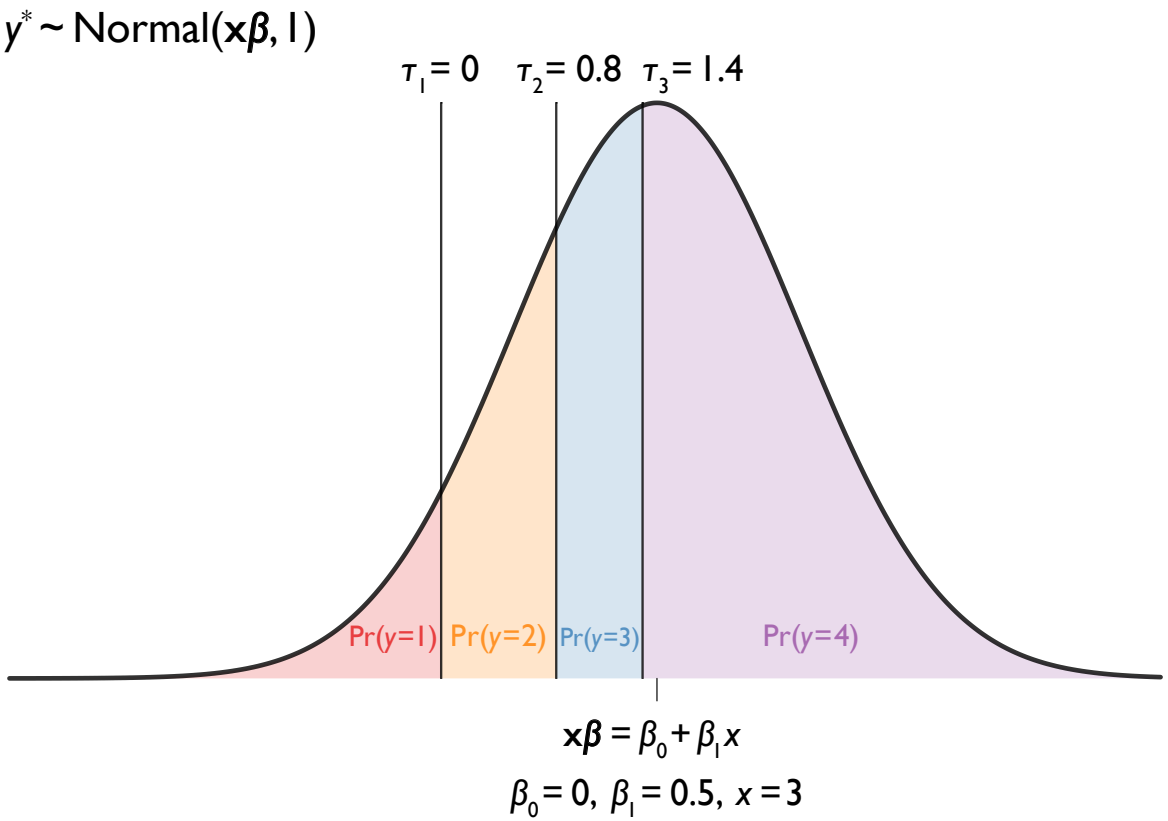


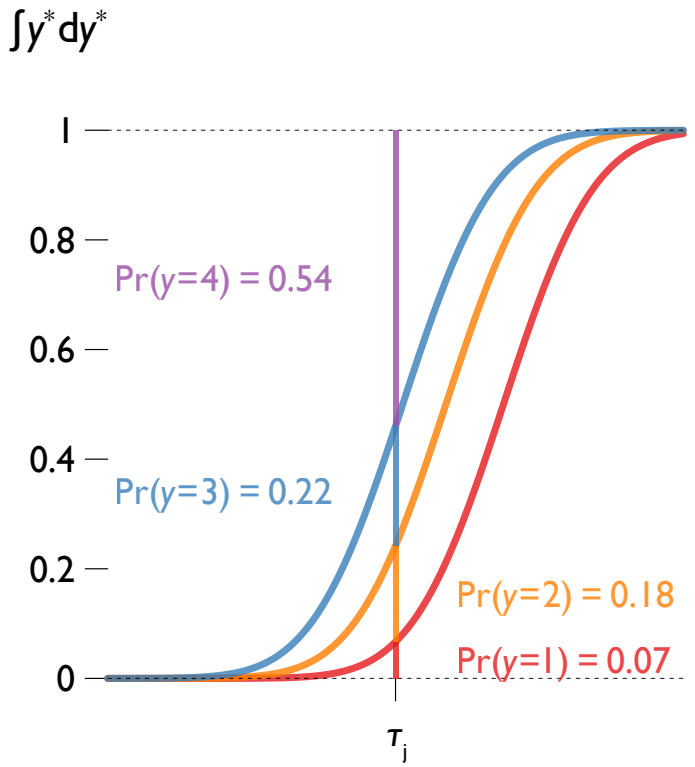


Now suppose that x_1 rises to 3:
because $\beta_1 = 0.5$, the most likely value of y^* is now 1.5

y^* is now most likely between $\tau_3 = 1.4$ and $\tau_4 = \infty$,
so now $y_i = 4$ is the median expected category

If we know that $\beta_1 = 0.5$, $\tau_2 = 0.8$, and $\tau_3 = 1.4$,
we can say a shift from $x_1 = 2$ to $x_1 = 3$ will probably raise y by two categories





Not totally satisfactory: "median category" is a bit vague, and all categories are still possible – but to what extent?

We'll devise better ways to interpret the model

But first, how do we estimate β and τ?

An ordered probit MLE

Let's build a likelihood for the ordered probit model

Start with the latent variable y_i^* and recall that

$$\begin{aligned} F_{\mathcal{N}}\left(y_i^* | \mu_i, \sigma^2\right) &= \int_{-\infty}^{y_i^*} f_{\mathcal{N}}\left(y_i^* | \mu_i, \sigma^2\right) dy_i^* \\ &= \int_{-\infty}^{y_i^*} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_i^* - \mu_i)^2}{2\sigma^2}\right) dy_i^* \\ &= \Phi\left(y_i^* | \mu_i, \sigma^2\right) \end{aligned}$$

We can thus write the CDF of the standard normal distribution as

$$\Phi\left(y_i^* | \mu_i, 1\right)$$

And we can have `R` calculate this:

```
pnorm(y, mean=mu, sd=1)
```

An ordered probit MLE

Next, we need a probability model of y_i

Let's recode y_i into a vector, s.t. $y_{ij} = 1$ if y_i^* falls into category j:

$$y_{ij} = \begin{cases} 1 & \text{if } \tau_{j-1} < y_i^* \leq \tau_j \\ 0 & \text{otherwise} \end{cases}$$

Now we can write the probability of y_{ij}:

$$\begin{aligned} \Pr(y_{ij} = 1) &= \Pr(\tau_{j-1} < y_i^* \leq \tau_j) \\ &= \int_{\tau_{j-1}}^{\tau_j} f_{\mathcal{N}}(y_i^* | \mu_i, 1)\, dy_i^* \\ &= \int_{-\infty}^{\tau_j} f_{\mathcal{N}}(y_i^* | \mu_i, 1)\, dy_i^* - \int_{-\infty}^{\tau_{j-1}} f_{\mathcal{N}}(y_i^* | \mu_i, 1)\, dy_i^* \\ &= \Phi(\tau_j | \mu_i, 1) - \Phi(\tau_{j-1} | \mu_i, 1) \end{aligned}$$

using `R`:

```
pnorm(tau[j], mean=mu[i], sd=1)
  - pnorm(tau[j-1], mean=mu[i], sd=1)
```

An ordered probit MLE

If y_i falls into category j, then the likelihood of y_i is proportional to $\Pr(y_{ij} = 1)$.
All other categories $\neq j$ are irrelevant.

This requirement suggests maximizing the following likelihood,

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\tau}|\mathbf{y}) = \prod_{i=1}^{n} \left\{ \prod_{j=1}^{m} \left[\Phi(\tau_j|\mathbf{x}_i\boldsymbol{\beta}, 1) - \Phi(\tau_{j-1}|\mathbf{x}_i\boldsymbol{\beta}, 1) \right]^{y_{ij}} \right\}$$

which "picks out" the probability of the right category for each observation.

Taking logs, we have

$$\log \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\tau}|\mathbf{y}) = \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij} \log \left[\Phi(\tau_j|\mathbf{x}_i\boldsymbol{\beta}, 1) - \Phi(\tau_{j-1}|\mathbf{x}_i\boldsymbol{\beta}, 1) \right]$$

When programming this likelihood, there is one caveat:
The $\boldsymbol{\tau}$'s must remain in order, and nothing in $\log \mathcal{L}(\cdot)$ guarantees this.
Solution: penalty functions – subtract 1,000,000 from $\log \mathcal{L}(\cdot)$ if order wrong.

Ordered probit quantities of interest

In models of ordered data, we get one expected probability for each category of y

These probabilities sum to one:

$$\begin{aligned}
\Pr(y = 1|\mathbf{x}_i\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\tau}}) &= \hat{\pi}_{1i} \\
\Pr(y = 2|\mathbf{x}_i\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\tau}}) &= \hat{\pi}_{2i} \\
\Pr(y = 3|\mathbf{x}_i\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\tau}}) &= \hat{\pi}_{3i} \\
&\dots \\
\Pr(y = m|\mathbf{x}_i\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\tau}}) &= 1 - \sum_{j=1}^{m-1} \hat{\pi}_{ji}
\end{aligned}$$

Ordered probit quantities of interest

In models of ordered data, we get one expected probability for each category of y

To calculate these $\hat{\pi}$'s we use the standard normal CDF:

$$
\begin{aligned}
\Pr(y = 1|\mathbf{x}_i\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\tau}}) &= \Phi\left(\hat{\tau}_1|\mathbf{x}_i\hat{\boldsymbol{\beta}}, 1\right) \\
\Pr(y = 2|\mathbf{x}_i\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\tau}}) &= \Phi\left(\hat{\tau}_2|\mathbf{x}_i\hat{\boldsymbol{\beta}}, 1\right) - \Phi\left(\hat{\tau}_1|\mathbf{x}_i\hat{\boldsymbol{\beta}}, 1\right) \\
\Pr(y = 3|\mathbf{x}_i\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\tau}}) &= \Phi\left(\hat{\tau}_3|\mathbf{x}_i\hat{\boldsymbol{\beta}}, 1\right) - \Phi\left(\hat{\tau}_2|\mathbf{x}_i\hat{\boldsymbol{\beta}}, 1\right) \\
&\dots \\
\Pr(y = m|\mathbf{x}_i\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\tau}}) &= 1 - \Phi\left(\hat{\tau}_{m-1}|\mathbf{x}_i\hat{\boldsymbol{\beta}}, 1\right)
\end{aligned}
$$

These are the colored vertical line segments from the CDF plots

If you wanted a counterfactual $\mathbf{x_c}$, you could substitute it above

How do we actually do this in R?

Ordered probit quantities of interest

In models of ordered data, we get one expected probability for each category of y

In `R`, for a four category probit, let `xib <- x[i,]%*%beta`, and then:

$$
\begin{aligned}
\Pr(y=1|\mathbf{x}_i\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\tau}}) &= \texttt{pnorm(0, mean=xib)} \\
\Pr(y=2|\mathbf{x}_i\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\tau}}) &= \texttt{pnorm(tau2, mean=xib) - pnorm(0, mean=xib)} \\
\Pr(y=3|\mathbf{x}_i\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\tau}}) &= \texttt{pnorm(tau3, mean=xib) - pnorm(tau2, mean=xib)} \\
\Pr(y=4|\mathbf{x}_i\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\tau}}) &= \texttt{1 - pnorm(tau3, mean=xib)}
\end{aligned}
$$

Predicted values are similar, but instead of a probability of each category,
we predict a single category will happen (we get only 1s and 0s for each $\tilde{y}_{ij}$)

Slightly different for ordered logit (substitute logistic CDF)

Confidence intervals's for $\hat{\pi}$ depend on uncertainty in $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\tau}}$:
We will need to *simulate*

The `simcf` package can help:
see `oprobitsimev()`, `oprobitsimfd()`, and `oprobitsimrr()`

Goodness of Fit

The gamut of GOF tests from last week applies to ordered choice models as well

Many are straightforward (LR, BIC, Wald)

Some may take a bit of thought:

Percent correctly predicted	*The right category vs. the wrong ones?*
Actual vs. Predicted plots	*One for each category?*
ROC plots	*Right vs. wrong? Something multidimensional?*
Concordance index / AUC	*Right vs. wrong? Something multidimensional?*

Final concern: validating the parallel probits / proportional logits assumption

Statistical tests exist but are widely criticized as too conservative

Instead, use week's model (multinomial logit) as a robustness check:
if it makes a substantive difference, use MNL instead of ordered probit

How to write the Ordered Probit model in a paper

When writing up an ordered probit (or ordered logit) in a paper,
don't write out a linear regression model (like $y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$)
and say you estimated it with "ordered probit"

Ordered probit is not a linear model and does not have an error term

Either just say you used ordered probit and avoid writing an equation,
or write the model compactly, e.g.:

I estimated an ordered probit model where the probability that
the ordered response y falls into the j category for case i is given by:

$$\Pr\left(y_i = j|\mathbf{x}_i\right) = \int_{\tau_{j-1}}^{\tau_j} \text{Normal}\left(\mathbf{x}_i\boldsymbol{\beta}, 1\right) \mathrm{d}\mathbf{x}_i\boldsymbol{\beta},$$

where x is a vector of covariates, $\boldsymbol{\beta}$ is a vector of coefficients,
and $\boldsymbol{\tau}$ is a $j+1$ vector of cutpoints with $t_0 = -\infty$, $t_1 = 0$,
and $t_j = \infty$ for identification.

Sexist Attitudes to Working Mothers

In 1977 and 1989, the General Social Survey read a statement to respondents (R's):

"A working mother can establish just as warm and secure of a relationship with her child as a mother who does not work."

R's could strongly disagree (D), disagree (d), agree (a), or strongly agree (A):

Percent who. . .	strongly disagree	disagree	agree	strongly agree
1977 sample (n=1379)	17.1	33.3	33.8	15.7
1989 sample (n=914)	6.7	28.8	42.7	21.9

Combining disagree's (d+D) and agree's (a+A) can be illuminating, but beware this throws away information!

Percent who. . .	disagree or strongly disagree	agree or strongly agree
1977 sample (n=1379)	50.5	49.5
1989 sample (n=914)	35.4	64.6

Sexist Attitudes to Working Mothers

In 1977 and 1989, the General Social Survey read a statement to respondents (R's):

"A working mother can establish just as warm and secure of a relationship with her child as a mother who does not work."

R's could strongly disagree (D), disagree (d), agree (a), or strongly agree (A):

Percent who. . .	strongly disagree	disagree	agree	strongly agree
1977 sample (n=1379)	17.1	33.3	33.8	15.7
1989 sample (n=914)	6.7	28.8	42.7	21.9

We also have five covariates that might explain variation in this response:

male	whether the respondent was male (1) or female (0)
white	whether the respondent was white (1) or some other race (0)
age	respondent's age in years
education	respondent's years of education completed
prestige	% of other survey respondents rating this R's job as prestigious

Sexist Attitudes to Working Mothers

In 1977 and 1989, the General Social Survey read a statement to respondents (R's):

"A working mother can establish just as warm and secure of a relationship with her child as a mother who does not work."

R's could strongly disagree (D), disagree (d), agree (a), or strongly agree (A):

Percent who. . .	strongly disagree	disagree	agree	strongly agree
1977 sample (n=1379)	17.1	33.3	33.8	15.7
1989 sample (n=914)	6.7	28.8	42.7	21.9

Research questions:

(1) Do attitudes vary by subgroup – sex, race, age, edu, prestige – all else equal?

(2) Do subgroup differences persist or attenuate over time: did subgroups with more sexist attitudes in 1977 "catch-up" to changing social attitudes by 1989?

(3) To what extent did changes over time result from attitudinal change within subgroups, versus shifts in population composition towards less sexist groups?

Sexist Attitudes: descriptive results

group 1 (group 2)	% agree or str agree, '77 1	(2)	1 − (2)	% agree or str agree, '89 1	(2)	1 − (2)
female (male)	57	41	16	69	59	11
age 28±5 (age 61±5)	63	39	24	74	52	22
nonwhite (white)	56	49	7	71	64	7
college grad (hs grad)	62	50	12	66	62	4
high prestige job (low)	48	47	2	69	65	4

As a first cut, look at subgroup means without even fitting a model:

Some subgroups are consistently less sexist, especially women and the young

All groups became less sexist over time (or at least more circumspect)

Perhaps some "catch-up" by groups the were more sexist in 1977 (high school grads and men)?

But do we trust these results?

Sexist Attitudes: descriptive results

group 1 (group 2)	% agree or str agree, '77 1	(2)	1 − (2)	% agree or str agree, '89 1	(2)	1 − (2)
female (male)	57	41	16	69	59	11
age 28±5 (age 61±5)	63	39	24	74	52	22
nonwhite (white)	56	49	7	71	64	7
college grad (hs grad)	62	50	12	66	62	4
high prestige job (low)	48	47	2	69	65	4

Lots of potential confounding bias and sampling error:

(1) Covariates are correlated by period, so age could absorb education differences, and education could mask race or job effects

(2) Observed differences across periods could result from changing views by group, changing confounders by group, or uniform changes in the population

(3) To estimate age, education, and job differences, we looked at small subgroups $\rightarrow$ greater risk of sampling error

I've omitted the subgroup n's to save space – you should show

Sexist Attitudes: descriptive results

group 1 (group 2)	% agree or str agree, '77 1	(2)	1 − (2)	% agree or str agree, '89 1	(2)	1 − (2)
female (male)	57	41	16	69	59	11
age 28±5 (age 61±5)	63	39	24	74	52	22
nonwhite (white)	56	49	7	71	64	7
college grad (hs grad)	62	50	12	66	62	4
high prestige job (low)	48	47	2	69	65	4

A further problem affecting efficiency:

(4) Before summarizing, we collapsed from 4 to 2 categories,
so we reduced information at the start

To fix these problems, we need a regression model for ordered categorical data

To understand what the model tells us,
we need a way to make estimation results at least this substantively relevant

	1977 sample	1989 sample		1977 sample	1989 sample
male	-0.397	-0.443	job prestige	0.002	0.005
	(0.058)	(0.073)		(0.003)	(0.003)
white	-0.238	-0.206	intercept	1.322	1.984
	(0.091)	(0.108)		(0.179)	(0.229)
age	-0.012	-0.013	τ_2	1.016	1.198
	(0.002)	(0.002)		(0.041)	(0.067)
education	0.045	0.028	τ_3	2.078	2.415
	(0.012)	(0.015)		(0.054)	(0.078)
			log likelihood	-1758.824	-1082.657
			AIC	3501.649	2149.314
			In-sample concordance (null=0.5)	0.634	0.626
			LOO-VC concordance (null=0.5)	0.625	0.609
			N	1379	914

Some questions to test your ability to read ordered probit coefficients

1. If a 1977 subject were in the middle of "agree,"
how many years older would they have to be to reach the middle of "disagree"?

	1977 sample	1989 sample		1977 sample	1989 sample
male	-0.397	-0.443	job prestige	0.002	0.005
	(0.058)	(0.073)		(0.003)	(0.003)
white	-0.238	-0.206	intercept	1.322	1.984
	(0.091)	(0.108)		(0.179)	(0.229)
age	-0.012	-0.013	τ_2	1.016	1.198
	(0.002)	(0.002)		(0.041)	(0.067)
education	0.045	0.028	τ_3	2.078	2.415
	(0.012)	(0.015)		(0.054)	(0.078)
			log likelihood	-1758.824	-1082.657
			AIC	3501.649	2149.314
			In-sample concordance (null=0.5)	0.634	0.626
			LOO-VC concordance (null=0.5)	0.625	0.609
			N	1379	914

Some questions to test your ability to read ordered probit coefficients

2. In either year, is there any covariate that can move someone from the middle of "disagree" to the middle of "agree" by itself?

	1977 sample	1989 sample		1977 sample	1989 sample
male	-0.397	-0.443	job prestige	0.002	0.005
	(0.058)	(0.073)		(0.003)	(0.003)
white	-0.238	-0.206	intercept	1.322	1.984
	(0.091)	(0.108)		(0.179)	(0.229)
age	-0.012	-0.013	τ_2	1.016	1.198
	(0.002)	(0.002)		(0.041)	(0.067)
education	0.045	0.028	τ_3	2.078	2.415
	(0.012)	(0.015)		(0.054)	(0.078)
			log likelihood	-1758.824	-1082.657
			AIC	3501.649	2149.314
			In-sample concordance (null=0.5)	0.634	0.626
			LOO-VC concordance (null=0.5)	0.625	0.609
			N	1379	914

Some questions to test your ability to read ordered probit coefficients

3. The coefficient for *male* is larger in 1989 than in 1977:
is this a bigger effect?

	1977 sample	1989 sample		1977 sample	1989 sample
male	-0.397	-0.443	job prestige	0.002	0.005
	(0.058)	(0.073)		(0.003)	(0.003)
white	-0.238	-0.206	intercept	1.322	1.984
	(0.091)	(0.108)		(0.179)	(0.229)
age	-0.012	-0.013	τ_2	1.016	1.198
	(0.002)	(0.002)		(0.041)	(0.067)
education	0.045	0.028	τ_3	2.078	2.415
	(0.012)	(0.015)		(0.054)	(0.078)
			log likelihood	-1758.824	-1082.657
			AIC	3501.649	2149.314
			In-sample concordance (null=0.5)	0.634	0.626
			LOO-VC concordance (null=0.5)	0.625	0.609
			N	1379	914

Some questions to test your ability to read ordered probit coefficients

4. If I forgot to include the estimated cutpoints (τ's) in the table, could you ever interpret these coefficients beyond their signs?

	1977 sample	1989 sample		1977 sample	1989 sample
male	-0.397	-0.443	job prestige	0.002	0.005
	(0.058)	(0.073)		(0.003)	(0.003)
white	-0.238	-0.206	intercept	1.322	1.984
	(0.091)	(0.108)		(0.179)	(0.229)
age	-0.012	-0.013	τ_2	1.016	1.198
	(0.002)	(0.002)		(0.041)	(0.067)
education	0.045	0.028	τ_3	2.078	2.415
	(0.012)	(0.015)		(0.054)	(0.078)
			log likelihood	-1758.824	-1082.657
			AIC	3501.649	2149.314
			In-sample concordance (null=0.5)	0.634	0.626
			LOO-VC concordance (null=0.5)	0.625	0.609
			N	1379	914

Some questions to test your ability to read ordered probit coefficients

5. What is the correct null hypothesis use in a test of τ_3's significance?

	1977 sample	1989 sample		1977 sample	1989 sample
male	-0.397	-0.443	job prestige	0.002	0.005
	(0.058)	(0.073)		(0.003)	(0.003)
white	-0.238	-0.206	intercept	1.322	1.984
	(0.091)	(0.108)		(0.179)	(0.229)
age	-0.012	-0.013	τ_2	1.016	1.198
	(0.002)	(0.002)		(0.041)	(0.067)
education	0.045	0.028	τ_3	2.078	2.415
	(0.012)	(0.015)		(0.054)	(0.078)
			log likelihood	-1758.824	-1082.657
			AIC	3501.649	2149.314
			In-sample concordance (null=0.5)	0.634	0.626
			LOO-VC concordance (null=0.5)	0.625	0.609
			N	1379	914

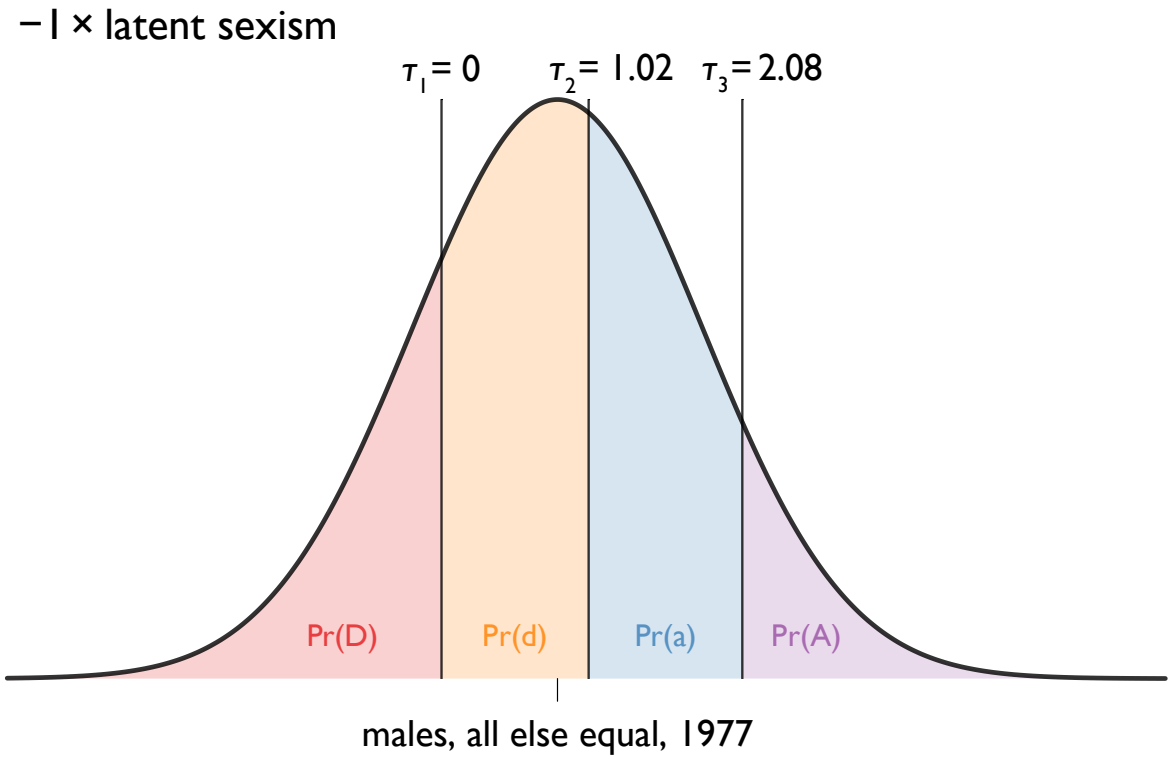
Some questions to test your ability to read ordered probit coefficients

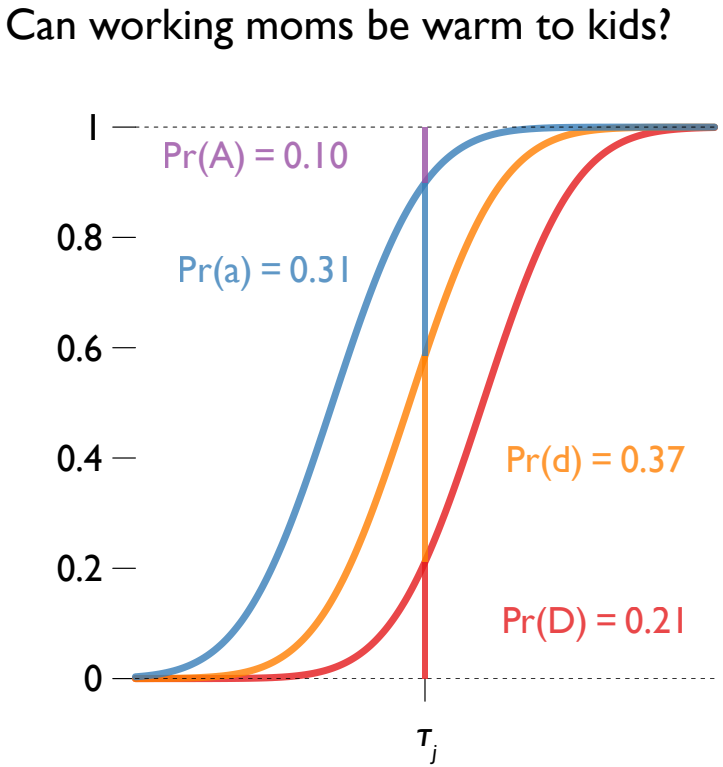
6. How well do the models fit? Can you say if one fits better than the other? What might concordance mean here?

	1977 sample	1989 sample		1977 sample	1989 sample
male	-0.397	-0.443	job prestige	0.002	0.005
	(0.058)	(0.073)		(0.003)	(0.003)
white	-0.238	-0.206	intercept	1.322	1.984
	(0.091)	(0.108)		(0.179)	(0.229)
age	-0.012	-0.013	τ_2	1.016	1.198
	(0.002)	(0.002)		(0.041)	(0.067)
education	0.045	0.028	τ_3	2.078	2.415
	(0.012)	(0.015)		(0.054)	(0.078)
			log likelihood	-1758.824	-1082.657
			AIC	3501.649	2149.314
			In-sample concordance (null=0.5)	0.634	0.626
			LOO-VC concordance (null=0.5)	0.625	0.609
			N	1379	914

Clearly need a better way to display ordered probit than a table of coefficients

For a start, let's go back to PDF and CDF plots



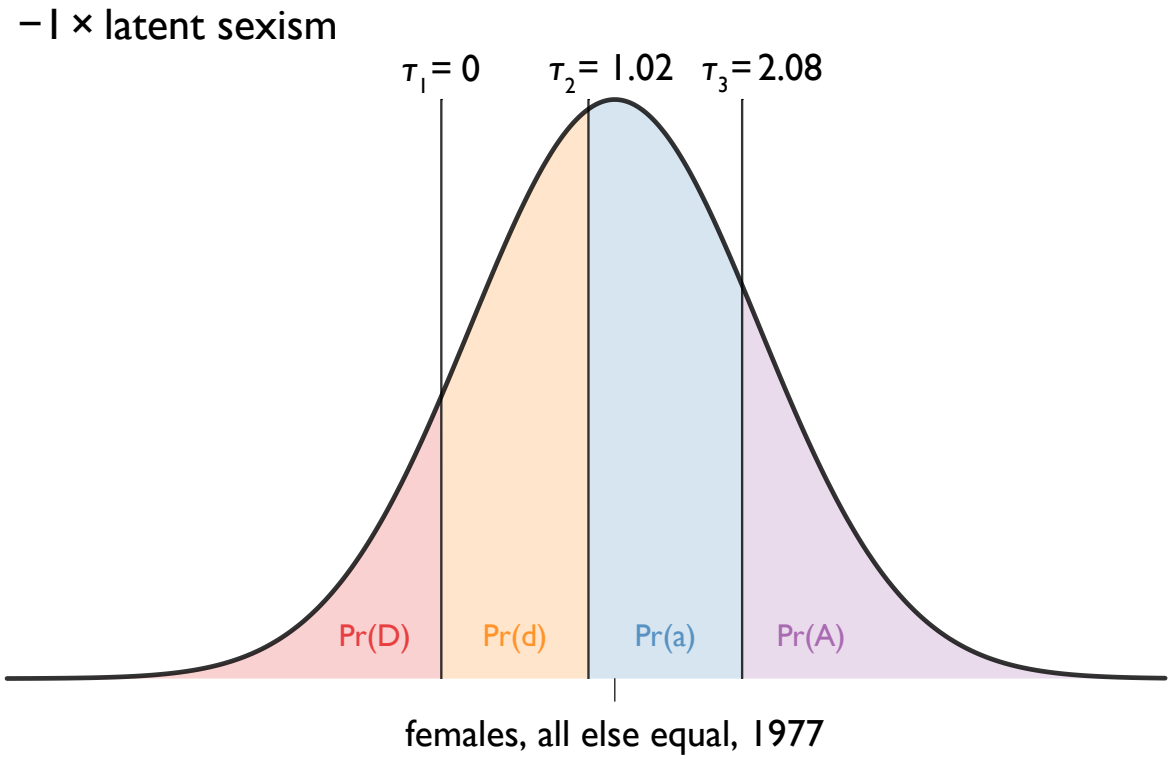


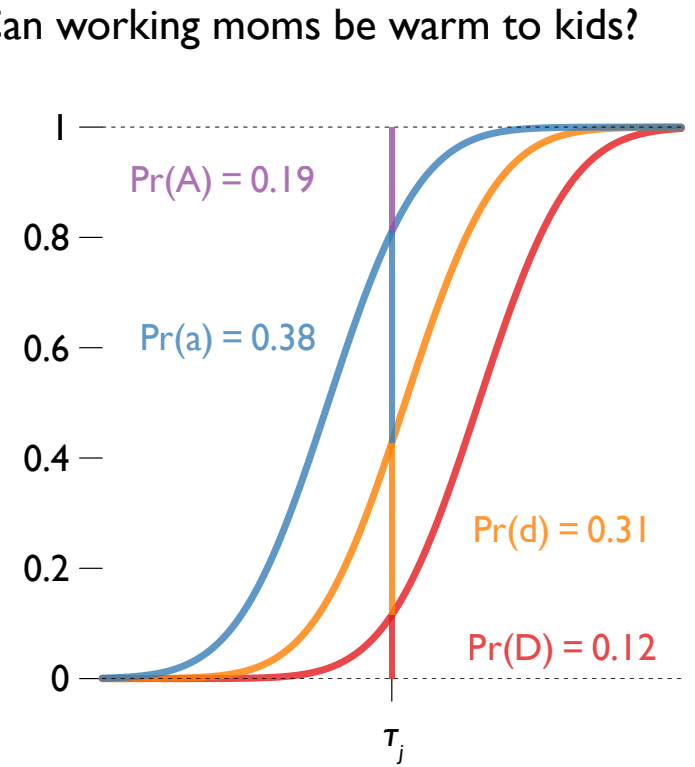
Consider someone in 1977 who is male but otherwise average on all covariates:

$$\begin{aligned} \mathbf{x}_{\text{hyp}}\hat{\boldsymbol{\beta}} &= \hat{\beta}_0 + \sum_{\forall k \neq \text{male}} \bar{x}_k \hat{\beta}_k + \hat{\beta}_{\text{male}} \times 1 \\ &= 1.322 - 0.123 - 0.397 \times 1 = 0.801 \end{aligned}$$

"Median" category is *disagree*, while the CDFs show the probability of each category

Combined probability of either *agree* or *strongly agree* is $31\% + 10\% = 41\%$



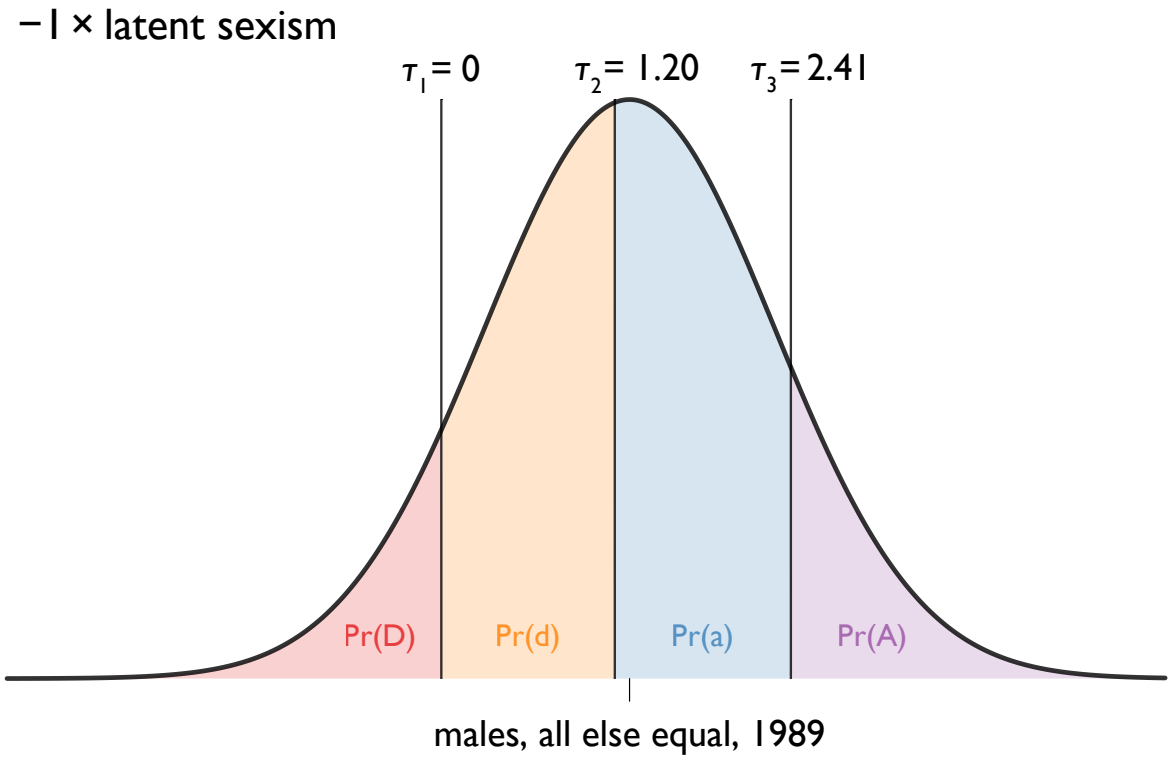


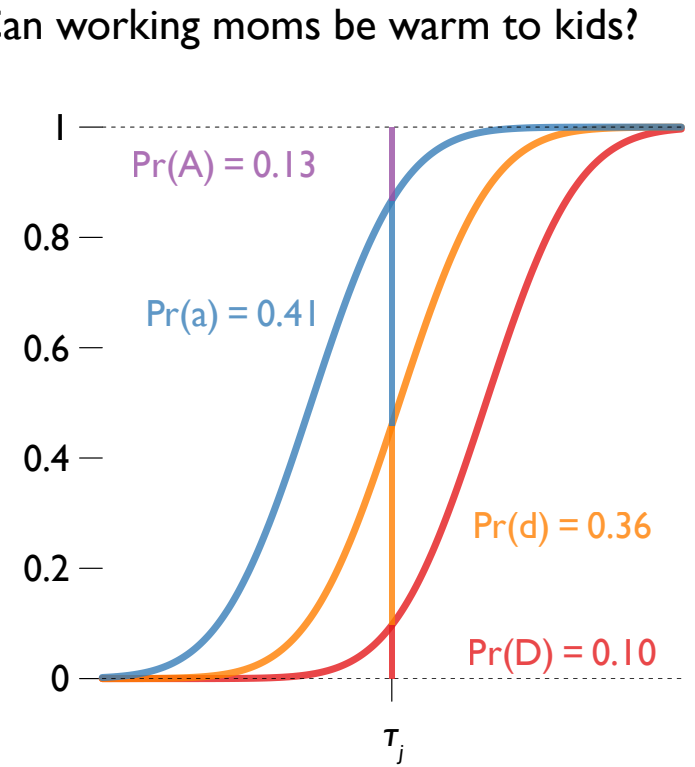
Consider someone in 1977 who is female but otherwise average on all covariates:

$$\begin{aligned} \mathbf{x}_{\text{hyp}}\hat{\boldsymbol{\beta}} &= \hat{\beta}_0 + \sum_{\forall k \neq \text{male}} \bar{x}_k \hat{\beta}_k + \hat{\beta}_{\text{male}} \times 0 \\ &= 1.322 - 0.123 - 0.397 \times 0 = 1.200 \end{aligned}$$

"Median" category is *agree*, while the CDFs show the probability of each category

Combined probability of either *agree* or *strongly agree* is 38% + 19% = 57%



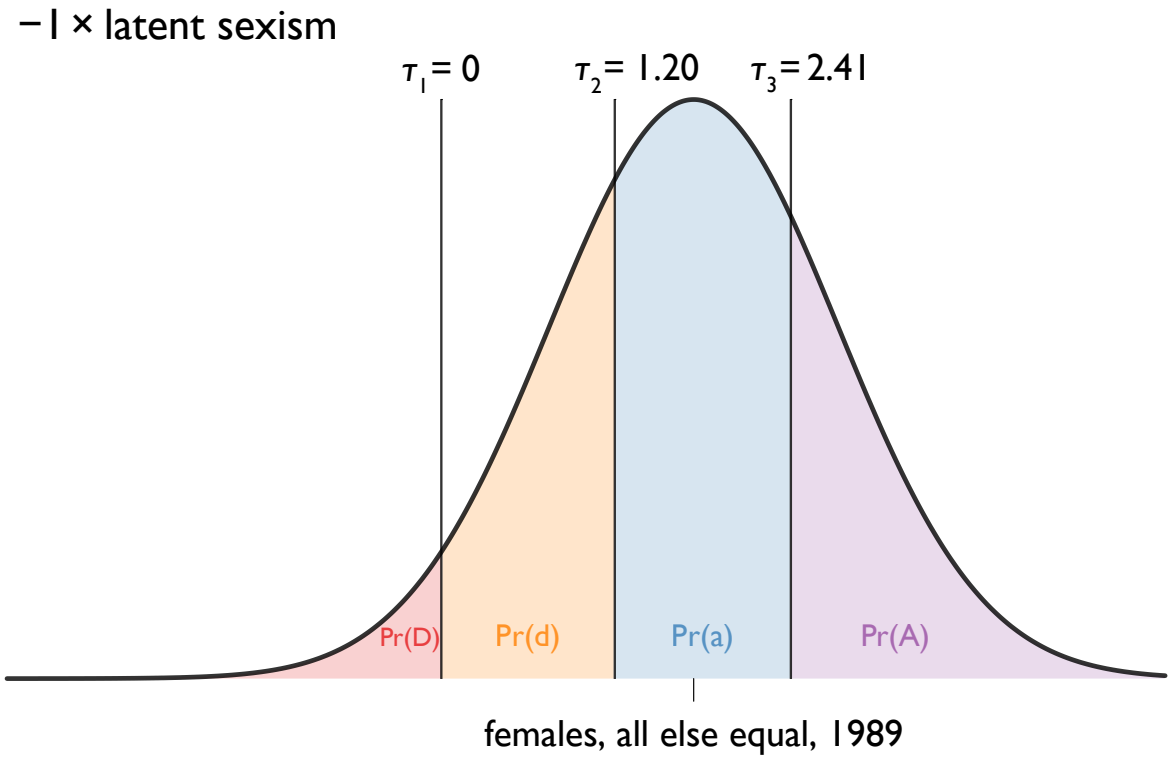


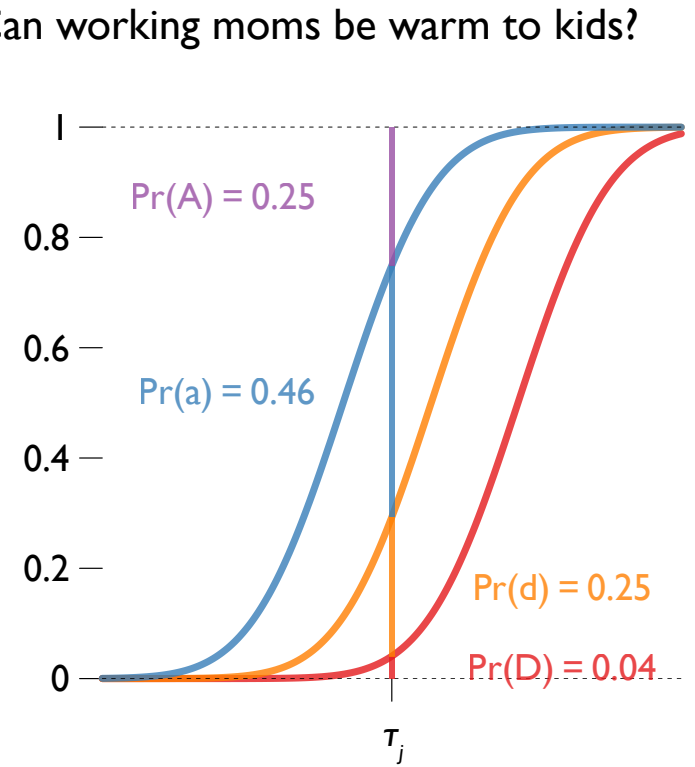
Consider someone in 1989 who is male but otherwise average on all covariates:

$$\begin{aligned} \mathbf{x}_{\text{hyp}}\hat{\boldsymbol{\beta}} &= \hat{\beta}_0 + \sum_{\forall k \neq \text{male}} \bar{x}_k \hat{\beta}_k + \hat{\beta}_{\text{male}} \times 1 \\ &= 1.984 - 0.195 - 0.442 \times 1 = 1.346 \end{aligned}$$

"Median" category is *agree*, while the CDFs show the probability of each category

Combined probability of either *agree* or *strongly agree* is 41% + 13% = 54%



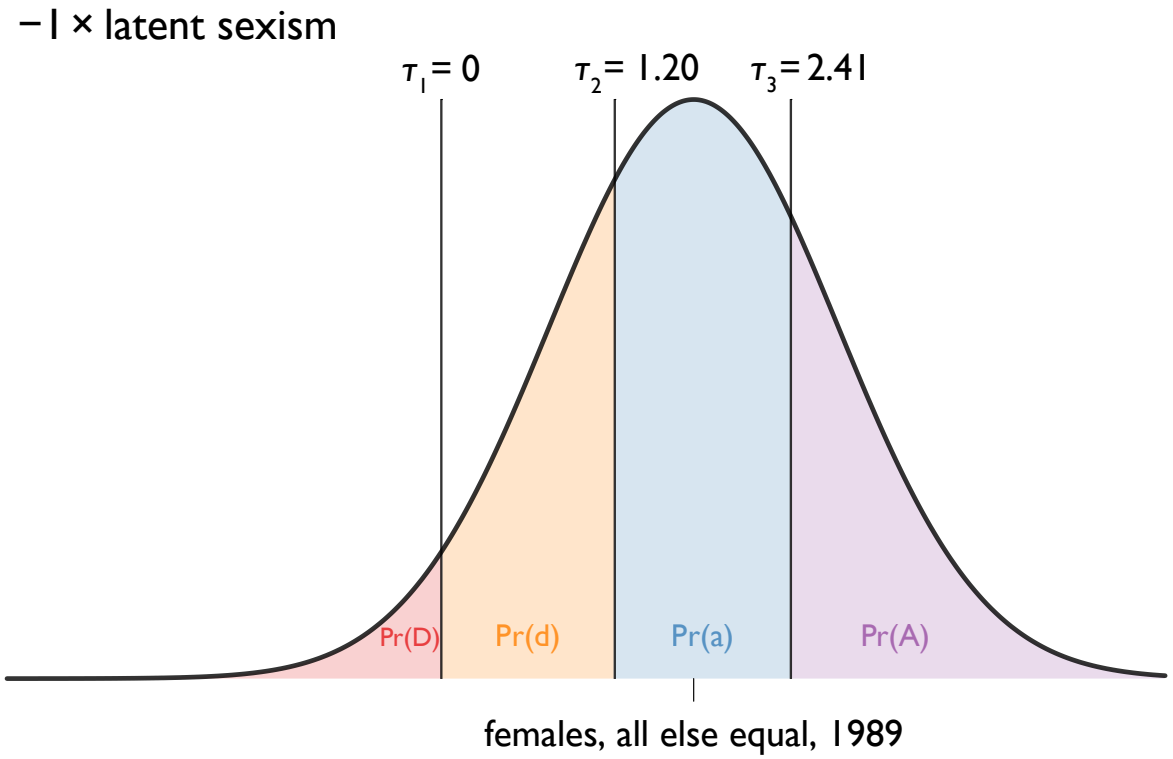


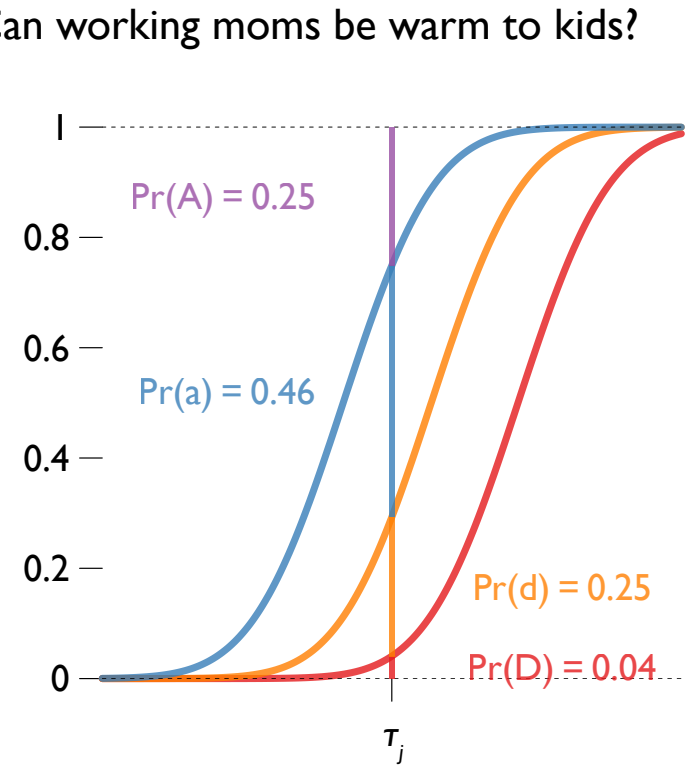
Consider someone in 1989 who is female but otherwise average on all covariates:

$$\begin{aligned} \mathbf{x}_{\text{hyp}}\hat{\boldsymbol{\beta}} &= \hat{\beta}_0 + \sum_{\forall k \neq \text{male}} \bar{x}_k \hat{\beta}_k + \hat{\beta}_{\text{male}} \times 0 \\ &= 1.984 - 0.195 - 0.442 \times 0 = 1.789 \end{aligned}$$

"Median" category is *agree*, while the CDFs show the probability of each category

Combined probability of either *agree* or *strongly agree* is 46% + 25% = 71%



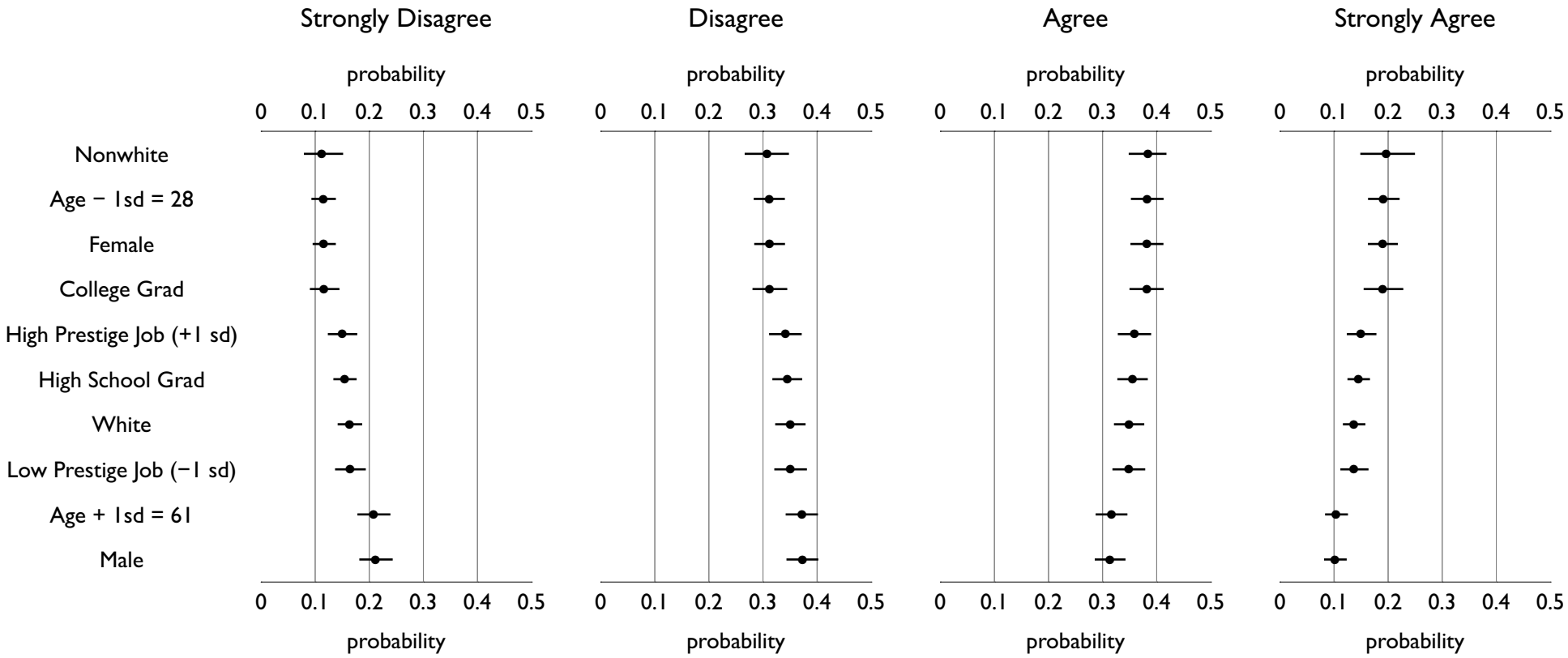


Pedagogically helpful but inefficient – and we have 4 more covariates!

Can we get this all one one slide? With confidence intervals?

Caveat on labeling latent variables

Technically, the latent y^* is just the propensity to agree with the GSS question; any other label one applies is a subjective assessment – I've given mine



Let's apply our usual techniques to the 1977 covariates:
simulate counterfactual predicted probabilities for each category

Simulations above done with `simcf`'s `oprobitsimev()`
and graphed in dotplots with CIs using `tile`'s `ropeladder()`

Flexible way to show a variety of specific counterfactuals at once

But do we really need to show all four categories every time?

Efficient & clear presentation for models of categorical data

With more than three categories, presenting ordered probit results can get tedious

One plot per category seems like a lot to wade through for 5 or 7 categories. . .

Two bad solutions:

1. Use linear regression instead (because it's "easier to interpret")

- assumes interval level outcome, which is unlikely
- what does an expected response of 4.2 on a 7 point scale really mean when 4 is "no opinion" and 5 is "slightly agree"?

2. Collapse categories before estimation (to "make things simpler")

- throws away information, so you results will have larger standard errors

To see why this is bad, think back to latent variables justification of binary logit

Why collapsing categories before estimation is inefficient

Suppose you were interested in risk factors for high blood pressure – arbitrarily defined as systolic $\mathrm{bp} > 140$ – and have measures of exact bp.

You could:

(*i*) dichotomize the data and run a logit with outcome $\mathbb{I}(\mathrm{bp} > 140)$

- throws away most information
- may badly estimate the intensity of effects on bp unless you have lots of data
- Tells you nothing about other thresholds (e.g., > 120)

(*ii*) run a linear regression on bp, then simulate the probability $\widehat{\mathrm{bp}} > 140$

- uses all information
- yields most precise available estimates of quantities of interest, including probability of high bp (the binary outcome)

Why collapsing categories before estimation is inefficient

Suppose we didn't have exact bp,
but did have an ordered measure (low, medium, high, very high)

While not as granular as the latent variable,
this is still more informative than a single dichotomy

Collapsing categories has the same disadvantages as discarding the latent variable

We should keep these categories in the estimation model to get better results

As in the thought experiment above,
we can always predict collapsed categories after efficient estimation
using the full set of categories

A better solution for modeling many ordered categories:

3. Use ordered probit with all categories;
simulate predictions for all categories;
sum up similar categories before computing CIs

Efficient & clear presentation for models of categorical data

Step 1. Estimate the four category ordered probit; retain $\hat{\boldsymbol{\beta}}$ and $\mathrm{Var}(\hat{\boldsymbol{\beta}})$

Step 2. Choose counterfactual scenarios

Step 3. Simulate a predicted probability for each category-scenario combination using draws from $\beta \sim \mathrm{MVN}(\hat{\boldsymbol{\beta}}, \mathrm{Var}(\hat{\boldsymbol{\beta}}))$. Repeat this step 1000 times:

Str. Disagree	Disagree	Agree	Str. Agree
12%	20%	35%	33%
11%	22%	34%	33%
14%	23%	32%	31%
⋮	⋮	⋮	⋮

Step 4: Sum up the simulations by groups of categories

Str. Disagree *or* Disagree	Agree *or* Str. Agree
32%	68%
33%	67%
37%	63%
⋮	⋮

Efficient & clear presentation for models of categorical data

Step 5: Present predicted probabilities, first differences, and relative risks for the collapsed categories:

"In scenario ___, the probability of any agreement is 66% [95% CI: 64%–68%]."

Advantages:

Estimation on the full categories uses all available information (more precise results)

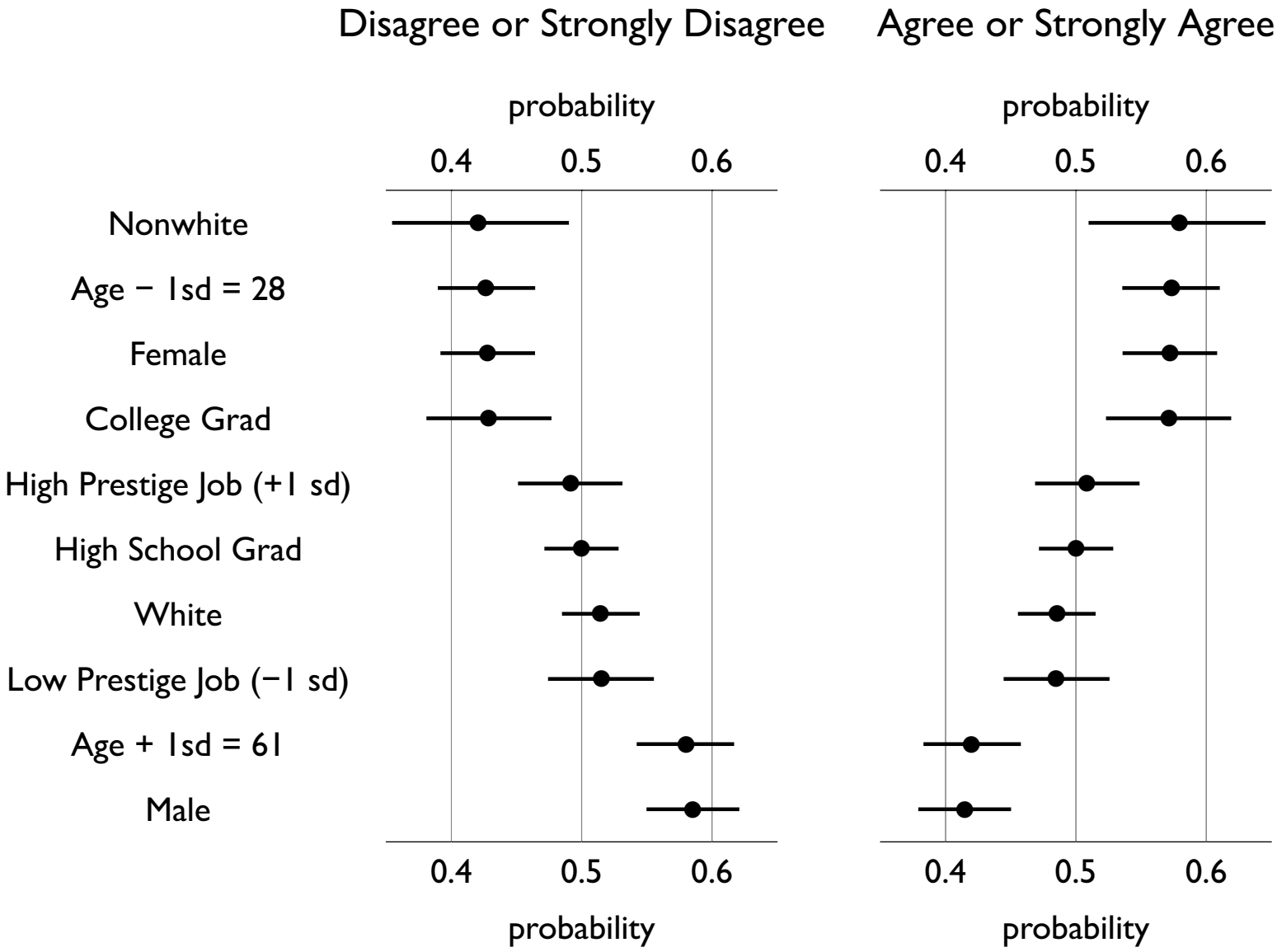
"Simulating out" to the probabilities of collapsed categories yields MLEs for a simpler model summary without making assumptions you don't believe

Results can be as easy to present as a linear regression,
but much more meaningful in terms of the substance of the question

Caveats:

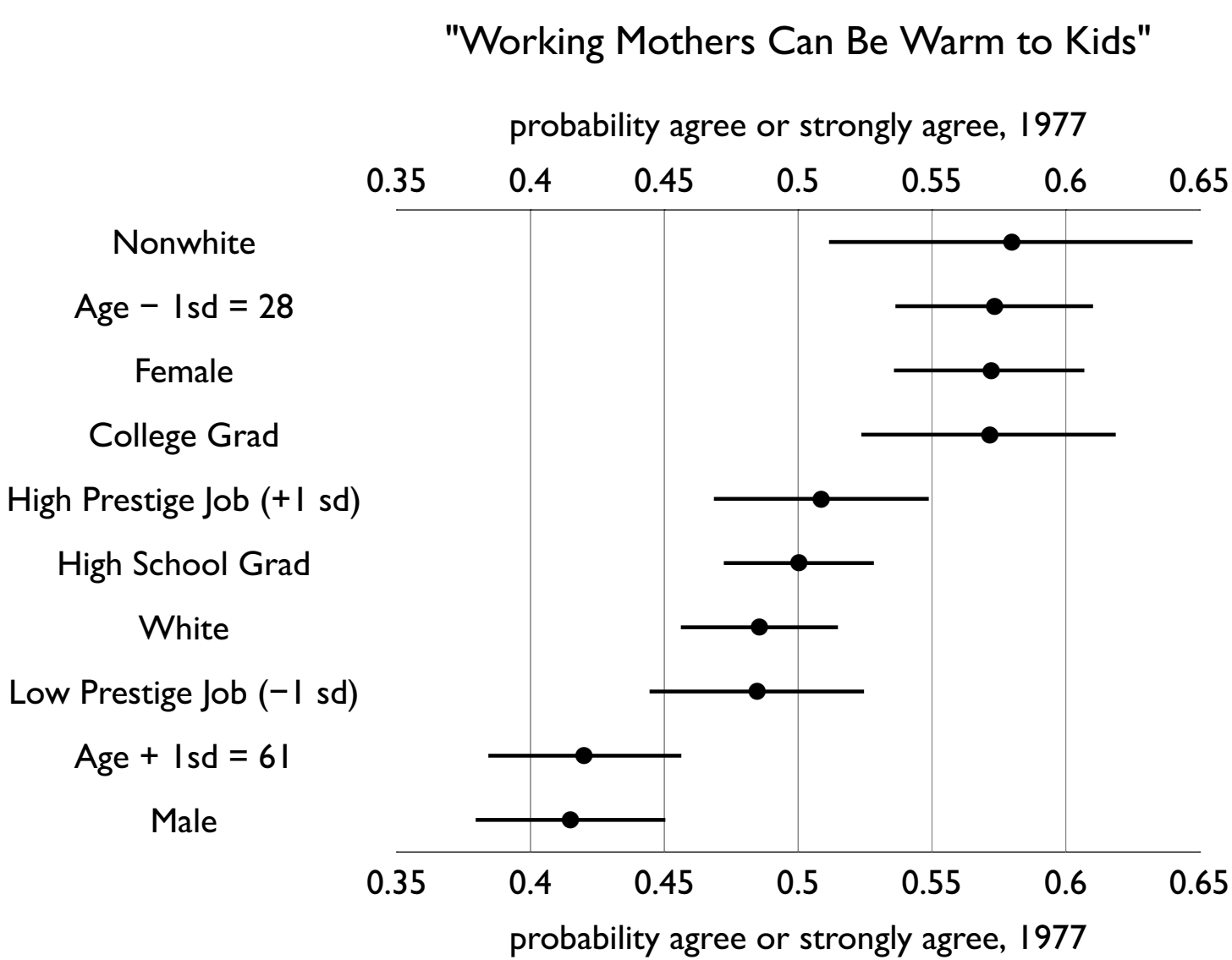
Worth looking at the results for all categories to make sure collapsing makes sense

Be wary of combining "unlike" results and hiding important relationships
(Not likely with ordered probit, could happen with next week)



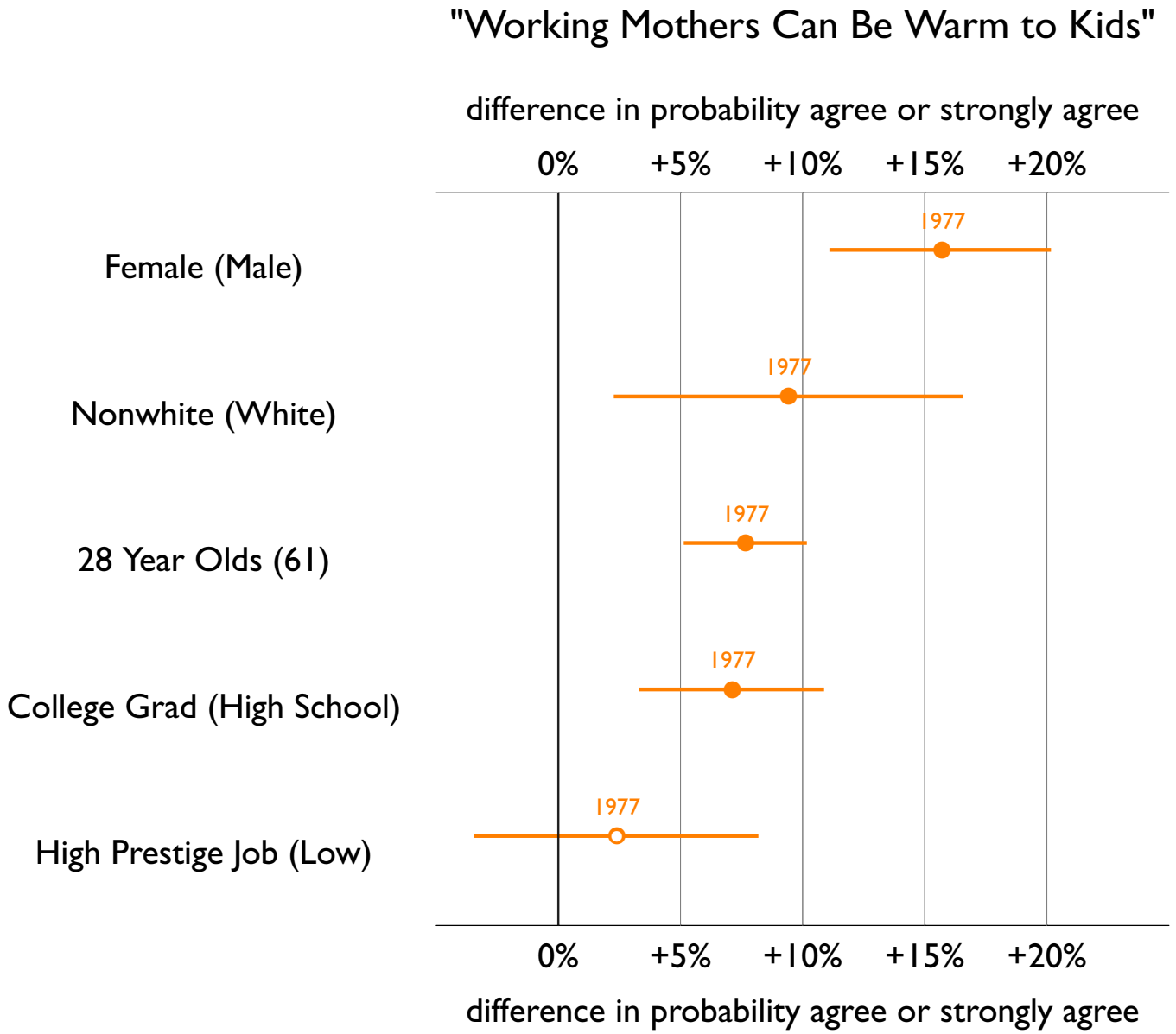
Simulated counterfactual probabilities aggregated after estimation into 2 categories

Note these are mirror images – they must be, as Pr(a) + Pr(A) = 1 - Pr(d) - Pr(D)



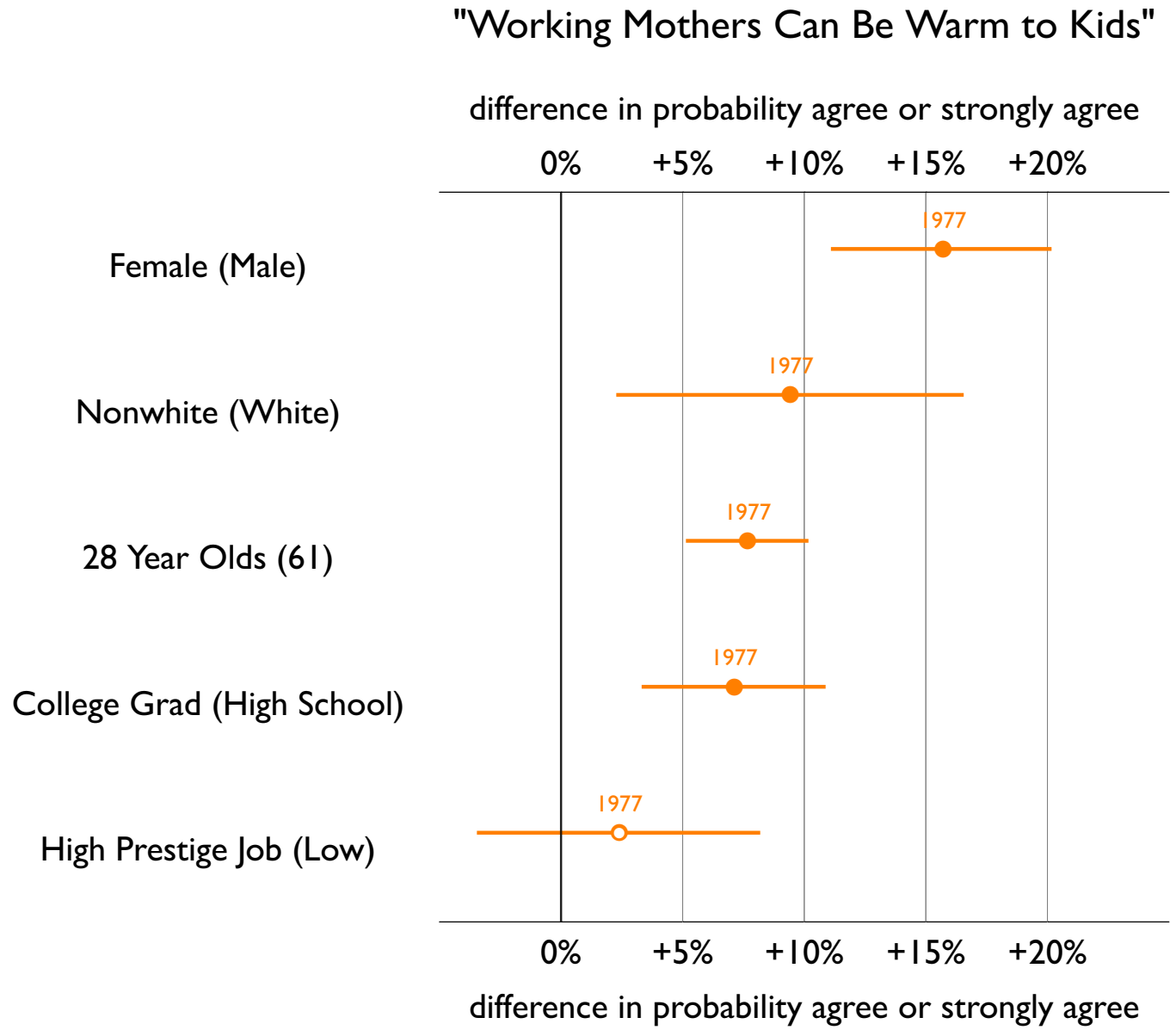
A complete summary of counterfactual expected probabilities for the 1977 model, given the chosen aggregation (judgment call for the analyst)

We used the full precision of the data in estimation,
but simplify presentation to focus on the key substantive findings



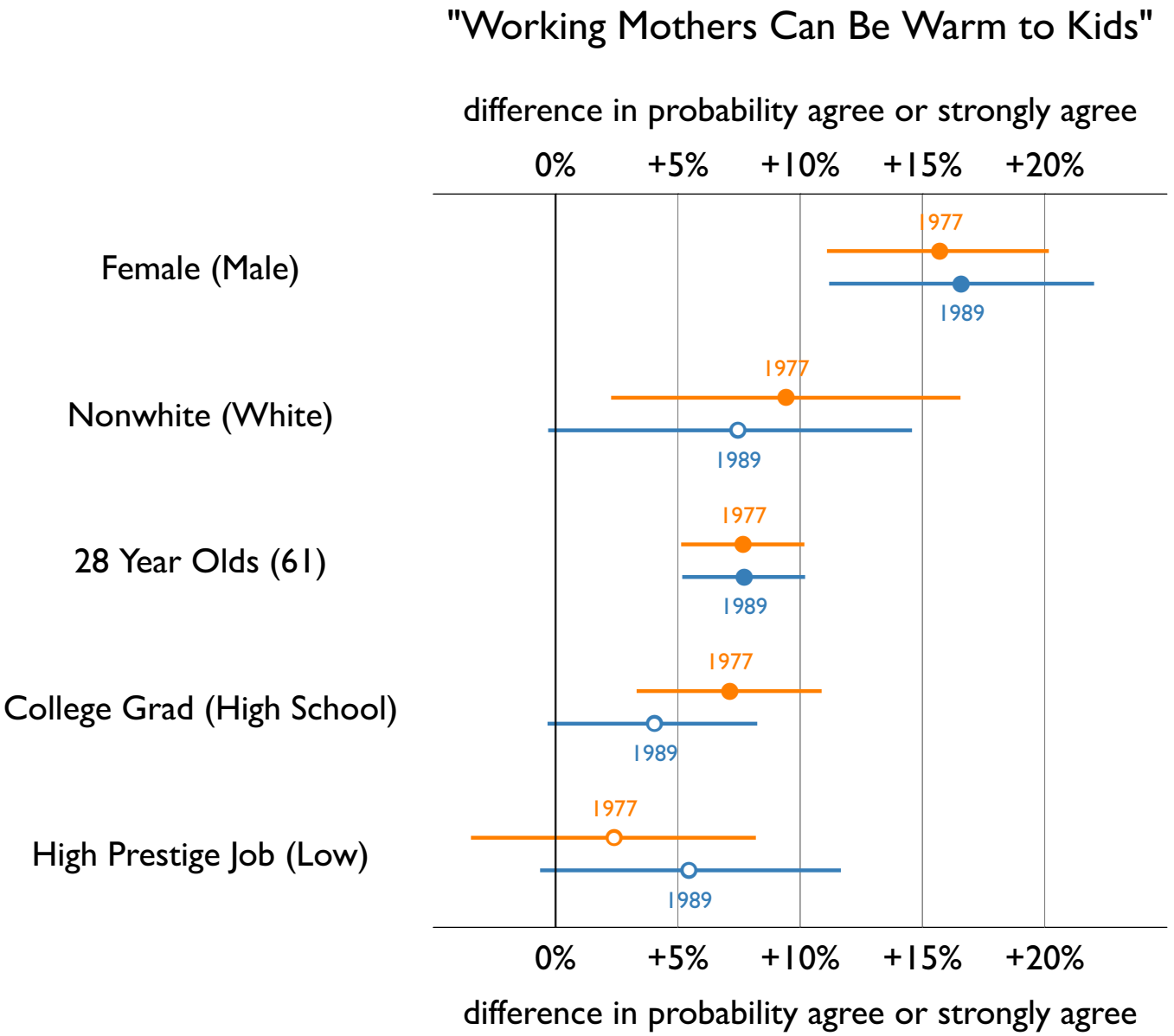
Usually first differences (or relative risks) are more interesting than EVs

Above: the change in Pr(a or A) given a shift from *1* to *(2)*, all else equal



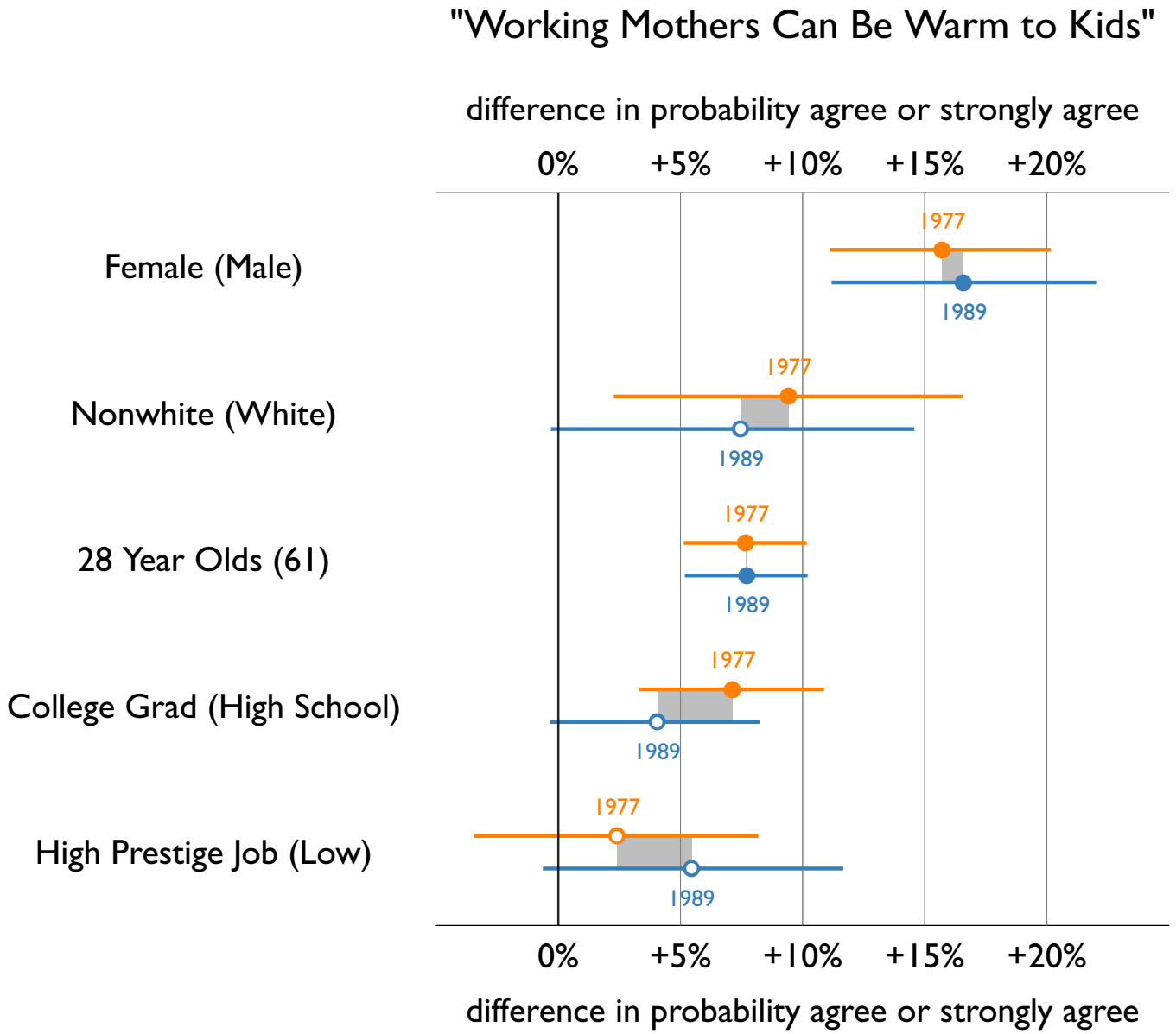
Answers the questions a table of coefficients answers for linear regression

If you can only give one summary of this model, use this one



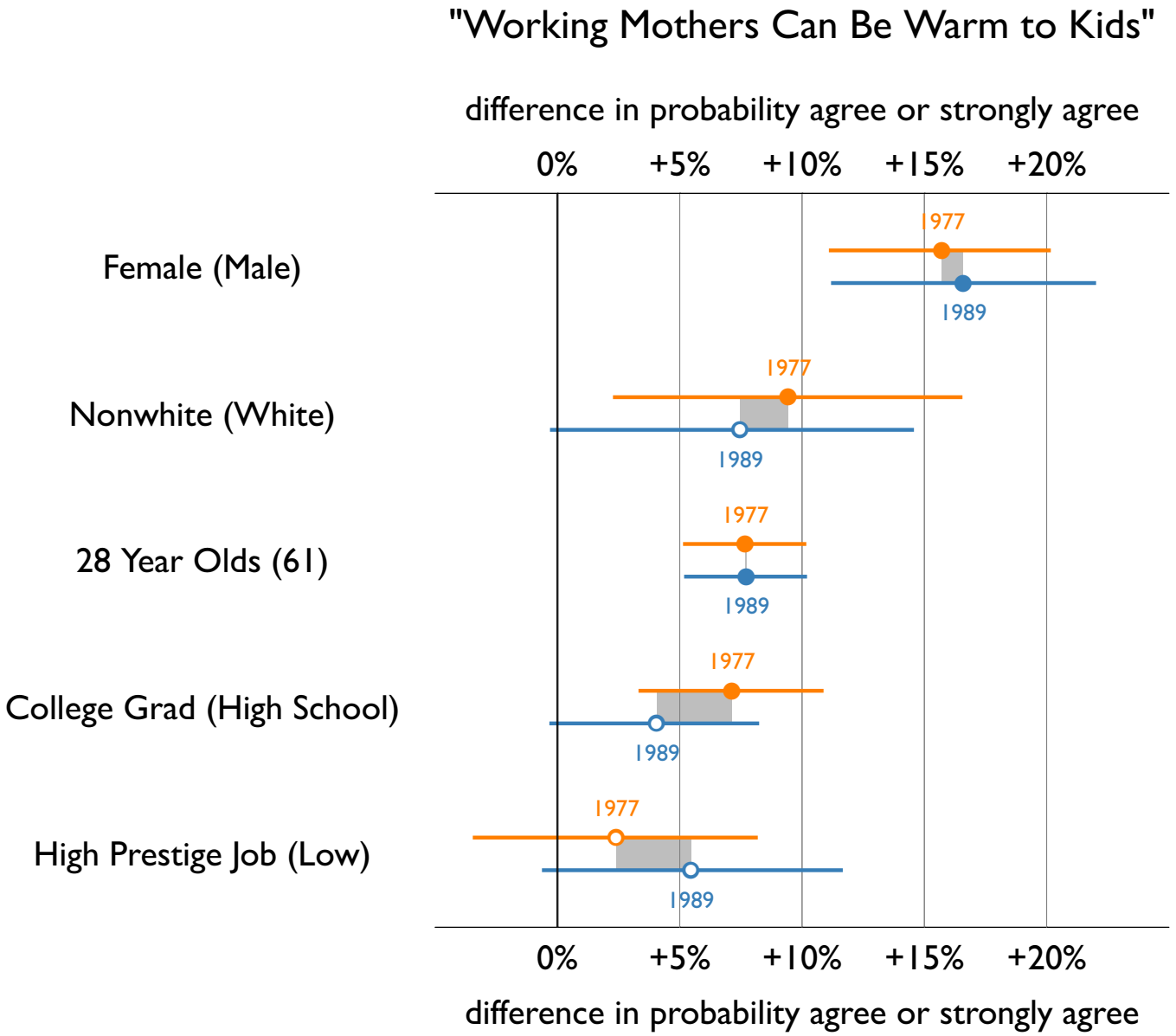
For comparability, all counterfactuals here use 1977 covariate means & sd's

Otherwise we'd conflate changing population composition with changing relationships



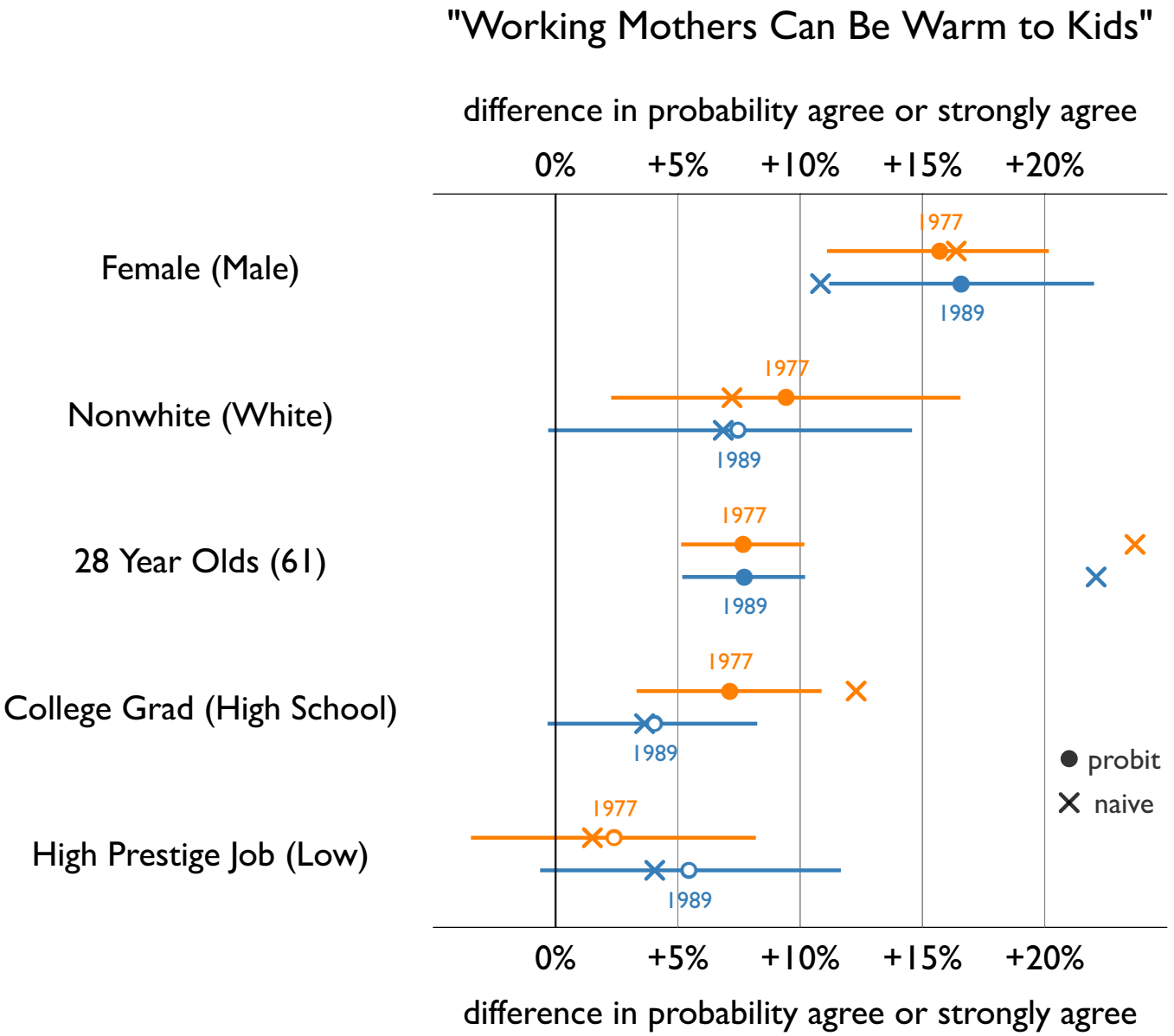
Almost no substantive change – little or no attenuation (resarch question 2)

Aside: Are any interyear differences *statistically* significant?



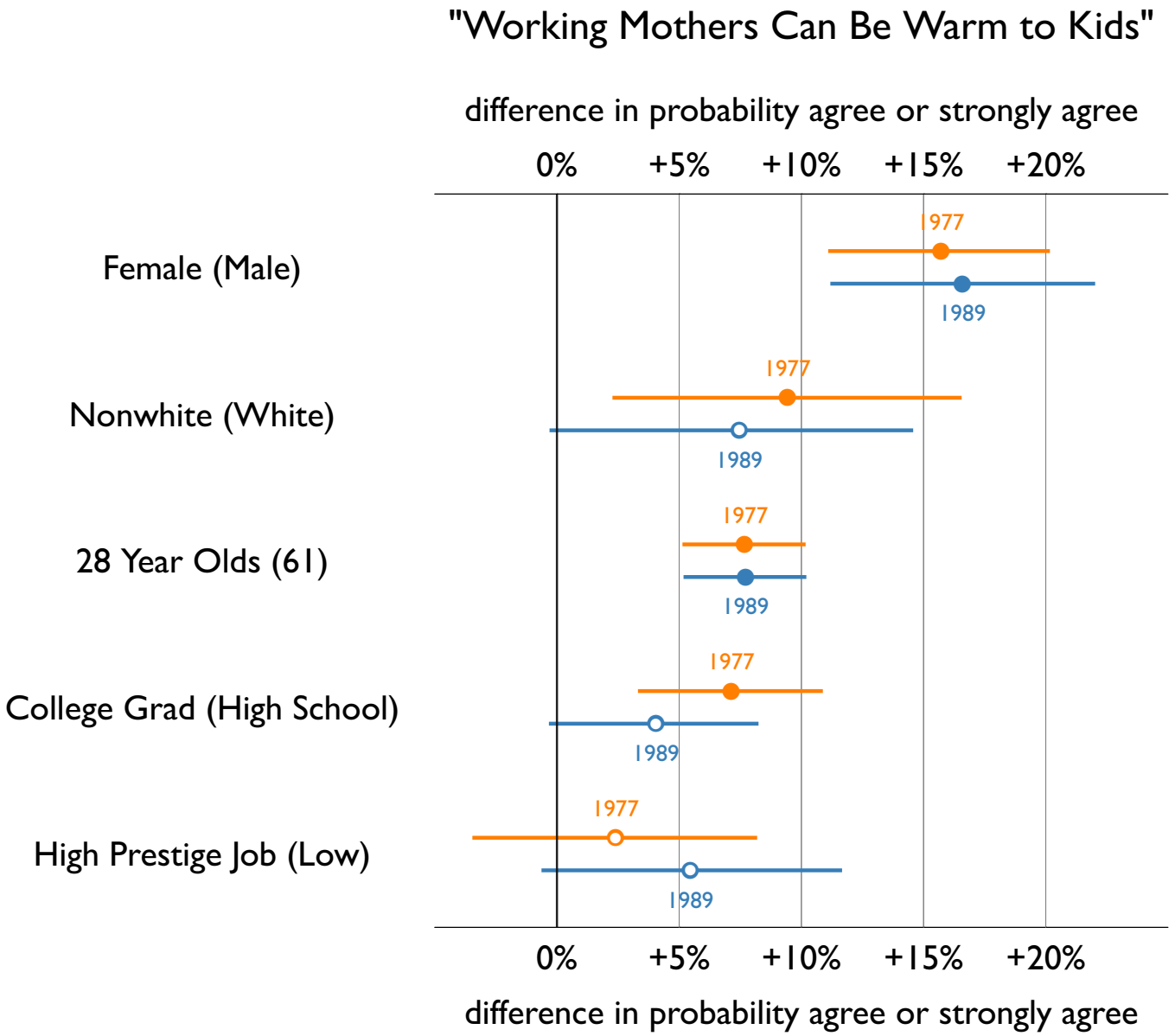
To find out, need a hierarchical model or interaction terms with $\mathbb{I}(\text{year} = 1989)$

(Why? Separate models like ours don't estimate interyear correlations in $\hat{\beta}$)



Good to use the same display for descriptive statistics & model results

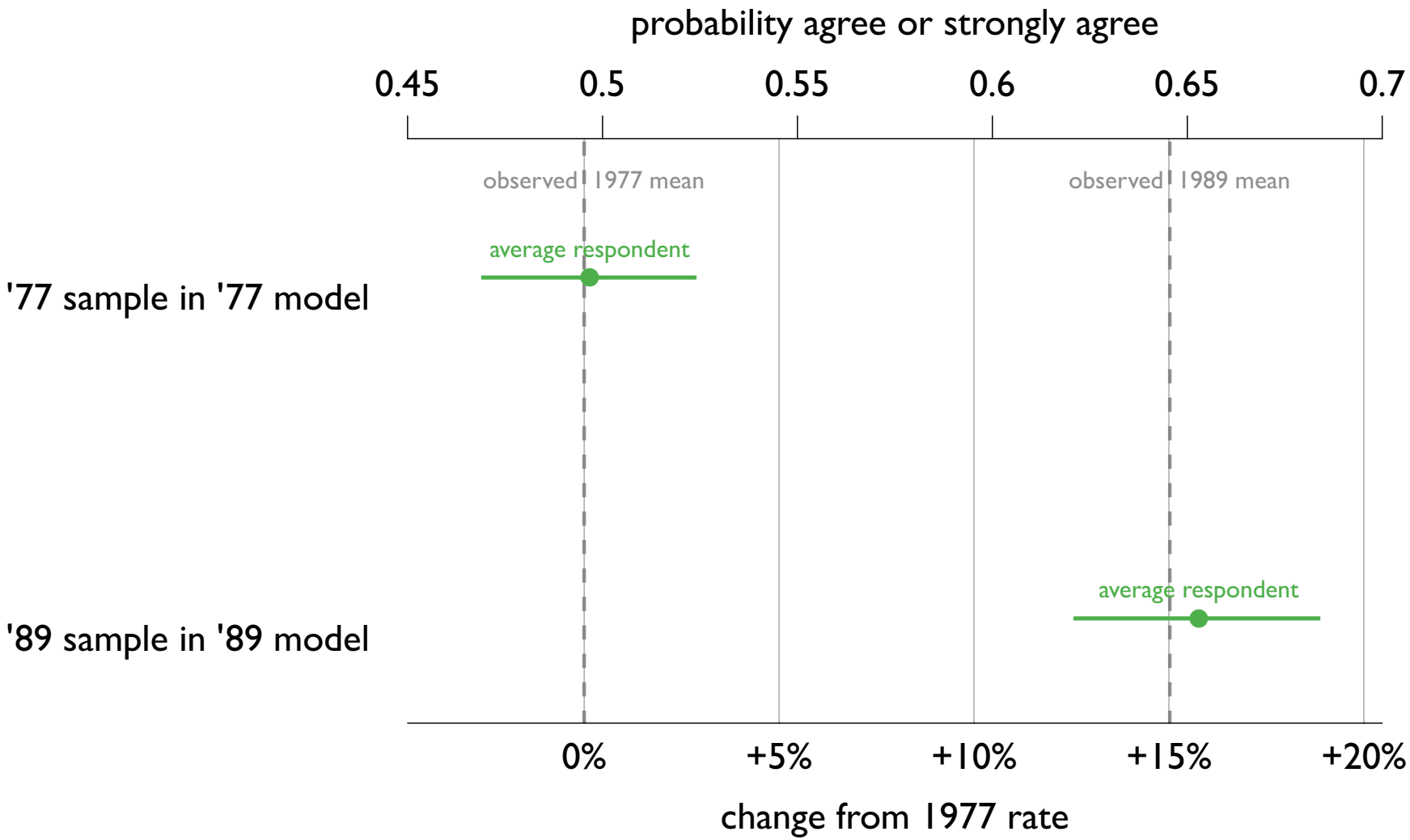
Here, reveals massive confounding bias in descriptive statistics



If the change from 1977 to 1989 isn't attenutation across subgroups, what is it?

Broad decline in sexism? Or growth in population share of less sexist groups?

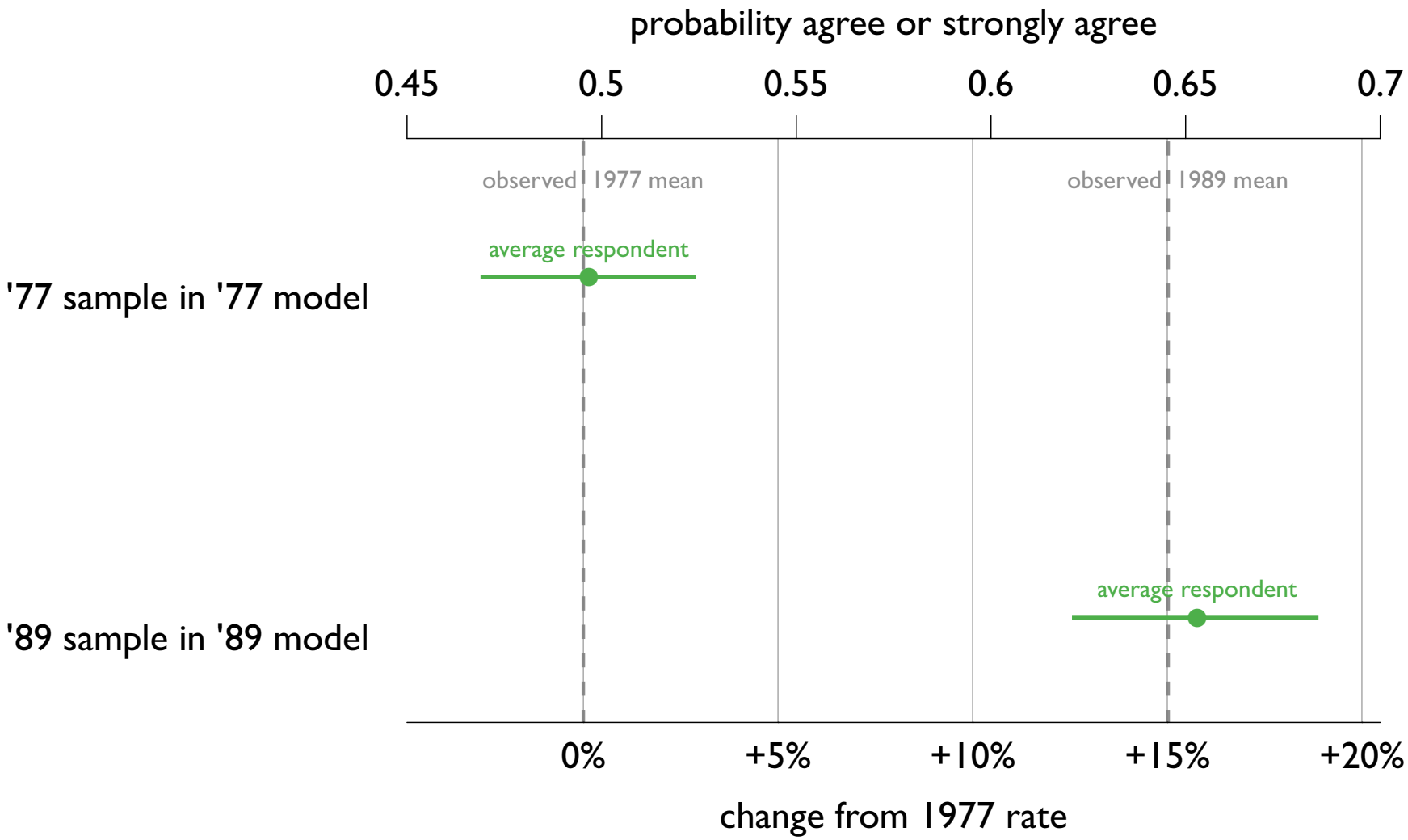
"Working Mothers Can Be Warm to Kids"



To find out: (1) simulate EVs for average person in 1977, using the 1977 model

(2) simulate EVs for the average person in 1989, using the 1989 model

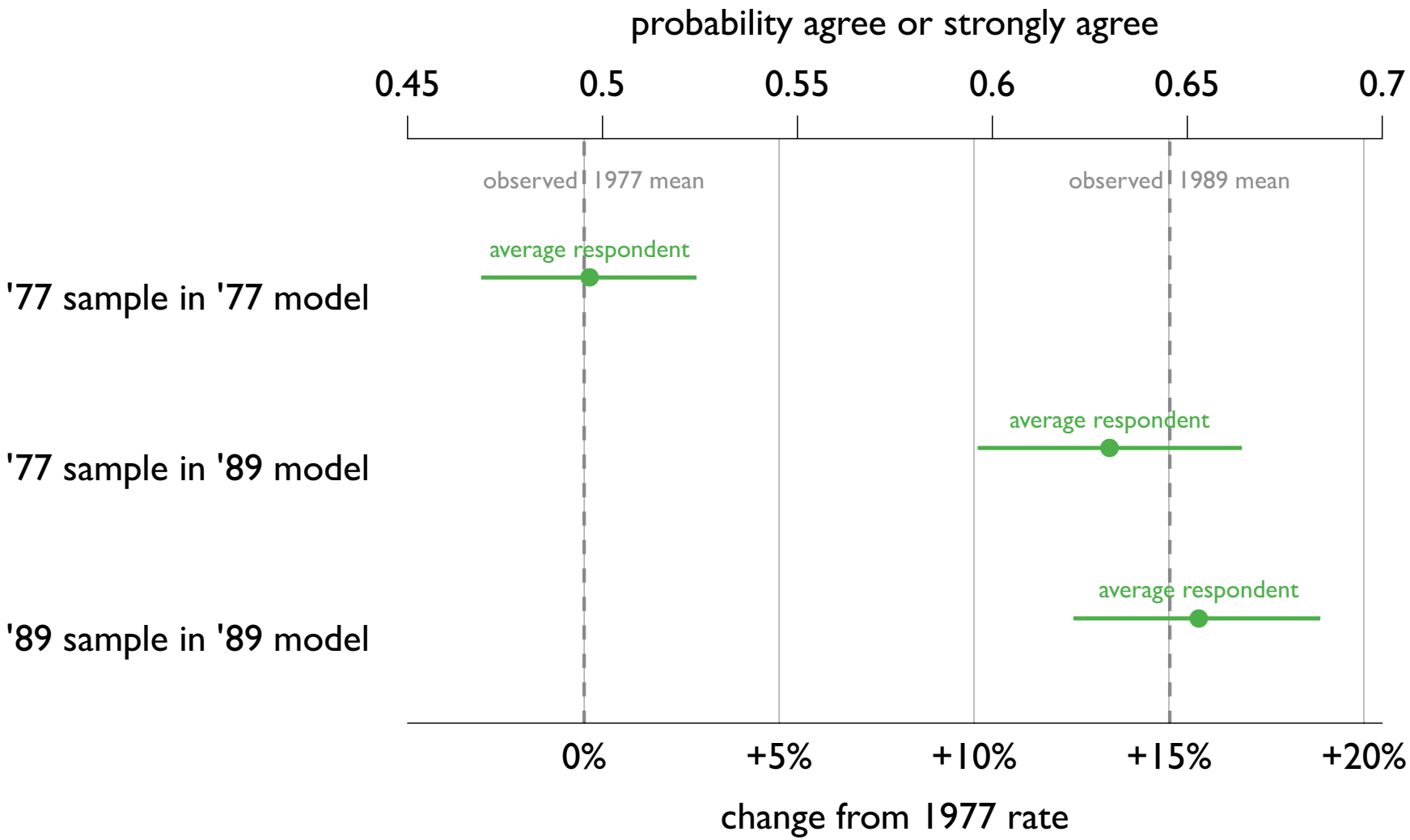
"Working Mothers Can Be Warm to Kids"



Close to the sample means. What if we (3) used 1989 model but 1977 sample?

$\Rightarrow$ show whether changes in population or parameters lay behind changing attitudes

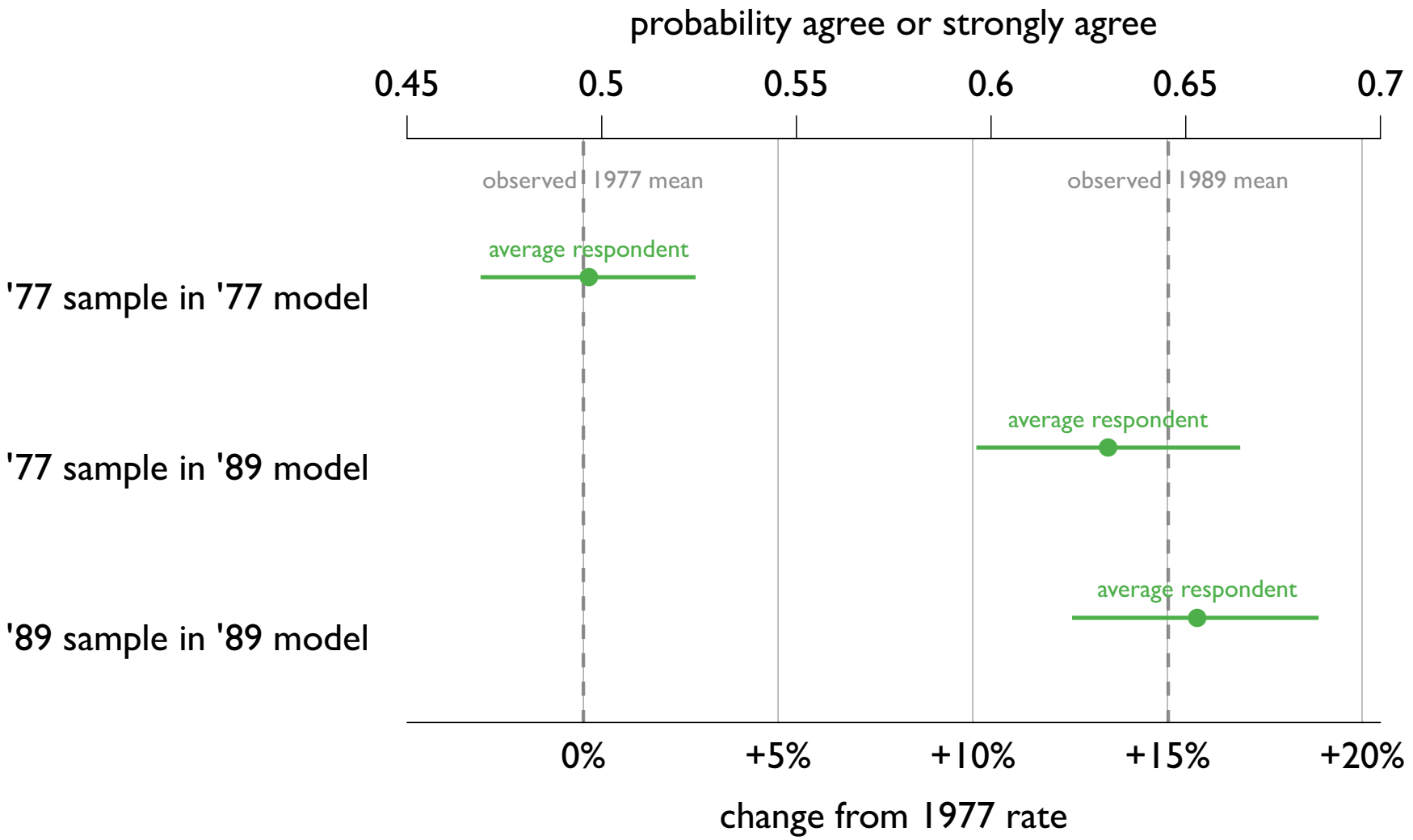
"Working Mothers Can Be Warm to Kids"



Changing model parameters responsible for almost all (85%?) of the difference

Supports broad attitudinal change over subgroup composition explanations

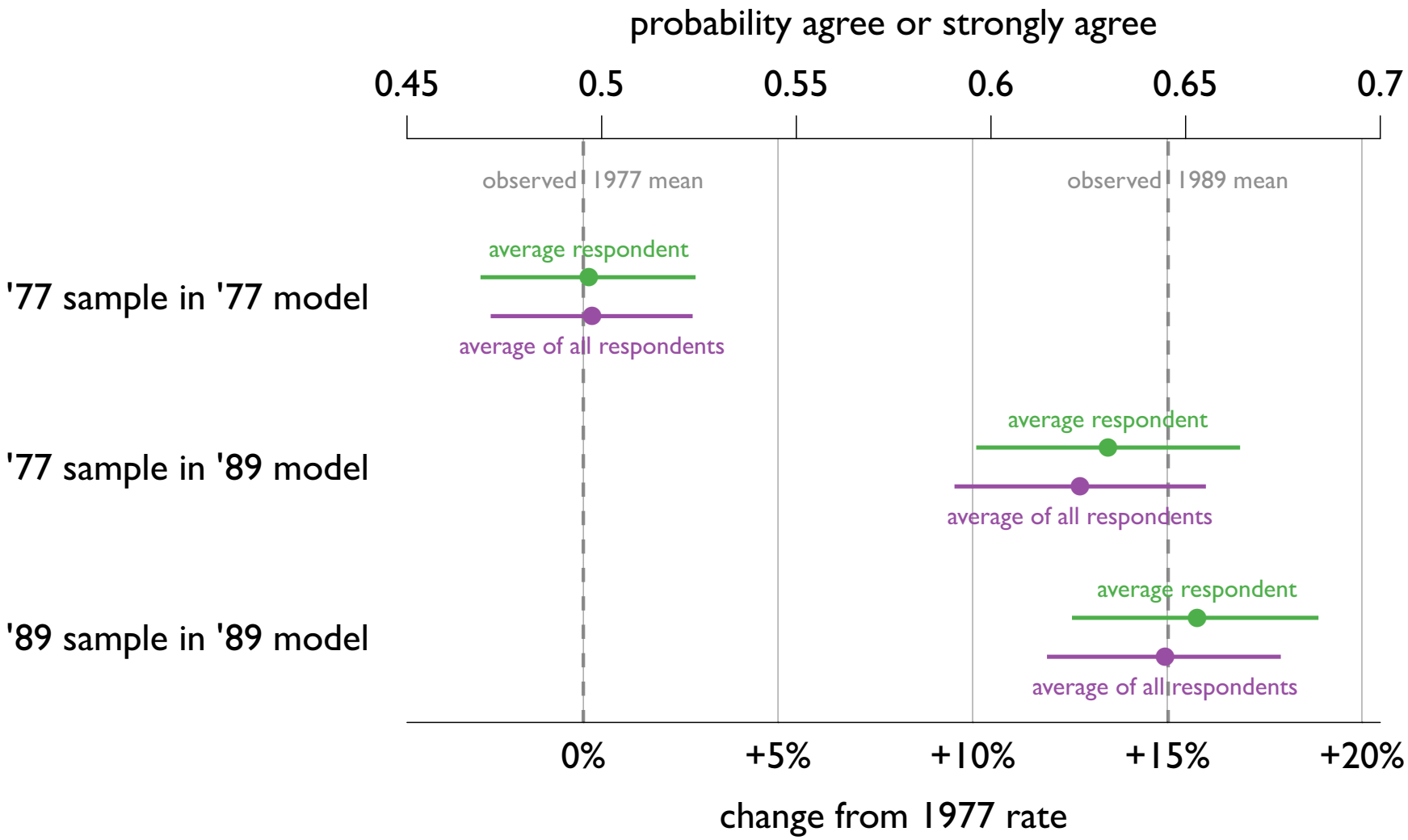
"Working Mothers Can Be Warm to Kids"



One worry: we've used simulations of a hypothetical person with average characteristics to draw conclusions about the whole sample

What if changes throughout the sample are relevant, not just at the mean?

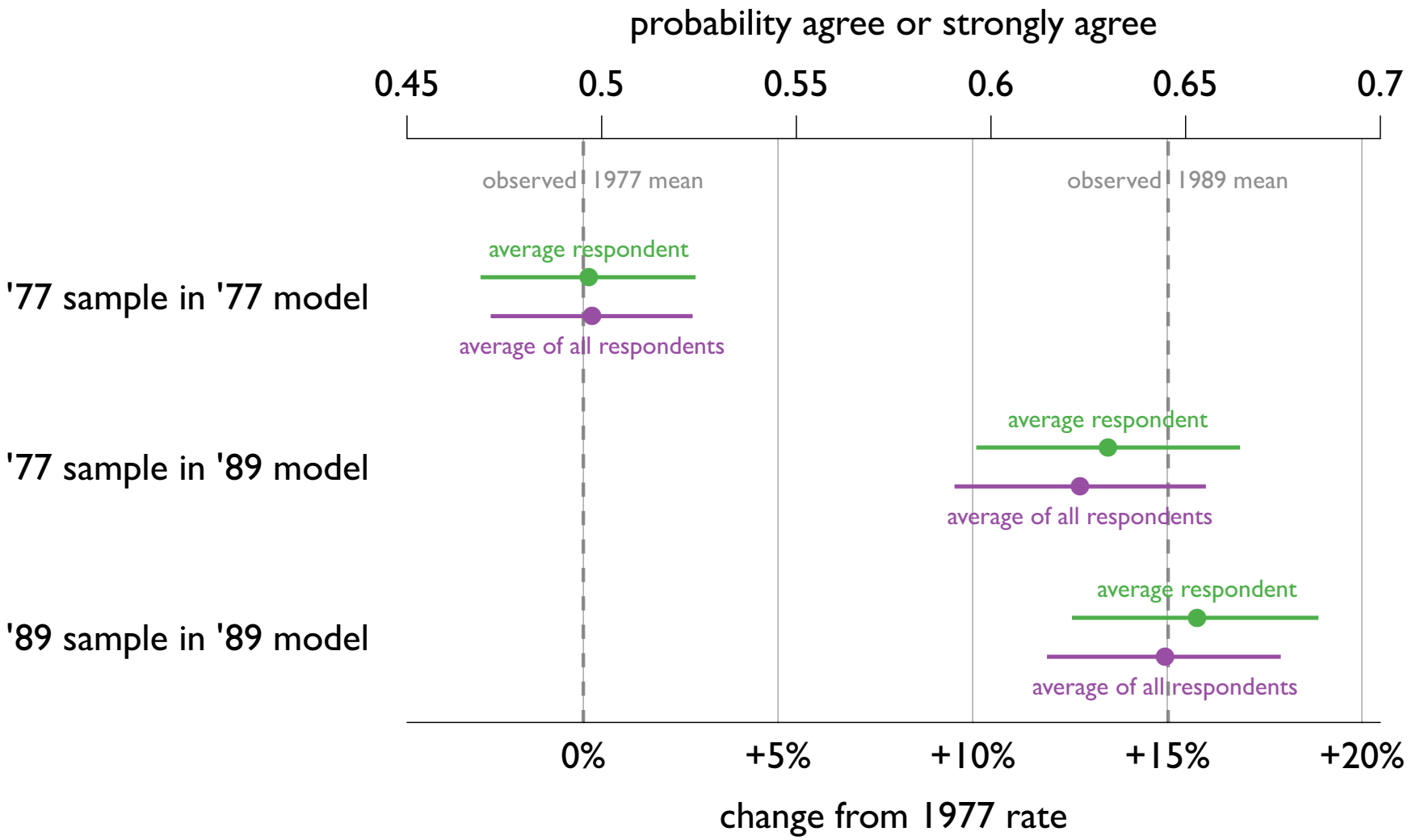
"Working Mothers Can Be Warm to Kids"



Alternative approach: "In-sample simulaton"

simulate every person in the sample and take the mean across ~~counter~~factual cases

"Working Mothers Can Be Warm to Kids"



In-sample simulation makes only a small difference here

More important when specific cases in sample are of direct interest

Concluding thoughts

We've learned

1. How to model ordered categorical data using ordered probit
2. How to interpret ordered probit results through simulation & visuals
3. A bit about model selection for these models

Everything we've done applies straightforwardly to ordered logit

No real difference between these models; coefficients just on different scales

Caveats

With a large number of categories, ordered probit may not be worth the effort compared to linear regression – but confirm roughly equal spacing!

If your outcome variable is the *sum* of ordered scales, ordered probit doesn't help: DGP in this case is quite murky, as steps on the scale are hard to characterize

If you can't order your outcome at all, ordered probit doesn't help –
you'll need a *multinomial* method (next week)