lecture 04

X

adding covariates to our models; logit, Poisson, negative binomial; quantities of interest

Paper and Workshop

review

an example exam question

Show that the scalar representation

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

is equivalent to the matrix representation $\mu_i = X_i\beta$.

Further, show that stacking the $\mu_1, \mu_2, ..., \mu_N$ into a column vector μ is equivalent to $\mu = X\beta$.

Define *maximum likelihood estimate*, *(observed) Fisher information*, *invariance property*, and *delta method*. Explain how we use each and how they all fit together into a workflow.

```
Rows: 487
Columns: 8
$ country           <chr> "Argentina", "Argentina", "Argentina", "Argentina", "…
$ year              <dbl> 1946, 1951, 1954, 1958, 1960, 1963, 1965, 1973, 1983,…
$ average_magnitude <dbl> 10.53, 10.53, 4.56, 8.13, 4.17, 8.35, 4.17, 10.13, 10…
$ eneg              <dbl> 1.342102, 1.342102, 1.342102, 1.342102, 1.342102, 1.3…
$ enep              <dbl> 5.750, 1.970, 1.930, 2.885, 5.485, 5.980, 5.155, 3.19…
$ upper_tier        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ en_pres           <dbl> 2.09, 1.96, 1.96, 2.65, 2.65, 3.90, 3.90, 2.66, 2.30,…
$ proximity         <dbl> 1.00, 1.00, 0.20, 1.00, 0.20, 1.00, 0.33, 1.00, 1.00,…
```

Suppose I have the data set above loaded into R. What **formula** do I need to replicate Clark and Golder's model specification below?

Table 2
The Strategic Modifying Effect of Electoral Laws

	Dependent Variable: Effective Number of Electoral Parties											
	Cross-Sectional Analysis								Pooled Analysis			
Regressor	1980s Amorim Neto & Cox Data[a]		1980s Amorim Neto & Cox Data[a]		1990s Whole Sample		1990s Established Democracies[b]		1946 to 2000 Whole Sample		1946 to 2000 Established Democracies[b]	
Ethnic			−0.05	(0.28)	0.06	(0.37)	−0.70	(0.68)	0.19	(0.13)	0.11	(0.14)
ln(Magnitude)			−0.08	(0.30)	0.51	(0.44)	−0.61	(0.59)	0.33*	(0.20)	0.08	(0.23)
UppertierSeats	0.04**	(0.01)	−0.07	(0.04)	0.01	(0.02)	−0.02	(0.06)	0.05***	(0.02)	−0.06*	(0.03)
PresidentCandidates			0.22	(0.27)	0.36	(0.26)	0.07	(0.22)	0.35**	(0.16)	0.26*	(0.15)
Proximity	−6.05***	(0.88)	−5.88***	(0.84)	−4.19***	(1.26)	−4.95***	(1.24)	−3.42***	(0.55)	−3.10***	(0.46)
Ethnic × ln(Magnitude)	0.39***	(0.07)	0.37*	(0.20)	−0.09	(0.17)	0.63*	(0.34)	0.08	(0.12)	0.26	(0.17)
Ethnic × UppertierSeats			0.07***	(0.02)	−0.005	(0.01)	0.01	(0.04)	−0.02**	(0.01)	0.06***	(0.02)
PresidentCandidates × Proximity	2.09***	(0.26)	1.84***	(0.43)	0.99**	(0.46)	1.42***	(0.44)	0.80***	(0.23)	0.68***	(0.23)
Constant	2.40***	(0.21)	2.60***	(0.51)	4.08***	(0.95)	5.15***	(1.32)	2.81***	(0.34)	2.92***	(0.35)
Observations	51		51		62		39		555		487	
R^2	.71		.77		.29		.48		.30		.40	

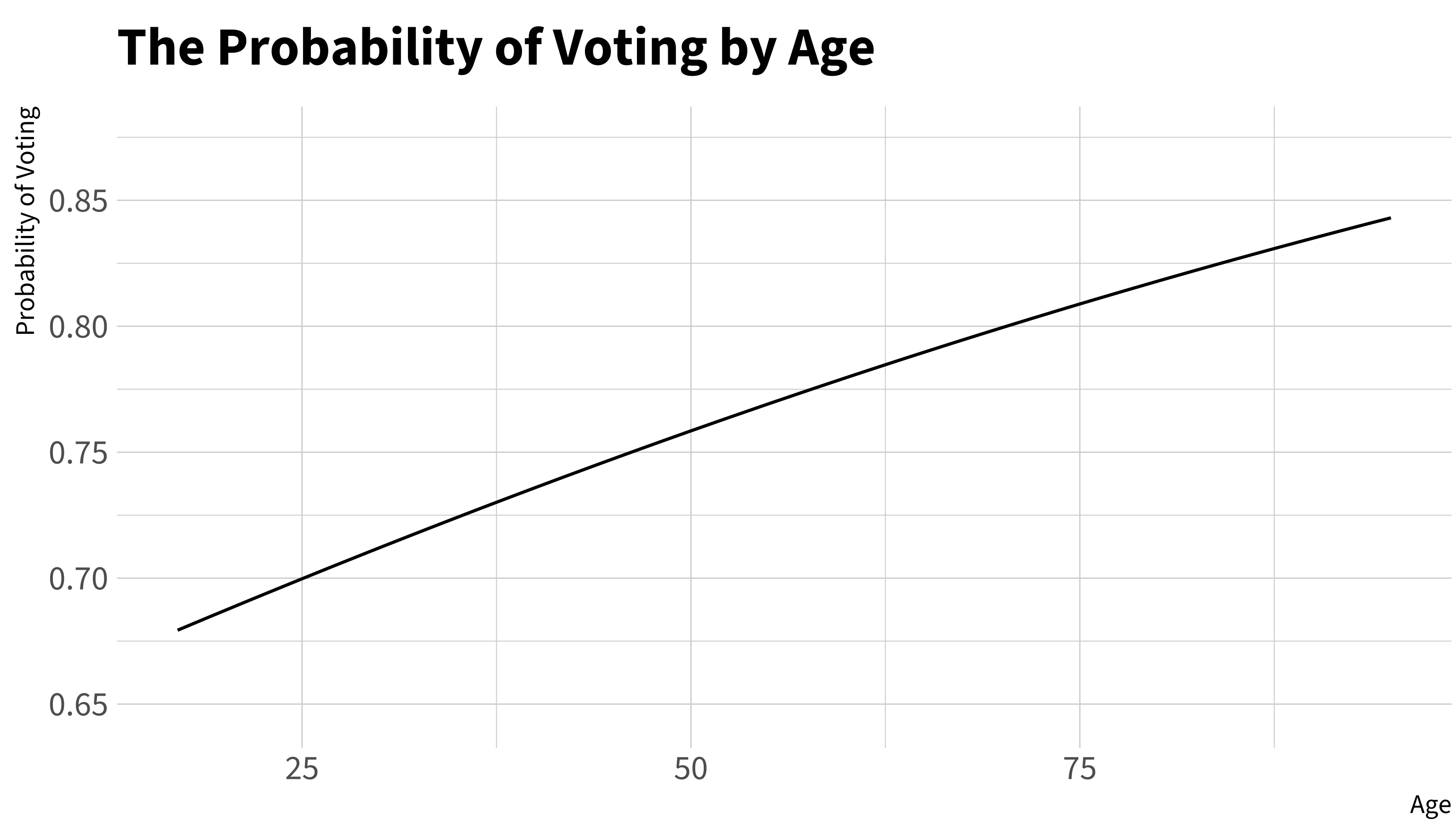
Note: Standard errors are given in parentheses for cross-sectional models; robust standard errors clustered by country are used for the pooled models.
a. See Amorim Neto and Cox (1997).
b. Established Democracies omits elections from countries that transitioned to democracy after 1989.
$^{*}p < .10$. $^{**}p < .05$. $^{***}p < .01$.

a Bernoulli model

with covariates

$$y_i \sim \text{Bernoulli}(\pi)$$

But what if we want π to depend on some **explanatory variables**?



We want something *like*:

$$\pi_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

If we did this, how would it work?

Linear Probability Model

easy to estimate

$$(X^\top X)^{-1} X^\top y$$

easy to interpret

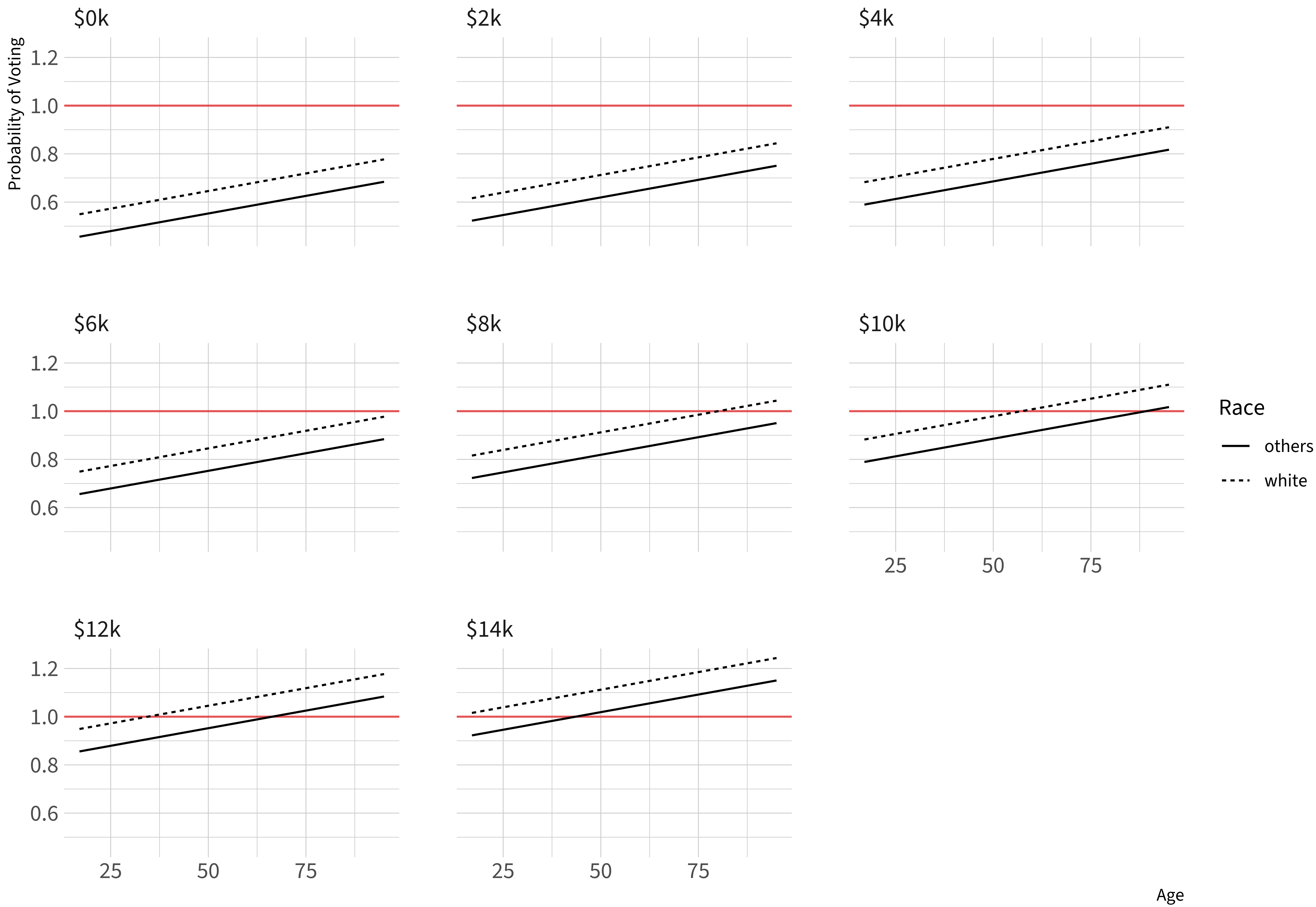
$$\frac{\partial \pi_i}{\partial x_j} = \beta_j$$

$\boxed{\text{Don't sleep on these!}}$

But there are

disadvantages.

The Probability of Voting by Age, Income, and Race



problem 1:

probabilities are not bounded between 0 and 1.

Note: Let's not overstate how problematic this problem is.

If y_i is binary, then $\mathrm{Var}(y_i) = \Pr(y_i)[1 - \Pr(y_i)]$, which, for the LPM, equals $X_i\beta(1 - X_i\beta)$.

problem 2:
non-constant variance

Note: Again, let's not overstate how problematic this problem is.

If y_i is binary, then the residual can take on only two values: $-Pr(y_i)$ or $1 - Pr(y_i)$.

problem 3:
non-normal errors

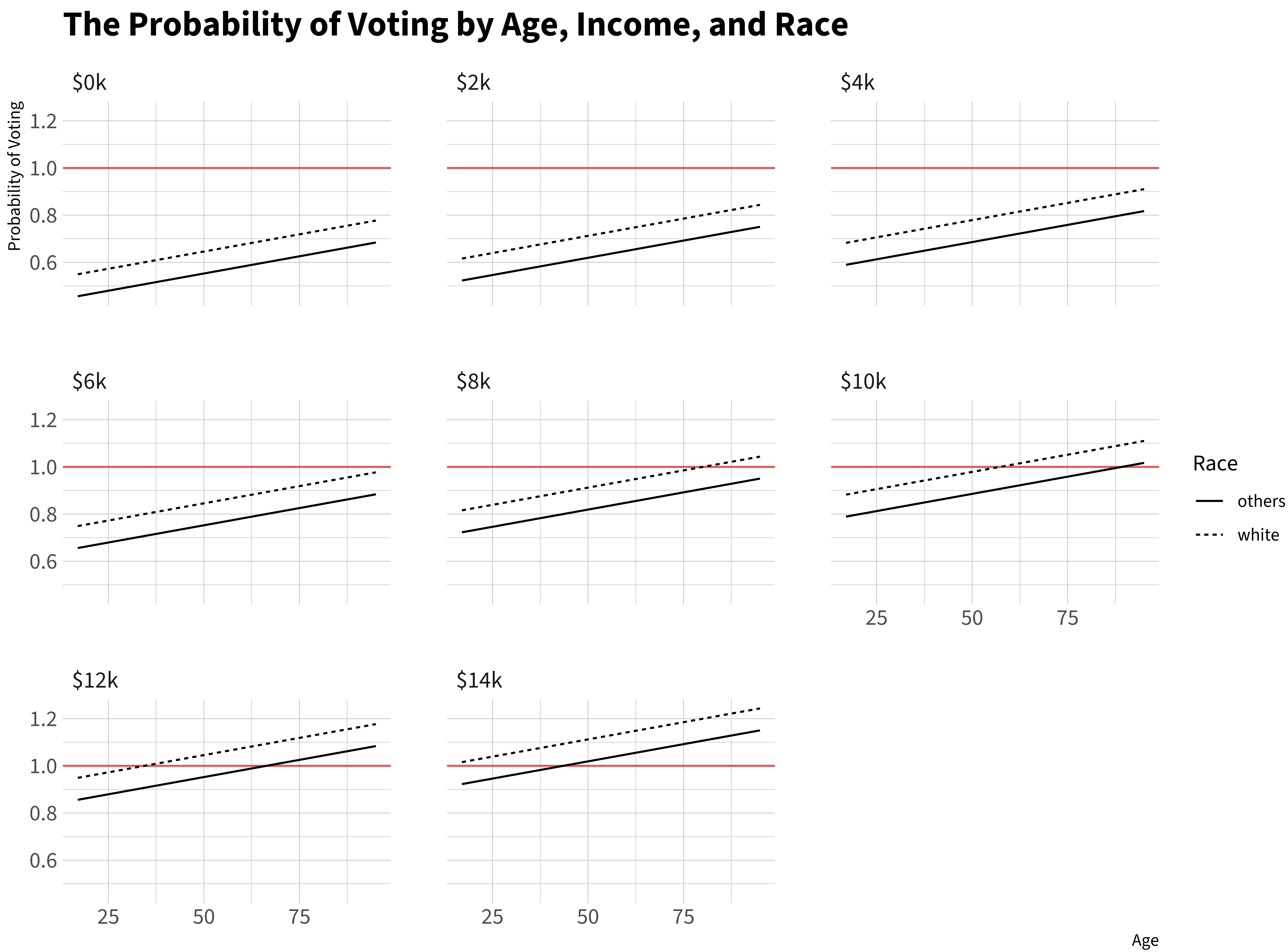
Note: Again, let’s not overstate how problematic this problem is.

You expect smaller effects as the probability of an event approaches zero or one.

problem 4:

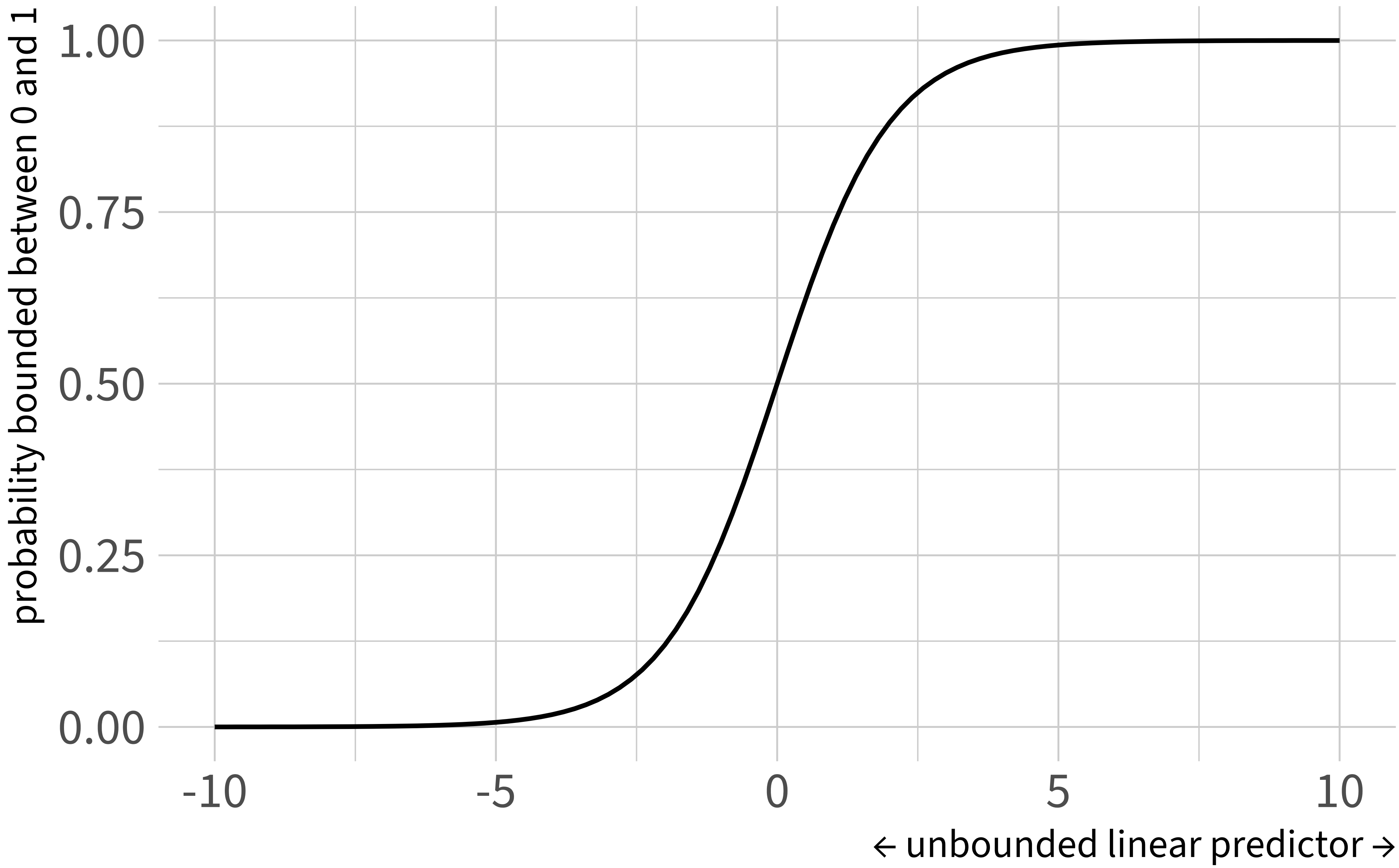
(non)compression

IMO, this is the most important problem.


The Probability of Voting by Age, Income, and Race
$0k
$2k
$4k
$6k
$8k
$10k
$12k
$14k
Probability of Voting
1.2
1.0
0.8
0.6
25
50
75
Age
Race
others
white

How can we include covariates, but keep the π between zero and one?

The Inverse Logit Functin



$$\mathrm{logit}^{-1}(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

This is the CDF of the logistic distribution, which is `plogis()` in R.

$$y_i \sim \mathrm{Bernoulli}(\pi_i),$$
$$\text{where } \pi_i = \mathrm{logit}^{-1}(X_i \beta)$$

```r
# data
devtools::install_github("jrnold/ZeligData")
turnout <- ZeligData::turnout
```

```
> as_tibble(turnout)
# A tibble: 2,000 × 5
   race     age educate income  vote
   <fct>  <int>   <dbl>  <dbl> <int>
 1 white     60      14  3.35      1
 2 white     51      10  1.86      0
 3 white     24      12  0.630     0
 4 white     38       8  3.42      1
 5 white     25      12  2.79      1
 6 white     67      12  2.39      1
 7 white     40      12  4.29      0
 8 white     56      10  9.32      1
 9 white     32      12  3.88      1
10 white     75      16  2.70      1
# ℹ 1,990 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```r
# formula
f <- vote ~ age + educate + income + race

# make X and y
mf <- model.frame(f, data = turnout)
X <- model.matrix(f, data = mf)
y <- model.response(mf)
```

```
> head(X)
  (Intercept) age educate income racewhite
1           1  60      14 3.3458         1
2           1  51      10 1.8561         1
3           1  24      12 0.6304         1
4           1  38       8 3.4183         1
5           1  25      12 2.7852         1
6           1  67      12 2.3866         1
```

```
> head(y)
1 2 3 4 5 6
1 0 0 1 1 1
```

```r
# log-likelihood function
logit_ll <- function(beta, y, X) {
  linpred <- X%*%beta  # perhaps denoted eta
  p <- plogis(linpred) # pi is special in R, so I use p
  ll <- sum(dbinom(y, size = 1, prob = p, log = TRUE))
  return(ll)
}
```

```r
# use optim
par_start <- rep(0, ncol(X))
opt <- optim(par_start,
             fn = logit_ll,
             y = y,
             X = X, # 👈 covariates! 🎉
             method = "BFGS",
             hessian = TRUE,
             control = list(fnscale = -1))
```

```r
# function to fit model
est_logit <- function(f, data) {

  # make X and y
  mf <- model.frame(f, data = data)
  X <- model.matrix(f, data = mf)
  y <- model.response(mf)

# create starting values
  par_start <- rep(0, ncol(X))

  # run optim()
  est <- optim(par_start,
               fn = logit_ll,
               y = y,
               X = X,
               hessian = TRUE,
               control = list(fnscale = -1),
               method = "BFGS")

  # check convergence; print warning if not
  if (est$convergence != 0) print("Model did not converge!")

  # create list of objects to return
  res <- list(beta_hat = est$par,
              var_hat = solve(-est$hessian))

  # return the list
  return(res)
}
```

```r
# fit model
fit <- est_logit(f, data = turnout)
print(fit, digits = 2)
```

```
> # fit model
> fit <- est_logit(f, data = turnout)
> print(fit, digits = 2)
$beta_hat
[1] -3.037  0.028  0.176  0.177  0.251

$var_hat
         [,1]     [,2]     [,3]     [,4]     [,5]
[1,]  0.10622 -8.2e-04 -5.1e-03 -4.2e-04 -1.0e-02
[2,] -0.00082  1.2e-05  2.9e-05  7.3e-06 -5.5e-05
[3,] -0.00514  2.9e-05  4.1e-04 -1.6e-04 -3.2e-04
[4,] -0.00042  7.3e-06 -1.6e-04  7.4e-04 -4.9e-04
[5,] -0.01016 -5.5e-05 -3.2e-04 -4.9e-04  2.1e-02
```

alternatively

```
# data
devtools::install_github("jrnold/ZeligData")
turnout <- ZeligData::turnout

# formula
f <- vote ~ age + educate + income + race

# fit model
fit <- glm(f, data = turnout, family = binomial)

# coefficient estimates
coef(fit)

### (Intercept)         age     educate      income   racewhite
### -3.03426101  0.02835433  0.17563360  0.17711176  0.25079764

# variance estimates
vcov(fit)

###              (Intercept)           age       educate        income     racewhite
### (Intercept)  0.1062494162 -8.242756e-04 -5.146302e-03 -4.236393e-04 -1.015639e-02
### age         -0.0008242756  1.197588e-05  2.907509e-05  7.289931e-06 -5.494225e-05
### educate     -0.0051463023  2.907509e-05  4.134177e-04 -1.583055e-04 -3.219270e-04
### income      -0.0004236393  7.289931e-06 -1.583055e-04  7.375674e-04 -4.896729e-04
### racewhite   -0.0101563932 -5.494225e-05 -3.219270e-04 -4.896729e-04  2.145222e-02
```

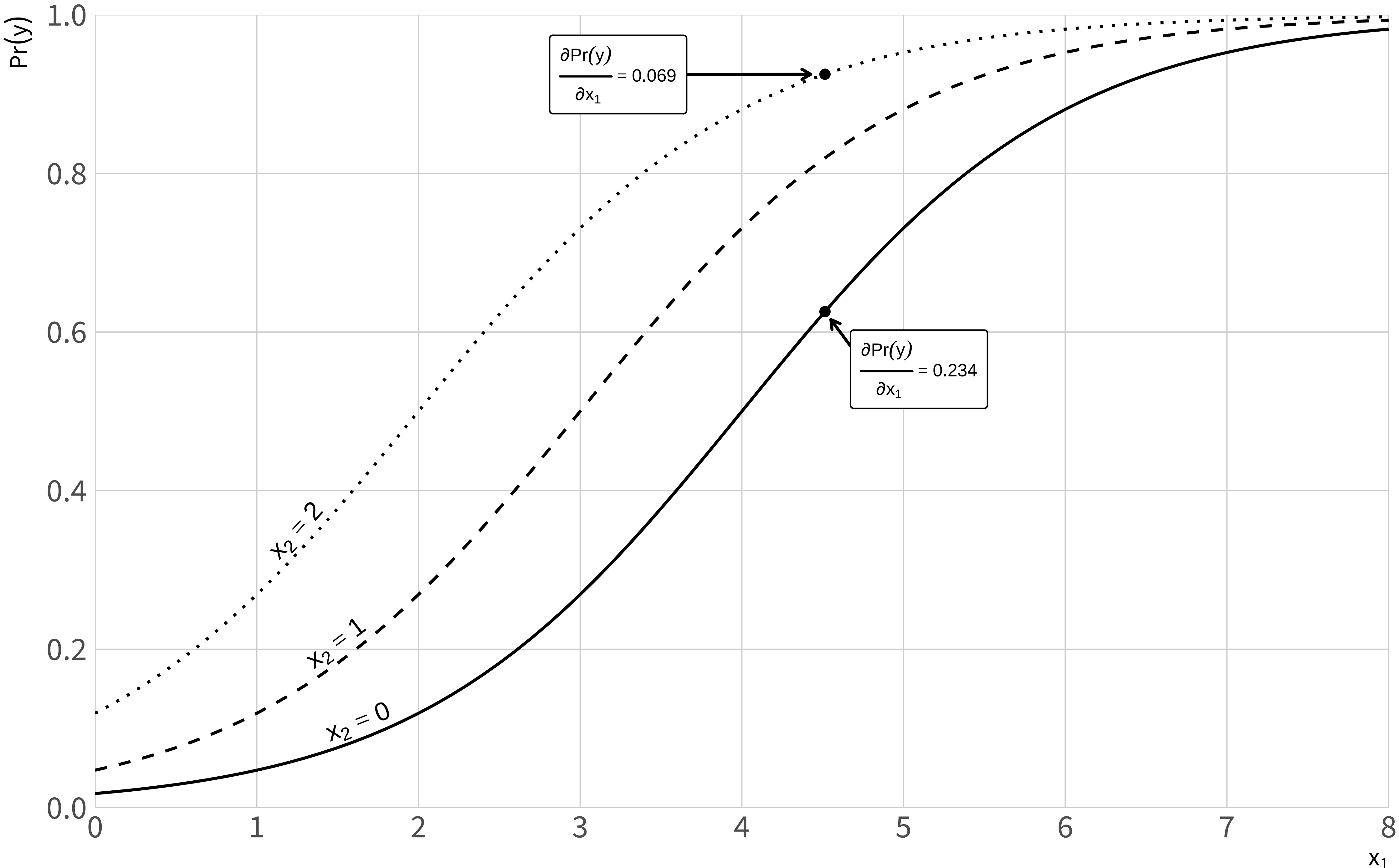
```
> # fit model
> fit <- est_logit(f, data = turnout)
> print(fit, digits = 2)
$beta_hat
[1] -3.037  0.028  0.176  0.177  0.251

$var_hat
         [,1]     [,2]     [,3]     [,4]     [,5]
[1,]  0.10622 -8.2e-04 -5.1e-03 -4.2e-04 -1.0e-02
[2,] -0.00082  1.2e-05  2.9e-05  7.3e-06 -5.5e-05
[3,] -0.00514  2.9e-05  4.1e-04 -1.6e-04 -3.2e-04
[4,] -0.00042  7.3e-06 -1.6e-04  7.4e-04 -4.9e-04
[5,] -0.01016 -5.5e-05 -3.2e-04 -4.9e-04  2.1e-02
```

HUGE

difference

Illustrative Logit Model (Without Product Term)

From Berry, DeMeritt, and Esarey (2010)



Model: $\pi_i = \text{logit}^{-1}(-4 + x_1 + x_2)$

Recall that $\pi_i = \text{logit}^{-1}(X_i\beta)$. To keep this derivative simple, we denote $\eta_i = X_i\beta$ and break it into three parts: (i) find $\frac{\partial \pi_i}{\partial \eta_i}$, (ii) find $\frac{\partial \eta_i}{\partial x_{i1}}$, and (iii) use the chain rule $\frac{\partial \pi_i}{\partial x_{i1}} = \frac{\partial \pi_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial x_{i1}}$.

Step 1: $\frac{\partial \pi_i}{\partial \eta_i}$

$$
\begin{aligned}
\pi_i &= \frac{e^{\eta_i}}{1+e^{\eta_i}} && \text{definition of inverse logit} \\
\frac{\partial \pi_i}{\partial \eta_i} &= \frac{(1+e^{\eta_i})\, e^{\eta_i} - e^{\eta_i}\, e^{\eta_i}}{(1+e^{\eta_i})^2} && \text{quotient rule} \\
&= \frac{e^{\eta_i}}{1+e^{\eta_i}} \left(1 - \frac{e^{\eta_i}}{1+e^{\eta_i}}\right) && \text{factor out } \tfrac{e^{\eta_i}}{1+e^{\eta_i}} \\
&= \pi_i\,(1-\pi_i) && \text{substitute } \pi_i = \tfrac{e^{\eta_i}}{1+e^{\eta_i}}
\end{aligned}
$$

Exercise 8 Inverse Logit

Define $p(\theta) = \frac{1}{1-e^{-\theta}}$ for $\theta \in \mathbb{R}$. Find $p'(\theta)$ and $p''(\theta)$.

Step 2: $\frac{\partial \eta_i}{\partial x_{i1}}$

$$
\frac{\partial \eta_i}{\partial x_{i1}} = \beta_1 \quad \text{since } \eta_i = \sum_j x_{ij}\beta_j
$$

(usual linear regression derivative)

Step 3: Chain rule

$$
\begin{aligned}
\frac{\partial \pi_i}{\partial x_{i1}} &= \frac{\partial \pi_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial x_{i1}} \\
&= \pi_i(1-\pi_i)\,\beta_1 \\
&= \left[\text{logit}^{-1}(X_i\beta)\right] \cdot \left[1 - \text{logit}^{-1}(X_i\beta)\right] \cdot \beta_1
\end{aligned}
$$

(putting the two above together)

The marginal effect of x_{i1} on the probability of an event is **not constant**. It depends on the value of x_{i1} *and all the other x's.*

The marginal effect of a variable on the probability of an event is **not constant**. It depends on the value of that variable and the values of all the other variables.

like a **polynomial**
in linear regression

like an **interaction**
in linear regression

TABLE 2. Models of Involvement in Militarized Disputes, 1950–1985: Assessing the Liberal Peace

Variable		*Eqn 1*	*Eqn 2*	*Eqn 3*	*Eqn 4*
Democracy score$_L$	ß	-0.0497	-0.0554	-0.0413	-0.0457
	SE$_ß$	0.0074	0.0077	0.0083	0.0076
	p	<.001	<.001	<.001	<.001
Economic growth rate$_L$		-0.0223	-0.0318	-0.0297	-0.0220
		0.0085	0.0093	0.0094	0.0090
		.009	<.001	<.001	.02
Allies		-0.821	-0.868	-0.864	-0.815
		0.080	0.092	0.092	0.083
		<.001	<.001	<.001	<.001
Contiguity		1.31	1.26	1.32	1.39
		0.08	0.09	0.09	0.08
		<.001	<.001	<.001	<.001
Capability ratio		-0.00307	-0.00345	-0.00356	-0.00279
		0.00042	0.00050	0.00051	0.00041
		<.001	<.001	<.001	<.001
Dyadic trade-to-GDP ratio$_L$		-66.1		-43.8	-81.0
		13.4		12.5	15.2
		<.001		<.001	<.001
Total trade-to-GDP ratio$_L$			-0.706	-0.512	
			0.192	0.193	
			<.001	.008	
Trend, dyadic trade-to-GDP ratio$_H$					-8.89
					2.93
					<.001
Constant		-3.29	-3.28	-3.20	-3.32
		0.08	0.10	0.10	0.08
		<.001	<.001	<.001	<.001
Chi^2		764.04	593.65	611.74	728.56
P of chi^2		<.0001	<.0001	<.0001	<.0001
Log likelihood		-3477.57	-2795.89	-2786.85	-3231.42
N		20,990	17,709	17,709	19,772

quantities of interest

Making the Most of Statistical Analyses: Improving Interpretation and Presentation

Gary King Harvard University
Michael Tomz Harvard University
Jason Wittenberg Harvard University

Social scientists rarely take full advantage of the information available in their statistical results. As a consequence, they miss opportunities to present quantities that are of greatest substantive interest for their research and express the appropriate degree of certainty about these quantities. In this article, we offer an approach, built on the technique of statistical simulation, to extract the currently overlooked information from any statistical method and to interpret and present it in a reader-friendly manner. Using this technique requires some expertise, which we try to provide herein, but its application should make the results of quantitative articles more informative and transparent. To illustrate our recommendations, we replicate the results of several published works, showing in each case how the au-

We show that social scientists often do not take full advantage of the information available in their statistical results and thus miss opportunities to present quantities that could shed the greatest light on their research questions. In this article we suggest an approach, built on the technique of statistical simulation, to extract the currently overlooked information and present it in a reader-friendly manner. More specifically, we show how to convert the raw results of any statistical procedure into expressions that (1) convey numerically precise estimates of the quantities of greatest substantive interest, (2) include reasonable measures of uncertainty about those estimates, and (3) require little specialized knowledge to understand.

The following simple statement satisfies our criteria: "Other things being equal, an additional year of education would increase your annual income by $1,500 on average, plus or minus about $500." Any smart high school student would understand that sentence, no matter how sophisticated the statistical model and powerful the computers used to produce it. The sentence is substantively informative because it conveys a key quantity of interest in terms the reader wants to know. At the same time, the sentence indicates how uncertain the researcher is about the estimated quantity of interest. Inferences are never certain, so any honest presentation of statistical results must include some qualifier, such as "plus or minus $500" in the present example.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	-2.67^{*}	-2.91^{*}	-3.02^{*}	-4.73^{*}	-4.73^{*}	-4.86^{*}
	(0.66)	(0.69)	(0.72)	(0.65)	(0.67)	(0.69)
District Competitiveness	3.69^{*}	3.45^{*}	3.32^{*}	4.07^{*}	4.07^{*}	3.93^{*}
	(0.78)	(0.80)	(8.1)	(0.94)	(0.96)	(0.96)
PR	0.13^{*}	-0.06	0.43	0.85	0.85	1.20^{*}
	(1.01)	(1.04)	(1.41)	(0.95)	(0.96)	(1.05)
District Competitiveness × PR	-2.34^{*}	-2.20^{*}	-1.98	-2.91^{*}	-2.91^{*}	-2.71^{*}
	(1.24)	(1.25)	(1.26)	(1.39)	(1.40)	(1.40)
ENEP		0.14^{*}	0.21^{*}		> 0.01	0.08
		(0.08)	(0.10)		(0.09)	(0.12)
PR × ENEP			-0.19			-0.15
			(0.16)			(0.18)
Age				0.01^{*}	0.01^{*}	0.01^{*}
				(> 0.01)	(> 0.01)	(> 0.01)
Male				0.07	0.07	0.07
				(0.07)	(0.07)	(0.07)
Education				0.16^{*}	0.16^{*}	0.16^{*}
				(0.02)	(0.02)	(0.02)
Married				0.02	0.02	0.02
				(0.08)	(0.08)	(0.08)
Union Member				0.24^{*}	0.24^{*}	0.24^{*}
				(0.08)	(0.08)	(0.08)
Household Income				0.10^{*}	0.10^{*}	0.10^{*}
				(0.03)	(0.03)	(0.03)
Close To Party				0.43^{*}	0.43^{*}	0.43^{*}
				(0.07)	(0.07)	(0.07)
Number of Respondents	7652	7651	7651	5219	5218	5218
Number of Districts	614	613	613	569	568	568
Number of Elections	5	5	5	5	5	5
AIC	8308	8307	8308	5529	5531	5532
BIC	8364	8369	8377	5627	5636	5644

Standard errors in parentheses

* indicates significance at $p < 0.05$

TABLE 2: A table showing a series of hierarchical models that demostrate the robustness of the conclusions to changes in model specification. Notice especially that the substantive results do not change with the inclusion of the effective number of parties.

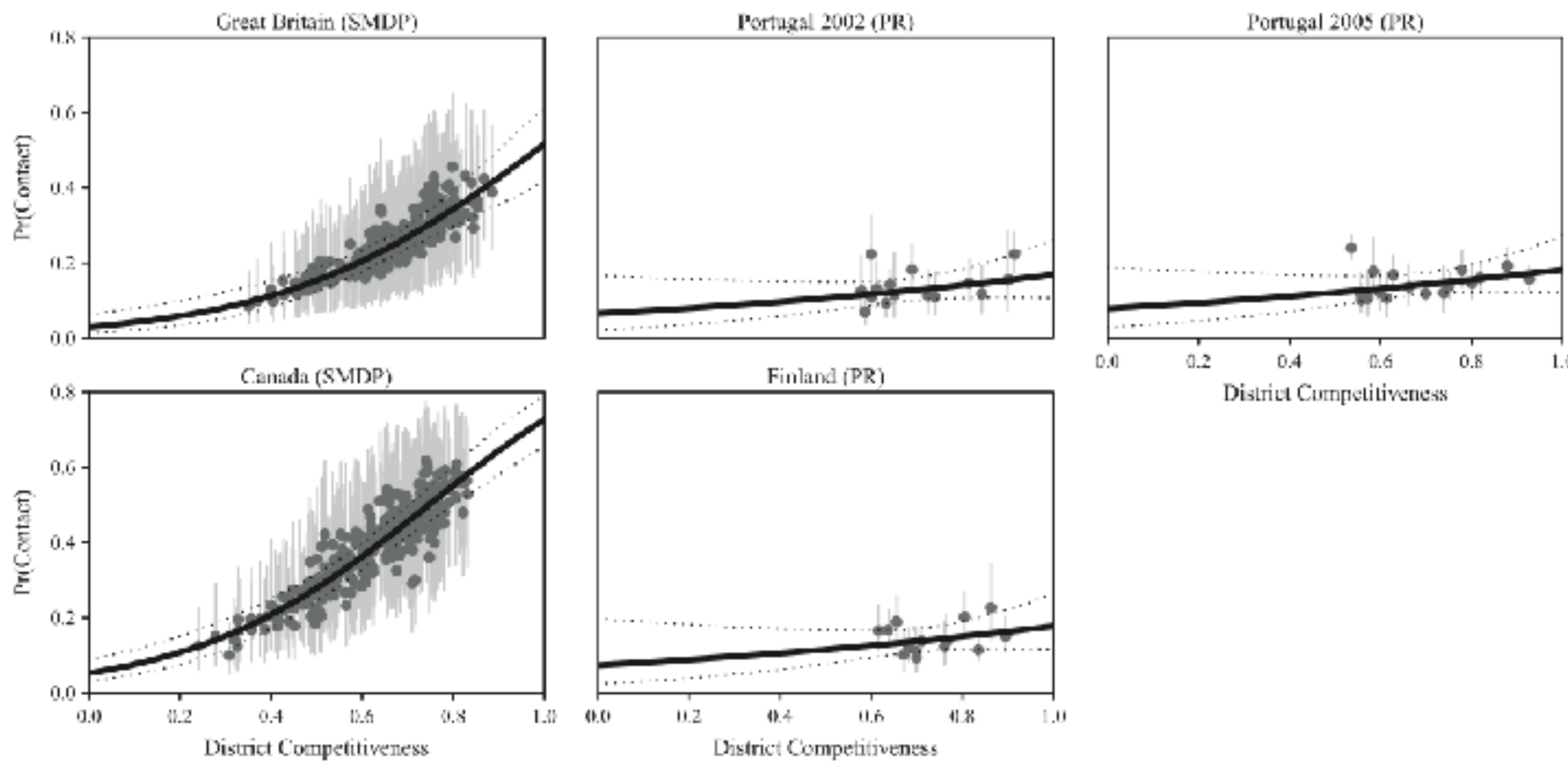


Fig. 5. This figure shows how the predicted probability of an individual receiving contact from a political party changes as competitiveness increases across the countries included in the analysis. The solid line indicates the estimated probability and the dotted lines show the 90% Bayesian credible interval around that estimate The figure shows that the results in Fig. 1 is robust. The two SMDP elections in the data (Canada and Great Britain) show large increases in the probability of receiving contact as competitiveness increases, while the three PR elections (Finland and Portugal) show no increase. This offers strong support for the Interaction Hypothesis.

We can be **very creative** with quantities of interest.

Any $\tau(\theta)$ works!

But we usually want the **expected value** and **first difference**.

average. The predicted value will be expressed in the same metric as the dependent variable, so it should require little specialized knowledge to understand.

Expected Values

Depending on the issue being studied, the *expected* or mean value of the dependent variable may be more interesting than a predicted value. The difference is subtle but important. A predicted value contains both fundamental

$$E(y \mid X_c) = \pi = \text{logit}^{-1}(X_c\beta)$$

```r
# create chosen values for X
# note 1: naming columns helps a bit later
# note 2: can also do with f, model.matrix(..., newdata = ...)
X_c <- cbind(
  "constant" = 1, # intercept
  "age"      = median(turnout$age),
  "educate"  = median(turnout$educate),
  "income"   = median(turnout$income),
  "white"    = 1 # white indicators = 1
)
```

```
> head(X)
  (Intercept) age educate income racewhite
1           1  60      14 3.3458         1
2           1  51      10 1.8561         1
3           1  24      12 0.6304         1
4           1  38       8 3.4183         1
5           1  25      12 2.7852         1
6           1  67      12 2.3866         1
> head(X_c)
     constant age educate income white
[1,]        1  42      12 3.3508     1
```

Warning: The way I'm doing this, R isn't helping use make sure these match up. See note 2.

```r
# function to compute qi
ev_fn <- function(beta, X) {
  plogis(X%*%beta)
}

# invariance property
ev_hat <- ev_fn(fit$beta_hat, X_c)
```

```
> ev_hat
          [,1]
[1,] 0.7517864
```

```r
# delta method
library(numDeriv)  # for grad()
grad <- grad(
  func = ev_fn,      # what function are we taking the derivative of?
  x = fit$beta_hat, # what variable(s) are we taking the derivative w.r.t.?
  X = X_c)          # what other values are needed?
se_ev_hat <- sqrt(grad %*% fit$var_hat %*% grad)
```

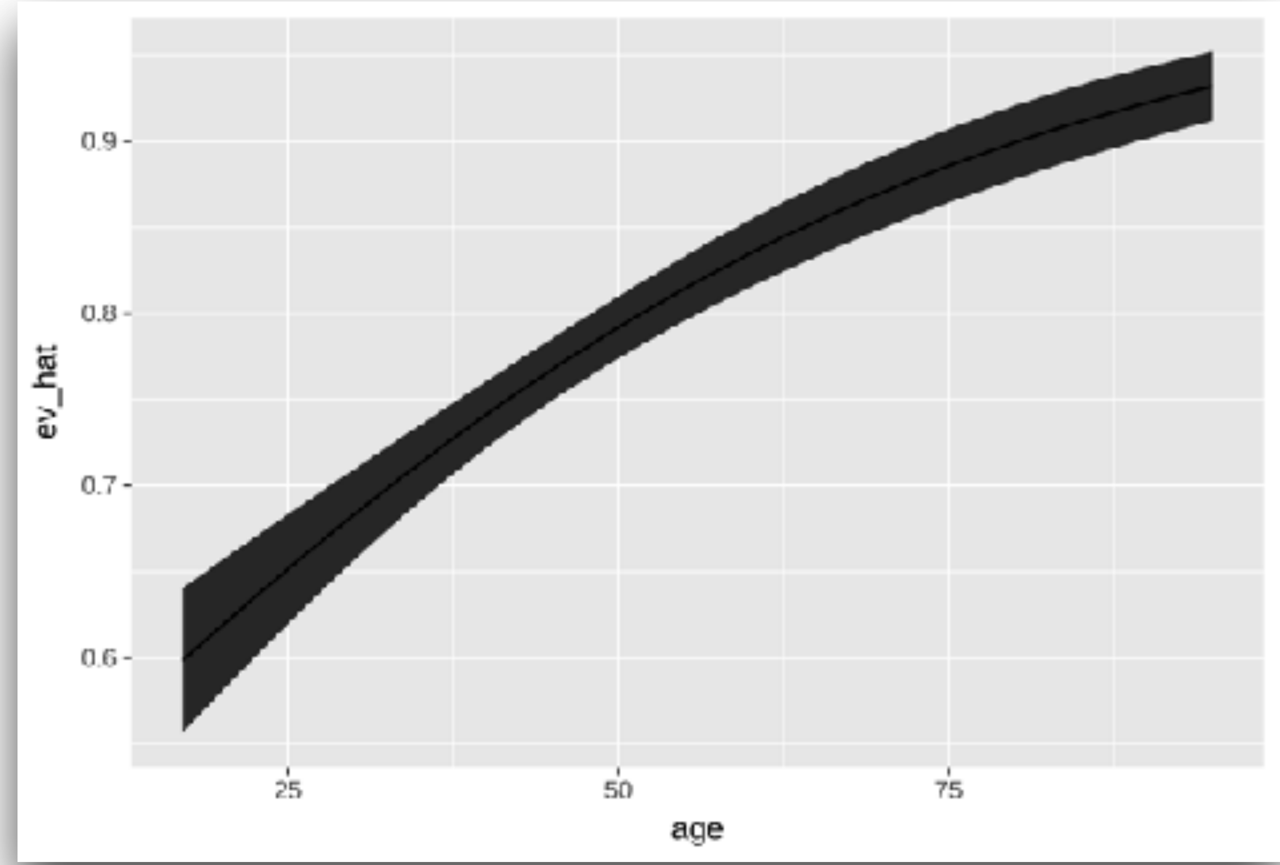
```
> se_ev_hat
          [,1]
[1,] 0.0114178
```

```r
# ---- compute the ev given X_c (w/ range of values) ----

# create chosen values for X
X_c <- cbind(
  "constant" = 1, # intercept
  "age"      = min(turnout$age):max(turnout$age),
  "educate"  = median(turnout$educate),
  "income"   = median(turnout$income),
  "white"    = 1 # white indicators = 1
)

# containers for estimated quantities of interest and ses
ev_hat <- numeric(nrow(X_c))
se_ev_hat <- numeric(nrow(X_c))

# loop over each row of X_c and compute qi and se
for (i in 1:nrow(X_c)) {   # for the ith row of X...
  # invariance property
  ev_hat[i] <- ev_fn(fit$beta_hat, X_c[i, ])
  # delta method
  grad <- grad(
    func = ev_fn,
    x = fit$beta_hat,
    X = X_c[i, ])
  se_ev_hat[i] <- sqrt(grad %*% fit$var_hat %*% grad)
}
```

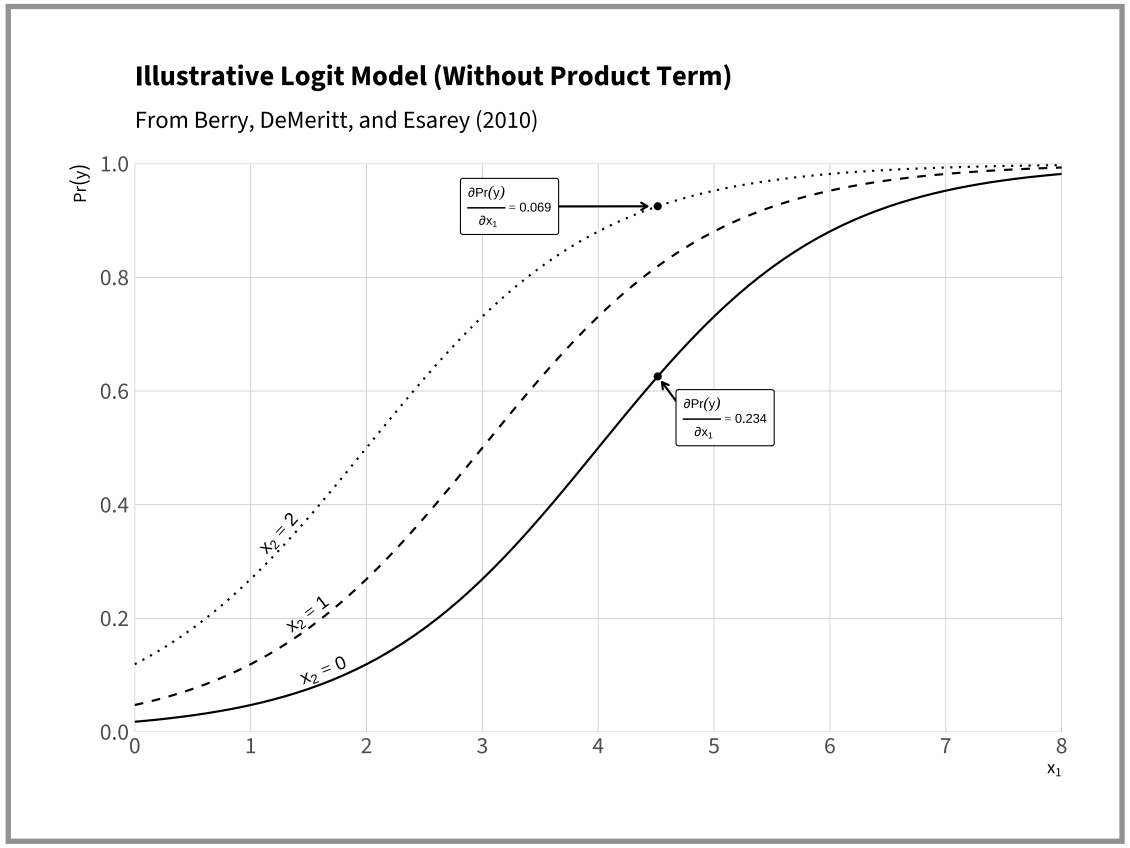
```
# put X_c, qi estimates, and se estimates in data frame
qi <- cbind(X_c, ev_hat, se_ev_hat) |>
  data.frame() |>
  glimpse()
```

```
> # put X_c, qi estimates, and se estimates in data frame
> qi <- cbind(X_c, ev_hat, se_ev_hat) |>
+   data.frame() |>
+   glimpse()
Rows: 79
Columns: 7
$ constant  <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ age       <dbl> 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36,…
$ educate   <dbl> 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12,…
$ income    <dbl> 3.3508, 3.3508, 3.3508, 3.3508, 3.3508, 3.3508, 3.3508, 3.3508, 3.3508, 3.3508,…
$ white     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ ev_hat    <dbl> 0.5983202, 0.6051232, 0.6118858, 0.6186055, 0.6252802, 0.6319075, 0.6384855, 0.…
$ se_ev_hat <dbl> 0.02562903, 0.02480870, 0.02399725, 0.02319621, 0.02240713, 0.02163161, 0.02087…
```

```r
# plot
ggplot(qi, aes(x = age, y = ev_hat,
               ymin = ev_hat - 1.64*se_ev_hat,
               ymax = ev_hat + 1.64*se_ev_hat)) +
  geom_ribbon() +
  geom_line()
```



Compare to this one from before.



gorithm, since steps 1–3 suffice to simulate one expected value. This shortcut is appropriate for the linear-normal and logit models in Equations 2 and 3.

First Differences

A first difference is the difference between two expected, rather than predicted, values. To simulate a first difference, researchers need only run steps 2–5 of the expected value algorithm twice, using different settings for the explanatory variables

$$E(y \mid X_{lo}) - E(y \mid X_{lo})$$

```r
# ---- compute first difference ----

# make X_lo
X_lo <- cbind(
  "constant" = 1, # intercept
  "age"      = quantile(turnout$age, probs = 0.25), # 31 years old; 25th percentile
  "educate"  = median(turnout$educate),
  "income"   = median(turnout$income),
  "white"    = 1 # white indicators = 1
)

# make X_hi by modifying the relevant value of X_lo
X_hi <- X_lo
X_hi[, "age"] <- quantile(turnout$age, probs = 0.75) # 59 years old; 75th percentile

# function to compute first difference
fd_fn <- function(beta, hi, lo) {
  plogis(hi%*%beta) - plogis(lo%*%beta)
}

# invariance property
fd_hat <- fd_fn(fit$beta_hat, X_hi, X_lo)

# delta method
grad <- grad(
  func = fd_fn,
  x = fit$beta_hat,
  hi = X_hi,
  lo = X_lo)
se_fd_hat <- sqrt(grad %*% fit$var_hat %*% grad)

# estimated fd
fd_hat

# estimated se
se_fd_hat

# 90% ci
fd_hat - 1.64*se_fd_hat  # lower
fd_hat + 1.64*se_fd_hat  # upper
```

```
> # estimated fd
> fd_hat
         [,1]
25% 0.1416257
> # estimated se
> se_fd_hat
          [,1]
[1,] 0.0170934
> # 90% ci
> fd_hat - 1.64*se_fd_hat  # lower
         [,1]
25% 0.1135925
> fd_hat + 1.64*se_fd_hat  # upper
         [,1]
25% 0.1696588
```

Your Turn!

https://gist.github.com/carlislerainey/7798659a9d5d8decb87352068a2d1655

models, generally

1. Choose a distribution.
2. Choose the parameters to model as functions of covariates and those to model as fixed.
3. Choose an inverse link function.
4. Fit the model.
5. Compute QIs.

models for counts

The Distributive Politics of Enforcement

Alisha C. Holland Harvard University

Why do some politicians tolerate the violation of the law? In contexts where the poor are the primary violators of property laws, I argue that the answer lies in the electoral costs of enforcement: Enforcement can decrease support from poor voters even while it generates support among nonpoor voters. Using an original data set on unlicensed street vending and enforcement operations at the subcity district level in three Latin American capital cities, I show that the combination of voter demographics and electoral rules explains enforcement. Supported by qualitative interviews, these findings suggest how the intentional nonenforcement of law, or forbearance, can be an electoral strategy. Dominant theories based on state capacity poorly explain the results.

In much of the developing world, a source of resources for the poor is the ability to violate property laws without state sanction. Squatters gain rent-free housing if their takings succeed. Street vendors secure a way to earn a living when the government ignores their unlicensed stands. The idea that enforcement has distributive consequences is not new. Yet conventional wisdom is that limited enforcement reflects a weak state unable to implement its laws due to budget constraints or principal-agent problems.

In contrast, this article argues that nonenforcement of law is often intentional—what I call *forbearance*—and explains why some governments tolerate violations of the law by the poor and others do not. The argument is sim-

An intuitive distributive logic thus provides greater leverage to understand enforcement (and its absence) than dominant capacity-based approaches.

Focusing on variation in enforcement against unlicensed street vendors at the city and subcity level, this article tests this electoral theory in two ways. I first examine time-series data on enforcement in a city that constitutes a single electoral district, Bogotá, Colombia. I show that city mayors with nonpoor core constituencies conduct almost five times more enforcement operations against street vendors than those with poor constituencies. Second, I collect original data on enforcement operations and unlicensed street vending in a sample of 89 subcity units, or districts, in three cities. I select cities that vary in tics. The Poisson distribution is appropriate given that enforcement is a count variable with a range restricted to positive integers.[13]

My first hypothesis is that enforcement operations drop off with the fraction of poor residents in an electoral district. So district poverty should be a negative and significant predictor of enforcement, but only in politically decentralized cities. Poverty should have no relationship with enforcement in politically centralized cities once one controls for the number of vendors.

I include the number of vendors as a covariate for the limited purpose of observing the difference depending on whether enforcement policy is locally or centrally

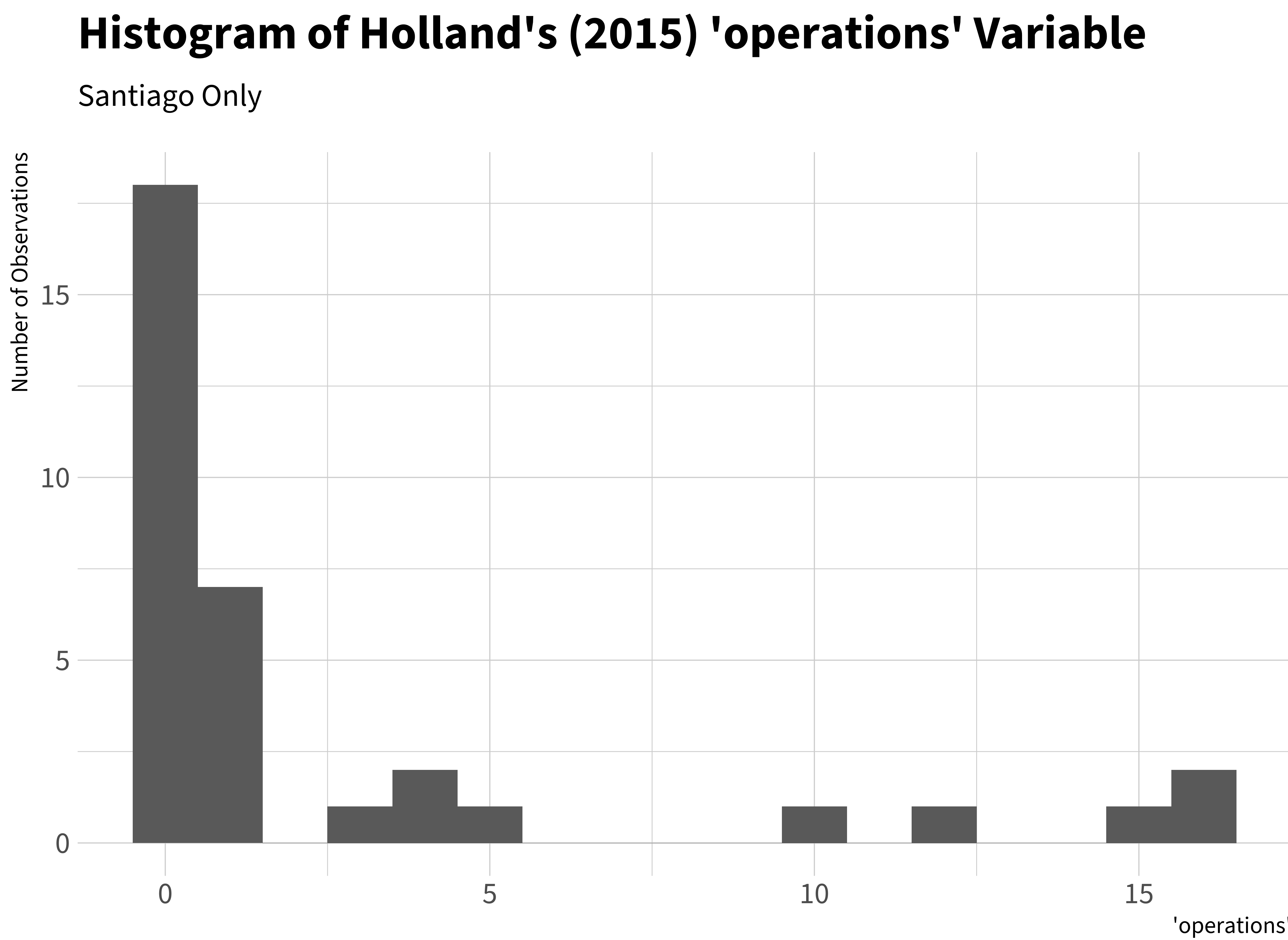
Table 1 Theoretical Hypotheses and Empirical Predictions

Hypothesis	Empirical Prediction
Hypothesis 1: Enforcement decreases with the poverty of an electoral district.	$\beta_{lower} < 0$, $\beta_{vendors\|lower} \approx 0$ in Lima and Santiago $\beta_{vendors} > 0$ in Bogotá
Hypothesis 2: District demographics are less relevant under limited political competition.	$\beta_{lower\|vendors} \approx 0$ in Bogotá $\beta_{margin*lower} > 0$ in Lima and Santiago
Hypothesis 3: Politicians enforce less when their core constituents are poor.	$\bar{E}_{nonpoor} > \bar{E}_{poor}$ in Bogotá $\beta_{right} > 0$ in Santiago
Alternative 1: Enforcement decreases with the poverty of a district due to capacity constraints.	$\beta_{lower} < 0$ in all cities $R^2_{budget} > R^2_{lower}$
Alternative 2: Enforcement decreases with the poverty of a district because the police are less responsive.	$\beta_{lower} < 0$ in all cities $\beta^{arrests}_{lower} < 0$ in Santiago
Alternative 3: Politicians enforce in proportion to the number of offenses.	$\beta_{vendor} > 0$ in all cities

Notes: $\bar{E}$ refers to the expected number of enforcement operations.

```
> # data; see ?crdata::holland2015
> holland <- crdata::holland2015 |>
+   filter(city == "santiago")
> glimpse(holland)
Rows: 34
Columns: 7
$ city       <chr> "santiago", "santiago", "santiago", "santiago", "santiago", "santiago", "santi…
$ district   <chr> "Cerrillos", "Cerro Navia", "Conchali", "El Bosque", "Estacion Central", "Huec…
$ operations <dbl> 0, 0, 0, 0, 12, 0, 0, 0, 1, 1, 0, 10, 1, 5, 0, 0, 0, 4, 4, 0, 1, 16, 1, 1, 0, …
$ lower      <dbl> 52.2, 69.8, 54.8, 58.4, 43.6, 58.3, 41.0, 38.3, 36.7, 60.1, 73.8, 16.4, 7.7, 2…
$ vendors    <dbl> 0.50, 0.60, 5.00, 1.20, 1.00, 0.30, 0.05, 1.25, 2.21, 0.70, 1.00, 0.50, 0.05, …
$ budget     <dbl> 337.24, 188.87, 210.71, 153.76, 264.43, 430.42, 312.75, 255.53, 149.48, 164.98…
$ population <dbl> 6.6160, 13.3943, 10.7246, 16.8302, 11.1702, 8.5761, 5.1277, 7.1443, 39.8355, 1…
```

Note: Santiago is “highly decentralized.”


Histogram of Holland's (2015) 'operations' Variable
Santiago Only
Number of Observations
0
5
10
15
0
5
10
15
'operations'

Poisson Regression

Element	**Details**
Outcome	Count (non-negative integers).
Model	$y_i \sim \text{Poisson}(\lambda_i)$, with $\lambda_i = \exp(X_i\beta)$. Note: $\text{Var}(y_i) = \lambda_i$.
Expected value	$\hat{\lambda} = \exp(X_c\hat{\beta})$.
First difference	$\hat{\Delta} = \hat{\lambda}_{\text{hi}} - \hat{\lambda}_{\text{lo}} = \exp(X_{\text{hi}}\hat{\beta}) - \exp(X_{\text{lo}}\hat{\beta})$.
Function	`glm()` with `family = poisson` for fitting models; `dpois()` for probabilities.
Parameterization notes	None.
Alternatives	Negative binomial regression (relaxes variance = mean), zero-inflated variants.

How would our logit code change?

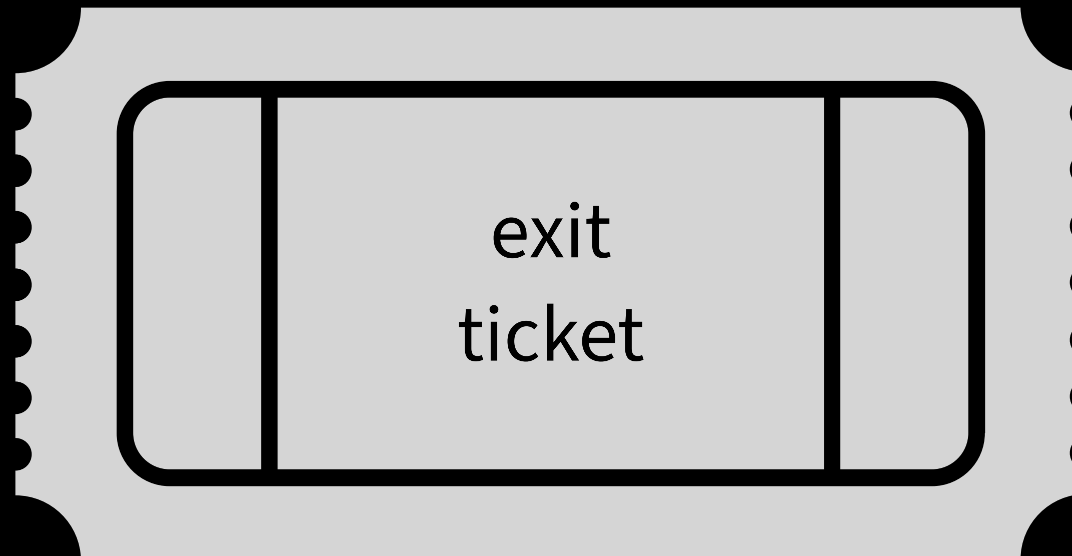
https://www.diffchecker.com/u0EHewB1/

Negative Binomial Regression

Element	**Details**
Outcome	Count (non-negative integers).
Model	$y_i \sim \text{NegBin}(\mu_i, \theta)$, with $\mu_i = \exp(X_i\beta)$. Note: $\text{Var}(y_i) = \mu_i + \mu_i^2/\theta$.
Expected value	$\hat{\mu} = \exp(X_c\hat{\beta})$.
First difference	$\hat{\Delta} = \hat{\mu}_{\text{hi}} - \hat{\mu}_{\text{lo}} = \exp(X_{\text{hi}}\hat{\beta}) - \exp(X_{\text{lo}}\hat{\beta})$.
Function	`MASS::glm.nb()` for fitting models; `dnbinom()` for densities.
Parameterization notes	Uses the mean–dispersion form where regression is parameterized by the mean μ_i with a exp inverse link. The dispersion parameter θ (`size` in R) controls overdispersion, where $\text{Var}(y_i) = \mu_i + \mu_i^2/\theta$. In `dnbinom()` you can use either (`size`, `prob`) or (`size`, μ); `glm.nb()` uses (μ_i, θ).
Alternatives	Poisson regression (variance equals mean), zero-inflated variants.

How would our Poisson code change?

https://www.diffchecker.com/67AIcHV3/



The last three weeks (ML, SE, and X) outline a coherent and complete toolkit for statistical modeling. What are the major ideas and how do they fit together?