Design and Analysis of Sample Surveys

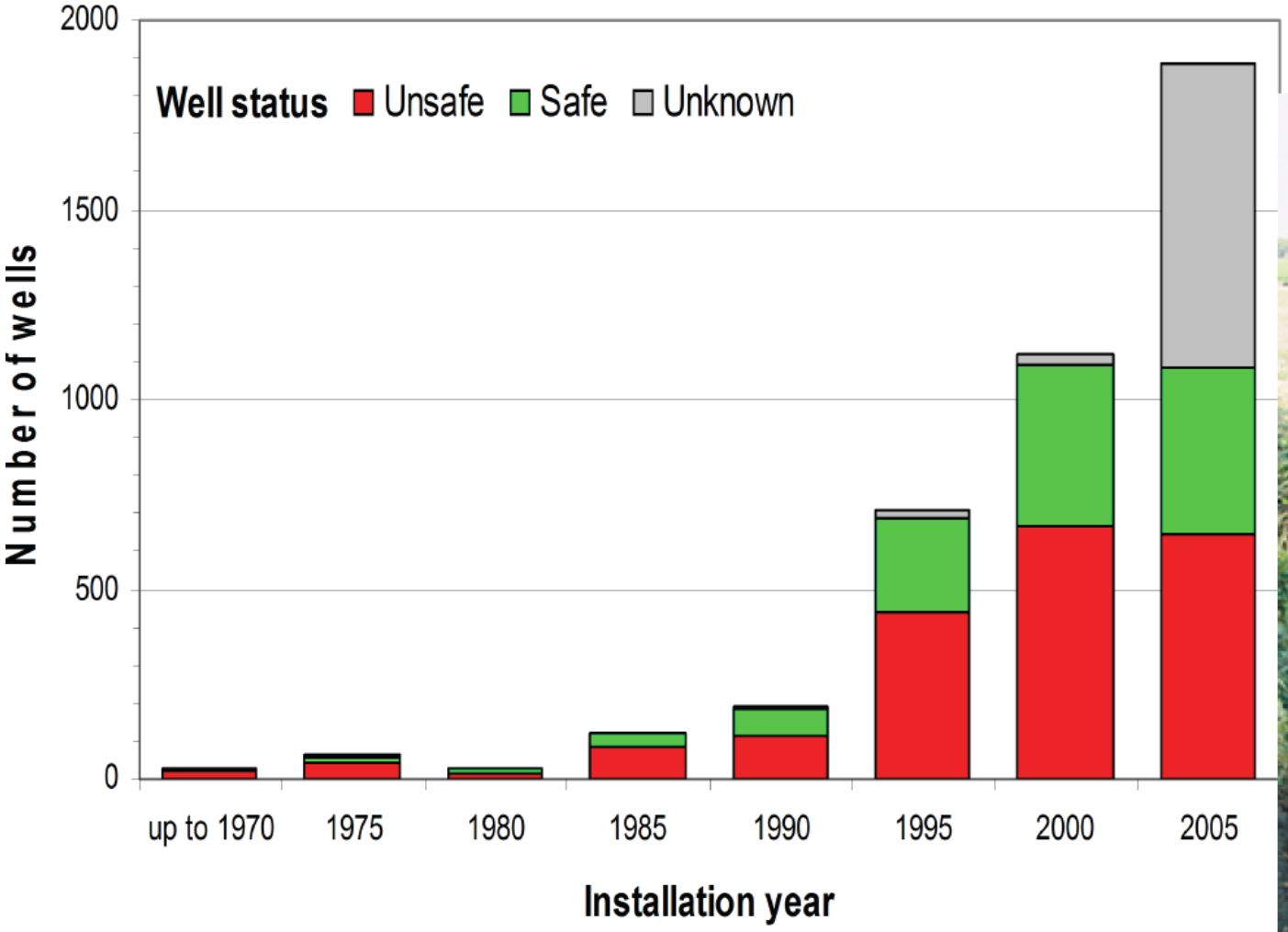
Andrew Gelman
Department of Statistics and Department of Political Science
Columbia University

Class 2a: Logistic regression

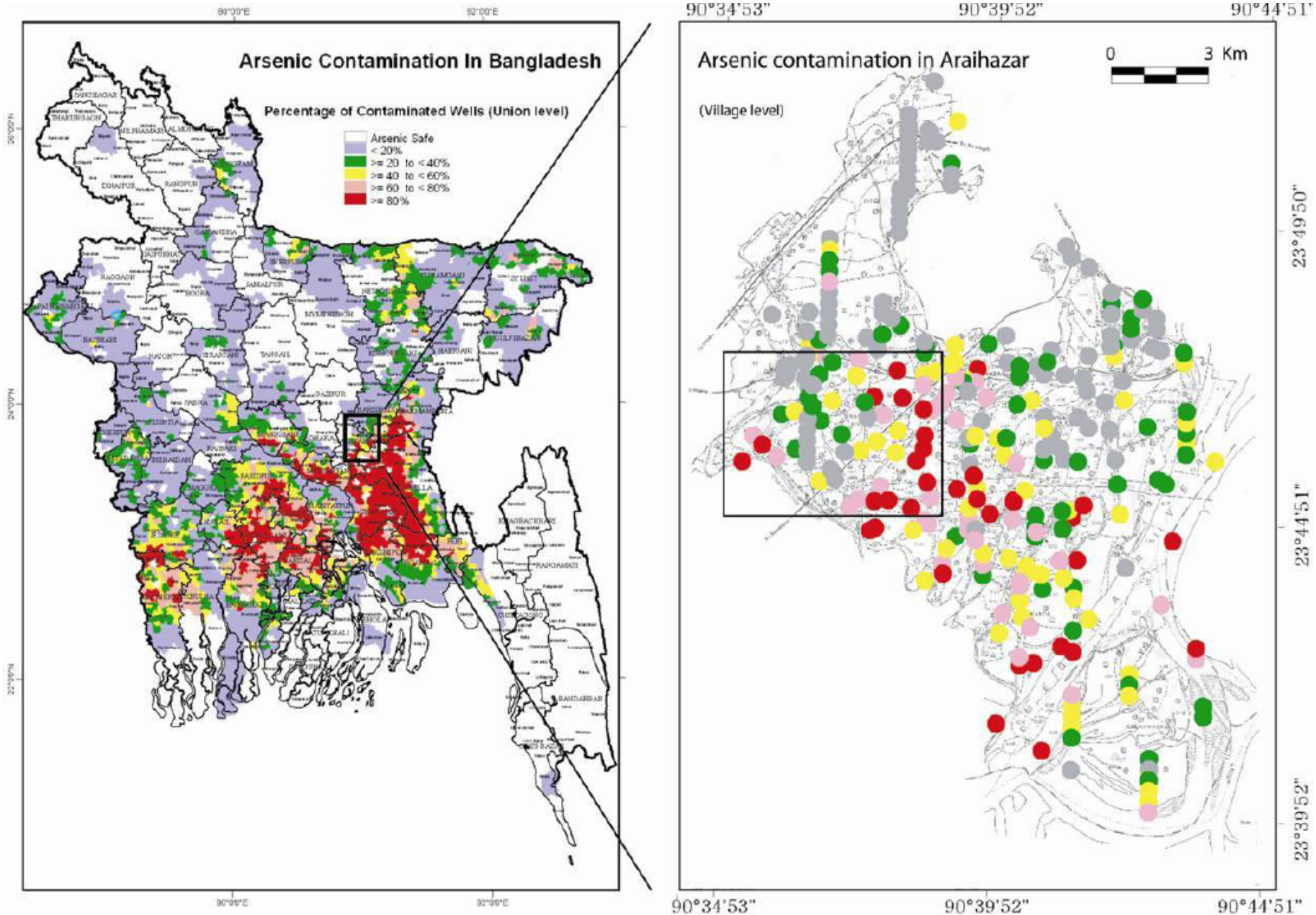
Logistic regression

- Example: switching wells to get safer drinking water in Bangladesh
- Building a logistic regression model
- Logistic regression with interactions
- Evaluating, checking, and comparing models

Natural arsenic in well water



Mix of high and low-arsenic wells



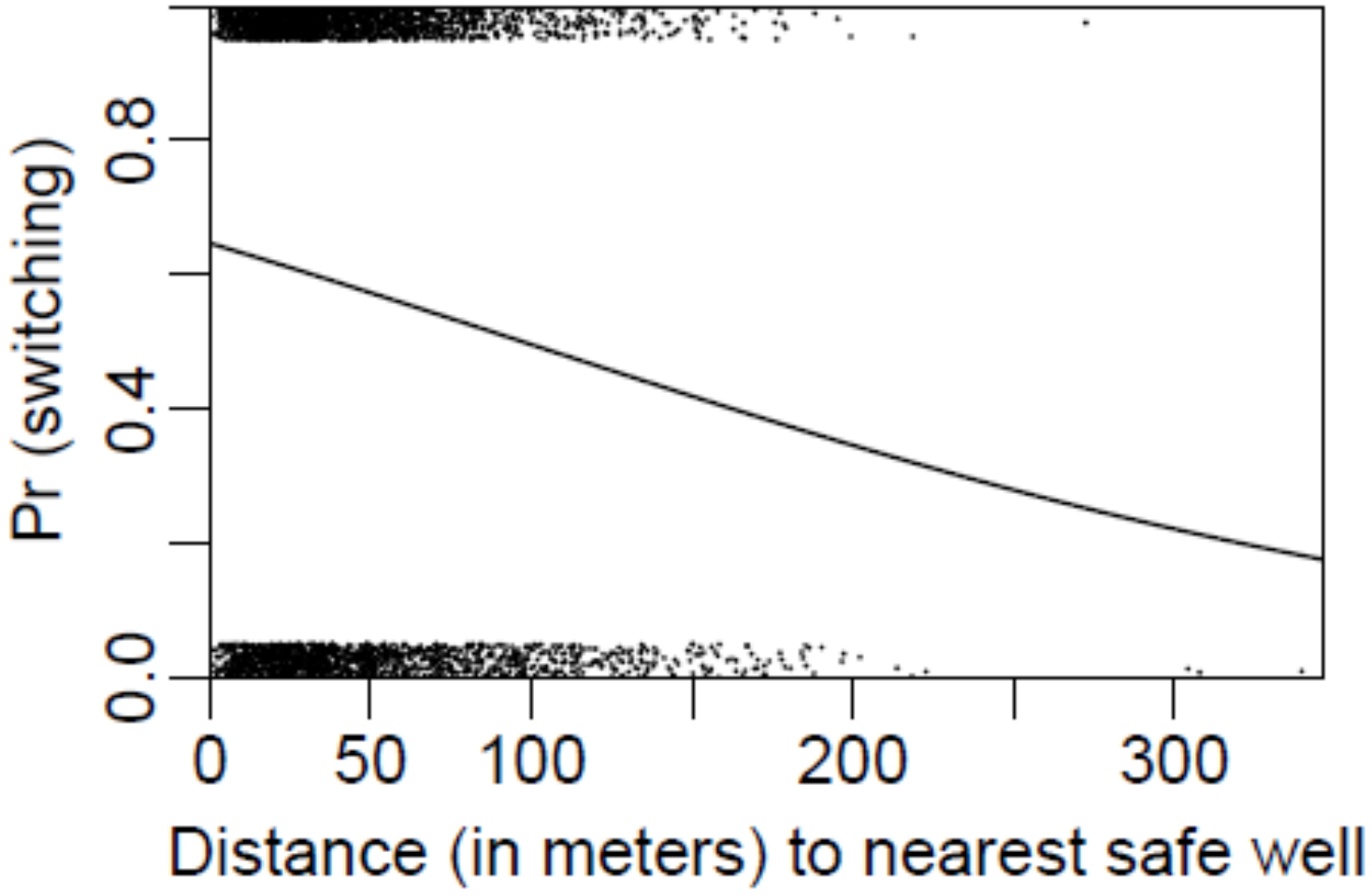
Digging new wells

Survey data: would you switch wells?

- Logistic regression
- Predictor variables:
  - Distance to nearest safe well
  - Arsenic level of your current well
  - Education
  - Membership in community organizations (not predictive)

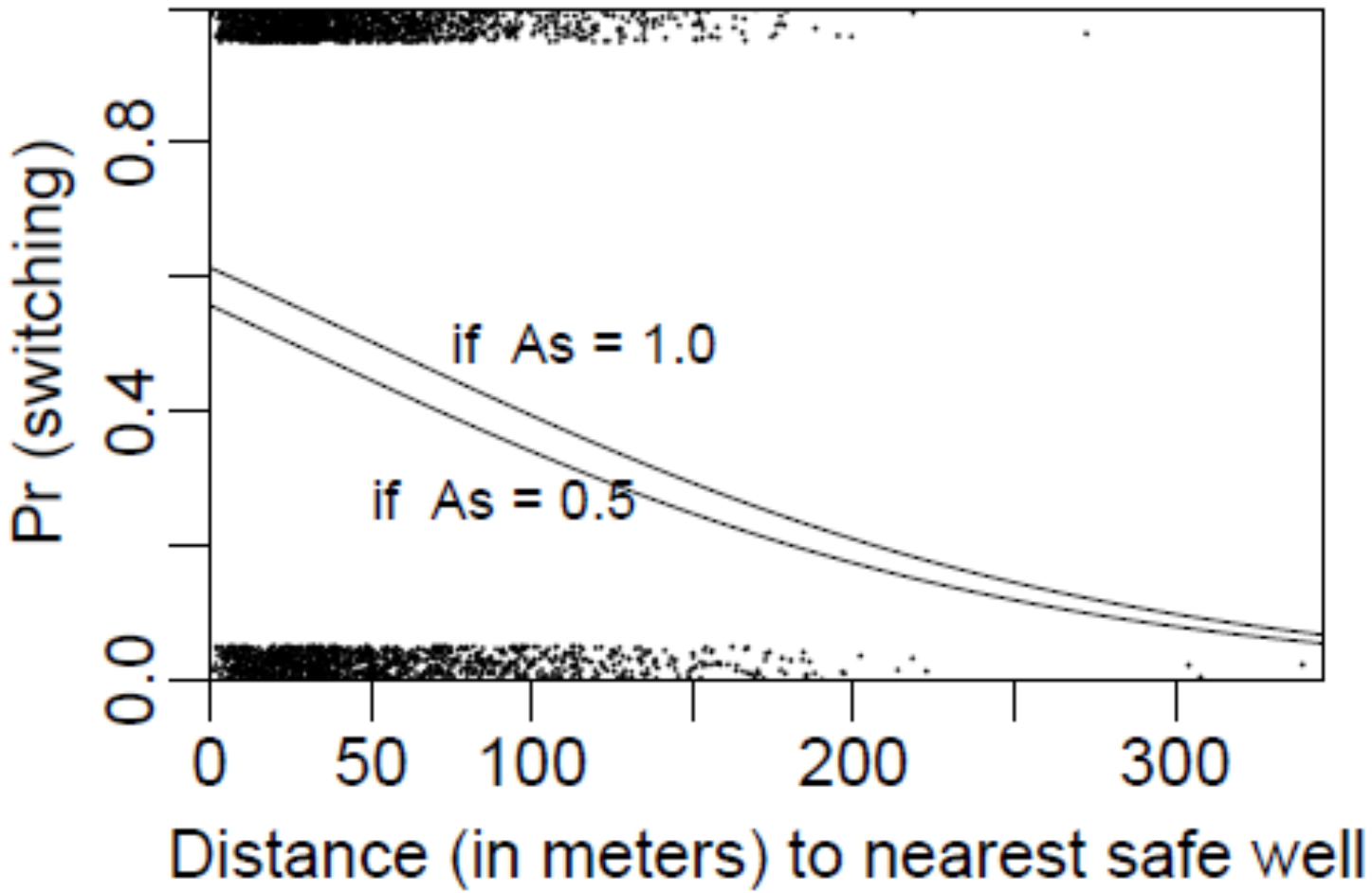
Predicting switching given distance

Pr (switch) = logit^{-1} (0.61 – 0.62*dist100)



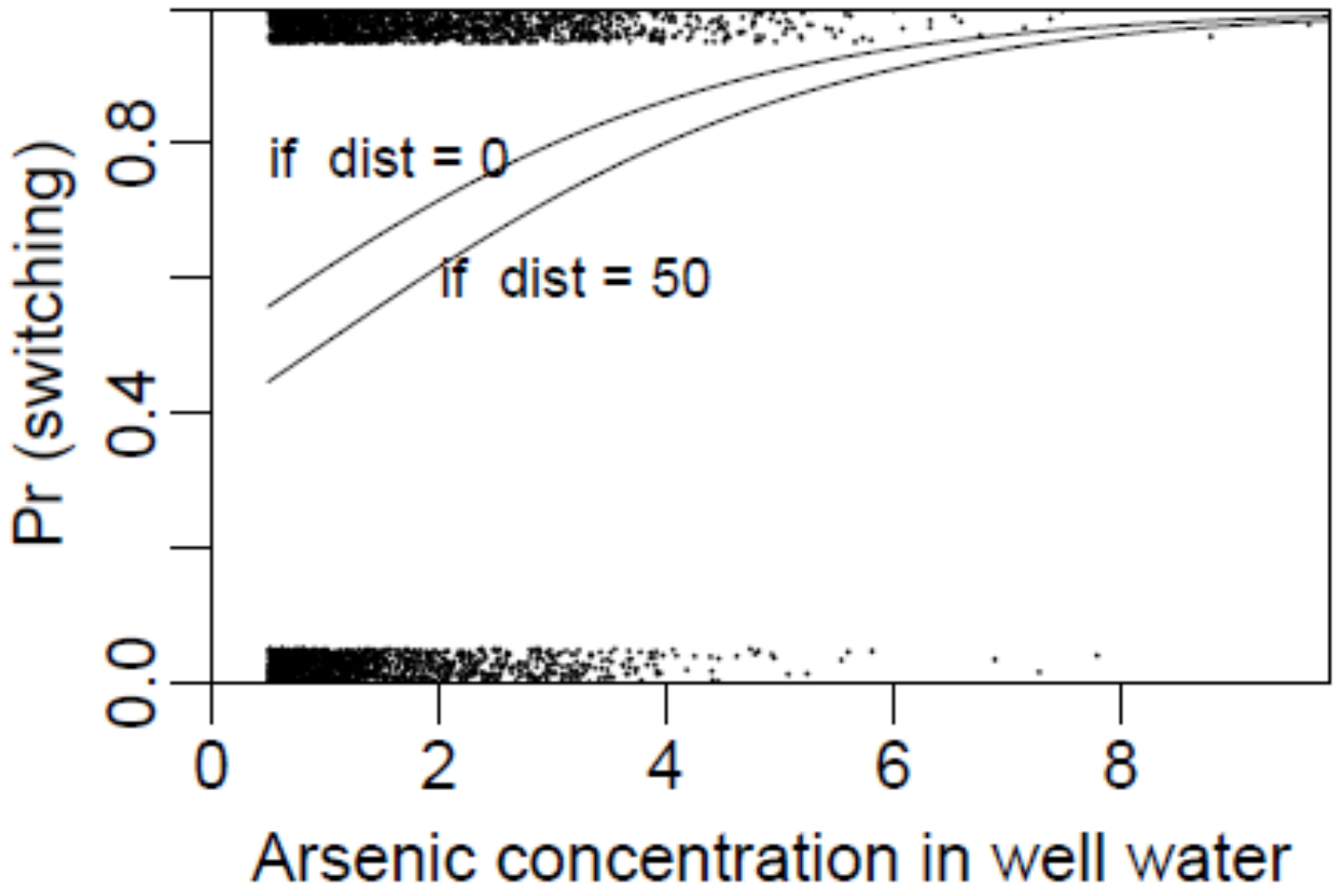
Predicting switching given distance and arsenic level

Pr (switch) = logit^{-1} (0.00 – 0.90*dist100 + 0.46*As)



Predicting switching given distance and arsenic level

Pr (switch) = logit^{-1} (0.00 – 0.90*dist100 + 0.46*As)



Adding the interaction

	coef.est	coef.se
(Intercept)	-0.15	0.12
dist100	-0.58	0.21
arsenic	0.56	0.07
dist100:arsenic	-0.18	0.10

Using centered inputs

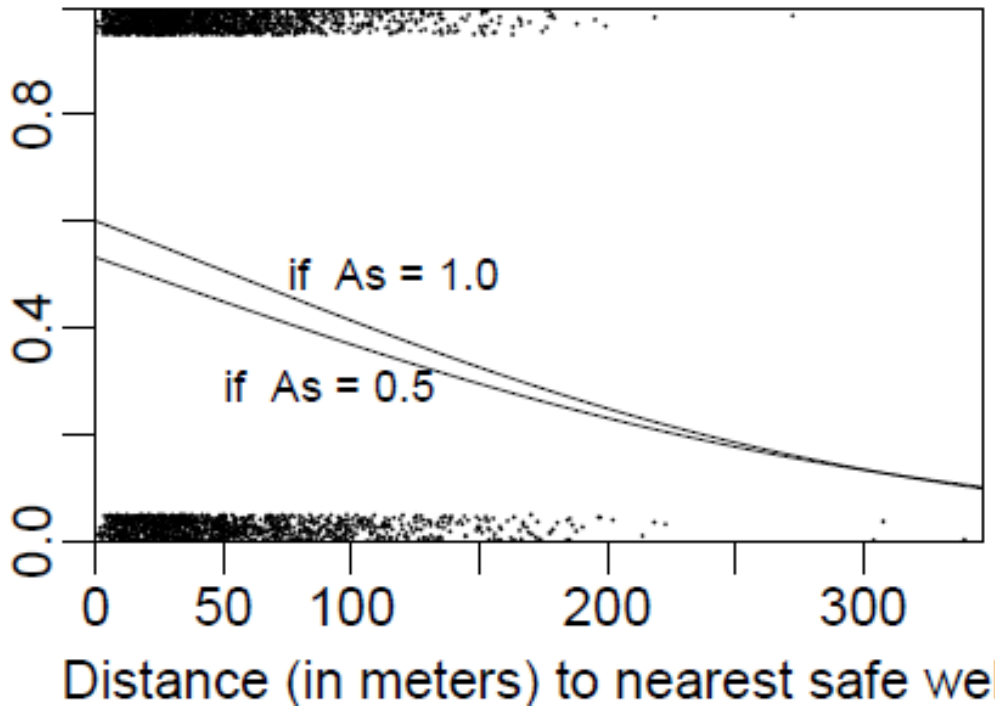
```
c.dist100 <- dist100 - mean(dist100)
c.arsenic <- arsenic - mean(arsenic)
```

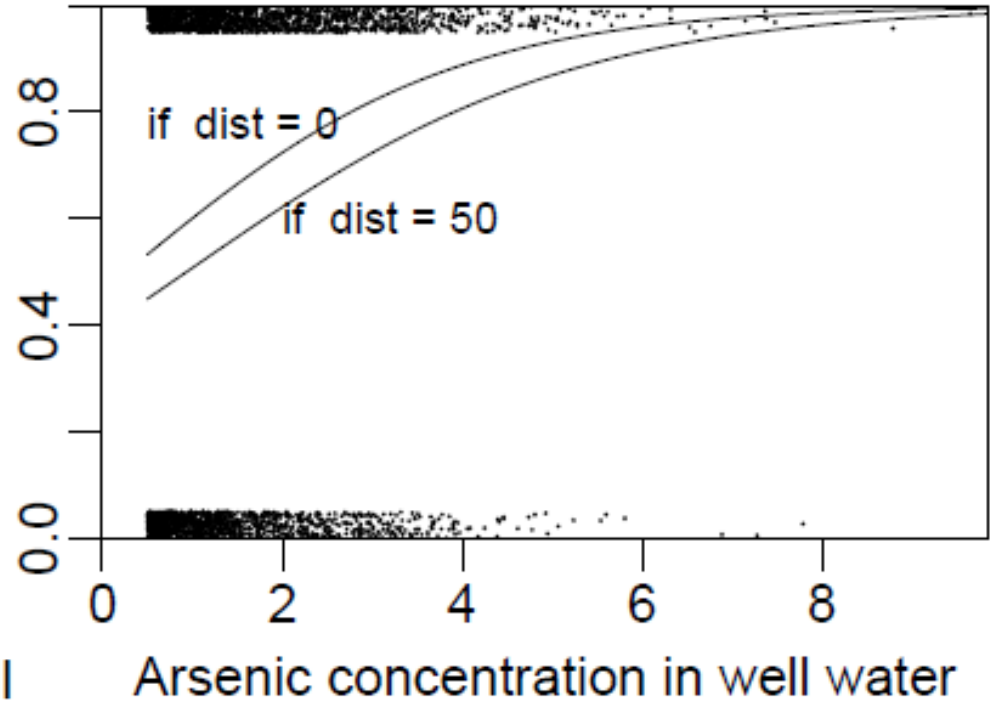
```
                   coef.est coef.se
(Intercept)           -0.15    0.12
dist100               -0.58    0.21
arsenic                0.56    0.07
dist100:arsenic       -0.18    0.10
```

```
                        coef.est coef.se
(Intercept)              0.35     0.04
c.dist100               -0.88     0.10
c.arsenic                0.47     0.04
c.dist100:c.arsenic     -0.18     0.10
```

Fitted model with interactions

- Nonparallel lines (on logit scale)





Adding social predictors

	coef.est	coef.se
(Intercept)	0.20	0.07
c.dist100	-0.88	0.11
c.arsenic	0.48	0.04
c.dist100:c.arsenic	-0.16	0.10
assoc	-0.12	0.08
educ4	0.17	0.04

"assoc" has wrong sign and is not statistically significant, so discard!

After discarding “assoc”

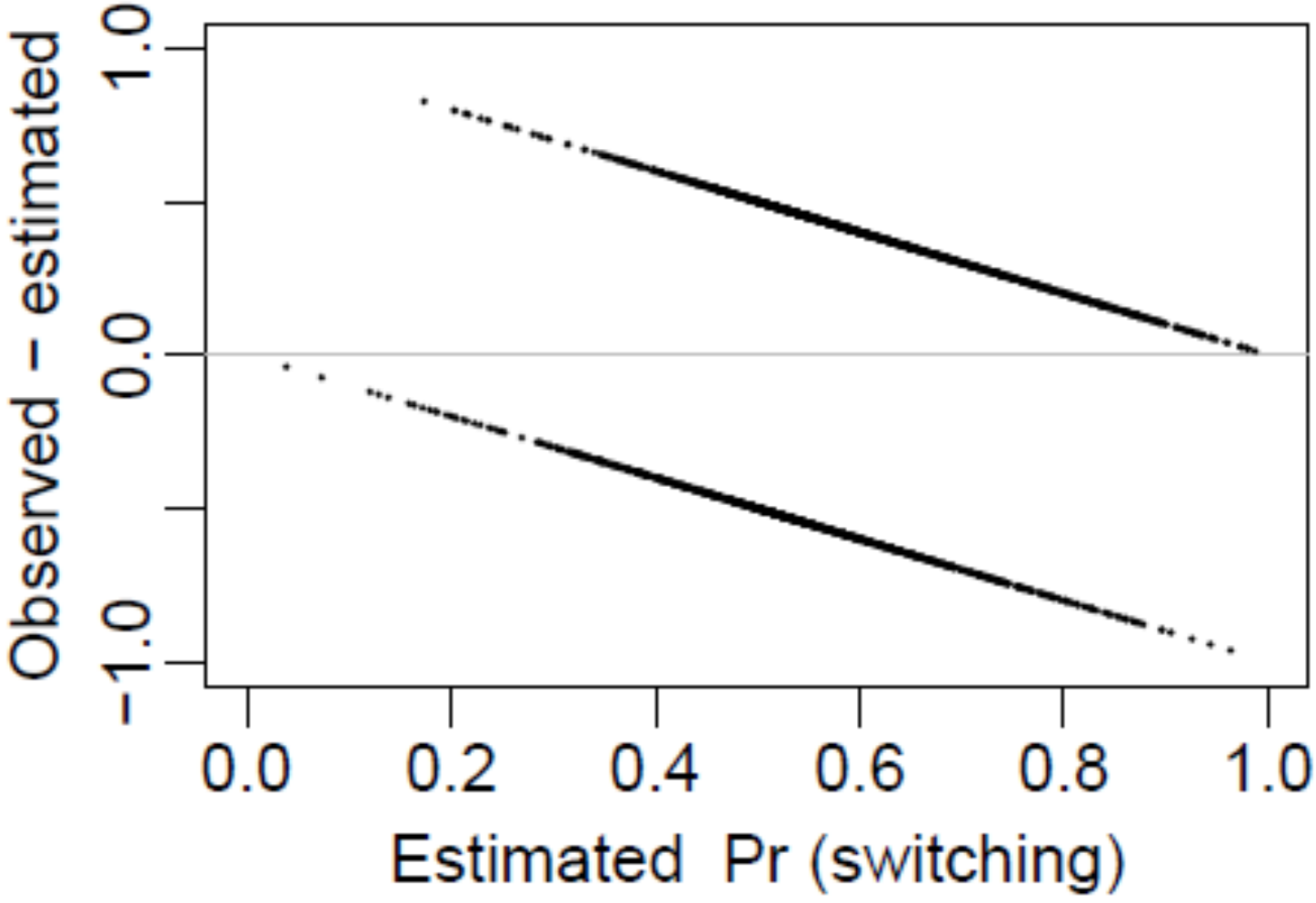
	coef.est	coef.se
(Intercept)	0.15	0.06
c.dist100	-0.87	0.11
c.arsenic	0.48	0.04
c.dist100:c.arsenic	-0.16	0.10
educ4	0.17	0.04

Try more interactions

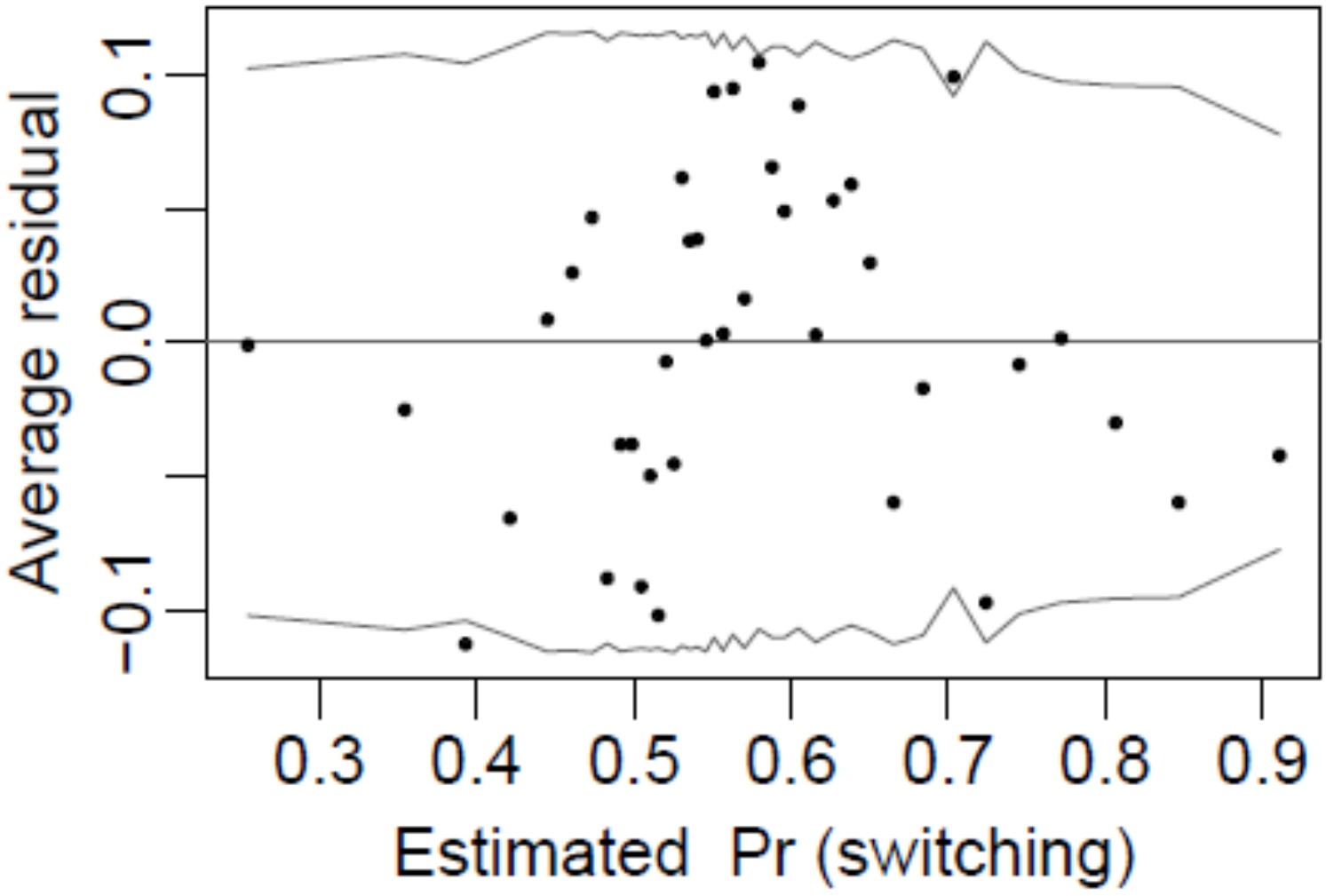
	coef.est	coef.se
(Intercept)	0.36	0.04
c.dist100	-0.90	0.11
c.arsenic	0.49	0.04
c.educ4	0.18	0.04
c.dist100:c.arsenic	-0.12	0.10
c.dist100:c.educ4	0.32	0.11
c.arsenic:c.educ4	0.07	0.04

(Interpret each coefficient)

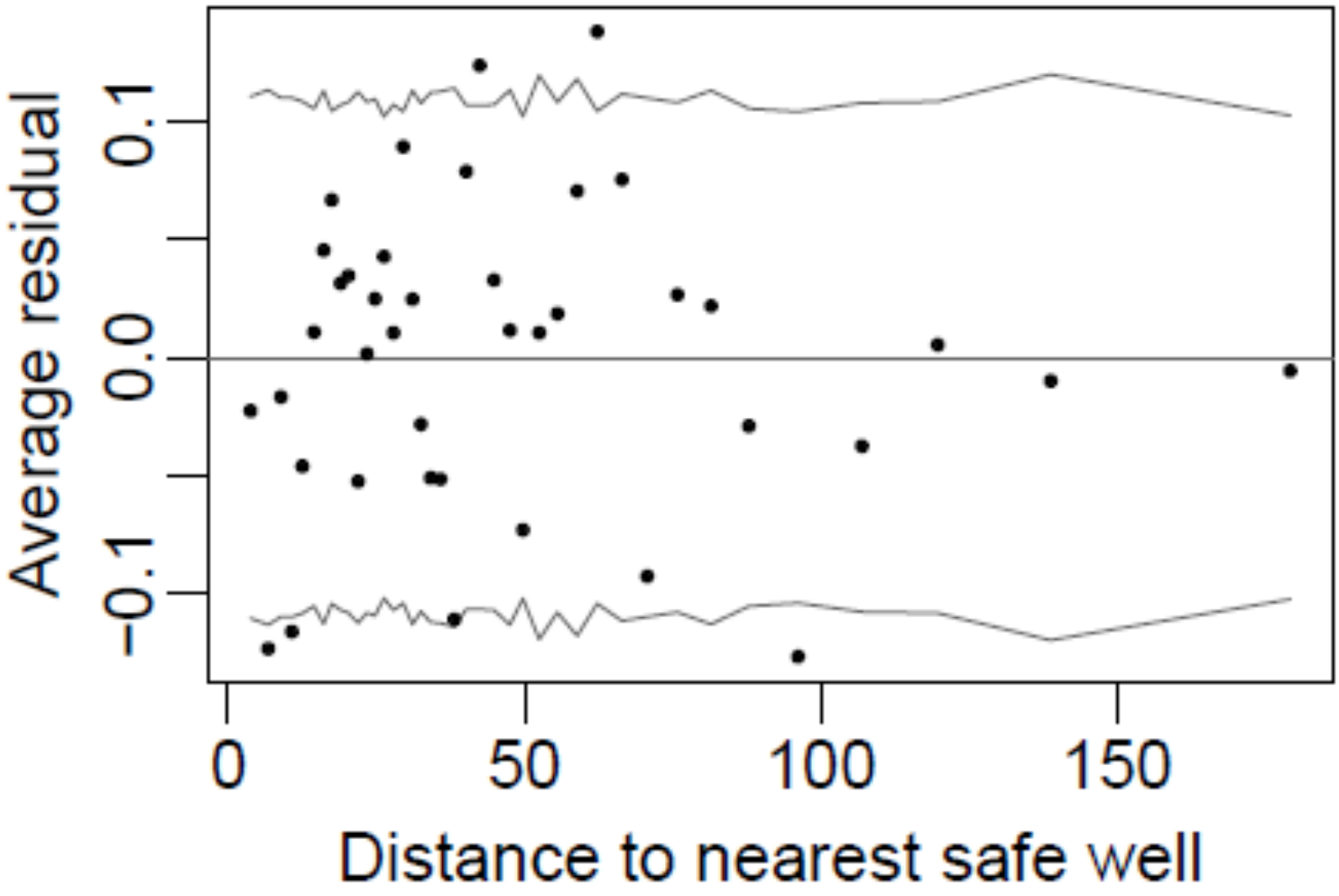
Model checking: residual plot



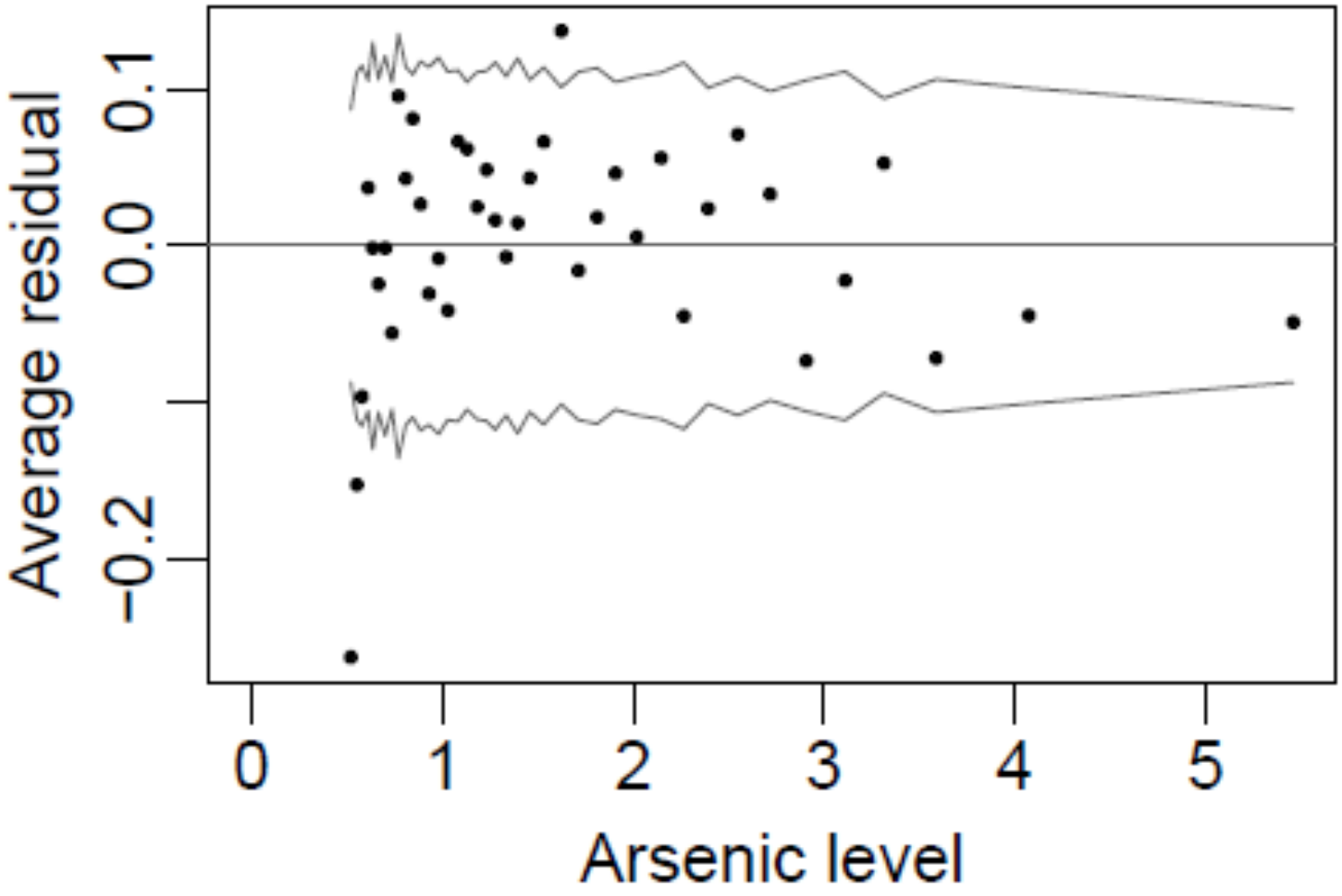
Binned residual plot

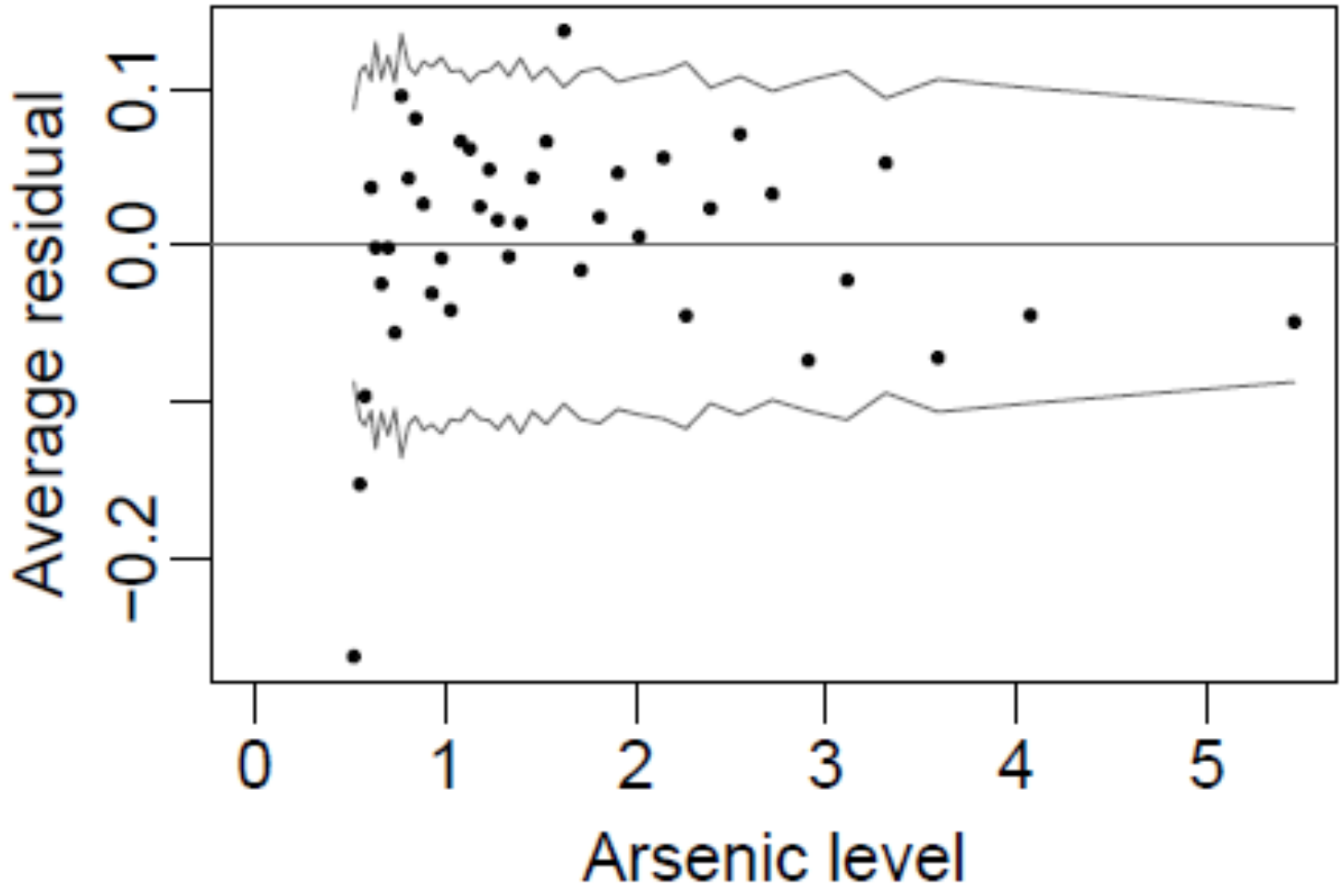


Binned residuals vs. distance



Binned residuals vs. arsenic





Try the log scale:

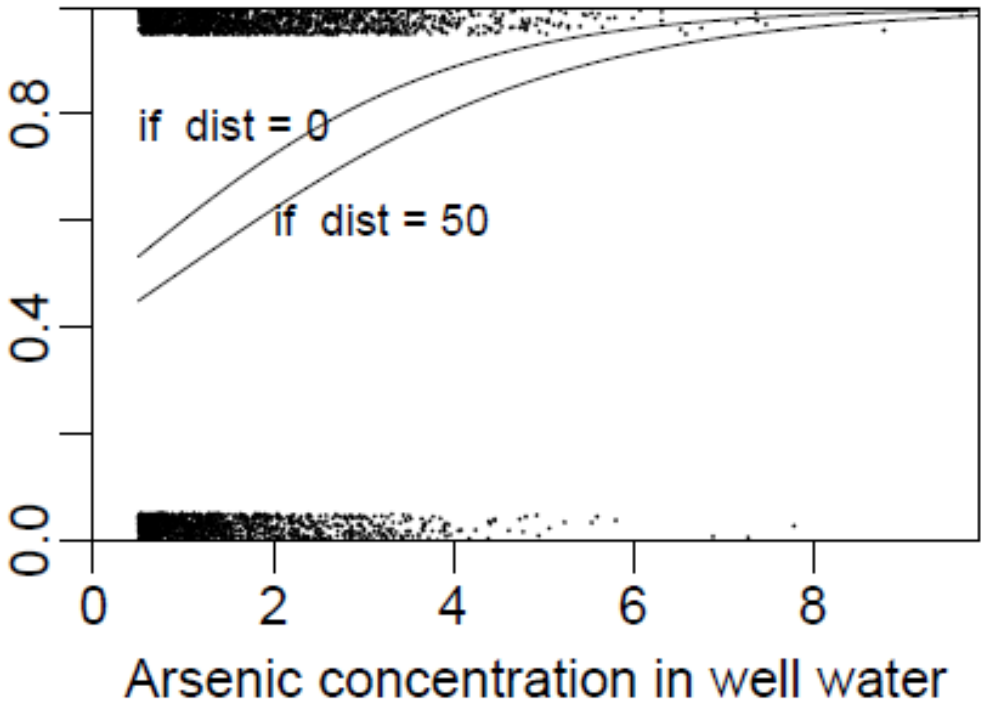
```
log.arsenic <- log(arsenic)
c.log.arsenic <- log.arsenic - mean (log.arsenic)
```

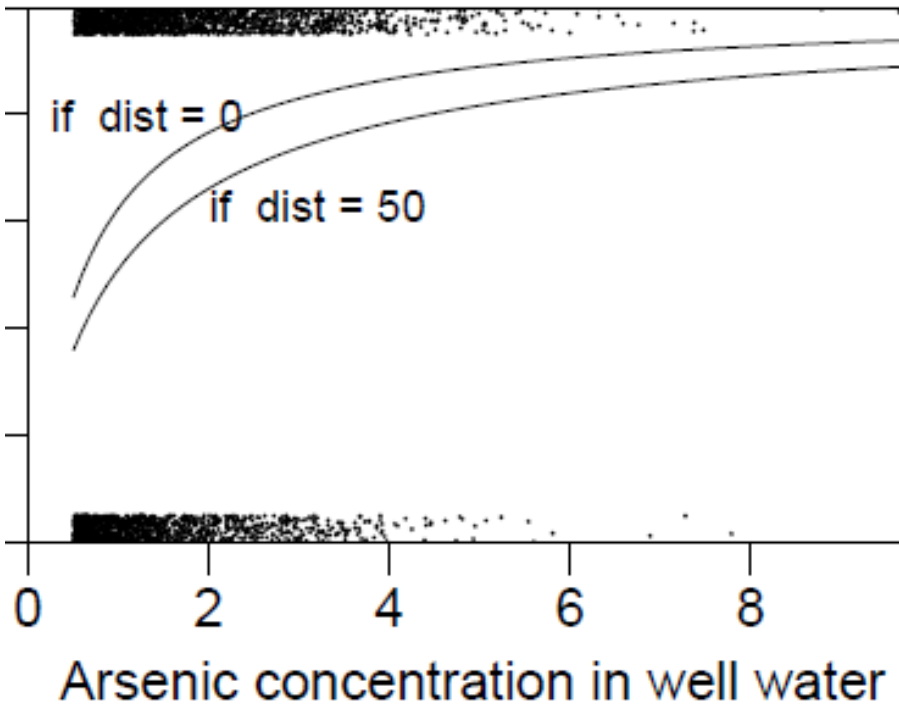
New model

	coef.est	coef.se
(Intercept)	0.35	0.04
c.dist100	-0.98	0.11
c.log.arsenic	0.90	0.07
c.educ4	0.18	0.04
c.dist100:c.log.arsenic	-0.16	0.19
c.dist100:c.educ4	0.34	0.11
c.log.arsenic:c.educ4	0.06	0.07

(Qualitatively similar to earlier model)

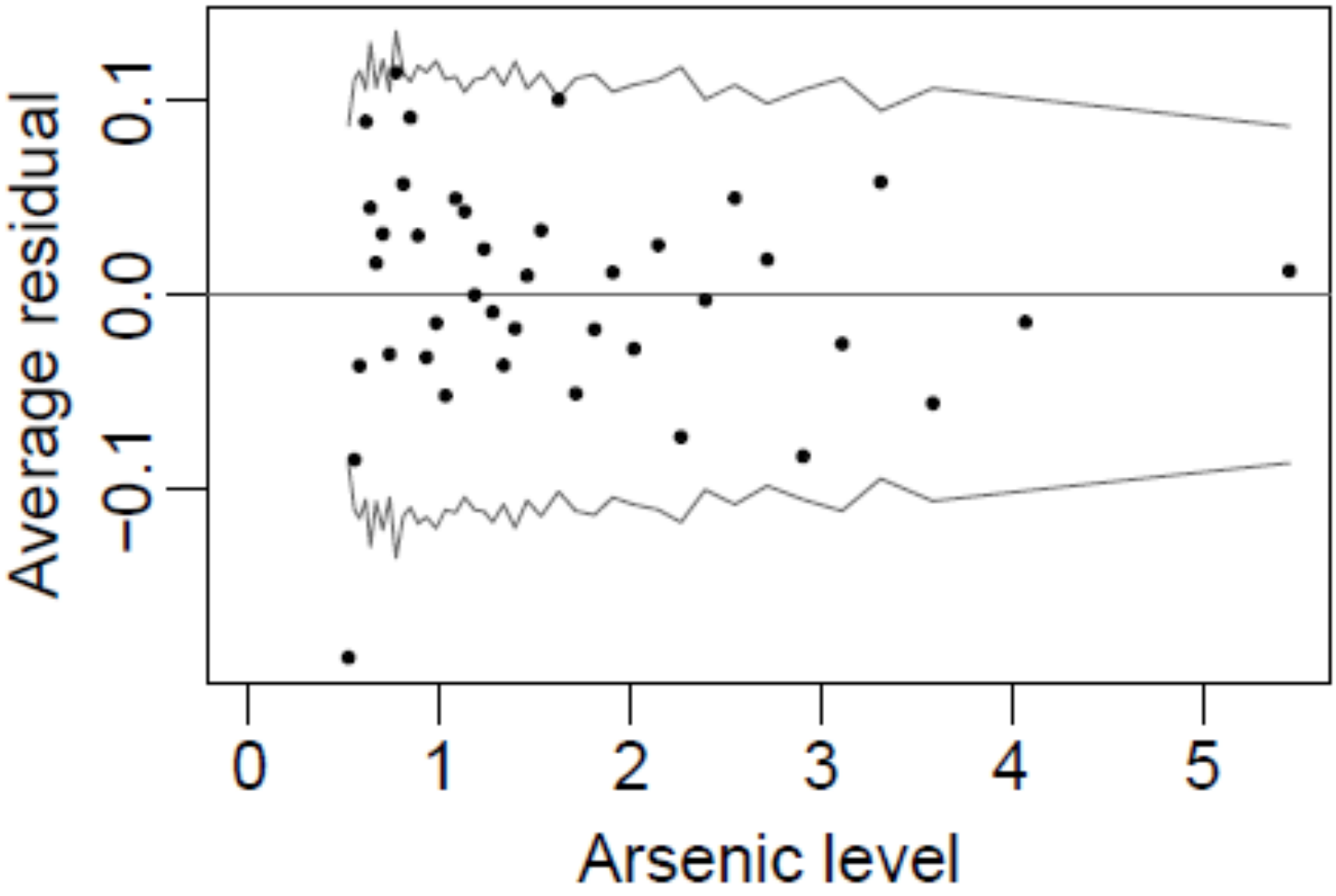
Comparing old and new models





(Education is held constant at its average value)

Binned residuals—new model



(Pretty good, not perfect)

Model for switching

- Distance to walk comes in linearly
  - Does this make sense?
  - Yes
- Current arsenic level comes in on the log scale
  - Does this make sense?
  - Yes (psychologically)
  - No (physically)
- Positive interaction between distance and arsenic
  - Does this make sense?
  - ?

Summary of logistic regression concepts

- Steps of model building
- Graphical model checking
- Continual model improvement

The divide-by-4 rule

- Here is the result of fitting a logistic regression to Republican vote in the 1972 National Election Study:

```
glm(formula = vote ~ income, family=binomial(link="logit"))
            coef.est coef.se
(Intercept)    -1.40    0.19
income          0.33    0.06
 n = 1179, k = 2
```

- Income is on a 1–5 scale.
- Approximately how much more likely is a person in income category 4 to vote Republican, compared to a person in income category 2?
- Give an approximate estimate, standard error, and 95% interval.

Choice models

- Probit regression and the latent-variable model
- Mapping the logit/probit regressions to a formal model of preference
- Treating discrete variables as continuous
- Cautionary example: death sentences and crime

Probit regression

Take the logistic regression coefficients and s.e.'s:

	coef.est	coef.se
(Intercept)	0.35	0.04
c.dist100	-0.98	0.11
c.log.arsenic	0.90	0.07
c.educ4	0.18	0.04
c.dist100:c.log.arsenic	-0.16	0.19
c.dist100:c.educ4	0.34	0.11
c.log.arsenic:c.educ4	0.06	0.07

and just divide everything by 1.6

Latent-data model

- Logit: $\Pr(y_i = 1) = \text{logit}^{-1}(X_i\beta)$

$$
\begin{aligned}
y_i &= \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i < 0 \end{cases} \\
z_i &= X_i\beta + \epsilon_i, \;\; \epsilon_i \sim \text{N}(0, 1.6^2)
\end{aligned}
$$

- Probit: Same model, but $\epsilon_i \sim \text{N}(0, 1)$
- Probit coefs are logit coefs divided by 1.6

Taking the latent data seriously

- Example: modeling political preferences

Pr (person i votes Republican) = $\text{logit}^{-1}(X_i\beta)$

- Latent data $z_i = X_i\beta + \epsilon_i$
- Interpret z_i as a continuous attitude
  - Can be measured using "feeling thermometer" questions
  - Can be modeled as being stable across issues or over time

Logistic regression as a formal model of choice

- Well-switching model:

	coef.est	coef.se
(Intercept)	0.61	0.06
dist100	-0.62	0.10

- Decision for household i with As level As_i
  - $a_i As_i$ = benefit of switching to a safe well
  - $b_i + c_i x_i$ = cost of switching to a well at distance x_i
  - $Pr(switch) = Pr(y_i=1) = Pr\ (a_i As_i > b_i + c_i x_i)$

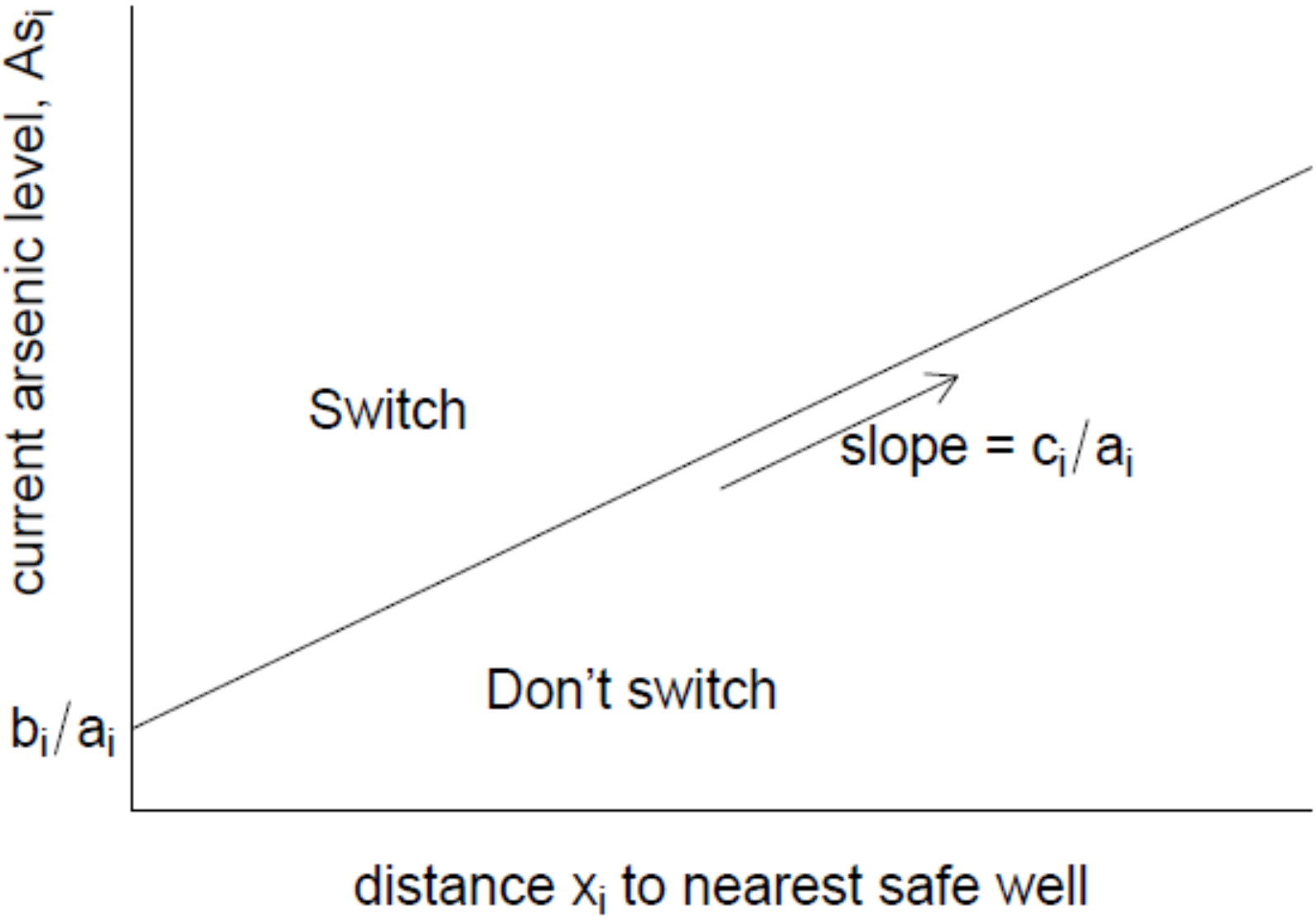
    $= Pr\ ((a_i As_i - b_i)/c_i > x_i)$

	coef.est	coef.se
(Intercept)	0.61	0.06
dist100	-0.62	0.10

- Decision for household i with As level As_i
  - $a_i As_i$ = benefit of switching to a safe well
  - $b_i + c_i x_i$ = cost of switching to a well at distance x_i
  - $\Pr(\text{switch}) = \Pr(y_i=1) = \Pr(a_i As_i > b_i + c_i x_i)$

    $= \Pr((a_i As_i - b_i)/c_i > x_i)$
- All depends on distribution of $(a_i As_i - b_i)/c_i$
- The net benefit of switching, divided by the cost per distance traveled to a new well
  - If $(a_i As_i - b_i)/c_i$ has an approximately normal dist in the population, then the logit/probit model makes sense

Discrete choice model



Treating discrete variables as continuous

- Binary variables
  - Often you can just fit with a linear model
- Ordered categories
  - Strong Dem, Dem, Weak Dem, ... , Strong Rep
  - Just treat them as 1-7 on a continuous scale

Cautionary example: death sentences and crime

- Deterrent effect of capital punishment
- Historically, when death penalty comes, crime rates go down
- Data show this at national and state levels
- Death penalty typically goes with other crime-fighting measures
- Regression predicting crime rate from death-sentencing rate and other predictors . . .
- . . . utility model of the choices of potential murderers

Deterrent effect of the death penalty

From New York Times article, 18 Nov 2007:

> *To many economists, then, it follows inexorably that there will be fewer murders as the likelihood of execution rises.*
>
> *"I am definitely against the death penalty on lots of different grounds," said Joanna M. Shepherd, a law professor at Emory with a doctorate in economics who wrote or contributed to several studies. "But I do believe that people respond to incentives."... The recent studies are, some independent observers say, of good quality, given the limitations of the available data.*
>
> *"These are sophisticated econometricians who know how to do multiple regression analysis at a pretty high level," Professor Weisberg of Stanford said.*

Faith in "sophisticated econometricians"

Robert Weisberg
Edwin E. Huddleson, Jr. Professor of Law

Biography

Faculty Co-Director, Stanford Criminal Justice Center

Robert Weisberg '79 works primarily in the field of criminal justice, writing and teaching in the areas of criminal law, criminal procedure, white collar crime, and sentencing policy. He also founded and now serves as faculty co-director of the Stanford Criminal Justice Center (SCJC), which promotes and coordinates research and public policy programs on criminal law and the criminal justice system, including institutional examination of the police and correctional systems. Professor Weisberg was a consulting attorney for the NAACP Legal Defense Fund and the California Appellate Project, where he worked on death penalty litigation in the state and federal courts. In addition, he served as a law clerk to Justice Potter Stewart of the U.S. Supreme Court and Judge J. Skelly Wright of the U.S. Court of Appeals for the District of Columbia Circuit. In 1979, Professor Weisberg received his J.D. from Stanford Law School, where he served as President of the Stanford Law Review. Professor Weisberg is a two-time winner of the law school's John Bingham Hurlbut Award for Excellence in Teaching.

Before joining the Stanford Law School faculty in 1981, Professor Weisberg received a PhD in English at Harvard and was a tenured English professor at

weisberg@stanford.edu

650 723.0612

Curriculum Vitae

Education

BA, City College of New York, 1966

MA, 1967; PhD (English), 1971, Harvard University Graduate School of Arts and Sciences

JD, Stanford Law School, 1979

Model of death penalty deterrence

From "Deterrence versus Brutalization," by Joanna M. Shepherd:

For technically-inclined readers, I express the system symbolically:

$$M_{i,t} = \alpha_i + \beta_1 Pa_{i,t} + \beta_2 Ps \mid a_{i,t} + \beta_3 SD_i Pe \mid s_{i,t} + \gamma_1 Z_{i,t} + \gamma_2 TD_t + \varepsilon_{i,t}, \quad (1)$$

$$Pa_{i,t} = \phi_{1,i} + \phi_2 M_{i,t} + \phi_3 PE_{i,t} + \phi_4 TD_t + \varsigma_{i,t}, \quad (2)$$

$$Ps \mid a_{i,t} = \theta_{1,i} + \theta_2 M_{i,t} + \theta_3 JE_{i,t} + \theta_4 PI_{i,t} + \theta_5 PA_{i,t} + \theta_6 TD_t + \xi_{i,t}, \quad (3)$$

$$Pe \mid s_{i,t} = \psi_{1,i} + \psi_2 M_{i,t} + \psi_3 JE_{i,t} + \psi_4 PI_{i,t} + \psi_5 TD_t + \zeta_{i,t}, \quad (4)$$

It continues:

> *The first equation measures the response of the behavior of criminals to the deterrent factors while controlling for a series of other factors found in the series . . .*

Don't take the choice model too seriously

- Death penalty affects incentives of potential murderers
- Also affects incentives of . . .
  - Judges
  - Juries
  - Prosecutors
  - General public

Summary of choice models

- Correspondence between descriptive regression models and underlying choice models
- But don't take the choice models too seriously!

Summary of regression modeling

- Build models, don't take data right out of the box
- Linear and logarithmic transformations
- Model checking: exploratory data analysis
- Understanding your fitted models: exploratory model analysis