Fitting discrete-data regression models in social science

Andrew Gelman

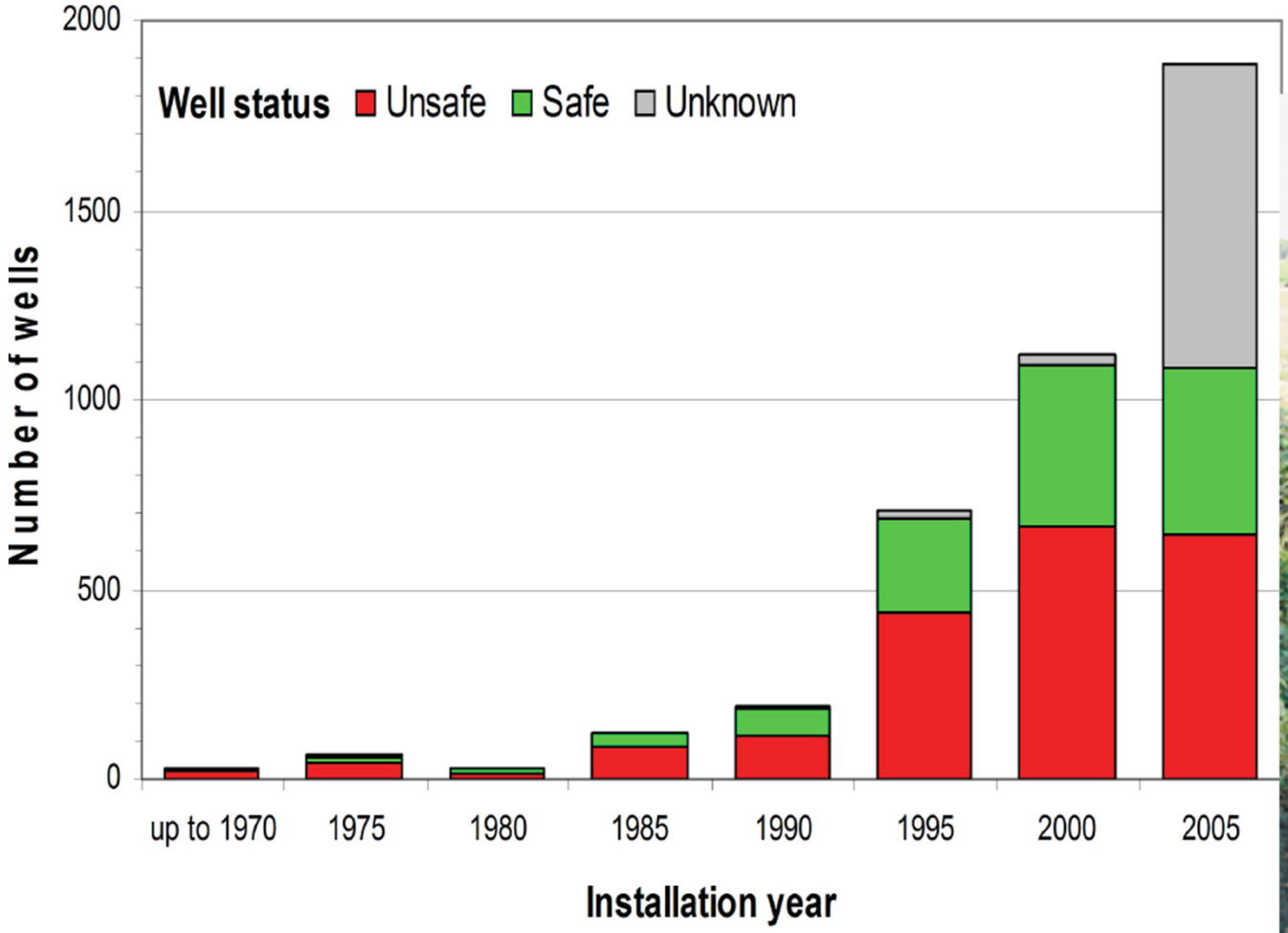
Dept. of Statistics and Dept. of Political Science
Columbia University

For Greg Wawro's class, 7 Oct 2010

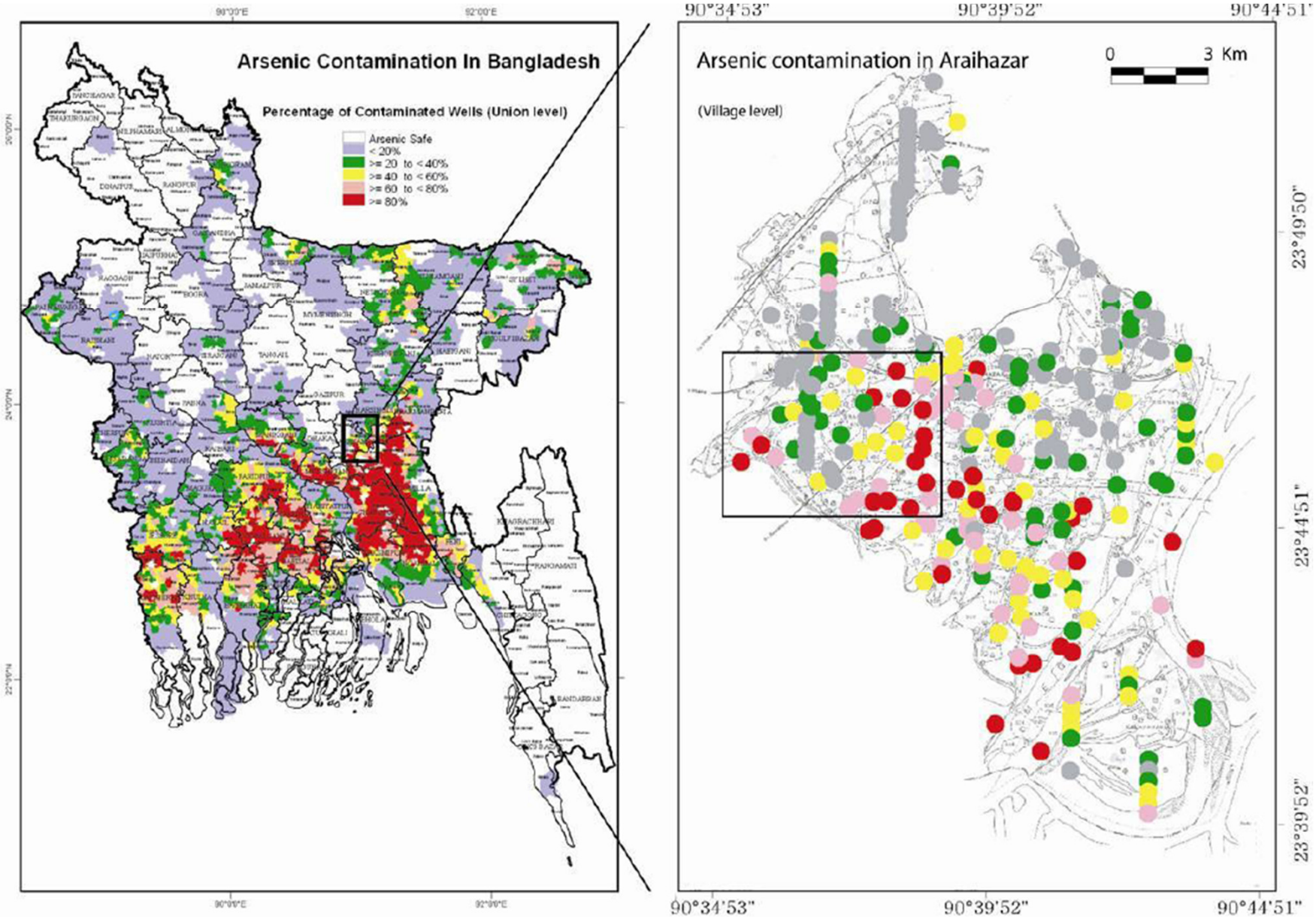
Today's class

- Example: wells in Bangladesh
  - Building a logistic regression model
  - Logistic regression with interactions
  - Evaluating, checking, and comparing models
  - Probit regression and the latent-variable model
  - Mapping the logit/probit regressions to a formal model of preferences
- Related methodological topics

Natural arsenic in well water



Mix of high and low-arsenic wells



Digging new wells

Today's class

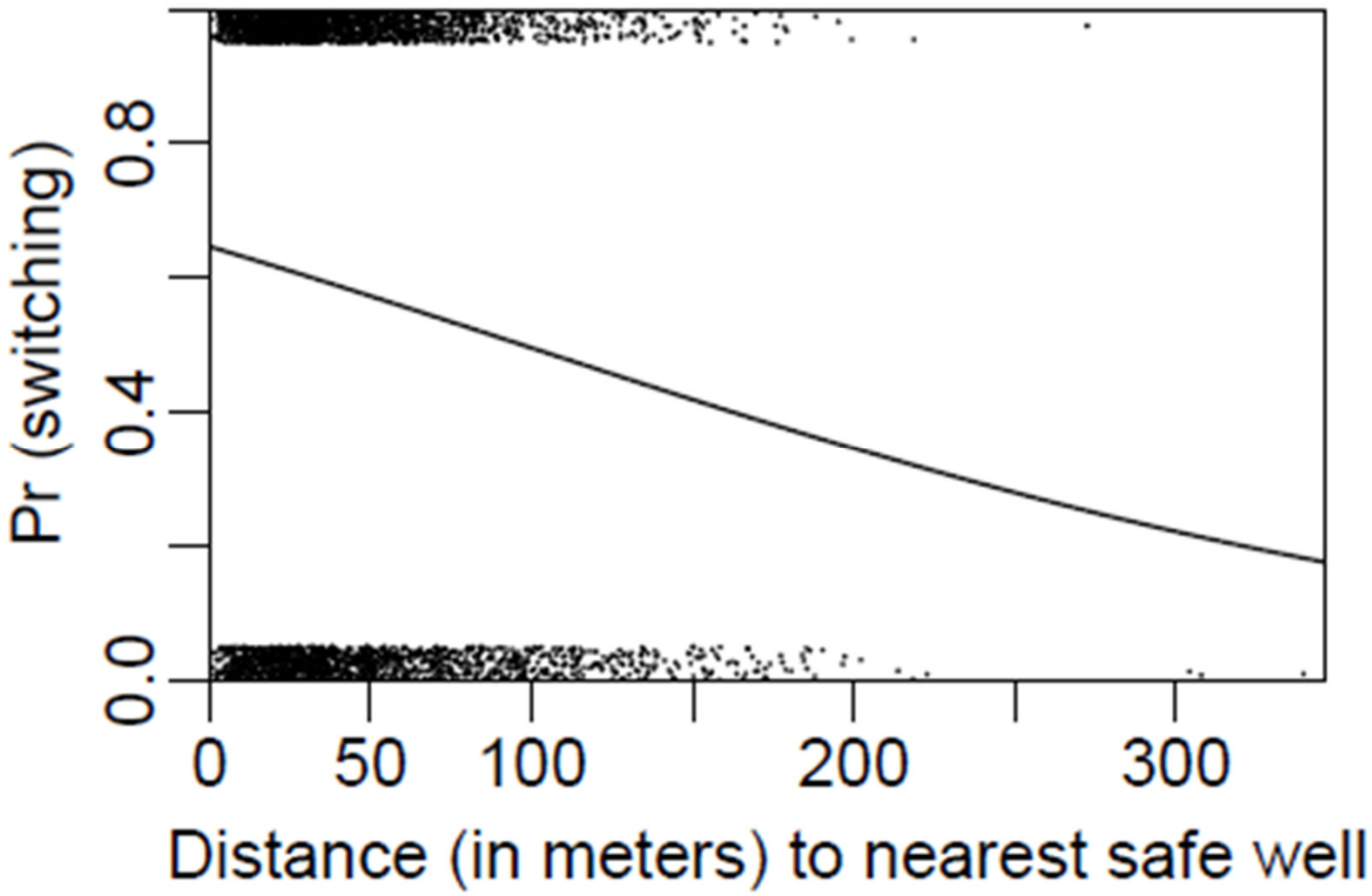
- Example: wells in Bangladesh
  - Building a logistic regression model
  - Logistic regression with interactions
  - Evaluating, checking, and comparing models
  - Probit regression and the latent-variable model
  - Mapping the logit/probit regressions to a formal model of preferences
- Related methodological topics

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)

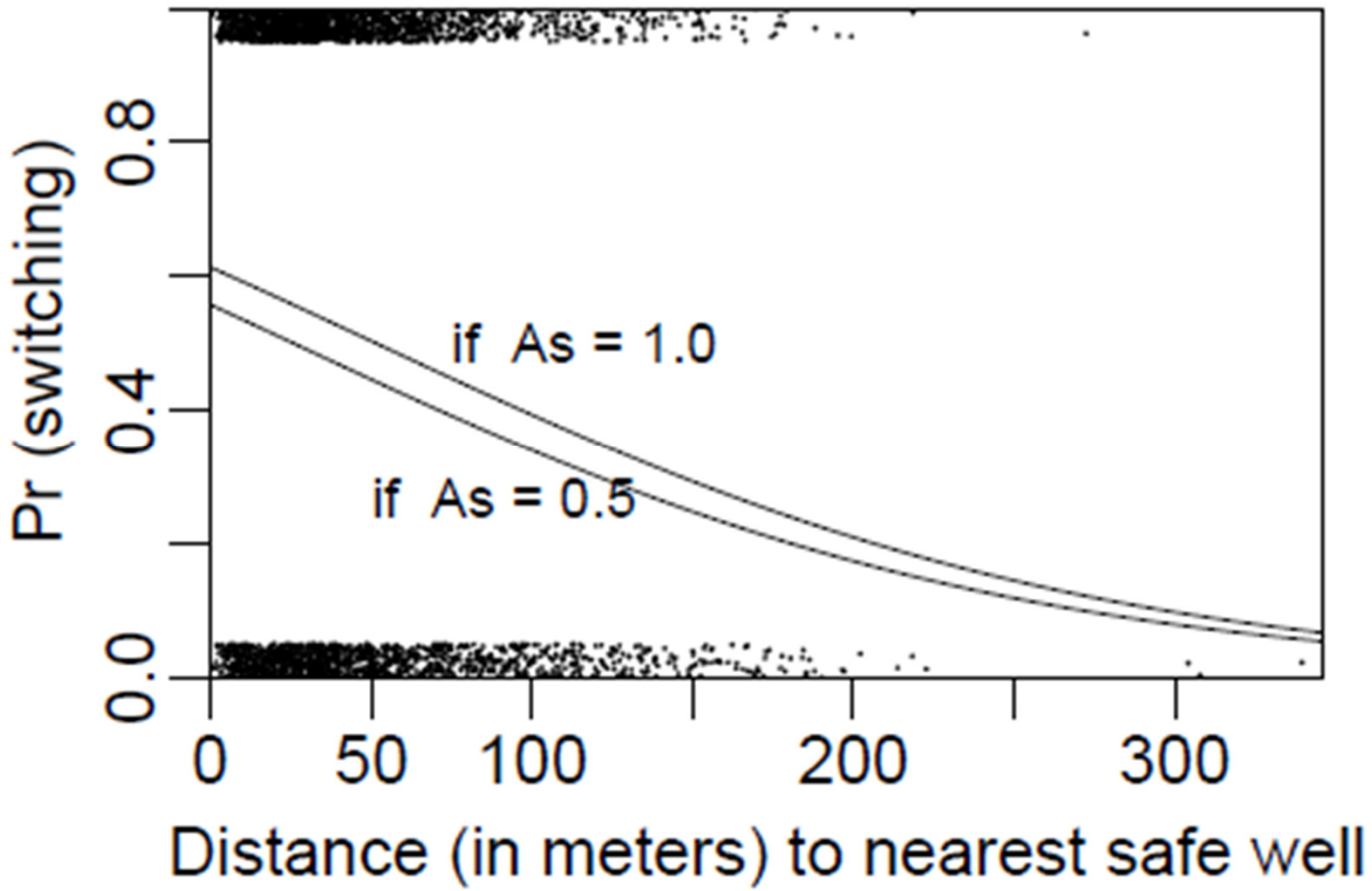
Predicting switching given distance

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



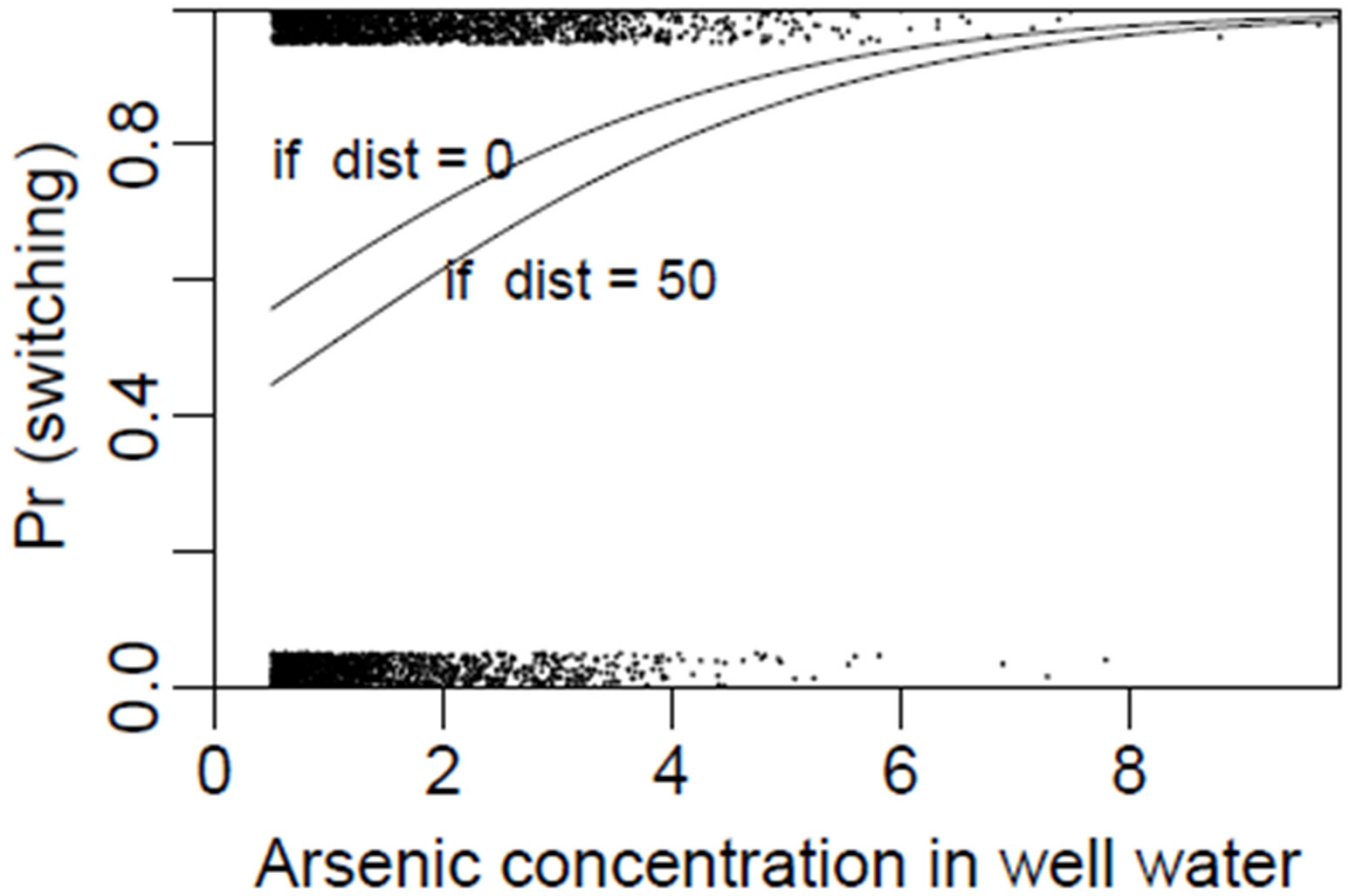
Predicting switching given distance and arsenic level

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



Predicting switching given distance and arsenic level

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



Today's class

- Example: wells in Bangladesh
  - Building a logistic regression model
  - Logistic regression with interactions
  - Evaluating, checking, and comparing models
  - Probit regression and the latent-variable model
  - Mapping the logit/probit regressions to a formal model of preferences
- Related methodological topics

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

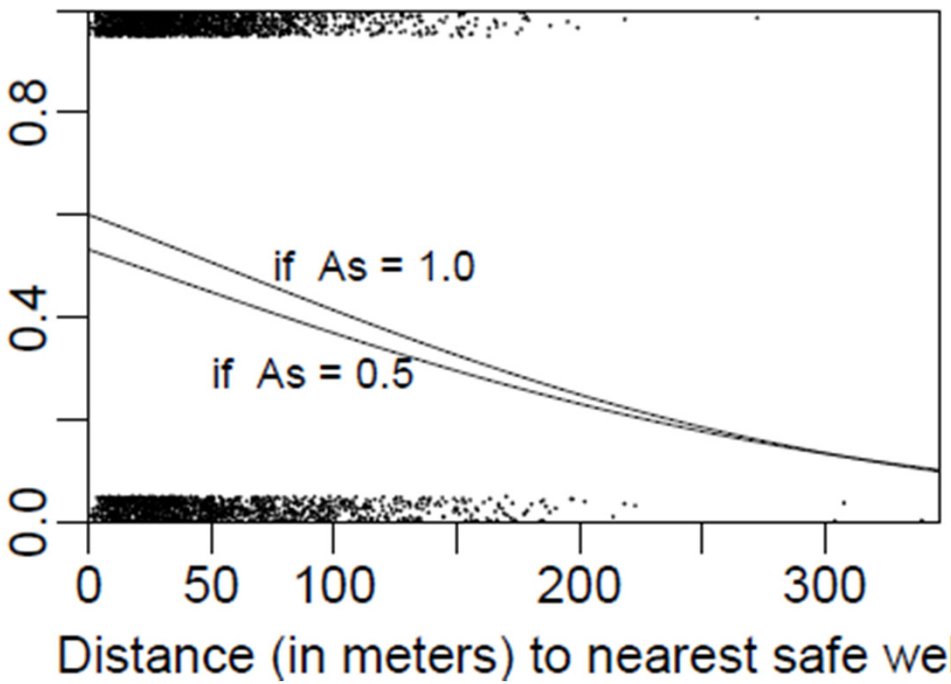
```
c.dist100 <- dist100 - mean(dist100)
c.arsenic <- arsenic - mean(arsenic)
```

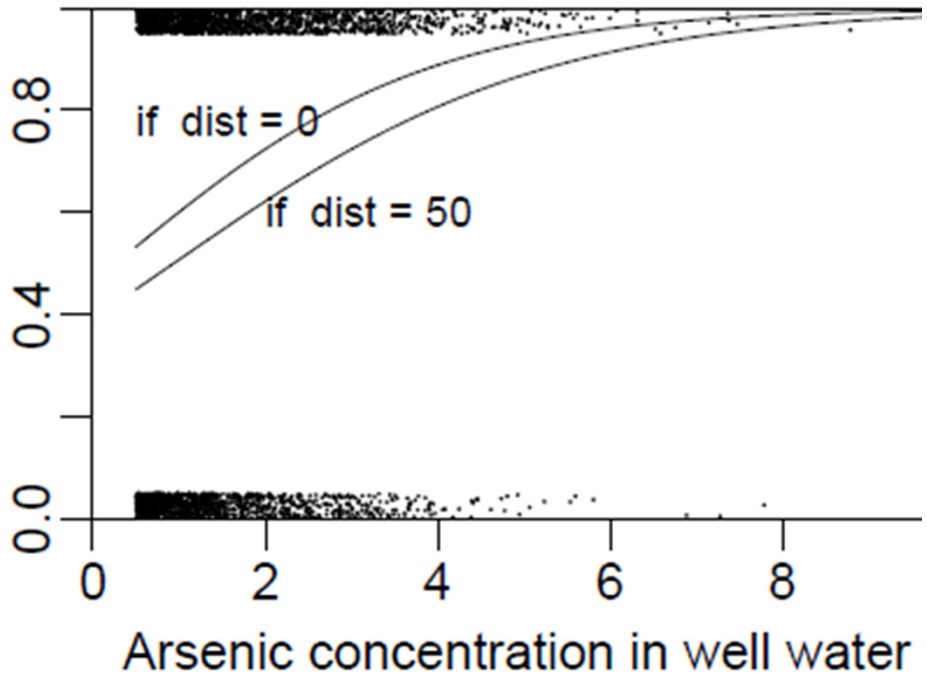
	coef.est	coef.se
(Intercept)	-0.15	0.12
dist100	-0.58	0.21
arsenic	0.56	0.07
dist100:arsenic	-0.18	0.10

	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

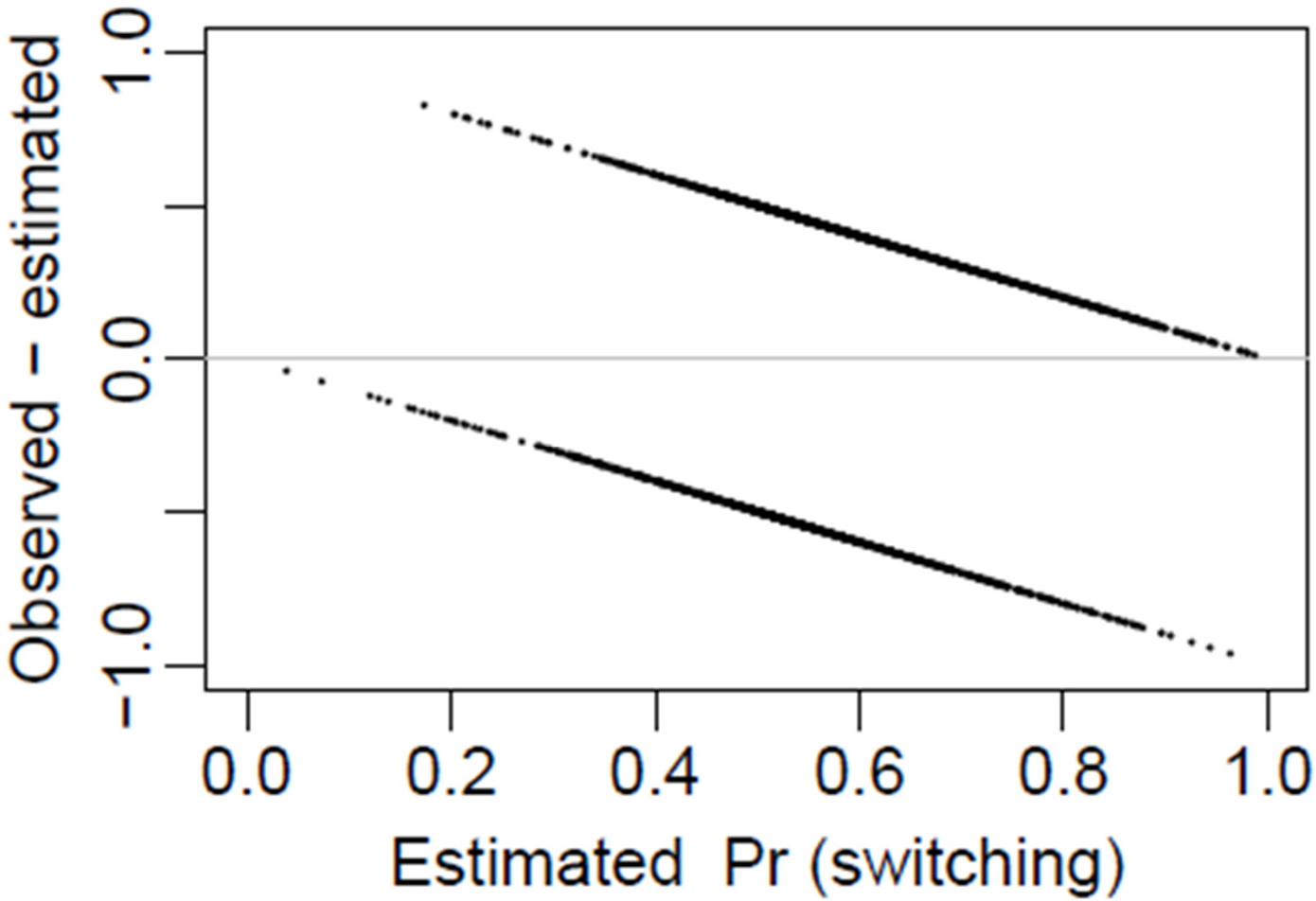
	coef.est	coef.se
(Intercept)	0.36	0.04
c.dist100	-0.90	0.11
c.arsenic	0.49	0.04
c.educ4	0.18	0.04
c.dist100:c.arsenic	-0.12	0.10
c.dist100:c.educ4	0.32	0.11
c.arsenic:c.educ4	0.07	0.04

(Interpret each coefficient)

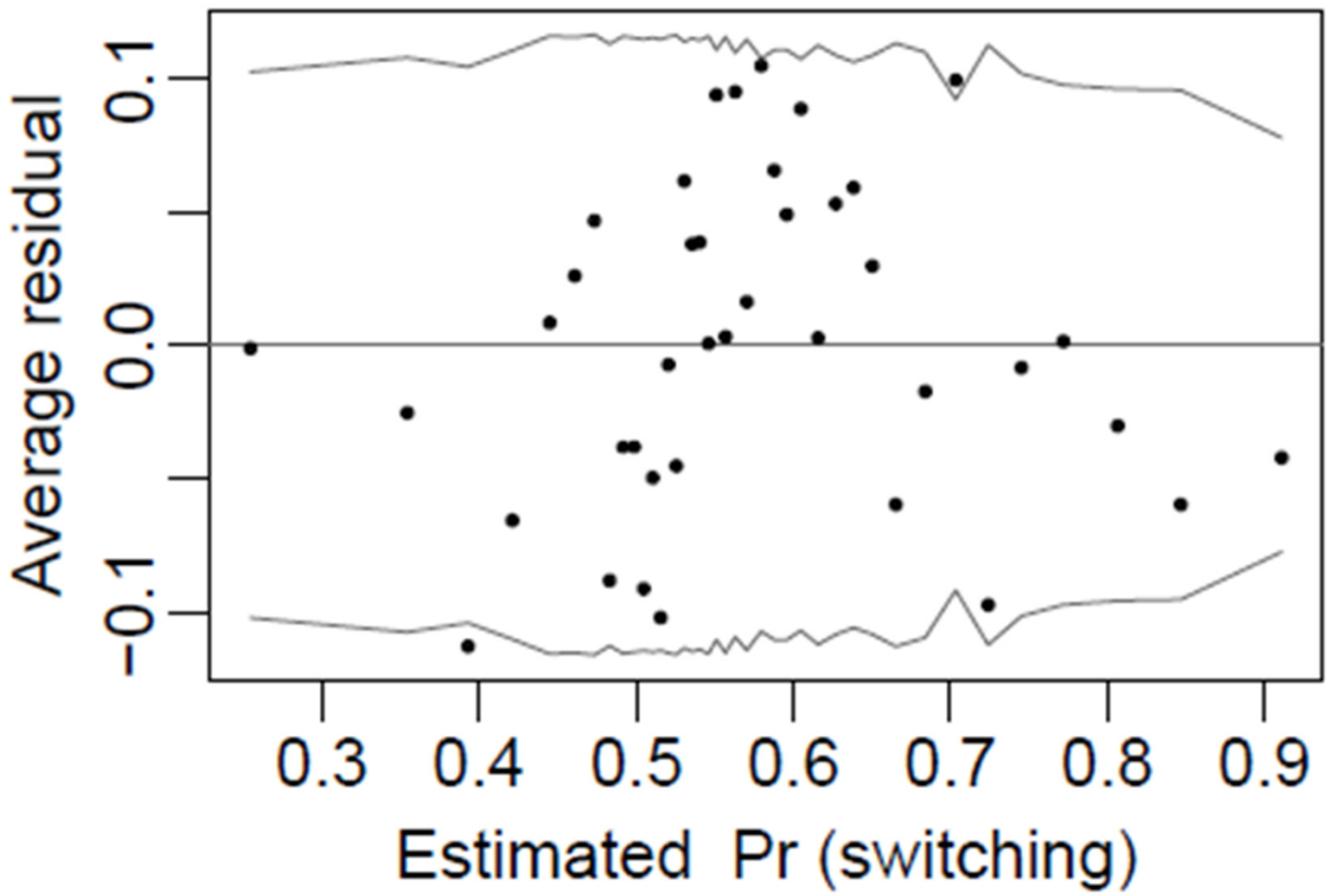
Today's class

- Example: wells in Bangladesh
  - Building a logistic regression model
  - Logistic regression with interactions
  - Evaluating, checking, and comparing models
  - Probit regression and the latent-variable model
  - Mapping the logit/probit regressions to a formal model of preferences
- Related methodological topics

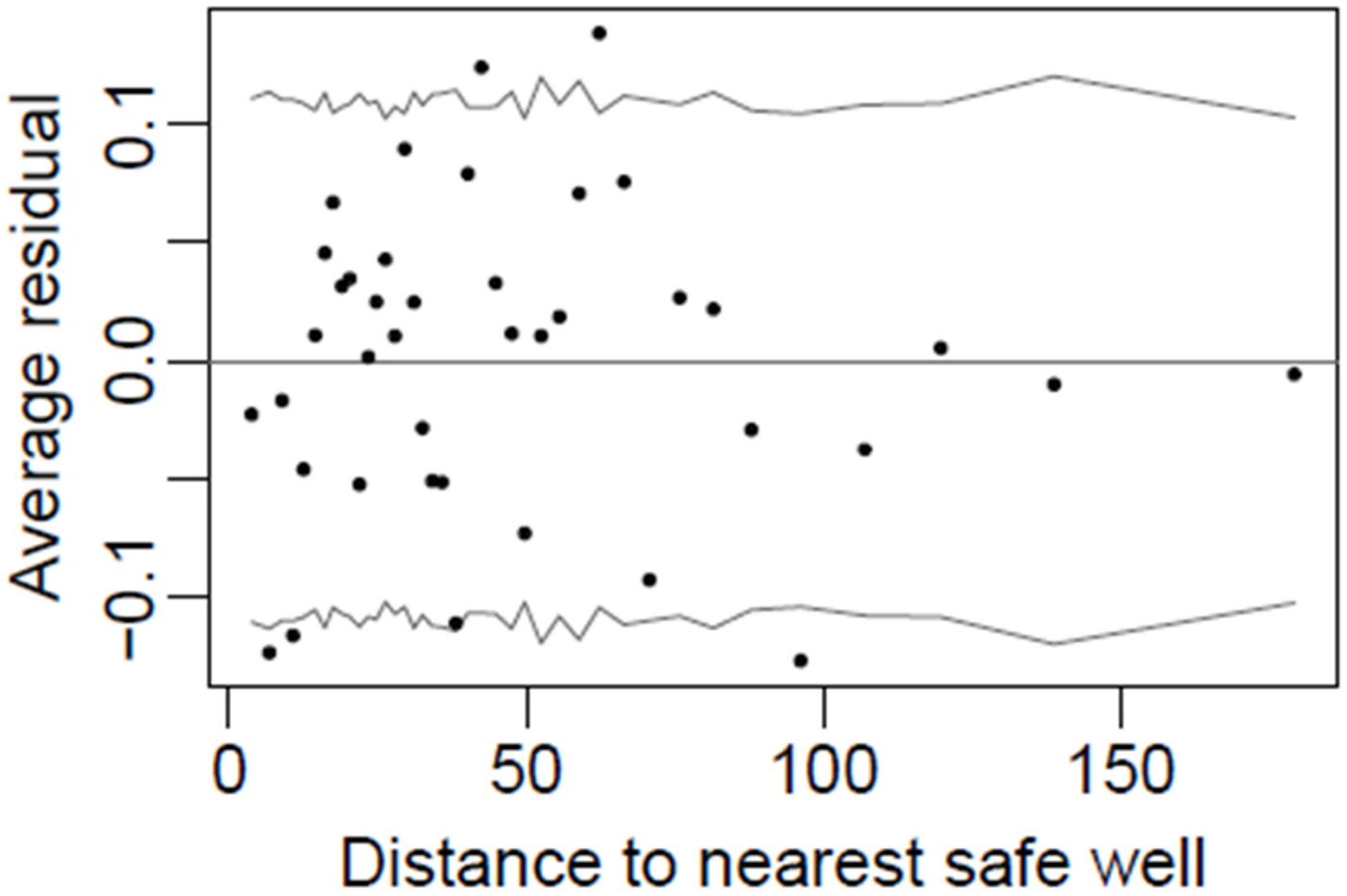
Residual plot



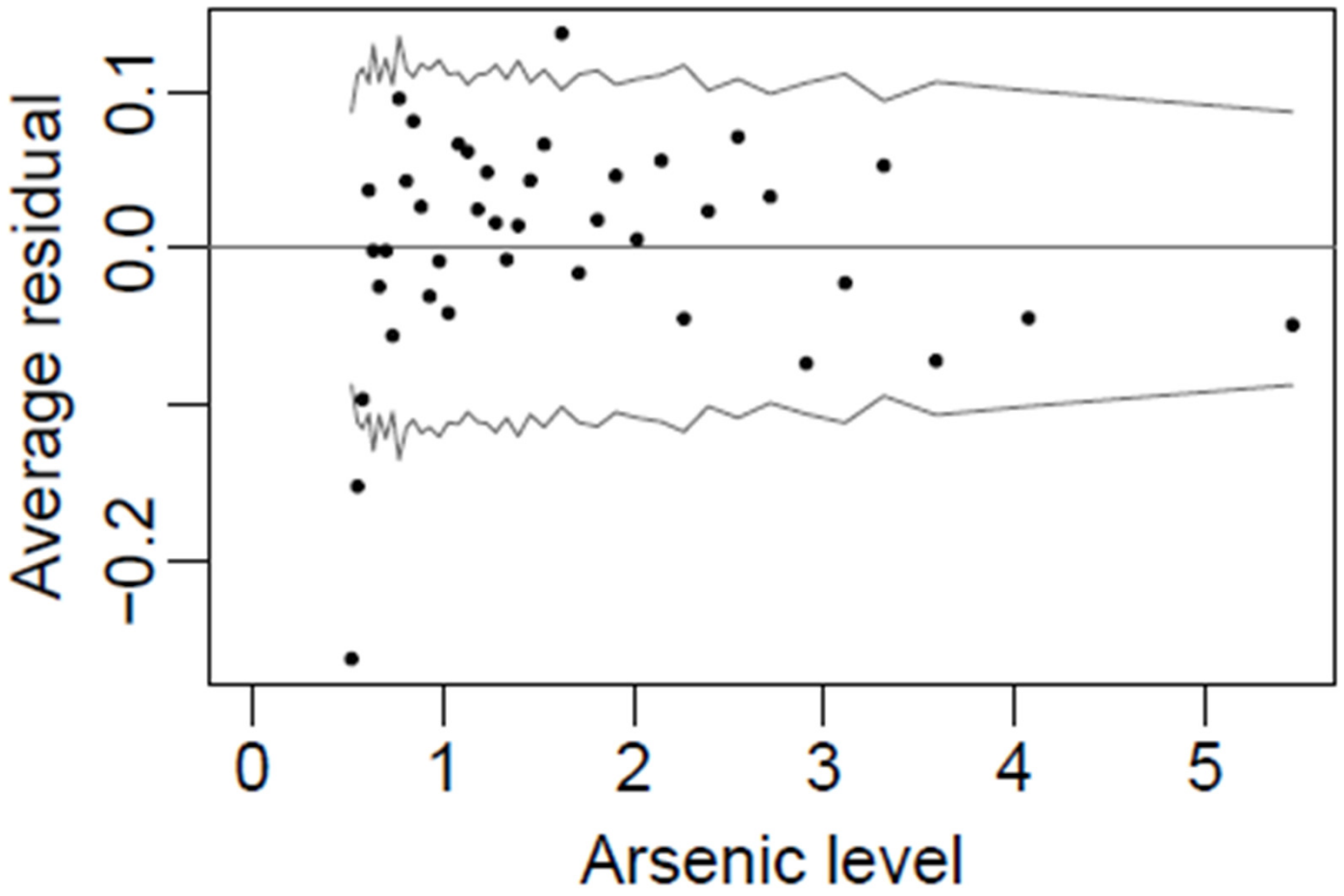
Binned residual plot

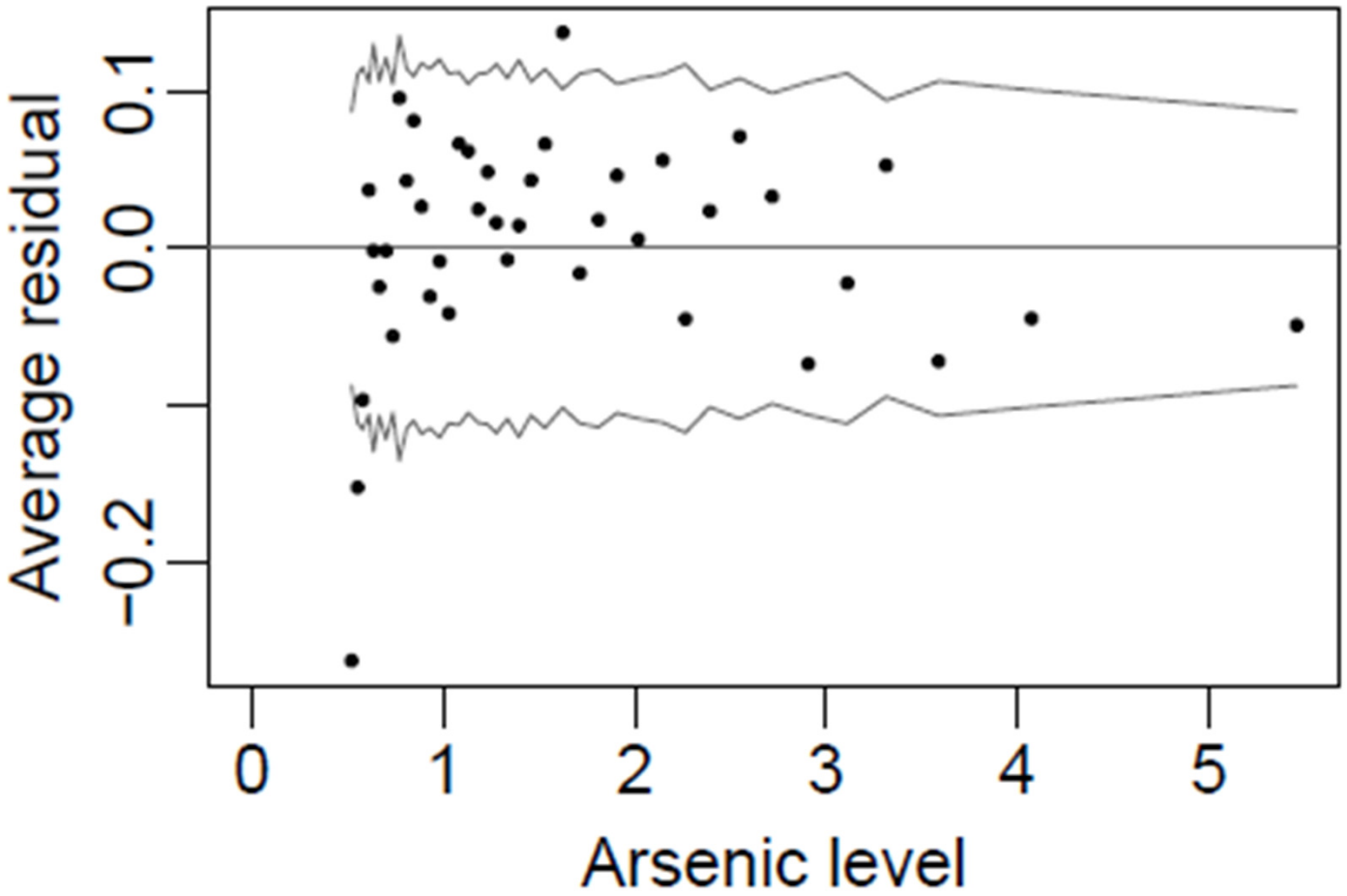


Binned residuals vs. distance



Binned residuals vs. arsenic





Try the log scale:

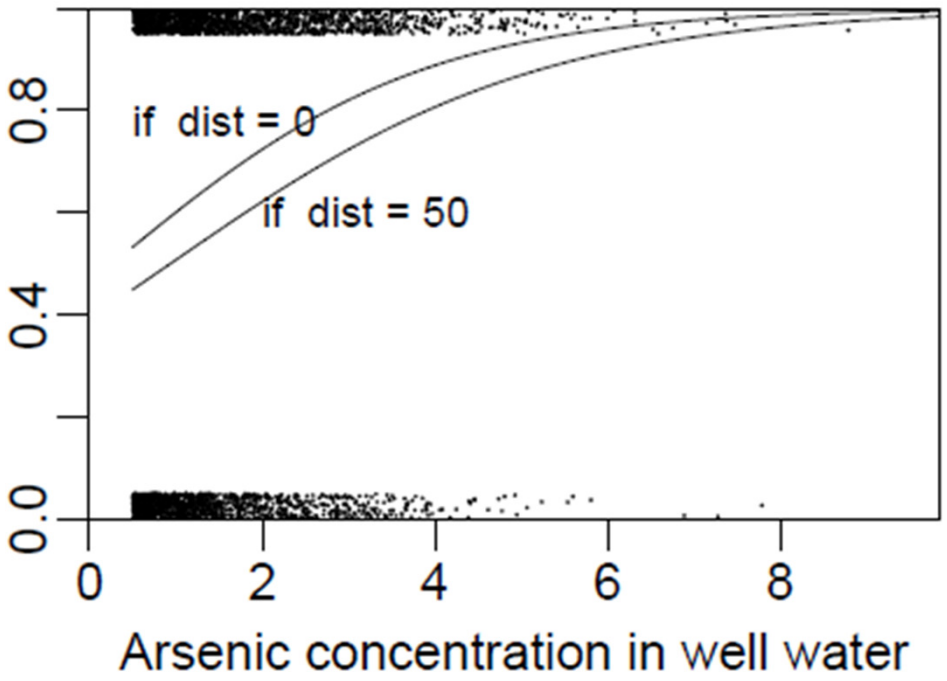
```
log.arsenic <- log(arsenic)
c.log.arsenic <- log.arsenic - mean (log.arsenic)
```

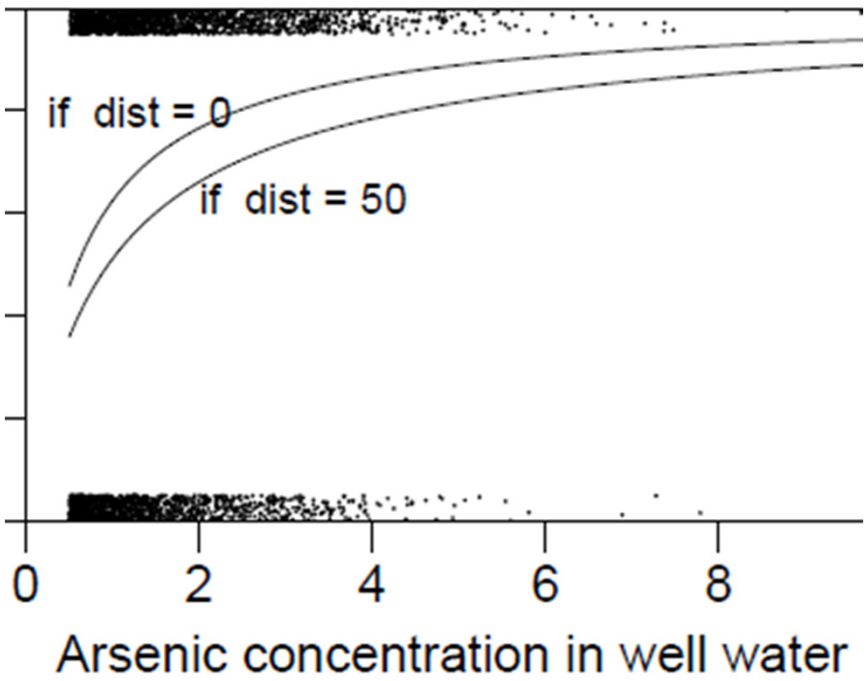
New model

	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)

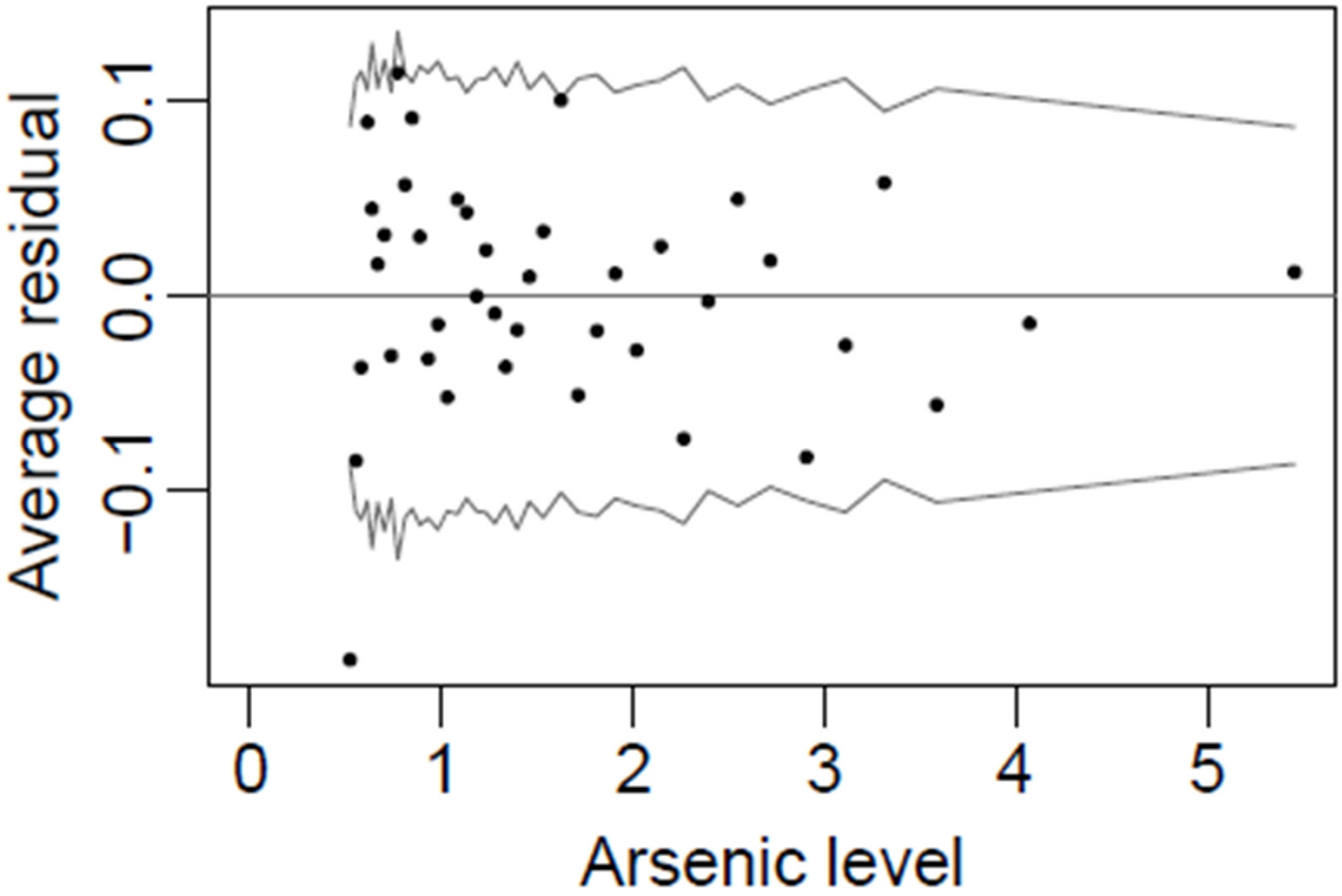
Comparing old and new models





(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?
  - ?

Today's class

- Example: wells in Bangladesh
  - Building a logistic regression model
  - Logistic regression with interactions
  - Evaluating, checking, and comparing models
  - Probit regression and the latent-variable model
  - Mapping the logit/probit regressions to a formal model of preferences
- Related methodological topics

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

Today's class

- Example: wells in Bangladesh
  - Building a logistic regression model
  - Logistic regression with interactions
  - Evaluating, checking, and comparing models
  - Probit regression and the latent-variable model
  - Mapping the logit/probit regressions to a formal model of preferences
- Related methodological topics

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
  - $\Pr(\text{switch}) = \Pr(y_i=1) = \Pr(a_i As_i > b_i + c_i x_i)$

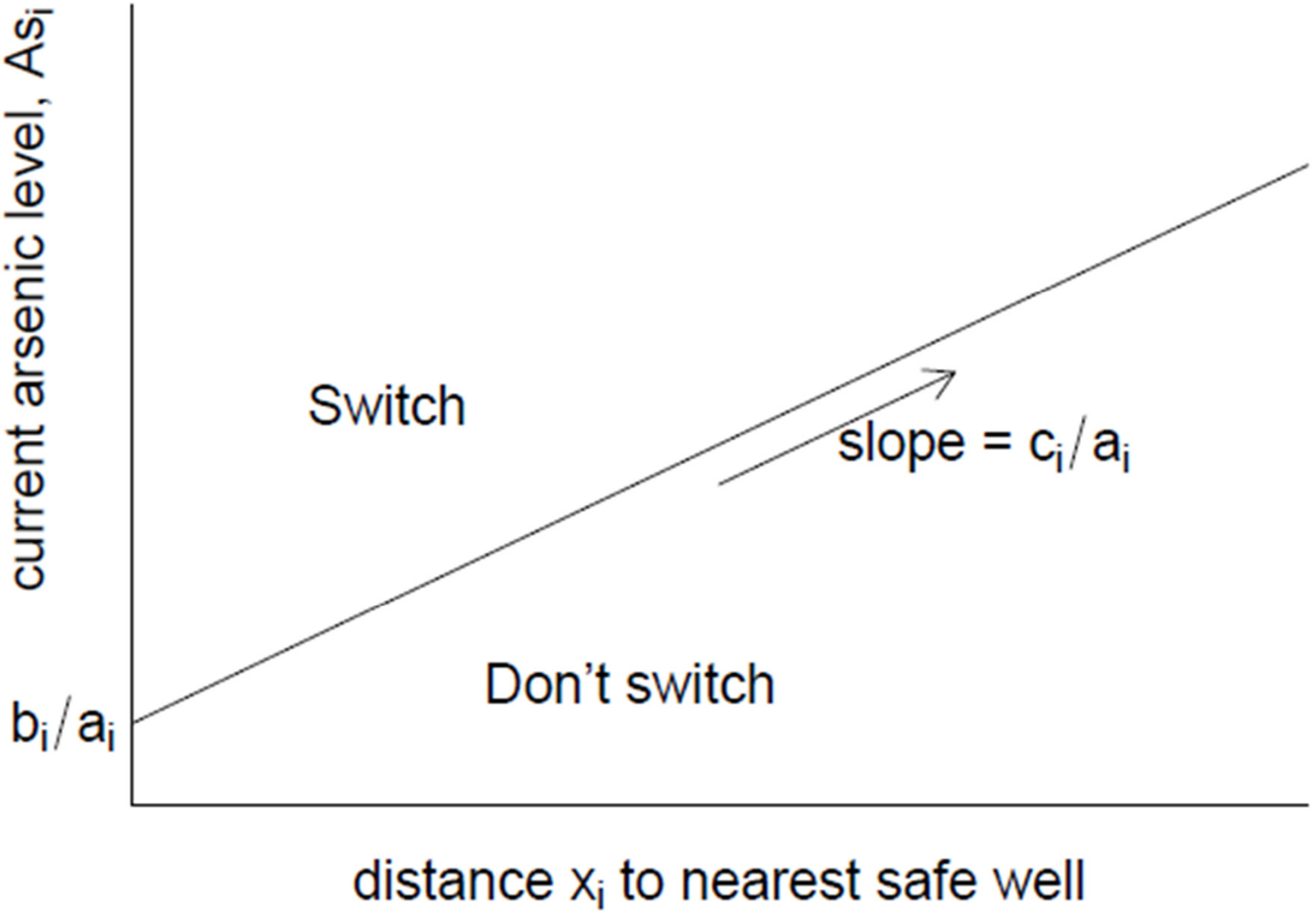
$$= \Pr((a_i As_i - b_i)/c_i > x_i)$$

	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
  - $Pr(switch) = Pr(y_i=1) = Pr\ (a_i As_i > b_i + c_i x_i)$

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



Today's class

- Example: wells in Bangladesh
  - Building a logistic regression model
  - Logistic regression with interactions
  - Evaluating, checking, and comparing models
  - Probit regression and the latent-variable model
  - Mapping the logit/probit regressions to a formal model of preferences
- Related methodological topics

Some topics in fitting discrete-data regression models

- Poisson models
- Treating discrete variables as continuous
- Unordered categorical regression
- Robust alternative to logit-probit

Poisson models

- Always Poisson regression, never Poisson distribution
- Always allow for overdispersion

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

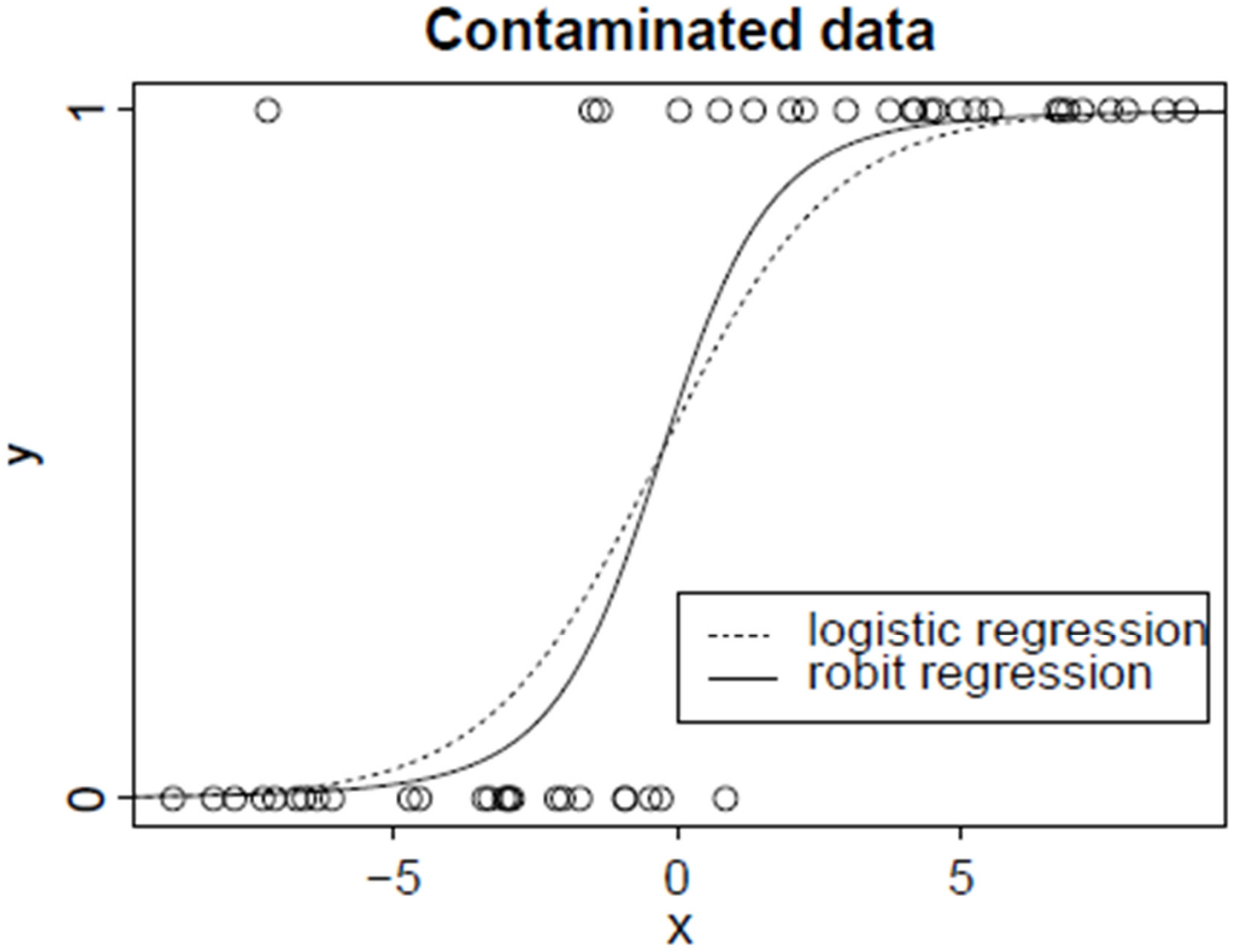
Unordered categorical regression

- Use an ordered model if possible or else try repeated binary splits
- Example: Bush/Clinton/Perot/no-vote
  - Vote or no-vote
  - If vote: major party or Perot
  - If major party: Bush or Clinton

Robust alternative to logit-probit

- Bound the probabilities away from 0 and 1 to allow for “contamination”

Robit (robust) regression



Today's class

- Example: wells in Bangladesh
  - Building a logistic regression model
  - Logistic regression with interactions
  - Evaluating, checking, and comparing models
  - Probit regression and the latent-variable model
  - Mapping the logit/probit regressions to a formal model of preferences
- Related methodological topics