

Risky odds or odd risks? Interpreting models for dichotomous outcomes

Ken Kleinman

Outline

- Motivation: dichotomous outcomes
- Some approaches
 - Linear, log, logistic
- Problems with logistic regression?
- “Solutions”
- Solution
- Recommendations

Setting

- You have a dichotomous outcome:
 - Did a patient succeed at quitting smoking?
 - Was diagnosis implied by electronic medical record review correct?
 - (Does a patient have high blood pressure?)

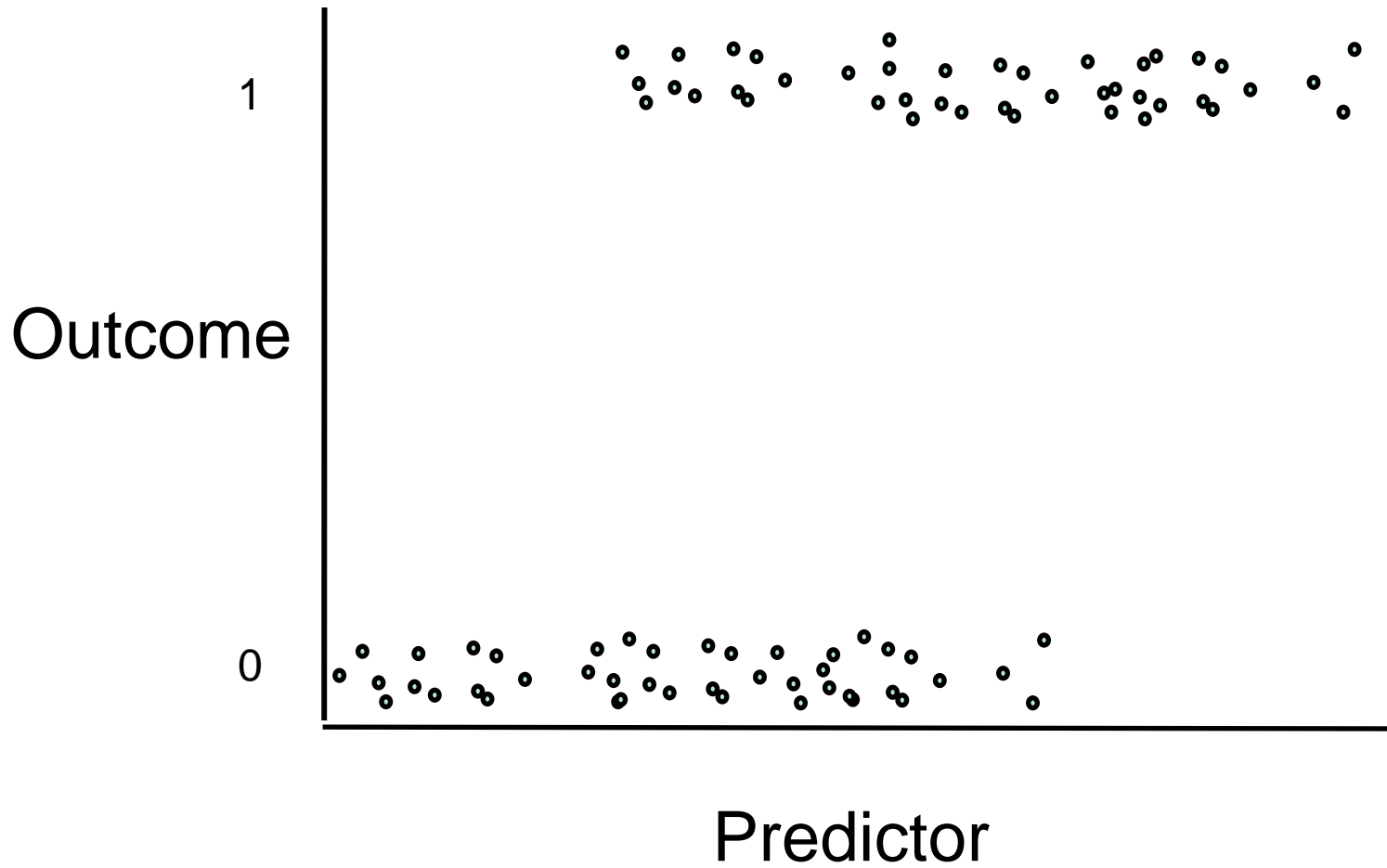
Setting

- What do you want to learn from your dichotomous outcome?
- What's the probability (risk) that it happens?
- Does a predictor (risk factor) affect that probability?
- How much does the risk factor affect that probability?

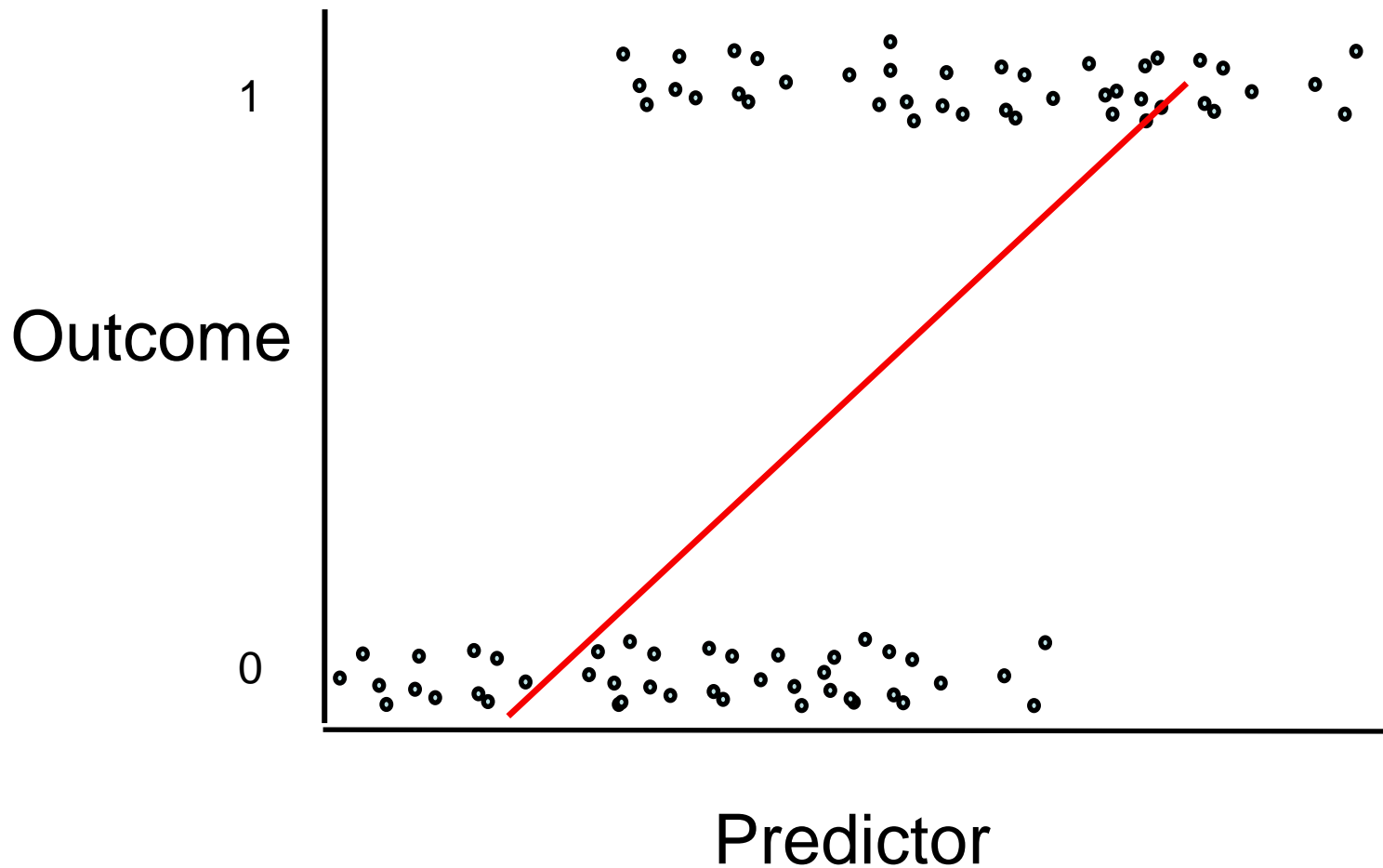
What to do with $y=(0,1)$?

- Logistic regression?
 - Now standard, and most just accept it
 - Results in odds ratios associated with predictors
- What are some alternatives?
 - Linear regression
 - “Log-binomial” regression
- First, alternatives, then review of logistic

Visual key



Modeling options: linear model



Modeling options

- Linear model
- Use $Y = 0$ or 1
- Model:

$$y = B_0 + x_1 B_1 + \dots + x_k B_k + e$$

- I.e.

$$E(y) = B_0 + x_1 B_1 + \dots + x_k B_k$$

Modeling options

$$E(y) = B_0 + x_1 B_1 + \dots + x_k B_k$$

- Values may be less than 0, greater than 1
- But: B_1 is difference in $E(y)$, i.e. the difference in the probability that $y=1$! (After controlling for other covars, even!)
- This is what we really want to know: How much more is my risk, due to x_1 ?

Modeling options

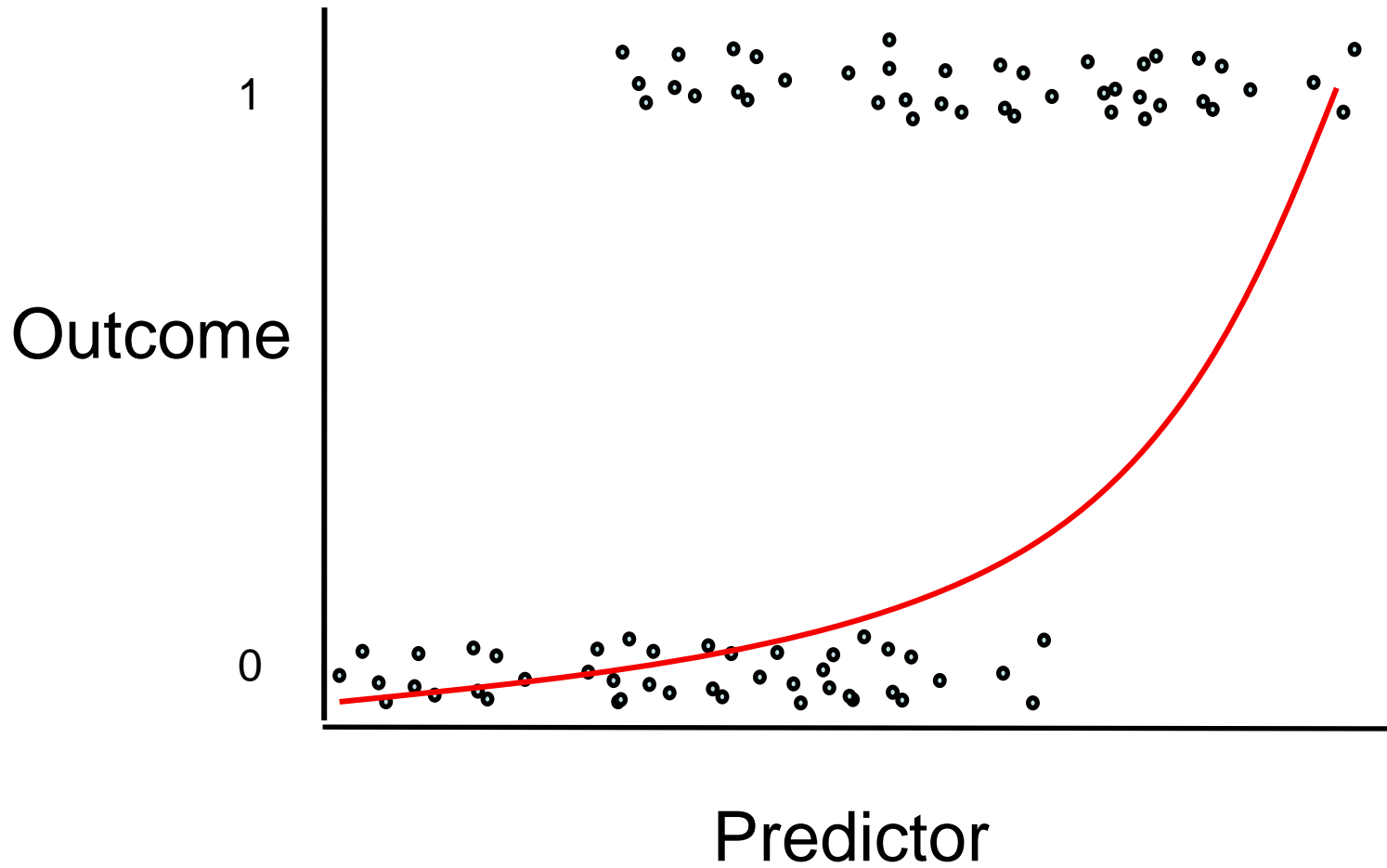
- Log-linear model

$$\log(E(y)) = B_0 + x_1 B_1 + \dots + x_k B_k$$

- Ensures all predicted values are > 0 :

$$E(y) = e^{B_0 + x_1 B_1 + \dots + x_k B_k}$$

Visual key



Modeling options

$$E(y) = e^{B_0 + x_1 B_1 + \dots + x_k B_k}$$

- Values greater than 1 are possible
- But: B_1 is the log of the relative risk:

$$\frac{E(y | x_1 = 1)}{E(y | x_1 = 0)} = \frac{e^{B_0 + B_1 + \dots + x_k B_k}}{e^{B_0 + x_2 B_2 + \dots + x_k B_k}} = \frac{e^{B_1} \cdot e^{B_0 + \dots + x_k B_k}}{e^{B_0 + x_2 B_2 + \dots + x_k B_k}} = e^{B_1}$$

- Not quite as good as knowing how much more my risk is, but useful

Modeling options

- Similar: use binomial (or Bernoulli) distribution (Generalized Linear Model) with the same log link :

$$\log(\Pr(y = 1)) = B_0 + x_1 B_1 + \dots + x_k B_k$$

- Retain nice interpretation of B_1
- Values > 1 now impossible!! (In range of observed data.)
- **“Log-Binomial” regression**

Modeling options

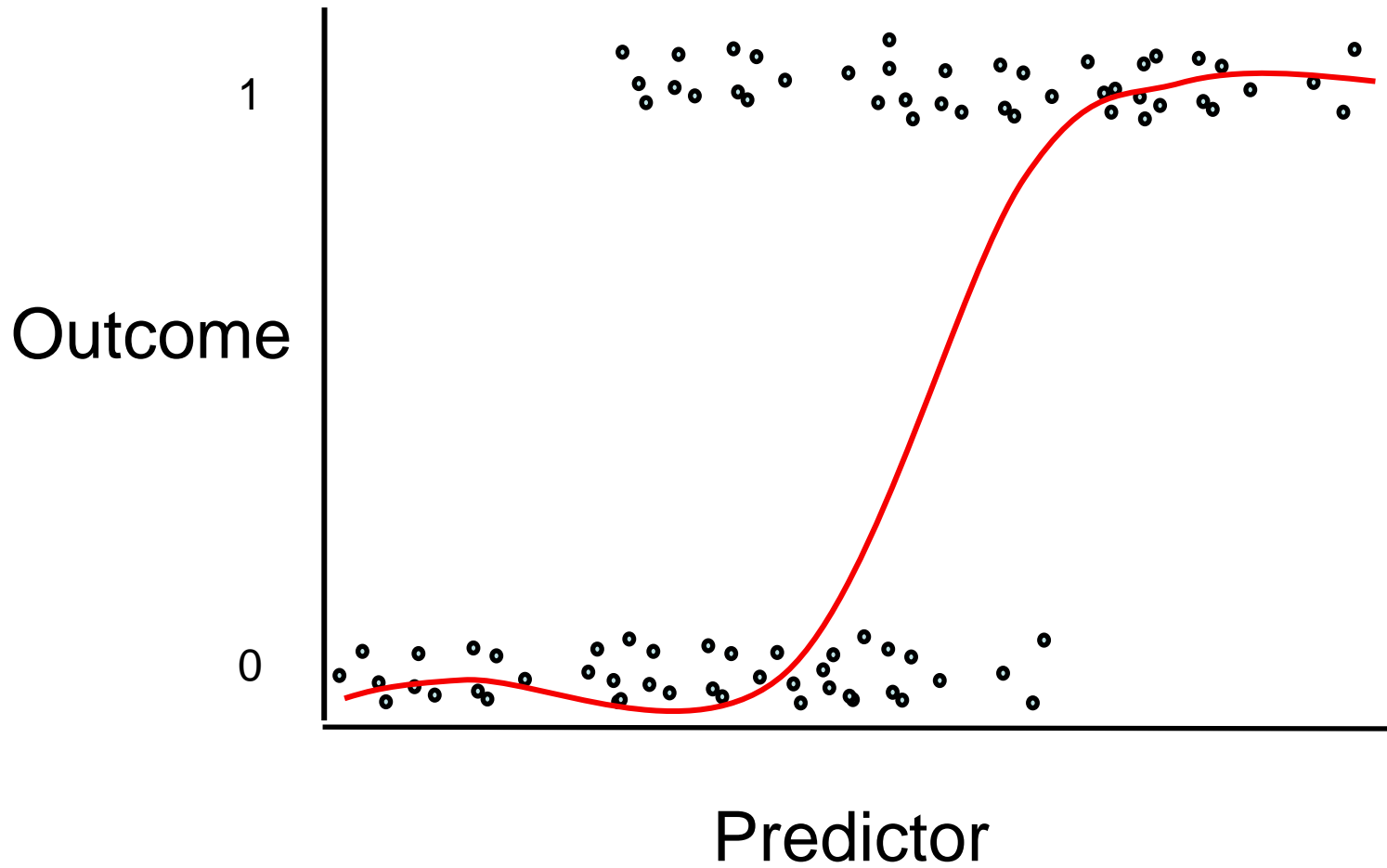
- Logistic regression
- Make sure all predicted values are >0 , <1

$$\Pr(y = 1) = \frac{e^{B_0 + x_1 B_1 + \dots + x_k B_k}}{1 + e^{B_0 + x_1 B_1 + \dots + x_k B_k}} = \text{expit}(\sum xB)$$

- I.e., after more algebra:

$$\log \left[\frac{\Pr(y = 1)}{\Pr(y = 0)} \right] = \log(\text{Odds}(y)) = \text{logit}(y) = \sum xB$$

Visual key



Modeling options

- Also, after more algebra, B_1 is log relative odds:

$$\frac{\text{logit}(y | x_1 = 1)}{\text{logit}(y | x_1 = 0)} = B_1 \quad \rightarrow \quad e^{B_1} = \frac{\text{Odds}(y | x = 1)}{\text{Odds}(y | x = 0)}$$

- This is the support for the interpretation that logistic regression results in the odds ratio
- Odds ratio is even less useful than the risk ratio, but at least it goes up when the risk goes up

Probability scales

- What is an odds?
- Odds = $\Pr(y=1)/\Pr(y=0)$
- Odds = How many times more likely am I to win than to lose?
- Having an **odds ratio** of 2, or doubling the odds, means I'm twice as many times more likely to win as to lose
- Ratio of a ratio. Yuck.

Probability scales

- If $\Pr(y=1) = .1$, then odds $= .1/.9 = .11$
- To make an OR of 2, the odds in the other case must be .22, which implies that the probability in the other case must be .18 (if $\Pr(y=1)=.18$, odds $= .18/.81 = .22$)
- $OR = .22/.11 = 2$
- Probability difference: $.18 - .1 = .08$
- Probability (risk) ratio of $.1818/.1 = 1.8$

Probability scales

- If $\Pr(y=1) = .98$, then odds = $.98/.02 = 49$
- To make an OR of 2, odds in the other case must be 98, which implies that the probability in the other case must be .99 (if $\Pr(y=1)=.99$, odds = $.99/.01 \approx 98$)
- $OR = 98/49 = 2$
- Probability difference of $.99 - .98 = .01$
- Risk ratio = $.99/.98 = 1.01$

Probability scales

- I showed that a single odds ratio (2) might mean a medium (1.8) or a miniscule (1.01) risk ratio
- I could also construct examples with equal risk ratios which result in different OR
- Could show equal risk differences with different OR and RR, too

Problems?

- Popular wisdom:
“If the baseline probability is small, the odds ratio is similar to the risk ratio”

Problem with logistic regression?

- Some have argued that when $\Pr(y=1)$ is “not small,” that the dissimilarity between the OR and the RR means that logistic regression should be “fixed”
- This is not true, of course—there’s nothing incorrect about logistic regression
- “Just” with its interpretation
- If you want to interpret OR as RR

“Solutions (1)”

- “Correction factor” (Zhang and Yu, JAMA 1998) will convert OR from a logistic regression into a RR
- Kleinman (not me) and Norton (Health Services Research 2009) get an ‘adjusted risk measure’ from logistic regression

“Solutions (1)”

- BUT when other variables are present in logistic regression, the values they take effectively change $\Pr(y=1)$ for the main exposure.
- What?

“Solutions (1)”

- The values covariates take effectively change $\Pr(y=1)$ for the main exposure.

$$\Pr(y = 1) = \text{expit}(\sum xB) = \frac{e^{B_0 + x_1 B_1 + x_2 B_2}}{1 + e^{B_0 + x_1 B_1 + x_2 B_2}}$$

- Suppose we're interested in the effect of x_1 , and that $B_0=1$, $B_2=4$
- Then, if $x_2=0$: $P(y=1|x_1=0) = e/(1+e) = .73$
- But, if $x_2=1$: $P(y=1|x_1=0) = e^5/(1+e^5) = .99$

“Solutions (1)”

- In other words, the (multivariable) logistic regression model **implies a range** of baseline probabilities (for different sets of covariates) with a ***constant OR*** for the risk factor of interest
- **BUT:** a constant **RR** implies a **different OR** for different baseline probabilities (as I showed)
- Therefore, getting a RR from logistic regression is pointless— it’s self-contradictory! – **if** logistic regression is the right thing to do.

“Solutions (1)”

- Getting a RR from logistic regression is pointless if logistic regression is the right thing to do, i.e. if there is a constant odds ratio
- What if instead, there is a constant risk ratio?
- Then logistic regression is the wrong thing to do. Instead, we should use the log-binomial regression

“Solutions (2)”

- “Log-binomial” model we saw earlier:

$$\log(\Pr(y = 1)) = B_0 + x_1 B_1 + \dots + x_k B_k$$

Like Zhang and Yu, generates the RR

- Easy fitting of this model in SAS was advertised by Spiegelman and Hertzmark (AJE 2005;165:199-200); also shown by Robbins et al. (Ann Epidemiol, 2002), others. Easy in R, also

“Solutions (2)”

- “Log-binomial” model

$$\log(\Pr(y = 1)) = B_0 + x_1 B_1 + \dots + x_k B_k$$

- As implied by Zhang and Yu, this model has the predicted probability going up by a fixed proportion (e^{B_1}) for each unit change in x_1
- As we just reviewed, this is not the case if the odds ratio is constant, but maybe it's not constant, and the risk ratio is constant?

“Solutions” (2)

- But there are reasons to suspect this might not be the case
- For example, consider a problem with two dichotomous predictors.
- The risk ratio for x_1 must be the same for each value of x_2 . Explicitly:

$$\frac{\Pr(y = 1 \mid x_1 = 1, x_2 = 0)}{\Pr(y = 1 \mid x_1 = 0, x_2 = 0)} = \frac{\Pr(y = 1 \mid x_1 = 1, x_2 = 1)}{\Pr(y = 1 \mid x_1 = 0, x_2 = 1)}$$

- Why is that troubling? Well...

“Solutions” (2)

- Suppose data are:

X_1	X_2	$\Pr(y=1)$	
0	0	.25	
1	0	.75	$RR(x_1 x_2=0) = .75/.25 = 3$

- When $x_2 = 0$, RR for x_1 is 3

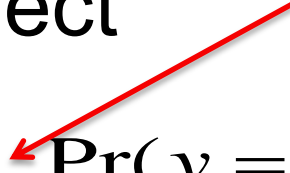
“Solutions” (2)

- Suppose:
- What's the RR for x_1 when $x_2 = 1$?
- It can't possibly be 3!! So the lob-bin model can't be right

X_1	X_2	$\Pr(y=1)$	
0	0	.25	
1	0	.75	$RR(x_1 x_2=0) = .75/.25 = 3$
0	1	.75	
1	1	?	$RR(x_1 x_2=1) = ?/.75 \neq 3$

“Solutions” (2)

- When $x_2 = 1$, RR can't be > 1.33 .
- In other words, the log-binomial or relative risk model can't be correct

$$\frac{\Pr(y = 1 \mid x_1 = 1, x_2 = 0)}{\Pr(y = 1 \mid x_1 = 0, x_2 = 0)} \neq \frac{\Pr(y = 1 \mid x_1 = 1, x_2 = 1)}{\Pr(y = 1 \mid x_1 = 0, x_2 = 1)}$$


- ORs, however, can be exactly the same!

“Solutions” (2)

- If the data look like that, the software will also often fail to fit the log-binomial model
- Spiegelman and Hertzmark suggest trying the log-binomial model, and if it doesn't work to replace it with a Poisson model:

$$E(y) = e^{B_0 + x_1 B_1 + \dots + x_k B_k}$$

(again, easy in SAS or R)

“Solutions (2)”

- Poisson model: $E(y) = e^{B_0 + x_1 B_1 + \dots + x_k B_k}$
- This will converge, but note $E(y) > 1$?
- Fixes bad data by allowing the expected probability to be greater than 1!
- This is a very bad idea. Better to take the failure of the log-binomial model to fit as diagnostic that relative risk does not describe reality very well

“Solutions” (2)

- Notice here that the problems with the limits of relative risks get worse quickly when the $\text{Pr}(y=1)$ can be large (especially with a wide range of baseline $\text{Pr}(y=1)$)
- Just the case when you might want to avoid logistic regression “because the OR is not similar to the RR”

Real solution

- The importance of our results may (should!) depend not just on demonstrating a relationship, but on its magnitude
- in terms that can be understood

Real solution: **Predicted probabilities**

Real solution

Predicted probabilities?

In each model, we linked $\Pr(y=1)$ with some function of x and B

So, take estimated values of B from the regression results and values of x that are interesting; plug them into the function and get predicted $\Pr(y=1|x)$ from the model

Easy to do, but more work to write up

Real solution

Why are predicted probabilities the real solution?

They are the **only** way to give a sense of what is the real effect of a changing covariate value (unless the linear risk-difference model fits well)

In other words, they are the only way to know if a RR or OR of 40 increases my actual risk enough to bother me

What does an OR or RR mean?

- Smoking and lung cancer
- Smoking raises my risk of incident cancer per year: $RR \approx OR \approx 40$
- If it's a change from 1% to 40% risk, that means something different than a change from a .00000001 to .00000040 risk
- I couldn't find the actual risks for lung cancer from smoking **anywhere** on-line. Only the risk ratio!

Maybe smoking is an acceptable risk?

- US: ~175k new lung cancers annually out of ~300M population, so $\text{pr}(\text{lung ca}) \approx 0.0006$
- 87% of lung cancers are from smoking
- About 50 million smokers
- $\text{Pr}(\text{lung ca}|\text{no smoking}) = .00009$
- $\text{Pr}(\text{lung ca}|\text{smoking}) = .003$
- (15% of smokers get lung CA, lifetime. Many other risks, of course)

Real solution

- So I need to know the actual probability of lung cancer due to smoking, not just the risk ratio
- This principle applies to any application in which the outcome is dichotomous, I claim

Real example: Asthma control

- PACE study, Tracy Lieu, PI (application from Smith et al., Pediatrics 2008 122:760-769)
- What factors are associated with suboptimal control of asthma?
- 754 families
- Model: 30 covariates, including 28 dummy variables, 2 continuous
- Race, age, income, employment, worry, discrimination, etc., etc.,

Which model?

- Results in the paper rely on logistic regression.
- Was that the right choice?
- How might results have changed if we used log-binomial regression?
- Linear-binomial regression?*
- How can we use predicted probabilities to understand the results?

Model results

	Logistic OR	LogBin RR	Poisson RR	Linear PD
Black	.855	.995	.946	-.026
Low symptoms	1.68	1.05	1.37	.09
Daily meds	3.84	1.12	2.10	.23
More (+10) Worry	1.57	1.03	1.21	.08

Model results

	Logistic OR		Poisson RR	Linear PD
Black	.855		.946	-.026
Low symptoms	1.68		1.37	.09
Daily meds	3.84		2.10	.23
More (+10) Worry	1.57		1.21	.08

Model results

	Logistic OR		Poisson RR	Linear PD
Black	.855		.946	-.026
Low symptoms	1.68		1.37	.09
Daily meds	3.84		2.10	.23
More (+10) Worry	1.57		1.21	.08

Predicted Probabilities

Logistic		Poisson	Poisson*	Linear*
.02		.06	.06	0
.14		.19	.19	.15
.27		.25	.25	.30

Predicted Probabilities

Logistic		Poisson	Poisson*	Linear*
.02		.06	.06	0
.14		.19	.19	.15
.27		.25	.25	.30
.53		.42	.42	.53

Predicted Probabilities

Logistic		Poisson	Poisson*	Linear*
.02		.06	.06	0
.14		.19	.19	.15
.27		.25	.25	.30
.53		.42	.42	.53
.88			1	.85

Predicted Probabilities

Logistic		Poisson	Poisson*	Linear*
.02		.06	.06	0
.14		.19	.19	.15
.27		.25	.25	.30
.53		.42	.42	.53
.88			1	.85
.93			1	.999

Predicted Probabilities

Logistic		Poisson	Poisson*	Linear*
.02		.06	.06	0
.14		.19	.19	.15
.27		.25	.25	.30
.53		.42	.42	.53
.88			1	.85
.93			1	.999
.99			1	1

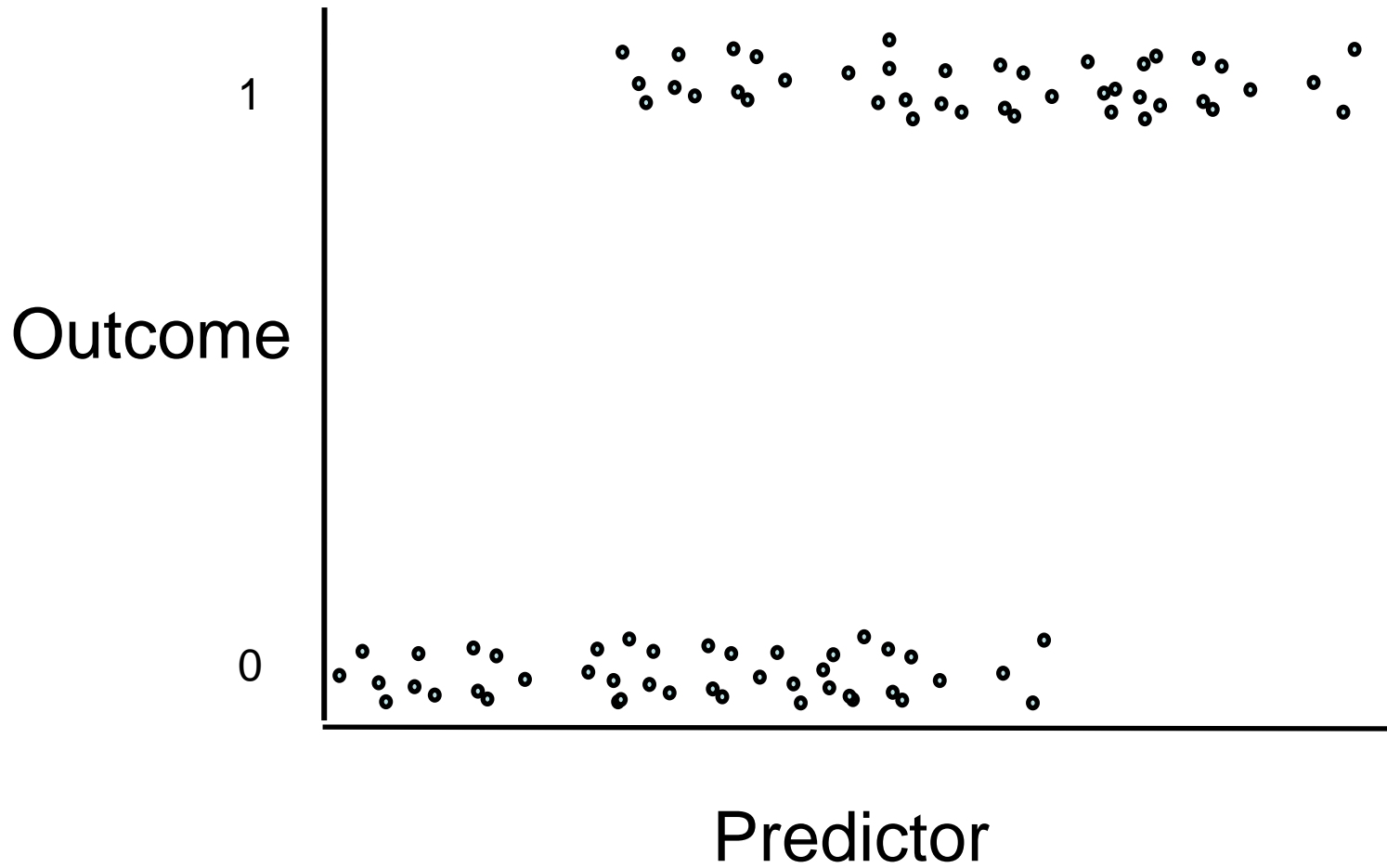
Predicted Probabilities

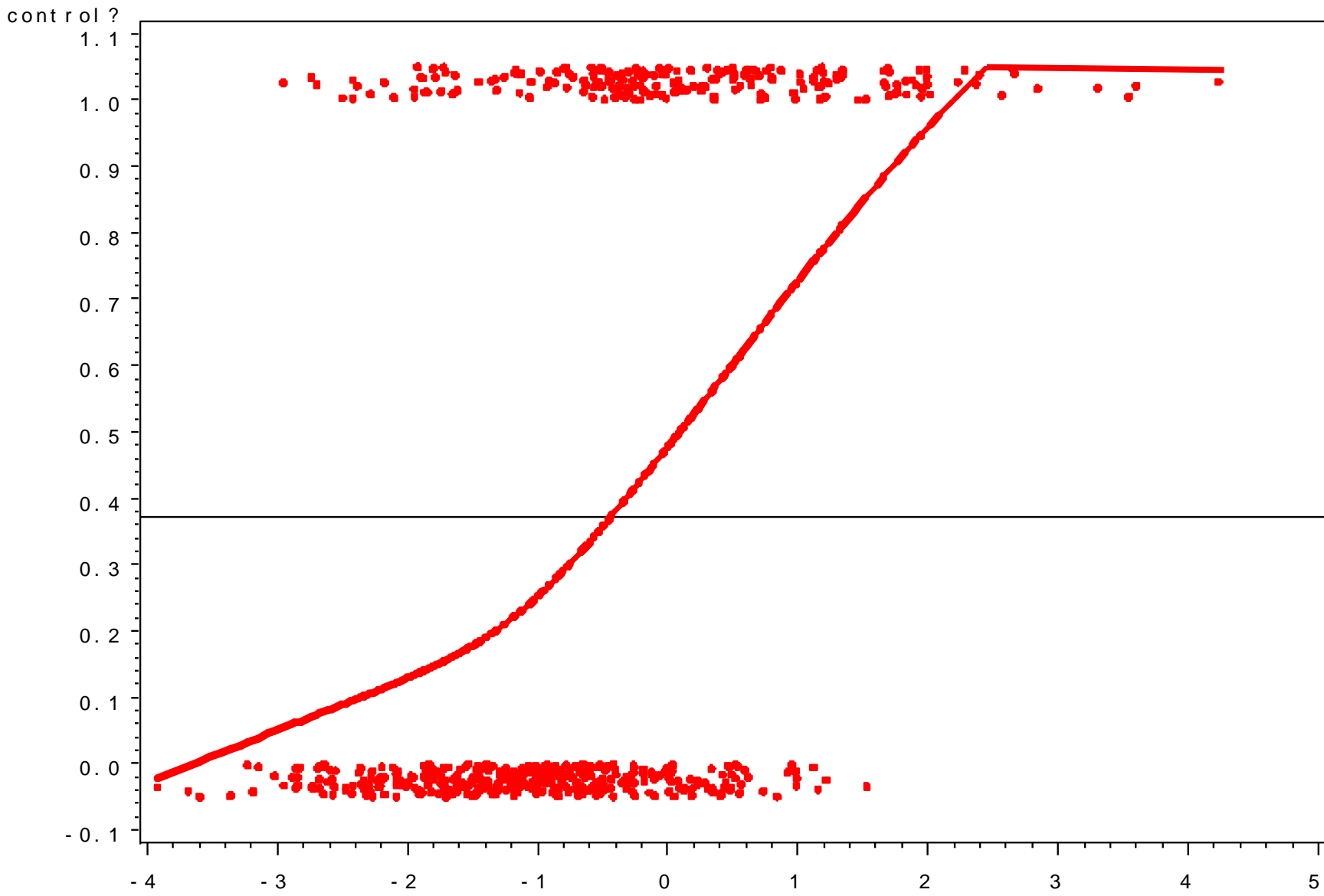
Logistic	LogBin	Poisson	Poisson*	Linear*
.02	.59	.06	.06	0
.14	.71	.19	.19	.15
.27	.74	.25	.25	.30
.53	.80	.42	.42	.53
.88	.92		1	.85
.93	.97		1	.999
.99	.99		1	1

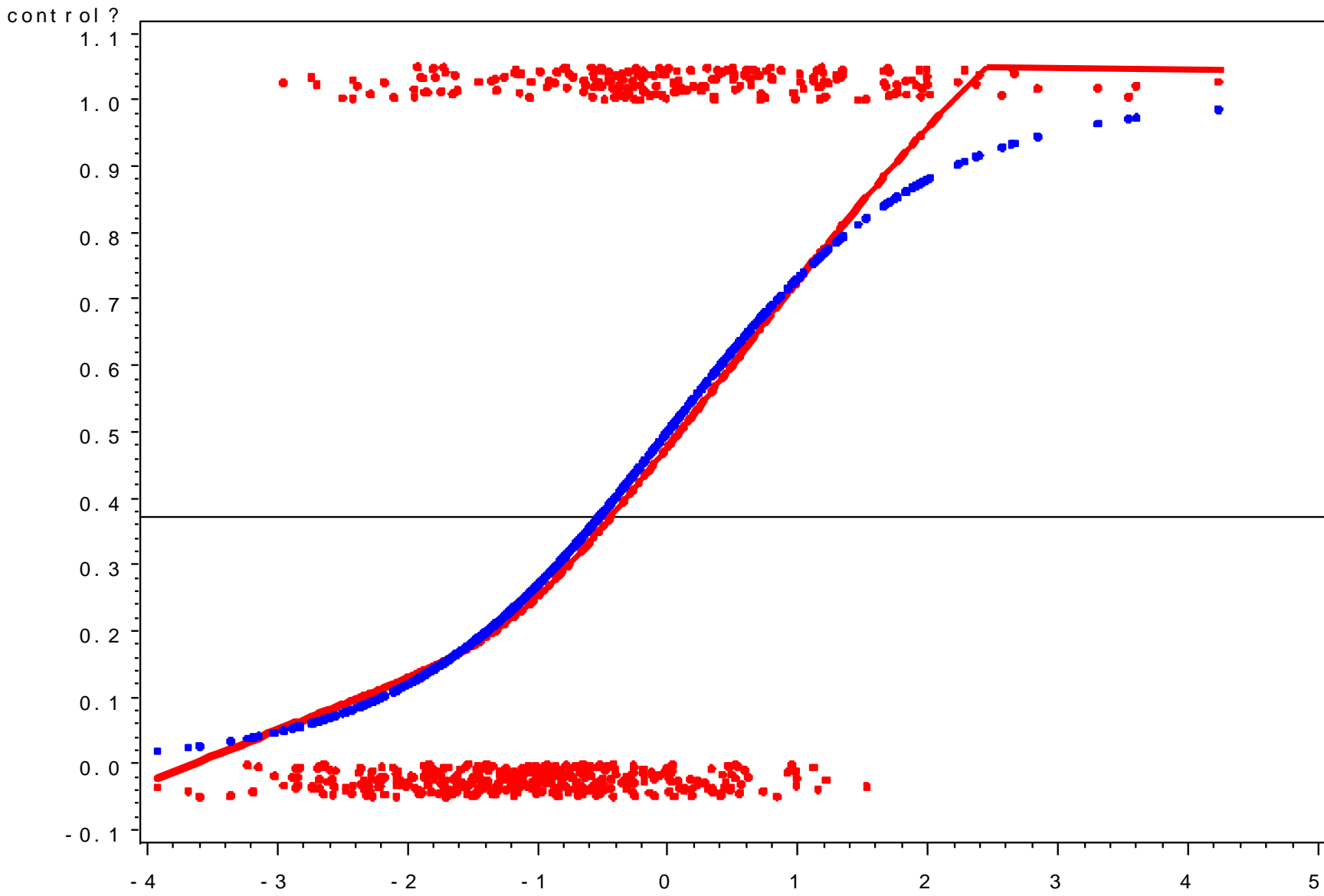
Predicted probabilities

- So, they're different.
- But which of them is best?
- Which of them is right?
- Is any?

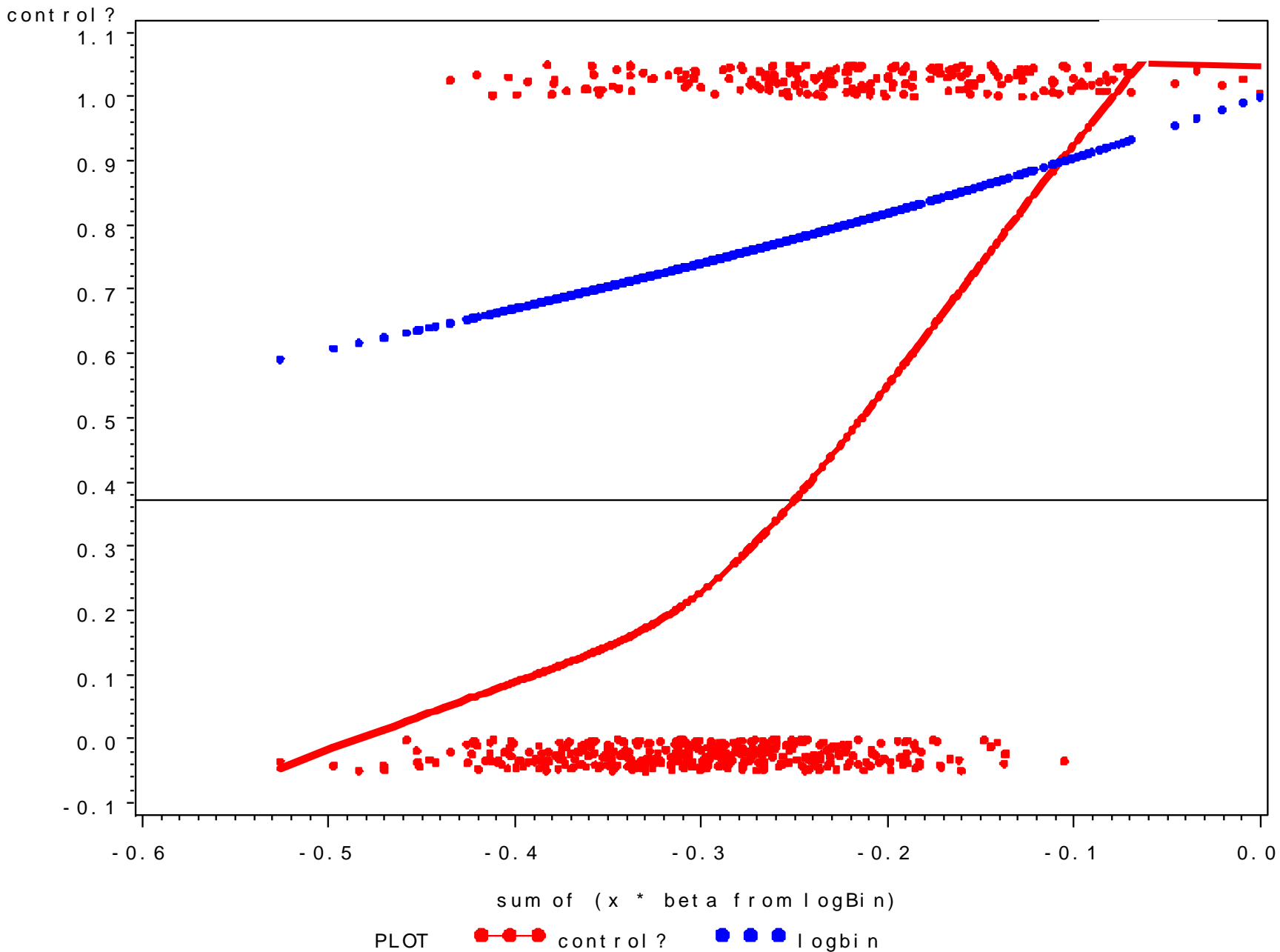
Visual key

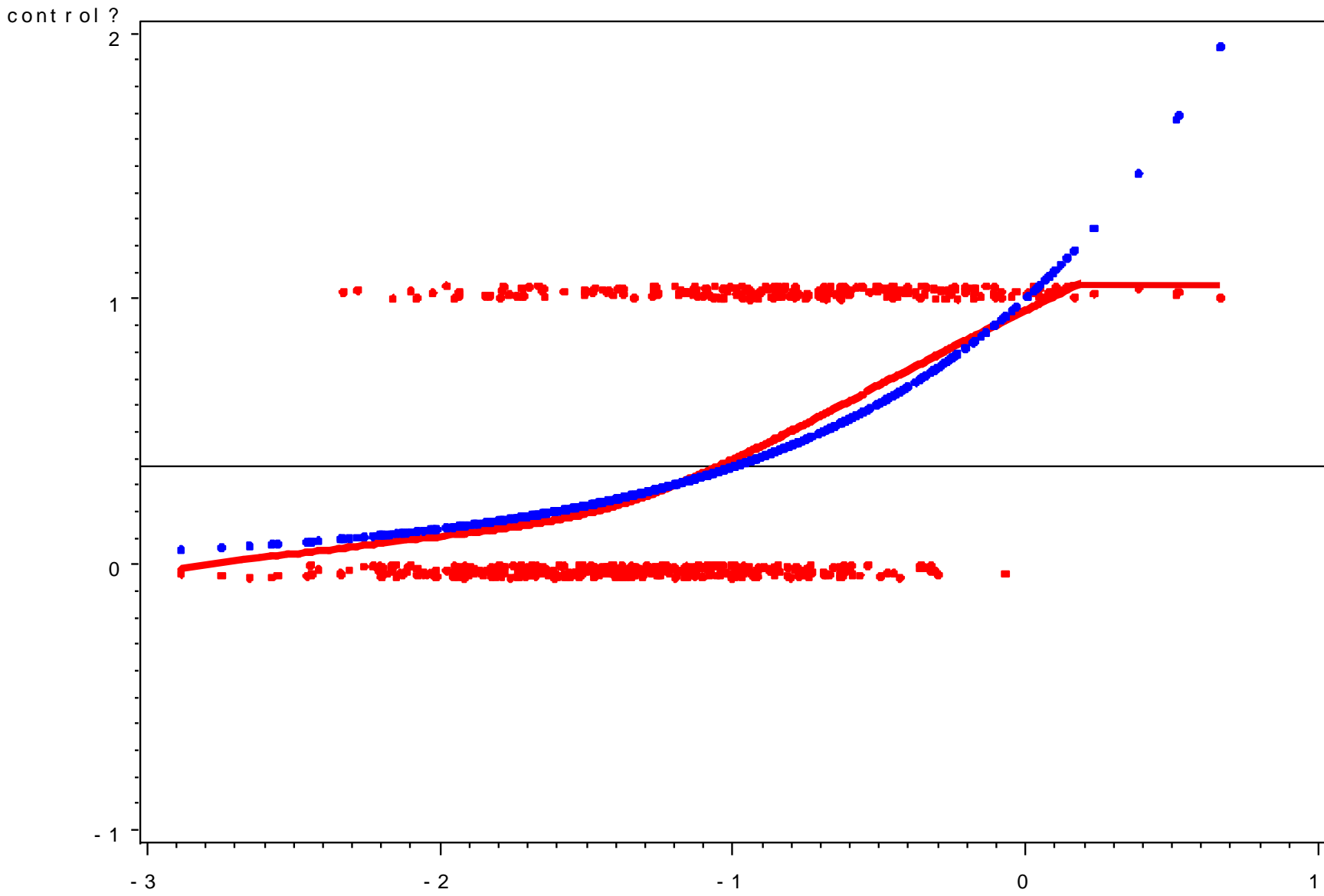






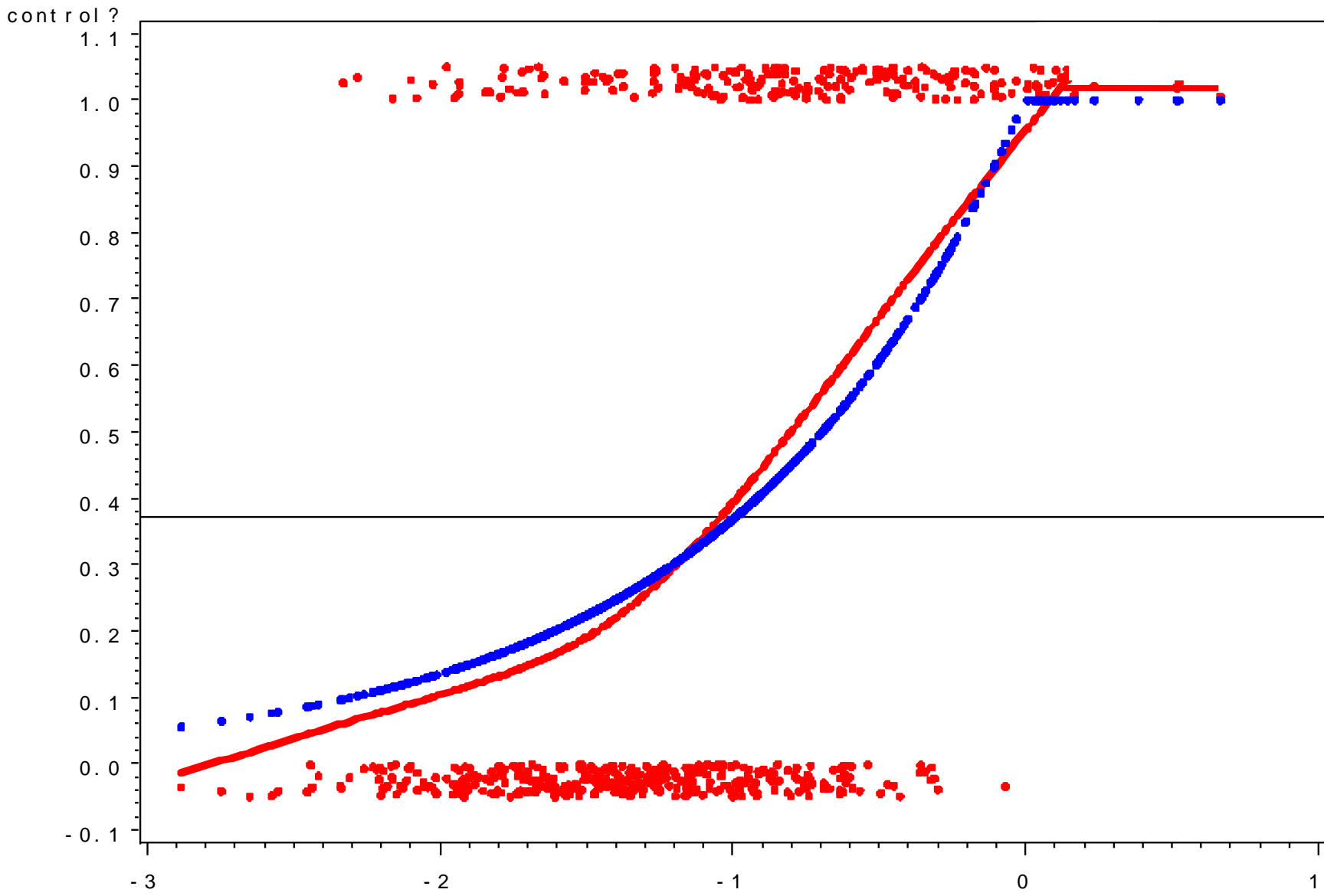
sum of (x * beta from logistic)
PLOT ●—● control? ●—● logistic





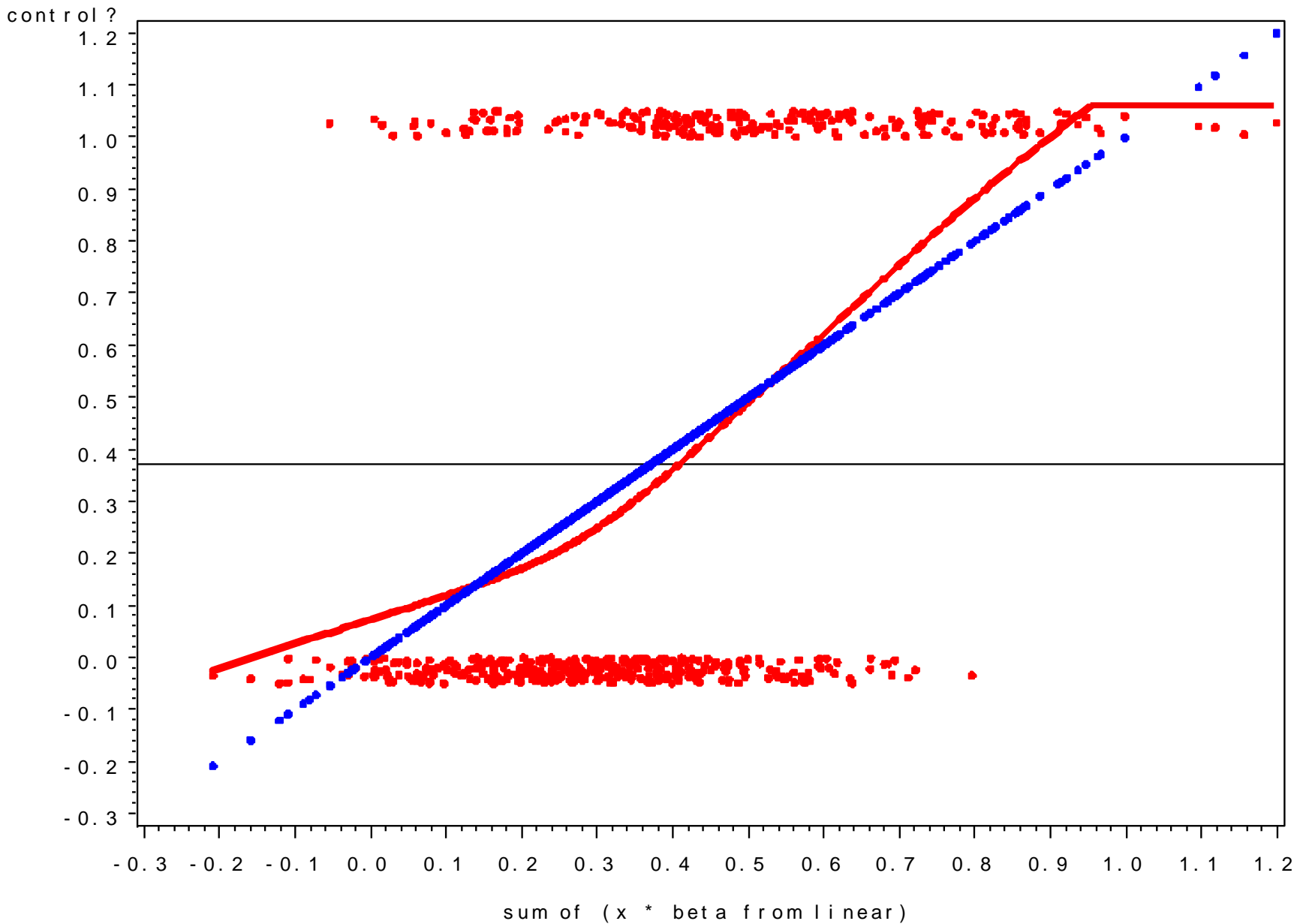
sum of (x * beta from Poisson)

PLOT ●—● control ? ●—● poisson

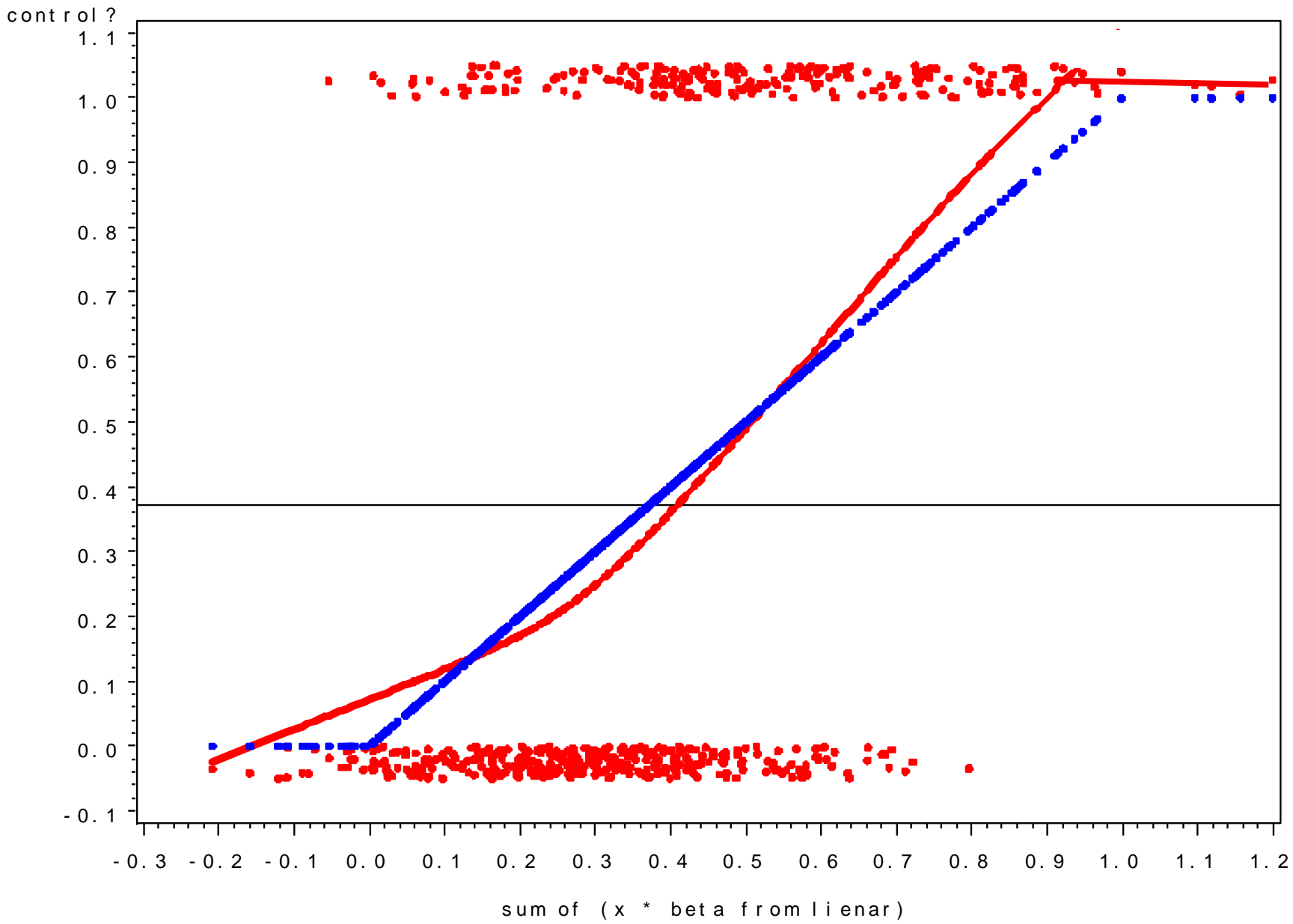


sum of (x * beta from Poisson)

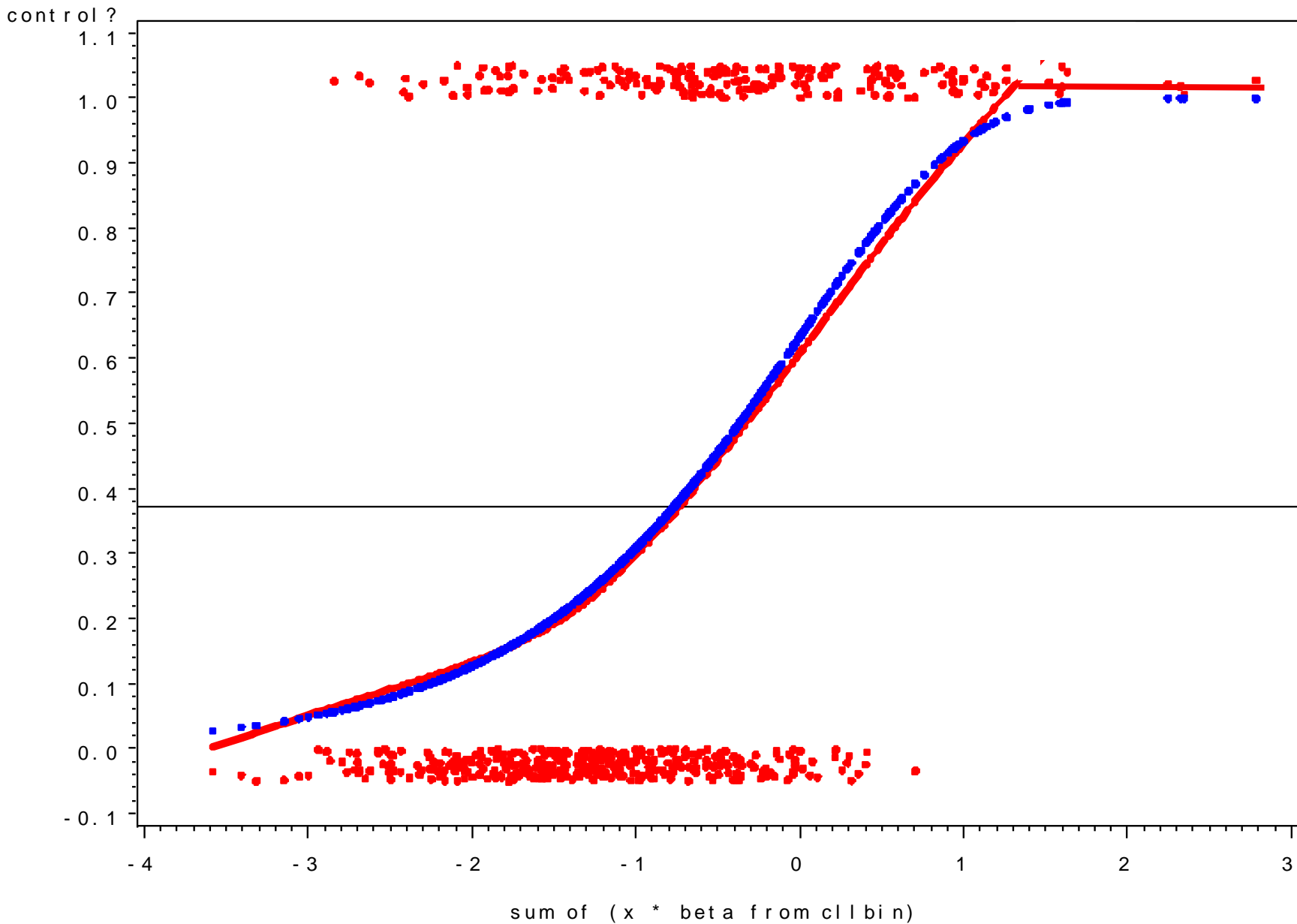
PLOT ●—●—● control ? ●—●—● poisson*

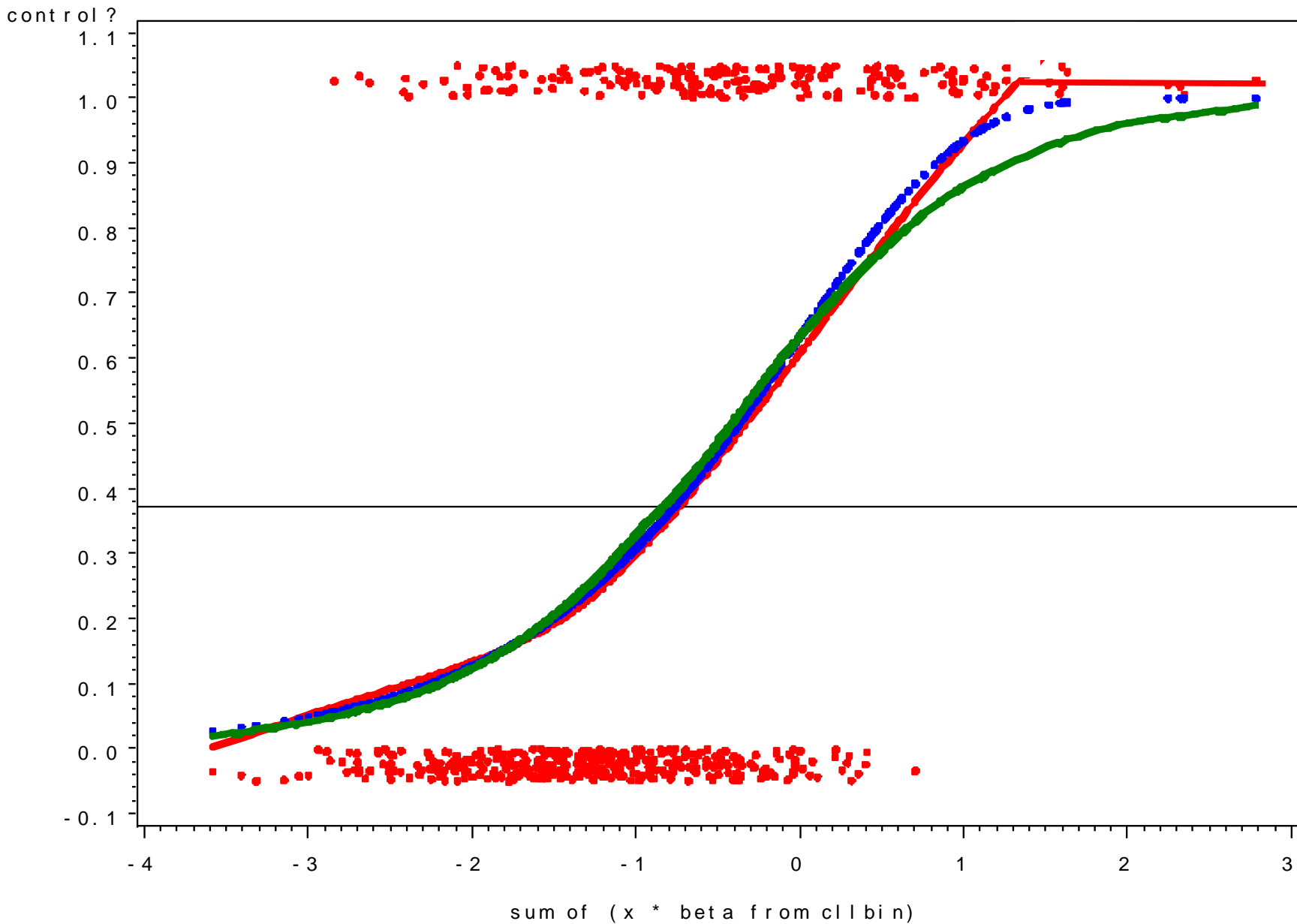


PLOT ●—● control ? ●—● linear



PLOT ●—●—● control ? ●—●—● linear *





Overall: Sum squared error*

Constant	352.0

*Statisticians prefer to use “deviance” or Pearson’s chi-square to assess fit, but in Poisson and linear models, these are not calculable

Overall: Sum squared error

Constant	352.0
Log-Bin	317.6

Overall: Sum squared error

Constant	352.0
Log-Bin	317.6
Log-Poisson	313.6

Overall: Sum squared error

Constant	352.0
Log-Bin	317.6
Log-Poisson	313.6
Linear	274.3

Overall: Sum squared error

Constant	352.0
Log-Bin	317.6
Log-Poisson	313.6
Linear	274.3
Logistic	208.4

Overall: Sum squared error

Constant	352.0
Log-Bin	317.6
Log-Poisson	313.6
Linear	274.3
Logistic	208.4
Cloglog	204.3

Overall: Sum squared error

Constant	352.0
Log-Bin	317.6
Log-Poisson	313.6
Linear	274.3
Logistic	208.4
Cloglog	204.3

Message: Logistic model fits pretty well, compared to RR, RD models.

Predicted Probabilities

- Note that you can also get CI for the predicted probabilities, but they are likely to be wide because they will incorporate variability for all of the Bs
- I suggest reporting p-values for the Bs to show whether the covariates contribute statistically (statistical significance), and predicted probabilities (without CI) to show what that contribution means (practical significance)

Predicted Probabilities

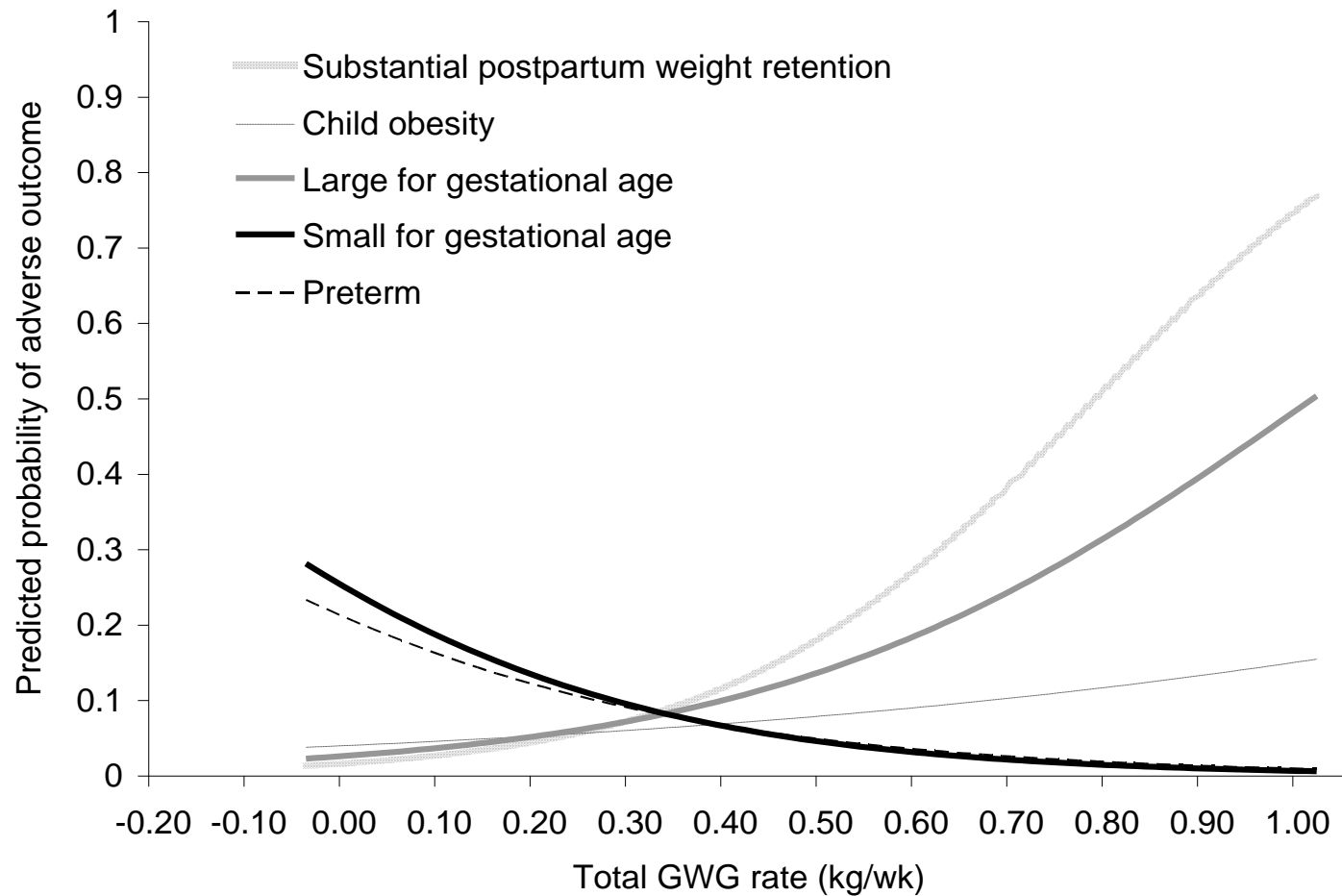
- You can get a predicted probability for any combination of covariate values
- To use covariate values that are far from the observed values is as dangerous as extrapolation in linear regression
- For publication, I suggest that you focus on modal subjects or others of particular interest, then vary just a few covariate values to demonstrate the implications of the model

Table 4a: Predicted probabilities of suboptimal asthma control when variables are varied from medium risk values to high and low risk values

Variable	Variable Value	Predicted probability of suboptimal control	Difference in probability lowest to highest value
Income	9-12,000	75 %	26 %
	15-20,000*	52 %	
	75,000	49 %	
Parental expectations of functioning score	4 (high expectations)	34 %	37 %
	8 *	52 %	
	12 (low expectations)	71 %	
Parental definition of good asthma control (days with symptoms /week)	0-1 day/week	33 %	22 %
	2-4 days/week	55 %	
	5-7 days/week*	52 %	
Competing family priorities score	6 (few competing priorities)	44 %	19 %
	13*	52 %	
	22 (many competing priorities)	63 %	

* Represents medium risk value. Predicted probabilities for other values of each variable calculated while holding all other variables at their medium risk value.

Example of predicted probabilities in practice



Oken, Kleinman, Belfort, Hammitt, and Gillman, AJE

Example of predicted probabilities in practice

Modifiable risk factors

Prenatal smoking

No	1.0 (Ref)
Yes	1.71 (0.90, 3.25)

Gestational weight gain^b

Inadequate or adequate	1.0 (Ref)
Excessive	1.32 (0.85, 2.03)

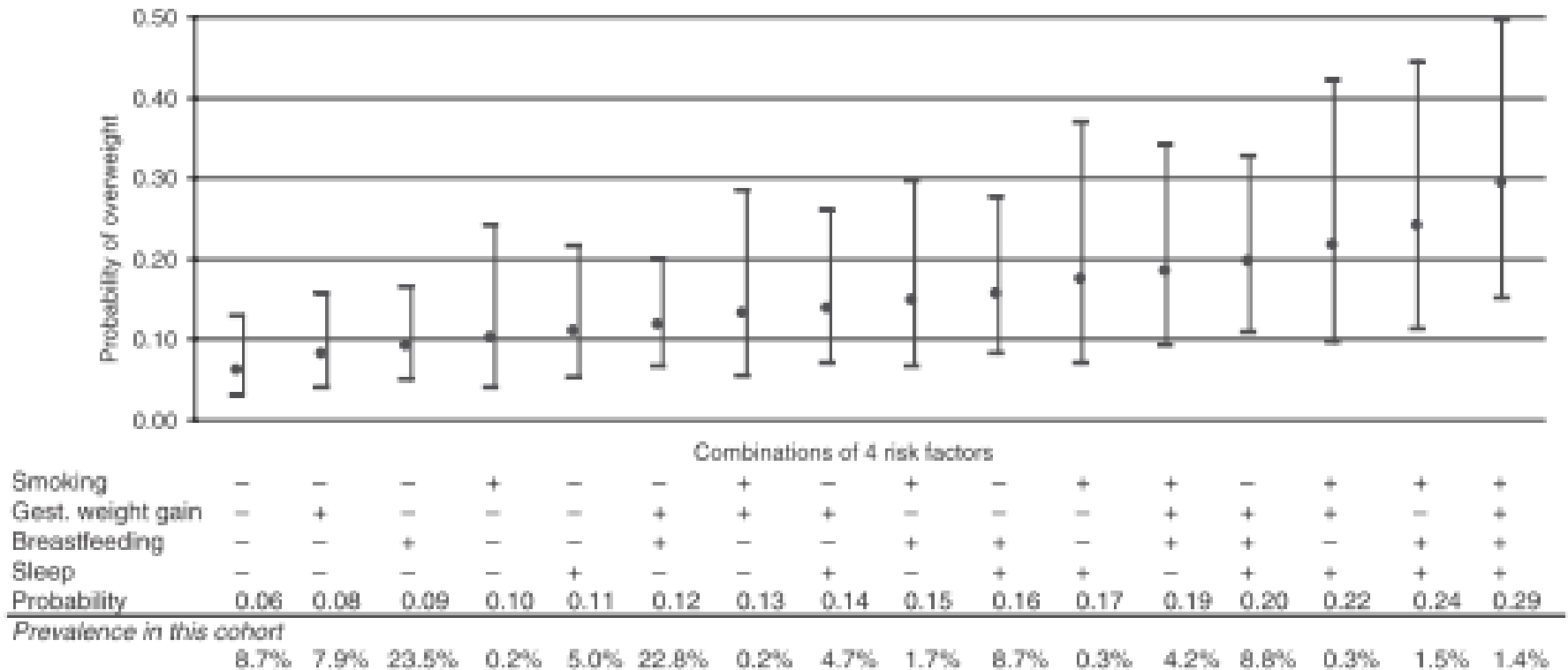
Breastfeeding duration (m)

>12	1.0 (Ref)
<12	1.50 (0.86, 2.64)

Infant sleep (h/day)

>12	1.0 (Ref)
<12	1.83 (1.17, 2.85)

Example of predicted probabilities in practice



Gillman, Rifas-Shiman, Kleinman, et al., Obesity 2008; 16:1651-1656

Recommendations

- If you do a logistic regression, remember that the estimated odds ratios are not estimated risk ratios
- Relative risk model (log Bin or Poisson) likely to perform poorly for multivariate models if the range of $\Pr(y=1)$ is large
- Regardless of which model you use, report predicted probabilities

Recommendations

- Don't use a model just because it's there!
(Powerful software is a blessing and a curse.)
- Choose models which fit the data better, not because the (wrong) parameter estimates are easier to understand
- **Use predicted probabilities to help make model results accessible**