

# Problem Set 2

Carrie Kathlyn Townley Flores, Filipe Recch, Kaylee Tuggle Matheny,  
Klint Kanopka, Kritphong Mongkhonvanit  
EDUC 252L

February 7, 2018

## 1 Breaking the Classical Test Theory Model

### 1.1 Coin Flips

Coin flips should not be reliable data - they're random! To look at this a little more analytically:

$$\alpha = \frac{K}{K-1} \left( 1 - \frac{\sum_{i=1}^K p_i(1-p_i)}{\sigma_X^2} \right)$$

The interesting thing to note here is that the probability of flipping heads is:

$$p_i = 0.5$$

And the variance on the sum of  $K$  coin flips will be:

$$\sigma_X = 0.25K$$

Substituting in the formula for Cronbach's Alpha:

$$\alpha = \frac{K}{K-1} \left( 1 - \frac{\sum_{i=1}^K (0.5)(1-0.5)}{0.25K} \right)$$

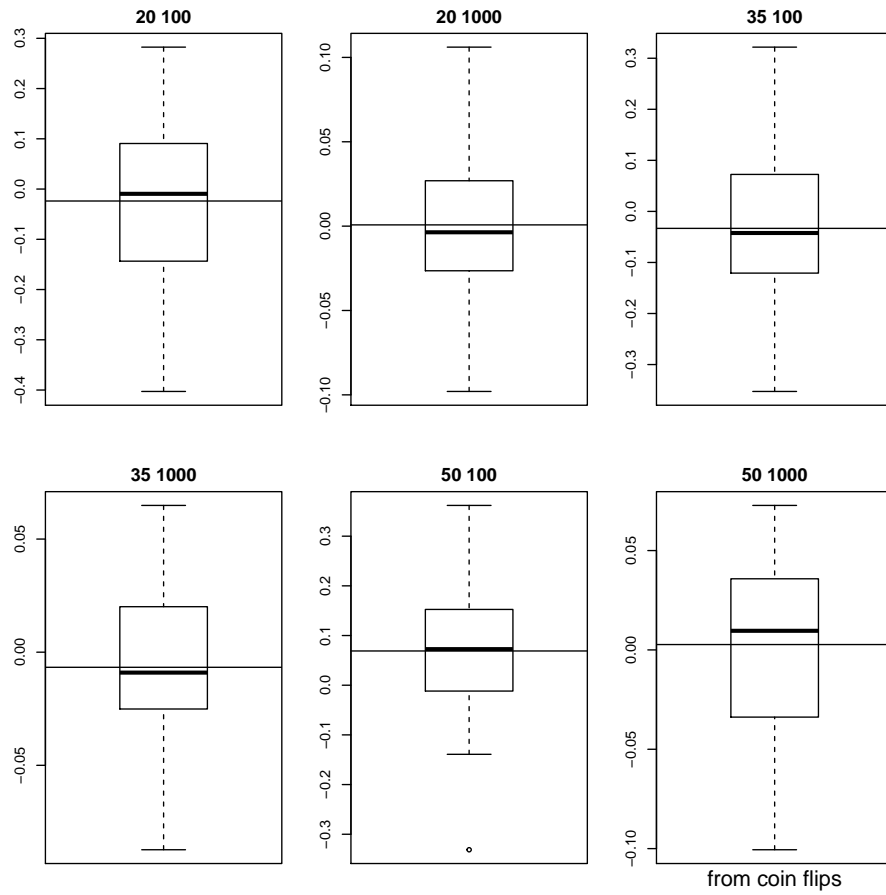
Cleaning up:

$$\alpha = \frac{K}{K-1} \left( 1 - \frac{0.25K}{0.25K} \right)$$

$$\alpha = \frac{K}{K-1} (1-1)$$

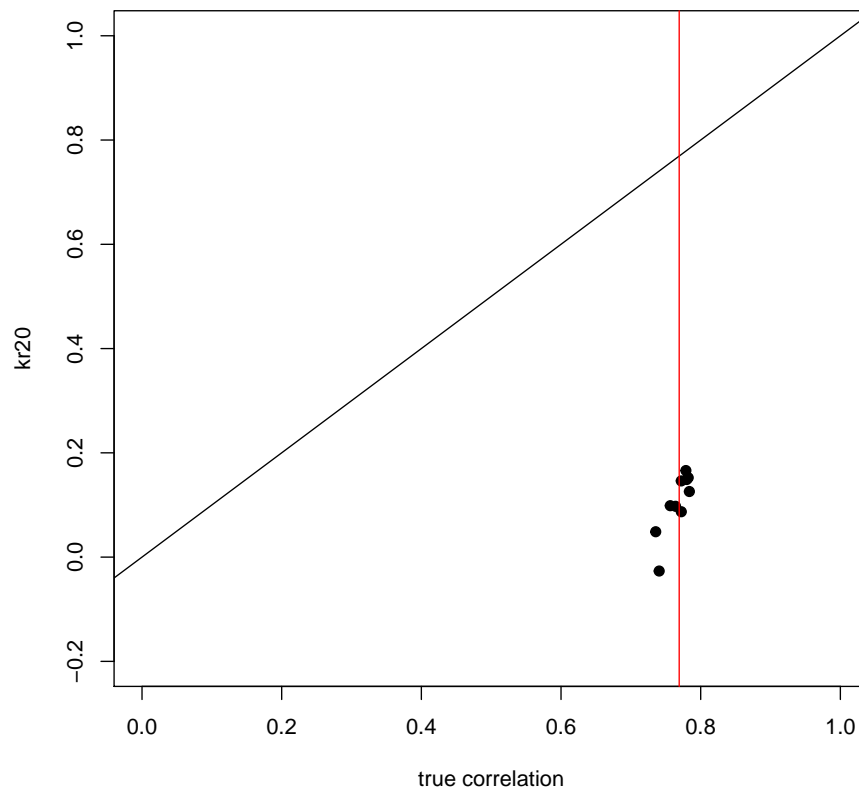
$$\alpha = 0$$

The expectation, then, is that  $\alpha$  should be zero for each situation.

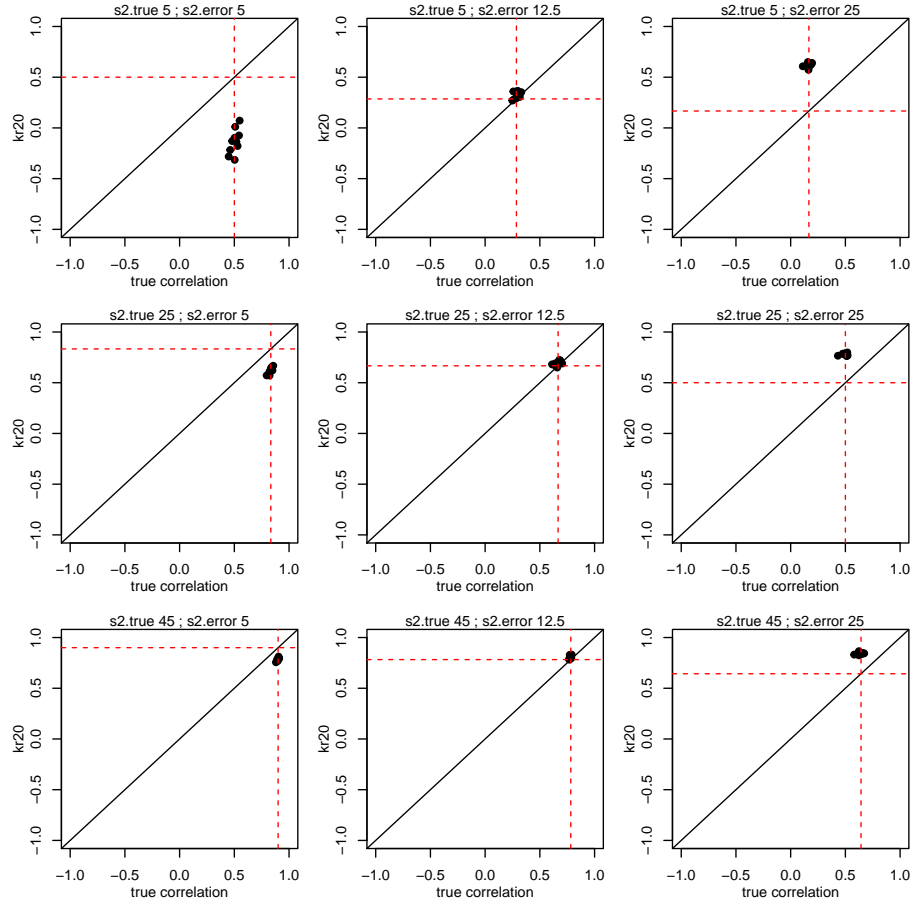


16  
 17 These  $\alpha$  plots make sense - they are centered around zero, as predicted, and  
 18 as the number of items increases,  $\alpha$  is more tightly clustered around zero.

## 19 1.2 Simulating Item Response Data



20  
 21 The feature of the data generation mechanism that makes the  $\alpha$  values super  
 22 low is that items are getting marked correctly (essentially) at random! Even  
 23 though the item responses generated correct p-values and test-level correlations,  
 24 the data generation disregarded any internal structure you would expect. More  
 25 clearly stated, respondents of similar ability levels did not have similar item  
 26 response profiles.

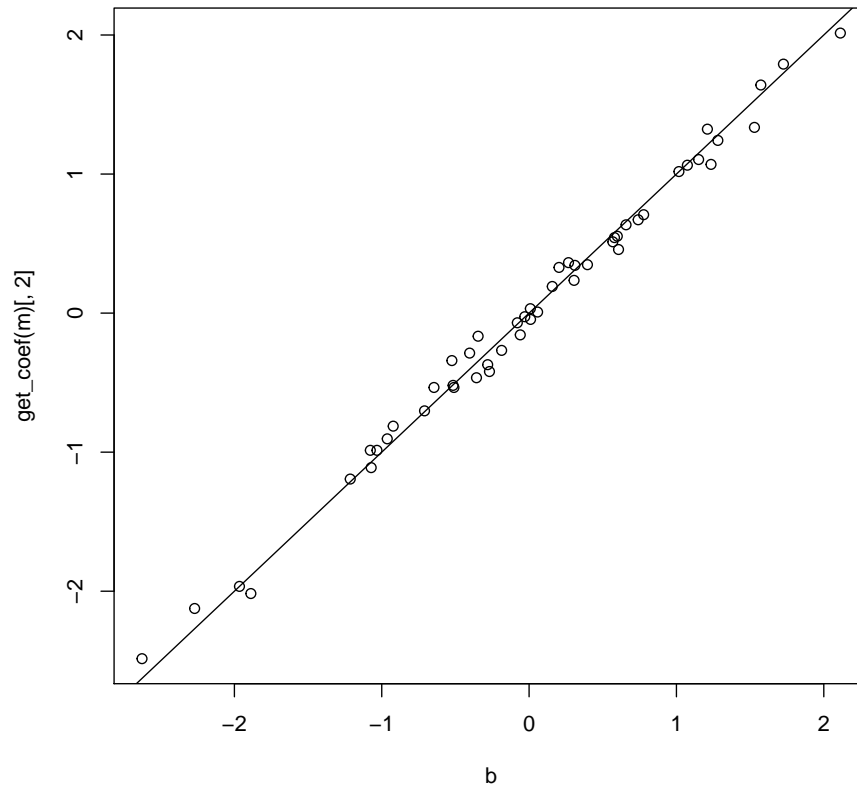


27

28 Looking at the resulting plots, it's clear that even with nonsensical item  
 29 response data, the KR-20 estimate of reliability increases as a function of both  
 30 true score variance and error variance. This *feels* very wrong. Increasing true  
 31 score variance can be done by applying an instrument to a population it may  
 32 not have been originally designed for. Increasing error variance can be done  
 33 by adding more items or manipulating the quality of items. The challenge  
 34 with feeling good about the CTT model is that KR-20 is both heavily valued  
 35 and easily manipulated. The worst part is that some of the behaviors that  
 36 would increase a KR-20 value could have negative impacts on the validity of the  
 37 instrument.

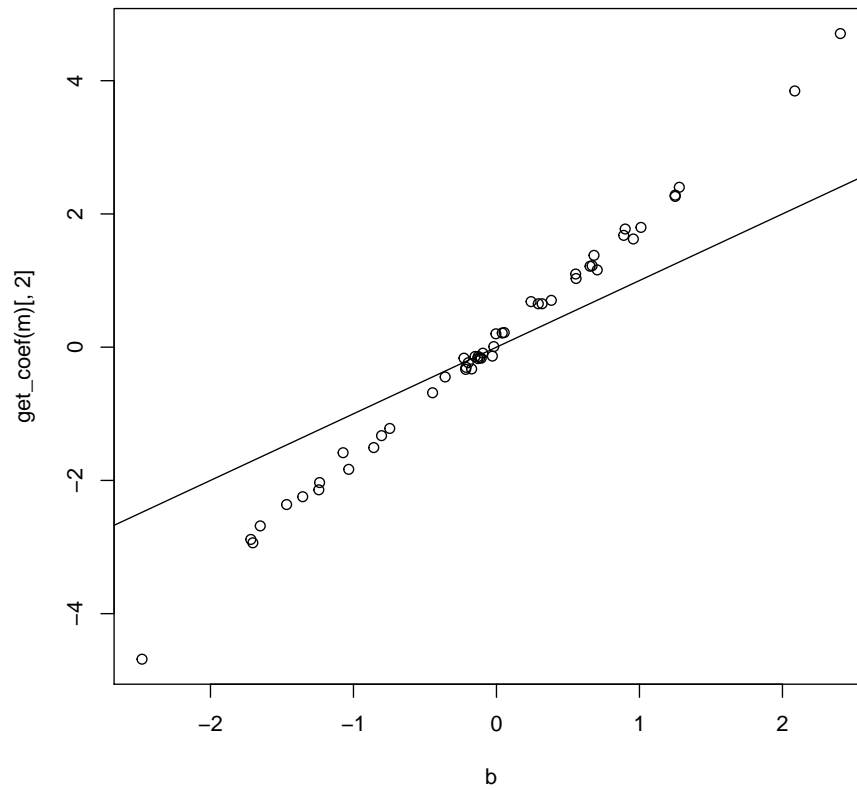
## 2 Different Link Functions

### 2.1 The Default



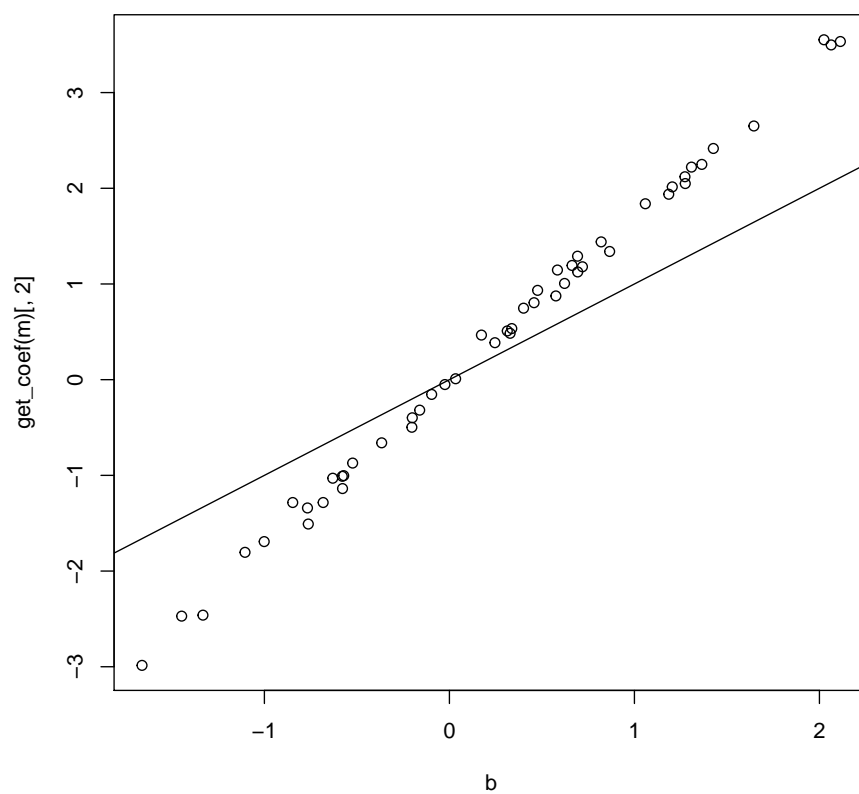
The default estimates item difficulties (or item easiness, specifically, because mirt) that are in line with the actual (specified) item difficulties.

## 43 2.2 The Normal



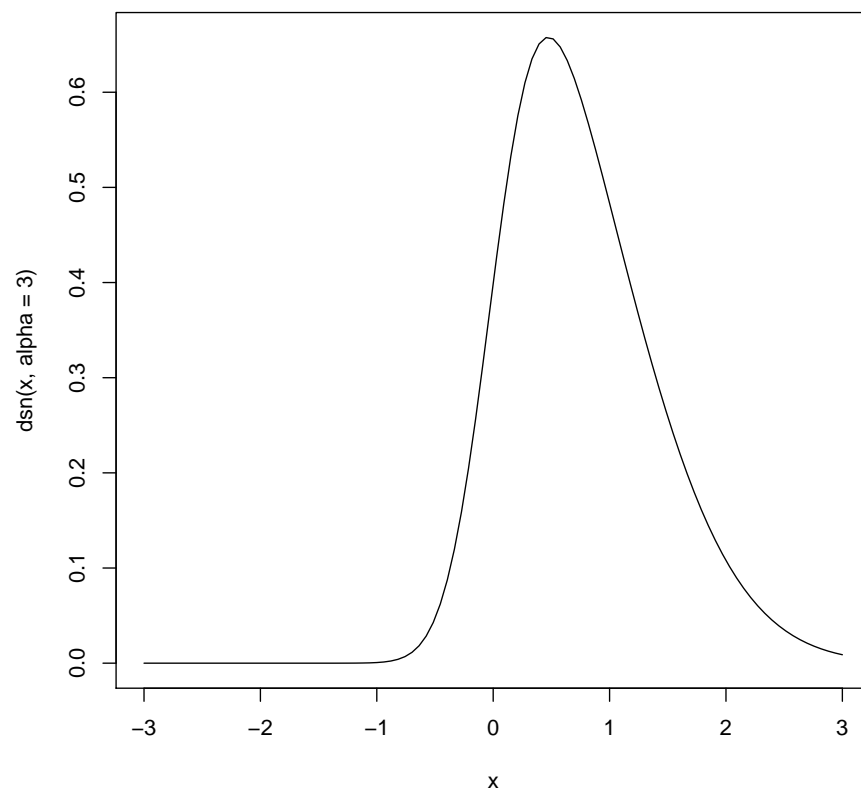
44  
 45 Using a normal link function, mirt predicts easy items are easier than they  
 46 actually are and hard items are harder than they actually are. The farther an  
 47 item is from zero, the larger the gap between estimated difficulty and specified  
 48 difficulty. The slope of the line here is 1.7, so it is possible to transform between  
 49 them, but the community tends to prefer the logistic because it is computation-  
 50 ally simpler (even if the normal is theoretically nicer).

## 51 2.3 Heavy Tails



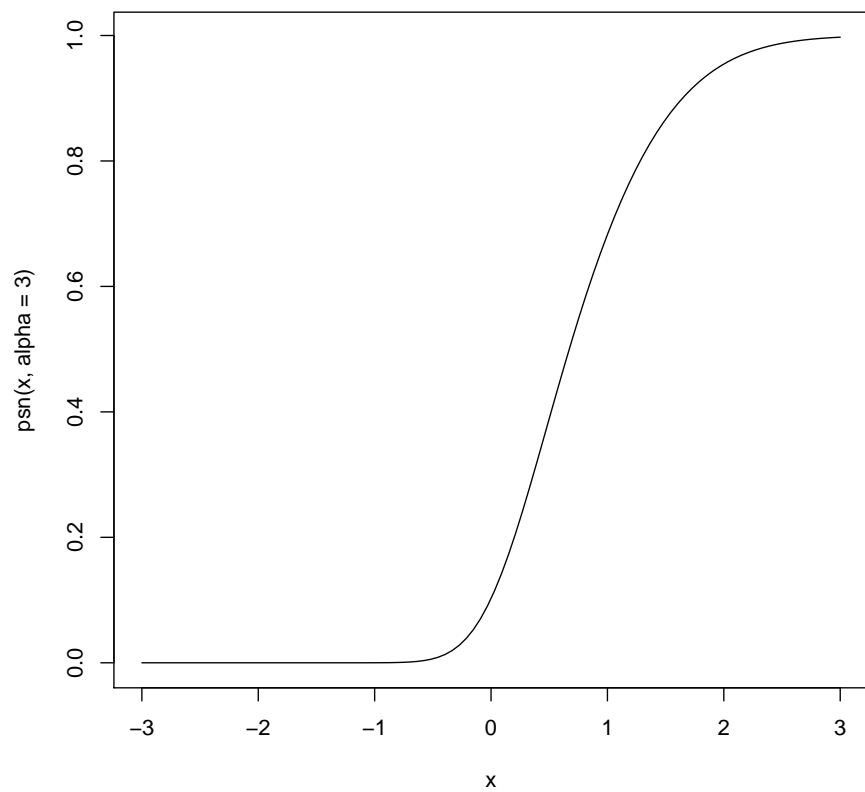
52  
 53 Using a link function with heavy tails, mirt does essentially the same as  
 54 above, just with more divergence farther from zero.

55 **2.4 Skewed**

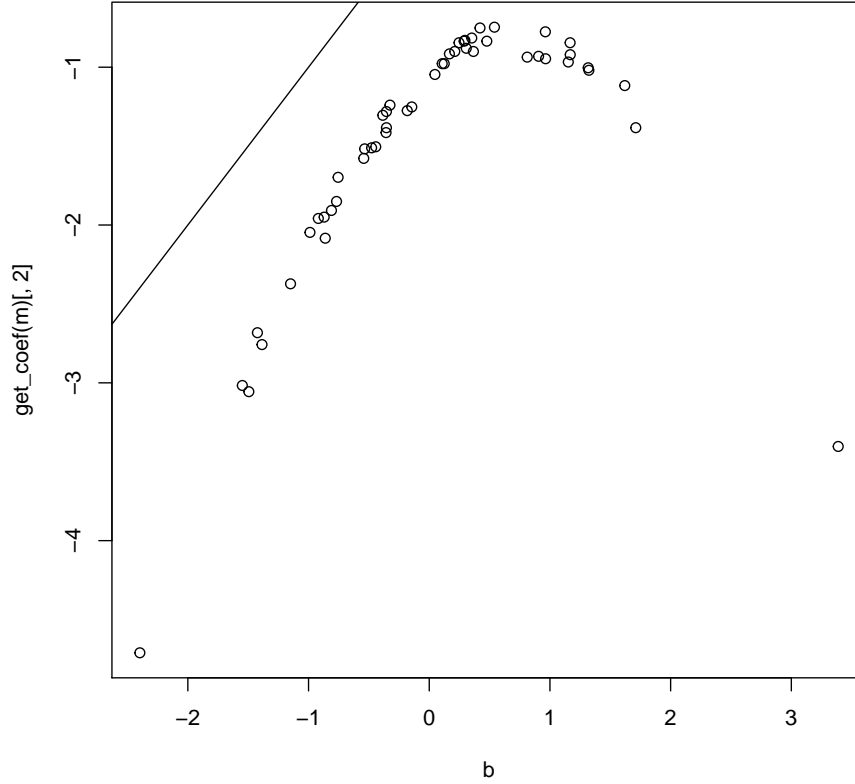


56





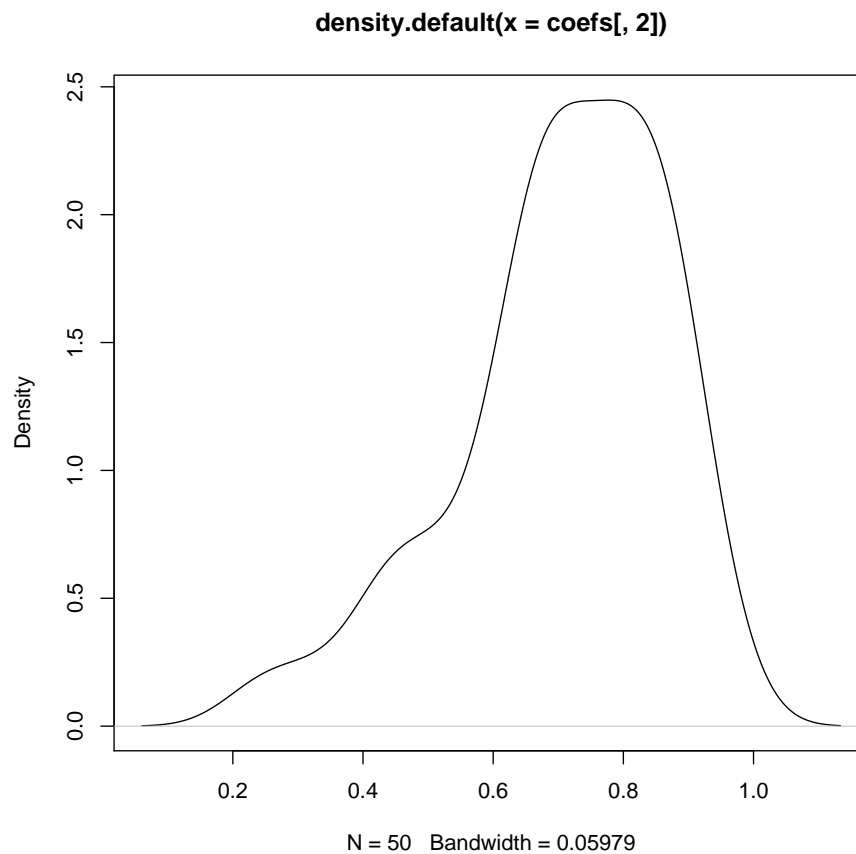
57



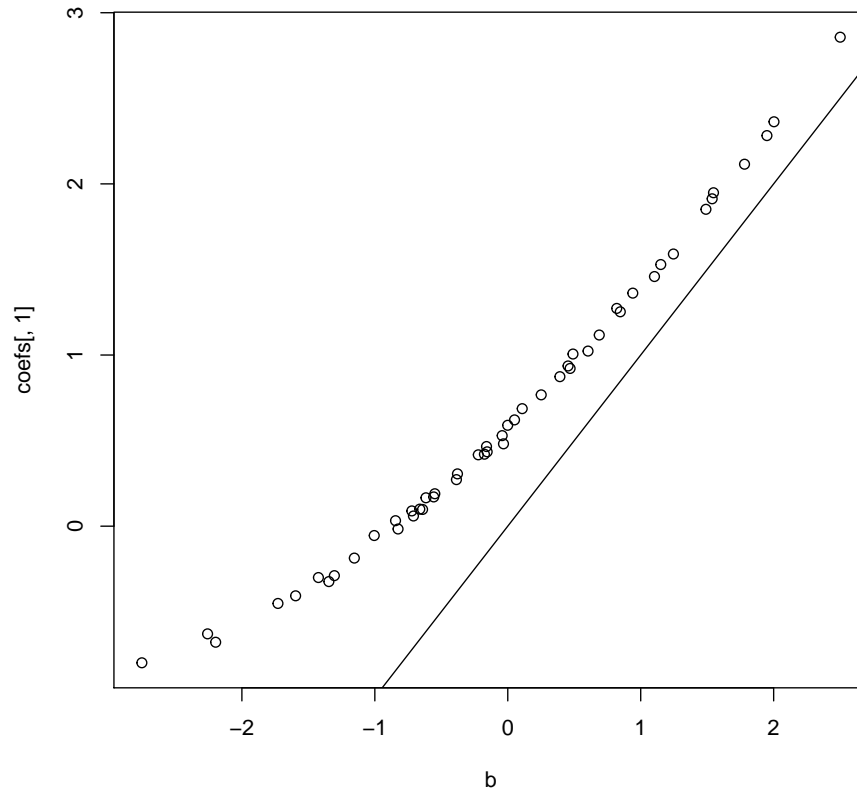
58

59 When using a skewed distribution as the link function, every item is esti-  
60 mated as being harder than it actually is, but where this model really tanks  
61 is for extremely easy items. It finds those to be significantly harder than they  
62 actually are and the difference between actual and estimated difficulty for easy  
63 items diverges wildly. The skew on the link function is set up in a way that  
64 there is essentially no impact of increased  $\theta$  on the respondent's probability of  
65 getting the item correct until  $\theta$  gets to a value of  $-1$ . Remember that the input  
66 to the model is of the form  $\theta_i + b_j$ . If trait values less than  $-1$  don't provide any  
67 increase in probability of scoring correctly on an item, for items with easiness  
68 above 1, the model will view those as significantly harder than they actually  
69 are. If you look at the  $b_{estimated}$  vs.  $b$  curve, you can see that at  $b = 1$ , the  
70 estimated easiness starts to take a turn away from actual easiness. This is a  
71 product of the link function being non-symmetric, essentially causing the two  
72 halves of the graph to be scaled differently.

### 73 3 Adding A Lower Asymptote



74  
 75 When moving the lower asymptote above zero, we simulate guessing. In  
 76 the original model, with no guessing parameter, the discrimination density plot  
 77 was centered around 1. Now, it is centered around 0.75. This happens because  
 78 raising the floor of the logistic curve fundamentally changes its shape, lowering  
 79 the discrimination.



80  
 81 When looking at the item easiness estimates, the model overestimates eas-  
 82 iness. This is especially true for the harder questions(easiness less than zero),  
 83 where the estimate starts to diverge from the actual difficulty. One explana-  
 84 tion for this is that for harder questions, a larger segment of the population is  
 85 benefiting from guessing, where on easier questions, a smaller segment of the  
 86 population benefits from guessing (because they were already getting that item  
 87 correct).