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:

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

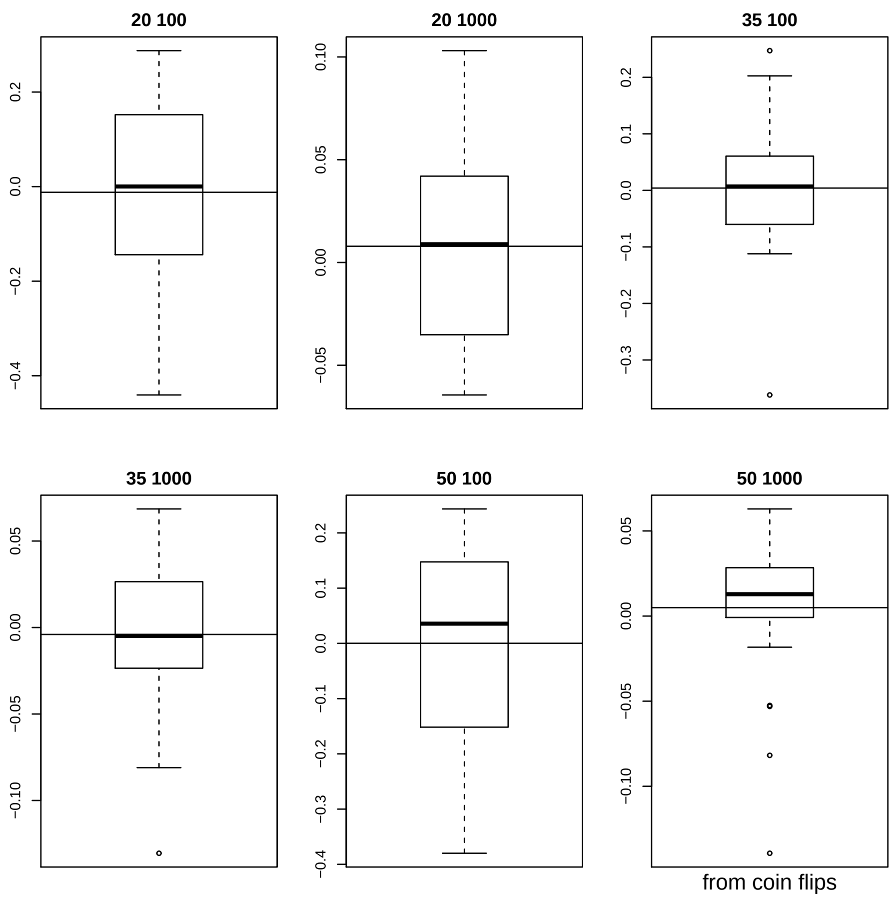
Cleaning up:

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

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

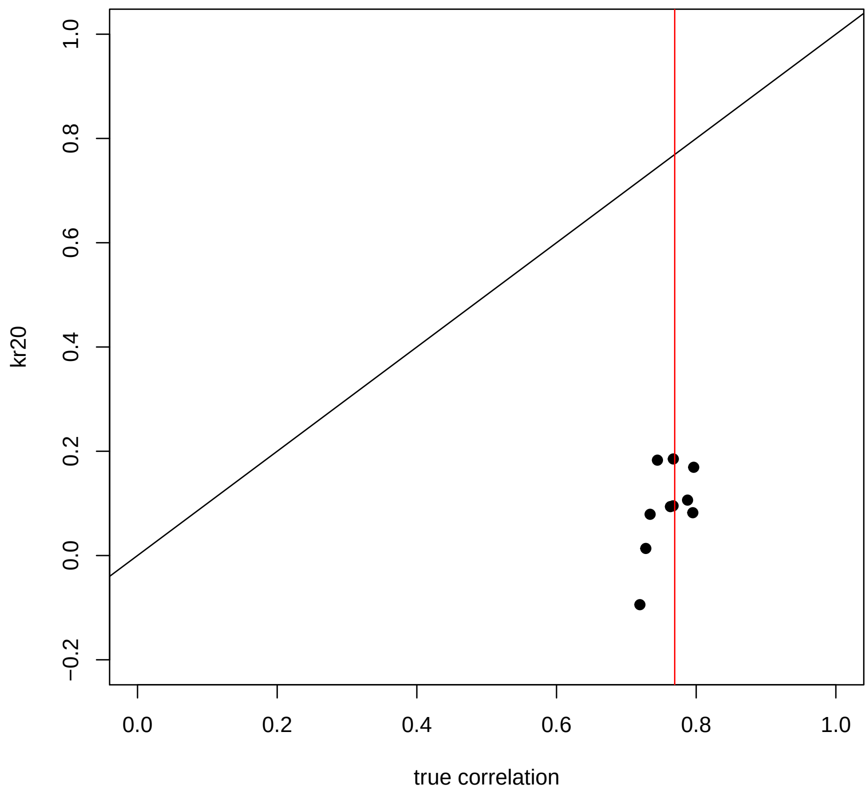
$$\alpha = 0$$

The expectation, then, is that α should be zero for each situation.

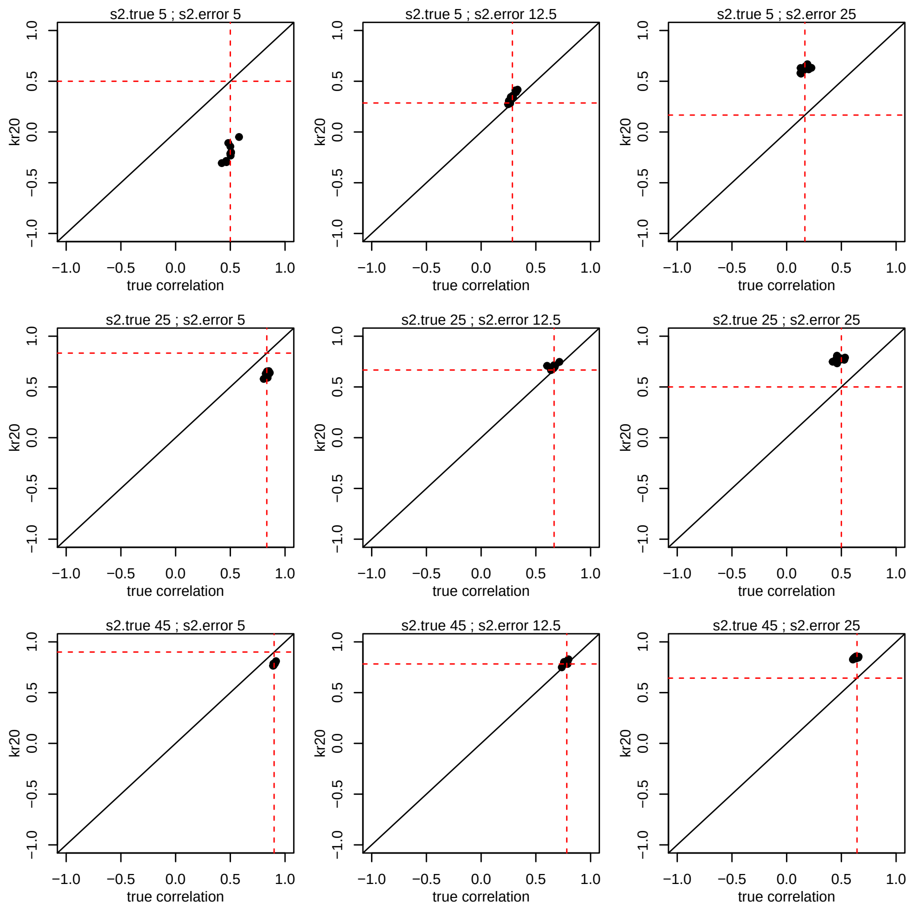


These α plots make sense - they are centered around zero, as predicted, and as the number of items increases, α is more tightly clustered around zero.

1.2 Simulating Item Response Data

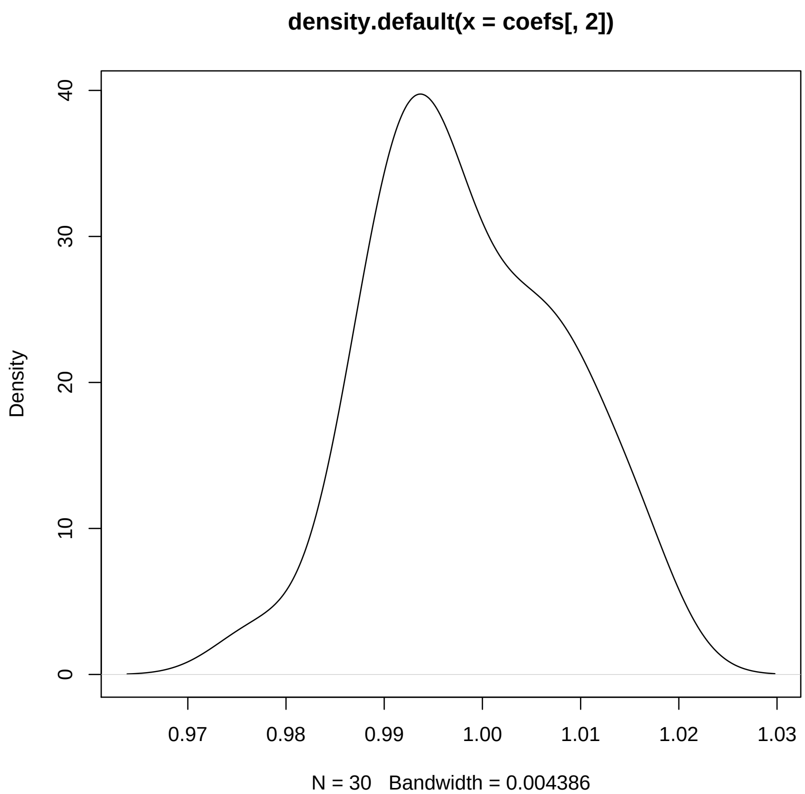


The feature of the data generation mechanism that makes the α values super low is that items are getting marked correctly (essentially) at random! Even though the item responses generated correct p-values and test-level correlations, the data generation disregarded any internal structure you would expect.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



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

2 Basic Structure of IR Models

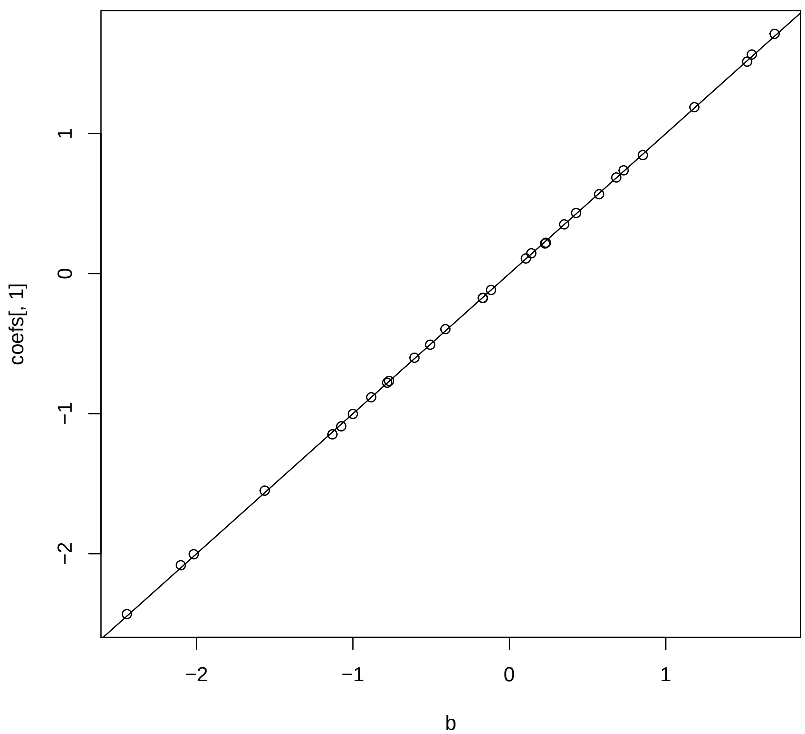


1.

The plot shows the density of the item discriminations (i.e. the coefficients of the thetas in the equation $\beta_0 + \beta_1\theta$), which means the density of the items' discrinations. The more the sample size is increased, the closer it appears to a normal distribution with a mean at 1. That means the discriminations are near 1. This makes sense because earlier in the code, we had set

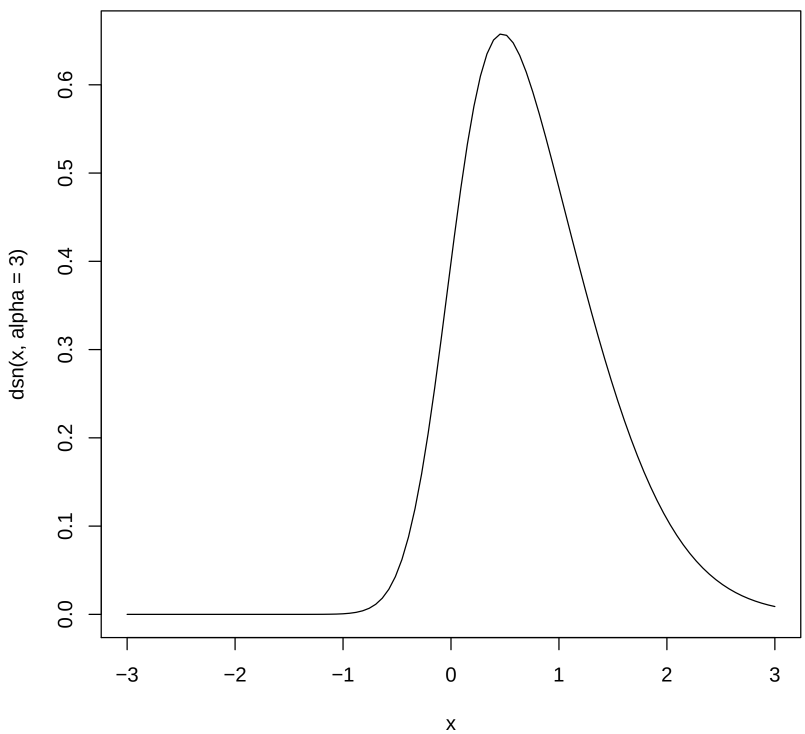
```
a <- 1
```

and a is the item discrimination coefficient.

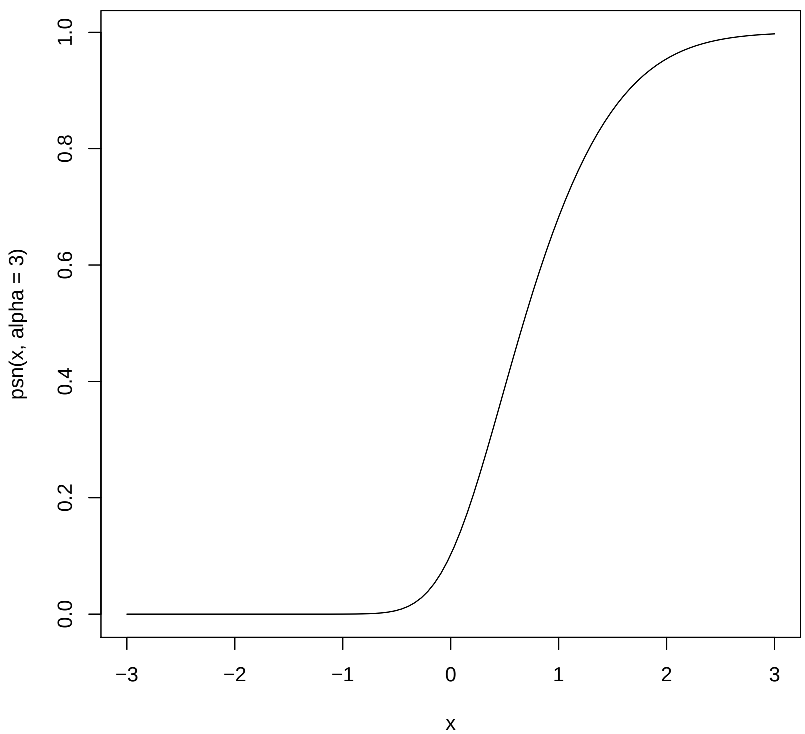


2.

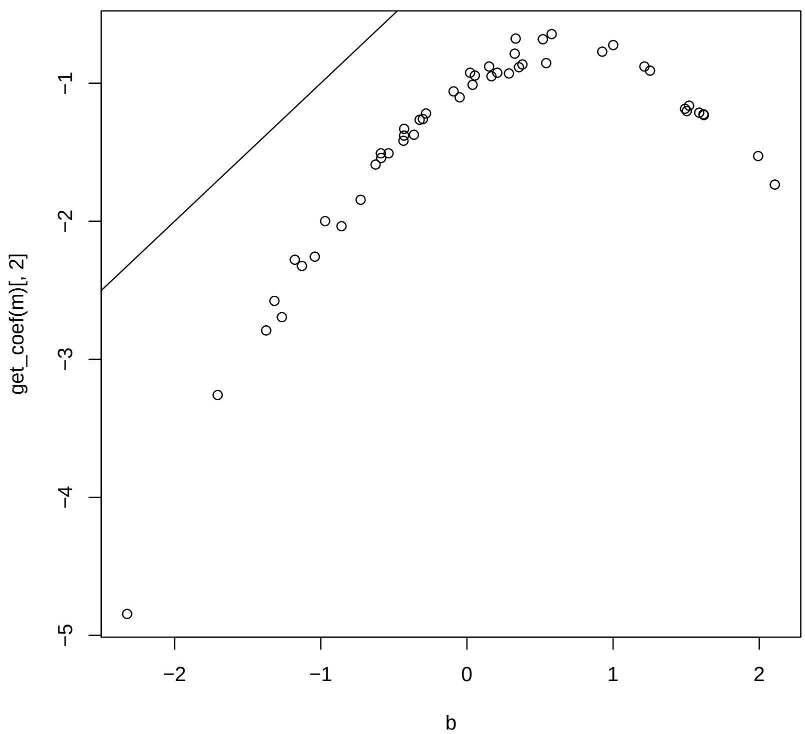
This plot shows that the b's are almost the same as the intercepts, which makes sense because it means that the actual item easiness is close to the estimated easiness.



3.



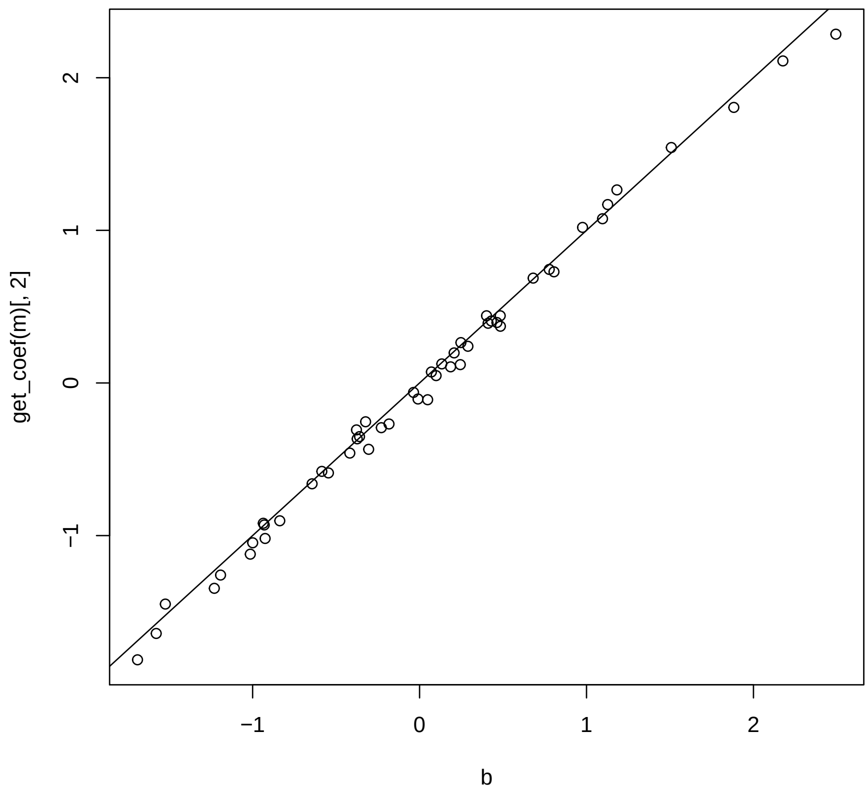
52



This is showing that when items get easier, the estimation gets worse and actually predicts that they're harder.

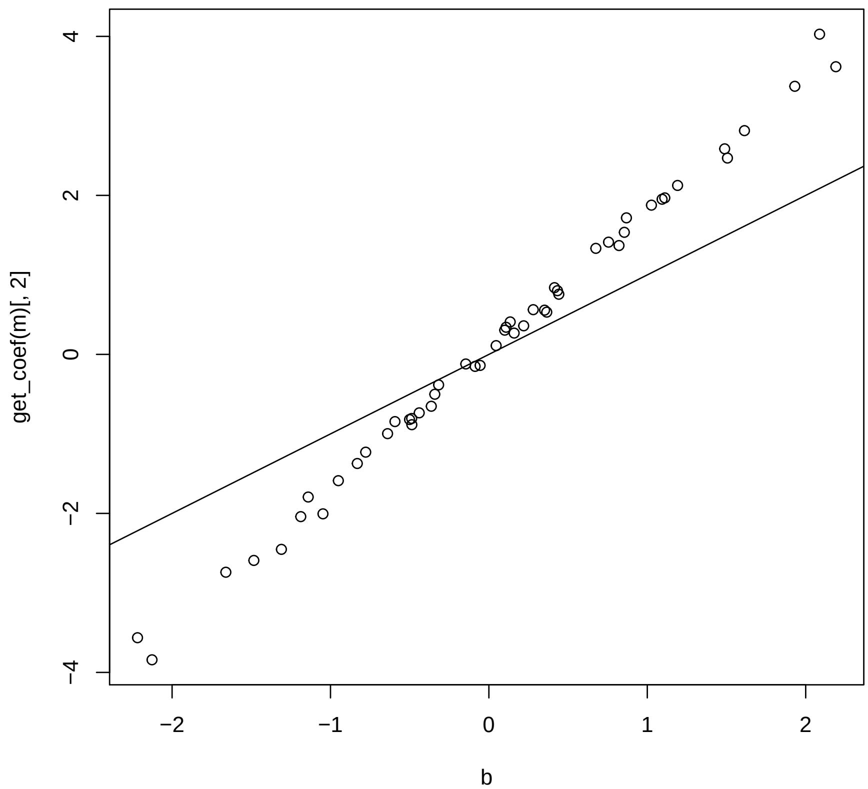
3 Different Link Functions

3.1 The Default



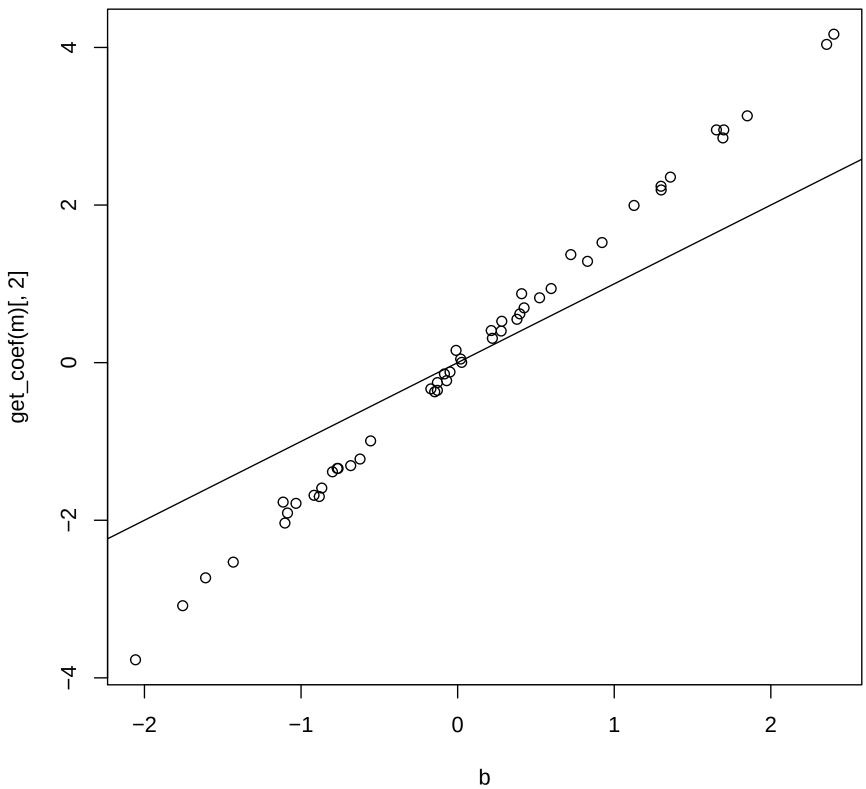
The default estimates item difficulties (or item easiness, specifically, because mirt) that are in line with the actual (specified) item difficulties.

3.2 The Normal



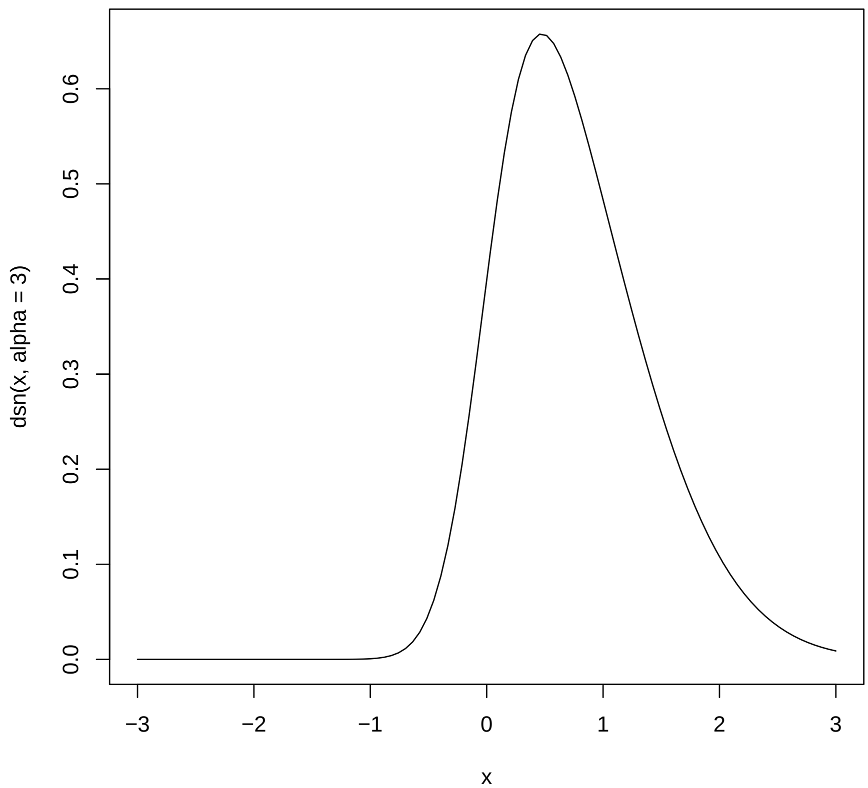
Using a normal link function, mirt predicts easy items are easier than they actually are and hard items are harder than they actually are. The farther an item is from zero, the larger the gap between estimated difficulty and specified difficulty. The slope of the line here is 1.7, so it is possible to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

3.3 Heavy Tails

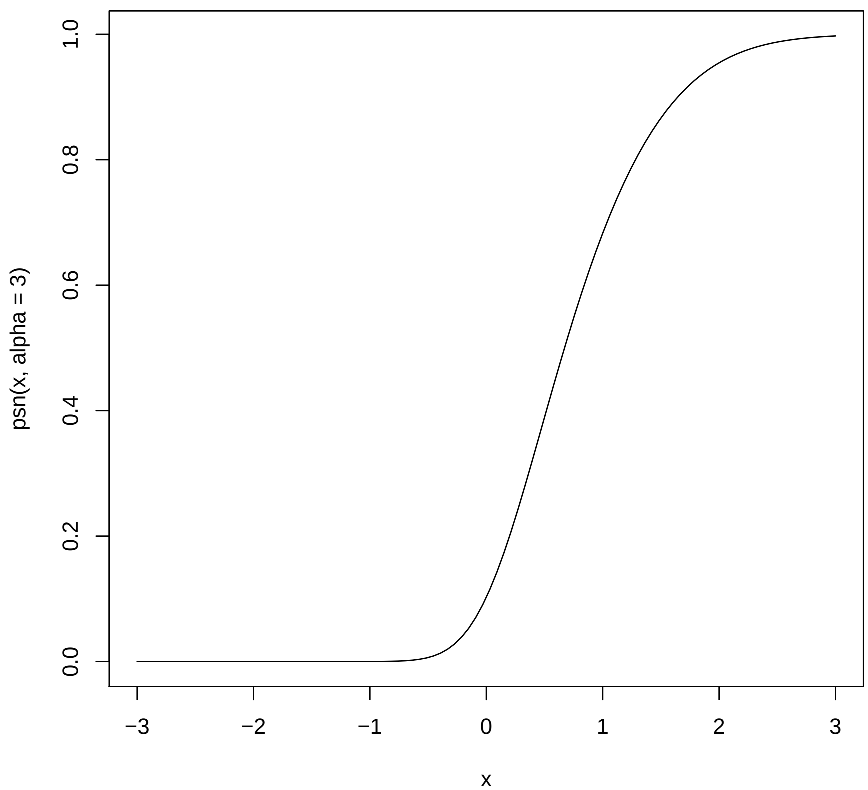


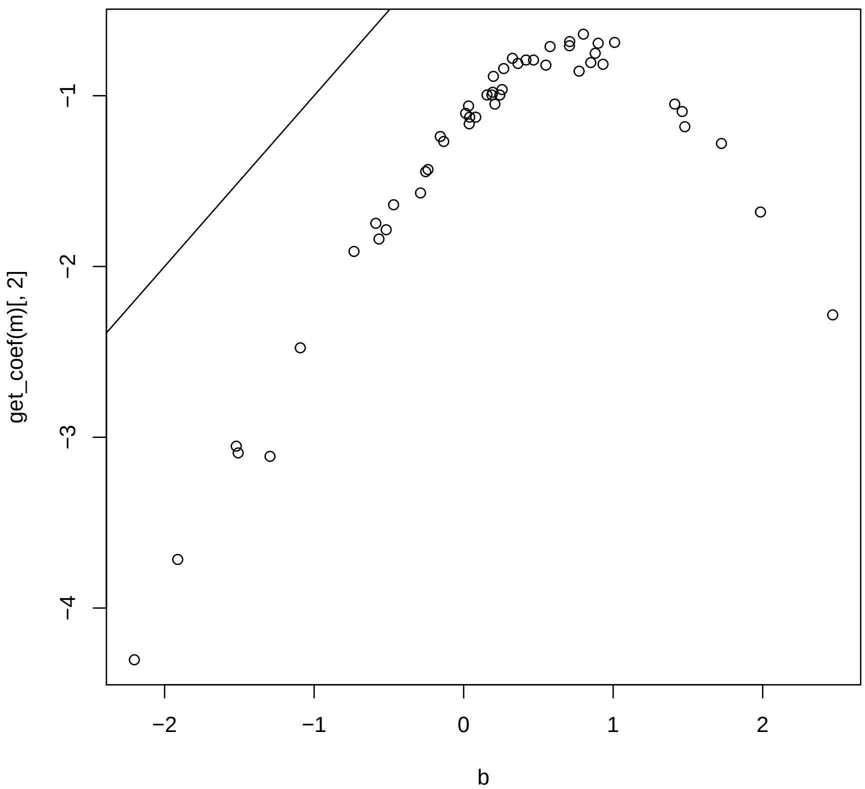
Using a link function with heavy tails, mirt does essentially the same as above, just with more divergence farther from zero.

3.4 Skewed



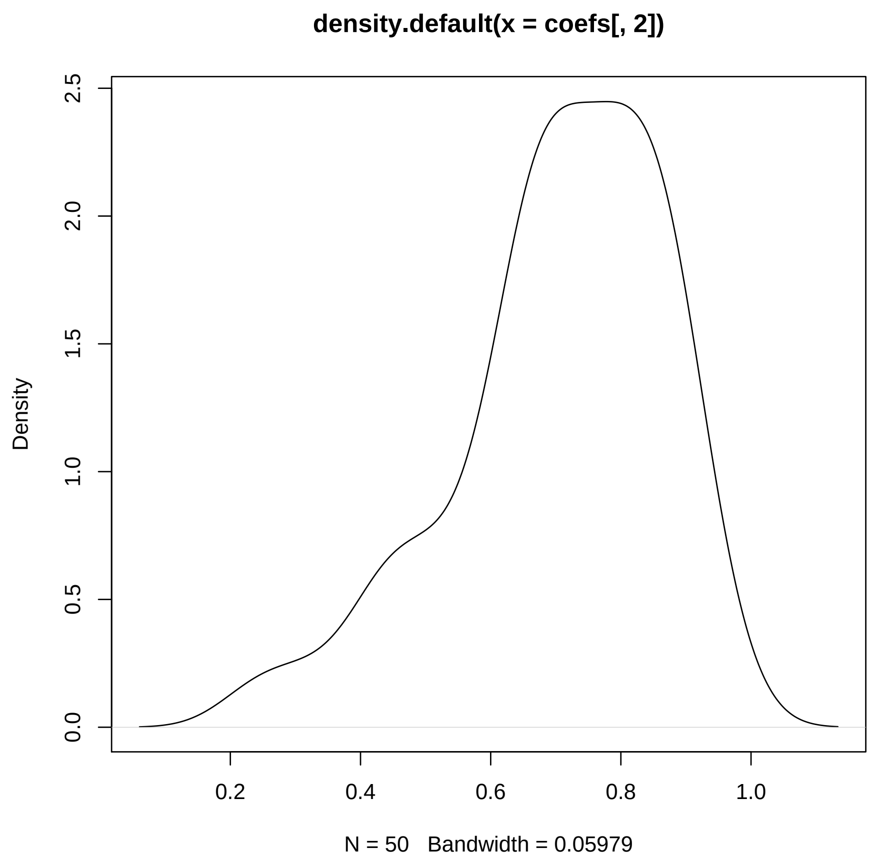
74


psn(x, alpha = 3)
0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3
x

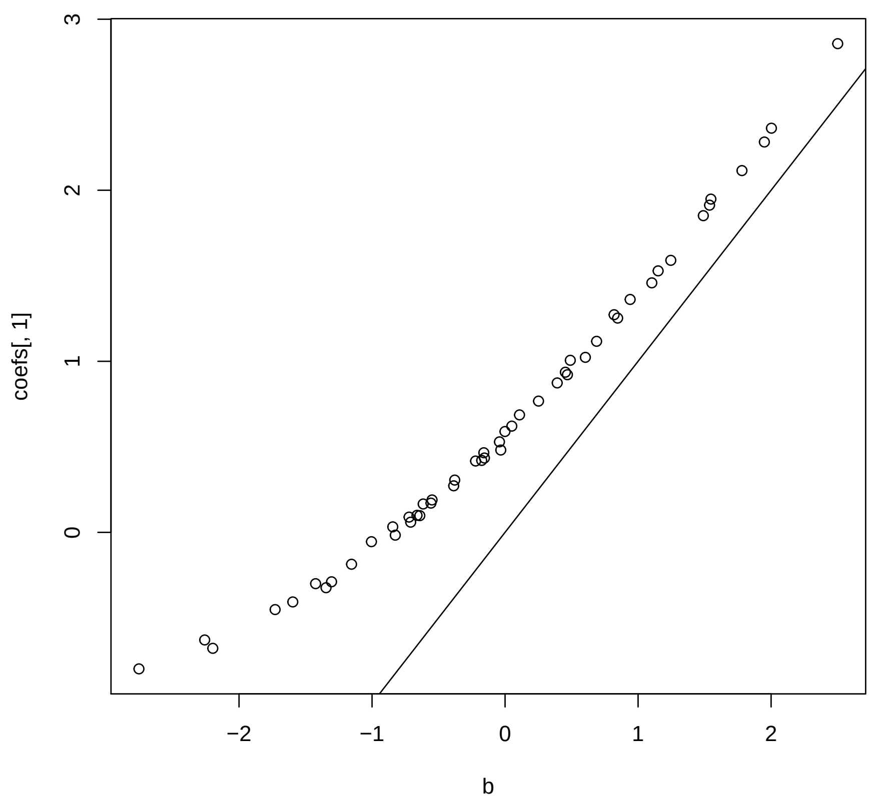


When using a skewed distribution as the link function, every item is estimated as being harder than it actually is, but where this model really tanks is for extremely easy items. It finds those to be significantly harder than they actually are and the difference between actual and estimated difficulty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

4 Adding A Lower Asymptote



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



When looking at the item easiness estimates, the model overestimates easiness. This is especially true for the harder questions(easiness less than zero), where the estimate starts to diverge from the actual difficulty. One explanation for this is that for harder questions, a larger segment of the population is benefiting from guessing, where on easier questions, a smaller segment of the population benefits from guessing (because they were already getting that item correct).