I was thinking about the selection of subjects and I am not sure about some things, I maybe a bit confused after we had tried and done so many things and even now the final results have been derived in several steps, so I feel like I might be missing the big picture.
So, first we had taken the residuals of :

$bmi=b_0+b_1*age+b_2*age^2+b_3*gender+b_4*year+b_5*ffq$….with all their errors….

and then we got the effect sizes:

$bmi\_residuals=b_0+b_1*nutrient1+b_2*nutrient2+….+b_{21}*nutrient21$ (b_0 and some other N.S.)….with all errors….

and constructed diet score:$b_1*nutrient1+b_2*nutrient2+….+b_{21}*nutrient21=diet\_score$

and did the same, but added PA(both scaled and 1 subtracted from PA):

$bmi\_residuals=b_0+b_1*diet\_score+b_2*PA$….with all errors….

and got :

b_0= -0.003038 N.S.
b_1= 0.201390 <2e-16
b_2= -0.111493 <2e-16

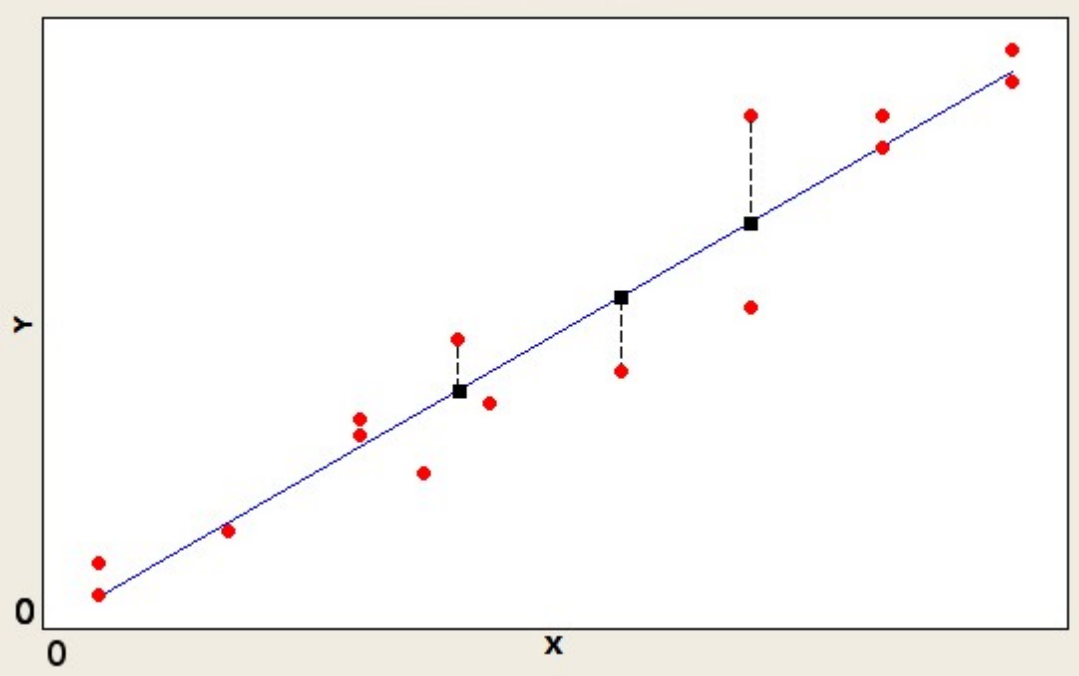
and constructed environment score: 0.201390*diet_score+-0.111493*PA= environment_score

this will give us :

bmi_residuals~environment_score , since :

b_0=-0.003038 N.S.
b_environment_score=1 <2e-16 since we just added the two terms together

and we end up being interested in these residuals:



except that I got this from google and our data looks nothing like that :)
when we look at the ratio of the bmi z-score and environment z-score, we take the above equation, but instead of the residuals we look at bmi z-score and environment z-score and look at

bmi_z_score=environment_z_score…..1=bmi_z_score/environment_z_score

so, where ever this ratio is bigger then one, then the residuals are above the line, but when the ratio is smaller then one, the residuals are below the line, spanning through entire bmi_z_score, so to ensure we only select the normal/underweight ones we take only those that have corresponding bmi < 25

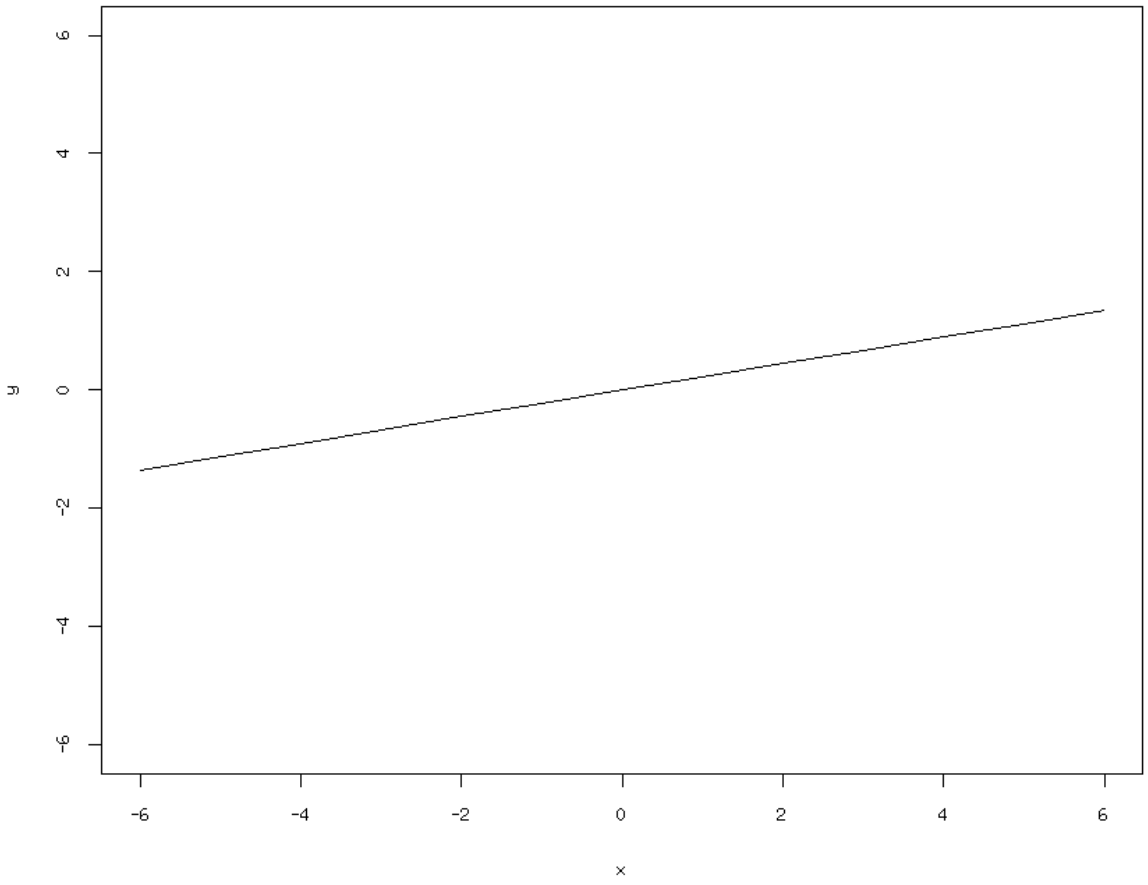
but if you look at the relationship between the z-score of the residuals from the basic covariates on which environment score was build on and the z-score of the environment, the equation is:

bmi_residuals_z_score=0.2329*environment_z_score….with all errors….

since we modeled the residuals ok, the equation for bmi_z_score and environment z-score is almost the same:

bmi_z_score=0.2246*environment_z_score….with all their errors….

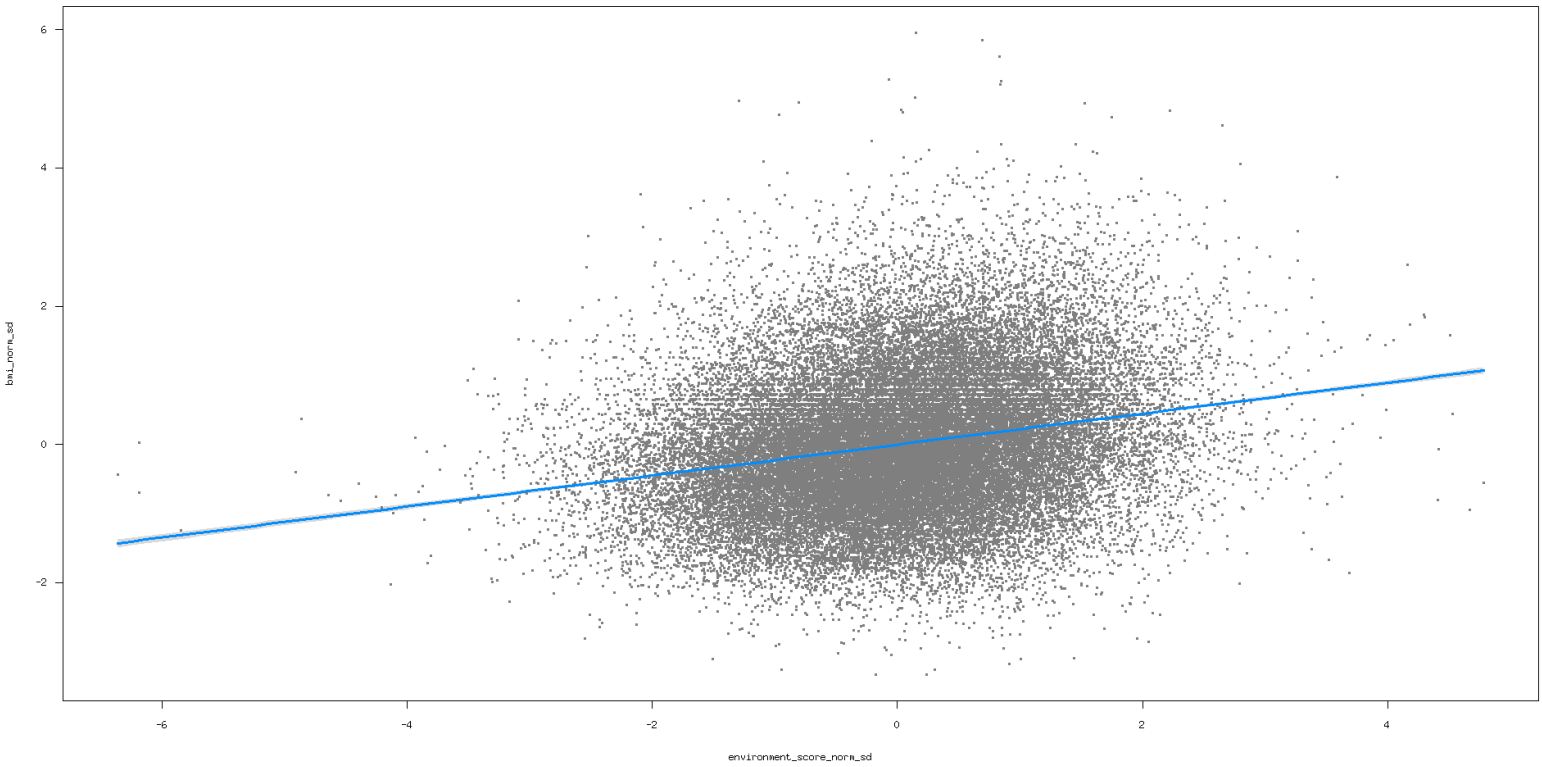
but that means that bmi_z_score/environment_z_score=4.45236:



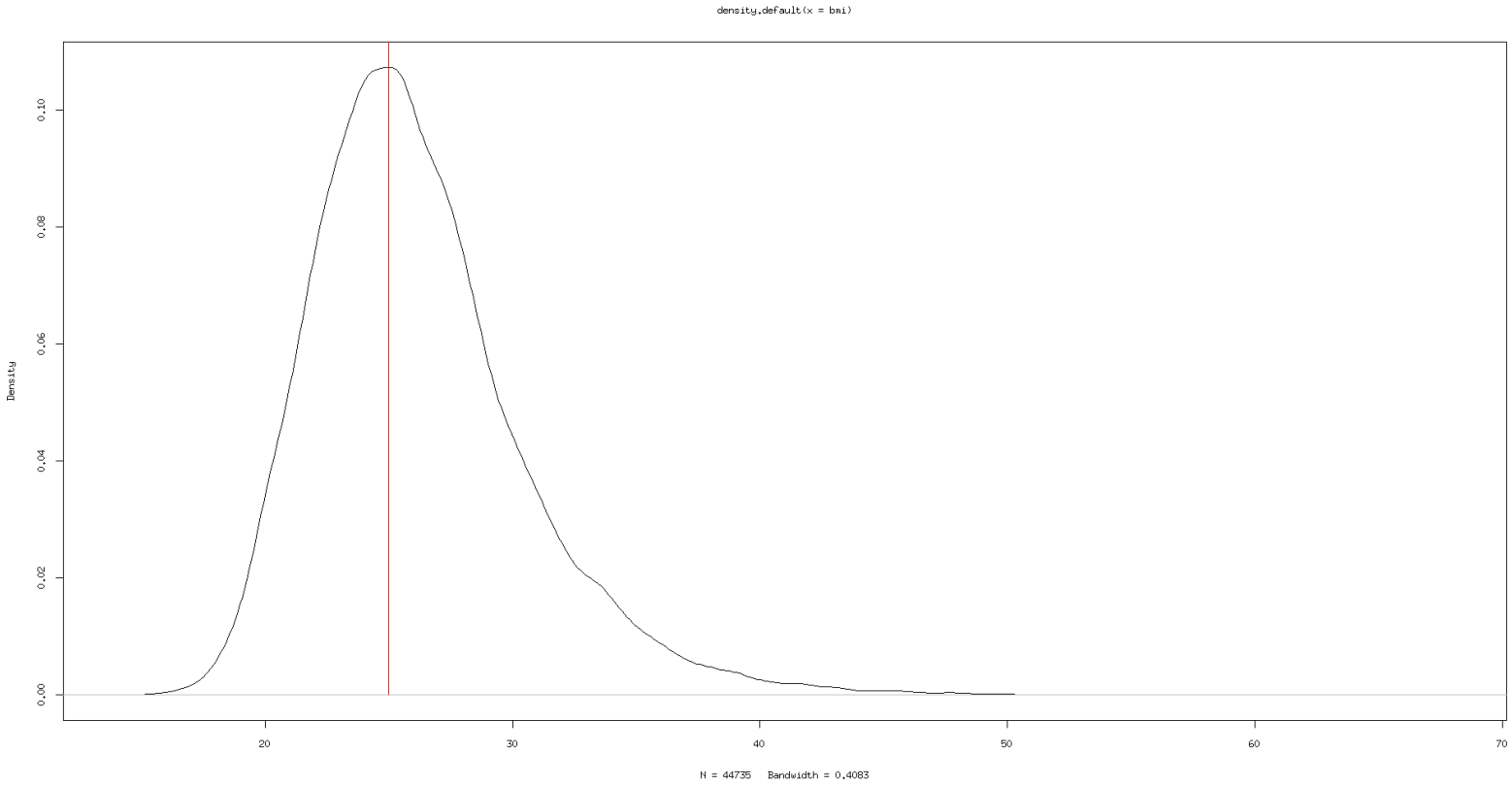
but we are looking at ratio smaller then 1 or bottom 25% of the z-score of this ratio...should we really do that?
Lets say we transform the environment variable into fitted values, like we have been doing all the time before, so new variable = 0.2246*environment_z_score, then we get bmi_z_score=new_variable and do what we want to do, get those that have smaller bmi_z_score...we are in fact looking at the residuals of the simple y=x model, which has been simplified from

bmi=age+agesqr+gender+year+ffq_factor+nutrient1+...+nutrientN+TEI+PA

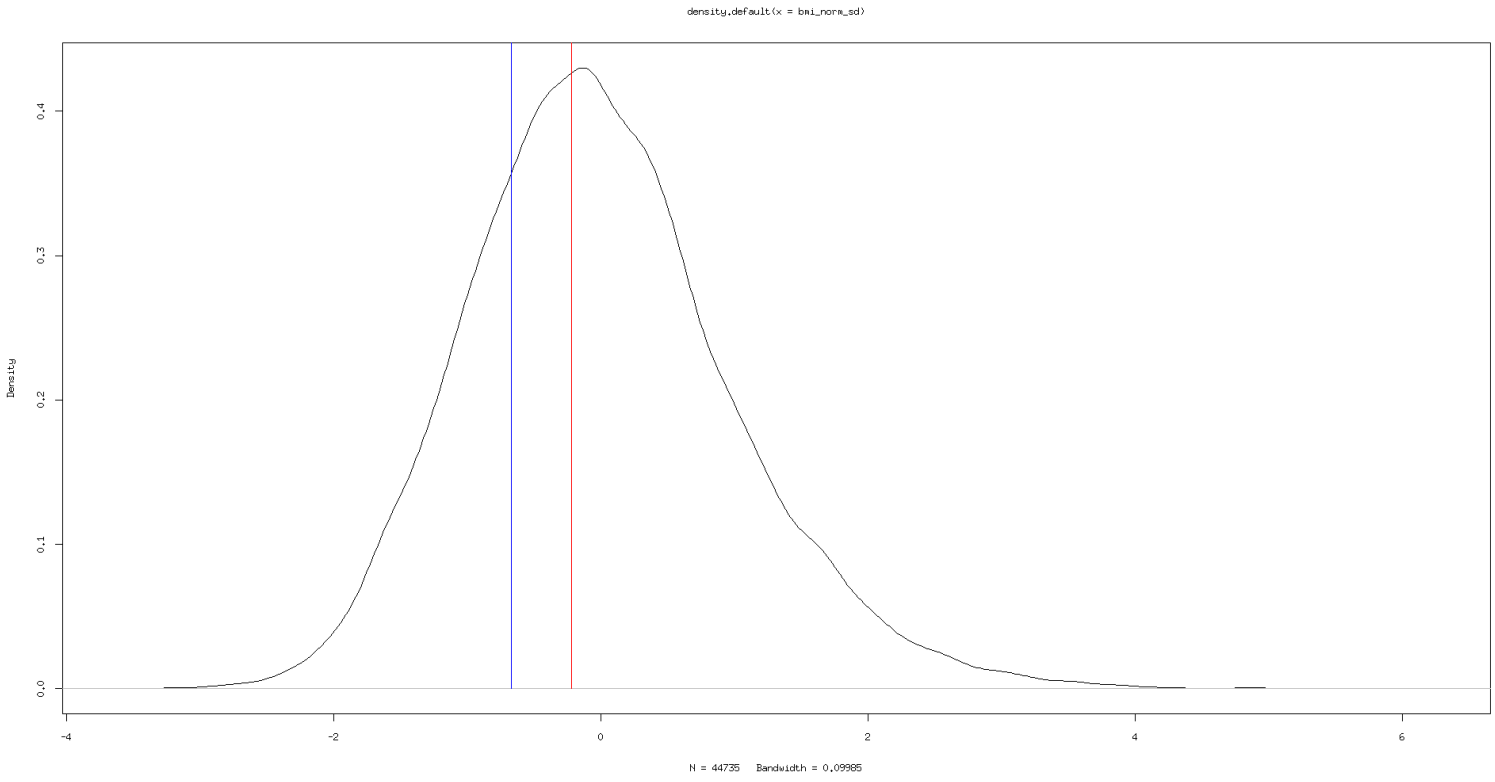
which is a simple multiple regression model and we are selecting those that dont fit that model, the way our data looks like:



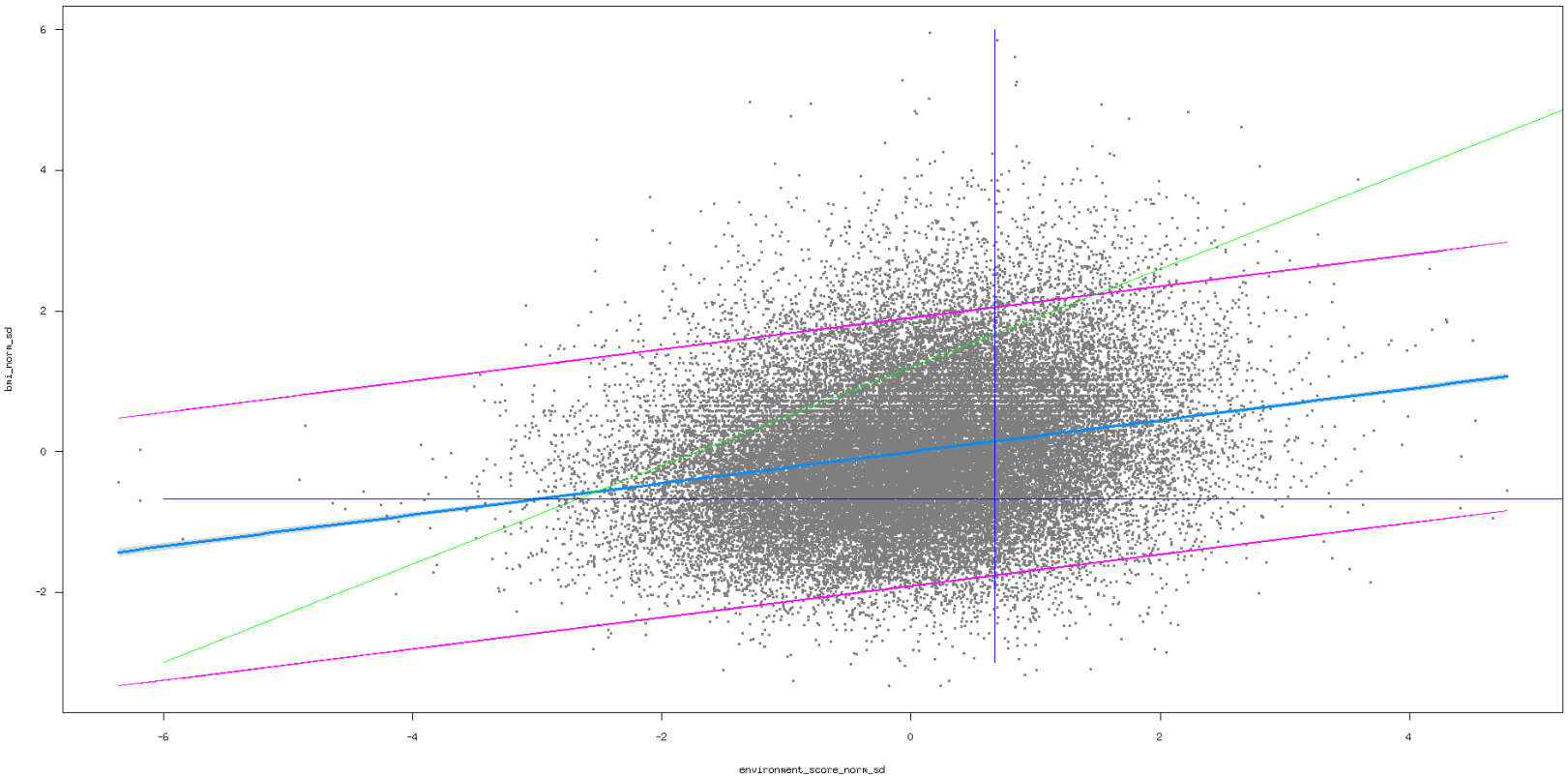
that slight gray thickness around the regression line are 0.95 confidence intervals, so basically almost all our data will fall outside, but we wish to concentrate only on certain individuals, by either looking somewhere below this line and taking those that have normal bmi or take those that are in the top 25% of environment z-score and bottom 25 % of the bmi z-score, which are quite lean people, since bmi looks like this with red line at 25:



which on the z-score is like this with blue at 25%:



So looking back at the model, confidence intervals in red, bottom 25% of bmi z-score and top 25% of environment z-score in blue, identity line in green, prediction intervals in magenta:



we are selecting those that are in bottom-right division, so very little as we have already observed with the actual number of selected individuals…
with the difference that in looking at the ratio smaller than one, we take entire bottom division(left and right) or take those that are in the bottom 25% of the ratio, which will take those that are more far from the fit, but still all in the entire bottom(left and right)

but this environment score is not really the environment score, it is our model of the environment score, so we are looking at the values that dont fit this model. We have such a large sample size that it is not hard to fit something significant, but we use that model to examine those that dont fit it, focusing on normal/underweight subjects and I am worried that just because we have a vector representing our "environment score" we are forgetting it is actually just a very simple model and we have no idea what kind of environment it really represents and what kind of error we are looking at, I know we have discussed this same thing at the meeting the other day, but we havent really resolved anything, apart from deciding to work with what ever we have and move on, but the way we select subject should nevertheless take into account those issues and I am not sure how to put those issues in the right context so we dont miss the big picture when selecting the subjects.