Bayesian Regression Methods

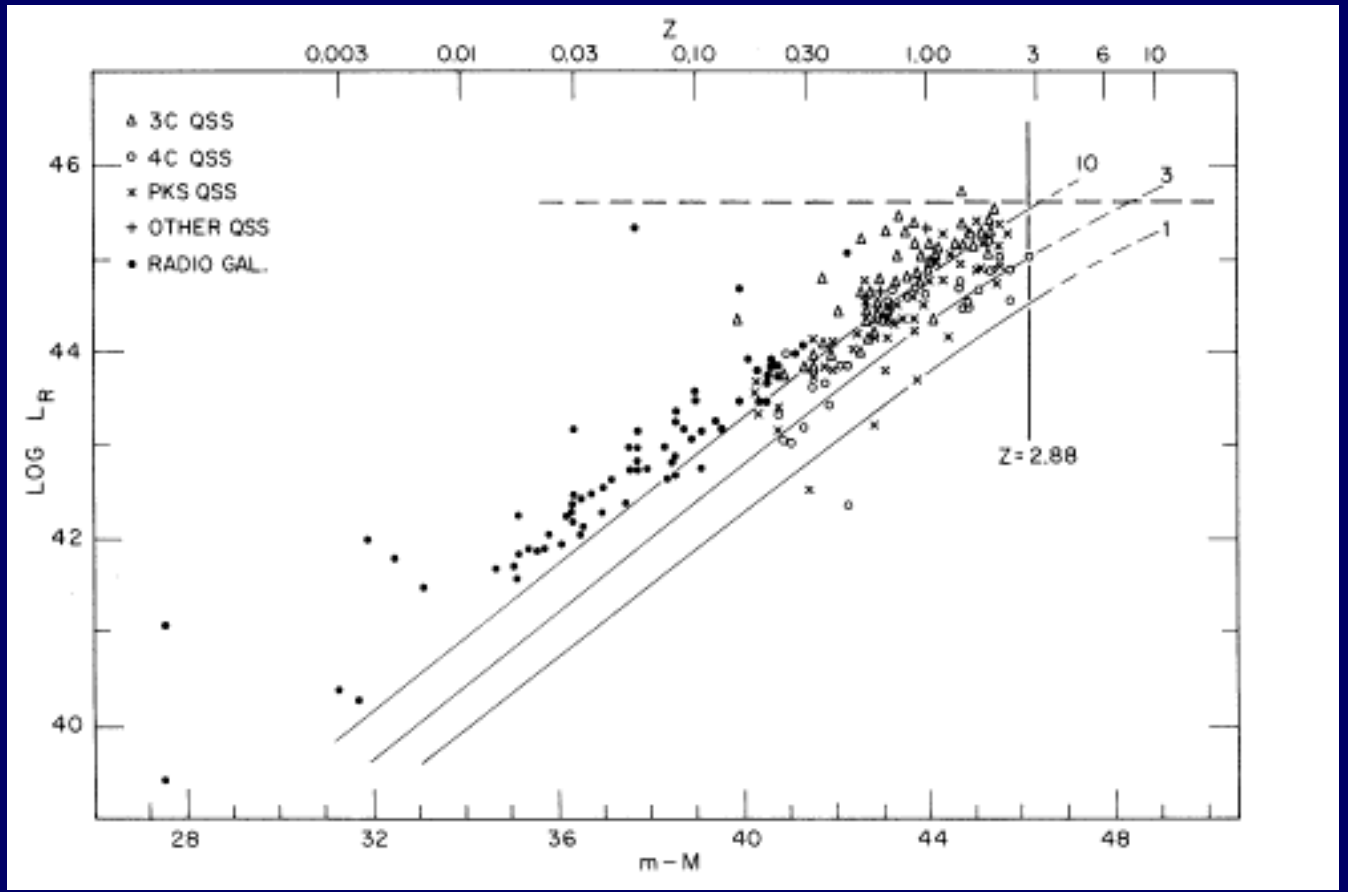
Stefano Andreon
Stefano.andreon@brera.inaf.it

- Simple on purpose. Emphasis is in allowing you to quickly re-use material for you own science problem.

- Stop me, ask, interact. I will show no sexy scientific results, my aim is to solve problems you will likely encounter. I put myself at your place, guessing what you may need. For this reason, my samples are small (how many of you have huge sample? Furthermore, there is no need of huge samples to learn), but methods applies to large sample too.

Resist from plotting x vs y blindly

- 1) Pay attention to selection effects. It seems there is a correlation below, however it is a selection effect (Sandage 1972, ApJ 178, 25) fig 7



- 2) suppose you have 10 columns and you plot each one against each other. There are about 40 plots. If no correlation is put in there (e.g. I put random numbers there), there will be 2 plots correlated at 95 % confidence, because the latter means “1 out 20”.

- Even if you have a billion rows each!

Why we regress x vs y ?

A) Prediction

- We want to use x as proxy of y, i.e. we have x, but not y, for some items, and we would like to have an idea of y. e.g. Mass proxies.
- As before, swapping x and y

B) Parameter estimation

We have both x and y, we are not interested in predict y (that we have) from x. We are interested in the relationship between them, how one goes vs the other one

C) Is there a trend?

We ask ourselves if there is a trend between x and y. Compared to case B), here we cannot suppose a trend is there: this is what we want to investigate!

D) Model selection

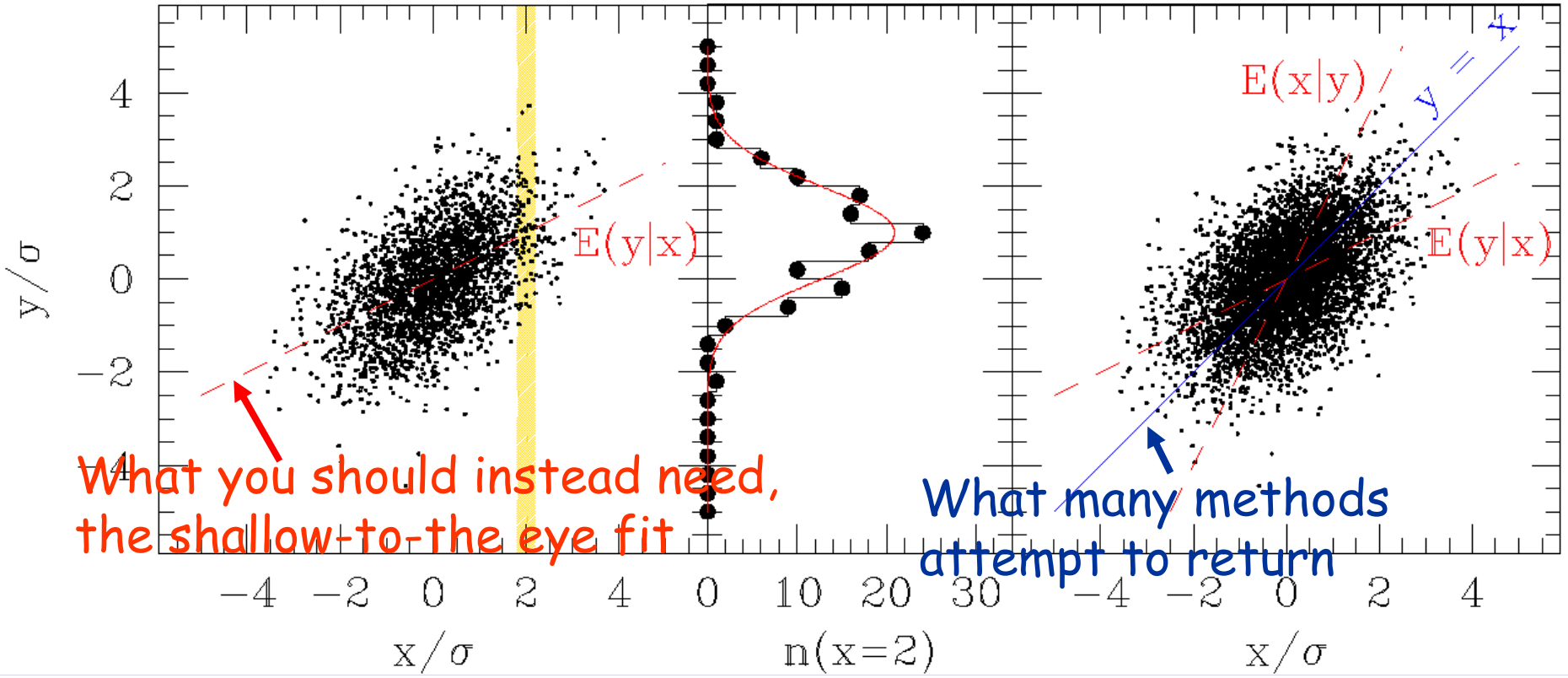
Are data better described by a given model (e.g y=x) or another one (e.g. y=arctan(x)?

These four things are conceptually different: do not expect you use the same tool to do these four different things. E.g. in case C you cannot assume a trend is there!

Why the (usual) best fit is wrong for prediction

... when there is an error structure on x values and one adopts usual methods. Usual methods return answers to a different problem.

Andreon & Hurn (2010, MNRAS 404, 1922)



Check it by yourself!

- Compute E(x|y) stepping y. Connect the resulting pairs.
- Compute E(y|x) stepping x. Connect the resulting pairs.
- Guess, by eye, the major axis of the data point cloud.
- Are the three lines equal?

A mess is there in some sub-fields of astronomy

- 1) the previous three things are mixed up
- 2) the most used tools return either the right answer to a different question, or a poor estimate of the requested number
- 3) most astronomers miss basic concept, e.g. the conceptual difference between a X|Y fit and a Y|X fit.

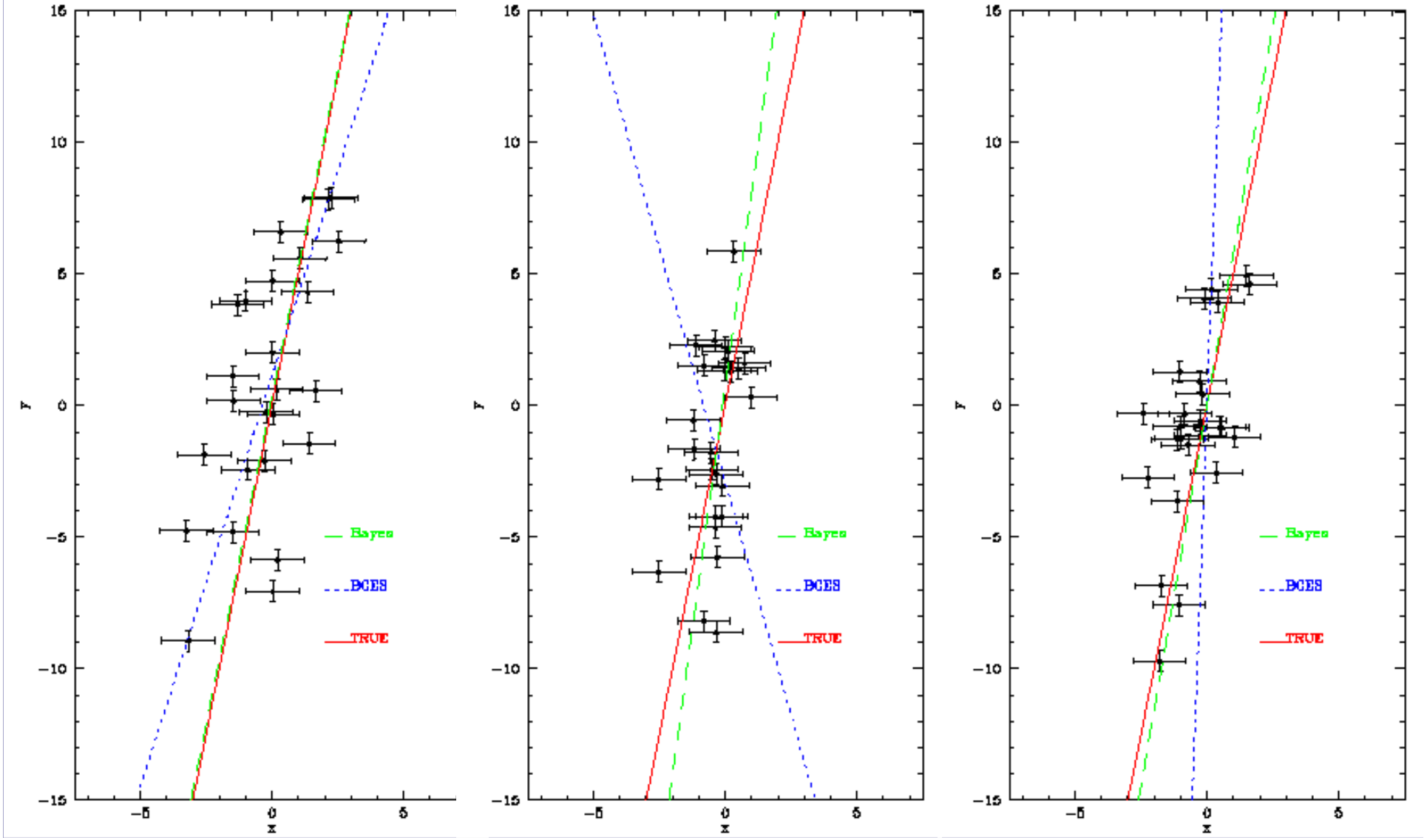
Regressions

Astronomers are interested in estimating as quantities vary as a function of each other: e.g. Tully-Fisher, Faber-Jackson, Magorrian relations, Fundamental-Plane, cluster scaling relations, GRBs (e.g. Amati) relations, etc.

How to determine the parameter of these regressions? There is no consensus in the astronomical literature: direct-, inverse-, orthogonal-, Bivariate Correlated Error and Intrinsic Scatter- (BCES), Measurement Error and Intrinsic Scatter- (MEIS), etc. -fit?

Performances

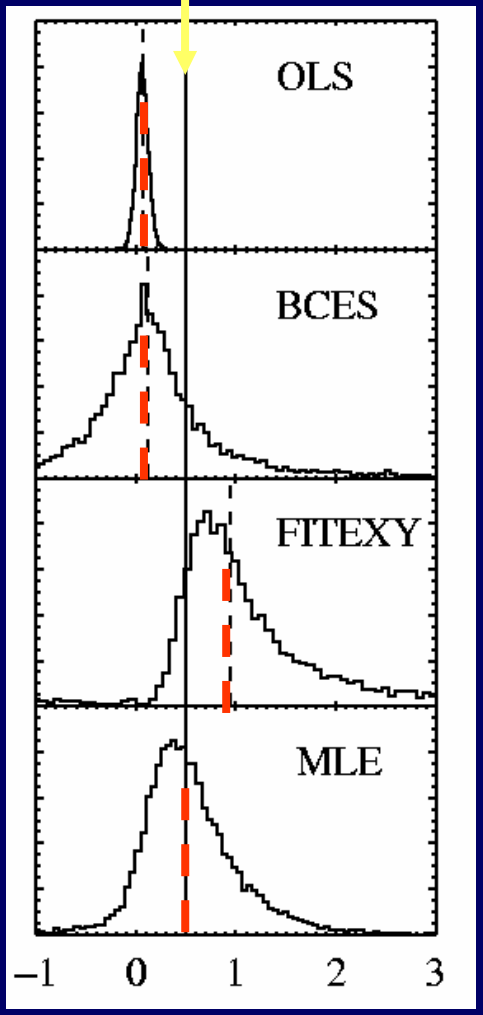
Andreon 2010, Bayesian Methods in Cosmology



Kelly et al 2007, ApJ 665, 1489

Why astronomer' techniques do not work

Input value



Estimated slope

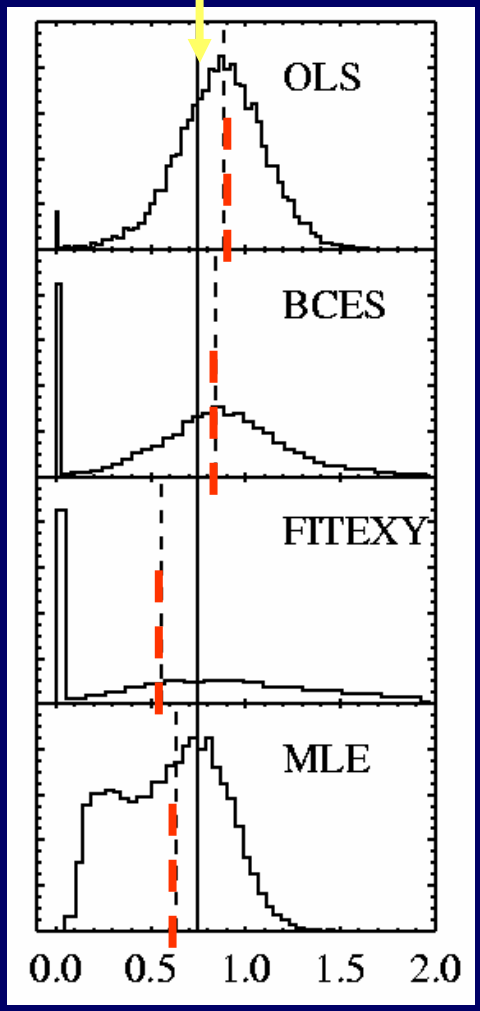
Ordinary Last Square fit

Bivariate Correlated Error and Intrinsic Scatter fit

Press et al. (numerical recipes)

Simplified bayesian solution

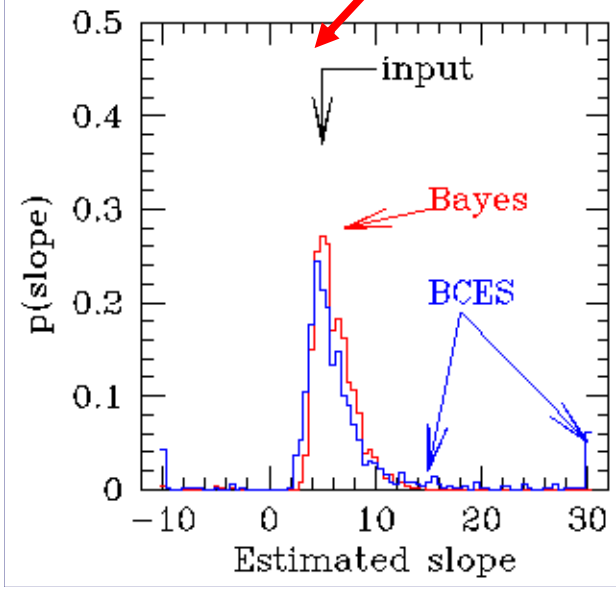
Input value

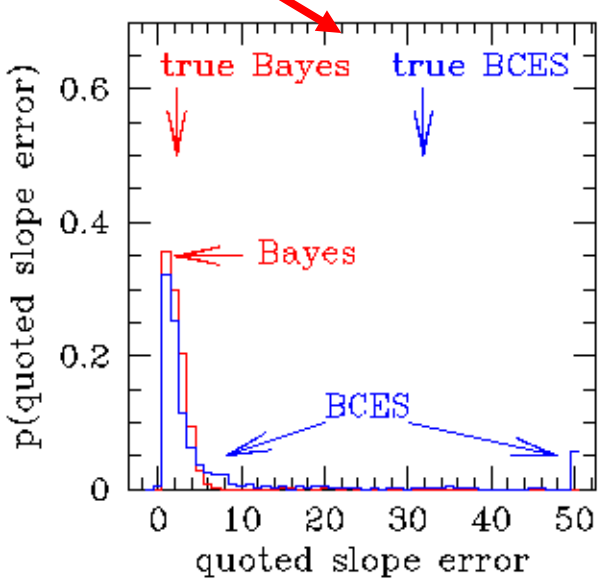


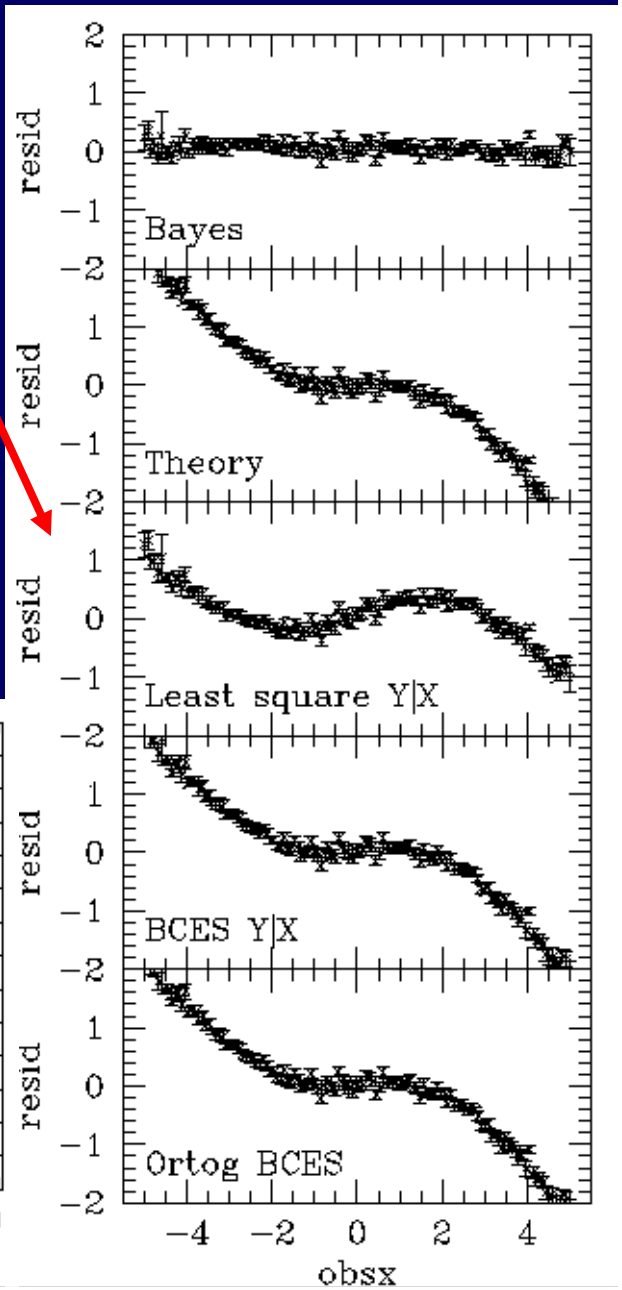
Estimated intrinsic dispersion

Performances

- **Prediction** (SA & Hurn, 2011, Statistical Analysis and Data Mining, invited review)
- **Recovering input parameters** (SA, 2010, Bayesian method for Cosmology)







The usual features of astronomical data

- Errors are heteroscedastic
- Data are not missing at random
- An intrinsic scatter is often there (non-heterogeneous population/systematcs)
- Relations can be non-linear
- Errors are often non-Gaussian
- Errors are sometime noisy
- (mixtures) Your sample include some unwanted object (or photon) but you cannot get rid of it (think to a weak signal over a background, a studies of quiescient early-type galaxies at z=2, etc.)
- (Prior) You known something about what you are studying (you don't observe it from your backyard telescope and with VLT, isn't?, and you will likely not ask time at a lambda were the source likely does not emit, isn't). You have priors on objects and parameters under study, and you may want include them in the analysis (say on H0)

The Bayesian way: you need to known four things only, that you already known!

a) Probabilities are in the 0 to 1 range

$$0 \leq p(E) \leq 1$$

E.g. the probability that tomorrow is sunny is 130 %, or -30%, make no sense (to me). Similarly, the probability that the Hubble constant, H_0, is between 50 and 100 is -12%, or 134 %, makes no sense

b) P=0 or 1

- $P(\Omega)=1$

If I throw a die, the probability of observing 1, 2, 3 … 6 is one. If all outcomes are in the considered set …

- $P(\Phi)=0$

If I throw a die, the probability of observing no face is zero

c) the sum rule/axiom

$$p(x) = \Sigma_y\, p(x,y) = \int p(x,y)dy$$

Table 2.1 Bag content

	ball	die	total
blue	4/22	5/22	9/22
red	10/22	3/22	13/22
total	14/22	8/22	22/22

Color probability distribution

Shape prob distrib.

Alias Margin-alisation

- Marginalisation is one of the only two things you need to remember. The other one is stored in JAGS. Nothing else is needed. Nothing else is allowed! Resist from introducing estimators, talking about optimal, best, … no freedom (good)

d) product rule/axiom

$$p(x,y) = p(x|y)*p(y)=p(y|x)*p(x)$$

ex: in a bag I have 4 blue balls and 10 red balls. If I extract two of them without replacement, what is the probability that both are red?

The probability of getting red the first ball, $p(x)$, is =10/14

The probability of getting red the second ball, after having get a red ball in the first extraction, $p(y|x)$, is = 9/13

The probability of getting red both, $p(x,y)$, is the product

10/14* 9/13= $p(x) * p(y|x)$

If now you change the order …

e) Bayes theorem

$$p(\theta|data) = c * p(data|\theta) * p(\theta)$$

Posterior = c* Likelihood* prior

Can be derived from the product rule, or, in alternative, assumed as axiom, and the product rule derived.

Central tool for parameter estimation

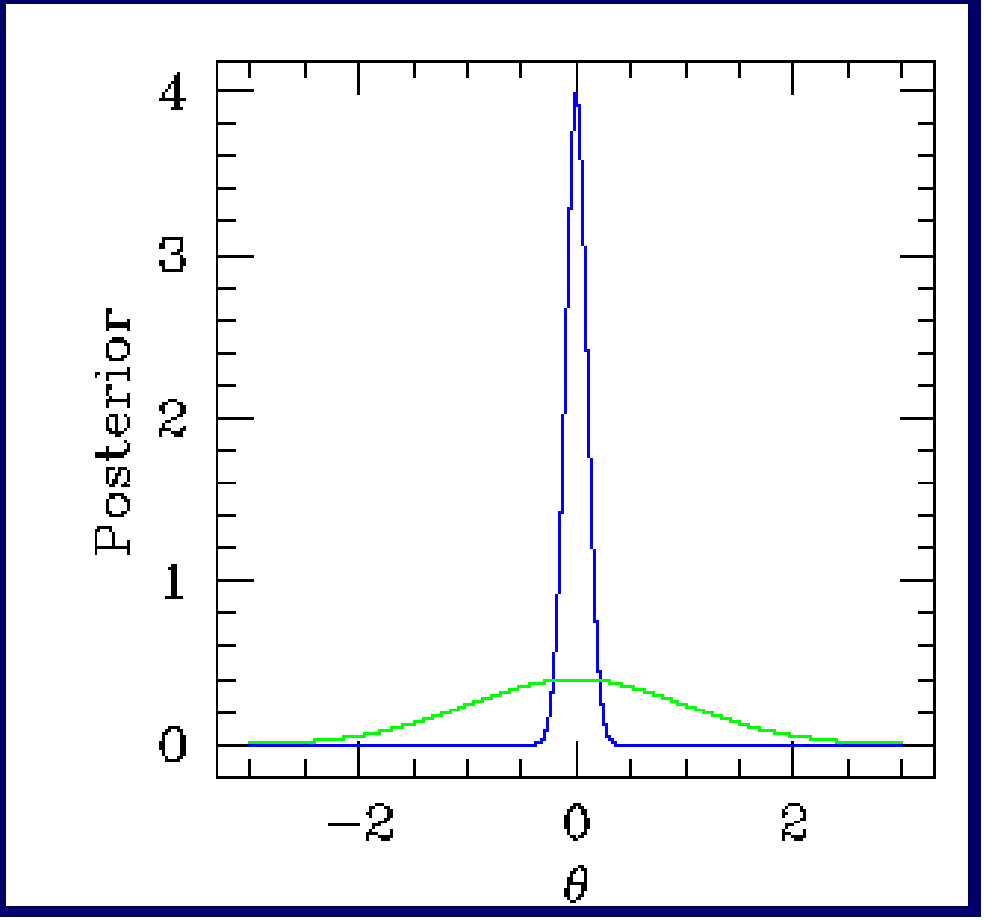
Second thing to remember (for a while only, later your computer remember it for you)

The posterior width quantifies the uncertainty

$$p(\theta|data) = c * p(data|\theta) * p(\theta)$$

if $p(\theta|data)$ is a narrow (almost delta) function, θ is very well determined
if $p(\theta|data)$ is a flat function, θ is badly determined

Do you want to known the uncertainty? Compute the posterior, and its width! This is the mantra of most applications: spell a prior, compute the likelihood, multiply them, and compute the width of the result.



- Everything comes from these axioms, no other ingredients used, no 'in the long run', no 'far from the boundaries', etc.
- Everything else needed, everything else allowed (advantage!)

posterior = prior* likelihood

Is an easy formula, don't need a computer for it, isn't ?

But:

a) Don't waste your time! writing the likelihood in closed form is often difficult (always impossible in research activities, unless you are Laplace), and always compsuming researcher precious time. This operation can be delegated to a software.

b) Don't waste CPU Time! Computing the posterior in a multidimensional space with a stupid strategy, such as stepping (and also with a smart one, a Monte Carlo) is time-compsuming and will not well explore the parameter space, unless the problem being solved is very simple.

Several Bayesian platforms, as JAGS, are able to write the likelihood in your place (well, they make something different, but this is just numerics) and furthermore have built inside smarter than an MC sampler (MCMC is used, see Stoica talk). Don't care about numerics, you have the posterior probability distribution in form of sampling: many values where the posterior is high, dew where the posterior is low.

Ready to deal with regressions (and whatever problem!), we only need to indicate the *logical* link (dependency and math expression) between quantities. Full stop.

In detail …

1) Write the mathematical model that describe how the data are generated: e.g.

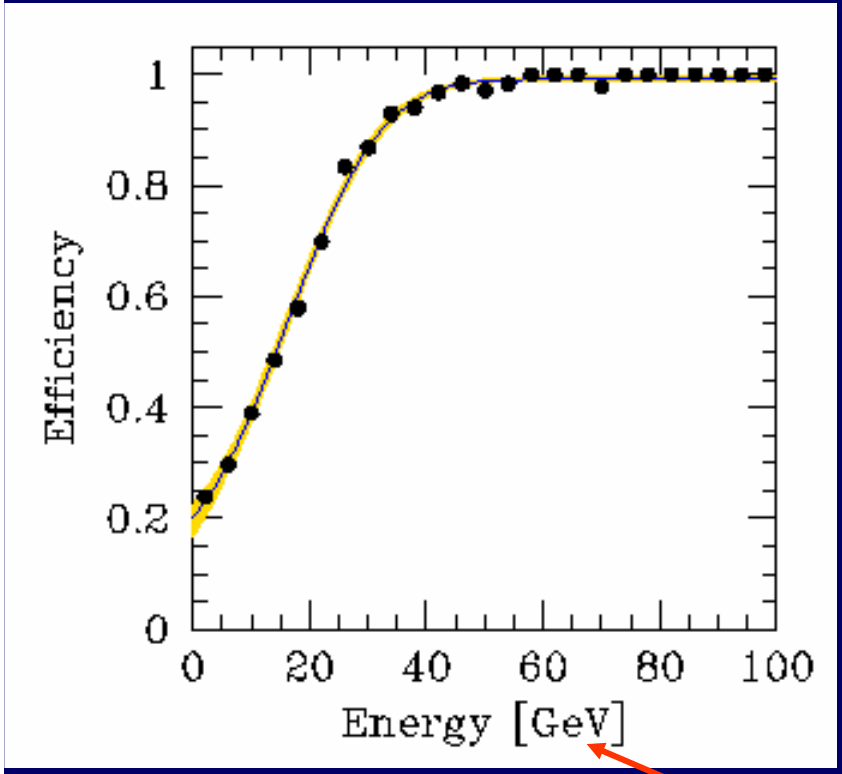
obsntot[i]~dpois(nbkg[i]/C[i]+nclus[i]). Sketching a graph of the logical link between quantities (bayesian graph) is useful.

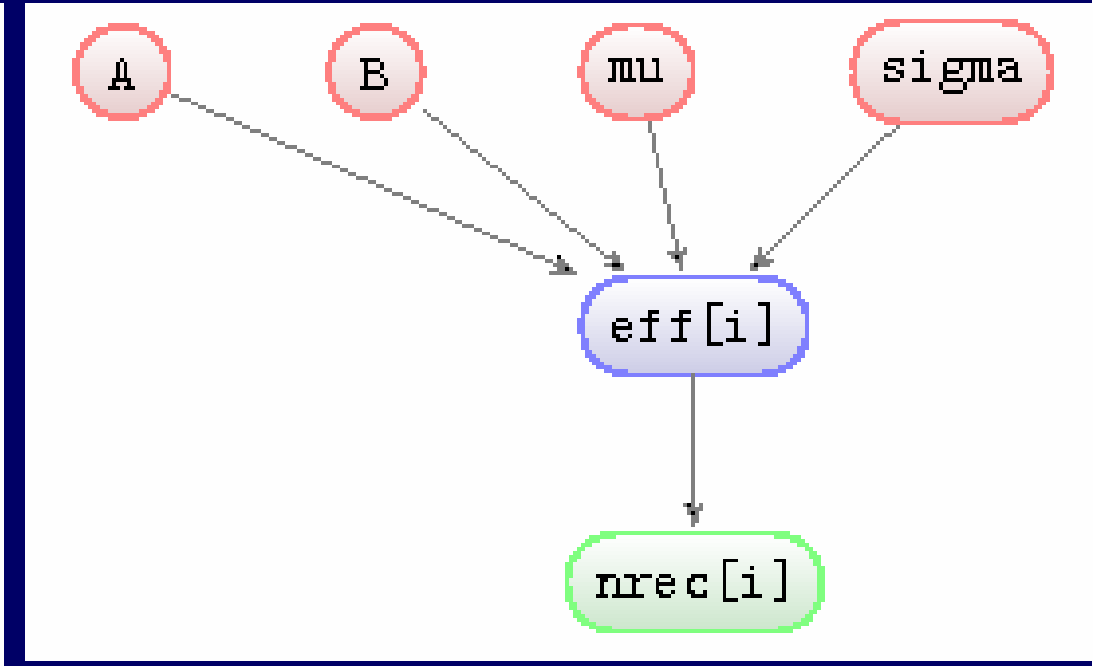
2) Write in mathematical words what you already known about objects/quantities of your interest (prior, e.g.

Sigma_v ~dunif(200,2000)

3) Let the computer to compute (why otherwise it is called computer?) and deal with the numirical-related problems.

Non-linear regressions exist!





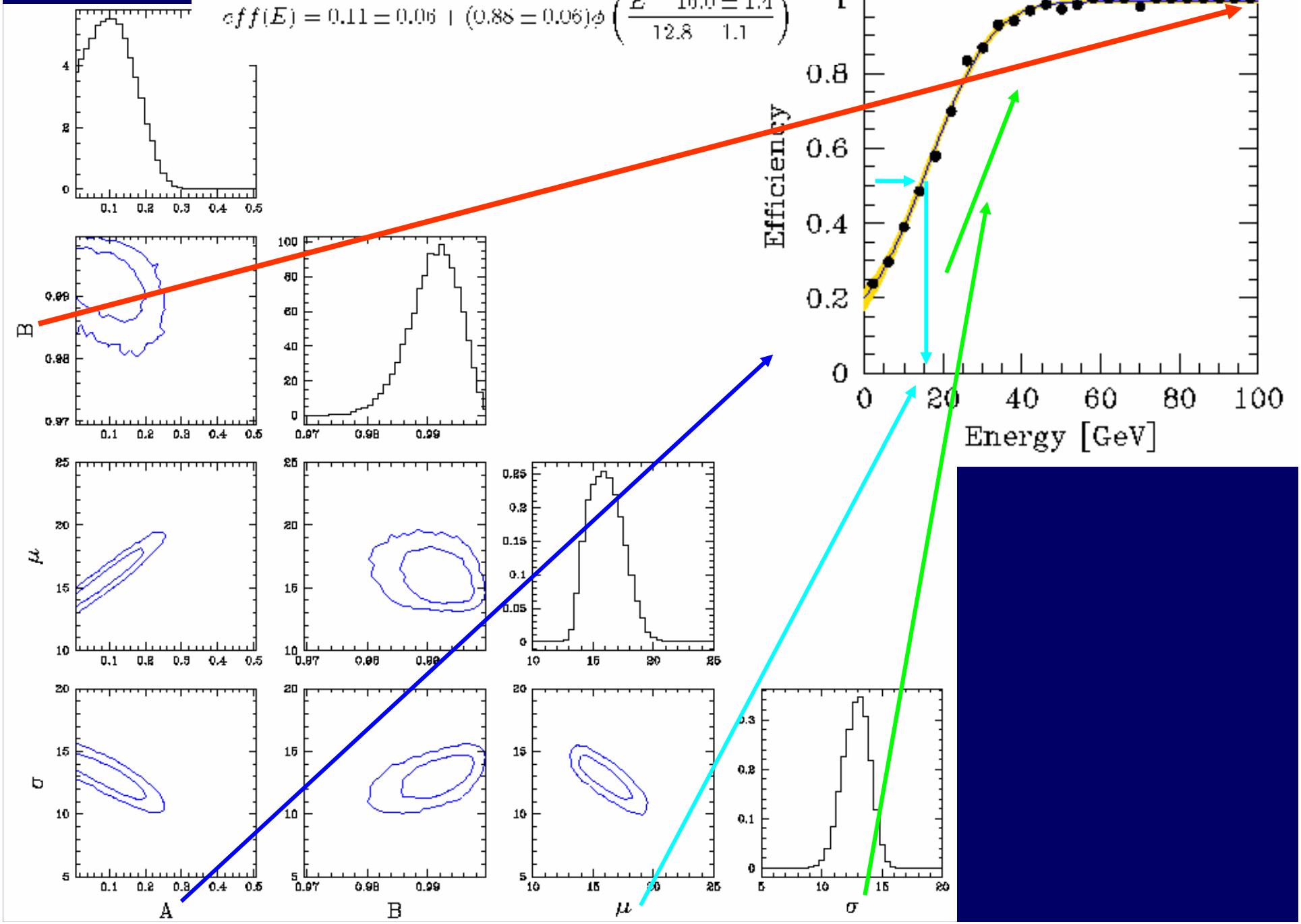
Energy dependency of your instrument, completeness as a function of flux, whatever else!

Etheroscedastic errors, here and everything else.

The model (running code)

```
model {
for (i in 1:length(nrec)) {            # foreach datum
nrec[i] ~ dbin(eff[i],ninj[i])         # binomial likelihood
 eff[i] <- A + (B-A)*phi((E[i]-mu)/sigma)  #math express of
                                           # fitted function
}
A~dunif(0,1)                           # weak priors
B~dunif(0,1)                           # taken so, not a suggestion
mu~dunif(0,100)                        # to always use them!
sigma~dunif(0,100)
}
```

$$eff(E) = 0.11 \pm 0.06 \; + \; (0.88 \pm 0.06)\phi\left(\frac{E \quad 16.0 \pm 1.4}{12.8 \quad 1.1}\right)$$



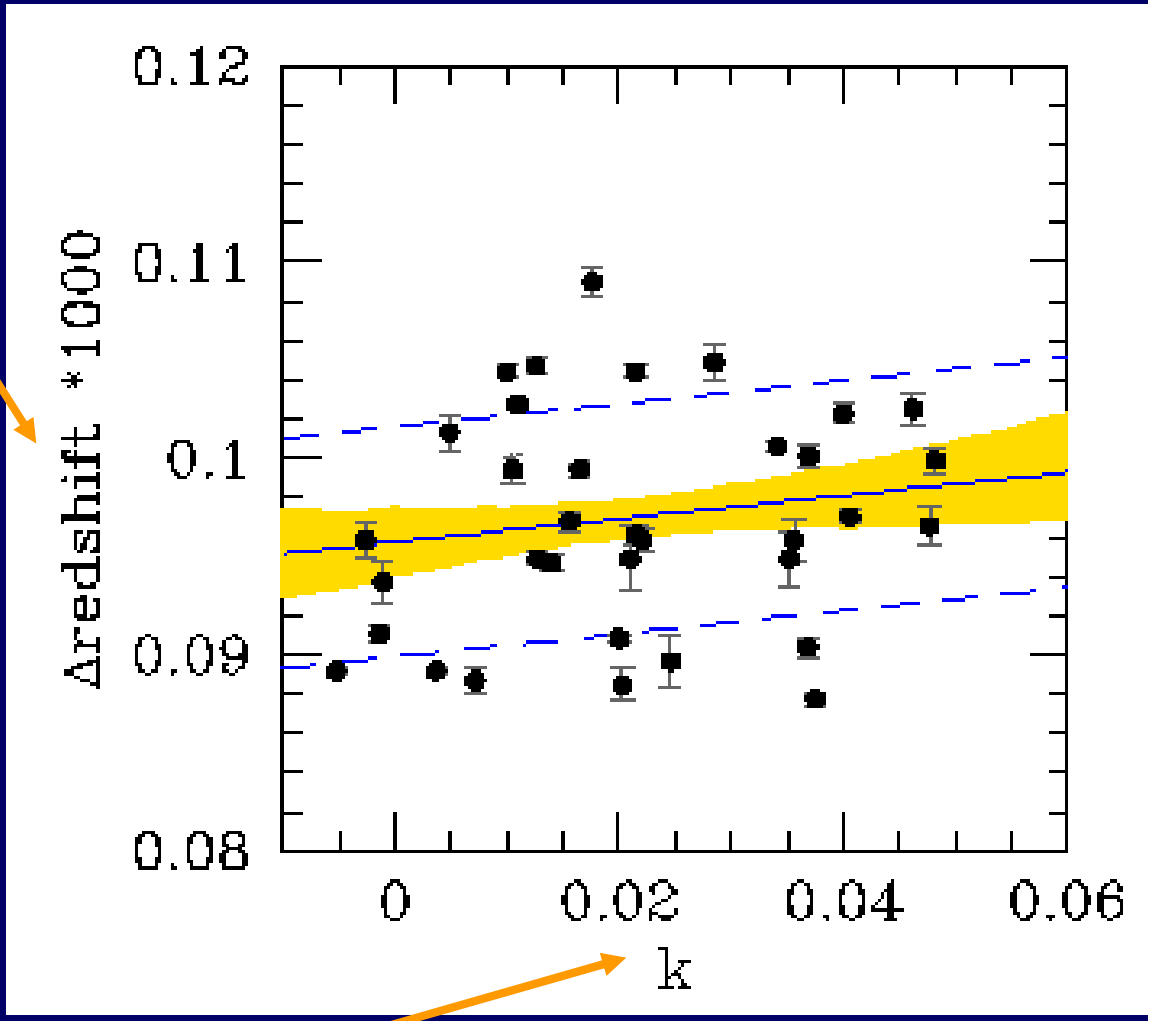
- If some bins (values at some energies) are missing, no problem (of course, larger errors!).
- True from now on, just write NA where you don't have the value if missing at random, otherwise wait some slides...

Modeling intrinsic scatter

$1000*(\lambda_{lab} / \lambda_{QSO} - cost)$

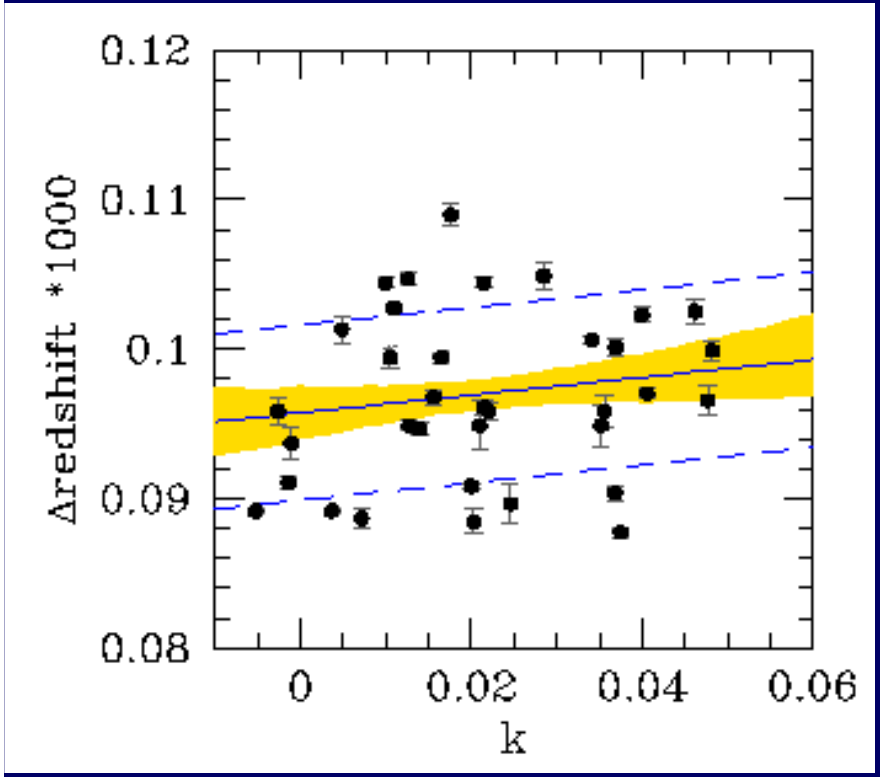
cost is ~$(1 + z_{QSO})$

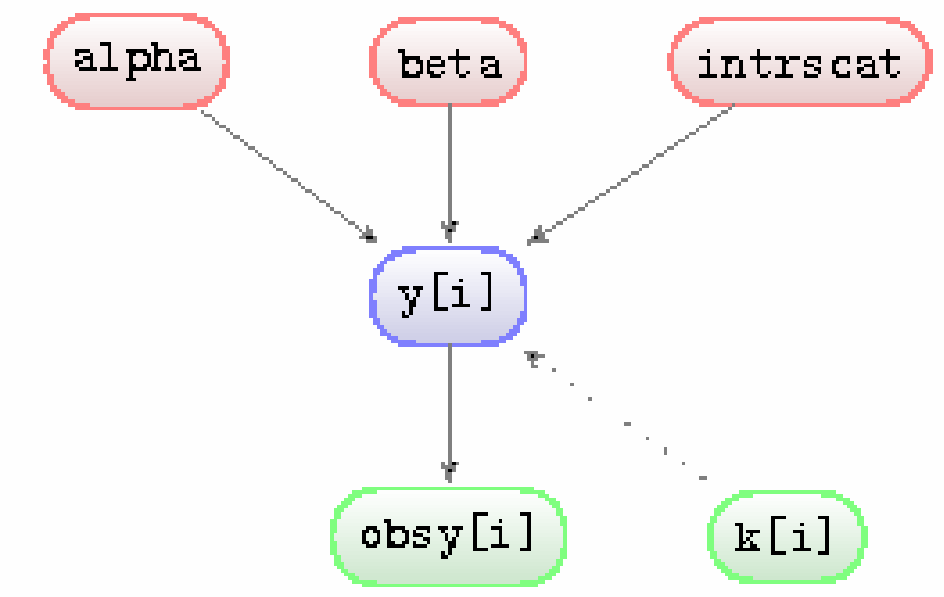
If physics at z=zQSO is the same as in our lab, the y axis should be just the QSO redshift, with no trend with k



Sensitivity of a given line ratio on to the value of the constant of fine structure (TBC). Data curtesy of Wendt & Molaro (2011)

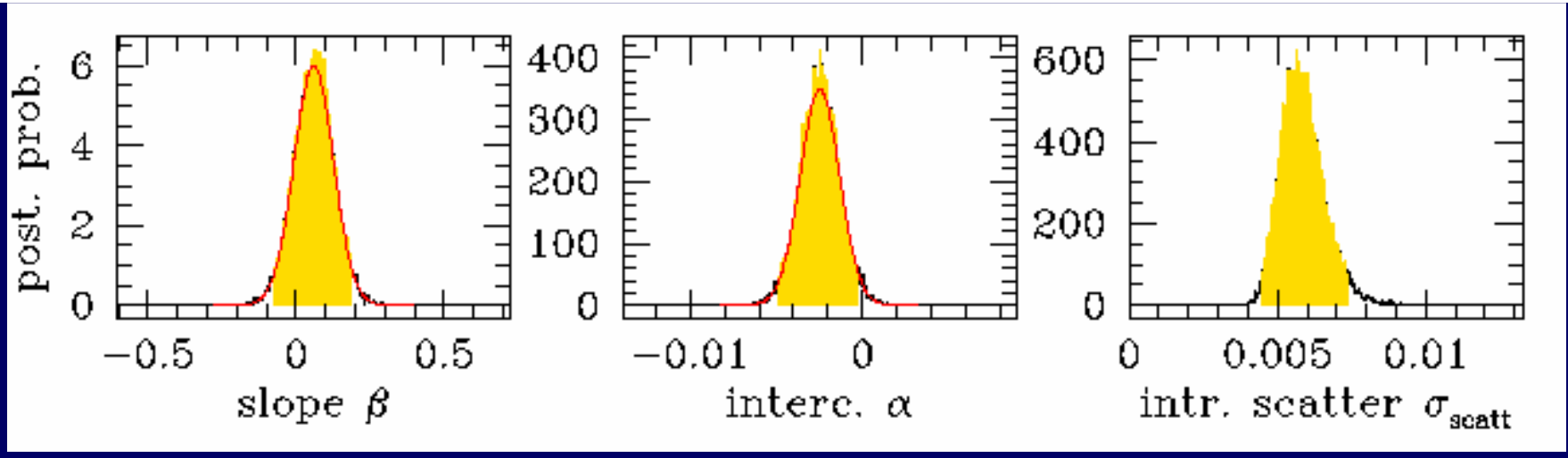
Let suppose a trend is there. What is the slope?





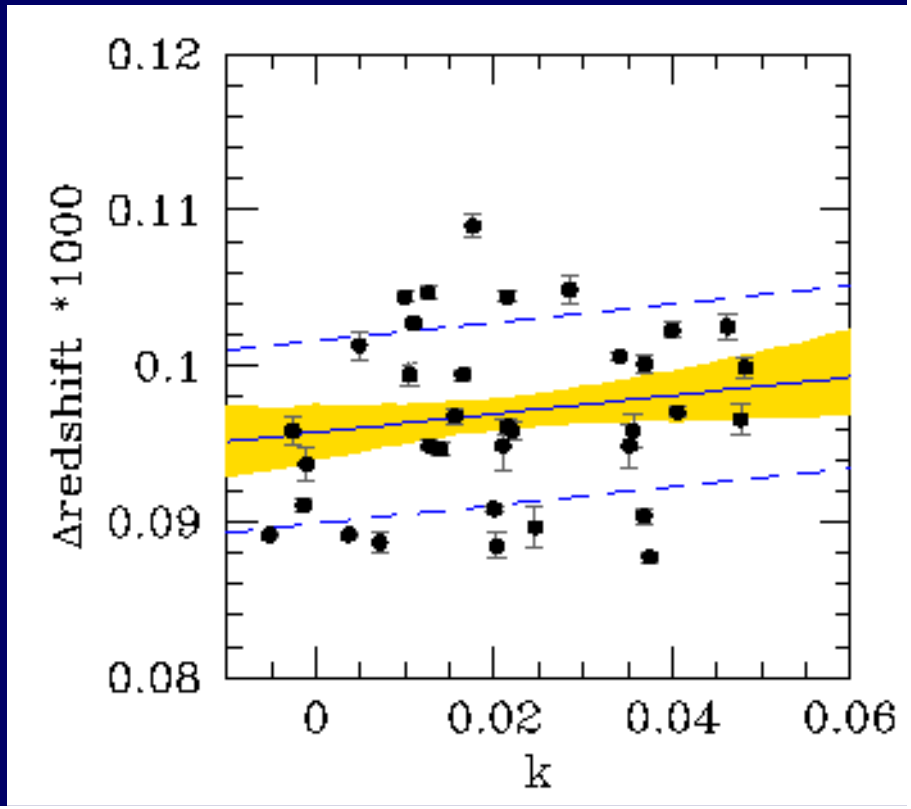
The model

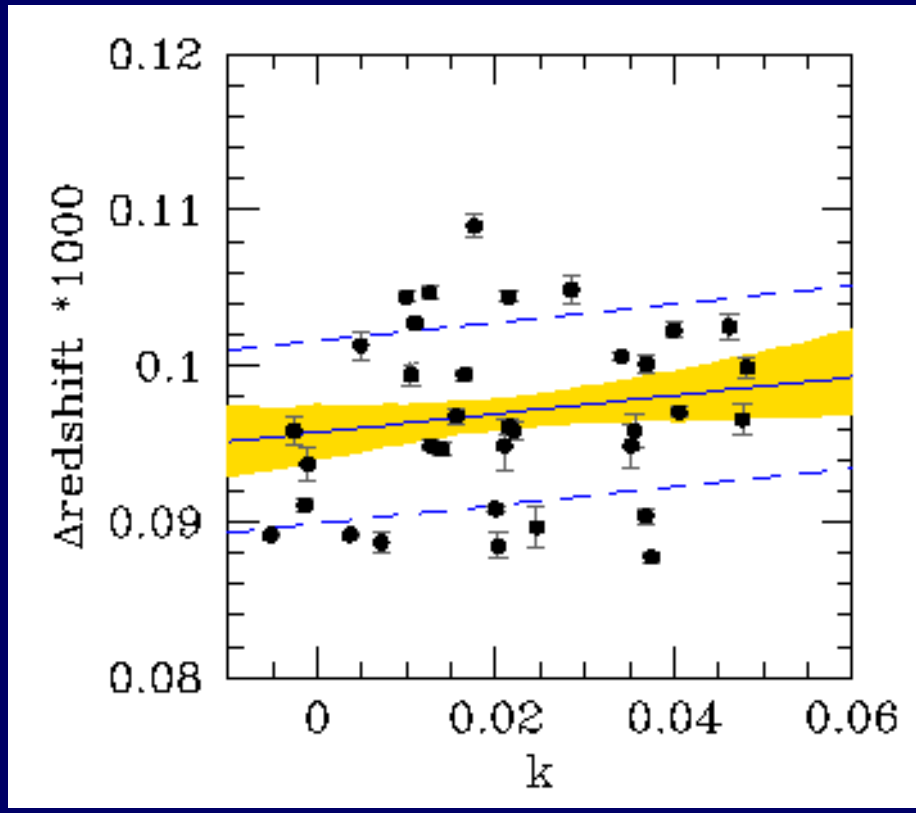
```
model {
intrscat ~ dunif(0,3)                #weak priors
alpha ~ dnorm(0.0,1.0E-4)
beta ~ dt(0,1,1)
for (i in 1:length(x)) {             #foreach data point
# modeling ordinate
obsy[i] ~ dnorm(y[i],pow(err.y[i], -2))  #gauss errors
y[i] ~ dnorm(z[i],pow(intrscat, -2))     # gauss intrinsic scatter
# modeling ordinate vs x
z[i] <- alpha+0.1+beta*(k[i]-0.03)     # linear y vs x
}
```



The found slope is half a sigma away from zero, and the total variation of y over the x range is comparable to the y uncertainty of the mean model. Not convincing evidence for a trend, but a (small) one still possible.

What about if you miss the intrinsic scatter, as in old analysis?

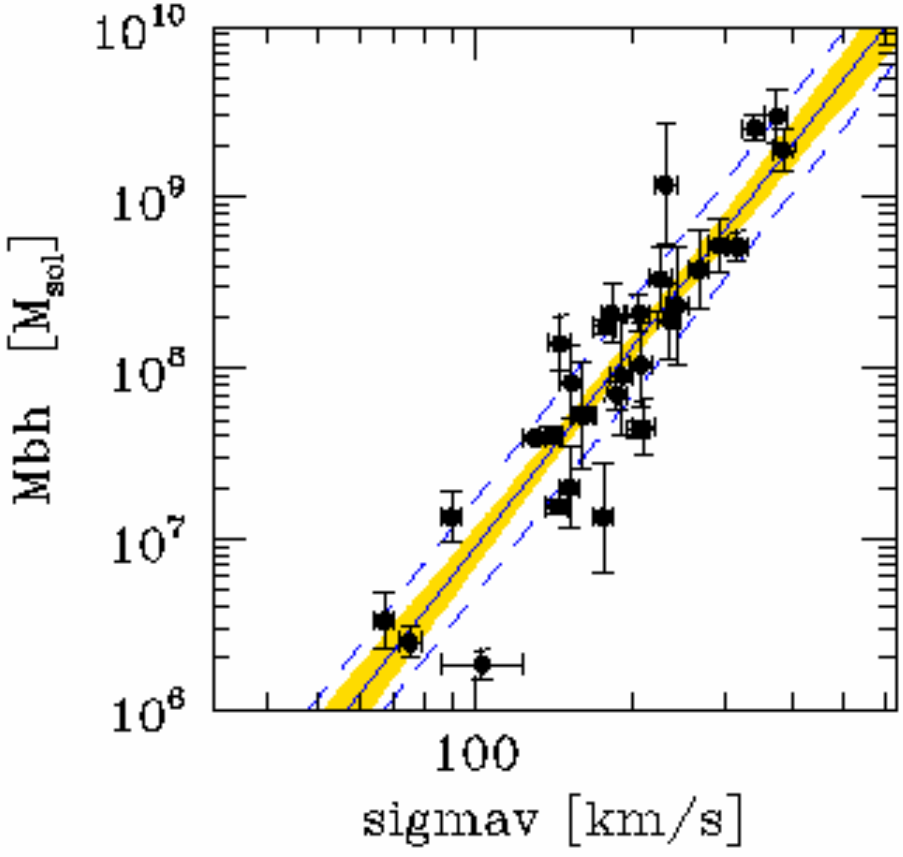


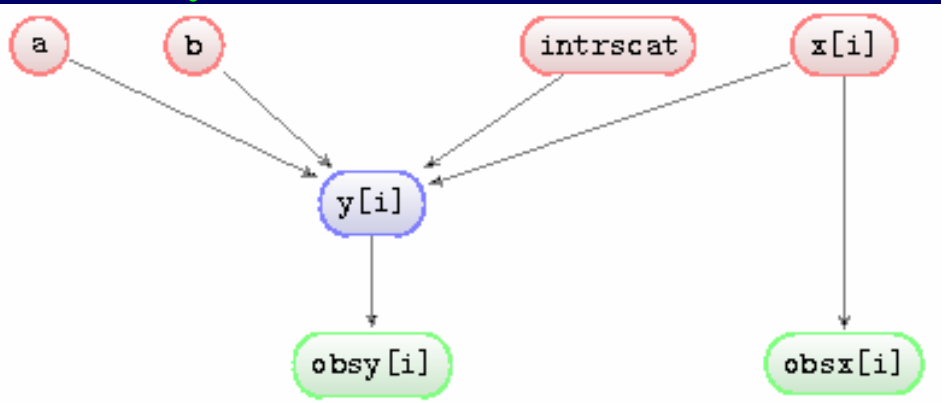


$$\Delta\lambda/\lambda = (3.0257975 \pm 1.1)\ 10^{-6} + (0.59 \pm 0.66)\ 10^{-4}(k - 0.03)$$

P.S. in this example intrinsic scatter is a synonim of sytematics, all measurements pertain to a single object (it is not a population spread)

The astronomer nightmare starts: errors on predictor



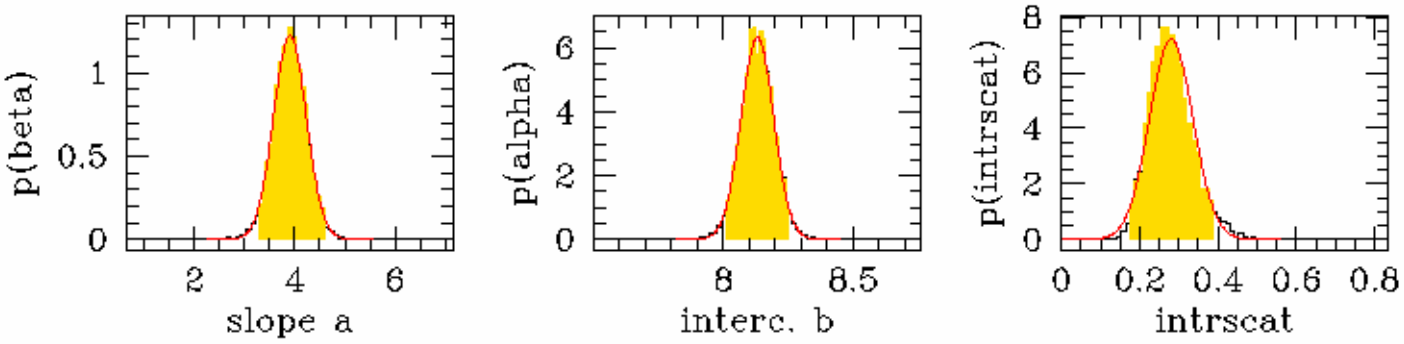


Magorrian relation, data from Tremaine et al. (2002).

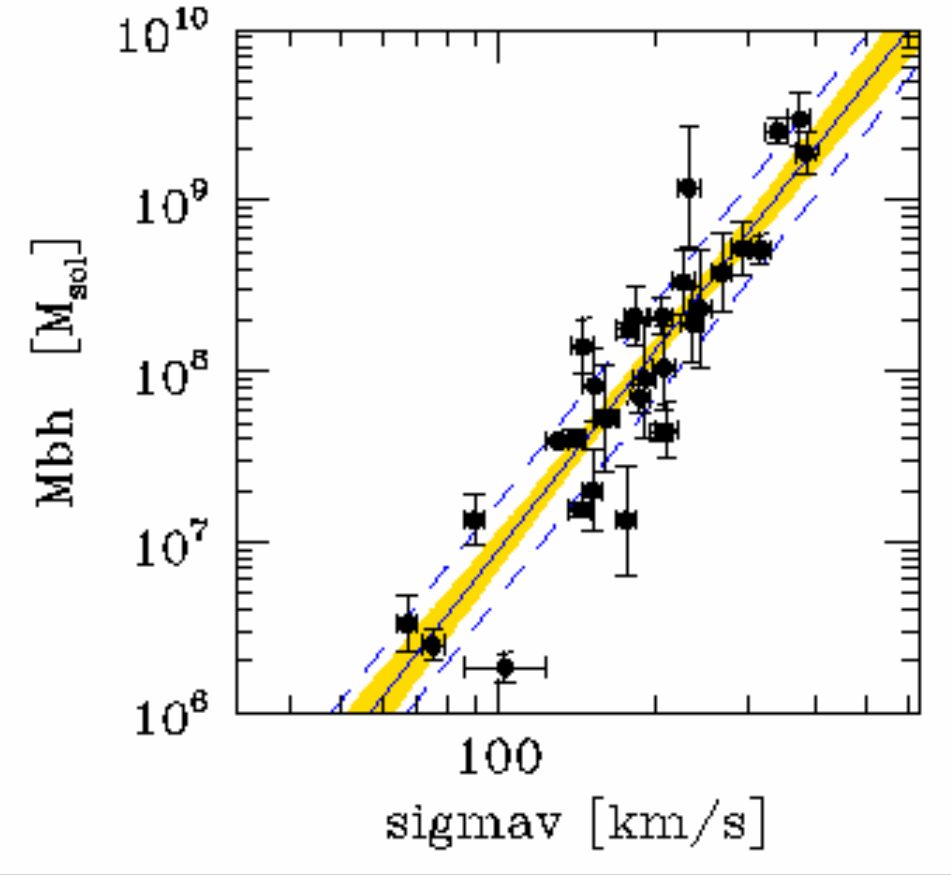
Nightmare because from here on a bias arizes if you slopply ignore what you should account for!

The model

```
model {
for (i in 1:length(obsx)) {
x[i] ~ dunif(-1.0E+4,1.0E+4)                # priors on x's
obsx[i] ~ dnorm(x[i],pow(errx[i],-2))       # Gauss err on x
y[i] ~dnorm(b+a*(x[i]-2.3), prec.scat)      # Gauss scatter
obsy[i] ~ dnorm(y[i],pow(erry[i],-2))       # Gauss errors
}
prec.scat ~ dgamma(1.0E-2,1.0E-2)           # weak priors
intrscat <- 1/sqrt(prec.scat)
b ~ dnorm(0.0,1.0E-4)
a ~ dt(0,1,1)
}
```

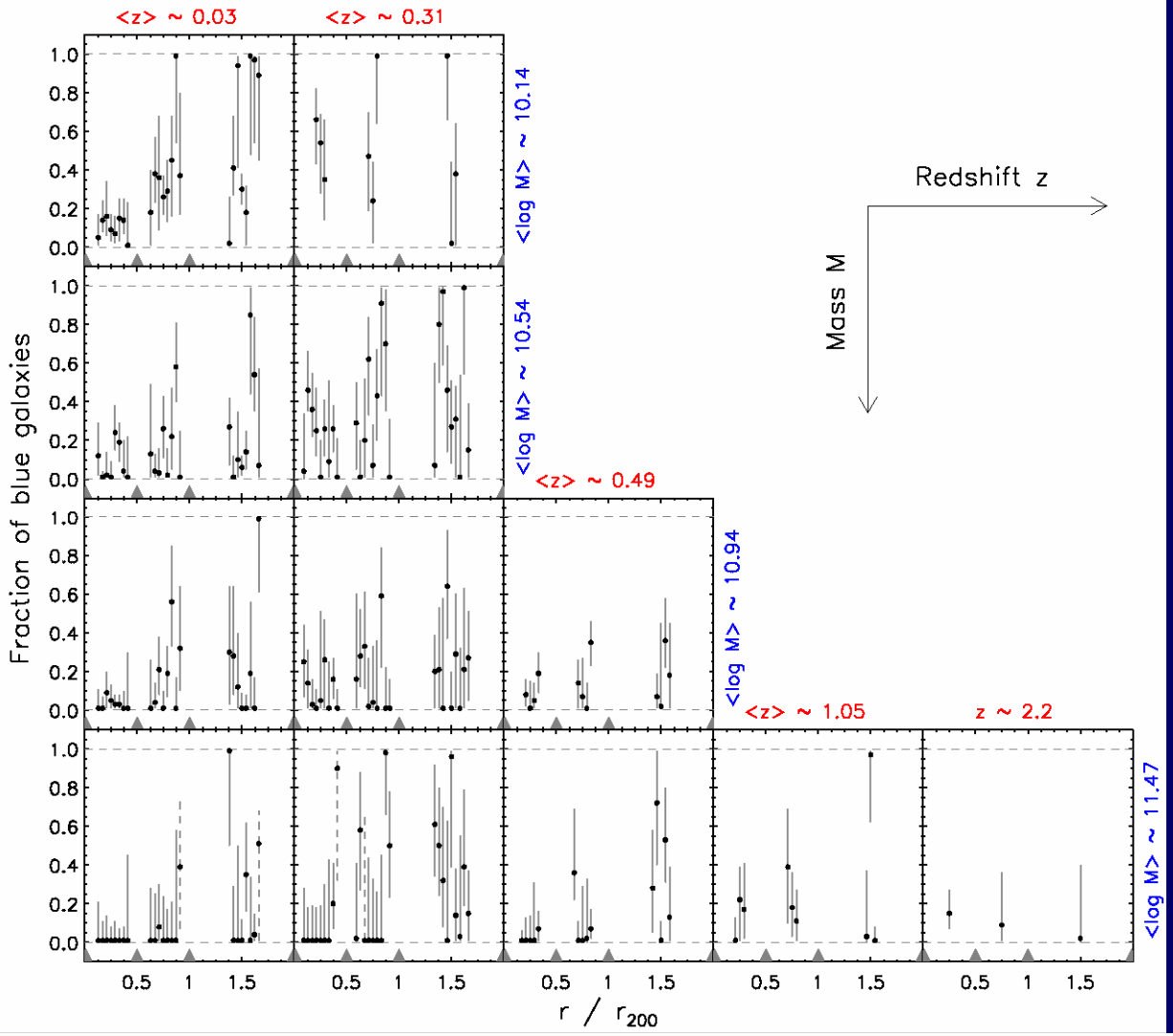


$$\log M_{bh} = (3.9 \pm 0.3)\ (\log \sigma_v - 2.2) + 8.13 \pm 0.06,$$



Intrinsic scatter is probably here a population spread.

Getting crazy: multiple non linear, regression, non-gauss & mixtures

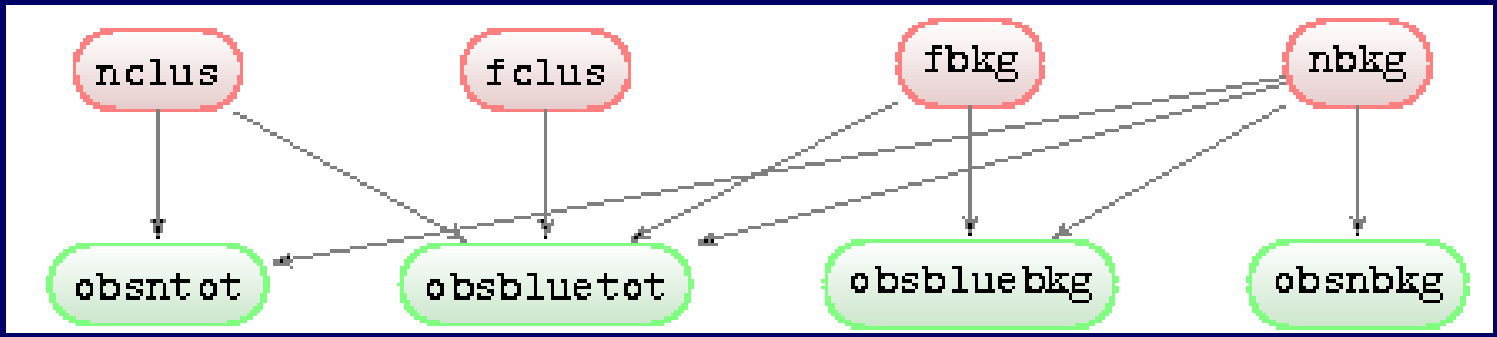


Data from Raichoor & Andreon (2012).

The fraction of blue galaxies depends on redshift, clustercentric distance and galaxy mass.

And this is not enough, a cross-term is needed.

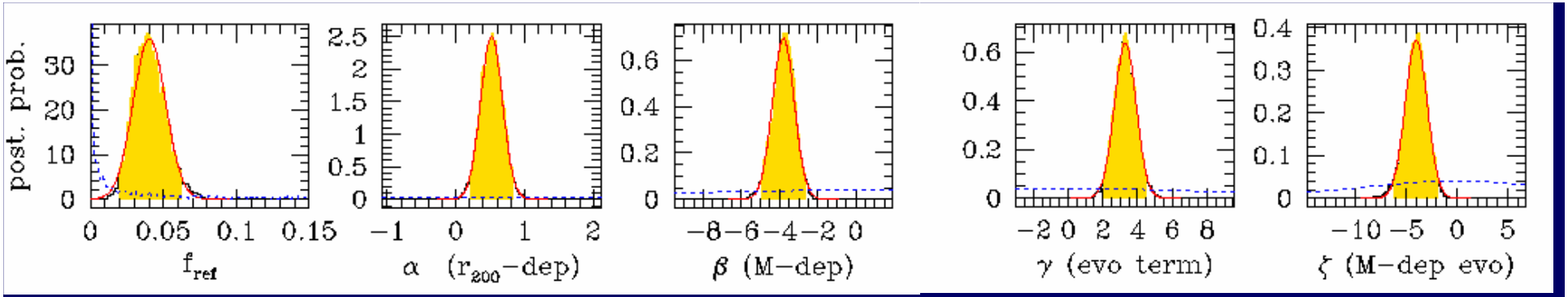
Unfortunately, when we measure we don't have individual membership of cluster galaxies!

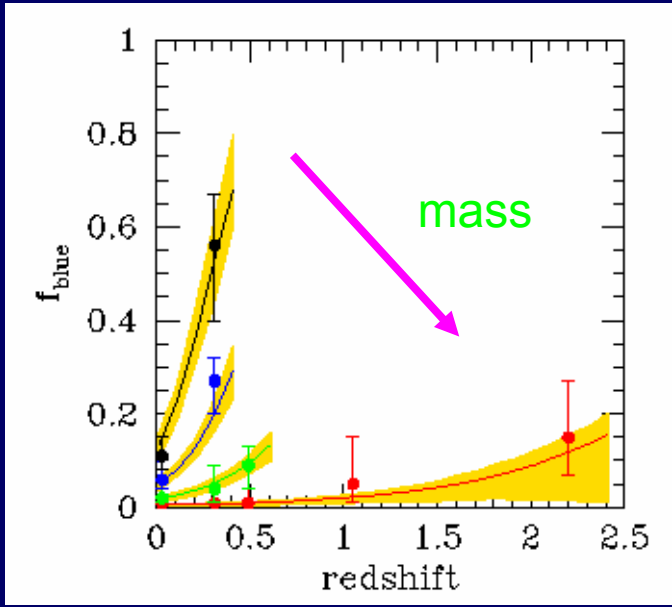


+ 3 dependencies (galaxy mass, redshift, clustercentric distance)

```
model {
for (i in 1:length(obsntot)){
obsnbkg[i]~dpois(nbkg[i])                          # Poiss bkg
obsnbluebkg[i]~dbin(fbkg[i],obsnbkg[i])  # Binom f_blue bkg
obsntot[i]~dpois(nbkg[i]/C[i]+nclus[i])     # Poiss bkg+clus
obsnbluetot[i]~dbin(f[i],obsntot[i])         # Binom f_blue tot
f[i] <- (fbkg[i]*nbkg[i]/C[i]+fclus[i]*nclus[i])/(nbkg[i]/C[i]+nclus[i]) #algebra

                    # 5 param fitted function (4 param insuff)
fclus[i] <- ilogit(lgfclus0+alpha*log(r200[i]/0.25)+beta*(lgM[i]-11)
                +gamma*(z[i]-0.3)+zeta*(lgM[i]-11)*(z[i]-0.3))
nbkg[i] ~ dunif(1,1e+7)
fbkg[i] ~ dbeta(1,1)
nclus[i]~ dunif(1,1e+7)
}
lgfclus0 ~dnorm(0,0.01)  # priors
alpha ~ dnorm(0,0.01)
beta ~ dnorm(0,0.01)
gamma ~ dnorm(0,0.01)
zeta ~ dnorm(0,0.01)
}
```

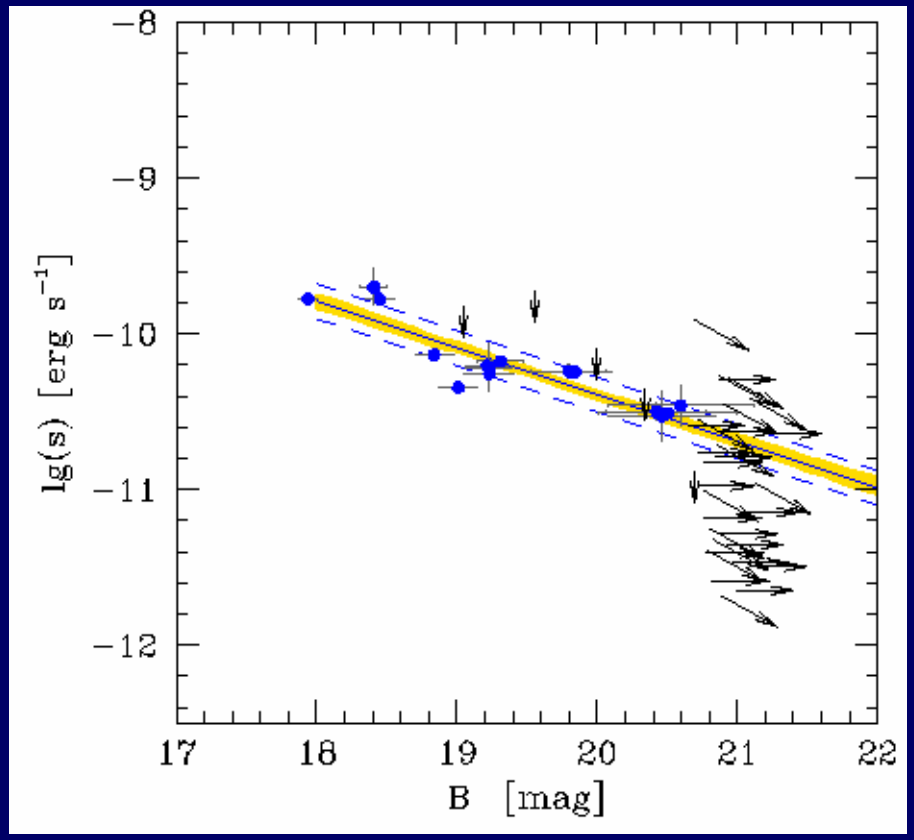




$$f_{blue}(r/r_{200}, M, z) = \texttt{ilogit}\Big[A_0 + \alpha \cdot \log(r/(0.25 \cdot r_{200})) + \beta \cdot (\log(M/M_\odot) - 11) + \gamma \cdot (z - 0.3) + \zeta \cdot (\log(M/M_\odot) - 11) \cdot (z - 0.3)\Big],$$

$$\texttt{ilogit}(x) = (1 + \exp(-x))^{-1}$$

Sometime we deal with upper/lower limits, isn't?

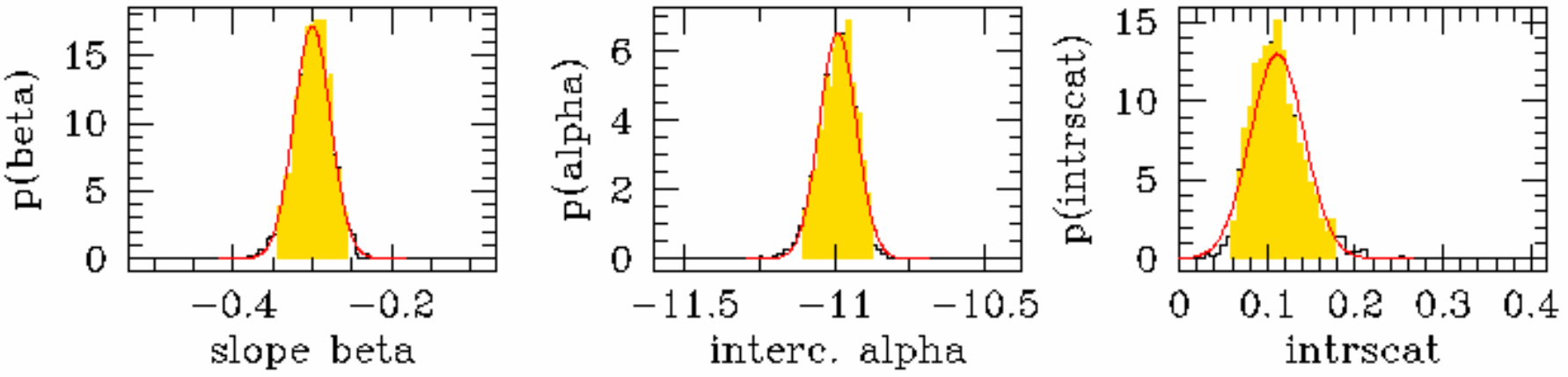


Optical to X-ray flux, fake data

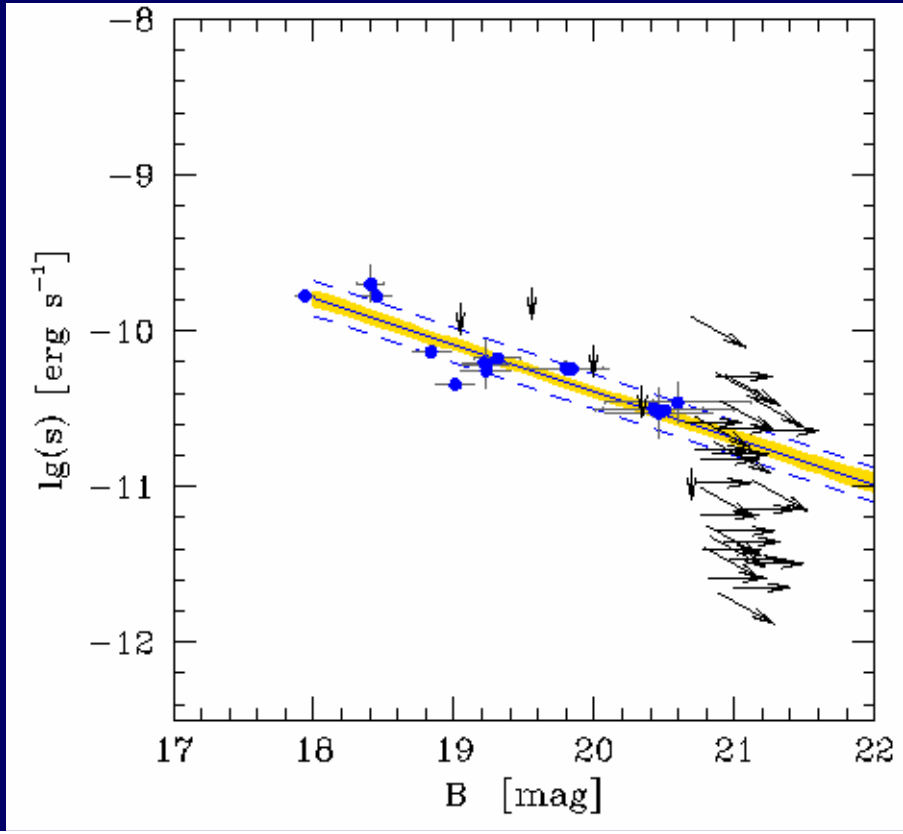
Upper limits (72 % of the total sample!) authomatically dealt with.

The model

```
model {
intrscat ~ dunif(0,10)                  # weak priors
alpha ~ dnorm(0.0,1.0E-4)
beta ~ dt(0,1,1)
zptB <-24                               # astronomers idosincrasies
for (i in 1:length(obstotB)){           # for each datum
 obstotB[i] ~ dpois(sB[i]+bkgB[i]/CB[i]) # Poiss fluct tot B band
 obsbkgB[i] ~ dpois(bkgB[i])             # Poiss fluct bkg B banc
 obstotS[i] ~ dpois(sS[i]+bkgS[i]/CS[i]) # Poiss fluct tot X band
 obsbkgS[i] ~ dpois(bkgS[i])             # Poiss fluct bkg X band
 magB[i] ~ dunif(18,25)                  # weak prior
 sB[i] <- pow(10,(zptB-magB[i])/2.5)     # mag definition
                                         # linear relation with intr scatt
 lgfluxS[i] ~ dnorm((magB[i]-22)*beta+alpha,pow(intrscat,-2))
 sS[i] <- pow(10,lgfluxS[i]-zptS[i])
 bkgB[i] ~ dunif(0,1e+7)                 # weak prior
 bkgS[i] ~ dunif(0,1e+7)                 # weak prior
}}
```



Input was: slope=-0.3, intecept=-11, intr scatter=0.10



With important data structure

One step back, the Eddington-Malquist-Bayes correction …

Let consider one source, 4 (X-ray) detected photons (per unit time). What is its rate? 4? No, there are more faint stuff scattered up in flux than bright stuff scattered down! (usual case of Kenter et al 2005, ApJS 161, 9, x-ray survey with sources as faint as 2-4 photons).

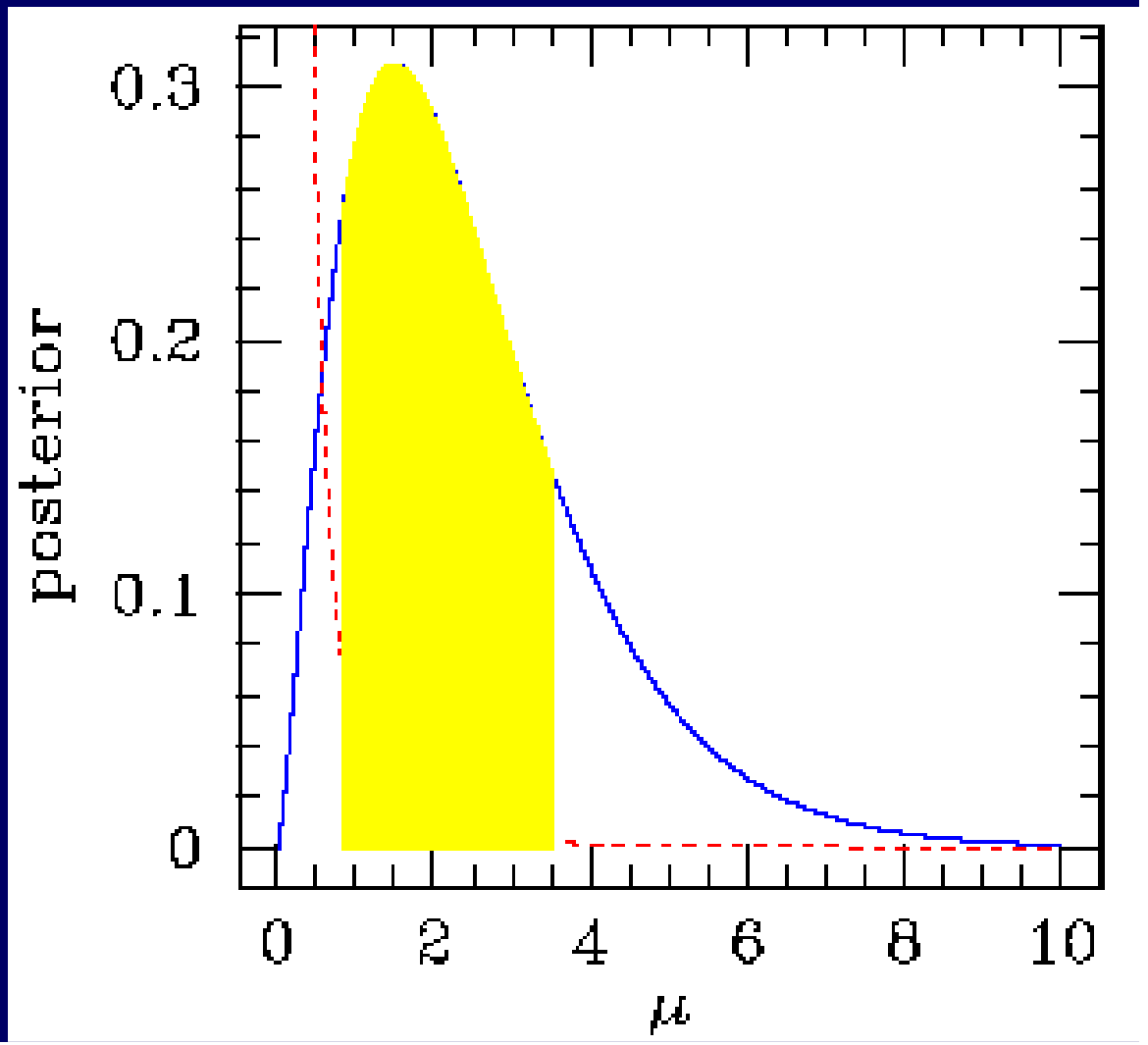
$p(\theta|data) = c * p(data|\theta) * p(\theta)$

$p(\mu|4) = c^* \, p(4|\,\mu)\, p(\mu)$

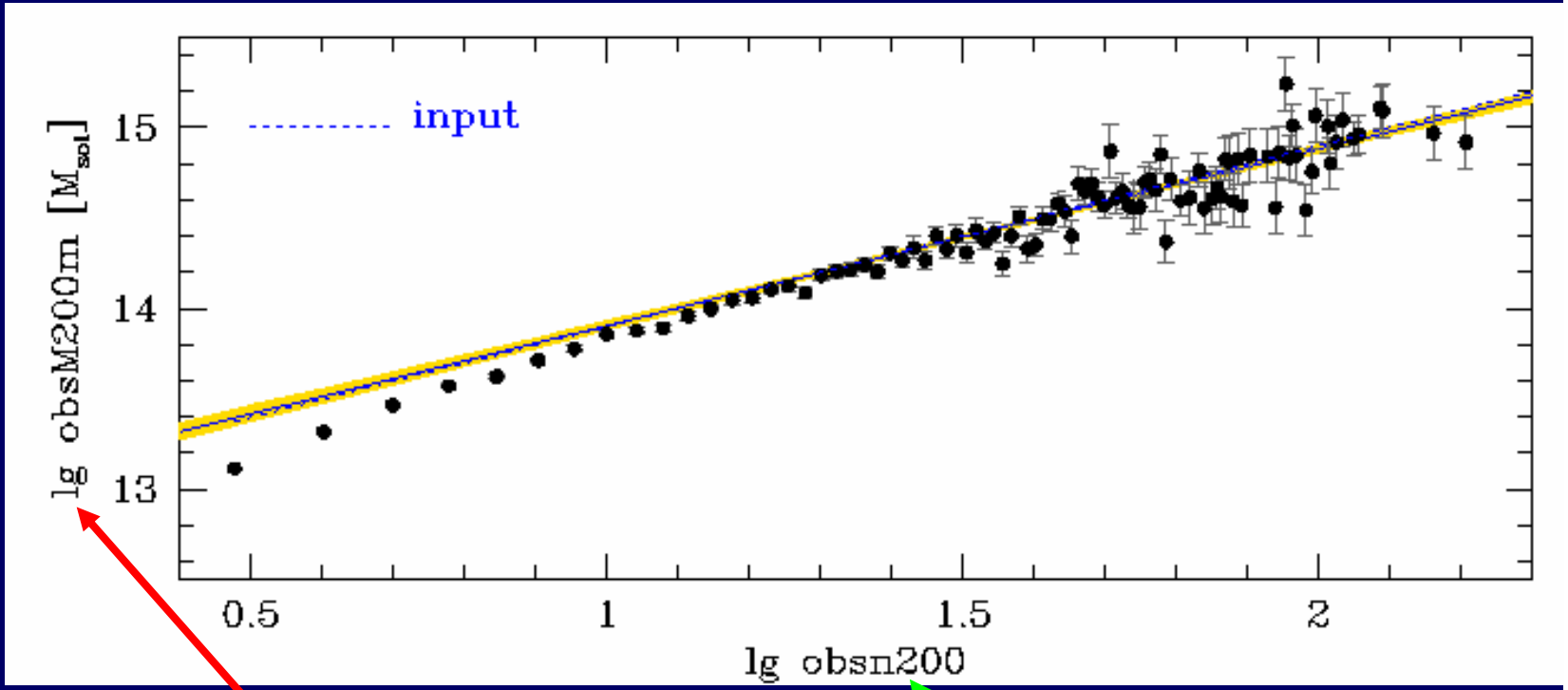
at the studied fluxes, the prior $p(\mu)$ (=number counts for astronomers) is well known, $p(\mu) = \mu^{\beta}$ with beta approx 2.5 (euclidian slope).

4 photons are observed but the maximum a posteriori (most probable) is about 1.5!

same holds true for cluster velocity dispersions



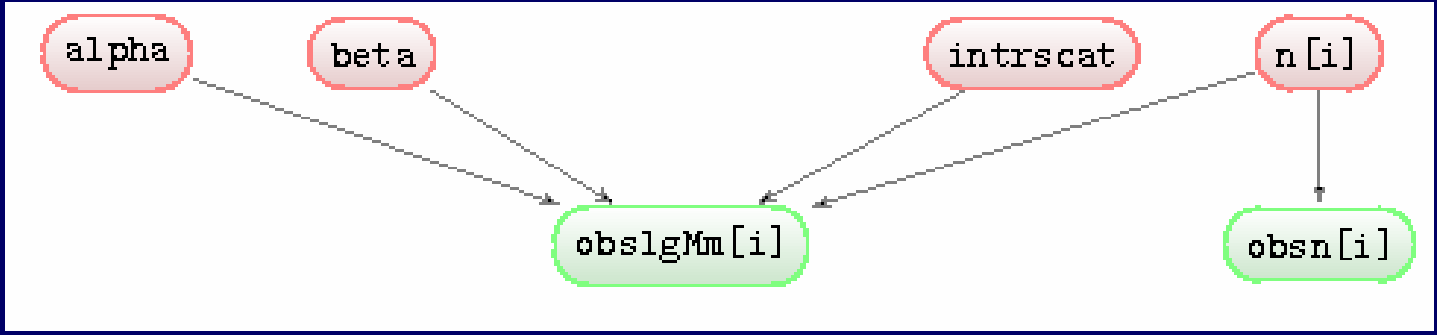
The astronomical addressed problem

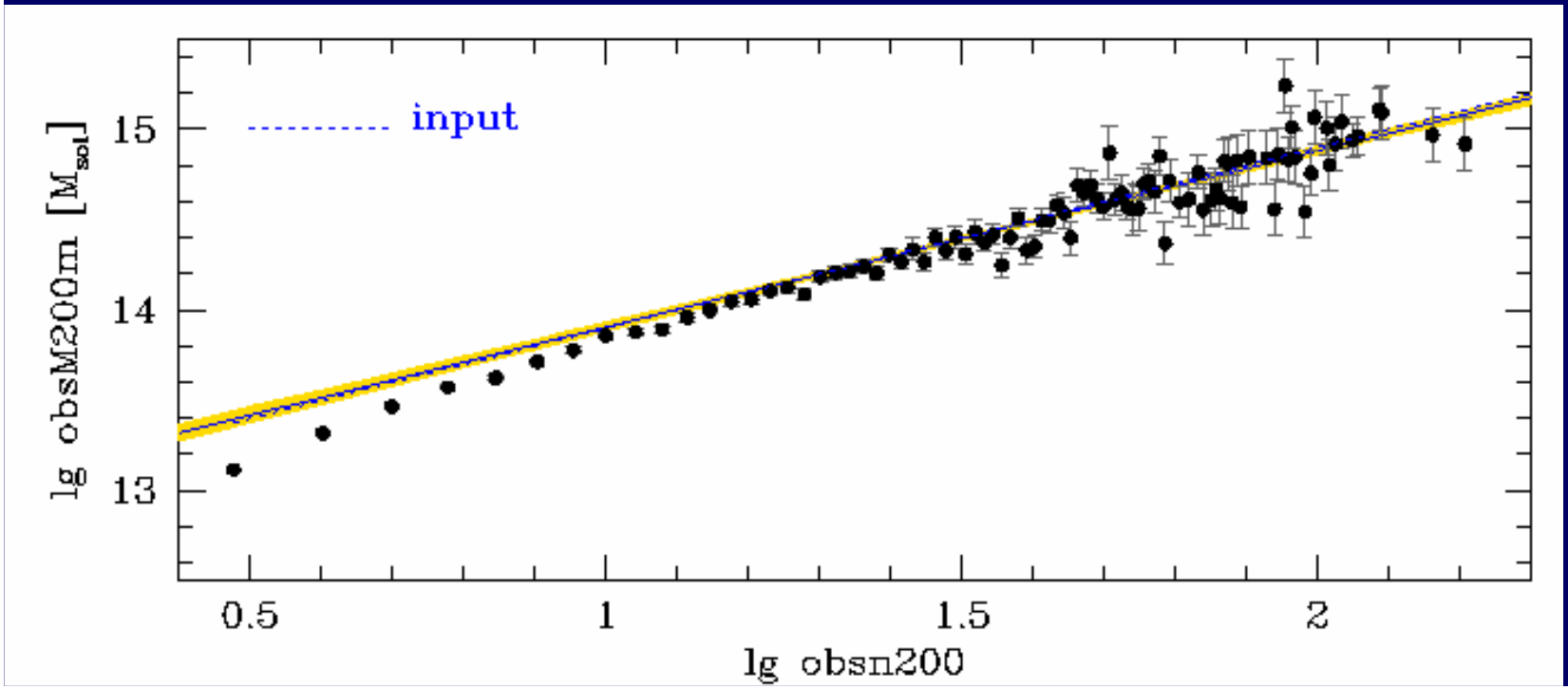


(Weak-lensing) mass of cluster stacks of a given observed value of richness

Fake data, based on a true case, fully described in Andreon & Hurn (2010).

The sheer simplicity of the analysis

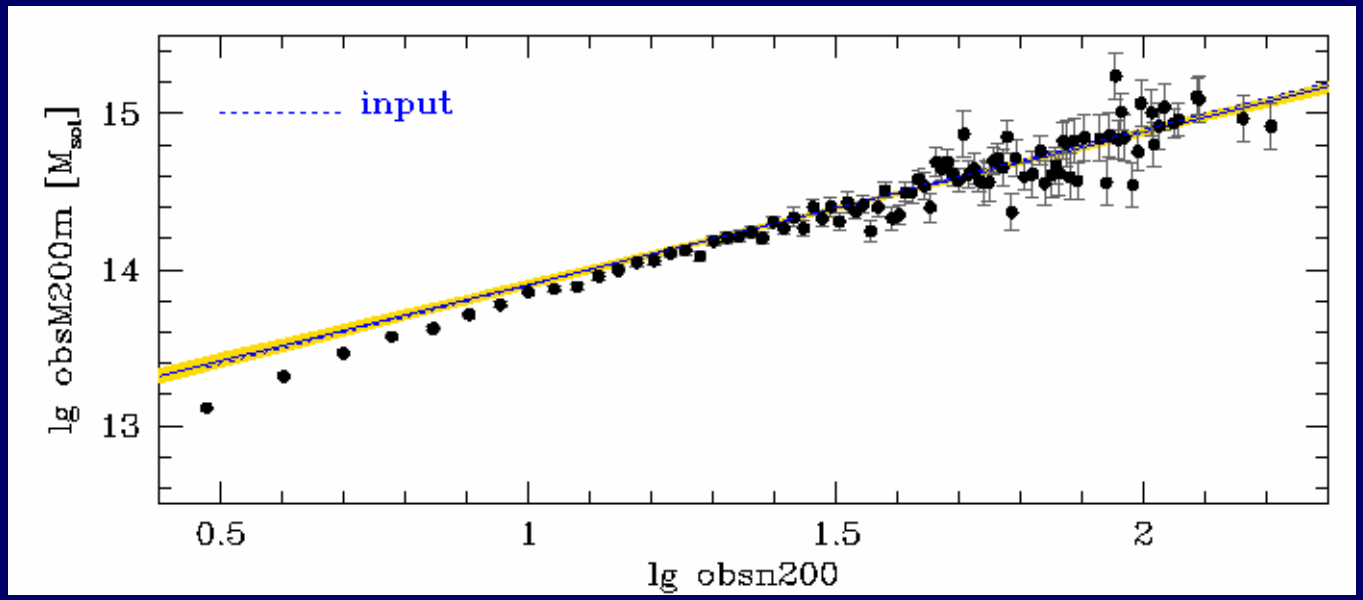




Implementation complex because Schechter is not in JAGS

```
data {                          # JAGS idiosincrasies
zeros<-obsn-obsn
C<-10 }
model {
for (i in 1:length(obsn)) {   #foreach datum
obslgMm[i] ~ dnorm(lgM[i],pow(err[i],-2))     # mass errors
lgM[i] <- alpha*(lgn[i]-1.5)+beta                    # linear relation
obsn[i] ~dpois(n[i])                                  # poiss errors
n[i]  <- pow(10,lgn[i])
# implementing a math distribution missing in JAGS
lgn[i] ~ dunif(-1,3)
phi[i] <- n[i]/10^2-(-2+1-1)*lgn[i]    # Jenkins mass function
zeros[i] ~ dpois(phi[i]+C)
}
alpha ~ dt(0,1,1)                          # prior
beta ~ dnorm(14.4,pow(3,-2))
}
```

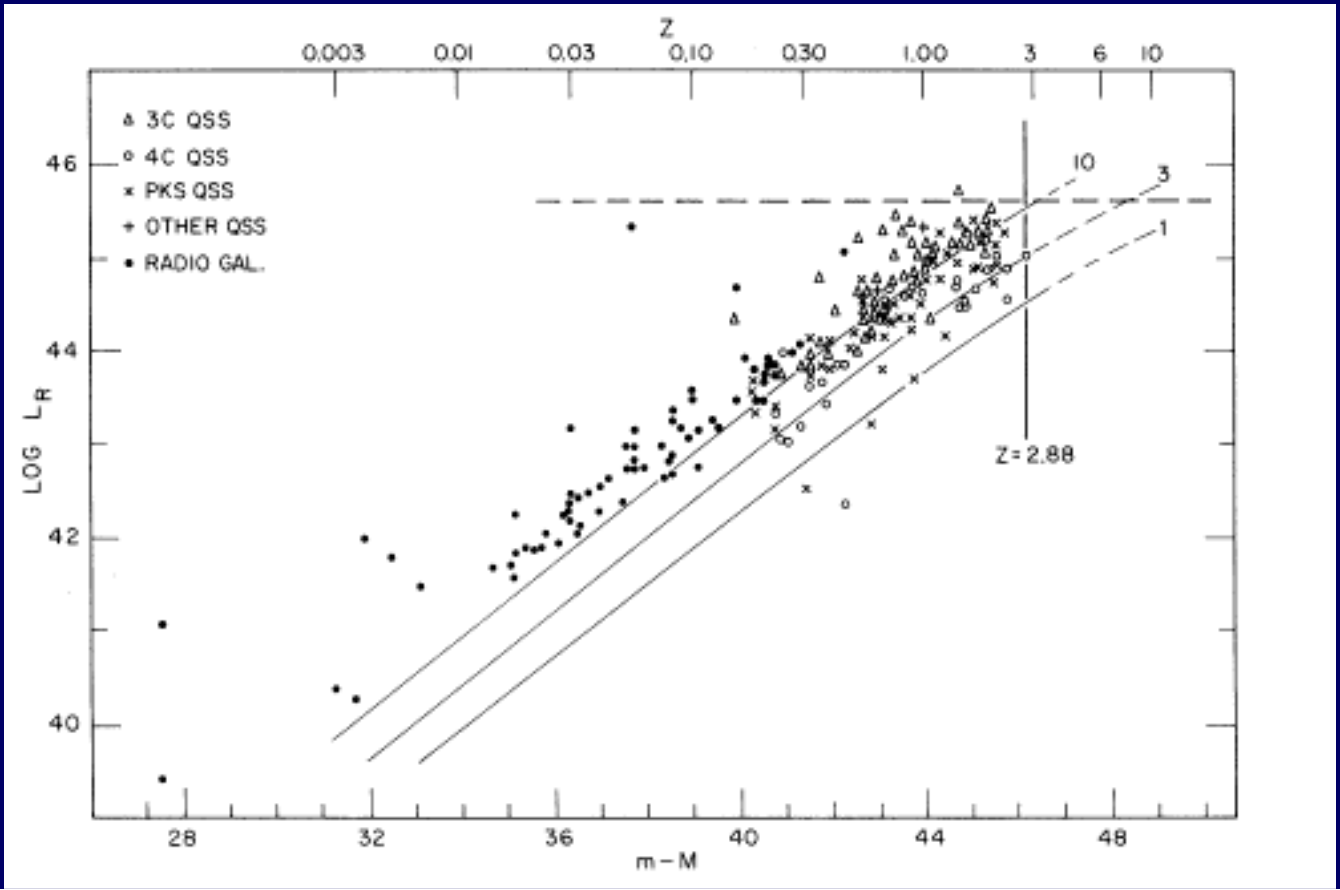
Result:

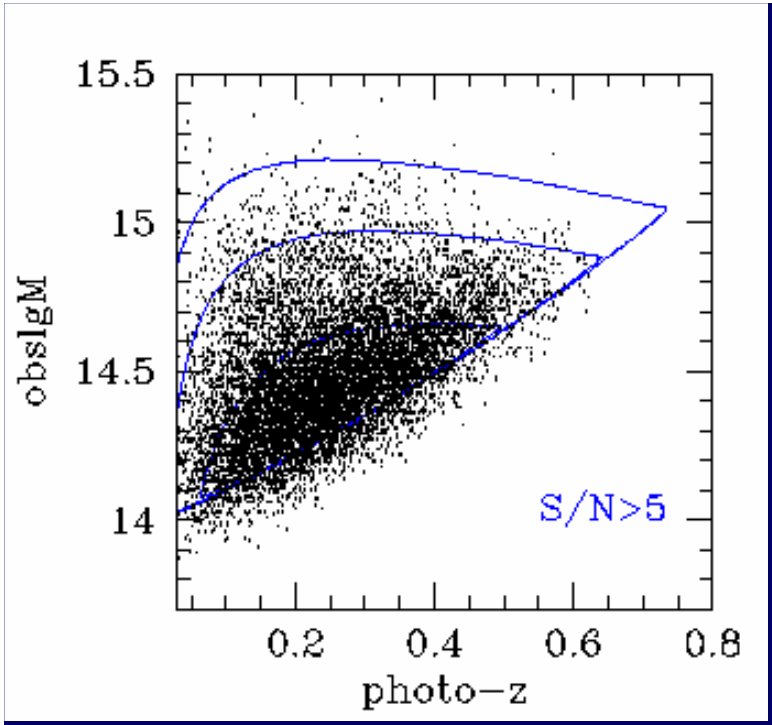


Don't confuse "fitting observed values" to "fitting true values".

Modeling selection effects (non random data collection).

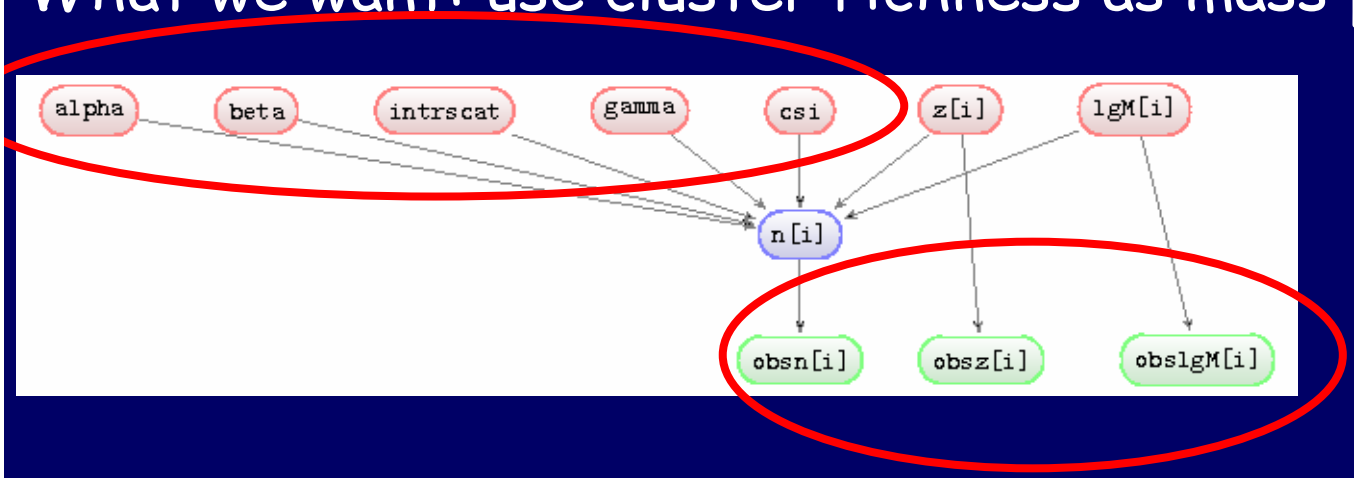
We started with this





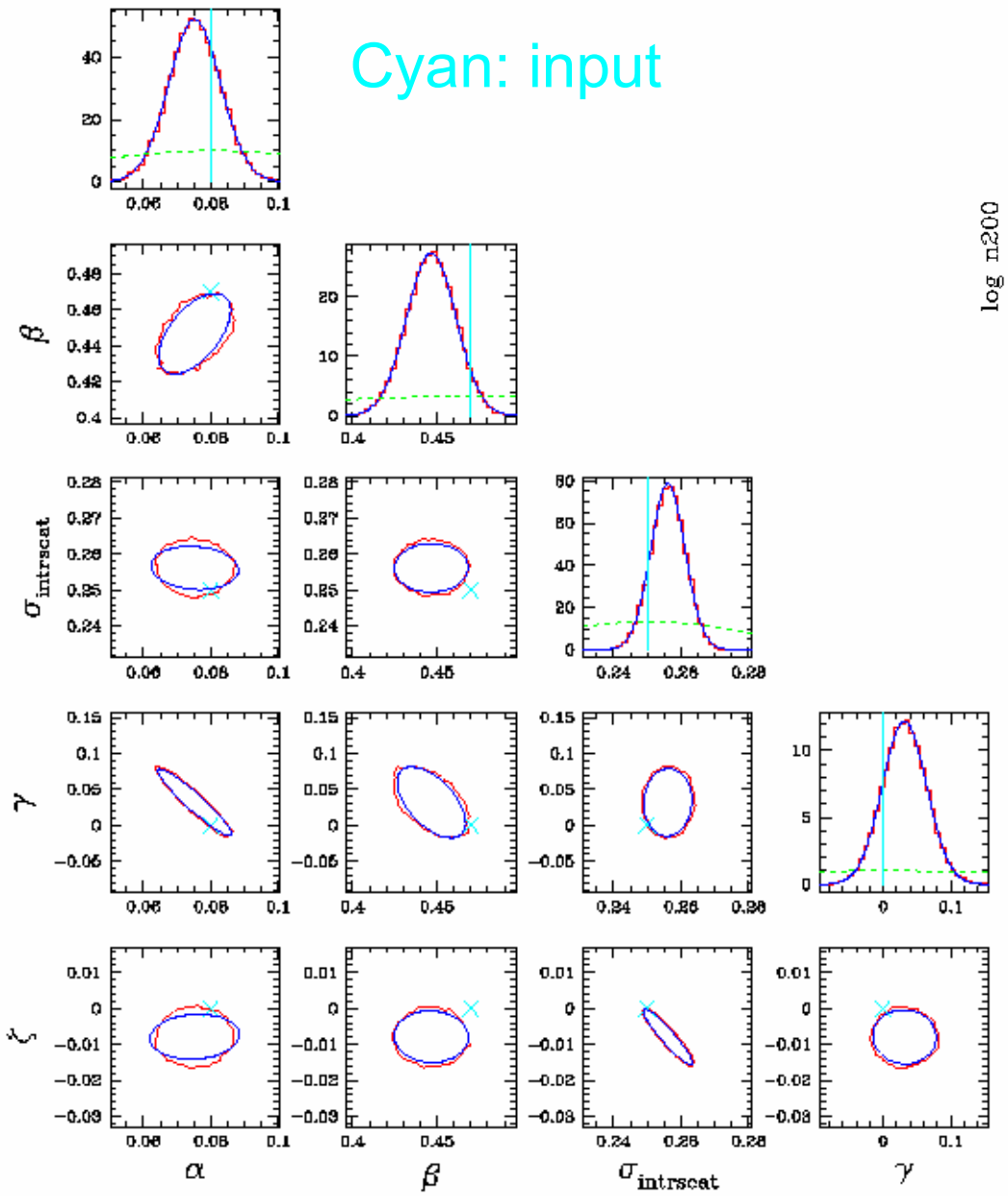
Fake (but realistic) weak lensing masses from the dark mission EUCLID, about 10000 clusters with noisy masses

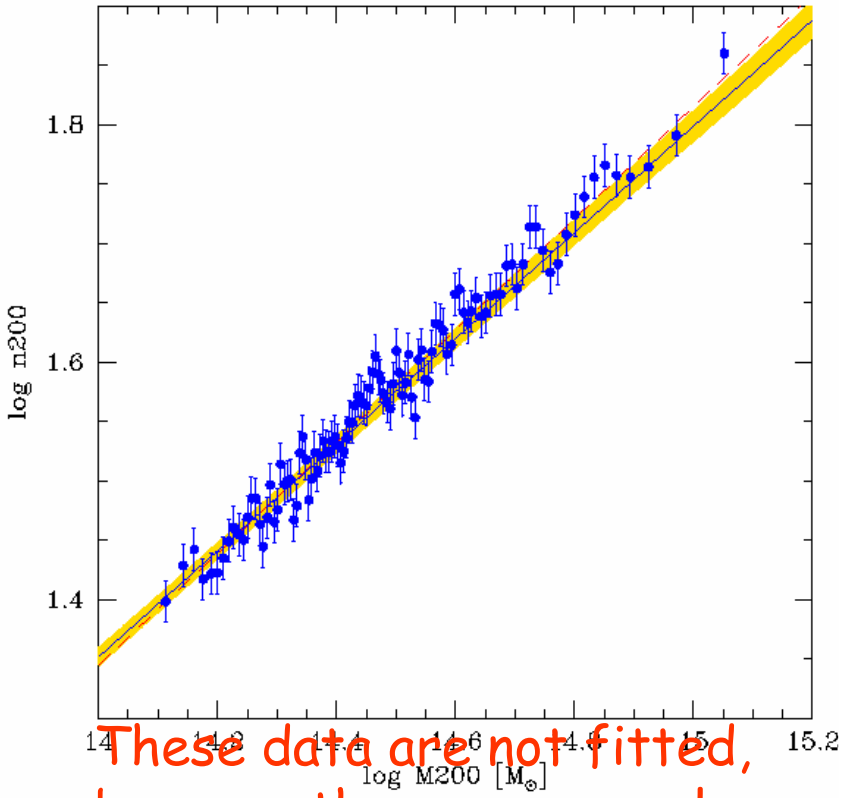
What we want: use cluster richness as mass proxy



The little we can measure. Code in Andreon & Berge 2012

Cyan: input





These data are not fitted, because they are error-less (and thus not available to astronomers). Noisy measurements are fitted!

Up to now focus on parameter estimation (finding relation parameter values)

Move to prediction: I have something (usually cheap to have) and a calibrating sample and I want something (usually costy to measure).

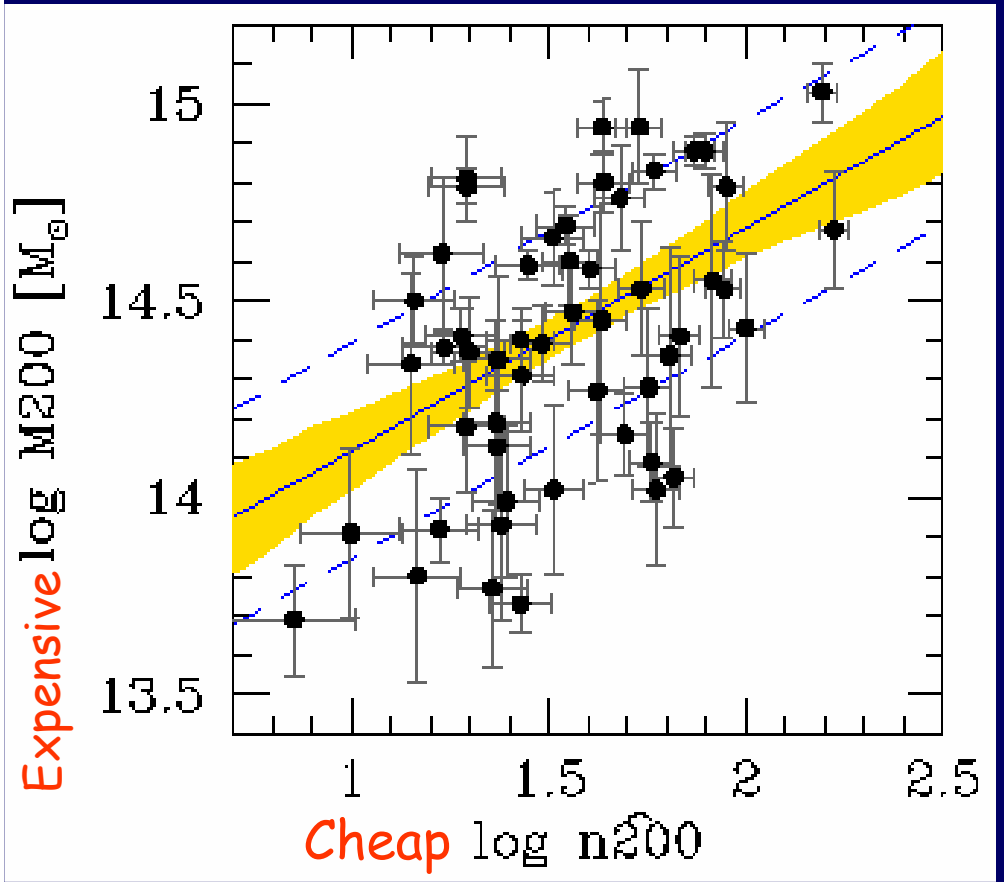
Predicted mass of a cluster.

- You can easily predict the mass of a cluster, knowing its n200, with zero character of model (or code) to type!
  In practique: add data to your file, with NA (not available) for mass. See Andreon & Hurn (2010).

- The theory (definition!) ...

How it works?

- If y is correlated to x, you can, at the first order, use y(xtilde) to predict y of a new xtilde value, reading the y|x relation at x=xtilde.



At first order, the error on y(tilde) should combine the uncertainty in y|x and the xtilde errors

Posterior Predictive Distribution

$$p(\widetilde{lgM}|lgM) = \int p(\widetilde{lgM}|\theta)p(\theta|lgM)d\theta$$

Predicted data

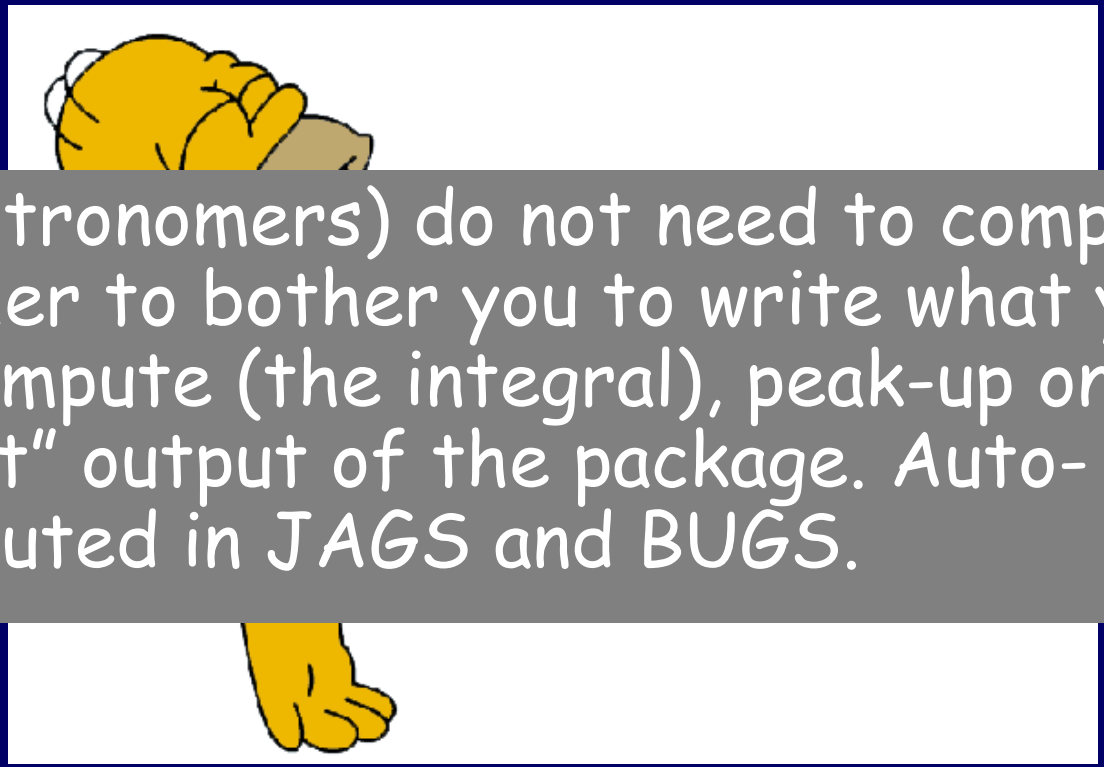
Observed data

What it is: the uncertainty on the predicted value is given by combining the uncertainty on regression parameters, p(θ|lgM), and the probability of new data given the regression parameters. It's just the product rule of probability. It is the error propagation formula, everything included.

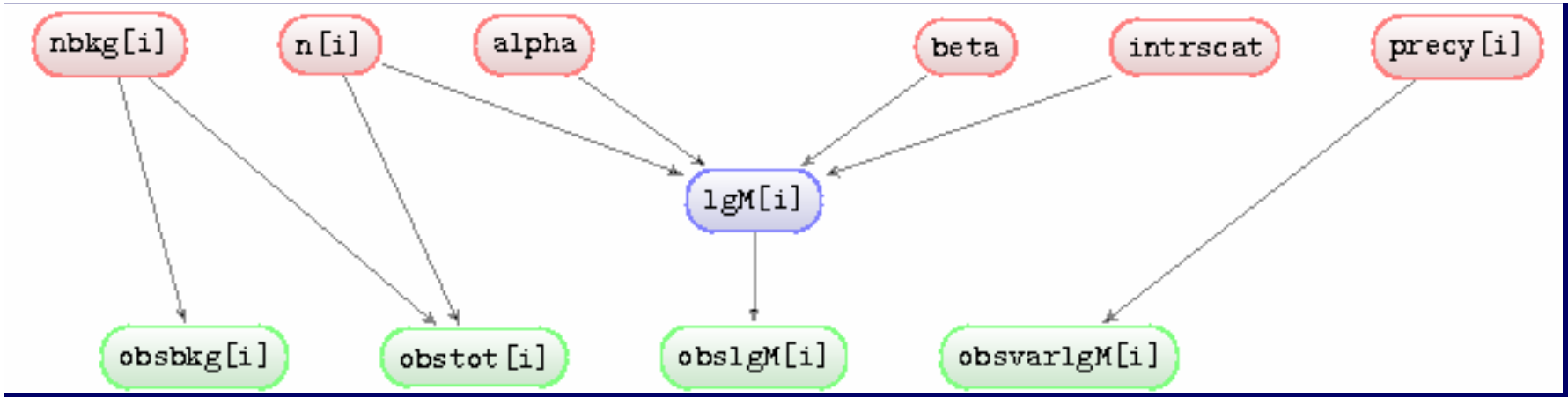
The integral captures:

- Many things in Nature are often non- Gaussian distributed.
- Proxy measurement errors.
- Probably, there is no comparison sample with identical lgM. With some luck, we have clusters of similar lgM.
- No-one known the true M values, we only have noisy (ie with errors) measurements, obslgM.
- Even errors have an error! Could you sincerely measure an error with infinity precision?

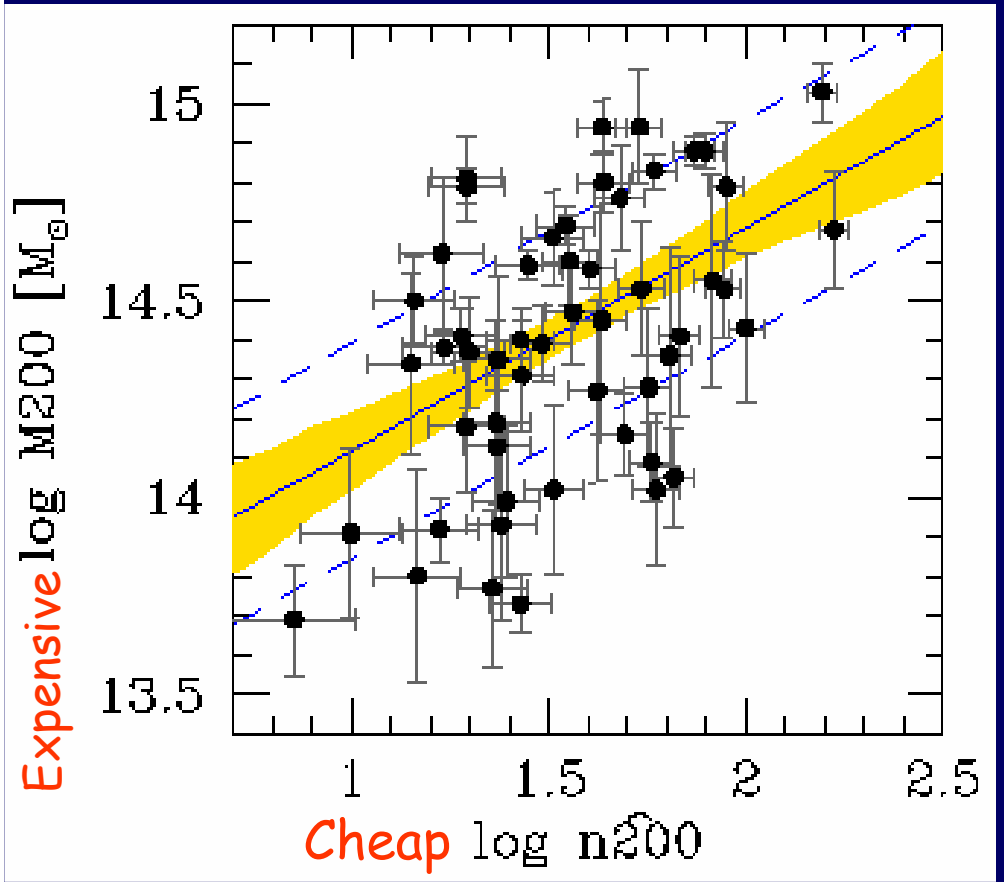
How to compute posterior predictive uncertainty in practice:



You (astronomers) do not need to computed it, neither to bother you to write what you want to compute (the integral), peak-up only the "right" output of the package. Auto-computed in JAGS and BUGS.



Additional complication:
errors have errors!



Model and code in
Andreon & Hurn (2010)

Summary:

- Efficiency: no Gauss err, no linear relation
- Varying physical constant: intrinsic scatter (systematics), heteroscedastic errors on target
- Magorrian relation: heteroscedastic errors on target *and* predictor, and intrinsic scatter
- Quenching: multiple non-linear regression with non-Gauss heteroscedastic errors, mixture.
- Optical-to-X ratio: dealing with upper limits (and heteroscedastic errors, intrinsic scatter, non-gauss errors)
- Mass-richness scaling: everything above+ x data structure (leading to the Eddington-Malquist bias)
- Euclid simulation: everything above + non-random data selection
- Mass-richness: prediction (posterior predictive distribution)

Not addressed for lack of time

- Do the model fit the data? (generate fake data and look if they resemble to real data)
- Role of the prior (plot it on the top of the posterior and use another prior).
- Model selection: the data support this model or that other (use Bayes factor)
- Mixture of regressions.

Suggested lectures

- Everything (and more) is in the >700 pages of Gelman et al. (2003) Bayesian Data Analysis.

Most of the material (and codes) of this lecture is taken from:

- Measurement errors and scaling relations in astrophysics: a review, Andreon & Hurn 2013, Statistical Analysis and Data Mining, 6, 15
- Understanding better (some) astronomical data using Bayesian methods, Andreon 2012, Chapter 2 of "Astrostatistical Challenges for the New Astronomy" (ed. J. Hilbe), Springer Series on Astrostatistics. Largely based on an invited talk at ISI 2011 - 58th World Statistics Congress, Dublin.
- Bayesian methods for galaxy evolution studies, Andreon 2009, Chapter 12 of "Bayesian methods in cosmology", Cambridge Universtiy press
- And/or my own first-author papers (after 2009).

All on arxiv.org