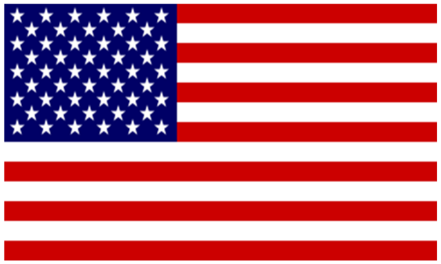
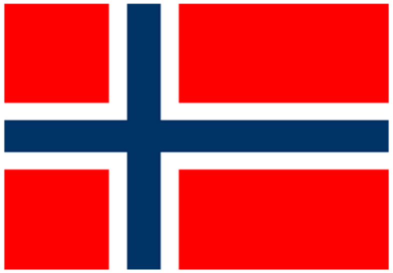
HOW RELIABLE ARE RELIABILITIES?

Daniel Gianola

University of Wisconsin-Madison
Norwegian University of Life Sciences





Georg-August-Universität Göttingen

Department of Animal Sciences: **Prof. Dr. Henner Simianer**

Deutsche
Forschungsgemeinschaft
DFG

Mercator-Gastprofessuren, 2006

Alexander von Humboldt
Stiftung / Foundation

Origin:
1830–40; un-^{1} + reliable

—Related forms
un·re·li·a·bil·i·ty, un·re·li·a·ble·ness, noun
un·re·li·a·bly, adverb

—Synonyms
undependable, irresponsible, untrustworthy.

ACCURACY: EXTENT TO WHICH ONE DEVIATES FROM “TRUTH”

RELIABILITY: PRECISION OF THE STATEMENT

IT IS POSSIBLE TO BE RELIABLE AND INACCURATE: A CONSISTENT LIER

ON THE SEMANTICS OF THE PROBLEM

$$z = \alpha + \beta x$$

$$x = \frac{z - \alpha}{\beta}$$

➡ Variable x is an inaccurate (biased) measure of z

$$Corr(x,z) = \frac{\beta Var(x)}{\sqrt{Var(x)\beta^2 Var(x)}} = 1$$

➡ Variable x is completely reliable (in the correlation sense)

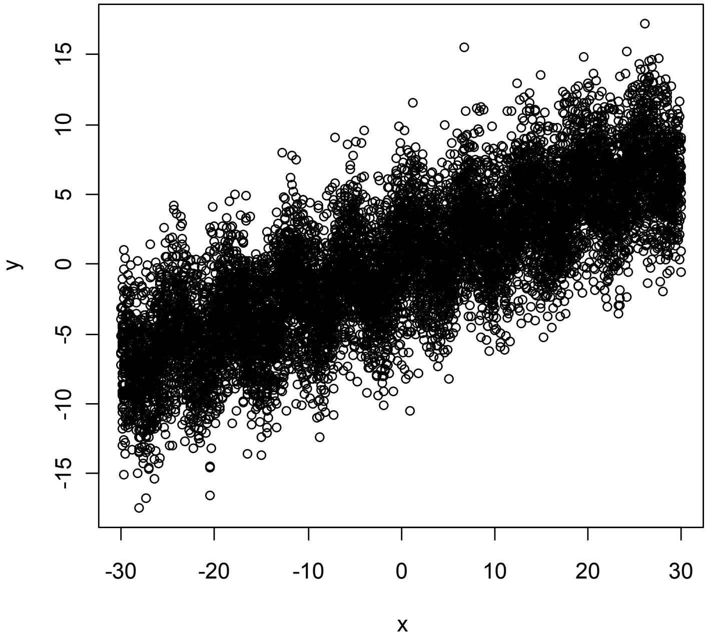
LEARNING “ARCHITECTURE” WITH LINEAR MODELS

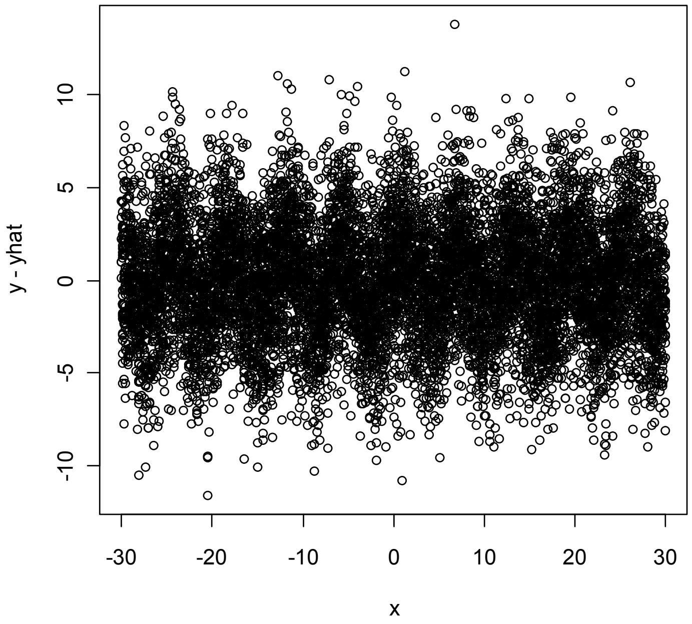
We are given (x,y) data (n=10,000). It looks like this:

```
yhat= 0.07936+0.24814*x

> cor(x,y)
[1] 0.8064256

> cor(y,yhat)
[1] 0.8064256
```





RESIDUALS DISPLAY
SINUSOIDAL BEHAVIOR
WHY?

TRUE MODEL

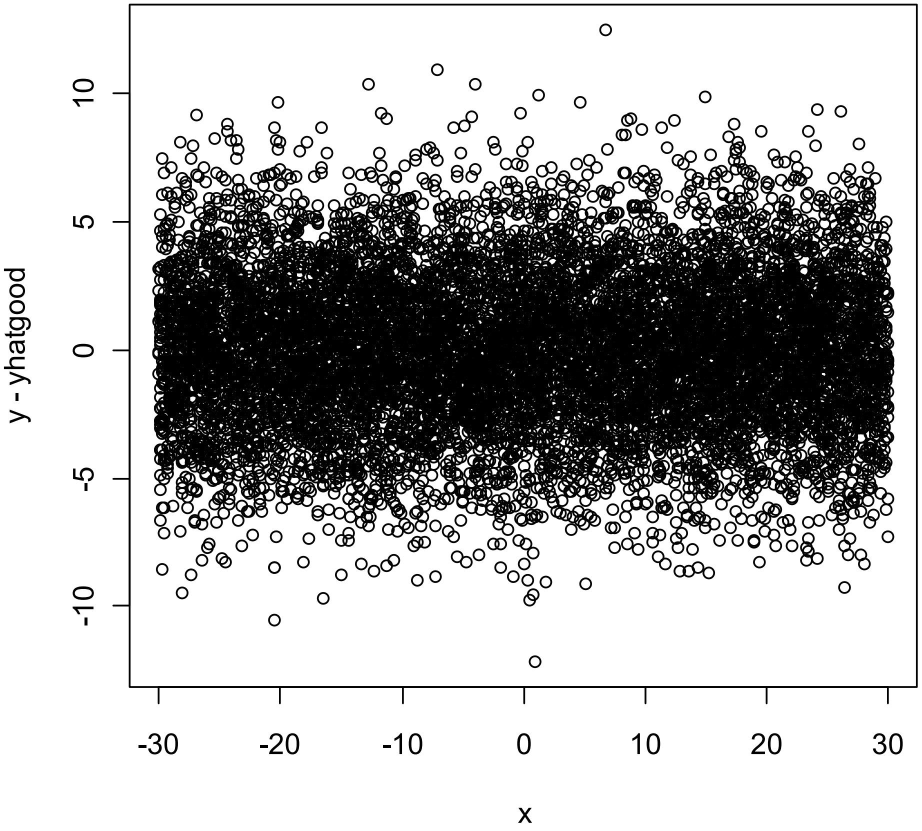
```
> e<-rnorm(10000,0,sqrt(9))
> x<-runif(10000,-30,30)
> a<-0.10
> b<-0.25
>y<-a+b*x+sin(x)+cos(x)+e

> model<-lm(y~x+sin(x)+cos(x))
```

>Estimated coefficients:

```
>(Intercept)          x       sin(x)      cos(x)
> 0.1030       0.2489      0.9518      0.9433
```

RESIDUALS LOOK RANDOM



WE GENERATE A TESTING SET (n=10,000) AT THE SAME VALUES OF X

```
> enew<-rnorm(10000,0,sqrt(9))
>ynew<-a+b*x+sin(x)+cos(x)+enew
```

CALCULATE PREDICTIVE MEAN SQUARED ERROR: “BAD” +“GOOD” MODELS

```
> msepredbadmodel<-sum((ynew-yhat)**2/10000)
> msepredbadmodel=  9.7257

> msepredgoodmodel<-sum((ynew-yhatgood)**2/10000)
> msepredgoodmodel= 8.7293
```

CALCULATE PREDICTIVE CORRELATIONS

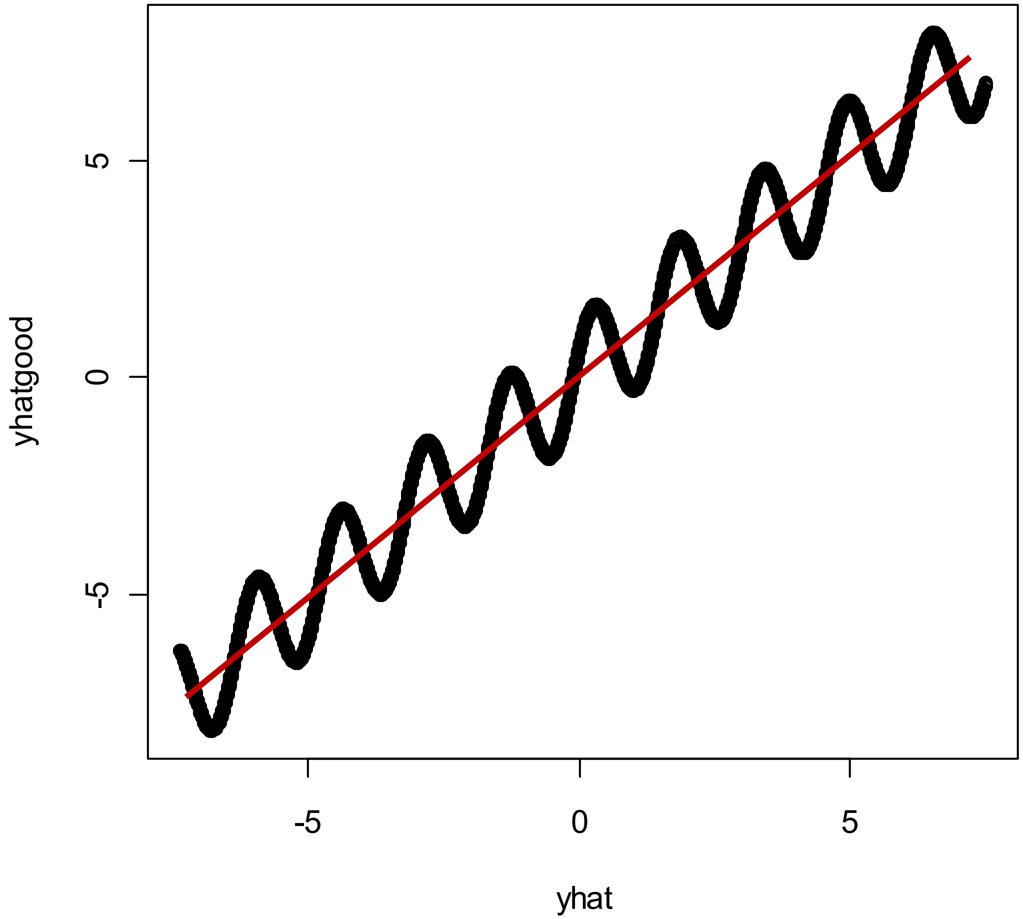
```
> cor(yhat,ynew)=          0.807
> cor(yhatgood,ynew)=   0.829
```

MSE(Good)/MSE(Bad)= 0.90
MSE(Bad)/MSE(Good)= 1.11

Cor(BAD)/Cor(GOOD)= 0.97

WHAT HAVE WE LEARNED FROM A LARGE CORRELATION?

```
> lm(yhat~yhatgood)
Coefficients:
(Intercept)     yhatgood
   0.005653      0.953468
```



<u>Point 1</u>: HAS THE LINEAR MODEL LED US TO UNDERSTAND “ARCHITECTURE”?

Reliability: standard formulae

$$reliability = 1 - \frac{PEV}{Var(u)}$$

$$unreliability = 1 - reliability = \frac{PEV}{Var(u)}$$

$$\begin{aligned} PEV &= Var(\hat{u} - u) = Var_{\hat{u}}[E(\hat{u} - u)|\hat{u}] + E_{\hat{u}}(Var[(\hat{u} - u)|\hat{u}]) \\ &= E_{\hat{u}}(Var[(\hat{u} - u)|\hat{u}]) \\ &= E_{\hat{u}}(Var[u|\hat{u}]) \end{aligned}$$

$$\begin{aligned} Var(u) &= E_{\hat{u}}(Var[u|\hat{u}]) + Var_{\hat{u}}(E[u|\hat{u}]) \\ E_{\hat{u}}(Var[u|\hat{u}]) &\leq Var(u) \end{aligned}$$

Under MULVN $\Rightarrow \; E_{\hat{u}}(Var[u|\hat{u}]) = Var[u|\hat{u}]$

IN SOME NON-GAUSSIAN MODELS, POSTERIOR VARIANCE CAN BE SOMETIMES LARGER THAN PRIOR VARIANCE, LEADING TO NEGATIVE RELIABILITY , AND POSITIVE UNRELIABILITY.

POINT 2: WHAT WE CALL “RELIABILITY” IS VERY MUCH TAILORED FOR NORMAL DISTRIBUTIONS AND LINEAR MODELS

ILLUSTRATION OF SOMEPOINTS

Standard analysis (fixed **X**)

Genotypic value (signal from genome)

Assumption

$$y = f + e = X\beta + e$$

Bayesian or Frequentist?
(more later)

$$\beta|\sigma^2_\beta \sim N(0, I\sigma^2_\beta)$$

$$E(y|X,\beta) = X\beta$$

$$E(y|X) = 0$$

$$Var(y|X,\sigma^2_e,\sigma^2_\beta) = XX'\sigma^2_\beta + I\sigma^2_e$$

Prediction of marker effects: BLUP (iid marker effects)

$$\left[X'X + \frac{\sigma_e^2}{\sigma_\beta^2} I\right]\widehat{\beta} = X'y$$

Assume inverse exists

$$\left[I + \frac{\sigma_e^2}{\sigma_\beta^2}(X'X)^{-1}\right]\widehat{\beta} = (X'X)^{-1}X'y$$

$$\widehat{\beta} = \left[I + \frac{\sigma_e^2}{\sigma_\beta^2}(X'X)^{-1}\right]^{-1}\widetilde{\beta}_{\text{OLS}} \Rightarrow \text{SHRINKAGE}$$

Prediction of signal ($X\beta$) to phenotype

$$Var(X\beta|y) = XVar(\beta|y)X'$$

$$= X\left[I + \frac{\sigma_e^2}{\sigma_\beta^2}(X'X)^{-1}\right]^{-1}X'\sigma_e^2$$

Prediction of future record

$$y^{*} = X^{*}\beta + e^{*}$$

$$E(X^{*}\beta + e^{*}|y, X, X^{*}) = X^{*}E(\beta|y, X)$$

$$= X^{*}\left[I + \frac{\sigma_e^2}{\sigma_\beta^2}(X'X)^{-1}\right]^{-1}\widetilde{\beta}_{\text{OLS}}$$

$$Var(X^{*}\beta + e^{*}|y, X, X^{*}) = X^{*}Var(\beta|y, X)X^{*} + I^{*}\sigma_e^2$$

GAUSSIAN PROCESS ANALYSIS (IID MARKER EFFECTS)

$$y = f + e = X\beta + e$$

$$\beta \sim N(0, I\sigma_\beta^2)$$

[Read Falconer and Mackay IQG]

$$X \sim F$$

$$E(y|X,\beta) = X\beta$$

[Genotypes vary at random: population Genetics]

$$\left.\begin{aligned} E(y|\beta) &= E_X E(y|X,\beta) = E(X)\beta \\ E(y) &= E_\beta[E(X)\beta] = E(X)E(\beta) = 0 \end{aligned}\right\}$$

Big assumption

Are frequencies effect-dependent? Are effects frequency dependent?
[TURELLI, ZHANG&HILL, MACKAY WITH MARKERS AND

$$Var(y) = Var(f) + Var(e) = Var(f) + I\sigma_e^2$$

$$\begin{aligned} Var(f) &= Var(X\beta) \\ &= E_X(Var(X\beta|X) + Var_X[E(X\beta|X)] \\ &= E_X[XVar(\beta)X']) + Var_X[XE(\beta)] \\ &= E_X[XX'\sigma_\beta^2] + Var_X(0) \\ &= \sigma_\beta^2 E_X[XX'], \end{aligned}$$

Covariance matrix of signal

BP= “best predictor”
(MULVN assumed)

$$\begin{aligned} \hat{f} &= BP(f) \\ \left[\frac{1}{\sigma_e^2} I + Var^{-1}(f)\right]\hat{f} &= \frac{1}{\sigma_e^2} y \\ \left[I + \frac{\sigma_e^2}{\sigma_\beta^2} E_X^{-1}[XX']\right]\hat{f} &= y \\ E_X^{-1}[XX']\left[E_X[XX'] + \frac{\sigma_e^2}{\sigma_\beta^2} I\right]\hat{f} &= y \\ \left[E_X[XX'] + \frac{\sigma_e^2}{\sigma_\beta^2} I\right]\hat{f} &= E_X[XX']y \end{aligned}$$

Under multivariate normality

$$\begin{aligned}
Var(f|y) &= Var(f) - Cov(f,y)Var^{-1}(y)Cov'(f,y) \\
&= Var(f) - Var(f)[Var(f) + I\sigma_e^2]^{-1}Var(f) \\
&= \sigma_\beta^2 E_X[XX'] - \sigma_\beta^2 E_X[XX'][\sigma_\beta^2 E_X[XX'] + I\sigma_e^2]^{-1}\sigma_\beta^2 E_X[XX'] \\
&= \sigma_\beta^2 E_X[XX'] - \sigma_\beta^2 E_X[XX']\frac{E_X^{-1}[XX']}{\sigma_\beta^2}\left[I + \frac{\sigma_e^2}{\sigma_\beta^2}E_X[XX']\right]^{-1}\sigma_\beta^2 E_X[XX'] \\
&= \left\{I - \left[I + \frac{\sigma_e^2}{\sigma_\beta^2}E_X[XX']\right]^{-1}\right\}\sigma_\beta^2 E_X[XX'].
\end{aligned}$$

Proper assessment of posterior uncertainty requires knowledge
of the genotypic distribution

$$X_{\text{ind,marker}} = \begin{bmatrix} x_{11} & . & x_{1p} \\ x_{21} & . & x_{2p} \\ . & . & . \\ x_{n1} & . & x_{np} \end{bmatrix}$$

$$XX' = \begin{bmatrix} x_{11} & . & x_{1p} \\ x_{21} & . & x_{2p} \\ . & . & . \\ x_{n1} & . & x_{np} \end{bmatrix} \begin{bmatrix} x_{11} & x_{21} & . & x_{n1} \\ . & . & . & . \\ x_{1p} & x_{2p} & . & x_{np} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{j=1}^{p} x_{1_j}^2 & \sum_{j=1}^{p} x_{1_j} x_{2_j} & & \sum_{j=1}^{p} x_{1_j} x_{n_j} \\ & \sum_{j=1}^{p} x_{2_j}^2 & & \\ & & . & \\ & & & \sum_{j=1}^{p} x_{n_j}^2 \end{bmatrix}$$

Under HW

$$
\begin{aligned}
E\left(\sum_{j=1}^{p} x_{ij}^{2}\right) &= \sum_{j=1}^{p} Var(x_{ij}) + \sum_{j=1}^{p} E^{2}(x_{ij}) \\
&= \sum_{j=1}^{p} 2p_j q_j + \sum_{j=1}^{p} (p_j - q_j)^2 \\
&= \sum_{j=1}^{p} (1 - 2p_j q_j) = p - \sum_{j=1}^{p} 2p_j q_j \\
E\left(\sum_{j=1}^{p} x_{1_j} x_{2_j}\right) &= \sum_{j=1}^{p} Cov(x_{1_j} x_{2_j}) + \sum_{j=1}^{p} E(x_{1_j}) E(x_{2_j}) \\
&= \sum_{j=1}^{p} 2\phi_{ij} p_j q_j + \sum_{j=1}^{p} (p_j - q_j)^2 \\
&= \sum_{j=1}^{p} p_j^2 + q_j^2 - 2p_j q_j (1 - \phi) \\
Cov(x_{1_j} x_{2_j}) &= p_j^2 + q_j^2 - 2p_j q_j (1 - \phi) - (p_j - q_j)^2 \\
&= 2pq\phi
\end{aligned}
$$

-How to obtain sensible estimates? Is **XX**’ a good estimate of E(**XX**’)?

-Should we assume HW and use estimates of allelic frequencies and of φ (i,j) as if there were no selection, etc.?

Future record:

$$f^* = X^*\beta + e^*$$

uno

$$E(f^*|f) = E(f^*) + Cov(X^*\beta, \beta X')Var^{-1}(f)f$$

dos

$$E(f^*|y) = E_{f|y}E(f^*|f,y) = E_{f|y}E(f^*|f)$$

tres

$$= E_{f|y}[Cov(X^*\beta, \beta X')Var^{-1}(f)f]$$

$$\begin{aligned} Cov(X^*\beta, \beta X') &= Cov[E(X^*\beta, \beta X'|X^*, X)] + E[Cov(X^*\beta, \beta X')|X^*, X] \\ &= \sigma_\beta^2 E[X^*X'] \end{aligned}$$

cuatro

$$\begin{aligned} E(f^*|y) &= E_{f|y}\left[\sigma_\beta^2 E[X^*X'](XX'\sigma_\beta^2 + I\sigma_e^2)^{-1} f\right] \\ &= \sigma_\beta^2 E[X^*X']E\left[(XX'\sigma_\beta^2 + I\sigma_e^2)^{-1}\right]\widehat{f} \end{aligned}$$

cinco

DOES ANYBODY KNOW HOW TO COMPUTE THE ABOVE?
(CALCULATING THE PEV IS EVEN MORE INVOLVED)

IS MY MODEL “RIGHT”?

TAKING MODEL UNCERTAINTY INTO ACCOUNT BY MODEL AVERAGING

$$
\begin{aligned}
p(\theta|y) &= \sum p(\theta|y, M)p(M|y) \\
&= \sum p(\theta|y, M)p(M|y)dM
\end{aligned}
$$

THE PUNCH LINE: VARIANCE OF PREDICTION ERRORS TAKING MODEL UNCERTAINTY INTO ACCOUNT

$$Var(\theta|y) = E_M[Var(\theta|y, M)] + Var[E_M[\theta|y, M]]$$

Average PEV

Variance among predictions from different models

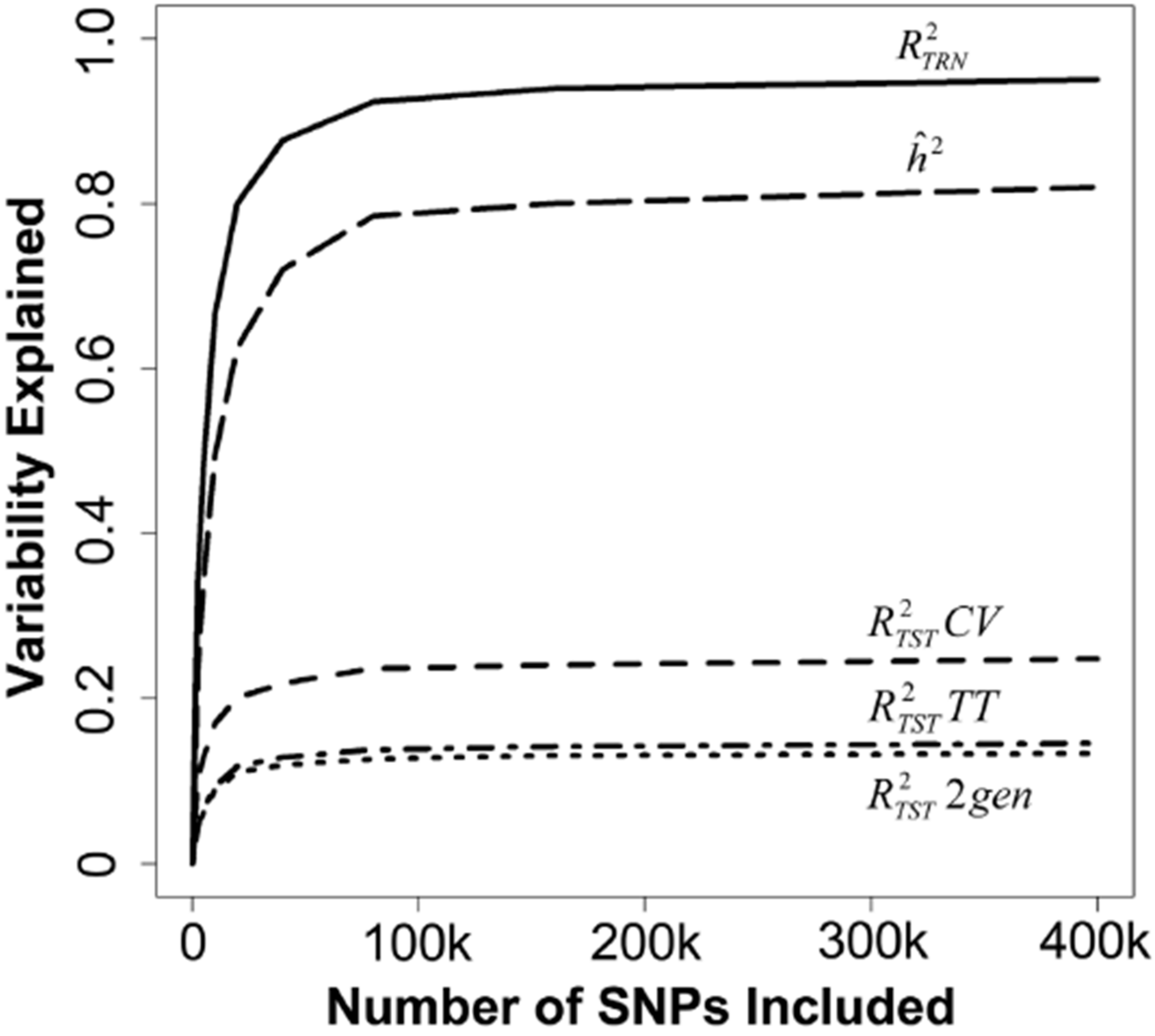
POINT 3: CROSS-VALIDATION

(take model uncertainty into account:
never did this in the BLUP era)

➔A. Prediction and goodness of fit are different ball games (a model that fits well to training data may have atrocious predictive ability

➔B. Any cross-validation scheme (e.g., k-folds) has a cross-validation distribution

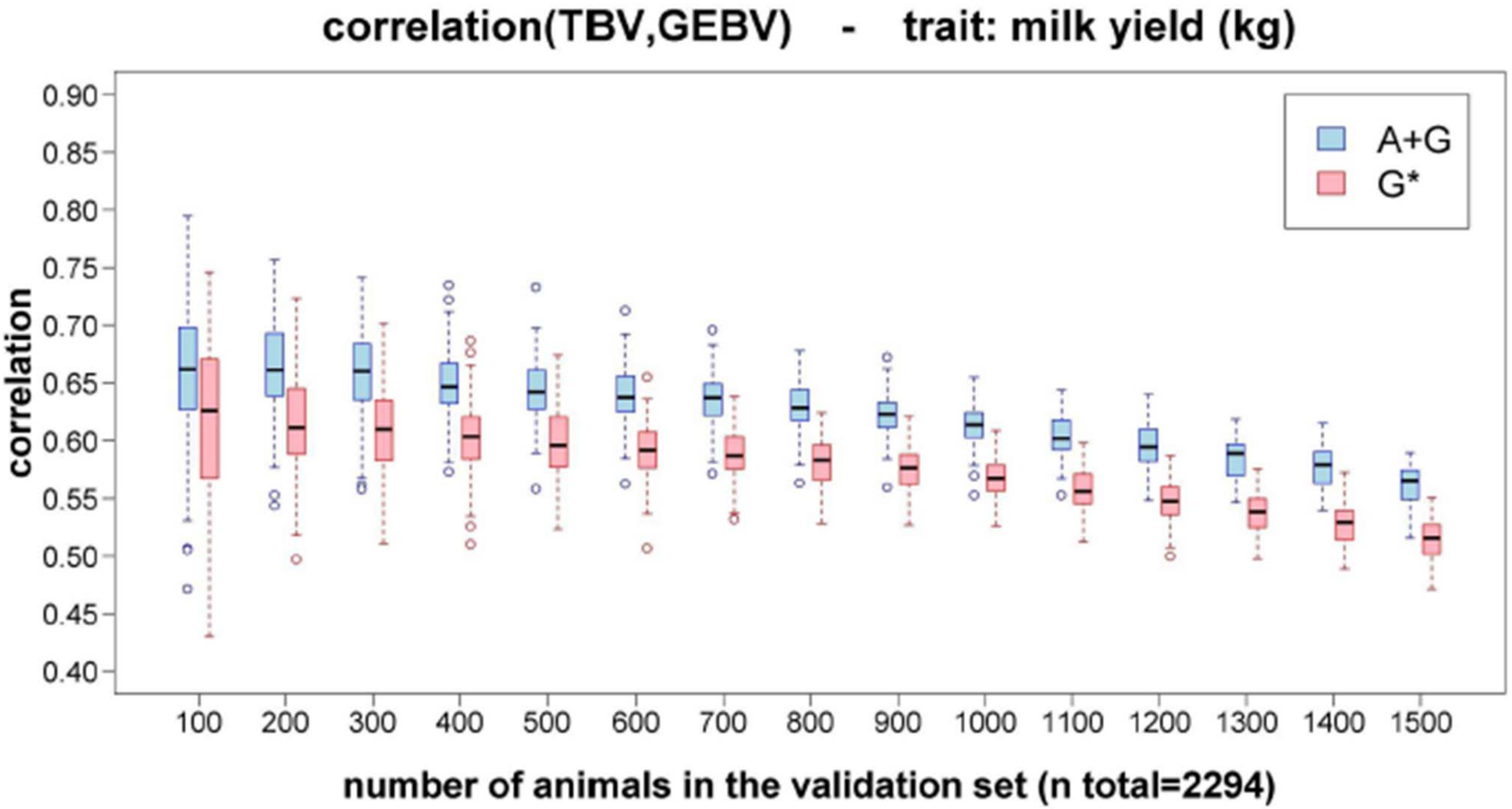
3A. GOODNESS OF FIT (TRAINING= TRN) vs. **PREDICTIVE ABILITY** (TESTING= TST)



HUMAN STATURE: MAKOWSKY et al. , Plos Genetics 2011

3B. CROSS-VALIDATION UNCERTAINTY

(Erbe et al. 2010)



Recapitulating:
How should “reliability” be measured?

- “Apparent reliability”: model dependent
- Formulae for linear model with Gaussian assumptions (even if “true”) do not carry universally to other models
- There is model uncertainty…
- There is cross-validation uncertainty, possibly stemming from heterogeneity in training and testing samples, or in environments
- Should strive for “empirical” measures of “reliability” : cautious with model-derived expressions
- Not obvious how to develop a, say, sire specific cross-validation reliability

Things perhaps one should do

1. Look at “hat” matrix

TRN

$$\widehat{\beta} = (X'X + I\lambda)^{-1}X'y$$

$$\widehat{y} = X(X'X + I\lambda)^{-1}X'y$$

$$\frac{\partial \widehat{y}}{\partial y} = X(X'X + I\lambda)^{-1}X'$$

$$H_{TRN} = X(X'X + I\lambda)^{-1}X'$$

Influence depends on data structure

TST

$$\widehat{\beta}_{TST} = (\widehat{y}'\widehat{y})^{-1}\widehat{y}'y_{TST}$$

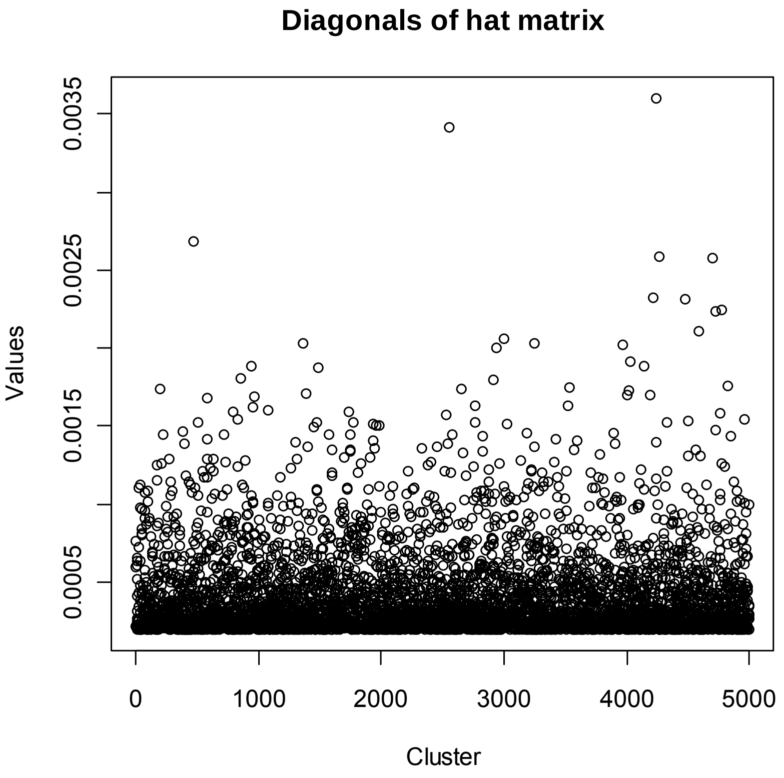
$$\widehat{y}_{TST} = \widehat{y}(\widehat{y}'\widehat{y})^{-1}\widehat{y}'y$$

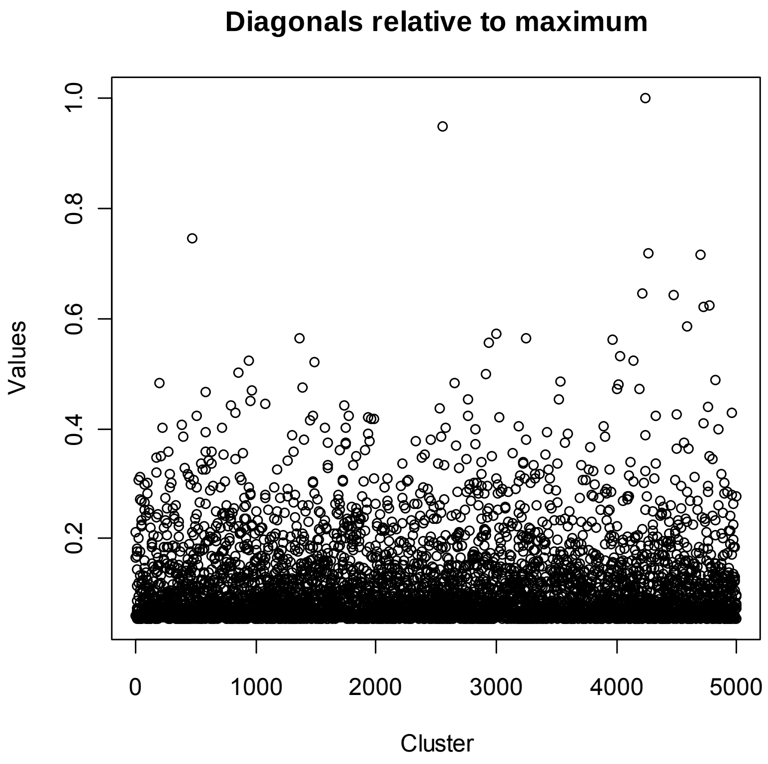
$$H_{TST} = \widehat{y}(\widehat{y}'\widehat{y})^{-1}\widehat{y}$$

Influence depends on training data

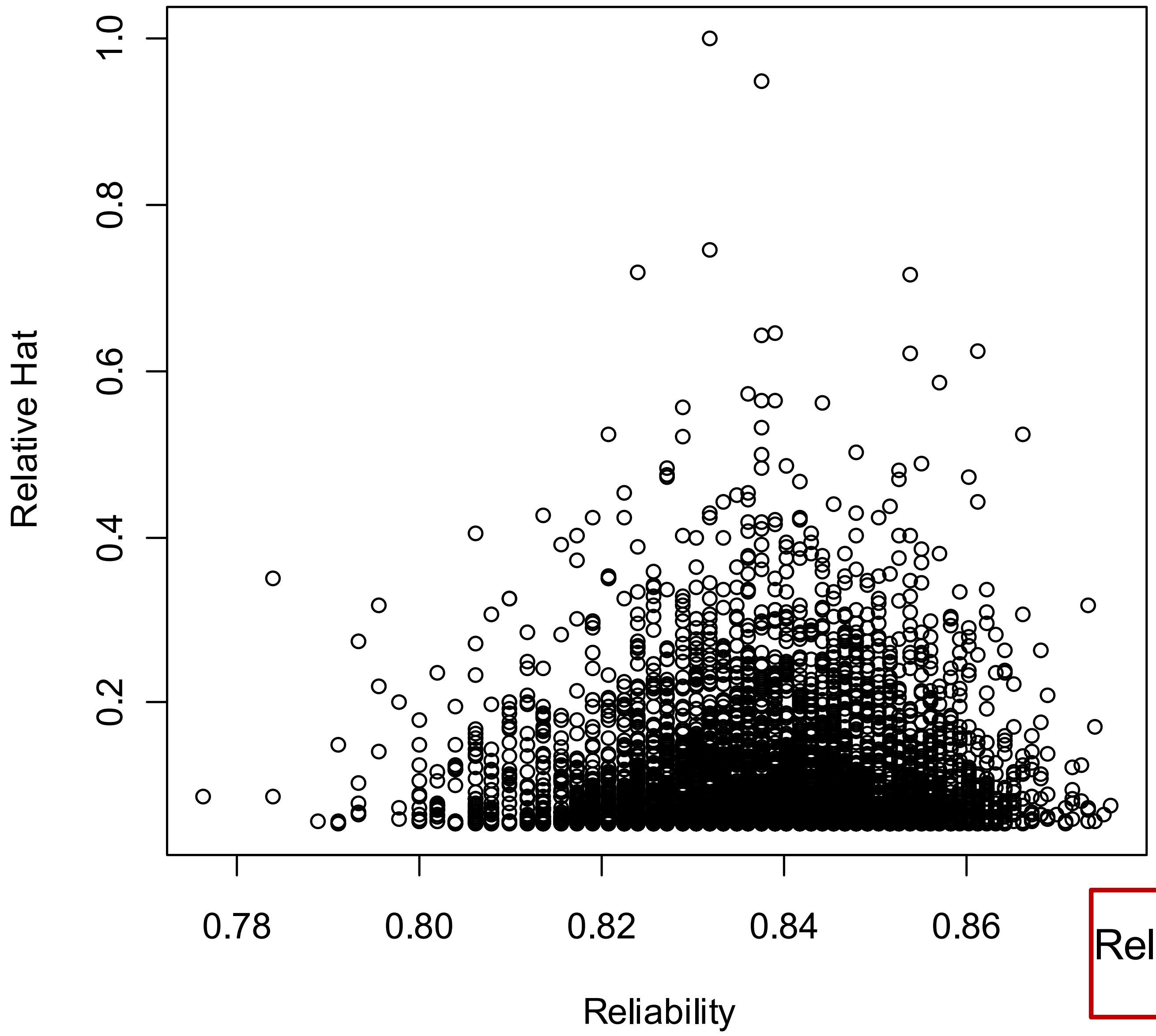
Example: 5000 sires, heritability= 0.20, Var(y)= 1000,
n(TRN)=Poisson(100), n(TST)=Poisson(150)
Predictor(BLP sire), Predictand(Future performance of sire progeny)

TESTING SET





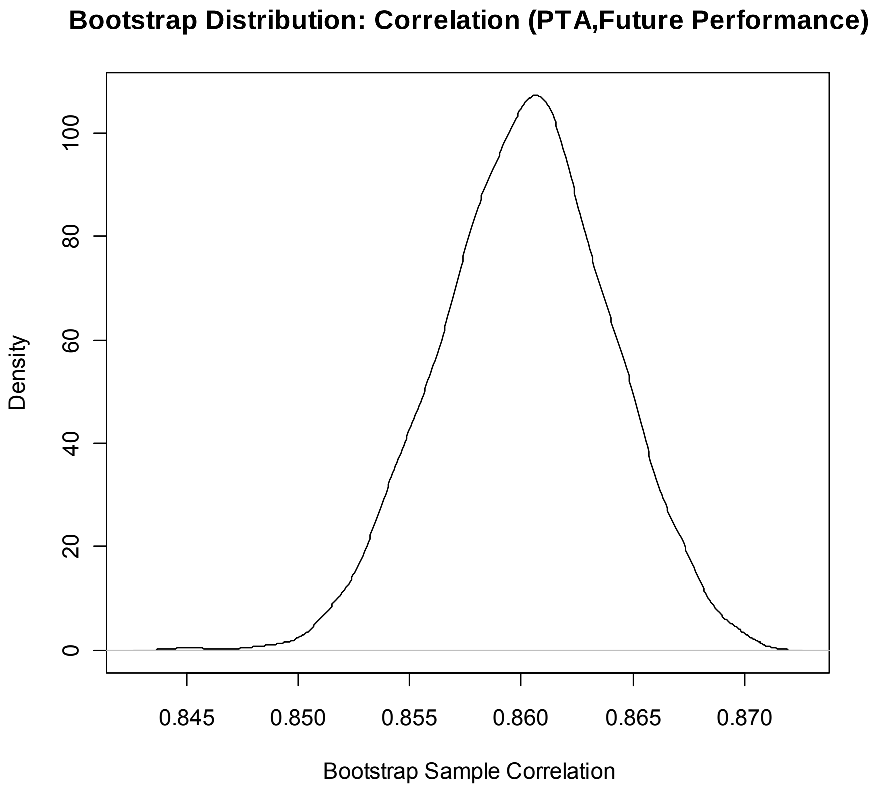
Reliability vs. Relative Hats

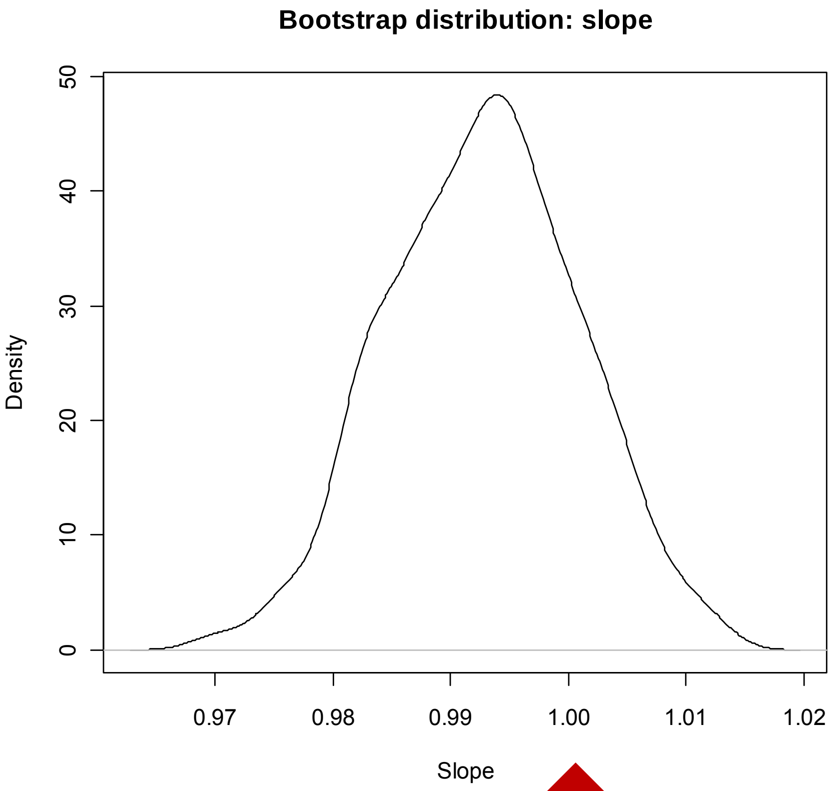


$$\mathrm{Rel}(cluster\ i) = \frac{n_i}{n_i + \frac{4-h^2}{h^2}}$$

Things perhaps one should do

2. Bootstrap CV regression line or CV correlation (no IID)


Bootstrap Distribution: Correlation (PTA,Future Performance)

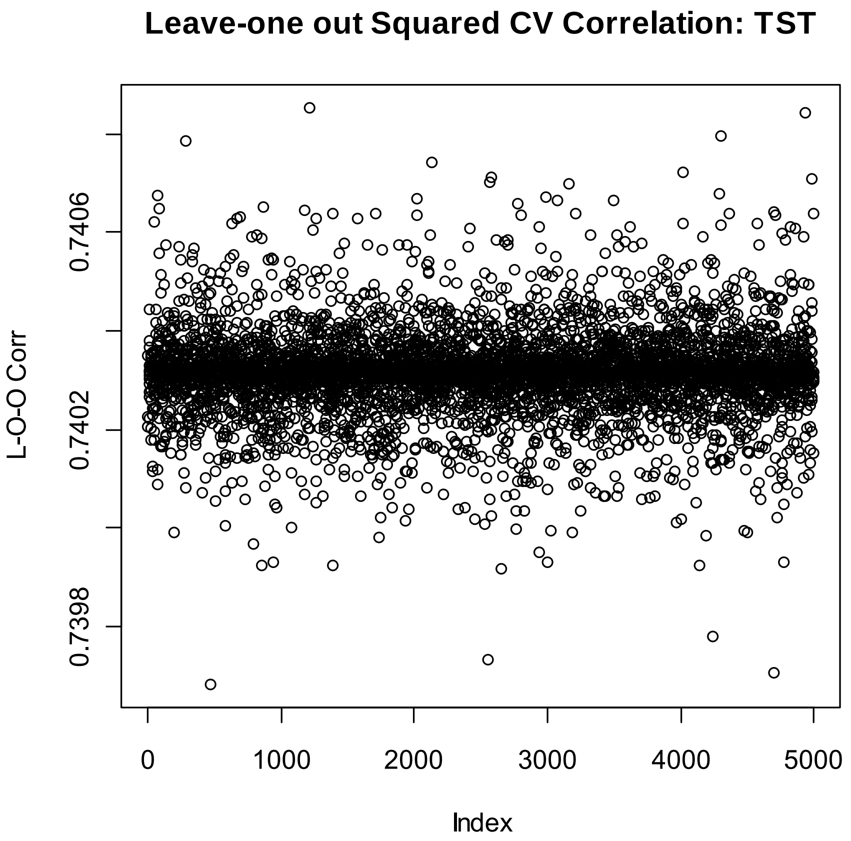

Bootstrap distribution: slope

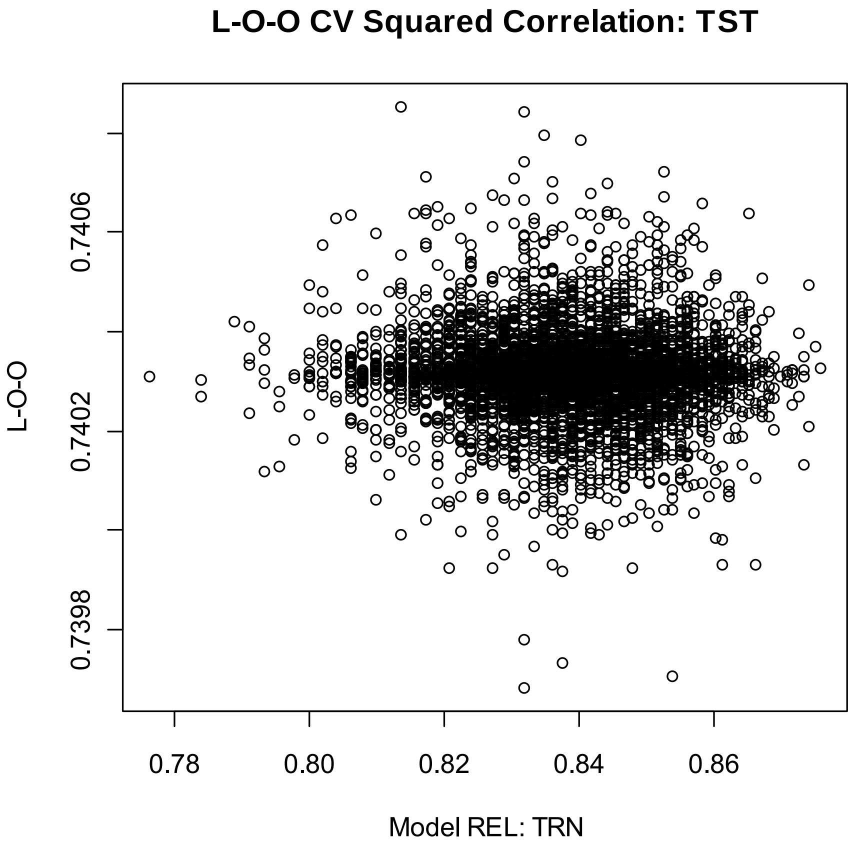
500 bootstrap samples

nice

Things perhaps one should do

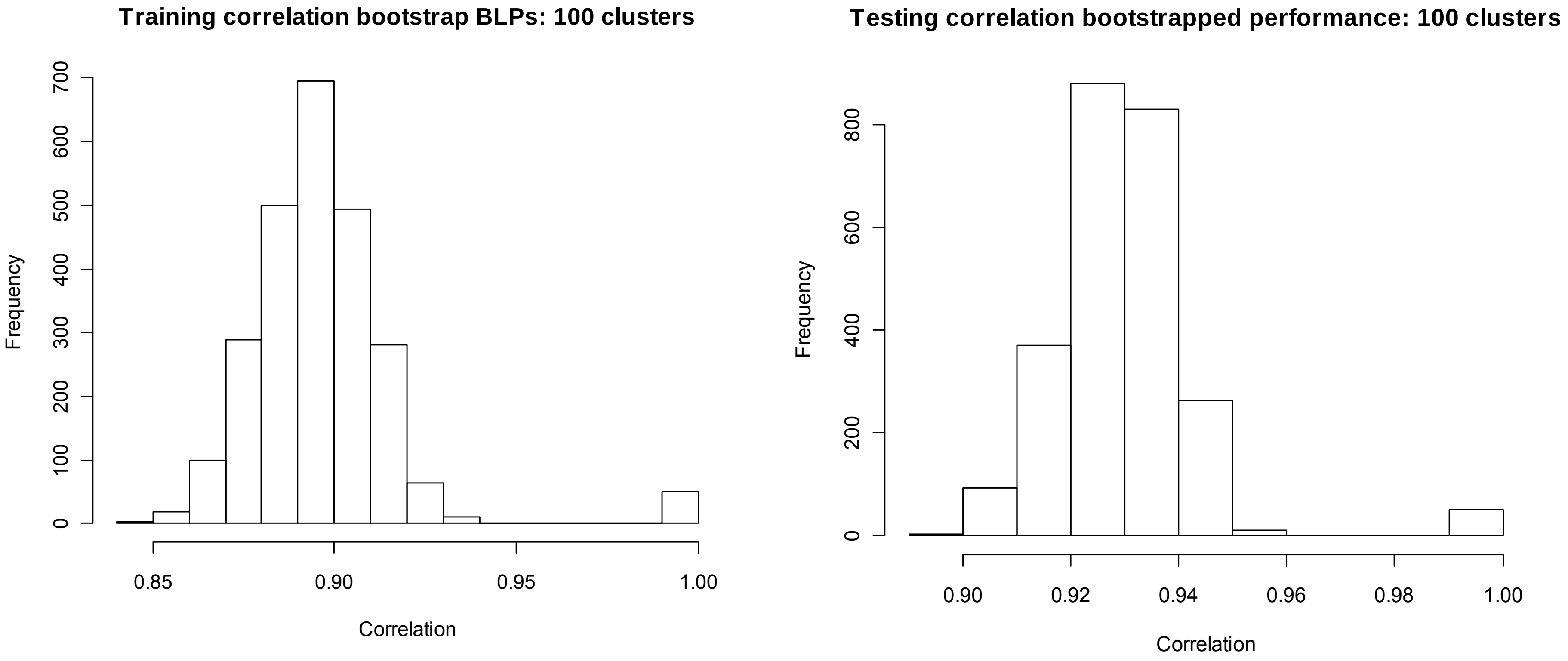
3. Calculate k-out statistics



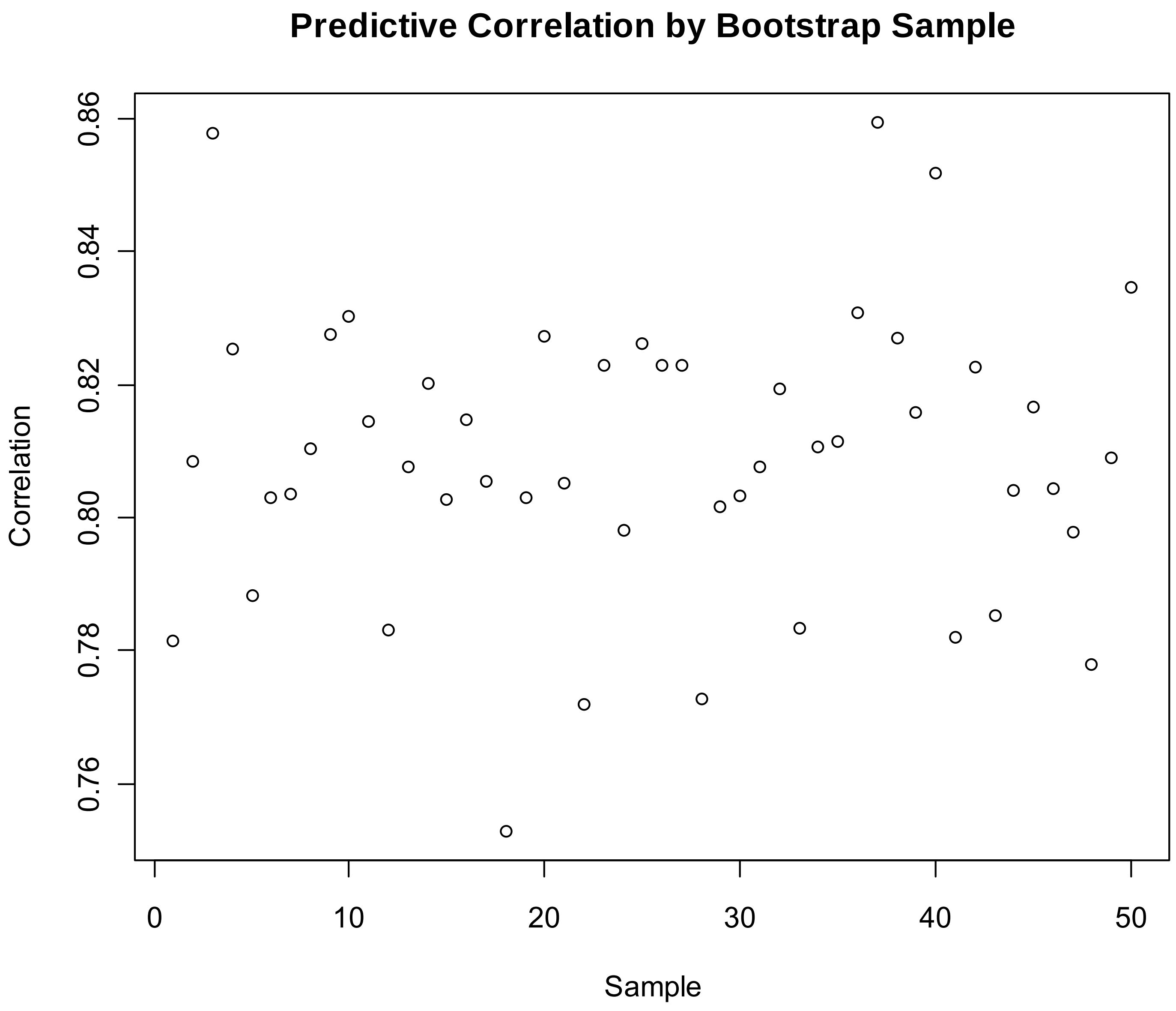


Things perhaps one should do

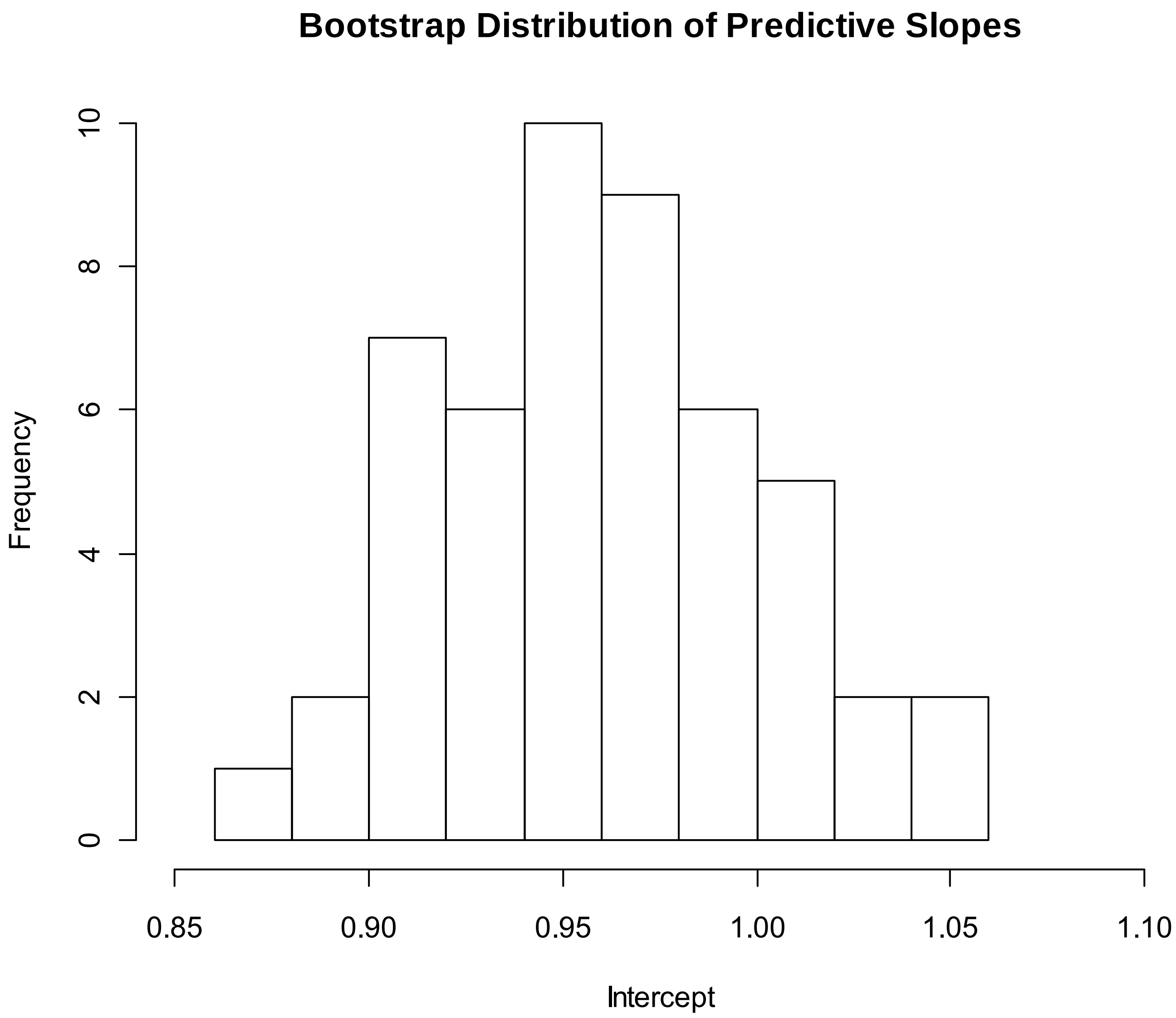
4. Bootstrap CV procedure (in K-folds K is low)



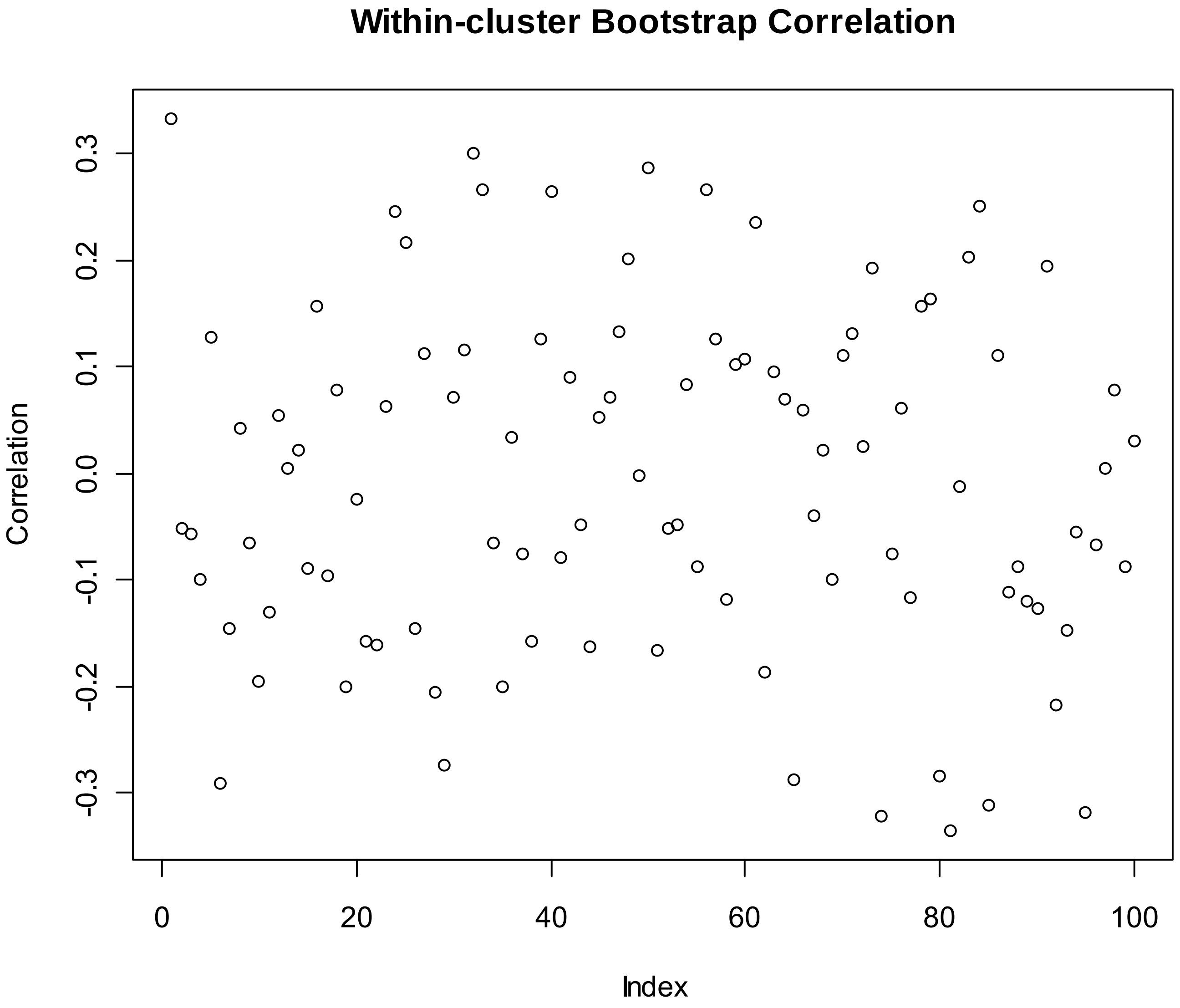
50 Bootstrap samples of each of 100 clusters



Predictive Correlation by Bootstrap Sample
Correlation
0.76
0.78
0.80
0.82
0.84
0.86
Sample
0
10
20
30
40
50



50 Bootstrap samples of each of 100 clusters



For each cluster, correlate bootstrapped BLP and bootstrapped future performance
B=50

Things perhaps one should do

5. Develop candidate-specific CV reliabilities
A procedure (have tried others…)

Two types of bootstrap correlations: among and within clusters

- For $i = bootstrap\ sample$ and $j = cluster$

$$z_{ij} = \frac{1}{2}[\log\frac{1+\rho_{i\cdot}}{1-\rho_{i\cdot}} + \log\frac{1+\rho_{\cdot j}}{1-\rho_{\cdot j}}]$$

- Estimate (ANOVA)

$$\lambda = \frac{Var\left(z_{ij}, \text{WITHIN CLUSTERS}\right)}{Var\left(z_{ij}, \text{AMONG CLUSTERS}\right)}$$

- Form

$$c_j = \bar{z}_{\cdot\cdot} + \frac{B}{B+\lambda}[\bar{z}_{\cdot j} - \bar{z}_{\cdot\cdot}]$$

- Invert

$$r_j = \frac{\exp(2c_j) - 1}{\exp(2c_j) + 1}$$

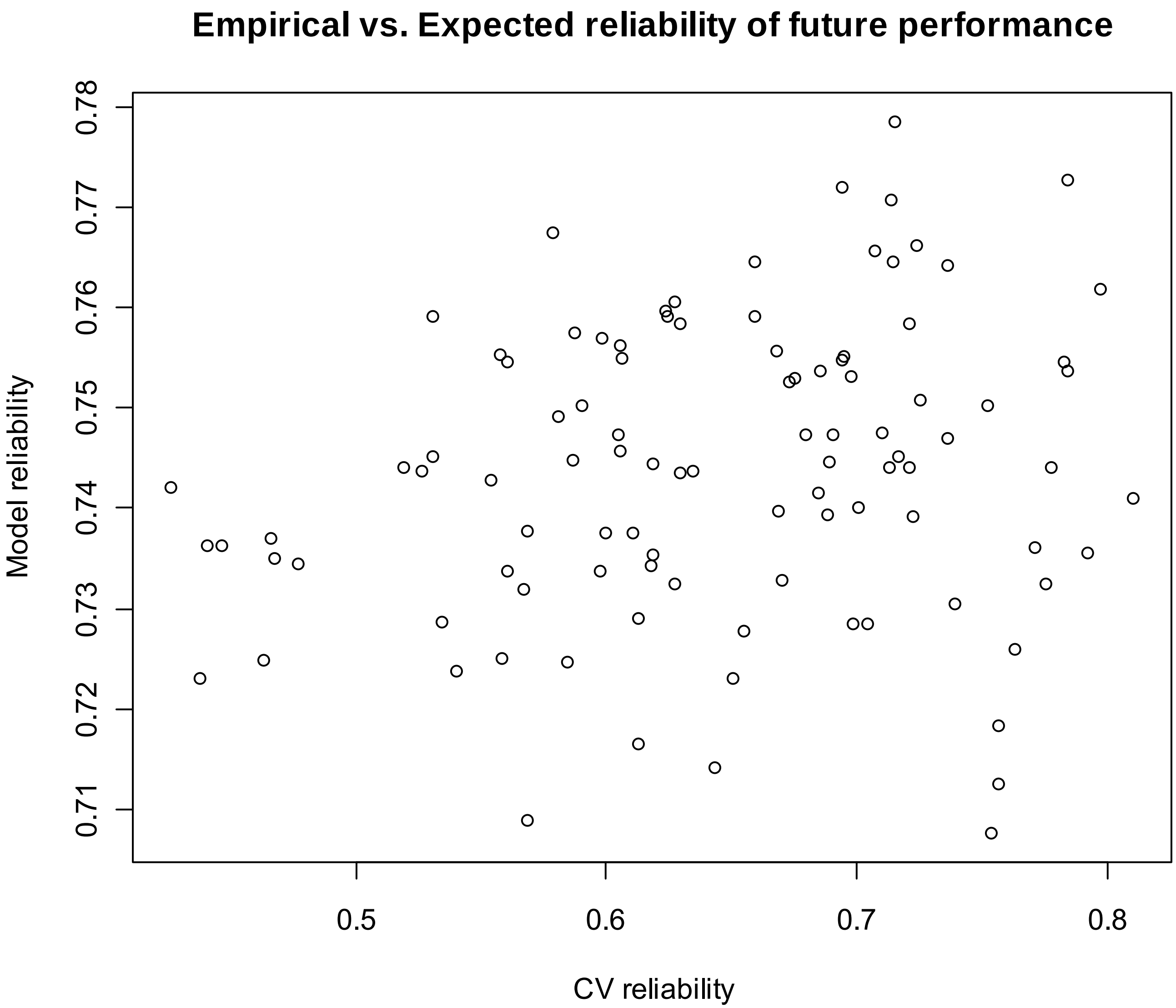
$$rel_j(CV) = r_j^2$$

THE LINEAR MODEL LEADS US TO EXPECT, FOR EACH CLUSTER

$$rel(BLP, future\ mean) = \frac{n_{TRN}}{n_{TRN+\frac{4-h^2}{h^2}}} \times \frac{n_{TST}}{n_{TST+\frac{4-h^2}{h^2}}}$$

$$\mathrm{logit}[rel(BLP, future\ mean)] = \log \frac{rel(BLP, future\ mean)}{1 - rel(BLP, future\ mean)}$$

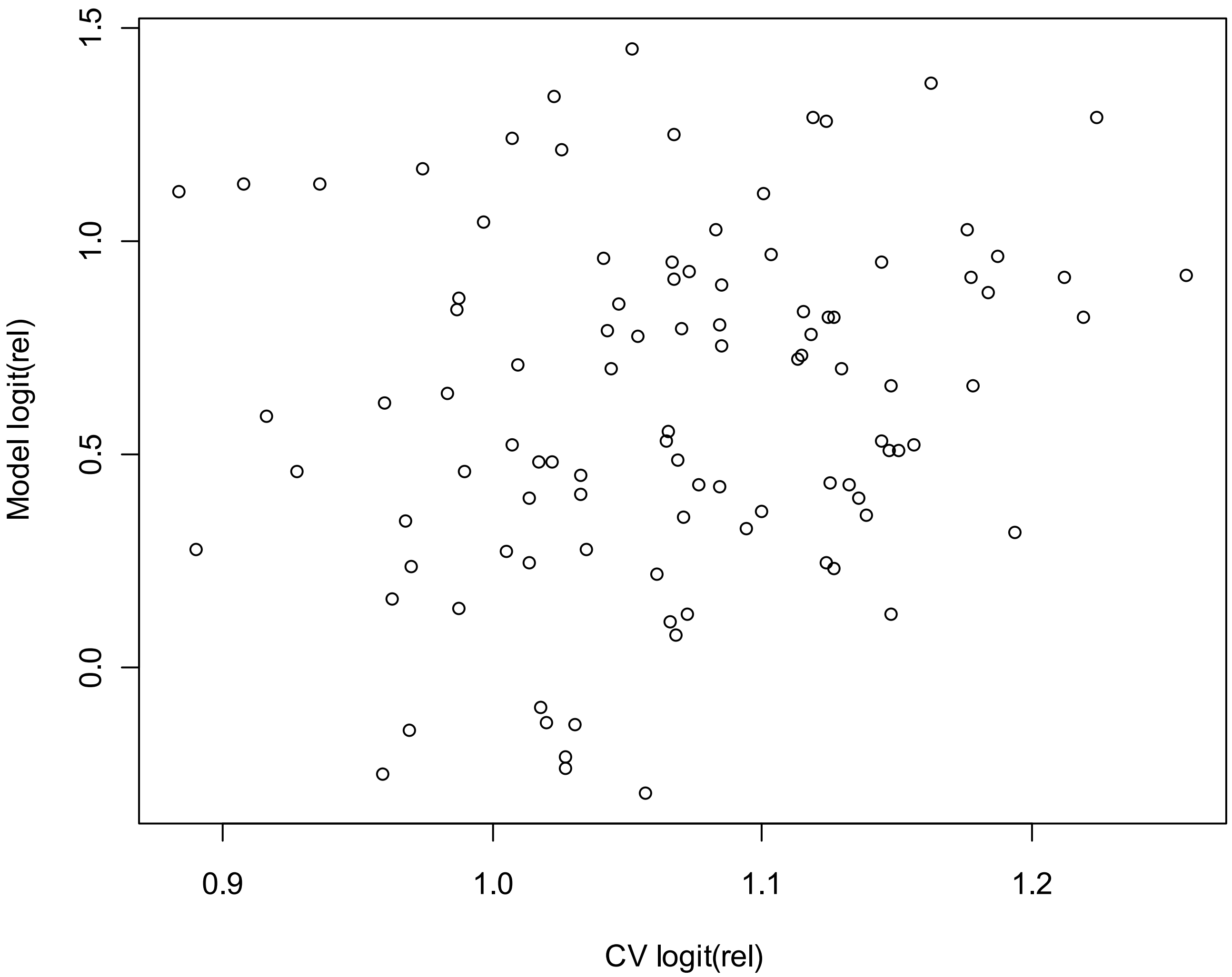
DO THE EMPIRICAL RELIABILITIES PREDICT THE EXPECTED RELIABILITIES WHEN THE MODEL IS TRUE?



Slope (Model|CV)= 0.033 P<0.04
B=50

Not convincing

Empirical vs. Expected logit(reliability of future performance)



Slope (Model|CV)= 0.038 P<0.05
B=50

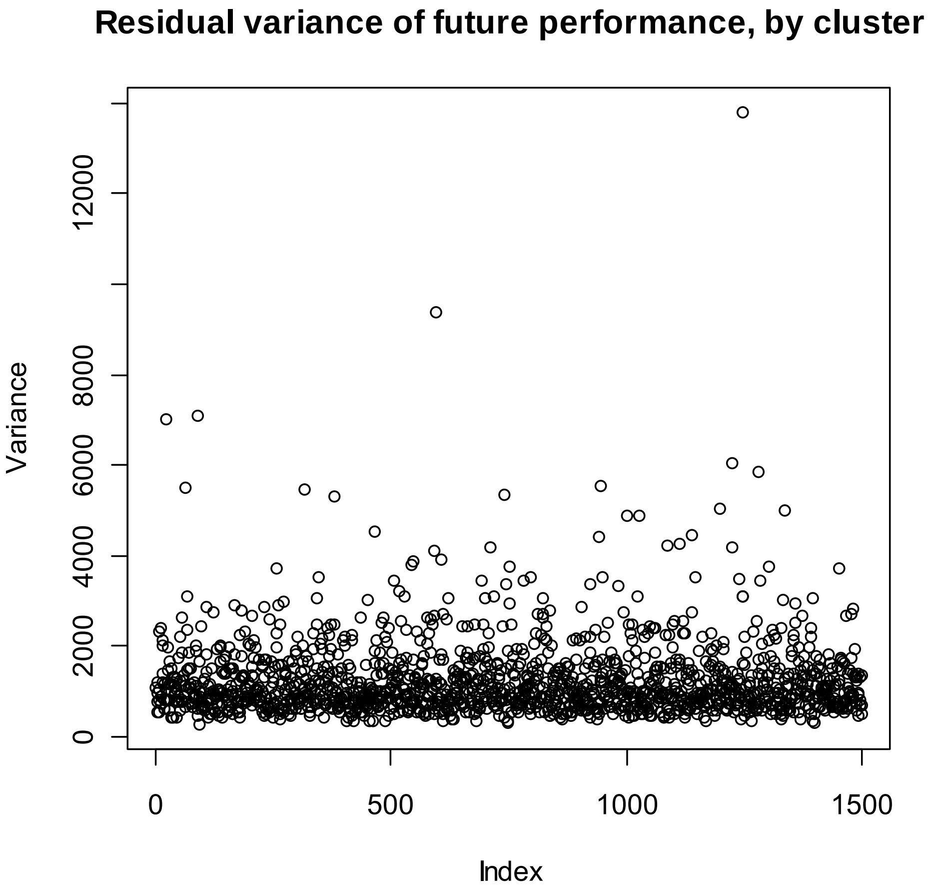
Not convincing

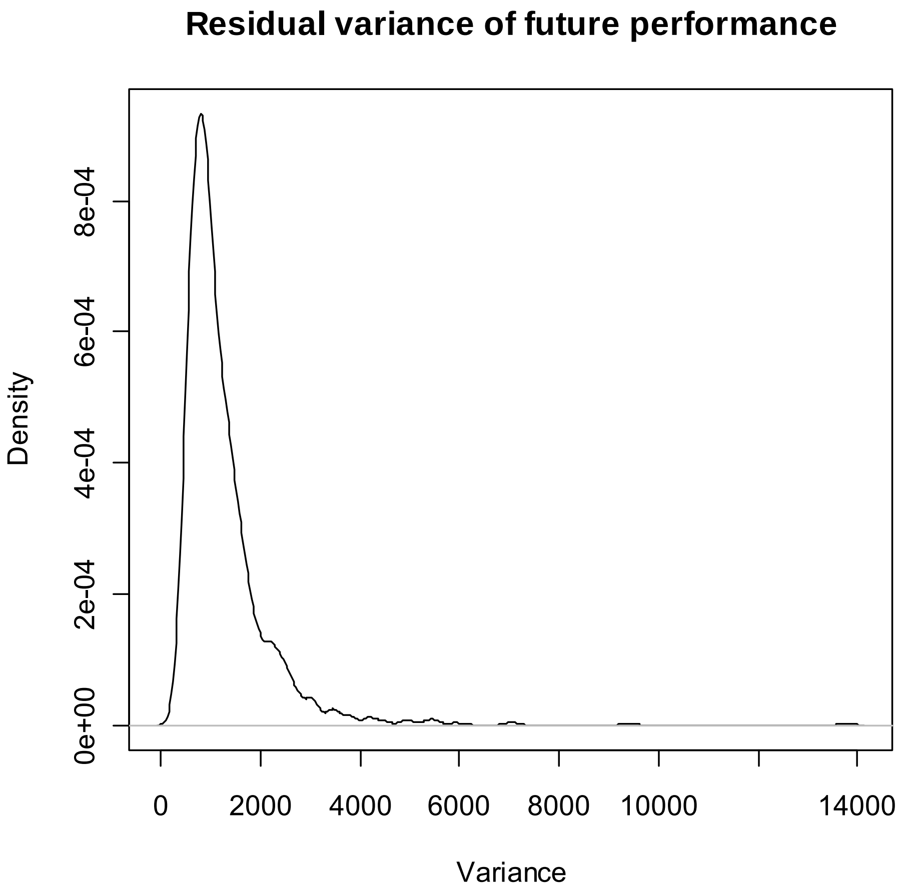
Creating havoc: model derived measures reliabilities ignore:

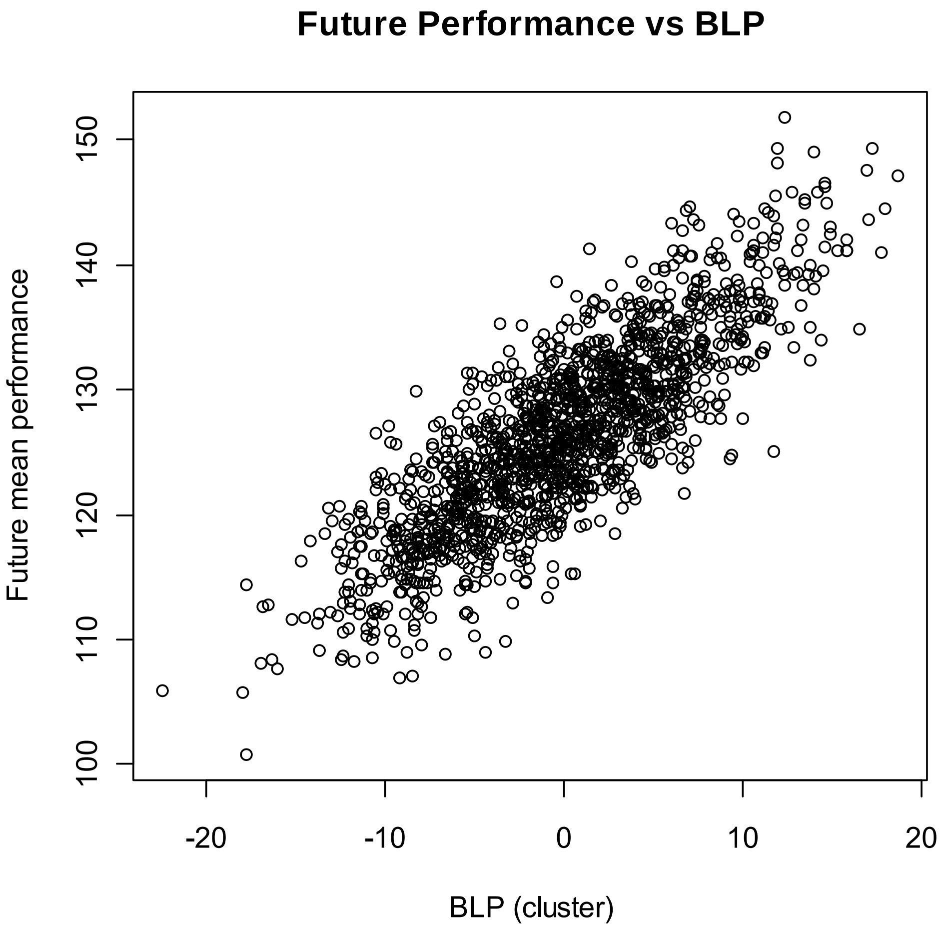
Heterogeneous variance
Selection
Preferential treatment

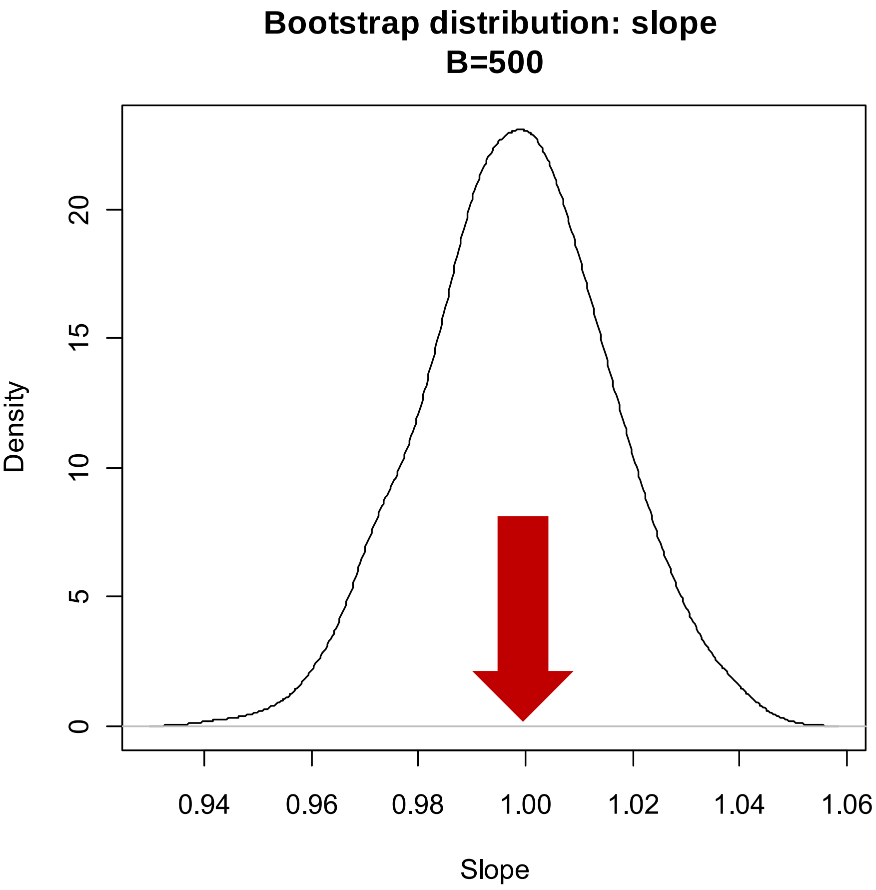
ALL THESE ARE ELEMENTS OF MODEL UNCERTAINTY

- 1500 CLUSTERS, h^2. TRAINING RESIDUAL VARIANCE= 950
- n_{TRN} ~Poisson(100), n_{TST}~Poisson(150)
- TESTING RESIDUAL VARIANCE VARIES AT RANDOM OVER CLUSTERS: χ^{-2} (scale=950, df=8)






Future Performance vs BLP
Future mean performance
100
110
120
130
140
150
-20
-10
0
10
20
BLP (cluster)


Bootstrap distribution: slope
B=500
Density
0
5
10
15
20
0.94
0.96
0.98
1.00
1.02
1.04
1.06
Slope

BUILDING SELECTION
(zero-inflated Poisson)

$$\lambda_{i,TST} = \lambda_{i,TRN} \exp\left(\frac{BLP(cluster\ i}{40}\right)$$

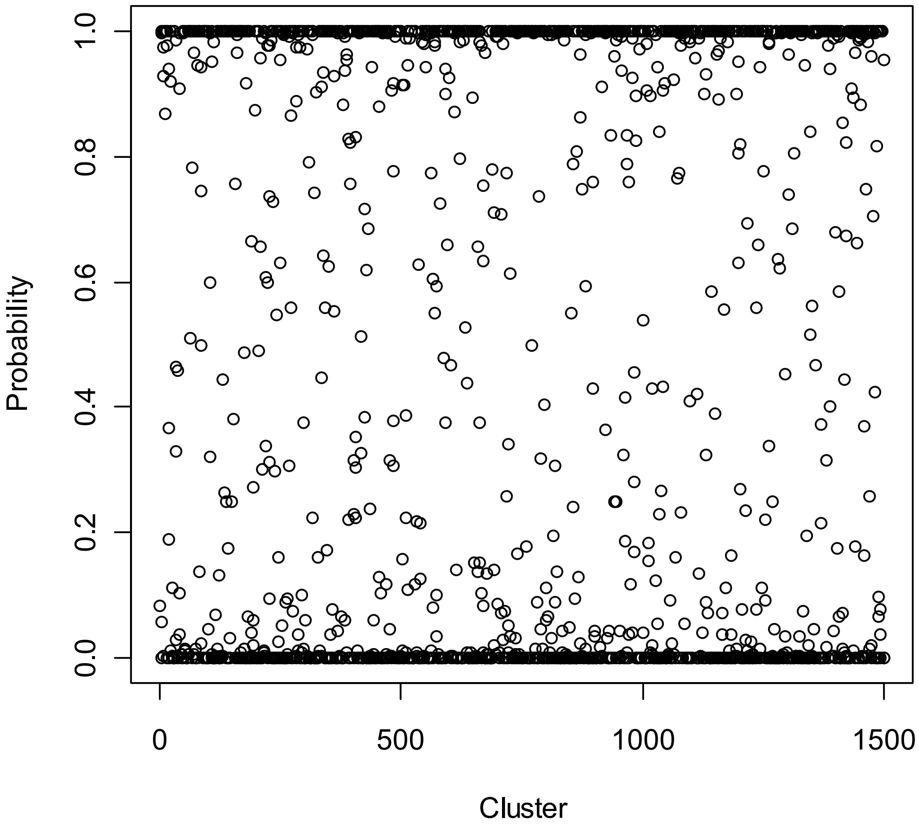
$$\text{Prob}(n_{i,TST} = 0|\text{Zero state}) = 1 - \Phi[BLP(cluster\ i)]$$

$$\delta_i = Bernoulli(1, \text{Prob}(n_{i,TST} = 0|\text{Zero state})$$

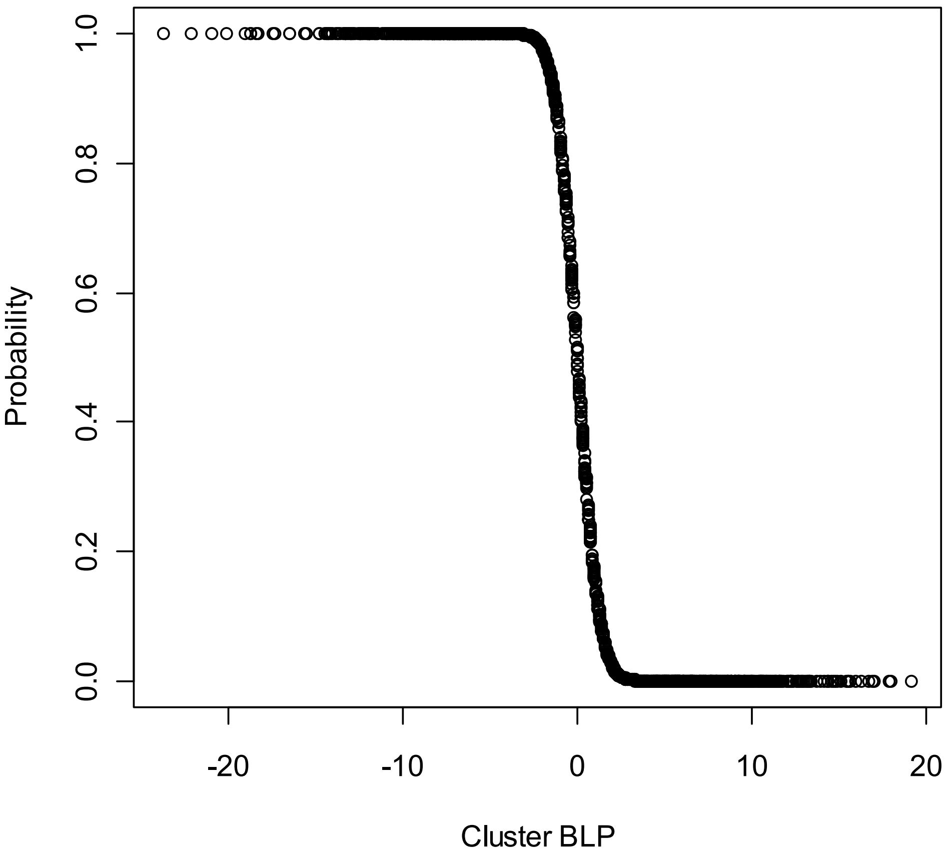
$$n_{i,TST} \sim 0 \times \delta_i + (1 - \delta_i) Poisson(\lambda_{i,TST})$$

<u>A “good” training BLP :</u>

- decreases probability of “structural” zero
- increases Poisson parameter of testing cluster size

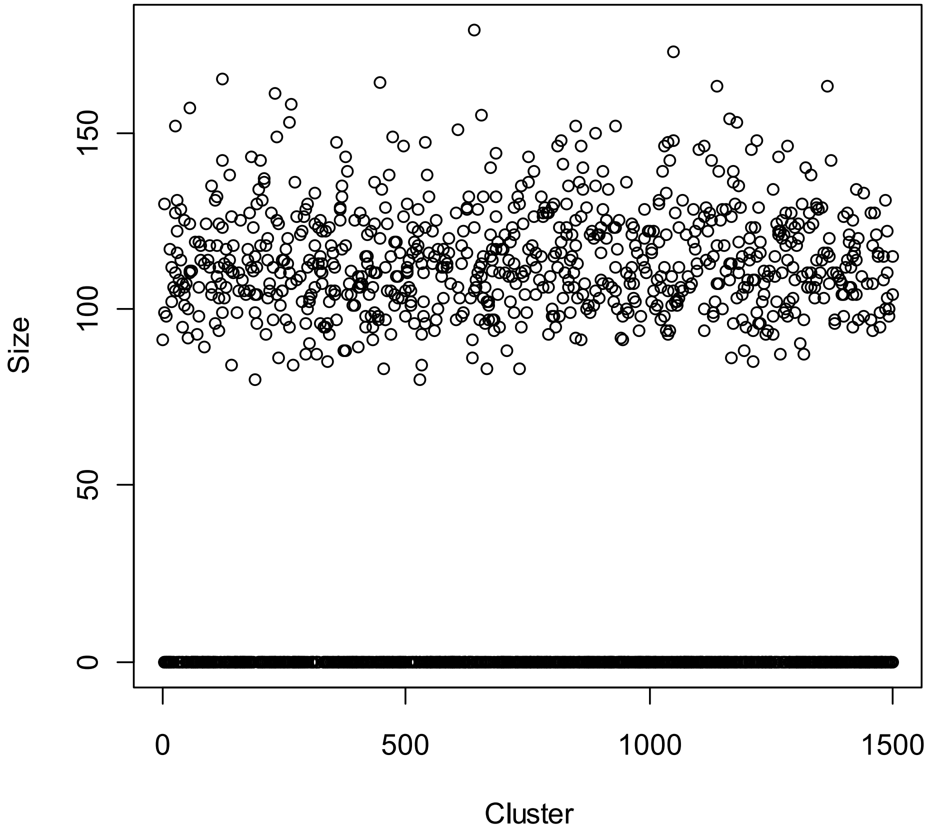


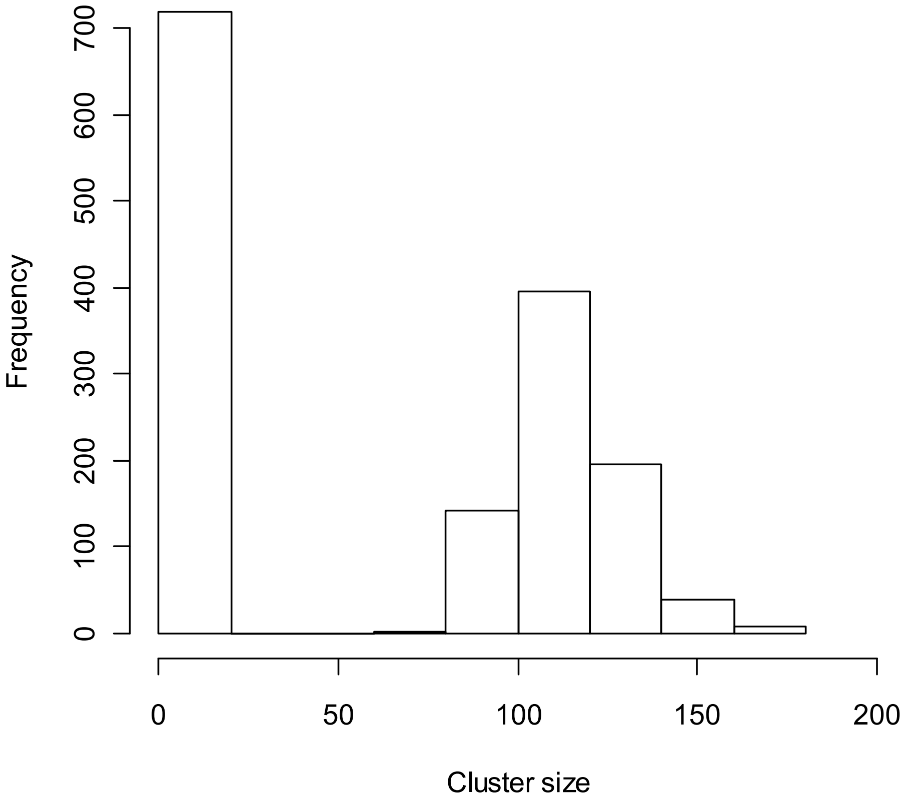
Perfect Zero probability, by cluster



Perfect zero probability vs BLP

Median probability= 0.37


Testing set cluster size

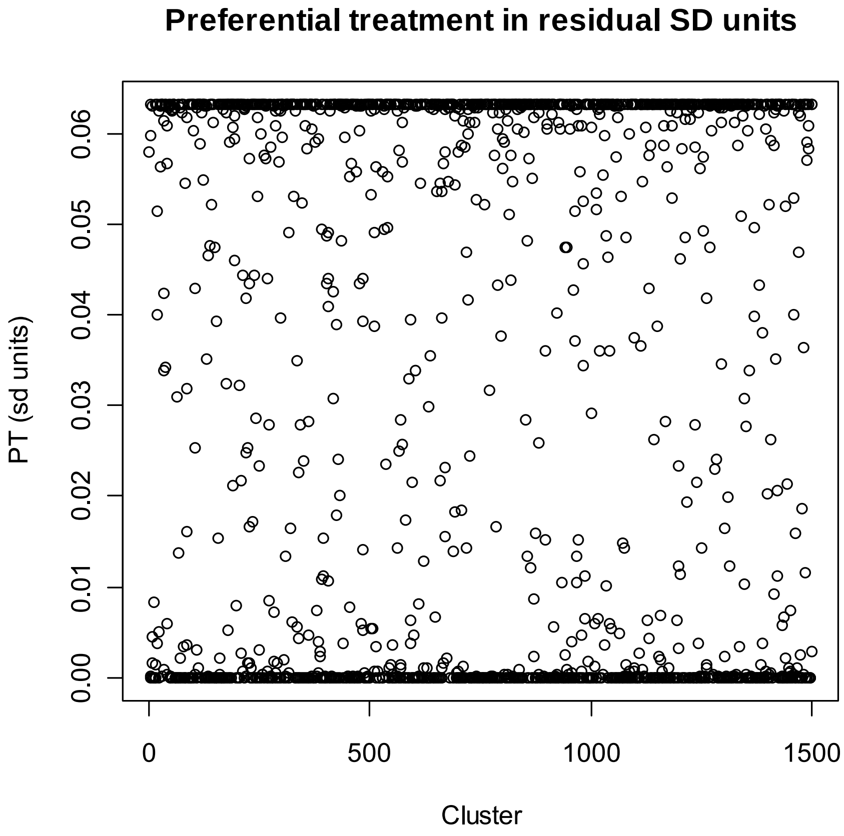

Testing set cluster size

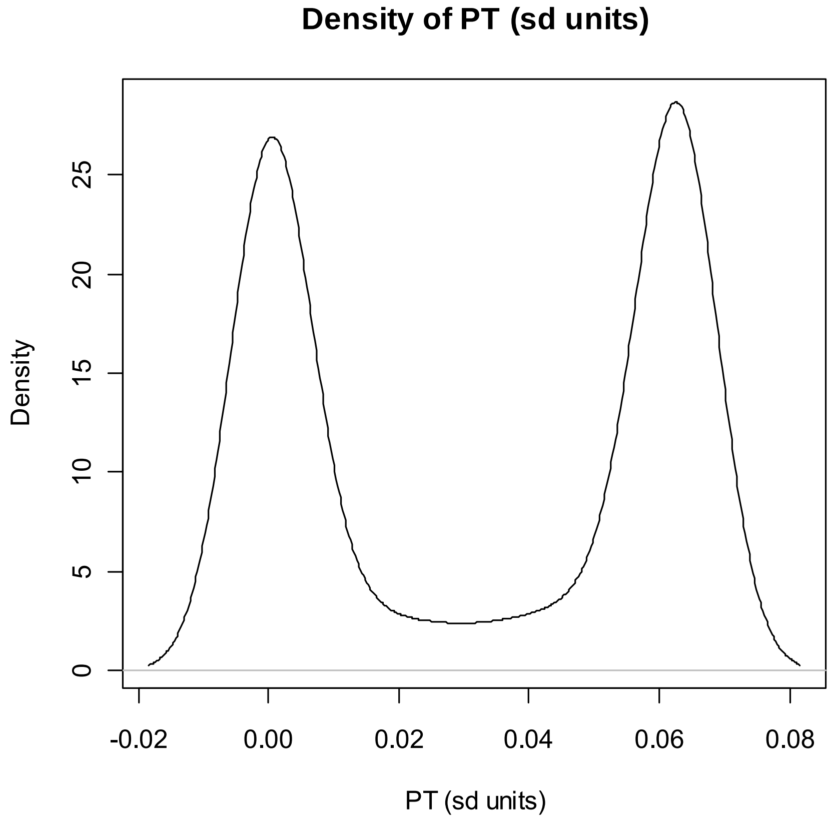
- 719 clusters do not have data in testing set
- Testing set cluster size varies between 2-179

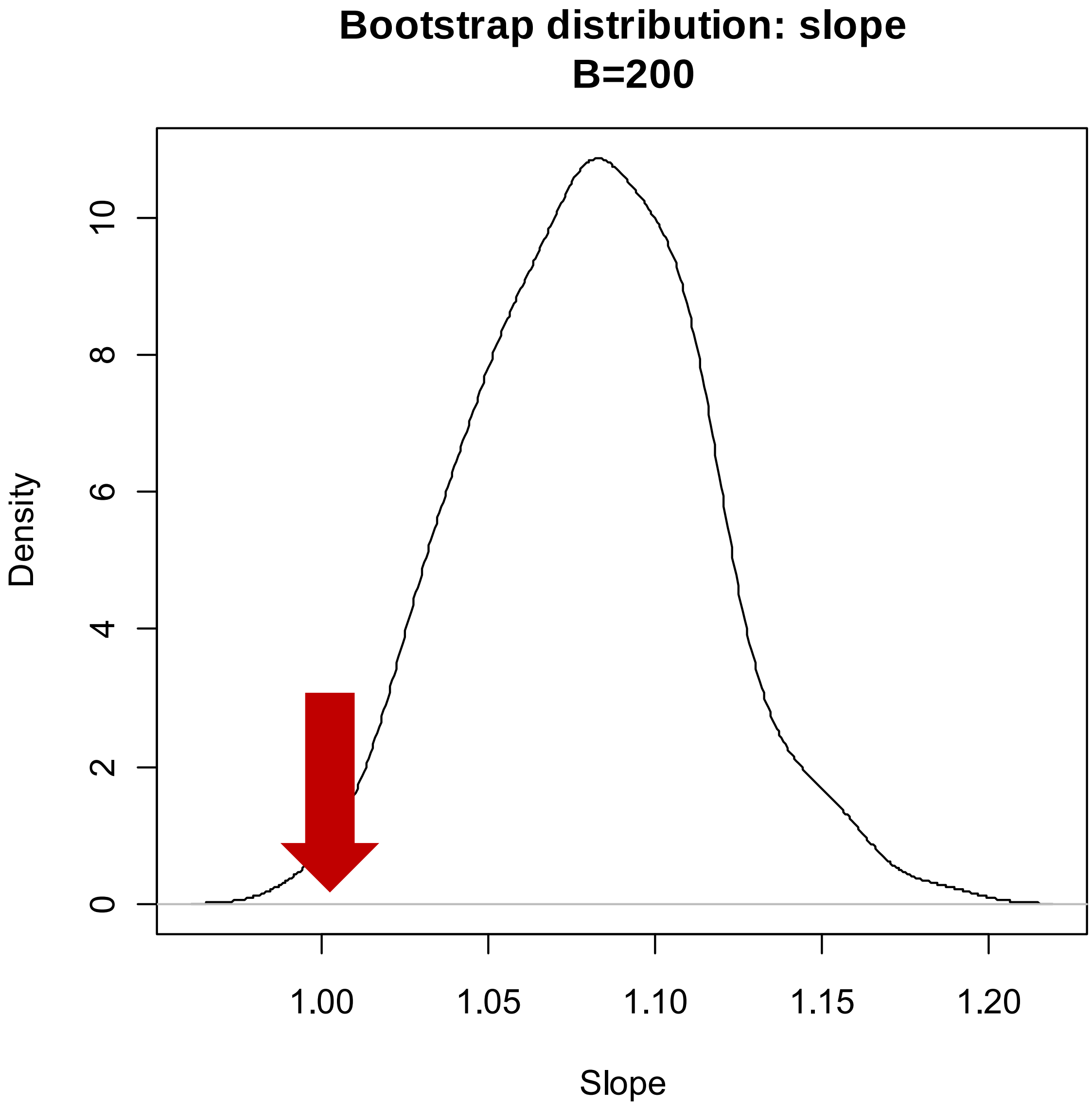
BUILDING PREFERENTIAL TREATMENT IN ADDITION TO SELECTION

$$\Delta = 0.1[\max(residual_{TST}) - \min(residual_{TST}]$$

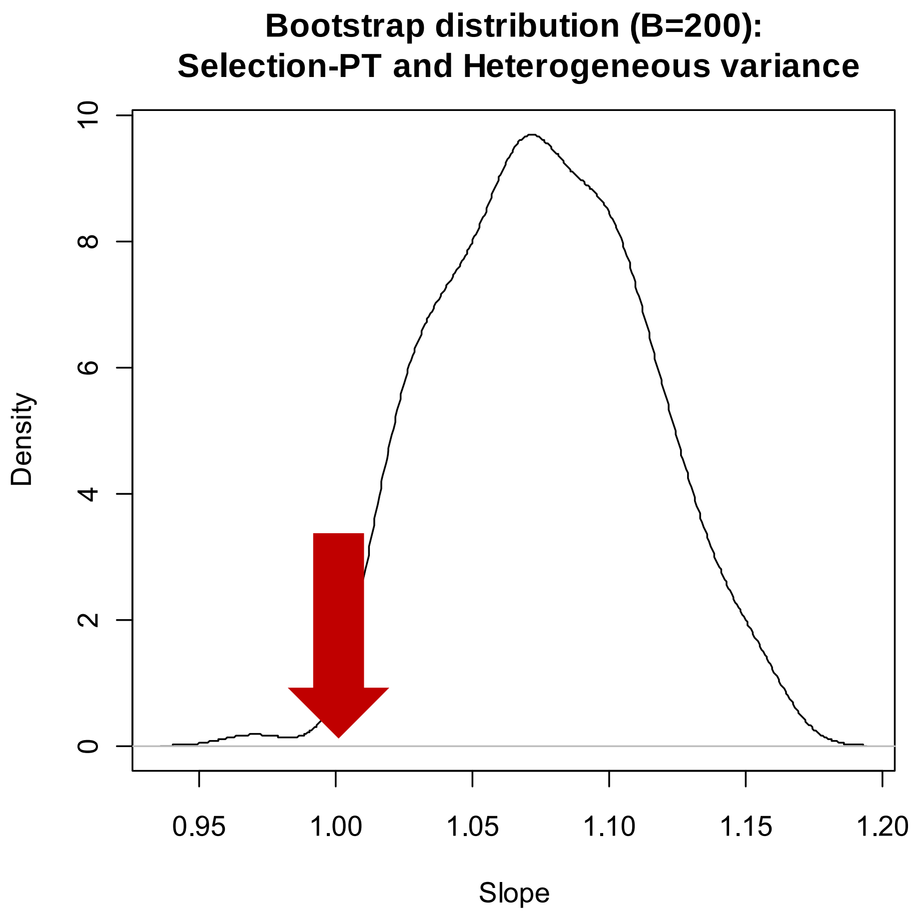
$$y_{ij,TST} = \mu + cluster_i + \Delta \times \Phi[BLP(cluster\ i)] + e_{ij,TST}$$



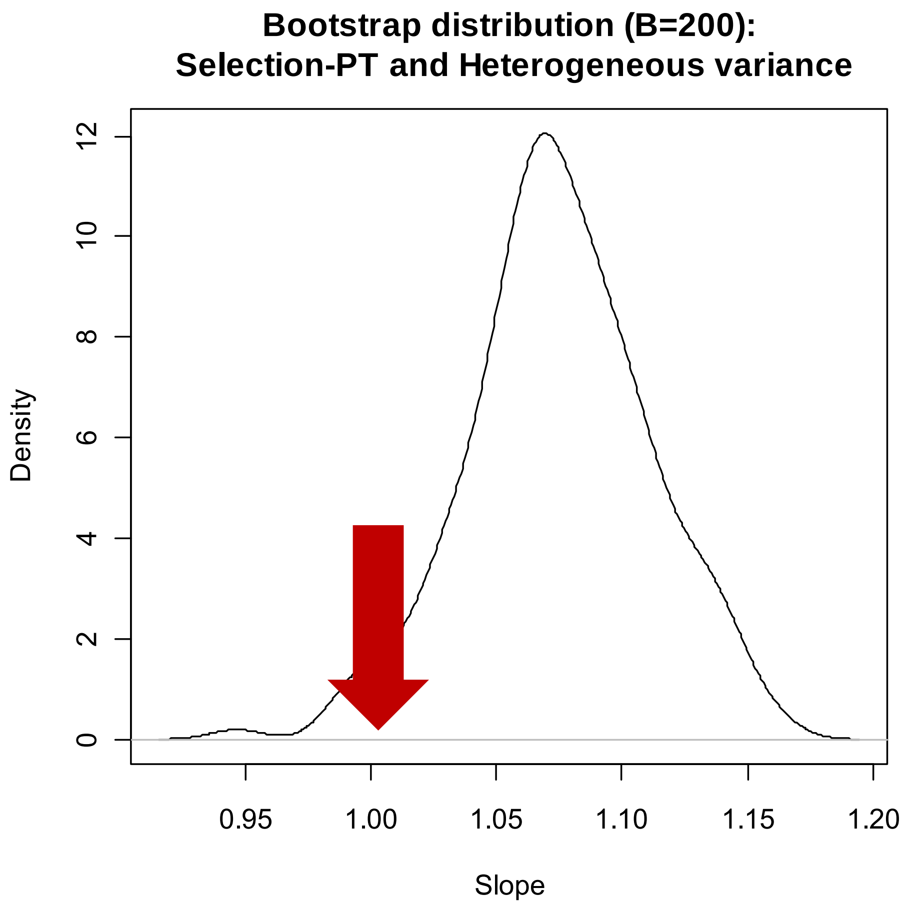




Median of bootstrap distribution= 1.08


Bootstrap distribution (B=200):
Selection-PT and Heterogeneous variance

Df=8 for scale inverted chi-square


Bootstrap distribution (B=200):
Selection-PT and Heterogeneous variance

Df=4 for scale inverted chi-square

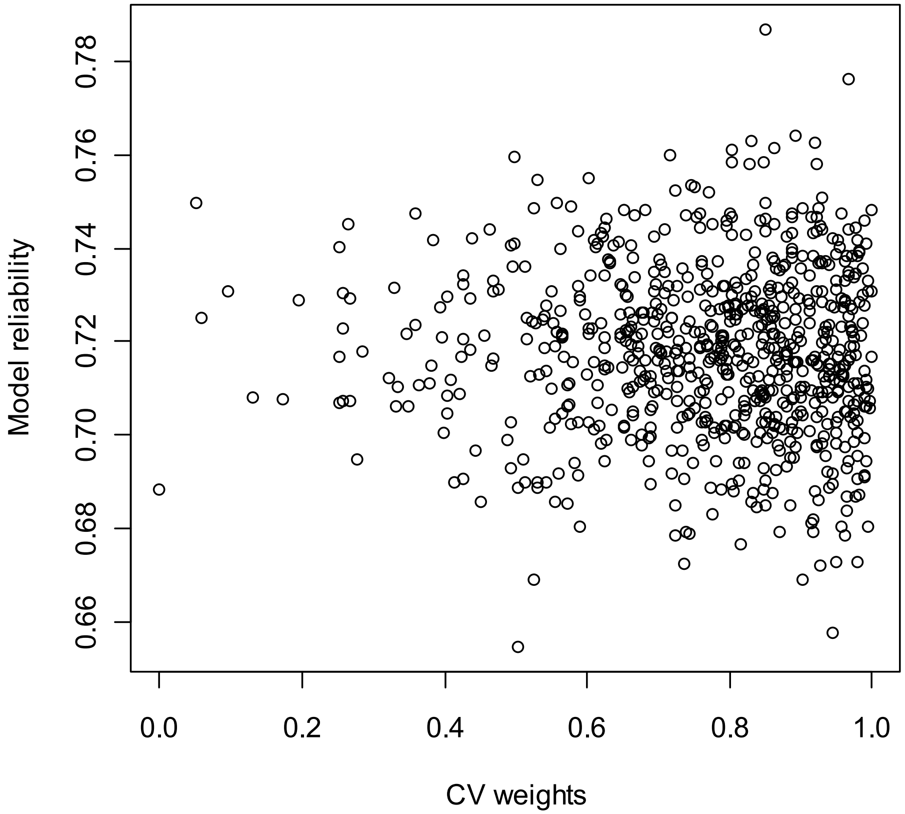
Median of bootstrap distribution= 1.08 in both cases

- These are just one of the few complications that one can think of
- How can we arrive at credible measures of reliability when one cannot
- even get good point estimates of the CV slope?
- Need to think about data driven measure of CV reliabilities instead of model driven quantities

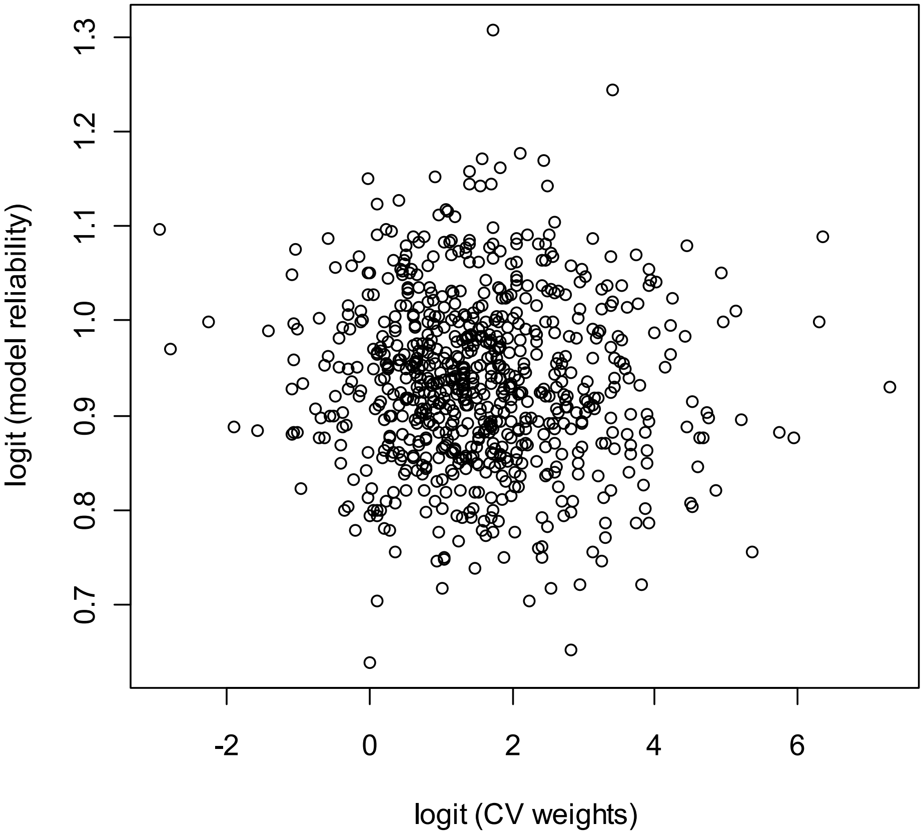
$$triweight_i = 1 - \frac{\left|residual_i\left(\text{CV regression}\right)\right|}{\max\left|residual_i\left(\text{CV regression}\right)\right|}$$

$$bisquare_i = 1 - \left(\frac{residual_i\left(\text{CV regression}\right)}{6 \times median\left[\left|residual_i\left(\text{CV regression}\right)\right|\right]}\right)^2$$

Relationship between CV weights and model derived reliability of future performance



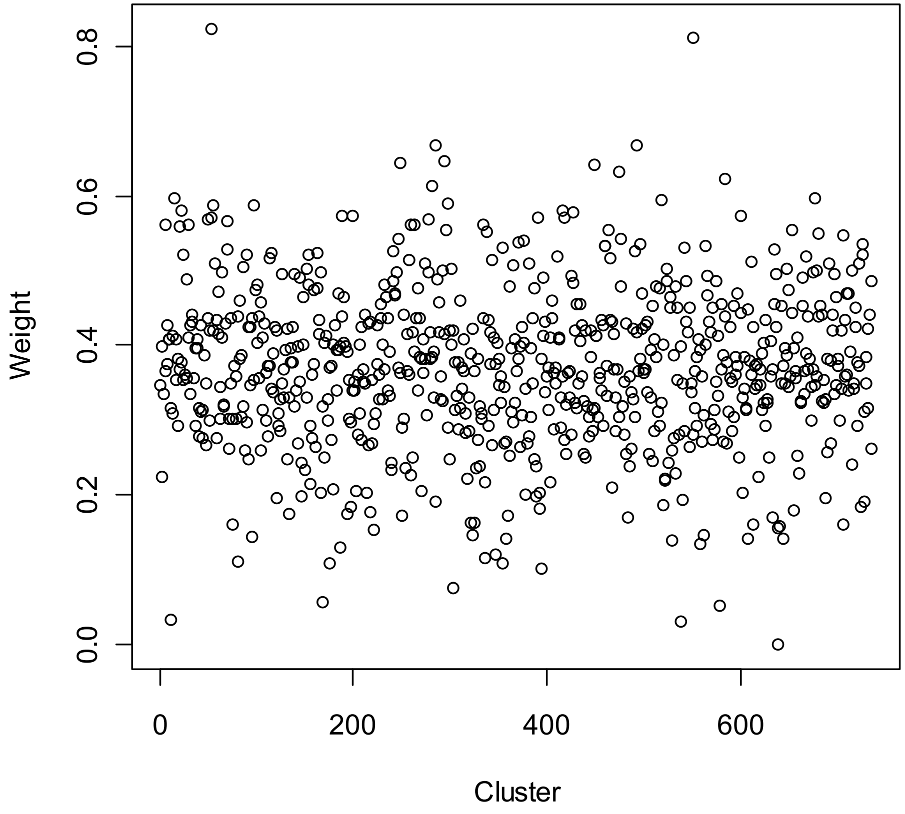
Relationship between CV logit (weights) and model derived logit (reliability) of future performance



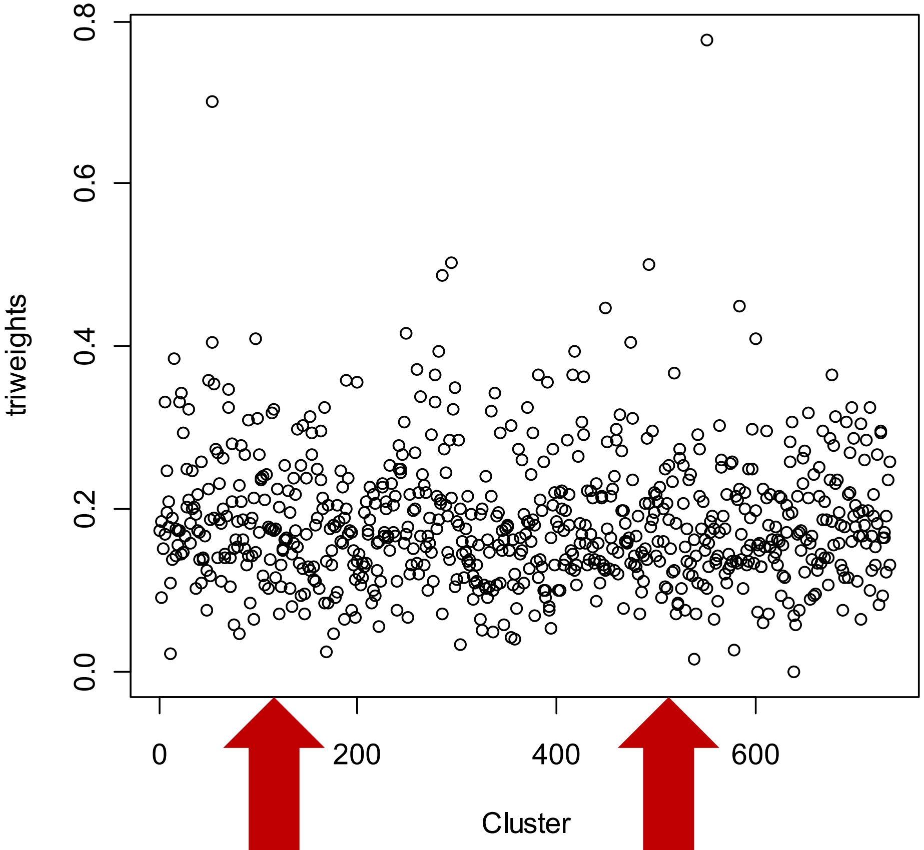
THERE IS NO RELATIONSHIP BETWEEN CV PERFORMANCE
(HIGHER WEIGHT➔ BETTER ALIGNMENT)
AND MODEL DERIVED RELIABILITY

BRING INTO THE PICTURE AMOUNT OF INFORMATION IN TRAINING AND TESTING SETS

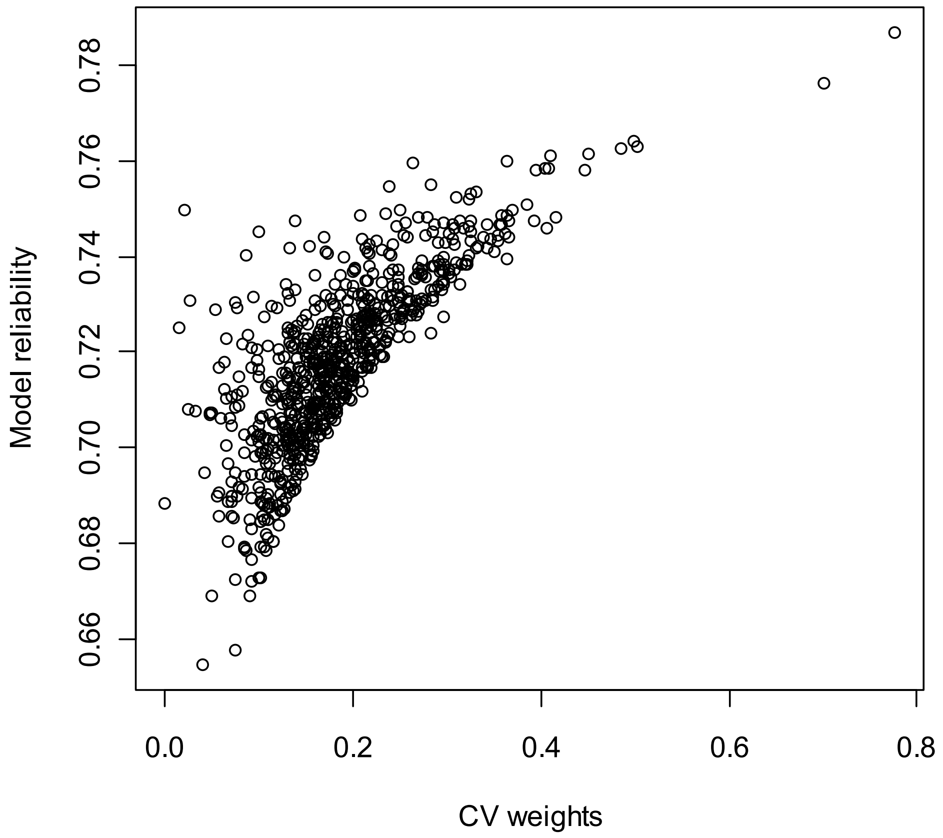
Triangular weights from CV regression

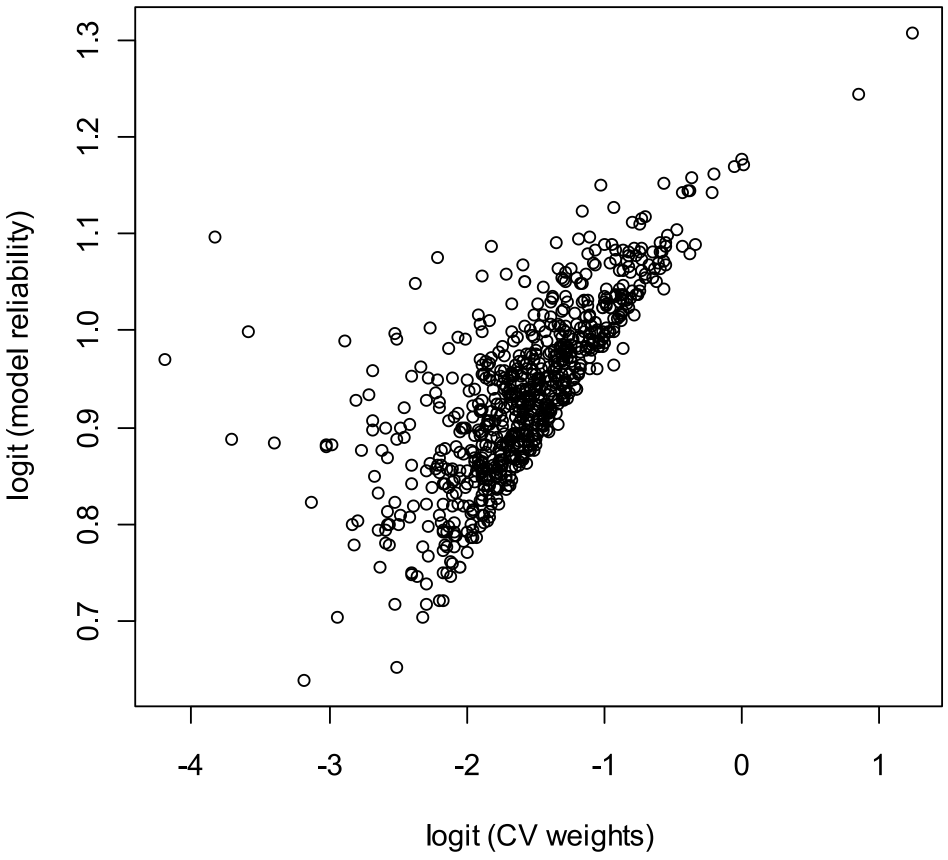


Triangular weights from CV regression: weighted by TRN and TST cluster size



$$w_{TRN} = n_{cluster,TRN} / \max(n_{cluster,TRN})$$
$$w_{TST} = n_{cluster,TST} / \max(n_{cluster,TST})$$


Relationship between CV weights and model derived reliability of future performance


Relationship between CV logit (weights) and model derived logit (reliability) of future performance

LIFE IS DIFFICULT…

- NOT IMMEDIATELY OBVIOUS HOW THESE RELATIONSHIPS GENERALIZE TO "ANIMAL" OR MORE GENERAL MODELS
- SAME HOLDS FOR GENOME-ENABLED PREDICTION
- CROSS-VALIDATION (SHOULD HAVE DONE THIS 40 YEARS AGO) GIVES AN OPPORTUNITY TO THINK "OUTSIDE OF THE BOX"
- SHOULD DO THIS MORE OFTEN (…*La vita è bella*)