Tips, tricks, and traps of Empirical Software Engineering

Emilia Mendes
University of Auckland, NZ
emilia@cs.auckland.ac.nz

Tim Menzies
West Virginia University, USA
tim@menzies.us

IEEE ASE'09, tutorial T3, Tuesday Nov.17, 2009

Objective

When you leave here today you will...

- Understand how to document ...
  - ... the value (or otherwise) of some proposed SE technique
- Better understand the strengths (and weaknesses) of the research reports you read in the literature

Target audience

Are you who we think you are?

- Someone who, in the near future, will be trying to:
  - Publish a paper
  - Review a paper
  - Assessing an empirical result in the SE literature.

Presenters

Are we who you think we are?

Mendes

- 10 years+ research into empirical SE
- Author:
  - Cost Estimation Techniques for Web Projects (2008)

- 120+ refereed papers; e.g.
  - ESE 2007; IEEE TSE 2007/8
- On numerous editorial boards

Menzies

- 20+ years research into empirical AI + SE
- Former NASA SE research chair
- Co-founder PROMISE conference
  - Predictive models in SE
- 170+ refereed papers e.g.
  - ESE 2009; IEEE TSE 2006/7
- On numerous editorial boards

Maybe, in the near future, we will review your papers.

Scope

What is covered

- Preliminaries
- Why
  - evidence-based SE
  - current (weak) state of the field
- How:
  - Do you know your ABCs
- Case studies....
- References

Case studies:

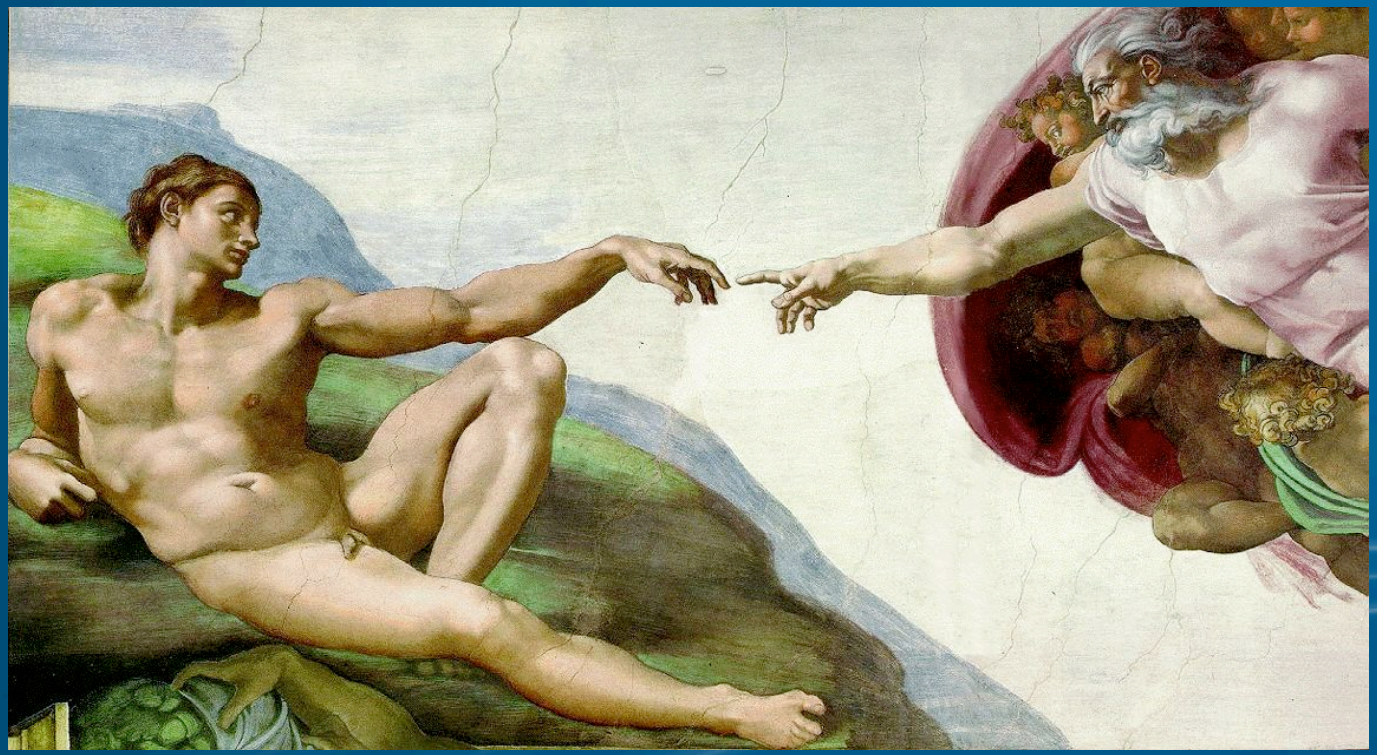
1. Ceiling + floor effects
2. Brittle conclusions
3. Conclusion instability
4. Beware the straight line
5. More maths

Out-of-scope

What is not covered here

- The early stages of experiments has been well documented by others:
  - Design of experiments and measurements
    - e.g. [Fenton96], [Easterbrook07]
- Lately, we are seeing are too many bad reports on those experiments
  - Researchers "dropping the ball" during "data analysis" stage
- Hence, this tutorial:
  - How not to fail in the end run.

and so,
it begins ...



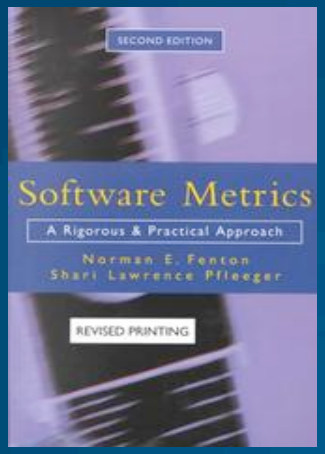
Roadmap

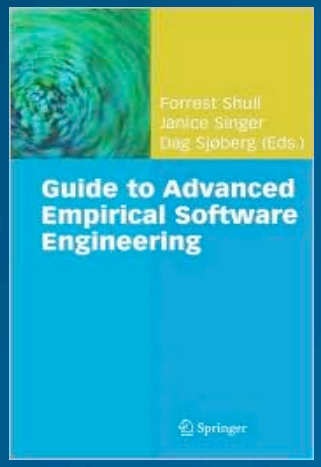
Where are we going now?

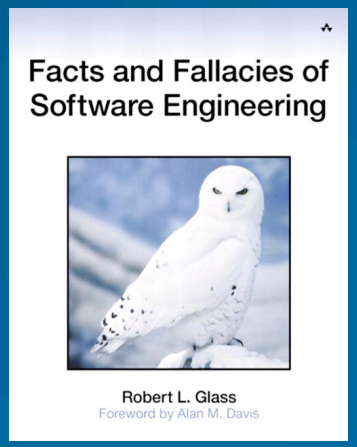
- **Preliminaries**
- Why
  - evidence-based SE
  - weak current state of the field
- How:
  - Do you know your ABCs?
- Case studies
- References

Empirical SE: state of the art

Too many "do-ings", not so much "learn-ings"







Strong on method, weak on conclusions from those methods:

- [Fenton96]: Software Metrics
- [Shull03]: Guide to Advanced Empirical SE

Strong on conclusions, not so strong on the justifications for those conclusions:

- [Endres03]: Handbook of Software and Systems Engineering
- [Glass02]: Facts and Fallacies of Software Engineering

These are important books that deserve your close study.

But they reflect the current state of the art

- So much analysis
- So few conclusions

Weak state of empirical SE

[Menzies08a]: most cost/effective software QA methods?

- Literature describes (about) 100 techniques for independent software V&V
- Very little relative cost-benefit assessment
  - Usually, reports that "I did X and it was ok".
  - Not ”X > Y”
- Very few generalizations over multiple methods or multiple projects. Sadly, these typically use:
  - Just a few projects
  - or "Delphi analysis" i.e. ask a few "experts"
  - or extensive data analysis, which they can't show us (confidentiality)

Empirical SE: state of the art

Progress towards general results.... slow

[Zannier06]:
- 5% * 1300 ICSE articles
- 45 "empirical papers"
  - 44 only review own methods

[Fenton07]:
- ".. much of the current software metrics research is inherently irrelevant to the industrial mix..."
- "... any software metrics program that depends on extensive metrics collection is doomed to failure ..."

[KitchenhamMendes07]:
- Model learned there not good here (for effort estimation)

[Zimmerman09]:
- ditto (for defect estimation)

[GreenMenzies09]:
- Conclusions from relative cost/ benefit analysis dependent on local value function.

[Basili09]:
- Still far to go
- But we should celebrate progress over last 30 years.
- And we are turning the corner

[Budgen09]:
- Empirical SE results
  - too immature for making policy
- Need for better reporting:
  - Systematic reviews
  - Structured abstracts

Evidence-based SE

Gather and Analyse Empirical Data Systematically

"The essence of the evidence-based paradigm is that of:

•systematically collecting and analyzing all of the available empirical data about a given phenomenon
• in order to obtain a much wider and more complete perspective than would be obtained from an individual study,
•not least because each study takes place within a particular context and involves a specific set of participants.

While these factors can bias the outcomes from one study,
•taking a wider view should make it possible to produce more reliable conclusions and to minimize the effects of local factors."

EBSE website
(http://www.dur.ac.uk/ebse/home_researchers.php)

Evidence-based SE

Gather and Analyse Empirical Data Systematically

"The core tool of the evidence-based paradigm is the Systematic Literature Review (SLR). "

- EBSE website (http://www.dur.ac.uk/ebse/home_researchers.php)

A systematic review is

- a method that enables the evaluation and interpretation of all accessible research relevant to a research question, subject matter, or event of interest (Kitchenham, 2004).

There are numerous motivations for carrying out a systematic literature review, amongst which the most common are (Kitchenham, 2004):

- To review the existing evidence regarding a treatment of technology, for example, to review existing empirical evidence of the benefits and limitations of a specific Web development method.
- To identify gaps in the existing research that will lead to topics for further investigation.
- To provide a context/framework so as to properly place new research activities.

Evidence-based SE

Gather and Analyse Empirical Data Systematically

A Systematic review generally comprises the following steps (Kitchenham, 2004):

- Formulation of a focused review question;
- Identification of the need for carrying out a systematic review;
- A comprehensive, exhaustive search and inclusion of primary studies;
- Quality assessment of included studies;
- Data extraction;
- Summary and synthesis of study results (meta-analysis);
- Interpretation of the results to determine their applicability;
- Report-writing.

But the good news is...

The field is ripe for improvement

- So, what are you writing right now?
  - And when you submit it, will it get accepted?

“Where facts are few, experts are many."

-- Donald R. Gannon

Roadmap

Where are we going next?

- **Preliminaries**
- Why
  - evidence-based SE
  - poor current state of the field
- How:
  - Do you know your ABCs?
  - Case studies
- References

Do you know your A,B,Cs ?

Estimating discrete class prediction using (A,B,C,D)

(historical logs)

	Really absent	Really there
Detector is silent	A (true negative)	B
Detector is triggered	C	D (true positive)

accuracy = (A+D)/(A+B+C=D)
precision = D / (C+D)
recall = pd = prob(detect)
= D / (B+D)
pf = prob(false alarm)
= C / (A+C)
pos/neg = (B+D) / (A+C)

If more than two classes:

- Need one table for each class X:
  - there=(class is X);
  - absent=not(class is X)

<-- classified as

happy	sad	tired	
100	3	2	happy
0	3	2	sad
80	50	120	tired

	sad= false	sad= true
not sad!	302	2
sad!	53	3

	happy= false	happy = true
not happy!	175	5
	80	100

	tired= false	tired= true
not tired!	106	130
tired!	4	120

	acc	pd	pf	prec
sad	85%	60%	15%	5%
happy	76%	95%	45%	56%
tired	63%	48%	4%	97%

Trade offs

There is no such thing as a free lunch

accuracy = (A+D)/(A+B+C=D)
precision = D / (C+D)
recall = pd = prob(detect) = D / (B+D)
pf = prob(false alarm) = C / (A+C)
pos/neg = (B+D) / (A+C)

$$pf = \frac{pos}{neg} \cdot \frac{(1 - prec)}{prec} \cdot recall$$

If target class rare pos/neg very small and

- high(accuracy, pd) does not mean high(pd,accuracy)
- high precision requires vanishingly small pf (v. hard to do)

If more than two classes, accuracy misleading

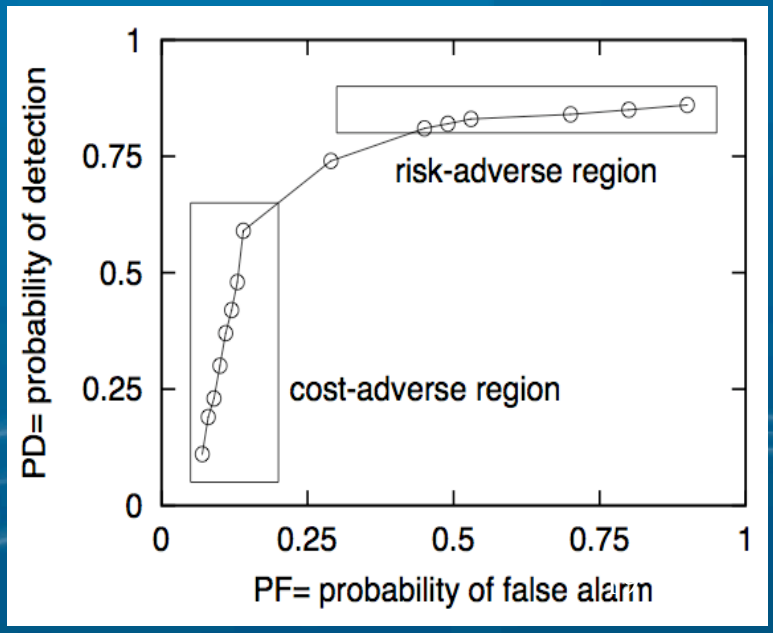
- Need to report separate pd,pf,prec for each class

Characteristic shape (see right) of pd vs (pf,prec) curves

- at pd = 0, no mistakes, no detection
- at pd = max, catch everything, and then some

AUC = area under curve : common performance measure.
"Best" detector is a domain-specific solution

- if risk-adverse, favor high pd
- if cost-adverse, favor low pf



Means vs Medians vs Quartiles

Surely, you can't "mean" that.

Bill Gates + 9 homeless people:

- Mean income = $5 billion
- This number characterizes the social standing of NO ONE in that sample.

Median = "the middle number"

- If an even number of observations, then average of the 2 middle numbers.

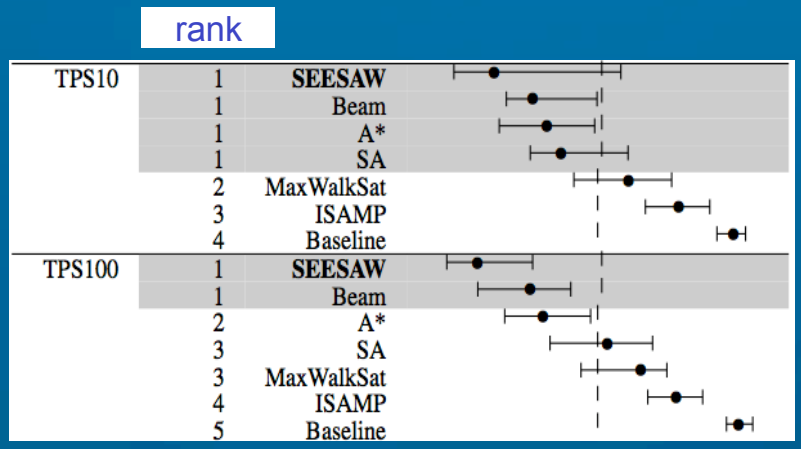
More generally

- Median = 50th percentile
  = second "quartile

Quartile charts:

- Show quartiles:
- Millions of observations can be displayed in a very small space.
- No need to descend into the quagmire of deciding whether or not to use
  - AR, RE, MRE, MMRE,...

- Dot = median
  - "Wings" = 2nd, 3rd quartile
- Sort treatments by median
- Add a "rank" column
  - row "i" has same rank as previous rows if they are statistically insignificantly different.



In the above, the first (four,two) treatments on (TPS10,TPS100) are statistically insignificantly different:

- Note the asymmetry in the distributions
  - "Means" not informative.

“Not everything that can be counted counts, and not everything that counts can be counted."

- Albert Einstein

Scope

What is covered

- Preliminaries
- Why
  - evidence-based SE
  - current (weak) state of the field
- How:
  - Do you know your ABCs?
- Case studies....
- References

Case studies:

1. Ceiling + floor effects
2. Brittle conclusions
3. Conclusion instability
4. Beware the straight line
5. More maths

Case studies

What can go wrong

And now...

- we come to the heart of the matter.

1. Always test for ceiling and floor effects

Check that your do better than some dumber alternative

Ceiling and Floor effects

Always check the roof

What does "83%" mean?

- Is that a "good" score?

But if everything scores 81 to 85?

- Then "83%" looks a little dull

Given some new better, more sophisticated method

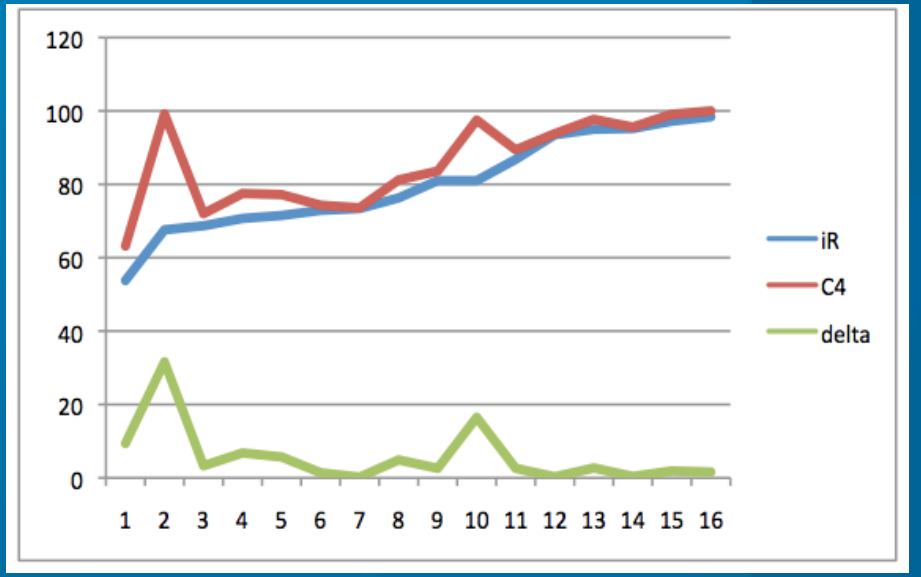
- good practice to compare against a seemingly dumber thing

- Floor effect:
  - Some inherent lower bound on performance in a domain
- Similary, ceiling effect:
  - Some inherent upper bound

E.g. ceiling and floor effects

[Holte85] : the "ONER" experiment

Decision tree learners build trees
- "N" levels deep
- Each node is one test on one attribute
- Each leaf is a class



Are they too clever?
- What if we restrict "N" to "one"?
- OneR: learns trees one level deep.

On data sets with two classes, performs nearly as well as standard tree learners.
- but when classes > 2, more complex learners are more valuable

Why would you use such a learner?
- Fast: may work when other methods crash
- Defines a "floor" effect: you've got to better than OneR
- Also useful as a fast feature subset selector
  - Determining what to prune before using a more elaborate learner

E.g. ceiling and floor effects

[Domingos97]: 6 learners on 28 data sets

Within one data set, very similar results for all learners :

- Performance determined by data, not algorithm

What is a "good" result depends on ceiling effects in the data:

- Scoring "50" on primary tumor is a great score.
- Scoring 100% on Soybean is a boring score.

Table 1. Classification accuracies and sample standard deviations, averaged over 20 random training/test splits. "Bayes" is the Bayesian classifier with discretization and "Gauss" is the Bayesian classifier with Gaussian distributions. Superscripts denote confidence levels for the difference in accuracy between the Bayesian classifier and the corresponding algorithm, using a one-tailed paired t test: 1 is 99.5%, 2 is 99%, 3 is 97.5%, 4 is 95%, 5 is 90%, and 6 is below 90%.

Data Set	Bayes	Gauss	C4.5	PEBLS	CN2	Def.
Audiology	73.0±6.1	73.0±6.1^{6}	72.5±5.8^{6}	75.8±5.4^{3}	71.0±5.1^{5}	21.3
Annealing	95.3±1.2	84.3±3.8^{1}	90.5±2.2^{1}	98.8±0.8^{1}	81.2±5.4^{1}	76.4
Breast cancer	71.6±4.7	71.3±4.3^{6}	70.1±6.8^{5}	65.6±4.7^{1}	67.9±7.1^{1}	67.6
Credit	84.5±1.8	78.9±2.5^{1}	85.9±2.1^{3}	82.2±1.9^{1}	82.0±2.2^{1}	57.4
Chess endgames	88.0±1.4	88.0±1.4^{6}	99.2±0.1^{1}	96.9±0.7^{1}	98.1±1.0^{1}	52.0
Diabetes	74.5±2.4	75.2±2.1^{6}	73.5±3.4^{5}	71.1±2.4^{1}	73.8±2.7^{6}	66.0
Echocardiogram	69.1±5.4	73.4±4.9^{1}	64.7±6.3^{1}	61.7±6.4^{1}	68.2±7.2^{6}	67.8
Glass	61.9±6.2	50.6±8.2^{1}	63.9±8.7^{6}	62.0±7.4^{6}	63.8±5.5^{6}	31.7
Heart disease	81.9±3.4	84.1±2.8^{1}	77.5±4.3^{1}	78.9±4.0^{1}	79.7±2.9^{3}	55.0
Hepatitis	85.3±3.7	85.2±4.0^{6}	79.2±4.3^{1}	79.0±5.1^{1}	80.3±4.2^{1}	78.1
Horse colic	80.7±3.7	79.3±3.7^{1}	85.1±3.8^{1}	75.7±5.0^{1}	82.5±4.2^{2}	63.6
Hypothyroid	97.5±0.3	97.9±0.4^{1}	99.1±0.2^{1}	95.9±0.7^{1}	98.8±0.4^{1}	95.3
Iris	93.2±3.5	93.9±1.9^{6}	92.6±2.7^{6}	93.5±3.0^{6}	93.3±3.6^{6}	26.5
Labor	91.3±4.9	88.7±10.6^{6}	78.1±7.9^{1}	89.7±5.0^{6}	82.1±6.9^{1}	65.0
Lung cancer	46.8±13.3	46.8±13.3^{6}	40.9±16.3^{5}	42.3±17.3^{6}	38.6±13.5^{3}	26.8
Liver disease	63.0±3.3	54.8±5.5^{1}	65.9±4.4^{1}	61.3±4.3^{6}	65.0±3.8^{3}	58.1
LED	62.9±6.5	62.9±6.5^{6}	61.2±8.4^{6}	55.3±6.1^{1}	58.6±8.1^{2}	8.0
Lymphography	81.6±5.9	81.1±4.8^{6}	75.0±4.2^{1}	82.9±5.6^{6}	78.8±4.9^{3}	57.3
Post-operative	64.7±6.8	67.2±5.0^{3}	70.0±5.2^{1}	59.2±8.0^{2}	60.8±8.2^{4}	71.2
Promoters	87.9±7.0	87.9±7.0^{6}	74.3±7.8^{1}	91.7±5.9^{3}	75.9±8.8^{1}	43.1
Primary tumor	44.2±5.5	44.2±5.5^{6}	35.9±5.8^{1}	30.9±4.7^{1}	39.8±5.2^{1}	24.6
Solar flare	68.5±3.0	68.2±3.7^{6}	70.6±2.9^{1}	67.6±3.5^{6}	70.4±3.0^{2}	25.2
Sonar	69.4±7.6	63.0±8.3^{1}	69.1±7.4^{6}	73.8±7.4^{1}	66.2±7.5^{5}	50.8
Soybean	100.0±0.0	100.0±0.0^{6}	95.0±9.0^{3}	100.0±0.0^{6}	96.9±5.9^{3}	30.0
Splice junctions	95.4±0.6	95.4±0.6^{6}	93.4±0.8^{1}	94.3±0.5^{1}	81.5±5.5^{1}	52.4
Voting records	91.2±1.7	91.2±1.7^{6}	96.3±1.3^{1}	94.9±1.2^{1}	95.8±1.6^{1}	60.5
Wine	96.4±2.2	97.8±1.2^{3}	92.4±5.6^{1}	97.2±1.8^{6}	90.8±4.7^{1}	36.4
Zoology	94.4±4.1	94.1±3.8^{6}	89.6±4.7^{1}	94.6±4.3^{6}	90.6±5.0^{1}	39.4

Ceiling effects in defect prediction

[Menzies07b]: when not to expect high precision

[Zhang07] commented that [Menzies07a]'s results had precision results that "were too low to be practical".

But, there are fundamental ceiling effects on the precision in the [Menzies07a] data sets.

- neg/pos ratios of 1,7,9,10,13,16,249.

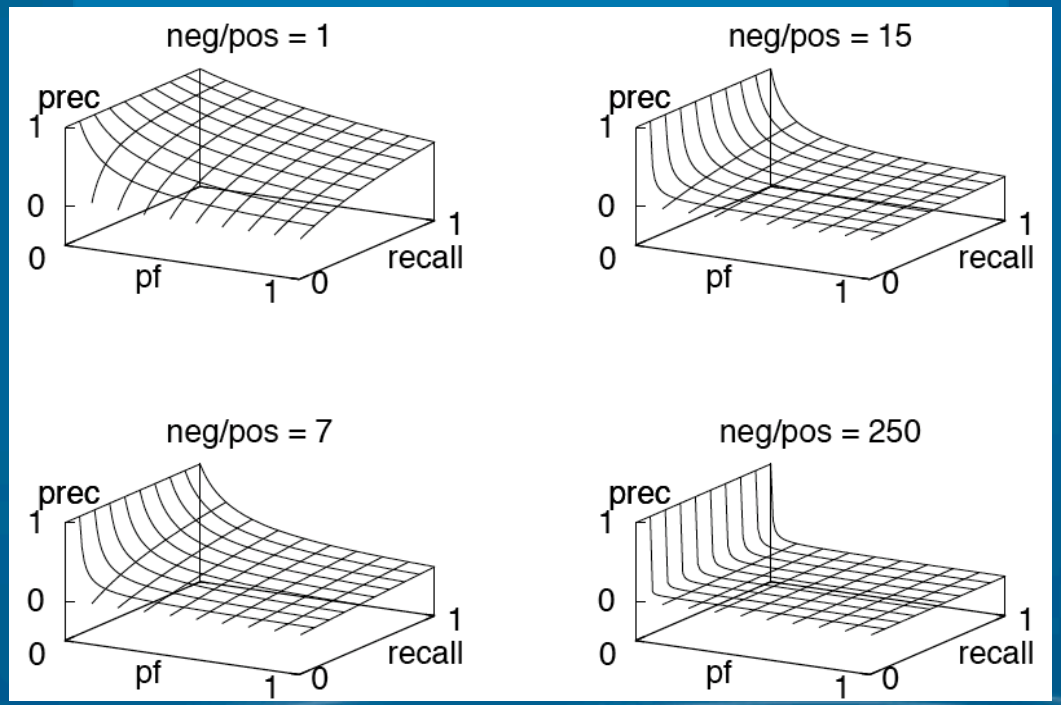
At neg/pos >= 15

- high precision requires very low pf
- but low pf implies low pd

So, if you want to *see* the target...

- It is fundamentally impossibly to require high precision

$$pf = \frac{pos}{neg} \cdot \frac{(1 - prec)}{prec} \cdot recall$$



2. Always test for solution stability

Several times, jiggle the data, recompute the model

Brittleness

[Harman07]: How robust are your solutions to change?

"In some software engineering applications, solution robustness may be as important as solution functionality.

- For example, it may be better to locate an area of the search space that is rich in stable solutions,
- Rather than identifying an even better solution that is surrounded by a set of far less fit solutions.

"Hitherto, research has tended to focus on the production of the fittest possible results.

- However, many application areas require solutions in a search space that may be subject to change.
- This makes robustness a natural second order property to which the research community could and should turn its attention."

Brittleness: try jiggling the training data

[Wu08]: Problem with decision tree learning

R = resubstitution error rate
- Error rate of a tree using the cases from which it was constructed

P = predictive error rate
- Error rate on cases not seen during construction.

P < E, sometimes dramatically:
- The "letter recognition dataset"
  - 20,000 cases:
- R(C4.5) = 4%,
- P(leave-one-out) = 20,000-fold cross-validation = 11.7%.

Leaving out a single case from 20,000 strong affects the constructed tree.

Brittleness: try jiggling the training data

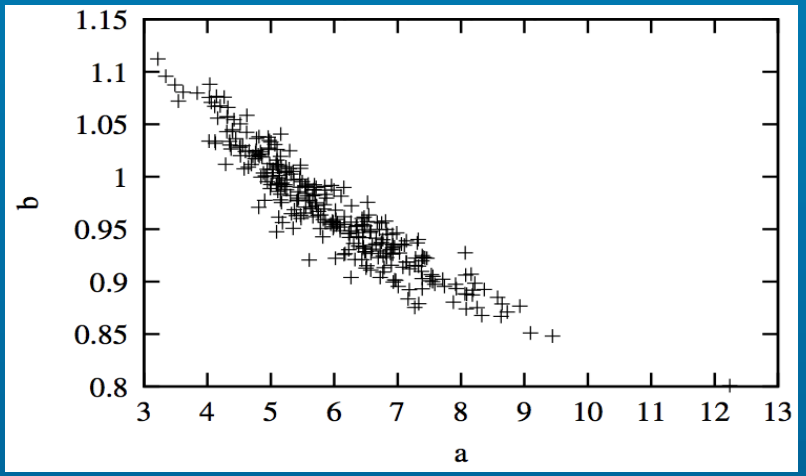
[Baker07]: Problem with effort estimation

COCOMO effort estimation

- effort = a * prod(EMi) * LOC b*sum(SFi)
- <a,b> are the "tuning parameters
- defaults to <a,b> = <2.94, 0.91>

Baker: 100 times, learn from 90% of some NASA cocomo data (selected at random)

Note that <2.94,0.91> not even on the chart.



Model instability an open and urgent issue in effort estimation [Foss05]

- Makes it difficult to claim that estimator A is better than estimator B

3. Throwing away data is a good thing

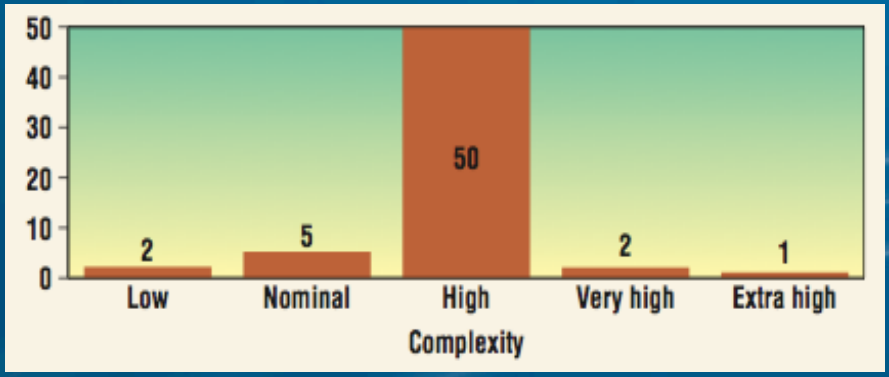
Try a little column pruning

Many reasons to prune columns

Not all data is useful data

- Some columns are noisy
  - contain signals unconnected to goal
  - E.g. due to Poorly collected data

- One column may be correlated to another (superflous)
- Reducing columns reduces variance in output model.
- All the values may be the same
- All the values center around a single value
  - Distribution of program complexity at NASA:



For example

http://iccle.googlecode.com/svn/trunk/share/data/arff/bn.arff

Linear regression reports

```
Defects =
82.2602 * S1=L,M,VH +
158.6082 * S1=M,VH +
249.407 * S1=VH +
41.0281 * S2=L,H +
68.9153 * S2=H +
151.9207 * S3=M,H +
125.4786 * S3=H +
257.8698 * S4=H,M,VL +
108.1679 * S4=VL +
134.9064 * S5=L,M +
-385.7142 * S6=H,M,VH +
115.5933 * S6=VH +
-178.9595 * S7=H,L,M,VL +
...
[ 50 lines deleted ]
```

Correlates 0.45 (badly) to "actual"

Column selection with WRAPPER

10 times,

- take 90% of the data
- run a best first search through combinations of attributes.
- At each step, call linear regression to asses a particular combination of attributes.

Report the number of times (out of 10) that WRAPPER selected for a variable

Wrapper results

(Recall we want to improve over correlation = 0.45)

- Linear regression reports five variables (out of 24) selected 50% or more.

8(80 %) KLoC
6(60 %) P5
6(60 %) S7
5(50 %) D3
4(40 %) Language
3(30 %) log(hours)
2(20 %) Hours
2(20 %) P7
2(20 %) D1
….
[snip]

A second run of a 10-way using just those variables

Results:

- much larger correlation (98%):
- a smaller model

Defects =
876.3379 * S7=VL +
-292.9474 * D3=L,M +
483.6206 * P5=M +
5.5113 * KLoC +
95.4278

4. Beware the Straight Line

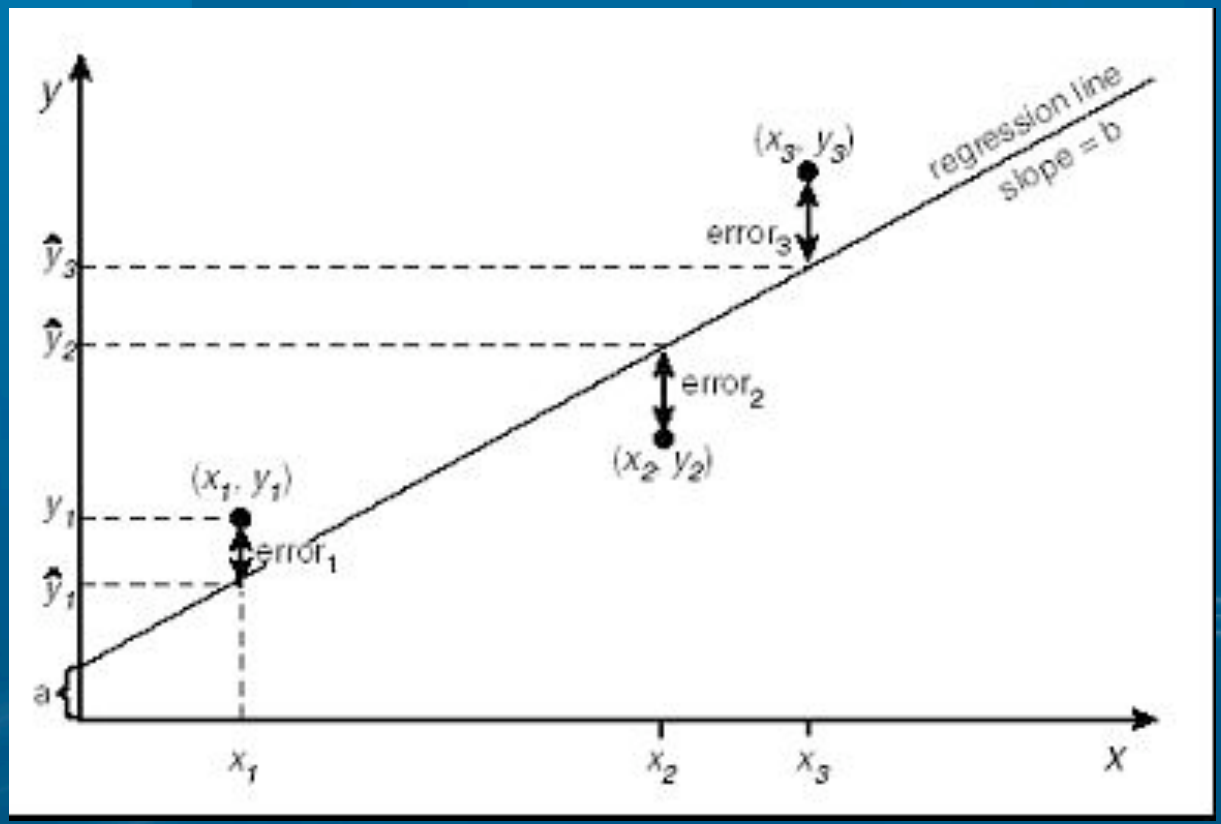
About correlation

Correlation Analysis

An Equation is obtained by fitting a straight line that minimizes the sum of the squared errors (linear or multivariate regression).

Errors represent the residuals, which are the differences between actual and predicted values.

$$\hat{y} = a + bx$$

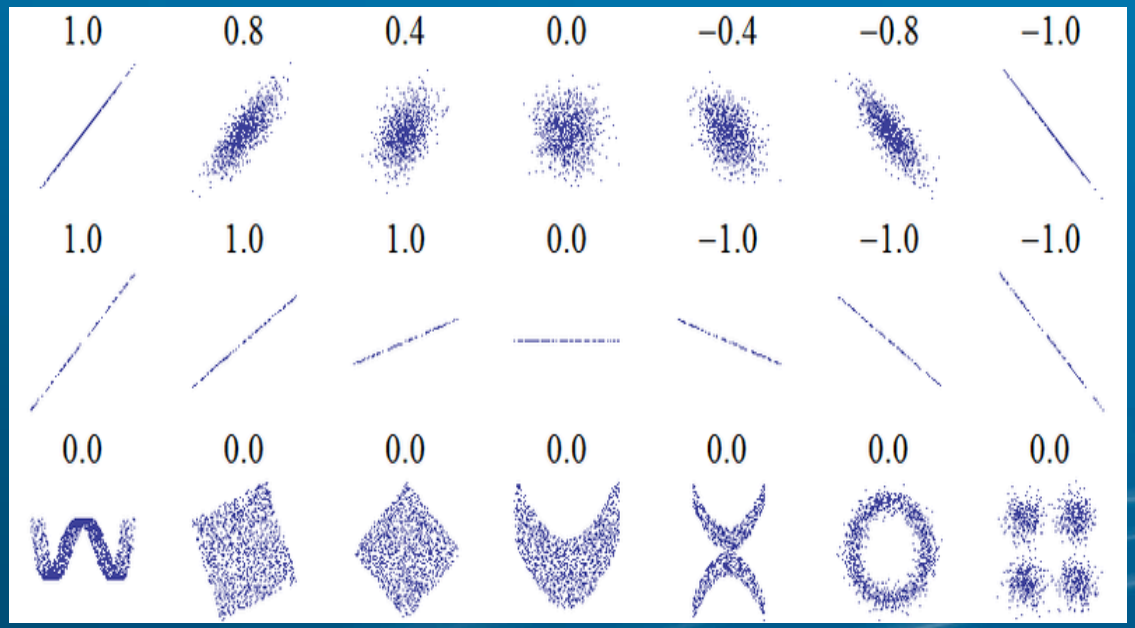


Correlation

Performance measures for continuous variables

Does "this" tend to move "that"?

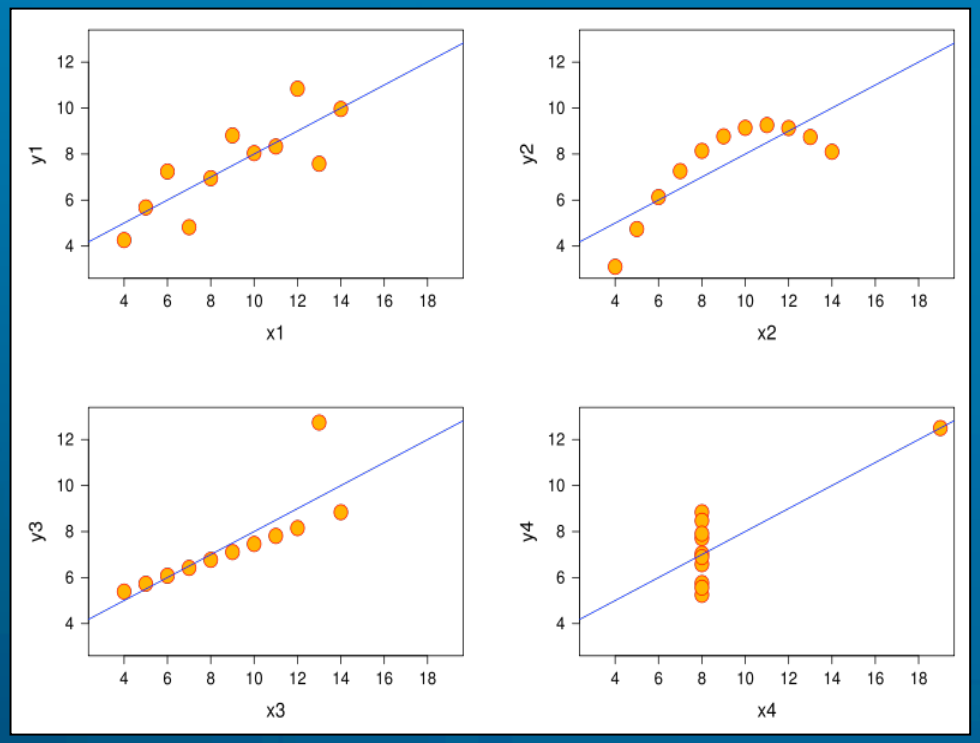
- correlation = 0 : no connection
- correlation = 1 : strong positive correlation
- correlation = 2 : strong negative connection



Correlation: traps for the unwary

Traps for the unwary

All these (x1,y1) values have a correlation of 0.8.



Correlation: traps for the unwary (#2)

Correlation does not predict for (pd,pf)

model 1: defects predicted by the "Halstead" measures (poor correlation)

$$
\begin{aligned}
defects_1 &= 0.231 + (0.00344 * N) + (8.88e-4 * V) \\
&\quad -(0.185 * L) - (0.0343 * D) - (0.00541 * I) \\
&\quad +(1.68e-5 * E) + (0.711 * B) \\
&\quad -(4.7e-4 * T) \\
c_1 &= -0.3616
\end{aligned}
$$

model 2: defects predicted by LOC (better correlation)

$$
\begin{aligned}
defects_2 &= 0.0164 + 0.0114 * LOC \\
c_2 &= 0.65
\end{aligned}
$$

predict "defect!" if modeli >= x

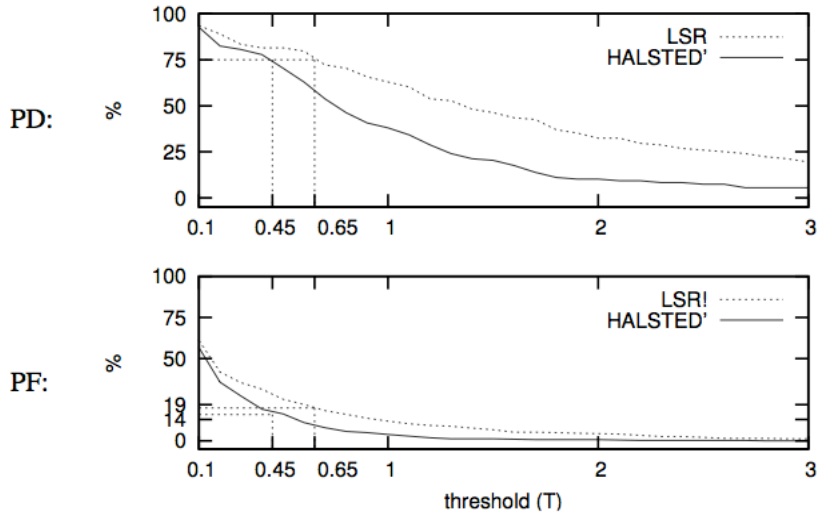


Figure 5. Effort, probability of false alarm, and probability of detection seen using $defects_i \geq T$ where $defects_i$ is one of Equation 1 (the "HALSTEAD" curves) or Equation 4 (the "LSR" curves) and T controls when the detector triggering (see the discussion around Equation 3).

"Correlation" is not "decision"

5. More Maths

Different evaluation measures: Accuracy measures

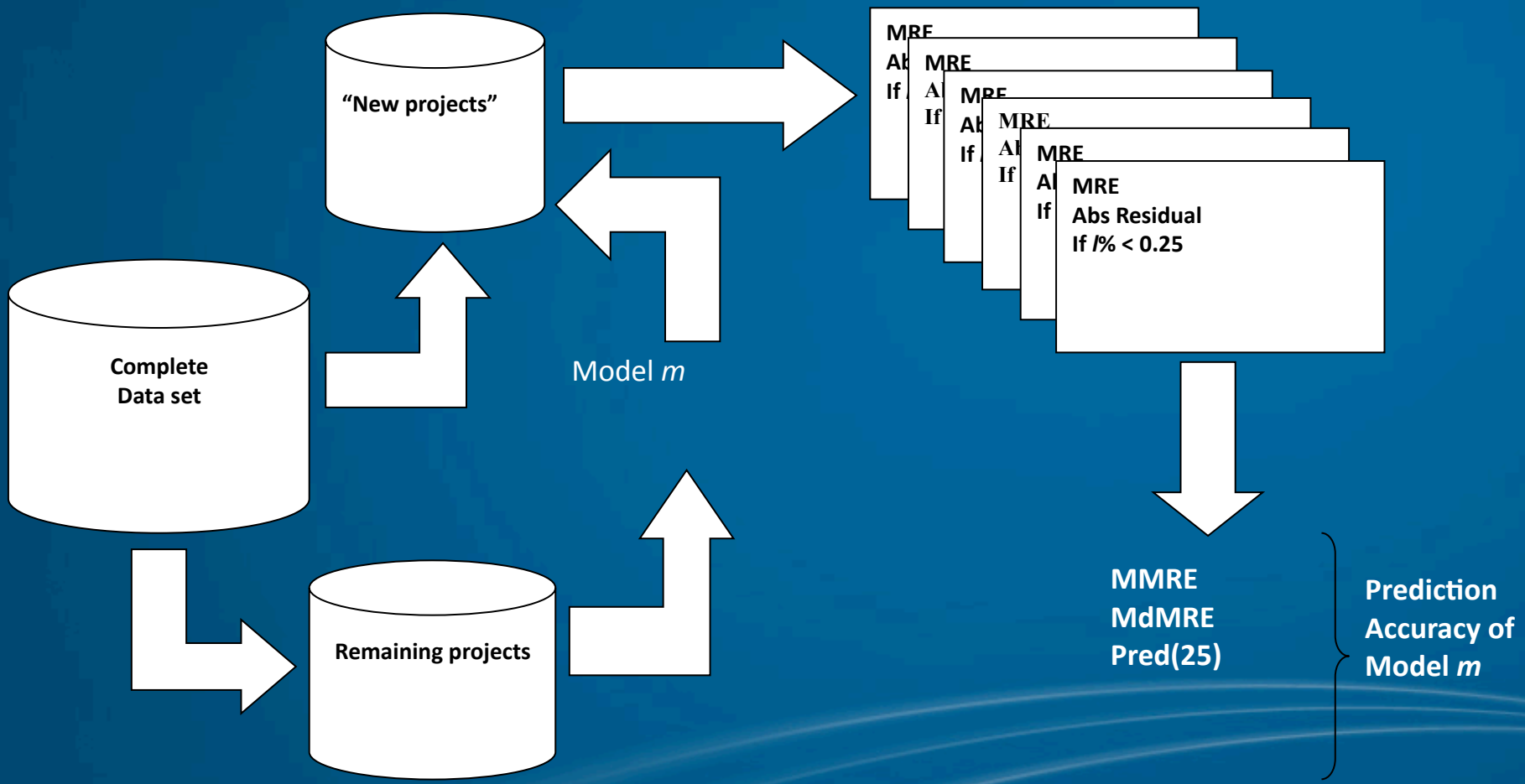
Measures based on the Magnitude of Relative Error (MRE)

$$MRE = \frac{|e - \hat{e}|}{e}$$

Mean MRE (MMRE)
Median MRE (MdMRE)
Relative error not greater than *l* (generally at 0.25) (Pred(25))

Absolute residuals => | e - $\hat{e}$ |

Different evaluation measures: Accuracy measures



AR, RE, MRE, PRED, MMRE, medMRE

How close/far away are you from "it"?

Given prediction "p" and actual value "a"

- AR = Absolute residual = a - p
- RE = Relative error = (a - p) / a
- MRE= Magntiude of releative error = abs(RE)

For predictions p1...pn and actuals a1...an

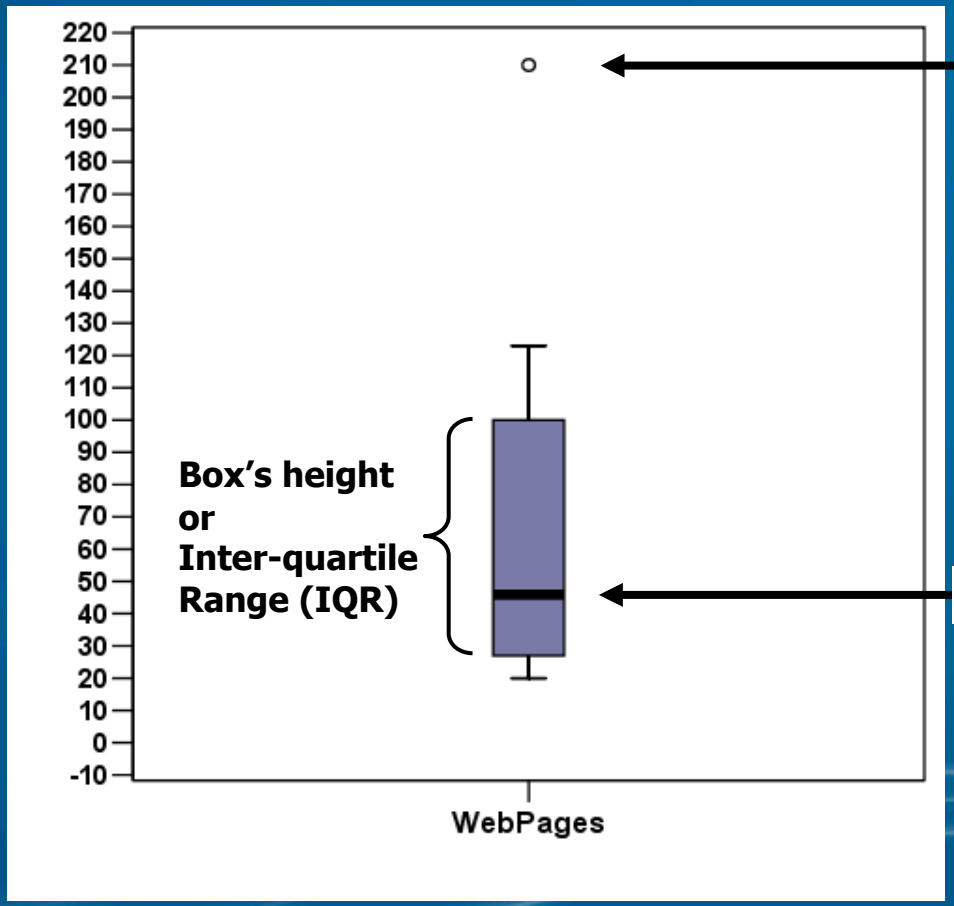
- PRED(N): percentage of predictions with MRE < N%
  - a.k.a. how close did you get
- MMRE: mean MRE ; a.k.a. how far away did you fall
- MedMRE = median MRE (50th percentile)

Notes:

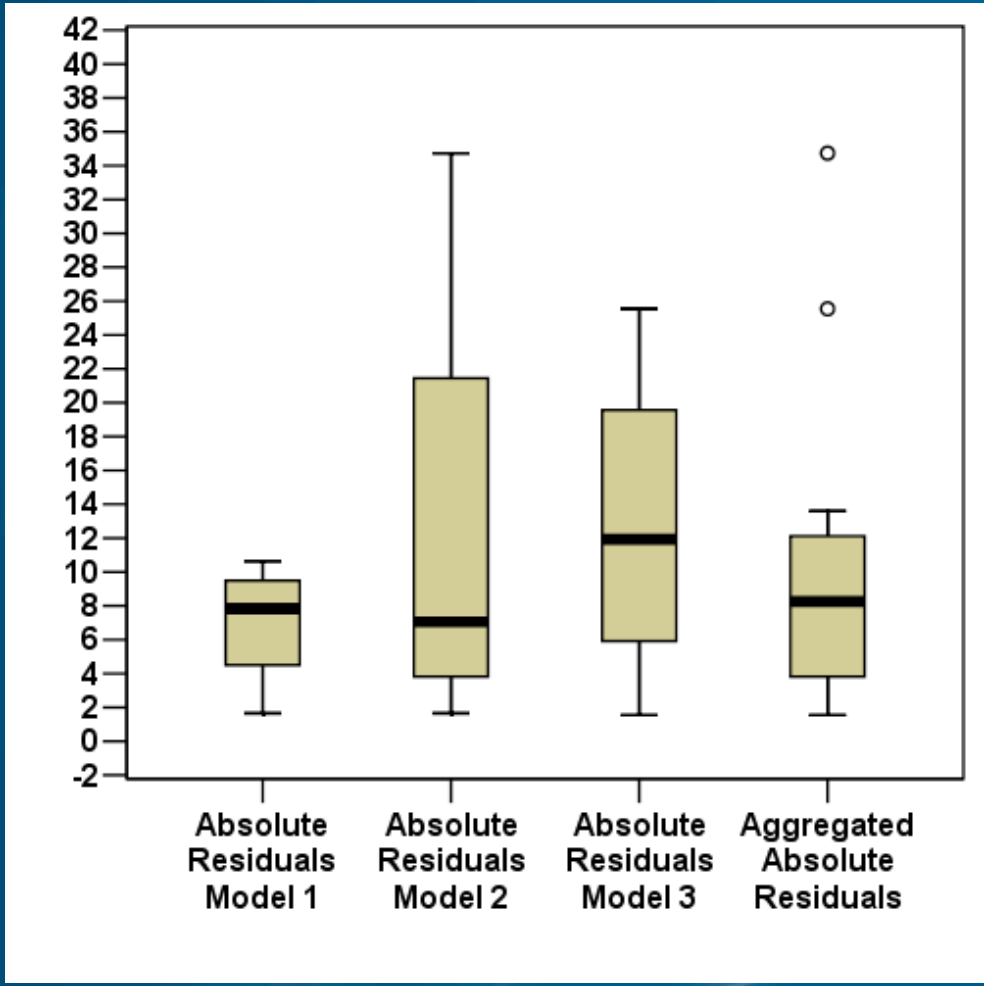
- Lower is better for AR, RE, MRE, MRE, medMRE
- Higher is better for PRED
- PRED(25) is a commonly reported statistic
  - Can be high, even if a small number of predictions are very bad (since PRED is blind to such bad cases)
- MMRE can be high, even if most errors are low
  - since any mean measure can be distorted by a few large outliers

Different evaluation measures: Accuracy measures

Webpages
20
20
25
26
1st quartile → 27
35
45
median → 45
46
67
89
3rd quartile → 90
100
100
100
123



Different evaluation measures: Accuracy measures



Reporting

- Parametric test
  - Compare Means
    - Independent samples T-test (2 means)
    - Paired Samples T-test
    - One-way ANOVA (comparing more than 2 means)
- Non-parametric test
  - Use ranks
    - Mann-Whitney U test
    - Wilcoxon test
    - Kruskal-Wallis H

Reporting

- Four types of validity that must be considered:
  - **Internal:** Unknown factors that may affect the dependent variable. E.g. confounding factors we're unaware of.
  - **External:** To what extent we can generalise the findings.
  - **Conclusion:** To be able to draw correct conclusions regarding the relationship between treatments and the experiment's outcome. E.g. use of adequate statistical test, use of proper measurement
  - **Construct:** Represents to what extent the independent and dependent variables precisely measure the concepts they claim to measure.

Case Studies: Kitchenham and Mendes (2009)

- Dataset Selection
  - Need to motivate the choice of datasets
  - Only two or three datasets are regularly used, despite the existence of at least 31 datasets in the public domain!
  - How representative are the projects from the two-three mostly used datasets?
- Non-repeatable Results
  - Several studies do not document the procedure used to select projects for analysis => results cannot be independently validated!
  - Example of problematic dataset: ISBSG dataset

Case Studies: Kitchenham and Mendes (2009)

- Expertise in competing methods
  - Proponents of new techniques may be experts in these new techniques, but are they also experts in statistical methods?
  - What happens if they simply reuse MMREs from the literature? How reliable are these?
- Failure to present statistical evidence
  - Important to use statistical significance tests.
  - Unfortunately, many papers base their conclusions solely on MMRE, MdMRE and Pred(25) values.
  - Another issue: shall we use training/validation sets, or base our predictions on the entire dataset?

Case Studies: Kitchenham and Mendes (2009)

- Using MMRE for model building and model comparison
  - What happens if the technique being compared 'learns' to choose a model that underestimates, thus optimising MMRE? How fair would a comparison be?
- Relevance to Real estimation processes
  - The data used to build a model should be representative of the projects for which the model will be used.
  - One should take into account any heterogeneity among projects by at least:
    - Appropriately classifying different types of projects (e.g. new and enhancement)
    - Considering the age of projects, in particular when dealing with companies focusing on SPI

Other lessons

- Describe
  - the experimental design used
  - the statistical analysis employed
  - the data filtering employed (e.g. removing outliers, filtering data)
  - the threats to the validity of the evaluation
  - the type of sample used (e.g. self-selected, random) and origin (students, professionals)
  - the contribution of the paper
  - previous work and how your work makes a contribution

Roadmap

Where are we going now?

- Preliminaries
- Why
  - evidence-based SE
  - current (weak) state of the field
- How:
  - Do you know your ABCs?
- Case studies....
- References

References

Read and read again...

- [Armstrong07] Armstrong JS. 2007. Significance tests harm progress in forecasting. International Journal of Forecasting 23: 21.
- [Baker07] Dan Baker "A hybrid approach to expert and model-based effort estimation". Master thesis, LCSEE, WVU,
- [Basili09] "Personnel Communication" by V. Basili, 2009
- [Budgen09] "Personnel Communication" by D. Budgen, 2009
- [Cohen88] J.Cohen 1988. The earth is round (p < .05). American Psychologist 49: 997-1003.
- [Demsar06] J. Demsar, "Statistical Comparisons of Classifiers over Multiple Data Sets" Journal of Machine Learning Research 7 (2006) 1-30
- [Easterbrook03] S. Easterbrook, J. Singer, M. Storey, and D. Damian "Selecting Empirical Methods for Software Engineering Research" by. In [Shull03] p285-311, 2003.
- [Endres03] "A Handbook of Software and Systems Engineering" by A. Endres and D. Rombach, Addison-Wesley, 2003.
- [Fenton96] "Software Metrics, a rigorous approach" by N. Fenton and S.L. Pfleeger. Pws Pub Co, 2nd Edition, 1996
- [Fenton07] "Invited talk" by N. Fenton. In PROMSE'07. In http://promisedata.org/?cat=130. 2007.
- [Foss05] T. Foss , E. Stensrud , B. Kitchenham , I. Myrtveit, A Simulation Study of the Model Evaluation Criterion MMRE, IEEE Transactions on Software Engineering, v.29 n.11, p.985-995, November 2003
- [Glass02] "Facts and Fallacies of Software Engineering", Addison Wesley, 2002.
- [GreenMenzies09] P. Green, T. Menzies, S. Williams, O. El-Rawas, "Understanding the Value of Software Engineering Technologies", IEE ASE 2009
- [Harman07] M. Harman. The current state and future of search based software engineering. In Future of¬† Software Engineering, ICSE,Äô07. 2007
- [Holte93] R.C. Holte "Very simple classification rules perform well on most commonly used datasets", Machine Learning, 1993, pp63-91.
- [Jiang07] Y. Jiang, B. Cukic, and T. Menzies. Fault prediction using early lifecycle data. In ISSRE,Äô07, 2007. Available from http://menzies.us/pdf/07issre.pdf.
- [Kitchenham04] B.A. Kitchenham, T. Dyba, M. Jorrgensen "Evidence-Based Software Engineering", ICSE 2004: 273-281

References

Read and read again...

- [Kitchenham07] B.A. Kitchenham, E. Mendes, G. Horta Travassos "Cross versus Within-Company Cost Estimation Studies: A Systematic Review:, IEEE Trans. Software Eng. 33(5): 316-329 (2007)
- [KitchenhamMendes09] Why Data Mining Studies may be Invalid, PROMISE 2009
- [Lessmann08] S. Lessmann, B. Baesens, C. Mues, S. Pietsch "Benchmarking Classification Models for Software Defect Prediction: A Proposed Framework and Novel Findings" IEEE Transactions on In Software Engineering, IEEE Transactions on, Vol. 34, No. 4. (2008), pp. 485-49
- [Mendes07] E. Mendes, "Cost Estimation Techniques For Web Projects" IGI Publishing, 2007
- [Menzies07a] T. Menzies, J. Greenwalk, A. Frank, "Data Mining Static Code Attributes to Learn Defect Predictors" IEEE Transactions on Software EngIneering, Vol. 32, No. 11, January 2007
- [Menzies07b] T. Menzies, A. Dekhtyar, J. Distefano, J. Greenwald, "Problems with Precision: A Response to 'Comments on Data Mining Static Code Attributes to Learn Defect Predictors", September 2007 (vol. 33 no. 9) pp. 637-640, 2007
- [Menzies08a] "Learning Better IV&V Practices" by T. Menzies,, M. Benson, K. Costello. C. Moats. M. Northey and J. Richarson. In Innovations in Systems and Software Engineering. http://menzies.us/pdf/07ivv.pdf. March 2008.
- [Shull03] "Guide to Advanced Empirical Software Engineering" F. Shull and J. Singer and D.I.K. Sjoberg. Springer, 2003.
- [Wu08] "Top 10 algorithms in data mining" by X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P S. Yu, Zhi-Z. Zhou, M. Steinbachm D. J. Hand Dan Steinberg in Knowl Inf Syst (2008) 14:1,37
- [Zannier06] C. Zannier, G. Melnik, and F. Maurer "On the Success of Empirical Studies in the International Conference on Software Engineering",. In ICSE'06.
- [Zhang07]: H. Zhang, X. Zhang, Comments on "Data Mining Static Code Attributes to Learn Defect Predictors", September 2007 (vol. 33 no. 9) pp. 635-637 IEEE Transactions on Software Engineering, vol. 33, no. 9, pp. 635-637, September, 2007.
- [Zimmerman09] T. Zimmermann, N. Nagappan, H. Gall, E. Giger, and B. Murphy, "Cross-project Defect Prediction", in Proceedings of the 7th joint meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering (ESEC/ FSE),
Association for Computing Machinery, Inc., August 2009