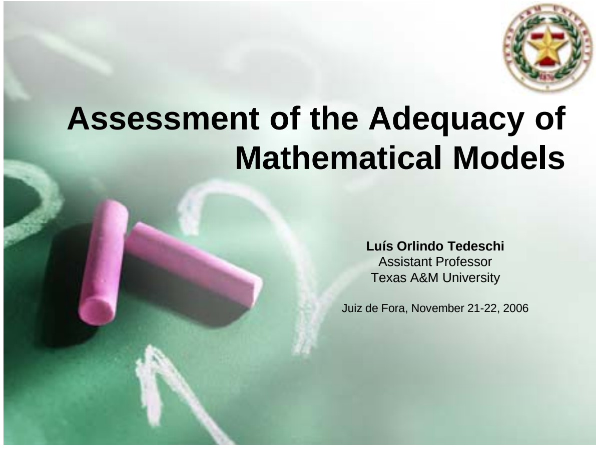
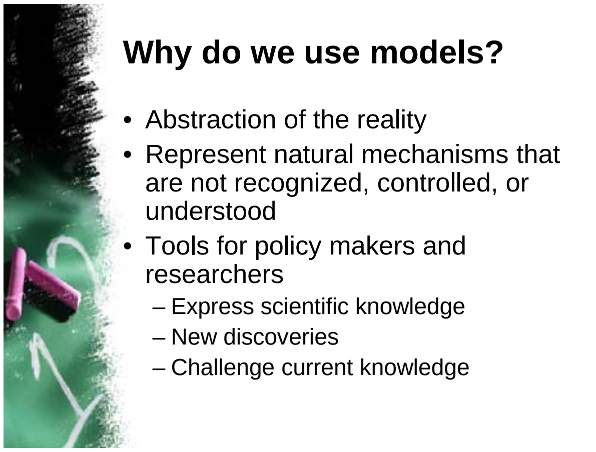


Assessment of the Adequacy of Mathematical Models

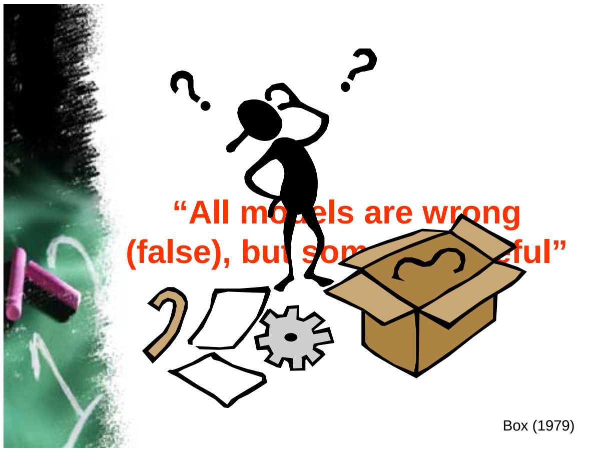
Luís Orlindo Tedeschi
Assistant Professor
Texas A&M University

Juiz de Fora, November 21-22, 2006



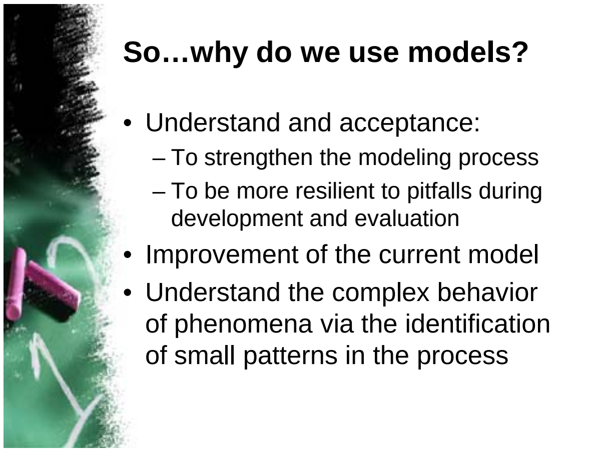
Why do we use models?

- Abstraction of the reality
- Represent natural mechanisms that are not recognized, controlled, or understood
- Tools for policy makers and researchers
  - – Express scientific knowledge
  - – New discoveries
  - – Challenge current knowledge



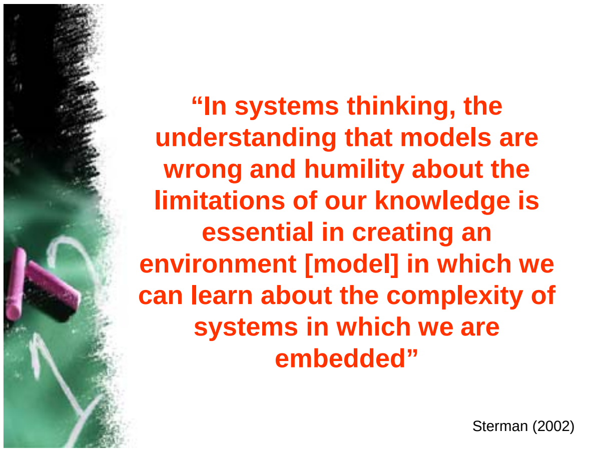
“All models are wrong (false), but some are useful”

Box (1979)



So…why do we use models?

- Understand and acceptance:
  - – To strengthen the modeling process
  - – To be more resilient to pitfalls during development and evaluation
- Improvement of the current model
- Understand the complex behavior of phenomena via the identification of small patterns in the process

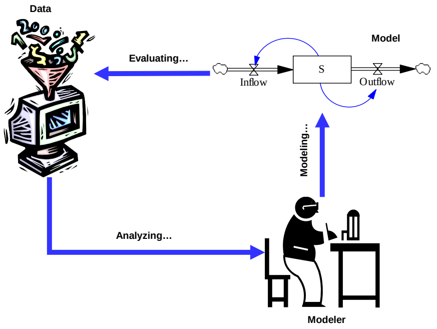


“In systems thinking, the understanding that models are wrong and humility about the limitations of our knowledge is essential in creating an environment [model] in which we can learn about the complexity of systems in which we are embedded”

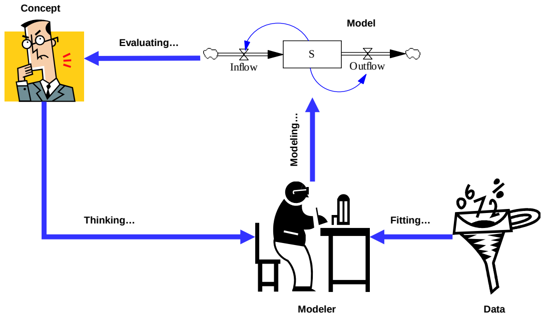
Sterman (2002)

Processes for Model Development using Systems Thinking

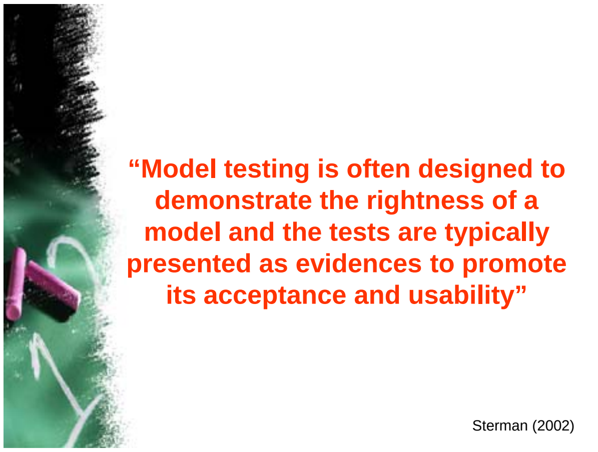
Empirical or Relational Models



Conceptual or Theoretical Models

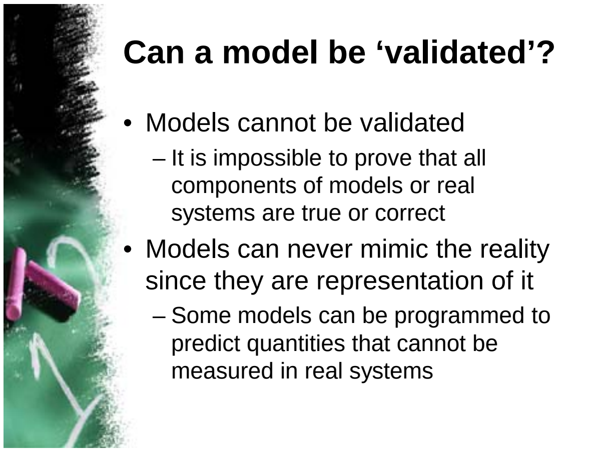


Model Evaluation



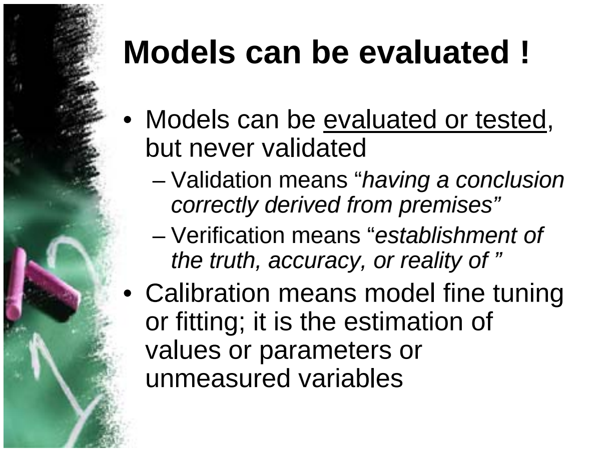
"Model testing is often designed to demonstrate the rightness of a model and the tests are typically presented as evidences to promote its acceptance and usability"

Sterman (2002)



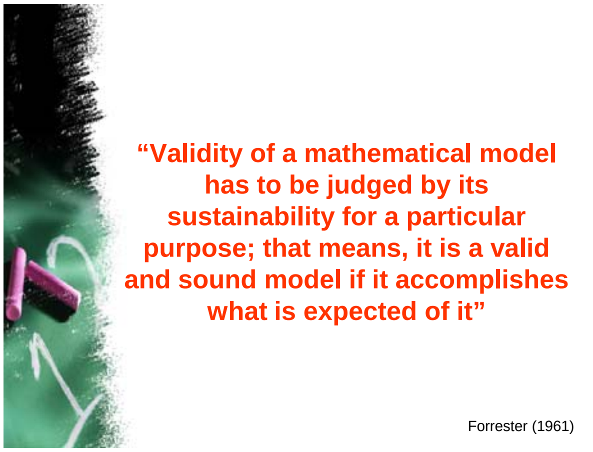
Can a model be 'validated'?

- Models cannot be validated
  - It is impossible to prove that all components of models or real systems are true or correct
- Models can never mimic the reality since they are representation of it
  - Some models can be programmed to predict quantities that cannot be measured in real systems



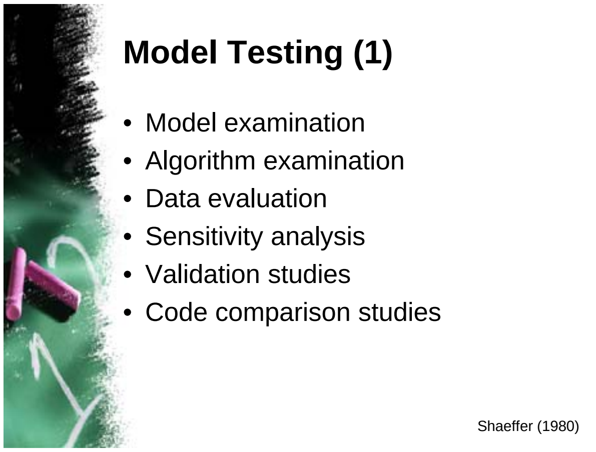
Models can be evaluated !

- Models can be evaluated or tested, but never validated
  - Validation means "*having a conclusion correctly derived from premises"*
  - Verification means "*establishment of the truth, accuracy, or reality of "*
- Calibration means model fine tuning or fitting; it is the estimation of values or parameters or unmeasured variables



“Validity of a mathematical model has to be judged by its sustainability for a particular purpose; that means, it is a valid and sound model if it accomplishes what is expected of it”

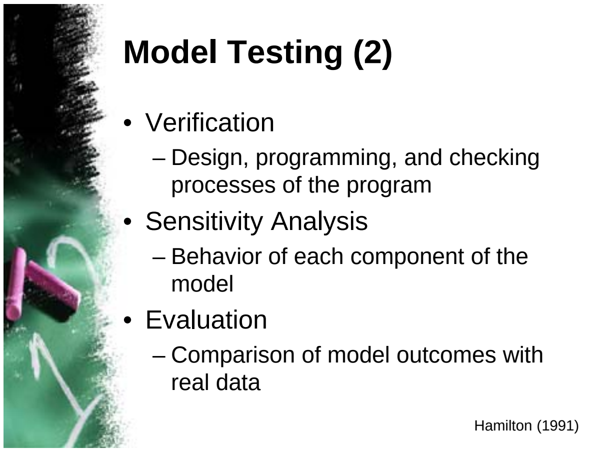
Forrester (1961)



Model Testing (1)

- Model examination
- Algorithm examination
- Data evaluation
- Sensitivity analysis
- Validation studies
- Code comparison studies

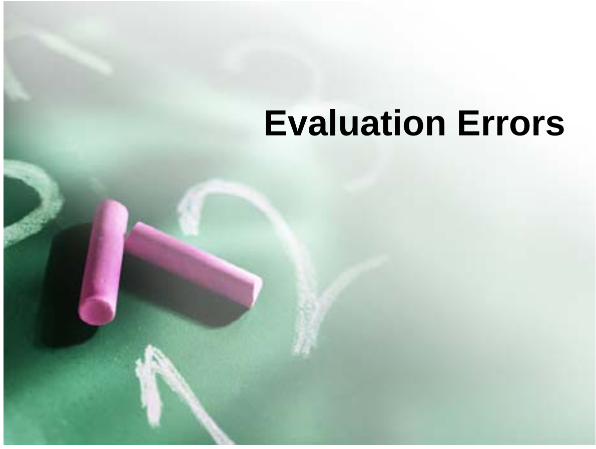
Shaeffer (1980)



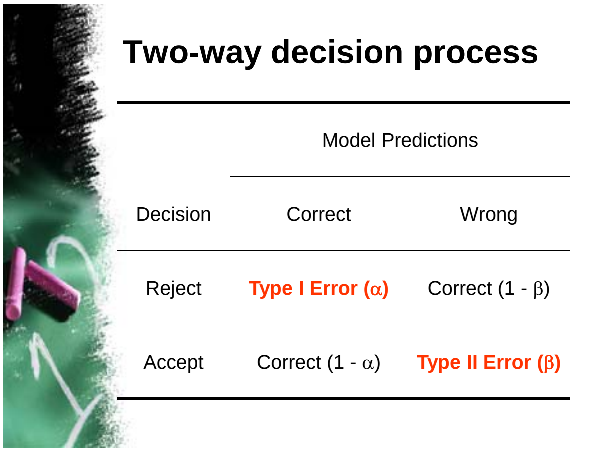
Model Testing (2)

- Verification
  - Design, programming, and checking processes of the program
- Sensitivity Analysis
  - Behavior of each component of the model
- Evaluation
  - Comparison of model outcomes with real data

Hamilton (1991)

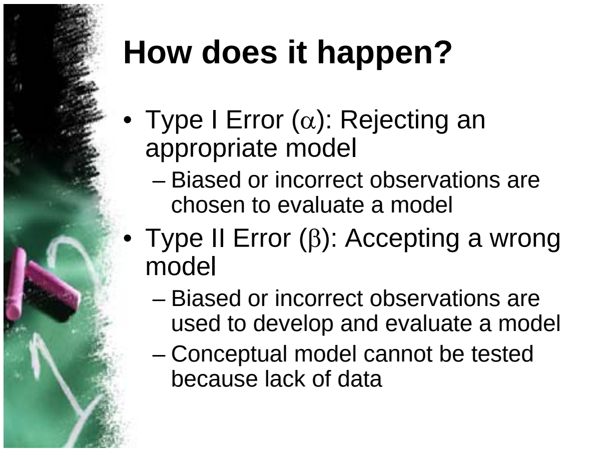


Evaluation Errors



Two-way decision process

	Model Predictions	
Decision	Correct	Wrong
Reject	**Type I Error (α)**	Correct (1 - β)
Accept	Correct (1 - α)	**Type II Error (β)**



How does it happen?

- Type I Error (α): Rejecting an appropriate model
  - Biased or incorrect observations are chosen to evaluate a model
- Type II Error (β): Accepting a wrong model
  - Biased or incorrect observations are used to develop and evaluate a model
  - Conceptual model cannot be tested because lack of data

Accuracy x Precision

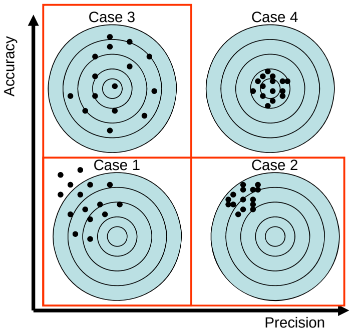
Definition

- Accuracy
  - It measures how closely model-predicted values are to the true values
  - Ability to predict the right values
- Precision
  - It measures how closely individual model-predicted values are within each other
  - Ability to predict similar values consistently

Definition

- Inaccuracy or bias
  - Systematic deviation from the truth
- Imprecision or uncertainty
  - Magnitude of the scatter about the average mean

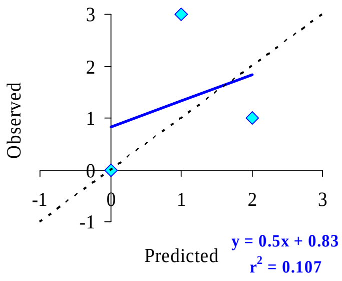
Accuracy x Precision



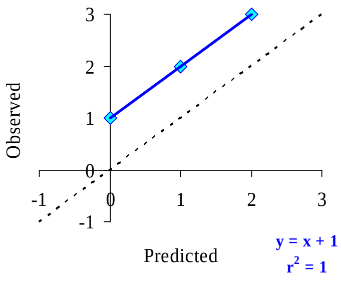
Imprecise or Uncertain

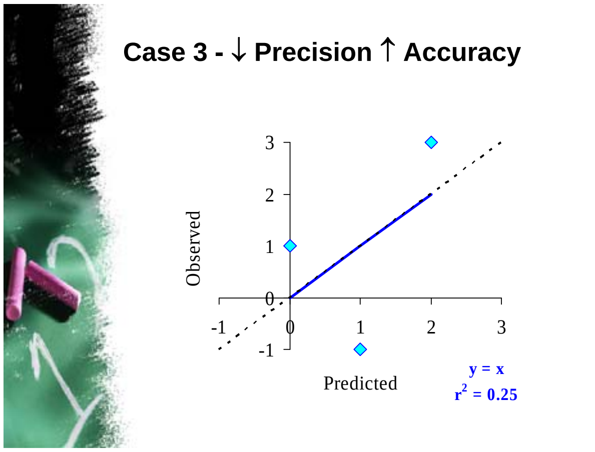
Inaccurate or biased

Case 1 - ↓ Precision ↓ Accuracy

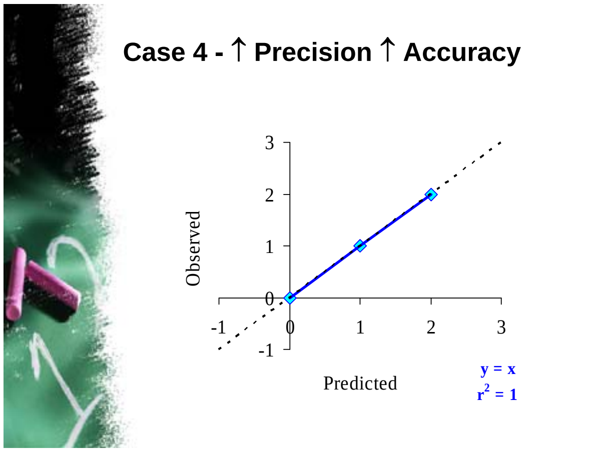


Case 2 - ↑ Precision ↓ Accuracy

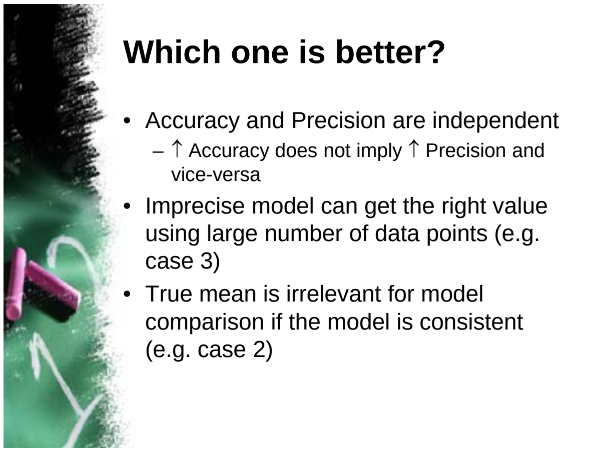




Case 3 - ↓ Precision ↑ Accuracy



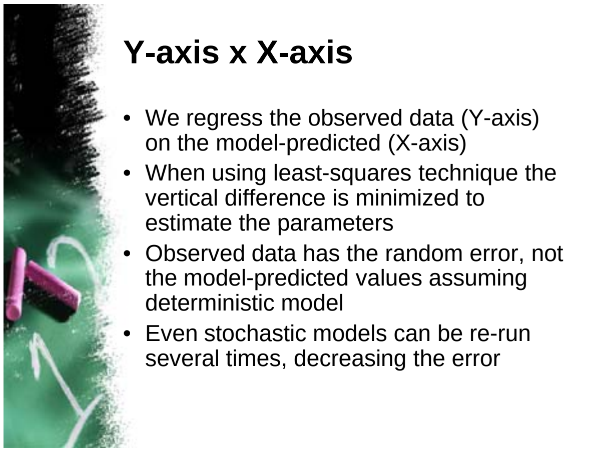
Case 4 - ↑ Precision ↑ Accuracy



Which one is better?

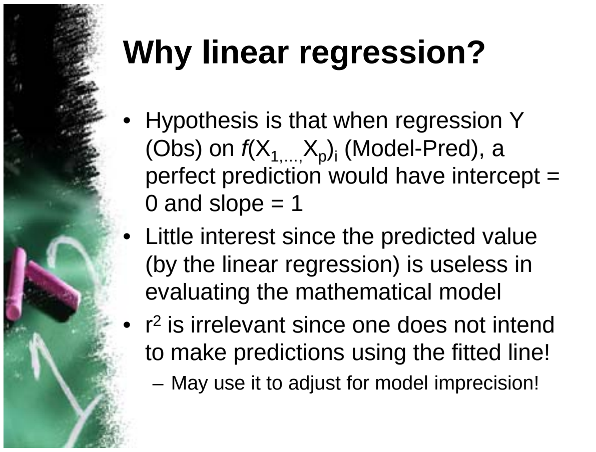
- Accuracy and Precision are independent
  - ↑ Accuracy does not imply ↑ Precision and vice-versa
- Imprecise model can get the right value using large number of data points (e.g. case 3)
- True mean is irrelevant for model comparison if the model is consistent (e.g. case 2)

Techniques for Model Evaluation: *Regression Analysis*



Y-axis x X-axis

- We regress the observed data (Y-axis) on the model-predicted (X-axis)
- When using least-squares technique the vertical difference is minimized to estimate the parameters
- Observed data has the random error, not the model-predicted values assuming deterministic model
- Even stochastic models can be re-run several times, decreasing the error



Why linear regression?

- Hypothesis is that when regression Y (Obs) on $f(X_{1,\ldots,}X_{p})_{i}$ (Model-Pred), a perfect prediction would have intercept = 0 and slope = 1
- Little interest since the predicted value (by the linear regression) is useless in evaluating the mathematical model
- r^2 is irrelevant since one does not intend to make predictions using the fitted line!
  - May use it to adjust for model imprecision!

Assumptions for LR

- The X-axis values are known without errors (deterministic)
- The Y-axis values have to be independent, random, and homoscedastic
- Residuals are independent and identically distributed ~ N(0,σ^2)

Caution about r^2

- A high coefficient of correlation (r) does not indicate that useful predictions can be made by a given mathematical model since it measures precision not accuracy
- A high r does not imply the estimated line is a good fit (curvilinear)
- An r near zero does not indicate that observed and model-predicted are not correlated since they may have a curvilinear shape

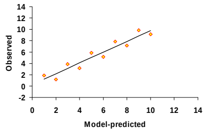
Mean square error (MSE)

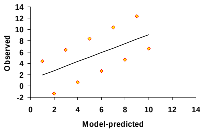
- Also known as residual mean square or standard error of the estimate
- This statistic may be used to compare model 'validity' when comparing models

$$MSE = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-2}$$

$$MSE = \frac{s_Y^2 \times (n-1) \times (1-r^2)}{n-2}$$

Comparison of Model Prediction





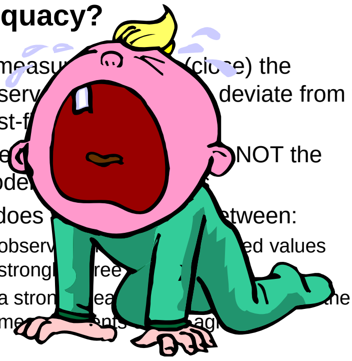
- $Y_1 = a + b \times X \pm N(0,1)_{\alpha=0.2}$
- a = 0.28 ± 0.63
- b = 0.95 ± 0.10
- *P* (a=0) = 0.67
- *P* (b=1) = 0.63
- *P* (a=0 & b=1) = 0.90
- **r^2 = 0.92**
- **MSE = 0.89**

- $Y_2 = a + b \times X \pm N(0,4)_{\alpha=0.2}$
- a = 1.12 ± 2.53
- b = 0.80 ± 0.41
- *P* (a=0) = 0.67
- *P* (b=1) = 0.63
- *P* (a=0 & b=1) = 0.90
- **r^2 = 0.32**
- **MSE = 13.7**

Concerns about LR

- Assumptions of normality and homoscedasticity are rarely satisfied
- Ambiguous results depending on the scatter of the data
- Regression lacks sensitivity to distinguish between random clouds and data points
- Stochastic models require different technique to derive the parameters

Is r^2 a good indicator of adequacy?



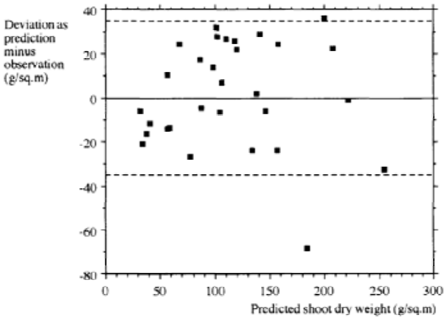
- r^2 measu (cio e) the observ deviate from the best-f
- The NOT the mode
- r^2 does etween:
  - observ ed values strongl ree
  - a stron ea he me nts ag

Fitting Errors: *Analysis of Deviation*

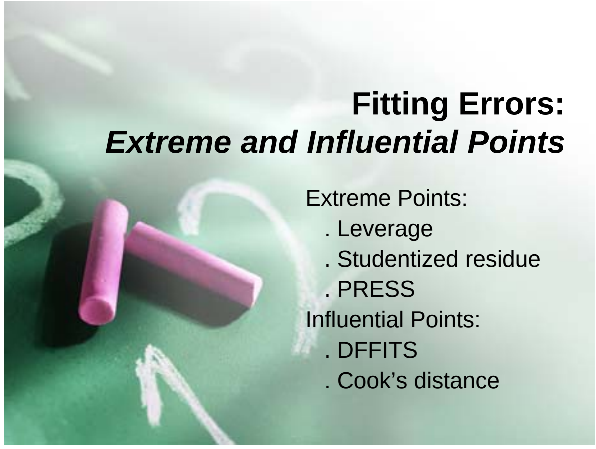
Analysis of Deviation

- Empirical but powerful analysis
- Deviation is the difference between **model-predicted** minus **observed** values
- Usually, an acceptable range is used to accept or not the model performance

Deviation Plot Analysis



Mitchell and Sheehy (1997)



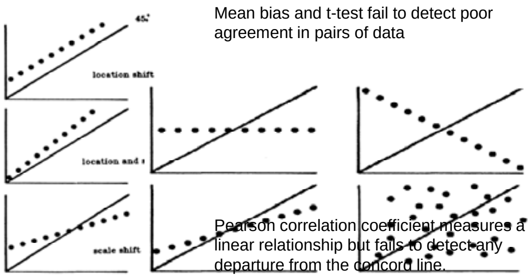
Fitting Errors: *Extreme and Influential Points*

Extreme Points:
. Leverage
. Studentized residue
. PRESS
Influential Points:
. DFFITS
. Cook's distance

Concordance Correlation Coefficient

Failure of Agreement Measures

Mean bias and t-test fail to detect poor agreement in pairs of data



Pearson correlation coefficient measures a linear relationship but fails to detect any departure from the concord line.

Lin (1989)

What is CCC?

- CCC aka reproducibility index
- Are the model-predicted values precise and accurate at the same time across a range and are tightly amalgamated along the unity line through the origin?
- CCC accounts for precision and accuracy at the same time
- Proposed initially by Krippendorff (1970) and modified by Lin (1989)

How CCC is computed?

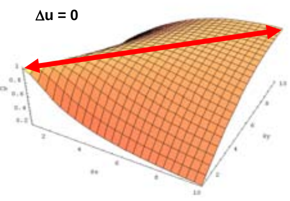
$$\hat{\rho}_c = \frac{2 \times s_{f(X_1,\ldots,X_p)Y}}{s_Y^2 + s_{f(X_1,\ldots,X_p)}^2 + (\bar{Y} - \bar{f}(X_1,\ldots,X_p))^2}$$

Decomposition:

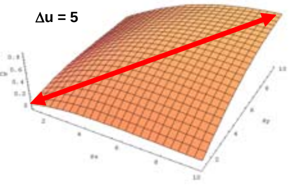
$$\hat{\rho}_c = \hat{\rho} \times C_b \quad \text{and} \quad C_b = \frac{2}{\left[v + \frac{1}{v} + \mu^2 \right]}$$

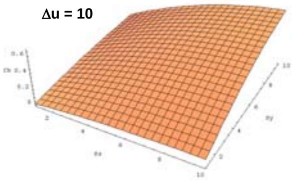
$$v = \frac{\sigma_1}{\sigma_2} \text{ for population} \qquad \mu = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1 \sigma_2}} \text{ for population}$$

$$v = \frac{s_Y}{s_{f(X_1,\ldots,X_p)}} \text{ for sample} \qquad \mu = \frac{\bar{Y} - \bar{f}(X_1,\ldots,X_p)}{\sqrt{s_Y s_{f(X_1,\ldots,X_p)}}} \text{ for sample}$$



Effects of Δu and Δv on Accuracy (C_b)





Limitations of CCC

- Assumes that each pair of data point are interchangeable, that means, the order of the data point does not matter; there is no covariance
- Nickerson (1997) suggested an adaptation to the CCC

An improved CCC estimate

- CCC uses squared perpendicular distance $(Y_1 – Y_2)^2$ of any paired data point to the unity line
- Unfortunately, it measures only how close the data point is to the unity line and not which direction it goes

An improved CCC estimate

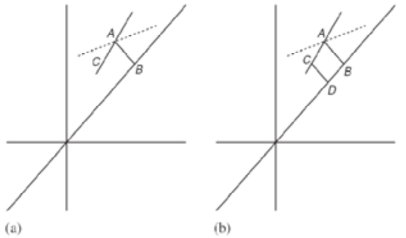


Figure 1. Comparison of the two criteria: (a) Lin's criteria; (b) new criterion.

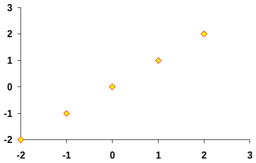
Liao (2003)

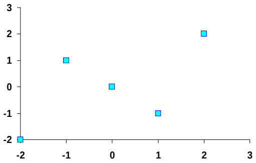
An improved CCC estimate

- It is a quadratic area function of ρ whereas in Lin's it is quadratic distance function of ρ
- Accuracy (A_ρ) includes ρ whereas in Lin's (C_b) it does not

$$A_\rho = \frac{4\times\left(\frac{s_{f(X_1,\dots,X_p)}}{s_Y}\right)-\rho\times\left[1+\left(\frac{s_{f(X_1,\dots,X_p)}}{s_Y}\right)^2\right]}{(2-\rho)\times\left[1+\left(\frac{s_{f(X_1,\dots,X_p)}}{s_Y}\right)^2\right]\times\left(\frac{\bar{Y}-\bar{f}(X_1,\dots,X_p)}{s_Y}\right)^2}$$

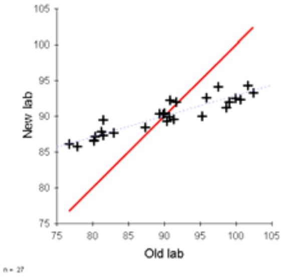
$$\gamma_\rho = \rho\times A_\rho$$





• Intercept = 0	• Intercept = 0
• Slope = 1	• Slope = 0.6
• r^2 = 1	• r^2 = 0.6
• C_b = 1 and A_ρ = 1	• C_b = 1 and A_ρ = 1
• ρ_c = 1 and γ_ρ = 1	• ρ_c = 0.6 and γ_ρ = 0.6
• r_2 = 1	• r_2 = 0.65

Comparison Lin's x Liao's



Liao (2003)

- Lin's CCC
  - C_b = 0.571
  - r_c = 0.527

  Liao's CCC
  - A_r (C_b) = 0.205
  - G_r (r_c) = 0.189

  Chinchilli's CCC
  - $GCCC_w$ = 0.179

Diverse Evaluation Measurements

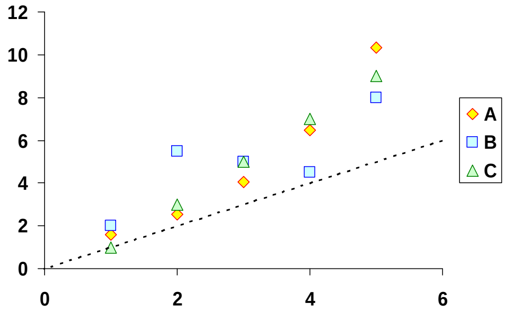
Mean Bias

- Likely to be the oldest and most used statistic to assess model accuracy

$$MB = \frac{\sum_{i=1}^{n}(Y_i - f(X_1,\dots,X_p)_i)}{n}$$

$$t_{MB} = \frac{MB}{\sqrt{\frac{\sum_{i=1}^{n}\left(\left(Y_i - f(X_1,\dots,X_p)_i\right)-MB\right)^2}{n\times(n-1)}}}$$

Which model has the lowest MB?



- All models (A, B, and C) have the same MB = 2
- *t*-test for Model A (exponential)
  - Assuming $\sigma_1 = \sigma_2$: *P* = 0.29
  - Assuming $\sigma_1 \neq \sigma_2$: *P* = 0.28
  - Assuming covariance: *P* = 0.09
- *t*-test for Model B
  - Assuming $\sigma_1 = \sigma_2$: *P* = 0.14
  - Assuming $\sigma_1 \neq \sigma_2$: *P* = 0.13
  - Assuming covariance: *P* = 0.02
- *t*-test for Model C (linear)
  - Assuming $\sigma_1 = \sigma_2$: *P* = 0.25
  - Assuming $\sigma_1 \neq \sigma_2$: *P* = 0.24
  - Assuming covariance: *P* = 0.05

Mean bias

- Has to be adjusted for covariance!
- Rejection rates of the H_0 hypothesis increases as correlated errors increase
- Cannot be used as the main statistics for model evaluation

Resistant r²

- Resistant means it is insensible to outliers or extreme points
- Uses the median instead of mean

$$r_r^2 = 1 - \left(\frac{\underset{i=1}{\overset{n}{\mathrm{M}}}\left(\left|Y_i - \hat{Y}_i\right|\right)}{\underset{i=1}{\overset{n}{\mathrm{M}}}\left(\left|Y_i - \overline{Y}\right|\right)} \right)^2$$

Modeling Efficiency

- Proportion of variation explained by the line Y = $f(X_{1,\ldots,}X_p)$
- Varies from [-∞ to 1]; MEF = 1 is better

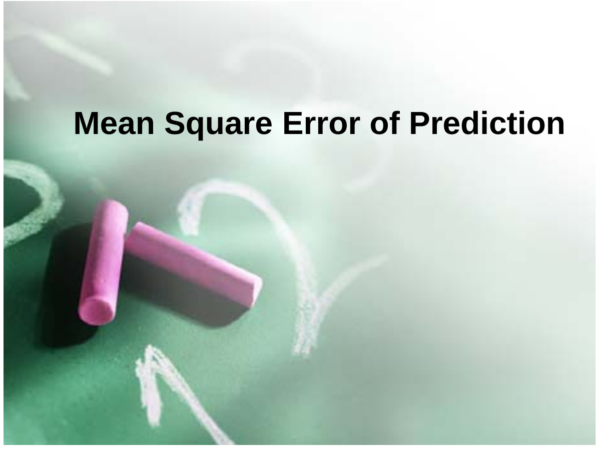
$$MEF = \frac{\left(\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 - \sum_{i=1}^{n}\left(Y_i - f(X_1,\ldots,X_p)_i\right)^2\right)}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \boxed{f(X_1,\ldots,X_p)_i}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}$$

$$r = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \boxed{\hat{Y}_i}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = \frac{s_{f(X_1,\ldots,X_p)Y}}{s_Y \times s_{f(X_1,\ldots,X_p)}}$$

Coefficient of Determination

- Ratio of total variance of observed data to the squared of the difference between model-predicted and mean of observed
- It is the proportion of the total variance of the observed values explained by the predicted data
- CD = 1 is better

$$CD = \frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(f(X_1,\ldots,X_p)_i - \overline{Y}\right)^2}$$



Mean Square Error of Prediction

MSEP x MSE

- MSE assesses the precision of the fitted linear regression using the difference between observed and regression-predicted values
- MSEP consists the difference between observed and model-predicted values

MSEP x MSE

$$MSE = \frac{\sum_{i=1}^{n}\left(Y_i - \boxed{\hat{Y}_i}\right)^2}{\boxed{n-2}}$$

$$MSEP = \frac{\sum_{i=1}^{n}\left(Y_i - \boxed{f(X_1,\ldots,X_p)_i}\right)^2}{\boxed{n}}$$

Limitations of MSEP

- Removes the negative sign
- Weights the deviation by their squares, thus giving more influence to larger data points

- Does not provide information about model precision

Decomposition of MSEP

- Work of Theil (1961)
- Expanded MSEP equation and solved for known linear measures of linear regression

$$MSEP = \frac{\sum_{i=1}^{n}\left(Y_i - f(X_1,\ldots,X_p)_i\right)^2}{n}$$

$$MSEP = \frac{\sum_{i=1}^{n}\left[\left(\bar{f}(X_1,\ldots,X_p) - \bar{Y}\right) + \boxed{f(X_1,\ldots,X_p)}_i - \bar{f}(X_1,\ldots,X_p)\big) - \big(\boxed{Y_i} - \bar{Y}\right)\right]^2}{n}$$

$$MSEP = \left(\bar{f}(X_1,\ldots,X_p) - \bar{Y}\right)^2 + s^2_{f(X_1,\ldots,X_p)} + s^2_Y - 2 \times r \times s_{f(X_1,\ldots,X_p)} \times s_Y$$

Understanding MSEP

$$MSEP_3 = \boxed{\left(f(X_1,\ldots,X_p) - Y\right)^2} + \boxed{s^2_{f(X_1,\ldots,X_p)} \times (1-b)^2} + \boxed{\left(1 - r^2\right) \times s^2_Y}$$

Mean Bias   Systematic Bias   Random

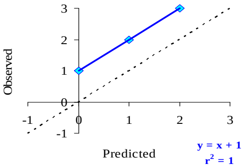
Inequality Proportions	Equations	Descriptions
U^M	$\left(\bar{f}(X_1,\ldots,X_p) - \bar{Y}\right)^2 / MSEP$	Mean bias
U^S	$\left(s_{f(X_1,\ldots,X_p)} - s_Y\right)^2 / MSEP$	Unequal variances
U^C	$2 \times (1-r) \times s_{f(X_1,\ldots,X_p)} \times s_Y / MSEP$	Incomplete (co)variation
U^R	$s^2_{f(X_1,\ldots,X_p)} \times (1-b)^2 / MSEP$	Systematic or slope bias
U^D	$\left(1 - r^2\right) \times s^2_Y / MSEP$	Random errors

[a] Note that $U^M + U^S + U^C = U^M + U^R + U^D = 1$

Understanding MSEP

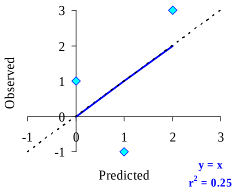
- Mean bias indicate the error in central tendency
- Systematic bias indicate how much the regression deviates from Y = X line, that means, errors due to regression
- Random errors indicate the unexplained variation that cannot be accounted for by the relationship

Case 2 - ↑ Precision ↓ Accuracy



- MSEP = 1
- Mean Bias = 100%

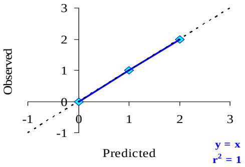
Case 3 - ↓ Precision ↑ Accuracy



- MSEP = 2
- Mean bias = 0
- Slope bias = 0
- Random = 100%

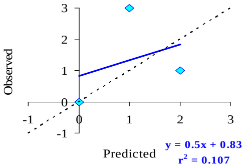
- Unequal variance = 33.3%
- Incomplete (co)variation = 66.7%

Case 4 - ↑ Precision ↑ Accuracy



- MSEP = 0

Case 1 - ↓ Precision ↓ Accuracy



- MSEP = 1.67
- Mean bias = 6.7%
- Slope bias = 10%
- Random = 83.3%

- Unequal variances = 11.1%
- Incomplete (co)variation = 82.2%

Nonparametric Analysis

Why nonparametric?

- One might be interested in the comparison of the ranking of real-observed values versus those predicted by models
  - Bull's EPD for efficiency
- More resilient to abnormalities of the data
  - Outliers and influential points

Nonparametric tests

- Spearman correlation is the linear correlation coefficient of the ranks

$$r_S = \frac{\sum_{i=1}^{n}\left(R_i - \bar{R}\right)\left(S_i - \bar{S}\right)}{\sqrt{\sum_{i=1}^{n}\left(R_i - \bar{R}\right)^2}\sqrt{\sum_{i=1}^{n}\left(S_i - \bar{S}\right)^2}}$$

- Kendall's coefficient measures the ordinal concordance of ½×n×(n-1) data points where a data point cannot be paired with itself

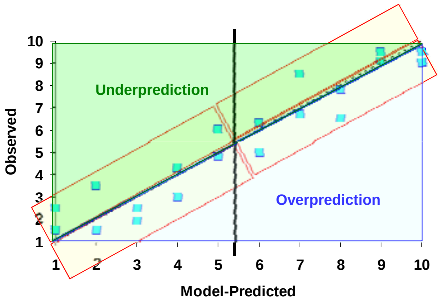
$$\tau = \frac{Concordant - Discordant}{\sqrt{Concordant + Discordant - ExtraY} \times \sqrt{Concordant + Discordant - ExtraX}}$$

Balance analysis

- Evaluates the balance of number of data points under- and overpredicted by the model above and below the observed and model-predicted mean

Model prediction	Observed or Model-Predicted Mean	
	Below	Above
Overpredicted	n_{11}	n_{12}
Underpredicted	n_{21}	n_{22}

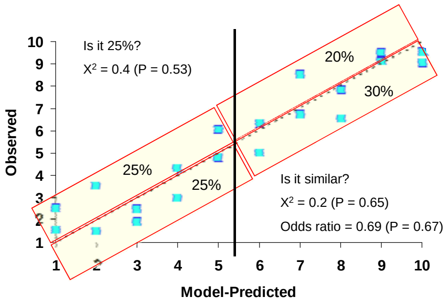
Balanced Analysis



What do we check in the BA?

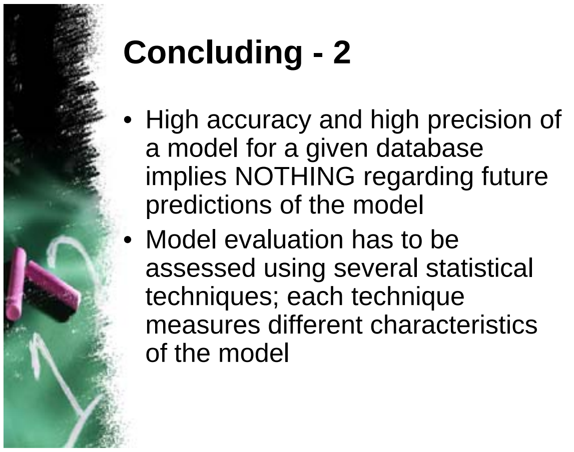
- Is the trend of under- or over-prediction similar?
- Is it similar below and above the mean?
- Use Chi^2 analysis to check if the number of points is not different
  - – Check if they are 25%
  - – Check if the distribution is similar

Balanced Analysis



Concluding - 1

- Acceptance of model wrongness is important to ensure more reliable and accurate models are developed
- Assessment of model adequacy requires a combination of several statistical analyses
- Usefulness of a model depends on the purpose it was developed for



Concluding - 2

- High accuracy and high precision of a model for a given database implies NOTHING regarding future predictions of the model
- Model evaluation has to be assessed using several statistical techniques; each technique measures different characteristics of the model