

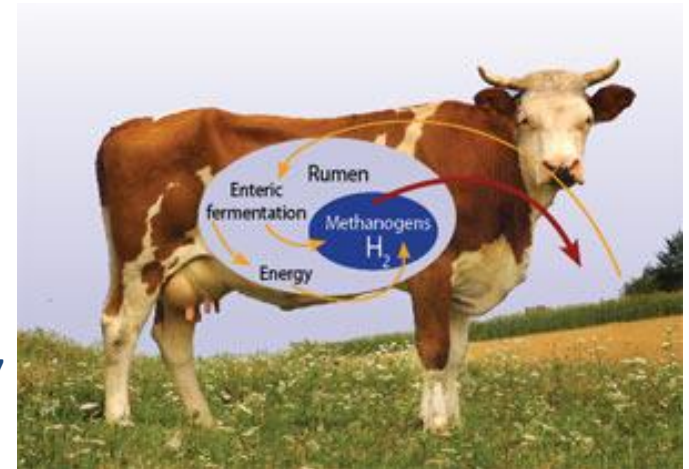
Model Evaluation

Ermias Kebreab

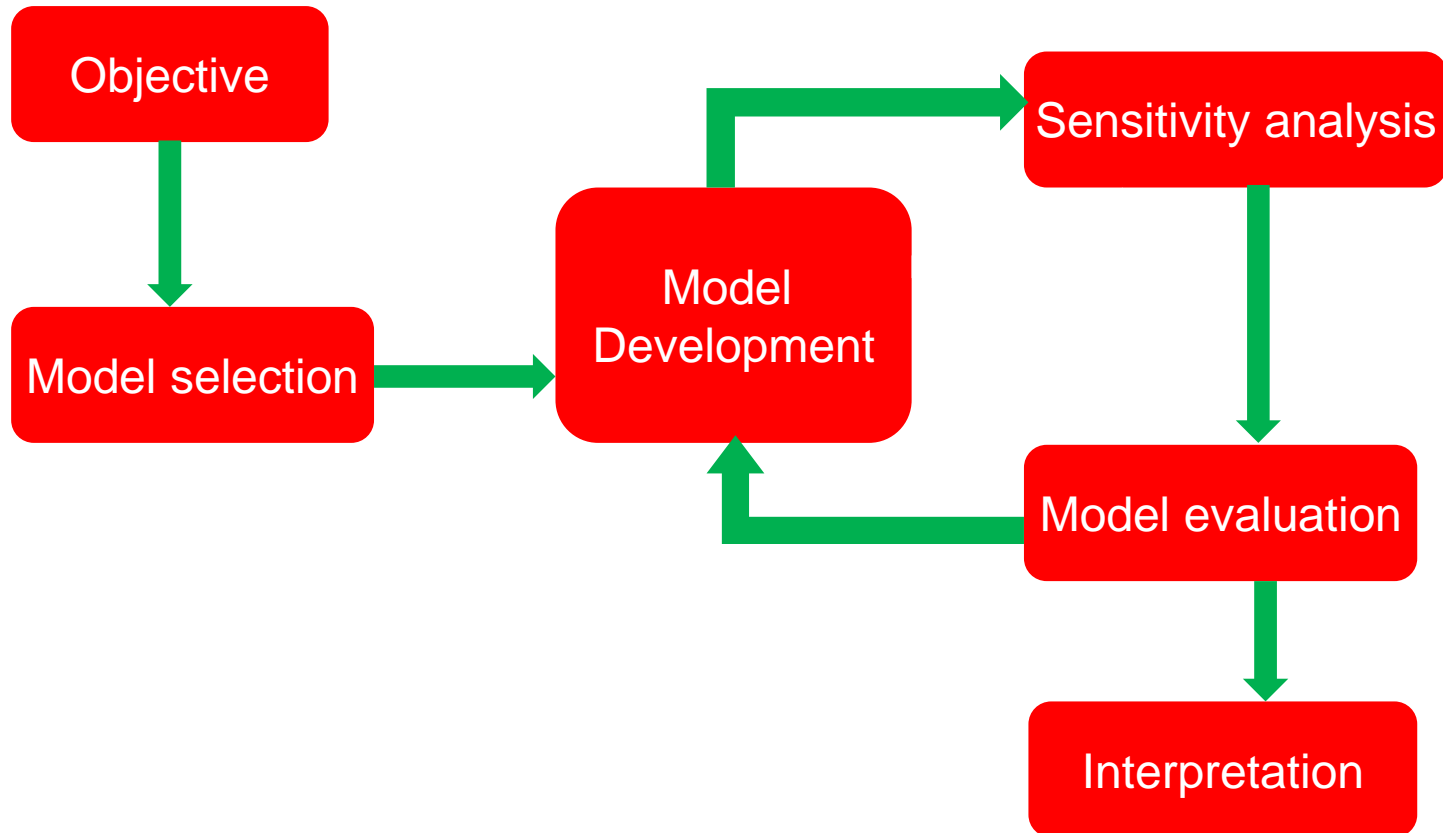
University of California, Davis



Pittsburgh, June 25, 2017



Modeling Process



Introduction

- Statistical measures of model performance commonly compare predictions with observations judged to be reliable
- Model evaluation indicates the level of accuracy and precision of model predictions
- Evaluates the credibility or reliability of a model by comparing it to real-world observations
- ‘Validation’ has been used to mean ‘Evaluation’ but no model can be validated completely because all of the infinite possibilities cannot be evaluated.

Model Evaluation Methods

There are three types of quantitative statistical model evaluation methods

- Standard Regression Statistics (SRS)
 - Determines strength of linear relationship.
E.g., Linear regression technique, analysis of residuals
- Error Index – quantifies deviation in obs. units
 - E.g., Mean square error of prediction (MSPE)
- Dimensionless – relative model evaluation
 - E.g. Concordance correlation coefficient (CCC), Nash-Sutcliffe Index (NSE)

SRS – Linear Regression

Linear regression:

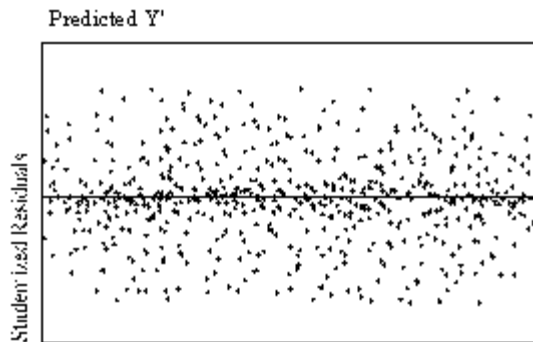
- Model predicted values are plotted on the X -axis
- A slope of 1 and intercept of 0 indicate perfect agreement
- Assumption – (1) all error variance is contained in predicted values and observed data are error free
 - Measured data is rarely, if at all, error free so care should be taken with this method

SRS - Linear Regression

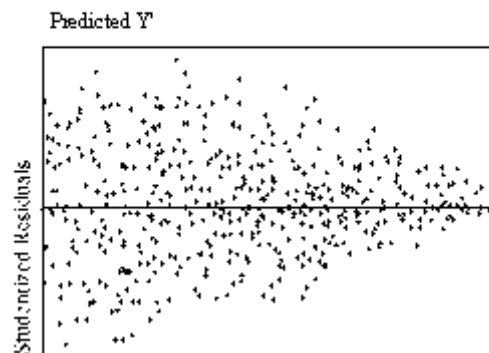
Assumptions (continued):

- (a) the Y-axis values have to be independent, random and homocedastic and
- (b) residuals are independent and identically distributed.

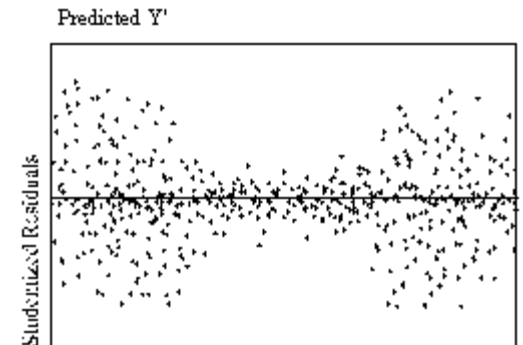
Homoscedasticity



Heteroscedasticity



Heteroscedasticity

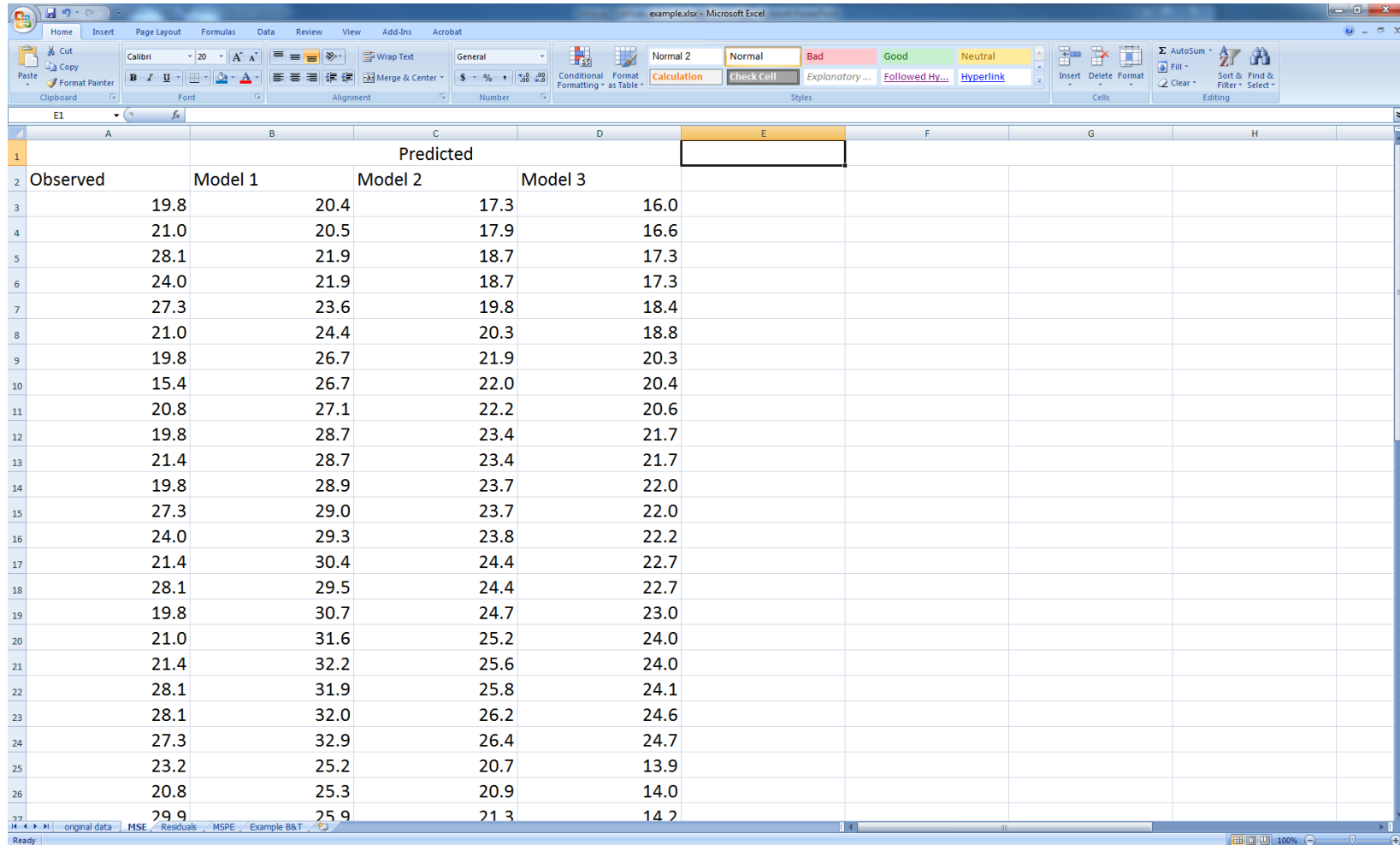


SRS - Correlation

Pearson correlation coefficient (r) or coefficient of determination (R^2):

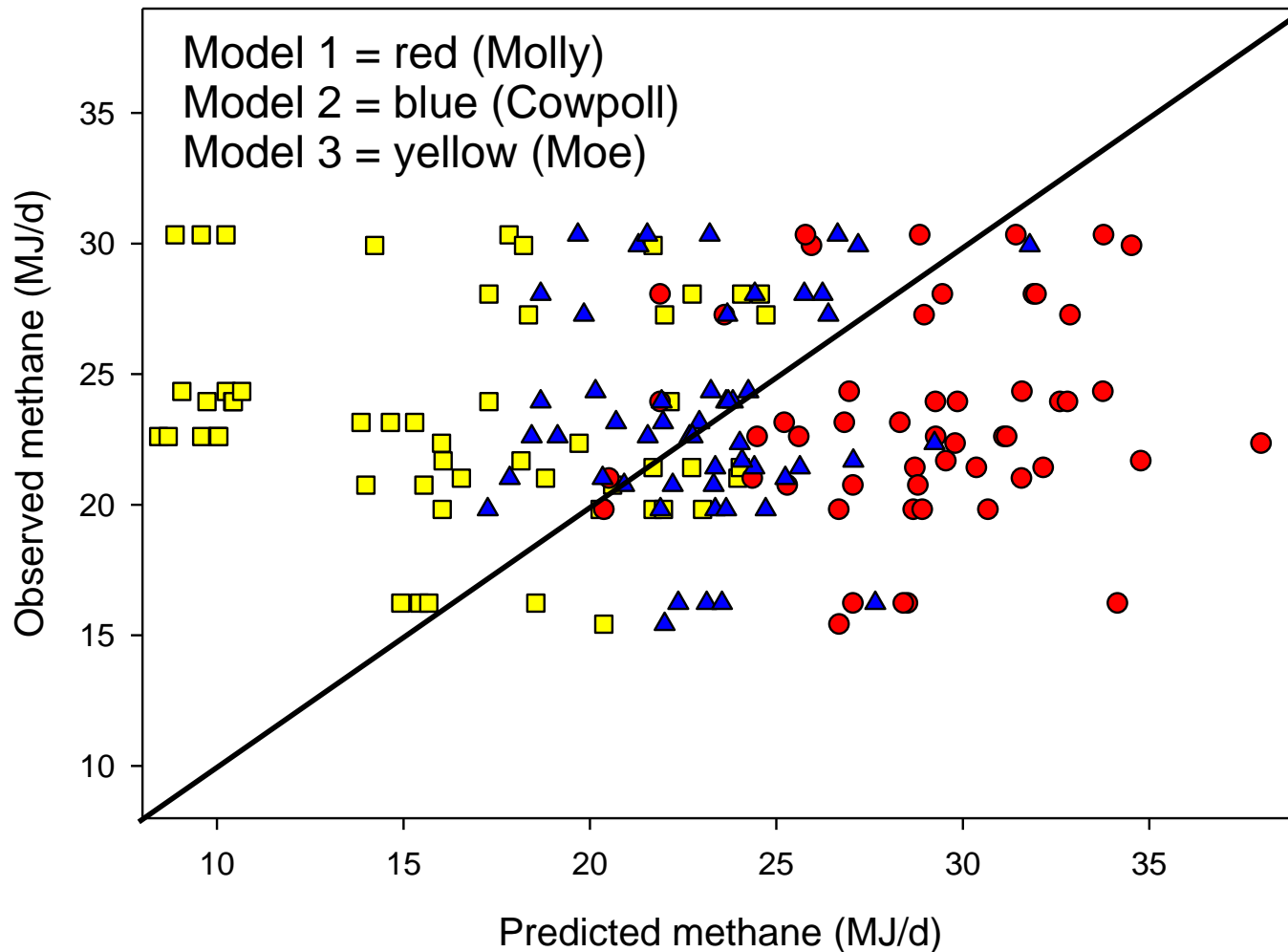
- Describe the degree of collinearity
- If $r=0$ no linear relationship exists, $r=1$ or -1 perfect positive or negative relationship
- R^2 described the proportion of variance in measured data explained by model
- Problem – Oversensitive to extreme values and insensitive to additive or proportional differences between predicted and observed values.

Linear Regression - Example



	A	B	C	D	E	F	G	H
1			Predicted					
2	Observed	Model 1	Model 2	Model 3				
3	19.8	20.4	17.3	16.0				
4	21.0	20.5	17.9	16.6				
5	28.1	21.9	18.7	17.3				
6	24.0	21.9	18.7	17.3				
7	27.3	23.6	19.8	18.4				
8	21.0	24.4	20.3	18.8				
9	19.8	26.7	21.9	20.3				
10	15.4	26.7	22.0	20.4				
11	20.8	27.1	22.2	20.6				
12	19.8	28.7	23.4	21.7				
13	21.4	28.7	23.4	21.7				
14	19.8	28.9	23.7	22.0				
15	27.3	29.0	23.7	22.0				
16	24.0	29.3	23.8	22.2				
17	21.4	30.4	24.4	22.7				
18	28.1	29.5	24.4	22.7				
19	19.8	30.7	24.7	23.0				
20	21.0	31.6	25.2	24.0				
21	21.4	32.2	25.6	24.0				
22	28.1	31.9	25.8	24.1				
23	28.1	32.0	26.2	24.6				
24	27.3	32.9	26.4	24.7				
25	23.2	25.2	20.7	13.9				
26	20.8	25.3	20.9	14.0				
27	29.9	25.9	21.3	14.2				

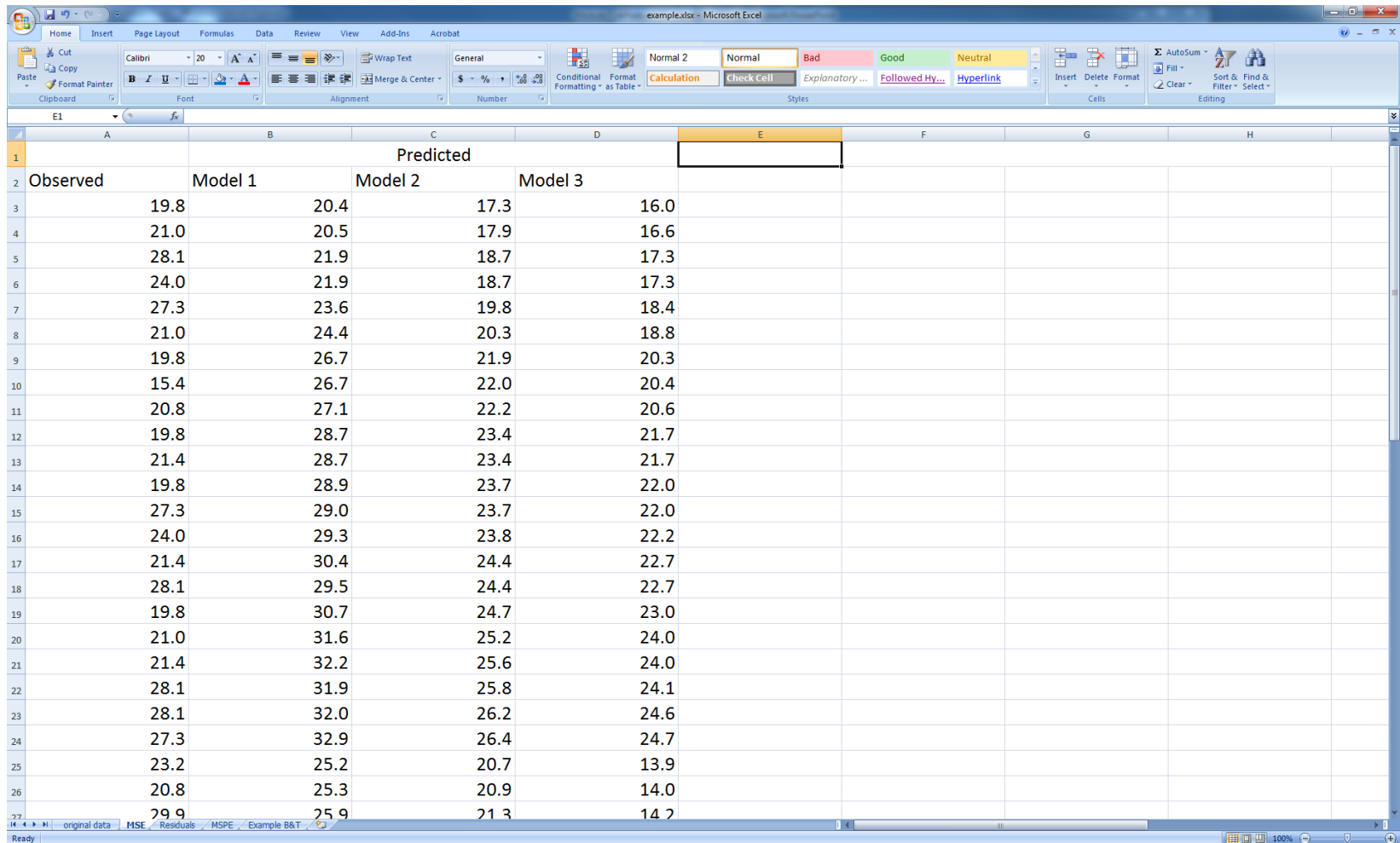
Linear Regression - Example



SRS - Analysis of Residuals

- The method involves regressing residuals (observed – predicted) against predicted or other model variables including model inputs, but not against observed
- Regressing residuals on observed values has been proved to be inadequate to properly identify biases with the simplest, most basic model (St Pierre, 2003)
- Residuals are not correlated with predictions and the slope of residuals regressed on predictions is zero if the model is unbiased

Analysis of Residuals - Data



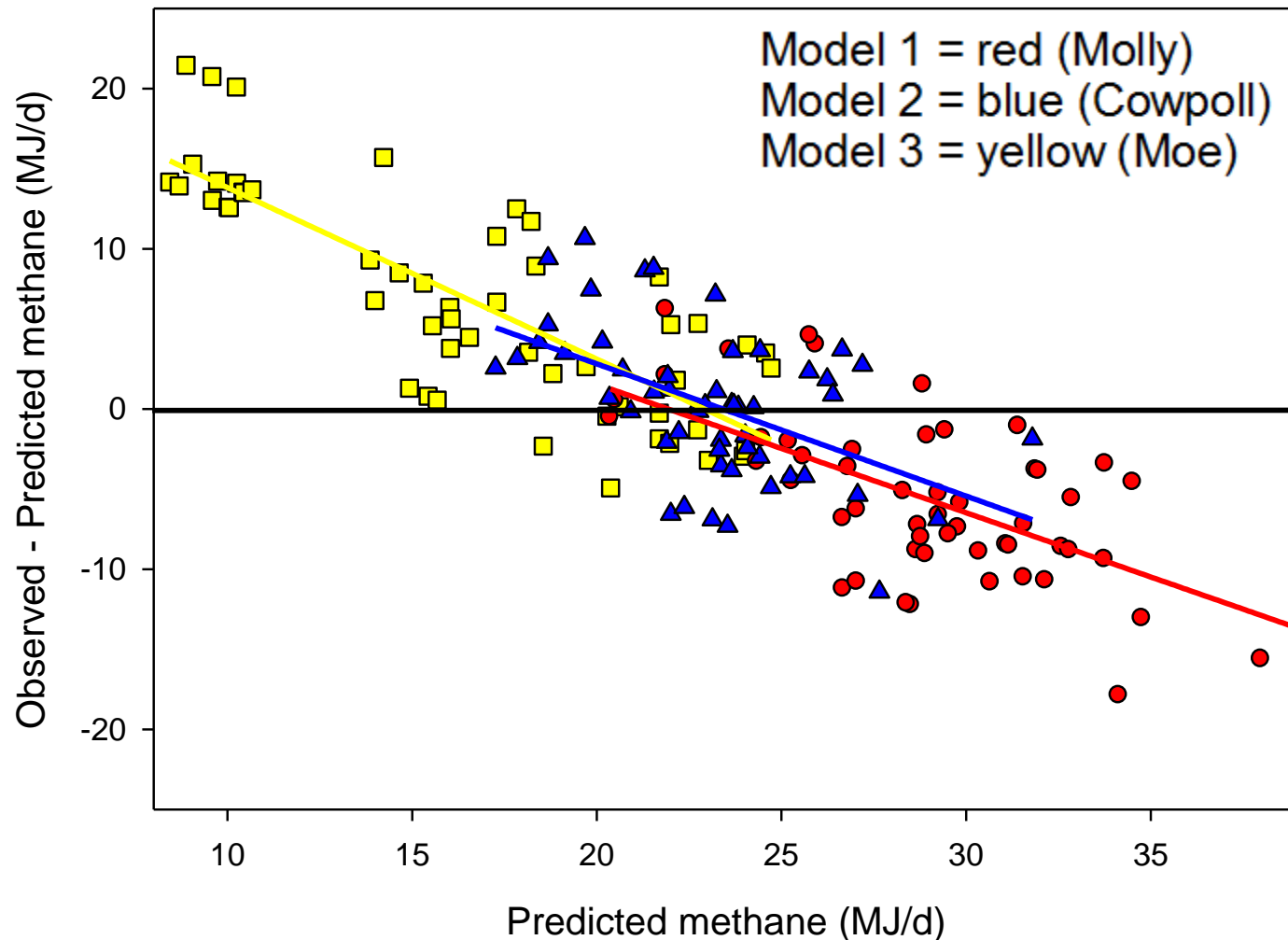
	A	B	C	D	E	F	G	H
			Predicted					
1	Observed	Model 1	Model 2	Model 3				
2	19.8	20.4	17.3	16.0				
3	21.0	20.5	17.9	16.6				
4	28.1	21.9	18.7	17.3				
5	24.0	21.9	18.7	17.3				
6	27.3	23.6	19.8	18.4				
7	21.0	24.4	20.3	18.8				
8	19.8	26.7	21.9	20.3				
9	15.4	26.7	22.0	20.4				
10	20.8	27.1	22.2	20.6				
11	19.8	28.7	23.4	21.7				
12	21.4	28.7	23.4	21.7				
13	19.8	28.9	23.7	22.0				
14	27.3	29.0	23.7	22.0				
15	24.0	29.3	23.8	22.2				
16	21.4	30.4	24.4	22.7				
17	28.1	29.5	24.4	22.7				
18	19.8	30.7	24.7	23.0				
19	21.0	31.6	25.2	24.0				
20	21.4	32.2	25.6	24.0				
21	28.1	31.9	25.8	24.1				
22	28.1	32.0	26.2	24.6				
23	27.3	32.9	26.4	24.7				
24	23.2	25.2	20.7	13.9				
25	20.8	25.3	20.9	14.0				
26	29.9	25.9	21.3	14.2				

Kebreab et al. 2008 JAS

Analysis of Residuals – Obs - Pred

example.xlsx - Microsoft Excel									
Model 1									
		Predicted				Observed-Predicted			
Observed	Model 1	Model 2	Model 3			Model 1	Model 2	Model 3	
19.82	20.37	17.26	16.04			-0.55	2.56	3.78	
21.02	20.51	17.85	16.55			0.51	3.17	4.47	
28.07	21.88	18.68	17.29			6.19	9.39	10.78	
23.95	21.88	18.68	17.29			2.07	5.27	6.66	
27.27	23.6	19.84	18.35			3.67	7.43	8.92	
21.02	24.35	20.34	18.81			-3.33	0.68	2.21	
19.82	26.67	21.89	20.27			-6.85	-2.07	-0.45	
15.43	26.68	22	20.37			-11.25	-6.57	-4.94	
20.75	27.05	22.22	20.61			-6.30	-1.47	0.14	
19.82	28.67	23.36	21.69			-8.85	-3.54	-1.87	
21.42	28.71	23.36	21.69			-7.29	-1.94	-0.27	
19.82	28.91	23.65	21.97			-9.09	-3.83	-2.15	
27.27	28.96	23.68	22			-1.69	3.59	5.27	
23.95	29.26	23.83	22.15			-5.31	0.12	1.80	
21.42	30.36	24.41	22.73			-8.94	-2.99	-1.31	
28.07	29.45	24.42	22.74			-1.38	3.65	5.33	
19.82	30.67	24.71	23.02			-10.85	-4.89	-3.20	
21.02	31.57	25.24	23.97			-10.55	-4.22	-2.95	
21.42	32.15	25.63	24.03			-10.73	-4.21	-2.61	
28.07	31.89	25.75	24.07			-3.82	2.32	4.00	
28.07	31.96	26.24	24.57			-3.89	1.83	3.50	
27.27	32.87	26.39	24.72			-5.60	0.88	2.55	
23.15	25.21	20.7	13.86			-2.06	2.45	9.29	
20.75	25.29	20.92	13.99			-4.54	-0.17	6.76	
29.93	25.94	21.3	14.23			3.99	8.63	15.70	

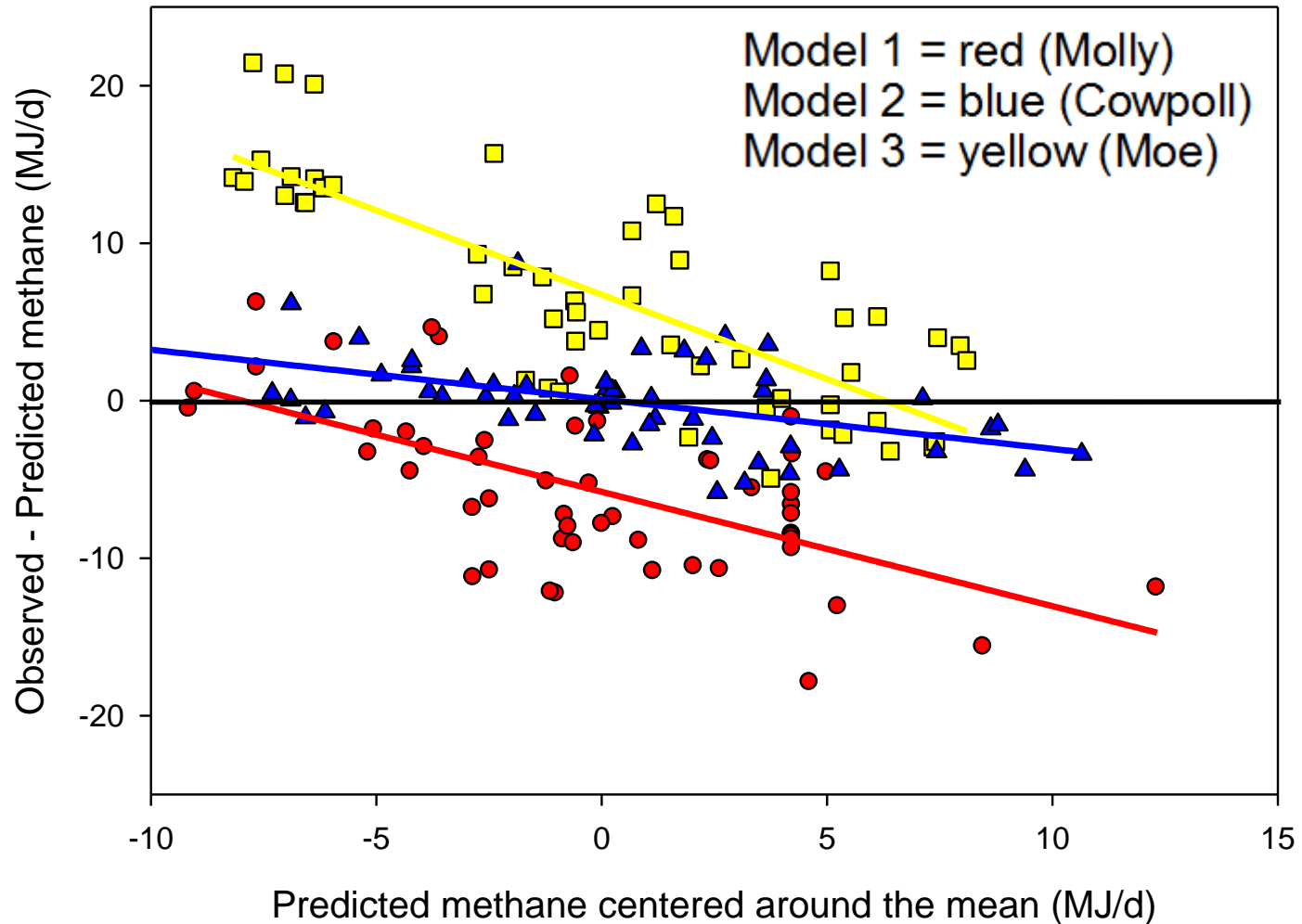
Analysis of Residuals - Graph



Analysis of Residuals – Centered Analysis

- Predicted values can be centered by subtracting the mean of all predicted values from each prediction
- This makes the slope and intercept estimates in the regression orthogonal and thus, independent (St Pierre, 2003)
- This allows for mean biases to be assessed using the intercepts of the regression equations, and the slopes to determine the presence of linear biases.

Analysis of Residuals - Centered



Analysis of Residuals - Equations

- Model 1

-5.78 (SE=0.56; $P < 0.001$) – 0.73 (SE=0.13; $P < 0.001$) (X-29.5) $r^2 = 0.38$

- Model 2

0.085 (SE=0.34; $P = 0.8$) – 0.32 (SE=0.11; $P < 0.001$) (X-23.1) $r^2 = 0.26$

- Model 3

6.73 (SE=0.56; $P < 0.001$) – 1.07 (SE=0.11; $P < 0.001$) (X-16.6) $r^2 = 0.65$

Error Index

Mean Square Error of Prediction (MSPE) and root mean square error (RMSPE):

- Valuable because they indicate error in the units (or squared units) of the observed value
- In general RMSPE values less than half of the SD of observed values may be considered a good agreement
- Therefore, RMSPE can be standardized by dividing it by SD of observed values (RSR).
- RSR varies from optimum of 0 to large positive values. The lower RSR the better the model performance

Mean Square Error of Prediction

An assessment of the error of prediction can be made by calculation of the root mean square error (MSPE):

$$MSEP = \frac{\sum_{i=1}^n (P_i - O_i)^2}{n}$$

where n is the number of runs and P_i and O_i are the predicted and observed values, respectively.

$$RSR = \frac{RMSEP}{STD_{obs}} = \frac{\sqrt{\frac{\sum_{i=1}^n (P_i - O_i)^2}{n}}}{\sqrt{\frac{\sum_{i=0}^n (O_i - \bar{O})^2}{n}}}$$

Mean Square Prediction Error

The MSPE can be decomposed into:

- error due to overall bias of prediction,
- error due to deviation of the regression slope from unity, and
- error due to the disturbance (random variation; Bibby and Toutenburg, 1977).

Root MSPE (RMSPE) which is in the same unit as observed value can be used as a measure of accuracy of prediction.

Mean Square Prediction Error

$$MSEP = \sum_{i=1}^n \left(P_i - O_i \right)^2 / n \quad [1]$$

$$MSEP = \frac{1}{n} \sum_i^n \left[\left(\bar{P} - \bar{O} \right) + \left(P_i - \bar{P} \right) - \left(O_i - \bar{O} \right) \right]^2 \quad [2]$$

where P_i is the predicted value and O_i is the observed value. This may be written as:

Mean Square Prediction Error

$$\text{MSEP} = \frac{1}{n} \sum_{i=1}^n \left[\left(\bar{P} - \bar{O} \right) + \left(P_i - \bar{P} \right) - \left(O_i - \bar{O} \right) \right]^2 \quad [2]$$

$$= \left(\bar{P} - \bar{O} \right)^2 + s_p^2 + s_o^2 - 2rs_p s_o, \quad [3]$$

$$\bar{P} = \frac{1}{n} \sum_{i=1}^n P_i \quad \bar{O} = \frac{1}{n} \sum_{i=1}^n O_i$$

$$s_p^2 = \frac{1}{n} \sum_{i=1}^n \left(P_i - \bar{P} \right)^2 \quad s_o^2 = \frac{1}{n} \sum_{i=1}^n \left(O_i - \bar{O} \right)^2$$

$$r = \frac{\frac{1}{n} \sum_{i=1}^n \left(P_i - \bar{P} \right) \left(O_i - \bar{O} \right)}{s_p s_o}$$



Mean Square Prediction Error

$$MSEP = \left(\bar{P} - \bar{O}\right)^2 + \left(s_p - rs_o\right)^2 + \left(1 - r^2\right)s_o^2$$

The decomposition of MSEP has some convenient interpretations. The first term is zero when $\bar{P} = \bar{O}$ i.e. when the average predicted value coincides with the average observed value.

Errors which lead to a positive value for this term may be called errors in central tendency (ECT) or mean bias.

Mean Square Prediction Error

$$MSEP = \left(\bar{P} - \bar{O}\right)^2 + \left(s_p - rs_o\right)^2 + \left(1 - r^2\right)s_o^2$$

- The second and the third term can be expressed as errors due to regression (ER) and errors due to disturbances (ED).
- The reason for this terminology is because the final term is the variation in O which is not accounted for by a least square regression of O and P – it is not the ‘unexplained variance’. It represents the portion of MSEP which cannot be eliminated by linear corrections of the predictions.

Mean Square Prediction Error

The penultimate term can be written as follows:

$$s_P^2 \left(1 - \frac{rS_O}{S_P} \right)^2$$

which measures the deviation of the least squares regression coefficient (rS_O/S_P) from one, the value it would have been if the predictions were completely accurate.

Root MSPE - Example

Let's work with a simple example first and then calculate the MSEP for previous example (i.e. methane emission models):

i	Observed	Predicted
1	10	5
2	2	-2
3	-7	-4
4	4	0
5	-3	1
6	6	4
7	4	7
8	-4	-2
9	-1	-2
10	3	2

Root MSPE - Example

i	Observed	Predicted	(obs - AveO)^2	(pred - AveP)^2	(Pred-Obs)^2
1	10	5	74.0	16.8	25
2	2	-2	0.4	8.4	16
3	-7	-4	70.6	24.0	9
4	4	0	6.8	0.8	16
5	-3	1	19.4	0.0	16
6	6	4	21.2	9.6	4
7	4	7	6.8	37.2	9
8	-4	-2	29.2	8.4	4
9	-1	-2	5.8	8.4	1
10	3	2	2.6	1.2	1
Sum	14	9	236.4	114.9	101
Average	1.4	0.9	23.64	11.49	10.1
SP	3.39			MSPE	10.1
SA	4.86			RMSPE	3.2
r	0.77				
				ECT	0.3
				ER	0.1
				ED	9.7
					10.1



MSPE – Methane Example

- Model 1

MSEP = 59.7 (RMSEP = 7.7 MJ/d; 26%)

ECT = 33.5 (56%), ER = 10.8, ED = 15.4 (26%)

- Model 2

MSEP = 21.4 (RMSEP = 4.6 MJ/d; 16%)

ECT = 0.1 (0.5%), ER = 5.54, ED = 15.8 (74%)

- Model 3

MSEP = 90.5 (RMSEP = 9.5 MJ/d; 32%)

ECT = 45.3 (50%), ER = 29.3, ED = 15.9 (18%)

Mean Square Prediction Error

$$MSEP = \sum_{i=1}^n \left(P_i - O_i \right)^2 / n \quad [1]$$

Problem: Squared differences magnify impact of outliers so Mean absolute error (MAE) can be used

$$MAE = \sum_{i=1}^n |P_i - O_i| / n$$

Dimensionless Evaluation Statistics

Index of agreement (d), Nash-Sutcliffe efficiency (NSE), Persistence model efficiency (PME), Prediction efficiency (Pe) and Concordance correlation coefficient or reproducibility index (CCC) are used to evaluate precision and accuracy of model predictions.

Accuracy measures how closely model-predicted values are to the true values. Model's ability to predict the right values

Precision measures how closely individual model-predicted values are within each other.

Model's ability to predict similar values consistently

34



Concordance correlation coefficient

CCC can be represented as a product of two components:

- A correlation coefficient estimate that measures precision (r) (Range 0 to 1, where 1 = perfect fit)
- A bias correction factor (C_b) that indicates how far the regression line deviates from the line of unity (Range from 0 to 1 and 1 indicates that no deviation from the line of unity has occurred).

Final thoughts

- Several model evaluation tools are available. Some such as k-fold also available for internal model evaluation
- When writing a modelling paper provide at least one dimensionless statistic and one error index statistic with additional information such as SD of measured data.

Practical

- Objective: Write model evaluation tool in R. Calculate an error index (MSPE) and dimensionless (CCC) based evaluation
- A simple data containing observed and predicted values is provided as the file: MSPE for R.csv
- Use the data to calculate MSEP, and its decomposition to ECT, ER and ED. Express it as a percentage of the total MSPE
- Calculate the RMSPE and express it as a percentage of the obs. mean. Calculate RSR
- Calculate CCC– does the result of RMSPE (or RSR) agree with CCC?