

VALIDATION / EVALUATION OF A MODEL



Validation or evaluation?

- Treat model as a scientific hypothesis
 - Hypothesis: does the model imitate the way the real world functions?
 - We want to validate or invalidate hypothesis - validation
- Treat model as engineering tool
 - The question is how good the tool is
 - We want to evaluate the quality of the model



The model as a scientific hypothesis



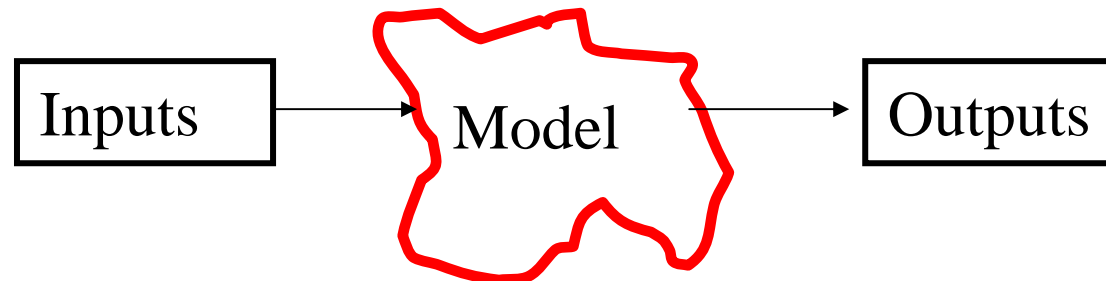
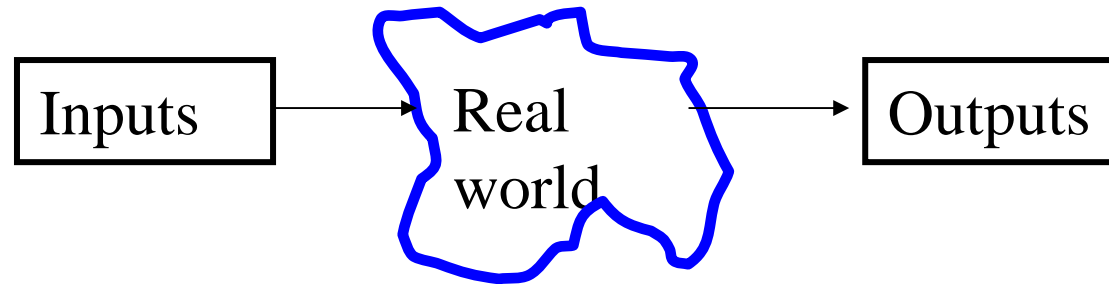
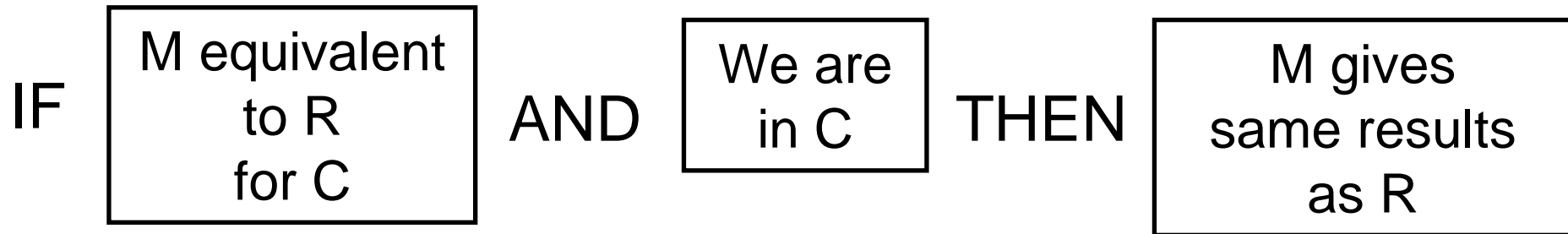
- Does the model behave in the same way as the real world for a set of conditions C.
 - “behaves like”: Each process gives results similar to measurements (within experimental error)

hypothesis

M
behaves like R
for C



- If the hypothesis is correct, then model predictions and observations will be the same



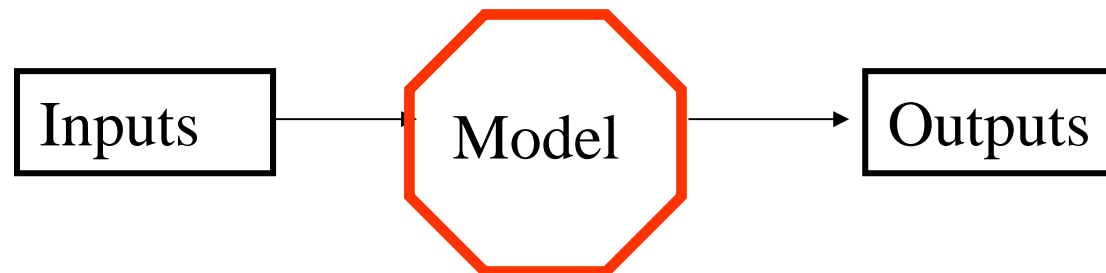
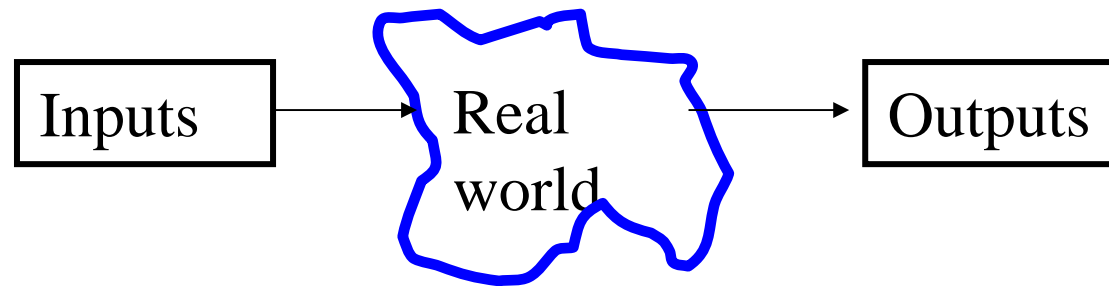
- What can we deduce from this syllogism?

IF M does NOT
give same
results as R AND We are
in C THEN M NOT
equivalent
to R for C

~~IF M gives
same results
as R AND We are
in C THEN M equivalent
To R
for C~~



- We can invalidate a model
- We cannot validate a model
 - The model may be right for the wrong reasons
 - e. g. even if aphid densities are correct, models of individual processes may e wrong



- Logically, we can't validate a model
- In any case, we know that all models are false
 - A model is a simplification of reality
 - e.g. aphid-ladybeetle model is extreme simplification, ignores other populations, plant growth, etc. etc. etc.
 - It is not meant to be exactly the same as reality



So a model as theory is useless?

- NO
 - Show that unlikely hypotheses are possible
 - Show that accepted hypotheses are wrong
 - Compare alternative hypotheses

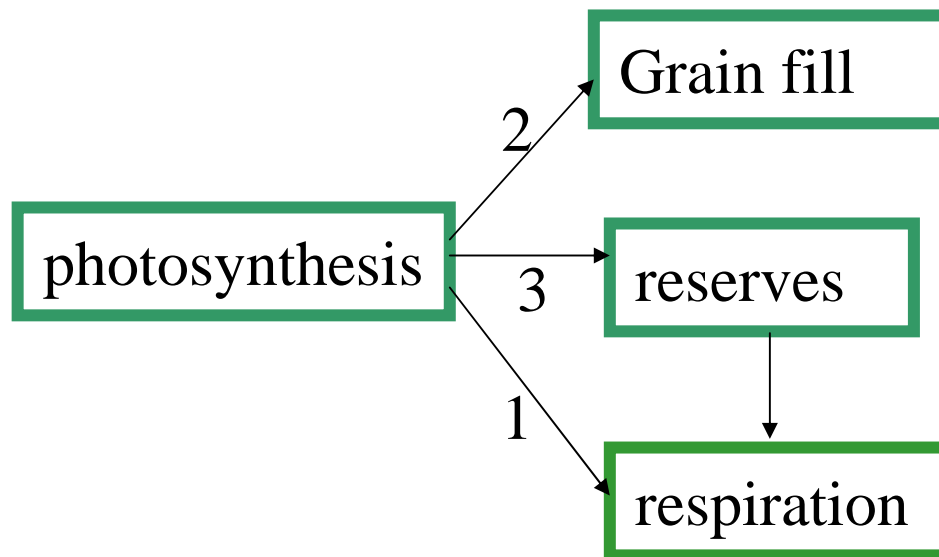


Example of comparison of hypotheses

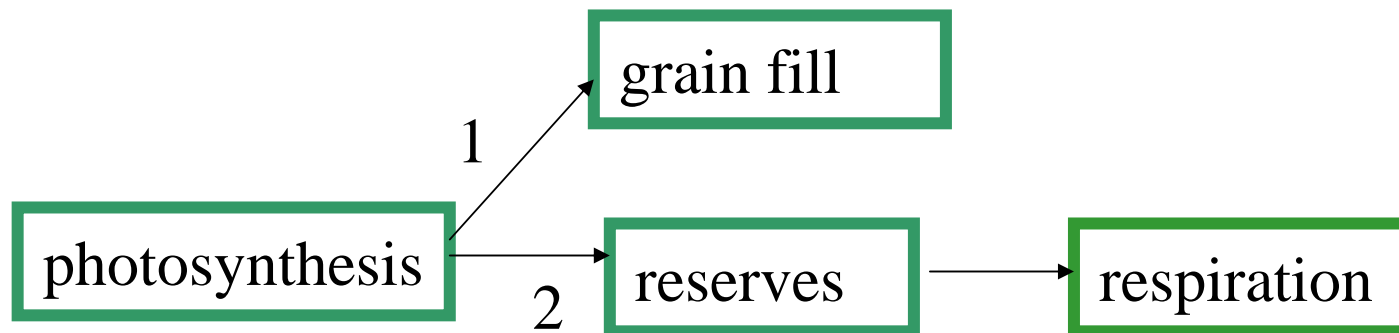
- Respiration of wheat during grain filling. Does the C for respiration come directly from photosynthesis, or from reserves?



- Hypothesis 1 : C for respiration comes from photosynthesis if possible.

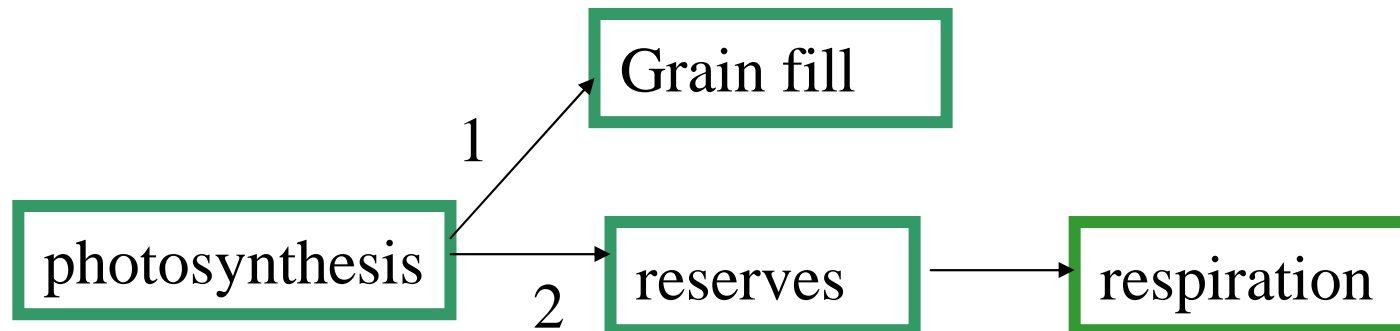
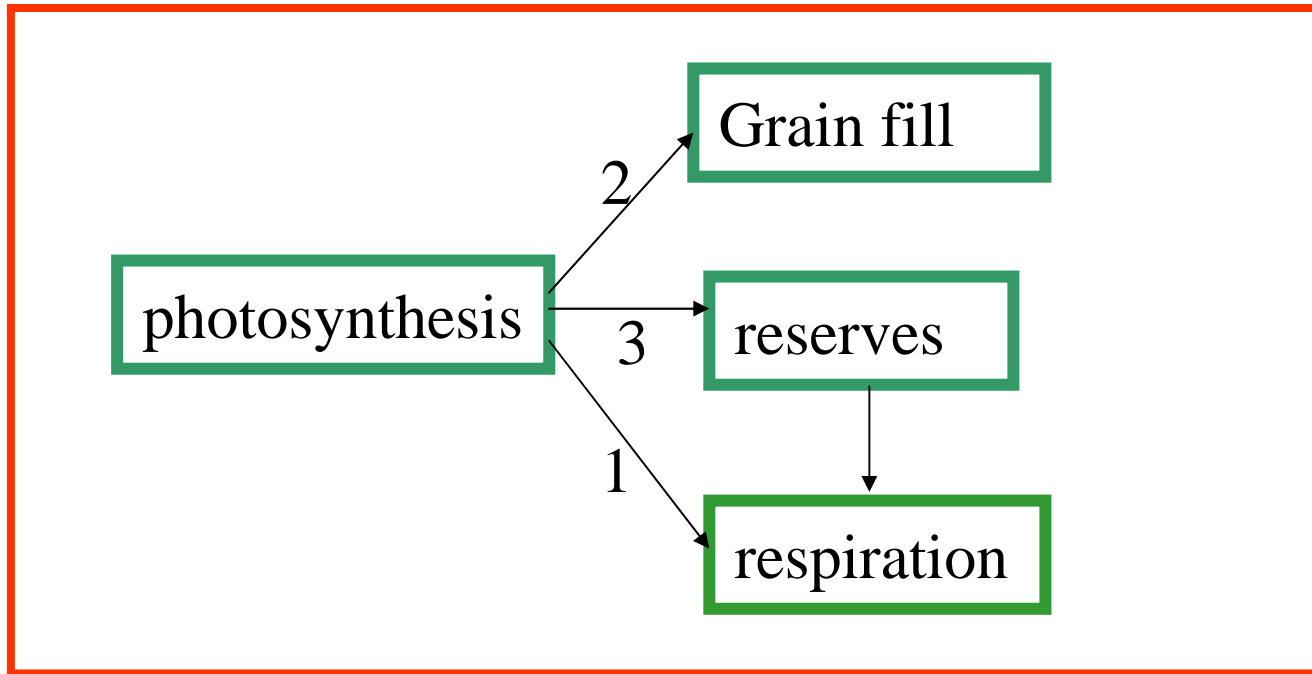


- Hypothesis 2: C for respiration comes from reserves



- Do experiments using pulses of ^{14}C marked air. Measure ^{14}C concentrations in grain and reserves.
- Develop 2 models, corresponding to above 2 hypotheses. Models predict ^{14}C concentrations in grain and reserves.
- Model based on hypothesis 1 is more consistent with data.





Conclusions?

- Hypothesis 1 is more apt to reproduce observed results.
- We don't accept it as exactly true, but as better working hypothesis
 - So this is like engineering model?
 - Yes and no.
 - Yes because we look at how well model reproduces results.
 - No because we have drawn conclusions about mechanisms.



Engineering model



Evaluation

- We don't treat the model as a hypothesis but as a tool.
- We want it to reproduce important aspects of reality (e. g. predict yield, predict response to fertilizer)
- How well does model do that? That's what we evaluate.



The role of evaluation

- At the start of a modelling project
 - Define objectives and therefore evaluation criteria
- During the project
 - To choose between alternatives, evaluate each
 - Evaluation may give indication of how to improve model
- At the end of the project (or of a cycle)
 - Evaluation provides measure of quality of model



The practice of evaluation

- Compare model to data, measure model quality
- Estimate how well model will predict for new cases
- Evaluation applies to all models. Both simple linear models and complex dynamic system models.
 - So we can use simple linear models to illustrate
 - We will point out specific aspects of dynamic system models

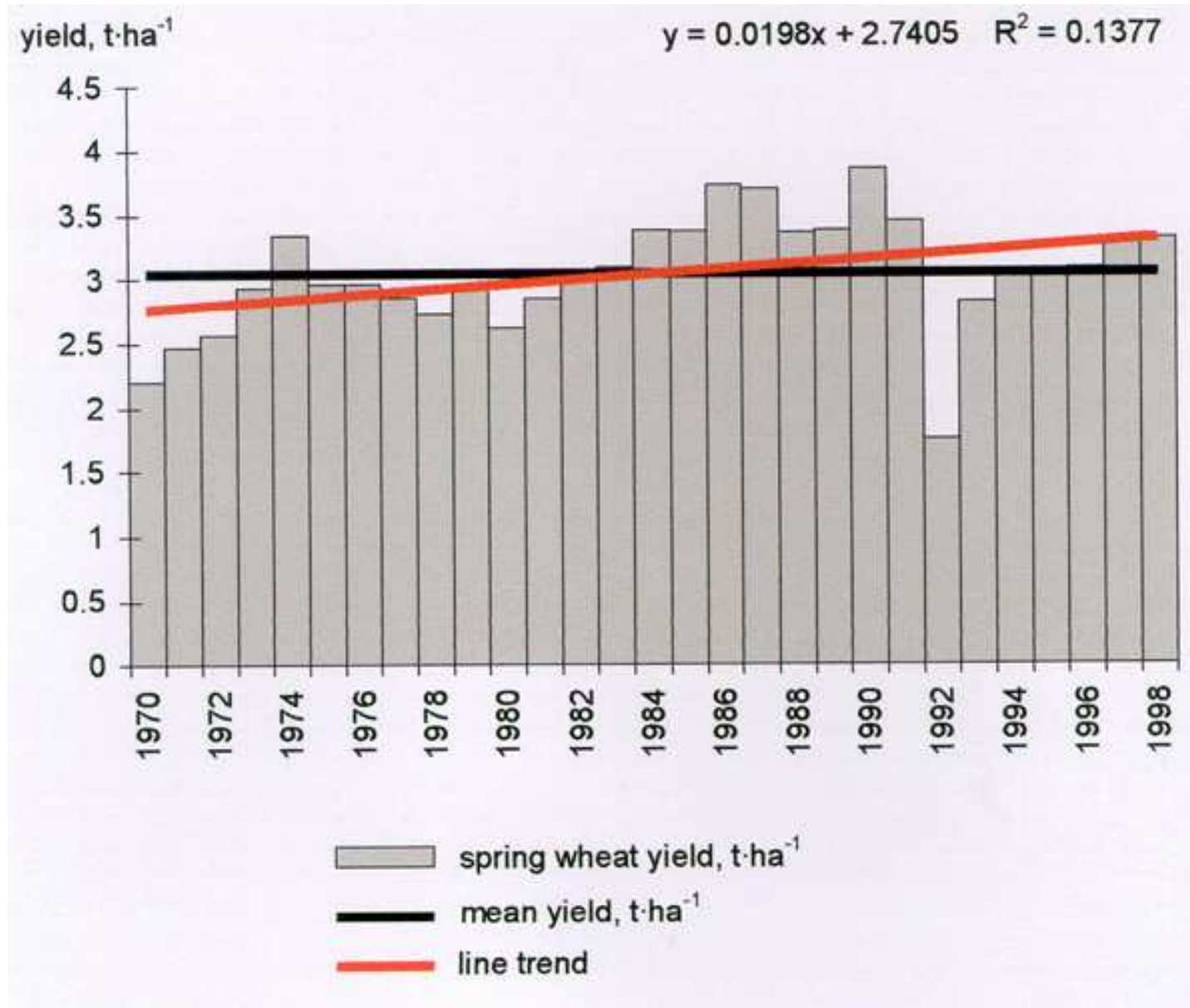


Examples of data and models

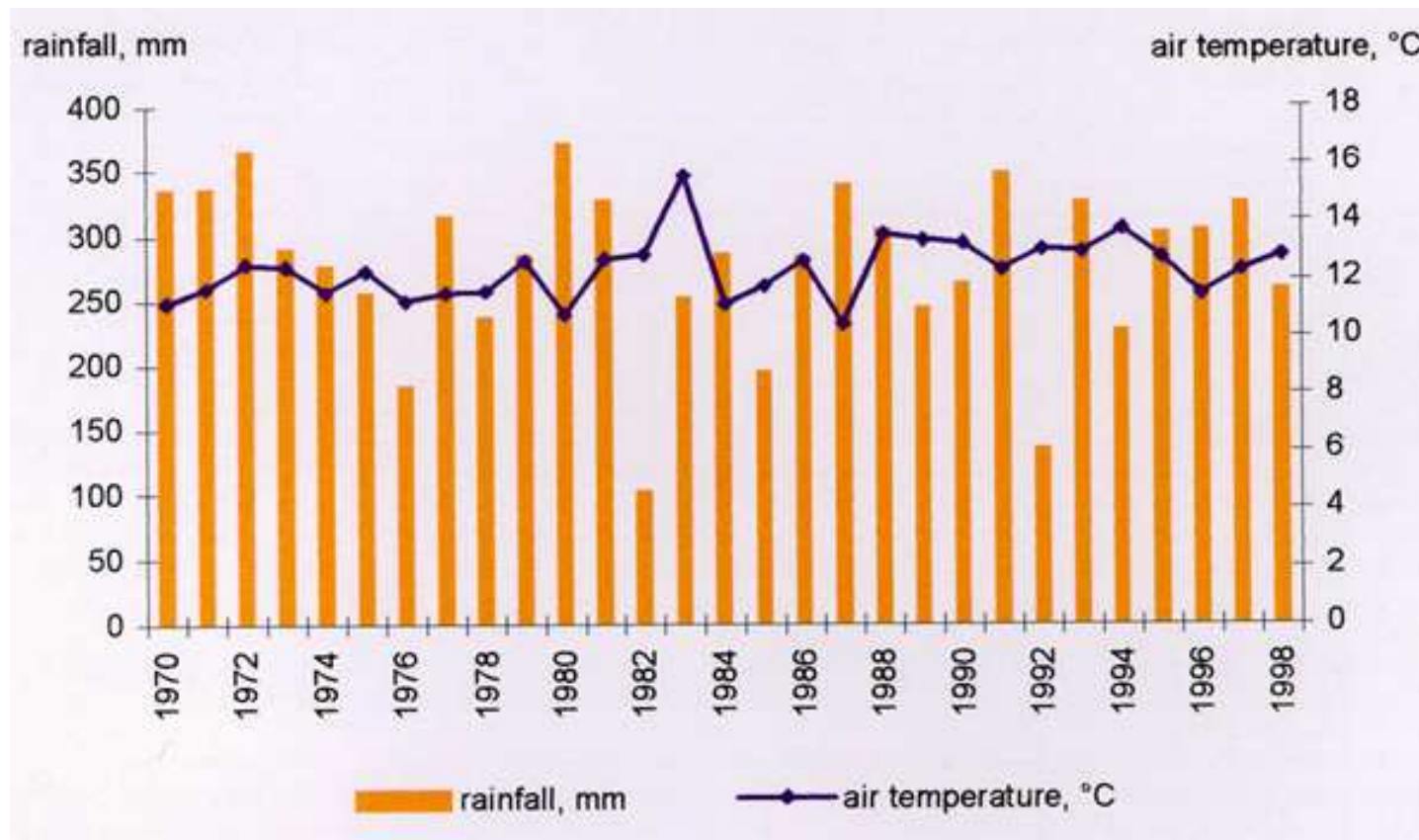
- Dynamic system models.
 - We have seen several examples
 - Dynamic system model for corn. Model is used to compare different irrigation strategies. We don't present model, just measured and calculated values.
- Static models
 - Predicting yield in a Polish region. To show that same problems arise for static and dynamic models.
 - An invented example, used to illustrate methods.



Spring wheat yield in the Zachodnie Pomorze Province 1970-1998.



Rainfall and mean air temperature from March 11 to August 20 in the Zachodnie Pomorze Province 1970-1998.



Relationship between spring wheat yield (t·ha⁻¹) in the Zachodnie Pomorze Province and weather components 1970-1998

Regression equations R²

Till April 30

$$y = 1.40129 + 0.2396x_1 - 0.0448x_2 + 0.130222x_3 - 0.002306x_4 \quad 67.84$$

Till May 31

$$y = 2.25358 + 0.2837x_1 - 0.01490x_2 + 0.15782x_3 - 0.01039x_5 - 0.020279x_6 \\ 79.31$$

Till June 30

$$y = 3.77033 + 0.2557x_1 - 0.01218x_2 + 0.13753x_3 - 0.01265x_5 - 0.0235x_7 - 0.00273x_8 \\ 88.85$$

Till July 31

$$y = 3.61260 + 0.02643x_1 - 0.01126x_2 + 0.13129x_3 - 0.01291x_5 - 0.02982x_7 + 0.03687x_9 - \\ 0.00215x_{10} + 0.03107x_{11} \quad 90.5$$



Artificial data and model

- Invent formula for generating data. This is « real world ». Generate sample of 8 data values. Those are « measurements ».

$$Y = \theta_1 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5 + \varepsilon$$

- Model has same form (linear model with 5 explanatory variables)
- Use data to estimate model parameters
- Estimate 0,2,3,4, or 6 parameters. Others have default values that are different than true values.

$$\hat{Y} = \hat{\theta}_1 + \hat{\theta}_1 x_1 + \hat{\theta}_2 x_2 + \hat{\theta}_3 x_3 + \hat{\theta}_4 x_4 + \hat{\theta}_5 x_5$$



Data for artificial example

$$f(X ; \hat{\theta}) = \hat{\theta}^{(0)} + \hat{\theta}^{(1)} * x^{(1)} + \hat{\theta}^{(2)} * x^{(2)} + \hat{\theta}^{(3)} * x^{(3)} + \hat{\theta}^{(4)} * x^{(4)} + \hat{\theta}^{(5)} * x^{(5)}$$

$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$	Y	\hat{Y}
-1.6339	0.7977	0.4416	-0.4463	-0.4728	-9.3896	-9.3144
-0.9485	1.0700	0.5047	0.5308	-0.3257	-3.2312	-3.0994
-0.2512	0.1952	0.5099	0.8226	0.4495	0.3732	0.7236
0.3789	1.0193	-0.2185	0.8163	-1.9263	7.1024	6.1808
0.1464	1.1373	1.0657	1.6325	-0.5528	5.8245	5.4485
-1.1984	-1.7925	0.3530	-0.2601	0.2617	-11.2130	-11.8425
-0.9720	-0.1533	0.1113	1.1251	0.1019	-5.8110	-5.9035
0.3931	-1.2031	2.0132	-0.7947	-1.4396	2.8158	0.4690

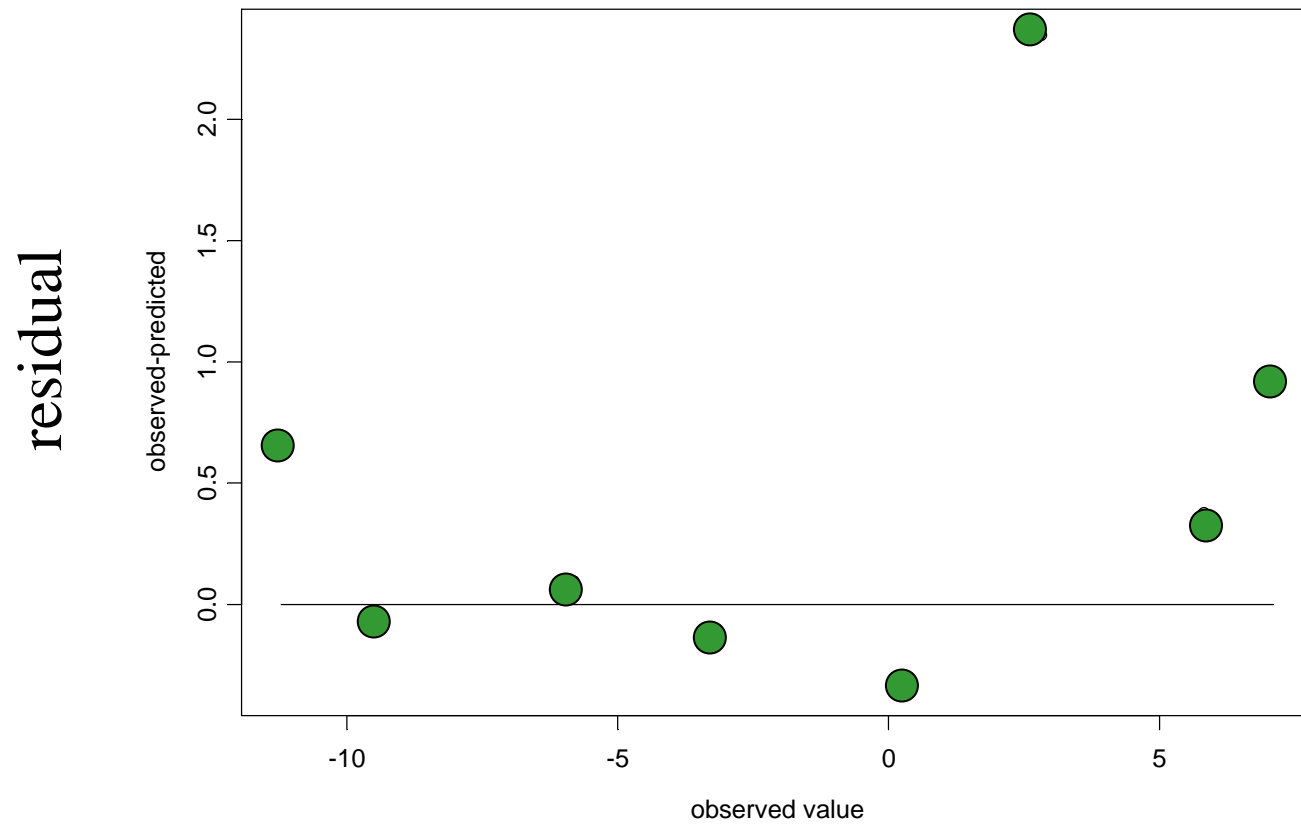


Graphs



Artificial model

Residuals (observed – calculated)

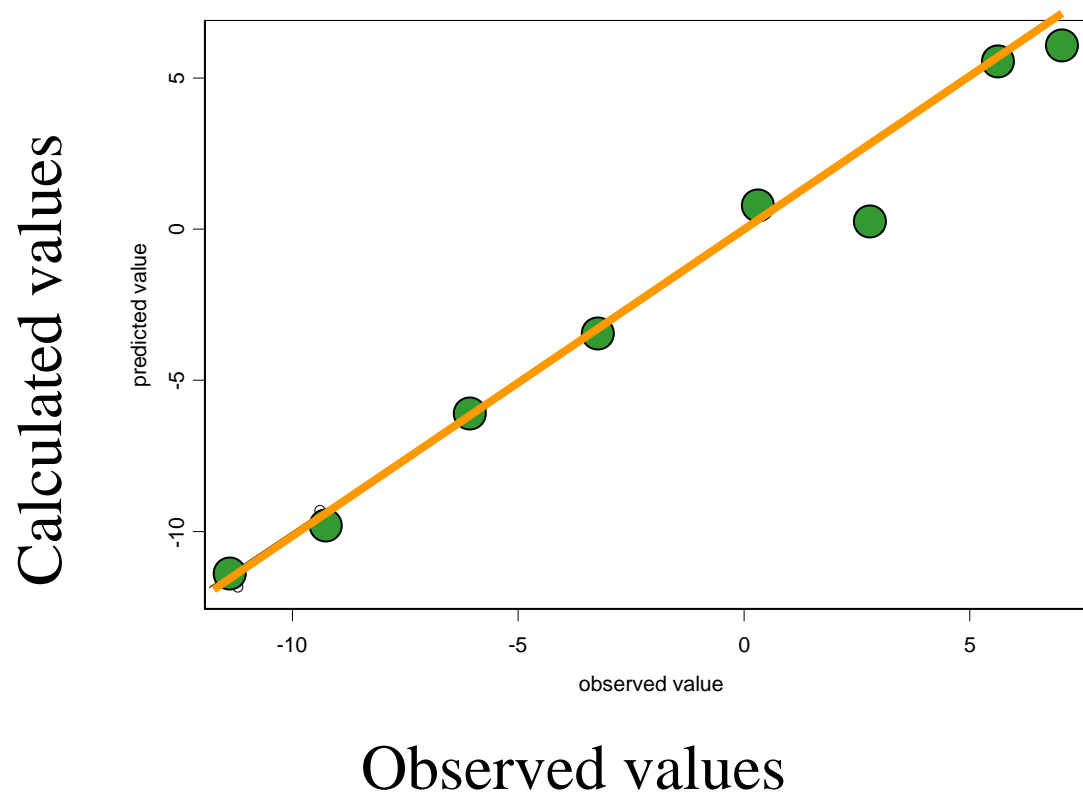


Observed values



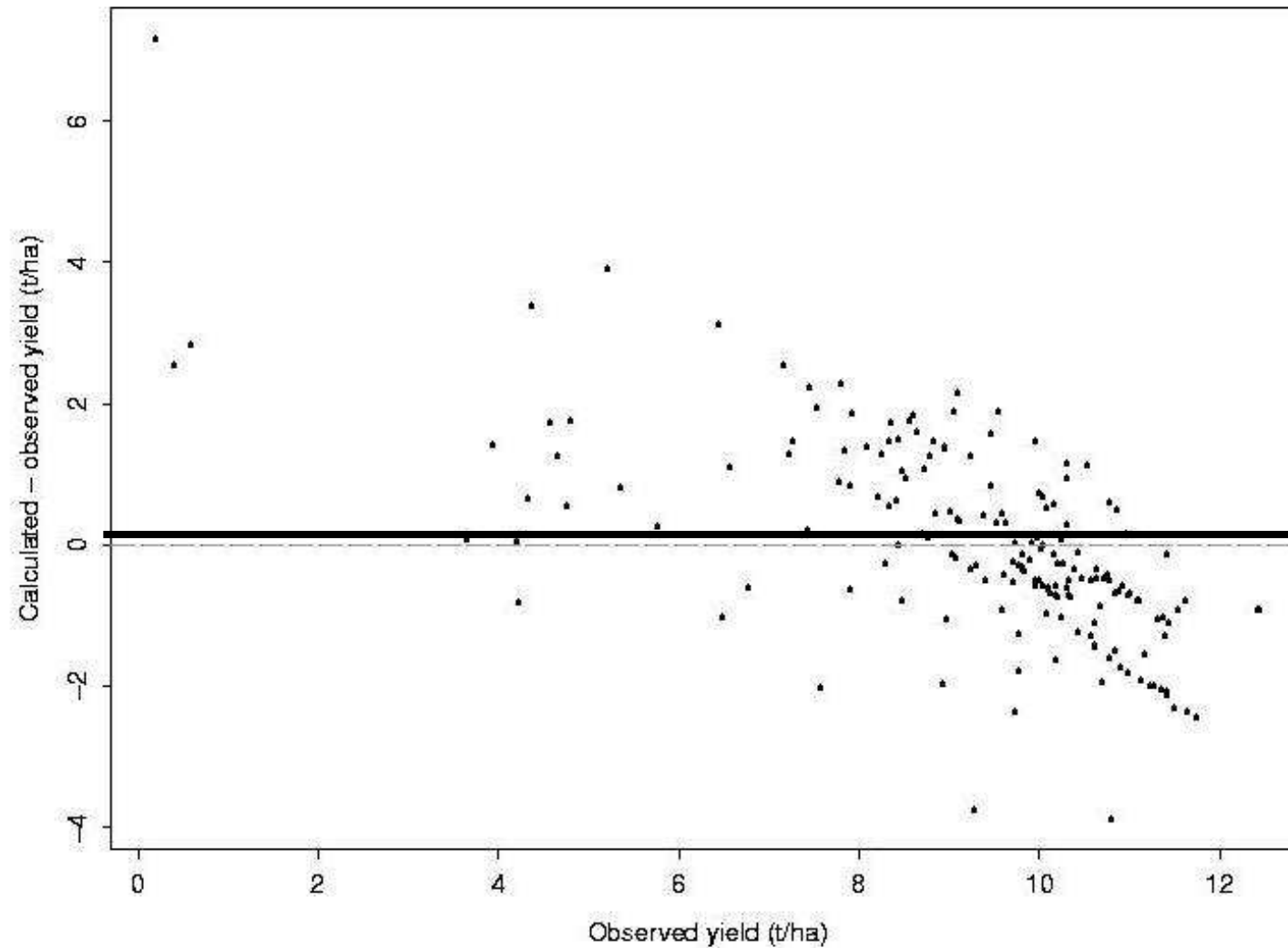
Artificial example

Calculated vs observed values



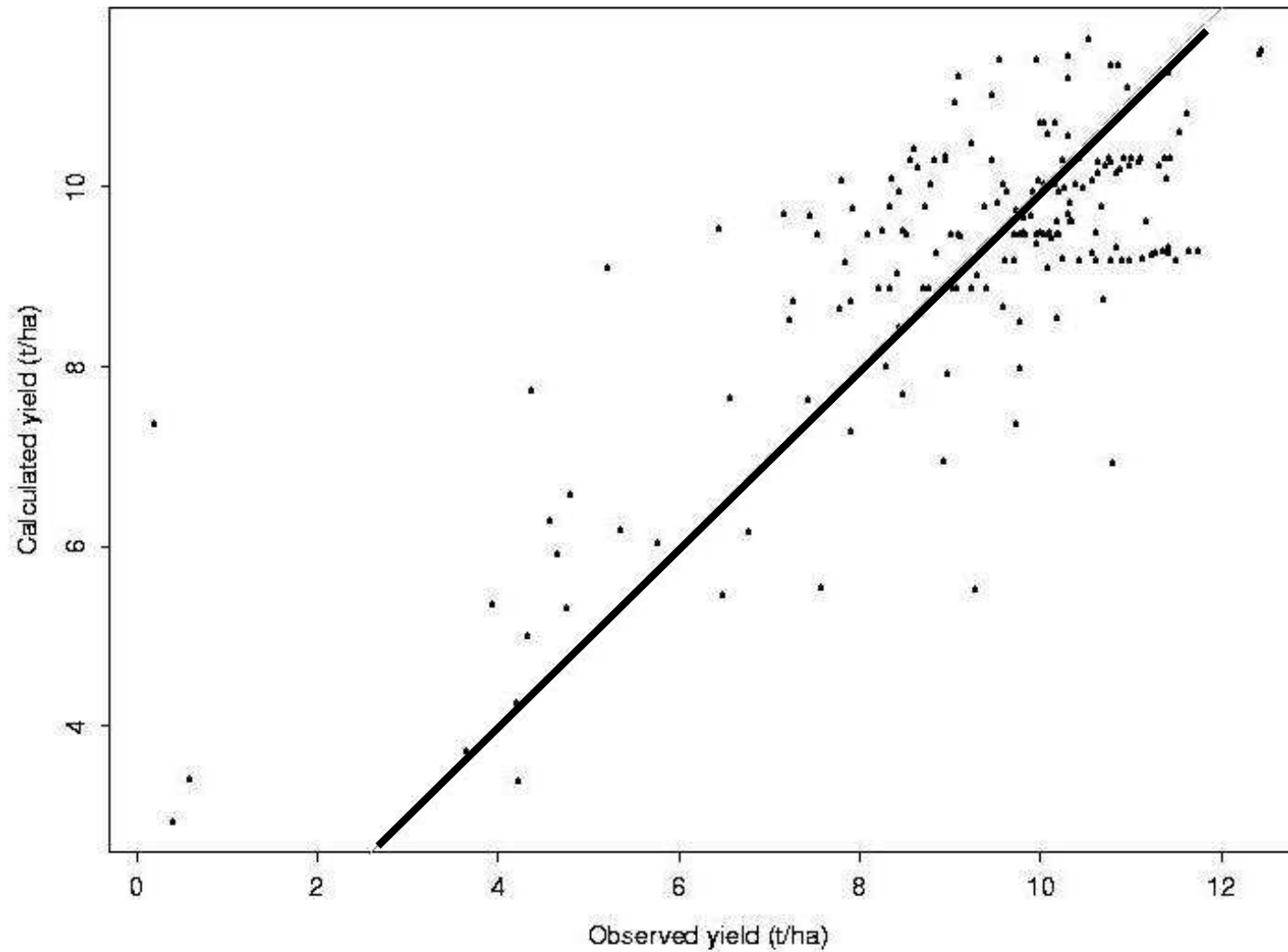
Corn model. Residuals

Model errors



Corn model. Observed vs calculated

Calculated versus observed yield



Measures of error

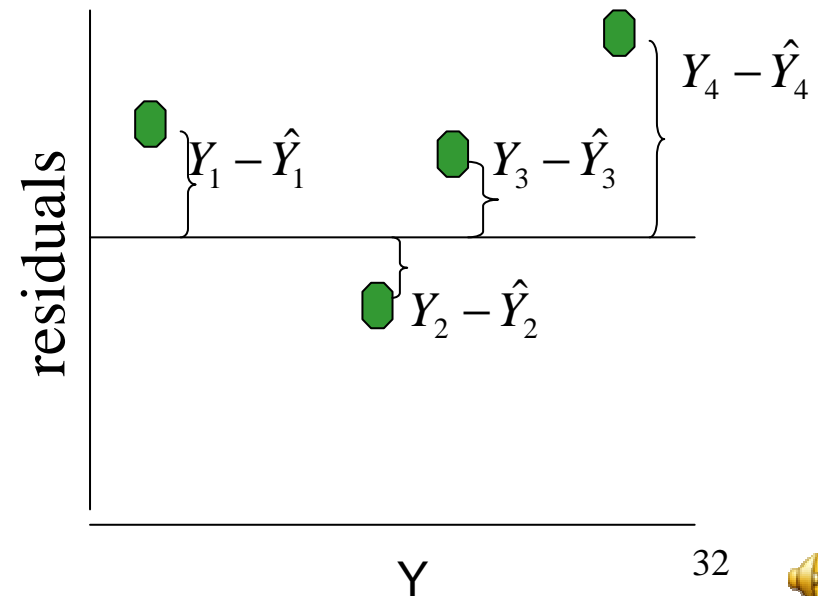
- Summarize information about differences between measured and calculated values



MSE

- Mean Squared Error (the most common measure)

$$MSE = (1/N) \sum_{i=1}^N (Y_i - \hat{Y}_i)^2$$



RMSE and MAE

name	equation	value for example
mean squared error	$MSE = \frac{\sum (Y_i - \hat{Y}_i)^2}{N}$	0.88 (t/ha) ²
root mean squared error	$RMSE = \sqrt{MSE}$	0.94 (t/ha)
mean absolute error	$MAE = \frac{\sum Y_i - \hat{Y}_i }{N}$	0.62 (t/ha)



R squared

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

- R^2 if model is perfect?
 - $R^2=1$
 - Can R^2 be > 1 ?
 - No.
- R^2 if model is just average of observed values?
 - $R^2=0$
 - Can R^2 be less than 0?
 - Yes, for complex models
- This criterion also called efficiency
- For yield example, $R^2=0.98$

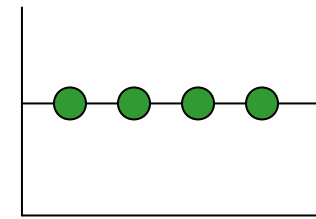
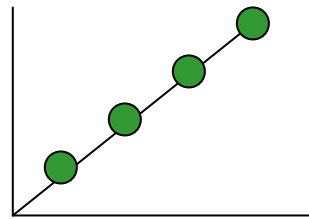


R² and graphs

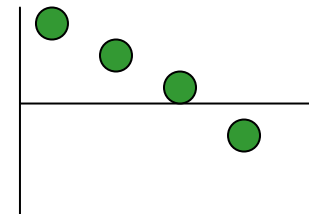
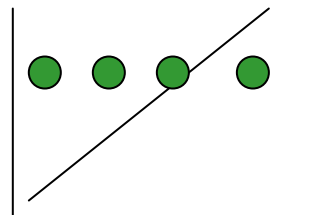
Calculated vs.
observed

Residuals

$$R^2=1 \quad Y_i=\hat{Y}_i$$



$$R^2=0 \quad Y_i \neq \hat{Y}_i$$



Components of model error

- To better understand origin of error
- May give ideas of how to improve model



Components of MSE

MSE=bias² + variability difference + remainder

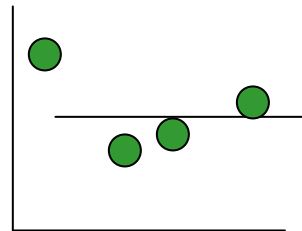


Bias term

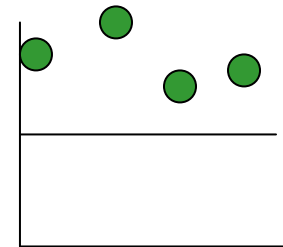
residuals

$$\text{bias}^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{N}$$

Small bias



Large bias



- If bias is large
 - Left out or underestimated a factor that systematically increases or decreases response
 - For example, underestimated harvest index (relation of yield to total biomass)
 - In example, $\text{bias}^2=0.23$ ($\text{MSE}=0.88$)

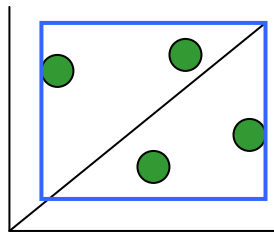


Variability difference

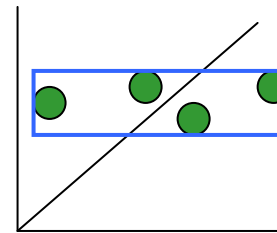
$$\text{variability difference} = (\sigma_Y - \sigma_{\hat{Y}})^2$$

Calculated vs observed values

small



large



- If SDSD is large
 - Left out or underestimated a factor that sometimes increases or decreases response
 - For example, effect of water stress
 - In example variability difference = 0.06 (MSE=0.88)

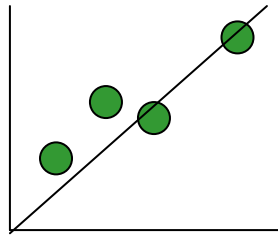


Remainder

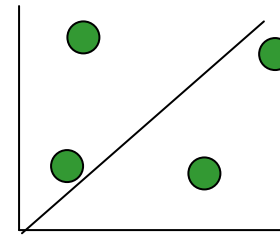
$$\text{remainder} = 2\sigma_Y \sigma_{\hat{Y}} (1 - \text{correlation coefficient})$$

Calculated vs observed values

small



large



- If LCS is large
 - Sorry, error is in details of model
 - In example, remainder=0.59 (MSE=0.88)



Criteria for model comparison

- Can we use criteria we have seen?
 - Graphs, MSE, R^2



MSE and R^2

- What will be effect of adding extra variables to model, and estimating their parameters, on MSE and R^2 ?
 - MSE will decrease, R^2 will increase
 - Because adding extra terms allows better fit
 - $\text{MSE}=0.16$ (was 0.21)
 - $R^2=0.56$ (was 0.45)



- Should we add extra variables in this case?
- Should we always add extra variables?
 - Is a more complex model always better than a simpler model?
 - Should we always put all our knowledge of the system into a model?
- The answer is no. Next we explain why.
 - Note that this implies that MSE and R^2 are not good for comparing models of different complexity.



Summary to here

- Common methods of evaluation
 - Graphs
 - MSE, R^2
- Decomposition of MSE can give indication of source of errors
- MSE and R^2 are not suited for comparing models of different complexity



EVALUATING PREDICTIONS



- Often, the goal is to predict for different situations
 - Could be future (prediction) or could be past (unobserved situations)
 - So we need to compare prediction errors of models. That is topic of this section.
 - (In other cases, we are interested in using a model to make decisions. In that case, we need to compare the quality of decisions based on different models. That is another lecture).
- In this case we are not really interested in MSE or R^2 per se.
 - We have data, don't need model for those situations
 - We would be interested in MSE if it gave information about predictions. Does it?



- First, define prediction quality



Prediction for what situations?

- Define target population = situations where we want to use model.
 - For model of animal metabolism rate, random selection of animals (of given race, age).
 - Corn yield in southwestern France. Random fields in region, random climate for region, certain management practices



Prediction of what variables?

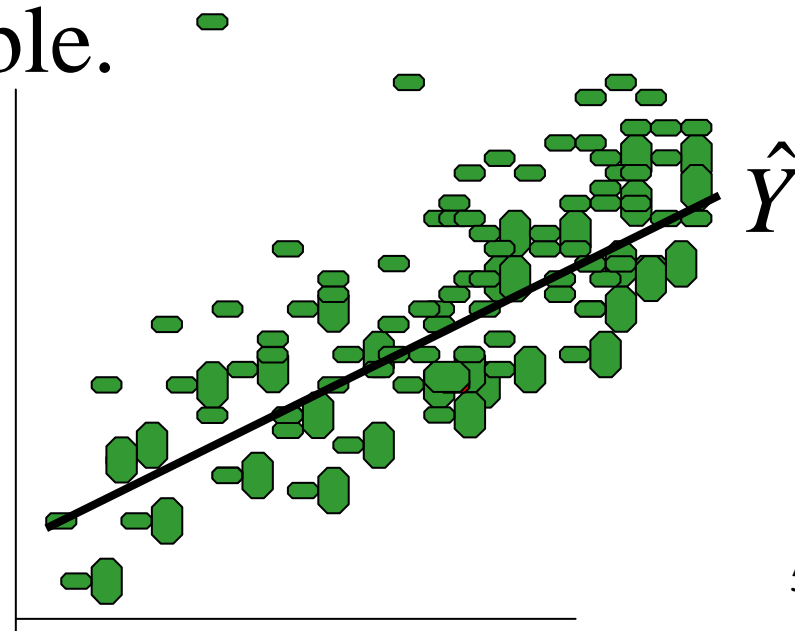
- Prediction quality depends on what we predict. Define target variables.
 - e. g. Aphid-ladybeetle model may have different error for prey population in margins, aphid population in wheat, ladybeetle population, total predation, etc.



A criterion of prediction error

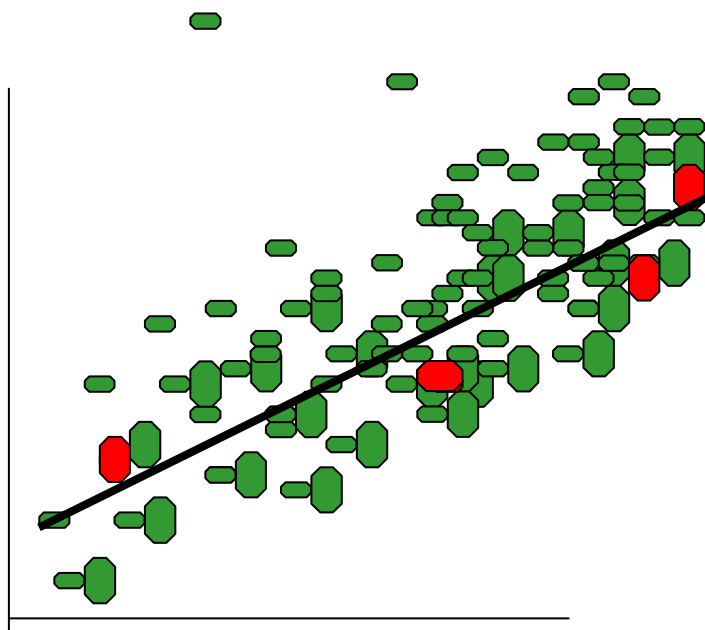
- A common measure of prediction error is $MSEP$ =mean squared error of prediction.
- Expectation over target population.
 Y is target variable.

$$MSEP = E \left[(Y - \hat{Y})^2 \right]$$



The difference MSE, MSEP

- MSE is adjustment error (based on measurements)
- MSEP is prediction error (for full target population)



- target population
- measurements

$$MSEP = E \left[\left(Y - \hat{Y} \right)^2 \right]$$

$$MSE = (1/N) \sum_{i=1}^N (Y_i - \hat{Y}_i)^2$$



- The difference between MSE and MSEP is very important.
 - Conceptually.
 - Practically. MSE and MSEP can be very different.



Estimate value of MSEP

- MSEP measures average squared error over target population. At best, we only have measurements for a sample.
- How can we measure MSEP?
 - We can't
- How can we estimate MSEP?
 - Based on measurements (no other choice)



- MSEP looks like MSE (a sum of squared errors).
- Is MSE a good estimator of MSEP?
 - We have a sample of measurements. On the average over possible samples, is $MSE = MSEP$?

$$MSEP = E \left[\left(Y - \hat{Y} \right)^2 \right]$$
$$MSE = (1/N) \sum_{i=1}^N (Y_i - \hat{Y}_i)^2$$



MSE estimates MSEP if...

- Our measurements are representative of the target population
- The measurements weren't used to develop the model
 - Often, measurements used to estimate parameter values
 - But could also be used to choose form of function etc.



Representative sample

- If data are not representative of target population, of course MSE is not a good measure of MSEP

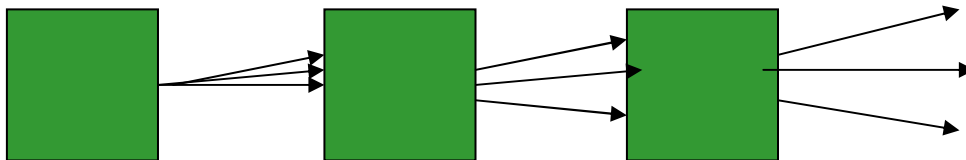


- Insure by random sampling
 - For complex systems, random sampling may not be possible.
 - e.g. agronomy experiments at field stations, not farmer fields.
 - With many explanatory variables, even random sample may not be representative
 - e.g. climate. With only a few years sampled, hard to say if this is representative sample.



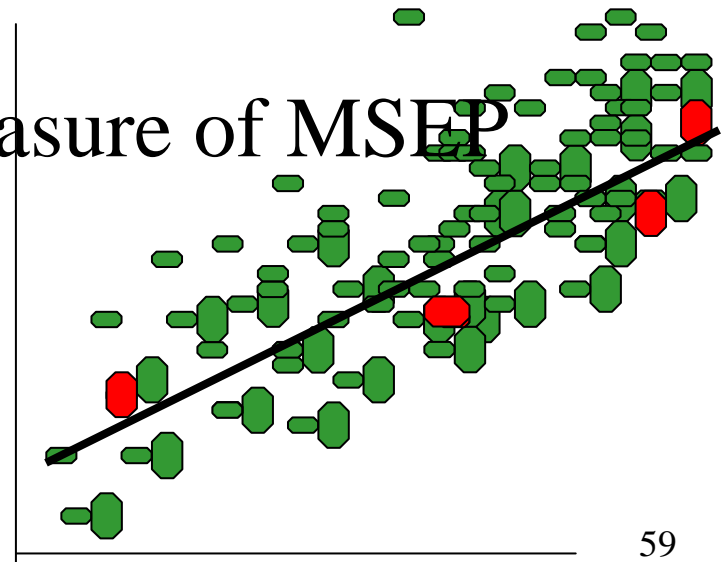
If sample from target population unavailable

- Can estimate error in each part of model, and use model to get overall error
 - In particular, parameter error
 - If we know possible distribution of parameter values, run model to get distribution of responses
 - See uncertainty analysis, Bayesian estimation



If measurements used to develop model?

- Typically, use measurements to estimate model parameters.
- Then model fits measurements better than new data
- So MSE isn't a good measure of MSEP



Example. $MSEP \neq MSE$

Adjusted parameters	$MSE(\hat{\theta})$	$MSEP(\hat{\theta})$
$\theta^{(0)}, \theta^{(1)}$	4.6077	4.30
$\theta^{(0)}, \theta^{(1)}, \theta^{(2)}$	0.0143	0.07
$\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \theta^{(3)}$	0.0119	0.06
$\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \theta^{(4)}$	0.0040	0.10
$\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \theta^{(4)}, \theta^{(5)}$	0.0003	0.42



Conclusions about MSE and MSEP

$MSEP \neq MSE$

For large p/n , $MSEP \gg MSE$

MSE always decreases as model complexity increases

MSEP has a minimum for some number of parameters



How can we estimate MSE if data is used in model development?

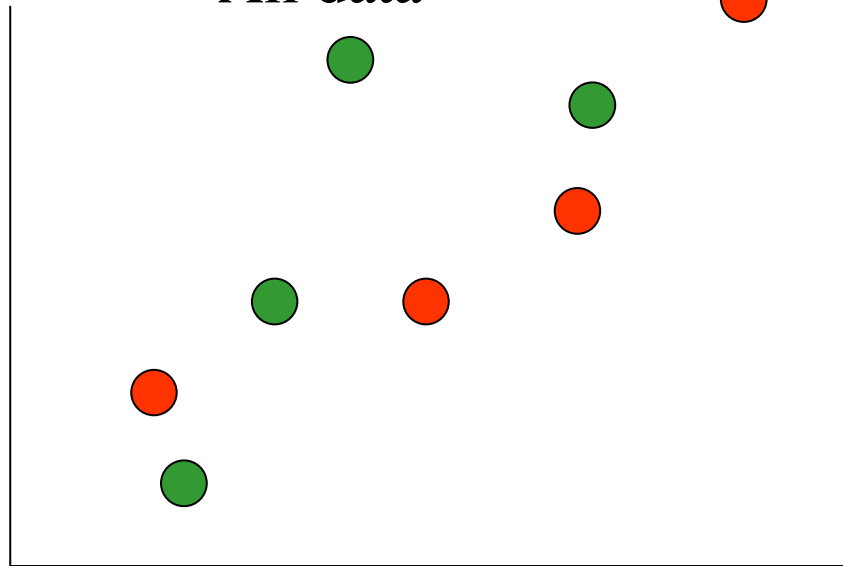
- This is important practical question
- We need estimate of error



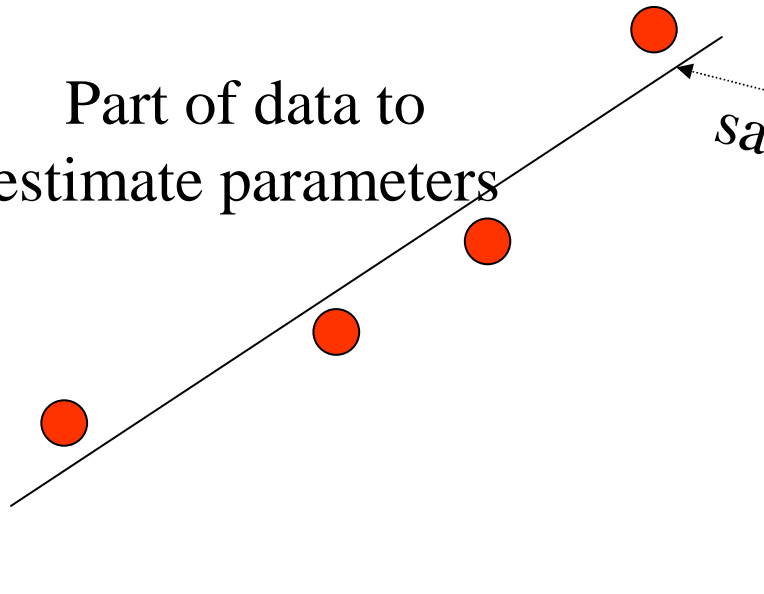
Data splitting



All data

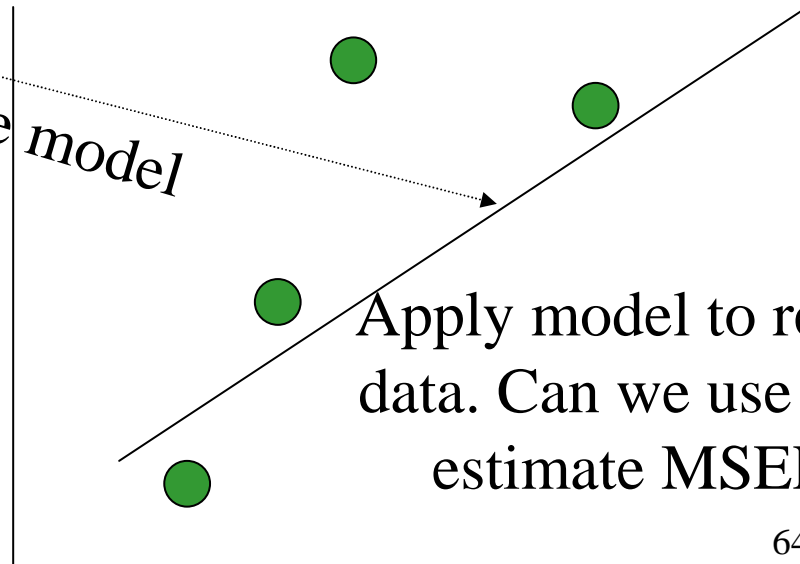


Part of data to
estimate parameters

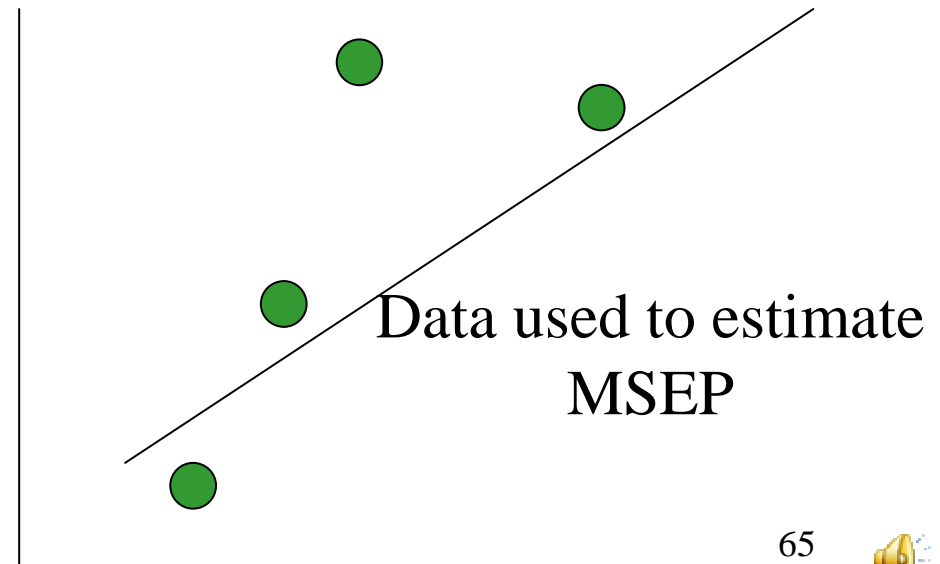
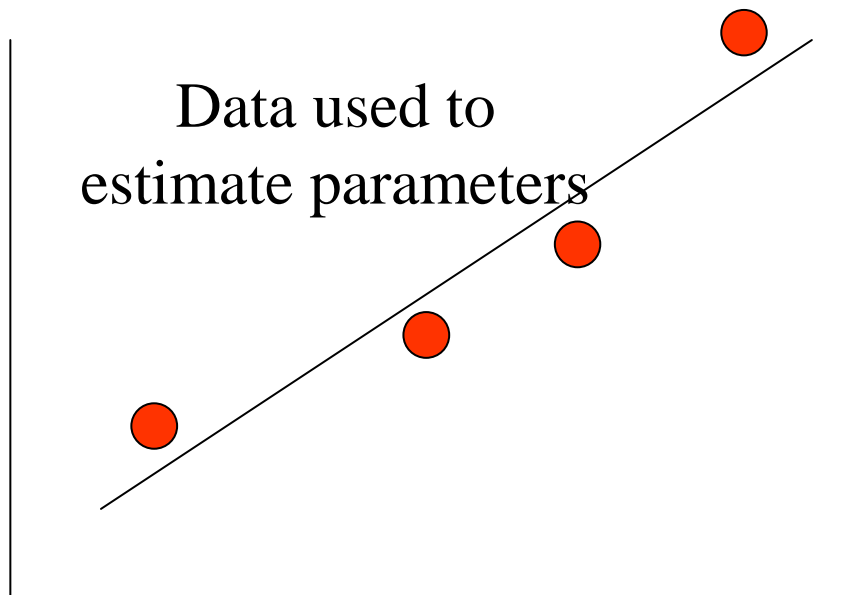


same model

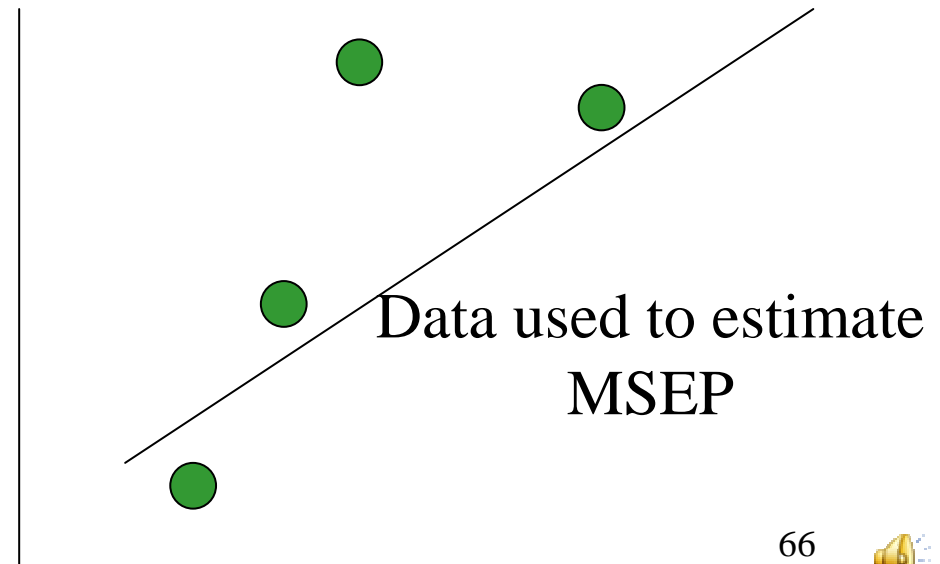
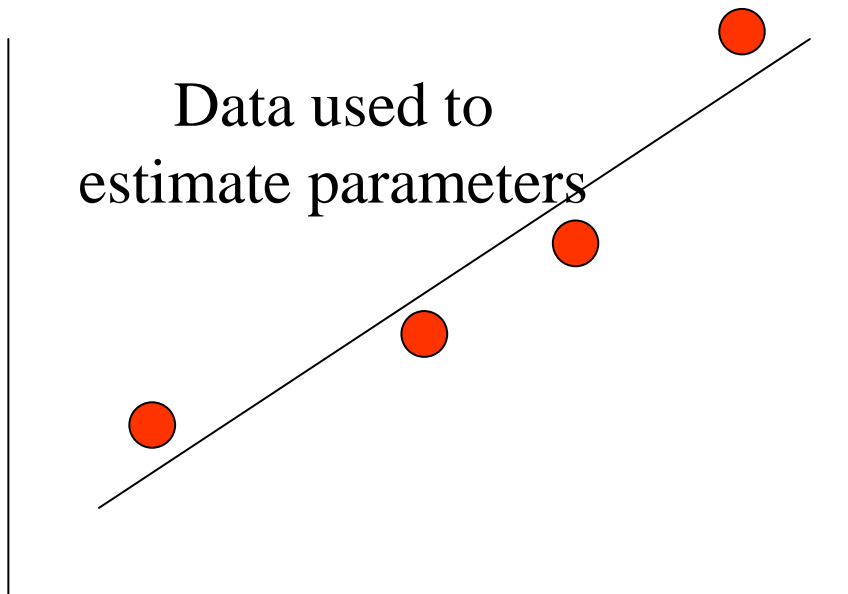
Apply model to rest of
data. Can we use fit to
estimate MSE?



- Yes, use MSE for second part of data to estimate MSEP (if data are from target distribution)
- Second part of data wasn't used to estimate parameters



- What are disadvantages of data splitting?
 - Arbitrary division of data into two parts
 - Use only part of data to estimate parameters
 - Use only part of data to estimate MSEP

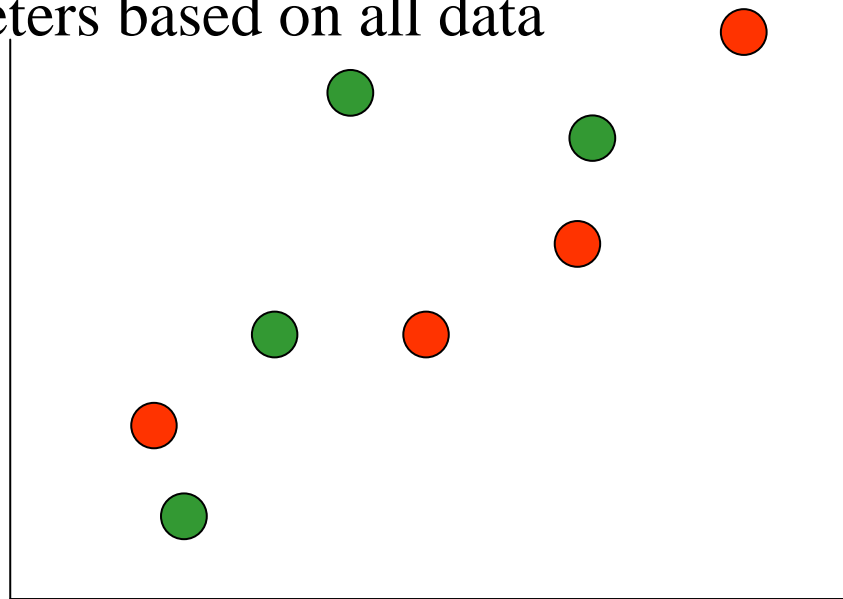


Other strategy

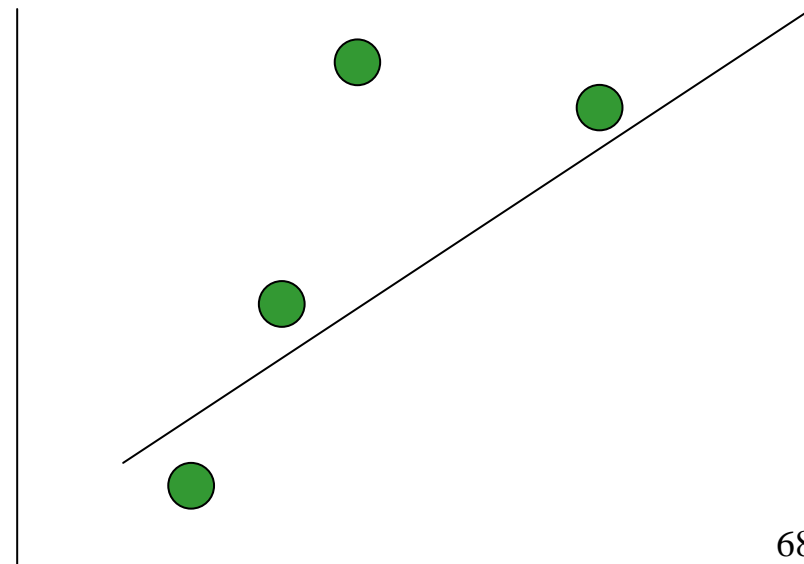
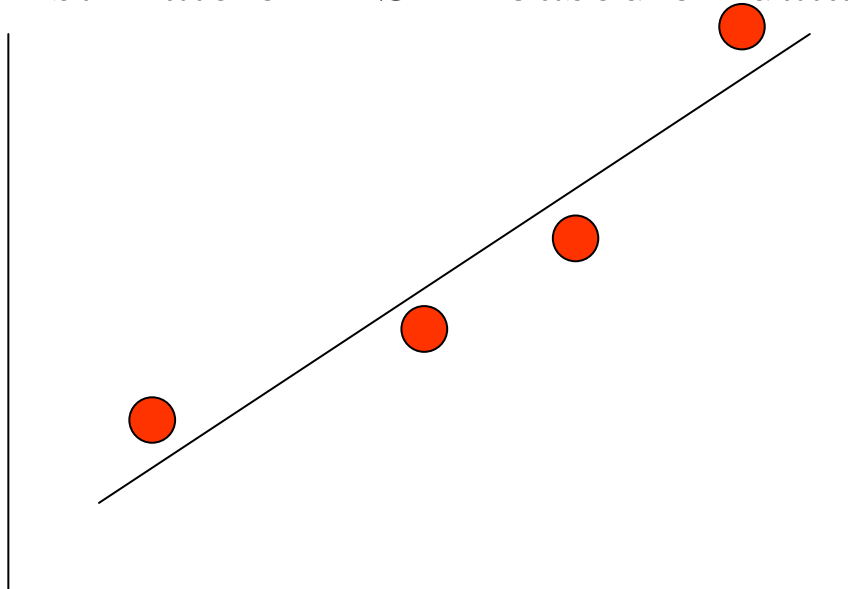
- Use all data to estimate parameters, then data splitting to estimate MSE_P



Proposed parameters based on all data



Estimate of MSE based on data splitting

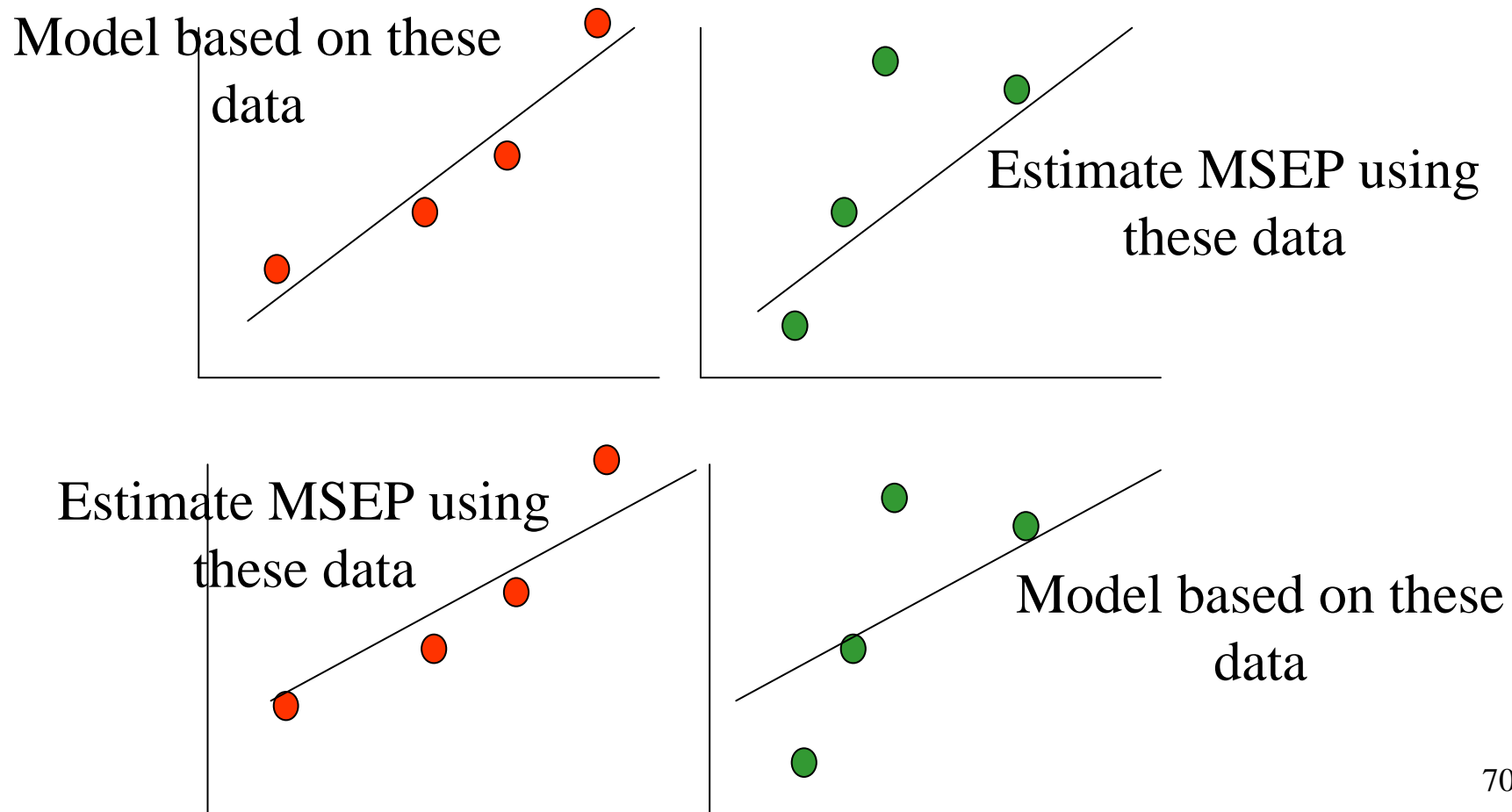


- What do you think of that?
- We want two things: parameter estimates for model and estimate of MSEP.
- This way, get best parameter estimates (use all data)
- And MSEP is correctly estimated.
 - The only problem is that MSEP refers to model based on half the data.
 - This probably overestimates MSEP for model based on all data.



Other strategy

- As above, but do data splitting twice. Then use average MSEP.



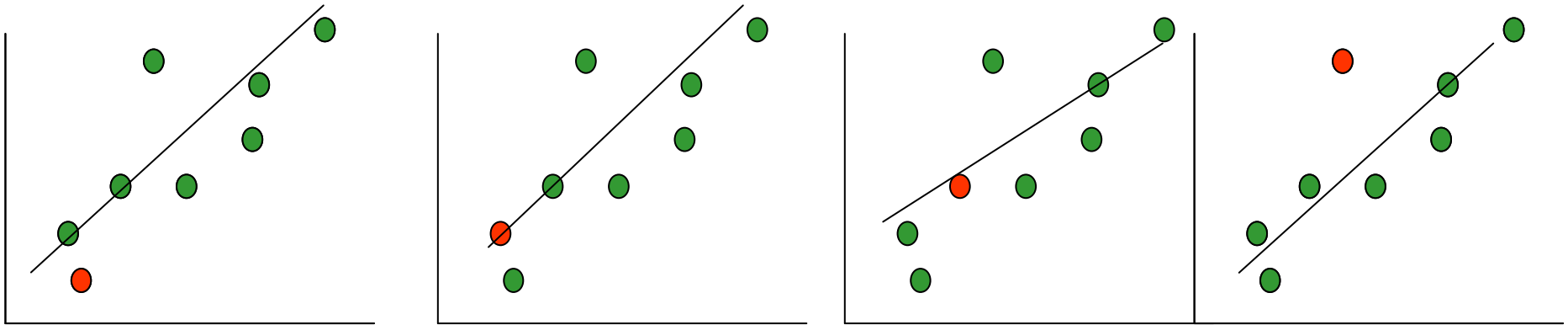
- What do you think of that?
- Less arbitrary
 - But split into two groups is still arbitrary
- Use all data to estimate MSEP
- But model for calculating MSEP isn't model that is proposed.
- Could we do better?



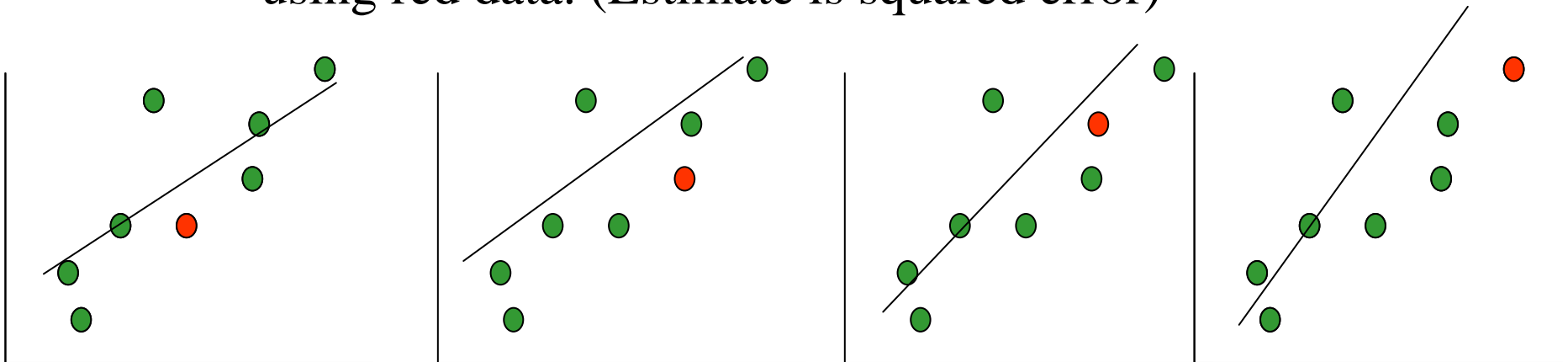
Cross validation

- Similar to above ideas.





Develop model using only green data. Estimate MSE_P using red data. (Estimate is squared error)



For N data values, repeat N times. Final estimate of MSE_P is average of N MSE_P estimates.



Calculation with cross validation

$$\boxed{Y_1} \quad Y_2 \quad Y_3 \quad \dots \quad Y_N \quad Y_{1-f_{-1}}(U_1)$$

$$Y_1 \quad \boxed{Y_2} \quad Y_3 \quad \dots \quad Y_N \quad Y_{2-f_{-2}}(U_2)$$

$$Y_1 \quad Y_2 \quad Y_3 \quad \dots \quad \boxed{Y_N} \quad Y_{N-f_{-N}}(U_N)$$

$$\hat{MSEP} = 1/n \sum [Y_i - f_{-i}(U_i)]^2$$



- What do you think of that?
- Proposed model based on all data.
- Evaluation based on model that uses all data but 1. So should be close to proposed model.



Decompose MSEP

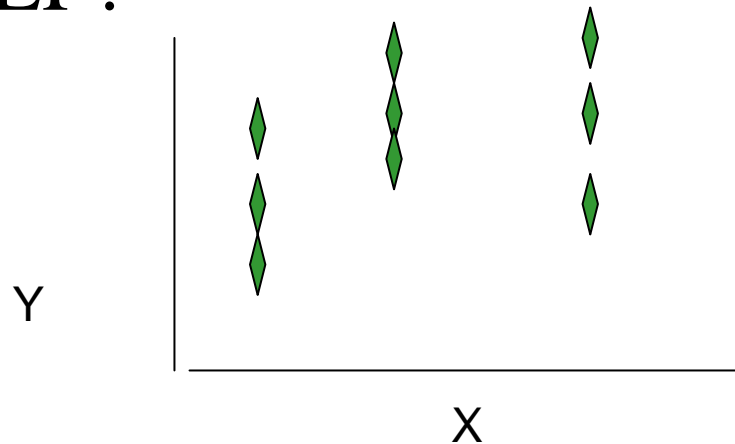


- MSEP can be written as the sum of two terms
- To help understand what determines predictive quality

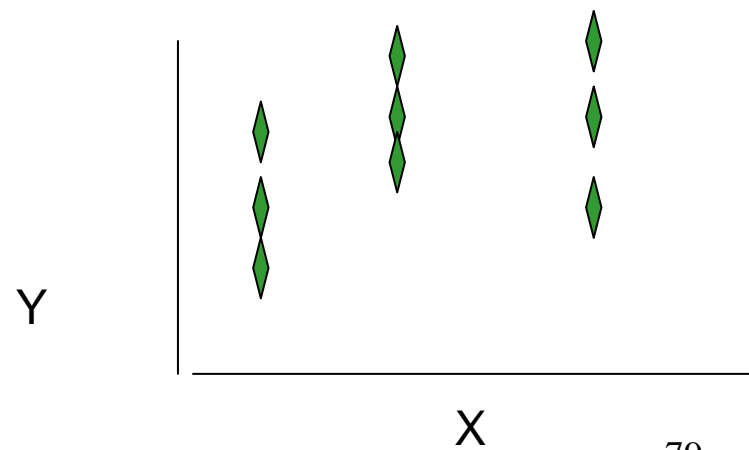


First term

- Model has some explanatory variables
- They do not explain all the variability in Y
 - e.g. Temp, geometry, initial values don't explain all aphid-ladybeetle dynamics
- What is relation between unexplained variability and MSEP?

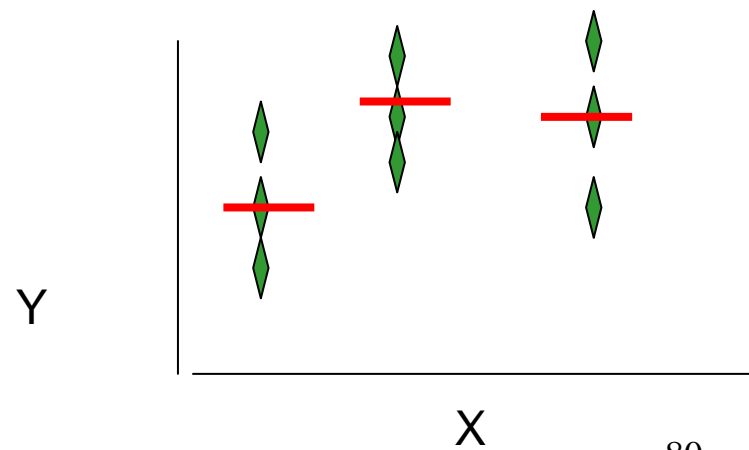


- For each value of explanatory variables X , model has unique prediction. Can't be exact for all
- What is best possible model?



- Best possible model (smallest MSE) equals average at each X.
- Remaining error is average variance for fixed X.

$$E_X \{ \text{var}(Y|X)^2 \}$$



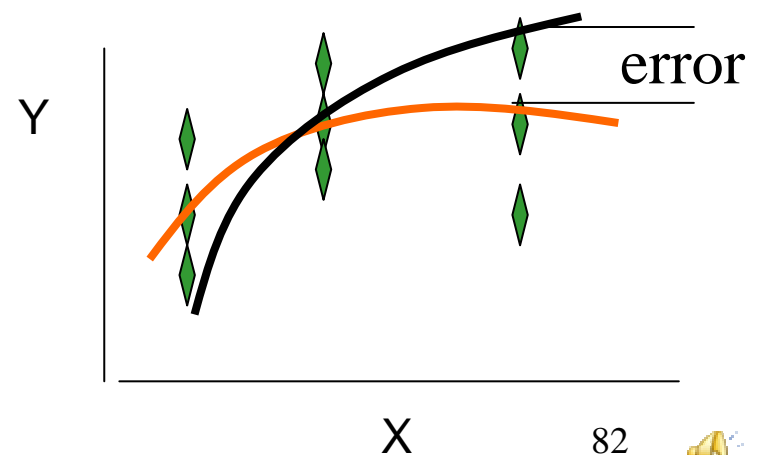
- Average variance for fixed X is lower limit for MSE. Just depends on choice of explanatory variables.
- What is effect of adding more explanatory variables (more detailed model)?
 - Adding explanatory variables always reduces average variance for fixed X .
 - But some explanatory variables are important, others less important or irrelevant.
- What is second term in MSE?



Second contribution to MSE

- Actual model will not be best model
 - Equations not exactly “correct”.
 - Parameters not exactly “correct”.
- Second term, model error for fixed X, measures difference between actual model and best model .

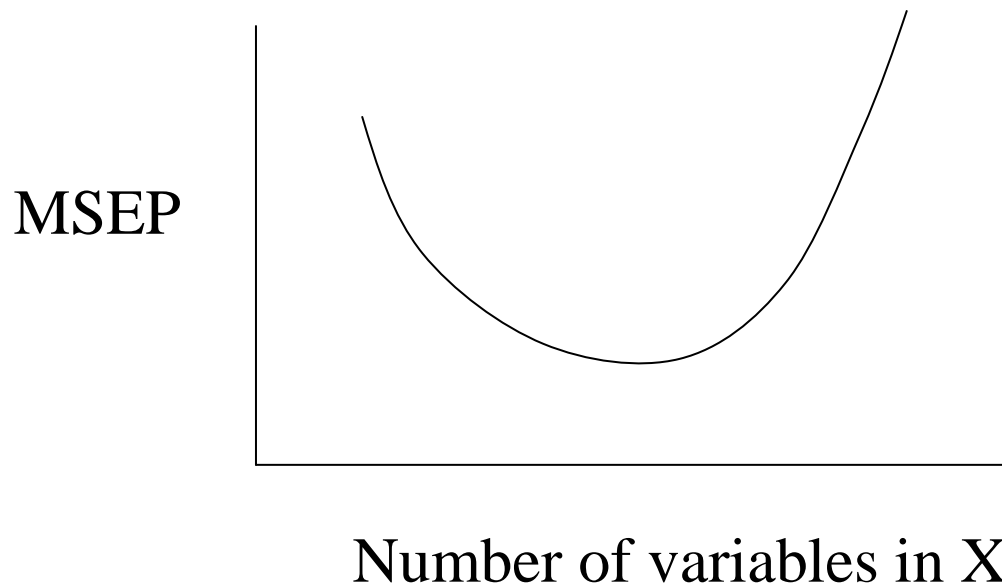
$$E_X \{ [E_Y(Y|X) - \hat{Y}(X)]^2 \}$$



- What is effect of extra detail (more variables in X or more equations) on second term?
 - This leads to more parameters. Each must be estimated. In general, more overall error.



- Overall effect of adding more variables in X ?
 - Reduces average variance for fixed X .
 - But in general increases model error for fixed X



- What is good strategy?
 - Add important variables, that reduces average variance for fixed X a lot.
 - Don't add unimportant variables.
 - Appropriate model complexity will depend on amount of data for estimating parameters.
 - This is particularly important for dynamic system models, where very complex models are possible



Example

Model	Variables in model Parameters in the model	<i>First term</i>	<i>2nd term</i>	<i>MSEP</i> ($\hat{\theta}$)
$f_1(X; \theta)$	$x^{(1)}$ $\theta^{(0)}, \theta^{(1)}$	4.04	0.36	4.40
$f_3(X; \theta)$	$x^{(1)} x^{(2)} x^{(3)}$ $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \theta^{(3)}$	0.04	0.01	0.05
$f_5(X; \theta)$	$x^{(1)} x^{(2)} x^{(3)} x^{(4)} x^{(5)}$ $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \theta^{(4)}, \theta^{(5)}$	0.04	0.35	0.39



Summary

- Common criterion of prediction error is MSEP
 - Specify target population, target variables
- MSE is not in general a good estimator of MSEP
 - In particular if measured sample is not representative of target population, or if sample is used for parameter estimation
 - The difference between MSE and MSEP depends on p/n
- MSEP is the result of two contributions
 - Variation due to fact that explanatory variables don't explain all variability
 - Differences between model and best model
- MSEP has a minimum for some intermediate level of complexity



Evaluating decisions based on a model

- Not exactly the same as a good model for prediction.
- See David Makowski lecture.



THE END



References for examples

- Gent, M. P. N., 1994, Photosynthate Reserves during Grain Filling in Winter Wheat, *Agron J* 86:159-167
- Michalska, B. and Witos, A. 2000. Weather-based spring wheat yielding forecasting. *EJPAU* online.
<http://www.ejpau.media.pl/volume3/issue2/agronomy/art-04.html>