Data Splitting and Resampling

how to use the data set to build a model

Introduction

- How predictive is the model we developed?
- Error on the training data is *not* a good indicator of performance on future data
  - ***Q: Why?***
  - A: Because new data will probably not be **exactly** the same as the training data!
- Overfitting – fitting the training data too precisely - usually leads to poor results on new data
- Also: we cannot be certain about unseen data. Stationarity is Dead!

How to check if a model fit is good?

- The R^2 statistic has become the almost universally standard measure for model fit in linear models.
- What is R^2?

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

← Model error

← Variance in the dependent variable

- It is the ratio of error in a model over the total variance in the dependent variable.
- Hence the lower the error, the higher the R2 value.

How to check if a model fit is good?

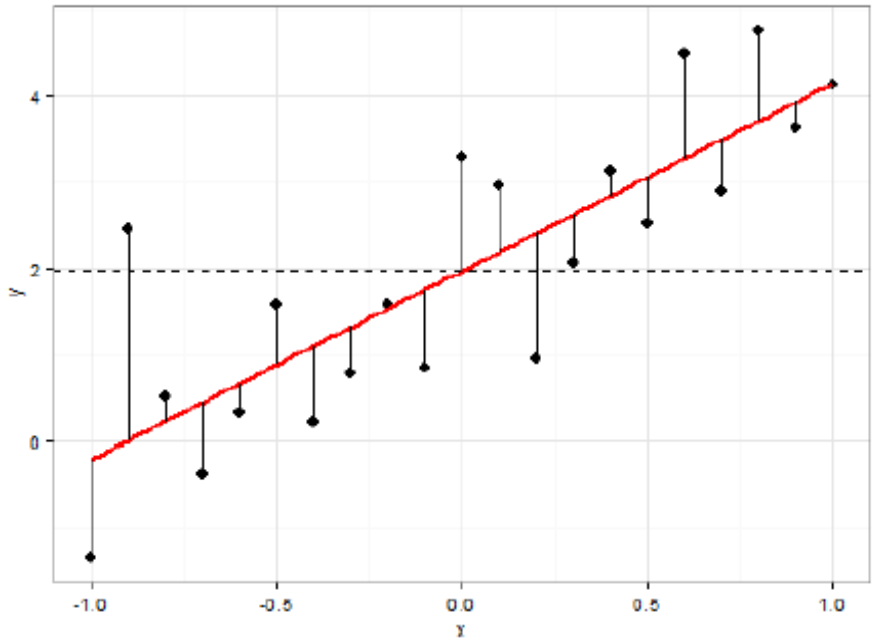
$$\sum(y_i - f_i)^2 = 18.568$$

$$\sum(y_i - \bar{y})^2 = 55.001$$

$$R^2 = 1 - \frac{18.568}{55.001}$$

$$R^2 = 0.6624$$

A decent model fit!



How to check if a model fit is good?

$$\sum(y_i - f_i)^2 = 15.276$$

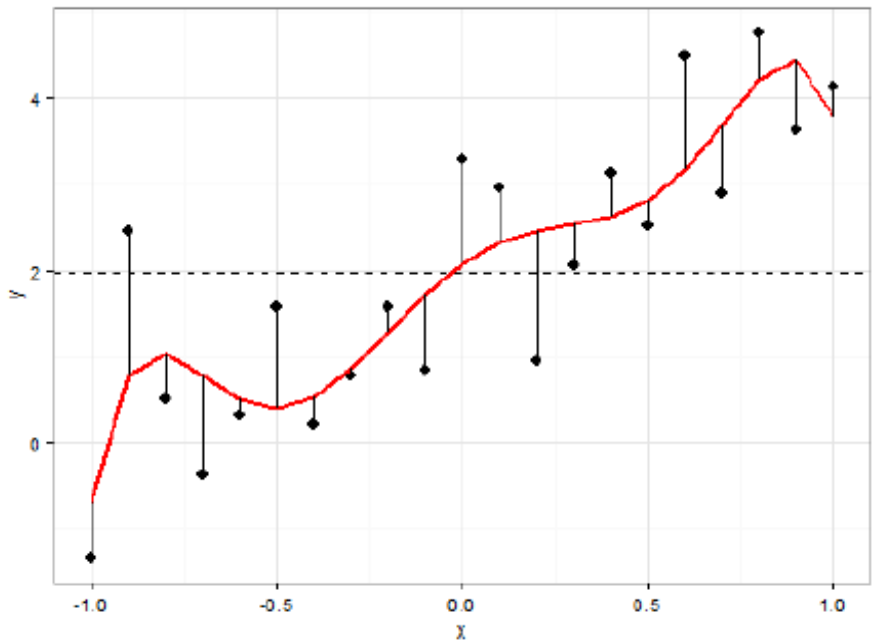
$$\sum(y_i - \bar{y})^2 = 55.001$$

$$R^2 = 1 - \frac{15.276}{55.001}$$

$$R^2 = 0.72$$

Is this a better model?

No, overfitting!



OVERFITTING

- Modeling techniques tend to overfit the data.
- Multiple regression:

✓*Every* time you add a variable to the regression, the model's R^2 goes up.

✓Naïve interpretation: *every* additional predictive variable helps to explain yet more of the target's variance. But that can't be true!

✓Left to its own devices, Multiple Regression will fit *too many* patterns.

✓A reason why modeling requires subject-matter expertise.

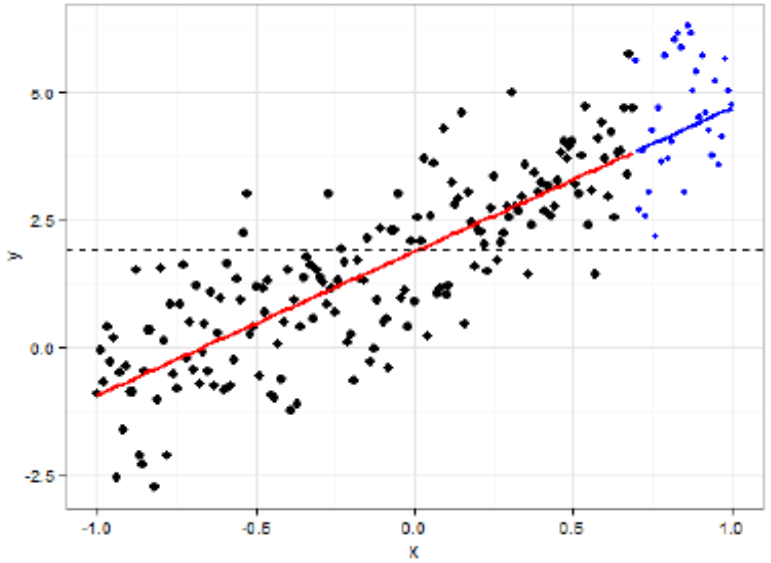
OVERFITTING

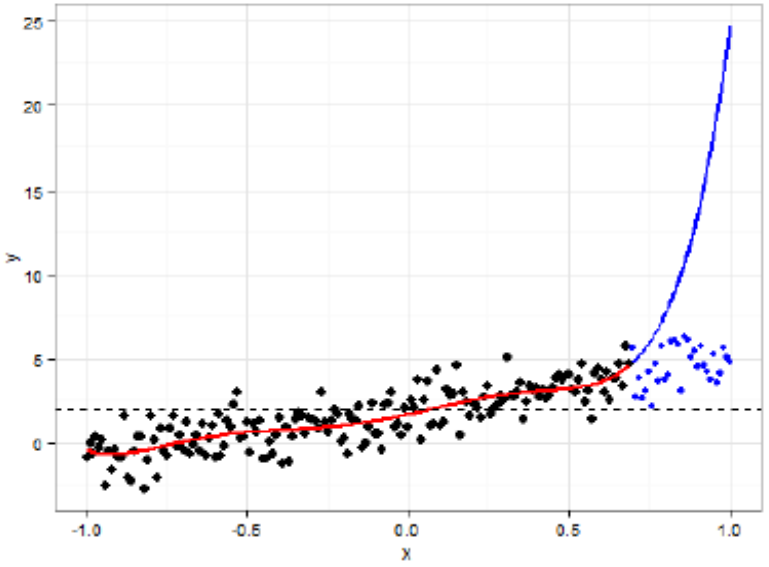
- Error on the dataset used to *fit* the model can be misleading
  - › Doesn't predict future performance.
- Too much complexity can diminish model's accuracy on future data.
  - › Sometimes called the Bias-Variance Tradeoff.

OVERFITTING

- What are the consequences of overfitting?

  "*Overfitted models will have high R^2 values, but will perform poorly in predicting out-of-sample cases*"





WHY WE NEED CROSS-VALIDATION?

- R^2 does not offer any significant insights into how well our regression model can predict future values.
- One thing that R-squared offers no protection against is overfitting. On the other hand, cross validation, by allowing us to have cases in our testing set that are different from the cases in our training set, inherently offers protection against overfitting.
- When a model is to be used for prediction purposes it is useful to obtain empirical evidence as to its generalizability, or its capacity to make accurate predictions for new samples of data. This process is sometimes referred to as “validating” the regression equation.

WHY WE NEED CROSS-VALIDATION?

- One way to address this issue is to literally obtain a new sample of observations. That is, after the model is developed from the original sample, the investigator conducts a new study, replicating the original one as closely as possible, and uses the new data to assess the predictive validity of the model.
- This procedure is usually impractical
- An alternative, more practical procedure is *cross-validation*.

CROSS-VALIDATION

- In cross-validation the original sample is split into two parts. One part is called the training sample, and the other part is called the *validation* sample.

1) **What portion of the sample should be in each part?**

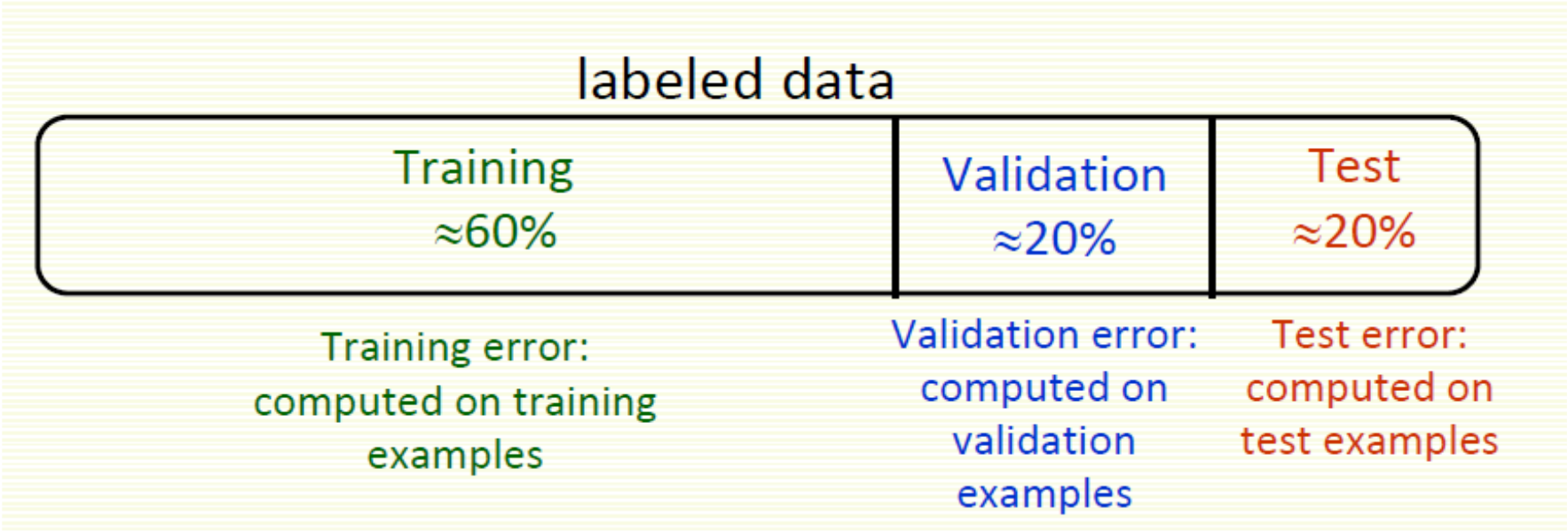
   If sample size is very large, it is often best to split the sample in half. For smaller samples, it is more conventional to split the sample such that 2/3 of the observations are in the training sample and 1/3 are in the validation sample.

CROSS-VALIDATION

2) How should the sample be split?

The most common approach is to divide the sample randomly, thus theoretically eliminating any systematic differences. However, one should also use some physical insight in making the splitting choices.

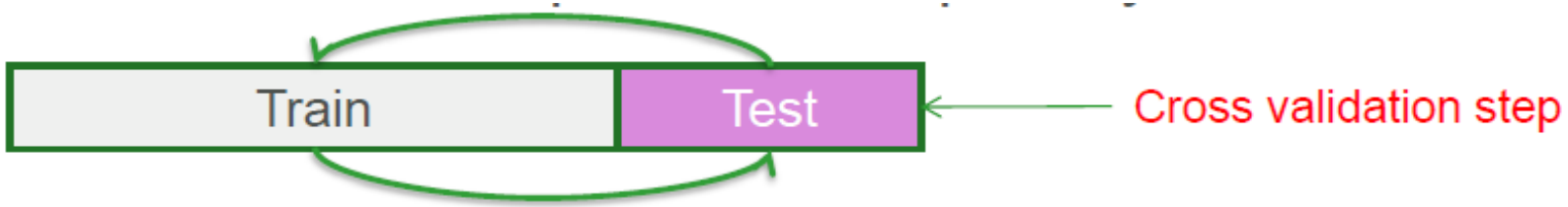
TRAINING/ VALIDATION – IDEAL PROCEDURE



CROSS VALIDATION – THE IDEAL PROCEDURE

1.Divide data into three sets, training, validation and test sets

2.Find the optimal model using the training set, and use the test set to check its predictive capability



3.See how well the model can predict the test set

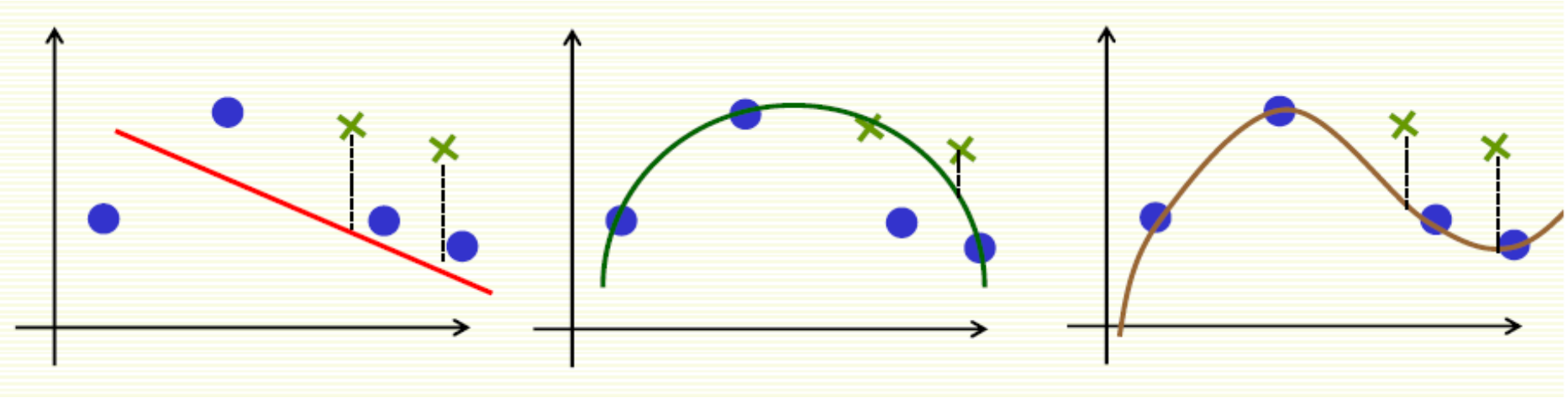
4.The validation error gives an unbiased estimate of the predictive power of a model

TRAINING/TEST DATA SPLIT

Talked about splitting data in training/test sets

- training data is used to fit parameters
- test data is used to assess how the model generalizes to new data

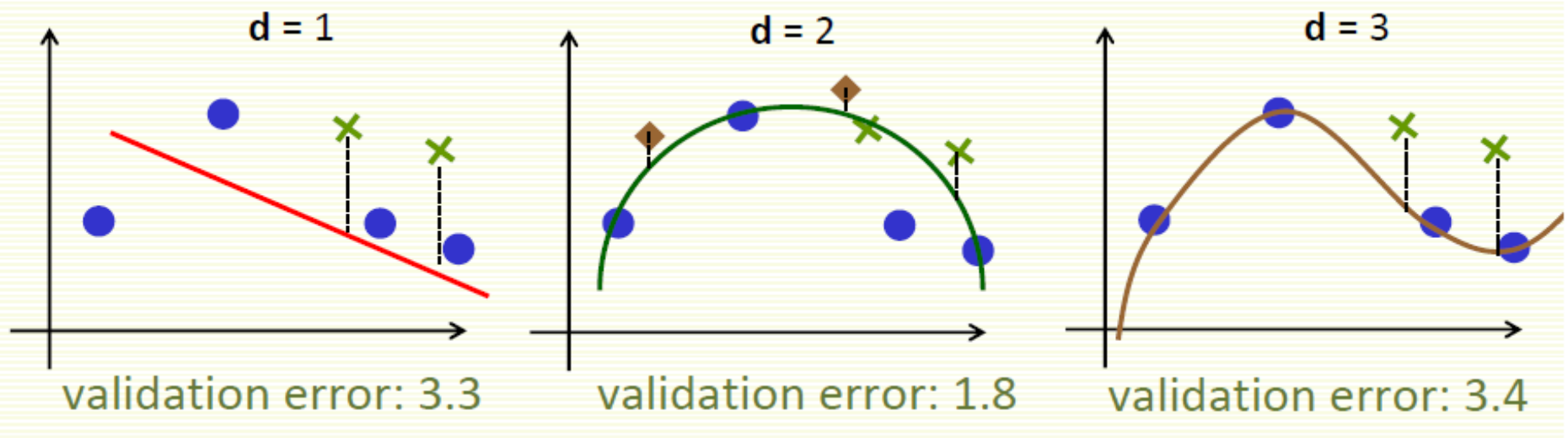
TRAINING/TEST DATA SPLIT



What about test error?

- Test error should be computed for data that was **not used for training at all**
- Not even in making choice of the model

Training/Validation/Test Data



- Training Data
- Validation Data

  d = 2 is chosen
- Test Data

  1.3 test error computed for **d** = 2

Improved cross-validation

- Even better approach: *repeated* cross-validation

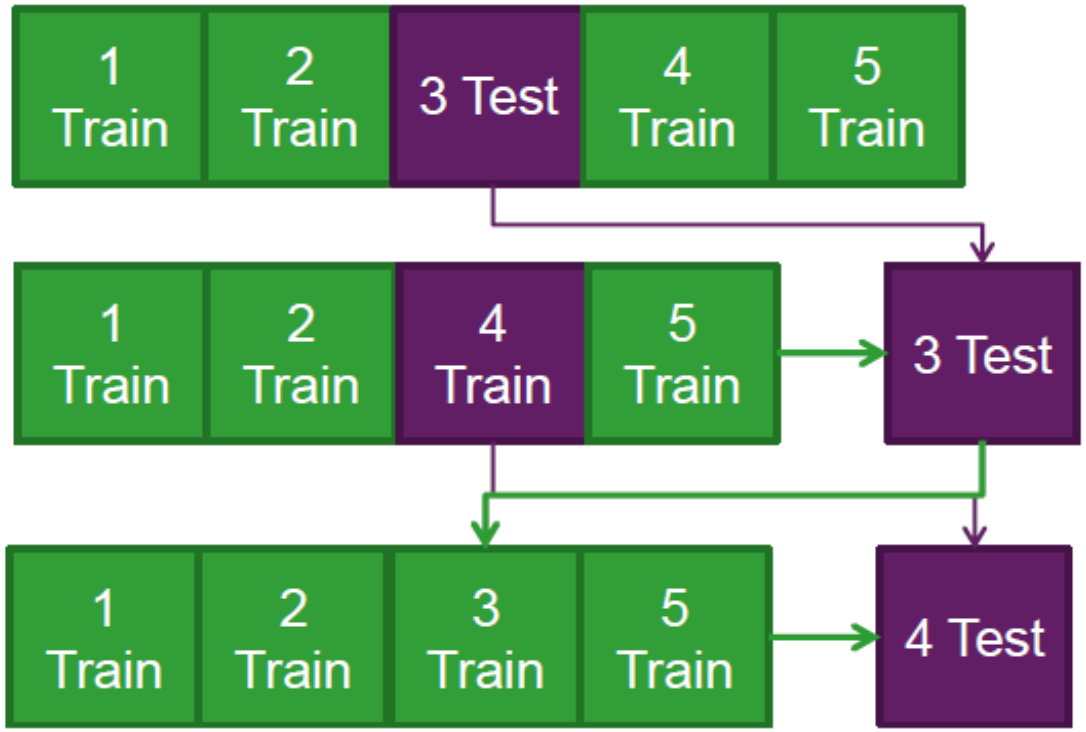
Example:

10-fold cross-validation is repeated 10 times and results are averaged (reduce the variance)

K-FOLD CROSS VALIDATION

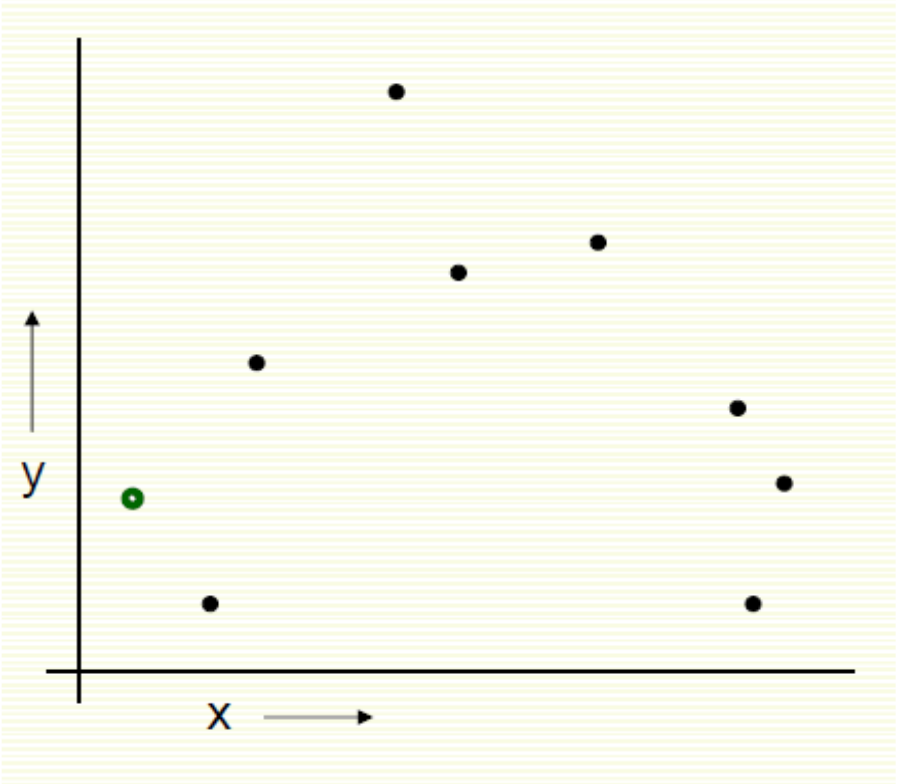
- The approach can be generalised to use multiple training-validation data splits
- k-fold CV works as follows:
    1. Split the sample into k subsets of equal size
    2. For each fold estimate a model on all the subsets except one
    3. Use the left out subset to test the model, by calculating a CV metric of choice
    4. Average the CV metric across subsets to get the CV error
- This has the advantage of using all data for estimating the model, however finding a good value for *k* can be tricky

K-fold Cross Validation Example



1. Split the data into 5 samples
2. Fit a model to the training samples and use the test sample to calculate a CV metric.
3. Repeat the process for the next sample, until all samples have been used to either train or test the model

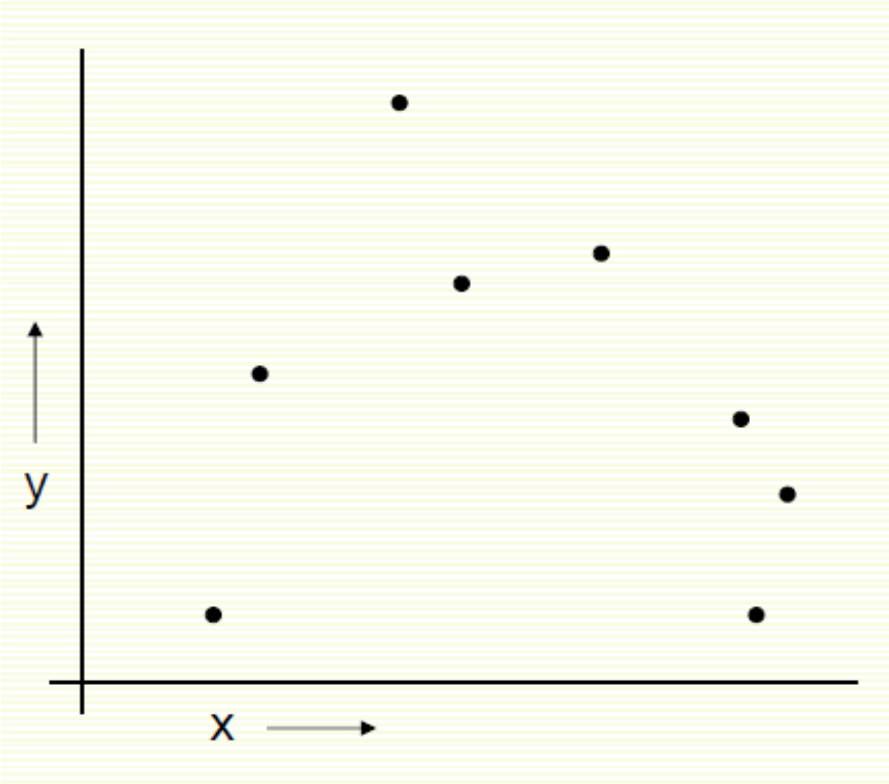
Leave-one-out Cross Validation is extreme case of k-fold Cross-Validation



- For k=1 to R

1. Let ($\mathbf{x}^k$,$\mathbf{y}^k$) be the **k** example

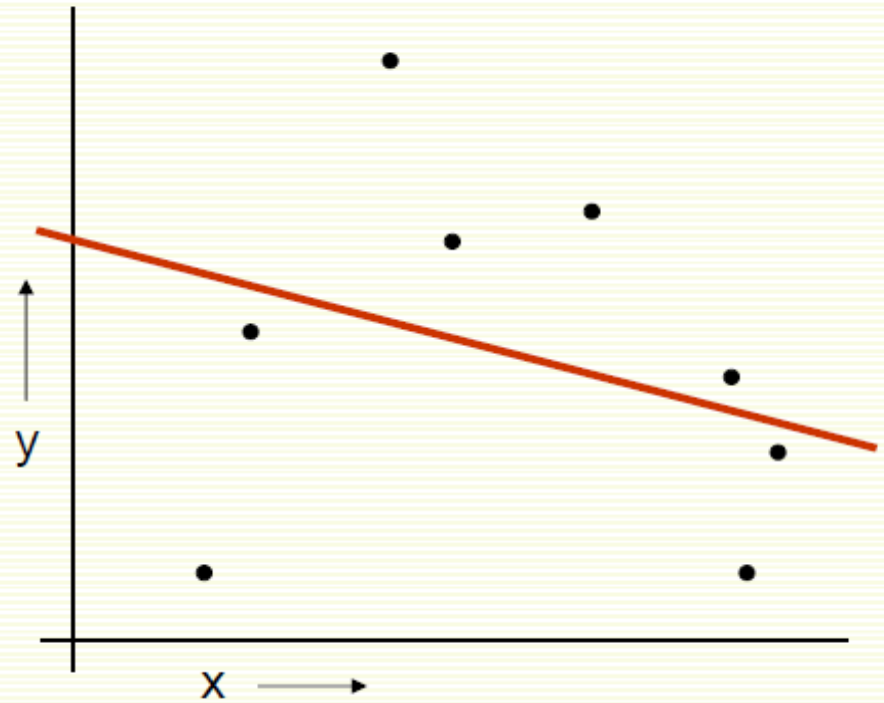
LOOCV (Leave-one-out Cross Validation)



For **k**=1 to **n**

1. Let ($\mathbf{x}^k$,$\mathbf{y}^k$) be the **k**th example
2. Temporarily remove ($\mathbf{x}^k$,$\mathbf{y}^k$) from the dataset

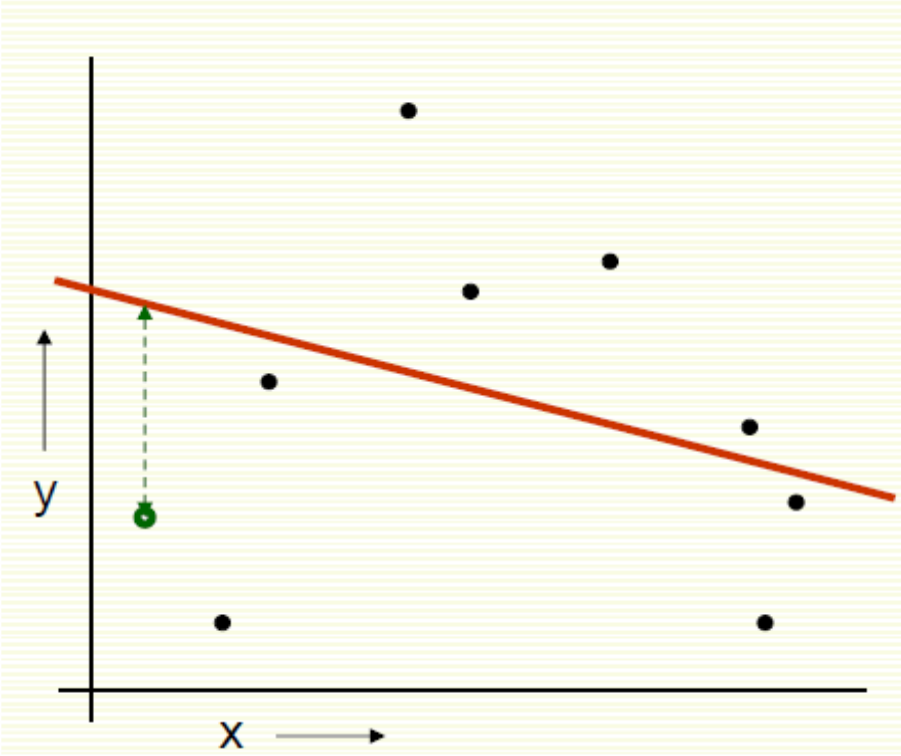
LOOCV (Leave-one-out Cross Validation)



For **k**=1 to **n**

1. Let ($\mathbf{x}^k$,$\mathbf{y}^k$) be the **k**th example
2. Temporarily remove ($\mathbf{x}^k$,$\mathbf{y}^k$) from the dataset
3. Train on the remaining **n**-1 examples

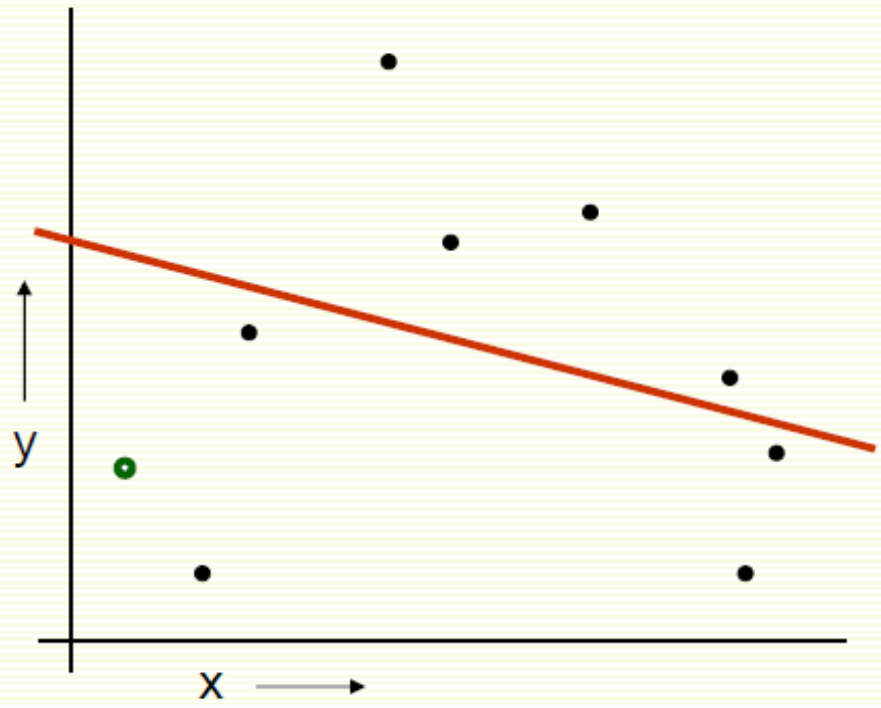
LOOCV (Leave-one-out Cross Validation)



For **k**=1 to **n**

1. Let ($\mathbf{x}^k$,$\mathbf{y}^k$) be the **k**th example
2. Temporarily remove ($\mathbf{x}^k$,$\mathbf{y}^k$) from the dataset
3. Train on the remaining **n**-1 examples
4. Note your error on ($\mathbf{x}^k$,$\mathbf{y}^k$)

LOOCV (Leave-one-out Cross Validation)

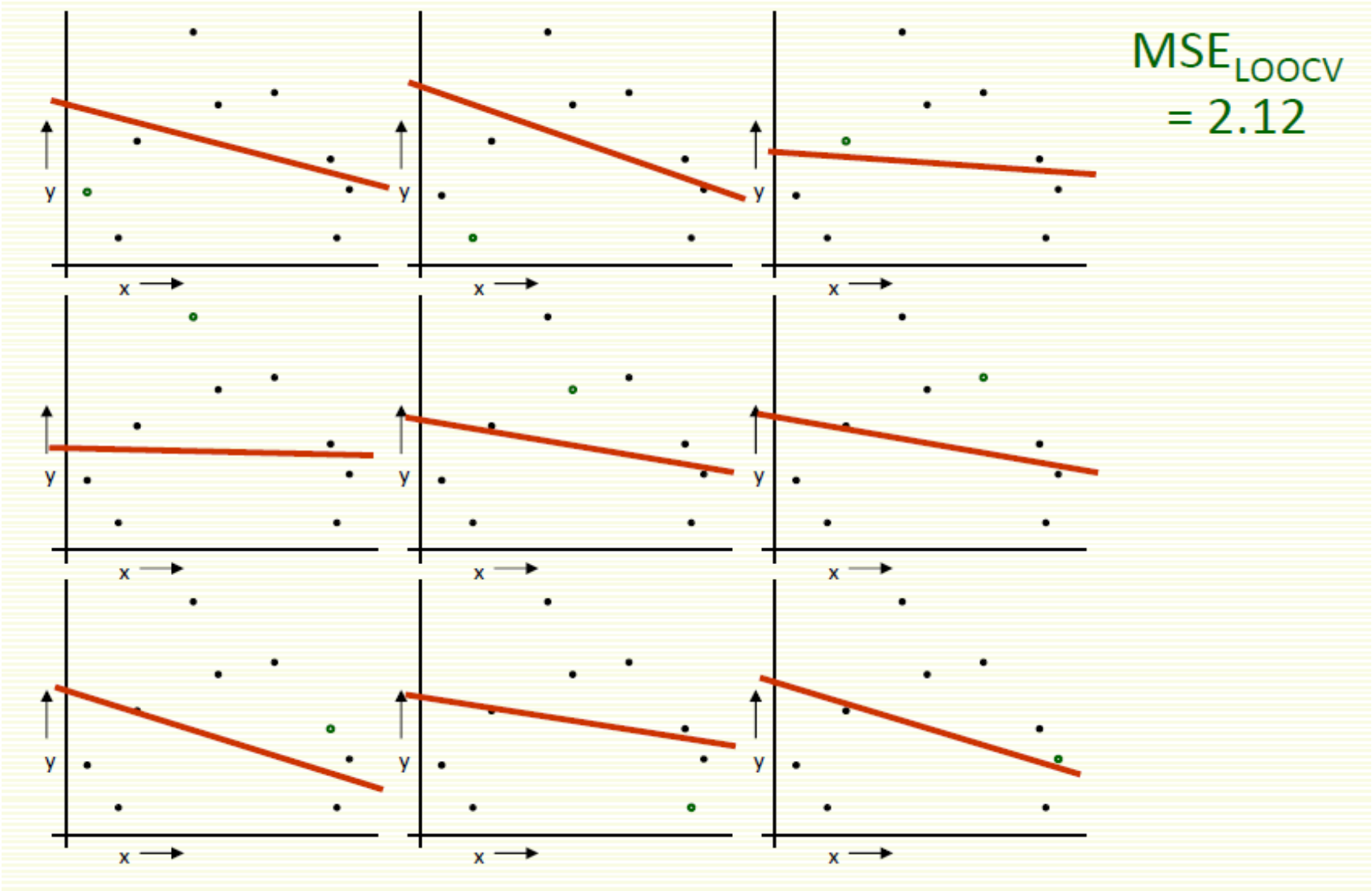


For **k**=1 to **n**

1. Let ($\mathbf{x}^k$,$\mathbf{y}^k$) be the **k**th example
2. Temporarily remove ($\mathbf{x}^k$,$\mathbf{y}^k$) from the dataset
3. Train on the remaining **n**-1 examples
4. Note your error on ($\mathbf{x}^k$,$\mathbf{y}^k$)

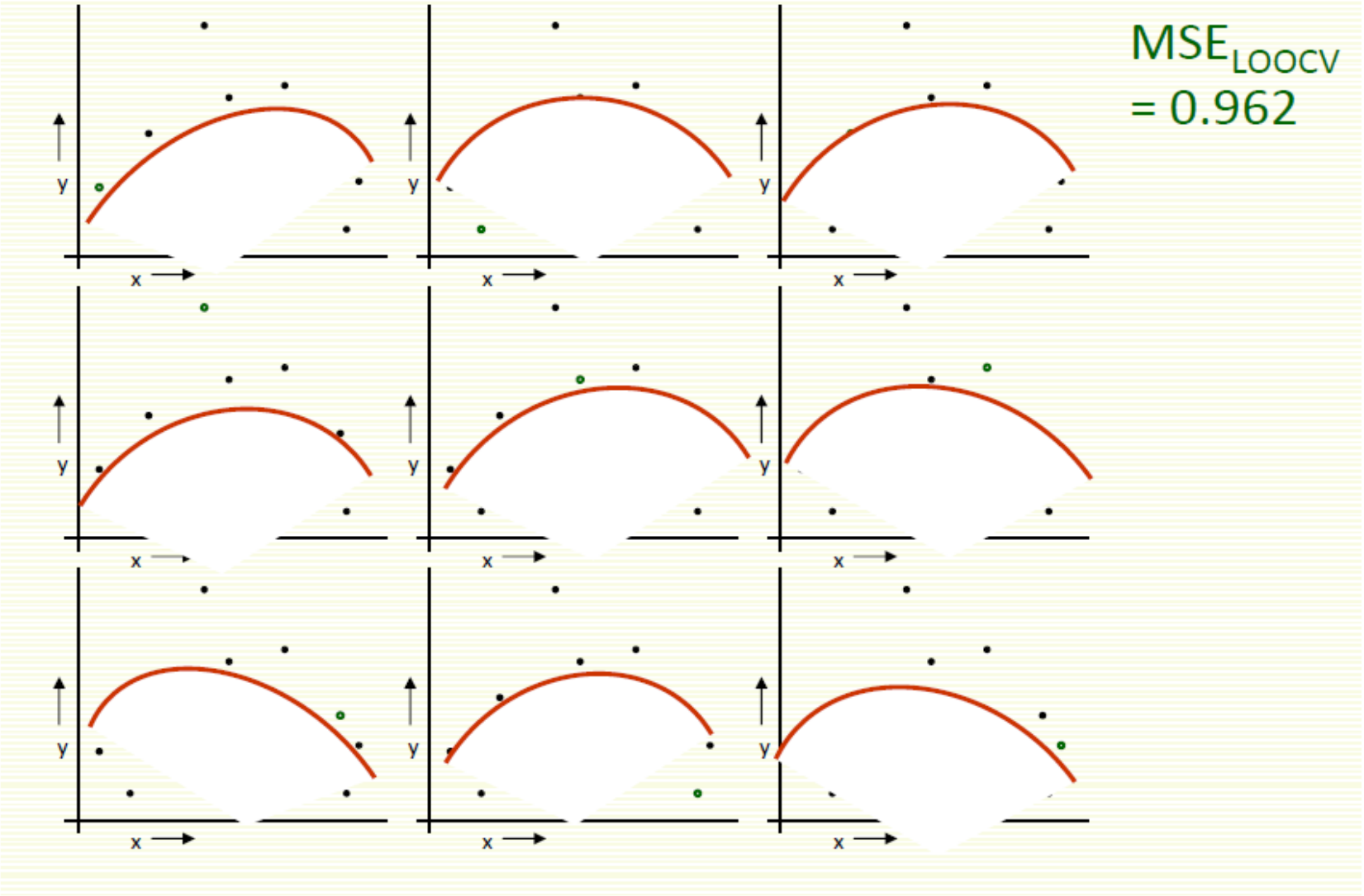
When you've done all points, report the mean error

LOOCV (Leave-one-out Cross Validation)



$MSE_{LOOCV} = 2.12$

LOOCV for Quadratic Regression

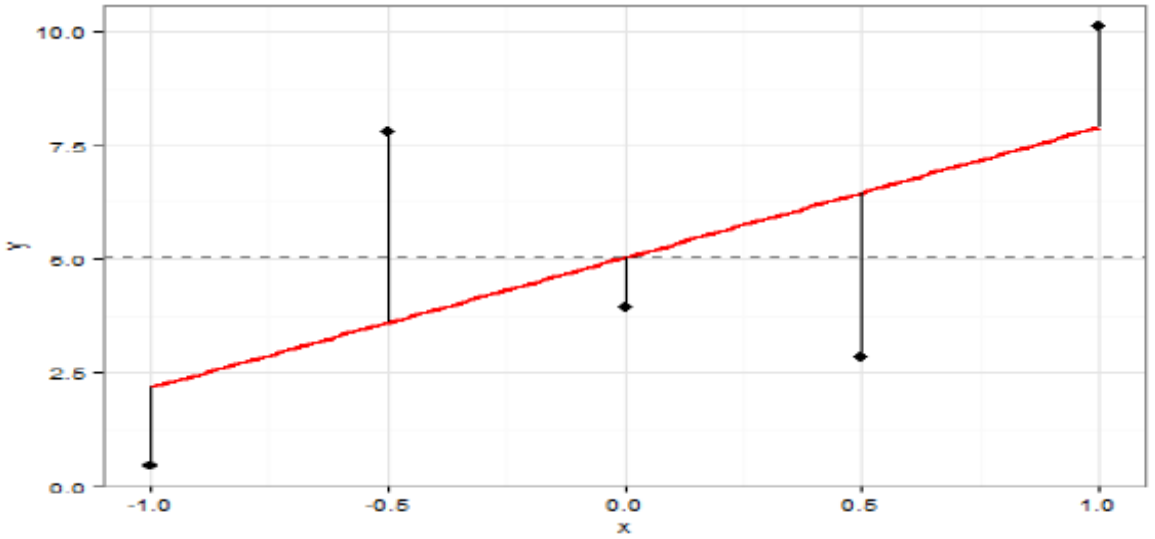


Cross Validation - Metrics

- How do we determine if one model is predicting better than another model?

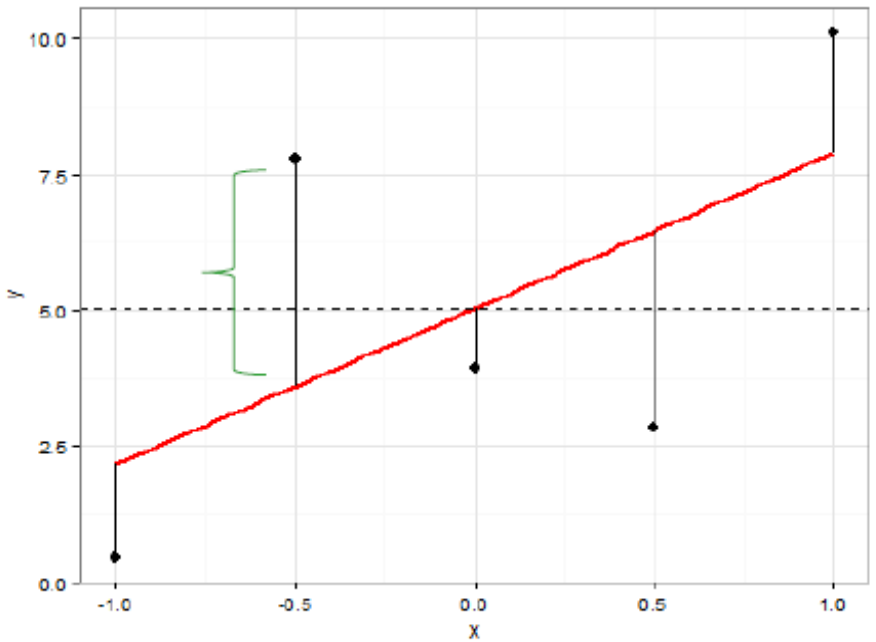
The basic relation:

› $Error_i = y_i - f_i$ ← The difference between observed (y) and predicted value (f), when applying the model to unseen data



Cross Validation Metrics

- Mean Squared Error (MSE)
  - $1/n\sum(y_i - f_i)^2$
  - 7.96
- Root Mean Squared Error (RMSE)
  - $\sqrt{1/n\sum(y_i - f_i)^2}$
  - 2.82
- Mean Absolute Percentage Error (MAPE)
  - $(1/n\sum|\frac{y_i - f_i}{y_i}|)*100$
  - 120%



Best Practice for Reporting Model Fit

1. Use Cross Validation to find the best model
2. Report the RMSE and MAPE statistics from the cross validation procedure
3. Report the R Squared from the model as you normally would.

The added cross-validation information will allow one to evaluate not how much variance can be explained by the model, but also the predictive accuracy of the model. Good models should have a high predictive AND explanatory power!

Which kind of Cross Validation?

	Downside	Upside
Test-set	may give unreliable estimate of future performance	cheap
Leave-one-out	expensive	doesn't waste data

Which kind of Cross Validation?

	Downside	Upside
Test-set	may give unreliable estimate of future performance	cheap
Leave-one-out	expensive	doesn't waste data
10-fold	wastes 10% of the data,10 times more expensive than test set	only wastes 10%, only 10 times more expensive instead of **n** times
3-fold	wastes more data than 10-fold, more expensive than test set	slightly better than test-set
N-fold	Identical to Leave-one-out	