Data, Inference & Applied Machine Learning

Course: 18-785

Patrick McSharry
patrick@mcsharry.net
www.mcsharry.net
Twitter: @patrickmcsharry

Fall 2022

ICT Center of Excellence
Carnegie Mellon University

Course outline

Week	Description
1	Measurement, data types, data collection, data cleaning
2	Data manipulation, data exploration, visualization techniques
3	Probability, statistical distributions, descriptive statistics
4	Statistical hypothesis testing, quantifying confidence
5	Time series analysis, autoregression, moving averages
6	Linear regression, parameter estimation, model selection, evaluation

Data & Inference

WEEK 6A

Today's Lecture

No.	Activity	Description	Time
1	Challenge	Understanding the past and forecasting the future	10
2	Discussion	Price discovery - Lego	10
3	Case study	Sales and marketing	10
4	Analysis	Linear regression	20
5	Demo	Techniques for linear regression	20
6	Q&A	Questions and feedback	10

Understanding the past

- The first step in data analytics should be to obtain a deeper understanding of the past.
- This involves investigating the relationship between some explanatory variables, **X**, and the dependent variable of interest, **y**.
- Depending on the problem at hand, there may only exist a limited number of candidate variables or we may need to select from a large collection.

Forecasting the future

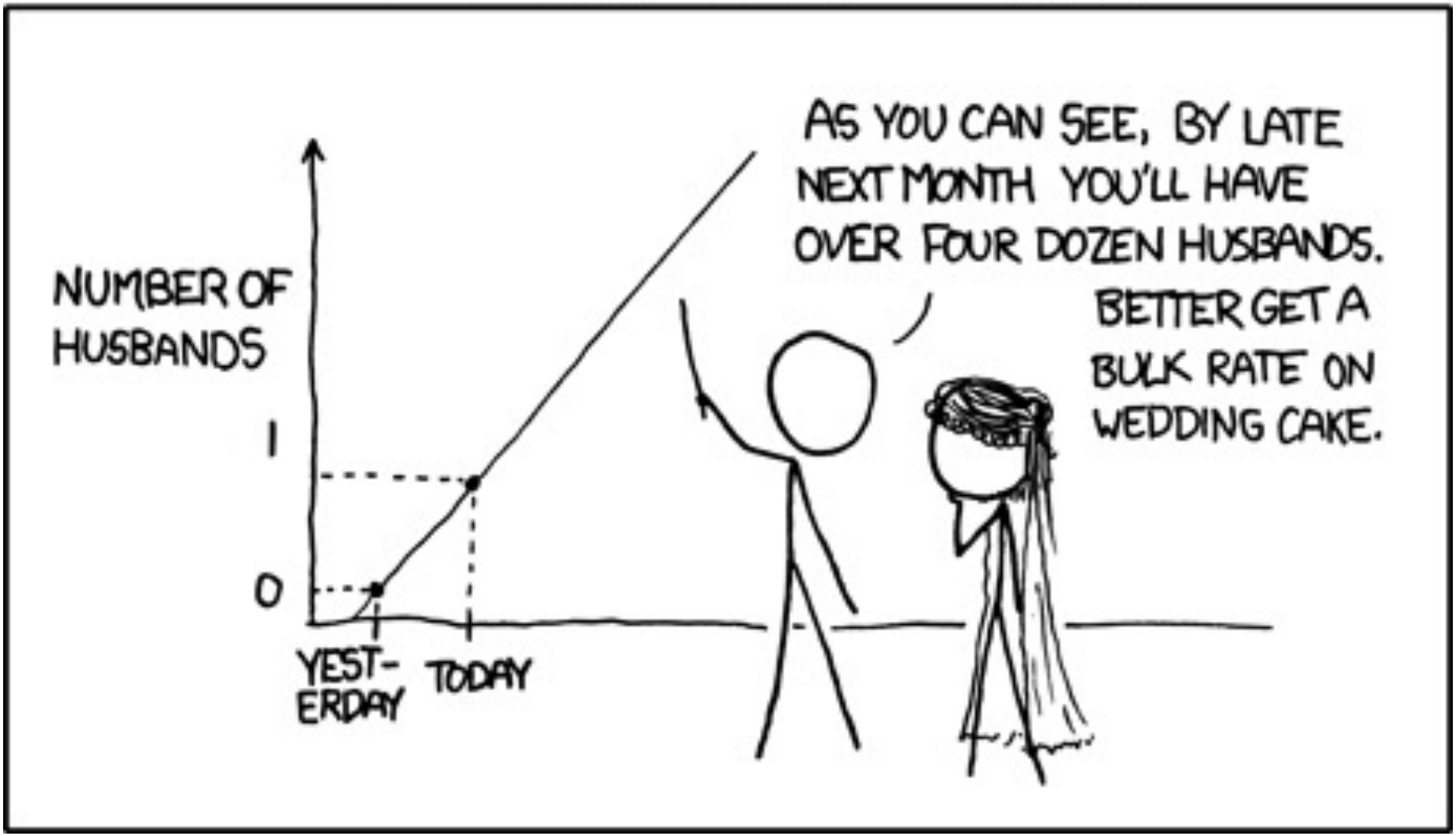
- Having understood the past, there should be some hope of being able to forecast the future.
- This is where we need to distinguish between the signal and the noise.
- Being able to identify the underlying signal offers a means of forecasting the future.
- In practice this will work as long as the data generating process does not change.

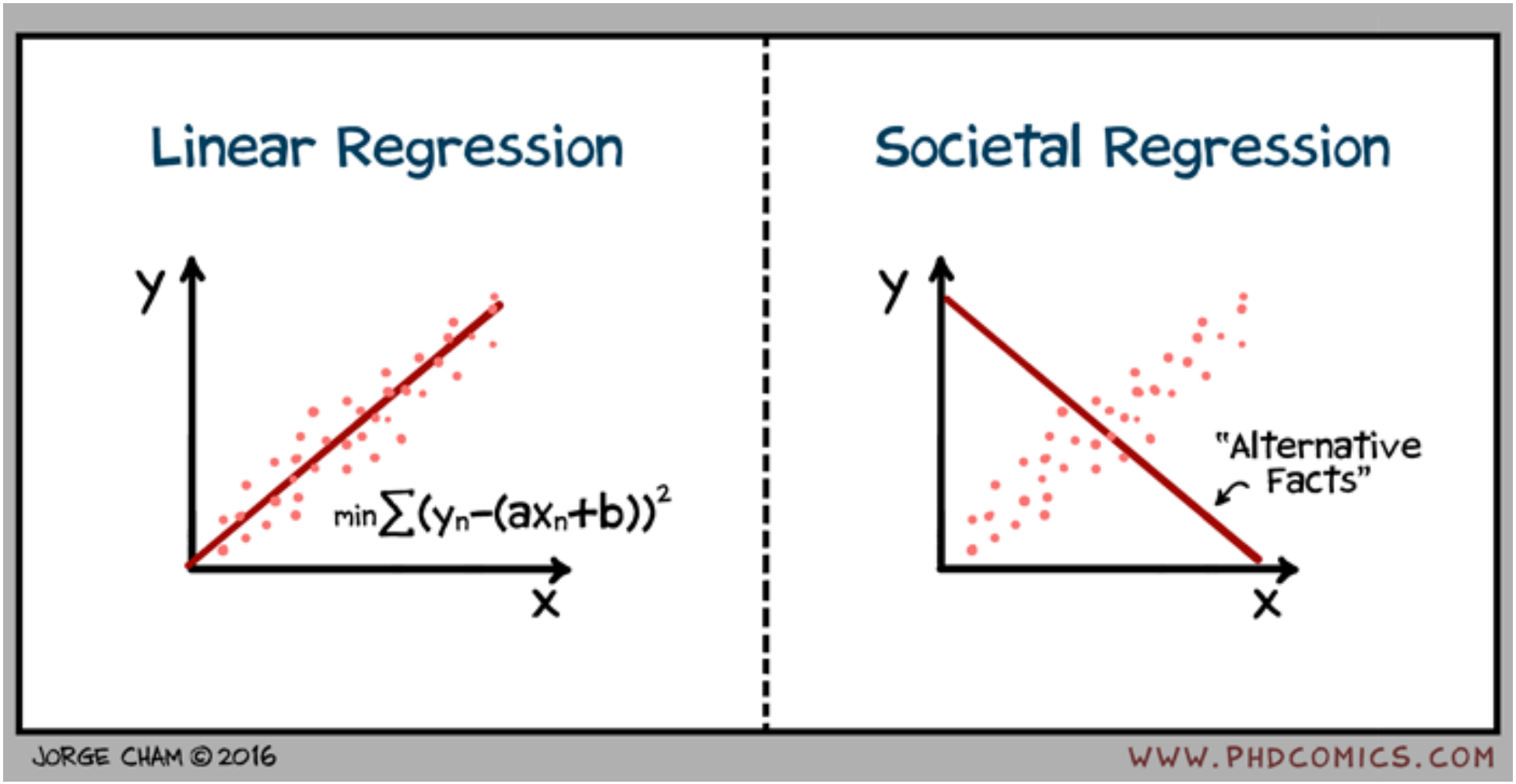
Regression analysis

FITS A STRAIGHT LINE TO THIS MESSY SCATTERPLOT. x IS CALLED THE *INDEPENDENT* OR *PREDICTOR* VARIABLE, AND y IS THE *DEPENDENT* OR *RESPONSE* VARIABLE. THE *REGRESSION* OR *PREDICTION* LINE HAS THE FORM

$$y = a + bx$$

MY HOBBY: EXTRAPOLATING




Linear Regression
y
x
$\min\sum(y_n-(ax_n+b))^2$
Societal Regression
y
x
"Alternative Facts"
JORGE CHAM © 2016
WWW.PHDCOMICS.COM

How much should Lego cost?

- Lego tends to behave a little like a commodity in that a certain weight of Lego has a particular price attached to it.
- Lego does not appear to lose value with age.
- We can use EBay to collect auction data and think about price discovery.
- How much should a given weight of Lego cost us when purchasing on EBay?

Lego price versus weight

- The price of lego increases linearly with the weight:

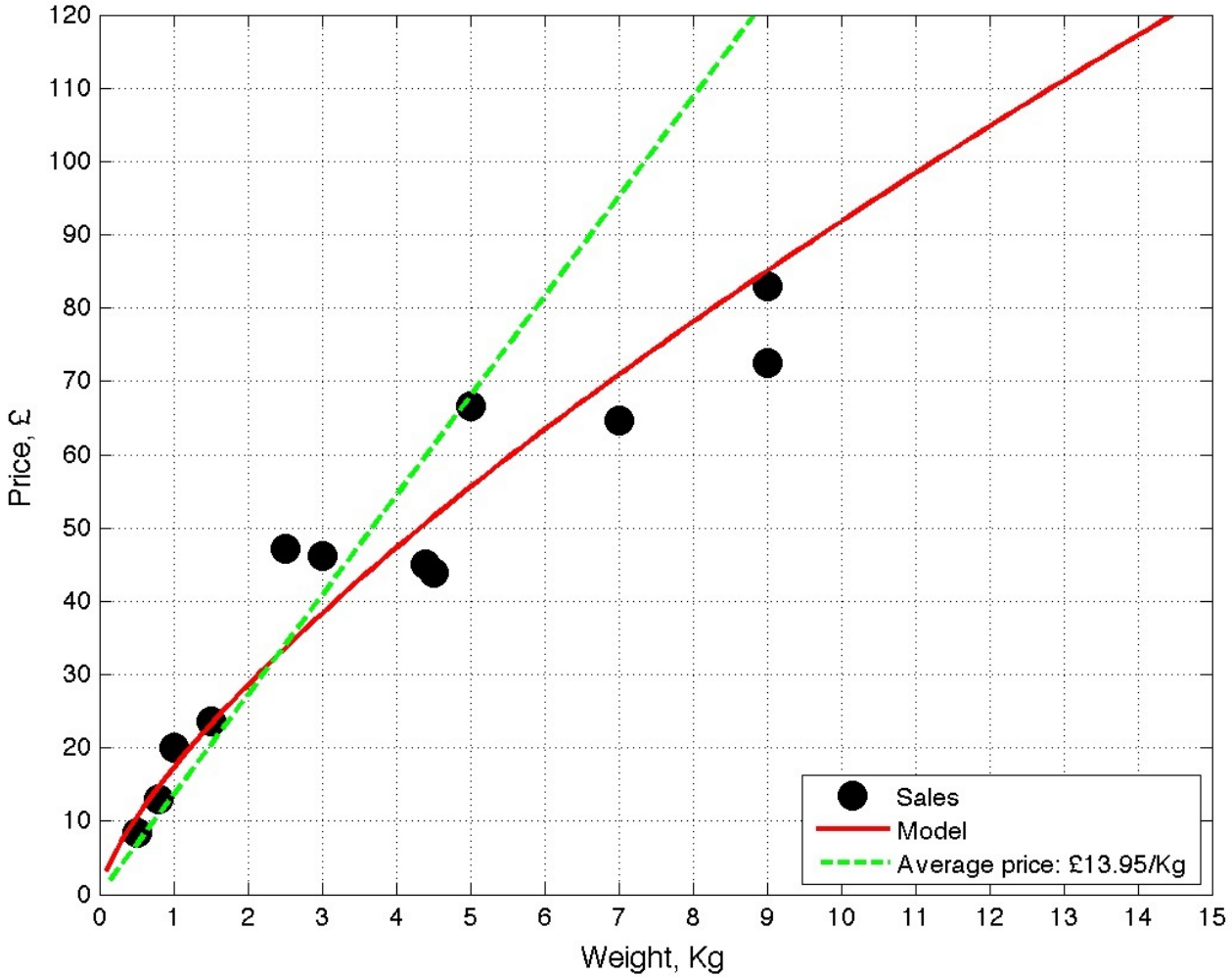
  Price = a*Weight

- www.slido.com
- #73026

Price versus Weight

- Model A: Price = a*Weight
- Model B: log(Price) = a + b*log(Weight)
- The nonlinearity in Model B allows the Price to decrease for relatively large weights of Lego.
- This reflects the fact that more people are interested in small quantities of Lego than large quantities.
- Supply and demand is relevant for Lego.
- Resellers can therefore purchase large quantities and then separate into smaller amounts.

Lego Analysis



Wine Quality

- Mouton Rothschild
- Vintage: 2000
- Price: $2,600

- In order to predict wine quality, which variables would you collect?
- www.slido.com

  #73026

Wine Quality

- The traditional approach for measuring the quality of wine involves the "swishing and spitting" technique of wine gurus such as Robert Parker to predict auction prices.
- Bordeaux are best when the grapes are ripe and their juice is concentrated.
- In years when the summer is hot, grapes get ripe.
- In years of below-average rainfall, the fruit gets concentrated.
- It's in the hot and dry years that you tend to get the legendary vintages.

Ayres, Ian (2007). Man vs machine - Grape expectations: the price of wine. FT 01-Sep-2007.

Wine formula

- A US Economist, Orley Ashenfelter decided to build on these facts.
- He put these facts into a formula that offered a means of predicting wine quality based on the weather without having to taset it:
- Wine quality = 12.145 + 0.00117 winter rainfall + 0.0614 average growing season temperature - 0.00386 harvest rainfall.

Ashenfelter (2008). Predicting the quality and prices of bordeaux wine. Economics Journal

Angry traditional wine critics

- Traditional wine critics were not pleased. Britain's Wine magazine said "the formula's self-evident silliness invite[s] disrespect".
- When Ashenfelter gave a wine presentation at Christie's wine department, dealers in the back hissed.
- And Parker said Ashenfelter was "rather like a movie critic who never goes to see the movie but tells you how good it is based on the actors and the director".

The result

- Bordeaux spend 18-24 months in oak casks before they are set aside for ageing in bottles.
- Experts have to wait four months just to have a first taste, after the wine is placed in barrels. And even then it's a rather foul, fermenting mixture.
- It's far from clear that tasting this undrinkable early wine offers accurate information about the wine's future quality.
- Ashenfelter's predictions were astonishingly accurate.
- To date, few wine experts have publicly acknowledged the power of Ashenfelter's predictions.
- Wine experts forecasts now correspond much more closely to the the outcome of Ashenfelter’s simple equation.

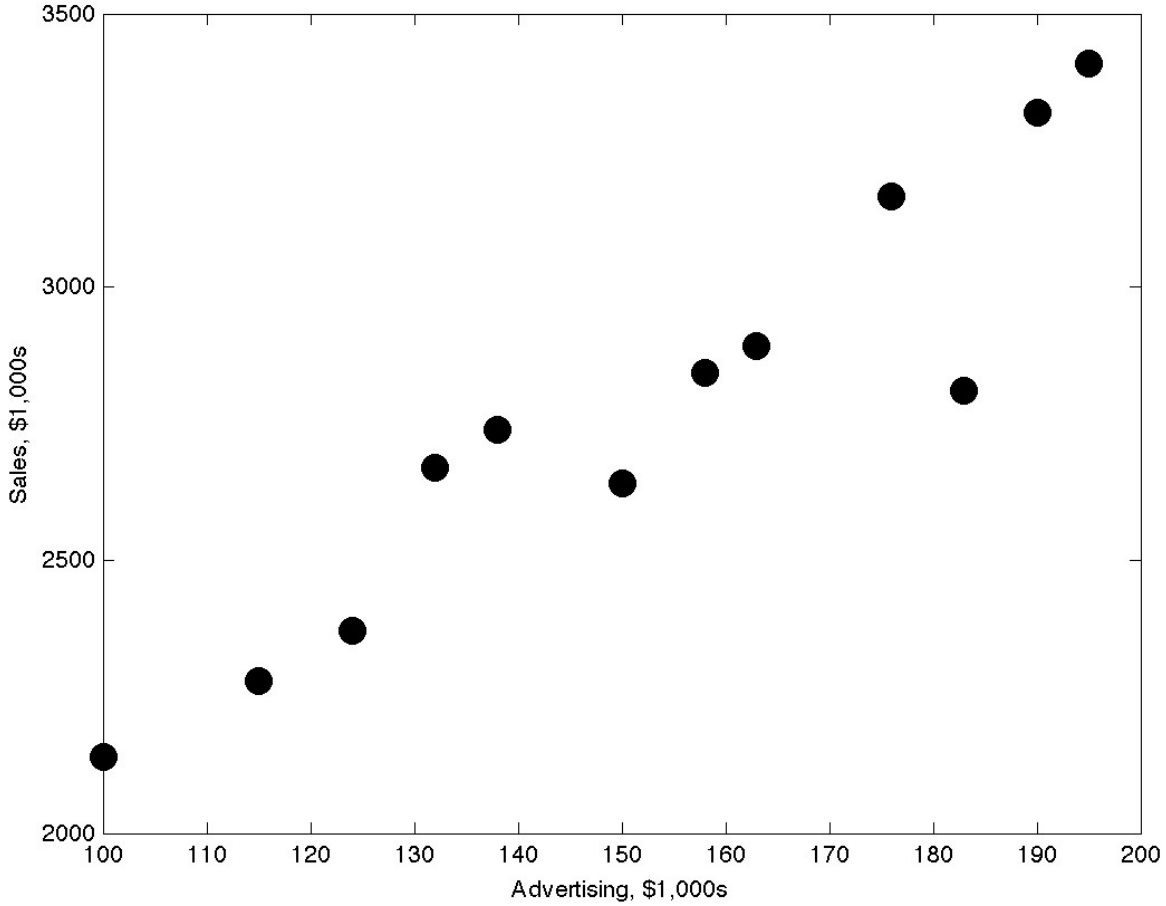
Case Study – Sales & Advertising

- The typical range of spending on marketing and advertising is between 1% and 10% of gross revenue.
- The U.S. Small Business Administration recommends spending 7 to 8 percent of your gross revenue for marketing and advertising if you're doing less than $5 million a year in sales.
- Start-ups and small businesses usually allocate between 2% and 3% of revenue for marketing and advertising.
- But this figure could be as much as 20% if you are operating in a competitive industry.

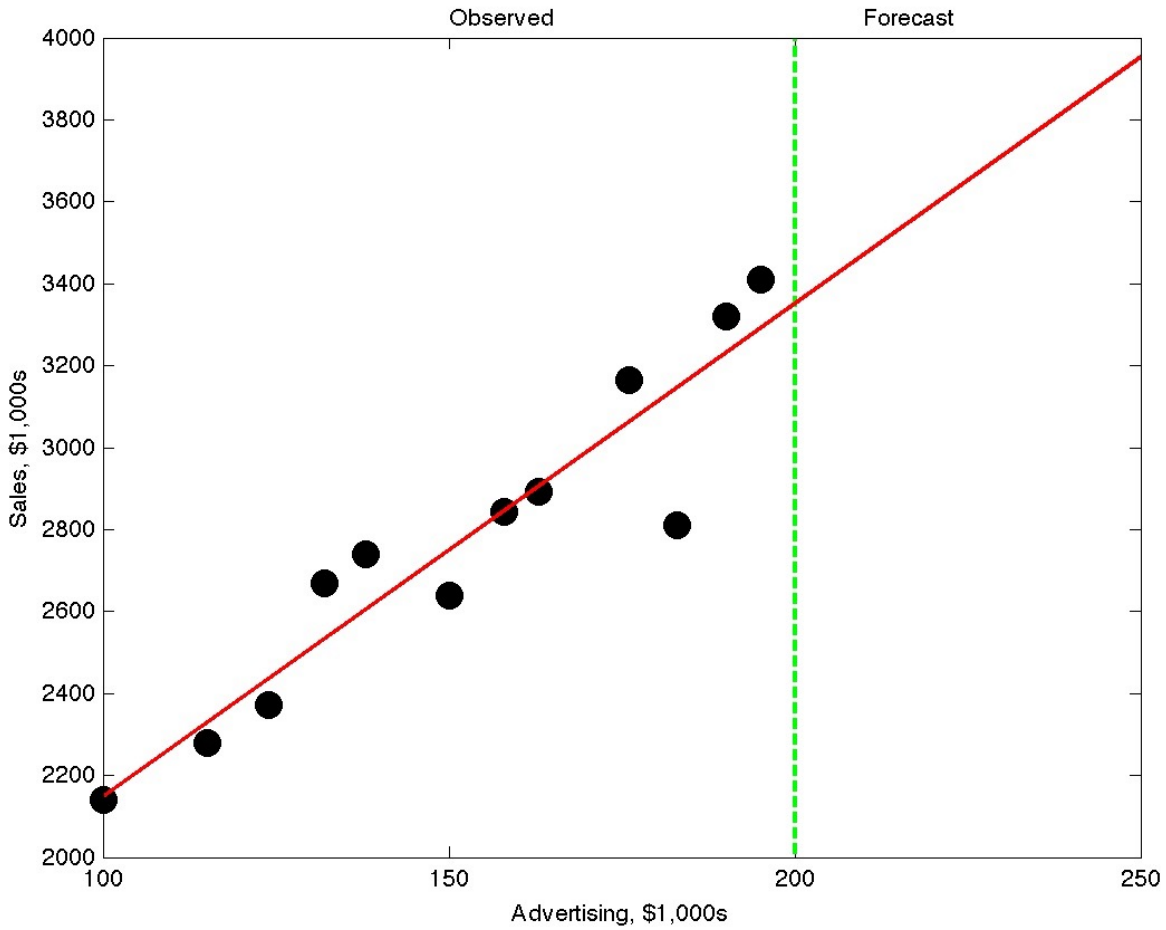
Sales and Advertising

Month	Advertising, $1000	Sales, $1000
Jan	100	2140
Feb	115	2279
Mar	132	2670
Apr	124	2371
May	150	2640
Jun	138	2739
Jul	163	2892
Aug	176	3166
Sep	158	2843
Oct	190	3320
Nov	183	2811
Dec	195	3410

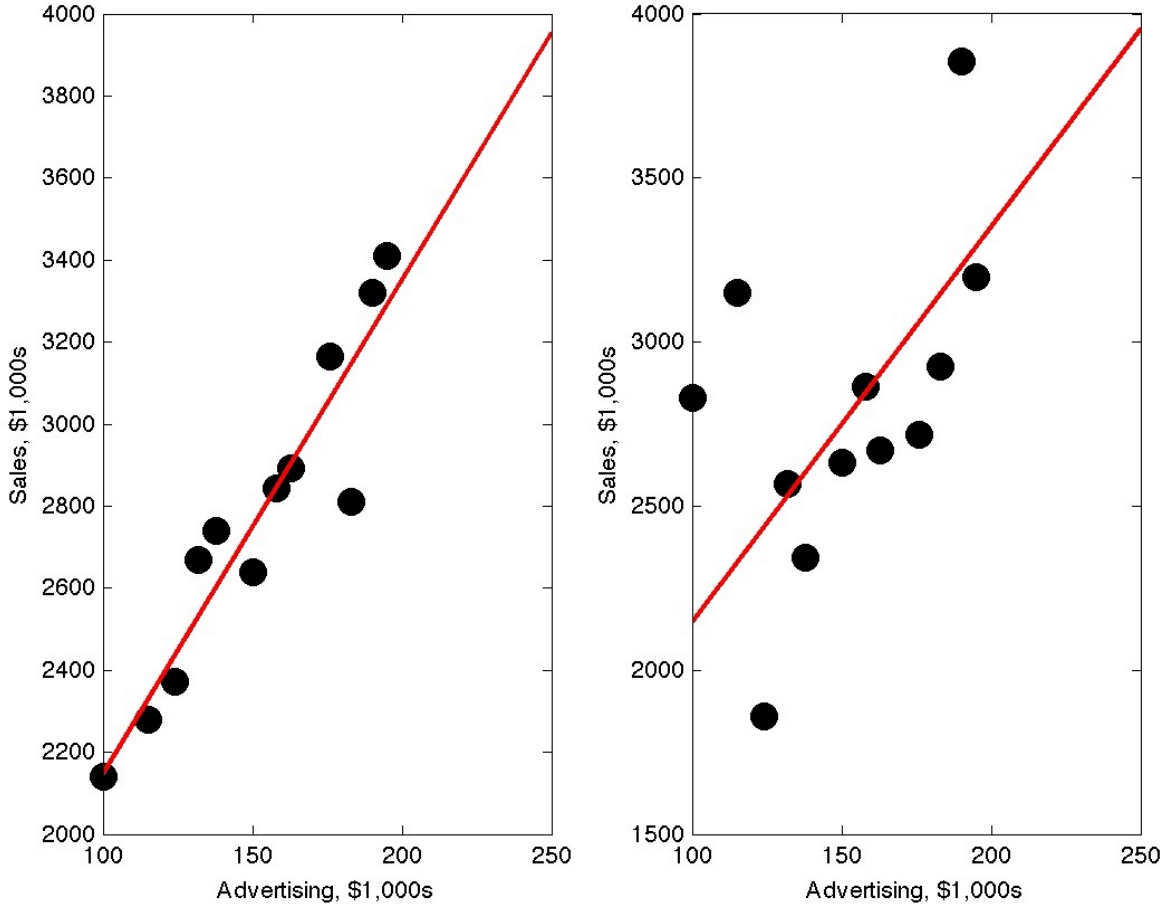
Graph of Sales versus Advertising



Sales = 943 + 12*Advertising



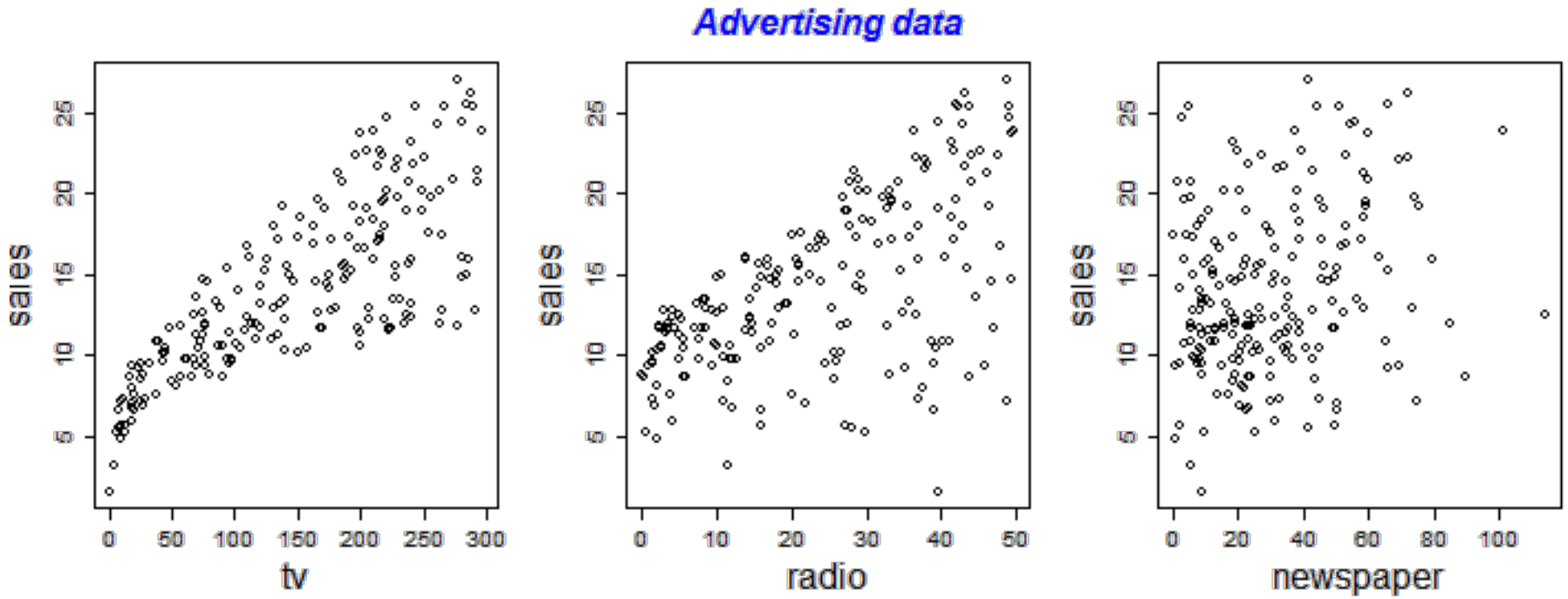
Comparison



As CEO, which graph would give you most confidence to increase the advertising spend? (Left or Right)

www.slido.com
#73026

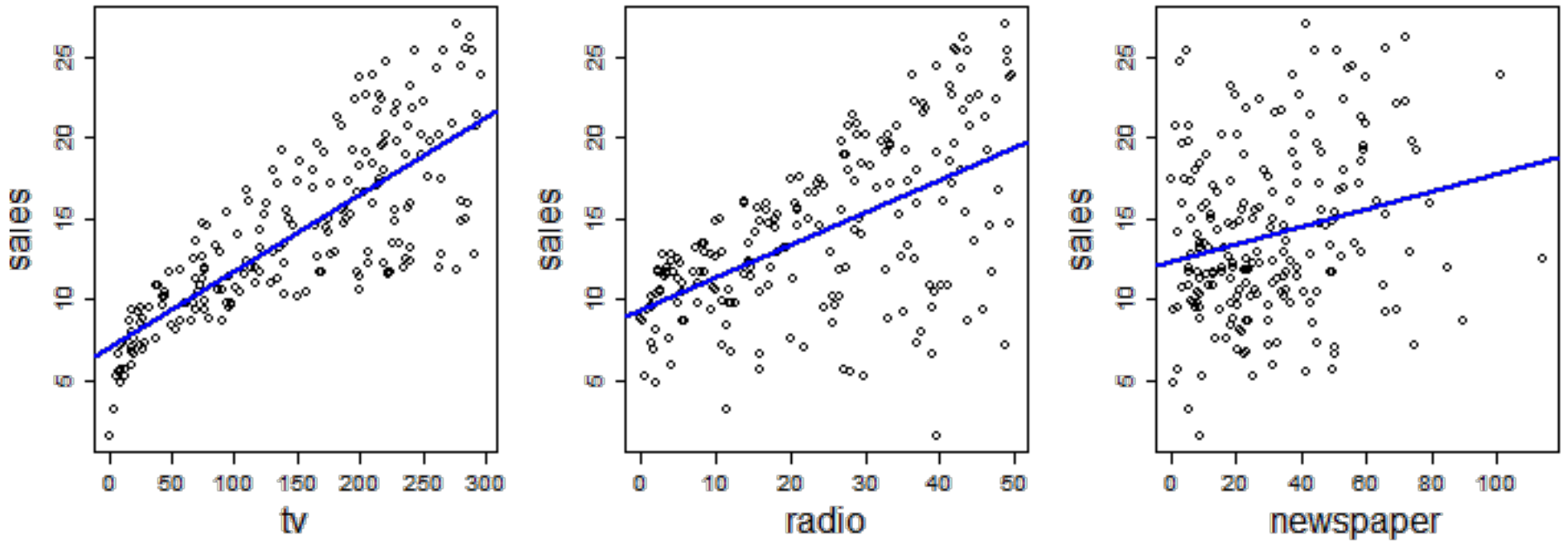
Advertising media



As CEO a recession causes you to stop spending on one type of advertising:
TV or Radio or Newspaper

www.slido.com
#73026

Linear regression



Source: http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv

Advertising Questions

- 1. Is there a relationship between advertising budget and sales?
- 2. How strong is the relationship between advertising budget and sales?
- 3. Which media contribute to sales?
- 4. How accurately can we estimate the effect of each medium on sales?
- 5. How accurately can we predict future sales?
- 6. Is the relationship linear?
- 7. Is there synergy (interaction) among the advertising media?

Correlation

- The correlation between two random variables *X* and *Y* is given by

$$\rho_{XY} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sigma_X \sigma_Y}$$

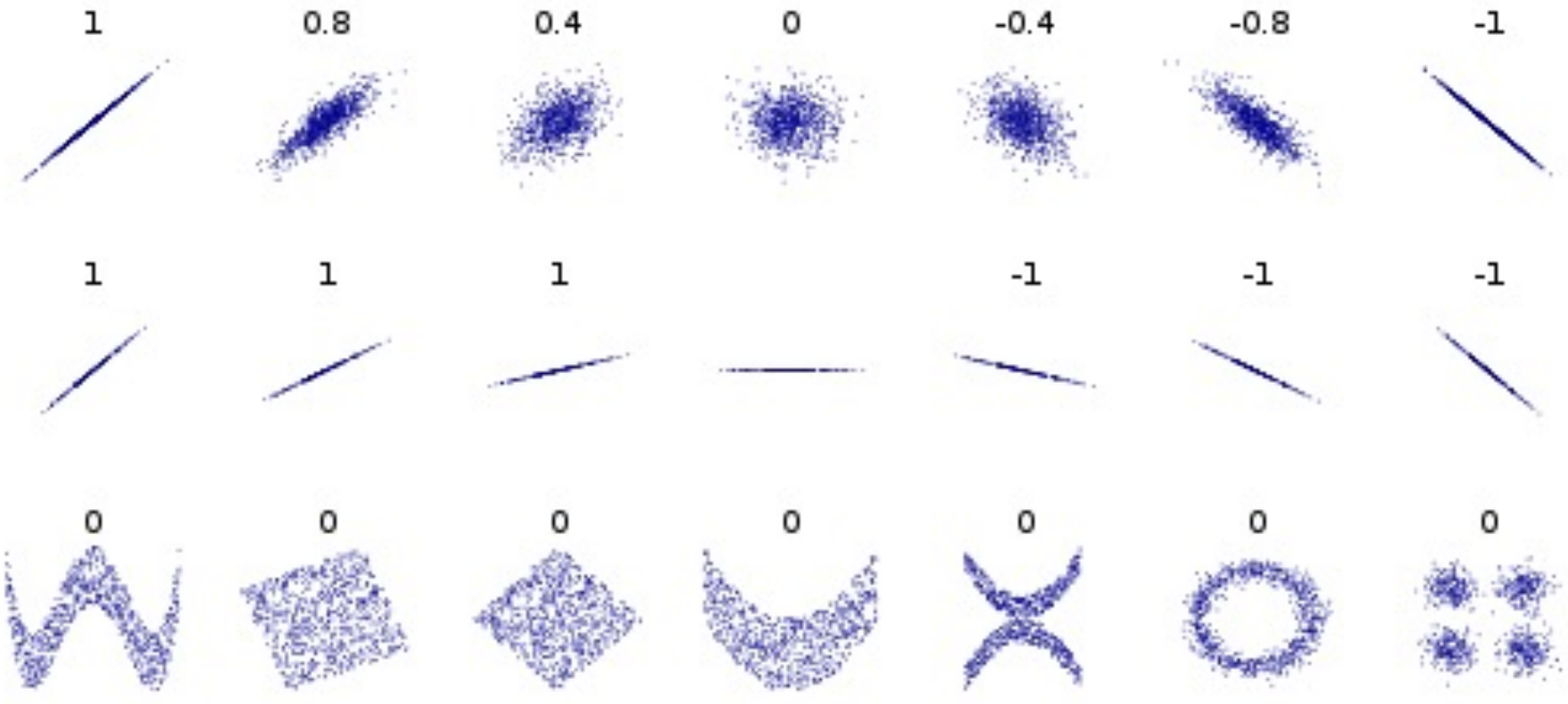
- For a time series of measurements of *X* and *Y* (x_i and y_i where $i = 1, 2, ..., n$), the sample correlation coefficient is

$$r_{xy} = \frac{\frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

  where $\bar{x}$ and $\bar{y}$ are the sample averages and s_x and s_y are sample standard deviations

- The correlation coefficient is 1 for perfectly correlated variables, -1 for anti-correlation and 0 for no correlation

Correlation examples



Linear regression outputs

- Given measurements of an independent variable x_n and dependent variable y_n, we can fit a linear model such that

$$y_n = a + bx_n + \varepsilon_n$$

- where
- a is the intercept (also known as constant);
- b is the slope (indicates how y depends on x);
- and the ε_n are the model errors or residuals.

Parameter estimation

- Given a dataset of interest and a model structure that we believe is appropriate, the next step is to estimate the model parameters.
- We wish to estimate the parameters such that the model provides a good fit to the data.
- This implies that the model is capable of describing the historical data that we have observed.

Maximum likelihood

- Given a set of observations and an underlying probability model, maximum likelihood identifies the values of the model parameters that are most likely to have generated the observations
- The likelihood function expresses the probability of generating the time series x_i with parameter θ:

$$L(\theta) = f_\theta(x_1, \ldots, x_n \mid \theta)$$

- If one assumes that the data drawn from a particular independent, identically distributed (IID) distribution:

$$L(\theta) = \prod_{i=1}^{N} f_\theta(x_i \mid \theta)$$

Maximum likelihood with E~N(0,σ^2)

- Assume that the model errors are normally distributed:

$$p(E_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{E_i^2}{2\sigma^2}\right)$$

- Form the likelihood based on the model errors:

$$L = \prod_{i=1}^{N} p(E_i)$$

- Take the logarithm:

$$\ln L = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N} E_i^2$$

- Maximising *lnL* corresponds to minimizing least squares subject to the errors being IND

Data: response and predictors

- Consider the response (dependent variable):

  $\mathbf{y} = (y_1, \ldots, y_n)^T$

  and a n x p model matrix $\mathbf{X}$ defined by:

  $\mathbf{X} = [\mathbf{x}_1 | \ldots | \mathbf{x}_p]$

  containing predictors (explanatory variables):

  $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T, j = 1, \ldots, p$

Linear regression

- Given *p* predictors $\mathbf{x}_1, \ldots, \mathbf{x}_p$, the response **y** is predicted by

$$\hat{\mathbf{y}} = \hat{\beta}_0 + \mathbf{x}_1\hat{\beta}_1 + \cdots + \mathbf{x}_p\hat{\beta}_p$$

- A model fitting procedure produces the vector of parameters

$$\hat{\beta} = \left[\hat{\beta}_0, \cdots, \hat{\beta}_p\right]$$

Ordinary Least Squares

- We define the ordinary least squares criterion as:

$$L(\beta) = \left|y - X\beta\right|^2$$

- The ordinary least squares estimator is given by

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, L(\beta)$$

Linear Regression

- Explanatory and dependent variables
- Design matrix;
- Fitting polynomials;
- Solving linear system of equations;
- Statistical significance of variables;

Matlab functions

- polyfit, polyval
- regress, regstats
- pinv
- stepwise

Q&A

Data & Inference

WEEK 6B

Today's Lecture

No.	Activity	Description	Time
1	Challenge	Obtaining an appropriate model	10
2	Discussion	How to select the optimal model?	10
3	Case study	Kaggle	10
4	Analysis	Model evaluation	20
5	Demo	Techniques for evaluation	20
6	Q&A	Questions and feedback	10

What is a good model?

- In general, we speak of a good model as one that is capable of describing the data that has been observed.
- This quality of a particular model is known as the “goodness-of-fit”.
- Goodness-of-fit statistics focus only on summarizing the errors generated by the model and neglect the complexity of the model.

What is an appropriate model?

- The terminology appropriate suggests that we desire to deploy the model for an application.
- Examples include classification or forecasting.
- An appropriate model is one that performs well on the task at hand.
- Having outstanding goodness-of-fit statistics is just one part of decided whether or not a model is appropriate.

Occam’s Razor

- Occam's razor is a principle attributed to the 14th-century English logician and Franciscan friar William of Ockham
- The principle states that a theory should rely on as few assumptions as possible, eliminating those that make no difference to the observable predictions of the theory
- Given multiple competing theories that are equally plausible, the principle of Occam’s Razor suggests selecting the theory that relies on the fewest assumptions

Model parsimony

- While increasing the complexity of a model naturally gives more freedom to provide a better fit to the observations, a model with too many parameters will not distinguish between the generative dynamics that we wish to extract and fluctuations due to factors such as measurement errors, non-stationarity and noise
- We should aim to identify the simplest model that is compatible with the observations
- This provides motivation for seeking a parsimonious model (one with as few parameters as possible)
- "Everything should be made as simple as possible, but not simpler" - Einstein

Overfitting = Memorizing

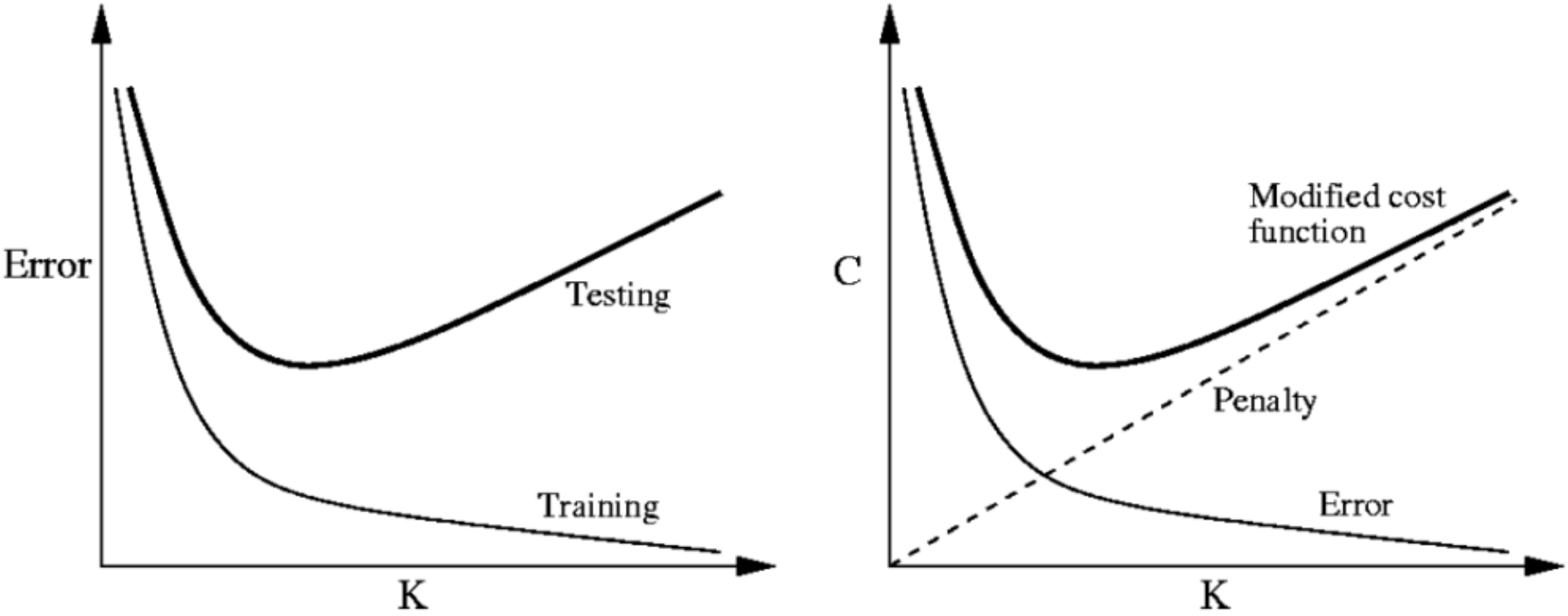
- Overfitting refers to a model that models the training data too well. Instead of learning the general distribution of the data, the model learns the expected output for every data point.
- This is the same a memorizing the answers to a maths problem instead of knowing the formulas. Because of this, the model cannot generalize. Everything is all good as long as you are in familiar territory, but as soon as you step outside, you're lost.

Hackenoon

Over-fitting

- Consider a time series as the sum of a signal from a dynamical process plus observational noise
- When fitting a model to a single sample of time series (in-sample), if we increase the complexity of the model, it will eventually begin to fit the noise
- While the MSE may decrease (in-sample) this will simply indicate that the model is learning about the particular realisation of noise in our one sample of the time series
- This ability of unnecessarily complex models to fit noise is known as over-fitting

In-sample and out-of-sample



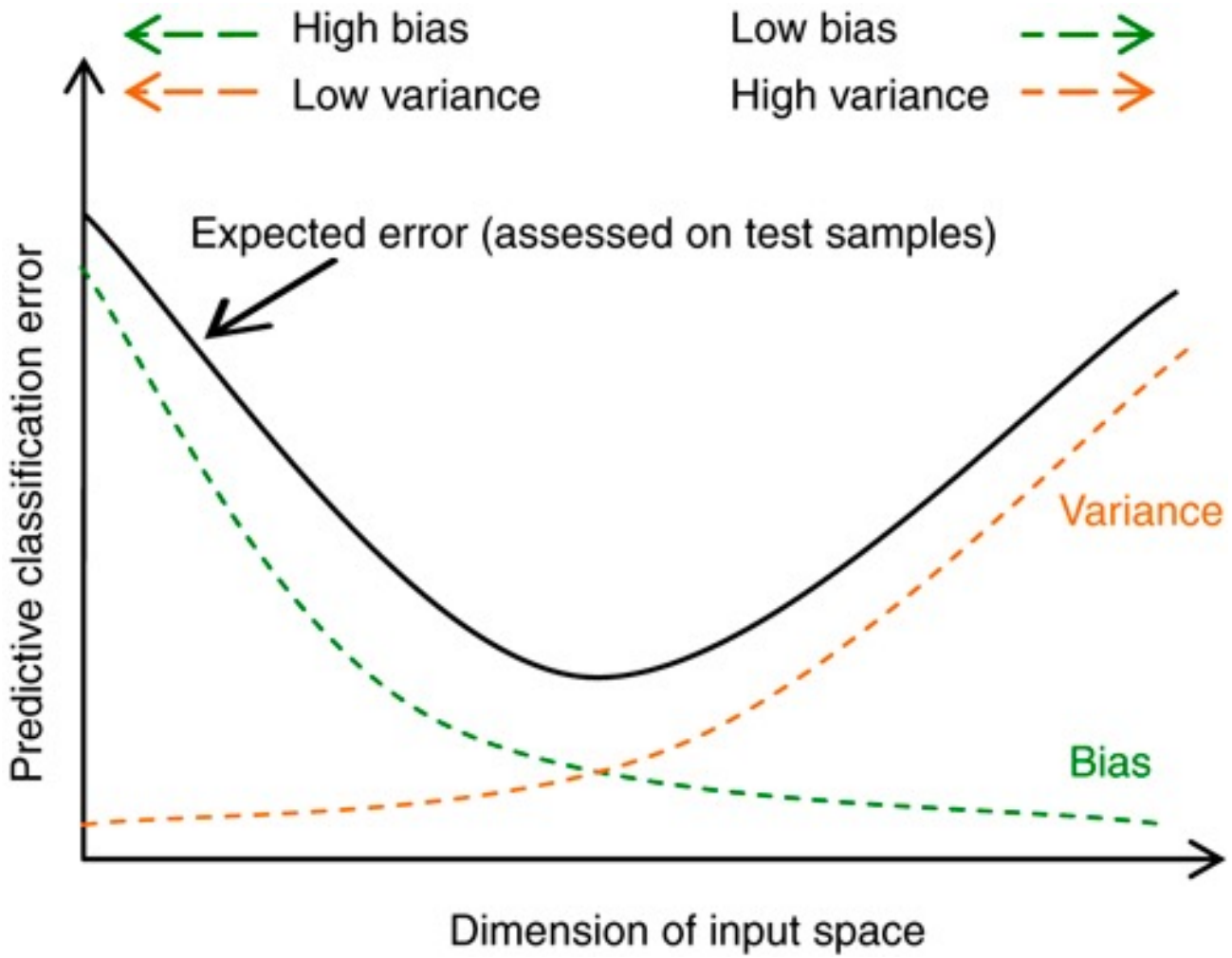
An obvious sign of model over-fitting is one that performs better on training (in-sample) data than testing (out-of-sample) data

Source: McSharry (2005), The danger of wishing for chaos

Bias and Variance

- Bias is error from erroneous assumptions in the learning algorithm.
- High bias can cause an algorithm to miss the relevant relations between features and target outputs (under-fitting).
- Variance is error from sensitivity to small fluctuations in the training set.
- High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (over-fitting).

Bias variance trade-off



Y Wang, D J Miller and R Clarke (2008). Approaches to working in high-dimensional data spaces.

Information criteria

- Information criteria are employed to avoid over-fitting whereby the complexity of the model serves only to fit the noise and not the underlying signal
- By penalising the complexity of the model, it is possible to select a model which is parsimonious
- IC aim to provide a balance between complexity and goodness of fit

Akaike Information Criteria

- Akaike (1974), proposed the AIC as a measure of the goodness of fit of a model:

$$AIC = -2\ln L + 2k$$

where L is the maximised value of the likelihood function for the model and k is the number of parameters

- For normally and independently distributed prediction errors, this may be expressed as

$$AIC = N\ln\left(\frac{RSS}{N}\right) + 2k$$

where RSS is the residual sum of squares with N observations

Akaike, H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19: 716–723.

Bayesian Information Criteria

- Schwarz (1978) proposed the Schwarz or Bayesian IC is a measure of the goodness of fit of a model:

$$BIC = -2 \ln L + k \ln N$$

where L is the maximised value of the likelihood function for the model with N observations and k is the number of parameters

- For normally and independently distributed prediction errors, this may be expressed as

$$BIC = N \ln\left(\frac{RSS}{N}\right) + k \ln N$$

where RSS is the residual sum of squares

- The BIC penalizes free parameters more strongly than does the Akaike information criterion

Schwarz, G. 1978. Estimating the dimension of a model. Annals of Statistics 6: 461–464.

Minimum descripton length

- Rissanen (1978) proposed the minimum description length principle as a formalisation of Occam's Razor in which the best hypothesis for a given set of data is the one that leads to the largest compression of the data
- The goal of statistical inference may be cast as trying to find regularity in the data
- Regularity may be identified with 'ability to compress'
- MDL combines these two insights by viewing learning as data compression: it tells us that, for a given set of hypotheses H and data set D, we should try to find the hypothesis or combination of hypotheses in H that compresses D the most
- In many cases, MDL model selection coincides with BIC

Mallows C_p Statistic

- Mallows C_p statistic provides a stopping rule for various forms of stepwise regression and is given by

$$C_p = SSres/MSres - N + 2p$$

- SSres is the residual sum of squares for the model with *N* observations and *p* regressors,
- MSres is the residual mean square when using all available variables,
- The model with the lowest C_p value is the most "adequate" model.

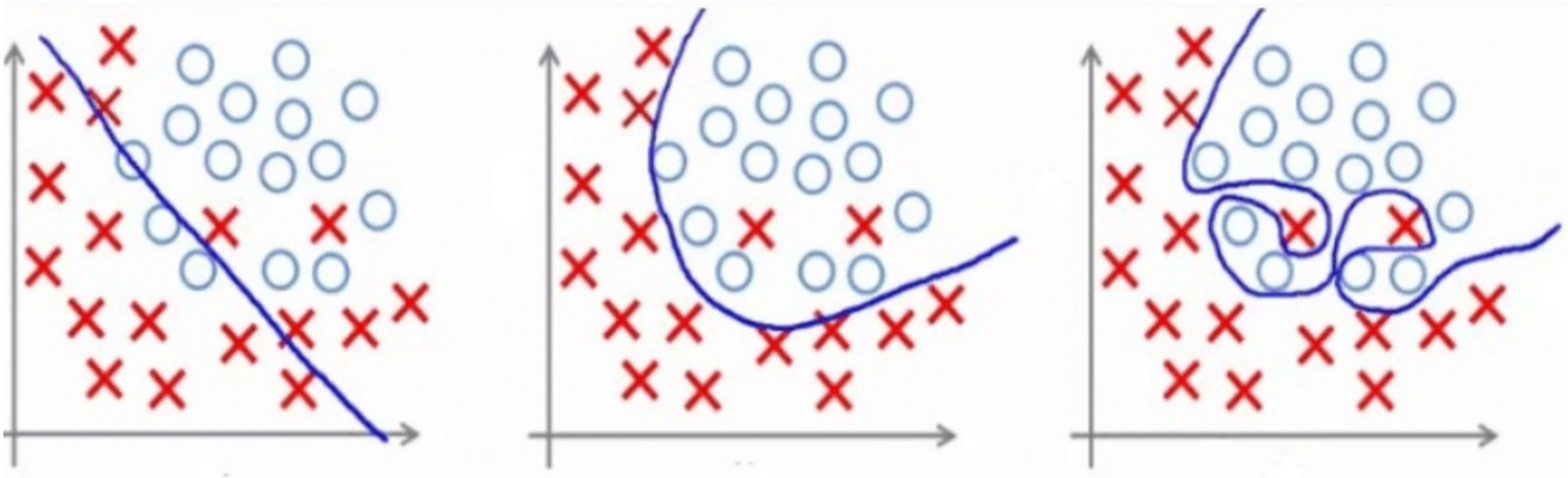
Step-wise variable selection

- There are many ways of selecting variable for inclusion in a model
- Backward or forward step-wise approaches tend to involve:
- General to specific: start with all the variables included and reject variables one by one
- Specific to general: start with no variable and include one variable at a time
- In both cases a condition for the optimal fit provides a stopping criterion

Step-wise variable selection

- This approach is similar to forward selection
- At each iteration variables which are made obsolete by new additions are removed
- The algorithm stops when nothing new is added or when a term is removed immediately after it was added
- Threshold p values are required for adding a variable ($p = p_{enter}$) and for removing a variable ($p = p_{remove}$)

Which model fit is optimal?



A

B

C

www.slido.com
#15489

Case Study - Kaggle

- Kaggle uses a crowdsourcing approach and runs public contests to obtain practical solutions for classification and forecasting problems.
- Core to the approach is a clearly specified challenge, data and competition.
- The competitors are motivated both the the financial reward and the glory of winning a prestigious competition.

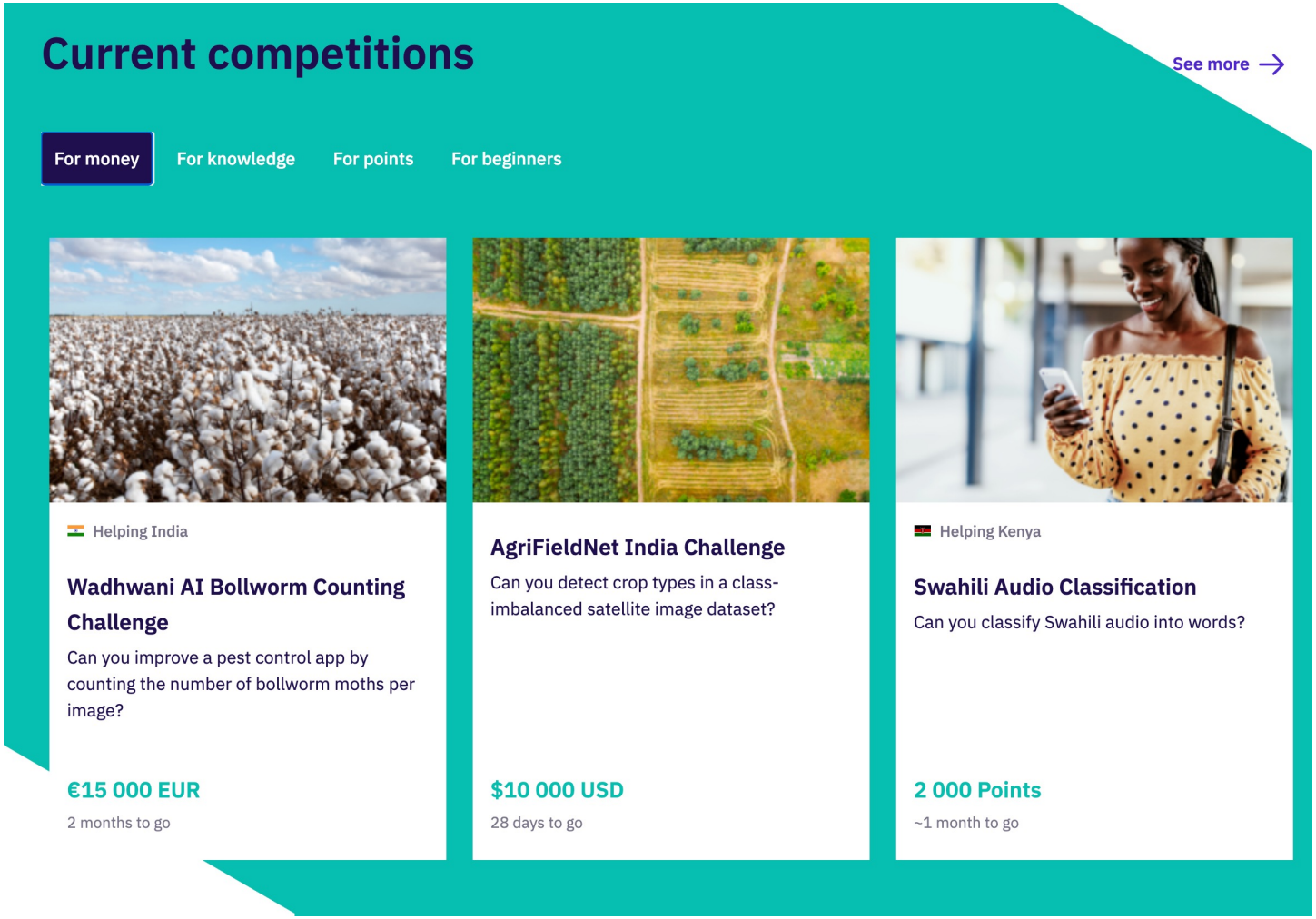
Netflix competition

- Netflix offered a $1 million prize to anyone who could significantly improve its movie recommendation system Cinematch (with an RMSE of 0.9525) by 10%
- The winning team, “BellKor’s Pragmatic Chaos”, a group of 7 individuals, achieved 10.06%
- The runners-up, “Ensemble”, formed from a collection of 28 teams, achieved 10.06%
- A 50/50 blend of the two would have achieved 10.19%

Kaggle: inspired by Netflix

- Anthony Goldbloom, CEO of Kaggle was motivated by the Netflix competition and saw that this approach had great potential.
- He set up Kaggle as a forecasting competition platform in 2010.
- Kaggle makes it easy for any organization to set up a competition and find the best approach for solving their particular challenge.

Zindi



www.zindi.africa

Class Poll

- Which metrics would you use to evaluate and compare predictive models?

www.slido.com
#15489

Forecast benchmarks

- Forecasting is like a horse race
- Any new method may appear useful until it is compared to some simple benchmarks
- These benchmarks serve to establish levels of forecast performance that can be easily achieved without a complicated mathematical model
- They should also be robust

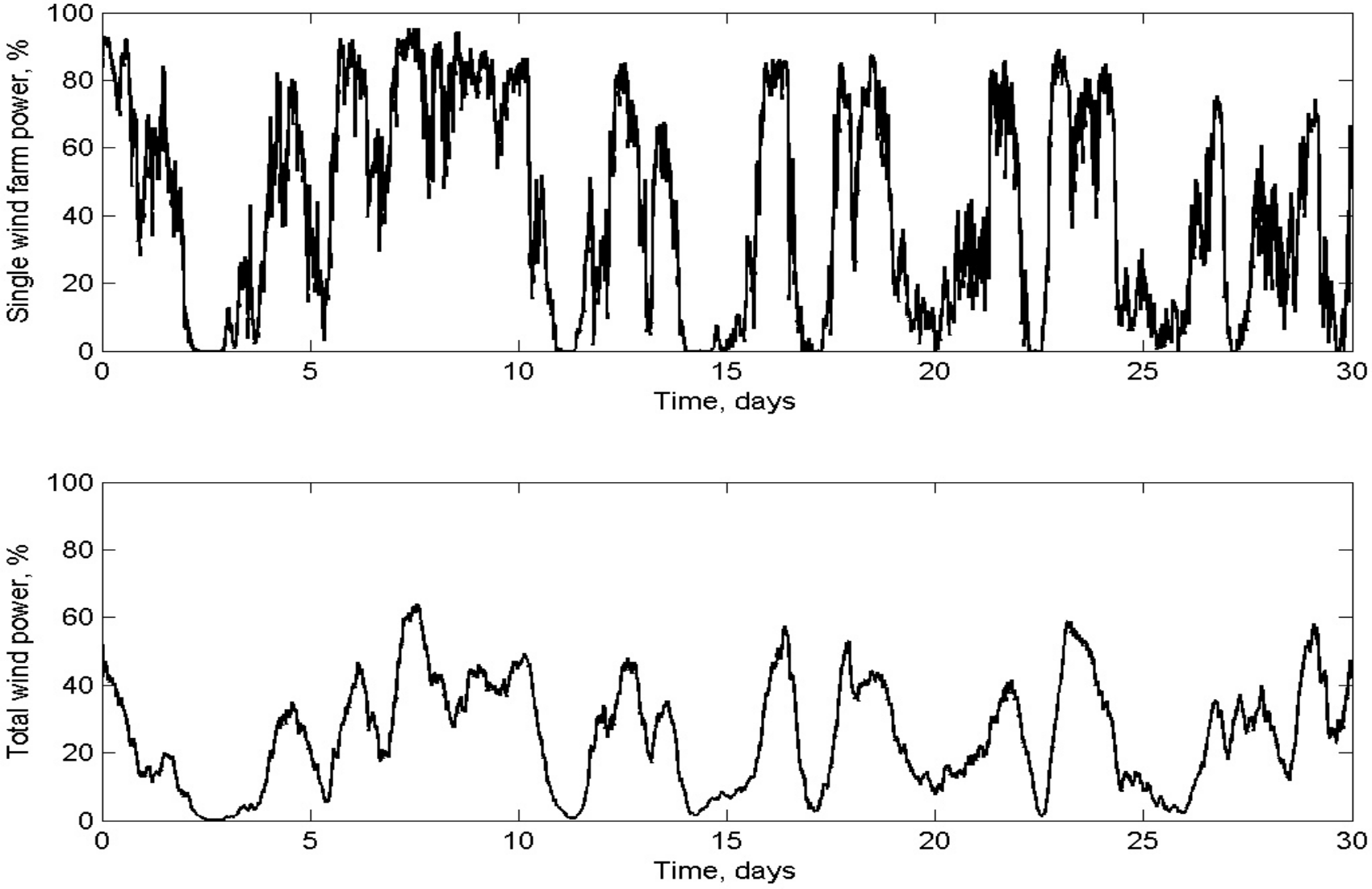
Persistence

- The persistence forecast corresponds to assuming that the underlying dynamics are generated by a random walk
- The best guess forecast is simply the last available observation
- The persistence benchmark is common in meteorology where we should be able to forecast temperatures better than simply looking out the window!
- If the time series are noisy we may take the average of the last *n* observations as a benchmark
- If the time series have an underlying seasonality, then the persistence forecast should take this into account

Unconditional average forecast

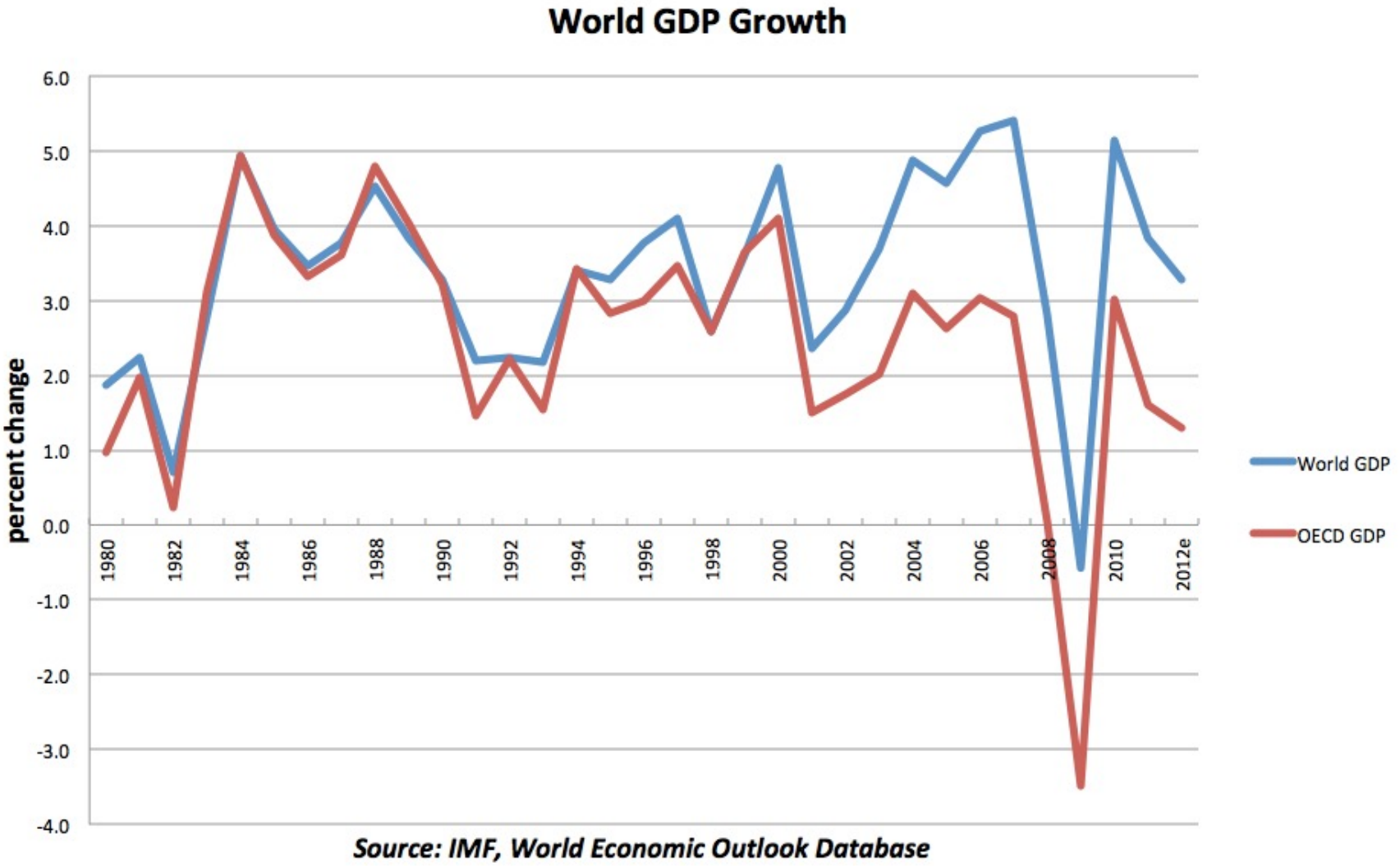
- The unconditional average represent the expected value of the unconditional distribution
- This forecast assumes that the time-ordering of the observations do not provide any additional information
- In meteorology, this is the climate forecast
- It is also important to include seasonal effects

Example of persistence



Source: EirGrid

Example of unconditional average



Benchmark Quiz

- Which of the following is not a suitable benchmark for forecasting?
- A: Moving Average of observations
- B: Neural Network
- C: Median of observations
- D: Average of observations

www.slido.com
#15489

R^2 - coefficient of determination

- The coefficient of determination, R^2, measures the proportion of variability in a data set that is accounted for by a statistical model
- In the case of linear regression, we can decompose the sum of the squares into a part due to the regression and the residuals, such that $SS_{tot} = SS_{reg} + SS_{res}$ where

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \quad SS_{reg} = \sum_i (\hat{y}_i - \bar{y})^2 \quad SS_{res} = \sum_i (y_i - \hat{y}_i)^2$$

- and R^2 is defined as

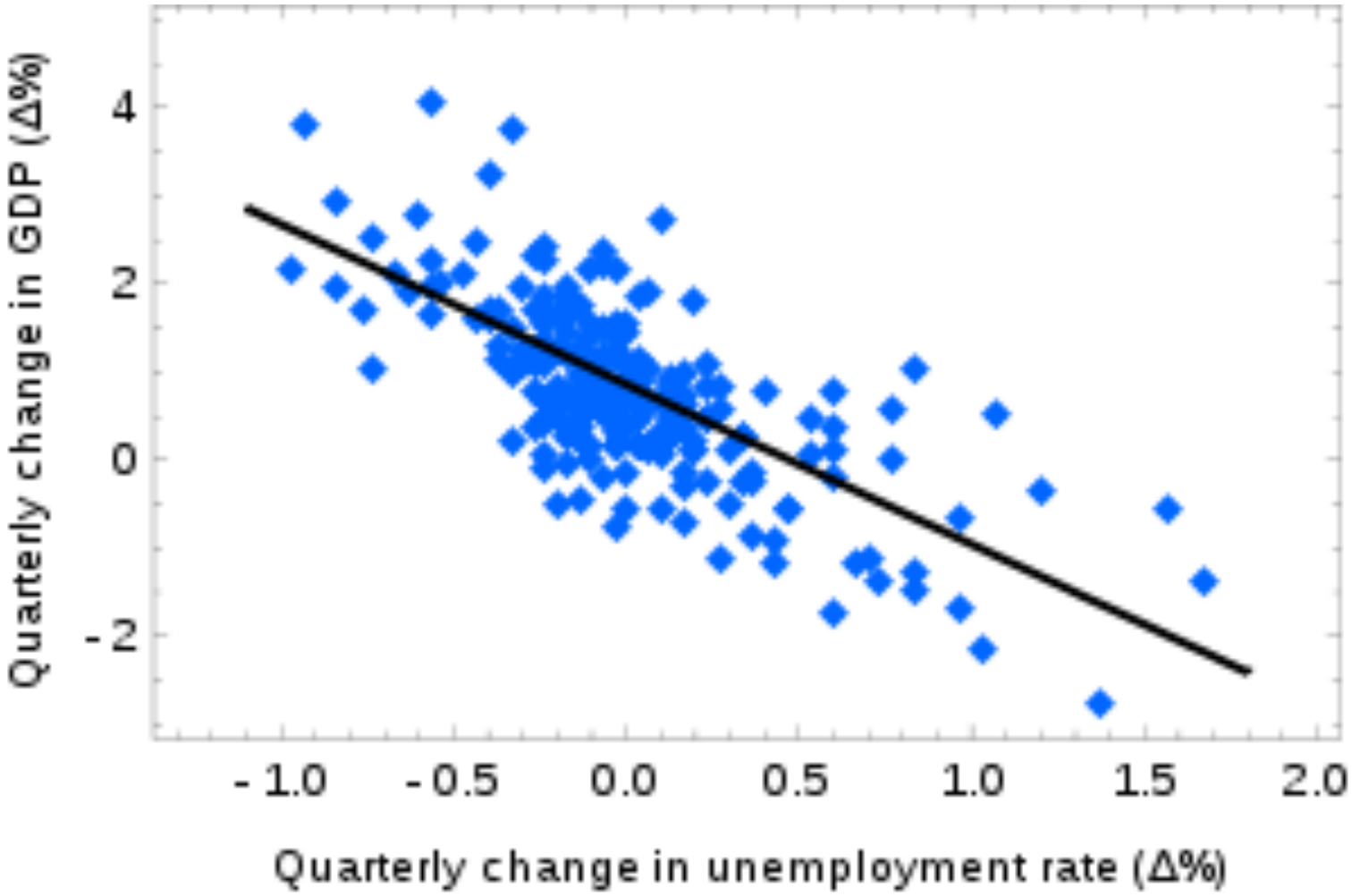
$$R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

- R^2 measures the amount of variance explained by the model given by the ratio of the explained variance (variance of the model's predictions) with the total variance (of the data)

R^2 and correlation

- Coefficient of determination, R^2 is related to the correlation coefficient
- Both attempt to quantify how well a linear model fits to a data set
- The further the points are scattered from the line, the smaller is the value of R^2
- R^2 is the square of the correlation coefficient which is often denoted by r

Okun's Law



Source: wikipedia

Adjusted R^2

- Adjusted R^2 accounts for the fact that the R^2 tends to spuriously increase when extra explanatory variables are added to the model
- R^2 can be written as $R^2 = VAR_{res}/VAR_{tot}$ where $VAR_{res} = SS_{res}/n$ and $VAR_{tot} = SS_{tot}/n$
- Replacing with statistically unbiased estimates

$VAR_{res} = SS_{res}/(n-p-1)$ and $VAR_{tot} = SS_{tot}/(n-1)$:

- Adjusted $R^2 = 1 - [(n-1)/(n-p-1)](1-R^2)$

Mean Squared Error

- The mean-square-error is given by

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

- It represents a measure of forecast performance which is analogous to the least squares parameter estimation technique
- If the forecast errors are not normally distributed, MSE may give misleading results

Mean Absolute Error

- The mean absolute error is given by

$$MAE = \frac{1}{N}\sum_{i=1}^{N} | \hat{y}_i - y_i |$$

- This forecast measure focuses on the magnitude of the errors
- It is more robust than MSE as the large errors are not squared
- It is commonly used in wind energy forecasting and may be given as a fraction of the total energy being generated

MAPE

- The mean absolute percentage error is given by

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{y}_i - y_i}{y_i}\right|$$

- Focusing on the percentage error is useful as a means of standardising the result
- It should only be used if the dependent variable is positive definitive
- This measure is commonly used in energy forecasting

Evaluation & Selection

- Occam's Razor;
- Over-fitting;
- Generalization;
- Model parsimony;
- Information criteria;
- Variable selection

Matlab functions

- stepwise
- dataset, table
- fitlm
- regstats

Q&A