Aalto University
School of Business

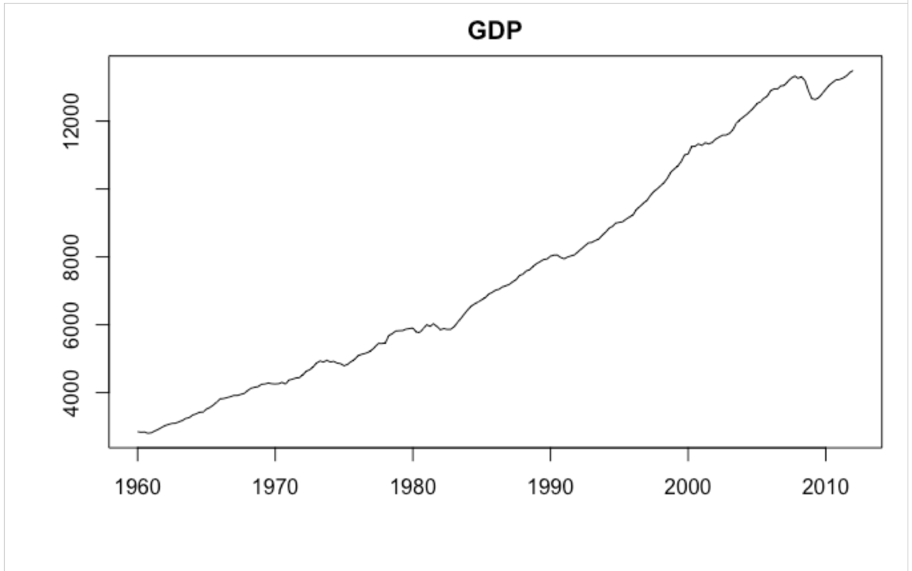
Time series modeling and predictive analytics

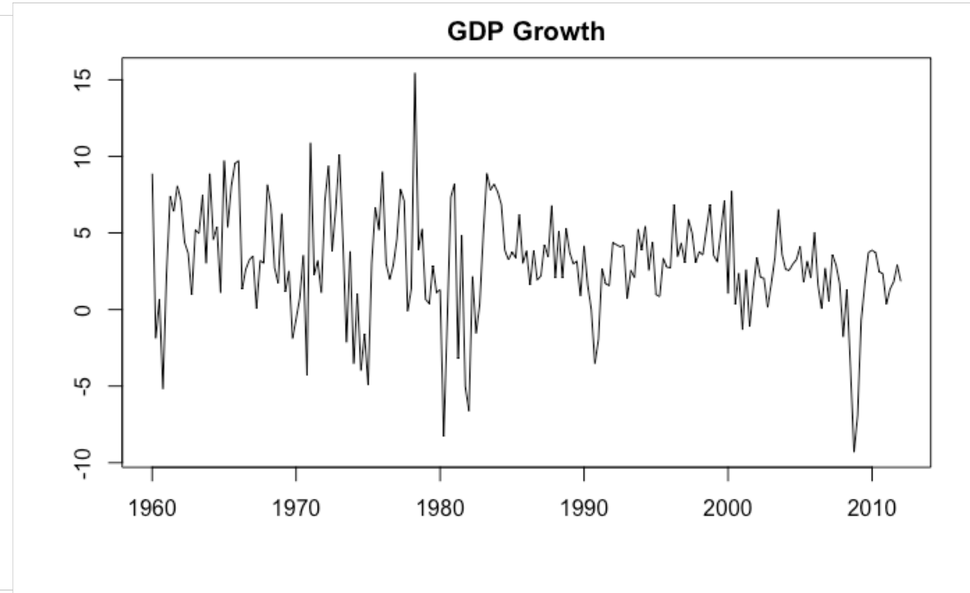
Pekka Malo, Associate Prof. (statistics)
Aalto BIZ / Department of Information and Service Management

Agenda

- **Basic concepts in time series analysis**
- **Stationarity**
- **Forecasting with simple time series models**
- **Feature selection**

A Time Series: US GDP



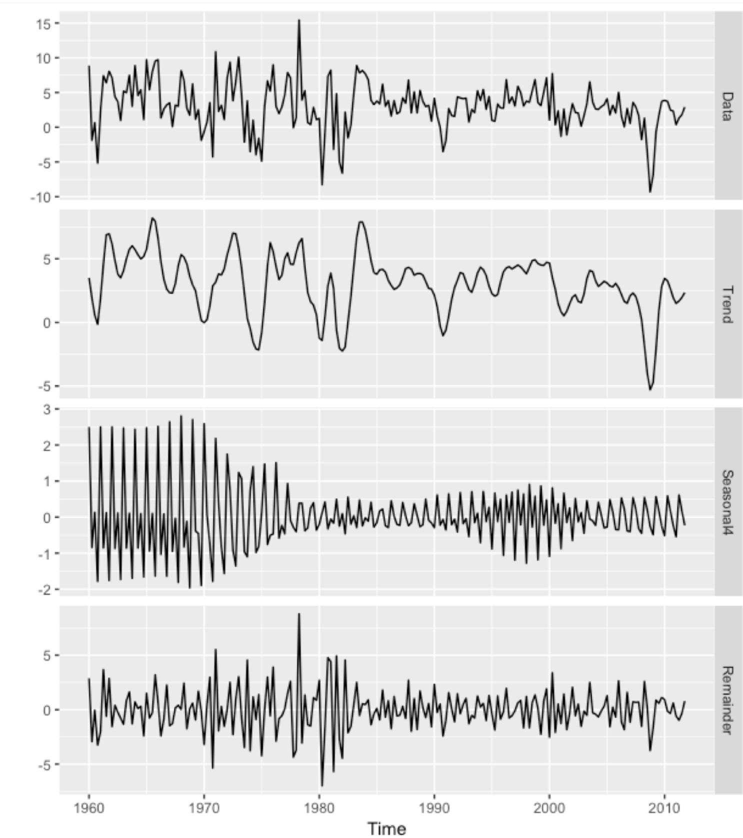


Objectives of time series analysis

1. Compact description of data.
2. Interpretation.
3. Forecasting.
4. Control.
5. Hypothesis testing.
6. Simulation.

Example: Decomposition of US GDP growth

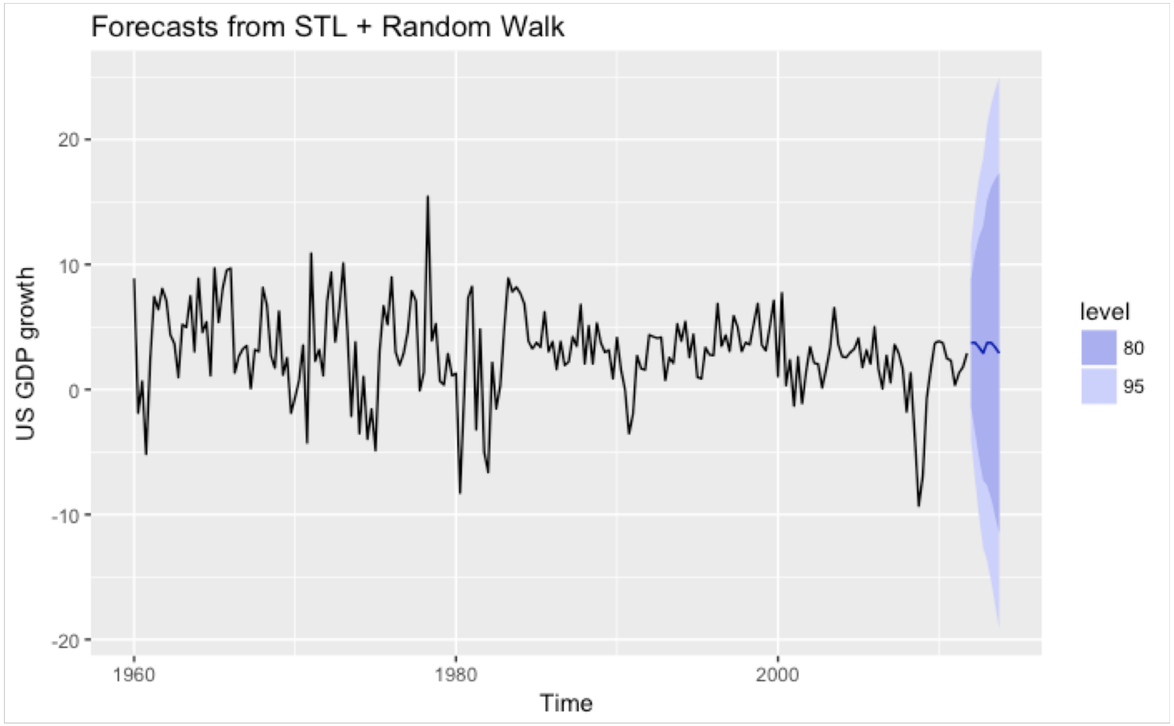
Objectives of time series analysis



1. Compact description of data.
   Example: Classical decomposition: $X_t = T_t + S_t + Y_t.$
2. Interpretation. Example: seasonal adjustment
3. Forecasting. Example: Predict US GDP
4. Control.
5. Hypothesis testing.
6. Simulation.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. J. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, *6*(1), 3–73.
http://www.jos.nu/Articles/abstract.asp?article=613

Naïve forecasts of adjusted data



Time series modeling: Chasing stationarity

Aalto University
School of Business

Typical steps in time series modeling

1. Plot the time series.
   Look for trends, seasonal components, step changes, outliers.
2. Transform data so that residuals are **stationary**.
   (a) Estimate and subtract T_t, S_t.
   (b) Differencing.
   (c) Nonlinear transformations (log, $\sqrt{\cdot}$).
3. Fit model to residuals.

Why do we need to test for non-stationarity?

- If a series is non-stationary, persistence of shocks to the system is infinite (i.e. they never die away)

- Spurious regressions: If two variables are trending over time, a regression of one on the other could have a high R^2 even if the two series are unrelated

- If variables in a regression model are non-stationary, it can be shown that the standard assumptions for asymptotic analysis are not valid --> the usual t-ratios don’t follow t-distribution and we cannot do any reliable hypothesis tests regarding regression parameters

Definition of a time series model

A **time series model** specifies the joint distribution of the sequence $\{X_t\}$ of random variables.
For example:

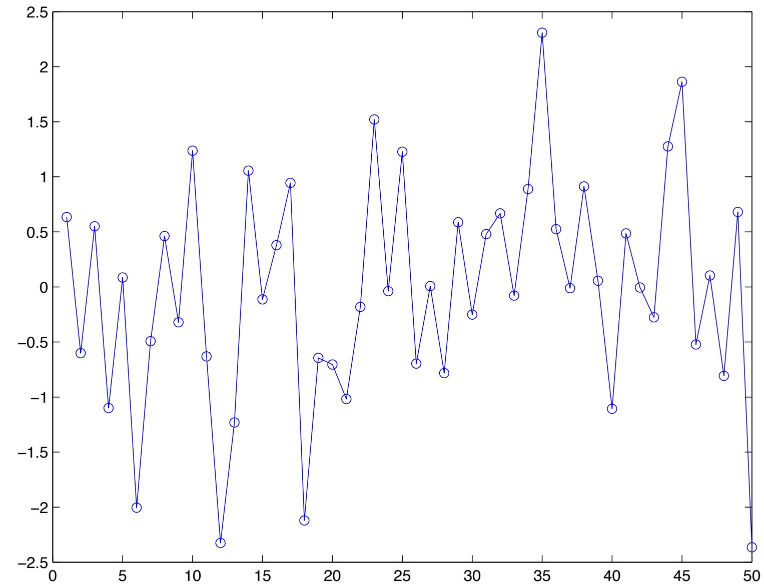
$$P[X_1 \leq x_1, \ldots, X_t \leq x_t] \text{ for all } t \text{ and } x_1, \ldots, x_t.$$

Notation:
$X_1, X_2, \ldots$ is a stochastic process.
$x_1, x_2, \ldots$ is a single realization.

Simple example: White Noise

Gaussian White Noise

$$P[X_t \le x_t] = \Phi(x_t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_t} e^{-x^2/2} dx.$$



Example: White noise: $X_t \sim WN(0, \sigma^2)$.
i.e., $\{X_t\}$ uncorrelated, $\mathrm{E}X_t = 0$, $\mathrm{Var}X_t = \sigma^2$.

Example: i.i.d. noise: $\{X_t\}$ independent and identically distributed.

$$P[X_1 \le x_1, \ldots, X_t \le x_t] = P[X_1 \le x_1] \cdots P[X_t \le x_t].$$

Not interesting for forecasting:

$$P[X_t \le x_t | X_1, \ldots, X_{t-1}] = P[X_t \le x_t].$$

Transformations

It is mathematically equivalent to forecast the given variable or any monotonic transformation of the variable and lagged values

- E.g., it is equivalent to forecast the level of GDP, its logarithm, or percentage growth rate
- Given a forecast of one, we can construct the forecast of the other

Statistically, it is best to forecast a transformation which is close to iid

- For variables such as output and prices, this typically means forecasting growth rates
- For rates, typically means forecasting changes

Stationarity

Definition

Strict stationarity: distributions are time-invariant.

Definition

Weak stationarity: the first two moments (mean and variance) are time-invariant.

Weak stationarity (formally)

Suppose that $\{X_t\}$ is a time series with $\mathrm{E}[X_t^2] < \infty$.
Its **mean function** is

$$\mu_t = \mathrm{E}[X_t].$$

Its **autocovariance function** is

$$\begin{aligned}\gamma_X(s,t) &= \mathrm{Cov}(X_s, X_t) \\ &= \mathrm{E}[(X_s - \mu_s)(X_t - \mu_t)].\end{aligned}$$

We say that $\{X_t\}$ is **(weakly) stationary** if

1. μ_t is independent of t, and
2. For each h, $\gamma_X(t+h, t)$ is independent of t.

Weak stationarity (in practice)

- When plotting a time series, we observe that the series varies around a fixed level within a a finite range!
- The first two moments of the distribution are constants (i.e., mean and variance do not depend on time index)

Example: White Noise model

Example: i.i.d. noise, $\mathrm{E}[X_t] = 0, \mathrm{E}[X_t^2] = \sigma^2$. We have

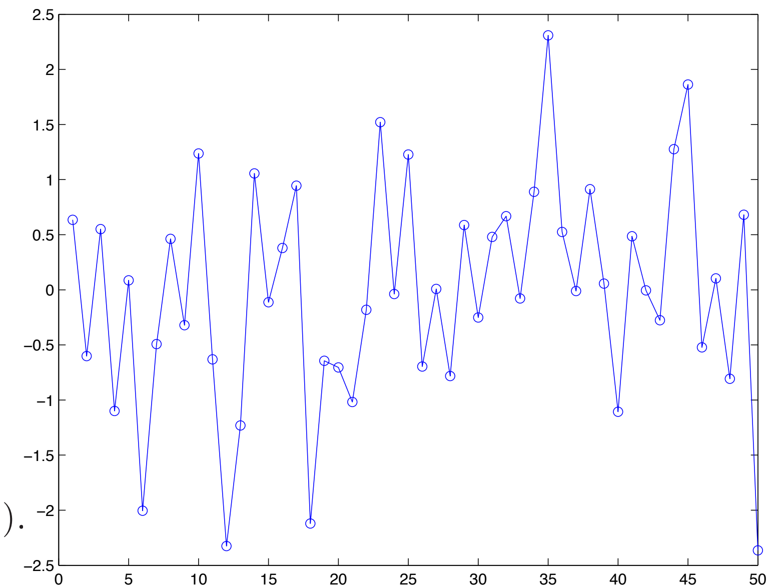
$$\gamma_X(t+h,t) = \begin{cases} \sigma^2 & \text{if } h = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Thus,

1. $\mu_t = 0$ is independent of t.
2. $\gamma_X(t+h,t) = \gamma_X(h,0)$ for all t.

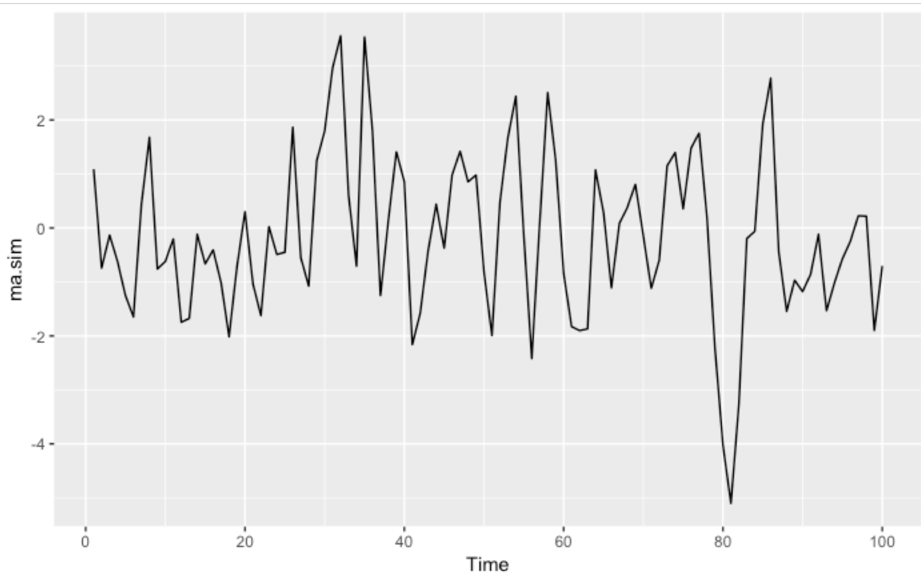
So $\{X_t\}$ is stationary.

Similarly for any white noise (uncorrelated, zero mean), $X_t \sim WN(0, \sigma^2)$.



Example: Moving Average MA(1)-model

MA(1) with $\theta = 0.7, \sigma = 0.1$



$$X_t = W_t + \theta W_{t-1}, \qquad \{W_t\} \sim WN(0, \sigma^2).$$

We have $\mathrm{E}[X_t] = 0$, and

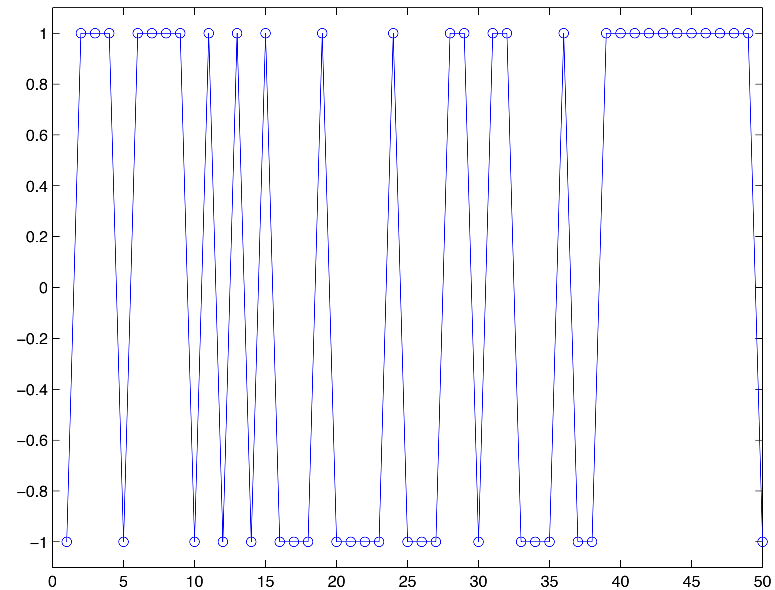
$$\begin{aligned}\gamma_X(t+h, t) &= \mathrm{E}(X_{t+h}X_t) \\ &= \mathrm{E}[(W_{t+h} + \theta W_{t+h-1})(W_t + \theta W_{t-1})] \\ &= \begin{cases} \sigma^2(1+\theta^2) & \text{if } h = 0, \\ \sigma^2\theta & \text{if } h = \pm 1, \\ 0 & \text{otherwise.} \end{cases}\end{aligned}$$

Thus, $\{X_t\}$ is stationary.

Example: Random Walk process

Suppose we use coin-flipping to decide whether to walk one step forward or one step backward. Statistically, we are then following a binary i.i.d model.

$$P[X_t = 1] = P[X_t = -1] = 1/2.$$



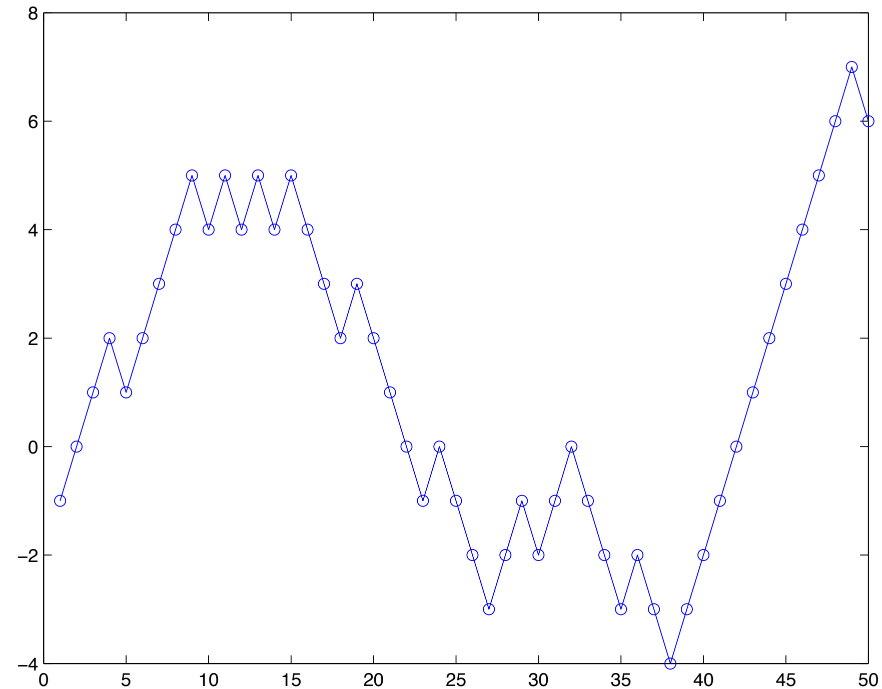
Ref: www.stat.berkeley.edu/~bartlett/courses/153-fall2010

Example random walk (2)

Our path or progress is then a sum of the steps that we have taken

$S_t = \sum_{i=1}^{t} X_i.$

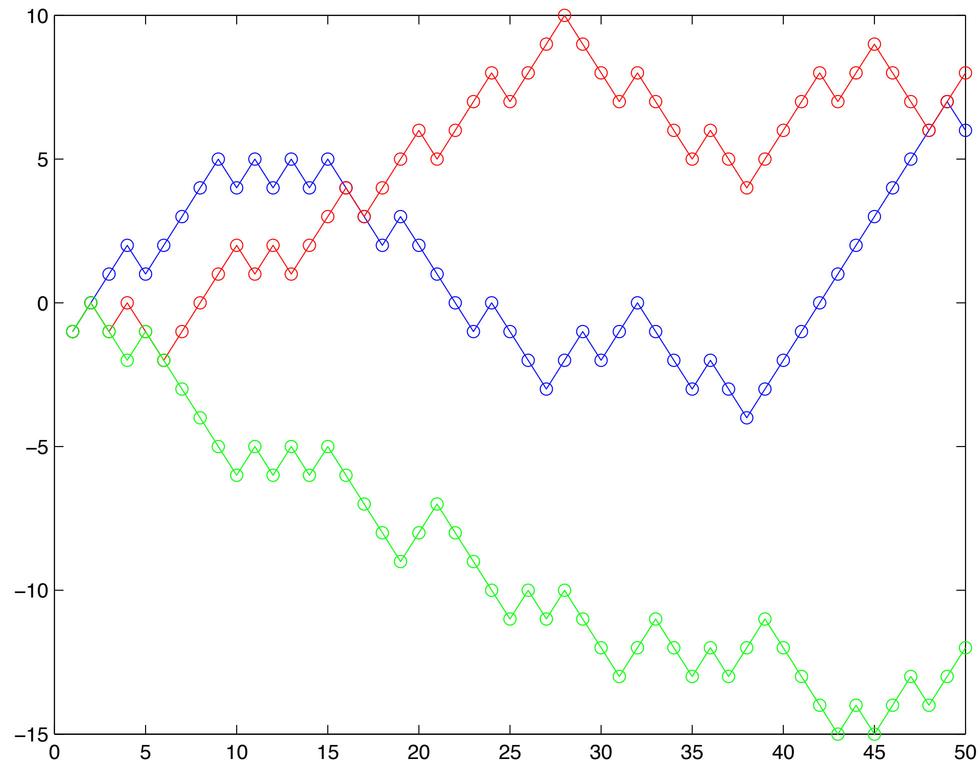
Differences: $\nabla S_t = S_t - S_{t-1} = X_t.$



Example random walk (3)

What is the mean and variance?

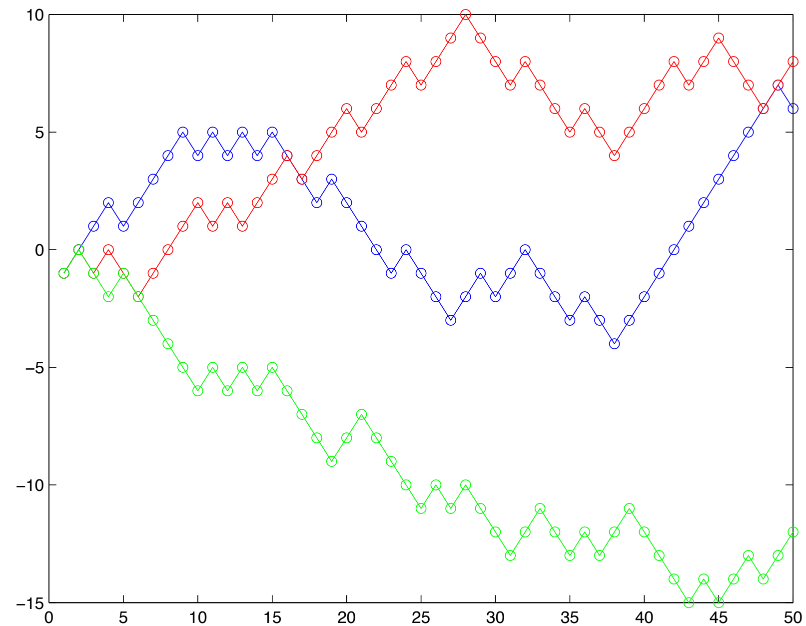
$\mathrm{E}S_t$? $\mathrm{Var}S_t$?



Example random walk (4)

Example: Random walk, $S_t = \sum_{i=1}^{t} X_i$ for i.i.d., mean zero $\{X_t\}$.
We have $\mathrm{E}[S_t] = 0, \mathrm{E}[S_t^2] = t\sigma^2$, and

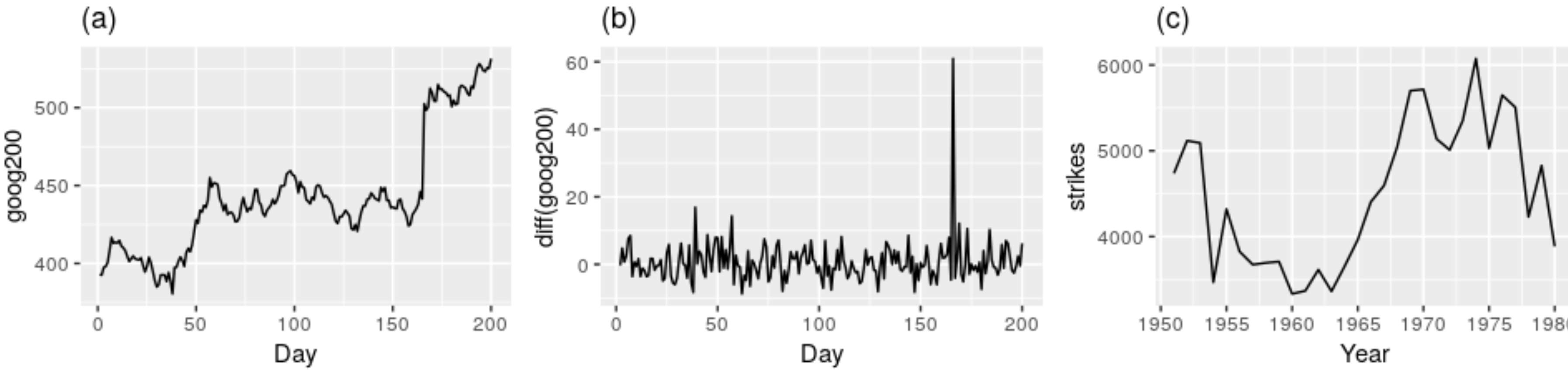
$$\begin{aligned}\gamma_S(t+h,t) &= \mathrm{Cov}(S_{t+h}, S_t)\\ &= \mathrm{Cov}\left(S_t + \sum_{s=1}^{h} X_{t+s}, S_t\right)\\ &= \mathrm{Cov}(S_t, S_t) = t\sigma^2.\end{aligned}$$



1. $\mu_t = 0$ is independent of t, but
2. $\gamma_S(t+h,t)$ is not.

So $\{S_t\}$ is not stationary.

Stationary or non-stationary?



Chasing stationarity: choosing the right method

Although trend-stationary and difference stationary series are both trending over time, the correct approach needs to be used each case

Deterministic trend process (e.g., $y_t = \alpha + \beta t + u_t$):

- Use detrending
- Differencing trend-stationary series would remove the non-stationarity but as a result it would introduce MA(1) structure into the errors

Random walk (or stochastic non-stationarity):

- Use "differencing" (e.g., instead of modeling levels, consider rate of change)
- Trying to detrend a series with a stochastic trend will not remove the non-stationarity

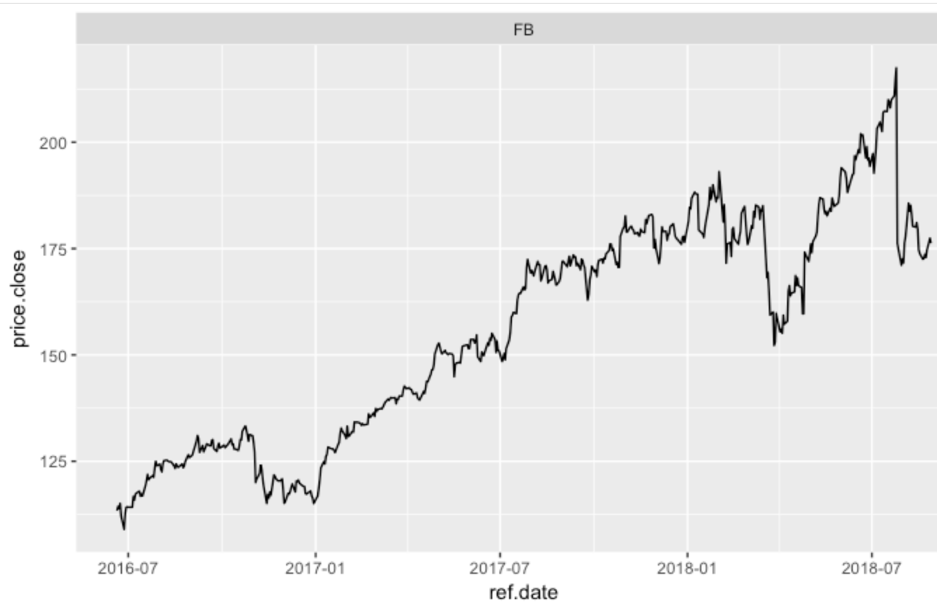
Testing for stationarity

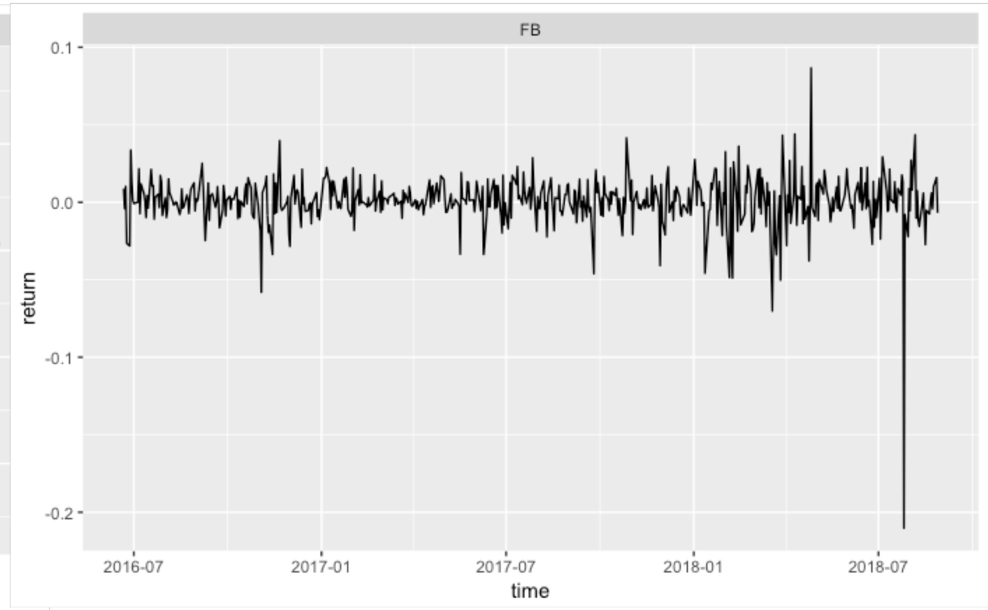
- One way to determine more objectively whether differencing is required is to use a *unit root test*.
- These are statistical hypothesis tests of stationarity that are designed for determining whether differencing is required.
  - *Dickey Fuller (DF) tests*
  - *Augmented Dickey Fuller (ADF) tests*
  - *Phillips-Perron test*
  - *KPSS test*

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, *54*(1-3), 159–178. https://doi.org/10.1016/0304-4076(92)90104-Y

Differencing time series

Facebook stockprice and returns





Forecasting with time series

Aalto University
School of Business

Notation

- y_t: time series to forecast
- n: last observation
- $n + h$: time period to forecast
- h: forecast horizon
- I_n: information available at time n to forecaset y_{n+h} (e.g., leading indicators, covariates, historical observations)

Forecast distribution

When we say that we would like to forecast y_{n+h} given I_n, we mean that y_{n+h} is uncertain

- y_{n+h} has a conditional distribution, $y_{n+h} \mid I_n \sim F(y_{n+h} | I_n)$
- The conditional distribution contains all information about the unknown y_{n+h}

A complete forecast of y_{n+h} is the conditional distribution or a its density $f(y_{n+h}|I_n)$

However, $F(y_{n+h}|I_n)$ is complicated (distribution), we typically report low dimensional summaries called forecasts

Different types of forecasts

Point forecast (the most common forecast type)

Variance forecast

Interval forecast

Density forecast

Etc.

Point forecasts

- Point forecast $f_{n+h|n}$ is “the best guess” for y_{n+h} given the distribution $F(y_{n+h}|I_n)$
- We can measure accuracy by a loss function, which is typically fore regressions the “squared error”: $l(f, y) = (y - f)^2$
- The risk is the expected loss:
- The “best” point forecast is defined to be the one with the smallest risk:

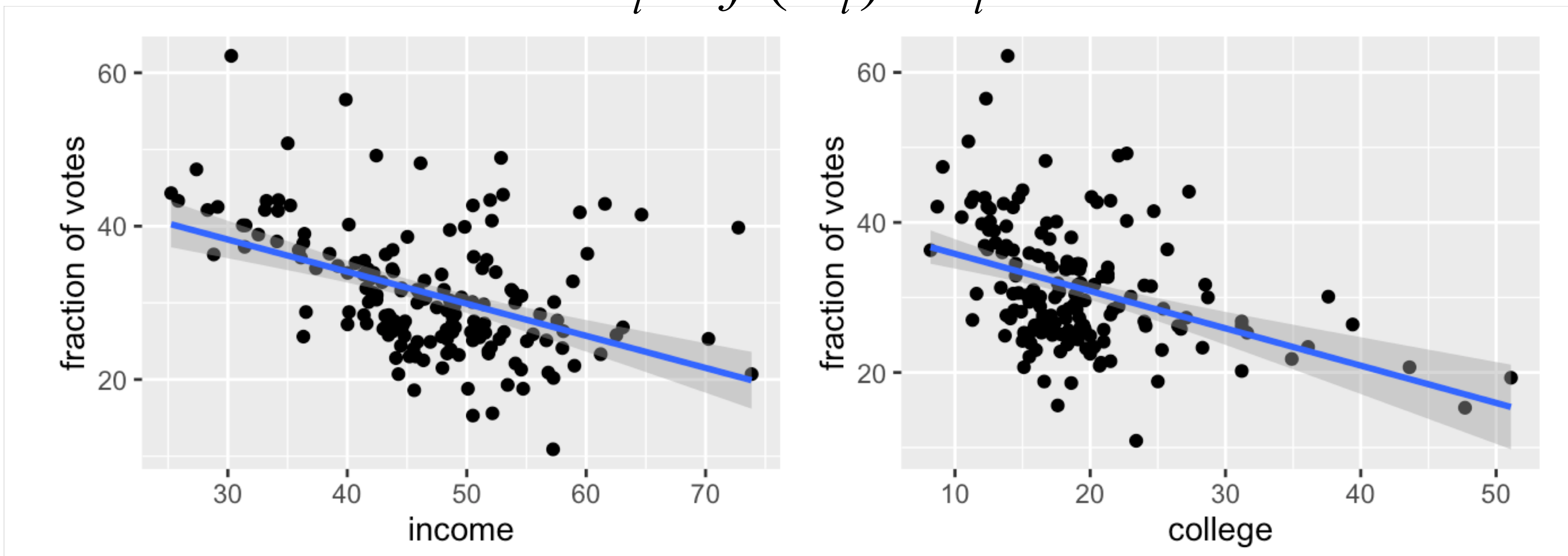
$$f = \text{argmin}_f \, E\left((y_{n+h} - f)^2 \,|I_n)\right)$$
$$= E(y_{n+h}|I_n)$$

The optimal point forecast is the true conditional expectation.

Point forecasts are estimates of the conditional expectation!

Ex. What function f(x) would predict the fraction of votes for Donald Trump?

$$Y_i = f(\mathbf{X}_i) + \varepsilon_i$$



Why estimate f?

$$\hat{Y} = \hat{f}(X)$$

Is there an ideal f?

The ideal or optimal predictor of Y with regard to mean-squared prediction error:

f(x) = E(Y |X = x) is the function that minimizes

$$E[(Y - g(X))^2 | X = x]$$

over all functions g at all points X = x

Estimation

Challenge: The conditional distribution and the ideal point forecast are unknown. They need to be estimated (approximated) from data using a suitable model.

Estimation involves typically the following issues:

- Approximation of $E(y_{n+h}|I_n)$ with a parametric family
- Selecting a model within the parametric family
- Selecting a sample period (window width)
- Estimation of parameters

Choice of information set

What features (or variables) should be included in the information set?

Past lags of the target variable?

Past observations of some other variables, “leading indicators”, covariates, dummy indicators?

$$I_n = (x_n, x_{n-1}, \dots)$$

Caution: Use of features in predicting

- It is not clear what the actual future values would be for the features (variables used as indicators)
- If the features are predictable (i.e., have some patterns that can be modeled), you can create a forecast for each of the features separately. However, remember that using these predicted values as features propagates their forecast errors to the target variable, which may further increase errors or produce biased forecasts
- A pure time series model (i.e., one that uses only past records of the target variable) may have similar or even better performance than a model that uses features

Markov approximation

- The conditional expectation depends on infinite past: $E(y_{n+1}|I_n) = E(y_{n+1}|x_n, x_{n-1}, \ldots)$

- Rather than attempting to grasp the inifinte past, we can typically replace it with Markov (finite memory) approximation

- For some p: $E(y_{n+1}|x_n, x_{n-1}, \ldots) \approx E(y_{n+1}|x_n, \ldots, x_{n-p})$

How to estimate f?

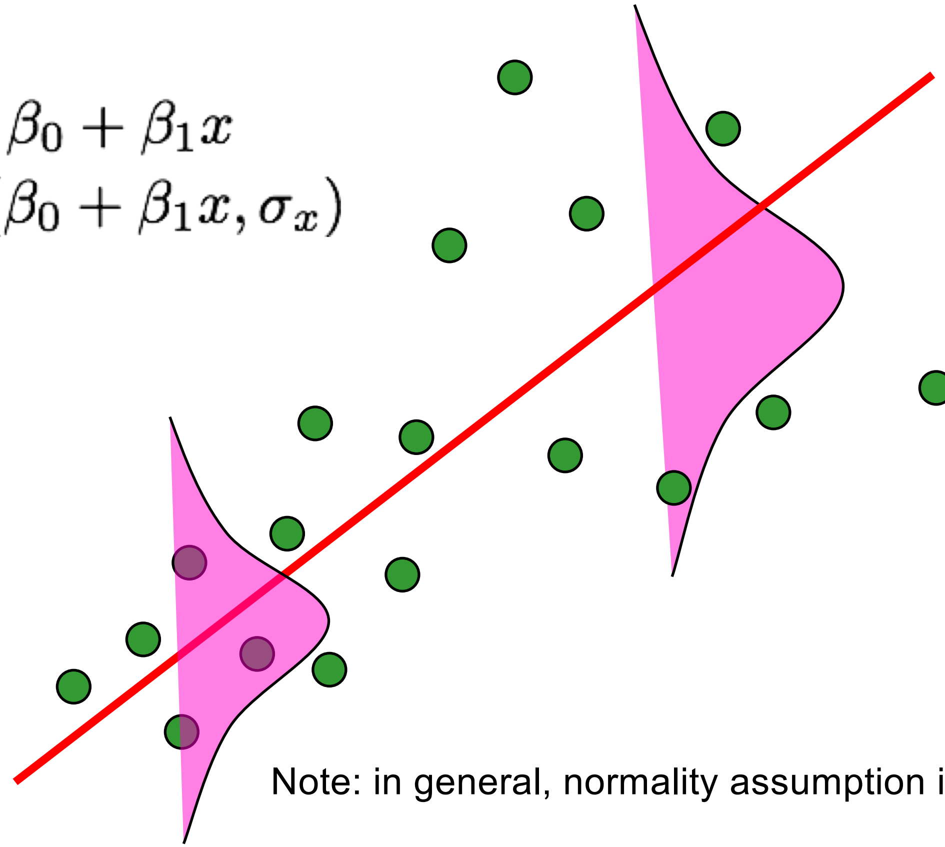
- We will assume we have observed a set of training data:

$$\{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_n, Y_n)\}$$

- We must then use the training data and a statistical method to estimate f.
- Perhaps the most common approach is to use linear regression
- Although linear models are almost never correct, they serve as a good and interpretable approximation to the unknown true function f(X)

Simple regression

$$E[y] = \beta_0 + \beta_1 x$$
$$y \sim N(\beta_0 + \beta_1 x, \sigma_x)$$



Note: in general, normality assumption is not required

Linear approximation and forecasting model

- The true $E(y_{n+1}|x_n, \dots, x_{n-p})$ is unknown and possibly non-linear
- However, in practice, linear approximations give a solid baseline model which often performs surprisingly well even in the presence of unknown non-linearities:

$$E\left(y_{n+1}\middle|x_n, \dots, x_{n-p}\right) \approx \beta_0 + \beta_1' x_n + \cdots + \beta_p' x_{n-p} = \beta' x_n$$

- The model error is defined as the difference between the actual observation y_{n+1} and the linear forecast

$$e_{n+1} = y_{n+1} - \beta' x_n$$

- As a result, we obtain the following linear point forecasting model:

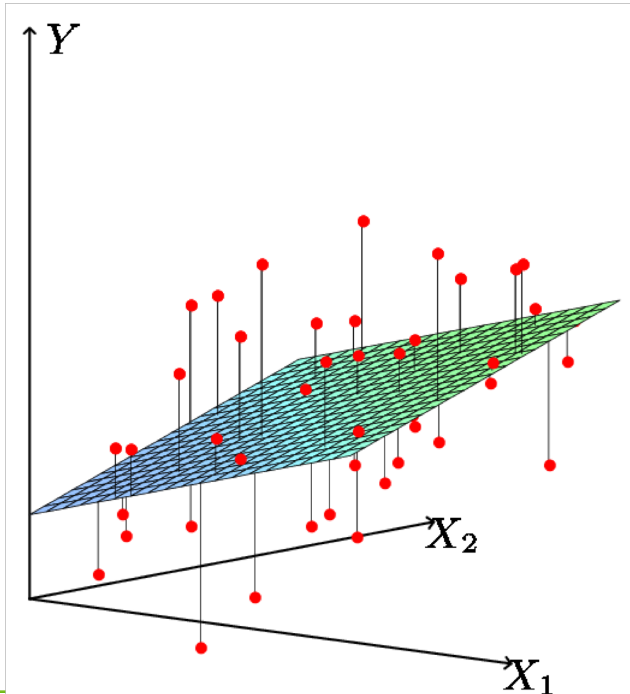
$$y_{n+1} = \beta' x_n + e_{n+1}$$

In Matrix Form: $y = X\beta + \varepsilon$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{np} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Least squares fit



minimize the residual sum of squares

$$\begin{aligned}\mathrm{RSS}(\beta) &= \sum_{i=1}^{N}(y_i - f(x_i))^2 \\ &= \sum_{i=1}^{N}\Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2.\end{aligned}$$

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j.$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Example: univariate autoregression

- Suppose $x_t = (y_t, y_{t-1}, \dots, y_{t-k+1})$
- A linear forecasting model is given by

$$y_{t+1} = \beta_0 + \beta_1 y_t + \beta_2 y_{t-1} + \cdots + \beta_k y_{t-k+1} + e_{t+1}$$

- This model is known as kth order autoregression AR(k)

$$\widehat{\beta} = \left(\sum_{t=0}^{n-1} \mathbf{x}_t \mathbf{x}_t' \right)^{-1} \left(\sum_{t=0}^{n-1} \mathbf{x}_t y_{t+1} \right)$$

$$\widehat{y}_{n+1|n} = \widehat{f}_{n+1|n} = \widehat{\beta}' \mathbf{x}_n$$

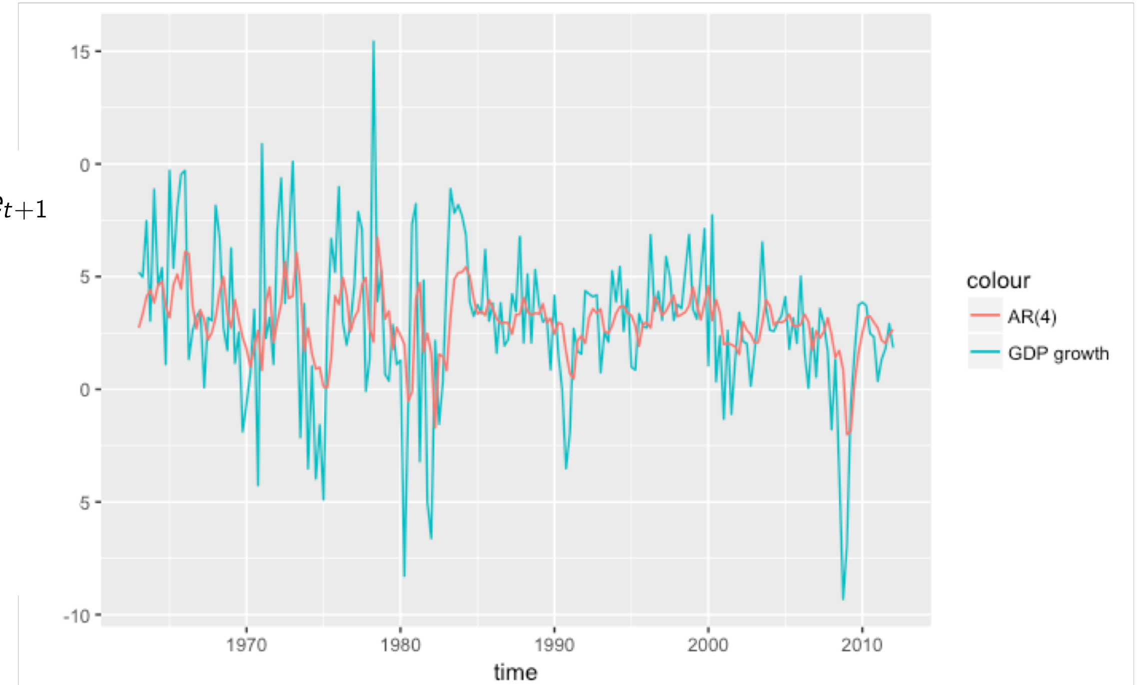
GDP example

$y_t = \Delta \log(GDP_t)$, quarterly

AR(4) (reasonable benchmark for quarterly data)

$$y_{t+1} = \beta_0 + \beta_1 y_t + \beta_2 y_{t-1} + \beta_3 y_{t-2} + \beta_4 y_{t-3} + e_{t+1}$$

	$\widehat{\beta}$	$s(\widehat{\beta})$
Intercept	1.54	(0.45)
$\Delta \log(GDP_t)$	0.29	(0.09)
$\Delta \log(GDP_{t-1})$	0.18	(0.10)
$\Delta \log(GDP_{t-2})$	-0.05	(0.08)
$\Delta \log(GDP_{t-3})$	0.06	(0.10)



GDP example (2)

One step-ahead forecast using AR(4) model

Uses the information from the last 4 quarters 2011:2 – 2012:1 to predict the unknown observation 2012:2 which is not included in the original dataset

Do we trust the obtained forecast?

	Actual	Forecast
2011:1	0.36	
2011:2	1.33	
2011:3	1.80	
2011:4	2.91	
2012:1	1.84	
2012:2		2.59

Forecast selection

- The choice of AR(4) model was arbitrary!
- Should we have considered an autoregression with different number of lags?
- Forecasts can be quite sensitive to these choices
- The goal is to produce accurate forecasts that minimize the empirical risk (low MSFE)
- Finding the true model is not relevant as this maybe a model with infinite number of parameters

Model	Forecast
AR(0)	2.99
AR(1)	2.59
AR(2)	2.65
AR(3)	2.68
AR(4)	2.59
AR(5)	2.83
AR(6)	2.83
AR(7)	2.83
AR(8)	2.78
AR(9)	2.87

⇒ **Model selection problem!**

Long-term prediction

Long-term prediction means predicting further into the future

Choices to implement or use the regression model in prediction:

- Recursive Prediction Strategy
- Direct Prediction Strategy
- And variants

Recursive Prediction Strategy

Predictions are made one step ahead at the time

$$\hat{x}_{t+1} = f(x_t, x_{t-1}, x_{t-2}, \ldots x_{t-d+1})$$
$$\hat{x}_{t+2} = f(\hat{x}_{t+1}, x_t, x_{t-1}, x_{t-2}, \ldots x_{t-d})$$

Benefits: Only one prediction model f to estimate

Disadvantages: Accumulation of errors in each step

Direct prediction strategy

Predictions are made k steps ahead at once:

$$\hat{x}_{t+k} = f_k(x_t, x_{t-1}, x_{t-2}, \ldots x_{t-d+1})$$

Benefits: The problem of k steps ahead prediction is solved directly

Disadvantages: Must train a model f_k for each k

Long-term prediction

- What is long-term prediction depends on the context!
- Interesting phenomena vary from milliseconds to centuries
- Prediction further into the future is more difficult
- Direct Prediction Strategy is preferred if long-term prediction is the main interest

Review questions

- What makes a time series analysis different from classification problems?
- What role does stationarity have in time series analysis?
- What is the definition of point forecast in this context?
- Describe one example of an autoregressive time series model

Testing vs. model selection

Our next topic will be to discuss the forecast selection problem further!

Historically, it has been common to use statistical tests to select empirical models, but more recent discussions suggest that use of statistical tests may not be a good idea when choosing forecasts

- Tests answer scientific questions (e.g., hypothesis regarding model parameters such as is some coefficient of interest zero)
- Tests are not designed to answer the question: Which estimate yields better forecast
- Standard errors are appropriate for measuring estimation precision but not the goodness of forecasts

For model selection, we want something different than the classical tests!

Model selection

What is a model? What is a good model?

- Why are we building models? Models are useful, because they help us to *answers questions* about the reality
- Models are abstractions of reality: some details are forgotten (or must be forgotten), example: describe your day!
- The modeler is always faced with a trade-off with fidelity to data and the level of abstraction
- Generalization was earlier defined as the ability to model (or predict) future, unseen data!

Model selection criteria

- If two models have the same error, which one is better?
- One approach is to use the simpler model, that is ”fewer parts”
- Model selection crtieria are used to give a number for model complexity, for instance ”number of parts in the model”
- Using more data in training results in better models usually
- Parts mentioned are usually model parameters, or coefficients
- Introduce model selection criteria that penalize for model complexity

Criteria for model selection

$$C_p = \frac{1}{n}\left(\text{RSS} + 2d\hat{\sigma}^2\right)$$

d = number of predictors

$$\text{AIC} = \frac{1}{n\hat{\sigma}^2}\left(\text{RSS} + 2d\hat{\sigma}^2\right)$$

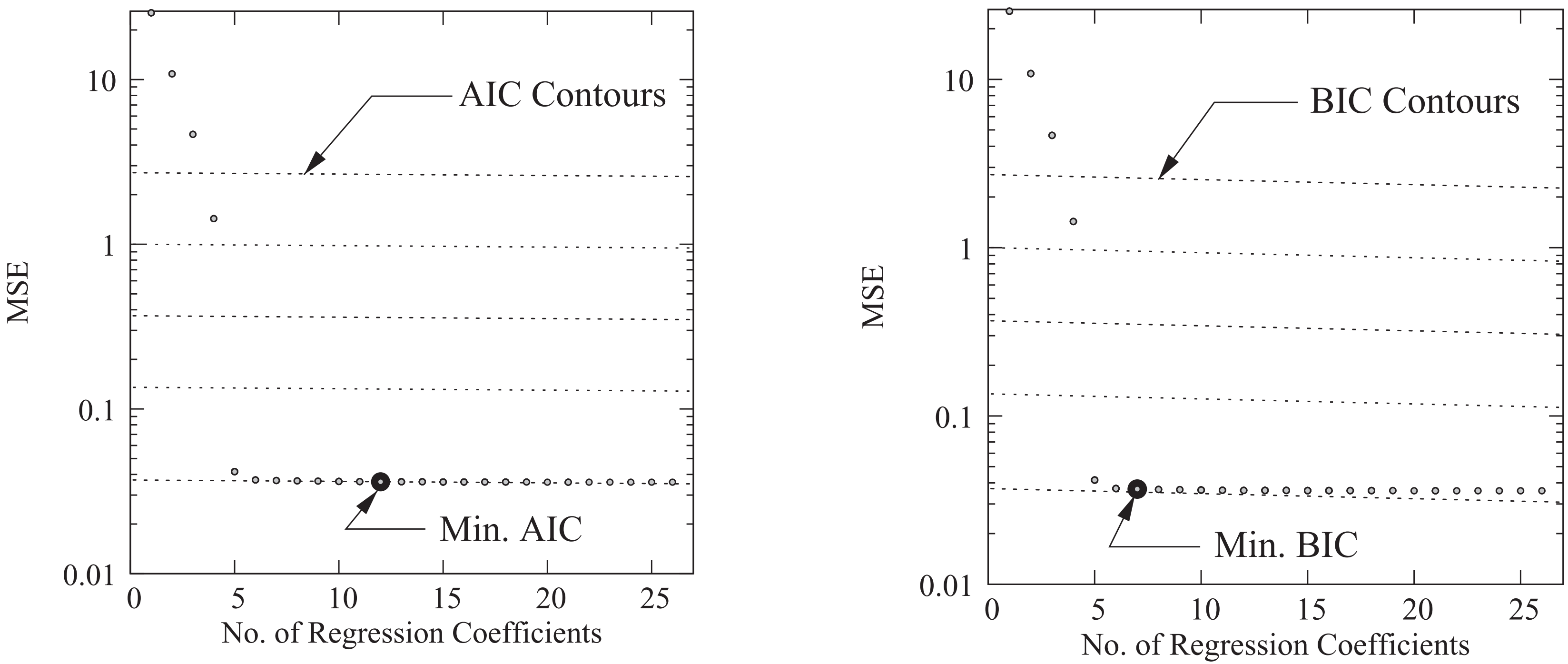
$$\text{BIC} = \frac{1}{n}\left(\text{RSS} + \log(n)d\hat{\sigma}^2\right)$$

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$

AIC and BIC

- Both AIC and BIC are relative measures for selecting models
- AIC leads to the following model selection ”rules of thumb”:
  i. If two models have the same error, select the one with less parameters (simpler)
  ii. If two models have the same number of parameters, select the one with smaller error
- **BIC leads to the following model selection ”rules of** thumb”:
  i. If the models have the same number of parameters and error, use the one learned from more data

AIC vs. BIC minimization



Source: Sinha et al. (2015): A multiobjective exploratory procedure for regression model selection

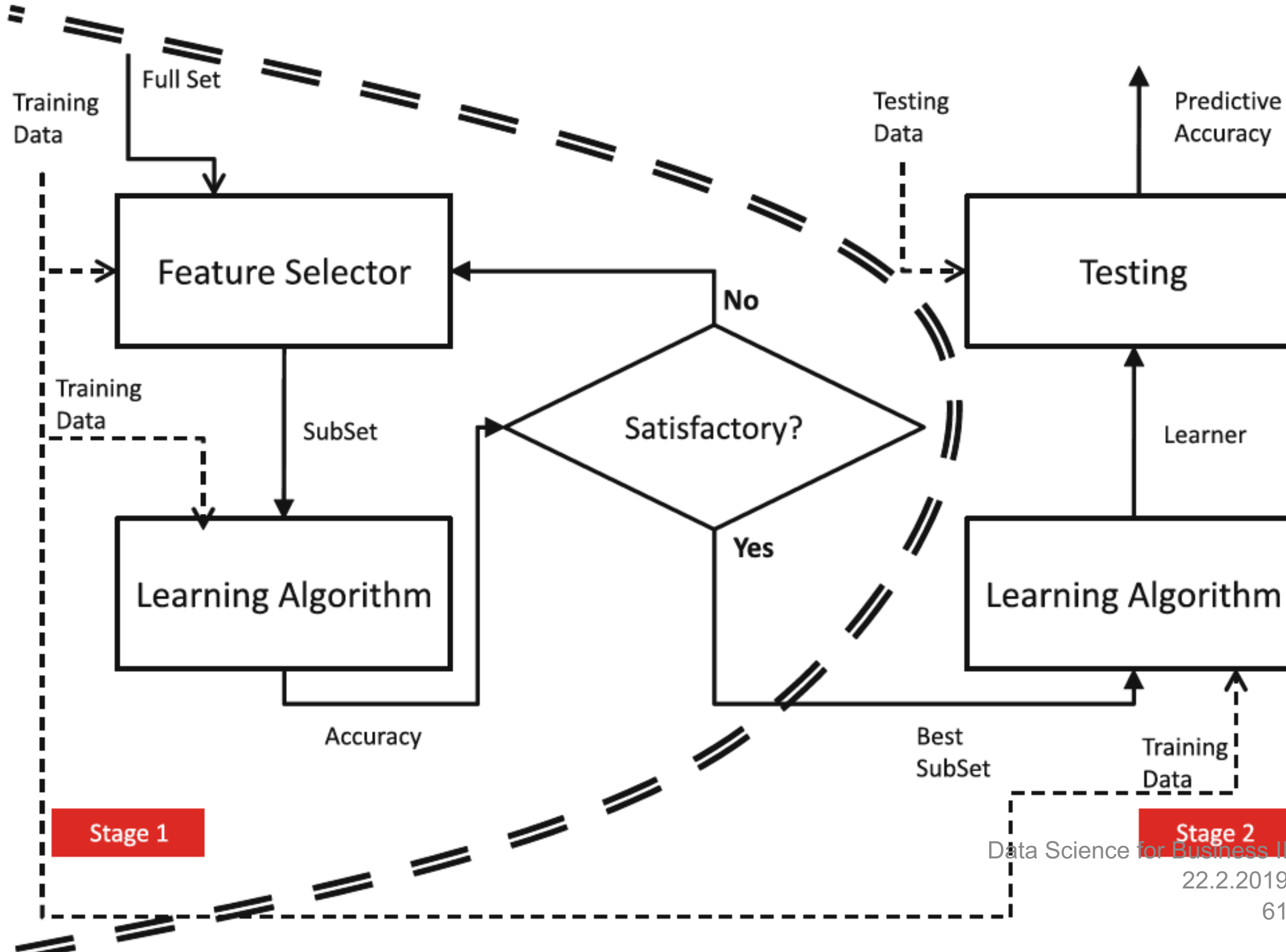
Recap: predictive accuracy and model selection

- Goodness of models can be judged from many perspectives
- Model can have good predictive accuracy, in a validation set. This does not guarantee good generalization into the future
- Relative merit of two models can be estimated using model selection criteria, like Akaike’s Information criterion (AIC), or Bayesian Information Criterion (BIC). In addition to predictive error, the model complexity and effective sample size determine what is good
- Business considerations can also be used a criterion to select the model!

Feature selection
(review from DSFB-1)

Aalto University
School of Business

Feature selection with learning in the loop
"Wrapper perspective"



Suggested solutions

- **Subset selection**
  - E.g., best subset selection, stepwise selection methods
  - Identifying a subset of all p predictors X that we believe to be related to the response Y, and then fitting the model using this subset
- **Shrinkage**
  - Involves shrinking the estimates coefficients towards zero
  - This shrinkage reduces the variance
- **Dimension reduction**
  - E.g. Principle Components and Factor Analysis
  - Involves projecting all p predictors into an M-dimensional space where $M < p$, and then fitting linear regression model

Aalto University
School of Business

Stepwise selection routines

Forward search

Forward-stepwise selection is a greedy algorithm, producing a nested sequence of models.

Forward Search

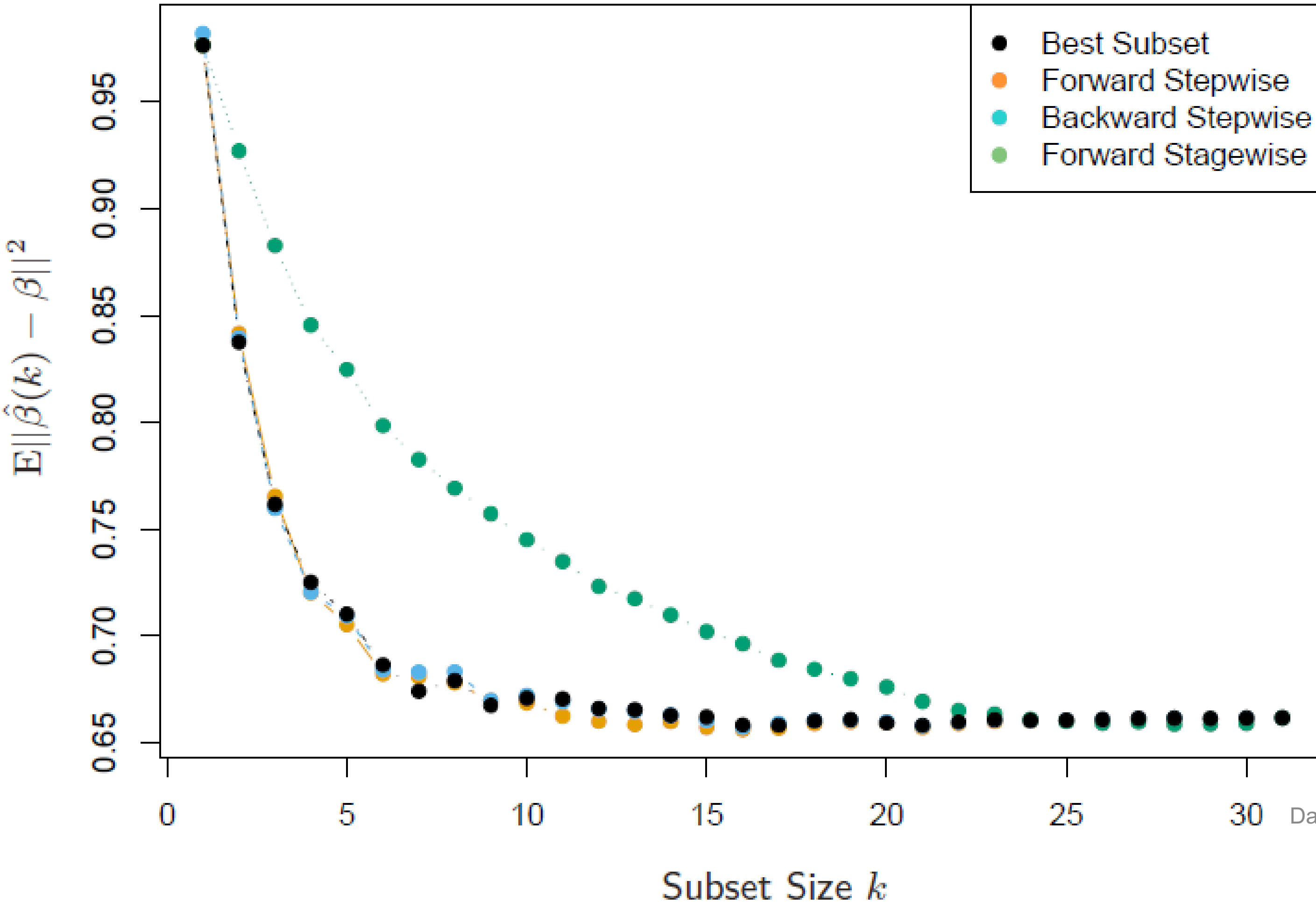
- Let $\mathcal{F} = \{\}$
- While not selected desired number of features
- For each unused feature f:
  - Estimate model's error on feature set $\mathcal{F} \bigcup f$ (using cross-validation)
- Add f with lowest error to $\mathcal{F}$

Backward search

Backward Search

- Let $\mathcal{F} = \{\text{all features}\}$
- While not reduced to desired number of features
- For each feature $f \in \mathcal{F}$:
    - Estimate model's error on feature set $\mathcal{F} \backslash f$ (using cross-validation)
- Remove f with lowest error from $\mathcal{F}$

Stepwise solutions as approximations



Comments

- Forwards / backward stepwise methods can be used when the number of variables p is too large for best subsets method
- Forward / backward methods are heuristics and are not guaranteed to find the best model containing a subset of predictors
- Backward selection requires that $n >> p \rightarrow$ full model can be fitted
- Forward selection can also be used when $n < p$. In fact, it is the only viable subset method for large p

Aalto University
School of Business

Regularization techniques

(aka ”shrinkage methods”)

Why shrinkage might be considered?

- OLS is good when the relationship between Y and X is linear and the number of observations n is way bigger than the number of predictors p i.e., $n > p$
- But, when p is almost as large as n, then the least squares fit can have high variance and may result in overfitting and poor estimates on unseen observations,
- And, when $n < p$, then the variability of the least squares fit increases dramatically, and the variance of these estimates in infinite (unique estimate doesn't exist!)

Ridge regression

Ridge estimates are found by minimizing:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \text{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2,$$

- As with least squares, ridge regression seeks coefficient estimates that fit the data well, by making the RSS small.
- The second term is a penalty that shrinks the coefficients towards zero
- Though not immediately obvious, shrinking can help to reduce variance

Problem with Ridge regression

- **Unlike subset selection, which will generally select models that involve just a subset of the variables, ridge regression will include all p predictors in the final model!**

The Lasso Estimator

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

- Similar to ridge regression, but the key difference in behavior follows from penalty

Why Lasso is good?

- **Ability to force some coefficients exactly to zero → performs variable selection**
- A model is called “sparse” when it involves only a subset of variables
- Can even be used when p > n, a situation where OLS fails completely!
- Computationally efficient: for any given “lambda”, we only need to fit one model and the computations turn out to be very simple

Lasso vs. Ridge

LASSO:

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

Ridge:

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s,$$

Lasso vs. Ridge

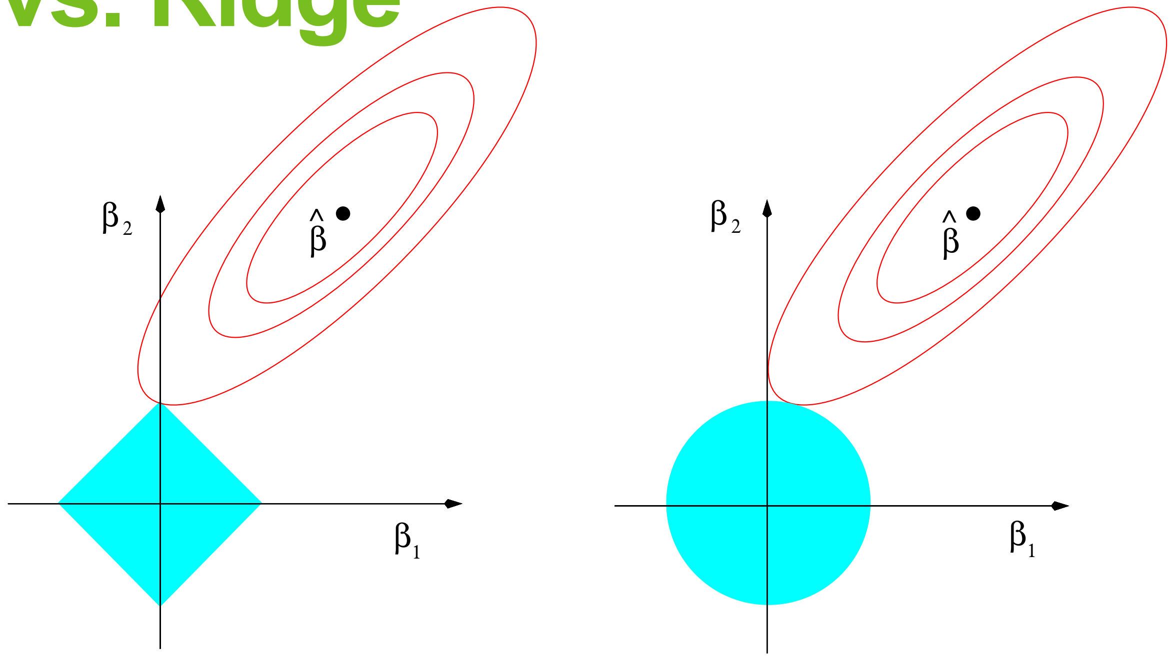
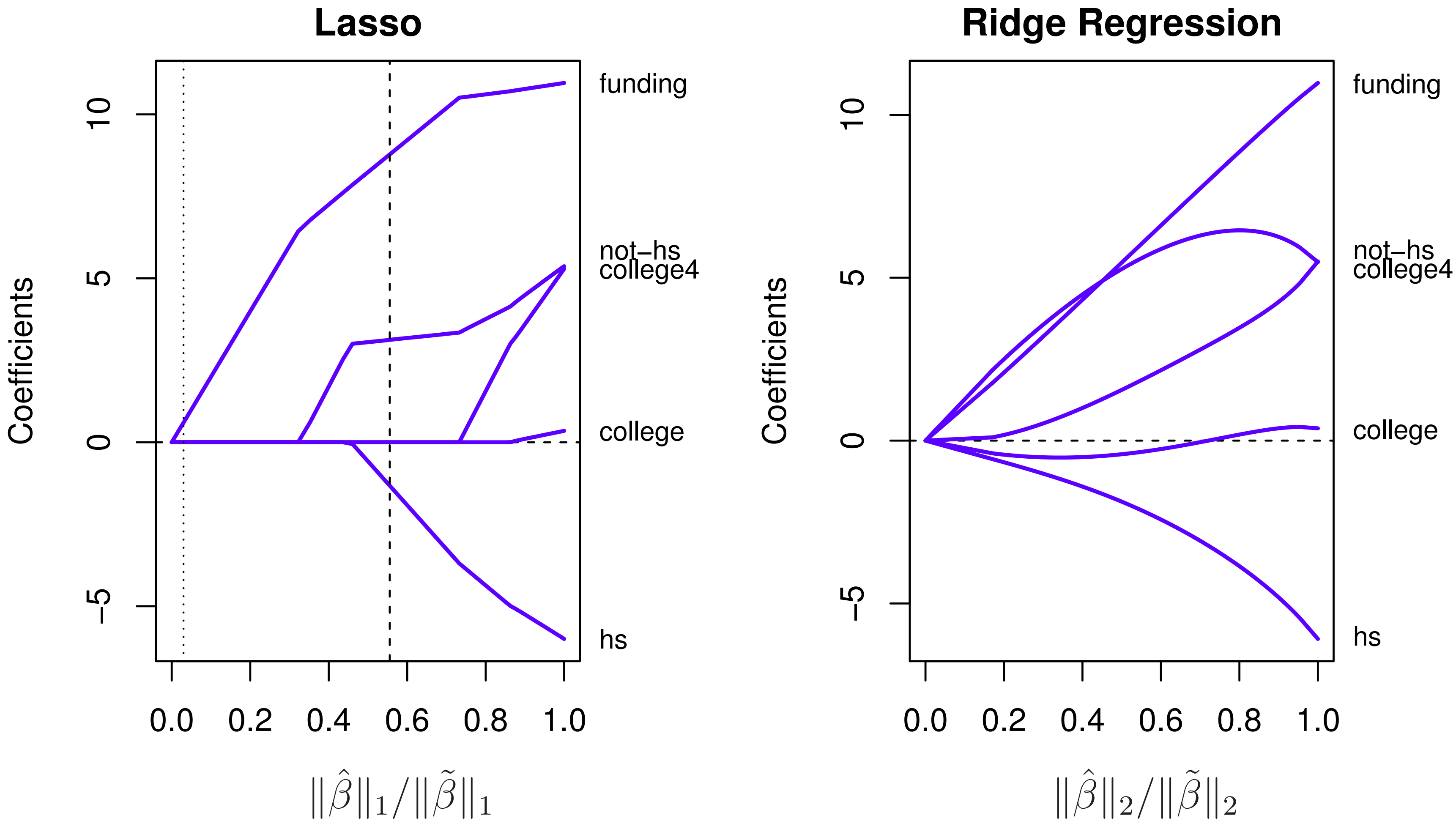


Figure 2.2 *Estimation picture for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions* $|\beta_1|+|\beta_2| \leq t$ *and* $\beta_1^2+\beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the residual-sum-of-squares function. The point* $\widehat{\beta}$ *depicts the usual (unconstrained) least-squares estimate.*

Example

Table 2.1 *Crime data: Crime rate and five predictors, for $N = 50$ U.S. cities.*

city	funding	hs	not-hs	college	college4	crime rate
1	40	74	11	31	20	478
2	32	72	11	43	18	494
3	57	70	18	16	16	643
4	31	71	11	25	19	341
5	67	72	9	29	24	773
⋮	⋮	⋮	⋮	⋮		
50	66	67	26	18	16	940


Lasso
Coefficients
10
5
0
−5
0.0 0.2 0.4 0.6 0.8 1.0
$\|\hat{\beta}\|_1/\|\tilde{\beta}\|_1$
funding
not−hs
college4
college
hs
Ridge Regression
Coefficients
10
5
0
−5
0.0 0.2 0.4 0.6 0.8 1.0
$\|\hat{\beta}\|_2/\|\tilde{\beta}\|_2$
funding
not−hs
college4
college
hs

Variable scales

- The standard OLS estimates are scale equivariant: multiplying a variable by constant "c" just leads to scaling of estimated coefficients by factor of "1/c" [i.e. scaling doesn't matter]

- In ridge regression, coefficient estimates can change substantially when multiplying a given predictor by a constant, due to the sum of squared coefficients term in the penalty part of the ridge regression objective function

Choice regularization coefficient

- **Consider use of information criteria**

$$\mathrm{AIC} = \frac{1}{n\hat{\sigma}^2}\left(\mathrm{RSS} + 2d\hat{\sigma}^2\right)$$

$$\mathrm{BIC} = \frac{1}{n}\left(\mathrm{RSS} + \log(n)d\hat{\sigma}^2\right)$$

- **Choose the model with best performance in training/validation set**