# ST3233: Applied Time Series Analysis

## AY2014/2015, Semester I

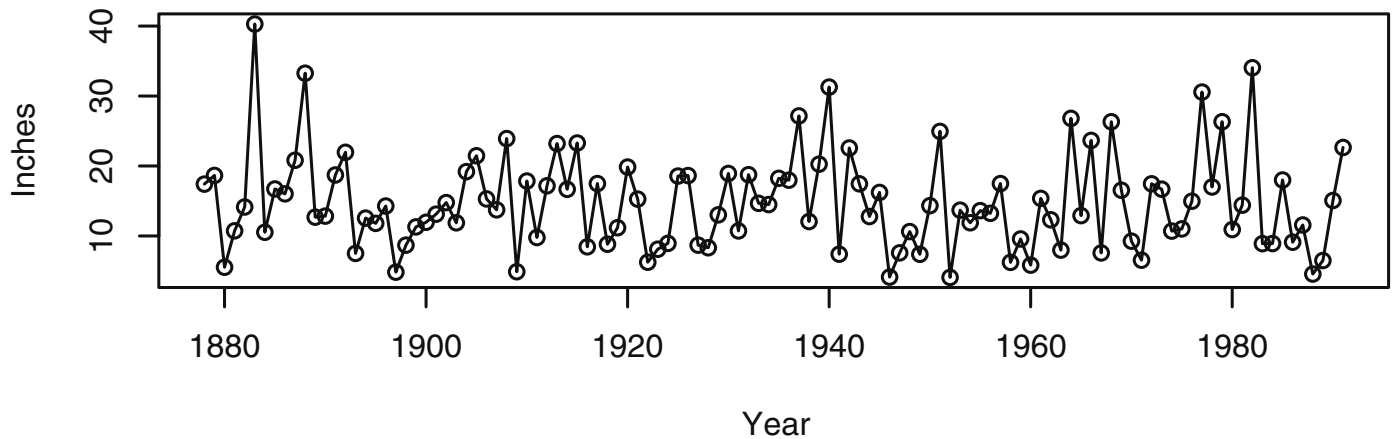
Dr. Alkema
Contact: alkema@nus.edu.sg

**Time series** are perhaps one of the most familiar forms of data from our everyday lives. They take the form of **observations that are linked over or indexed by time**. Examples include: (i) the daily movements of a stock index; (iv) temperature fluctuations at a weather recording station over a time course of days to years; (iii) annual counts of an endangered species; (iv) the monthly sales figures of a merchandise. This linking of observations in time means that observations cannot be considered to be an independent random sample as in many other applications of statistics. Instead this dependency structure must be accounted for. One way to do this is via a family of **models** called ARIMA—auto-regressive (i.e. regressing upon itself) integrated moving averages. This course will focus on ARIMA models, and you will learn what exactly an ARIMA model is, how to estimate their parameters, and how to perform **predictions** of the future evolution of the time series using the fitted ARIMA model.

# You will learn how to use models to (attempt to) answer questions such as ...

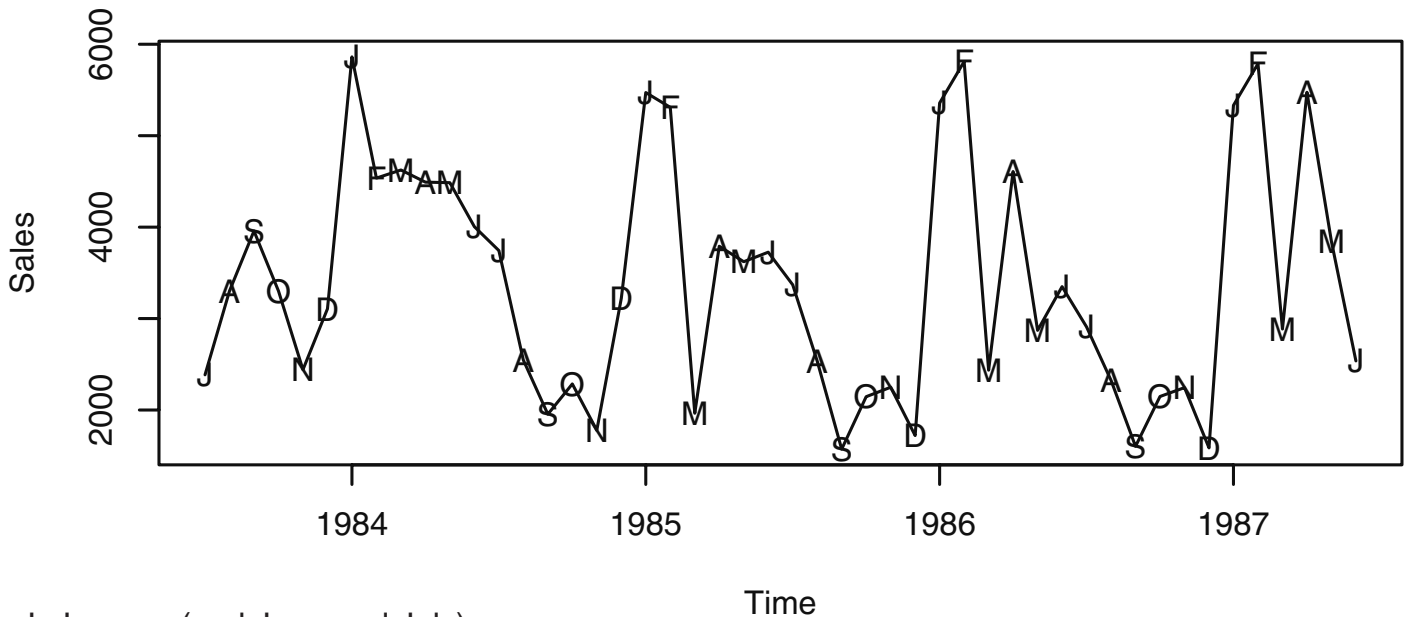
- How much rain do we expect to fall in Los Angeles (Singapore) next year?

**Exhibit 1.1 Time Series Plot of Los Angeles Annual Rainfall**



# You will learn how to use models to (attempt to) answer questions such as ...

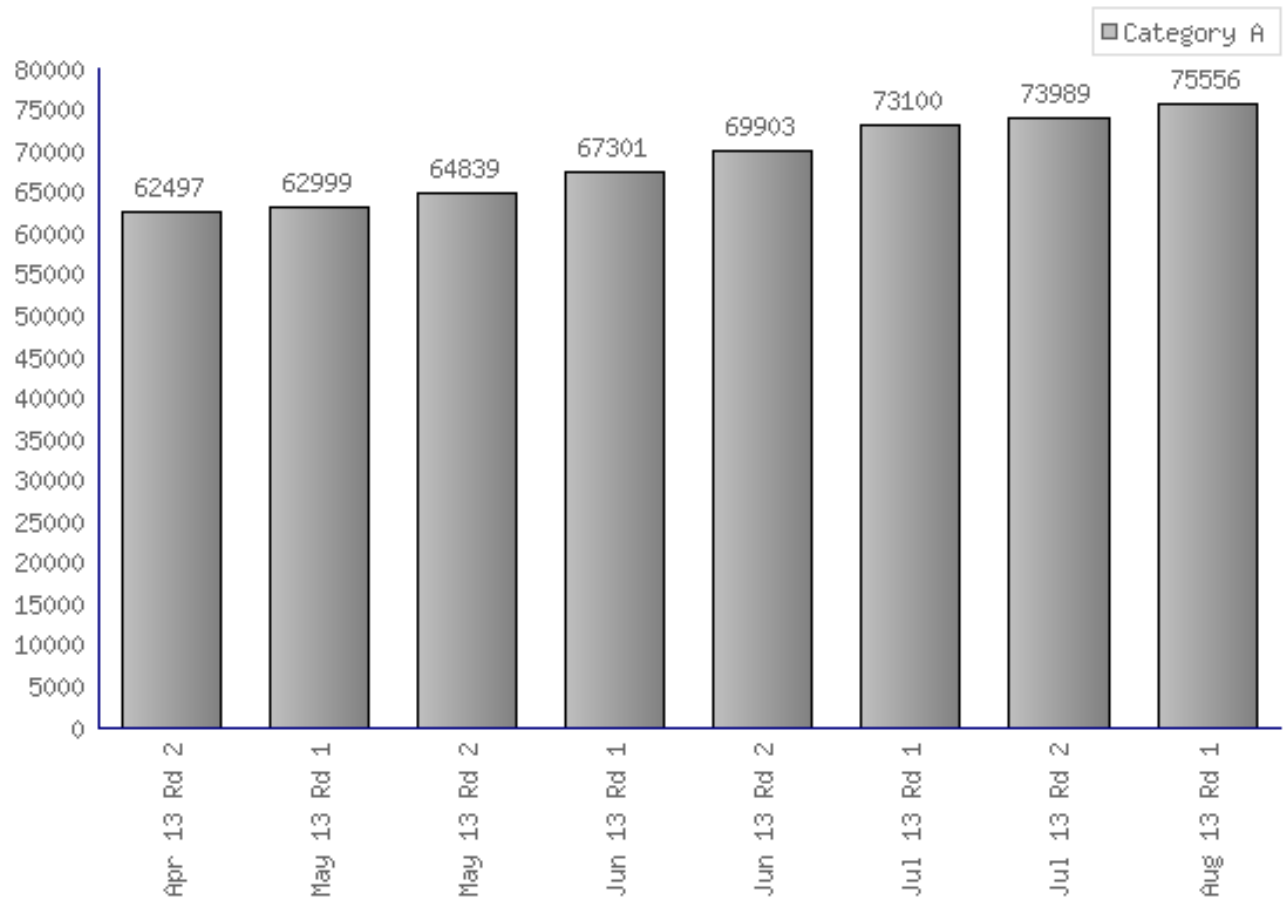
- How do sales (of oil filters) change with time?




J=January (and June and July),
F=February, M=March (and May), and so forth

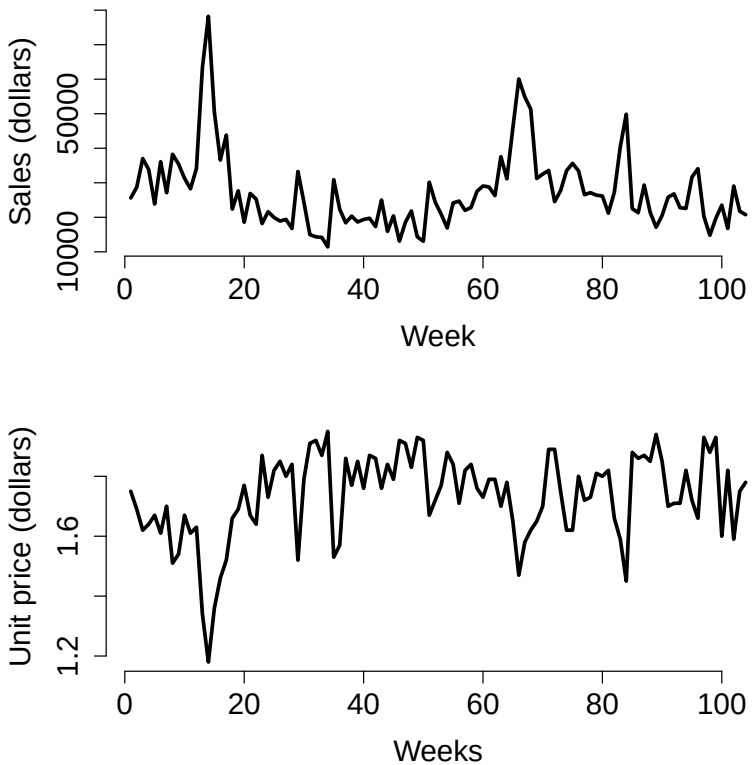
# You will learn how to use models to (attempt to) answer questions such as ...

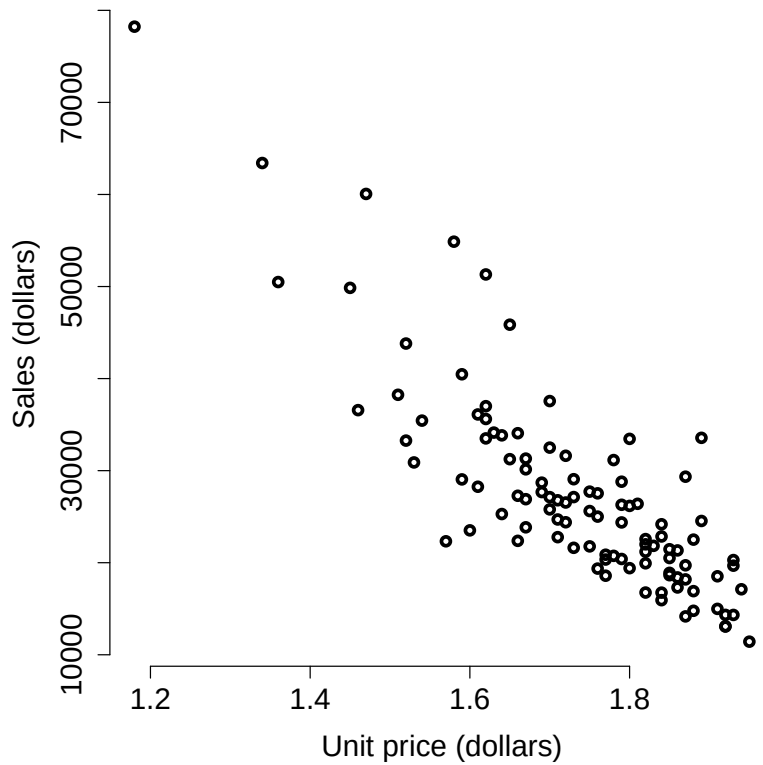
- Shall I buy a car soon or wait for the COE to decrease?



# A more detailed example: Bluebird potato chips (ch.11)

- On day 1 in your new job, your boss gives you a dataset with information on total sales and the unit price of potato chips.
- Your boss' question: Predict total sales for the next four weeks if:
  - the company keeps the unit sales price constant at 1.78 dollar,
  - the company increases the price to 2 dollars.
- Help, what to do?????



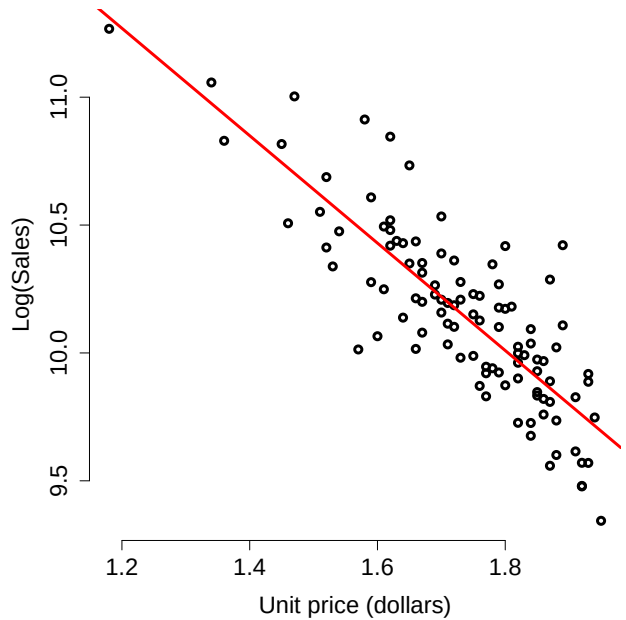



# Analyzing the sales dataset

- You may be familiar with regression analysis and wonder if you could use the following linear regression model (fit illustrated below):

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t,$$

where $Y_t$ denotes the log(sales), $X_t$ the price, $\beta_0$ and $\beta_1$ are unknown regression coefficients and $\varepsilon_t$ are random error terms, which are assumed to have mean zero, constant variance and be uncorrelated.




- Is it ok to use the regression model for this dataset?
- Let's use the residuals to check whether it is reasonable to assume that the error terms are uncorrelated over time.

# Review: Correlation

- The correlation between two random variables $Z$ and $W$, e.g. sales $Z$ and sales price $W$, is defined as follows:

$$Corr(Z, W) = \frac{Cov(Z, W)}{\sqrt{Var(Z)Var(W)}} = \frac{E(Z - E(Z))(W - E(W))}{\sqrt{Var(Z)Var(W)}}.$$

  It measures the direction and strength of the linear relation between $Z$ and $W$.

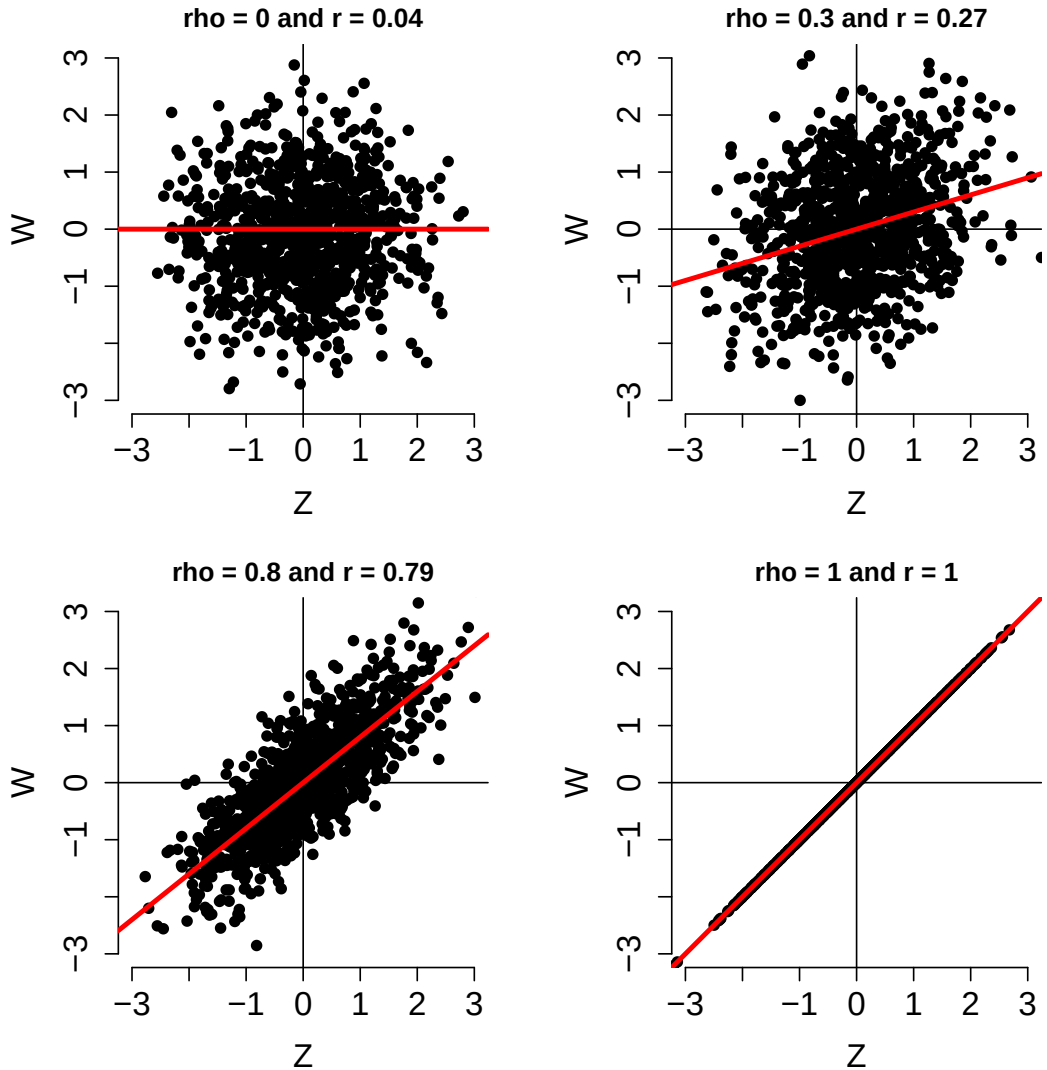
- The sample correlation between random variables $Z$ and $W$ based on sampled pairs $(z_1, w_1), .., (z_n, w_n)$, is given by:

$$r(Z, W) = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(w_i - \bar{w})}{\sqrt{\sum_{i=1}^{n}(z_i - \bar{z})^2 \sum_{i-1}^{n}(w_i - \bar{w})^2}}.$$

  The sample correlation is an estimate for the true correlation between $Z$ and $W$. For a large sample size, a sample correlation that is close to 1 (or -1) suggests that $Z$ and $W$ are positively (or negatively) correlated.

# Illustration of correlation

Simulation of $(z_1, w_1), .., (z_{1000}, w_{1000})$ from $(Z, W) \sim N_2((0,0), \Sigma)$, with $\sigma_Z = \sigma_Z = 1$ and $Corr(Z, W) = \rho$.



# Time series analysis: Autocorrelation

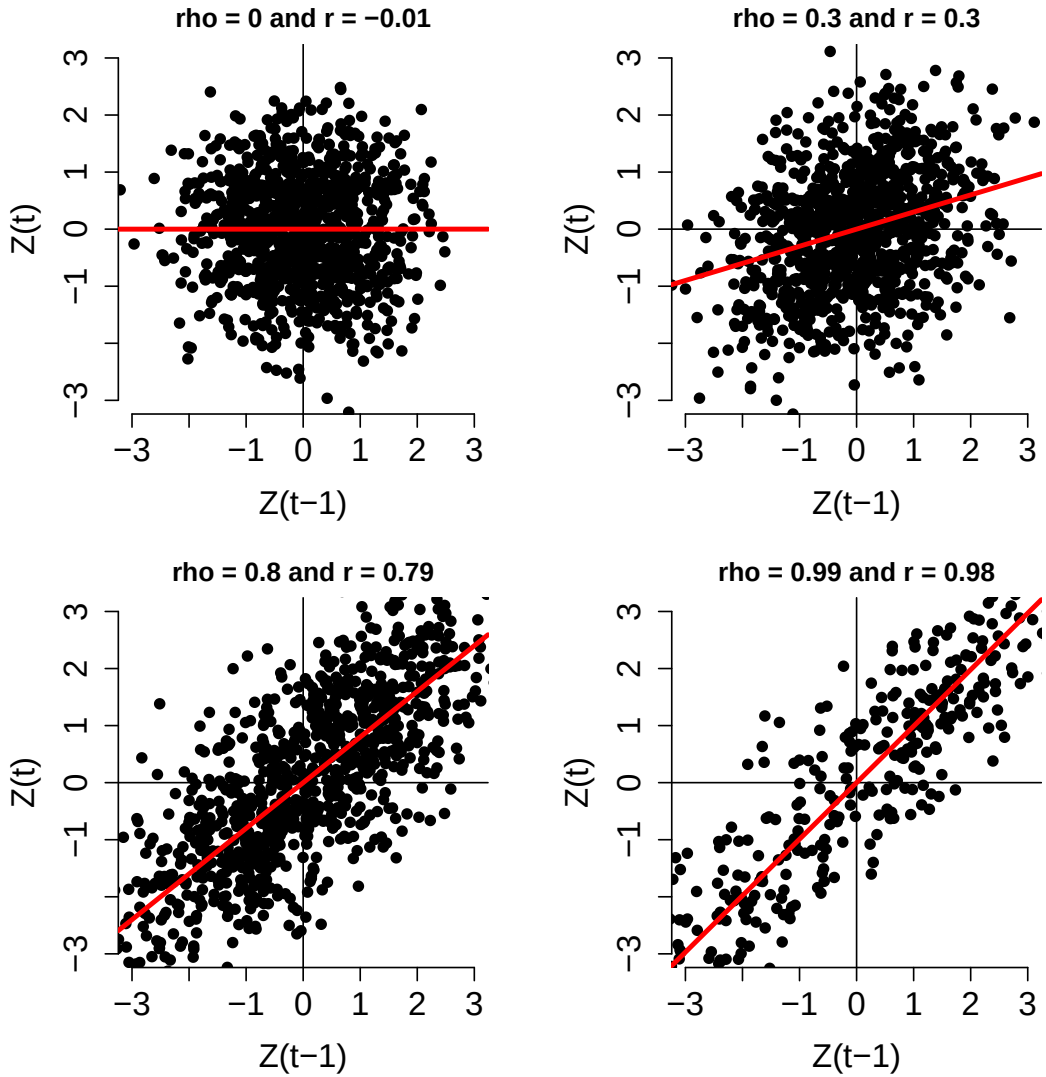
- For one time series of random variables $Z_1, Z_2, \ldots, ..$ (e.g., sales), the correlation between $Z_t$ and $Z_{t-k}$ for some time $t$ and time lag $k$ is referred to as autocorrelation.
- For an observed time series $z_1, z_2, \ldots z_n$, the sample autocorrelation at lag $k$ (e.g. $k = 1$) is given by:

$$r_k(Z) = \frac{\sum_{t=k+1}^{n}(z_t - \bar{z})(z_{t-k} - \bar{z})}{\sqrt{\sum_{t=1}^{n}(z_t - \bar{z})^2}}.$$

If the autocorrelation between $Z_t$ and $Z_{t-k}$ is the same for all times $t = k+1, k+2, \ldots$, then $r_k(Z)$ measures the autocorrelation between random variables $Z_t$ and $Z_{t-k}$ for any $t$.

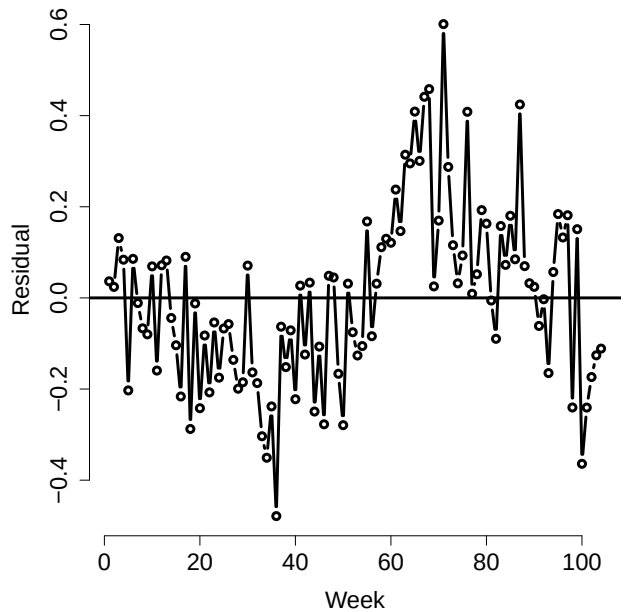
# Illustration of autocorrelation

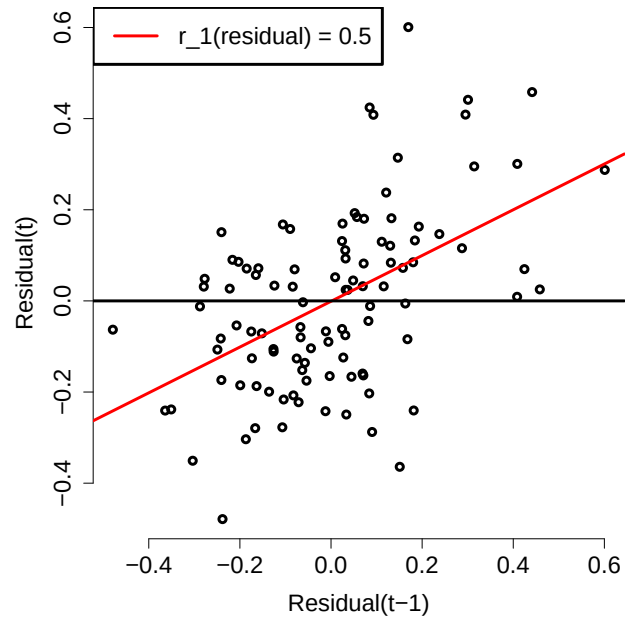
Simulation of $(z_1, z_2, \ldots, z_{1000})$ with $E(Z_t) = 0$, $Var(Z_t) = 1$ and $cor(Z_t, Z_{t-1}) = \rho$ for all $t$ (details of this simulation to be discussed later in the course).



# Model assumptions for linear regression

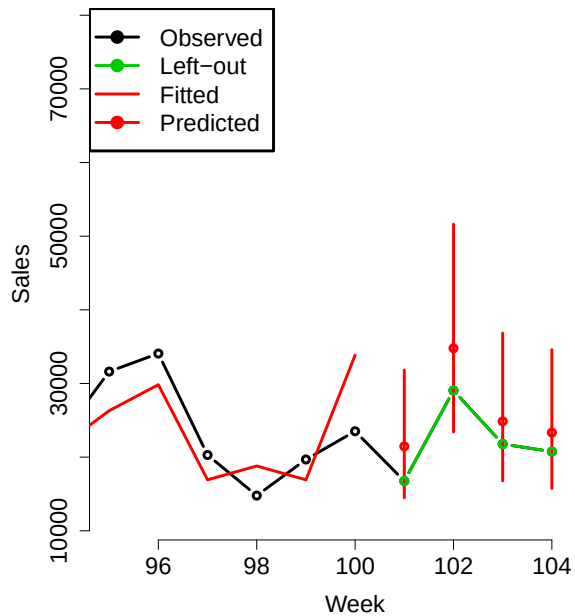
- In the linear regression model $Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$, we assume that the error terms $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are uncorrelated.
- For the sales data, the residuals turn out to be autocorrelated. This suggests that the assumption of uncorrelated error terms may be violated.
- Does autocorrelation of error terms matter for your analysis? In particular, does it matter for obtaining predictions of future sales?
- Let's illustrate what may happen with a validation exercise.



# Model validation exercise

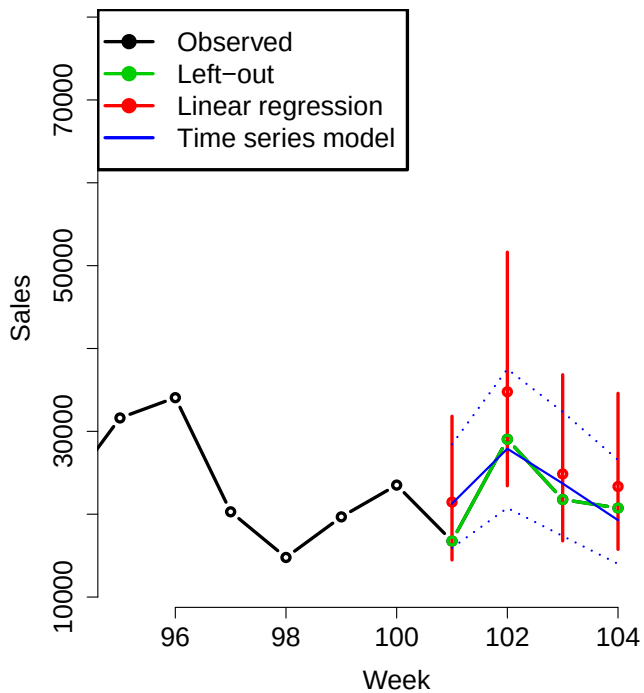
- Suppose that we would have used the linear regression model to predict sales for the most recent (observed) four weeks; how well would be model have predicted the outcomes?
- Let's check through a validation exercise:
  - Fit the regression model to a training data set, which is the full data but excluding the last 4 observations.
  - Use the model to predict the last 4 observations.




- Result: the regression model overpredicts the last 4 observations. (Is that what you expected?)
- Is there a more appropriate model to construct predictions?

# Time series models

- Time series models are used for analyzing (describing) time series data and making predictions.
- Example: the predictions below for the sales data were obtained a time series model.




- The time series model gives predictions which are closer to the left-out values (as compared to the regression model) because it accounts for autocorrelation in the time series (the part of the time series that is not explained by the covariate).

# A bit more detail

- The predictions on the previous slide were obtained from the same linear regression model as discussed so far

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t,$$

but instead of assuming that the errors are independent (uncorrelated over time), which we realized may not be accurate, the errors were modeled as follows:

  - $\varepsilon_t = \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \phi_3 \varepsilon_{t-3} + e_t$ where $e_t$ are independent random variables with mean zero.

This model for $\varepsilon_t$ is referred to as an autoregressive model of order 3 (you are regressing the outcome $\varepsilon_t$ on its past values). This model outperforms the linear regression model because it takes into account the autocorrelation in the error terms.

- No worries if this is not very clear yet, we will introduce and discuss these types of models at much greater length later in the course.

# Applied Time Series Analysis

- You will learn how to use models to (attempt to) answer questions such as ...
  - How much rain do we expect to fall in Los Angeles (Singapore) next year?
  - How do sales (of oil filters) change with time?
  - Shall I buy a car soon or wait for the COE to decrease?
  - How to predict sales of potato chips, given information about the per-unit sales price?
  - ...
- More generally, our goals are to:
  1. understand/model (a simplified version of) the stochastic mechanism that give rise to an observed time series,
  2. predict/forecast future values.
- We will cover the 3 main steps to find an appropriate model:
  - Model identification/specification
    - We will discuss various models for time series, in particular, the autoregressive integrated moving average (ARIMA) models.
  - Model fitting
  - Model diagnostics

# Module organization

- See IVLE!
- Lecture times: Mon and Th, 8.15 – 9.30 (with one 7 min break).
- Class material (e-books):
  - (Main) Cryer and Chan (2008). Time series analysis with applications in R (2nd ed). Springer (ISBN: 978-0-387-75958-6; e-ISBN: 978-0-387-75959-3).
  - (Reference material) Brockwell and Davis (2002). Introduction to Time Series and Forecasting (2nd ed). Springer (ISBN: 978-0-387-95351-9).

# Module overview

## Course overview, subject to change

| Wk (Monday date) | | Topic | Material | Tut | Remark |
|---|---|---|---|---|---|
| 1 | 11-Aug | Intro & Fundamental Concepts & R | Ch 1 & 2, Ch 16.0 - 16.2 | | |
| 2 | 18-Aug | Models for stationary time series | Ch 4 | | |
| 3 | 25-Aug | Models for nonstationary time series | Ch 5, Ch 3 | 1 | |
| 4 | 1-Sep | Model specification | Ch 6 | 2 | |
| 5 | 8-Sep | Parameter estimation | Ch 7 | 3 | |
| 6 | 15-Sep | Model diagnostics | Ch 8 | 4 | |
| | 20-Sep | Recess week | | | |
| 7 | 29-Sep | Forecasting | Ch 9 | 5 | |
| 8 | 6-Oct | Forecasting | Ch 9 | | NO Mon-lec or tutorials (Hari Raya Haji holiday) |
| 9 | 13-Oct | Seasonal models | Ch 10 | 6 | |
| 10 | 20-Oct | Time series regression models | Ch 11 | 7 | |
| 11 | 27-Oct | Time series models of heteroskedasticity | Ch 12 | 8 | |
| 12 | 3-Nov | TBD | | 9 | |
| 13 | 10-Nov | Review | | 10 | |

- Midterm will be scheduled in week 7 or 8. Tentative date is Th Oct 2.

# Notes on module content and organization

- Slides will be uploaded before start of every lecture:
  - The goal of the slides is to aid the presentation of the material.
  - The slides are concise: they are only a summary of the text and not intended to give all details.
  - Please read the book!
- *Applied* does not imply there will not be any theory!
  - We won't be able to carry out a full analysis/construct predictions until end of chapter 9!
- Datasets will be related to various areas.
  If you have a data set that you are particularly interested in, send it to me and I'll check if it can be included as an example in lecture/tutorials.

# Computing

- We'll use the statistical software R
  - as a tool to fit models, check model fits, obtain predictions,
  - to get insight into class material through simulations, visualization.

  Let R work for you!
- Most R-code will be provided to you:
  - R-code is available with the book, and explained in Ch. 16.
  - Supplementary R-code will be uploaded on IVLE (e.g., check out ch0_intro2R.R if you are new to R).
- I hope you will improve your R programming skills in this module (e.g. useful for FYP or future job).
- R-code will be checked during (a subset of) tutorials, and may be tested on exams.
- I highly recommend using R-studio to work with R!

# Grading

- Break-down:
  - Project 1 or midterm: 25% (week 7 or 8)
  - Project 2: 20% (2nd half of semester)
  - Final: 50%
  - Tutorial presentations and checks: 5%
- Tutorials (see pdf on IVLE):
  - Students will present solutions
  - Tutors check R-code
  - Goal: learning!

# Important

- Ask questions before/after and DURING class, via email (alkema@nus.edu.sg), or consultation.
- If you have questions, please ask! If you are struggling, please let me know as soon as possible, so that we can resolve the situation. If you have suggestions to make the class better, please let me know as well.
- Learn concepts, understand what's going on, don't just memorize formulas.