Lecture 8 Logistic Regression and Time-Series Forecasting

Yingda Zhai (Dr.)

School of Computing, NUS

October 4, 2022

BT1101 Roadmap: Predictive (7-10), Prescriptive (11-12)

- Week 1 - 6 – Descriptive Analytics
- Week 7 – Linear Regression
- Week 8 – Logistic Reg & Time Series
- Week 9 – Data Mining Basics
- Week 10 – *Assessment*
- Week 11 – Linear Optimization
- Week 12 – Integer Optimization & Summary
- Week 13 – Tutorials and Consultation
- Exam Wk – *Final Exam*

1 Logistic Regression

- Regress with Binary Dependent Variable
- Maximum Likelihood Estimator (MLE)
- Interpretation for Logistic Regression

2 Concepts in Time Series

- Cross-Section and Time-Series Data
- Trend and Seasonality of Time Series
- Stationary and Weakly Dependent Process

3 OLS Regression in Time Series

- Static and Finite Distributed Lag Model (FDL)
- Assumptions of OLS Regression in Time Series Analysis
- Dealing with Trend and Seasonality

4 Exponential Smoothing Models and Auto-Regressive

- Prediction with Univariate Time-Series
- Smoothing Models: Moving Averages
- Smoothing Models: Exponential Smoothing Models
- Train-Test Split

Learning Objectives

- Be ready to handle binary outcome variable with logistic regression and interpret the its coefficients.
- Understand basic concepts that are important to time-series analysis such as difference between time series and cross-sectional data, stationarity, trend, seasonality, etc.
- Be able to understand and use moving average, exponential smoothing, Holt-Winter methods, and autoregressive model (AR) for univariate time-series.

Logistic Regression

Logistic Regression: A Binary Dependent Variable

- We have consider the case where independent variables in linear regression are continuous and categorical.

  What if the dependent variable is categorical or a binary dummy, e.g. yes/no decision or success/fail?

Customer	Previous Spending	Marital Status	#Ads Displayed	*Purchased*
Andy	\$476	Married	3	Yes
Charlie	\$169	Single	2	No
Ashley	\$23	Married	6	No

- Can we build a regression model to predict a consumer's online purchase decision based on the data we collect in the e-commerce platform?

$$\texttt{purchased} \sim \texttt{spending} + \texttt{marital} + \texttt{ads} + \cdots \tag{1}$$

RM: Observe that $\texttt{purchased} \in \{0, 1\}$ while right hand side of (1) ranges typically in real values $\mathbb{R}$.

Logistic Regression: A Binary Dependent Variable

- A *general linear model* (GLM) is a more generalized linear model where a link function is used to map the dependent variable to a linear combination of independent variables.
- In particular, a *logistic regression* or logit regression uses a logit link function to map the probability of a successful event, e.g. $p \equiv \Pr(\texttt{purchased} = 1)$, into a linear combination of predictors X's.

$$\operatorname{logit}(p) = \log \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \tag{2}$$

RM:
1. A logit function is defined as $\operatorname{logit}(x) \equiv \log x/(1-x)$ and logit function maps any number between $(0, 1)$ to $\mathbb{R}$.
2. $p/(1-p)$ is called the "odds" of such successful event, e.g. `purchased = 1` and $\log p/(1-p)$ is thus called the "log-odds".
3. Logistic regression (2) predicts the log-odds of an event occurrence $Y = 1$ rather than predicting a binary variable Y directly.

How do we get MLE estimators?

NUS National University of Singapore

Running Logistic Regression in R

- Data file `titanic.csv` contains passenger's information and if they survived the sinking of the Titanic in April 15, 1912.
- In R, call general linear model `glm()` function with specified parameter `family = binomial` for logistic regression.

```
# read 'titanic.csv' file into data frame object 'titanic'.
  titanic = read.csv('titanic.csv', header = TRUE)
# use 'glm()' with specified parameter 'family = binomial' for
    logistic regression.
  fit_surv = glm(survived ~ sex + age + sibsp + parch + fare +
      embarked, family = binomial, data = titanic, control = list
      (maxit = 50))
# display the output of logistic regression
  summary(fit_surv)
```

See `titanic` data manual for details of these variables.

Interpreting Coefficients in Logistic Regression

$\text{logit}(p) \equiv \log \frac{p}{1-p} = b_0 + b_1\,\texttt{sex} + b_2\,\texttt{age} + \cdots$, where $p = \Pr(\texttt{survived=1})$.

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.991142   0.335272   5.939 2.87e-09 ***
sexmale     -2.635345   0.190231 -13.853  < 2e-16 ***
age         -0.020467   0.007201  -2.842 0.004482 **
...
```

- b_0 Log-odds when all X's are zero. Baseline odds of survival is $\exp(1.991) = 7.32$.
- b_1 Being a male decreases the log-odds of survival by $|b_1|$, holding all other constant. Or, being a male multiplies the odds by $\exp(-2.635) = 0.072$, i.e. the odds of survival decreases by 92.8%!
- b_2 Being each year older decreases the log-odds of survival by $|b_2|$, holding all other constant. Or, it multiplies the odds by $\exp(-0.0205) = 0.9797$, i.e. the odds of survival decreases by 2.03%.

RM: In general, b_k is the marginal effect of X_k on log-odds of event $Y = 1$ (e.g. survival). Or, $\exp(b_k) - 1$ is the marginal change of X_k on odds of survival, not probability of survival!

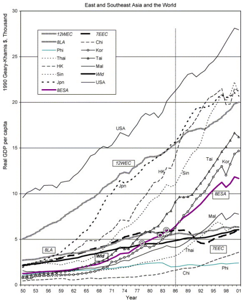
Few Things about Logistic Regression

- Note that the fitted logistic regression is a classifier. It can be used for classification in terms of $\Pr(y_\nu = 1)$ given new data points of $\boldsymbol{x}_\nu$, e.g. loyal customer, dog in the picture, survived in virus outbreak?

  In R, call `predict(model, newdata, type = 'response')` for prediction of **probability** after logit model fit.
- The estimators b's in logistic regression is *maximum-likelihood estimators* (MLE) rather than OLS estimators.
- Unlike multivariate linear regression, the only key assumption of logistic regression model is *independent sample*, i.e. observations (x_i, y_i) are independent from each other for all $i = 1, 2, \ldots, n$.
- z-distribution (standard normal) instead of t-distribution is used in statistical inference in logistic regression; yet hypothesis testing remains similar.

Applications of Time-Series Analysis

Economic Metric

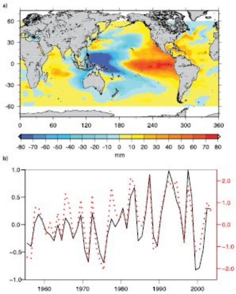


GDPs in South East Asian Countries.

Financial Market

S&P500 Market Index

Environment



Sea Level Change

Concepts in Time Series

Cross Sectional Data, Time Series and Panel Data

- Cross-Sectional data: cross section of information (variables) of numerous subjects or entities at a *certain* time stop. e.g. `mroz`.
- Time-Series data: a series of information (variables) of one subject or entity across *multiple* time stops in a temporal ordering. e.g. `hseinv`.
- Panel data: contains multiple entities' information at multiple stops. e.g. `jtrain`.

RM:
1. Cross-sectional: fixing at a time spot, data "snapshot" of multiple entities.
2. Time-series: fixing one entity, series of data "snapshot" across time.
3. Panel: a series of "snapshots" of multiple entities across time.
4. Panel data analysis is beyond the scope of the course but it is actually similar to cross-sectional data rather than time series, in term of analysis.

Cross Sectional Data, Time Series and Panel Data

entity i	time t	$X_{i,t}$	X_i
1	1	$X_{1,1}$	X_1
2	1	$X_{2,1}$	X_2
3	1	$X_{3,1}$	X_3

(a) Cross-Sectional Data

entity i	time t	$X_{i,t}$	X_t
1	1	$X_{1,1}$	X_1
1	2	$X_{1,2}$	X_2
1	3	$X_{1,3}$	X_3

(b) Time-Series Data

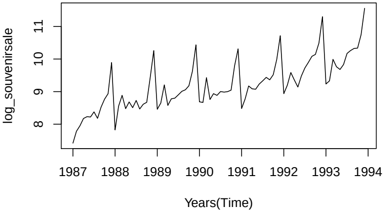
entity i	time t	$X_{i,t}$
1	1	$X_{1,1}$
1	2	$X_{1,2}$
1	3	$X_{1,3}$
2	1	$X_{2,1}$
2	2	$X_{2,2}$
2	3	$X_{2,3}$
3	1	$X_{3,1}$
3	2	$X_{3,2}$
3	3	$X_{3,3}$

(c) Panel Data

Table: Typical Wide-Form Data Tables for One Generic Variable X

Trend and Seasonality

- Trend: a tendency of upward or downward movement of time series in the long-run, e.g. newborn's weight gain, GDP.
- Seasonality: patterns repeats at certain lengths of intervals, e.g. precipitation, diurnal temperature, box-office sales.
- Cyclicality: long-term pattern that shows fluctuation with no fixed intervals.



(a) Souvenir Sale: trend and seasonality

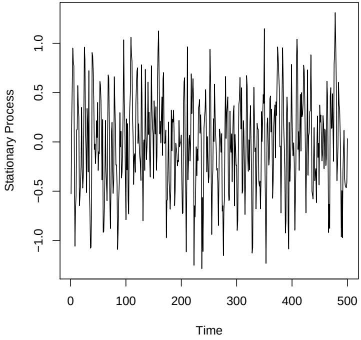
(b) Historical S&P500: cyclicality

Stationary Time Series

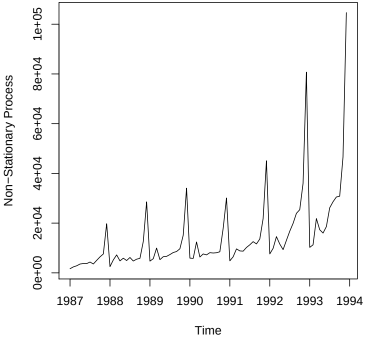
- Stationarity: a time series whose probabilistic behavior is stable over time. The probability distribution governing X_t and X_{t+h} is the same, regardless of h, e.g., $\mathbb{E}(X_t) = \mathbb{E}(X_{t+h})$ for all h.
- Time-series data is much harder to deal with, compared with (supposedly independent) cross-sectional data due to the auto-correlation among the sample points, e.g., $\text{Corr}(X_t, X_{t-1}) > 0$.
- Time series analysis and prediction is much about extracting its structure of autocorrelations.

RM:
1. As an example, a time series with any trend or seasonality is nonstationary since its mean $\mathbb{E}(X_t)$ is changing, at least.
2. It turns out stationarity is the key to any time series analysis.
3. If the time series is nonstationary, some forms of correction need to be done to make it stationary, such as "de-trending" first.

Stationary Time Series



(a) A "white noise" is a stationary series with no auto-correlation.



(b) A non-stationary process often shows clear pattern like seasonality and trend.

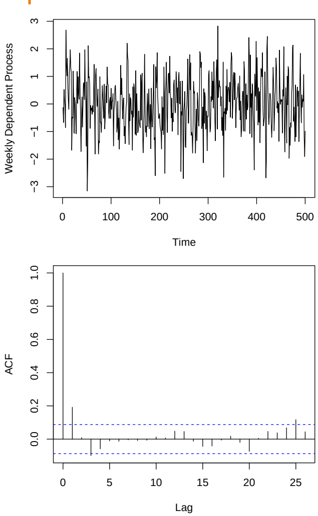
RM:
1. Plotting the time series is often the quickest way to tell stationarity.
2. There are few ways to test stationarity such as `adf.test()` in R.

Weakly Dependent Time Series

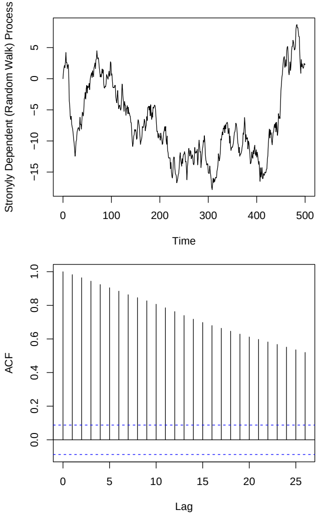
- Stationarity has to do with probability distribution of X_t as it moves across time. A different concept we need is *weakly dependent*.
- Weakly dependence: a *stationary* time series process is *weakly dependent* if correlation between X_t and X_{t+h} goes to zero sufficiently quickly as $h \to \infty$, or "**asymptotically uncorrelated**".
- The reason we need both stationarity and weakly dependence is to use OLS regression with large sample in time series analysis.

RM:
1. A nonstationary series leaves no hope to study its statistical property as it is elusively ever-changing. A persistent high autocorrelation among X_t makes all x_t one "same" observation.
2. For your interest, stationarity and weakly dependence are needed for law of large numbers (LLN) and central limit theorem (CLT) for large sample time series analysis.

Weakly Dependent Time Series



(a) a weakly dependent process and its autocorrelation plot



(b) a highly persistent process and its autocorrelation plot

RM: `acf()` plots autocorrelation between X_t and X_{t-h}, to uncover its structure.

OLS Regression in Time Series

Static and Finite Distributed Lag Model (FDL)

- Static Model: contemporaneous relationship among $(\boldsymbol{X}_t, Y_t, t = 1, 2, \ldots)$,

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_k X_{kt} + \epsilon_t \tag{3}$$

- Finite Distributed Lag Model (FDL) of order q: allows X_t and its q-order lags to affect Y_t,

$$Y_t = \alpha_0 + \delta_0 X_t + \delta_1 X_{t-1} + \cdots + \delta_q X_{t-q} + \epsilon_t \tag{4}$$

RM:
1. How do we interpret δ's in FDL? Suppose that at time t, x has a *permanent* increase of 1 unit. Then compared with the level of y_{t-1} (right before the change), $y_t - y_{t-1} = \delta_0$; $y_{t+1} - y_{t-1} = \delta_0 + \delta_1$; and so on up until q-periods after the change, $y_{t+q} - y_{t-1} = \delta_0 + \cdots + \delta_q$.
2. δ_0 is called impact propensity measuring the immediate effect of the change and $(\delta_0 + \cdots + \delta_q)$ is called long-run propensity of FDL.

Assumptions of OLS Regression in Time Series Analysis

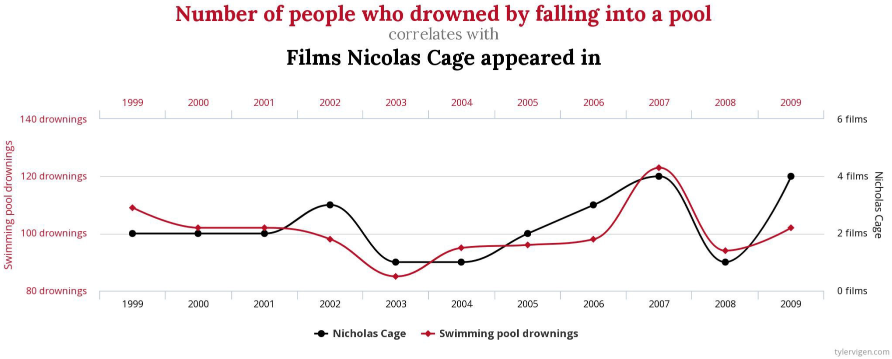
- Regression model in time series: $Y_t = \beta_0 + \beta_1 X_{1t} + \cdots + \beta_k X_{kt} + \epsilon_t$.

Assumption	Math Expression
TS1. Mean-Zero Error	$\mathbb{E}(\epsilon_t \| \boldsymbol{X}_t) = 0$, for all i
TS2. Homoskedasticity	$\text{Var}(\epsilon_t \| \boldsymbol{X}_t) = \sigma_\epsilon^2$, for all t
TS3. Uncorrelated Error	$\text{Cov}(\epsilon_t, \epsilon_s \| \boldsymbol{X}_t, \boldsymbol{X}_s) = 0$, for all $t \neq s$
TS4. Weakly Dependence	$\{(\boldsymbol{X}_t, Y_t), t = 1, 2, \ldots\}$ is stationary and weakly dependent
TS5. Linearity	$Y_t = \beta_0 + \beta_1 X_{1t} + \cdots + \beta_k X_{kt} + \epsilon_t$

RM:
1. Boldface $\boldsymbol{X}_t$ indicates that $\boldsymbol{X}_t$ is a vector, i.e. $\boldsymbol{X}_t = (X_{1t}, \ldots, X_{kt})$. Note that subscription k refers to k-th predictor, not entity.
2. TS1 - TS5 will make sure that OLS estimators in time series work exactly the same as before in cross-sectional data.

Spurious Relationship

- (To Nicolas Cage) Stop filming bad movies and save lives!?



- An example of spurious relationships.
  - Variables are strongly correlated simply because of shared time trend.

Regression in Time Series with Trend

- Many time series data often comes with a time *trend*.
- One easy mistake to say two or more trending time series Y_t and $\boldsymbol{X}_t$ have relationship simply because each happens to grow/shrink over time. An typical example of spurious regression.
- To solve this problem, simply **add a time trend variable t as a covariate**:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \cdots + \beta_k X_{kt} + \gamma t + \epsilon_t \qquad (5)$$

RM:
1. Just treating $X_{t+1} = t$, it fits into our multivariate regression framework as long as it satisfies assumption TS1-TS5.
2. In another word, omitting covariate t in (5) potentially yields biased estimator $\boldsymbol{\beta}$'s if Y_t and one of $\boldsymbol{X}_t$ are trending. (**recall that ϵ correlates with $X \Rightarrow$ biased OLS estimators**)

Run OLS Regression for Time Series in R

- Let's see an example of spurious regression in time series using data set `hseinv` on house investment time series data, where `invpc` and `price` are housing investment per capita and price index, respectively.
- Nothing special from what are doing in multivariate linear regression. Compare `invpc ~ price` vs. `invpc ~ price + t`.

```
fit_ip = lm(invpc ~ price, data = hseinv)
# fit OLS regression with additional time trend variable 't'.
fit_ipt = lm(invpc ~ price + t, data = hseinv)
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1366     0.2010   -0.68  0.50064
price         0.7209     0.2198    3.28  0.00215 **
---
(Intercept)  0.609042   0.313477   1.943   0.0593 .
price       -0.222725   0.378973  -0.588   0.5601
t            0.005375   0.001829   2.939   0.0055 **
```

RM: Trending variable t makes once significant price coef insignificant. Even the sign of `price` flips!

Another Way to Interpret: Detrending

- As we have emphasized that any trending time series is **nonstationary**, how adding a trend covariate t makes it stationary?
- Take example of (5), after OLS regressing Y_t on $\boldsymbol{X}_t$ and t:

$$\hat{y}_t = b_0 + b_1 x_{1t} + \cdots + b_k x_{kt} + \hat{\gamma} t \tag{6}$$

- "Magically" we can reproduce $(b_1, \ldots, b_k)$ by doing the following:
  1. Regress each y_t, x_{1t}, ... and x_{kt} on an intercept and the time trend, and save the residual from each regression, denoted as $\ddot{y}_t$, $\ddot{x}_{1t} \ldots$ and $\ddot{x}_{kt}$, e.g. $\ddot{y}_t \equiv y_t - a_0 - a_1 t$ from the regression $y_t = \alpha_0 + \alpha_1 t + \epsilon_t$. The residual $e_t \equiv \ddot{y}_t$, have the time trend removed, or being linearly detrended.
  2. Run the regression model: $\ddot{y}_t$ on $\ddot{x_{1t}} \ldots \ddot{x_{kt}}$. The estimated coefficients before $\ddot{\boldsymbol{x}}_t$ are exactly $\boldsymbol{b} = (b_1, \ldots, b_k)$

RM:
1. The "detrended" time series $\ddot{y}_t$ and $\ddot{\boldsymbol{x}}_t$ become stationary.
2. This is much more general result: residual in regressions can be seen as "after-treated" y with the treatment being the "model".

Regression in Time Series with Seasonality

- Compared to trending of time series, seasonality is less common simply because many time series have been *seasonally adjusted* at the source.
- In case you have raw data that is seasonally unadjusted, simply **include a set of seasonal dummy variables** in the regression. For instance,

$$\begin{aligned} Y_t = & \beta_0 + \beta_1 X_{1t} + \cdots + \beta_k X_{kt} \\ & + \delta_1 \mathrm{summer}_t + \delta_2 \mathrm{fall}_t + \delta_3 \mathrm{winter}_t + \epsilon_t \end{aligned}$$

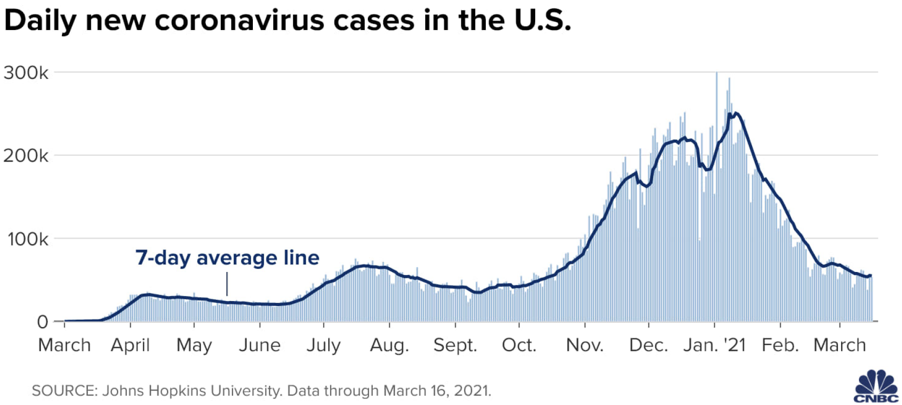
RM:
1. The seasonal dummy labels to which season this observation t belongs.
2. In this formulation, Spring_t is the reference level. Don't include it in the regression, otherwise you have a so-called perfect multicollinearity problem since four seasonal dummies always add up to one, or perfected correlated, for any observation. One of four needs to be excluded as reference level.
3. Similar to detrending, we can do the same exercise for "de-seasoning".

Exponential Smoothing Models and Auto-Regressive

Univariate Time Series Analysis

- Many time series alone contains useful information. Future value of the series can be predicted using its own past values (its own *lag* terms). A typical example is stock price prediction in financial market.
- Instead of introducing other $\boldsymbol{X}_t$ in the model, we now focus on how to extract useful information from one time series process alone, i.e. univariate time series analysis.
- Note that such univariate time series process still has to be both **stationary** and **weakly dependent** for valid analysis.
- A family of popular univariate time series models is exponential smoothing models.

How Would You Predict Y_{t+1} with Time Series?



- Using observe only: $(y_1, \ldots, y_t, \ldots, y_{T-1}, y_T)$. What would be $\hat{y}_{T+1}$?

How Would You Predict y_{T+1} with Time Series?

Question

If you could only use observed time series $(y_1, y_2, \ldots, y_{10})$, what should be $\hat{y}_{11}$?

Prediction $\hat{y}_{11}$	Formula for $\hat{y}_{11}$	Model
today's observed value	$\hat{y}_{11} = y_{10}$	Naïve
avg. of all past values	$\hat{y}_{11} = (y_1 + \cdots + y_{10})/10$	Simple average
avg. of 3 immediate lags	$\hat{y}_{11} = (y_8 + y_9 + y_{10})/3$	Moving average
avg. between today's and all previous values with fixed weight $(0.6, 0.4)$	$\hat{y}_{11} = 0.6 \times y_{10} + 0.4 \times \hat{y}_{10}$, $\hat{y}_{10} = 0.6 \times y_9 + 0.4 \times \hat{y}_8$, $\cdots$, $\hat{y}_1 = y_1$.	Exponential smoothing

How Would You Predict y_{T+1} with Time Series?

- Observe only: $(y_1, y_2, \ldots, y_{T-1}, y_T)$. How to predict $\hat{y}_{T+1}$?

Prediction $\hat{y}_{T+1}$	Formula for $\hat{y}_{T+1}$	Model
today's observed value	$\hat{y}_{T+1} = y_T$	Naïve
avg. of all past values	$\hat{y}_{T+1} = (y_1 + \cdots + y_T)/T$	Simple average
avg. of K-immediate lags	$\hat{y}_{T+1} = (y_{T-K+1} + \cdots + y_T)/K$	Moving average (of K-period window)
avg. with exponential weight α	$\hat{y}_1 = y_1$, $\hat{y}_2 = \alpha \cdot y_1 + (1-\alpha) \cdot \hat{y}_1$, $\cdots$, $\hat{y}_{T+1} = \alpha \cdot y_T + (1-\alpha) \cdot \hat{y}_T$	Exponential smoothing

- Think about the following questions:
  - Differences between the models?
  - What is the parameter we are using for prediction of Y_{T+1}?
  - Why do you think we need stationarity for a time series?

Smoothing A Time-Series: Moving Average

- Moving average is a simple technique to smooth the series by computing the average of a moving widow of K-period.

$$m_t = (y_t + y_{t-1} + \cdots + y_{t-K+1})/K \tag{MA}$$

  - m_t is the smoothed series by moving average.

- To compute the series of (MA), use `TTR::SMA(df$y, n = k)` for a window of k-period.

RM:
1. Taking average of a moving window, smooths the original series and dampens its idiosyncratic noises.
2. The width of the moving window K, determines to what degree historical information is incorporated (at an equal weight).

Forecasting A Time-Series: Moving Average

- Stationarity makes sure that a stable Y_t was generating the observed $\{y_t\}$. We leveraged the mean behavior of Y_t for prediction, i.e. $\hat{Y}_{t+1} = \mathbb{E}(Y_t)$, which is inferred by the moving averages $\{m_t\}$.
- To forecast the time series based on moving average series m_t, simply use today's moving average values for tomorrow's predictions, i.e.,

$$\hat{y}_{t+1} = m_t \tag{MA-f}$$

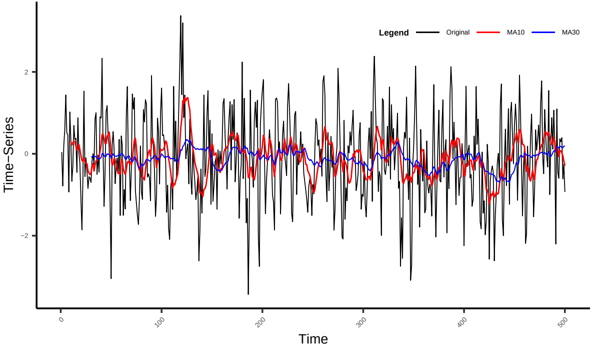
  - In this time series lecture, we always use $\hat{y}_t$ to denote the forecast values.

RM:
1. We are assuming that m_t is a good "summary" of historical information in recent observation, i.e., an estimate for the mean.
2. Then we say a good prediction for tomorrow is simply the this m_t.
3. In the form of dataframe, we simply "shift" m_t by one row to get $\hat{y}_t$.

Forecasting A Time-Series: Moving Average

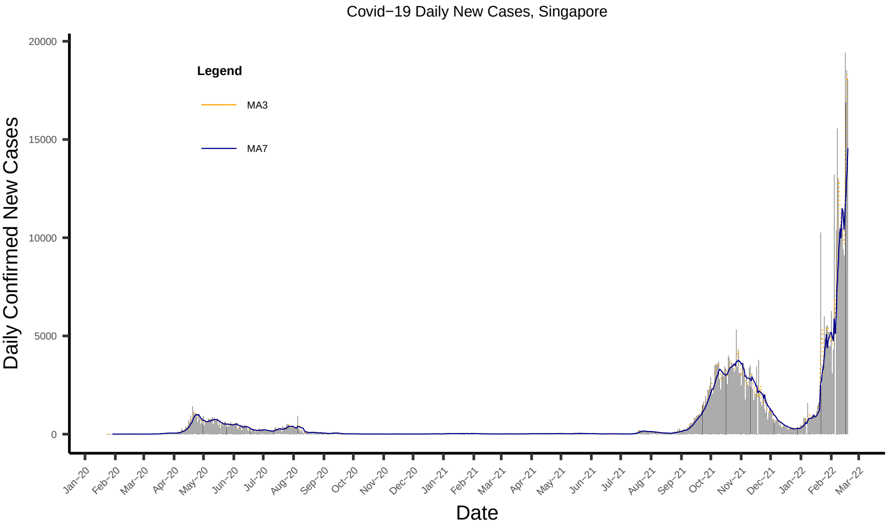
t	y_t	m_t	$\hat{y}_t$
1	y_1	NA	NA
2	y_2	m_2	NA
3	y_3	m_3	$\hat{y}_3$
4	y_3	m_4	$\hat{y}_4$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
9	y_9	m_9	$\hat{y}_9$
10	y_{10}	m_{10}	$\hat{y}_{10}$

Obs. vs. MA vs. Pred



Larger K, more smooth the MA.

Confirmed New Covid-19 Cases in SG



Smoothing A Time-Series: Exponential Smoothing Models

- (Simple) exponential smoothing computes the averages with all previous data, and a fixed $\alpha \in (0, 1)$ weight on today's value.

$$s_t = \alpha \cdot y_t + (1 - \alpha) \cdot s_{t-1} \qquad \text{(EXP1)}$$
$$s_1 = y_1$$

  - s_t and s_{t-1} are the smoothed values by exponential smoothing for today and yesterday, respectively.

- To see why named "exponential": exponential weights.

- To forecast (one-step) with exponential smoothing: $\hat{y}_{t+1} = s_t$.

RM:
1. Contrast to MA, exponential smoothing leverages all past information but with more weight (i.e., α) on recent observations.
2. When $\alpha = 1$, we have naïve forecast. When $\alpha = 0$, $s_t = y_1$ for all t.

Smoothing A Time-Series: Exponential Smoothing Models

- Simple exponential fails when the original series exhibits trend or/and seasonality (nonstationary).

- Double exponential smoothing takes trend into consideration by incorporating a trend-"slope" that is updating by exponential smoothing.

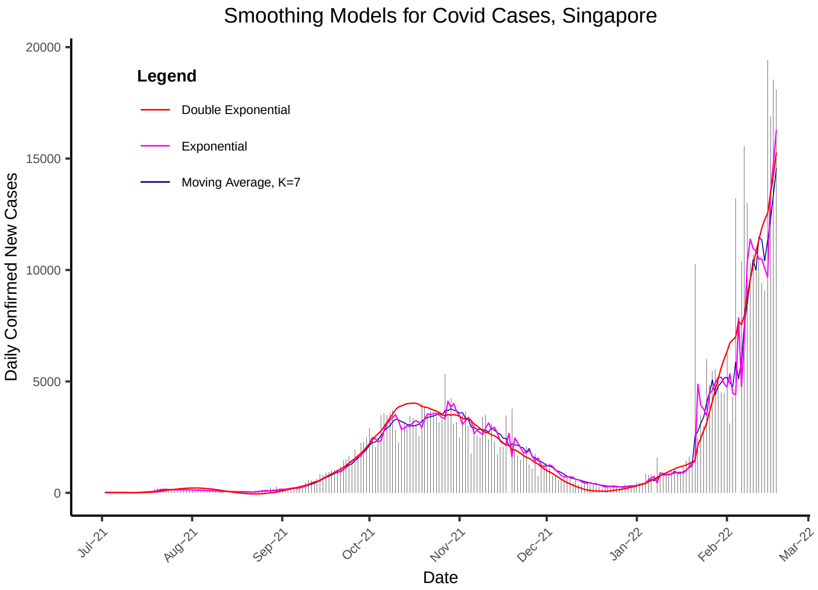
$$s_t = \alpha \cdot y_t + (1-\alpha) \cdot (s_{t-1} + b_{t-1}) \tag{EXP2}$$
$$b_t = \beta \cdot (s_t - s_{t-1}) + (1-\beta) \cdot b_{t-1}$$
$$s_1 = y_1 \text{ and } b_1 = y_2 - y_1$$

  - b_t are the "slopes" for the trend, an (exponentially) weighted average between the recent trend, $(s_t - s_{t-1})$ and all past trends, summarized by b_{t-1}.

- To forecast m-step into the future, with double exponential smoothing: $\hat{y}_{t+m} = s_t + m \cdot b_t$.

RM: 1 Double exponential smoothing has two parameters α and β for smoothed level s_t and trend b_t, respectively.

Exponential Smoothing for Covid-19 Cases SG



Smoothing A Time-Series: Exponential Smoothing Models

- Triple exponential smoothing (Holt-Winters) takes one step further to account for seasonality.
- Similar to double exponential, one additional parameter γ governs the exponential smoothing process for an updated seasonal cycle corrections.
- In R, use `HoltWinters(x, alpha, beta, gamma, ..)`.
- Forecasting with Holt-Winters is based on both trending and seasonal factors.

Model	Trend	Seasonality	Parameters	Calling `HoltWinters(..)`
Single	No	No	α	`x, beta=FALSE, gamma=FALSE`
Double	Yes	No	α, β	`x, gamma=FALSE`
Holt-Winters	Yes	Yes	α, β, γ	`x`

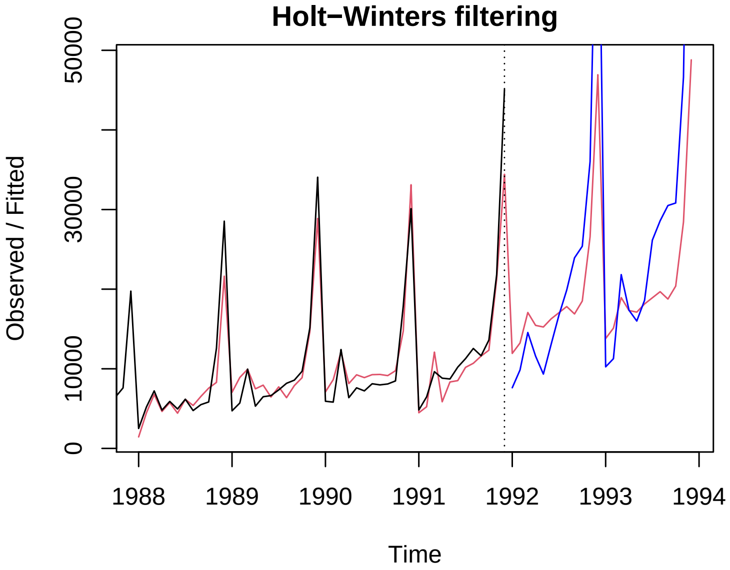
RM:
1. Model parameters (α, β, γ) could be specified by analyst, or estimated.
2. They are estimated by minimizing the sum square residuals, $\sum_t (y_t - \hat{y}_t)^2$. Call `HoltWinters$SSE` for SSR.

Split Data into Training and Testing

- To test the predictive accuracy of the model, a common practice is to split the original data into train vs. test sets.
  - Model is trained on the training set and its predictions are compared to the "holdout" testing set.

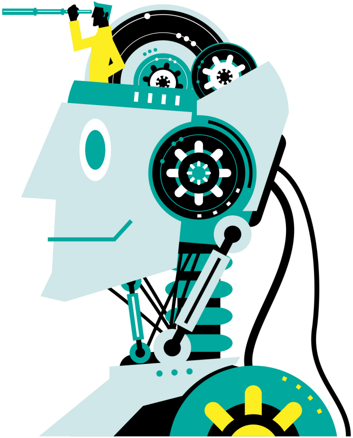
```
# split the souvenir into training set Jan87-Dec91 and test set Jan92-Dec93
souvenir_train = window(souvenirsale, start = 1987, end = c(1991,12))
souvenir_test = window(souvenirsale, start = c(1992,1), end = c(1993,12))

# train the HoltWinters on the training date
souvenir_hw_train = HoltWinters(souvenir_train)
# let's predict Jan1992-Dec1993 with the Holt-Winters model
souvenir_pred_train = predict(souvenir_hw_train, n.ahead = 24)

# visually comparison
plot(souvenir_hw_train, souvenir_pred_train)
lines(souvenir_test, col = "blue")

# quantify the difference in terms of sum square errors
sqrt(mean(souvenir_pred_train - souvenir_test)^2)
```

Split Data into Training and Testing



Summary

- Logistic regression is one of the most popular classifiers in either academia or industry. The binary Y is nonlinear but log-odds of Y is still linear in in $\boldsymbol{X}\beta$.
- Don't be fooled by spurious relationship!
- Machine is dumb. It is up to analyst's discretion for correct choice of models. Apply proper smoothing model based on your observation for trend and seasonality.



Online Assessment

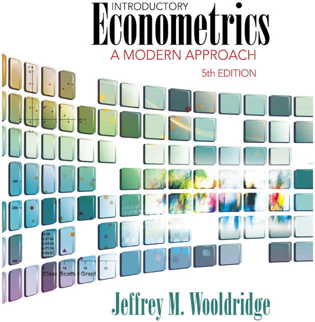
- Online Assessment: **Tuesday in two weeks, Oct 18**.
  - Exam window: **12:00 - 1:00 PM**.
  - Time to finish: **1 hour**.
  - Place: **Examplify** and online proctoring
  - Coverage: Week 1 - Week 8; more on descriptive analytics.
  - Format: 10 MCQ and 2 Short Answers.
  - Open book/note: YES.
  - Individual assessment: YES.

Some Preparation

- Getting familiar with Examplify.
  - Common Briefing (strongly recommended): **Oct 10, 10-11a**. Join this link with password:742374.
  - Online instruction: https://wiki.nus.edu.sg/display/DA/Download+and+Install+Examplify.
- Online Proctoring policy and screen recording.
  - Policy and guideline on proctoring: https://wiki.nus.edu.sg/display/DA/Proctoring+Remote+Assessments+-+Student.
  - Screen recording tools: https://www.comp.nus.edu.sg/images/Panopto.pdf and https://cit.nus.edu.sg/services/software/screen-recording/.
  - Guide to upload screen recording in Canvas: https://wiki.nus.edu.sg/pages/viewpage.action?pageId=404358262.
  - Try screen recording yourself and upload sth onto the Canvas folder.

Recommended Reading (Optional)

- Chapter 7 and 10.
  - Wooldridge, J.M. (2013). *Introductory Econometrics: A Modern Approach*. Cengage Learning. ISBN: 9781111531041.



Maximum Likelihood Estimators (MLE) for Logistic Regression

- Instead of OLS, logistic regression is estimated with maximum likelihood estimation (MLE).
- We assume that binary $y_i \in \{0, 1\}$ follows independent Bernoulli event of success with prob $p_i \equiv \mathrm{P}(y_i = 1|\boldsymbol{X}_i)$ for data point $i = 1, 2, \ldots, n$.
- $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_n)$ is called maximum likelihood estimators since $\hat{\boldsymbol{\beta}}$ maximize the joint probability (or likelihood):

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i} \qquad \text{(Likelihood)}$$

RM:
1. Observe that the "success" probability $p_i = p_i(\boldsymbol{X_i}, \beta)$.
2. MLE estimator $\hat{\beta}$ is the solution to $\max_\beta L(\beta)$.
3. A good read for maximum likelihood estimation here.

Back

Why the Name of "Exponential" Smoothing?

- The exponential smoothing model puts α on today's obs and $(1-\alpha)$ on s_{t-1}, a "summary" of all history up to yesterday.
- Equivalently, exponential smoothing (EXP2) is a weighted average of all past obs. with a geometric weights.

$$
\begin{aligned}
s_t &= \alpha y_t + (1-\alpha) s_{t-1} \\
&= \alpha y_t + (1-\alpha)\left(\alpha y_{t-1} + (1-\alpha) s_{t-2}\right) \\
&= \alpha y_t + \alpha(1-\alpha) y_{t-1} + (1-\alpha)^2 s_{t-2} \\
&= \quad \cdots \\
&= \alpha\left(y_t + (1-\alpha) y_{t-1} + (1-\alpha)^2 y_{t-2} + \cdots + (1-\alpha)^{t-2} y_2\right) \\
&\quad\quad + (1-\alpha)^{t-1} y_1
\end{aligned}
$$

- The geometric weights, $1, (1-\alpha), (1-\alpha)^2, \ldots, (1-\alpha)^t, \ldots$, is the discrete version of exponential function, $f(x) = e^x$, hence the name.

Back