Model-Based TS Forecasting
Simple Regression

Nagiza F. Samatova, samatova@csc.ncsu.edu
Professor, Department of Computer Science
North Carolina State University

Senior Scientist, Computer Science & Mathematics Division
Oak Ridge National Laboratory

TS Data Analysis Methods

TS Data Analysis & Forecasting

Data-Driven

Data-driven methods are used when model assumptions are likely to be violated, or when the structure of time series changes over time.

- **Baseline**: average, naive, seasonal naive, drift
- **Differencing**
- **Smoothing**: moving average, exponential smoothing

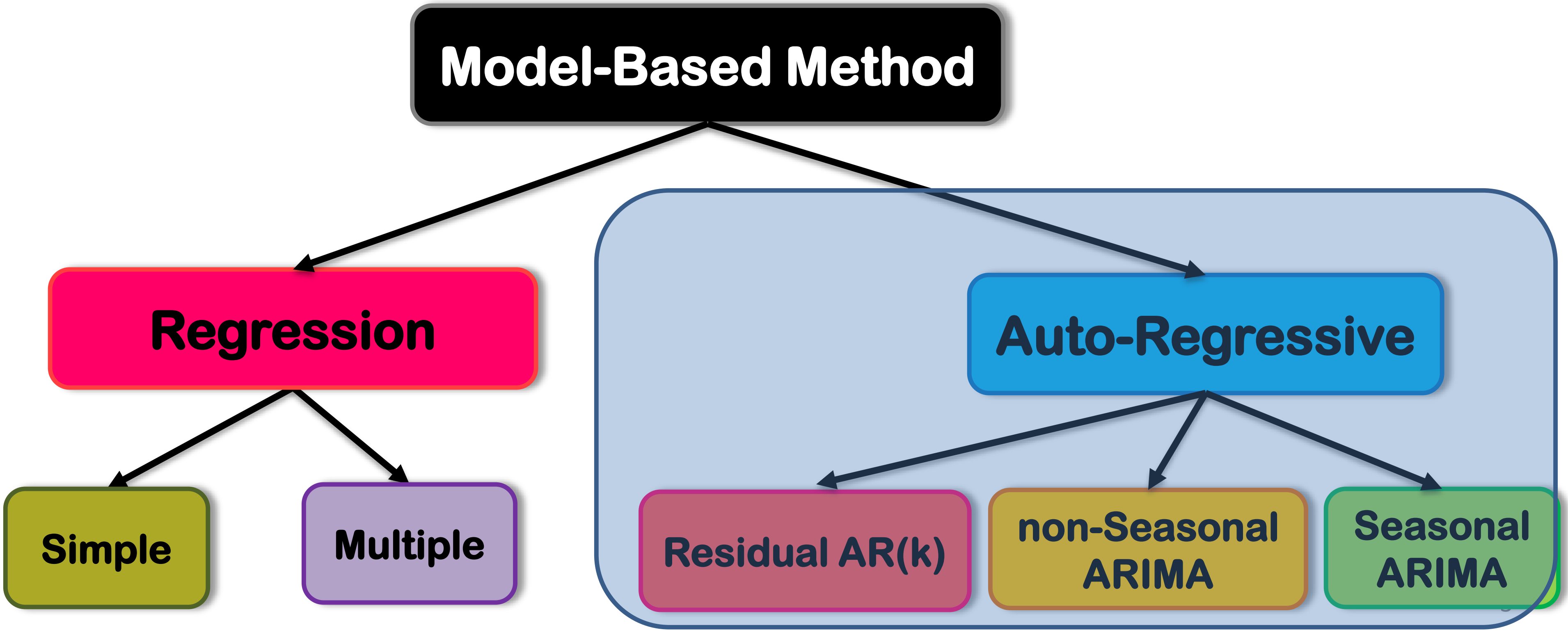
Model-based

Training data is used to estimate model parameters, and then the model with these parameters is used to generate forecasts.

- **ARIMA**
- **Linear Regression**
- **Logistic Regression**
- **Neural Networks**

TS Forecasting

REGRESSION



TS Parts: Systematic vs Non-systematic

TS Part	Definition	Detection	How to deal w/
Level	Average value of ts		
Trend	Long-term increase decrease in the data	lag.plot	De-trend via lag-1 differencing
Seasonality	Variations occurring during known periods of the year (monthly, quarterly, holidays)	lag.plot, Acf plots	De-seasonalize via lag-k differencing
Cycles	Other oscillating patterns about the trend (e.g., business or economic conditions)		
Auto-correlation	Correlation between neighboring points in ts	Acf, lag.plot	
Noise	Residuals after level, trend, seasonality, and cycles are removed	Normality tests	

Regression-based Models

- Linear regression model can be set up to capture a time series with a trend and/or seasonality.
- Common trends
  - linear
  - exponential
  - polynomial
- Common seasonality
  - additive
  - multiplicative
- Use of sine and cosine terms
- Regression model can be used to quantify the correlation between neighboring values in a time series (called autocorrelation)
  - This type of model called an autoregressive (AR) model
  - It is useful for evaluating the predictability of a series
    - Is it just a random walk?

Model with Trend

- Linear regression can be used to fit a global trend that applies to the entire series and will apply in the forecasting period.
  - A linear trend means that the values of the series increase or decrease linearly in time
  - An exponential trend captures an exponential increase or decrease.
  - Quadratic functions or higher order polynomials can be used to capture more complex trend
- How can linear regression be set up for all these common trend types?

Simple Regression with Linear Trend

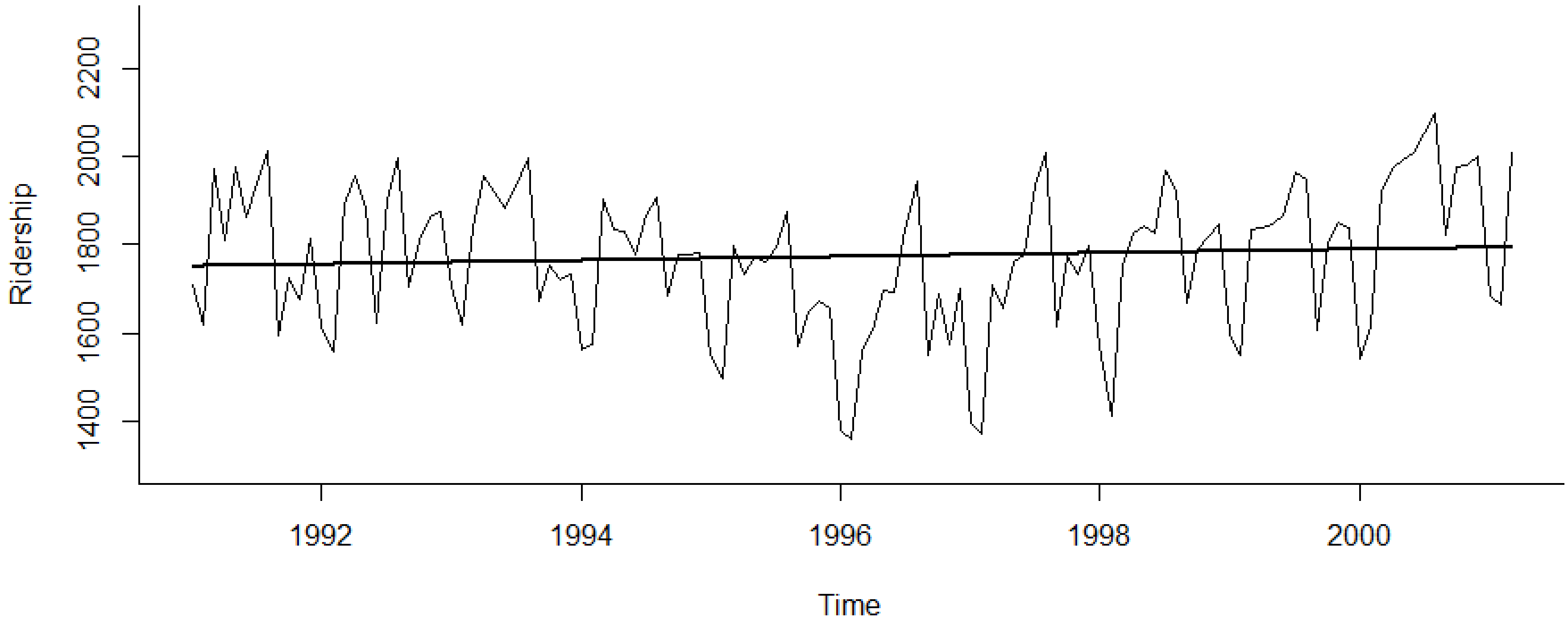
- **Example: fitting a linear trend to the Amtrak ridership**
  - **create a new column that is a time series index $t = 1,2,3$ to serve as a predictor variable**
  - **partition the time series into training and validation periods**
    - keep 12 month in the validation set to
      - provide monthly forecasts for the year ahead
      - allow evaluation of forecasts on different months of the year
- **To fit a linear relationship between Ridership and Time, we set output variable y as the Amtrak ridership and the predictor as time index t in the regression model:**

$$\boldsymbol{y_t = \beta_0 + \beta_1 t + \epsilon}$$

  - where y_t is the Ridership at time point t and ϵ is the standard noise term in a linear regression
- **Thus three of the four time series components are modeled:**
  - **level (β_0), trend (β_1) , and noise (ϵ)**

Fitting Linear Trend

```
train.lm <- tslm(train.ts ~ trend)

plot(train.ts, xlab = "Time", ylab = "Ridership",
      ylim = c(1300, 2300), bty = "l")
lines(train.lm$fitted, lwd = 2)
```



Is it Significant?

- **A significant coefficient for trend does not mean that a linear fit is adequate.**
- **An insignificant coefficient does not mean that there is no trend in the data.**
  - **In the example, the slope coefficient (0.3514) is insignificant (p-value=0.39), yet there may be a trend in the data**
    - often once we control for seasonality
    - when run a linear regression on the de-seasonalized data we find the trend coefficient (0.8294) is statistically significant ($p\text{-value} < 0.001$)
- **To determine suitability of any trend shape, look at the time plot of the (de-seasonalized) time series with the trend overlaid**
  - **examine the residual time plot**
  - **look at performance measures on the validation period**

Regression with Exponential Trend

- **Exponential trend implies a multiplicative increase/decrease of the series over time**
  - $y_t = ce^{\beta_1 t + \epsilon}$
- **To fit an exponential trend simply replace the output variable y with $log(y)$ and fit a linear regression:**
  - $\log(y_t) = \beta_0 + \beta_1 t + \epsilon$
  - **log is the natural logarithm with base of e**
  - In example we would fit a linear regression of $log(Ridership)$ on the index variable t.
- **Exponential trends are popular in sales data, where they reflect percentage growth**

Regression with Polynomial Trend

- Polynomial trend is easy to fit via linear regression as well
- In particular, a quadratic relationship of the form
  - $\boldsymbol{y_t = \beta_0 + \beta_1 t + \beta_2 t^2} + \epsilon$
  - This is done by creating an additional predictor t^2 and fitting a multiple linear regression with the two predictors t and t^2

Model with Seasonality

- **Examples are**
  - **day-of-week patterns,**
  - **monthly patterns**
    - Amtrak ridership example exhibits strong monthly seasonality with highest traffic during summer months
  - **quarterly patterns**
- **The most common way to capture seasonality in a regression model is by creating a new categorical variable that denotes the season for each observation.**
  - This categorical variable is then turned into dummy variables, which in turn are included as predictors in the regression model
  - For m seasons we create $m - 1$ dummy variables, which are binary variables that take on the value of 1 if the record falls in that particular season, and 0 otherwise.
  - The m^{th} season does not require a dummy, since it is identified when all the $m - 1$ dummies take on zero values

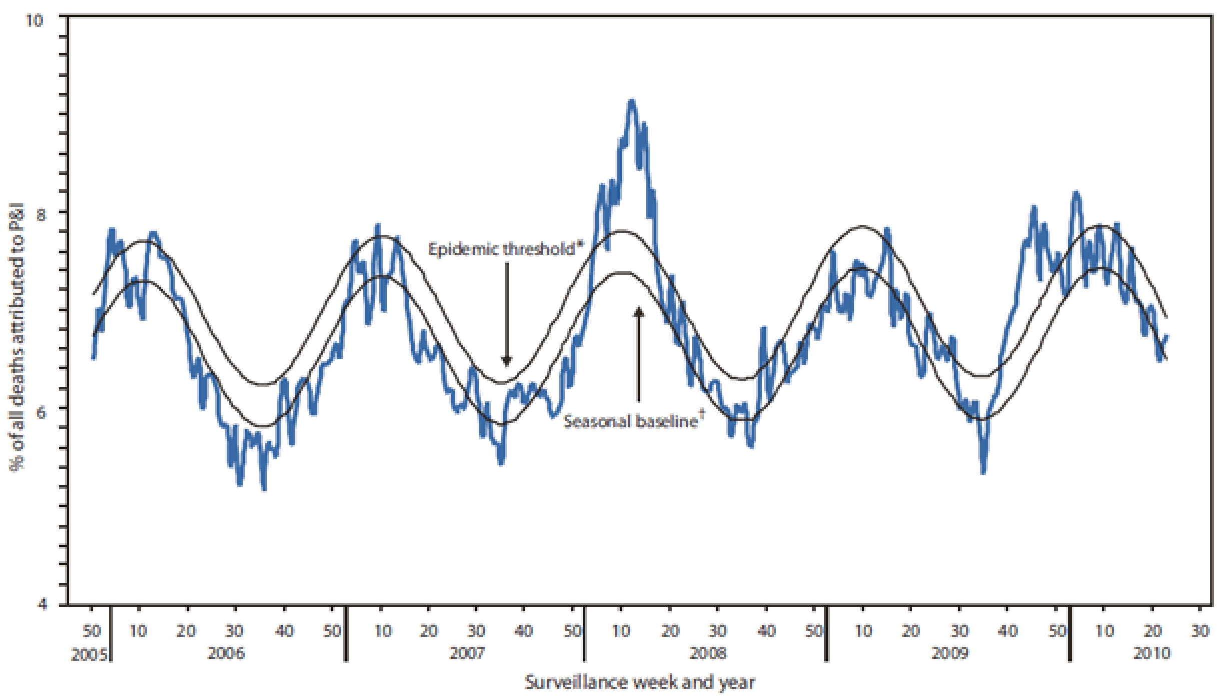
Model with Trend and Seasonality

- **Amtrak example:**
- **We fit a model with 13 predictors: 11 dummy variables for month, and t and t^2 for trend.**

```
> train.lm.trend.season <- tslm(train.ts ~ trend + I(trend^2) + season)
> summary(train.lm.trend.season)

Call:
lm(formula = formula, data = "train.ts", na.action = na.exclude)

Residuals:
     Min       1Q   Median       3Q      Max 
-213.775  -39.363    9.711   42.422  152.187 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.697e+03  2.768e+01  61.318  < 2e-16 ***
trend       -7.156e+00  7.293e-01  -9.812  < 2e-16 ***
I(trend^2)   6.074e-02  5.698e-03  10.660  < 2e-16 ***
season2     -4.325e+01  3.024e+01  -1.430  0.15556    
season3      2.600e+02  3.024e+01   8.598 6.60e-14 ***
season4      2.606e+02  3.102e+01   8.401 1.83e-13 ***
season5      2.938e+02  3.102e+01   9.471 6.89e-16 ***
season6      2.490e+02  3.102e+01   8.026 1.26e-12 ***
season7      3.606e+02  3.102e+01  11.626  < 2e-16 ***
season8      4.117e+02  3.102e+01  13.270  < 2e-16 ***
season9      9.032e+01  3.102e+01   2.911  0.00437 ** 
season10     2.146e+02  3.102e+01   6.917 3.29e-10 ***
season11     2.057e+02  3.103e+01   6.629 1.34e-09 ***
season12     2.429e+02  3.103e+01   7.829 3.44e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 70.92 on 109 degrees of freedom
Multiple R-squared:  0.8246, Adjusted R-squared:  0.8037 
F-statistic: 39.42 on 13 and 109 DF,  p-value: < 2.2e-16
```

Sinusoidal Functions for Smooth Seasonality

- **Example**
  - CDC regression model for the percent of weekly deaths attributed to pneumonia & influenza in 122 cities.
  - The model includes a quadratic trend as well as sine and cosine functions for capturing the smooth seasonality pattern
    - $y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 \sin\left(\frac{2\pi t}{52.18}\right) + \beta_4 \cos\left(\frac{2\pi t}{52.18}\right) + \epsilon$



http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5929a2.htm

Sinusoidal Fit in CDC Example

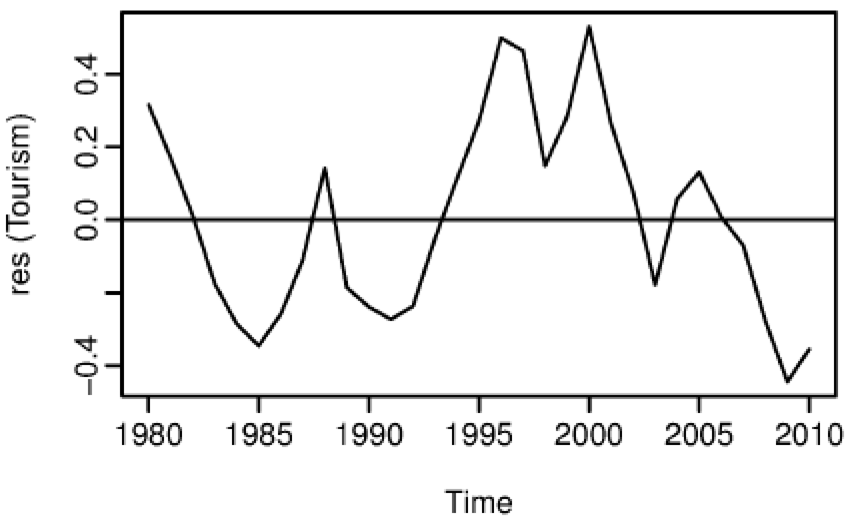
- The trend terms t and t^2 accommodate long-term linear and curvilinear changes in the background proportion of pneumonia & influenza death arising from factors such as population growth or improved disease prevention or treatment.
- The sine and cosine terms capture the yearly periodicity of weekly data (with 52.18 weeks per year).

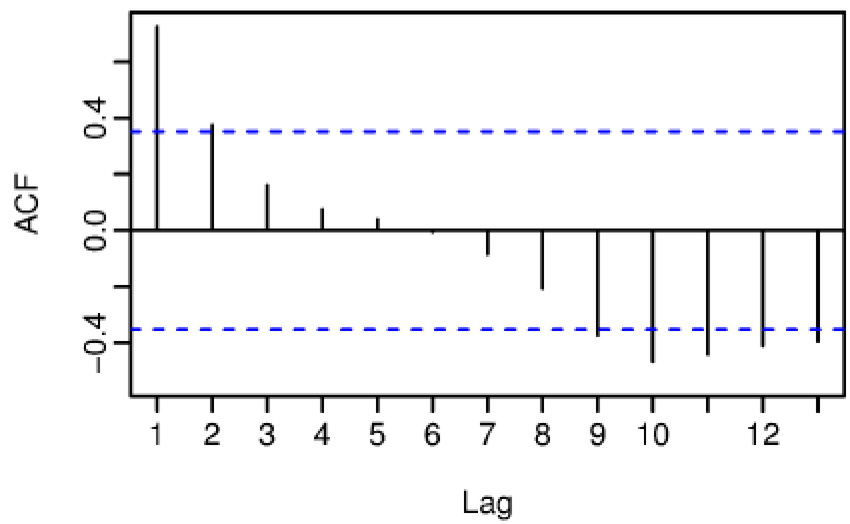
Sinusoidal Fit in Amtrak Example

```
> train.lm.trig <- tslm(train.ts ~ trend + I(trend^2) + I(sin(2*pi*trend/12)) + I(cos(2*pi*trend/12)))
> summary(train.lm.trig)

Call:
lm(formula = formula, data = "train.ts", na.action = na.exclude)

Residuals:
    Min      1Q  Median      3Q     Max 
-301.29  -76.47   25.29   91.41  235.12 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                1.899e+03  3.339e+01  56.891  < 2e-16 ***
trend                     -6.833e+00  1.243e+00  -5.497 2.26e-07 ***
I(trend^2)                 5.852e-02  9.712e-03   6.025 1.97e-08 ***
I(sin(2 * pi * trend/12)) -5.435e+01  1.542e+01  -3.524 0.000606 ***
I(cos(2 * pi * trend/12)) -1.100e+02  1.553e+01  -7.083 1.10e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 121.1 on 118 degrees of freedom
Multiple R-squared:  0.4465, Adjusted R-squared:  0.4277 
F-statistic:  23.8 on 4 and 118 DF,  p-value: 1.911e-14
```

Residual Autocorrelation in Linear Regression

- **Autocorrelation in TS Data**:
  - The value of a variable observed in the current time period will be influenced by its value in the previous period(s)
- **Regression for TS Data:**
  - When fitting a regression model to time series data, it is very common to find **autocorrelation in the residuals**.
    - Estimated model violates the assumption of no autocorrelation in the errors
    - Forecasts may be inefficient: there is some information left over that should be utilized in order to obtain better forecasts.
    - Forecasts from a model with autocorrelated errors are still unbiased, and so are not wrong, but usually have larger prediction intervals.



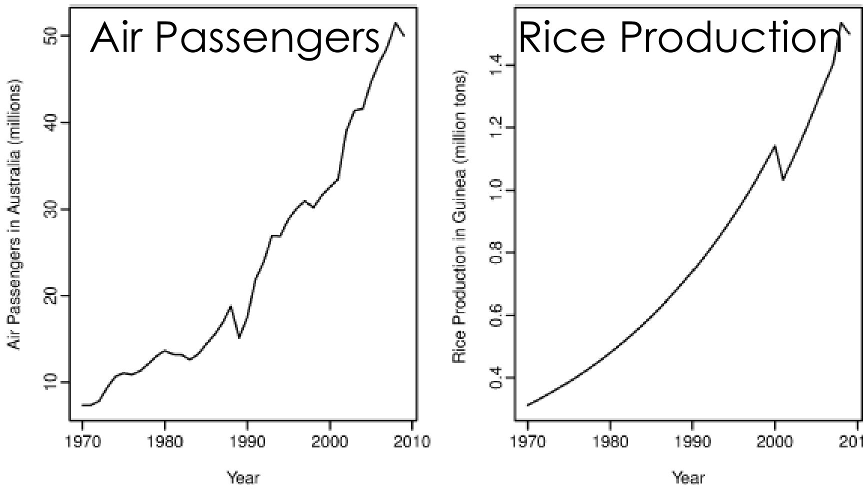


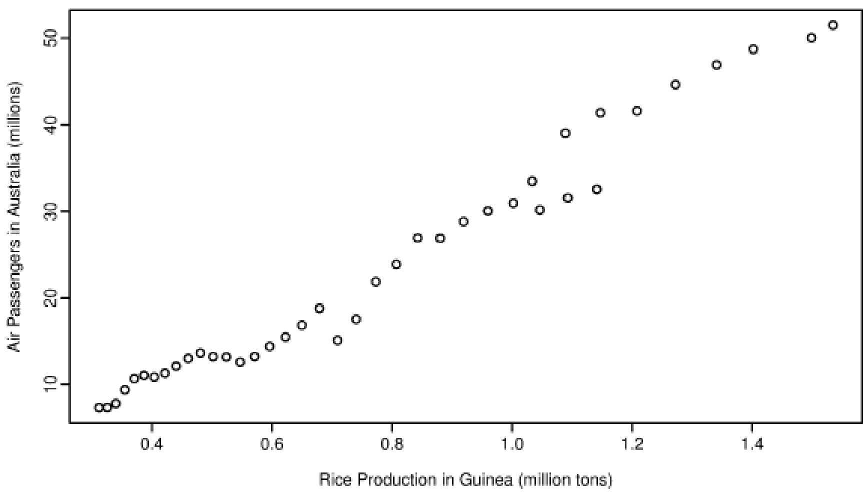
Non-stationarity Effect on Linear Regression

- **Non-stationary TS Data**:
  - Time series data are often non-stationary: values do not fluctuate around a constant mean or with a constant variance.
  - Regressing non-stationary ts can lead to spurious regression

- **Spurious Regression:**
  - Two trending time series may appear to be related.

Air passengers in Australia are regressed against rice production in Guinea.

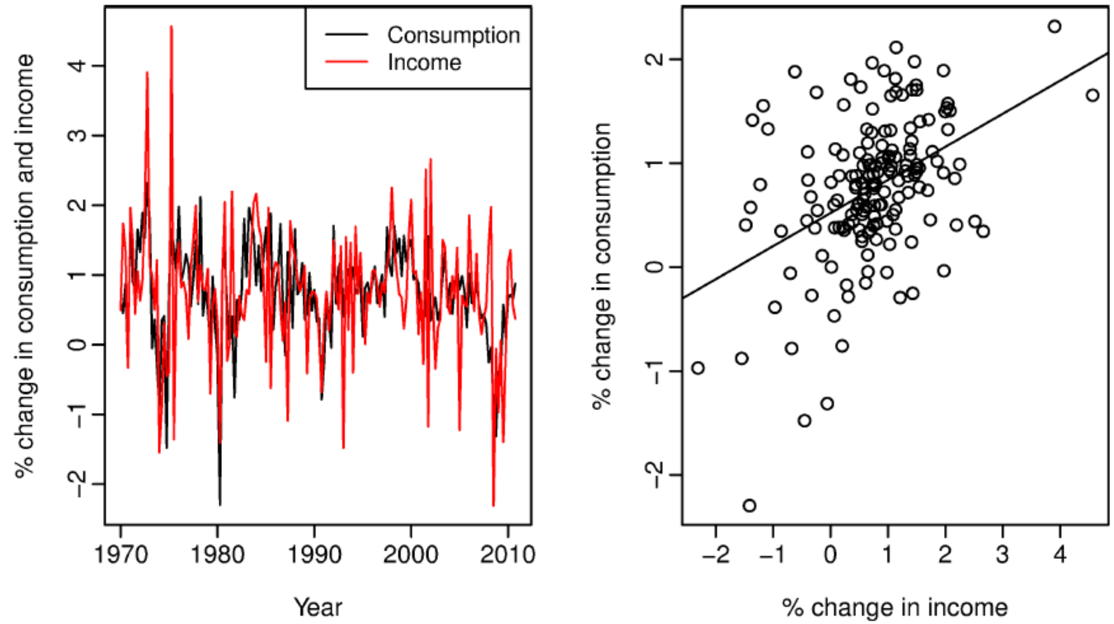




Unavailable Future Values for Predictors

- Using a regression model to forecast time series data poses a challenge:
  - **Future values of the predictor variable (e.g. Income) are needed to be input into the estimated model, but these are not known in advance.**

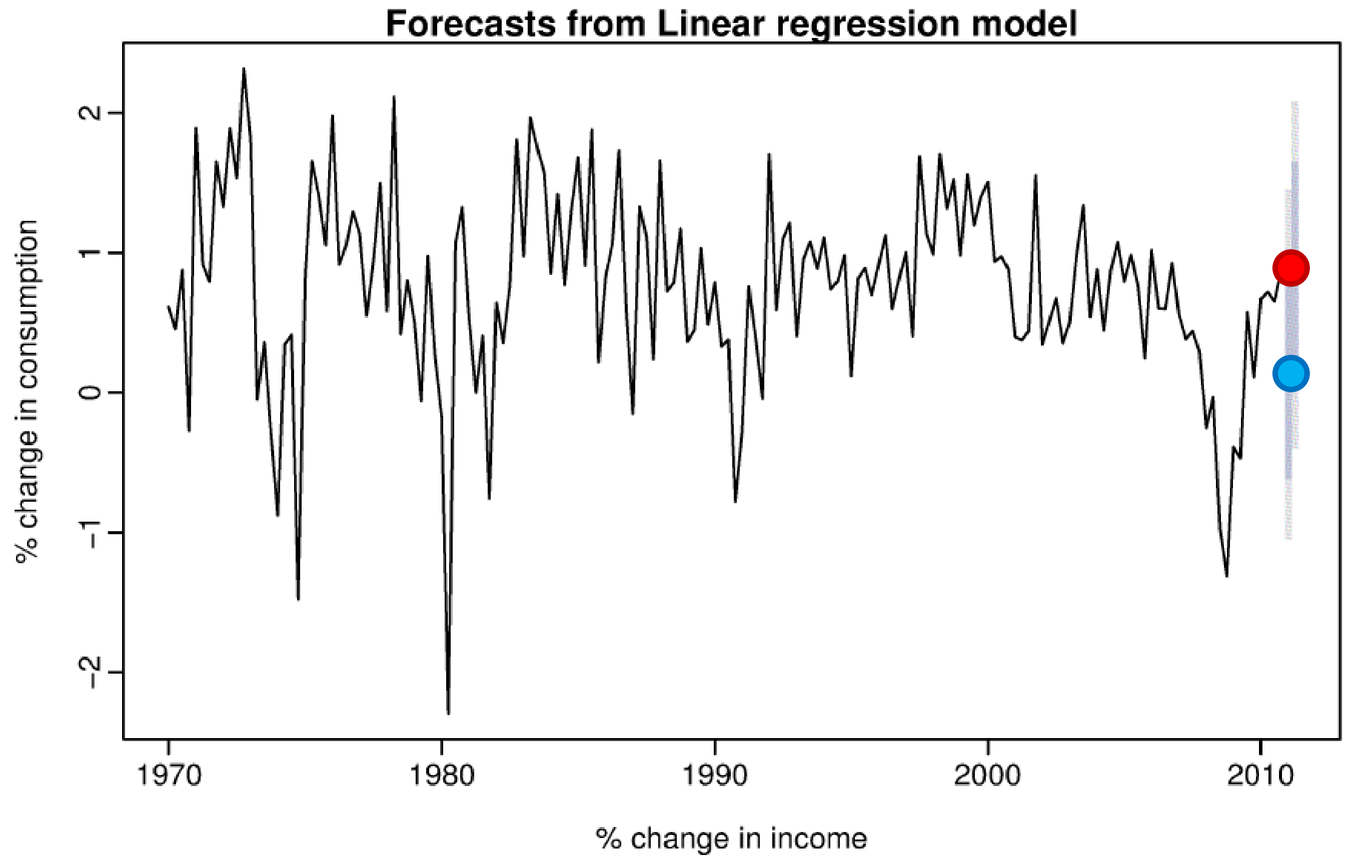
Percentage changes (growth rates) of real personal consumption expenditure (C) and real personal disposable income (I) for the US for the period March 1970 to Dec 2010.



- Response (C): Percentage change in consumption
- Predictor (I): Percentage change in income
- Regression Model: $\hat{C} = 0.52 + 0.32 \; * I$
  - A 1% increase in personal disposable income will result in an average increase of 0.84% in personal consumption expenditure.

Scenario-based Forecast for Unavailable Predictor

- The forecaster assumes **possible scenarios** for the unavailable predictor:
  - For example the US policy maker may want to forecast consumption if there is a **1% growth in income** for each of the quarters in 2011.
  - Alternatively, a **1% decline** in income for each of the quarters may be of interest.



Acknowledgements

- Books
  - Free and online (otexts.com/fpp): Forecasting Principles & Practice by R. Hyndman, G. Athanasopoulos ← Excellent Book!!!
  - Practical Time Series Forecasting with R: A Hand-on Guide by Shmueli & Lichtendahl
- Packages
  - R: fpp (install.packages ("fpp", dependencies=TRUE))