DS-GA 3001.009
Modeling Time Series Data
Lab 6

Artie Shen | Center for Data Science

NYU

Agenda

- Recap
  - SARIMA
  - Preprocessing
  - Model Selection

- Programming - CO_2 Proportion Trend
  - Pre-processing
  - Model Selection

Preprocessing

Goal Make sure the mean and variance structure is regular and satisfy the conditions of stationary.

Detrend

- Trend Stationary $x_t = \mu_t + y_t$, where μ_t is a function of t denoting the trend and y_t is a stationary process.
- Run linear regression on x_t to obtain an estimator for trend $\hat{\mu}_t$
- Detrend the process $\hat{x}_t = x_t - \hat{\mu}_t$

Differecing

- Random Walk $\mu_t = \mu_{t-1} + \delta + w_t, x_t = \mu_t + y_t = \delta + \mu_{t-1} + w_t + y_t$
- Differencing $\bigtriangledown x_t = x_t - x_{t-1}$
- $x_t - x_{t-1} = \delta + w_t + y_t - y_{t-1}$

Preprocessing

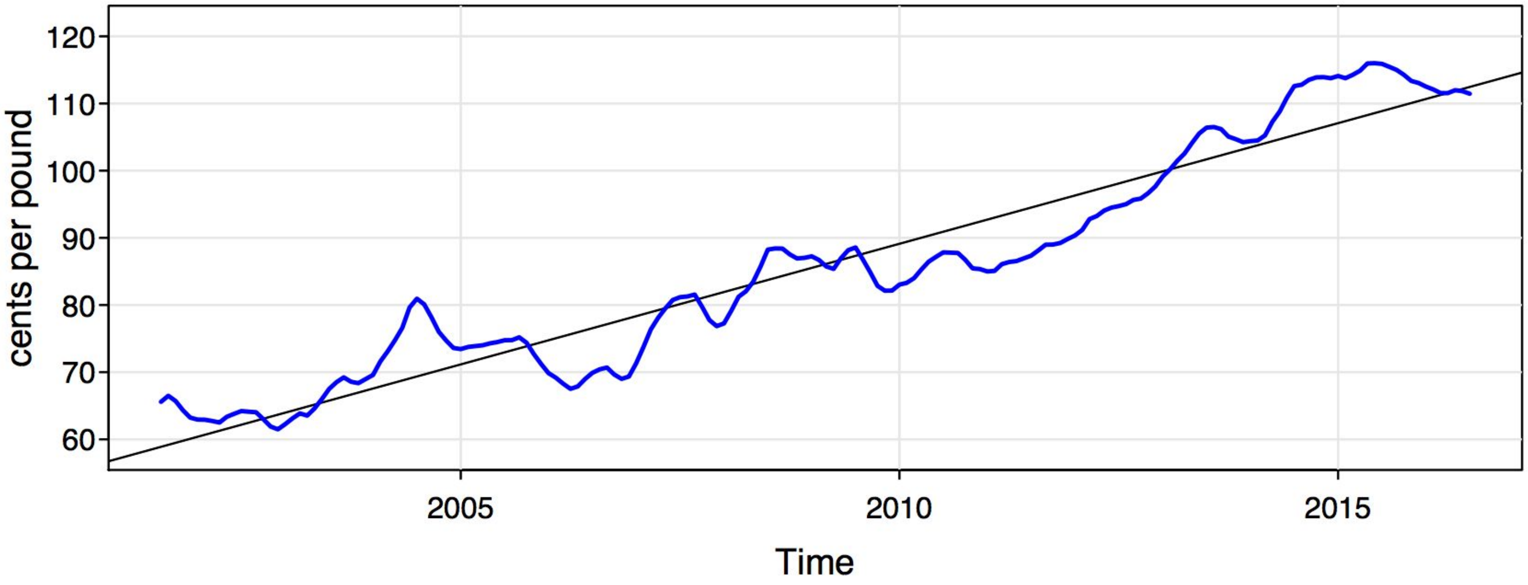
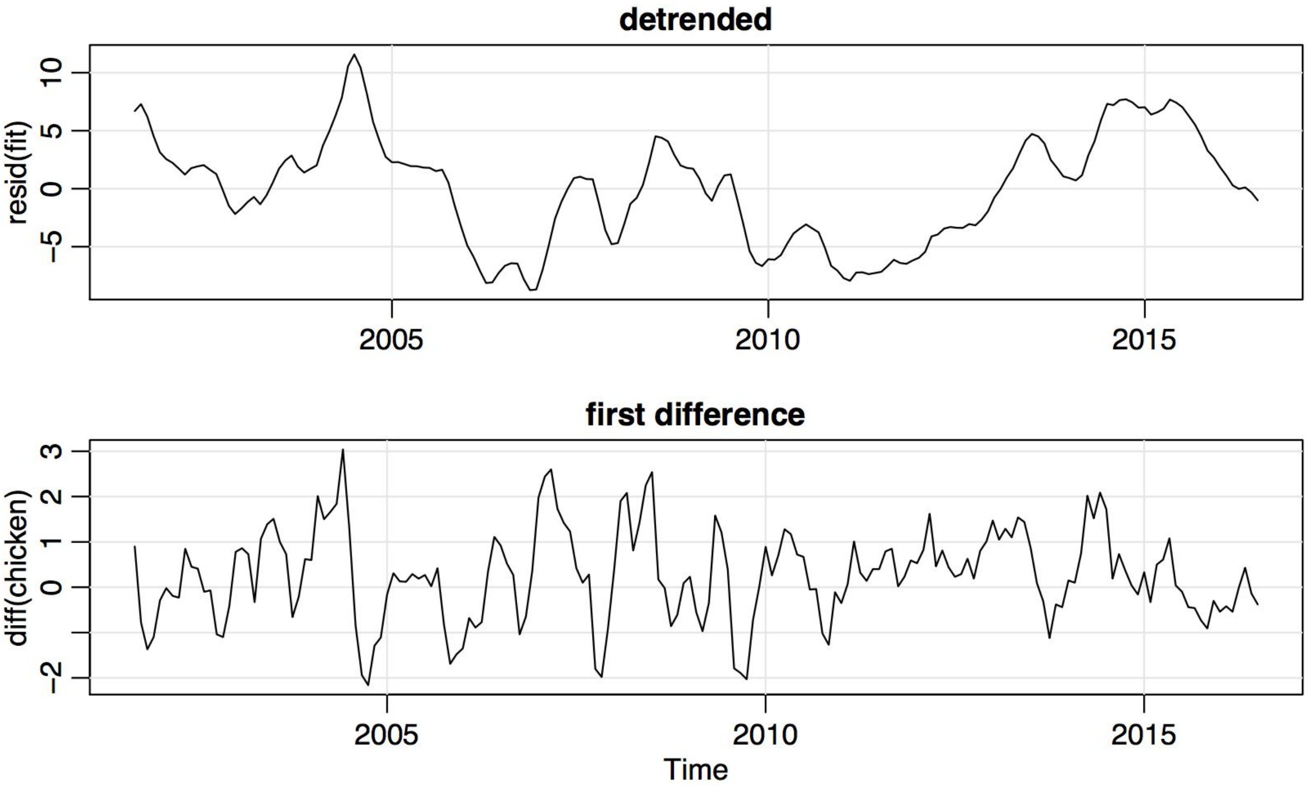


Fig. 2.1. *The price of chicken: monthly whole bird spot price, Georgia docks, US cents per pound, August 2001 to July 2016, with fitted linear trend line.*

Preprocessing



*
Fig. 2.4. *Detrended (top) and differenced (bottom) chicken price series. The original data are shown in Figure 2.1.*

Preprocessing

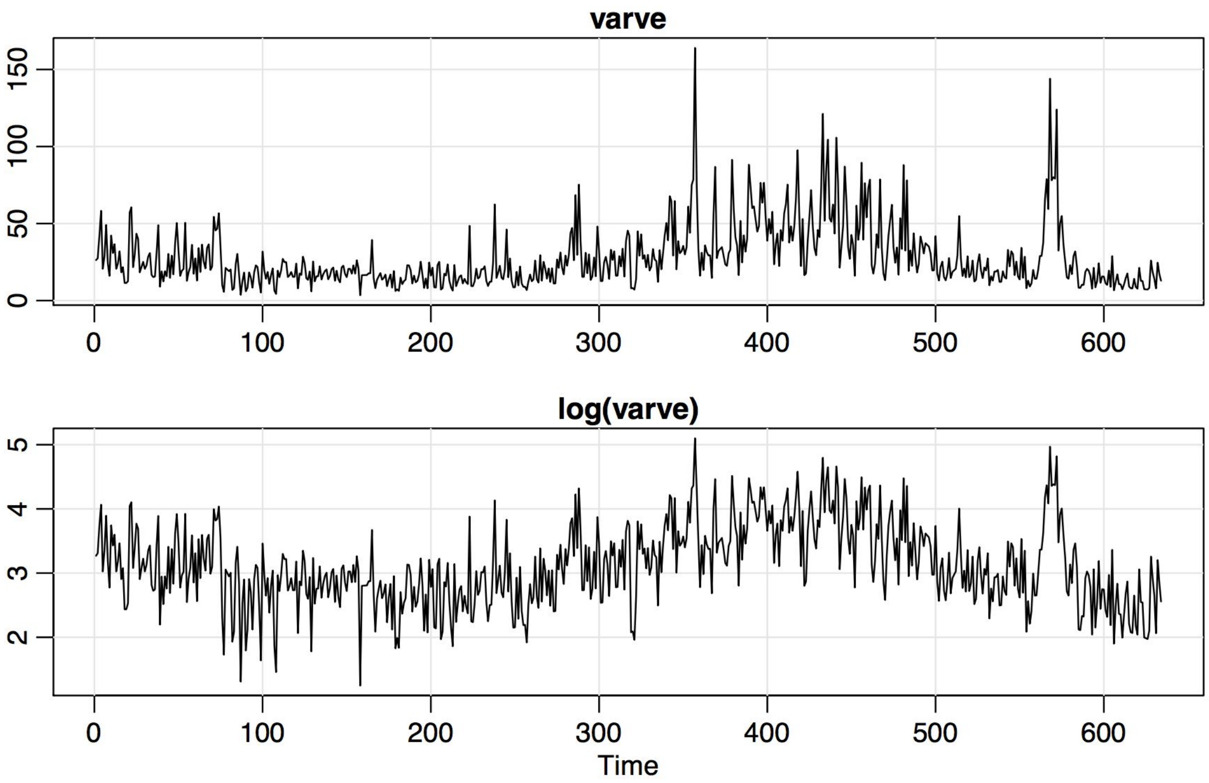


Fig. 2.7. *Glacial varve thicknesses (top) from Massachusetts for* $n = 634$ *years compared with* log *transformed thicknesses (bottom).*

Model Selection

Goal Because the famous Bias-variance trade off, we need a measure that takes both model performance and model complexity into account when selecting the hyper-parameters of a model.

Akaike Information Criterion

- $AIC = -2logL_k + 2k$ where L_k is the maximized likelihood.
- For normal regression problem, $AIC = log\hat{\sigma}_k^2 + \frac{n+2k}{n}$.
- $SSE = \sum_{t=1}^{n}(x_t - \hat{x}_t)^2$
- $\hat{\sigma}_k^2 = \frac{SSE(k)}{n}$, where n is the sample size and k is the number of parameters

Model Selection

AIC, Bias Corrected (AICc)

- When the sample size n is small, it's proved that AIC have biased for complex models and therefore is vulnerable to overfitting.
- $AICc = log\hat{\sigma}_k^2 + \frac{n+2k}{n-k-2}$

Bayesian Information Criterion

- $BIC = log\hat{\sigma}_k^2 + \frac{klogn}{n}$
- The penalty term in BIC is much larger than AIC. BIC comparably prefers simple model.

Seasonal ARIMA

$ARIMA(p, d, q) \times (P, D, Q)S$

- p = non-seasonal AR order
- d = non-seasonal differencing
- q = non-seasonal MA order
- P = seasonal AR order
- D = seasonal differencing
- Q = seasonal MA order
- S = time span of repeating seasonal pattern.
- Example $ARIMA(1, 0, 0) \times (2, 0, 0)12$, $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-12} + \phi_3 x_{t-24}$

Evaluation

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

Programming

- **Github:**
  - **https://github.com/charlieblue17/timeseries2018**
- **No submission required.**