IST772 Chapter Notes Template: After Completing Please Submit as a PDF.
Originality Assertion: By submitting this file you affirm that this writing is your own.

Name: Wanyue Xiao
Date: November 3, 2020
Chapter Number: #11
Title of Chapter: Analyzing Change over Time

Time Series Analysis

Cross-sectional data: data that collected in terms of many subjects (such as individuals, countries, regions) at the same one point or period of time.

Repeated Measures Analysis: subjects are measured at more than one specific point in time. Such a measure usually gets only a few different points in time.

- **Example:** repeated test on student who took an internship before and after.
- **Non-independence of observation:** the observations should be dependent with others. In such a condition, people can expect that those observations are corelated.
- t.test(X, Y, **paired = FALSE**) indicates that X and Y are independent. A narrower band means greater certainty about the result. If the lower bound is far from 0, this result suggests that the difference between X and Y is likely.
- The **df of the t.test** (having paired as True) equals to (the total number of df – 1(1 df lost for the calculation of mean of the difference score)).
- The **limitation of paired sample t-test** is that it is used to compare one time point to one other time point. Therefore, we **need to use the repeated measures ANOVA**. **For a balanced design, aov() works best. For unbalanced data, one can use ezANOVA.**

Repeated measures ANOVA: one might need to get a balanced data in order to run a aov() test.

- Specifically, one can use **error() function** in the aov() test to specify an individual difference as error variance. Therefore, one is being able to eliminate the confounding effects of the individual difference from the overall error term.

Time Series Analysis: usually measures a single subject or a small amount of closely related phenomena repeatedly at regular intervals over time.

- **Example:** price fluctuation of a stock market; measure one chicken's weight at 12 points in time.
- Repeated measure analysis takes the effects of individual differences into consideration while time series analysis examines and control for the effects of trends and cycles.
- Time series consists of four components:
    - **trend** – increase or decrees across time
    - **seasonality** – regular fluctuations that occur frequently across a period of time
    - **cyclicality** – *cycle* occurs when the data exhibit rises and falls that are not of a fixed frequency. One example is the recession.
    - **Irregular component** – noise
- One can use decompose(ts(X, frequency =)) to decompose time series data.
- **Stationarity**: means that the statistical properties (mean, variance, autocorrelation, etc.) of a process generating a time series do not change over time.
- **Autocorrelation** function: correlates a variable with itself at a later time period.
    - ACF plot: show serial correlation in data that changes over time. The Pearson's correlation coefficient is a number between -1 and 1 that describes a negative or

positive correlation respectively. A value of zero indicates no correlation. The height of each line is the correlation of the original variable correlated with itself at different amounts of lag. **The first line always equals to 1.0 since it will be correlated perfectly with itself at zero lags.** The horizontal line is the statistical significance for correlation, regardless of positive one or negative one.
    - A stationary time series plot's lagged correlation should be non-significant as well as no specific pattern or variance between positive and negative correlations.
    - **Whiting**: the process of removing seasonality and trend from a time series. If one examines the ACF plot of noise, strong pattern of significant multiple autocorrelations should be avoided in the result.
    - **The adf.test()** function from tseries package is able for the detection of stationary process. The alternative hypothesis is that the process is stationary.

Differencing: a process that remove the rend. One can use the diff() function to do that.
Heteroscedasticity: big difference in variability from the early stage compared to the later stage.

Change Point Analysis: to detect whether any changes have occurred. It determines the number of changes and estimates the time of each change by examining the mean of time series.

- The cpt.mean() function is available from changepoint package. The penalty is used to measure how sensitive the algorithm is to the change. No penalty means strictest way.
- **Removing trends is not required in this analysis.**
- **Probability in change-point analysis:** use cpt.mean() to get the confidence level which represents he "strength of belief". The closer the value approaches to 1, the better the result (change in the mean level is not due to chance).

ARIMA and exANOVA

More information about ezANOVA, please see the R Code Fragment and Explanation section.
ARIMA:
Auto-Regressive, Integrated, Moving Average, usually abbreviated as ARIMA. Any 'non-seasonal' time series that exhibits patterns and is not a random white noise can be modeled with ARIMA models.

- An ARIMA model is characterized by 3 terms: p, d, q
    - p is the order of the AR term
    - q is the order of the MA term
    - d is the number of differencing required to make the time series stationary
- For the purpose of. Making a series stationary, the most common approach is to difference it. That is, subtract the previous value from the current value. Sometimes, depending on the complexity of the series, more than one differencing may be needed.

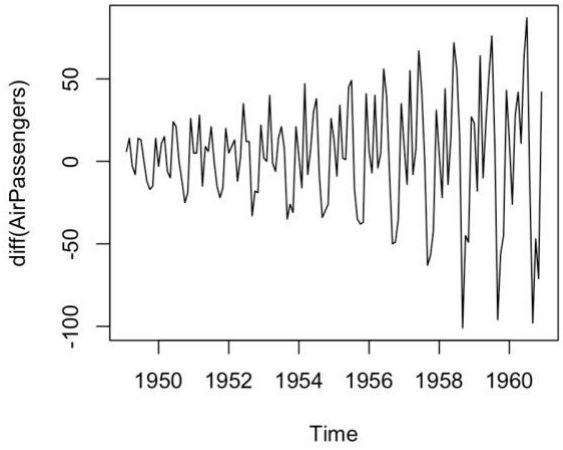
Exercise Review

2.
The nlme package is not available for current version.

IST772 Chapter Notes Template: After Completing Please Submit as a PDF.
Originality Assertion: By submitting this file you affirm that this writing is your own.

```
> library(nlme)
Warning message:
package 'nlme' was built under R version 3.6.2
> data()
> data(Blackmore)
Warning message:
In data(Blackmore) : data set 'Blackmore' not found
```

5.
plot(diff(AirPassengers))

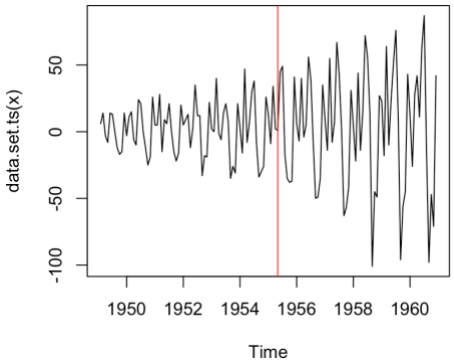


The variance or the fluctuation increase gradually across the time.

```
> cpt.var(diff(AirPassengers))
Class 'cpt' : Changepoint Object
       ~~   : S4 class containing 12 slots with names
              cpttype date version data.set method test.stat pen.type pen.value minseglen cpts
 ncpts.max param.est

Created on  : Fri Apr 26 15:58:35 2019

summary(.)  :
----------
Created Using changepoint version 2.2.2
Changepoint type      : Change in variance
Method of analysis    : AMOC
Test Statistic  : Normal
Type of penalty       : MBIC with value, 14.88853
Minimum Segment Length : 2
Maximum no. of cpts   : 1
Changepoint Locations : 76
```



The major change point occurs at 76, which seems to be between 1955 and 1956.

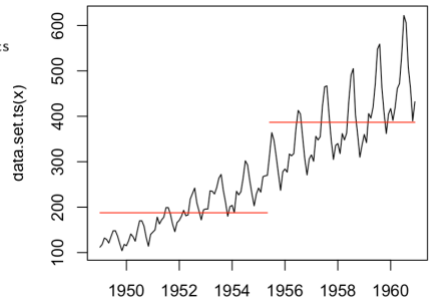
6.

IST772 Chapter Notes Template: After Completing Please Submit as a PDF.
Originality Assertion: By submitting this file you affirm that this writing is your own.

```
> cpt.mean(AirPassengers)
Class 'cpt' : Changepoint Object
      ~~   : S4 class containing 12 slots with names
             cpttype date version data.set method test.stat pen.type pen.value minseglen cpts
 ncpts.max param.est

Created on  : Fri Apr 26 15:58:35 2019

summary(.)  :
----------
Created Using changepoint version 2.2.2
Changepoint type      : Change in mean
Method of analysis    : AMOC
Test Statistic  : Normal
Type of penalty       : MBIC with value, 14.90944
Minimum Segment Length : 1
Maximum no. of cpts   : 1
Changepoint Locations : 77
```



Similarly, the change point is 77.

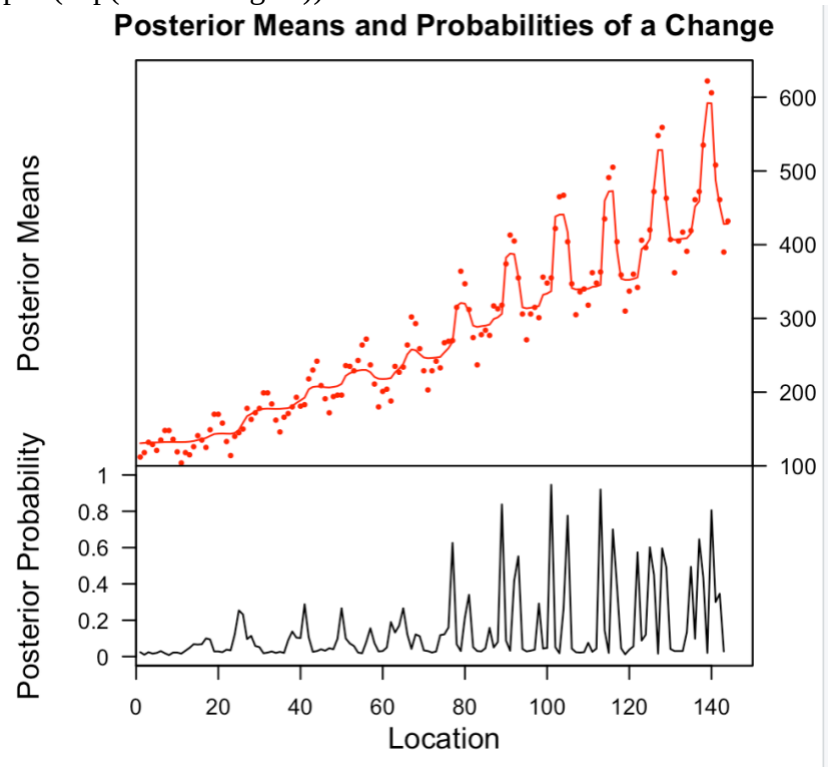
Question: **Plot and interpret the results. Com- pare the change point of the mean that you uncovered in this case to the change point in the variance that you uncovered in Exercise 5. What do these change points suggest about the history of air travel?**

7.

According to the changepoint discovered in question 5, 1955 seemed to be a breaking point for the airplane. More and more passages chose airplane as transportation mean while the fluctuation / variance of passages also increased gradually.

8.

plot(bcp(AirPassengers))



The lower section has greater variance in the latter stage, especially after location 77 which is consistent with the result of question 6.

R Code Fragment and Explanation

Code: ezANOVA(data, dv, wid, within, detailed)

IST772 Chapter Notes Template: After Completing Please Submit as a PDF.
Originality Assertion: By submitting this file you affirm that this writing is your own.

ezANOVA: analysis of data from factorial experiments, including purely within-Ss designs (a.k.a. "repeated measures"), purely between-Ss designs, and mixed within-and-between-Ss designs, yielding ANOVA results, generalized effect sizes and assumption checks.

- dv: contains the dependent variable, values must be numeric.
- wid: contains the variable specifying the case/Ss identifier.
- within: contain predictor variable that are manipulated within-Ss.
- detailed: If TRUE, return extra information about ANOVA.

Example: the example listed below is extracted from the textbook ("Reasoning with Data"), chapter 11.

```
ezANOVA(data=chwBal,dv=.(weight),within=.(TimeFact),wid=.(Chick),detailed=TRUE)
$ANOVA
       Effect DFn DFd     SSn      SSd        F             p p<.05       ges
1 (Intercept)   1  44 8431251 429898.6 862.9362  1.504306e-30     * 0.9126857
2    TimeFact  11 484 1982388 376697.6 231.5519 7.554752e-185     * 0.7107921

$`Mauchly's Test for Sphericity`
    Effect            W             p p<.05
2 TimeFact 1.496988e-17 2.370272e-280     *

$`Sphericity Corrections`
    Effect       GGe        p[GG] p[GG]<.05       HFe       p[HF] p[HF]<.05
2 TimeFact 0.1110457 7.816387e-23         * 0.1125621 4.12225e-23         *
```

Explanation of Result:

- DFn: degrees of freedom in the numerator
- DFd: df in the denominator
- SSn: sum of squares in the numerator
- SSd: sum of squares in the denominator
- ges: generalized Eta-Squared measure of effect size, **which represent the proportion of variance that is accounted for by the predictor.**
- W: Mauchly's W statistic

1. Given that the p-vlaue of TimeFact is lower than 0.05, one can reject that hypothesis that no change in weight over time.
2. The next section is used to test the homogeneity of variance. If the p-value is lower than 0.05, it represents a violation of such an assumption. Specifically, in this case, the range weight of chicken is narrower for younger chicken but increase gradually in each time group. Therefore, one can speculate that the F-test is incorrect.
3. The third section provides test on the violation of test of sphericity. GGe is more conservative than HFe. Since both the corrections show the same result (significant), then one can conclude that any inflation in the F-ratio has not adversely affected the decision to reject the null hypothesis.

Question for Class

1. Is it possible that the result of t.test and that of BESTmcmc are contradictory?
2. How to interpret the seasonal component of a time series, especially for the cycle and lag?

IST772 Chapter Notes Template: After Completing Please Submit as a PDF.
Originality Assertion: By submitting this file you affirm that this writing is your own.