**Lecture 7 Trend Estimation and Residual Analysis**

**7.1 Estimating deterministic trends by regression methods**

A trendin a time series may be loosely defined as “long-term change in the mean level”. Deterministic trends may well be characterized by deterministic functions, while stochastic trends may be represented by stochastic models. In this talk, we consider fitting the deterministic trends by the regression method. A variety of stochastic models will be introduced and studied in Lectures 8 and 9.

We will use the ordinary least-squares method (see Lecture 5), which assumes independent observations, to estimate the trend of a process, although we recognize that, in general, the observations are not independent. There are two justifications for this shortcut: (1) for large sample size, the effect of the simple least-squares estimation is similar with the effect of the ***generalized least-squares method***, which assumes dependent observations, for the types of trends to be considered; (2) the unknown correlation (dependence) that is not detected by the simple least-squares method may be retained in the stochastic trend for further analysis.

**Example 7.1**  The data set *uspop* in {datasets} gives the population of the United States (in millions) as recorded by the decennial census for the period 1790–1970. The time series plot suggests a possible quadratic trend.

|  |
| --- |
| plot.ts(uspop,type='b') |
|  |
| popyear1<-c(1:19)  popyear2<-c(1:19)^2  pop.fit<-lm(uspop~popyear1+popyear2)  summary(pop.fit) |
| Call:  lm(formula = uspop ~ popyear1 + popyear2)  Residuals:  Min 1Q Median 3Q Max  -6.5997 -0.7105 0.2669 1.4065 3.9879  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 6.30914 2.13387 2.957 0.00928 \*\*  popyear1 -1.90193 0.49132 -3.871 0.00135 \*\*  popyear2 0.63446 0.02387 26.584 1.14e-14 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 2.78 on 16 degrees of freedom  Multiple R-squared: 0.9983, Adjusted R-squared: 0.9981  F-statistic: 4645 on 2 and 16 DF, p-value: < 2.2e-16 |
| plot.ts(pop.fit$fitted.values,type='b') |
|  |
| plot.ts(pop.fit$residuals,type='b') |
|  |

The quadratic function fits the data well (the predictors are highly significant and is very high). However, there is a certain pattern in the residuals. The regression equation is given by

Fitted *uspop* = 6.31 – 1.90 *t* + 0.63 , *t* = 1, 2, … █

**Example 7.2**  Re Example 6.2., the average monthly temperatures in a city over 12 years, the plot shows clearly a seasonal trend with 12 “seasons”. Let  be the temperature series. Then

 =  + .

The seasonal trend  is estimated by the regression method and the regression equation is given by

 =  = - 8.55 Jan - 6.31 Feb + 0.264 Mar + 8.07 Apr

+ 14.5 May + 19.7 Jun + 22.1 Jul + 20.7 Aug

+ 16.1 Sep + 10.5 Oct + 2.58 Nov - 4.64 Dec

|  |
| --- |
| avtemp.fit<-lm(AvTemperature~0+factor(Months)) |
| Call:  lm(formula = AvTemperature ~ 0 + factor(Months))  Residuals:  Min 1Q Median 3Q Max  -4.5950 -1.2490 0.0642 1.0513 5.4550  Coefficients:  Estimate Std. Error t value Pr(>|t|)  factor(Months)1 -8.5508 0.5482 -15.597 < 2e-16 \*\*\*  factor(Months)2 -6.3058 0.5482 -11.502 < 2e-16 \*\*\*  factor(Months)3 0.2650 0.5482 0.483 0.63  factor(Months)4 8.0700 0.5482 14.720 < 2e-16 \*\*\*  factor(Months)5 14.4958 0.5482 26.441 < 2e-16 \*\*\*  factor(Months)6 19.7233 0.5482 35.976 < 2e-16 \*\*\*  factor(Months)7 22.0650 0.5482 40.248 < 2e-16 \*\*\*  factor(Months)8 20.7400 0.5482 37.831 < 2e-16 \*\*\*  factor(Months)9 16.1250 0.5482 29.413 < 2e-16 \*\*\*  factor(Months)10 10.5417 0.5482 19.229 < 2e-16 \*\*\*  factor(Months)11 2.5833 0.5482 4.712 6.13e-06 \*\*\*  factor(Months)12 -4.6433 0.5482 -8.470 4.18e-14 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 1.899 on 132 degrees of freedom  Multiple R-squared: 0.9814, Adjusted R-squared: 0.9797  F-statistic: 580.5 on 12 and 132 DF, p-value: < 2.2e-16 |
| plot.ts(avtemp.fit$fitted.values,type='b') |
|  |
| plot.ts(avtemp.fit$residuals,type='b') |
|  |

The seasonal means function fits the data well with large value of . Note that the residual plot shows a rectangular shape. █

**Example 7.3** Re: data file *MaunaLoaCO2.csv*, atmospheric CO2 concentrations (ppmv) collected at Mauna Loa Observatory, Hawaii, years 1990-2003 (data source: *co2* in {datasets}), an upward linear trend and a seasonal trend are apparent in the time series plot below.

|  |
| --- |
| plot.ts(MaunaLoaCO2$co2.14) |
|  |

**Estimating the trends:** Let  be the CO2 concentration,  =  +  + . The regression equation is given by

 = 

|  |
| --- |
| co2.14.fit<-lm(co2.14~0+time+factor(month.14))  summary(co2.14.fit) |
| Call:  lm(formula = co2.14 ~ 0 + time + factor(month.14))  Residuals:  Min 1Q Median 3Q Max  -1.3685 -0.6198 -0.1270 0.5914 1.7524  Coefficients:  Estimate Std. Error t value Pr(>|t|)  time 1.384e-01 1.221e-03 113.4 <2e-16 \*\*\*  factor(month.14)1 3.523e+02 2.263e-01 1556.9 <2e-16 \*\*\*  factor(month.14)2 3.529e+02 2.268e-01 1556.3 <2e-16 \*\*\*  factor(month.14)3 3.537e+02 2.273e-01 1555.9 <2e-16 \*\*\*  factor(month.14)4 3.548e+02 2.278e-01 1557.2 <2e-16 \*\*\*  factor(month.14)5 3.552e+02 2.284e-01 1555.5 <2e-16 \*\*\*  factor(month.14)6 3.545e+02 2.289e-01 1548.6 <2e-16 \*\*\*  factor(month.14)7 3.528e+02 2.295e-01 1537.6 <2e-16 \*\*\*  factor(month.14)8 3.506e+02 2.300e-01 1524.1 <2e-16 \*\*\*  factor(month.14)9 3.487e+02 2.306e-01 1512.2 <2e-16 \*\*\*  factor(month.14)10 3.487e+02 2.312e-01 1508.5 <2e-16 \*\*\*  factor(month.14)11 3.500e+02 2.317e-01 1510.6 <2e-16 \*\*\*  factor(month.14)12 3.513e+02 2.323e-01 1512.4 <2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.7657 on 155 degrees of freedom  Multiple R-squared: 1, Adjusted R-squared: 1  F-statistic: 2.919e+06 on 13 and 155 DF, p-value: < 2.2e-16 |
| plot.ts(co2.14.fit$fitted.values) |
|  |
| plot.ts(co2.14.fit$residuals) |
|  |

The linear and seasonal functions fit the data well with near perfect  value. However, the residual plot shows a certain drift. █

**7.2 Residual Analysis**

In time series modelling it is most important to examine the independence of the residuals (see the comments at the beginning of Lecture 6). The model diagnostics presented in Lectures 4 and 5 are also necessary and useful in model validation.

The residual for the additive time series model is defined as

 =  - =  - ( +  + ), t = 1, 2, …, n.

For a given sequence of residuals,, , …, , which is a time series, the sample *acf* is given by

 = , k = 0, 1, 2, …

If the residuals are independent, the theoretical *acf* (see Section 6.3) is zero for all k > 0 and thus the values of the sample *acf* should be close to zero at all lags.

**Example 7.4** Re example 7.1, the *acf* plot of the residuals is shown below, where the dotted lines are the lower and upper confidence limits for the sample *acf*. The sample *acf* seems to be large at lags 2 and 3, though just within the limits. The diagnostics reveal some certain pattern in the residuals.

|  |
| --- |
| acf(pop.fit$residuals) |
|  |
| plot(pop.fit) |
|  |

A note on calculation of the *acf* for short time series: The formula for the sample *acf* is different from the formula for the usual sample correlation (see Lecture 4). When the sample sizes are large, the results from the two formulas are almost identical. When the sample sizes are small, however, the results may vary. In this time series, the entire length is 19. Below are the *acf* at lags 1, 2 and 3 and related tests obtained by *cor.test*, which indicate that the autocorrelation within the residuals is highly significant.

|  |
| --- |
| cor.test(pop.fit$residuals[1:18], pop.fit$residuals[2:19]) |
| Pearson's product-moment correlation  data: pop.fit$residuals[1:18] and pop.fit$residuals[2:19]  t = 1.3736, df = 16, p-value = 0.1885  alternative hypothesis: true correlation is not equal to 0  95 percent confidence interval:  -0.1674739 0.6874235  sample estimates: cor 0.3247911 |
| cor.test(pop.fit$residuals[1:17], pop.fit$residuals[3:19]) |
| Pearson's product-moment correlation  data: pop.fit$residuals[1:17] and pop.fit$residuals[3:19]  t = -1.8952, df = 15, p-value = 0.0775  alternative hypothesis: true correlation is not equal to 0  95 percent confidence interval:  -0.7596893 0.0521175  sample estimates: cor -0.4395377 |
| cor.test(pop.fit$residuals[1:16], pop.fit$residuals[4:19]) |
| Pearson's product-moment correlation  data: pop.fit$residuals[1:16] and pop.fit$residuals[4:19]  t = -2.7351, df = 14, p-value = 0.01611  alternative hypothesis: true correlation is not equal to 0  95 percent confidence interval:  -0.8400863 -0.1334730  sample estimates: cor -0.5901332 |

**█**

**Example 7.5** Re example 7.3, the *acf* plot shows no evidence of autocorrelation in the residuals and the diagnostics expose no violations of the model assumptions.

|  |
| --- |
| acf(avtemp.fit$residuals) |
|  |
| plot(avtemp.fit) |
|  |

**█**

**Example 7.6** Re example 7.3, it is clear that the sample *acf* for the residuals is highly significant at various lags based on the *acf* plot below and the residuals are auto-correlated, although the model fits well to the data with near perfect value of. The diagnostics also indicate some certain pattern in the residuals, which may not be normally distributed.

|  |
| --- |
| acf(co2.14.fit$residuals) |
|  |
| plot(co2.14.fit) |
|  |

█

**Exercises**

7.1 Re-do the examples in this talk.

7.2 Re the data file *AirPollution.Summer.1994.csv*, there are 120 observations on the following 5 variables.

*O3*  - Daily maximum ozone in parts per billion.

*NO2*  - Daily maximum NO2 in parts per billion.

*NO*  - Daily maximum NO in parts per billion.

*SO2*  - Daily maximum SO2 in parts per billion.

*PM10* - Daily maximum PM10 in micrograms/metre^3

1. Use *O3* as the response variable to establish a linear regression model against the four predictors.
2. Find the sample *acf* for the residuals. Are the residuals auto-correlated?
3. Perform model diagnostics for the regression model and state your results.

In this exercise we included variables in the model that vary with time without explicitly including time as a predictor variable.

**References**

* Chatfield, C. (2004), *The Analysis of Time Series: An Introduction*, CHAPMAN & HALL.
* Cryer, J. D. (1986), *Time Series Analysis*, PWS-KENT.
* Millard, S.P. and Neerchal, N. K. (2000), *Environmental Statistics with S-PLUS*, Chapman & Hall.