Handout 4

Time series with trend (From Sections 1.32,1.3.3 and 1.5 in the text)

Consider the temperature data (years 1850-2012) from Handout 1. The temperature series can be described as

$$obs = smooth + rough,$$

where the smooth part is called the trend and the rough part fluctuates about zero. For notational convenience, we will write year 1850 as time 1 and year 2012 as time 163. Mathematically, speaking if we call the observed average temperature at time t as Y_t, the trend as m_t and the rough part as X_t, then we can write

$$Y_t = m_t + X_t,\ t = 1, ..., n = 163,$$

where the year 1850 is time 1 and the year 2012 is time 163. [**Note**: in the text, the temperature is denoted by X_t and the rough part is denoted by Y_t. Here the notations are exactly the opposite.]

You will find a graph which presents, the trend part along with the data and a plot with rough part. We will talk about methods for estimating the trend later. Note that if we want to guess (forecast) temperature for time 164 (i.e., year 2013), then we will do have to obtain estimates of the trend and the rough parts at time $t = 164$. In such a case, predicted temperature $\hat{Y}_{164}$ at time $t = 164$ will be

$$\hat{Y}_{164} = \hat{m}_{164} + \hat{X}_{164},$$

where $\hat{m}_{164}$ and $\hat{X}_{164}$ are the estimated trend and the rough at time $t = 164$.
Similarly, if we want to forecast (guess) Y_{n+h} (say $h = 2$, then $n + h = 165$), then the guessed value is

$$\hat{Y}_{n+h} = \hat{m}_{n+h} + \hat{X}_{n+h},$$

where $\hat{m}_{n+h}$ and $\hat{X}_{n+h}$ are the guessed values of m_{n+h} and X_{n+h}. Obtaining an estimate of m_{n+h} is not difficult. However, obtaining an estimate of X_{n+h} requires modeling which we will discuss in a few weeks. It is important to note that in order to guess m_{n+h} and X_{n+h}, we will need to obtain estimates of the entire trend $\{m_t : t = 1, ..., n\}$ and the entire rough $\{X_t : t = 1, ..., n\}$. This note is only concerned with methods for obtaining to estimates of the trend and the rough parts.

There are many methods for obtaining the trend, but we will focus on only three of them:
a) polynomial fit, b) loess, c) two-sided moving average.

Estimate of rough

Note that the rough $X_t = Y_t - m_t$. Hence we can obtain estimate of the rough by subtracting $\hat{m}_t$ from Y_t.

Polynomial trend:

This is a simple method. For instance if we want to obtain a linear trend (polynomial of degree 1), simply carry out a regression of temperature on time (dependent variable is Y_t and independent variable is t), i.e., fit the model

$$Y_t = \beta_0 + \beta_1 t + \varepsilon_t.$$

Note that the trend is being modeled as $m_t = \beta_0 + \beta_1 t$. Once you obtain the least squares estimate of β_0 and β_1, then the estimated trend is

$$\hat{m}_t = \hat{\beta}_0 + \hat{\beta}_1 t.$$

Similarly if we want to model the trend by a quadratic, i.e., $m_t = \beta_0 + \beta_1 t + \beta_2 t^2$, we fit a quadratic regression model

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t.$$

Then the estimated trend is

$$\hat{m}_t = \hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_2 t^2.$$

In general, if we can fit a polynomial of degree p

$$Y_t = \beta_0 + \beta_1 t + \cdots + \beta_p t^p + \varepsilon_t,$$

and, if $\hat{\beta}_0, \hat{\beta}_1, ...$ are the least squares estimates of $\beta_0, \beta_1, ...$, then the estimated trend is

$$\hat{m}_t = \hat{\beta}_0 + \hat{\beta}_1 t + \cdots + \hat{\beta}_p t^p.$$

The choice of the degree of polynomial is very important and there are methods for selecting the degree. But we will ignore that issue at present.

Loess

Loess (locally weighted polynomial regression) is a popular method for obtaining the trend. The method is basically this. At every time point t, fit a linear regression of X on time for the time points in a window of q i.e., regress Y_s on s when $s = t - q, ..., t + q$, and then fitted value of the estimated regression at t is $\hat{m}_t$. This method works quite well. However, as in the first two cases, the choice of q is important and there are methods which can guide the choice of q.

In the packages, you need to specify "span" instead of q. Span is the proportion of data being used for each locan linear regression. Relation between span and q is:

$$span = (2q + 1)/n.$$

Note: The packages carry out a local regression with unequal weights at each point t and that is the default. Usual regression method uses equal weights when carrying out a local regression and this may produce an unsmooth trend. The default option in the package is to use unequal weight resulting in a smooth trend estimate.

Two-sided moving average.

In this method, estimate of m_t at time t is the average of of $2q+1$ observatons $Y_{t-q}, Y_{t-q+1}, ..., Y_{t+q}$, i.e.,

$$\hat{m}_t = \sum_{j=-q}^{q} Y_{t+j}/(2q+1).$$

Note that in order to be able to obtain an estimate of m_t at time t, you will need to have q time points to the left and to the right of t. This is not possible when t is beween 1 and q or t is between $n-q+1$ and n. One can use the average of available observations to the left and to the right of t when $t = 1, .., q$ or when $t = n-q+1, ..., n$, but those estimates are not usually very good. For this reason, in this method, estimate of m_t is obtained for $t = q+1, ..., n-q$, and no estimate is usually provided for $\{m_1, , , , m_q\}$ and $\{m_{n-q+1}, ..., m_n\}$.

Temperature data

We have fitted a polynomial of degree 6 and a loess with $q = 20$ (i.e., $span = (2q+1)/n = 41/163 \approx 0.25$). Note that the two fits are somewhat different.

R Commands:

Suppose the temperature data is stored as y. Create a new variable 'tm' for time.

> tm=1:163

Run a polynomial fit of degree 6 and a loess with span=0.25. Note that the command 'lm' is for linear model. All regressions are done by this command.

> polytrnd=lm(y∼poly(tm,6))

> loesstrnd=loess(y∼tm, span=0.25)

Note that fitted ploynomial trend is in polytrnd$fit and the fitted loess trend is in loesstrnd$fit. The residuals are in polytrnd$resid and loesstrnd$resid, respectively. The residuals are estimates of the rough $\{X_t\}$.

First a plot of the temperature data with polynomial and loess fits. There are many ways to do this. Here is one such way.

> plot(tm, y, type='l', lty=1, xlab="Time", ylab="Temp", main="Teperature series with poly and loess trends")

> points(tm, polytrnd$fit, type='l', lty=1)

> points(tm, loesstrnd$fit, type='l', lty=2)

> legend(1,0.4, c("temp","polyfit","loess"), lty=c(1,1,2))

Note: 'xlab' and 'ylab': labels on the x and the y axes.

'main': title of the graph.

'type': different types - line (l), circle (o), ...(here we have chosen 'l').

'lty' is linetype. solid=1, dotted=2,.....

'legend': legends for the various components of the plot. We have placed the legend at the location (1,0.4) (x=1,y=0.4).

We will now create multiple plots in one page. We will plot estimates X_t against time for polynomial and loess models, obtain histograms. First you need a command to be able to create four subplots and then use the commands for plotting.

```
> par(mfrow=c(2,2))
> plot(tm, polytrnd$resid, type='l', ylim=c(-0.15,0.35), xlab="Time", ylab="Residual", main="Polyfit: residual")
> plot(tm, loesstrnd$resid, type='l', xlab="Time", ylab="Residual", main="Loess: residual")
> hist(polytrnd$resid, freq=FALSE,xlim=c(-0.15,0.35), ylim=c(0,4.5), xlab="Residual", main="Histogram: polyfit resid")
> hist(loesstrnd$resid, freq=FALSE, xlim=c(-0.15,0.35), ylim=c(0,4.5), xlab="Residual", main="Histogram: loess resid")
```

Note: Normally, the plotter in R will decide the axes on its own unless you specify them using 'xlim' and 'ylim'. It is useful when comapring several plots. Normally, the vertical axis in histogram displays the 'frequency' (count). In order to override it we have used 'freq=FALSE'. This will force the vertical axis of the histogram to be density.

Figure 1: Temperature Series (1850-2012)

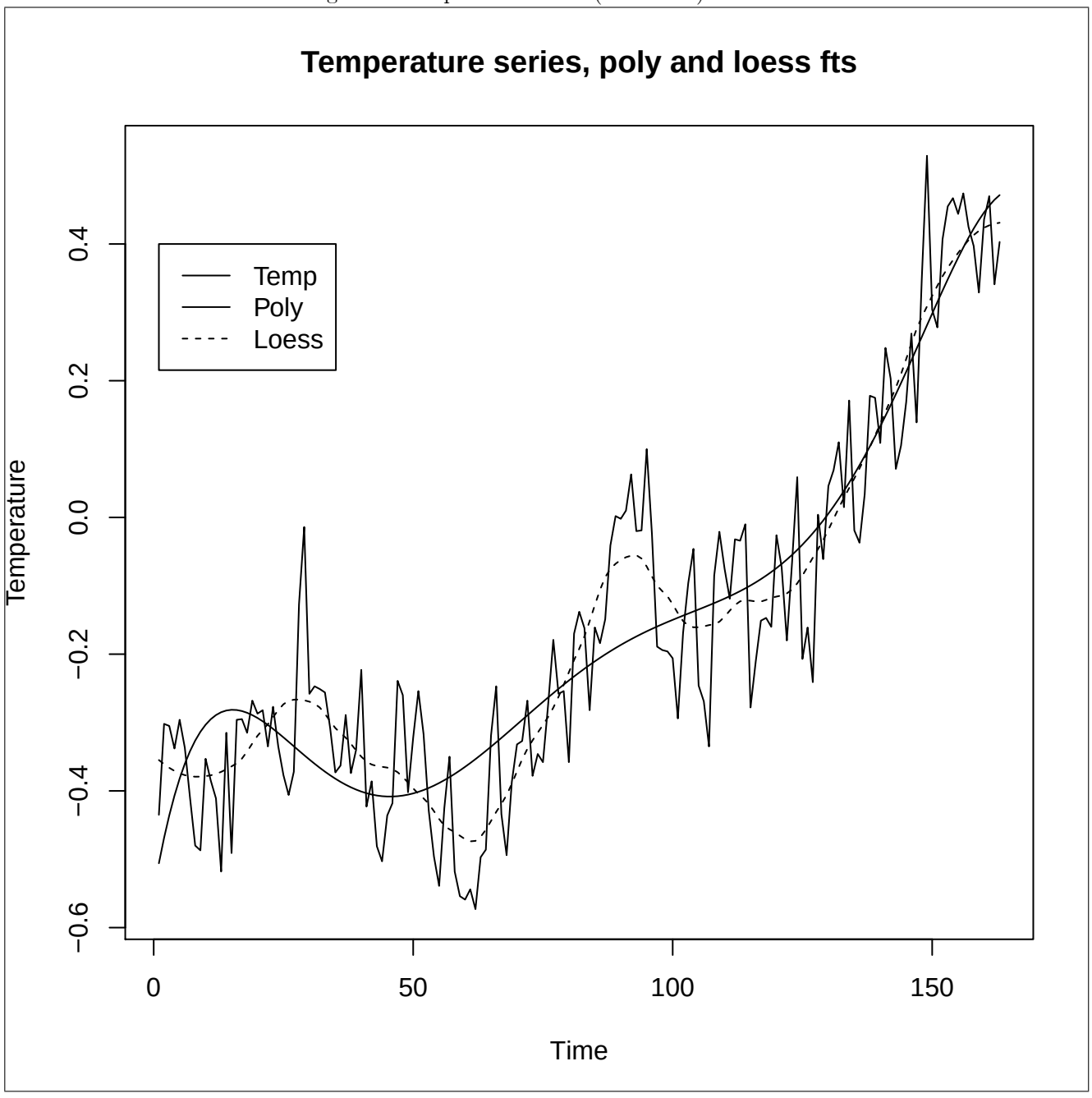

Figure 2: Temperature Series (1850-2012)