Analysis of Global Plastics Production Data
SCIE1500

Atakelty Hailu

February 20, 2022

INSTRUCTIONS

Read in the data on global plastics production, visualise and explore it, and then determine what type of equation or mathematical model would describe the relationship between production and time. The exercises can also be done in Excel or other software (e.g. Python).

1. Step 1: read in downloaded data and explore

   Read the global plastics production data from: `https://ourworldindata.org/plastic-pollution?utm_source=newsletter`

```
gpp = read.csv("global-plastics-production.csv")
#head(gpp)
summary(gpp[,4])

##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   2000000  20750000  76500000 118530303 198500000 381000000
```

   What is the min, max, mean of gpp?

   What are the log values of these statistics (min, max, mean)?

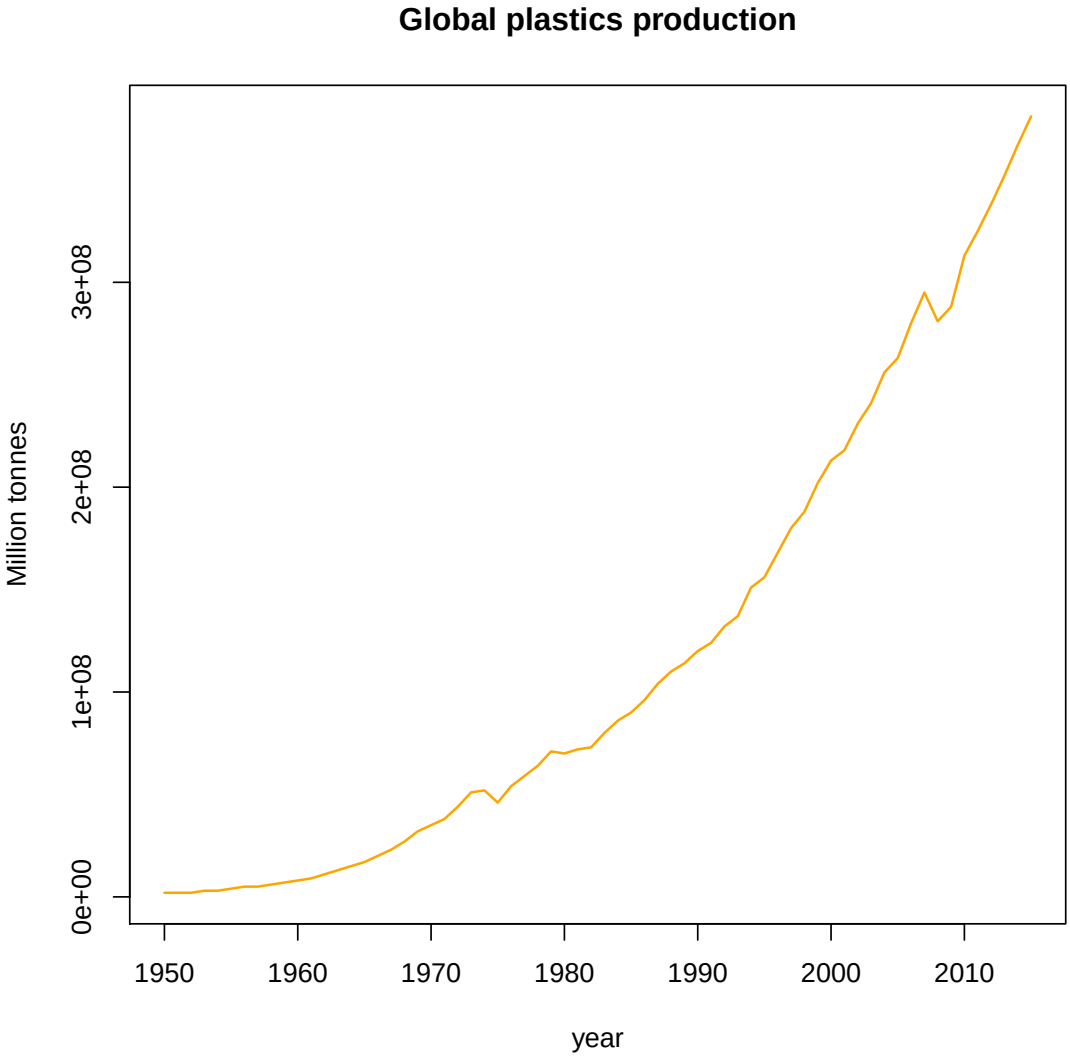
2. Step 2: extract variables and use shorter simpler names

   Also, create a time trend based on year (for easier modelling) as it is the passage of time rather than the actual calendar year that we are interested in. t is also smaller than calendar year.

```
#time trend
t = gpp$Year - 1949
#plastics production
ppd.million.tonnes = gpp$Global.plastics.production..million.tonnes.
```

3. Step 3: visualise data

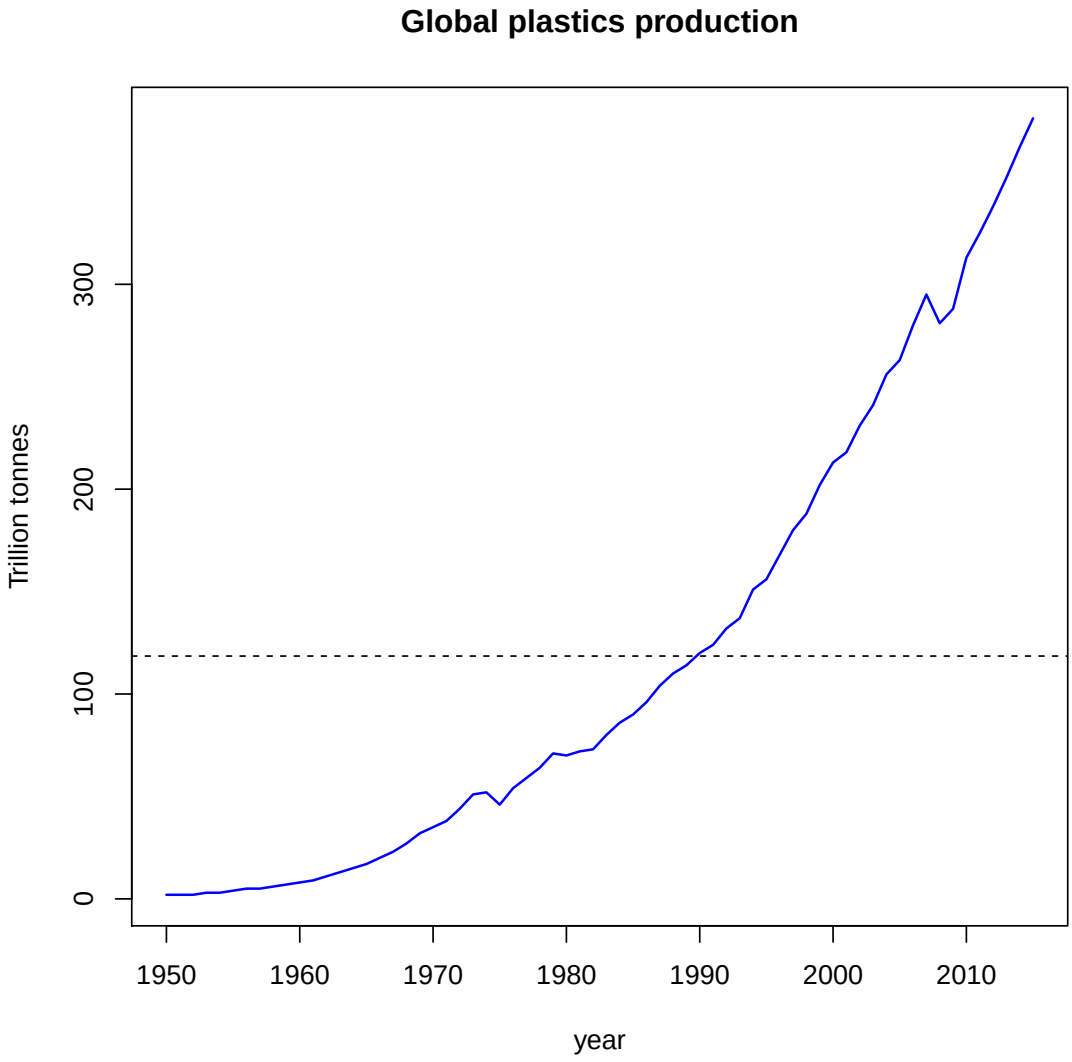
```
#plot data
plot(t+1949, ppd.million.tonnes,
     xlab="year", ylab="Million tonnes",
     main ="Global plastics production",
     type="l", col = "orange", lwd=1.5)
```



4. Step 4: Explore effect of scaling and transformation (logs)

   Plastic production in trillion tonnes

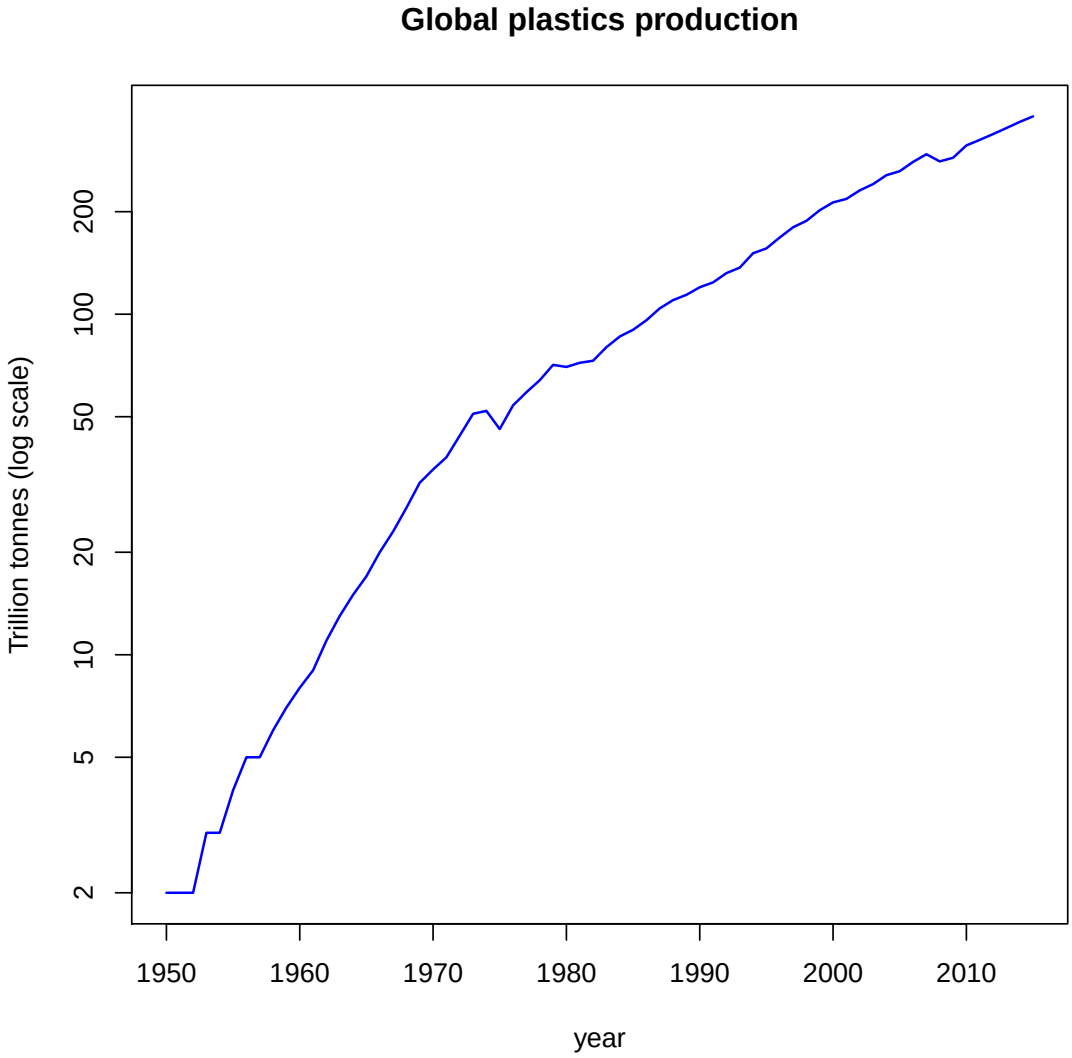
```
ppd.trillion.tonnes = ppd.million.tonnes/1000000
plot(t+1949, ppd.trillion.tonnes,
     xlab="year", ylab="Trillion tonnes",
     main ="Global plastics production",
     type="l", col = "blue", lwd=1.5)
#draw a horizontal line at the mean value
abline(h=mean(ppd.trillion.tonnes), lty=2)
```



Why do think we need to scale data?

Let's now use a log scale for the y-axis, where equal distances (50 to 100, and 100 to 200) represent a doubling of the value, rather than the addition of the same amount.

```
plot(t+1949, ppd.trillion.tonnes,
     xlab="year", ylab="Trillion tonnes (log scale)",
     main ="Global plastics production",
     type="l", col = "blue", lwd=1.5,
     log="y")
```



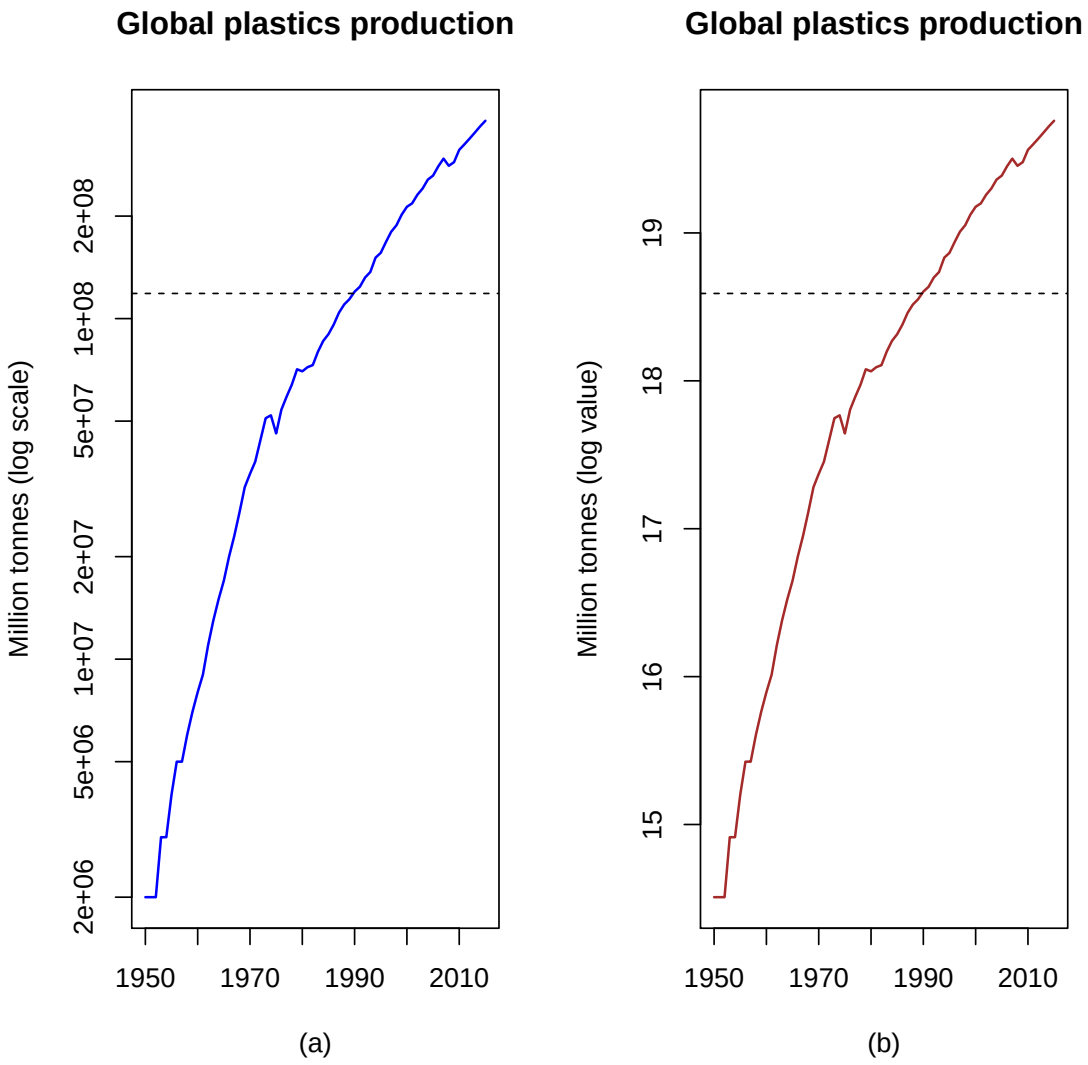
Let's create logged value for the production data.

```
#log transform: plastics production in log
log.ppd.million.tonnes = log(ppd.million.tonnes)
```

Now, we have two options. We could plot the raw data (million tonnes) and use a log scale on the y-axis, or we could actually plot the logged value of the data (log.ppd.million.tonnes). Compare the effects of using log scale in plotting against transforming the data (using logs) and then plotting the transformed data.

```
#comparision is easier if we plot side by side
par(mfrow=c(1,2))
plot(t+1949, ppd.million.tonnes,
     xlab="(a)", ylab="Million tonnes (log scale)",
     main ="Global plastics production",
     type="l", col = "blue", lwd=1.5,
     log="y")
#draw a horizontal line at the mean value
abline(h=mean(ppd.million.tonnes), lty=2)
print(mean(ppd.million.tonnes))

## [1] 118530303
```

```
plot(t+1949, log.ppd.million.tonnes,
     xlab="(b)", ylab="Million tonnes (log value)",
     main ="Global plastics production",
     type="l", col = "brown", lwd=1.5)
#draw a horizontal line at the mean value
abline(h=log(mean(ppd.million.tonnes)), lty=2)
```




```
par(mfrow=c(1,1))
```

5. Step 5: What is the story about the growth rates in plastic production?

   Calculate the growth rates in plastic production for each of the years between 1951 and 2015 (inclusive)? Summarise the growth rates: determine minimum, maximum, median and mean growth rates? Write down your observations about the growth rates.

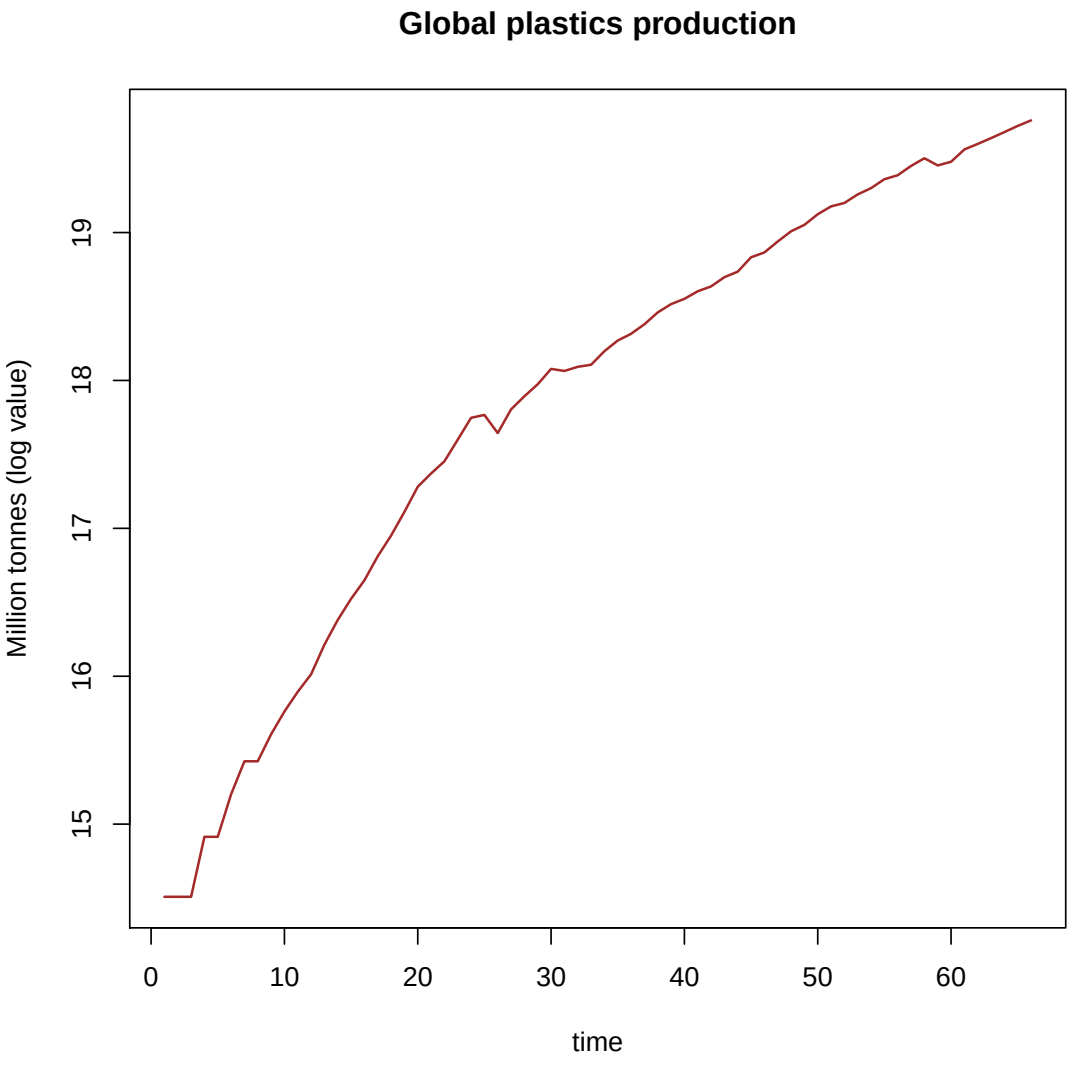
6. Step 6: Determine if an exponential relationship is appropriate

   Do the plots done so far suggest a functional form? What do we know about how plastic production increases over time. One thing we know for sure is that the increase in production is increasing over time. Therefore, a linear relationship (such as $gpp = a + b * t$) is not appropriate. And that is clear from the plot which is closer to being exponential or a quadratic curve opening up than to a linear curve.

We can also assess whether an exponential relationship is appropriate by plotting the log of the gpp against time and checking if that appears linear. This is because if gpp grown exponentially over time (i.e. $gpp = a.e^{rt}$, where r is the rate of exponential growth), then that relationship would be a linear relationship between time and the logged value of gpp. Why? $gpp = a.e^{rt}$ and $log(gpp) = a + rt$ represent the same process or relationship. In other words, an explonential relationship implies that the rate of growth or proportional increase in gpp is constant over time.

So, let's check if what we have looks like this: $log(gpp) = a + rt$, i.e. $log(gpp)$ is a linear function of time. Does the plot below appear linear or very close to being linear?

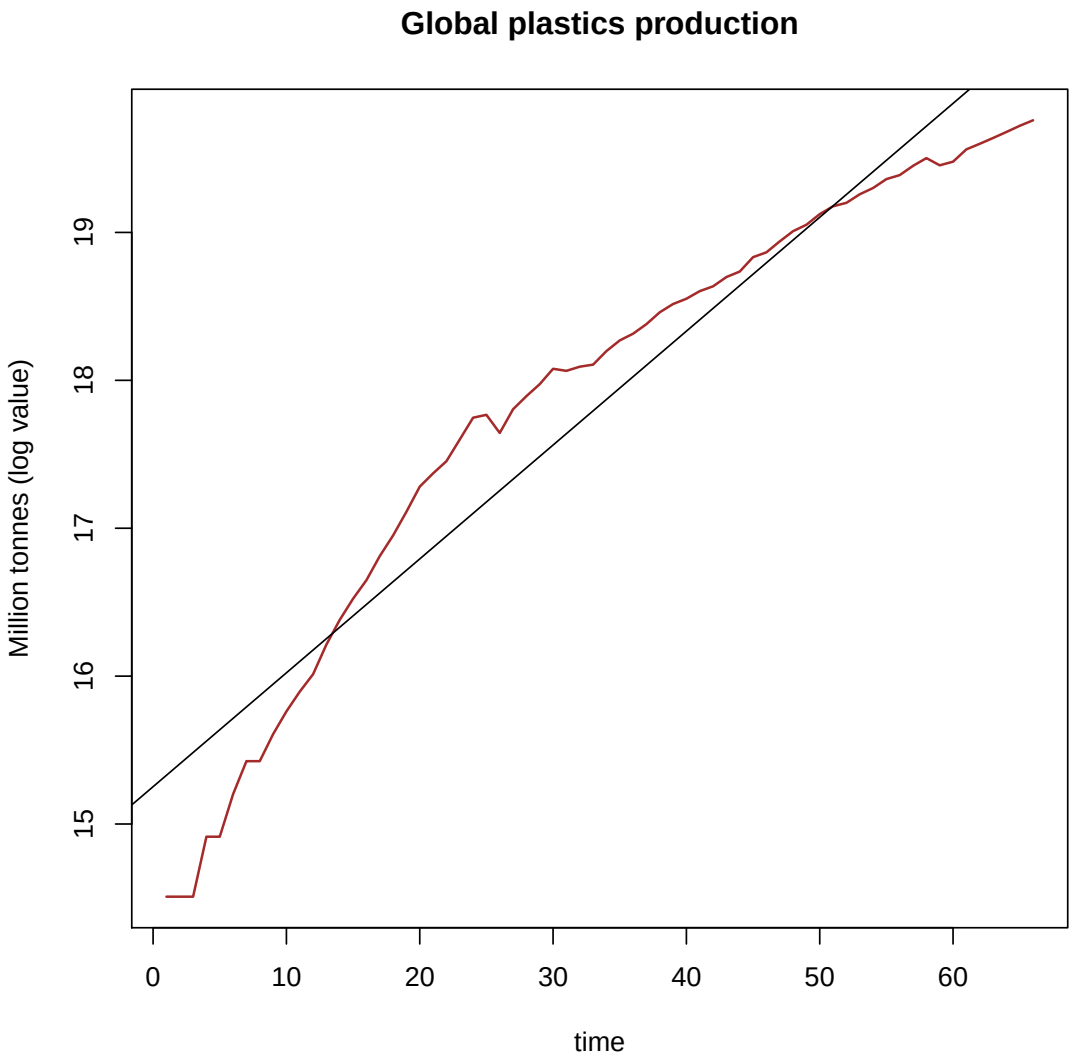
```
#comparision is easier if we plot side by side
par(mfrow=c(1,1))
plot(t, log.ppd.million.tonnes,
     xlab="time", ylab="Million tonnes (log value)",
     main ="Global plastics production",
     type="l", col = "brown", lwd=1.5)
```



How would you fit such a curve? If you have no knowledge of curve, you could fit the curve by trial and error (but different people would end up with different stories) or you could fix a to the log of production at the start of the period and fix r at the average annual growth rate that would bring the gpp up to its 2015 value (ask if you are not sure how this works). However, the focus of our exercise is not curve fitting and we will not get into details here. We can also use curve fitting methods (statistical techniques) to estimate the parameters for exponential relationship by

regressing $log(gpp)$ on t in Excel or some other software. If we did that, the fitted model would be $log(gpp) = 15.252292 + 0.077005 * t$ and the linear approximation plotted against the actual data would look like as follows. We will call this model 1 (**M1**).

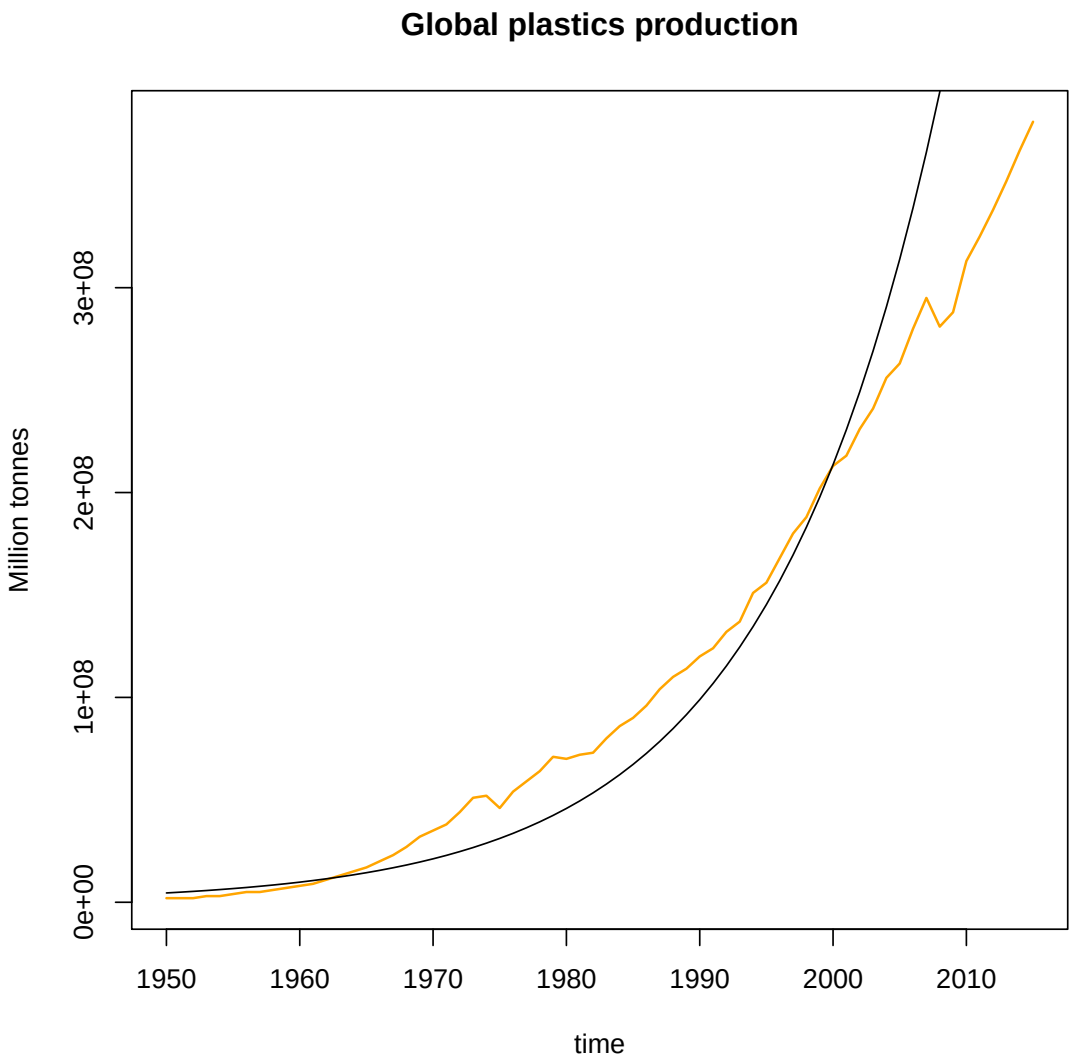
```
#comparision is easier if we plot side by side
par(mfrow=c(1,1))
plot(t, log.ppd.million.tonnes,
     xlab="time", ylab="Million tonnes (log value)",
     main ="Global plastics production",
     type="l", col = "brown", lwd=1.5)

lm.gpp.linear = lm(log.ppd.million.tonnes ~ t)
#summary(lm.gpp.linear)
abline(lm.gpp.linear)
```



If we do the plots in natural units (not logs), this is what we have. Is the model good?

```
#comparision is easier if we plot side by side
par(mfrow=c(1,1))
plot(t+1949, ppd.million.tonnes,
     xlab="time", ylab="Million tonnes",
     main ="Global plastics production",
```

```
     type="l", col = "orange", lwd=1.5)
points(t+1949, exp(15.252292 + 0.077005*t), type="l", col="black")
```



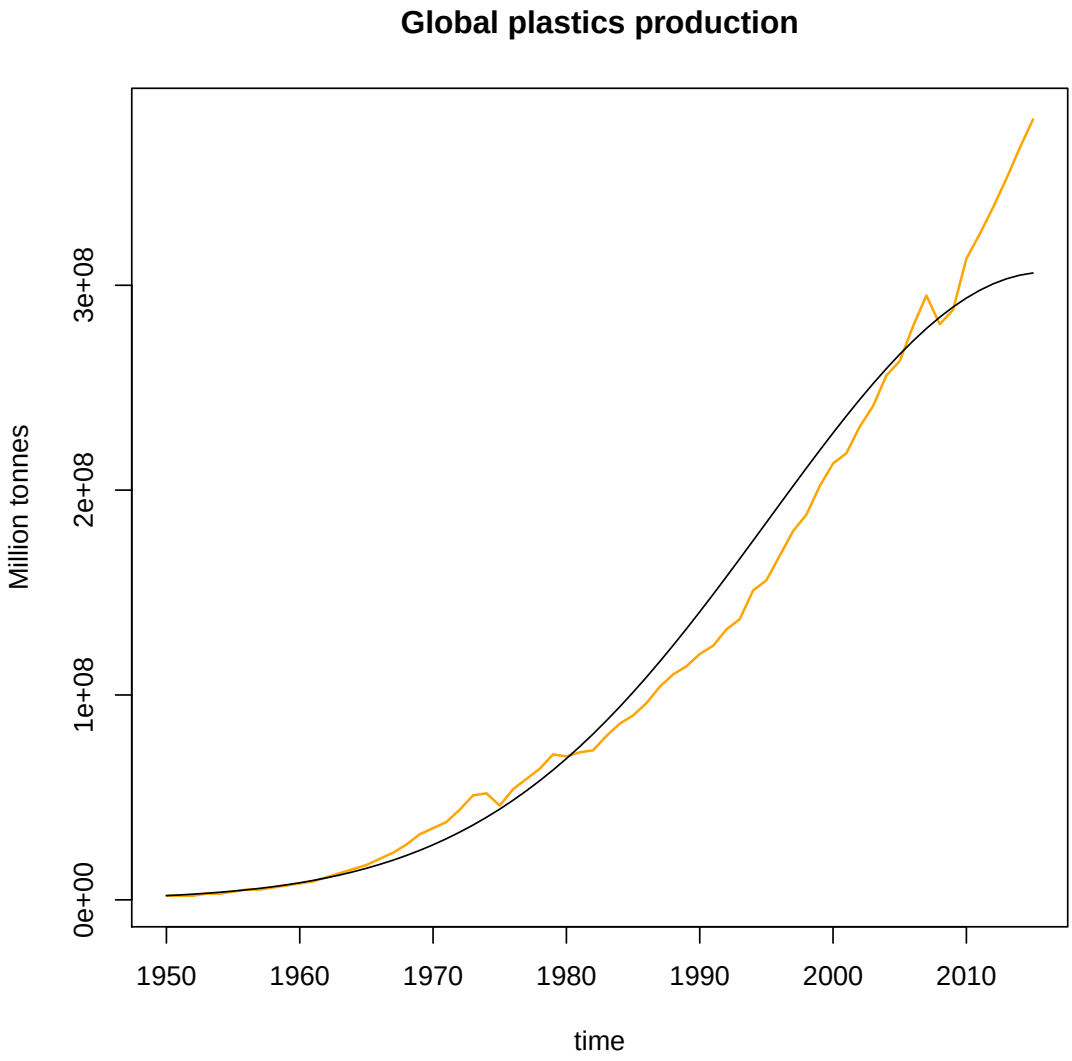
The fitted exponential curve provides a decent approximation to the actual change in plastic production but tends to grossly overestimate trens in production towards the end of the period. Therefore, we need to be cautious about using such a fit. In fact, we could explore other forms that would improve out description of the process.

7. Step 6: Have a go at another functional form

   Based on the plot of log of plastic production against time, we can see that the relationship between the logged value of plastic production and time appears quadratic. Therefore, $log(gpp) = a+bt+ct^2$, where b is positive but c is negative could be better. We could fit such a curve by trial and error (e.g. in Excel) but we could also estimate is appropriately using curve fitting techniques in Excel or other software, to obtain the following: $log(gpp) = 14.38 + 0.1539t - 0.001147t^2$ or $gpp = 1760519e^{0.1539t-0.001147t^2}$. We will call this model 2 (**M2**).

```
#comparision is easier if we plot side by side
par(mfrow=c(1,1))
plot(t+1949, ppd.million.tonnes,
     xlab="time", ylab="Million tonnes",
     main ="Global plastics production",
```

```
     type="l", col = "orange", lwd=1.5)

lm.gpp.quad = lm(log.ppd.million.tonnes ~ t + I(t^2))
#summary(lm.gpp.quad)
fittedv = fitted.values(lm.gpp.quad)
points(t+1949, exp(fittedv), type="l", col="black")
```



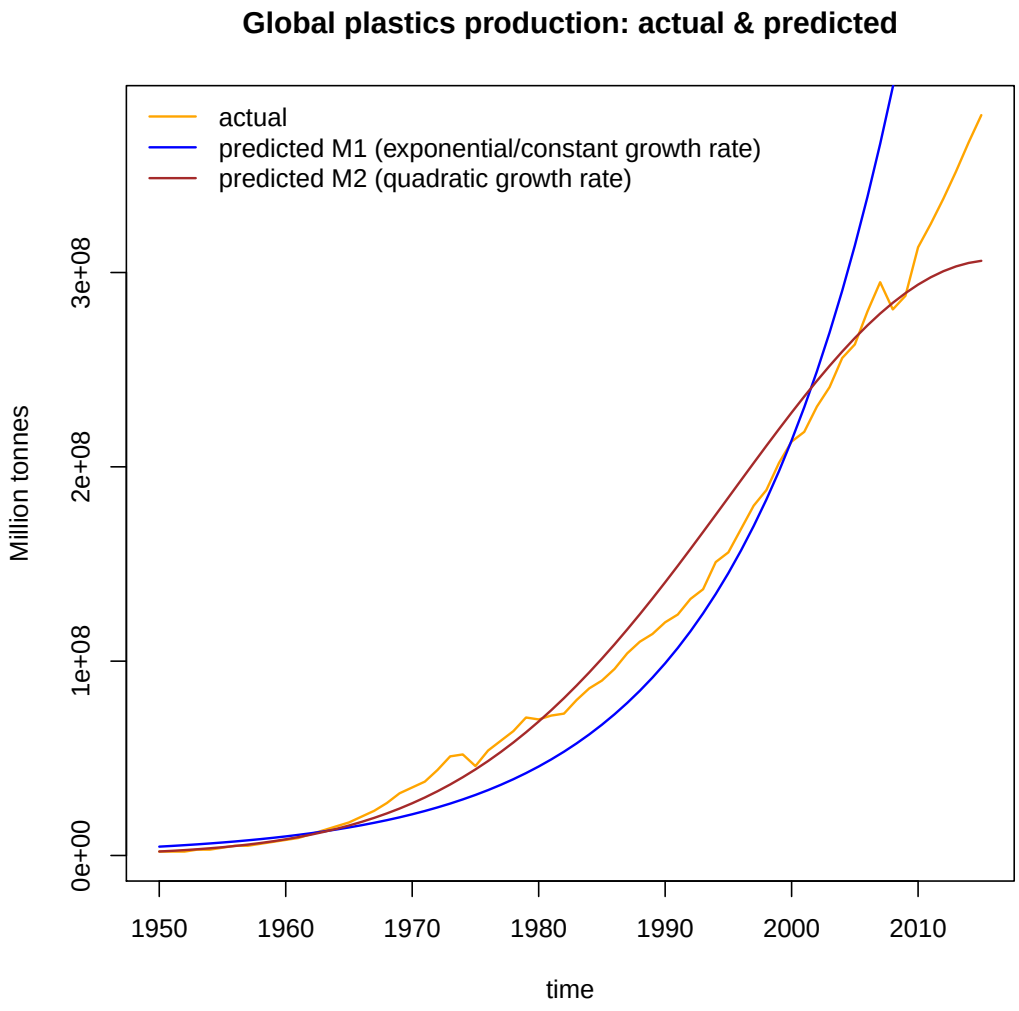
8. Step 7: Student exercise

   - What does the model $log(gpp) = 14.38 + 0.1539t - 0.001147t^2$ imply about the predicted growth rate plastic production between 1960 and 1961, between 2000 and 2001 and between 2010 and 2011? How do these compare with the actual growth rates and with the prediction from the exponential equation?
   - What does this model tell us about the maximum level global plastic production could reach? When would that be reached according the model?

```
## [1] 2016
## [1] -4337.059
```

   - Generate a plot that includes the following three curves: a plot of the actual plastics production, the predicted values accoring to M1 and M2. See the following:

```r
#comparision is easier if we plot side by side
par(mfrow=c(1,1))
plot(t+1949, ppd.million.tonnes,
     xlab="time", ylab="Million tonnes",
     main ="Global plastics production: actual & predicted",
     type="l", col = "orange", lwd=1.5)
points(t+1949, exp(15.252292 + 0.077005*t), type="l", col="blue", lwd=1.5)
points(t+1949, exp(fittedv), type="l", col="brown", lwd=1.5)
legend("topleft", c("actual",
                    "predicted M1 (exponential/constant growth rate)",
                    "predicted M2 (quadratic growth rate)"),
       col=c("orange","blue","brown"), lwd=rep(1.5,3), lty=rep(1,3), bty="n")
```



- Is it wise to use a model (M1, M2, or any other similar model) to predict plastic production far into the future? Explian why?