NTHU STAT 5410 - Linear Models

Assignment 7 Report

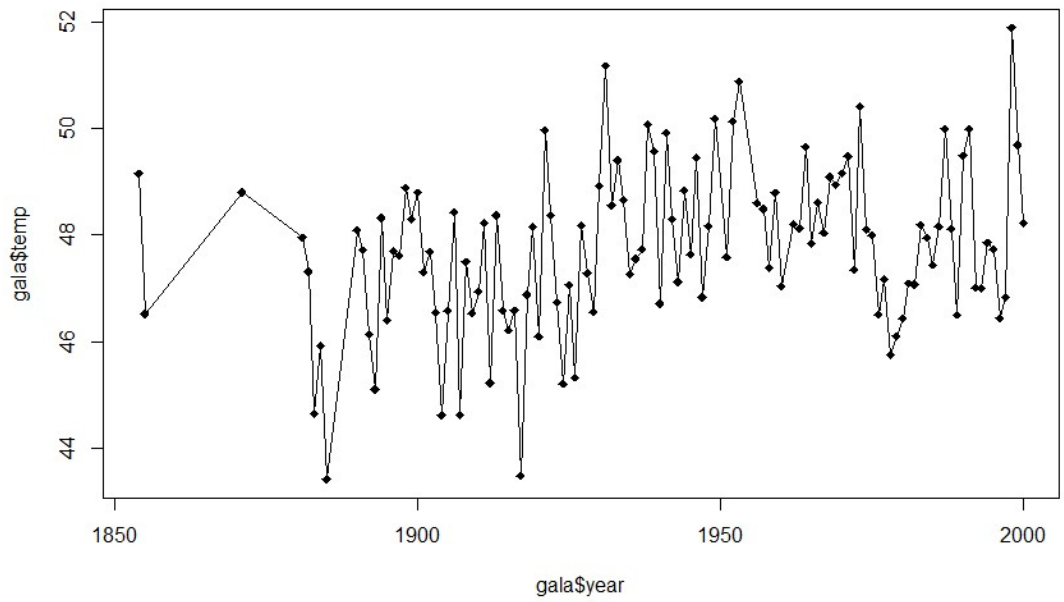
105061110 周柏宇

1.
```
> gala <- read.table("C:/Users/Thomas/Downloads/Linear_models/hw7/climate.txt", header=T)
```

i.

From the plot we can see a slight positive correlation before 1950 and negative correlation after 1950, maybe these variables has quadratic relationship and the linear trend is not too clear. This can also be confirmed by their correlation value.

```
> plot(gala$year, gala$temp)
```




```
> cor(gala$year, gala$temp)
[1] 0.2921634
```

ii.

Because the temperature data are time series, it is nature to have correlation in successive observations. However, the time in this dataset is not distributed regularly i.e. there are missing values for several years. Before we calculated the autocorrelation, we first need some preprocessing.

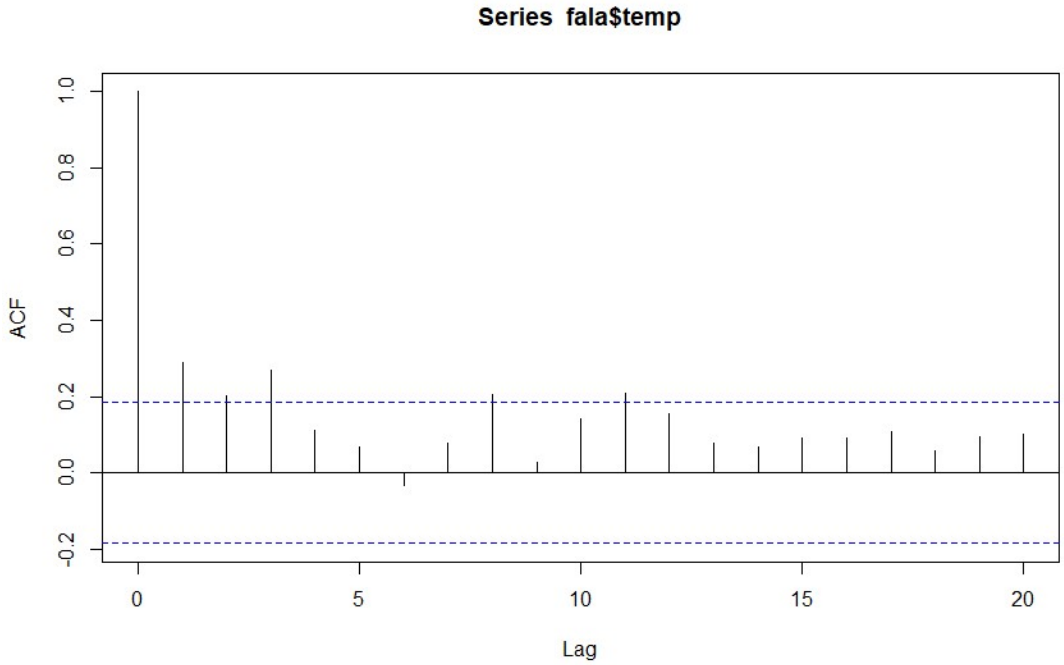
I create a processed version of the dataset “climate_fix.txt” with these procedures:

```
> dyear <- diff(gala$year)
> dyear[dyear > 1] # gap in the missing years
[1] 16 10  5  2  3  2
> which(dyear > 1) # indices where the gap happens
[1]  2  3  8 68 71 76
```

For those gaps being too large, there is no way we to remedy. Therefore, I ditch the data before year 1890. The rest of gap consists of one or two missing years. We can just use nearest year or averaging to interpolate the missing data.

```
> fala <- read.table("C:/Users/Thomas/Downloads/Linear_models/hw7/climate_fix.txt", header=T)
```

Then we can take a look at the autocorrelation in the temperature.



There are indeed higher correlations within successive observations. But the correlation is strong enough to indicate a linear trend.

iii.

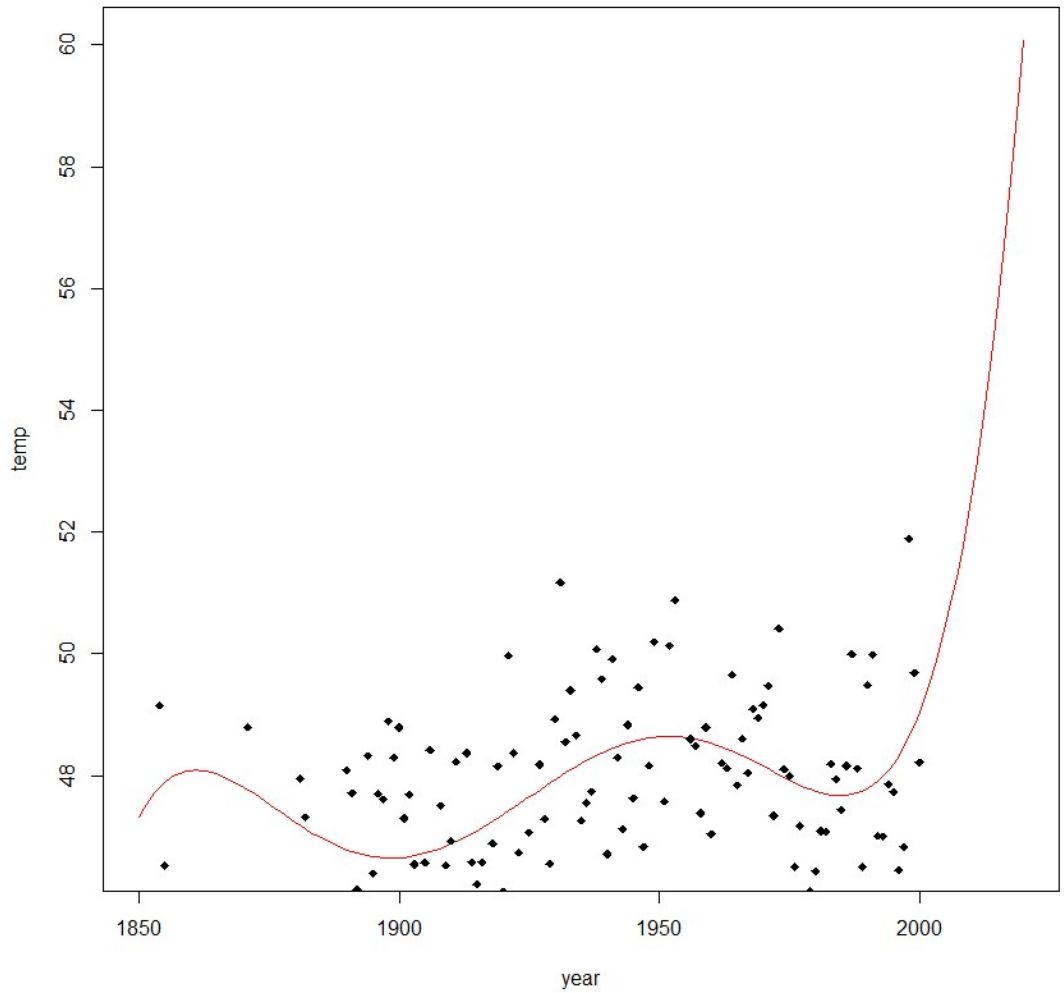
```
> fit10 <- lm(temp ~ poly(year, 10), data=gala)
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
```

```
poly(year, 10)10   0.3474     1.4146   0.246  0.80652
```

```
> fit9 <- lm(temp ~ poly(year, 9), data=gala)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
poly(year, 9)9   1.3994     1.4083   0.994  0.32265

> fit8 <- lm(temp ~ poly(year, 8), data=gala)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
poly(year, 8)8  -1.1011     1.4082  -0.782  0.43600

> fit7 <- lm(temp ~ poly(year, 7), data=gala)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
poly(year, 7)7  -0.9373     1.4056  -0.667 0.506341

> fit6 <- lm(temp ~ poly(year, 6), data=gala)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
poly(year, 6)6   1.2124     1.4020   0.865 0.389067

> fit5 <- lm(temp ~ poly(year, 5), data=gala)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
poly(year, 5)5   3.3824     1.4004   2.415 0.017384 *
```

By the backward elimination, we stop at 5th order polynomial.

```
> x <- c(1850:2020)
> xx <- data.frame(year=x)
> pred <- predict(fit5, xx)
> plot(x, pred, type="l", xlab="year", ylab="temp",
 col=2)
> points(gala$year, gala$temp, pch=18)
```



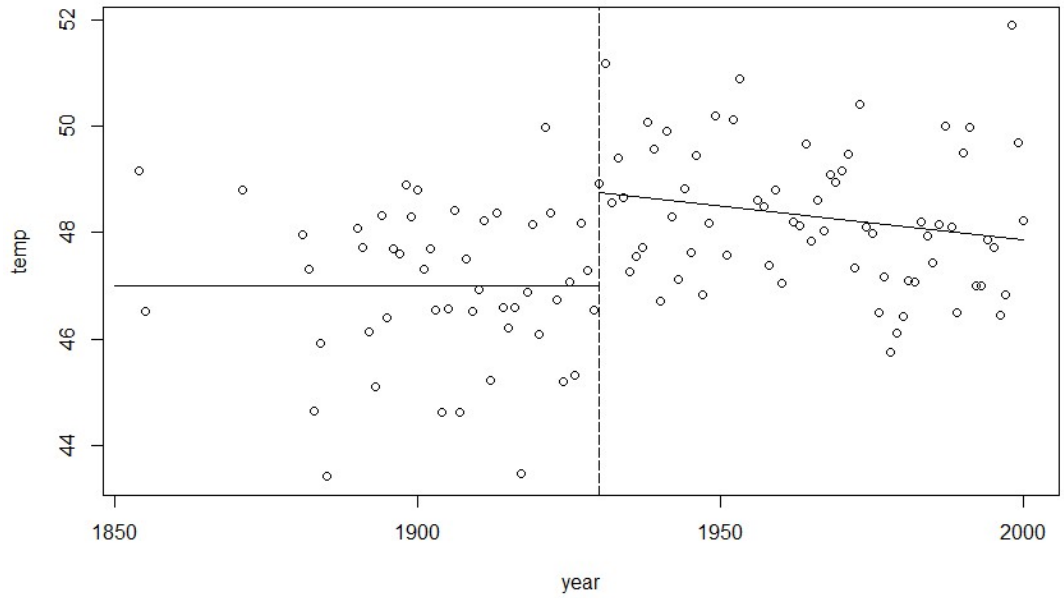

```
> predict(fit5, data.frame(year=2020))
   1
60.07774
```

The prediction for temperature in 2020 is 60.07774 degree. Since it is an extrapolation, the prediction is not so reliable as we can see in the plot.

iv.

```
> g1 <- lm(temp ~ 1, data=gala,subset=(year<=1930))
> g1$coefficients
(Intercept)
   46.99673
> g2 <- lm(temp~year, data=gala,subset=(year>1930))
> g2$coefficients
```

```
(Intercept)        year
72.90952792 -0.01251854
> plot(gala$year,gala$temp,xlab="year",ylab="temp
")
> segments(1850, g1$coef[1], 1930, g1$coef[1])
> abline(v=1930, lty=5)
> segments(2000, g2$coef[1]+g2$coef[2]*2000, 1930,
g2$coef[1]+g2$coef[2]*1930)
```

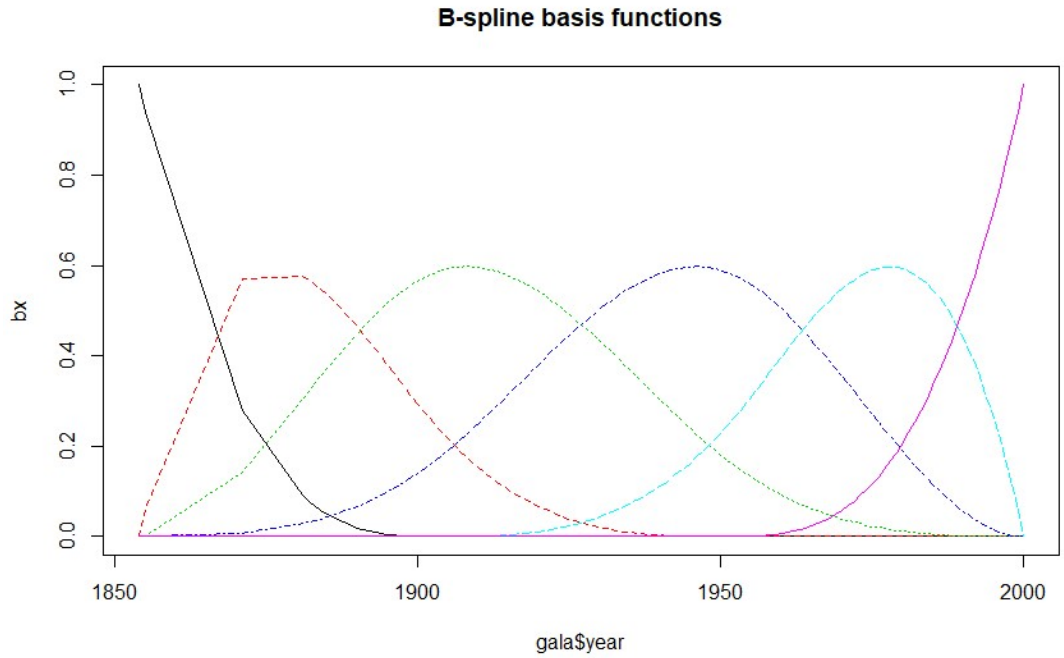


The discontinuous prediction for the temperature is ridiculous. I don't think broken stick regression is suitable for this task.

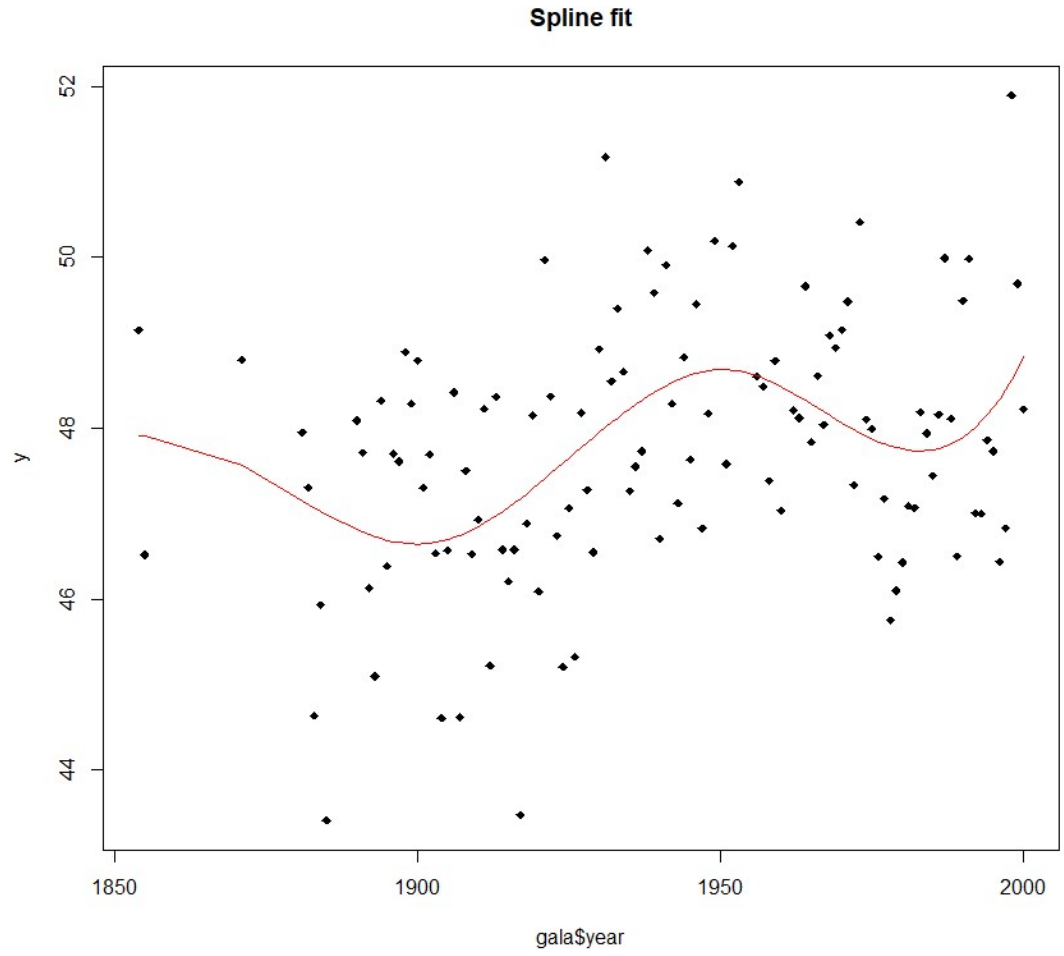
v.

The range of years is 1854 to 2000. To get 6 basis, we need to divide the range into 4 regions and repeats the head and tail 3 times each.

```
> library(splines)
> knots <- c(1854, 1854, 1854, 1854, 1902.5, 1951.5,
 2000, 2000, 2000, 2000)
> year_spline <- splineDesign(knots, gala$year)
> matplot(gala$year,year_spline,type="l",main="B-s
pline basis functions")
```



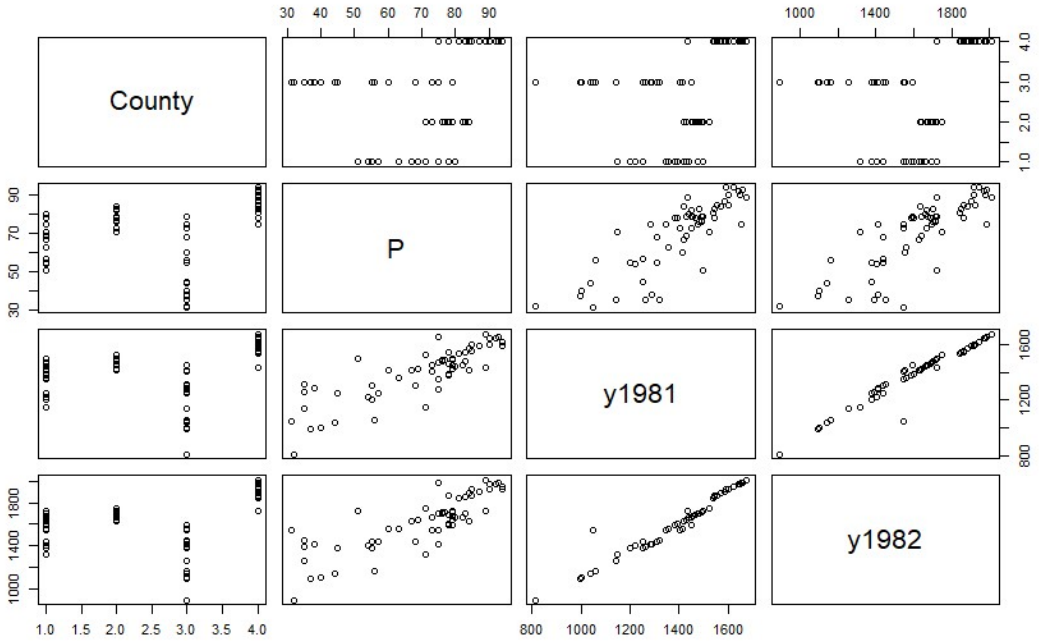

```
B_fit <- lm(gala$temp ~ year_spline)
```



I think this model capture the oscillation of temperature along the year pretty well, which I did not expect such trend to exist. In this case, I think we can really obtain some knowledge from the fitted model about which we do not know before.

2. We first observe the data set.

```
> gala <- read.table("C:/Users/Thomas/Downloads/Linear
_models/hw7/soil.txt", header=T)
> pairs(gala)
```

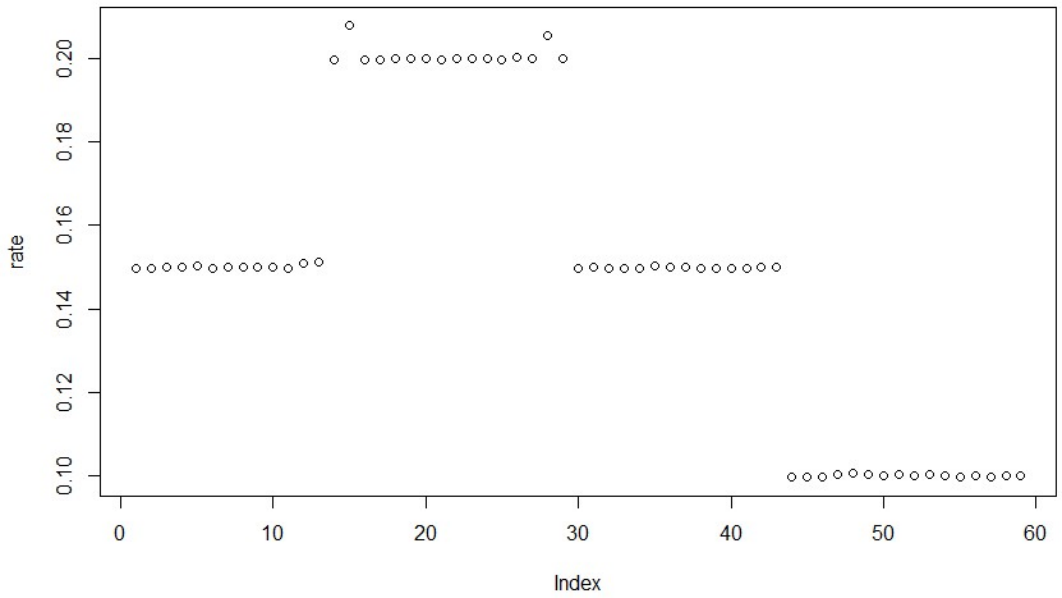


- For the four countries, the variables exhibit different mean and variance with little outlier.
- y1981 and y1982 are strongly correlated (cor=0.9677936). There seem to be an outlier being way above the line at y1981=1047. It turns out to be the observation “44 Meeker 31 1047 1548”. We will exclude this point for the rest of analysis.
- P and y1981 are positively correlated. (cor=0.8361775), so does P and y1982 (cor=0.8206177).

Since we would like to use the information P (soil productivity score) to predict farmland values. We might need to incorporate the price difference between 1981 and 1982 as a rate of growth to predict values for arbitrary year.

```
> county <- gala[,1]
> P <- gala[,2]
```

```
> y1981 <- gala[,3]
> y1982 <- gala[,4]
> delta <- y1982 - y1981
> rate <- delta / y1981
```



It turns out that the price growth rate is quite consistent in the same counties. But I don't think we have a way to use this property to predict country beyond these four.

The way I approach this is to use the soil productivity score P to predict y1982 (for prediction latter than 1982; otherwise, predict y1981). And using the fitted y1982 we predict the grow rate. Finally, we assume a constant growth rate for the farmland values to extrapolate the value for arbitrary year.

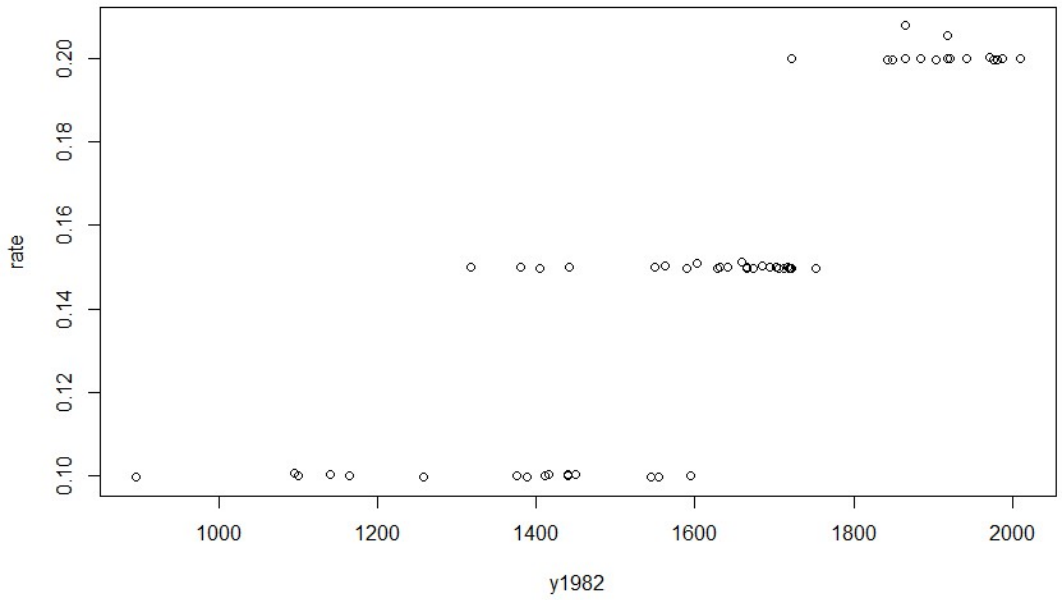
```
> fit_82 <- lm(y1982 ~ P)
> summary(fit_82)

Call:
lm(formula = y1982 ~ P)

Residuals:
    Min      1Q  Median      3Q     Max
-309.59 -102.60   10.56   76.59  337.59

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  753.653     74.890   10.06 2.98e-14 ***
P            12.309      1.031   11.94  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.
1 ‘ ’ 1

Residual standard error: 136.4 on 57 degrees of freedom
Multiple R-squared:  0.7144, Adjusted R-squared:  0.709
4
F-statistic: 142.6 on 1 and 57 DF,  p-value: < 2.2e-16
```

We obtain the predicted farmland values in 1982 with

$$\widehat{y1982} = 753.653 + 12.309 * \mathrm{P}$$

Next, we need to predict the growth rate



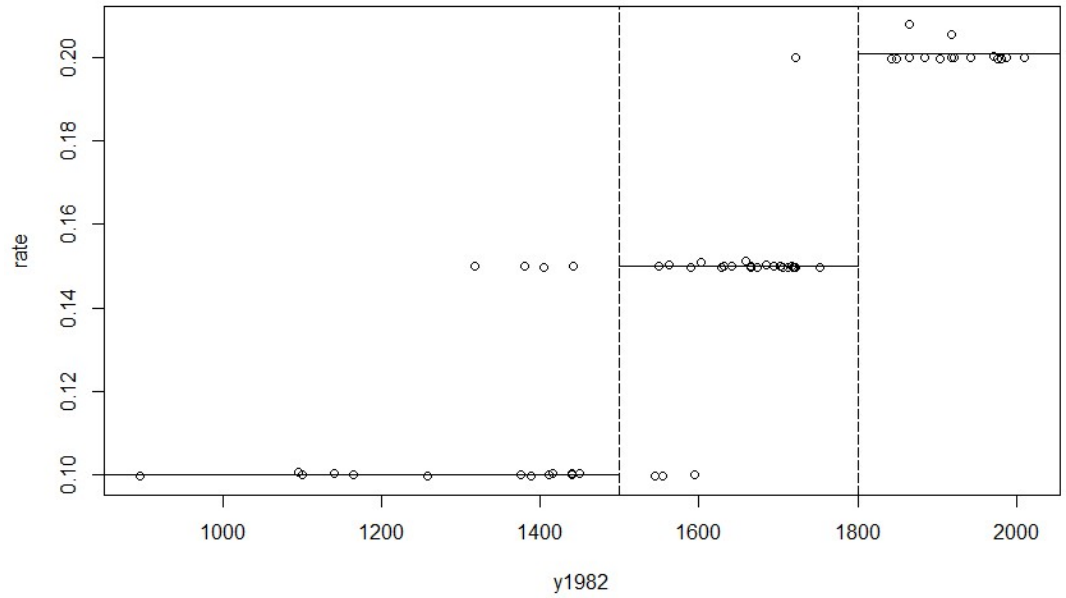
We can tell that there are 3 kinds of rate corresponding to the region roughly divided by y1982=1500 and 1800. We will fit a constant for each region (excluding the outliers).

```
> r1 <- lm(rate ~ 1, subset=(y1982 < 1500 & rate < 0.1
2))
> r2 <- lm(rate ~ 1, subset=(y1982 >= 1500 & y1982 < 180
0 & rate >= 0.12 & rate < 0.18))
> r3 <- lm(rate ~ 1, subset=(y1982 >= 1800 & rate >= 0.1
8))
```

```
> segments(800, r1$coef[1], 1500, r1$coef[1])
> abline(v=1500, lty=5)
> segments(1500, r2$coef[1], 1800, r2$coef[1])
> abline(v=1800, lty=5)
> segments(1800, r3$coef[1], 2200, r3$coef[1])
> r1

Call:
lm(formula = rate ~ 1, subset = (y1982 < 1500 & rate <
0.12))

Coefficients:
(Intercept)
       0.1

> r2

Call:
lm(formula = rate ~ 1, subset = (y1982 >= 1500 & y1982 <
 1800 &
    rate >= 0.12 & rate < 0.18))

Coefficients:
(Intercept)
      0.15

> r3

Call:
lm(formula = rate ~ 1, subset = (y1982 >= 1800 & rate >=
 0.18))

Coefficients:
(Intercept)
    0.2008
```

The predicted growth rate is

$$\widehat{rate} = \begin{cases} 0.1, & \widehat{y1982} < 1500 \\ 0.15, & 1500 \le \widehat{y1982} < 1800 \\ 0.2008, & 1800 \le \widehat{y1982} \end{cases}$$



Finally, the extrapolation for the farmland values at certain year x is

$$\widehat{yx} = \widehat{y1982} * \widehat{rate}^{x-198}$$

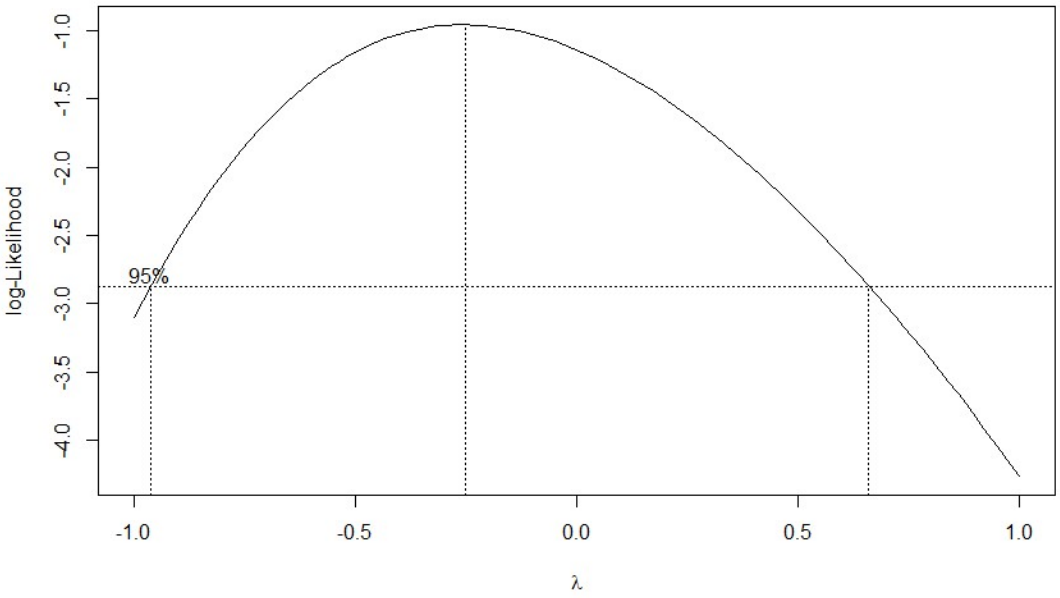
and same procedure for year before 1981.

Because we are provided with too little information about the trend of value, I don't expect the naïve constant growth rate model to be very accurate. Also, the capability of generalizing the growth rate of one area to another is questionable (as shown in previous plots, the growth rate has spatial homogeneity). In summary, this model can only service as a rough estimation of the farmland values given solely the soil productivity score as predictor.

3.

```
> gala <- read.table("C:/Users/Thomas/Downloads/Linear
_models/hw7/car.txt", header=T)
>
> AO <- gala[,2]
> GNP <- gala[,3]
> CP <- gala[,4]
> OP <- gala[,5]
>
> DAO <- diff(AO)
```

```
> DGNP <- diff(GNP)
> DCP <- diff(CP)
> DOP <- diff(OP)
>
> AO <- AO[1:length(AO)-1]
> GNP <- GNP[1:length(GNP)-1]
> CP <- CP[1:length(CP)-1]
> OP <- OP[1:length(OP)-1]
>
> g <- lm(DAO ~ DGNP + DCP + DOP)
> summary(g)

Call:
lm(formula = DAO ~ DGNP + DCP + DOP)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.265e-04  5.566e-04   0.227   0.8292
DGNP         8.097e-06  4.731e-06   1.711   0.1477
DCP         -6.066e-07  5.191e-07  -1.169   0.2952
DOP         -2.773e-06  1.048e-06  -2.645   0.0457 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0004083 on 5 degrees of freedom
Multiple R-squared:  0.7154,    Adjusted R-squared:  0.5447
F-statistic:  4.19 on 3 and 5 DF,  p-value: 0.0785

> library(MASS)
> boxcox(g, plotit=T, lambda=seq(-1,1,by=0.1))
```



Since the confidence interval does not contains 1, this implies we do need transformation. Furthermore, 0 is in the confidence interval, we can choose it as it is a more interpretable than the max log-likelihood λ, which is around -0.25. (note that $\lambda = 0$ corresponds to logarithm transformation)

```
> log_DAO <- log(DAO)
> g2 <- lm(log_DAO ~ DGNP + DCP + DOP)
> summary(g2)

Call:
lm(formula = log_DAO ~ DGNP + DCP + DOP)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.5261084  0.6920065 -12.321 6.24e-05 ***
DGNP         0.0129407  0.0058827   2.200  0.07912 .
DCP         -0.0007489  0.0006454  -1.160  0.29825
DOP         -0.0056283  0.0013036  -4.318  0.00759 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 0.5076 on 5 degrees of freedom
```

```
Multiple R-squared:  0.8374,    Adjusted R-squared:  0.7398
F-statistic: 8.584 on 3 and 5 DF,  p-value: 0.02041
```

With the transformation, we achieve

- Significance for DGNP and DOP increase noticeably
- Multiple R-squared increase from 0.7154 to 0.8374