NTHU STAT 5410 - Linear Models Assignment 7 Report

105061110 周柏宇

1. i.

ii.

iii.

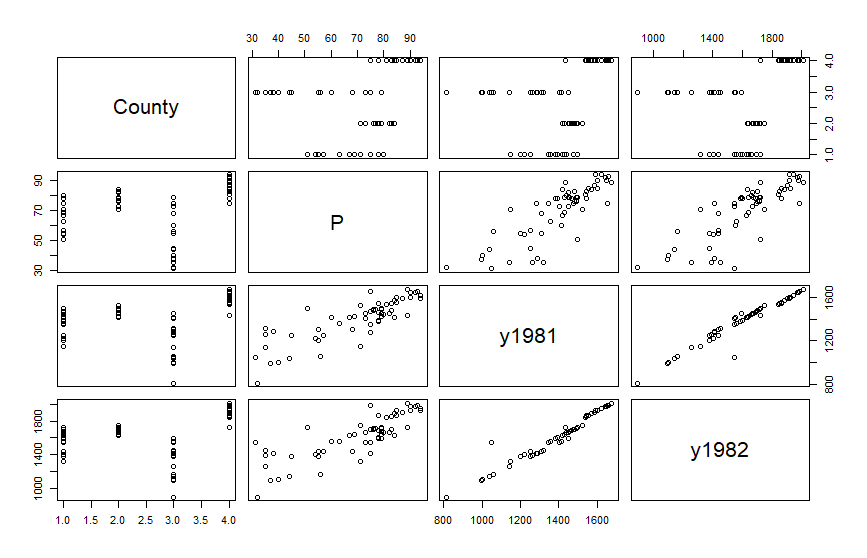
iv.

v.

1. 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 following prediction.
* 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]

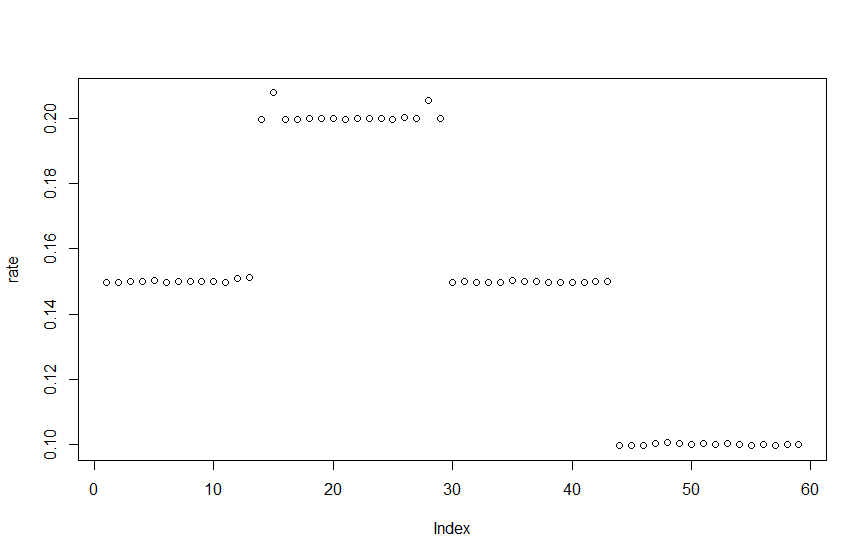
> P <- gala[,2]

> 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

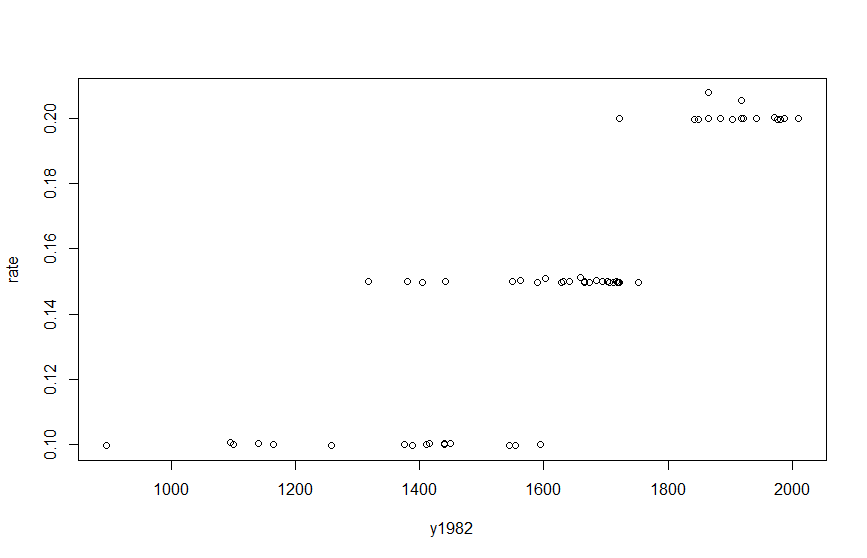
Residual standard error: 136.4 on 57 degrees of freedom

Multiple R-squared: 0.7144, Adjusted R-squared: 0.7094

F-statistic: 142.6 on 1 and 57 DF, p-value: < 2.2e-16

We obtain the predicted farmland values in 1982 with

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 < 0.12))

> r2 <- lm(rate ~ 1, subset=(y1982 >= 1500 & y1982 < 1800 & rate >= 0.12 & rate < 0.18))

> r3 <- lm(rate ~ 1, subset=(y1982 >= 1800 & rate >= 0.18))

> 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:

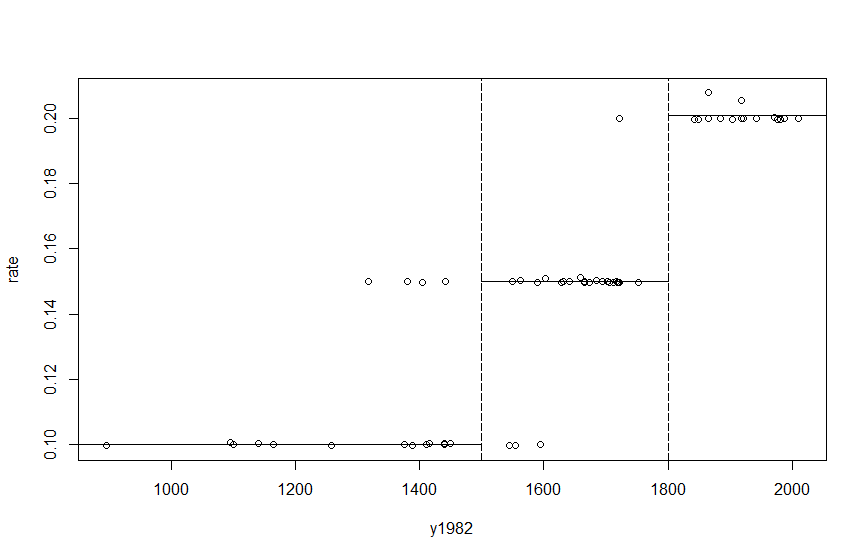
lm(formula = rate ~ 1, subset = (y1982 >= 1800 & rate >= 0.18))

Coefficients:

(Intercept)

0.2008

The predicted growth rate is



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

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.