Piecewise Regression

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Piecewise Regression

Introduction

- Here is a fascinating data set, originally distributed as part of an article in the online *Journal of Statistics Education*.
- The reference is Rouncefield, M. (1995). The statistics of poverty and inequality. *Journal of Statistics Education*, 3.
  The article, which you will not need to complete this exercise, may be downloaded from *http://www.amstat.org/publications/jse/v3n2/datasets.rouncefield.html*
  The article contains a substantial amount of background information on the meaning of the variables contained in the data file.
- We'll load in the data directly from a text file and remove the missing values.

```
> poverty.data <- read.table ("poverty.dat",header=T)
> poverty.data <- na.omit(poverty.data)
> attach(poverty.data)
> names(poverty.data)
[1] "Birth.Rate"       "Death.Rate"       "Infant.Mortality" "Male.Life.Exp"
[5] "Female.Life.Exp"  "GNP"              "Region"           "Country"
```

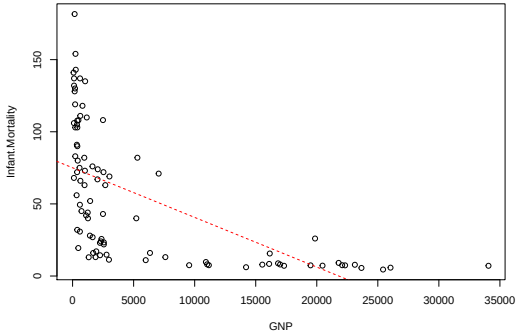
Infant Mortality vs. GNP

- In this exercise, we are particularly interested in the relationship between `Infant.Mortality` and `GNP` (gross national product), often taken as an indicator of overall economic productivity.
- Using techniques demonstrated previously in class, we begin by producing a scatterplot showing `Infant.Mortality` predicted from `GNP`. We fit the data with linear regression and add the line to the scatterplot in dotted red.

Infant Mortality vs. GNP

- The graph appears to be extremely nonlinear.

```
> fit <- lm(Infant.Mortality ~ GNP)
> plot(GNP,Infant.Mortality)
> abline(fit,lty=2,col="red")
```



Infant Mortality vs. GNP

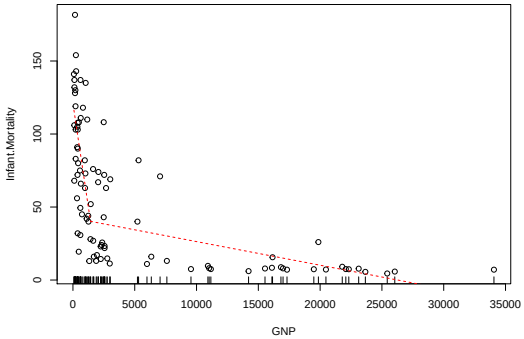
- Let's digress to scan sections of the graph using an interactive regression plotter.
- We're back!
- We discovered during our analysis that several Middle Eastern oil-producing nations seemed not to follow the model that worked well for the rest of the data.
- Now let's try an automatic piecewise (or "segmented") regression program, in the `segmented` package.
- The program picks one or more points automatically, reports back with the results, and plots the regression lines.

Infant Mortality vs. GNP

```
> library(segmented)
> out.lm <- lm(Infant.Mortality ~ GNP)
> o<-segmented(out.lm,seg.Z=~GNP,psi=list(GNP=c(4000)),
+              control=seg.control(display=FALSE))
> summary(o)

        ***Regression Model with Segmented Relationship(s)***

Call:
segmented.lm(obj = out.lm, seg.Z = ~GNP, psi = list(GNP = c(4000)),
    control = seg.control(display = FALSE))

Estimated Break-Point(s):
   Est. St.Err
1419.0  239.5

t value for the gap-variable(s) V:  0

Meaningful coefficients of the linear terms:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 120.85388    8.00815  15.091  < 2e-16 ***
GNP          -0.05683    0.01245  -4.565 1.63e-05 ***
U1.GNP        0.05522    0.01246   4.432       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.83 on 87 degrees of freedom
Multiple R-Squared: 0.6252,  Adjusted R-squared: 0.6123

Convergence attained in 4 iterations with relative change -3.421471e-15
```

Infant Mortality vs. GNP

```
> plot(GNP,Infant.Mortality)
> plot(o,col="red",lty=2,add=TRUE)
```



Infant Mortality vs. GNP

- We can see how the outliers distorted the results.
- We can remove them and recheck.

Electricity Consumption vs. Weather

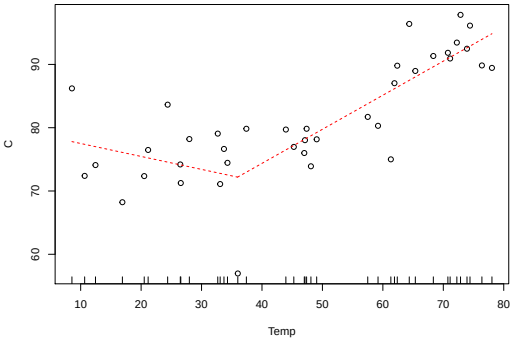
- The segreg data set presents day-by-day records of the average temperature and the electricity consumption for 39 months in a building on the University of Minnesota campus.
- Of course, the average temperature is an imperfect indicator of "degree days," which are more relevant for prediction of electricity consumption.
- This building used electricity for cooling, but not for heating. Hence, one might expect the graph to be relatively flat up to the point where electricity starts "kicking in".
- When we plot the data and do a segmented regression, we see the effect of one observation once more.

Electricity Consumption vs. Weather

```
> data(segreg)
> attach(segreg)
> out.lm <- lm(C~Temp)
>   o<-segmented(out.lm,seg.Z=~Temp,psi=list(Temp=c(40)),
+                control=seg.control(display=FALSE))
>   summary(o)

        ***Regression Model with Segmented Relationship(s)***

Call:
segmented.lm(obj = out.lm, seg.Z = ~Temp, psi = list(Temp = c(40)),
    control = seg.control(display = FALSE))

Estimated Break-Point(s):
   Est. St.Err
35.960  4.582

t value for the gap-variable(s) V:  0

Meaningful coefficients of the linear terms:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  79.5467     4.2165  18.866   <2e-16 ***
Temp         -0.2043     0.1686  -1.212    0.234
U1.Temp       0.7428     0.1886   3.939       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.343 on 35 degrees of freedom
Multiple R-Squared: 0.6816,  Adjusted R-squared: 0.6543

Convergence attained in 3 iterations with relative change 2.287381e-16
```

Electricity Consumption vs. Weather

```
>   plot(Temp,C)
>   plot(o,col="red",lty=2,add=TRUE)
```



Electricity Consumption vs. Weather

- We'll remove data point 15 and replot.

```
>  Tempr <- Temp[-15]
>  Cr <- C[-15]
>   out.lm <- lm(Cr~Tempr)
>   o<-segmented(out.lm,seg.Z=~Tempr,psi=list(Tempr=c(40)),
+               control=seg.control(display=FALSE))
>   summary(o)

        ***Regression Model with Segmented Relationship(s)***

Call:
segmented.lm(obj = out.lm, seg.Z = ~Tempr, psi = list(Tempr = c(40)),
    control = seg.control(display = FALSE))

Estimated Break-Point(s):
   Est. St.Err
48.010  5.378

t value for the gap-variable(s) V:  0

Meaningful coefficients of the linear terms:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 74.97864    2.67970  27.980   <2e-16 ***
Tempr        0.04445    0.08307   0.535    0.596
U1.Tempr     0.57165    0.15000   3.811       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.517 on 34 degrees of freedom
Multiple R-Squared: 0.7259,  Adjusted R-squared: 0.7017

Convergence attained in 4 iterations with relative change 1.640044e-16
```

Electricity Consumption vs. Weather

```
> plot(Tempr,Cr)
> plot(o,col="red",lty=2,add=TRUE)
```