Project

Xuehan Zhao

February 23, 2018

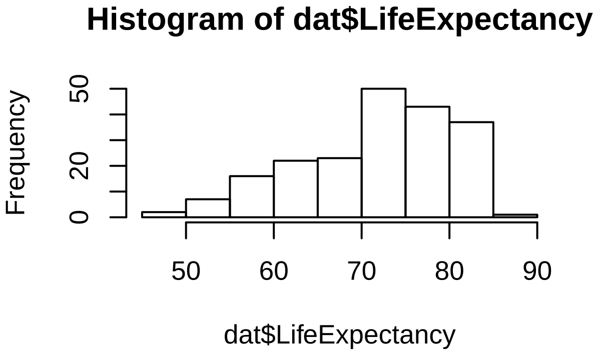
1. Overview

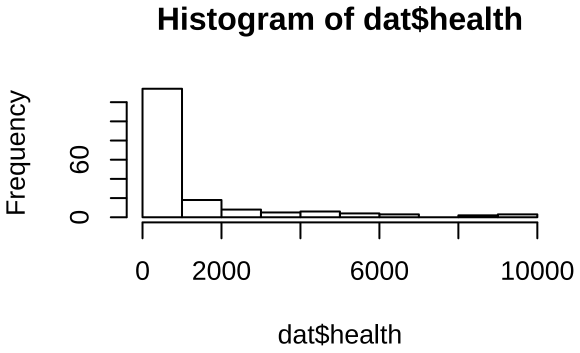
201 observations.

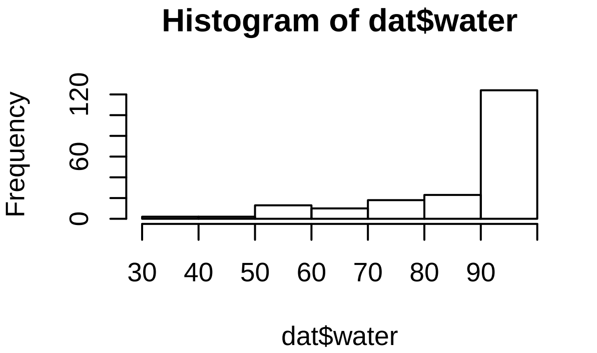
Dependent variable: Life Expectancy

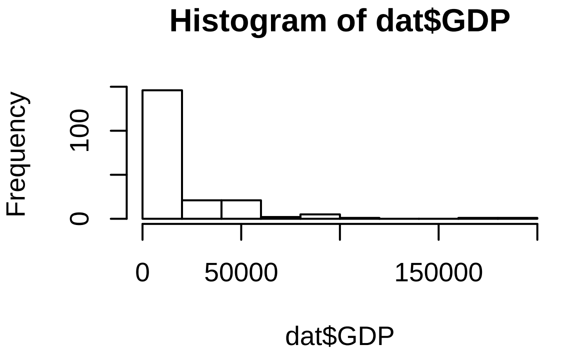
Independent Variables: Health, water, GDP, Region, measles, pm2.5, food.

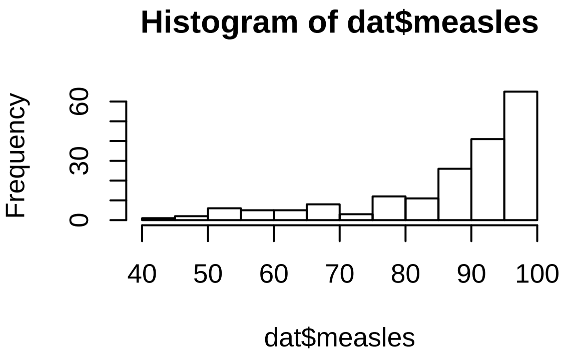
histogram of variables

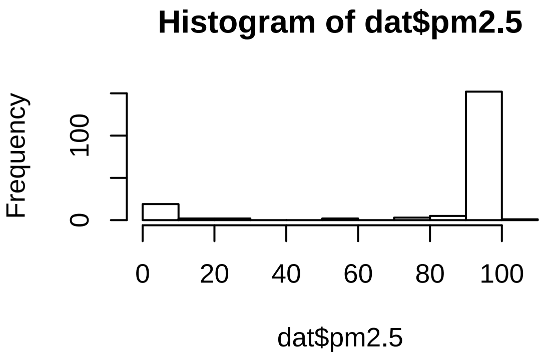


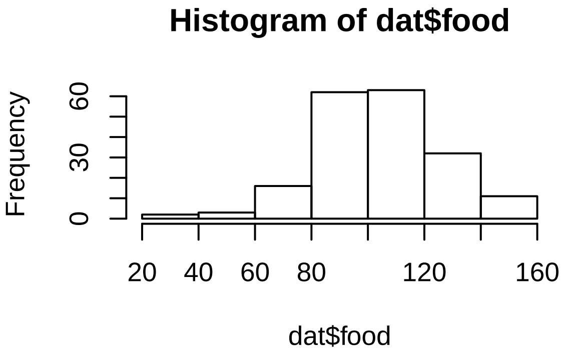










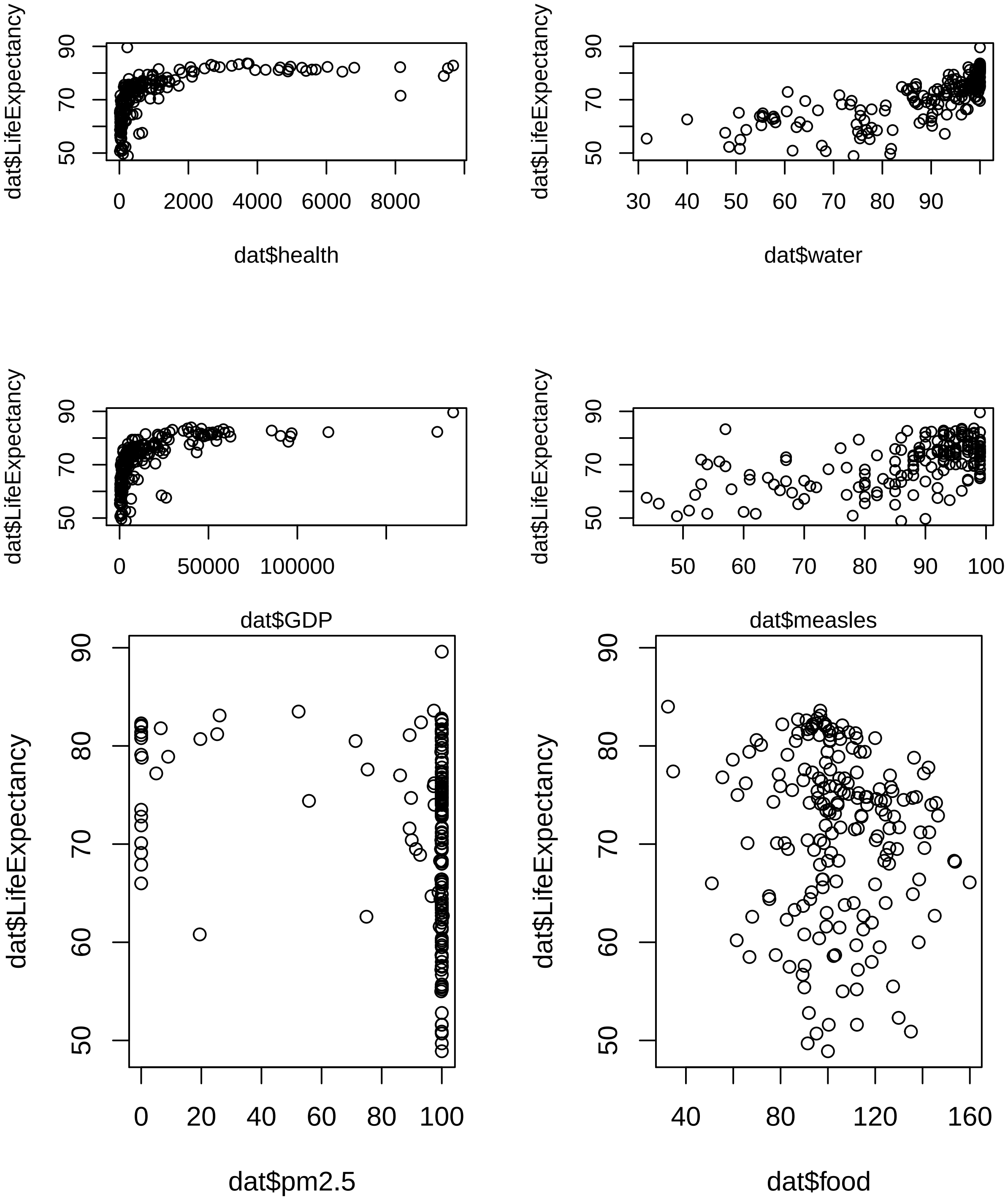


Only food is roughly symmetric.

Life expectancy, water, measles, pm2.5 are left-skewed.

health, GDP are right-skewed.

scatter plot of x vs y



2. Transformation

Life Expectancy - unchanged

health - take log

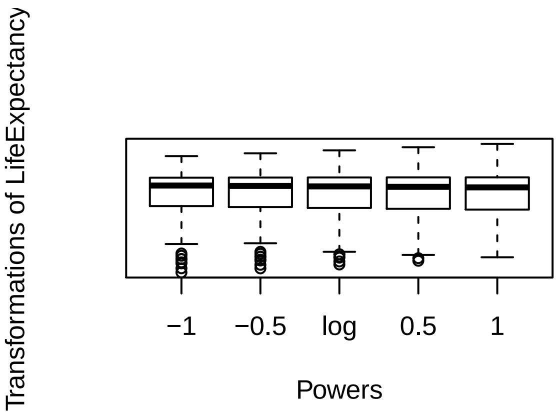
water - p = 9

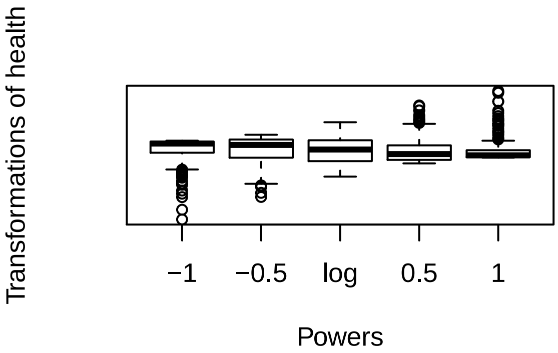
GDP - take log

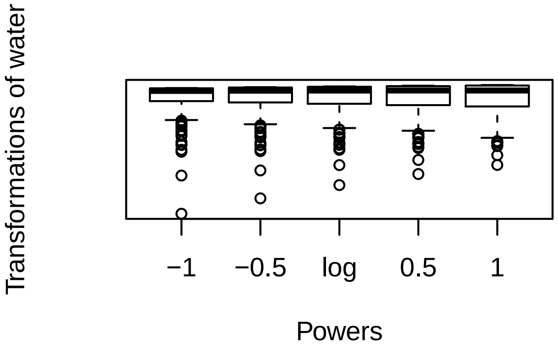
measles - turn into categorical data, cut at 50,75

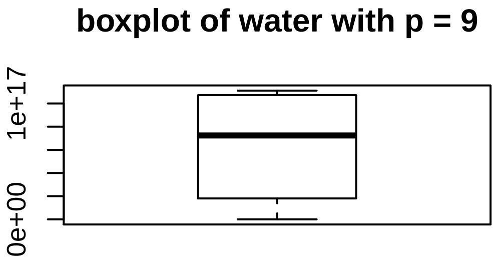
pm2.5 - logit

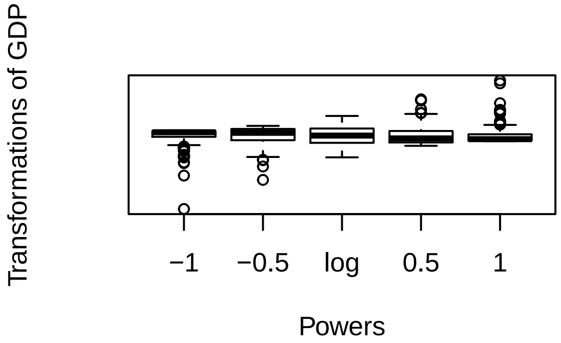
food - unchanged

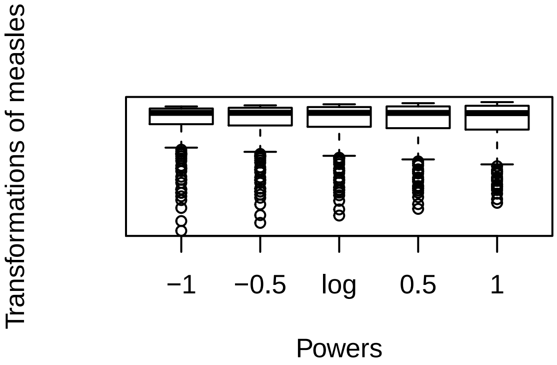


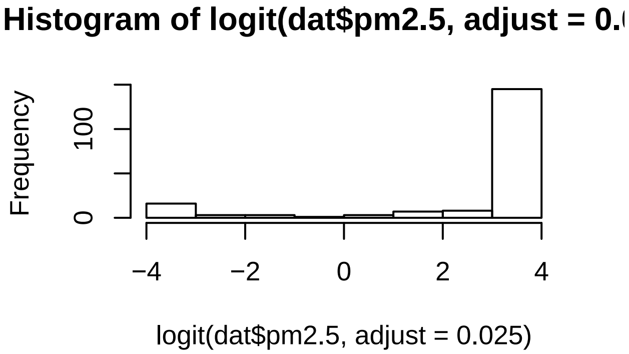


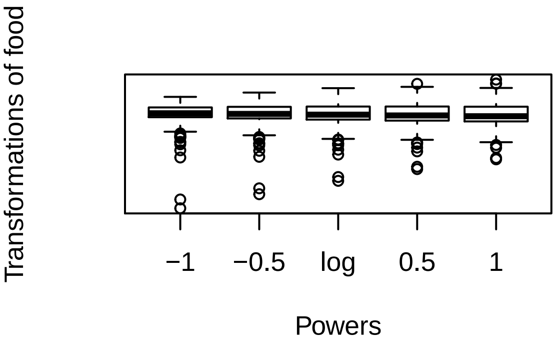




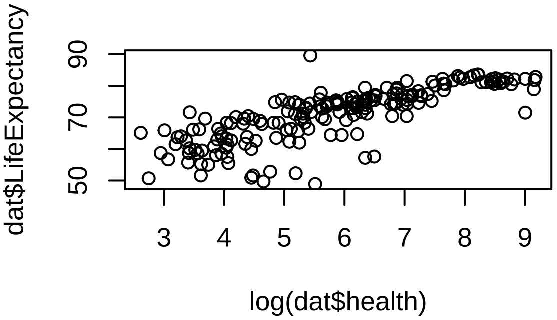


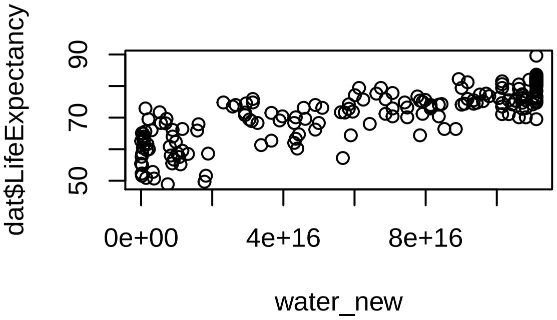


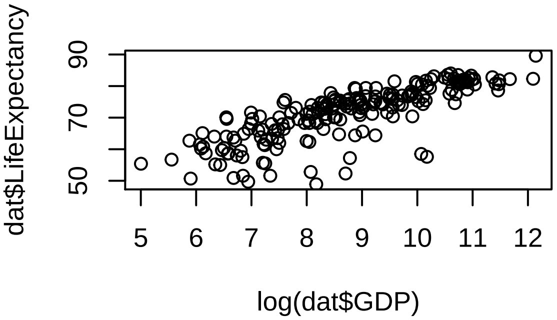


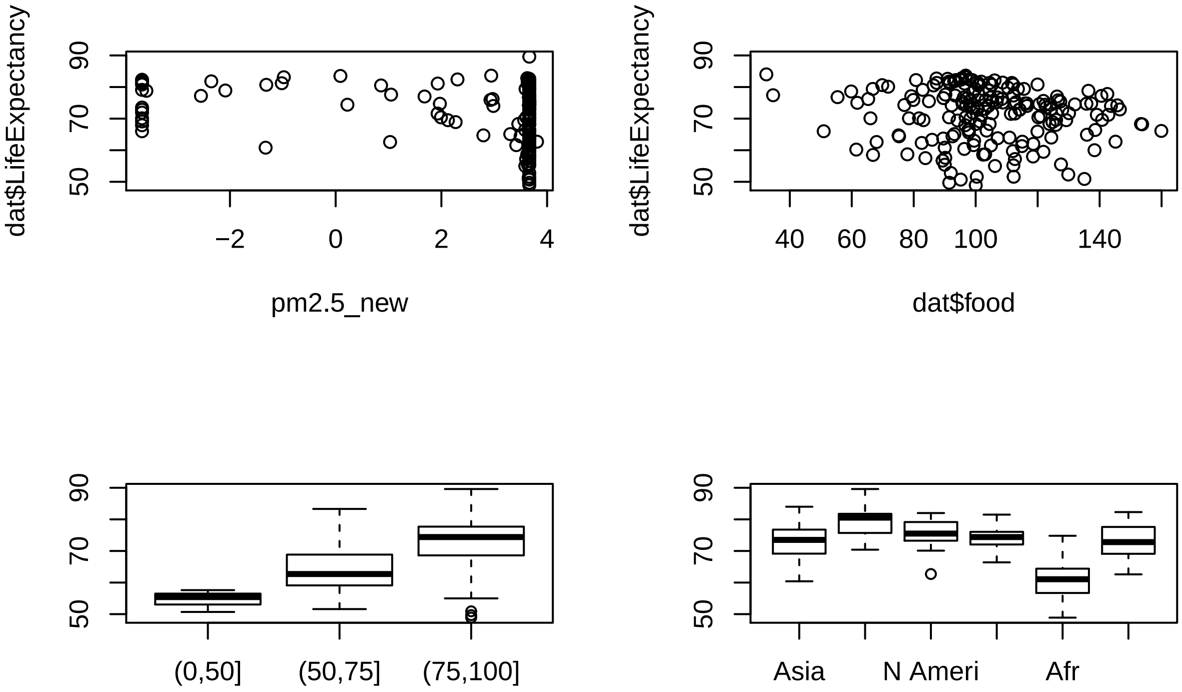


scatter plot after transformation.









3. Regression

(a). check multicollinearity

```
# correlation
cor(na.omit(data[,-c(1,2,6,9)]))
```

```
##             health      water        GDP       pm2.5        food
## health   1.0000000  0.8050738  0.9548396 -0.31296788 -0.15918982
## water    0.8050738  1.0000000  0.7764994 -0.20820046 -0.14621714
## GDP      0.9548396  0.7764994  1.0000000 -0.26555123 -0.19459324
## pm2.5   -0.3129679 -0.2082005 -0.2655512  1.00000000  0.09194482
## food    -0.1591898 -0.1462171 -0.1945932  0.09194482  1.00000000
```

```
# check Generalized VIF
mymodel1 <- lm(exp~health+water+GDP+measles+pm2.5+food+region, data = data)
vif(mymodel1) # VIF > 5 indicates presence of multicollinearity
```

```
##              GVIF Df GVIF^(1/(2*Df))
## health  16.009989  1        4.001248
## water    3.429650  1        1.851931
## GDP     13.115436  1        3.621524
## measles  1.290803  2        1.065896
## pm2.5    2.344783  1        1.531269
## food     1.218473  1        1.103845
## region   5.993727  5        1.196106
```

```
# Model Respecification
# x <- cbind(log(dat$health), water_new, log(dat$GDP), measles_new,
# pm2.5_new, dat$food, as.factor(dat$Region))
x <- na.omit(data[,-c(1,2,6,9)])
```

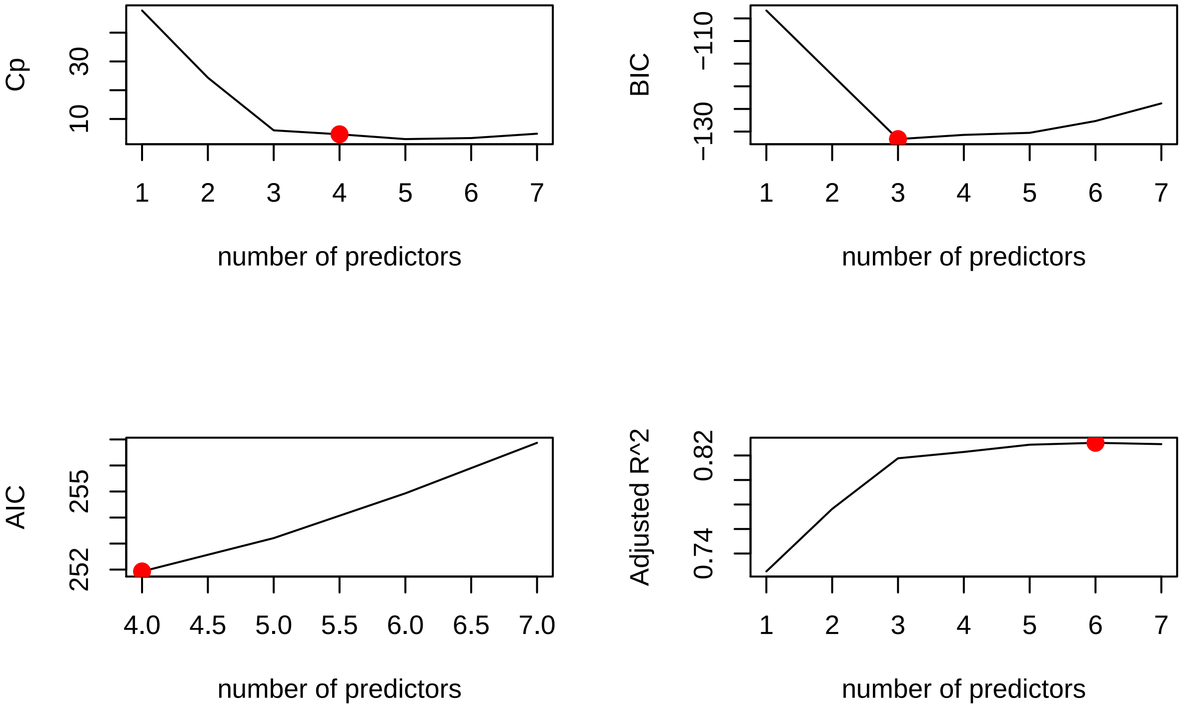
```
pca <- princomp(na.omit(x)) # principle component
summary(pca)
```

```
## Importance of components:
##                           Comp.1       Comp.2       Comp.3       Comp.4
## Standard deviation     4.186471e+16 1.769221e+01 1.126613e+01 1.751881e+00
## Proportion of Variance 1.000000e+00 1.785944e-31 7.241912e-32 1.751109e-33
## Cumulative Proportion  1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##                           Comp.5
## Standard deviation     1.609553e+00
## Proportion of Variance 1.478136e-33
## Cumulative Proportion  1.000000e+00
```

```
round(pca$loadings, 2)
```

```
##
## Loadings:
##        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## health                               1.00
## water  -1.00
## GDP           -0.45   0.70   0.55
## pm2.5          0.30  -0.46   0.84
## food           0.84   0.54
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## SS loadings       1.0  0.998  0.993  1.008    1.0
## Proportion Var    0.2  0.200  0.199  0.202    0.2
## Cumulative Var    0.2  0.400  0.598  0.800    1.0
```

(b). Model Selection



(c). check interaction terms:

```
fit.interact = lm(exp~health*water*measles*region, data = data)
round(summary(fit.interact)$coef, 3)
```

```
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            56.978      9.295   6.130    0.000
## health                                  1.018      1.967   0.518    0.606
## water                                   0.000      0.000  -0.385    0.700
## measles(50,75]                         21.937     43.708   0.502    0.617
## measles(75,100]                         7.574      7.509   1.009    0.315
## region2                                52.035     47.542   1.095    0.276
## region3                               -10.467     25.758  -0.406    0.685
## region4                                15.840     35.264   0.449    0.654
## region5                               -11.560      6.851  -1.687    0.094
## region6                                 6.404     24.964   0.257    0.798
## health:water                            0.000      0.000   1.163    0.247
## health:measles(50,75]                  -4.663     10.443  -0.446    0.656
## health:measles(75,100]                 -0.241      1.582  -0.153    0.879
## water:measles(50,75]                    0.000      0.000  -0.450    0.653
## health:region2                         -7.590      7.267  -1.044    0.298
## health:region3                          3.207      4.643   0.691    0.491
## health:region4                         -1.884      5.735  -0.329    0.743
## health:region5                          0.854      1.506   0.567    0.572
## health:region6                         -1.685      4.932  -0.342    0.733
## water:region2                           0.000      0.000  -1.123    0.263
## water:region3                           0.000      0.000   0.204    0.839
## water:region4                           0.000      0.000  -1.000    0.319
## water:region5                           0.000      0.000   0.086    0.932
## water:region6                           0.000      0.000  -0.372    0.711
## measles(50,75]:region2                -15.885     20.193  -0.787    0.433
## measles(50,75]:region3                  9.051     49.126   0.184    0.854
## measles(50,75]:region5                 15.491     43.871   0.353    0.725
## measles(50,75]:region6               -478.410    490.033  -0.976    0.331
## health:water:measles(50,75]             0.000      0.000   0.507    0.613
## health:water:region2                    0.000      0.000   1.056    0.293
## health:water:region3                    0.000      0.000  -0.611    0.543
## health:water:region4                    0.000      0.000   0.827    0.410
## health:water:region5                    0.000      0.000   0.146    0.884
## health:water:region6                    0.000      0.000   0.396    0.693
## health:measles(50,75]:region3          -3.167     10.777  -0.294    0.769
## health:measles(50,75]:region5          -4.319     10.572  -0.409    0.684
## health:measles(50,75]:region6         106.025    108.722   0.975    0.331
## water:measles(50,75]:region5            0.000      0.000   0.388    0.699
## water:measles(50,75]:region6            0.000      0.000  -0.906    0.366
## health:water:measles(50,75]:region5     0.000      0.000  -0.327    0.744
```

```
#Anova(fit.interact, type = "II")
```

(d). regression

```
##
## Call:
```

```
## lm(formula = exp ~ health + water + measles + region, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.6044  -1.3181   0.0094   1.3761   9.8919
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.937e+01  4.642e+00  10.636  < 2e-16 ***
## health           2.192e+00  4.223e-01   5.191 1.69e-06 ***
## water            6.217e-17  1.733e-17   3.586  0.00059 ***
## measles(50,75]   5.837e+00  3.909e+00   1.493  0.13957
## measles(75,100]  7.605e+00  3.774e+00   2.015  0.04744 *
## region2         -1.514e+00  1.260e+00  -1.201  0.23339
## region3         -1.133e+00  1.399e+00  -0.810  0.42055
## region4         -2.698e+00  1.459e+00  -1.849  0.06838 .
## region5         -6.023e+00  1.323e+00  -4.551 1.99e-05 ***
## region6         -1.595e+00  1.543e+00  -1.033  0.30465
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.5 on 76 degrees of freedom
## Multiple R-squared:  0.8432, Adjusted R-squared:  0.8246
## F-statistic: 45.41 on 9 and 76 DF,  p-value: < 2.2e-16
```

```
## [1] 14.02623
```

health water measles region 5

drop pm2.5, food.

$$LifeExpectancy = 49.37 + 2.192log(health) + 6.217 * 10^{-17}water^9/9$$

$$+\begin{cases} 0, & measles \in (0,50] \\ 5.837, & measles \in (50,75] \\ 7.605, & measles \in (75,100] \end{cases} + \begin{cases} 0, & Asia \\ -1.514, & Europe \\ -1.133, & NorthAmerica \\ -2.698, & SouthAmerica \\ -6.023, & Afica \\ -1.595, & Oceania \end{cases}$$

So, no need for interaction terms.

4. Prediction

examples & CI

```
predict(fit.t, newdata = test[33,], interval = "prediction", level = 0.95) # Canada
```

```
##         fit      lwr      upr
## 79 70.42985 63.01401 77.84569
```

```
test[33,]$exp # actual value
```

```
## [1] 73.1
```

```
predict(fit.t, newdata = test[83,], interval = "prediction", level = 0.95) # UK
```

```
##          fit      lwr      upr
## 190 80.51524 73.3555 87.67499
```

```
test[83,]$exp # actual value
```

```
## [1] 81.1
```

```
predict(fit.t, newdata = test[37,], interval = "prediction", level = 0.95) # Japan
```

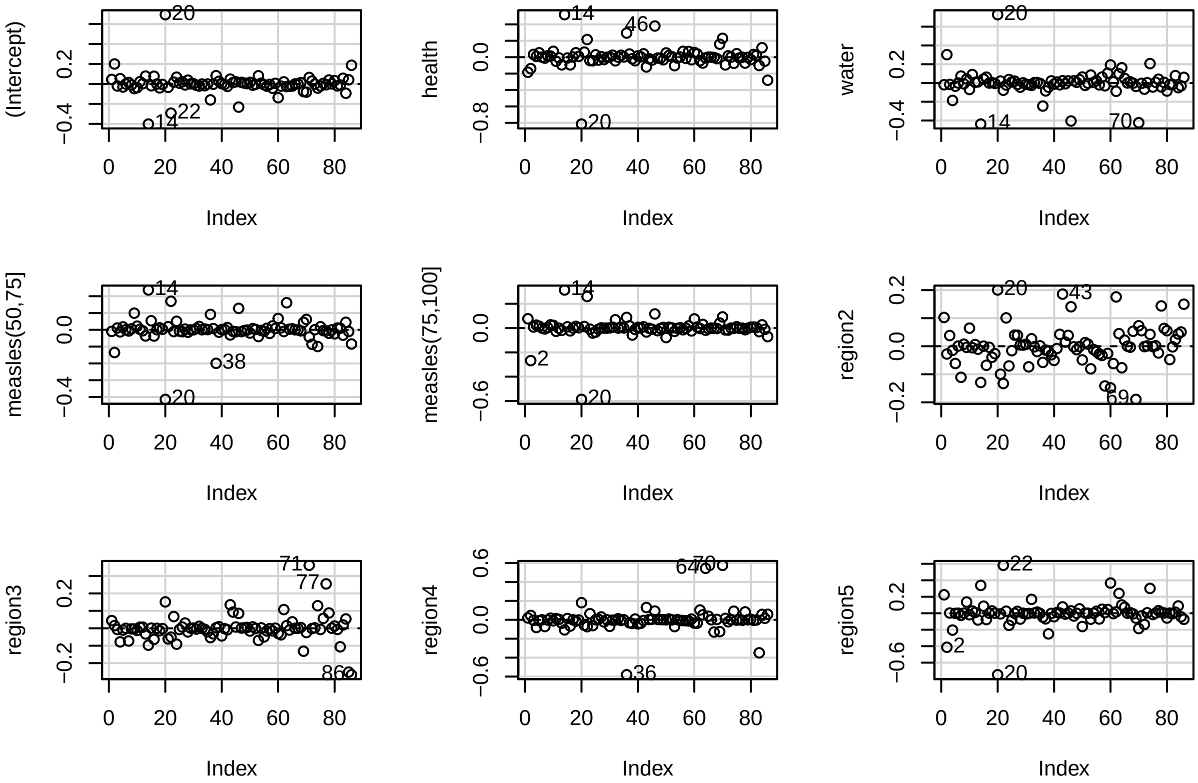
```
##         fit      lwr      upr
## 91 81.89623 74.57761 89.21486
```

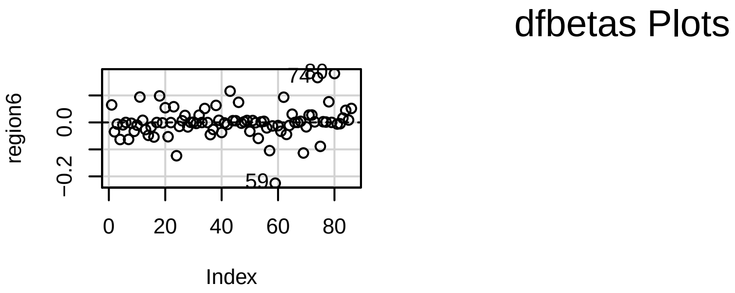
```
test[37,]$exp # actual value
```

```
## [1] 83.6
```

5. Diagnose

```
fit.t = lm(exp~health+water+measles+region, data = train)
## DFBETAS D^{*}
cutoff = 2/sqrt(nrow(train))
dfbetasPlots(fit.t, intercept = T, id.n = 3)
```





```
temp = dfbetas(fit.t)

order(abs(temp[,1]),decreasing = T)[c(1,2)]
```

```
## [1] 20 14
```

```
order(abs(temp[,2]),decreasing = T)[c(1,2)]
```

```
## [1] 20 14
```

```
order(abs(temp[,3]),decreasing = T)[c(1,2)]
```

```
## [1] 20 14
```

```
order(abs(temp[,4]),decreasing = T)[c(1,2)]
```

```
## [1] 20 14
```

```
order(abs(temp[,5]),decreasing = T)[c(1,2)]
```

```
## [1] 20 14
```

```
order(abs(temp[,6]),decreasing = T)[c(1,2)]
```

```
## [1] 20 69
```

```
order(abs(temp[,7]),decreasing = T)[c(1,2)]
```

```
## [1] 71 86
```

```
order(abs(temp[,8]),decreasing = T)[c(1,2)]
```

```
## [1] 36 70
```

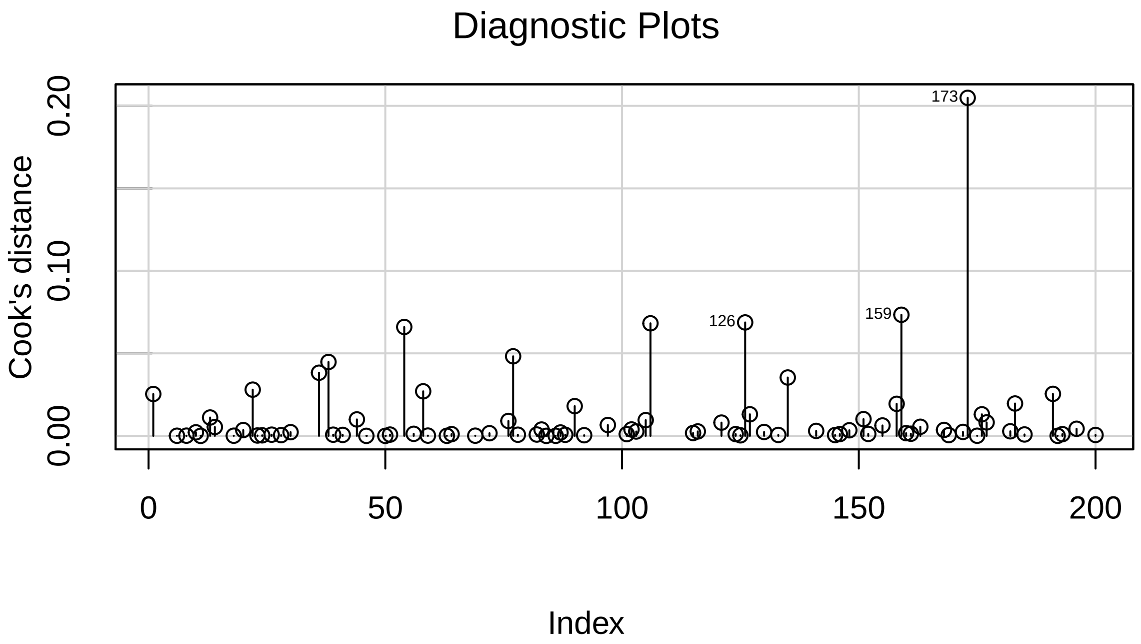
```
order(abs(temp[,9]),decreasing = T)[c(1,2)]
```

```
## [1] 20 22
```

```
order(abs(temp[,10]),decreasing = T)[c(1,2)]
```

```
## [1] 59 80
```

```
## Cook's distance
influenceIndexPlot(fit.t, vars="Cook", id.n = 3, id.cex = 0.5)
```



```
D_i = cooks.distance(fit.t)
order(D_i,decreasing = T)[c(1,2)]
```

```
## [1] 20  2
```

```
D_i[20]
```

```
##       173
## 0.2048794
```

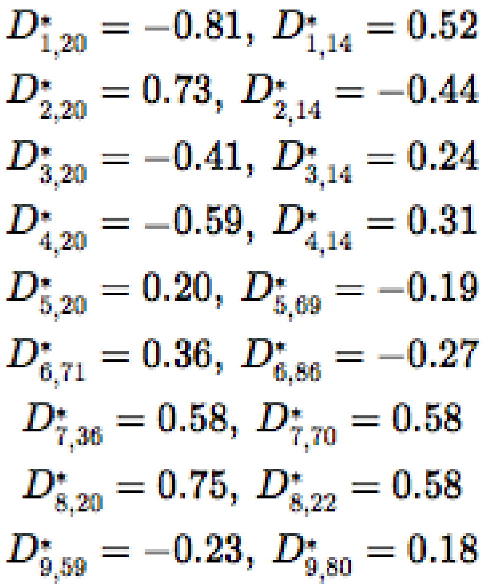
```
D_i[2]
```

```
##        159
## 0.07338397
```

Influence Analysis Summary

Cook's distance $D_{20} = 0.2$ (next largest $D_2 = 0.07$)

DFBETAS



$$D^*_{1,20} = -0.81,\ D^*_{1,14} = 0.52$$
$$D^*_{2,20} = 0.73,\ D^*_{2,14} = -0.44$$
$$D^*_{3,20} = -0.41,\ D^*_{3,14} = 0.24$$
$$D^*_{4,20} = -0.59,\ D^*_{4,14} = 0.31$$
$$D^*_{5,20} = 0.20,\ D^*_{5,69} = -0.19$$
$$D^*_{6,71} = 0.36,\ D^*_{6,86} = -0.27$$
$$D^*_{7,36} = 0.58,\ D^*_{7,70} = 0.58$$
$$D^*_{8,20} = 0.75,\ D^*_{8,22} = 0.58$$
$$D^*_{9,59} = -0.23,\ D^*_{9,80} = 0.18$$

Figure 1:

Overall, observation 20 (Swaziland, country with the lowest life expectancy), 14 (Libya) appear to have the most influence on the results.

appendix

```
dat = read.csv("/Users/tcc/Desktop/Winter/stat423/project/423data.csv")
nrow(dat)
head(dat)
```

```
par(mfrow = c(2,2))
hist(dat$LifeExpectancy)
hist(dat$health)
hist(dat$water)
hist(dat$GDP)
par(mfrow = c(2,2))
hist(dat$measles)
hist(dat$pm2.5)
hist(dat$food)
```

```
par(mfrow = c(2,2))
plot(dat$health, dat$LifeExpectancy)
plot(dat$water, dat$LifeExpectancy)
plot(dat$GDP, dat$LifeExpectancy)
plot(dat$measles, dat$LifeExpectancy)
par(mfrow = c(1,2))
plot(dat$pm2.5, dat$LifeExpectancy)
plot(dat$food, dat$LifeExpectancy)
```

```
par(mfrow = c(2,2))
symbox(~LifeExpectancy, data = dat) # no transformation needed
```

```r
symbox(~health, data = dat) # log
symbox(~water, data = dat) # try positive large p like 9
water_new = (dat$water^9 - 1)/9
boxplot(water_new, main = "boxplot of water with p = 9")

par(mfrow = c(2,2))
symbox(~GDP, data = dat) # log
symbox(~measles, data = dat)

hist(logit(dat$pm2.5, adjust = 0.025))
pm2.5_new = logit(dat$pm2.5, adjust = 0.025)
symbox(~food, data = dat) # no need to change
```

```r
par(mfrow = c(2,2))
plot(log(dat$health), dat$LifeExpectancy)
plot(water_new, dat$LifeExpectancy)
plot(log(dat$GDP), dat$LifeExpectancy)
par(mfrow = c(2,2))
plot(pm2.5_new, dat$LifeExpectancy)
plot(dat$food, dat$LifeExpectancy)


measles_new = cut(dat$measles, breaks = c(0,50,75,100))
boxplot(dat$LifeExpectancy ~ measles_new)
boxplot(dat$LifeExpectancy ~ dat$Region,names = c("Asia","Euro","N Ameri","S Ameri", "Afr", "Ocea"))
```

```r
data = data.frame(country = dat$Country, exp = dat$LifeExpectancy,
                  health = log(dat$health),
                  water = water_new, GDP = log(dat$GDP),
                  measles = measles_new, pm2.5 = pm2.5_new,
                  food = dat$food, region = as.factor(dat$Region))
data_have_na = data
data = data[complete.cases(data), ]
```

```r
# cross-validation
set.seed(123)
n = nrow(data)
w = sample(n,n/2)
train = data[w,-1]
test = data[-w,]

## Model Selection ##
# fit all
regfit.full = regsubsets(exp~health+water+measles+pm2.5+food+GDP+region, data = train, nvmax=7)

# AIC
fit_full = lm(exp~., data=train)
fit_null = lm(exp~1, data=train)
AIC = step(fit_full, scope=list(lower=fit_null), direction="both")
aic = c(256.87, 254.93,253.21,251.93)

num = c(7,6,5,4)

par(mfrow = c(2,2))
```

```r
# Cp
plot(summary(regfit.full)$cp, ylab = "Cp", xlab ="number of predictors", type = "l")
num_var = 1:7
index2=which.min(abs(summary(regfit.full)$cp - (num_var+1)))
points(index2,summary(regfit.full)$cp[index2],col="red",cex=2,pch=20)

# BIC
plot(summary(regfit.full)$bic,ylab="BIC",xlab="number of predictors",type="l")
index3=which.min(summary(regfit.full)$bic)
points(index3,summary(regfit.full)$bic[index3],col="red",cex=2,pch=20)

# AIC
plot(num, aic,ylab = "AIC", xlab ="number of predictors", type = "l")
index4 = which.min(aic)
points(num[index4], aic[index4],col="red",cex=2,pch=20)
# adjusted R^2
plot(summary(regfit.full)$adjr2,ylab="Adjusted R^2",xlab="number of predictors",type="l")
index1=which.max(summary(regfit.full)$adjr2)
points(index1,summary(regfit.full)$adjr2[index1],col="red",cex=2,pch=20)

#coeffcients of the best model obtained#
coef(regfit.full,4)
```

```r
#Total MSE
fit.t = lm(exp~health+water+measles+region, data = train)
summary(fit.t)
pred = predict(fit.t,newdata =test)
mean((pred - test$exp)^2)
```