State-of-Life-Predicting-Life-Expectancy-from-1970s-Census-Data

XunSun

2024-12-14

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(GGally)

## Warning: 程序包'GGally'是用R版本4.4.2 来建造的

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(patchwork)  
library(gt)

## Warning: 程序包'gt'是用R版本4.4.2 来建造的

library(leaps)

## Warning: 程序包'leaps'是用R版本4.4.2 来建造的

library(caret)

## Warning: 程序包'caret'是用R版本4.4.2 来建造的

## 载入需要的程序包：lattice  
##   
## 载入程序包：'caret'  
##   
## The following object is masked from 'package:purrr':  
##   
## lift

library(ggplot2)  
library(dbplyr)

##   
## 载入程序包：'dbplyr'  
##   
## The following objects are masked from 'package:dplyr':  
##   
## ident, sql

R dataset state.x77 from library(faraway) contains information on 50 states from 1970s collected by US Census Bureau. The goal is to predict ‘life expectancy’ using a combination of remaining variables.

library(faraway)

## Warning: 程序包'faraway'是用R版本4.4.2 来建造的

##   
## 载入程序包：'faraway'

## The following object is masked from 'package:lattice':  
##   
## melanoma

## The following object is masked from 'package:GGally':  
##   
## happy

data.state <- as.data.frame(state.x77)|>   
 janitor::clean\_names()   
head(data.state)

## population income illiteracy life\_exp murder hs\_grad frost area  
## Alabama 3615 3624 2.1 69.05 15.1 41.3 20 50708  
## Alaska 365 6315 1.5 69.31 11.3 66.7 152 566432  
## Arizona 2212 4530 1.8 70.55 7.8 58.1 15 113417  
## Arkansas 2110 3378 1.9 70.66 10.1 39.9 65 51945  
## California 21198 5114 1.1 71.71 10.3 62.6 20 156361  
## Colorado 2541 4884 0.7 72.06 6.8 63.9 166 103766

view(data.state)

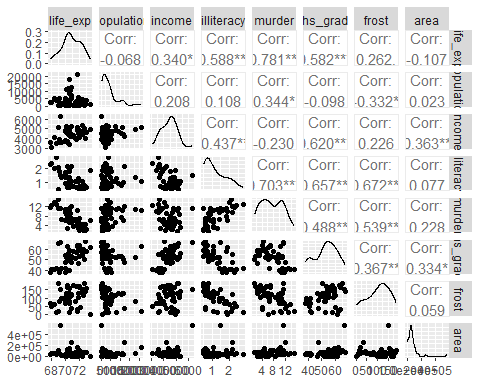
1. Provide descriptive statistics for all variables of interest – no test required

summary(data.state)

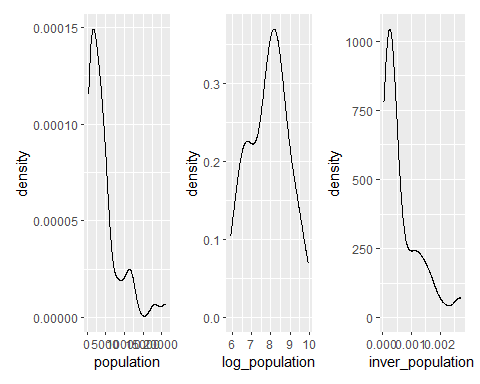
## population income illiteracy life\_exp   
## Min. : 365 Min. :3098 Min. :0.500 Min. :67.96   
## 1st Qu.: 1080 1st Qu.:3993 1st Qu.:0.625 1st Qu.:70.12   
## Median : 2838 Median :4519 Median :0.950 Median :70.67   
## Mean : 4246 Mean :4436 Mean :1.170 Mean :70.88   
## 3rd Qu.: 4968 3rd Qu.:4814 3rd Qu.:1.575 3rd Qu.:71.89   
## Max. :21198 Max. :6315 Max. :2.800 Max. :73.60   
## murder hs\_grad frost area   
## Min. : 1.400 Min. :37.80 Min. : 0.00 Min. : 1049   
## 1st Qu.: 4.350 1st Qu.:48.05 1st Qu.: 66.25 1st Qu.: 36985   
## Median : 6.850 Median :53.25 Median :114.50 Median : 54277   
## Mean : 7.378 Mean :53.11 Mean :104.46 Mean : 70736   
## 3rd Qu.:10.675 3rd Qu.:59.15 3rd Qu.:139.75 3rd Qu.: 81163   
## Max. :15.100 Max. :67.30 Max. :188.00 Max. :566432

Examine exploratory plots, e.g., scatter plots, histograms, boxplots to get a sense of the data and possible variable transformations.

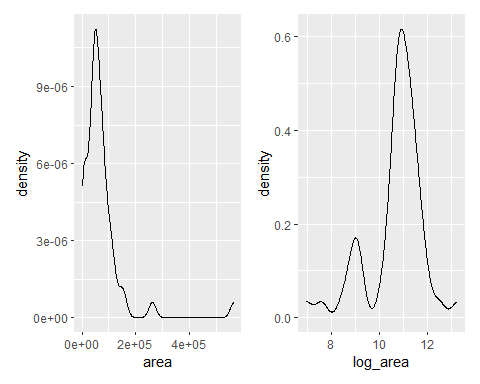
data.state |>  
 relocate(`life\_exp`) |>  
 ggpairs()

 ## look for appropriate transformations

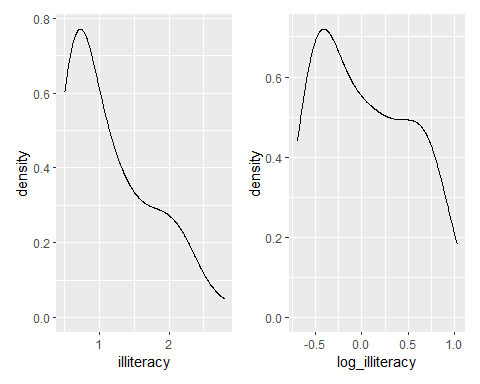
density\_plot\_population <-data.state |>  
 ggplot(aes(x = population)) +  
 geom\_density()  
  
density\_plot\_logpopulation <-data.state |>  
 mutate(log\_population = log(population)) |>  
 ggplot(aes(x = log\_population)) + geom\_density()  
  
density\_plot\_inver\_population <-data.state |>  
 mutate(inver\_population = 1/(population)) |>  
 ggplot(aes(x = inver\_population)) + geom\_density()  
  
  
density\_plot\_population + density\_plot\_logpopulation + density\_plot\_inver\_population



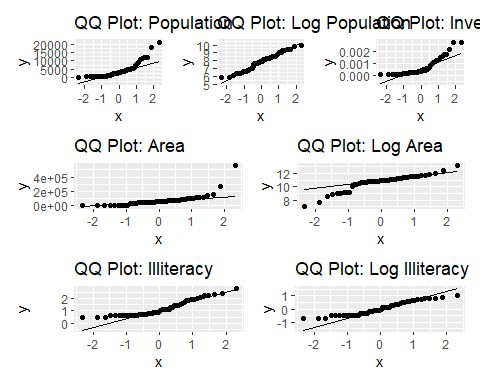
density\_plot\_area <- data.state |>  
 ggplot(aes(x = area)) +  
 geom\_density()  
  
density\_plot\_logarea <- data.state |>  
 mutate(log\_area = log(area)) |>  
 ggplot(aes(x = log\_area)) + geom\_density()  
  
density\_plot\_area + density\_plot\_logarea



density\_plot\_illiteracy <- data.state |>  
 ggplot(aes(x = illiteracy)) +  
 geom\_density()  
  
density\_plot\_logilliteracy <- data.state |>  
 mutate(log\_illiteracy = log(illiteracy)) |>  
 ggplot(aes(x = log\_illiteracy)) + geom\_density()  
  
density\_plot\_illiteracy + density\_plot\_logilliteracy



qq\_population <- data.state |>  
 ggplot(aes(sample = population)) +  
 stat\_qq() +  
 stat\_qq\_line() +  
 ggtitle("QQ Plot: Population")  
  
qq\_log\_population <- data.state |>  
 mutate(log\_population = log(population)) |>  
 ggplot(aes(sample = log\_population)) +  
 stat\_qq() +  
 stat\_qq\_line() +  
 ggtitle("QQ Plot: Log Population")  
  
qq\_inver\_population <- data.state |>  
 mutate(inver\_population = 1 / (population)) |>  
 ggplot(aes(sample = inver\_population)) +  
 stat\_qq() +  
 stat\_qq\_line() +  
 ggtitle("QQ Plot: Inverse Population")  
  
# QQ Plots for Area  
qq\_area <- data.state |>  
 ggplot(aes(sample = area)) +  
 stat\_qq() +  
 stat\_qq\_line() +  
 ggtitle("QQ Plot: Area")  
  
qq\_log\_area <- data.state |>  
 mutate(log\_area = log(area)) |>  
 ggplot(aes(sample = log\_area)) +  
 stat\_qq() +  
 stat\_qq\_line() +  
 ggtitle("QQ Plot: Log Area")  
  
# QQ Plots for Illiteracy  
qq\_illiteracy <- data.state |>  
 ggplot(aes(sample = illiteracy)) +  
 stat\_qq() +  
 stat\_qq\_line() +  
 ggtitle("QQ Plot: Illiteracy")  
  
qq\_log\_illiteracy <- data.state |>  
 mutate(log\_illiteracy = log(illiteracy)) |>  
 ggplot(aes(sample = log\_illiteracy)) +  
 stat\_qq() +  
 stat\_qq\_line() +  
 ggtitle("QQ Plot: Log Illiteracy")  
  
# Combine QQ Plots with Patchwork for Comparison  
library(patchwork)  
  
(qq\_population | qq\_log\_population | qq\_inver\_population) /   
(qq\_area | qq\_log\_area) /   
(qq\_illiteracy | qq\_log\_illiteracy)



data.state <- data.state |>  
 mutate(log\_Population = log(population)) |>  
 select(-population)  
view(data.state)

# Define the response variable and dataset  
response <- "life\_expectancy" # Replace with your actual response variable  
data <- data.state # Your dataset  
  
# Null model  
null\_model <- lm(life\_exp ~ 1, data = data)  
  
# Full model (all predictors)  
full\_model <- lm(life\_exp ~ ., data = data)

library(MASS)

## Warning: 程序包'MASS'是用R版本4.4.2 来建造的

##   
## 载入程序包：'MASS'

## The following object is masked from 'package:patchwork':  
##   
## area

## The following object is masked from 'package:dplyr':  
##   
## select

library(olsrr)

## Warning: 程序包'olsrr'是用R版本4.4.2 来建造的

##   
## 载入程序包：'olsrr'

## The following object is masked from 'package:MASS':  
##   
## cement

## The following object is masked from 'package:faraway':  
##   
## hsb

## The following object is masked from 'package:datasets':  
##   
## rivers

# Forward Selection  
forward\_aic <- stepAIC(null\_model,   
 scope = list(lower = null\_model, upper = full\_model), direction = "forward")

## Start: AIC=30.44  
## life\_exp ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + murder 1 53.838 34.461 -14.609  
## + illiteracy 1 30.578 57.721 11.179  
## + hs\_grad 1 29.931 58.368 11.737  
## + income 1 10.223 78.076 26.283  
## + frost 1 6.064 82.235 28.878  
## <none> 88.299 30.435  
## + log\_Population 1 1.054 87.245 31.835  
## + area 1 1.017 87.282 31.856  
##   
## Step: AIC=-14.61  
## life\_exp ~ murder  
##   
## Df Sum of Sq RSS AIC  
## + hs\_grad 1 4.6910 29.770 -19.925  
## + frost 1 3.1346 31.327 -17.378  
## + log\_Population 1 2.9854 31.476 -17.140  
## + income 1 2.4047 32.057 -16.226  
## <none> 34.461 -14.609  
## + area 1 0.4697 33.992 -13.295  
## + illiteracy 1 0.2732 34.188 -13.007  
##   
## Step: AIC=-19.93  
## life\_exp ~ murder + hs\_grad  
##   
## Df Sum of Sq RSS AIC  
## + log\_Population 1 4.6350 25.135 -26.387  
## + frost 1 4.3987 25.372 -25.920  
## <none> 29.770 -19.925  
## + illiteracy 1 0.4419 29.328 -18.673  
## + area 1 0.2775 29.493 -18.394  
## + income 1 0.1022 29.668 -18.097  
##   
## Step: AIC=-26.39  
## life\_exp ~ murder + hs\_grad + log\_Population  
##   
## Df Sum of Sq RSS AIC  
## + frost 1 2.21416 22.921 -28.998  
## + illiteracy 1 1.10754 24.028 -26.640  
## <none> 25.135 -26.387  
## + income 1 0.11819 25.017 -24.623  
## + area 1 0.00175 25.134 -24.391  
##   
## Step: AIC=-29  
## life\_exp ~ murder + hs\_grad + log\_Population + frost  
##   
## Df Sum of Sq RSS AIC  
## <none> 22.921 -28.998  
## + illiteracy 1 0.051595 22.870 -27.111  
## + area 1 0.015956 22.905 -27.033  
## + income 1 0.010673 22.911 -27.021

# Summary of the selected model  
summary(forward\_aic)

##   
## Call:  
## lm(formula = life\_exp ~ murder + hs\_grad + log\_Population + frost,   
## data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.41760 -0.43880 0.02539 0.52066 1.63048   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 68.720810 1.416828 48.503 < 2e-16 \*\*\*  
## murder -0.290016 0.035440 -8.183 1.87e-10 \*\*\*  
## hs\_grad 0.054550 0.014758 3.696 0.000591 \*\*\*  
## log\_Population 0.246836 0.112539 2.193 0.033491 \*   
## frost -0.005174 0.002482 -2.085 0.042779 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7137 on 45 degrees of freedom  
## Multiple R-squared: 0.7404, Adjusted R-squared: 0.7173   
## F-statistic: 32.09 on 4 and 45 DF, p-value: 1.17e-12

n <- nrow(data.state)   
forward\_bic <- step(null\_model,   
 scope = list(lower = null\_model, upper = full\_model),   
 direction = "forward",   
 k = log(n))

## Start: AIC=32.35  
## life\_exp ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + murder 1 53.838 34.461 -10.785  
## + illiteracy 1 30.578 57.721 15.004  
## + hs\_grad 1 29.931 58.368 15.561  
## + income 1 10.223 78.076 30.107  
## <none> 88.299 32.347  
## + frost 1 6.064 82.235 32.702  
## + log\_Population 1 1.054 87.245 35.659  
## + area 1 1.017 87.282 35.680  
##   
## Step: AIC=-10.79  
## life\_exp ~ murder  
##   
## Df Sum of Sq RSS AIC  
## + hs\_grad 1 4.6910 29.770 -14.1894  
## + frost 1 3.1346 31.327 -11.6415  
## + log\_Population 1 2.9854 31.476 -11.4039  
## <none> 34.461 -10.7852  
## + income 1 2.4047 32.057 -10.4900  
## + area 1 0.4697 33.992 -7.5593  
## + illiteracy 1 0.2732 34.188 -7.2712  
##   
## Step: AIC=-14.19  
## life\_exp ~ murder + hs\_grad  
##   
## Df Sum of Sq RSS AIC  
## + log\_Population 1 4.6350 25.135 -18.739  
## + frost 1 4.3987 25.372 -18.271  
## <none> 29.770 -14.189  
## + illiteracy 1 0.4419 29.328 -11.025  
## + area 1 0.2775 29.493 -10.746  
## + income 1 0.1022 29.668 -10.449  
##   
## Step: AIC=-18.74  
## life\_exp ~ murder + hs\_grad + log\_Population  
##   
## Df Sum of Sq RSS AIC  
## + frost 1 2.21416 22.921 -19.438  
## <none> 25.135 -18.739  
## + illiteracy 1 1.10754 24.028 -17.080  
## + income 1 0.11819 25.017 -15.063  
## + area 1 0.00175 25.134 -14.831  
##   
## Step: AIC=-19.44  
## life\_exp ~ murder + hs\_grad + log\_Population + frost  
##   
## Df Sum of Sq RSS AIC  
## <none> 22.921 -19.438  
## + illiteracy 1 0.051595 22.870 -15.639  
## + area 1 0.015956 22.905 -15.561  
## + income 1 0.010673 22.911 -15.549

Backward Selection

# Backward selection based on AIC  
backward\_aic <- step(full\_model,   
 direction = "backward",trace = TRUE)

## Start: AIC=-23.15  
## life\_exp ~ income + illiteracy + murder + hs\_grad + frost + area +   
## log\_Population  
##   
## Df Sum of Sq RSS AIC  
## - area 1 0.0092 22.859 -25.134  
## - income 1 0.0159 22.866 -25.120  
## - illiteracy 1 0.0359 22.885 -25.076  
## <none> 22.850 -23.154  
## - frost 1 1.0933 23.943 -22.817  
## - log\_Population 1 2.1947 25.044 -20.569  
## - hs\_grad 1 3.1607 26.010 -18.677  
## - murder 1 23.6107 46.460 10.329  
##   
## Step: AIC=-25.13  
## life\_exp ~ income + illiteracy + murder + hs\_grad + frost + log\_Population  
##   
## Df Sum of Sq RSS AIC  
## - income 1 0.0109 22.870 -27.111  
## - illiteracy 1 0.0518 22.911 -27.021  
## <none> 22.859 -25.134  
## - frost 1 1.1073 23.966 -24.769  
## - log\_Population 1 2.1994 25.058 -22.541  
## - hs\_grad 1 3.8468 26.706 -19.358  
## - murder 1 26.7410 49.600 11.598  
##   
## Step: AIC=-27.11  
## life\_exp ~ illiteracy + murder + hs\_grad + frost + log\_Population  
##   
## Df Sum of Sq RSS AIC  
## - illiteracy 1 0.0516 22.921 -28.9980  
## <none> 22.870 -27.1107  
## - frost 1 1.1582 24.028 -26.6405  
## - log\_Population 1 2.3302 25.200 -24.2594  
## - hs\_grad 1 5.2719 28.141 -18.7389  
## - murder 1 26.9930 49.863 9.8624  
##   
## Step: AIC=-29  
## life\_exp ~ murder + hs\_grad + frost + log\_Population  
##   
## Df Sum of Sq RSS AIC  
## <none> 22.921 -28.998  
## - frost 1 2.214 25.135 -26.387  
## - log\_Population 1 2.450 25.372 -25.920  
## - hs\_grad 1 6.959 29.881 -17.741  
## - murder 1 34.109 57.031 14.578

# Display summary of the final model  
summary(backward\_aic)

##   
## Call:  
## lm(formula = life\_exp ~ murder + hs\_grad + frost + log\_Population,   
## data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.41760 -0.43880 0.02539 0.52066 1.63048   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 68.720810 1.416828 48.503 < 2e-16 \*\*\*  
## murder -0.290016 0.035440 -8.183 1.87e-10 \*\*\*  
## hs\_grad 0.054550 0.014758 3.696 0.000591 \*\*\*  
## frost -0.005174 0.002482 -2.085 0.042779 \*   
## log\_Population 0.246836 0.112539 2.193 0.033491 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7137 on 45 degrees of freedom  
## Multiple R-squared: 0.7404, Adjusted R-squared: 0.7173   
## F-statistic: 32.09 on 4 and 45 DF, p-value: 1.17e-12

# Number of observations  
n <- nrow(data.state)  
  
# Backward selection based on BIC  
backward\_bic <- step(full\_model,   
 direction = "backward",   
 k = log(n),   
 trace = TRUE)

## Start: AIC=-7.86  
## life\_exp ~ income + illiteracy + murder + hs\_grad + frost + area +   
## log\_Population  
##   
## Df Sum of Sq RSS AIC  
## - area 1 0.0092 22.859 -11.7503  
## - income 1 0.0159 22.866 -11.7355  
## - illiteracy 1 0.0359 22.885 -11.6919  
## - frost 1 1.0933 23.943 -9.4333  
## <none> 22.850 -7.8583  
## - log\_Population 1 2.1947 25.044 -7.1846  
## - hs\_grad 1 3.1607 26.010 -5.2923  
## - murder 1 23.6107 46.460 23.7129  
##   
## Step: AIC=-11.75  
## life\_exp ~ income + illiteracy + murder + hs\_grad + frost + log\_Population  
##   
## Df Sum of Sq RSS AIC  
## - income 1 0.0109 22.870 -15.6385  
## - illiteracy 1 0.0518 22.911 -15.5491  
## - frost 1 1.1073 23.966 -13.2970  
## <none> 22.859 -11.7503  
## - log\_Population 1 2.1994 25.058 -11.0691  
## - hs\_grad 1 3.8468 26.706 -7.8855  
## - murder 1 26.7410 49.600 23.0703  
##   
## Step: AIC=-15.64  
## life\_exp ~ illiteracy + murder + hs\_grad + frost + log\_Population  
##   
## Df Sum of Sq RSS AIC  
## - illiteracy 1 0.0516 22.921 -19.4379  
## - frost 1 1.1582 24.028 -17.0804  
## <none> 22.870 -15.6385  
## - log\_Population 1 2.3302 25.200 -14.6993  
## - hs\_grad 1 5.2719 28.141 -9.1788  
## - murder 1 26.9930 49.863 19.4225  
##   
## Step: AIC=-19.44  
## life\_exp ~ murder + hs\_grad + frost + log\_Population  
##   
## Df Sum of Sq RSS AIC  
## <none> 22.921 -19.438  
## - frost 1 2.214 25.135 -18.739  
## - log\_Population 1 2.450 25.372 -18.271  
## - hs\_grad 1 6.959 29.881 -10.093  
## - murder 1 34.109 57.031 22.226

# Display summary of the final model  
summary(backward\_bic)

##   
## Call:  
## lm(formula = life\_exp ~ murder + hs\_grad + frost + log\_Population,   
## data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.41760 -0.43880 0.02539 0.52066 1.63048   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 68.720810 1.416828 48.503 < 2e-16 \*\*\*  
## murder -0.290016 0.035440 -8.183 1.87e-10 \*\*\*  
## hs\_grad 0.054550 0.014758 3.696 0.000591 \*\*\*  
## frost -0.005174 0.002482 -2.085 0.042779 \*   
## log\_Population 0.246836 0.112539 2.193 0.033491 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7137 on 45 degrees of freedom  
## Multiple R-squared: 0.7404, Adjusted R-squared: 0.7173   
## F-statistic: 32.09 on 4 and 45 DF, p-value: 1.17e-12

# Stepwise selection (both directions, default AIC)  
stepwise\_aic <- step(full\_model,   
 direction = "both",trace = TRUE)

## Start: AIC=-23.15  
## life\_exp ~ income + illiteracy + murder + hs\_grad + frost + area +   
## log\_Population  
##   
## Df Sum of Sq RSS AIC  
## - area 1 0.0092 22.859 -25.134  
## - income 1 0.0159 22.866 -25.120  
## - illiteracy 1 0.0359 22.885 -25.076  
## <none> 22.850 -23.154  
## - frost 1 1.0933 23.943 -22.817  
## - log\_Population 1 2.1947 25.044 -20.569  
## - hs\_grad 1 3.1607 26.010 -18.677  
## - murder 1 23.6107 46.460 10.329  
##   
## Step: AIC=-25.13  
## life\_exp ~ income + illiteracy + murder + hs\_grad + frost + log\_Population  
##   
## Df Sum of Sq RSS AIC  
## - income 1 0.0109 22.870 -27.111  
## - illiteracy 1 0.0518 22.911 -27.021  
## <none> 22.859 -25.134  
## - frost 1 1.1073 23.966 -24.769  
## + area 1 0.0092 22.850 -23.154  
## - log\_Population 1 2.1994 25.058 -22.541  
## - hs\_grad 1 3.8468 26.706 -19.358  
## - murder 1 26.7410 49.600 11.598  
##   
## Step: AIC=-27.11  
## life\_exp ~ illiteracy + murder + hs\_grad + frost + log\_Population  
##   
## Df Sum of Sq RSS AIC  
## - illiteracy 1 0.0516 22.921 -28.9980  
## <none> 22.870 -27.1107  
## - frost 1 1.1582 24.028 -26.6405  
## + income 1 0.0109 22.859 -25.1344  
## + area 1 0.0041 22.866 -25.1197  
## - log\_Population 1 2.3302 25.200 -24.2594  
## - hs\_grad 1 5.2719 28.141 -18.7389  
## - murder 1 26.9930 49.863 9.8624  
##   
## Step: AIC=-29  
## life\_exp ~ murder + hs\_grad + frost + log\_Population  
##   
## Df Sum of Sq RSS AIC  
## <none> 22.921 -28.998  
## + illiteracy 1 0.052 22.870 -27.111  
## + area 1 0.016 22.905 -27.033  
## + income 1 0.011 22.911 -27.021  
## - frost 1 2.214 25.135 -26.387  
## - log\_Population 1 2.450 25.372 -25.920  
## - hs\_grad 1 6.959 29.881 -17.741  
## - murder 1 34.109 57.031 14.578

# Summary of the final model  
summary(stepwise\_aic)

##   
## Call:  
## lm(formula = life\_exp ~ murder + hs\_grad + frost + log\_Population,   
## data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.41760 -0.43880 0.02539 0.52066 1.63048   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 68.720810 1.416828 48.503 < 2e-16 \*\*\*  
## murder -0.290016 0.035440 -8.183 1.87e-10 \*\*\*  
## hs\_grad 0.054550 0.014758 3.696 0.000591 \*\*\*  
## frost -0.005174 0.002482 -2.085 0.042779 \*   
## log\_Population 0.246836 0.112539 2.193 0.033491 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7137 on 45 degrees of freedom  
## Multiple R-squared: 0.7404, Adjusted R-squared: 0.7173   
## F-statistic: 32.09 on 4 and 45 DF, p-value: 1.17e-12

# Number of observations  
n <- nrow(data.state)  
  
# Stepwise selection (both directions, BIC)  
stepwise\_bic <- step(full\_model,   
 direction = "both",   
 k = log(n),trace = TRUE) # BIC penalty

## Start: AIC=-7.86  
## life\_exp ~ income + illiteracy + murder + hs\_grad + frost + area +   
## log\_Population  
##   
## Df Sum of Sq RSS AIC  
## - area 1 0.0092 22.859 -11.7503  
## - income 1 0.0159 22.866 -11.7355  
## - illiteracy 1 0.0359 22.885 -11.6919  
## - frost 1 1.0933 23.943 -9.4333  
## <none> 22.850 -7.8583  
## - log\_Population 1 2.1947 25.044 -7.1846  
## - hs\_grad 1 3.1607 26.010 -5.2923  
## - murder 1 23.6107 46.460 23.7129  
##   
## Step: AIC=-11.75  
## life\_exp ~ income + illiteracy + murder + hs\_grad + frost + log\_Population  
##   
## Df Sum of Sq RSS AIC  
## - income 1 0.0109 22.870 -15.6385  
## - illiteracy 1 0.0518 22.911 -15.5491  
## - frost 1 1.1073 23.966 -13.2970  
## <none> 22.859 -11.7503  
## - log\_Population 1 2.1994 25.058 -11.0691  
## - hs\_grad 1 3.8468 26.706 -7.8855  
## + area 1 0.0092 22.850 -7.8583  
## - murder 1 26.7410 49.600 23.0703  
##   
## Step: AIC=-15.64  
## life\_exp ~ illiteracy + murder + hs\_grad + frost + log\_Population  
##   
## Df Sum of Sq RSS AIC  
## - illiteracy 1 0.0516 22.921 -19.4379  
## - frost 1 1.1582 24.028 -17.0804  
## <none> 22.870 -15.6385  
## - log\_Population 1 2.3302 25.200 -14.6993  
## + income 1 0.0109 22.859 -11.7503  
## + area 1 0.0041 22.866 -11.7355  
## - hs\_grad 1 5.2719 28.141 -9.1788  
## - murder 1 26.9930 49.863 19.4225  
##   
## Step: AIC=-19.44  
## life\_exp ~ murder + hs\_grad + frost + log\_Population  
##   
## Df Sum of Sq RSS AIC  
## <none> 22.921 -19.438  
## - frost 1 2.214 25.135 -18.739  
## - log\_Population 1 2.450 25.372 -18.271  
## + illiteracy 1 0.052 22.870 -15.639  
## + area 1 0.016 22.905 -15.561  
## + income 1 0.011 22.911 -15.549  
## - hs\_grad 1 6.959 29.881 -10.093  
## - murder 1 34.109 57.031 22.226

# Summary of the final model  
summary(stepwise\_bic)

##   
## Call:  
## lm(formula = life\_exp ~ murder + hs\_grad + frost + log\_Population,   
## data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.41760 -0.43880 0.02539 0.52066 1.63048   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 68.720810 1.416828 48.503 < 2e-16 \*\*\*  
## murder -0.290016 0.035440 -8.183 1.87e-10 \*\*\*  
## hs\_grad 0.054550 0.014758 3.696 0.000591 \*\*\*  
## frost -0.005174 0.002482 -2.085 0.042779 \*   
## log\_Population 0.246836 0.112539 2.193 0.033491 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7137 on 45 degrees of freedom  
## Multiple R-squared: 0.7404, Adjusted R-squared: 0.7173   
## F-statistic: 32.09 on 4 and 45 DF, p-value: 1.17e-12

library(olsrr)  
  
# Perform stepwise selection based on p-values (default uses AIC)  
stepwise\_model <- ols\_step\_both\_p(full\_model,   
 pent = 0.05, # Entry significance level  
 prem = 0.05) # Removal significance level  
  
# Print stepwise selection results  
print(stepwise\_model)

##   
##   
## Stepwise Summary   
## ---------------------------------------------------------------------------------  
## Step Variable AIC SBC SBIC R2 Adj. R2   
## ---------------------------------------------------------------------------------  
## 0 Base Model 174.329 178.153 30.094 0.00000 0.00000   
## 1 murder (+) 129.285 135.021 -13.541 0.60972 0.60159   
## 2 hs\_grad (+) 123.968 131.616 -18.458 0.66285 0.64850   
## 3 log\_Population (+) 117.507 127.067 -23.743 0.71534 0.69677   
## 4 frost (+) 114.896 126.368 -25.200 0.74041 0.71734   
## ---------------------------------------------------------------------------------  
##   
## Final Model Output   
## ------------------  
##   
## Model Summary   
## ---------------------------------------------------------------  
## R 0.860 RMSE 0.677   
## R-Squared 0.740 MSE 0.458   
## Adj. R-Squared 0.717 Coef. Var 1.007   
## Pred R-Squared 0.659 AIC 114.896   
## MAE 0.571 SBC 126.368   
## ---------------------------------------------------------------  
## RMSE: Root Mean Square Error   
## MSE: Mean Square Error   
## MAE: Mean Absolute Error   
## AIC: Akaike Information Criteria   
## SBC: Schwarz Bayesian Criteria   
##   
## ANOVA   
## -------------------------------------------------------------------  
## Sum of   
## Squares DF Mean Square F Sig.   
## -------------------------------------------------------------------  
## Regression 65.378 4 16.344 32.088 0.0000   
## Residual 22.921 45 0.509   
## Total 88.299 49   
## -------------------------------------------------------------------  
##   
## Parameter Estimates   
## -------------------------------------------------------------------------------------------  
## model Beta Std. Error Std. Beta t Sig lower upper   
## -------------------------------------------------------------------------------------------  
## (Intercept) 68.721 1.417 48.503 0.000 65.867 71.574   
## murder -0.290 0.035 -0.798 -8.183 0.000 -0.361 -0.219   
## hs\_grad 0.055 0.015 0.328 3.696 0.001 0.025 0.084   
## log\_Population 0.247 0.113 0.192 2.193 0.033 0.020 0.474   
## frost -0.005 0.002 -0.200 -2.085 0.043 -0.010 0.000   
## -------------------------------------------------------------------------------------------

library(modelsummary)

## Warning: 程序包'modelsummary'是用R版本4.4.2 来建造的

## `modelsummary` 2.0.0 now uses `tinytable` as its default table-drawing  
## backend. Learn more at: https://vincentarelbundock.github.io/tinytable/  
##   
## Revert to `kableExtra` for one session:  
##   
## options(modelsummary\_factory\_default = 'kableExtra')  
## options(modelsummary\_factory\_latex = 'kableExtra')  
## options(modelsummary\_factory\_html = 'kableExtra')  
##   
## Silence this message forever:  
##   
## config\_modelsummary(startup\_message = FALSE)

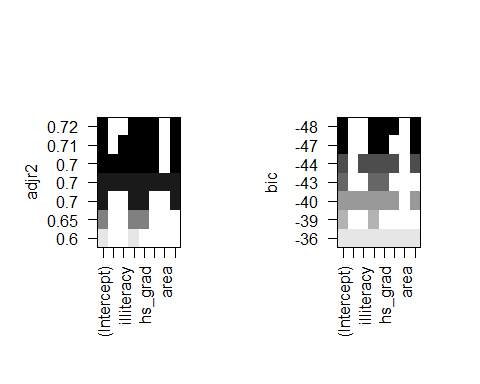
modelsummary(list(stepwise\_AIC\_Model = stepwise\_aic,stepwisw\_BIC\_Model = stepwise\_bic,backward\_AIC\_Model = backward\_aic, backward\_BIC\_Model = backward\_aic,forkward\_AIC\_Model = forward\_aic, forward\_BIC\_Model = forward\_aic),   
 output = "markdown",   
 statistic = c("std.error", "p.value"))

|  | stepwise\_AIC\_Model | stepwisw\_BIC\_Model | backward\_AIC\_Model | backward\_BIC\_Model | forkward\_AIC\_Model | forward\_BIC\_Model |
| --- | --- | --- | --- | --- | --- | --- |
| (Intercept) | 68.721 | 68.721 | 68.721 | 68.721 | 68.721 | 68.721 |
|  | (1.417) | (1.417) | (1.417) | (1.417) | (1.417) | (1.417) |
|  | (<0.001) | (<0.001) | (<0.001) | (<0.001) | (<0.001) | (<0.001) |
| murder | -0.290 | -0.290 | -0.290 | -0.290 | -0.290 | -0.290 |
|  | (0.035) | (0.035) | (0.035) | (0.035) | (0.035) | (0.035) |
|  | (<0.001) | (<0.001) | (<0.001) | (<0.001) | (<0.001) | (<0.001) |
| hs\_grad | 0.055 | 0.055 | 0.055 | 0.055 | 0.055 | 0.055 |
|  | (0.015) | (0.015) | (0.015) | (0.015) | (0.015) | (0.015) |
|  | (<0.001) | (<0.001) | (<0.001) | (<0.001) | (<0.001) | (<0.001) |
| frost | -0.005 | -0.005 | -0.005 | -0.005 | -0.005 | -0.005 |
|  | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) |
|  | (0.043) | (0.043) | (0.043) | (0.043) | (0.043) | (0.043) |
| log\_Population | 0.247 | 0.247 | 0.247 | 0.247 | 0.247 | 0.247 |
|  | (0.113) | (0.113) | (0.113) | (0.113) | (0.113) | (0.113) |
|  | (0.033) | (0.033) | (0.033) | (0.033) | (0.033) | (0.033) |
| Num.Obs. | 50 | 50 | 50 | 50 | 50 | 50 |
| R2 | 0.740 | 0.740 | 0.740 | 0.740 | 0.740 | 0.740 |
| R2 Adj. | 0.717 | 0.717 | 0.717 | 0.717 | 0.717 | 0.717 |
| AIC | 114.9 | 114.9 | 114.9 | 114.9 | 114.9 | 114.9 |
| BIC | 126.4 | 126.4 | 126.4 | 126.4 | 126.4 | 126.4 |
| Log.Lik. | -51.448 | -51.448 | -51.448 | -51.448 | -51.448 | -51.448 |
| RMSE | 0.68 | 0.68 | 0.68 | 0.68 | 0.68 | 0.68 |

all\_submodel = regsubsets(life\_exp ~., data = data.state)  
summary(all\_submodel)

## Subset selection object  
## Call: regsubsets.formula(life\_exp ~ ., data = data.state)  
## 7 Variables (and intercept)  
## Forced in Forced out  
## income FALSE FALSE  
## illiteracy FALSE FALSE  
## murder FALSE FALSE  
## hs\_grad FALSE FALSE  
## frost FALSE FALSE  
## area FALSE FALSE  
## log\_Population FALSE FALSE  
## 1 subsets of each size up to 7  
## Selection Algorithm: exhaustive  
## income illiteracy murder hs\_grad frost area log\_Population  
## 1 ( 1 ) " " " " "\*" " " " " " " " "   
## 2 ( 1 ) " " " " "\*" "\*" " " " " " "   
## 3 ( 1 ) " " " " "\*" "\*" " " " " "\*"   
## 4 ( 1 ) " " " " "\*" "\*" "\*" " " "\*"   
## 5 ( 1 ) " " "\*" "\*" "\*" "\*" " " "\*"   
## 6 ( 1 ) "\*" "\*" "\*" "\*" "\*" " " "\*"   
## 7 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*"

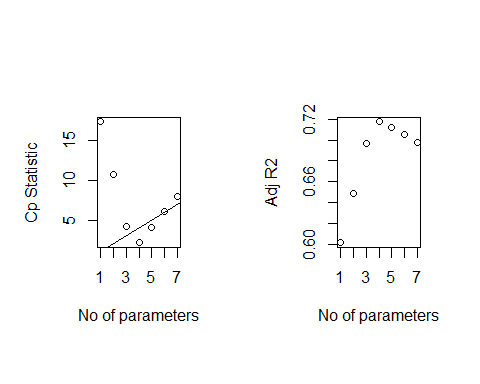
library(leaps)  
par(mfrow=c(1,2))  
plot(all\_submodel, scale = "adjr2")   
  
plot(all\_submodel, scale = "bic")



submodel\_summary <- summary(all\_submodel)  
  
print(submodel\_summary$which)

## (Intercept) income illiteracy murder hs\_grad frost area log\_Population  
## 1 TRUE FALSE FALSE TRUE FALSE FALSE FALSE FALSE  
## 2 TRUE FALSE FALSE TRUE TRUE FALSE FALSE FALSE  
## 3 TRUE FALSE FALSE TRUE TRUE FALSE FALSE TRUE  
## 4 TRUE FALSE FALSE TRUE TRUE TRUE FALSE TRUE  
## 5 TRUE FALSE TRUE TRUE TRUE TRUE FALSE TRUE  
## 6 TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE  
## 7 TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE

sum1 = summary(all\_submodel)  
par(mfrow=c(1,2))  
plot(1:7, sum1$cp, xlab = "No of parameters", ylab = "Cp Statistic")  
abline(0,1)  
plot(1:7, sum1$adjr2, xlab = "No of parameters", ylab = "Adj R2")



library(glmnet)

## Warning: 程序包'glmnet'是用R版本4.4.2 来建造的

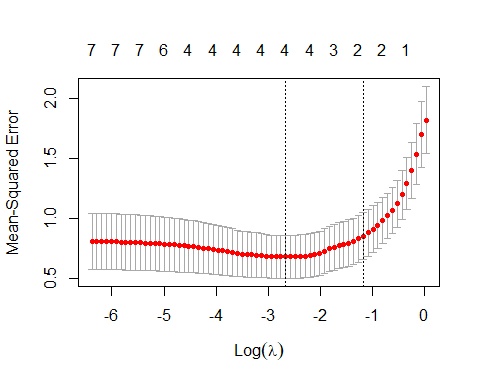
## 载入需要的程序包：Matrix

##   
## 载入程序包：'Matrix'

## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack

## Loaded glmnet 4.1-8

# Convert predictors to a matrix (exclude the intercept)  
X <- model.matrix(life\_exp ~ ., data = data.state)[, -1]  
  
# Response variable  
y <- data.state$life\_exp  
  
# Fit LASSO model with cross-validation  
set.seed(123) # Set seed for reproducibility  
lasso\_cv <- cv.glmnet(X, y, alpha = 1, nfolds = 10) # alpha = 1 for LASSO  
  
# Plot cross-validation results  
plot(lasso\_cv)



# Best lambda based on minimum cross-validation error  
best\_lambda <- lasso\_cv$lambda.min  
  
# Lambda within 1 standard error of the minimum error  
lambda\_1se <- lasso\_cv$lambda.1se  
  
cat("Best lambda (minimum error):", best\_lambda, "\n")

## Best lambda (minimum error): 0.06987808

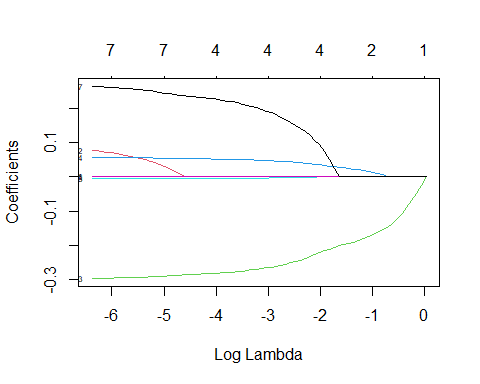
cat("Best lambda (1-SE rule):", lambda\_1se, "\n")

## Best lambda (1-SE rule): 0.3096033

# Fit the final LASSO model  
final\_lasso <- glmnet(X, y, alpha = 1, lambda = best\_lambda)  
  
# Extract coefficients as a matrix  
lasso\_coefficients <- as.matrix(coef(final\_lasso))  
  
# Get the row names of variables with non-zero coefficients  
selected\_variables <- rownames(lasso\_coefficients)[lasso\_coefficients[, 1] != 0]  
  
# Print the selected variables  
print(selected\_variables)

## [1] "(Intercept)" "murder" "hs\_grad" "frost"   
## [5] "log\_Population"

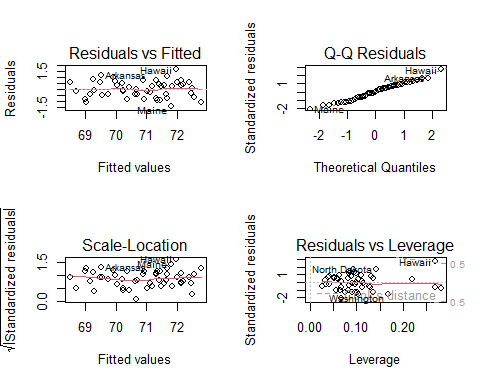
plot(lasso\_cv$glmnet.fit, xvar = "lambda", label = TRUE)



# Define the formula  
selected\_formula <- life\_exp ~ murder + hs\_grad + frost + log\_Population  
  
# Fit the regression model with selected variables  
selected\_model <- lm(selected\_formula, data = data.state)  
  
# Display the summary of the regression  
summary(selected\_model)

##   
## Call:  
## lm(formula = selected\_formula, data = data.state)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.41760 -0.43880 0.02539 0.52066 1.63048   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 68.720810 1.416828 48.503 < 2e-16 \*\*\*  
## murder -0.290016 0.035440 -8.183 1.87e-10 \*\*\*  
## hs\_grad 0.054550 0.014758 3.696 0.000591 \*\*\*  
## frost -0.005174 0.002482 -2.085 0.042779 \*   
## log\_Population 0.246836 0.112539 2.193 0.033491 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7137 on 45 degrees of freedom  
## Multiple R-squared: 0.7404, Adjusted R-squared: 0.7173   
## F-statistic: 32.09 on 4 and 45 DF, p-value: 1.17e-12

par(mfrow = c(2,2))  
plot(selected\_model)



set.seed(111)  
  
train = trainControl(method = "cv", number = 10)  
  
model\_10fold = train(selected\_formula,  
data = data.state,  
trControl = train,  
method = 'lm',  
na.action = na.pass)  
  
model\_10fold

## Linear Regression   
##   
## 50 samples  
## 4 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 44, 45, 45, 46, 44, 45, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 0.7514293 0.7768153 0.6340191  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

model\_10fold$resample

## RMSE Rsquared MAE Resample  
## 1 0.7260482 0.5544736 0.5823172 Fold01  
## 2 0.6477251 0.8407706 0.6219278 Fold02  
## 3 0.6144612 0.8668420 0.5363928 Fold03  
## 4 0.6187864 0.8727852 0.5046897 Fold04  
## 5 0.7397143 0.8005523 0.6327371 Fold05  
## 6 0.7129359 0.9011760 0.6074712 Fold06  
## 7 0.8472056 0.6158807 0.6999753 Fold07  
## 8 0.4983759 0.8025908 0.4776243 Fold08  
## 9 0.9325622 0.6421142 0.8548502 Fold09  
## 10 1.1764783 0.8709673 0.8222052 Fold10