p8130_hw5_yl5508

Yifei LIU (yl5508)

2023/12/7

```
library(tidyverse)
library(faraway)
library(glmnet)
library(caret)

set.seed(123)
```

Problem 1

(a)

```
#load data set
sta_data =
  as.data.frame(state.x77)

'sta_data |>
  mutate(state = rownames(state.x77))
rownames(sta_data) = NULL'
```

```
## [1] "sta_data |>\n  mutate(state = rownames(state.x77))\nrownames(sta_data) = NULL"
```

```
sum_sta =
  sta_data |>
  skimr::skim() |>
  select(skim_variable, numeric.mean, numeric.sd,
         numeric.p0, numeric.p25, numeric.p50, numeric.p75, numeric.p100)

colnames(sum_sta) = c("Variable", "Mean", "SD", "Min", "Q1", "Median", "Q3", "Max")

knitr::kable(x = sum_sta, caption = "Variables pre-analysis", digits = 1)
```

Table 1: Variables pre-analysis

Variable	Mean	SD	Min	Q1	Median	Q3	Max
Population	4246.4	4464.5	365.0	1079.5	2838.5	4968.5	21198.0
Income	4435.8	614.5	3098.0	3992.8	4519.0	4813.5	6315.0
Illiteracy	1.2	0.6	0.5	0.6	0.9	1.6	2.8
Life Exp	70.9	1.3	68.0	70.1	70.7	71.9	73.6

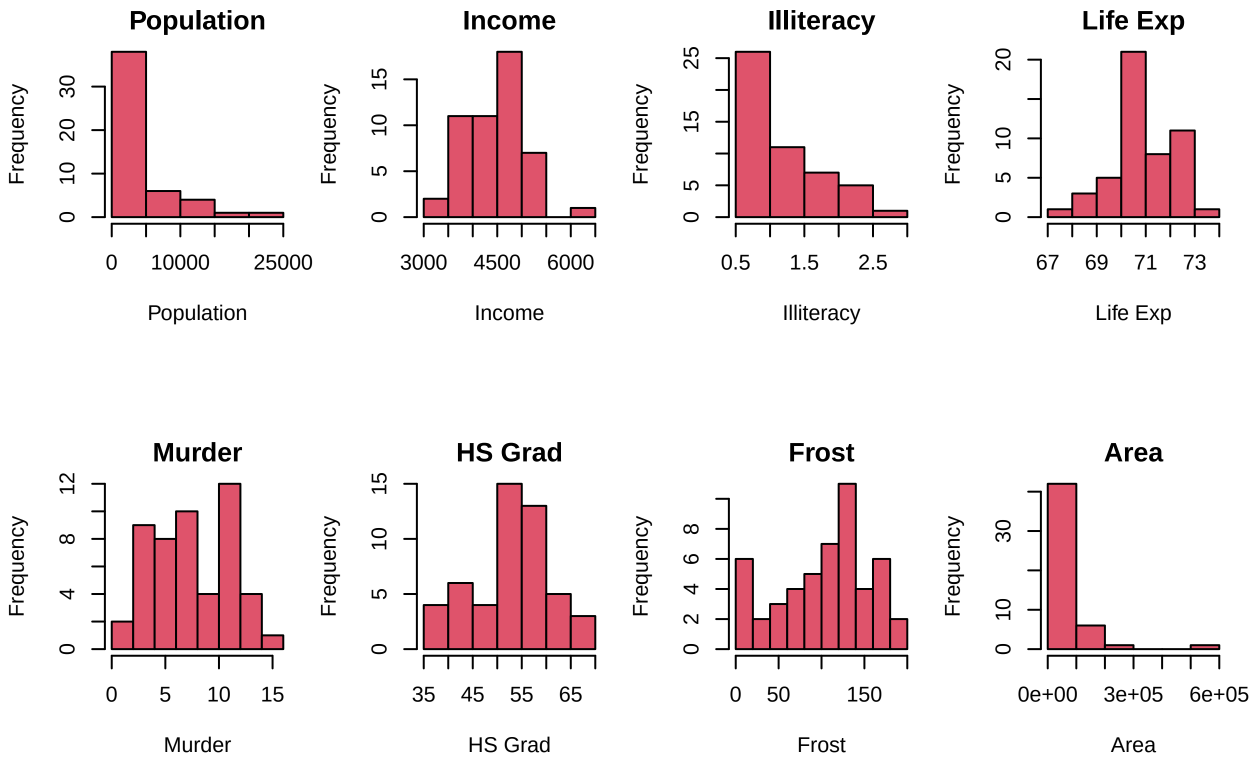
Variable	Mean	SD	Min	Q1	Median	Q3	Max
Murder	7.4	3.7	1.4	4.3	6.8	10.7	15.1
HS Grad	53.1	8.1	37.8	48.0	53.2	59.2	67.3
Frost	104.5	52.0	0.0	66.2	114.5	139.8	188.0
Area	70735.9	85327.3	1049.0	36985.2	54277.0	81162.5	566432.0

Continuous variables includes `Population`, `Income`, `Illiteracy`, `Life Exp`, `Murder`, `HS Grad`, `Frost`, `Area`.

Only variable `state` in the data set is categorical.

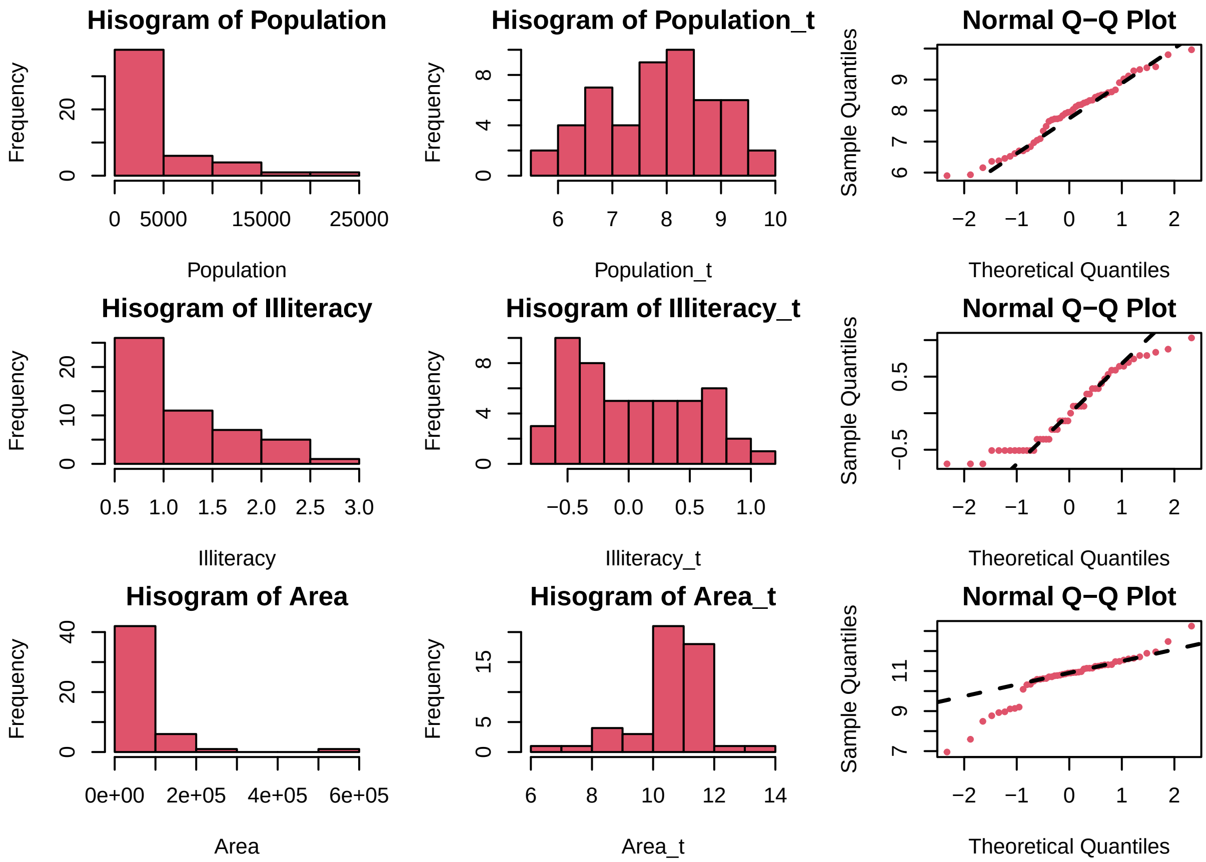
(b)

```r
#histogram of variables
par(mfrow = c(2, 4), mar = c(8, 4, 2, 1))

for (i in 1:8) {
  sta_data[,i] |>
  hist(main = colnames(sta_data[i]), xlab = colnames(sta_data[i]), freq = T, col = 2)
}
```



From the histograms, we notice that `Population`, `Illiteracy`, `Area` need to be transformed in order to get a normal distribution.

```r
#log transformation
sta_transformed =
  sta_data |>
  mutate(
    Population_t = log(Population),
    Illiteracy_t = log(Illiteracy),
    Area_t = log(Area)) |>
  select(Population, Population_t, Illiteracy, Illiteracy_t, Area, Area_t)

sta_tidy =
  sta_data |>
  mutate(
    Population_t = log(Population),
    Illiteracy_t = log(Illiteracy),
    Area_t = log(Area)) |>
  select(-Population, -Illiteracy, -Area) |>
  select(`Life Exp`, everything())

par(mfrow = c(3, 3), mar = c(4, 4, 2, 2))

for (i in seq(1, 5, 2)) {
  #untransformed variables
  sta_transformed[,i] |>
    hist(main = str_c("Hisogram of ", colnames(sta_transformed[i])),
         xlab = colnames(sta_transformed[i]), freq = T, col = 2)

  #log transformed variables
  sta_transformed[,i+1] |>
    hist(main = str_c("Hisogram of ", colnames(sta_transformed[i+1])),
         xlab = colnames(sta_transformed[i+1]), freq = T, col = 2)

  #Q-Q plot
  qqnorm(sta_transformed[,i+1], col = 2, pch = 19, cex = 0.5)
  qqline(sta_transformed[,i+1], col = 1, lwd = 2, lty = 2)
}
```



(c)

```r
#global variables
lm(`Life Exp` ~ ., data = sta_tidy) |>
  summary()
```

```
##
## Call:
## lm(formula = `Life Exp` ~ ., data = sta_tidy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.44702 -0.42901  0.04546  0.50742  1.68911
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.799e+01  1.798e+00  37.809  < 2e-16 ***
## Income        -4.417e-06  2.475e-04  -0.018   0.9858
## Murder        -3.114e-01  4.659e-02  -6.684 4.12e-08 ***
## `HS Grad`      5.482e-02  2.552e-02   2.148   0.0375 *
## Frost         -4.669e-03  3.173e-03  -1.471   0.1487
## Population_t   2.537e-01  1.311e-01   1.936   0.0597 .
## Illiteracy_t   1.883e-01  4.204e-01   0.448   0.6565
## Area_t         7.314e-02  1.102e-01   0.663   0.5107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.7335 on 42 degrees of freedom
## Multiple R-squared:  0.7441, Adjusted R-squared:  0.7014
## F-statistic: 17.45 on 7 and 42 DF,  p-value: 1.368e-10
```

```
#forward stepwise
model_fw = lm(`Life Exp` ~ ., data = sta_tidy) |>
  step(direction = "forward")
```

```
## Start:  AIC=-23.71
## `Life Exp` ~ Income + Murder + `HS Grad` + Frost + Population_t +
##     Illiteracy_t + Area_t
```

```
model_fw |> summary()
```

```
##
## Call:
## lm(formula = `Life Exp` ~ Income + Murder + `HS Grad` + Frost +
##     Population_t + Illiteracy_t + Area_t, data = sta_tidy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.44702 -0.42901  0.04546  0.50742  1.68911
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.799e+01  1.798e+00  37.809  < 2e-16 ***
## Income       -4.417e-06  2.475e-04  -0.018   0.9858
## Murder       -3.114e-01  4.659e-02  -6.684 4.12e-08 ***
## `HS Grad`     5.482e-02  2.552e-02   2.148   0.0375 *
## Frost        -4.669e-03  3.173e-03  -1.471   0.1487
## Population_t  2.537e-01  1.311e-01   1.936   0.0597 .
## Illiteracy_t  1.883e-01  4.204e-01   0.448   0.6565
## Area_t        7.314e-02  1.102e-01   0.663   0.5107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7335 on 42 degrees of freedom
## Multiple R-squared:  0.7441, Adjusted R-squared:  0.7014
## F-statistic: 17.45 on 7 and 42 DF,  p-value: 1.368e-10
```

```
#backward stepwise
model_bk = lm(`Life Exp` ~ ., data = sta_tidy) |>
  step(direction = "backward")
```

```
## Start:  AIC=-23.71
## `Life Exp` ~ Income + Murder + `HS Grad` + Frost + Population_t +
##     Illiteracy_t + Area_t
##
##                Df Sum of Sq    RSS     AIC
## - Income        1    0.0002 22.596 -25.712
## - Illiteracy_t  1    0.1079 22.704 -25.475
## - Area_t         1    0.2368 22.833 -25.192
## <none>                        22.596 -23.713
## - Frost          1    1.1645 23.760 -23.200
## - Population_t   1    2.0155 24.611 -21.441
## - `HS Grad`      1    2.4822 25.078 -20.502
## - Murder         1   24.0347 46.631  10.512
##
## Step:  AIC=-25.71
## `Life Exp` ~ Murder + `HS Grad` + Frost + Population_t + Illiteracy_t +
##     Area_t
##
##                 Df Sum of Sq    RSS      AIC
## - Illiteracy_t   1    0.1095 22.705 -27.4708
## - Area_t         1    0.2616 22.858 -27.1370
## <none>                        22.596 -25.7125
## - Frost          1    1.2628 23.859 -24.9936
## - Population_t   1    2.3859 24.982 -22.6937
## - `HS Grad`      1    4.4112 27.007 -18.7959
## - Murder         1   24.4834 47.079   8.9907
##
## Step:  AIC=-27.47
## `Life Exp` ~ Murder + `HS Grad` + Frost + Population_t + Area_t
##
##                 Df Sum of Sq    RSS     AIC
## - Area_t         1    0.2157 22.921 -28.998
## <none>                        22.705 -27.471
## - Population_t   1    2.2792 24.985 -24.688
## - Frost          1    2.3760 25.082 -24.495
## - `HS Grad`      1    4.9491 27.655 -19.612
## - Murder         1   29.2296 51.935  11.899
##
## Step:  AIC=-29
## `Life Exp` ~ Murder + `HS Grad` + Frost + Population_t
##
##                 Df Sum of Sq    RSS     AIC
## <none>                        22.921 -28.998
## - Frost          1     2.214 25.135 -26.387
## - Population_t   1     2.450 25.372 -25.920
## - `HS Grad`      1     6.959 29.881 -17.741
## - Murder         1    34.109 57.031  14.578
```

```r
model_bk |> summary()
```

```
##
## Call:
## lm(formula = `Life Exp` ~ Murder + `HS Grad` + Frost + Population_t,
##     data = sta_tidy)
##
## 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 *  
## Population_t  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
```

```r
#both
model_bth = lm(`Life Exp` ~ ., data = sta_tidy) |>
  step(direction = "both")
```

```
## Start:  AIC=-23.71
## `Life Exp` ~ Income + Murder + `HS Grad` + Frost + Population_t +
##     Illiteracy_t + Area_t
##
##                 Df Sum of Sq    RSS     AIC
## - Income         1    0.0002 22.596 -25.712
## - Illiteracy_t   1    0.1079 22.704 -25.475
## - Area_t         1    0.2368 22.833 -25.192
## <none>                       22.596 -23.713
## - Frost          1    1.1645 23.760 -23.200
## - Population_t   1    2.0155 24.611 -21.441
## - `HS Grad`      1    2.4822 25.078 -20.502
## - Murder         1   24.0347 46.631  10.512
##
## Step:  AIC=-25.71
## `Life Exp` ~ Murder + `HS Grad` + Frost + Population_t + Illiteracy_t +
##     Area_t
##
##                 Df Sum of Sq    RSS      AIC
## - Illiteracy_t   1    0.1095 22.705 -27.4708
## - Area_t         1    0.2616 22.858 -27.1370
## <none>                       22.596 -25.7125
## - Frost          1    1.2628 23.859 -24.9936
## + Income         1    0.0002 22.596 -23.7129
## - Population_t   1    2.3859 24.982 -22.6937
## - `HS Grad`      1    4.4112 27.007 -18.7959
## - Murder         1   24.4834 47.079   8.9907
##
## Step:  AIC=-27.47
## `Life Exp` ~ Murder + `HS Grad` + Frost + Population_t + Area_t
##
##                 Df Sum of Sq    RSS     AIC
## - Area_t         1    0.2157 22.921 -28.998
## <none>                       22.705 -27.471
## + Illiteracy_t   1    0.1095 22.596 -25.712
## + Income         1    0.0017 22.704 -25.475
## - Population_t   1    2.2792 24.985 -24.688
## - Frost          1    2.3760 25.082 -24.495
## - `HS Grad`     1    4.9491 27.655 -19.612
## - Murder        1   29.2296 51.935  11.899
##
## Step:  AIC=-29
## `Life Exp` ~ Murder + `HS Grad` + Frost + Population_t
##
##                 Df Sum of Sq    RSS     AIC
## <none>                       22.921 -28.998
## + Area_t         1     0.216 22.705 -27.471
## + Illiteracy_t   1     0.064 22.858 -27.137
## + Income         1     0.011 22.911 -27.021
## - Frost          1     2.214 25.135 -26.387
## - Population_t   1     2.450 25.372 -25.920
## - `HS Grad`      1     6.959 29.881 -17.741
## - Murder         1    34.109 57.031  14.578
```

```
model_bth |> summary()
```

```
##
## Call:
## lm(formula = `Life Exp` ~ Murder + `HS Grad` + Frost + Population_t,
##     data = sta_tidy)
##
## 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 *
## Population_t  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
```

Based on AIC, the function reduces the set of potential predictors. The model with the smallest value would be deemed as appropriate.

Actually the model shown after variables selection would not be the final result. We need to trim some variables off based on the p-value listed in the tables.

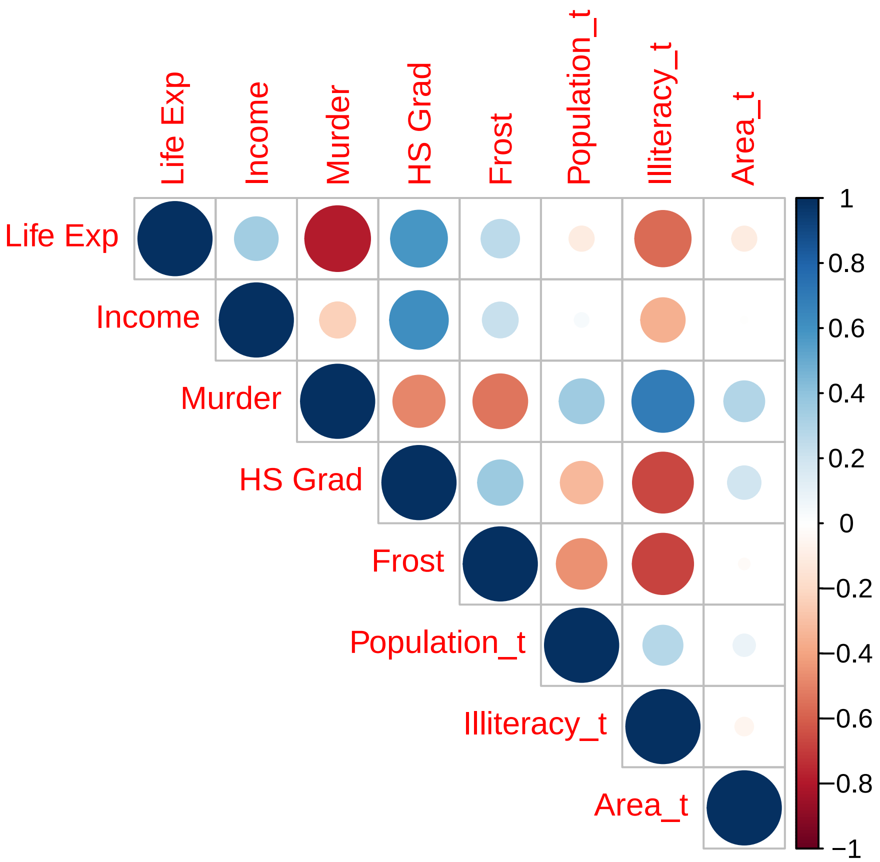
For forward selection, the model shows that only variables `Murder` and `HS Grad` is significantly effective (p < 0.05).

For backward selection, `Murder`, `HS Grad`, `Frost`, `Population_t` are significant variables.

For method concerning both forward and backward selection, the result is the same as backward selection.

So, I would pick `Murder`, `HS Grad`, `Frost`, `Population_t` as my predictors.

```
corrplot::corrplot(cor(sta_tidy), type = "upper")
```



We notice that there is a strong negative correlation between `Illiteracy` and `HS Grad` (approximately 0.8). My variables subset doesn't include both, for the variables selection process can partly deal with multicollinearity issue.

(d)

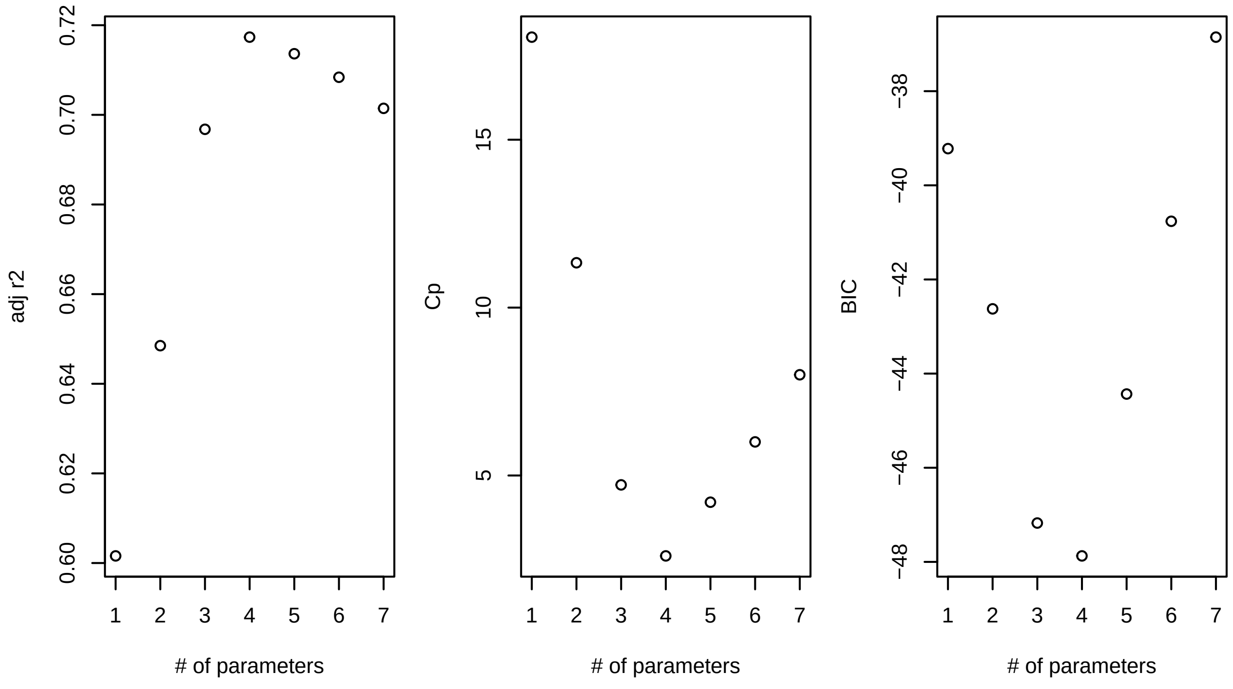
```
library(leaps)
```

```
## Warning:   'leaps' R 4.3.2
```

```
mat = as.matrix(sta_tidy)
leaps(x = mat[, 2:8], y = mat[, 1], method = "adjr2", nbest = 2)
```

```
## $which
##       1     2     3     4     5     6     7
## 1 FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
## 1 FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
## 2 FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE
## 2 FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE
## 3 FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE
## 3 FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE
## 4 FALSE  TRUE  TRUE  TRUE  TRUE FALSE FALSE
## 4 FALSE  TRUE  TRUE FALSE  TRUE  TRUE FALSE
## 5 FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
## 5 FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
## 6 FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## 6  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
## 7  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##
## $label
## [1] "(Intercept)" "1"           "2"           "3"           "4"
## [6] "5"           "6"           "7"
##
## $size
##  [1] 2 2 3 3 4 4 5 5 6 6 7 7 8
##
## $adjr2
##  [1] 0.6015893 0.3252044 0.6484991 0.6301232 0.6967729 0.6939230 0.7173392
##  [8] 0.7031061 0.7136360 0.7117179 0.7083894 0.7069987 0.7014485
```

```r
model_cri = regsubsets(`Life Exp` ~ ., data = sta_tidy)
res =
  model_cri |>
  summary()

par(mfrow = c(1, 3), mar = c(8, 4, 4, 1))
plot(1:7, res$adjr2, xlab = "# of parameters", ylab = "adj r2")
plot(1:7, res$cp, xlab = "# of parameters", ylab = "Cp")
plot(1:7, res$bic, xlab = "# of parameters", ylab = "BIC")
```



```
res$outmat[4,]
```

```
##       Income       Murder    `HS Grad`        Frost Population_t Illiteracy_t 
##          " "          "*"          "*"          "*"          "*"          " " 
##       Area_t 
##          " "
```

From the criterion-based procedures, using `adjusted r squared/Cp criterion/BIC`, we conclude that the best model contain 4 parameters and the parameters are `Murder`, `HS Grad`, `Frost`, `Population_t`.

(e)

```
#explore possible lambda
fit1 = glmnet(x = as.matrix(sta_tidy[2:8]), y = sta_tidy$`Life Exp`, data = sta_tidy, lambda = 1)
coef(fit1)
```

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                       s0
## (Intercept)  70.95464716
## Income         .        
## Murder       -0.01030729
## HS Grad        .        
## Frost          .        
## Population_t   .        
## Illiteracy_t   .        
## Area_t         .
```

```
fit2 = glmnet(x = as.matrix(sta_tidy[2:8]), y = sta_tidy$`Life Exp`, data = sta_tidy, lambda = 0.1)
coef(fit2)
```

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                        s0
## (Intercept)  69.623968632
## Income         .         
## Murder       -0.238460282
## HS Grad       0.040072350
## Frost        -0.001483129
## Population_t  0.132353805
## Illiteracy_t   .         
## Area_t         .
```

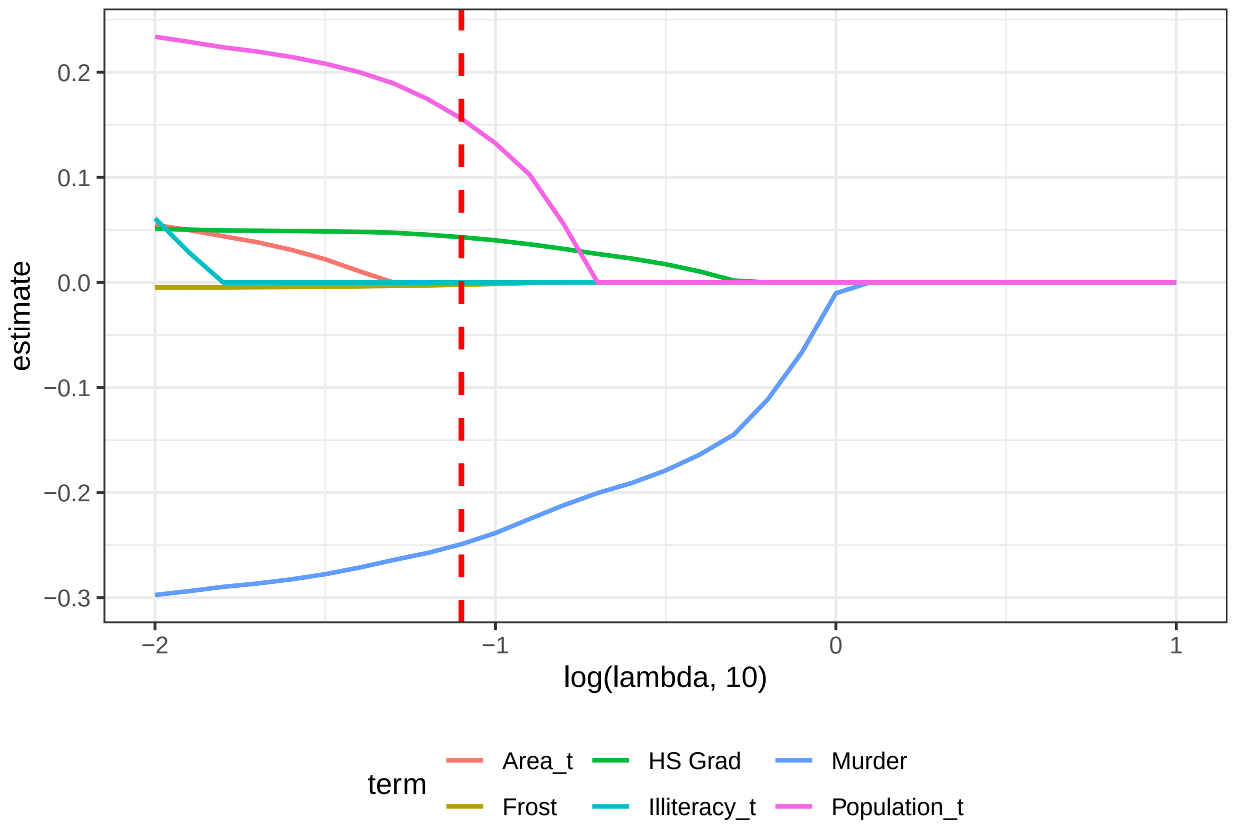
```
fit3 = glmnet(x = as.matrix(sta_tidy[2:8]), y = sta_tidy$`Life Exp`, data = sta_tidy, lambda = 0.01)
coef(fit3)
```

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                        s0
## (Intercept)  68.426042158
## Income         .         
## Murder       -0.297414030
## HS Grad       0.051183064
## Frost        -0.004748876
## Population_t  0.233986320
## Illiteracy_t  0.062909458
## Area_t        0.054660434
```

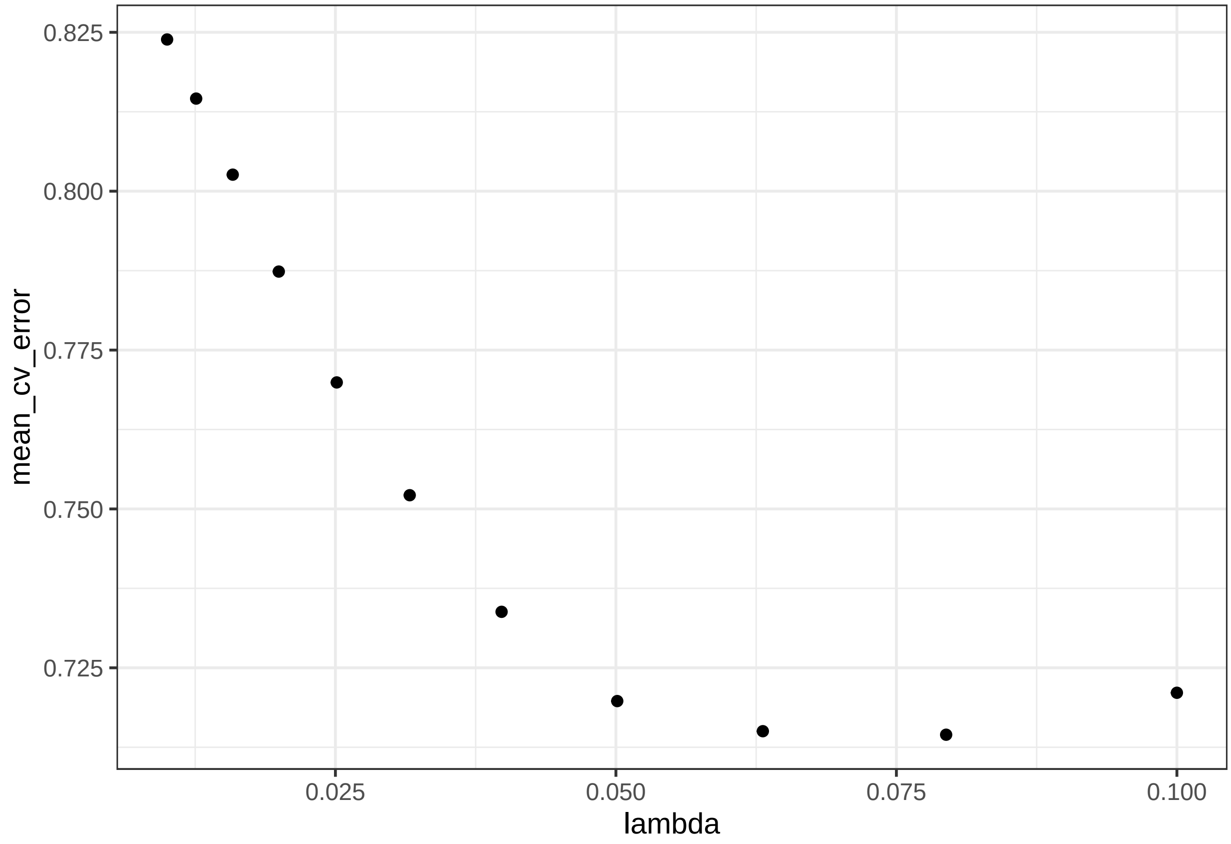
We would consider setting the range of lambda at the interval of (0.01, 0.1).

```
#grid search
lambda_seq = 10^seq(-2, 1, by = 0.1)
```

```
cv_res =
  cv.glmnet(x = as.matrix(sta_tidy[2:8]), y = sta_tidy$`Life Exp`, data = sta_tidy,
            lambda = lambda_seq, nfolds = 5)

opt_lambda = cv_res$lambda.min

#variables contraction
glmnet(x = as.matrix(sta_tidy[2:8]), y = sta_tidy$`Life Exp`, data = sta_tidy, lambda = lambda_seq) |>
  broom::tidy() |>
  select(term, lambda, estimate) |>
  complete(term, lambda, fill = list(estimate = 0) ) |>
  filter(term != "(Intercept)") |>
  ggplot(aes(x = log(lambda, 10), y = estimate, group = term, color = term)) +
  geom_path(size = 0.8) +
  geom_vline(xintercept = log(opt_lambda, 10), color = "red", linetype = "dashed", size = 1) +
  theme_bw() +
  theme(legend.position = "bottom")
```



```
tb_res = tibble(
  lambda = cv_res$lambda,
  mean_cv_error = cv_res$cvm) |>
  filter(lambda < 0.1)

#choosing optimal lambda
tb_res |>
  ggplot(aes(x = lambda, y = mean_cv_error)) +
  geom_point() +
  theme_bw()
```



The optimal lambda we have chosen is 0.08. And the variables we determine are `Murder`, `HS Grad`, `Frost`, `Population_t`.

(f)

```
#stepwise
fit_bth = lm(`Life Exp` ~ ., data = sta_tidy) |>
  step(direction = "both")
```

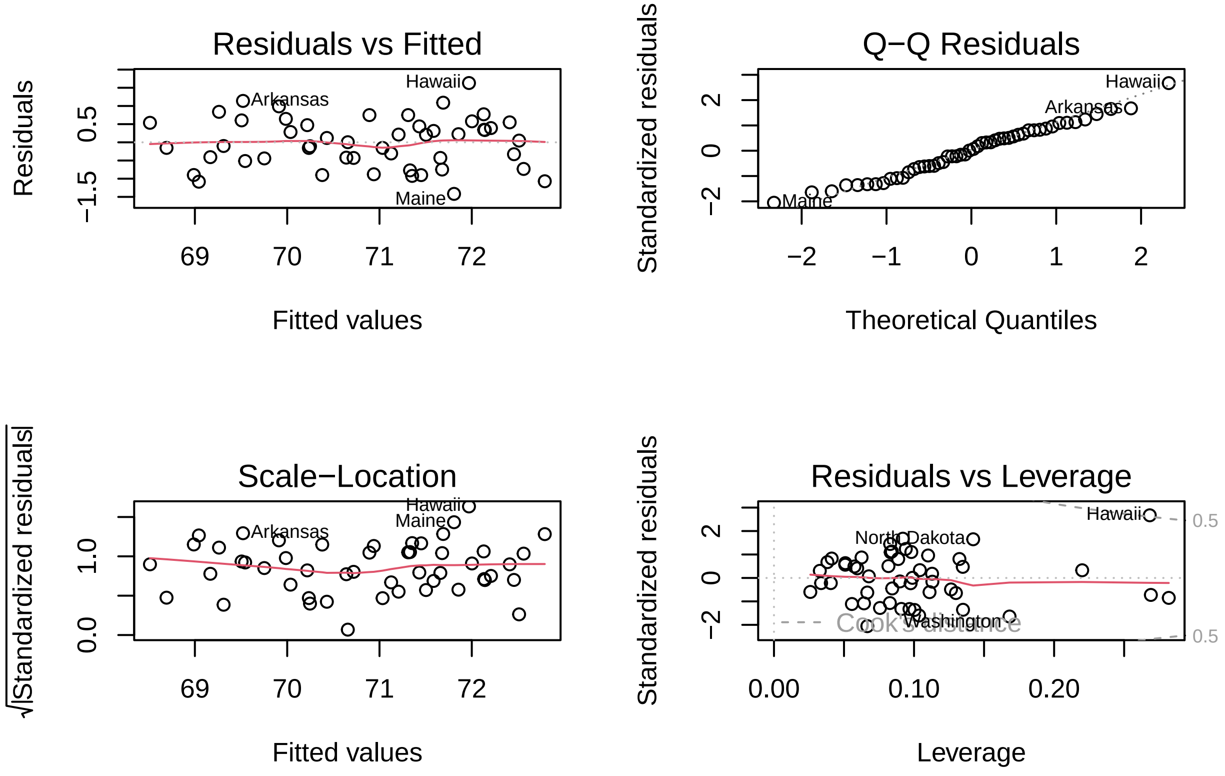
```
## Start:  AIC=-23.71
## `Life Exp` ~ Income + Murder + `HS Grad` + Frost + Population_t +
##     Illiteracy_t + Area_t
##
##                Df Sum of Sq    RSS     AIC
## - Income        1    0.0002 22.596 -25.712
## - Illiteracy_t  1    0.1079 22.704 -25.475
## - Area_t        1    0.2368 22.833 -25.192
## <none>                      22.596 -23.713
## - Frost          1    1.1645 23.760 -23.200
## - Population_t  1    2.0155 24.611 -21.441
## - `HS Grad`     1    2.4822 25.078 -20.502
## - Murder        1   24.0347 46.631  10.512
##
## Step:  AIC=-25.71
## `Life Exp` ~ Murder + `HS Grad` + Frost + Population_t + Illiteracy_t +
##     Area_t
##
##                 Df Sum of Sq    RSS      AIC
## - Illiteracy_t  1    0.1095 22.705 -27.4708
## - Area_t        1    0.2616 22.858 -27.1370
## <none>                      22.596 -25.7125
## - Frost         1    1.2628 23.859 -24.9936
## + Income        1    0.0002 22.596 -23.7129
## - Population_t  1    2.3859 24.982 -22.6937
## - `HS Grad`     1    4.4112 27.007 -18.7959
## - Murder        1   24.4834 47.079   8.9907
##
## Step:  AIC=-27.47
## `Life Exp` ~ Murder + `HS Grad` + Frost + Population_t + Area_t
##
##                 Df Sum of Sq    RSS     AIC
## - Area_t        1    0.2157 22.921 -28.998
## <none>                      22.705 -27.471
## + Illiteracy_t  1    0.1095 22.596 -25.712
## + Income        1    0.0017 22.704 -25.475
## - Population_t  1    2.2792 24.985 -24.688
## - Frost         1    2.3760 25.082 -24.495
## - `HS Grad`     1    4.9491 27.655 -19.612
## - Murder        1   29.2296 51.935  11.899
##
## Step:  AIC=-29
## `Life Exp` ~ Murder + `HS Grad` + Frost + Population_t
##
##                 Df Sum of Sq    RSS     AIC
## <none>                      22.921 -28.998
## + Area_t        1     0.216 22.705 -27.471
## + Illiteracy_t  1     0.064 22.858 -27.137
## + Income        1     0.011 22.911 -27.021
## - Frost         1     2.214 25.135 -26.387
## - Population_t  1     2.450 25.372 -25.920
## - `HS Grad`     1     6.959 29.881 -17.741
## - Murder        1    34.109 57.031  14.578
```

```r
summary(fit_bth)$adj.r.squared
```

```
## [1] 0.7173392
```

From above exploring, we find that all models or criteria suggest a variables combination of `Murder`, `HS Grad`, `Frost`, `Population_t`. So, I recommend final model to be a multiple linear model of `Life Exp` ~ `Murder` + `HS Grad` + `Frost` + `Population_t` (r squared = 0.72).
(1) Check the model assumptions.

```r
#checking assumptions
par(mfrow = c(2, 2))
plot(fit_bth)
```



For `Residuals vs Fitted plot`, it is used to detect unequal error variance (heteroscedasticity) and outliers. Basically, the plot shows residual values bounce around 0. Except `Hawaii`, they are in the range of (-1.5, 1.5) indicating only `Hawaii` could be a potential outlier.

For `Q-Q plot`, it is used to detect non-normality of residuals and outliers. The plot shows a straight line indicating residuals are normal. `Hawaii` is an extreme point to some degree.

For `Scale-Location plot` (`Standardized Residuals ~ Fitted plot`), it is used to detect unequal error variance (heteroscedasticity). The variances are equal.

For `Residuals vs Leverage plot`, it is used to detect influential cases. We notice that `Hawaii` is just on the Cook's distance line at the upper right corner. `Hawaii` could be an influential case.

Overall, assumptions of regression are met.

```r
#delete extreme outlier `Hawaii`
fit_bth_de = lm(`Life Exp` ~ ., data = sta_tidy[- 11, ]) |>
  step(direction = "both")
```

```
## Start:  AIC=-32.69
## `Life Exp` ~ Income + Murder + `HS Grad` + Frost + Population_t +
##     Illiteracy_t + Area_t
##
```

```
##                 Df Sum of Sq    RSS     AIC
## - Illiteracy_t  1    0.0085 18.148 -34.669
## - Income        1    0.0175 18.157 -34.644
## - Frost         1    0.1374 18.277 -34.322
## <none>                      18.140 -32.692
## - Area_t        1    0.9406 19.080 -32.215
## - `HS Grad`     1    1.0527 19.193 -31.927
## - Population_t  1    3.4542 21.594 -26.151
## - Murder        1   21.8173 39.957   4.003
##
## Step:  AIC=-34.67
## `Life Exp` ~ Income + Murder + `HS Grad` + Frost + Population_t +
##     Area_t
##
##                 Df Sum of Sq    RSS     AIC
## - Income        1    0.0219 18.170 -36.610
## - Frost         1    0.1363 18.285 -36.302
## <none>                      18.148 -34.669
## - Area_t        1    0.9529 19.101 -34.161
## - `HS Grad`     1    1.5219 19.670 -32.723
## + Illiteracy_t  1    0.0085 18.140 -32.692
## - Population_t  1    3.7370 21.885 -27.494
## - Murder        1   27.4827 45.631   8.510
##
## Step:  AIC=-36.61
## `Life Exp` ~ Murder + `HS Grad` + Frost + Population_t + Area_t
##
##                 Df Sum of Sq    RSS     AIC
## - Frost         1    0.1719 18.342 -38.148
## <none>                      18.170 -36.610
## - Area_t        1    1.1178 19.288 -35.685
## + Income        1    0.0219 18.148 -34.669
## + Illiteracy_t  1    0.0128 18.157 -34.644
## - `HS Grad`     1    2.2386 20.409 -32.917
## - Population_t  1    4.1279 22.298 -28.579
## - Murder        1   29.3768 47.547   8.525
##
## Step:  AIC=-38.15
## `Life Exp` ~ Murder + `HS Grad` + Population_t + Area_t
##
##                 Df Sum of Sq    RSS     AIC
## <none>                      18.342 -38.148
## - Area_t        1     1.117 19.459 -37.251
## + Frost         1     0.172 18.170 -36.610
## + Income        1     0.057 18.285 -36.302
## + Illiteracy_t  1     0.006 18.336 -36.165
## - `HS Grad`     1     2.105 20.447 -34.825
## - Population_t  1     5.792 24.134 -26.702
## - Murder        1    32.027 50.370   9.351
```

```r
tibble(
  original_fit = summary(fit_bth)$adj.r.squared,
  deletion_fit = summary(fit_bth_de)$adj.r.squared
) |>
```

```
  knitr::kable(digits = 2)
```

original_fit	deletion_fit
0.72	0.75

(2) Test the model predicative ability using a 10-fold cross-validation.

```
#cross validation
train = trainControl(method = "cv", number = 5)

model_cv = train(`Life Exp` ~ Murder + `HS Grad` + Frost + Population_t,
                 data = sta_tidy[- 11, ], method = 'lm', na.action = na.pass)

model_cv$finalModel
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Coefficients:
##     (Intercept)           Murder  `\\`HS Grad\\``            Frost
##       67.906960        -0.276679         0.046799        -0.001632
##    Population_t
##        0.337449
```

```
print(model_cv)
```

```
## Linear Regression
##
## 49 samples
##  4 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 49, 49, 49, 49, 49, 49, ...
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.7322097  0.7084789  0.6263773
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
summary(model_cv)$adj.r.squared
```

```
## [1] 0.7393987
```

R squared of cross validation is 0.74, suggesting the model fits quite well and it explains a great proportion of outcome `Life Exp`.

(g)

From analysis concerning with relevant data set, we find out several indicators which would predict a large proportion of the target outcome.

In pre-processing procedure, we use log function to transform some variables which do not follow a normal distribution.

By using automatic variables selection procedure (stepwise), criterion-based procedure and LASSO, we get the best subset of the variables, which contains `Murder`, `HS Grad`, `Frost`, `Population_t`. These variables are with the highest adjusted r squared, lowest Cp and BIC.

After generating our final model, we check the basic assumptions. We notice that assumption of equal variance is met and assumption of residual normality is met. But one extreme case (`Hawaii`) is detected and it has been removed in later analysis.

Lastly, a 10-fold cross-validation is taken into account and the model is shown fitting quite well according to the adjusted r squared result. This suggests that the final model can successfully predict life expectancy among states as the investigator requests.