

Homework5

Yuki Joyama

2023-12-12

Problem 1

a) The following table shows the descriptive statistics for all variables of interest in 50 States.

Characteristic	N = 50 ¹
Population	4,246.4 / 2,838.5 (4,464.5)
Income per capita	4,435.8 / 4,519.0 (614.5)
Illiteracy (%)	1.2 / 1.0 (0.6)
Life Expectancy (years)	70.9 / 70.7 (1.3)
Murder rate (per 100,000)	7.4 / 6.9 (3.7)
High graduates (%)	53.1 / 53.3 (8.1)
Number of days below freezing	104.5 / 114.5 (52.0)
Land area (mile ²)	70,735.9 / 54,277.0 (85,327.3)

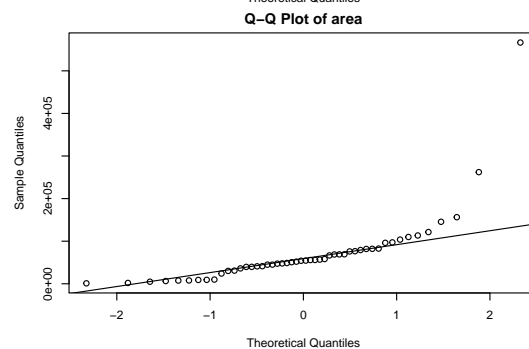
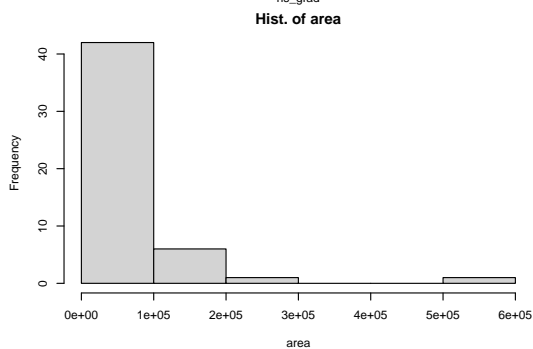
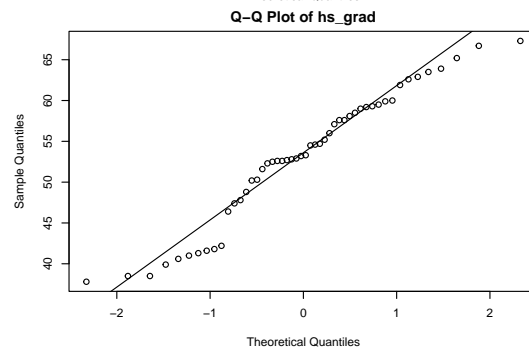
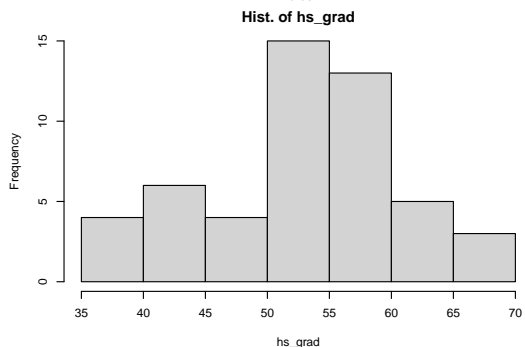
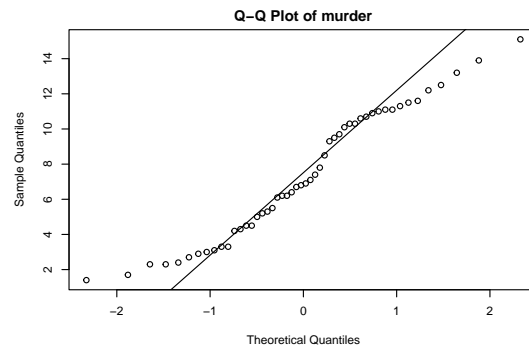
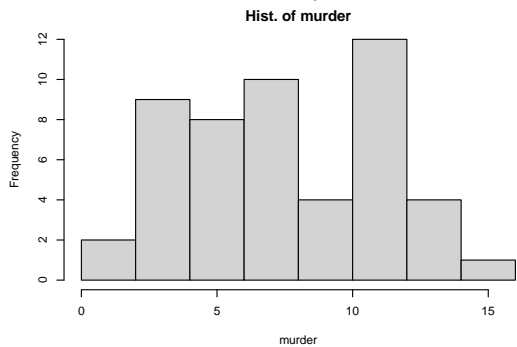
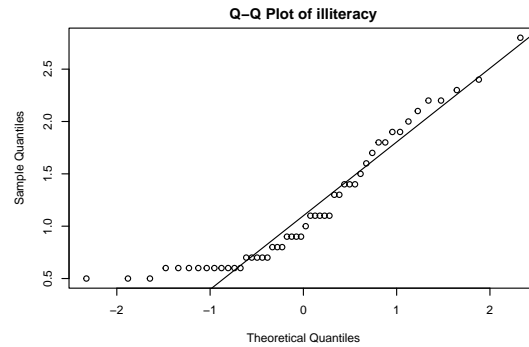
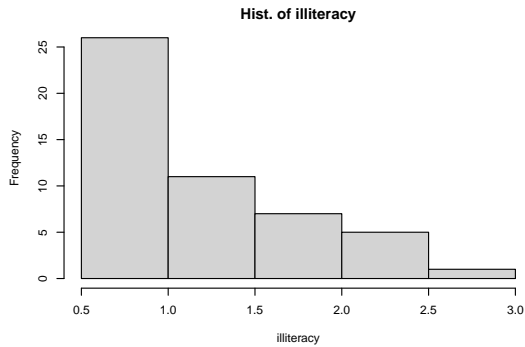
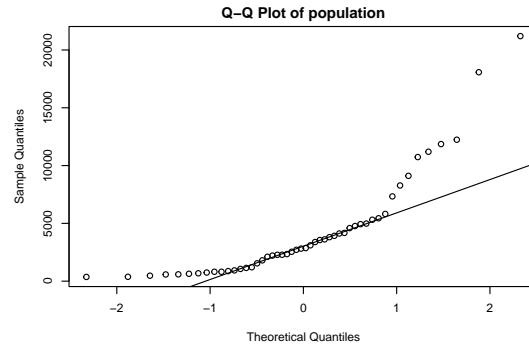
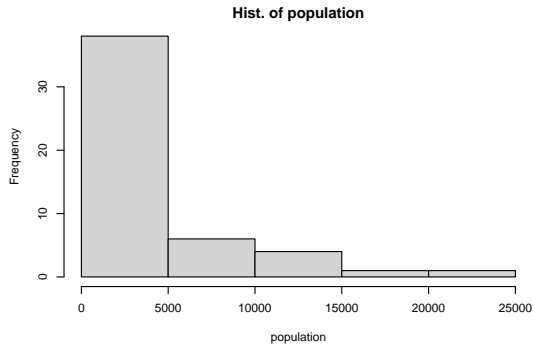
¹Mean / Median (SD)

b)

var	statistic	p.value
population	0.770	<0.001
income	0.977	0.43
illiteracy	0.883	<0.001
life_exp	0.977	0.442
murder	0.953	0.047
hs_grad	0.953	0.046
frost	0.955	0.053
area	0.572	<0.001

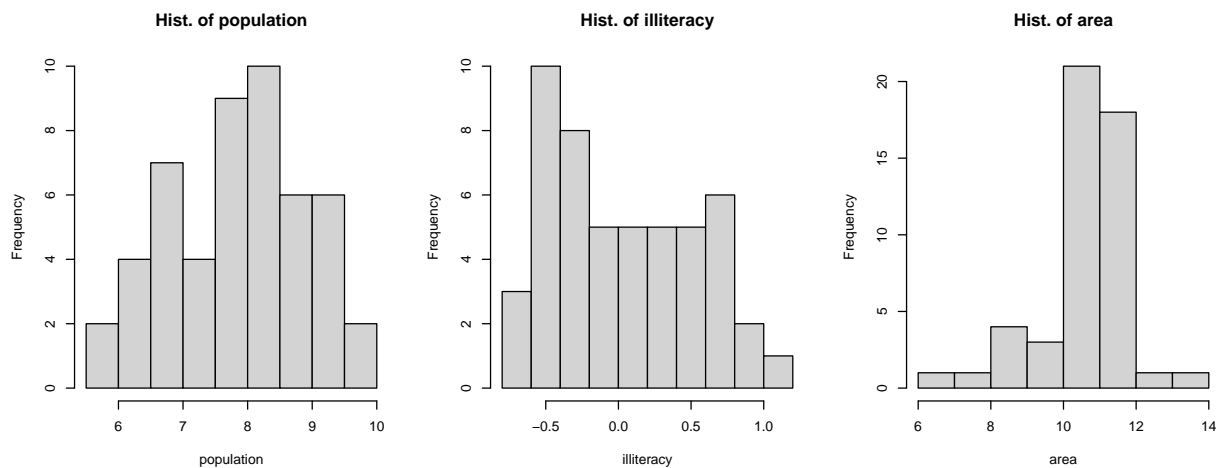
The results of Shapiro-Wilk test indicates that variable `population`, `illiteracy`, `murder`, `hs_grad`, and `area` is not normally distributed.

The histogram and Q-Q plots for these variables are as follows:

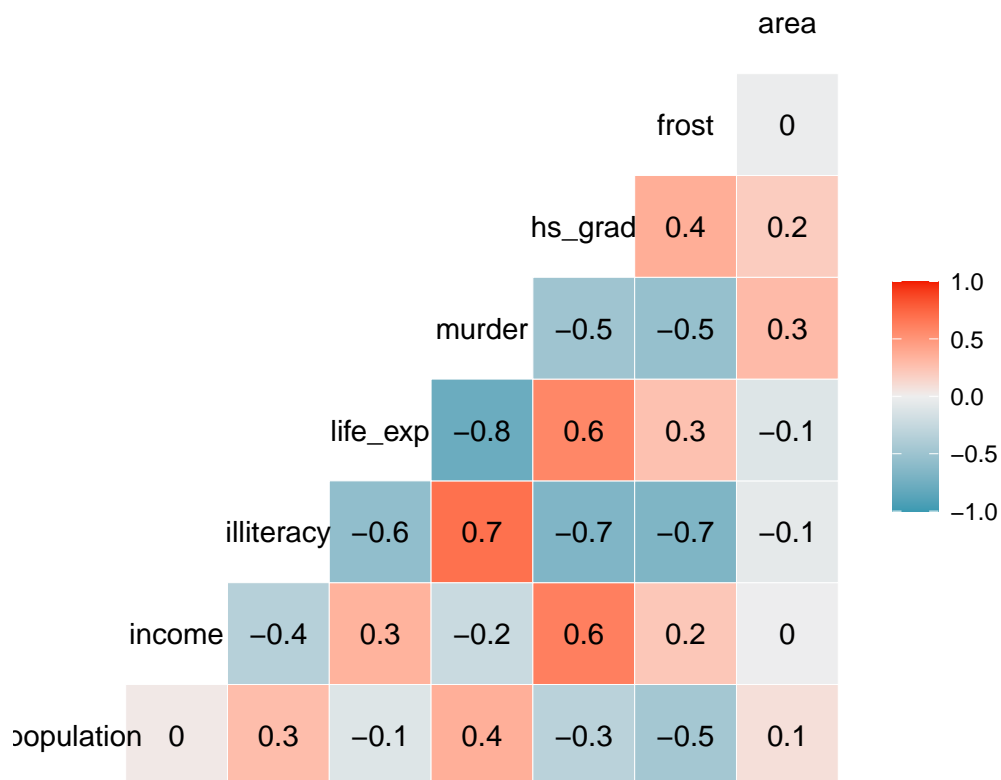


Given the shape of the histograms, I will log-transform population, illiteracy, and area.

Now, let's check these histograms.



I will use the data set including the log-transformed variables for the later analysis. Let's check the correlation between each variable and linear regression model including all variables.



```
##
## Call:
## lm(formula = life_exp ~ ., data = df_val)
##
```

```
## 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 ***
## population   2.537e-01  1.311e-01   1.936  0.0597 .
## income      -4.417e-06  2.475e-04  -0.018  0.9858
## illiteracy   1.883e-01  4.204e-01   0.448  0.6565
## 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
## area         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
```

c) Automatic procedures In this section, I will use backward and forward procedures.

```
## Start:  AIC=-23.71
## life_exp ~ population + income + illiteracy + murder + hs_grad +
##      frost + area
##
##              Df Sum of Sq    RSS    AIC
## - income      1    0.0002 22.596 -25.712
## - illiteracy   1    0.1079 22.704 -25.475
## - area         1    0.2368 22.833 -25.192
## <none>                22.596 -23.713
## - frost       1    1.1645 23.760 -23.200
## - population  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 ~ population + illiteracy + murder + hs_grad + frost +
##      area
##
##              Df Sum of Sq    RSS    AIC
## - illiteracy   1    0.1095 22.705 -27.4708
## - area         1    0.2616 22.858 -27.1370
## <none>                22.596 -25.7125
## - frost       1    1.2628 23.859 -24.9936
## - population  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 ~ population + murder + hs_grad + frost + area
##
##              Df Sum of Sq    RSS    AIC
```

```

## - area          1      0.2157 22.921 -28.998
## <none>                                22.705 -27.471
## - population    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 ~ population + murder + hs_grad + frost
##
##           Df Sum of Sq    RSS    AIC
## <none>                22.921 -28.998
## - frost         1      2.214 25.135 -26.387
## - 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
##
## Call:
## lm(formula = life_exp ~ population + murder + hs_grad + frost,
##     data = df_val)
##
## Coefficients:
## (Intercept)  population      murder    hs_grad      frost
##   68.720810    0.246836   -0.290016    0.054550   -0.005174
##
## Start:  AIC=30.44
## life_exp ~ 1
##
##           Df Sum of Sq    RSS    AIC
## + murder        1     53.838 34.461 -14.609
## + hs_grad        1     29.931 58.368  11.737
## + illiteracy     1     28.688 59.611  12.791
## + income         1     10.223 78.076  26.283
## + frost          1      6.064 82.235  28.878
## <none>                88.299  30.435
## + population    1      1.054 87.245  31.835
## + area          1      1.042 87.257  31.842
##
## 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
## + population     1      2.9854 31.476 -17.140
## + income         1      2.4047 32.057 -16.226
## + area          1      1.4583 33.003 -14.771
## <none>                34.461 -14.609
## + illiteracy     1      0.1292 34.332 -12.797
##
## Step:  AIC=-19.93
## life_exp ~ murder + hs_grad
##

```

```
##           Df Sum of Sq   RSS   AIC
## + population 1    4.6350 25.135 -26.387
## + frost      1    4.3987 25.372 -25.920
## <none>                29.770 -19.925
## + illiteracy 1    0.8366 28.934 -19.351
## + area       1    0.1236 29.647 -18.134
## + income     1    0.1022 29.668 -18.097
##
## Step: AIC=-26.39
## life_exp ~ murder + hs_grad + population
##
##           Df Sum of Sq   RSS   AIC
## + frost      1    2.21416 22.921 -28.998
## + illiteracy 1    1.05998 24.075 -26.542
## <none>                25.135 -26.387
## + income     1    0.11819 25.017 -24.623
## + area       1    0.05387 25.081 -24.495
##
## Step: AIC=-29
## life_exp ~ murder + hs_grad + population + frost
##
##           Df Sum of Sq   RSS   AIC
## <none>                22.921 -28.998
## + area      1    0.215741 22.706 -27.471
## + illiteracy 1    0.063655 22.858 -27.137
## + income    1    0.010673 22.911 -27.021
##
##
## Call:
## lm(formula = life_exp ~ murder + hs_grad + population + frost,
##     data = df_val)
##
## Coefficients:
## (Intercept)      murder      hs_grad  population      frost
##   68.720810   -0.290016    0.054550    0.246836   -0.005174
```

The both procedures generated the same model (variables included in the final model: **murder**, **hs_grad**, **population**, **frost**). There does not appear to be a close call, as the elimination/addition of each variable consistently decreases the AIC value and indicates a better model fit. Therefore, I would keep all the variables suggested by the procedure.

Intuitively, we could assume that there is an association between **illiteracy** and **HS graduation rate**. My subset does not include both, so instead of checking for multicollinearity, let's examine correlation.

```
## [1] -0.6688091
```

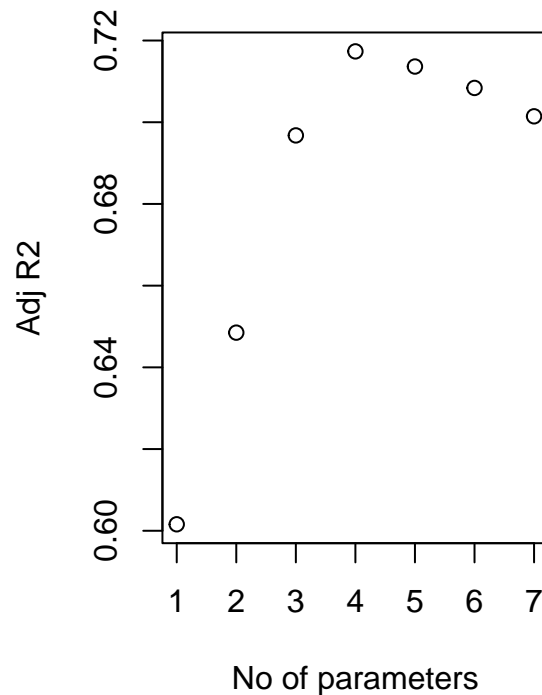
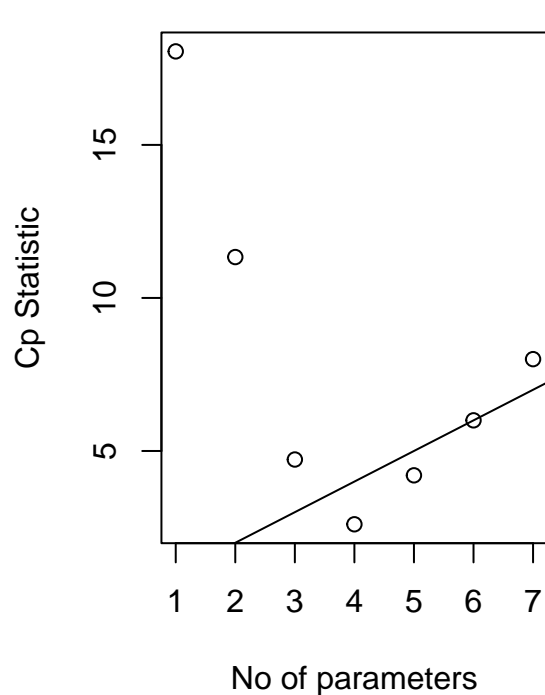
There seems to be a moderate negative association between the two variables.

d) Criterion-based procedures

```
## Subset selection object
## Call: regsubsets.formula(life_exp ~ ., data = df_val)
## 7 Variables (and intercept)
```

```
##           Forced in Forced out
## population      FALSE      FALSE
## income          FALSE      FALSE
## illiteracy      FALSE      FALSE
## murder          FALSE      FALSE
## hs_grad         FALSE      FALSE
## frost           FALSE      FALSE
## area            FALSE      FALSE
## 1 subsets of each size up to 7
## Selection Algorithm: exhaustive
##           population income illiteracy murder hs_grad frost area
## 1 ( 1 ) " "           " "           " "           "*"      " "      " "      " "
## 2 ( 1 ) " "           " "           " "           "*"      "*"      " "      " "
## 3 ( 1 ) "*"           " "           " "           "*"      "*"      " "      " "
## 4 ( 1 ) "*"           " "           " "           "*"      "*"      "*"      " "
## 5 ( 1 ) "*"           " "           " "           "*"      "*"      "*"      "*"
## 6 ( 1 ) "*"           " "           "*"           "*"      "*"      "*"      "*"
## 7 ( 1 ) "*"           "*"           "*"           "*"      "*"      "*"      "*"

```



The Mallows's Cp criterion and Adjusted R^2 suggests that the model with four parameters (population, murder, hs_grad, frost) is the best fit.

e) The LASSO method

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##           s0

```

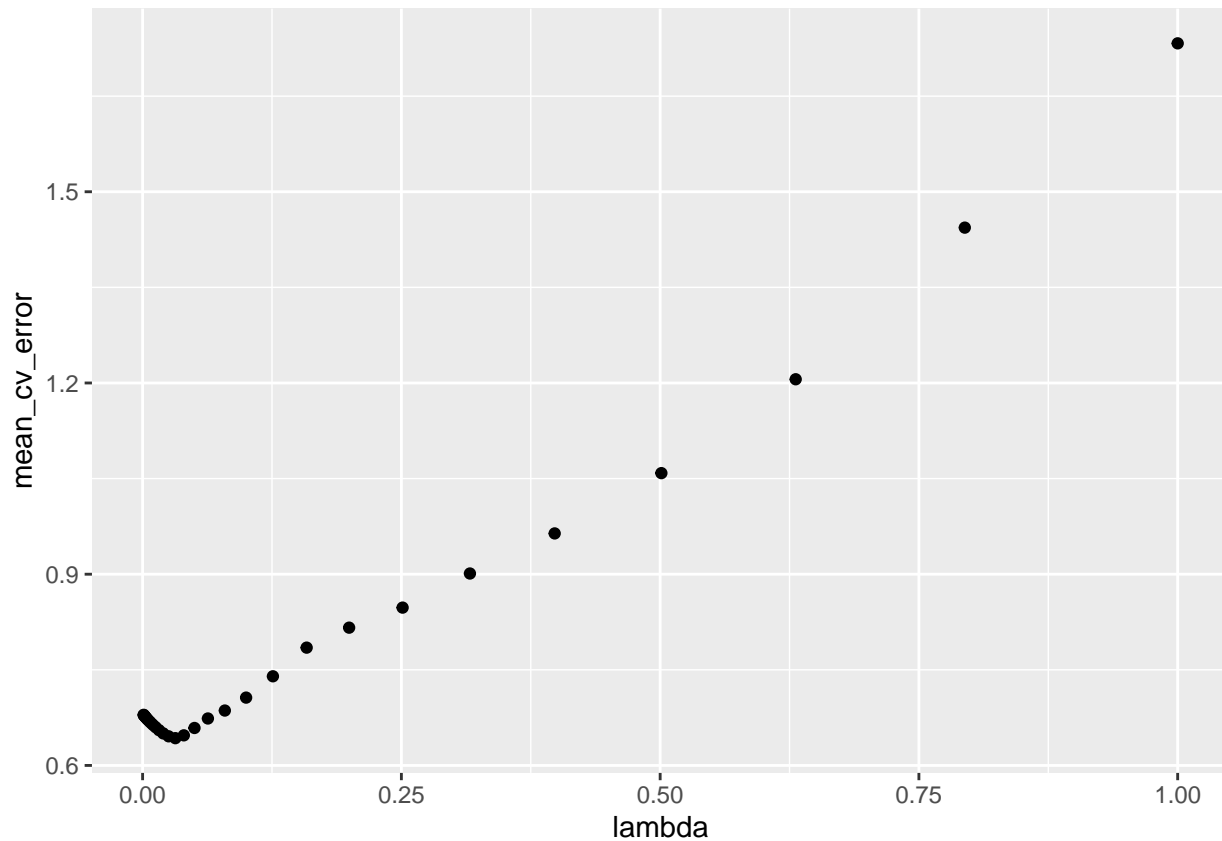
```

## (Intercept) 70.8786
## population  0.0000
## income      .
## illiteracy  .
## murder      .
## hs_grad     .
## frost       .
## area        .

## 8 x 1 sparse Matrix of class "dgCMatrix"
##              s0
## (Intercept) 69.624719212
## population  0.132311050
## income      .
## illiteracy  .
## murder      -0.238481455
## hs_grad     0.040070763
## frost       -0.001484793
## area        .

##
## Call:  cv.glmnet(x = as.matrix(df_val[2:8]), y = df_val$life_exp, lambda = lambda_seq,      nfolds =
##
## Measure: Mean-Squared Error
##
##      Lambda Index Measure      SE Nonzero
## min 0.03162    16  0.6428 0.1960         5
## 1se 0.19953     8  0.8161 0.3026         2

```

```
## [1] 0.03162278
```

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##               s0
## (Intercept) 68.904248047
## population   0.208096331
## income      .
## illiteracy   .
## murder      -0.277654060
## hs_grad      0.048555587
## frost        -0.004090184
## area         0.022049544
```

The lambda value that minimizes the test MSE turns out to be 0.0316228. The final model produced by the optimal lambda value does not include **income** and **illiteracy** because they were not influential enough.

f) Compare the subsets from c, d, and e

```
## Linear Regression
##
## 50 samples
## 4 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
```

```

## Summary of sample sizes: 46, 45, 46, 43, 46, 45, ...
## Resampling results:
##
##      RMSE      Rsquared    MAE
##    0.743866  0.7127871  0.6417565
##
## Tuning parameter 'intercept' was held constant at a value of TRUE

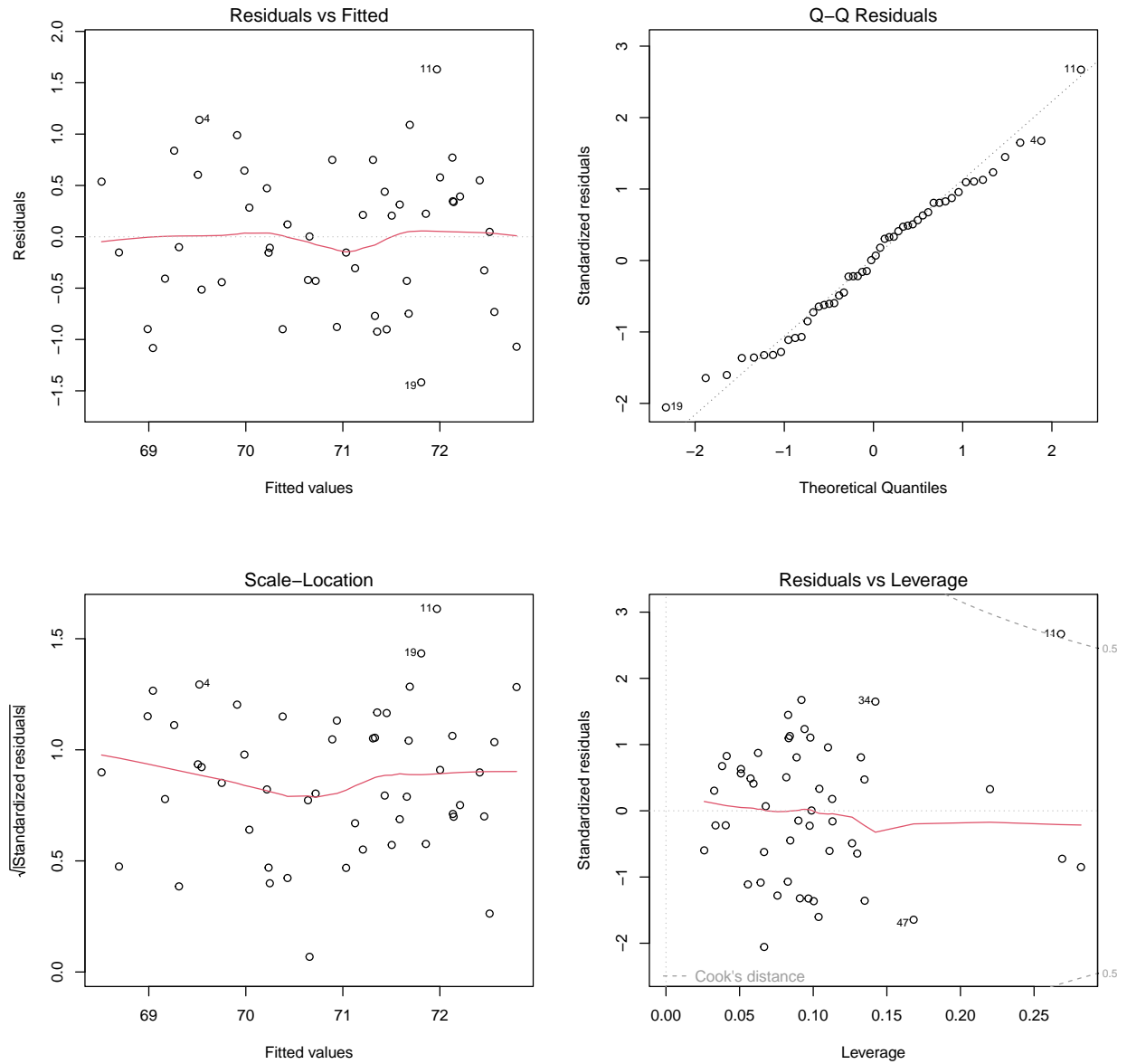
## glmnet
##
## 50 samples
## 5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 46, 46, 46, 45, 44, 43, ...
## Resampling results:
##
##      RMSE      Rsquared    MAE
##    0.7471661  0.7437034  0.6434308
##
## Tuning parameter 'alpha' was held constant at a value of 1
## Tuning
## parameter 'lambda' was held constant at a value of 0.03162278

```

The Root Mean Square Errors (RMSEs) from linear regression model (selected by c and d) and LASSO regression model (by e) indicate that the linear model has slightly better predictive ability in the testing data set. Thus, I will employ the linear regression model as my final model.

Let's check the model assumptions.

In the Residuals vs Fitted plot, the residual values bounce around 0, indicating that the error variances are equal. In Q-Q plot, the points are mostly aligned on a straight line, so we can say that the residuals are normally distributed. Variance of residuals is almost constant over the range of X given the Scale-Location plot. Finally, the Residuals vs Leverage plot shows that there are no significant outliers or influential observations in this model.



g) Findings

A multiple regression model was constructed to investigate life expectancy and its associated variables. As a first step, **population**, **literacy**, and **area** were log-transformed due to skewed distributions. Then, three methods (automatic procedure, criterion-based procedure, and LASSO method) were used to determine which variables to include in the final model. To check the model predictive ability, we conducted a 10-fold cross-validation. The performance of the models was compared by looking at the root mean square error (RMSE): the linear regression models by **murder**, **hs_grad**, **population** and **frost** showed the best predictive ability. The model diagnostics were also tested to ensure that there were no significant violations of the model assumptions. The results of these models are summarized in the table below.

With $\alpha = 0.05$, all variables included in the final model showed statistically significant associations with life expectancy. We see that **murder** and **frost** have a negative impact on life expectancy, while **hs_grad** and **population** can positively affect the outcome.

term	estimate	std.error	statistic	p.value
(Intercept)	68.721	1.417	48.503	<0.001
murder	-0.290	0.035	-8.183	<0.001
hs_grad	0.055	0.015	3.696	0.001
population	0.247	0.113	2.193	0.033
frost	-0.005	0.002	-2.085	0.043