Homework 5

Mari Sanders

2024-12-08

a)

```
state_data <-
  state.x77 %>% as_tibble() %>% janitor::clean_names()
state_data %>% summary() %>% knitr::kable()
```

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

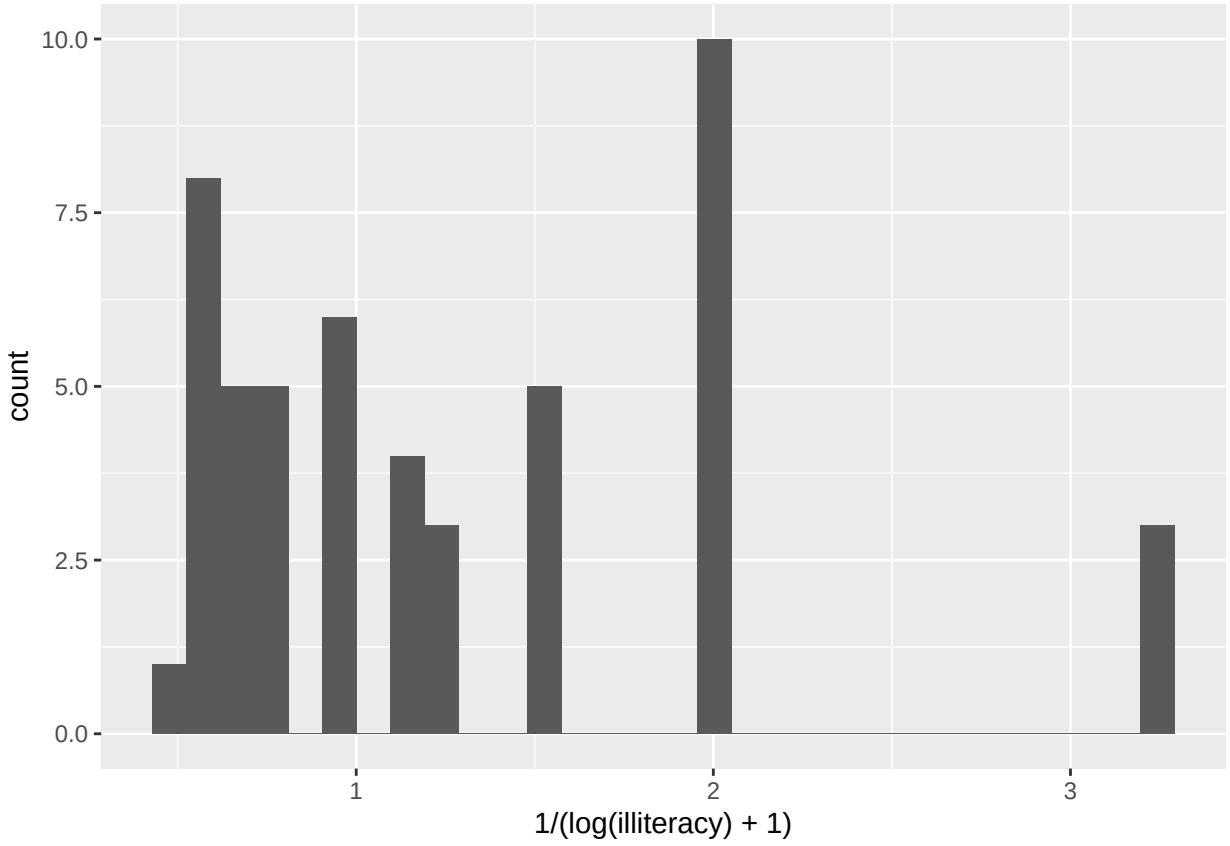
```
state_data %>%
  summarize(population_sd = sd(population, na.rm = TRUE),
            income_sd = sd(income, na.rm = TRUE),
            illiteracy_sd = sd(illiteracy, na.rm = TRUE),
            lifeexpec_sd = sd(life_exp, na.rm = TRUE),
            murder_sd = sd(murder, na.rm = TRUE),
            hsgrad_sd = sd(hs_grad, na.rm = TRUE),
            frost_sd = sd(frost, na.rm = TRUE),
            area_sd = sd(area, na.rm = TRUE)) %>% knitr::kable()
```

population_sd	income_sd	illiteracy_sd	lifeexpec_sd	murder_sd	hsgrad_sd	frost_sd	area_sd
4464.491	614.4699	0.6095331	1.342394	3.69154	8.076998	51.98085	85327.3

b)

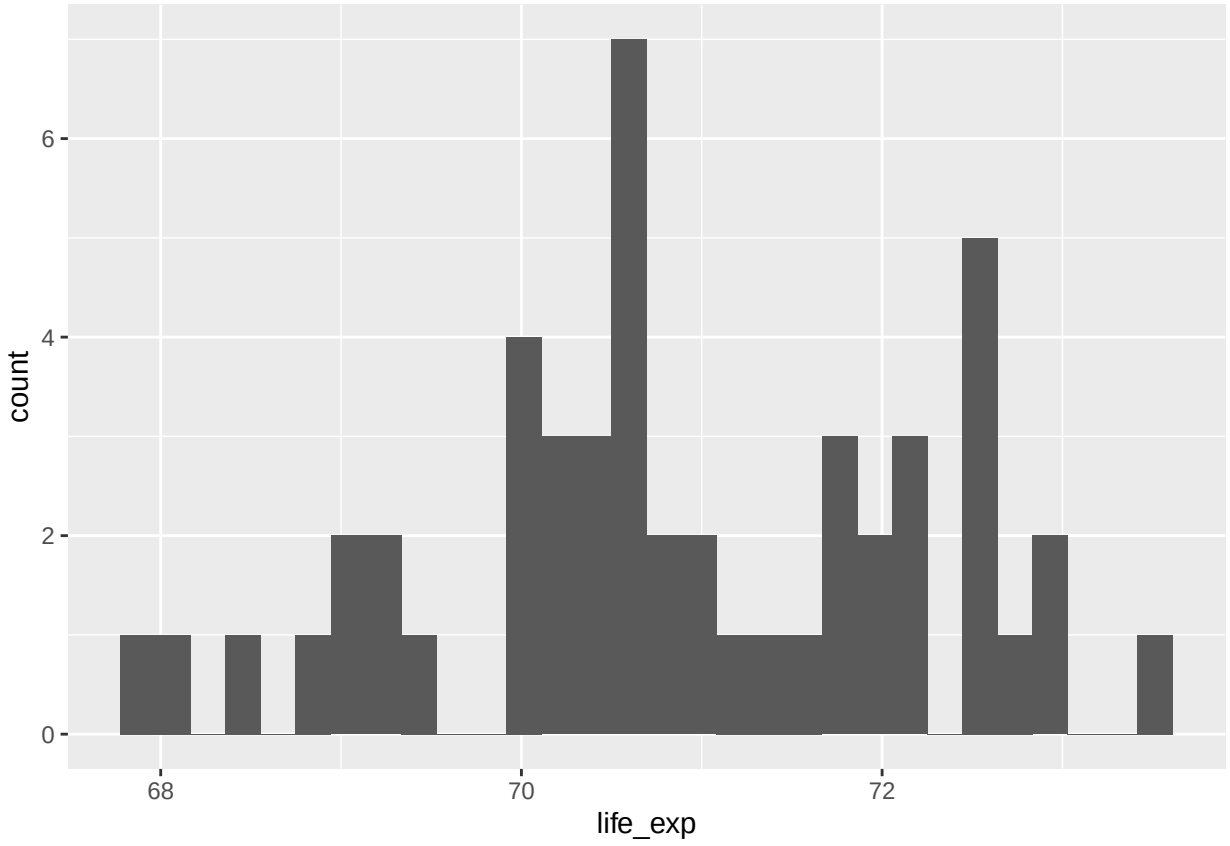
```
pop <- ggplot(state_data, aes(x = log(population))) + geom_histogram()

income <- ggplot(state_data, aes(x = income)) + geom_histogram()

ggplot(state_data, aes(x = 1/(log(illiteracy) + 1))) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



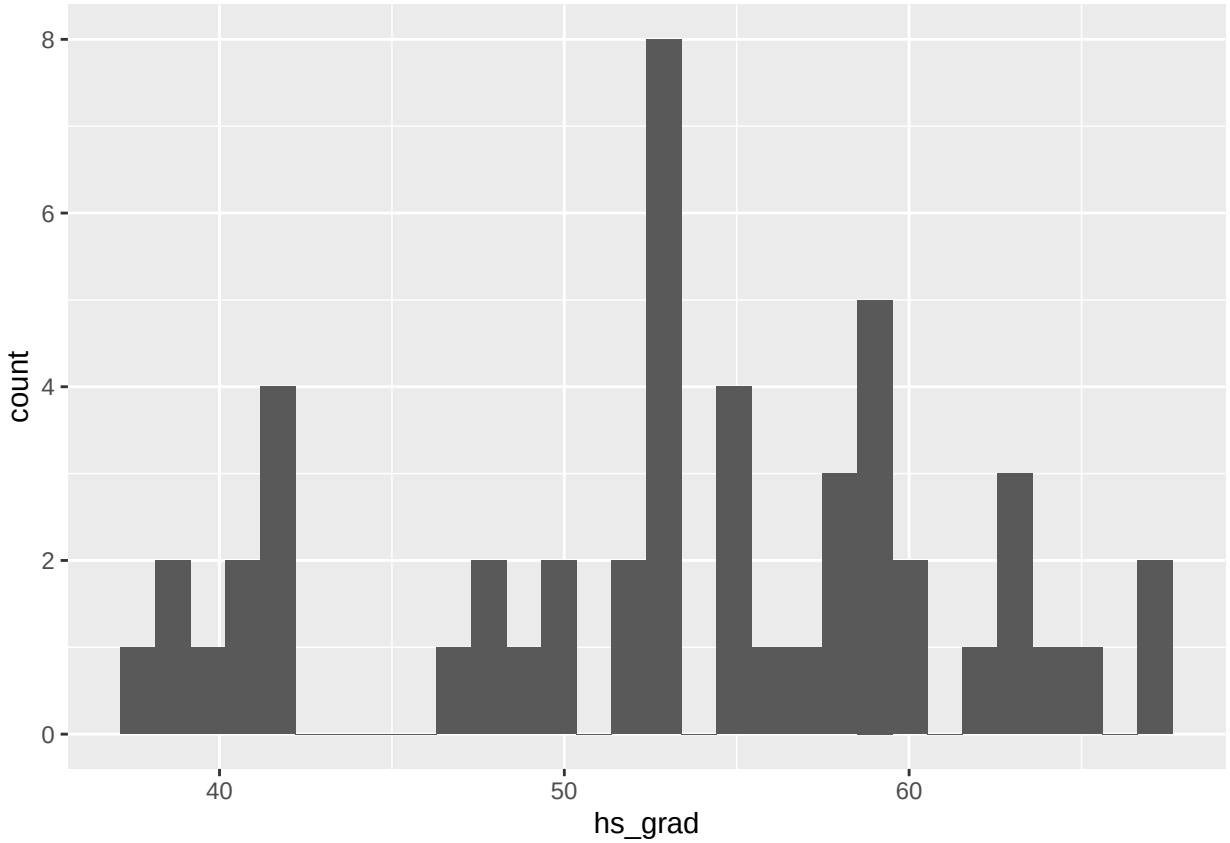

```
ggplot(state_data, aes(x = life_exp)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



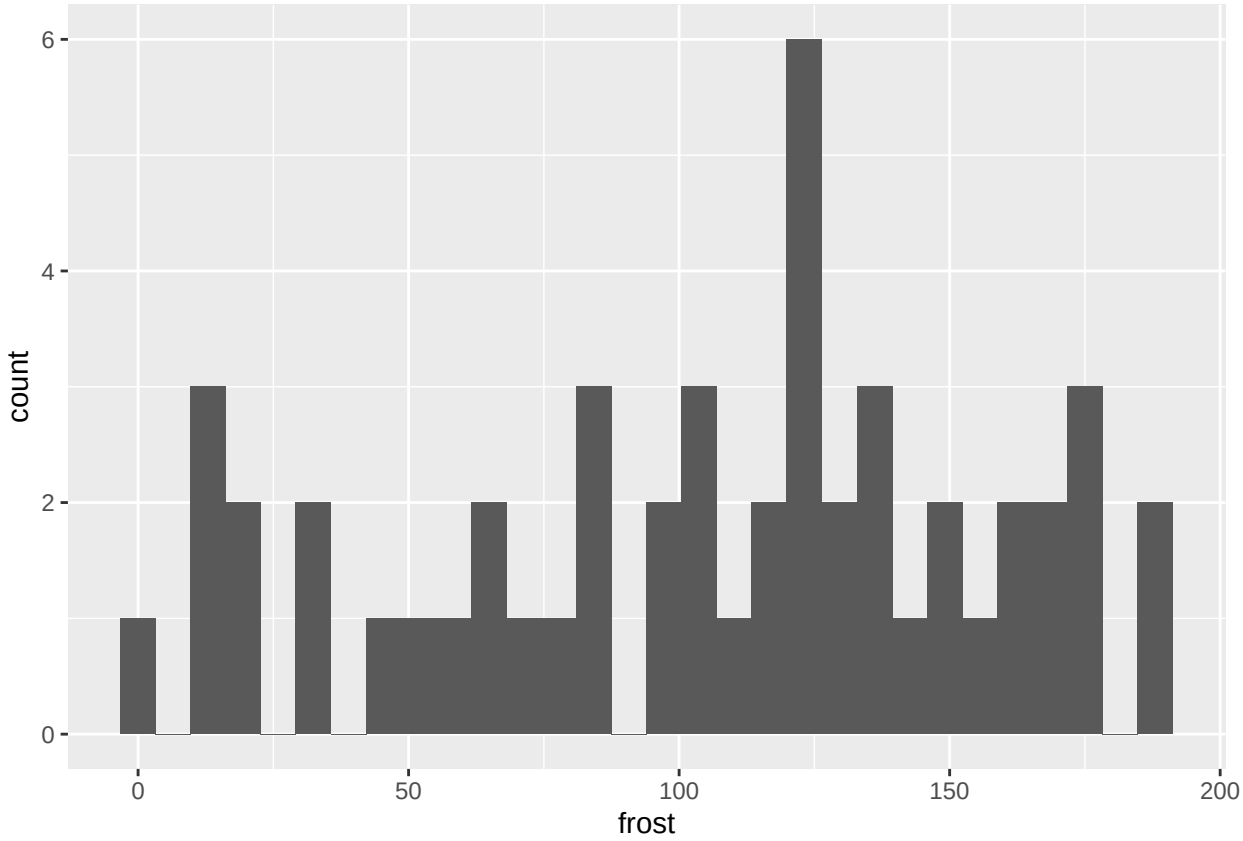

```
ggplot(state_data, aes(x = hs_grad)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



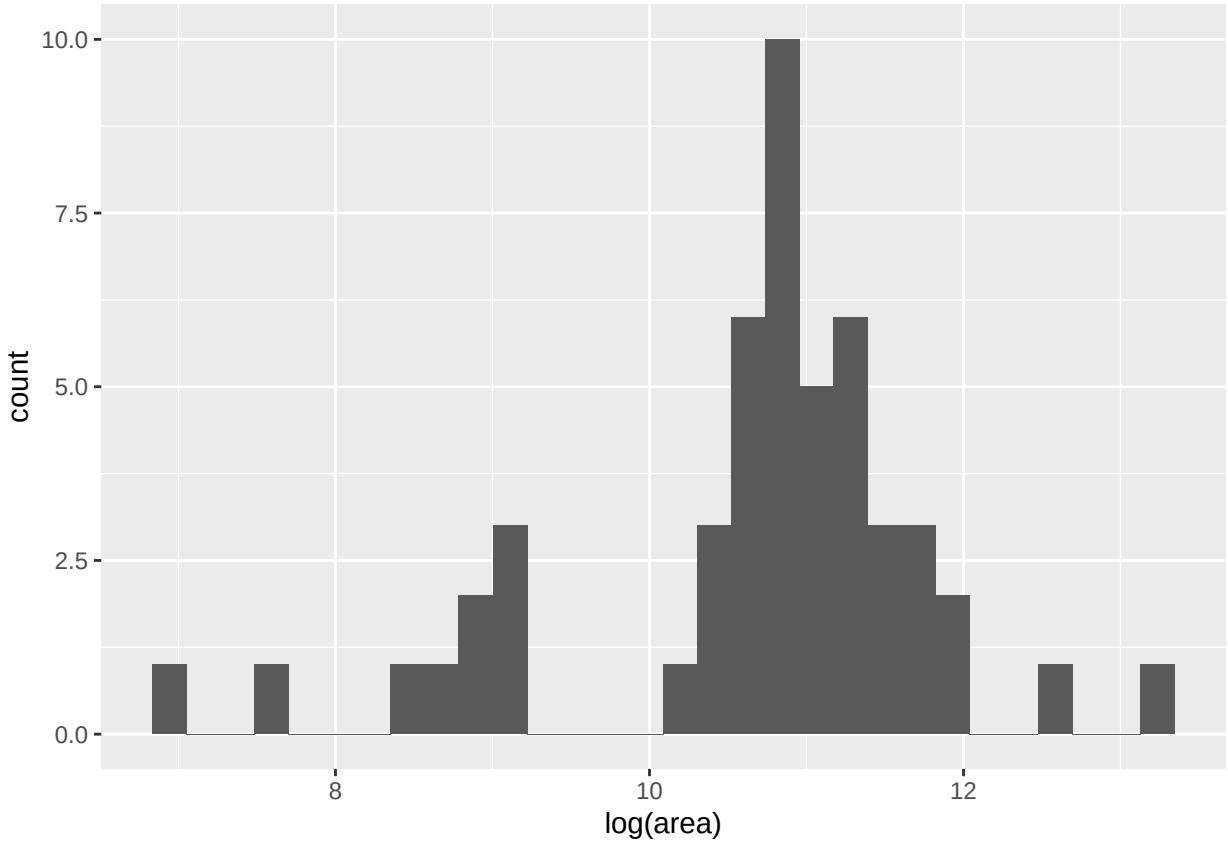

```
ggplot(state_data, aes(x = frost)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(state_data, aes(x = log(area))) + geom_histogram()
```

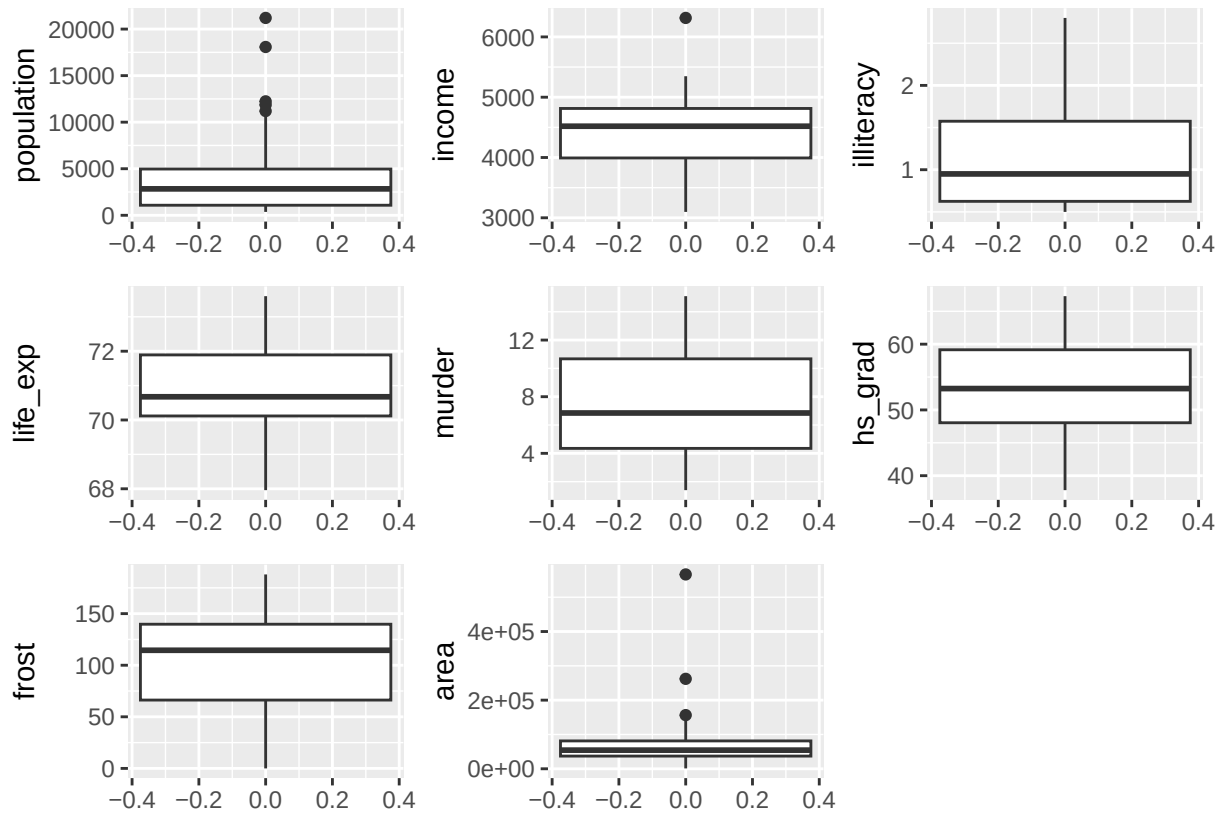
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
par(nfrow = c(2,3))
```

```
## Warning in par(nfrow = c(2, 3)): "nfrow" is not a graphical parameter
```

```
pop <- ggplot(state_data, aes(y = population)) + geom_boxplot()
income <- ggplot(state_data, aes(y = income)) + geom_boxplot()
illiteracy <- ggplot(state_data, aes(y = illiteracy)) + geom_boxplot()
life_exp <- ggplot(state_data, aes(y = life_exp)) + geom_boxplot()
murder <- ggplot(state_data, aes(y = murder)) + geom_boxplot()
grad <- ggplot(state_data, aes(y = hs_grad)) + geom_boxplot()
frost <- ggplot(state_data, aes(y = frost)) + geom_boxplot()
area <- ggplot(state_data, aes(y = area)) + geom_boxplot()

pop + income + illiteracy + life_exp + murder + grad + frost + area
```



Life expectancy, and murder seem to have a relationship, as well as life expectancy and high school grad and life expectancy and illiteracy. There seems to be a slight relationship between life expectancy and frost. We ended up transforming `population`, `illiteracy` and `area` because they looked skewed when plotted.

```
state_data <-
  state_data %>%
  mutate(population = log(population),
         illiteracy = 1/(log(illiteracy + 1)),
         area = log(area))
```

c)

```
full_model <- lm(life_exp ~ ., data = state_data)
only_1 <- lm(life_exp ~ 1, data = state_data)
summary(full_model)
```

```
##
## Call:
## lm(formula = life_exp ~ ., data = state_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4595 -0.4220  0.0533  0.4947  1.7016
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.833e+01  1.550e+00  44.081  < 2e-16 ***
## population   2.457e-01  1.273e-01   1.930   0.0604 .
## income      -1.397e-05  2.530e-04  -0.055   0.9562
## illiteracy  -1.612e-01  3.839e-01  -0.420   0.6767
## murder      -3.099e-01  4.568e-02  -6.784 2.97e-08 ***
## hs_grad      5.448e-02  2.552e-02   2.135   0.0386 *
## frost       -4.818e-03  3.049e-03  -1.580   0.1216
## area         7.662e-02  1.116e-01   0.687   0.4961
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7337 on 42 degrees of freedom
## Multiple R-squared:  0.744,  Adjusted R-squared:  0.7013
## F-statistic: 17.43 on 7 and 42 DF,  p-value: 1.384e-10
```

```
forward_model <- step(only_1, direction = "forward", scope = formula(full_model))
```

```
## 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    25.216 63.083  15.621
## + 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    0.0568 34.405 -12.692
##
## 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    1.0108 28.759 -19.653
## + 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
## <none>                    25.135 -26.387
## + illiteracy  1   0.73828 24.397 -25.878
## + 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  0.030298 22.891 -27.064
## + income      1  0.010673 22.911 -27.021
```

```
summary(forward_model)
```

```
##
## Call:
## lm(formula = life_exp ~ murder + hs_grad + population + frost,
##     data = state_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 ***
## 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
```

```
backward_model <- step(full_model, direction = "backward")
```

```
## Start:  AIC=-23.68
## life_exp ~ population + income + illiteracy + murder + hs_grad +
##     frost + area
##
##              Df Sum of Sq    RSS     AIC
## - income      1    0.0016 22.610 -25.680
## - illiteracy  1    0.0949 22.704 -25.475
```

```
## - area        1    0.2538 22.863 -25.126
## <none>                    22.609 -23.684
## - frost       1    1.3437 23.953 -22.797
## - population  1    2.0052 24.614 -21.435
## - hs_grad     1    2.4536 25.062 -20.533
## - murder      1   24.7715 47.380  11.309
##
## Step:  AIC=-25.68
## life_exp ~ population + illiteracy + murder + hs_grad + frost +
##     area
##
##              Df Sum of Sq    RSS      AIC
## - illiteracy  1    0.0950 22.705 -27.4708
## - area        1    0.2804 22.891 -27.0641
## <none>                    22.610 -25.6804
## - frost       1    1.5196 24.130 -24.4281
## - population  1    2.3351 24.946 -22.7663
## - hs_grad     1    4.6277 27.238 -18.3702
## - murder      1   25.0696 47.680   9.6245
##
## 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
```

```
summary(backward_model)
```

```
##
## Call:
## lm(formula = life_exp ~ population + murder + hs_grad + frost,
##     data = state_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 ***
## population   0.246836   0.112539   2.193 0.033491 *
## 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 *
## ---
## 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
```

```
both_model <- step(full_model, direction = "both")
```

```
## Start:  AIC=-23.68
## life_exp ~ population + income + illiteracy + murder + hs_grad +
##     frost + area
##
##              Df Sum of Sq    RSS     AIC
## - income      1    0.0016 22.610 -25.680
## - illiteracy  1    0.0949 22.704 -25.475
## - area        1    0.2538 22.863 -25.126
## <none>                    22.609 -23.684
## - frost       1    1.3437 23.953 -22.797
## - population  1    2.0052 24.614 -21.435
## - hs_grad     1    2.4536 25.062 -20.533
## - murder      1   24.7715 47.380  11.309
##
## Step:  AIC=-25.68
## life_exp ~ population + illiteracy + murder + hs_grad + frost +
##     area
##
##              Df Sum of Sq    RSS      AIC
## - illiteracy  1    0.0950 22.705 -27.4708
## - area        1    0.2804 22.891 -27.0641
## <none>                    22.610 -25.6804
## - frost       1    1.5196 24.130 -24.4281
## + income      1    0.0016 22.609 -23.6841
## - population  1    2.3351 24.946 -22.7663
## - hs_grad     1    4.6277 27.238 -18.3702
## - murder      1   25.0696 47.680   9.6245
##
## 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
## + illiteracy  1    0.0950 22.610 -25.680
## + income      1    0.0017 22.704 -25.475
## - 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
## + area        1     0.216 22.705 -27.471
## + illiteracy  1     0.030 22.891 -27.064
## + income      1     0.011 22.911 -27.021
## - 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
```

```
summary(both_model)
```

```
##
## Call:
## lm(formula = life_exp ~ population + murder + hs_grad + frost,
##     data = state_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 ***
## population   0.246836   0.112539   2.193 0.033491 *
## 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 *
## ---
## 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
```

```
model_nofrost <- lm(life_exp ~ population + murder + hs_grad + frost, data = state_data)

summary(model_nofrost)
```

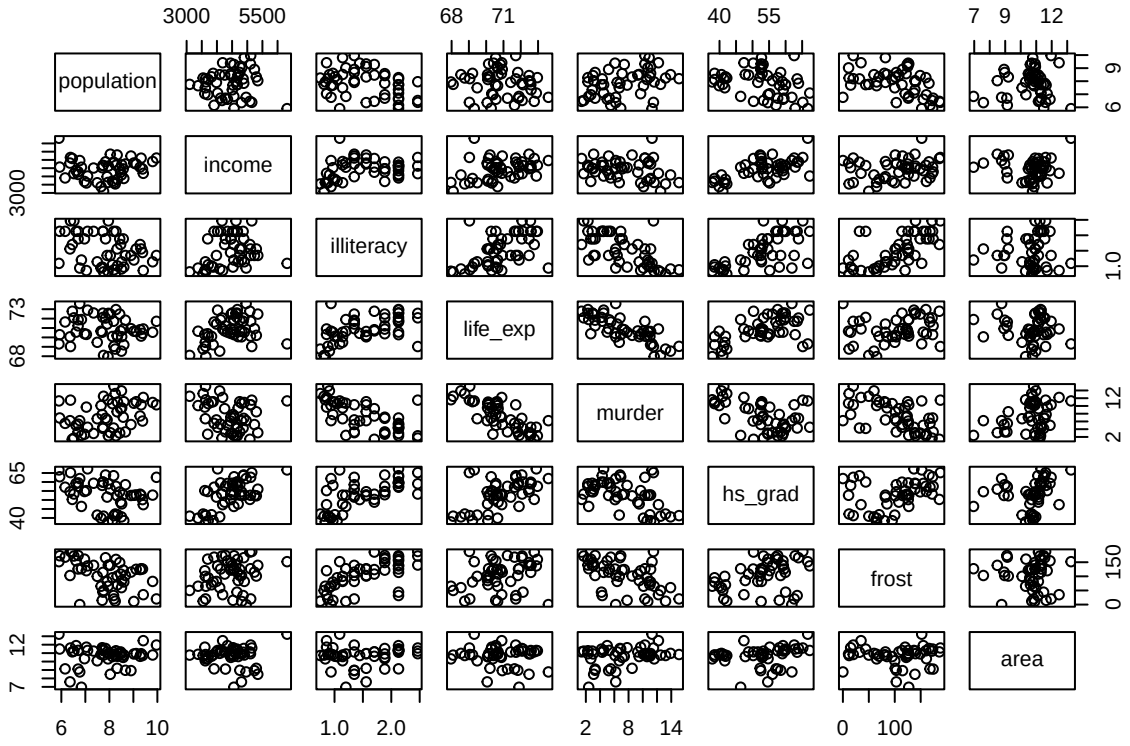
```
##
## Call:
## lm(formula = life_exp ~ population + murder + hs_grad + frost,
##     data = state_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.41760 -0.43880  0.02539  0.52066  1.63048
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 68.720810   1.416828  48.503  < 2e-16 ***
## population   0.246836   0.112539   2.193 0.033491 *
## 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 *
## ---
## 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
```

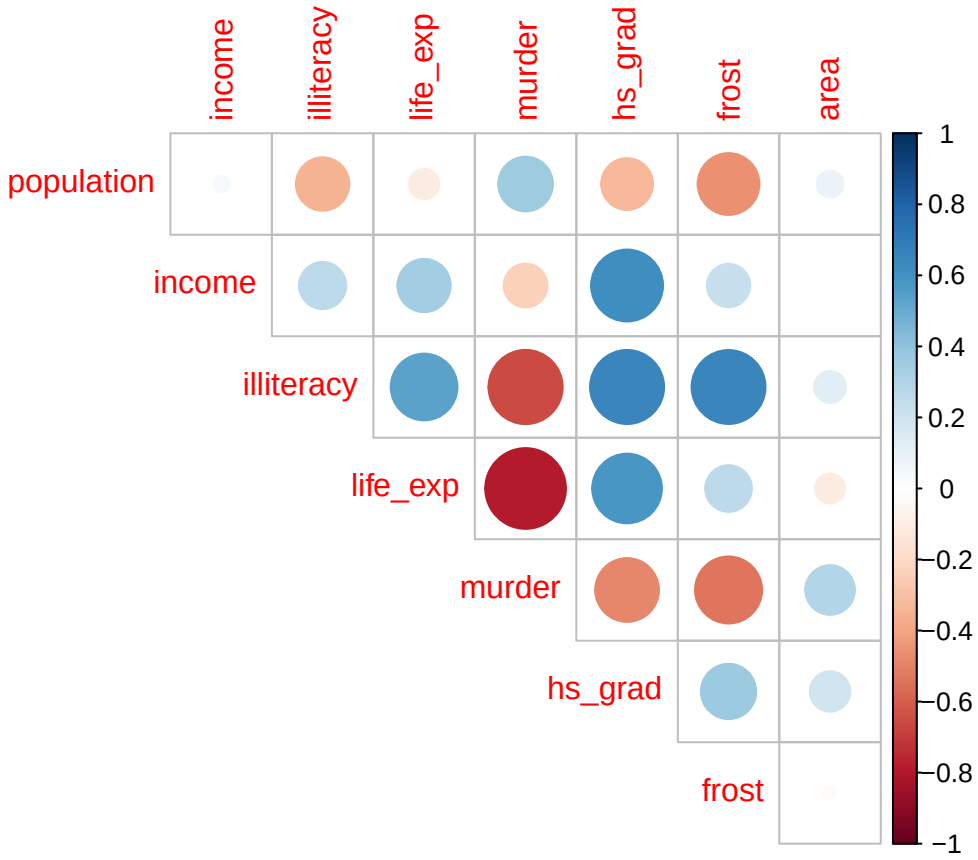
Doing backward, both, and regular subsetting gets the same result to include population, murder, hs_grad, and frost, area. `life_exp` = `population`β_1 + `murder` β_2 + `hs_grad` β_3 + `frost` β_4

The only "close call" variable is `frost`, but the adjusted r-squared value decreased when you take `frost` out, so I chose to keep it in the model.

```
pairs(state_data)
```




```
corrplot(cor(state_data), type = "upper", diag = FALSE)
```



There is a strong relationship between `illiteracy` and `hs_grad`. Our subset only contains `hs_grad` and not `illiteracy`.

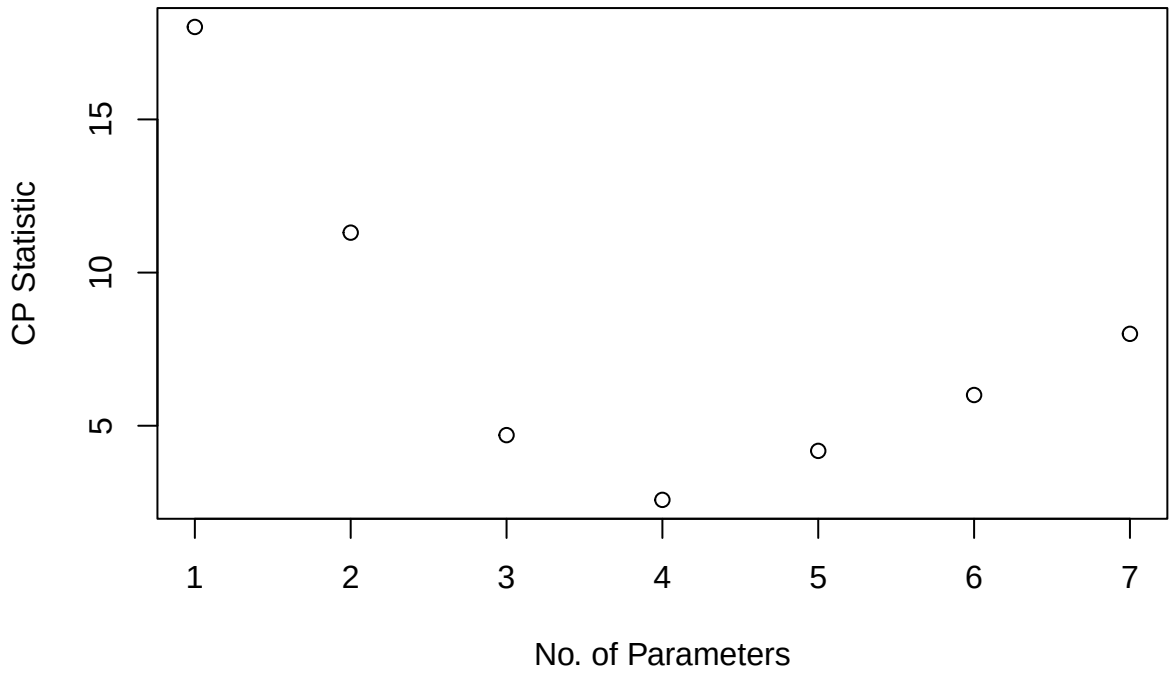
d)

```
subsets = regsubsets(life_exp ~., data = state_data)
subset_res = summary(subsets)

subset_res
```

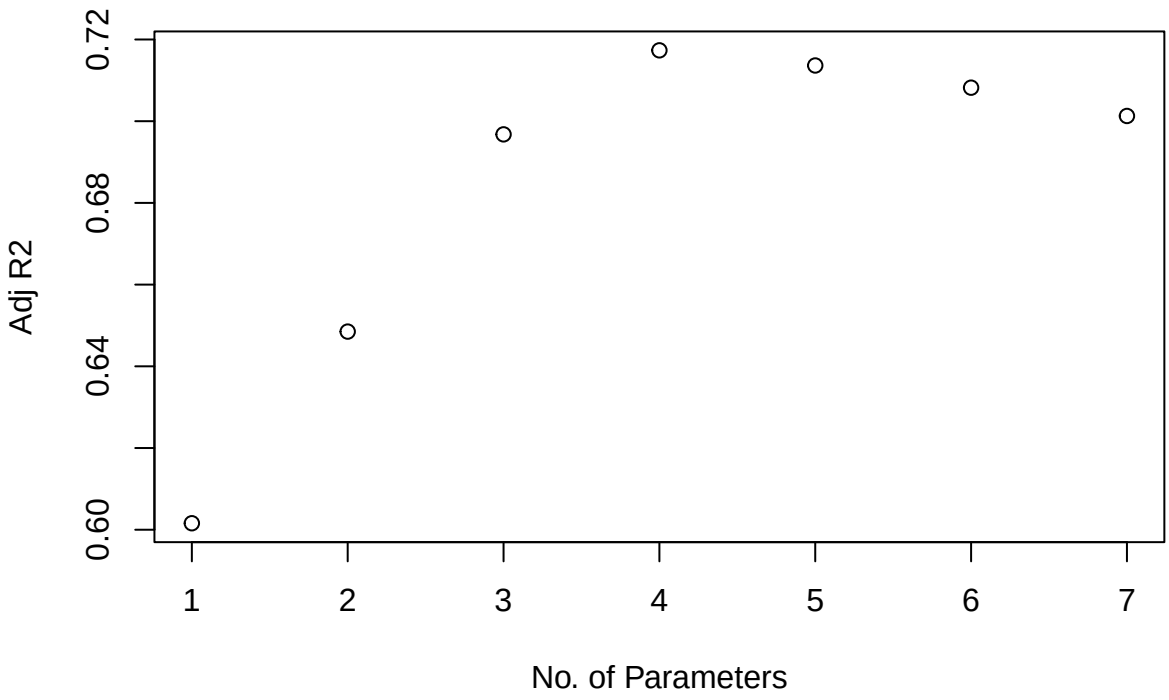
```
## Subset selection object
## Call: regsubsets.formula(life_exp ~ ., data = state_data)
## 7 Variables  (and intercept)
##            Forced in Forced out
## population     FALSE      FALSE
## income         FALSE      FALSE
## illiteracy     FALSE      FALSE
## murder         FALSE      FALSE
## 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 ) "*"        "*"    "*"        "*"    "*"     "*"   "*"
```

```
plot(subset_res$cp, xlab = "No. of Parameters", ylab = "CP Statistic")
```




```
plot(subset_res$adjr2, xlab = "No. of Parameters", ylab = "Adj R2")
```



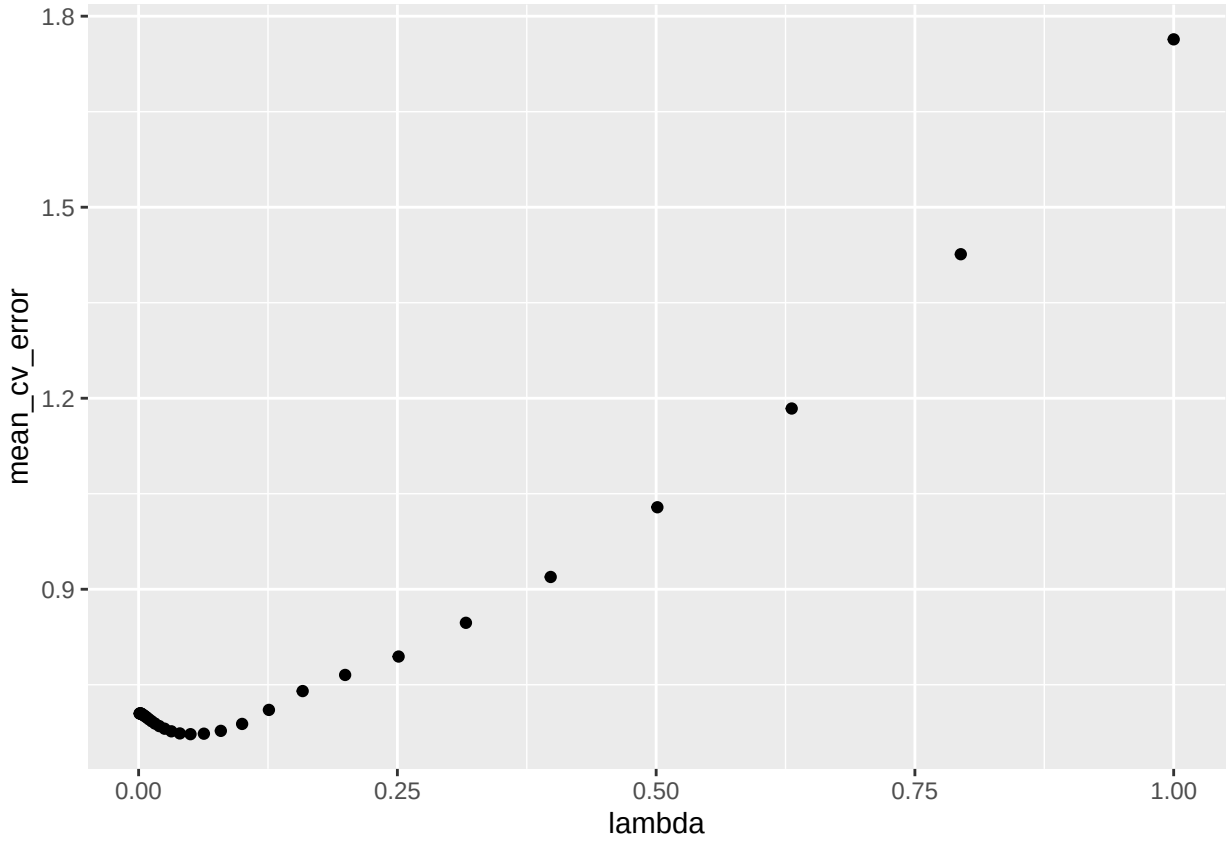
From the regular subsetting, we can see that 4 parameters is the best model, which is what we got from forward, backward, and a combination of both in subsetting for a model.

e)

```
data <- state_data %>% select(-life_exp)
set.seed(5)
lambda_seq <- 10^seq(-3,0, by = 0.1)
cv_object <- cv.glmnet(as.matrix(data[1:7]), state_data$life_exp, lambda = lambda_seq, nfolds = 5)

cv_object
```

```
##
## Call:  cv.glmnet(x = as.matrix(data[1:7]), y = state_data$life_exp,      lambda = lambda_seq, nfolds
##
## Measure: Mean-Squared Error
##
##      Lambda Index Measure     SE Nonzero
## min 0.05012    14  0.6724 0.1240       4
## 1se 0.25119     7  0.7943 0.1719       2
```

```
tibble(lambda = cv_object$lambda,
       mean_cv_error = cv_object$cvm) %>%
  ggplot(aes(x = lambda, y = mean_cv_error)) + geom_point()
```




```
cv_object$lambda.min
```

```
## [1] 0.05011872
```

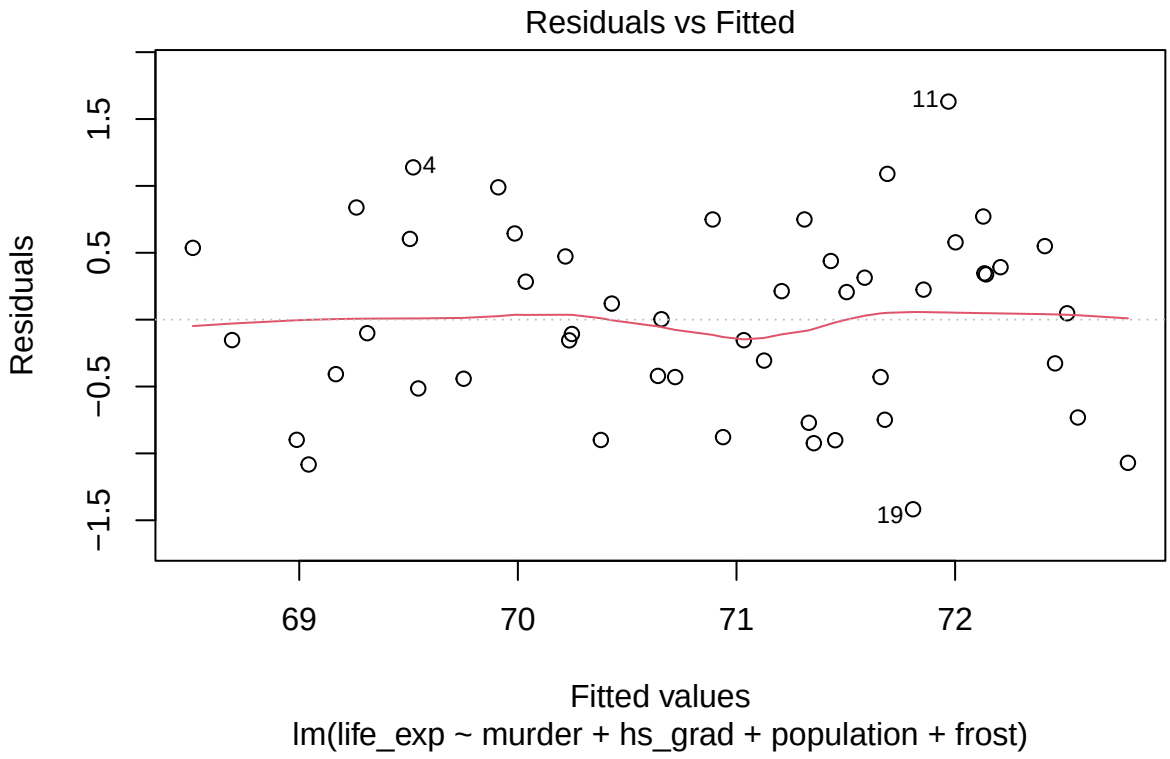
```
fit_bestcv <- glmnet(as.matrix(data[1:7]), state_data$life_exp, lambda = cv_object$lambda.min)
coef(fit_bestcv)
```

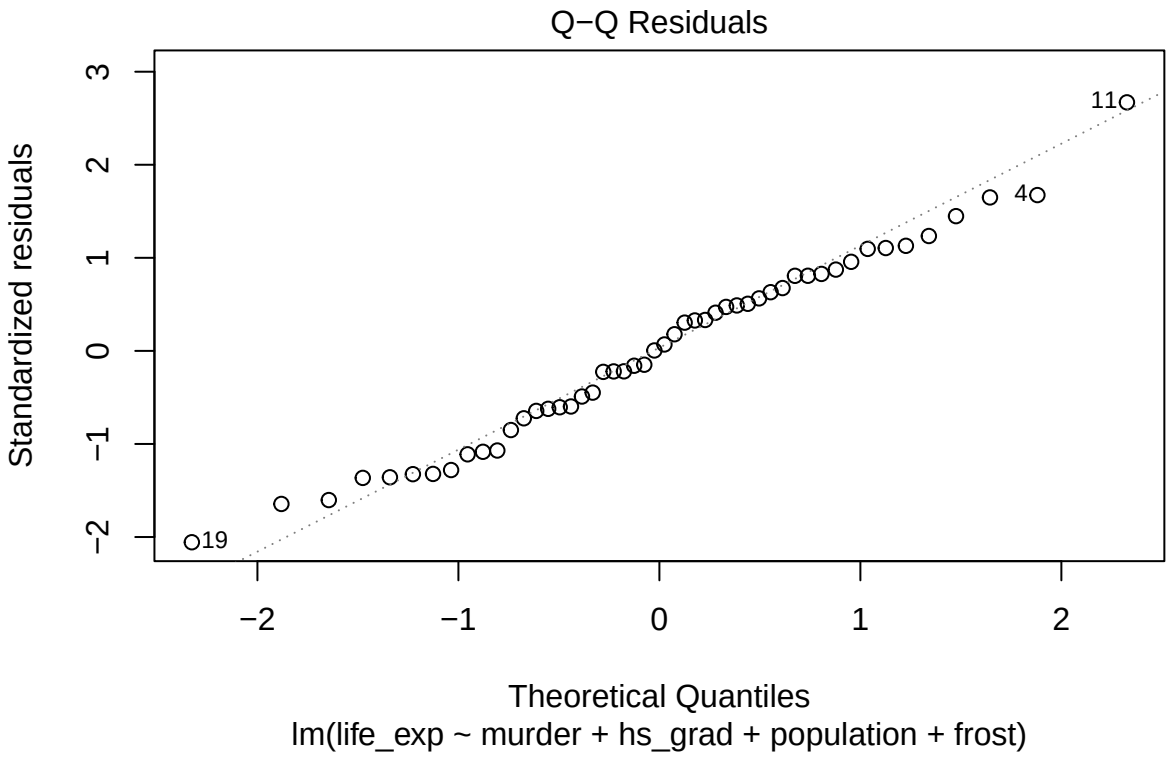
```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                       s0
## (Intercept) 69.17386637
## population   0.18944379
## income       .
## illiteracy   .
## murder      -0.26418480
## hs_grad      0.04729095
## frost       -0.00332490
## area         .
```

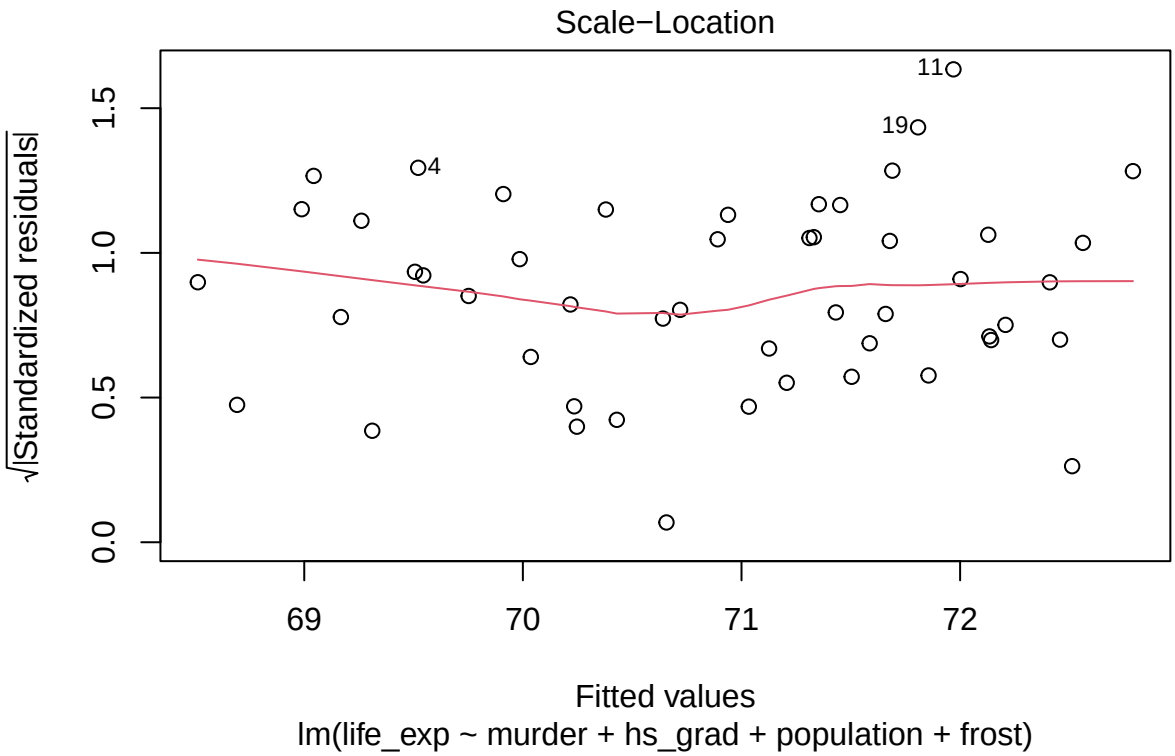
The best lambda for our model is 0.0501187.

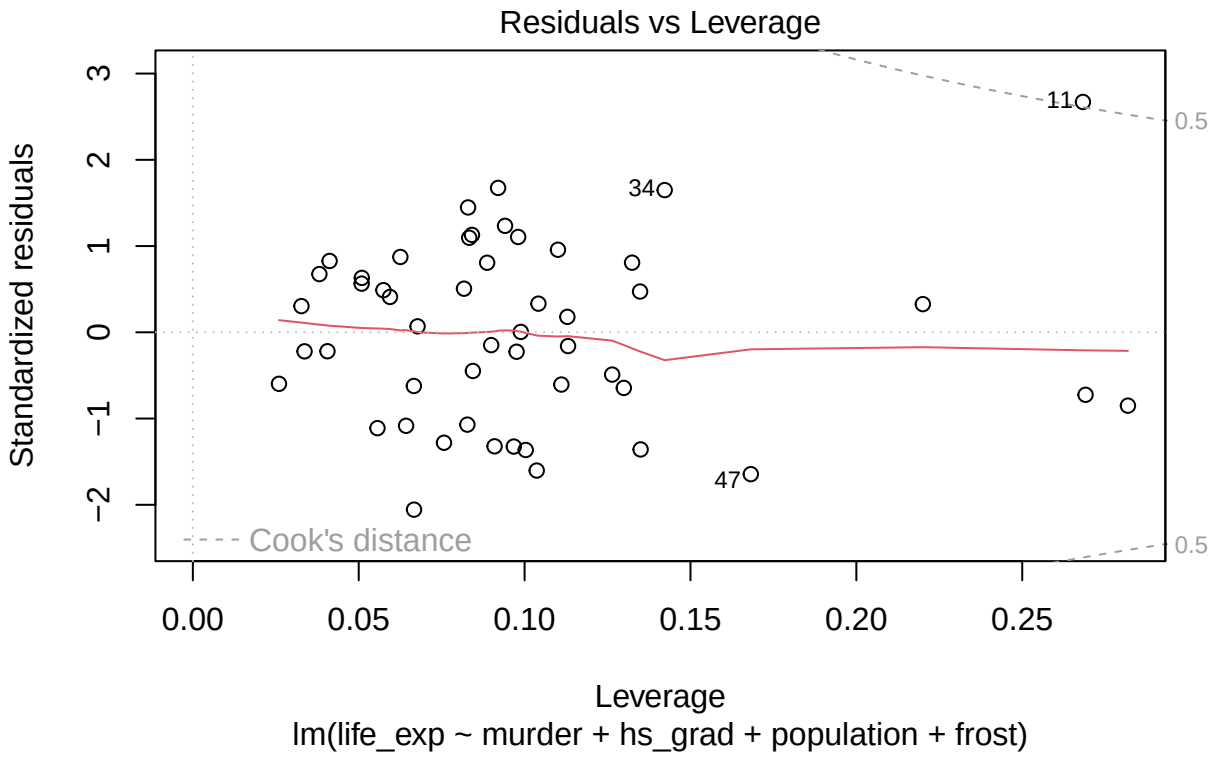
f)

```
plot(forward_model)
```

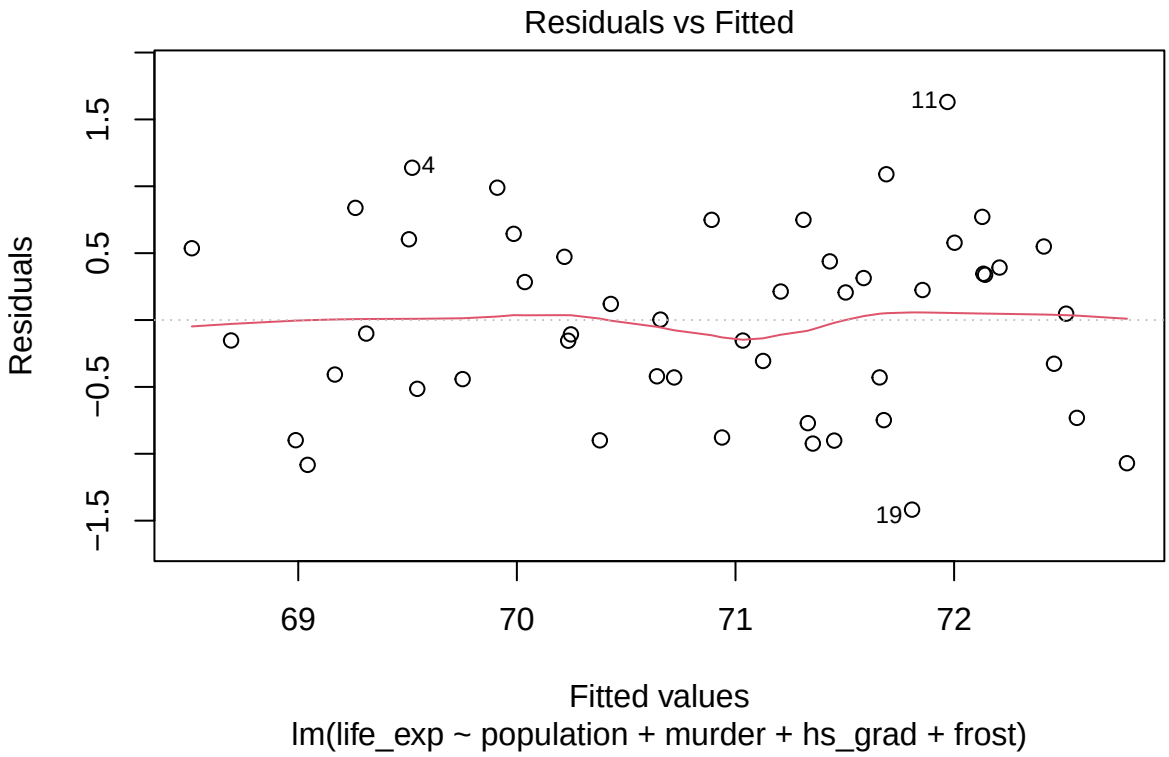


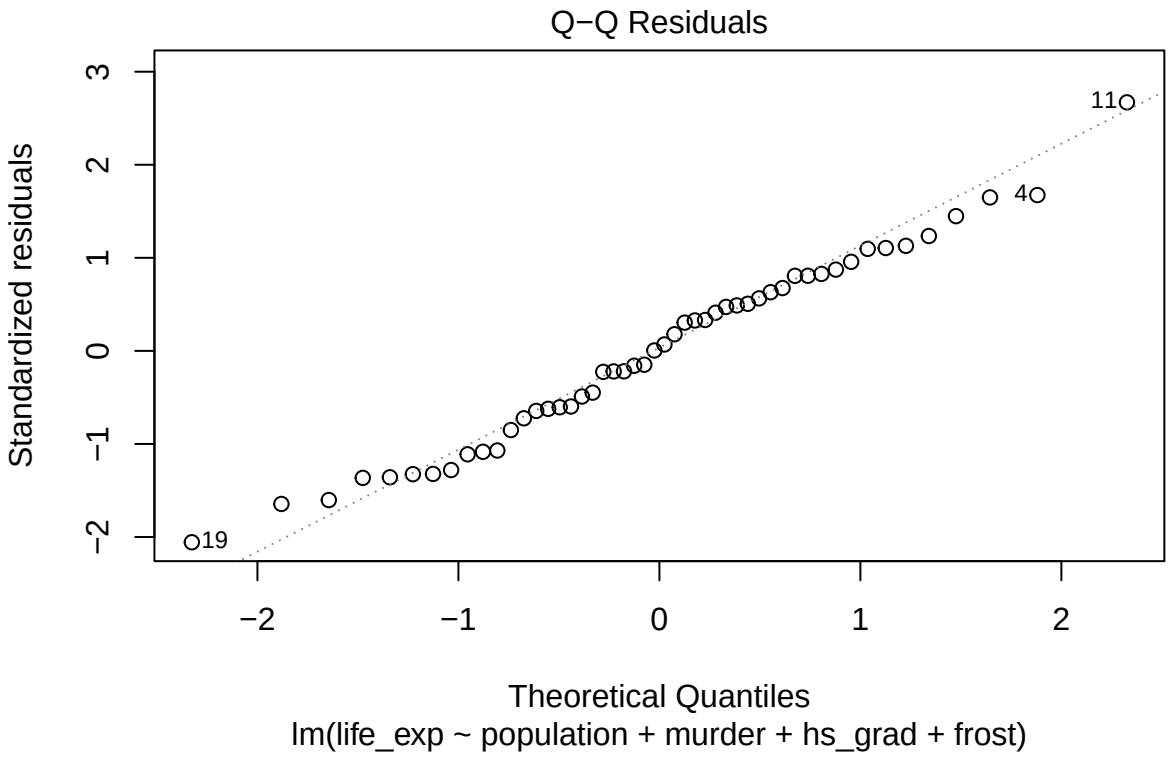

Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(life_exp ~ murder + hs_grad + population + frost)
11
4
19
−2
−1
0
1
2
3

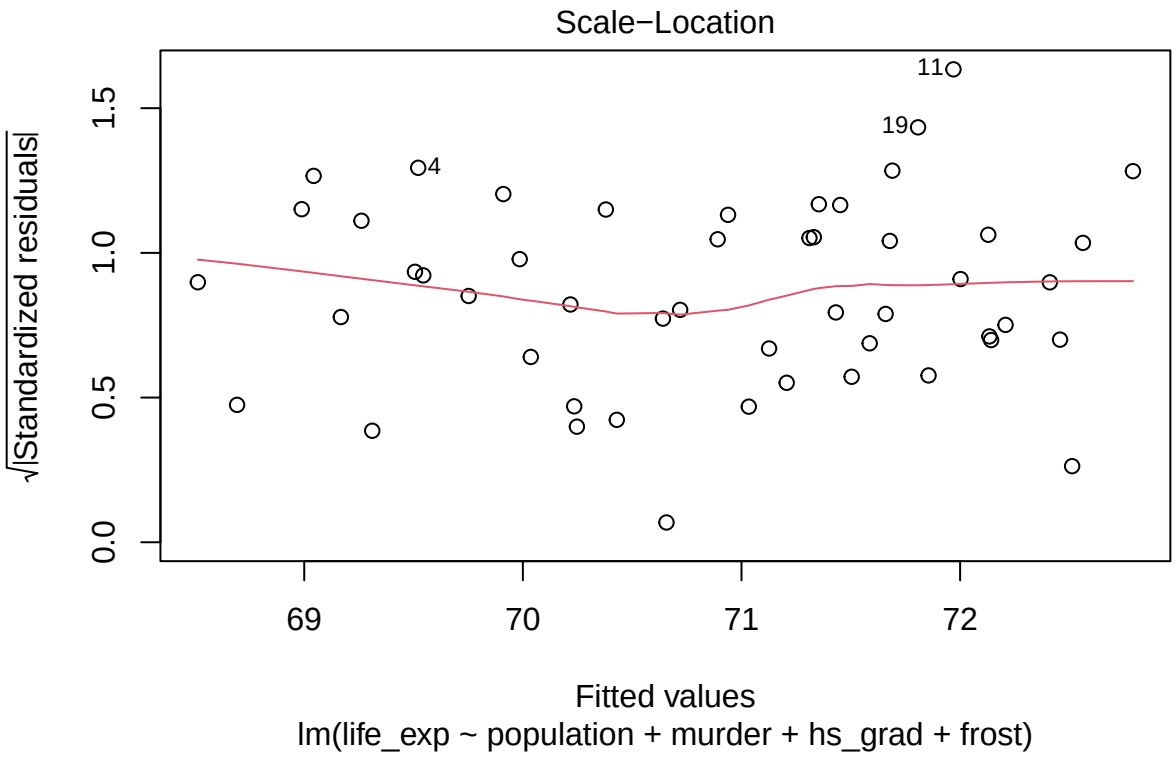

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
69
70
71
72
11
19
4
Fitted values
lm(life_exp ~ murder + hs_grad + population + frost)

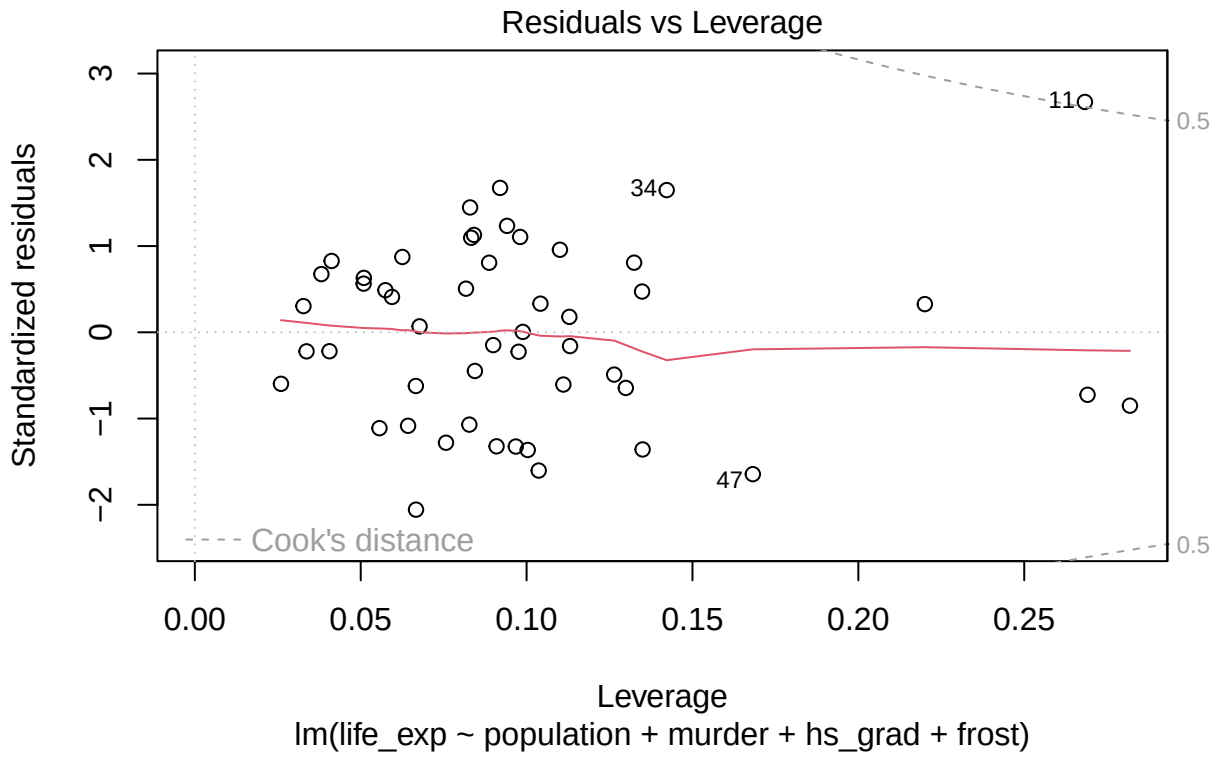



```
plot(backward_model)
```

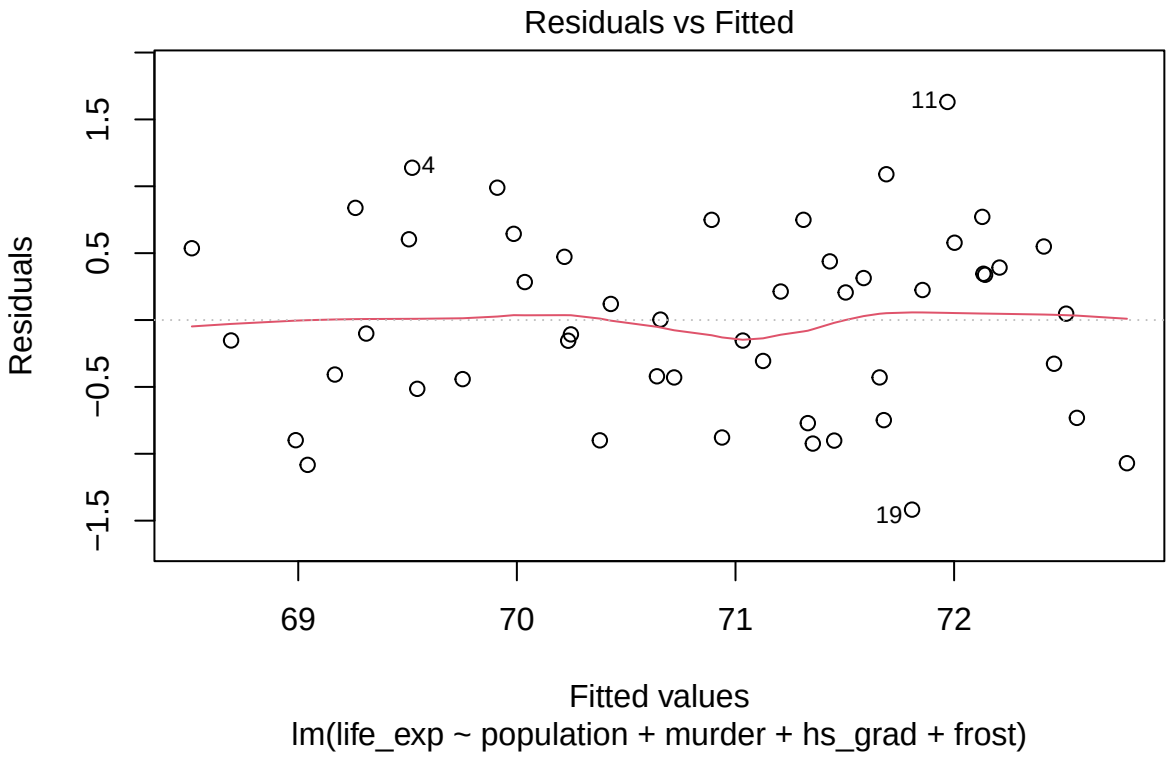

Residuals vs Fitted
Residuals
1.5
0.5
−0.5
−1.5
4
11
19
69
70
71
72
Fitted values
lm(life_exp ~ population + murder + hs_grad + frost)

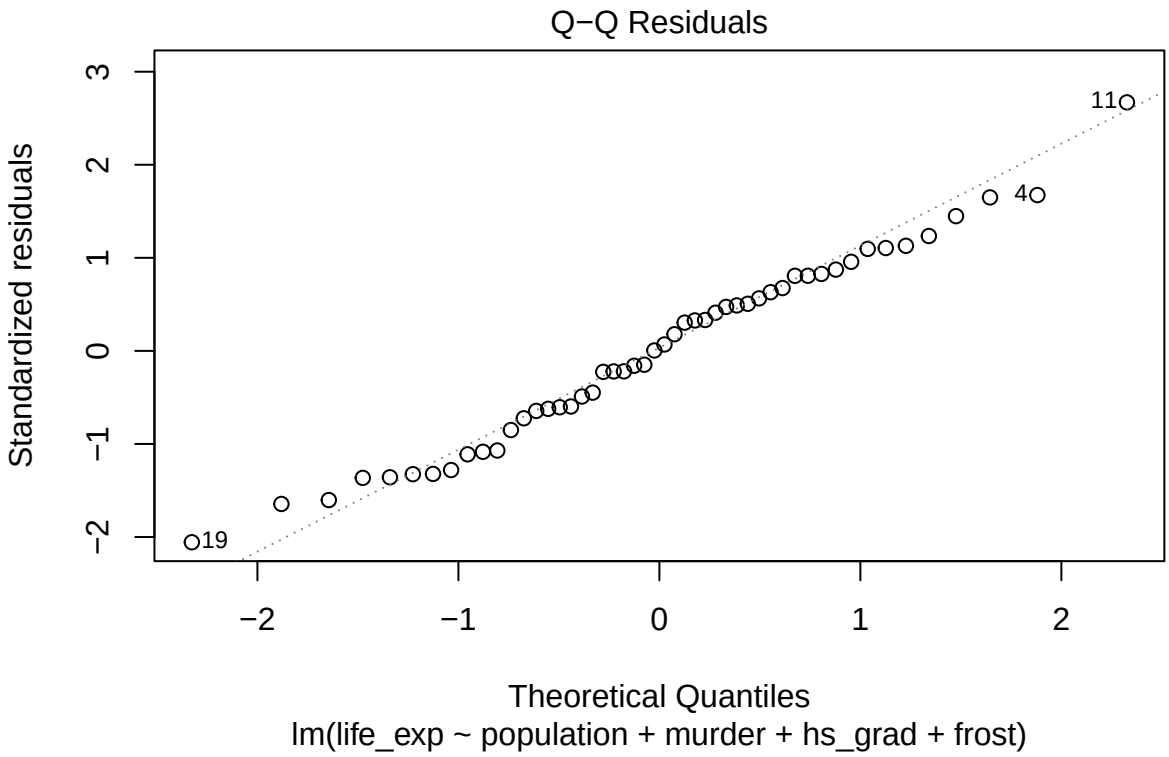

Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(life_exp ~ population + murder + hs_grad + frost)
11
4
19
−2
−1
0
1
2
3

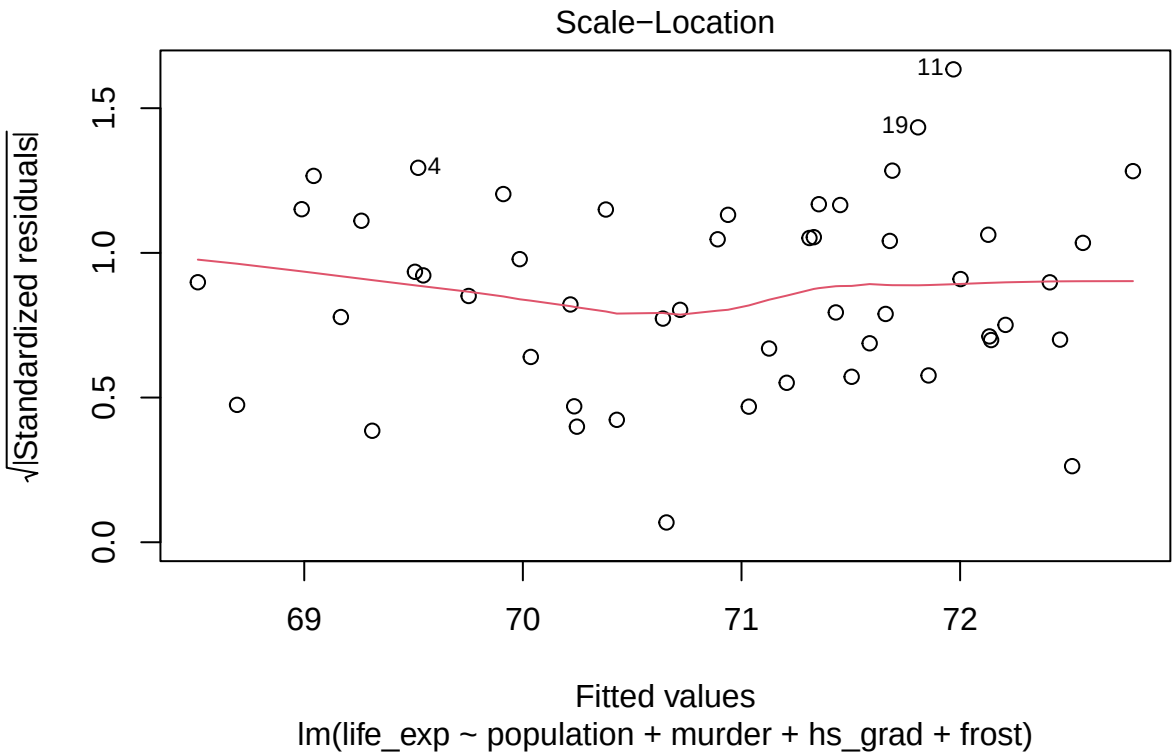

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
69
70
71
72
11
19
4
Fitted values
lm(life_exp ~ population + murder + hs_grad + frost)

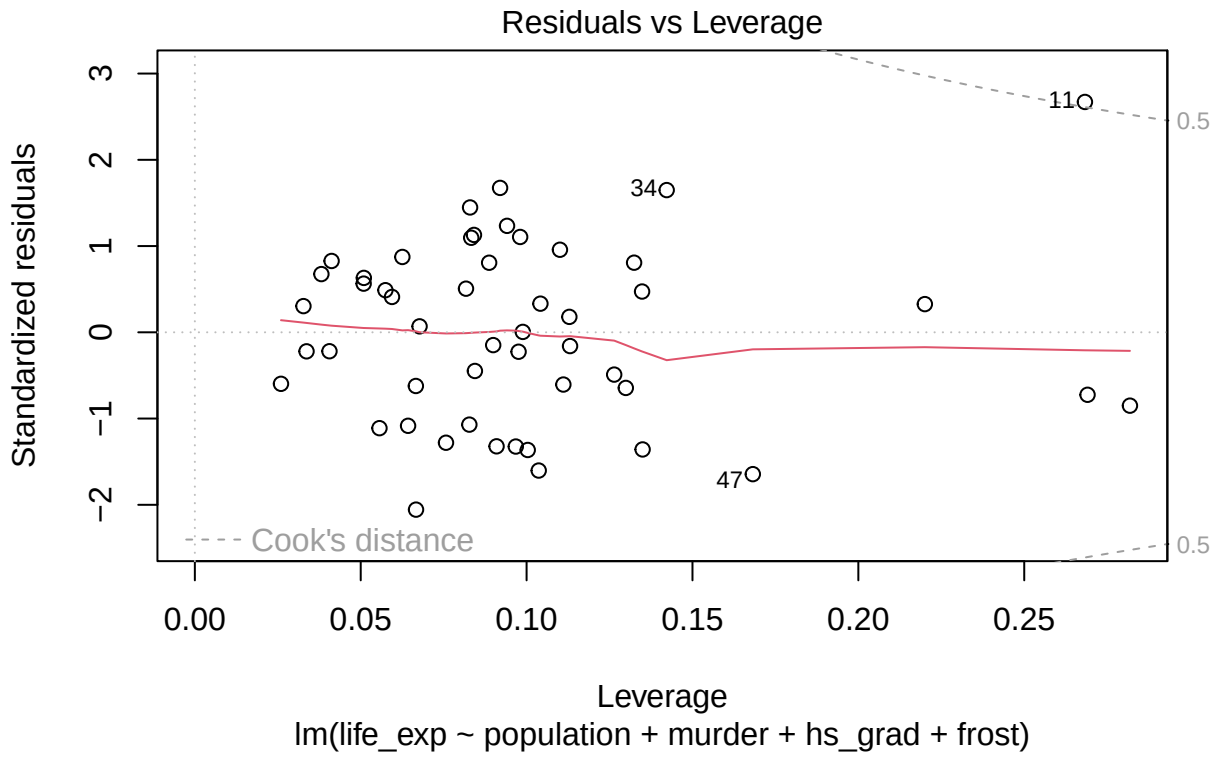



```
plot(both_model)
```

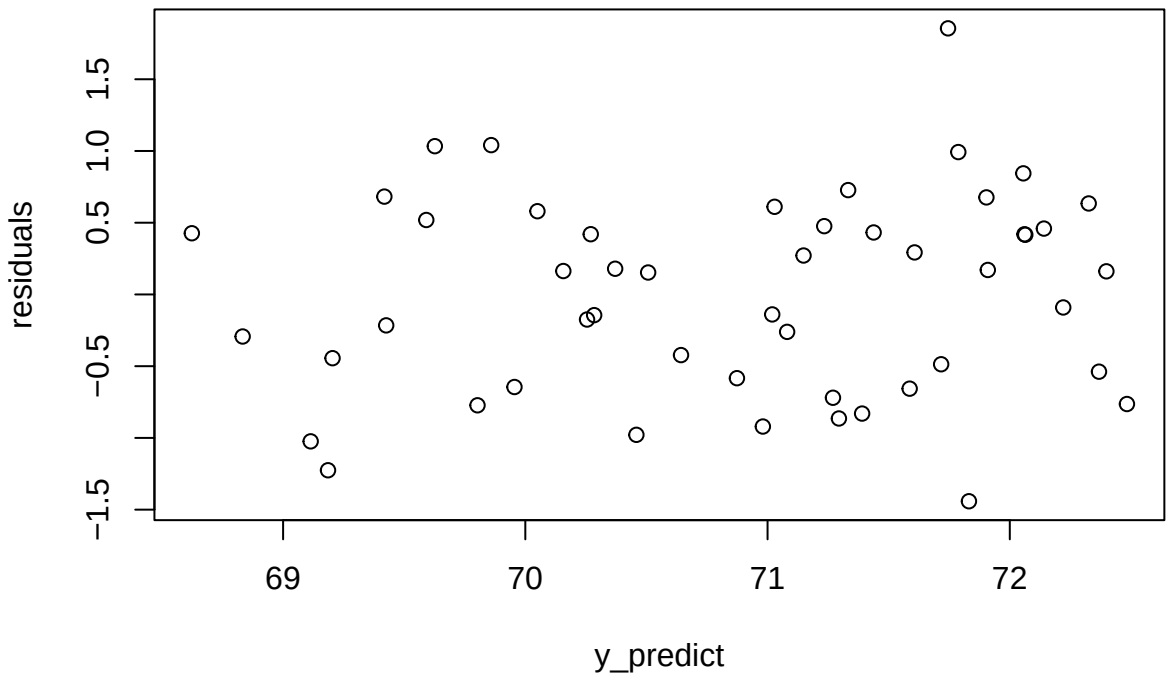

Residuals vs Fitted
Residuals
1.5
0.5
−0.5
−1.5
4
11
19
69
70
71
72
Fitted values
lm(life_exp ~ population + murder + hs_grad + frost)


Q−Q Residuals
Standardized residuals
3
2
1
0
−1
−2
11
4
19
−2
−1
0
1
2
Theoretical Quantiles
lm(life_exp ~ population + murder + hs_grad + frost)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
69
70
71
72
11
19
4
Fitted values
lm(life_exp ~ population + murder + hs_grad + frost)




```
y_predict <- predict(cv_object, as.matrix(state_data[-4]), s = "lambda.min")
residuals <- state_data$life_exp - y_predict
plot(y_predict, residuals)
```




```
train <- trainControl(method = "cv", number = 10)

model_caret = train(life_exp ~ population + murder + hs_grad + frost,
                    data = state_data, trControl = train, method = 'lm',
                    na.action = na.pass)

model_caret$finalModel
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Coefficients:
## (Intercept)   population       murder      hs_grad        frost
##   68.720810     0.246836    -0.290016     0.054550    -0.005174
```

```
print(model_caret)
```

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

```
## Summary of sample sizes: 44, 44, 45, 44, 46, 44, ...
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.7092074  0.7141432  0.6238165
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Each of the models seem to have constant variance, are approximately normal and variance seems to be normal.

e)

Looking at a bunch of different models, we recommend using a model that includes `population` `murder` `hs_grad` and `frost`. This was the one that was recommended to use by regular subsetting, which picks a model based on the adjusted r^2, which will only increase if you add a parameter that is helpful. Additionally, when we did forward, backward, and a combination of both subsetting, we got that the parameters should be 4. We also did lasso, which helps to fit a model when there are multicollinearities in the model. We found in this as well that 4 parameters should be included in the model. The RMSE is 0.79, which means that our model is off by less than 1 year at predicting `life_expectancy`. The RMSE tells us how good the model is at predicting values in a data set, so having a smaller value is best.