Homework 6

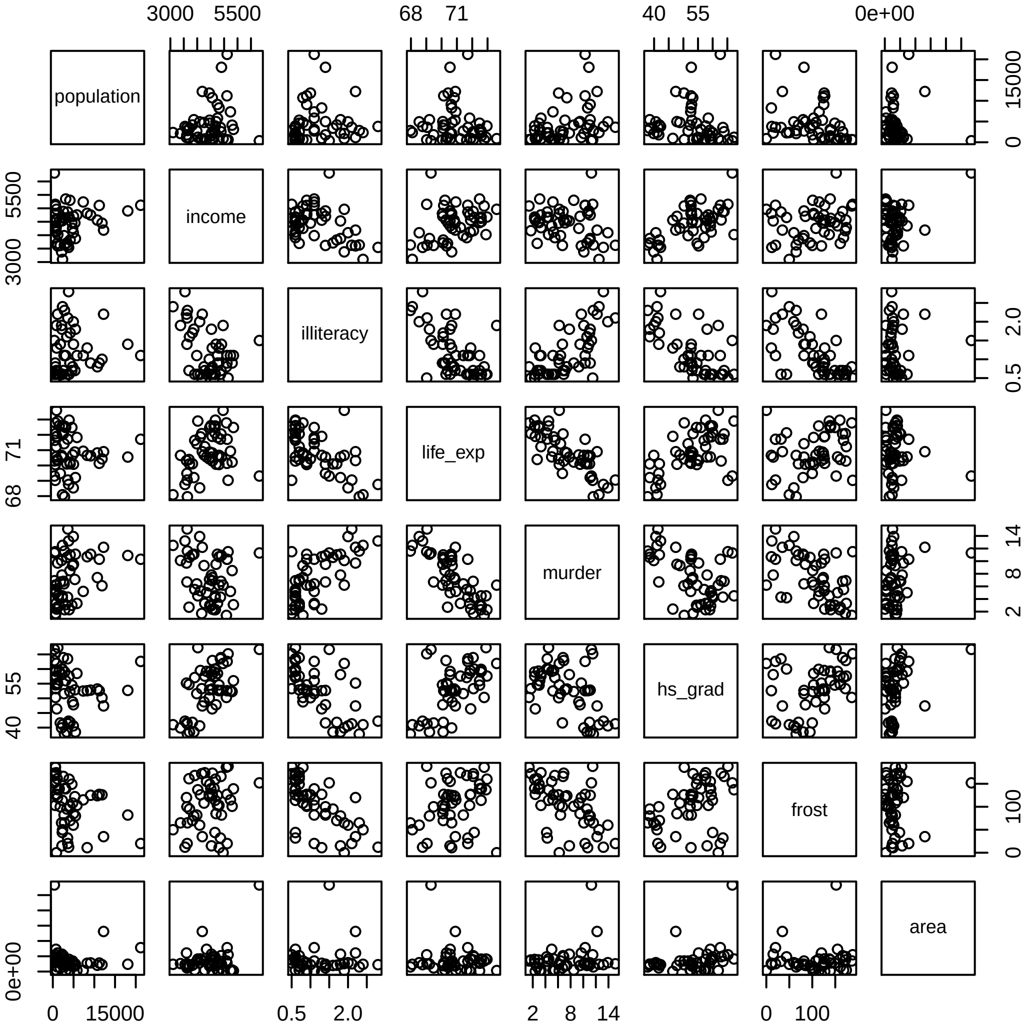
Alyssa Vanderbeek (amv2187)

3 December 2018

1. Explore the dataset and generate appropriate descriptive statistics and relevant graphs for all variables of interest (continuous and categorical) – no test required. Be selective! Even if you create 20 plots, you don't want to show them all.

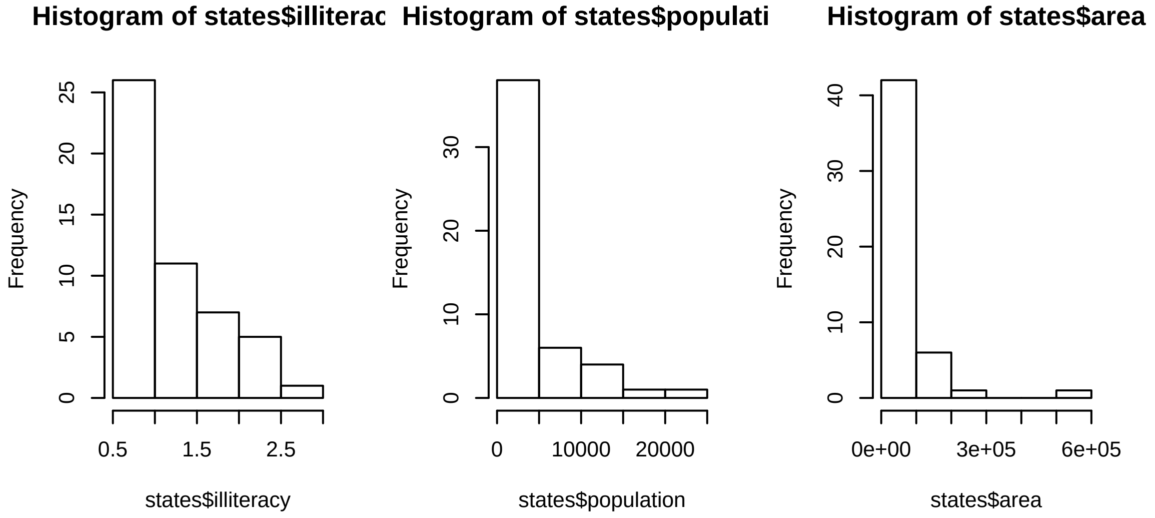
Table 1: Summary statistics

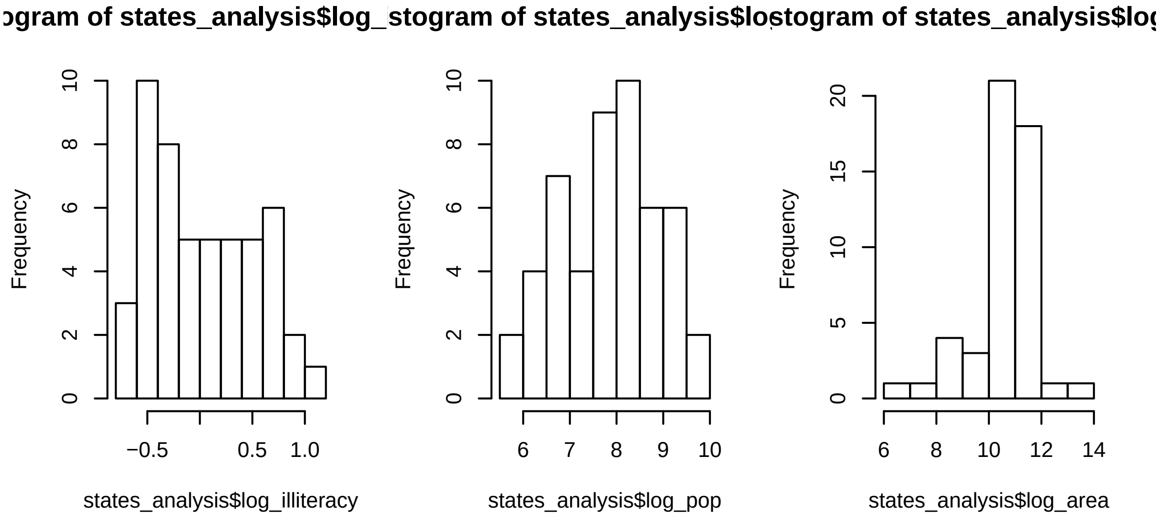
	NA	Mean	Std. Dev.	Min	1st Q	Median	3rd Q	Max
area	0	70735.88	85327.3	1049	36985.25	54277	81162.5	566432
frost	0	104.46	51.98	0	66.25	114.5	139.75	188
hs_grad	0	53.11	8.08	37.8	48.05	53.25	59.15	67.3
illiteracy	0	1.17	0.61	0.5	0.62	0.95	1.58	2.8
income	0	4435.8	614.47	3098	3992.75	4519	4813.5	6315
life_exp	0	70.88	1.34	67.96	70.12	70.67	71.89	73.6
murder	0	7.38	3.69	1.4	4.35	6.85	10.67	15.1
population	0	4246.42	4464.49	365	1079.5	2838.5	4968.5	21198



It looks like murder is correlated both with life expectancy (-0.7808458) and illiteracy (0.7029752), suggesting that it is a potential confounder. Specifically, murder is positively associated with illiteracy (higher murder rate = higher illiteracy rate) and negatively associated with life expectancy (higher murder rate = lower life expectancy).

After examining the distribution of each variable in the dataset, I chose to perform a log transformation on the estimates for area size, illiteracy rate, and population size, which were all skewed.





2. Use automatic procedures to find a 'best subset' of the full model. Present the results and comment on the following

Final model using backwards elimination:

```
##
## Call:
## lm(formula = life_exp ~ murder + hs_grad + frost + log_pop, data = states_analysis)
##
## 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_pop      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
```

Final model using forwards process:

```
##
## Call:
## lm(formula = life_exp ~ murder + hs_grad + log_pop + frost, data = states_analysis)
##
## 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_pop      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
```

Final model using a stepwise process:

```
## Start:  AIC=-23.71
## life_exp ~ income + murder + hs_grad + frost + log_area + log_illiteracy +
##     log_pop
##
##                   Df Sum of Sq    RSS     AIC
## - income          1    0.0002 22.596 -25.712
## - log_illiteracy  1    0.1079 22.704 -25.475
## - log_area        1    0.2368 22.833 -25.192
## <none>                        22.596 -23.713
## - frost           1    1.1645 23.760 -23.200
## - log_pop         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 + log_area + log_illiteracy +
##     log_pop
##
##                   Df Sum of Sq    RSS      AIC
## - log_illiteracy  1    0.1095 22.705 -27.4708
## - log_area        1    0.2616 22.858 -27.1370
## <none>                        22.596 -25.7125
## - frost           1    1.2628 23.859 -24.9936
## - log_pop         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 + log_area + log_pop
##
##            Df Sum of Sq    RSS     AIC
## - log_area  1    0.2157 22.921 -28.998
## <none>                  22.705 -27.471
## - log_pop   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 + log_pop
##
##           Df Sum of Sq    RSS     AIC
## <none>                 22.921 -28.998
## - frost    1     2.214 25.135 -26.387
## - log_pop  1     2.450 25.372 -25.920
## - hs_grad  1     6.959 29.881 -17.741
## - murder   1    34.109 57.031  14.578
```

All automatic processes conclude the same model, using percent increase in population size (log(population)), rate of high school graduation (hs_grad), murder rate per 100,000 (murder), and average number of days annually with temperatures below freezing (frost) as predictors of life expectancy.

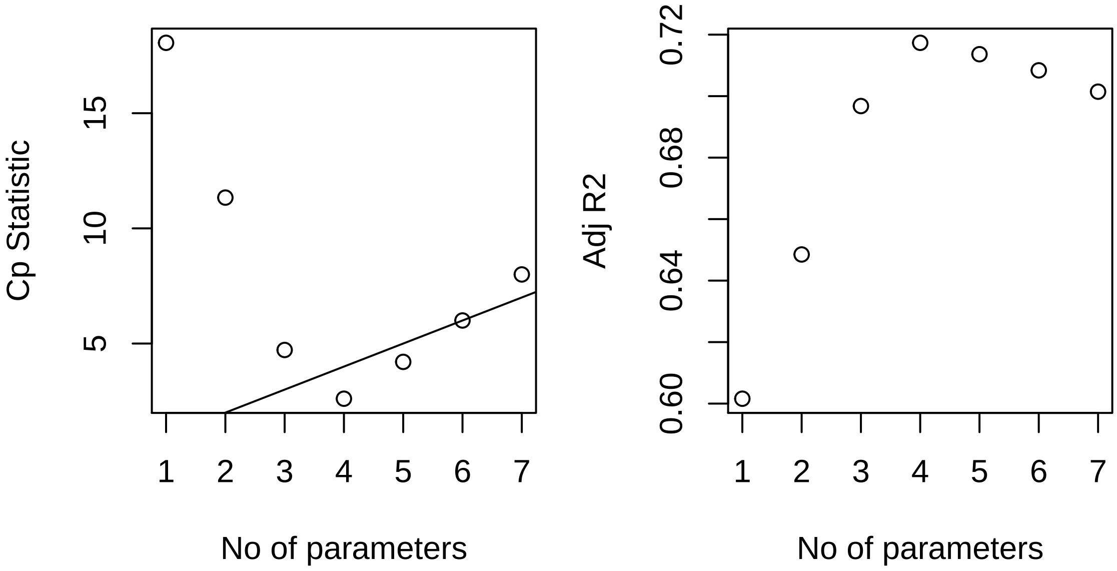
'Frost' is the least significant predictor, with a p-value of 0.043. However, no variables were seen to be a "close call" at the 5% significance level.

There is an observed correlation between illiteracy (with and without log transformation) and high school graduation rate (-0.6571886), but my model includes only high school graduation rate as a predictor.

3. Use criterion-based procedures studied in class to guide your selection of the 'best subset'. Summarize your results (tabular or graphical)

Table 2: Criterion-based model building

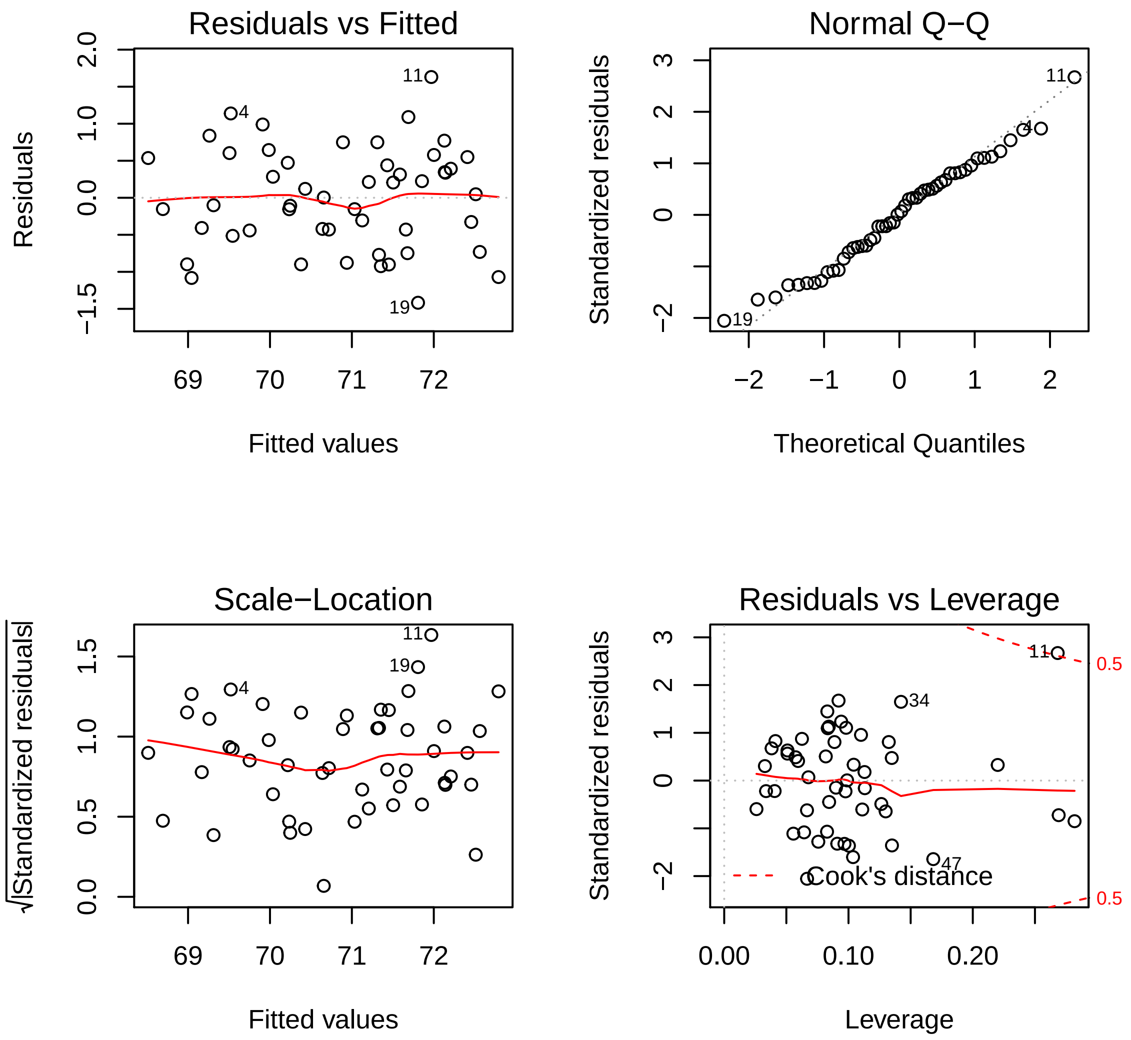
p	(Intercept)	income	murder	hs_grad	frost	log_area	log_illiteracy	log_pop	rss	rsq	adjr2	cp	bic
1	1	0	1	0	0	0	0	0	34.46133	0.6097201	0.6015893	18.054999	-39.22051
2	1	0	1	1	0	0	0	0	29.77036	0.6628461	0.6484991	11.335656	-42.62472
3	1	0	1	1	0	0	0	1	25.13538	0.7153378	0.6967729	4.720403	-47.17452
4	1	0	1	1	1	0	0	1	22.92123	0.7404135	0.7173392	2.604837	-47.87315
5	1	0	1	1	1	1	0	1	22.70549	0.7428568	0.7136360	4.203829	-44.43397
6	1	0	1	1	1	1	1	1	22.59600	0.7440968	0.7083894	6.000318	-40.76364
7	1	1	1	1	1	1	1	1	22.59583	0.7440987	0.7014485	8.000000	-36.85199



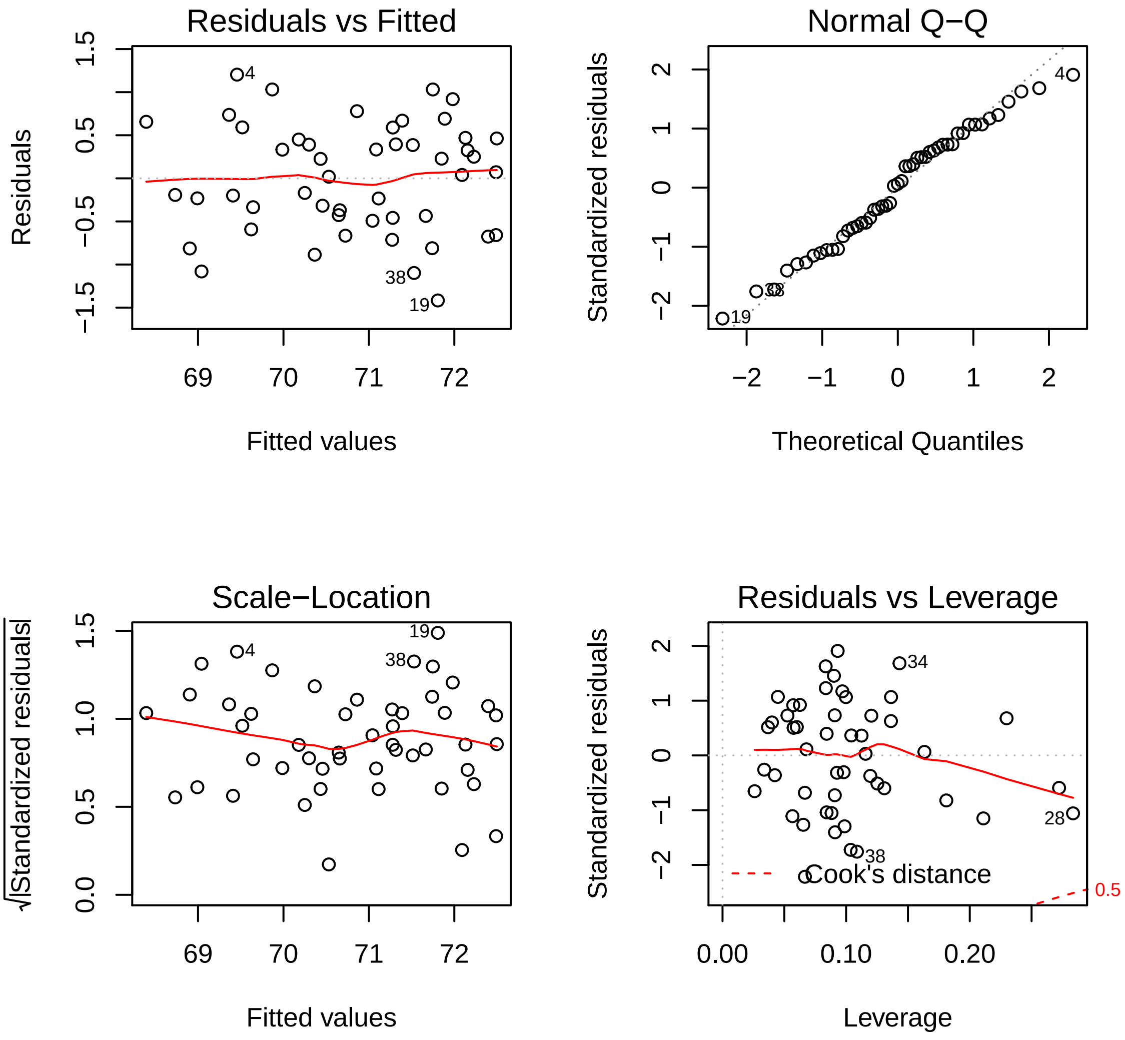
According to the Cp statistics and Adjusted R^2, the ideal number of parameters is 4; as seen in the table above, those parameters are murder, hs_grad, frost, and log_pop - the same as what was concluded in the automatic process.

4. Compare the two ‘subsets’ from parts 2 and 3 and recommend a ‘final’ model. Using this ‘final’ model do the following. a) Identify any leverage and/or influential points and take appropriate measures. b) Check the model assumptions.

All analyses above recommend the same model using percent increase in population size (log(population)), rate of high school graduation (hs_grad), murder rate per 100,000 (murder), and average number of days annually with temperatures below freezing (frost) as predictors of life expectancy.



```
##
## Call:
## lm(formula = life_exp ~ murder + hs_grad + log_pop + frost, data = states_analysis[-11,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.41708 -0.45880  0.03924  0.46286  1.20332
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 67.906960   1.344438  50.510  < 2e-16 ***
## murder      -0.276679   0.033203  -8.333 1.35e-10 ***
## hs_grad      0.046799   0.013953   3.354  0.00165 **
## log_pop      0.337449   0.109043   3.095  0.00342 **
## frost       -0.001632   0.002610  -0.625  0.53499
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6621 on 44 degrees of freedom
## Multiple R-squared:  0.7611, Adjusted R-squared:  0.7394
## F-statistic: 35.05 on 4 and 44 DF,  p-value: 3.709e-13
```



The 11th entry showed evidence of being an influential outlier (according to Cook's distance and measure of influence). To check to see whether this entry had a significant impact on the model and its assumptions, I compared diagnostics of the model with and without the 11th point. The diagnostic plots above show that, in fact, the model assumptions (1. residuals have mean zero, 2. residuals have equal variance, 3. residuals are independent) are met for both models - with and without the potential influential point. However, we can see that without the point, the 'frost' variable is no longer a significant predictor of life expectancy, with a p-value of 0.53.

Using the 'final' model chosen in part 4, focus on MSE to test the model predictive ability a) Use a 10-fold cross-validation

```
##          RMSE  Rsquared       MAE    RMSESD RsquaredSD     MAESD
## 1   0.7453512 0.7975389 0.6311168 0.2527274 0.16559687 0.2010312
## 2   0.7220211 0.7660915 0.6123710 0.1876809 0.12191794 0.1324566
## 3   0.7420167 0.7644641 0.6400099 0.2022501 0.19846543 0.1861321
## 4   0.7523093 0.7422363 0.6563839 0.3016137 0.20706197 0.2598988
## 5   0.7534132 0.7574106 0.6561317 0.1840462 0.22030745 0.1733027
## 6   0.7427202 0.7495104 0.6443664 0.2136023 0.13783553 0.1872413
## 7   0.7396154 0.8120813 0.6402588 0.2266164 0.09432498 0.1934086
## 8   0.7551494 0.7264384 0.6362972 0.1993740 0.12746459 0.1800144
## 9  0.7270407 0.7647401 0.6238596 0.2332786 0.15516225 0.1873079
## 10 0.7419066 0.8007227 0.6261783 0.1771696 0.11810094 0.1467010
```

(b) Experiment a new, but simple bootstrap technique called "residual sampling". Summarize the MSE.

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot::boot(data = states_analysis, statistic = boot.res, R = 10)
##
##
## Bootstrap Statistics :
##      original        bias    std. error
## t1* 0.5275224 -0.008270433  0.06540224
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot::boot(data = states_analysis, statistic = boot.res, R = 1000)
##
##
## Bootstrap Statistics :
##      original       bias    std. error
## t1* 0.5924127 -0.09234814   0.0436284
```

The MSE of the bootstrap method is notably smaller than that of the 10-fold cross validation. This is intuitive on a smaller dataset such as the one used here (n = 50); a 10-fold cross validation using 50 data points trains the model on subsets of ~45, whereas the bootstrapping method uses sampled sets of size 50. It may be that at this sample size, this size difference of ~5 results in the bootstrap method having less MSE. This can be verified if we perform a LOOCV process:

```
##         mse
## 1 0.6431721
```

Increasing the size of the training subsets in the k-fold validation improves the MSE, though the bootstrap still performs better. Because of the senstivity of these two methods of CV, I would recommend running both to assess model performance. Both suggest that the model we selected above is subject to a fair bit of error, so we would want to revisit the model-building process. Ideally, we want a model whose performance does not vary notably by CV method AND has a small MSE. However, I think that the size of the dataset use here may unavoidably result in sensitive CV performance. A improved model will be one that reduces the MSE across both methods of CV used here.