HUDM5123_Lab02_Diagnostics

Zhuqian Karen Zhou

February 13, 2020

Task 1: Fit the model

The linear model looks like this: HS.Grad[i] = b0 + b1·Illiteracy[i] + b2·Income[i] + b3·Area[i] + error[i]

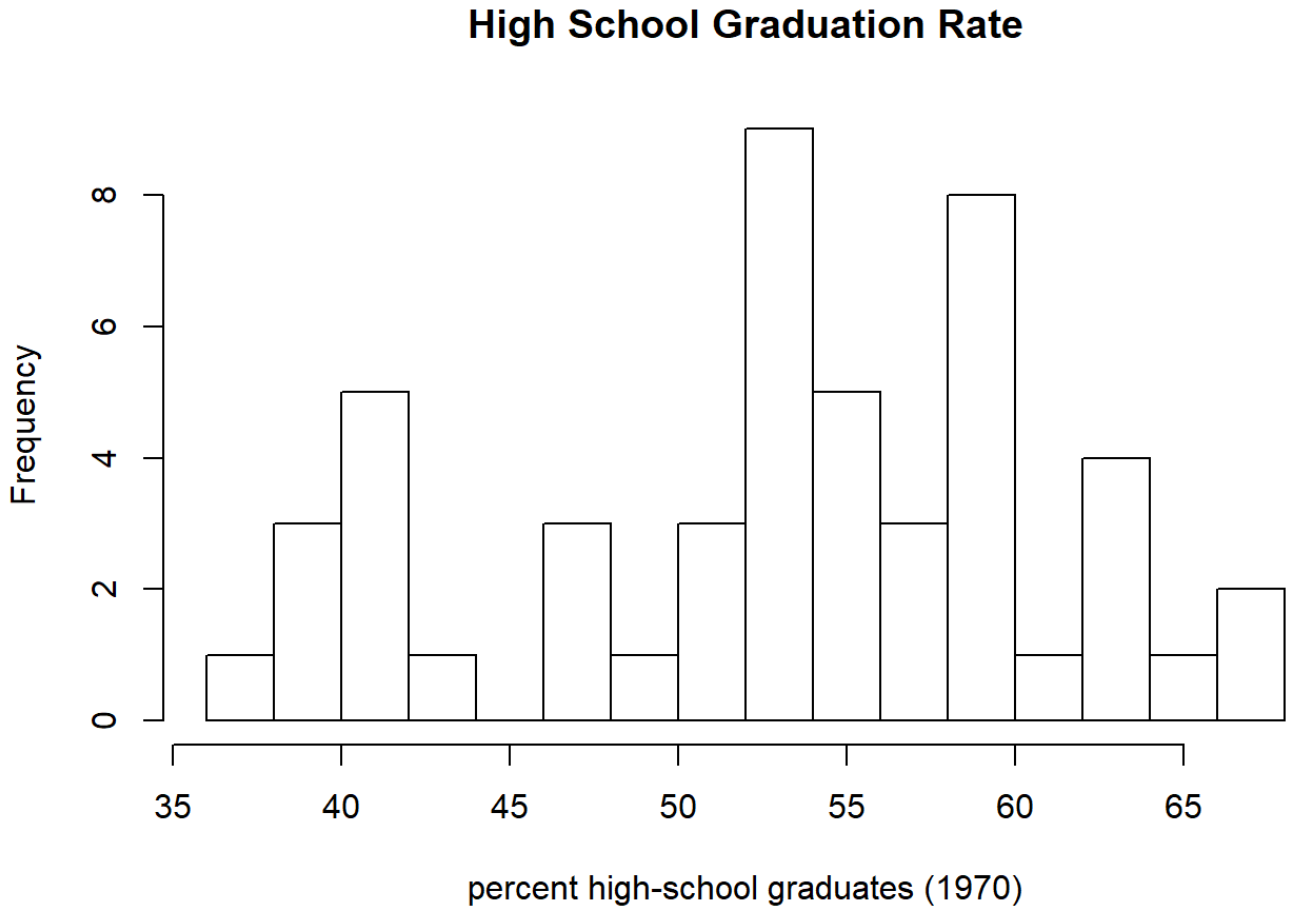
```
##
## Call:
## lm(formula = HS.Grad ~ Illiteracy + Income + Area, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9928 -3.6970 -0.4094  3.1467 13.9656
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.385e+01  7.185e+00   6.102 2.03e-07 ***
## Illiteracy  -7.396e+00  1.377e+00  -5.370 2.52e-06 ***
## Income       3.621e-03  1.462e-03   2.477  0.01700 *
## Area         2.618e-05  9.499e-06   2.756  0.00836 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.072 on 46 degrees of freedom
## Multiple R-squared:  0.6298, Adjusted R-squared:  0.6057
## F-statistic: 26.09 on 3 and 46 DF,  p-value: 5.224e-10
```

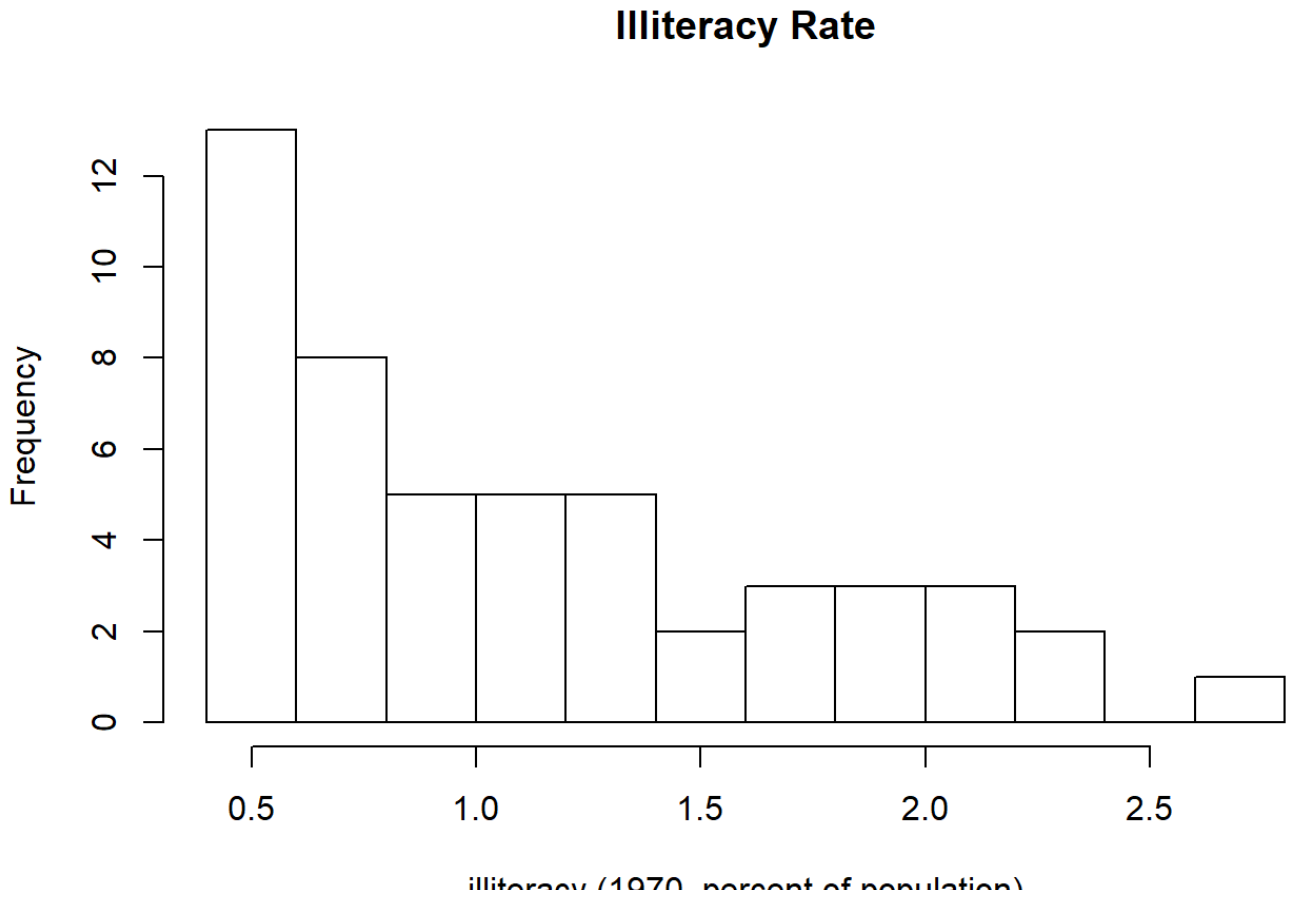
The prediction equation with estimated coefficients is as follows: HS.Grad[i] = 43.846 + (-7.396)·Illiteracy[i] + 0.004·Income[i] + 0.000026·Area[i] + error[i]

The R-squared of the estimated model is 0.6298, which means that the three predictors, illiteracy rate, income, and state area, accounts for around 62.98% of variance in high school graduation rate.

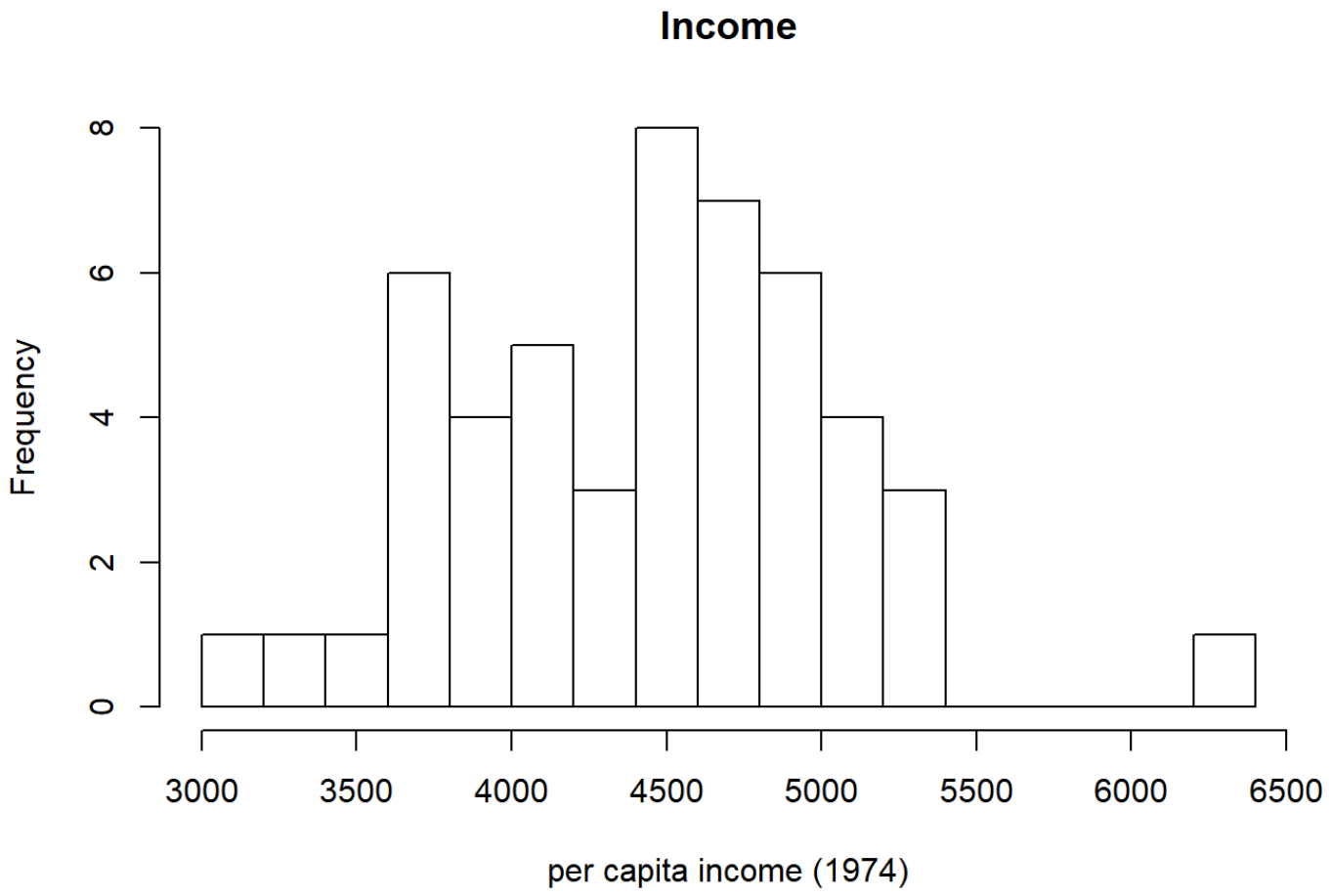
The residual standard error, 5.972, shows the average amount that high school graduation rate in the dataset deviates from the regression line. Its degree of freedom, 46, means there are 46 pieces of independent information among the total 50 data points (i.e. states) going into the estimation of the four parameters, i.e. b0, b1, b2, and b3, in the current model.

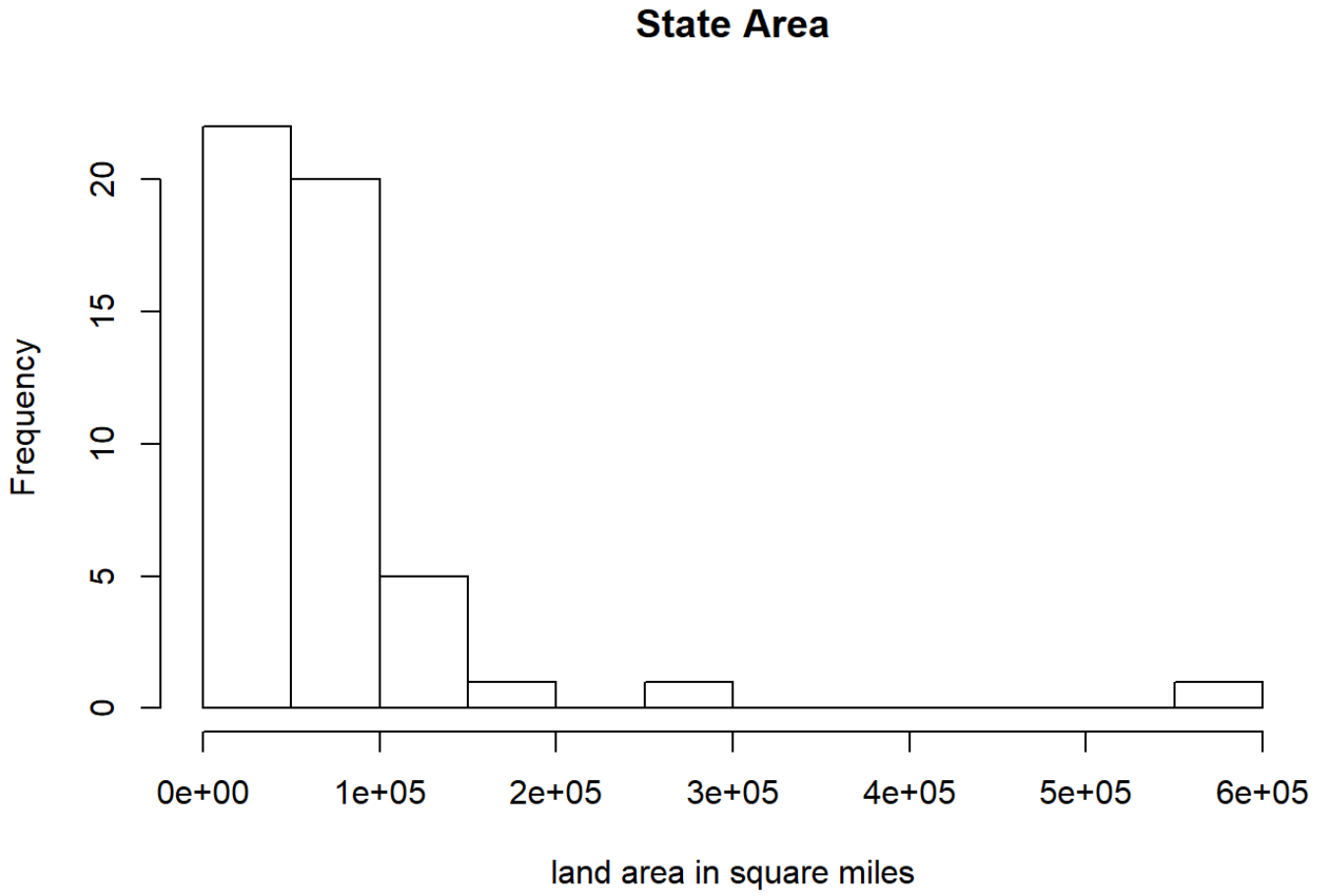
Task 2: Draw plots


High School Graduation Rate
Frequency
0
2
4
6
8
35
40
45
50
55
60
65
percent high-school graduates (1970)

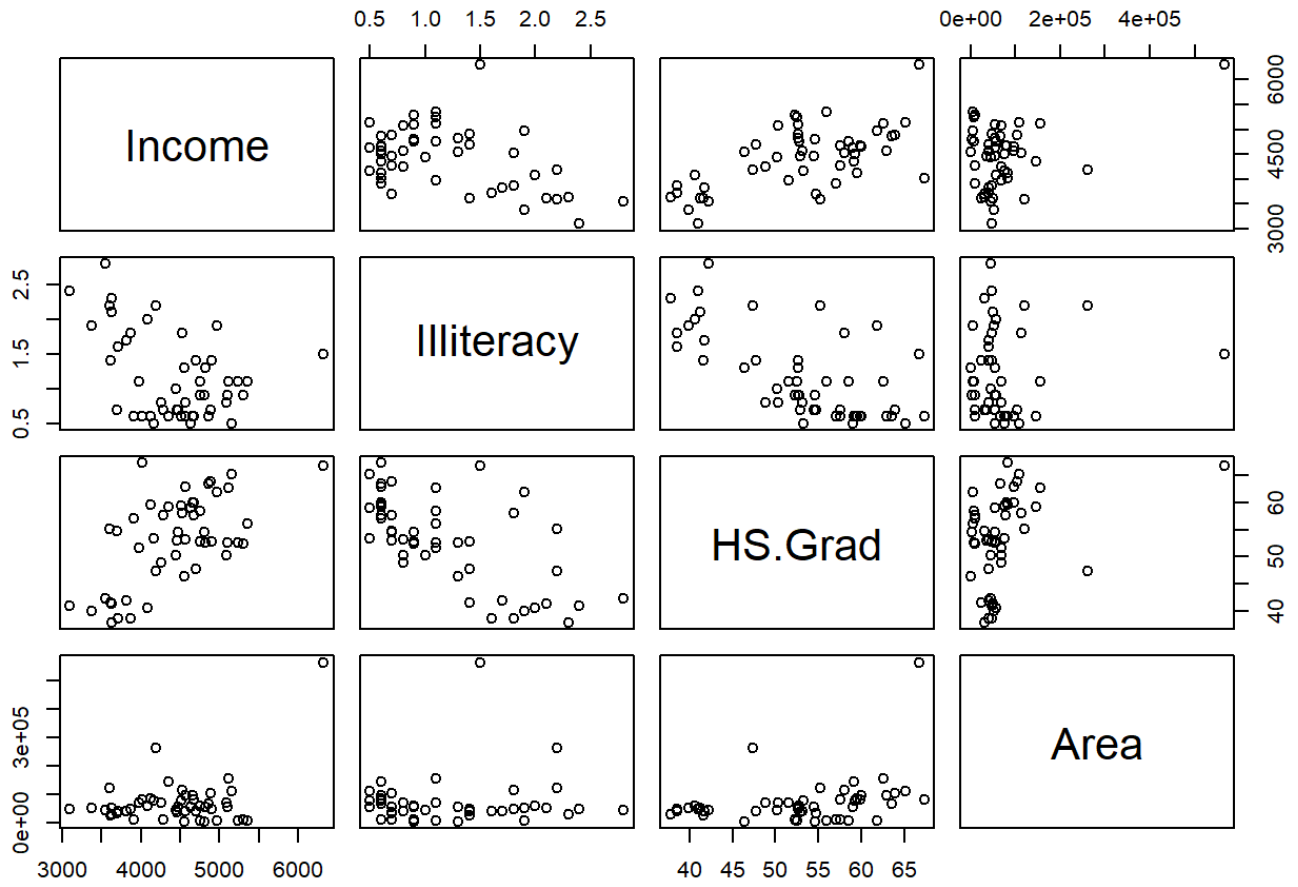

Illiteracy Rate
Frequency
0
2
4
6
8
10
12
0.5
1.0
1.5
2.0
2.5

illiteracy (1970, percent of population)


Income
Frequency
0
2
4
6
8
3000
3500
4000
4500
5000
5500
6000
6500
per capita income (1974)



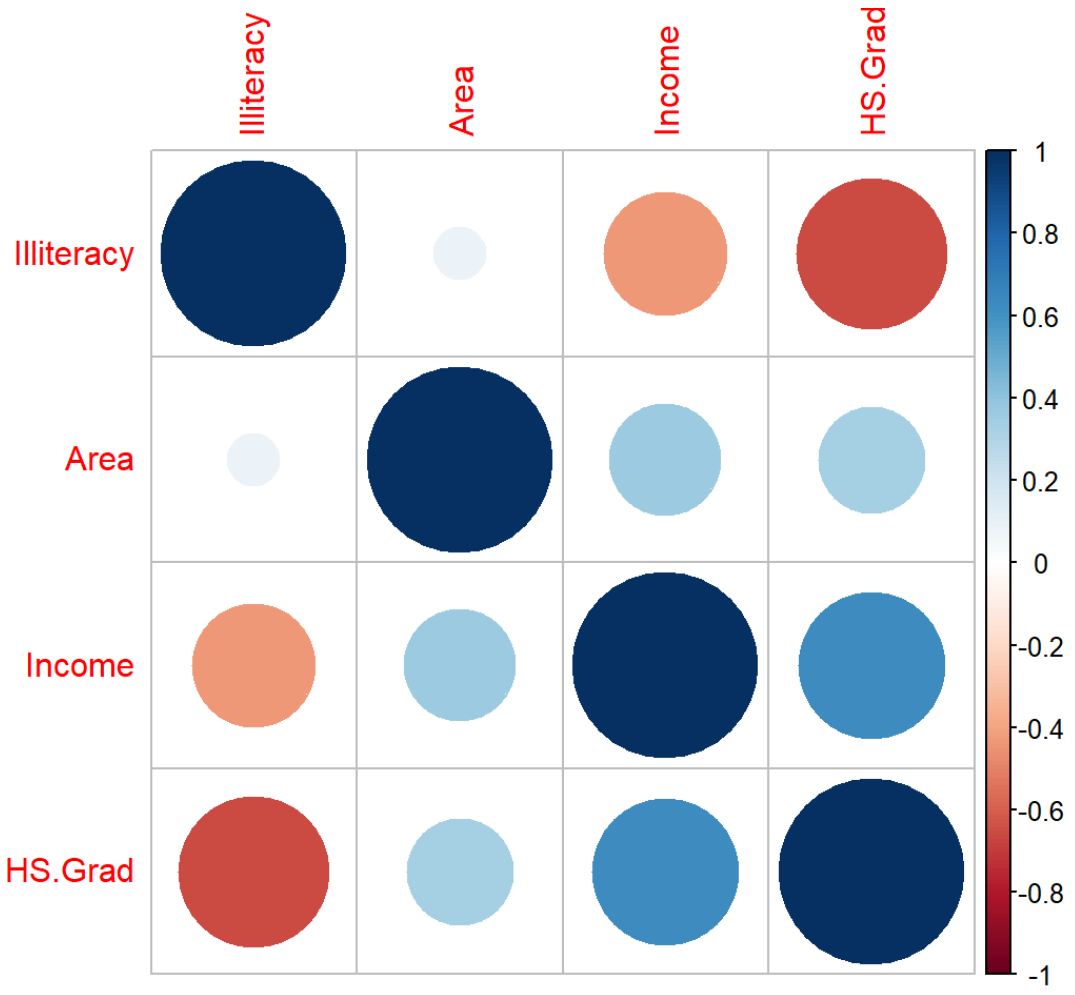
The histograms show us that the distrubution of the four variables. The central tendencies of high school graduation rate and of income per capita are high since their modes are around medians while illiteracy rate and state area are less so because their modes are aroud their minimal values.



The bivariate scatterplot shows us a roughly *positive* linear relationship between high school graduation rate and income percapita, a roughly *negative* linear relationship between high school graduation rate, and a almost random relationship between high school graduation rate and state area.

```
##            Income Illiteracy HS.Grad Area
## Income       1.00      -0.44    0.62 0.36
## Illiteracy  -0.44       1.00   -0.66 0.08
## HS.Grad      0.62      -0.66    1.00 0.33
## Area         0.36       0.08    0.33 1.00
```

```
## corrplot 0.84 loaded
```



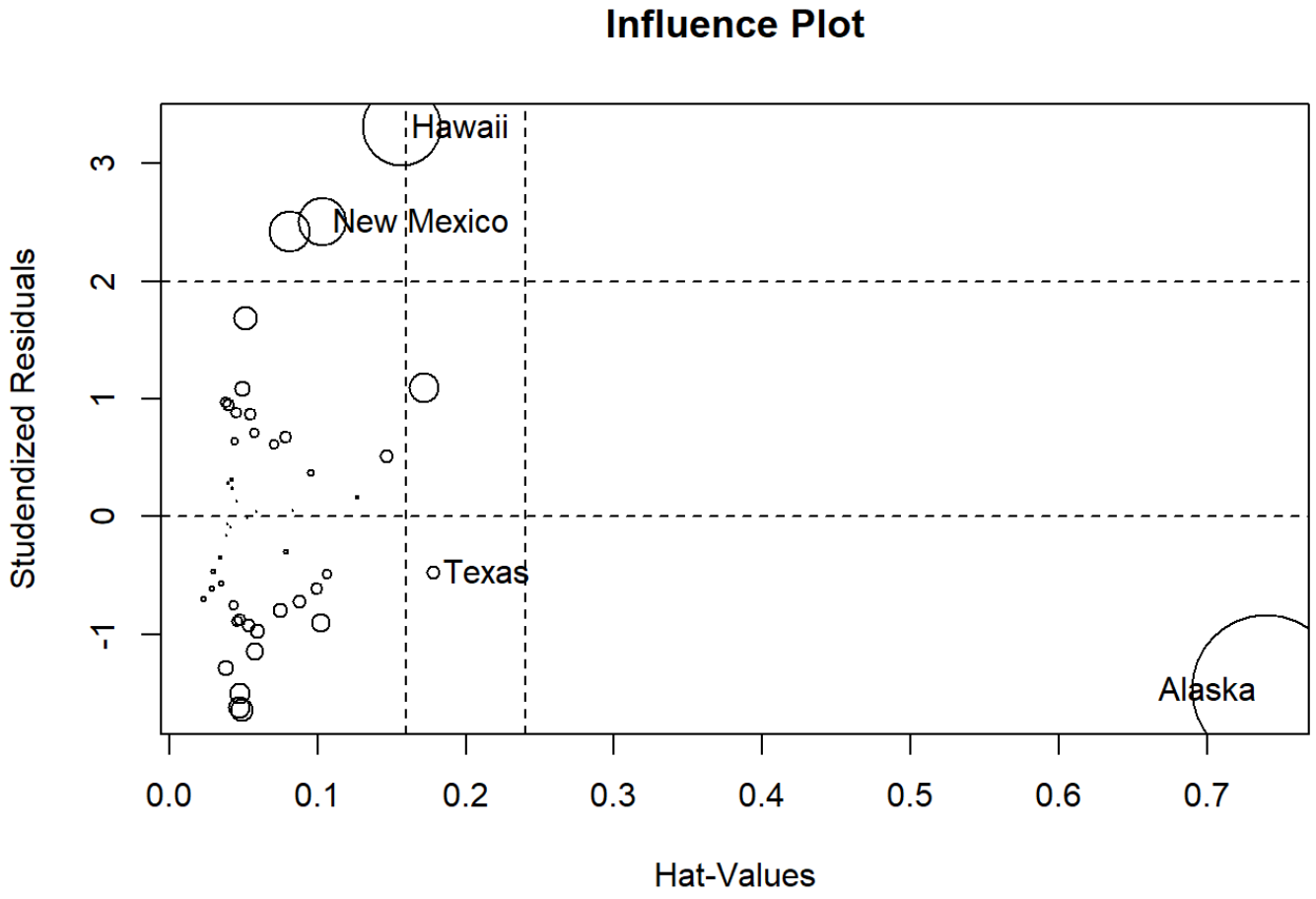
The correlation matrix and the correlation plot help us quantify the linear relationships shown in the scatterplot. The postive relationship between high school graduation rate and income per capita and the negative relationship between high school graduation rate and illiteracy rate are highly linear with both correlation coefficients greater than 0.6 while the relationship between high school graduation rate and state area is less linear with the correlation coefficient around 0,33.

Task 3: Draw a influence plot

```
## Warning: package 'car' was built under R version 3.5.3
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 3.5.3
```

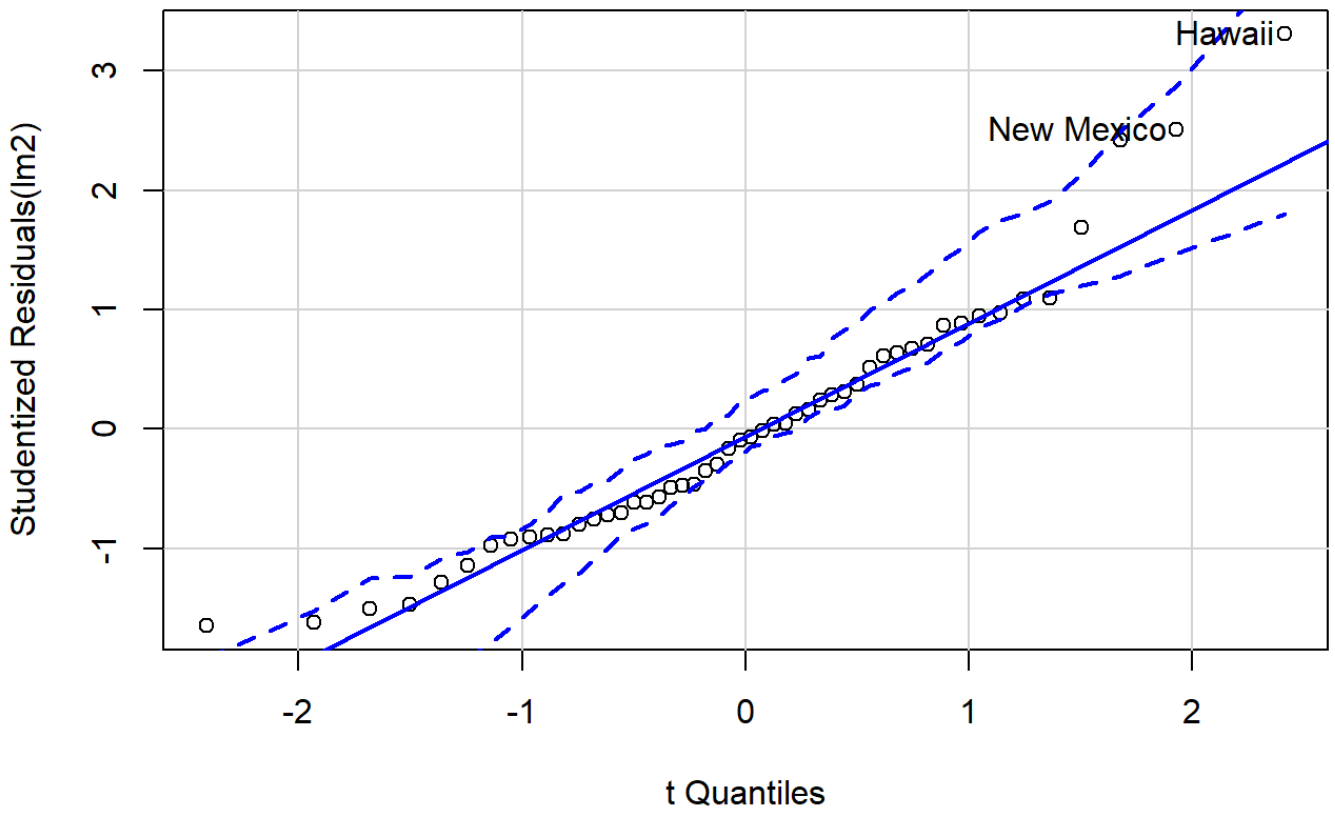



```
##                StudRes       Hat      CookD
## Alaska     -1.4673106 0.7397358 1.49243280
## Hawaii      3.3065656 0.1567540 0.41787414
## New Mexico  2.5073603 0.1030226 0.16191097
## Texas      -0.4751608 0.1780181 0.01243354
```

Leverage is messuared by hat-values (i.e. x-axis on the plot). Discrepancy is mesared by studentdized residuals (i.e. y-axis on the plot). And the influence of a data point is visualized as the size of the circle in the plot. Therefore, we can tell that State Alaska has the highest leverage and is most influential while Sate Hawaii has the most discrepancy.

Although we need to be cautious everytime we throw out an outlier, I do think the most influential point, State Alaska, can be excluded in this case since it is the largest state of the US but with the third fewest polution and is far from the US main land. Its geographical and historical specialities make it very likely to be an outlier.

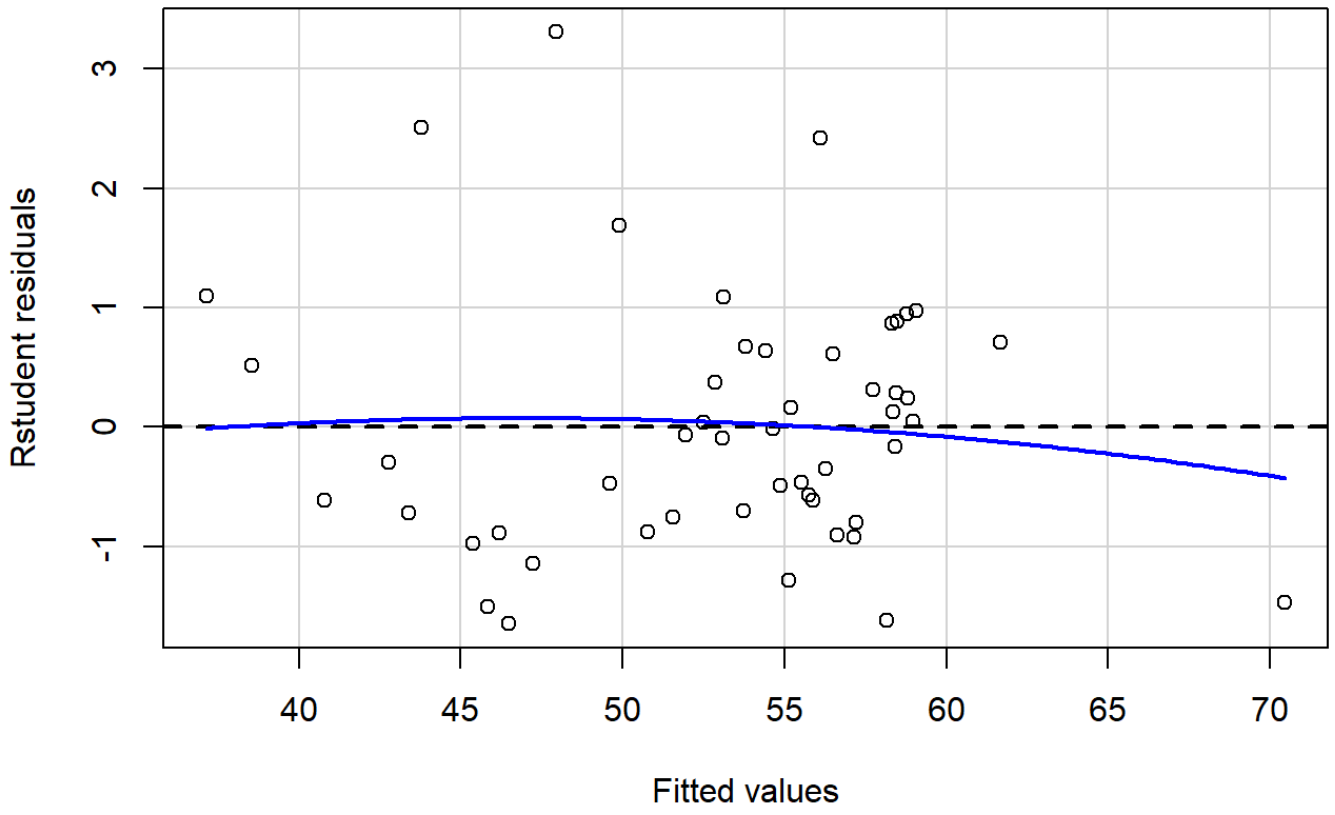
Task 4: Draw a QQ plot for checking the normality of error term




```
##     Hawaii New Mexico
##         11         31
```

The plot shows evidence of normality since almost all points fall within the 95% confidence bands, which means our sample does not differ much from the theoretical expectation that the error term of the model is normally distributed.

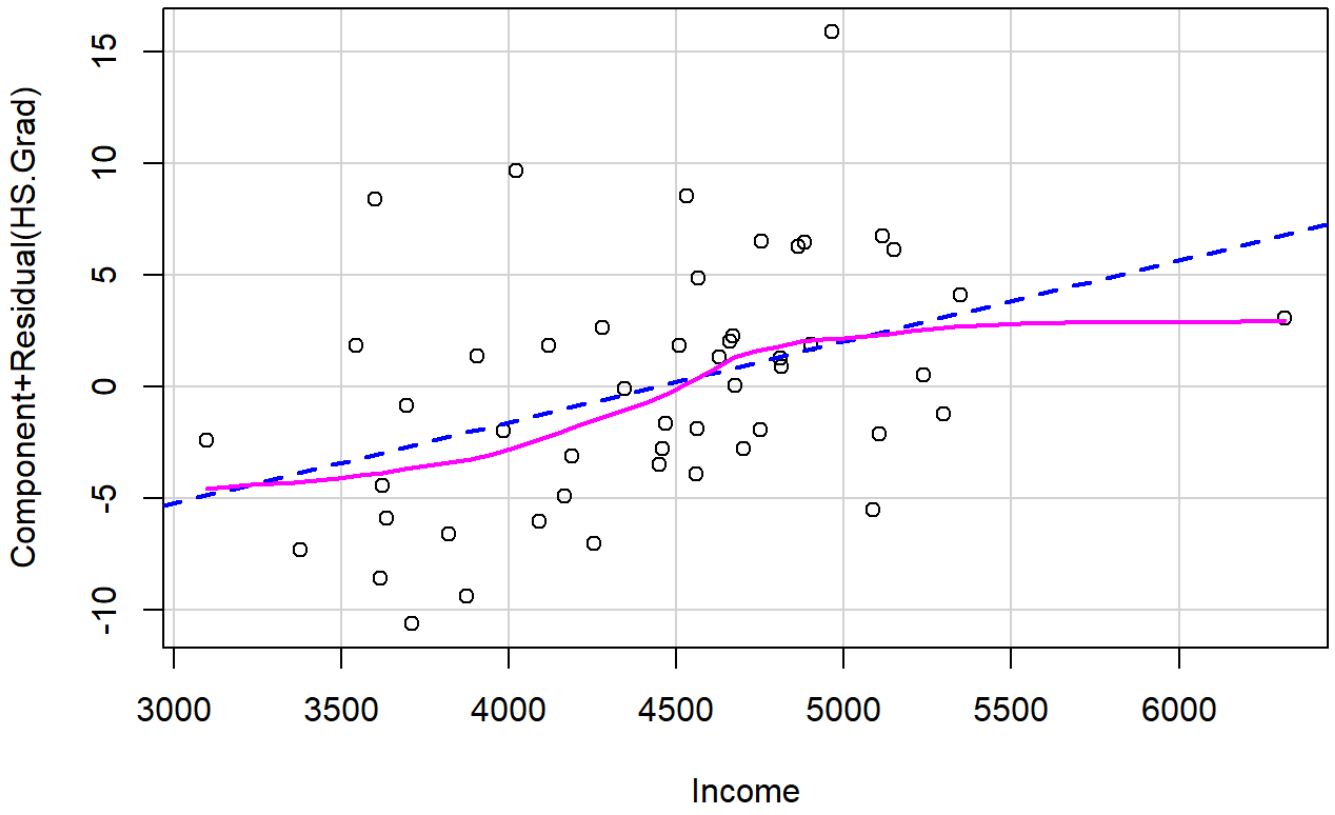
Task 5: Draw the residual plot for checking constant error variance

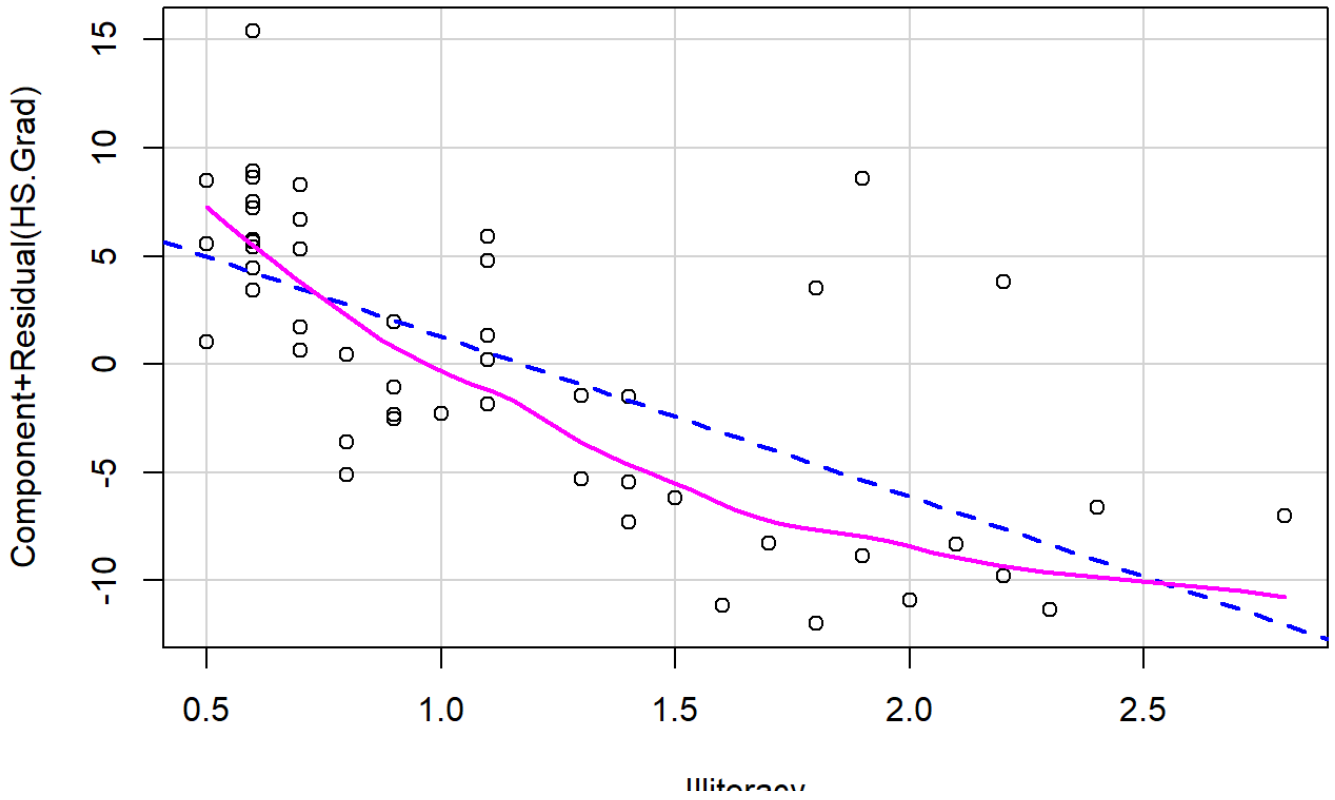


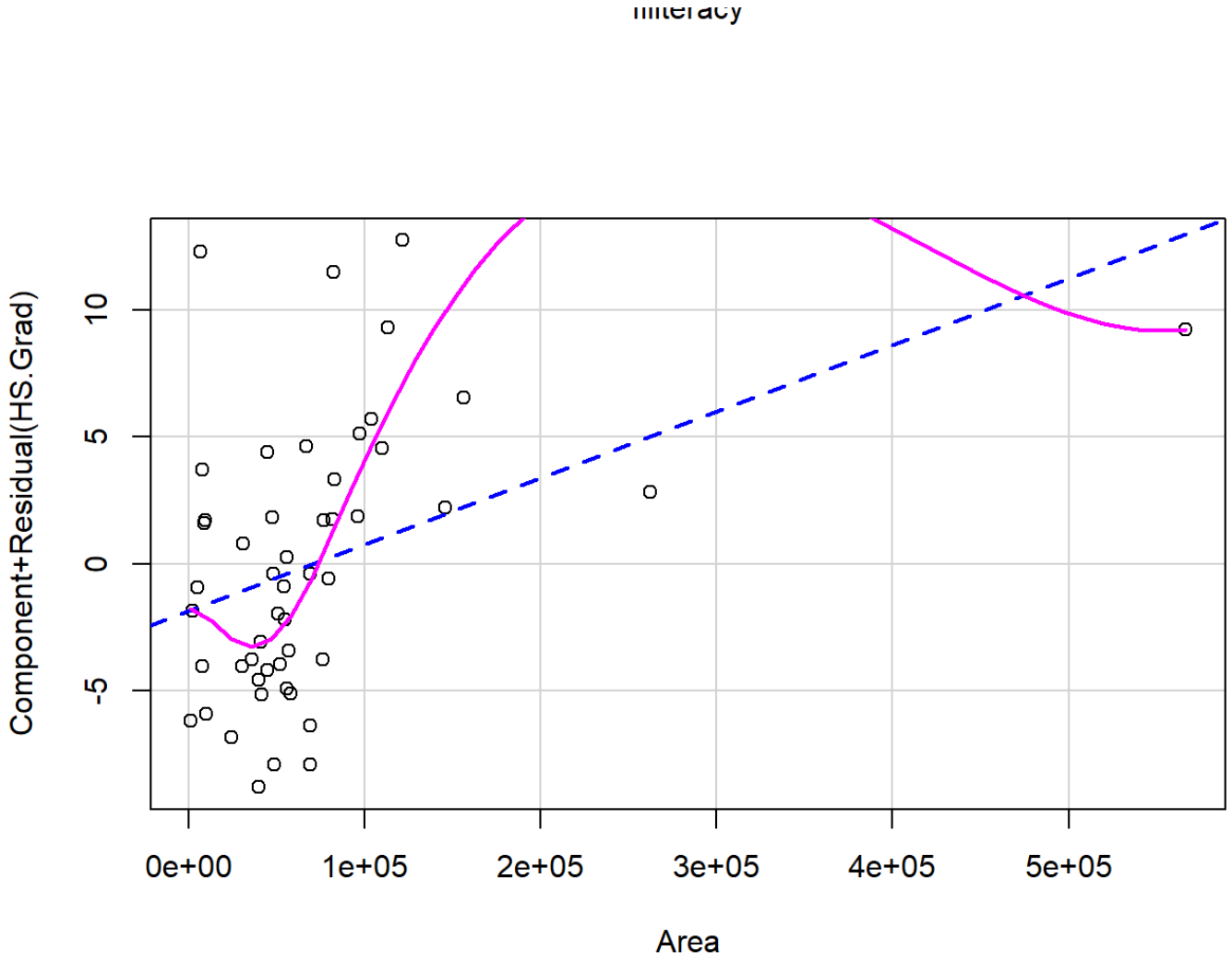
If the error variance is not constant, we will expect to see the distrubution of the data point with a trumpt-like shape, i.e. they center around the line of zero studendentized residual when the predicted values are small but tend to spread around the line to a greater extent when the predicted values go up.

Since we do not see such a shape on the residual plot we just drew, there is no evidence for non-constant error variance.

Task 6: Draw CR-plots for checking linearity


Component+Residual(HS.Grad)
15
10
5
0
-5
-10
3000
3500
4000
4500
5000
5500
6000
Income


Component+Residual(HS.Grad)
15
10
5
0
-5
-10
0.5
1.0
1.5
2.0
2.5



According to John Fox, the component-plus residual (CR) plot, also called the partial-residual plot, depict the relationship between the partial residual (i.e. E[i,j] + E[i]+B[j]·x[i,j]) and the predictor (i.e. X[i,j]). Comparing the lowess smooth line (i.e the solid line) and the line for the linear least-squares fit (i.e. the dash line), we can not only check the linearity between variables but also distinguish between monotone and non-monotone nonlinearity when nonlinearity is present.

Looking at the three CR-plots above, it is obvious that all three predicors, income, illiteracy, and area, have somehow nonlinear relationships with the prediction, high school graduation rate. The difference among them is that the nonlinearity of income and illiteracy are monotone while the nonlinearity of state area is non-monotone, which mainly results from the suspected outliner, Alaska.

Task 7: Calculate VIF for detecting multicollinearity

```
## Illiteracy     Income       Area
##   1.342699   1.537651   1.251376
```

As the variance inflation factor shows, all three predictors in the model have VIF less than 4, which means multicolinearity is not a big concern in the current model. Among them, the variable, income, has the largest VIF. Let's take it as an example to demonstrate how VIF is calculated.

First, fit the linear model by modeling "Income" on other two predictors, "Illiteracy" and "Area".

```
##
## Call:
## lm(formula = Income ~ Illiteracy + Area, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -848.97 -342.37  -74.39  395.14 1068.68
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.784e+03  1.637e+02  29.222  < 2e-16 ***
## Illiteracy  -4.717e+02  1.189e+02  -3.966 0.000248 ***
## Area         2.877e-03  8.496e-04   3.386 0.001442 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 506 on 47 degrees of freedom
## Multiple R-squared:  0.3497, Adjusted R-squared:  0.322
## F-statistic: 12.63 on 2 and 47 DF,  p-value: 4.063e-05
```

After getting the multiple r-squared of this model, 0.3497, calculte VIF using the forumla, VIF=1/(1-multiple r-squred).

```
## [1] 1.537752
```

The result of such calculation, 1.537752, is almost the same as the one computed by the VIF function, 1.537651.

Task 8: Fit a model with a categorical predictor

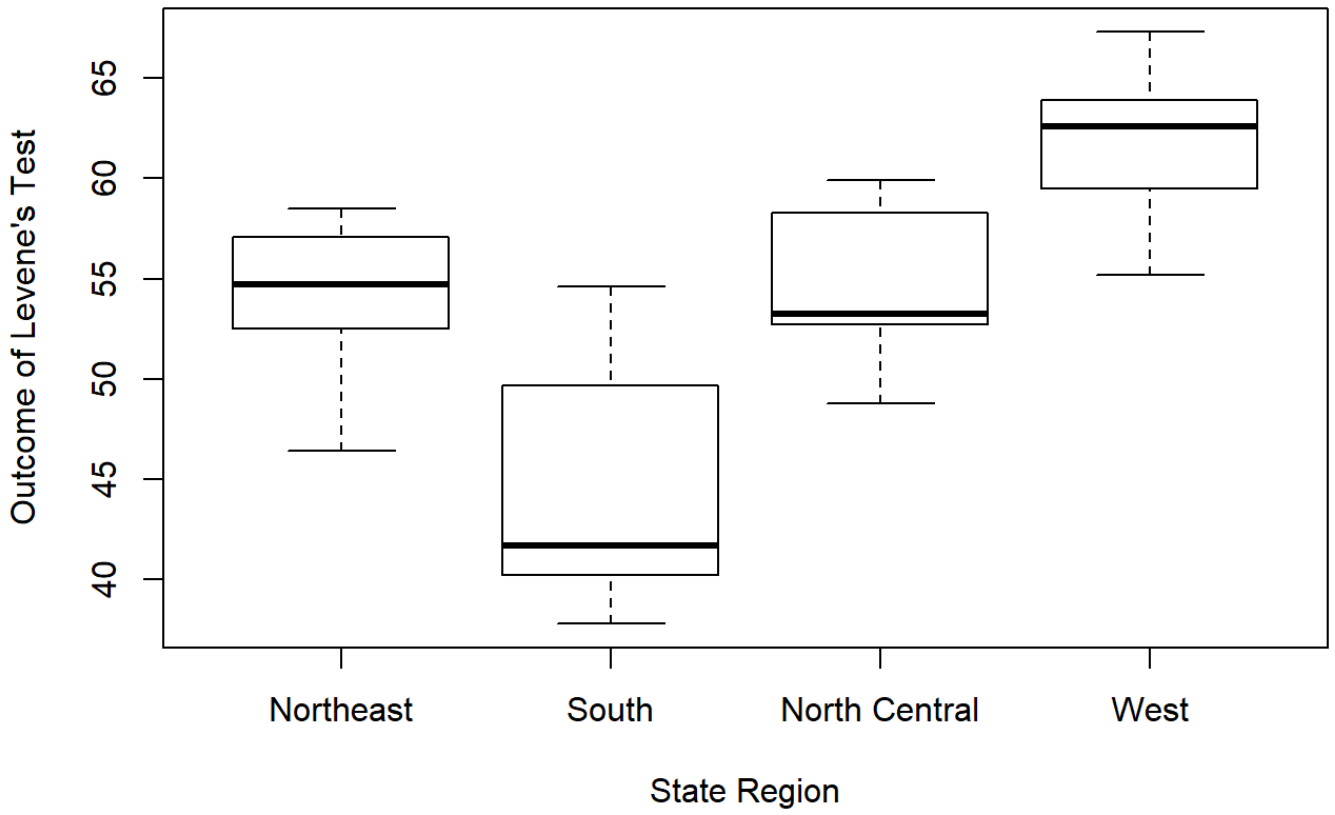
```
##
## Call:
## lm(formula = HS.Grad ~ state.region, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.567 -2.983 -1.242  3.183 10.256
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  53.967      1.482  36.414  < 2e-16 ***
## state.regionSouth            -9.623      1.853  -5.194 4.57e-06 ***
## state.regionNorth Central     0.550      1.961   0.281 0.780328
## state.regionWest              8.033      1.928   4.167 0.000135 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.446 on 46 degrees of freedom
## Multiple R-squared:  0.7155, Adjusted R-squared:  0.697
## F-statistic: 38.57 on 3 and 46 DF,  p-value: 1.3e-12
```

Coeffcients in this model mean that 1) states in the northeast region have average high school graduation rate at 53.967%; 2) state in the south region have average high school graduation rate at 44.344%(=53.967%-9.623%); 3) state in the north central region have average high school graduation rate at 54.517%(=53.967%+0.550%); 4) state in the west region have average high school graduation rate at 62.000%(=53.967%+8.033%).

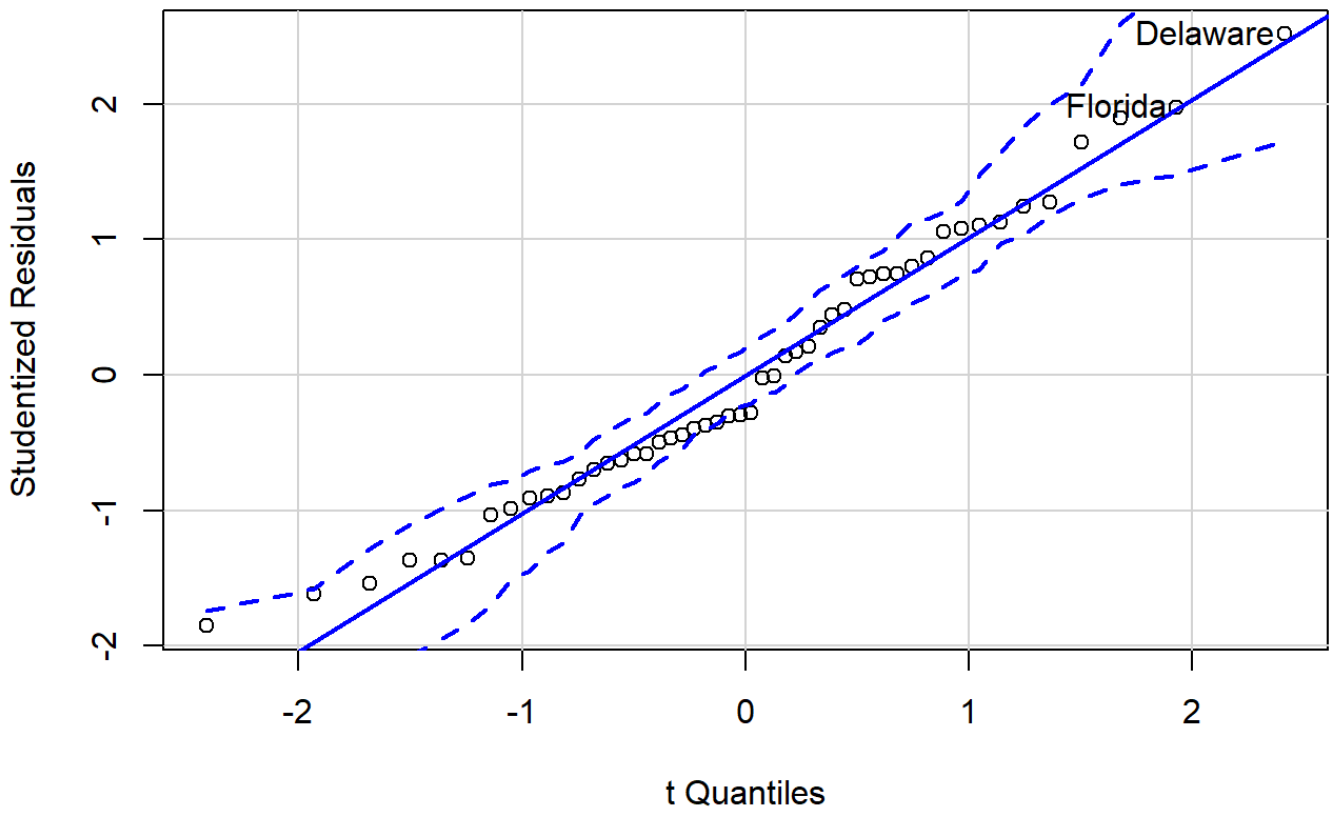
Task 9: Check the constant variance assumption by running Levene's test and the normality assumption by drawing a QQ plot

We need to check the constant variance assumption and the normality assumption but not the linearity assumption because the last one is trivially maintained when the predictor is a categorical variable.

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  0.9671 0.4164
##       46
```



The output from Levene's test suggests that the constant variance assumption is NOT tenable here (p=0.4164). The boxplot above shows the unequal variance clearly.




```
## Delaware  Florida
##        8        9
```

The plot shows evidence of normality since almost all points fall within the 95% confidence bands, which means our sample does not differ much from the theoretical expectation that the error term of the model is normally distributed.