HUDM 5123 - Linear Models and Experimental Design
Lab 02 - OLS Diagnostics

1 The Data

For lab today we will use the `state.x77` data that is built into R. Begin by accessing the help information on the data set by typing `help(state.x77)` or `?state.x77`.
The data:

```
> state.x77
               Population Income Illiteracy Life Exp Murder HS Grad Frost   Area
Alabama              3615   3624        2.1    69.05   15.1    41.3    20  50708
Alaska                365   6315        1.5    69.31   11.3    66.7   152 566432
Arizona              2212   4530        1.8    70.55    7.8    58.1    15 113417
Arkansas             2110   3378        1.9    70.66   10.1    39.9    65  51945
California          21198   5114        1.1    71.71   10.3    62.6    20 156361
Colorado             2541   4884        0.7    72.06    6.8    63.9   166 103766
Connecticut          3100   5348        1.1    72.48    3.1    56.0   139   4862
Delaware              579   4809        0.9    70.06    6.2    54.6   103   1982
Florida              8277   4815        1.3    70.66   10.7    52.6    11  54090
Georgia              4931   4091        2.0    68.54   13.9    40.6    60  58073
Hawaii                868   4963        1.9    73.60    6.2    61.9     0   6425
Idaho                 813   4119        0.6    71.87    5.3    59.5   126  82677
Illinois            11197   5107        0.9    70.14   10.3    52.6   127  55748
Indiana              5313   4458        0.7    70.88    7.1    52.9   122  36097
Iowa                 2861   4628        0.5    72.56    2.3    59.0   140  55941
Kansas               2280   4669        0.6    72.58    4.5    59.9   114  81787
Kentucky             3387   3712        1.6    70.10   10.6    38.5    95  39650
Louisiana            3806   3545        2.8    68.76   13.2    42.2    12  44930
Maine                1058   3694        0.7    70.39    2.7    54.7   161  30920
Maryland             4122   5299        0.9    70.22    8.5    52.3   101   9891
Massachusetts        5814   4755        1.1    71.83    3.3    58.5   103   7826
Michigan             9111   4751        0.9    70.63   11.1    52.8   125  56817
Minnesota            3921   4675        0.6    72.96    2.3    57.6   160  79289
Mississippi          2341   3098        2.4    68.09   12.5    41.0    50  47296
Missouri             4767   4254        0.8    70.69    9.3    48.8   108  68995
Montana               746   4347        0.6    70.56    5.0    59.2   155 145587
Nebraska             1544   4508        0.6    72.60    2.9    59.3   139  76483
Nevada                590   5149        0.5    69.03   11.5    65.2   188 109889
New Hampshire         812   4281        0.7    71.23    3.3    57.6   174   9027
New Jersey           7333   5237        1.1    70.93    5.2    52.5   115   7521
New Mexico           1144   3601        2.2    70.32    9.7    55.2   120 121412
New York            18076   4903        1.4    70.55   10.9    52.7    82  47831
North Carolina       5441   3875        1.8    69.21   11.1    38.5    80  48798
North Dakota          637   5087        0.8    72.78    1.4    50.3   186  69273
Ohio                10735   4561        0.8    70.82    7.4    53.2   124  40975
Oklahoma             2715   3983        1.1    71.42    6.4    51.6    82  68782
Oregon               2284   4660        0.6    72.13    4.2    60.0    44  96184
Pennsylvania        11860   4449        1.0    70.43    6.1    50.2   126  44966
Rhode Island          931   4558        1.3    71.90    2.4    46.4   127   1049
South Carolina       2816   3635        2.3    67.96   11.6    37.8    65  30225
South Dakota          681   4167        0.5    72.08    1.7    53.3   172  75955
Tennessee            4173   3821        1.7    70.11   11.0    41.8    70  41328
Texas               12237   4188        2.2    70.90   12.2    47.4    35 262134
Utah                 1203   4022        0.6    72.90    4.5    67.3   137  82096
Vermont               472   3907        0.6    71.64    5.5    57.1   168   9267
Virginia             4981   4701        1.4    70.08    9.5    47.8    85  39780
Washington           3559   4864        0.6    71.72    4.3    63.5    32  66570
West Virginia        1799   3617        1.4    69.48    6.7    41.6   100  24070
Wisconsin            4589   4468        0.7    72.48    3.0    54.5   149  54464
Wyoming               376   4566        0.6    70.29    6.9    62.9   173  97203
```

Examine the structure via `str(state.x77)` and note that it is a numeric matrix, not a data frame. Convert it to a data frame and call it "dat" via `dat <- data.frame(state.x77)`. Use the functions `names()`, `head()`, `tail()`, `dim()`, and `str()` to examine the data frame. Note that the `data.frame()` function changes white space in variable names to dots. For example, "Life Exp" becomes "Life.Exp". We will begin our analyses by running a multiple regression of life expectancy on murder rate, high school graduation rate, frost, and illiteracy rate; assign it to the name `lm1`.

```
lm1 <- lm(Life.Exp ~ Murder + Illiteracy + Frost, data = dat)

> summary(lm1)

Call:
lm(formula = Life.Exp ~ Murder + Illiteracy + Frost, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-1.59010 -0.46961  0.00394  0.57060  1.92292

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 74.556717   0.584251 127.611  < 2e-16 ***
Murder      -0.280047   0.043394  -6.454 6.03e-08 ***
Illiteracy  -0.601761   0.298927  -2.013  0.04998 *
Frost       -0.008691   0.002959  -2.937  0.00517 **
---

Residual standard error: 0.7911 on 46 degrees of freedom
Multiple R-squared:  0.6739,Adjusted R-squared:  0.6527
F-statistic: 31.69 on 3 and 46 DF,  p-value: 2.915e-11
```

Task 1 *Run a multiple regression of state high school graduation rate on illiteracy rate, income, and state area. Write out (a) the model and (b) the prediction equation with estimated coefficients. Report and interpret the* R^2 *value and the residual standard error and its degrees of freedom.*

Before moving to diagnostics, it is a good idea to examine the data graphically to the extent possible. Since we are working with three predictors, it is not easy to visualize the complete data relationships. Instead, we can use some univariate and multivariate summaries to get some basic sense about how the data interrelate and how they look on a variable-by-variable basis. Beginning with univariate plots, create univariate histograms for the four variables in our model.

```
hist(dat$Murder,
     breaks = 15,
     xlab = "Number of Murders per 100,000",
     main = "State Murder Rates")
hist(dat$Life.Exp,
```

```
     breaks = 15,
     xlab = "Life Expectancy (yrs)",
     main = "State Life Expectancies")
hist(dat$Frost,
     breaks = 15,
     xlab = "Avg. # Days Below Freezing",
     main = "State Frost")
hist(dat$Illiteracy,
     breaks = 15,
     xlab = "Percent Illiterate (1970)",
     main = "State Illiteracy Rates")
```

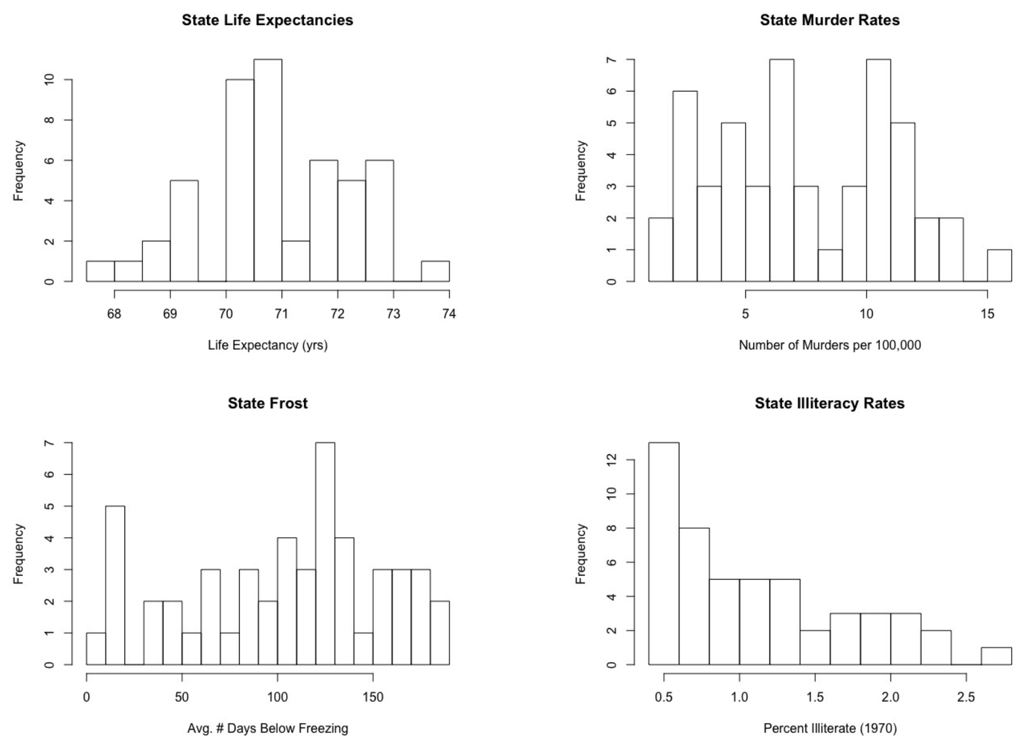

Figure 1: Univariate histograms of state data

Next, create bivariate scatterplots that show the interrelationships between these four variables. Note that we are only concerned with these four variables, so we will subset `dat` to only include them and leave out the rest.
Another way to examine the strength of linear relationships between variables is with the correlation matrix.

```
> (c1 <- round(cor(dat[,c(3,4,5,6)]), 2))
           Illiteracy Life.Exp Murder HS.Grad
Illiteracy       1.00    -0.59   0.70   -0.66
Life.Exp        -0.59     1.00  -0.78    0.58
Murder           0.70    -0.78   1.00   -0.49
HS.Grad         -0.66     0.58  -0.49    1.00
```

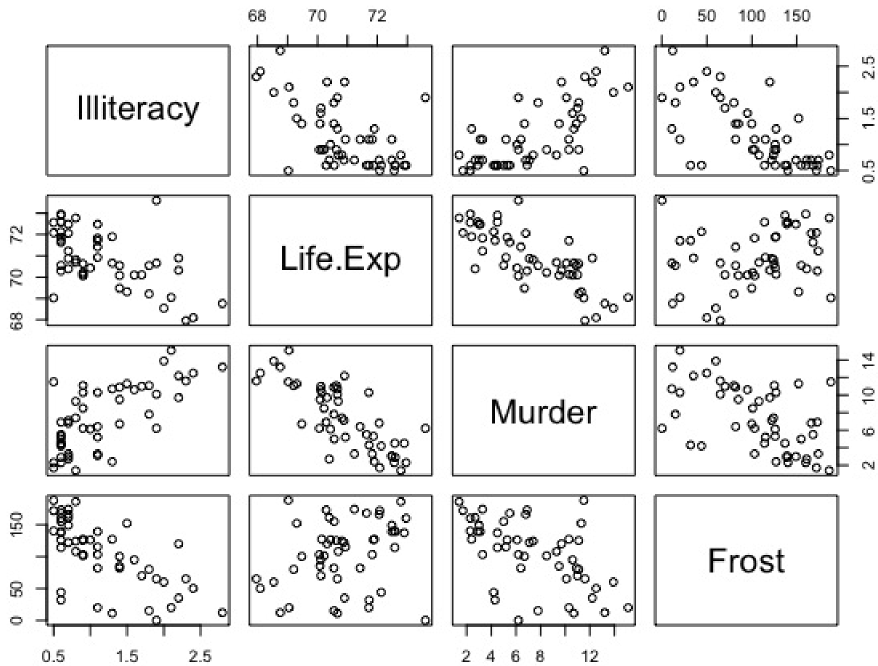

Figure 2: Bivariate scatterplots of state data

Install and load package **corrplot**, which can help to visualize correlation matrices.

```
install.packages("corrplot")
library(corrplot)
corrplot(corr = c1,
         method = "circle",
         order = "hclust")
```

Task 2 *Create univariate and bivariate plots and report the correlation matrix both numerically rounded to two decimal places and visually using package* ***corrplot*** *with ordering determined by hierarchical cluster analysis. Do this for the variables used in Task 1.*

2 Diagnostics

The package **car**, written by the author of our textbook, has most of the functions in it we will use for diagnostics in lab today. Install (if you haven't already) and load the package:

```
install.packages("car")
library(car)
```

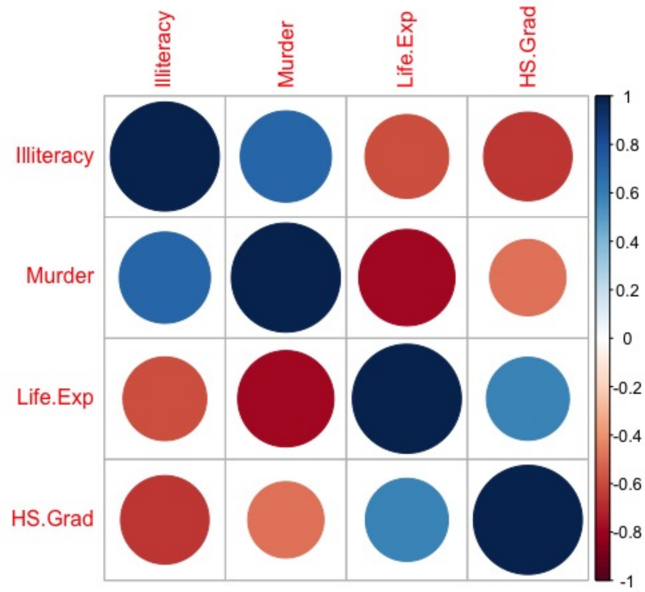

Figure 3: Visualization of the correlation matrix via **corrplot**

2.1 Leverage, Discrepancy, and Influence

In line with the notes, we will begin by checking for points with high influence. Access the help file on the `influencePlot()` function with `help(influencePlot)` or `?influencePlot` and read the description. Run the function on the output `lm1`.

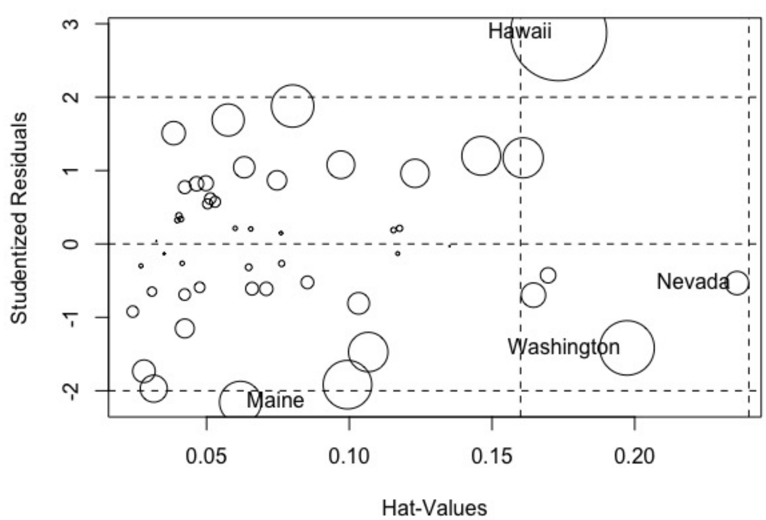

Figure 4: Influence plot output from package **car**

Task 3 *Create and influence plot with the regression model you fit in Task 1 and paste a* *copy of the plot into your lab write-up. Describe the points that (a) have highest leverage, (b) most discrepancy, and (c) are most influential. Should influential points be thrown out here? Why or why not?*

2.2 Normality of Error Term

With only 50 observations, it will be a challenge to assess normality using a histogram (see the top left panel in Figure 1 for a histogram of the outcome). Instead, we will use a QQ plot using the `qqPlot()` function from package **car**.

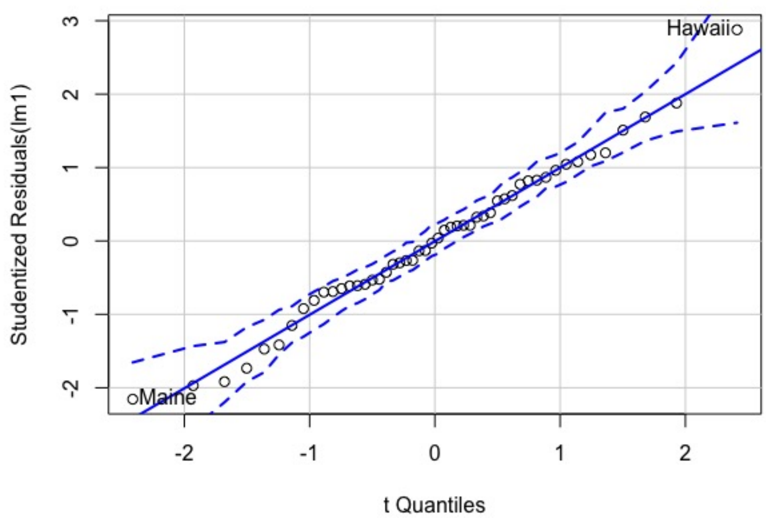

Figure 5: QQ plot from package **car**

Task 4 *Create and interpret a QQ plot of studentized residuals for `lm1` with the function `qqPlot(lm1)`. Does the plot show evidence of non-normality or not? Save the QQ plot as a jpeg and copy and paste it into your lab document.*

2.3 Constant Error Variance

To check for constant error variance we will examine a plot of studentized residuals against the ordered fitted (i.e., predicted) values. To create this plot, use the function `residualPlot(lm1, type = "rstudent")`.

Task 5 *Create a similar plot with data from the regression in Task 1. Is there evidence for non-constant error variance ? Why or why not? Save the residual plot as a jpeg and copy and paste it into your lab document.*

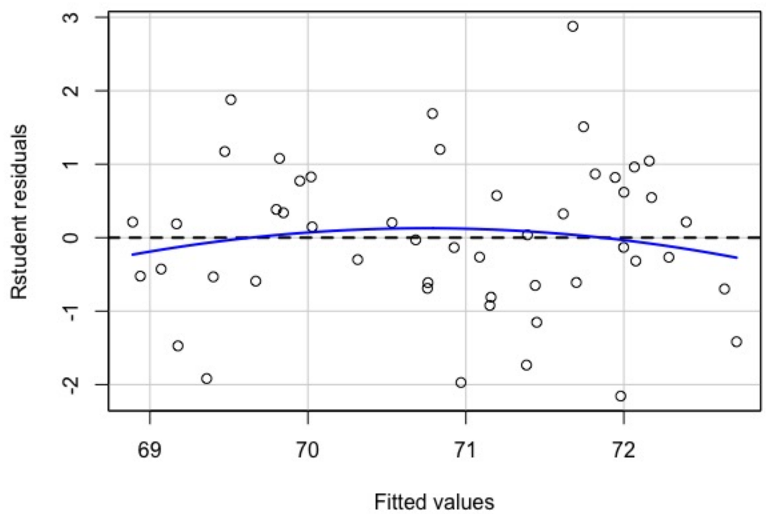

Figure 6: Plot of studentized residuals against predicted, aka "fitted", values from package **car**

2.4 Linearity

Component-plus-residual plots allow us to check on the linearity assumption for each predictor variable. Code to create the CR plots uses the `crPlot()` function in **car**.

```
crPlot(lm1, variable = "Illiteracy")
crPlot(lm1, variable = "Frost")
crPlot(lm1, variable = "Murder")
```

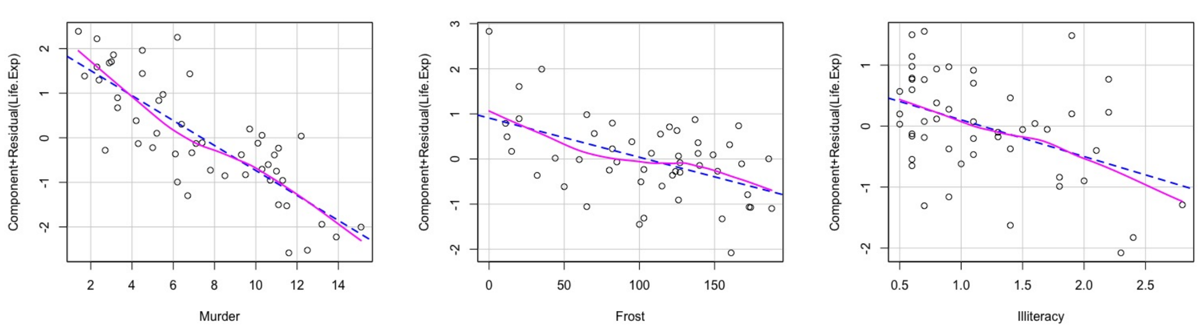

Figure 7: CR-plots from package **car**

Task 6 *Refer back to the reading in Fox and briefly summarize what a CR plot is. Then, create CR plots for the predictors used in the model fit in Task 1. After examining the CR plots, what do you conclude about the tenability of the linearity assumption? Be specific using language motivated by Fox.*

2.5 Multicollinearity

Use the vif() function to get VIFs.

```
vif(lm1)
    Murder Illiteracy      Frost
  2.009008   2.599152   1.852747
```

Task 7 *Calculate and report VIFs for the model you fit in Task 1. For the variable with the largest VIF, demonstrate how to calculate the VIF value for that variable by first calculating* R_j^2 *for that variable and then using* R_j^2 *to determine VIF.*

3 Diagnostics with a Categorical Predictor

The variable `state.region` is a factor that denotes whether each state is in the Northeast, South, North Central, or West. Add it to the data frame by using the column bind function `cbind()` via `dat <- cbind(dat, state.region)`.

```
str(dat)
'data.frame': 50 obs. of  9 variables:
 $ Population  : num  3615 365 2212 2110 21198 ...
 $ Income      : num  3624 6315 4530 3378 5114 ...
 $ Illiteracy  : num  2.1 1.5 1.8 1.9 1.1 0.7 1.1 0.9 1.3 2 ...
 $ Life.Exp    : num  69 69.3 70.5 70.7 71.7 ...
 $ Murder      : num  15.1 11.3 7.8 10.1 10.3 6.8 3.1 6.2 10.7 13.9 ...
 $ HS.Grad     : num  41.3 66.7 58.1 39.9 62.6 63.9 56 54.6 52.6 40.6 ...
 $ Frost       : num  20 152 15 65 20 166 139 103 11 60 ...
 $ Area        : num  50708 566432 113417 51945 156361 ...
 $ state.region: Factor w/ 4 levels "Northeast","South",..: 2 4 4 2 4 4 1 2 2 2 ...
```

Run a regression of life expectancy on state region (the categorical factor).

```
> summary(lm2)

Call:
lm(formula = Life.Exp ~ state.region, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2046 -0.8836  0.3638  0.8083  2.3654

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               71.26444    0.36068 197.584  < 2e-16 ***
state.regionSouth         -1.55819    0.45085  -3.456  0.00119 **
state.regionNorth Central  0.50222    0.47713   1.053  0.29803
state.regionWest          -0.02983    0.46920  -0.064  0.94958
```

```
---

Residual standard error: 1.082 on 46 degrees of freedom
Multiple R-squared:  0.3901,Adjusted R-squared:  0.3503
F-statistic: 9.806 on 3 and 46 DF,  p-value: 4.083e-05
```

Task 8 *Run a regression of high school graduation rate on state region factor and interpret the coefficients in context.*

With categorical factors, we will primarily focus on the constant variance assumption and the normality assumption. Create a categorical factor using the population variable for demonstration by cutting the variable at the values of 1852 and 4164, which mark the 33rd and 67th percentiles, respectively.

```
dat$pop_cat <- cut(x = dat$Population,
                   breaks = c(0, 1852, 4164, Inf),
                   labels = c("Small", "Moderate", "Large"))
```

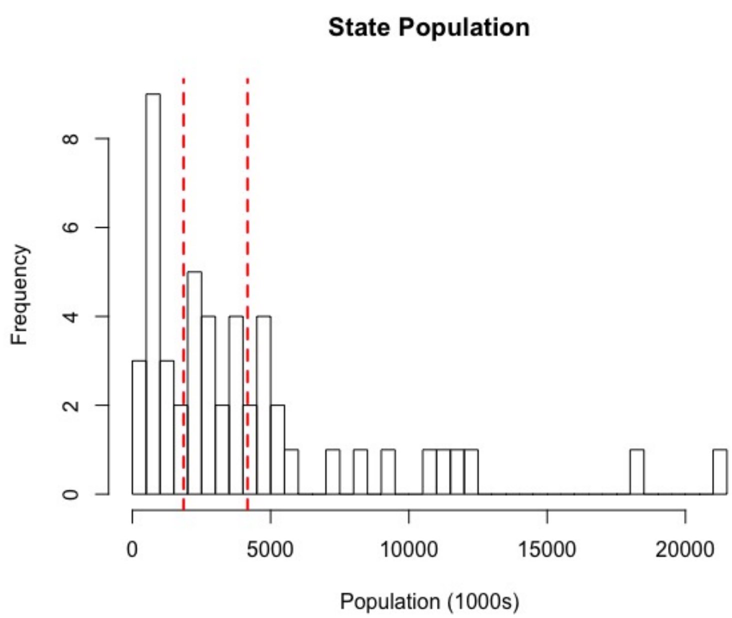

Figure 8: Cut points for population factor

Run a regression analysis of high school graduation rate on the state population factor.

```
lm3 <- lm(HS.Grad ~ pop_cat,
+            data = dat)
summary(lm3)

Call:
lm(formula = HS.Grad ~ pop_cat, data = dat)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-15.623  -3.740   1.944   5.426  12.488

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       57.224      1.857  30.818   <2e-16 ***
pop_catModerate   -5.811      2.667  -2.179   0.0344 *
pop_catLarge      -6.635      2.626  -2.527   0.0149 *
---

Residual standard error: 7.656 on 47 degrees of freedom
Multiple R-squared:  0.1382,Adjusted R-squared:  0.1016
F-statistic: 3.769 on 2 and 47 DF,  p-value: 0.03032
```

Check constant variance by running Levene's test, calculating max variance ratio, and visually inspecting the data. Output from Levene's test suggests the constant variance assumption is not tenable here ($p = .046$).

```
> leveneTest(lm3)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value  Pr(>F)
group  2  3.2996 0.04561 *
      47
```

Group sample variances may be calculated as follows.

```
by(data = dat$HS.Grad, INDICES = dat$pop_cat, FUN = var)
dat$pop_cat: Small
[1] 47.78316
------------------------------------------------------------------------
dat$pop_cat: Moderate
[1] 92.69317
------------------------------------------------------------------------
dat$pop_cat: Large
[1] 37.4911
```

The maximum variance ratio is $92.7 / 37.5 = 2.47$. The boxplot may be produced as follows.

```
boxplot(HS.Grad ~ pop_cat,
        data = dat,
        xlab = "State Population Category",
        ylab = "Outcome")
```

Check normality by examining the QQ plot.

```
qqPlot(lm3,
       xlab = "t Quantiles",
       ylab = "Studentized Residuals")
```

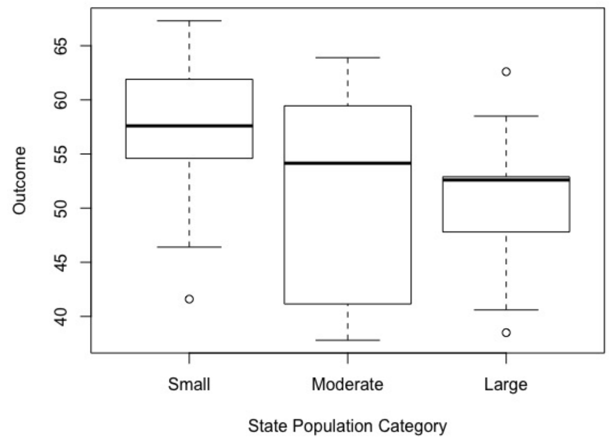

Figure 9: Boxplots of high school graduation rate by state population category

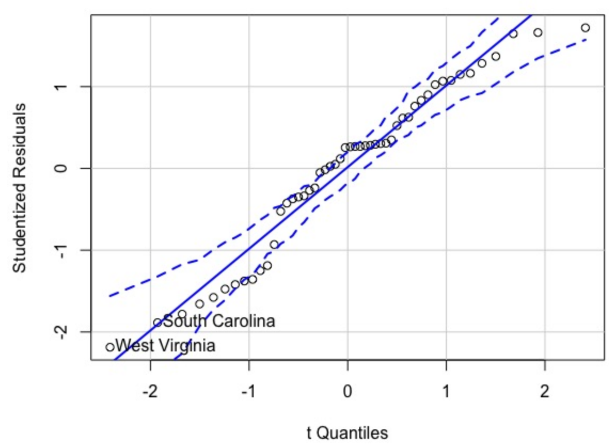

Figure 10: QQ plot for the residuals due to regressing high school graduation rate on state population

Task 9 *Discuss which assumptions need to be checked in this case and which do not, and why. Then check the assumptions that should be checked.*