Linear Regression Models

Contents

1 Simple Linear Regression Model **2**

2 Polynomial Regression Models **4**

3 Multiple Linear Regression Models **6**
- 3.1 Exploratory Data Analysis (EDA) . . . 7
- 3.2 Model Fitting . . . 8
- 3.3 Model Diagnostics . . . 9
- 3.4 Model Selection . . . 13

4 Regression Models with Interaction Terms **16**

5 Robust Regression Models **17**

6 References **18**

1 Simple Linear Regression Model

```r
# fitting a simple linear regression model
fit <- lm(weight~height,data = women)

# get the summary output of the model
summary(fit)
```

```
##
## Call:
## lm(formula = weight ~ height, data = women)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7333 -1.1333 -0.3833  0.7417  3.1167
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -87.51667    5.93694  -14.74 1.71e-09 ***
## height        3.45000    0.09114   37.85 1.09e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.525 on 13 degrees of freedom
## Multiple R-squared:  0.991,  Adjusted R-squared:  0.9903
## F-statistic:  1433 on 1 and 13 DF,  p-value: 1.091e-14
```

```r
# the fitted values
fitted(fit)
```

```
##        1        2        3        4        5        6        7        8
## 112.5833 116.0333 119.4833 122.9333 126.3833 129.8333 133.2833 136.7333
##        9       10       11       12       13       14       15
## 140.1833 143.6333 147.0833 150.5333 153.9833 157.4333 160.8833
```

```r
fit$fitted.values
```

```
##        1        2        3        4        5        6        7        8
## 112.5833 116.0333 119.4833 122.9333 126.3833 129.8333 133.2833 136.7333
##        9       10       11       12       13       14       15
## 140.1833 143.6333 147.0833 150.5333 153.9833 157.4333 160.8833
```

```r
# the residuals
residuals(fit)
```

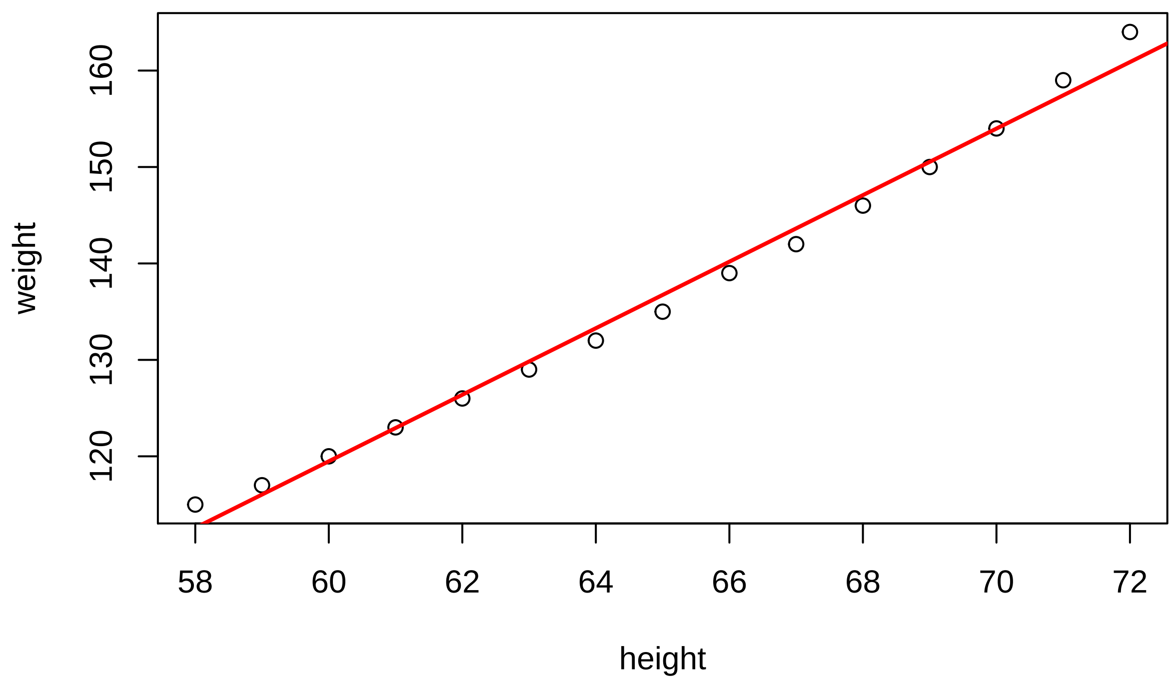
```
##           1           2           3           4           5           6
##  2.41666667  0.96666667  0.51666667  0.06666667 -0.38333333 -0.83333333
##           7           8           9          10          11          12
## -1.28333333 -1.73333333 -1.18333333 -1.63333333 -1.08333333 -0.53333333
##          13          14          15
##  0.01666667  1.56666667  3.11666667
```

```r
fit$residuals
```

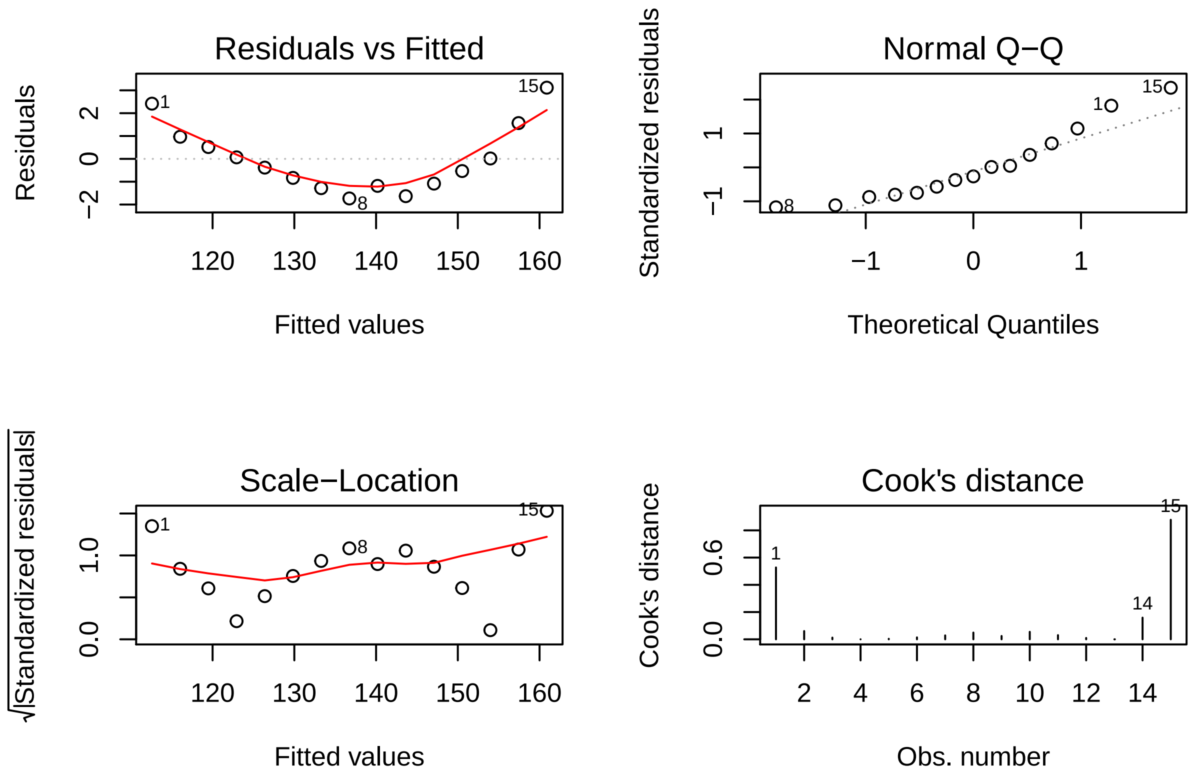
```
##           1           2           3           4           5           6
##  2.41666667  0.96666667  0.51666667  0.06666667 -0.38333333 -0.83333333
##           7           8           9          10          11          12
```

```
## -1.28333333 -1.73333333 -1.18333333 -1.63333333 -1.08333333 -0.53333333
##          13          14          15
##  0.01666667  1.56666667  3.11666667
```

```
# the fitted regression line on the scatterplot
plot(weight ~ height , data = women)
abline(fit, col = "red", lwd = 2)
```




```
# residuals analysis plots
par(mfrow = c(2,2))
plot(fit, 1:4)
```




```
# make predictions
predict(fit, newdata = data.frame(height = 66), interval = "prediction", level = 0.95)
```

```
##        fit     lwr      upr
## 1 140.1833 136.775 143.5916
```

```
predict(fit, newdata = data.frame(height = 66), interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 140.1833 139.3102 141.0565
```

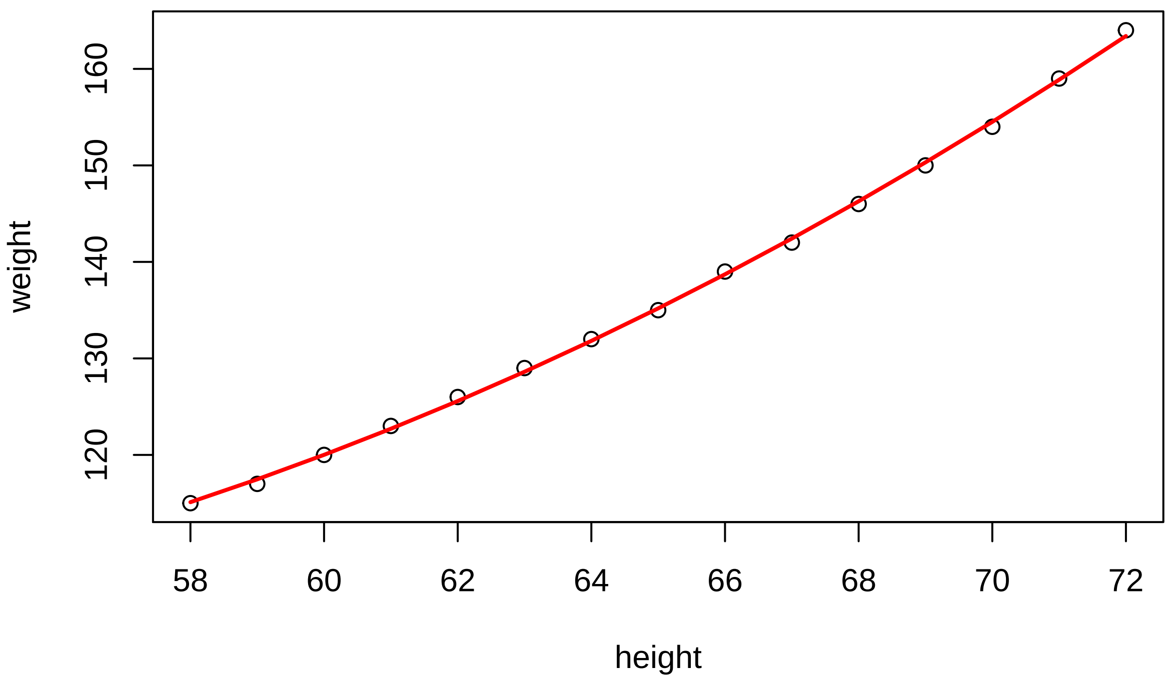
2 Polynomial Regression Models

```
# the predictors contain some polynomial terms
fit2 <- lm(weight ~ height + I(height^2), data = women)
summary(fit2)
```

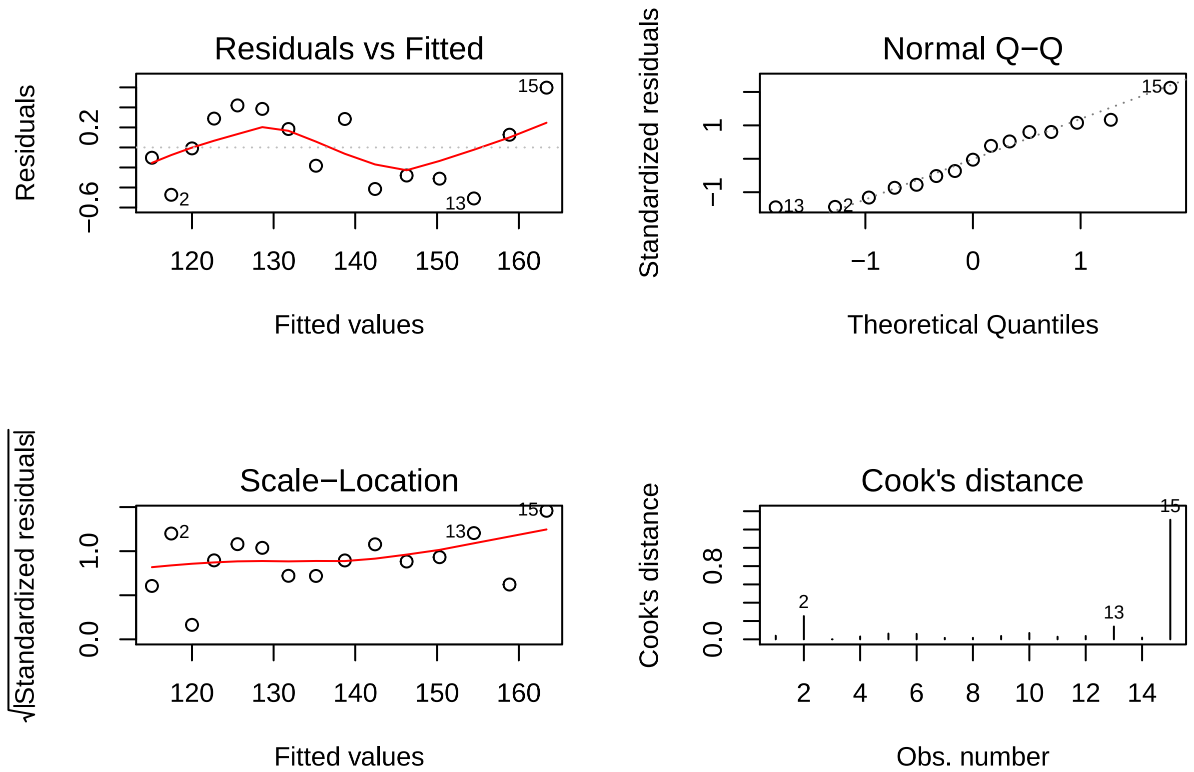
```
##
## Call:
## lm(formula = weight ~ height + I(height^2), data = women)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50941 -0.29611 -0.00941  0.28615  0.59706
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 261.87818   25.19677  10.393 2.36e-07 ***
## height       -7.34832    0.77769  -9.449 6.58e-07 ***
## I(height^2)   0.08306    0.00598  13.891 9.32e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3841 on 12 degrees of freedom
## Multiple R-squared:  0.9995, Adjusted R-squared:  0.9994
## F-statistic: 1.139e+04 on 2 and 12 DF,  p-value: < 2.2e-16
```

```r
# fitted curves on the scatterplot
plot(weight ~ height , data = women)
lines(women$height, fitted(fit2), col = "red", lwd = 2)
```




```r
# residual analysis
par(mfrow = c(2,2))
plot(fit2, 1:4)
```




```
# test if the quadratic term of height is effective
anova(fit, fit2)
```

```
## Analysis of Variance Table
##
## Model 1: weight ~ height
## Model 2: weight ~ height + I(height^2)
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     13 30.2333
## 2     12  1.7701  1    28.463 192.96 9.322e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# since the p values is significant, the quadratic term of height is effective.
```

3 Multiple Linear Regression Models

```
# The data provided in R
library("dplyr")
state.x77 <- as.data.frame(state.x77)
glimpse(state.x77)
```

```
## Rows: 50
## Columns: 8
## $ Population <dbl> 3615, 365, 2212, 2110, 21198, 2541, 3100, 579, 8277, 493...
```

```
## $ Income     <dbl> 3624, 6315, 4530, 3378, 5114, 4884, 5348, 4809, 4815, 40...
## $ Illiteracy <dbl> 2.1, 1.5, 1.8, 1.9, 1.1, 0.7, 1.1, 0.9, 1.3, 2.0, 1.9, 0...
## $ `Life Exp` <dbl> 69.05, 69.31, 70.55, 70.66, 71.71, 72.06, 72.48, 70.06, ...
## $ Murder     <dbl> 15.1, 11.3, 7.8, 10.1, 10.3, 6.8, 3.1, 6.2, 10.7, 13.9, ...
## $ `HS Grad`  <dbl> 41.3, 66.7, 58.1, 39.9, 62.6, 63.9, 56.0, 54.6, 52.6, 40...
## $ Frost      <dbl> 20, 152, 15, 65, 20, 166, 139, 103, 11, 60, 0, 126, 127,...
## $ Area       <dbl> 50708, 566432, 113417, 51945, 156361, 103766, 4862, 1982...
```

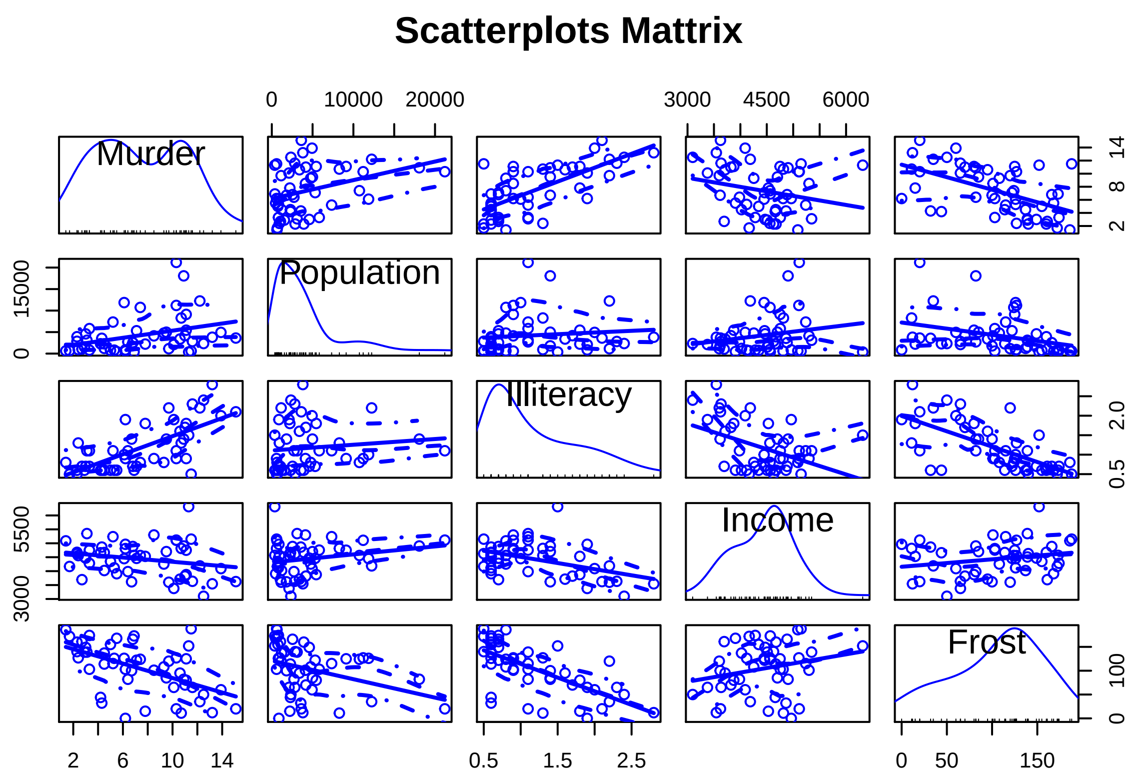
```
# select some variables to form a new dataset named states
states <- state.x77 %>%
   select(Murder, Population, Illiteracy, Income, Frost)
```

3.1 Exploratory Data Analysis (EDA)

```
## EDA
# obtain the correlation of variables of interest
cor(states)
```

```
##                Murder Population Illiteracy     Income      Frost
## Murder      1.0000000  0.3436428  0.7029752 -0.2300776 -0.5388834
## Population  0.3436428  1.0000000  0.1076224  0.2082276 -0.3321525
## Illiteracy  0.7029752  0.1076224  1.0000000 -0.4370752 -0.6719470
## Income     -0.2300776  0.2082276 -0.4370752  1.0000000  0.2262822
## Frost      -0.5388834 -0.3321525 -0.6719470  0.2262822  1.0000000
```

```
# scatterplot matrix
library(car)
scatterplotMatrix(states,spread = FALSE, smoother.args = list(lty = 2),
                  main ="Scatterplots Mattrix")
```



3.2 Model Fitting

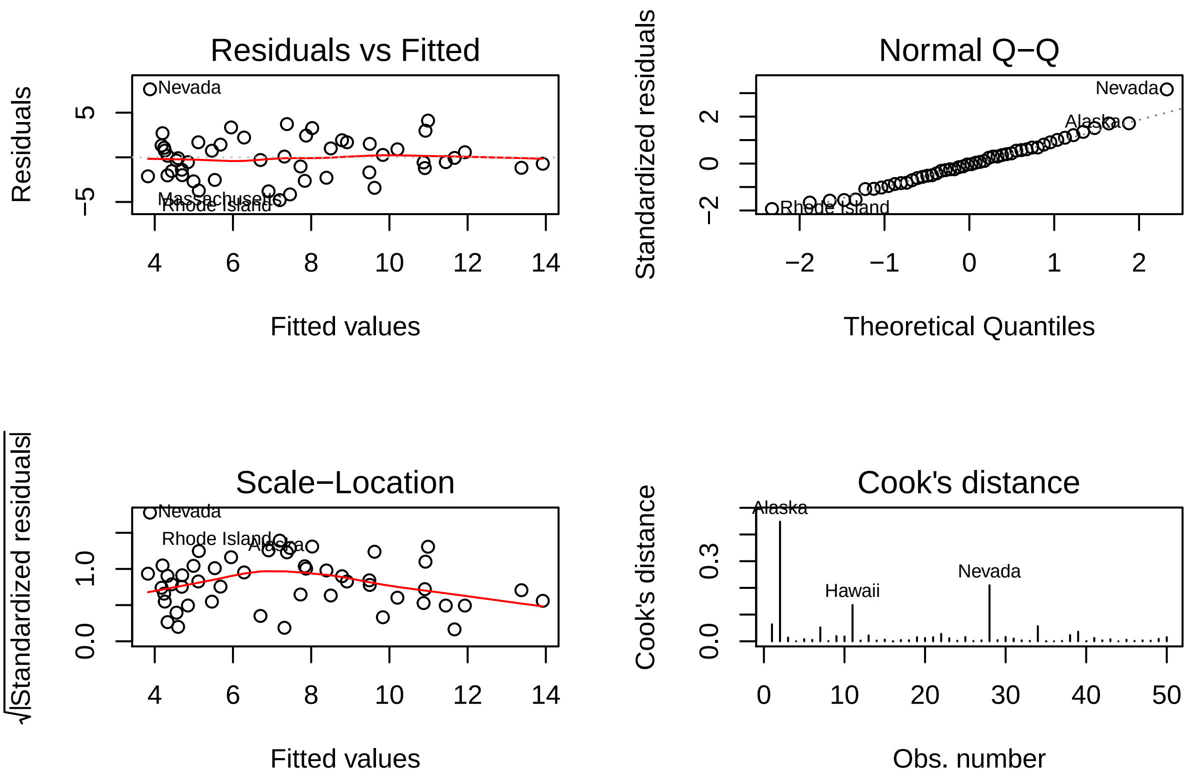
```
# fit a multiple linear regression model
fit <- lm(Murder ~ Population + Illiteracy + Income + Frost,data=states)
# or
# fit <- lm(Murder ~ .,data=states)
summary(fit)
```

```
##
## Call:
## lm(formula = Murder ~ Population + Illiteracy + Income + Frost,
##     data = states)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7960 -1.6495 -0.0811  1.4815  7.6210
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.235e+00  3.866e+00   0.319   0.7510
## Population  2.237e-04  9.052e-05   2.471   0.0173 *
## Illiteracy  4.143e+00  8.744e-01   4.738 2.19e-05 ***
## Income      6.442e-05  6.837e-04   0.094   0.9253
## Frost       5.813e-04  1.005e-02   0.058   0.9541
## ---
```

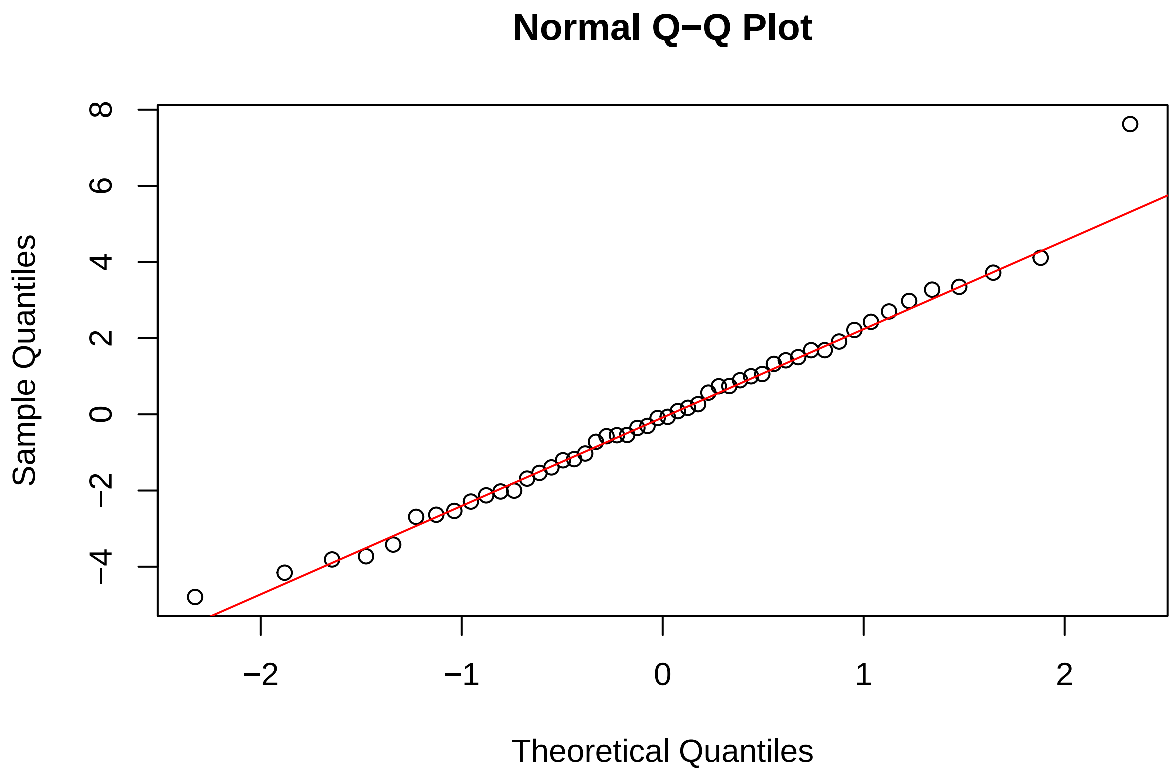
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.535 on 45 degrees of freedom
## Multiple R-squared:  0.567,  Adjusted R-squared:  0.5285
## F-statistic: 14.73 on 4 and 45 DF,  p-value: 9.133e-08
```

3.3 Model Diagnostics

```
# model diagnostics
par(mfrow = c(2,2))
plot(fit, 1:4)
```




```
# check the assumption of normality of residuals
par(mfrow = c(1,1))
qqnorm(residuals(fit))
qqline(residuals(fit), col = "red")
```



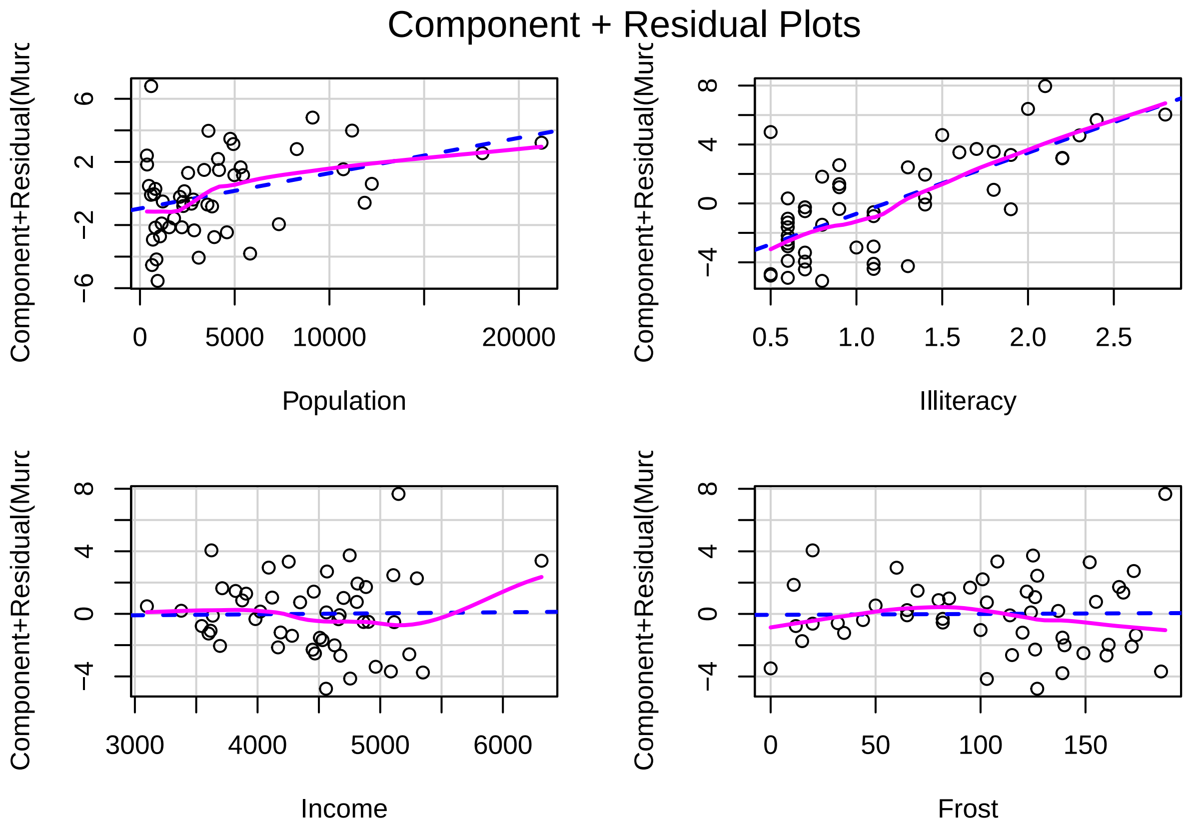

```r
# shapiro Wilk test for normality
shapiro.test(residuals(fit))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit)
## W = 0.98264, p-value = 0.6672
```

```r
# check the assumption of independence of residuals
durbinWatsonTest(fit)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.2006929      2.317691   0.262
##  Alternative hypothesis: rho != 0
```

```r
# check the assumption of linearity
library(car)
crPlots(fit)
```




```r
# check the constant variance of residuals
library(car)
ncvTest(fit)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.746514, Df = 1, p = 0.18632
```

```r
# Test if the multi-collinearity exists
library(car)
vif(fit)
```

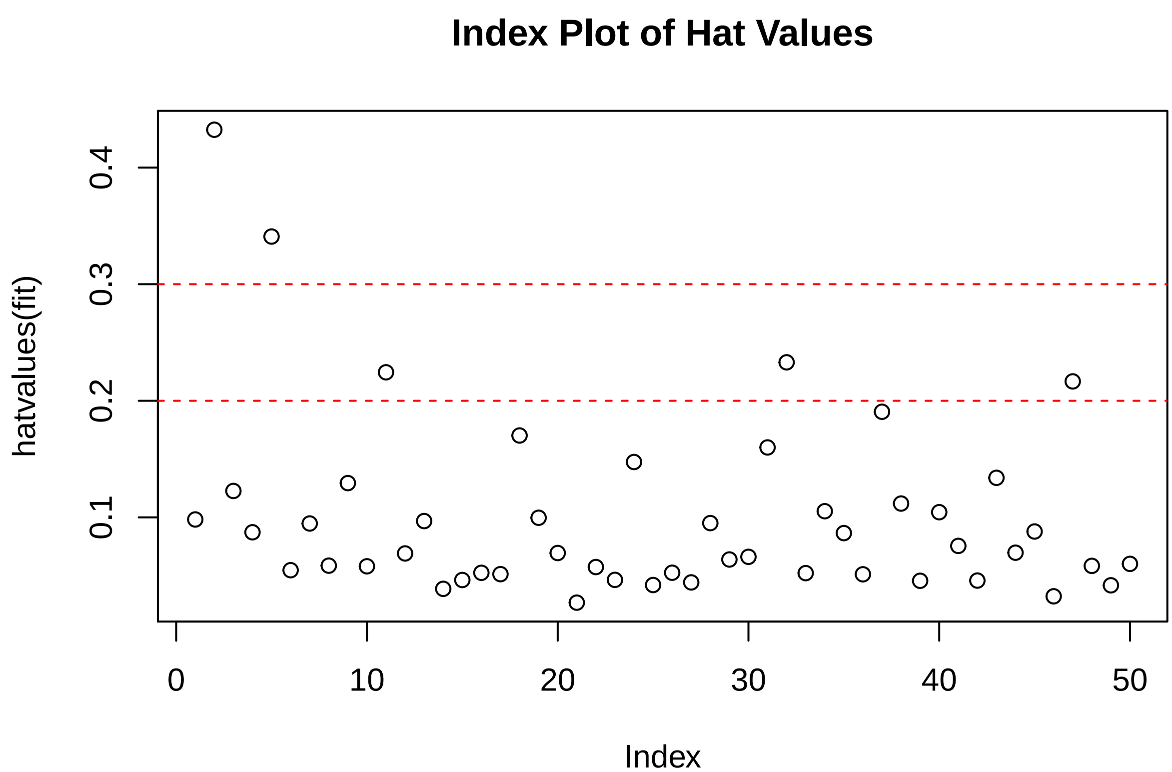
```
## Population Illiteracy     Income      Frost
##   1.245282   2.165848   1.345822   2.082547
```

```r
# Detecting outliers
library(car)
outlierTest(fit)
```

```
##        rstudent unadjusted p-value Bonferroni p
## Nevada 3.542929         0.00095088     0.047544
```

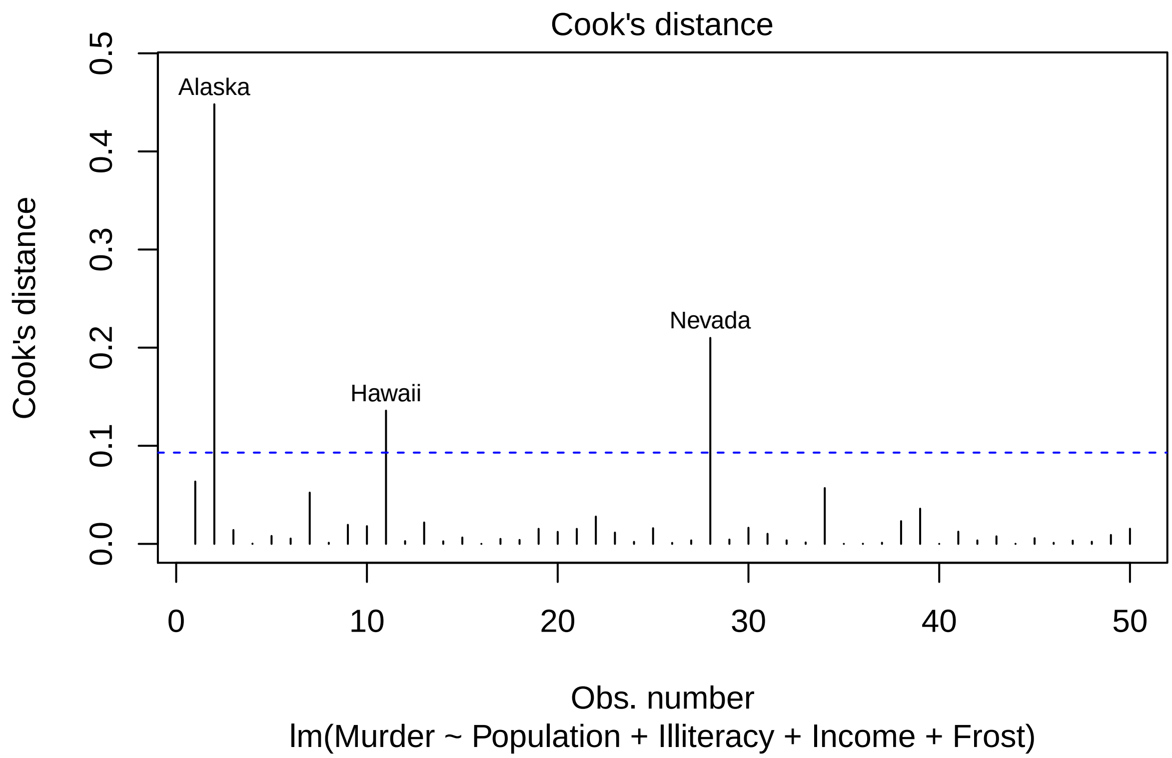
```r
# find high-leverage points
hat.plot <- function(fit){
     p <- length(coefficients(fit))
     n <- length(fitted(fit))
     plot(hatvalues(fit), main="Index Plot of Hat Values")
     abline(h=c(2,3)*p/n, col="red", lty=2)
     identify(1:n, hatvalues(fit), names(hatvalues(fit)))
```

```
}
hat.plot(fit)
```




```
## integer(0)
```

```
# find influential points
cutoff <- 4/(nrow(states)-length(fit$coefficients)-2)
plot(fit,which = 4, cook.levels = cutoff)
abline(h = cutoff, lty = 2, col="blue")
```



3.4 Model Selection

```
#  stepwise regression
library(MASS)
stepAIC(fit, direction="backward")
```

```
## Start:  AIC=97.75
## Murder ~ Population + Illiteracy + Income + Frost
##
##              Df Sum of Sq    RSS     AIC
## - Frost       1     0.021 289.19  95.753
## - Income      1     0.057 289.22  95.759
## <none>                    289.17  97.749
## - Population  1    39.238 328.41 102.111
## - Illiteracy  1   144.264 433.43 115.986
##
## Step:  AIC=95.75
## Murder ~ Population + Illiteracy + Income
##
##              Df Sum of Sq    RSS     AIC
## - Income      1     0.057 289.25  93.763
## <none>                    289.19  95.753
## - Population  1    43.658 332.85 100.783
## - Illiteracy  1   236.196 525.38 123.605
##
```

```
## Step:  AIC=93.76
## Murder ~ Population + Illiteracy
##
##              Df Sum of Sq    RSS     AIC
## <none>                    289.25  93.763
## - Population  1    48.517 337.76  99.516
## - Illiteracy  1   299.646 588.89 127.311
```

```
##
## Call:
## lm(formula = Murder ~ Population + Illiteracy, data = states)
##
## Coefficients:
## (Intercept)   Population   Illiteracy
##   1.6515497    0.0002242    4.0807366
```

```r
stepAIC(fit, direction="forward")
```

```
## Start:  AIC=97.75
## Murder ~ Population + Illiteracy + Income + Frost
```

```
##
## Call:
## lm(formula = Murder ~ Population + Illiteracy + Income + Frost,
##     data = states)
##
## Coefficients:
## (Intercept)   Population   Illiteracy       Income        Frost
##   1.235e+00    2.237e-04    4.143e+00    6.442e-05    5.813e-04
```

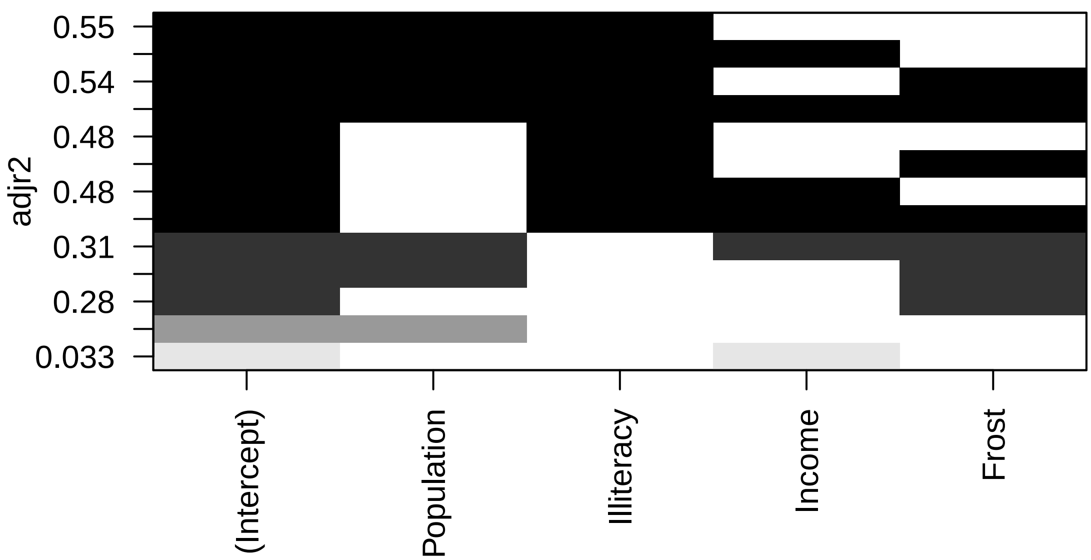
```r
stepAIC(fit, direction="both")
```

```
## Start:  AIC=97.75
## Murder ~ Population + Illiteracy + Income + Frost
##
##              Df Sum of Sq    RSS     AIC
## - Frost       1     0.021 289.19  95.753
## - Income      1     0.057 289.22  95.759
## <none>                    289.17  97.749
## - Population  1    39.238 328.41 102.111
## - Illiteracy  1   144.264 433.43 115.986
##
## Step:  AIC=95.75
## Murder ~ Population + Illiteracy + Income
##
##              Df Sum of Sq    RSS     AIC
## - Income      1     0.057 289.25  93.763
## <none>                    289.19  95.753
## + Frost       1     0.021 289.17  97.749
## - Population  1    43.658 332.85 100.783
## - Illiteracy  1   236.196 525.38 123.605
##
## Step:  AIC=93.76
## Murder ~ Population + Illiteracy
##
##              Df Sum of Sq    RSS     AIC
```

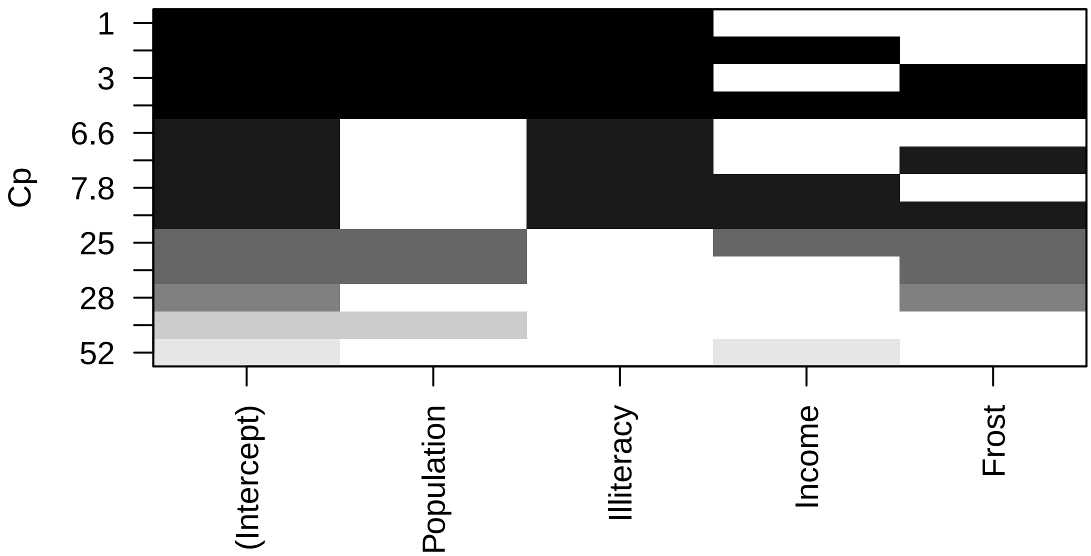
```
## <none>                      289.25  93.763
## + Income      1     0.057 289.19  95.753
## + Frost       1     0.021 289.22  95.759
## - Population  1    48.517 337.76  99.516
## - Illiteracy  1   299.646 588.89 127.311
```

```
##
## Call:
## lm(formula = Murder ~ Population + Illiteracy, data = states)
##
## Coefficients:
## (Intercept)   Population   Illiteracy
##   1.6515497    0.0002242    4.0807366
```

```
# best subset regression model
library(leaps)
leaps_mod <- regsubsets(Murder ~ Population + Illiteracy + Income + Frost,
                        data=states, nbest=4)
plot(leaps_mod, scale="adjr2")
```




```
plot(leaps_mod, scale="Cp")
```



4 Regression Models with Interaction Terms

```
fit1 <- lm(Murder ~ Population + Illiteracy + Income + Frost + Population:Income,
          data=states)
summary(fit1)
```

```
##
## Call:
## lm(formula = Murder ~ Population + Illiteracy + Income + Frost +
##     Population:Income, data = states)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5999 -1.7083 -0.0403  1.4839  7.4744
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -5.807e-01  4.328e+00  -0.134 0.893881
## Population         1.246e-03  1.094e-03   1.139 0.260783
## Illiteracy         3.858e+00  9.266e-01   4.164 0.000144 ***
## Income             5.143e-04  8.359e-04   0.615 0.541565
## Frost             -3.470e-04  1.012e-02  -0.034 0.972794
## Population:Income -2.109e-07  2.249e-07  -0.938 0.353421
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.538 on 44 degrees of freedom
## Multiple R-squared:  0.5754, Adjusted R-squared:  0.5272
## F-statistic: 11.93 on 5 and 44 DF,  p-value: 2.52e-07
```

```
anova(fit, fit1)
```

```
## Analysis of Variance Table
##
## Model 1: Murder ~ Population + Illiteracy + Income + Frost
## Model 2: Murder ~ Population + Illiteracy + Income + Frost + Population:Income
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     45 289.17
## 2     44 283.50  1    5.6677 0.8796 0.3534
```

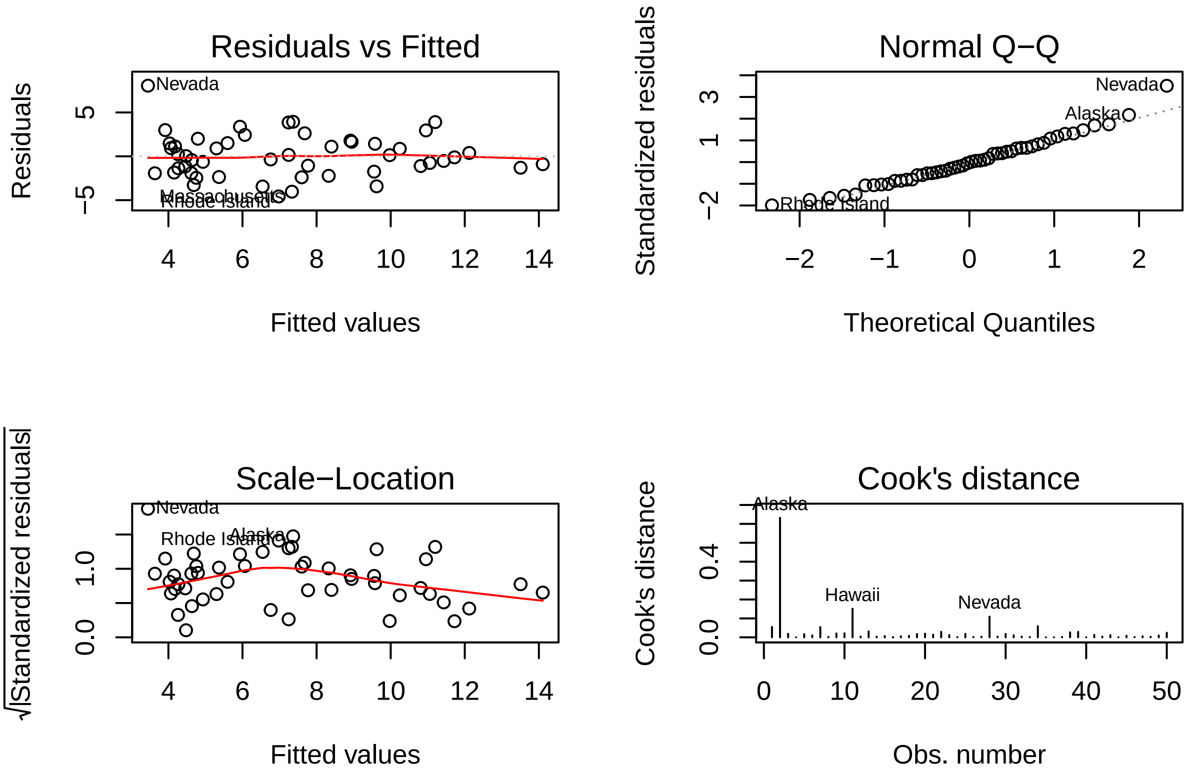
```
# showing the interaction not work
```

5 Robust Regression Models

```
library("MASS")
fit3 <- rlm(Murder ~ Population + Illiteracy + Income + Frost,data=states)
summary(fit3)
```

```
##
## Call: rlm(formula = Murder ~ Population + Illiteracy + Income + Frost,
##     data = states)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.57792 -1.65709 -0.04884  1.49383  8.05141
##
## Coefficients:
##             Value   Std. Error t value
## (Intercept)  2.3910  4.0897     0.5846
## Population   0.0002  0.0001     2.4327
## Illiteracy   4.0439  0.9249     4.3721
## Income      -0.0001  0.0007    -0.1828
## Frost       -0.0022  0.0106    -0.2107
##
## Residual standard error: 2.34 on 45 degrees of freedom
```

```
par(mfrow = c(2,2))
plot(fit3, 1:4)
```



6 References

https://www.datacamp.com/community/tutorials/linear-regression-R

MH Kutner, CJ Nachtsheim, J Neter, W Li (2005), Applied linear statistical models.

Gareth James, Daniela Witten, Trevor Hastie Robert Tibshirani (2013), An Introduction to Statistical Learning with Applications in R.

https://rpubs.com/bryangoodrich/

https://www.scribbr.com/statistics/linear-regression-in-r/

https://stats.oarc.ucla.edu/r/dae/robust-regression/