STOR 590 HW1 Solution

Taebin Kim

Part 1.

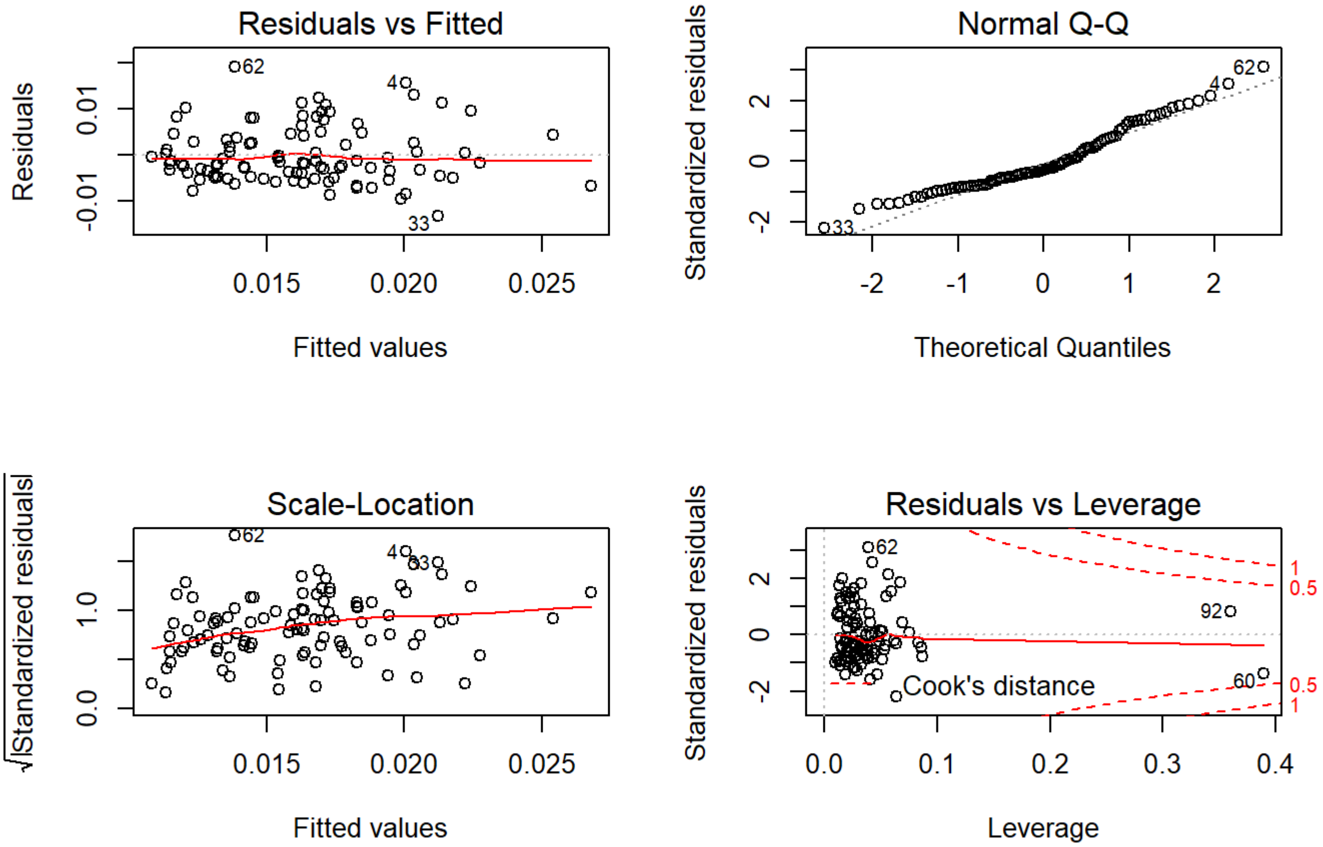
We can construct a regression model to predict PNR as a function of the other 9 numerical variables with the following code. We use forward or backward variable selection to determine the optimal model. Note that we are using a “weight” vector that has entries 0 in places 9 and 78, and a 1 in every other entry.

```
X <- read.csv('ProportionNotReturned.csv', header = T)
wts <- rep(1,100)
wts[c(9,78)] <- c(0,0)
lm1 <- lm(PNR ~ Pop + Rural + MedAge + Travel + Hsgrad + Collgrad
          + MedInc + Black + Hisp, data = X, weights = wts)
step(lm1, trace = FALSE)
```

```
##
## Call:
## lm(formula = PNR ~ Pop + Hsgrad + Black, data = X, weights = wts)
##
## Coefficients:
## (Intercept)          Pop       Hsgrad        Black
##  -7.239e-03    7.561e-09    2.277e-04    1.833e-04
```

It shows that the optimal linear regression model obtained by forward or backward variable selection is `PNR ~ Pop + Hsgrad + Black`. We also use diagnostics to assess various measures of fit, such as whether the residuals appear to be normally distributed.

```
lm1 <- lm(PNR ~ Pop + Hsgrad + Black, data = X, weights = wts)
par(mfrow=c(2,2))
plot(lm1)
```



We can see that the majority of the residuals are normally distributed.

Part 2.

For each of Bladen and Robeson counties, we obtain a 99% prediction interval, based on the results in the other 98 counties.

```
pr1 <- predict(lm1, se.fit = T, interval = 'prediction', level = 0.99, weights = 1)
```

```
## Warning in predict.lm(lm1, se.fit = T, interval = "prediction", level = 0.99, : predictions o
n current data refer to _future_ responses
```

```
pr1$fit[c(9,78),]
```

```
##           fit           lwr        upr
## 9  0.01728530  0.0005795402 0.03399106
## 78 0.01555308 -0.0013119958 0.03241816
```

The result shows that the 99% prediction intervals for the PNR of Bladen and Robeson counties are respectively [0.0005795402, 0.03399106] and [−0.0013119958, 0.03241816].

Part 3.

Based on the result in part 2, we estimate a lower bound on the excess PNR for Bladen and Robeson counties that cannot be explained by natural variability.

```
X$PNR[9] - 0.03399106
```

```
## [1] 0.07910894
```

```
X$PNR[78] - 0.03241816
```

```
## [1] 0.07758184
```

We can observe that the excess PNR of Bladen and Robeson counties are 0.07910894 and 0.07758184, respectively.

Part 4.

Then with the information that the numbers of absentee ballots requested in Bladen and Robeson counties are respectively 8,110 and 16,069, we estimate the total number of absentee ballots that are unaccounted for.

```
8110*(X$PNR[9] - 0.03399106)
```

```
## [1] 641.5735
```

```
16069*(X$PNR[78] - 0.03241816)
```

```
## [1] 1246.663
```

It shows that the estimated total number of absentee ballots that are unaccounted for in Bladen and Robeson counties are 641.5735 and 1246.663, respectively.

Part 5.

Since the the estimated total number of absentee ballots far exceeds the upper bound of its 99% prediction interval in Bladen and Robeson counties, we conlude that there were voting irregularities in both counties.