HW1

Ted Henson

1/14/2020

```
setwd('~/Advanced Linear Models/HW 1/')
voters <- read_csv("ProportionNotReturned.csv")
```

```
## Parsed with column specification:
## cols(
##   County = col_character(),
##   PNR = col_double(),
##   Pop = col_double(),
##   Rural = col_double(),
##   MedAge = col_double(),
##   Travel = col_double(),
##   Hsgrad = col_double(),
##   Collgrad = col_double(),
##   MedInc = col_double(),
##   Black = col_double(),
##   Hisp = col_double(),
##   AbsBal = col_double()
## )
```

Problem 1

```
weights = ifelse(voters$County %in% c('ROBESON', 'BLADEN') == F, 1, 0)
full.model = lm(PNR ~ ., data = voters[,2:ncol(voters)], weights = weights)
```

```
confint(full.model)
```

```
##                     2.5 %       97.5 %
## (Intercept) -7.264826e-02 1.700502e-02
## Pop         -2.862510e-08 1.010211e-07
## Rural       -4.723762e-05 1.227360e-04
## MedAge      -4.989209e-04 2.951927e-04
## Travel      -4.072824e-04 7.783876e-04
## Hsgrad      -1.312853e-04 9.871861e-04
## Collgrad    -4.403782e-04 3.704111e-04
## MedInc      -3.318743e-07 3.522623e-07
## Black        9.572748e-05 2.857638e-04
## Hisp        -2.223038e-04 5.871728e-04
## AbsBal      -4.630629e-07 1.934654e-07
```

```
none = lm(PNR ~ 1, data = voters)
step.mod = stepAIC(none, scope = list(upper = full.model),  direction = 'both')
```

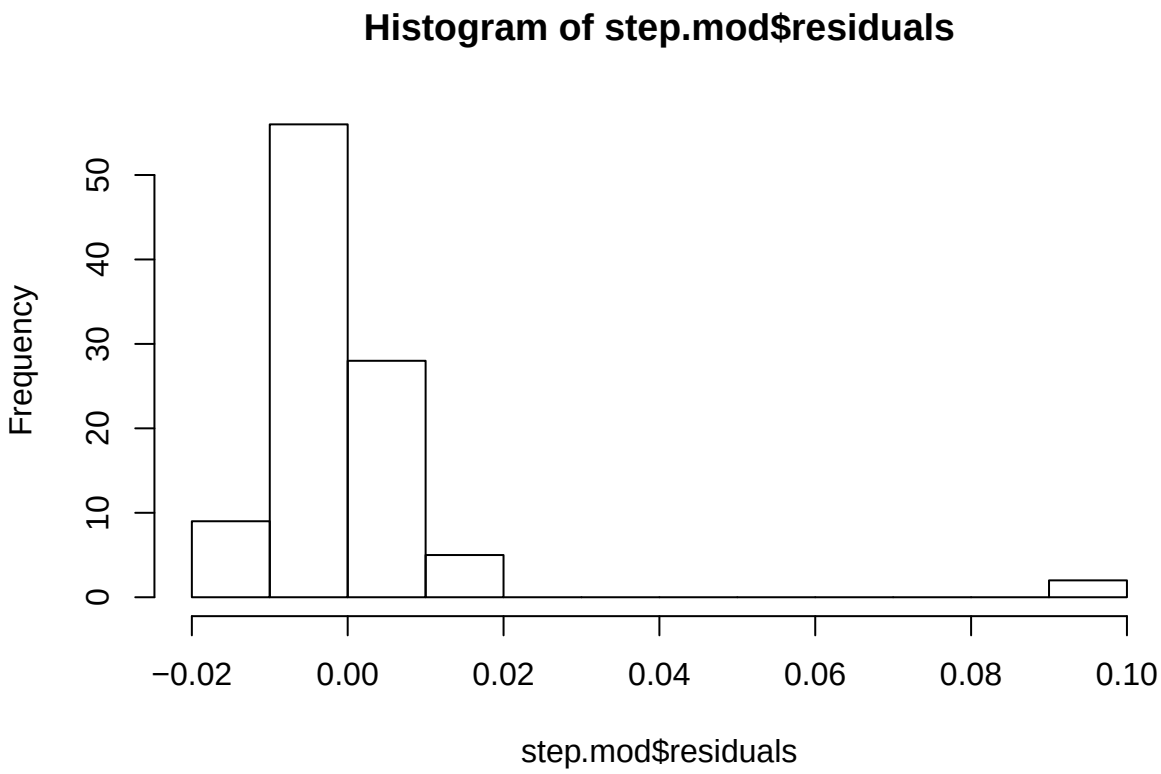
```
## Start:  AIC=-837.59
```

```
## PNR ~ 1
##
##            Df  Sum of Sq      RSS     AIC
## + Black     1 0.00138120 0.021198 -841.90
## + MedAge    1 0.00072221 0.021857 -838.84
## <none>                   0.022580 -837.59
## + MedInc    1 0.00038528 0.022194 -837.31
## + Hsgrad    1 0.00034363 0.022236 -837.12
## + Pop       1 0.00023380 0.022346 -836.63
## + Hisp      1 0.00017338 0.022406 -836.36
## + Collgrad  1 0.00011878 0.022461 -836.12
## + AbsBal    1 0.00010924 0.022470 -836.07
## + Travel    1 0.00003466 0.022545 -835.74
## + Rural     1 0.00000129 0.022578 -835.59
##
## Step:  AIC=-841.9
## PNR ~ Black
##
##            Df  Sum of Sq      RSS     AIC
## <none>                   0.021198 -841.90
## + MedAge    1 0.00039665 0.020802 -841.79
## + Hisp      1 0.00020613 0.020992 -840.88
## + Pop       1 0.00020608 0.020992 -840.88
## + AbsBal    1 0.00010566 0.021093 -840.40
## + MedInc    1 0.00007419 0.021124 -840.25
## + Hsgrad    1 0.00003431 0.021164 -840.06
## + Travel    1 0.00002421 0.021174 -840.01
## + Rural     1 0.00000629 0.021192 -839.93
## + Collgrad  1 0.00000001 0.021198 -839.90
## - Black     1 0.00138120 0.022580 -837.59
```

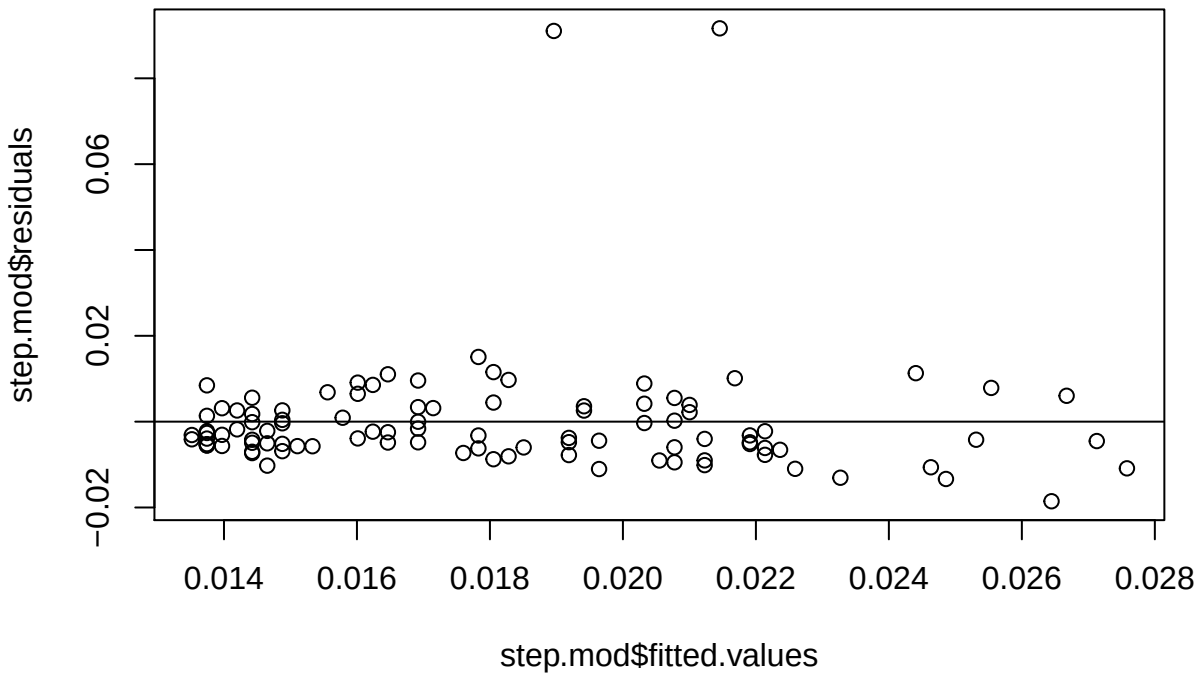
```
summary(step.mod)
```

```
##
## Call:
## lm(formula = PNR ~ Black, data = voters)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.018546 -0.005792 -0.003218  0.003136  0.091644
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.352e-02  2.351e-03   5.750 1.02e-07 ***
## Black       2.268e-04  8.977e-05   2.527   0.0131 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01471 on 98 degrees of freedom
## Multiple R-squared:  0.06117,    Adjusted R-squared:  0.05159
## F-statistic: 6.385 on 1 and 98 DF,  p-value: 0.01311
```

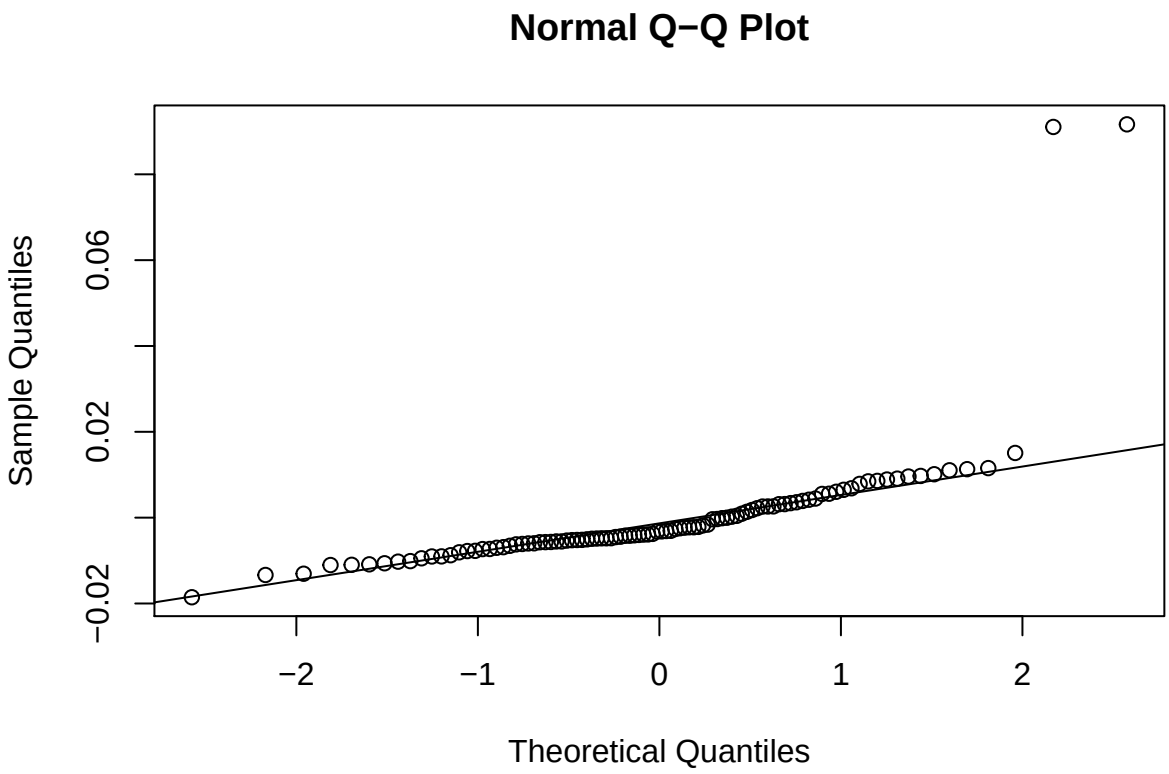
```
hist(step.mod$residuals)
```




```
plot(step.mod$fitted.values, step.mod$residuals)
abline(a = 0, b= 0)
```




```
qqnorm(step.mod$residuals)
qqline(step.mod$residuals)
```



Problem 2

```
intervals = predict.lm(full.model,
                       set.fit = T,
                       interval = 'prediction',
                       level = .99,
                       weights = 1)
cbind(voters[which(weights == 0),'County'], intervals[which(weights == 0),])
```

```
##     County        fit           lwr        upr
## 9   BLADEN 0.01751575  0.0004757346 0.03455576
## 78 ROBESON 0.01662637 -0.0009860686 0.03423880
```

Problem 3

```
dat = data.frame(County = voters$County[which(weights == 0)],
Excess.PNR = voters$PNR[which(weights == 0)] - intervals[which(weights == 0),'upr'])
dat
```

```
##     County Excess.PNR
## 9   BLADEN 0.07854424
## 78 ROBESON 0.07576120
```

Problem 4

```
data.frame(County = voters[which(weights == 0),'County'],
           Unaccounted.Absentee = voters$AbsBal[which(weights == 0)]*dat$Excess.PNR)
```

```
##    County Unaccounted.Absentee
## 1  BLADEN              636.9938
## 2 ROBESON             1217.4067
```

Problem 5

Using a linear regression model built on on all counties that were not Bladen or Robeson and stepwise selection, with AIC as the criterion, a model to predict the PNR was composed of an intercept and the number of black voters. Confidence intervals of a model of all variables confirmed that this was the only statistically significant variable. Plots of the residuals showed that the residuals were mostly normally distributed aside from a couple of outliers which are our counties of question. Using this model, a 99% prediction interval for the PNR was computed for Bladen and Robeson county. The true PNR for these counties was far higher than the upper 99% prediction interval. About 11% of absentee ballots were not returned for each of these counties, and the upper 99% interval for these counties was about 3.5%. So over 7.5% of the ballots missing from these counties are unexpected based on the model. This proportion times the number of absentees requested equates to about 637 unexplainable missing ballots for Bladen and 1217 for Boeson. These counties had many more missing ballots than other counties; moreover, the models were unable to explain these unexpected results based on the voting data. More investigation into these counties voting results is needed to explain the excess missing ballots.