HOMEWORK 1 - Proportion Not Returned

Rui Li

8/20/2020

# Import and Summarize the Dataset

#Import Proportion Not Returned.csv  
library(readr)  
dataset <- 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()  
## )

#Have a general idea about the dataset  
head(dataset)

## # A tibble: 6 x 12  
## County PNR Pop Rural MedAge Travel Hsgrad Collgrad MedInc Black Hisp  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 ALAMA~ 0.0116 163339 28.6 40 23.5 84.7 22.1 54263 19 11  
## 2 ALEXA~ 0.0044 38206 72.8 42 25.3 80.6 13.3 51893 5 4  
## 3 ALLEG~ 0.0082 11387 100 49 26.1 81 18.6 45244 1 9  
## 4 ANSON 0.0357 25460 78.5 40 28 80.4 9.5 42500 48 3  
## 5 ASHE 0.0082 27418 84.9 47 26.8 83.3 19.5 47509 1 5  
## 6 AVERY 0.0162 17953 88.8 44 22.3 79 19.8 46516 4 4  
## # ... with 1 more variable: AbsBal <dbl>

print(summary(dataset))

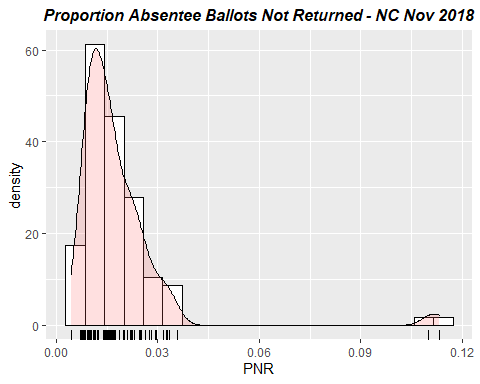
## County PNR Pop Rural   
## Length:100 Min. :0.00440 Min. : 4310 Min. : 1.10   
## Class :character 1st Qu.:0.01105 1st Qu.: 25102 1st Qu.: 42.35   
## Mode :character Median :0.01495 Median : 56534 Median : 62.25   
## Mean :0.01815 Mean : 102833 Mean : 61.21   
## 3rd Qu.:0.02133 3rd Qu.: 117801 3rd Qu.: 84.92   
## Max. :0.11310 Max. :1074596 Max. :100.00   
## MedAge Travel Hsgrad Collgrad   
## Min. :26.00 Min. :19.30 Min. :72.00 Min. : 8.20   
## 1st Qu.:40.00 1st Qu.:22.80 1st Qu.:80.28 1st Qu.:14.45   
## Median :42.00 Median :24.35 Median :83.40 Median :18.80   
## Mean :41.96 Mean :24.94 Mean :83.30 Mean :20.68   
## 3rd Qu.:45.00 3rd Qu.:26.80 3rd Qu.:87.22 3rd Qu.:23.65   
## Max. :51.00 Max. :36.70 Max. :92.50 Max. :57.70   
## MedInc Black Hisp AbsBal   
## Min. :36958 Min. : 0.00 Min. : 1.00 Min. : 532   
## 1st Qu.:46459 1st Qu.: 5.00 1st Qu.: 3.75 1st Qu.: 4284   
## Median :51774 Median :18.50 Median : 6.00 Median : 9710   
## Mean :53305 Mean :20.43 Mean : 6.45 Mean : 21118   
## 3rd Qu.:58652 3rd Qu.:32.25 3rd Qu.: 8.25 3rd Qu.: 21208   
## Max. :88887 Max. :62.00 Max. :21.00 Max. :225409

# Find the optimal model to predict PNR

#Summarize the PNR variable  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.6.3

ggplot(dataset, aes(x=PNR)) +   
 geom\_histogram(aes(y=..density..), bins = 20,fill = "white", col = "black") +   
 geom\_density(alpha=.2, fill="#FF6666") +  
 geom\_rug() +  
 labs(title = 'Proportion Absentee Ballots Not Returned - NC Nov 2018') +   
 theme(plot.title = element\_text(hjust = 0.5, size=12, face="bold.italic"))



#Omitting Bladen and Robeson counties with weights  
wts = dataset$PNR<0.1  
#Extract PNR and other predicted variables  
pnr.df = dataset[,2:11]  
#Build up full model  
full.lm = lm(formula = PNR ~ ., data = pnr.df, weights=as.numeric(wts))  
print(summary(full.lm))

##   
## Call:  
## lm(formula = PNR ~ ., data = pnr.df, weights = as.numeric(wts))  
##   
## Weighted Residuals:  
## Min 1Q Median 3Q Max   
## -0.011413 -0.003979 -0.001140 0.003621 0.018394   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.031e-02 2.230e-02 -1.359 0.177628   
## Pop 9.990e-09 5.701e-09 1.752 0.083198 .   
## Rural 4.073e-05 4.252e-05 0.958 0.340743   
## MedAge -1.727e-04 1.796e-04 -0.962 0.338798   
## Travel 2.161e-04 2.954e-04 0.732 0.466414   
## Hsgrad 5.177e-04 2.585e-04 2.003 0.048266 \*   
## Collgrad -1.196e-04 1.753e-04 -0.682 0.496791   
## MedInc -1.397e-08 1.692e-07 -0.083 0.934397   
## Black 1.825e-04 4.662e-05 3.913 0.000179 \*\*\*  
## Hisp 1.969e-04 2.025e-04 0.973 0.333428   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.006275 on 88 degrees of freedom  
## Multiple R-squared: 0.2739, Adjusted R-squared: 0.1996   
## F-statistic: 3.688 on 9 and 88 DF, p-value: 0.0005898

#Apply backward selection model  
full.backward = step(full.lm, direction = "backward")

## Start: AIC=-1003.02  
## PNR ~ Pop + Rural + MedAge + Travel + Hsgrad + Collgrad + MedInc +   
## Black + Hisp  
##   
## Df Sum of Sq RSS AIC  
## - MedInc 1 0.00000027 0.0034653 -1009.01  
## - Collgrad 1 0.00001833 0.0034834 -1008.49  
## - Travel 1 0.00002107 0.0034861 -1008.41  
## - Rural 1 0.00003613 0.0035012 -1007.98  
## - MedAge 1 0.00003642 0.0035014 -1007.97  
## - Hisp 1 0.00003725 0.0035023 -1007.95  
## <none> 0.0034650 -1007.02  
## - Pop 1 0.00012091 0.0035859 -1005.59  
## - Hsgrad 1 0.00015796 0.0036230 -1004.56  
## - Black 1 0.00060297 0.0040680 -992.98  
##   
## Step: AIC=-1005.01  
## PNR ~ Pop + Rural + MedAge + Travel + Hsgrad + Collgrad + Black +   
## Hisp  
##   
## Df Sum of Sq RSS AIC  
## - Travel 1 0.00003364 0.0034989 -1010.05  
## - MedAge 1 0.00003622 0.0035015 -1009.97  
## - Hisp 1 0.00003713 0.0035024 -1009.95  
## - Collgrad 1 0.00003877 0.0035041 -1009.90  
## - Rural 1 0.00004073 0.0035060 -1009.84  
## <none> 0.0034653 -1009.01  
## - Pop 1 0.00012087 0.0035862 -1007.58  
## - Hsgrad 1 0.00016061 0.0036259 -1006.48  
## - Black 1 0.00063826 0.0041036 -994.11  
##   
## Step: AIC=-1006.05  
## PNR ~ Pop + Rural + MedAge + Hsgrad + Collgrad + Black + Hisp  
##   
## Df Sum of Sq RSS AIC  
## - MedAge 1 0.00003743 0.0035364 -1010.98  
## - Hisp 1 0.00004405 0.0035430 -1010.80  
## - Collgrad 1 0.00004682 0.0035458 -1010.72  
## <none> 0.0034989 -1010.05  
## - Rural 1 0.00010637 0.0036053 -1009.05  
## - Pop 1 0.00015082 0.0036498 -1007.83  
## - Hsgrad 1 0.00019829 0.0036972 -1006.53  
## - Black 1 0.00068003 0.0041790 -994.29  
##   
## Step: AIC=-1006.98  
## PNR ~ Pop + Rural + Hsgrad + Collgrad + Black + Hisp  
##   
## Df Sum of Sq RSS AIC  
## - Collgrad 1 0.00005003 0.0035864 -1011.58  
## <none> 0.0035364 -1010.98  
## - Rural 1 0.00007514 0.0036115 -1010.88  
## - Hisp 1 0.00008153 0.0036179 -1010.70  
## - Pop 1 0.00015908 0.0036954 -1008.58  
## - Hsgrad 1 0.00020594 0.0037423 -1007.32  
## - Black 1 0.00078865 0.0043250 -992.85  
##   
## Step: AIC=-1007.58  
## PNR ~ Pop + Rural + Hsgrad + Black + Hisp  
##   
## Df Sum of Sq RSS AIC  
## - Hisp 1 0.00006453 0.0036509 -1011.79  
## <none> 0.0035864 -1011.58  
## - Rural 1 0.00008404 0.0036704 -1011.26  
## - Pop 1 0.00011727 0.0037037 -1010.36  
## - Hsgrad 1 0.00016430 0.0037507 -1009.10  
## - Black 1 0.00085991 0.0044463 -992.09  
##   
## Step: AIC=-1007.79  
## PNR ~ Pop + Rural + Hsgrad + Black  
##   
## Df Sum of Sq RSS AIC  
## - Rural 1 0.00005063 0.0037015 -1012.42  
## <none> 0.0036509 -1011.79  
## - Hsgrad 1 0.00011645 0.0037674 -1010.65  
## - Pop 1 0.00016668 0.0038176 -1009.33  
## - Black 1 0.00080052 0.0044514 -993.97  
##   
## Step: AIC=-1008.42  
## PNR ~ Pop + Hsgrad + Black  
##   
## Df Sum of Sq RSS AIC  
## <none> 0.0037015 -1012.42  
## - Hsgrad 1 0.00007514 0.0037767 -1012.41  
## - Pop 1 0.00011641 0.0038180 -1011.32  
## - Black 1 0.00075157 0.0044531 -995.93

print(summary(full.backward))

##   
## Call:  
## lm(formula = PNR ~ Pop + Hsgrad + Black, data = pnr.df, weights = as.numeric(wts))  
##   
## Weighted Residuals:  
## Min 1Q Median 3Q Max   
## -0.013330 -0.004575 -0.001610 0.003681 0.019052   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.239e-03 1.391e-02 -0.520 0.6041   
## Pop 7.561e-09 4.398e-09 1.719 0.0888 .   
## Hsgrad 2.277e-04 1.648e-04 1.381 0.1704   
## Black 1.833e-04 4.195e-05 4.369 3.22e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.006275 on 94 degrees of freedom  
## Multiple R-squared: 0.2243, Adjusted R-squared: 0.1996   
## F-statistic: 9.061 on 3 and 94 DF, p-value: 2.498e-05

#Apply forward selection model  
full.forward <- step(lm(PNR ~ 1, data=pnr.df), list(upper=full.lm), direction='forward')

## 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  
## + 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  
## + 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

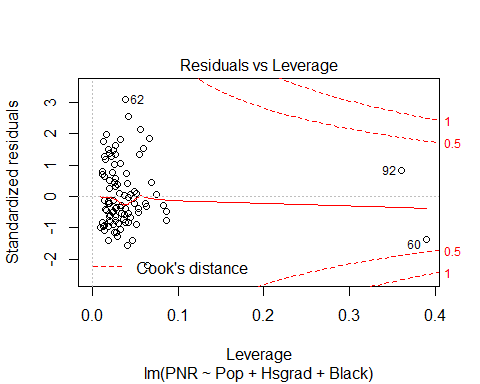
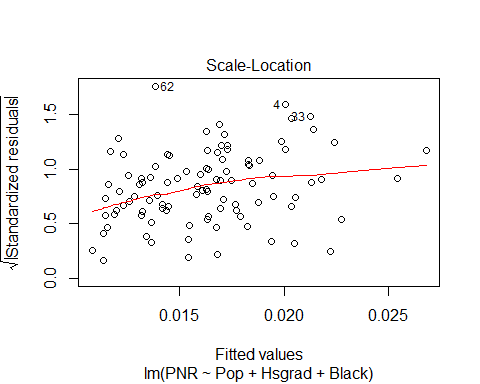
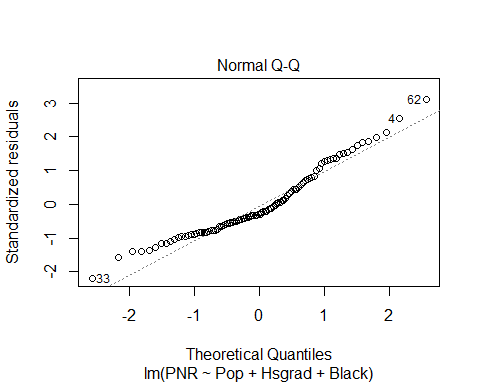
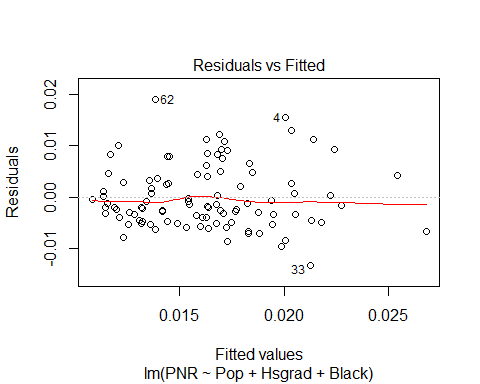
print(summary(full.forward))

##   
## Call:  
## lm(formula = PNR ~ Black, data = pnr.df)  
##   
## 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

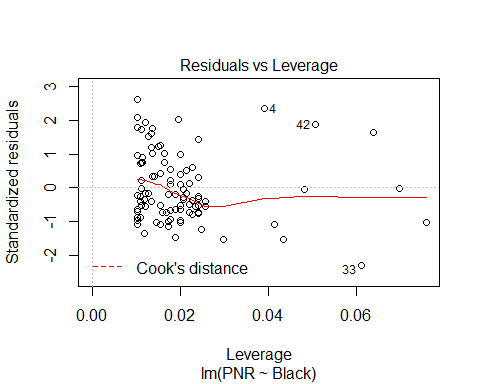
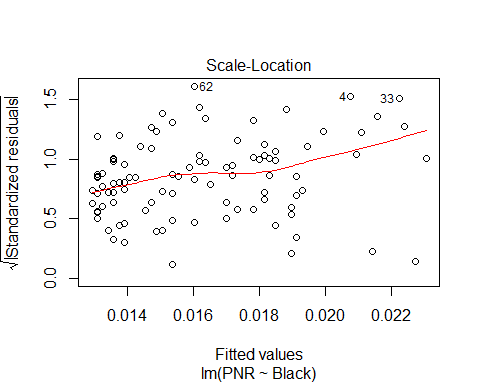
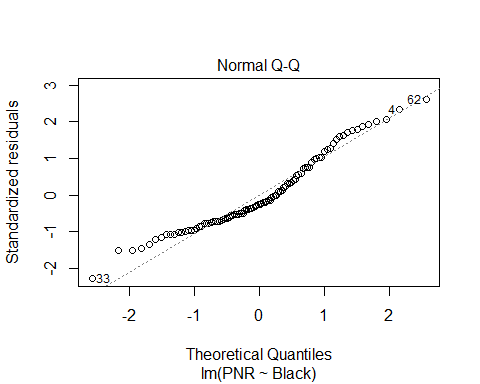
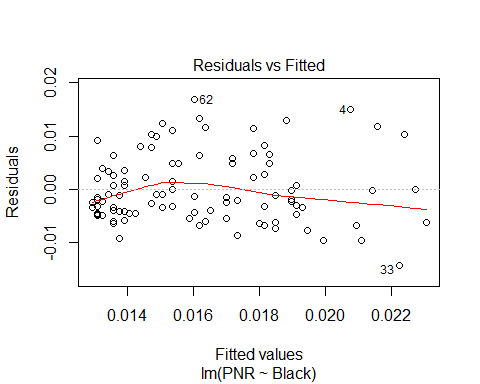
I apply both forward selection and backward selection to the full model, and get two different optimal models. The optimal model of backward selection is *PNR ~ Pop + Hsgrad + Black*, while the one of forward selection is *PNR ~ Black*.

In the following step, I will use diagnostics to assess various measures of fit and choose the better model.

#Optimal model of backward selection  
backward.lm = lm(formula = PNR ~ Pop + Hsgrad + Black, data = pnr.df, weights = as.numeric(wts))  
#Diagnostics of backward optimal model  
plot(backward.lm)



#Optimal model of forward selection  
forward.lm = lm(formula = PNR ~ Black, data = pnr.df, weights = as.numeric(wts))  
#Diagnostics of backward optimal model  
plot(forward.lm)



**Diagnosis Results:** From the *Residuals vs. fitted values* plots of both models, we can see that both scatterplots distribute randomly, with some exception outliers in the bottom right corner, but the backward model’s is more haphazard.

From the *Normal Probability plot* of both models, the errors of the backward model is more distributed normally, with a better straight line than the forward model.

From the *Scale Lotion* plots, both models show radom patterns.

From the *Cook’s distance* plots, we can see there are some outliers affecting both models.

Thus, according to the results above, I will pick the backward model,*PNR ~ Pop + Hsgrad + Black*, as my optimal model for the future steps.

# PNR prediction interval for Bladen and Robeson counties

#Choose backward model as the optimal model  
opt.lm = backward.lm  
#Predict based on 99% prediction interval  
opt.pre.99=predict(opt.lm,se.fit=T,interval='prediction',level=0.99,weights=1)

## Warning in predict.lm(opt.lm, se.fit = T, interval = "prediction", level = 0.99, : predictions on current data refer to \_future\_ responses

opt.pre.99$fit[c(9,78),]

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

The 99% prediction interval of Bladen and Robeson are *(0.00058, 0.03399)*, and *(-0.00131, 0.03242)*.

# Estimate excess PNR for Bladen and Robeson counties

excess.PNR=pnr.df[c(9,78),'PNR']-opt.pre.99$fit[c(9,78),'upr']  
excess.PNR

## PNR  
## 1 0.07910894  
## 2 0.07758184

The excee PNR of Bladen and Robeson are *0.079*, and *0.078* resepectively.

# Estimate the total number of absentee ballots that are unaccounted for

total=sum(excess.PNR\*dataset[c(9,78),'AbsBal'])  
print(total)

## [1] 1888.236

The total number of absentee ballots that are unaccounted for is 1888.236.

**Question** The actual number of votes by which Mark Harris was leading at the time the count was stopped was 905. The Harris campaign responded to the allegations by asserting that the number of potentially missing votes was very small and certainly less than 905. Does your analysis support that conclusion - why or why not?

print(905/(905+total))

## [1] 0.323997

**Answer** Compared to the predicted missing votes of 1888, 905 only represents a minor group of people. According to my analysis, the valid votes only represent 32.40% of the whole. Thus, my analysis does not support Harris’s conclusion.