STOR 590, FALL 2020
HOMEWORK 1, DUE FRIDAY, AUGUST 21.

Refer to this data file:

"Proportion Not Returned" datafile

This dataset refers to the proportion of absentee ballots that were not returned (out of all absentee ballots requested) in the 100 counties of North Carolina in the November 2018 election. There are 12 variables in the dataset, defined as follows:

County: Name of county
PNR: Proportion Not Returned
Pop: Population (2017)
Rural: Percent of population classified as rural
MedAge: Median age
Travel: Mean travel time to work
Hsgrad: Percent high school graduates
Collgrad: Percent college graduates
MedInc: Median income
Black: Percent black (non-hispanic)
Hisp: Percent hispanic
AbsBal: Absentee ballots applied for.

The problem

Two counties that are both part of District 9 (Bladen and Robeson - counties 9 and 78) faced allegations of voter fraud. Part of the alleged evidence to support this claim is that both counties showed an exceptionally high proportion of absentee ballots that were not returned. In brief, the allegation centers on the possibility that those ballots were stolen or otherwise misappropriated and used to increase the votes of Republican Mark Harris. The intent of this exercise is to investigate the statistical plausibility that such anomalous figures could have occurred by chance. This should take into account the influence of various factors such as population size, education level and racial proportions have on the handling of absentee ballots.

Within R, you can read this datafile using the "read.csv" command. For example,
X=read.csv('ProportionNotReturned.csv',header=T)
(you may need to insert a path before the file name) will load the data into a dataframe X. names(X) will then give you the names of the variables in X.

What I want you to do.

1. Omitting Bladen and Robeson counties, construct a regression model to predict PNR as a function of the other 9 numerical variables. You can use any standard methods of regression analysis that you prefer, but the intention is that you should use the "lm" command in R and use forward or

backward variable selection to determine the optimal model. Also use diagnostics to assess various measures of fit, such as whether the residuals appear to be normally distributed. (Note: Rather than delete Robeson and Bladen from the dataset, a better way to do it is to create a "weight" vector that has entries 0 in places 9 and 78, and a 1 in every other entry. Then use "weights" as an option in the lm command. You will find this works better when you try to use the "predict" (or "predict.lm") function in R.)

2. For each of Bladen and Robeson counties, find a *prediction interval* for the PNR, based on the results in the other 98 counties. You are free to experiment with different probability values, but I suggest calculating a 99% prediction interval.

3. Based on your answer to part 2, estimate a lower bound on the *excess PNR for Bladen and Robeson counties* that cannot be explain by natural variability. For instance, in Bladen county the actual PNR was 0.113; to quote a hypothetical example, if the upper bound of the 99% prediction interval for Bladen county was 0.042, then the excess PNR would be 0.071 (0.113 minus 0.042).

4. The numbers of absentee ballots requested in Bladen and Robeson counties were respectively 8,110 and 16,069. Combining this information with your answer to 3, estimate the *total number* of absentee ballots that are unaccounted for.

5. 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?

Your answer should be in the form of a report that summarizes your conclusions and supports them, where necessary, with relevant statistical analyses. You *may* use R Markdown for this, if you are familiar with R Markdown, but other forms of word processing (e.g. MS Word, Latex) are equally acceptable. Whatever format of report you adopt, it's important that you state your conclusions with clear verbal summaries and don't rely solely on output from R or R-Studio.

Hand in via the "Assignments" tab on sakai no later than **1:00 pm, Friday August 21.**

Notes 1: The variable "AbsBal" gives the number of absentee ballots requested in all 100 counties.

Notes 2: The quick way to compute prediction intervals in conjunction with the "lm" command in R is a command of the form

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

applied to the object "lm1" that you got from the "lm" command. However, the default version of this used the same weights as were used for the model fit. If you followed my earlier suggestion and defined weights through some command such as wts=as.numeric(Y$PNR<0.1), which gives weight 0 to Bladen and Robeson counties, the resulting prediction intervals will be infinite! The best way round this seems to be to amend the above "predict" command to

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

which resets all the weights to 1. From this, you can extract the correct prediction intervals for Bladen and Robeson counties.