ZHOU Kevin

Student ID : A0197122H

NUS - IE5101 - Applied Forecasting Methods

Professor : Chen Nan

Project 1 : US 2016 Presidential Primary Election

Table of Contents

2.1 Step 1: Simple Regression Model

2.1.1 model proposed

For the first simple regression, we will try to have a simple interpretable model. To ease the interpretation :

1- We will opt only for 5 variables

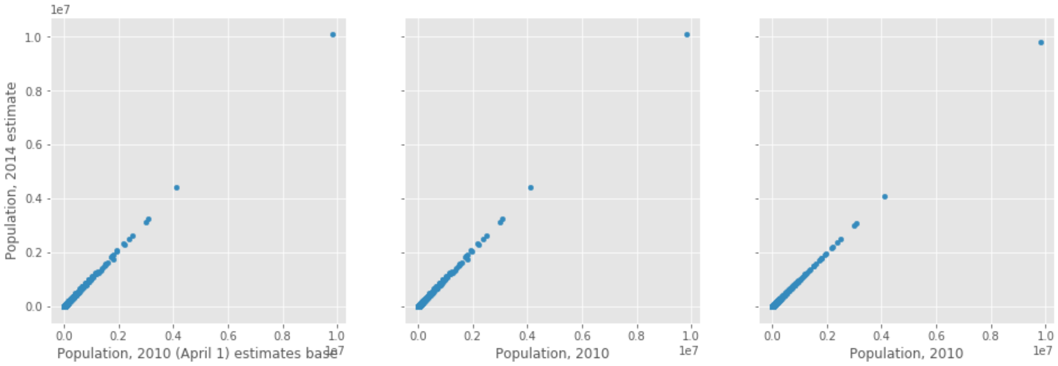


Illustration 1: Correlation between the different Population variables

2- The variable chosen should be as much non collinear as possible. We can easily suspect some level of collinearity between variables of the same category such as :

- **race related variables (“**White alone, percent, 2014”, “Black or African American alone, percent, 2014”, “White alone, not Hispanic or Latino, percent”, “Black-owned firms, percent, 2007”...)
- **population number related variables,**
- **or wealthiness related variables...**

This is particularly true for the population number across the years in the 3 scatter plots below where we observe a nearly perfect identity function when plotting against each other.

3- We choose the variables one by one with the maximum R squared value to have for the maximum explained variance (a simple linear regression with the variable). This is however just a simplification and would surely not yield the most accurate results.

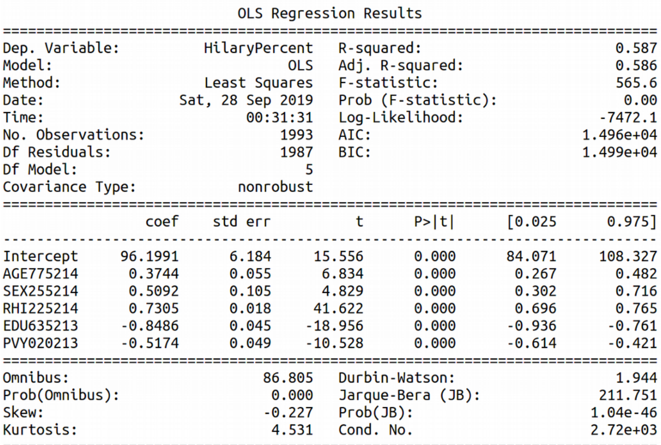
With the below 3 criteria: our model will use the following 5 features, as they will represent respectively, age, gender, race, education level and wealthiness of the county and have a level of non collinearity mutually.

1	Q("Persons 65 years and over, percent, 2014")	'AGE775214'
2	Q("Female persons, percent, 2014")	'SEX255214'
3	Q("Black or African American alone, percent, 2014")	'RHI225214'
4	Q("High school graduate or higher, percent of persons age 25+, 2009-2013")	'EDU635213'

5	Q("Persons below poverty level, percent, 2009-2013")	‘PVY020213'

2.1.2 Results Analysis

The OLS regression with the 5 variables above and a intercept value gives the following results.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:          HilaryPercent   R-squared:                       0.587
Model:                            OLS   Adj. R-squared:                  0.586
Method:                 Least Squares   F-statistic:                     565.6
Date:                Sat, 28 Sep 2019   Prob (F-statistic):               0.00
Time:                        00:31:31   Log-Likelihood:                -7472.1
No. Observations:                1993   AIC:                         1.496e+04
Df Residuals:                    1987   BIC:                         1.499e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     96.1991      6.184     15.556      0.000      84.071     108.327
AGE775214      0.3744      0.055      6.834      0.000       0.267       0.482
SEX255214      0.5092      0.105      4.829      0.000       0.302       0.716
RHI225214      0.7305      0.018     41.622      0.000       0.696       0.765
EDU635213     -0.8486      0.045    -18.956      0.000      -0.936      -0.761
PVY020213     -0.5174      0.049    -10.528      0.000      -0.614      -0.421
==============================================================================
Omnibus:                       86.805   Durbin-Watson:                   1.944
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              211.751
Skew:                          -0.227   Prob(JB):                     1.04e-46
Kurtosis:                       4.531   Cond. No.                     2.72e+03
==============================================================================
```

Illustration 2: OLS Regression results with training set (train-test split of 80% and 20%)

With 5 variables, the **R-squared** indicates that 58.7 % of the variation are explained by this model containing 5 variables.

All the variables’ **p-value** are close to zero and their associated **t-score** big meaning that they all are relevant in the model.

The big value of the F-statistic also goes along that way.

2.1.2 model Checking and Diagnosis

We will mainly check the validity of the below assumptions made by the ordinary least squared regression:

- Normality assumptions checks based on residuals
- Uncorrelated predictors assumptions
- Homoscedasticity assumption i.e. the constant variance is the predictors and the dependent variable.

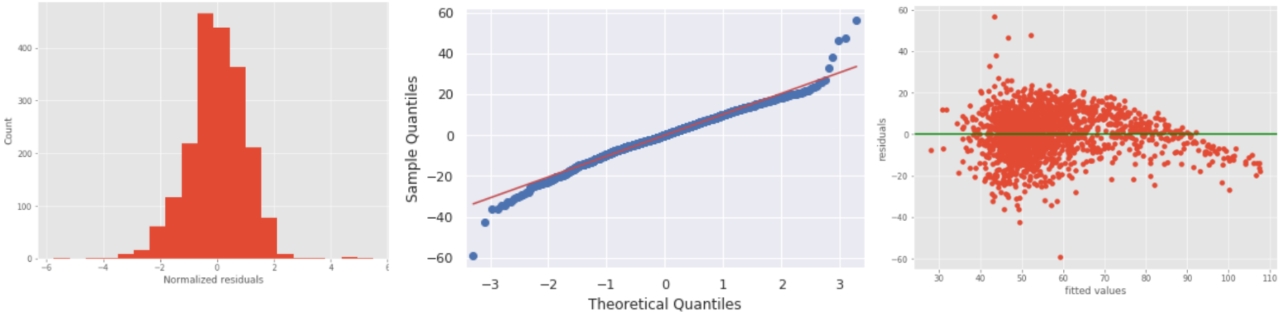


Illustration 3: (from left to right) histogram of normalized residuals, qq plot of the residuals, scatter plot of the residuals against the fitted values

The **quantile-quantile plot of the residuals** shows a rather close fit to the normal distribution in the middle. The **histogram plot** also display a pretty good normally distributed values of residuals.
However in the upper and lower tails, we notice departure from the fitted normal line making it a long-tail distribution. Either data does not satisfy normal assumption or the model is not sufficient enough to describe the dependent variable.

The **scatter plot of the residuals against the fitted values** shows more clearly some curving and also a slight non-constant variance in the distribution of residuals with smaller variance on higher end of fitted values. The homoscedasticity and normality assumptions seem not to be respected.

The partial plots of residuals against the 4 predictors ('AGE775214', 'SEX255214', 'EDU635213', 'PVY020213') don't show any particular violation of assumptions.

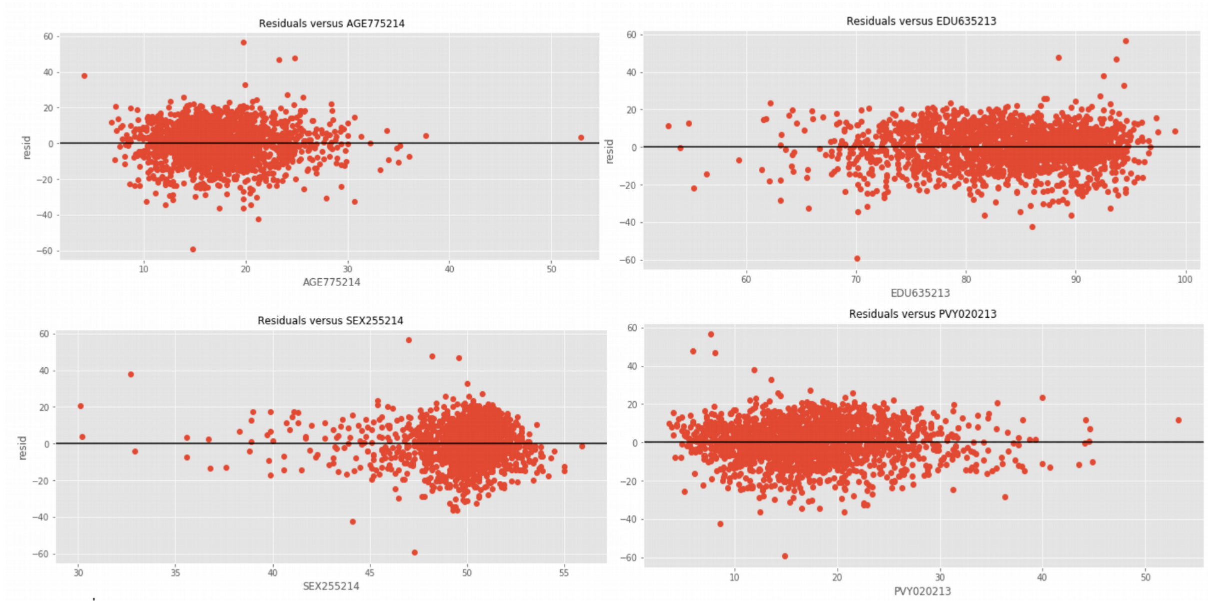


Illustration 4: Partial plots of residuals vs predictors : 'AGE775214', 'SEX255214', 'EDU635213', 'PVY020213'

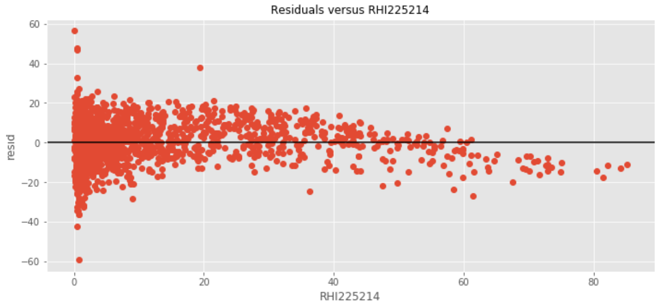


Illustration 5: Partial plots of residuals vs predictors ‘RHI225214’

The partial plot of residuals against predictor ‘RHI225214’ shows however some clear curvature.

We can apply a **square root transformation** to the predictor which is easier than the log transformation since the predictor variable contains values equal to 0.

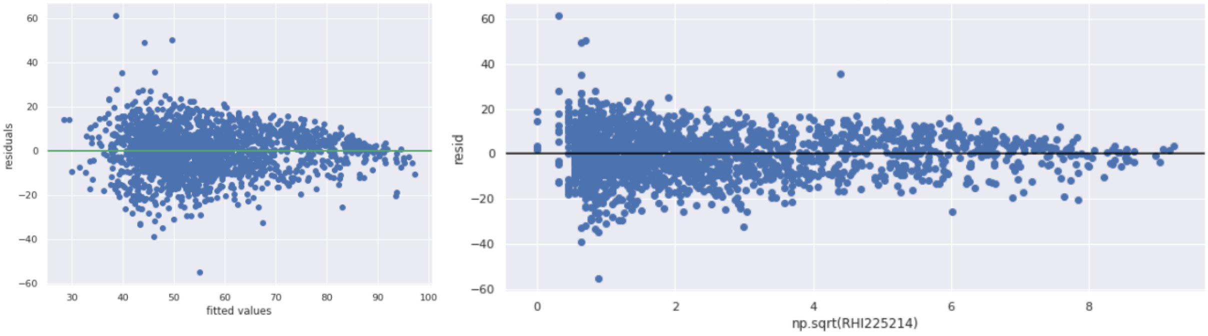


Illustration 6: the residuals plot against fitted values and against the predictor

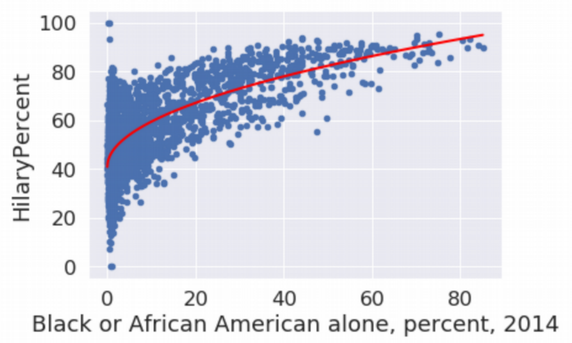


Illustration 7: Scatter plot of HilaryPercent against RHI225214

The model with the square root of RHI225214 instead of RHI225214 has indeed **removed the curvature** from both the residuals plot against fitted values and against the predictor. This is due to the non linearity of the predictor as shown in the scatter plot.

The variance explained has been improved to **R²=60.1%**. However we still observe the **non constant-variance** of the residuals as variance of residuals tends to decrease when the predictor increases.

Outliers

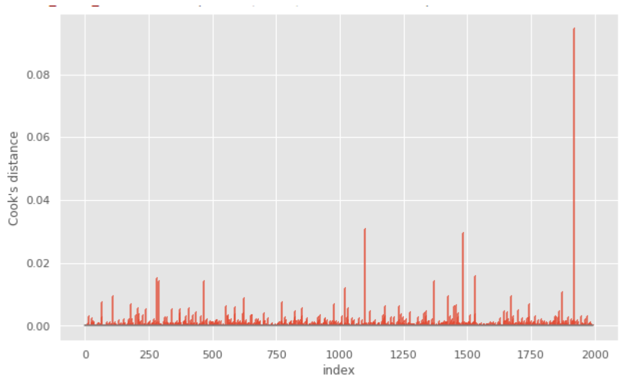


Illustration 8: Outlier identification with the Cook's distance

Assessing the magnitudes of the Cook's distance of the observations does not yield any particular results. Based on this There are no points with high enough values to be considered to be a possible problem (all distance are inferior to 0.1)

The influence plots gives however points that may need our attentions. There are quite a number of **low leverage influential points** that may distort our regression model in the blue rectangles below.

We can also distinguish some **high leverage and high influential points** that may cause even more issues displayed below in green rectangles.

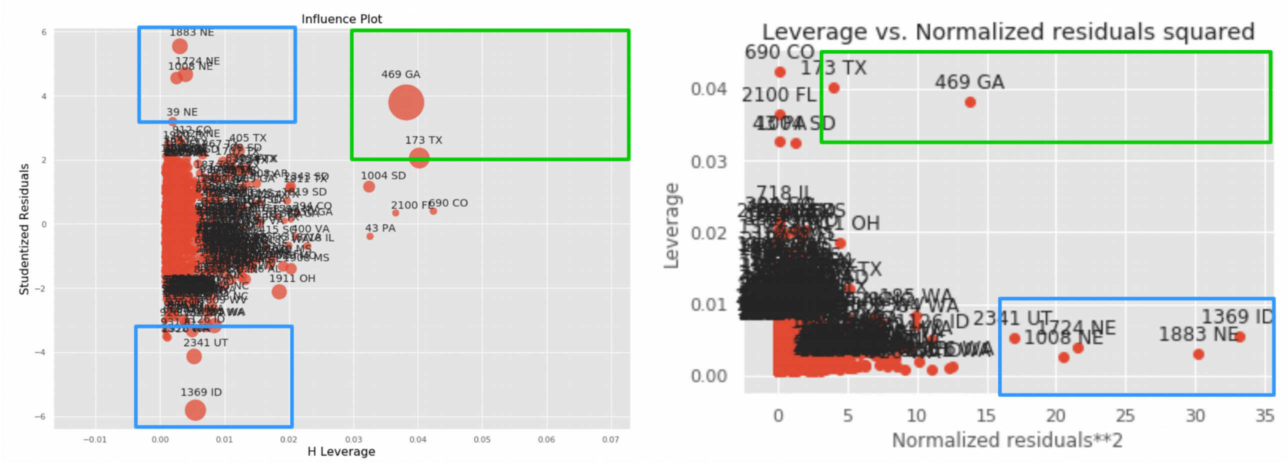


Illustration 9: Outliers identification by comparing values of leverage and residual

Those points have typically low number of population. In the next model we will build, we may want to include the number of population/voters in our model.

Collinearity Diagnosis

The collinearity level between the predictors of our model could be assessed by computing the VIF (Variance inflation factor). We can claim the multicollinearity problem is neglectable in our modal since they all have a VIF below 5.

```
VIF for Intercept:719.0136454574009
VIF for AGE775214:1.0541349259195134
VIF for SEX255214:1.0577333771388462
VIF for RHI225214:1.3271481179783713
VIF for EDU635213:1.7933280510573104
VIF for PVY020213:1.9228507223070948
```

To conclude, the current modal still have a quite low **R^2 square. Lots of variance is not explained by the variables chosen as it would be logic to think so since there is more than 40 unexploited variables.**

2.2 Full Regression Model

The metrics used to evaluate the model will be the weighted mean squared errors WMSE.

To optimize our modal parameters according to this error function, we will opt for the weighted least square regression. (sm.WLS of the statsmodal library)

I implemented a cross validation as the modal validation technique for assessing the accuracy of our modal.

2.2.1 Full Run

A run of every combination of features (52 features in total) would be computationally too costly to execute. We would have to fit the linear regression n times with n = $\sum_{1\leq n\leq 52}\binom{52}{n}$ = 4.5 10^{15} combinations.

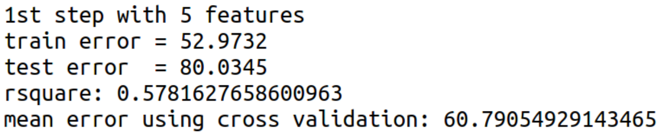
My computer computes roughly R=10 linear regressions per seconds. It would have taken n/R=4.5 10^{14} seconds to perform the full run. This is only considering only untransformed combination of features and only considering the principal effects of each features.

2.2.1 Greedy search

With the packages provided in the case study in class, I firstly performed a forward and backward selection search based on the the AIC neglecting the interaction effects.

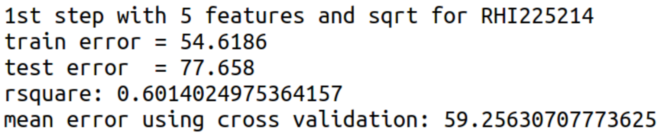
The train and test error are computed based on a 80-20 % split. Are also calculate the R square and the WSME using a 10-fold cross validation.

1



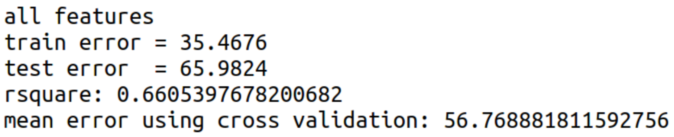
```
1st step with 5 features
train error = 52.9732
test error  = 80.0345
rsquare: 0.5781627658600963
mean error using cross validation: 60.79054929143465
```

2



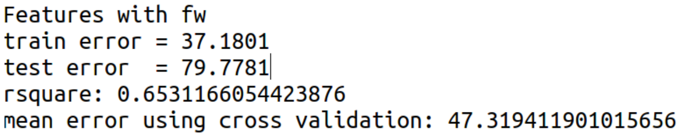
```
1st step with 5 features and sqrt for RHI225214
train error = 54.6186
test error  = 77.658
rsquare: 0.6014024975364157
mean error using cross validation: 59.25630707773625
```

3



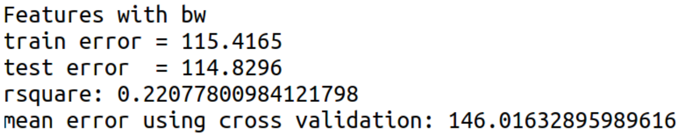
```
all features
train error = 35.4676
test error  = 65.9824
rsquare: 0.6605397678200682
mean error using cross validation: 56.768881811592756
```

4



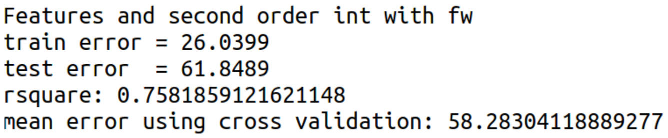
```
Features with fw
train error = 37.1801
test error  = 79.7781
rsquare: 0.6531166054423876
mean error using cross validation: 47.319411901015656
```

5



```
Features with bw
train error = 115.4165
test error  = 114.8296
rsquare: 0.22077800984121798
mean error using cross validation: 146.01632895989616
```

6



```
Features and second order int with fw
train error = 26.0399
test error  = 61.8489
rsquare: 0.7581859121621148
mean error using cross validation: 58.28304118889277
```

We can notice that all models trained seems to have overfitting patterns. The test errors obtained is much bigger than the training error. The best results comes out to be the modal resulted from the forward selection based on the cross validation error.

Applying a Linear regression weighted (WLS) by the number of democratic voters (Clinton voters + Sander voters) yields much better results on the modal obtained with forward selection: The mean cv error is reduced to 35%.

Conclusion

Due to time constraints, I did not have time to run the greedy search on WLS. Further works would include: running the greedy search on WLS by using different evaluation metrics and an in depth diagnosis of the modal as done in terms of outliers, collinearity, and other linear regression assumptions.