ANLY 699: Assignment 3

Code

Subhash Pemmaraju

06/07/2020

Summary output of variables as is

Hide

```
summary(merged_data[,c(6,9,13,16,18,19,20,21,22,23,24,25,26,27)])
```

```
##      Value           yrs_plr        perc_obese    perc_smokers
##  Min.   :  0.00   Min.   : 2731   Min.   :12.0   Min.   : 6.00
##  1st Qu.:  7.00   1st Qu.: 6792   1st Qu.:29.0   1st Qu.:15.00
##  Median : 18.00   Median : 8309   Median :33.0   Median :17.00
##  Mean   : 25.94   Mean   : 8572   Mean   :32.9   Mean   :17.45
##  3rd Qu.: 38.00   3rd Qu.:10076   3rd Qu.:37.0   3rd Qu.:20.00
##  Max.   :100.00   Max.   :29138   Max.   :58.0   Max.   :41.00
##                   NA's   :295     NA's   :2      NA's   :2
##  perc_uninsured   perc_college     perc_unemp     perc_child_pov
##  Min.   : 2.00   Min.   : 15.0   Min.   : 1.300   Min.   : 3.00
##  1st Qu.: 7.00   1st Qu.: 50.0   1st Qu.: 3.100   1st Qu.:15.00
##  Median :11.00   Median : 58.0   Median : 3.900   Median :20.00
##  Mean   :11.47   Mean   : 57.9   Mean   : 4.126   Mean   :21.14
##  3rd Qu.:14.00   3rd Qu.: 66.0   3rd Qu.: 4.800   3rd Qu.:26.00
##  Max.   :34.00   Max.   :100.0   Max.   :18.100   Max.   :68.00
##  NA's   :3       NA's   :2       NA's   :3        NA's   :3
##  perc_diabetes   median_income      perc_65up       perc_black
##  Min.   : 2.00   Min.   : 25385   Min.   : 4.80   Min.   : 0.000
##  1st Qu.: 9.00   1st Qu.: 40002   1st Qu.:16.27   1st Qu.: 0.700
##  Median :12.00   Median : 46843   Median :18.90   Median : 2.200
##  Mean   :12.17   Mean   : 51734   Mean   :19.28   Mean   : 8.999
##  3rd Qu.:15.00   3rd Qu.: 59350   3rd Qu.:21.80   3rd Qu.:10.225
##  Max.   :34.00   Max.   :125933   Max.   :57.60   Max.   :85.400
##  NA's   :2       NA's   :2595     NA's   :2       NA's   :2
##   perc_female     perc_18less
##  Min.   :26.80   Min.   : 0.00
##  1st Qu.:49.40   1st Qu.:20.00
##  Median :50.30   Median :22.10
##  Mean   :49.89   Mean   :22.06
##  3rd Qu.:51.00   3rd Qu.:23.90
##  Max.   :56.90   Max.   :42.00
##  NA's   :2       NA's   :2
```

Summary output of variables with normalization

Hide

```
merged_data1<- scale(merged_data[,c(6,9,13,16,18,19,20,21,22,23,24,25,26,27)])
summary(merged_data1)
```

```
##      Value            yrs_plr          perc_obese        perc_smokers
##  Min.   :-1.0911   Min.   :-2.2720   Min.   :-3.83225   Min.   :-3.1950
##  1st Qu.:-0.7966   1st Qu.:-0.6923   1st Qu.:-0.71549   1st Qu.:-0.6843
##  Median :-0.3339   Median :-0.1024   Median : 0.01787   Median :-0.1263
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000
##  3rd Qu.: 0.5075   3rd Qu.: 0.5851   3rd Qu.: 0.75122   3rd Qu.: 0.7106
##  Max.   : 3.1159   Max.   : 7.9993   Max.   : 4.60134   Max.   : 6.5690
##                    NA's   :295       NA's   :2          NA's   :2
##  perc_uninsured     perc_college         perc_unemp      perc_child_pov
##  Min.   :-1.8399   Min.   :-3.628898   Min.   :-1.9201   Min.   :-2.0419
##  1st Qu.:-0.8688   1st Qu.:-0.668259   1st Qu.:-0.6973   1st Qu.:-0.6911
##  Median :-0.0920   Median : 0.008459   Median :-0.1538   Median :-0.1282
##  Mean   : 0.0000   Mean   : 0.000000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.4906   3rd Qu.: 0.685177   3rd Qu.: 0.4576   3rd Qu.: 0.5472
##  Max.   : 4.3748   Max.   : 3.561226   Max.   : 9.4930   Max.   : 5.2750
##  NA's   :3         NA's   :2           NA's   :3         NA's   :3
##  perc_diabetes      median_income       perc_65up         perc_black
##  Min.   :-2.50305   Min.   :-1.6335   Min.   :-3.08157   Min.   :-0.6293
##  1st Qu.:-0.78086   1st Qu.:-0.7274   1st Qu.:-0.63872   1st Qu.:-0.5803
##  Median :-0.04278   Median :-0.3032   Median :-0.07989   Median :-0.4754
##  Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000
##  3rd Qu.: 0.69530   3rd Qu.: 0.4721   3rd Qu.: 0.53747   3rd Qu.: 0.0857
##  Max.   : 5.36982   Max.   : 4.5998   Max.   : 8.15875   Max.   : 5.3422
##  NA's   :2          NA's   :2595      NA's   :2          NA's   :2
##   perc_female        perc_18less
##  Min.   :-10.1097   Min.   :-6.38615
##  1st Qu.: -0.2129   1st Qu.:-0.59657
##  Median :  0.1812   Median : 0.01134
##  Mean   :  0.0000   Mean   : 0.00000
##  3rd Qu.:  0.4877   3rd Qu.: 0.53240
##  Max.   :  3.0714   Max.   : 5.77197
##  NA's   :2          NA's   :2
```

Hide

```
#Multiple regression model with demographic variables
model1<-lm(yrs_plr~Value+perc_65up+perc_18less+perc_female+perc_black, data=data.frame(merged_da
ta1))
summary(model1)
```

```
##
## Call:
## lm.default(formula = yrs_plr ~ Value + perc_65up + perc_18less +
##     perc_female + perc_black, data = data.frame(merged_data1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2731 -0.5625 -0.0851  0.4453  7.4146
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.01216    0.01653   0.736   0.4620
## Value       -0.18337    0.01772 -10.348   <2e-16 ***
## perc_65up    0.37123    0.02406  15.431   <2e-16 ***
## perc_18less  0.31109    0.02297  13.544   <2e-16 ***
## perc_female -0.04152    0.01820  -2.282   0.0226 *
## perc_black   0.37161    0.01709  21.750   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8734 on 2841 degrees of freedom
##   (295 observations deleted due to missingness)
## Multiple R-squared:  0.2384, Adjusted R-squared:  0.2371
## F-statistic: 177.9 on 5 and 2841 DF,  p-value: < 2.2e-16
```

Regression Analysis: Model with demographic variables

There is a clear statistically significant negative correlation of -0.18337 between Access to parks and years of potential life lost. What this means is that for every 1 unit increase in access to public parks, there is a 0.18 unit reduction in years of potential life lost. Furthermore, we can see that it is influenced by demographic variables. Higher is the percentage of population above 65 or below 18, more is the years of potential life lost. More is the percentage of black population, more is the potential life lost. However, as % of females in the population improves, then years of potential life improves. Hence the negative correlation.

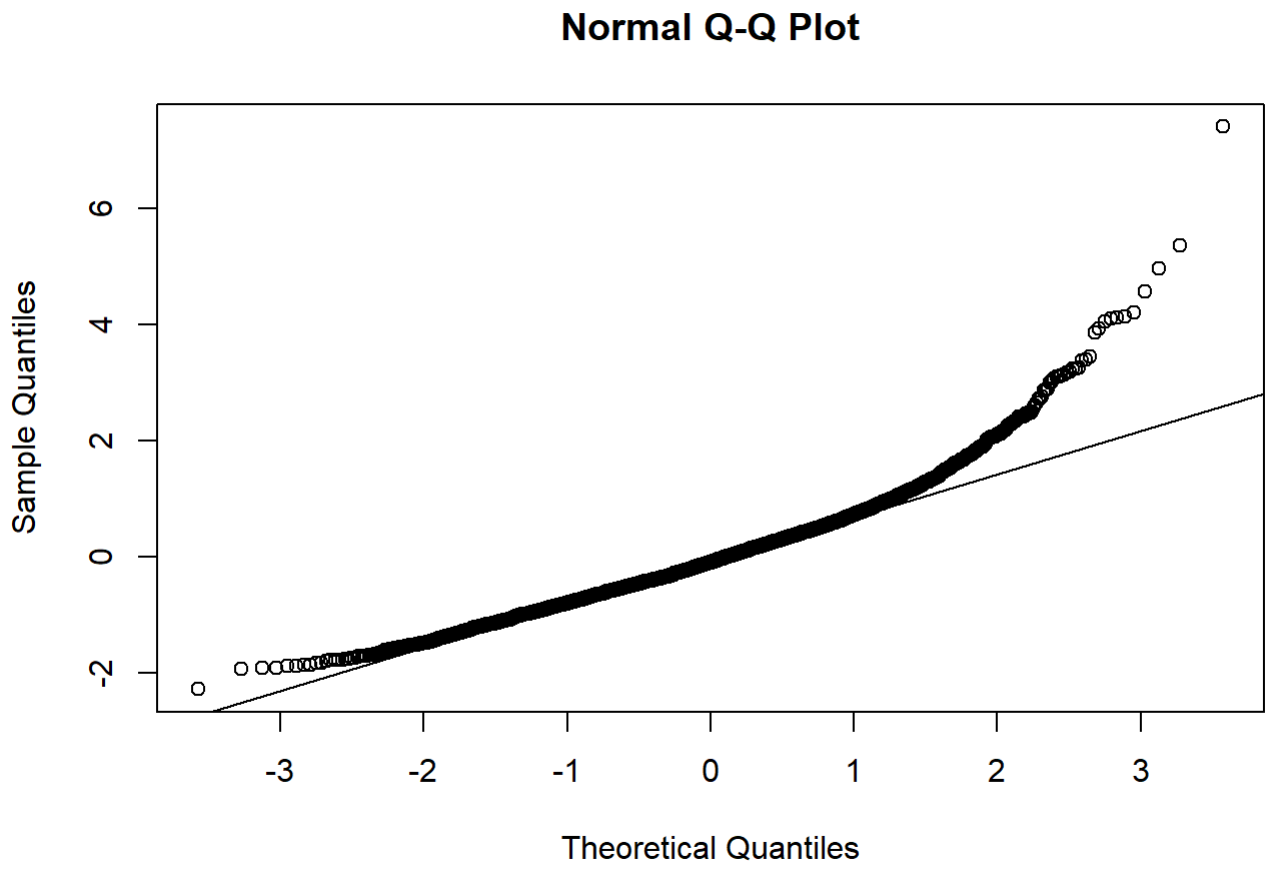
Hide

```
#Multiple regression model with demographic variables
vif(model1)
```

```
##       Value   perc_65up perc_18less perc_female  perc_black
##    1.074030    1.864459    1.864337    1.137498    1.141879
```

Hide

```
qqnorm(model1$residuals)
qqline(model1$residuals)
```



Hide

```
shapiro.test(model1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model1$residuals
## W = 0.94276, p-value < 2.2e-16
```

Regression Diagnostics

The VIF has all values low <<5. Therefore multicollinearity is not a problem. The QQ plot shows some evidence of non-normality. The Shapiro-Wilkes test confirms that non-normality is a strong problem. The way to proceed in this case would be to take log variables or add square of independent variables as an additional correlate.

Hide

```
#Multiple regression model with demographic variables and pre-existing health conditions
model2<-lm(yrs_plr~Value+perc_65up+perc_18less+perc_female+perc_black+perc_obese+perc_smokers+perc_diabetes, data=data.frame(merged_data1))

summary(model2)
```

```
##
## Call:
## lm.default(formula = yrs_plr ~ Value + perc_65up + perc_18less +
##     perc_female + perc_black + perc_obese + perc_smokers + perc_diabetes,
##     data = data.frame(merged_data1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8558 -0.3921 -0.0475  0.3232  4.2731
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.0354282  0.0115329  -3.072 0.002147 **
## Value          0.0480473  0.0134171   3.581 0.000348 ***
## perc_65up      0.3197813  0.0172797  18.506  < 2e-16 ***
## perc_18less    0.2611773  0.0163318  15.992  < 2e-16 ***
## perc_female   -0.0007474  0.0127250  -0.059 0.953165
## perc_black     0.1786696  0.0128331  13.923  < 2e-16 ***
## perc_obese     0.0009953  0.0151306   0.066 0.947557
## perc_smokers   0.6235873  0.0139920  44.568  < 2e-16 ***
## perc_diabetes  0.1408816  0.0148484   9.488  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6073 on 2838 degrees of freedom
##   (295 observations deleted due to missingness)
## Multiple R-squared:  0.6322, Adjusted R-squared:  0.6312
## F-statistic: 609.7 on 8 and 2838 DF,  p-value: < 2.2e-16
```

Regression Analysis: Model with demographic variables and variables for pre-existing health conditions

From the regression analysis, we can see that in this case Years of potential lives lost and access to parks are positively correlated which does not make intuitive sense. What can be seen however, is that pre-existing health conditions like smoking and diabetes are strongly correlated and increase years of potential life lost. Percentage of obese population is not statistically significant. This is possibly because obesity and diabetes are strongly corrrelated. The model may suffer from multicolliearity. However, the VIF does not provide any such indication.

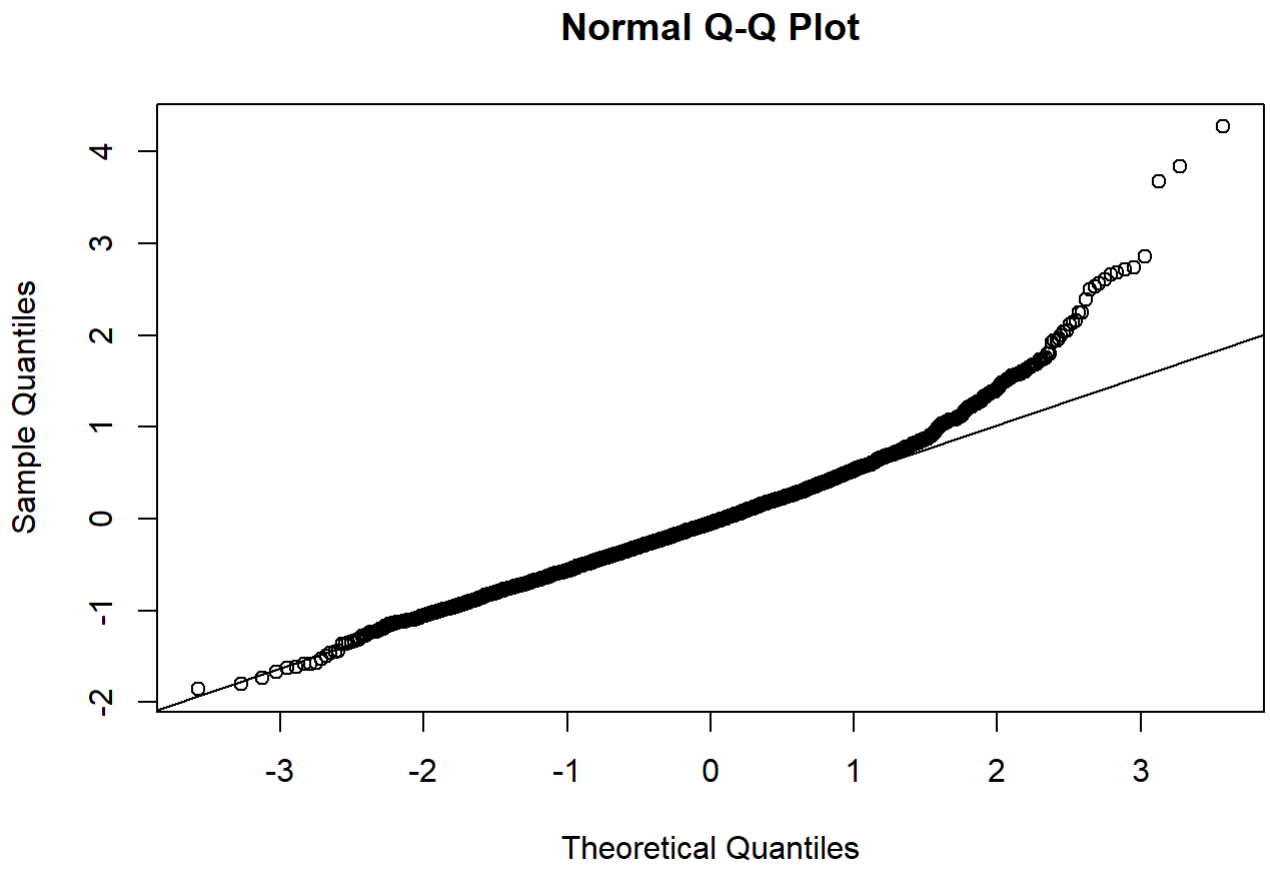
Hide

```
vif(model2)
```

```
##         Value     perc_65up   perc_18less   perc_female    perc_black
##      1.273452      1.989440      1.949630      1.150768      1.332491
##    perc_obese  perc_smokers perc_diabetes
##      1.752844      1.520025      1.679328
```

Hide

```
qqnorm(model2$residuals)
qqline(model2$residuals)
```



Hide

```
shapiro.test(model2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model2$residuals
## W = 0.96234, p-value < 2.2e-16
```

Regression Diagnostics

Strong evidence of non-normality in the data based on Q-Q plot and Shapiro-Wilkes test. Further refinement of the model is needed.

Hide

```
anova(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: yrs_plr ~ Value + perc_65up + perc_18less + perc_female + perc_black
## Model 2: yrs_plr ~ Value + perc_65up + perc_18less + perc_female + perc_black +
##     perc_obese + perc_smokers + perc_diabetes
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   2841 2167.4
## 2   2838 1046.8  3    1120.7 1012.8 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparison of two models

Comparison of two models indicates that model2 with pre-existing conditions added is superior to the other model and has incremental predictive power over it.