Results & Conclusions

Zayed Shahjahan

12/9/2020

Our final model is as follows:

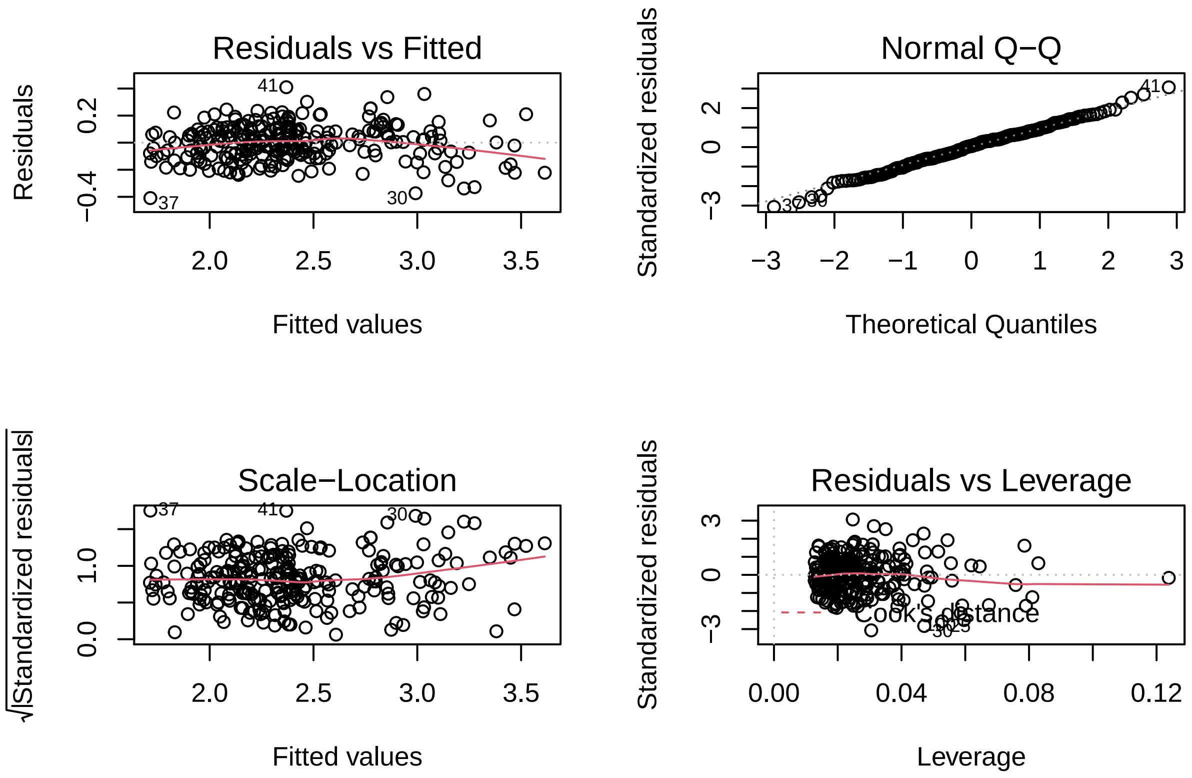
```
##
## Call:
## lm(formula = sqrt(adjobese) ~ phys + nutri + vrich + region,
##     data = df2[-c(256, 40), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40919 -0.08231  0.00711  0.08606  0.40955
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.285745   0.039243  32.764  < 2e-16 ***
## phys             0.060109   0.007344   8.185 1.45e-14 ***
## nutri            0.117194   0.008526  13.745  < 2e-16 ***
## vrich           -0.213661   0.050507  -4.230 3.29e-05 ***
## regionNortheast -0.046626   0.025437  -1.833  0.06800 .
## regionSouth      0.062854   0.022773   2.760  0.00621 **
## regionWestern   -0.116762   0.027866  -4.190 3.88e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1354 on 247 degrees of freedom
## Multiple R-squared:  0.9011, Adjusted R-squared:  0.8987
## F-statistic: 375.1 on 6 and 247 DF,  p-value: < 2.2e-16
```

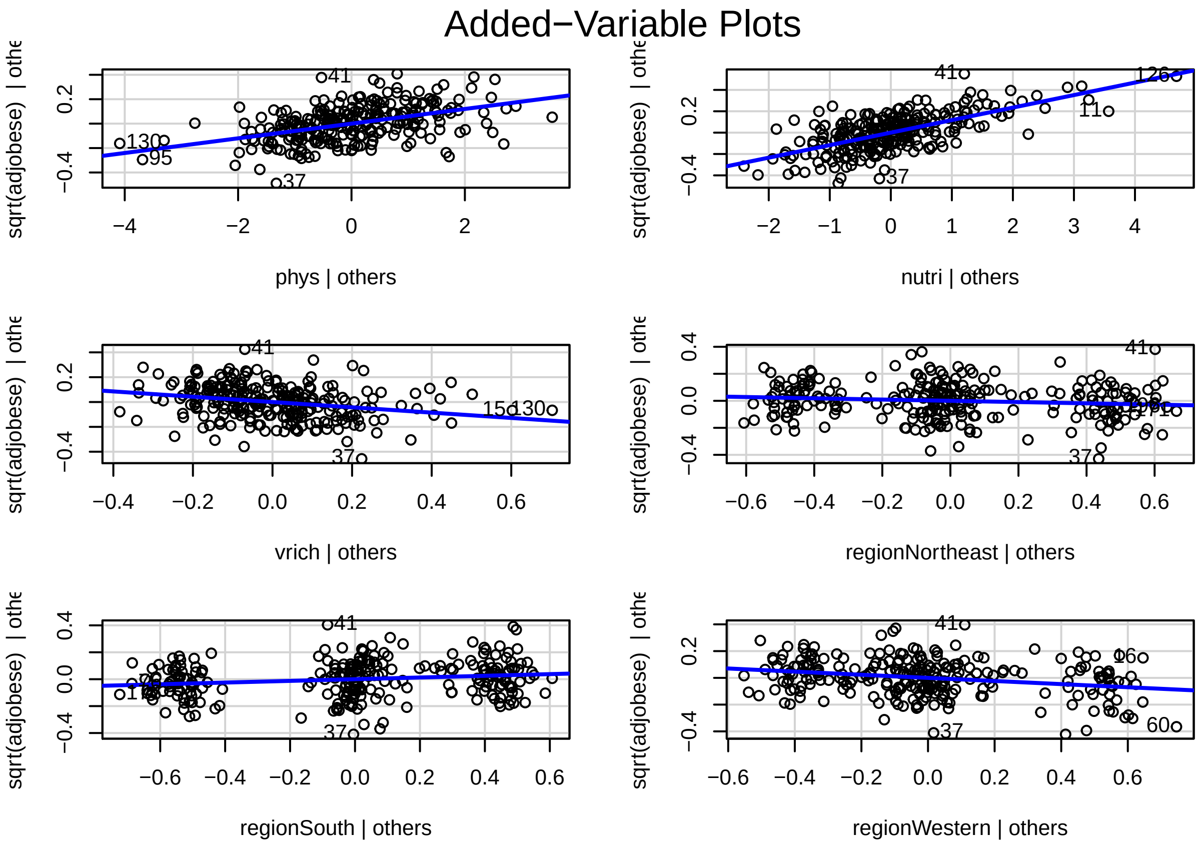
The VIF is as follows:

Again we have more evidence that income and physical activity move together. Income was a continuous variable we would have been able to work with this better.

```
##            phys           nutri           vrich regionNortheast     regionSouth
##       10.450252        4.668810        5.589855        1.599166        1.600491
##   regionWestern
##        1.568901
```

The residuals are we expect them after removing the 2 influential observations


Residuals vs Fitted
Residuals
Fitted values
−0.4
0.2
2.0
2.5
3.0
3.5
41
37
30
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−3
0
2
−3
−2
−1
0
1
2
3
41
37
30
Scale−Location
√|Standardized residuals|
Fitted values
0.0
1.0
2.0
2.5
3.0
3.5
37
41
30
Residuals vs Leverage
Standardized residuals
Leverage
−3
0
3
0.00
0.04
0.08
0.12
Cook's distance
30
25



From the Added variable plots we see that on their own, Physical activity and nutrition do well in fitting the data.

We have also left out observations, 256 and 40.

Our Final model leads us to the following conclusions:

Physical activity is a factor, but due to its Variance inflation factor, we cannot trust the coefficient on this variable. And even if this was taken into account the practical significance of this is questionnable, especially when compared to the other variables.

Without a doubt lack nutrition and wealth are statistically and practically significant in the regression model. However it is worth noting that in the original box-and-whisker plot, (see Data Description) we saw a different relationship between Income brackets and Obesity. In that context, as incomes increased, obesity decreased until the income bracket was the largest. In that setting obesity increased again. However, in our final model we see that the relationship between Obesity and the highest income bracket is inverted. Which is very strange. It is expected in the context of the US. It is unexpected given our understanding from the Box-And-Whisker plot. When taken together, it leaves us with more questions than answers.

As for regional variation: Being in the Southern US increases the rate of obesity but in terms of practical significance it is not that important a predictor. Health outcomes are much better if located in the Coastal regions i.e. Northeast and West.

Our model explains almost 90% of the variability in the data. This comes after the square root adjustment as well as the introduction of the categorical variables for Income and region.

For future work, examining the interactions between income and physical activity will yield to a better understanding of Obesity. This can be achieved by using income on a continuous scale instead of categorical variables.

Also, we used population variation as a weight to ensure more accurate variables but in hindsight is could have also distorted our interpretations. But it is worth noting that the structure of our data was such that we weren't working with individual level data to begin with. We started with what could best be described as a dataset of surveys containing percentages aggregated across income brackets and States.

As a result, utilizing every aspect of the dataset, from the State codes to the sample_sizes was not an unwarranted approach.