Adult Obesity vs Income and political affiliation

Lucy Lin

2023-07-13

This is a small project that uses multiple linear regression to create a model of the percentage of obesity given the information from the CDC's 2021 Obesity/Weight status by income dataset.

The link to this data can be found here: https://nccd.cdc.gov/dnpao_dtm/rdPage.aspx?rdReport=DNPAO_DTM.ExploreByTopic&islClass=OWS&islTopic=OWS1&go=GO

I have also used the party affiliation by state chart from Pew Research.

The link to this data can be found here: https://www.pewresearch.org/religion/religious-landscape-study/compare/party-affiliation/by/state/

```
#Formatting and preparing the data
raw_data <- read.csv("C:/Users/Ycull/Downloads/AdultObesity_Income_2021.csv")
data <- raw_data[8:364,c(1,6,15,29)]
names(data)[2] <- 'State'
#I filtered the data and renamed "LocationDesc" to "State"

political_data <-read.csv("C:/Users/Ycull/Downloads/Party affiliation by state - Sheet1.csv")

#install.packages("dplyr")
#comment out install functions for R Markdown
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#Using dplyr package to merge the two tables

combined_data <- inner_join(data,political_data, by = "State")
combined_data <- combined_data[,c(1:5,7)]

combined_data$Republican.lean.Rep. <- as.numeric(sub("%", "", combined_data$Republican.lean.Rep.,fixed=
combined_data$Democrat.lean.Dem. <- as.numeric(sub("%", "", combined_data$Democrat.lean.Dem.,fixed=TRUE
```

```r
combined_data$Lean_value <- combined_data$Republican.lean.Rep. - combined_data$Democrat.lean.Dem.
#Positive values of lean_value imply that the state has more Republicans than Democrats and negative va

combined_data <- combined_data[!grepl("Data not reported", combined_data$Stratification1),]
#removed columns of unknown income

combined_data <- combined_data[!grepl("-", combined_data$Data_Value),]
combined_data <- combined_data[!grepl("~", combined_data$Data_Value),]
#removed columns of unknown obesity values

names(combined_data)[names(combined_data) == 'Stratification1'] <- 'Income_Bracket'
#renamed "Stratification1" to "Income_Bracket"

combined_data$Income_Bracket <- as.factor(combined_data$Income_Bracket)
#Income bracket is now a factor with 6 levels
```

My hypothesis is that residents with higher income in Democratic-leaning states are most likely to have lower obesity rates. My assumptions are that higher incomes enable healthier lifestyles for the residents and Democratic-leaning states are more likely to enforce progressive regulations on healthful foods.

```r
levels(combined_data$Income_Bracket)
```

```
## [1] "$15,000 - $24,999"  "$25,000 - $34,999"  "$35,000 - $49,999"
## [4] "$50,000 - $74,999"  "$75,000 or greater" "Less than $15,000"
```

```r
combined_data$Income_Bracket1 = relevel(combined_data$Income_Bracket, ref = 'Less than $15,000')
#combined_data$Income_Bracket1 = factor(combined_data$Income_Bracket, levels = c("Less than $15,000","$
levels(combined_data$Income_Bracket1)
```

```
## [1] "Less than $15,000"  "$15,000 - $24,999"  "$25,000 - $34,999"
## [4] "$35,000 - $49,999"  "$50,000 - $74,999"  "$75,000 or greater"
```

```r
#Choosing a baseline constraint of the lowest income bracket by reordering baseline levels
#Expectation is that as income levels increase further from the baseline, the variable will probably be

fit1 = lm(Data_Value ~ Income_Bracket1 + Lean_value + Income_Bracket1:Lean_value , data = combined_data)
summary(fit1)
```

```
##
## Call:
## lm(formula = Data_Value ~ Income_Bracket1 + Lean_value + Income_Bracket1:Lean_value,
##     data = combined_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7752  -3.0411   0.2809   3.1610  10.8611
##
## Coefficients:
##                                             Estimate Std. Error t value
## (Intercept)                                  39.6010     0.6410  61.781
## Income_Bracket1$15,000 - $24,999             -1.6877     0.9065  -1.862
```

```
## Income_Bracket1$25,000 - $34,999              -2.3219     0.9065  -2.561
## Income_Bracket1$35,000 - $49,999              -3.5769     0.9065  -3.946
## Income_Bracket1$50,000 - $74,999              -3.5636     0.9097  -3.917
## Income_Bracket1$75,000 or greater             -5.0613     0.9065  -5.583
## Lean_value                                     2.3923     3.7821   0.633
## Income_Bracket1$15,000 - $24,999:Lean_value    0.1691     5.3486   0.032
## Income_Bracket1$25,000 - $34,999:Lean_value    5.1109     5.3486   0.956
## Income_Bracket1$35,000 - $49,999:Lean_value   10.0835     5.3486   1.885
## Income_Bracket1$50,000 - $74,999:Lean_value   13.2418     5.3614   2.470
## Income_Bracket1$75,000 or greater:Lean_value  13.4062     5.3486   2.506
##                                              Pr(>|t|)
## (Intercept)                                   < 2e-16 ***
## Income_Bracket1$15,000 - $24,999             0.063654 .
## Income_Bracket1$25,000 - $34,999             0.010937 *
## Income_Bracket1$35,000 - $49,999             0.000100 ***
## Income_Bracket1$50,000 - $74,999             0.000112 ***
## Income_Bracket1$75,000 or greater            5.47e-08 ***
## Lean_value                                   0.527535
## Income_Bracket1$15,000 - $24,999:Lean_value  0.974801
## Income_Bracket1$25,000 - $34,999:Lean_value  0.340100
## Income_Bracket1$35,000 - $49,999:Lean_value  0.060406 .
## Income_Bracket1$50,000 - $74,999:Lean_value  0.014099 *
## Income_Bracket1$75,000 or greater:Lean_value 0.012747 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.422 on 287 degrees of freedom
## Multiple R-squared:  0.2578, Adjusted R-squared:  0.2293
## F-statistic: 9.062 on 11 and 287 DF,  p-value: 6.838e-14
```

```
fit2 = lm(Data_Value ~ Income_Bracket1 + Lean_value, data = combined_data)
summary(fit2)
```

```
##
## Call:
## lm(formula = Data_Value ~ Income_Bracket1 + Lean_value, data = combined_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7133  -2.9803   0.2697   3.0280  10.5391
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      39.8611     0.6367  62.604  < 2e-16 ***
## Income_Bracket1$15,000 - $24,999   -1.6940     0.8967  -1.889  0.05986 .
## Income_Bracket1$25,000 - $34,999   -2.5120     0.8967  -2.801  0.00543 **
## Income_Bracket1$35,000 - $49,999   -3.9520     0.8967  -4.407 1.47e-05 ***
## Income_Bracket1$50,000 - $74,999   -4.0418     0.9012  -4.485 1.05e-05 ***
## Income_Bracket1$75,000 or greater  -5.5600     0.8967  -6.201 1.91e-09 ***
## Lean_value                          9.3844     1.5667   5.990 6.16e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.483 on 292 degrees of freedom
```

```
## Multiple R-squared:  0.2237, Adjusted R-squared:  0.2078
## F-statistic: 14.03 on 6 and 292 DF,  p-value: 4.987e-14
```

```
fit3 = lm(Data_Value ~ Lean_value, data = combined_data)
summary(fit3)
```

```
##
## Call:
## lm(formula = Data_Value ~ Lean_value, data = combined_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.7529  -3.6542   0.3012   3.3025  11.8021
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.9042     0.2844 129.778  < 2e-16 ***
## Lean_value    9.3694     1.6772   5.586 5.25e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.8 on 297 degrees of freedom
## Multiple R-squared:  0.09508,    Adjusted R-squared:  0.09203
## F-statistic: 31.21 on 1 and 297 DF,  p-value: 5.247e-08
```

```
fit4 = lm(Data_Value ~ Income_Bracket1, data = combined_data)
summary(fit4)
```

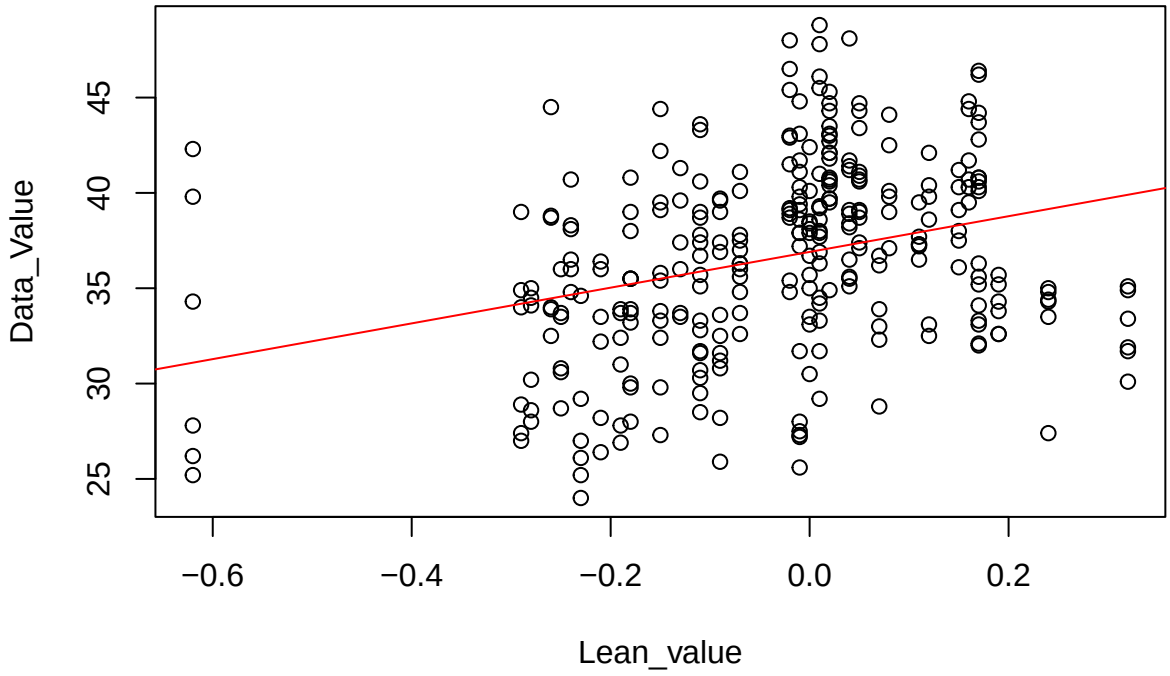
```
##
## Call:
## lm(formula = Data_Value ~ Income_Bracket1, data = combined_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.112  -3.426   0.088   3.591  10.982
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        39.5120     0.6707  58.910  < 2e-16 ***
## Income_Bracket1$15,000 - $24,999   -1.6940     0.9485  -1.786  0.07515 .
## Income_Bracket1$25,000 - $34,999   -2.5120     0.9485  -2.648  0.00853 **
## Income_Bracket1$35,000 - $49,999   -3.9520     0.9485  -4.166 4.08e-05 ***
## Income_Bracket1$50,000 - $74,999   -4.0202     0.9534  -4.217 3.31e-05 ***
## Income_Bracket1$75,000 or greater  -5.5600     0.9485  -5.862 1.23e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.743 on 293 degrees of freedom
## Multiple R-squared:  0.1284, Adjusted R-squared:  0.1135
## F-statistic: 8.629 on 5 and 293 DF,  p-value: 1.209e-07
```

```r
library(knitr)
kable(anova(fit3), caption='ANOVA Table')
```

Table 1: ANOVA Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Lean_value	1	718.9035	718.90347	31.20561	1e-07
Residuals	297	6842.1787	23.03764	NA	NA

```r
plot(Data_Value ~ Lean_value, data = combined_data)
abline(fit3, col = 'red')
```




```r
#plot of obesity rates vs lean value
```

Based on the multiple regression model using the lowest income bracket of "Less than $15,000" as the baseline constraint, there is a weak correlation present between income and obesity. As I have predicted, obesity rates seem to decrease as income increases. In other words, as income increases farther from the baseline income, each income bracket, as a factor, becomes more statistically significant. This is also suggested by the national averages by income brackets in the first few rows of the CDC's dataset.

There is an even weaker correlation between the state's political affiliation but this variable still remains statistically significant. Specifically, Republican leaning states tend to have higher rates of obesity.

The interaction variable between income and political affiliation appear to less significant. Yet, if this is taken into consideration, increased income results in a lower increase in obesity rates.

While my r^2 values are very low, my p-values are statistically significant and therefore do suggest there is some level of correlation.

There are obvious limitations to these models and data. The sample sizes vary greatly for each state and the CDC has disclaimers about the possible inaccuracies of this dataset. My fitted models rely on only two variables and state governments might have a lot less to do with obesity rates in their state. Obesity is a complex issue and is rampant across the entire country.