Harrisburg University of Science and Technology

Factors Influencing Diabetes incidence in the US

ANLY510-92 (Spring): Analytics II – Principle & Applications

Group 7: Subhash Bharadwaj Pemmaraju

4-14-2019

Contents

Introduction

Diabetes is a major disease affecting millions of Americans. Not only do Americans spend millions of dollars on healthcare to manage the disease, they also spend millions of dollars on studying the disease – reasons for its prevalence in certain people, genetic predisposition, alternative treatments, effective preventive measures etc. What excites me about this topic is the existence of granular data for the US that few other countries have. As a data scientist, there is ample scope for research and analysis into the factors that effect diabetes prevalence. This can help researchers identify where they should spend their time and money to help Americans deal with the problem of diabetes and other such lifestyle diseases.

Problem Statement

Through this analysis, I am trying to answer the following questions

1. What are the factors effecting incidence of diabetes in the United States?
2. Among these factors, which ones are more important?
3. Do these factors interact with one another to reinforce or dampen their effect on the population?
4. Can we use this information to predict the prevalence of diabetes in a part of the US?

Significance of the study

There are a vast multitude of variables that can effect the prevalence of diabetes. This study is one among many attempts to identify and isolate the most important factors. This will help scientists focus their time and research on the right areas to help address this major health crisis. Perhaps research like this can also generate further ideas and follow up research and kickstart a policy discussion on the disease and its impact on the US.

Hypotheses

The key hypothesis that this study makes is the following:

Prevalence of diabetes in a county is influenced by different levels of the following factors – Income, poverty, unemployment, physical inactivity, obesity, percentage of population over 65

My prediction for the outcome of the experiment is that not all these variables matter as much, some matter more than others, such as physical inactivity and obesity prevalence.

This prediction is logical because:

- If people are obese, then they are more likely to get diabetes later in their life. This is a scientific fact and should be reflected in the data such that obesity prevalence and diabetes prevalence are positively correlated
- If people are inactive, then they are more likely to get diabetes. This is also supported by the fact that diabetes is a "lifestyle disease" in that people living sedentary lifestyles and not getting enough exercise are more likely to get diabetes. So inactivity and diabetes prevalence should be positively correlated
- If people are earning more, then they are less likely to get diabetes. It is an economic fact that as people's income levels grow, they are more likely to consume protein and can afford the diet and healthcare that a higher income provides. So they are likely to suffer from diabetes less. So incidence of diabetes and income levels should be negatively correlated

Identification of Variables

The dataset, which was put together by the CDC, uses a variety of sources including:

- CDC's data on diabetes prevalence, obesity, leisure-time physical inactivity at a county level

- US Census bureau social and economic data at a county level

The key variables are:

Income – Median household income ($1000s)

Unemploy – Unemployment Rate (%)

Ob_prev_adj – Obesity percentage (%)

Dm_prev_adj – Diagnosed Diabetes percentage (%)

Ltpia_prev_adj – Leisure time physical inactivity percentage (%)

Perc65_up – Age 65 years and above percentage (%)

Povpct – people living in poverty percentage (%)

The dependent variable in this analysis is the % of people having diabetes (dm_prev_adj) and the independent variables are all the other variables listed above.

Furthermore, a few modified variables are used, which are factor variables with two levels of the independent variables = 1 for values below the median, 2 for values at or above the median

Identification of Factors and Levels

All the independent variables are bucketed into low and high with values 1 and 2 respectively based on whether the values lie below the median or at/above the median. The factor variables are listed below along with their levels and what the levels mean.

Variable	Level	Value	Level	Value
Inc_level	1	<46.4	2	>=46.4
Unemp_level	1	<4.1%	2	>=4.1%

Ob_level	1	<31.5%	2	>=31.5%
Inact_perc_level	1	<25.5%	2	>=25.5%
Perc65up_level	1	<17.2%	2	>=17.2%
Pov_level	1	<15%	2	>=15%

Population Specification

The population is US county level data for the dependent and independent variables. The data used for analysis is selected on the basis of presence of data for the dependent and independent variables. The counties for which the data is missing is deleted from the dataset.

Literature Review

Marian et al. (2013) examine how the prevalence of diabetes varies by socioeconomic status within migrant groups of Australia. They find that prevalence of Type 2 diabetes was higher in men than women and across socio-economic strata, the prevalence of diabetes is higher among migrant populations than native populations. Wild et al. (2010) in their study, find that outcomes associated with diagnosed diabetes are worse for people from lower socio-economic strata of the population. They conduct this study in Scotland. They find that people in the lower quartiles of socio-economic variables are more likely to have hospital admissions for diabetes related complications. Lida et al. (2017) conduct a study on the changes in diagnosed diabetes, obesity and physical inactivity in the US between 2004 and 2012. They find that compared to 2004–2008, the year-on-year changes for diabetes, obesity, and physical inactivity were lower in 2008–2012 and these differences varies by regions and counties. They conclude that levels of these risk factors remain high in the US population and need to be reduced. Gucciardi et al. (2014) examine an interesting interaction between food insecurity and prevalence of diabetes in the US. They find that food insecurity is higher among households with diabetics. This results in a difficult situation where households have competing priorities between the need for food, diabetes medication, healthcare etc. which makes their situation worse. Rosella et al. (2016) study the impact of diabetes on healthcare costs. They find that diabetes disproportionately impacts healthcare costs for households and call for an allocation of resources to deal with that problem.

Procedures of the Experimental Testing

Design selection

In this paper, we have five different socio-economic variables that are the treatment variables and the two treatments they undergo are values below the median and at/above the median respectively. The dependent variable on which these treatments are applied is the diabetes percentage of the population.

Experimental design

The experiment is conducted as 2^k factorial design where k = 6 and each of the 6 variables have two levels. K are all continuous variables bucketed into factor variables

Unit observation

Each observation corresponds to a particular county being subject to a certain level of each of the 6 dependent variables. The impact of these treatments is felt on the dependent variable, which is the prevalence of diabetes in the county.

Data Collection Procedures and Methods

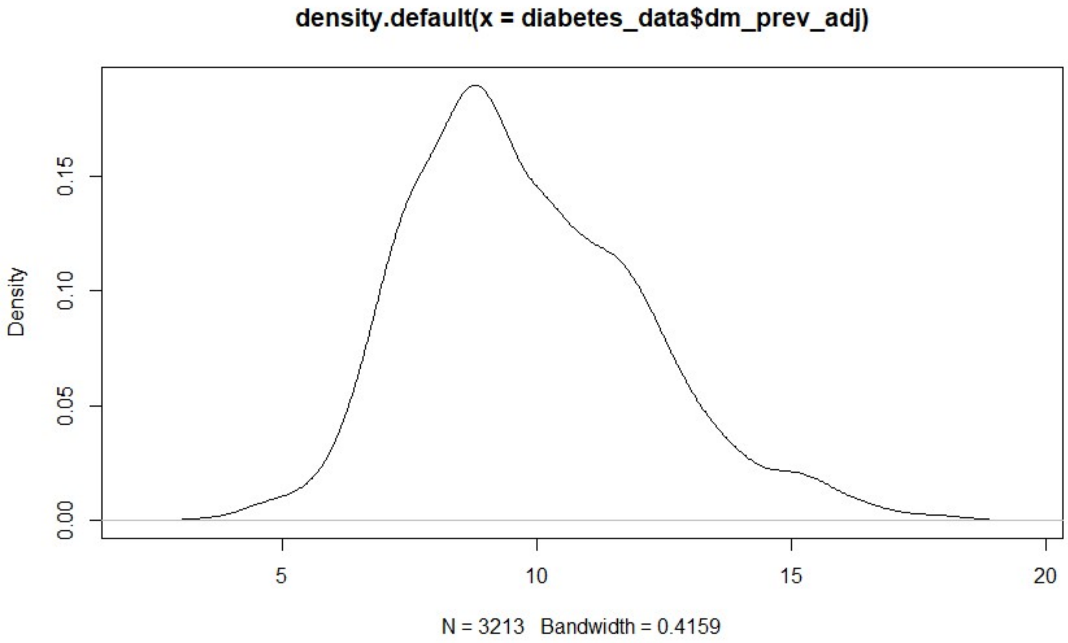
The data has been downloaded from the CDC, which put together the data from a multitude of sources including surveys and US census bureau data for the socio-economic variables.

Data Analysis and Findings

ANOVA Analysis

Before carrying out ANOVA analysis, we first need to make sure that the assumptions of ANOVA are met.

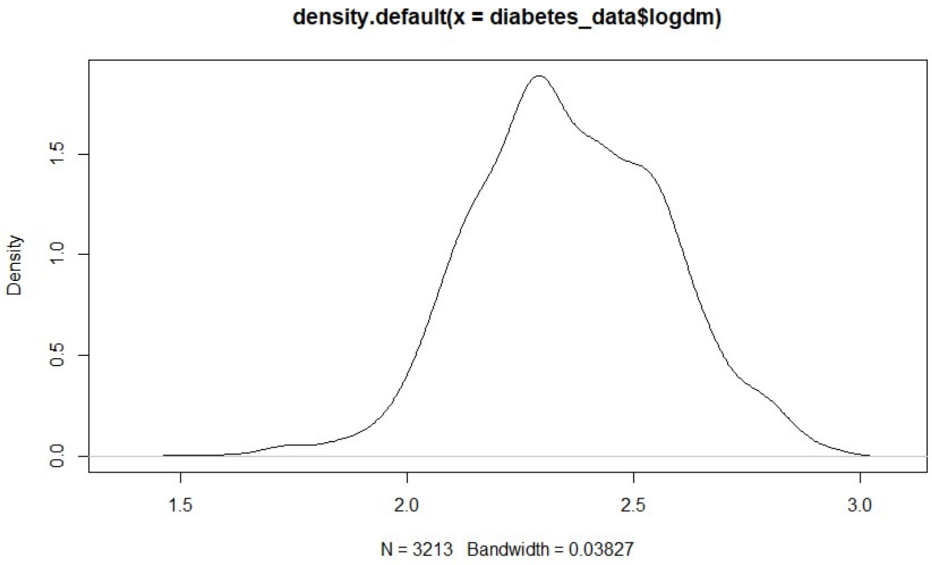
1. Normality of the variables



```
D'Agostino skewness test


data:  diabetes_data$dm_prev_adj

skew = 0.56452, z = 12.23200, p-value < 2.2e-16

alternative hypothesis: data have a skewness
```

We can see that there is a problem of skew in the data, so we do a log transformation with an adjustment

Logdm = log(diabetes prevalence+1)



```
            D'Agostino skewness test


data:  diabetes_data$logdm

skew = -0.055443, z = -1.285000, p-value = 0.1988

alternative hypothesis: data have a skewness
```

The D'Agostino test indicates that the skew problem has been resolved with the transformation, as we can see from the density plot and from the normality test.

2. Assumption of equality of variance

Bartlett test is conducted for the dependent variable and each of the independent variables.

```
> bartlett.test(diabetes_data$logdm, diabetes_data$inc_level)

            Bartlett test of homogeneity of variances

data:  diabetes_data$logdm and diabetes_data$inc_level

Bartlett's K-squared = 46.006, df = 1, p-value = 1.179e-11

> bartlett.test(diabetes_data$logdm, diabetes_data$pov_level)
```

```
        Bartlett test of homogeneity of variances

data:  diabetes_data$logdm and diabetes_data$pov_level

Bartlett's K-squared = 6.5596, df = 1, p-value = 0.01043

> bartlett.test(diabetes_data$logdm, diabetes_data$unemp_level)

        Bartlett test of homogeneity of variances

data:  diabetes_data$logdm and diabetes_data$unemp_level

Bartlett's K-squared = 2.9747, df = 1, p-value = 0.08458

> bartlett.test(diabetes_data$logdm, diabetes_data$perc65up_level)

        Bartlett test of homogeneity of variances

data:  diabetes_data$logdm and diabetes_data$perc65up_level

Bartlett's K-squared = 8.9811, df = 1, p-value = 0.002728

> bartlett.test(diabetes_data$logdm, diabetes_data$ob_level)

        Bartlett test of homogeneity of variances

data:  diabetes_data$logdm and diabetes_data$ob_level

Bartlett's K-squared = 21.301, df = 1, p-value = 3.924e-06

> bartlett.test(diabetes_data$logdm, diabetes_data$inact_perc_level)

        Bartlett test of homogeneity of variances

data:  diabetes_data$logdm and diabetes_data$inact_perc_level

Bartlett's K-squared = 21.343, df = 1, p-value = 3.84e-06
```

Bartlett’s test seems to indicate a problem with the variance of the data. There seems to be evidence of heteroskedasticity. We will see if this is a problem for the analysis or not.

We compare the variance at each of the level of the variables and compare to see if there is a big difference.

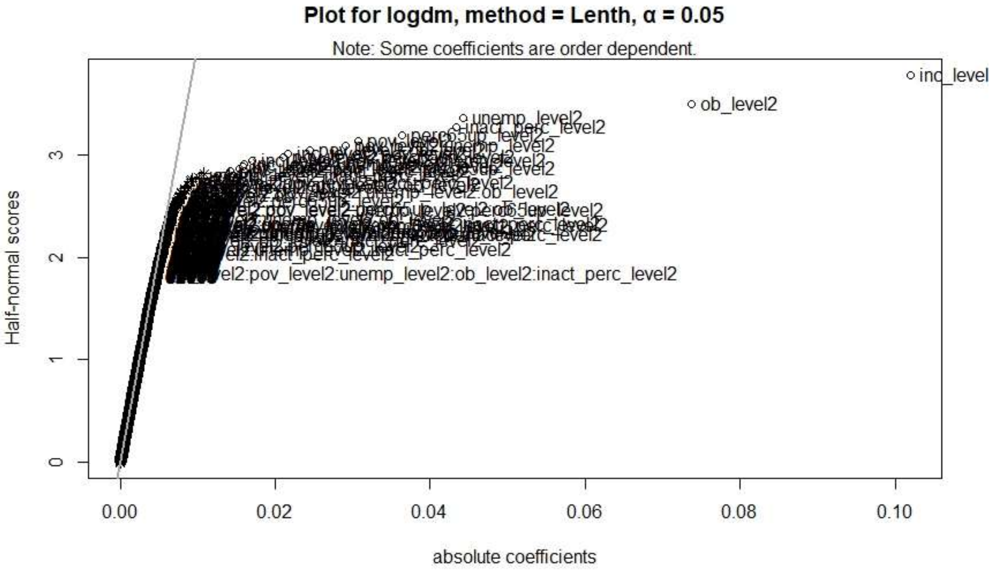
```
> tapply(diabetes_data$logdm, diabetes_data$inc_level, var)
         1          2
0.04132490 0.02943708
> tapply(diabetes_data$logdm, diabetes_data$pov_level, var)
         1          2
0.03979680 0.03501951
> tapply(diabetes_data$logdm, diabetes_data$unemp_level, var)
         1          2
0.03660690 0.03990179
> tapply(diabetes_data$logdm, diabetes_data$perc65up_level, var)
         1          2
0.04838676 0.04166201
> tapply(diabetes_data$logdm, diabetes_data$ob_level, var)
         1          2
0.04044393 0.03211693
> tapply(diabetes_data$logdm, diabetes_data$inact_perc_level, var)
         1          2
0.03878216 0.03078980
```

We can see that the difference is very small around 1-2x so heteroskedasticity is not a big enough problem to be concerned about.

3. Next we develop the **full ANOVA model**

   model<-aov(logdm~inc_level*pov_level*unemp_level*perc65up_level*ob_level*inact_perc_level, data=diabetes_data)

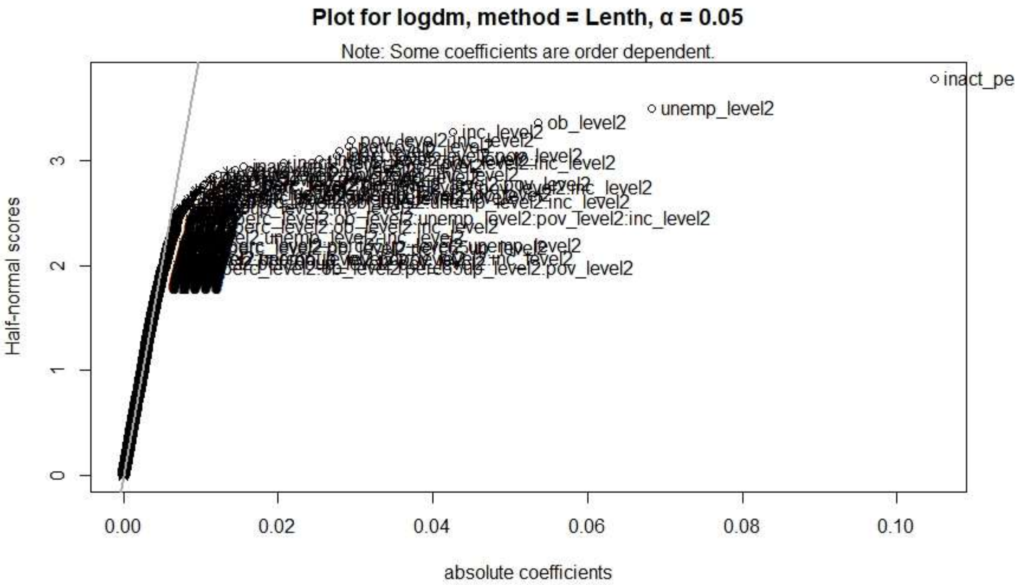
4. The **halfnormal plot** reveals the most important variables in the analysis



5. We also attempt the **reverse order model** to account for the fact that differences in sequence can change the result

   model_rev<-aov(logdm~inact_perc_level*ob_level*perc65up_level*unemp_level*pov_level*inc_level, data=diabetes_data)

6. The **halfnormal plot** reveals the most important variables in that regression

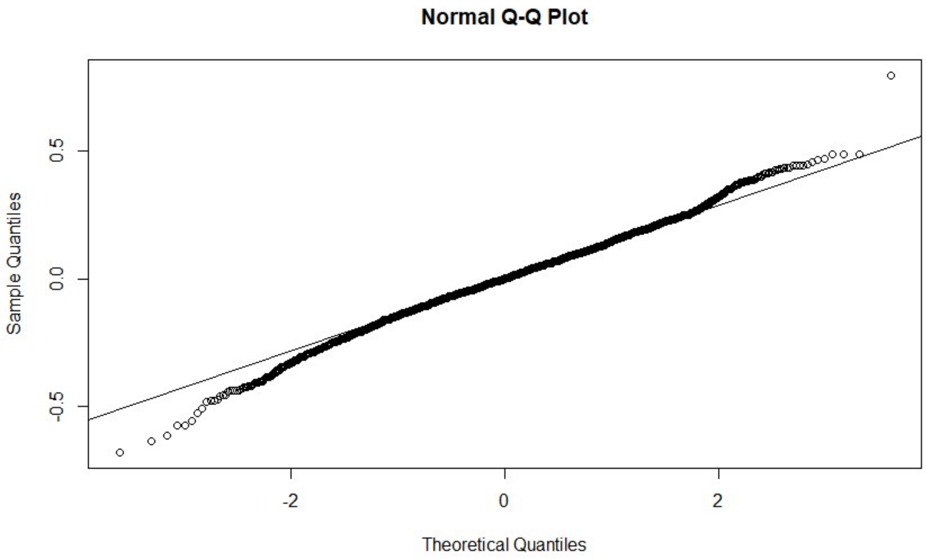


7. From the above two half normal plots, we can see that the most important variables for the analysis are:
    a. Physical inactivity levels
    b. Obesity level
    c. Unemployment level
    d. Income level
8. A **reduced ANOVA model** is now developed with the significant variables model2<-aov(logdm~inact_perc_level*ob_level*unemp_level*inc_level, data=diabetes_data)

```
Anova Table (Type III tests)

Response: logdm

                              Sum Sq  Df  F value   Pr(>F)

(Intercept)                   543.77   1 22534.5089 < 2.2e-16 ***

inact_perc_level                1.11   1    46.0331 1.380e-11 ***

ob_level                        0.54   1    22.2043 2.556e-06 ***
```

```
unemp_level                              4.88    1   202.0531 < 2.2e-16 ***
inc_level                              0.00    1     0.0948 0.7581585
inact_perc_level:ob_level                     0.01    1     0.4748 0.4908454
inact_perc_level:unemp_level                    0.90    1    37.3162 1.126e-09 ***
ob_level:unemp_level                          0.39    1    16.2822 5.585e-05 ***
inact_perc_level:inc_level                     0.13    1     5.2984 0.0214087 *
ob_level:inc_level                          0.03    1     1.4476 0.2290060
unemp_level:inc_level                         1.86    1    77.0339 < 2.2e-16 ***
inact_perc_level:ob_level:unemp_level           0.35    1    14.3897 0.0001514 ***
inact_perc_level:ob_level:inc_level             0.00    1     0.1824 0.6693288
inact_perc_level:unemp_level:inc_level          0.62    1    25.5928 4.454e-07 ***
ob_level:unemp_level:inc_level                 0.51    1    20.9960 4.779e-06 ***
inact_perc_level:ob_level:unemp_level:inc_level   0.30    1    12.4837 0.0004163 ***
Residuals                              77.14 3197
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

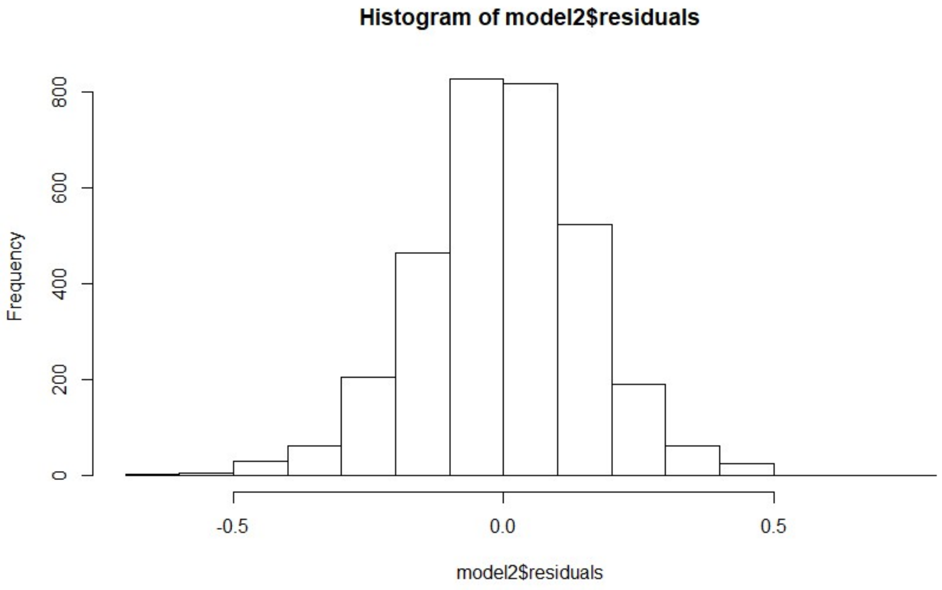
9. From **the F-test values and p-values** we can make the following conclusions:
    a. Individual dependent variables except for income level are all statistically significant
    b. All interaction effects are statistically significant except for the interaction between obesity levels and income levels; obesity levels, income levels and inactivity levels
10. **Residual analysis** is carried out:



Q-Q indicates some evidence of non-normality but it remains to be seen if it is a problem

```
Shapiro-Wilk normality test


data:  model2$residuals
W = 0.99411, p-value = 4.109e-10
```

Shapiro test indicates that there is non-normality. But looking at the histogram/density plot indicates that it is not a big problem.

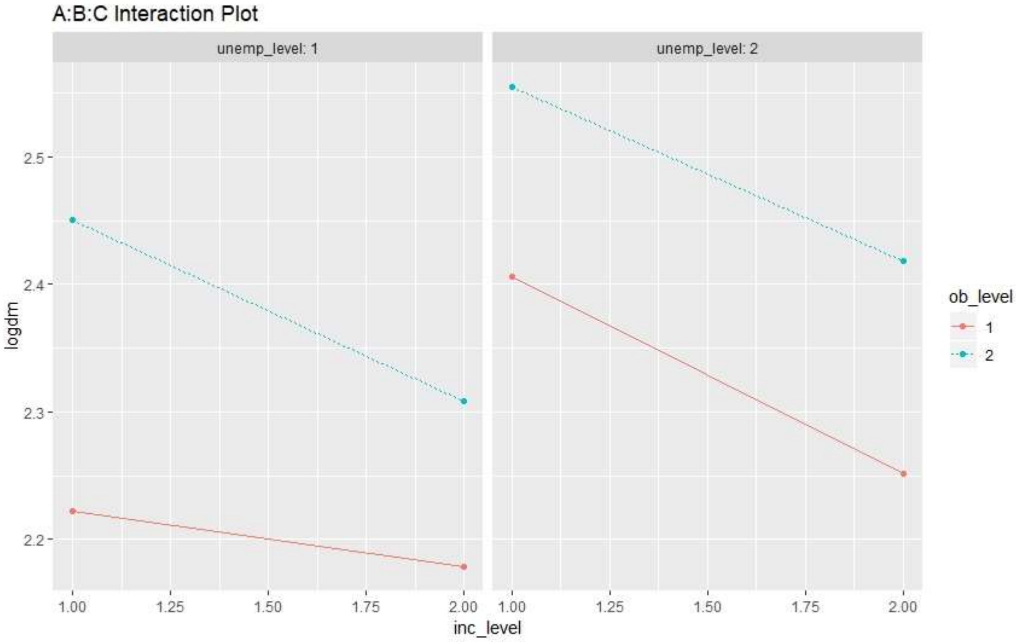


11. **Mean analysis** of the significant variables is conducted.

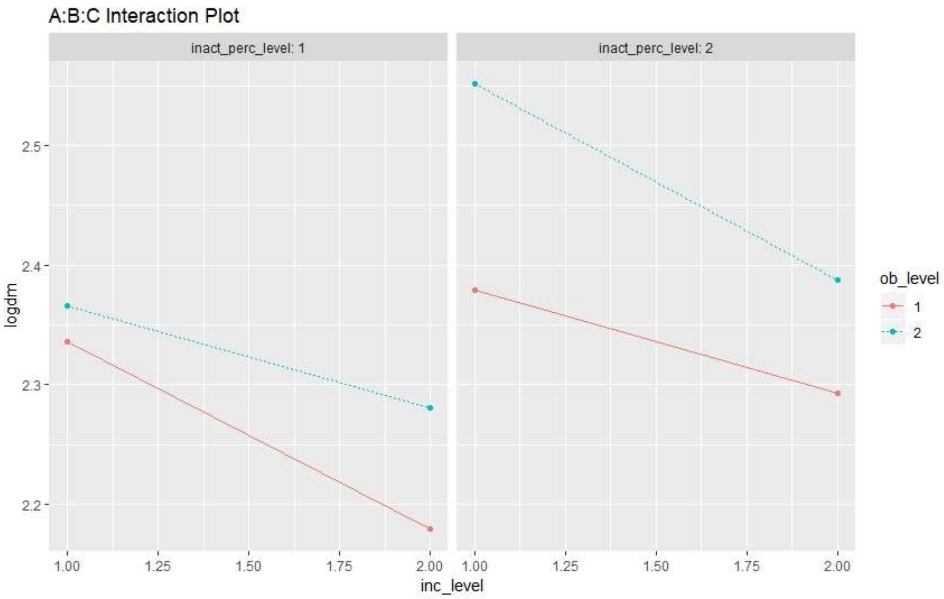
```
> tapply(diabetes_data$logdm, diabetes_data$inc_level, mean)

       1        2

2.461244 2.257303

> tapply(diabetes_data$logdm, diabetes_data$ob_level, mean)

       1        2

2.258601 2.453647

> tapply(diabetes_data$logdm, diabetes_data$unemp_level, mean)

       1        2

2.269505 2.441691

> tapply(diabetes_data$logdm, diabetes_data$inact_perc_level, mean)

       1        2

2.251562 2.461182
```

As expected, higher levels of obesity, unemployment and inactive percentage lead to higher incidence of diabetes, while higher income leads to lower incidence of diabetes.

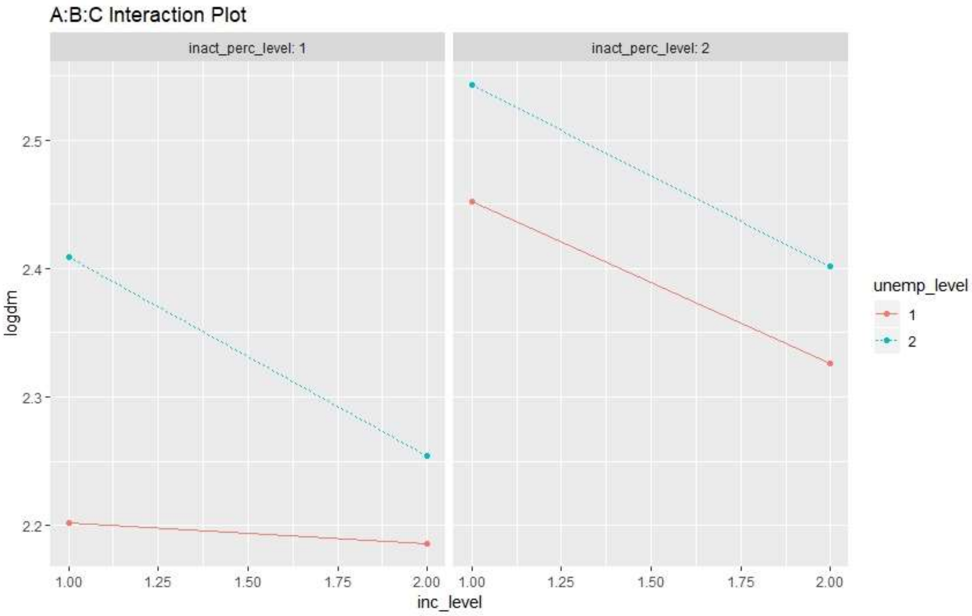
12. **Interaction effects** are also studied for the variables



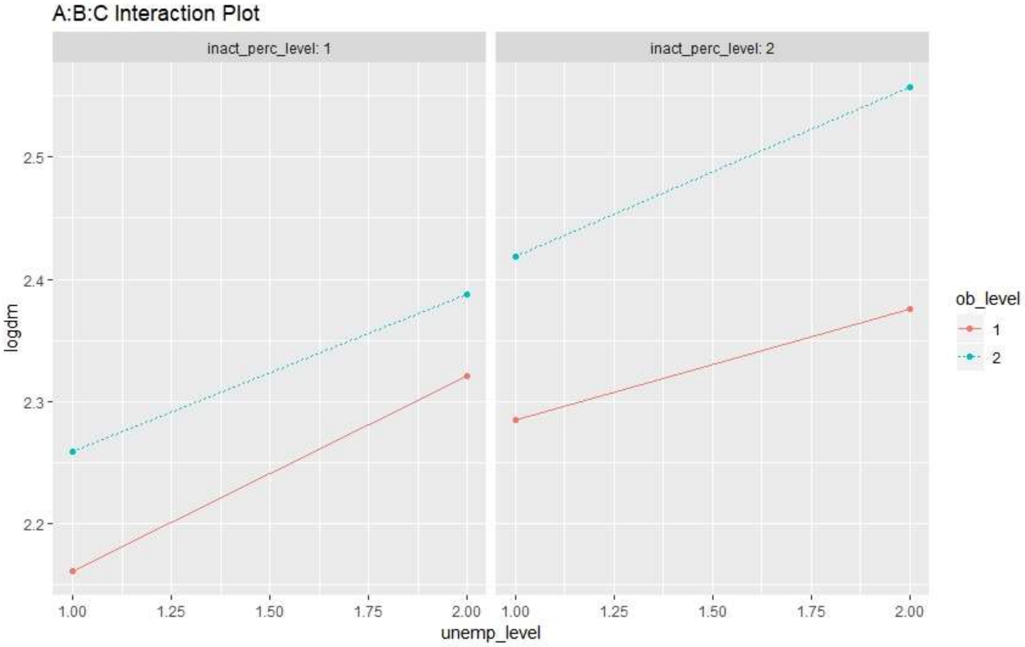
At higher obesity levels, diabetes incidence is higher. As income increases, diabetes incidence is lower. The impact of income on diabetes incidence is higher at low levels of obesity and high levels of unemployment when compared to low levels of obesity and low levels of unemployment.



At low levels of inactivity, the diabetes prevalence is on average lower than at higher levels of inactivity. Furthermore, the gap between diabetes incidence at low and high levels of obesity is smaller. And as seen before, diabetes incidence decreases with increase in income.



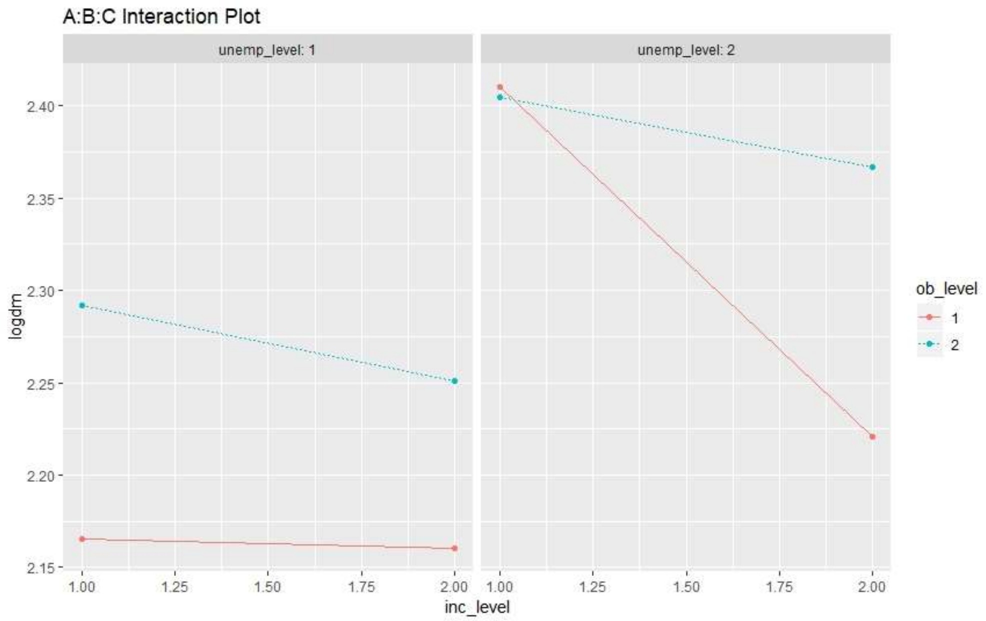
At low levels of unemployment and low levels of inactivity, diabetes prevalence is almost unchanged at different levels of income. At high levels of inactivity, while the incidence of diabetes is on average higher at higher levels of unemployment, the relationship of diabetes prevalence to income is relatively the same as at lower levels of unemployment.



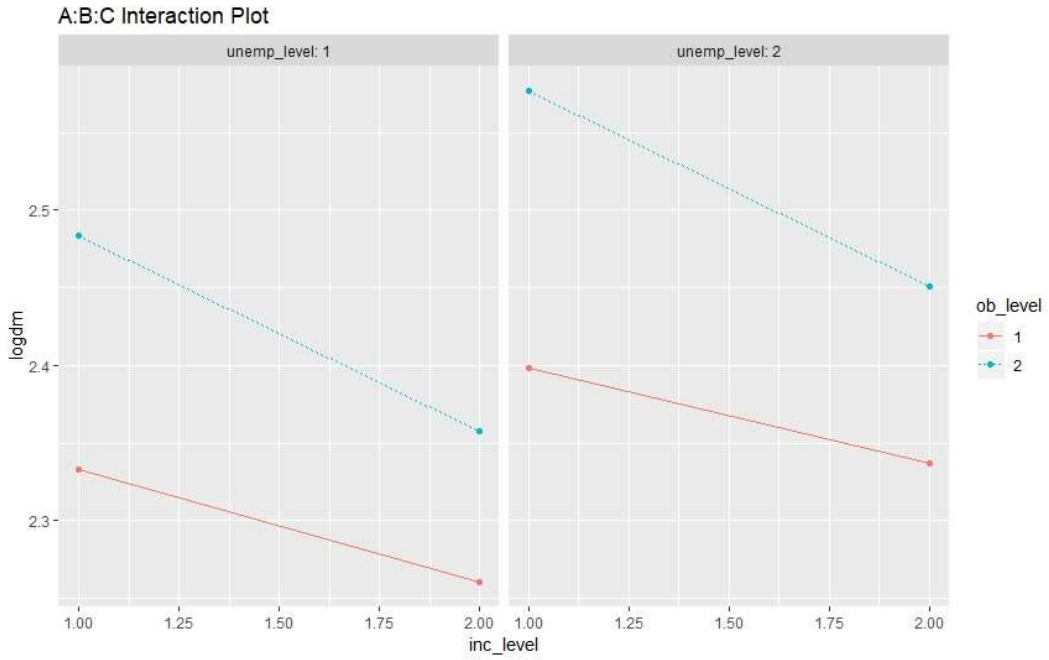
The slope of the relationship between diabetes prevalence and unemployment is relatively unchanged at different levels of obesity or inactivity. However, at low levels of inactivity, the gap between diabetes prevalence is narrow at different levels of obesity.

For **4-variable interactions**, we consider 3 variable interactions at two levels of the 4th variable, namely inactivity percentage = low/high

At low levels of inactivity,



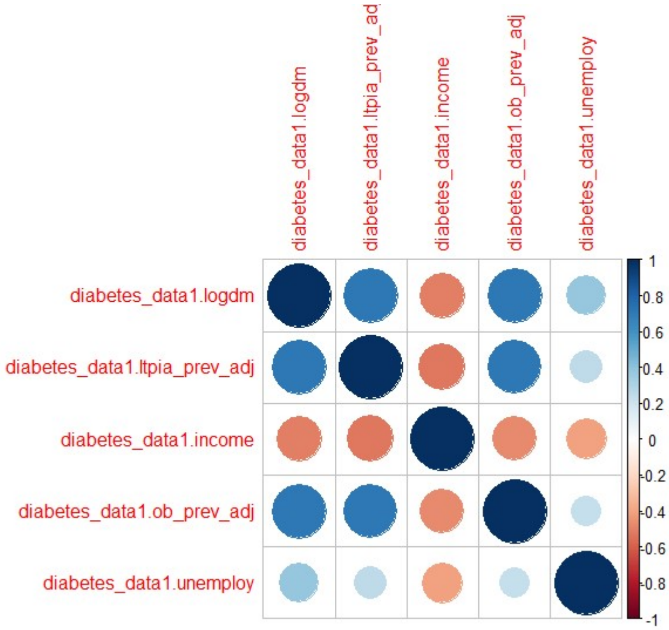
At low level of obesity and unemployment, the relationship between diabetes prevalence and income is relatively flat. The relationship is very steep at high levels of unemployment and low levels of obesity. At high levels of obesity, levels of unemployment does not impact the relationship between diabetes prevalence and income, other than shifting it to a higher level on average



At high levels of inactivity, for different levels of unemployment, the slope of the diabetes prevalence vs income level is relatively unchanged. The level of diabetes prevalence is higher on average at higher unemployment levels.

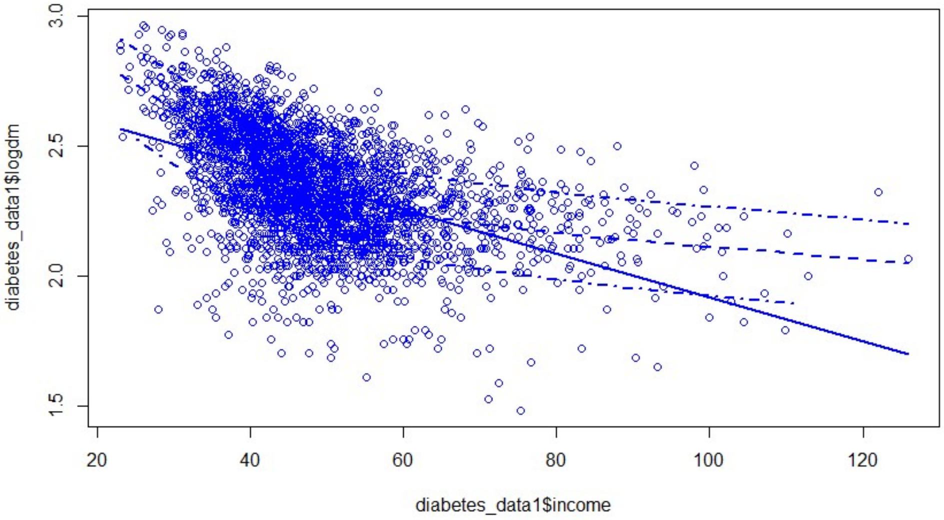
Regression Analysis

1. Regression Analysis needs to be carried out next to evaluate the relationship between the base variables of Income, Unemployment, Obesity, Inactivity and Diabetes prevalence
2. First, some analysis is carried out to see if they are in fact correlated and if yes, what the relationship is between them
3. **Correlation Plot**:

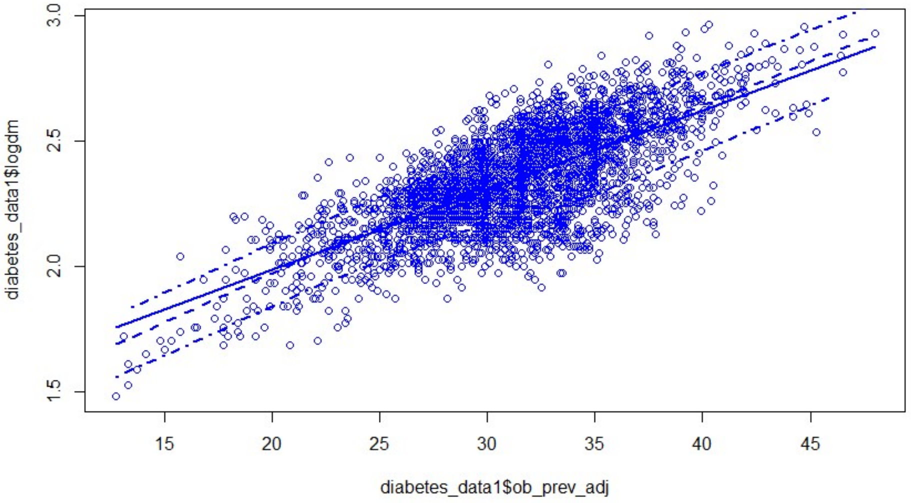


From the correlation plot, we can see that all the explanatory variables except for income are positively correlated and the correlation for Income and unemployment is smaller than the correlation for physical inactivity and obesity prevalence. Some of the explanatory variables are also correlated to each other such as inactivity and obesity.

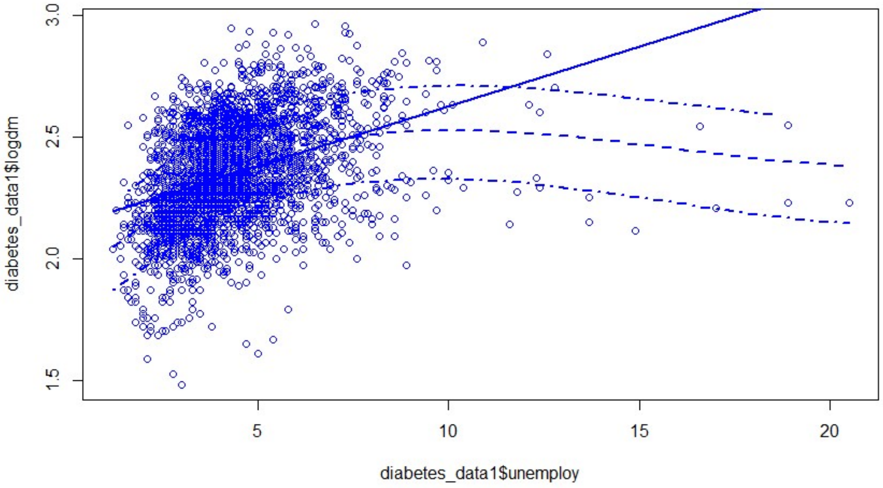
4. **Scatterplots** to check whether the relationship is largely linear

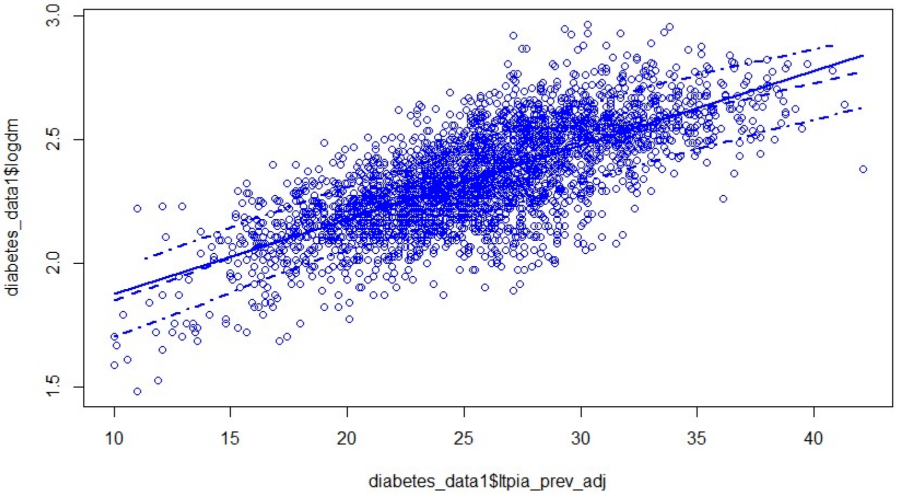


diabetes_data1$logdm
diabetes_data1$income
1.5
2.0
2.5
3.0
20
40
60
80
100
120



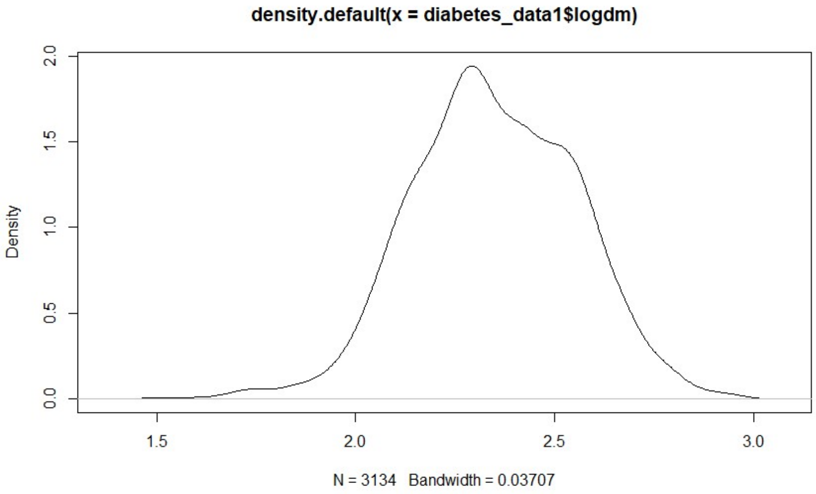
diabetes_data1$logdm
diabetes_data1$ob_prev_adj
1.5
2.0
2.5
3.0
15
20
25
30
35
40
45

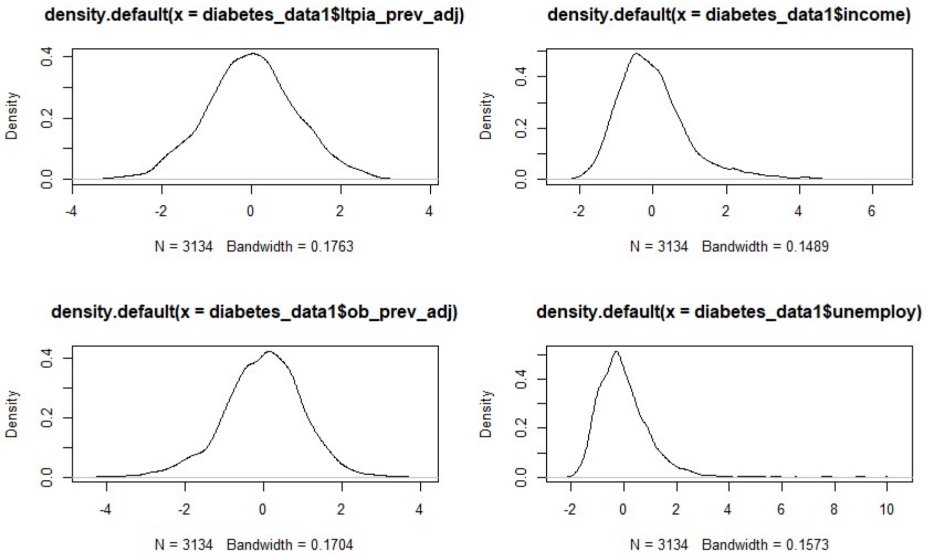




From the plots, we can see that income and unemployment appear to have non-linear relationships. So we first start with a simple linear model and examine its fit and then incorporate non-linear elements to the relationship. The variables are first scaled for convenience of analysis.

5. **Density plots** of the explanatory variables and the diabetes variable, reveals that the data is approximately normal with a little skew for income and unemployment





6. Regression analysis

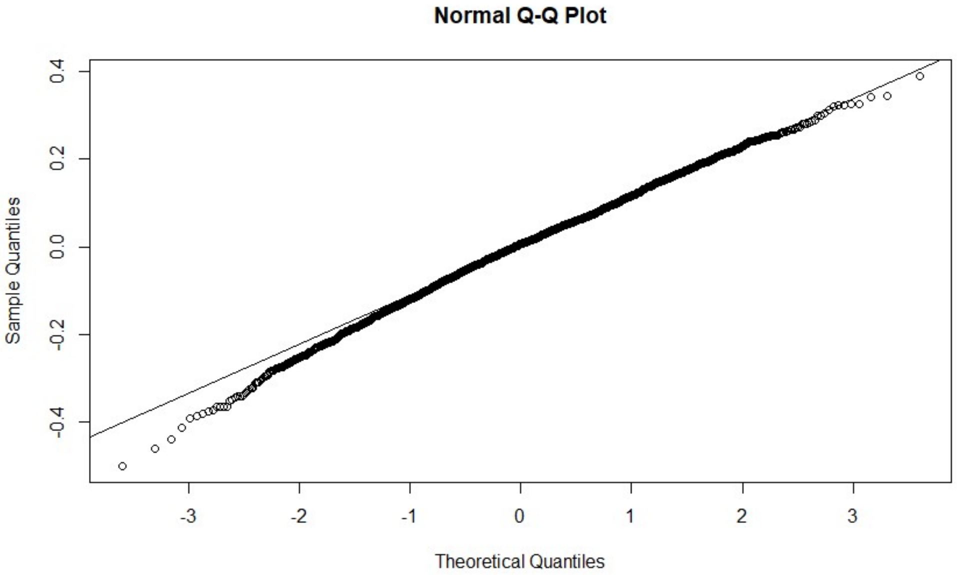
```
Call:

lm.default(formula = logdm ~ income * ob_prev_adj * unemploy *

   ltpia_prev_adj, data = diabetes_data1)

Residuals:

    Min       1Q   Median       3Q      Max

-0.49989 -0.07434  0.00705  0.07645  0.39005
```

```
Coefficients:

                              Estimate Std. Error t value Pr(>|t|)

(Intercept)                        2.344260   0.002692 870.694  < 2e-16 ***

income                            -0.029189   0.003326  -8.776  < 2e-16 ***

ob_prev_adj                         0.066772   0.003525  18.944  < 2e-16 ***

unemploy                           0.039605   0.002978  13.301  < 2e-16 ***

ltpia_prev_adj                      0.072203   0.003513  20.552  < 2e-16 ***

income:ob_prev_adj                     -0.033012   0.003543  -9.317  < 2e-16 ***

income:unemploy                        0.004497   0.003110   1.446 0.148277

ob_prev_adj:unemploy                    -0.004905   0.003905  -1.256 0.209191

income:ltpia_prev_adj                  -0.014814   0.003772  -3.928 8.75e-05 ***

ob_prev_adj:ltpia_prev_adj               -0.020046   0.002409  -8.321  < 2e-16 ***

unemploy:ltpia_prev_adj                  0.001211   0.003953   0.306 0.759475

income:ob_prev_adj:unemploy               -0.008542   0.003489  -2.448 0.014407 *

income:ob_prev_adj:ltpia_prev_adj          -0.009214   0.001680  -5.485 4.47e-08 ***

income:unemploy:ltpia_prev_adj             0.013321   0.003718   3.583 0.000345 ***

ob_prev_adj:unemploy:ltpia_prev_adj        -0.004499   0.002358  -1.908 0.056447 .

income:ob_prev_adj:unemploy:ltpia_prev_adj  0.001824   0.001237   1.474 0.140499

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.1196 on 3118 degrees of freedom
```

Multiple R-squared: 0.6645,	Adjusted R-squared: 0.6628

F-statistic: 411.6 on 15 and 3118 DF, p-value: < 2.2e-16

All the main effects and most of the interaction effects are significant as can be seen from the regression results. VIF output of the model shows no problem of multicollinearity. As expected, except for the income effect which is negative, all the other main variables show positive correlation to diabetes prevalence. It needs to be noted that the dependent variable here is a log variable, so in order to be able to use this model for prediction, the predicted log value must be transformed to its original form to get the precited value of diabetes prevalence.



Q-Q plot shows some evidence of skewness, but it doesn't appear to be a problem. Shapiro-Wilk test confirms non-normality.

Shapiro-Wilk normality test

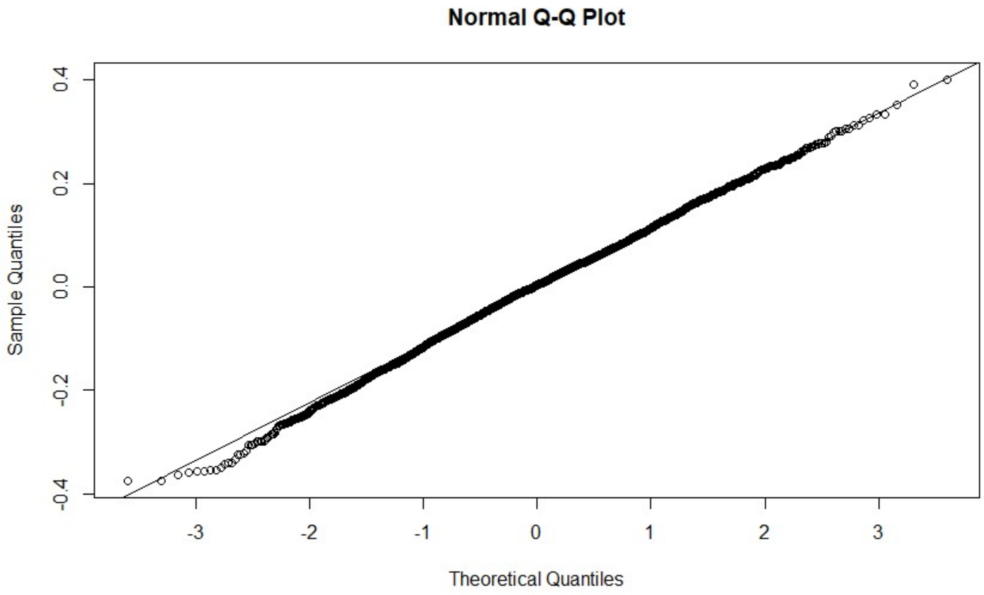
```
data:  model3$residuals
W = 0.99647, p-value = 9.916e-07
```

Now a model is built incorporating non-linearity of unemployment into the mix by adding square of unemployment as a dependent variable.

```
Call:
lm.default(formula = logdm ~ income * ob_prev_adj * unemploy *
    ltpia_prev_adj + unemploy2, data = diabetes_data1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37587 -0.07571  0.00365  0.07540  0.40198

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                            2.3515762  0.0026793 877.670  < 2e-16 ***
income                                -0.0234183  0.0032676  -7.167 9.53e-13 ***
ob_prev_adj                            0.0686845  0.0034344  19.999  < 2e-16 ***
unemploy                               0.0584830  0.0032342  18.083  < 2e-16 ***
ltpia_prev_adj                         0.0707320  0.0034219  20.670  < 2e-16 ***
unemploy2                             -0.0098142  0.0007458 -13.159  < 2e-16 ***
income:ob_prev_adj                    -0.0272733  0.0034765  -7.845 5.90e-15 ***
income:unemploy                       -0.0079404  0.0031714  -2.504 0.012339 *
```

```
ob_prev_adj:unemploy                    -0.0008993  0.0038140  -0.236 0.813611

income:ltpia_prev_adj                   -0.0129646  0.0036743  -3.528 0.000424 ***

ob_prev_adj:ltpia_prev_adj              -0.0183950  0.0023486  -7.832 6.52e-15 ***

unemploy:ltpia_prev_adj                 -0.0028366  0.0038609  -0.735 0.462575

income:ob_prev_adj:unemploy             -0.0056188  0.0034038  -1.651 0.098892 .

income:ob_prev_adj:ltpia_prev_adj       -0.0089023  0.0016356  -5.443 5.65e-08 ***

income:unemploy:ltpia_prev_adj           0.0098435  0.0036286   2.713 0.006709 **

ob_prev_adj:unemploy:ltpia_prev_adj     -0.0060688  0.0022982  -2.641 0.008316 **

income:ob_prev_adj:unemploy:ltpia_prev_adj  0.0016082  0.0012042   1.336 0.181808

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.1165 on 3117 degrees of freedom

Multiple R-squared:  0.6821,       Adjusted R-squared:  0.6805

F-statistic:   418 on 16 and 3117 DF,  p-value: < 2.2e-16
```

The square of unemployment also becomes statistically significant and the adjusted R-squared is higher now. The residuals analysis again reveals that multicollinearity is not a problem and the Q-Q plot, shapiro wilk reveals less non-normality problem.



```
Shapiro-Wilk normality test


data:  model4$residuals

W = 0.99856, p-value = 0.007368
```

7. **Comparison of the two models**, shows the model with non-linear unemployment added is superior

```
Analysis of Variance Table


Model 1: logdm ~ income * ob_prev_adj * unemploy * ltpia_prev_adj

Model 2: logdm ~ income * ob_prev_adj * unemploy * ltpia_prev_adj + unemploy2

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)

1   3118 44.629

2   3117 42.280  1    2.3487 173.15 < 2.2e-16 ***

---
```

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

8. Similar analysis of non-linearity of income reveals that it is not significant, so the above model is the correct model

Bias Control Strategy

The data is available for over 3000 counties out of 3142 counties/county equivalents of the US. Therefore, in terms of diversity of the population from which the data has been sourced, it is fairly representative of the population of the US. The key to a successful analysis is that the data sourced is random and representative of the population as a whole. The US census bureau data meets these requirements.

Summary of Findings

From the analysis, we can see that the prevalence of diabetes is influenced by several key socio-economic variables. While some variables such as obesity prevalence, physical inactivity and unemployment are more important than other variables such as income, all of these variables do play an important role in predicting incidence of diabetes, particularly when they start interacting with each other to amplify the effect on prevalence of diabetes. For example, we have seen how physical inactivity a very important magnifier of the effect on diabetes prevalence at all levels of the other explanatory variables can be.

Furthermore, having identified the factors that influence the prevalence of diabetes in a county the most, we have also built a regression model to help predict that prevalence. We have seen that unemployment has a non-linear effect and in the absence of such non-linear effects, we would be overestimating its influence on diabetes prevalence (since the coefficient on the non-linear term is negative).

Recommendations

Based on the analysis, we recommend that in order to be able to effectively predict and subsequently manage the incidence of diabetes in a county, we need to track some key variables such as prevalence of obesity in the county, the income levels, unemployment levels, physical inactivity levels. What would be worth exploring further is differences across gender and race as empirical research historically shows that diabetes prevalence varies by gender and race.

References

1. Abouzeid, M., Philpot, B., Janus, E. D., Coates, M. J., & Dunbar, J. A. (2013). Type 2 diabetes prevalence varies by socioeconomic status within and between migrant groups: analysis and implications for Australia. *BMC Public Health, 13*(1), 1–9.
2. Wild, S., Mcknight, J., Mcconnachie, A., & Lindsay, R. (2010). Socio-economic status and diabetes-related hospital admissions: A cross-sectional study of people with diagnosed diabetes. *Journal of Epidemiology and Community Health, 64*(11), 1022-1024.
3. Geiss, L., Kirtland, K., Lin, J., Shrestha, S., Thompson, T., Albright, A., & Gregg, E. (2017). Changes in diagnosed diabetes, obesity, and physical inactivity prevalence in US counties, 2004-2012. *PLoS One, 12*(3), E0173428.
4. Gucciardi, E., Vahabi, M., Norris, N., Monte, J., & Farnum, P. (2014). The Intersection between Food Insecurity and Diabetes: A Review. *Current Nutrition Reports, 3*(4), 324-332.
5. Rosella, L., Lebenbaum, M., Fitzpatrick, T., O'Reilly, D., Wang, J., Booth, G., Wodchis, W. (2016). Impact of diabetes on healthcare costs in a population-based cohort: A cost analysis. *Diabetic Medicine, 33*(3), 395-403.
6. Data Source: Interactive Atlas of Heart Disease and Stroke, Centers for Disease Control and Prevention, https://nccd.cdc.gov/DHDSPAtlas/?state=County&ol=[10]

Appendix

R Code:

ANLY510-Project_final.R

Data:

County Level
Diabetes Data.csv