Life Expectancy Prediction

A Binary Classification Problem

Group 4: Tim Chen, Kahyun Jo, Valeria Lopez Robles, Hao Qiu, Michael Xu

Abstract

There has been a growing interest in using life expectancy not only because it is one of the crucial indicators of the overall health of a population, but it also plays a significant role in the health sector and the economy of a country. After utilizing the correlation matrix and forward and backward selections, we were able to select GDP from Consumption, Education, Calories from Plant, Obesity, Depression, Smoking and Sanitation Levels, and explore how these factors impact life expectancy. With the help of a generalized linear regression (logistic regression/binary classification)model, we conclude that life expectancy is positively correlated with GDP from consumption, and Safe Access to Sanitation, while negatively correlated with depression and . The model has an 90% accuracy in testing.

Data Source

01 **World Bank**

- Population

02 **United Nations Data**

- Lifetime
- GDP (per Capita)
- Education

03 **Our World in Data**

- Diet
- Mental
- Smoking
- Sanitation
- Obesity

Questions

Since we set Country and Year as combined unit, Keep in mind that all values in data is based on the unit of per country per year. This time we focus on longevity aspect(>72 Life Expectancy).

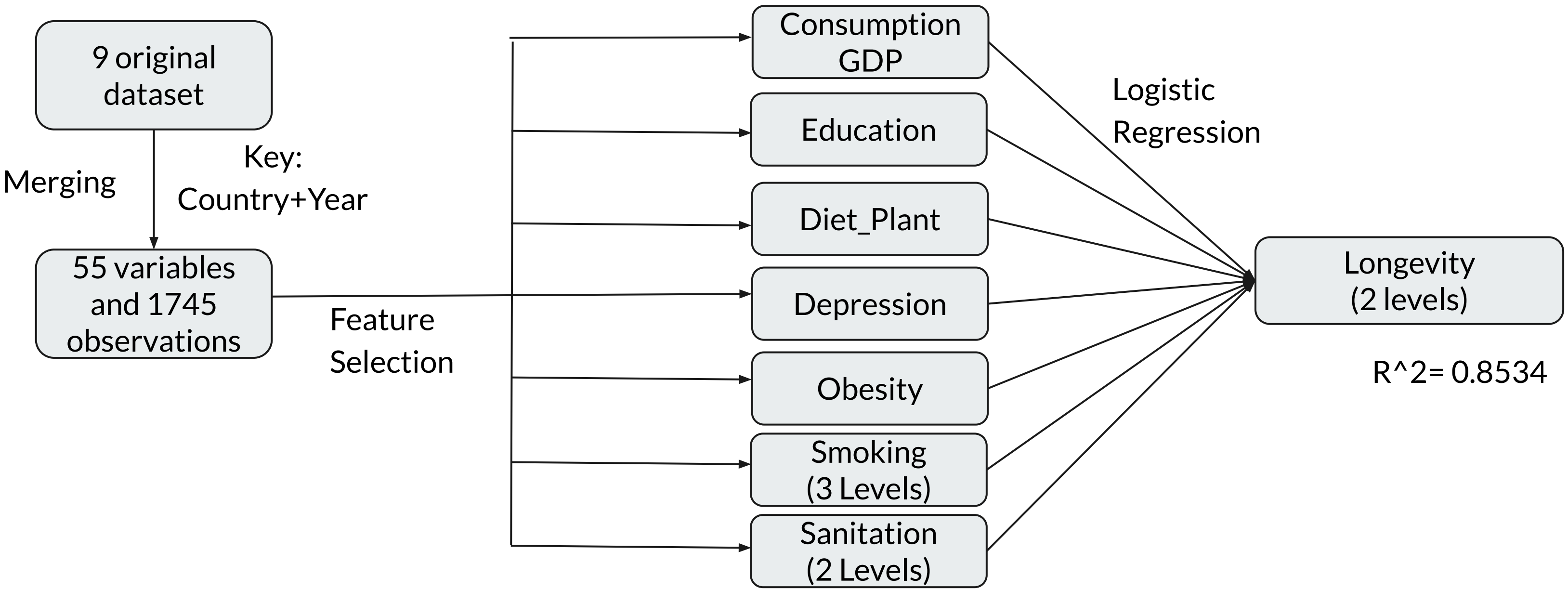
1. Scholarly articles have given some predictors that are said to impact the individual’s longevity. Is there a significant relationship between life expectancy greater than 72 and eg. obesity or amount of calories of Vegetable consumed in a country ?

2. What is the strongest predictor of a country’s overall Longevity?

3. How well are the percentage of daily smokers and sanitation accessibility as predictors for longevity?

4. How well can we predict on future data?

Schematic



9 original dataset

Merging

Key:
Country+Year

55 variables and 1745 observations

Feature Selection

Consumption GDP

Education

Diet_Plant

Depression

Obesity

Smoking
(3 Levels)

Sanitation
(2 Levels)

Logistic Regression

Longevity
(2 levels)

R^2= 0.8534

Research & Related Papers

- “Generally, wealthier countries have a higher average life expectancy than poorer countries [2,3,4], which can be argued, are achieved through higher standards of living, more effective **health systems**, and more resources invested in determinants of health (e.g. **sanitation, housing,** **education**) [5]” (Freeman etc. 2).

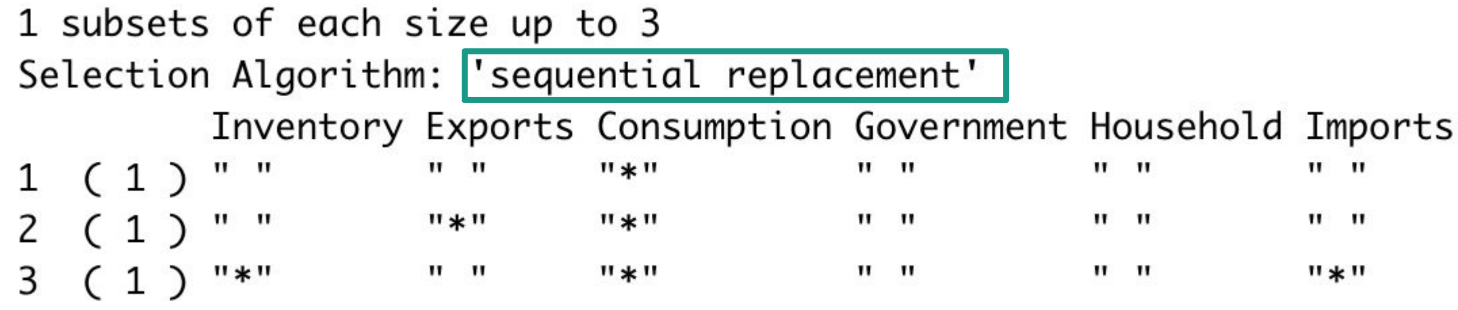
GDP per Capita

Access to Sanitation

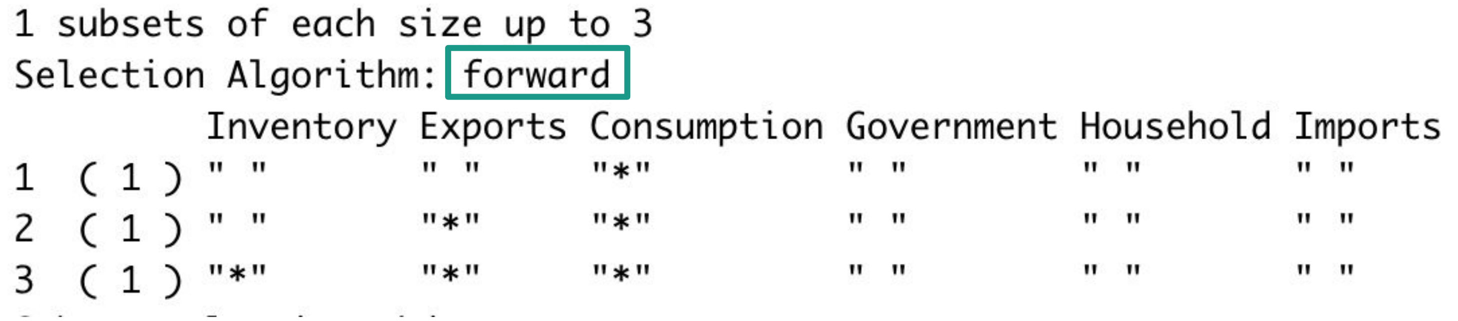
Years of Education

Forward/Backward Selection

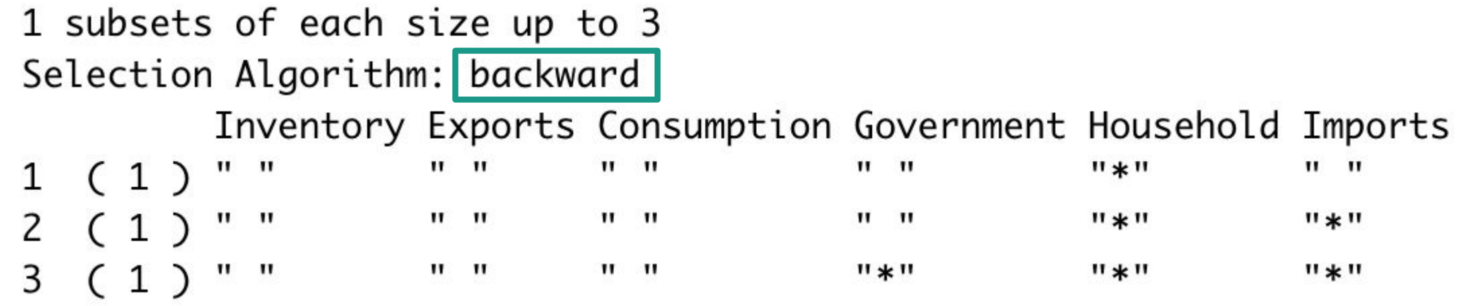
- Choosing between all the GDP factors
  - highly correlated
- function imported from leap library : regsubsets()
- When we set nvmax = 3, we use the following:
  - sequential replacement forward; backward
- Finally decided on using **GDP-Consumption** as the predictor variable for GDP



```
1 subsets of each size up to 3
Selection Algorithm: 'sequential replacement'
         Inventory Exports Consumption Government Household Imports
1  ( 1 ) " "       " "     "*"         " "        " "       " "
2  ( 1 ) " "       "*"     "*"         " "        " "       " "
3  ( 1 ) "*"       " "     "*"         " "        " "       "*"
```



```
1 subsets of each size up to 3
Selection Algorithm: forward
         Inventory Exports Consumption Government Household Imports
1  ( 1 ) " "       " "     "*"         " "        " "       " "
2  ( 1 ) " "       "*"     "*"         " "        " "       " "
3  ( 1 ) "*"       "*"     "*"         " "        " "       " "
```



```
1 subsets of each size up to 3
Selection Algorithm: backward
         Inventory Exports Consumption Government Household Imports
1  ( 1 ) " "       " "     " "         " "        "*"       " "
2  ( 1 ) " "       " "     " "         " "        "*"       "*"
3  ( 1 ) " "       " "     " "         "*"        "*"       "*"
```

Individual Factors

- Body mass index
- Systolic blood pressure
- Smoking
- Diabetes
- Depression
- Diet

Predictors

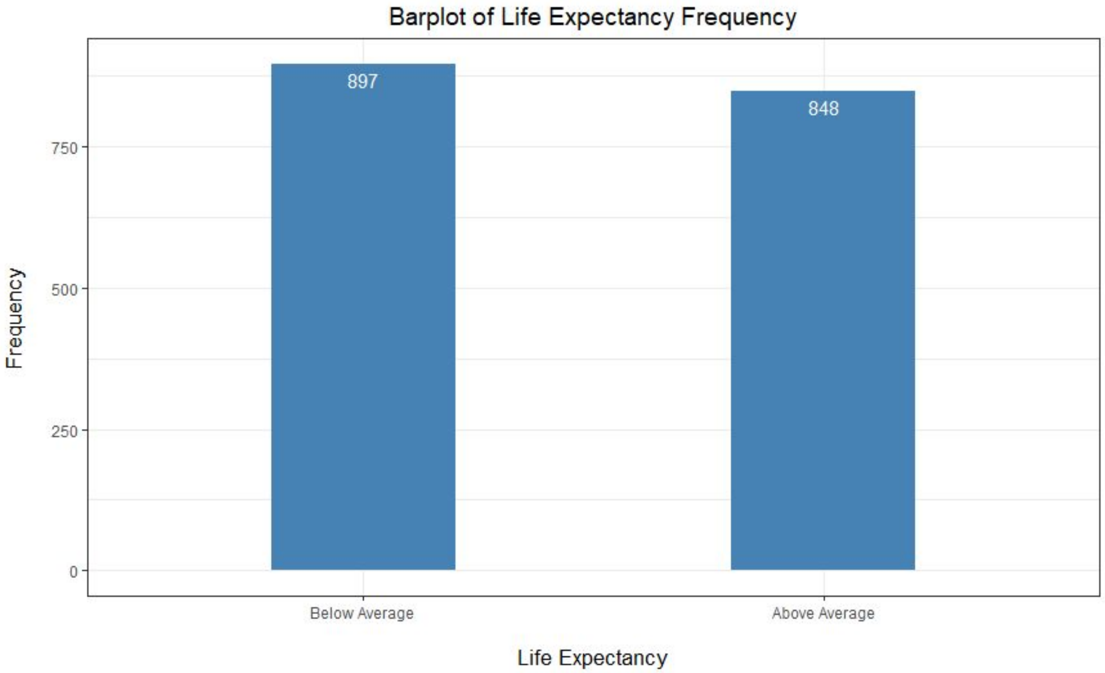
Overweight Population

Smoking Population

Mental Disease

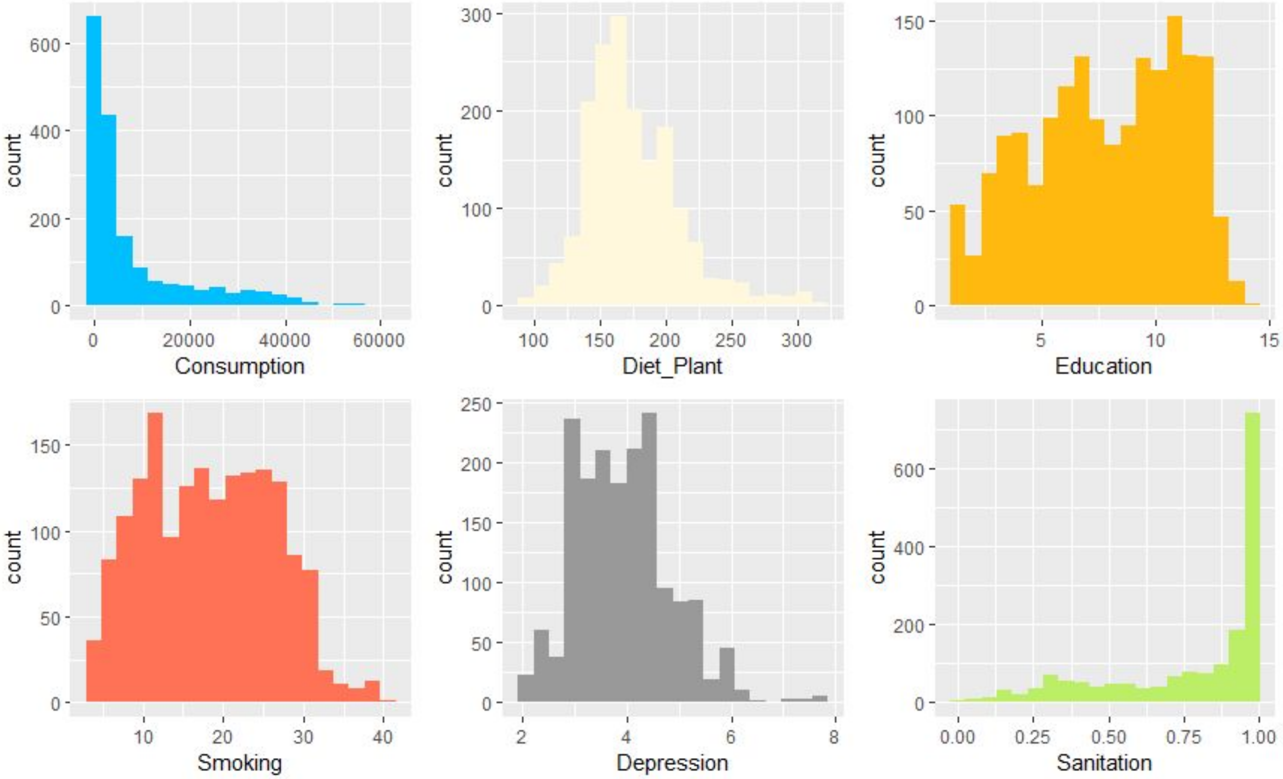
Diet Composition

Outcome Variable: Life_Expectancy



The frequency of each two level of the outcome variable is balanced because we divide the original lifetime column into the binary categorical variable proportionally based on the quantile statistics.

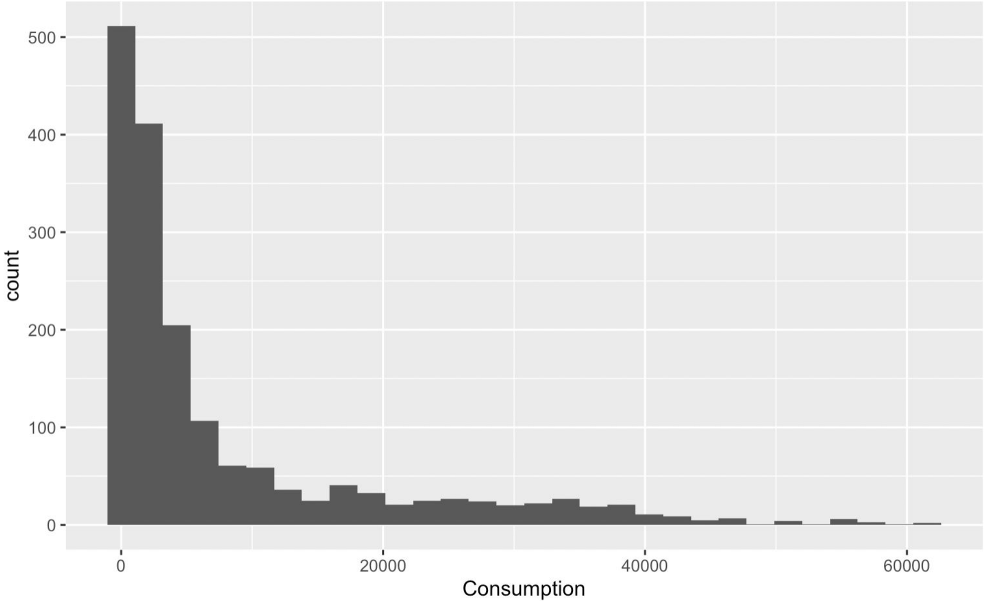
Numeric Variables Overview



Variable- Consumption

- Heavily right-skewed

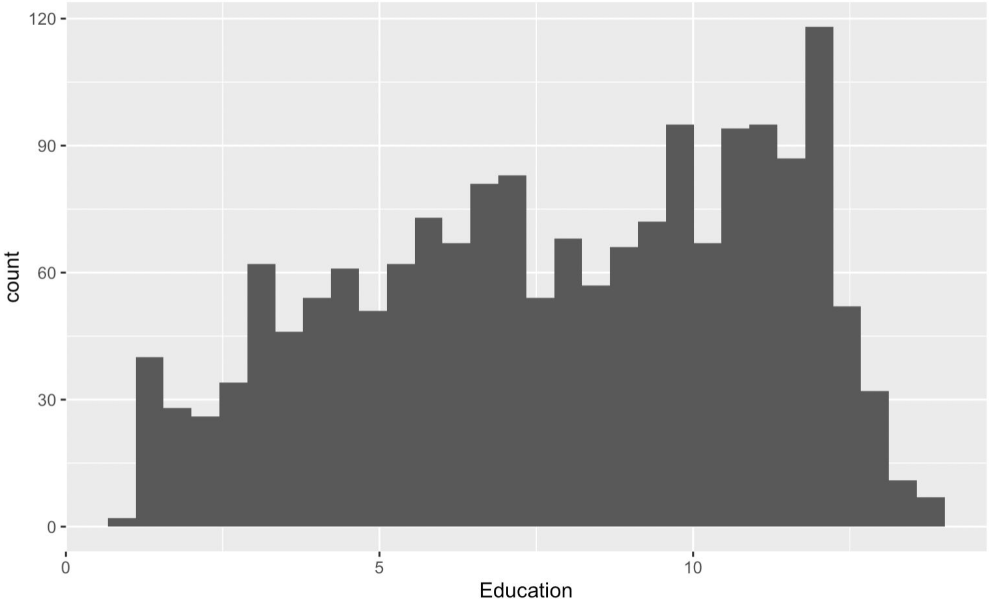
- Log Transformation



Variable - Education

- Follow Normal Distribution

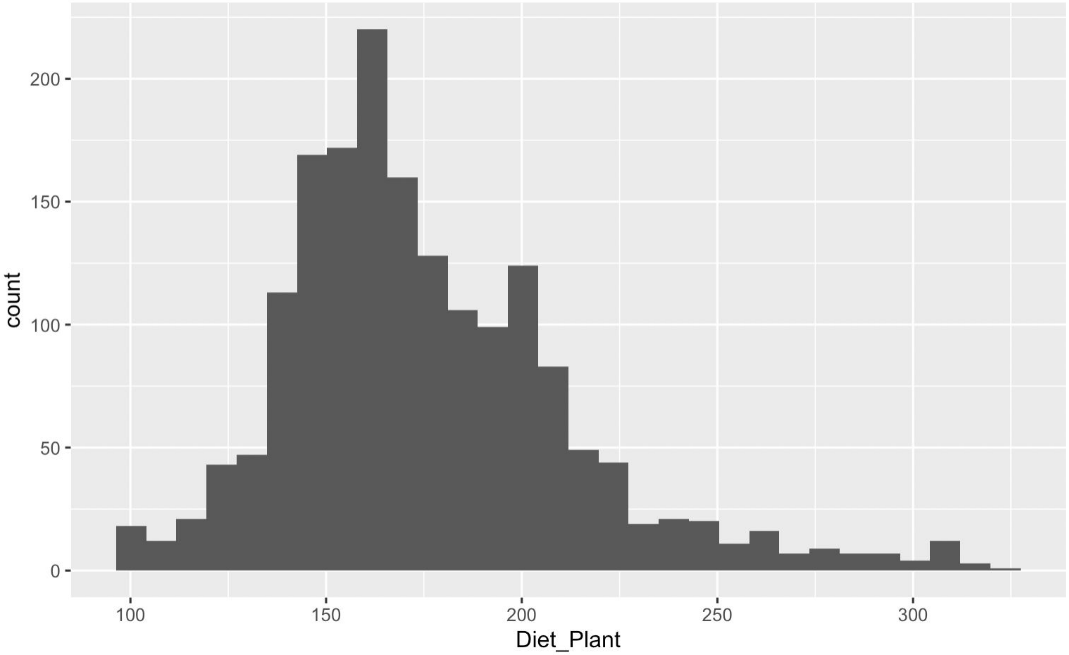
- No need to adjust



Variable - Diet_Plant

- A little right-skewed

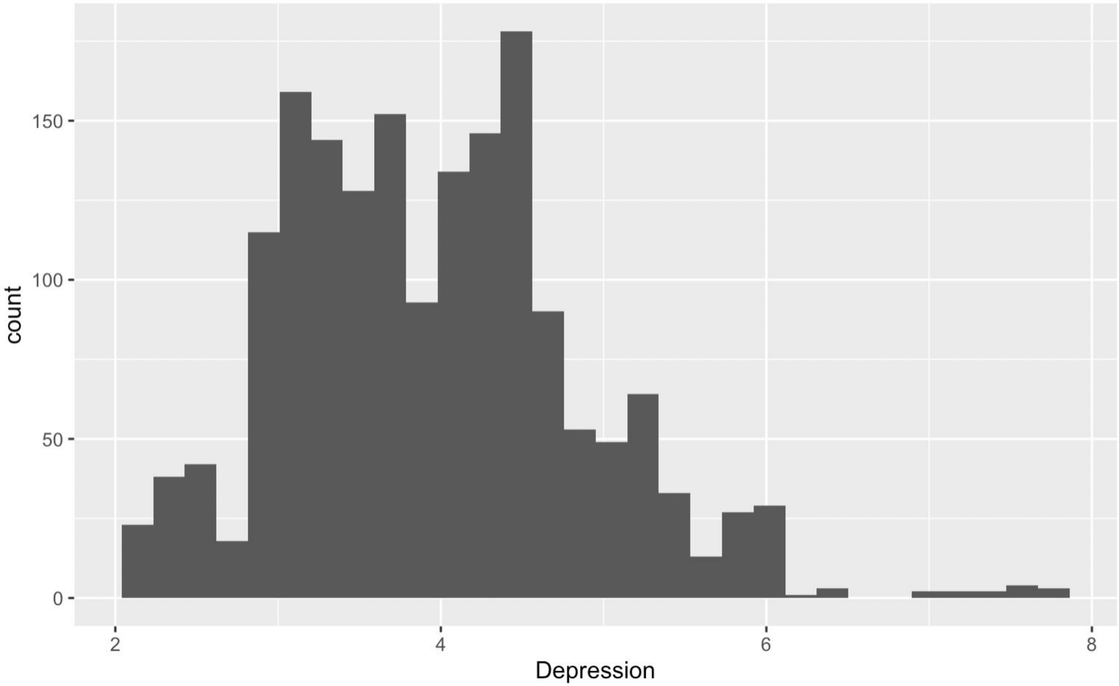
- keep as it is due to large # observations



Variable - Depression

- Right Skewed

- keep as it is due to large # observations

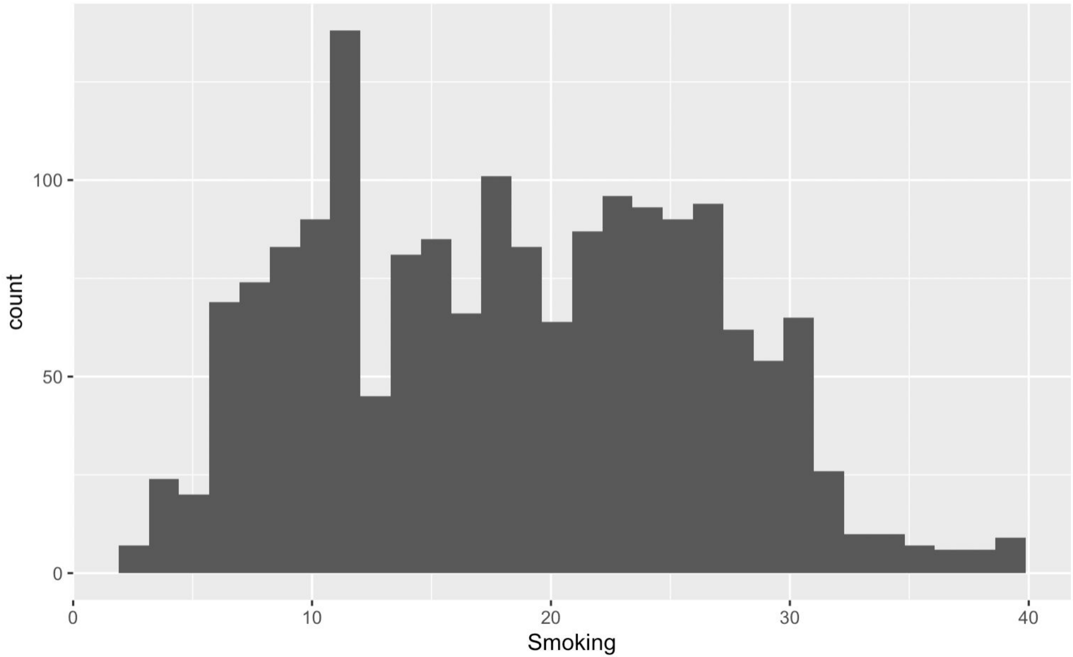


Categorical Variable: Smoking

- Roughly follow Normal Distribution

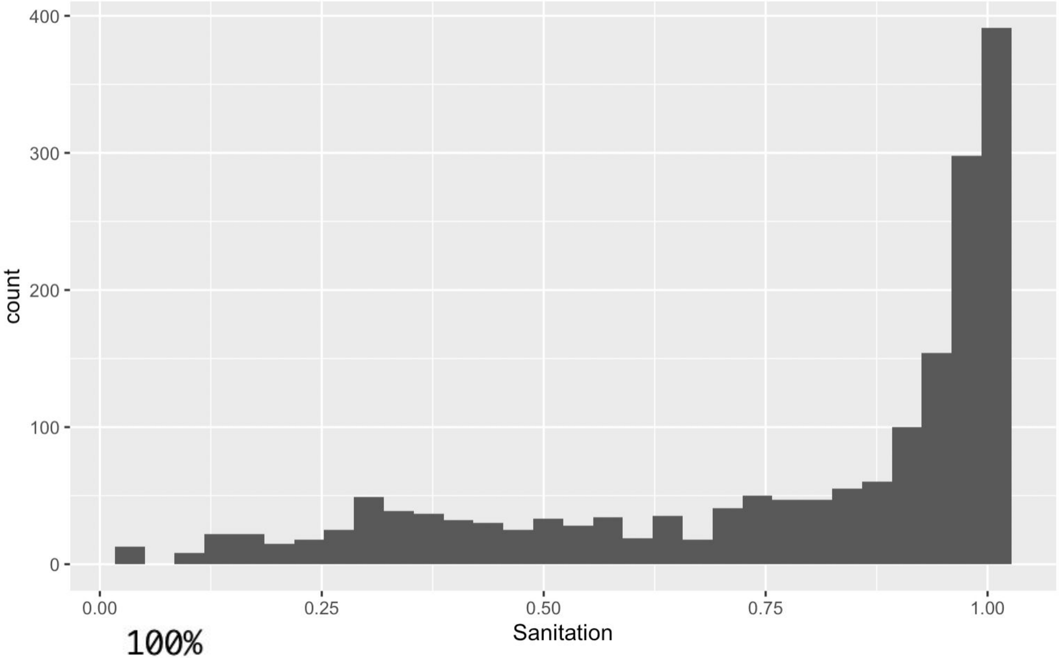
Smoking (3 levels)
Quantile Information:

```
 0%   33.33% 66.666%     100%
3.0     13.6     22.6    39.7
```



Categorical Variable: Sanitation

- Left-skewed

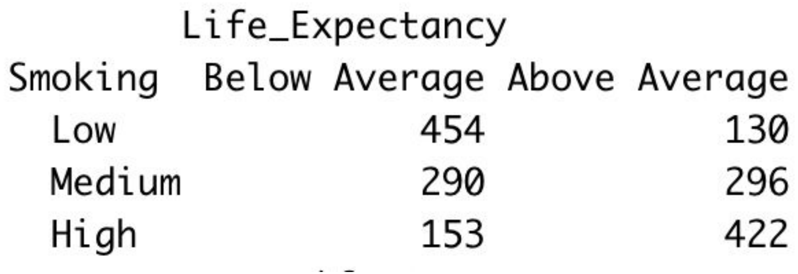


Sanitation (2 levels)
Quantile Information:

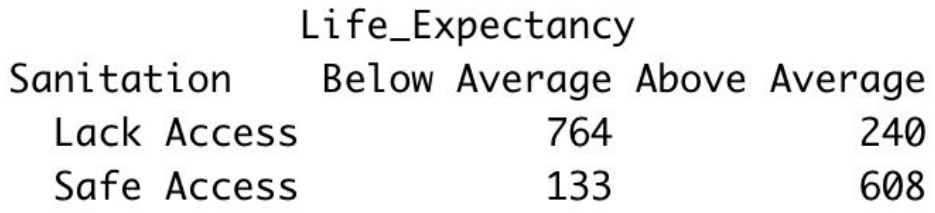
```
        0%        25%        50%        75%       100% 
0.02373481 0.59912743 0.91719534 0.98949089 1.00000000 
```

Contingency Tables

- Some numbers quite small in the second table that we need to be careful about

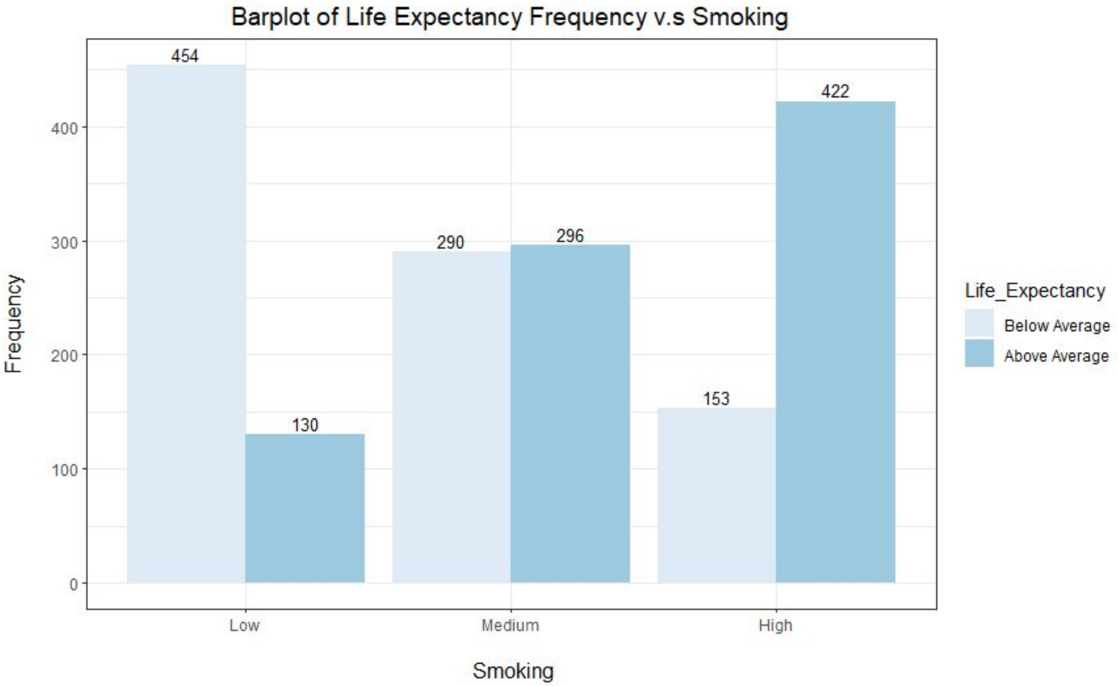


```
           Life_Expectancy
Smoking   Below Average Above Average
  Low               454           130
  Medium            290           296
  High              153           422
```



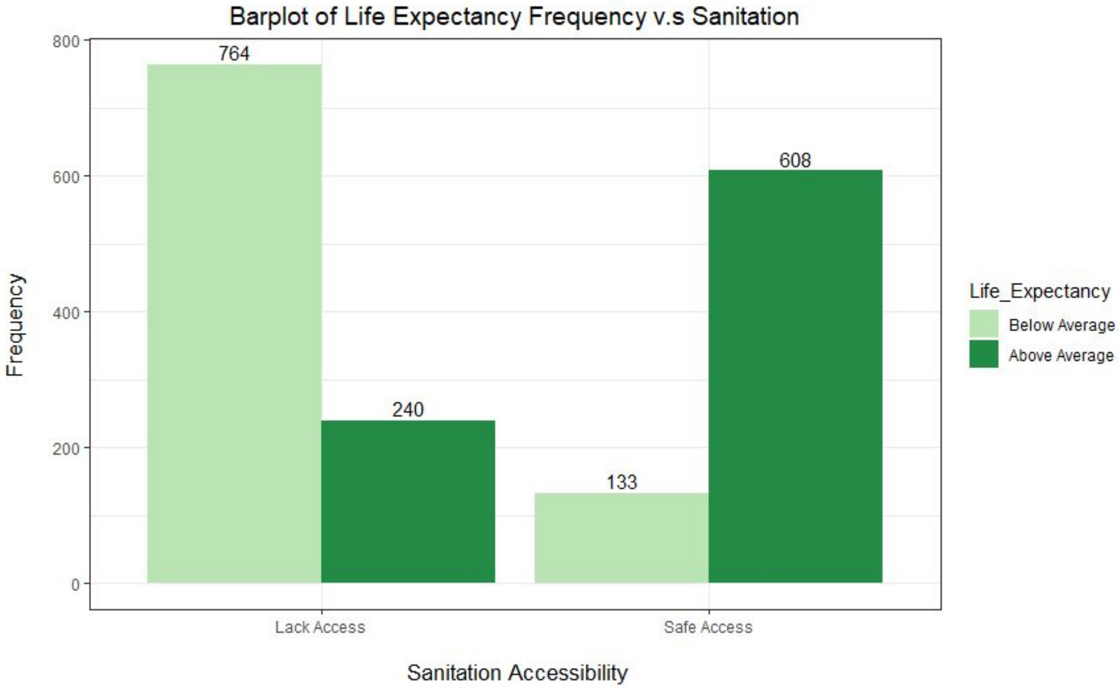
```
              Life_Expectancy
Sanitation     Below Average Above Average
  Lack Access            764           240
  Safe Access            133           608
```

Life Expectancy v.s Smoking



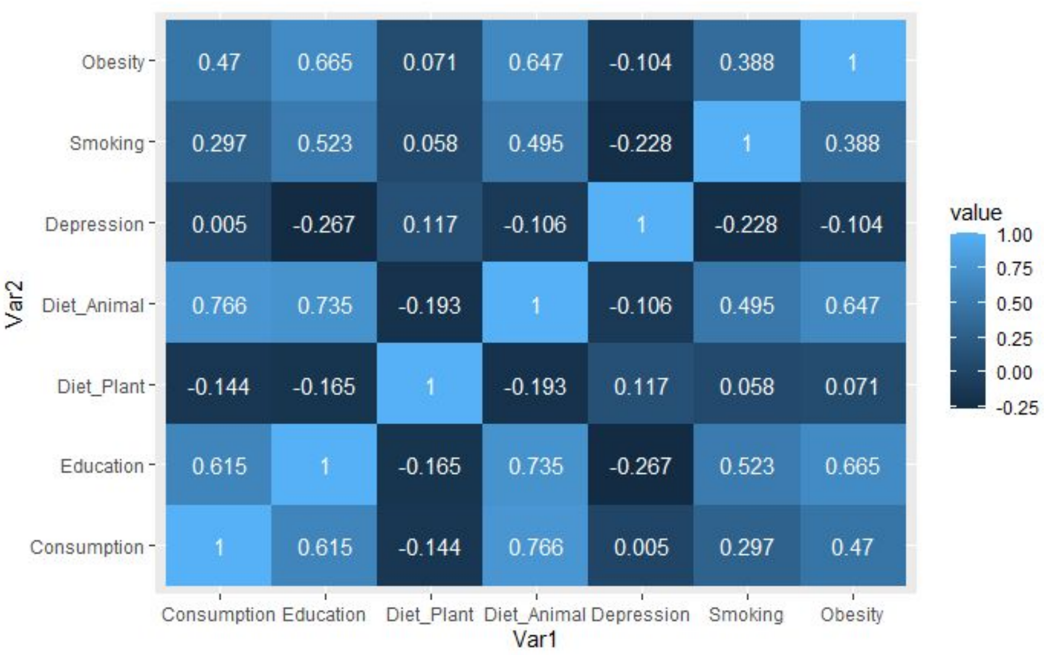
From the plot, we could observe that for country with low percentage of smokers, the proportion of below average life expectancy is very high. While for country with high percentage of smokers, most countries have above average life expectancy

Life Expectancy v.s Sanitation



From the plot, we could observe that for lack of sanitation access, the proportion of below average life expectancy is very high. While for safe sanitation access, the proportion of high life expectancy takes the major part.

Correlation Heatmap



	Consumption	Education	Diet_Plant	Diet_Animal	Depression	Smoking	Obesity
Obesity	0.47	0.665	0.071	0.647	-0.104	0.388	1
Smoking	0.297	0.523	0.058	0.495	-0.228	1	0.388
Depression	0.005	-0.267	0.117	-0.106	1	-0.228	-0.104
Diet_Animal	0.766	0.735	-0.193	1	-0.106	0.495	0.647
Diet_Plant	-0.144	-0.165	1	-0.193	0.117	0.058	0.071
Education	0.615	1	-0.165	0.735	-0.267	0.523	0.665
Consumption	1	0.615	-0.144	0.766	0.005	0.297	0.47

This is the correlation heatmap for all chosen variables. From this plot, we can see that the correlations between some of the variables are quite high (e.g. Consumption & Diet_Animal, close to 1.0). Therefore, we decide to delete Diet_Animal and we need to pay attention to other variables with higher correlation when training the models to avoid multicollinearity.

Model 1

Lifetime = sigmoid(

β_0 + β_1*Consumption + β_2*Education + β_3*Diet_Plant +
β_4*Depression + β_5*Obesity + β_6*Smoking_Med +
β_7*Smoking_High + β_8*Sanitation_Safe_Access

)

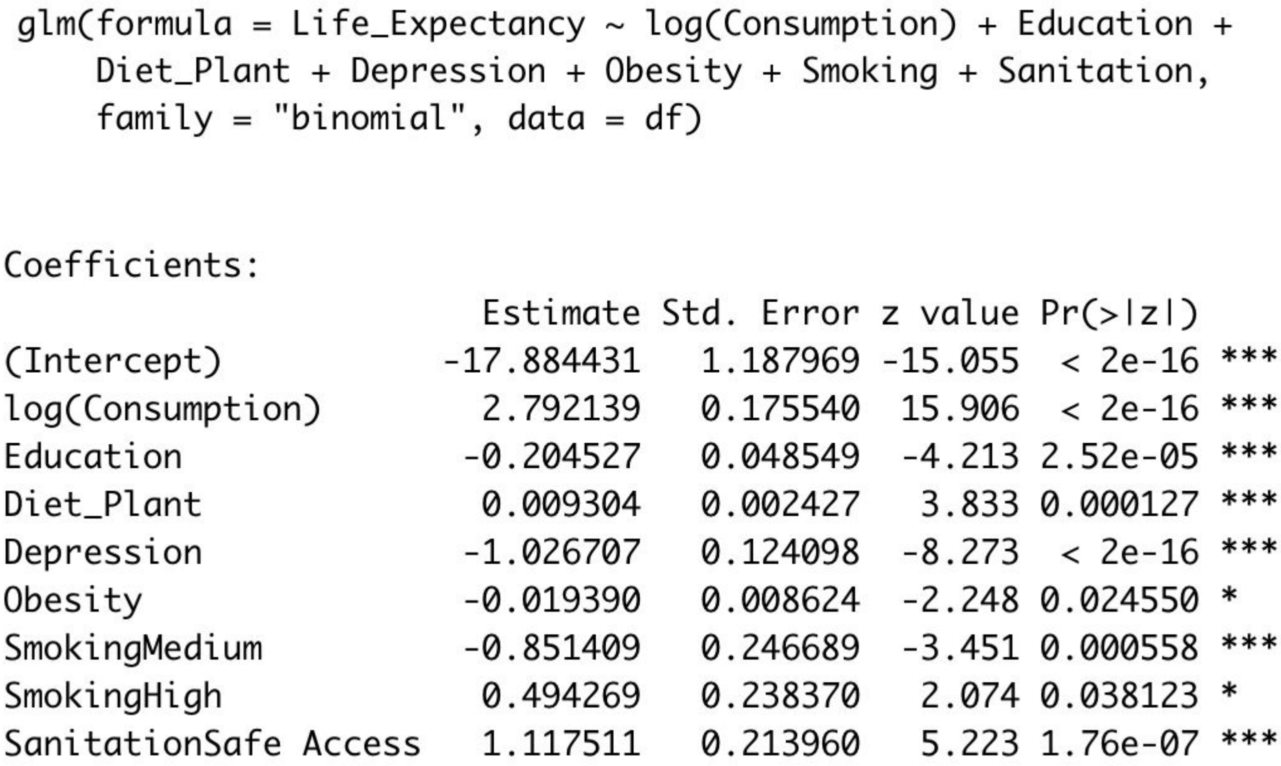
Predictors
—Numerical
—Categorical

Model 1 Name Matrix

Variable Name	Type	Meaning (per country per year)
Consumption	Numerical	GDP of Consumption per capita
Education	Numerical	Average years of schooling
Diet_plant	Numerical	Average daily per capita supply of Calories from plant-based food
Depression	Numerical	Percent of Adults with depression disorder
Obesity	Numerical	Percent of population whose BMI >= 25
Smoking	Categorical	Percent of Daily Smoker : “High”, “Med”, “Low”
Sanitation	Categorical	Percent of People with levels of access to Safe Sanitation : “High”, “Med”, “Low”

Model 1

- Obesity and High Level of Smoking are not as significant



```
glm(formula = Life_Expectancy ~ log(Consumption) + Education + 
    Diet_Plant + Depression + Obesity + Smoking + Sanitation, 
    family = "binomial", data = df)
```

```
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)           -17.884431   1.187969 -15.055  < 2e-16 ***
log(Consumption)        2.792139   0.175540  15.906  < 2e-16 ***
Education              -0.204527   0.048549  -4.213 2.52e-05 ***
Diet_Plant              0.009304   0.002427   3.833 0.000127 ***
Depression             -1.026707   0.124098  -8.273  < 2e-16 ***
Obesity                -0.019390   0.008624  -2.248 0.024550 *  
SmokingMedium          -0.851409   0.246689  -3.451 0.000558 ***
SmokingHigh             0.494269   0.238370   2.074 0.038123 *  
SanitationSafe Access   1.117511   0.213960   5.223 1.76e-07 ***
```

VIF: Model 1

```
vif(glm(Life_Expectancy ~ log(Consumption)+
Education + Diet_Plant + Depression + Obesity,
df, family="binomial"))
```

log(Consumption)	Education	Diet_Plant	Depression	Obesity
2.008312	1.585473	1.269884	1.300972	1.647141

- Removed the categorical variables: smoking, sanitation
- The VIF is smaller than 5 for all four predictors. Thus, the associated regression coefficients are NOT poorly estimated due to multicollinearity

Modeling Methodology

- Based on Scholarly articles about those significant factors impacting individual’s health and longevity, we find related datasets to explore

- We use selection techniques(fwd,bwd) along with Confusion Matrix to eliminate possible highly correlated/confounding factors. Use VIF to double check to prevent multicollinearity

- We add interaction terms between categorical variables in the model.

Final Model

$$p = P(\text{Life\_Expectancy} = \text{“Below 72”})$$

$$\text{Log}(p/(1-p)) =$$

$$\beta_0 + \beta_1{}^*\text{Consumption} + \beta_2{}^*\text{Education} + \beta_3{}^*\text{Diet\_Plant} +$$
$$\beta_4{}^*\text{Obesity} + \beta_5{}^*\text{Depression} + \beta_6{}^*\text{Smoking\_Med}$$

$$+ \beta_7{}^*\text{Smoking\_High} + \beta_8{}^*\text{Sanitation\_SafeAccess}$$

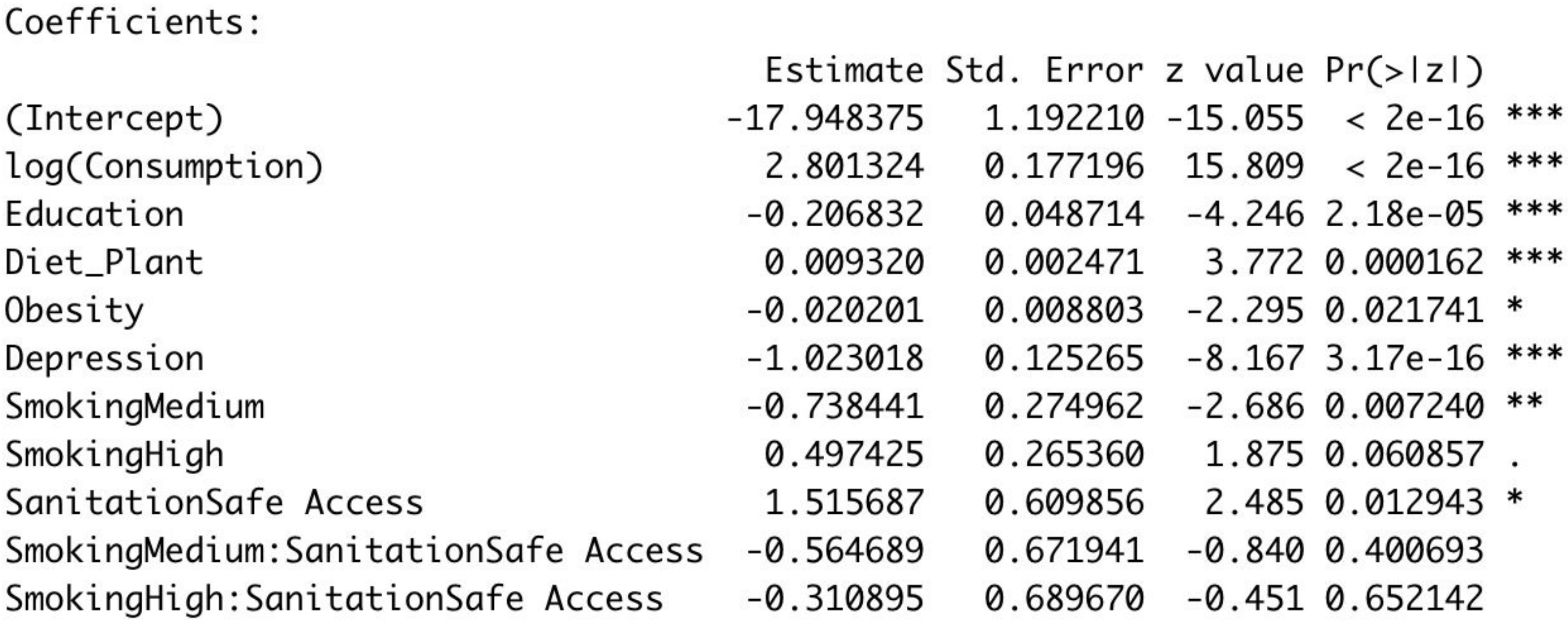
$$+\beta_9{}^*\text{Smoking\_Med*Sanitation\_SafeAccess} +$$
$$\beta_{10}{}^*\text{Smoking\_High*Sanitation\_SafeAccess}$$

Predictors
—Numerical
—Categorical
—Interaction

Final Model Name Matrix

Variable Name	Type	Meaning (per country per year)
Consumption	Numerical	GDP of Consumption per capita
Education	Numerical	Average years of schooling
Diet_plant	Numerical	Average daily per capita supply of Calories from plant-based food
Depression	Numerical	Percent of Adults with depression disorder
Obesity	Numerical	Percent of population whose BMI >= 25
Smoking	Categorical	Percent of Daily Smoker : “High”, “Med”, “Low”
Sanitation	Categorical	Percent of People with levels of access to Safe Sanitation : “High”, “Med”, “Low”

Final Model Summary



```
Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                         -17.948375   1.192210 -15.055  < 2e-16 ***
log(Consumption)                      2.801324   0.177196  15.809  < 2e-16 ***
Education                            -0.206832   0.048714  -4.246 2.18e-05 ***
Diet_Plant                            0.009320   0.002471   3.772 0.000162 ***
Obesity                              -0.020201   0.008803  -2.295 0.021741 *
Depression                           -1.023018   0.125265  -8.167 3.17e-16 ***
SmokingMedium                        -0.738441   0.274962  -2.686 0.007240 **
SmokingHigh                           0.497425   0.265360   1.875 0.060857 .
SanitationSafe Access                 1.515687   0.609856   2.485 0.012943 *
SmokingMedium:SanitationSafe Access  -0.564689   0.671941  -0.840 0.400693
SmokingHigh:SanitationSafe Access    -0.310895   0.689670  -0.451 0.652142
```

Model Interpretation

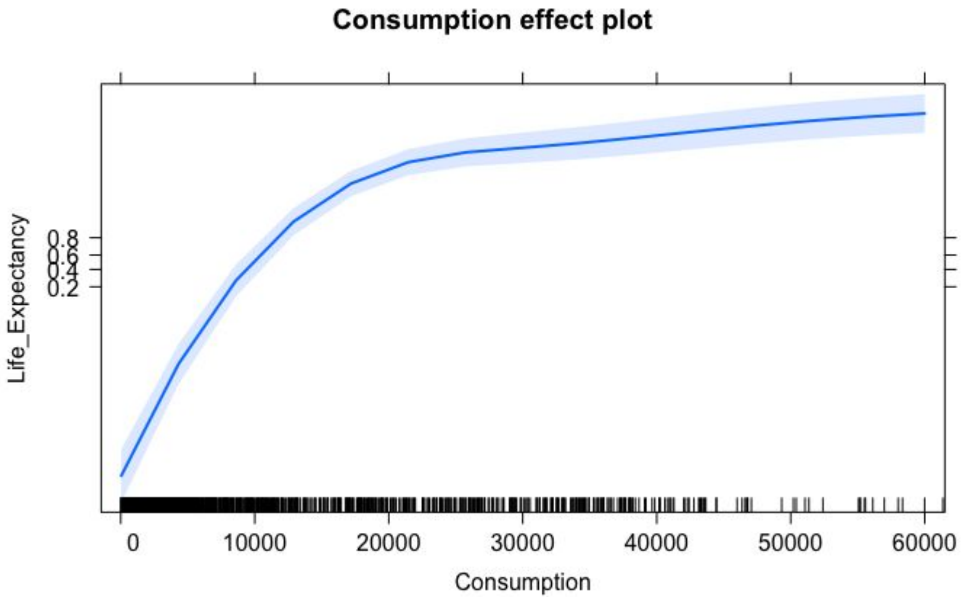
Lifetime = sigmoid(

-17.95 + 2.80 * log(Consumption) - 0.21 * Education

+ 0.01 * Diet_Plant - 0.02 * Obesity - 1.02 * Depression
- 0.74 * Smoking_Med
+ 0.50* Smoking_High + 1.51 * Sanitation_SafeAccess
- 0.56* Smoking_Med*Sanitation_SafeAccess
- 0.31 * Smoking_High*Sanitation_SafeAccess)

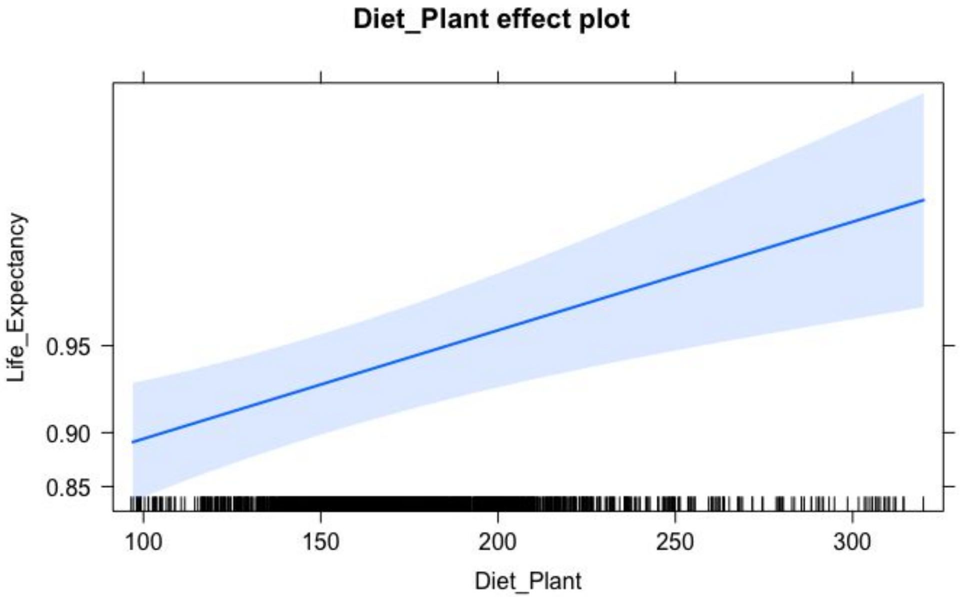
Effect Plot - Consumption

Generally, GDP from consumption shows an increasing relationship with life expectancy. However, the connection between GDP from consumption and life expectancy weakens after the consumption reaches a certain level, around 25000.



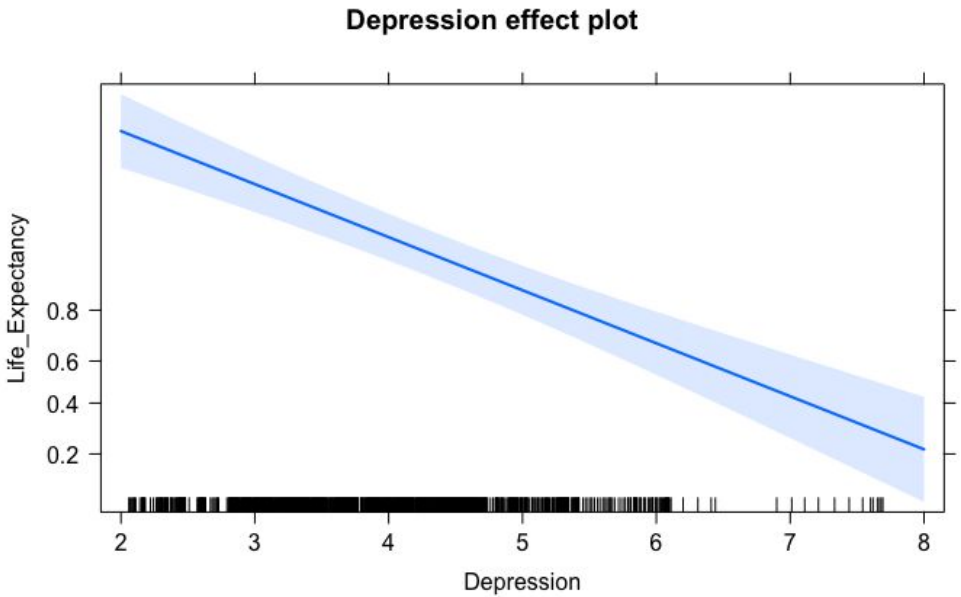
Effect Plot -Diet_Plant

Those who consume more vegetables are more likely to live longer.



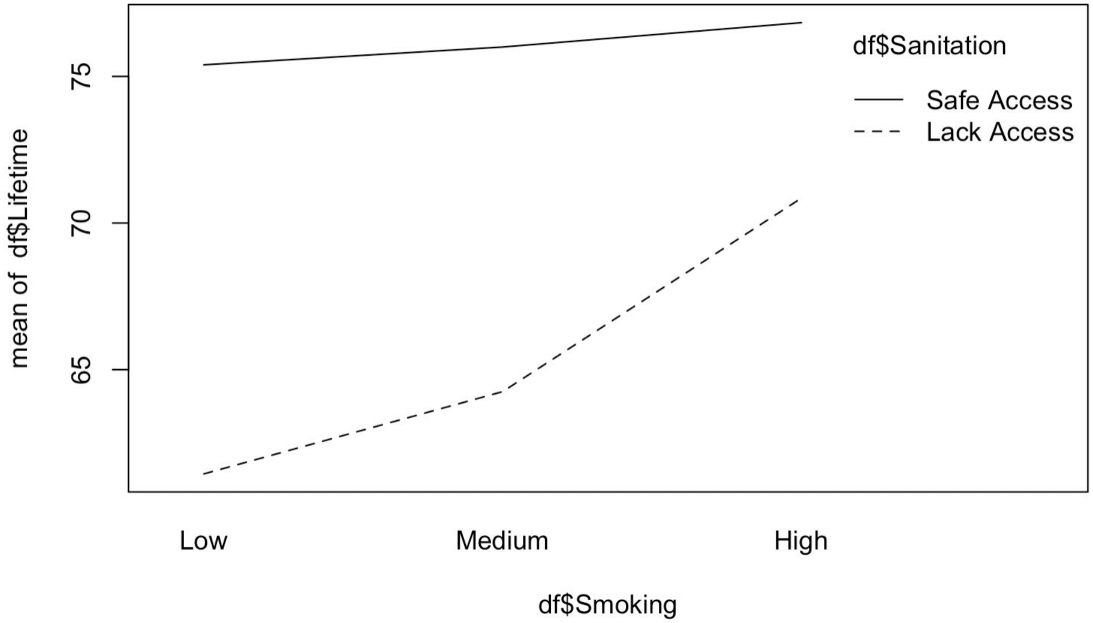
Effect Plot - Depression

A higher level of depression results in a reduction of life expectancy.



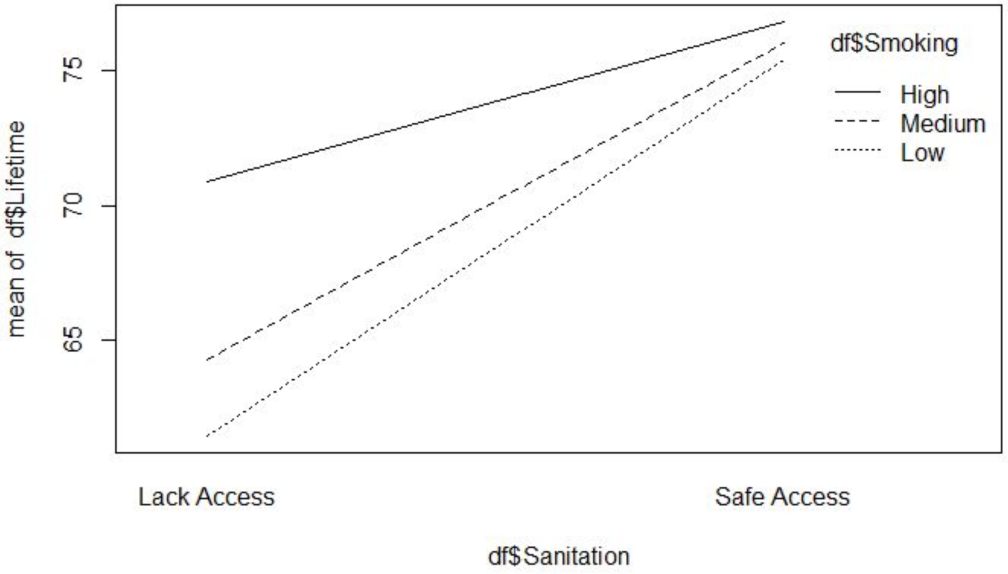
Interaction Plot

- In general, Safe access to Sanitation has a mean higher than Lack Access no matter what the Smoking level is.



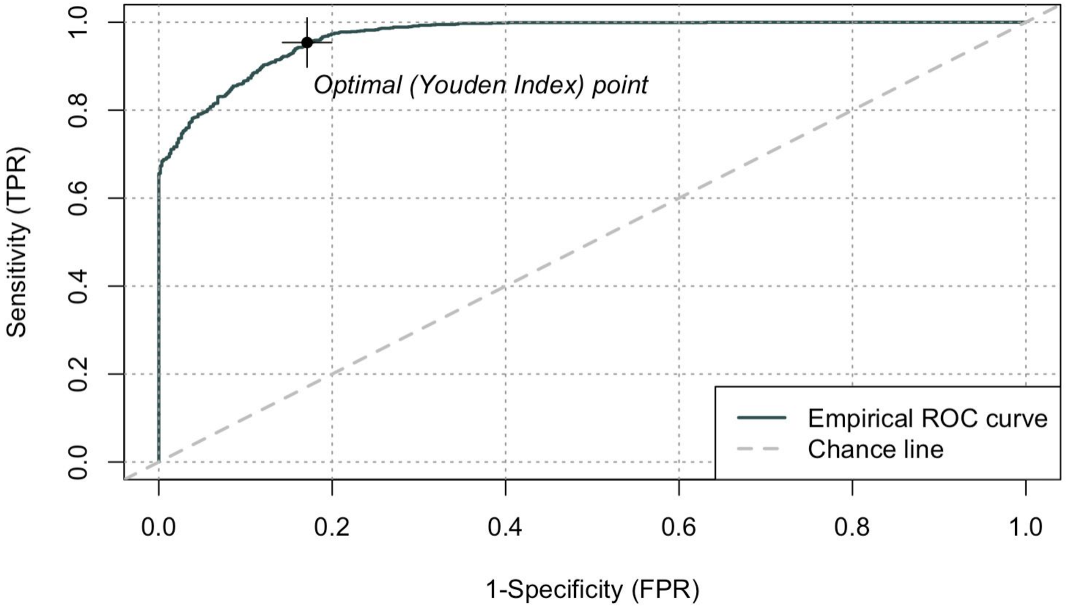
Interaction Plot

- In general, Safe access to Sanitation is a good indicator of a higher mean life expectancy, independent of percentage level of daily smokers.



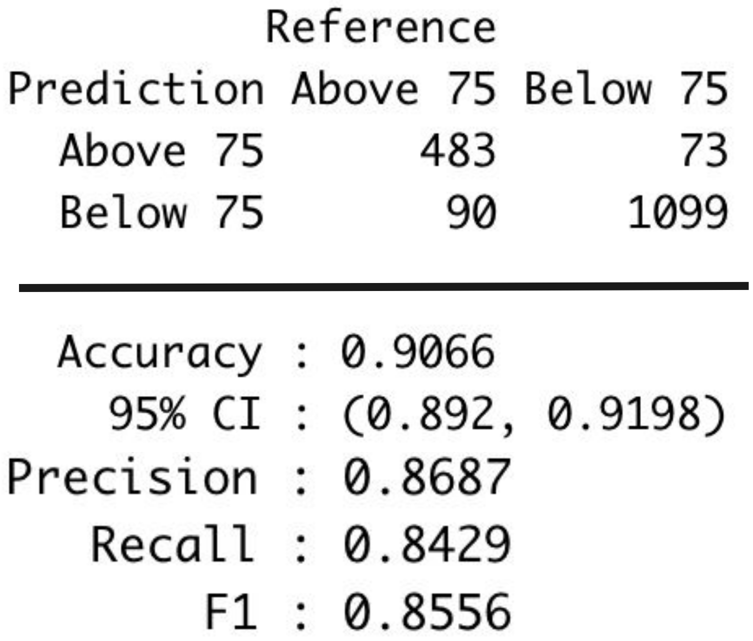
Measure of Accuracy

- ROC curve
  - AUC is 0.9688
- The model is highly capable of distinguishing between classes
- The higher the AUC score, the better the classification of the predicted values is.



Measure of Accuracy

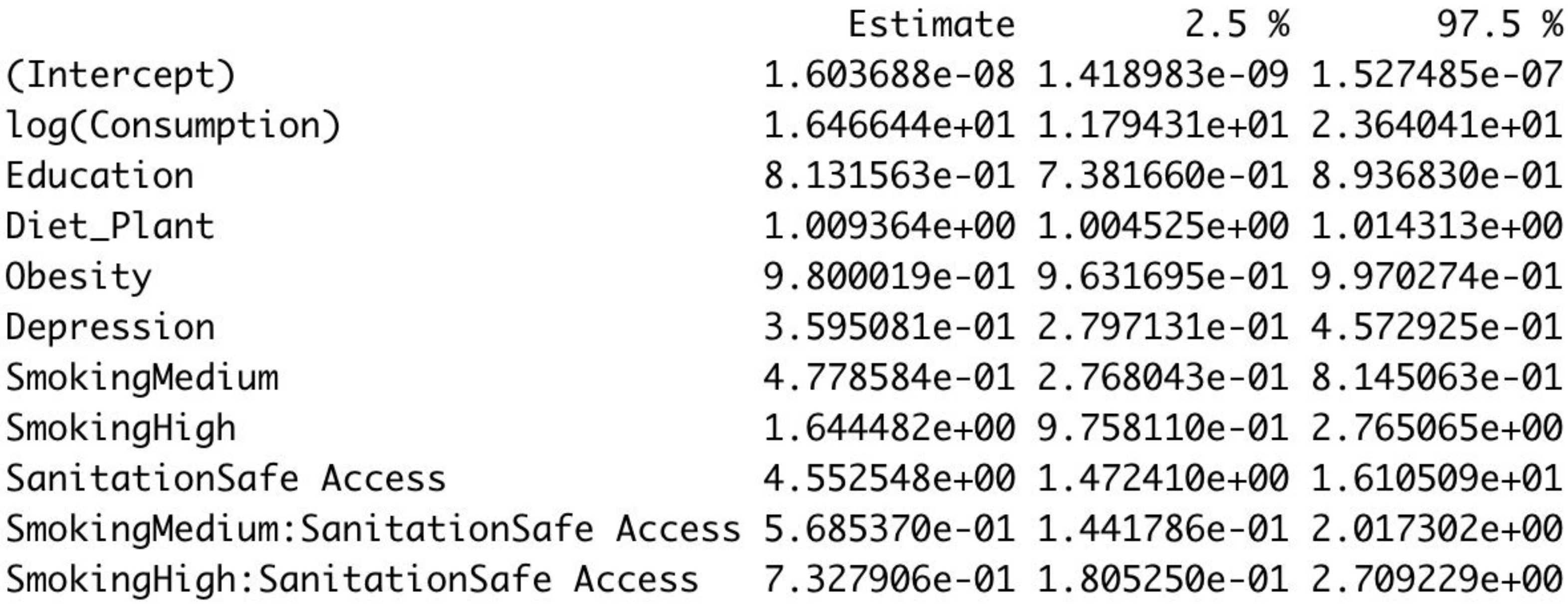
- Confusion Matrix
  - Diagonal High
- Accuracy: 90.66%
- Other Metrics for Reference



```
           Reference
Prediction Above 75 Below 75
  Above 75      483       73
  Below 75       90     1099
```

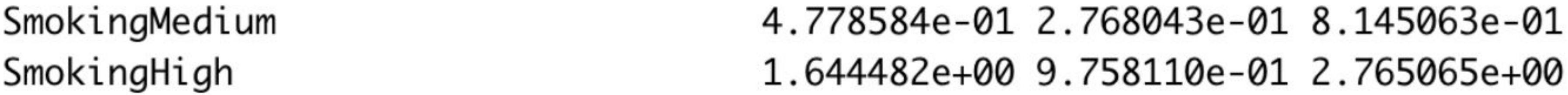
```
  Accuracy : 0.9066
    95% CI : (0.892, 0.9198)
Precision : 0.8687
    Recall : 0.8429
        F1 : 0.8556
```

Model Interpretation (Exponentiate Coefficient)



```
                                        Estimate        2.5 %       97.5 %
(Intercept)                         1.603688e-08 1.418983e-09 1.527485e-07
log(Consumption)                    1.646644e+01 1.179431e+01 2.364041e+01
Education                           8.131563e-01 7.381660e-01 8.936830e-01
Diet_Plant                          1.009364e+00 1.004525e+00 1.014313e+00
Obesity                             9.800019e-01 9.631695e-01 9.970274e-01
Depression                          3.595081e-01 2.797131e-01 4.572925e-01
SmokingMedium                       4.778584e-01 2.768043e-01 8.145063e-01
SmokingHigh                         1.644482e+00 9.758110e-01 2.765065e+00
SanitationSafe Access               4.552548e+00 1.472410e+00 1.610509e+01
SmokingMedium:SanitationSafe Access 5.685370e-01 1.441786e-01 2.017302e+00
SmokingHigh:SanitationSafe Access   7.327906e-01 1.805250e-01 2.709229e+00
```

Model Interpretation



```
SmokingMedium                    4.778584e-01 2.768043e-01 8.145063e-01
SmokingHigh                      1.644482e+00 9.758110e-01 2.765065e+00
```

Keeping all other variable constant, the odds of having a life expectancy above 72 is 0.48 times lower for countries with medium percentage of Smokers than those with low percentage of Smokers.

Keeping all other variable constant, the odds of having a life expectancy above 72 is 1.64 times higher for countries with high percentage of Smokers than those with low percentage of Smokers.

Explanation for Countries with high percentage of smokers having high odds of longer life expectancy

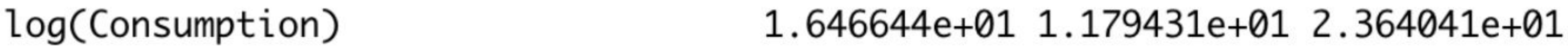
The model shows that the country with high amount of daily smokers actually has higher odds of longer life expectancy, which seems counterintuitive. The timing and the impact of the smoking epidemic vary by country, and this can possibly explain why. The countries with medium percentage of smokers may have already passed their smoking epidemic, which would already affect life expectancy negatively, given that smoking-related diseases and deaths appear usually 30-40 years later. The countries with high percentage of smokers may have experienced not long ago or currently experiencing the smoking epidemic. In this case, the resulting outcome of smoking will come in years later.

The Role of Smoking in Country Differences in Life Expectancy Across Europe, 1985–2014

Author: Fanny Janssen, PhD

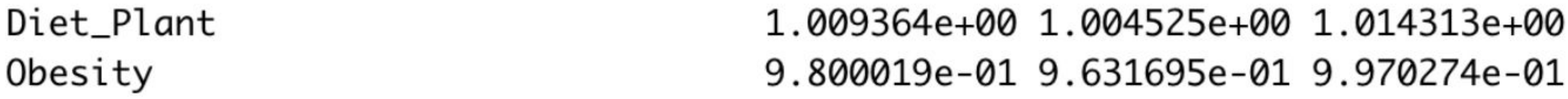
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7789949/

Model Interpretation



```
log(Consumption)                    1.646644e+01 1.179431e+01 2.364041e+01
```

Keeping all other variable constant, the odds of having a life expectancy above 72 increases by 1.64 times when log of Consumption in GDP increases by 1%



```
Diet_Plant                  1.009364e+00 1.004525e+00 1.014313e+00
Obesity                     9.800019e-01 9.631695e-01 9.970274e-01
```

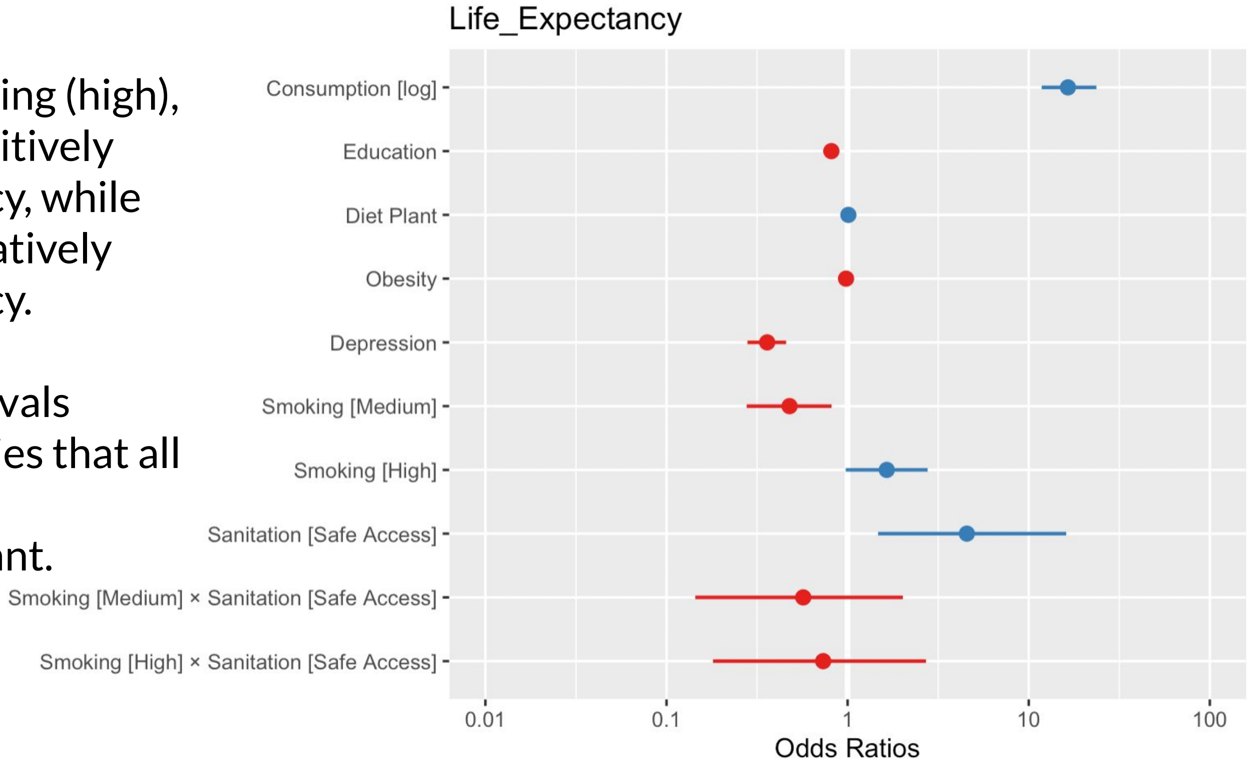
Keeping all other variable constant, the odds of having a life expectancy above 72 increases by 1% when Calories from plant consumed in a country increases by 1%.

Keeping all other variable constant, the odds of having a life expectancy above 72 is 2% less when percent of overweight population increases by 1%.

Plot of Odds

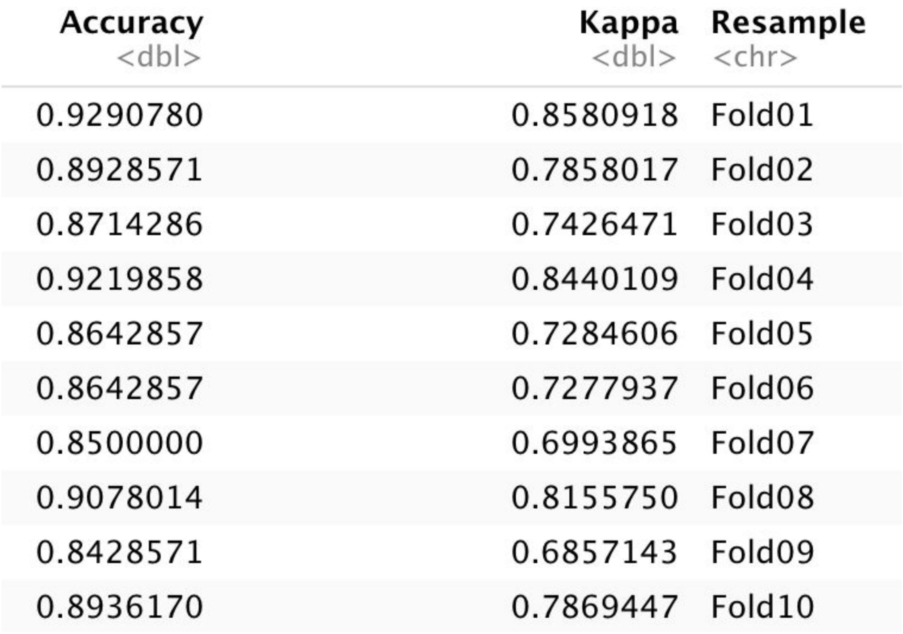
- Consumption, smoking (high), and sanitization positively affect life expectancy, while other variables negatively affect life expectancy.

- No confidence intervals crossed, which implies that all of the variables are statistically significant.



Test for Overfitting

- We use 10-fold Cross-Validation
- The resulting Accuracy for each resampled fold ranges from 85-92%

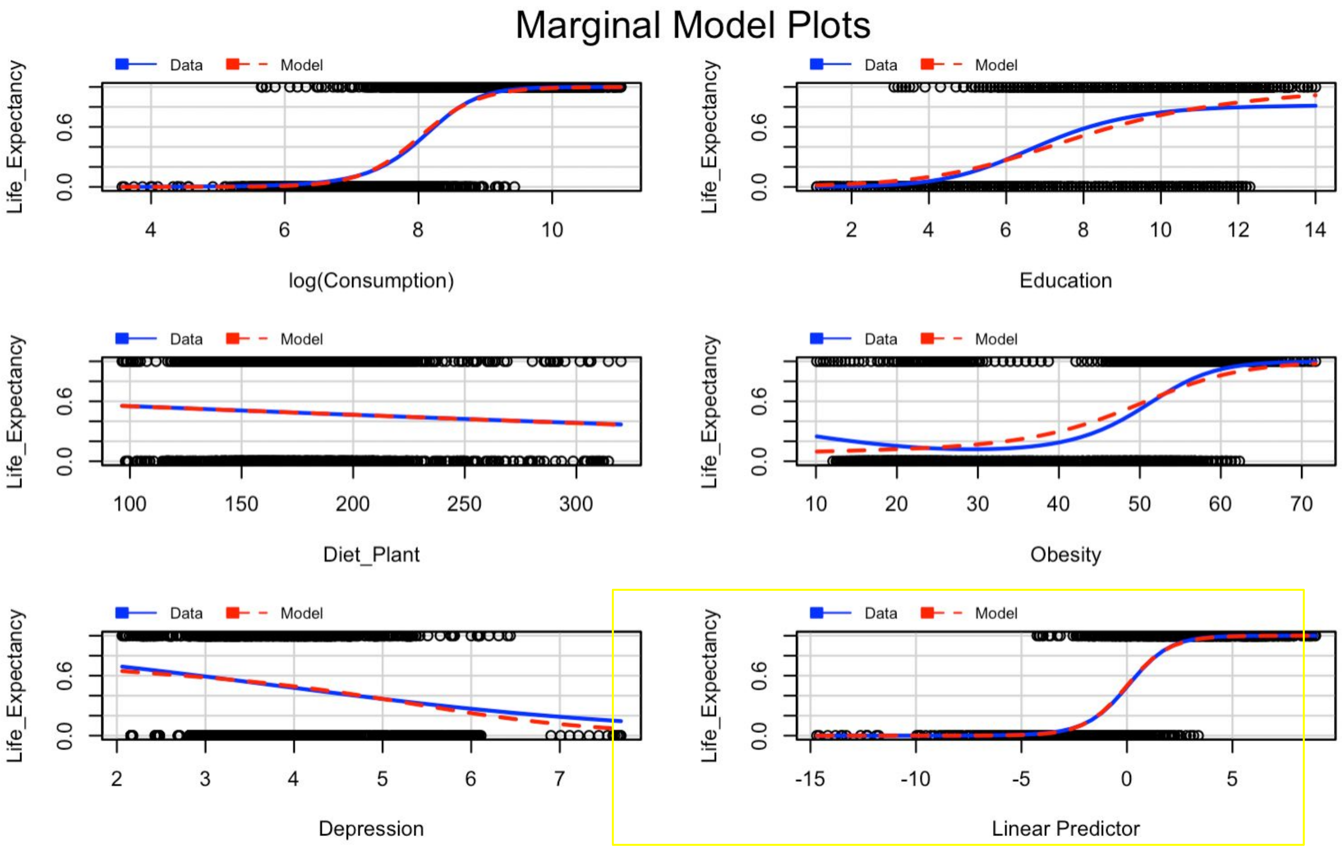


Accuracy	Kappa	Resample
<dbl>	<dbl>	<chr>
0.9290780	0.8580918	Fold01
0.8928571	0.7858017	Fold02
0.8714286	0.7426471	Fold03
0.9219858	0.8440109	Fold04
0.8642857	0.7284606	Fold05
0.8642857	0.7277937	Fold06
0.8500000	0.6993865	Fold07
0.9078014	0.8155750	Fold08
0.8428571	0.6857143	Fold09
0.8936170	0.7869447	Fold10

MMP

- MMP for consumption, diet, and depression represents our real data pretty well.

- MMP for education and obesity show that deviations between our data and the model do exist.
    - For example, it becomes more difficult to predict life expectancy when a person's education level is above 6.



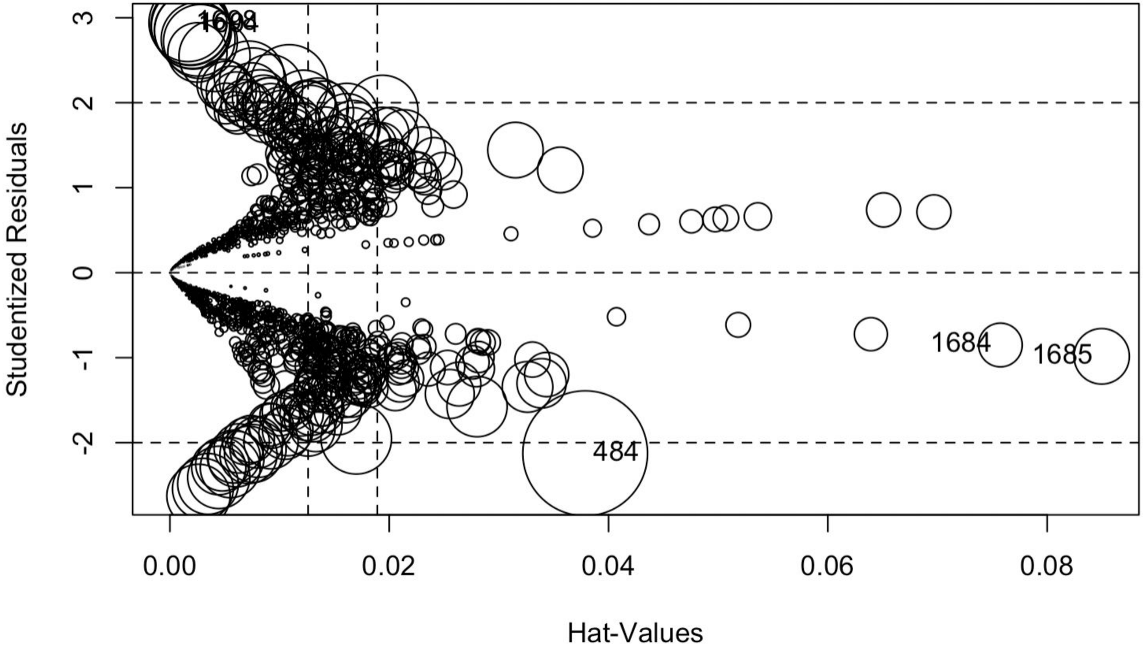
Influential Plot

- y-axis indicates how unusual a data point is ; the x-axis shows its leverage on the coefficient
- the size of the bubbles in the plot indicates the Cook's D value of each data points
- the plot used to visually assess heteroscedasticity

point 484 has a high standardized residual and leverage, which is a point we should worry about

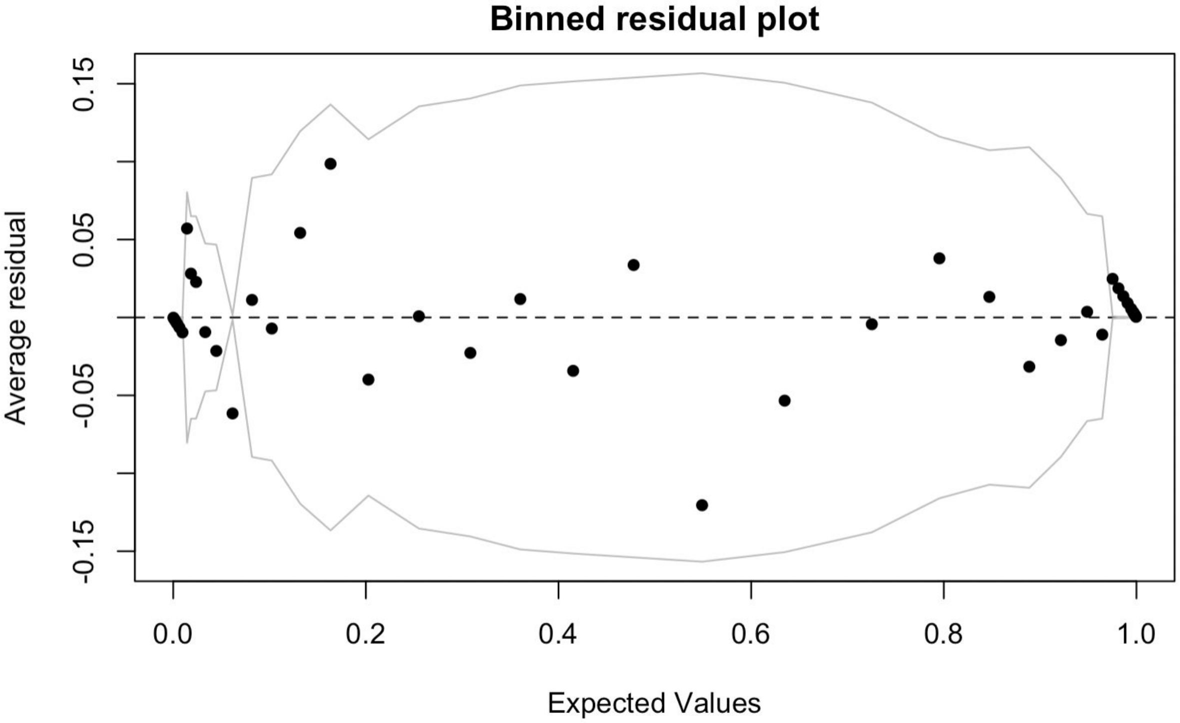
points 1684 and 1685 have high leverage but do not have a high standardized residual, and thus, aren't bad leverages

the symmetry about the y-axis might be due to us averaging life expectancy throughout all the years and therefore we obtain equally incorrect results



Variance Analysis

- The grey lines represent +/- 2 standard error
- 95% of data inside the box



Conclusions

Our purpose of the study is to learn which factors will lead to the longevity of a nation's life expectancy. After we run the logistic model with 5 numerical terms, 2 categorical variables and one interaction effect, we had a pretty good model with around 90% accuracy, and is tested to prevent overfitting by cv.

Out model shows that an increase in National Consumption has the largest positive impact on longevity, and the percent of people with depression disorder clearly has a negative impact on longevity. Also National consumption of plant(measured in calories) also has a tiny positive effect on national longevity.

As we assumed, country with medium amount of daily smokers has higher odds of having lower life expectancy, but what's surprising is that the country with high amount of daily smokers actually has higher odds of longer life expectancy.

Shortcomings

- Cutting point for numerical to categorical can be studied more
    - More levels and interval ranges impact final model results
- Interaction effect is not significant in the model
    - possibly analyze more factors and explore interaction effect
- Country-Year unit fails to provide meaningful insight
  (smoking,education,etc.)
    - Individual Data would be more interesting to analyze for audience's interest to prolong Longevity
- Averaging life expectancy for all the years in one model may remove information as life expectancy changes every year. Running our model per average life expectancy per year may help us observe different trends.

Data Cited

Population: https://data.worldbank.org/indicator/SP.POP.TOTL

Democracy:
https://ourworldindata.org/explorers/democracy?tab=table&country=ARG~AUS~BWA~CHN&Dataset=Varieties+of+Democracy&Metric=Electoral+democracy&Sub-metric=Main+index

Cause of Death: https://ourworldindata.org/grapher/annual-number-of-deaths-by-cause?time=earliest

Education: https://ourworldindata.org/grapher/mean-years-of-schooling-long-run?tab=table

GDP by Type($):
https://data.un.org/Data.aspx?q=gdp&d=SNAAMA&f=grID:101;currID:USD;pcFlag:0&c=2,3,5,6&s=_crEngNameOrderBy:asc,yr:desc&v=1

Life Expectancy:
https://data.un.org/Data.aspx?q=life&d=PopDiv&f=variableID:68&c=2,4,6,7&s=_crEngNameOrderBy:asc,_timeEngNameOrderBy:desc,_varEngNameOrderBy:asc&v=1

Data Cited

Obesity: https://ourworldindata.org/obesity

Smoking: https://ourworldindata.org/smoking

Diet: https://ourworldindata.org/diet-compositions

Mental Health: https://ourworldindata.org/mental-health,

Image and Paper Cited

Country to Continent Conversion Table:

https://github.com/dbouquin/IS_608/blob/77c1523be1684e04ed7d3c1a5fb584cbfcf9196e/NanosatDB_munging/Countries-Continents.csv

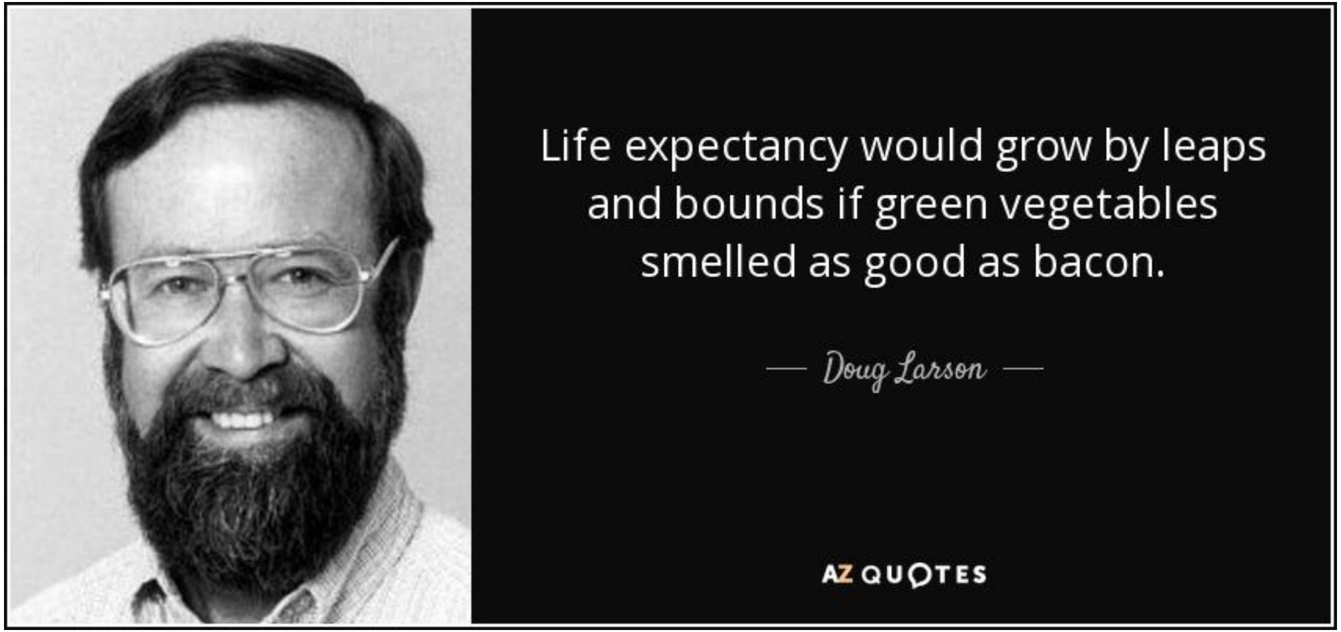
https://static.vecteezy.com/system/resources/previews/003/331/185/original/high-resolution-world-map-with-continent-in-different-color-free-vector.jpg

https://i.pinimg.com/originals/85/25/3e/85253e366cd6aafb6ff687d60c0e6e76.jpg

Freeman, T., Gesesew, H.A., Bambra, C. *et al.* Why do some countries do better or worse in life expectancy relative to income? An analysis of Brazil, Ethiopia, and the United States of America. *Int J Equity Health* 19, 202 (2020). https://doi.org/10.1186/s12939-020-01315-z

Q&A

Questions?
Questions?
Questions?



Life expectancy would grow by leaps and bounds if green vegetables smelled as good as bacon.

— Doug Larson —