HEALTH DATA SCIENCE PROJECT

Factors influencing obesity among adults in the US

PROBLEM DEFINITION:

In this project, we are trying to identify the various determinants contributing to the prevalence of obesity among adults in the United States, including social, economic, cultural, environmental, and behavioural factors.

INTRODUCTION:

We will use the BRFSS (Behavioral Risk Factor Surveillance System) dataset. It is a cross-sectional telephone survey that state health departments conduct monthly over landline telephones and cellular telephones with a standardised questionnaire and technical and methodologic assistance from the CDC (Centers for Disease Control and Prevention).BRFSS is used to collect prevalence data among U.S. residents regarding their risk behaviours and preventive health practices that can affect their health status. We decided on the topic of the factors that influence obesity among adults in the U.S. Studying the factors influencing obesity among adults in the U.S. is crucial.

Obesity is a significant public health issue in the U.S., contributing to various chronic diseases like heart disease, diabetes, and certain cancers. Understanding the factors contributing to obesity can inform public health policies and interventions aimed at prevention and management. Obesity-related healthcare costs are substantial, impacting both individuals and the healthcare system as a whole.

By identifying the factors influencing obesity, healthcare providers can develop more effective strategies for prevention and treatment, potentially reducing healthcare expenditures. Obesity can have profound social and economic consequences, including reduced quality of life, discrimination, and decreased productivity. Exploring the factors behind obesity can help address underlying societal issues such as access to healthy foods, socioeconomic disparities, and cultural influences.

How do we measure obesity in simple terms?
Obesity is typically defined by the Body Mass Index (BMI) with a threshold value (e.g., BMI ≥ 30 as obese).

VARIABLE	QUESTION	VALUE
1. **_AGEG5YR**	Fourteen-level age category	

2. _STATE	Do you currently live in___(state)____?	Values from 1 to 78

3. _SEX	Sex	Male = BIRTHSEX=1 or BIRTHSEX notin (1,2) and SEXVAR=1
		Female = BIRTHSEX=2 or BIRTHSEX notin (1,2) and SEXVAR=2

4. educa		1 no school, maybe just kindergarten 2 elementary (1-8) 3 high (9-11) 4 (12 or GED high school graduate) 5 1- 3 college, technical 6 college GRAD 4 years< 9 refuse to answer BLANK not asked
5. marital		1 married 2 divorced 3 widowed (lost wife) 4 separated 5 never married (single) 6 member.unmarried couple 9 refuse to answer BLANK
6. employ1	Are you currently…?	
7. income3	Is your annual household income from all sources:	
8. _BMI5	Body Mass Index (BMI)	
9. _BMI5CAT	Four-categories of Body Mass Index (BMI)	
10. genhlth	General Health (physical health + Mental health)	1 excellent 2 very good 3 good 4 Fair 5 Poor 7 not know, not sure 9 refuse BLANK not asked, missed
11. poorhlth	During 30 days, how many days, you could not do self-care, work, recreation	1-30 number of days 88 None of days 77 not know 99 refused BLANK
12. _TOTINDA	Adults who reported doing physical activity or exercise during the past 30 days	

	other than their regular job	
13. diabete4	(Ever told) (you had) diabetes?	
14. cvdcrhd4	(Ever told) (you had) angina or coronary heart disease?	
15. smokday2	do you now smoke cigarettes every day, some days, or not at all?	

Contingency tables:

Factors influencing obesity among adults in the US

The Models:

THE MODELS WE CAN INCLUDE IN OUR ANALYSIS ARE 3:
1. Logistic Regression, → Arnela Halili
2. Multivariate logistic regression, → Maha
3. Generalized Linear Mixed Models→ Amal

Model 1: Logistic Regression

Objective*: Predict whether an individual is obese based on typical BRFSS collected data.*

Variables:

- **Dependent Variable**: Obesity status (1 = BMI ≥ 30, 0 otherwise).
- **Independent Variables**:
  - **Age** (continuous).
  - **Gender** (male, female).
  - **Physical Activity** (number of days active per week).
  - **Education Level** (less than high school, high school graduate, some college, college graduate).
  - **Income Level** (categories typically divided by BRFSS income thresholds).

Model Output & Interpretation:

- The output provides coefficients indicating the log odds of being obese for a unit change in the continuous variables or for the categorical changes from the reference group.
- Interpretation involves calculating the odds ratio by exponentiating the coefficient. For example, an odds ratio for "Physical Activity" less than 1 suggests that more activity decreases the likelihood of obesity.

Model 2: Multivariate Logistic Regression

Objective: Analyze the impact of the same predictors on multiple binary health outcomes related to obesity (e.g., has diabetes DIABETE4, has heart disease CVDCRHD4).

The purpose of the Multivariate logistic regression, more commonly referred to as multinomial logistic regression, is used when the dependent variable is categorical with more than two levels. It models the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables.

Variables: As in Model 1 but with multiple binary outcomes.

Model Output & Interpretation:

- Outputs multiple sets of coefficients, one for each health condition.
- Coefficients for each condition are interpreted independently, considering how predictors influence each condition while accounting for their correlation.

Model 3: Generalized Linear Mixed Models (GLMM)

Objective: Consider regional effects by accounting for clustering within states or counties.

Variables:

- **Dependent Variable**: Obesity status.
- **Fixed Effects**: Individual-level variables as above.
- **Random Effects**: State or county to adjust for clustering effects that may influence obesity.

Model Output & Interpretation:

- Outputs include fixed effects for individual predictors and random effects indicating variability between clusters.

- Interpretation should focus on both the individual-level predictors and the importance of geographic variations.

Additional Information on GLMM:

Model Brief

Objective:

The objective of this analysis is to investigate the factors associated with obesity status while considering the regional effects by accounting for clustering within states or counties. This approach will help identify both individual-level predictors and geographic variations influencing obesity.

Variables:

- **Dependent Variable:**
    - **Obesity Status:** Typically derived from the BMI categories.
        - You can use the `_BMI5CAT` variable for this purpose, categorizing individuals into obese and non-obese groups.
- **Fixed Effects (Individual-level variables):**
    - **Age Category:** `_AGEG5YR`
    - **Sex:** `_SEX`
    - **Education Level:** `educa`
    - **Marital Status:** `marital`
    - **Employment Status:** `employ1`
    - **Income Level:** `income3`
    - **Body Mass Index (BMI):** `_BMI5`
    - **General Health:** `genhlth`
    - **Physical Activity:** `_TOTINDA`
    - **Diabetes Status:** `diabete4`
    - **Coronary Heart Disease Status:** `cvdcrhd4`
    - **Smoking Status:** `smokday2`
- **Random Effects (Clustering effects):**
    - **Geographic Clustering:** `_STATE` (or a more granular geographic identifier if available, such as county).

Model Output & Interpretation:

- **Outputs:**
    - Fixed effects estimates for individual predictors.

    - Random effects estimates indicating variability between clusters (states or counties).
- **Interpretation:**
    - Focus on the significance and direction of the fixed effects to understand how individual-level predictors influence obesity status.
    - Examine the random effects to assess the importance of geographic variations and identify regions with higher or lower obesity rates.

Steps for Analysis

1. **Specify the Research Question:**
    - What individual-level factors are associated with obesity status?
    - How do geographic variations (state or county) influence obesity rates?
2. **Data Preparation:**
    - Ensure all variables are correctly coded and missing values are handled.
    - Create a binary obesity status variable from `_BMI5CAT` if not already available.
3. **Model Specification:**
    - Define the GLMM with obesity status as the dependent variable.
    - Include individual-level variables as fixed effects.
    - Incorporate state or county as random effects to adjust for clustering.
4. **Model Fitting:**
    - Fit the GLMM using appropriate software (e.g., `lme4` package in R).
    - Check model diagnostics to ensure the model fits the data well.
5. **Model Interpretation:**
    - Interpret the fixed effects to understand the impact of individual predictors.
    - Analyze the random effects to understand the extent of geographic variability.
6. **Reporting Results:**
    - Present the results in a clear and concise manner, highlighting key findings.
    - Discuss the implications of both individual-level predictors and geographic variations on obesity.

Additional Notes:

- **Variable Availability**: Before proceeding, confirm the availability of these variables in the specific BRFSS dataset you are using. Variables can vary slightly between survey years.
- **Handling Missing Data**: BRFSS datasets often have missing values; decide on imputation or exclusion strategies based on the extent and pattern of missing data.
- **Complex Survey Design**: BRFSS employs a complex survey design with weighting, which should be considered in your analysis to ensure representativeness and correct inference.

```
[1] "state"          "genhlth"        "poorhlth"       "hrtdisease"
"diabetic"
 [6] "marital"        "edu"             "employstatus"   "income"
"smokeday"
[11] "totinda"        "sex"             "age5cat"        "bmi5"
"bmi"
[16] "index"          "obesity_status"
```

WHAT DO WE NEED TO DO?

Contingency tables

Logistic regression

Generalized linear models

Multivariate model ???

THE POWER POINT PRESENTATION
(select the most relevant sentences)

→ change the names of the column names (to be easier)