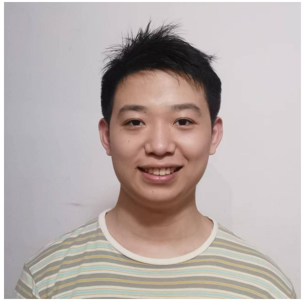
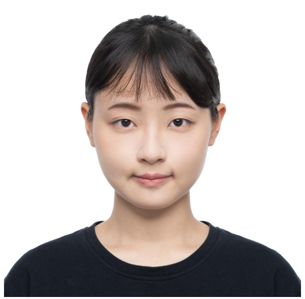
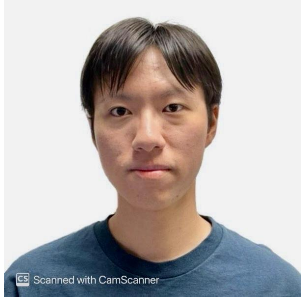
Categorical Analysis on Diabetes and BMI over Pima Indian Patients



Liangjie Lu



Weiting Lin



Luyang Zhang

Yaocao Chen

Data Collection Process - Kaggle

Data Source:

- Originated from the National Institute of Diabetes and Digestive and Kidney Diseases.
- Contributed by Vincent Sigillito from Johns Hopkins University, received on May 9, 1990.

Data Description:

- Aimed at predicting diabetes presence in ***female Pima Indian patients aged 21 or older***.
- Includes variables like *BMI*, pregnancies, plasma glucose, blood pressure, skin fold thickness, insulin, diabetes pedigree function, and age.
- Features a binary variable for diabetes status (0 = free of diabetes or 1 otherwise).

1. Data Source: https://www.kaggle.com/datasets/mathchi/diabetes-data-set
2. The Human Genome Project and Diabetes: Genetics of Type II Diabetes. New Mexico State University. 1997. June 1, 2006. "Diabetes and Genes in Disease". Archived from the original on June 16, 2006. Retrieved June 1, 2006.

Primary Problem

The dataset highlights the importance of diagnostic measures in predicting diabetes risk, focusing on the link between BMI and diabetes incidence.

- BMI, a key indicator of body fat based on height and weight, is crucial in assessing diabetes risk.
- The CDC provides standardized BMI categories for adults, applicable across all demographics.

Our study seeks to explore *the relationship between BMI category and diabetes status, aiming to determine statistical independence.*

BMI	**BMI Category**
Below 18.5	Underweight
18.5 – 24.9	Healthy Weight
25.0 – 29.9	Overweight
30.0 and Above	Obesity

Table 1. Standard BMI Categories and Their Health Implications.

CDC: Centers for Disease Control and Prevention
Table 1 Source:https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

Model/Test Usage

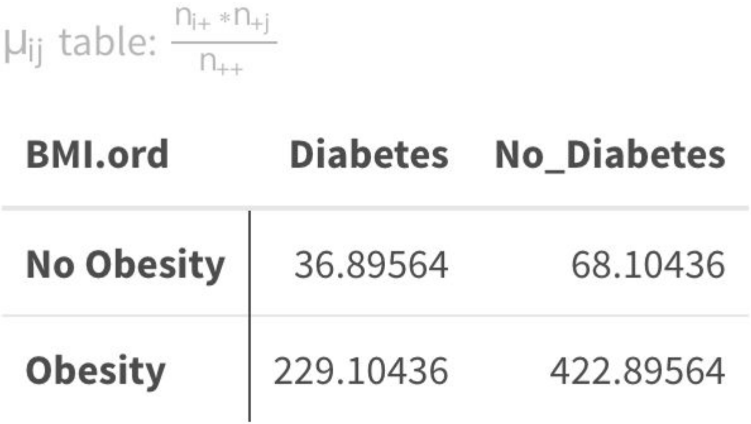
- Independent Test: Pearson's Test (X^2) / Log Likelihood Test (G^2)
- Ordinal Correlation Test: Assessing independence by using sample correlation as the test statistic, with mean BMI values for each category as the basis.
- Fisher's Test: Apply on the whole dataset. Apply on healthy BMI and other three categories to find the relationship between abnormal BMI with healthy BMI.
- Logistics Model with Binomial: Regress diabetes status on BMI raw scores
- Odds ratio for diabetes among non-obesity and among obesity subjects

Independence Test:

$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$

H_α : at least one $\pi_{ij} \neq \pi_{i+}\pi_{+j}$

i=Obesity, No-Obesity, j=Diabetes, No Diabetes



μ_{ij} table: $\frac{n_{i+} * n_{+j}}{n_{++}}$

BMI.ord	Diabetes	No_Diabetes
No Obesity	36.89564	68.10436
Obesity	229.10436	422.89564

1. Pearson’s chi square Test (X^2)

$$X^2 = \sum_{j=1}^{c} \frac{(n_j - \mu_j)^2}{\mu_j}$$

X^2 = 43.3614

P-value: 2.8295e-11

2. Log Likelihood Test (G^2)

$$G^2 = 2\sum n_j \log\left(n_j / \mu_j\right)$$

G^2 = 53.9656

P-value: 1.945e-13

Result: P-value<0.001 reject null hypothesis, and Diabetes variable is dependent to BMI ordinal variable.

Note: pchisq(q=value, df=1, lower.tail=FALSE)

Ordinal Correlation Test

$$H_0 : \rho = 0 \text{ vs } H_1 : \rho \neq 0$$

Test statistic: $M = (n-1)r^2$

Under $H_0, M \sim \chi_1^2$

$$r = \frac{\sum_i \sum_j (v_i - \bar{v})(w_j - \bar{w}) p_{ij}}{\sqrt{\sum_i (v_i - \bar{v})^2 p_{i+} \sum_j (w_j - \bar{w})^2 p_{+j}}}$$

=0.320

$$M = 77.3 > 10.8 = F^{-1}_{\chi_1^2}(0.001)$$

So reject H0 at significance level 0.001

BMI Category	BMI	BMI mean
Underweight	18.2	(18.2+18.2 +18.2+18.4) /4 =18.25
Underweight	18.2	
Underweight	18.2	
Underweight	18.4	
Healthy weight	…	22.75
Overweight	…	27.42
Obesity	…	36.48

Fisher’s Exact Test Assumptions

Assumptions:

- ***Categorical Variables***: The test is applied to categorical data arranged in a 2x2 contingency table.
- ***Independence***: Each observation is independent, meaning the selection of one individual does not influence the selection of another.
- ***Fixed Margins***: The row and column totals are fixed, or 'conditioned'.
- ***Sample Size***: Fisher's Exact Test is traditionally used for small sample sizes, it can also be used for any larger sample size.
- ***Mutually Exclusive Categories***: Each subject can only be in one category for each variable, and each outcome can only fall into one category.

Based on the assumptions, we decided to apply the fisher’s exact test on the healthy BMI with other two categories(overweight and obesity) to find out the relationship between proportion of diabetes cases for those with a 'Healthy BMI' and those with BMI levels considered outside of the healthy range

Fisher’s Exact Test Result

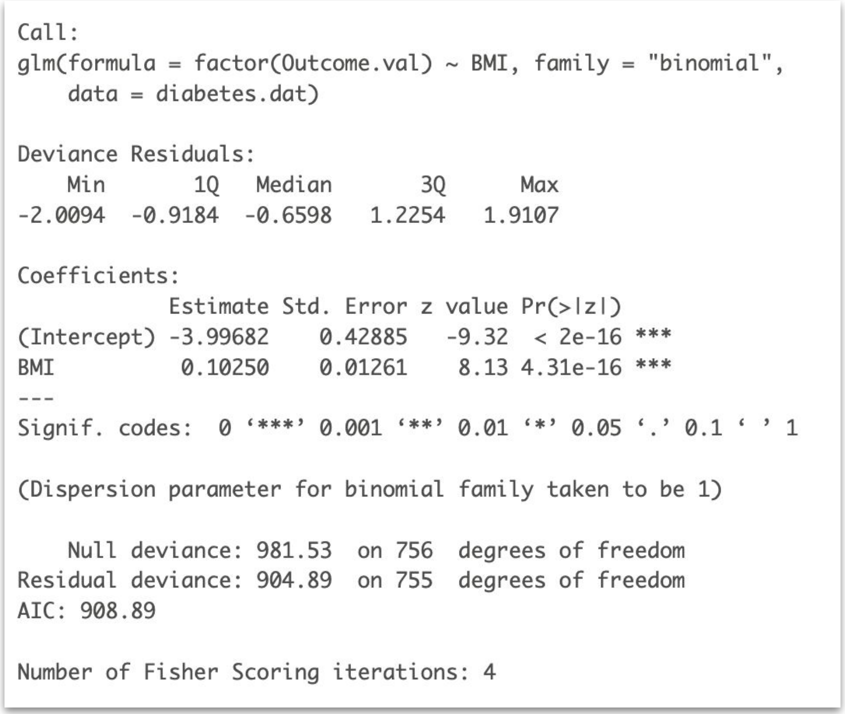
	odds ratio	P value
Healthy vs. Overweight	3.70928	1.1751e-03
Healthy vs. Obesity	11.556069	1.8037e-15

From the results above, if the null hypothesis is odds ratio equals to 1 then we reject the null hypothesis for both overweight and obesity. We can draw the conclusion that there’s significant correlation between diabetes and healthy BMI level. We can also find that the odds ratio increases when the the BMI level increase. The outcome indicates that the odds of having diabetes among the overweight people is higher than the odds of having diabetes among the healthy individuals.

Logistic Regression

BMI (Predictor)	Diabetes (Response)
33.6	1
26.6	0
23.3	1
28.1	0
43.1	1
25.6	0

- Let π be the probability that patients have diabetes.
- Let X be the BMI score



```
Call:
glm(formula = factor(Outcome.val) ~ BMI, family = "binomial",
    data = diabetes.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0094  -0.9184  -0.6598   1.2254   1.9107

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.99682    0.42885   -9.32  < 2e-16 ***
BMI          0.10250    0.01261    8.13 4.31e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 981.53  on 756  degrees of freedom
Residual deviance: 904.89  on 755  degrees of freedom
AIC: 908.89

Number of Fisher Scoring iterations: 4
```

$$logit(\pi) = -4 + 0.103X$$

Logistic Regression

Model Validation

$$H_0 : \beta_1 = 0 \text{ against } H_1 = \beta_1 \neq 0$$

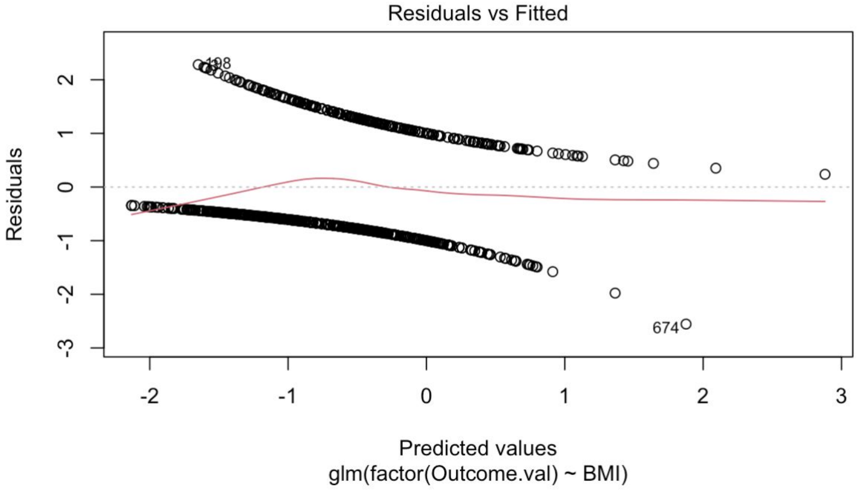
test statistics for Likelihod Ratio Test:

$$G^2 = (\text{Null Deviance}) - (\text{Residaul Deviance})$$

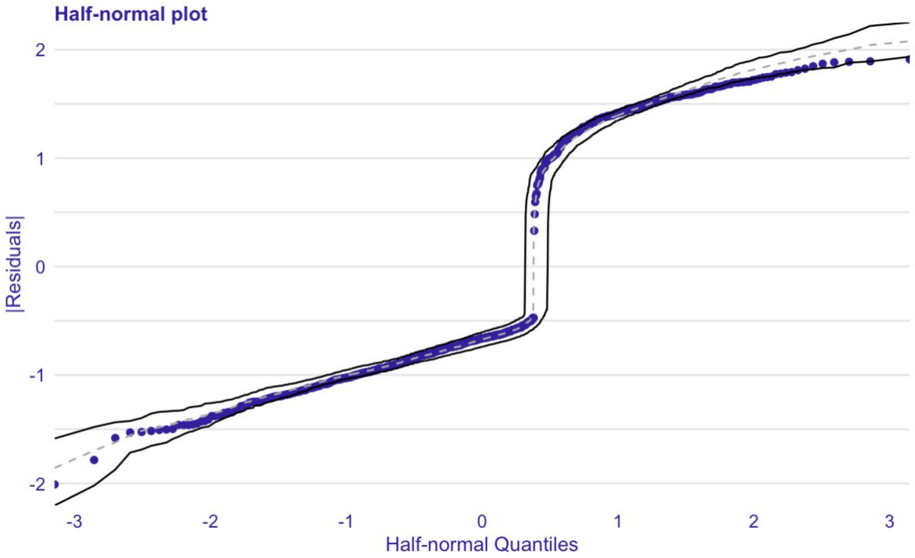
$$= 981.53 - 904.89 = 76.64 \sim \chi^2_1$$

Rejection Region: > 3.8416

Conclusion: We reject the null hypothesis



Loess of residuals seem close to zero



No specific outliers

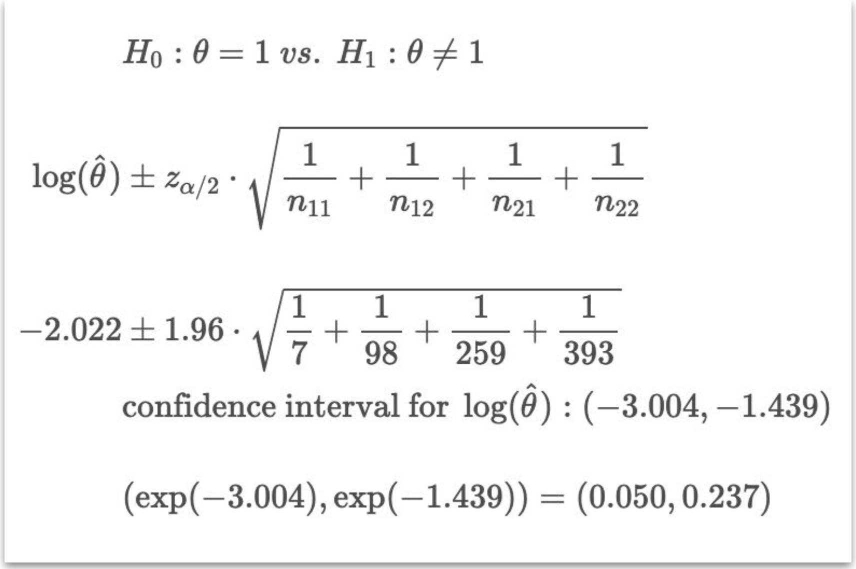
Odds Ratio

	Diabetes	No Diabetes	Total
No Obesity	7	98	105
Obesity	259	393	652
Total	266	491	757

- π_1 = 7/105 = 0.0667, π_2 = 259/652=0.397
- Odds for diabetes among no obesity:0.0667/(1-0.0667) =0.0715
- Odds for diabetes among obesity: 0.397/(1-0.397)=0.658
- Estimated odds ratio = 0.0715/0.658=0.108

Odds Ratio – Confidence interval

	Diabetes	No Diabetes	Total
No Obesity	7	98	105
Obesity	259	393	652
Total	266	491	757



$$H_0 : \theta = 1 \; vs. \; H_1 : \theta \neq 1$$

$$\log(\hat{\theta}) \pm z_{\alpha/2} \cdot \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

$$-2.022 \pm 1.96 \cdot \sqrt{\frac{1}{7} + \frac{1}{98} + \frac{1}{259} + \frac{1}{393}}$$

confidence interval for $\log(\hat{\theta}) : (-3.004, -1.439)$

$$(\exp(-3.004), \exp(-1.439)) = (0.050, 0.237)$$

- The 95% confidence interval for estimated odds ratio is (0.05, 0.237)
- Since 1 is not included in the confidence interval, we reject the null hypothesis and conclude that the estimated odds ratio for having diabetes among obesity patients is higher than the corresponding for non-obesity patients.

Conclusion of Project

1. Based on the **Pearson's test** and **Log-likelihood test**, we conclude that the Diabetes variable (Diabetes/No Diabetes) and BMI Ordinal variable (Obesity/No Obesity) are dependent and both with p-value less than 0.0001.
2. Reject the null hypothesis **asserting zero correlation** at significance level 0.001. The population correlation between Diabetes outcomes and BMI categories is not 0. Diabetes outcomes are dependent on and BMI categories!
3. The **Fisher's Exact Test** indicates a statistically significant association between BMI categories and diabetes outcome. The result also shows when BMI increases, the odds of diabetes also increase.
4. The **logistic regression** suggests that BMI is significant in explaining diabetes.
5. The estimated odds for having diabetes among obesity patients is higher than the corresponding odds for non-obesity patients.

The status of diabetes is determined by the BMI (Body Mass Index) categories corresponding to the particular participants (female Pima Indian patients aged 21 or older).

Thank you for your attention!