Association between Type II Diabetes and Body Mass Index (BMI)

Cheng Peng

West Chester University

Contents

1 Introduction

The association between a high BMI and Type 2 Diabetes is causal and robust. Excess body fat, particularly in the abdominal area, drives the development of insulin resistance and beta-cell failure, the hallmarks of the disease. While BMI has its limitations as a metric, it remains a powerful indicator of diabetes risk at both a population and individual level. Weight management through lifestyle modification is the single most effective strategy for preventing and managing Type 2 Diabetes.

2 Data Source

The diabetes data set in this case study contains 768 observations on 9 variables. The data set is available in the UCI machine learning data repository. R library {mlbench} has two versions of this data. The data set contains a significant number of missing values.

There are 9 variables in the data set.

1. **pregnant**: Number of times pregnant
2. **glucose**: Plasma glucose concentration (glucose tolerance test)
3. **pressure**: Diastolic blood pressure (mm Hg)
4. **triceps**: Triceps skin fold thickness (mm)
5. **insulin**: 2-Hour serum insulin (mu U/ml)
6. **mass**: Body mass index (weight in kg/(height in m)^2)
7. **pedigree**: Diabetes pedigree function

8. **age**: Age (years)
9. **diabetes**: Class variable (test for diabetes)

The following code loads the `PimaIndiansDiabetes2` dataset from the R **mlbench** library.

```
#library(mlbench)
data(PimaIndiansDiabetes2)           # load the data to R work-space
diabetes.0 = PimaIndiansDiabetes2    # make a copy of the data for data cleansing
diabetes = na.omit(diabetes.0)       # Delete all records with missing components
y0=diabetes$diabetes
diabete.01 = rep(0, length(y0))      # define a 0-1 to test which probability is used in glm()
diabete.01[which(y0=="pos")] = 1
diabetes$diabetes.01 = diabete.01
head(diabetes)
```

```
   pregnant glucose pressure triceps insulin mass pedigree age diabetes
4         1      89       66      23      94 28.1    0.167  21      neg
5         0     137       40      35     168 43.1    2.288  33      pos
7         3      78       50      32      88 31.0    0.248  26      pos
9         2     197       70      45     543 30.5    0.158  53      pos
14        1     189       60      23     846 30.1    0.398  59      pos
15        5     166       72      19     175 25.8    0.587  51      pos
   diabetes.01
4            0
5            1
7            1
9            1
14           1
15           1
```

Records with missing values were excluded from the analysis, resulting in a final analytic dataset of 392 complete cases.

3 Clinical Question

Many studies indicated that body mass index (BMI) is a more powerful risk factor for diabetes than genetics. The objective of this case study is to explore the **association** between BMI and diabetes.

The general interpretation of BMI for adults is given below:

- **underweight**: < 18.5
- **Normal and Healthy Weight**: [18.5, 24.9]
- **Overweight**: [25.0, 29.9]
- **Obese**: > 30.0

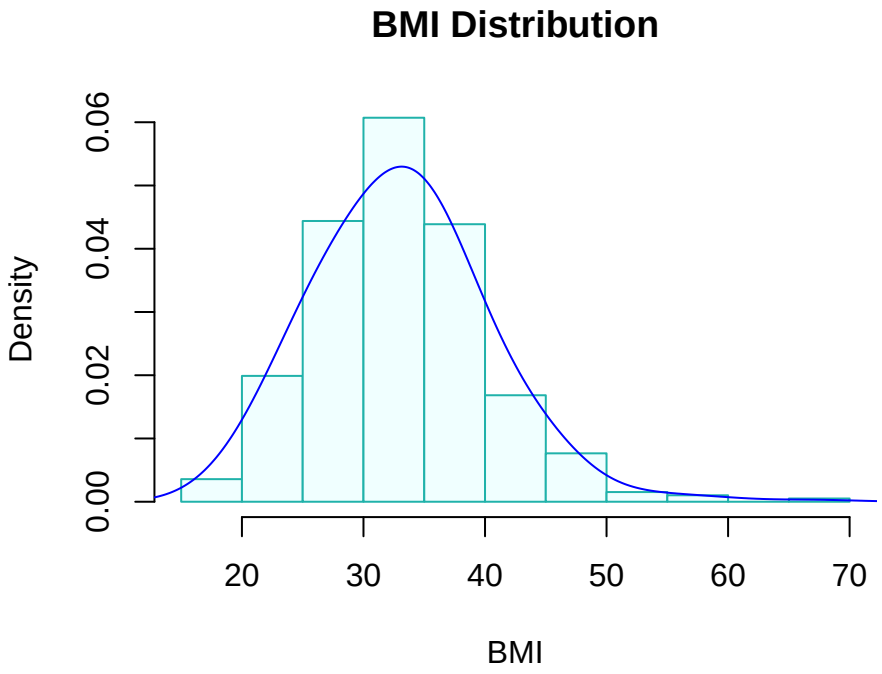
In this study, we focus only on simple logistic regression modeling, we stay with the original (continuous) numerical variable BMI. The above discretization will be used in the case study in the next note.

4 Exploratory Data Analysis

Since we only study the simple logistic regression model, only one predictor variable is included in the model. We first perform exploratory data analysis on the predictor variable to make sure the variable is not extremely skewed.

```
ylimit = max(density(diabetes$mass)$y)   # find the highest point on the density curve.
hist(diabetes$mass, probability = TRUE,
```

```
    main = "BMI Distribution",
    xlab="BMI",
    col = "azure1",
    border="lightseagreen")
  lines(density(diabetes$mass, adjust=2), col="blue")
```



The histogram above shows that the BMI distribution is slightly right-skewed. However, no transformation is necessary for this predictor variable at this stage.

The response variable, `diabetes`, is a binary categorical variable. In the dataset of 392 subjects, 130 (33.3%) were diagnosed with Type 2 diabetes. From a modeling perspective, this distribution does not indicate a class imbalance that would bias the results.

Exploratory Data Analysis (EDA) for other variables will be performed in a subsequent case study on multiple logistic regression, as needed.

5 Building Logistic Regression Model

Since the simple logistic regression contains only one continuous variable of a binary categorical variable as the predictor variable, no there is no issue of potential imbalance. We will not transform BMI and fit a logistic regression directly to the data.

```
s.logit = glm(diabetes ~ mass,
          family = binomial(link = "logit"),  # family is the binomial, logit(p) = log(p/(1-p))!
          data = diabetes)                    # the data frame is a subset of the original iris data
result = summary(s.logit)
result
```

```
Call:
```

```
glm(formula = diabetes ~ mass, family = binomial(link = "logit"),
    data = diabetes)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.60614    0.59173  -6.094 1.10e-09 ***
mass         0.08633    0.01705   5.062 4.14e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 498.10  on 391  degrees of freedom
Residual deviance: 469.03  on 390  degrees of freedom
AIC: 473.03

Number of Fisher Scoring iterations: 4
```

The response variable diabetes is a binary factor. Using R's default alphabetical ordering for factors, the reference level is "neg" (coded as 0) and the event level is "pos" (coded as 1). The model thus estimates the probability of the event, P(diabetes = "pos"). The logistic regression model was fitted as follows:

The summary of major statistics is given below.

```
model.coef.stats = summary(s.logit)$coef        # output stats of coefficients
conf.ci = confint(s.logit)                      # confidence intervals of betas
sum.stats = cbind(model.coef.stats, conf.ci.95=conf.ci)   # rounding off decimals
pander(sum.stats,caption = "The summary stats of regression coefficients")
```

Table 1: The summary stats of regression coefficients

	Estimate	Std. Error	z value	Pr(>|z|)	2.5 %	97.5 %
(Intercept)	-3.606	0.5917	-6.094	1.1e-09	-4.806	-2.482
mass	0.08633	0.01705	5.062	4.143e-07	0.05383	0.1208

As shown in the table, BMI demonstrates a positive association with diabetes status, with a coefficient (β_1) of 0.0863 and a p-value close to zero. This finding is supported by the 95% confidence interval of [0.0538, 0.1208], which excludes the null value of zero, and is consistent with existing literature.

From a practical perspective, it is more common to interpret the results using the odds ratio. Therefore, we now convert the estimated regression coefficients into odds ratios.

```
# Odds ratio
model.coef.stats = summary(s.logit)$coef
odds.ratio = exp(coef(s.logit))
out.stats = cbind(model.coef.stats, odds.ratio = odds.ratio)
pander(out.stats,caption = "Summary Stats with Odds Ratios")
```

Table 2: Summary Stats with Odds Ratios

	Estimate	Std. Error	z value	Pr(>|z|)	odds.ratio
(Intercept)	-3.606	0.5917	-6.094	1.1e-09	0.02716
mass	0.08633	0.01705	5.062	4.143e-07	1.09

The odds ratio for BMI is 1.09, indicating that for each one-unit increase in BMI, **the odds of a positive diabetes test increase by approximately 9%**. This represents a practically significant risk factor for diabetes.

Several global goodness-of-fit measures for the model are summarized in the following table.

```
## Other global goodness-of-fit
dev.resid = s.logit$deviance
dev.0.resid = s.logit$null.deviance
aic = s.logit$aic
goodness = cbind(Deviance.residual =dev.resid, Null.Deviance.Residual = dev.0.resid,
      AIC = aic)
pander(goodness)
```

Deviance.residual	Null.Deviance.Residual	AIC
469	498.1	473

Since the above global goodness-of-fit is based on the **log-likelihood (LL) function**, we don't have other candidate models with corresponding likelihood at the same scale to compare in this simple logistic regression model, we will not interpret these goodness-of-fit measures.

Comments: Definitions of key terms for understanding the output.

1. The **Saturated Model** is a model that assumes each data point has its own parameters (which means you have n parameters to estimate.)
2. The **Null Model** assumes the exact "opposite", in that is assumes one parameter for all of the data points, which means you only estimate 1 parameter.
3. The **Proposed Model** assumes you can explain your data points with p parameters + an intercept term, so you have $p + 1$ parameters.
4. **log-likelihood**: In simple terms, the log-likelihood is a measure of how well a statistical model fits a set of data. A higher (less negative) log-likelihood indicates a better fit.
5. **-2 Log-Likelihood (-2LL)**: Multiplying by 2 gives us a statistic whose sampling distribution is known and very useful.
6. **Null Deviance** $= 2(LL(\text{Saturated Model}) - LL(\text{Null Model}))$ on $df = df_{\text{Sat}} - df_{\text{Null}}$
7. **Residual Deviance** $= 2(LL(\text{Saturated Model}) - LL(\text{Proposed Model}))$ with $df = df_{\text{Sat}} - df_{\text{Proposed}}$

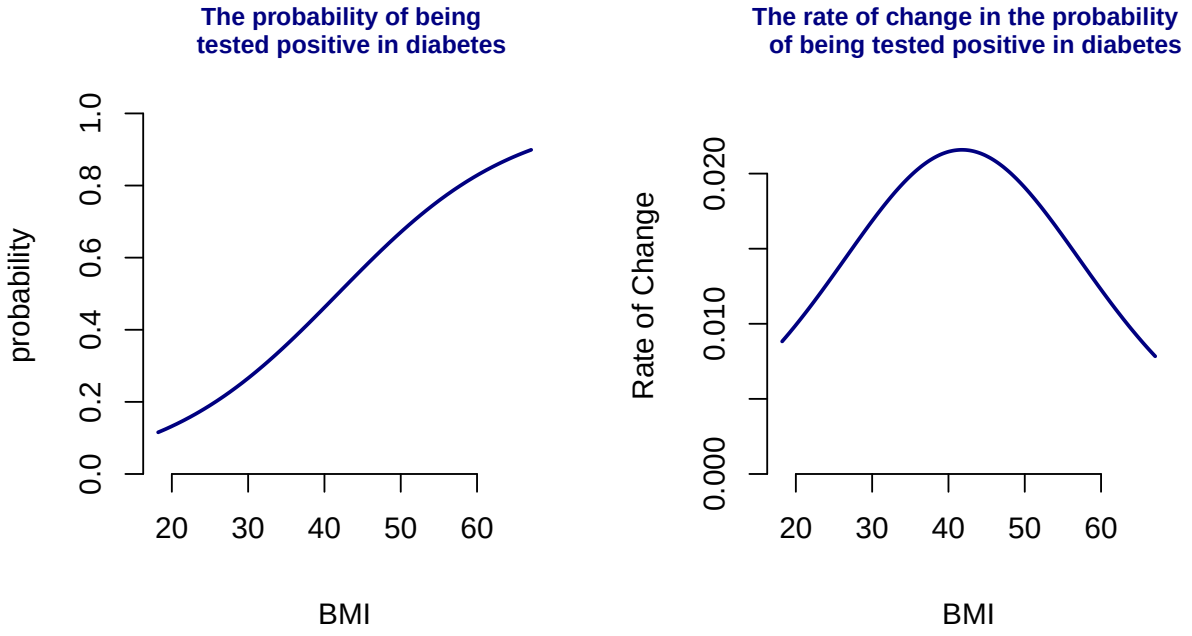
It is important to know that both **Null Deviance** and **Residual Deviance** are asymptotic χ^2_{df}!

If your Null Deviance is really small, it means that the Null Model explains the data pretty well. Likewise with your Residual Deviance.

The success probability curve (so-called S curve) is given below.

```
bmi.range = range(diabetes$mass)
x = seq(bmi.range[1], bmi.range[2], length = 200)
beta.x = coef(s.logit)[1] + coef(s.logit)[2]*x
success.prob = exp(beta.x)/(1+exp(beta.x))
failure.prob = 1/(1+exp(beta.x))
ylimit = max(success.prob, failure.prob)
##
beta1 = coef(s.logit)[2]
success.prob.rate = beta1*exp(beta.x)/(1+exp(beta.x))^2
##
```

```r
##
par(mfrow = c(1,2))
plot(x, success.prob, type = "l", lwd = 2, col = "navy",
     main = "The probability of being \n  tested positive in diabetes",
     ylim=c(0, 1.1*ylimit),
     xlab = "BMI",
     ylab = "probability",
     axes = FALSE,
     col.main = "navy",
     cex.main = 0.8)
# lines(x, failure.prob,lwd = 2, col = "darkred")
axis(1, pos = 0)
axis(2)
# legend(30, 1, c("Success Probability", "Failure Probability"), lwd = rep(2,2),
#        col = c("navy", "darkred"), cex = 0.7, bty = "n")
##
y.rate = max(success.prob.rate)
plot(x, success.prob.rate, type = "l", lwd = 2, col = "navy",
     main = "The rate of change in the probability \n  of being tested positive in diabetes",
     xlab = "BMI",
     ylab = "Rate of Change",
     ylim=c(0,1.1*y.rate),
     axes = FALSE,
     col.main = "navy",
     cex.main = 0.8
     )
axis(1, pos = 0)
axis(2)
```



The left-hand side plot in the above figure is the standard **S curve** representing how the probability of a

positive test increases as the BMI increases. After diving deeper to see the rate of change in the probability of a positive test, we obtain the curve on the right-hand side that indicates that the rate of change in the probability of positive test increases when BMI is less than 40 and decreases when BMI is greater than 40. The turning point is about 40.

6 Conclusion and Discussion

This case study explores the relationship between diabetes and Body Mass Index (BMI). The results show a positive association: for each one-unit increase in BMI, the odds of developing diabetes increase by approximately 9%. It is important to note that the development of diabetes is complex and involves many factors. This analysis could be improved by incorporating additional variables. While the results might be adjusted, the positive association between BMI and diabetes would be expected to persist based on existing clinical studies.