Investigating Risk Factors for Diabetes Diagnosis: A Cross-Sectional Study

ECON F342 : Applied Econometrics

Under the Supervision of

Dr. Rishi Kumar

Name	**ID**
Chinmay Anand	2020B3A70776H
Harsh Loya	2020B3A40831H
Sabyasachi Bhoi	2020B3A72147H
Dhruv Deshmukh	2020B3A72160H
Arya V Singalwar	2020B3A71861H
Ayush Varshney	2020B3A32146H
Nimish Agrawal	2020B3A71857H
Subodh Kumar	2020B3A31165H

Abstract

This paper investigates the relationship between various independent variables and diabetes diagnosis, using a cross-sectional dataset of various individuals. The research question is whether certain independent variables, such as age, BMI, and family history, are predictive of diabetes diagnosis. The regression was performed using data released by NIDDK. The regression analysis revealed that glucose levels, age, pregnancy count, BMI and family history were all significant predictors of diabetes diagnosis, with older age and higher weight increasing the likelihood of diagnosis, and family history acting as a risk factor. Other independent variables, such as Insulin levels, Skin Thickness and Blood Pressure, were found to have no significant effect on diagnosis. These findings have important implications for diabetes prevention and management, as they suggest that targeting certain risk factors, such as weight and family history, may be effective in reducing the incidence of diabetes. However, the study is limited by its cross-sectional design.

Contents

1 Introduction

Diabetes is one of the top causes of death worldwide. Combined with obesity and hypertension, it tends to take the form of cardiovascular diseases, the leading causes of death worldwide (Khan et al., 2017). India has the second-largest population of people who have diabetes. 64.5 million people had diabetes in 2017, and the number of diabetics is projected to increase to 87 million by 2030 (Dudeja, Singh, Gadekar, & Mukherji, 2017). The number of people who have diabetes is higher in the urban parts of India compared to the rural regions. In a metropolitan area, a person is twice as likely to suffer from diabetes. The more susceptible people are the ones living in urban slums rather than the wealthy urban class. This is mainly caused due to lack of information about nutrition and proper diet.

In diabetes, the most common type of diabetes found is type two diabetes, accounting for more than 90 percent of the diabetic population. Among the people who have type two diabetes, over 50 percent never get diagnosed. According to the International Diabetes Federation, about 66 percent of Indians making up the diabetic population (44 Lakhs) were unaware they had diabetes. This makes this group of individuals have an increased risk of complications. Kahn et al. (2010), in a study, have determined that screening people for diabetes improves their quality of life.

A few known reasons for the escalation of the diabetes problem in India are greater resistance to insulin, strong genetic factors, and environmental factors due to urbanization: Studies by Yajnik et al. (1995) have shown that low birth weight is associated with higher insulin resistance among Indians. His studies have also demonstrated that Indian newborns have higher insulin levels than European newborns. One more factor attributed to this is accelerated childhood growth. A follow-up study has supported these findings and has concluded that lower birth weight, with obesity, leads to very high rates of diabetes. Genetic susceptibility does play an essential role in the occurrence of Type two diabetes. Biologically, Indians are known to have unfavorable phenotypes and risk profiles for type 2 diabetes and coronary artery diseases, thus making them more prone to the conditions. Due to these factors, diabetes also develops at least a decade or two earlier among Indians than Europeans, which is considerably higher. Despite all the evidence suggesting genetic factors' role among Indians, the genetic shift is a slow process. Without long-term follow-up studies, a conclusive statement cannot be made about the current diabetes situation with respect to genetic factors. The role of environmental factors is also a serious factor that cannot be overlooked in a study for diabetes. India is undergoing rapid urbanization, with urbanization rates at 35 percent. This has led to increased consumption of foods rich in fats, sugars, and calories. This, combined with high levels of mental stress, is not a good sign regarding insulin sensitivity and obesity. Studies have also shown that the possibility of incurring diabetes in India was lower among people with lower incomes and much more among people from affluent groups.

It has also been observed that physical inactivity is linked with a higher probability of incurring diabetes. This is one of the reasons the rural population of India, which has a physically active lifestyle, does not suffer from diabetes as much as the urban population, which has a comparatively sedentary lifestyle. A study by Misra et al. (2001) has reported that migration from rural areas to slums in urban areas leads to obesity and glucose intolerance, which intensify the problem of diabetes.

In India, urbanization and economic growth have caused a change in dietary patterns. Traditional patterns are being replaced by industrially manufactured packaged food. Greater intake of such high-calorie items has put many at risk of developing diabetes and coronary artery diseases.

The objective of this study is to identify the determinants that exhibit a significant impact on the likelihood of diabetes incidence. To this end, a dataset from the National Institute of Diabetes and Kidney Diseases, comprising variables such as age, BMI, and number of pregnancies, among others, has been utilized. We have employed Multiple Linear Regression to assess the effect of these variables. Our model scrutinizes the significance of independent variables at 10 percent, 5 percent, and 1 percent levels on the outcome, namely, the propensity of diabetes occurrence. Diagnostic tests, including Heteroskedasticity, Omitted Variable Bias, and Multicollinearity have been conducted to ensure the credibility of our model.

2 Literature Review

2.1 Paper 1

"Non-Communicable Disease Risk Factors and their Trends in India"

By Suzanne Nethan, Dhirendra Sinha, and Ravi Mehrotra

About the Paper

In this paper, Nethan, Sinha, & Mehrotra (2017) examine non-communicable diseases (NCD). It accounts for about 38 million (68%) deaths and 5.87 million (60%) worldwide. The paper discusses various national/state-level surveys in India, including single or multiple risk factors. Being overweight, obesity and tobacco use are all covered nationally.

Use Case

As a result of reading this paper, we gained a deeper understanding of what we can expect from a research model as diabetes is a non-communicable disease.

Methodology

Indicator definitions from the World Health Organization (WHO) for both urban and rural populations were used in the research. The percentage of the population covered by the polls for each risk factor was then calculated by adding the state-by-state population proportion and dividing it by the total Indian population. The previous and present data from the periodic surveys were also contrasted to evaluate changes in prevalence. PubMed, Google, and various surveillance systems were searched for this systematic study. Forty-one papers/survey reports from the search results were ultimately determined to be eligible-data on NCD risk variables from states and the nation.

Conclusion

India has a much-delayed response to NCD risk factors surveillance, and information on the exact needs to be more consistent and complete. According to the result, India should plan for a cost and time-effective NCD surveillance system.

2.2 Paper 2

> "Performance of Indian Diabetes Risk Score (IDRS) as a screening tool for diabetes in an urban slum"

By Lt Col Puja Dudeja, Maj Gurpreet Singh, Maj Tukaram Gadekar, Air Cmde Sandip Mukherji

About the Paper

To identify undetected Type 2 diabetes, the Madras Diabetes Research Foundation (MDRF) created the Indian Diabetes Risk Score (IDRS). As well as examining the prevalence of undiagnosed Type 2 diabetes in urban slums, this paper will examine how well the IDRS performs as a diagnostic tool.

Use Case

As a result of reading this paper, we gained a deeper understanding of what we can expect from a research model.

Methodology

Surveys were carried out in urban slum areas. A total of 155 IDRS observation tools were used to assess diabetes risk, including variables that can be adjusted (waist circumference, physical activity) and variables that cannot be modified (age, family history). Anthropometry data was collected. Using fasting blood sugar levels, diabetes was diagnosed.

Conclusion

Using IDRS to the community can effectively detect un-diagnosed diabetes.

2.3 Paper 3

> "Prevalence of Diabetes Mellitus and its risk factors"

By Akula Sanjeevaiah, Akula Sushmitha, Thota Srikanth

About the Paper

In this paper, Sanjeevaiah, Sushmitha, & Srikanth (2019) explore the prevalence of Diabetes Mellitus (DM) and its risk factors among the population of India. The study found a high prevalence of DM, particularly in urban areas and among those with a family history of the disease. The major risk factors identified were age, obesity, sedentary lifestyle, hypertension, and smoking. The paper highlights the need for early detection and effective management of DM to prevent complications and improve health outcomes.

Methodology

The authors conducted a cross-sectional study for a period of 4 months with a sample size of 250. The basis for the selection of subjects was age greater than 15 years of both genders who are identified with diabetes. Measurements of height and weight were taken to estimate BMI, waist circumference and blood pressure.

Conclusion

This paper concluded that age, waist circumference, hypertension, BMI, smoking habit and total cholesterol are noteworthy when comparing diabetic and non-diabetic subjects.

2.4 Paper 4

> "Assessment of Diabetes Risk in an Adult Population Using Indian Diabetes Risk Score in an Urban Resettlement Colony of Delhi."

By Acharya, Anita Shankar and Singh, Anshu and Dhiman, Balraj

About the Paper

Acharya, Singh, & Dhiman (2017) employed the Indian Diabetes Risk Score(IDRS) to assess the risk of individuals getting a positive diabetes diagnosis. Various socio-economic factors were taken into consideration alongside the IDRS score.

Methodology

In a cross-sectional survey of urban settlements in East Delhi, researchers sampled adults aged 30 or above of both genders, resulting in a total of 580 subjects. Employing the analytical capabilities of SPSS, the gathered data was examined. The survey assessed several factors, including but not limited to socio-economic status, physiological and psychological assessments, addiction habits, BMI, anthropometry, and IDRS scores. The socio-economic status was determined by utilizing the widely recognized *"Kuppuswamy Scale"*.

Conclusion

The authors have arrived at the conclusion that over 90% of the surveyed study subjects were found to be at risk of developing diabetes, thereby emphasizing the criticality of early-stage screening to enable prompt interventions.

2.5 Paper 5

> "Validity of Indian Diabetes Risk Score and its association with body mass index and glycosylated hemoglobin for screening of diabetes in and around areas of Lucknow"

By Khan, Mohammad Mustufa and Sonkar, Gyanendra Kumar and Alam, Roshan and Mehrotra, Sudhir and Khan, M Salman and Kumar, Ajay and Sonkar, Satyendra Kumar

About the Paper

Khan et al. (2017) employ various risk factors such as waist size, age, level of physical activity amongst the individuals and their family history for screening obesity and abdominal obesity. Their primary aim in this paper is to assess the validity of *Indian Diabetes Risk Score (IDRS)* and its association with BMI.

Methodology

The authors employed the a cross-sectional dataset along with the MDRF-IDRS risk score. They put heavy emphasis on the BMI level and the average glucose levels (which was retrieved using the *HbA1c* blood test) of the individuals in the dataset. These variables were then regressed and the regression coefficients were obtained; which were then tested for statistical significance using two-tailed tests at 5% level of significance.

Conclusion

The authors found that 63.9% of 24% pre-diabetics and 77% percent of the 50.4% diabetics were at high risk of diabetes and diabetes complications respectively as per the MDRF-IDRS metric. The combination of the IDRS metric with BMI value as well as the HbA1c results can be used for monitoring for diabetes and obesity.

3 Data and Methodology

It is with utmost pleasure that we elucidate upon the dataset made available by the esteemed National Institute of Diabetes and Digestive and Kidney Diseases, which is designed to predict the likelihood of a patient's susceptibility to diabetes based on specific diagnostic measurements. This dataset is a carefully selected subset of a larger database, wherein all patients are of Pima Indian ancestry and female, aged 21 years or older. The rationale behind this selection criteria is owing to the fact that the Pima Indians exhibit the highest incidence of type 2 diabetes worldwide (Booth, Nourian, Weaver, Gull, & Kamimura, 2017).

Our analysis aims to comprehensively examine the most significant factors that contribute to diabetes onset. The dataset comprises nine variables, which were meticulously chosen to ensure their relevance in predicting diabetes. We shall now deliberate on the chosen variables in detail.

Table 1: Summary statistics of the dataset

Statistic	N	Mean	St. Dev.	Min	Max
Pregnancies	392	3.301	3.211	0	17
Glucose	392	122.628	30.861	56	198
BloodPressure	392	70.663	12.496	24	110
SkinThickness	392	29.145	10.516	7	63
Insulin	392	156.056	118.842	14	846
BMI	392	33.086	7.028	18.200	67.100
DiabetesPedigreeFunction	392	0.523	0.345	0.085	2.420
Age	392	30.865	10.201	21	81

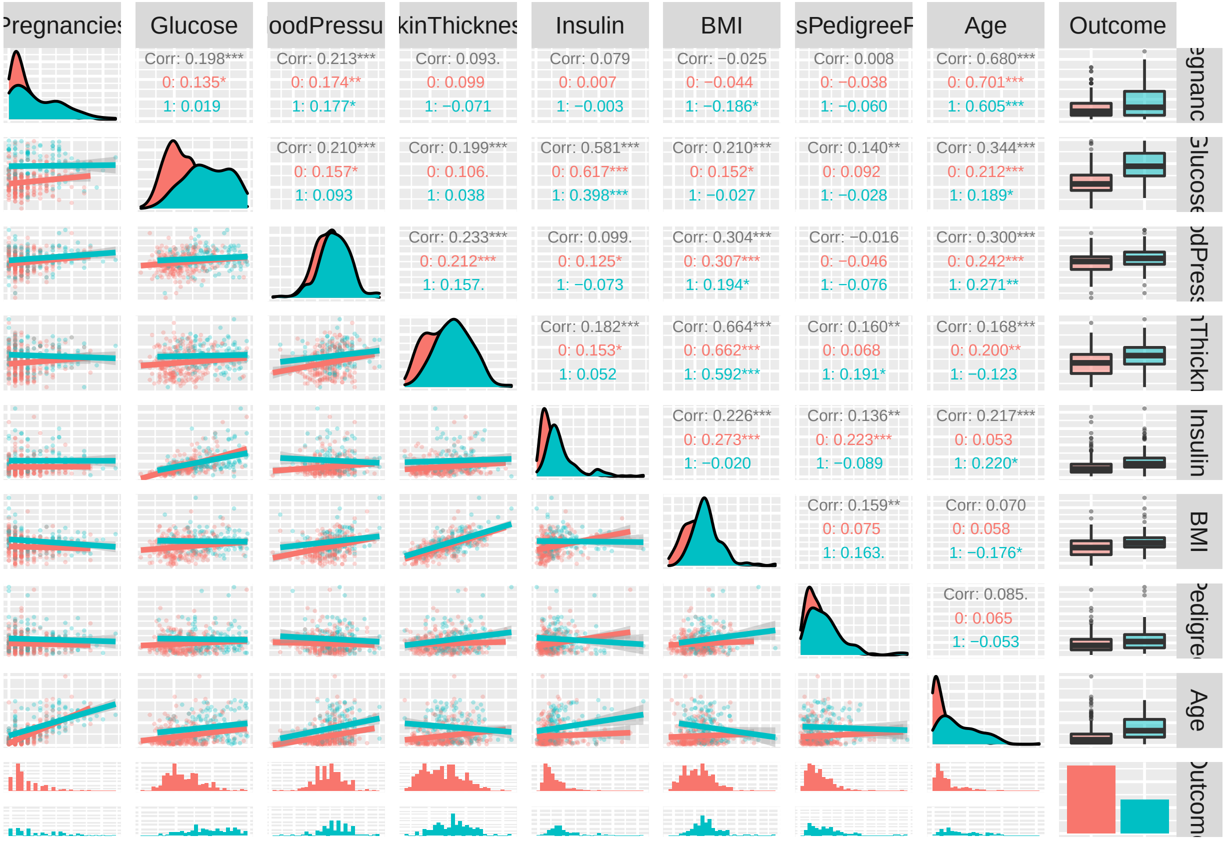

Figure 1: Correlation Plot

3.1 Outcome

The first variable is the dependent variable, “Outcome”, which takes binary values of 0 or 1, signifying non-diabetic and diabetic status, respectively.

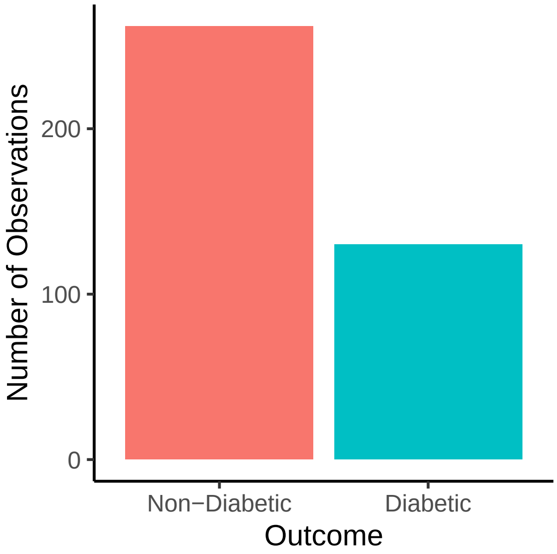

Figure 2: Outcome Variable Frequency

The visualization under consideration provides a clear representation of the distribution of observations across two categories of an outcome variable. Specifically, we observe that out of the total 768 observations in the dataset, 500 individuals do not have diabetes, while 268 individuals do have diabetes.

3.2 Pregnancies

The variable “Pregnancies” records the total number of pregnancies that the patient has experienced. Prior research has established a positive correlation between gestational diabetes mellitus and multiple pregnancies, with women with four or more pregnancies displaying a higher susceptibility to diabetes (Vaajala et al., 2023).

Pregnancy Dummy

In analyzing our dataset, we have identified the presence of a categorical variable denoting pregnancy, with possible values ranging from 0 to 17. In order to appropriately model the influence of this variable, we have created a set of indicator variables that partition the range of pregnancies into three distinct intervals: [0,4], (4,10], and (10,17]. Within this framework, we have elected to designate the range [0,4] as the reference group, against which we will compare the effects of other pregnancy categories.

Var1	Freq
[0,4]	285
(4,10]	91
(10,17]	16

3.3 Glucose

The "Glucose" variable records the plasma glucose concentration (mg/dL) over two hours following an oral glucose tolerance test. High fasting blood glucose concentrations are an indicator of an increased likelihood of developing diabetes. Normal fasting blood glucose concentrations range from 70 mg/dL to 100 mg/dL. Lifestyle changes are recommended if the value lies between 100-125 mg/dL, and diabetes is diagnosed if it exceeds this range.

3.4 Blood Pressure

The "BloodPressure" variable records the average diastolic blood pressure level (mm/Hg) of the patient. Patients with hypertension exhibit insulin resistance and are at a higher risk of developing diabetes. A staggering 66% of diabetic individuals worldwide have blood pressure levels greater than 130/80 mm Hg or are diagnosed with hypertension (Petrie, Guzik, & Touyz, 2018).

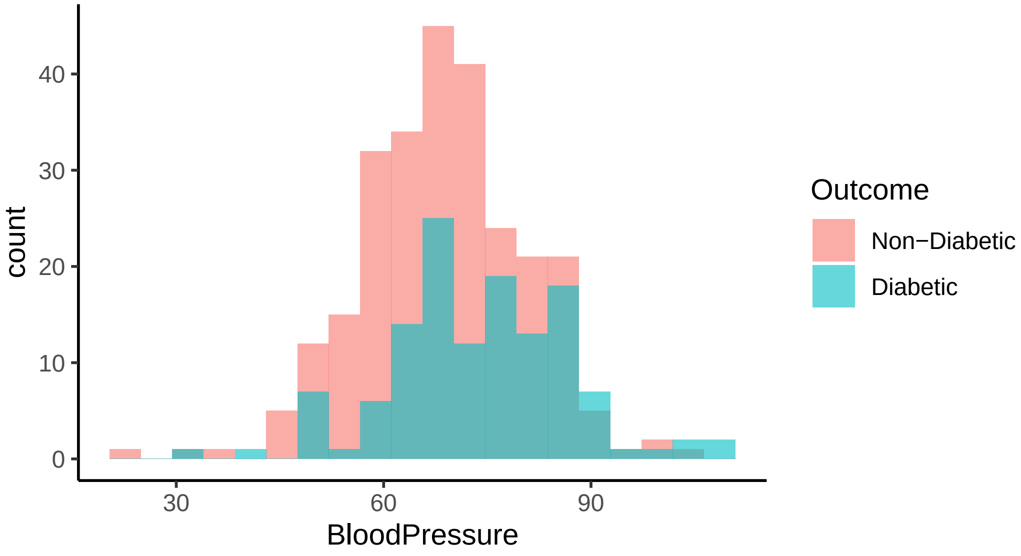

Figure 3: Blood Pressure Distribution

Blood Pressure Dummy

Our analysis has incorporated a classification scheme that distinguishes between individuals with varying blood pressure levels. Specifically, we have defined four distinct categories based on the observed blood pressure values: hypotension for values between 0 and 60, normal for values between 60 and 80, pre-hypertension for values between 80 and 89, and hypertension for values exceeding 90. For the purposes of our analysis, we have selected the normal blood pressure range (60-80) as the reference group, against which we will assess the relative effects of other blood pressure categories.

Var1	Freq
Normal	231
Hypo	83
Pre	56
Hyper	22

3.5 Skin Thickness

The "SkinThickness" variable records the triceps skin fold thickness (mm) of the patient. In type-2 diabetes, insulin resistance affects subcutaneous tissue thickness. Previous research has indicated an association between diabetic patients and increased skin thickness (Jain, Pandey, Lahoti, & Rao, 2013).

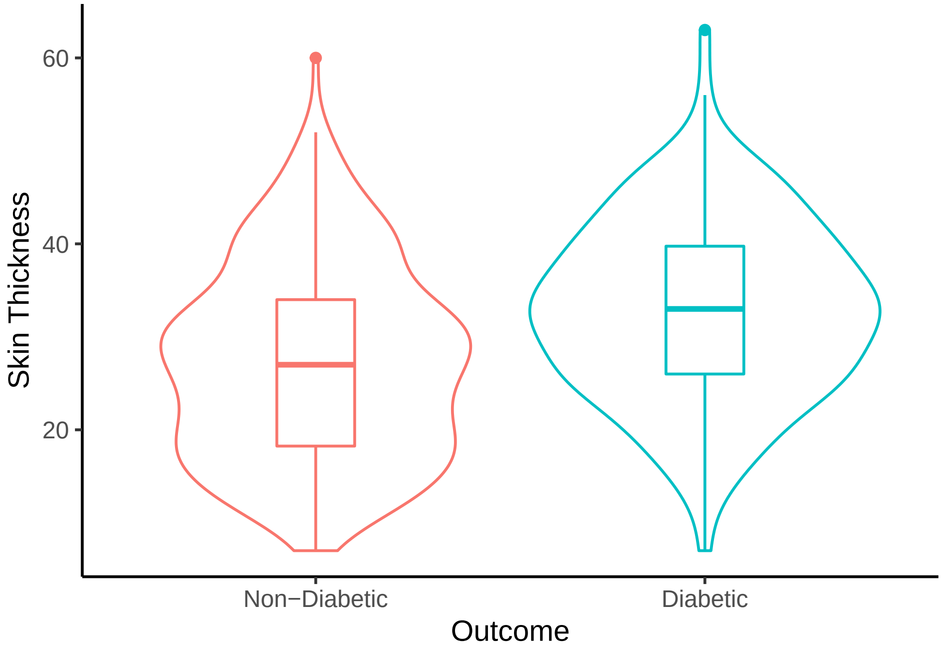

Figure 4: Skin Thickness Distribution

The mean skin thickness in diabetic patients is found to be higher than that of non-diabetic individuals. The visual representation of the distribution through a violin plot indicates a distinct pattern, with the skin thickness of diabetic patients exhibiting a high degree of concentration around the median value, falling within the range of 30-40 mm. This finding suggests a potential avenue for further exploration of the mechanisms underlying the observed relationship between diabetes and skin thickness.

3.6 Insulin

The “Insulin” variable records the result of the 2-hour serum insulin (µU/mL) level. Healthy individuals exhibit insulin levels between 5 and 15 µU/mL, whereas higher insulin levels are directly correlated with diabetes.

3.7 BMI

The “BMI” variable records the body mass index of the patient. Overnutrition and obesity may result in insulin resistance, leading to increased glucose levels in the blood and eventually, diabetes. Research has shown that individuals with a BMI greater than 25 are more likely to develop diabetes.

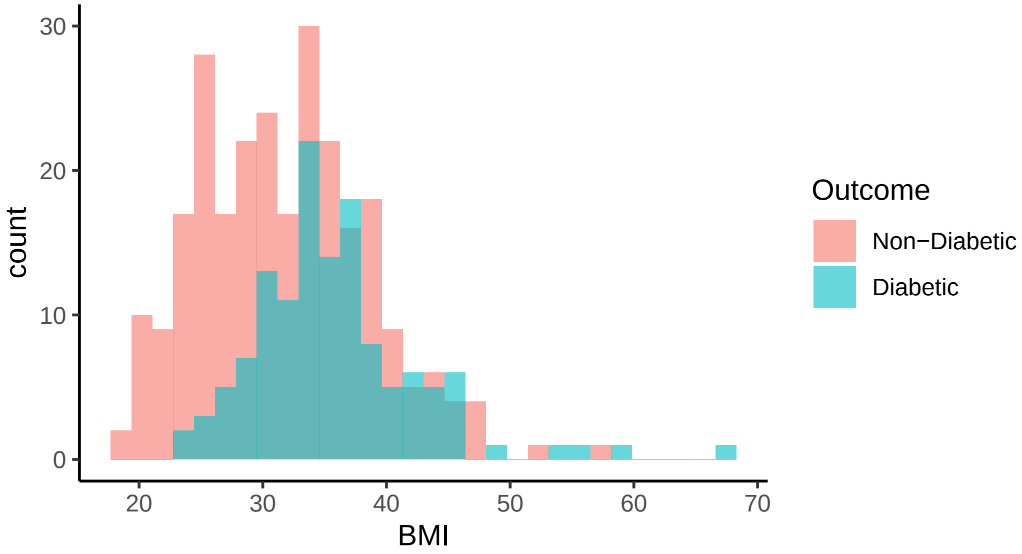

Figure 5: BMI Distribution

In the dataset, the BMI variable has values from 15 to around 70. The maximum number of patients without diabetes lie in the BMI range of 25 to 30. While a majority of those having diabetes have BMI in the range of 35 to 45.

3.8 Diabetes Pedigree Function

The “DiabetesPedigreeFunction” variable measures the likelihood of becoming diabetic based on family history. A family tree is generated based on which a function is generated, assigning a high or low value based on the number of diabetic patients in the patient’s family history (Massaro, Magaletti, Cosoli, Giardinelli, & Leogrande, 2022).

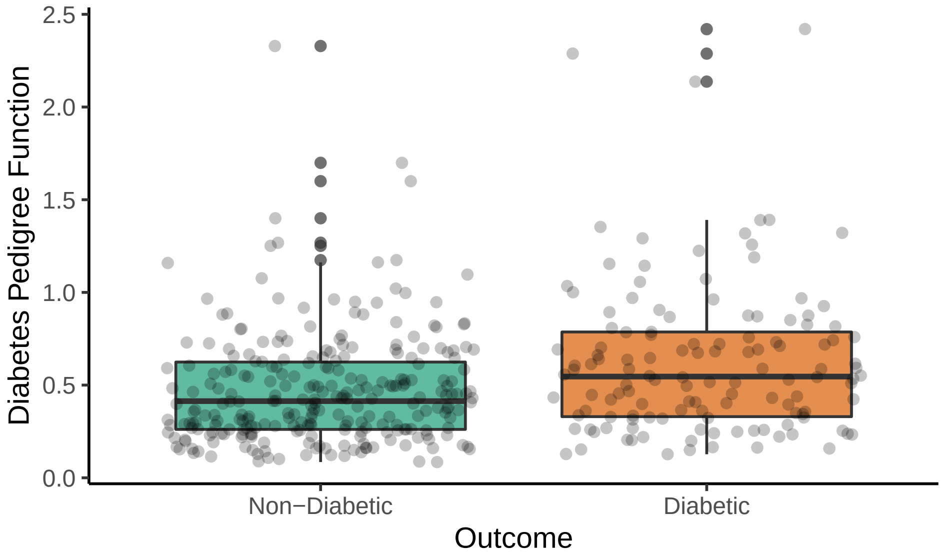

Figure 6: Diabetes Pedigree Function Distribution vs Output

This function gives the risk or probability of getting diabetes based on family history, and diabetic patients have a higher mean value of diabetes pedigree function. This data indicates that there is a higher chance of the patient having diabetes if someone in his/her pedigree was diabetic.

3.9 Age

Finally, the "Age" variable records the age of the patient. With age, individuals are more likely to develop multiple medical conditions, including high blood pressure and high cholesterol. Current literature suggests that middle-aged individuals are more likely to develop diabetes due to these indirect factors.

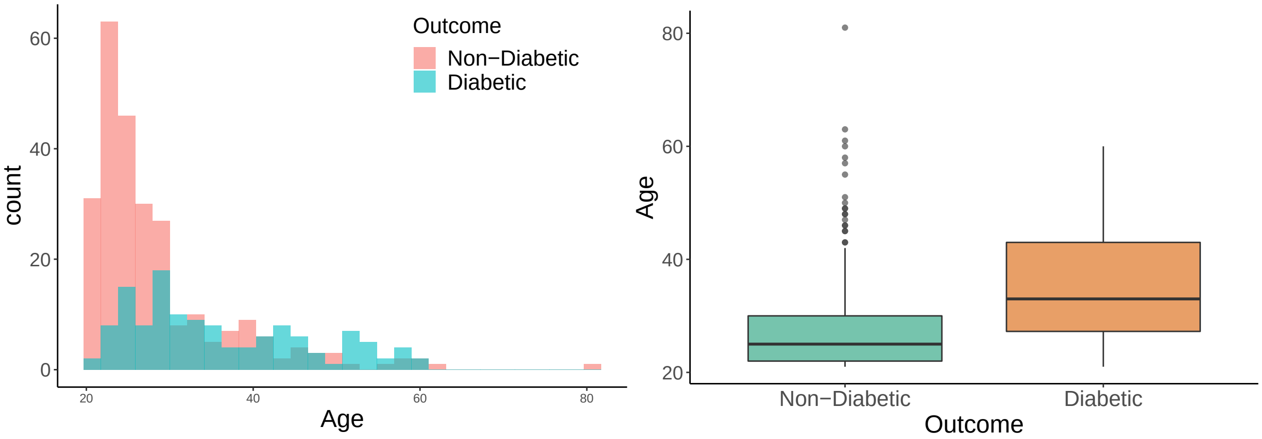

Figure 7: Age distribution

In conclusion, this dataset is a valuable resource for predicting diabetes onset based on specific diagnostic measurements. Our analysis of the chosen variables underscores the significance of hypertension, insulin resistance, and family history in determining an individual's susceptibility to diabetes.

3.10 Methodology

In consideration of the cross-sectional nature of the data, where the outcome variable assumes binary values of either 1 (diabetic) or 0 (non-diabetic), the use of a linear probability model is justified. The regression coefficients reflect the extent to which a one-unit change in an independent variable influences the probability of the dependent variable.

4 Regression

Table 4: Robust Regression Results

	Dependent variable:
	Outcome
PregnancyRange(4,10]	−0.034
	(0.060)
PregnancyRange(10,17]	0.237**
	(0.114)
SkinThickness	0.002
	(0.003)
Glucose	0.007***
	(0.001)
BloodDummyHypo	−0.014
	(0.052)
BloodDummyPre	0.013
	(0.059)
BloodDummyHyper	0.115
	(0.090)
Insulin	−0.0001
	(0.0002)
log(DiabetesPedigreeFunction)	0.098***

	(0.032)
Age	0.008^{***} (0.003)
BMI	0.007^{*} (0.004)
Constant	-1.039^{***} (0.147)
Observations	392
Residual Std. Error	0.333 (df = 380)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 5: Regression Statistics

	Value
R-square	0.3467
Adj. R-square	0.3278
F-statistic	419.2018
p-value for F-test	0.0000

4.1 Interpretation

- The probability of a patient being diabetic increases by 0.007 in case his/her plasma glucose concentration (mg/dL) increases by 1 mg/dL. This result is along the expected lines.
- The probability of a person being diabetic increases by $(0.098/100)\% \approx 0.00098\%$ as their Diabetes Pedigree Function increases by 1 unit. This is again on expected lines of common literature.
- The probability of a person being diabetic increases by 0.008 with an increase of 1 year in the person's age. This supports our initial assumptions that with increase in age, the body is more likely to be affected by multiple medical conditions, including high blood pressure and high cholesterol and hence more prone to be diabetic.
- If a patient's pregnancy range lies between 10 to 14, the probability of her being diabetic increases by 0.237.
- The increase in BMI by 1 unit increases the probability of the person being diabetic by 0.007.
- The p-value for F test is close to zero which means we can reject our null hypothesis at all significance levels implying that all the variables together are jointly significant.

Significant Variables

- At **99% Confidence Interval**: Glucose, log(DiabetesPedigreeFunction), Age
- At **95% Confidence Interval**: PregnancyRange(10,17]
- At **90% Confidence Interval**: BMI

Insignificant Variables

- Insulin, Skin Thickness, Blood Pressure, PregnancyRange(4,10]

4.2 Diagnostics

Test for Heteroskedasticity

We're employing the **Breusch-Pagan Test** in order to check for heteroskedasticity.

$$H_0 : \text{Residuals have constant variance}$$
$$H_a : \text{Residuals have variable variance}$$

Table 6: BP Test Results

statistic	43.5433
p.value	8.733105e-06
parameter	11
method	studentized Breusch-Pagan test

Given the negligible p-value, the null hypothesis is rejected, and it may be inferred that Heteroskedasticity prevails in the system. In response, we have applied Robust Linear Regression as a remedy.

Test for Omitted Variable Bias

We're going to employ the **Ramsey RESET Test** (Ramsey, 1969) to test for Omitted Variable Bias.

$$H_0 : \text{Model has no omitted variables}$$
$$H_a : \text{Model has some omitted variables}$$

Table 7: RESET Test Results

df1	1
df2	379
statistic	0.31132

p.value	0.5772008
method	RESET test

With a p-value exceeding the 10% threshold, we are unable to reject the null hypothesis at a 10% level of significance. Thus, we may conclude that our model is not impacted by omitted variable bias.

Multicollinearity

The presence of multicollinearity may be ascertained by computing the **Variation Inflation Factor** for each variable. Typically, a VIF exceeding 10 indicates the presence of multicollinearity amongst the variables.

Table 8: VIF Values

Variables	Tolerance	VIF
PregnancyRange(4,10]	0.593	1.687
PregnancyRange(10,17]	0.760	1.316
SkinThickness	0.532	1.881
Glucose	0.600	1.666
BloodDummyHypo	0.863	1.158
BloodDummyPre	0.902	1.109
BloodDummyHyper	0.907	1.103
Insulin	0.641	1.561
log(DiabetesPedigreeFunction)	0.955	1.047
Age	0.482	2.073
BMI	0.505	1.978

In light of the VIF values of less than 10 for all variables, it is reasonable to assert with confidence the absence of multicollinearity amongst the said variables.

5 Conclusion

In light of the aforementioned empirical inquiry, our endeavor was to explore the relation between distinct variables such as age, BMI, family history, insulin levels and the probability of receiving a favorable diagnosis for diabetes mellitus amongst Indian individuals.

Drawing on the tenets of (robust) regression analysis performed on cross-sectional data, we demonstrated that blood glucose levels, age, family history and high number of pregnancies exhibited statistically significant effects on the likelihood of a positive diagnosis. These results are aligned with extant literature on the subject that also postulated analogous relationships between the said variables and the eventual diagnosis.

That being said, we must underscore several limitations that warrant attention. First and foremost, we utilized a rudimentary cross-sectional dataset that precludes us from accounting for several non-quantitative variables such as sedentary behavior, psychological and physiological stress levels, among others, that could significantly impact blood glucose levels and thereby affect the final diagnosis. Second, we employed a linear probability model for the analysis, which may not be the most appropriate choice for binary dependent variables.

6 References

Acharya, A. S., Singh, A., & Dhiman, B. (2017). Assessment of diabetes risk in an adult population using indian diabetes risk score in an urban resettlement colony of delhi. *The Journal of the Association of Physicians of India*, *65*(3), 46—51. Retrieved from http://europepmc.org/abstract/MED/28462543

Booth, C., Nourian, M. M., Weaver, S., Gull, B., & Kamimura, A. (2017). Policy and social factors influencing diabetes among pima indians in arizona, USA. *Policy*, *7*(3).

Dudeja, P., Singh, G., Gadekar, T., & Mukherji, S. (2017). Performance of indian diabetes risk score (IDRS) as screening tool for diabetes in an urban slum. *Medical Journal Armed Forces India*, *73*(2), 123–128. Retrieved from https://www.sciencedirect.com/science/article/pii/S0377123716301137

Hlavac, M. (2022). *Stargazer: Well-formatted regression and summary statistics tables*. Bratislava, Slovakia: Social Policy Institute. Retrieved from https://CRAN.R-project.org/package=stargazer

Jain, S. M., Pandey, K., Lahoti, A., & Rao, P. K. (2013). Evaluation of skin and subcutaneous tissue thickness at insulin injection sites in indian, insulin naive, type-2 diabetic adult population. *Indian journal of endocrinology and metabolism*, *17*(5), 864–870. LWW.

Kahn, R., Alperin, P., Eddy, D., Borch-Johnsen, K., Buse, J., Feigelman, J., Gregg, E., et al. (2010). Age at initiation and frequency of screening to detect type 2 diabetes: A cost-effectiveness analysis. *The Lancet*, *375*(9723), 1365–1374. Elsevier.

Khan, M. M., Sonkar, G. K., Alam, R., Mehrotra, S., Khan, M. S., Kumar, A., & Sonkar, S. K. (2017). Validity of indian diabetes risk score and its association with body mass index and glycosylated hemoglobin for screening of diabetes in and around areas of lucknow. *Journal of Family Medicine and Primary Care*, *6*(2), 366. Wolters Kluwer–Medknow Publications.

Massaro, A., Magaletti, N., Cosoli, G., Giardinelli, V., & Leogrande, A. (2022). The prediction of diabetes. *Available at SSRN 4135264*.

Misra, A., Pandey, R., Rama Devi, J., Sharma, R., Vikram, N., & Khanna, N. (2001). High prevalence of diabetes, obesity and dyslipidaemia in urban slum population in northern india. *International journal of obesity*, *25*(11), 1722–1729. Nature Publishing Group.

Nethan, S., Sinha, D., & Mehrotra, R. (2017). Non communicable disease risk factors and their trends in india. *Asian Pacific journal of cancer prevention: APJCP*, *18*(7), 2005. Shahid Beheshti University of Medical Sciences.

Petrie, J. R., Guzik, T. J., & Touyz, R. M. (2018). Diabetes, hypertension, and cardiovascular disease: Clinical insights and vascular mechanisms. *Canadian Journal of Cardiology*, *34*(5), 575–584. Elsevier.

Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, *31*(2), 350–371. [Royal Statistical Society, Wiley]. Retrieved March 4, 2023, from http://www.jstor.org/stable/2984219

Sanjeevaiah, A., Sushmitha, A., & Srikanth, T. (2019). Prevalence of diabetes mellitus and its risk factors. *IAIM*, *6*(3), 319–324.

Vaajala, M., Liukkonen, R., Ponkilainen, V., Kekki, M., Mattila, V. M., & Kuitunen, I. (2023). Higher odds of gestational diabetes among women with multiple pregnancies: A nationwide register-based cohort study in finland. *Acta Diabetologica*, *60*(1), 127–130. Springer.

Yajnik, C., Fall, C., Vaidya, U., Pandit, A., Bavdekar, A., Bhat, D., Osmond, C., et al. (1995). Fetal growth and glucose and insulin metabolism in four-year-old indian children. *Diabetic Medicine*, *12*(4), 330–336. Wiley Online Library.