Project Report: Diabetes Analysis in Pima Native American Females

Yidan Zhu

Scientific Background

The Pima, or “ River People” are a group of Native Americans living in central and southern Arizona. There have been community marriages for over 2000+ years. According to a research paper by Schulz and Chaudhari (2015) about high-risk populations, native heritage is associated with higher diabetes prevalence, and the situation with the Pima was quite unique. By 1970, the prevalence of type II diabetes was about 40% among Pimas aged 35 and older and currently affects about half of all Pimas over age 35[1]. This population has been under continuous studies since 1965 by the National Institute of Diabetes and Digestive and Kidney Diseases because of its high incidence rate of diabetes. Each community resident over 5 years of age was asked to undergo a standardized examination every two years. Diabetes was diagnosed according to World Health Organization Criteria; that is, if the 2-hour post-load plasma glucose was at least 200 mg/dl (11.1 mmol/l) at any survey examination or if the Indian Health Service Hospital serving the community found a glucose concentration of at least 200 mg/dl during the course of routine medical care[2]. This data set provided a well-validated data resource in which to explore prediction of diabetes in a longitudinal manner.

Description of Variables

In this diabetes research, patients who volunteered were from Gila River Community. The data was originally taken from the National Institute of Diabetes and Digestive and Kidney Disease[2]. Dataset had a sample size of 768 female patients who were at least 21 years old. The dataset contains 8 integer variables shown in Table 1 and Table 2. Those variables were chosen because they have been found to be significant risk factors for diabetes among Pimas or other populations[2].

Detailed descriptions of each integer variable were displayed in Table 2. All integer variables are continuous, whereas the outcome variable is binary, returning either 0 or 1, indicating not having or having diabetes respectively. As shown in Table 2, variables “glucose” and “insulin” are directly explaining the diagnosis of diabetes, which may not be significantly meaningful to analyze them. Therefore, “Glucose” and “Insulin” were excluded completely for the following analysis. In this project, “pregnancies” is the main variable of interest, and the remaining variables: “Blood Pressure”, “Skin Thickness”, “BMI”, “Age” and “Diabetes Pedigree Function” are the confounder variables.

Table 1: This is a partial table displaying a capped dataset consisting of 5 observations with 8 integer variables and 1 response variable

Pregnancies	Glucose (GTIT)	BloodPressure (mm Hg)	SkinThickness (mm)	Insulin (µU/ml)	BMI (weight in kg/(height in m)^2)	DiabetesPedigreeFunction	Age (years)	Outcome
1	103	80	11	82	19.4	0.491	22	0
1	101	50	15	36	24.2	0.526	26	0
5	88	66	21	23	24.4	0.342	30	0
8	176	90	34	300	33.7	0.467	58	1
7	150	66	42	342	34.7	0.718	42	0

Table 2: Each statement from the left column contains the description of the corresponding variable of the right column

Variables	Description of each variable
Pregnancies	Numbers of Pregnancies
Glucose	Plasma Glucose Concentration at 2 Hours in an Oral Glucose Tolerance Test
Blood pressure	Diastolic Blood Pressure
Skin thickness	Triceps Skin Fold Thickness
Insulin	2-Hour Serum Insulin
BMI	Body Mass Index
Diabetes Pedigree Function	Function which scores likelihood of diabetes based on family history
Age	Age in years
Outcome	Binary outcome: 0 for No diabetes, 1 for Diabetes

Questions of interest

1. To analyze the influence of numbers of pregnancies on diabetes
2. To analyze how age affects the influence of pregnancies on diabetes

Initial Analysis

Correlation

This heat map shows the correlations by the shade of red and purple. The darker the color, the stronger the correlation. Red represents a positive correlation while purple indicates a negative correlation between variables. As shown in Figure 1, “Age” and “Pregnancies”, as well as “BMI” and “Skin Thickness” have relatively strong positive correlations with each other.

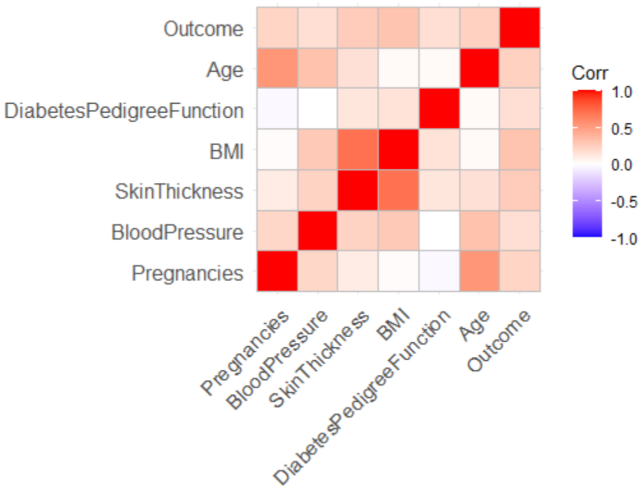


Figure 1: The heat map is showing the correlation among all the variables

Imputation of data

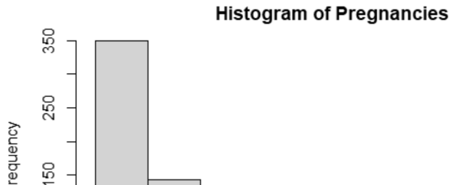
There is some incomplete data in the dataset, i.e. when Skin Thickness =0, BMI=0 and Blood Pressure=0, that needs to be treated in order to avoid misleading conclusions. For “BMI” and “Blood Pressure”, there were only 10-30 out of 768 missing values. Thus, mean values were calculated separately for “BMI” and “Blood Pressure” and the missing value 0s were replaced by their corresponding mean values. Since there were 227 out of 768 missing values for “Skin Thickness”, this variable was treated in a more complex manner. First, a linear regression model was built based on the other confounders, and “Skin Thickness” is the response variable in this model. Imputation of “Skin Thickness” was conducted by replacing the missing value 0s with the predicted response values.

Table 3: The generated linear regression model was utilized for imputing missing values in “Skin Thickness”, the model has an adjusted R^2 value of 0.4

	Imputation Model		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-6.926	-12.060 – -1.791	**0.008**
Age	0.115	0.042 – 0.188	**0.002**
BloodPressure	-0.011	-0.078 – 0.057	0.759
BMI	1.005	0.883 – 1.127	**<0.001**
DiabetesPedigreeFunction	0.609	-1.647 – 2.866	0.596
Observations	460		
R^2 / R^2 adjusted	0.410 / 0.405		

Histogram

The displayed histograms show the distribution of numbers of pregnancies and distribution of having diabetes vs not having diabetes.



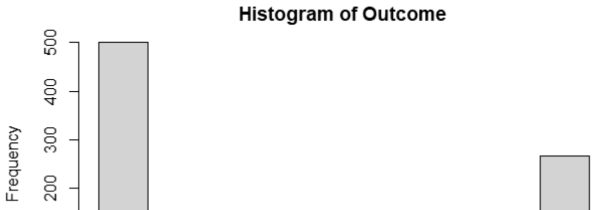
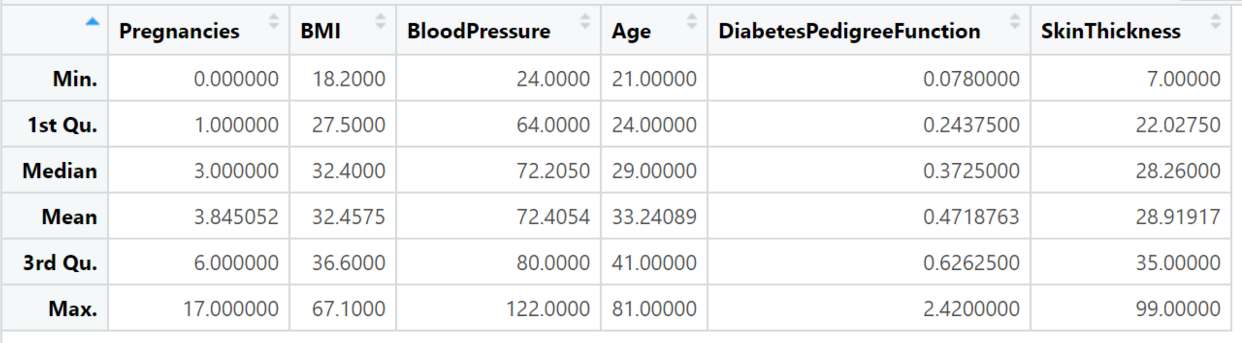


Figure 2: The histogram to the left shows the distribution of numbers of pregnancies with frequency on the x-axis.

Figure 3:The histogram to the right shows the frequency of the outcome. 0s indicate not having diabetes and 1s indicate having diabetes

After the imputation of data, the minimum value of “BMI” was 18.2. The minimum of “Age” was 21 since only female patients who were at least 21 years old volunteered in this research. The range of “Pregnancies” was from 0 to 17, with a mean value of 3.84. Note that this specific population, the Pimas, has been under continuous study since 1965 by the National Institute of Diabetes and Digestive and Kidney Diseases. For a population that has a cultural tradition of community marriage, having many children was not uncommon back then.

Table 4: The generated table shows the mean values, medians, minimum values and maximum values, as well as the 1st and 3rd quartiles of each variable



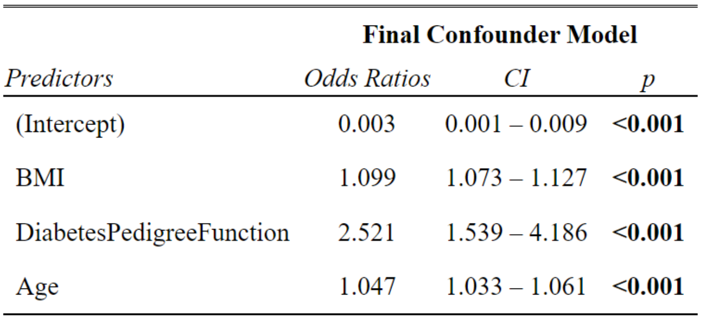
	Pregnancies	BMI	BloodPressure	Age	DiabetesPedigreeFunction	SkinThickness
Min.	0.000000	18.2000	24.0000	21.00000	0.0780000	7.00000
1st Qu.	1.000000	27.5000	64.0000	24.00000	0.2437500	22.02750
Median	3.000000	32.4000	72.2050	29.00000	0.3725000	28.26000
Mean	3.845052	32.4575	72.4054	33.24089	0.4718763	28.91917
3rd Qu.	6.000000	36.6000	80.0000	41.00000	0.6262500	35.00000
Max.	17.000000	67.1000	122.0000	81.00000	2.4200000	99.00000

Model Selection

Backward Selection

Since the response was binary, a logistic model from generalized linear regression was chosen to perform data analysis. The initial model was fitted with only confounders which were “Skin Thickness”, “BMI”, “Age” and “Diabetes Pedigree Function”. A backward selection with threshold p=0.2 was applied to this initial confounder model, and the result suggested keeping “BMI”, “Age” and “Diabetes Pedigree Function” as the remaining confounders. With the remaining confounders, the final confounder model was obtained(see figure 3).

Table 5: The final confounder model and a summary of exponential ORs, exponential CIs, and p-value



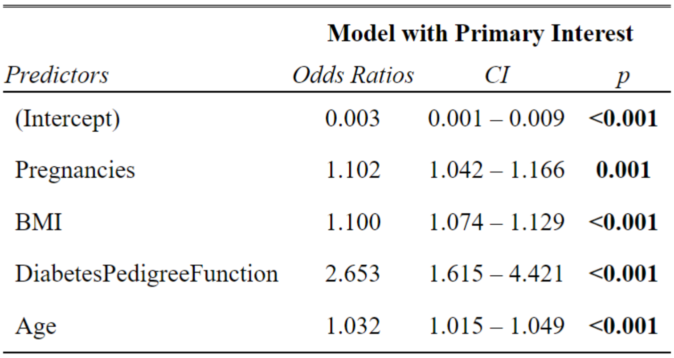
	Final Confounder Model		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.003	0.001 – 0.009	**<0.001**
BMI	1.099	1.073 – 1.127	**<0.001**
DiabetesPedigreeFunction	2.521	1.539 – 4.186	**<0.001**
Age	1.047	1.033 – 1.061	**<0.001**

Main Variable of Interest and Interaction

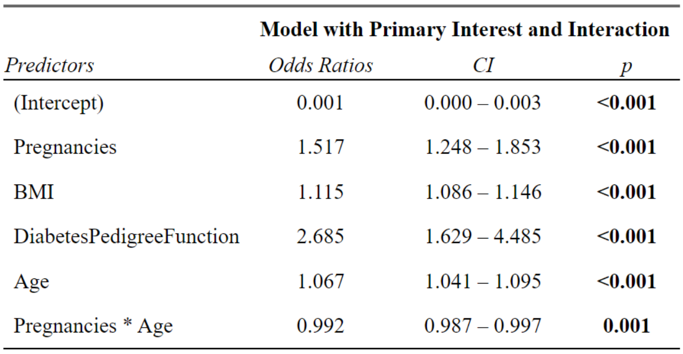
The main variable of interest “Pregnancies” was added to the final confounder model. The summary of this new model showed the p-values of all the predictors were highly significant(less than 0.05). The interaction term of “Age: Pregnancies” was added to the model with the variable of primary interest to analyze how age would potentially affect the influence of pregnancies on diabetes.

Table 6: The exponentialed CIs, p-value and exponential ORs of the model with primary variable

Table 7:The exponentialed CIs, p-value and exponential ORs of the model with primary variable and interaction term

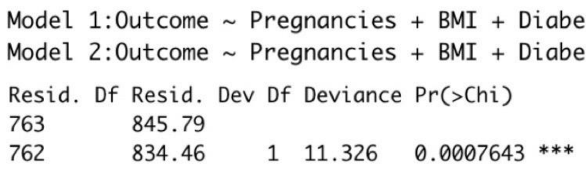


	Model with Primary Interest		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.003	0.001 – 0.009	**<0.001**
Pregnancies	1.102	1.042 – 1.166	**0.001**
BMI	1.100	1.074 – 1.129	**<0.001**
DiabetesPedigreeFunction	2.653	1.615 – 4.421	**<0.001**
Age	1.032	1.015 – 1.049	**<0.001**



	Model with Primary Interest and Interaction		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.001	0.000 – 0.003	**<0.001**
Pregnancies	1.517	1.248 – 1.853	**<0.001**
BMI	1.115	1.086 – 1.146	**<0.001**
DiabetesPedigreeFunction	2.685	1.629 – 4.485	**<0.001**
Age	1.067	1.041 – 1.095	**<0.001**
Pregnancies * Age	0.992	0.987 – 0.997	**0.001**

Next, ANOVA test with LRT was conducted to compare the model without the interaction term and the model with the interaction term. The null hypothesis of the test was that the coefficient for interaction in the more complex model was zero. The test result showed a p-value less than 0.05 which implied that there was not enough evidence to support the null hypothesis. Therefore, the model with the interaction term “Age: Pregnancies" provided a better fit for the data set. The best model so far was *Outcome ~ Pregnancies + BMI + Diabetes Pedigree Function + Age + Age: pregnancies* and the following analysis will be conducted for this model.



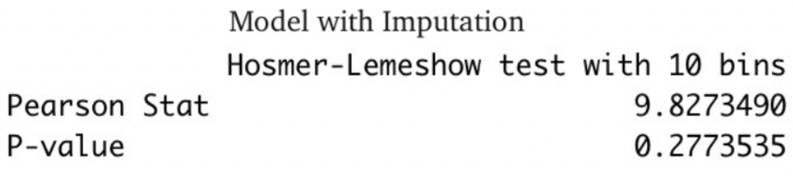
```
Model 1:Outcome ~ Pregnancies + BMI + DiabetesPedigreeFunction + Age
Model 2:Outcome ~ Pregnancies + BMI + DiabetesPedigreeFunction + Age + Pregnancies:Age

Resid. Df Resid. Dev Df Deviance Pr(>Chi)
763       845.79
762       834.46     1  11.326   0.0007643 ***
```

Figure 4: ANOVA test and a p-value of 0.00076

Goodness of Fit & Diagnostics

To determine how closely the model mirrors observed data, the Hosmer Lemeshow Goodness of fit test was performed. The null hypothesis of the test was that the model fitted well with the data. As the p-value was greater than 0.05, the null hypothesis was not rejected. Thus, there was not enough evidence to conclude that the model did not fit the data well.



```
                    Model with Imputation
                    Hosmer-Lemeshow test with 10 bins
Pearson Stat                             9.8273490
P-value                                  0.2773535
```

Figure 5: Hosmer-Lemeshow test and a p-value of 0.277

To assess the model's compliance with its assumptions, model diagnostics were performed. Based on the Residual vs Fitted plot, the residual was approximately equally distributed around 0. With the Normal Q-Q plot, the deviance deviated from the normal distribution at the lower tail.

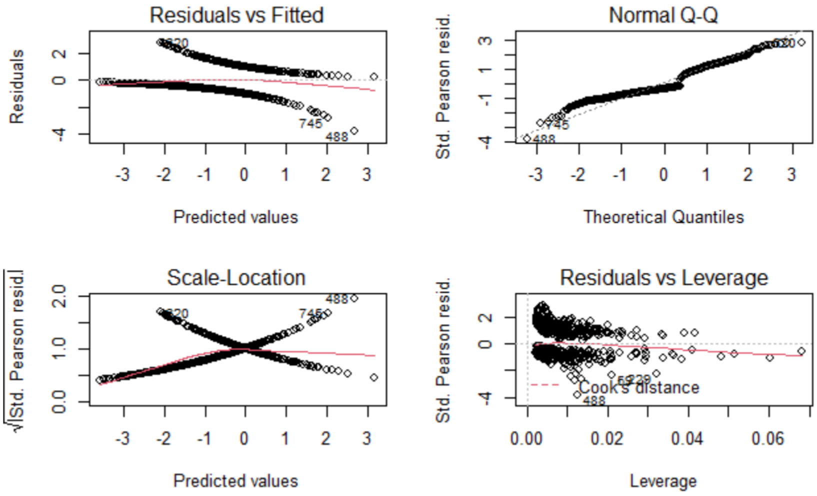


Figure 6: Top left subplot: the Residual vs Fitted plot. Top right subplot:Normal Q-Q plot.

Bottom left subplot: Scale-Location plot. Bottom right subplot: Residuals vs Leverage plot

The outlier detection with Cook's distance was plotted, and the plot suggested three potential outliers: observation #59, # 229 and #488. These three outliers had deviated Diabetes Pedigree Function values, which means that the likelihood score of diabetes is based on family history. The mean value of Diabetes Pedigree Function was 0.47, whereas #59 had a value of 1.78, #229 had a value of 2.32 and # 488 had a value of 1.16. Since these three outliers had large Diabetes Pedigree Function values but were still informative about the influence of pregnancies on diabetes, these three outliers were included in the final model.

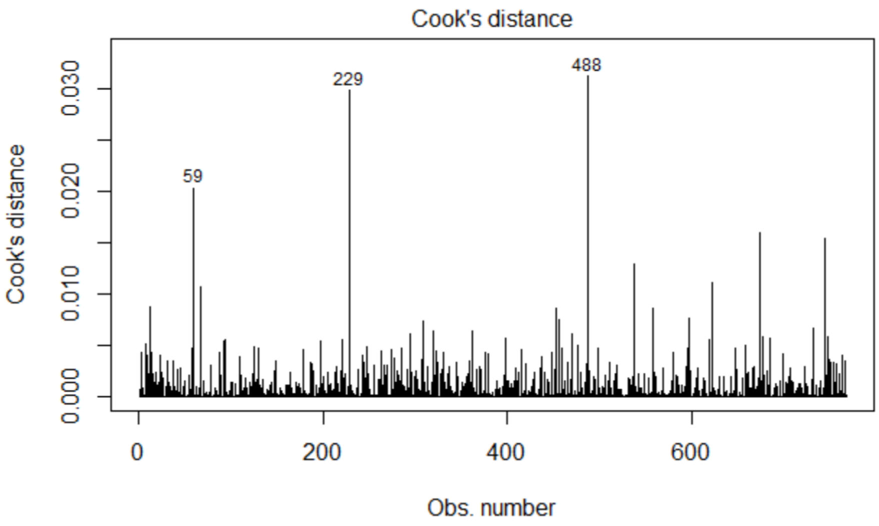
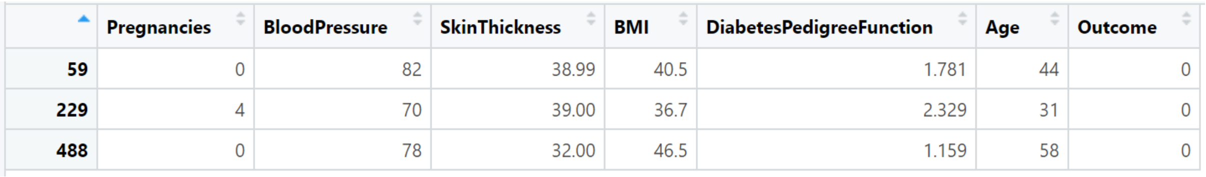


Figure 7: Cook's Distance for outlier detection. #59, #229 and #488 were shown as outliers

Table 8: The description of observations #59, #229 and #488 after imputation

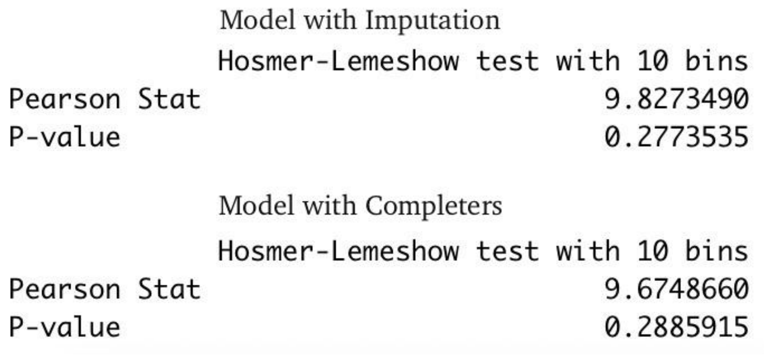


	Pregnancies	BloodPressure	SkinThickness	BMI	DiabetesPedigreeFunction	Age	Outcome
59	0	82	38.99	40.5	1.781	44	0
229	4	70	39.00	36.7	2.329	31	0
488	0	78	32.00	46.5	1.159	58	0

Sensitivity Analysis

To determine if imputation of missing data affects the outcome, a sensitivity analysis was conducted. 462 complete case observations were extracted from the original 768 observations. The same model selection process was applied to the complete case data. A confounder model was fitted, and backward selection (with a threshold p=0.2) was conducted. Then, the main variable of interest "Pregnancies" and the interaction term "Age: Pregnancies" were introduced to the model. The goodness of fit test, outlier detection, and model diagnostics were also performed in the same order as in the previous section. As a result, the obtained final model provided a good fit with complete case data, the diagnostics plots were similar to the previous final model. Comparing final models with imputation data and complete case data, the difference of each predictor's odds ratio (exponential of estimated coefficients) was insignificant. Therefore, the imputation model was consistent with the original data. Moreover, it has a larger sample size (n=768) than the model fitted with complete case data (n=460). Thus, the final model was chosen to be the model with imputation: *Outcome ~ Pregnancies + BMI + Diabetes Pedigree Function + Age + Age: pregnancies.*

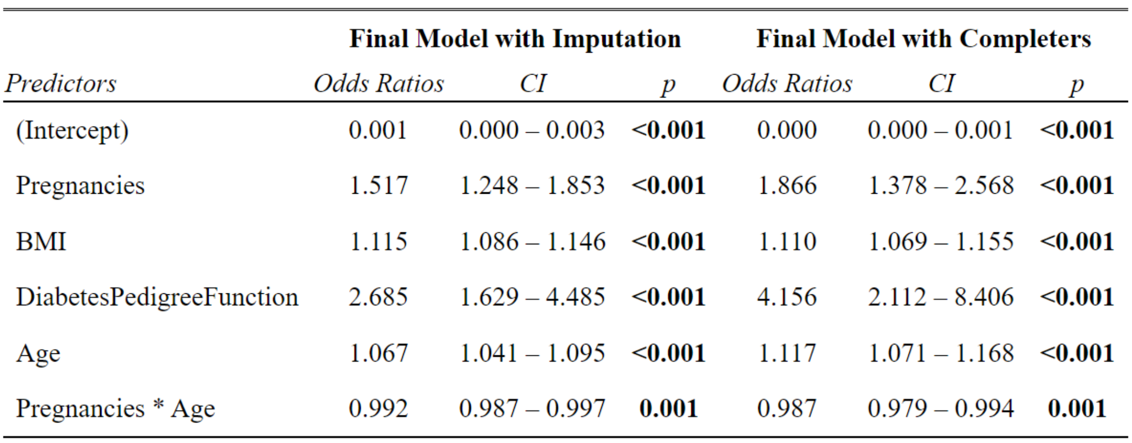
Table 9: Comparison of Hosmer-Lemeshow test result between model with imputation and model with completers



```
                 Model with Imputation
                 Hosmer-Lemeshow test with 10 bins
Pearson Stat                            9.8273490
P-value                                 0.2773535

                 Model with Completers
                 Hosmer-Lemeshow test with 10 bins
Pearson Stat                            9.6748660
P-value                                 0.2885915
```

Table 10: Comparison table of model with imputation and model with completers in terms of exponential ORs, exponential CIs and p-values



	Final Model with Imputation			**Final Model with Completers**		
Predictors	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*
(Intercept)	0.001	0.000 – 0.003	**<0.001**	0.000	0.000 – 0.001	**<0.001**
Pregnancies	1.517	1.248 – 1.853	**<0.001**	1.866	1.378 – 2.568	**<0.001**
BMI	1.115	1.086 – 1.146	**<0.001**	1.110	1.069 – 1.155	**<0.001**
DiabetesPedigreeFunction	2.685	1.629 – 4.485	**<0.001**	4.156	2.112 – 8.406	**<0.001**
Age	1.067	1.041 – 1.095	**<0.001**	1.117	1.071 – 1.168	**<0.001**
Pregnancies * Age	0.992	0.987 – 0.997	**0.001**	0.987	0.979 – 0.994	**0.001**

Conclusion & Discussion

According to Table 7, for every additional pregnancy, there is a 51.7% (95% CI: 24.8% - 85.3%, LRT) increase in the odds of having diabetes, controlling for the effects of others. Since there have been marriages within the Pima Native American communities for over 2000+ years, and the communities have been marked for high risk of diabetes in heritage[1], every additional pregnancy would be very risky for the females. When age increases by one unit, there is a 6.7% (95% CI: 4.1% - 9.5%, LRT) increase in the odds of having diabetes, controlling for the effects of others. A literature study at *Johns Hopkins Medicine* also confirms that aging increases the risk of having diabetes [3].

When age increases by 10 units, the effect of the number of pregnancies on diabetes decreases by 8% (95% CI: 3% - 13%, LRT), controlling for the effects of others. Even though aging and pregnancy both increase the odds of having diabetes, the effect of pregnancy on diabetes lessens as the females get older. This result could be reasoned that older females have less probability to be pregnant. A woman's peak reproductive years are between the late teens and late 20s. By age 30, fertility (the ability to get pregnant) starts to decline[4].

We are currently treating “Age” as a continuous variable. For further analysis, “Age” can be treated in a categorical format. Specifically, the dataset can be stratified to analyze the effect of pregnancies under different age groups. This might be informative for the targeting Pima females in different age groups for calling more attention to type II diabetes.

References

1. Schulz, L. O., & Chaudhari, L. S. (2015). High-Risk Populations: The Pimas of Arizona and Mexico. *Current obesity reports*, *4*(1), 92–98. https://doi.org/10.1007/s13679-014-0132-9
2. Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. Proceedings of the Annual Symposium on Computer Application in Medical Care, 261–265. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245318/
3. Gestational Diabetes Mellitus (GDM). (2020). Johns Hopkins Medicine. https://www.hopkinsmedicine.org/health/conditions-and-diseases/diabetes/gestational-diabetes
4. Having a Baby After Age 35: How Aging Affects Fertility and Pregnancy. (2020, October). ACOG. https://www.acog.org/womens-health/faqs/having-a-baby-after-age-35-how-aging-affects-fertility-and-pregnancy