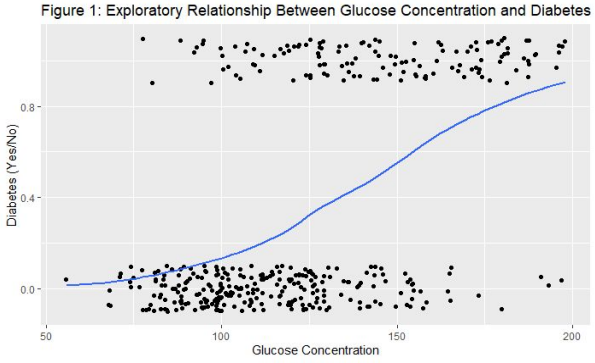
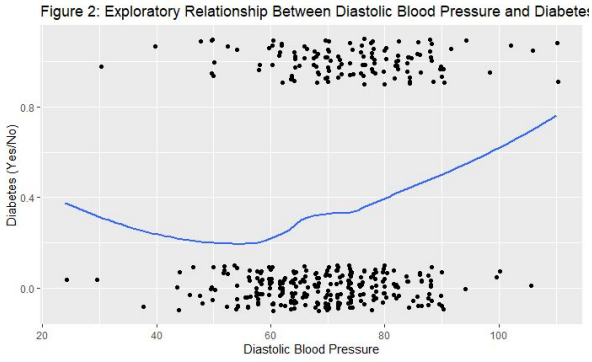
Jacob Andros (section 1) + Chloe Crawford (section 2)

STAT 330: Homework 7 – Diabetes

1. Type 2 diabetes is a disorder that can cause a number of complications in one's body because it prevents the body from making enough insulin to control blood sugar levels. The longer type 2 diabetes goes undiagnosed, the more health complications it causes. For that reason, it is important for individuals and their doctors to better understand what factors are significant in determining how likely one is to have type 2 diabetes. Statistical modeling through logistic regression will help doctors better understand the nature of the relationship of type 2 diabetes with the variables in this dataset. But more importantly, it will also give them the ability to predict whether or not an individual is likely to have type 2 diabetes given their glucose concentration, blood pressure, BMI, and other easy-to-obtain measurements. Improving doctors' understanding of this disease will, in the long run, lead to earlier diagnosis of type 2 diabetes and patients with less complications from the disease.

2. The plots below show exploratory relationships between (1) glucose concentration and diabetes, and (2) diastolic blood pressure and diabetes. It is clear in the first plot that there is some relationship between glucose concentration and having diabetes - it appears that women with higher glucose concentrations are far more likely to have the disease. The second plot does not show as clear of a relationship - it seems possible that, when we fit a model later on in this analysis, blood pressure will not be part of it.





The reason we do not use traditional multiple linear regression for an analysis like this is because in this dataset, the response variable (whether or not someone is diagnosed with type 2 diabetes) is not quantitative, but categorical. Since we want to predict the probability of someone falling into a certain category rather than a raw number, we use logistic regression instead of multiple regression.

3. To create a logistic regression model that is effective in helping doctors predict the likelihood of a patient having diabetes, we used the AIC-minimizing, best subsets variable selection method. We chose to aim to minimize AIC because the doctors are

more concerned with predicting whether or not someone has diabetes than they are with inference, and AIC is designed to make the most effective predictive model. We used the best subsets method (instead of forward or backward selection) because there are only 8 explanatory variables in the dataset, so it only takes a few seconds for the best subsets algorithm to run.

The variables that are deemed most important in explaining the presence of diabetes, according to this selection method, are the glucose concentration, BMI, pedigree (family history of diabetes), and age. It is important to note than the dataset used to obtain this model had any observations with zero's removed from it, as well as anyone over the age of 70.

4. The model created in the last step can be mathematically expressed as follows:

$$\log(p_i / (1-p_i)) = \beta_0 + \beta_1 * \text{glucose} + \beta_2 * \text{BMI} + \beta_3 * \text{pedigree} + \beta_4 * \text{age}$$
$$\text{where } y_i \sim_{iid} \text{Bern}(p_i)$$

In using this model, we are assuming two things. First, we are assuming that the log-odds ratio has a linear (or at least monotonic) relationship with each explanatory variable. Second, we assume that all of the observations used in this dataset are independent of one another. Let's start by examining the first assumption by examining the monotonicity of the log-odds ratio with each of the explanatory variables in the selected model.

Figure 3: Monotonic Relationship Between Glucose Concentration and Diabetes

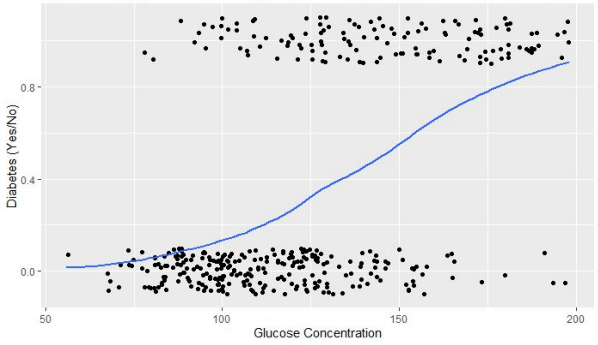


Figure 4: Monotonic Relationship Between BMI and Diabetes

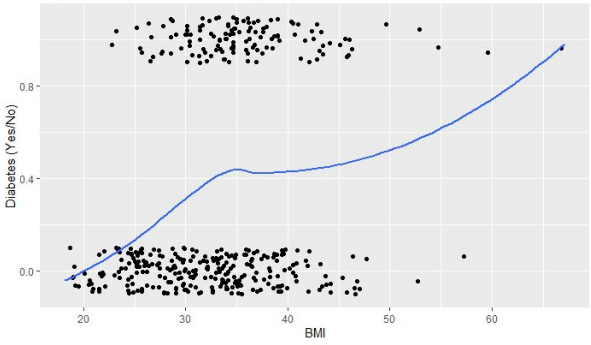


Figure 5: Monotonic Relationship Between Family History and Diabetes

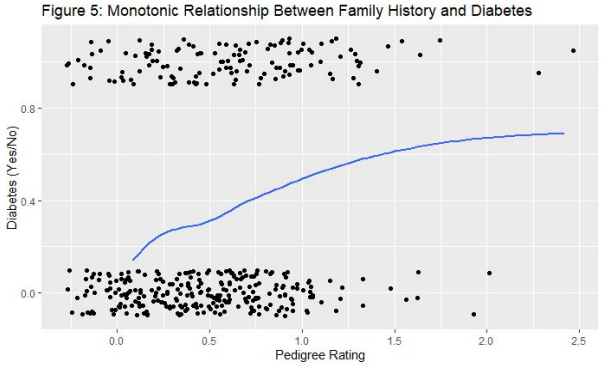
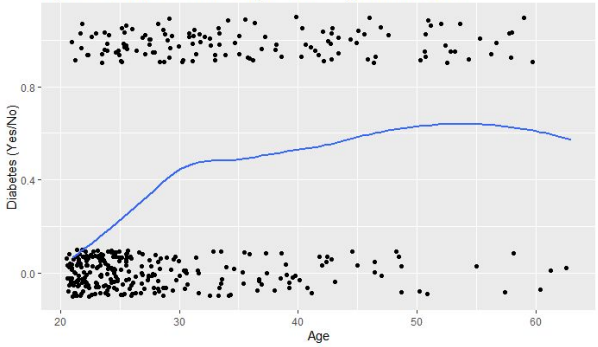


Figure 6: Monotonic Relationship Between Age and Diabetes



Since the smooth curve in each of these four plots is monotonic (there is a slight dip at the end of the age one, but barely), we can say the linearity assumption is met.

As for independence, we don't know much about the data collection methods used, so we will assume that the data was collected from individuals who are not related and whose health is not in any way connected. Therefore, this data meets the independence condition.

5. The logistic regression model we have fitted estimates the four specified explanatory variables to have the following effects on the log-odds ratio of having diabetes (all intervals given are 95% confidence interval).

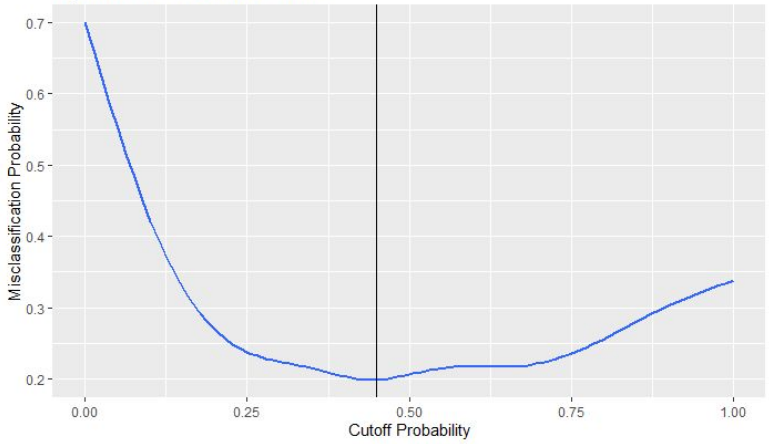
Variable	**Symbol**	**Estimate**	**Interval for β_i**	**Interval for $\exp(\beta_i)$**	**Interval for $100 * (\exp(\beta_i) - 1)$**
Intercept	β_0	-10.23	(-12.49, -8.19)	(.0000038, .00028)	(-99.9996, -99.972)
Glucose	β_1	.036	(.027, .046	(1.027, 1.047)	(2.687, 4.72)
BMI	β_2	.073	(.034, .114)	(1.034, 1.12)	(3.449, 12.049)
Pedigree	β_3	1.08	(.271, 1.92)	(1.311, 6.82)	(31.13, 581.58)
Age	β_4	.06	(.033, .089)	(1.034, 1.093)	(3.398, 9.27)

As an example of how to interpret these estimates for each beta term (slope term), we can examine the glucose concentration variable. According to this model, we would estimate that as a patient's glucose concentration increases by 1 mm hg, they would be e^.036, or 1.04, times more likely to have diabetes. To account for the uncertainty in this

estimate, we can say we are 95% confident that as a patient's glucose level goes up by 1 mm hg, they are between 1.027 and 1.047 times more likely to have diabetes (where 1.027 is e^.027 and 1.047 is e^.046). These interpretations are under the assumption that we are holding all variables besides glucose constant.

6. The threshold for classification that minimizes the misclassification rate is a probability of 0.45. The following plot shows that misclassifications are minimized when we use 0.45 as the "cutoff probability". This means that if the model gives a patient a probability of at least 0.45 for having diabetes, then we would classify them as having the disease.

Figure 7: Misclassification Plot



7. From R, we can find that the area under the ROC curve (AUC) is about 0.863, and the pseudo-r-squared is about 0.31. The confusion matrix for all of the data, along with its summary statistics, are as follows:

Pred.Class --> True.Class (below)	0	1	Sum
0	229	32	261
1	44	86	130
Sum	273	118	391

Sensitivity	0.662
Specificity	0.877
Positive Predictive Value	0.729
Negative Predictive Value	0.839

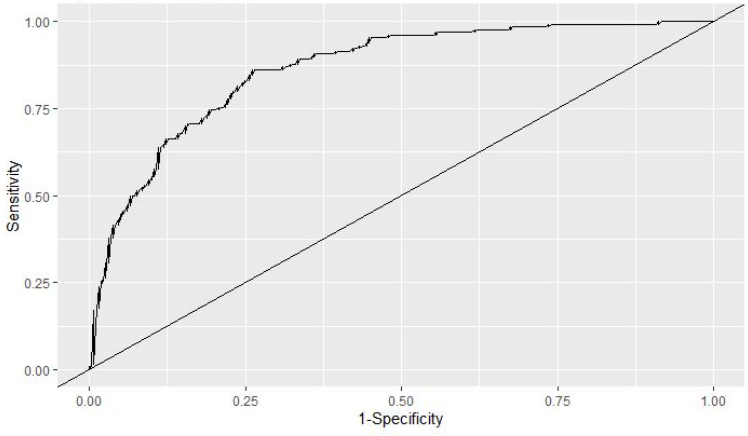
There are a few things to take away from these figures regarding the logistic regression model and the 45% cutoff threshold.

- 66% sensitivity means that based on this dataset, if a patient has diabetes, then there is a 66% chance of us correctly classifying the patient as having diabetes given their measurements.
- 88% specificity means that if a patient does not have diabetes, then there is an 88% chance of us correctly classifying them as not having diabetes based on their measurements.
- The 73% positive predictive value means that if we classify a patient as having diabetes, then there is a 73% chance that they actually have it.
- The 84% negative predictive value means that if we classify a patient as not having diabetes, then there is an 84% chance that they are in fact free of the disease.

The fact that all four of these measures, as well as the area under the curve (AUC), are all well above 50%, is a good indicator that this model is correctly classifying patterns in the dataset more often than not. As another measure of fit we calculated a psuedo-R^2 0.309. This means that 30.9% of the variation in the logged-odds ratio of having diabetes is explained by the model. Although this may seem like a fairly low R^2 value, pseudo-R^2's are notorious for being lower than they should be and we are not concerned.

The AUC term represents the area under the ROC curve. The ROC curve (below) is a plot measuring the predictive effectiveness of the model. The diagonal line is called the "coin-flip model" - meaning that if we had no predictive power whatsoever and randomly guessed whether each patient has diabetes, our curve would guess correctly according to that line, about 50% of the time. In other words, the closer the AUC is to 0.5, the less effective the model is, and the closer it is to 1, the more effective the model is. The AUC of .86, which is closer to 1 than it is to 0.5, means that the model is predicting diabetes status far more effectively than we would be able to by simply guessing.

Figure 8: ROC Curve



8. Using cross validation techniques, we were able to calculate the predictive power of our model. As previously mentioned, the model does predict correctly the majority of cases.. The model predicts the best when diagnosing people who truly do not have the disease as negative for the disease with an average specificity of 87.3%. The model also does well minimizing false negatives with an average negative predictive value of 83.3%. The model predicts the worst when diagnosing people who truly have the disease as positive with an average sensitivity of 64.7% and does slightly better at minimizing false positives with a positive predictive value of 72.1%. Overall, this model attains its goal of prediction quite effectively. The fact that these four values for the cross-validation are all quite close to their correspondents in the confusion matrix for the whole data is a good sign.

Sensitivity	0.647
Specificity	0.873
Positive Predictive Value	0.721
Negative Predictive Value	0.833

9. According to the model, the probability of a patient with these measurements having diabetes is .092. Since this is far below the classification threshold of 0.45, and quite low in general, we do not think this patient has diabetes.