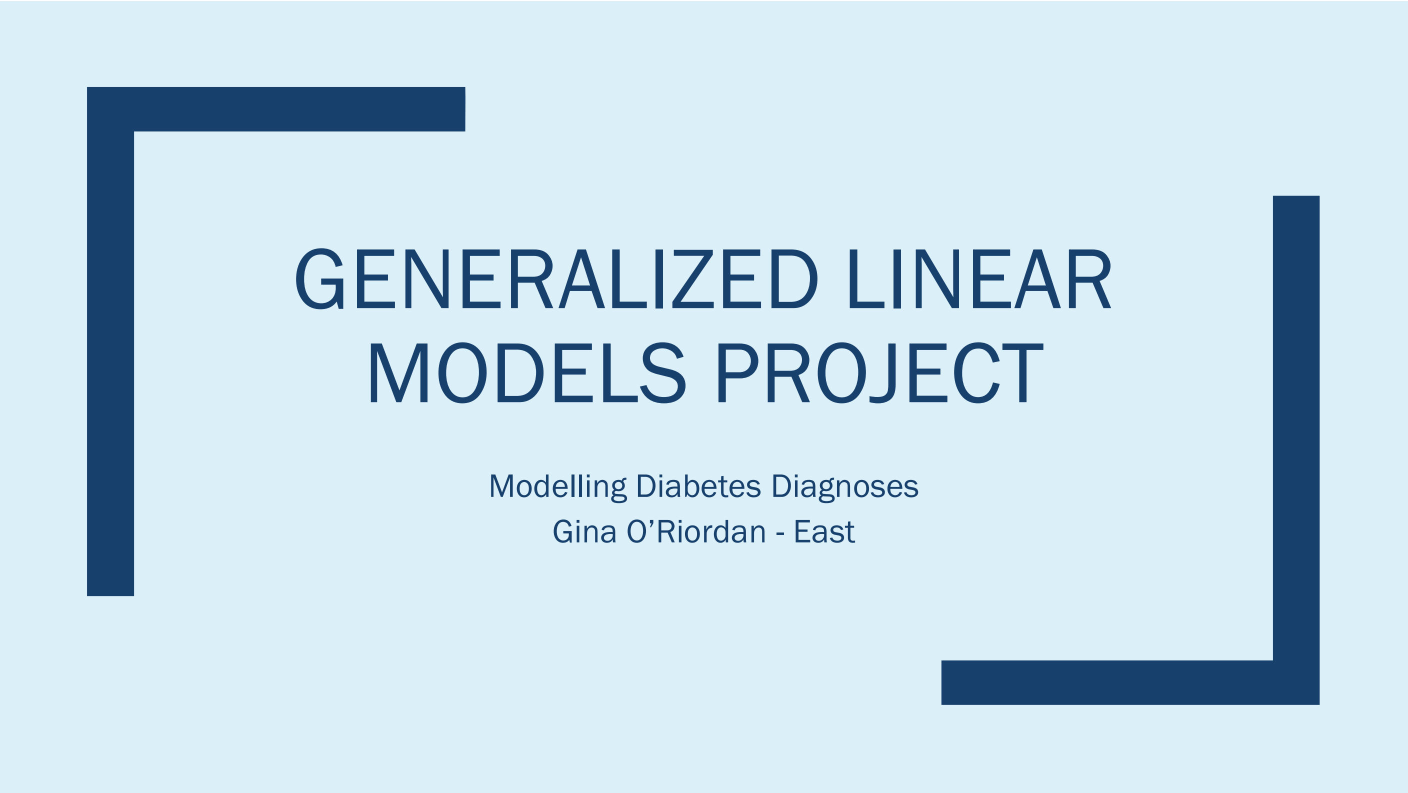
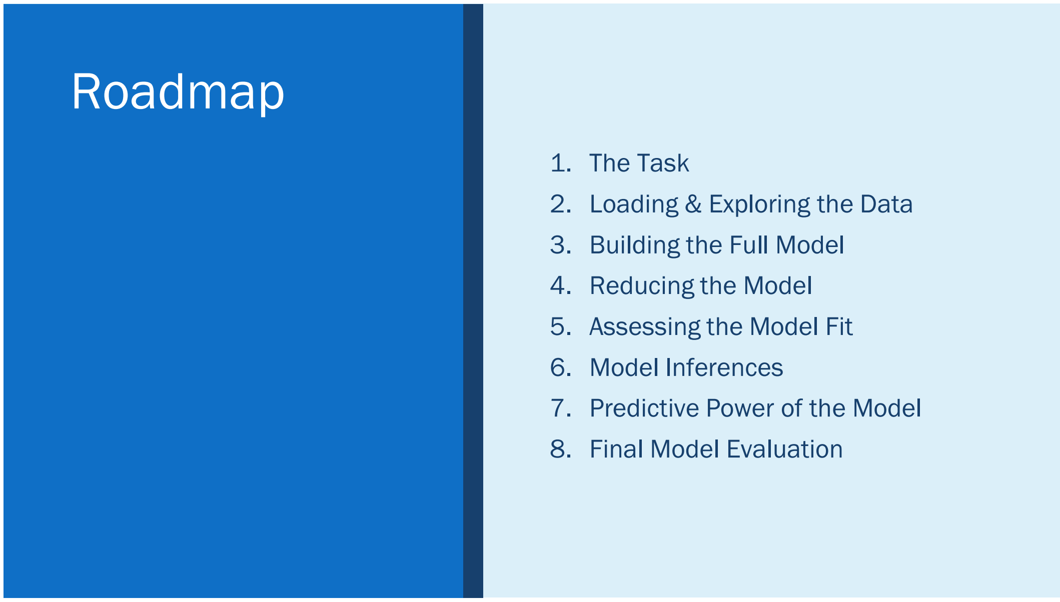


GENERALIZED LINEAR MODELS PROJECT

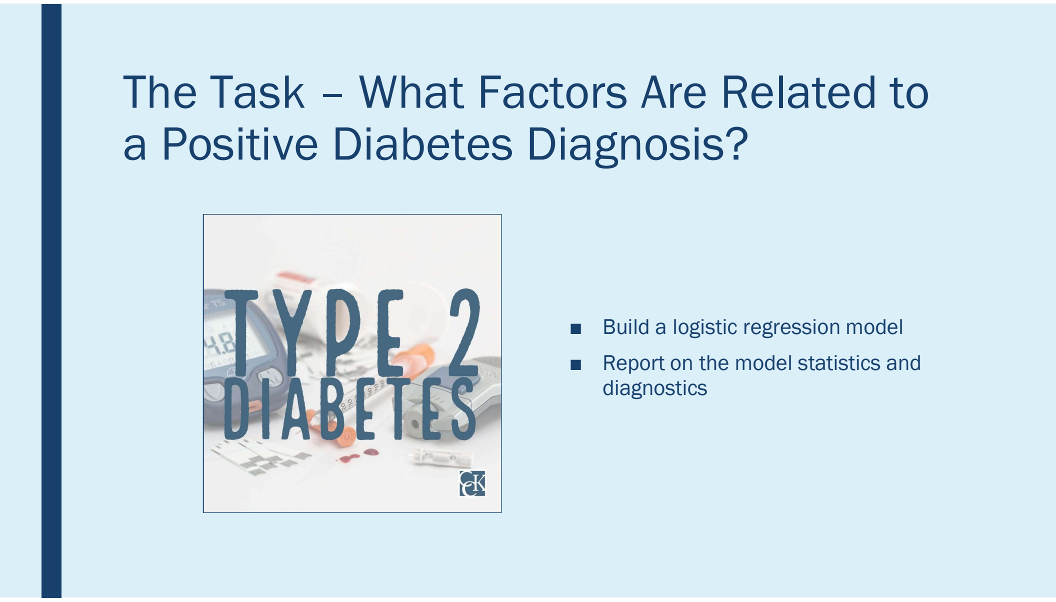
Modelling Diabetes Diagnoses
Gina O’Riordan - East



Roadmap

1. The Task
2. Loading & Exploring the Data
3. Building the Full Model
4. Reducing the Model
5. Assessing the Model Fit
6. Model Inferences
7. Predictive Power of the Model
8. Final Model Evaluation

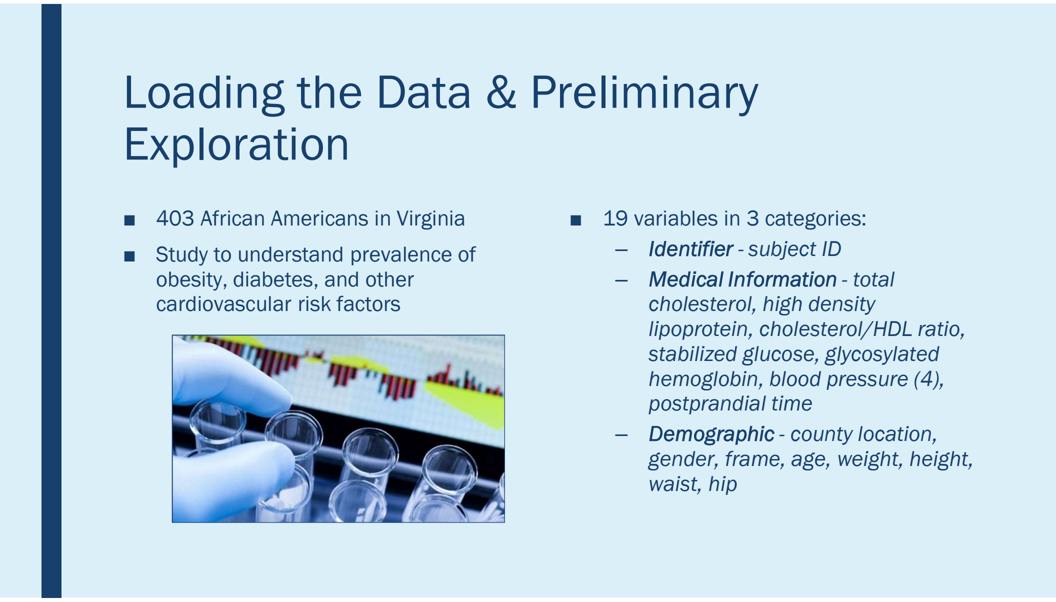
The agenda includes the following topics: (see slide)



The Task – What Factors Are Related to a Positive Diabetes Diagnosis?

- Build a logistic regression model
- Report on the model statistics and diagnostics

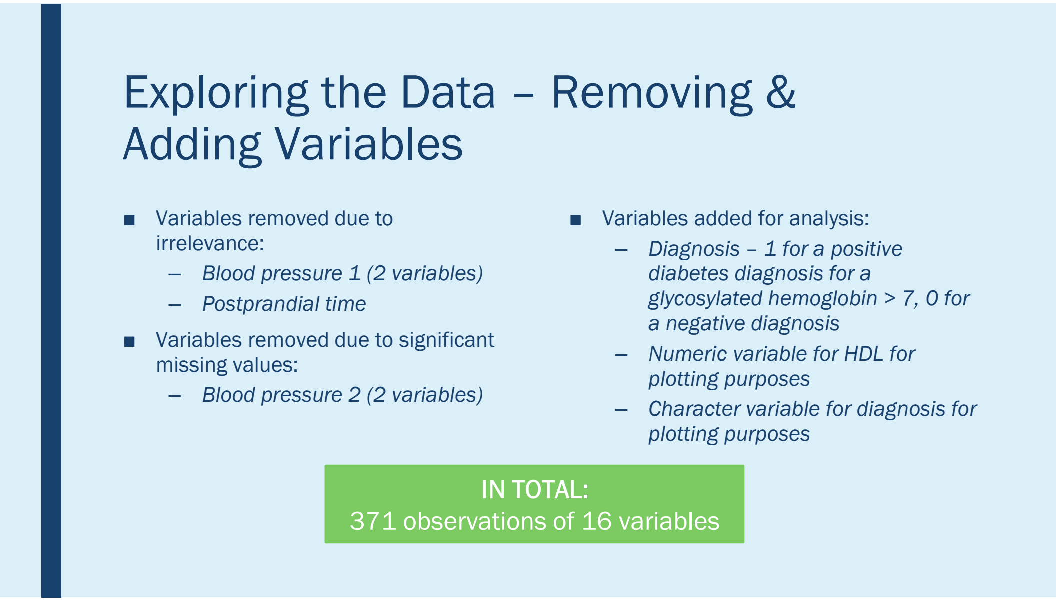
- The goal of this project is to build a logistic regression model to explain what factors are related to a positive diabetes diagnosis. Once that model is built, we'll analyze its statistics and diagnostics to determine its level of adequateness.
- Let's begin by loading and exploring the data.



Loading the Data & Preliminary Exploration

- 403 African Americans in Virginia
- Study to understand prevalence of obesity, diabetes, and other cardiovascular risk factors
- 19 variables in 3 categories:
  - ***Identifier*** *- subject ID*
  - ***Medical Information*** *- total cholesterol, high density lipoprotein, cholesterol/HDL ratio, stabilized glucose, glycosylated hemoglobin, blood pressure (4), postprandial time*
  - ***Demographic*** *- county location, gender, frame, age, weight, height, waist, hip*

- The diabetes data was a study done with 403 African Americans in Virginia to understand the prevalence of obesity, diabetes, and other cardiovascular risk factors.
- There are 19 variables that can be subset into three main categories:
  - Identifier variables – **Subject ID** identifies the individual anonymously.
  - Medical information variables – **Total cholesterol, high density lipoprotein (HDL), and the cholesterol/HDL ratio** are all standard cholesterol measures taken via blood sample. **Stabilized glucose** is a blood sugar value taken while fasting, while **glycosylated hemoglobin** measures blood sugar over a three month period. **Blood pressure** includes systolic and diastolic measuring the pressure on blood vessels with blood flowing out of and into the heart, respectively; note that there are two measures of blood pressure, systolic and diastolic, totaling four variables. **Postprandial time** is the time in minutes from when the blood labs were drawn until the analysis.
  - Demographic variables – **County location** is either Buckingham or Louisa, **gender** is either male or female, and **frame** is be small, medium, or large. **Age** is measured in years, **weight** in pounds, and **height, waist**, and **hip** are measured in inches.

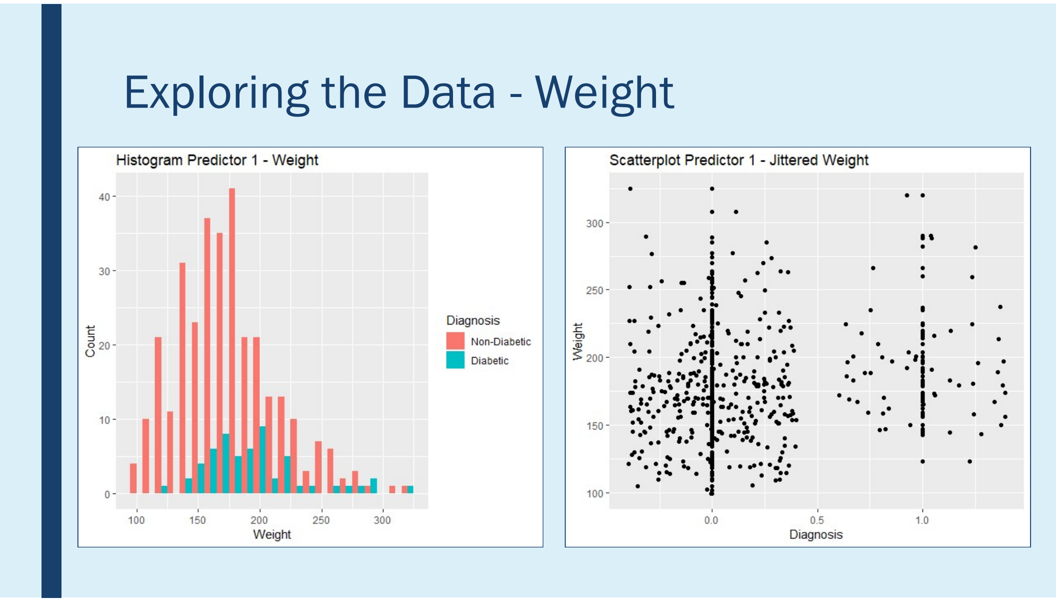


Exploring the Data – Removing & Adding Variables

- Variables removed due to irrelevance:
  - *Blood pressure 1 (2 variables)*
  - *Postprandial time*
- Variables removed due to significant missing values:
  - *Blood pressure 2 (2 variables)*
- Variables added for analysis:
  - *Diagnosis – 1 for a positive diabetes diagnosis for a glycosylated hemoglobin > 7, 0 for a negative diagnosis*
  - *Numeric variable for HDL for plotting purposes*
  - *Character variable for diagnosis for plotting purposes*

IN TOTAL:
371 observations of 16 variables

- Since the ask is to analyze the factors of diabetes diagnosis, blood pressure values in general are of little value and are also removed. The postprandial time is also irrelevant; stabilized glucose is taken while fasting so timing doesn’t matter, but if there was a post-meal glucose reading, lab timing could be important.
- Upon reviewing the data, the second blood pressure values both systolic and diastolic have 262 missing values out of the 403 in the original dataset. Therefore, they were removed for analysis as well.
- Variables that were added for analysis included the main response variable ‘diagnosis.’ This is a binary variable; in other words, an individual is given a ‘1’ for a positive diabetes diagnosis when their glycosylated hemoglobin was greater than 7 and a ‘0’ if the glucose measure was less than 7. Then a numeric variable for HDL and a character variable for diagnosis were added to ease plotting efforts on the following pages.
- Overall, there are 371 observations of 16 variables in the final dataset used for analysis. Let’s begin!

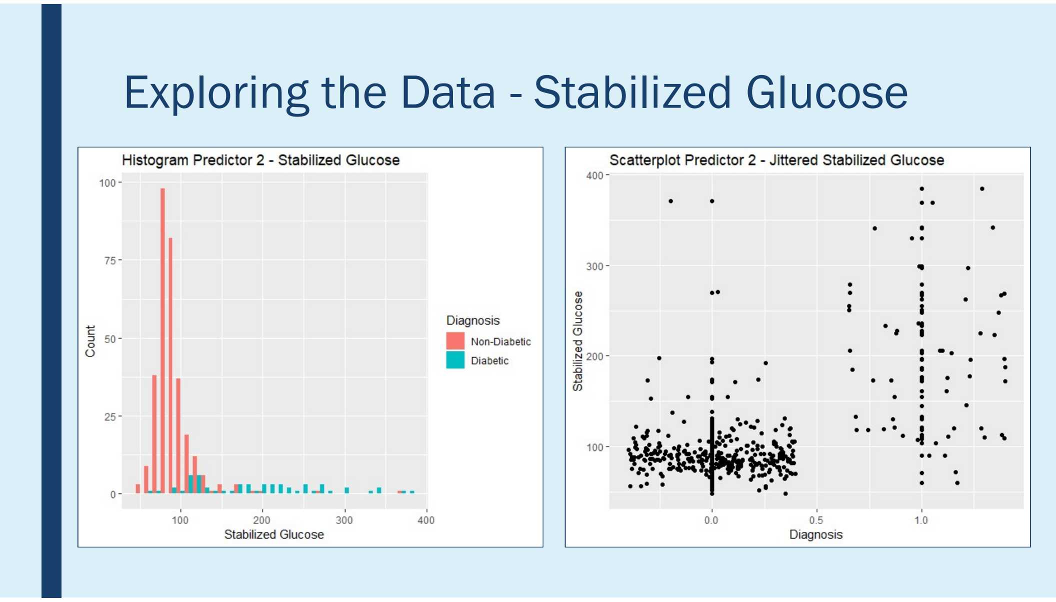


Exploring the Data - Weight

Histogram Predictor 1 - Weight

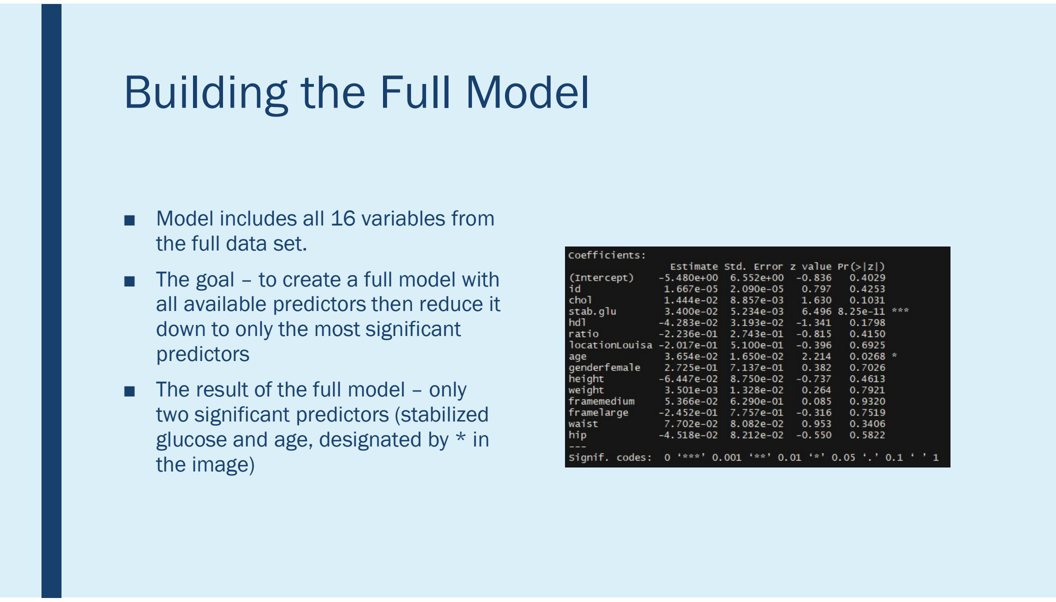
Scatterplot Predictor 1 - Jittered Weight

- To dig into the data a bit more, let’s explore the two most influential factors of type 2 diabetes diagnosis – weight and stabilized glucose values. Here we see two plots of the weight predictor variable.
- The first is an interleaved histogram on the left of the page; this shows the number of people (y-axis) who are either non-diabetic (red) or diabetic (blue) and their respective weight in pounds from left to right (x-axis). We can see from the graph that the average non-diabetic weight is lower (~180 pounds) than a diabetic (~200 pounds) and that there are more non-diabetics overall than diabetics.
- The scatterplot on the right of the page is a bit altered from the original plot. Essentially, because the diagnosis variable on the x-axis is only two values, either 0 or 1, its hard to gain value from the scatterplot when all the dots overlap on either the 0 or 1 x-value. Therefore, we jitter plot and shake the points a bit from side to side, keeping the middle of the graph (~ value 0.5) clear to designate separation between diabetics (1 = diagnosis) and non-diabetics (0 = diagnosis). The y-axis is showing weight values, so overall, this plot is showing the same as the histogram that there are more non-diabetics (more dots at 0 = diagnosis), and that the diabetics have higher weights, mostly above 150 pounds.



Exploring the Data - Stabilized Glucose

- Now we see two plots of the stabilized glucose predictor variable.
- The concepts of the left side's interleaved histogram still apply, but the x-axis along the bottom is now the stabilized glucose of the individuals, increasing as we move to the right. The red non-diabetic individuals have a very frequent low glucose at 100, while the diabetics have a more distributed frequency but with a higher mean stabilized glucose. This is very common as diabetics have uncontrolled blood glucose levels and thus take medication to manage this issue.
- Again, the concepts of the jittered scatterplot to the right are also the same as the weight plot we saw previously. Here the cluster of dots for non-diabetics is very centralized around 100 with few unusual values (or outliers) towards the high glucose values, whereas diabetics have a very far ranging stabilized glucose distribution.

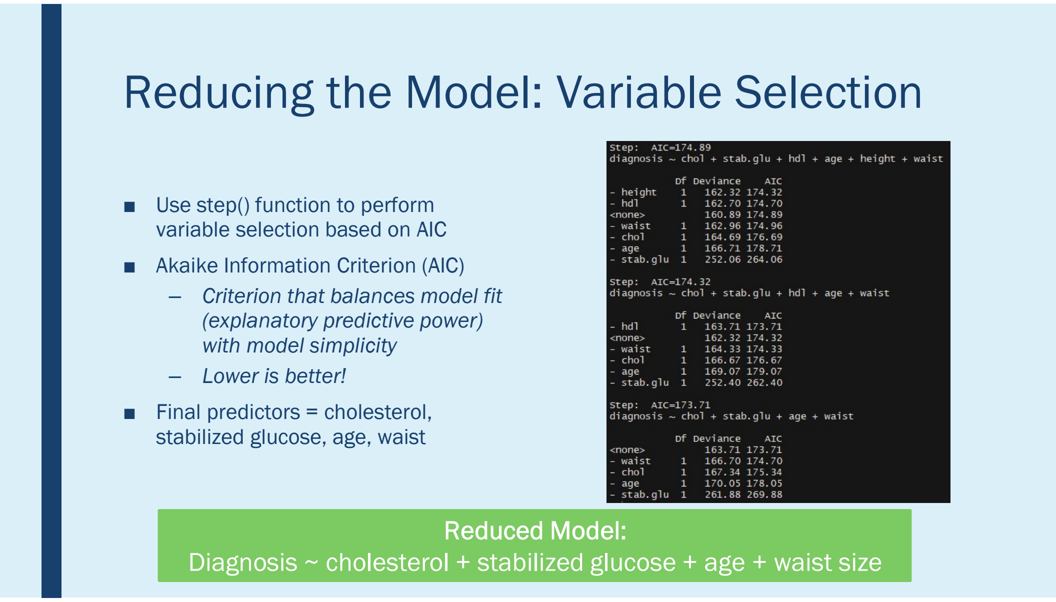


Building the Full Model

- Model includes all 16 variables from the full data set.
- The goal – to create a full model with all available predictors then reduce it down to only the most significant predictors
- The result of the full model – only two significant predictors (stabilized glucose and age, designated by * in the image)

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)    -5.480e+00  6.552e+00  -0.836   0.4029    
id              1.667e-05  2.090e-05   0.797   0.4253    
chol            1.444e-02  8.857e-03   1.630   0.1031    
stab.glu        3.400e-02  5.234e-03   6.496 8.25e-11 ***
hdl            -4.283e-02  3.193e-02  -1.341   0.1798    
ratio          -2.236e-01  2.743e-01  -0.815   0.4150    
locationLouisa -2.017e-01  5.100e-01  -0.396   0.6925    
age             3.654e-02  1.650e-02   2.214   0.0268 *  
genderfemale    2.725e-01  7.137e-01   0.382   0.7026    
height         -6.447e-02  8.750e-02  -0.737   0.4613    
weight          3.501e-03  1.328e-02   0.264   0.7921    
framemedium     5.366e-02  6.290e-01   0.085   0.9320    
framelarge     -2.452e-01  7.757e-01  -0.316   0.7519    
waist           7.702e-02  8.082e-02   0.953   0.3406    
hip            -4.518e-02  8.212e-02  -0.550   0.5822    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- The next step is to begin variable selection by building a full model. This ‘full’ model includes all 16 variables from the full data set.
- The goal here is to create a model with all available predictors then use a function to reduce that model down to only the most significant predictors. The simpler model is what will be analyzed for accuracy.
- Once the full model was run, we found that only two predictors were significant. The image to the right is the model output, with p-values on the far right column, and the asterisks showing the significant predictors.



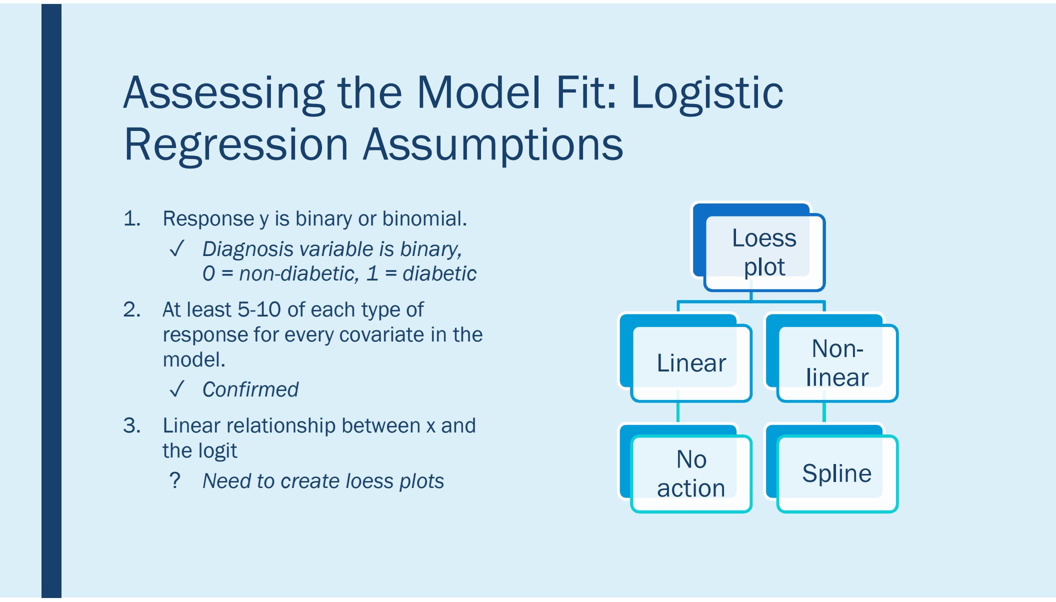
Reducing the Model: Variable Selection

- Use step() function to perform variable selection based on AIC
- Akaike Information Criterion (AIC)
  - *Criterion that balances model fit (explanatory predictive power) with model simplicity*
  - *Lower is better!*
- Final predictors = cholesterol, stabilized glucose, age, waist

```
Step:  AIC=174.89
diagnosis ~ chol + stab.glu + hdl + age + height + waist

           Df Deviance    AIC
- height    1   162.32 174.32
- hdl       1   162.70 174.70
<none>          160.89 174.89
- waist     1   162.96 174.96
- chol      1   164.69 176.69
- age       1   166.71 178.71
- stab.glu  1   252.06 264.06

Step:  AIC=174.32
diagnosis ~ chol + stab.glu + hdl + age + waist

           Df Deviance    AIC
- hdl       1   163.71 173.71
<none>          162.32 174.32
- waist     1   164.33 174.33
- chol      1   166.67 176.67
- age       1   169.07 179.07
- stab.glu  1   252.40 262.40

Step:  AIC=173.71
diagnosis ~ chol + stab.glu + age + waist

           Df Deviance    AIC
<none>          163.71 173.71
- waist     1   166.70 174.70
- chol      1   167.34 175.34
- age       1   170.05 178.05
- stab.glu  1   261.88 269.88
```

Reduced Model:
Diagnosis ~ cholesterol + stabilized glucose + age + waist size

- To reduce the model, we used the step() function, which performs variable selection based on the AIC, or Akaike Information Criterion.
- This criterion balances model fit, also known as explanatory predictive power, with model simplicity. The goal is to have the lowest AIC, demonstrating the optimal subset of predictors to use in the regression model.
- The image to the right demonstrates how the step() function evaluates and show the AIC; the function starts with the full model then removes one predictor per step, only stopping when removing another variable doesn't lower AIC further. The image shows the last three steps of this backward selection and the final formula of the reduced model.
- This model we'll use moving forward includes four predictors – cholesterol, stabilized glucose, age, and waist size. These results make sense; the key indicators of type two diabetes are problems around weight gain. These problems include (but are not limited to) high cholesterol and waist size. Age is also a huge factor for type two diabetics as the majority are diagnosed after the age of 45. Finally, a high stabilized glucose is the number one indicator of diabetes and is the first test a doctor will run to determine if an individual has the disease.



Assessing the Model Fit: Logistic Regression Assumptions

1. Response y is binary or binomial.
   - ✓ *Diagnosis variable is binary, 0 = non-diabetic, 1 = diabetic*
2. At least 5-10 of each type of response for every covariate in the model.
   - ✓ *Confirmed*
3. Linear relationship between x and the logit
   - ? *Need to create loess plots*

Loess plot

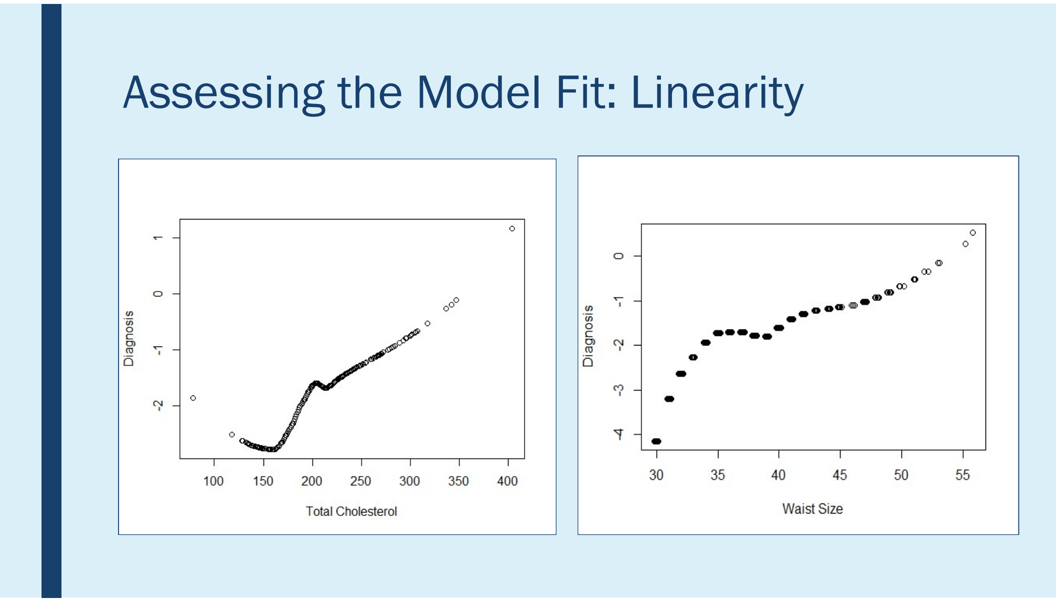
Linear

Non-linear

No action

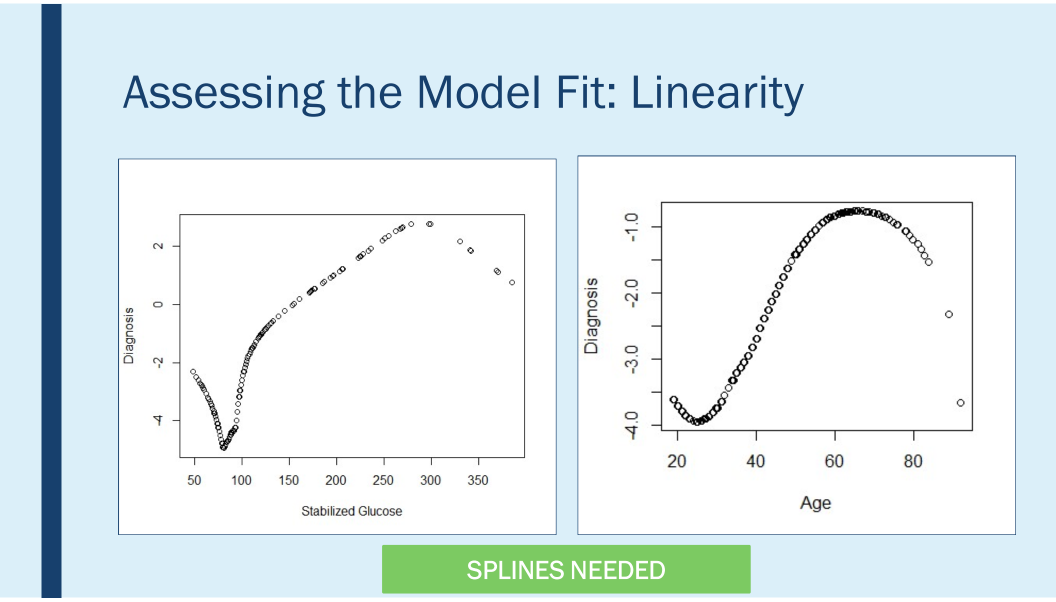
Spline

- As found previously, we have the reduced model: diagnosis predicted on cholesterol, stabilized glucose, age, and waist size. Now we need to assess the model fit.
- The logistic regression model assumes the response y is binary or binomial, which we have in our diagnosis variable with 0 meaning non-diabetic and 1 meaning diabetic.
- The model also assumes there are at least 5-10 of each type of response for every covariate in the model, which we have also upheld.
- Finally, the logistic regression model assumes a linear relation between x and the logit; to do this we create a smoothed scatterplot, or loess plot, for the continuous predictors to ensure they have a linear relation with the logit.
- Using the decision tree on the right, we see that if the loess plot is linear, there is no further action to take and the predictor can be used in the model. However, if the predictor is non-linear, a spline must be utilize to account the non-linearity. We'll explain splines in a few pages.



Assessing the Model Fit: Linearity

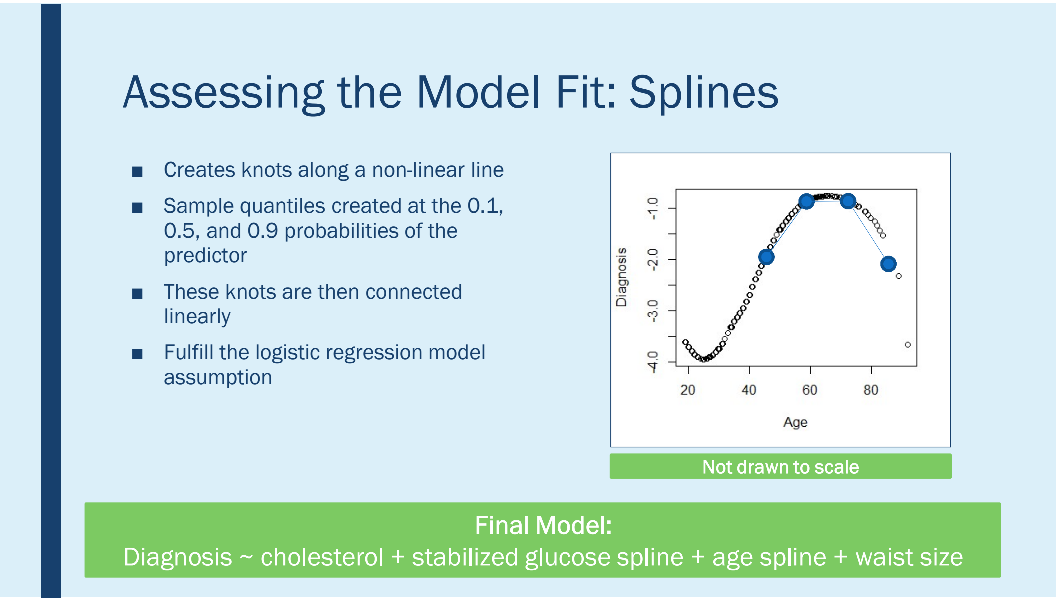
- Here we see the loess plots for total cholesterol on the left and waist size on the right. Though not a perfect line, these predictors will be considered linear.
- Note the axis of the loess plot include the predictor in its original measures, e.g. cholesterol, on the x-axis and the logit of the diagnosis variable on the y-axis.



Assessing the Model Fit: Linearity

SPLINES NEEDED

- The stabilized glucose (left) and age (right) predictors are not linear. Therefore, we must utilize splines to account for the non-linearity.



Assessing the Model Fit: Splines

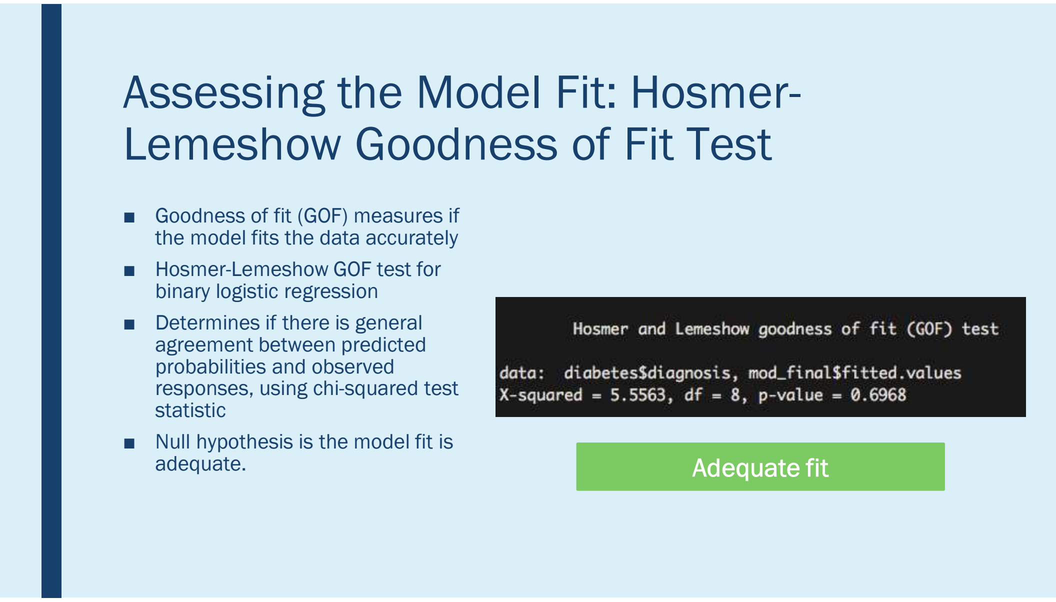
- Creates knots along a non-linear line
- Sample quantiles created at the 0.1, 0.5, and 0.9 probabilities of the predictor
- These knots are then connected linearly
- Fulfill the logistic regression model assumption

Diagnosis
-1.0
-2.0
-3.0
-4.0
20
40
60
80
Age

Not drawn to scale

Final Model:
Diagnosis ~ cholesterol + stabilized glucose spline + age spline + waist size

- Splines are a function that creates knots along a non-linear line. These knots are sample quantiles created at the 0.1, 0.5, and 0.9 probabilities of the predictor. Then they are connected linearly to fulfill the logistic regression model assumption.
- As an example, we see the age loess plot here to the right. I've super imposed a few blue-colored knots to demonstrate the splines, or linear lines, between those knots that would then be used in the regression model. Note these are not drawn to scale.
- By utilizing splines for stabilized glucose and age, the final model predicts the diagnosis binary variable on cholesterol, the stabilized glucose spline, the age spline, and waist size.



Assessing the Model Fit: Hosmer-Lemeshow Goodness of Fit Test

- Goodness of fit (GOF) measures if the model fits the data accurately
- Hosmer-Lemeshow GOF test for binary logistic regression
- Determines if there is general agreement between predicted probabilities and observed responses, using chi-squared test statistic
- Null hypothesis is the model fit is adequate.

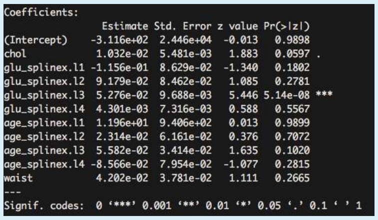
```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  diabetes$diagnosis, mod_final$fitted.values
X-squared = 5.5563, df = 8, p-value = 0.6968
```

Adequate fit

- Now we'll test the model for goodness of fit, measuring if the model fits the data accurately. Because our response variable, diagnosis, is a binary measure, we use a Hosmer-Lemeshow goodness of fit test for binary logistic regression. This determines if there is general agreement between the predicted probabilities of the model and the observed responses in the data using a chi-squared test statistic. The null hypothesis of this test is that model fit is accurate.
- When the test was run, see the image to the right, the p-value of the test was very high, thus the null hypothesis is not rejected and the model fit is adequate.

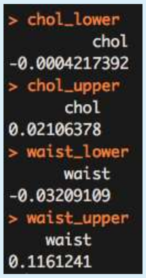
Model Inferences: Analyzing Significant Predictors & Influential Observations

P-values



```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -3.116e+02  2.446e+04  -0.013   0.9898
chol            1.032e-02  5.481e-03   1.883   0.0597 .
glu_splinex.l1 -1.156e-01  8.629e-02  -1.340   0.1802
glu_splinex.l2  9.179e-02  8.462e-02   1.085   0.2781
glu_splinex.l3  5.276e-02  9.688e-03   5.446 5.14e-08 ***
glu_splinex.l4  4.301e-03  7.316e-03   0.588   0.5567
age_splinex.l1  1.196e+01  9.406e+02   0.013   0.9899
age_splinex.l2  2.314e-02  6.161e-02   0.376   0.7072
age_splinex.l3  5.582e-02  3.414e-02   1.635   0.1020
age_splinex.l4 -8.566e-02  7.954e-02  -1.077   0.2815
waist           4.202e-02  3.781e-02   1.111   0.2665
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

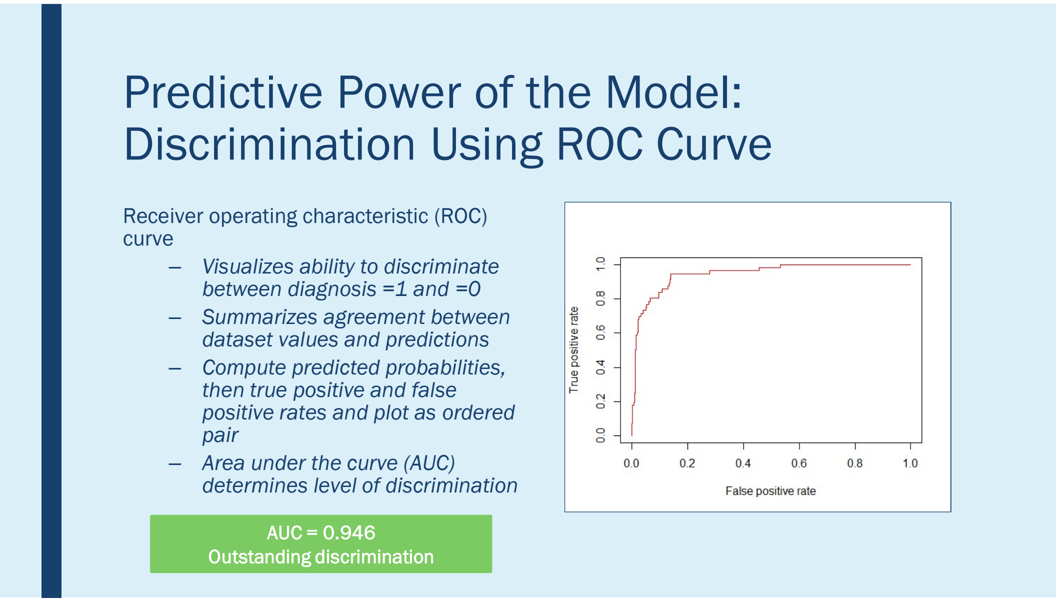
Confidence Intervals



```
> chol_lower
         chol
-0.0004217392
> chol_upper
      chol
0.02106378
> waist_lower
      waist
-0.03209109
> waist_upper
    waist
0.1161241
```

No influential observations found using the Cook's distance comparison to the F distribution threshold.

- By analyzing the p-values and confidence intervals of the significant predictors, we don't see anything out of the ordinary to be concerned about.
  - Looking solely at the p-values of the non-spline predictors, we see cholesterol is significant. Waist does have a high p-value and might be considered insignificant; however, with subject matter knowledge, I have chosen to keep it in the model.
  - The confidence intervals demonstrate that we are 95% confident that the true value of the beta_cholesterol and beta_waist lie in those intervals.
- Then we checked for influential observations, or influential points that have a disproportionately large effect on the results of a regression analysis. To check for such observations, we calculate the Cooks' Distance and compare which of those distances has exceeded the 50th percentile of the F distribution. We found there were no such observations to be removed from the model.



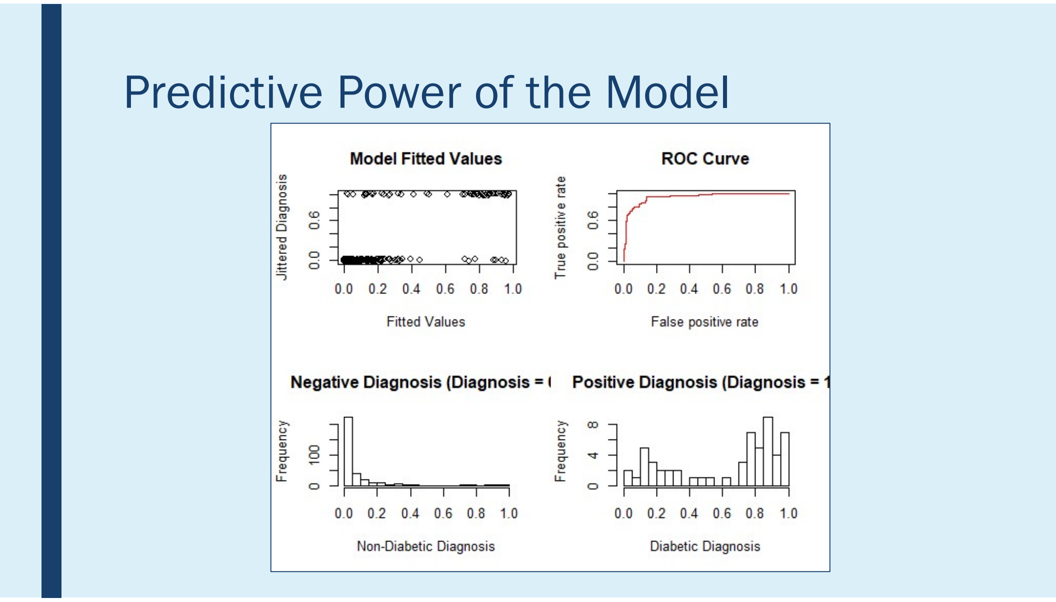
Predictive Power of the Model: Discrimination Using ROC Curve

Receiver operating characteristic (ROC) curve

- *Visualizes ability to discriminate between diagnosis =1 and =0*
- *Summarizes agreement between dataset values and predictions*
- *Compute predicted probabilities, then true positive and false positive rates and plot as ordered pair*
- *Area under the curve (AUC) determines level of discrimination*

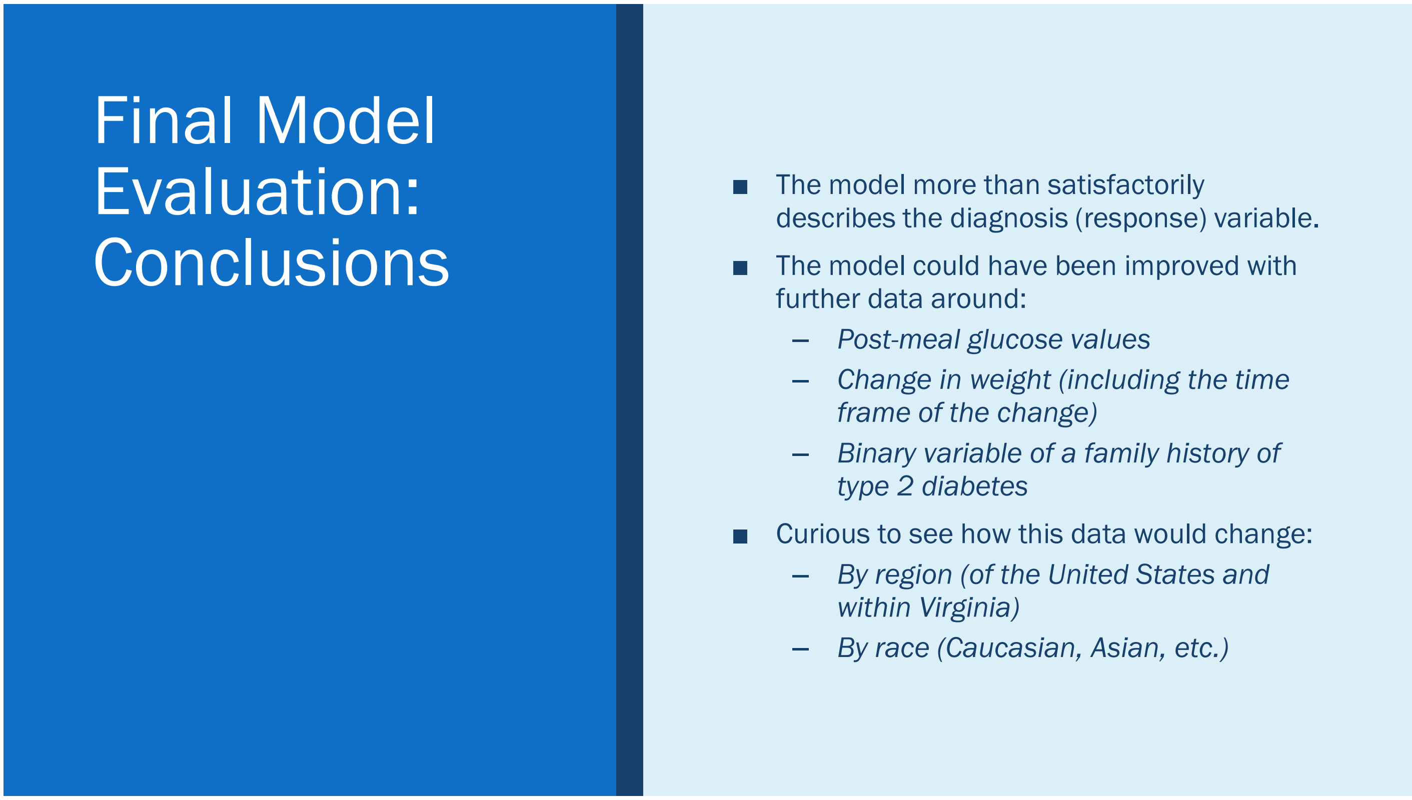
AUC = 0.946
Outstanding discrimination

- When measuring the predictive power of the model, we wish to measure the model's ability to distinguish, or discriminate, between the response variable (diagnosis) observations equal to 1 and 0. The receiver operating characteristic (ROC) curve visualizes and quantifies this ability and summaries the agreement between the dataset values and the predicted values of diagnoses. Thus we need to fit the model and obtain the predicted probabilities, then compute the true positive and false positive rate and plot these.
- The plot on the right is the ROC curve, with the true positive rate on the y-axis and the false positive rate on the x-axis. Discrimination is quantified as the area under the curve (AUC), with a higher AUC showing outstanding model discrimination between the two diagnosis outcomes (0 or 1). The AUC of this curve is 0.946, which means the model has outstanding discrimination, near perfect even. In other words, the four predictors in the model are highly effective at sorting the diabetic from the non-diabetic diagnoses.



Predictive Power of the Model

- Here we see a few other graphs; top right is the model’s fitted values, of the predicted diagnosis of either 0 or 1 for the individuals. Then the ROC curve is to the right and was previously discussed.
- Then the bottom row of graphs show the distribution of negative and positive diabetes diagnoses. The negative diagnosis sees the same distribution of fitted values as the first graph, with a cluster at 0 and then a left-tail distribution. The positive diagnosis is also reflective of the model fitted values graph showing a more even distribution and then a spike at 1.
- Overall, this demonstrates strong predictive power of the model.



Final Model Evaluation: Conclusions

- The model more than satisfactorily describes the diagnosis (response) variable.
- The model could have been improved with further data around:
  - *Post-meal glucose values*
  - *Change in weight (including the time frame of the change)*
  - *Binary variable of a family history of type 2 diabetes*
- Curious to see how this data would change:
  - *By region (of the United States and within Virginia)*
  - *By race (Caucasian, Asian, etc.)*

- Now for a few final conclusions to sum up this modeling exercise.
- The model more than satisfactorily describes the diagnosis response variable. It has passed the tests of model fit and predictive power with flying colors.
- The model could have been improved with further data around post-meal glucose values (in comparison to the current stabilized glucose values); an individual’s change in weight and over what period of time that change occurred as type 2 diabetes is directly related to lifestyle habits; and a binary variable indicating if the individual had a family history of type 2 diabetes as it is genetic.
- I’d be curious to see how this model would change by region of the United States and other counties in Virginia and by race, as only African Americans in two Virginia counties were included in the study.

THANK YOU!

Gina O'Riordan - East