STA3064: Project C

Team 10: Kyle Lam, Elijah McLaughlin, Damian Jean Baptiste, Alexander Khan

1. Motivation

- Kyle and Elijah both have a history of diabetes in their families.
- Fitting a logistic regression model for this data set will help predict diabetes for females, or potentially both males and females, from a number of predictor variables.

2. Data Description

- This dataset was sourced from Kaggle and was originally produced by the *National Institute of Diabetes and Digestive and Kidney*
- Data reflects 250 female patients and their respective medical measurements.
- Consists of 8 quantitative predictor variables and 1 binary response variable
- 250 observations

Link to data set:
https://www.kaggle.com/datasets/vikasukani/diabetes-data-set/data

2. Data Description (Cont.)

Quantitative Predictor Variables:

1. **Pregnancies:** Number of pregnancies
2. **Glucose:** Glucose level in blood
3. **BloodPressure:** Blood pressure measurement
4. **SkinThickness:** Thickness of the skin
5. **Insulin:** Insulin level in blood
6. **BMI:** Body mass index
7. **DiabetesPedigreeFunction:** Diabetes likelihood percentage depending age and patient's diabetic family history.
8. **Age:** Age of patient

Binary Response Variable:

9. Outcome: Diabetes (1 = Event that patient has diabetes, 0 = Event patient does not have diabetes)

3. Data Exploration

PROC IMPORT code after dataset file was uploaded into SAS:

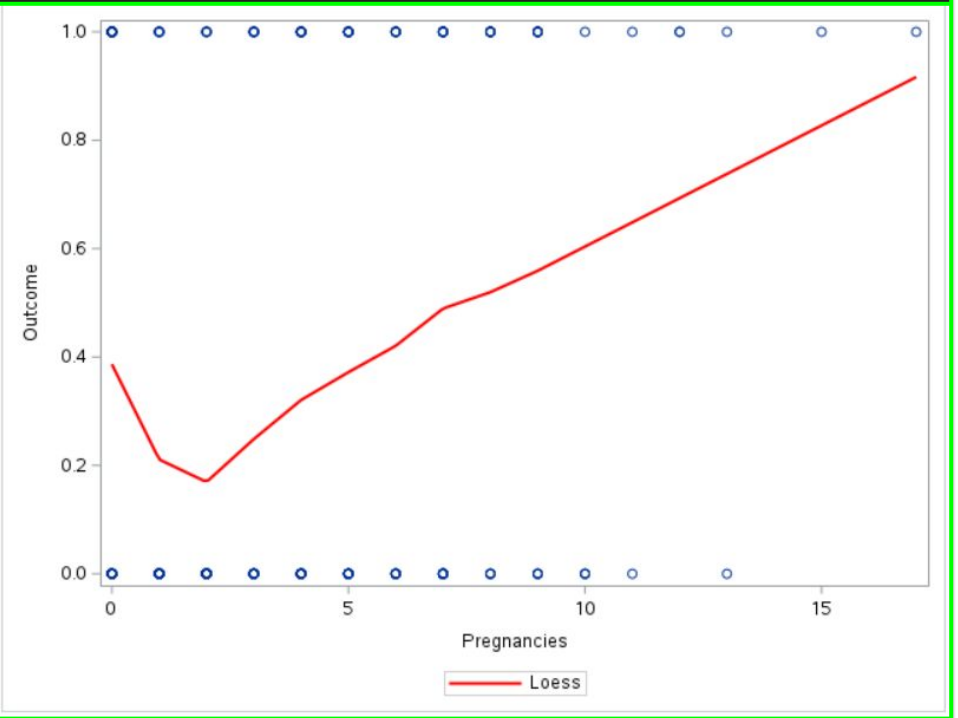
```
1 FILENAME REFFILE '/home/u62117076/sasuser.v94/STA3064/Project C/Diabetes.csv';
2 PROC IMPORT DATAFILE=REFFILE
3     DBMS=CSV
4     OUT=Diabetes;
5     GETNAMES=YES;
6 RUN;
```

*Preserved variable names from first row of CSV file

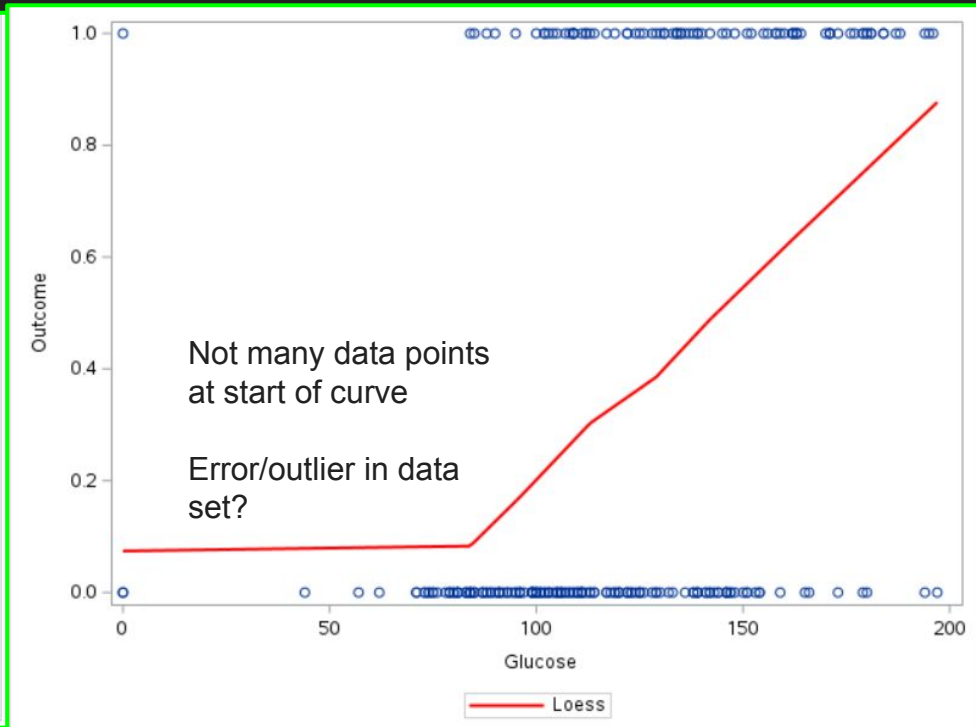
3. Data Exploration: Scatterplots with LOESS smooth on each curve

Green indicates an appearance of a general positive trend

Pregnancies vs Outcome

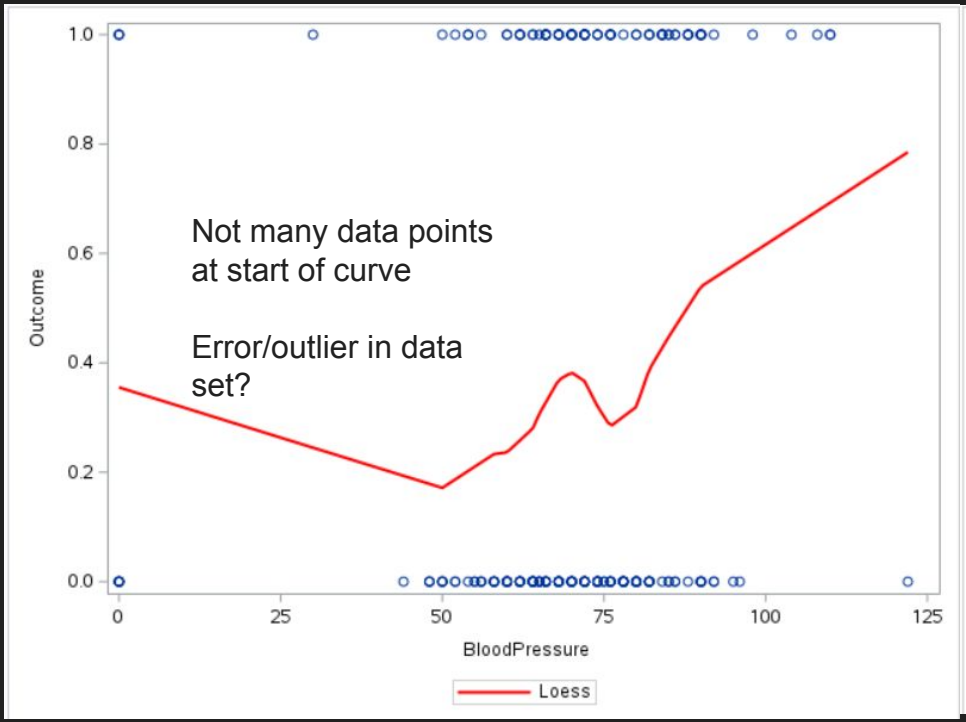


Glucose vs Outcome

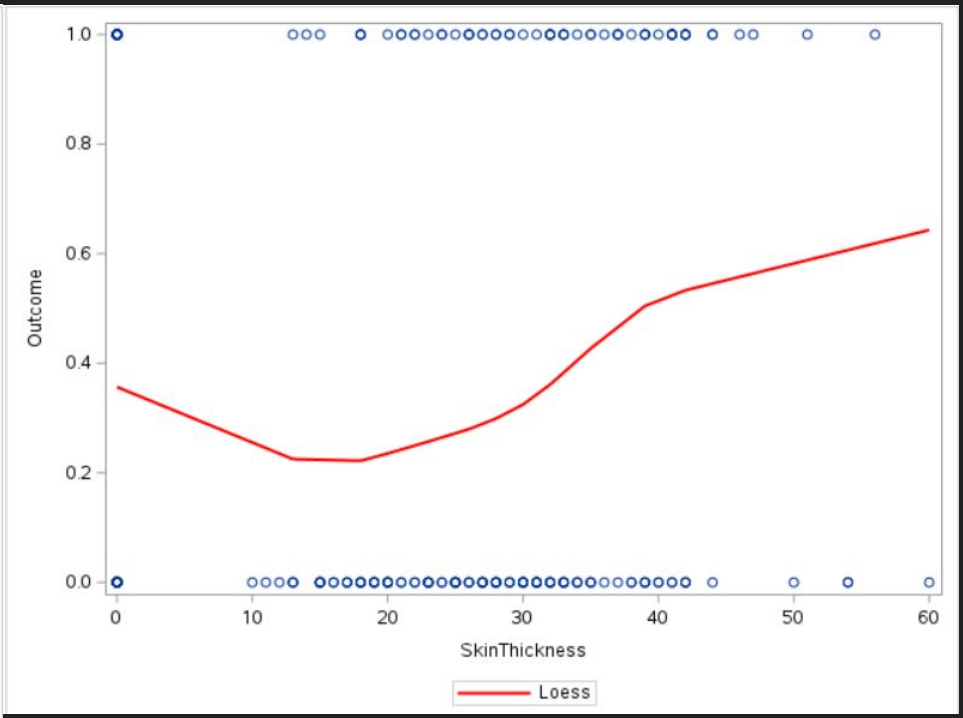


3. Data Exploration: Scatterplots with LOESS smooth on each curve

Blood Pressure vs Outcome



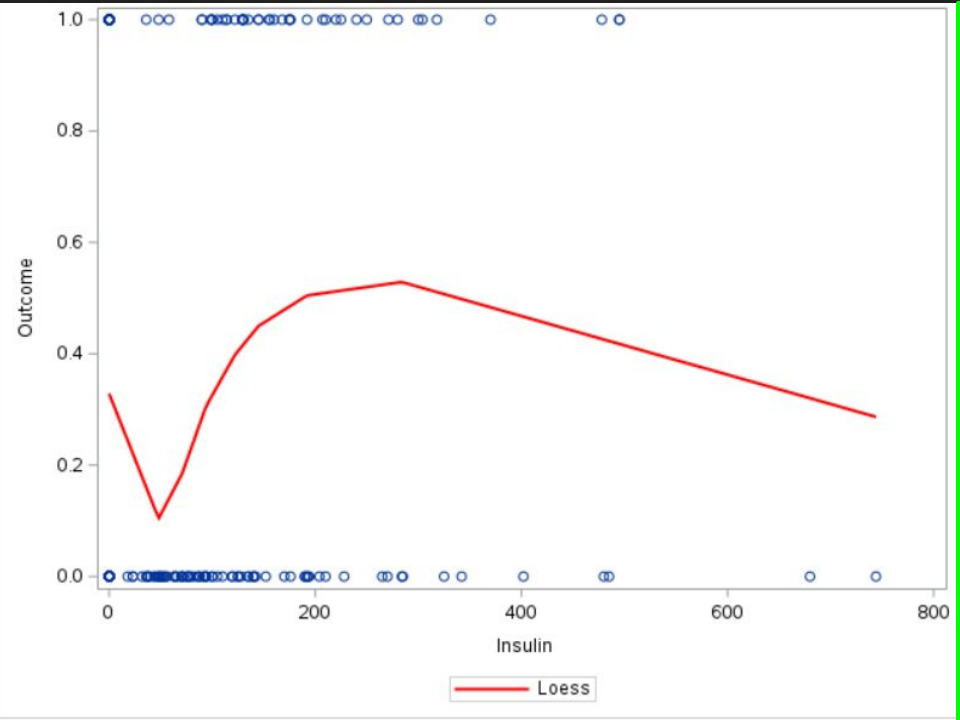
Skin Thickness vs Outcome



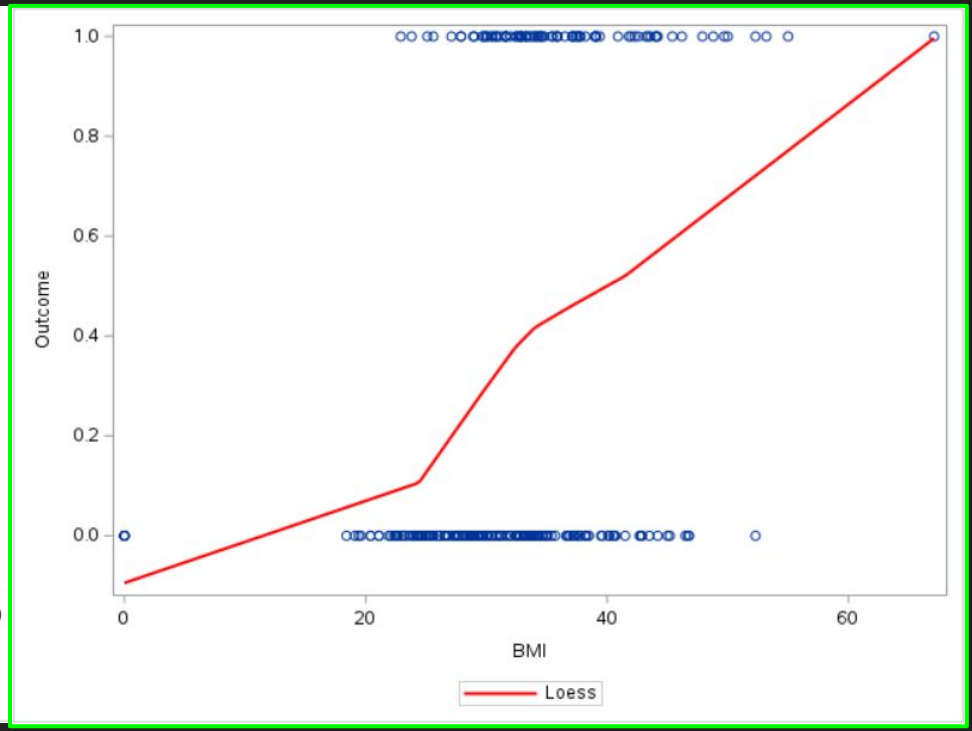
3. Data Exploration: Scatterplots with LOESS smooth on each curve

Green indicates an appearance of a general positive trend

Insulin vs Outcome



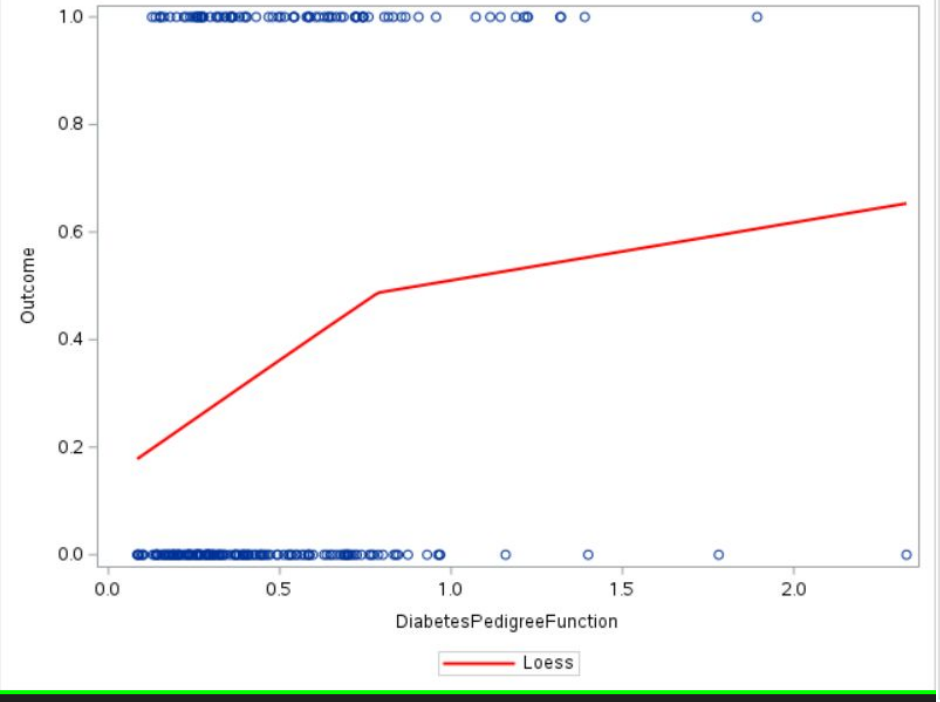
BMI vs Outcome



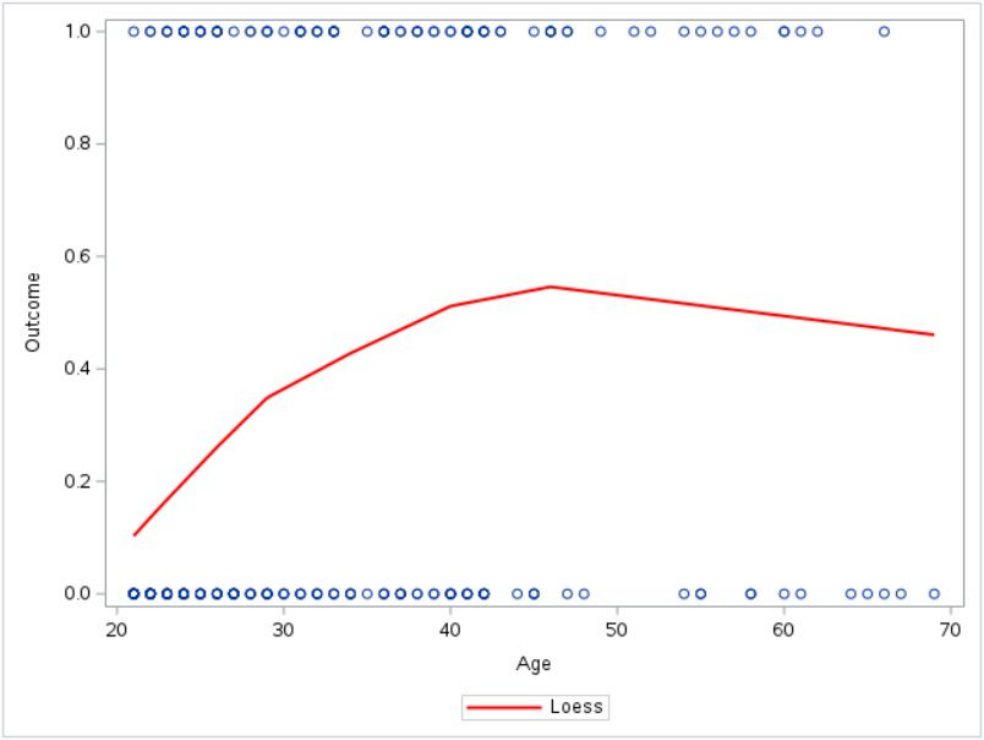
3. Data Exploration: Scatterplots with LOESS smooth on each curve

Green indicates an appearance of a general positive trend

Diabetes Pedigree Function vs Outcome



Age vs Outcome



4. Model Fitting and Analysis

One Predictor Model

4a. Fitting logistic regression model using one predictor

Predictor Variable Selected: BMI (Looked most promising based on LOESS)

Intercept Estimate = -4.4526 (Log-odds when BMI = 0)

BMI Estimate - 0.1146 (Change in log-odds of Event=1 per 1 unit change in BMI)

Parameter Estimates Table:

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.4526	0.7475	35.4797	<.0001
BMI	1	0.1146	0.0217	27.8477	<.0001

← Significant

← Significant

4b. Statistical Significance of one-predictor-model

Intercept Estimate P-Value = <0.0001. Reject H0. Intercept is significantly different from zero

BMI Estimate P-Value = <0.0001. Reject H0. Significant association between BMI and the log-odds of the event.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.4526	0.7475	35.4797	<.0001
BMI	1	0.1146	0.0217	27.8477	<.0001

4b. Statistical Significance of one-predictor-model (cont.)

Likelihood Ratio Test P-value = <0.0001

H0: Model only including the intercept is good enough for this data set. REJECTED

Ha: Inclusion of BMI as predictor significantly improves the fit of the logistic regression model

Significant relationship between BMI and Diabetes

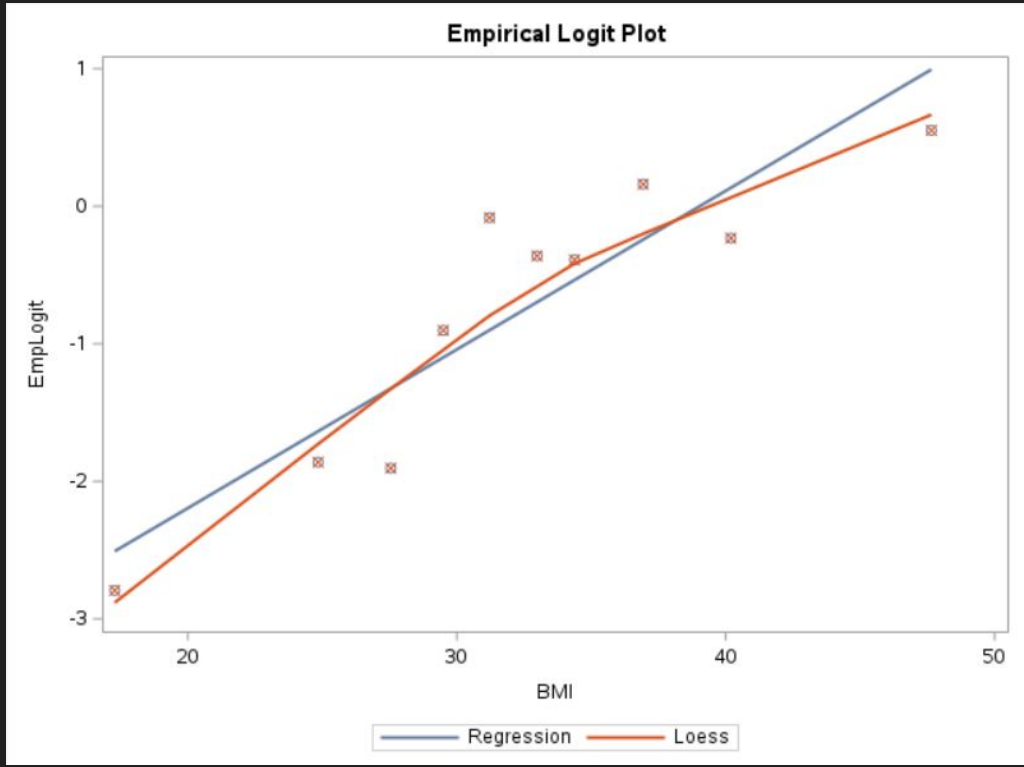
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	36.9688	1	<.0001
Score	31.8826	1	<.0001
Wald	27.8477	1	<.0001

4c. Empirical logit plot of one-predictor-model

- 10 groupings/bins created
  - On average, the ten points are increasing
- Line of best fit appears moderate and LOESS matches up fairly

Conclusion:

- Assumption of linearity for the log-odds response in the logistic regression equation is met.



All Predictor Model

4d. Impact of Multicollinearity

Since the Variance Inflation (VIF) of all the predictors each lie between [1.0207, 1.77866], they are considered low (< 5)

We can conclude that multicollinearity is not a major concern.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-0.78457	0.14044	-5.59	<.0001	0
Pregnancies	1	0.02492	0.00963	2.59	0.0103	1.49747
Glucose	1	0.00448	0.00089577	5.00	<.0001	1.44095
BloodPressure	1	-0.00160	0.00158	-1.01	0.3137	1.25046
SkinThickness	1	0.00020947	0.00201	0.10	0.9171	1.49291
Insulin	1	-0.00046174	0.00026414	-1.75	0.0817	1.52687
BMI	1	0.01503	0.00346	4.34	<.0001	1.36657
DiabetesPedigreeFunction	1	0.22290	0.08229	2.71	0.0072	1.10207
Age	1	0.00146	0.00296	0.49	0.6211	1.77866

4e. Fitting Best Multiple Logistic Regression Model

Using a forward selection method, we get that the best multiple logistic model includes Glucose, BMI, Pregnancies, DiabetesPedigreeFunction, and Insulin as predictors.

The parameters estimates table confirms the statistically significance of these variables.

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	Glucose	1	1	44.9129	<.0001
2	BMI	1	2	17.6304	<.0001
3	Pregnancies	1	3	10.2710	0.0014
4	DiabetesPedigreeFunc	1	4	6.2187	0.0126
5	Insulin	1	5	4.1454	0.0417

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-8.8912	1.2069	54.2761	<.0001
Pregnancies	1	0.1748	0.0531	10.8399	0.0010
Glucose	1	0.0272	0.00581	21.9114	<.0001
Insulin	1	-0.00281	0.00141	3.9984	0.0455
BMI	1	0.1107	0.0255	18.9105	<.0001
DiabetesPedigreeFunc	1	1.4268	0.5380	7.0327	0.0080

4f. Assessing best model

Both models are statistically significant, and reject the null hypothesis. However, when comparing Association of Predicted Probabilities and Observed Responses, we see that Model 1 explains more variability than Model 2.

Model 1 includes all significant predictors.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	84.2	Somers' D	0.683
Percent Discordant	15.8	Gamma	0.683
Percent Tied	0.0	Tau-a	0.310
Pairs	14104	c	0.842

Model 2 includes only *BMI* and *Glucose.*

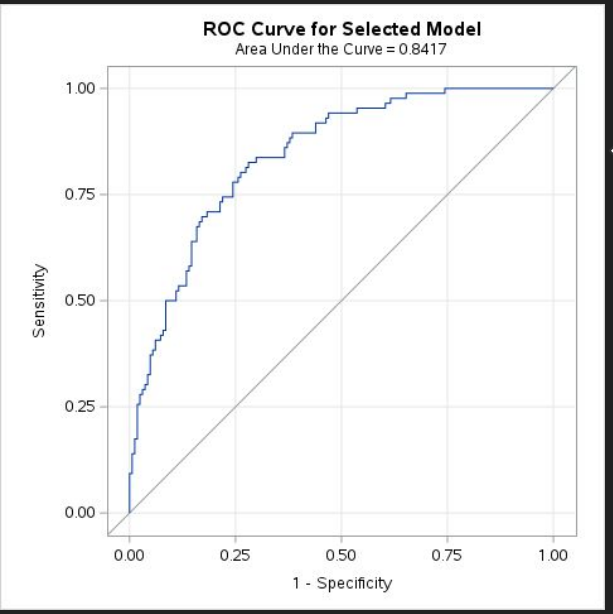
Association of Predicted Probabilities and Observed Responses			
Percent Concordant	80.4	Somers' D	0.608
Percent Discordant	19.6	Gamma	0.608
Percent Tied	0.0	Tau-a	0.276
Pairs	14104	c	0.804

4g. Prediction of Probability of Success for Set Values In Best Model

With an 80-20 split, this graphic shows the probability of success for set values for Model 1.

Obs	Selected	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome	F_Outcome	I_Outcome	P_0	P_1
1	0	2	138	62	35	0	33.6	0.127	47	1	1	0	0.71119	0.28881
2	0	0	135	68	42	250	42.3	0.365	24	1	1	0	0.68787	0.31213
3	0	4	99	72	17	0	25.6	0.294	28	0	0	0	0.87581	0.12419
4	0	2	117	90	19	71	25.2	0.313	21	0	0	0	0.89816	0.10184
5	0	10	75	82	0	0	33.3	0.263	38	0	0	0	0.61803	0.38197
6	0	8	120	78	0	0	25	0.409	64	0	0	0	0.65382	0.34618
7	0	9	91	68	0	0	24.2	0.2	58	0	0	0	0.79962	0.20038
8	0	3	99	54	19	86	25.6	0.154	24	0	0	0	0.92725	0.07275
9	0	0	180	66	39	0	42	1.893	25	1	1	1	0.06660	0.93340
10	0	2	71	70	27	0	28	0.586	22	0	0	0	0.91095	0.08905
11	0	1	103	80	11	82	19.4	0.491	22	0	0	0	0.95645	0.04355
12	0	1	73	50	10	0	23	0.248	21	0	0	0	0.96855	0.03145
13	0	2	109	92	0	0	42.7	0.845	54	0	0	1	0.41633	0.58367
14	0	4	129	86	20	270	35.1	0.231	23	0	0	0	0.74851	0.25149
15	0	0	101	65	28	0	24.6	0.237	22	0	0	0	0.94722	0.05278
16	0	2	110	74	29	125	32.4	0.698	27	0	0	0	0.77327	0.22673
17	0	1	80	55	0	0	19.1	0.258	21	0	0	0	0.97540	0.02460
18	0	2	142	82	18	64	24.7	0.761	21	0	0	0	0.75711	0.24289
19	0	1	151	60	0	0	26.1	0.179	22	0	0	0	0.82579	0.17421
20	0	1	81	72	18	40	26.6	0.283	24	0	0	0	0.94977	0.05023
21	0	2	85	65	0	0	39.6	0.93	27	0	0	0	0.59521	0.40479
22	0	3	171	72	33	135	33.3	0.199	24	1	1	0	0.57406	0.42594
23	0	4	76	62	0	0	34	0.391	25	0	0	0	0.80671	0.19329
24	0	7	160	54	32	175	30.5	0.588	39	1	1	1	0.41227	0.58773
25	0	4	97	60	23	0	28.2	0.443	22	0	0	0	0.82307	0.17693
26	0	4	99	76	15	51	23.2	0.223	21	0	0	0	0.91863	0.08137
27	0	3	120	70	30	135	42.9	0.452	30	0	0	0	0.50998	0.49002
28	0	8	84	74	31	0	38.3	0.457	39	0	0	1	0.48170	0.51830
29	0	0	93	60	25	92	28.7	0.532	22	0	0	0	0.92364	0.07636
30	0	5	105	72	29	325	36.9	0.159	28	0	0	0	0.80711	0.19289
31	0	1	136	74	50	204	37.4	0.399	24	0	0	0	0.71640	0.28360
32	0	0	114	80	34	285	44.2	0.167	27	0	0	0	0.79609	0.20391
33	0	3	148	66	25	0	32.5	0.256	22	0	0	0	0.61042	0.38958
34	0	3	111	90	12	78	28.4	0.495	29	0	0	0	0.82611	0.17389
35	0	2	75	64	24	55	29.7	0.37	33	0	0	0	0.92194	0.07806
36	0	6	85	78	0	0	31.2	0.382	42	0	0	0	0.75895	0.24105
37	0	0	129	110	46	130	67.1	0.319	26	1	1	1	0.12478	0.87522
38	0	0	119	64	18	92	34.9	0.725	23	0	0	0	0.74230	0.25770
39	0	1	128	98	41	58	32	1.321	33	1	1	0	0.53838	0.46162
40	0	9	123	70	44	94	33.1	0.374	40	0	0	1	0.45065	0.54935
41	0	5	158	84	41	210	39.4	0.395	29	1	1	1	0.37198	0.62802
42	0	4	109	64	44	99	34.8	0.905	26	1	1	0	0.56786	0.43214
43	0	5	111	72	28	0	23.9	0.407	27	0	0	0	0.82382	0.17618
44	0	8	196	76	29	280	37.5	0.605	57	1	1	1	0.14207	0.85793
45	0	5	109	62	41	129	35.8	0.514	25	1	1	0	0.62865	0.37135
46	0	2	158	90	0	0	31.6	0.805	66	1	1	1	0.46452	0.53548
47	0	3	74	68	28	45	29.7	0.293	23	0	0	0	0.91368	0.08632
48	0	7	181	84	21	192	35.9	0.586	51	1	1	1	0.21134	0.78866
49	0	10	122	68	0	0	31.2	0.258	41	0	0	1	0.43210	0.56790
50	0	9	124	70	33	402	35.4	0.282	34	0	0	0	0.60750	0.39250

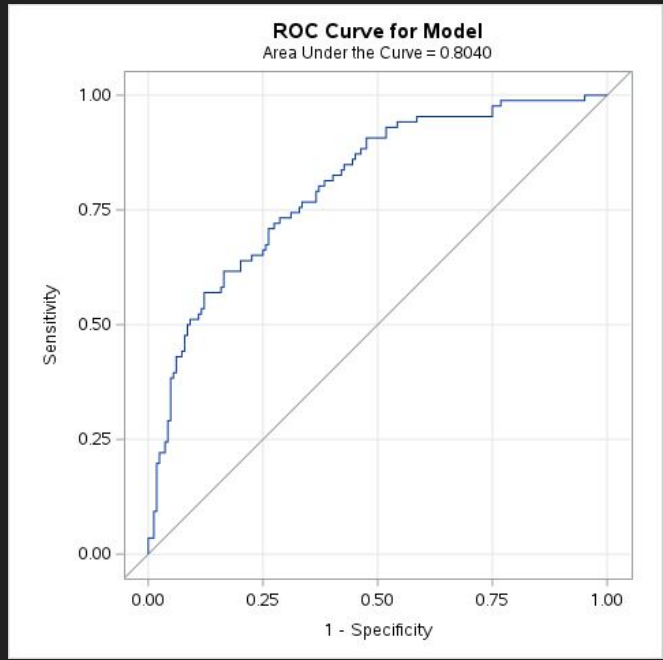
4h. ROC Curve for Best Model and Additional Model



Model 1, or the Best Model, ROC Curve shows that c = 0.8417, or explains 84.17% of the variability

Roughly 4% difference

Model 2, ROC Curve shows that c = 0.8040, or only explains 80.40% of the variability



5. Conclusion

We came into this project looking to find ways to predict and prevent diabetes in women and with our results we have found that the significant variables used to predict diabetes are glucose levels, BMI, number of pregnancies, diabetic family history, and insulin levels. With this knowledge, we could inform people on how to better protect themselves and prevent the development of diabetes. Interventions such as classes on healthy eating and exercising could further advance the efficiency of this information.

Certification

We, the project team members, certify that the percentage of the effort listed by each of our names below is an accurate account of the original effort contributed by each team member in the producing of this project and report.

Name (Printed)	Percent of Total Effort
Kyle Lam	25 %
Alexander Khan	25 %
Elijah McLaughlin	25 %
Damian Jean Baptiste	25 %