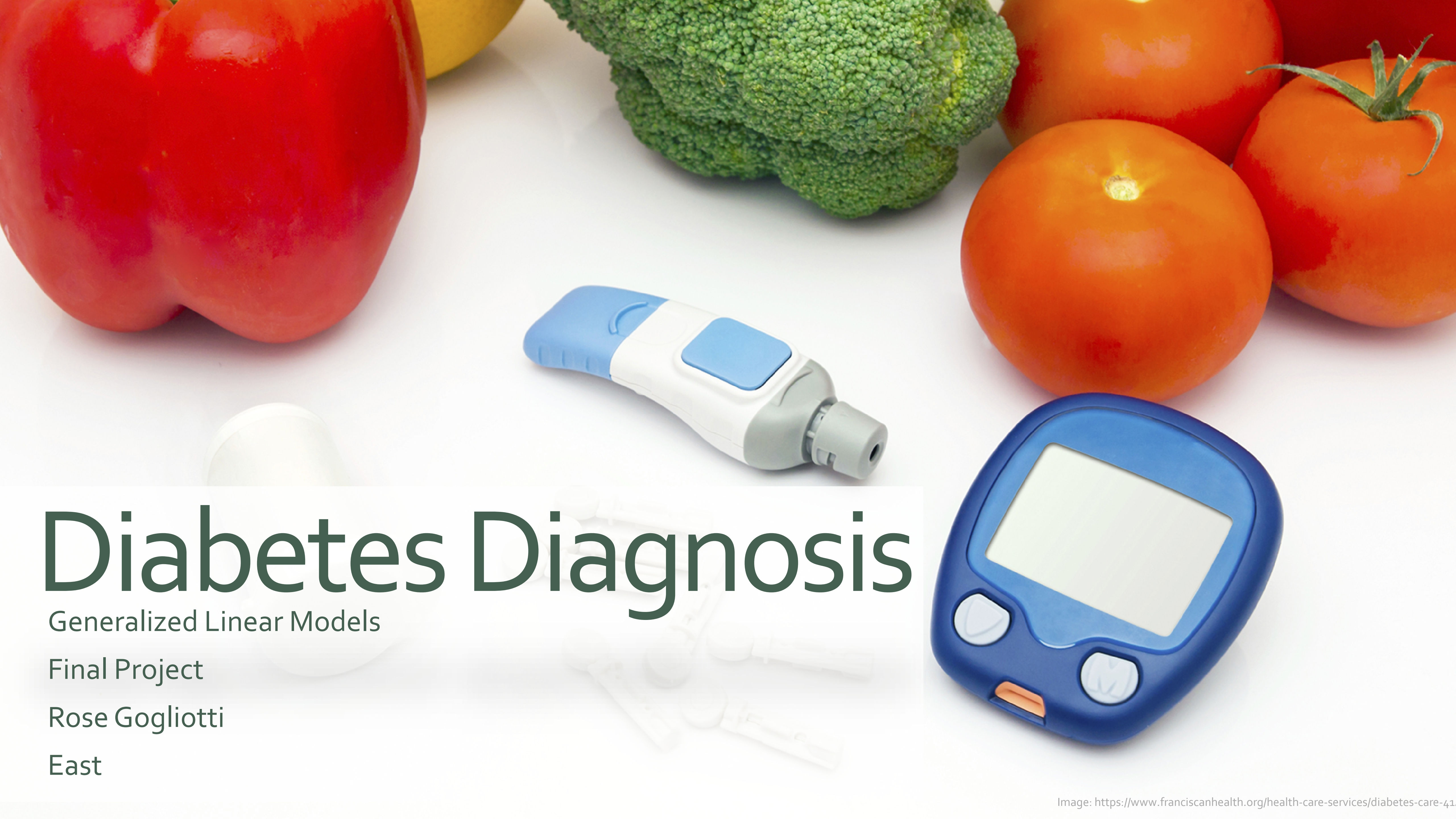


Diabetes Diagnosis

Generalized Linear Models

Final Project

Rose Gogliotti

East

Image: https://www.franciscanhealth.org/health-care-services/diabetes-care-41

Goal: Build a logistic regression model to explain what factors are related to a positive diagnosis.

Dataset: Diabetes from the Faraway package

19 variables
403 Observations

Create Diagnosis variable from glyhb

Negative (no diabetes)
Diagnosis = 0 if glyhb < 7

Positive (have diabetes)
Diagnosis = 1 if glyhb > 7

Variables	Description
Id	Subject id
Chol	Total cholesterol
Stab.glu	Stabilized glucose
hdl	High density lipoprotein
Ratio	Cholesterol/hdl
glyhb	Glycosolated Hemoglobin
Location	County - a factor with levels Buckingham Louisiana
Age	age in years
gender	a factor with levels male female
Height	height in inches
weight	weight in pounds
frame	a factor with levels small medium large
Bp.1s	First Systolic Blood Pressure
Bp.1d	First Diastolic Blood Pressure
Bp.2s	Second Systolic Blood Pressure
Bp.2d	Second Diastolic Blood Pressure
waist	Waist in inches
hip	Hip in inches
time.ppn	Postprandial Time (in min) when Labs were Drawn
diagnosis	Factor indicating diagnosis

Some variables removed due to high # NA and non-predictive abilities

Variables	Description
~~id~~	~~Subject id~~
Chol	Total cholesterol
Stab.glu	Stabilized glucose
hdl	High density lipoprotein
ratio	Cholesterol/hdl
~~glyhb~~	~~Glycosolated Hemoglobin~~
~~Location~~	~~County - a factor with levels Buckingham Louisiana~~
age	age in years
gender	a factor with levels male female
height	height in inches
weight	weight in pounds
frame	a factor with levels small medium large
Bp.1s	First Systolic Blood Pressure
Bp.1d	First Diastolic Blood Pressure
~~Bp.2s~~	~~Second Systolic Blood Pressure~~
~~Bp.2d~~	~~Second Diastolic Blood Pressure~~
waist	Waist in inches
hip	Hip in inches
~~time.ppn~~	~~Postprandial Time (in min) when Labs were Drawn~~
diagnosis	Factor indicating diagnosis (calculated value)

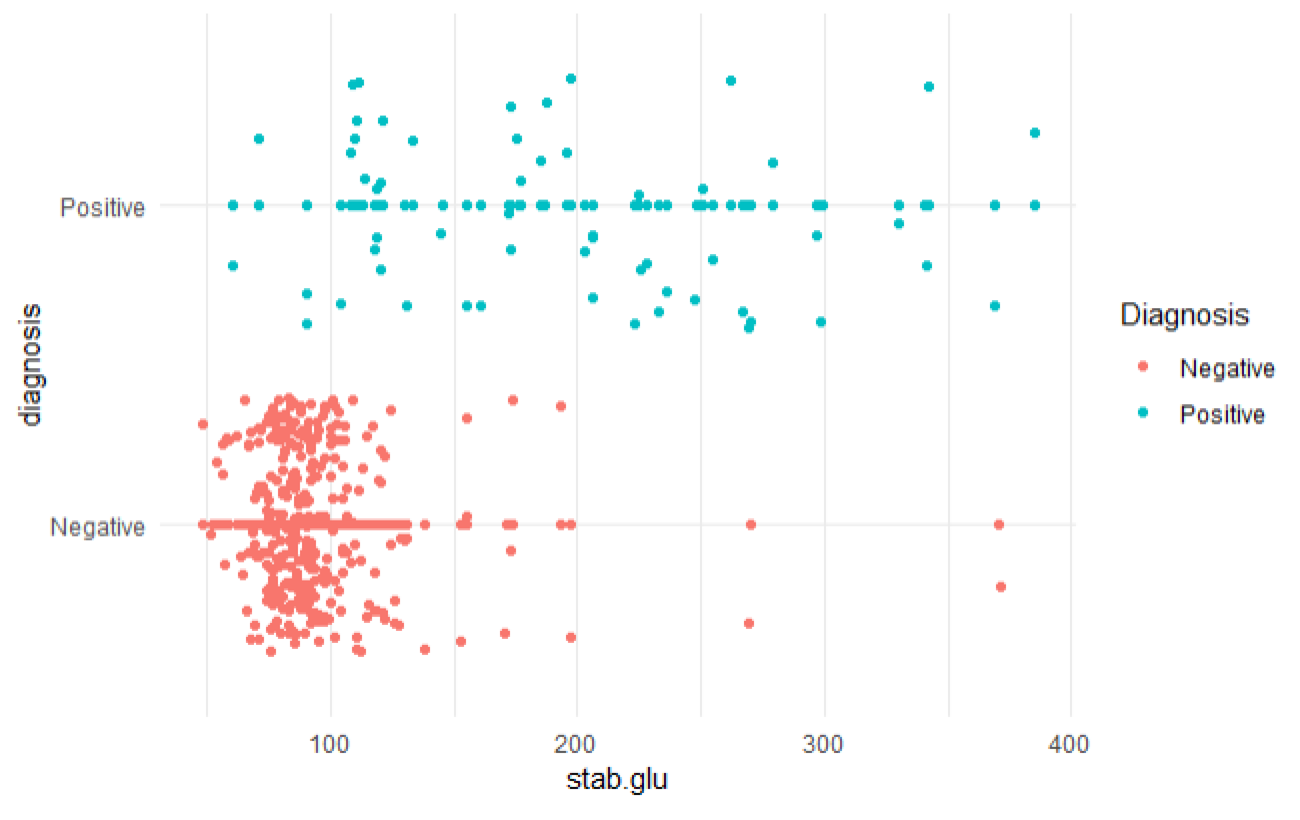
Remaining :

- 367 Observations (after dropping NA values)
- diagnosis response variable
- 13 predictor variables

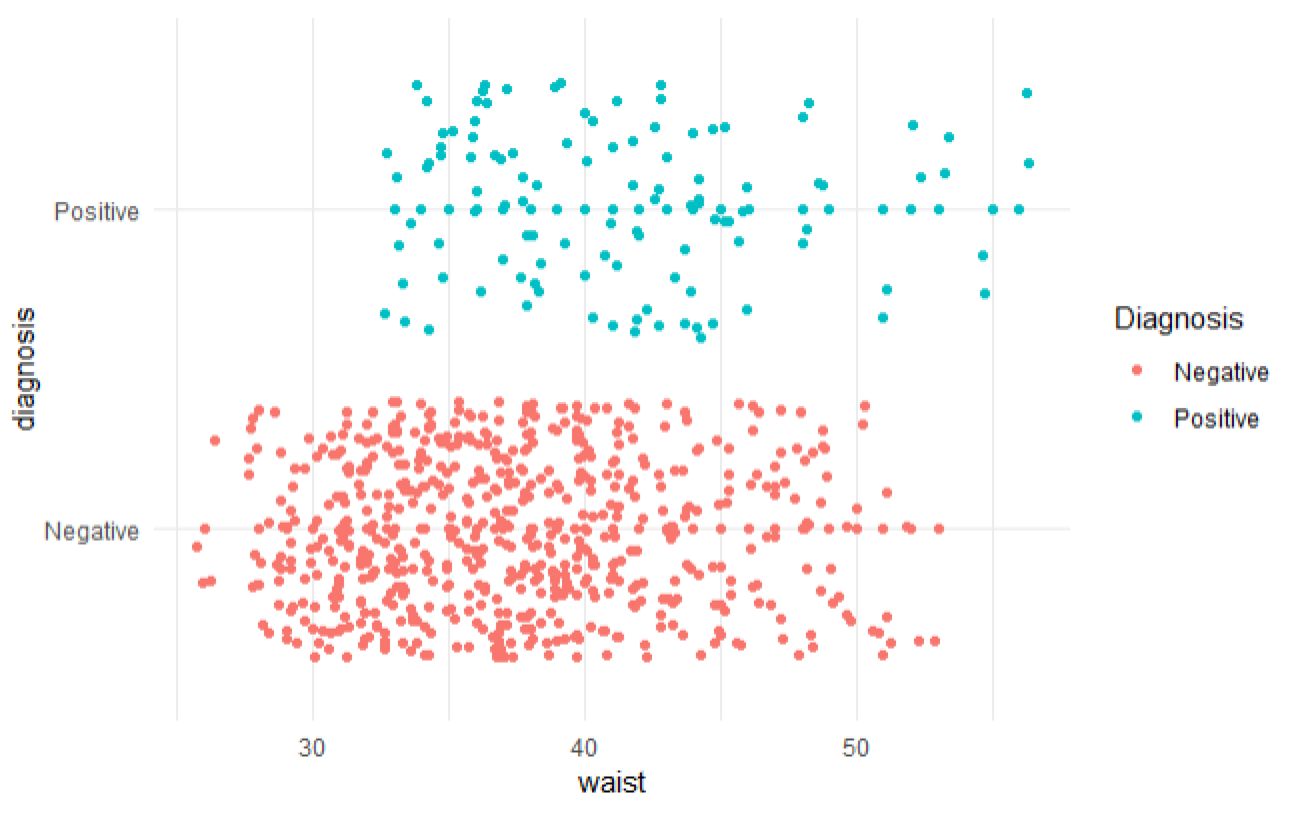
Diagnosis

Negative (0)	**Positive (1)**
311	56

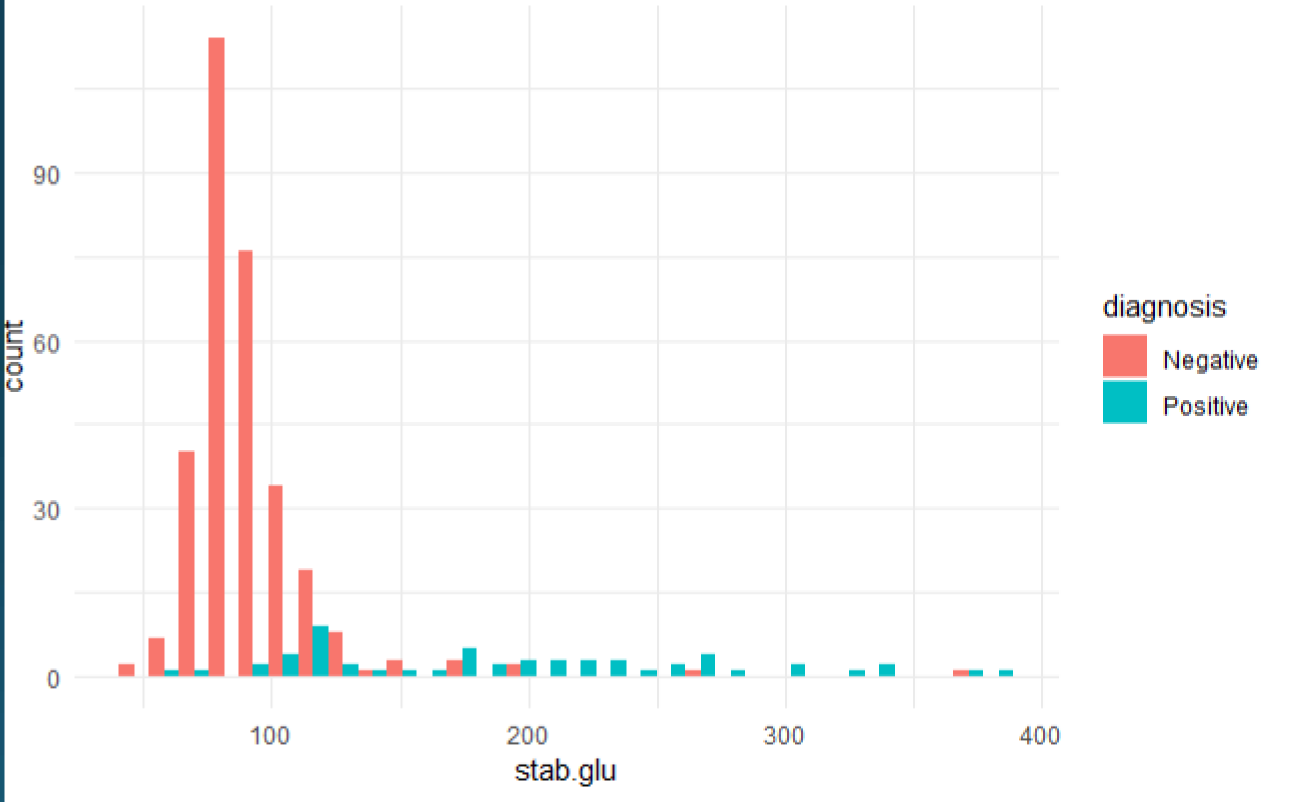
Step 1: Stabilized Glucose vs.Diagnosis



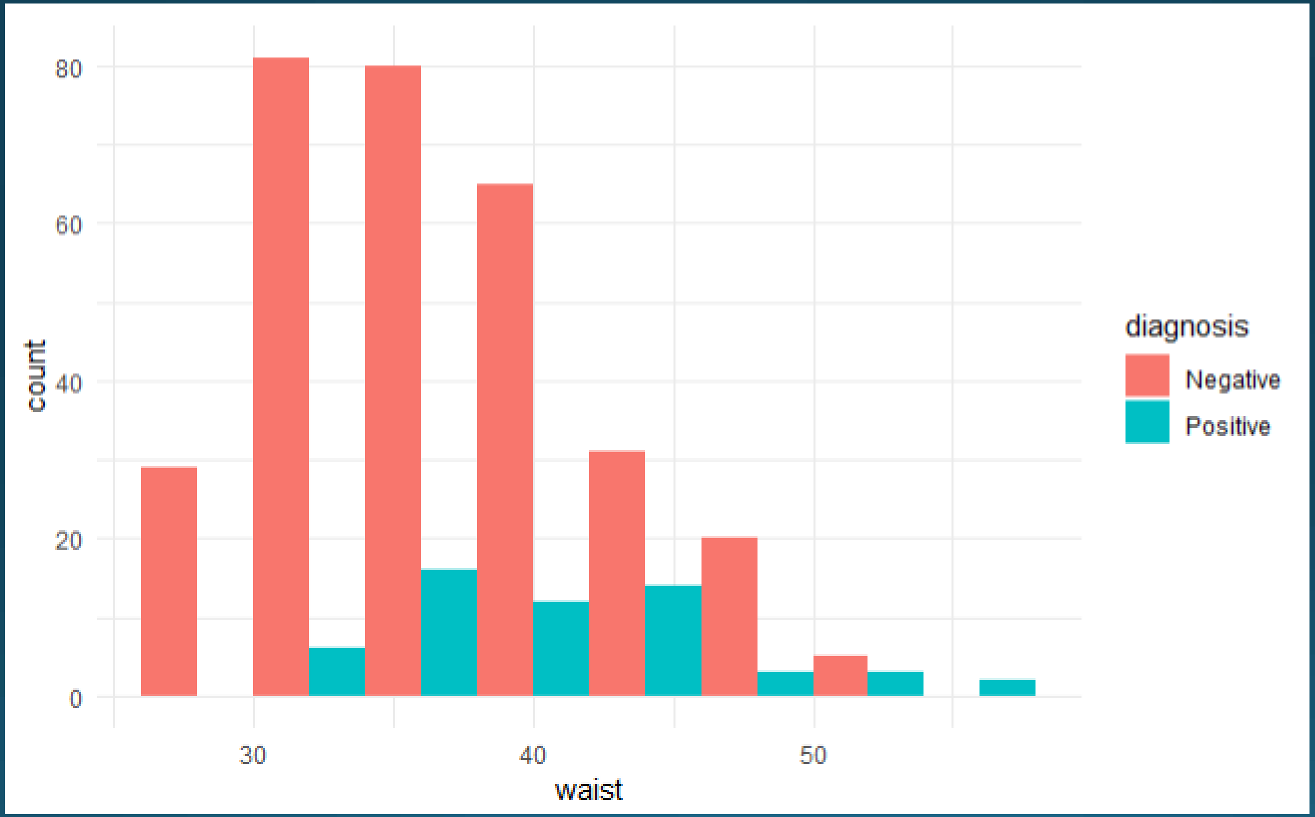
Step 1: Waist vs Diagnosis



Step 1: Stabalized Glucose by Diagnosis

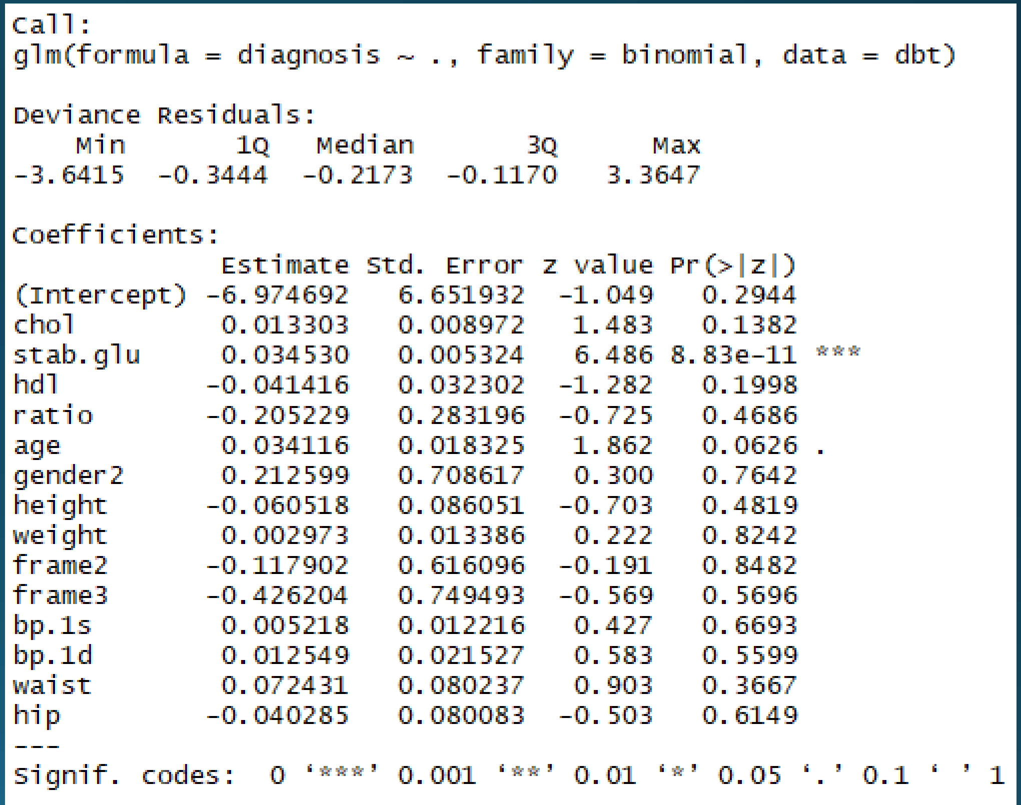


Step 1: Waist by Diagnosis



Step 2: Fit a model

- Initially all remaining variables modeled
- Violates “Rule of 5” so variable reduction needed

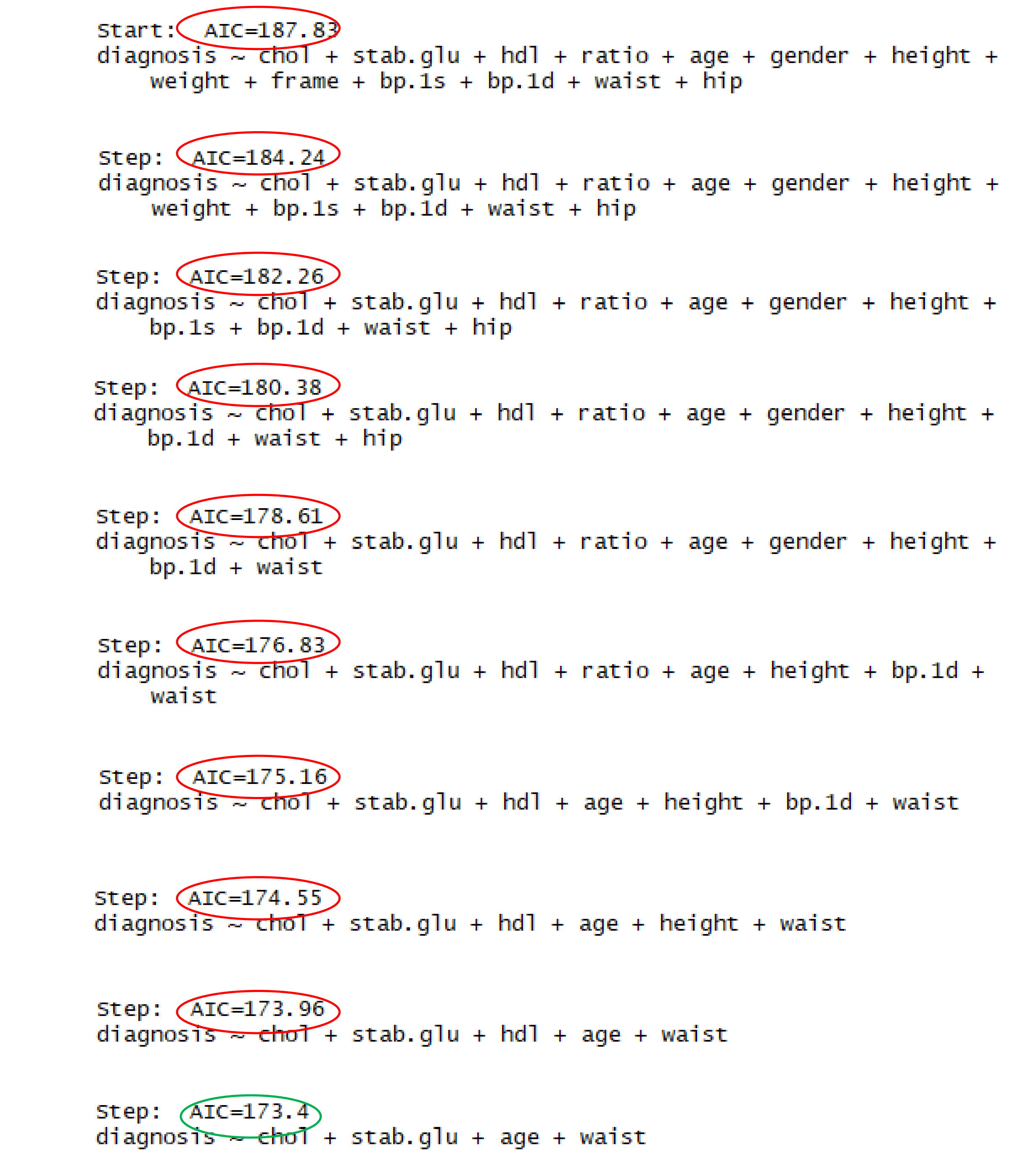


```
Call:
glm(formula = diagnosis ~ ., family = binomial, data = dbt)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.6415  -0.3444  -0.2173  -0.1170   3.3647  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -6.974692   6.651932  -1.049   0.2944    
chol         0.013303   0.008972   1.483   0.1382    
stab.glu     0.034530   0.005324   6.486 8.83e-11 ***
hdl         -0.041416   0.032302  -1.282   0.1998    
ratio       -0.205229   0.283196  -0.725   0.4686    
age          0.034116   0.018325   1.862   0.0626 .  
gender2      0.212599   0.708617   0.300   0.7642    
height      -0.060518   0.086051  -0.703   0.4819    
weight       0.002973   0.013386   0.222   0.8242    
frame2      -0.117902   0.616096  -0.191   0.8482    
frame3      -0.426204   0.749493  -0.569   0.5696    
bp.1s        0.005218   0.012216   0.427   0.6693    
bp.1d        0.012549   0.021527   0.583   0.5599    
waist        0.072431   0.080237   0.903   0.3667    
hip         -0.040285   0.080083  -0.503   0.6149    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Step 2: Variable Selection

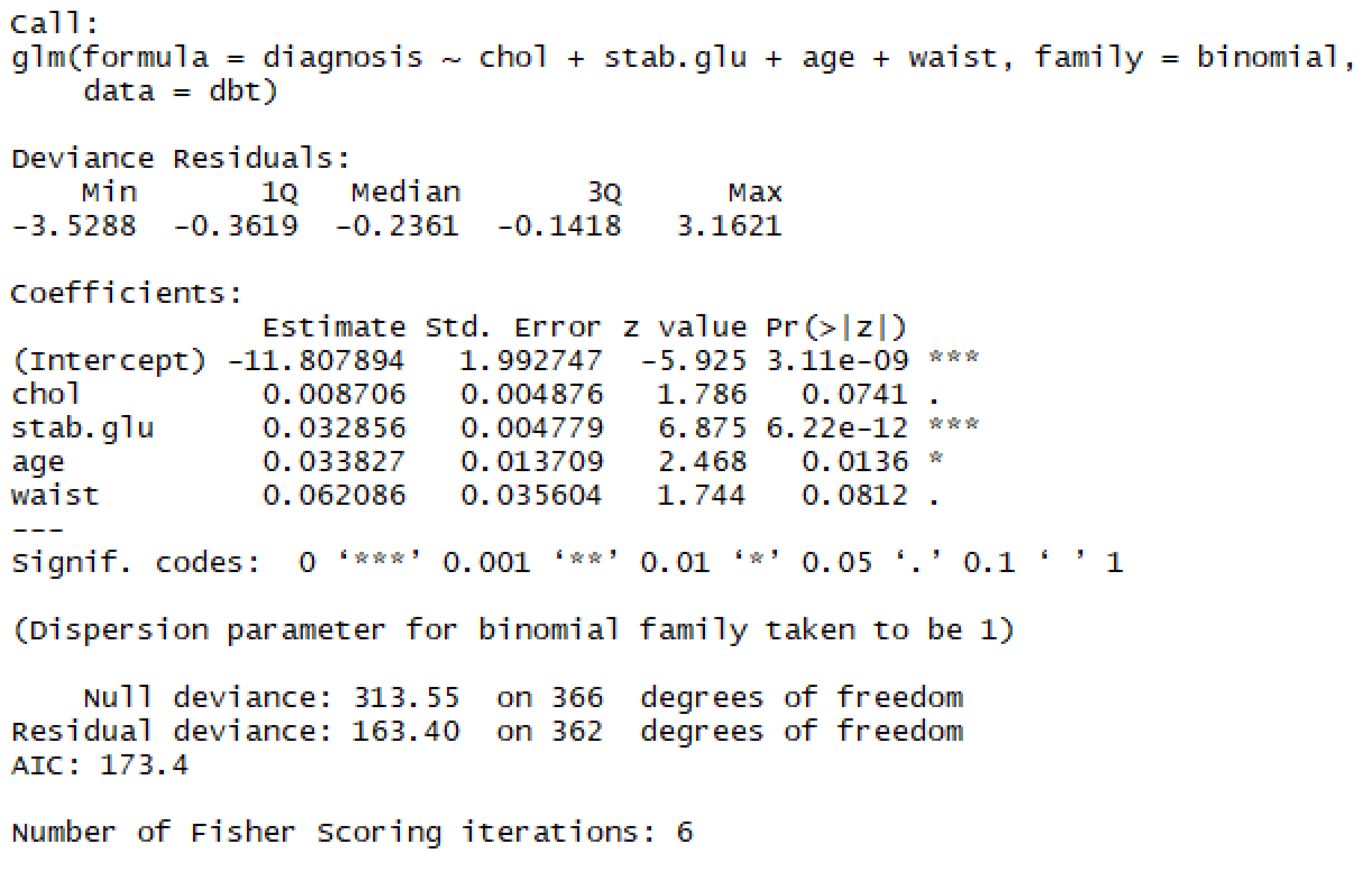
- Step() uses backward selection to remove variables
- Removes variables one at a time but only if removal reduced AIC
- Stops when further removal does not decrease AIC

Final Model Selected



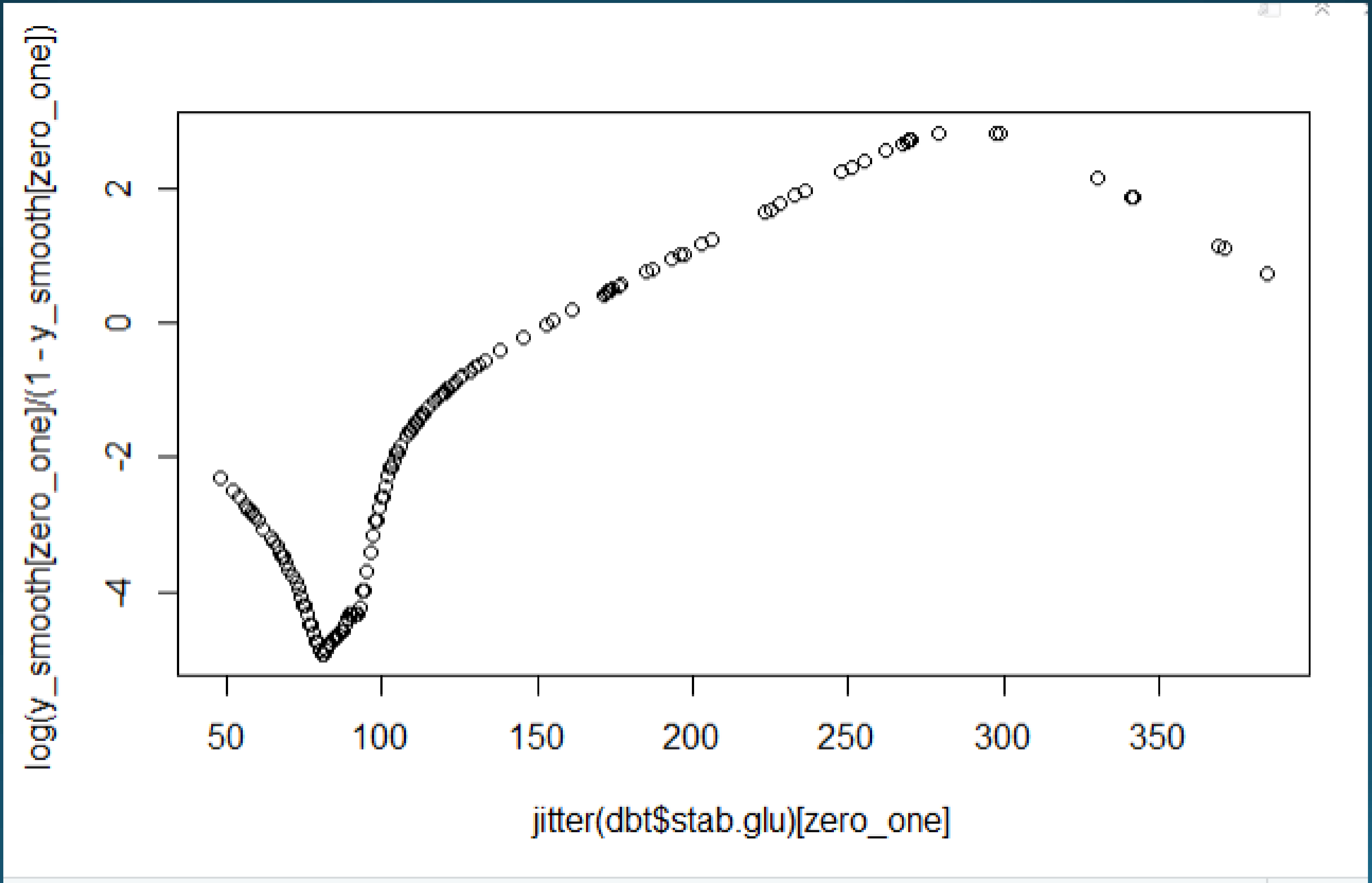
```
Start:  AIC=187.83
diagnosis ~ chol + stab.glu + hdl + ratio + age + gender + height +
    weight + frame + bp.1s + bp.1d + waist + hip

Step:  AIC=184.24
diagnosis ~ chol + stab.glu + hdl + ratio + age + gender + height +
    weight + bp.1s + bp.1d + waist + hip

Step:  AIC=182.26
diagnosis ~ chol + stab.glu + hdl + ratio + age + gender + height +
    bp.1s + bp.1d + waist + hip

Step:  AIC=180.38
diagnosis ~ chol + stab.glu + hdl + ratio + age + gender + height +
    bp.1d + waist + hip

Step:  AIC=178.61
diagnosis ~ chol + stab.glu + hdl + ratio + age + gender + height +
    bp.1d + waist

Step:  AIC=176.83
diagnosis ~ chol + stab.glu + hdl + ratio + age + height + bp.1d +
    waist

Step:  AIC=175.16
diagnosis ~ chol + stab.glu + hdl + age + height + bp.1d + waist

Step:  AIC=174.55
diagnosis ~ chol + stab.glu + hdl + age + height + waist

Step:  AIC=173.96
diagnosis ~ chol + stab.glu + hdl + age + waist

Step:  AIC=173.4
diagnosis ~ chol + stab.glu + age + waist
```

Step 2: Model Selected

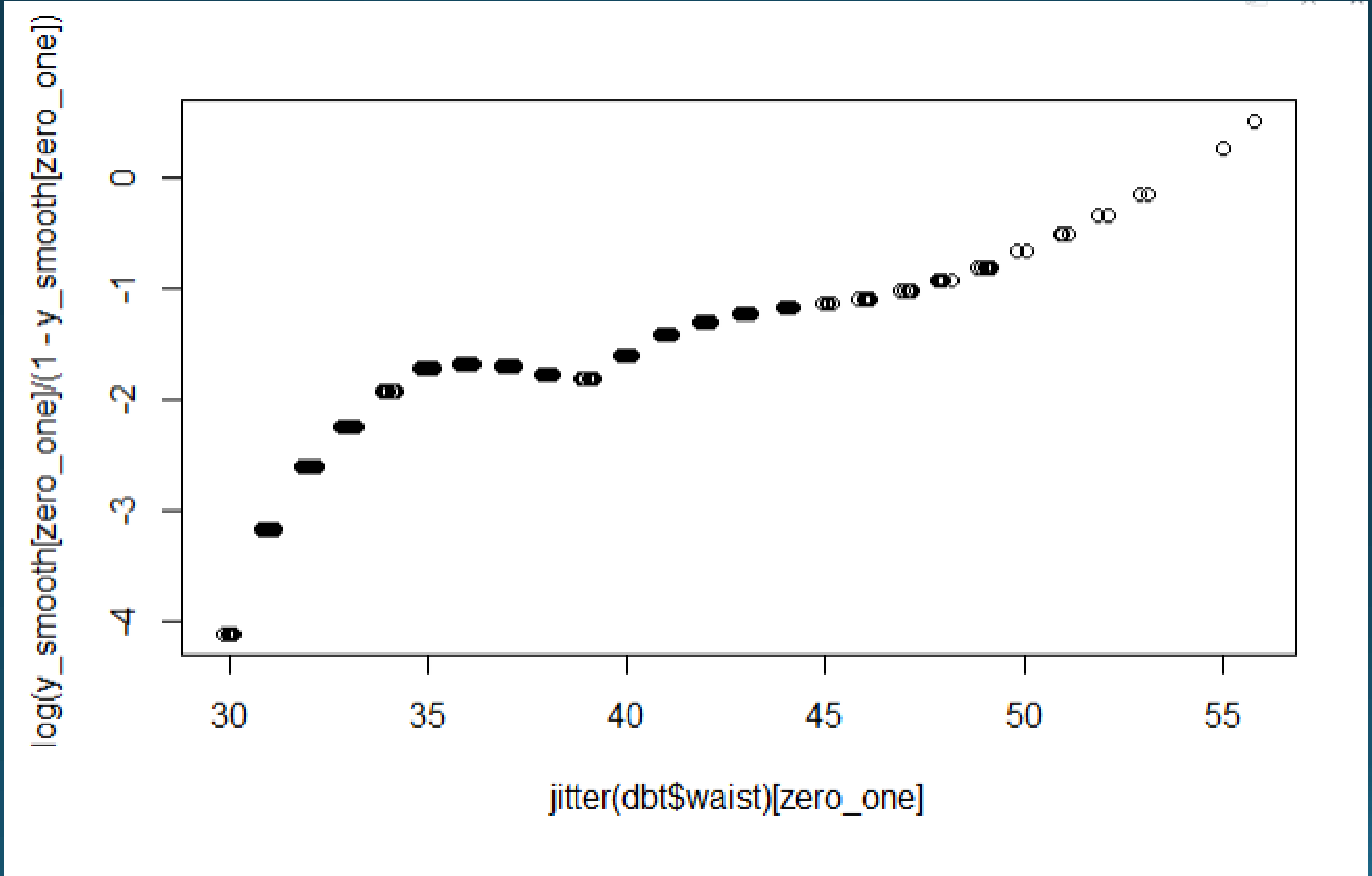


```
Call:
glm(formula = diagnosis ~ chol + stab.glu + age + waist, family = binomial,
    data = dbt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5288  -0.3619  -0.2361  -0.1418   3.1621

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.807894   1.992747  -5.925 3.11e-09 ***
chol          0.008706   0.004876   1.786   0.0741 .
stab.glu      0.032856   0.004779   6.875 6.22e-12 ***
age           0.033827   0.013709   2.468   0.0136 *
waist         0.062086   0.035604   1.744   0.0812 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 313.55  on 366  degrees of freedom
Residual deviance: 163.40  on 362  degrees of freedom
AIC: 173.4

Number of Fisher Scoring iterations: 6
```

Step 3: Stabilized Glucose

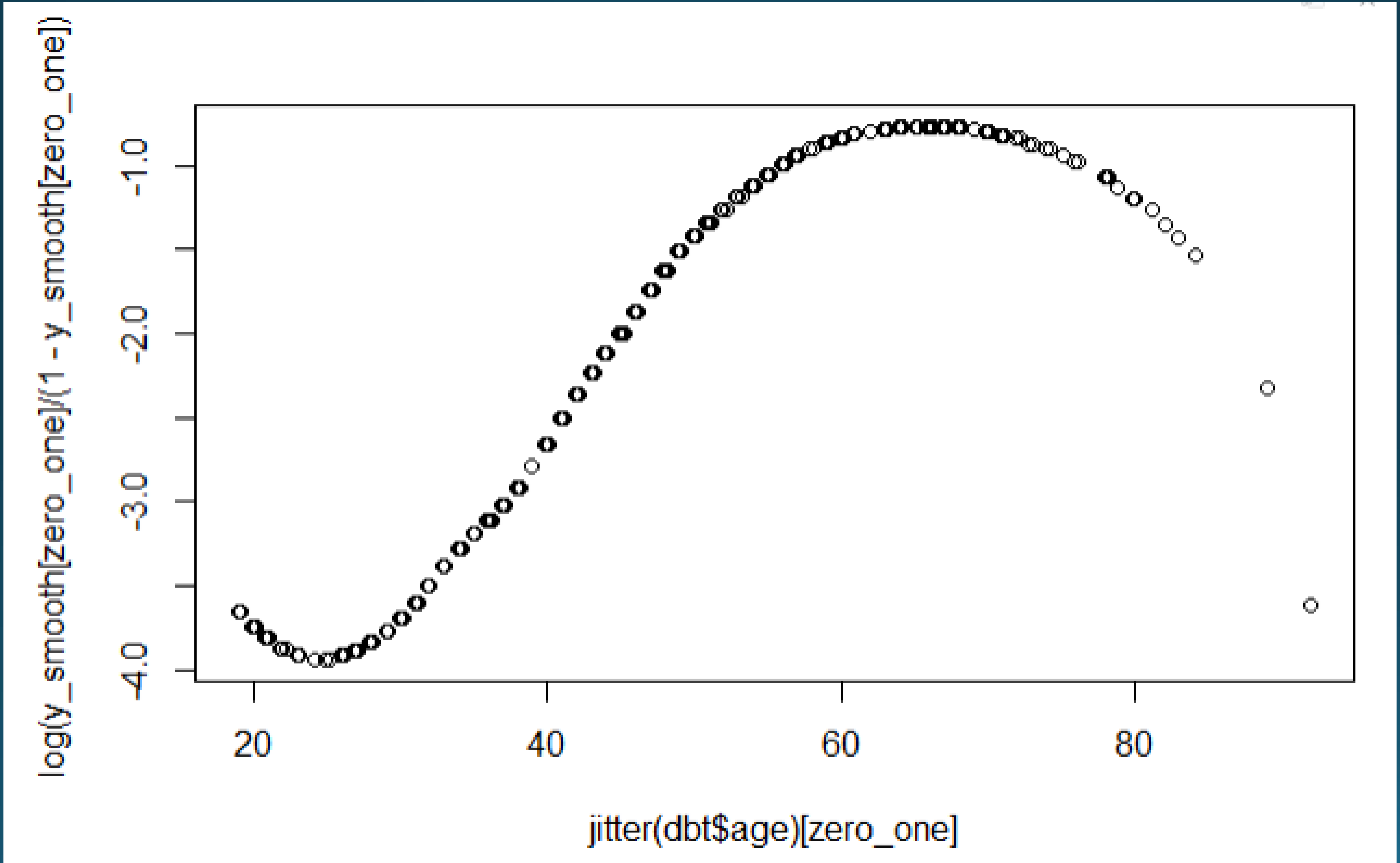


Step 3: Waist

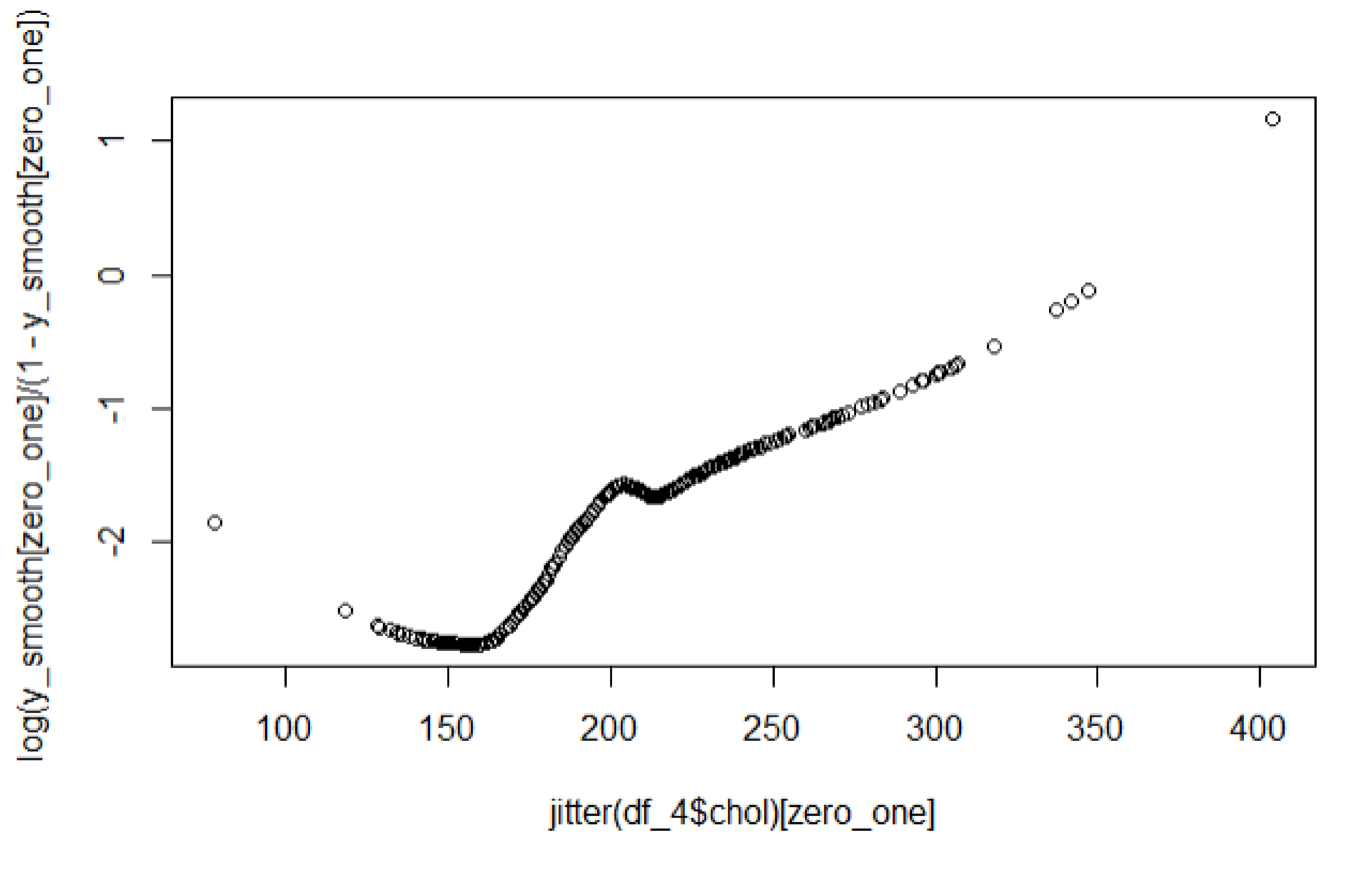


log(y_smooth[zero_one]/(1 - y_smooth[zero_one]))
0
-1
-2
-3
-4
30
35
40
45
50
55
jitter(dbt$waist)[zero_one]

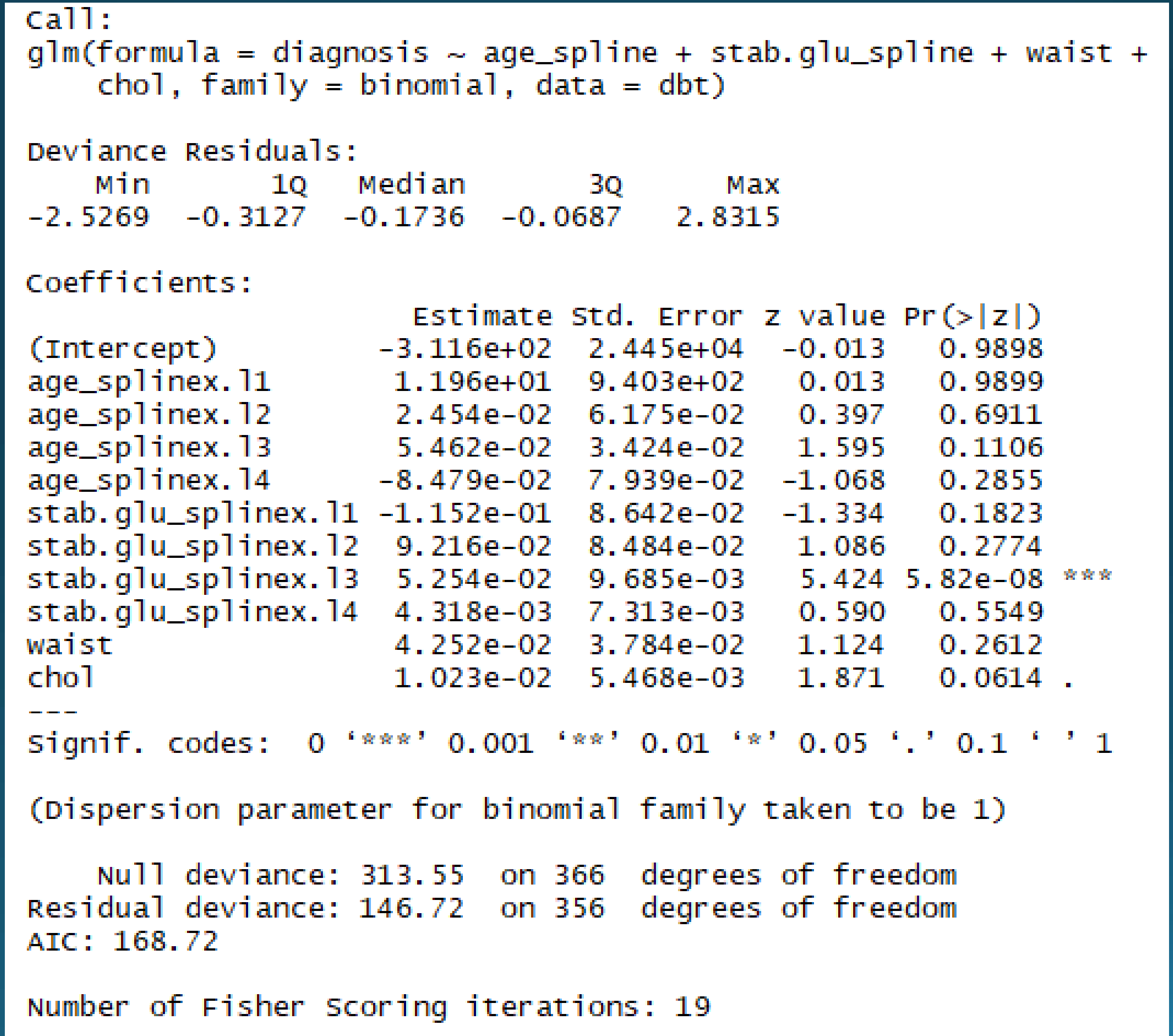
Step 3: Age


log(y_smooth[zero_one]/(1 - y_smooth[zero_one])
-1.0
-2.0
-3.0
-4.0
20
40
60
80
jitter(dbt$age)[zero_one]

Step 3: Cholesterol



Step 4: Refit Model with Splines



```
Call:
glm(formula = diagnosis ~ age_spline + stab.glu_spline + waist +
    chol, family = binomial, data = dbt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5269  -0.3127  -0.1736  -0.0687   2.8315

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -3.116e+02  2.445e+04  -0.013   0.9898
age_splinex.l1       1.196e+01  9.403e+02   0.013   0.9899
age_splinex.l2       2.454e-02  6.175e-02   0.397   0.6911
age_splinex.l3       5.462e-02  3.424e-02   1.595   0.1106
age_splinex.l4      -8.479e-02  7.939e-02  -1.068   0.2855
stab.glu_splinex.l1 -1.152e-01  8.642e-02  -1.334   0.1823
stab.glu_splinex.l2  9.216e-02  8.484e-02   1.086   0.2774
stab.glu_splinex.l3  5.254e-02  9.685e-03   5.424 5.82e-08 ***
stab.glu_splinex.l4  4.318e-03  7.313e-03   0.590   0.5549
waist                4.252e-02  3.784e-02   1.124   0.2612
chol                 1.023e-02  5.468e-03   1.871   0.0614 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 313.55  on 366  degrees of freedom
Residual deviance: 146.72  on 356  degrees of freedom
AIC: 168.72

Number of Fisher Scoring iterations: 19
```

Step 3: Homer-Lemeshow Goodness of Fit Test

- Ho: Model fit is adequate
- Ha: Model fit is not adequate

- Test Statistic: 5.6728
- Degrees of Freedom: 8

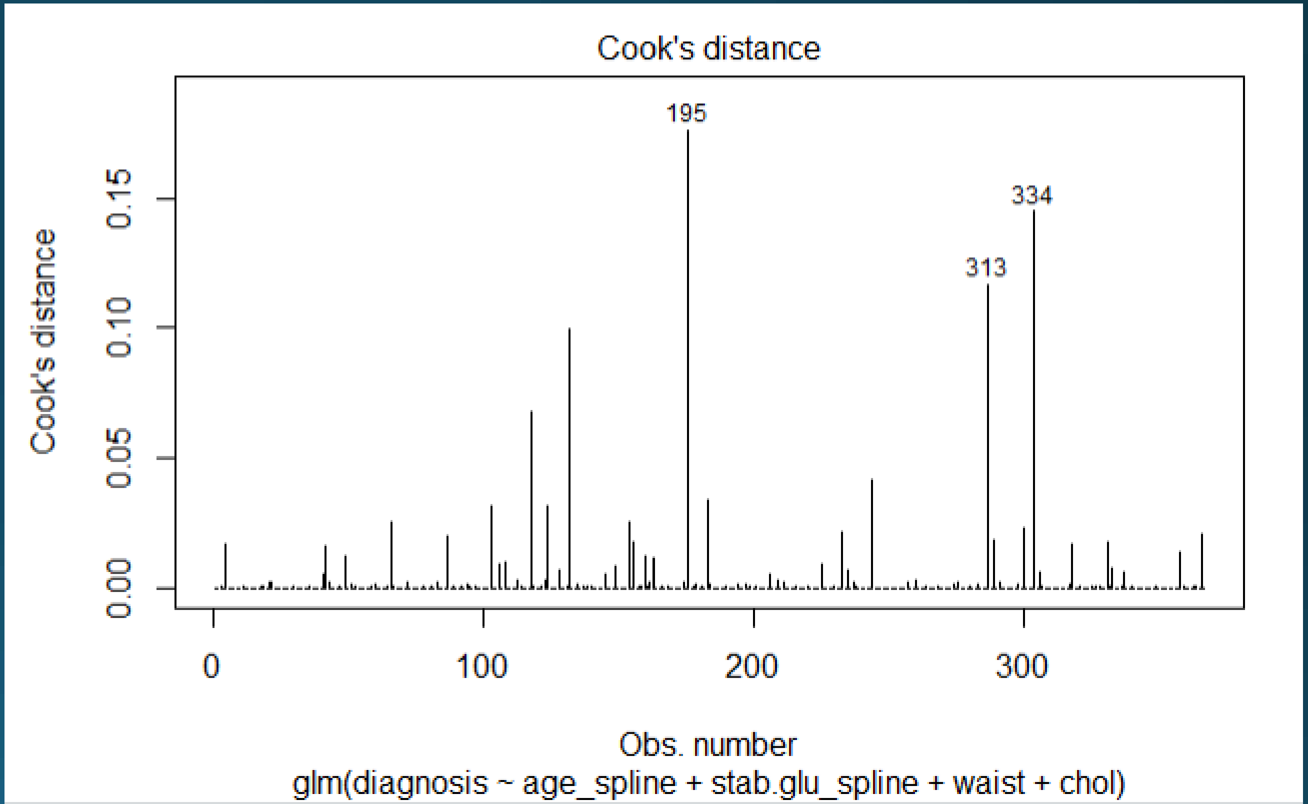
- P-value: 0.6838

Fail to Reject Ho

Model Fit is Adequate

Step 4: Influential Observations

- Cook's Distance shows no outliers (i.e. no values >1)
- Observation removal would not change model
- Therefore no changes were made to the data



Step 4: 95% Confidence Intervals and P-values

Cholesterol: β_{chol}= 0.00102

(0.0002,0.0215)

We are 95% confident that the true population value of cholesterol slope falls within the interval

p-value= 0.061

Assuming Ho is true, there is a 0.061 chance of obtaining a cholesterol slope with a magnitude of ≥ 0.00102

Step 4: 95% Confidence Intervals and P-values

Waist: β_{waist}= 0.00425

(-0.321, 0.1173)

We are 95% confident that the true population value of waist slope falls within the interval

p-value=0.261

Assuming Ho is true, there is a 0.061 chance of obtaining a waist slope with a magnitude of ≥ 0.00425

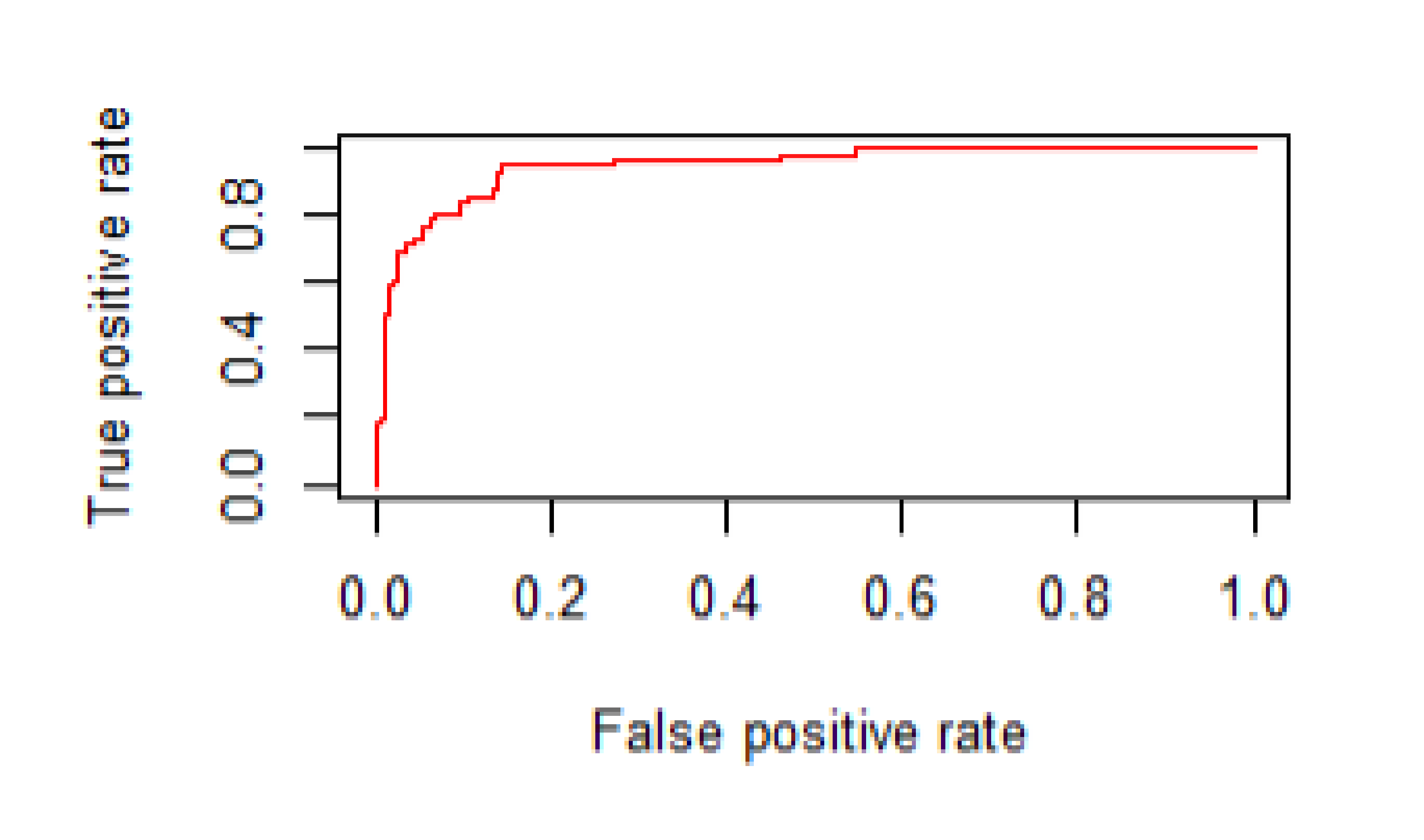
Step 4: 95% Confidence Intervals and P-values

These statistics are not meaningful for splines so no confidence intervals were computed for the age and stabilized glucose splines

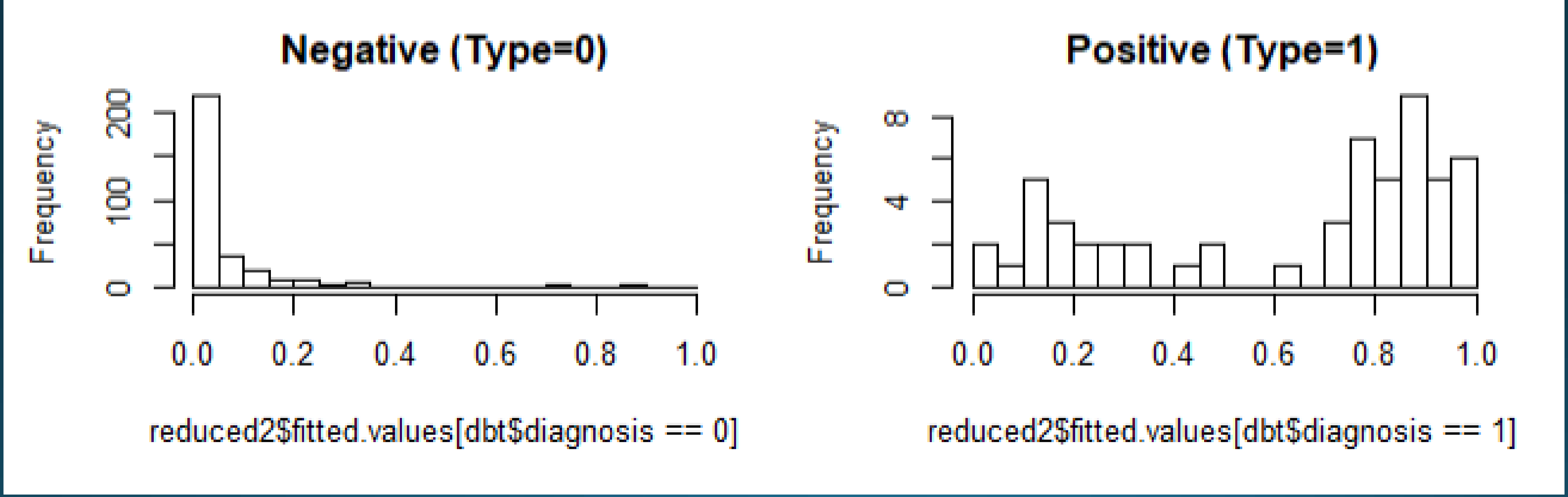
Step 5: ROC Curve

Excellent Model Discrimination

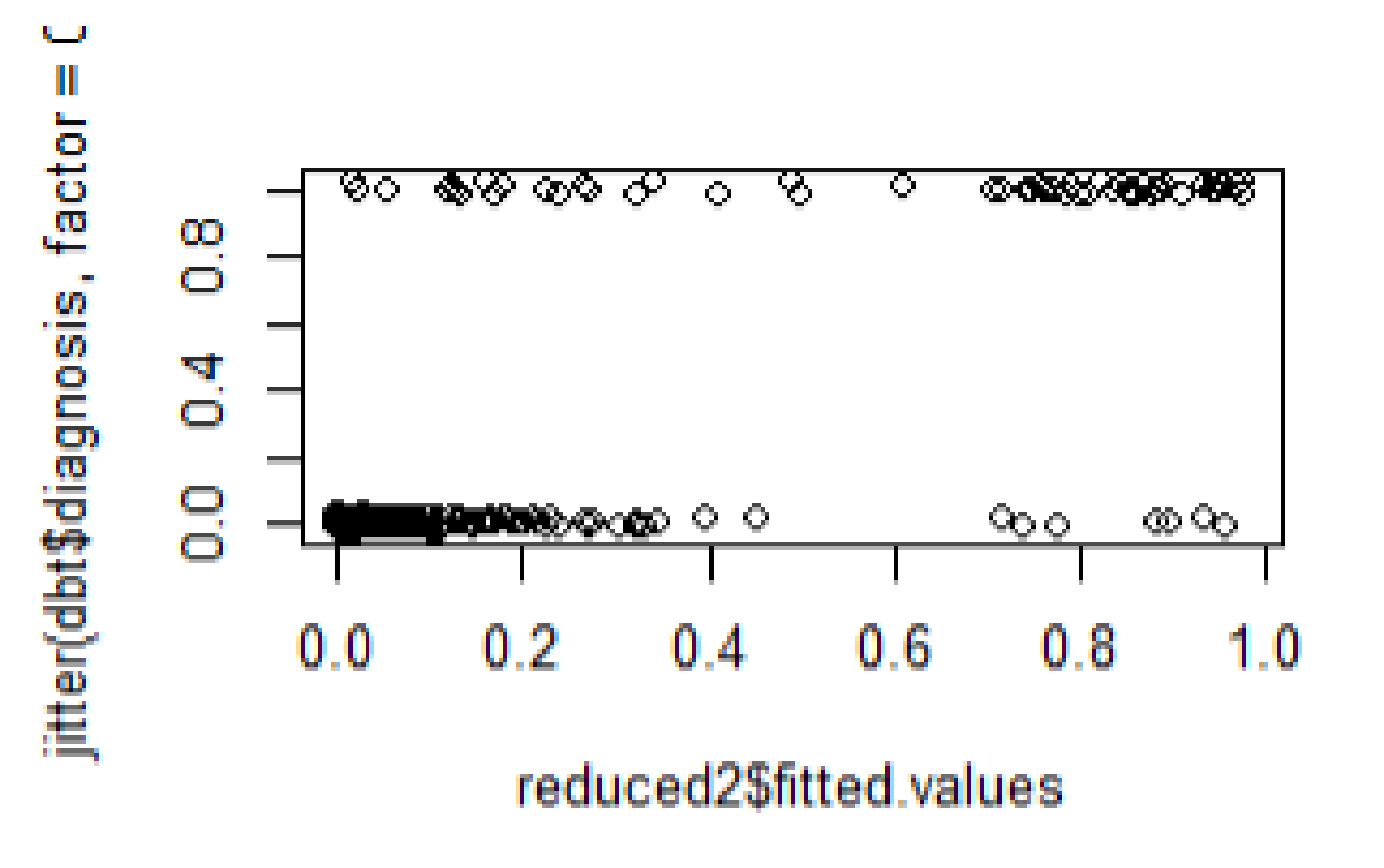
AUC= 0.946



Step 5: Predicted Probabilities Histogram



Step 5: Fitted Values vs True Diagnosis



Conclusions

- Resulting model performs well
  - Good discrimination
  - Most positives identified

- Room for improvement
  - Use testing/training data to prevent overfitting
  - Use more data for more balanced sample

- Future Areas of Exploration
  - Forward selection with splines in place to test if same variables selected
  - Try BMI in modeling
  - Speak with a subject matter expert