Quantifying the Risk of Being Diagnosed with Diabetes Using Generalized Linear Mixed Effect Models

By: Matthew Brigham

STA 531

Final Project

Fall 2020

Table of Contents

Background

Diabetes is a common disease affecting 1 in 10 Americans. Diabetes can be dangerous because it is known to harm nerves and blood vessels, leading to serious health complications, such as strokes, heart disease, and blindness. It is known that diabetes affects cohorts of age, gender, and race differently. However, there are other individual factors that can increase the risk of diabetes, such as obesity.

Diabetes is more common among non-Hispanic blacks than whites and Asian Americans. (ADA) It is also known that obesity, BMI, HDL cholesterol, gender, waist-to-hip ratio, and height can all be determinants of diabetes, as well. These are all simple and cheap measures that could be used to determine the risk that an individual has for getting diabetes. The development of a linear model could help to quantify the risk associated with these factors and being diagnosed with diabetes.

Problem Description

Using publicly available data, a generalized mixed effect model will be built to help assess the risk of being diagnosed with diabetes.

Methods

Data Description

The data was retrieved from the Biostatistics Department at Vanderbilt and was originally collected by Dr. John Schorling of the University of Virginia School of Medicine for a

study on church-based smoking cessation interventions for rural African Americans. (SCHORLING, 1997). Data was collected from the African American populations from two different counties in Virginia in the 1990s. There were 19 variables and 403 individuals studied; Table 1 describes the variables measured.

VARIABLE	DESCRIPTION
ID	Subject ID
CHOL	Total cholesterol
STAB.GLU	Stabilized glucose conc. In blood
HDL	High density lipoprotein
RATIO	Ratio of cholesterol to HDL
GLYHB	Glycosylated hemoglobin
LOCATION	County observation resides in; either Buckingham or Louisa
AGE	Age of participant in years
GENDER	Gender of participant
HEIGHT	Height of participant in inches
WEIGHT	Weight of participant in pounds
FRAME	Description of body frame as small, medium, or large
BP.1S	First systolic blood pressure reading
BP.1D	First diastolic blood pressure reading
BP.2S	Second systolic blood pressure reading
BP.2D	Second diastolic blood pressure reading
WAIST	Measure of waist circumference in inches
HIP	Measure of hip circumference in inches
TIME.PPN	Post-prandial time between when labs were drawn recorded in minutes

Table 1: Description of variables in dataset.

Figure 1 below shows scatterplots of the response (as a continuous variable) and some of the continuous predictors. There are clear linear relationships between some (Stabilized Glucose) and there are weak relationships between others (total cholesterol to HDL ratio). The

scatterplots also suggest that the data might be skewed for some of the predictors, such as age and BMI.

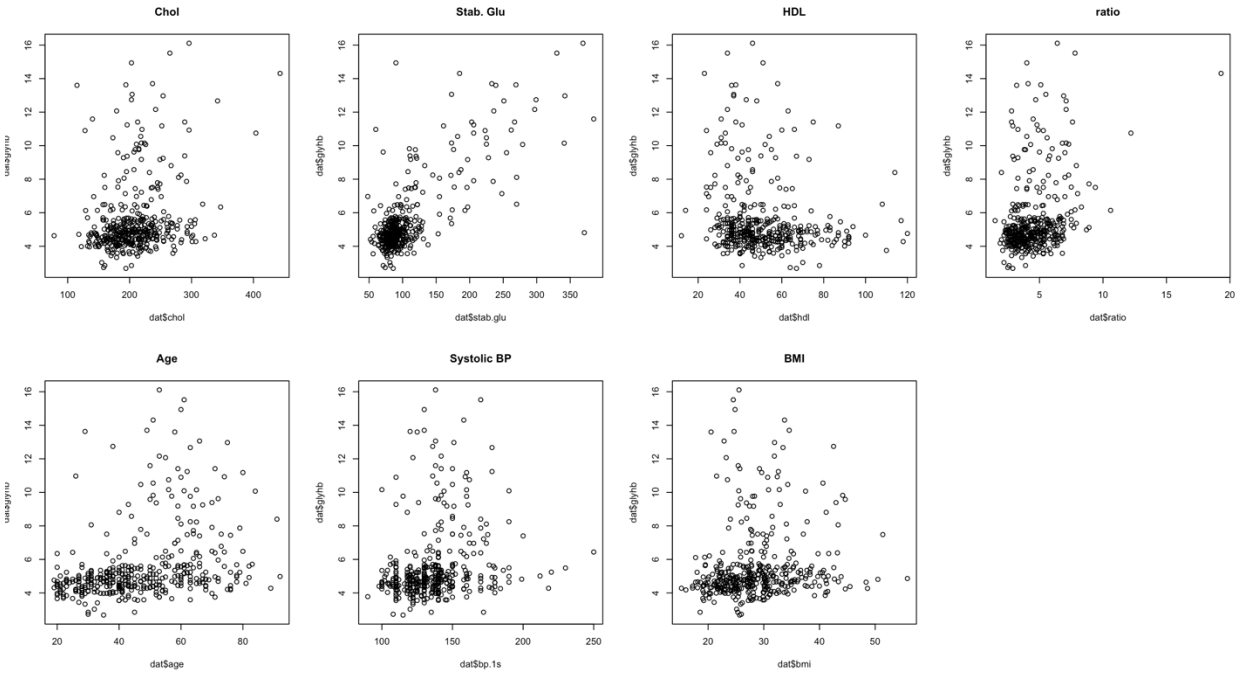

Figure 1: Scatterplots show the relationship between the predictors and the response. Some of the data appears to be skewed and there is some linear relationship.

Missing Values and Imputation

There are a substantial number of missing values in this data set. There are 575 NA values and 524 of those belong to two variables: the second readings of systolic and diastolic blood pressure (262 missing values each). That means that approximately 65% of the measurements for these two variables are missing. For this reason, these variables were omitted from the analysis. The other variables had at most 3.2% missingness. The amount of missing variables are illustrated in Figure 2 below. Upon examination of missing values, it is assumed that they are missing completely at random (MCAR). The missing data values were imputed using the "mice" package in R. There were a total of 5 imputed datasets using a

predictive mean matching imputation method. Figure 3 shows a density plot of the imputed values overlayed with the density of the original values. This figure shows that the imputed values (colored magenta) are similar to the original.

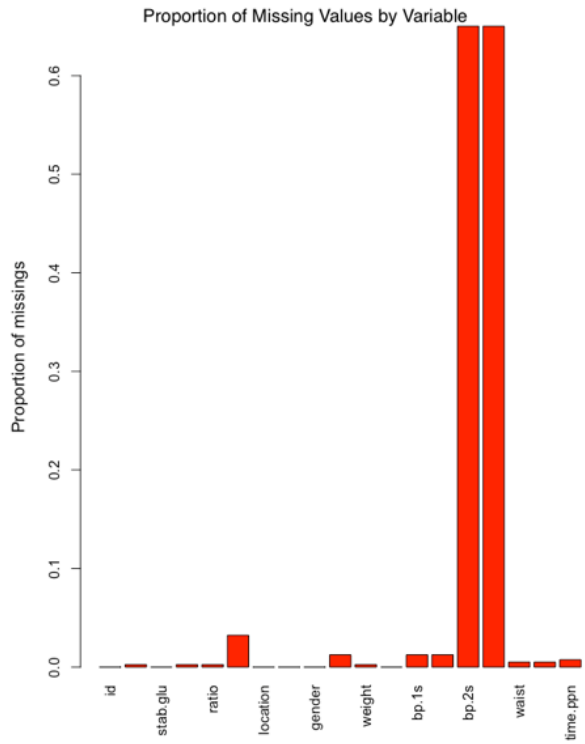

Figure 2: The proportion of missing values is shown for each variable. For data imputation, it is advised that no more than 5% of the data be missing.

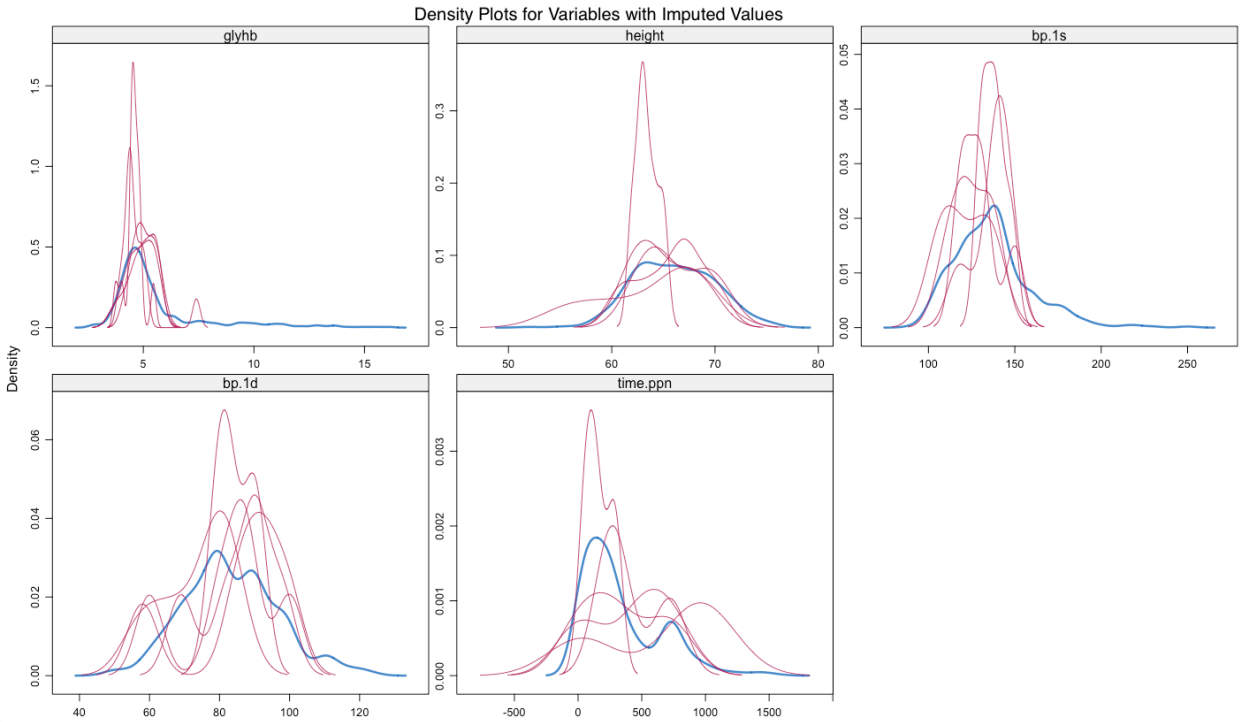

Figure 3: Density plots are shown for the variables with imputed missing values. The blue distributions represent original data and the magenta distributions (for the 5 imputed datasets) represent imputed values. It is desirable to see peaks in the imputed data near the peaks of the original data, suggesting that the imputed data is similar to the original data.

Introduction of New Variables

The original variables were manipulated from the original. A new variable of BMI (Body Mass Index) was calculated for each observation. The formula below in Equation 1 shows how to calculate BMI from weight and height using pounds and inches. It has been observed that BMI is positively associated with having diabetes mellitus. (Bays et al, 2007) For this reason, BMI was included in the analysis.

$$BMI = 703\frac{weight}{(height)^2} \quad (1)$$

The response variable, glycosylated hemoglobin, is a recorded as continuous variable and is used to diagnose diabetes. A positive diagnosis of diabetes is given for values greater than 7.0. A new categorical variable was created to represent a positive diagnosis of diabetes (glyhb > 7.0).

Preliminary Data Exploration

A preliminary analysis of the linear relationship between continuous predictor variables and response variable (glyhb) was performed using ANOVA and restricted cubic splines (utilizing 3 knots). The ANOVA uses a Chi-Square test to determine the significance of a non-linear relationship between the predictor and response. Figure 4 shows the output for the Chi-Square test. Since the p-value is greater than 0.05, the non-linear term is not significant and a linear relationship exists. This analysis was performed for the remaining continuous variables and the results are summarized in Table 2. Terms determined to have significant non-linear relationships with diabetes used restricted cubic splines in the regression model discussed later.

Wald Statistics Response: glyhb

Factor	Chi-Square	d.f.	P
chol	14.17	2	0.0008
Nonlinear	0.01	1	0.9129
TOTAL	14.17	2	0.0008

Figure 4: The results of fitting a restricted cubic splines model of cholesterol onto glycosylated hemoglobin. The p-value of the Chi-Square test suggests that a linear relation will suffice in predicting the response.

VARIABLE	χ^2 STATISTIC	DF	P-VALUE	RELATIONSHIP TO RESPONSE
CHOLESTEROL	0.01	1	0.913	Linear
STABILIZED GLUCOSE	10.20	1	0.001	Non-linear
HDL	3.83	1	0.050	Linear
RATIO CHOL/HDL	0.32	1	0.570	Linear
AGE	5.74	1	0.0167	Non-Linear
HEIGHT	0.00	1	0.947	Linear
WEIGHT	2.57	1	0.109	Linear
SYSTOLIC B.P.	5.58	1	0.018	Non-Linear
DIASTOLIC B.P.	0.35	1	0.556	Linear
WAIST	1.96	1	0.161	Linear
HIP	2.97	1	0.085	Linear
POST-PRANDIAL TIME	0.00	1	0.968	Linear
BMI	3.27	1	0.071	Linear

Table 2: The results of a Chi-Square test are shown; the results describe the linearity of the relationship between the continuous predictors and the response.

Model Building

Model Selection

A generalized linear mixed model (GLMM) and multiple logistic regression model (Generalized Linear Model) were chosen to model the data. A comparison of these models' fit is explored. The data was collected from a group of African Americans in a congregated region of Virginia and those who interacted regularly at the same churches. It is suspected that the data may be clustered for this reason. The incorporation of a random variable in the GLMM helps to model the possible correlation of the observations. GLMMs make several assumptions: 1) normality, 2) the logit link function is appropriate, and 3) there is constant variance among residual errors and random effects. GLMMs use Gauss-Hermite Quadrature to calculate maximum likelihood estimates of the parameters. A random intercept model was chosen because longitudinal data was not considered. If longitudinal data was considered, it

may have been desirable to include a random slope term to model the random aspect to the change in blood pressure measurements for each individual. However, since the only data considered is the data that the participants show up with, a random intercept model was chosen because the intercept represents the participants' baseline measurements and no change is measured. Multiple logistic regression models are easier to interpret and work with in R, therefore a likelihood ratio test was used to compare the fits of the GLMM and multiple logistic regression models. Multiple logistic regression utilizes maximum likelihood estimation to approximate the model parameters.

A full GLMM model was fit and then backwards selection was used to reduce the model. A random intercept was incorporated based on participant location. For the sake of interpretability, a multiple logistic regression model was fit and a likelihood ratio test was used to compare the two models. The test is used to determine of the difference in log-likelihoods of the two models is statistically significant (the reduced models were used in the comparison after confirming that both the reduced models were adequate compared to their respective full models using likelihood ratio tests). The result was a p-value of 0.990, suggesting that the two models are not significantly different. Therefore, a multiple logistic regression model was used and those results are presented. The R output for the full and reduced GLMM models are shown in Figures 5 and 6, respectively. The R output for the full model is shown below in Figure 7. There are some drawbacks to using backwards elimination, including that some predictors that are important may be eliminated based on p-value. Another hazard of backwards selection is that variables may become insignificant after other variables are removed and then may become significant again. (Chowdhury & Turin, 2020) For this reason, outside research

was used to double-check the validity of the findings of backwards selection. Some variables were kept solely due to the findings of other researchers. These variables include BMI, waist-to-hip ratio, HDL, obesity (measured with BMI in our case), and height. (Schmidt et al, 1992; Bays et al, 2007; Farbstein & Levy, 2012; WILLER; Schulze et al 1970) For this reason, BMI and the waist-to-hip ratio were kept in the model even though their coefficients were insignificant. The final reduced model is shown in Figure 8. A significance level of 0.15 was used for variables that weren't suggested as having a relationship with diabetes in other research.

```
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature, nAGQ = 10) [glmerMod]
 Family: binomial  ( logit )
Formula: glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + ratio +
    age + rcs(age, 3) + height + weight + frame + bp.1s + rcs(bp.1s,
    3) + bp.1d + waist * hip + gender * bmi + w.h.ratio + (1 |      location)
   Data: dat
Control: glmerControl(optimizer = "bobyqa")

     AIC      BIC   logLik deviance df.resid
   193.8    285.1    -73.9    147.8      368

Scaled residuals:
    Min      1Q  Median      3Q     Max
-12.766  -0.186  -0.095  -0.030  55.409

Random effects:
 Groups   Name        Variance Std.Dev.
 location (Intercept) 0        0
Number of obs: 391, groups:  location, 2

Fixed effects:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                 17.378312  34.788031   0.500   0.6174
chol                         0.013604   0.009213   1.477   0.1398
stab.glu                     0.112609   0.031808   3.540   0.0004 ***
rcs(stab.glu, 3)stab.glu' -0.202096   0.081004  -2.495   0.0126 *
hdl                         -0.050608   0.034526  -1.466   0.1427
ratio                       -0.274750   0.298238  -0.921   0.3569
age                          0.085484   0.060292   1.418   0.1562
rcs(age, 3)age'             -0.066389   0.063936  -1.038   0.2991
height                      -0.353933   0.396287  -0.893   0.3718
weight                       0.052933   0.066866   0.792   0.4286
framemedium                 -0.411135   0.651062  -0.631   0.5277
framelarge                  -0.835901   0.815884  -1.025   0.3056
bp.1s                        0.041050   0.035885   1.144   0.2526
rcs(bp.1s, 3)bp.1s'         -0.028029   0.031843  -0.908   0.3636
bp.1d                        0.002631   0.022976   0.115   0.9088
waist                        0.678614   0.665461   1.020   0.3078
hip                         -0.486744   0.577057  -0.843   0.3990
gendermale                   5.497504   3.646371   1.508   0.1316
bmi                         -0.241231   0.373216  -0.646   0.5180
w.h.ratio                  -23.487465  28.590671  -0.822   0.4114
waist:hip                   -0.001260   0.003418  -0.369   0.7124
gendermale:bmi              -0.194605   0.122699  -1.586   0.1127
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5: A full GLMM model was fit and the results are shown.

```
Control: glmerControl(optimizer = "bobyqa")

     AIC      BIC   logLik deviance df.resid
   181.3    233.2    -77.6    155.3      390

Scaled residuals:
    Min      1Q  Median      3Q     Max
-7.308 -0.200 -0.102 -0.043 49.821

Random effects:
 Groups   Name        Variance Std.Dev.
 location (Intercept) 0        0
Number of obs: 403, groups:  location, 2

Fixed effects:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -20.682689   7.486508  -2.763 0.005733 **
chol                         0.007519   0.005134   1.465 0.143041
stab.glu                     0.111573   0.030647   3.641 0.000272 ***
rcs(stab.glu, 3)stab.glu'  -0.200547   0.078178  -2.565 0.010310 *
hdl                        -0.016004   0.013897  -1.152 0.249503
age                         0.024570   0.016441   1.494 0.135070
height                     -0.012205   0.087677  -0.139 0.889287
bp.1s                       0.012636   0.009569   1.321 0.186632
gendermale                  2.308531   2.212538   1.043 0.296770
bmi                         0.042620   0.036265   1.175 0.239899
w.h.ratio                   3.498648   3.177621   1.101 0.270885
gendermale:bmi             -0.098308   0.074387  -1.322 0.186307
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) chol   stb.gl r(.,3) hdl    age    height bp.1s  gndrml bmi    w.h.rt
chol        -0.109
stab.glu    -0.533  0.010
rcs(s.,3).' 0.541  0.009 -0.988
hdl         -0.135 -0.165  0.023 -0.027
age         -0.178 -0.103 -0.074  0.054 -0.089
height      -0.785  0.005  0.131 -0.164  0.089  0.278
bp.1s       -0.103 -0.039  0.129 -0.096 -0.049 -0.362 -0.092
gendermale  -0.007  0.049 -0.009  0.047 -0.036 -0.088 -0.214  0.129
bmi         -0.275 -0.034 -0.065  0.068  0.148  0.242  0.178 -0.079  0.457
w.h.ratio   -0.249 -0.031  0.073 -0.067 -0.035 -0.175 -0.154 -0.013  0.188 -0.043
gendermal:bm  0.176 -0.031 -0.037  0.000  0.097  0.048  0.023 -0.120 -0.952 -0.430 -0.248
```

Figure 6: A reduced GLMM model was fit. A summary of the fit is shown.

```
Call:
glm(formula = glyhb.cat ~ chol + stab.glu + hdl + ratio + age +
    height + weight + bp.1s + bp.1d + waist * hip + gender *
    bmi, data = dat)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.23397  -0.11029  -0.03622   0.03830   1.11588

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.261e-02  1.367e+00  -0.009    0.993
chol            3.797e-04  5.631e-04   0.674    0.501
stab.glu        4.301e-03  2.635e-04  16.323   <2e-16 ***
hdl            -6.457e-06  1.758e-03  -0.004    0.997
ratio           1.183e-02  1.960e-02   0.603    0.547
age             1.382e-03  1.016e-03   1.360    0.175
height         -3.208e-04  2.050e-02  -0.016    0.988
weight         -7.739e-06  3.775e-03  -0.002    0.998
bp.1s           7.611e-04  8.254e-04   0.922    0.357
bp.1d          -1.967e-04  1.264e-03  -0.156    0.876
waist          -9.012e-03  1.540e-02  -0.585    0.559
hip            -1.991e-02  1.493e-02  -1.334    0.183
gendermale      9.711e-02  1.767e-01   0.549    0.583
bmi             2.904e-03  2.144e-02   0.135    0.892
waist:hip       3.462e-04  3.361e-04   1.030    0.304
gendermale:bmi -5.260e-03  6.227e-03  -0.845    0.399
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.06532482)

    Null deviance: 51.067  on 402  degrees of freedom
Residual deviance: 25.281  on 387  degrees of freedom
AIC: 61.8

Number of Fisher Scoring iterations: 2
```

Figure 7: A full multiple logistic regression model was fit using most of the original variables.

```
Call:
glm(formula = glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) +
    hdl + age + bp.1s + gender * bmi + w.h.ratio, family = binomial(),
    data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8046  -0.2787  -0.1431  -0.0596   3.9586

Coefficients: (1 not defined because of singularities)
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -21.503234   4.638557  -4.636 3.56e-06 ***
chol                        0.007525   0.005134   1.466 0.142771
stab.glu                    0.112142   0.030368   3.693 0.000222 ***
rcs(stab.glu, 3)stab.glu          NA         NA      NA       NA
rcs(stab.glu, 3)stab.glu' -0.202352   0.077033  -2.627 0.008619 **
hdl                       -0.015832   0.013843  -1.144 0.252746
age                         0.025209   0.015811   1.594 0.110842
bp.1s                       0.012514   0.009540   1.312 0.189598
gendermale                  2.242253   2.159681   1.038 0.299161
bmi                         0.043515   0.035698   1.219 0.222863
w.h.ratio                   3.431615   3.139357   1.093 0.274352
gendermale:bmi            -0.098068   0.074321  -1.320 0.186996
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 339.14  on 402  degrees of freedom
Residual deviance: 155.28  on 392  degrees of freedom
AIC: 177.28

Number of Fisher Scoring iterations: 7
```

Figure 8: After backwards elimination, a reduced model was determined. Not all of the predictors are significant, but some were kept because of significance determined in other studies.

Table 3 summarizes the results of the multiple logistic regression. The confidence intervals suggest that some of the parameters are not useful in the explanation of the response and can be removed since the interval contains 0. The coefficients were kept because other research suggests that these are significant in explaining diabetes. Based on the confidence intervals, only one of the variables had a significant relationship with predicting diabetes: stabilized glucose levels (zero is not in the 95% confidence interval)l. The estimated odds of contracting diabetes increases by $e^{0.112}$ = 1.119 times (an 11.9% increase) higher for each unit increase in stabilized glucose.

COEFFICIENT	ESTIMATE	STANDARD ERROR	CONFIDENCE INTERVAL
INTERCEPT	-21.503	4.639	(-31.286, -12.962)
CHOL	0.008	0.005	(-0.002, 0.018)
STAB.GLU	0.112	0.030	(0.0564, 0.176)
RCS(STAB.GLU, 3)STAB.GLU'	-0.202	0.077	(-0.361, -0.057)
HDL	-0.016	-0.015	(-0.044, -0.011)
AGE	0.025	0.016	(-0.006, 0.057)
BP.1S	0.013	0.010	(-0.007, 0.031)
GENDER: MALE	2.242	2.242	(-2.047, 6.482)
BMI	0.044	0.036	(-0.028, 0.114)
WAIST-TO-HIP RATIO	3.432	3.139	(-2.729, 9.647)
GENDER:MALE*BMI	-0.098	-0.098	(-0.245, 0.048)

Table 3: This table shows the same results as Figure 7. Parameter estimates are given with standard error for the reduced model.

Model Validation

A likelihood ratio test was used to compare the utility of the full and reduced models. A p-value of 0.999 suggests that the reduced model is adequate in explaining the variance in the response variable.

Singularity was an issue encountered with modeling the GLMM. Singularity in GLMMs occurs when the model may be overfit and the variance of the effects is close to zero. There is no consensus on how to handle singularity in this context. This may have impacted the fit and decision to use a Generalized Linear Model.

There were 5 data sets with different imputed values. It is customary to perform the regression on all of the datasets and then pool the results. All of the regressions produced equivalent values for the parameter estimates.

A Receiver Operating Characteristic (ROC) curve is useful in determining the overall measure of fit of the model and how well it discriminates between outcomes. Figure 9 shows the ROC curve for the multiple linear regression model. The area under the curve (AUC) is a

measure of the fit of the model. A value of 0.5 implies a 50/50 guess, therefore a high AUC is desirable because it implies a higher prediction accuracy. The multiple logistic model has an AUC value of 0.942, which suggests high predictive accuracy for the data used to build it. Such a high value for AUC raises concern for overfitting.

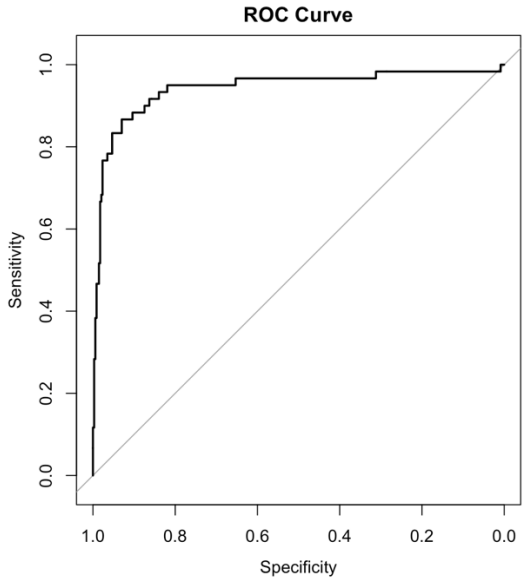

Figure 9: A ROC curve graphically shows the how well the model fits the data used to build it. The model estimates the data far better than a simple 50/50 guess.

The dispersion of the data was measured using a Chi-Square Test with a Pearson ratio test statistic. This test had a p-value of 0, suggesting the measure of overdispersion is significantly different from zero. The presence of overdispersion may have impacted the results of fit and may impact the predictive ability of the model on new data, considering the GLM has such a high AUC (the model may be overfit).

Model Assumption Check

According to SCHIELZETH (2020), GLMMs are robust to "objective" failures of the normality assumption, therefore, there is not much concern about this violation.

The constant variance assumption was evaluated by testing for the existence of overdispersion. A Chi-Square test using the Pearson statistic was performed and a p-value of 0.999 was calculated, suggesting there is no significant difference between the amount of overdispersion and 0. The constant variance assumption is presumed to be satisfied. There may be under dispersion, and this should also be tested for.

Discussion

There was a degree of relaxation when the predictors were being selected to build the model. A high significance level of 15% was used to capture more variability in the model. Also, features were chosen based on the importance discovered by the work of other researchers, even if their statistical significance did not mee the 15% threshold.

The data was collected from two small, rural counties and from African Americans who attended the same churches. Correlation between observations requires the use of a generalized mixed model. Since the data used to build the model was collected over 20 years ago, the prediction should not be relied upon for current patients. A new model should be developed based on recent data and can use the techniques employed in this paper. If new data is to be collected to explore the risk factors of diabetes, a larger dataset representing people from other areas of the country would be an improvement to help get rid of the cluster effect. The original data used in this paper was intended to solve a different research question

about church-going African Americans in rural locations and smoking programs, as it pertains to coronary disease.

Some of the assumptions of GLMMs were violated. Solutions to this issue include retrieving new data or using a different model. It has been shown by SCHIELZETH (2020) that GLMMs are robust in their parameter estimation when the normality and variance assumptions were not met. Other researchers have shown that different models, such as Artificial Neural Networks, Random Forests, and K-means clustering, are all suitable methods for predicting diabetes. (ALAM, 2019) Due to time constraints, the model was not evaluated in terms of under dispersion. Referring to the scatterplots in Figure 1, visually, it can be seen there is a possibility of under dispersion.

One issue encountered in building the model was that of singularity. A singular mixed model means that the parameters are on a boundary of the parameter space and that the variances of the effects are close to zero. The existence of singularity suggests that the model may be overfit. Singularity is a common issue encountered when using the GLMMs with large data sets. There is no clear method of handling singularity in this case, but several options have been suggested. Reducing the complexity of the model may allow the variance-covariance matrix of the GLMM to be estimated. (Matuschek et al, 2017) Other solutions involve model selection: 1) use a selection method that considers predictive accuracy and overfitting (Bates et al 2015, Matuschek et al 2017) and 2) use backwards selection prioritizing singularity rather than significance of terms. (Barr et al, 2013) Even with these suggestions, there is no consensus on how to deal with singularity in GLMMs.

All of the imputed datasets were used with the model and yielded similar results, even though they are not shown in the report (shown in the R code in Appendix I). Only the first imputed data set was used to build the regression model. Since so few data points were missing (other than the two variables that were dropped), there was very little difference in the model parameter estimates.

A multiple logistic regression model was successfully developed and shown to be a viable replacement to the GLMM in modeling the data. Further model calibration and validation is needed to assess the fit of the model and to determine the extent of multicollinearity.

References

Schorling, J. (1997). *Diabetes Dataset*. Retrieved 2020, from http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets

Schorling 1997 - Schorling JB, Roach J, Siegel M, Baturka N, Hunt DE, Guterbock TM, Stewart HL: A trial of church-based smoking cessation interventions for rural African Americans. *Preventive Medicine* 26:92-101; 1997.

Bays, H., Chapman, R., & Grandy, S. (2007, April 10). The relationship of body mass index to diabetes mellitus, hypertension and dyslipidaemia: Comparison of data from two national surveys. Retrieved December 08, 2020, from https://onlinelibrary.wiley.com/doi/full/10.1111/j.1742-1241.2007.01336.x

Chowdhury, M. Z., & Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. Retrieved 2020, from https://fmch.bmj.com/content/fmch/8/1/e000262.full.pdf

Schmidt, M., Duncan, B., Canani, L., Karohl, C., & Chambless, L. (1992, July 01). Association of Waist-Hip Ratio With Diabetes Mellitus: Strength and Possible Modifiers. Retrieved December 08, 2020, from https://care.diabetesjournals.org/content/15/7/912

Farbstein, D., & Levy, A. (2012, March). HDL dysfunction in diabetes: Causes and possible treatments. Retrieved December 08, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3332215/

Kautzky-Willer, A., Harreiter, J., & Pacini, G. (2016, June). Sex and Gender Differences in Risk, Pathophysiology and Complications of Type 2 Diabetes Mellitus. Retrieved December 08, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4890267/

MB. Schulze, C., N. Eckel, K., M. Janghorbani, F., N. Stefan, H., MB. Schulze, K., K. Muhlenbruch, T., . . . SG. Wannamethee, A. (1970, January 01). Associations of short stature and components of height with incidence of type 2 diabetes: Mediating effects of cardiometabolic risk factors. Retrieved December 08, 2020, from https://link.springer.com/article/10.1007/s00125-019-04978-8

Dansinger, M. (2019, May 18). Diabetes & High Blood Pressure: Managing Diabetic Hypertension. Retrieved December 08, 2020, from https://www.webmd.com/diabetes/high-blood-pressure

National Diabetes Statistics Report, 2020. (2020, February 11). Retrieved December 08, 2020, from https://www.cdc.gov/diabetes/library/features/diabetes-stat-report.html

Statistics About Diabetes. (n.d.). Retrieved December 08, 2020, from https://www.diabetes.org/resources/statistics/statistics-about-diabetes

Schielzeth, H., Dingemanse, N., Nakagawa, S., Westneat, D., Allegue, H., Teplitsky, C., . . . Araya-Ajoy, Y. (2020, July 16). Robustness of linear mixed-effects models to violations of distributional assumptions. Retrieved December 08, 2020, from https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.13434

Alam, T., Iqbal, M., Ali, Y., Wahab, A., Ijaz, S., Baig, T., . . . Abbas, Z. (2019, July 09). A model for early prediction of diabetes. Retrieved December 08, 2020, from https://www.sciencedirect.com/science/article/pii/S2352914819300176

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal; *Journal of Memory and Language* **68**(3), 255–278.

Douglas Bates, Reinhold Kliegl, Shravan Vasishth, and Harald Baayen (2015). *Parsimonious Mixed Models*; preprint (https://arxiv.org/abs/1506.04967).

Hannes Matuschek, Reinhold Kliegl, Shravan Vasishth, Harald Baayen, and Douglas Bates (2017). Balancing type I error and power in linear mixed models. *Journal of Memory and Language* **94**, 305–315.

Appendix I: R Code and Output

```
> #Matthew Brigham
> #STA 531
> #Final Project
> #
> # Diabetes With Generalized Linear Mixed Models and Multiple Logistic Regression.
> # MICE missing value imputation was used and one data set was used to beuild the models.
> # At the end of the code, the final regression model was used on the remaining imputed data sets
> # and it is shown that all of the coefficients are similar.
> #
> # Note: glyhb > 7 is diagnosed diabetes
> #
> # fit_08 is reduced GLM model
>
>
> ########################################
> ############# Import Data #############
> ########################################
>
> orig_dat = read.csv("diabetes.csv")
> head(orig_dat)
   id chol stab.glu hdl ratio glyhb   location age gender height weight  frame bp.1s bp.1d bp.2s bp.2d waist hip
1 1000  203      82  56   3.6  4.31 Buckingham  46 female     62    121 medium   118    59    NA    NA    29  38
2 1001  165      97  24   6.9  4.44 Buckingham  29 female     64    218  large   112    68    NA    NA    46  48
3 1002  228      92  37   6.2  4.64 Buckingham  58 female     61    256  large   190    92   185    92    49  57
4 1003   78      93  12   6.5  4.63 Buckingham  67   male     67    119  large   110    50    NA    NA    33  38
5 1005  249      90  28   8.9  7.72 Buckingham  64   male     68    183 medium   138    80    NA    NA    44  41
6 1008  248      94  69   3.6  4.81 Buckingham  34   male     71    190  large   132    86    NA    NA    36  42
  time.ppn
1      720
2      360
3      180
4      480
5      300
6      195
>
> dat0 = orig_dat
>
> ########################################
> ########## Data Manipulations #########
> ########################################
>
>
> # Identify and Impute Missing Values
>
>    library(mice)
>    library(VIM)
>
>    # Look for Pattern of Missingness
>    md.pattern(dat0)  #tells us number of observations with that missing data structure
    id stab.glu location age gender frame chol hdl ratio weight waist hip time.ppn height bp.1s bp.1d glyhb bp.2s
136  1       1        1   1      1     1    1   1     1      1     1   1        1      1     1     1     1     1
239  1       1        1   1      1     1    1   1     1      1     1   1        1      1     1     1     1     0
3    1       1        1   1      1     1    1   1     1      1     1   1        1      1     1     1     0     1
10   1       1        1   1      1     1    1   1     1      1     1   1        1      1     1     1     0     0
3    1       1        1   1      1     1    1   1     1      1     1   1        1      1     0     0     1     0
4    1       1        1   1      1     1    1   1     1      1     1   1        1      0     1     1     1     0
1    1       1        1   1      1     1    1   1     1      1     1   1        1      0     0     0     1     0
1    1       1        1   1      1     1    1   1     1      1     1   1        0      1     1     1     1     1
1    1       1        1   1      1     1    1   1     1      1     1   1        0      1     1     1     1     0
1    1       1        1   1      1     1    1   1     1      1     1   1        0      1     0     0     1     0
1    1       1        1   1      1     1    1   1     1      1     0   0        1      1     1     1     1     1
1    1       1        1   1      1     1    1   1     1      1     0   0        1      1     1     1     1     0
```

```
1   1      1      1  1    1    1   1  1    1    0    1  1      1    1    1    1    1    0
1   1      1      1  1    1    1   0  0    0    1    1  1      1    1    1    1    1    0
   0      0      0  0    0    0   1  1    1    1    2  2      3    5    5    5   13  262
  bp.2d
136    1  0
239    0  2
3      1  1
10     0  3
3      0  4
4      0  3
1      0  5
1      1  1
1      0  3
1      0  5
1      1  2
1      0  4
1      0  3
1      0  5
    262 563
>    agg_plot = aggr(dat0, combined = F) #histogram showing proportion of missing data by variable
>
>
>    # Impute NA values
>    dat1=dat0
>    dat1 = dat0[,-c(15, 16)] #remove categorical variables and 2nd measurements of BP
>    imp = mice(dat1, maxit = 0) # 0 iterations
Warning message:
Number of logged events: 3
>
>    pred_mat = imp$predictorMatrix  #get predictor matrix
>    method = imp$method  #get methods, mostly pmm: predictive mean matching
>
>    imp2 = mice(dat1, m=5, maxit = 50, predictorMatrix = pred_mat, method = method, seed = 500, print = F) #creates 5 datasets of imputed
values
>
>    # Analyze the distribution with and without imputed values for each of 5 imputed datasets
>    newdat_imp_1 = complete(imp2,1)
>    newdat_imp_2 = complete(imp2,2)
>    newdat_imp_3 = complete(imp2,3)
>    newdat_imp_4 = complete(imp2,4)
>    newdat_imp_5 = complete(imp2,5)
>
>    #Density Plot of Imputed Data
>    par(mfrow=c(2,7))
>    densityplot(imp2)
>
> #Add new variables
>
>    dat = newdat_imp_1
>
>    # Rescale and square nonlinear terms (determined later)
>    dat$stab.glu.2 = ((dat$stab.glu - median(dat$stab.glu))/sd(dat$stab.glu))^2  #center before squaring
>    dat$age.2 =  ((dat$age - median(dat$age))/sd(dat$age))^2
>    dat$bp.1s.2 = ((dat$bp.1s - median(dat$bp.1s))/sd(dat$bp.1s))^2
>
>    # Change glyhb to categorical: glyhb>7 = 1 , glyhb<7 = 0
>
>    for (i in 1:length(dat$glyhb)){
+      if (dat$glyhb[i] >= 7){
+        dat$glyhb.cat[i] = 1
+      } else if (dat$glyhb[i] < 7){
+        dat$glyhb.cat[i] = 0
+      }
+    }
>
>
```

```
>    #calculate waist to hip ratio
>    dat$w.h.ratio = (dat$waist)/dat$hip
>
>    #Add new variable BMI
>    dat$bmi = dat$weight/((dat$height)^2)*703
>
>    # Change "frame" to a factor small, medium, large
>    dat$frame = factor(dat$frame, levels = c("small", "medium", "large"))
>
> ################################
> ######## Data Exploration #####
> ################################
>
>    dat = newdat_imp_1 #first imputed data set
>
> #summary statistics of original and imputed data sets
>
>    summary(dat) #contains imputed values and new values
       id             chol           stab.glu        hdl             ratio          glyhb
 Min.   : 1000   Min.   : 78.0   Min.   : 48.0   Min.   : 12.00   Min.   : 1.50   Min.   : 2.680
 1st Qu.: 4792   1st Qu.:179.0   1st Qu.: 81.0   1st Qu.: 38.00   1st Qu.: 3.20   1st Qu.: 4.385
 Median :15766   Median :204.0   Median : 89.0   Median : 46.00   Median : 4.20   Median : 4.840
 Mean   :15978   Mean   :207.7   Mean   :106.7   Mean   : 50.43   Mean   : 4.52   Mean   : 5.567
 3rd Qu.:20336   3rd Qu.:230.0   3rd Qu.:106.0   3rd Qu.: 59.00   3rd Qu.: 5.40   3rd Qu.: 5.600
 Max.   :41756   Max.   :443.0   Max.   :385.0   Max.   :120.00   Max.   :19.30   Max.   :16.110
  location            age           gender            height         weight         frame
 Length:403         Min.   :19.00   Length:403         Min.   :52.00   Min.   : 99.0   Length:403
 Class :character   1st Qu.:34.00   Class :character   1st Qu.:63.00   1st Qu.:150.5   Class :character
 Mode  :character   Median :45.00   Mode  :character   Median :66.00   Median :172.0   Mode  :character
                    Mean   :46.85                      Mean   :66.02   Mean   :177.5
                    3rd Qu.:60.00                      3rd Qu.:69.00   3rd Qu.:200.0
                    Max.   :92.00                      Max.   :76.00   Max.   :325.0
    bp.1s          bp.1d           waist           hip          time.ppn
 Min.   : 90.0   Min.   : 48.00   Min.   :26.00   Min.   :30.00   Min.   :   5.0
 1st Qu.:121.5   1st Qu.: 75.00   1st Qu.:33.00   1st Qu.:39.00   1st Qu.:  90.0
 Median :136.0   Median : 82.00   Median :37.00   Median :42.00   Median : 240.0
 Mean   :136.9   Mean   : 83.34   Mean   :37.88   Mean   :43.02   Mean   : 341.3
 3rd Qu.:146.0   3rd Qu.: 90.00   3rd Qu.:41.00   3rd Qu.:46.00   3rd Qu.: 525.0
 Max.   :250.0   Max.   :124.00   Max.   :56.00   Max.   :64.00   Max.   :1560.0
>    summary(dat0) #contains NA values
       id             chol           stab.glu        hdl             ratio          glyhb
 Min.   : 1000   Min.   : 78.0   Min.   : 48.0   Min.   : 12.00   Min.   : 1.500   Min.   : 2.68
 1st Qu.: 4792   1st Qu.:179.0   1st Qu.: 81.0   1st Qu.: 38.00   1st Qu.: 3.200   1st Qu.: 4.38
 Median :15766   Median :204.0   Median : 89.0   Median : 46.00   Median : 4.200   Median : 4.84
 Mean   :15978   Mean   :207.8   Mean   :106.7   Mean   : 50.45   Mean   : 4.522   Mean   : 5.59
 3rd Qu.:20336   3rd Qu.:230.0   3rd Qu.:106.0   3rd Qu.: 59.00   3rd Qu.: 5.400   3rd Qu.: 5.60
 Max.   :41756   Max.   :443.0   Max.   :385.0   Max.   :120.00   Max.   :19.300   Max.   :16.11
                 NA's   :1       NA's   :1       NA's   :1        NA's   :13
  location            age           gender            height         weight         frame
 Length:403         Min.   :19.00   Length:403         Min.   :52.00   Min.   : 99.0   Length:403
 Class :character   1st Qu.:34.00   Class :character   1st Qu.:63.00   1st Qu.:151.0   Class :character
 Mode  :character   Median :45.00   Mode  :character   Median :66.00   Median :172.5   Mode  :character
                    Mean   :46.85                      Mean   :66.02   Mean   :177.6
                    3rd Qu.:60.00                      3rd Qu.:69.00   3rd Qu.:200.0
                    Max.   :92.00                      Max.   :76.00   Max.   :325.0
                                                       NA's   :5       NA's   :1
    bp.1s          bp.1d           bp.2s           bp.2d           waist          hip          time.ppn
 Min.   : 90.0   Min.   : 48.00   Min.   :110.0   Min.   : 60.00   Min.   :26.0   Min.   :30.00   Min.   :   5.0
 1st Qu.:121.2   1st Qu.: 75.00   1st Qu.:138.0   1st Qu.: 84.00   1st Qu.:33.0   1st Qu.:39.00   1st Qu.:  90.0
 Median :136.0   Median : 82.00   Median :149.0   Median : 92.00   Median :37.0   Median :42.00   Median : 240.0
 Mean   :136.9   Mean   : 83.32   Mean   :152.4   Mean   : 92.52   Mean   :37.9   Mean   :43.04   Mean   : 341.2
 3rd Qu.:146.8   3rd Qu.: 90.00   3rd Qu.:161.0   3rd Qu.:100.00   3rd Qu.:41.0   3rd Qu.:46.00   3rd Qu.: 517.5
 Max.   :250.0   Max.   :124.00   Max.   :238.0   Max.   :124.00   Max.   :56.0   Max.   :64.00   Max.   :1560.0
 NA's   :5       NA's   :5        NA's   :262     NA's   :262      NA's   :2      NA's   :2       NA's   :3
>
> #Linearity - restricted cubic splines for continuous variables.
```

```
>
>    fit_chol = lrm(glyhb.cat ~ rcs(chol,3), x=T, y=T, data =dat)
Error in eval(predvars, data, callenv) : object 'glyhb.cat' not found
>    anova(fit_chol)  #linear p-value = 0.9129
Error in anova(fit_chol) : object 'fit_chol' not found
>    plot(dat$chol,dat$glyhb, main = "Chol")
Error in plot.new() : figure margins too large
>
>    fit_stab.glu = lrm(glyhb.cat ~ rcs(stab.glu,3), x=T, y=T, data =dat)
Error in eval(predvars, data, callenv) : object 'glyhb.cat' not found
>    anova(fit_stab.glu)  #NONLINEAR p-value = 0.0014
Error in anova(fit_stab.glu) : object 'fit_stab.glu' not found
>    plot(dat$stab.glu,dat$glyhb, main = "Stab. Glu")
Error in plot.new() : figure margins too large
>
>    fit_hdl = lrm(glyhb.cat ~ rcs(hdl,3), x=T, y=T, data =dat)
Error in eval(predvars, data, callenv) : object 'glyhb.cat' not found
>    anova(fit_hdl)  #LINEAR or NONLINEAR p-value = 0.0502
Error in anova(fit_hdl) : object 'fit_hdl' not found
>    plot(dat$hdl,dat$glyhb, main = "HDL")
Error in plot.new() : figure margins too large
>
>    fit_ratio = lrm(glyhb.cat ~ rcs(ratio,3), x=T, y=T, data =dat)
Error in eval(predvars, data, callenv) : object 'glyhb.cat' not found
>    anova(fit_ratio)  #LINEAR p-value = 0.5695
Error in anova(fit_ratio) : object 'fit_ratio' not found
>    plot(dat$ratio,dat$glyhb, main = "ratio")
Error in plot.new() : figure margins too large
>
>    fit_age = lrm(glyhb.cat ~ rcs(age,3), x=T, y=T, data =dat)
Error in eval(predvars, data, callenv) : object 'glyhb.cat' not found
>    anova(fit_age)  #NONLINEAR p-value = 0.0166
Error in anova(fit_age) : object 'fit_age' not found
>    plot(dat$age,dat$glyhb, main = "Age")
Error in plot.new() : figure margins too large
>
>    fit_height = lrm(glyhb.cat ~ rcs(height,3), x=T, y=T, data =dat)
Error in eval(predvars, data, callenv) : object 'glyhb.cat' not found
>    anova(fit_height)  #LINEAR p-value = 0.9470
Error in anova(fit_height) : object 'fit_height' not found
>    plot(dat$height,dat$glyhb, main = "Height")
Error in plot.new() : figure margins too large
>
>    fit_weight = lrm(glyhb.cat ~ rcs(weight,3), x=T, y=T, data =dat)
Error in eval(predvars, data, callenv) : object 'glyhb.cat' not found
>    anova(fit_weight)  #LINEAR or NONLINEAR p-value = 0.1086
Error in anova(fit_weight) : object 'fit_weight' not found
>    plot(dat$weight,dat$glyhb, main = "Weight")
Error in plot.new() : figure margins too large
>
>    fit_bp.1s = lrm(glyhb.cat ~ rcs(bp.1s,3), x=T, y=T, data =dat)
Error in eval(predvars, data, callenv) : object 'glyhb.cat' not found
>    anova(fit_bp.1s)  #NONLINEAR p-value = 0.0182
Error in anova(fit_bp.1s) : object 'fit_bp.1s' not found
>    plot(dat$bp.1s,dat$glyhb, main = "Systolic BP")
Error in plot.new() : figure margins too large
>
>    fit_bp.1d = lrm(glyhb.cat ~ rcs(bp.1d,3), x=T, y=T, data =dat)
Error in eval(predvars, data, callenv) : object 'glyhb.cat' not found
>    anova(fit_bp.1d)  #LINEAR p-value = 0.5562
Error in anova(fit_bp.1d) : object 'fit_bp.1d' not found
>    plot(dat$bp.1d,dat$glyhb, main = "Diastolic BP")
Error in plot.new() : figure margins too large
>
>    fit_waist = lrm(glyhb.cat ~ rcs(waist,3), x=T, y=T, data =dat)
Error in eval(predvars, data, callenv) : object 'glyhb.cat' not found
```

```
>    anova(fit_waist)  #LINEAR p-value = 0.1611
Error in anova(fit_waist) : object 'fit_waist' not found
>    plot(dat$waist,dat$glyhb, mian = "Waist")
Error in plot.new() : figure margins too large
>
>    fit_hip = lrm(glyhb.cat ~ rcs(hip,3), x=T, y=T, data =dat)
Error in eval(predvars, data, callenv) : object 'glyhb.cat' not found
>    anova(fit_hip)  #LINEAR or NONLINEAR p-value = 0.0846
Error in anova(fit_hip) : object 'fit_hip' not found
>    plot(dat$hip,dat$glyhb, main = "Hip")
Error in plot.new() : figure margins too large
>
>    fit_time.ppn = lrm(glyhb ~ rcs(time.ppn,3), x=T, y=T, data =dat)
>    anova(fit_time.ppn)  #LINEAR p-value = 0.9678
             Wald Statistics          Response: glyhb

 Factor     Chi-Square d.f. P
 time.ppn   0.52       2    0.7729
  Nonlinear 0.51       1    0.4772
 TOTAL      0.52       2    0.7729
>    plot(dat$time.ppn,dat$glyhb, main = "Time.ppn")
Error in plot.new() : figure margins too large
>
>    fit_bmi = lrm(glyhb ~ rcs(bmi,3), x=T, y=T, data =dat)
Error in rcs(bmi, 3) : object 'bmi' not found
>    anova(fit_bmi)  #LINEAR or NONLINEAR p-value = 0.0706
Error in anova(fit_bmi) : object 'fit_bmi' not found
>    plot(dat$bmi,dat$glyhb, main = "BMI")
Error in xy.coords(x, y, xlabel, ylabel, log) :
  'x' and 'y' lengths differ
>
> #Scatterplots
>
>    x1 = dat[ ,c("glyhb", "chol", "stab.glu", "stab.glu.2", "hdl", "ratio", "age",
+              "bp.1s", "bmi", "age.2", "w.h.ratio" )]
Error in `[.data.frame`(dat, , c("glyhb", "chol", "stab.glu", "stab.glu.2",  :
  undefined columns selected
>    pairs(x1)
Error in pairs(x1) : object 'x1' not found
>
>    par(mfrow= c(2,4))
>    plot(dat$chol,dat$glyhb, main = "Chol")
>    plot(dat$stab.glu,dat$glyhb, main = "Stab. Glu")
>    plot(dat$hdl,dat$glyhb, main = "HDL")
>    plot(dat$ratio,dat$glyhb, main = "ratio")
>    plot(dat$age,dat$glyhb, main = "Age")
>    plot(dat$bp.1s,dat$glyhb, main = "Systolic BP")
>    plot(dat$bmi,dat$glyhb, main = "BMI")
Error in xy.coords(x, y, xlabel, ylabel, log) :
  'x' and 'y' lengths differ
>
> ########################################################
> ######## GLMM W/ and W/O Random Location and Restr. Cubic Splines
> ########################################################
>
> #Fit glmer with random effect location
>
>    #Full Model
>    fit_1 = glmer(glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + ratio + age + rcs(age, 3) + height
+               + weight + frame + bp.1s + rcs(bp.1s, 3) + bp.1d + waist*hip + gender*bmi + w.h.ratio
+               + (1 | location), data = dat, family = binomial,
+               control = glmerControl(optimizer = "bobyqa"),
+               nAGQ = 10)
Error in eval(predvars, data, env) : object 'glyhb.cat' not found
>    sumfit_1 = summary(fit_1);
>    sumfit_1
```

```
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature, nAGQ = 10) [glmerMod]
 Family: binomial  ( logit )
Formula: glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + ratio +
    age + rcs(age, 3) + height + weight + frame + bp.1s + rcs(bp.1s,
    3) + bp.1d + waist * hip + gender * bmi + w.h.ratio + (1 |      location)
   Data: dat
Control: glmerControl(optimizer = "bobyqa")

     AIC      BIC   logLik deviance df.resid
   193.8    285.1    -73.9    147.8      368

Scaled residuals:
    Min      1Q  Median      3Q     Max
-12.766  -0.186  -0.095  -0.030  55.409

Random effects:
 Groups   Name        Variance Std.Dev.
 location (Intercept) 0        0
Number of obs: 391, groups:  location, 2

Fixed effects:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            17.378312  34.788031   0.500   0.6174
chol                    0.013604   0.009213   1.477   0.1398
stab.glu                0.112609   0.031808   3.540   0.0004 ***
rcs(stab.glu, 3)stab.glu'  -0.202096   0.081004  -2.495   0.0126 *
hdl                    -0.050608   0.034526  -1.466   0.1427
ratio                  -0.274750   0.298238  -0.921   0.3569
age                     0.085484   0.060292   1.418   0.1562
rcs(age, 3)age'          -0.066389   0.063936  -1.038   0.2991
height                 -0.353933   0.396287  -0.893   0.3718
weight                  0.052933   0.066866   0.792   0.4286
framemedium             -0.411135   0.651062  -0.631   0.5277
framelarge              -0.835901   0.815884  -1.025   0.3056
bp.1s                   0.041050   0.035885   1.144   0.2526
rcs(bp.1s, 3)bp.1s'       -0.028929   0.031843  -0.908   0.3636
bp.1d                   0.002631   0.022976   0.115   0.9088
waist                   0.678614   0.665461   1.020   0.3078
hip                    -0.486744   0.577057  -0.843   0.3990
gendermale              5.497504   3.646371   1.508   0.1316
bmi                    -0.241231   0.373216  -0.646   0.5180
w.h.ratio              -23.487465  28.590671  -0.822   0.4114
waist:hip               -0.001260   0.003418  -0.369   0.7124
gendermale:bmi          -0.194605   0.122699  -1.586   0.1127
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation matrix not shown by default, as p = 22 > 12.
Use print(x, correlation=TRUE)  or
    vcov(x)       if you need it

fit warnings:
fixed-effect model matrix is rank deficient so dropping 3 columns / coefficients
Some predictor variables are on very different scales: consider rescaling
convergence code: 0
boundary (singular) fit: see ?isSingular

>
>    #remove bp.1d
>    fit_2 = glmer(glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + ratio + age + rcs(age, 3) + height
+               + weight + frame + bp.1s + rcs(bp.1s, 3) + waist*hip + gender*bmi + w.h.ratio
+               + (1 | location), data = dat, family = binomial,
+               control = glmerControl(optimizer = "bobyqa"),
+               nAGQ = 10)
Error in eval(predvars, data, env) : object 'glyhb.cat' not found
>    sumfit_2 = summary(fit_2);
```

```
Error in h(simpleError(msg, call)) :
  error in evaluating the argument 'object' in selecting a method for function 'summary': object 'fit_2' not found
>    sumfit_2
Error: object 'sumfit_2' not found
>
>    #remove frame
>    fit_3 = glmer(glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + ratio + age + rcs(age, 3) + height
+              + weight + bp.1s + rcs(bp.1s, 3) + waist*hip + gender*bmi + w.h.ratio
+              + (1 | location), data = dat, family = binomial,
+              control = glmerControl(optimizer = "bobyqa"),
+              nAGQ = 10)
Error in eval(predvars, data, env) : object 'glyhb.cat' not found
>    sumfit_3 = summary(fit_3);
Error in h(simpleError(msg, call)) :
  error in evaluating the argument 'object' in selecting a method for function 'summary': object 'fit_3' not found
>    sumfit_3
Error: object 'sumfit_3' not found
>
>    #remove waist*hip
>    fit_4 = glmer(glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + ratio + age + rcs(age, 3) + height
+              + weight + bp.1s + rcs(bp.1s, 3) + gender*bmi + w.h.ratio
+              + (1 | location), data = dat, family = binomial,
+              control = glmerControl(optimizer = "bobyqa"),
+              nAGQ = 10)
Error in eval(predvars, data, env) : object 'glyhb.cat' not found
>    sumfit_4 = summary(fit_4);
Error in h(simpleError(msg, call)) :
  error in evaluating the argument 'object' in selecting a method for function 'summary': object 'fit_4' not found
>    sumfit_4
Error: object 'sumfit_4' not found
>
>    #remove waist*hip
>    fit_4 = glmer(glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + ratio + age + rcs(age, 3) + height
+              + weight + bp.1s + rcs(bp.1s, 3) + gender*bmi + w.h.ratio
+              + (1 | location), data = dat, family = binomial,
+              control = glmerControl(optimizer = "bobyqa"),
+              nAGQ = 10)
Error in eval(predvars, data, env) : object 'glyhb.cat' not found
>    sumfit_4 = summary(fit_4);
Error in h(simpleError(msg, call)) :
  error in evaluating the argument 'object' in selecting a method for function 'summary': object 'fit_4' not found
>    sumfit_4
Error: object 'sumfit_4' not found
>
>    #remove weight
>    fit_5 = glmer(glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + ratio + age + rcs(age, 3) + height
+              + bp.1s + rcs(bp.1s, 3) + gender*bmi + w.h.ratio
+              + (1 | location), data = dat, family = binomial,
+              control = glmerControl(optimizer = "bobyqa"),
+              nAGQ = 10)
Error in eval(predvars, data, env) : object 'glyhb.cat' not found
>    sumfit_5 = summary(fit_5);
Error in h(simpleError(msg, call)) :
  error in evaluating the argument 'object' in selecting a method for function 'summary': object 'fit_5' not found
>    sumfit_5
Error: object 'sumfit_5' not found
>
>    #remove ratio
>    fit_6 = glmer(glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + age + rcs(age, 3) + height
+              + bp.1s + rcs(bp.1s, 3) + gender*bmi + w.h.ratio
+              + (1 | location), data = dat, family = binomial,
+              control = glmerControl(optimizer = "bobyqa"),
+              nAGQ = 10)
Error in eval(predvars, data, env) : object 'glyhb.cat' not found
>    sumfit_6 = summary(fit_6);
Error in h(simpleError(msg, call)) :
```

```
 error in evaluating the argument 'object' in selecting a method for function 'summary': object 'fit_6' not found
>    sumfit_6
Error: object 'sumfit_6' not found
>
>    #remove rcs(bp.1s, 3)
>    fit_7 = glmer(glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + age + rcs(age, 3) + height
+              + bp.1s + gender*bmi + w.h.ratio
+              + (1 | location), data = dat, family = binomial,
+              control = glmerControl(optimizer = "bobyqa"),
+              nAGQ = 10)
Error in eval(predvars, data, env) : object 'glyhb.cat' not found
>    sumfit_7 = summary(fit_7);
Error in h(simpleError(msg, call)) :
 error in evaluating the argument 'object' in selecting a method for function 'summary': object 'fit_7' not found
>    sumfit_7
Error: object 'sumfit_7' not found
>
>    #remove rcs(age, 3)
>    fit_8 = glmer(glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + age + height
+              + bp.1s + gender*bmi + w.h.ratio
+              + (1 | location), data = dat, family = binomial,
+              control = glmerControl(optimizer = "bobyqa"),
+              nAGQ = 10)
Error in eval(predvars, data, env) : object 'glyhb.cat' not found
>    sumfit_8 = summary(fit_8);
>    sumfit_8
Generalized linear mixed model fit by maximum likelihood (Adaptive Gauss-Hermite Quadrature, nAGQ = 10) [glmerMod]
 Family: binomial  ( logit )
Formula: glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + age +
   height + bp.1s + gender * bmi + w.h.ratio + (1 | location)
  Data: dat
Control: glmerControl(optimizer = "bobyqa")

    AIC      BIC   logLik deviance df.resid
  181.3    233.2    -77.6    155.3      390

Scaled residuals:
   Min     1Q Median     3Q    Max
-7.308 -0.200 -0.102 -0.043 49.821

Random effects:
 Groups   Name        Variance Std.Dev.
 location (Intercept) 0        0
Number of obs: 403, groups:  location, 2

Fixed effects:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -20.682689   7.486508  -2.763 0.005733 **
chol                   0.007519   0.005134   1.465 0.143041
stab.glu               0.111573   0.030647   3.641 0.000272 ***
rcs(stab.glu, 3)stab.glu'  -0.200547   0.078178  -2.565 0.010310 *
hdl                   -0.016004   0.013897  -1.152 0.249503
age                    0.024570   0.016441   1.494 0.135070
height                -0.012205   0.087677  -0.139 0.889287
bp.1s                  0.012636   0.009569   1.321 0.186632
gendermale             2.308531   2.212538   1.043 0.296770
bmi                    0.042620   0.036265   1.175 0.239899
w.h.ratio              3.498648   3.177621   1.101 0.270885
gendermale:bmi        -0.098308   0.074387  -1.322 0.186307
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
          (Intr) chol   stb.gl r(.,3) hdl    age    height bp.1s  gndrml bmi    w.h.rt
chol       -0.109
stab.glu   -0.533  0.010
```

```
rcs(s.,3).'  0.541  0.009 -0.988
hdl        -0.135 -0.165  0.023 -0.027
age         -0.178 -0.103 -0.074  0.054 -0.089
height      -0.785  0.005  0.131 -0.164  0.089  0.278
bp.1s       -0.103 -0.039  0.129 -0.096 -0.049 -0.362 -0.092
gendermale  -0.007  0.049 -0.009  0.047 -0.036 -0.088 -0.214  0.129
bmi         -0.275 -0.034 -0.065  0.068  0.148  0.242  0.178 -0.079  0.457
w.h.ratio   -0.249 -0.031  0.073 -0.067 -0.035 -0.175 -0.154 -0.013  0.188 -0.043
genderml:bm  0.176 -0.031 -0.037  0.000  0.097  0.048  0.023 -0.120 -0.952 -0.430 -0.248
fit warnings:
fixed-effect model matrix is rank deficient so dropping 1 column / coefficient
convergence code: 0
boundary (singular) fit: see ?isSingular

>
>    #likelihood ratio test for full and reduced GLMM data sets
>    res_dev_glmm_loc_reduced = 155.3
>    res_dev_glmm_full =
+    df_glmm_loc_reduced = 390
>    df_glmm_loc_full =
+    AIC_glmm_loc_reduced = 181.3
>    AIC_glmm_loc_full =
+    pchisq(res_dev_glmm_loc_reduced - res_dev_glmm_full, df_glmm_loc_reduced - df_glmm_loc_full,
+           lower.tail = F) #p-value is 0.990 - therefore can use either
> ######################################
> ######## Multiple Logistic Regression Assessment
> ######################################
>
> #Compare to glmm random location with likelihood ratio test
> #Restricted cubic splines were used to model terms identified as non-linear
>
>    md.pattern(dat)  #tells us number of observations with that missing data structure
 /\     /\
{  `---'  }
{  O   O  }
==>  V <==  No need for mice. This data set is completely observed.
 \  \|/  /
  `-----'

   id chol stab.glu hdl ratio glyhb location age gender height weight frame bp.1s bp.1d waist hip time.ppn
403  1    1       1  1     1     1       1  1     1     1     1     1     1     1     1  1       1  0
     0    0       0  0     0     0       0  0     0     0     0     0     0     0     0  0       0  0
>    agg_plot = aggr(dat, combined = F) #histogram showing proportion of missing data by variable
>
>    #full model - removed frame, rcs(bp.1s, 3), rcs(stab.glu, 3), rcs(age, 3), w.h.ratio
>    fit_01 = glm(glyhb.cat ~ chol + stab.glu + hdl + ratio + age +  height
+                 + weight + bp.1s + bp.1d + waist*hip + gender*bmi,
+                 data = dat)
Error in eval(predvars, data, env) : object 'glyhb.cat' not found
>    sumfit_01 = summary(fit_01)
>    sumfit_01

Call:
glm(formula = glyhb.cat ~ chol + stab.glu + hdl + ratio + age +
    height + weight + bp.1s + bp.1d + waist * hip + gender *
    bmi, data = dat)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.23397  -0.11029  -0.03622   0.03830   1.11588

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.261e-02  1.367e+00  -0.009    0.993
chol           3.797e-04  5.631e-04   0.674    0.501
stab.glu       4.301e-03  2.635e-04  16.323   <2e-16 ***
```

```
hdl          -6.457e-06  1.758e-03  -0.004    0.997
ratio         1.183e-02  1.960e-02   0.603    0.547
age           1.382e-03  1.016e-03   1.360    0.175
height       -3.208e-04  2.050e-02  -0.016    0.988
weight       -7.739e-06  3.775e-03  -0.002    0.998
bp.1s         7.611e-04  8.254e-04   0.922    0.357
bp.1d        -1.967e-04  1.264e-03  -0.156    0.876
waist        -9.012e-03  1.540e-02  -0.585    0.559
hip          -1.991e-02  1.493e-02  -1.334    0.183
gendermale    9.711e-02  1.767e-01   0.549    0.583
bmi           2.904e-03  2.144e-02   0.135    0.892
waist:hip     3.462e-04  3.361e-04   1.030    0.304
gendermale:bmi -5.260e-03  6.227e-03  -0.845    0.399
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.06532482)

    Null deviance: 51.067  on 402  degrees of freedom
Residual deviance: 25.281  on 387  degrees of freedom
AIC: 61.8

Number of Fisher Scoring iterations: 2

>
>    #remove bp.1d
>    fit_02 = glm(glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + ratio + age + rcs(age, 3) + height
+              + weight + frame + bp.1s + rcs(bp.1s, 3)  + waist*hip + gender*bmi + w.h.ratio,
+              data = dat, family = binomial() )
Error in eval(predvars, data, env) : object 'glyhb.cat' not found
>    sumfit_02 = summary(fit_02)
Error in h(simpleError(msg, call)) :
  error in evaluating the argument 'object' in selecting a method for function 'summary': object 'fit_02' not found
>    sumfit_02
Error: object 'sumfit_02' not found
>
>    #remove waist*hip
>    fit_02 = glm(glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + ratio + age + rcs(age, 3) + height
+              + weight + frame + bp.1s + rcs(bp.1s, 3) + gender*bmi + w.h.ratio,
+              data = dat, family = binomial() )
Error in eval(predvars, data, env) : object 'glyhb.cat' not found
>    sumfit_02 = summary(fit_02)
Error in h(simpleError(msg, call)) :
  error in evaluating the argument 'object' in selecting a method for function 'summary': object 'fit_02' not found
>    sumfit_02
Error: object 'sumfit_02' not found
>
>    #remove rcs(bp.1s, 3)
>    fit_03 = glm(glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + ratio + age + rcs(age, 3) + height
+              + weight + frame + bp.1s + gender*bmi + w.h.ratio,
+              data = dat, family = binomial() )
Error in eval(predvars, data, env) : object 'glyhb.cat' not found
>    sumfit_03 = summary(fit_03)
Error in h(simpleError(msg, call)) :
  error in evaluating the argument 'object' in selecting a method for function 'summary': object 'fit_03' not found
>    sumfit_03
Error: object 'sumfit_03' not found
>
>    #remove frame
>    fit_04 = glm(glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + ratio + age + rcs(age, 3) + height
+              + weight + bp.1s + gender*bmi + w.h.ratio,
+              data = dat, family = binomial() )
Error in eval(predvars, data, env) : object 'glyhb.cat' not found
>    sumfit_04 = summary(fit_04)
Error in h(simpleError(msg, call)) :
  error in evaluating the argument 'object' in selecting a method for function 'summary': object 'fit_04' not found
```

```
>    sumfit_04
Error: object 'sumfit_04' not found
>
>    #remove weight
>    fit_05 = glm(glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + ratio + age + rcs(age, 3) + height
+               + bp.1s + gender*bmi + w.h.ratio,
+               data = dat, family = binomial() )
Error in eval(predvars, data, env) : object 'glyhb.cat' not found
>    sumfit_05 = summary(fit_05)
Error in h(simpleError(msg, call)) :
  error in evaluating the argument 'object' in selecting a method for function 'summary': object 'fit_05' not found
>    sumfit_05
Error: object 'sumfit_05' not found
>
>    #remove ratio
>    fit_06 = glm(glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + age + rcs(age, 3) + height
+              + bp.1s + gender*bmi + w.h.ratio,
+              data = dat, family = binomial() )
Error in eval(predvars, data, env) : object 'glyhb.cat' not found
>    sumfit_06 = summary(fit_06)
Error in h(simpleError(msg, call)) :
  error in evaluating the argument 'object' in selecting a method for function 'summary': object 'fit_06' not found
>    sumfit_06
Error: object 'sumfit_06' not found
>
>    #remove rcs(age, 3)
>    fit_07 = glm(glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + age + height
+              + bp.1s + gender*bmi + w.h.ratio,
+              data = dat, family = binomial() )
Error in eval(predvars, data, env) : object 'glyhb.cat' not found
>    sumfit_07 = summary(fit_07)
Error in h(simpleError(msg, call)) :
  error in evaluating the argument 'object' in selecting a method for function 'summary': object 'fit_07' not found
>    sumfit_07
Error: object 'sumfit_07' not found
>
>    #looks good, try removing height (other research says height is important)
>
>    #remove height
>    fit_08 = glm(glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + age
+              + bp.1s + gender*bmi + w.h.ratio,
+              data = dat, family = binomial() )
Error in eval(predvars, data, env) : object 'glyhb.cat' not found
>    sumfit_08 = summary(fit_08)
>    sumfit_08

Call:
glm(formula = glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) +
   hdl + age + bp.1s + gender * bmi + w.h.ratio, family = binomial(),
   data = dat)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.8046  -0.2787  -0.1431  -0.0596   3.9586

Coefficients: (1 not defined because of singularities)
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -21.503234   4.638557  -4.636 3.56e-06 ***
chol                   0.007525   0.005134   1.466 0.142771
stab.glu               0.112142   0.030368   3.693 0.000222 ***
rcs(stab.glu, 3)stab.glu        NA        NA      NA       NA
rcs(stab.glu, 3)stab.glu'  -0.202352   0.077033  -2.627 0.008619 **
hdl                   -0.015832   0.013843  -1.144 0.252746
age                    0.025209   0.015811   1.594 0.110842
bp.1s                  0.012514   0.009540   1.312 0.189598
gendermale             2.242253   2.159681   1.038 0.299161
```

```
bmi                    0.043515   0.035698   1.219 0.222863
w.h.ratio              3.431615   3.139357   1.093 0.274352
gendermale:bmi        -0.098068   0.074321  -1.320 0.186996
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 339.14  on 402  degrees of freedom
Residual deviance: 155.28  on 392  degrees of freedom
AIC: 177.28

Number of Fisher Scoring iterations: 7

>
>    confint(fit_08)
Waiting for profiling to be done...
                           2.5 %       97.5 %
(Intercept)           -31.285765012 -12.96193827
chol                  -0.001990458   0.01814252
stab.glu               0.056363399   0.17618307
rcs(stab.glu, 3)stab.glu          NA          NA
rcs(stab.glu, 3)stab.glu'  -0.360606016  -0.05666169
hdl                   -0.043772938   0.01059884
age                   -0.005540259   0.05684134
bp.1s                 -0.006800703   0.03093685
gendermale            -2.047247704   6.48225930
bmi                   -0.027951454   0.11366166
w.h.ratio             -2.728663414   9.64719862
gendermale:bmi        -0.245265050   0.04803552
>
>    #likelihood ratio test to compare GLM full and reduced Mult Log Regression
>    res_dev_log_full = 147.84
>    res_dev_log_red = 155.28
>    df_log_full = 369
>    df_log_red = 392
>    AIC_log_full = 191.84
>    AIC_log_red = 177.28
>    pchisq(res_dev_log_red - res_dev_log_full, df_log_red - df_log_full,
+         lower.tail = F) #p-value is 0.999 - therefore can use reduced
[1] 0.9990926
>
>
>    #likelihood ratio test for GLMM random location and Mult Log. Regression
>    res_dev_glmm_loc = 155.3
>    res_dev_logistic = 155.28
>    df_glmm_loc = 390
>    df_logistic = 392
>    AIC_glmm_loc = 181.3
>    AIC_logistic = 177.28
>    pchisq(res_dev_glmm_loc - res_dev_logistic, df_glmm_loc - df_logistic,
+         lower.tail = F) #p-value is 0.990 - therefore can use either
[1] NaN
Warning message:
In pchisq(res_dev_glmm_loc - res_dev_logistic, df_glmm_loc - df_logistic,  :
  NaNs produced
>
> ###############################################################################
> ##########   Assessment of Reduced Multiple Logistic Regression Model #####
> ###############################################################################
>
>
>    #fit a logistic regression model using terms from glm (which gave me p values)
>    require(rms)
>    dd = datadist(dat)
>    options(datadist = 'dd')
```

```
>    lrm.reduced = lrm(glyhb.cat ~ chol + stab.glu + hdl + age
+                 + bp.1s + gender*bmi, data = dat)
Error in eval(predvars, data, callenv) : object 'glyhb.cat' not found
>
>
>    #Calibration using bootstraps and Calibration Plot - DOES NOT WORK (can't use glm)
>    cal_log = calibrate(lrm.reduced, B=100) #does not work with glm or lrm models
Error in (function (classes, fdef, mtable)  :
  unable to find an inherited method for function ‘calibrate’ for signature ‘"lrm"’
>    boot_strap = boot(dat,
+                 predict(fit_08),
+                 R = 1000,
+                 cor.type = 's')
Error in statistic(data, original, ...) :
  could not find function "statistic"
>    boots = bootstrap(dat, 100)
>    probs = predict(fit_08, newdata = dat[boots[1,],])
Error in xj[i] : invalid subscript type 'list'
>    predict_matrix = data.matrix(dat[,])
>
>    #discrimination - ROC Curve
>    library(pROC)
>    pred = predict(fit_08, type = c("response"))
>    roccurve = roc(dat$glyhb.cat ~ pred)
Error in model.frame.default(formula = dat$glyhb.cat ~ pred, na.action = "na.pass") :
  invalid type (NULL) for variable 'dat$glyhb.cat'
>    plot(roccurve, main = "ROC Curve")
>    auc(roccurve) #Area Under Curve = 0.9419 - good discrimination
Area under the curve: 0.9419
>
>    #validation
>    library(purrr) #for map()
>    validate(lrm.reduced, method = "boot", B= 1000, data = dat, x = TRUE, y = TRUE)
Error in validate.lrm(lrm.reduced, method = "boot", B = 1000, data = dat,  :
  fit did not use x=TRUE,y=TRUE
>    require(lme4)
>    require(languageR)
>    somers.mer(fit_08)
Error in somers.mer(fit_08) : could not find function "somers.mer"
>
>    x = bootstrap(dat, 100)
>    map(x, lrm.reduced)
Error: Can't convert a `lrm/rms/glm` object to function
Run `rlang::last_error()` to see where the error occurred.
>    somers2()
Error in somers2() : argument "y" is missing, with no default
>
>
> ################################################
> ############# Assess Model GLM #############
> ################################################
>
>
>    #Residuals
>    res = residuals(fit_08, "pearson")
>    hist(res)
>    plot(res) #resid vs index
>    min(res)
[1] -7.075276
>    max(res)
[1] 50.27119
>    which(res>25) #observation 334 has extremely large residual comparatively
334
334
>
>    #try removing observation 334 - Does not improve fit
```

```
>    fit_09 = glm(glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + age
+              + bp.1s + gender*bmi + w.h.ratio,
+              data = dat[-334,], family = binomial() )
Error in eval(predvars, data, env) : object 'glyhb.cat' not found
>    sumfit_09 = summary(fit_09)
>    sumfit_09

Call:
glm(formula = glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) +
   hdl + age + bp.1s + gender * bmi + w.h.ratio, family = binomial(),
   data = dat[-334, ])

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.5075  -0.2291  -0.0902  -0.0305   3.5245

Coefficients: (1 not defined because of singularities)
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -26.657327   5.515142  -4.833 1.34e-06 ***
chol                   0.007229   0.005375   1.345 0.178605
stab.glu               0.155242   0.036976   4.198 2.69e-05 ***
rcs(stab.glu, 3)stab.glu        NA        NA      NA       NA
rcs(stab.glu, 3)stab.glu'  -0.306901   0.090449  -3.393 0.000691 ***
hdl                   -0.021562   0.015111  -1.427 0.153600
age                    0.032681   0.016989   1.924 0.054393 .
bp.1s                  0.013644   0.010363   1.317 0.187973
gendermale             1.580012   2.285483   0.691 0.489361
bmi                    0.043526   0.036946   1.178 0.238761
w.h.ratio              4.159627   3.345645   1.243 0.213759
gendermale:bmi        -0.084494   0.078259  -1.080 0.280293
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 335.31  on 401  degrees of freedom
Residual deviance: 137.40  on 391  degrees of freedom
AIC: 159.4

Number of Fisher Scoring iterations: 7

>    plot(fit_09)
>
>    #Overdispersion - no consensus on how to calc, cite source bookmarked
>    r_df = df.residual(fit_08)
>    r_pearson = residuals(fit_08, "pearson")
>    test_stat_over = sum(r_pearson^2)
>    pearson_ratio = test_stat_over/r_df #indicates possibility of underdispersion. usually
>    p = pchisq(test_stat_over, df = r_df, lower.tail = F)
>    p #p=0 reject null, dispersion is significantly different from zero
[1] 0
>
>
> ############################################################
> ############# Fit Remaining Imputed Data Sets #############
> ############################################################
> #All models were ran on first imputed data set.
> #Show that imputed data sets are not significantly different due to small
> #percentage of missing values.
>
>    #Final Fit for remaining imputed data sets
>
>    #assume same final features as calculated from 1st imputed data set
>    dat2 = complete(imp2,2)
>    dat3 = complete(imp2,3)
>    dat4 = complete(imp2,4)
```

```
>    dat5 = complete(imp2,5)
>
> #Add new variables to data sets
>    #dat2
>    dat2$stab.glu.2 = ((dat2$stab.glu - median(dat2$stab.glu))/sd(dat2$stab.glu))^2
>    dat2$age.2 =  ((dat2$age - median(dat2$age))/sd(dat2$age))^2
>    dat2$bp.1s.2 = ((dat2$bp.1s - median(dat2$bp.1s))/sd(dat2$bp.1s))^2
>    for (i in 1:length(dat2$glyhb)){ # Change glyhb to categorical: glyhb>7 = 1 , glyhb<7 = 0
+      if (dat2$glyhb[i] >= 7){
+        dat2$glyhb.cat[i] = 1
+      } else if (dat$glyhb[i] < 7){
+        dat2$glyhb.cat[i] = 0
+      }
+    }
>    dat2$w.h.ratio = (dat2$waist)/dat2$hip #waist-hip ratio
>    dat2$bmi = dat2$weight/((dat2$height)^2)*703 #Add new variable BMI
>    dat2$frame = factor(dat2$frame, levels = c("small", "medium", "large"))  # Change "frame" to a factor small, medium, large
>
>    #dat3
>    dat3$stab.glu.2 = ((dat3$stab.glu - median(dat3$stab.glu))/sd(dat3$stab.glu))^2
>    dat3$age.2 =  ((dat3$age - median(dat3$age))/sd(dat3$age))^2
>    dat3$bp.1s.2 = ((dat3$bp.1s - median(dat3$bp.1s))/sd(dat3$bp.1s))^2
>    for (i in 1:length(dat3$glyhb)){ # Change glyhb to categorical: glyhb>7 = 1 , glyhb<7 = 0
+      if (dat3$glyhb[i] >= 7){
+        dat3$glyhb.cat[i] = 1
+      } else if (dat$glyhb[i] < 7){
+        dat3$glyhb.cat[i] = 0
+      }
+    }
>    dat3$w.h.ratio = (dat3$waist)/dat3$hip #waist-hip ratio
>    dat3$bmi = dat3$weight/((dat3$height)^2)*703 #Add new variable BMI
>    dat3$frame = factor(dat3$frame, levels = c("small", "medium", "large"))  # Change "frame" to a factor small, medium, large
>
>    #dat4
>    dat4$stab.glu.2 = ((dat4$stab.glu - median(dat4$stab.glu))/sd(dat4$stab.glu))^2
>    dat4$age.2 =  ((dat4$age - median(dat4$age))/sd(dat4$age))^2
>    dat4$bp.1s.2 = ((dat4$bp.1s - median(dat4$bp.1s))/sd(dat4$bp.1s))^2
>    for (i in 1:length(dat4$glyhb)){ # Change glyhb to categorical: glyhb>7 = 1 , glyhb<7 = 0
+      if (dat4$glyhb[i] >= 7){
+        dat4$glyhb.cat[i] = 1
+      } else if (dat$glyhb[i] < 7){
+        dat4$glyhb.cat[i] = 0
+      }
+    }
>    dat4$w.h.ratio = (dat4$waist)/dat4$hip #waist-hip ratio
>    dat4$bmi = dat4$weight/((dat4$height)^2)*703 #Add new variable BMI
>    dat4$frame = factor(dat4$frame, levels = c("small", "medium", "large"))  # Change "frame" to a factor small, medium, large
>
>    #dat5
>    dat5$stab.glu.2 = ((dat5$stab.glu - median(dat5$stab.glu))/sd(dat5$stab.glu))^2
>    dat5$age.2 =  ((dat5$age - median(dat5$age))/sd(dat5$age))^2
>    dat5$bp.1s.2 = ((dat5$bp.1s - median(dat5$bp.1s))/sd(dat5$bp.1s))^2
>    for (i in 1:length(dat5$glyhb)){ # Change glyhb to categorical: glyhb>7 = 1 , glyhb<7 = 0
+      if (dat5$glyhb[i] >= 7){
+        dat5$glyhb.cat[i] = 1
+      } else if (dat$glyhb[i] < 7){
+        dat5$glyhb.cat[i] = 0
+      }
+    }
>    dat5$w.h.ratio = (dat5$waist)/dat5$hip #waist-hip ratio
>    dat5$bmi = dat5$weight/((dat5$height)^2)*703 #Add new variable BMI
>    dat5$frame = factor(dat5$frame, levels = c("small", "medium", "large"))  # Change "frame" to a factor small, medium, large
>
>
>    fitimp2.2 = glm(glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + age
+                  + bp.1s + gender*bmi + w.h.ratio,
```

```
+                 data = dat2, family = binomial() )
>    sumfitimp2.2 = summary(fitimp2.2);
>    sumfitimp2.2

Call:
glm(formula = glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) +
   hdl + age + bp.1s + gender * bmi + w.h.ratio, family = binomial(),
   data = dat2)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.7980  -0.2793  -0.1425  -0.0595   3.9583

Coefficients: (1 not defined because of singularities)
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -21.562380   4.655701  -4.631 3.63e-06 ***
chol                   0.007597   0.005147   1.476 0.139973
stab.glu               0.112539   0.030406   3.701 0.000215 ***
rcs(stab.glu, 3)stab.glu        NA        NA     NA       NA
rcs(stab.glu, 3)stab.glu'  -0.203366   0.077115  -2.637 0.008360 **
hdl                   -0.015976   0.013841  -1.154 0.248414
age                    0.024863   0.015823   1.571 0.116119
bp.1s                  0.012679   0.009528   1.331 0.183303
gendermale             2.269839   2.130793   1.065 0.286760
bmi                    0.042398   0.035890   1.181 0.237474
w.h.ratio              3.484139   3.145612   1.108 0.268026
gendermale:bmi        -0.099523   0.073295  -1.358 0.174513
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 339.14  on 402  degrees of freedom
Residual deviance: 155.24  on 392  degrees of freedom
AIC: 177.24

Number of Fisher Scoring iterations: 7

>    plot(fitimp2.2)
>
>    fitimp2.3 = glm(glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + age
+                 + bp.1s + gender*bmi + w.h.ratio,
+                 data = dat3, family = binomial() )
>    sumfitimp2.3 = summary(fitimp2.3);
>    sumfitimp2.3

Call:
glm(formula = glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) +
   hdl + age + bp.1s + gender * bmi + w.h.ratio, family = binomial(),
   data = dat3)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.7999  -0.2785  -0.1429  -0.0596   3.9582

Coefficients: (1 not defined because of singularities)
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -21.561625   4.647818  -4.639  3.5e-06 ***
chol                   0.007547   0.005141   1.468 0.142111
stab.glu               0.112500   0.030382   3.703 0.000213 ***
rcs(stab.glu, 3)stab.glu        NA        NA     NA       NA
rcs(stab.glu, 3)stab.glu'  -0.203276   0.077059  -2.638 0.008341 **
hdl                   -0.015953   0.013840  -1.153 0.249050
age                    0.024870   0.015820   1.572 0.115922
bp.1s                  0.012780   0.009495   1.346 0.178314
gendermale             2.263816   2.144153   1.056 0.291055
```

```
bmi                   0.042798   0.035818   1.195 0.232145
w.h.ratio               3.465085   3.141939   1.103 0.270093
gendermale:bmi          -0.099010   0.073771  -1.342 0.179553
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 339.14  on 402  degrees of freedom
Residual deviance: 155.20  on 392  degrees of freedom
AIC: 177.2

Number of Fisher Scoring iterations: 7

>    plot(fitimp2.3)
>
>    fitimp2.4 = glm(glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + age
+                + bp.1s + gender*bmi + w.h.ratio,
+                data = dat4, family = binomial() )
>    sumfitimp2.4 = summary(fitimp2.4);
>    sumfitimp2.4

Call:
glm(formula = glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) +
   hdl + age + bp.1s + gender * bmi + w.h.ratio, family = binomial(),
   data = dat4)

Deviance Residuals:
   Min      1Q   Median      3Q     Max
-2.8041  -0.2796  -0.1430  -0.0597   3.9606

Coefficients: (1 not defined because of singularities)
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -21.487366   4.640000  -4.631 3.64e-06 ***
chol                   0.007517   0.005132   1.465 0.142984
stab.glu               0.112438   0.030364   3.703 0.000213 ***
rcs(stab.glu, 3)stab.glu        NA        NA     NA      NA
rcs(stab.glu, 3)stab.glu'  -0.203051   0.077020  -2.636 0.008381 **
hdl                   -0.015892   0.013834  -1.149 0.250652
age                    0.025093   0.015806   1.588 0.112379
bp.1s                  0.012588   0.009535   1.320 0.186764
gendermale              2.212110   2.161305   1.024 0.306068
bmi                    0.042472   0.035890   1.183 0.236650
w.h.ratio               3.420605   3.137245   1.090 0.275572
gendermale:bmi          -0.097178   0.074379  -1.307 0.191375
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 339.14  on 402  degrees of freedom
Residual deviance: 155.35  on 392  degrees of freedom
AIC: 177.35

Number of Fisher Scoring iterations: 7

>    plot(fitimp2.4)
>
>
>    fitimp2.5 = glm(glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) + hdl + age
+                + bp.1s + gender*bmi + w.h.ratio,
+                data = dat5, family = binomial() )
>    sumfitimp2.5 = summary(fitimp2.5);
>    sumfitimp2.5

Call:
```

```
glm(formula = glyhb.cat ~ chol + stab.glu + rcs(stab.glu, 3) +
   hdl + age + bp.1s + gender * bmi + w.h.ratio, family = binomial(),
   data = dat5)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.8881  -0.3016  -0.1605  -0.0748   3.8649

Coefficients: (1 not defined because of singularities)
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -19.740987   4.370964  -4.516 6.29e-06 ***
chol                   0.005693   0.004816   1.182 0.237119
stab.glu               0.100647   0.028404   3.543 0.000395 ***
rcs(stab.glu, 3)stab.glu        NA        NA      NA       NA
rcs(stab.glu, 3)stab.glu'  -0.175902   0.072787  -2.417 0.015664 *
hdl                   -0.010000   0.013349  -0.749 0.453772
age                    0.031693   0.015410   2.057 0.039719 *
bp.1s                  0.007145   0.009452   0.756 0.449665
gendermale             2.099593   2.106941   0.997 0.319001
bmi                    0.054356   0.034580   1.572 0.115975
w.h.ratio              3.007525   3.078233   0.977 0.328555
gendermale:bmi        -0.091468   0.072412  -1.263 0.206529
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 342.61  on 402  degrees of freedom
Residual deviance: 162.86  on 392  degrees of freedom
AIC: 184.86

Number of Fisher Scoring iterations: 7

>   plot(fitimp2.5)
>
>   #Pool Paramters for 5 datasets
>   paramdat1 = fit_08$coefficients
>   paramdat2 = fitimp2.2$coefficients
>   paramdat3 = fitimp2.3$coefficients
>   paramdat4 = fitimp2.4$coefficients
>   paramdat5 = fitimp2.5$coefficients
>
>   paramdat1
        (Intercept)                chol           stab.glu  rcs(stab.glu, 3)stab.glu
       -21.503234392         0.007524847         0.112141876                  NA
rcs(stab.glu, 3)stab.glu'                hdl                 age              bp.1s
       -0.202352186        -0.015832158         0.025209309         0.012514372
         gendermale                 bmi           w.h.ratio       gendermale:bmi
         2.242253263         0.043514522         3.431615007        -0.098067885
>   paramdat2
        (Intercept)                chol           stab.glu  rcs(stab.glu, 3)stab.glu
       -21.562380046         0.007596636         0.112539367                  NA
rcs(stab.glu, 3)stab.glu'                hdl                 age              bp.1s
       -0.203365804        -0.015975674         0.024862979         0.012678948
         gendermale                 bmi           w.h.ratio       gendermale:bmi
         2.269839073         0.042397759         3.484138667        -0.099523453
>   paramdat3
        (Intercept)                chol           stab.glu  rcs(stab.glu, 3)stab.glu
       -21.561624893         0.007547253         0.112500436                  NA
rcs(stab.glu, 3)stab.glu'                hdl                 age              bp.1s
       -0.203275914        -0.015953009         0.024870192         0.012780495
         gendermale                 bmi           w.h.ratio       gendermale:bmi
         2.263816324         0.042797623         3.465085189        -0.099010288
>   paramdat4
        (Intercept)                chol           stab.glu  rcs(stab.glu, 3)stab.glu
       -21.487365803         0.007517253         0.112437770                  NA
```

```
rcs(stab.glu, 3)stab.glu'                hdl                age               bp.1s
        -0.203051336       -0.015891909        0.025093377        0.012587716
         gendermale                bmi          w.h.ratio      gendermale:bmi
         2.212110402        0.042471793        3.420605167       -0.097177911
>    paramdat5
         (Intercept)               chol         stab.glu  rcs(stab.glu, 3)stab.glu
       -19.740987442        0.005693271        0.100646727                 NA
rcs(stab.glu, 3)stab.glu'                hdl                age               bp.1s
        -0.175901856       -0.010000192        0.031692786        0.007145219
         gendermale                bmi          w.h.ratio      gendermale:bmi
         2.099593258        0.054355909        3.007524505       -0.091468255
>
>    all.equal(paramdat5, paramdat4, paramdat3, paramdat2, paramdat1) #returns TRUE, all equal
Error in all.equal.numeric(paramdat5, paramdat4, paramdat3, paramdat2,  :
  all(scale > 0) is not TRUE
In addition: Warning messages:
1: In if (countEQ) { :
  the condition has length > 1 and only the first element will be used
2: In if (!countEQ) N <- length(target) :
  the condition has length > 1 and only the first element will be used
```