HOMEWORK ANALYSIS #8 – DIABETES

Type 2 diabetes is a problem with your body that causes blood sugar levels to rise higher than normal (hyperglycemia) because your body does not use insulin properly. Specifically, your body can't make enough insulin to keep your blood sugar levels normal. Type 2 diabetes is associated with various health complications such as neuropathy (nerve damage), glaucoma, cataracts and various skin disorders. Early detection of diabetes is crucial to proper treatment so as to alleviate complications.

The dataset Diabetes.txt contains information on 768 women who are at risk for diabetes. The dataset contains the following variables:

Variable Name	Description
pregnant	Number of times pregnant
glucose	Plasma glucose concentration at 2 hours in an oral glucose tolerance test
diastolic	Diastolic blood pressure (mm Hg)
triceps	Triceps skin fold thickness (mm)
insulin	2 hour serum insulin (mu U/ml)
bmi	Body mass index
pedigree	Numeric strength of diabetes in family line (higher numbers mean stronger history)
age	Age
diabetes	Does patient have diabetes (0 if "No", 1 if "Yes")

1. In your own words, summarize the overarching problem. Discuss how statistical modeling will be able to answer the posed questions.

Ultimately, we want a way to help doctors predict diabetes earlier so they can provide proper treatment, and alleviate complications better. We will need to check if can use the variable in the risk factor variable table to accurately predict diabetes. If we can, we can then build a model that takes in those variable and use the model to predict diabetes earlier.

2. Explore the data using basic exploratory graphics and summary statistics. Include scatterplots with smooth curves to show the relationship between 2 covariates and the response (diabetes). Comment on any potential relationships you see through this exploratory analysis. Explain why traditional multiple linear regression methods are not suitable for this problem.

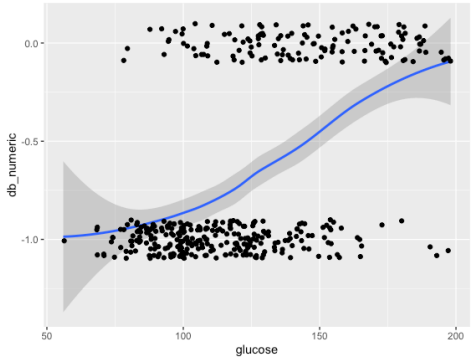

Figure 1

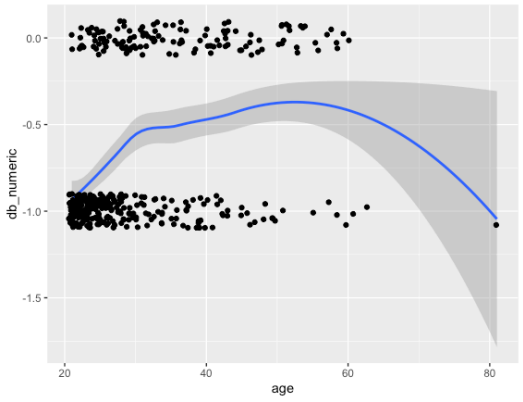

Figure 2

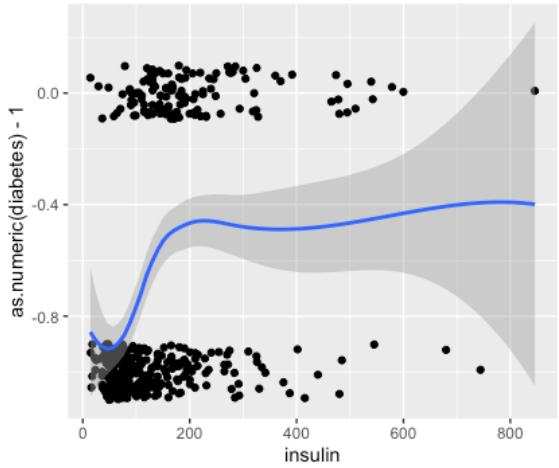

Figure 3

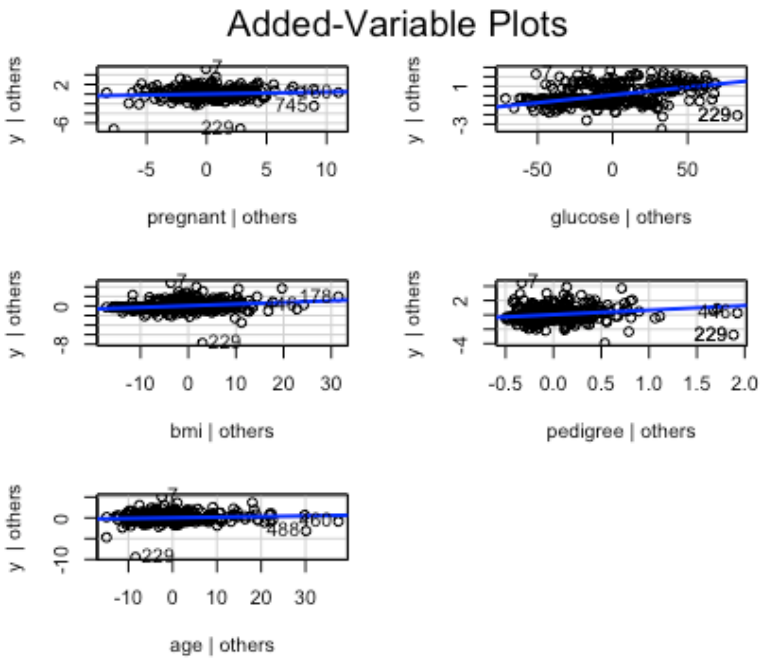

Figure 4

From figures 1, 2, and 3 it appears that as glucose increases, the risk of diabetes seems to increase; as age increases, the risk of diabetes seems to increase; and as insulin increases, the risk of diabetes also seems to increase.

Traditional multiple linear regression methods are not suitable for this problem because what we are trying to predict is binary. People either have diabetes, or they do not. This is different from outputs that are quantitative, and there for different techniques need to be used to understand the data.

3. Use variable selection to choose which variables to use in a logistic regression model for diabetes. Provide a justification of your choice in criteria (AIC or BIC) and algorithm (forward vs. backward vs. exhaustive). What factors do you find are important in explaining the presence of diabetes?

Generally AIC is used when prediction is the goal, and BIC is used when inference is the goal. Because predicting diabetes is the goal, AIC was used. Exhaustive always better when possible. It was better so we used it. Pregnancy, glucose, bmi, pedigree and age are the important factors in explaining the presence of diabetes.

4. Write out a logistic regression model (using greek letters) that includes your chosen covariates. Describe and justify any assumptions that you use in writing out your model.

We let y_i be the binary response variable, where 1 indicates that they have diabetes, and 0 indicates that they don't have diabetes.

$Y_i \sim^{ind} Bernoulli(p_i)$

$Log(p_i / (a - p_i)) = Beta_0 + Beta_1 * pregnant + Beta_2 * glucose + Beta_3 * bmi + Beta_4 * pedigree + Beta_5 * age$

I assumed
1 That the log-odds of diabetes are linear with respect to our covariates
2 The response variables are independent given the covariate information available.
3 Our responses are distributed as Bernouli RVs.

To justify the linearity in the log-odds assumption, I provide added variable plots (figure 4) and smoothed scatterplots (figures 1,2, and 3). These don't show major violations of linearity in the log-odds. To justify the independence, we don't think that anyone's diabetes status would affect another's given that we don't have a lot of info about the data collection process. Although, if some observations were taken from two or more individuals in the same household there could be a possibility of non-independence when considering things like shared eating habits.

5. Fit the corresponding logistic regression model and give a 95% confidence interval for each effect therein. Interpret at least one (but not the intercept) of these intervals in the context of the problem.

I am 95% confident that holding all other variables constant if the individual is Pregnant their log-odds chances of having diabetes changes by between -2.304 and 21.297.

I am 95% confident that holding all other variables constant as glucose increases by one unit the individual's log-odds chances of having diabetes changes by between 2.740 and 4.771.

I am 95% confident that holding all other variables constant as BMI increases by one unit the individual's log-odds chances of having diabetes goes up by between 3.950 and 12.732.

I am 95% confident that holding all other variables constant as Pedigree increases by one unit the individual's log-odds chances of having diabetes goes up by between 40.166 and 639.816.

I am 95% confident that holding all other variables constant as Age increases by one unit the individual's log-odds chances of having diabetes goes up by between -0.0023, and 7.266.

6. Determine an appropriate threshold for classification that minimizes the misclassification rate. Provide an appropriate plot showing that this is indeed the minimum.

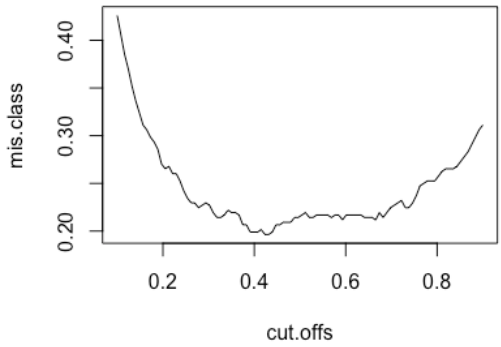

Figure 5

The appropriate threshold for classification that minimizes the misclassification rate is 0.4232323. See figure 5.

7. Assess the model fit by build a confusion matrix from all the data (i.e. not cross-validated) using the classification threshold that you found in the previous problem, report the pseudo-R2 and AUC for your logistic regression model. Comment on how well your model fits the data by its ability to correctly classify patients in the dataset. State your results in terms of sensitivity, specificity, positive predictive value and negative predictive value.

	0	1
0	230	45
1	32	85

Table 1

Sensitivity	0.6538462

Specificity	0.8778626
Positive predictive value	0.7264957
Negative predictive value	0.8313673

Table 2

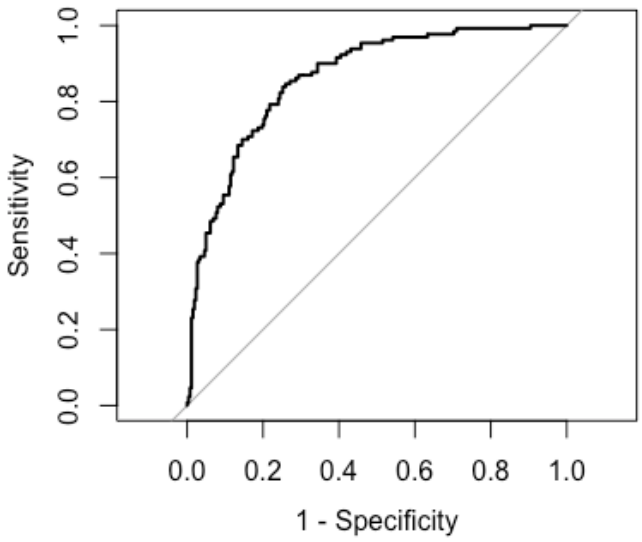

Figure 6

The Pseudo-R2 = 0.3075956
Area under the curve: 0.8631

The model fit is good because the sensitivity, specificity, positive predictive value and negative predictive value are all high; especially specificity and negative predictive value. Also the area under the curve is close to 1, which means the model fits most of the data well. The Pseudo-R2 is a touch low though meaning we can only account for 31% of the variance in diabetes based on input parameters.

8. Assess the predictive ability of your model by running a cross-validation study where you classify the "test" patients using the threshold you found above. Report your results in terms of the average sensitivity, specificity, positive predictive value and negative predictive value for the test sets.

Sensitivity	0.6436225
Specificity	0.8565715
Positive predictive value	0.6936048
Negative predictive value	0.8279712

Table 3

The model fit is good because the sensitivity, specificity, positive predictive value and negative predictive value are all high; especially specificity and negative predictive value.

9. Predict the probability of diabetes for the following patient: pregnant= 1, glucose= 90, diastolic= 62, triceps= 18, insulin= 59, bmi= 25:1, pedigree= 1:268 and age= 25. Do you think patient has diabetes? Why or why not?

This model predicts that with the given values above for the variables the individual does **not** have diabetes because their log -dds score was **0.0872821** which was **below** the **0.4232323** cut off.

Appendix 1

Source code: Is an R mark down file.
This code was derived from in class examples, TA help, and Stack Overflow suggestions.

```
---
title: "Diabetes3"
author: "Colin White"
date: "4/1/2021"
output: pdf_document
editor_options:
  chunk_output_type: console
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
##Clean and load data
```{r, message=FALSE, warning=FALSE}
library(ggplot2)
library(bestglm)
library(car)
library(knitr)
library(pROC)

setwd("~/Desktop/1A School/1A Winter 2021/STAT330/HW8")

rm(list = ls()) #### delete all variables
par(mfrow=c(1,1))

diabetes = read.table("fileDownload.txt", header=TRUE)
flag = which(apply(diabetes[,2:6],1,function(x){any(x == 0 )}))
diabetes_use = diabetes[-flag,]

rm(diabetes)
attach(diabetes_use)

```

##question 1
```

##question 2
```{r}
scatter.smooth(insulin,diabetes)
scatter.smooth(glucose,diabetes)
scatter.smooth(age,diabetes)

ggplot(diabetes_use,aes(x=insulin,y=as.numeric(diabetes)-1)) + geom_jitter(width=.5,height=.1) +
  geom_smooth()

ggplot(diabetes_use,aes(x=glucose,y=as.numeric(diabetes)-1)) + geom_jitter(width=.5,height=.1) +
  geom_smooth()

ggplot(diabetes_use,aes(x=age,y=as.numeric(diabetes)-1)) + geom_jitter(width=.5,height=.1) +
  geom_smooth()

```

#question 3
```{r}


var.select  = bestglm(diabetes_use,IC = 'AIC', family=binomial,method="exhaustive")
var.select$BestModel
best.model = var.select$BestModel
```

#question 4
```{r}
avPlots(best.model)
```

#problem 5
```{r}
CI = confint(best.model)

kable(CI)
kable(exp(CI)) # actual conf
kable(100*(exp(CI) - 1))
```

#problem 6
```{r}
my.preds = predict.glm(best.model, type="response")
cut.offs = seq(.1,.9,length = 100)
mis.class = numeric(length(cut.offs))

for(i in 1:length(cut.offs)){
  cutoff = cut.offs[i]
  classify = ifelse(my.preds>cutoff,1,0)
  mis.class[i] = mean(classify != diabetes_use$diabetes)
}
```

```
plot(cut.offs, mis.class,type="l")

cutoff_use = cut.offs[which.min(mis.class)]
cutoff_use
```

##problem 7
```{r}

r2 = 1 - best.model$deviance/best.model$null.deviance
r2
pred.class = ifelse(predict.glm(best.model,type="response")>cutoff_use,1,0)
conf.mat = table(predicted = pred.class,True = diabetes_use$diabetes)
kable(conf.mat)

sens = conf.mat[2,2] /(conf.mat[2,2]+ conf.mat[1,2])
spec = conf.mat[1,1] /(conf.mat[2,1]+ conf.mat[1,1])
ppv = conf.mat[2,2] /(conf.mat[2,2]+ conf.mat[2,1])
npv = conf.mat[1,1] /(conf.mat[1,1]+ conf.mat[1,2])

sens
spec
ppv
npv
```



```{r}
my.roc = roc(diabetes_use$diabetes,my.preds)
plot(my.roc,legacy.axes=TRUE)
auc(my.roc)
```

##problem8
```{r}

n.cv = 1000
n.test = round(.2 * nrow(diabetes_use))

sens <- spec <- ppv <- npv <- numeric(n.cv)

for(i in 1:n.cv){
  idx_test = sample(1:nrow(diabetes_use),n.test)
  train.data = diabetes_use[-idx_test,]
  test.data = diabetes_use[idx_test,]

  train.mod = glm(diabetes ~pregnant+ glucose+bmi+pedigree+age,
                  data = train.data, family = binomial)
  pred = predict.glm(train.mod,newdata = test.data, type = "response")
  pred.class = ifelse(pred > cutoff_use, 1, 0)
```

```
  conf.mat = table(predicted = pred.class, True = test.data$diabetes)

  sens[i] = conf.mat[2,2] /(conf.mat[2,2]+ conf.mat[1,2])
  spec[i] = conf.mat[1,1] /(conf.mat[2,1]+ conf.mat[1,1])
  ppv[i] = conf.mat[2,2] /(conf.mat[2,2]+ conf.mat[2,1])
  npv[i] = conf.mat[1,1] /(conf.mat[1,1]+ conf.mat[1,2])
}

print("Sensitivity is"); mean(sens)
mean(sens)
mean(spec)
mean(ppv)
mean(npv)
```

#9
```{r}
predict.glm(best.model,newdata = data.frame(pregnant=1, glucose=90,
diastolic=62, triceps=18, insulin=59, bmi=25.1, pedigree=1.268, age=25),type = "response")
```