Stat 6021: Guided Question Set 11

The Western Collaborative Group Study (WCGS) is one of the earliest studies regarding heart disease. Data were collected from 3154 males aged 39 to 59 in the San Francisco area in 1960. They all did not have coronary heart disease at the beginning of the study. The data set comes from the `faraway` package and is called `wcgs`. We will focus on predicting the likelihood of developing coronary heart disease based on the following predictors:

- `age`: age in years
- `sdp`: systolic blood pressure in mm Hg
- `dbp`: diastolic blood pressure in mm Hg
- `cigs`: number of cigarettes smoked per day
- `dibep`: behavior type, labeled A and B for aggressive and passive respectively.

The response variable is `chd`, whether the person developed coronary heart disease during annual follow ups in the study. Read the data in. We will also randomly split the data into two: half the data will be the training data set, and the remaining half will be the test data set. We will explore the training-test split in more detail in the next module. For this exercise, perform all analysis on the training data. The code below will randomly split the data into two halves.

```
library(faraway)
Data<-wcgs
set.seed(6021) ##for reproducibility to get the same split
sample<-sample.int(nrow(Data), floor(.50*nrow(Data)), replace = F)
train<-Data[sample, ] ##training data frame
test<-Data[-sample, ] ##test data frame
```

1. Before fitting a model, create some data visualizations to explore the relationship between these predictors and whether a middle-aged male develops coronary heart disease.
2. Use R to fit the logistic regression model using all the predictors listed above, and write the estimated logistic regression equation.

3. Interpret the estimated coefficient for `cigs` in context.

4. Interpret the estimated coefficient for `dibep` in context.

5. What are the estimated odds of developing heart disease for an adult male who is 45 years old, has a systolic blood pressure of 110 mm Hg, diastolic blood pressure of 70 mm Hg, does not smoke, and has type B personality? What is this person's corresponding probability of developing heart disease?

6. Carry out the relevant hypothesis test to check if this logistic regression model with five predictors is useful in estimating the odds of heart disease. Clearly state the null and alternative hypotheses, test statistic, and conclusion in context.

7. Suppose a co-worker of yours suggests fitting a logistic regression model without the two blood pressure variables. Carry out the relevant hypothesis test to check if this model without the blood pressure variables should be chosen over the previous model with all four predictors.

8. Based on the Wald test, is diastolic blood pressure a significant predictor of heart disease, when the other predictors are already in the model?

9. Based on all the analysis performed, which of these predictors would you use in your logistic regression model?