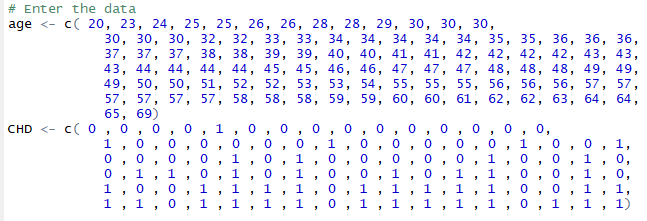
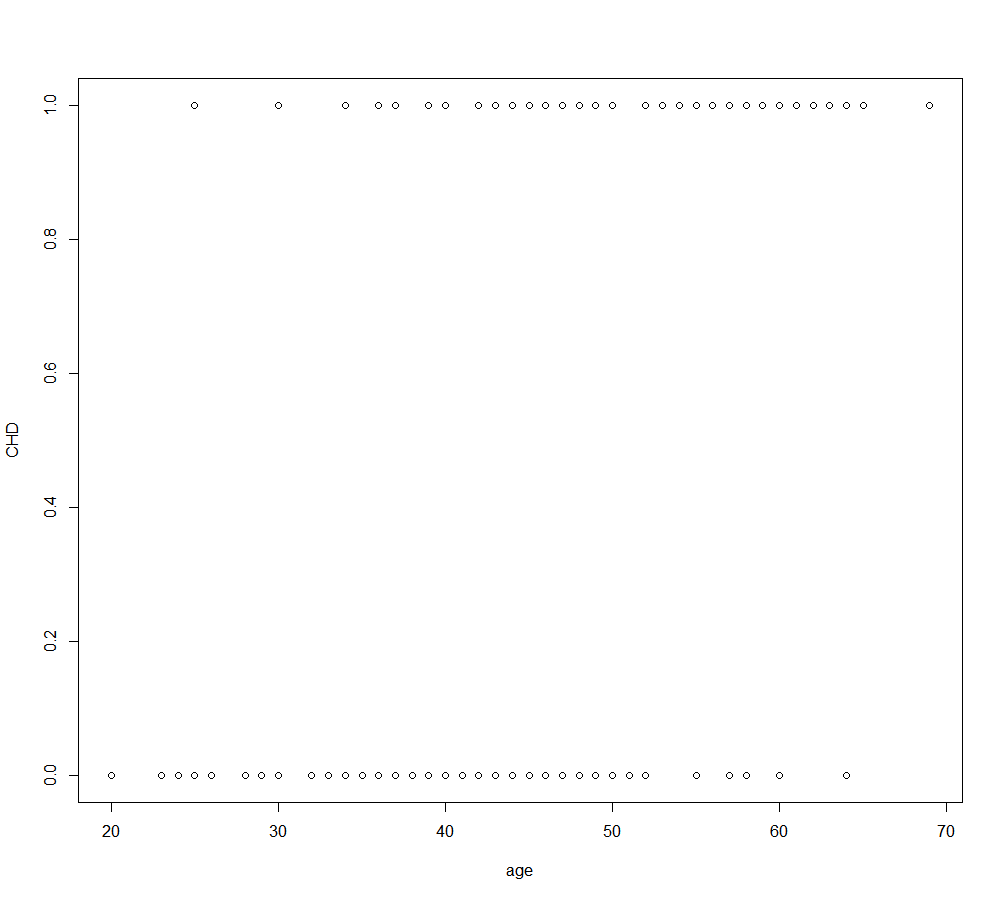
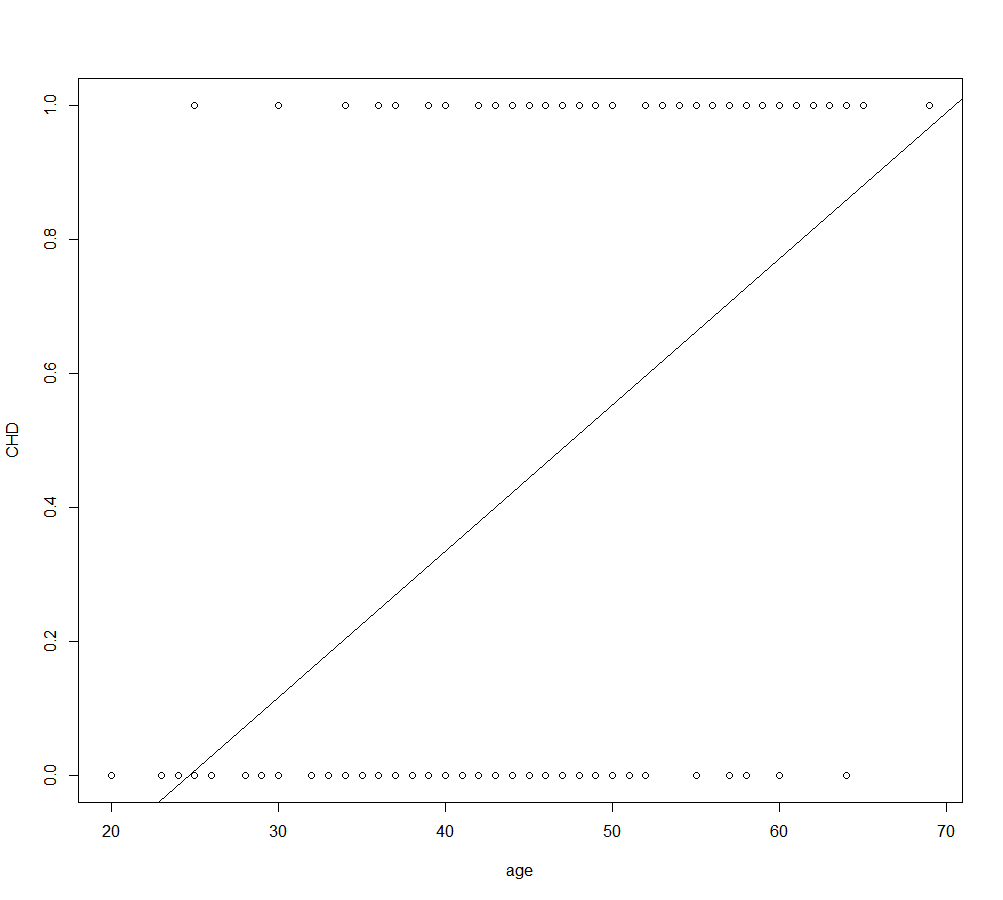
**Part A:** Predicting a binary response using single predictor (univariate).



1. Plot a scatter plot between age and CHD (Coronary Heart Disease).



1. Plot a linear line for scatter plot in Part (1).



1. Explain the problem encountered in Part (2).

Any time a straight line is fit to a binary response that is coded as 0 or 1, in principle we can always predict for some values of and for others (unless the range of is limited). To avoid this problem, we must model using a function that gives outputs between 0 and 1 forall values of . Many functions meet this description. In logistic regression, we use the ***logistic function.***

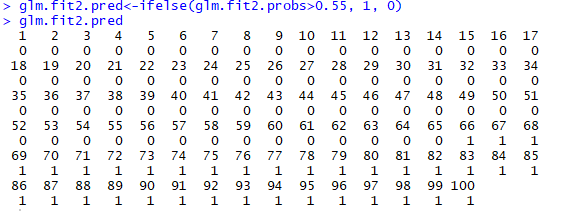
1. Use a logistic regression to model the prediction of CHD event rates using risk factor age.

glm.fit2 <- glm(CHD ~ age, family = binomial)

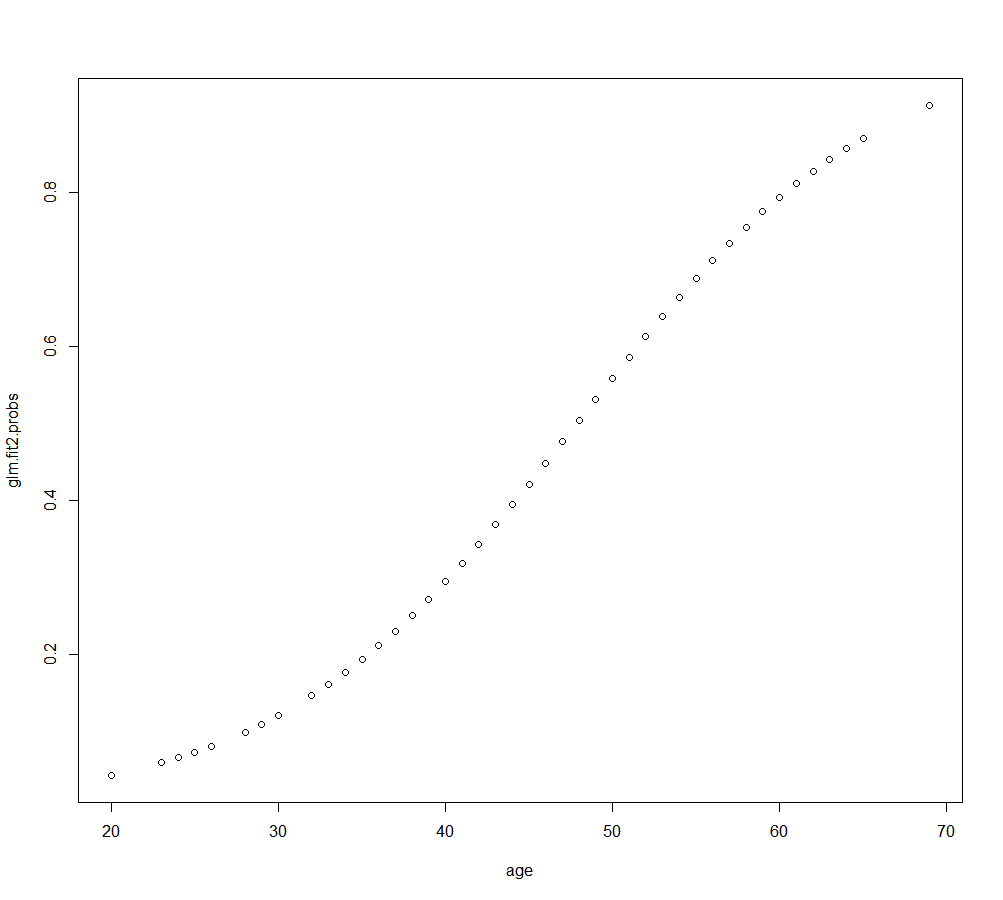
1. What is the estimated regression equation for the model in Part (4).

= -5.3094534 + (0.1109211\*age)

1. Make predictions for each subject in the dataset.



1. Plot age versus fitted value for age.



1. Predict the occurrence of CHD for a patient with age=30 (assuming that the threshold for a positive CHD is 0.55).

glm.fit2.pred<-ifelse(glm.fit2.probs>0.55, 1, 0)



For age = 30, as probability value is less than 0.55, CHD = 0.

**Part B:** Predicting a binary response using multiple predictors (multivariate).

The default data set resides in the ISLR package of the *R* programming language. It contains selected variables and data for 10,000 credit card users. Some of the variables present in the default data set are:

**student** - A binary factor containing whether or not a given credit card holder is a student.

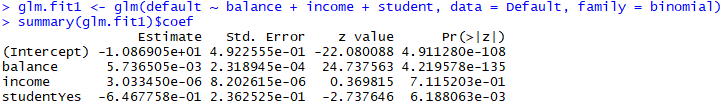
**income** - The gross annual income for a given credit card holder.

**balance** - The total credit card balance for a given credit card holder.

**default** - A binary factor containing whether or not a given user has defaulted on his/her

credit card.

1. Find the coefficient estimates for a logistic regression model that uses balance, income (in thousands of dollars), and student status to predict probability of default.



1. Interpret the coefficient for balance and student in Part (1).

As income predictor has low significance, we don’t consider it in equation.

Logit(p) = -1.087e+01 + (5.737e-03\*balance) + (-6.468e-01\*student)

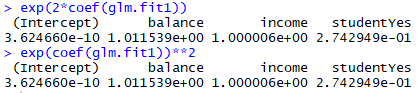
Balance: A one-unit increase in balance is associated with an increase in the log-odds of default by 5.736505e-03 units

Student: A one-unit increase in student is associated with an increase in the log-odds of default by -6.468e-01 units.

1. Compute the odds ratio, for balance and income. Note that the coefficient is such that is the odds ratio for a unit change in .



1. If we change income by two units, compute its odds ratio (for a change of units, the =.



1. Estimate the probability of default for
   1. A student with a credit balance of $1,500 and an income of $40,000.



* 1. A non-student with a credit balance of $1,500 and an income of $40,000.

