**PROBLEM 1.**

The problem deals with a chemical process whose yield(Y) can be determined on the basis of two factors : temperature($X_1$) and pressure($X_2$) through the following non-linear regression model :

$Y_i = \theta_0.X_{i1}^{\theta_1}.X_{i2}^{\theta_2} + \epsilon_i$, where $\epsilon_i$'s are the independent additive Gaussian error term assumed to have homogeneous error variance and mean equal to 0.

To get an essence of the data relating to the already in use old machine, I plot out the data :

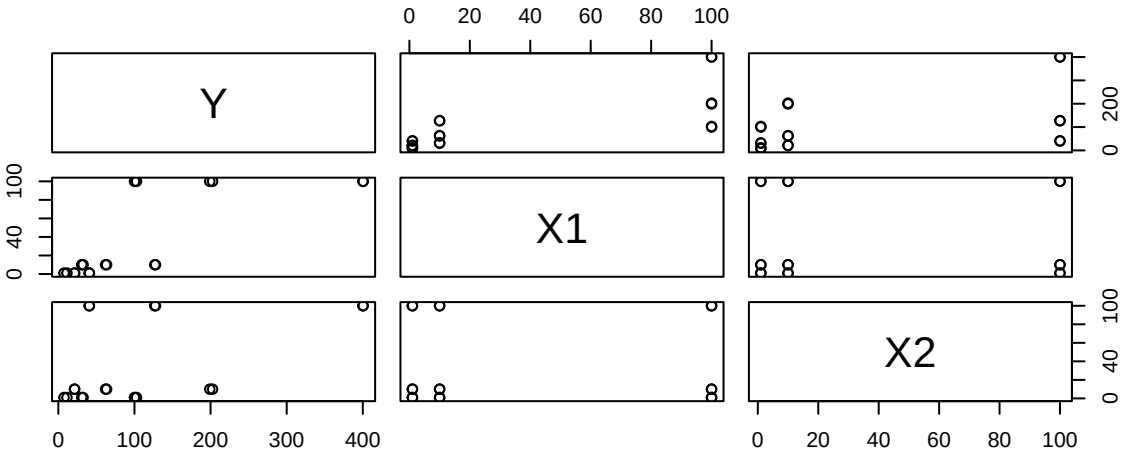



Figure 1: Pairs Plot

It seems from Figure 1 that there is no multicollinearity present among temperature and pressure.

The main goal of the problem is to find estimates of the parameters $\theta_0, \theta_1$ and $\theta_2$ on the basis of some initial starting values, which I have determined by assuming a linear model. The data is obtained from an old machine already in use.

Here, I obtain the starting values for $\theta_0, \theta_1$ and $\theta_2$ respectively as 9.8379473, 0.5012646, 0.3076603.

To get hold of the final converged estimates of the parameters, I have used the Gauss-Newton optimization algorithm and they read as follows : 10.1576718, 0.4984098, 0.2994003. Also, my algorithm converged at the 3rd iteration. I have also used the `nls()` to check the results of my optimization algorithm and they seem to be sharing similar results.

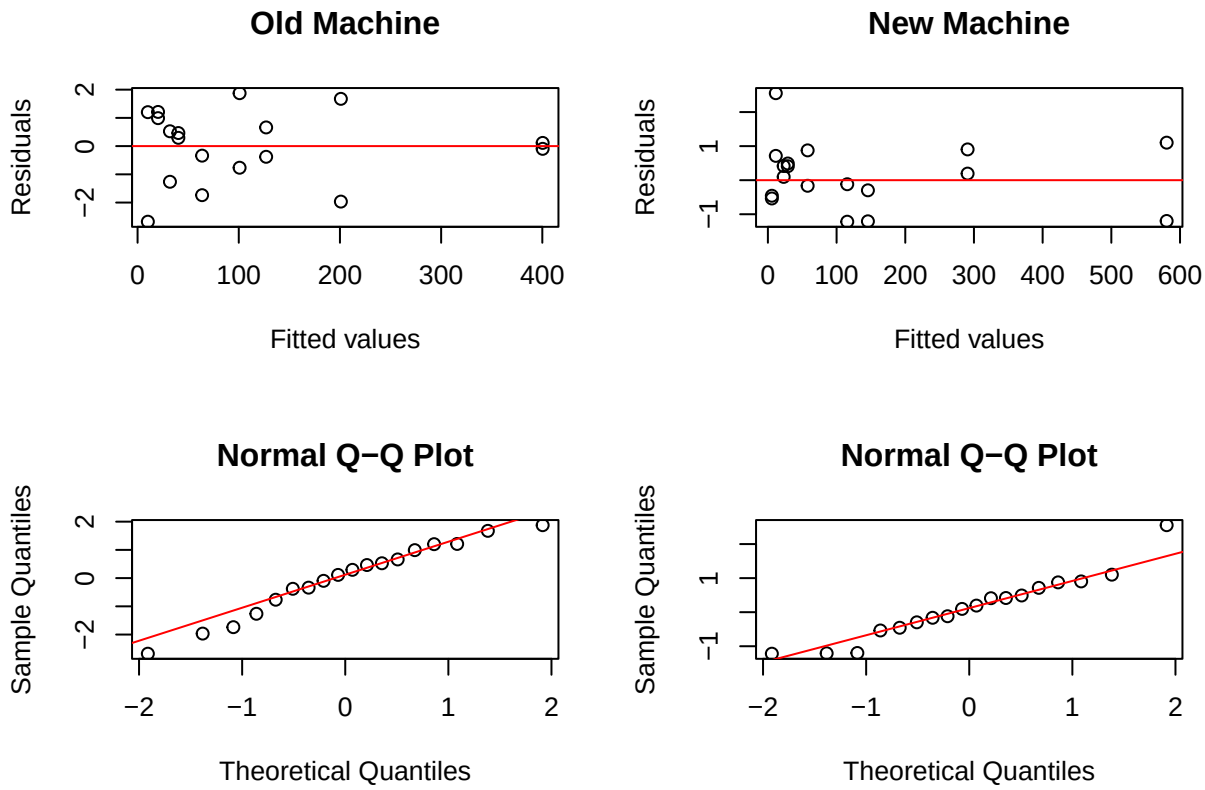
Also, it seems from the performed testing of hypothesis where $H_0 : X1/X2 = 1$ vs $H_a$ :not $H_0$ that there is strong evidence against the two predictors affecting the response similarly.

Again, the parameters of the model seem to be quite significant as the confidence intervals for each of the parameters seem to contain each of them respectively.

Due to increasing demand, the factory buys another similar machine and now, the problem of interest lies in whether the data obtained from this machine is similar in nature to the data of the old machine which was a result of the fit of the previous model. So, I repeat the same procedure as for the old machine and obtain the final parameter estimates.

The final estimates for the new machine are 5.7673885, 0.7011172, 0.3003161.My algorithm converged at the 3rd iteration. The `nls()` again gives similar results.

To compare the behaviour of the model on the two given datasets, I plot out the residuals and check if the model assumptions are successfully met :




The points bounce off the zero line to an optimum distance in the residuals vs fitted without following any specific pattern plots. In the QQ-plots for the residuals, the points are quite perfectly aligned to the slope. So, it is evident from the plots and the pseudo R-squared values of 0.9998882 and 0.8483418 for the old and new machine datasets respectively that the model fits the data quite well and the model assumptions of homogeneous variance, normality of the residuals are met. The pseudo R-squared value for the new machine falls a bit due to the fact of fitting the exact same model, though which might not be the case in reality, but, it can be surely said that the data from the new machine would be obtained from a model with very similar structure(changed coefficients).

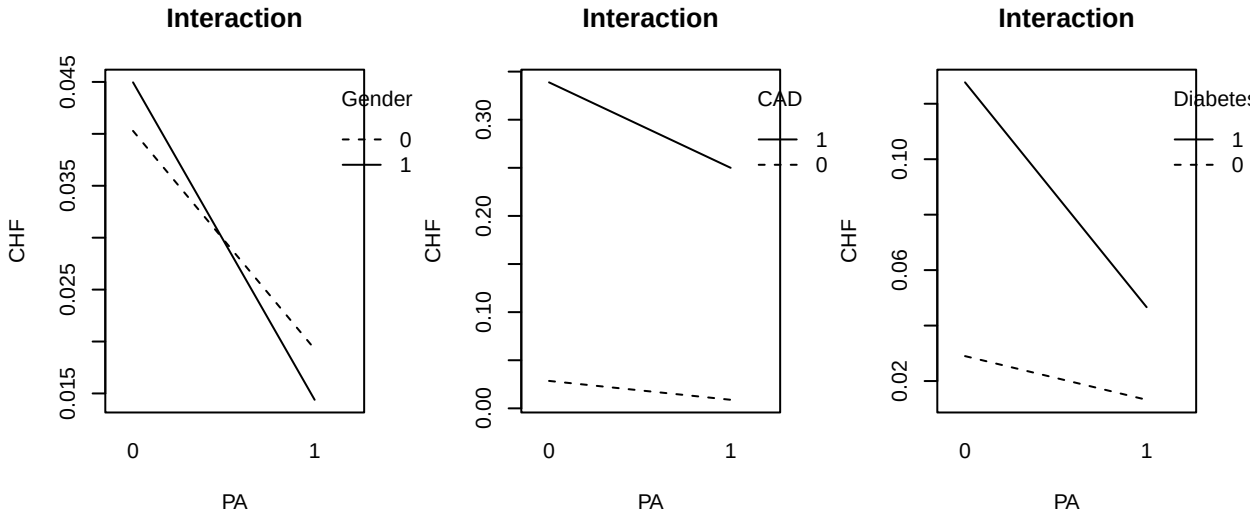
In conclusion, I can say that the model fits the datasets of the old and the new machine quite well,though they might not come from exactly the same model due to the fact of different coefficients. The specified model can be relied upon in the future to study the mentioned chemical process in the presence of temperature and pressure as covariates. Also, as a model can never be completely perfect, there is always room for improvement in the fact that some transformations can be done on the response variable so that the normality assumption of the residual can be met more perfectly as compared to the given model.

**PROBLEM 2.**

The problem deals with the fact of finding out the most prominent risk factors for congestive heart failure (CHF) and to what extent do they affect the occurence of CHF. For this purpose, I am using a dataset consisting of 5513 individuals from the US, who were surveyed during 2013-2014. The data consists of 12 predictor variables, v.i.z, ID, Gender(0=male,1=female), Age, Race(0=not Africa-American,1=African-American), Educ(Education level-binary), BP(daignosed with hypertension or not), Chol(high cholesterol or not), Diabetes(Diagnosed with diabetes or not), CAD(diagnosed with coronary artery disease or not), PA(Does at least 10 minutes of moderate physical exercise once a week or not), BMI and Alcohol(abnormal alcohol consumption or not). For this purpose, I am separating the entire data into the training set(4500 data points) and testing set(rest of the data),solely for prediction purposes.

Before moving on to model fitting, I would like to explore the data more to get a better understanding of the possible significant covariates. Primarily, I would like to consider the pairwise scatterplot and correlation matrix :

There does not seem to be any problem of multicollinearity. Also, I feel that understanding the possible effect of the covariates on the response variables should majorly rely upon the discretion of the researcher. From this understanding, I feel that ID and Education level does not have any significant effect on whether a person has been diagnosed with CHF or not and thus, I discard thesse two covariates. Now, I would try to find the possible interaction effects between Age, Diabetes and Coronary Artery Disease with all other covariates through interaction plots.




It can be seen that Diabetes:BMI, Race:Alcohol, CAD:Alcohol etc. have possible interactions among themselves, which means that the effect of other predictors on the response change for some covariates.

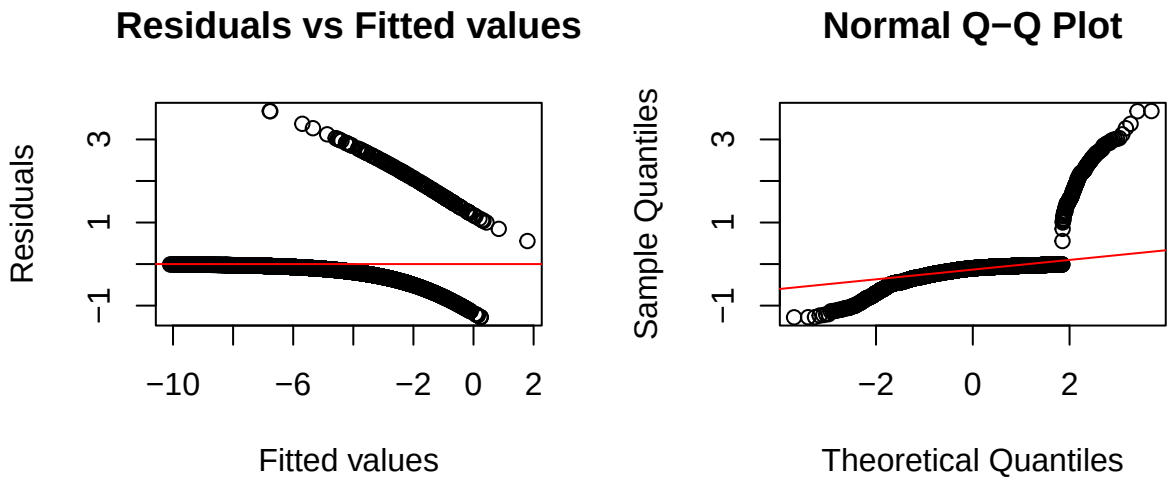
Now to model the data, it can be seen that the responses are binary in nature(i.e, outcome is either 0 or 1). I specify the model as :

$Y_i = \frac{exp(x_i^{'}\beta)}{1+exp(x_i^{'}\beta)} + \epsilon_i$, where $\epsilon_i$ is the non-Gaussian error term with mean 0 and heterogeneous variance.

The model can be alternatively written as a logit function : $logit(p_i) = log(\frac{p_i}{1-p_i})$ , where $p_i = \frac{exp(x_i^{'}\beta)}{1+exp(x_i^{'}\beta)}$ and $0 < p_i < 1$.

Primarily, I have fitted the model with the selected covariates along with the first order interactions in RStudio and considered it as the full model. Then, I use stepwise regression to find an optimal model with the interactions present, which also resulted in a lower AIC value. This final model(CHF ~ Age + Race + BP + Diabetes + CAD + PA + BMI + Alcohol + Age:Diabetes + Age:CAD + Race:CAD + Race:Alcohol + Diabetes:CAD + Diabetes:BMI + Diabetes:Alcohol + CAD:BMI + CAD:Alcohol + PA:Alcohol + BMI:Alcohol) partly matches with the graphical representations I have done earlier, but there are some differences for sure.

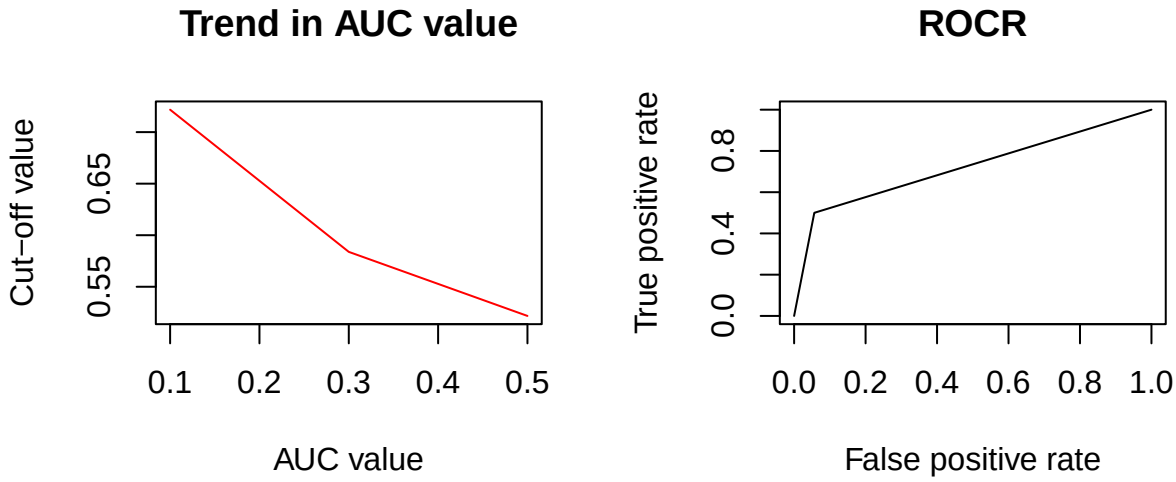
It is evident from the residual vs fitted plot that the points are not randomly scattered and there is pattern in them. This signifies the presence of heterogeneous error variances. The QQ-plot also detects the lack of normality in the residuals. All these are desired as it perfectly meets with assumptions of the generalized linear model : non-constancy error variance, non-normality of the error terms, no autocorrelation among the errors.




Next, I have performed different testing of hypothesis to validate the goodness of fit of the model, v.i.z, the Deviance test and the Hosmer-Lemeshow test.

The deviance test yields a p-value less than 0.05 leading to stong evidence against the null and thus the selected covariates have significant effect on the response. The Hosmer-Lemeshow test yielded a p-value greater than the significance level of 0.05, leading to to strong evidence in favour of the null hypothesis and thus concluding that the final model fits the data well.

Coming to the predictive ability of the model, I have used the test dataset to compute the confusion matrices with cut-off values 0.5, 0.3 and 0.1 respectively and plotted out the value of area under the curve.




It is evident from the curves that as I decrease the cut-off value for prediction, I get a better sensitivity and specificity rate with an AUC value of approximately 72% for cut-off value 0.1. I have intentionally lowered the cut-off value to reduce the bias towards 0's as it can be observed that there is a tremendous unbalance in the data(more number of 0's than 1's).

In conclusion, I can say that the selected predictor variables along with the significant interactions are the most prominent risk factors for CHF in consideration with the generalized linear model with a decent prediction ability. Furthermore, some improvements can be done to the model by reducing the imbalance in the dataset(by `ROSE()` in R) rather than reducing the cut-off value, which would help to keep essence of the original data intact.

References :


(i) Statistics 8320: Data Analysis II Generalized Linear Models by Christopher K. Wikle

(ii) Statistics 8320: Data Analysis II Nonlinear Regression by Christopher K. Wikle

(iii) Non-linear regression in R by Christian Ritz & Jens Carl Streibig

(iv) I have used Google to get better insight about the documentation regarding few R functions.