Statistical Modelling 2 Coursework

Juliette Limozin, CID: 01343907

March 9, 2020

---

1 Introduction

We are given a data containing 169 patients records of their resting heart pulse, their BMI (Body Mass Index), an categorical factor indicating whether or not they consume coffee regularly, and their heart rate 30 min after receiving a stimulant that has similar effects to caffeine. The goal of this report is to explore different linear models in order to best predict the stimulated pulse.

2 Exploratory analysis of the data set

The data set consists of 169 patients, including 109 coffee consumers. Here is a statistical summary of the data:

```
  rest_pulse    stimulated_pulse      bmi        coffee_reg
Min.   :58.00   Min.   : 61.00   Min.   :16.40   0: 60
1st Qu.:62.00   1st Qu.: 69.00   1st Qu.:20.00   1:109
Median :64.00   Median : 73.00   Median :21.00
Mean   :65.28   Mean   : 74.17   Mean   :20.91
3rd Qu.:68.00   3rd Qu.: 78.00   3rd Qu.:22.10
Max.   :79.00   Max.   :100.00   Max.   :24.60
```

Figure 1: Summary of data set

Rest pulse SD <dbl>	Stimulated pulse SD <dbl>	BMI SD <dbl>
4.512131	7.442647	1.646771

Figure 2: Standard deviations of the data set

As you can see from the means, standard deviations and minimum and maximum values, our data seems to have a consistent range of patients, we don't have patients with significantly higher or lower health levels. We can also look at the histogram of the BMIs (figure 3) to notice that we have an almost normal-looking distribution of the patiens' BMIs which is consistent with what we observe in real life.

We also find out that the stimulated pulse is not very correlated to the rest pulse (*cor* about 0.53) nor BMI (*cor* about 0.33).

3 Results

We have summarised the diagnostics of each considered model into the following table. We may assume that other than for the clinicians' model, all response outputs are the difference in pulse before and after the stimulant.

There are always more goodness-of-fit statistics to consider such as R^2, the F-statistic, the Mean Squared Error (MSE), Cook's distance etc. We have selected the ones present in this table as the most telling about the fit of the data. Essentially, the lower the deviance residuals and Akaike's Information

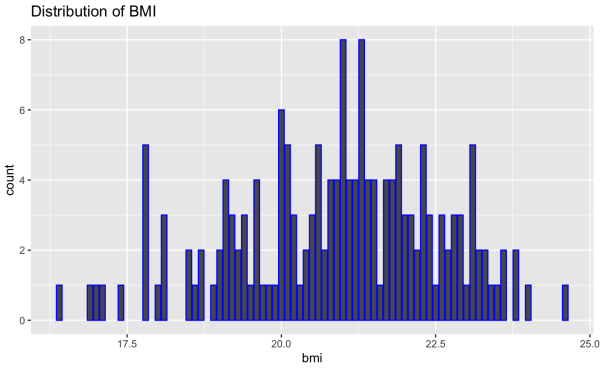

Figure 3: Distribution of BMI

Criteria (AIC), the better the fit of the data. Additionally we would prefer standard errors as close to zero as possible and p value below 0.05 (when p-value is less than significance level (¡ 0.05), we can safely reject the null hypothesis that the coefficient beta of the predictor is zero).

Although the statisticians' model is an improvement in the clinicians' initial model, there is still room for improvement in the deviance residuals, standard erros and AIC.

We tried fitting other models such as the Binomial and Poisson GLMs, with no improvement.

We proceeded to include the coffee factor as a second covariate of the statisticians' initial model and saw a significant improvement in the diagnostics. In the end, our best model was the **Gamma GLM with BMI and rest pulse as covariates, using the link function**.

Our best model has estimate $\beta = (-0.009247, 0.118886, -0.580313)$. The confidence interval of the beta parameters is shown in figure 4.

```
Confidence interval of beta:
                        2.5 %     97.5 %
(Intercept)     -1.23588849  1.2173942
dat$bmi          0.06322983  0.1745425
dat$coffee_reg1 -0.77127938 -0.3893467
```

Figure 4: Model fit for Statisticians' model

The figures 5, 6 and 7 show the way the model's fitted values compare to the data for different models.

Model	Deviance residuals	Standard Error	p-value	AIC
Clinicians' model with covariates BMI and Rest pulse	NA	(-2.2, 9.03, 5.86)	(0.04, 2.86e-16, 2.36e-8)	1077.589
Statisticians' Gamma GLM model with BMI covariate and inverse link function	59.132	(7.9, -6.4)	(7.04e-13, 1.41e-09)	981.67
Binomial GLM with BMI covariate and canonical link	59.132	(-19.36, 11.77)	(0.00, 0.00)	1184
Poisson GLM with BMI covariate and identity link	59.131	(-10.32, 13.69)	(0.00, 0.00)	1151
Gamma GLM with BMI and coffee factor covariates and inverse link	47.274	(0.08, 0.00, 0.01)	(9.22e-06, 0.00036, 9.99e-09)	944.45
Gamma GLM with BMI and coffee factor covariates and log link	46.759	(0.62, 0.02, 0.09)	(0.988, 4.05e-05, 1.2e-8)	941.96

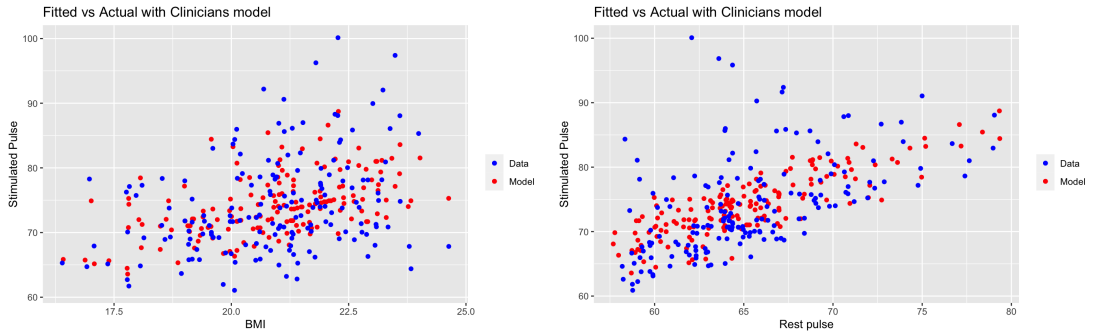

Figure 5: Model fit for Clinicians' model

4 Limitations

As we can see in the best model's fitted values figures, the residuals (vector of data - fitted values) are smaller for coffee consumers, meaning the model is better at predicting the stimulated pulse if the patient consumes coffee, because the stimulant had a higher influence on non-coffee drinkers, making their stimulated thus higher. This is one important factor to consider when running future clinical tests.

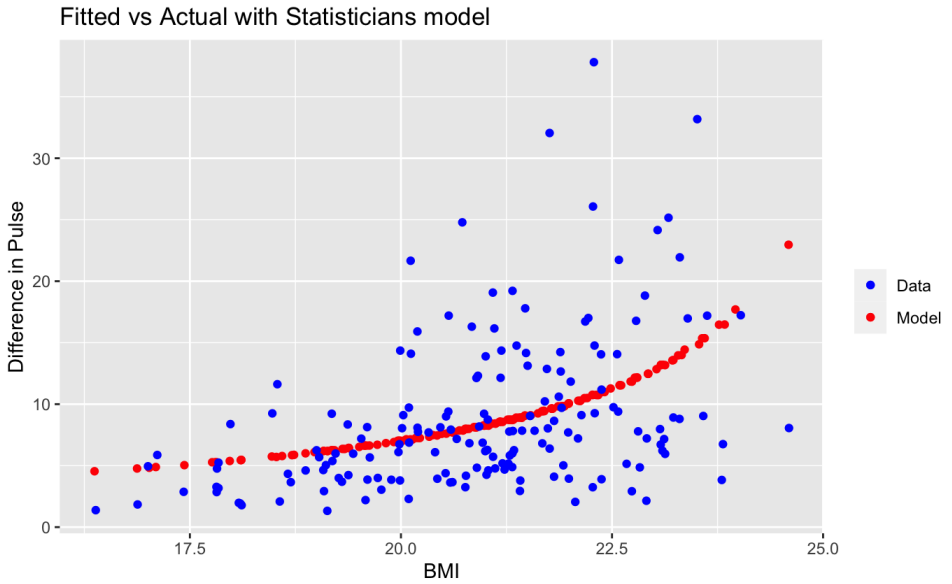

Figure 6: Model fit for Statisticians' model

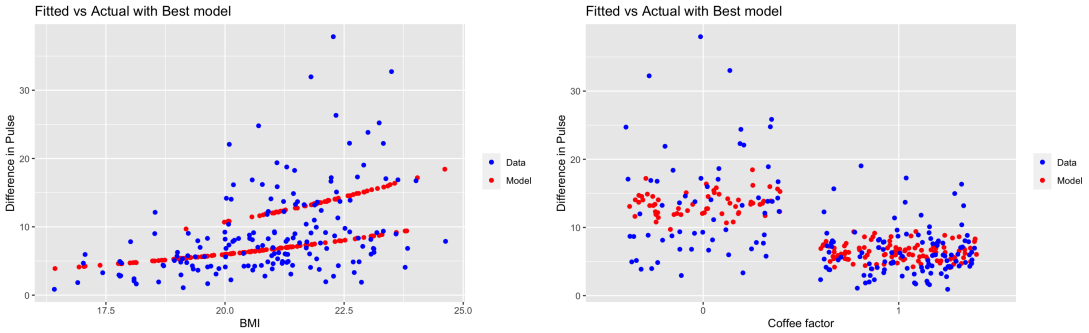

Figure 7: Model fit for our best model

For future models it is probably better to have different models for each category.

Another limitation is that we didn't have a perfectly balanced data as we had a few dropouts during the trial, affecting the ratio of coffee consumers.

Additionally, our data set isn't representative of a large population, not only due to simply the size of the data set, but also the range of BMIs. We were only dealing with health (or some underweight) patients, so the model mnight not perform as well on a data set of patients with higher BMIs.

Another thing to consider would be how well the data is able to predict a new set of data. The clinicians might want to consider running another trial with patients of similar health and test the model on this new data set.

5 Conclusion

Althought the Statisticians' model was an improvement on the clinicians' initial model, we have fitted a better-performing model: the Gamma GLM with covariates BMI and coffee factor, with log link function. Note that the model was better at predicting the data of coffee consumers, as non coffee consumers were for sensitive to the stimulant. In future it is worth considering having a model of each category of patients, and to sample a larger group of people.

6 Abstract: Numerical scheme used to fit the models

In the R code attached, we have used the inbuilt function lm() and glm() to fit our models. However I have also created my own function that mimics what glm() produces. Here is the code:

```
#Manually made GLM function
statglm <- function(d, tt = TRUE){ #inputs: d = dataset, tt is indicator for if we want a summary or not
  beta <- c(0.05,0) #initial estimate is 1/mean(Bmi), 0
  X <- cbind(1, d$bmi) #design matrix
  y <- d$stimulated_pulse-d$rest_pulse #response output
  jj <- 0
  #Inverse of the link function (in this case it's the inverse function too)
  inv.link <-function(u){
    1/u}
  #Deviance function = 2*(l(y) - l(mu))
  D <- function(u){
    a = -1-log(y)
    b = y/u + log(u)
    2*sum(a+b)}
  oldD <- D(inv.link(as.numeric(X%*%beta))) #Current deviance
  while(jj ==0){
    eta = X%*%beta
    mu <- inv.link(eta) #mu = 1/eta
    detadmu <- -1/(mu^2) #deta/dmu
    z <- eta +(y-mu)*detadmu #z
    w = mu^2 #weights
    lmod <- lm(z~d$bmi, weights=w) # regress z onto x with weights w
    beta <- as.vector(lmod$coeff) #new estimate of parameters beta
    newD <- D(inv.link(X%*%beta)) #new deviance
    control <- abs(newD-oldD)/(abs(newD)+0.1)
    if (isTRUE(control<1e-8)){
      jj<-1} #stop the algorithm when the control factor is small enough
    oldD <- newD}
  #Calculate statistics for diagnostics:
  J<- t(X)%*%diag(as.vector(w))%*%X
  invJ <-solve(J)
  beta.sd <- sqrt(as.vector(diag(invJ))) #Standard errors
  t_value<- beta/beta.sd #t value
  p_value <- 2*pt(-abs(t_value), df=nrow(d)-ncol(d)) #p value
  errors <- as.vector(y - inv.link(X%*%beta)) #Residuals
  RSS <- t(errors)%*%errors #Residual sum of squares
  sst <- t(y-mean(y))%*%(y-mean(y)) #Total sum of squres
  R2 <- 1- RSS/sst #R-squared
  AR2 <- 1- (1-R2)*(nrow(d)-1)/(nrow(d)-ncol(X)) #Adjusted R-squared
  MSE <- RSS/(nrow(d)-ncol(X)) #MSE
  MSR <- (sst-RSS)/(ncol(X)-1) #MSR
  F_value <- MSR/MSE #F-statistic
  coeff <- data.frame(beta, beta.sd, t_value, p_value)
  dev <- c(sign(y[1]-mu[1])*sqrt(D(inv.link(X[1,]%*%beta))))
  for (i in 2:nrow(d)){dev <- c(dev,sign(y[i]-mu[i])*sqrt(D(inv.link(X[i,]%*%beta)))) #Deviance residuals}
  dev <- summary(dev) #Deviance residuals
  phi <- newD/(nrow(d)-2) #estimated dispersion parameters
  AIC <- -2*sum(dgamma(y, 1/phi, 1/(phi*mu), log = TRUE)) + 2*2 #AIC
  if (tt == TRUE){
    cat("Deviance residuals:\n")
    print(dev)
    cat("\n")
    cat("Coefficients:\n")
    printCoefmat(coeff)
    cat("\n")
    cat("Residuals deviance: ",newD)
    cat("\n")
    cat("R-Squared: ", R2)
    cat(",\tAdjusted R-Squared: ", AR2,
      "\nF-statistic:", F_value,
      "with", nrow(d)-2-1, "DF")
    cat("\n")
    cat("AIC : ", AIC)}
  else{return(list("model" = inv.link(X%*%beta), "y" = y))}}
```

Report for the clinical team

We considered your initial model and the statisticians' model, and evaluated how good they were for predicting the stimulated heart rate using what we call goodness of fit measures. They are essentially statistics that measure the quality of a model's prediction. We managed to create a model that performed better that both of these: we used the Gamma GLM (like the statisticians), except we changed what we call the link function to the log function instead of the inverse, and added a second input, the coffee factor.

You can get this model with the following R code:

```
y <- data$stimulated_pulse - data$rest_pulse
fit <- glm(y~data$bmi + data$coffee_reg, family = Gamma(link = "log"))
```

Note that the model was better at predicting the difference in pulse for coffee drinkers, as non-coffee drinkers had higher sensitivity to the stimulant. For those who don't consume coffee, the difference in their pulse might not be linearly related to their bmi but rather quadraticly so it is worth considering having a different model for each type of consumer in future clinical trials.

You should also test the model on a new set of data taken for a similar population and see how accurate the prediction is.