STATISTICAL MODELLING 2 COURSEWORK

ANDREW MELVILLE 01200147

Date: 8 March 2020.

1. Aims

In this report we aim to evaluate the efficacy of several models that describe how BMI and regular coffee consumption affect an individuals response to $100mg$ of a stimulant. Firstly we explore a linear model fit by the clinical team, with BMI as our only covariate to predict an individual's stimulated pulse rate. Secondly we fit and evaluate the generalised linear model suggested by the statistician which instead predicts the absolute change in pulse rate from BMIs. Finally we search for any improvement on this in the space of generalised linear models with the response variable specified as being Gamma distributed. This is done by again looking for a motivated choice of response variable and considering the effects of coffee consumption.

2. Exploratory Analysis

First we use the summary command to display a description of the data.

```
1  rest_pulse      stim_pulse      bmi             coffee_reg
2  Min.   :58.00   Min.   :62.00   Min.   :15.90   0: 48
3  1st Qu.:61.00   1st Qu.:68.00   1st Qu.:19.40   1:109
4  Median :64.00   Median :71.00   Median :20.70
5  Mean   :65.26   Mean   :73.08   Mean   :20.61
6  3rd Qu.:68.00   3rd Qu.:76.00   3rd Qu.:21.80
7  Max.   :84.00   Max.   :98.00   Max.   :24.70
```

The summary shows that the pulse data is positive, with each quartile of the resting pulse being lower than that of the stimulated pulse. Additionally, we see that the BMI factor is distributed similarly to the distribution in the general population, and that the data is unbalanced with respect to the number of coffee drinkers, with much more regular consumers than not.

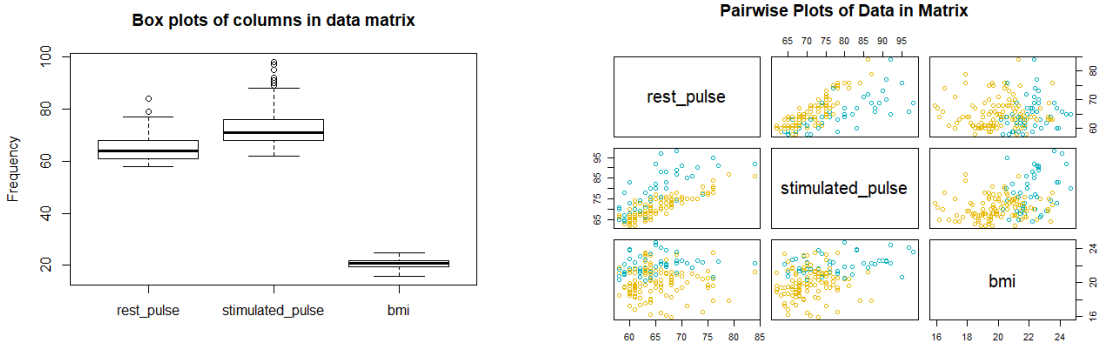

Figure 1. (a) Boxplots of the data that visualise the information given by R's 'summary' command. (b) Pairwise plots of the data in the matrix, with coffee drinkers plotted in blue, and non-coffee drinkers plotted in yellow.

Next we produce histograms and pairwise plots of each column in our data matrix. From this we hope to understand the nature of our data and uncover correlation between factors.
From Fig. 1(b) we see correlation between rest pulse and stimulated pulse as one would expect. As the pair plot shows little correlation between the resting pulse and BMI, it is valid to hope that including both as factors in the model will lead to a better fit of the model while remaining parsimonious.

3. Clinician's Linear Model

The first model we will consider is the clinician's linear model:

$$Y_i = \beta_0 + \beta_1 r_i + \beta_2 b_i + \epsilon_i \tag{3.1}$$

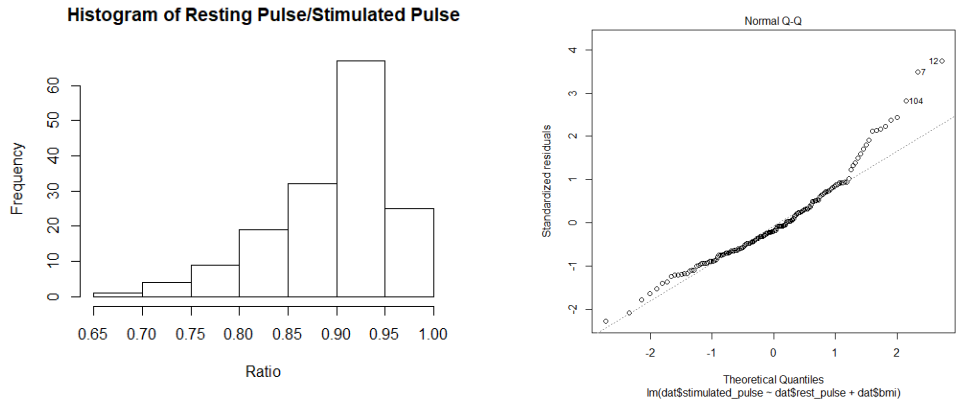

FIGURE 2. (a) Plot showing the distribution of the ratio increase of pulserate against the BMI of an individual. (b) Plot of the quartiles of the residuals against theoretical quartiles from a normal distribution.

where Y_i = stimulated pulse, r_i= resting pulse on day 1, b_i = the BMI in kg/m^2, and the ϵ i.i.d Gaussian random variables with mean 0 for individual i. This model assumes that the stimulated pulse is a linear combination of the individuals resting pulse and their BMI, and that an increase (or decrease) in either of these covariates is linearly proportional to a change in the mean response variable. Later, we will evaluate models that do not make these assumptions, and instead link a linear increase in the covariates to a multiplicative increase in the mean response.

The 'lm' function used to fit this model uses a traditional least squares scheme to calculate $\hat{\beta}$. As with all LSE, this is available in closed form:, with $\hat{\beta} = (X^T X)^{-1} XY$.

From this calculation, we find $\hat{\beta} = (-21.60, 0.98, 1.48)$, with significance values of $0.000984, < 2 \times 10^{-16}$, and 1.04×10^{-10} respectively. At first glance this appears to be evidence in favour of the model, but analysis of the residuals disagree. The assumptions made in the choice of the model mean that we expect our residuals to be approximately normally distributed. However, the comparison of the residual quartiles with the theoretical quartiles of a normal distribution shown in Figure 2(b) suggests that this is not the case.

Clearly the residuals are not normally distributed as we would expect from a model such as (3.1). Additionally, it appears as though the residuals have larger variance for larger fitted values. This motivates the fitting of the Statistician's GLM, which allows for a different choice of response distribution, along with a more flexible relationship between the covariates and the response.

4. STATISTICIAN'S GLM

The second model we will consider improves on the Clinician's Linear model by instead taking the difference in pulse rates as the response variable. Doing so no longer assumes a linear relationship between the resting pulse and covariates and the mean response. The model can be expressed as:

$$E(Y_i) = \mu_i = \frac{1}{\eta_i}, \qquad \eta_i = \beta_0 + \beta_1 b_i + \epsilon_i$$

with the response variable Y_i now considered as the *increase* from resting pulse to stimulated pulse. Implicit in the expression above is the choice of *link function* as the inverse link function. This makes the assumption that, from an estimated intercept term, individuals with a larger BMI will on average experience an inversely proportional decrease in the effect of the stimulant on raising their pulserate.

Unlike Linear Models like the one fitted above, Generalised Linear Models do not offer a closed-form solution for $\hat{\beta}$. Instead we must implement an Iterated Weighted least Squares (IWLS) method that

begins with an initial estimate for $\hat{\beta}$ and proceeds maximise the likelihood function by performing a weighted regression. The new estimate for β is computed as

$$\hat{\beta}_{n+1} = (X^T W_n X)^{-1} X^T W_n z_n, \tag{4.1}$$

where the weights $W_n^{-1} = \left(\frac{d\eta}{d\mu}\right)_n^2 V_n = \left(\frac{1}{\mu^2}\right)^2 \mu^2 = \frac{1}{\mu^2}$ are computed by substituting our current $\hat{\mu}_n = g^{-1}(X\hat{\beta}_n)$ for μ. This algorithm requires an initial choice for $\hat{\beta}_0$ to begin the sequence. For this, we must use an appropriate choice; too large a β may result in divergence. Hence, our choice of initial value is motivated by maximising the likelihood function of the y_i as a function of $\mu = g^-1(\eta)$, with $\theta(\mu) = \frac{1}{\mu}$, $b(\theta) = -\log(-\theta)$, and $a(\phi) = \phi$ all taken from the Gamma distribution expressed in exponential family form:

$$\begin{aligned} l(\mu, \phi;\ y) &= \log\left(\Pi_{i=1}^n \exp\left\{\frac{y_i\theta(\mu_i) - b(\theta(\mu_i))}{a(\phi)} + c(y,\phi)\right\}\right) \\ &= \sum_{i=1}^n \left(\frac{\frac{-y_i}{\mu_i} - \log(\mu_i)}{\phi} + c(y,\phi)\right) \end{aligned}$$

Here we assume that the μ_i are constant (i.e. $\mu_i = \mu$ for some μ). We will see later that this is equivalent to assuming a uniform distribution of weights for the first round of weighted least squares regression. This allows us to take the derivative of the log-likelihood with respect to μ and set $\frac{dl}{d\mu} = 0$ to find a minimum:

$$\begin{aligned} \frac{dl}{d\mu} = \frac{1}{\phi}\sum_{i=1}^n \frac{y_i}{\mu^2} - \frac{1}{\mu} &= 0 \\ \implies \frac{1}{n}\sum_{i=1}^n y_i &= \mu \end{aligned}$$

in other words, that our choice for μ is just the mean value of the observed response. Choosing μ in this way is equivalent to choosing $W = \frac{1}{\mu^2} = \frac{1}{\bar{y}^2}$.

While we now have a starting point for our algorithm, we must decide when the algorithm has deemed to converge. From lectures, a good stopping criterion is when the change in deviance D from estimate n to $(n+1)$ falls below some control value. First, we clarify what we mean by deviance of a model:

$$D = \frac{2}{\phi}\left\{l(y, \phi;\ y) - l(\hat{\mu}, \phi;\ y)\right\}, \tag{4.2}$$

or equivalently, the scaled difference in log-likelihoods of the saturated model and the current model fit by the algorithm. In our implementation, we will use the scaled deviance, defined as $D^* = \phi D$. For our Gamma model with inverse link function, this is just $D^* = 2\sum_{i=1}^n \log\left(\frac{\mu_i}{y_i}\right) + \frac{y_i - \mu_i}{\mu_i}$.
This algorithm was implemented in R (see accompanying code file), and converged after 5 iterations to $\hat{\beta} = (0.64953985 - 0.02456083)$, with final deviance value $D^* = 51.575$. Figure 3(a) shows the fit of the line against the data.
Here, we calculate the standard errors, t values and significance values for these estimates. Using $\hat{\phi}_D = \frac{D}{n-p}$, we find estimates for the standard errors $\hat{\beta} = (0, 5.613288 \times 10^{13})$, with the estimate of 0 likely due to a very small value falling victim to rounding. Using $\hat{\beta} \dot{\sim} N(\beta, \mathcal{J}^{-1})$ and our estimate of the standard errors, we can construct 95% confidence intervals for $\hat{\beta}$ as $(-0.78402845, -0.51505125)$ for $\hat{\beta}_1$ and $(0.01844631, 0.03067536)$ for $\hat{\beta}_2$.

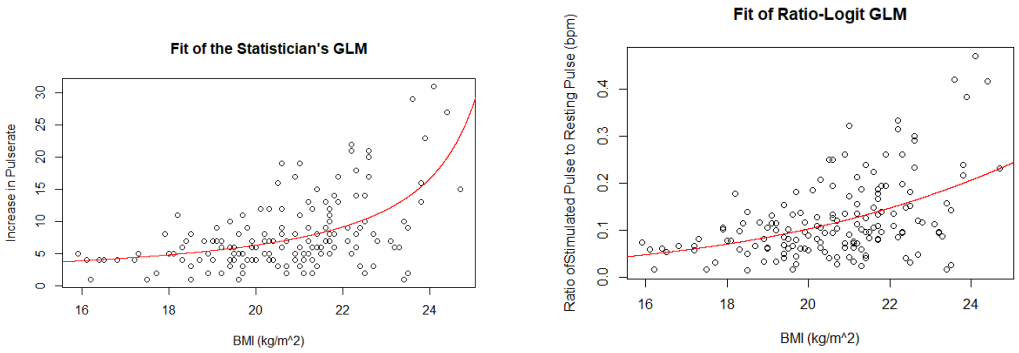

FIGURE 3. (a) Plot of the individual's BMI against their observed increase in pulse after 100mg of a stimulant. Overlaid is the line fit by the Statistician's GLM. (b) Plot of the individual's BMI against the observed pulse ratio r_i after 100mg of a stimulant. Overlaid is the line fit by the Ratio-Logit GLM.

5. Model Exploration

In this section, we attempt to find a GLM that improves on the one suggested by the Statistician by using R's glm function. This function fits the model using the same numerical process outlined earlier, which we implemented to fit the Statistician's GLM. Firstly, we will first consider the ratio of the resting pulse to the stimulated pulse as our response variable, $r_i = \frac{p_{r_i}}{p_{s_i}}$ with p_{r_i} resting pulse and p_{s_i} stimulated pulse of individual i. The exploratory plot (Figure 2(a)) of this variable suggests a Gamma distribution is valid, but taking the ratio as the response variable motivates a change in link function. Here, we choose the logit link function:

$$\mu_i = g^{-1}(\eta_i) = \frac{1}{1+e^{-\eta_i}}$$

which is a pertinent choice as $g : \mathbb{R} \to (0,1)$. This necessitates the assumption that an individual's pulse rate does not decrease after receiving 100mg of the stimulant. Additionally, the function maps linear changes in BMI to an exponential increase in pulse rate ratio up to a certain point, after which linear increases yield logarithmically diminishing returns. However, this means that extrapolation of this model may be invalid for values outside of the range of the given data (BMI values above 24) as we do not have data outside of this range to validate the choice of model outside of this range The linear predictor is thus defined:

$$\eta_i = \beta_0 + \beta_1 b_i \tag{5.1}$$

with b_i as above. Fitting this model, we find:

```
1                Est       Std. Err    t val    Pr(>|t|)
2 (Intercept) -6.28495      0.55350 -11.355  < 2e-16 ***
3 dat$bmi      0.20567      0.02695   7.632 2.21e-12 ***
```

At first glance we can see that the logit model is influenced much less by the cluster of data with both high BMI and pulse increase, at the cost of increasing their residuals. It is not valid to compare the deviance residuals of these two models as they do not fit the same response variable. Also, Figure 4 shows that no point has a Cook's distance large enough to warrant removal from the model.
Exploratory analysis shows that our sample has more than twice as many coffee drinkers than non-coffee drinkers, with the latter group having a much larger average BMI. This will almost certainly affect the fit of our GLM, so we may explore how including this information as a factor changes our fit. Firstly, we attempt to fit a model which constrains both groups' linear predictors to have the same intercepts. This is motivated by both groups having similar resting pulse rate distributions, but drastically different responses to the stimulant.

$$\eta_i = \beta_{0_j} + \beta_1 b_i \qquad \text{for } j \in \{0,1\} \tag{5.2}$$

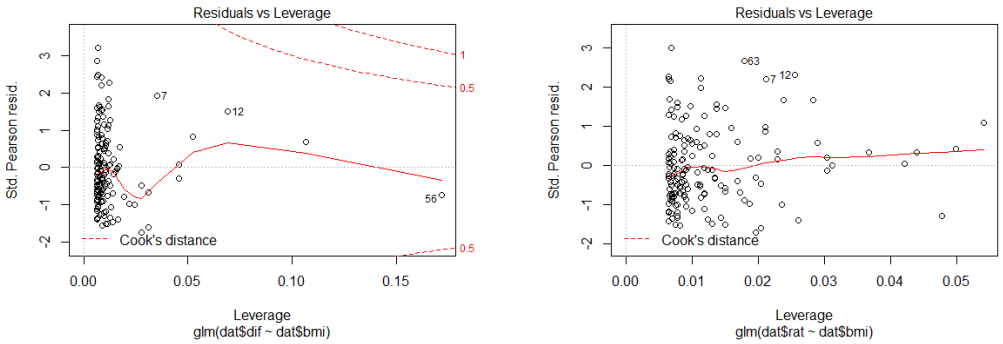

FIGURE 4. Plots of the leverage of each data point against their Standardised Pearson residuals. Left shows the plot for the Statistician's GLM, while the right shows the plot for the Ratio model fitted in this section.

with $j = 1$ and $j = 0$ denoting coffee drinkers and non coffee drinkers respectively.
The last model we fit assumes both groups have the same response to the stimulant, but fits distinct intercept terms for each group's linear predictor:

$$\eta_{i_j} = \beta_0 + \beta_{1_j} b_i \qquad \text{for } j \in \{0, 1\}$$

with $j = 1$ and $j = 0$ denoting coffee drinkers and non coffee drinkers respectively.
The Coffee Intercept model reports an AIC of -506.62, while the Coffee Coefficient model reports -508.76. Additionally, the residual deviance for each of these models is reported as 0.43 and 0.42 respectively. As the Coefficient model reports the lowest AIC out of all models fitted, we choose this as our final model.

6. CONCLUSION

The most chosen model returned an estimate of 4.276 for the shared intercept, with a 95% confidence interval of $(3.0760494, 5.51059349)$, -0.08845366 for the Coffee Drinker's coefficient, with a 95% confidence interval of $(-0.1478422, -0.02941564)$, and -0.11932025 for the coefficient of the Non-Coffee Drinkers, with a confidence interval of $(-0.1745322, -0.06521717)$. Whilst such wide confidence intervals point to the need of a much larger, balanced sample, the lower coefficient for non-coffee drinkers suggests a larger response to the effects of the stimulant than that experienced by regular caffeine consumers. This matches our intuition, as we would expect some tolerance to stimulants from this group.

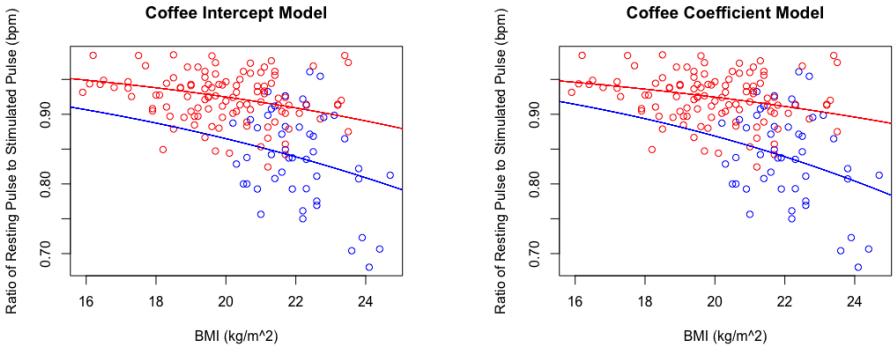

FIGURE 5. Plots of the lines fit to each model that considers regular coffee consumption as a factor. Coffee drinkers are shown in blue and non coffee drinkers in red.

7. Guidance for Clinicians

The largest issue with the data provided is the unbalanced data set. With twice as many coffee drinkers as non-coffee drinkers included, any model that is fit will struggle to report the effects of the stimulant on the smaller group. To combat this, ensure the workable dataset is made up of an approximately equal number of both groups. This difference is clearest in the range of BMI between the two groups, with Coffee drinkers having on average a much lower BMI. As a result, any conclusions drawn from the linear model will not be able to distinguish between an individual with low BMI for their group but no tolerance to stimulants and an individual with high BMI for their group and a tolerance to stimulants.

The linear model is not the wisest choice of model as it relates an increase in body mass to a proportional increase in reaction to the drug, which we know anecdotally to be false, nor does it account for tolerance to stimulants built up by regular coffee consumption. Instead we may attempt to fit a generalised form of the linear model, which corresponds an increase in body mass to a non-linear change in pulserate. Again, this relates our intuition about how heavier individuals require more of a stimulant to experience the same effects.

Using resting pulse as a covariate results in an estimate of it's coefficient as close to 1, which in real terms means that individuals with higher resting pulse rates will have higher stimulated resting pulses, and by the exact same amount. As a result, this covariate does little to explain the individual's health profile and it's effect on their response to the stimulant. As this is a stated aim of the study, one may use the resting pulse rate information to predict the *proportional* change in heart rate. In this way we may hope to capture information about both pulse rates in a useful way.