Statistical Modelling 2

Lorenz Wolf

March 9, 2020

1. Aims

In this report we aim to explore the effect of an artificial stimulant on heart rate and how this varies for individuals with different health profiles. Through data exploration and the analysis of different model fits we want to gain an understanding of the correlation between the bmi and the stimulant's effect on the heart rate. Since the stimulant targets the same receptors as caffeine it will be of special interest to investigate the effect of regular coffee consume on the effectiveness of the artificial stimulant.

2. Methods

2.1 Normal Linear Model

To model the response $\mathbf{Y}$ we specify the model as $\mathbf{Y} = \mathbf{X}\beta + \epsilon$. Here $\mathbf{X}$ is the design matrix containing covariates, β is the parameter vector to be estimated and ϵ the error term assumed to follow a Multivariate Normal distribution with mean μ=0 and covariance matrix Σ^2=σ^2 $\mathbf{I}$.

2.2 Generalized Linear Model

In a generalized linear model we assume our components of response $\mathbf{Y}$ to be independent and from the same distribution of the exponential family with mean E($\mathbf{Y}$)=μ. We then use the covariates $\mathbf{x_1}, ..., \mathbf{x_p}$ to form the linear predictor

$$\eta = X\beta = \sum_{j=1}^{p} \mathbf{x_j}\beta_j$$

A link function g(.) is used to relate μ to the linear predictor by $\eta_i = g(\mu_i)$ for $i = 1, ..., n$. Here g can be any monotonic differentiable function. The link is needed to map the linear predictor $\eta_i \in \mathbb{R}$ to the possibly restricted space of μ_i.

2.3 Numerical Scheme to fit GLM

To fit the GLM we use the iterative weighted least squares algorithm. As described in lectures the algorithm is as follows:

1. Given a current estimate $\hat{\beta}$, form the linear predictor $\hat{\eta}$ and the fitted values $\hat{\mu}$
2. Form the adjusted dependent variable: $z_i = \hat{\eta}_i + (y_i - \hat{\mu}_i)\frac{d\eta}{d\mu}\Big|_{\mu=\hat{\mu}_i}$
3. Form the estimated weights: $\widetilde{w}_{ii}^{-1} = \left(\frac{d\eta}{d\mu}\right)^2 V(\mu)\Big|_{\mu=\hat{\mu}_i}$
4. Regress z_i on x_i with weights $\widetilde{w}_{ii}^{-1}$ and obtain new estimate $\hat{\beta}$
5. Repeat steps 1 to 4 until convergence

Let us now apply this algorithm to our particular case of a gamma distribution with the inverse link (canonical link). For the gamma distribution we have

$$f(y, \alpha, \beta) = exp\Big(- \beta y - \frac{-log(\beta)}{\frac{1}{\alpha}} + (\alpha - 1)log(y) - log(\Gamma(\alpha))\Big)$$

Now using the transformation $\phi = \frac{1}{\alpha}$ and $\theta = \frac{\beta}{\alpha}$ we obtain

$$f(y, \phi, \theta) = exp\Big(\frac{\theta y - log(\theta)}{-\phi} + \frac{log(\phi)}{-\phi} + (\frac{1}{\phi} - 1)log(y) - log(\Gamma(\frac{1}{\phi}))\Big)$$

We therefore have that $a(\phi) = -\phi$, $b(\theta) = log(\theta)$, and $c(y,\phi) = -\Big(\frac{log(\phi)}{-\phi} + (\frac{1}{\phi} - 1)log(y) - log(\Gamma(\frac{1}{\phi}))\Big)$. It then follows that $z_i = 2\eta_i - y_i\eta_i^2$ and $\widetilde{w}_{ii}^{-1} = \frac{-1}{\mu_i^2}$.

Now we still need an appropriate starting point. We have $\eta = \mathbf{X}\beta$ and so the least squares estimator is $\hat{\beta} = (X^TX)^{-1}X^T\eta$. Since at the starting point we only have the observations y_i, we set $\mu_i = y_i$ so that $\eta_i = \frac{1}{\mu_i} = \frac{1}{y_i}$ as initial guess. As criterion for convergence we use the difference in the deviance of the model with the previous estimate and the new estimate, defined by $D = 2\phi(l(y,\phi,y) - l(\hat{\mu},\phi,y))$. So we need to find the log likelihood of the saturated and the estimated model.

$$l(\mu,\phi,y) = \sum_{i=1}^{n} \frac{-y_i}{\phi\mu_i} + \Big(\frac{1}{\phi} - 1\Big)log(y_i) - \frac{log(\mu_i)}{\phi} - nlog\Big(\Gamma\Big(\frac{1}{\phi}\Big)\Big) \tag{1}$$

$$= \sum_{i=1}^{n} \Big[-\frac{1}{\phi}\Big(\frac{y_i}{\mu_i} - log\Big(\frac{y_i}{\mu_i}\Big)\Big) - log(y_i)\Big] - nlog\Big(\Gamma\Big(\frac{1}{\phi}\Big)\Big) \tag{2}$$

Now for the saturated model we set $\mu_i = y_i$ and for the fitted model $\mu_i = \hat{\mu}_i$. Thus, we obtain the deviance to be

$$D = 2\Big(-n + \sum_{i=1}^{n} \frac{y_i}{\hat{\mu}_i} - log\Big(\frac{y_i}{\hat{\mu}_i}\Big)\Big)$$

The convergence criterion is then set to be

$$\frac{|D_{new} - D_{old}|}{|D_{new}| + 0.1}$$

which we check each iteration. Finally, we can first estimate the dispersion with $\hat{\phi}_D = \frac{D}{n-p}$ and then use the results from the IWLS to estimate $cov(\hat{\beta}) \approx \hat{\phi}(X^T\widetilde{W}X)^{-1}$ and obtain the standard errors s_i as the square root of the diagonal entries. We can then obtain 95% confidence intervals for the parameters $(\beta_i - 1.96s_i, \beta_i + 1.96s_i)$.The implementation of this algorithm for a gamma distribution with inverse link is displayed below.

Listing 1: Manual IWLS for Gamma with Inverse Link

```
D <- function(p){2* sum(y/p - log(y/p) -1)} #function for deviance
# function to compute convergence criterion from old and new deviance
convergence <- function(o,n){abs(n-o)/(abs(n)+0.1)}
# least squares estimator as initial guess
eta <- 1/(dat$stimulated_pulse-dat$rest_pulse)
beta_start <- solve(t(X)%*%X)%*%t(X)%*%eta
beta <- beta_start #initial guess

oldD <- D(y) #initial deviance
conv_crit <- 1 #initialize convergence criterion
# IWLS
while(conv_crit>10^(-8)){
 eta <- X%*%beta #estimated linear predictor
 mu <- 1/eta #estimated mean response
 z <- 2*eta - y * eta^2 #form the adjusted variate
 w <- (mu^2) #weights
 lmod <- lm(z~x, weights=w) #regress z on x with weights w
 beta <- as.numeric(lmod$coeff) #new beta
 print(beta) #print out the beta estimate every iteration
 newD <- D(1/(X%*%beta)) #compute new deviance
 conv_crit <- convergence(oldD, newD) #compute convergence criterion
 oldD <- newD #update old deviance
 }
# estimate deviance and obtain standard errors for estimates
phi_hat <- oldD/(length(dat$bmi)-2) #estimate of dispersion
cov_beta <- phi_hat * solve( t(X) %*% diag(as.vector(w))%*%X) #cov(beta)
beta.sd <- sqrt(as.vector(diag(cov_beta))) #standard errors of beta
```

To evaluate the model fit we consider residual plots and for model comparison we use the residual deviance, ANOVA, and AIC (which uses the number of parameters to penalise more complex models) defined as:

$$AIC = -2l(\hat{\beta}) + 2p \text{ ,where p is the number of parameters}$$

3. Data Exploration

According to the study design we should have an equal number of coffee drinkers and non-coffee drinkers. However, the number of coffee drinkers is 124, while there are only 53 non-coffee drinkers. Considering that the stimulant targets the same receptors as caffeine this imbalance might cause an underestimation of the effect of the stimulant, since intuitively regular exposure to caffeine leads to an adaptation to stimulation of the corresponding receptors. We would need to consult with one of the clinicians for further input on this matter.

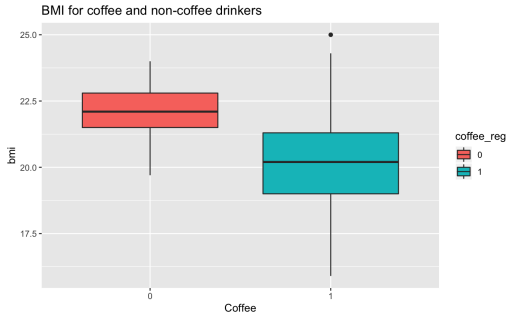

Figure 1: We can observe that among the subjects non-coffee drinkers tend to have a larger bmi.

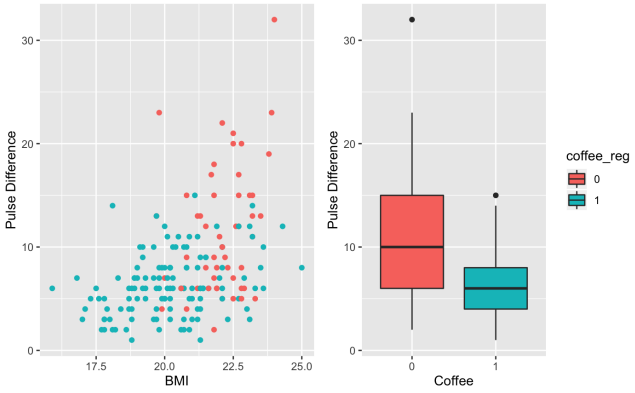

Figure 2: Scatter plot: The difference between the pulses seems to increase with increasing bmi. Box plot: The difference of pulses appears to be bigger for non-coffee drinkers than for coffee drinkers.

The observations of figure 1 and figure 2 are crucial for the analysis. For example suppose the stimulant has a larger effect on non-coffee drinkers, then this could lead to the wrong conclusion that the stimulant has larger effect on people with higher bmi. Hence we want to fit a model taking the coffee consume of a subject into account and see what happens. We don't want to confuse the influences of bmi and coffee consume on the effectiveness of the artificial stimulant.

4 Results

Model 1: Linear Model

The clinicians fit a normal linear model with the stimulated pulse as response and the rest pulse and bmi as predictors. It yields an AIC of 1028 and all predictors have small p-values indicating significance. The QQ plot shows that the residuals do not follow a normal distribution. This suggests that a normal linear model is not suitable for this data. Considering the residual plot we can observe a trend in the variance of the residuals, suggesting heteroscedasticity which violates the assumption of homoscedasticity in the residuals (NTA). Looking at the Cooks distance we do not observe any outliers with strong influence on the model fit. Furthermore, since the clinicians are interested in the effect of the artificial stimulant, which is described by the difference between stimulated pulse and rest pulse we should treat this difference as the response variable. This reduces the complexity of the model and makes it more interpretable.

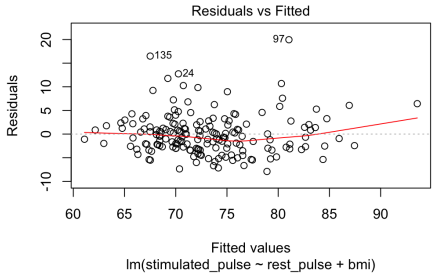

Figure 3:

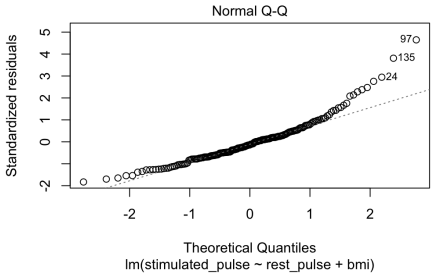

Figure 4:

Model 2: GLM suggested by statistician

The statistician suggests to fit a GLM with the stimulated pulse - rest pulse as response and bmi as predictor. This model addresses both problematic aspects of the clinician's model since the gamma distribution can model mean and variance. The model yields an AIC of 958, a residual deviance of 48, and the p values for the intercept and bmi predictor are very small indicating high significance. The standard errors are small. Considering the plot of the residuals we can not observe an obvious pattern suggesting homoscedasticity. The QQ plot looks much better than for Model 1, with some small deviance from the theoretical quantiles at the extremes. Considering the Cook's distance we can not observe any outliers with extreme influence on the model fit. However, this model does not take into account the potential influence of coffee consume. As discussed in the data exploration this could drastically change the results and lead to wrong conclusions regarding the effect of the stimulant on people with different health profiles.

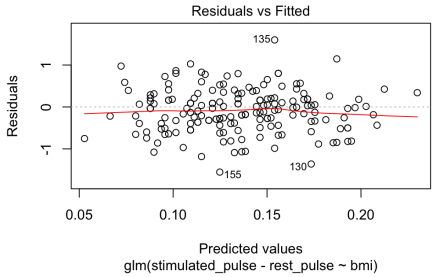

Figure 5:

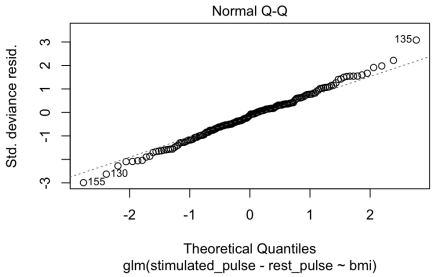

Figure 6:

Model 3: GLM with coffee consume

We now fit a GLM with both bmi and coffee consume as predictors. Model 3 yields an AIC of 941, a residual deviance of 43, and both the covariates and the intercept are significant. The residual plots and the QQ plot look as good as for model 2, if not slightly better. The small p-values for both predictors suggest that bmi and coffee consume are statistically significant predictors. This model takes into account the fact that more of the subjects with higher bmi are non-coffee drinkers, thus providing the clinicians with better insights. Comparing the bmi coefficient (-0.014) to the one in Model 2 (-0.02) we can conclude that Model 2 seems to overestimate the effect of bmi. Both smaller AIC and residual deviance as well as ANOVA with a chi-squared test suggest that Model3 is an improvement.

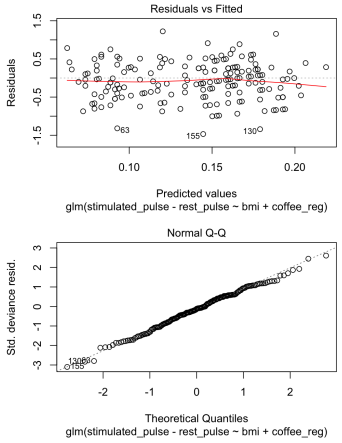

Figure 7: Can not observe pattern in residual plot. Standardized deviance residuals close to theoretical quantiles

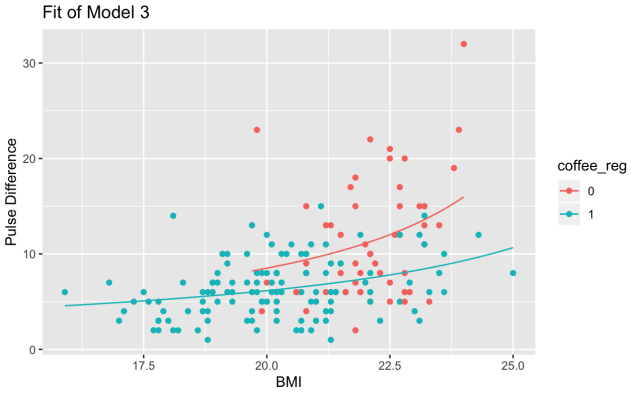

Figure 8: We show the fit of model 3 for coffee and non-coffee drinkers in the range of bmi for which we have data. Note the significant difference between the predictions depending on coffee consume.

We have tested further generalized linear models with Gamma distribution and the link functions $\eta_i = log(\mu_i)$ and $\eta_i = \mu_i$, but those did not lead to any better fitting models. We continue with Model3 and obtain 95% confidence intervals for the parameter estimates of model 3:

- Intercept($\hat{\beta}_1$): (0.256,0.526)
- bmi ($\hat{\beta}_2$): (-0.02, -0.008)

- coffee ($\hat{\beta}_3$): (0.026, 0.065)

How can Model3 be interpreted? We have $\hat{\mu}_i = \frac{1}{\hat{\eta}_i} = \frac{1}{\hat{\beta}_1 + x_{bmi}\hat{\beta}_3 + x_{coffee}\hat{\beta}_3}$. Now since $\hat{\beta}_2 < 0$ we have that an increase in bmi leads to an increase in $\hat{\mu}_i$ and thus a higher expected difference in pulses. On the other hand, since $\hat{\beta}_3 > 0$, regular coffee consume leads to a decrease in $\hat{\mu}_i$ and a lower expected difference in pulses. This effect can be observed in figure 8. Due to this relationship we have that for coffee drinkers the model blows up at a bmi of 31.8 (non-coffee drinkers 28.5). This is a clear limitation of the model. We can not extend our model for larger bmi getting close to those crucial values. Note that our data does not support the model for bmi larger than 25 anyways. The experimental design could be improved in the following ways:

1. Obtain more data, and ensure that coffee drinkers and non-coffee drinkers are evenly distributed across bmi and equally represented. Could for example ensure same number of subjects for bins of bmi.

2. More detailed data on the coffee consume, for example on a scale. This could give more insight into the relationship to coffee consume.

3. It could be interesting to obtain several measurements over time for each subject, since e.g. the resting heart rate might be higher on a stressful day. This could also provide an insight into how the body reacts to the stimulant after several times of taking it.

5 Conclusion

Model 1 does not appear to be a good fit and assumptions of the normal linear model are not satisfied. While Model 2 shows strong improvements over the linear model in terms of satisfaction of its assumptions, it fails to model a crucial feature of the data and overestimates the influence of bmi. As investigated in the data exploration the subjects with a larger bmi and larger pulse difference are mainly non-coffee drinkers. Thus, we fit a model including the regular coffee consume as predictor in addition to the bmi. We find that both predictors are highly significant. The data suggests that a regular coffee consume leads to a smaller effect of the stimulant. Furthermore, we find evidence that an increase in bmi leads to an increased effect of the stimulant.

Summary for Clinicians

There are 3 main problems with the model you initially tried to fit:

1. The relationship between the stimulated pulse and the rest pulse and bmi is not linear. Therefore it is not suitable to model stimulated pulse as $\hat{\beta}_1 + \hat{\beta}_2 pulse_{rest} + \hat{\beta}_3 bmi$.

2. When fitting the model you assume that the stimulated pulses measured come from a certain distribution. This assumption includes equal variance and independence which the data does not satisfy.

3. You did not take into account the effect of coffee consume on a subject's reaction to the stimulant.

4. Treating the difference between stimulated pulse and rest pulse as the outcome will still give you all the information you need to investigate the effect of the stimulant, but makes the model simpler.

We fit a different model, which uses the bmi and the coffee consume to model the difference between stimulated pulse and rest pulse. Given a bmi and coffee consume we predict the expected difference in pulses as $\frac{1}{\hat{\beta}_1 + x_{bmi}\hat{\beta}_3 + x_{coffee}\hat{\beta}_3}$. We could find the following main trends:

1. The coffee consume appears to play a significant role in understanding the effect of the stimulant. The stimulant seems to have a smaller effect on coffee drinkers.

2. The effect of bmi on the difference of pulses appears to be equally important. A larger bmi seems to lead to an increase in the effect of the stimulant

It is important to note that your data only really supports the model for non-coffee drinkers with a bmi between 19.7 and 24, for coffee drinkers between 15.9 and 25. We recommend to obtain more data (in which the coffee drinkers and non-coffee drinkers make up equal proportions of the data and are more evenly distributed) for more significant results.