Stat 437
December 14, 2022

# An Analysis of the Correlation Between Brain Structure and Political Orientation

Kimberly Peever

*Department of Mathematics and Statistics, University of Alberta, Canada*

kpeever@ualberta.ca

## 1. Introduction

A psychological experiment was conducted to examine the relationships between the volume of certain brain regions and political views. Specifically, the brain regions of the amygdala and anterior cingulate complex (acc) were studied. These regions of the brain are responsible for emotional processing as well as decision making. Research in this area is key for developing an understanding of the biological components that affect an individual's political orientation (Kanai *and others*, 2011). Kanai *and others* (2011), studied this relationship and they concluded that increased activation in these brain regions are a possible factor for the formation of political attitudes. In this report, the relationship between brain anatomy and political orientation is investigated and modelled. An exploratory analysis is conducted to explore any univariate or bivariate patterns before constructing any models. This is followed by fitting the data to both a logistic regression model and a generalized additive model. These models are then compared by appropriate model fitting and prediction analysis tools to find which model most accurately fits the data presented.

## 2. EXPLORATORY ANALYSIS

The dataset presented contains 90 observations. Each observation is sampled from a university student and consists of the variables; acc, amygdala, and orientation. Acc and amygdala are recorded as the residuals from the predicted volume, after adjusting for height, sex, and similar body-type variables. Orientation is coded as an ordinal variable on a five-point scale from 1 (very conservative) to 5 (very liberal), based on each student's score. All observations were complete in the dataset. The package 'mgcv' was used to fit the generalized additive model.

### 2.1 *Univariate analysis*

First, a univariate analysis was conducted to examine if there are any patterns or trends in the data. Histograms are provided in Figure 1 of the distribution of the volumes of amygdala and acc.

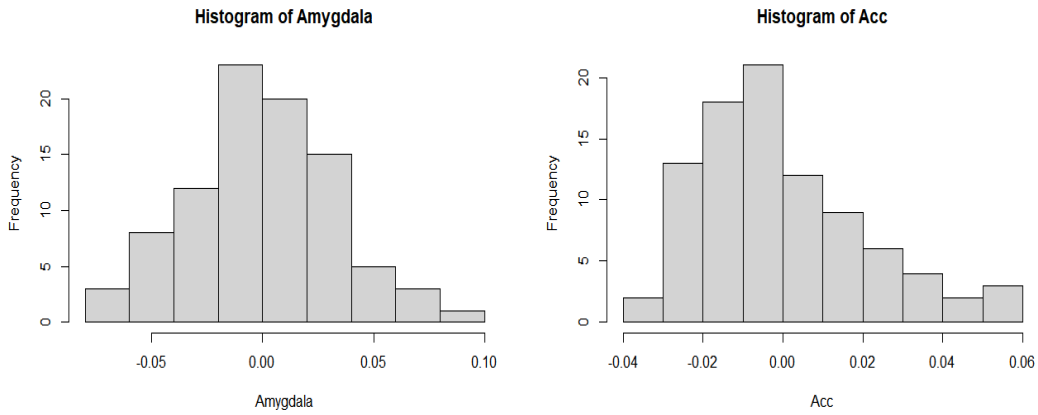


Fig. 1. Histograms of Amygdala and Acc, showing the frequency of observed volumes recorded in the amygdala and acc respectively.

The distribution of the volume in the amygdala appears to be normally distributed, while the volume of acc is slightly left skewed. It is possible acc activation rates are not normally distributed, this is not a cause for concern as the data can be fitted to models that do not require

the normality assumption.

The response variable orientation is plotted in Figure 2 with counts to get a sense of how many students fall where on the political spectrum. Here it can be seen that the majority of students fall on the moderate to liberal side of the spectrum, making the distribution right skewed. Only 13 students are moderately conservative and none are very conservative. This could pose a potential threat to the analysis because not only is the sample size small at 90 observations, but having a skewed frequency of responses may impact how the selected models perform.

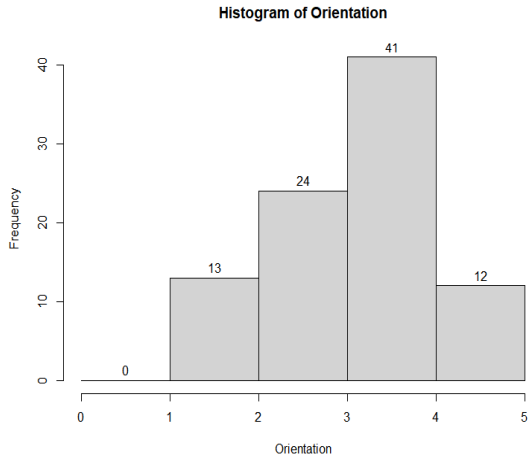


Fig. 2. Histogram of orientation, showing the frequency of responses for each value of orientation.

## 2.2 *Bivariate Analysis*

Next, a bivariate analysis is performed to get a general idea of how these variables interact with each other. First, the relationship between the volumes of amygdala and acc is investigated.

The correlation between the two brain regions was found to be -0.128484. Looking at Figure 3 there does not seem to be a pattern or obvious relationship between amygdala and acc. This low correlation is beneficial as multicollinearity can be an issue for obtaining reliable results.

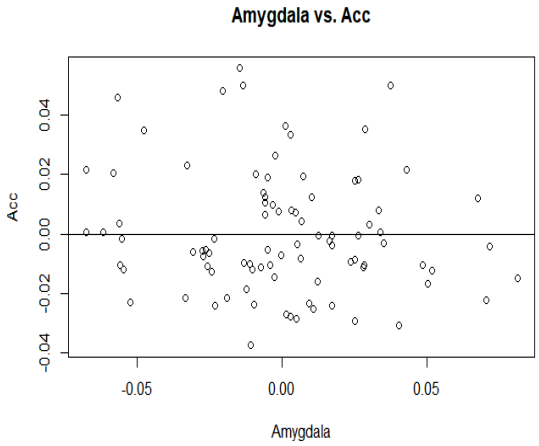


Fig. 3. Scatterplot of Amygdala and Acc, showing the relationship between the volumes in brain regions of amygdala and acc.

The correlation between amygdala and orientation was calculated using the Spearman's coefficient and was found to be -0.1998061, indicating a slightly negative relationship with orientation. Spearman's coefficient is a more accurate way to calculate correlation for an ordinal response variable. The correlation between acc and orientation was found to be 0.2923142, suggesting a slight positive relationship. It is difficult to say without fitting a model if these relationships are accurate but they provide an idea that acc may be more related to higher score on the spectrum and amygdala may possibly be related to lower scores. Plots of these relationships are provided in Figure 4.

Investigating these relationships further, case-resampling bootstrapping is conducted and 95% confidence intervals are obtained for these correlations. Case-resampling is used here due to the robustness against model assumptions. The data may not be normally distributed and the response variable is ordinal so case-resampling will provide a more reliable result for non-parametric data. For amygdala, the confidence interval of (-0.38663481, 0.01294573) was found. There is a strong possibility that amygdala has a negative correlation with orientation, however, due to 0 being contained in the interval, there is a small possibility that there is a positive or insignificant

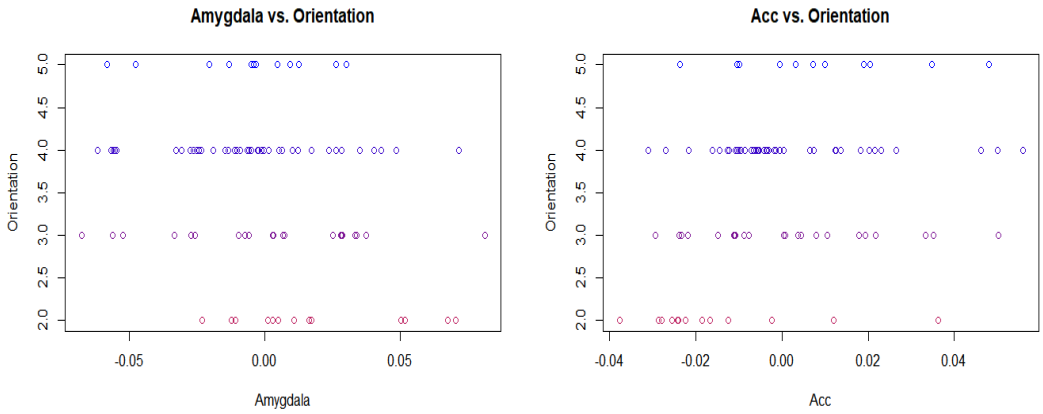


Fig. 4. Scatterplots of amygdala vs orientation and acc vs orientation respectively. Blue points indicate the liberal responses. Blue-purple indicate moderately liberal responses. Purple-red points indicate neutral responses. Red points indicate the moderately conservative responses.

correlation. For acc, the confidence interval of (0.09016436, 0.47744299) was found. Here it can be seen that the chance that the correlation is positive is very strong. 0 is not contained in the interval so it is very likely that acc has a positive correlation with orientation. These findings would indicate that activation of the amygdala is likely associated with conservatism and the activation of acc is likely associated with liberalism.

### 2.3 *Linear Model*

To further test if these relationships will hold, a simple linear model is fitted (2.1) and a summary is provided in Table 1.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{2.1}$$

Amygdala has an estimate of -5.23855 with a p-value of 0.0661. While this is not significant at the 5% level, it is still quite a small p-value. The negative estimate is consistent with the bivariate analysis where amygdala was correlated with a lower orientation. Acc has an estimate of 10.80198

| | Estimate | Std. Error | t value | Pr(>\| $t$ \|) |
|---|---|---|---|---|
| Intercept | 3.57784 | 0.09052 | 39.525 | <2e-16 |
| amygdala | -5.23855 | 2.81516 | -1.861 | 0.0661 |
| acc | 10.80198 | 4.49169 | 2.405 | 0.0183 |

Table 1. Summary table for linear regression model.

and a p-value of 0.0183. This is statistically significant at the 5% level and is also consistent with the findings above. Acc was correlated with higher scores of orientation and the estimate for the linear regression reflects that. Linear regression however is not the best suited model for the data. Linear regression assumes a normally distributed continuous response but the response variable is ordinal and from Figure 2 it can be seen that it is not normally distributed. The predictions achieved by this model simply don't make sense as orientation is not on a continuous scale.

### 2.4 *Next Steps*

To help further identify where students fall on the political spectrum, a binary response variable is created called 'conservative', classified as 1 when orientation $\leqslant 2$, and 0 when orientation $\geqslant 3$. Conservative is only equal to 1 for 13 observations. As stated earlier, this could be a potential issue when fitting a model as the dataset presented is already small.

## 3. MAIN ANALYSIS

### 3.1 *Logistic Regression Model*

The first model fitted to the data is logistic regression (3.2). The summary of the model is provided in Table 2.

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_i)}} \tag{3.2}$$

From Table 2 we can see that both amygdala and acc are significant at the 5% level. Amygdala

| | Estimate | Std. Error | z value | Pr(>\| z \|) |
|---|---|---|---|---|
| Intercept | -2.4582 | 0.4909 | -5.008 | 5.5e-07 |
| amygdala | 22.0246 | 10.9199 | 2.017 | 0.04370 |
| acc | -65.5247 | 25.2105 | -2.599 | 0.00935 |

Table 2. Summary table for linear regression model.

is positively related to conservative with an estimate of 22.0246 and acc is negatively related with an estimate of -65.5247.

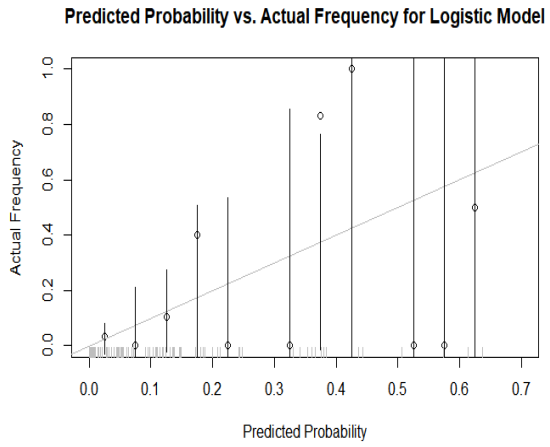


Fig. 5. Calibration plot for logistic regression. Vertical lines indicate confidence intervals for each point.

Model calibration is conducted and the plot above is obtained. The point below x=0.4 is just outside the confidence interval but the rest of the points fall within the confidence intervals. The intervals are quite wide due to the lack of points in the later frequencies. This makes is difficult to tell whether or not the data presented would be generalizable to a larger population.

Following model calibration, the in- sample misclassification rate is found at the 0.5 cutoff to see how well the model is classifying the points into the binary categories. The confusion matrix given by Table 3 is created and the in-sample misclassification rate for Table 3 is given by equation 3.3:

| | 0 | 1 |
|---|---|---|
| 0 | 74 | 3 |
| 1 | 12 | 1 |

Table 3. Confusion matrix for logistic model.

$$\frac{\#\text{ of missclassified points}}{n} = \frac{12+3}{90} = 0.1667 \tag{3.3}$$

The error rate is 0.1667 or approximately 16.67%. This model seems to do a good job at classifying the liberal points correctly as only 3 out of 74 are wrongly classified. However, due to there being 12 observations out of 13 that are misclassified as not being conservative, it seems like this model gives up on the minority class as only 1 conservative observation is properly classified as conservative.

To improve the generalizability of the error rate, 5-fold cross-validation is also performed. The error rate is recalculated and interestingly enough is also found to be 0.1667. So far the logistic model seems to fit the data moderately well with some room for improvement. The data fits logistic regression well as the observations are independent of each other and there is little multicollinearity between acc and amygdala. However, to properly fit the logistic model, amygdala and acc need to have linear effects. These effects can be found by fitting a generalized additive model (GAM).

### 3.2 *Generalized Additive Model*

The generalized additive model is fit (3.4), and the summary as well as model effect plots are stated in Table 4

$$Y_i = f_1(x_i) + f_2(z_i) + \epsilon_i \tag{3.4}$$

| | edf | Ref.df | Chi.sq | p-value |
|---|---|---|---|---|
| s(amygdala) | 1.000 | 1.000 | 4.422 | 0.03548 |
| s(acc) | 2.167 | 2.718 | 12.893 | 0.00555 |

Table 4. Summary table for linear regression model.

According to the summary of the GAM model, both the smoothing terms of amygdala and acc are significant at the 5% significance level. These results are very similar to the logistic model above. The model effect plots in Figure 6 show that amygdala has a linear smoothing effect and that acc has a curved smoothing effect that bends at approximately 0. This curved effect violates the logistic regression assumption of linear effects.

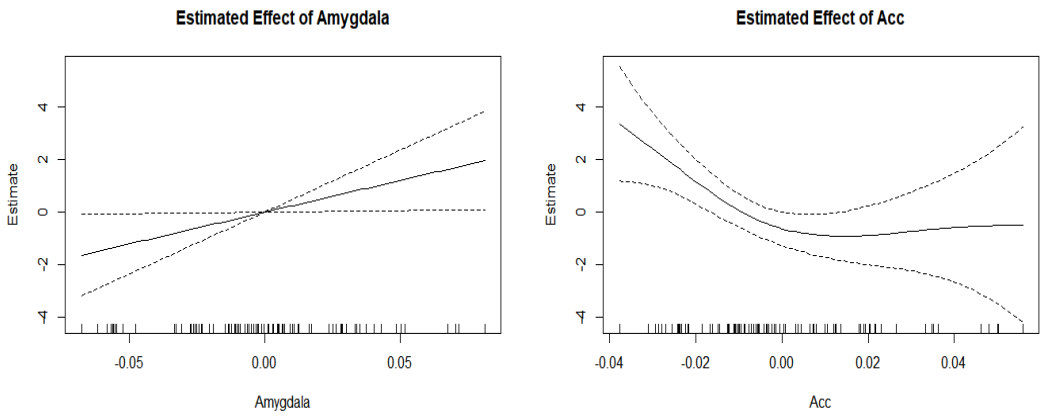


Fig. 6. Model effect plots for Amygdala and Acc respectively.

Model calibration was then performed to further investigate model fitting and Figure 7 was obtained. All the points on this plot fit into the confidence intervals, which is not difficult as the intervals are so wide due to low sample size. This calibration plot has similar issues to the logistic regression calibration plot. The GAM model seems to perform slightly better in regards to model calibration. It is important to test how the GAM model performs against the logistic model in terms of error rate as well, therefore the in-sample misclassification rates are compared next.

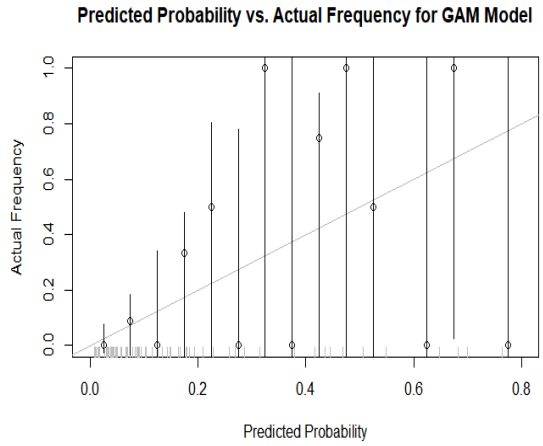


Fig. 7. Calibration plot of the generalized additive model. Vertical lines indicate confidence intervals for each point.

The confusion matrix is computed and found in Table 5. Using the same formula as specified in Equation 3.3, the In-sample misclassification error rate is found to be 0.1444 (Equation 3.5) or approximately 14.44%. This is slightly better than the previous model's rate of 0.1666. This model does a better job at classifying conservative points as there are 2 more points classified correctly. However, a large portion of the conservative points are still incorrectly classified, showing preference for the majority class.

| | 0 | 1 |
|---|---|---|
| 0 | 74 | 3 |
| 1 | 10 | 3 |

Table 5. Confusion matrix for GAM model.

$$\frac{10+3}{90} = 0.1444 \tag{3.5}$$

When this process is repeated using 5-fold cross validation, the classification error rate is calculated to be 0.2. This is slightly higher than the logistic regression cross validated rate of 0.1666. It is unclear which model would be best to generalize the data to a larger population. The GAM model however appears to fit the provided data better. To further investigate the accuracy of the fitted models, a hypothesis test is conducted with deviance being used as the test statistic.

The test is defined as $H_0$= Logistic model, and $H_a$= GAM model. It was found that the p-value= 0.24. With $\alpha = 0.05$, the p-value is too large to be rejected so there's not enough evidence in this case to say there's a significant difference between models.

## 4. Conclusion

Given the results presented above, neither model is a great fit for the data given. However, in consideration of the small sample size, the GAM model does a better job at providing informative interpretations for the data. It is more appropriate given that it accounts for the non-linear effect of Acc and has a lower in-sample misclassification rate. It is possible that logistic regression may generalize better to the population as it had a lower cross-validatation classification error rate, but for the data presented, the GAM model is more suitable. More data and model fitting is needed to find an accurate final model applicable for generalization. Both models evaluate the relationship between brain regions and political orientation similarly. Acc and amygdala volumes were both found to be significant for predicting political views. The exploratory and main analysis found that acc was associated with higher levels on the five-point-scale and that amygdala was found to be associated with lower levels. These findings are consistent with Kanai *and others* (2011), stating that, "greater liberalism was associated with increased gray matter volume in the anterior cingulate cortex, whereas greater conservatism was associated with increased volume of the right amygdala."

## APPENDIX

```
#Loading in dataset
brain_responses=read.csv("n90_pol.csv")

#Exploratory data analysis

#Univariate analysis
hist(brain_responses$amygdala, main="Histogram of Amygdala", xlab="Amygdala")
summary(brain_responses$amygdala)
hist(brain_responses$acc, main="Histogram of Acc", xlab="Acc")
summary(brain_responses$acc)
hist(brain_responses$orientation, breaks = 5:0, labels=TRUE,
     main="Histogram of Orientation",xlab="Orientation")
summary(brain_responses$orientation)

#The correlation between volumes of amygdala and acc
plot(brain_responses$amygdala, brain_responses$acc, main="Amygdala vs. Acc"
     ,xlab="Amygdala", ylab="Acc")
abline(0,0)
cor(brain_responses$amygdala, brain_responses$acc)

#The correlation between orientation and the volume of the amygdala
palette=colorRampPalette(c("#FF3030", "#0000FF"))(5)
plot(brain_responses$amygdala, brain_responses$orientation,
     main="Amygdala vs. Orientation",xlab="Amygdala", ylab="Orientation",
     col=palette[brain_responses$orientation])
cor(brain_responses$amygdala, brain_responses$orientation,method="s")

#The correlation between orientation and the volume of the acc.
plot(brain_responses$acc, brain_responses$orientation,
     main="Acc vs. Orientation",xlab="Acc", ylab="Orientation",
     col=palette[brain_responses$orientation])
cor(brain_responses$acc, brain_responses$orientation, method="s")

#95% bootstrap confidence intervals for above correlations.

#For amygdala
nboot=500
boot_vec=rep(NA,nboot)
set.seed(444)
for (b in 1:nboot) {
  boot_data=brain_responses[sample(1:90,replace = T),]
  boot_cor=cor(boot_data$amygdala, boot_data$orientation,method = "s")
  boot_vec[b]=boot_cor
}
quantile(boot_vec,probs=c(0.025,0.975))

#For acc
boot_vec=rep(NA,nboot)
for (b in 1:nboot) {
  boot_data=brain_responses[sample(1:90,replace = T),]
  boot_cor=cor(boot_data$acc, boot_data$orientation,method = "s")
  boot_vec[b]=boot_cor
}
quantile(boot_vec,probs=c(0.025,0.975))

#Linear regression model for orientation on the volumes of amygdala and acc
brain_lm=lm(orientation~amygdala+acc, data=brain_responses)
summary(brain_lm)
```

```
#Main analysis
#creating a binary response variable 'conservative', which is 1 when the
#student has orientation less than or equal to 2, and 0 otherwise.
brain_responses$conservative=as.numeric(brain_responses$orientation<=2)
mean(brain_responses$conservative)*90

#Fitting a logistic regression of conservative on the linear effects of the

#volumes of amygdala and acc.
brain_responses_glm=glm(conservative~amygdala+acc, data = brain_responses,
                        family = binomial)
summary(brain_responses_glm)

#Performing model calibration for logistic model
brain_responses$pred=predict(brain_responses_glm,type="response")
brain_responses$pred_bin=cut(brain_responses$pred,breaks = seq(0,1,0.05))
cali_table=aggregate(brain_responses$conservative,
                     by=list(bin=brain_responses$pred_bin)
                     ,FUN=function(x)c(mean=mean(x),count=length(x)))
cali_table$pred=c(0.025, 0.075, 0.125, 0.175, 0.225, 0.325, 0.375, 0.425,
                  0.525, 0.575, 0.625)
cali_table$se= sqrt(cali_table$pred *
                      (1 - cali_table$pred)/cali_table$x[,"count"])

#Plotting calibration plot for logistic model
plot(cali_table$x[,"mean"]~cali_table$pred,
     main="Predicted Probability vs. Actual Frequency for Logistic Model",
     xlab="Predicted Probability", ylab="Actual Frequency",
     xlim = c(0, 0.7), ylim = c(0, 1))
abline(0, 1, col = "grey")
rug(fitted(brain_responses_glm), col = "grey")
segments(x0 = cali_table$pred, y0 = cali_table$pred - 1.96 * cali_table$se,
         y1 = cali_table$pred +  1.96 * cali_table$se)


#Finding in-sample mis-classification rate using 0.5 cutoff for logistic
#model
cons_pred_prob=data.frame(predict(brain_responses_glm,type="response"))
brain_responses_merge=merge(x=brain_responses,y=cons_pred_prob,all=TRUE,
                            by="row.names")
brain_responses_merge$cons_pred=
  as.numeric(brain_responses_merge$
               predict.brain_responses_glm..type....response..>0.5)
table(brain_responses_merge$conservative,brain_responses_merge$cons_pred)
(3+12)/90

#Recalculating the classification error rate for logistic model
#using cross-validation
set.seed(444)
folds=matrix(sample(nrow(brain_responses)),ncol=5)
cv_error=function(f_fit,form,...){
  cv_cer=rep(NA,5)
  for (i in 1:5) {
    mod=f_fit(form, data = brain_responses[c(folds[,-i]),],
              family = binomial,...)
    pred=as.numeric(predict(mod,brain_responses,type="response")
```

```
                    [folds[,i]]>0.5)
    cons=brain_responses$conservative[folds[,i]]
    cv_cer[i]=mean(pred!=cons)
  }
  return(mean(cv_cer))
}
cv_error(glm,conservative~amygdala+acc)

#Fitting a generalized additive model for conservative on the volumes of
#amygdala and acc.
library(mgcv)
brain_responses_gam=gam(conservative~s(amygdala)+s(acc),
                        data=brain_responses, family=binomial)
summary(brain_responses_gam)
plot(brain_responses_gam,select=1, main="Estimated Effect of Amygdala",
     ylab="Estimate", xlab="Amygdala")
plot(brain_responses_gam,select=2, main="Estimated Effect of Acc",
     ylab="Estimate", xlab="Acc")

#Performing model calibration for GAM model
brain_responses$pred=predict(brain_responses_gam,type="response")
brain_responses$pred_bin=cut(brain_responses$pred,breaks = seq(0,1,0.05))
cali_table=aggregate(brain_responses$conservative,
                     by=list(bin=brain_responses$pred_bin)
                     ,FUN=function(x)c(mean=mean(x),count=length(x)))
cali_table$pred=c(0.025, 0.075, 0.125, 0.175, 0.225, 0.275, 0.325, 0.375,
                  0.425, 0.475, 0.525, 0.625, 0.675, 0.775)
cali_table$se= sqrt(cali_table$pred *
                      (1 - cali_table$pred)/cali_table$x[,"count"])

#Plotting calibration plot for Gam model.
plot(cali_table$x[,"mean"]~cali_table$pred,
     main="Predicted Probability vs. Actual Frequency for GAM Model",
     xlab="Predicted Probability", ylab="Actual Frequency",
     xlim = c(0, 0.8), ylim = c(0, 1))
abline(0, 1, col = "grey")
rug(fitted(brain_responses_gam), col = "grey")
segments(x0 = cali_table$pred, y0 = cali_table$pred - 1.96
         * cali_table$se, y1 = cali_table$pred +  1.96 * cali_table$se)

#Finding in-sample mis-classification rate using 0.5 cutoff for GAM model
cons_pred_prob2=data.frame(predict(brain_responses_gam,type="response"))
brain_responses_merge2=merge(x=brain_responses,y=cons_pred_prob2,all=TRUE,
                             by="row.names")
brain_responses_merge2$cons_pred2=
  as.numeric(brain_responses_merge2$
               predict.brain_responses_gam..type....response..>0.5)
table(brain_responses_merge2$conservative,brain_responses_merge2$cons_pred2)
(10+3)/90

#Recalculating the classification error rate for GAM model using
#cross-validation
cv_error(gam, conservative~s(amygdala)+s(acc) )

#Calculating deviance
set.seed(444)
simulate.from.logr <- function(df, mdl) {
```

```
  probs <- predict(mdl, newdata = df, type = "response")
  df$conservative <- rbinom(n = nrow(df), size = 1, prob = probs)
  return(df)
}
delta.deviance.sim <- function(df, mdl) {
  sim.df <- simulate.from.logr(df, mdl)
  GLM.dev <- glm(conservative~amygdala+acc, data = sim.df,
                 family = "binomial")$deviance
  GAM.dev <- gam(conservative~s(amygdala)+s(acc), data=sim.df,
                family='binomial')$deviance
  return(GLM.dev - GAM.dev)
}
(delta.dev.observed <- brain_responses_glm$deviance -
    brain_responses_gam$deviance)
delta.dev <- replicate(100, delta.deviance.sim(brain_responses,
                                               brain_responses_glm))
mean(delta.dev.observed <= delta.dev)
```

## References

Kanai, R., Feilden, T., Firth, C. and Rees, G. (2011). Political orientations are correlated with brain structure in young adults. *Current Biology* **21**, 677–680.