Summative

Liam Hughes

2023-12-05

Analysis 1 - Binary Logistic

Introduction

The first analysis will use the "wdbc.csv" data file available in the "Data" subfolder of the main project directory. This data can also be retrieved online at https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic. This dataset provides the maximum value, average (mean) value and standard error for 10 different features of masses found in women's breasts. Analysing these features has been shown to predict whether or not a mass is malignant or benign with up to 97% accuracy in previous studies ("Street (1999)). This study also indicates that the maximum (or 'worst') values are likely to be the strongest predictors of malignancy. As such, the present analysis will consider only the maximum values of each feature when attempting to create a model. The aim of this analysis will be to create a statistical model that can accurately predict whether or not a tumour is benign or malignant based on the 10 different features provided by the data.

Exploratory Analysis

Boxplots for each possible predictor can be seen in Figure 2. As can be seen, the general trend across all predictors tends to indicate that more extreme/larger measured values are associated with a malignant diagnosis, with the perimeter, concave points and radius measurements displaying the greatest differences between the two groups.

Due to the binary nature of the data, a binary logistic regression should be seen as most appropriate for modelling. However, there are some considerations that are required before proceeding. Firstly, as the data used are the largest values for each respective variable, there is likely to be some level of collinearity across the predictors. This is particularly true for the area, radius and perimeter variables, as the cells measured are typically round, and the formulae for calculating the area and perimeter are reliant on the radius value. Figure 1 displays a heatmap showing the correlation between predictor variables.

Furthermore, there are a large number of potential predictors, all of which could be important in diagnosis, making it irresponsible to not consider variables at first glance. The issue with using all of these predictors is the likelihood of complete or quasi-complete separation. That is, that the model predicts the outcome (benign/malignant) with perfect or near-perfect accuracy. Whilst this can sound positive, the biggest issue here is that the model likely loses generisability, as the accuracy of the model is dependant on the specific dataset. As such, penalised linear regression will be applied in order to account for both of these possible issues whilst still maintaining viability of as many predictors as possible.

Data Modelling

As previously mentioned, the outcome variable in this data set is binary (benign/malignant), therefore a binary logistic regression is the most logical choice for modelling. The general form for a binary logistic

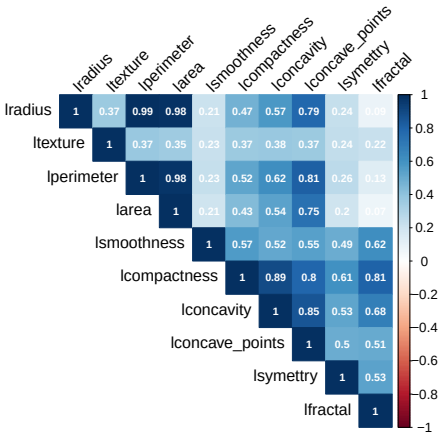

Figure 1: A heatmap displaying the correlation between predictor variables. Darker colours indicate higher correlation.

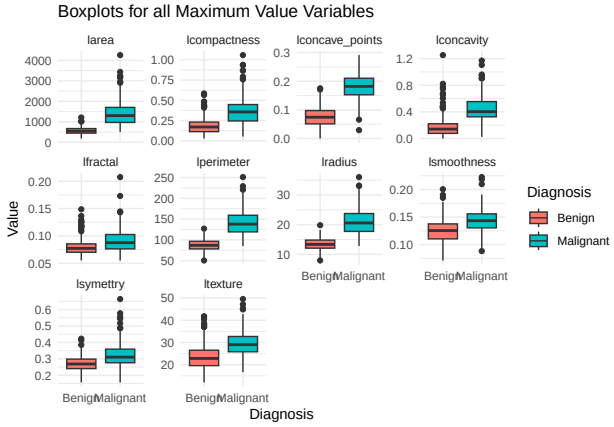

Figure 2: Boxplots displaying the differences between benign and malignant diagnoses, for each maximum value variable.

regression is denoted as:

$$y_i \sim \text{Bernoulli}(\theta_i), \quad \theta_i = \text{ilogit}(\phi_i), \quad \phi_i = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ki}, \quad \text{for } i \, \epsilon \, 1...n.$$

This indicates that each independent outcome, y_i, has its own probability of success, θ_i modelled by the bernoulli distribution.

As was seen in figure 2, the scales of each variable were drastically different, which could influence the results of the model, as variables with scales of greater magnitude tend to have larger regression coefficients. In order to make comparisons across coefficients (and thus across different predictors), the predictor variables will be standardised.

Now the variables are standardised, binary logistic regression can be applied. This model will use the lasso method of penalised binary logistic regression. In this method, coefficients are initially calculated using maximum likelihood estimation (values of the parameters which give the data the highest likelihood of occurence), then a penalty term, λ, is applied which shrinks all coefficients equally. Some coefficients may be shrunk to zero values, effectively removing them from the model. As such, it is unnecessary to perform nested model comparison with this method of penalised regression as the optimal model arises naturally as a result of the penalisation.

The data will be split into two sets; a training set and a testing set. This is done primarily to assess whether the model performs well with unseen data. The training set will be assigned two thirds of the data at random, and the testing set will be assigned the remaining third. Cross-validation will be used on the training set to determine the optimal value for λ.

```
#Fit model
train_model <- cv.glmnet(x_train, y_train, alpha = 1, family = 'binomial', type.measure = 'deviance')
#Make predictions using model
bin_predictions <- as.numeric(predict(train_model, s = train_model$lambda.1se, newx = x_test, type = 'r
```

After applying 10-fold cross validation, the optimal penalty term (λ) was calculated to be 0.01. The intercept term, β_0, was determined to be -0.83. As a result of the lasso penalisation, the 'area', 'compactness' and 'fractal' variables were removed from the model entirely. Therefore, the final model can be denoted as:

$$\phi_i = \beta_0 + \beta_1 rad_i + \beta_2 tex_i + \beta_3 per_i + \beta_4 smooth_i + \beta_5 conc_i + \beta_6 concpoints_i + \beta_7 sym \quad \text{for } i \, \epsilon \, 1...n.$$

Where β_0 = -0.83 and β_1 to β_7 are 3.15, 0.86, 0.01, 0.47, 0.26, 1.21, 0.12, respectively. These coefficients indicate the change in log odds of the diagnosis of an individual patient, when the respective predictor variable increases by one unit (and all other variables are held constant). This can be somewhat difficult to intuitively interpret, so the coefficients will be transformed into incidence rate ratios through exponentiation.

Incidence rate ratios represent the (multiplicative) change in odds of an event if the respective predictor increases by one unit, and all others are held constant. For example, in the context of our model, the incidence rate ratio for the radius variable is 23.34. This means that for each one unit increase of the radius variable, asssuming all other predictors are held constant, the tumour is 2234% more likely to be malignant/cancerous.

However, what needs to be considered in this case is that the data were standardised. As such, the coefficients represent the change in log odds for a one standard deviation increase in the respective predictor, rather than a single unit. This is also true for the incidence rate ratios, so in the previous example of the radius variable, the percentage increase is associated with a one standard deviation increase in the radius of the cell.

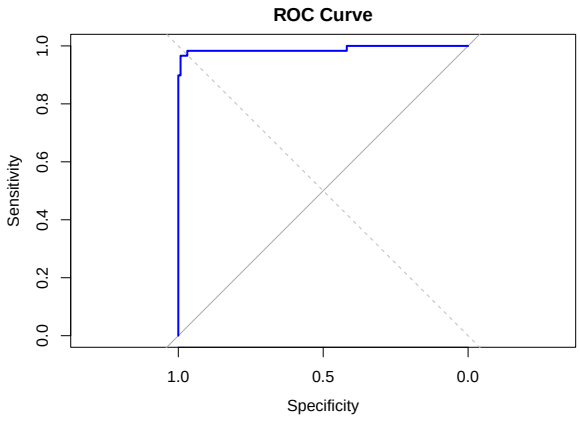

Figure 3: A plot to show the true positive rate against the false positive rate.

Model Performance

Now that the final model has been established, the performance of the model can be assessed. Figure 3 displays the true positive rate against the false positive rate, using the testing data (data that the model has not encountered). As can be seen, the model performs very well at predicting malignancy, with the area under the curve being 0.99. The area under the curve represents how well the model can distinguish between a tumour being benign and malignant. An area of 1 indicates perfect discrimination, whilst a score of 0.5 indicates a performance no better than random chance.

Conclusion

This model is able to effectively discriminate between benign and malignant tumours based on the 10 features provided by the dataset. It can therefore be concluded that using the cell features is definitely a viable method that can be used to diagnose cancers, although it should be used in conjunction with other methods due to the fatal consequences of a false-negative (i.e, a tumour identified by the model as benign but is actually malignant).

Analysis 2 - Poisson, Zero Inflated or Negative Binomial

Introduction

This analysis will use the "bike_counts.csv" data file available in the "Data" subfolder of the main directory, or alternately online at https://www.kaggle.com/datasets/new-york-city/nyc-east-river-bicycle-crossings. This dataset provides daily counts for the number of bicycles crossing various bridges across the eastern side of New York City. There are a total of 11 variables: X (ID for each day), Date, Day (a repeat of the Date variable), temp_high (highest recorded temperature on a given day), temp_low (lowest recorded temperature on a given day), Precipitation (level of rainfall recorded on given day), Brooklyn.Bridge (first of 5 count variables for number of cyclists crossing respective bridge), Manhattan.Bridge, Williamsburg.Bridge, Queensboro.Bridge and Total (sum of all cyclist bridge crossings). The present analysis will use only the "Total" variable as the outcome variable as it is representative of all mentioned bridges. This data presents the question: can the daily total number of bridge crossings be predicted by the highest recorded temperature and precipitation level?

Exploratory Analysis

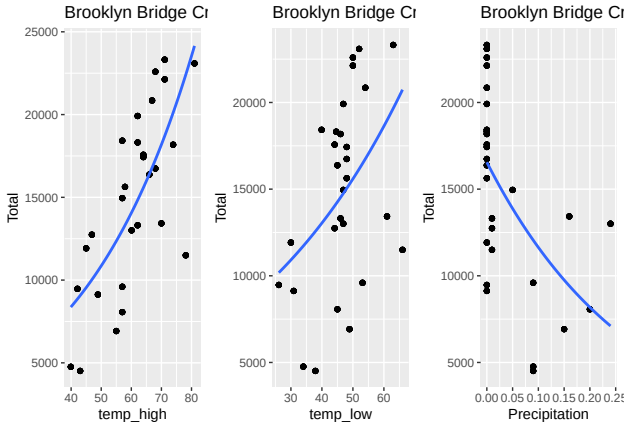

Figure 4: Scatterplots showing the number of bridge crossers per day by according to the day's highest recorded temperature, lowest recorded temperature and precipitation level.

Figure 4 shows the relationships between the total number of cyclists crossing all bridges and highest daily temperature, lowest daily temperature and daily precipitation, respectively. As can be seen, the temperature variables are both positively associated with the number of people crossing the bridges, whereas precipitation is negatively associated. Due to the high level of correlation between the temperature variables however, only one should be included in the model. The highest daily temperature seems to have the stronger relationship (albeit minimally) and more tightly distributed residuals, and thus will be the variable considered in the model.

Data Modelling

Due to the nature of the data being 'count' data, the most appropriate models will be based on either poisson or negative binomial regression. However, as the variance of the outcome variable is significantly greater than the mean ($\boldsymbol{m} = 14780$, $\sigma^2 = 29055557$), in this case the negative binomial model will be more appropriate in order to account for overdispersion. The general model for a negative binomial regression is:

$$y_i \sim \text{NegBinomial}(\mu_i, r), \quad \log(\mu_i) = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ki}, \quad \text{for } i \, \epsilon \, 1...n.$$

Where y_i is the outcome variable, μ_i is the mean and r is a unknown, constant, dispersion parameter.

```
#Initial negative binomial model
p_mod1 <- glm.nb(Total ~ 1, data = poisson_data)

#Extract r
r1 <- round(p_mod1$theta, 2)

#Extraxt mu
mu1 <- round(p_mod1$coefficients[1], 2)

#Calculate theta
theta <- round((r1 / (r1 + mu1)), 2)
```

```r
#Calculate deviance
deviance_mod1 <- -2 * logLik(p_mod1)
```

In order to better understand the data in the absence of any predictors, an intercept-only model will first be considered. The intercept only model assumes that on a given day, the total number of cyclists crossing all considered bridges is a negative binomial distribution with parameters $\mu = 9.6$, $r = 6.14$ and $\theta = 0.39$.

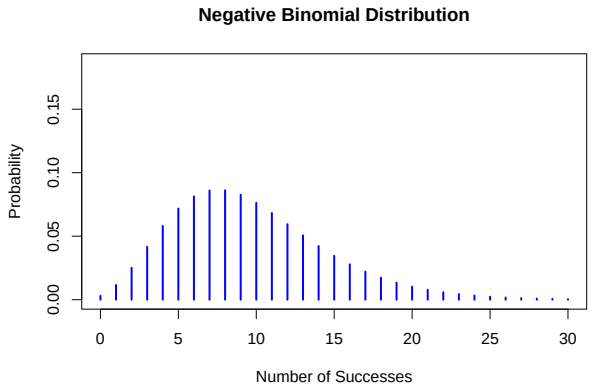

Figure 5: A negative binomial distribution based off the intercept only model.

To build on this initial model, the precipitation and temperature variables are now added in as predictors. This model is denoted as:

$$\log(\mu_i) = \beta_0 + \beta_1 temp_i + \beta_2 Precipitation_i$$

```r
#Plot second model
p_mod2 <- glm.nb(Total ~ temp_high + Precipitation, data = poisson_data)

#Compare models
anova_1v2 <- anova(p_mod2, p_mod1)

#Extract model comparison result
anova1_lr<- anova_1v2$`LR stat.`
```

In this model, both temp_high ($\beta_1 = 0.03$, SE = 0, z = 16.34, 95% CI{0.02, 0.03}) and precipitation ($\beta_2 =$ -2.41, SE = 0.26, z = -9.4, 95% CI{-2.88, -1.94}) were calculated to be significant predictors at the 5% level.

The coefficients indicate the change in the log count of total people crossing the bridges for a one unit increase in a given predictor variable (assuming all others are held constant). This can be difficult to interpret, so for simplification the coefficients will be transformed into incidence rate ratios through exponentiation.

These ratios are 1.03 and 0.09 for temp_high and precipitation respectively. Therefore, for each one unit increase in temp_high, we would expect to see a 3% increase in the total number of bridge crossings, whereas we would expect a 91% decrease in numbers for every unit increase in precipitation, assuming all other variables are held constant.

Deviance is a value that quantifies the difference between what the data show and what the model predicts, and as such is a key metric for assessing model fit. The deviance for this model is 3730.5, much lower than that of the intercept only model (3941.95), indicating better fit. An ANOVA test confirms that the model with predictors is significantly better at predicting the number of bridge crossings ($\Delta_D = 211.46$, $p < 0.01$)

As displayed in the exploratory analysis, the precipitation variable had a large number of zero values, indicating days where there was no rain. As such, it would be interesting to explore this variable as a factor,

Table 1: A table displaying the model coefficients and corresponding errors, Z values and p values.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	8.26	0.13	63.31	0.00
temp_high	0.02	0.00	11.64	0.00
Prec_Factorrain	-0.36	0.19	-1.91	0.06
temp_high:Prec_Factorrain	0.00	0.00	-0.24	0.81

separating days of rain vs no rain and seeing how this has an impact on the total number of bridge crossings. An interaction effect between the two predictors will also be considered, that is, does the fact that it rained or not change the relationship between the maximum temperature and total number of bridge crossings?

This new model can be denoted as:

$$\log(\mu_i) = \beta_0 + \beta_1 temp_i + \beta_2 rain_i$$

In this model, the 'rain' variable is binary and can only take the values of 0 or 1; 0 on days where there is no rain and 1 on days there is.

The coefficients and respective standard errors, z values and p values can be found in table 1. As can be seen, both predictors were once again considered significant at the 5% level, but the interaction effect between the two was not. This indicates that the relationship between the number of bridge crossings and highest recorded temperature is likely to hold true regardless of whether it rained or not.

The deviance of this new model (3699.92) is lower than that of the previous (3730.5), indicating a better fit. An ANOVA test confirms the newer model is significantly better at predicting the total number of bridge crossings ($\Delta_D = 30.57$, $p < 0.01$), thus confirming that it is better to interpret the 'precipitation' variable as binary, where the only notable information is whether it rained or not at all, opposed to a continuous variable which indicates the amount of rain on a given day.

As with the previous model, it is easier to interpret incidence rate ratios, which are 1.02 and 0.7 for the 'temp_high' and 'rain' variables respectively. This indicates that for a one unit increase in the 'temp_high' variable, we would expect a 2% increase in crossings, whereas we would expect a 30% decrease on days where rain is present (assuming all other variables are held constant).

Model Prediction

Now that the final model has been established, it can be used to make predictions of the total counts of bridge crossings based on the highest recorded daily temperature and whether or not it rained that day. These predictions are shown in Figure 6.

```
#Make predictions
negbin_predictions <- ggpredict(p_mod3, c("temp_high", "Prec_Factor"))

#Plot predictions
plot(negbin_predictions) +
  theme_minimal() +
  labs(title = "Predicted Counts for Negative Binomial Model",
       x = "Highest Recorded Temperature",
       y = "Predicted Counts",
       color = "Rain")
```

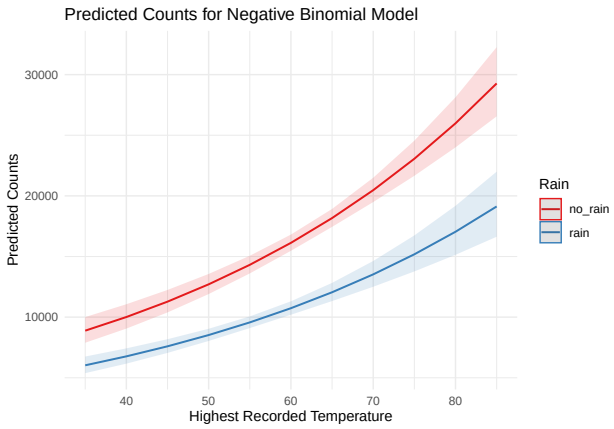

Figure 6: Predicted Total Bridge Crossings based on the final negative binomial model. The lines indicate the predicted values and the shaded surrounding area indicate the confidence intervals.

Conclusion

In conclusion, it should be seen that bridge crossings can be modelled somewhat effectively from the maximum daily recorded temperature and precipitation levels, however precipitation levels should be interpreted simply as zero or non-zero in order to optimise the model. It is also likely that more information is required in order to model with greater accuracy, however the model presented in this analysis can at least capture the overall trend of the data to understand ball-park figures.

Analysis 3 - Multi-Level Linear

Introduction

This analysis will use the JSP.DAT data file Mortimore (1988) available in the data subfolder of the main working directory, or alternatively online at https://www.bristol.ac.uk/cmm/learning/support/datasets/. The data are from 50 schools in inner London, and measure the performance of school children over a 3 year period. This dataset contains 3236 observations of 9 variables: School (ID of school attended by child), Class (class in school child attended), Sex (sex of child; boy or girl), Social (9-level factor indicating self-rated social class of child), Ravens (score on Raven's test; general ability measure taken prior to 3 year period), ID (individual ID of child), English (english test score, possible outcome variable), Math (math test score, possible outcome variable) and Year (year of study). For simplicity, and to get the best idea of a child's overall ability across maths and english, the Math and English variables will be combined into a weighted average outcome variable. This analysis will attempt to discover how an individual child's overall score varies as a function of their sex, social class and their Raven's test score, as well as to investigate whether these effects differ across schools.

In light of the aims of the analysis, data from only the first year of study will be considered. The continuous variables, Ravens and Overall, will also be standardised in order to maximise interpretability of the model coefficients and prevent convergence issues. Additionally, the data will be split into training and testing sets, with the training set being assigned 80% of the data. The testing set will then be used to make predictions and assess the model's predictive ability.

The coding for each level of the Social variable, relating to the highest earner in a child's household, is as follows:

1. Upper class

2. Upper middle class
3. Middle class, non-manual job
4. Middle class, manual job
5. Lower middle class
6. Working class
7. Long-term unemployed
8. Not currently employed
9. Father absent

Exploratory Analysis

The data takes on a hierarchical structure, as there are individual students, the classes that these students are from and the schools that these classes are from.

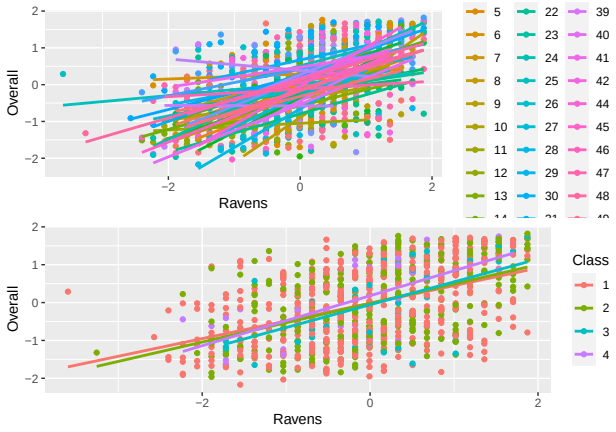

Figure 7: Scatterplots including regression lines to show the distribution of the weighted average test scores against the Ravens test scores, by school and class, respectively.

As can be seen in Figure 7, there are significant differences between schools in terms of performance, both in terms of the slopes and intercepts of each regression line. The class variable seems to have little differentiation across groups, however this is somewhat irrelevant unless considered within the individual schools.

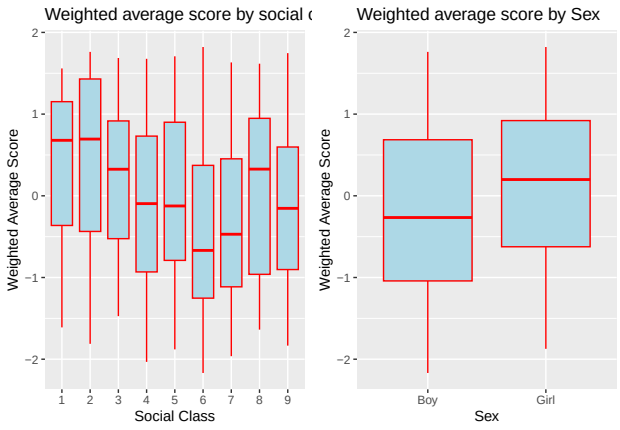

Figure 8: Box plots to show the distribution of the weighted average variable by categorical variables Social and Sex, respectively.

Table 2: Fixed Effects

	Estimate	Std. Error	t value
(Intercept)	-0.0501473	0.0497756	-1.007467
Ravens	0.5062506	0.0344697	14.686824

What can also be recognised from figure 8 are the differences in weighted average scores between both the Sex and Social variables, with girls generally performing better than boys albeit with similarly distributed scores, and a great range of performance based on social class, although once again the spread is somewhat consistent among all levels.

Data Modelling

As previously mentioned, this dataset provides data with a hierarchical structure, and as such is fit for multi-level (linear) regression. The general form of a multi level linear model is:

$$\text{for} i \epsilon 1...n, \quad y_i \sim N(\mu_i, \sigma^2), \quad \mu_i = b_0 + b_1 x_i + \zeta_{[s_i]0} + \zeta_{[s_i]1} x_i, \quad \text{for} j \epsilon 1...J, \quad \vec{\zeta_j} \sim N(0, \Sigma).$$

What this indicates is that, each independent observation y_i is modelled from a normal distribution with mean μ_i and standard deviation σ. μ_i, which also represents the the expected value of the outcome variable for the ith observation, can be calculated through the above formula, with b_0 representing the intercept term and b_1 representing the coefficient for predictor x_i. This part of the formula represents the "fixed" effects, that is, effects that are expected across the entire population without regards to the hierarchical structure.

The second part of the formula can therefore be considered as the "random" effects, that is, effects that vary based on the hierarchical structure of the data. These effects are denoted by $\zeta_{[s_i]0}$, which is the "random" intercept term, and $\zeta_{[s_i]1}$, the "random" coefficient. These are not strictly random, but rather are represented by a matrix of normal distributions with mean 0 and covariance matrix Σ.

To put this in the context of the present analysis, the Sex and Social variables would be part of the fixed effects; the effect of each is expected to be roughly similar across the entire population. In contrast, the school, class and year variables are all considered part of the "random" effects, as each level is expected to have slightly different effects. For example, if we have a student in a particular class, that individual class is expected to have its own distribution of scores based on class-level factors, such as the quality of teaching. The school from which the class is from is also expected to have its own distribution, based on something such as the amount of funding the school gets. The variation in these hierarchical layers can be best captured through a multi-level model.

Initially, a basic model will be plotted. Additional complexity will be added until the optimal model is found. The initial model will consider only 'ravens' variables as a predictor, with school as the only hierarchy level.

```
#Initial mixed effects model
lme1 <- lmer(Overall ~ Ravens + (Ravens | School), data = train_data_lme)

#Extract model summary
lme1_sum <- summary(lme1)

#Extract fixed effects
lme1_fixed <- lme1_sum$coefficients[, c("Estimate", "Std. Error", "t value")]
```

```
## $School
```

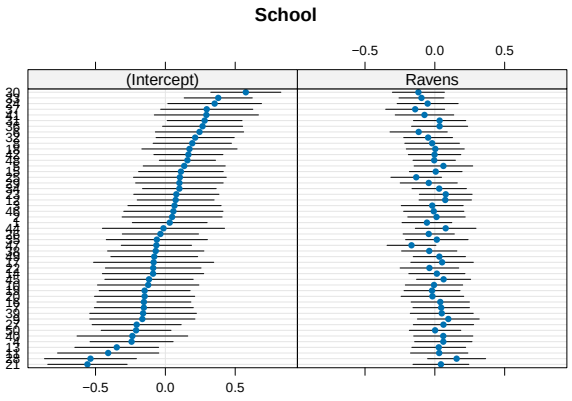

Figure 9: A dotplot displaying the initial model's random effects.

The fixed and random effects of the initial model can be in Table 2 and Figure 9 respectively. The fixed effects can be interpreted similar to previous models; the intercept represents the value of the 'Overall' variable when all predictors are zero. The coefficient for "Ravens" represents the change in the outcome variable for a one-unit increase in the Raven's variable, assuming all other variables are held constant (irrelevant in this model's context as it is the only predictor). However, as both continuous variables have been standardised, the units in this case represent standard deviation changes, rather than individual units.

Interpreting the random effects is slightly more complex. The random effects have a variance and standard deviation, so for the intercept, the variance and standard deviations represent the variability in the intercept across all schools. This is similar for the "Ravens" variable, where the variance and standard deviations represent the variability in the slope of the "Ravens" variable across all schools. As such, the random effects capture the differences across schools.

Now the initial model has been fitted, extra complexity will be added and models will be compared in order to find the optimal model. First, the "Sex" variable will be added.

```
#Second mixed model
lme2 <- lmer(Overall ~ Ravens + Sex + (Ravens | School), data = train_data_lme)

#Compare two models
lme_anova_1 <- anova(lme1, lme2)
```

The log-likelihood of the second model is -1133.67, significantly lower than that of the initial model (-1154.9). An ANOVA test confirms that the second model is a better fit for the data at the 5% level ($p = 0$). As such, the initial model will be rejected.

Now the model will be made as complex as possible, adding the Social variable to the previous model, and also considering the Class variable as a hierarchical layer.

```
#Attempt to create third model (boundary issue)
lme3 <- lmer(Overall ~ Ravens + Sex + Social + (Ravens | School/Class), data = train_data_lme)

#Create final model
lme4 <- lmer(Overall ~ Ravens + Sex + Social + (Ravens | School), data = train_data_lme)

#Compare models 2 and 4
lme_anova_2 <- anova(lme2, lme4)
```

Table 3: Fixed Effects

	Estimate	Std. Error	t value
(Intercept)	-0.0413650	0.1696268	-0.2438587
Ravens	0.4863164	0.0326749	14.8834818
SexGirl	0.3714352	0.0539732	6.8818424
Social2	0.0429390	0.1793324	0.2394380
Social3	-0.0332191	0.1864487	-0.1781677
Social4	-0.2604528	0.1696126	-1.5355750
Social5	-0.1270088	0.1883111	-0.6744629
Social6	-0.3842831	0.1894964	-2.0279176
Social7	-0.4075640	0.2026354	-2.0113167
Social8	-0.1421041	0.2647835	-0.5366804
Social9	-0.2451696	0.1779543	-1.3777112

However, when attempting to fit this model, an error is returned, specifying a boundary (singular) fit. This is likely due to overcomplicating the model with the class level of hierarchy. As such, this model will be disregarded. Instead, only the "Social" variable will be added as an additional predictor.

```
## $School
```

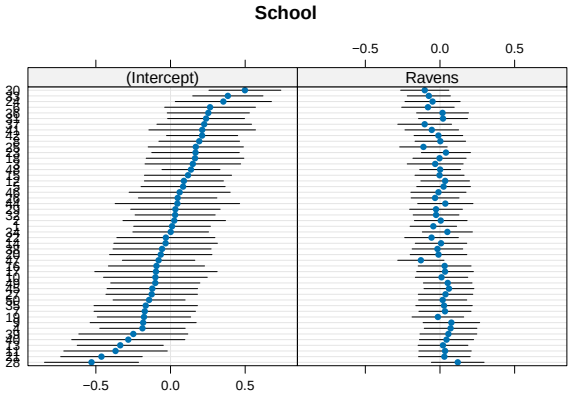

Figure 10: A dot plot displaying the final model's random effects.

The log-likelihood of the final model is -1132.15, compared to -1133.67 of the previous viable model. An ANOVA test confirms that the more complex model is a significantly better fit to the data ($p = 0$). This model will therefore be considered the optimal model. The fixed and random effects for this model can be seen in table 3 and 10 respectively. These can be interpreted similar to the initial model, however the interpretations of factor variables (i.e., Sex and Social) are in reference to the base level. In this analysis, the base level for "Sex" is being a boy, and the base level for "Social" is level 1. Thus, all coefficients for these two variables represent the standard deviation change of the outcome variable, assuming all other variables are held constant. For example, the "Sex" variable's base level is "Boy", so the coefficient in table 3 associated with "SexGirl" indicates the standard deviation increase in the "Overall" (outcome) variable if an individual is a girl (assuming all other variables are held constant). To put it as simply as possible, the model generally expects girls to achieve a higher weighted average score than boys.

Model Prediction

Now that the final model has been established, it's predictive ability will be tested using the "testing" set that was established earlier in the analysis. In order to assess the model's ability to capture the differing effects by school, 10 schools will be chosen at random to have predicted vs actual values graphs plotted.

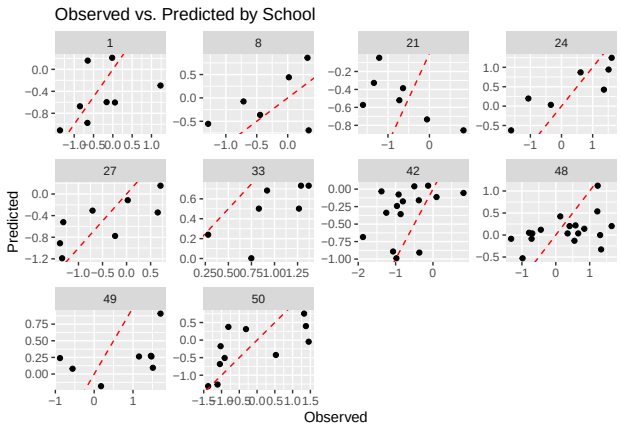

Figure 11: A graph showing the predicted vs actual values for the model. The dotted red line indicates where the predicted and actual values are the same. 10 schools were picked at random due to

As can be seen in Figure 11, the model performed reasonably well at predicting the values across different schools. The dotted line in these graphs represent where the predicted and actual values are the same, thus indicating perfect prediction. Whilst most of the values do not lie exactly on this line, most show a distribution around the line, which shows that the model has done well to capture the underlying pattern across the different schools.

Conclusion

This analysis attempted to create a model which could predict how overall score varies as a function of sex, social class and Raven's test score, as well as whether these effects differ across schools. In light of the predicted vs actual graph, this model should be considered a success. The model was able to identify the differing relationships across differing schools with a good level of accuracy (at least as far as identifying the underlying patterns), as well as highlighting the overall pattern (fixed effects) that the data seemed to follow. Therefore, multi-level linear models should be seen as adequate when attempting to address differences between schools with respect to academic achievement.

Analysis 4 - Nonlinear Regression

Introduction

The final analysis will use the 'manufacturing.csv' data file available in the Data subfolder of the main project directory. This data file contains 3957 observations of 6 variables: 'Temperature. . . C', 'Pressure..kPa', 'Temperature.x.Pressure', 'Material.Fusion.Metric', 'Material.Transformation.Metric' and 'Quality.Rating'. All variables are related to particular manufacturing processes, with the latter 5 being conditions of the process and the 'Quality.Rating' intuitively being the rated quality of the outcome of said process. As such, the latter 5 are possible predictors and 'Quality.Rating' is the intended outcome variable. This analysis will attempt to use non-linear regression to create a model that accurately predicts the 'Quality.Rating' variable.

In order to later assess model quality, the dataset will be split at random into a training and testing set, with training being assigned 80% of the data. The variables will also be given simpler names that are just as easy to interpret: 'Temperature', 'Pressure', 'TempxPres', 'MFM', 'MTM' and 'Quality'.

Exploratory Analysis

As with all previous analyses, exploratory analysis will be performed to assess potential relationships within the data.

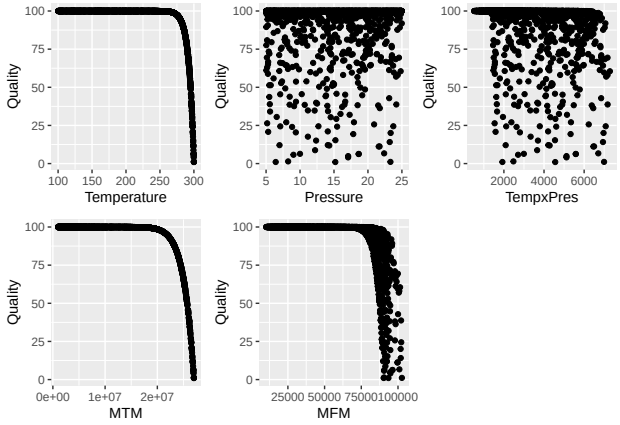

Figure 12: Scatterplots to show the distribution of all possible predictor variables with the Quality variable

As can be seen in Figure 12, the Temperature, MTM (Material Transformation Metric) and MFM (Material Fusion Metric) all seem to have similar relationships with the quality variable, showing a sharp decrease once a certain threshold is met. In contrast, both variables relating to pressure (pressure and pressure multiplied by temperature) seem to have no clear relationship whatsoever. As such, it makes sense to exclude these variables from the initial model.

However, what is also highlighted by Figure 12 is the possible issue of multicollinearity. This can be highlighted by the beforementioned extremely similar relationships between the quality variable and the temperature, MTM and MFM variables.

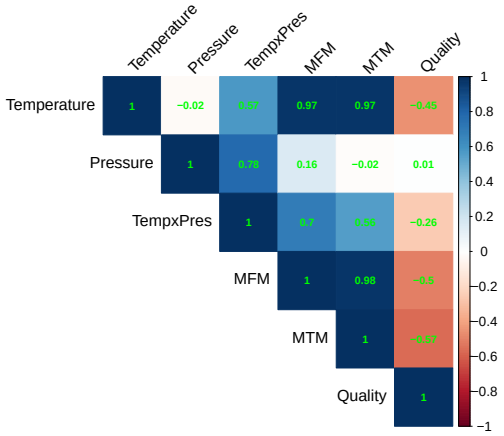

Figure 13: A correlation matrix indicating the correlation between all variables within the dataset

As suspected, the temperature, MTM and MFM variables show near perfect correlation, highlighted in figure 13. This is an issue that will require consideration further into the analysis. Furthermore, the MFM variable

shows the greatest variability with respect to the temperature and MTM variables, and as such will also be excluded from the model.

Data Modelling

The nonlinear approach taken will be that of a generalised additive model - a sum of smooth functions rather than a sum of linear terms. Therefore, whereas the general form for linear regression can be denoted as:

$$y_i \sim N(\mu_i, \sigma^2), \quad \mu_i = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ki}, \text{for } i \, \epsilon \, 1...n.$$

Indicating that each independent observation of the outcome variable, y_i, comes from a normal distribution with mean μ_i and standard deviation σ, where the mean is a linear function of the predictors.

Whereas the model for a generalised additive model is:

$$y_i \sim D(g(\mu_i)), \quad \mu_i = f_1(x_{1i}) + f_2(x_{2i}, x_{3i}, x_{4i}) + ... + f_L(x_{Li}), \quad \text{for } i \, \epsilon \, 1...n.$$

Whilst the two are similar, this new model denotes that each independent value of the outcome instead comes from some undefined distribution with parameter μ_i. μ_i can then be defined from the addition of a smoothed intercept term, $f_1(x_{1i})$ and of smooth functions of the predictor variables.

The 'mgcv' R package (Wood (2017)) is used to calculate the optimal model for the data. Parameters are calculated through a process similar to maximum likelihood estimation: restricted marginal likelihood. This process balances likelihood with complexity, in order to avoid the pattern of the data being overly influenced by noise. This is reflected in two additional hyperparameters, k and sp, which represent the number of basis functions and the smoothing penalty, respectively. In this model, the basis functions will be thin plate splines.

The temperature and MTM variables will be considered individually due to their correlation, and the models compared to determine the best fit.

```
#Model data with only temperature as predictor
polmod1 <- gam(Quality ~ s(Temperature), data = poly_train)
```

```
#Plot second model with MTM as predictor
polmod2 <- gam(Quality ~ s(MTM), data = poly_train)

#Anova test between models
pol_anova <- anova(polmod1, polmod2)

#Extract deviance values
pol_dev1 <- pol_anova$`Resid. Dev`[1]
pol_dev2 <- pol_anova$`Resid. Dev`[2]

#Extract df
pol_df <- summary(polmod2)$s.table[1]

#Extract smoothing parameter
pol_sp <- format(polmod2$sp, scientific = FALSE)
```

After fitting the two models, the model with the MTM variable as a predictor had a significantly lower residual deviance ($\boldsymbol{D} = 4706$) when compared to the model with temperature as a predictor ($\boldsymbol{D} = 33613$) indicating better model fit. As such, this will be considered the optimal model.

The final model had 9 effective degrees of freedom, indicating the smooth term is a ninth-order polynomial. The smoothing penalty was determined to be extremely small ($\text{sp} = 4.1763012 \times 10^{-5}$).

Model Predictions

Now the model is finalised, it's predictive ability should be evaluated using the testing data.

```
#Make predictions using model
polypredict <- predict(polmod2, newdata = poly_test, type = "response")

#Plot predictions
plot(poly_test$Quality, polypredict, main = "Actual vs. Predicted", xlab = "Actual", ylab = "Predicted")
abline(a = 0, b = 1, col = "red", lty = 2)
```

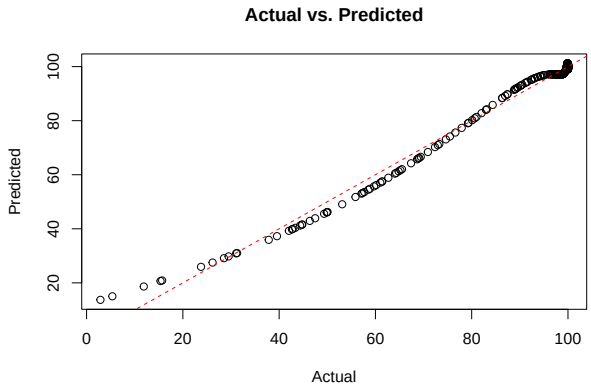

Figure 14: A graph showing the predicted values of the final model vs the actual values

As can be seen in Figure 14, the model performed excellently at predicting the actual values in the testing set. The red dashed line indicates perfect prediction; where the model's predicted value and the actual value are equal. As such, the close proximity of the data points to this line indicate very good predictive performance.

Conclusion

It can be concluded that a GAM model is very effective at modelling the quality rating of manufacturing processes, at least within this particular dataset. Whilst the data were split into training and testing sets, the near-perfect predictive performance could be indicative of an overfit model, a common issue among non-linear regression models. Therefore, the generalisability of the model is questionable at best - more data would be required to validate the model further.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Friedman, J., Tibshirani, R., & Hastie, T. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Kleiber, C., & Zeileis, A. (2008). *Applied econometrics with R*. Springer-Verlag. https://CRAN.R-project.org/package=AER

Lüdecke, D. (2018). Ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, *3*(26), 772. https://doi.org/10.21105/joss.00772

Lüdecke, D. (2023). *sjPlot: Data visualization for statistics in social science*. https://CRAN.R-project.org/package=sjPlot

Mortimore, P. (1988). *School matters: The junior years*. Open Books. https://books.google.co.uk/books?id=jdg0AAAAMAAJ

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for r and s+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77.

Robinson, D., & Hayes, A. (2021). *Broom: Convert statistical objects into tidy tibbles*. https://CRAN.R-project.org/package=broom

Roman Tsegelskyi, Gergely Daróczi. (2022). *Pander: An r pandoc writer*. https://CRAN.R-project.org/package=pander

Sarkar, D. (2008). *Lattice: Multivariate data visualization with R*. Springer. http://lmdvr.r-forge.r-project.org

"Street, "Wolberg., Nick". (1999). Nuclear feature extraction for breast tumor diagnosis. *Proc. Soc. Photo-Opt. Inst. Eng.*, *1993*. https://doi.org/10.1117/12.148698

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth). Springer. https://www.stats.ox.ac.uk/pub/MASS4/

Wei, T., & Simko, V. (2021). *R package 'corrplot': Visualization of a correlation matrix*. https://github.com/taiyun/corrplot

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation*. https://dplyr.tidyverse.org

Wolberg, M., William, & Street, W. (1995). *Breast Cancer Wisconsin (Diagnostic)*. UCI Machine Learning Repository.

Wood, S. N. (2017). *Generalized additive models: An introduction with r*. Chapman; Hall/CRC.

Xie, Y. (2023). *Knitr: A general-purpose package for dynamic report generation in r*. https://yihui.org/knitr/