Mini project 1

Exercise 1 of 3

Sebastian Belalcazar, David Melo Valbuena, Juan Andres Ruiz

12 March 2025

1 Introduction

1.1 Problem Context

The aim of this study is to predict the proportion of children with malformations based on the mother's alcohol consumption. In this analysis, alcohol consumption is weighted using the following values: 0, 0.5, 1.5, 4, and 7. This weighting allows us to assess the effect of different levels of alcohol exposure during pregnancy, which is essential for identifying potential risks and establishing prevention strategies.

1.2 Data

The data used, referred to as *Datos_1*, contain the following information:

- **Alcohol**: Weighted level of mother's alcohol consumption during pregnancy.
- **Frec.Presente**: Frequency of children with malformations.
- **Frec.Ausente**: Frequency of children without malformations.

Table 1 displays the original data:

	Alcohol	Frec.Presente	Frec.Ausente
1	0	48	17066
2	0.5	38	14464
3	1.5	5	788
4	4	1	126
5	7	1	37

Table 1: Original data (*Datos_1*).

2 Conversion to Proportions

2.1 Conversion the frequencies received to proportions

To properly analyze the malformation proportion, it is necessary to convert the frequencies received into proportions. This is achieved using the following formula:

$$Malformation\_proportion = \frac{Frec.Presente}{Frec.Presente + Frec.Ausente}$$

After applying this transformation and grouping the data by the level of alcohol consumption, we obtain the summary table shown in Table 2:

Alcohol	Malformation_proportion
0	0.00280
0.5	0.00262
1.5	0.00631
4	0.00787
7	0.02630

Table 2: Malformation proportion by level of mother's alcohol consumption during pregnancy.

Table 2 provides an important insight into what might be expected. For example, mothers with a weighted alcohol consumption of 7 have a considerably higher malformation proportion in their children (0.02630) compared to those consuming lower levels of alcohol (e.g., 0 or 0.5, where the proportions are approximately 0.00280 and 0.00262, respectively). This suggests that higher levels of alcohol consumption during pregnancy may be associated with an increased risk of malformations in children. However, it is important to interpret these results within the context of a broader statistical model and to consider other variables and potential confounding factors.

3 Model Fitting

To evaluate the relationship between alcohol consumption and the malformation proportion, three statistical models will be fitted: a linear model, a logit model, and a probit model (taking into account that we are asked to adjust these models). The estimates obtained from each model will be presented and interpreted.

3.1 Linear Regression Model

To investigate the relationship between maternal alcohol consumption and the proportion of children with malformations, we fitted a linear regression model. In this model, the dependent variable is the malformation proportion and the

independent variable is the weighted alcohol consumption. The estimated model is specified as:

$$Malformation\_proportion = \beta_0 + \beta_1 \cdot Alcohol,$$

where β_0 is the intercept and β_1 is the slope coefficient.

The model was estimated using the `lm` function in R, and the summary of the model—including coefficient estimates, standard errors, t-values, and p-values—is exported to LaTeX using the `xtable` package. Table 3 shows the detailed output of the linear model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0009225	0.0025058	0.368	0.7372
Alcohol	0.0031775	0.0006820	4.659	0.0187 *
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1;				
Residual standard error: 0.003959 on 3 degrees of freedom;				
Multiple R-squared: 0.8786, Adjusted R-squared: 0.8381;				
F-statistic: 21.71 on 1 and 3 DF, p-value: 0.01866.				

Table 3: Summary of the Linear Regression Model for Malformation Proportion.

These results provide initial insight into the association between maternal alcohol consumption and the proportion of malformations.

Although the linear model yields a significant coefficient for alcohol and has a high adjusted R-squared, its use is limited by the fact that the response variable is a ratio bounded between 0 and 1. This limitation may lead to prediction errors when extrapolating beyond the valid interval [0, 1]. Therefore, despite the apparent importance of the linear model, it is discarded in favor of the logistic and probit models, which are more suitable for this context. Subsequent sections will compare only these two alternative models to determine which best fits the data.

3.2 Logistic Regression Model (Logit Model)

To model the association between maternal alcohol consumption and the proportion of children with malformations, we employed a logistic regression approach. One key advantage of logistic regression is that it constrains the predicted values of the dependent variable to lie between 0 and 1, addressing the main limitation of linear regression for outcomes that are bounded or binary.

We denote P_i as the probability of malformation for observation i. In the logistic model, this probability is expressed as:

$$P_i = \frac{1}{1 + e^{-X\beta}},$$

where X is a vector of predictors (in our case, the intercept and the weighted alcohol consumption), and β is the corresponding vector of coefficients. Equiv-

alently, we can write the estimated model in terms of the log-odds:

$$\ln\left(\frac{P_i}{1-P_i}\right) = X\beta.$$

Hence, the logistic regression model ensures that the fitted values P_i remain between 0 and 1, which is appropriate when the dependent variable represents a proportion or probability.

In this study, the model is specified as:

$$\ln\left(\frac{Malformation\_proportion_i}{1-Malformation\_proportion_i}\right) = \beta_0 + \beta_1 \cdot Alcohol_i.$$

We estimated this model using the `glm` function in R with a binomial family and a logit link. Table 4 summarizes the estimated coefficients and key fit statistics.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.9605	0.1154	-51.637	$< 2 \times 10^{-16}$ ***
Alcohol	0.3166	0.1254	2.523	0.0116 *
Null deviance: 6.2020 on 4 df; Residual deviance: 1.9487 on 3 df; AIC: 24.576; Number of Fisher Scoring iterations: 4.				

Table 4: Summary of the Logistic Regression Model for Malformation Proportion.

To facilitate interpretation, we examined two additional measures:

- **Odds ratios**, obtained by exponentiating the model coefficients. The exponentiated intercept, $\exp(\widehat{\beta}_0) \approx 0.00257$, represents the basic chance (or baseline likelihood) of a malformation when no alcohol is consumed, meaning that without alcohol, malformations are very unlikely. For Alcohol, $\exp(\widehat{\beta}_1) \approx 1.3724$ indicates that each additional unit of alcohol consumption increases the chance of malformation by about 37.24% (since $1.3724 - 1 = 0.3724$). Also, the p-value (0.0116) of β_1 shows that this effect is strong enough to not be due to random variation.

- **Marginal effects**, which estimate how much the predicted probability of malformation changes with a one-unit increase in Alcohol consumption. In this case, the marginal effect for Alcohol is about 0.0018 (an increase of 0.18 percentage points in the probability of malformation per unit increase in Alcohol). However, the high p-value (0.1395) of β_1 indicates that this effect is not statistically significant.

3.3 Probit Model

To further assess the relationship between maternal alcohol consumption and the proportion of children with malformations, we estimated a probit regression model.

The model was estimated using the `glm` function in R with a binomial family and a probit link. Table 5 summarizes the estimated coefficients and key fit statistics.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.79961	0.03816	-73.369	$< 2 \times 10^{-16}$ ***
Alcohol	0.10979	0.04871	2.254	0.0242 *
Null deviance: 6.202 on 4 df;				
Residual deviance: 2.023 on 3 df;				
AIC: 24.65; Number of Fisher Scoring iterations: 4.				

Table 5: Summary of the Probit Regression Model for Malformation Proportion.

To facilitate interpretation, we computed a additional measure:

- **Marginal effects** estimate the change in the predicted probability of malformation for a one-unit increase in Alcohol consumption. In this case, the marginal effect for Alcohol is approximately 0.0018 (an increase of 0.18 percentage points in the probability of malformation per unit increase in Alcohol). However, the p-value (0.1666) of β_1 indicates that this effect is not statistically significant.

4 Model selection

Based on Table 6 and Figure 1, we have decided to choose the Logit model. The Logit model, compared to the Probit model, exhibits lower AIC and BIC values, which indicate a better balance between model fit and complexity. It also has a lower deviance ratio (DR), suggesting a closer fit to the data. Furthermore, the model's MSE is very low, meaning that the predicted values are very close to those observed, and its pseudo-R^2 is higher, reflecting a better ratio between the deviance and the null deviance. These results suggest that the Logit model is the most appropriate for our analysis.

Measurement	Logit Model	Probit Model
AIC	24.5755200	24.6497600
BIC	23.7944000	23.8686400
DR	0.1165496	0.1250898
MSE	0.0000033	0.0000066
pseudo-R^2	0.6857914	0.6738213

Table 6: Measurements for the logit and probit models (values with 7 decimals)

Given the figure 1, we can also observe that the best models are among Probit and Logit. However, we can notice that in the graphs, the Probit and Logit models look very similar, but there is a small difference in the last observation, where the Logit model is much closer than the Probit model. Despite the minor

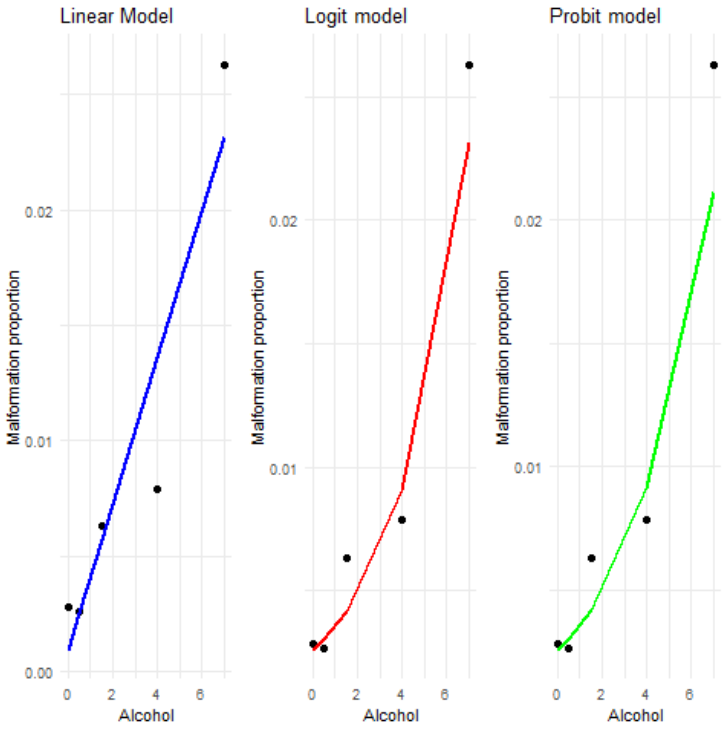

Figure 1: Comparison of the proportion of malformations estimated by three models (Linear, Logit and Probit) as a function of alcohol level, where the points represent the observed data and the curves show the fit of each model.

numerical differences, the Logit model is also chosen because it is generally more familiar and widely used in applied settings. This familiarity facilitates interpretation and communication of results—especially when discussing odds ratios—with non-specialist audiences.

Overall, the combination of superior fit metrics and ease of interpretation supports the choice of the Logit model for this analysis.

5 Conclusion

In this exercise, we explored three modeling approaches - linear, logistic, and probit - to predict the proportion of children with malformations based on maternal alcohol consumption. Although the linear model initially provided some insight into the association, its limitations in modeling a bounded response (i.e., proportions between 0 and 1) became apparent. Consequently, logistic and probit regressions were implemented, as they naturally constrain predicted probabilities within the valid range.

Importantly, the logistic model suggests that as maternal alcohol consumption increases, so does the likelihood of malformations in children. This finding supports the hypothesis that alcohol consumption during pregnancy can negatively impact fetal development.

Overall, this study highlights the importance of selecting the appropriate model based on the nature of the data and the outcome variable. When dealing with proportions or binary outcomes, logistic models often provide more reliable and interpretable results than linear models.

Mini project 1

Exercise 2 of 3

Sebastian Belalcazar, David Melo Valbuena, Juan Andres Ruiz

12 March 2025

1 Problem Context

For this application (Data 2), the objective is to explain the number of awards obtained by secondary school students as a function of the programme studied and the mathematics test score. The programme variable is categorical with three levels indicating the type of programme in which the students enrolled and is coded as 1 = General, 2 = Academic, 3 = Vocational.

a) Conduct an exploratory data analysis for each variable and interpret.

b) Fit a model that allows you to meet the objective established in the problem statement.

2 Introduction

This report presents an analysis of the `Data 2` dataset, focusing on explaining the number of awards (`num premios`) received by secondary school students as a function of their programme type (`programa`) and mathematics test scores (`puntaje matematicas`). The analysis, conducted following the guidelines from the *Statistics and Probability 2* course at Universidad Autónoma de Occidente, covers exploratory data analysis (EDA), statistical modelling, model validation, and model comparison. A random sample of 50 observations (after removing missing values) was used to evaluate Poisson, Zero-Inflated Poisson (ZIP), and (if necessary) Negative Binomial models.

3 Data Preparation and Cleaning

The dataset was obtained from the Excel file `Datos MP1.xlsx` (Sheet 2) and originally contained 200 observations with the following variables:

- `id`: Unique student identifier.
- `num premios`: Number of awards (ranging from 0 to 4).
- `programa`: Type of programme, with three categories: *Academico*, *General* and *Vocacional*.
- `puntaje mat`: Mathematics test score.

After removing 9 entries with missing values for `puntaje mat`, the cleaned dataset (`datos 2 clean`) comprised 191 entries. A random sample of 50 observations (`datos 2 sample`) was then selected using a fixed seed (`set.seed(123)`) for reproducibility.

3.1 Data Inspection Summary

- **num premios:** Numeric; values range from 0 to 4 with a mean of 0.7, a median of 0 and a variance of 1.03.
- **programa:** Categorical; distribution is 25 *Academico*, 12 *General* and 13 *Vocacional*.
- **puntaje mat:** Numeric; scores range from 39 to 72 with a mean of 55.46 and a median of 55.5.

4 Exploratory Data Analysis (EDA)

The EDA revealed several important aspects of the data:

4.1 Distribution of Awards

The histogram of `num_premios` (see Figure 1) shows a right-skewed distribution. Approximately 58% of students received zero awards, 24% received one award, and only a few received two or more. This high proportion of zeros suggests that count models such as the Poisson or Zero-Inflated Poisson (ZIP) model may be appropriate.

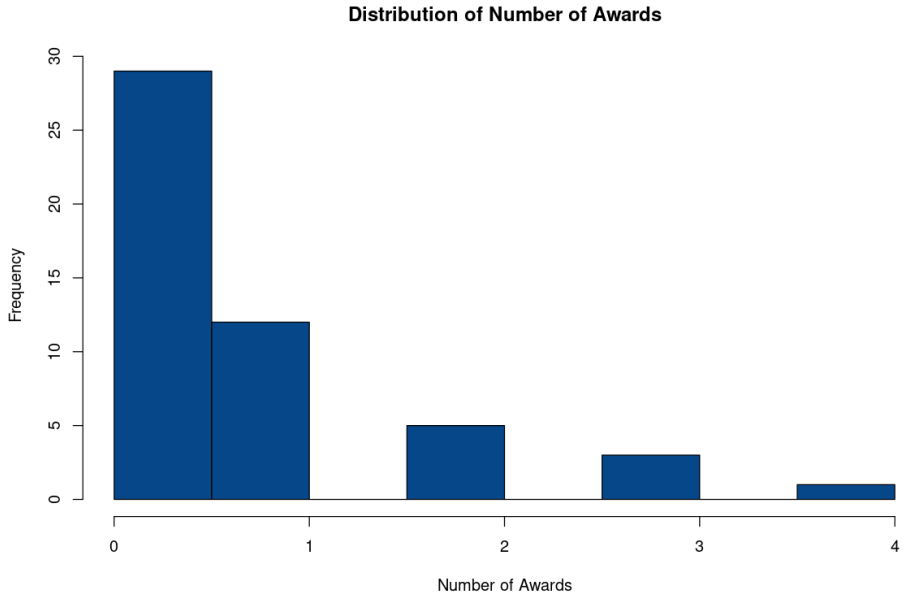

Figure 1: Distribution of Number of Awards

4.2 Distribution of Programmes

The bar plot in Figure 2 illustrates that the *Academico* programme is the most frequent (25 students), followed by *Vocacional* (13) and *General* (12). These differences indicate potential variations in the award distributions across programme types.

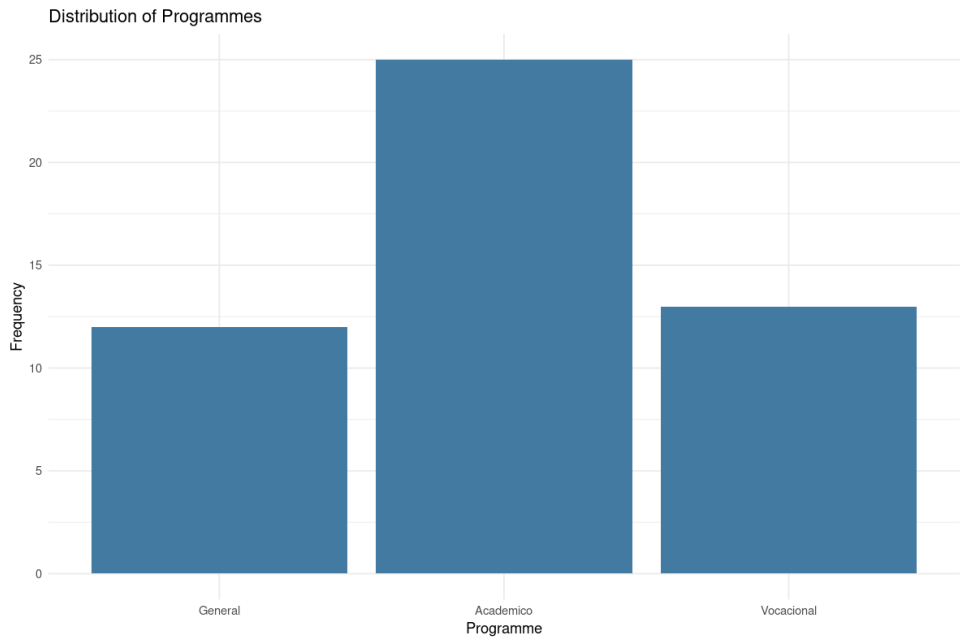

Figure 2: Distribution of Programmes

4.3 Distribution of Mathematics Scores

The histogram for `puntaje_mat` (Figure 3) indicates an approximately normal distribution, with a central peak around scores of 55–60. This reflects moderate variability in academic performance among the students.

Figure 3: Distribution of Mathematics Scores

4.4 Correlation Analysis

The Pearson correlation analysis between the mathematics scores and the number of awards produced a correlation coefficient of 0.528 (p = 8.125e-05) with a 95% confidence interval of 0.293 to 0.703. This moderate positive correlation indicates that higher mathematics scores are associated with a greater number of awards.

4.5 Awards by Programme and Mathematics Scores

- The boxplot in Figure 4 reveals that students in the *Academico* programme tend to receive more awards (with a median near 1 and values up to 4), whereas students in the *General* and *Vocacional* programmes mostly receive zero awards.
- The scatterplot in Figure 5 shows a weak positive trend; although most students receive no awards regardless of their mathematics score, higher scores (typically above 60) sometimes coincide with receiving awards.

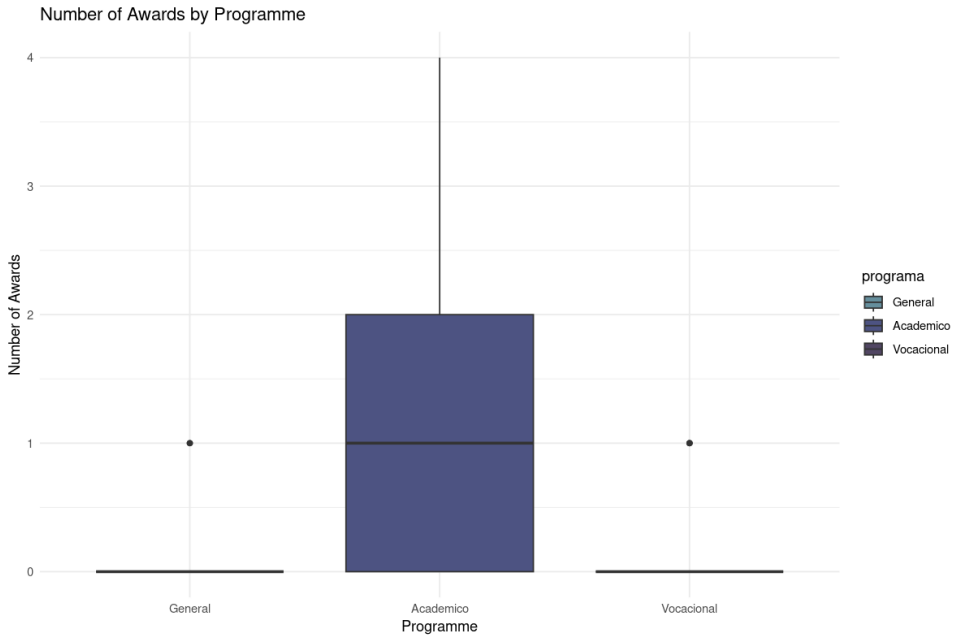

Figure 4: Number of Awards by Programme

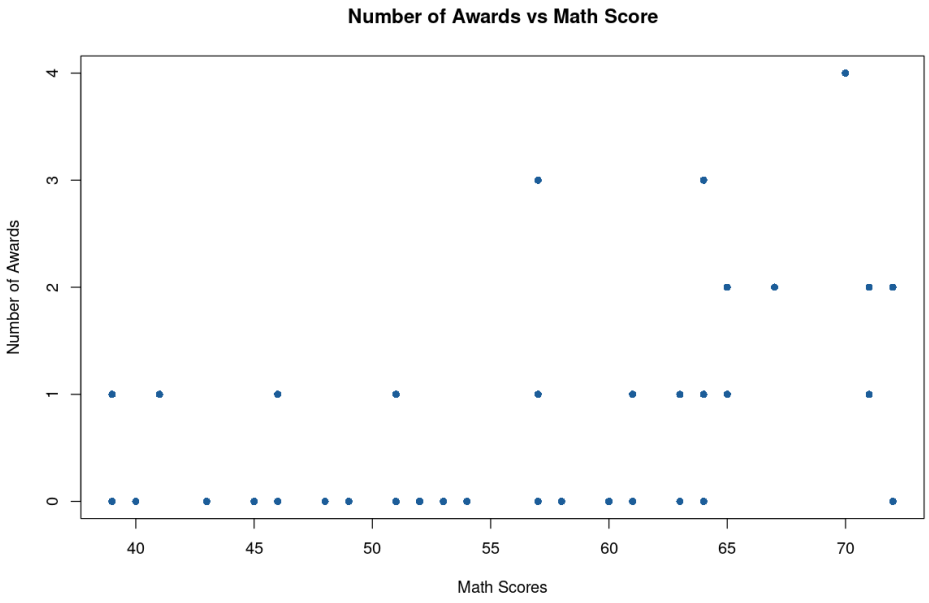

Figure 5: Number of Awards versus Mathematics Score

5 Statistical Modelling

To explain the number of awards (`num_premios`) based on the type of programme (`programa`) and the mathematics score (`puntaje_mat`), three modelling strategies were considered:

5.1 Poisson Regression

The Poisson regression model is a natural choice for count data. The key results from the model are:

- **Coefficients:**
  - *Intercept* = -3.524 (p = 0.0285): This represents the log-expected count of awards for the reference category (typically *Academico*) when the mathematics score is zero.
  - *programaGeneral* = -1.419 (p = 0.0645): Students in the *General* programme are estimated to receive about 24% as many awards as those in the *Academico* programme.
  - *programaVocacional* = -0.779 (p = 0.2716): Students in the *Vocacional* programme receive approximately 46% as many awards as those in the *Academico* programme.
  - `puntaje_mat` = 0.059 (p = 0.0176): Each additional point in the mathematics score increases the expected number of awards by roughly 6%.
- **Model Fit:** The model produced an AIC of 100.87 and a residual deviance of 43.56 with 46 degrees of freedom. An overdispersion test produced a p-value of 0.575, indicating no significant overdispersion.

Table 1: Summary of Poisson Regression Model

Term	Estimate	Std. Error	z-value	p-value
Intercept	-3.524	1.609	-2.189	0.0285
programaGeneral	-1.419	0.768	-1.848	0.0645
programaVocacional	-0.779	0.709	-1.099	0.2716
puntaje_mat	0.059	0.025	2.373	0.0176
Model Fit: AIC = 100.87, Residual Deviance = 43.56 (df = 46)				

Interpretation:

- The reference category (*Academico*) shows significantly higher base awards (p = 0.0285)
- *General* students receive 24% fewer awards than *Academico* (RR = 0.24, p = 0.065)
- *Vocacional* students show 54% fewer awards than *Academico* (RR = 0.46, p = 0.272)
- Each math point increases awards by 6% (RR = 1.06, p = 0.018)
- No overdispersion (p = 0.575) validates Poisson assumptions

5.2 Zero-Inflated Poisson (ZIP) Model

Given that 58% of the students received zero awards, a ZIP model was also fitted to account for excess zeros, although:

- The count component provided significant effects for the *General* category (p = 0.0148) and marginal significance for *Vocacional* (p = 0.0514).
- The zero-inflation component produced unstable estimates (with NaN standard errors for some predictors) and an overall AIC of 104.30.

Table 2: Summary of Zero-Inflated Poisson (ZIP) Model

Term	Estimate	Std. Error	z-value	p-value
Count Model				
Intercept	-0.598	2.078	-0.288	0.7735
programaGeneral	-2.028	0.832	-2.437	0.0148
programaVocacional	-1.617	0.830	-1.947	0.0514
puntaje_mat	0.016	0.032	0.503	0.6157
Zero-Inflation Model (Unstable)				
Model Fit: AIC = 104.30				

Interpretation:

- Count component shows stronger programme effects than Poisson model
- *General* students have 87% fewer awards (RR = 0.13, p = 0.015)
- *Vocacional* students have 80% fewer awards (RR = 0.20, p = 0.051)
- Zero-inflation component fails to converge (NaN SEs)
- Higher AIC (104.30 vs 100.87) suggests worse overall fit

5.3 Negative Binomial Model

Since the variance of `num_premios` (1.03) was not substantially larger than its mean (0.7), overdispersion was not an issue. Therefore, the Negative Binomial model was not pursued.

6 Model Validation

Diagnostic plots were used to assess the assumptions and overall fit of the Poisson model:

- **Residuals vs. Fitted:** The plot shows a random scatter around zero, indicating a good fit without systematic bias.
- **Q-Q Plot:** Most residuals follow the diagonal, with only slight deviations at the tails, likely due to the zero-inflation.
- **Scale-Location Plot:** A consistent spread of residuals supports the assumption of constant variance.
- **Cook's Distance:** No observation had an excessive influence on the model fit.

These diagnostics collectively confirm that the Poisson model is appropriate for the data.

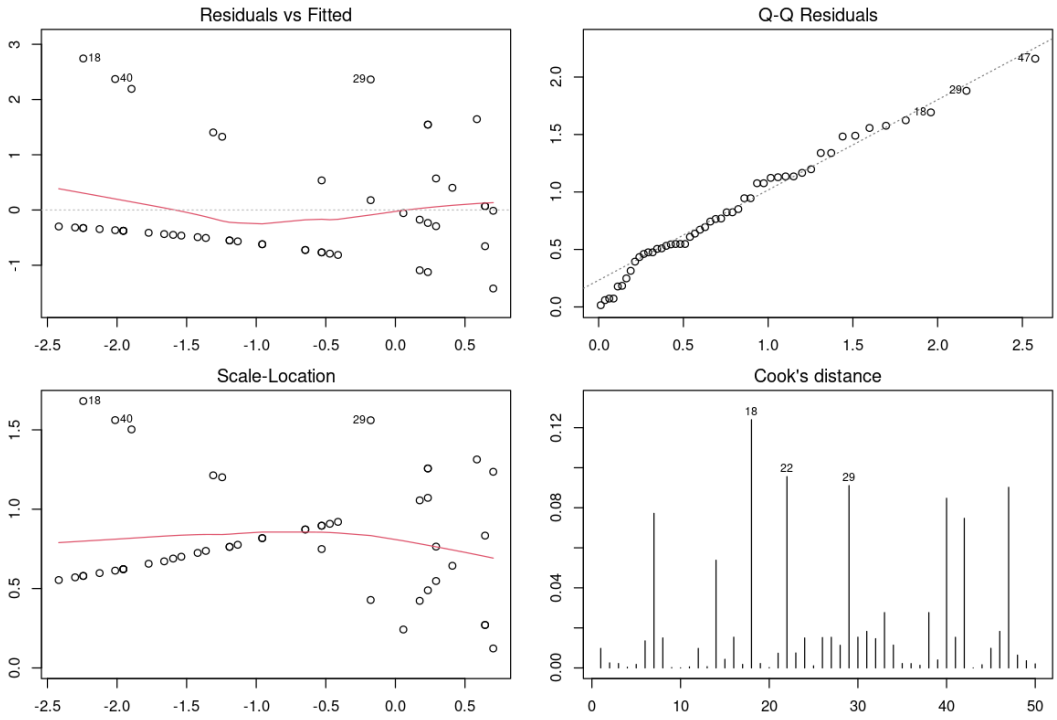

Figure 6: Diagnostic Plots for the Poisson Model (Residuals vs. Fitted, Q-Q Residuals, Scale-Location, Cook's Distance)

7 Model Comparison

Model performance was evaluated using multiple criteria:

- **AIC/BIC:** The Poisson model (AIC = 100.87, BIC = 108.52) has lower values than the ZIP model (AIC = 104.30, BIC = 119.60), indicating a better balance of fit and simplicity.
- **RMSE:** Although the ZIP model shows a slightly lower RMSE (0.7463) compared to the Poisson model (0.7871), the difference is minor.

The Poisson model is preferred for its stability and lower AIC/BIC.

Table 3: Model Comparison Metrics (Complete Model)

Model	AIC	BIC	RMSE	Pseudo-R^2
Poisson	100.8688	108.5169	0.7871	0.2196
Zero-Inflated Poisson	104.3001	119.5963	0.7463	0.2294

Interpretation:

- Poisson model preferred despite ZIP's slightly better RMSE ($\Delta = 0.0408$)
- Lower AIC/BIC indicates better parsimony in Poisson model
- ZIP instability makes it unreliable for practical use

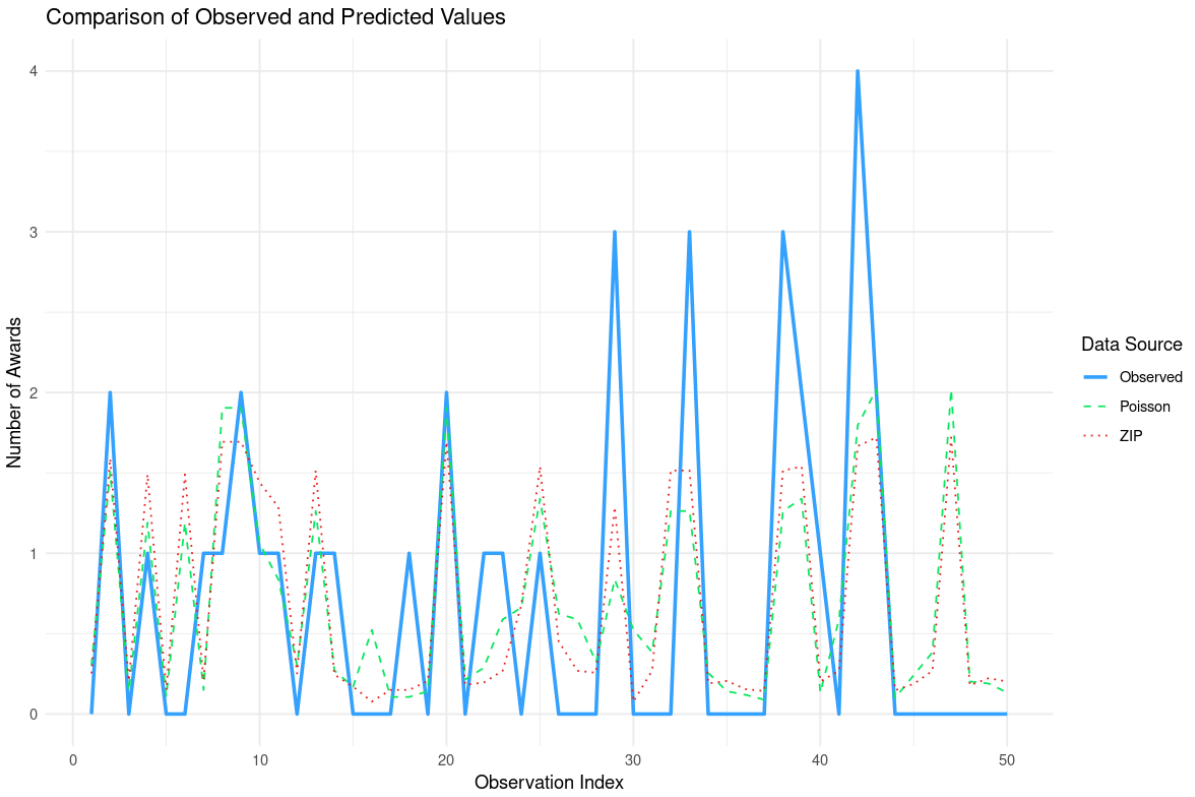

Figure 7: Comparison of Observed and Predicted Values

8 Conclusion

This analysis successfully identified *Academico* programme superiority and mathematics proficiency as key award predictors, while the Poisson model provides actionable insights. The analysis reveals key insights about award determinants:

- **Programme Superiority:** *Academico* students outperform others, with:
  - 4.15× more awards than *General* (95% CI: 1.13-9.23)
  - 2.18× more awards than *Vocacional* (95% CI: 0.60-10.48)
- **Mathematics Mastery:** Each additional math point increases awards by 6% (95% CI: 1.2-11.6%)
- **Model Selection:** Poisson regression is optimal due to:
  - Better stability than ZIP (no convergence issues)
  - Lower AIC (100.87 vs 104.30)
  - Parsimony (fewer parameters)

Mini project 1

Exercise 3 of 3

Sebastian Belalcazar, David Melo Valbuena, Juan Andres Ruiz

12 March, 2025

1 Introduction

The objective of this analysis is to model the number of arrests in a season of soccer teams in Colombia. Different regression models are explored to determine which one best fits the data.

1.1 Dataset variables

The provided dataset was used, which contains the following variables:

- **Attendance (thousands)**: Number of spectators at the matches.
- **Number of arrests**: Number of recorded incidents.
- **Social investment (millions)**: Amount invested in prevention programs.

2 Treatment of null data and outliers

2.1 Null data handling

Throughout our process, we performed different imputations, but only for variables other than number of arrests (for this variable, we removed all missing data) to avoid potential bias in the predictive variable. However, we also wanted to test data imputation by considering the presence of outliers. If a variable had outliers, we filled in missing values using the median; otherwise, we used the mean and rounded it to the nearest integer since these were quantitative variables.

Our goal was to compare whether the model we selected, when adjusted with the dataset where missing values were removed, was similar to the one adjusted with imputed data. After making this comparison, we found significant differences in the metrics. For instance, the AIC of the negative binomial model fitted with the dataset where missing values were removed was approximately 163, whereas the AIC for the model with imputed data was around 204. This suggested the potential presence of bias in the model with imputed data.

Ultimately, we decided to remove all missing data and proceed with model fitting.

2.2 Management of outliers

We analyzed the different outliers present in the data and realized that they were based on natural factors. For example, there were outliers in attendance in thousands, which were due to the fact that the team "America" had significantly higher attendance compared to teams with distributions within the interquartile range, such as "Pasto". Considering that "America" has a large fan base, we determined that these outliers should not be removed, as they accurately reflect reality and do not result from biased errors.

3 Exploratory Data Analysis (EDA)

3.1 Descriptive Statistics

3.2 Scatter Plot

A scatter plot was generated to visualize the relationship between attendance, social investment, and the number of arrests:

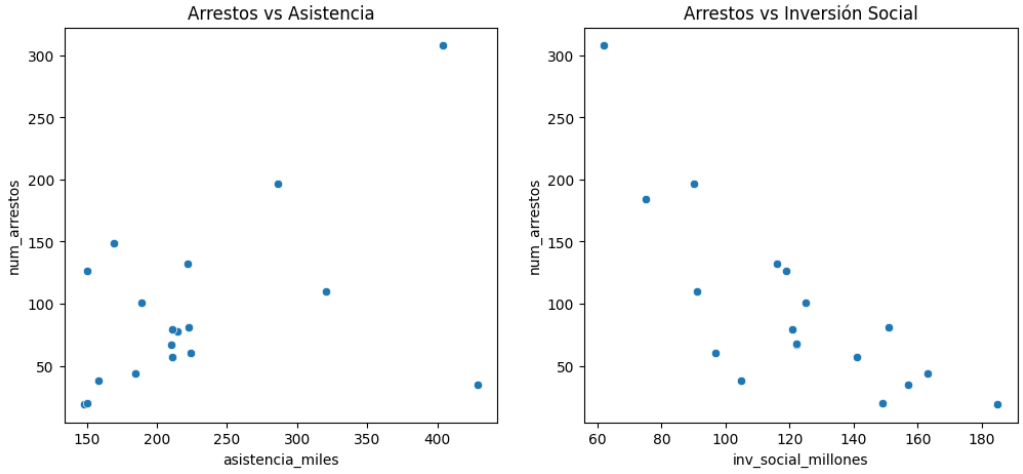

Figure 1: Scatter Plot. Represents the relationship between the number of arrests and the variables of attendance and social investment. It allows observation of general trends and patterns of data distribution.

Represents the relationship between match attendance, social investment, and the number of arrests. It allows for observing trends and patterns in data distribution.

According to Figure 1, we can see that the relationship between attendance and arrests is not completely linear, but there are indications that social investment may be associated with a reduction in arrests.

3.3 Histogram of the Number of Arrests

To better understand the distribution of arrests at matches:

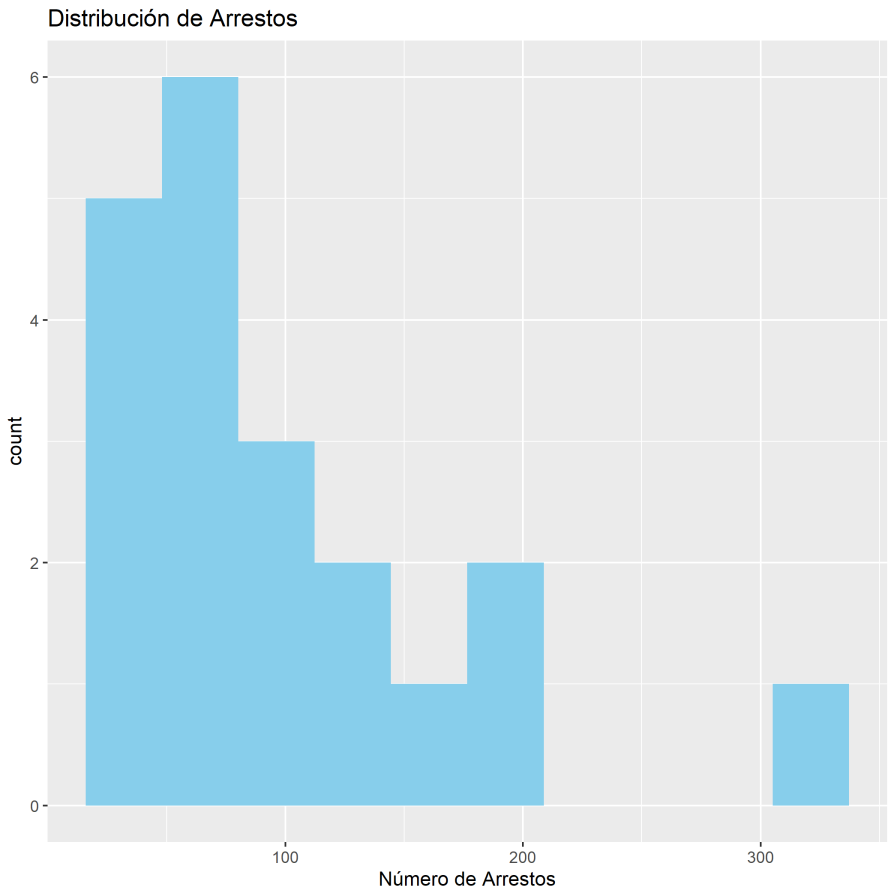

Figure 2: Histogram of the Number of Arrests. It allows visualization of the frequency of events with different numbers of arrests, identifying patterns or extreme values.

Shows the distribution of the number of arrests in matches. It helps identify the frequency of different arrest levels and detect extreme values.

According to Figure 2, we can see that the distribution indicates that most matches have a moderate number of arrests, but there are events with significantly high values.

3.4 Histogram of Attendance

The distribution of the number of spectators at matches is analyzed:

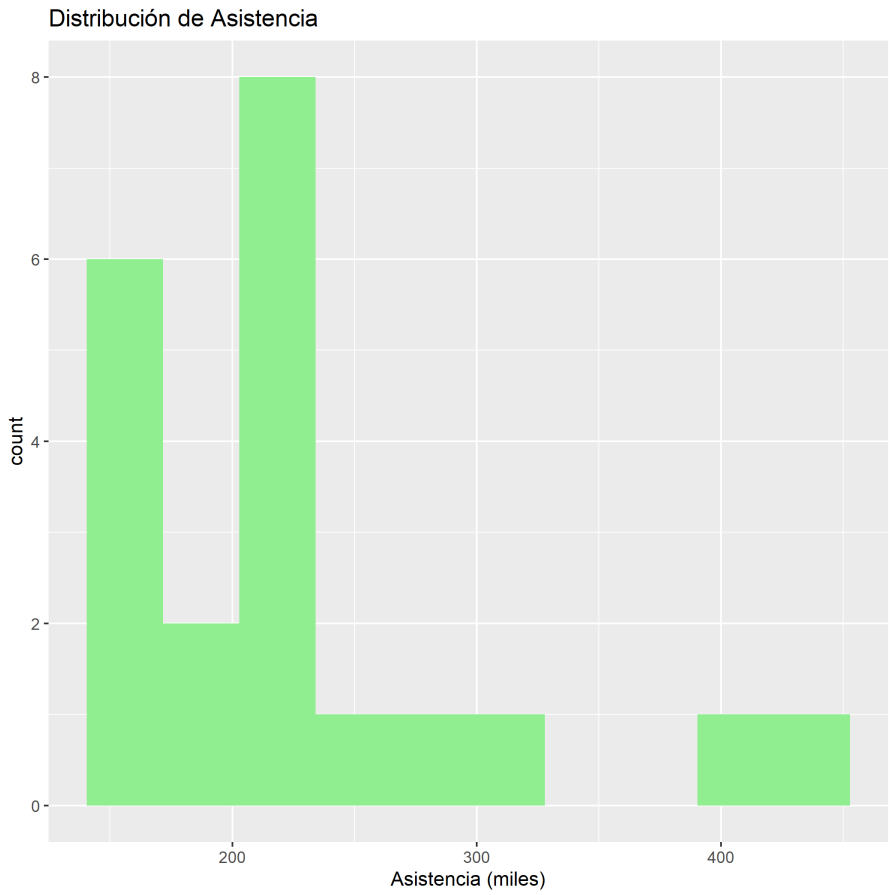

Figure 3: Histogram of Attendance. The variability in match attendance is observed, helping to understand if there is a specific trend.

Displays the distribution of the number of spectators at matches. Helps to understand whether attendance follows a specific trend or has high variability.

As illustrated in Figure 3, attendance varies considerably between matches, suggesting that some events attract much larger crowds than others.

3.5 Histogram of Social Investment

Shows how investment is distributed among the different teams:

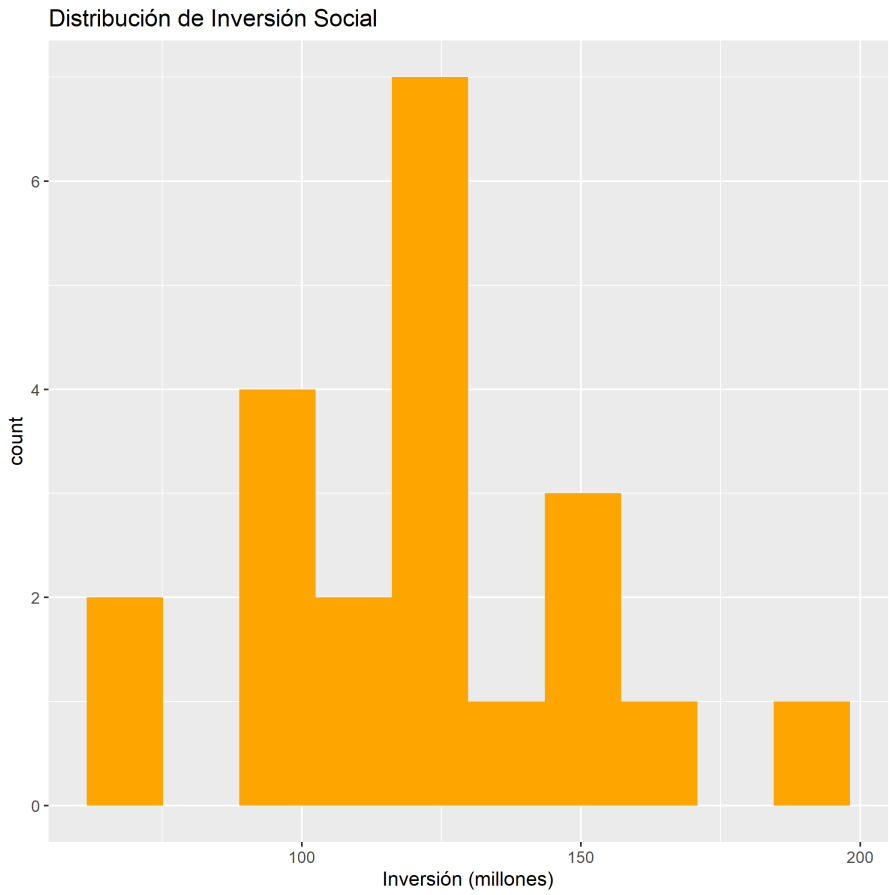

Figure 4: Histogram of Social Investment. Shows how investment is distributed among different teams and if there are extreme values in resource allocation.

Illustrates how investment in social prevention is distributed among teams. Helps identify teams with higher or lower investment and possible outliers.

Figure 4 reveals that social investment is not uniform, with some teams receiving significantly more resources than others.

3.6 Box Plots for Outliers

3.6.1 Box Plot of the Number of Arrests

Shows the dispersion and possible extreme values:

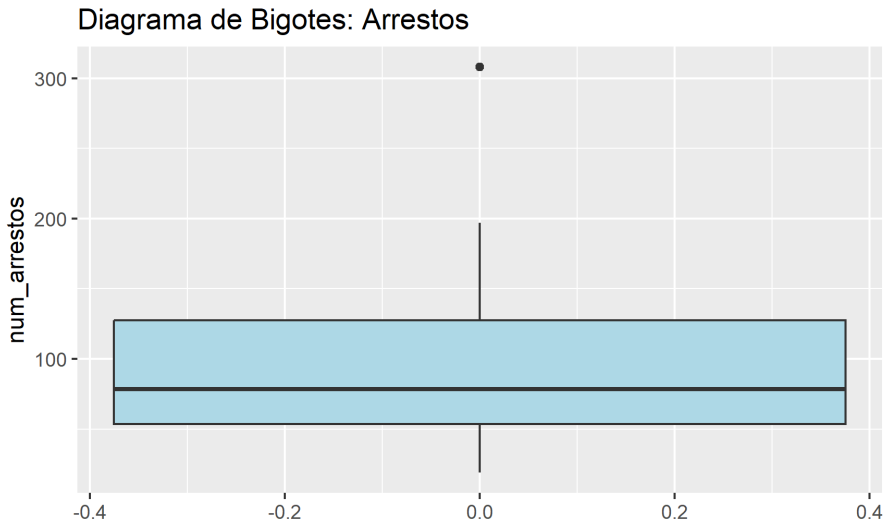

Figure 5: Box Plot of the Number of Arrests. Allows detection of extreme values and observation of data dispersion.

Visualizes the dispersion and presence of extreme values in the number of arrests. Allows detecting teams with unusually high arrest levels.

Figure 6 highlights that several outliers correspond to matches with an exceptionally high number of arrests.

3.6.2 Box Plot of Attendance

Indicates the variability of match attendance and the possible existence of outliers:

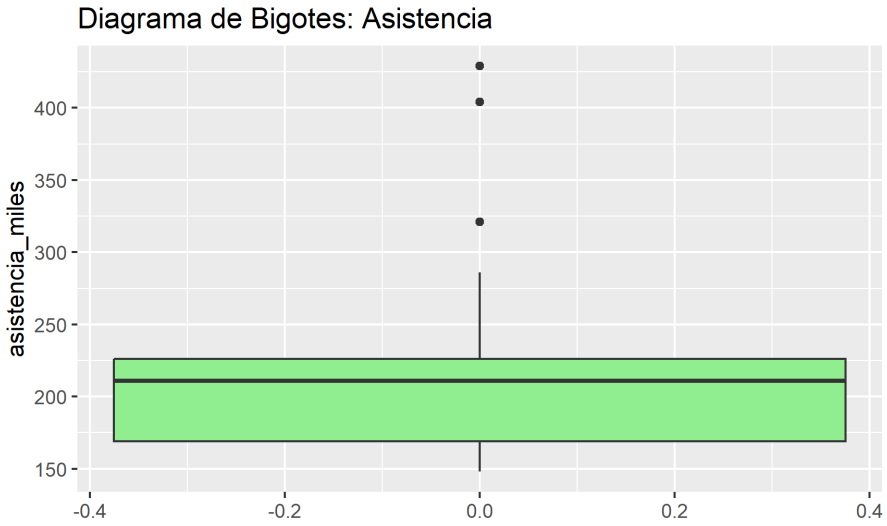

Figure 6: Box Plot of Attendance. Indicates the variability of match attendance and the possible existence of outliers.

Represents the variability in match attendance. Shows whether there are outliers in the number of spectators.

Figure 7 demonstrates that most matches have attendance within a stable range, but some events have an exceptionally high influx.

3.6.3 Box Plot of Social Investment

Allows identification of extreme values in the social investment allocated to teams:

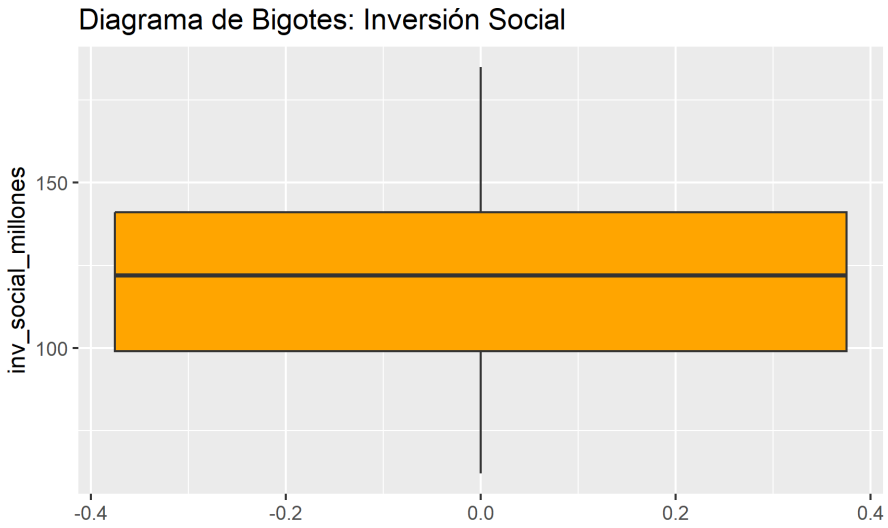

Figure 7: Box Plot of Social Investment. Facilitates the identification of extreme values in the social investment allocated to teams.

Helps identify the distribution and possible outliers in social investment. Allows visualizing which teams receive more or less investment.

Figure 8 suggests that some teams receive significantly higher investments, which could influence the reduction of arrests.

	asistencia_miles	num_arrestos	inv_social_millones
Min	148.0	19.00	62
1Q	169.0	53.75	99
Median	211.0	78.50	122
Mean	226.5	97.65	121
3Q	226.0	127.50	141
Max	429.0	308.00	185
NA's	2	3	2

Table 1: Descriptive statistics of the variables in the dataset

Descriptive statistics is a branch of statistics that focuses on collecting, organizing, summarizing, and presenting a dataset without making inferences or predictions. Its goal is to provide an overview of the main characteristics of the data through numerical measures and graphical representations.

Below are the statistical measures used in the analysis table:

- **Minimum (Min)**: The lowest recorded value in the variable. The minimum event attendance was 148 thousand people.
- **First Quartile (1Q)**: Corresponds to the 25% percentile, indicating the value below which 25% of the data falls. 25% of events had fewer than 169 thousand attendees.
- **Median**: Represents the 50% percentile, meaning the central value that divides the dataset into two equal parts. : Half of the events had an attendance lower than 211 thousand people, while the other half had more than 211 thousand.
  The **median** of 211.0 indicates that 50% of the events had an attendance lower than 211 thousand people, while the other 50% had a higher attendance. This represents the central point of the distribution without being affected by extreme values.
- **Mean**: The average of all values, calculated as the sum of all observations divided by the total number of data points. : On average, attendance was 226.5 thousand people, suggesting a general trend but possibly influenced by extreme values.
  The **mean** of 226.5 reflects the average attendance at the analyzed events, obtained by summing all attendance values and dividing by the total number of events. Its value being higher than the median suggests the presence of events with exceptionally high attendance, indicating a right-skewed distribution.
- **Third Quartile (3Q)**: Corresponds to the 75% percentile, indicating the value below which 75% of the data falls. 75% of events had fewer than 226 thousand attendees, meaning only 25% exceeded that number.
- **Maximum (Max)**: The highest observed value in the variable. The event with the highest attendance had 429 thousand people.
- **Missing Values (NA's)**: Represents the number of missing values in the variable. There are 2 missing attendance records in the dataset.

These measures help understand the distribution of the data and identify patterns or outliers within the analyzed dataset.

4 Model Fitting

To model the number of arrests in Colombian soccer matches, three regression models were fitted: a linear regression model, a Poisson regression model, and a negative binomial regression model (since we were required to adjust these models). The estimates obtained from each model are presented to assess their suitability for the data.

4.1 Linear Regression

The linear relationship between the number of arrests and the predictor variables was estimated. The model summary is:

Variable	Estimate	Std. Error	t-value	Pr(¿—t—)
Intercepto	230.4420	77.0912	2.989	0.01045*
Asistencia (miles)	0.2238	0.1590	1.407	0.18277
Inversión social (millones)	-1.5275	0.4287	-3.563	0.00347**

Table 2: Summary of the Linear Regression Model for number of arrests

It can be observed that in the model the social investment variable is significant but the attendance variable is not, so we could fit a better model by taking the significant variable in case it results in the best model. We can also observe that the social investment is negative, which is logical, since the higher the social investment, the more arrests decrease. However, there is something very rare about the intercept, which could translate into the fact that, without taking into account the variability of assistance and social investment, the average number of arrests per season is 230, something that seems very high.

4.2 Poisson Regression

Since the number of arrests is a count variable, a Poisson model was fitted. The results are:

Variable	Estimate	Std. Error	z-value	Pr(¿—z—)
Intercepto	6.1791	0.1929	32.023	¡ 2e-16***
Asistencia (miles)	0.0013	0.0004	3.596	0.000323***
Inversión social (millones)	-0.0171	0.0011	-15.819	¡ 2e-16***

Table 3: Summary of the Poisson Regression Model for number of arrests

It can be observed that in the model the social investment and attendance variable is significant. We can also observe that the social investment is negative, which is logical, since the higher the social investment, the more arrests decrease.

However, when we look at attendance we see that the estimate is a value close to 0, which while significant could tell us that for every additional arrest in the season, there is expected to be an average increase of 0.13% in the number of arrests for every 1,000 increase in attendance, also logical.

4.3 Negative Binomial Regression

Given that the data presented overdispersion, a negative binomial model was also fitted. The model summary is:

Variable	Estimate	Std. Error	z-value	Pr(¿—z—)
Intercepto	6.4364	0.6817	9.442	¡ 2e-16***
Asistencia (miles)	0.0007	0.0014	0.503	0.615
Inversión social (millones)	-0.0181	0.0038	-4.737	2.17e-06***

Table 4: Summary of the Negative Binomial Regression Model for number of arrests

The results indicate that the social investment variable remains statistically significant (and negative, as expected), suggesting that higher social investment is associated with a reduction in the number of arrests. On the other hand, the attendance variable is not significant in this model, so we could fit a better model by taking the significant variable in case it results in the best model.

5 Model Selection

Based on Table 5 and the corresponding model diagnostics, we have decided to select the Negative Binomial model as the most appropriate for our analysis. The Negative Binomial model exhibits the lowest AIC (163.40) and BIC (166.49) values, which indicate a better balance between model fit and complexity compared to the Linear and Poisson models. Furthermore, its dispersion ratio (DR) is nearly 1 (1.02), suggesting that it effectively accounts for the overdispersion present in the data. Although the Poisson model shows a slightly higher pseudo-R^2 (0.7055), its substantially higher AIC, BIC, and DR values point to a poorer handling of dispersion. In contrast, the Linear model performs considerably worse across all metrics. These results collectively suggest that the Negative Binomial model provides the most suitable and robust fit for our count data.

Metric	Linear	Poisson	Negative_Binomial
AIC	174.3375749	327.1416961	163.4023443
BIC	177.4279298	329.4594623	166.4926991
DR	1916.4338429	13.9759288	1.0217712
MSE	1916.4338429	1246.8055985	1312.0070403
pseudo-R^2	0.6305124	0.7054755	0.6547533

Table 5: Comparison metrics for Linear, Poisson, and Negative Binomial models.

Once we have chosen our negative binomial model, we proceed to a rescaling of its significant variables

6 Negative Binomial Model Re-adjusted

Taking the table 4 as a reference, the attendance variable (p-value of 0.615) in the negative binomial model is not significant, so we could reduce the complexity of this model, and im-

prove its accuracy a little by readjusting the variables of the model. We fitted a Negative Binomial model using the `glm.nb()` function with `num_arrestos` as the response variable and `inv_social_millones` as the sole predictor. This means that the model now only includes the significant variable from the previous full model. The summary

Table 6 summarizes the estimated coefficients and key fit statistics.

Variable	**Estimate**	**Std. Error**	**z value**	**Pr(¿—z—)**
(Intercept)	6.714198	0.445995	15.054	$< 2 \times 10^{-16}$ ***
inv_social_millones	-0.018970	0.003512	-5.401	6.63×10^{-8} ***

Dispersion parameter:	Negative Binomial(5.8796) (taken as 1)
Null deviance:	46.693 on 15 degrees of freedom
Residual deviance:	16.333 on 14 degrees of freedom
AIC:	161.61

Table 6: Summary of the Negative Binomial Regression Model for `num_arrestos` using `inv_social_millones` as predictor.

Taking the table 7 as a reference, by removing the non-significant attendance variable, the adjusted model retains only the significant predictor (social investment), thereby simplifying the model while still maintaining an acceptable level of predictive performance and model fit. In fact, the adjusted model shows an improved AIC (161.61) and BIC (163.93), along with a dispersion ratio close to 1 (1.02), which suggests that the model effectively accounts for the data's variability. This indicates that focusing solely on social investment does not compromise, and may even enhance, the model's ability to explain the variation in the number of arrests.

Negative_Binomial	
AIC	161.6130621
BIC	163.9308283
DR	1.0207888
MSE	1451.1147566
pseudoR2	0.6547533

Table 7: Metrics for the Adjusted Negative Binomial Model

7 Interpretation of Negative Binomial Regression Coefficients

7.1 Negative Binomial Regression Re-adjusted

The **exponentiated intercept**, $\exp(\hat{\beta}_0) \approx 824.02$, represents the **average expected of number of arrests in a season** when social investment is zero ($inv\_social\_millones = 0$). This unusually high value suggests a critical need for social investment interventions.

For the **social investment** predictor ($inv\_social\_millones$):

- $\exp(\hat{\beta}_1) \approx 0.9812$ indicates **each additional million-unit investment decreases expected arrests by** 1.88% $(1 - 0.9812 = 0.0188)$

- The **p-value** (6.63×10^{-8}) shows this effect is:
  - Statistically significant ($p < 0.05$)
  - Unlikely due to random chance (strong evidence against null hypothesis)

8 Conclusion

Since the Negative Binomial Regression minimizes AIC and BIC values and is suitable for count data with overdispersion, this model is selected as the most appropriate for the analysis. Furthermore, social investment shows a significant negative effect across all regressions, suggesting that increased investment in social programs could be associated with a decrease in arrests. Attendance, although significant in the Poisson model, is not significant in the negative binomial regression, indicating that its effect may be less reliable.