tutorial

2024-03-18

The academic paper titled "Modelling association football scores" by M.J. Maher, published in Statistica Neerlandica in 1982 discusses the application of the Poisson model to model football scores, contrasting previous studies that preferred the Negative Binomial distribution. The author incorporates parameters representing teams' attacking and defensive strengths into the Poisson model, exploring various model structures to find the most suitable one. It's concluded that while there are some discrepancies, an independent Poisson model reasonably fits football scores. We built a simplified model and fitted it on a dataset containing International football results from 1872 to 2024 retrieved from https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017.

```
# Load libraries
library(tidyverse)
library(lubridate)
library(broom)
library(MASS)

# Load the data
matches <- read.csv("results.csv")
```

Just like the paper, we limited our scope to a limited period of time.

```
# Filter for recent data
recent_matches <- matches %>%
  mutate(date = ymd(date)) %>%
  filter(date > as.Date("2015-01-01"))
```

First we fitted a Poisson regression model on Home scores without any covariates, as the purpose is to compare Poisson and Negative Binomial on this count data. Summary of the estimated model is shown below.

```
# Poisson Regression Model for Home Scores
home_poisson_model <- glm(home_score ~ 1, family = poisson(link = "log"), data = recent_matches)
summary(home_poisson_model)
```

```
##
## Call:
## glm(formula = home_score ~ 1, family = poisson(link = "log"),
##     data = recent_matches)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.477712   0.008593    55.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 13770  on 8399  degrees of freedom
## Residual deviance: 13770  on 8399  degrees of freedom
## AIC: 29444
##
## Number of Fisher Scoring iterations: 5
```

```
library(AER)
overdispersion_test <- dispersiontest(home_poisson_model, trafo = 1, alternative = "greater")
print(overdispersion_test)
```

```
##
##  Overdispersion test
##
## data:  home_poisson_model
## z = 14.627, p-value < 2.2e-16
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##     alpha
## 0.6788181
```

The statistical test of presence of overdispersion of the Poisson model is highly significant, indicating significant overdispersion in the data as detected through the model.

Next we fitted a Negative binomial model to account for overdispersion in the data and compare with the Poisson model. Summary of the estimated model is shown below.‘

```
# Negative binomial model for over-dispersion in count data
home_nb_model <- glm.nb(home_score ~ 1, data = recent_matches)
summary(home_nb_model)
```

```
##
## Call:
## glm.nb(formula = home_score ~ 1, data = recent_matches, init.theta = 2.802001551,
##     link = log)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.47771    0.01079   44.29   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(2.802) family taken to be 1)
##
##     Null deviance: 9245.3  on 8399  degrees of freedom
## Residual deviance: 9245.3  on 8399  degrees of freedom
## AIC: 28405
##
## Number of Fisher Scoring iterations: 1
##
##
```

```
##                Theta:  2.802
##            Std. Err.:  0.126
##
##  2 x log-likelihood:  -28401.087
```

We did not fit a logistic regression model since the data is count type, not binary.

The AIC of the Poisson model is 29444 while that of the Negative Binomial is 28405, making the latter better in terms of model fit to the data.

In the analysis of football scores, the Negative Binomial regression model is preferred over the Poisson regression due to significant evidence of overdispersion present in the data. Overdispersion occurs when the observed variance is greater than the mean, a scenario common in count data like football scores where factors such as team strategy, player performance, and other unmeasured variables can introduce extra variability. The Negative Binomial model, by introducing an extra parameter to account for this overdispersion, provides a more flexible and accurate fit for such data. Additionally, the Akaike Information Criterion (AIC), a measure of model quality that penalizes for complexity, is lower for the Negative Binomial model than for the Poisson model in this case. A lower AIC indicates a better model fit when comparing models on the same dataset, suggesting that the Negative Binomial model not only addresses the overdispersion more effectively but also improves the overall model fit without overfitting. Therefore, considering both the statistical evidence of overdispersion and the lower AIC value, the Negative Binomial regression emerges as a more suitable choice for modeling football scores under these conditions.