Eden fagerstrom

2022-05-10

Abstract

Shooting is the one of the most important skills in basketball as it is essential in order to score points for your team which will ultimately determine the outcome of the game. This report examines the effects of distance and ball type on making a successful shot into a basketball hoop. The 3 ball types tested in this experiment are Basketball, Soccerball and Netball. This experiment helped produce a model to predict making a successful shot based on distance. The Statistical analysis showed Ball type did not effect the success of making a shot although basketball was the preferred ball there was not enough statistical evidence to suggest ball type was a statistically significant contributing factor to the success of making a shot. Making a successful shot was greatly affected by distance as increased distance significantly decreased the chance of making a successful shot.

Introduction

Basketball requires shooting the ball in order to secure points for your team which will ultimately decide the outcome of the game. The results of the effects of distance on ball type will be an important contributing factor since shots can be made from any distance within the court in a basketball game. Ball type is also an interesting variable as we can now investigate why the ball type basketball is exclusively used in the sport over other varying ball types and if its a significant contributing factor in making a successful shot. This experiment will investigate the effects of the ball types Soccerball and Netball compared to Basketball. The 2nd predictor variable is the distance of the shot where shots were taken at three different distances; 1.5m, 3m, 4.5m.

This experiment will provide the ability to predict the effects of distance and differing ball types on making a successful shot. This experiment will generate insight and essential information for aspiring basketball players as to where they may want to practice shooting for games or may lead them to take more in game shots at distances with a higher chance of successfully making the shot. Once the experiment has concluded, some analysis will be conducted using a binomial distribution to examine the effects of the 2 predictor variables previously described on making a successful shot. This distribution will be conducted with a generalized linear model, a binomial family as well as a link function of logit. This model will provide useful insight into the effects of distance and ball type on making a successful shot.

Experimental Design/ Observation (Including design and data collection method)

aim

This experiment aims to investigate the effects of different ball types at different distances has on making a successful shot in a basketball hoop.

method

This experiment will be undertaken at a local indoor leisure centre on indoor basketball courts in order to reduce any existing natural variables ie wind resistance etc. The experiment has 2 predictor variables which are distance and Ball type. For Ball type there are 3 different ball types used; basketball, Soccerball and Netball. These 3 balls are the standard sizes used in games Basketball(size 7), Soccerball(Size 5) and Netball(Size 5) weighing about 600 g, 450 g, and 450g respectively. The 2nd predictor variable is distance which will be three measurements made from the back of the rim using a tape measure to distinguish where to mark each point with some duct tape. Once each shooting point is marked the three balls will be shot by one person from the smallest distance to the largest distance and then repeated. At each distance the balls will be shot in a randomly selected order to ultimately limit any potential effects of friction and heat within the ball so the results will maintain some validity. Each distance will be repeated 4 times with 3 balls shot at each distance each time, so we have a total of 36 observations (3*4*3). Each time a shot was made it will be marked down as a miss (0) or make (1) and the corresponding distance and ball type will be recorded with it. This gives rise to a binomial response which will be analysed while utilising a generalized linear model with a logit link function

Variables

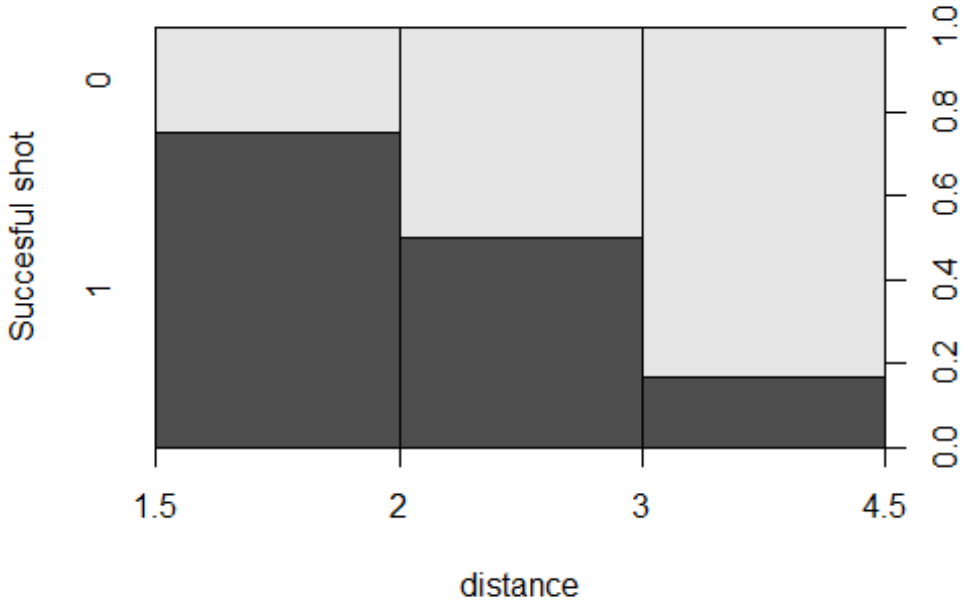
Predictors: Distance: 1.5 metres, 3 metres, 4.5 metres (All measured from the back of the rim of hoop) Ball Type: Basketball (size 7) 600g Soccerball (size 5) 450 grams Netball (size 5) 450 grams

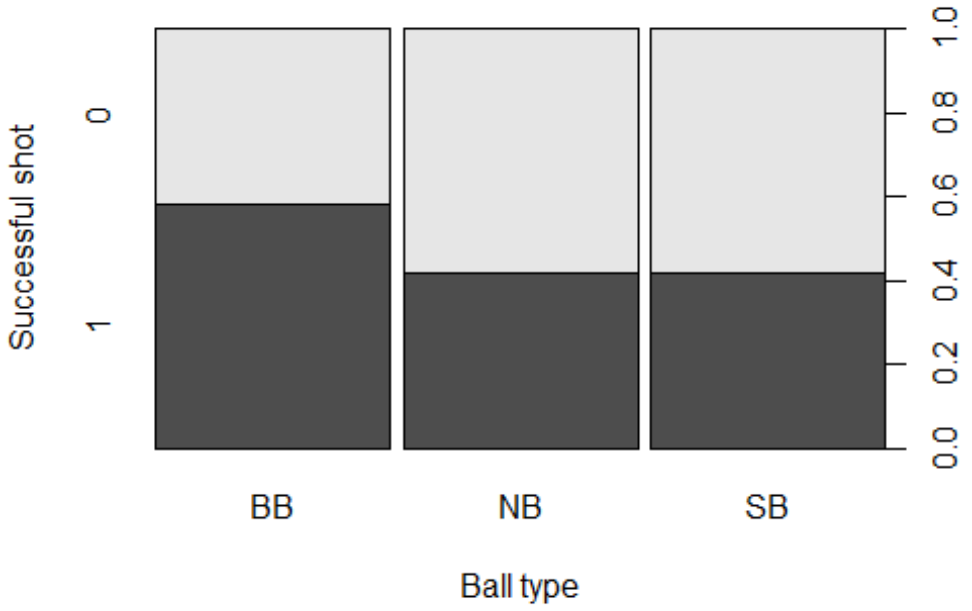
Response: The response variable is whether the shot goes through the hoop or not to see the effects of different ball types and distances on making a successful shot. Hence, 0 for missed shot, 1 for made shot

Exploratory Data Analysis

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
##     distance   shot1  ball1
##  Min.   :1.5   0:19   BB:12
##  1st Qu.:1.5   1:17   NB:12
##  Median :3.0          SB:12
##  Mean   :3.0
##  3rd Qu.:4.5
##  Max.   :4.5
```





From the above plot of distance against making a successful shot we can observe a decrease in successful shots as distance increases which could indicate a strong negative linear relationship.

The plot of ball type against making a successful shot displays an even distribution of successful shots between netball and soccerball with the highest success rate being basketball. There is no obvious relationship visible here.

Statistical Analysis (Model selection process)

The response of making the basket in each case will be a binary response in the form of a count as 0 or 1, where 0 is a missed shot and 1 is successful shot in the basket. This means the response will follow a binomial distribution. To model the binomial distribution, a logistic regression will be used which is a member of the Generalized Linear Model with a binomial response and contains a logistic link. The formula for a logistic regression is:

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Where π_i is the probability of success (Making a shot) β_0 is the intercept β_i is the coefficient for each variable, where i= 1:n $X_{1:n}$ is each of the predictor variables

Dummy Variables

In order to use categorical predictive variables in logistic regression a dummy variable will be created for Basketball (ball type) as this is the preferred ball type this will be the dummy variable against soccer ball and netball to be able to see the difference in making a shot between this ball type and the other ball types.

Logistic Regression Model

The first model was built using Generalised Linear Models with the family argument of binomial distribution and a link of logit. This model will contain the distance of the shot and the ball type without any interactions.

```
##
## Call:
## glm(formula = shot1 ~ ball1 + distance, family = binomial(link = "logit"),
##     data = shots)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5980  -0.8556  -0.5444   0.8087   1.9912
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.2263     1.3027   2.477  0.01326 *
## ball1NB      -0.8845     0.9589  -0.922  0.35633
## ball1SB      -0.8845     0.9589  -0.922  0.35633
## distance     -0.9280     0.3469  -2.675  0.00747 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 49.795  on 35  degrees of freedom
## Residual deviance: 39.889  on 32  degrees of freedom
## AIC: 47.889
##
## Number of Fisher Scoring iterations: 4
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: shot1
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                        35     49.795
## ball1     2   0.8935        33     48.902 0.639699
## distance  1   9.0127        32     39.889 0.002681 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Test for Ball type and Successful Shots:

Null Hypothesis: $H_0: \beta_i = 0$ There is no difference between different ball type and making a shot.

Alternative Hypothesis: $H_A: \beta_i \neq 0$ There is at least one ball type which differs from the ball type basketball on making a shot.

Conclusion:

From the above results we can observe a p-value of 0.639699 which is greater than the 5% significance level and hence we fail to reject the null and we can conclude that There is no statistically significant difference between different ball type and making a shot.

Hypothesis Test for Distance and Successful shots:

Null Hypothesis: $H_0: \beta_1 = 0$ Distance does not have a statistically significant effect on successfully making a shot.

Alternative Hypothesis: $H_A: \beta_1 \neq 0$ Distance does have a statistically significant effect on successfully making a shot.

Conclusion:

From the above results we can observe a p-value of 0.002681 which is less than the 5% significance level. Therefore we reject the null and we can conclude that distance has a statistically significant effect on successfully making a shot.

Model Explanation

Looking at the above results, it can be observed that ball types soccerball and netball don't have statistically difference against ball type basketball in making a successful shot as they both had a p-value greater than the 5% significance level. However since these are included in the model as they are basketball types, a decrease of about 0.9 in the log odds of making a successful shot for both soccerball and netball.

For the results above we can also observe that distance of shot does have a statistically significant effect on making a successful shot as the p-value was less than the 5% significance level. Here we can observe that distance decreased the log odds of making a successful shot by 0.9280. From this model we will further investigate whether we can produce a better model using the stepAIC() function.

Model Formula

$$logit(Y_i) = 3.2263 + -0.8845X_1 - 0.8845X_2 - 0.9280X_3$$

Where: $X_1 = Soccerball, X_2 = Netball$ and $X_3 = Distance$

Stepwise AIC Approach for Model Selection:

```
##
## Call:
## glm(formula = shot1 ~ distance, family = binomial(link = "logit"),
##     data = shots)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7093  -0.7584  -0.6390   0.7266   1.8381
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.5380     1.0556   2.404  0.01620 *
## distance     -0.8940     0.3363  -2.658  0.00786 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 49.795  on 35  degrees of freedom
## Residual deviance: 41.055  on 34  degrees of freedom
## AIC: 45.055
##
## Number of Fisher Scoring iterations: 4

## Analysis of Deviance Table
##
```

```
## Model: binomial, link: logit
##
## Response: shot1
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                        35     49.795
## distance  1   8.7403        34     41.055 0.003113 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Logistic regression Model (Final Model Selected)

The next model will be constructed with the exclusion of ball type. This approach was also used via the stepwise AIC function and the same model was produced as the best fit to the data.

```
##
## Call:
## glm(formula = shot1 ~ distance, family = binomial(link = "logit"),
##     data = shots)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7093  -0.7584  -0.6390   0.7266   1.8381
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.5380     1.0556   2.404  0.01620 *
## distance     -0.8940     0.3363  -2.658  0.00786 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 49.795  on 35  degrees of freedom
## Residual deviance: 41.055  on 34  degrees of freedom
## AIC: 45.055
##
## Number of Fisher Scoring iterations: 4

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: shot1
##
## Terms added sequentially (first to last)
```

```
##
##
##          Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                        35     49.795
## distance  1   8.7403        34     41.055 0.003113 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Test for Distance and Successful shots:

Null Hypothesis: $H_0: \beta_1 = 0$ Distance does not have a statistically significant effect on successfully making a shot.

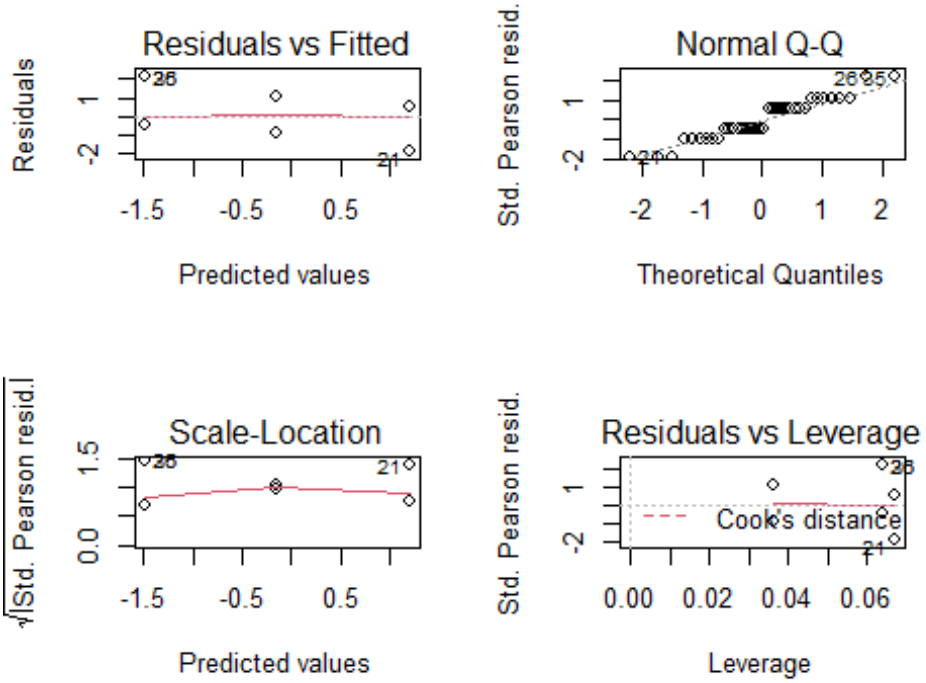
Alternative Hypothesis: $H_A: \beta_1 \neq 0$ Distance does have a statistically significant effect on successfully making a shot.

Conclusion:

From the above results we can observe a p-value of 0.003113 which is far less than the 5% significance, hence we can reject the null hypothesis and conclude that distance has a statistically significant effect on making a successful shot.

Diagnostic Plots

```
par(mfrow=c(2,2))
plot(shots.glm.d)
```



Residuals vs fitted

The residual vs fitted plot shows an even number of points above and below the red line through the centre and the red line is flat, hence showing no heteroscedasticity. A pattern is visible, although the residuals vs fitted plot will always show a pattern based on the nature of the success response.

Normal QQ

From the QQ plot we can see the observations vary from the line and have longer tails than what were expected.

Scale Location

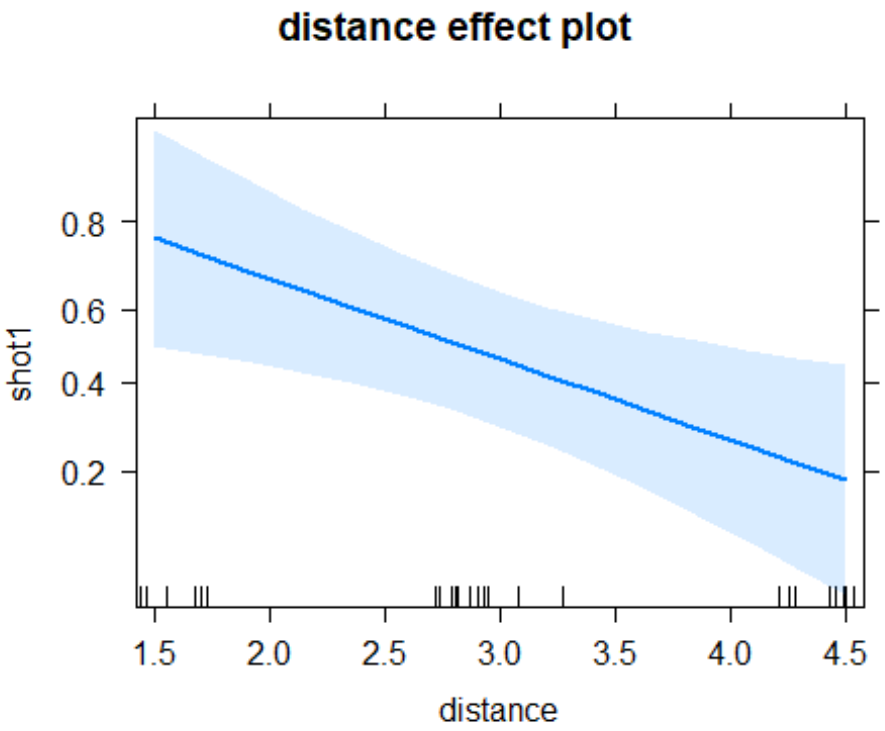
There is a slight pattern in the Scale location plot although as discussed in the residual vs fitted plot patterns tend to occur in logistic regression plots due to the nature of the discrete counts of success in the model.

Residuals vs Leverage

The observed observations are all seen to be within cooks distance.

Effects plots

```
library(effects)

## Warning: package 'effects' was built under R version 4.1.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.1.3

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.

plot(allEffects(shots.glm.d))
```



From the distance effects plot above we can observe that as shooting distance increased the shooting success rate decreased.

Discussion

From the analysis we saw that the 2nd model produced the smallest deviance and was chosen by the Stepwise AIC approach. The 2nd model provided a deviance of about 8.7403, whereas the first model provided a deviance of 9.0127. The diagnostic plots above informs us that the model picked satisfies the conditions and can therefore be chosen to help describe and interpret the dataset produced from the experiment although since our ball type variable wasn't statistically significant this model will be used only to explain the effects of distance on making a successful shot.

Final Model Explanation

The final model formula:

$$logit(Y) = 2.5380 + -0.8940X$$

Where X is the predictor variable distance of shot.

Here we can observe that distance decreased the log odds of making a successful shot by 0.9280. Distance was always likely to have a greater effect on making a successful shot in comparison to ball type as we saw from the results and chosen model. Although we did see from our exploratory analysis that Basketball had more shots made than the other 2 balls which were exactly equal in making successful shots. For further study this experiment could be undertaken by a professional Basketball player who has a higher shot accuracy

than the non-athlete who conducted this experiment and is more familiar with shooting a basketball so we could possibly see some more interesting results and larger variation between differing ball types since someone shooting a ball who is relatively unskilled or under practiced in shooting a ball will not vary much from different ball types.

Conclusion

In conclusion, distance was proven to be the only statstically significant contributing factor in relation to making a successful shot. This is due to many factors such as more strength being required to keep the ball in the air for longer and with more strength applied to the shot accuracy will typically be sacrificed especially by an experienced basketball shooter. The model produced with distance as the only predictor variable provided a lower AIC and deviance then the original model including ball type. Overall the results produced from the final model can help generate insight and essential information for aspiring basketball players as to where they may want to practice shooting for games or may lead them to take more in game shots at distances with a higher chance of successfully making the shot.

References

Intro2r.info. 2022. Binomial data. [online] Available at: https://www.intro2r.info/unit4/binomial-data.html [Accessed 15 May 2022].

Timeless Basketball. 2022. The Importance of Sharp Shooting Skills — Timeless Basketball. [online] Available at: <https://timelessbasketball.com/basketball-shooting-skills/#:~:text=Shooting%20is%20the%20most%20essential,all%20those%20skills%20to%20matter.> [Accessed 15 May 2022].

Eastly, A., 2022. STAT3030 Project. [report, PDF] p.16. Available at: http://file:///C:/Users/edenf/Downloads/Example%20Report.pdf [Accessed 15 May 2022].