Likelihood of Injury for NBA Players

Math 499 Final Project

Erika Iwule

May 18th, 2024

I. Introduction

The NBA is a multi-billion dollar industry, estimated to be worth around $120 billion. NBA players' salaries average around $10 million. Each team has 15 men on their active roster for an 82-game regular season. Because of the limited roster and long game season, it is important for teams to track their player's injuries. Most importantly, it is essential for teams to predict the likelihood of injury, due to the large investment per NBA Player. The goal of this project is to analyze in RStudio using logistic regression, which model and predictors are best to determine the likelihood of injury of an NBA player.

II. The Dataset

I have collected a dataset from Kaggle.com that provides NBA players' statistics to conduct several analyses. The injury dataset provides 6 explanatory variables; Player's Age, Player's Weight, Player's Height, Previous Injuries, Training Intensity, and Recovery Time. The dataset also provides a binary response variable, y = Likelihood of Injury. There are 1000 observations in the dataset.

	Player_Age	Player_Weight	Player_Height	Previous_Injuries	Training_Intensity	Recovery_Time	Likelihood_of_Injury
1	24	66.25193	175.7324	1	0.457928994	5	0
2	37	70.99627	174.5817	0	0.226521626	6	1
3	32	80.09378	186.3296	0	0.613970306	2	1
4	28	87.47327	175.5042	1	0.252858118	4	1
5	25	84.65922	190.1750	0	0.577631754	1	1
6	38	75.82055	206.6318	1	0.359208747	4	0
7	24	70.12605	177.0446	0	0.823552227	2	0
8	36	79.03821	181.5232	1	0.820696161	3	1
9	28	64.08610	183.7948	1	0.477350398	1	1
10	28	66.82999	198.1150	1	0.350819109	1	0
11	38	90.09771	179.1735	0	0.362559797	3	0
12	21	79.02034	171.7098	0	0.805714526	4	0

III. Model Fitting and Selection

To start my analysis, I fitted all the predictors to a logistic regression model. I used the GLM method in RStudio to produce a summary of the model. The model is

$$\widehat{\pi} = -1.171427 - 0.001084x1 - 0.001053x2 + 0.005254x3 + 0.160139x4 + 0.625049x5 + 0.015224\,x6$$

Out of the six predictors, only the predictor, Training Intensity, seems to be significant with a p-value of 0.00524.

```
Call:
glm(formula = Likelihood_of_Injury ~ Player_Age + Player_Weight +
    Player_Height + Previous_Injuries + Training_Intensity +
    Recovery_Time, family = binomial, data = injury)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.171427   1.273131  -0.920  0.35751
Player_Age        -0.001084   0.009772  -0.111  0.91165
Player_Weight     -0.001053   0.006462  -0.163  0.87061
Player_Height      0.005254   0.006453   0.814  0.41550
Previous_Injuries  0.160139   0.127466   1.256  0.20900
Training_Intensity 0.625049   0.223865   2.792  0.00524 **
Recovery_Time     -0.015224   0.037536  -0.406  0.68505
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1386.3  on 999  degrees of freedom
Residual deviance: 1375.9  on 993  degrees of freedom
AIC: 1389.9
```

Next, I wanted to fit the best model by selecting the best subset of variables. I used backward stepwise selection method on the injury dataset. Backward stepwise selection is a stepwise regression approach that begins with a fully saturated model and at each step gradually eliminates variables from the regression model to find a reduced model that best summarizes the

data. Backward stepwise selection uses akaike information criterion (AIC) to determine which is the best. The model with the lowest AIC is deemed to be the best model.

```
Step:  AIC=1387.89
Likelihood_of_Injury ~ Player_Weight + Player_Height + Previous_Injuries +
    Training_Intensity + Recovery_Time

                     Df Deviance    AIC
- Player_Weight       1   1375.9 1385.9
- Recovery_Time       1   1376.0 1386.0
- Player_Height       1   1376.5 1386.5
- Previous_Injuries   1   1377.5 1387.5
<none>                    1375.9 1387.9
- Training_Intensity  1   1383.7 1393.7
```


```
Step:  AIC=1387.89
Likelihood_of_Injury ~ Player_Weight + Player_Height + Previous_Injuries +
    Training_Intensity + Recovery_Time

                     Df Deviance    AIC
- Player_Weight       1   1375.9 1385.9
- Recovery_Time       1   1376.0 1386.0
- Player_Height       1   1376.5 1386.5
- Previous_Injuries   1   1377.5 1387.5
<none>                    1375.9 1387.9
- Training_Intensity  1   1383.7 1393.7
```


```
Step:  AIC=1385.92
Likelihood_of_Injury ~ Player_Height + Previous_Injuries + Training_Intensity +
    Recovery_Time

                     Df Deviance    AIC
- Recovery_Time       1   1376.1 1384.1
- Player_Height       1   1376.6 1384.6
- Previous_Injuries   1   1377.5 1385.5
<none>                    1375.9 1385.9
- Training_Intensity  1   1383.7 1391.7

Step:  AIC=1384.07
Likelihood_of_Injury ~ Player_Height + Previous_Injuries + Training_Intensity

                     Df Deviance    AIC
- Player_Height       1   1376.7 1382.7
- Previous_Injuries   1   1377.7 1383.7
<none>                    1376.1 1384.1
- Training_Intensity  1   1384.0 1390.0
```

```
Step:  AIC=1382.7
Likelihood_of_Injury ~ Previous_Injuries + Training_Intensity

                     Df Deviance    AIC
- Previous_Injuries   1   1378.3 1382.3
<none>                    1376.7 1382.7
- Training_Intensity  1   1384.8 1388.8
```

```
Step:  AIC=1382.31
Likelihood_of_Injury ~ Training_Intensity

                     Df Deviance    AIC
<none>                    1378.3 1382.3
- Training_Intensity  1   1386.3 1388.3

Call:  glm(formula = Likelihood_of_Injury ~ Training_Intensity, family = binomial,
    data = injury)

Coefficients:
       (Intercept)  Training_Intensity
           -0.3076              0.6271

Degrees of Freedom: 999 Total (i.e. Null);  998 Residual
Null Deviance:      1386
Residual Deviance: 1378         AIC: 1382
```

The best model with the lowest AIC of 1382.31 is the model with Training Intensity as the only explanatory variable. The new model is $\widehat{\pi} = \ -0.3076 + 0.6271x$.

```
Call:
glm(formula = Likelihood_of_Injury ~ Training_Intensity, family = binomial,
    data = injury)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.3076     0.1264  -2.434  0.01492 *
Training_Intensity  0.6271     0.2227   2.815  0.00487 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1386.3  on 999  degrees of freedom
Residual deviance: 1378.3  on 998  degrees of freedom
AIC: 1382.3
```

IV. Statistical Inferences

In order to analyze the new model for the likelihood of injury, I must conduct several statistical inferences to describe the data. I began with a Wald hypothesis test. My hypotheses were; H0: Beta = 0 vs. H1 = Beta did not equal 0. Alpha is equal to 0.05. The test concludes that the z-value is 2.815 and the p-value is 0.00487. Since the p-value is less than 0.05, we reject H0. Thus, there is significant evidence that training intensity is not independent from the likelihood of injury for NBA Players. Next, I conducted a 95% Wald confidence interval for beta. The 95% Wald CI is (0.1905512, 1.0636297). We are 95% confident that beta is between (0.1905512, 1.0636297).

Secondly, I conducted the Likelihood Ratio Test and Confidence interval. My hypotheses were; H0: Beta = 0 vs. H1 = Beta did not equal 0. Alpha is equal to 0.05. The test concludes that the LRT statistic is 7.9851 with 1 degree of freedom, and the p-value is 0.00471. Since the p-value is less than 0.05, we reject H0. Thus, there is significant evidence that training intensity is not independent from the likelihood of injury for NBA Players. The 95% LRT Confidence interval is (0.1917668, 1.06537200). We are 95% confident that beta is between (0.1917668, 1.06537200).

Next, I analyzed the marginal effect which is the probability rate of change for describing the effect of an explanatory variable depends on the value of $\widehat{\pi}$. As well as, the Wald test for the training intensity effect and a 95% Wald confidence interval for the odds ratio corresponding to a 1-unit increase training effect. The average marginal effect is 0.022584. My hypotheses were; H0: Beta = 0 vs. H1 = Beta did not equal 0. Alpha is equal to 0.05. The test concludes that the

z-value is 2.815 and z^2 = 7.929028 with one degree of freedom. The p-value is 0.00487. Since the p-value is less than 0.05, we reject H0. There is significant evidence that beta does not equal 0. The 95% Wald CI is (1.209973, 2.89673). We are 95% confident that the odds ratio corresponding to a 1-unit increase in training intensity has at least a 20.9% increase and at most a 289% increase in the odds of the likelihood of injury.

Lastly, I performed the Likelihood Ratio test for the training intensity effect and a 95% LRT confidence interval for the odds ratio corresponding to a 1-unit increase training effect. My hypotheses were; H0: Beta = 0 vs. H1 = Beta did not equal 0. Alpha is equal to 0.05. The test concludes that the LRT = 7.9851 with one degree of freedom. The p-value is 0.004716. Since the p-value is less than 0.05, we reject H0. There is significant evidence that beta does not equal 0. The 95% Likelihood-Ratio CI is (1.2113880, 2.9019183). We are 95% confident that the odds ratio corresponding to a 1-unit increase in training intensity has at least 21% increase and at most a 290% increase in odds of the likelihood of injury.

V. Making Predictions

To describe the effect of an explanatory variable x, it sets the other variables at their sample means and finds at the smallest and largest x values. The effect is summarized by reporting those values $\widehat{P}\,(Y = 1)$. Using the mean of Training Intensity, $\widehat{P}\,(Y = 1) = 0.5000048$. For Training Intensity's min $\widehat{P}\,(Y = 1) = 0.4237072$, and at the max $\widehat{P}\,(Y = 1) = 0.5788578$. At the 1st quartile $\widehat{P}\,(Y = 1) = 0.4609701$, and the 3rd quartile $\widehat{P}\,(Y = 1) = 0.5375385$.

VI. Confusion Tables and Accuracy

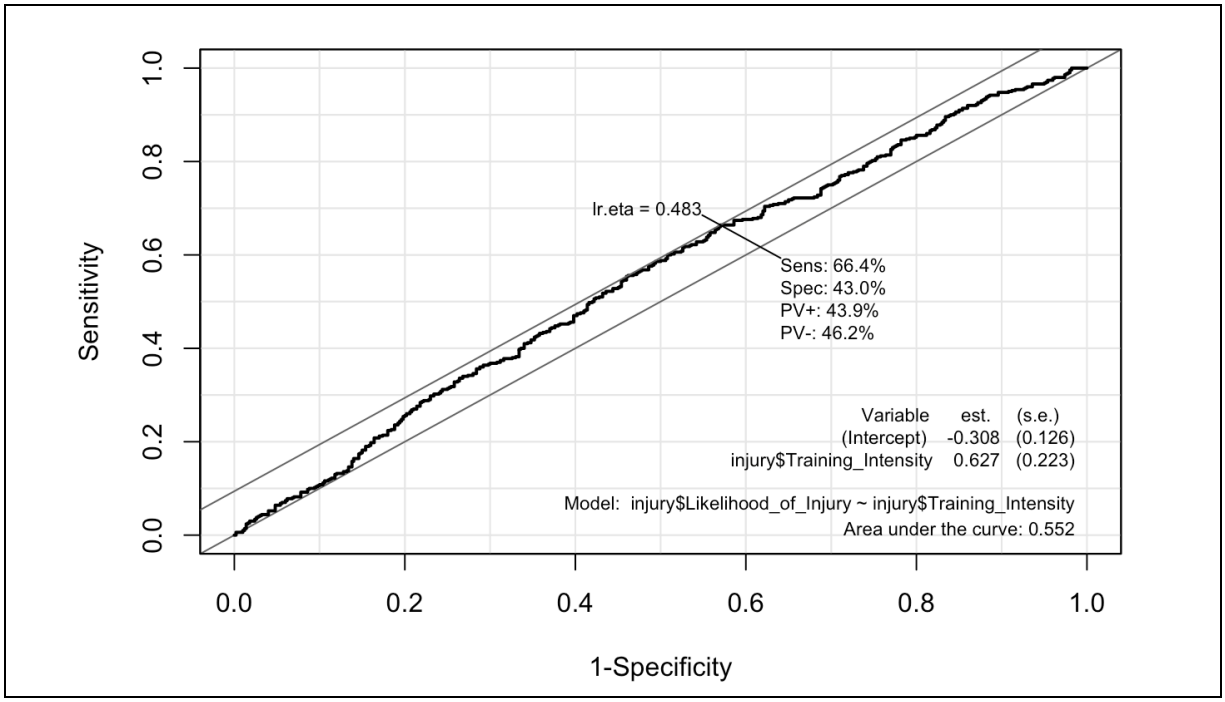
Confusion tables cross-classify the binary outcome y, likelihood of injury, with a prediction of whether y = 0 or 1. From this, we can determine the accuracy of the model based on a cut-off point of pi_0. The first cut-off point was 0.5 which yielded an accuracy of 0.542. The second cut-off point was 0.55 which yielded an accuracy of 0.519. Finally, the third cut-off point was 0.45 which yielded an accuracy of 0.529. The first cut-off point of 0.5 had the highest accuracy of the three cut-off points.

```
                            predicted
injury$Likelihood_of_Injury   0   1
                           0 273 227
                           1 231 269
[1] 0.542
                            predicted2
injury$Likelihood_of_Injury   0   1
                           0 418  82
                           1 399 101
[1] 0.519
                            predicted3
injury$Likelihood_of_Injury   0   1
                           0 104 396
                           1  75 425
[1] 0.529
```

VII. ROC Curve, AUC, Best Cut-Off Point

A receiver operating characteristic (ROC) curve is a plot that shows the sensitivity and specificity of the predictions for all the possible cutoffs for the injury model. The ROC curve will also show the best cut-off point with the highest accuracy. The area under the curve (AUC) summarizes the predictive power of the model. The closer the AUC is to 1, the better the model.

As shown in the plot below, the AUC is .552 which is close to .50. The model is weak and barely better than a random guess. The best cut-off point is 0.483, which has an accuracy of 0.546.



VIII. LOOCV and K-Fold Cross-Validation

To estimate the performance of the logistic regression model of the likelihood of injury, I performed Leave-One-Out Cross Validation and K-fold Cross-Validation. LOOCV happens by splitting the data into a training set and a validation set. This is performed n times. K-fold validation is preferred because it is only performed 5 or 10 times. The LOOCV accuracy is .54. The K-fold Cross-validation accuracy is .538. However, for this model, LOOCV is more accurate than K-fold Cross-Validation.

IX. Probit Link and Identity links to Model Data

Other than logistic regression, we can use probit link and identity link functions to model the injury dataset. Using fully saturated models, both models had the same significant predictor of Training Intensity. The Training Intensity p-value for the identity link model is 0.00482, and 0.00515 for the probit link model. Both models have the same residual deviance of 1375.9 with 993 degrees of freedom and an AIC of 1389.9.

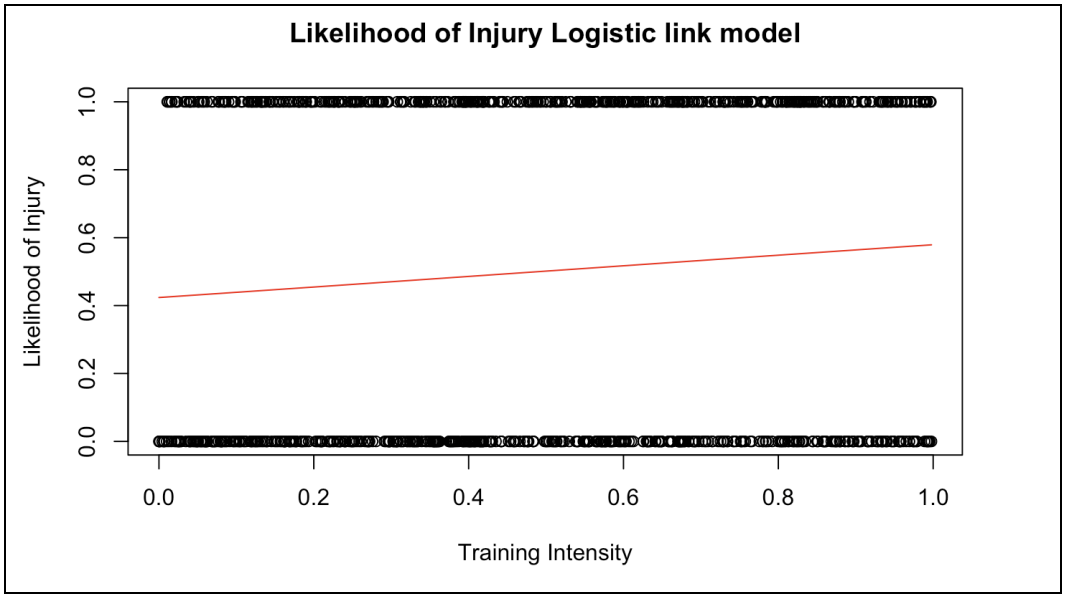
```
Call:
glm(formula = Likelihood_of_Injury ~ Player_Age + Player_Weight +
    Player_Height + Previous_Injuries + Training_Intensity +
    Recovery_Time, family = binomial(link = "identity"), data = injury)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.2191067  0.3147270   0.696  0.48632
Player_Age         -0.0002753  0.0024176  -0.114  0.90934
Player_Weight      -0.0002867  0.0015980  -0.179  0.85760
Player_Height       0.0012588  0.0015965   0.788  0.43043
Previous_Injuries   0.0398073  0.0315265   1.263  0.20671
Training_Intensity  0.1550602  0.0550110   2.819  0.00482 **
Recovery_Time      -0.0037404  0.0092868  -0.403  0.68712
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1386.3  on 999  degrees of freedom
Residual deviance: 1375.9  on 993  degrees of freedom
AIC: 1389.9
```

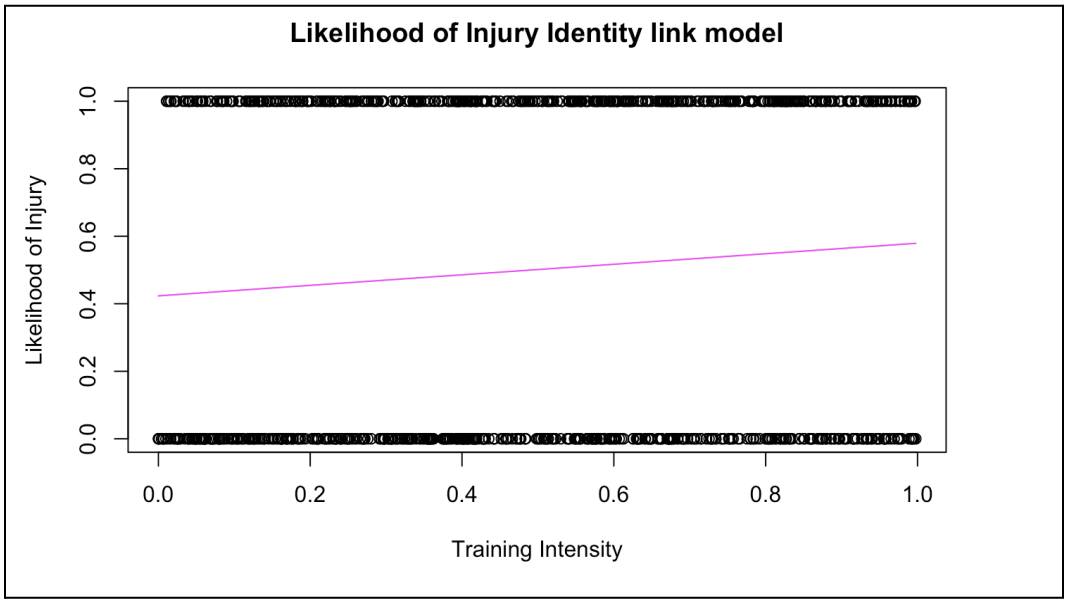
```
Call:
glm(formula = Likelihood_of_Injury ~ Player_Age + Player_Weight +
    Player_Height + Previous_Injuries + Training_Intensity +
    Recovery_Time, family = binomial(link = "probit"), data = injury)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.7276511  0.7959019  -0.914  0.36059
Player_Age         -0.0006818  0.0061101  -0.112  0.91115
Player_Weight      -0.0006725  0.0040401  -0.166  0.86779
Player_Height       0.0032633  0.0040347   0.809  0.41863
Previous_Injuries   0.1002299  0.0796946   1.258  0.20851
Training_Intensity  0.3910398  0.1397728   2.798  0.00515 **
Recovery_Time      -0.0095060  0.0234702  -0.405  0.68546
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1386.3  on 999  degrees of freedom
Residual deviance: 1375.9  on 993  degrees of freedom
AIC: 1389.9
```

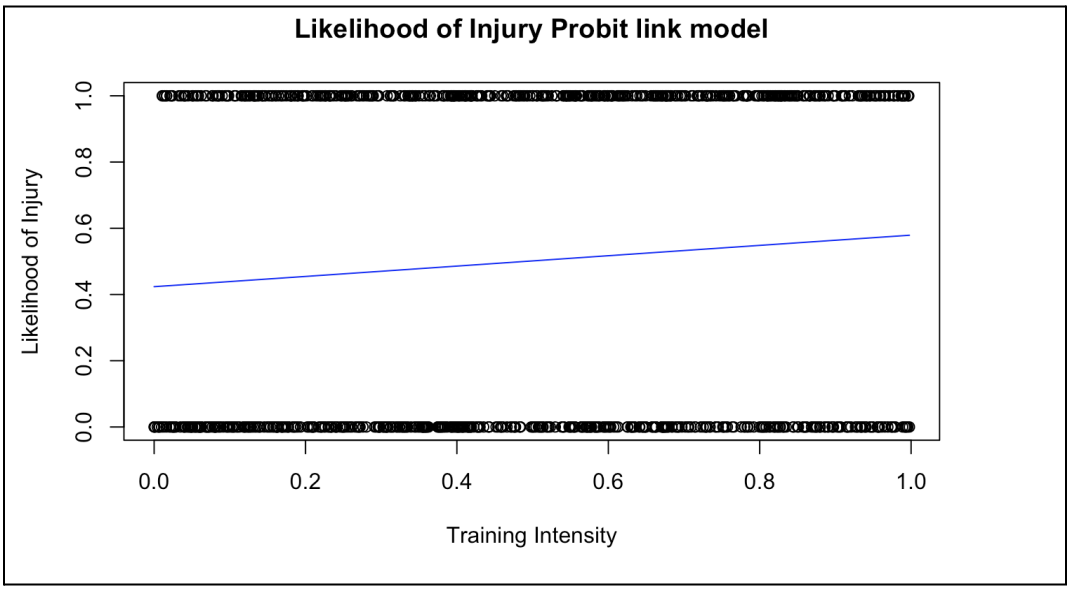
To determine which model is better, I conducted a likelihood ratio test to compare the models using the ANOVA function in RStudio. It yielded a deviance of 0.012225, meaning neither model is better than the other. Next, I plotted the ROC of both models to compare their AUCs. The identity link model AUC is 0.5592. The probit link model AUC is 0.5594, the same as the logistic regression model. Thus, the probit and logistic models are minimally better than the identity link model for the likelihood of injury.

X. Data Visualization

Furthermore, I plotted the prediction curve for each model to depict how similar all three methods are for the injury dataset. For these plots, Training Intensity is the only explanatory variable with y = Likelihood of Injury.


Likelihood of Injury Logistic link model
Likelihood of Injury
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Training Intensity


Likelihood of Injury Identity link model
Likelihood of Injury
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Training Intensity


Likelihood of Injury Probit link model
Likelihood of Injury
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Training Intensity

XI. Conclusion

For the NBA, there is a critical need to prevent the injuries of players in the league, because of money, longevity, winning games, health, and safety. To accomplish better practices to keep NBA players safe, it is important to know which factors have an impact on the likelihood of injury. After conducting multiple analyses, the most significant factor is the training intensity of the player. The best model is $\widehat{\pi} = -0.3076 + 0.6271x$, with training intensity as the only explanatory variable using logistic regression.

XII. Citations

MrSimple. "Injury Prediction Dataset." *Kaggle*, 24 Feb. 2024, www.kaggle.com/datasets/mrsimple07/injury-prediction-dataset?resource=download.

Agresti, Alan. *An Introduction to Categorical Data Analysis*. Wiley, 2019.