Uncovering the Metrics That Matter:

A Linear Modeling Approach to Hamstring Injury Risk in College Football

Jinwoo Choi
University of Arizona
October 12,2025

Abstract

This study aims to investigate the key physical and performance-based metrics that are collected from the advanced sports science technology that tracks athlete's performance. Using the metrics collected by this product, a linear regression model was developed to identify which variables most strongly predict injury occurrence .Predictor variables included measures such as total duration, player load, high-speed distance, acceleration and deceleration counts, total contacts and maximum velocity. The model was trained and validated to evaluate its predictive accuracy and to rank feature importance. Results highlight the metrics most associated with elevated injury risk, providing actionable insights for athletic training staff to optimize workload management and injury prevention strategies. This analysis demonstrates how data-driven modeling can support evidence-based decision-making in collegiate sports performance and health management.

Introduction

At the competitive level of collegiate football, a program's success is ultimately measured by its ability to win games. Maintaining consistent access to top-performing athletes is therefore essential, underscoring the importance of minimizing injuries and keeping key players on the field. The physical demand of training and competition in a sport like Football can be overload for athletes, which most programs have started investing heavily in Sports Science departments in recent years to ensure the players are well-taken care of, optimize their performance, and most importantly to prevent injuries. As player-tracking technologies and performance analytics became more advanced, the sports science industry has been able to benefit from visualizing insights in workload, movement, and metrics. While certain injuries are unavoidable, such as those resulting from high-impact collisions, awkward landings, or underlying health conditions—many performance-related injuries, like hamstring strains, can be mitigated through proper training, conditioning, and workload management.

This project examines which training and workload metrics are most strongly associated with hamstring performance-related injuries in collegiate football athletes, using a linear regression approach. The analysis intentionally excludes external factors such as genetics, lifestyle, and diet, and focuses instead on indicators of training load within the program's structure that may contribute to

physical strain and injury risk. The goal is to move beyond descriptive summaries toward actionable insights that enable coaching and sports science staff to better manage player workloads and support performance and longevity.

Dataset

The dataset was obtained from the University of Arizona's NCAA Division I football program, the Arizona Wildcat Football team, and was collected using the Catapult Vector 8[1] athlete monitoring system. Data were recorded during both practice sessions and competitive games, covering the period from the start of fall camp on July 31, 2025, through the October 4, 2025 game against Oklahoma State University, representing the early portion of the 2025–2026 NCAA football season. The dataset used for this analysis had been previously cleaned and standardized prior to modeling. A summary of the key features and performance metrics included in the dataset is presented below.

- Player: Name of the individual athlete.
- Velocity_Band_5_Total_Effort_Count: Number of instances in which the athlete reached between 80% and 95% of their maximum velocity during a session.
- Velocity_Band_5_Total_Distance: Total distance (in yards) covered by the athlete while moving at 80%–95% of their maximum velocity.
- Velocity_Band_6_Total_Effort_Count: Number of instances in which the athlete exceeded 95% of their maximum velocity during a session.
- Velocity_Band_6_Total_Distance: Total distance (in yards) covered by the athlete while exceeding 95% of their maximum velocity.
- Distance: Total distance (in yards) covered by the athlete during the session.
- Player_Load: A composite measure of total physical exertion, calculated by Catapult's inertial measurement algorithm.
- PL_Min: Player load normalized per minute of activity, representing intensity relative to session duration.
- Hi_Accel: Total count of high-intensity acceleration efforts.
- Hi_Decel: Total count of high-intensity deceleration efforts.
- Date: The date on which the practice or game session occurred.
- Game: Binary indicator denoting session type (1 = game, 0 = practice).
- DB, WR, LB, RB, TE, DL, OL: Binary positional indicators representing whether the athlete's primary position is Defensive Back (DB), Wide Receiver (WR), Linebacker (LB), Running Back (RB), Tight End (TE), Defensive Line (DL), or Offensive Line (OL), respectively. If all of the position indicators have a value of 0, this indicates that the player's position is a QB (Quarterback).

[1] Catapult. (n.d.). *Vector 8*. Retrieved October 12, 2025, from https://www.catapult.com/vector8

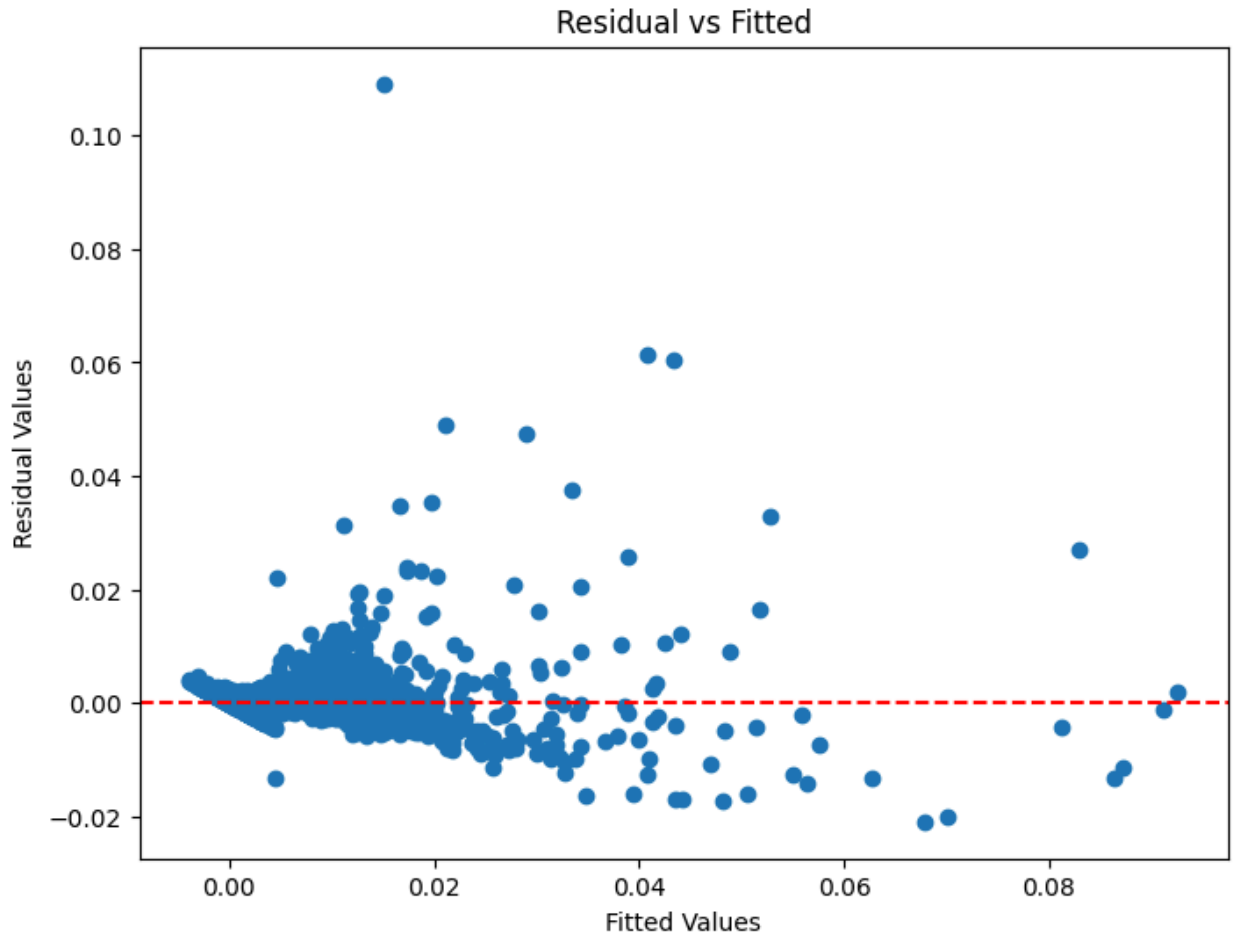
Methodology

To guide the analytical approach, I consulted with the Sports Science department of the Arizona Wildcats Football program to better understand which performance metrics are most relevant for identifying and mitigating injury risk. Through these discussions, staff emphasized that hamstring injuries are among the most common soft-tissue injuries observed in collegiate football and typically occur during high-speed sprinting activities. As a result, the high-speed distance metric was identified as a primary variable of interest for this study. Additionally, using the Catapult's product, I divided the high-speed distance into two levels based on intensity. The first group included movements between 80% and 95% of a player's maximum speed, and the second group included any movement at higher than 95% of maximum speed. This helped capture how much time players spent in different sprinting zones, which may relate to the chance of injury.

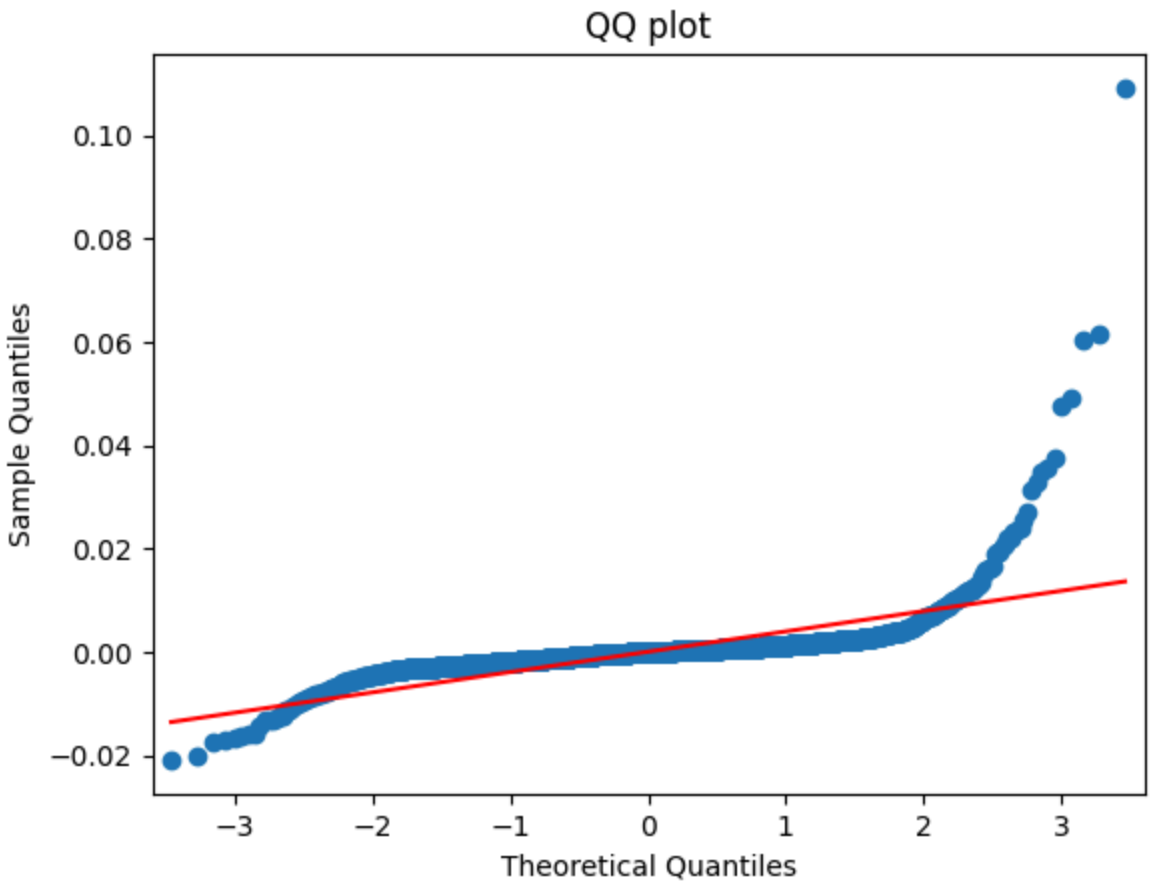
To develop the linear regression model for this project, the first step was to define the appropriate response variable. The goal was to measure the intensity of each practice session by examining the proportion of total distance covered at high speeds. To capture this, a high-speed ratio was created by combining the total distances from Velocity Band 5 (80–95% of max velocity) and Velocity Band 6 (above 95% of max velocity), then dividing that combined value by the total distance of the session, but subtract the total distance by 500 yd of constant value to exclude the non-intensive distance such as walking during practices, jogging to different stations during practices, or walking in the sidelines, etc. This ratio represents the overall sprint workload for each player during a given session.

Visualization:

Residual vs Fitted & QQ Plot:



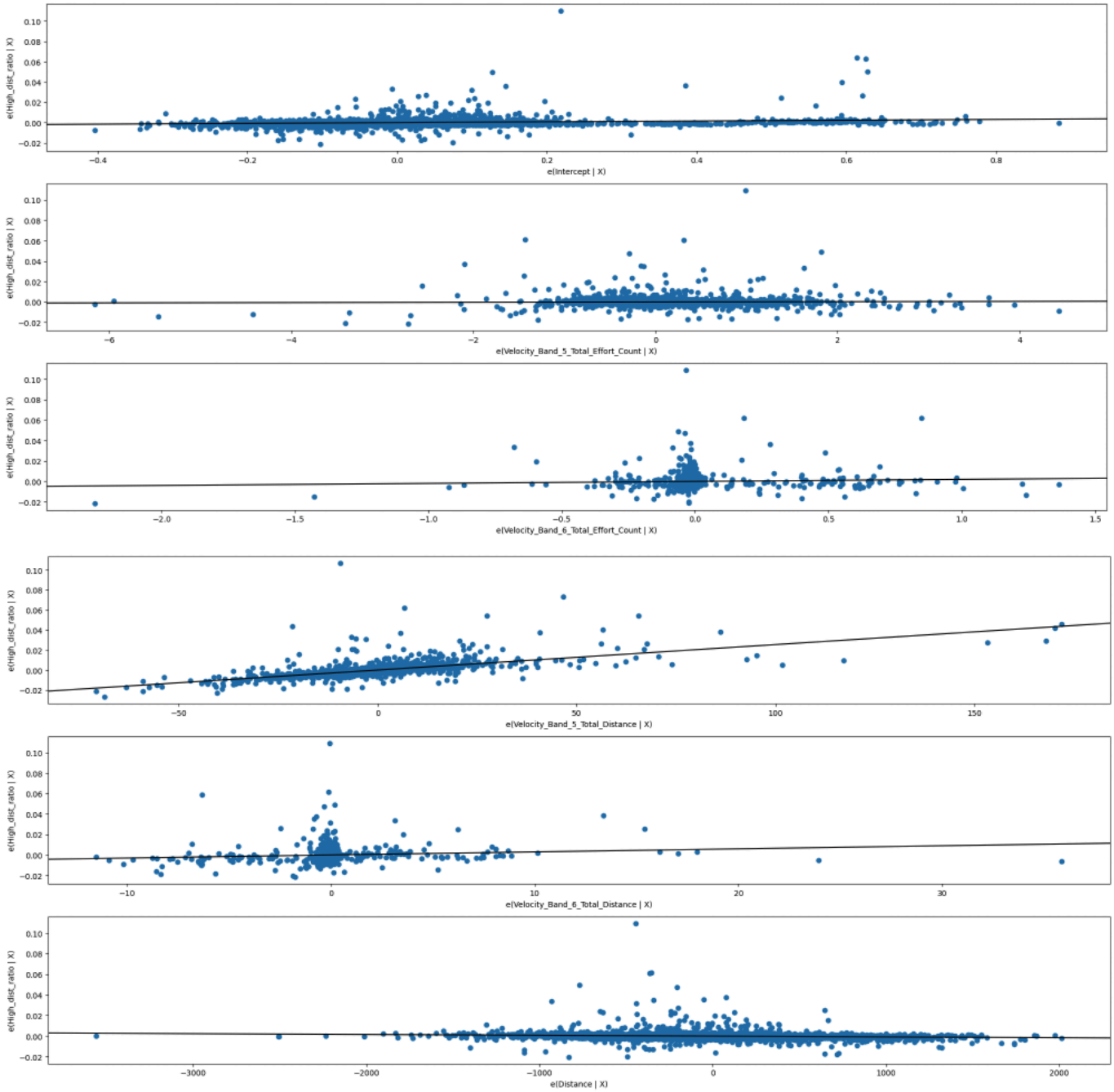
The residuals versus fitted values plot shows that most residuals are centered around zero, suggesting that the model captures the general trend of the data. However, a visible increase in residual spread at higher fitted values indicates mild heteroscedasticity. This means the model's prediction error tends to increase for sessions with higher high-speed ratios. Such behavior is expected in performance datasets where variability naturally increases with workload intensity. Overall, the model meets basic linearity assumptions, but there could be possible transformation techniques to improve the non-constant variance issue.

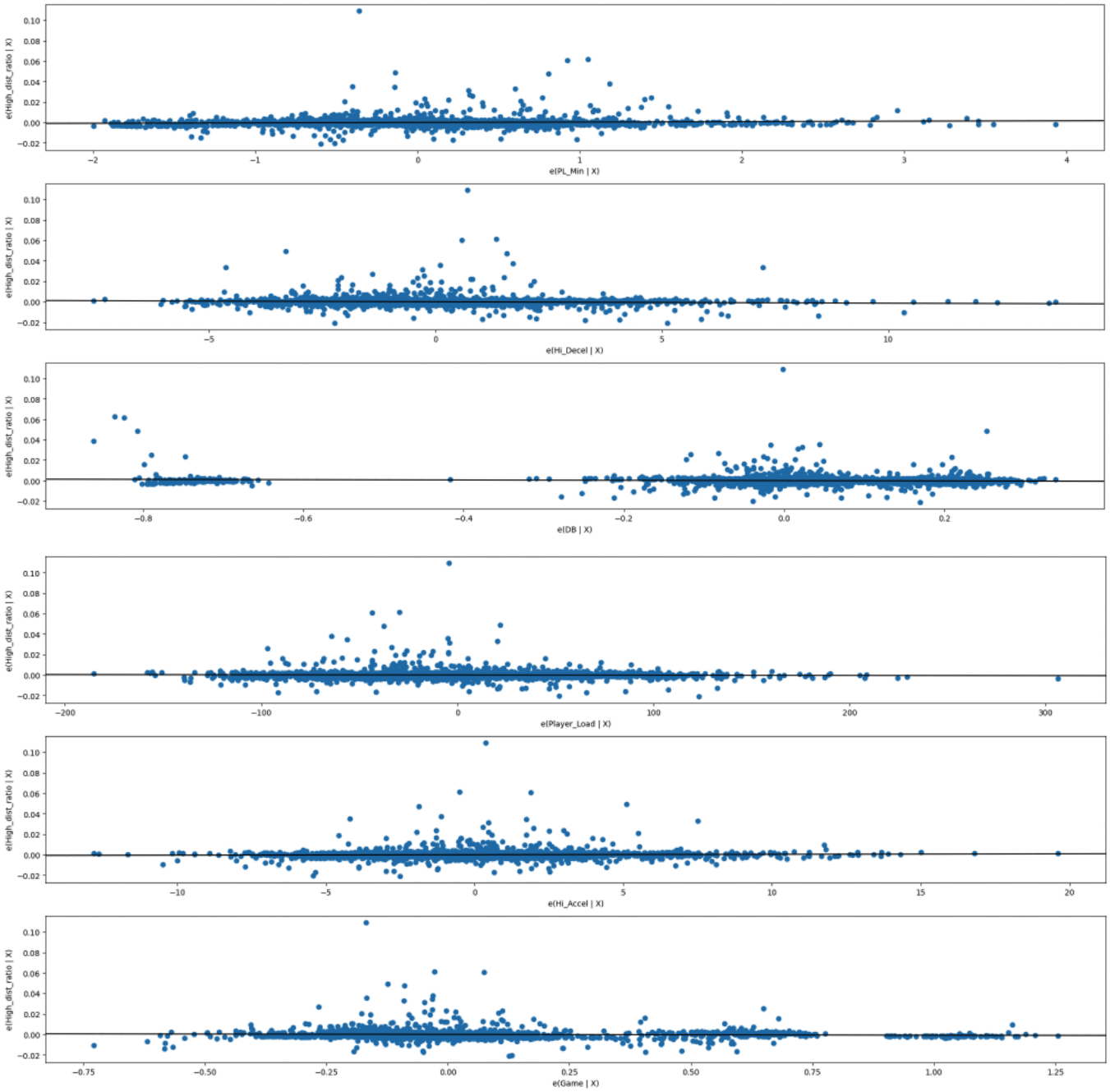


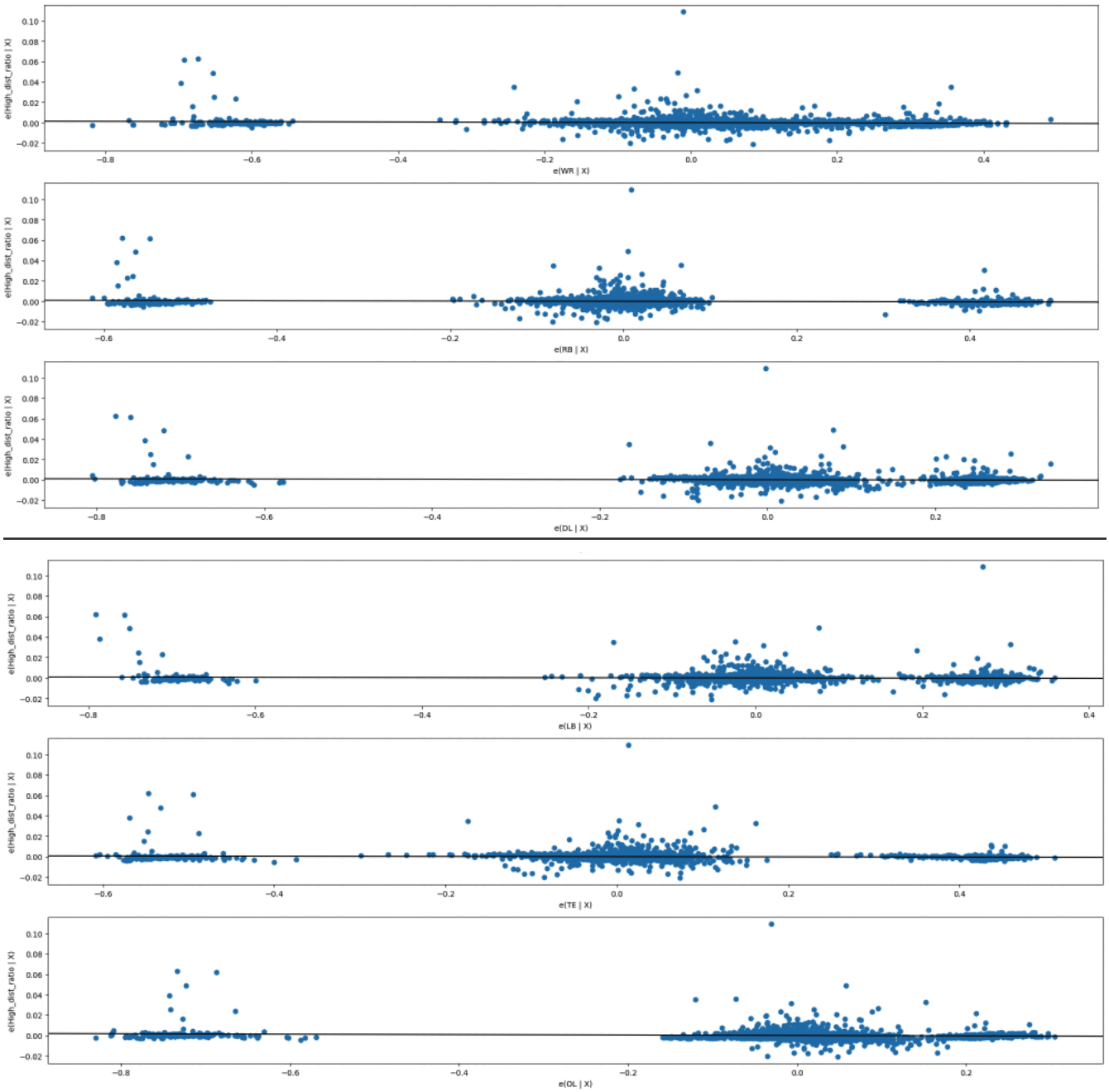
The Q–Q plot of the residuals indicates that most data points align closely with the reference line, suggesting that residuals are approximately normally distributed in the central portion of the data. However, deviations at both tails, particularly in the upper tail, indicate the presence of a few extreme residuals, likely corresponding to sessions with unusually high high-speed ratios. This suggests mild right-skewness in the residual distribution. While this represents a minor concern from normality, it is not uncommon in sports performance data and does not substantially affect the reliability of the regression model given the sample size.

Additionally, calculating the normality through the Shapiro-Wilk Test Statistic, the t-test had a value of 0.486 with a p-value of 0. This calculation indicates that p-value is significantly low enough to fail the null hypothesis, which states that the dataset is not normally distributed. We will further investigate whether the linear regression model is the appropriate fit or not.

Partial Regression Plot:




e(High_dist_ratio | X)
e(PL_Min | X)
e(Hi_Decel | X)
e(DB | X)
e(Player_Load | X)
e(Hi_Accel | X)
e(Game | X)



The partial regression plots collectively show that variables related to high-speed distance have generally positive relationships with the high-speed ratio. Among these, Velocity Band 5 Effort Count displays the steepest slope, indicating a stronger influence on high-speed running intensity, while other variables show flatter slopes, suggesting weaker effects. Although some predictors exhibit only modest trends due to the small range of response values, the overall positive direction across multiple variables suggests that greater workload metrics are associated with higher proportions of high-speed running. The position indicator plots show the relationship between the

high-speed distance ratio for each position. These findings are consistent with on-field movement patterns and confirm that both workload and player position are key factors influencing sprint intensity among collegiate football athletes.

Although there seems to be some outliers and influential points on all those plots that could mislead the data, it is important to take those players into consideration as well to get a greater understanding of this hamstring workload. For this model, we will not be removing or adjusting any of the values to understand the relationship between the high speed ratio and workload metrics for all players in the dataset. Therefore, no calculations to determine the outliers, and

Statistic Summary:

OLS Regression Results

Dep. Variable:	High_dist_ratio	R-squared:	0.798
Model:	OLS	Adj. R-squared:	0.797
Method:	Least Squares	F-statistic:	879.8
Date:	Mon, 13 Oct 2025	Prob (F-statistic):	0.00
Time:	05:06:51	Log-Likelihood:	15672.
No. Observations:	3803	AIC:	-3.131e+04
Df Residuals:	3785	BIC:	-3.120e+04
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0039	0.000	10.482	0.000	0.003	0.005
Velocity_Band_5_Total_Distance	0.0003	5.36e-06	47.444	0.000	0.000	0.000
Velocity_Band_5_Total_Effort_Count	0.0001	0.000	1.241	0.215	-8.22e-05	0.000
Velocity_Band_6_Total_Distance	0.0003	4.7e-05	6.378	0.000	0.000	0.000
Velocity_Band_6_Total_Effort_Count	0.0020	0.001	3.301	0.001	0.001	0.003
Distance	-7.857e-07	1.23e-07	-6.372	0.000	-1.03e-06	-5.44e-07
Player_Load	-1.912e-06	1.28e-06	-1.489	0.137	-4.43e-06	6.05e-07
PL_Min	0.0004	9.43e-05	4.290	0.000	0.000	0.001
Hi_Accel	4.382e-05	2.18e-05	2.015	0.044	1.17e-06	8.65e-05
Hi_Decel	-0.0001	3.3e-05	-4.169	0.000	-0.000	-7.29e-05
Game	-0.0008	0.000	-3.020	0.003	-0.001	-0.000
DB	-0.0015	0.000	-4.599	0.000	-0.002	-0.001
WR	-0.0016	0.000	-4.574	0.000	-0.002	-0.001
LB	-0.0010	0.000	-2.907	0.004	-0.002	-0.000
RB	-0.0014	0.000	-3.742	0.000	-0.002	-0.001
TE	-0.0011	0.000	-2.735	0.006	-0.002	-0.000
DL	-0.0011	0.000	-3.396	0.001	-0.002	-0.000
OL	-0.0021	0.000	-6.347	0.000	-0.003	-0.001

Omnibus:	6088.857	Durbin-Watson:	1.940
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7090490.234
Skew:	10.024	Prob(JB):	0.00
Kurtosis:	213.582	Cond. No.	5.25e+04

Based on the regression summary, the model achieved an R-squared value of 0.798, indicating a strong relationship between the predictor variables and the high-speed ratio. Metrics related to high-speed distance and effort count showed clear positive associations with the response variable, which aligns with expectations since greater high-speed running directly increases the

proportion of sprint distance. Conversely, variables such as total distance, player load, and high acceleration and deceleration counts displayed negative relationships with the high-speed ratio. From a sports science perspective, this can be explained by fatigue and workload effects—players who accumulate greater total distance or load are less likely to sustain high-intensity sprinting, reflecting decreased sprint output over time. Similarly, higher acceleration and deceleration counts suggest frequent stop-and-go movements, limiting opportunities to reach maximum velocity and therefore reducing the proportion of high-speed distance. Ironically, All positional indicators produced relatively small or negative coefficients, suggesting that no single position showed a substantially greater relationship on the high-speed ratio once other workload factors were accounted for. The Game variable also displayed a slight negative coefficient, indicating that game sessions did not significantly differ from practice sessions in terms of high-speed distance proportion. Overall, these coefficient results are consistent with expectations of sports context and provide meaningful insights into how training workload and movement patterns influence sprint activity in collegiate football players.

When evaluating feature significance, most predictor variables exhibited low p-values, indicating statistically significant relationships with the high-speed distance ratio. This suggests that these metrics meaningfully contribute to explaining variations in high-speed running activity. However, the Velocity Band 5 Effort Count showed a comparatively high p-value, implying that its relationship with the high-speed ratio is not statistically significant. This result contrasts with the partial regression plot interpretation, where this variable appeared to have a stronger positive trend, suggesting that its influence may be less consistent once the effects of other predictors are controlled for.

Results

The performance of this linear regression model was evaluated using the Mean Squared Error(MSE) and Mean Absolute Error(MAE) metric, which measures the average squared difference between the predicted and actual value of the high-speed distance ratio, and the average absolute difference between the predicted and actual value of the high-speed distance ratio. This model achieves the MSE value of 1.54 EE-5, and the MAE value of 1.68EE-03, indicating that the errors are extremely small and that the model's fitted values align closely with the actual observed

data. Through this metric, the low error value suggests that the model is highly effective and accurate at capturing the relationship between the workload and position features with the high-speed ratio. Overall, the MSE and MAE supports that dataset has a strong linear relationship that aligns with the linear regression model.

Analysis of individual predictors revealed that Velocity_Band_5_Total_Distance and Velocity_Band_6_Total_Distance are the most influential predictors as expected, but ironically Velocity Band 5 Effort Count did not have a strong relationship with the response variable according to the p-value. The overall work metrics such as Total distance and player load had slightly negative correlation, indicating that too much workload will likely fatigue players to not reach the maximum speed effectively, and also increasing the denominator of the ratio in addition to the dotal distance at the same time. Explosive movement such as high acceleration and high deceleration also showed a slightly negative coefficient, meaning that many stop-and-go movements will restrict the players from reaching their near maximum velocity.

Positional indicators were not a significant factor for this response variable, implying that once workload metrics are considered, certain positions do not have any substantial impact and explain the additional variation in sprint intensity. Similarly, game variables also show a slight negative relationship, suggesting not much difference between game and practice sessions in terms of proportion of high-speed running. Overall the model showed a strong fit, and low prediction error, indicating that selected features captured the fundamentals of high-speed workload in collegiate football athletes effectively.

Conclusion

Although the model demonstrated a strong fit and low prediction error, a couple of linear regression assumptions were not fully satisfied, suggesting that a more advanced modeling approach could better capture the complexity of the dataset. Nevertheless, the model performed exceptionally well for a linear framework and provided interpretable results that can offer valuable insights for coaching and sports science staff. The findings highlight key workload variables that influence sprint intensity, which can help inform training load management and performance monitoring strategies. In future work, I plan to extend this analysis by incorporating historical injury data and developing more sophisticated predictive models to estimate injury risk more directly. This would enable a

deeper investigation into the relationship between workload metrics and specific injuries, such as hamstring strains, at the individual athlete level.

Reference

- University of Arizona, Arizona Wildcat Football Team, Sports Science Department, Strength & Conditioning Program. (2025). *Player workload dataset collected using Catapult Vector 8* [Unpublished raw data]. Provided by Catapult Specialist & Assistant Director of Football Strength & Conditioning James Perez.

Code

- https://colab.research.google.com/drive/1VC1wBTaj2Sw7BZaomteZ3g-sEAS6sdAp?usp=sharing
- https://github.com/jinwoo1015/hamstring_injury_metrics