HW 4 - Simpson's Paradox in NFL Combine Data

Emma Schmidt

STA 610 - Fall 2022

1 Introduction

An NFL coach desires a tall football team because he believes it offers offensive players better field vision and defensive players an advantage in disrupting the opposing team's offense. A scout warns him against this strategy, suggesting that an increase in height correlates to a decrease in power and explosiveness. The scout backs this claim with NFL combine data spanning the years 2009-2019. He argues that over this time frame height has been negatively correlated with broad jump distance (a measure that is often used to assess a player's power and explosiveness). I argue that the scout has made an error in his analysis, and that by adopting a hierarchical modeling approach the data will contradict the conclusions of the scout. This phenomenon is called Simpson's Paradox, and is the focus of this paper.

2 Linear Regression Model

The model adopted by the scout is a simple linear regression model that explores the relationship between height (m) and broad jump distance (cm). The model is expressed below:

$$y_i = \mu + \beta x_i + \epsilon_i$$

$$y_i : Broad\ Jump\ (cm), \quad x_i : Height\ (m), \quad \epsilon_i \sim N(0, \sigma^2)$$

Utilizing this model and the NFL combine data, the scout arrives at the conclusion that for every 1m increase in athlete height, the coach can expect to see a 158.20cm decrease in broad jump distance. This relationship can be seen in the model output below:

$$y_i = 589.48 - 158.20 x_i$$

The coach is now worried about his strategy, fearing that the advantages of a taller team may not outweigh the consequences of lost power in his athletes. To further convince the coach of his findings the scout presents him with Figure 1, which suggests a strong negative correlation between the two indicators.

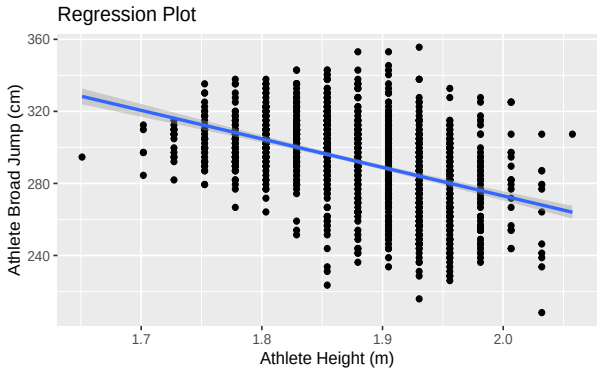

Figure 1: Height vs. Broad Jump Regression Analysis

3 Multilevel Model

The scout has failed to consider how a multilevel modeling structure might impact his findings. Before finalizing any decisions, I offer the coach a different approach. Football players come in all shapes and sizes, and thus each offer the team a different skill set. These skill sets suit different positions, and for this reason I group by position type. To achieve this grouping structure I employ a model with a random intercept and slope for athlete height, grouped by position type. The five position types are back/receivers, defensive backs, defensive linemen, linebackers, and offensive linemen. The modeling structure can be seen below:

$$y_{ij} = \mu_j(Position\ Type_j) + \beta_j x_{ij}(Height_{ij}) + \epsilon_{ij}$$

$$y_{ij} : Broad\ Jump\ (cm), \quad x_{ij} : Height\ (m), \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

$$\mu_j : \text{random intercept for each position type}, \quad \mu_j = \mu + \alpha_j, \quad \mu_j \sim N(\mu, \tau^2)$$

$$\beta_j : \text{random slope for each position type}, \quad \beta_j \sim N(0, \tau_\beta^2)$$

The revised multilevel model yields much different results. In all but one of the positions height is positively correlated with broad jump distance. In particular, defensive linemen exhibit this relationship the strongest with a 1m increase in height leading to a 189.43cm expected increase in broad jump distance. Defensive backs, linebackers, and offensive lineman are the other position types that show a positive relationship between height and broad jump distance. The last position type, backs/receivers, shows little relationship between height and broad jump distance. Overall, these findings make for great news because they align with the coach's initial intuition that height offers an advantage to a football team. Figure 2 solidifies these findings for the coach with a clear visual representation of the multilevel model output.

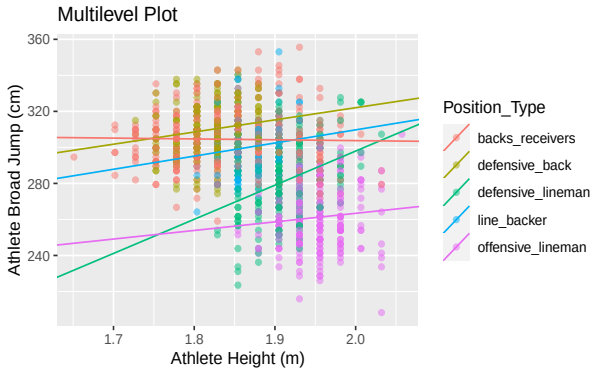

Figure 2: Height vs. Broad Jump Multilevel Analysis by Position Type

4 Simpson's Paradox

The phenomenon exemplified in this exploration is known as Simpson's Paradox. Simpson's Paradox is formally defined as the statistical phenomenon where an association between two variables in a population emerges, disappears or reverses when the population is divided into sub populations. While the scout conducted a sound linear regression analysis of the data, his conclusions were incorrect because he did not account for differences between players. Had he considered the multilevel model that I adopted in section three of this paper, he would have seen that his original determination regarding the relationship between height and broad jump was reversed. The reversal in association between the variables from model to model can be more explicitly seen in Figure 3, which shows the two regression models side by side. In conclusion, the data supports the coach's initial strategy and he should not worry about athlete height compromising their power and explosiveness.

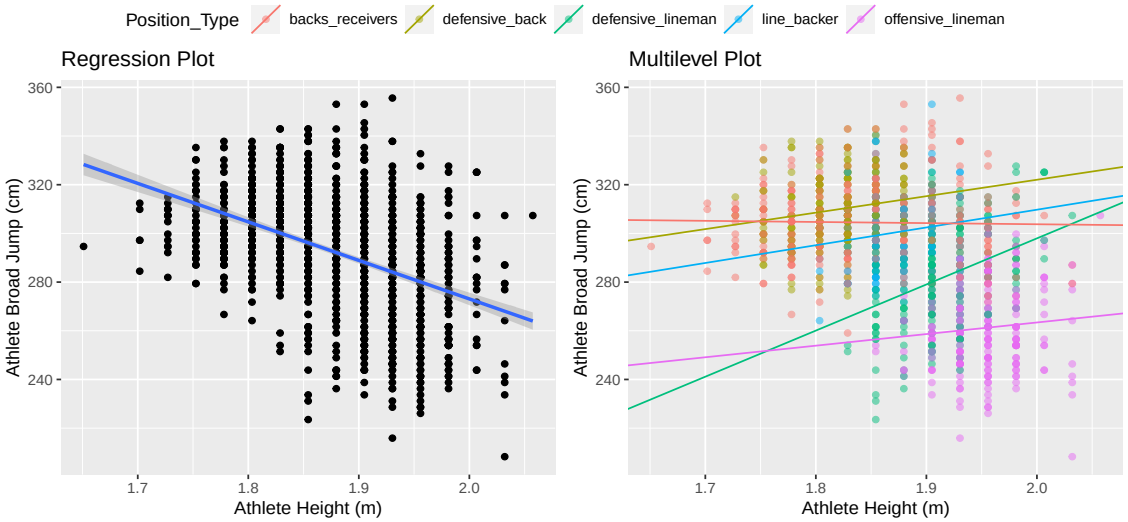

Figure 3: Simpson's Paradox