# 2019 NFL COMBINE ANALYSIS

Inferential Methods Project

Sean McConnell

December 12, 2019

# Introduction

Each year, a limited number of college football players from across the nation are invited to attend the NFL combine to impress scouts for their chance to play professionally. These players are examined head to toe, measured, and tested for speed and strength. I chose the 2019 NFL combine metrics & results (Fernandez, 2019) because it showcased the that year's best college players who have yet to become anything in the NFL. The combine participants' future in the NFL was undecided at this point; they could become hall of fame players or one season stars.

This analysis did not look at talent or skill, but rather inherited physical attributes shared among players. The overall research goal was to assess if certain inherited physical attributes impacted player's positions and performance. I used ANVOA to compare mean hand size by position zone and determined that mean hand size was different by position zone. I then used simple linear regression to assess the impact that height had on 40-yard dash times and determined that height is a significant predictor of 40-yard dash times; tall players were slower than short players.

I was astounded by the results because I believed that effort and determination was the only factor to success. There's no doubt that each of these players put in a significant amount of their lives training and playing for this opportunity, but at the end of the day, their inherited physical attributes played an enormous role as well.

## ANOVA: Hand Size vs Position Zone

### Purpose

The first step in the overall research goal involved player hand size and position zone. Since players come in all shapes and sizes for every position, I reasoned that every position zone will have different mean hand sizes. If the average hand size differed across position zones, it would not insinuate that players with specific hand sizes fit to one zone, but rather that each zone had an "ideal" hand size for which players were better suited. The question was: did each position zone have a different mean hand size?

### Variables

During the combine, the hand size of each player was measured and ranged from 8.25 inches to 10.88 inches. The measurement is taken from the tip of the thumb to the tip of the pinky finger of the dominant hand. Since hand size was the continuous variable being analyzed, there was no reason to further stratify these measurements.

The combine data also contained each player's college position. Since there were 18 positions, and many of these were similar on offense and defense (Offensive Tackle vs Defensive Tackle, Line Backer vs Running Back, etc..), I wanted to gain a better understanding of each position's "zone" and the corresponding hand size of players in that zone. Positions were classified into the following zones:

- Offense and Defensive Line positions were classified as "Line"
- Linebackers, Running Backs, and Full Backs were classified as "Mid"
- Quarterbacks were classified as "QB"

- Wide Receivers, Cornerbacks, and Safeties were classified as “Skilled”
- Kickers, Punters, and Long Snappers were classified as “Special”

## Method and Assumptions

In order to assess the mean hand size across position zones, I determined that the ANOVA method would the most effective. In order to use ANOVA method, I needed to ensure the two main assumptions were met (see appendix 1 – method 1):

- **Independence of cases –** hand size is an independent variable
- **Distribution of residuals are normal** – the residuals were normally distributed

Since both assumptions were satisfied, the analysis is reliable.

## Level of significance and hypotheses

For this analysis, I used a level of significance of 0.01 with the following hypotheses being tested:

> **$H_0$ =** There is no difference in the mean hand size between position zones of players that participated in the 2019 NFL combine.
>
> **$H_A$ =** There is a difference in the mean hand size between position zones of players that participated in the 2019 NFL combine.

## Results

After analyzing the ANOVA table *(see fig. 3)*, the Pr > F was <0.0001 *(see figure 3)* which is less than the significance level of 0.05, so I rejected the null hypothesis; there was a statistically significant difference in the mean hand size between position zones of players that participated in the 2019 NFL combine.

# Simple Linear Regression: Height vs. 40-Yard Dash Times

## Purpose

The second step in the overall research goal involved player height and their 40-yard dash time. It stood to reason that taller players would have longer strides, but tall players typically weight more in football, which would slow them down. The same reasoning went for smaller players too: shorter players would have shorter strides but would typically weigh less. The question was: did height have a significant impact on 40-yard dash times for players at the 2019 Combine?

## Variables

During the combine, the height of each player was measured while standing flat-footed. The measurement was taken to the tip of the player's head and ranged from 67 inches to 87 inches.

The 40-yard dash time is taken via laser for the most accurate measurement. The player starts in a sprinter's stance, begins running when they are comfortable, and sprints through the 40-yard mark. This test measures the player's explosion off the line, acceleration, and top speed.

## Method and Assumptions

In order to assess if height is a predictor of 40-yard dash time, I decided to use simple linear regression because both factors are continuous independent variables. In order to use simple linear regression, I had to ensure two main assumptions were met *(see appendix 1 – method 2):*

- **Linearity** - The relationship between height and 40-yard dash times is linear. I inspected the scatterplot *(see fig. 4.),* and the data appeared to follow a straight upward line with no curvature. This indicated that the relationship between the two variables was linear.
- **Residuals are normally distributed** – the residuals are normally distributed. I inspected the Q-Q plot and the histogram of the residuals *(see fig. 5)*, and both appeared normal.

Since both assumptions were satisfied, the analysis is reliable.

## Level of significance and hypotheses

For this analysis, I used a level of significance of 0.01 with the following hypotheses being tested:

**$H_0$ =** Height is not a significant predictor for 40-yard dash times of players that participated in the 2019 NFL combine.

**$H_A$ =** Height is a significant predictor for 40-yard dash times of players that participated in the 2019 NFL combine.

## Results

After analyzing the simple linear regression model *(see figure 6)*, the Pr > |t| for height is <0.0001, which is less than the significance level of 0.05, so I rejected null hypothesis; height is a statistically significant predictor of 40-yard dash times for players that participated in the 2019 NFL combine.

## Conclusions

Through the ANOVA analysis of hand size vs position zone, and the simple linear regression of height vs 40-yard dash time, we can conclude these physical attributes have an impact on predicting a player's position zone and speed. These findings support general knowledge of human anatomy and the resulting performance of inherited traits, but I believe this analysis is just the tip of the iceberg. In order to gain more insight, I would increase the population to include multiple years of the combine, as well as current and past NFL players.

The ANOVA analysis suggests that each position zone is better suited for players with specific hand sizes. I believe this can be generalized for each position zone across the NFL and football in general. Some positions are better suited for players with large hands (QB, Wide Receiver, Linemen), while hand size is not as important to be successful in other positions (Cornerback, Linebacker, Kicker). The next logical step to improve this analysis is stratification of position zones based on offense or defense, and possibly even down the position itself.

The regression analysis suggests that height is a predictor of 40-yard dash times, but it also showed that shorter players ran faster times than taller players. One would expect taller players with longer strides to run faster than shorter players with shorter strides. I believe this may be generalized with players in the NFL, but certainly not for the general population. A multivariate analysis with height, weight, shoe-size, stride length, and power clean max weight would be the logical step to narrow in on the full scale of predictors.

There is much more analysis that needs to be done, but these findings begin to support the claim that certain inherited physical attributes impact player’s positions and performance.

## Reference List

Fernandez, D. (2019). *2019 NFL Scouting Combine.* Retrieved from Kaggle: https://www.kaggle.com/dtrade84/2019-nfl-scouting-combine/version/1

*Linear regression with SAS*. (2010). Retrieved from Purdue: https://www.stat.purdue.edu/~tqin/system101/method/method_linear_sas.htm#regression-assumption

Melfi, V. (2004). *Lab 7: Proc GLM and one-way ANOVA.* Retrieved from Michigan State University: https://www.stt.msu.edu/~melfi/teaching/summer04/422/sas/lab7.pdf

mhlinder. (2015). *GLM Residual plots?* Retrieved from SAS communities: https://communities.sas.com/t5/Graphics-Programming/GLM-residual-plots/td-p/150248

Navidi, W. (2017). *Statistics for Engineers and Scientists 4th Edition.*

*Simple Linear Regression*. (2016). Retrieved from http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression4.html

*What Goes on at the Combine*. (2019). Retrieved from NFL.com: http://www.nfl.com/combine/workouts

# Appendix 1 – Data, Method, and Assumptions

## Method 1

The data obtained from the source was extremely clean. There was no need to alter or change any of the values in the original format. Position zone was the only created variable, which I used personal knowledge to populate. In total there were 18 positions represented, which were consolidated into 5 zones:

- Offense and Defensive Line positions were classified as “Line”
- Quarterbacks were classified as “QB”
- Linebackers and Running Backs were classified as “Mid”
- Wide Receivers, Cornerbacks, and Safeties were classified as “Skilled”
- Kickers, Punters, and Long Snappers were classified as “Special”

In order to reliably use ANOVA, I had to assess the following assumptions:

- **Independence of cases –** One player’s hand size has no effect on another player’s hand size. Each player inherits their hand size genetically from their parents, and each measurement is taken individually by an unbiased combine professional.
- **Distribution of residuals are normal** – Using the PROC GLM residuals plot in SAS, the distribution (see *fig. 1*) and Q-Q plot (see *fig. 2*) of the residuals is relatively normal.

Once the assumptions were met, I entered the SAS code for PROC GLM (see appendix 2), defined the position zone as the class, and set the model to compare the hand size with the defined class of position zone. Once SAS processed the code, I assessed the

ANOVA table, and inspected the box and whisker plot to ensure the conclusions appeared correct between the table and chart.

## Method 2

The data obtained from the source was extremely clean. There was no need to change the original format or create new variables, however there were 61 null values for 40-yard dash times. In order to have an unbiased data set, the null values were removed from the dataset before the simple linear regression analysis was performed. Even without these values, there were 213 players the analysis covered.

In order to reliably use simple linear regression, I had to assess the following assumptions:

- **Linearity** - The relationship between height and 40-yard dash times is linear. Looking at the regression plot (see fig. 4), there's a clear trend between height and 40-yard dash times. The plot is not curved; it followed an upward, straight line from the lowest to the highest height.
- **Residuals are normally distributed**- The residuals Q-Q plot and the residuals distribution (see fig. 5) appear normal. The Q-Q plot follows a straight, upward line with very little deviation at the ends. The histogram is approximately normal, with a slight peak in the middle.

Once the null values were removed from the population and the assumptions were confirmed, I entered the SAS code for PROC REG (see appendix 2) and set the model to compare 40-yard dash time to height. Once SAS processed the code, I assessed the fit models once more, and inspected the results of the table.

# Appendix 2 – Programming Code

**All analysis was performed in SAS Studio 3.8**

```
/* Import the datafile into SAS */

PROC IMPORT DATAFILE='/folders/myfolders/7100/project/2019_nfl_combine_results - Clean.csv'
        OUT=combine
        DBMS=CSV replace;
        GETNAMES=YES;
RUN;
```

## Method 1

```
PROC GLM data = combine PLOTS(UNPACK)=DIAGNOSTICS; /*direct GLM to use the combine dataset and show all plots for residuals of the dependent variable*/
class zone; /* define the class (x-axis of the plots) as "zone"*/
model hand_size = zone; /*direct GLM to use hand_size as the dependent variable (y-axis) and zone and the independent variable (x-axis)*/
run;
```

## Method 2

```
proc Reg data=combine2; /* direct reg to use the combine2 dataset, which removed all null values of 40-yard dash times */
model _40_yard = height; /*direct reg to set the model as 40-yard dash time as the dependent variable and height as the independent variable*/
run;
```

# Appendix 3 – Relevant Output

## Method 1

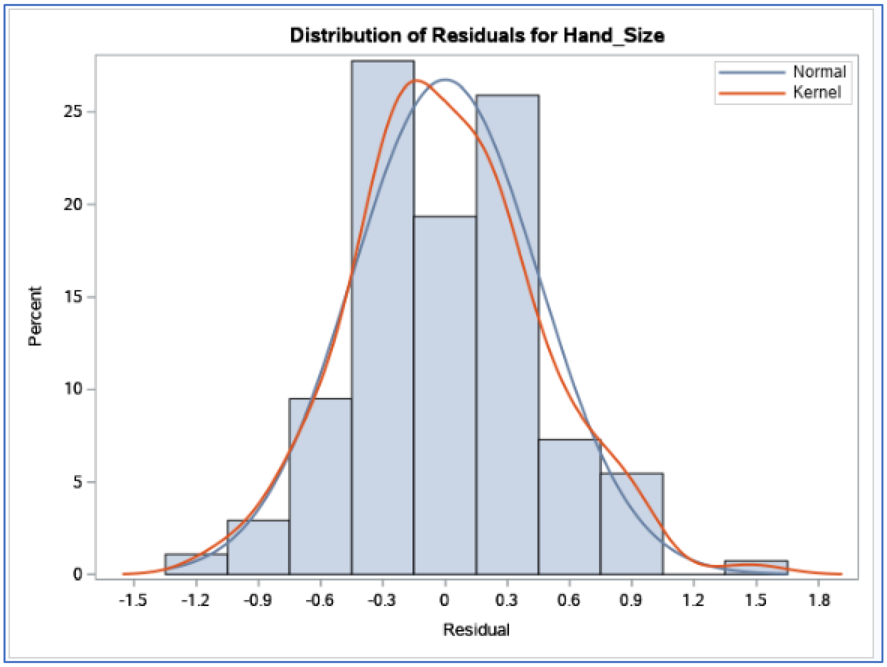


*Figure 1*

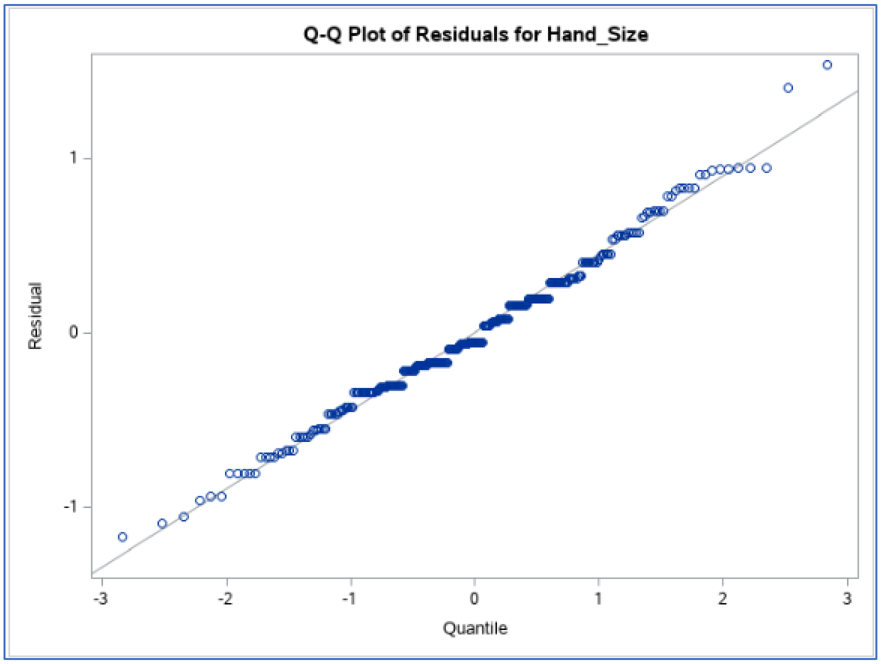
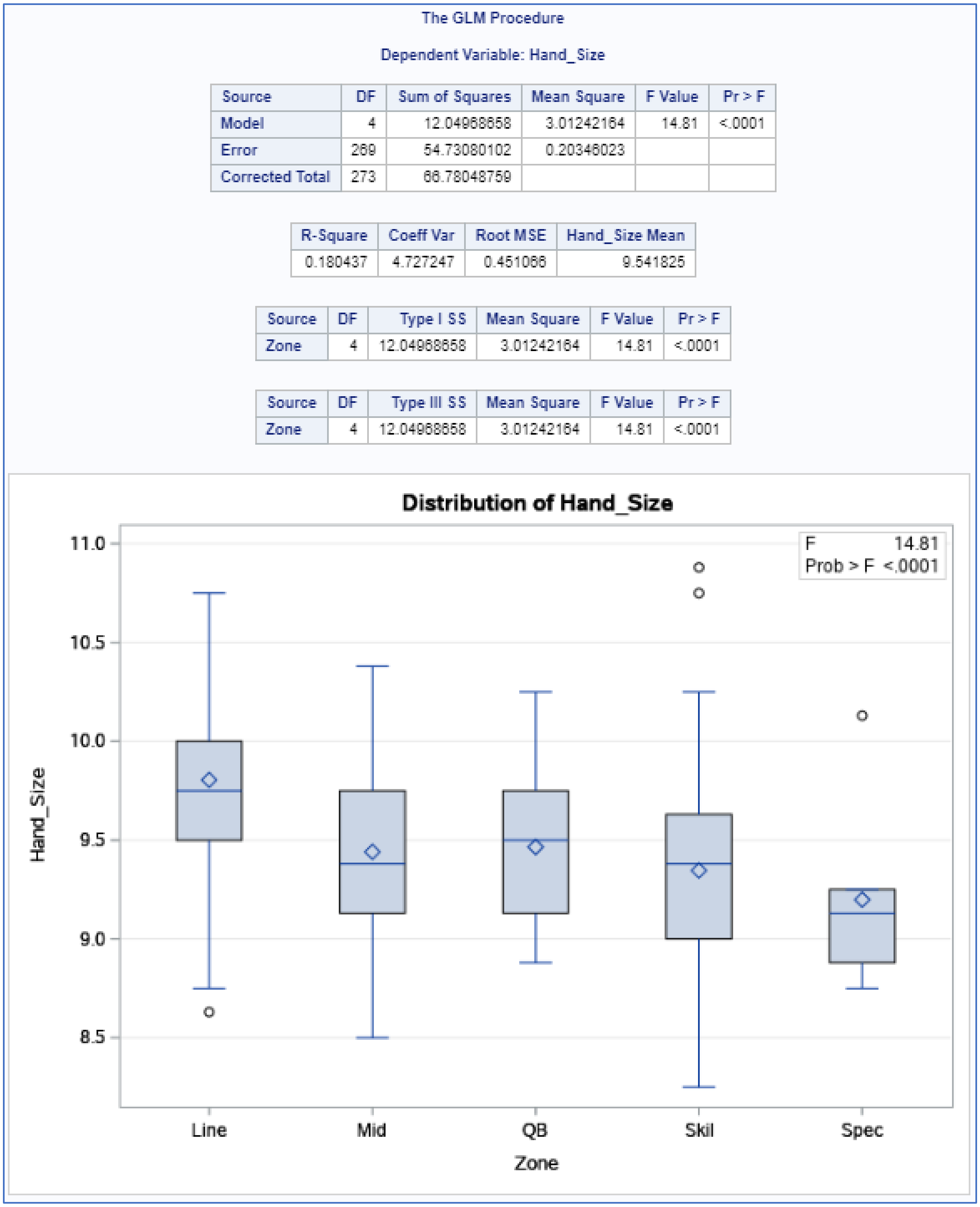


*Figure 2*



The GLM Procedure

Dependent Variable: Hand_Size

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 4 | 12.04968658 | 3.01242164 | 14.81 | <.0001 |
| Error | 269 | 54.73080102 | 0.20346023 | | |
| Corrected Total | 273 | 66.78048759 | | | |

| R-Square | Coeff Var | Root MSE | Hand_Size Mean |
|---|---|---|---|
| 0.180437 | 4.727247 | 0.451066 | 9.541825 |

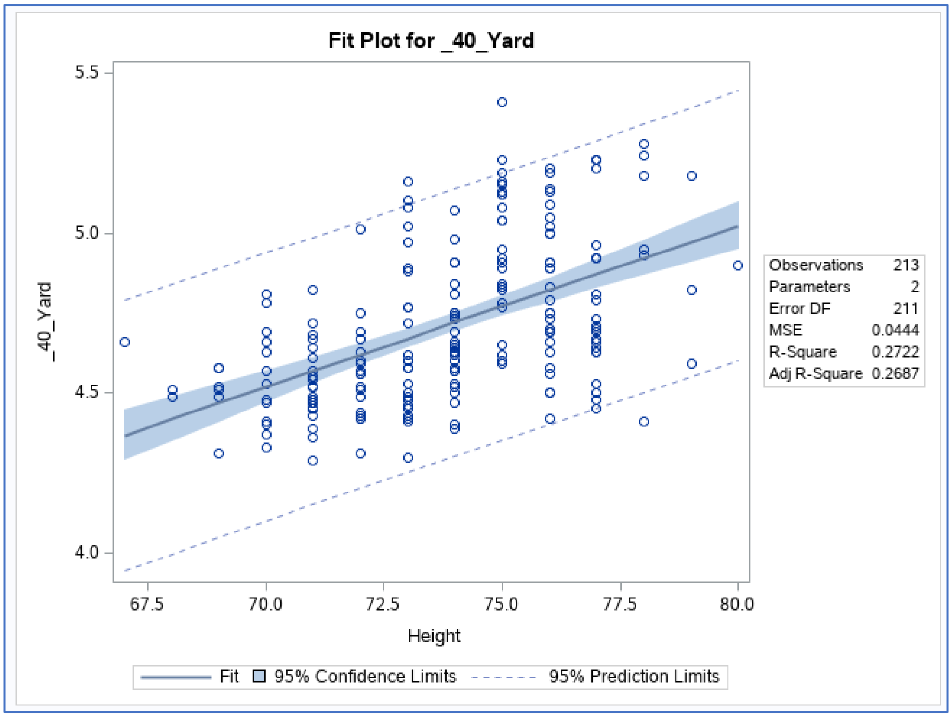
| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Zone | 4 | 12.04968658 | 3.01242164 | 14.81 | <.0001 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Zone | 4 | 12.04968658 | 3.01242164 | 14.81 | <.0001 |

Distribution of Hand_Size

*Figure 3*

## Method 2



| | |
|---|---|
| Observations | 213 |
| Parameters | 2 |
| Error DF | 211 |
| MSE | 0.0444 |
| R-Square | 0.2722 |
| Adj R-Square | 0.2687 |

*Figure 4*

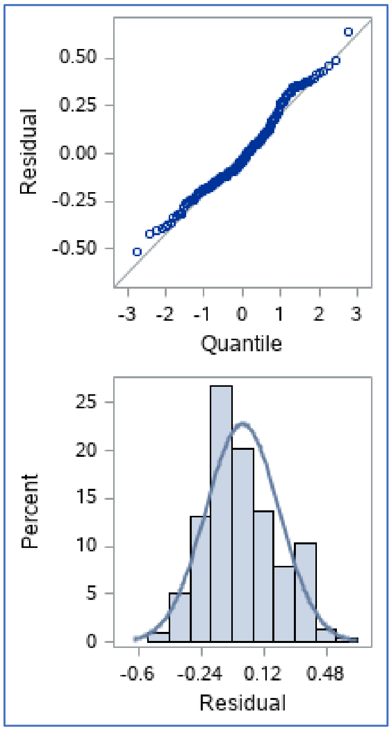
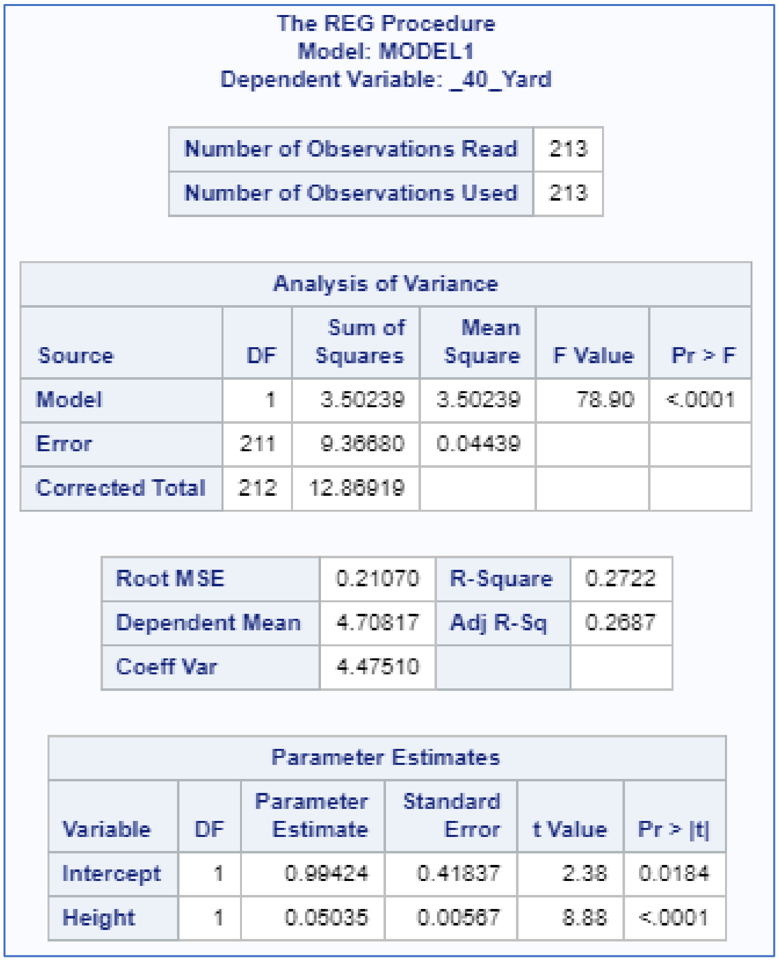


*Figure 5*



The REG Procedure
Model: MODEL1
Dependent Variable: _40_Yard

| | |
|---|---|
| Number of Observations Read | 213 |
| Number of Observations Used | 213 |

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 1 | 3.50239 | 3.50239 | 78.90 | <.0001 |
| Error | 211 | 9.36680 | 0.04439 | | |
| Corrected Total | 212 | 12.86919 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.21070 | R-Square | 0.2722 |
| Dependent Mean | 4.70817 | Adj R-Sq | 0.2687 |
| Coeff Var | 4.47510 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 0.99424 | 0.41837 | 2.38 | 0.0184 |
| Height | 1 | 0.05035 | 0.00567 | 8.88 | <.0001 |

*Figure 6*