Pross - Lab 5

Sebastian Pross

2022-05-05

INTRODUCTION

One of the many sports offered at the collegiate level is Track and Field. Personally, I am a member of Ramapo College's track and field team and specialize in the hammer throw. The purpose of competing in a throwing event is to throw the implement as far as possible. Currently, it is difficult to predict exactly where a person's best throw could be just by looking at them, but this lab poses the question: What can we use to predict a person's best throw, or PR, specifically focusing on a person's hammer throw PR. This lab will test whether or not three different predictors are significant in determining a person's hammer PR in meters. Those predictors being the amount of collegiate years they have spent throwing, their Discus PR (measured in meters), and their collegiate division (Division 1, 2, or 3).

DATA COLLECTION

In order to collect the data, the method used is known as stratified random sampling. Using an online database for track and field meet results known as TFRRS, I randomly chose 15 athletes from the top 500 ranking in each collegiate division (I, II, and III), and took note of their year (Freshman, Sophomore, Junior, or Senior, relating to 1 year - 4 years collegiate throwing experience) as well as their discus PR and their hammer PR, both being measured in meters. (Side note: The last athlete for Division III is actually me, and it happened by complete chance).

MODEL + MODEL ASSUMPTIONS

Because division is technically a categorical predictor, we need to use dummy variables. Because there are three categories, Division 1, 2, and 3, we need to use two dummy variables. This analysis will use a linear regression model. The following is the mode of the relationship for the response variable, Hammer PR, and the three predictors.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon,\ \epsilon\ N(0, \sigma)$$

Where the response variable and coefficients are defined as:

Y = Hammer PR
x_1 = Collegiate Years Throwing
x_2 = Discus PR
$x_3 = \begin{cases} 1 \text{ if D1} \\ 0 \text{ Otherwise} \end{cases}$
$x_4 = \begin{cases} 1 \text{ if D2} \\ 0 \text{ Otherwise} \end{cases}$

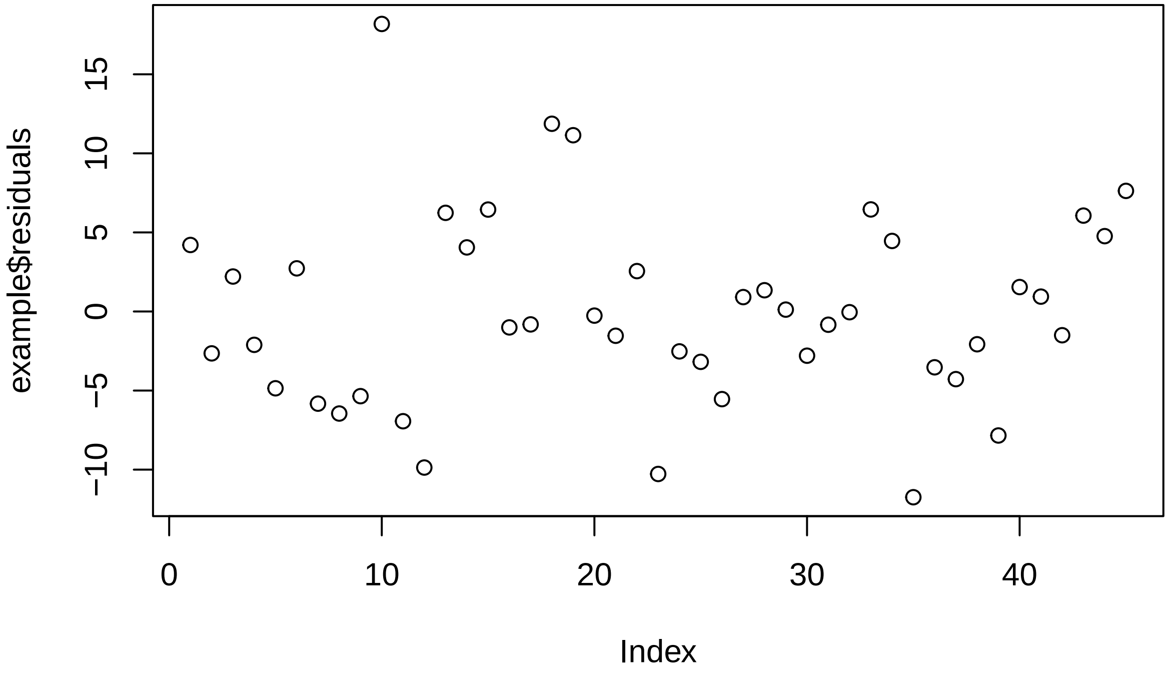
$$\begin{cases} H_o : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_a : \text{not all } \beta_i \text{ are } 0 \end{cases}$$

In terms of model assumptions, it is important to check if the residuals have any form of normality, shown in the following Shapiro-Wilk normality test. The Shapiro-Wilk normality test is defined with the hypothesis:

$$\begin{cases} H_o : \text{erors are normally distributed} \\ H_a : \text{errors are being drawn from a non-normal distribution} \end{cases}$$

```
##
##  Shapiro-Wilk normality test
##
## data:  example$residuals
## W = 0.97753, p-value = 0.5233
```

As we can see from the Shapiro-Wilk normality test, the p-value is 0.5233. And an an alpha level of 0.05, the p-value is much greater than the alpha level. Therefore, we can conclude that the residuals are indeed normally distributed. However, we must now check that the residuals have constant variance with the following residual plot:



Since there seems to be no pattern to the residual plot, and the points are all relatively randomly dispersed, we can conclude that a linear regression model is indeed appropriate.

MODEL PARAMETERS

We can find and interpret the model parameters with the summary function of the linear regression model:

```
##
## Call:
## lm(formula = Hammer ~ Years + Discus + d1 + d2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.7450  -3.5305  -0.8172   4.0512  18.1815
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.9167     6.7571   3.244  0.00239 **
## Years         2.5867     0.9620   2.689  0.01040 *
## Discus        0.4893     0.1604   3.051  0.00404 **
## d1            3.6576     2.8123   1.301  0.20085
## d2           -1.8934     2.3851  -0.794  0.43199
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.327 on 40 degrees of freedom
## Multiple R-squared:  0.4718, Adjusted R-squared:  0.419
## F-statistic: 8.933 on 4 and 40 DF,  p-value: 2.981e-05
```

As we can see, the coefficients for the model are given in the linear regression model in the "estimates" column. $\hat{\beta}_1 = 2.5867$ (Collegiate Years Throwing), $\hat{\beta}_2 = 0.4893$ (Discus PR), $\hat{\beta}_3 = 3.6576$ (Divison 1), and $\hat{\beta}_4 = -1.8934$ (Division 2).

The beta coefficient is the degree of change in the response variable for every unit change of the predictors. It seems that Hammer PR tends to increase with more years throwing as well as with a higher discus PR. Although, from the model we can see that collegiate division is not significant in predicting hammer PR. However, from going to Division 3 to Division 2, Hammer PR tends to decrease by 1.8934 meters, and from D3 to D1, it tends to increase by 3.6576 meters. For collegiate years throwing, this model shows that for each year increase, the athlete's hammer PR increases on average by 2.5867 meters. Similarly, for every meter increase in discus PR, an athlete's hammer PR increases by 0.4893 meters on average.

Since the divisions do not seem to be significant predictors because of their relatively high p-value, we can iteratively remove the Division 2 predictor. This results in a decrease in the p-value of the Division 1 predictor, to the point where it is indeed significant. That means that an athlete being a member of a division 1 college rather than division 2 or 3 is a significant effect an athlete's on hammer PR.

CONCLUSIONS

Overall, this model was not very well fit. The r squared value is 0.4718, meaning 47.18% of the data is represented by the model of linear regression. While this model does account for some, 47.18% is not a good fit. In terms of predicting hammer PR, we can see that farther and better performances in the other events (specifically Discus) and more experience seems to indicate positive change in hammer PR, division does not play a large role from Division 3 to Division 2, but from those it differs largely into Division 1, which makes sense, because those programs have larger budgets, recruit much more and much stricter, overall just have better facilities and resources to train their athletes to perform better.