STAT 350 Term Project

Identify the Factors That Affects a Striker's Rating on FIFA 18

Youneng Zou

November 30, 2017

Table of Contents

Abstract

In this paper, we will investigate the factors that influence the player ratings based on FIFA 18. In the dataset, there are 15,973 players' data and 7 variables. we will analyize the scatterplots matrix and the full linear regression model first, then we will use AIC and BIC methods to do model selection. Finally, I will get a model to show that the ratings of right striker are affected by 6 main player's technical attributes and the interactions between them.

Keywords:

Right striker rating, pace, dribbling, shooting, defending, passing, physical

Introduction

FIFA 18 is one of the most famous football game developed by EA sports. In FIFA 18, a player's rating is unaffected by the "left" or "right". That means, a player's rating is the same in LB (Left Back) and RB (Right Back). So we only need to consider these 11 positions: GK, RWB, RB, RCB, RDM, RM, RCM, RAM, RF, RW, RS. A player's overall rating is the highest one between these 11 positions. There is a different linear combination of a player's main attributes (e.g. speed, acceleration etc.) for each position, and it also adds a multiplier of the player's

international reputation. In this project, we will use linear regression model to identify which factors had significant effects on the rating of right striker.

Dataset:

The dataset we got had 307 observations with 6 variables, it is shared at Kaggle.com. It contains the complete dataset of players in FIFA 18. It contains player information such as age, club, league, nationality, salary, and all player attributes, even special attributes and traits and specialties etc. So we have to do data cleaning to remove a lots of columns before we can use the data.

Response variables:

1. rs: Rating on right striker position (Numerical)

Explanatory variables:

1. IR: International Reputation of player (Categorical)
2. PAC: Pace (Numerical)
3. DRI: Dribbling (Numerical)
4. SHO: Shooting (Numerical)
5. PAS: Passing (Numerical)
6. PHY: Physical (Numerical)

Analysis to do:

1. To construct and determine an appropriate linear regression model as the final model.
2. To check the goodness of fit of the final model.
3. To interpret the effects of the explanatory variables in the final linear regression model, and test explanatory variables significance.

Methods and Models.

In this project, we used multiple linear regression models to analyze the rating of player on right striker position. Based on the lecture materials, if we want to use multiple linear regression model to fit a data, then we will have following function:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2), \varepsilon \ is \ independent \ and \ identically \ distributed \ (iid)$$

And we need to satisfy the following five assumptions to use the model:

1. Assume that y_i is a linear function of $x_{i1}, x_{i2}, \dots, x_{ip}$.
2. Assume that ε_i has zero mean.
3. Assume that ε_i has constant variance.
4. Assume that ε_i are uncorrelated.
5. Assume that ε_i follows normal distribution.

So we need to use scatterplot matrix for the dataset to check whether there are linear

relationships between the response variable and explanatory variables, which is the first assumption. In this process, we also want to check whether there are outliers in the dataset and whether we need transformations for the variables.

Firstly, we will create our full model. Meanwhile, we will use correlation plot to select the suitable interactions between explanatory variables. It includes all the explanatory variables and some important interaction terms:

$$M_0: E(rs) = \beta_0 + \beta_1 * IR + \beta_2 * PAC + \beta_3 * DRI + \beta_4 * SHO + \beta_5 * PAS + \beta_6 * PHY + \beta_7 * (PAC * SHO) + \beta_8 * (PAC * PAS) + \beta_9 * (PAC * DRI) + \beta_{10} * (PAC * PHY) + \beta_{11} * (PAC * IR) + \beta_{12} * (SHO * PAS) + \beta_{13} * (SHO * DRI) + \beta_{14} * (SHO * PHY) + \beta_{15} * (SHO * IR) + \beta_{16} * (PAS * DRI) + \beta_{17} * (PAS * PHY) + \beta_{18} * (DRI * PHY) + \beta_{19} * (DRI * IR) + \beta_{20} * (PHY * IR)$$

Where IR is categorical variable contains number from 0 to 5.

We studied the residuals plot, normal Q-Q plot and the histogram of the residuals for the full model. We also checked whether we need transformation for the variables. Then we analyzed the summary table of the full model.

In order to get a better linear regression model, we also did model selection to get a better model. In this project, both Akaike's Information Criterion (AIC) and Bayesian's Information Criterion (BIC) are used.

Following are the summary of the model selection:

Information Criteria	Direction	Model	AIC/BIC Value
AIC	Forward	rs ~ SHO + IR + PHY + PAC + SHO:IR + SHO:PAC + IR:PHY + SHO:PHY + IR:PAC	145.7
AIC	Backward	rs ~ IR + PAC + SHO + DRI + PAS + PHY + PAC:SHO + SHO:DRI + SHO:PHY + IR:SHO + PAS:PHY	136.91
AIC	Both	rs ~ IR + PAC + SHO + DRI + PAS + PHY + PAC:SHO + SHO:DRI + SHO:PHY + IR:SHO + PAS:PHY	136.91
BIC	Forward	rs ~ SHO + IR + PHY + PAC + SHO:IR + SHO:PAC + IR:PHY	176.21
BIC	Backward	rs ~ IR + PAC + SHO + DRI + PHY + PAC:SHO + SHO:DRI + SHO:PHY + IR:SHO	174.61
BIC	Both	rs ~ IR + PAC + SHO + DRI + PHY + PAC:SHO + SHO:DRI + SHO:PHY + IR:SHO	174.61

For the AIC method, forward direction gave the AIC value 145.7, backward and both direction gave the same model with the value 136.91. Since it was smaller, we decided to choose this model. The result of AIC method is below:

$$M_{AIC}: E(rs) = \beta_0 + \beta_1 * IR + \beta_2 * PAC + \beta_3 * DRI + \beta_4 * SHO + \beta_5 * PAS + \beta_6 * PHY + \beta_7 * (PAC * SHO) + \beta_8 * (SHO * DRI) + \beta_9 * (SHO * PHY) + \beta_{10} * (SHO * IR) + \beta_{11} * (PAS * PHY)$$

Where IR is categorical variable contains number from 0 to 5.

For BIC method, forward direction had the value 176.21, but backward and both direction had smaller value which is 174.61. The result is below:

$$M_{BIC}: E(rs) = \beta_0 + \beta_1 * IR + \beta_2 * PAC + \beta_3 * DRI + \beta_4 * SHO + \beta_5 * PHY + \beta_6 * (PAC * SHO) + \beta_7 * (SHO * DRI) + \beta_8 * (SHO * PHY) + \beta_9 * (SHO * IR)$$

Where IR is categorical variable contains number from 0 to 5.

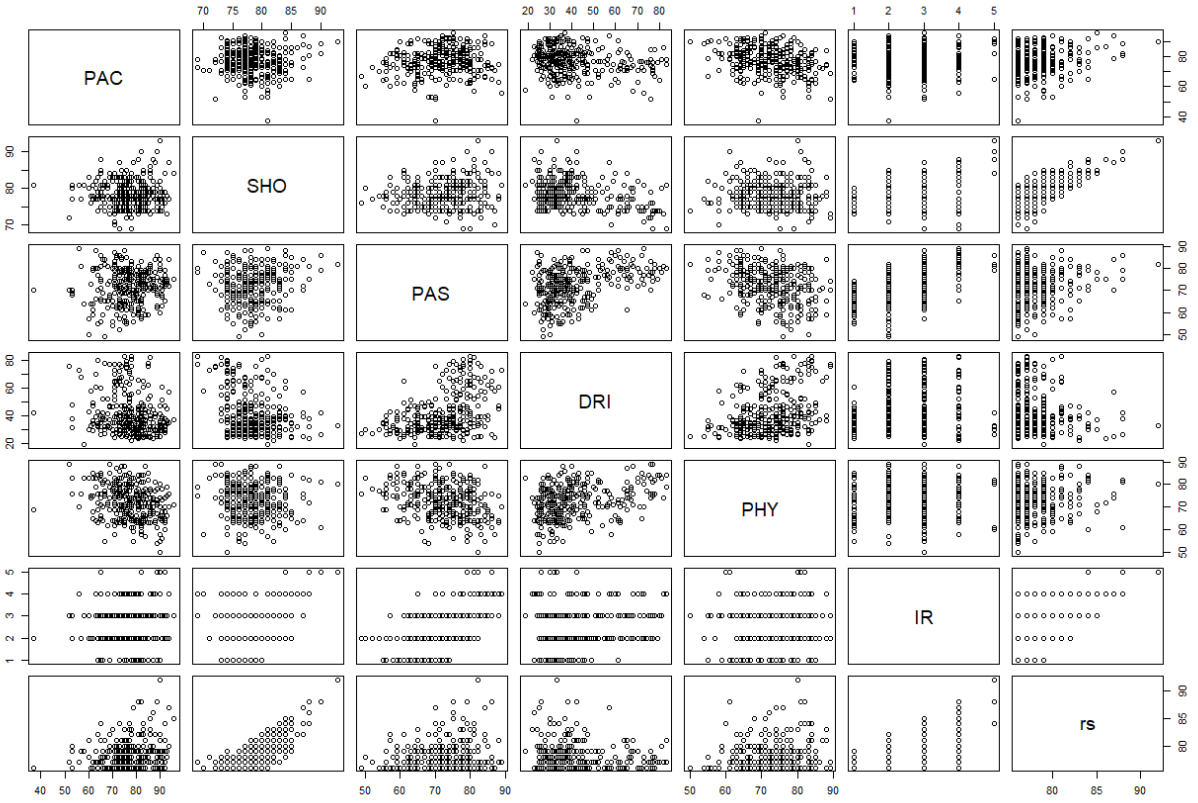
Based on the summary of the AIC and BIC model, we chose AIC model. After that, we constructed its residuals plot, normal Q-Q plot and the histogram from the distribution of the residuals to check whether the AIC model satisfied the five assumption of the multiple linear regression model.

Results

Correlation Plot Study

This is the correlation plot for all the variables in the dataset. Since we want to determine the suitable interaction between explanatory variables. We will choose the explanatory variables has less correlations, which is the absolute value less than 0.5. Therefore, we decided the interactions in the full model.

Scatterplot Matrix Study



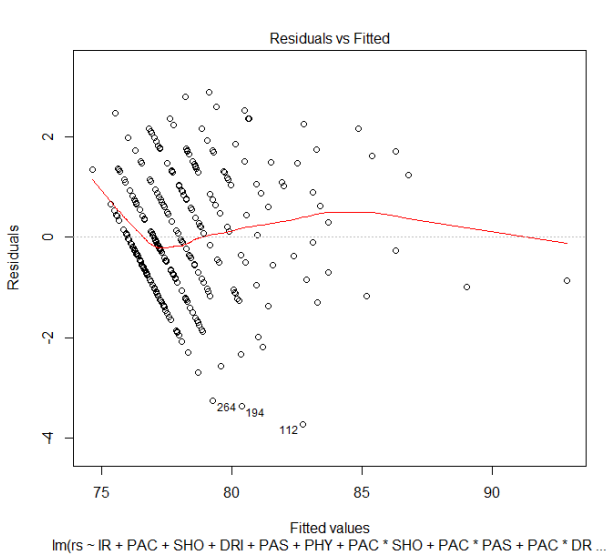
This is the scatterplot matrix of the response variable. From this plot, we can see that the explanatory variables PAC, SHO, PAS, DRI, PHY and IR may be linearly related to the response variable rs. And there are no obvious outliers and nonlinear pattern, so we won't use any transformation for the model.

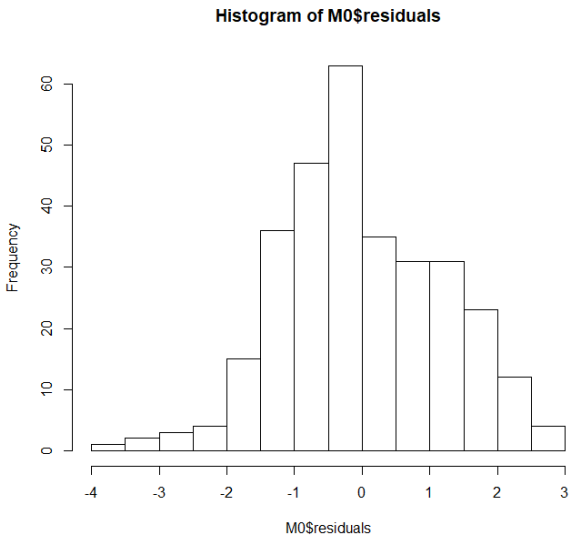
Analysis on the Full Model

The full model we constructed has 6 explanatory variables and 14 interactions, with the summary of the full model below:

Summary of full model				
Estimated σ	1.237	Significant Predictors	Estimated Value	p-value
R^2	0.7832	(Intercept)	114.0	0.00982**
F-statistic	51.67	IR	-8.142	0.00810**
F-stat p-value	<2.2e-16	PAC	-0.6373	0.01640*
		PHY	-0.6967	0.04237*
		PAC:SHO	0.007629	0.01016*
		SHO:DRI	0.003740	0.03029*
		IR:SHO	0.08980	0.00272**

In the full model, 3 of the 6 explanatory variables and 3 of the 14 interactions are significant with p-value < 0.05. So we can conclude that the model satisfies the first assumption of the multiple linear regression model.





From the residuals plot, the residuals were randomly located around at 0, so we concluded the

error term ε_i had zero mean and constant variance, also the error terms are independent to each other. In the histogram of residuals, the data was normally distributed. Therefore, we concluded that the full model satisfied the five assumptions of multiply linear regression model.

Model Selection

In order to get better model, we used AIC and BIC method. For AIC method, the selected model is below:

$$M_{AIC}: E(rs) = \beta_0 + \beta_1 * IR + \beta_2 * PAC + \beta_3 * DRI + \beta_4 * SHO + \beta_5 * PAS + \beta_6 * PHY + \beta_7 * (PAC * SHO) + \beta_8 * (SHO * DRI) + \beta_9 * (SHO * PHY) + \beta_{10} * (SHO * IR) + \beta_{11} * (PAS * PHY)$$

Where IR is categorical variable contains number from 0 to 5.

Here is the summary table for AIC model:

Estimated σ	1.226	Significant Predictors	Estimated Value	p-value
R^2	0.7803	(Intercept)	152.110477	2.8e-09***
F-statistic	95.25	IR	-4.831464	0.001995**
F-stat p-value	<2.2e-16	PAC	-0.648033	0.001017**
		SHO	-0.902592	0.002728**
		DRI	0.374280	0.000681***
		PAS	-0.190580	0.039198*
		PHY	-0.815020	0.000328***

	PAC:SHO	0.008957	0.000343***
	SHO:DRI	-0.004818	0.000695***
	SHO:PHY	0.008918	0.001018**
	IR:SHO	0.069728	0.000435***
	PAS:PHY	0.002541	0.042418*

For BIC method, the selected model is:

$$M_{BIC}: E(rs) = \beta_0 + \beta_1 * IR + \beta_2 * PAC + \beta_3 * DRI + \beta_4 * SHO + \beta_5 * PHY + \beta_6 * (PAC * SHO) + \beta_7 * (SHO * DRI) + \beta_8 * (SHO * PHY) + \beta_9 * (SHO * IR)$$

Where IR is categorical variable contains number from 0 to 5.

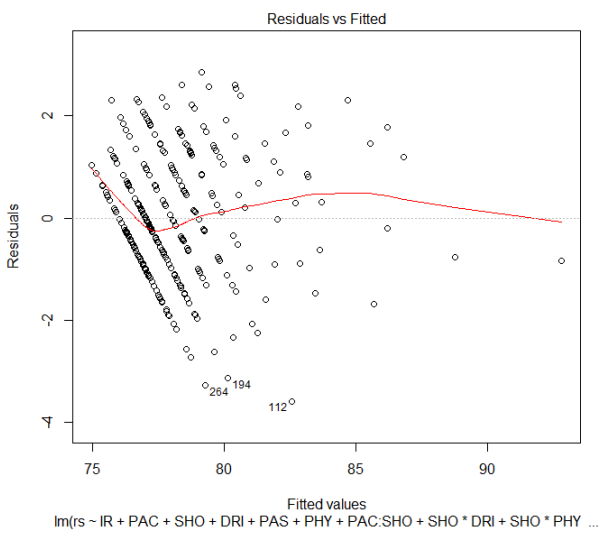
Here is the summary table for BIC model:

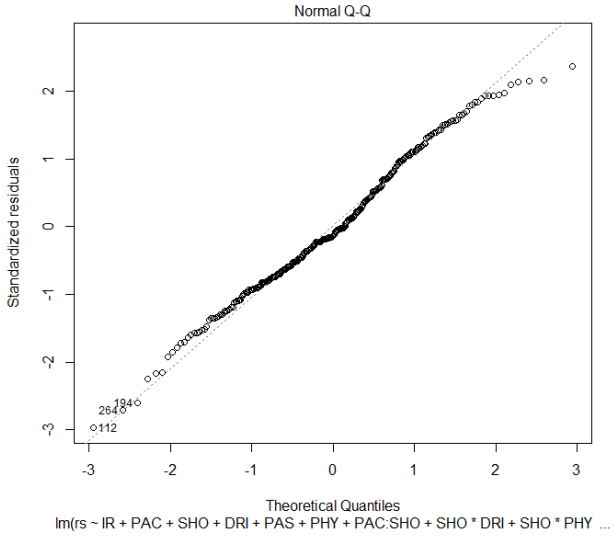
Estimated σ	1.219	Significant Predictors	Estimated Value	p-value
R^2	0.7814	(Intercept)	115.497005	0.006926**
F-statistic	118	IR	-5.688228	0.009543**
F-stat p-value	<2.2e-16	PAC	-0.780063	0.000258***
		PAC:SHO	0.010433	0.000125***
		IR:SHO	0.077957	0.005185**

From the summary, we can see that all the predictors of AIC model are significant, so we choose it as our final model.

Final Model Analysis

The final model we selected was from AIC method, which is M_{AIC}. The R^2 value is 0.7814 and the p-value of the model is <2.2e-16, which is significant. One important fact is that all predictors of this model are significant. The following two plots showed that this model also satisfied the five assumptions of multiply linear regression model, we concluded that the final model is good enough.





Conclusion

In conclusion, the player rating at right striker position is linear related to the following predictors:

1. International reputation level
2. Pace attribute rating
3. Shooting attribute rating

4. Dribbling attribute rating
5. Passing attribute rating
6. Physical attribute rating
7. The interaction between pace and shooting
8. The interaction between shooting and dribbling
9. The interaction between shooting and physical
10. The interaction between international reputation and shooting
11. The interaction between passing and physical

And all of these 11 predictors are significant. In conclusion, the player rating at right striker position has strong relationship with all the 5 main attributes, international reputation and other 5 interactions. And our final model is below:

$$M_{AIC}: E(rs) = 152.110477 - 4.831464 * IR - 0.648033 * PAC + 0.374280 * DRI - 0.902592 * SHO - 0.190580 * PAS - 0.815020 * PHY + 0.008957 * (PAC * SHO) - 0.004818 * (SHO * DRI) + 0.008918 * (SHO * PHY) + 0.069728 * (SHO * IR) + 0.002541 * (PAS * PHY)$$

Where IR is categorical variable contains number from 0 to 5.

Potential Problems

Although our final model is quiet well, there are still potential problems:

1. Since our data from the real dataset of FIFA 18, which means it is a game data, we can basically conclude that there must be a linear combination of the player's attributes for right striker position.
2. Our model is only work for multiple linear regression model, so we don't know whether there is model other than linear relationship is better than this one.

Improvement

In order to get more accuracy function for player's rating at right striker position, we can improve our investigation by including all attributes in the calculation. These will contain more than 40 variables, that will make our model too complicated, so I decided to leave it for further improvement.

References

Kevin. (n.d.). Retrieved December 02, 2017, from https://www.kaggle.com/kevinmh/evaluating-the-sofifa-com-rating-calculator/data

Appendix

Summary of full model:

```
Call:
lm(formula = rs ~ IR + PAC + SHO + DRI + PAS + PHY + PAC * SHO +
    PAC * PAS + PAC * DRI + PAC * PHY + PAC * IR + SHO * PAS +
    SHO * DRI + SHO * PHY + SHO * IR + PAS * DRI + PAS * PHY +
```

```
   DRI * PHY + DRI * IR + PHY * IR, data = RS)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7206 -0.7614 -0.1451  0.8683  2.8804

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.140e+02  4.387e+01   2.599  0.00982 **
IR          -8.142e+00  3.053e+00  -2.667  0.00810 **
PAC         -6.373e-01  2.640e-01  -2.414  0.01640 *
SHO         -3.657e-01  5.080e-01  -0.720  0.47224
DRI          3.401e-01  1.872e-01   1.817  0.07032 .
PAS          2.906e-01  4.347e-01   0.668  0.50442
PHY         -6.967e-01  3.417e-01  -2.039  0.04237 *
PAC:SHO      7.629e-03  2.948e-03   2.588  0.01016 *
PAC:PAS     -2.361e-05  1.730e-03  -0.014  0.98912
PAC:DRI     -3.969e-04  8.290e-04  -0.479  0.63251
PAC:PHY      1.084e-03  1.448e-03   0.748  0.45498
IR:PAC       1.310e-02  1.446e-02   0.906  0.36577
SHO:PAS     -5.067e-03  4.686e-03  -1.081  0.28052
SHO:DRI     -3.740e-03  1.718e-03  -2.177  0.03029 *
SHO:PHY      6.591e-03  3.753e-03   1.756  0.08011 .
IR:SHO       8.980e-02  2.970e-02   3.024  0.00272 **
DRI:PAS     -6.350e-04  1.042e-03  -0.609  0.54276
PAS:PHY      1.756e-03  1.765e-03   0.995  0.32061
DRI:PHY      3.421e-04  8.328e-04   0.411  0.68154
IR:DRI       8.919e-04  8.202e-03   0.109  0.91348
IR:PHY       9.668e-03  1.552e-02   0.623  0.53387
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.237 on 286 degrees of freedom
Multiple R-squared:  0.7832,    Adjusted R-squared:  0.7681
F-statistic: 51.67 on 20 and 286 DF,  p-value: < 2.2e-16
```

Summary of M_{AIC}:

```
Call:
lm(formula = rs ~ IR + PAC + SHO + DRI + PAS + PHY + PAC:SHO +
```

```
    SHO * DRI + SHO * PHY + IR * SHO + PAS * PHY, data = RS)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5793 -0.8342 -0.1635  0.8581  2.8481

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 152.110477  24.812311   6.130  2.8e-09 ***
IR           -4.831464   1.549086  -3.119 0.001995 **
PAC          -0.648033   0.195254  -3.319 0.001017 **
SHO          -0.902592   0.298638  -3.022 0.002728 **
DRI           0.374280   0.109001   3.434 0.000681 ***
PAS          -0.190580   0.092009  -2.071 0.039198 *
PHY          -0.815020   0.224226  -3.635 0.000328 ***
PAC:SHO       0.008957   0.002473   3.622 0.000343 ***
SHO:DRI      -0.004818   0.001406  -3.428 0.000695 ***
SHO:PHY       0.008918   0.002687   3.319 0.001018 **
IR:SHO        0.069728   0.019595   3.558 0.000435 ***
PAS:PHY       0.002541   0.001247   2.038 0.042418 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.226 on 295 degrees of freedom
Multiple R-squared:  0.7803,  Adjusted R-squared:  0.7721
F-statistic: 95.25 on 11 and 295 DF,  p-value: < 2.2e-16
```

Summary of M_{BIC}:

Call:

```
lm(formula = rs ~ IR + PAC + SHO + DRI + PHY + PAC * SHO + SHO *
    DRI + SHO * PHY + IR * SHO)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4915 -0.8043 -0.0704  0.7274  3.1236

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 115.497005  42.472524   2.719 0.006926 **
IR           -5.688228   2.180297  -2.609 0.009543 **
PAC          -0.780063   0.210915  -3.698 0.000258 ***
```

```
SHO         -0.708165   0.540324  -1.311 0.190995
DRI          0.392743   0.463642   0.847 0.397631
PHY         -0.364009   0.239692  -1.519 0.129914
PAC:SHO      0.010433   0.002684   3.887 0.000125 ***
SHO:DRI     -0.003925   0.005910  -0.664 0.507098
SHO:PHY      0.005952   0.003072   1.937 0.053638 .
IR:SHO       0.077957   0.027682   2.816 0.005185 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.219 on 297 degrees of freedom
Multiple R-squared:  0.7814,  Adjusted R-squared:  0.7748
F-statistic:   118 on 9 and 297 DF,  p-value: < 2.2e-16
```