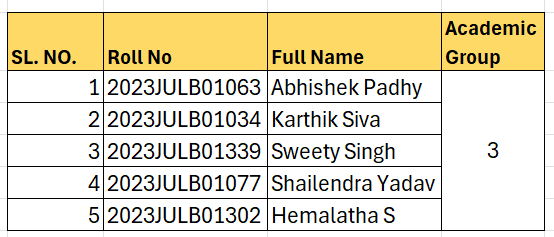
**Case 2 : TRANSFER VALUE OF SOCCER PLAYERS**

****

**1. Who is Kayla Nightingale? What does she want to do?**

Kayla Nightingale is a soccer fanatic and keeps a close eye on European league scores, standings, and player transfers. She wants to figure out what factors make a player good and how clubs can use that info when deciding to transfer players between teams.

**2. What does the transfer value of a player mean?**

The transfer value in football refers to the monetary worth of a player when they are bought or sold by a club. It's influenced by factors like the player's skill, age, contract status, market demand, and recent performance. Transfers can range from a few thousand to hundreds of millions of dollars for top players.

**3. How many observations and variables are there in the dataset that Kayla has put together?**

Kayla's dataset comprises 77 instances of football player transfers, encompassing 20 distinct variables.

4. **What is the dependent variable? Which are the independent variables?**

The dependent variable is the "**GBP\_M**" column, which represents the transfer fee in British pounds. The independent variables are the other columns, such as **"Y\_5," "Y\_4," "Y\_3," "Y\_2," "Y\_1," "WA," "Goals," "App," "GA," "Age," "Height," "Pos," "Foot," "CR," and "NR."**

**5. What are the continuous, binary, and categorical variables in the dataset?**

**Continuous variables:**

GBP\_M (Transfer fee in British pounds)

Y\_5, Y\_4, Y\_3, Y\_2, Y\_1 (Performance metrics over the past five years)

WA (Weighted average performance metric)

Goals (Number of goals scored)

App (Number of appearances)

GA (Goals against, possibly for goalkeepers)

Age (Player's age)

Height (Player's height)

**Binary variables:**

CR (Club ranking)

POS (Position of a player)

**Categorical variables:**

Year (Year of the transfer)

Player Name (Name of the player)

From (Previous club)

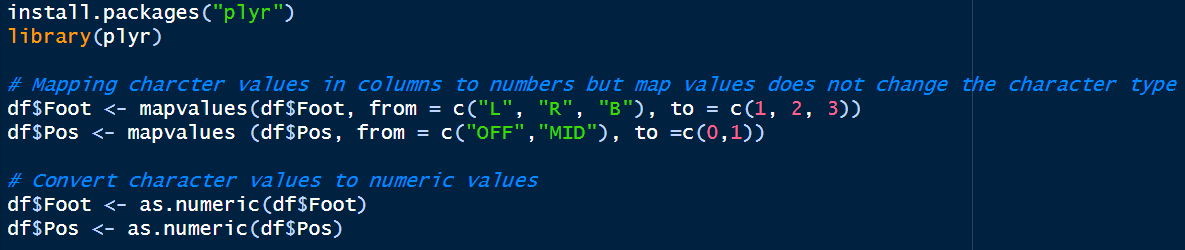
To (New club)

Foot (Preferred foot: Categorical)

NR (National team ranking).

**6. Make suitable conversions, so that all the variables to be analyzed are numeric.**





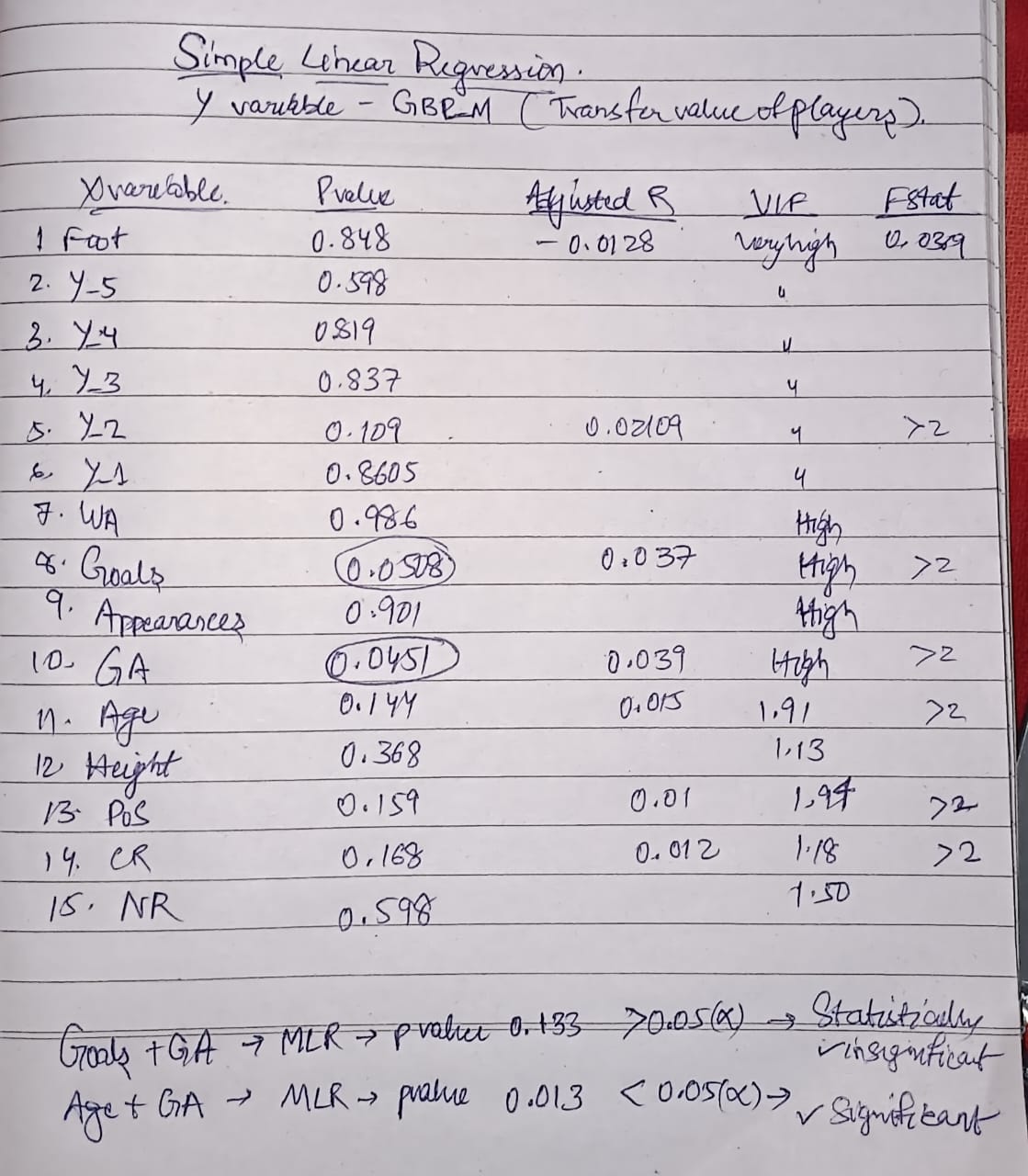
The binary variable Pos is transformed into a numeric form, while the category variable Foot is also converted into a numeric representation. This is done by introducing **Dummy variables** to the values given.

***7. Does the bivariate analysis indicate potential relationships? Explain with suitable evidence.***

Residual standard error: 16.09 on 75 degrees of freedom

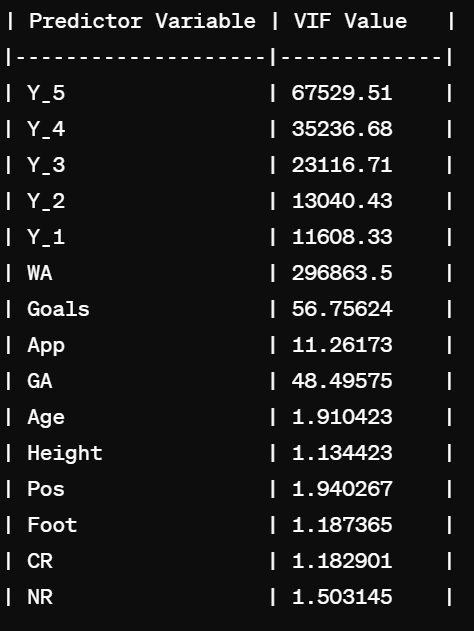
Multiple R-squared: 0.02813, Adjusted R-squared: 0.01517

F-statistic: 2.17 on 1 and 75 DF, p-value: 0.1449



The regression analysis results suggest that the model lacks strong explanatory power, with only about 2.8% of the variability in the dependent variable explained by the independent variable. Both the multiple R-squared and adjusted R-squared are low, indicating poor fit. Additionally, the F-statistic, used to test the overall significance of the model, is not statistically significant at the typical significance level of 0.05. Consequently, the relationship between the independent and dependent variables may not be significant.

***8. Which predictor variables are correlated with each other? Therefore, what changes will you make in your analysis? Explain.***



Based on the provided VIF values:

* The variables with higher VIF values, such as 'Y\_5', 'Y\_4', 'Y\_3', 'Y\_2', 'Y\_1', and 'WA', indicate potential multicollinearity issues with these predictors.
* Variables with VIF values closer to 1, such as 'Height', 'Pos', 'Foot', 'CR', and 'NR', suggest lower multicollinearity with other predictors.

There are a few ways to handle correlated predictor variables in your analysis. One option is to remove one of the correlated variables from your model. However, this can be problematic if the removed variable is an important predictor of the outcome variable.

So, I would recommend following these before making a decision:

* Look at how strongly the predictor variables are connected.
* Think about how important each predictor variable is for your research question.

***9. Which independent variables are statistically significant?***

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 49.9889 60.7569 0.823 0.4138

Y\_5 -1331.8071 1368.8353 -0.973 0.3344

Y\_4 -1333.2014 1369.4941 -0.973 0.3341

Y\_3 -1338.8078 1368.6279 -0.978 0.3318

Y\_2 -1323.9563 1370.9191 -0.966 0.3380

Y\_1 -1333.8251 1368.2042 -0.975 0.3335

WA 6668.1446 6851.3552 0.973 0.3343

Goals 0.3862 0.5053 0.764 0.4477

App -0.0299 0.2127 -0.141 0.8887

GA -24.3308 72.6910 -0.335 0.7390

Age -2.2227 0.9354 -2.376 0.0206 \*

Height 0.2002 0.3073 0.651 0.5172

POS2 2.1703 5.2283 0.415 0.6795

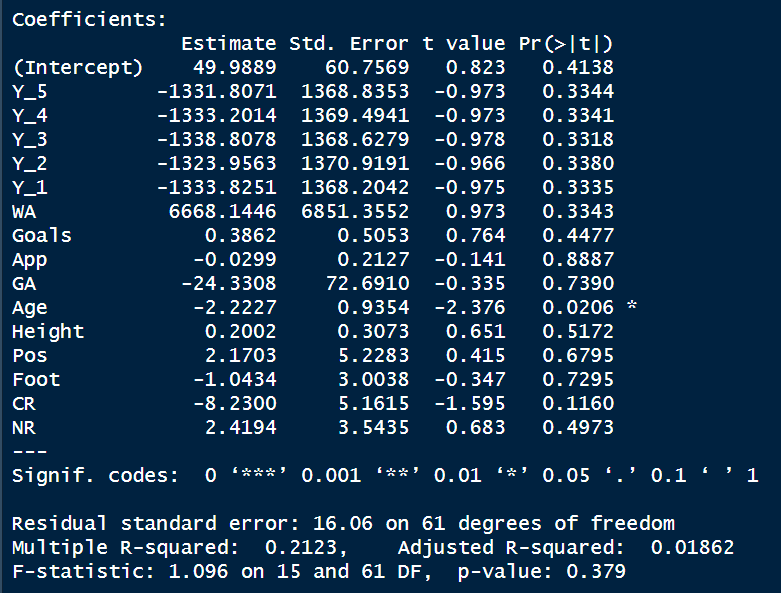
Foot3 -1.0434 3.0038 -0.347 0.7295

CR -8.2300 5.1615 -1.595 0.1160

NR 2.4194 3.5435 0.683 0.4973

* Among the independent variables listed in the table, Age stands out as statistically significant at the 5% level.
* Its p-value of 0.0206 is lower than the common threshold of 0.05, indicating its significance.
* This suggests that Age has a stronger association with market transfer value compared to the other variables.
* The lower p-value implies a higher dependency of market transfer value on Age.

***10. Do the necessary analysis and finalize a suitable model.***

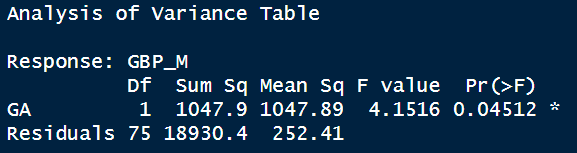
****

**INITIAL ANALYSIS**

* In this case, the variable **"Age"** stands out as significant with a p-value of **0.0206.**
* Based on these results, the final analysis suggests that the model, as currently specified, does not adequately explain the variability in the dependent variable.
* The low multiple R-squared **(21.23%)**indicates that the additional independent variables included in the model did not contribute meaningfully to explaining the variation in the dependent variable.
* To finalize the model, we considered to revise the model specification by possibly removing non-significant variables , using VIF to determine highly correlated variables which affect the multiple regression model.
* Additionally, it might be beneficial to collect more data or consider different variables that could better explain the variability in the dependent variable.

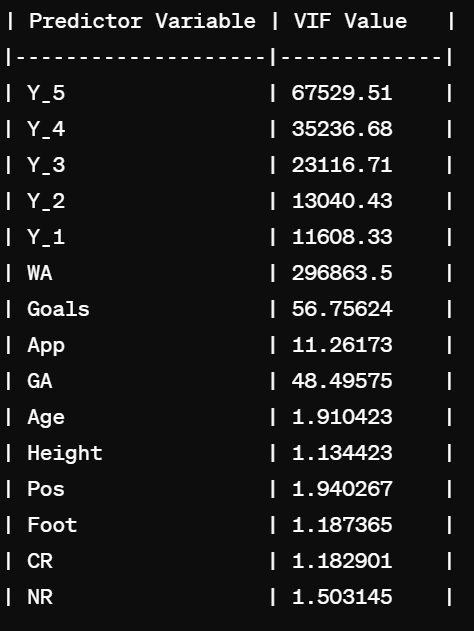
A screen shot of a computer

Description automatically generated



While doing SLR only Goals and GA independent variables were found significant.

Multicollinaearity



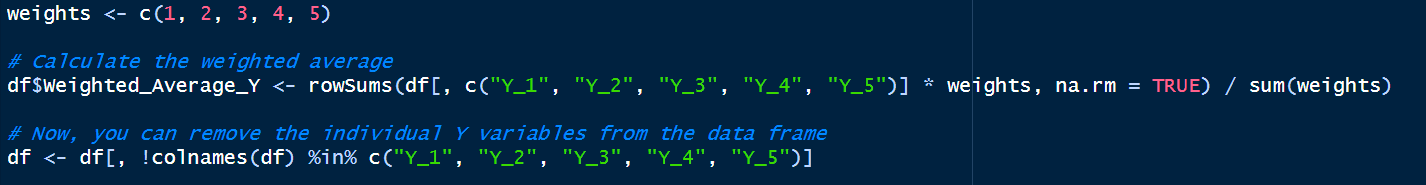
VIF is a measure of multicollinearity among predictor variables in a regression model. It quantifies how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

Interpreting the VIF values:

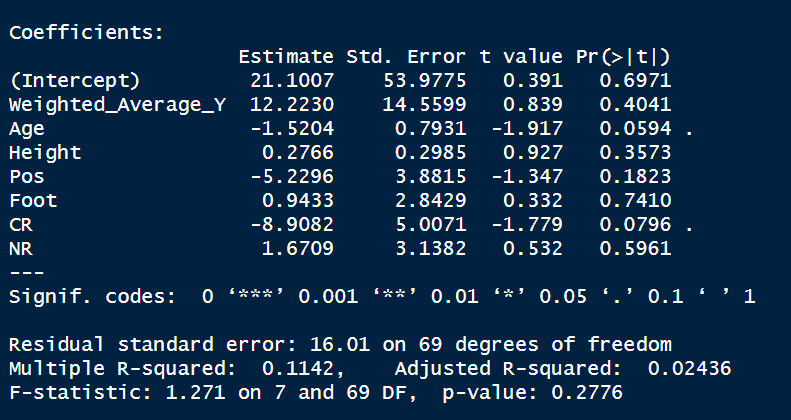
* Generally, a VIF value greater than 4 indicates high multicollinearity, suggesting that the predictor variable may be too highly correlated with other predictor variables in the model.
* VIF values close to 1 indicate low multicollinearity, meaning the predictor variable is not correlated with other predictor variables.

Based on the provided VIF values:

* The variables with higher VIF values, such as 'Y\_5', 'Y\_4', 'Y\_3', 'Y\_2', 'Y\_1', and 'WA', indicate potential multicollinearity issues with these predictors.
* Variables with VIF values closer to 1, such as 'Height', 'Pos', 'Foot', 'CR', and 'NR', suggest lower multicollinearity with other predictors.

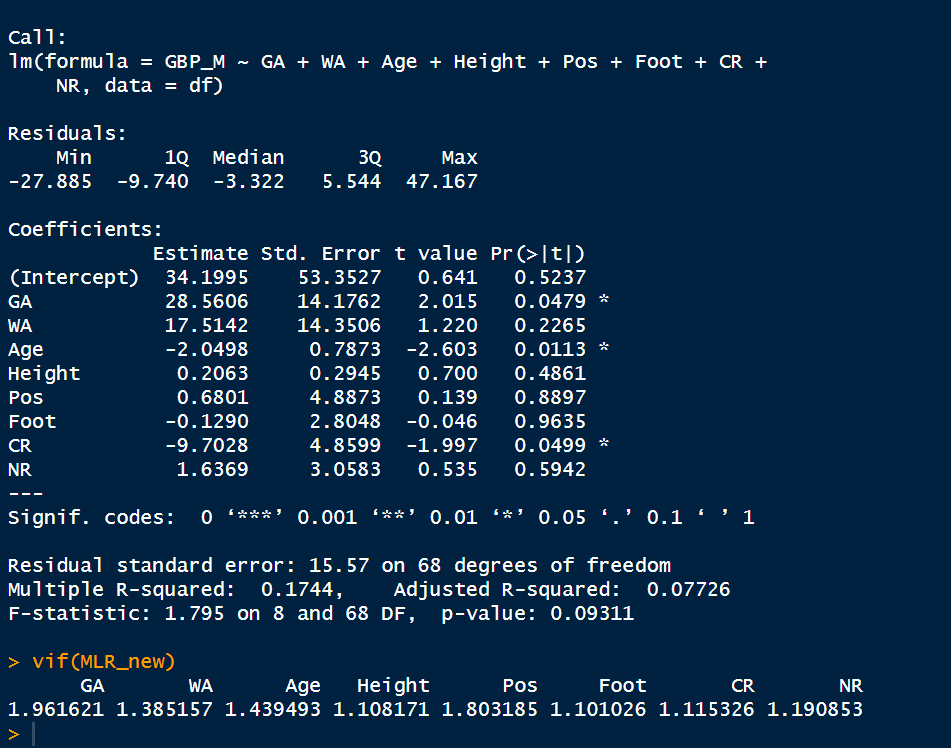


Since Years Y\_1 to Y\_5 are highly correlated , we combine them by giving appropriate weights. Here higher weights is given to Player having five year performance record will have more sustained performance than lesser years.

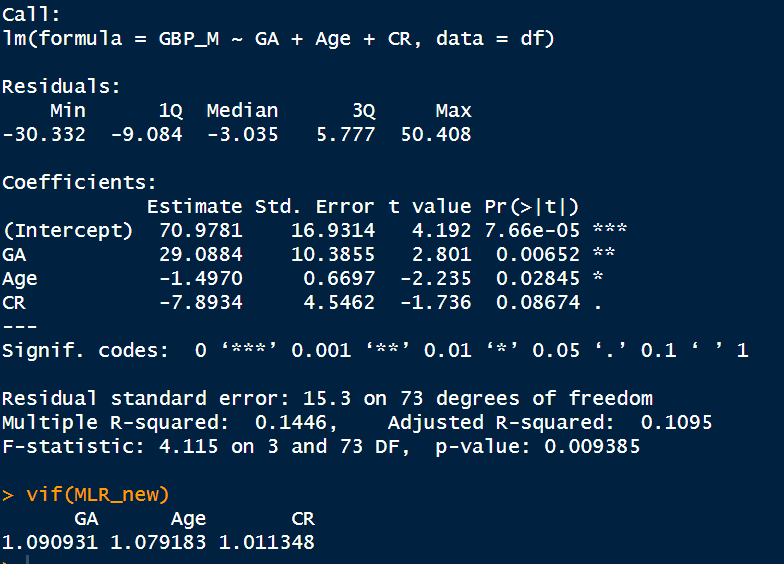


Finally, after removing highly correlated variables, we found that now Age has p value 0.05.

* The Age variable has an estimated coefficient of approximately -1.5204, indicating a negative association with the response variable GBP\_M. Similarly Pos and CR has negative association with the dependent variable - GBP\_M (Player’s transfer price in Pounds) whereas Weighted\_Average\_Y, Height , Foot, NR has positive association with GBP\_M.
* For every one-unit increase in Age, the response variable is expected to decrease by approximately 1.5204 units.
* For every one-unit increase in Height, the response variable is expected to increase by approximately 0.2766 units.
* The p-value for Age is 0.088, suggesting that Age may or may not be statistically significant at the significance level of 0.05. Similarly, the interpretation applies to other variables as well. Foot being least significant at p value 0.760.

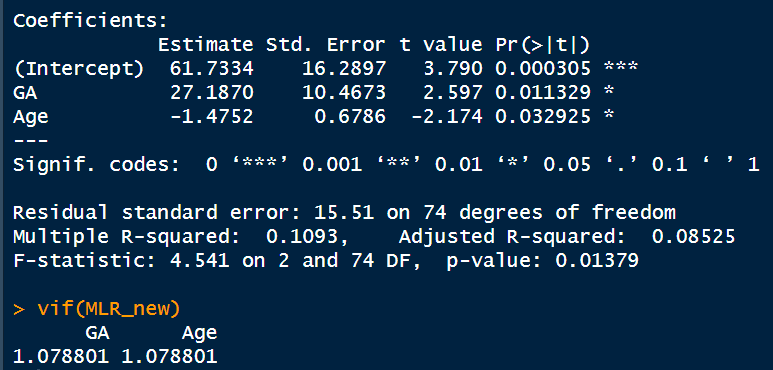


Since GA, Age and CR having p values less than 0.05 , so these IVs were taken for MLR.



CR became less significant as p value is higher than 0.05. Now we consider only GA and Age independent variables for analysis.

FINAL MODEL only 2 IVs were considered GA, Age.



Model is significant with 2 IVs GA and Age having p values less than 0.05,

* The coefficient for "GA" is 27.1870. This means that for each one-unit increase in "GA", the predicted value of the outcome variable increases by 27.1870 units, on average, holding "Age" constant.
* In this case, both "GA" and "Age" have p-values less than 0.05, so they are both statistically significant predictors of the outcome variable.

**The fit of the model:**

* The R-squared value of 0.1093 indicates that the model explains about 10.9% of the variance in the outcome variable. This is a relatively low R-squared value, which suggests that there may be other factors that are also important for predicting the outcome variable.
* The F-statistic of 4.541 and its corresponding p-value of 0.01379 indicate that the model is statistically significant overall, meaning that it is better than a model with no predictors at explaining the outcome variable.

**The vif values:**

* In this case, both vif values are very close to 1, which suggests that there is no significant multicollinearity between "GA" and "Age".

**Overall:**

* This model appears to be statistically significant and shows that both "GA" and "Age" are important predictors of the outcome variable – Transfer Price of players. However, the model only explains a small amount of the variance in the outcome variable, so there may be other important factors that are not included in the model.

***11. What is the final model equation?***

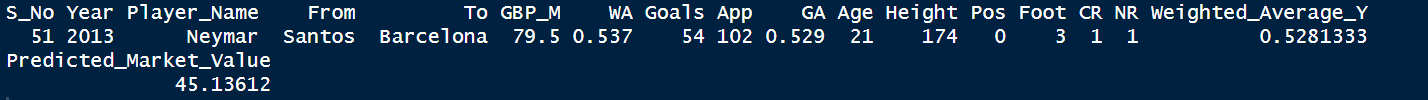
GBP\_M = 61.7334 + 27.1870\*(GA) – 1.4752\*(Age)

***12. What are the findings based on the final model?***

* This model appears to be statistically significant and shows that both "GA" and "Age" are important predictors of the outcome variable – Transfer Price of players. However, the model only explains a small amount of the variance in the outcome variable, so there may be other important factors that are not included in the model.

***13.Choose an offensive soccer player in a top European league (e.g. the Premier League or the La Liga League) and compute his market value using the regression model.***

Ans: Pos is “0” for Offensive player. Predicted market value is found out to be 45.136.



***14. What are your thoughts based on the previous model?***

Ans: Previous model had lots of variables and there was multi-collinearity among variables. It was difficult to analyze how much IVs were influencing DV. P values were lot higher than 0.05 only Age had significant p value. Adjusted R square were very low. Overall model’s p value was very high making the model statistically insignificant.

***15. Discuss the limitations of the model and how the model can be improved.***

**Final Model:**

**Limitations:**

* **Loss of information:** Removing non-significant variables might have excluded important information, even if they were individually not statistically significant. This could lead to biased estimates and reduced model generalizability.
* **Overfitting risk:** Focusing only on significant variables can increase the risk of overfitting, meaning the model might perform well on the training data but not generalize well to unseen data.

**Improvements:**

* **Consider alternative variable selection methods:** While using only significant variables is a common approach, it might not always be optimal. We can explore other methods like stepwise regression with cross-validation or information criteria-based methods for more robust variable selection.
* **Regularization techniques:** Apply regularization techniques like L1 or L2 regularization to penalize model complexity and reduce the risk of overfitting, especially if the number of variables is high relative to the number of observations.
* **Model comparison:** Compare the performance of this model with the original model on an independent test set to see if the trade-off between bias and variance is worthwhile.

**Additional considerations:**

* **Domain knowledge:** Incorporate your domain knowledge when interpreting the results and selecting variables. Statistical significance is not the only factor to consider; the practical meaning and relationships between variables are also crucial.
* **Visualization:** Visualize the relationships between variables and the outcome variable to identify potential non-linearities, outliers, or patterns that the models might not capture.
* **Alternative model types:** Depending on the nature of your data and research question, consider exploring other model types like decision trees, random forests, or gradient boosting machines, which might be better suited for capturing complex relationships or handling non-linear data.