Введение

В настоящее время проблема определения результатов футбольных команд довольно актуальна. Всем любителям футбола интересно: почему в конце сезона складывается то или иное положение команд в турнирной таблице ? Каждый футбольный сезон анализируется большим количеством футбольных экспертов, тренеров, а также и экономистов. Ведь и им, интересно же, от чего зависел успех команд в пройденном сезоне? Поэтому аналитики и ищут объяснение победам одних и поражениям других команд.

Вдохновило меня провести исследование на данную тему феноменальное выступление двух команд: Лестера в английской футбольной премьер- лиге и Ростова в российской футбольной премьер- лиге.

Перед исследованием выдвинем гипотезу о том, что ключевые факторы, влияющие на результаты команд в этом сезоне, это количество забитых голов и трансферная политика клубов.

Целью данной исследовательской работы является анализ результатов футбольных команд шести ведущих европейских лиг в сезоне 2015/2016.

Глава 1. Обзор литературы.

На данную тему написано очень много книг и статей, приведём некоторые из них.

В книге Саймона Купера и Стефана Шимански "Футболономика" приводятся исследования, в которых факторами, определяющими результаты футбольных сборных, были численность населения страны, уровень национального дохода и опыт матчей на крупных международных турнирах.

В качестве результатов (зависимой переменной) исследователями была взята разница между забитыми и пропущенными голами. Набор данных по которому проводился анализ, содержал в себе информацию о 189 странах.

В результате проведённого исследования было выяснено, что данные три фактора в совокупности лишь чуть больше чем на четверть объясняют разницу между забитыми и пропущенными голами. Тем не менее сам факт, что эти три показателя способны в такой степени объяснять исход матчей, говорит, что до определённой степени футбол рационален и прогнозируем.

В статье Вячеслава Данилова "А побеждают всегда..." в журнале "Экономическая политика" предпринята попытка ревизии эконометрических подходов предсказания результатов выступления футбольных сборных на чемпионатах мира. По мнению Данилова, нужно включить в такие модели помимо экономических, демографических, социокультурных и спортивных параметров, факторы, учитывающие эффект публичной политики, в частности наличие в стране комплексной программы подготовки молодых футболистов и готовность футбольной федерации натурализовывать игроков.

Также предлагается учитывать повышение вероятности победы на чемпионате мира команды, которая уже хотя бы однажды его выигрывала, и фактор команды- династии, который состоит в том, что сборная, выигравшая крупный международный турнир, с высокой долей вероятности побеждает и на следующем турнире.

По итогам проведённого исследования автор статьи делает следующие выводы:

1) Подушевой ВВП непосредственно отражается на спортивных результатах.

Он играет роль комплексного показателя развития спортивно материальной и организационной инфраструктуры, а также системы поиска и воспитания молодых спортсменов.

Зависимость показателей национальных сборных от ВВП имеет вид квадратичной функции: по мере роста ВВП показатели страны растут, но до определённой точки, после которой график функции опускается вниз.

2) Чем больше населения проживает в стране, тем большее число молодых талантов национальные федерации могут привлечь к занятиям футболом и у сборной такой страны существенно больше шансов на победу. Здесь, правда, многое зависит от культурных традиций страны и региона в целом. Поэтому в данной модели переменная, учитывающая численность населения, уравновешивается переменной, отсылающей к особенностям спортивной культуры страны и её футбольным традициям.

3) Наилучших результатов добиваются футбольные сборные стран с умеренным климатом. При прочих равных, среднегодовая дневная температура в столице страны около 14 °C имеет позитивную корреляцию со спортивными результатами. Именно такая среднегодовая дневная температура обеспечивает наилучшие условия для занятия футболом.

4) Фактор своего поля является немаловажным при прогнозировании итогов соревнований. Давление этого фактора в последнее время снижается из-за расширения футбольной географии.

5) Шанс выиграть крупный турнир возрастает после первого выигрыша.

6) Шанс выиграть следующий турнир у команды, только что завоевавшей титул, выше, чем у остальных участников турнира.

Глава 2.Описание исходных данных.

2.1 Для исследования собраны данные по результатам футбольных лиг 6 стран в сезоне 2015\2016: Англии, Испании, Италии, Германии, Франции и России. Переменные, использованные в этом наборе данных, представлены в таблице 1.1.

Таблица 1.1-Описание исходных данных.

Код переменной	Описание переменной	Тип переменной	Единицы измерения	Комментарий
M	Место в турнирной таблице	Количественная	Место	Определяет результат команды, зависимая переменная
NG	Количество матчей	Количественная	шт.	
Victory	Количество побед	Количественная	шт.	

Goals	Количество забитых голов	Количественная	шт.	
DG	Разница голов	Количественная	г.	Разница между забитыми и пропущенными голами
Star	Звёздные игроки	Качественная	чел.	0-"Нет звёздных игроков"; 1-"Одна, максимум две звезды"; 2-"Звёздная команда"
Transfers	Трансферные приобретения	Количественная	млн.евро	Бюджет, выделенный на трансферы

2.2 Изучение исходных данных.

Найдём значения описательных статистик по каждой из переменной нашего набора данных.

Найдём значения описательных статистик в RStudio с помощью функции summary().

```
summary(Futball$M)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    5.00   10.00   10.06   15.00   20.00
```

В исследуемых чемпионатах количество команд составляет: 20,18,16 команд. Исходя из этого, зависимая переменная - место в турнирной таблице имеет следующие показатели: Самое лучшее выступление, это первое место, в свою очередь, самое худшее 20 место. В зависимости от количества команд, худшим выступлением в лиге может быть и 18, и 16 место.

```
summary(Futball$Ng)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  30.00   34.00   38.00   36.25   38.00   38.00
```

Минимальное количество матчей, сыгранных за сезон командами в исследуемых чемпионатах, равно 30. Максимальное, в свою очередь, 38 матчей. Больше чем половина

команд, а именно 75 %, провели в этом сезоне от 34 до 38 матчей. Четверть команд сыграли от 30 до 34 матчей.

Среднее количество матчей, проведённых данными командами, равняется 36.25 матча. Так как в большинстве представленных лиг команды сыграли за сезон 38 матчей, значение медианы по данной переменной равняется 38. В силу того что, медиана больше чем среднее значение по данной переменной, можно сделать вывод о том, что распределение переменной количества матчей имеет отрицательную асимметрию.

```
summary(Futball$Victory)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   3.00   10.00   12.00   13.39   16.75   30.00 
```

Минимальное количество побед, одержанных командами в этом сезоне, равняется 3 победы. В таком плане «отличились» следующие команды:

«Астон Вилла» и «Труа». У них совсем не получился данный сезон, по его итогам они отправились на дивизион ниже.

Наибольшее количество побед, а именно 30, одержал ФК «ПСЖ», который выиграл Чемпионат Франции за явным преимуществом.

Четверть команд одержало от 3 до 10 побед, половина команд от 10 до 16.75 побед, ещё четверть от 16.75 до 30 побед. Среднее количество побед, одержанных командами, равняется 13.39, медиана по данной переменной 12 побед, то есть наибольшее количество команд одержало по 12 побед, что довольно средний показатель по итогам сезона. Так как медиана меньше чем среднее значение по данной переменной, то распределение этой переменной имеет положительную асимметрию.

```
summary(Futball$Goals)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  19.00   38.00   45.00   47.94   52.00  112.00 
```

Минимальное количество голов, забитых командами, равно 19. Так мало забил ФК «Крылья Советов». Максимальное количество голов, а именно 112, забила «Барселона», очень близко к ней оказался и другой испанский гранд «Реал Мадрид», на его счету 110 забитых голов. Четверть команд забило от 19 до 38 мячей за весь сезон, что конечно, низкий показатель, в среднем меньше чем гол за игру. Половина команд забила от 38 до 52 голов, ещё четверть команд от 52 до 112 голов. Среднее количество голов в этом сезоне у одной команды равняется 47.94 гола, наибольшее количество команд забило по 45 голов. Так как медиана данной переменной меньше чем среднее значение, то распределение этой переменной имеет положительную асимметрию.

```
summary(Futball$GD)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-55.00  -14.75   -6.00    0.00   11.50   83.00 
```

Наихудшая разница мячей минус 55 у французской команды «Труа», что ещё раз доказывает, что эта команда справедливо оказалась на последнем месте в турнирной таблице и покинула главную лигу Франции. Наилучшая разница мячей плюс 83 мяча у команды, которая забила наибольшее количество голов, у ФК «Барселона». В среднем разница мячей у команд расположилась в диапазоне от -14.75 гола до 11.50 гола. Медиана данной переменной равна -6, среднее значение 0. Так как медиана меньше чем среднее значение, то распределение данной переменной имеет положительную асимметрию.

```
summary(Futball$Star)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  0.0000  0.2719  0.0000  2.0000 
```

По данной качественной переменной можно заметить, что меньше чем в половине команд присутствуют звёздные игроки. Основная масса команд всё-таки обладает равным составом, без мега ярких футболистов.

```
summary(Futball$Transfers)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   2.075  11.400  29.210  43.900 210.500 
```

По данной переменной можно заметить, что есть команды, которые в силу разных причин, не приобрели ни одного футболиста. И, наоборот, есть команды, которые очень постарались на трансферном рынке: наибольшее количество 210,5 млн евро потратил ФК «Манчестер Сити», также выделяются ФК «Манчестер Юнайтед», который выделил на трансферы 149,5 млн евро, и ФК «Ювентус» с выделенными на покупки 129,150 млн.евро. В среднем команды потратили на трансферы 29,21 млн. евро.

Медиана по данной переменной равняется 11.4 млн.евро. Так как медиана меньше среднего значения, распределение данной переменной имеет положительную асимметрию.

Чтобы визуально убедиться в описательных статистиках переменных, построим различные диаграммы и объясним их.

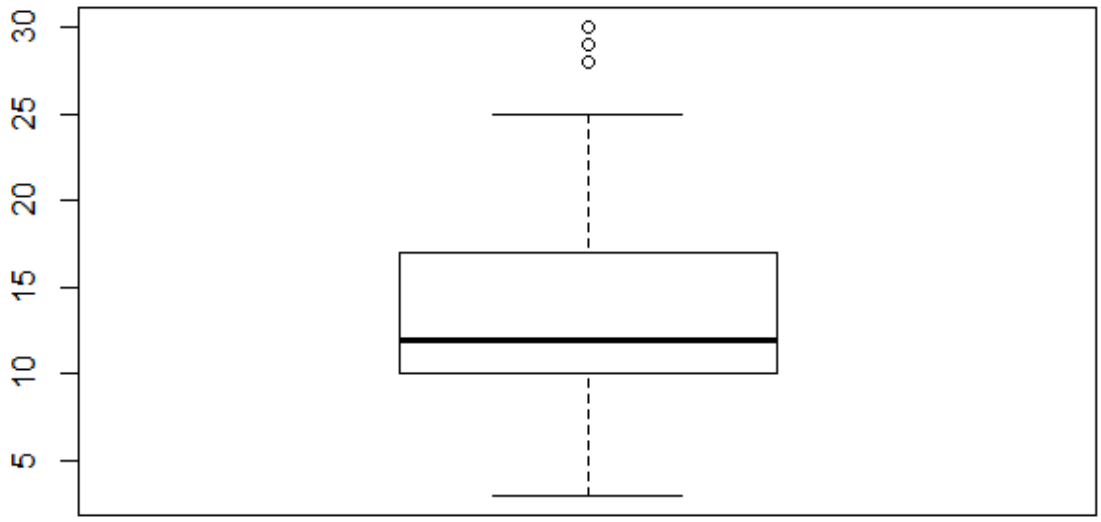

Рисунок 1-Диаграмма ящик с усами по количеству побед.

На рисунке 1 представлена диаграмма ящик с усами по переменной количество побед. На нём видно, что присутствуют три выброса, то есть несколько команд, имеют в своём активе такое количество побед, которое намного больше чем у других команд. Также на данной диаграмме подтверждается выявленная при изучении описательных статистик положительная асимметрия, то есть большее значение количества побед лежит выше медианы.

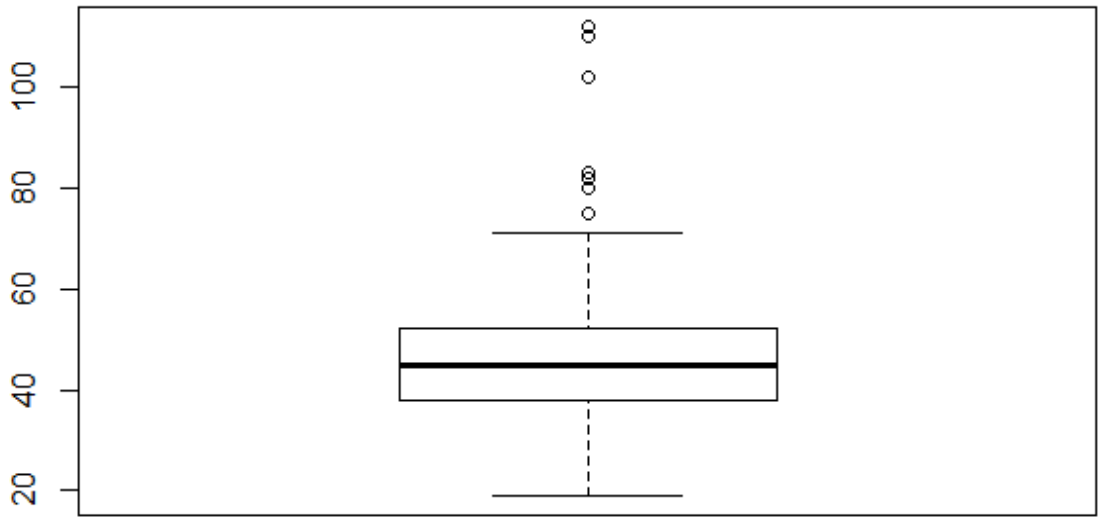

Рисунок 2-Диаграмма ящик с усами по количеству забитых голов.

На данной диаграмме можно заметить, что присутствуют выбросы, то есть количество забитых голов командами в этом сезоне распределено довольно неравномерно, есть команды которые забили намного больше, чем все в среднем. Также заметно, что распределение данной переменной имеет небольшую положительную асимметрию.

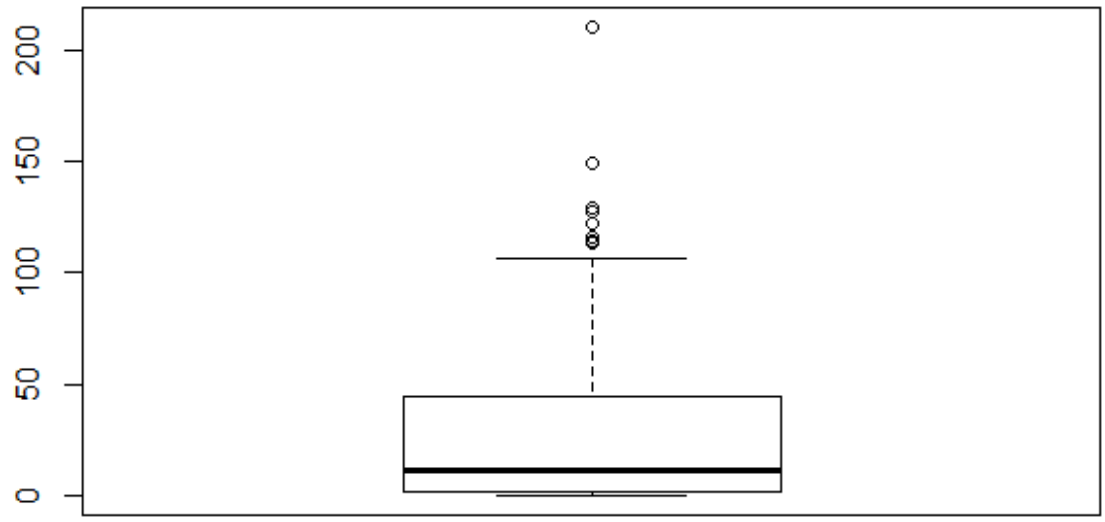

Рисунок 3-Диаграмма ящик с усами по трансферным затратам.

На диаграмме по затратам на трансферы видно, что распределение данной переменной имеет ярко выраженную положительную асимметрию, почти все значения лежат выше медианы. Также присутствуют выбросы, самый яркий это затраты на трансферы ФК «Манчестер Сити» в размере 210.5 млн.евро, который расположен сверху. Интересно, что значение минимума по данной переменной очень близко лежит с квантилью первого уровня. Это означает, что четверть всех значений лежит в очень маленьком диапазоне, от 0 до 2.075 млн.евро.

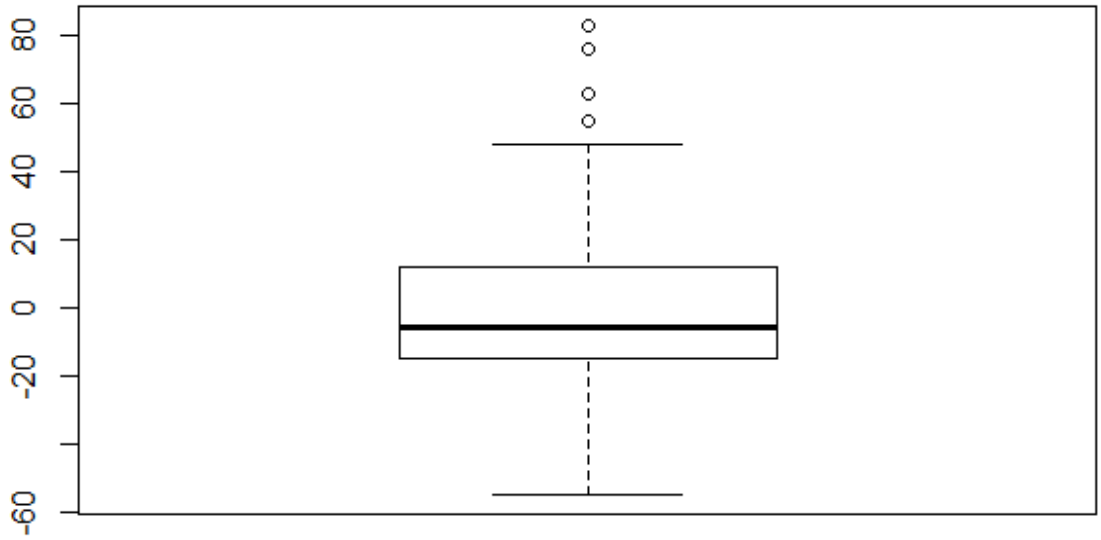

Рисунок 4-Диаграмма ящик с усами по разнице мячей.

По данной диаграмме можно отметить, что присутствуют команды, которые много забили голов в ворота соперников, и при этом довольно хорошо сыграли в обороне. Также видно, что распределение данной переменной скошено вправо, то есть присутствует положительная асимметрия.

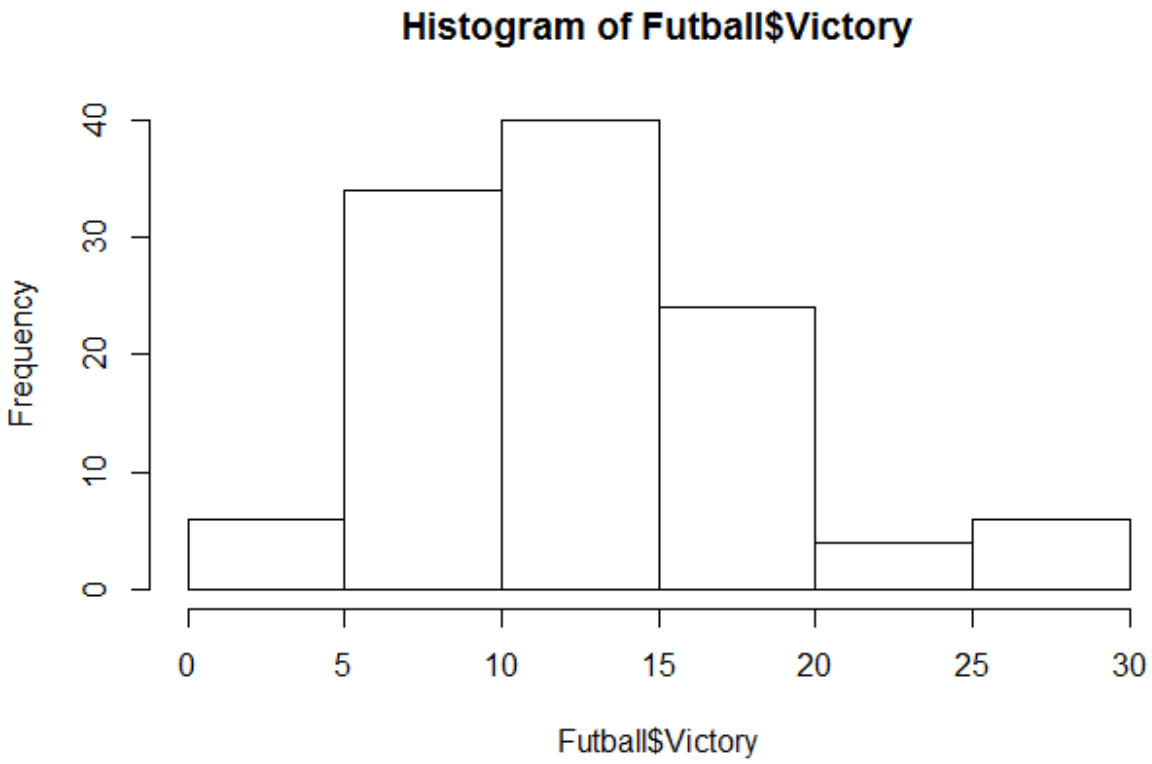

Рисунок 5-Гистограмма по количеству побед.

На данной гистограмме заметно, что распределение переменной количества побед очень близко к нормальному распределению.

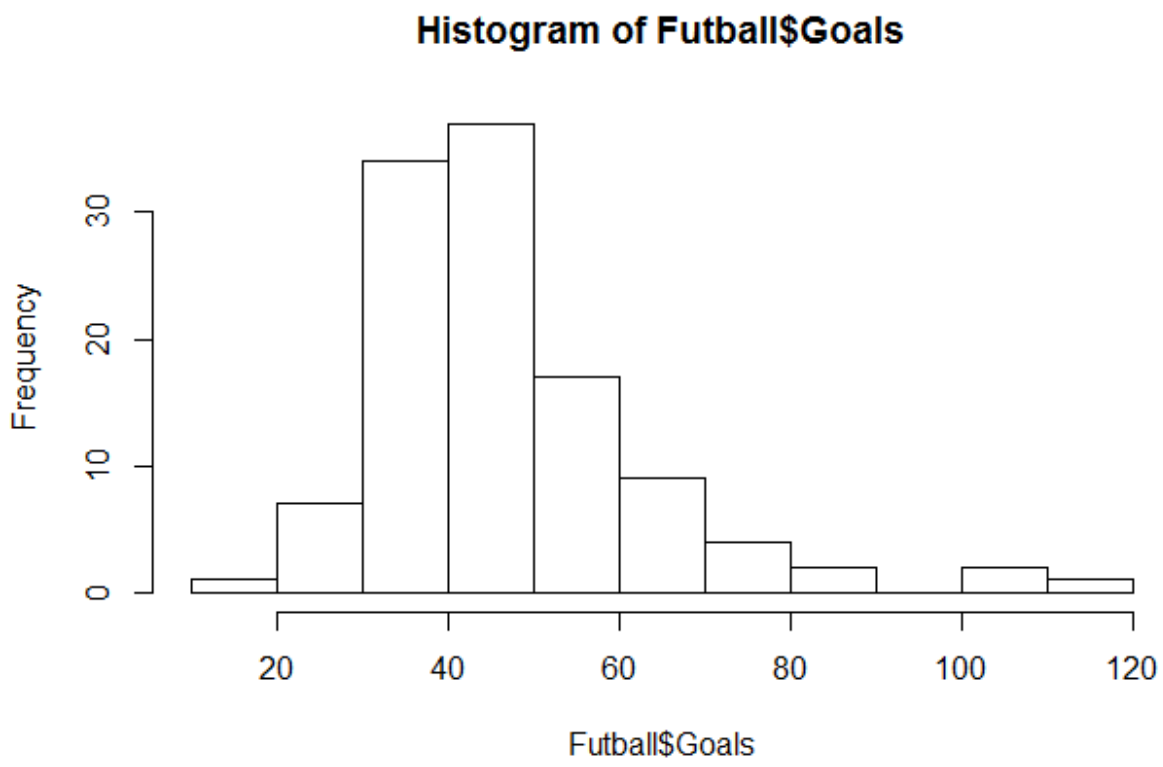

Рисунок 6-Гистограмма по количеству забитых голов.

На рисунке 6 видно, что количество забитых голов всеми командами относительно среднего значения расположено неравномерно. Это происходит, потому что большинство команд отличилось больше среднего показателя по данной переменной.

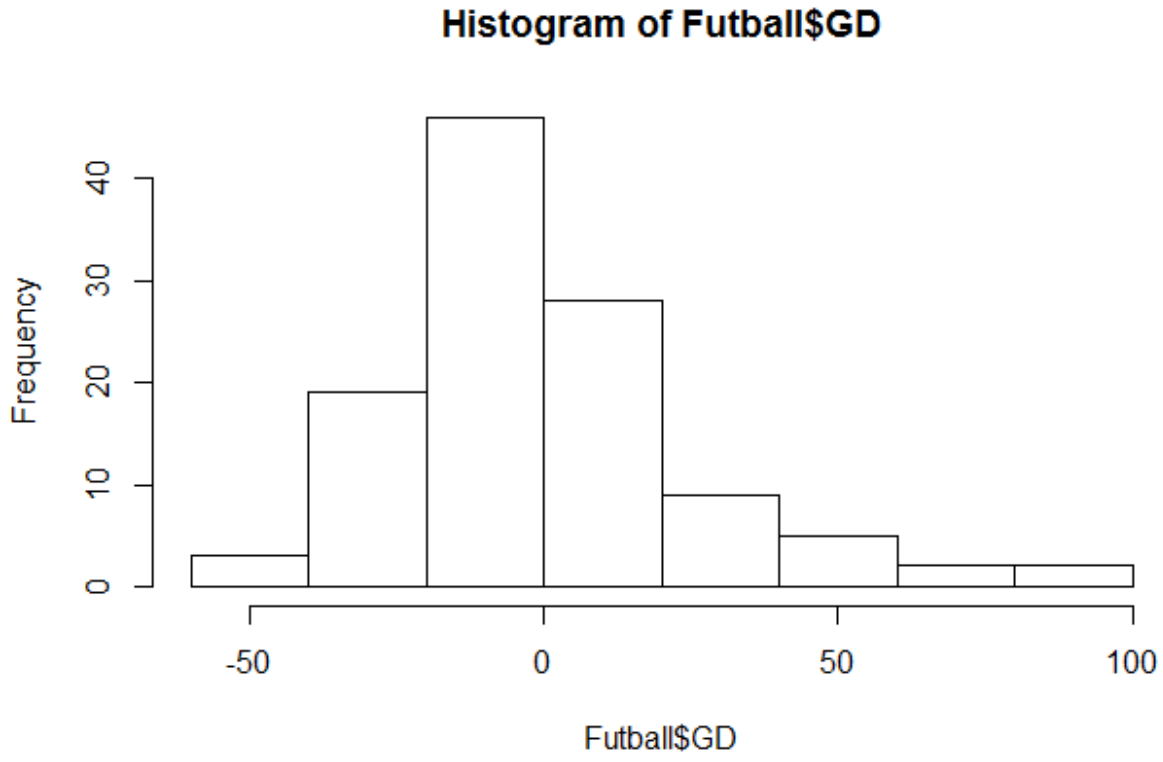

Рисунок 7-Гистограмма по разнице мячей.

На гистограмме по данной переменной можно заметить, что всё-таки большее количество рассматриваемых команд имеет положительную разницу между забитыми и пропущенными мячами.

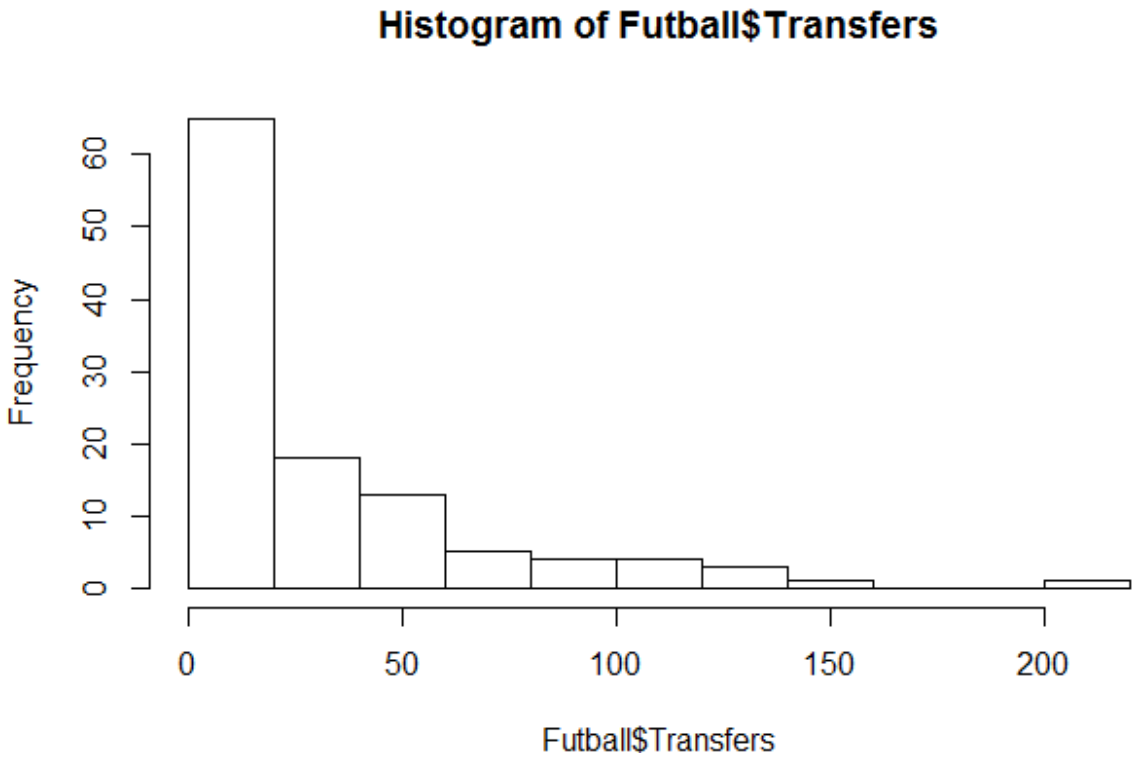

Рисунок 8-Гистограмма по затратам на трансферы.

На гистограмме по данной переменной подтверждается, то, что её распределение имеет положительную асимметрию.

Теперь исследуем парные взаимосвязи между переменными:

а) Корреляционная матрица.

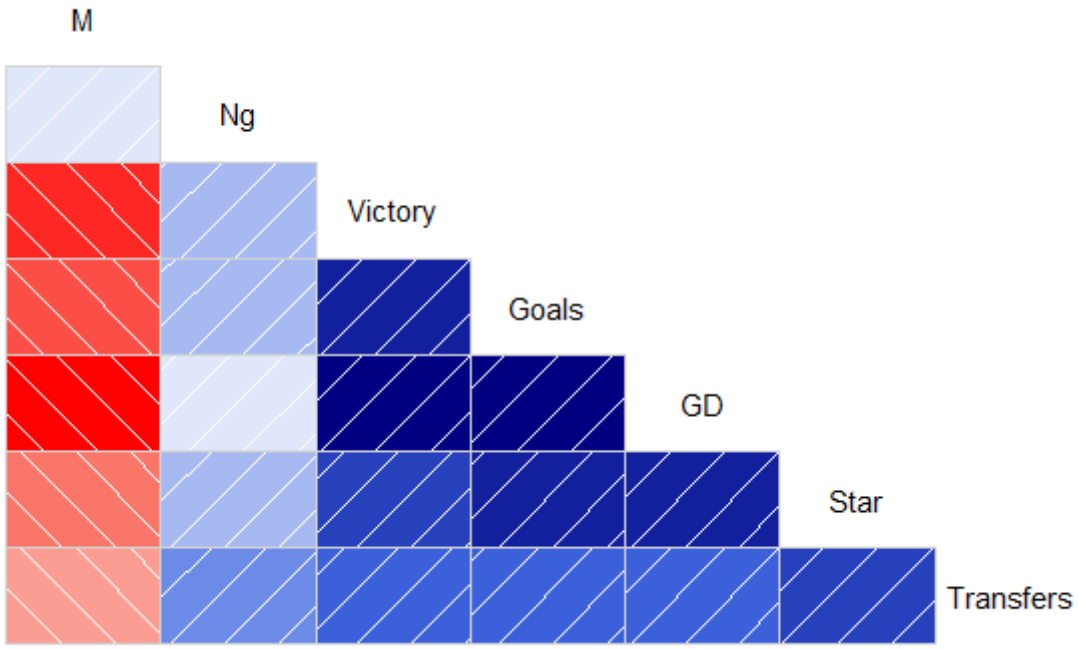

Рисунок 9-Корреляционная матрица.

Сразу выделяется сильная положительная корреляция между переменными Victory и GD, Victory и Goals, то есть между количеством побед и разницей забитых и пропущенных мячей, а также между количеством побед и количеством забитых голов командами за этот сезон. Ещё видна сильная прямая взаимосвязь между количеством забитых голов и разницей голов.

б) Матрица диаграмм рассеивания.

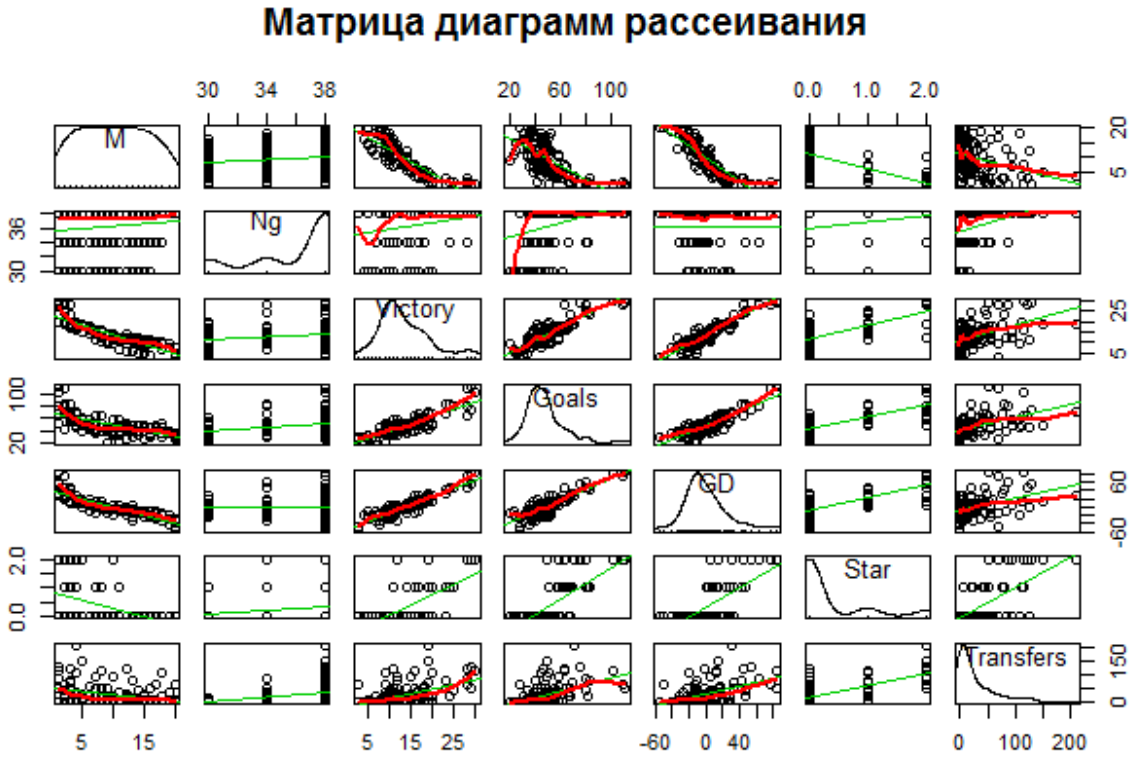

Рисунок 10-Матрица диаграмм рассеивания.

На диаграммы рассеяния тут наложены аппроксимирующие линии и сглаженные кривые, а по главной диагонали изображены диаграммы ядерной оценки функции плотности.

В основном распределение переменных имеет унимодальный характер, но присутствует и бимодальное распределение.

в) Визуальный анализ данных с помощью показателя VIF.

Мультиколлинеарность можно выявить при помощи статистики, называемой фактором инфляции дисперсии. Квадратный корень, извлеченный из этой статистики для любой независимой переменной, указывает на степень увеличения доверительного интервала для параметра регрессии данной переменной по сравнению с моделью без скоррелированных независимых переменных.

```
vif(fit)
       Ng   Victory     Goals        GD      Star Transfers
 1.843658  9.336752  6.139752 13.984731  3.322991  2.008672

sqrt(vif(fit))>2
       Ng   Victory     Goals        GD      Star Transfers
    FALSE      TRUE      TRUE      TRUE     FALSE     FALSE
```

Из этой модели видно, что присутствует проблема мультиколлинеарности, поэтому необходимо убрать переменную Goals, так как она сильно коррелирует с переменной GD и Victory.

```
vif(fit1)
       Ng   Victory        GD      Star Transfers
 1.445956  9.333687 10.243062  3.111141  1.989812

 sqrt(vif(fit1))>2
       Ng   Victory        GD      Star Transfers
    FALSE      TRUE      TRUE     FALSE     FALSE
```

В данной модели также присутствует мультиколлинеарность, поэтому уберём из модели переменную GD, которая коррелирует с переменной Victory.

```
vif(fit2)
       Ng   Victory      Star Transfers
 1.107222  1.968575  2.774347  1.980779

 sqrt(vif(fit2))>2
       Ng   Victory      Star Transfers
    FALSE     FALSE     FALSE     FALSE
```

Получив модель fit2, мы избавились от проблемы мультиколлинеарности.

Глава 3

3.1 ПОСТРОЕНИЕ И ИНФЕРЕНЦИЯ О МОДЕЛИ РЕГРЕССИИ.

Представим далее оценку модели линейной регрессии и выводы по построенной модели.

```
fit1<-lm(M~Ng+Victory+Goals+GD+Star+Transfers,data=Futball)
> summary(fit1)

Call:
lm(formula = M ~ Ng + Victory + Goals + GD + Star + Transfers,
    data = Futball)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9517 -1.5018 -0.1717  1.6169  5.0893

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.480054   2.717663   0.913 0.363521
Ng           0.332104   0.097415   3.409 0.000920 ***
Victory     -0.748057   0.111927  -6.683 1.09e-09 ***
Goals        0.111354   0.032036   3.476 0.000737 ***
GD          -0.115621   0.031461  -3.675 0.000373 ***
Star         1.191244   0.636554   1.871 0.064022 .
Transfers   -0.003579   0.007554  -0.474 0.636576
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.224 on 107 degrees of freedom
Multiple R-squared:  0.8507,   Adjusted R-squared:  0.8423
F-statistic: 101.6 on 6 and 107 DF,  p-value: < 2.2e-16
```

Как видно, получилась модель с незначимой переменной Transfers, исключим её из модели и построим новую регрессию.

```
fit1<-lm(M~Ng+Victory+Goals+GD+Star,data=Futball)
> summary(fit1)

Call:
lm(formula = M ~ Ng + Victory + Goals + GD + Star, data = Futball)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8750 -1.5853 -0.1884  1.6351  5.0732

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.73646    2.65367   1.031 0.304750
Ng           0.32263    0.09500   3.396 0.000957 ***
Victory     -0.75133    0.11131  -6.750 7.67e-10 ***
Goals        0.11282    0.03177   3.551 0.000570 ***
GD          -0.11551    0.03135  -3.685 0.000359 ***
Star         1.02901    0.53471   1.924 0.056932 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.216 on 108 degrees of freedom
Multiple R-squared:  0.8503,   Adjusted R-squared:  0.8434
F-statistic: 122.7 on 5 and 108 DF,  p-value: < 2.2e-16
```

По отношению к прошлой модели в полученной все переменные значимы, поэтому можно сделать вывод о том, что модель улучшена.

Из полученных результатов следует, что уравнение регрессии имеет следующий вид:

$$M = 2.73646 + 0.32263 * Ng - 0.75133 * Victory + 0.11282 * Goals - 0.11551 * GD + 1.02901 * Star$$

По полученному уравнению можно сделать следующие выводы:

Свободный член, равный 2.73646, показывает прогнозируемый уровень зависимой переменной (место команды в турнирной таблице), если значения объясняющих переменных при этом равны 0.Если же значения объясняющих переменных находятся далеко от 0, то данная интерпретация не будет иметь место.

Регрессионный коэффициент перед переменной Ng (количество матчей) равен 0.32263. Это означает, что при увеличении количества сыгранных игр в сезоне на одну игру, место занятое в турнирной таблице увеличивается на 0.32263. Данный коэффициент статистически значим, так как (Pr>|t|) = 0.000957 , меньше чем 0.05.

Регрессионный коэффициент перед переменной Victory (победы) равен -0.75133. Это означает, что при увеличении количества побед на одну, место в турнирной таблице уменьшается на 0.75133. Данная обратная зависимость показывает, что чем больше побед было у команды в сезоне, тем выше она была в турнирной таблице, а значит и результат был лучше. Этот коэффициент статистически значим на уровне 0.001, так как (Pr>|t|)= 0.00000767.

Регрессионный коэффициент перед переменной Goals (количество забитых голов) равен 0.11282. Это означает, что при увеличении забитых голов за сезон на один, место в турнирной таблице увеличивается на 0.11282. Данный коэффициент статистически значим на уровне 0.001, так как Pr(>|t|)=0.000570.

Регрессионный коэффициент перед переменной GD (разница мячей) равен -0.11551. Значит, что при увеличении разницы на 1 мяч, итоговый результат команды улучшается на 0.11551. Данный коэффициент статистически значим на уровне 0.001, так как (Pr>|t|)= 0.000359.

Регрессионный коэффициент перед переменной Star равен 1.02901. Это означает, что если в команде присутствуют звёздные игроки, то итоговый результат команды улучшается на 1.02901 места в турнирной таблице. Данный коэффициент статистически значим на уровне 0.1, так как (Pr>|t|)=0.056932.

Множественный коэффициент детерминации (Multiple R-squared: 0.8503) означает, что наша модель объясняет 85,03 % дисперсии значений результатов команд в прошедшем сезоне.

Скорректированный коэффициент детерминации (Adjusted R-squared) равен 0.8434.

Стандартная ошибка остатков (Residual standart error) равна 2.216. Она означает усреднённую ошибку предсказания результатов команд по итогам сезона 2015/16 с использованием данной модели.

3.2 ПРОВЕРКА ДАННЫХ НА НАЛИЧИЕ НЕОБЫЧНЫХ НАБЛЮДЕНИЙ

а) Проверка наличия выбросов. Выброс – это значение, которое плохо предсказывается подобранной моделью (то есть имеет большой положительный или отрицательный остаток).

Построим графики остатков.

```
> rstud<-rstandard(fit1)
> rjack<-rstudent(fit1)
> par(mfrow=c(2,2))
>  plot(fit1$res,ylab="raw residuals")
>  plot(rstud,ylab="studentized residuals")
>  plot(rjack,ylab="jacknife residuals")
```

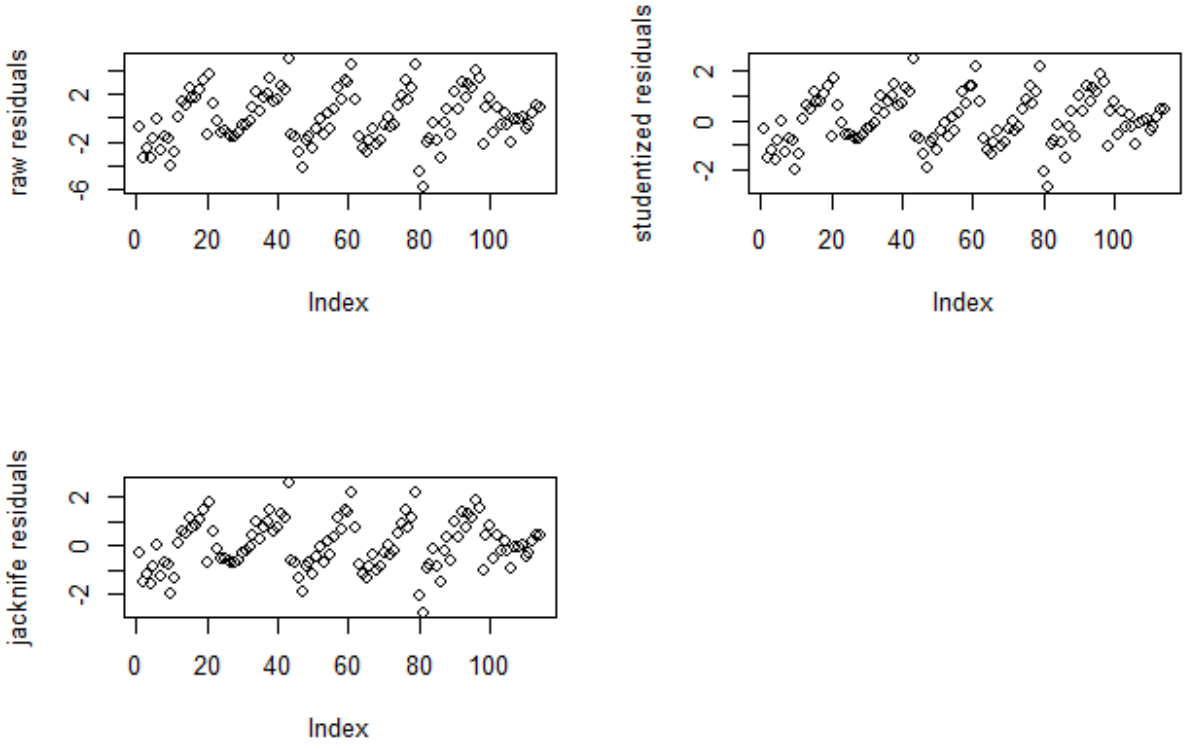

Рисунок 11-«Стандартные и Стьюдентизированные остатки по методу складного ножа от наблюдения»

Стьюдентизированные остатки представляют собой частное от деления обычного остатка на оценку его стандартного отклонения. На графике заметно, что особых выбросов нет, все наблюдения расположены довольно близко друг с другом.

Степень уверенности в результатах зависит от степени соответствия данных допущениям, лежащим в основе статистических тестов. Данные не очень хорошо

укладываются в 95% доверительные границы, это значит, что требование нормального распределения выполняется недостаточно хорошо.

На диаграмме qqPlot видно, что в основном все значения находятся близко к линии уравнения регрессии. Есть несколько значений, которые достаточно отдалены от этой линии и выходят за пределы доверительных границ.

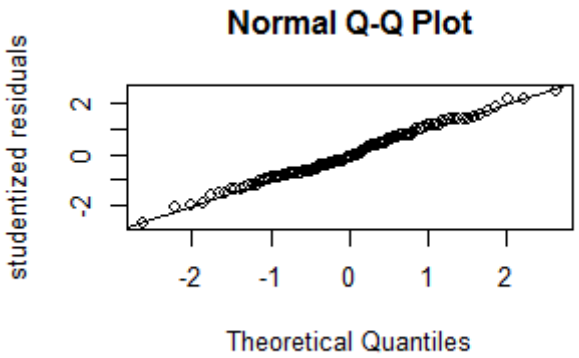

Рисунок 12-«Диаграмма Q-Q plot»

Дальше вычислим значение вероятности статистической ошибки первого рода с поправкой Бонферрони для наибольших остатков Стьюдента:

```
> library(car)
> fit<-lm(M~Ng+Victory+Goals+GD+Star,data=Futball)
> outlierTest(fit)

No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
81 -2.780988          0.0064054      0.73022
```

Функция outlierTest проверяет на значимость самый большой выброс в указанной модели методом Бонферрони, в котором двусторонняя вероятность нулевой гипотезы умножается на размер выборки. Смысл процедуры Бонферрони в том, что по теории вероятности мы должны ожидать какое-то количество выбросов, например, 5% из 100%; чем больше выборка, тем больше будет выбросов. Если у нас всего одно наблюдение и вероятность выброса 5%, то при 2-х наблюдениях вероятность выброса уже 10% и т.д. Вероятность по Бонферрони поэтому правильнее. Как видно тест по нашей модели предсказал, что нет таких значений стьюдентизированных остатков для которых значение вероятности Бонферрони меньше чем 0.05. Значение вероятности в данном случае у нас равно 0.73 (больше чем 0,05) , поэтому она не является значимой и выбросов не имеется.

б) Проверка наличия влиятельных наблюдений.

Существуют два метода обнаружения влиятельных наблюдений: расстояние Кука (или D-статистика) и диаграммы добавленных переменных. Можно сказать, что значения

расстояния Кука, превышающие 4/(n – k – 1), где n – объем выборки, а k – число независимых переменных, свидетельствуют о влиятельных наблюдениях. Построить диаграмму расстояний Кука можно при помощи следующего программного кода:

```
> library(car)
> fit<-lm(M~Ng+Victory+Goals+GD+Star,data=Futball)
> cutoff <- 4/(nrow(Futball)-length(fit1$coefficients)-2)
> plot(fit, which=4, cook.levels=cutoff)
> abline(h=cutoff, lty=2, col="green")
```

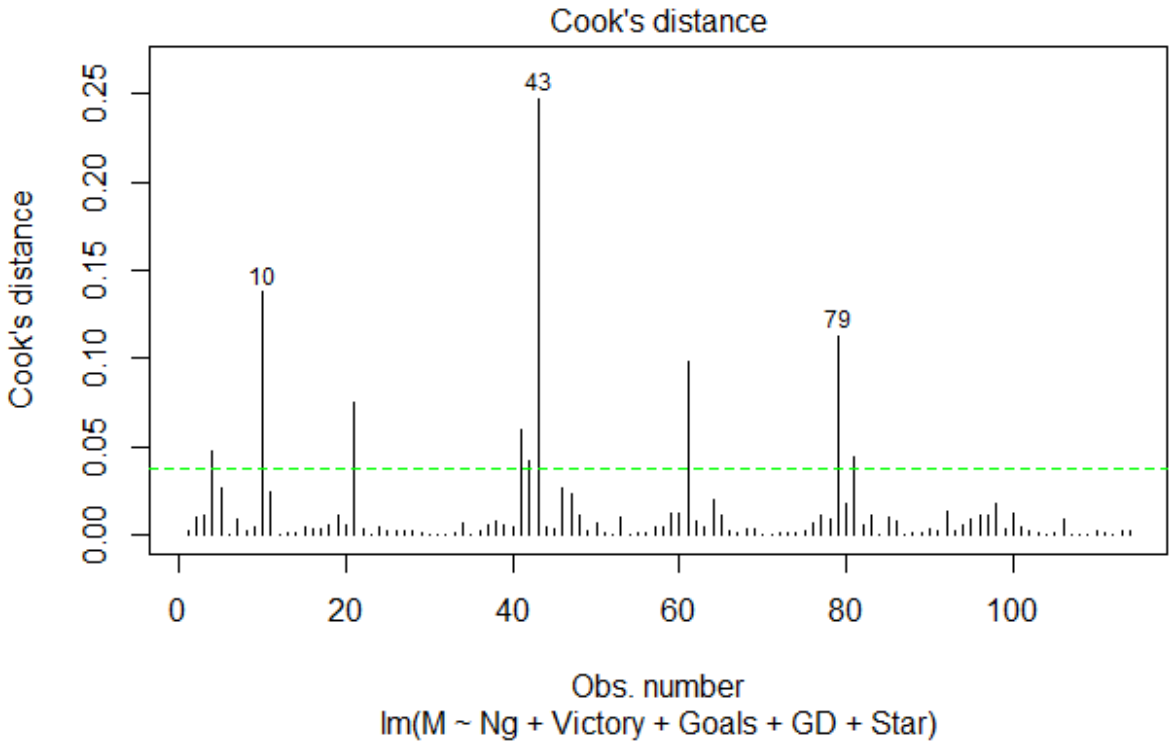

Рисунок 13- «Диаграмма расстояний Кука»

На рисунке 13 заметно, что сильно превышают линию 10, 43 и 79 значения. Это влиятельные наблюдения. Удаление данных наблюдений из модели заметно влияет на значение свободного члена и угловых коэффициентов в регрессионной модели.

Диаграммы расстояний Кука позволяют обнаружить влиятельные наблюдения, но при этом они не позволяют понять, как эти наблюдения влияют на модель. Определить это помогают диаграммы добавленных переменных. Для одной зависимой и k независимых переменных описанным ниже способом создается k диаграмм добавленных переменных.

Для каждой независимой переменной Xk отображаются остатки от регрессии зависимой переменной по остальным k – 1 независимым переменным. Такие диаграммы добавленных переменных можно построить при помощи функции avPlots() из пакета car:

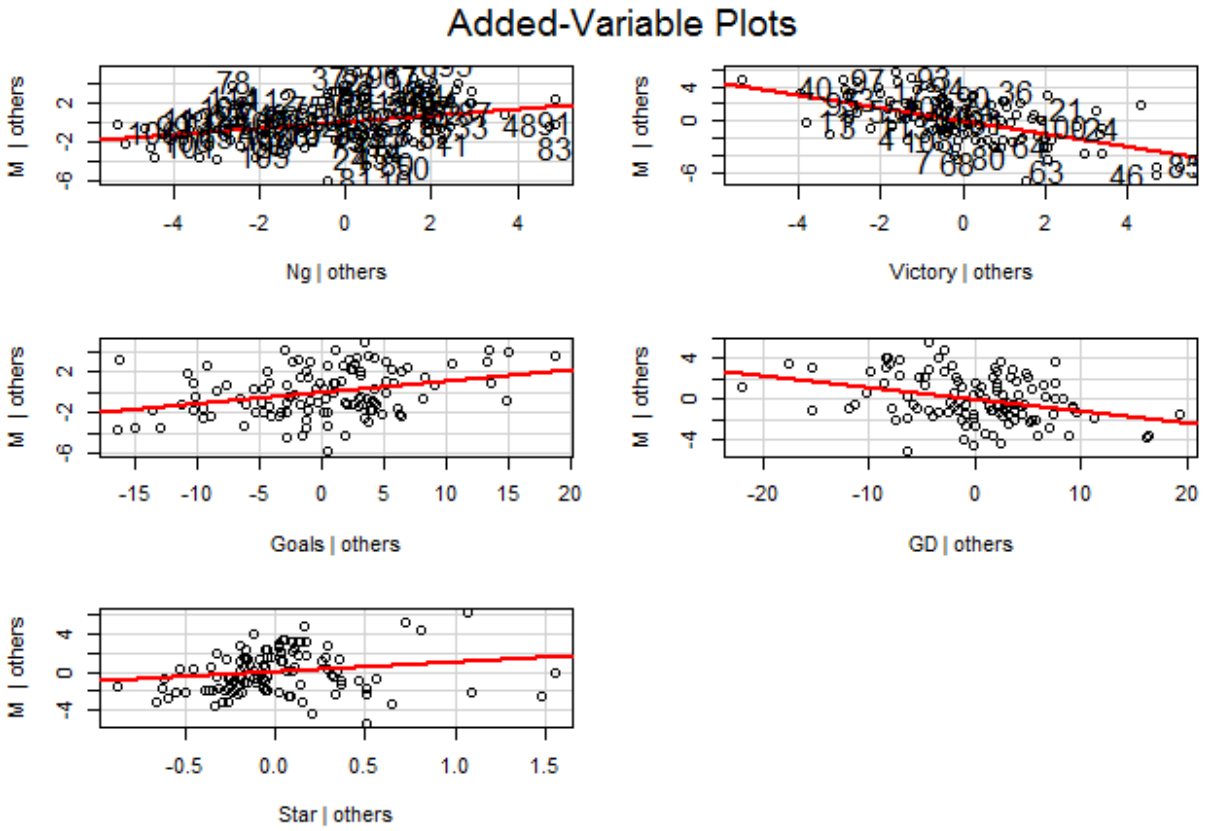

Рисунок 14- «Диаграмма добавленных переменных»

Прямая на каждой диаграмме – это регрессионный коэффициент для данной независимой переменной. Вклад влиятельных наблюдений можно оценить, если представить, как изменится линия, если удалить точку, соответствующую данному наблюдению.

в) Проверка наличия точек высокой напряжённости.

Точки с высокой напряженностью – это выбросы в отношении других независимых переменных. Они характеризуются необычным сочетанием значений независимых переменных. Значение зависимой переменной не используется при вычислении напряженности.

Наблюдения с высокой напряженностью идентифицируются при помощи показателя влияния (hatstatistic). Для определенного набора данных среднее значение этой статистики вычисляется как p/n, где p –это число параметров в модели (включая свободный член), а n –размер выборки. Наблюдения, для которых значение этой статистики превышает среднее в два или три раза, должны быть проанализированы.

Программный код, при помощи которого можно построить диаграмму для значений показателя влияния, таков:

```
> hat.plot<-function(fit) {
    p<-length(coefficients(fit))
    n<-length(fitted(fit))
    plot(hatvalues(fit),main="Диаграмма значений показателя влияния")
     abline(h=c(2,3)*p/n,col="green",lty=2)
```

```
    identify(1:n, hatvalues(fit), names(hatvalues(fit)))
}
> hat.plot(fit)
```

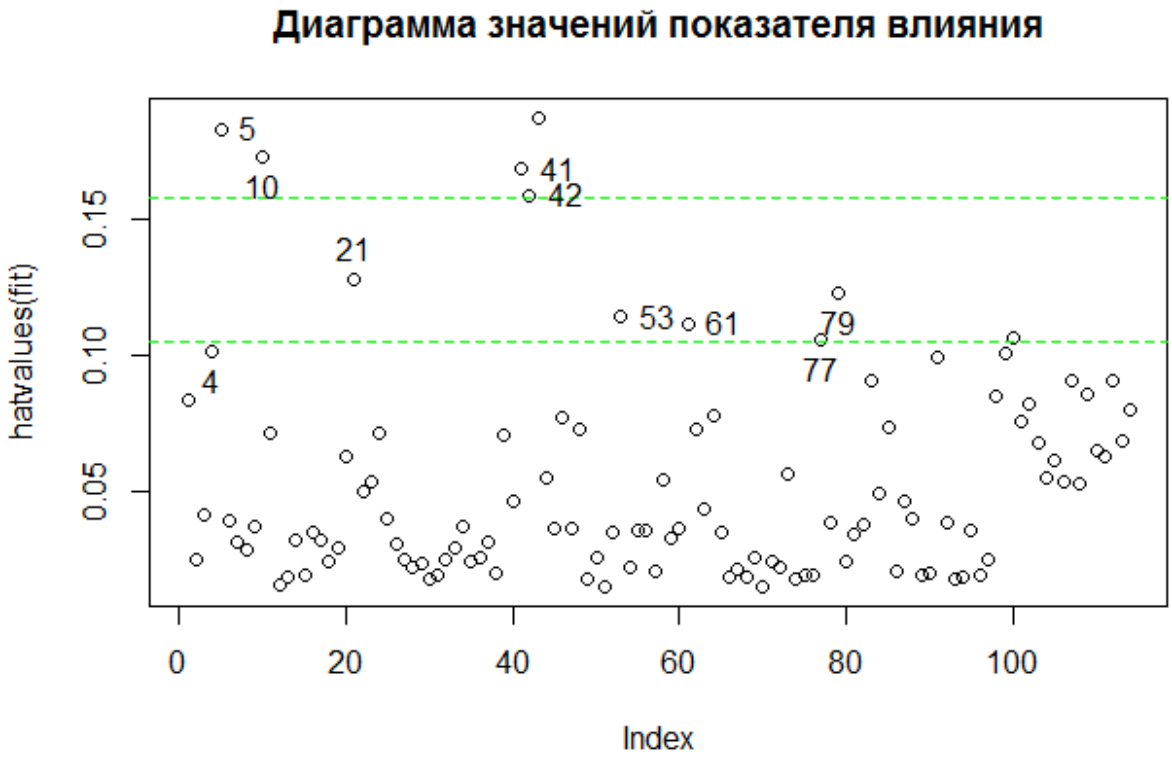

Рисунок 15- «Диаграмма значений показателя влияния»

На рисунке 15 видно, что особо необычными являются наблюдения 4,5,10,21,41,42,43,53,61,77,79. Можно объяснить это тем, что например 4 и 5 наблюдения- это две команды из английской футбольной премьер- лиги отличились в этом сезоне выделенными деньгами на приобретение новых футболистов. 41 и 42 наблюдения- это две команды испанской футбольной лиги, они в свою очередь отличились наибольшим количеством забитых голов за сезон среди всех команд ведущих европейских футбольных лиг.

Теперь сведём информацию о выбросах, влиятельных наблюдениях, точках с высокой напряжённостью на одну очень информативную диаграмму при помощи функции influencePlot.

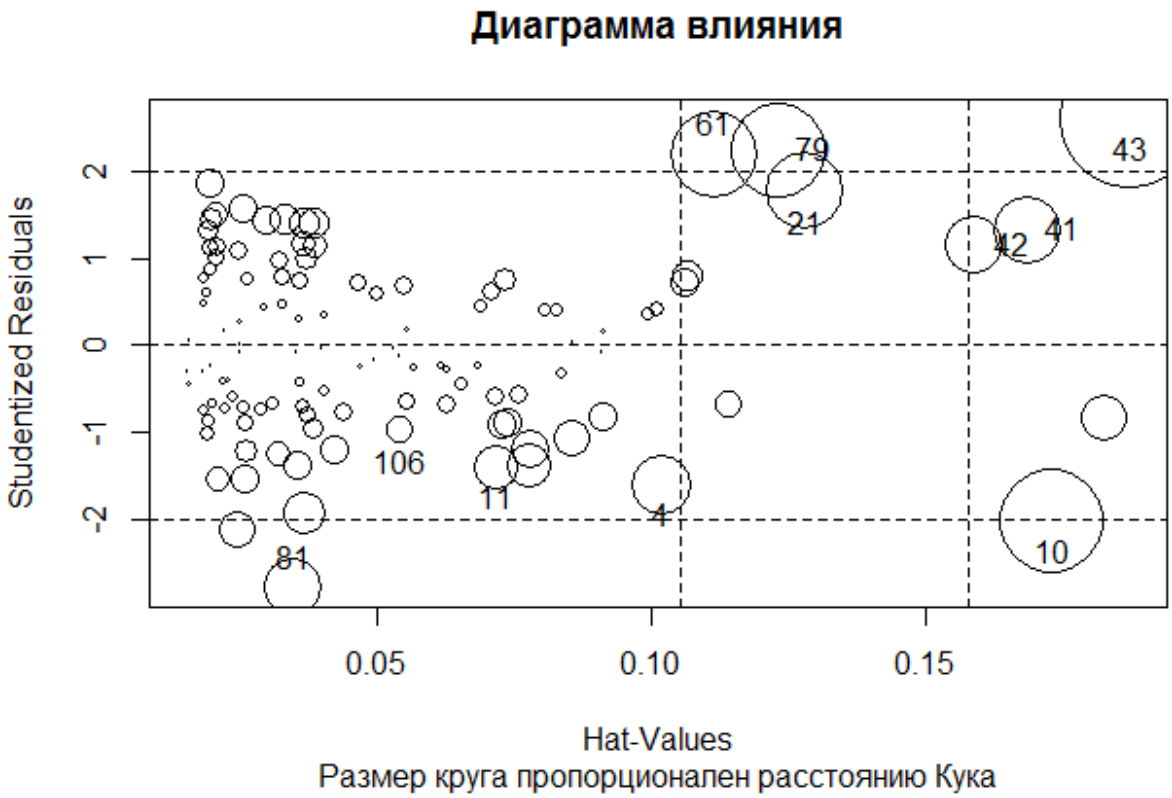

Рисунок 16- «Диаграмма влияния»

На рисунке 16 видно, что 10, 43,79 наблюдения - это влиятельные наблюдения, в свою очередь,4, 41,42, 61, наблюдения – это точки с высокой напряжённостью.

3.3 ДИАГНОСТИКА РЕГРЕССИОННЫХ МОДЕЛЕЙ НА ВЫПОЛНЕНИЕ СТАНДАРТНЫХ УСЛОВИЙ НА ОСТАТКИ

а) Проведём тест на гомоскедастичность (Тест Уайта).

```
library(car)
> ncvTest(fit)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 3.431638    Df = 1     p = 0.06395862
```

Функция ncvTest() позволяет проверить гипотезу о постоянстве дисперсии остатков как альтернативу тому, что дисперсия остатков изменяется в зависимости от подобранных значений. Статистически значимый результат свидетельствует о гетероскедастичности (неоднородности дисперсии остатков).

Так как в данном случае p-value>0.05, то гипотезу о гетероскедастичности отклоняем, значит наша модель гомоскедастична. Также стало ясно, что нам не нужно проводить здесь никаких преобразований.

Также осуществим проверку наличия гетероскедастичности с помощью теста Голдфельдта-Квандта. Тест Голдфельдта-Квандта — процедура тестирования гетероскедастичности случайных ошибок регрессионной модели, применяемая в случае, когда есть основания полагать, что стандартное отклонение ошибок может быть пропорционально некоторой переменной. Тест также основывается на предположении нормальности распределения случайных ошибок регрессионной модели.

```
> gqtest(fit)

        Goldfeld-Quandt test

data:  fit
GQ = 1.0919, df1 = 51, df2 = 51, p-value = 0.3773
```

Так как p-value=0.3773>0.05, то данный результат является незначимым. На основании этого делаем вывод о том, что в нашей модели отсутствует проблема гетероскедастичности остатков.

б) Проведём тест на автокорреляцию (Тест Дарбина- Уотсона).

Тест Дарбина-Уотсона – статистический критерий, используемый для нахождения автокорреляции остатков первого порядка регрессионной модели. Наблюдения, сделанные в короткие отрезки времени, более сильно коррелируют друг с другом, чем разнесенные во времени наблюдения.

```
> durbinWatsonTest(fit)

 lag Autocorrelation D-W Statistic p-value
   1       0.5785149      0.840559       0

 Alternative hypothesis: rho != 0
```

Так как значение p-value, получилось равным нулю, то делаем вывод о том, что в нашей модели присутствует проблема автокорреляции, то есть зависимость между остатками есть.

Проведём ещё один тест на проверку автокорреляции, а именно тест Бройша-Годфри.

В данном тесте случайные ошибки не обязательно должны быть нормально распределены. Тест является асимптотическим, то есть для достоверности выводов требуется большой объём выборки. Особенность данного теста заключается в том, что его можно использовать практически всегда. Если значение статистики превышает критическое значение, то автокорреляция признаётся значимой, в противном случае она незначима.

```
bgtest(fit)

        Breusch-Godfrey test for serial correlation of order up to 1

data:  fit
LM test = 42.267, df = 1, p-value = 7.963e-11
```

Так как значение p-value получилось меньше чем 0.05, то делаем вывод о присутствии автокорреляции.

в) Проверка на нормальность распределения остатков.

Проведём проверку на нормальность распределения остатков с помощью функции qqPlot.

Функция qqPlot() является более аккуратным методом проверки предположения о нормальности. Она изображает связь между остатками Стьюдента и квантилями распределения Стьюдента с n – p – 1 степенями свободы, где n –это объем выборки, а p –число параметров регрессии (включая свободный член).

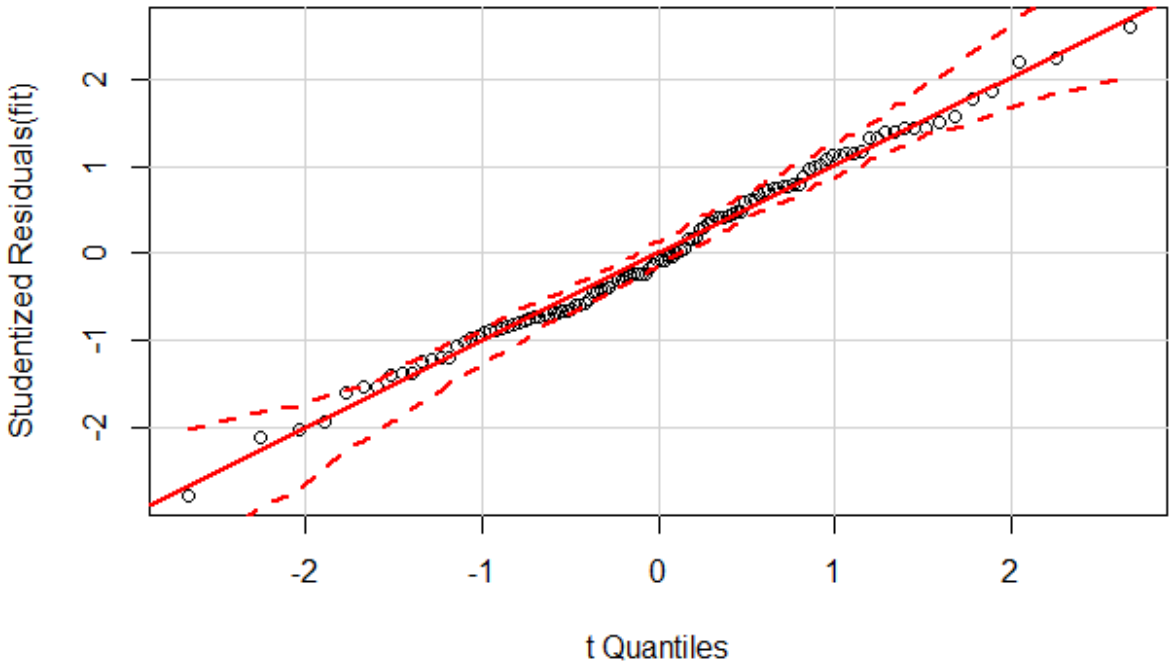

Рисунок 17- « Диаграмма qqPlot»

Как видно на рисунке 17, большая часть наблюдений попадает в рамки доверительного интервала, что свидетельствует о нормальности распределения.

3.4 ВЫБОР «ЛУЧШЕЙ РЕГРЕССИОННОЙ МОДЕЛИ»

Проведём проверку на наличие нелинейной связи между зависимой и объясняющими переменными с помощью диаграммы остатков и компонентов. Нелинейность показывает то, что возможно некорректно смоделирована функциональная форма этой независимой переменной в уравнении.

```
crPlots(fit)
```

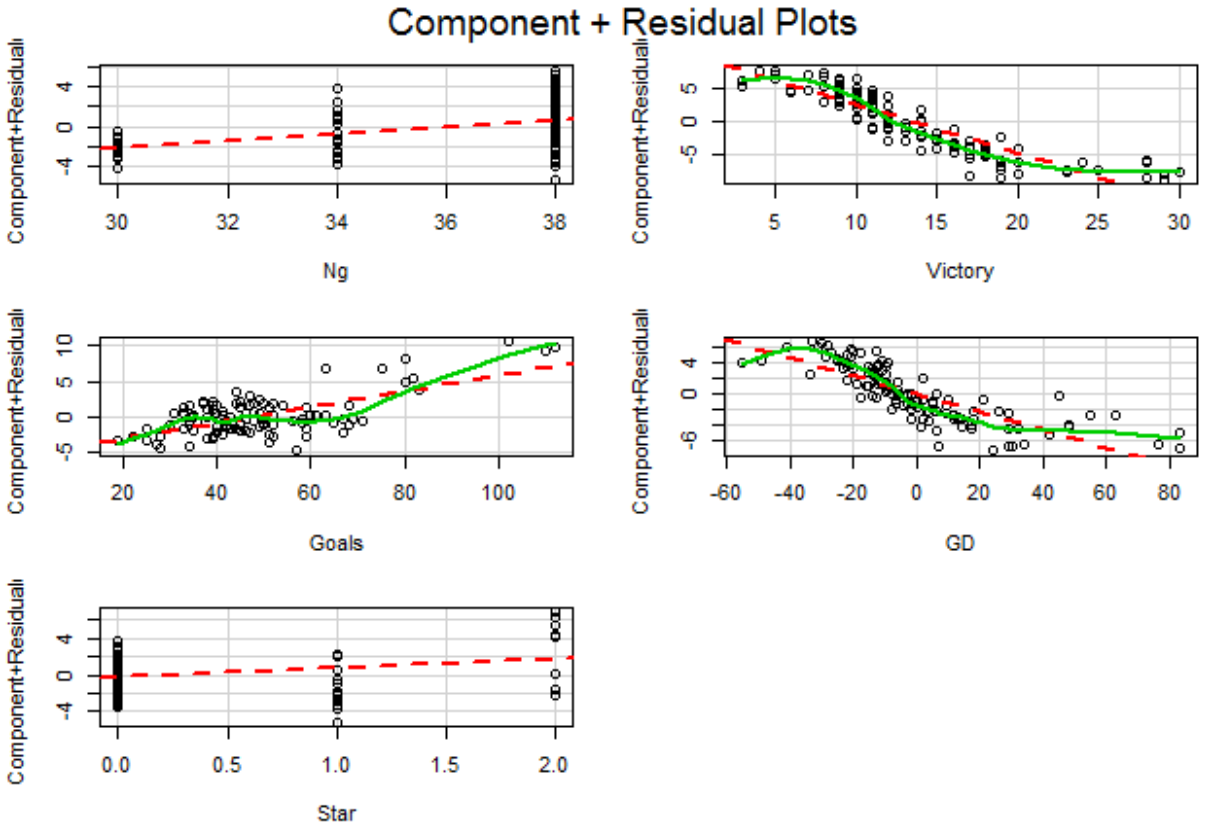

Рисунок 18- «Диаграмма частных остатков»

Теперь проведём проверку необходимости преобразования Бокса-Кокса к нормальному виду.

```
> summary(powerTransform(Futball$M))
bcPower Transformation to Normality

          Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
Futball$M    0.6993    0.135           0.4348           0.9638

Likelihood ratio tests about transformation parameters
                             LRT df         pval
LR test, lambda = (0) 31.726033  1 0.0154224
LR test, lambda = (1)  4.607168  1 0.189
```

Из результата применения данной функции следует, что зависимую переменную M можно нормализовать, заменив её на M в степени 0.69. Однако в данном случае гипотеза о lambda=1 не может быть отвергнута (p-value=0.189), поэтому необходимость такого преобразования не нужна.

Теперь проведём выбор «лучшей» регрессионной модели с помощью нескольких методов.

Начнём с метода сравнения моделей с помощью функции anova().

```
> fit1<-lm(M~Ng+Victory+Goals+GD+Star+Transfers,data=Futball)
> fit2<-lm(M~Ng+Victory+Goals+GD+Star,data=Futball)
> anova(fit2,fit1)
Analysis of Variance Table

Model 1: M ~ Ng + Victory + Goals + GD + Star
Model 2: M ~ Ng + Victory + Goals + GD + Star + Transfers
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    108 530.45
2    107 529.34  1    1.1108 0.2245 0.6366
```

В данном случае модель 1 вложена в модель 2. Функция anova() одновременно проверяет, занижает или завышает модель без переменной Transfers, предсказанные значения по сравнению с полным набором переменных. Поскольку результат теста получился незначим (p-value=0.6366), мы заключаем, что эта переменная не увеличивает предсказательную силу модели, так что мы правильно её исключили из нашей модели.

Информационный критерий Акаике – это другой способ сравнения моделей. При расчёте этого критерия учитывается статическое соответствие модели данным и число необходимых для достижения этого соответствия параметров. Предпочтения нужно отдавать моделям с меньшим значением AIC.

```
> AIC(fit1,fit2)
     df      AIC
fit1  8 514.5566
fit2  7 512.7956
```

С помощью этого критерия также подтвердилась необходимость исключения из модели переменной Transfers.

Теперь реализуем метод пошаговой регрессии, а именно метод пошагового исключения переменных.

```
> stepAIC(fit1,direction="backward")
Start:  AIC=189.04
M ~ Ng + Victory + Goals + GD + Star + Transfers

            Df Sum of Sq    RSS    AIC
- Transfers  1     1.111 530.45 187.28
<none>                   529.34 189.04
- Star       1    17.325 546.66 190.71
- Ng         1    57.497 586.83 198.79
- Goals      1    59.768 589.11 199.23
- GD         1    66.815 596.15 200.59
- Victory    1   220.978 750.31 226.81

Step:  AIC=187.28
M ~ Ng + Victory + Goals + GD + Star

          Df Sum of Sq    RSS    AIC
<none>                  530.45 187.28
- Star     1    18.190 548.64 189.12
- Ng       1    56.650 587.10 196.84
- Goals    1    61.939 592.39 197.87
- GD       1    66.693 597.14 198.78
```

```
- Victory  1   223.773 754.22 225.40

Call:
lm(formula = M ~ Ng + Victory + Goals + GD + Star, data = Futball)

Coefficients:
(Intercept)           Ng      Victory        Goals           GD         Star
     2.7365       0.3226      -0.7513       0.1128      -0.1155       1.0290
```

В данном случае мы начали с модели, содержащей все 6 независимых переменных. В столбце AIC приведено значение одноименного критерия для модели, из которой удалена указанная в соответствующей строке переменная. На первом шаге была удалена переменная Transfers, что привело к уменьшению AIC с 189.04 до 187.28. Удаление остальных переменных увеличивает значение данного критерия, поэтому процесс остановлен.

Также проведём регрессию по всем подмножествам с помощью функции regsubsets().

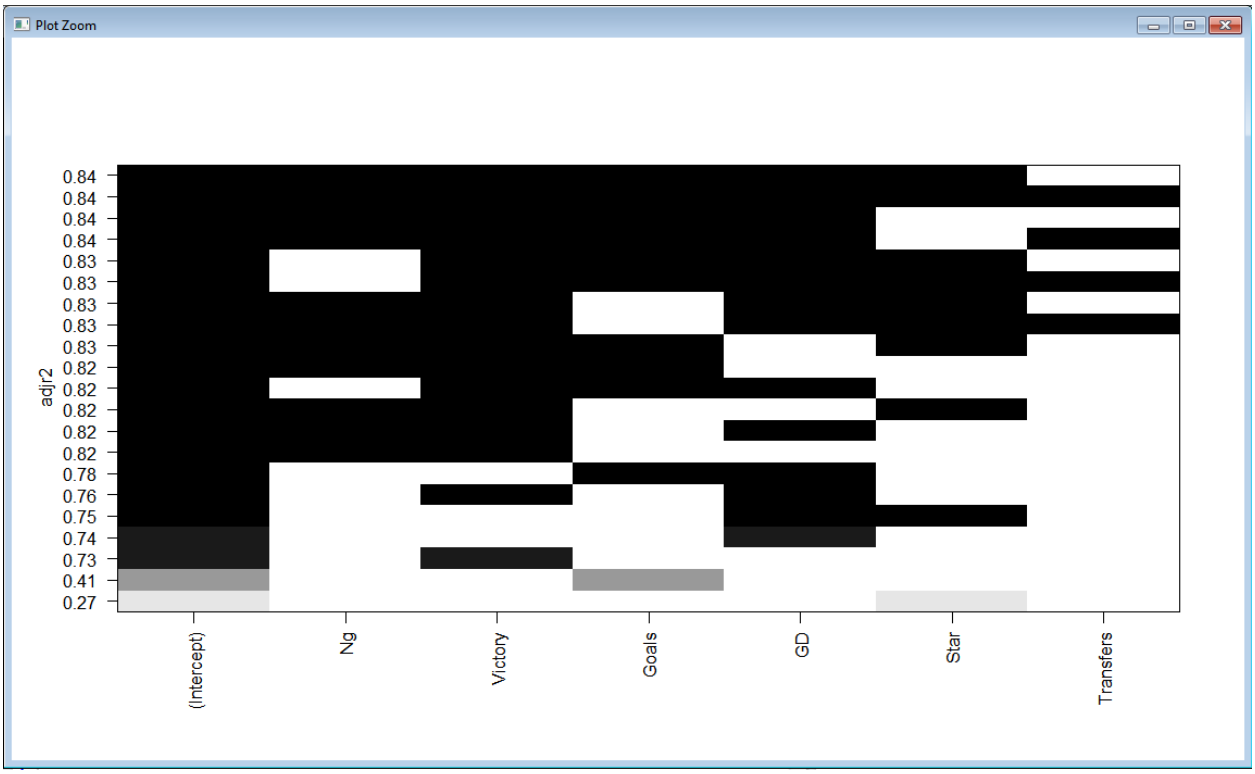

Рисунок 19- "Диаграмма регрессии по подмножествам"

На рисунке 19 видно, что модель, включающая 5 переменных, является наилучшей и имеет наибольший скорректированный коэффициент детерминации.

Также проанализируем регрессию по подмножествам с помощью показателя статистики Мэллоуса (Рисунок 20).

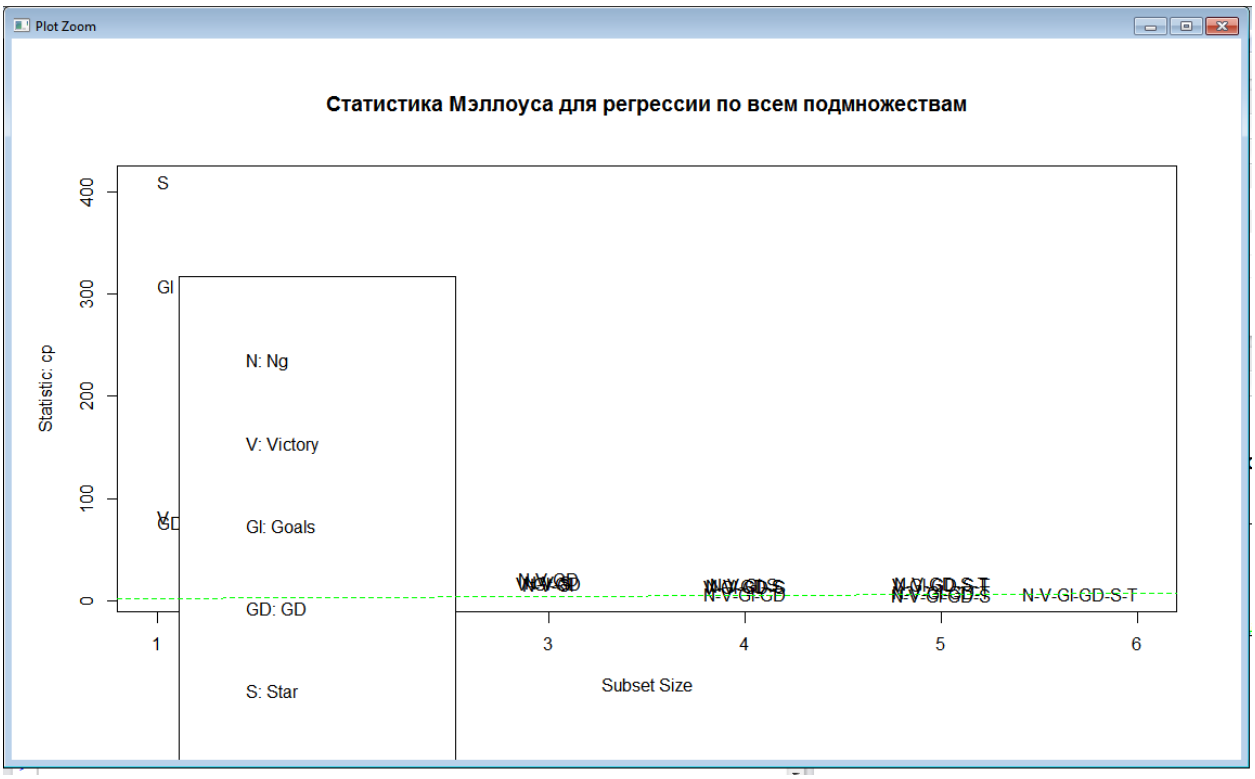

Рисунок 20- "Статистика Мэллоуса"

На данной диаграмме видно, что лучше всего нашу зависимость отражает модель, включающая в себя 5 переменных, без переменной Transfers.

Теперь проведём тест Чоу на проверку устойчивости параметров модели. Гипотеза H0 в данном случае утверждает, что качество общей модели регрессии без ограничений лучше качества частных моделей регрессии.

```
##      F value        d.f.1        d.f.2      P value
## 1.064675e+02 2.000000e+00 1.064000e+03 1.311216e-42
```

Видно, что p-value получилось очень маленьким, поэтому делаем вывод, что качество общей модели регрессии лучше, чем частных моделей.

3.5 АНАЛИЗ ОТНОСИТЕЛЬНОЙ ВАЖНОСТИ НЕЗАВИСИМЫХ ПЕРЕМЕННЫХ

Проведём сравнение стандартизованных коэффициентов регрессии. Стандартизованные коэффициенты регрессии- это коэффициенты, делённые на стандартное отклонение.

Можно сказать, что приведённое сравнение абсолютных величин стандартизованных коэффициентов регрессии позволяет получить не самое точное, но довольно наглядное представление о важности рассматриваемых факторов.

```
> futball<-as.data.frame(scale(Futball))
> zfit<-lm(M~Ng+Victory+Goals+GD+Star+Transfers,data=futball)
> coef(zfit)
  (Intercept)            Ng       Victory         Goals            GD
-3.607047e-17  1.729343e-01 -7.629402e-01  3.217576e-01 -5.134299e-01
Star
1.274449e-01
    Transfers
-2.508902e-02
```

Исходя из полученных стандартизованных коэффициентов регрессии, переменная Transfers (выделенный бюджет на трансферные приобретения) –наименее важный параметр модели, а переменная Victory (количество побед) наиболее важный параметр модели. Что следует признать логичным, ведь чем больше побед одерживает команда в сезоне, тем выше она будет в итоговой турнирной таблице.

ЗАКЛЮЧЕНИЕ

По итогам проведённого исследования, заметим, что трансферные приобретения клубов, оказались не значимым фактором в определении результатов команд в данном сезоне.

Это можно связать с тем, что новые игроки сразу не могут влиться в коллектив и играть так же хорошо, как в своём предыдущем клубе. Очень важно здесь, чтобы футболист смог как можно быстрее адаптироваться к новым условиям, а именно к новым партнёрам по полю, тренеру, стадиону, болельщикам, даже к обстановке в новом городе. Ведь, в противном случае, будет очень большая вероятность потраченных в никуда денег. Данное явление сейчас очень распространено в футбольном мире. Чтобы избежать пустой траты денег людям, ответственным за трансферную политику клуба, нужно учесть фактор адаптации игрока как один из основных при поиске игроков. Вариантом решения данной проблемы, может стать предоставление футболисту человека из клуба, который бы помог ему и его семье быстрее привыкнуть к новой стране.

СПИСОК ЛИТЕРАТУРЫ

1. Кабаков Р.И. R в действии. Анализ и визуализация данных в программе R – М.: ДМК Пресс, 2014. – 588 с.

2.Купер С. Футболономика/ Купер С., Шимански С.-Москва: Альпина Паблишер, 2016.-520 с.