統計學一下期末報告

B10705050 陳禹翰 B11705027 陳承妤 B11705004 周子馨
B11705026 宋凱翔 B11605019 陳詣斌

June 10, 2024

1 分析主題與動機

1.1 研究動機

在職業籃球領域中，球員薪資不僅代表了球員的市場價值，還反映了他們在球隊中的貢獻和潛力。了解哪些因素對球員薪資有顯著影響，不僅能幫助球隊更好地進行薪資管理和預算規劃，還能爲球員自身的職業發展提供參考。因此我們列舉了一些可能影響薪資的變數，並通過分析 NBA2023-2024 賽季的數據，來驗證我們的猜想。

1.2 研究問題

本研究旨在研究影響 NBA 球員薪資關鍵因素，以下將分爲二個面向進行分析，並且找出恰當的回歸模型：

1. NBA 球隊所在地區經濟水平與球隊相關數據對球隊薪資總額的影響
2. NBA 球員位置各項指標對薪資的影響

2 資料描述

2.1 NBA 球員資料

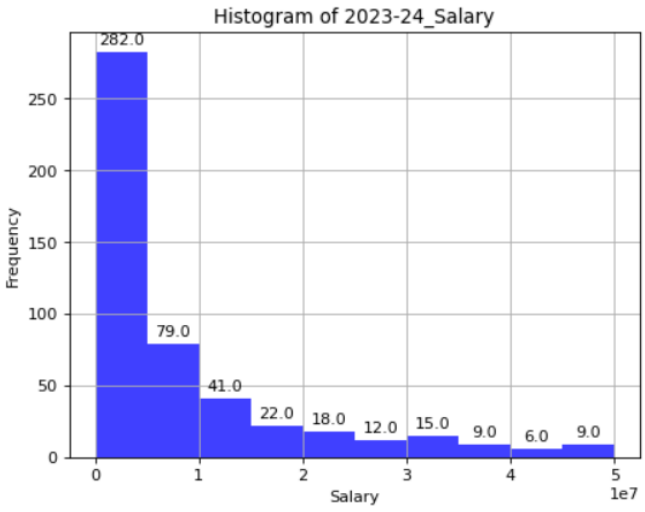
資料來源爲 basketball-reference.com 和 hoopshype.com，其中提供 NBA 2023-24 賽季有簽署合約且有上場比賽的 494 位球員相關資料，包含球員先發場次、投籃命中率、年齡等一般數據外，也含有籃板率、失誤率等相關進階數據。對於該資料集我們進行的前處理有：合併球員轉隊前後的資料；合併傳統數據、進階數據與薪水至同一檔案；計算自定義變數 (城市經濟水平、球員位置劃分、dummy variables 等)。

2.2 NBA 球隊所在城市資料

資料來源爲 american-growth-project-january-01042023r.pdf (unc.edu),2023 Fall Statement | In Brief (ontario.ca) 和 hoopshype.com，其中提供 NBA 球隊所在的 28 座城市之經濟表現與球隊的相關數據，包含城市人口數、GDP、該城市的球隊薪資及收益... 等。資料前處理爲當該城市擁有兩支球隊，該城市的球隊的相關資料則取其平均。

3 資料敍述統計與資料視覺化

對於 2023-34 賽季球員的薪水，我們畫出以下直方圖。我們可以發現圖形右偏情況明顯，代表整體低薪球員比例高。針對傳統累積數據畫出的直方圖，也有右偏的情況發生。




```
2023-24_Salary:
average:  8593873.4048583
standar deviation:  11067602.72773844
skewness:  1.8500508210381228
```

Figure 1: 2023-24 賽季球員薪水直方圖與相關統計量數

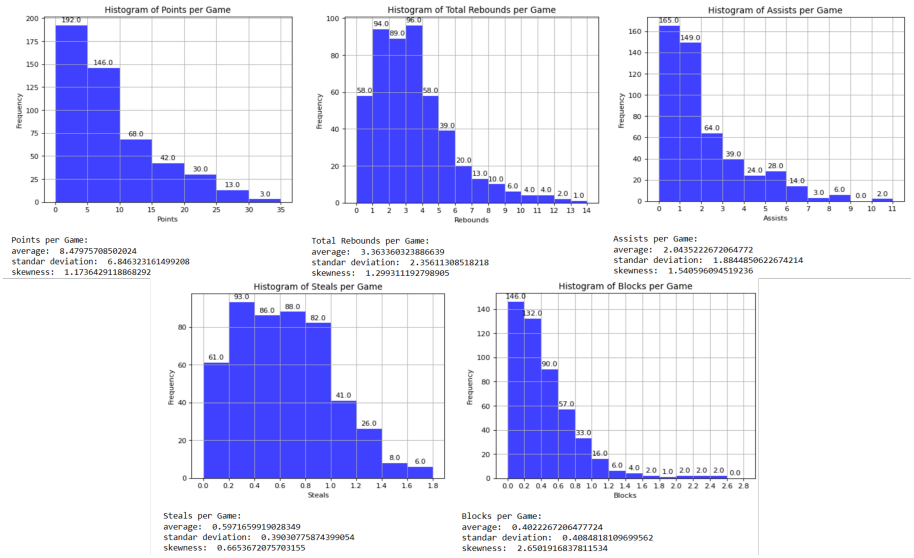
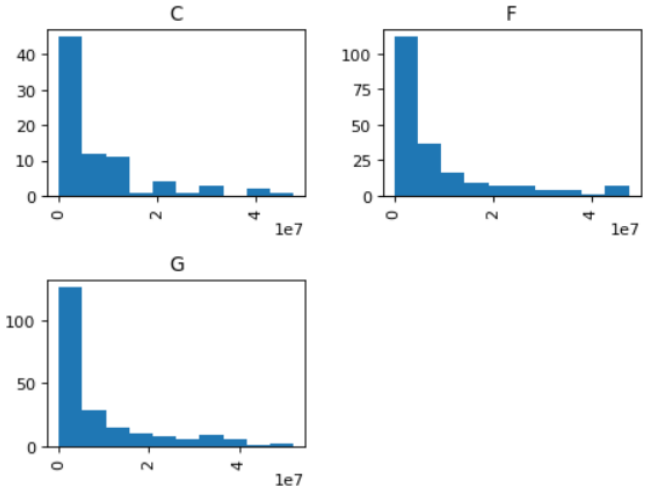


Figure 2: 2023-24 賽季球員傳統數據直方圖與相關統計量數

接著，我們畫出 2023-34 賽季各場上位置球員薪水盒鬚圖後，我們發現各位置之間的薪水之間沒有顯著差異，因此我們做以下檢定以驗證我們的猜想。首先，我們先使用直方圖與 Shapiro 檢定後發現資料不爲常態，故使用 Kruskal Wallis Test 做母數位置的檢定。

Figure 3: 2023-24 賽季球員薪水與各位置盒鬚圖




```
For Pos = C
Shapiro statistic = 0.721105 and p_value = 0.000000
For Pos = F
Shapiro statistic = 0.727757 and p_value = 0.000000
For Pos = G
Shapiro statistic = 0.738136 and p_value = 0.000000
```

Figure 4: 2023-24 賽季球員薪水直方圖（按位置）與 Shapiro Test

以下建立虛無假設與對立假設：

H_0: 所有母體的位置皆相同

H_1: 至少有兩個母體位置不同

	C	F	G
0	47607350.0	47649433.0	51915615
1	41000000.0	47607350.0	46741590
2	40600080.0	45640084.0	45640084
3	32600060.0	45640084.0	40806300
4	32459438.0	45640084.0	40064220
...	...	...	...
205	NaN	NaN	134863
206	NaN	NaN	120250
207	NaN	NaN	70687
208	NaN	NaN	64343
209	NaN	NaN	64343

210 rows × 3 columns

```
T1 = 20628.500000
T2 = 51044.000000
T3 = 50592.500000
H = 0.941534
pvalue = 0.624523
```

Figure 5: Kruskal Wallis Test 檢定結果

計算結果 p-value = 0.6245，不拒絕虛無假設，故沒有足夠證據顯示母體位置有不相同的情況，吻合我們的推測。

4 回歸模型建立與分析

我們針對以下兩個主題進行回歸模型的建立與分析：

1. NBA 球隊所在地區經濟水平與球隊相關數據對球隊薪資總額的影響
2. NBA 球員位置各項指標對薪資的影響

4.1 NBA 球隊薪資總額分析與模型

針對城市與球隊的相關資料集，首先觀察自變數對球隊薪資的散佈圖（Figure 6），可以發現 Win, Revenue (million) 與球隊總薪資有較明顯的線性關係，GDP (million) 與球隊總薪資可能有對數關係，而將該變數取 log 後發現有明顯線性關係，故此後我們選用 GDP (million) log 進行分析；Net Worth (m) 與球隊總薪資因 outliers 的存在故線性關係較不明顯。

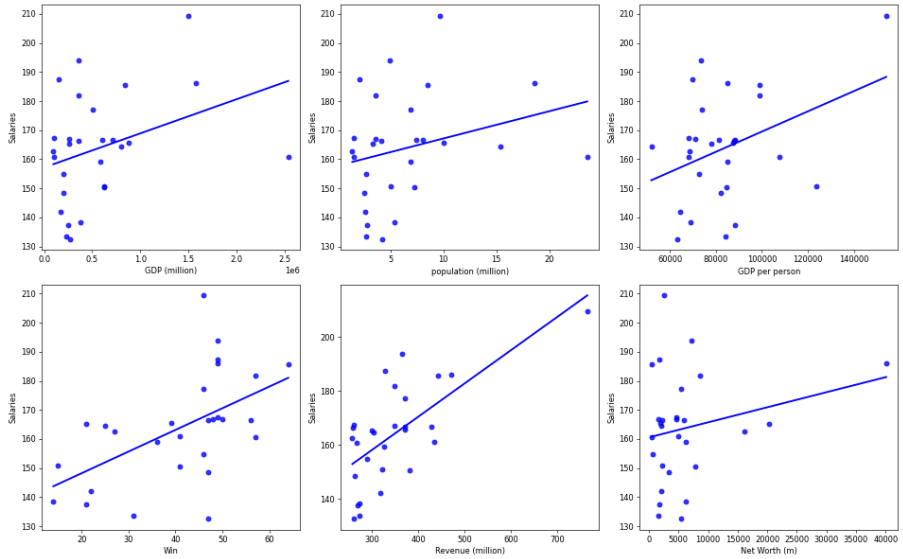


Figure 6: 所有自變數與 Team Salaries (millions) 之散佈圖

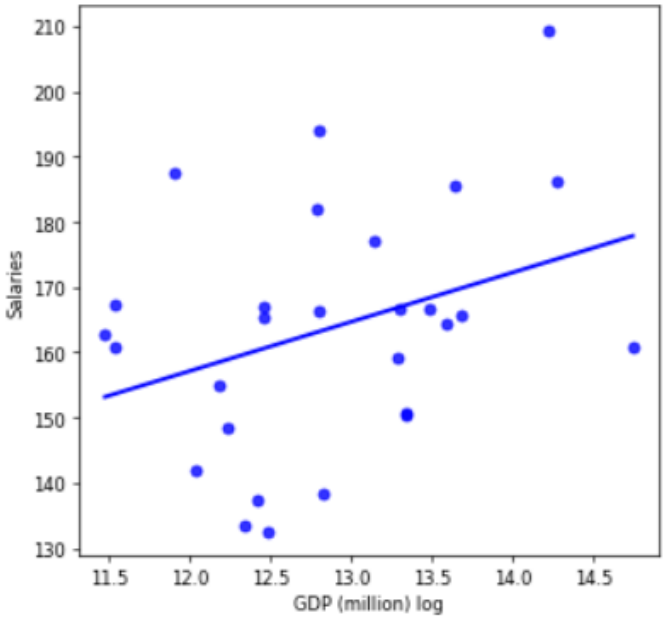
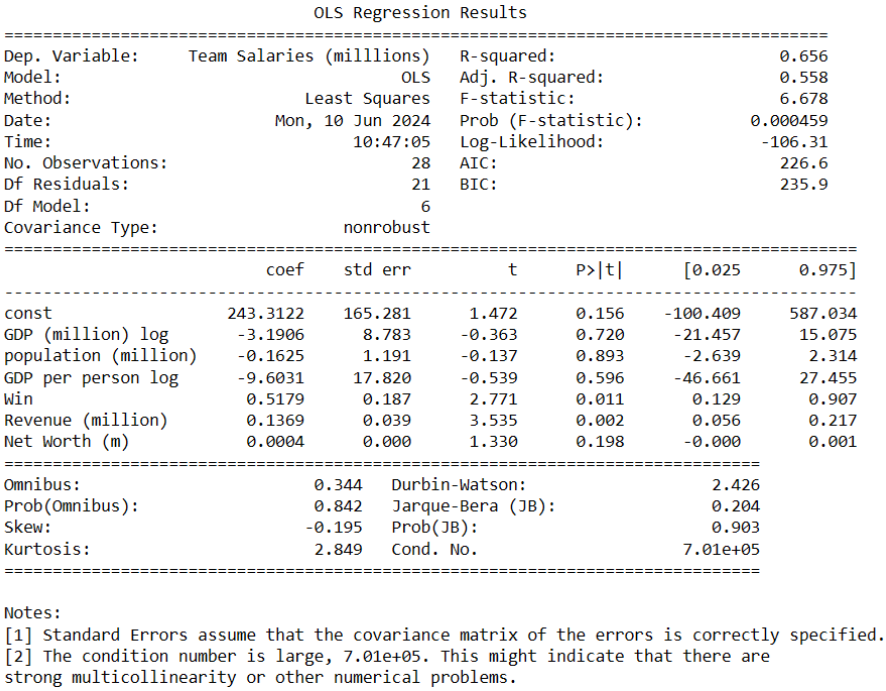


Figure 7: GDP (million) log 與 Team Salaries (millions) 之散佈圖

Figure 8: Team Salary 熱力圖

倘若我們將所有變數納入回歸模型（下方 Figure 9），可以發現 GDP (million) log, population (million), GDP per person log 三項變數有多元共線性的問題，且 t-statistic 的 p-value 很高，代表他們對於模型的關聯程度低。



OLS Regression Results

Dep. Variable:	Team Salaries (milllions)	R-squared:	0.656
Model:	OLS	Adj. R-squared:	0.558
Method:	Least Squares	F-statistic:	6.678
Date:	Mon, 10 Jun 2024	Prob (F-statistic):	0.000459
Time:	10:47:05	Log-Likelihood:	-106.31
No. Observations:	28	AIC:	226.6
Df Residuals:	21	BIC:	235.9
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	243.3122	165.281	1.472	0.156	-100.409	587.034
GDP (million) log	-3.1906	8.783	-0.363	0.720	-21.457	15.075
population (million)	-0.1625	1.191	-0.137	0.893	-2.639	2.314
GDP per person log	-9.6031	17.820	-0.539	0.596	-46.661	27.455
Win	0.5179	0.187	2.771	0.011	0.129	0.907
Revenue (million)	0.1369	0.039	3.535	0.002	0.056	0.217
Net Worth (m)	0.0004	0.000	1.330	0.198	-0.000	0.001

Omnibus:	0.344	Durbin-Watson:	2.426
Prob(Omnibus):	0.842	Jarque-Bera (JB):	0.204
Skew:	-0.195	Prob(JB):	0.903
Kurtosis:	2.849	Cond. No.	7.01e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.01e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Figure 9: 回歸模型結果（全變數）

故我們使用 Forward stepwise, Backward stepwise, Best subset 找出對球隊總薪資相對重要的變數。Forward stepwise 和 Backward stepwise 所找出的變數皆爲球隊勝場數與球隊收入，而 Best subest 則是城市 GDP 的對數、球隊勝場數、球隊收入、和球隊老闆身價。使用這兩種建構模型後，我們發現 Best subset 的模型（下方 Figure 10）在城市 GDP 的對數這項依舊有多元共線性的問題，倘若刪去此變數的話（下方 Figure 11）則有自回歸的問題，且此模型與 Forward/Backward stepwise 的模型相比，兩者解釋力的差異不大，故我們採用以球隊勝場數與球隊收入爲自變數之線性回歸模型（下方 Figure 12）。

OLS Regression Results

Dep. Variable:	Team Salaries (milllions)	R-squared:	0.651
Model:	OLS	Adj. R-squared:	0.590
Method:	Least Squares	F-statistic:	10.72
Date:	Mon, 10 Jun 2024	Prob (F-statistic):	4.69e-05
Time:	10:47:06	Log-Likelihood:	-106.51
No. Observations:	28	AIC:	223.0
Df Residuals:	23	BIC:	229.7
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	156.0599	44.076	3.541	0.002	64.881	247.239
GDP (million) log	-4.6900	3.888	-1.206	0.240	-12.732	3.352
Win	0.5242	0.179	2.926	0.008	0.154	0.895
Revenue (million)	0.1271	0.033	3.881	0.001	0.059	0.195
Net Worth (m)	0.0004	0.000	1.528	0.140	-0.000	0.001

Omnibus:	0.122	Durbin-Watson:	2.448
Prob(Omnibus):	0.941	Jarque-Bera (JB):	0.019
Skew:	-0.002	Prob(JB):	0.990
Kurtosis:	2.871	Cond. No.	1.94e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.94e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Figure 10: 回歸模型結果（Best Subset）

OLS Regression Results

Dep. Variable:	Team Salaries (milllions)	R-squared:	0.629
Model:	OLS	Adj. R-squared:	0.583
Method:	Least Squares	F-statistic:	13.56
Date:	Mon, 10 Jun 2024	Prob (F-statistic):	2.23e-05
Time:	11:17:49	Log-Likelihood:	-107.37
No. Observations:	28	AIC:	222.7
Df Residuals:	24	BIC:	228.1
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	104.1585	9.667	10.775	0.000	84.207	124.110
Win	0.5578	0.179	3.123	0.005	0.189	0.926
Revenue (million)	0.0995	0.024	4.198	0.000	0.051	0.148
Net Worth (m)	0.0004	0.000	1.359	0.187	-0.000	0.001

Omnibus:	0.170	Durbin-Watson:	2.558
Prob(Omnibus):	0.919	Jarque-Bera (JB):	0.009
Skew:	0.011	Prob(JB):	0.995
Kurtosis:	2.915	Cond. No.	4.20e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.2e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Figure 11: 回歸模型結果（Best Subset 刪除城市 GDP 的對數）

OLS Regression Results

Dep. Variable:	Team Salaries (milllions)	R-squared:	0.600
Model:	OLS	Adj. R-squared:	0.568
Method:	Least Squares	F-statistic:	18.78
Date:	Sat, 08 Jun 2024	Prob (F-statistic):	1.05e-05
Time:	02:53:17	Log-Likelihood:	-108.41
No. Observations:	28	AIC:	222.8
Df Residuals:	25	BIC:	226.8
Df Model:	2		
Covariance Type:	nonrobust		

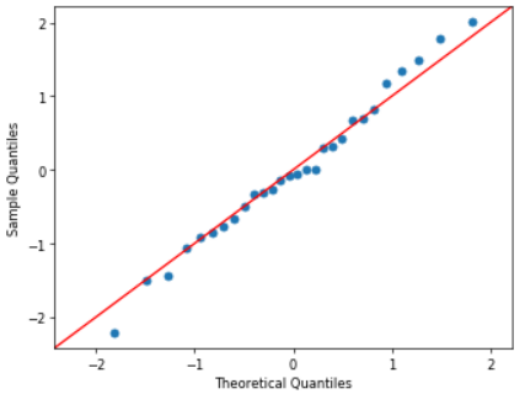
	coef	std err	t	P>|t|	[0.025	0.975]
const	105.7787	9.754	10.844	0.000	85.689	125.868
Win	0.5346	0.181	2.957	0.007	0.162	0.907
Revenue (million)	0.1045	0.024	4.389	0.000	0.055	0.154

Omnibus:	0.030	Durbin-Watson:	2.398
Prob(Omnibus):	0.985	Jarque-Bera (JB):	0.167
Skew:	0.068	Prob(JB):	0.920
Kurtosis:	2.647	Cond. No.	1.52e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.52e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

Figure 12: 回歸模型結果（Forward/Backward Stepwise）

最後，我們對此模型進行殘差分析以確認模型是否有效，該模型通過所有檢定，且無多元共線性與自回歸（Figure 14-17）。



Shapiro Test
Statistics=0.988, p=0.980

Figure 13: Q-Q Plot and Shapiro Test

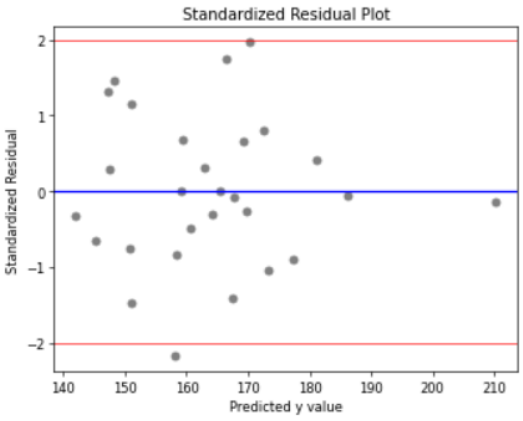


Figure 14: Scatter Plot for Standardized Residual (Predicted y Value)

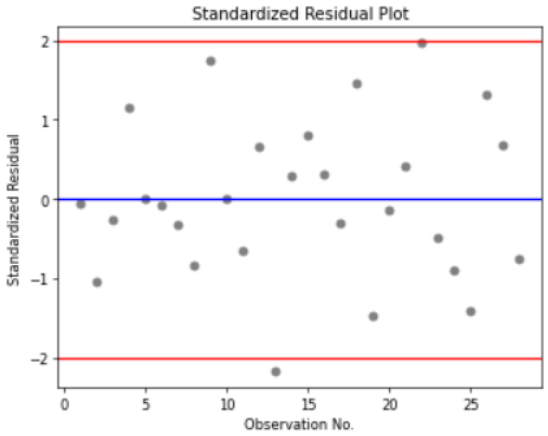


Figure 15: Scatter Plot for Standardized Residual (Observations)

```
runs =  16
n1 =  14
n2 =  14
runs_exp =  15.0
stan_dev =  2.596293654566205
z =  0.38516444325982163
pval_z =  0.7001155851110985
p_value for Z-statistic=  0.7001155851110985
```

Figure 16: Run Test Result

```
x_square_sum =  26.969221321761975
size =  28
x_d =  [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0.]
x_d =  [ 0.         -0.97970919  0.77160199  1.40699371 -1.13469788 -0.08810805
 -0.23911161 -0.52263868  2.59379519 -1.74933055 -0.66546271  1.32511968
 -2.82816529  2.45848321  0.49910056 -0.4827041  -0.61359355  1.76244483
 -2.92701268  1.3293859   0.55777064  1.54903809 -2.44756794 -0.413809
 -0.52370688  2.73240697 -0.63155564 -1.44366067]
d =  2.3947708052382803
```

Figure 17: Durbin Waston Test Result

4.2 球員位置對薪資的影響

我們先將球員位置分爲三大類，分別爲前鋒 (Forward)、中鋒 (Center)、後衛 (Guard)。我們使用 Stepwise Forward Regression 跟 Best Subsets Regression，找出對所有球員 (All)、前鋒 (Forward)、中鋒 (Center)、後衛 (Guard) 影響較大的變數，來推測各位置大致會重視哪些數據。

4.2.1 數據選擇與變數轉換

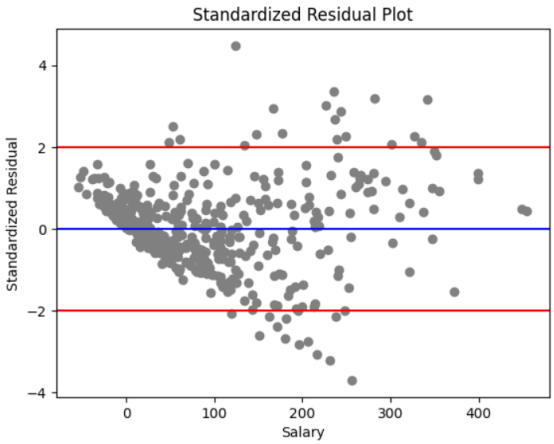


Figure 18: Salary

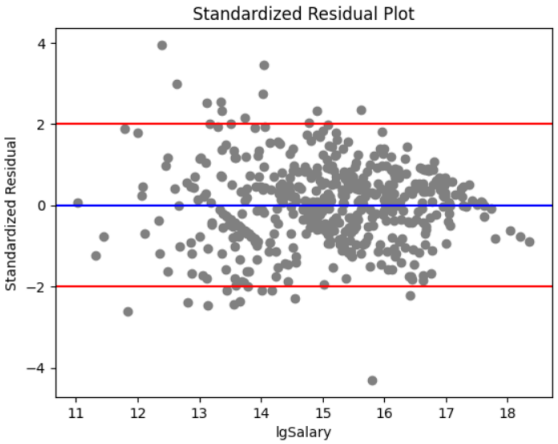


Figure 19: Log Salary

爲了使資料的殘差分析符合同質性，我們將薪資取 log，從 Figure 18 跟 Figure 19 的對比中，可看出取 log 前後的差別。

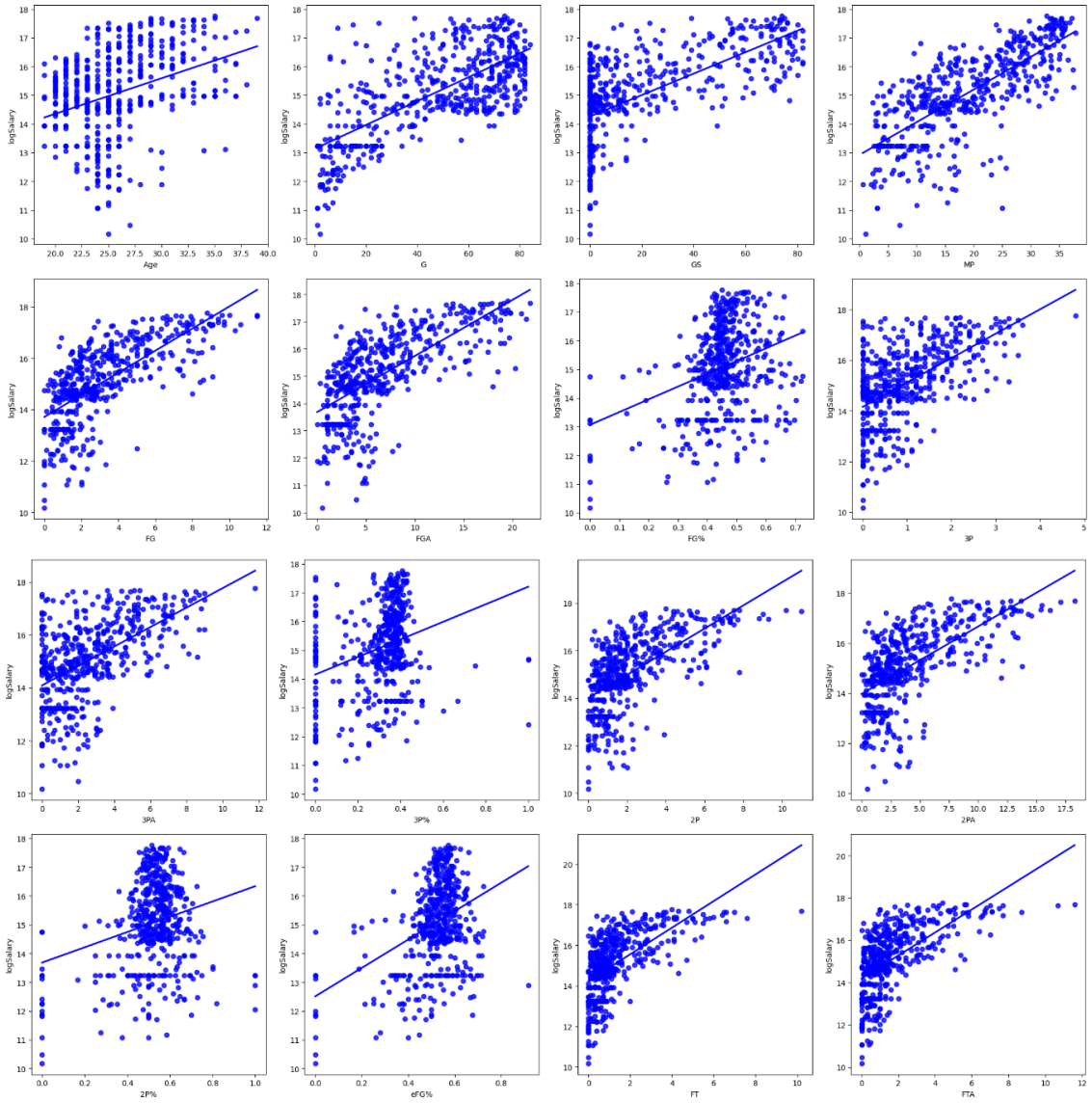


Figure 20: 所有變數散佈圖 (上)

從 Figure 20 中，可觀察到有些呈現 log 曲線，如 FG、FGA 等，像這些資料我們有做取 log 處理。有些資料既非線性也非 log 曲線，如 2P%、eFG% 等，我們刪除該變數。

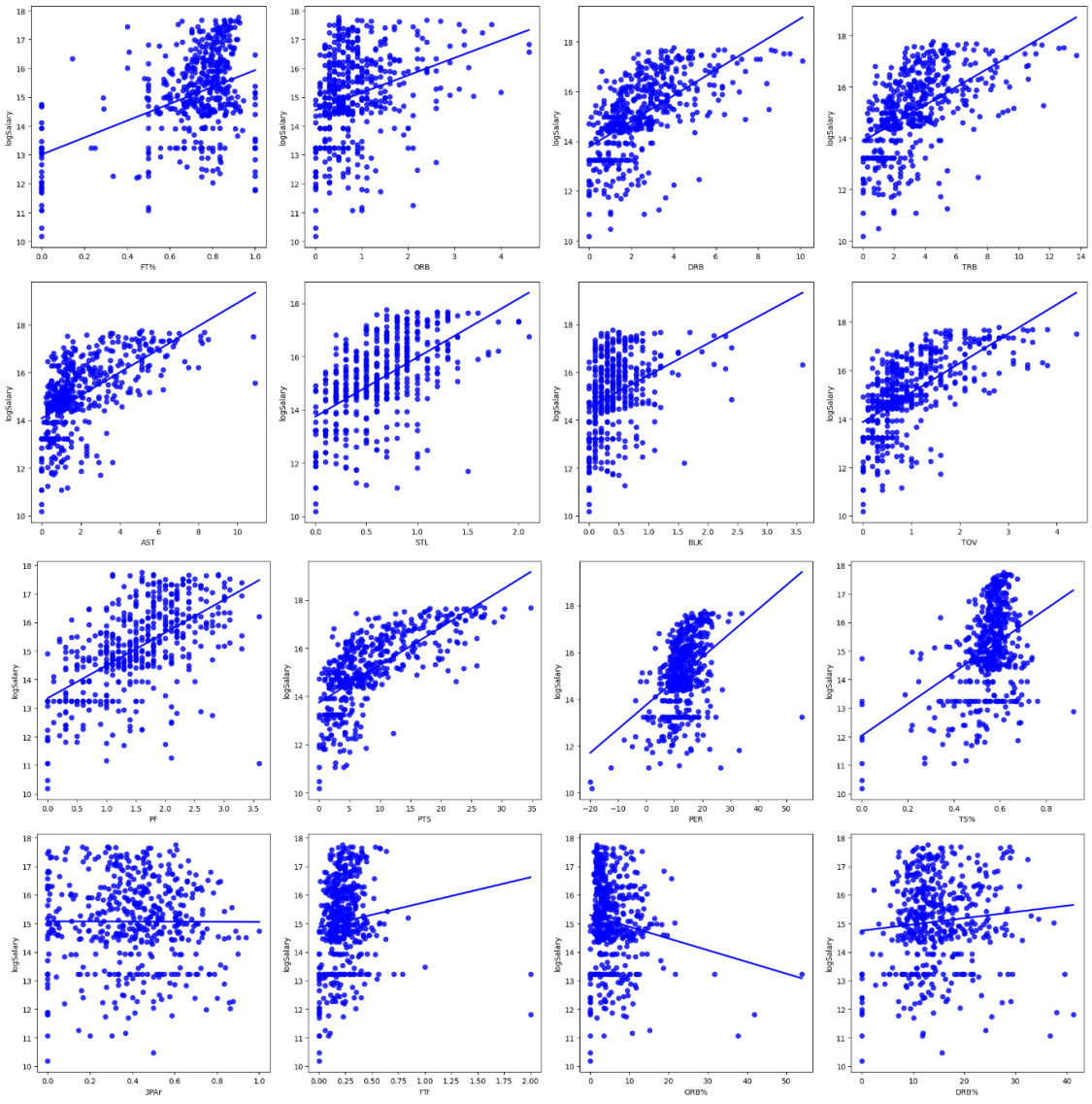


Figure 21: 所有變數散佈圖 (中)

從 Figure 21 中，可觀察到有些呈現 log 曲線，如 AST、TOV 等，像這些資料我們有做取 log 處理。有些資料既非線性也非 log 曲線，如 ORB%、DRB% 等，我們刪除該變數。

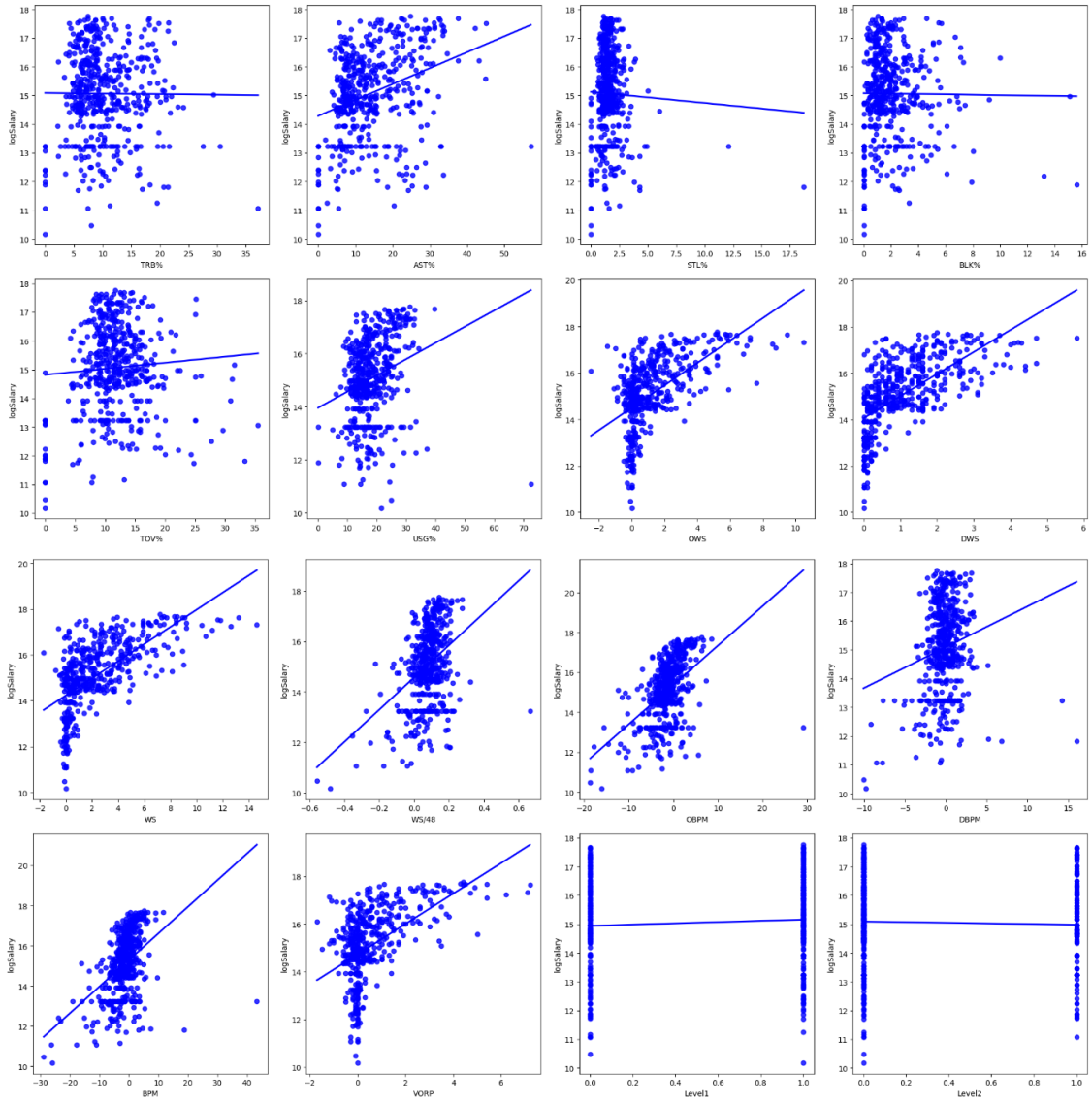


Figure 22: 所有變數散佈圖 (下)

從 Figure 22 中，可觀察到有些呈現 log 曲線，如 OWS、DWS 等，但我們做取 log 處理後，該變數依然非線性分佈，因此我們刪除該變數。

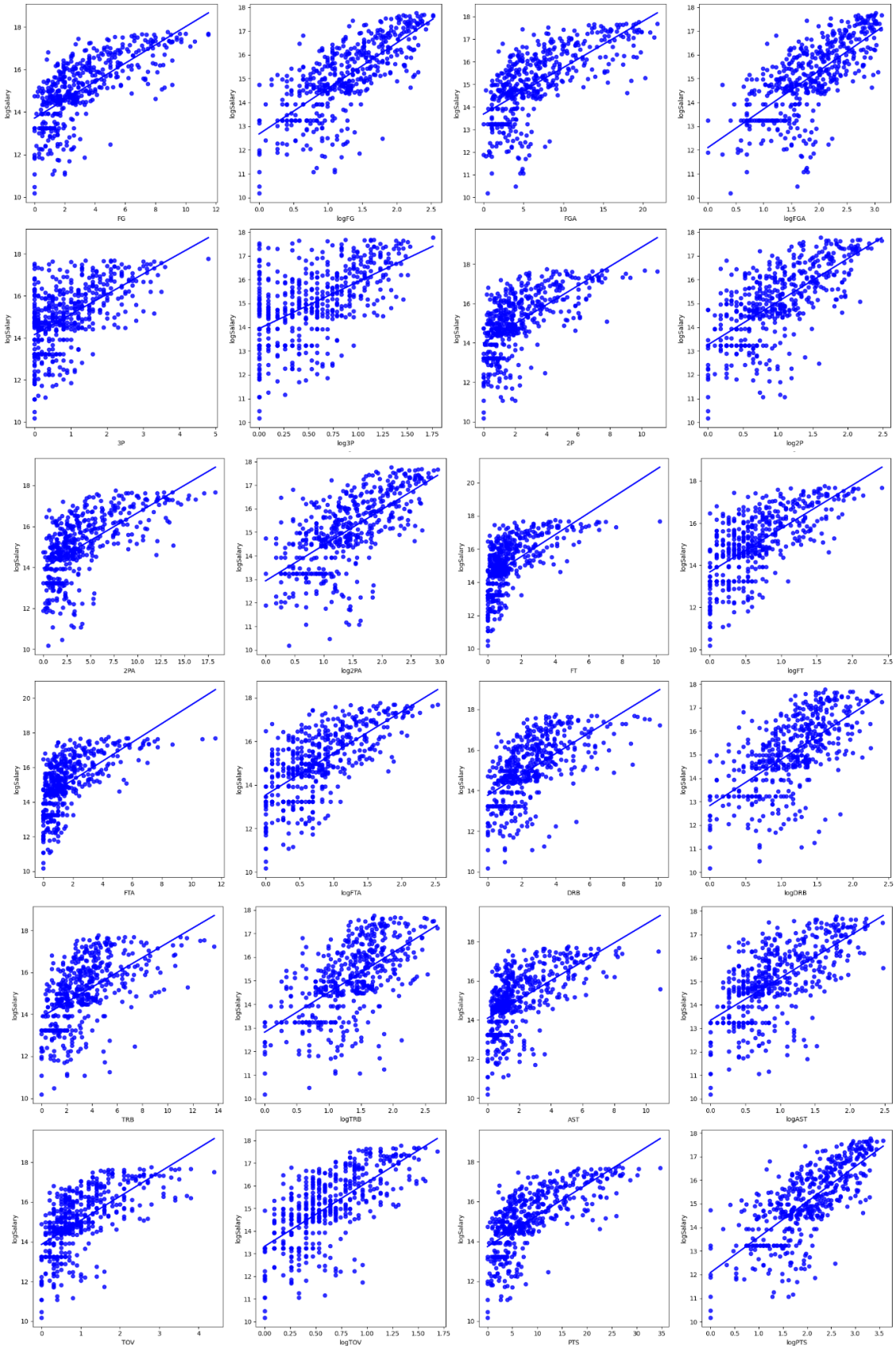


Figure 23: 所選變數與取 log 後的散佈圖對照

Figure 23 爲我們選擇的變數與轉換後變數的對照，分別爲 FG, logFG, FGA, logFGA, 3P, log3P, 2P, log2P, 2PA, log2PA, FT, logFT, FTA, logFTA, DRB, logDRB, TRB, logTRB, AST, logAST, TOV, logTOV, PTS, logPTS。

最後我們選擇 19 個自變數，分別爲 Age、G、GS、MP、3P%、STL、PF、logFG、logFGA、log3P、log2P、log2PA、logFT、logFTA、logDRB、logTRB、logAST、logTOV、logPTS。

4.2.2 所有球員 (All)

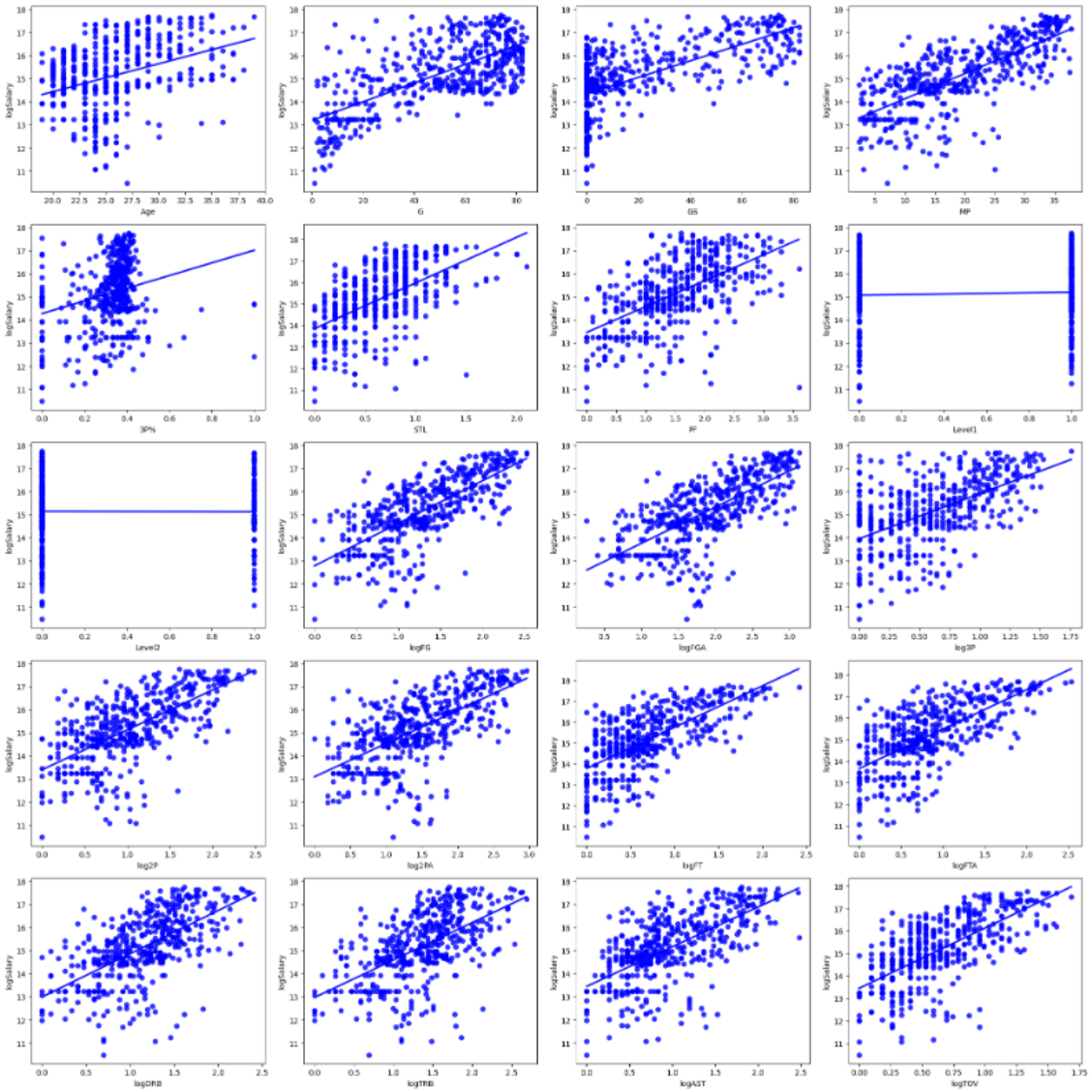


Figure 24: 所有球員—薪水與各變數的散佈圖

Stepwise Forward Regression:

```
from stepwise_regression import step_reg
forwardselect = step_reg.forward_regression(X_data2, y_data_1, 0.1, verbose=False)
print(forwardselect)
```

```
['const', 'MP', 'G', 'Age', 'logFT']
```

Figure 25: 所有球員—Stepwise Forward Regression 選擇的變數

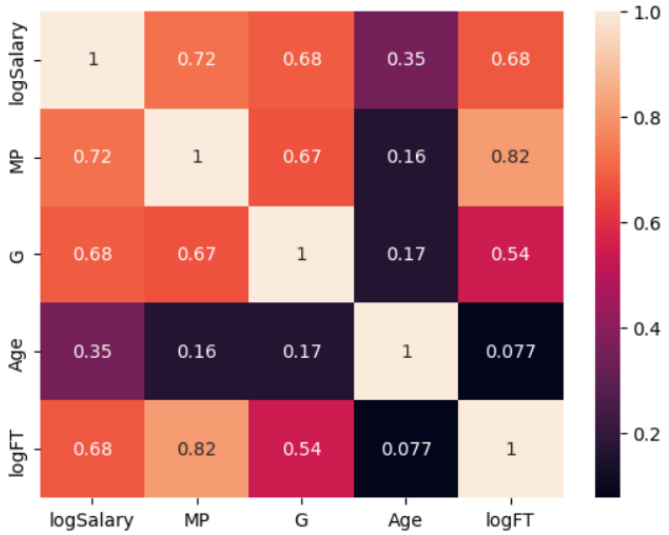


Figure 26: 所有球員—log 球員薪水與 Stepwise Forward 選擇變數之熱力圖

從 Figure 26 可得 corr(MP,logFT)$>$ 0.7，存在多元共線性的問題，因此我們試著刪除二者其一來解決多元共線性問題。

OLS Regression Results

Dep. Variable:	logSalary	R-squared:	0.645
Model:	OLS	Adj. R-squared:	0.643
Method:	Least Squares	F-statistic:	279.7
Date:	Sun, 09 Jun 2024	Prob (F-statistic):	1.85e-103
Time:	17:14:09	Log-Likelihood:	-600.65
No. Observations:	466	AIC:	1209.
Df Residuals:	462	BIC:	1226.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.8755	0.254	42.860	0.000	10.377	11.374
MP	0.0691	0.006	12.364	0.000	0.058	0.080
G	0.0202	0.002	9.187	0.000	0.016	0.025
Age	0.0759	0.010	7.816	0.000	0.057	0.095

Omnibus:	15.859	Durbin-Watson:	2.049
Prob(Omnibus):	0.000	Jarque-Bera (JB):	19.345
Skew:	-0.346	Prob(JB):	6.30e-05
Kurtosis:	3.720	Cond. No.	389.

Figure 27: 所有球員—回歸模型結果（Stepwise Forward 所選變數刪除 log 罰球命中數）

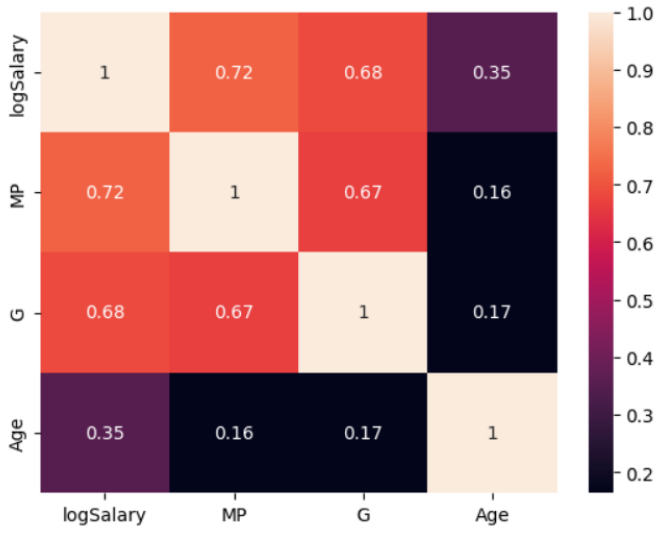


Figure 28: 所有球員—log 球員薪水與 Stepwise Forward 所選變數刪除 log 罰球命中數之熱力圖

Figure 27 爲刪除 log 罰球命中數變數的結果，可得 $R^2 = 0.645$、adjusted$R^2 = 0.643$，，二者差異之絕對值小於 0.06，且無發生 corr>0.7、F-test 的 p-value 很小但變數的 t-value 不拒絕、各變數係數與相關係數正負相反的情況。

OLS Regression Results

Dep. Variable:	logSalary	R-squared:	0.663
Model:	OLS	Adj. R-squared:	0.661
Method:	Least Squares	F-statistic:	303.1
Date:	Sun, 09 Jun 2024	Prob (F-statistic):	9.80e-109
Time:	17:15:18	Log-Likelihood:	-588.38
No. Observations:	466	AIC:	1185.
Df Residuals:	462	BIC:	1201.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.8620	0.247	43.953	0.000	10.376	11.348
G	0.0241	0.002	12.565	0.000	0.020	0.028
Age	0.0866	0.009	9.177	0.000	0.068	0.105
logFT	1.2827	0.094	13.642	0.000	1.098	1.467

Omnibus:	3.220	Durbin-Watson:	2.034
Prob(Omnibus):	0.200	Jarque-Bera (JB):	3.077
Skew:	0.141	Prob(JB):	0.215
Kurtosis:	3.281	Cond. No.	368.

Figure 29: 所有球員—回歸模型結果（Stepwise Forward 所選變數刪除上場時間）

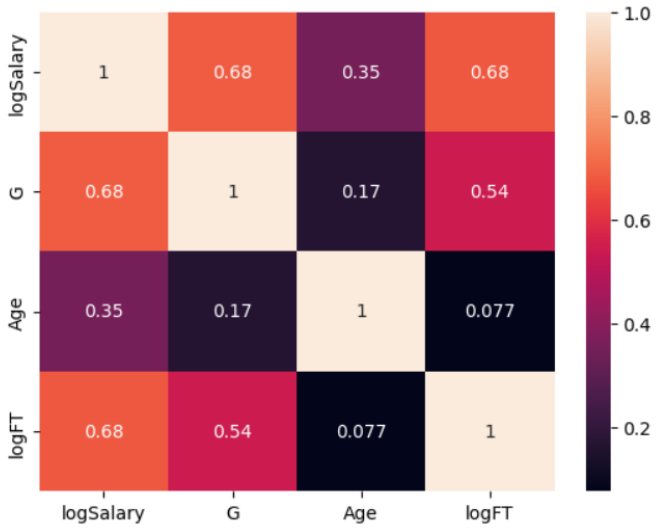


Figure 30: 所有球員—log 球員薪水與 Stepwise Forward 所選變數刪除上場時間之熱力圖

Figure 29 爲刪除上場時間變數的結果，可得 $R^2 = 0.663$、adjusted$R^2 = 0.661$，二者差異之絕對值小於 0.06，且無發生 corr$>$ 0.7、F-test 的 p-value 很小但變數的 t-value 不拒絕、各變數係數與相關係數正負相反的情況。

綜上所述，我們選擇 R^2 較佳的第二個模型（$R^2 = 0.663$, Stepwise Forward 所選

變數刪除上場時間)，作爲我們使用 Stepwise Forward 方法與下方使用 best subset 方法比較的模型。以下我們驗證該模型的殘差是否符合所有條件：

```
1. Zero mean
H0: Errors have zero mean.
H1: Errors do not have zero mean.
mean = -0.0000
std. dev. = 1.0001
Number of observation = 466
Hypothesized mean = 0
Significant level = 0.05
t-stat = -0.0002
t critical value one tail = -1.6481
p-value (one-tail) = 0.4999
t critical value two tail = -1.9651
p-value (two-tail) = 0.9998
Since the p_value = 0.9998 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that the errors do not have zero mean.
```

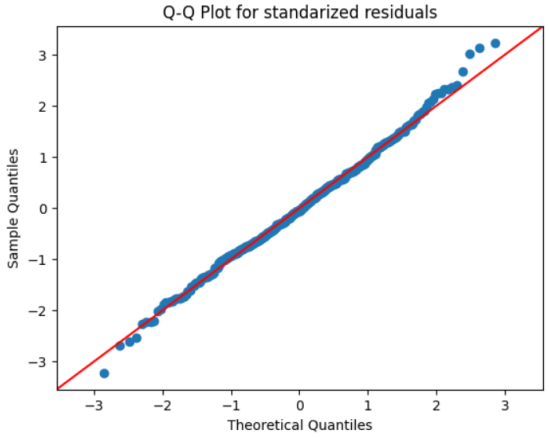
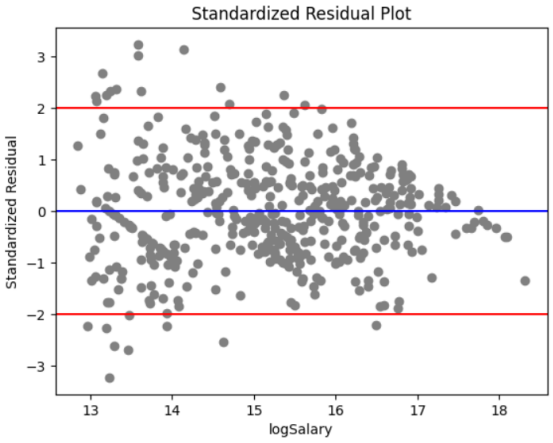


Figure 31: 所有球員—Stepwise Forward Q-Q Plot

```
3. Homoskedasticity
H0: Homoskedasticity
H1: Heteroskedasticity
```




```
Do not rejected H0. Heteroscedasticity does not appear to be a problem.
```

Figure 32: 所有球員—Stepwise Forward Homoscedasticity and Heteroscedasticity

```
2. Normality
H0: Errors are normally distributed.
H1: Errors are not normally distributed.
Shapiro-Wilk test for normality:
H0: The distribution is normal.
H1: The distribution is not normal.
For population = standarized residuals
Shapiro statistic = 0.996807 and p_value = 0.492368
Since the p_value = 0.4924 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that the distribution is not normal.
```

Figure 33: 所有球員—Stepwise Forward Shapiro Test

```
H0 : Randomness exists.
H1 : Randomness does not exist.
runs =  245
n1 =  233
n2 =  233
runs_exp =  234.0
stan_dev =  10.781904394196388
z =  1.020227929856344
pval_z =  0.30762037440778867
p_value for Z-statistic=  0.30762037440778867
Since the p_value = 0.3076 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that randomness does not exist.
```

Figure 34: 所有球員—Stepwise Forward Run Test Result

可得殘差分析條件皆符合，且 Durbin-Watson test = 2.034 無自相關的問題。

Best Subsets Regression:

此方法選出的最佳有效模型 (Figure 35) 包含四個自變數：Age、G、MP、logFT，$R^2 = 0.675$，adjusted $R^2 = 0.672$，二者差異之絕對值小於 0.06，且通過殘差分析條件 (Figure 36 至 40)、無多元共線性，Durbin-Watson test = 2.021 無自相關的問題。

OLS Regression Results

Dep. Variable:	logSalary	R-squared:	0.675
Model:	OLS	Adj. R-squared:	0.672
Method:	Least Squares	F-statistic:	239.1
Date:	Sun, 09 Jun 2024	Prob (F-statistic):	5.57e-111
Time:	01:25:31	Log-Likelihood:	-580.17
No. Observations:	466	AIC:	1170.
Df Residuals:	461	BIC:	1191.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.8303	0.243	44.533	0.000	10.352	11.308
Age	0.0820	0.009	8.764	0.000	0.064	0.100
G	0.0202	0.002	9.593	0.000	0.016	0.024
MP	0.0319	0.008	4.068	0.000	0.016	0.047
logFT	0.8808	0.135	6.508	0.000	0.615	1.147

Omnibus:	4.625	Durbin-Watson:	2.021
Prob(Omnibus):	0.099	Jarque-Bera (JB):	5.596
Skew:	-0.079	Prob(JB):	0.0609
Kurtosis:	3.513	Cond. No.	389.

Figure 35: 所有球員—Best Subsets 回歸模型結果

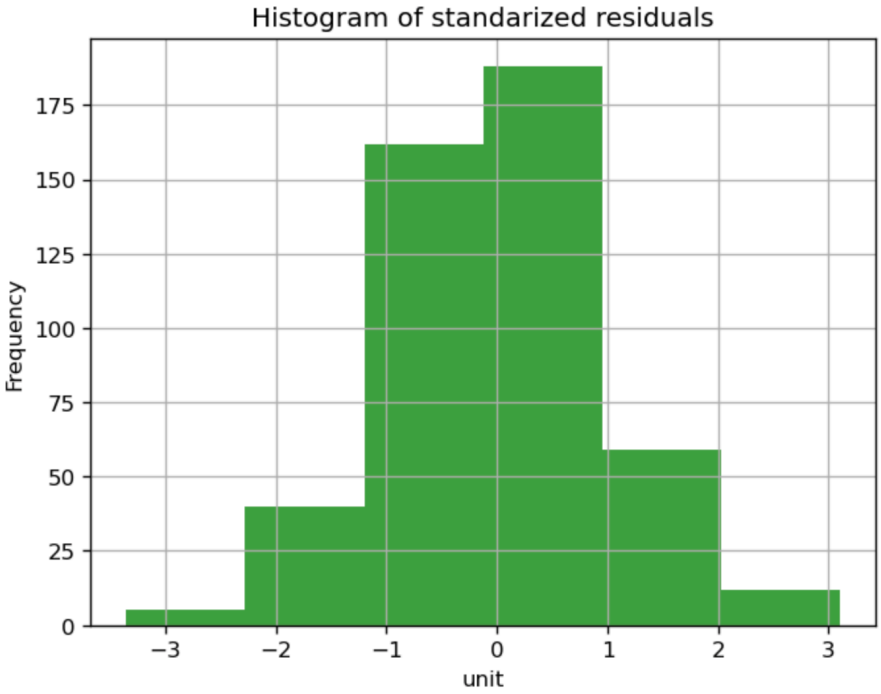


Figure 36: 所有球員—Best Subsets Normality Test

```
Shapiro statistic = 0.995413 and p_value = 0.187431
Since the p_value = 0.1874 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that the distribution is not normal.
```

Figure 37: 所有球員—Best Subsets Shapiro Test

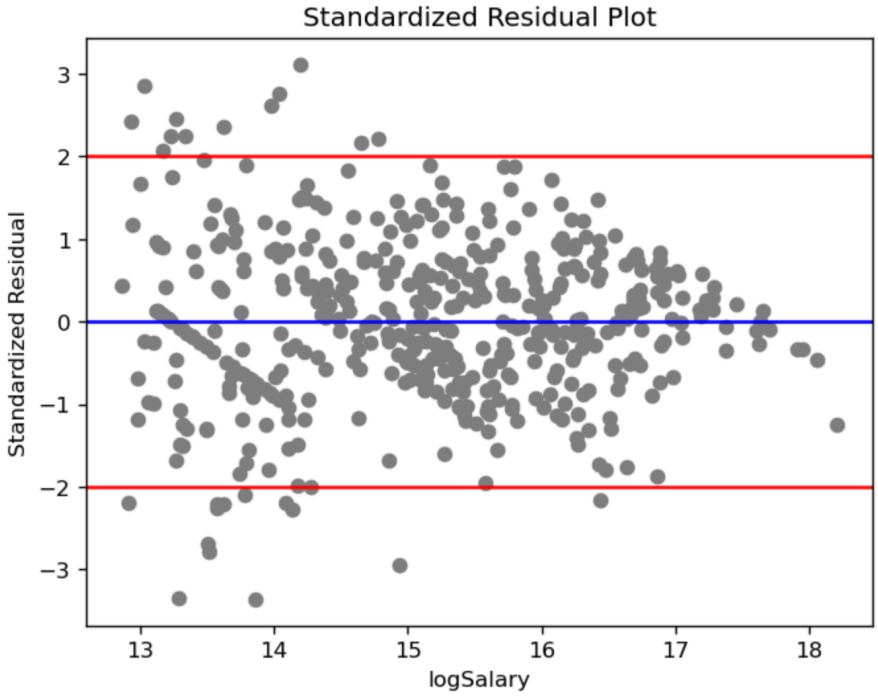


Figure 38: 所有球員—Best Subsets Homoscedasticity and Heteroscedasticity

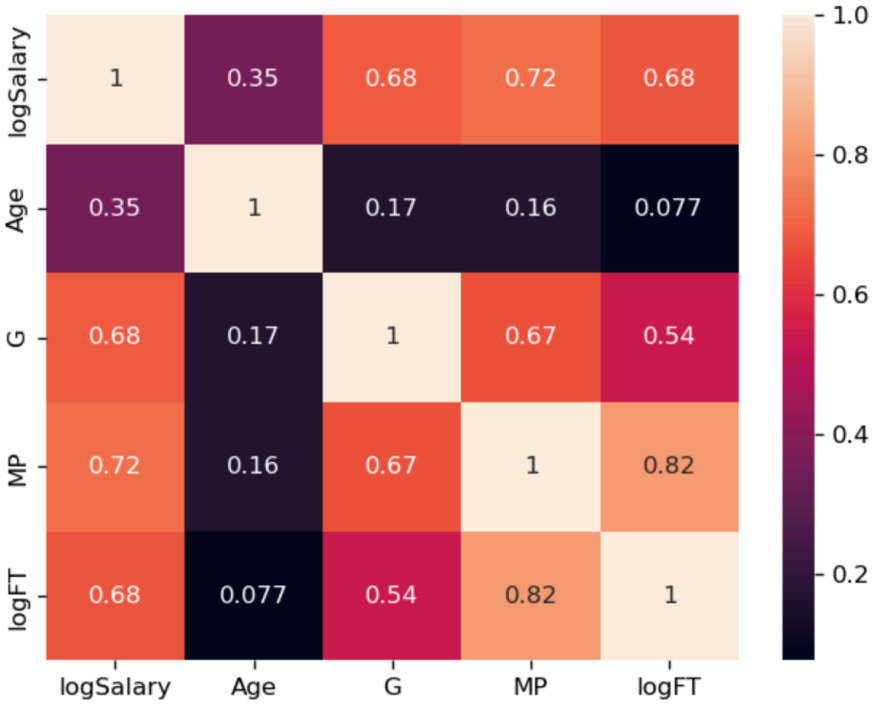


Figure 39: 所有球員—log 球員薪水與 Best Subsets 所選變數之熱力圖

```
runs =  232
n1 =  233
n2 =  233
runs_exp =  234.0
stan_dev =  10.781904394196388
z =  -0.185495987246608
pval_z =  0.8528400273659733
p_value for Z-statistic=  0.8528400273659733
Since the p_value = 0.8528 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that randomness does not exist.
```

Figure 40: 所有球員—Best Subsets Run Test Result

4.2.3 前鋒球員 (Forward)

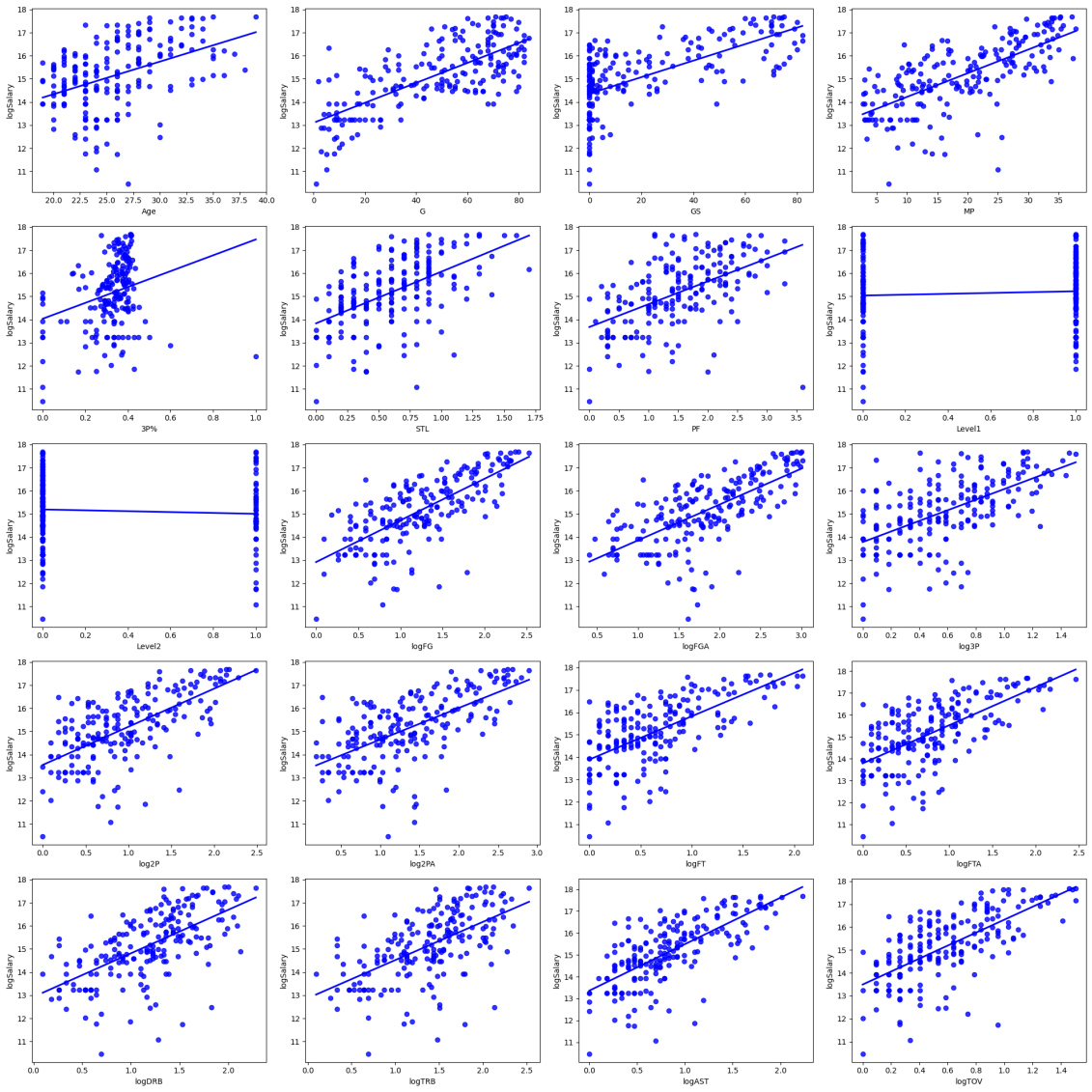


Figure 41: 前鋒球員的薪水與各變數的散佈圖

Stepwise Forward Regression:

```python
from stepwise_regression import step_reg
forwardselect = step_reg.forward_regression(X_data2, y_data_1, 0.1, verbose=False)
print(forwardselect)
```

```
['const', 'G', 'logAST', 'Age', 'logFT']
```

Figure 42: 前鋒球員-Stepwise Forward Regression 選擇的變數

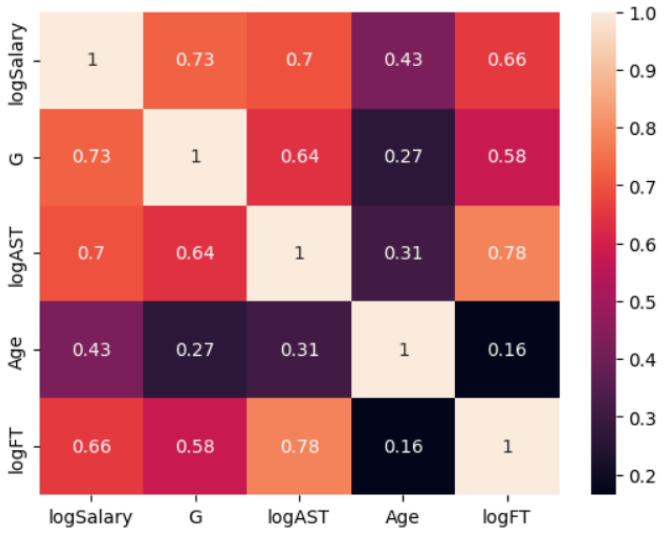


Figure 43: 前鋒球員-log 球員薪水與 Stepwise Forward 選擇變數之熱力圖

從 Figure 43 可得 corr(logAST,logFT)>0.7，存在多元共線性的問題，因此我們試著刪除二者其一來解決多元共線性問題。

OLS Regression Results

Dep. Variable:	logSalary	R-squared:	0.655
Model:	OLS	Adj. R-squared:	0.650
Method:	Least Squares	F-statistic:	124.7
Date:	Tue, 11 Jun 2024	Prob (F-statistic):	2.67e-45
Time:	17:02:05	Log-Likelihood:	-253.97
No. Observations:	201	AIC:	515.9
Df Residuals:	197	BIC:	529.2
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	11.2948	0.364	31.005	0.000	10.576	12.013
G	0.0267	0.003	8.227	0.000	0.020	0.033
logAST	1.0687	0.169	6.339	0.000	0.736	1.401
Age	0.0656	0.015	4.472	0.000	0.037	0.095

Omnibus:	3.445	Durbin-Watson:	2.064
Prob(Omnibus):	0.179	Jarque-Bera (JB):	3.028
Skew:	-0.271	Prob(JB):	0.220
Kurtosis:	3.261	Cond. No.	349.

Figure 44: 前鋒球員-回歸模型結果（Stepwise Forward 所選變數刪除 log 罰球命中數）

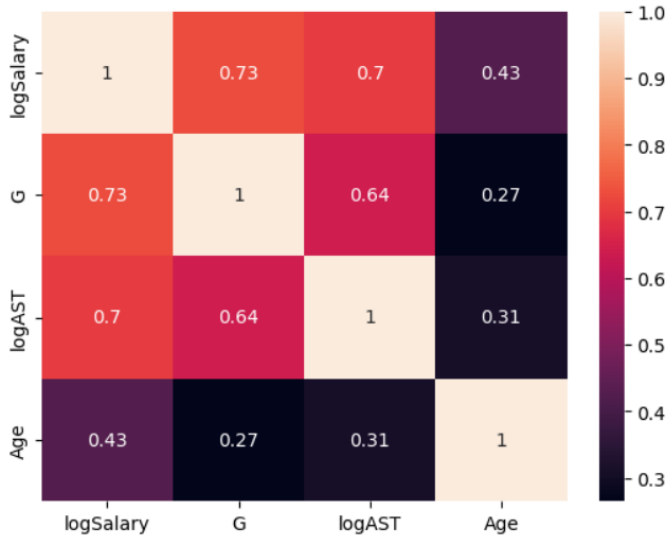


Figure 45: 前鋒球員-log 球員薪水與 Stepwise Forward 所選變數刪除 log 罰球命中數之熱力圖

Figure 44 爲刪除 log 罰球命中數變數的結果，可得 $R^2 = 0.655$、adjusted$R^2 = 0.650$，二者差異之絕對值小於 0.06，且無發生 corr> 0.7、F-test 的 p-value 很小但變數的 t-value 不拒絕、各變數係數與相關係數正負相反的情況。

OLS Regression Results

Dep. Variable:	logSalary	R-squared:	0.666
Model:	OLS	Adj. R-squared:	0.661
Method:	Least Squares	F-statistic:	130.9
Date:	Tue, 11 Jun 2024	Prob (F-statistic):	1.16e-46
Time:	17:03:27	Log-Likelihood:	-250.77
No. Observations:	201	AIC:	509.5
Df Residuals:	197	BIC:	522.7
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	11.0825	0.358	30.963	0.000	10.377	11.788
G	0.0273	0.003	8.867	0.000	0.021	0.033
Age	0.0818	0.014	5.763	0.000	0.054	0.110
logFT	1.0313	0.149	6.919	0.000	0.737	1.325

Omnibus:	3.703	Durbin-Watson:	2.122
Prob(Omnibus):	0.157	Jarque-Bera (JB):	3.848
Skew:	-0.157	Prob(JB):	0.146
Kurtosis:	3.600	Cond. No.	349.

Figure 46: 前鋒球員-回歸模型結果（Stepwise Forward 所選變數刪除 log 助攻數）

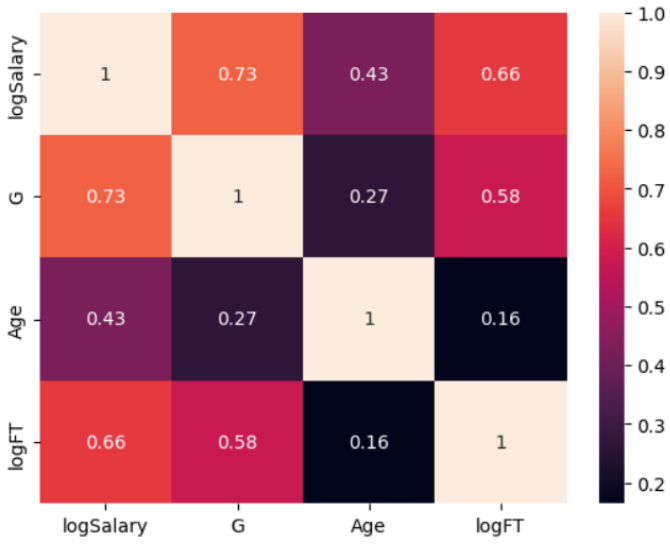


Figure 47: 前鋒球員-log 球員薪水與 Stepwise Forward 所選變數刪除 log 助攻數之熱力圖

Figure 46 爲刪除 log 助攻數變數的結果，可得 $R^2 = 0.666$、adjusted$R^2 = 0.661$，二者差異之絕對值小於 0.06，且無發生 corr$>$ 0.7、F-test 的 p-value 很小但變數的 t-value 不拒絕、各變數係數與相關係數正負相反的情況。

綜上所述，我們選擇 R^2 較佳的第二個模型（$R^2 = 0.666$, Stepwise Forward 所選

變數刪除 log 助攻數)，作爲我們使用 Stepwise Forward 方法與下方使用 Best Subsets 方法比較的模型。以下我們驗證該模型的殘差是否符合所有條件：

```
1. Zero mean
H0: Errors have zero mean.
H1: Errors do not have zero mean.
mean = -0.0001
std. dev. = 1.0001
Number of observation = 201
Hypothesized mean = 0
Significant level = 0.05
t-stat = -0.0019
t critical value one tail = -1.6525
p-value (one-tail) = 0.4992
t critical value two tail = -1.9719
p-value (two-tail) = 0.9985
Since the p_value = 0.9985 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that the errors do not have zero mean.
```

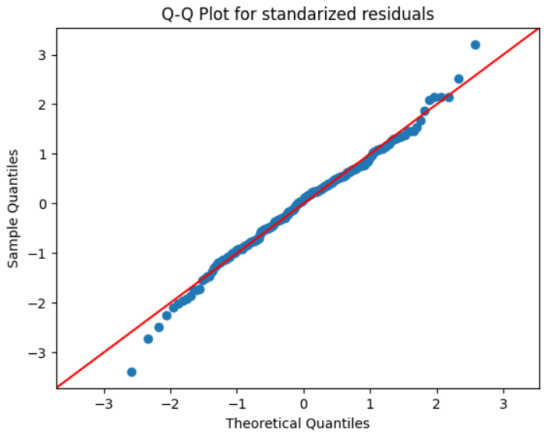
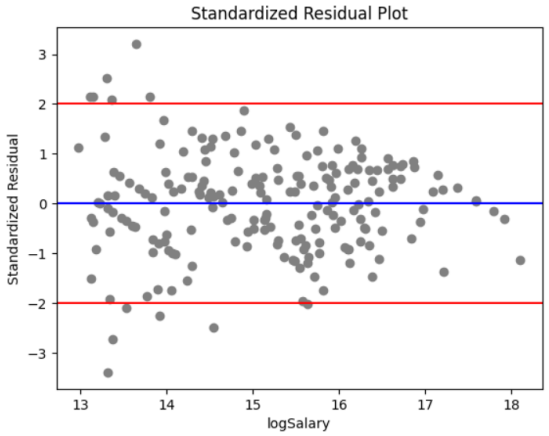


Figure 48: 前鋒球員-Stepwise Forward Q-Q Plot

```
3. Homoskedasticity
H0: Homoskedasticity
H1: Heteroskedasticity
```




```
Do not rejected H0. Heteroscedasticity does not appear to be a problem.
```

Figure 49: 前鋒球員-Stepwise Forward Scatter Plot for Standardized Residual

```
2. Normality
H0: Errors are normally distributed.
H1: Errors are not normally distributed.
Shapiro-Wilk test for normality:
H0: The distribution is normal.
H1: The distribution is not normal.
For population = standarized residuals
Shapiro statistic = 0.993143 and p_value = 0.475176
Since the p_value = 0.4752 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that the distribution is not normal.
```

Figure 50: 前鋒球員-Stepwise Forward Shapiro Test

```
4-1. Randomness
H0 : Randomness exists.
H1 : Randomness does not exist.
runs =  111
n1 =  101
n2 =  100
runs_exp =  101.49751243781094
stan_dev =  7.070803523557331
z =  1.343904908477551
pval_z =  0.17897913857498782
p_value for Z-statistic=  0.17897913857498782
Since the p_value = 0.1790 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that randomness does not exist.
```

Figure 51: 前鋒球員-Stepwise Forward Run Test Result

可得殘差分析條件皆符合，且 Durbin-Watson test = 2.122 無自相關的問題。

Best Subsets Regression:

此方法選出的最佳有效模型 (Figure 52) 包含四個自變數：Age、G、logFT、logAST，$R^2 = 0.676$，adjusted $R^2 = 0.669$，二者差異之絕對值小於 0.06，且通過殘差分析條件 (Figure 53 至 57)、無多元共線性，Durbin-Watson test = 2.085 無自相關的問題。

OLS Regression Results

Dep. Variable:	y	R-squared:	0.676
Model:	OLS	Adj. R-squared:	0.669
Method:	Least Squares	F-statistic:	102.2
Date:	Sun, 09 Jun 2024	Prob (F-statistic):	7.46e-47
Time:	01:26:42	Log-Likelihood:	-247.71
No. Observations:	201	AIC:	505.4
Df Residuals:	196	BIC:	521.9
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	11.1779	0.356	31.439	0.000	10.477	11.879
Age	0.0733	0.014	5.081	0.000	0.045	0.102
G	0.0247	0.003	7.699	0.000	0.018	0.031
logFT	0.7042	0.198	3.550	0.000	0.313	1.095
logAST	0.5432	0.221	2.460	0.015	0.108	0.979

Omnibus:	4.209	Durbin-Watson:	2.085
Prob(Omnibus):	0.122	Jarque-Bera (JB):	4.078
Skew:	-0.232	Prob(JB):	0.130
Kurtosis:	3.521	Cond. No.	354.

Figure 52: 前鋒球員—Best Subsets 回歸模型結果

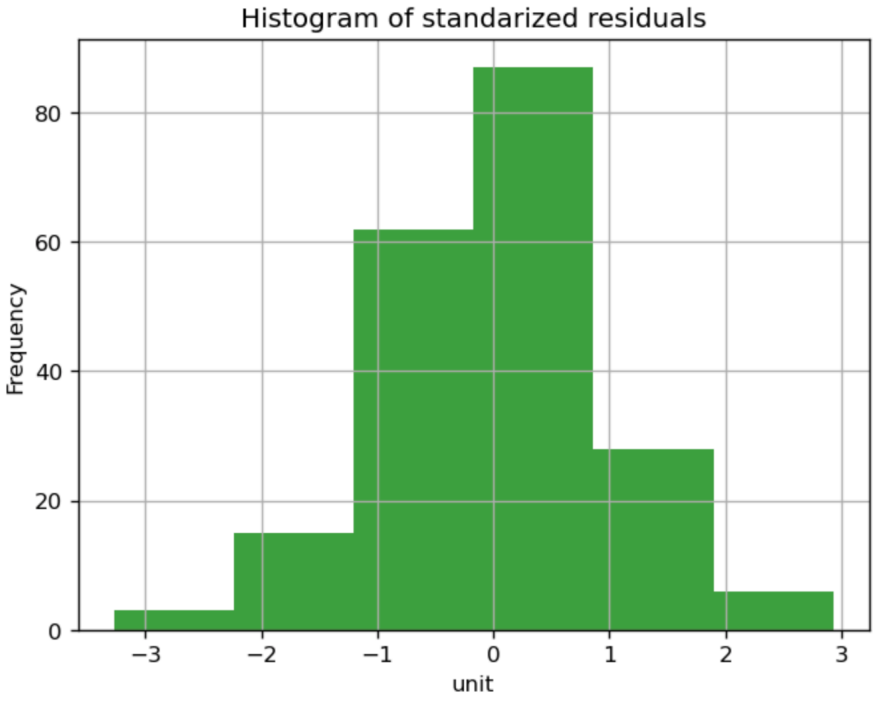


Figure 53: 前鋒球員—Best Subsets Normality Test

```
Shapiro statistic = 0.990888 and p_value = 0.236840
Since the p_value = 0.2368 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that the distribution is not normal.
```

Figure 54: 前鋒球員—Best Subsets Shapiro Test

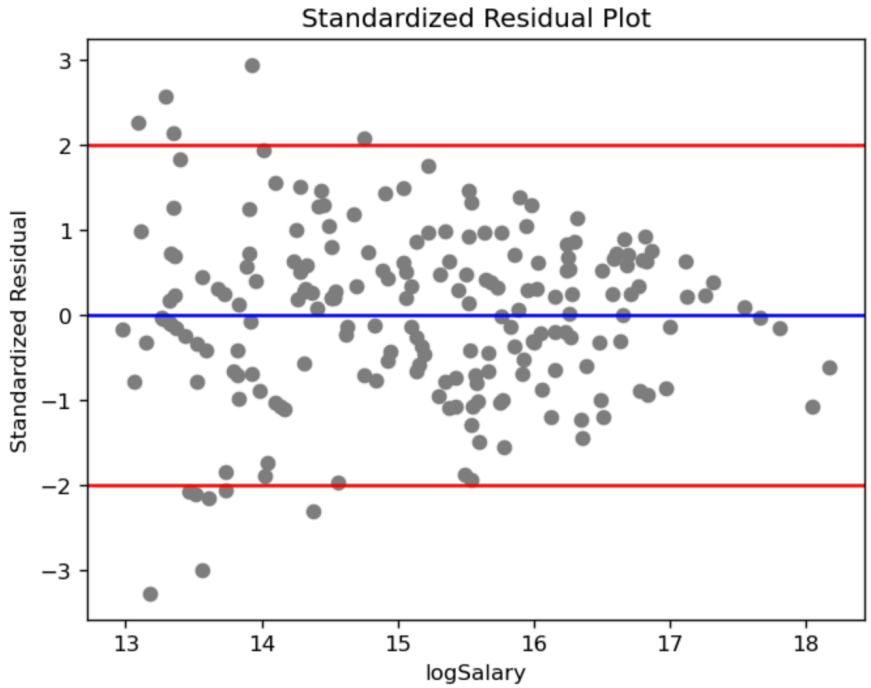


Figure 55: 前鋒球員—Best Subsets Homoscedasticity and Heteroscedasticity

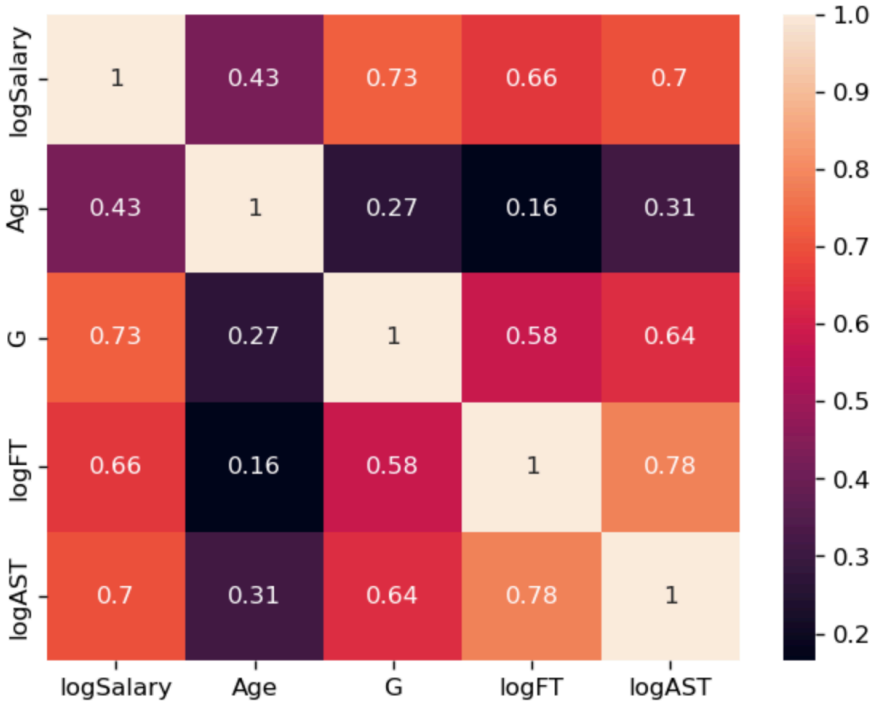


Figure 56: 前鋒球員—log 球員薪水與 Best Subsets 所選變數之熱力圖

```
runs =  109
n1 =  101
n2 =  100
runs_exp =  101.49751243781094
stan_dev =  7.070803523557331
z =  1.0610516240754697
pval_z =  0.2886664408883812
p_value for Z-statistic=  0.2886664408883812
Since the p_value = 0.2887 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that randomness does not exist.
```

Figure 57: 前鋒球員—Best Subsets Run Test Result

4.2.4 中鋒球員 (Center)

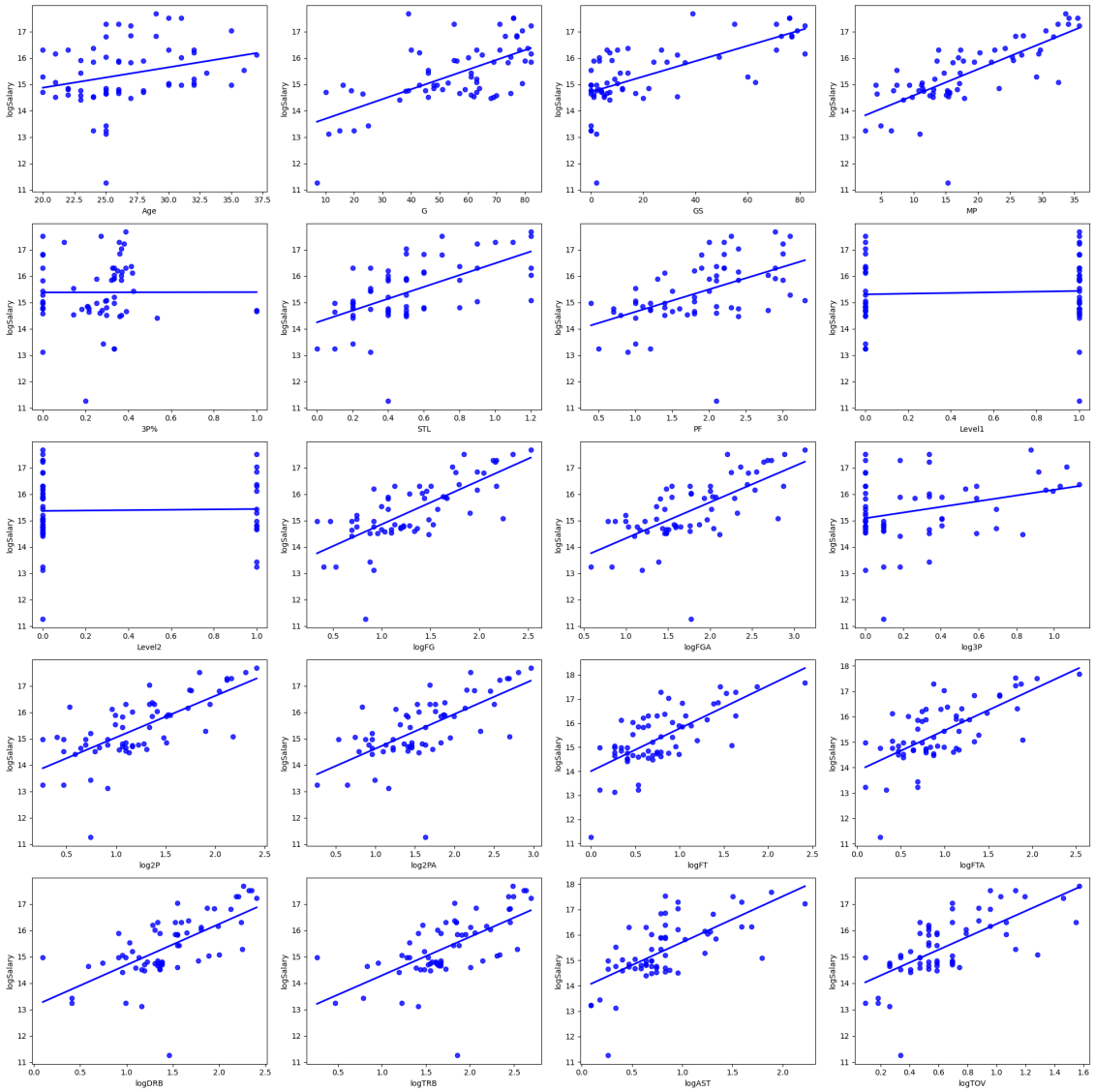


Figure 58: 中鋒球員的薪水與各變數的散佈圖

Stepwise Forward Regression:

```
from stepwise_regression import step_reg
forwardselect = step_reg.forward_regression(X_data2, y_data_1, 0.1, verbose=False)
print(forwardselect)
```

```
['const', 'MP', 'Age', 'logFG', 'logFGA', 'G']
```

Figure 59: 中鋒球員-Stepwise Forward Regression 選擇的變數

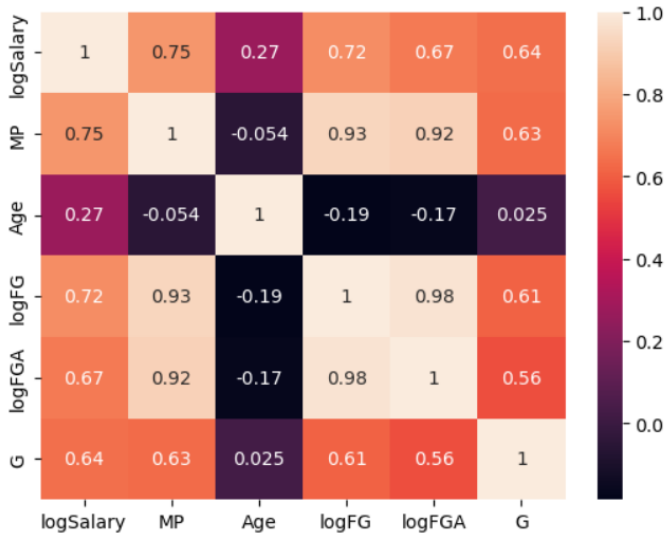


Figure 60: 中鋒球員-log 球員薪水與 Stepwise Forward 選擇變數之熱力圖

從 Figure 60 可得 corr(MP,logFG)、corr(MP,logFGA)、corr(logFGA,logFG)> 0.7，存在多元共線性的問題，因此我們試著刪除三者其二來解決多元共線性問題。

OLS Regression Results

Dep. Variable:	logSalary	R-squared:	0.669
Model:	OLS	Adj. R-squared:	0.652
Method:	Least Squares	F-statistic:	39.71
Date:	Tue, 11 Jun 2024	Prob (F-statistic):	3.56e-14
Time:	19:19:37	Log-Likelihood:	-65.649
No. Observations:	63	AIC:	139.3
Df Residuals:	59	BIC:	147.9
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	9.6413	0.691	13.960	0.000	8.259	11.023
Age	0.1020	0.022	4.618	0.000	0.058	0.146
logFGA	1.1203	0.190	5.885	0.000	0.739	1.501
G	0.0188	0.005	3.523	0.001	0.008	0.029

Omnibus:	29.911	Durbin-Watson:	2.176
Prob(Omnibus):	0.000	Jarque-Bera (JB):	81.033
Skew:	-1.382	Prob(JB):	2.53e-18
Kurtosis:	7.820	Cond. No.	501.

Figure 61: 中鋒球員-回歸模型結果（Stepwise Forward 所選變數刪除上場時間、log 投籃成功次數）

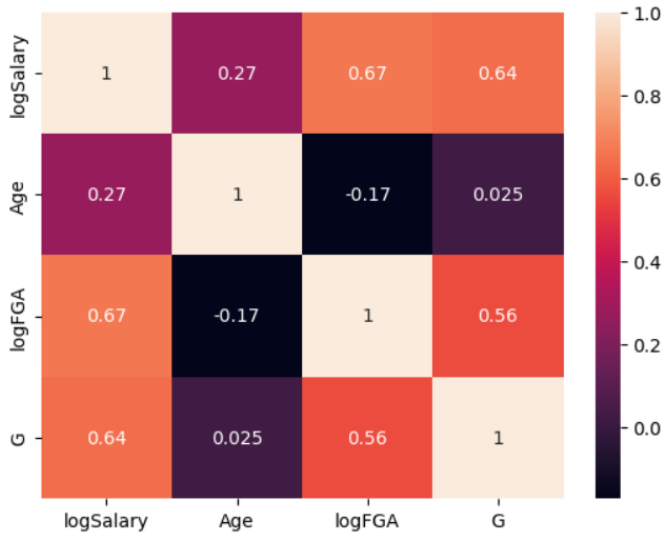


Figure 62: 中鋒球員-log 球員薪水與 Stepwise Forward 所選變數刪除上場時間、log 投籃成功次數之熱力圖

Figure 61 爲刪除上場時間、log 投籃成功次數變數的結果，可得 $R^2 = 0.669$、adjusted$R^2 = 0.652$，二者差異之絕對值小於 0.06，且無發生 corr> 0.7、F-test 的 p-value 很小但變數的 t-value 不拒絕、各變數係數與相關係數正負相反的情況。

OLS Regression Results

Dep. Variable:	logSalary	R-squared:	0.719
Model:	OLS	Adj. R-squared:	0.704
Method:	Least Squares	F-statistic:	50.21
Date:	Tue, 11 Jun 2024	Prob (F-statistic):	3.03e-16
Time:	19:32:11	Log-Likelihood:	-60.523
No. Observations:	63	AIC:	129.0
Df Residuals:	59	BIC:	137.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	9.7271	0.620	15.696	0.000	8.487	10.967
Age	0.1101	0.021	5.366	0.000	0.069	0.151
logFG	1.4890	0.208	7.154	0.000	1.073	1.905
G	0.0138	0.005	2.684	0.009	0.004	0.024

Omnibus:	20.682	Durbin-Watson:	2.264
Prob(Omnibus):	0.000	Jarque-Bera (JB):	36.135
Skew:	-1.106	Prob(JB):	1.42e-08
Kurtosis:	5.979	Cond. No.	488.

Figure 63: 中鋒球員-回歸模型結果（Stepwise Forward 所選變數刪除上場時間、log 投籃出手數）

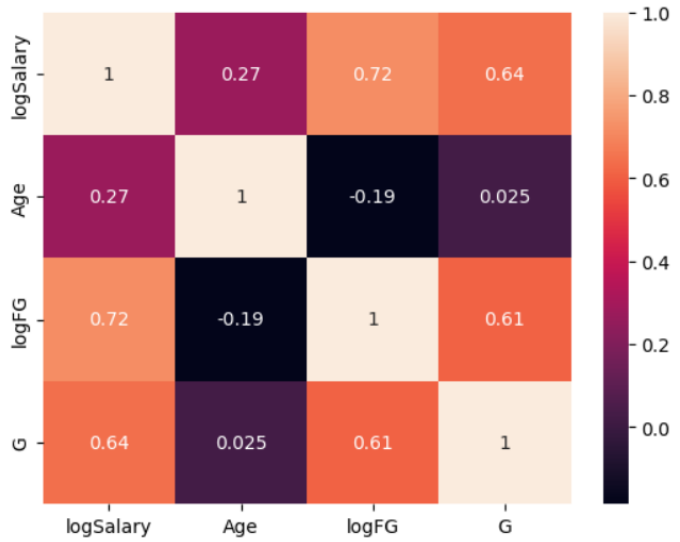


Figure 64: 中鋒球員-log 球員薪水與 Stepwise Forward 所選變數刪除上場時間、log 投籃出手數之熱力圖

Figure 63 爲刪除上場時間、log 投籃出手數變數的結果，可得 $R^2 = 0.719$、adjusted$R^2 = 0.704$，二者差異之絕對值小於 0.06，且無發生 corr> 0.7、F-test 的 p-value 很小但變數的 t-value 不拒絕、各變數係數與相關係數正負相反的情況。

OLS Regression Results

Dep. Variable:	logSalary	R-squared:	0.690
Model:	OLS	Adj. R-squared:	0.674
Method:	Least Squares	F-statistic:	43.70
Date:	Tue, 11 Jun 2024	Prob (F-statistic):	5.32e-15
Time:	19:40:39	Log-Likelihood:	-63.602
No. Observations:	63	AIC:	135.2
Df Residuals:	59	BIC:	143.8
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.8513	0.609	17.832	0.000	9.634	12.069
MP	0.0806	0.013	6.397	0.000	0.055	0.106
Age	0.0849	0.021	4.057	0.000	0.043	0.127
G	0.0147	0.005	2.694	0.009	0.004	0.026

Omnibus:	31.800	Durbin-Watson:	2.136
Prob(Omnibus):	0.000	Jarque-Bera (JB):	96.897
Skew:	-1.421	Prob(JB):	9.10e-22
Kurtosis:	8.369	Cond. No.	472.

Figure 65: 中鋒球員-回歸模型結果（Stepwise Forward 所選變數刪除 log 投籃出手數、log 投籃出手數）

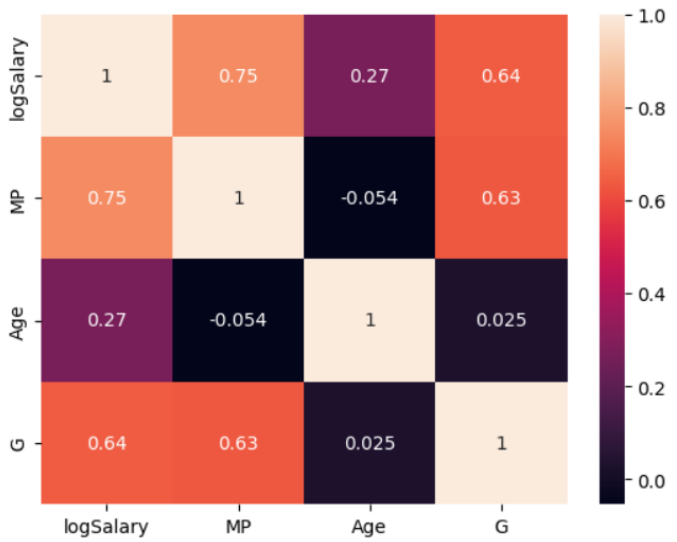


Figure 66: 中鋒球員-log 球員薪水與 Stepwise Forward 所選變數刪除 log 投籃出手數、log 投籃出手數之熱力圖

Figure 65 爲刪除 log 投籃出手數、log 投籃出手數變數的結果，可得 $R^2 = 0.690$、adjusted$R^2 = 0.674$，二者差異之絕對值小於 0.06，且無發生 corr$>$ 0.7、F-test 的 p-value 很小但變數的 t-value 不拒絕、各變數係數與相關係數正負相反的情況。

綜上所述，我們選擇 R^2 較佳的第二個模型（$R^2 = 0.719$, Stepwise Forward 刪除上場時間、log 投籃出手數），作爲我們使用 Stepwise Forward 方法與下方使用 Best Subsets 方法比較的模型。以下我們驗證該模型的殘差是否符合所有條件：

```
1. Zero mean
H0: Errors have zero mean.
H1: Errors do not have zero mean.
mean = 0.0005
std. dev. = 1.0072
Number of observation = 63
Hypothesized mean = 0
Significant level = 0.05
t-stat = 0.0036
t critical value one tail = 1.6698
p-value (one-tail) = 0.4986
t critical value two tail = 1.9990
p-value (two-tail) = 0.9971
Since the p_value = 0.9971 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that the errors do not have zero mean.
```

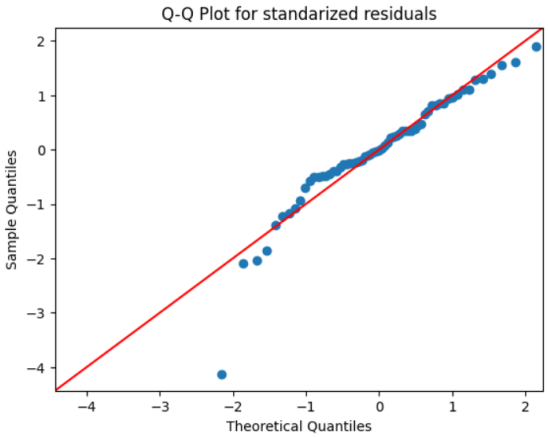


Figure 67: 中鋒球員-Stepwise Forward Q-Q Plot

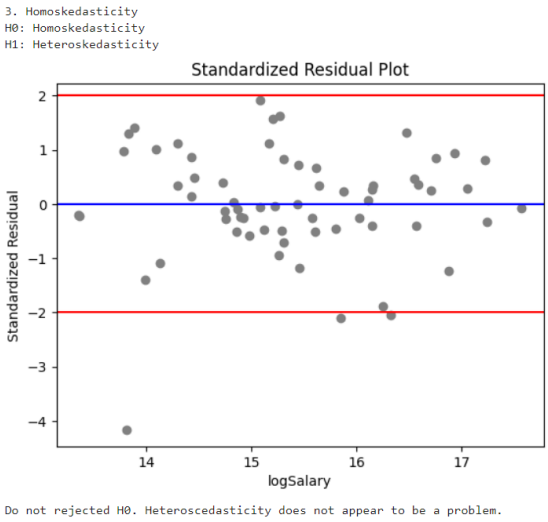


Figure 68: 中鋒球員-Stepwise Forward Scatter Plot for Standardized Residual

```
2. Normality
H0: Errors are normally distributed.
H1: Errors are not normally distributed.
Shapiro-Wilk test for normality:
H0: The distribution is normal.
H1: The distribution is not normal.
For population = standarized residuals
Shapiro statistic = 0.929480 and p_value = 0.001399
Since the p_value = 0.0014 < 0.05, we reject the null hypothesis.
That is, we have sufficient evidence to claim that the distribution is not normal.
```

Figure 69: 中鋒球員-Stepwise Forward Shapiro Test

```
4-1. Randomness
H0 : Randomness exists.
H1 : Randomness does not exist.
runs =  38
n1 =  32
n2 =  31
runs_exp =  32.492063492063494
stan_dev =  3.9354837377299643
z =  1.3995576846452813
pval_z =  0.1616458129961914
p_value for Z-statistic=  0.1616458129961914
Since the p_value = 0.1616 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that randomness does not exist.
```

Figure 70: 中鋒球員-Stepwise Forward Run Test Result

可得殘差分析條件皆符合，且 Durbin-Watson test = 2.264 無自相關的問題。

Best Subsets Regression:

此方法選出的模型若包含三個以上變數，會有多元共線性的問題，選擇一或二個變數則都拒絕殘差分析中的常態分佈檢定，故在中鋒這個位置，我們沒有使用 best subset 方法找出有效的模型。

4.2.5 後衛球員 (Guard)

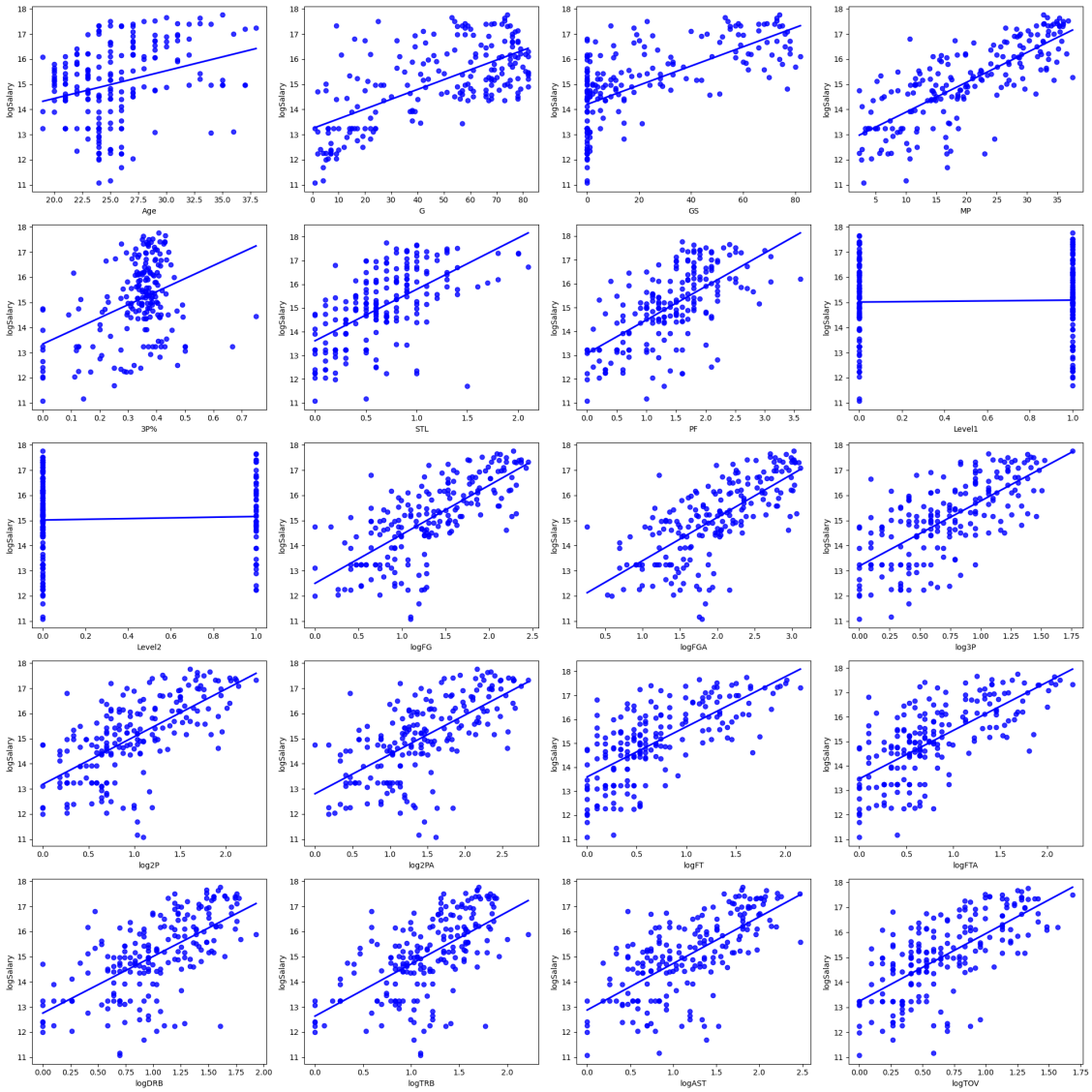


Figure 71: 後衛球員的薪水與各變數的散佈圖

Stepwise Forward Regression:

```
from stepwise_regression import step_reg
forwardselect = step_reg.forward_regression(X_data2, y_data_1, 0.1, verbose=False)
print(forwardselect)
```

```
['const', 'MP', 'G', 'Age', 'logFTA', 'logTRB']
```

Figure 72: 後衛球員-Stepwise Forward Regression 選擇的變數

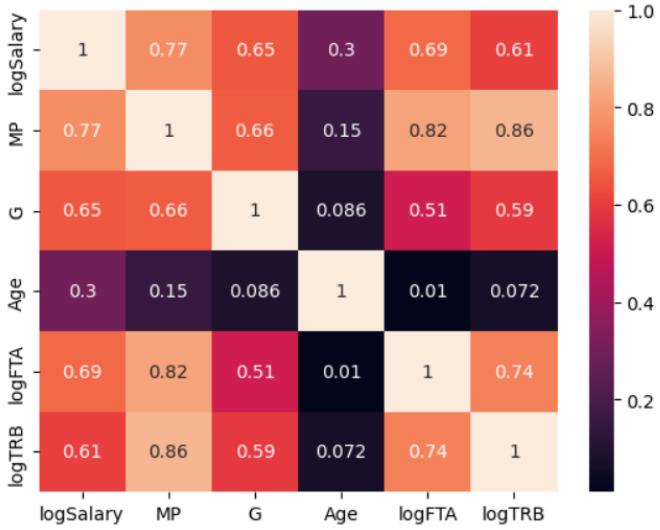


Figure 73: 後衛球員-log 球員薪水與 Stepwise Forward 選擇變數之熱力圖

從 Figure 73 可得 corr(MP,logFTA)、corr(MP,logTRB)、corr(logTRB,logFTA)> 0.7，存在多元共線性的問題，因此我們試著刪除三者其二來解決多元共線性問題。

OLS Regression Results

Dep. Variable:	logSalary	R-squared:	0.558
Model:	OLS	Adj. R-squared:	0.551
Method:	Least Squares	F-statistic:	83.25
Date:	Tue, 11 Jun 2024	Prob (F-statistic):	6.97e-35
Time:	20:09:30	Log-Likelihood:	-293.75
No. Observations:	202	AIC:	595.5
Df Residuals:	198	BIC:	608.7
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.2770	0.484	21.224	0.000	9.322	11.232
G	0.0260	0.003	7.417	0.000	0.019	0.033
Age	0.0874	0.018	4.928	0.000	0.052	0.122
logTRB	1.1375	0.199	5.703	0.000	0.744	1.531

Omnibus:	0.285	Durbin-Watson:	2.207
Prob(Omnibus):	0.867	Jarque-Bera (JB):	0.383
Skew:	0.083	Prob(JB):	0.826
Kurtosis:	2.867	Cond. No.	386.

Figure 74: 後衛球員-回歸模型結果（Stepwise Forward 所選變數刪除上場時間、log 罰球出手數）

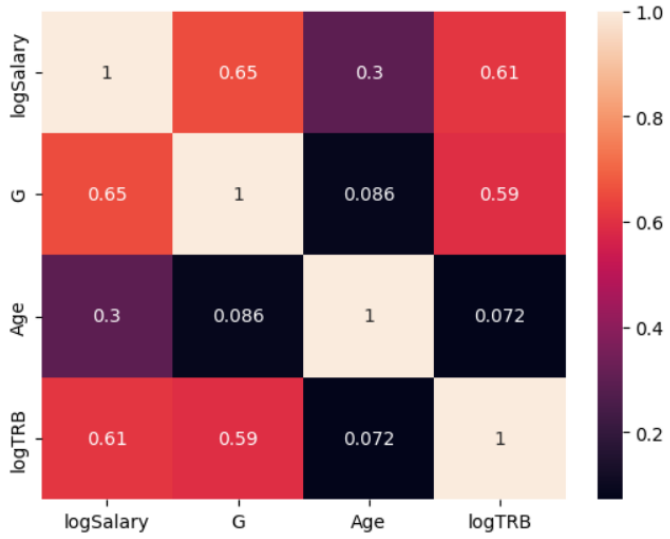


Figure 75: 後衛球員-log 球員薪水與 Stepwise Forward 所選變數刪除上場時間、log 罰球出手數之熱力圖

Figure 74 爲刪除上場時間、log 罰球出手數變數的結果，可得 $R^2 = 0.558$、adjusted$R^2 = 0.551$，二者差異之絕對值小於 0.06，且無發生 corr> 0.7、F-test 的 p-value 很小但變數的 t-value 不拒絕、各變數係數與相關係數正負相反的情況。

OLS Regression Results

Dep. Variable:	logSalary	R-squared:	0.663
Model:	OLS	Adj. R-squared:	0.658
Method:	Least Squares	F-statistic:	130.0
Date:	Tue, 11 Jun 2024	Prob (F-statistic):	1.47e-46
Time:	20:14:58	Log-Likelihood:	-266.23
No. Observations:	202	AIC:	540.5
Df Residuals:	198	BIC:	553.7
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.4024	0.412	25.232	0.000	9.589	11.215
G	0.0225	0.003	7.819	0.000	0.017	0.028
Age	0.0963	0.015	6.224	0.000	0.066	0.127
logFTA	1.4105	0.138	10.233	0.000	1.139	1.682

Omnibus:	4.046	Durbin-Watson:	2.131
Prob(Omnibus):	0.132	Jarque-Bera (JB):	3.803
Skew:	0.334	Prob(JB):	0.149
Kurtosis:	3.078	Cond. No.	375.

Figure 76: 後衛球員-回歸模型結果（Stepwise Forward 所選變數刪除上場時間、log 總籃板數）

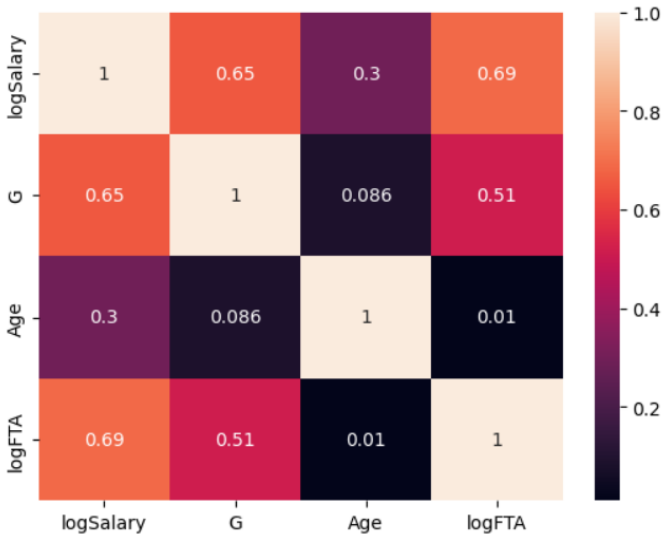


Figure 77: 後衛球員-log 球員薪水與 Stepwise Forward 所選變數刪除上場時間、log 總籃板數之熱力圖

Figure 76 爲刪除上場時間、log 總籃板數變數的結果，可得 $R^2 = 0.663$、adjusted$R^2 = 0.658$，二者差異之絕對值小於 0.06，且無發生 corr> 0.7、F-test 的 p-value 很小但變數的 t-value 不拒絕、各變數係數與相關係數正負相反的情況。

OLS Regression Results

Dep. Variable:	logSalary	R-squared:	0.658
Model:	OLS	Adj. R-squared:	0.653
Method:	Least Squares	F-statistic:	127.1
Date:	Tue, 11 Jun 2024	Prob (F-statistic):	6.52e-46
Time:	20:26:49	Log-Likelihood:	-267.76
No. Observations:	202	AIC:	543.5
Df Residuals:	198	BIC:	556.7
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.8161	0.412	26.270	0.000	10.004	11.628
MP	0.0870	0.009	10.009	0.000	0.070	0.104
G	0.0158	0.003	4.785	0.000	0.009	0.022
Age	0.0694	0.016	4.415	0.000	0.038	0.100

Omnibus:	0.380	Durbin-Watson:	2.175
Prob(Omnibus):	0.827	Jarque-Bera (JB):	0.219
Skew:	-0.074	Prob(JB):	0.896
Kurtosis:	3.064	Cond. No.	394.

Figure 78: 後衛球員-回歸模型結果（Stepwise Forward 所選變數刪除 log 罰球出手數、log 總籃板數）

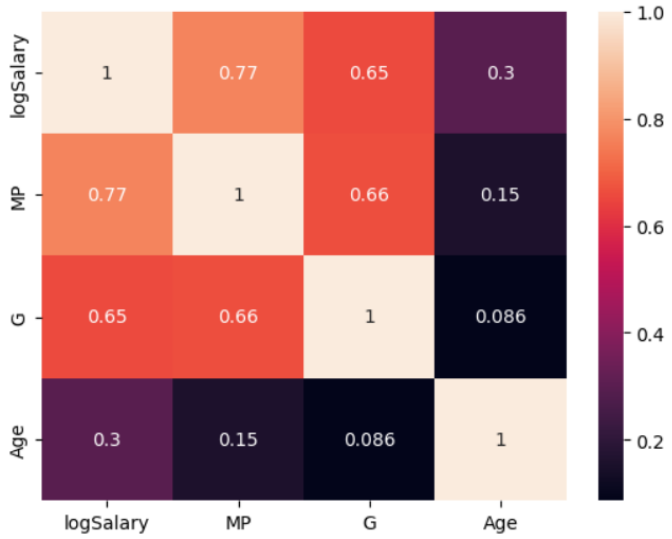


Figure 79: 後衛球員-log 球員薪水與 Stepwise Forward 所選變數刪除 log 罰球出手數、log 總籃板數之熱力圖

Figure 78 爲刪除 log 罰球出手數、log 總籃板數變數的結果，可得 $R^2 = 0.658$、adjusted$R^2 = 0.653$，二者差異之絕對值小於 0.06，且無發生 corr> 0.7、F-test 的 p-value 很小但變數的 t-value 不拒絕、各變數係數與相關係數正負相反的情況。

綜上所述，我們選擇 R^2 較佳的第二個模型（$R^2 = 0.663$, Stepwise Forward 所選變數刪除上場時間、log 總籃板數），作爲我們使用 Stepwise Forward 方法與下方使用 Best Subsets 方法比較的模型。以下我們驗證該模型的殘差是否符合所有條件：

```
1. Zero mean
H0: Errors have zero mean.
H1: Errors do not have zero mean.
mean = -0.0001
std. dev. = 1.0001
Number of observation = 202
Hypothesized mean = 0
Significant level = 0.05
t-stat = -0.0015
t critical value one tail = -1.6525
p-value (one-tail) = 0.4994
t critical value two tail = -1.9718
p-value (two-tail) = 0.9988
Since the p_value = 0.9988 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that the errors do not have zero mean.
```

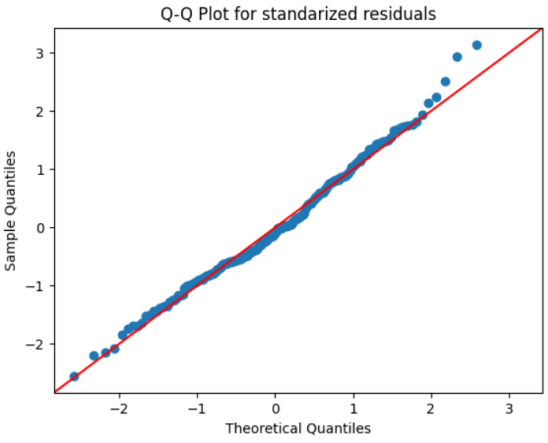


Figure 80: 後衛球員-Stepwise Forward Q-Q Plot

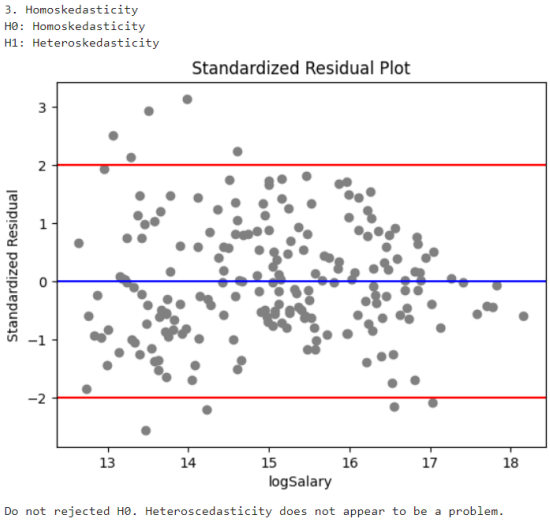


Figure 81: 後衛球員-Stepwise Forward Scatter Plot for Standardized Residual

```
2. Normality
H0: Errors are normally distributed.
H1: Errors are not normally distributed.
Shapiro-Wilk test for normality:
H0: The distribution is normal.
H1: The distribution is not normal.
For population = standarized residuals
Shapiro statistic = 0.990091 and p_value = 0.178835
Since the p_value = 0.1788 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that the distribution is not normal.
```

Figure 82: 後衛球員-Stepwise Forward Shapiro Test

```
4-1. Randomness
H0 : Randomness exists.
H1 : Randomness does not exist.
runs =  101
n1 =  101
n2 =  101
runs_exp =  102.0
stan_dev =  7.088635709281827
z =  -0.14107086906590566
pval_z =  0.8878139565934929
p_value for Z-statistic=  0.8878139565934929
Since the p_value = 0.8878 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that randomness does not exist.
```

Figure 83: 後衛球員-Stepwise Forward Run Test Result

可得殘差分析條件皆符合，且 Durbin-Watson test = 2.131 無自相關的問題。

Best Subsets Regression:

此方法選出的最佳有效模型 (Figure 84) 包含四個自變數：Age、G、MP、logFTA，$R^2 = 0.685$，adjusted $R^2 = 0.678$，二者差異之絕對值小於 0.06，且通過殘差分析條件 (Figure 85 至 89)、無多元共線性，Durbin-Watson test = 2.142 無自相關的問題。

OLS Regression Results

Dep. Variable:	y	R-squared:	0.685
Model:	OLS	Adj. R-squared:	0.678
Method:	Least Squares	F-statistic:	107.0
Date:	Sun, 09 Jun 2024	Prob (F-statistic):	2.81e-48
Time:	01:27:30	Log-Likelihood:	-259.56
No. Observations:	202	AIC:	529.1
Df Residuals:	197	BIC:	545.7
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.5466	0.402	26.247	0.000	9.754	11.339
Age	0.0826	0.015	5.338	0.000	0.052	0.113
G	0.0168	0.003	5.264	0.000	0.011	0.023
MP	0.0471	0.013	3.667	0.000	0.022	0.073
logFTA	0.8384	0.205	4.081	0.000	0.433	1.244

Omnibus:	1.201	Durbin-Watson:	2.142
Prob(Omnibus):	0.548	Jarque-Bera (JB):	0.930
Skew:	0.154	Prob(JB):	0.628
Kurtosis:	3.123	Cond. No.	402.

Figure 84: 後衛球員-Best Subsets 回歸模型結果

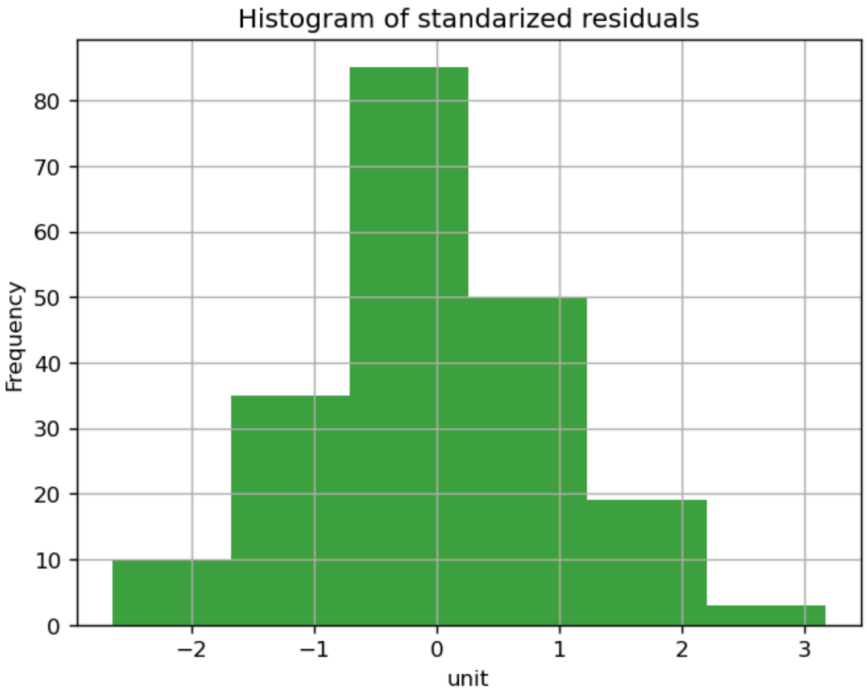


Figure 85: 後衛球員-Best Subsets Normality Test

```
Shapiro statistic = 0.995964 and p_value = 0.874441
Since the p_value = 0.8744 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that the distribution is not normal.
```

Figure 86: 後衛球員-Best Subsets Shapiro Test

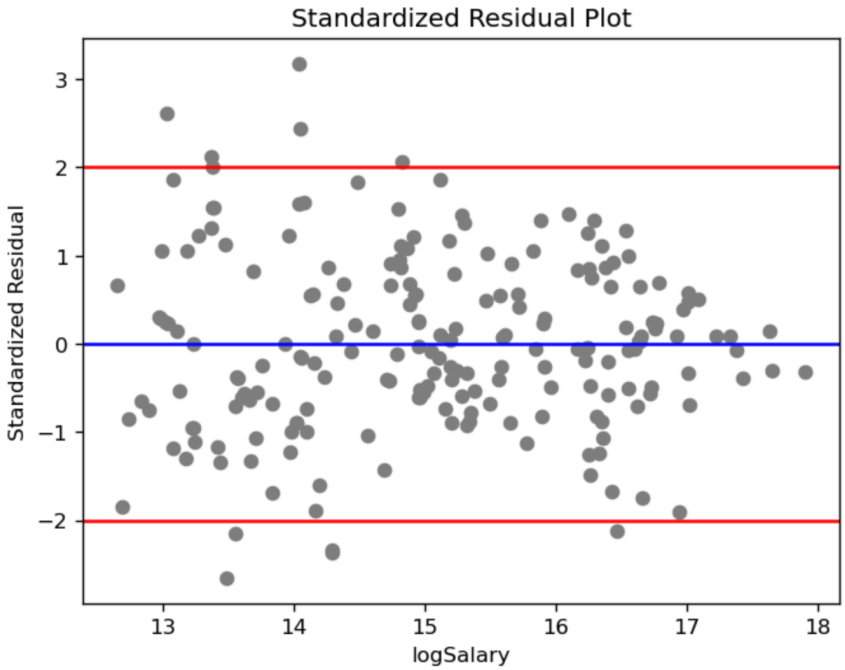


Figure 87: 後衛球員-Best Subsets Homoscedasticity and Heteroscedasticity

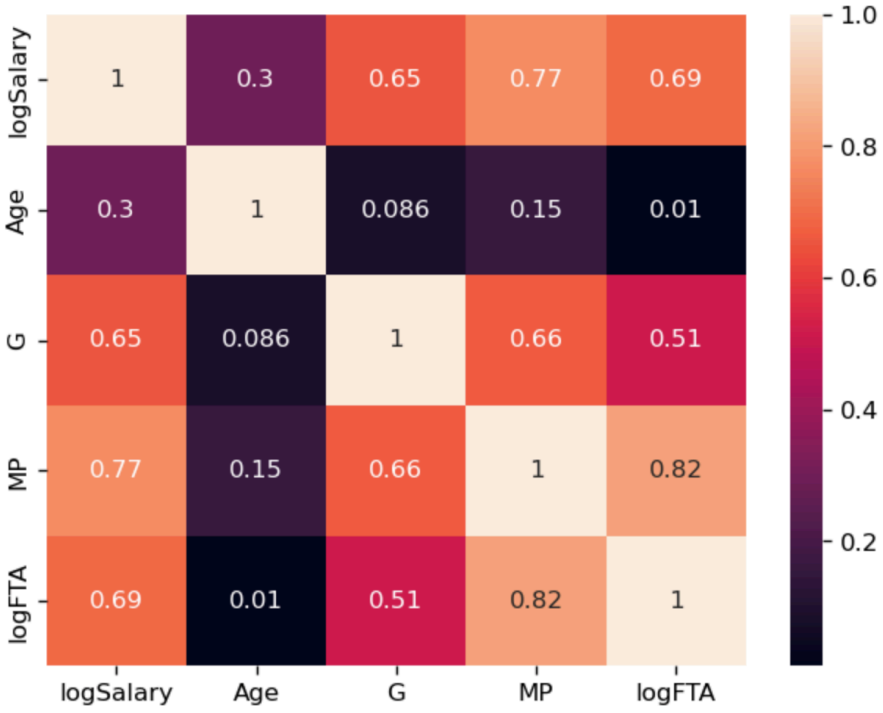


Figure 88: 後衛球員-log 球員薪水與 Best Subsets 所選變數之熱力圖

```
runs =  99
n1 =  101
n2 =  101
runs_exp =  102.0
stan_dev =  7.088635709281827
z =  -0.42321260719771703
pval_z =  0.6721401493594235
p_value for Z-statistic=  0.6721401493594235
Since the p_value = 0.6721 > 0.05, we do not reject the null hypothesis.
That is, we do not have sufficient evidence to claim that randomness does not exist.
```

Figure 89: 後衛球員-Best Subsets Run Test Result

4.2.6 小結

在所有球員的模型中，Best Subsets 的 $R^2 = 0.675$ 最好，其變數有 Age、G、MP、logFT。在前鋒球員的模型中，Best Subsets 的 $R^2 = 0.676$ 最好，其變數有 Age、G、logFT、logAST。在中鋒球員的模型中，Stepwise Forward 的 $R^2 = 0.719$ 最好，其變數有 Age、logFGA、G。在後衛球員的模型中，Best Subsets 的 $R^2 = 0.685$ 最好，其變數有 Age、G、MP、logFTA。

5 結論

於 NBA 球隊薪資總額與地區的影響中，我們使用 Forward/Backward Stepwise 找出沒有多元共線性與自回歸的變數，選擇以球隊勝場數、球隊收入爲自變數，建立我們的線性回歸模型，此模型通過所有檢定，且無多元共線性與自回歸，預測模型如下：

球隊 (by stepwise regression, $R^2 = 0.600, \text{Adj. } R^2 = 0.568$)：

$$\text{球隊總薪資} = 105.7787 + 0.5346\ \text{球隊勝場數} + 0.1045\ \text{球隊收入} \tag{1}$$

在球員位置對薪資的影響中，我們篩選出了適合的變數之後，分別對全體球員、後衛、前鋒、中鋒四個群體進行迴迴歸分析，預測模型如下：

後衛 (by best subset regression, $R^2 = 0.685, \text{Adj. } R^2 = 0.678$)：

$$\begin{aligned}\log(\text{球員薪資}) =& 10.5466 + 0.0826\ \text{球員年紀} + 0.0168\ \text{出賽場次} \\ &+ 0.0471\ \text{場均上場時間} + 0.8384 \log(\text{場均罰球出手數})\end{aligned} \tag{2}$$

前鋒 (by best subset regression, $R^2 = 0.676, \text{Adj. } R^2 = 0.669$)：

$$\begin{aligned}\log(\text{球員薪資}) =& 11.1779 + 0.0733\ \text{球員年紀} + 0.0247\ \text{出賽場次} \\ &+ 0.7042 \log(\text{場均罰球命中數}) + 0.7042 \log(\text{場均助攻數})\end{aligned} \tag{3}$$

中鋒 (by forward stepwise regression, $R^2 = 0.719, \text{Adj. } R^2 = 0.704$)：

$$\begin{aligned}\log(\text{球員薪資}) =& 9.7271 + 0.1101\ \text{球員年紀} + 0.0138\ \text{出賽場次} \\ &+ 1.4890 \log(\text{場均投籃命中數})\end{aligned} \tag{4}$$

全體球員 (by best subset regression, $R^2 = 0.675, \text{Adj. } R^2 = 0.672$)：

$$\begin{aligned}\log(\text{球員薪資}) =& 10.8303 + 0.0820\,\text{球員年紀} + 0.0202\,\text{出賽場次} \\ &+ 0.0319\,\text{場均上場時間} + 0.8808\log(\text{場均罰球命中數})\end{aligned} \tag{5}$$

從以上結果，我們可以發現到 best subset regression 通常能得到較好的 R^2，而不管是對四個群體中的哪一個群體進行觀察，球員年紀和出賽場次這兩個變數都有出現在回歸模型當中。關於球員年紀，我們提供的解釋是年紀小的球員可能因爲剛進入 NBA 尚未擁有良好表現，因此球隊毋須開出高價合約就能將其簽下。隨著年紀漸長，也越能適應 NBA 高強度的競技水平，薪水也因此升高。至於出賽場次，我們提供的解釋就很直觀，出賽場次越高的球員應更被教練看重，對球隊有貢獻高，薪資水平也因此較高。

除此之外有趣的是，可以發現在後衛、前鋒、全體球員三個模型之中都出現了罰球相關的變數，我們對此一現象有兩個詮釋。其一，製造對手犯規的能力在聯盟中是極受重視地；其二，在後衛、前鋒、中鋒三個群體之中，進攻的責任普遍落在後衛及前鋒身上。

另外，前鋒的回歸模型當中出現場均助攻數，這說明了能否送出助攻對前鋒來說也是一大標的。至於中鋒的模型中出現了場均投籃命中數，因爲近十年來聯盟規則的改動，使得中鋒在禁區的進攻能力式微，中鋒早已不如以往是球隊的進攻核心，而僅是擔任球隊防守、抓籃板的藍領勞工。因此，還仍有進攻能力的中鋒在現今是較爲罕見的球員，使得球隊予以器重。