Prediction of NBA free agents' salary

Amy Murray, Bin Fang, Jose Luis Ramirez, Tan Kentaro

1. Introduction

Every year some players in NBA will sign new contracts with teams when their previous contracts expire (they may stay in their original teams or jump into new ones). Thus, it is important for free agents (players who are still seeking for contracts) to predict that how much salary they can get in their future contracts and if some teams tend to offer them similar contracts. Besides, the model is also meaningful for the teams to predict a free agent's value and check whether they have sufficient salary space to offer him an appropriate contract. This is the reason why we seek to build this model.

According to our understanding of basketball and the NBA league, we assume that free agents' salaries are correlated with their on-court performances last season, their age and salary spaces of the team before new contracts are signed. Thus we build the model to check if all or some of those factors have linear association with free agents' salary. Unfortunately, free agents' salary does not follow normal distribution, so we make a Box-Cox transformation to remedy its normality. Then we use different criterions and automatic search procedure to search for a "good" model that can be used for prediction. Afterwards we use VIF to see if there is multicollinearity in the selected model and use externally standardized residual, leverage, DEFFITS, Cook's Distance and DFBETAS to represent outliers and influential points.

2. Variables included in our study

Table 1 (on Page 5) describes all the variables in our study, including 2 response variables and 26 predictor variables.

From the official website-http://stats.nba.com/, we find salary statistics of players who have signed new contracts with their teams in 2013-2014 season, and use these data as our response variable, indicated as "Salary" .

There are totally 26 predictor variables in our full model, and most of them are players' on court performance last season, such as points, rebounds, assists, etc. "ReAge" is a measure of the distance between players' current age and their primes, which equals the absolute value of their current age minus all players' average age. Here we assume that most players are at their primes at the average age: if a player is too young or too old, he tends to be less competitive. "TeamS" indicates the total salary of a team just before it offers a contract to a player. As we know, there is a salary cap each year: except for some special cases, a team can only provide a minimum salary for a player if its total salary has surpassed the salary cap. In other words, the most total salary a team has, the less likely it might provide a big contract for a player.

Figure 1 (on Page 6) is the Descriptive Statistics for all the variables, and we can see the mean, median, standard deviation, minimum and maximum make sense for each variable.

The reason why we do not include all the players' salary in our study is that for players whose previous contracts are still valid, their salaries are seemingly not dependent of their performance last season and their current age and the team's current salary space. We do not think their data can be used for prediction of free agents' salary this season.

Players' positions might be another factor that might influence their salaries. We have background information that Center and Power forward tend to be better paid because there are now few good "big men" in the league. Besides, players from different positions have different duties on court, so the real coefficient of the predictor variables might diverse among players of different positions (for example, rebounds might have larger influence on a center's salary than on a point guard's). In ideal sense, it is necessary to include four indicator variables to differentiate players' positions. But that will largely increase the number of predictor variables when interaction terms are included, and make our model less powerful and contradicts the principle of multiple linear regression that the sample size should be much larger than the number of predictors. Consequently we do not include players' positions as predictors in our model because we only have 127 observations.

3. Box-Cox Transformation and Test for normality

In figure 1 we can see for response variable "Salary", there is large difference between mean and median, which implies that "Salary" may not follow normal distribution. Texts for normality in Figure 2 confirm that the distribution of the response variable is skewed, so transformation is necessary.

The results of Box-Cox Transformation are shown in Figure 3: when =0.25, the likelihood function has a maximum value. Thus, we will make the transformation as: TransS = $(\text{Salary}^{0.25}\text{-}1) / 0.25$.

When we use the transformed salary as the new response variable to fit the model, we can see that QQ-plot in Figure 4-1 is relatively straight and passes through (0, 0) point, which indicates that "TransS" follows normal distribution. Texts for normality in Figure 4-2 show similar results.

4. Model Selection

Figure 5 shows some of the parameter estimates for the full model. However, we can see some of the parameters make no sense, because the p-value for individual t-test is very large and VIF is very large. That shows significant multicollinearity within the full model: it is predictable, for example, RPG=ORPG+DRPG, so these three variables are highly correlated.

Thus, the full model is not a good model for our problem. We seek to select a good reduced model by different criterions (Cp Value, Adjusted R-Square, AIC, BIC and SBC) and automatic model selection procedures (Forward Selection, Backward Elimination and Stepwise Selection).

Figure 6 briefly shows the model selection results of Cp value. The blue line indicated Cp=P+1. We can see that for models including 12 parameters, the "best" model has the smallest Cp Value, but for models including 5 or 6 parameters, the "best" model has a Cp-value close to the line Cp=P+1. Therefore Cp Criterion suggests a model including 5 or 6 parameters, whose Cp Value is both relatively small and close to P+1, which renders the prediction small variance and biasness.

Figure 7 represents model selection by different criterions. For models including 16-17 parameters, Adjusted R-square attains the maximum value; for models including 10-12 parameters, AIC and BIC attains the minimum value; for models including 5 parameters, SBC attains the minimum value. Therefore, for our data set, different criterions prefer different models.

Figure 8 – 10 represents the models suggested by forward selection, backward elimination and stepwise selection respectively. The model suggested by forward selection is not a good model, because we can many coefficients' individual t-tests are not all significant (actually some of them are very large, such as the coefficients of TOPG, STPG, PFPG). VIFs for those variables are very large, which indicates that there is multicollinearity in the model, so those values of coefficients make no sense and there will be large variation if this model is used. Besides, we can see that selection criterions do not quite support this model: Adjusted R-square gets maximum in the 13^{th} step, while AIC, BIC and SBC have minimums in the 12^{th}, 7^{th}, 4^{th} step respectively.

The model suggested by backward elimination is better: as it shows in Figure 9, most coefficients' individual tests are significant, although the p-values of FG and PFPG are slightly larger than 0.05. VIFs for PTS and MPG are larger than 7, and VIFs for APG are larger than 5, and the other VIFs are relatively small. Moreover, selection criterions to larger extent support this model: Adjusted R-square gets maximum in the 14^{th} step (last second step), and AIC, BIC and SBC all have minimums in the final step.

The model suggested by stepwise selection is the most parsimonious model: only five variables (TeamS ReAge MPG PTS RPM) are selected in this model. All coefficients' individual tests are significant, except that the p-value of MPG is slightly larger than 0.05. Meanwhile, selection criterions also support this model: Adjusted R-square gets maximum in the final step, and AIC, BIC also have minimums in the final step. SBC have minimums in the 4^{th} step, but the value does not change dramatically in the last step. Figure 10 also shows residual plots of this model, from which we can see that error variance is relatively constant and regression function is approximately linear because all the shape fitted by Lowess Method is close to the line "Residual=0".

Based on all criterions mentioned above, first we will reject the model suggested by forward selection, because is has much multicollinearity. The models suggested by backward elimination and stepwise selection are both good models, but we prefer the model suggested by stepwise selection for at least two reasons: (1) As we mentioned before, Cp citerion suggests a model including 5 or 6 parameters, whose Cp Value is both relatively small and close to P+1. Actually the model suggested by stepwise selection includes 6 parameters (5 variables), and its Cp value is 4.8559 (data not shown), which is small and close to 6; (2) the model suggested by stepwise selection has less multicollinearity than the one suggested by backward elimination. In the model suggested by backward elimination, some coefficient values are inconsistent with our understanding of basketball: for example, the coefficient of APG is negative, and that means if a player has more assists in a game, his salary will decrease, which is seemingly ridiculous. The explanation is that VIFs for APG is large, so multicollinearity makes the value of this coefficient unreliable.

Finally, the model we selected is Transformed Salary = 114.17 – (3.65432E-7) TeamS - 2.71 ReAge + 1.327 MPG + 3.91 PTS + 1.59 RPM. Because TransS = ($Salary^{0.25}$-1) / 0.25, the model will be: Salary = (29.54 – (9.14E-8) TeamS – 0.68 ReAge + 0.33 MPG + 0.97 PTS

+ 0.40 RPM) [4].

From standardized regression model we can see that PTS have largest influence on free agents' salary among all variables. Besides, free agents salary are dependent of teams' total salary before the player is added, players' age, and players' on court minutes per game and rebounds per 48 minutes.

5. Outliers and Influential Points

We will study the outliers and influential points in 127 observations. Outlying Y observations are diagnosed by standardized deleted residuals; outlying X observations are diagnosed by Leverage (H_{ii}); influence on All Fitted Values are diagnosed by Cook's Distance; influence on single fitted values are diagnosed by DEFFITS; influence on six regression coefficients are diagnosed by DFBETAS. All the results are represented from Figure 11 to 14.

As it shows in Figure 11, there are 10 outlying Y observations, which are Observation 5, 13, 19, 23, 39, 45, 55, 73, 78 and 118; there are 2 outlying X observations, which are Observation 7 and 82.

Figure 10 and 11 represents 10 influential points in the data set, which are Observation 5, 7, 10, 13, 23, 45, 55, 56, 73 and 118. Seven of those influential points are also outlying Y observations, which are Observation 5, 13, 23, 45, 55, 73, and 118. Figure 12 represents observation having influence on six regression coefficients respectively.

We check our data and confirm the reliability of those outliers and influential points. Therefore we have no reason to delete those observations. Actually they will help us to find a better model to fit the data in future study.

6. Conclusion

Our multiple linear regression model to predict free agents' salaries in 2013-2014 season is: Salary = (29.54 – (9.14E-8) TeamS – 0.68 ReAge + 0.33 MPG + 0.97 PTS + 0.40 RPM) [4]. This model is helpful for players who are still wandering in the free-agent market and teams who are willing to recruit some free agents.

Table 1: Variables in the regression model (including 2 response variables and 26 predictor variables)

Response Variables	**Salary**		Salary of individual players in 2013-2014 season
	TransS		Box-Cox transformed Salary
Predictor Variables	**TeamS**		Total Salary of a team before a new player is added
	ReAge		| Players' Age - Average Age |
	players' on court statistics in 2012-2013 season	**GP**	Games Played
		MPG	Minutes per Game Played
		PTS	Points per Game
		FG	Field Goal Percentage
		threeP	Three Points
		ORPG	Offense Rebounds per Game
		DRPG	Defense Rebounds per Game
		RPG	Rebounds per Game
		RPM	Rebounds per 48 Minutes
		APG	Assists per Game
		APM	Assists per 48 Minutes
		BLKPG	Blocks per Game
		BLKPM	Blocks per 48 Minutes
		STPG	Steals per Game
		STPM	Steals per 48 Minutes
		TOPG	Turnovers per Game
		TOPM	Turnovers per 48 Minutes
		PFPG	Personal Fouls per Game
		PFPM	Personal Fouls per 48 Minutes
		ASTvsTO	Assists vs Turnovers
		BLKvsPF	Blocks vs Personal Fouls
		STvsTO	Steals vs Turnovers
		STvsPF	Steals vs Personal Fouls

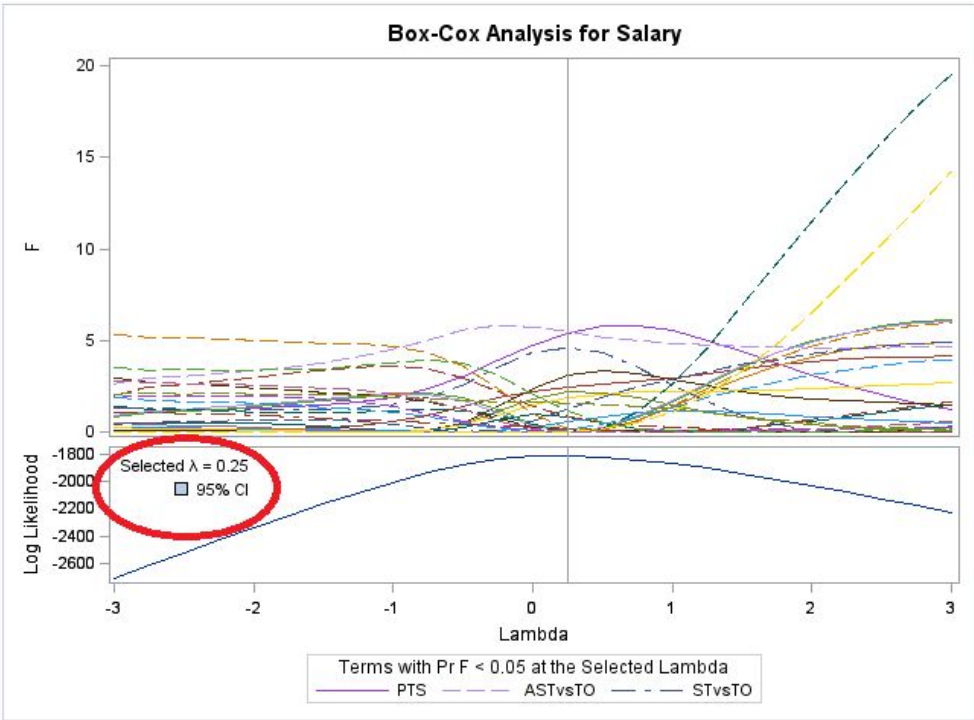
Figure 1: Descriptive Statistics for the variables

Variable	N	Mean	Median	Std Dev	Minimum	Maximum
Salary	127	3850761.87	2652000.00	3888128.43	83228.00	20513178.00
TransS	127	158.9755504	157.4185238	40.1361875	63.9402837	265.1957682
TeamSa	127	55597118.40	49928027.00	17586547.71	21513122.00	85675757.00
Age	127	28.8110236	28.0000000	3.7727769	22.0000000	39.0000000
ReAge	127	3.0731496	2.8100000	2.1713231	0.1900000	10.1900000
GP	127	62.0944882	66.0000000	18.3118632	6.0000000	82.0000000
MPG	127	21.0362205	21.3000000	7.8962888	5.1000000	37.5000000
PTS	127	8.2661417	8.0000000	4.4242876	1.2000000	19.2000000
FG	127	0.4472126	0.4440000	0.0604630	0.3080000	0.6140000
threeP	127	0.2661575	0.3300000	0.1756446	0	1.0000000
FT	127	0.7314961	0.7620000	0.1424206	0.2500000	0.9380000
ORPG	127	0.9425197	0.6000000	0.8806130	0	5.5000000
DRPG	127	2.7102362	2.2000000	1.7297047	0.1000000	9.1000000
RPG	127	3.6574803	2.8000000	2.5126827	0.3000000	14.4000000
RPM	127	8.2984252	7.3000000	4.1191189	2.3000000	19.2000000
APG	127	1.8354331	1.4000000	1.4866884	0	7.2000000
APM	127	3.9055118	3.2000000	2.4911688	0	11.5000000
BLKPG	127	0.4336220	0.3100000	0.4158118	0	2.4500000
BLKPM	127	1.0366929	0.6700000	0.9829725	0	4.7300000
STPG	127	0.6870866	0.6100000	0.3816425	0.0500000	2.0600000
STPM	127	1.5153543	1.4600000	0.5415759	0.4300000	2.7600000
TOPG	127	1.1846457	1.1200000	0.6342147	0.0800000	3.1000000
TOPM	127	2.6204724	2.5000000	0.8347929	0.7000000	4.9000000
PFPG	127	1.7960630	1.8000000	0.6110623	0.4000000	3.8000000
PFPM	127	4.3708661	4.0000000	1.5533941	1.9000000	9.8000000
ASTvsTO	127	1.5162205	1.4500000	0.7975925	0	4.3300000
BLKvsPF	127	0.2281890	0.1900000	0.1807015	0	1.0100000
STvsTO	127	0.6376378	0.6000000	0.3173988	0.1400000	2.3900000
STvsPF	127	0.3893701	0.3500000	0.1923611	0.0500000	1.3000000

Figure 2: Texts for normality (variable “Salary”)

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.926140	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.109586	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.374990	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.251909	Pr > A-Sq	<0.0050

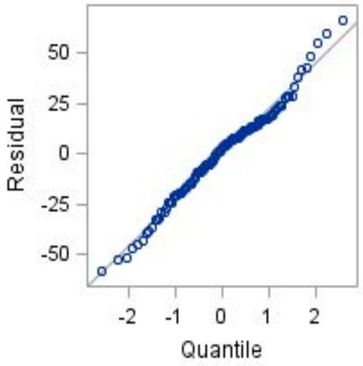
Figure 3: Box-Cox Transformation



Box-Cox Transformation Information for Salary			
Lambda	R-Square	Log Like	
-3.00	0.19	-2717.45	
-2.75	0.18	-2617.95	
-2.50	0.18	-2524.26	
-2.25	0.18	-2431.77	
-2.00	0.18	-2341.02	
-1.75	0.19	-2252.44	
-1.50	0.20	-2166.54	
-1.25	0.22	-2084.16	
-1.00	0.26	-2006.73	
-0.75	0.35	-1936.86	
-0.50	0.46	-1878.90	
-0.25	0.57	-1837.99	
0.00	0.64	-1816.56	
0.25	0.68	-1812.53	<
0.50	0.69	-1821.95	
0.75	0.69	-1841.57	
1.00	0.68	-1869.06	
1.25	0.67	-1902.69	
1.50	0.66	-1941.09	
1.75	0.64	-1983.25	
2.00	0.63	-2028.43	
2.25	0.62	-2076.07	
2.50	0.62	-2125.78	
2.75	0.61	-2177.24	
3.00	0.61	-2230.20	

< - Best Lambda
* - 95% Confidence Interval
+ - Convenient Lambda

Figure 4: Q-Q plot and Texts for normality (variable “TransS”)

<table>
<tr><th>Q-Q Plot</th><th>Texts for normality (variable “TransS”)</th></tr>
<tr><td>


</td><td>
<table>
<tr><th colspan="5">Tests for Normality</th></tr>
<tr><th>Test</th><th colspan="2">Statistic</th><th colspan="2">p Value</th></tr>
<tr><td>Shapiro-Wilk</td><td>W</td><td>0.990121</td><td>Pr < W</td><td>0.5023</td></tr>
<tr><td>Kolmogorov-Smirnov</td><td>D</td><td>0.055063</td><td>Pr > D</td><td>>0.1500</td></tr>
<tr><td>Cramer-von Mises</td><td>W-Sq</td><td>0.051947</td><td>Pr > W-Sq</td><td>>0.2500</td></tr>
<tr><td>Anderson-Darling</td><td>A-Sq</td><td>0.343202</td><td>Pr > A-Sq</td><td>>0.2500</td></tr>
</table>
</td></tr>
</table>

Figure 5: Parameter Estimates and VIF for Full Model (the table is truncated, and some data are not shown)

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	58.05931	54.57346	1.06	0.2900	0
Team Sa	1	-5.06673E-7	1.493922E-7	-3.39	0.0010	1.34254
Age	1	0.99543	0.75142	1.32	0.1883	1.56312
ReAge	1	-2.23219	1.22401	-1.82	0.0712	1.37381
GP	1	0.22198	0.19078	1.16	0.2474	2.37386
MPG	1	1.50953	0.95087	1.59	0.1156	10.96479
PTS	1	3.84272	1.54871	2.48	0.0148	9.13146
FG	1	77.48045	47.93817	1.62	0.1092	1.63400
threeP	1	-16.79806	17.75310	-0.95	0.3463	1.89116
FT	1	-4.82577	23.67535	-0.20	0.8389	2.21131
ORPG	1	-4.37256	49.91103	-0.09	0.9304	375.72872
DRPG	1	-3.91766	49.61345	-0.08	0.9372	1432.36539
RPG	1	0.36186	49.71936	0.01	0.9942	3035.55009
RPM	1	2.95401	2.30726	1.28	0.2034	17.56754
APG	1	-8.01611	8.96860	-0.89	0.3736	34.57807
APM	1	-6.48800	5.52179	-1.17	0.2428	36.80254

Figure 6: Model Selection by Cp Criterion

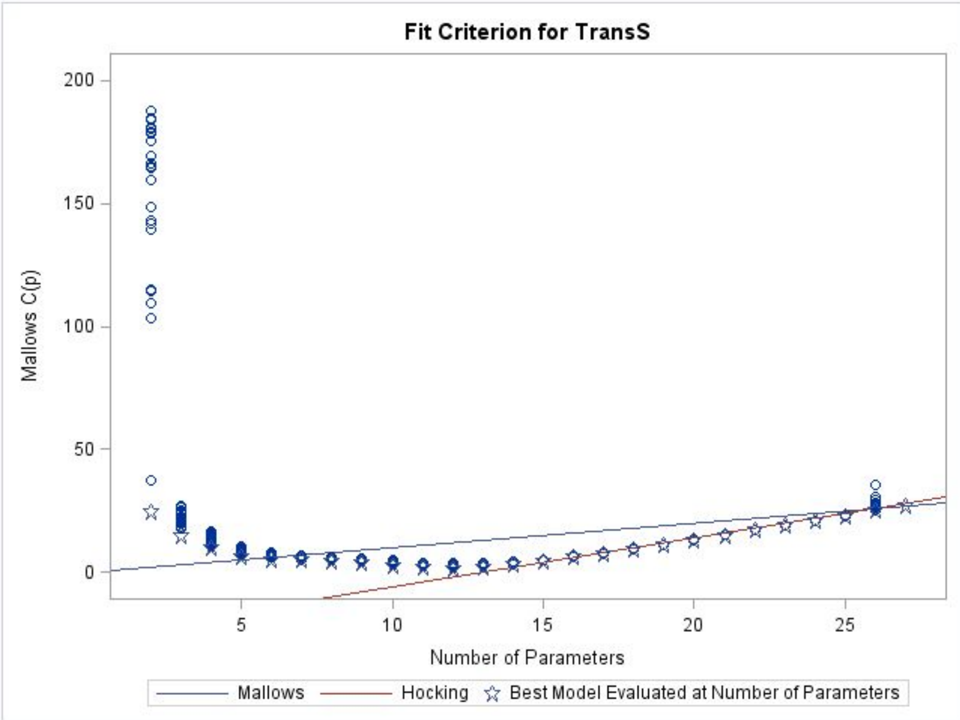


Figure 7: Model Selection by Different Criterions

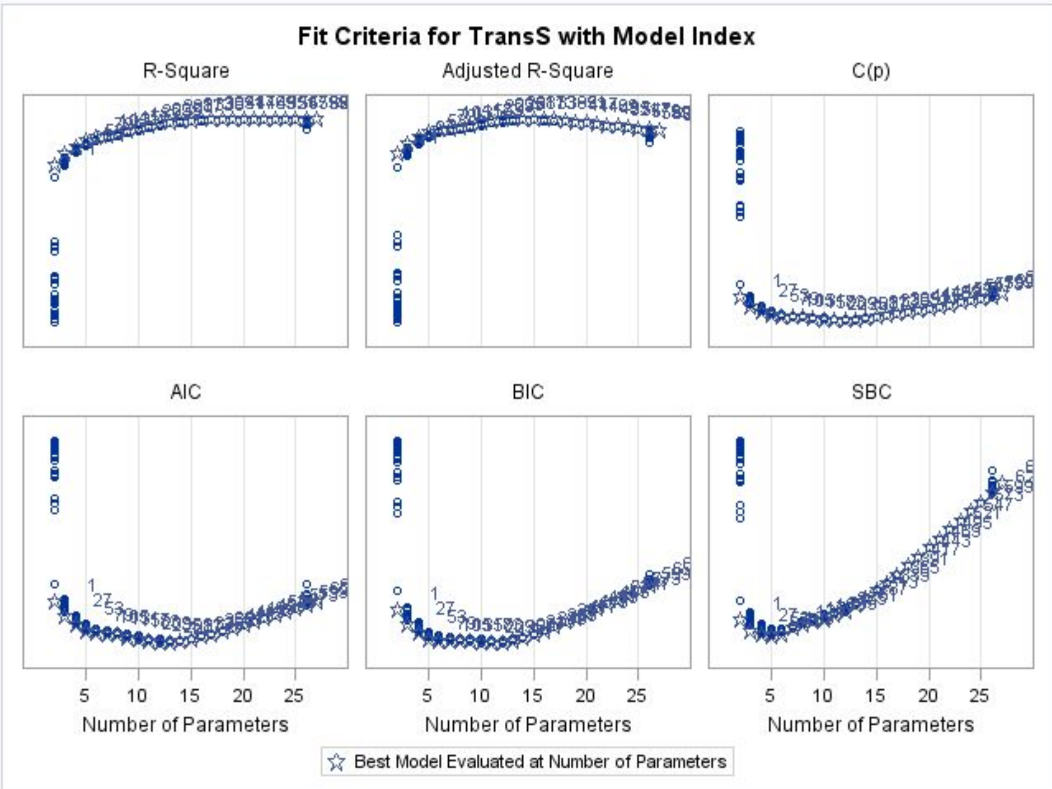


Figure 8: Model Selection by Forward Selection

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	72.42679	30.96424	2.34	0.0211	0	0
PTS	1	3.02299	1.37768	2.19	0.0303	0.33323	7.73901
TeamSa	1	-4.77745E-7	1.388374E-7	-3.44	0.0008	-0.20933	1.24187
RPM	1	1.26644	0.74249	1.71	0.0909	0.12997	1.94844
ReAge	1	-2.01497	1.11866	-1.80	0.0744	-0.10901	1.22898
MPG	1	1.93829	0.83676	2.32	0.0224	0.38133	9.09382
PFPG	1	-3.60892	10.03739	-0.36	0.7199	-0.05494	7.83635
FG	1	70.69816	42.01357	1.68	0.0953	0.10650	1.34419
APG	1	-8.16740	6.89488	-1.18	0.2387	-0.30253	21.88740
TOPG	1	5.61456	11.40144	0.49	0.6234	0.08872	10.89163
ASTvsTO	1	20.36984	7.79979	2.61	0.0103	0.40479	8.06173
STvsTO	1	-54.36050	20.66163	-2.63	0.0097	-0.42989	8.95861
STPG	1	-3.88954	27.94264	-0.14	0.8895	-0.03698	23.68913
STPM	1	21.97746	13.96851	1.57	0.1185	0.29655	11.92119
APM	1	-5.42114	4.63756	-1.17	0.2449	-0.33648	27.80260
GP	1	0.17370	0.16635	1.04	0.2987	0.07925	1.93300
STvsPF	1	25.08376	36.52026	0.69	0.4936	0.12022	10.28023

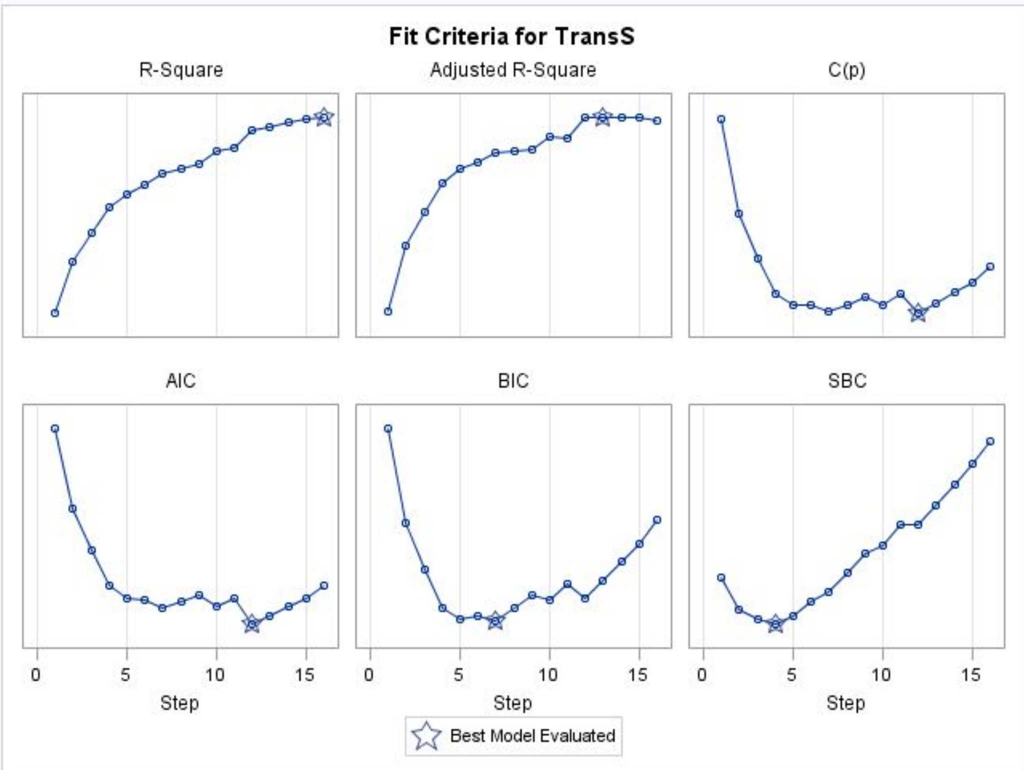


Figure 9: Model Selection by Backward Elimination

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	89.23180	26.45415	3.37	0.0010	0	0
Team Sa	1	-4.57103E-7	1.309632E-7	-3.49	0.0007	-0.20029	1.12712
ReAge	1	-2.46752	1.04555	-2.36	0.0200	-0.13349	1.09508
MPG	1	1.88734	0.74289	2.54	0.0124	0.37131	7.31138
PTS	1	3.66806	1.29932	2.82	0.0056	0.40434	7.02143
FG	1	77.02519	40.80328	1.89	0.0616	0.11603	1.29324
RPM	1	1.60404	0.71726	2.24	0.0273	0.16462	1.85471
APG	1	-11.47088	3.31899	-3.46	0.0008	-0.42489	5.17324
STPM	1	17.96266	6.97427	2.58	0.0113	0.24238	3.03128
PFPM	1	-3.53846	1.88495	-1.88	0.0630	-0.13695	1.82168
ASTvsTO	1	12.13987	5.22751	2.32	0.0220	0.24125	3.69369
STvsTO	1	-37.86919	12.52069	-3.02	0.0031	-0.29947	3.35565

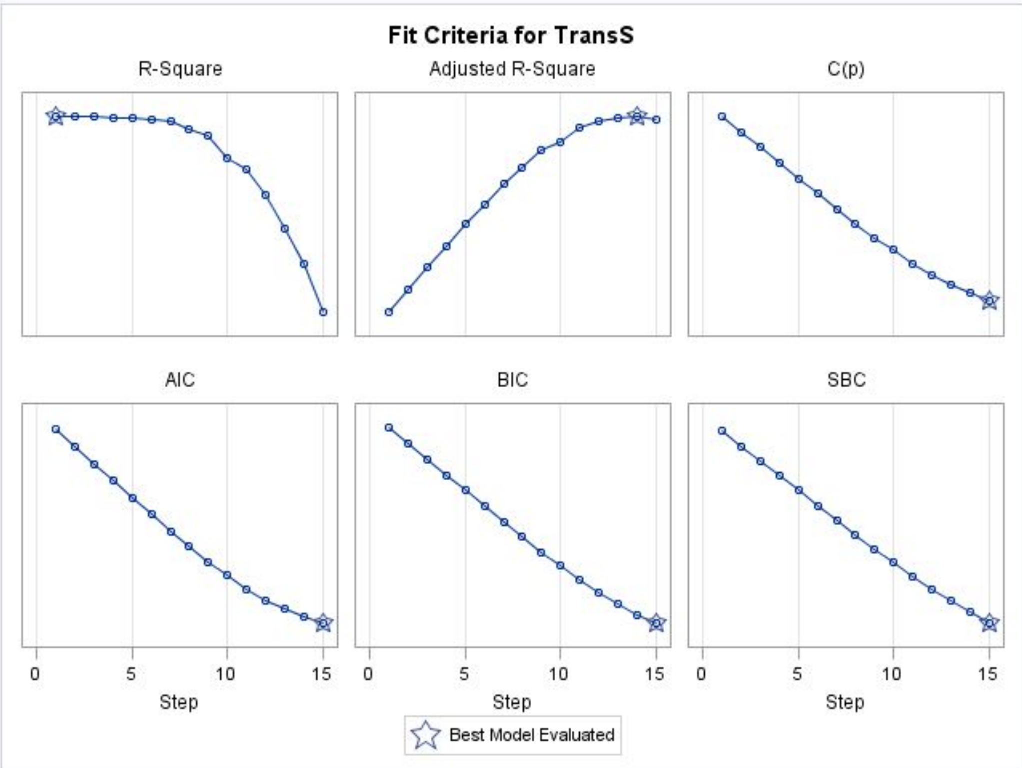
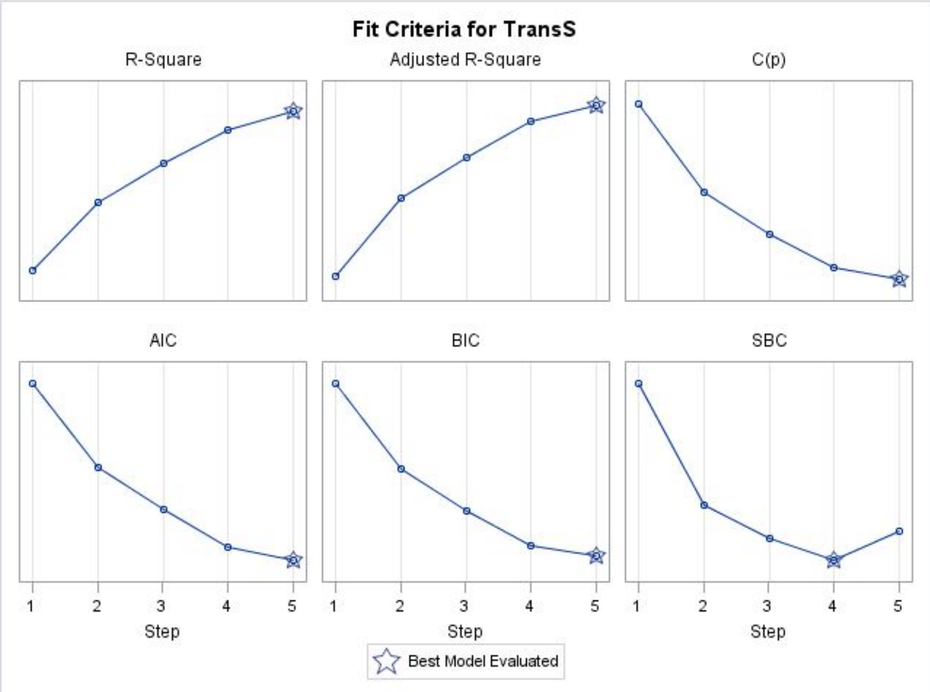


Figure 10: Model Selection by Stepwise Selection

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	114.17704	12.40383	9.20	<.0001	0	0
TeamSa	1	-3.65432E-7	1.315279E-7	-2.78	0.0063	-0.16012	1.04272
ReAge	1	-2.71554	1.07672	-2.52	0.0130	-0.14691	1.06519
MPG	1	1.32756	0.73171	1.81	0.0721	0.26118	6.50565
PTS	1	3.91133	1.31553	2.97	0.0036	0.43115	6.60170
RPM	1	1.59094	0.55781	2.85	0.0051	0.16328	1.02886



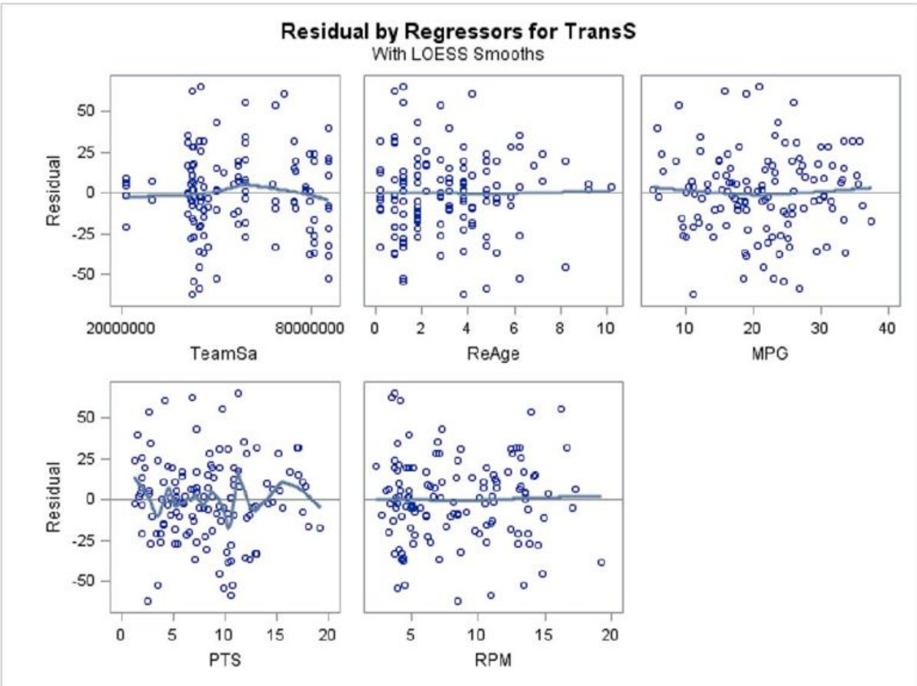


Figure 11: Outliers and Leverage Diagnostics for TranS

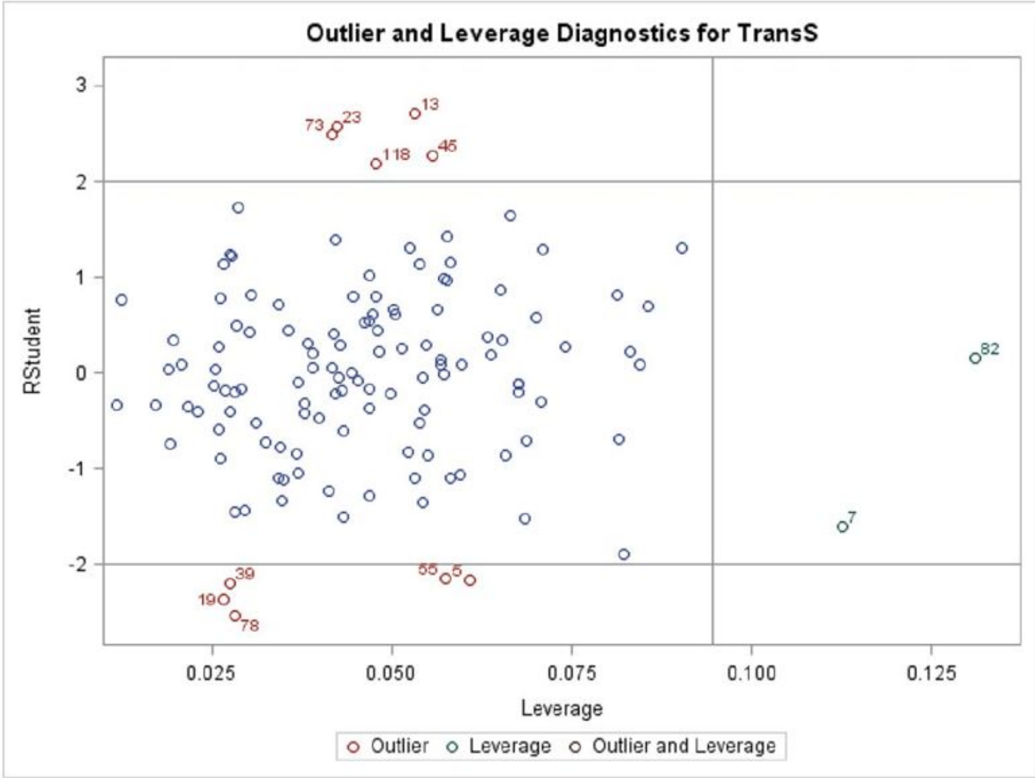


Figure 12: Cook's Distance for TranS

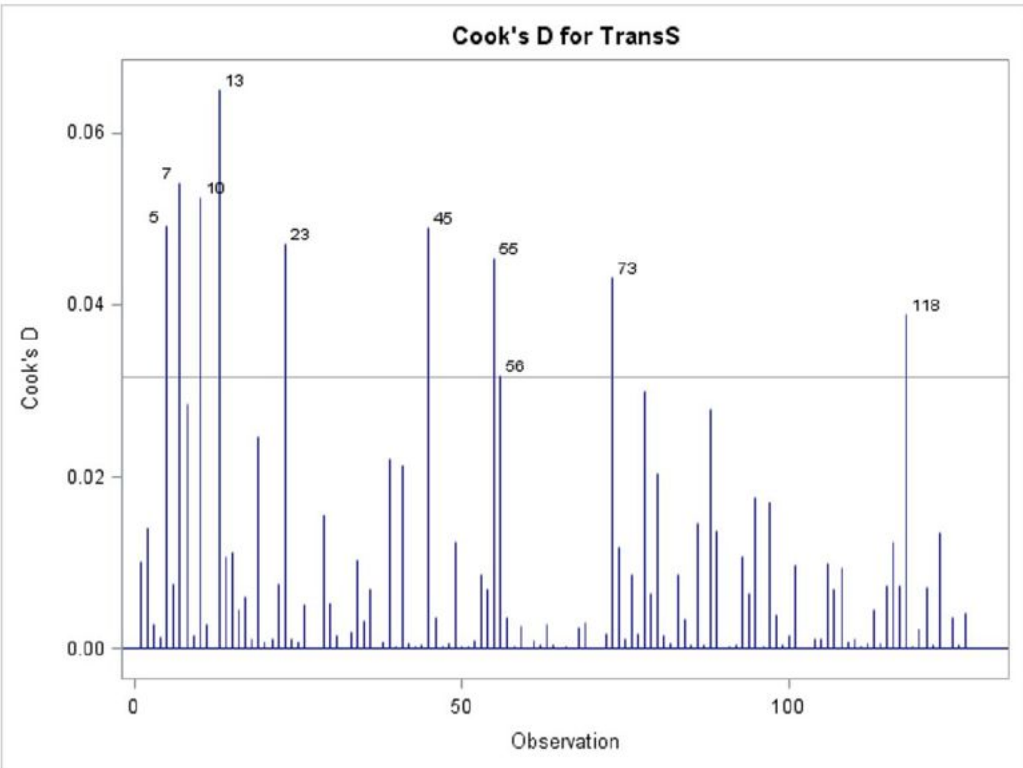


Figure 13: DEFFITS for TranS

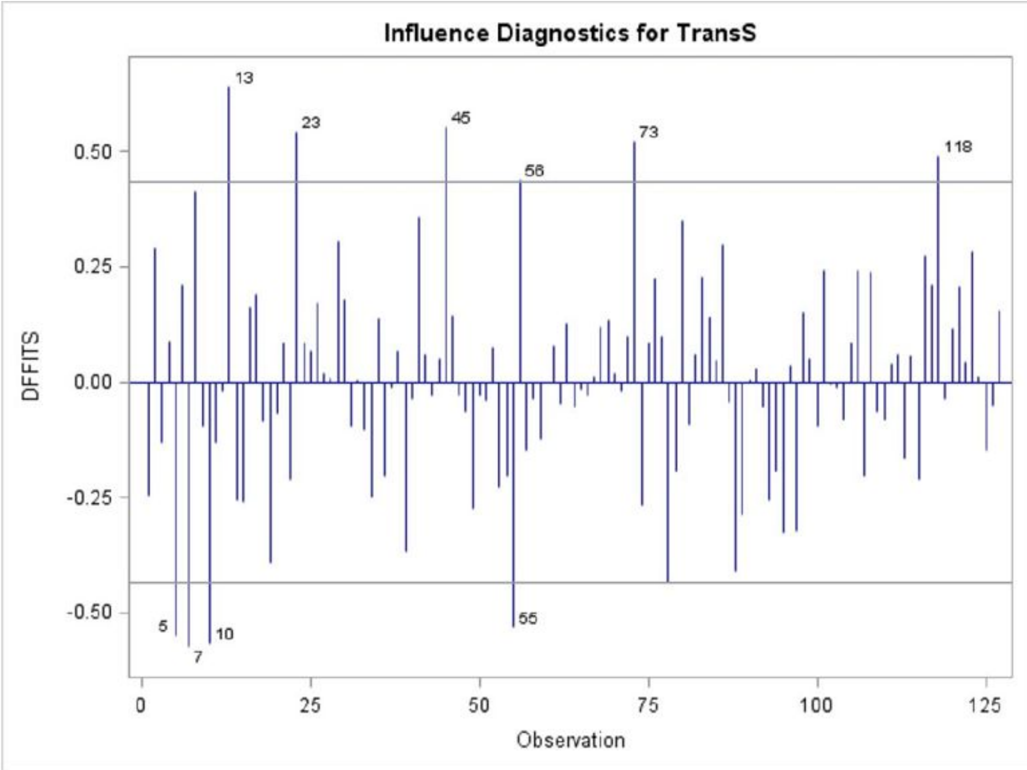


Figure 14: DFBETAS for TranS