응용통계연구 제 **21**권 3호, *2008*년, pp. *441 453*

한국 프로스포츠 선수들의 연봉에 대한 다변량적 분석*

송종우[1)]

요 약

프로스포츠 선수들의 연봉은 선수들의 개인 성적과 팀에 대한 기여도등으로 결정된다는 가정하에 프로농구와 프로야구 선수들의 전년도 성적으로 다음해 연봉을 예측 분석하였다. 분석에 있어서 data visualization 기법을 통해 변수사이의 관계, 이상점 발견, 모형진단등을 하였다. 다중선형회귀 모형(Multiple Linear Regression)과 트리모형(Regression Tree)을 이용해서 자료를 분석하고 모델간 비교를 했으며, Cross-Validation을 이용해서 최적모델을 선택하였다. 특히, 자동으로 변수선택을 하는 stepwise regression방법을 그냥 사용하기 보다는 먼저 설명변수들 사이의 관계나 설명변수와 반응변수 사이의 관계등을 조사하고 나서 이를 통해 선택된 변수들을 가지고 stepwise regression과 regression tree 방법론을 이용해서 적절한 변수 및 최종 모형을 선택하였다. 분석결과, 프로농구의 경우에는 경기당 득점, 어시스트, 자유투 성공수, 경력 등이 중요한 변수였고, 프로야구 투수의 경우에는 경력, 9이닝 당 삼진 수, 방어율, 피홈런 수 등이 중요한 변수였고, 프로야구 타자의 경우에는 경력, 안타 수, FA(자유계약)유무 여부 등이 중요한 변수였다.

주요용어: 프로스포츠, 연봉, 다중선형회귀, 트리모형, 모델최적화.

1. 서론

프로스포츠가 오래전부터 정착된 미국의 경우, 선수들의 연봉과 개인성적과의 관계에 대한 연구는 여러 차례 있어왔다. 특히 미국 프로야구 선수들의 연봉과 개인성적에 대한 관계는 1988년 ASA 미팅에서 15개 팀이 같은 데이터를 독립적으로 분석하고 나서 그 결과를 비교하는 기회도 있었다 (Hoaglin과 Velleman 1995; Watnik, 1998). 우리나라 프로스포츠도 이제 프로야구는 1982년에 창설되어 25년을 넘어가고, 프로농구는 1997년에 창설되어 10년째가 되어가고 있다. 우리는 우리나라 프로스포츠 선수들의 연봉도 과연 선수들 개인성적과 팀의 기여도등으로 예측 가능한지를 알아보고, 또 그렇다면 중요한 변수들은 어떤 것들이 있는지 알아보고자 한다. 프로축구나 프로배구가 배제된 이유는 두 종목의 경우 선수들의 연봉이 공개되어 있지 않기 때문이다. 본 연구에서는 작년 1년간의 데이터를 가지고 올해의 연봉을 예측하는데, 기본적으로 두가지 모형(다중선형회귀 모형, 트리모형)을 사용한다. 프로야구의 경우, 타자와 투수는 개인성적이나 팀의 기여도를 나타내는 변수의 종류가 매우 다르므로 각각 다른 모델을 사용하고, 프로농구의 경우는 하나의 모델을 사용하였다. 2장에서는 분석에

* 본 연구는 과학고 R&E 2007 연구비 지원에 의해 수행되었음. 참가학생들은 한성과학고, 이지은, 전상일, 김연응, 안병민 총 4명임.

1) (120-750) 교신저자. 서울시 서대문구 대현동 11-1, 이화여자대학교, 통계학과, 조교수.
E-mail: josong@ewha.ac.kr

사용된 모형을 간단히 설명하고, 각 모형에서 변수선택에 대해서 설명한다. 3장에서는 실제 자료를 사용해서 모형 적합을 하고, 최종 모델에서 선택된 모형에 대한 설명을 한다. 또 cross-validation을 사용하여 두 모형간의 성능을 비교한다. 마지막장에서는 최종결론을 내리고 미국 프로야구(MLB)의 경우와 비교해본다.

2. 분석 모형

이 절에서는 본 연구에서 사용된 두 모형, 다중회귀 모형과 트리모형에 대해서 간단히 설명하고, 각각의 모형에서 최적화 모형을 선택 하는 방법을 설명한다.

2.1. 다중선형회귀 모형(Multiple Linear Regression Model)

다중선형회귀 모형은 하나의 연속형 반응변수와 여러개의 설명변수가 있는 경우에 설명변수의 선형모형으로 반응변수를 설명하는 모형이다. p개의 설명변수가 있는 경우, 모형은 다음과 같은 식으로 표현된다.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

여기서 ϵ은 보통 평균 0, 분산 $\sigma^2 I$를 따르는 등분산 모형을 가정한다. 모수 추정은 잔차의 제곱합을 최소화하는 최소제곱추정법(Least-squares method)이 가장 많이 사용되는데, 반응변수가 정규분포를 따른다고 가정하면 최소제곱 추정량(Least-squares estimator)이 최대우도 추정량(Maximum likelihood estimator)이 된다. 일반적으로 설명변수의 수가 많을수록 반응변수의 변동에 대한 모델의 설명력(R^2)이 높아지는 반면, 설명변수들간의 다중공선성이 있을 확률이 많고 모델이 복잡해 지므로 적절한 수의 변수를 선택하는 변수선택 방법론이 많이 사용된다. 변수선택 방법론은 많이 있지만, 본 연구에서는 그중에서 가장 많이 사용되는 stepwise regression을 사용해서 변수 선택을 하기로 하였다. 특히 기존의 stepwise regression보다는 AIC(Akaike Information Criterion)값을 최소화하는 AIC stepwise regression을 사용한다 (Venables와 Ripley, 2002; Torgo, 2002). 본 연구에서 특히 중점을 둔 것은 data visualization 이다. 비통계 전문가가 회귀분석을 할 때 많이 범하는 실수 중 하나가 data visualization을 아예 안하거나 아주 제한적으로 사용한다는 것이다. 실제 자료분석에 있어서 data visualization은 아주 중요한 부분으로서, 실제 모형 가정이 맞는지를 체크할 때, 변수사이의 적절한 관계식을 찾을 때, 이상점들은 발견할 때, 변수사이의 상관관계를 점검할 때 등 많은 경우에서 아주 효과적인 툴로 사용될 수 있다 (Cleveland, 1993; Murrell, 2006). 다음 장에서 자세히 밝히겠지만, 본 연구에 사용된 자료에서도 data visualization을 통해서 자료의 명백한 실수를 발견할 수 있었고, 다중공선성이 높은 변수들도 찾아낼 수 있었다.

2.2. 트리 모형(Regression Tree Model)

트리모형은 그 모형의 단순함과 좋은 예측력으로 최근에 아주 많이 쓰이는 방법론이다. 트리 방법론은 자료를 반응변수의 이질성을 최소화하는 쪽으로 계속 이분해 나가면서(이 때 설명변수를 사용해서 이분한다) 자료를 나누고 최종 노드에서 그 노드에 속한 자료만을 가지고

모델 적합을 한다 (Breiman 등, 1984). 보통 미리 정의된 최종 노드의 크기보다 작은 수의 자료가 남으면 이분 분리를 중지한다. 트리 방법론은 반응변수가 어떤 종류인가에 따라 분류 트리(classification tree) 또는 회귀 트리(regression tree)로 나누어 지는데, 반응변수가 범주형 자료인 경우에는 분류방법론이 되고 연속형인 경우에는 회귀방법론으로 사용될 수 있다. 회귀 트리인 경우, 최종노드가 확정되면 그 노드에 속한 반응변수의 평균이 적합값(fitted value)이 된다. 트리 방법론에서 중요한 요소중 하나는 트리의 크기이다. 마치 다중회귀분석에서 설명변수의 수와 같이, 트리를 계속 확장해 나가서 트리의 크기가 커지면 잔차의 크기를 줄일 수 있지만 모형이 복잡해지고, 트리의 크기가 너무 작으면 모형은 간단하지만 잔차의 크기가 커질 수 있다. 트리의 최적화된 크기를 찾기 위해서 여러가지 기준이 있을 수 있지만, 가장 많이 사용되는 것은 비용-복잡 기준이다(cost-complexity criterion). 이 기준은 잔차 제곱합과 트리 크기(최종 노드의 수)의 가중평균의 형태를 가지고 있으며, 이 기준을 최소화 하는 트리를 최적화된 트리라고 한다 (Breiman 등, 1984). 트리 방법론은 모형 적합도 아주 빠르고, 최종 결과물도 이해하기 쉬운 장점들이 있지만, 그 계층적 구조때문에 안정적이지 못한 결과를 가져올 수도 있다. 다시 말해서, 트리의 상위층에서 약간의 변화가 생기면 그 하위구조는 아주 다른 결과가 나올 수도 있다. 이 문제점을 해결하기 위해서 bagging 방법론 등이 사용되기도 한다 (Hastie 등, 2003).

3. 분석 결과

3.1. 프로 농구

프로농구의 경우 2006-2007년의 선수별 개인 성적을 가지고 2007–2008년 선수 연봉을 예측해보았다. 현재 우리나라에는 총 10개의 프로농구 구단이 있고, 본 연구에 사용된 총 선수수는 84명이다. 외국 용병선수는 모두 제외하였고, 한번도 출장하지 않았거나, 시즌 중간에 군대에 가거나 그만둔 선수는 모두 제외하였다. 사용가능한 설명변수는 총 25개가 있다; G: 게임수, TO: 턴오버, Ast: 어시스트, APG: 경기당 어시스트, X2P&X2PA: 2점슛 성공&시도, X3P&X3PA: 3점슛 성공&시도, Stl: 스틸, RPG: 경기당 리바운드, PPG: 경기당 평균 득점, GD: 굿디펜스, FG: 야투성공률, BS: 블록슛, FT&FTA: 자유투 성공&시도, w.FT: 파울수(자유투 유), w.oFT: 파울수(자유투 무), DK&DKA: 덩크슛 성공&시도, Time: 총플레이시간(초), Def: 수비 리바운드, Off: 공격 리바운드, PTS: 득점, year: 경력(2007−데뷔년도/년). 그리고 반응 변수는 Salary(연봉)이다. 전체 선수들의 연봉분포는, 그림 3.1 (a)를 보면 알 수 있듯이, 거의 모든 집단에서의 연봉분포와 같이 오른쪽으로 긴 꼬리를 가지는(right-skewed) 분포를 가지고 있었다. 즉, 많은 선수들이 상대적으로 적은 연봉을 받고, 소수의 선수들이 높은 연봉을 받는 분포이다. 팀간 연봉 분포는 분산이 상당히 커서 한 팀이 다른 팀보다 연봉이 높거나 낮다고 말하기 쉽지 않다(그림 3.1 (b)). 우리는 모든 변수들 사이의 산점도를 그려서 변수사이의 관계들을 조사했다. 그림 3.1 (c)와 (d)를 보면 알 수 있듯이 2점슛 성공횟수와 시도횟수사이에는 아주 높은 상관관계가 있었고(0.991), 마찬가지로 3점슛 성공횟수와 시도 사이에도 아주 높은 상관관계가 있었다(0.985). 다중공선성을 피하기 위해서 상관관계가 높은 변수들은 모형에서 제외하였고, 덩크슛은 우리가 사용하려는 자료중 오직 3명의

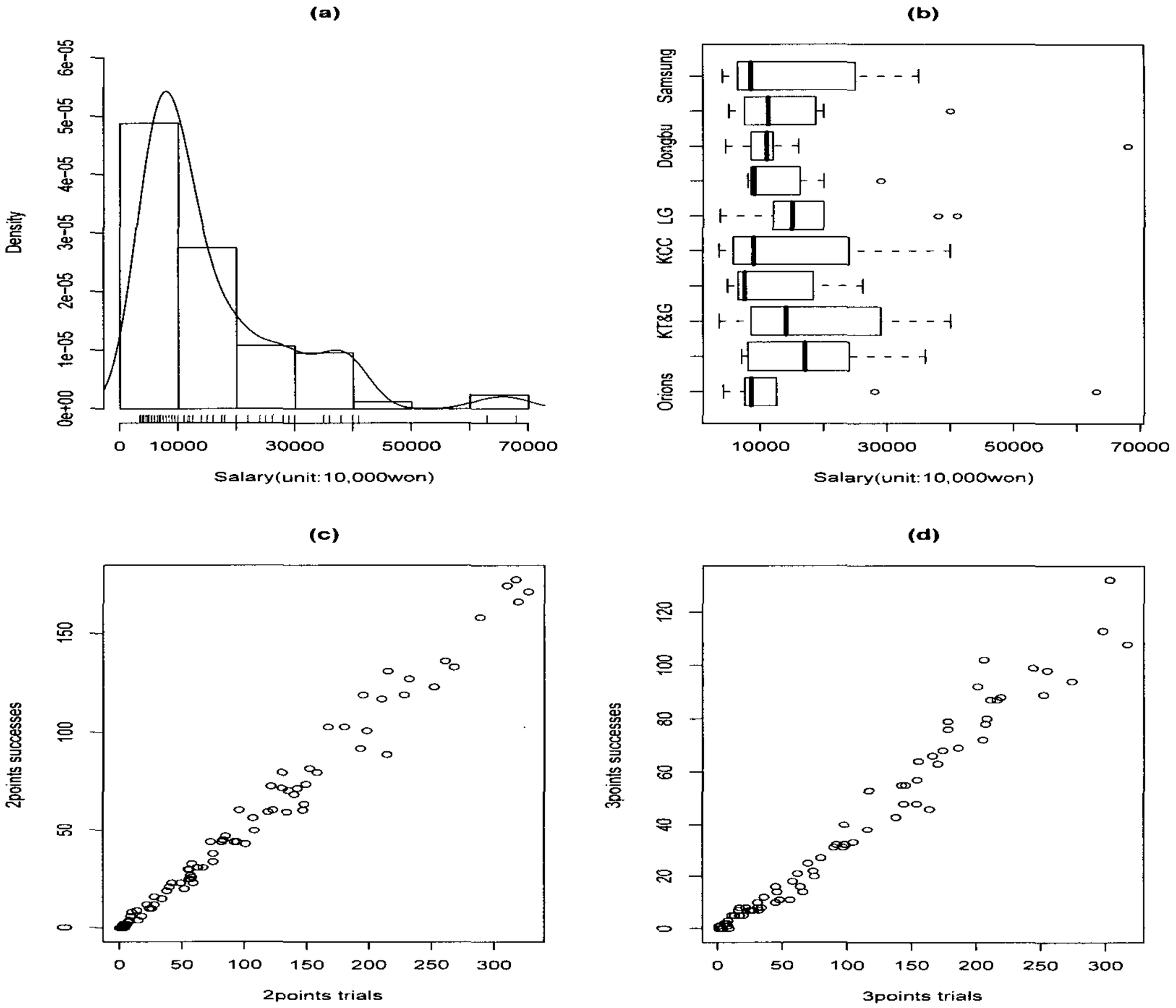

그림 3.1: 프로농구 ((a) 프로농구 선수 연봉 분포, (b) 팀별 연봉 분포, (c) 2점슛 시도횟수 대 성공횟수, (d) 3점슛 시도횟수 대 성공횟수)

선수만이 한 기록이 있고 또 그 선수들의 연봉도 분산이 커서 모형에서 제외하였다. 그래서 FTA, X2PA, X3PA, DK, DKA, Ast 등은 다중회귀 모형에서 제외되었다. 총 19개의 설명변수를 모두 모형에 포함해서 다중 선형회귀모형을 적합해 본 결과, 대부분의 변수들의 계수가 유의하지도 않았고 그 계수의 부호들도 이해하기 어려운 점이 많았다. 그러므로 적절한 수의 변수를 선택하기 위해서 stepwise regression방법론을 사용하여 모형적합을 하였다.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -404.05246 1770.67538  -0.228  0.82011
time          -0.14237    0.04296  -3.314  0.00141 **
FT            80.93310   47.61854   1.700  0.09329 .
```

```
RPG          1752.15301 1048.23077   1.672  0.09873 .
APG          1990.83630  722.51477   2.755  0.00733 **
BS            198.39483  118.92227   1.668  0.09938 .
PPG          1301.26924  439.44916   2.961  0.00409 **
year          756.32457  227.88224   3.319  0.00139 **
---
Signif. codes:  0 ‘ ***’  0.001 ‘ **’  0.01 ‘ *’  0.05 ‘ .’  0.1 ‘ ’  1

Residual standard error: 6245 on 76 degrees of freedom
Multiple R-Squared: 0.7927,     Adjusted R-squared: 0.7736
F-statistic: 41.52 on 7 and 76 DF,  p-value: < 2.2e-16
```

최종적으로 선택된 변수들은 총경기시간수. 자유투성공수, 경기당 리바운드, 경기당 어시스트, 경기당 득점. 블록슛 그리고 경력이다. 이상한 것은 총경기시간수의 계수가 음수로 나온 것이다. 우리는 여러 종류의 잔차그림을 통해 모형의 가정이 크게 어긋나지 않음을 확인할 수 있었다. 이는 뒤에 나오는 모든 다중선형모형에서도 마찬가지 였다. 경기시간수와 연봉만을 보면 양의 상관관계를 가지고 있지만 상관계수가 크지 않고, 또 최고 연봉 선수 2명의 총경기수가 상대적으로 낮아서 그점을 보정하려고 음수가 나왔다고 생각할 수 있다.

우리는 또 트리 모형을 사용해서 모형적합을 시도해 보았다. 트리모형의 경우 변수 선택이 자동적으로 되므로 아무 변수도 제거하지 않고 모형 적합을 시도하였다. 트리 방법론에서도 최적 트리 크기를 구하는 기법을 사용할 수 있지만, 자료의 수가 상대적으로 적고, 또 분류가 목적이 아닌 회귀모형에서는 트리크기가 아주 크지 않는 한 최대 크기를 사용하는 것이 예측오차를 줄이는 데 도움이 되므로 최대크기 트리를 사용하였다. 트리모형에서는 일단 2점슛 득점이 가장 중요하고, 그 다음에 경기당 득점, 파울수, 야투성공율과 경력등의 변수가 선택되었다(그림 3.2). 기본적으로 득점력이 있고 경력이 높은 선수가 높은 연봉을 받는다는 이해하기 쉬운 결과를 볼 수가 있다.

3.2. 프로 야구

프로야구의 경우에 타자와 투수는 개인기록 변수가 아주 다르므로 하나의 모형으로 설명될 수 없다. 그래서 우리는 타자와 투수를 각각 다른 모형으로 적합하였다. 타자, 투수 모두 2006년 성적과 2007년 연봉을 분석 자료로 사용하였다. 타자와 투수의 연봉 분포는 모두 오른쪽에 긴 꼬리를 가지는 분포를 가지고 있고 팀간 연봉분포는 큰 차이가 있다고 말할 순 없지만, 삼성의 연봉이 상대적으로 높았다. 총 8개의 팀이 있으며 그 팀은 다음과 같다(a: 두산, b: 롯데, c: 삼성, d: 한화, e: 현대, f: 기아, g: LG, h: SK).

3.2.1. 타자

원래 자료에서 타석수가 10이하인 경우 또는 타율이 5할 이상인 타자는 분석에서 제외하고 총 167명의 타자를 분석에 이용하였다. 자료에 몇개의 이상점들이 있었는데, 타율이 1할

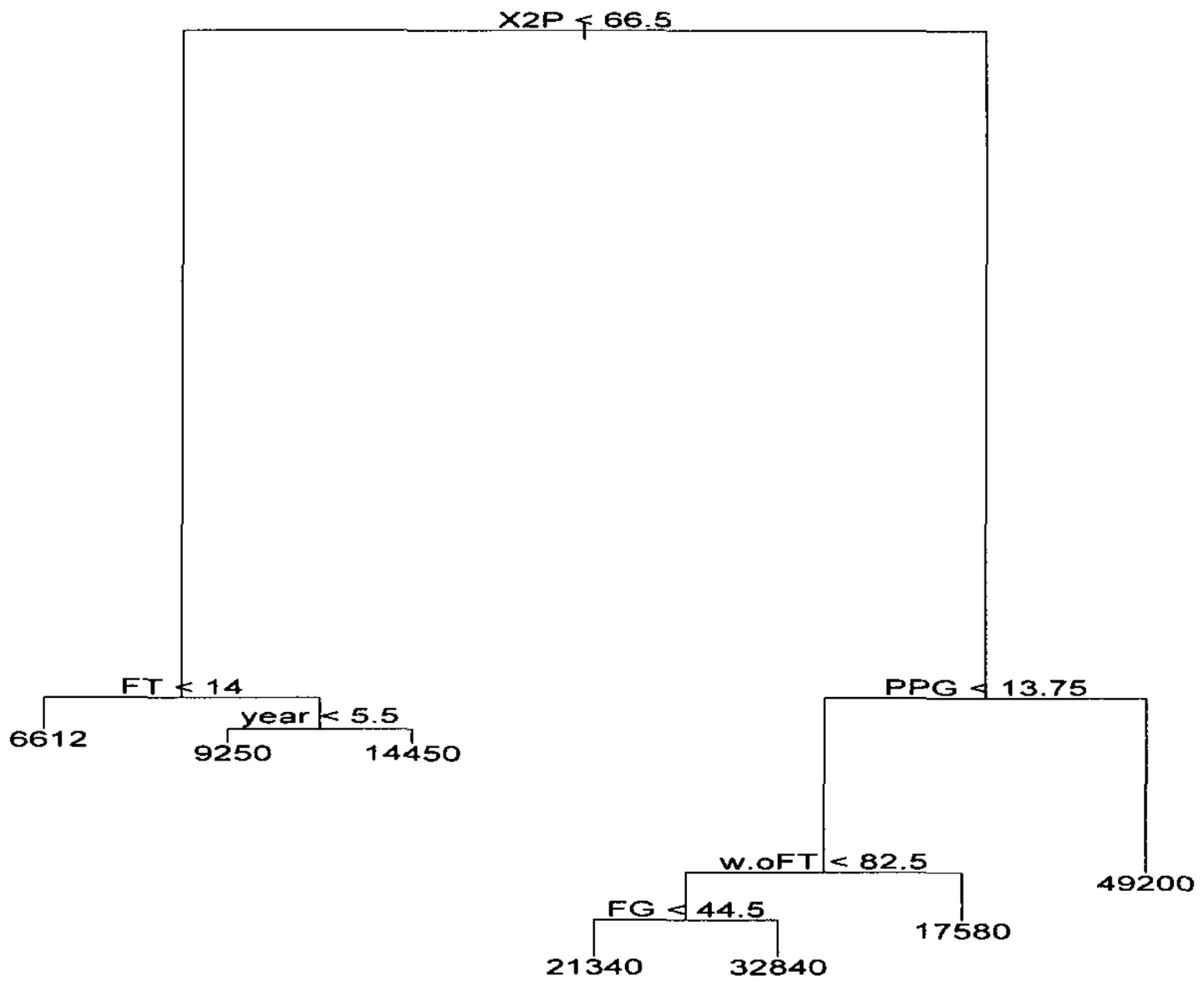

그림 3.2: 프로농구(트리 결과)

을 넘거나 OPS(출루율+장타율)값이 2를 넘는 등의 이상점들은 data visualization을 통해서 발견하고 효과적으로 수정할 수 있었다. 또, 최고연봉 선수인 심정수 선수는 부상 등으로 총 타석수가 적고 개인기록도 좋지 않아서 전체적인 패턴과 아주 동떨어진 자료이므로 이번 분석에서는 제외하였다. 사용가능한 설명변수는 다음과 같다; AVG: 타율(안타수/타수), R: 득점, BB: 볼넷, HBP: 사구, OPS: 출루율+장타율, SO: 삼진수, TPA: 타석수, GDP: 병살타, HIT: 안타수, AB: 타수(타석수−사사구−희생타), SB: 도루, SH: 희생타, X2B: 2루타수, SF: 희생플라이, X3B: 3루타수, FO/GO: 플라이아웃/땅볼아웃, HR: 홈런수, RBI: 타점수, SLG: 장타율, FA: 자유계약선수, OBP: 출루율, Career: 경력, XR: 득점공헌도, RC27: 모든 타자를 그 선수라고 가정했을 때 한경기에서 득점할 수 있는 점수의 기대치. 이 변수들 중에서 다른 변수들의 선형결합으로 표현될 수 있는 AB, XR, OPS는 제외하였다. 또 산점도를 통해서 연봉과의 관계를 조사해본 결과 연봉과 상관관계가 전혀 없는 FO/GO는 제외하였고, SH와 SF는 두 변수의 합인 HF라는 변수를 사용하였다. 대부분의 변수가 값이 클수록 개인 성적이나 팀에 대한 기여도가 높다고 할 수 있으므로(삼진수, 병살타수 제외) 다중선형회귀모형에서 계수의 부호가 양수로 나올 것을 예상할 수 있다. 하지만 위에서 제외한 변수 모두를 가지고 다중회귀모형 적합을 하면 많은 계수가 음의 부호를 갖고, 또 상당수의 계수가 유의하지 않았다. 앞서와 마찬가지로 stepwise regression방법으로 변수 선택을 해서 나온 최종 모형은 다음과 같다.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    755.76    1471.19   0.514 0.608173
TPA            119.05      29.87   3.986 0.000102 ***
HIT           -177.77     104.73  -1.697 0.091589 .
X3B          -1817.25     463.47  -3.921 0.000131 ***
HBP           -375.04     209.46  -1.791 0.075285 .
SO            -202.87      53.16  -3.816 0.000194 ***
RC.27           39.12      18.71   2.091 0.038118 *
FA1          13750.25    1633.89   8.416 2.21e-14 ***
HF            -471.98     129.10  -3.656 0.000349 ***
---
Signif. codes:  0 ‘ ***’  0.001 ‘ **’  0.01 ‘ *’  0.05 ‘ .’  0.1 ‘  ’  1

Residual standard error: 6822 on 158 degrees of freedom
Multiple R-Squared: 0.6891,     Adjusted R-squared: 0.6734
F-statistic: 43.78 on 8 and 158 DF,  p-value: < 2.2e-16
```

최종모형을 자세히 보면 일단 FA(자유계약)선수이고 득점기여도와 타석수가 높은 선수가 높은 연봉을 받고, 3루타수, 사구(몸에 맞는 볼), 삼진수, 희생타+희생플라이수, 안타수가 많을 수록 연봉이 낮음을 볼 수 있다. 특이한 것은 안타수의 계수가 음수인 점인데, 이는 선택된 변수들 사이에 양의 상관관계가 있기 때문이다. 대부분의 변수가 연봉과 양의 상관관계를 가진다는 것을 생각해보면 이를 알 수 있다. 3루타수가 많다고 연봉이 낮아지는 것도 특이한 점중 하나지만, 일단 고액연봉자를 포함한 대부분의 선수들이 3루타를 한번도 치지 못했고(66%) 일반적으로 연봉이 높은 장타자들이 발이 빠르지 않다는 사실을 생각해보면 이해할 수 있다. 같은 자료를 회귀트리모형으로 적합한 결과는 다음과 같다(그림 3.3). 트리 모형에서 선택된 변수들은 안타수(HIT), 볼넷수(BB), 타석수(TPA), 도루수(SB), 팀(Team), 자유계약선수(FA), 삼진수(SO), 병살타수(GDP) 등이다. 트리모형 결과를 보면, 안타수가 많고 자유계약선수이면서 팀이 두산, 롯데, 한화, SK인(Team: a, b, d, h) 경우에 연봉이 높고, 안타수가 적고 볼넷수가 적고, 타석수가 적을 수록 연봉이 낮음을 볼 수 있다.

3.2.2. 투수

투수 자료에서는 총 등판 수가 5경기 이하인 선수는 분석에서 제외하였고 총 119명의 투수들을 분석에 이용하였다. 사용가능한 설명변수는 다음과 같다; APP: 등판 경기수, IP: 투구이닝, ERA: 방어율, WHIP: 이닝당 출루 허용률, WIN: 승리, LOSE: 패전, HOLD: 홀드, POV: 승률, QS: 선발 6이닝 이상 3자책 이하, QS+: 선발 7이닝 이상 3자책 이하, R: 실점, ER: 자책점, OAVG: 피안타율, OOPS: 피출루율+피장타율, TPA: 상대타석, AB: 상대타수, HIT: 피안타, HR: 피홈런, K: 탈삼진, BB: 볼넷 허용, HBP: 몸에 맞는 공 허용, GDP: 병살타 유도,

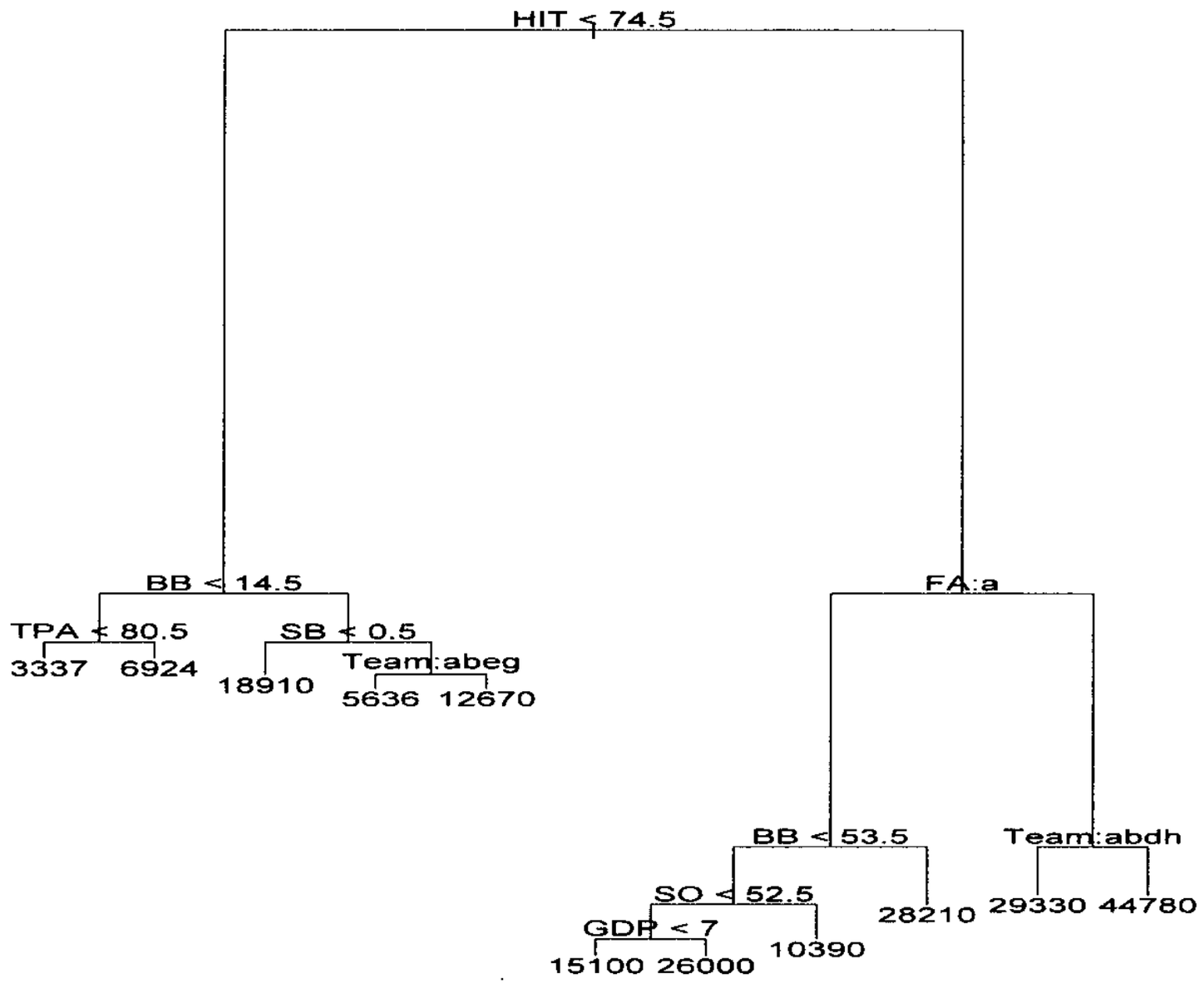

그림 3.3: 프로야구 타자(트리 결과)

IRS: 기출루자 득점 허용 비율, FOGO: 뜬공아웃 땅볼아웃 비율, CAR: 경력, FA: FA여부, FRGN: 외국인 선수. 팀별 연봉분포는 앞에서와 마찬가지로 한팀의 연봉이 다른팀보다 높거나 낮다고 말하기 힘들지만, 삼성과 한화의 연봉이 가장 높은 편이다. 투수자료에서 중요하게 생각되는 변수중의 하나는 승리경기수나 패전경기수 이지만, 이는 주로 선발투수에게만 적용시킬 수 있는 변수이다. 그러므로 선발투수와 마무리투수로 자료를 나눌 필요가 있다. 우리는 홀드와 세이브수의 합이 10을 넘는 선수를 마무리투수라고 하고, 그 합이 10이하인 선수는 선발투수로 나누는 변수를 정의해서 모형적합에 이용하였다. 또 많은 변수들이 변수값 자체보다는 이닝당 변수값이 의미 있는 경우가 대부분이므로 변수중에서 실점, 자책점, 피안타수, 피홈런수, 탈삼진수, 볼넷수, 몸에맞는 공수, 병살타 유도수, 기출루자 득점허용수 등은 실제 투구이닝수로 나눈 이닝당 수를 변수로 사용하였다. Stepwise regression을 사용한 다중선형회귀모형의 결과는 다음과 같다.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  22953.07    4915.62   4.669 8.74e-06 ***
APP           -173.69      83.06  -2.091 0.038865 *
IP            -512.93     339.37  -1.511 0.133606
HOLD           442.79     247.46   1.789 0.076364 .
```

```
SAVE          481.80     133.65   3.605 0.000474 ***
TPA           142.64      83.68   1.705 0.091156 .
K.9IP       24392.25   11727.44   2.080 0.039899 *
BB.9IP     -48282.91   26685.50  -1.809 0.073181 .
CAR          750.58     185.06   4.056 9.46e-05 ***
FAX         -7953.58    2777.73  -2.863 0.005037 **
FRGNX      -13614.19    3564.42  -3.819 0.000224 ***
---
Signif. codes:  0 ‘ ***’  0.001 ‘ **’  0.01 ‘ *’  0.05 ‘ .’  0.1 ‘ ’  1

Residual standard error: 7284 on 108 degrees of freedom
Multiple R-Squared: 0.5792,     Adjusted R-squared: 0.5402
F-statistic: 14.86 on 10 and 108 DF,  p-value: 2.535e-16
```

최종적으로 선택된 변수는 등판경기수(APP), 투구이닝수(IP), 홀드수(HOLD), 세이브수(SAVE), 상대타석수(TPA), 이닝당 삼진수, 이닝당 볼넷수, 경력, 자유계약여부, 외국인여부 등이다. 위의 결과에서 볼 수 있듯이 경력이 길고, 자유계약선수, 외국인, 이닝당 탈삼진수, 홀드, 세이브수가 높을 수록 연봉이 높다. 또 볼넷수가 많거나 등판경기수나 투구이닝수가 많을 수록 연봉이 낮게 나오는데 등판경기수나 투구이닝수의 계수가 음수인 것은 이해하기 어려우나, 실제 자료를 보면 상당수의 고액연봉 투수들의 실제 등판경기와 투구이닝수가 높지 않음을 볼 수 있다. 같은 자료를 회귀트리모형으로 적합한 결과는 다음과 같다(그림 3.4). 선택된 변수들은 이닝당 삼진수, 패전수, 경력, 이닝당 자책점, 피출루율+피장타율, 외국인여부, 이닝당 출루허용율 그리고 이닝당 피홈런수이다. 적합된 모형을 자세히 보면 이닝당 삼진수가 많고, 경력이 길고, 피홈런수가 적을 수록 연봉이 높고, 외국인 선수가 국내선수보다, 이닝당 출루허용율이 낮을 수록 연봉이 높다. 연봉이 낮은 선수들은 기본적으로 경력이 짧고 이닝당 자책점이 높고, 피출루율+피장타율이 높은 선수들이다.

3.3. 모형비교

우리는 프로농구선수와 프로야구 타자, 투수에 대해서 각각 다중선형회귀모형과 회귀트리모형으로 적합하고, 적절한 변수선택후에 채택된 결과를 설명하였다. 각각의 자료에 대해서 2개의 모형을 적합하였는데, 이 장에서는 다중선형회귀모형과 회귀트리모형을 비교하고자 한다. 복잡도가 다른 모형간의 비교는 일반적으로 잔차의 크기로 표현되는 적합도(goodness of fit)와 추정해야할 모수의 수 또는 자유도로 표현되는 모형의 복잡도(model complexity)의 가중평균이 많이 쓰인다. AIC나 BIC가 대표적인 예이다. 우리는 다중선형회귀모형에서는 stepwise regression방법론으로 적절한 변수선택후에 최종모형을 결정하였고, 회귀트리모형에서는 가능한 최대크기를 사용하였다. 트리모형에서도 크기 최적화가 가능하지만 자료수가 많지 않으므로 default size를 사용하였다. 우리는 두 모형의 성능을 비교할 때 Cross-Validation(CV)값을 이용하고자 한다. Cross-validation방법에서는 원자료를 training set과 test set으로 임의로 일단 나누고 모형적합은 training set만을, 모형평가는 test set만을 사

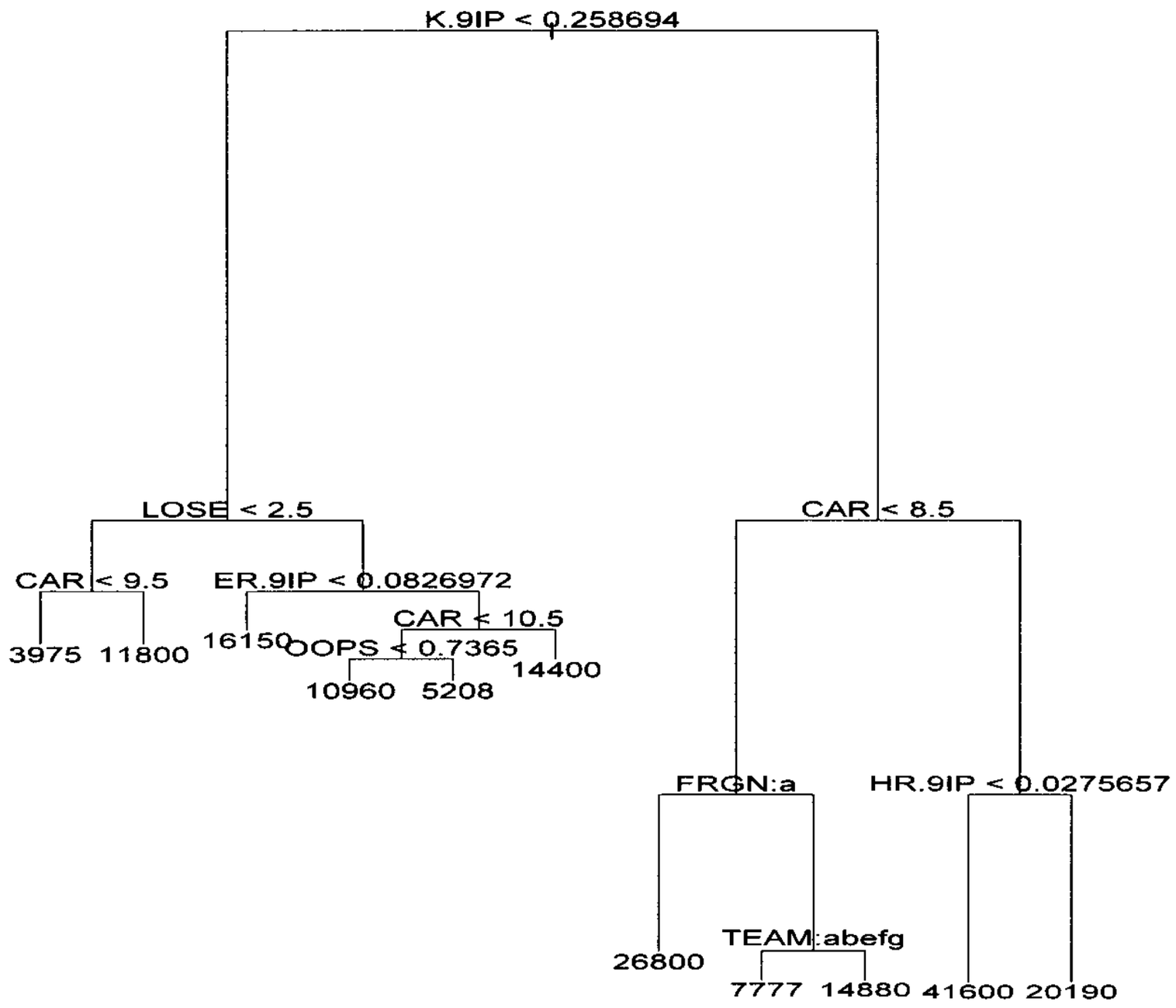

그림 3.4: 프로야구 투수(트리 결과)

표 3.1: 모형별 100번의 cross-validation 결과의 평균(괄호안은 표준편차)

	프로농구	프로야구 타자	프로야구 투수
다중선형모형	0.387(0.158)	0.401(0.114)	0.699(0.301)
트리모형	0.604(0.227)	0.893(0.357)	1.188(0.515)

용한다. 이렇게 하는 이유는 모형이 자료를 너무 쫓아가는 overfitting 모형이 실제 비슷한 자료가 왔을 때 적합도가 별로 좋지 않기 때문이다. 모형평가에는 표준화 잔차제곱평균(Normalized Mean Squared Error: NMSE)을 사용하였다. 우리는 70%의 자료를 training set으로 30%의 자료를 test set으로 사용하여 NMSE를 계산하였는데, 자료의 선택이 임의적이므로 총 100번을 선택해서 NMSE의 평균과 분산을 계산하였다.

위의 표 3.1에서 볼 수 있듯이 모든 경우에 stepwise regression을 사용한 다중선형회귀모형의 결과가 더 좋게 나왔다. 그 이유중의 하나는 자료의 수가 많지 않아서 트리모형의 크기(최종 노드수)가 크지 않기 때문이다. 예를 들어 트리모형의 최종 노드수가 7개이면 모든 예측치는 7개중의 하나인 값을 취해야만 한다. 그러므로 다양한 예측치가 나올 수 있는 다중선형모형에 비해서 최종노드수가 작으면 트리모형의 잔차제곱이 클 가능성이 높다. 특히 프로야

구 투수의 경우에 트리모형의 결과는 아주 좋지 않았다. 결과를 자세히 분석해본 결과, 트리모형을 사용한 프로야구의 경우, 경력과 9이닝당 삼진수, 승패수, FA여부 등이 중요한 변수인데, 고연봉을 받는 몇몇 선수들 가운데 트리모형과 잘 맞지 않는 선수들이 있었다. 예를들어 배영수 선수의 경우 연봉이 3억으로 고연봉 선수이지만, FA도 아니고 경력은 7년, 승수는 8승 등으로 트리모형에서 고연봉 집단으로 분류되는 그룹에 들지 못하였다. 다시 말해서, 트리모형의 경우에 잔차의 제곱합이 아주 큰 관측치들을 볼 수가 있는 반면, 다중선형회귀의 경우에는 그런 경우가 거의 없었다.

4. 결론 및 토의

우리는 본 연구를 통해서 프로농구와 프로야구 선수들의 연봉에 영향을 미치는 선수 개인 성적들이 어떤 것들이 있는지를 알아보았다. 다중회귀 모형과 트리모형에 따라 선택된 변수들이 약간 차이는 있었지만, 프로농구의 경우에는 경기당 득점, 어시스트, 자유투 성공수, 경력 등이 중요한 변수였고, 프로야구 투수의 경우에는 경력, 9이닝 당 삼진수, 방어율, 피홈런수 등이 중요한 변수였고, 프로야구 타자의 경우에는 경력, 안타수, FA유무 여부 등이 중요한 변수였다. 두 모형간의 비교에서는 모두 다중회귀모형이 트리모형보다 예측력이 뛰어난 결과를 보였지만, 트리모형이 결과를 설명하기에는 더 용이하였다. 트리모형의 예측력이 좋지 않은 이유는, 총 자료수가 많지 않아서 트리의 크기가 예측력이 좋아질 때까지 충분히 클 수 없었던 것으로 보인다. 트리방법론의 성과는 회귀모형보다는 분류(Classification)에서 더 잘 볼 수 있는 경향이 있는데, 분류의 경우에는 일반적으로 class의 수가 아주 많은 경우가 별로 없으므로 적당한 크기의 트리모형도 좋은 결과를 가져올 수 있다고 생각한다. 실제 연봉 데이터를 보면 오른쪽으로 긴 꼬리를 가지는 분포를 가지고 있다. 일반적으로 이런 분포의 데이터는 변수변환, 예를 들어 log 변환 후에 모형적합을 하는 것이 좋은 결과를 나타내는 경우가 많다. 우리도 log 변환 후에 다중 선형 회귀 모형적합을 시도해 보았는데 stepwise regression결과 같은 변수들이 선택되었고, 연봉데이타를 그대로 사용한 모형의 diagnostic plots들이 좋은 편이었으므로, 변수변환하지 않은 데이터를 그냥 사용하였다. 프로야구의 경우. 미국 MLB의 데이타도 분석 해보았다. American League에서 5개의 팀, national league에서 5개의 팀을 임의로 선택한 후에 트리모형으로 적합을 하였다. 마찬가지로 2007년 선수성적과 2008년 연봉데이터를 이용하였는데, 그 결과 우리나라 프로야구와는 좀 다른 결과가 나왔다. 경력이 중요한 요소인 점은 비슷했으나, 타자의 경우에 타점과 장타력이 아주 중요한 변수로 선택되었다. 이는 팀의 기여도에 아주 중요한 요소인 타점과 장타자를 우대하는 미국 프로야구의 성격을 잘 나타내고 있다고 생각한다.

참고문헌

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*, Chapman & Hall/CRC, New York.

Cleveland, W. S. (1993). *Visualizing Data*, Hobart Press, New Jersey.

Hastie, T., Tibshirani, R. and Friedman, J. H. (2003). *The Elements of Statistical Learning*, Springer, New York.

Hoaglin, D. C. and Velleman, P. F. (1995). A critical look at some analyses of major league baseball salaries, *The American Statistician*, **49**, 277–285.

Murrell, P. (2006). *R Graphics*, Chapman & Hall/CRC, New York.

Torgo, L. (2002). *Data Mining with R*, http://www.liaad.up.pt/ ltorgo/DataMiningWithR/

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*, Springer, New York.

Watnik, M. R. (1998). Pay for play: Are baseball salaries based on performance?, *Journal of Statistics Education*. **6**.

[2008년 3월 접수, 2008년 4월 채택]

A Multivariate Analysis of Korean Professional Players Salary*

Jongwoo Song[1)]

ABSTRACT

We analyzed Korean professional basketball and baseball players salary under the assumption that it depends on the personal records and contribution to the team in the previous year. We extensively used data visualization tools to check the relationship among the variables, to find outliers and to do model diagnostics. We used multiple linear regression and regression tree to fit the model and used cross-validation to find an optimal model. We check the relationship between variables carefully and chose a set of variables for the stepwise regression instead of using all variables. We found that points per game, number of assists, number of free throw successes, career are important variables for the basketball players. For the baseball pitchers, career, number of strike-outs per 9 innings, ERA, number of homeruns are important variables. For the baseball hitters, career, number of hits, FA are important variables.

Keywords: Professional sports, salary, multiple linear regression, regression tree, model optimization.

* This research was supported by science-highschool R&E 2007

1) Assistant Professor, Dept. of Statistics, Ewha Womans University, 11-1 Dahyun-Dong, Seodaemoon-Gu, Seoul 120-750, Korea.
E-mail: josong@ewha.ac.kr