Multiple Data Analysis Assignment #1 MLR

Predicting NBA player Salaries

2023.04.06

2018170809 오민제

Q1.

데이터셋 Overview

스포츠 팬들 사이에는 FA로이드라는 말이 있다. FA로이드란 스포츠 선수들이 FA(자유계약) 직전 해에 갑자기 급격하게 실력이 좋아져, 이듬해 더 좋은 계약을 받아내는 것을 통칭하는 FA와 스테로이드의 합성어이다. 어떻게 보면 이런 단어의 존재가 스포츠에서 선수의 가장 최근년도 퍼포먼스와 이후 계약 연봉 사이의 상관관계를 역설한다. 그렇다면 정말로 어떤 관계가 있을까? 본 보고서에서는 전미 농구 협회(NBA)의 최근 10년간 선수들의 계약 연봉과 계약 직전해 스탯의 관계를 다변량 회귀분석을 통해서 알아본다.

본 보고서는 KAGGLE에서 제공하는 현역 NBA 선수들의 2010/2011 시즌부터 2019/2020 시즌 사이에 체결된 계약에 대한 정보와 전년도 정규시즌 스탯에 관한 데이터셋을 활용하였으며, 다년계약인 선수의 경우 계약연봉은 계약기간 내 평균연봉으로 계산하였다. 위 데이터셋은 10년간 138명의 플레이어에 대한 199개의 계약과, 24개의 스탯에 관한 변수를 포함한다.

<데이터셋: https://www.kaggle.com/datasets/jarosawjaworski/current-nba-players-contracts-history?select=nba_contracts_history.csv**>**

AGE	나이	**FTM**	성공한 자유투 수
GP	플레이한 게임 수 (Max: 82)	**FTA**	시도한 자유투 수
W	승리한 게임 수	**FT%**	자유투 야투율 (FTM/FTA)
L	패배한 게임 수	**OREB**	공격 리바운드 수
MIN	플레이한 총 시간 (min)	**DREB**	수비 리바운드 수
PTS	득점한 총 점수	**REB**	리바운드 수 (OREB + DREB)
FGM	성공시킨 필드골 수	**AST**	어시스트 수
FGA	시도한 필드골 수	**TOV**	턴오버 수
FG%	야투율 (FGM/FGA)	**STL**	스틸 수
3PM	성공시킨 3 점슛 수	**BLK**	블락 수
3PA	시도한 3 점슛 수	**PF**	퍼스널 파울 수
3P%	3 점 야투율 (3PM/3PA)	**+/-**	출전 시간 내 득점 마진 (MARGIN)

Figure 1 데이터셋의 설명 변수들

이후 원활한 모델링을 위해 몇 개의 변수들을 수정하였다. 우선 승패 수는 서로 밀접한 연관이 있는 변수이기 때문에 승률 WR(win rate, 승률)이라는 한가지 변수로 통합하였다. 두번째

로, 득점 수, 어시스트 수 등의 변수들을 시즌 단위에서 경기 단위로 수정했다. 농구에서 82경기를 모두 뛰면서 평균 20득점을 한 선수(총 득점 1640)보다 50경기만 출전해도 평균 32득점(총 득점 1600)을 한 선수가 더 잘하는 선수이기에 일반적으로 경기당 평균 스탯으로 선수를 평가한다. 경기수(GP)로 나눈 변수 목록은 다음과 같다: MIN, PTS, FGM, FGA, 3PM, 3PA, FTM, FTA, OREB, DREB, REB, AST, TOV, STL, BLK, PF, +/-.

종속변수인 연봉은 돈의 특성상 그 가치가 시간과 밀접한 연관이 있기 때문에 일반적으로 인플레이션의 영향을 보정한다. 하지만 NBA선수의 시간에 따른 연봉 변화와 가장 밀접한 연관이 있는 것은 샐러리 캡이다. 샐러리 캡이란 한 해 구단이 사용할 수 있는 자본의 상한선으로, NBA 시장의 확대와 자본유입, 인플레이션 등 여러 요소들을 복합적으로 고려해 사무국이 매년 책정하는 값이다. 예를 들어 15-16시즌 연봉과 16-17시즌 연봉의 차이는 단순 인플레이션만으로는 설명하기 보다는 **Figure 2**를 샐러리 캡의 변화로 인한 것이라고 해석하는 것이 자연스럽다. 따라서 본 보고서는 선수의 연봉과 그 해 샐러리 캡의 비율을 종속변수로 사용하였다.

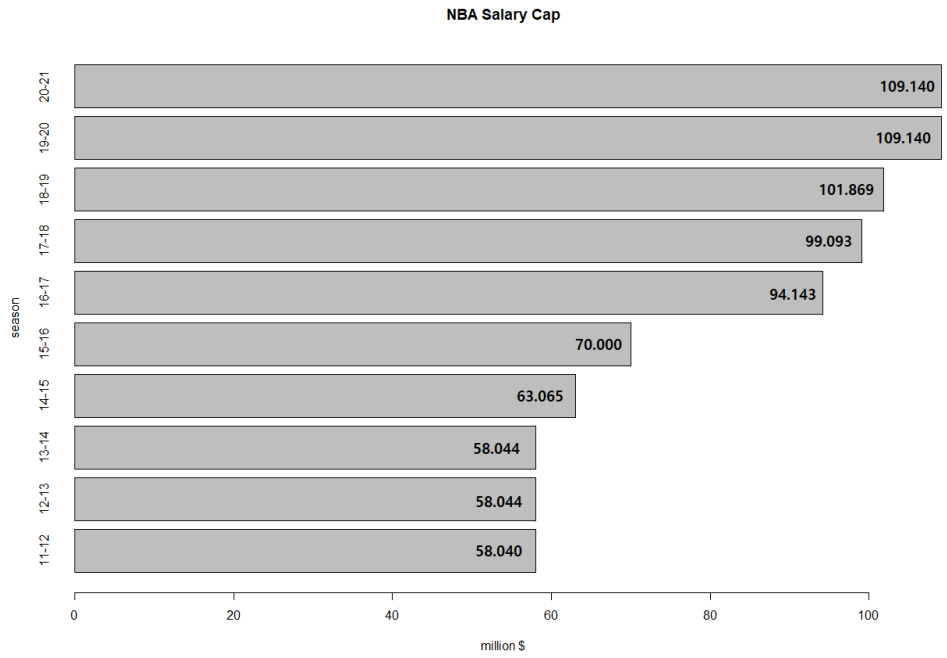


Figure 2 연도별 NBA 샐러리 캡

Q2.

중요변수 예측과 선형관계에 대한 가정

농구는 5대5라는 적은 인원수로 진행되기 때문에 특정 스타 플레이어가 큰 영향력을 발휘하기 쉬운 스포츠이다. 일반적으로 스타 플레이어는 단신으로 많은 점수, 리바운드와 어시스트를 올리고, 많은 시간을 플레이한다. 농구에서 이런 양과 관련된 변수를 볼륨이라고 하며, 스타 플레이어일수록 높은 볼륨을 기록한다. 반대로 농구에서 야투율처럼 플레이의 질과 관련된 변수는 효율이라고 하며, 일반적으로 효율과 볼륨은 trade-off 관계에 있다고 한다. 팀 내에서 스타 플레이어는 공격을 많이 시도하고, 이에 비례하는 상대팀의 집중견제를 받으며, 체력적으로 부담

또한 작용해 효율의 감소로 이어진다. 그에 비해서 롤 플레이어(스타 플레이어의 반대말)의 경우 상대가 약한 강도로 수비를 하기 때문에 볼륨은 낮지만 효율은 스타 플레이어와 큰 차이가 나지 않거나 앞서기도 한다. 승률과 득실마진의 경우 개인성적보다는 팀 성적의 영향이 크고, TOV, STL, BLK, PF같은 변수는 실력보다는 포지션에 따른 차이가 훨씬 크다. 마지막으로 보통 스포츠에서 연봉은 증가함수이지만, 농구에서는 기량 꺾인 이후에도 연봉을 감량하며 뛰는 선수가 많이 있기 때문에 (이런 선수들을 위한 연봉 제도도 존재한다) 상관이 없을 것이다. 이를 바탕으로 설명변수들 중 높은 상관관계가 있는 변수와 필요 없을 것으로 예상하는 변수를 나눌 수 있다.

상관관계가 있음	**필요 없음**		
볼륨관련 변수	*기타*	*효율관련 변수*	*성적, 포지션관련 변수*
MIN, PTS, FGM, FGA, 3PM, 3PA, FTM, FTA, OREB, DREB, REB, AST,	AGE	FG%, 3P%, FT%	GP, WR, STL, TOV, BLK, PF, +/-

Figure 3 설명변수들에 대한 예측

NBA에서 스타 선수일수록 높은 연봉을 받고, 높은 볼륨을 기록하기 때문에 종속변수는 설명변수에 따른 증가함수이다. 이 둘 선형성을 가정하기 앞서 샐러리 캡 제도로 돌아가야 한다. 샐러리 캡 제도는 구단뿐만 아니라 개별 선수의 최대 연봉에도 구단 샐러리 캡의 30-35%라는 상한선을 두며 이 최대연봉을 받는 선수를 맥스라고 한다. 따라서 선수가 아무리 특출나도 연봉은 샐러리캡에 수렴하며, 만약 실력이 설명변수라면 연봉에 대한 선형성을 가정할 수 없을 것이다.

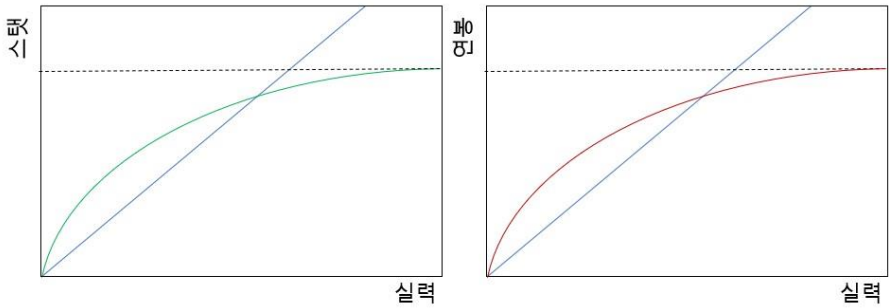


Figure 4 실력-연봉(빨간색), 실력-스탯(초록색) 각각의 예상 관계

하지만, 본 보고서에서 사용하는 설명변수는 실력이 아닌 스탯이며, 농구에서 스탯은 연봉과 마찬가지로 실력을 전부 반영하기보다는 수렴하는 성질이 있다. 그 이유는 6개월간 82경기라는 NBA 정규시즌의 살인적인 스케줄에서 찾을 수 있는데, 아무리 특정 선수가 뛰어나도 체력적인 한계로 매 공격을 직접 진행할 수 없다. 예시로 들어 르브론 제임스, 케빈 듀란트와 같이 현 리그에서 가장 뛰어난 선수들의 시즌 평균 득점은 다른 맥스급 선수들은 큰 차이가 나지 않는 26-27점이다. 축구와 같이 다른 스포츠에서는 메시라는 독보적인 선수가 시즌 70골을 넣으면서 기록 측면에서도 다른 스타 선수들을 두배 차이로 따돌리지만, 농구에서는 그 마이클 조던의 커리어 최고 시즌 평균 득점조차 36점으로, 맥스급 선수보다 30%정도 높은 수치이다.

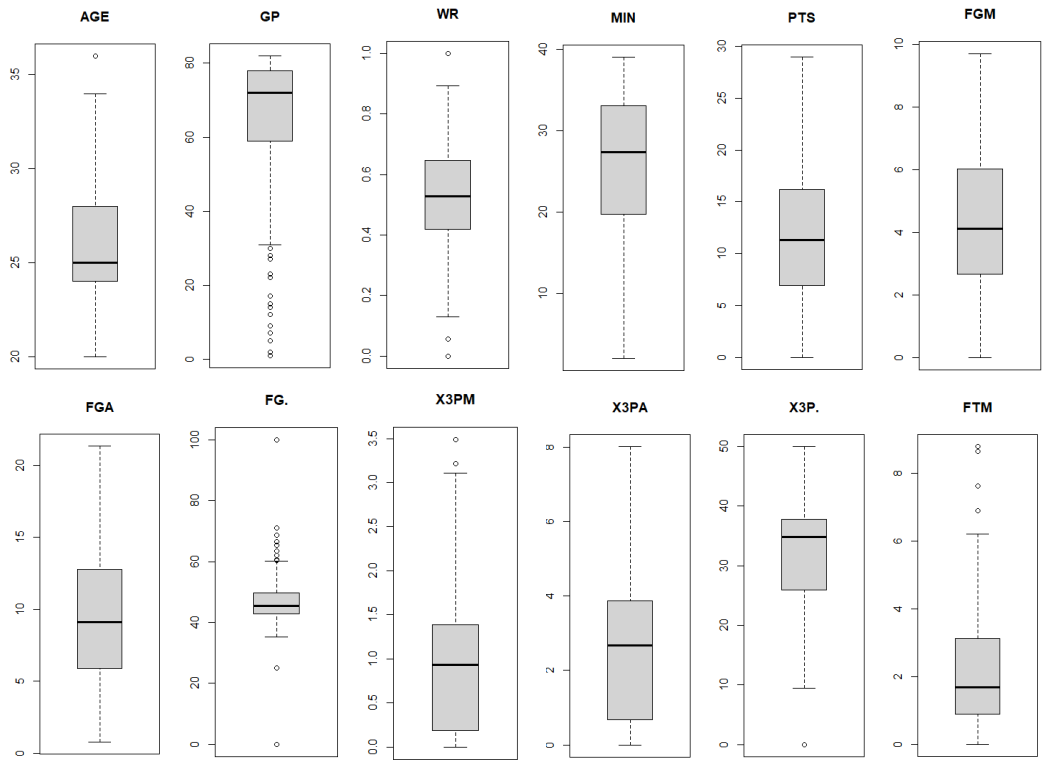
물론 실력이란 것은 굉장히 추상적인 개념이며 **Figure 4**은 예상일 뿐이다. 다만, 선수의 실력이 따라 연봉은 수렴하는 양상을 보며, 스탯 또한 수렴한다. 그렇다면 정량화 가능한 연봉(종속변수)과 스탯(설명변수)은 일정 수준에서 수렴하는 성질을 공유하며, 연봉은 스탯이 증가하면 연봉이 증가하는 증가함수 관계에 있으므로 본 보고서는 이 둘이 선형적인 관계라고 가정하겠다.

Q3.

단변량 통계량과 정규성

	mean	sd	kurtosis	skewness		mean	sd	kurtosis	skewness
AGE	25.93	2.84	**3.24**	**0.77**	**FTA**	2.92	2.18	**3.65**	**1.08**
GP	64.17	19.57	**4.72**	**-1.54**	**FT%**	74.03	14.98	13.07	-2.68
WR	0.53	0.17	**3.61**	**-0.26**	**OREB**	1.15	0.95	**4.16**	**1.27**
MIN	26.07	8.47	**2.58**	**-0.61**	**DREB**	3.64	2.03	**3.31**	**0.83**
PTS	12.09	6.34	**2.71**	**0.6**	**REB**	4.8	2.81	**3.72**	**0.94**
FGM	4.46	2.23	**2.42**	**0.43**	**AST**	2.59	2.25	**5.09**	**1.58**
FGA	9.54	4.69	**2.36**	**0.45**	**TOV**	1.54	0.91	**3.35**	**0.91**
FG%	46.74	8.09	17.04	**0.88**	**STL**	0.86	0.48	**3.01**	**0.62**
3PM	0.95	0.79	**2.97**	**0.66**	**BLK**	0.57	0.57	8.61	**2.04**
3PA	2.61	2	**2.45**	**0.41**	**PF**	2.06	0.67	**2.84**	**-0.06**
3P%	29.72	13.24	**3.51**	**-1.27**	**+/-**	0.69	3.49	**3.14**	**0.34**
FTM	2.23	1.75	**4.24**	**1.24**					

Figure 5 각 변수의 단변량 통계량



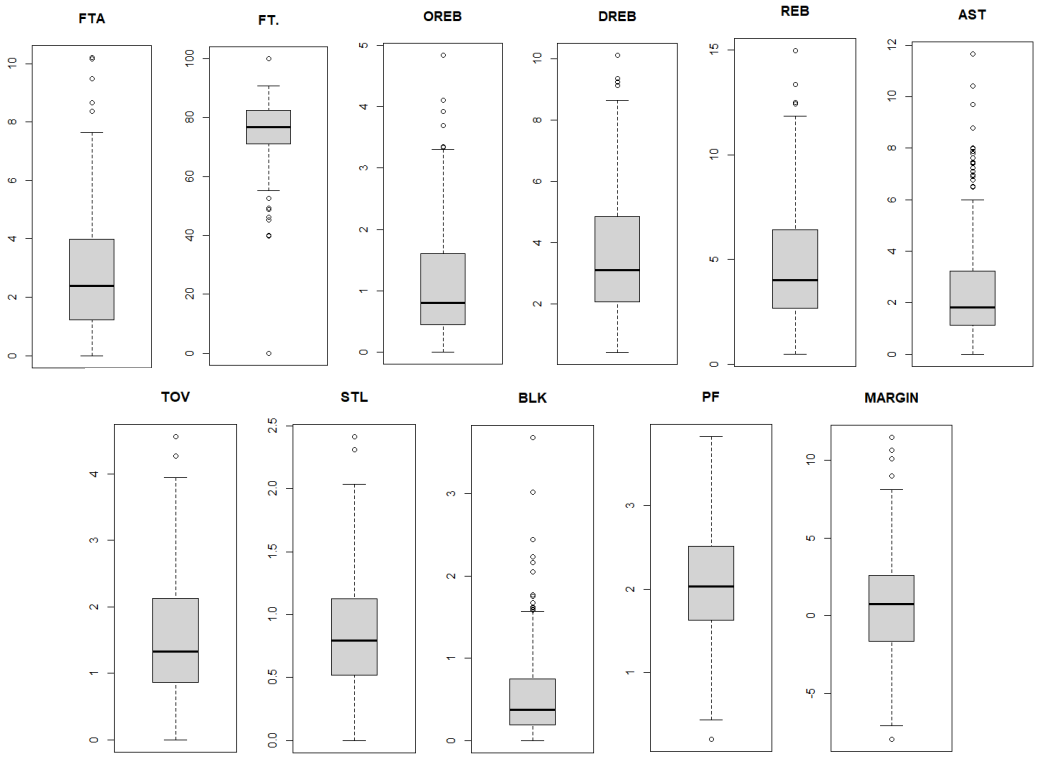


Figure 6 각 변수의 boxplot

Curran et al. (1996) 에 따르면 kurtosis와 skewness가 각각 (-2,2), (-7,7) 범위 내 일 때 정규성이 가정이 가능하다고 한다. **Figure 4**에서 해당 조건이 만족되는 값들은 볼드체로 처리했으며 두 조건을 모두 만족하는 변수들은 FG%, FT%, BLK을 제외한 나머지 20개이다. 해당 변수들에 한정해 추가적인 정규성을 확인하기 위해서 Shapiro-Wilks 검정과 QQ-plot을 진행한다.

	W	p-value		W	p-value
AGE	0.943	4.13×10^{-7}	**AST**	0.884	3.09×10^{-11}
GP	0.806	5.48×10^{-15}	**FGA**	0.903	4.50×10^{-10}
WR	0.988	$\mathbf{9.48\times10^{-2}}$	**3PM**	0.871	5.85×10^{-12}
MIN	0.948	1.31×10^{-6}	**3PA**	0.945	6.89×10^{-7}
PTS	0.960	2.27×10^{-5}	**3P%**	0.938	1.51×10^{-7}
FGM	0.969	2.11×10^{-4}	**FTM**	0.824	3.02×10^{-14}
FTA	0.965	7.50×10^{-5}	**TOV**	0.933	4.84×10^{-8}
OREB	0.929	3.00×10^{-8}	**STL**	0.967	1.118×10^{-4}
DREB	0.946	8.18×10^{-7}	**PF**	0.996	$\mathbf{8.70\times10^{-1}}$
REB	0.811	8.70×10^{-15}	**+/-**	0.991	$\mathbf{2.44\times10^{-1}}$

Figure 7 Shapiro-Wilks 검정 결과 (유의수준 = 5×10^{-2})

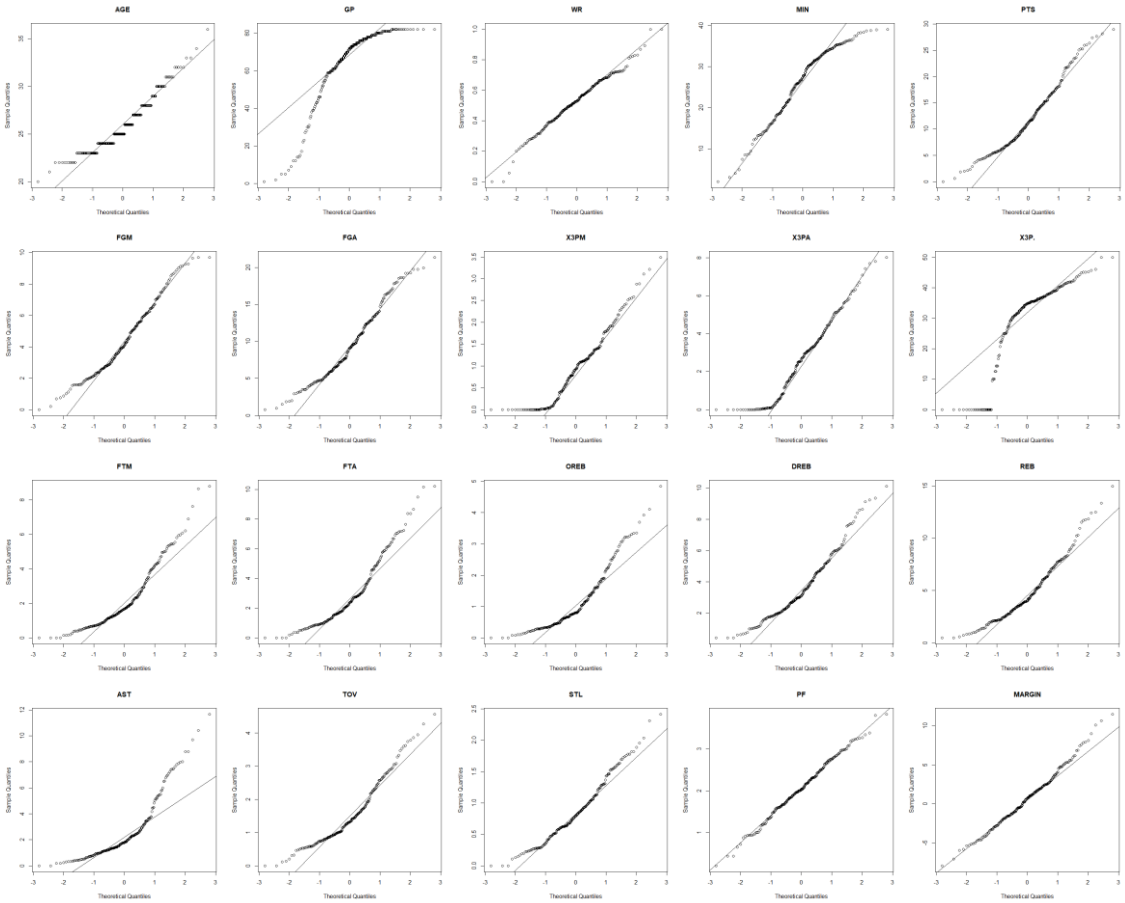


Figure 8 QQ-plot 결과

유의수준 0.05에서 Shapiro-Wilks 검정 **Figure 7.** 에 따르면 WR, PF, +/- 세가지 변수만이 정규성을 가진다는 것을 확인이 가능하다. **Figure 8.** QQ-plot 또한 WR, PF, +/-, 그리고 MIN, 3PM, 3PA에서 정규성을 확인이 가능하다.

Q4

Outlier 정의 및 제거

우선 Boxplot을 통해서 각 변수별로 (Q1 – 1.5IQR, Q3 + 1.5IQR) 외부에 위치한 관측치를 outlier로 판단할 수 있다. 하지만 23개의 변수 중 하나의 이상의 변수에서 outlier인 관측치의 개수는 총 79개로, 데이터 셋의 40%에 해당하는 양이기 때문에 모두 제거하기에는 무리가 있다.

	AGE	GP	WR	FG%	3PM	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	TOV	STL	BLK	PF	+/-
Over	1	-	1	9	2	-	4	5	1	6	4	4	20	2	2	15	-	4
Under	-	14	2	2	-	1	-	-	8	-	-	-	-	-	-	-	1	1

Figure 9 변수 별 outlier 개수

가능한 다른 접근법은 해당 관측치가 23개의 변수 중 outlier로 분류된 개수를 확인해, 많은 변수에서 outlier인 관측치만 drop하는 방법이다. **Figure 10**을 통해서 3+ outlier부터 frequency가 급격히 감소함을 확인할 수 있고, 따라서 Threshold를 2 미만으로 설정 시 데이터 양 측면에서 큰 손해를 볼 것이다. 본 보고서는 관측치 outlier의 Threshold를 3으로 설정하고, 13개의 관측치를 제거한 186개의 관측치를 사용하겠다.

Outlier 수	1	2	3	4	5	6	Total
Frequency	47	19	4	2	6	1	79

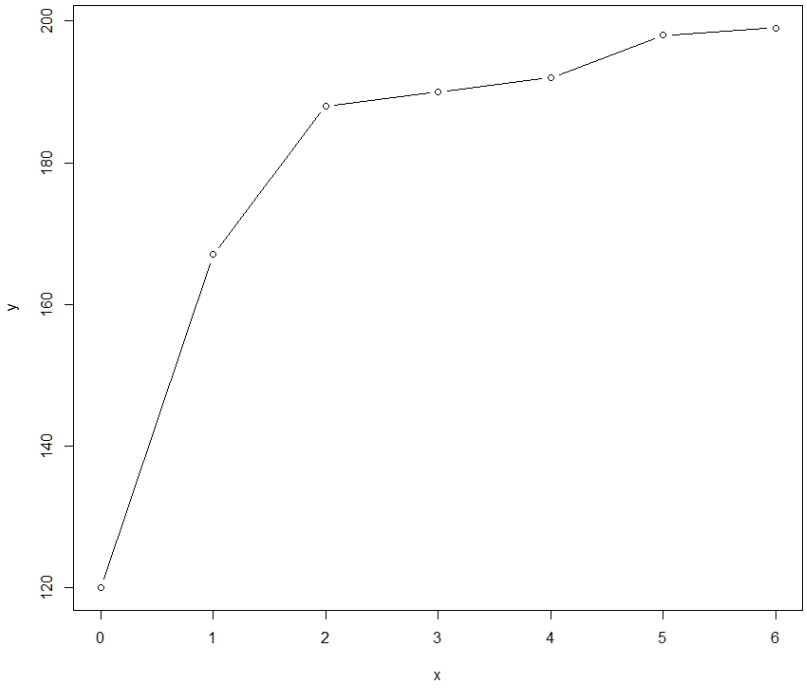


Figure 10 관측치별 outlier 개수, Threshold 설정 별 관측치 수

Q5

변수간 관계: 산점도, 상관계수

본 보고서에서 사용하는 변수들은 현재 서로 독립이 아니다. 리바운드 수는 공격과 수비 리바운드의 합이고, 경기를 뛰는 시간이 길수록 더 높은 볼륨을 기록할 것이다. WR와 +/-는 서로 밀접한 연관이 있을 것이고, 어시스트를 많이 기록하는 선수는 그만큼 턴오버를 기록할 것이다. MLR에서 서로 독립이 아닌 변수는 모델의 성능 저하로 이어질 수 있기 때문에 향후 제거를 위해서 각 설명변수 간 관계를 산점도와 상관계수를 통해서 확인한다.

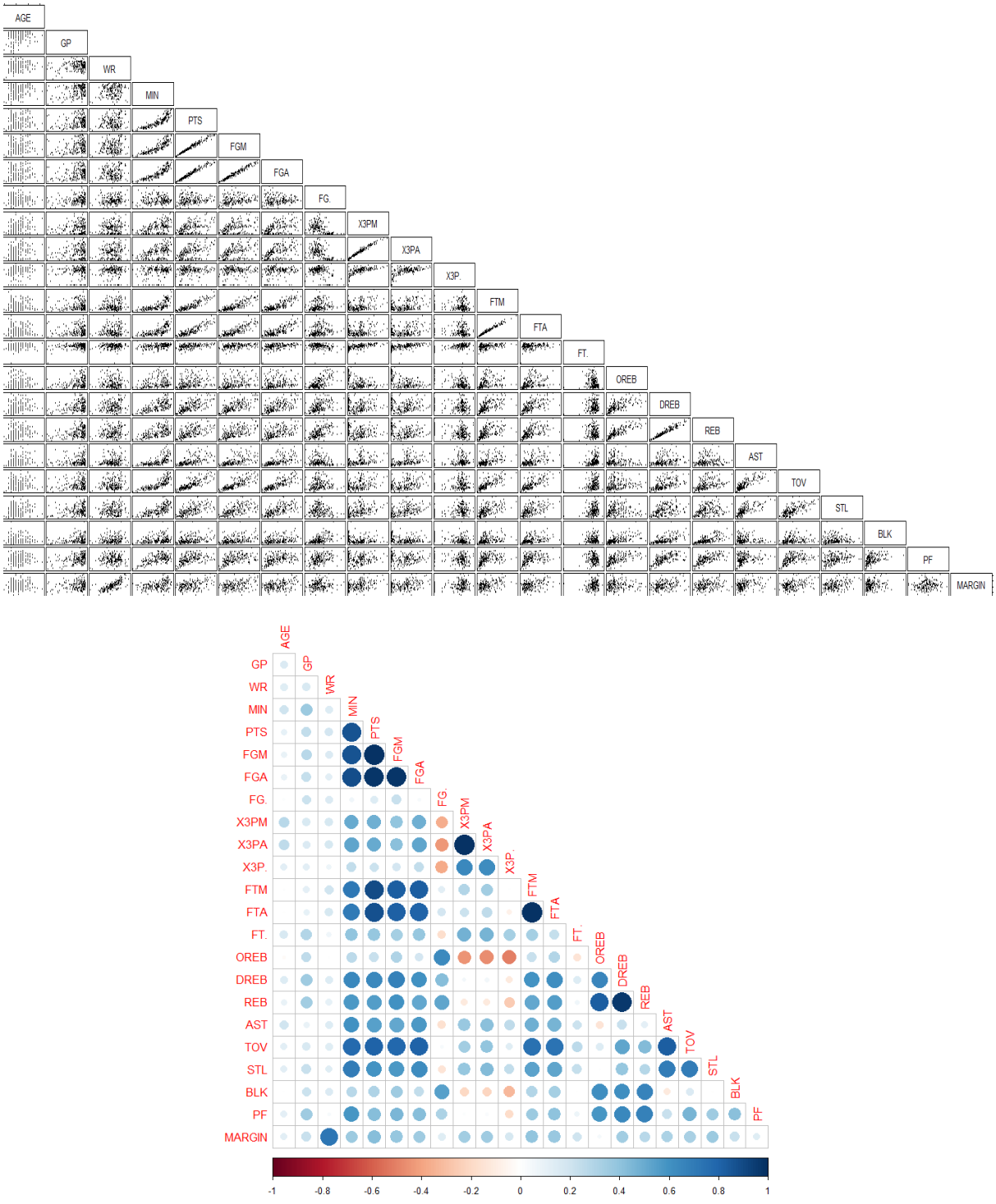


Figure 11 모든 변수에 대한 Scatterplot & Correlation

MIN, PTS, FGM, FGA는 각각 서로에게 강한 상관관계를 보이며, 이는 당연하다고 볼 수 있다. 많이 뛸수록 점수를 많이 낼 기회가 주어질 것이고, 기회가 많을수록 많은 점수를 올릴 것이며, 충분한 기회를 받아도 점수를 못 올리는 선수는 결국 많이 뛸 시간을 보장받지 못하게 될 것이기 때문이다. (FTM, FTA), (3PA, 3PM), (DREB, OREB, REB) 이 각각 높은 상관관계를 보이는

것 또한 자명하다; 시도를 많이 할수록 누적 성공 횟수가 높을 것이다. (AST, TOV, STL)은 모두 포인트가드 포지션과 연관된 스탯이다. 포인트가드 포지션은 농구에서 공을 오래 소유하며 어시스트를 올리는 역할로, 턴오버의 대부분이 실패한 어시스트 시도에서 발생하기 때문이다. 스틸의 경우 포인트가드 포지션을 성공적으로 수비할 때 주로 기록하는데, 농구에서 일반적으로 같은 포지션을 수비하기 때문에 포인트가드를 막는 선수 또한 포인트가드라 그렇다고 해석할 수 있다. (REB, DREB, OREB)와 (BLK, PF)는 모두 센터와 관련된 스탯들이다. 키가 큰 센터 선수가 골대 아래에서 리바운드를 잡는 역할로 리바운드 관련한 스탯이 높고, 골대 밑에서 성공적인 수비는 블락으로 연결되기 때문에 블락 또한 높다. 또한 골대 아래는 농구 코트에서 가장 몸싸움이 치열한 곳으로 센터 선수들의 퍼스널 파울이 일반적으로 높다. 농구는 점수를 많이 넣어야 이기는 게임이기 때문에 승률과 마진(+/-)은 예상처럼 관련이 있다. 이외에도 (MIN, PTS, FGM, FGA) 와 (OREB, DREB, REB, AST, TOV, STL)은 모두 볼륨 관련 스탯이기 오래 뛸수록(MIN) 다같이 높아질 것이다. 다만, 블락(BLK)만이 연관성이 없는 것이 특이하다.

마지막으로 눈여겨볼 점은 FGM, FGA는 FG%와 관계가 없고, FTM와 FTA는 FT%와 상관관계가 없지만, 3PM와 3PA, 3P%는 서로 강한 상관관계를 보이는 것이다. 이는 못 던져도 던지게 되는 다른 슛과 다르게 3점슛은 선택의 영역이기 때문에 잘못 던지는 (low 3P%) 선수는 3점을 던지지 않아서 (low 3PM) 연관성이 존재한다고 해석이 가능하다.

Q6

MLR 모델링

186개의 데이터를 7:3 비율인 training 데이터 130개와 validation 데이터 56개로 무작위로 분할하였다. 130개의 training 데이터에 대한 MLR 결과는 **Figure 12**와 같다. 이때 OREB, AST 단 두개의 변수의 p-value만이 threshold (0.05) 이하임을 볼 수 있다. 또한 FTM은 FTA를 통해서, REB는 DREB과 OREB을 통해서 얻을 수 있는 값들이기 때문에 NA를 반환한 것을 볼 수 있다. 이러한 현상들은 현재 변수들이 서로 강한 상관관계가 존재하기 때문이라고 추측할 수 있다.

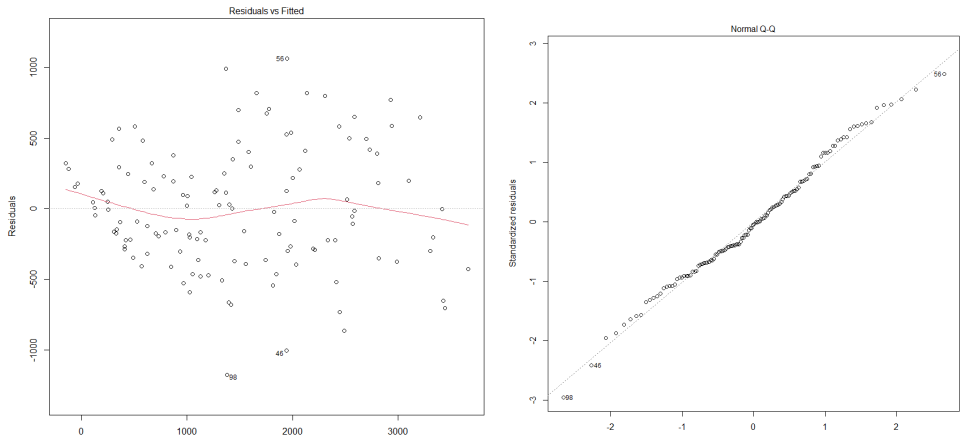
Variable	Estimate	Std.Error	t value	Pr(>|t|)
AGE	-5.31	19.29	-0.28	0.7835
GP	2.13	2.80	0.76	0.4491
WR	690.55	364.45	1.90	0.0608
MIN	13.54	17.16	0.79	0.4318
PTS	387.57	237.26	1.63	0.1053
FGM	-732.04	486.98	-1.50	0.1357
FGA	62.87	129.78	0.48	0.6291
FG.	-8.98	20.26	-0.44	0.6584
X3PM	-650.96	449.19	-1.45	0.1502
X3PA	127.76	163.84	0.78	0.4372
X3P.	-3.33	6.02	-0.55	0.5814

FTM	NA	NA	NA	NA
FTA	-120.94	197.12	-0.61	0.5408
FT.	-4.69	5.13	-0.92	0.3623
OREB	337.64	123.40	2.74	**0.0073**
DREB	-38.90	51.30	-0.76	0.4499
REB	NA	NA	NA	NA
AST	111.00	44.10	2.52	**0.0133**
TOV	-236.87	136.75	-1.73	0.0861
STL	-30.74	159.88	-0.19	0.8479
BLK	196.78	132.46	1.49	0.1403
PF	-119.65	111.95	-1.07	0.2876
MARGIN	32.89	21.17	1.55	0.1232

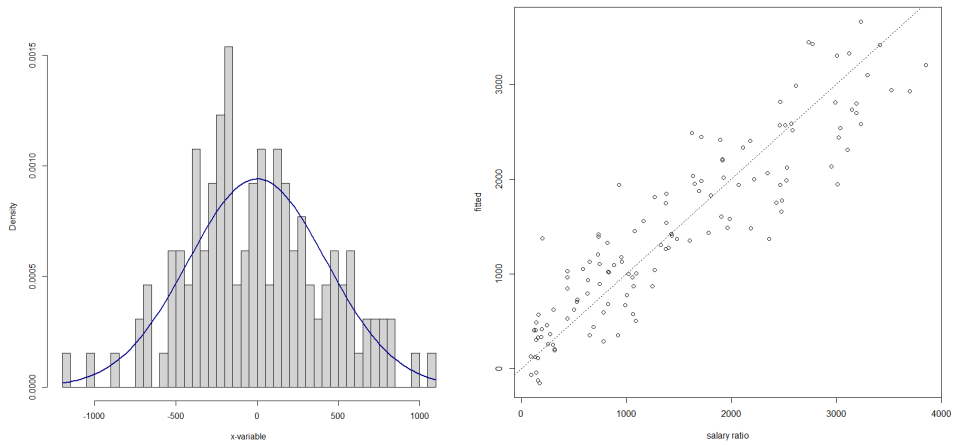
Adjusted R-squared = 0.7996

Figure 12 MLR 결과

(a) **(b)**



(c) **(d)**



Skewness=0.001, Kurtosis = 2.832

Figure 13 (a)잔차의 homoskedasticity, **(b)**QQ-plot, **(c)**정규분포와 비교, **(d)**모델의 선형성

OLE방식의 MLR은 다음 일반적으로 다음 4가지 가정을 만족해야 한다.

(1) 오차항이 정규분포를 따름: 잔차 분포의 Skewness와 kurtosis가 정규분포의 0,3 과 비슷하고 **(b)**와 **(c)**를 통해서 정규성을 만족함을 확인할 수 있다.

(2) 설명변수와 종속변수 사이에 선형관계가 성립함: **(d)**를 통해서 **Q2**의 선형성에 가정이 성립함을 확인할 수 있다.

(3) 각 관측치는 서로 독립: 포인트가드 인재가 희귀한 상황에서 준수한 선수가 시장에 2명이 있다고 가정하자. 이때 어느 구단이 둘 중 한 선수에게 높은 계약을 제시하면 다른 한 선수의 몸값 또한 상대적으로 증가할 것이다. 현실적으로 독립이라고 보기는 힘들다.

(4) 종속변수에 대한 오차항은 homoskedasticity를 만족함: **(a)**의 분포가 약간의 추세를 보이긴 하지만, 큰 틀에서 homoskedasticity를 확인할 수 있다

Adjusted R-squared 값이 0.8로 높고, 현실적으로 불가능한 (3)번 가정을 제외한 가정을 모두 만족하기 때문에 현 데이터가 MLR에 적합하다고 판단할 수 있다. 하지만 p-value가 너무 높게 나왔기 때문에 차원의 수를 줄일 필요가 있다.

Q7

변수 추출

Q6의 MLR 모델에서 유의수준(0.05)을 만족한 변수는 단 두개였다. 이는 서로 종속적인 변수들이 너무 많기 때문이라고 해석이 가능하였고, 변수 개수를 절반 이하로 감소시켜 모델링을 했을 시 더 나은 결과를 기대할 수 있다. 다만, 변수 두개만 사용할 수는 없기 때문에, 유의수준을 조금 높였을 때 상기한 모델에서 그나마 낮은 p-value(0.2 이하)를 보인 변수들은 다음과 같다: WR, PTS, FGM, 3PM, OREB, AST, TOV, BLK, MARGIN. 이를 Q5에서의 예측과 합치면 다음과 같은 목록을 얻을 수 있다. 이 체크리스트를 통해서 최대한 예측과 모델이 겹치는 선에서 서로 연관된 변수가 없도록 최종적으로 사용할 변수들을 선택하였다.

변수	예측	모델	연관된 다른 변수	최종 선택
WR	O	X	MARGIN	O
MIN	X	O	PTS, FGM, FGA, 3PM, 3PA, FTM, FTA, REB, OREB, DREB, AST, TOV, BLK,	X
PTS	O	O	MIN, FGM, 3PM, FTM	X
FGM	O	O	FGA, MIN, 3PM	O

FGA	X	O	FGM, MIN, 3PA	X
3PM	O	O	MIN, 3PA	X
3PA	X	O	MIN	X
FTM	X	O	MIN, FTA	O
FTA	X	O	MIN	X
REB	X	O	MIN, OREB, DREB	O
OREB	O	O	MIN, REB,	X
DREB	X	O	MIN, REB,	X
AST	O	O	MIN	O
TOV	O	X	MIN	X
BLK	O	X	MIN	X
MARGIN	X	O	WR	X

Figure 14 변수 추출 기준

Q9

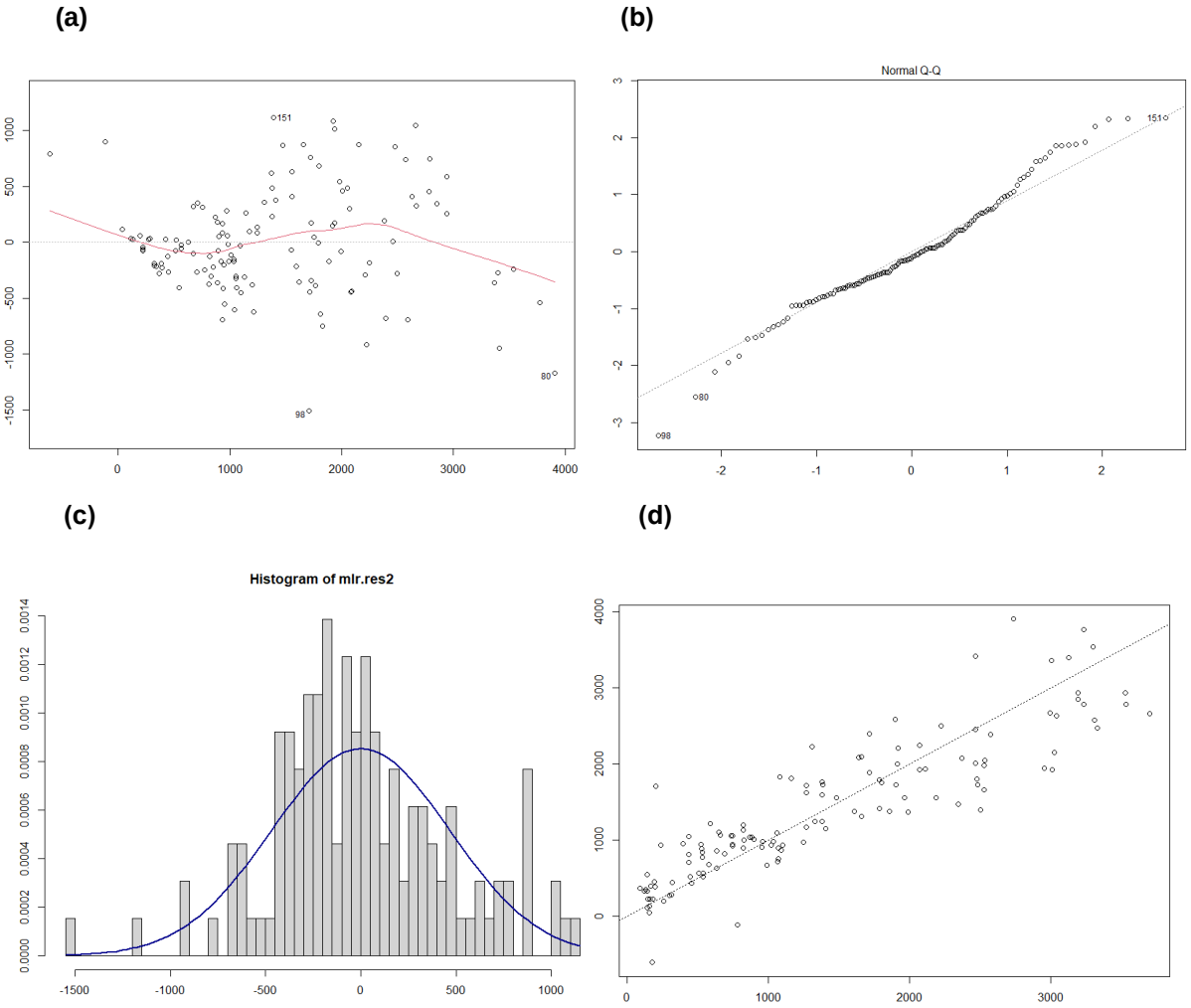
MLR 모델링 (with 추출된 변수)

	Estimate	Std.Error	t value	Pr(>|t|)
WR	1214.14	266.93	4.549	**0.00**
FGM	181.65	39.17	4.638	**0.00**
FTM	206.65	49.95	4.137	**0.00**
REB	86.91	23.5	3.699	**0.00**
AST	59.78	23.42	2.553	**0.01**

Adjusted R-squared = 0.80

Figure 15 추출된 변수들에 대한 MLR 결과

변수 추출 전에 비해서 사용된 모든 변수들이 통계적으로 유의미한 p-value를 가지는 것을 확인할 수 있다. 일반적으로 더 높은 야투, 자유투를 넣고, 팀이 더 좋은 성적을 기록하고, 더 많은 리바운드와 어시스트를 기록할수록 선수의 이듬해 연봉이 높아지는 것을 알 수 있다. 다만, 개인의 연봉과 팀의 성적은 무관할 것이라는 **Q2**예상과는 다르게 다른 요소가 동일할 때 전년도 팀의 성적이 우수했다면 그 공로를 인정받아 다음해 더 좋은 계약을 얻을 수 있다는 것을 확인할 수 있다.



skewness = 0.126, kurtosis = 3.455

Figure 16 (a)잔차의 homoskedasticity, **(b)**QQ-plot, **(c)**정규분포와 비교, **(d)**모델의 선형성

OLE의 4가지 가정

(1) 오차항이 정규분포를 따름: 잔차 분포의 Skewness와 kurtosis가 정규분포의 0,3 과 비슷하지만 **(b)**와 **(c)**를 통해서 정규성을 가정하기에는 힘들다는 것을 확인할 수 있다.

(2) 설명변수와 종속변수 사이에 선형관계가 성립함: **(d)**를 통해서 Q2의 선형성에 가정 또한 성립함을 확인할 수 있다.

(3) 각 관측치는 서로 독립: **Q7**과 마찬가지로 가정하기는 어렵다.

(4) 종속변수에 대한 오차항은 homoskedasticity를 만족함: **(a)**에서 잔차의 homoskedasticity를 확인할 수 있다

Adjusted R-squared 값이 0.8 정도로 높은 편이고, (1), (3)을 만족하지 못하지만 (1)의 정규성이 완전 없는 것은 아니며, (3)은 현실적으로 불가능하기 때문에 현 데이터를 MLR함에 큰 무리가 없다.

Q8 & Q10

모델 평가

Model	MSE	MAE	MAPE
Full variables	78.1 (254.4%)	**388.5**	**49.0**
Reduced variables	**30.7**	398.3 (102.5%)	57.5 (117.3%)

Figure 17 두 MLR에 대한 평가지표 비교, ()안의 값은 가장 좋은 값에 대한 비

모든 변수들을 사용한 모델이 MAE와 MAPE에서는 앞서지만, MSE에서는 추출된 변수들을 사용한 모델이 앞선다. 이는 추출된 모든 변수를 사용한 모델이 일반적으로는 더 잘 맞추지만, 한번 틀릴 때 큰 차이로 틀리기 때문에 MSE에서 이러한 차이를 보인다고 생각이 가능하다. 다만 관측치 수 자체가 그리 크지 않기 때문에 이런 평가 지표들이 정확하다고 말하기 힘든 부분이 존재한다.

Extra Q

시드에 따라 어떻게 테스트 데이터와 훈련 데이터를 구성하는지가 달라지고 그에 따라서 평가지표와 adjusted R-squared 또한 달라진다. 따라서 랜덤으로 7:3으로 나누는 과정을 10번씩 진행하고, 평균치를 비교하는 것이 더 신뢰성이 있다.

Model	**MSE**	**MAE**	**MAPE**	**Adjusted R^2**
Full variables	72.6 (116.0%)	393.3 (104.4%)	56.9 (108.2%)	**8.363**
Reduced variables	**62.6**	**376.6**	**52.6**	8.028 (96.0%)

Figure 18 MLR 평가지표의 10회 평균, ()안의 값은 가장 좋은 값에 대한 비

모든 지표에서 변수 수를 감소시킨 모델이 우월함 알 수 있다. Q8&10의 모든 변수를 사용한 모델이 높은 MSE값을 기록한 이유가 sampling과정에서 적은 sample(=1)로 인해 생긴 것임을 알 수 있다.

머신 러닝에서 서로 다른 두 모델이 비슷한 성능을 보인다면 더 단순한 모델이 우월하다고 한다. 본 보고서의 두 모델을 비교하면 추출된 변수만을 사용한 모델이 약 10%정도 우월한 성능지표를 기록했고, adjusted R-squared 값이 비슷하며, 변수들이 통계적으로 더 의미 있다고 해석할 수 있다. 따라서 더 단순한 모델임에도 모든 면에서 현재 데이터셋에 대해서는 추출된 변수를 사용한 모델이 우월하다고 할 수 있다.