REPORT

과목명 :	다변량해석
담당교수 :	안홍엽 교수님
제출일 :	2022년 6월 2일
이과대학	통계학과
학번/이름	2015110532 송호영 2018110481 이승준 2019110488 선주영 2020110494 김민준

목차

1. 서론 ····· 2

2. 본론

1) 자료 수집 ····· 2

2) 자료 처리 ····· 3

3) 자료 시각화 ····· 4

4) 주성분분석 ····· 5

5) 다변량분산분석 ····· 6

3. 결론 및 논의

1) 정리 ····· 8

2) 한계 ····· 9

4. 부록

1) 부록 ·····

2) 참고문헌 ·····

3) R Code ·····

운동 수행 능력 기준의 성별 집단 세분화

1. 서론

일상생활 속 우리는 남녀의 체력적인 차이를 느낄 수 있다. 평균적으로 남성이 여성보다 더 강한 근력을 가지고 있고, 여성은 남성보다 일반적으로 더 유연함을 보인다. 올림픽, 월드컵과 같은 국제 운동 경기에서도 남녀는 따로 구분하여 경기를 진행한다. 이처럼 남녀 간의 운동 수행 능력에 차이가 있다는 것은 모두가 공감하는 사실이다.

하지만 예외적인 경우도 분명 존재한다. 남성보다 힘이 센 여성, 여성보다 유연한 남성들도 존재한다. 성별로만 구분하는 것에 한계가 있기 때문에, 복싱, 레슬링 등의 시합은 체중에 따른 체급을 구분한다.

그리하여 본 조는 이 연구를 통해, 신체 정보를 가지고 있는 변수들에 주성분 분석(PCA)을 적용하여 추가적인 척도를 생성하고자 한다. 이후, 생성된 척도로 구분된 새로운 집단들이 성별과 함께 운동 수행 능력의 차이에 추가적인 영향을 주는지 다변량분산분석(MANOVA)을 통하여 판단할 것이다. 결과적으로 운동 수행에 있어 집단을 효과적으로 구분하고자 한다.

2. 본론

1) 자료수집

이 연구에서는 한국 사람 13,393명을 대상으로 조사한 7가지 신체변수와 4가지 운동 수행 변수, 그리고 운동 수행 능력에 따른 등급을 나타내는 자료를 사용하였다. 본 연구에서 사용한 자료의 출처는 다음과 같다.

자료 이름	페이지 url
bodyPerformance.csv	https://www.kaggle.com/datasets/kukuroo3/body-performance-data

<자료 출처, Table 2.1>

분석에 최종적으로 활용한 변수는 다음과 같다. 실제 자료에서는 class(운동 수행 능력에 따른 등급) 변수가 추가로 있으나 이 연구에 필요한 데이터는 아니라고 판단하여 분석을 할 때는 <Table 2.2>의 변수만 활용하였다.

변수 종류	변수 이름	변수 순번	변수 설명
신체 변수	age	1	나이
	gender	2	성별
	height_cm	3	키(cm)
	weight_kg	4	몸무게(kg)
	body.fat_.	5	체지방률(%)
	diastolic	6	이완기 혈압
	systolic	7	수축기 혈압
운동 수행 변수	gripForce	8	악력
	sit.and.bend.forward_	9	앉아 윗몸 앞으로 굽

	cm		히기(cm)
	sit.ups.counts	10	윗몸 일으키기
	broad.jump_cm	11	제자리 멀리뛰기(cm)

<변수 설명, Table 2.2>

2) 자료 처리

본격적인 분석에 앞서, 자료에 결측치나 이상치가 있는지 확인하기 위해 변수별 요약 통계를 살펴보았다. 모든 숫자는 소수점 넷째자리에서 반올림하였다.

변수 이름	표본 크기	F	M
Gender	13393	4926	8467

<남녀 비율, Table 2.3>

변수 이름	표본평균	표본 표준편차	최솟값	중위수	최댓값
age	36.775	13.626	21	32	64
height_cm	168.560	8.427	125	169.2	193.8
weight_kg	67.447	11.950	26.3	67.4	138.1
body.fat_.	23.240	7.257	3	22.8	78.4
diastolic	78.797	10.742	0	79	156.2
systolic	130.235	14.714	0	130	201
gripForce	36.964	10.625	0	37.9	70.5
sit.and.bend.forward_cm	15.209	8.457	-25	16.2	213
sit.ups.counts	39.771	14.277	0	41	80
broad.jump_cm	190.130	39.868	0	193	303

<변수 별 요약통계량, Table 2.4>

변수 별 요약통계<Table 2.4>를 확인한 결과, 결측치는 존재하지 않지만 체지방률, 앉아 윗몸 앞으로 굽히기 변수의 최댓값과 이완기 혈압, 수축기 혈압, 악력, 제자리 멀리뛰기 변수의 최솟값에 잠재적인 이상치가 존재한다. 따라서 각각 상위 10개, 하위 10개의 값을 살펴보았다.

body.fat_.	78.4	54.9	53.5	50.6	50.3	50.2	49.8	49.3	49.2	48.9
sit.and.bend.forward	213.0	185.0	42.0	40.0	40.0	37.0	35.2	35.2	35.2	35.2
diastolic	0	6	8	30	37	40	41	41	42	42
systolic	0.0	14.0	43.9	77.0	82.0	84.0	86.0	86.0	86.0	88.0
gripForce	0.0	0.0	0.0	1.6	2.1	3.5	4.4	5.3	6.7	7.9
broad.jump	0	0	0	0	0	0	0	0	0	0

<상위 10개-하위 10개, Table 2.5>

b_fat	gender	h_cm	w_kg	diastolic	class
78.4	M	177.6	74.5	69	A
54.9	M	172.8	95.0	90	D
53.5	F	162.5	113.3	82	D
50.6	F	160.0	109.2	94	D
50.3	F	160.0	76.9	82	D

<체지방률 상위 5개 관측치, Table 2.6>

먼저 체지방률의 최댓값이 나올 수 없는 값이라고 판단했다. 78.4는 다른 값들과 큰 격차를 가지고 있고, 체지방률 50% 이상인 5명의 사람들의 정보<Table 2.6>를 보았을 때, 혼자서 A등급

을 맞은 것을 근거로 이상치라고 판단했다.

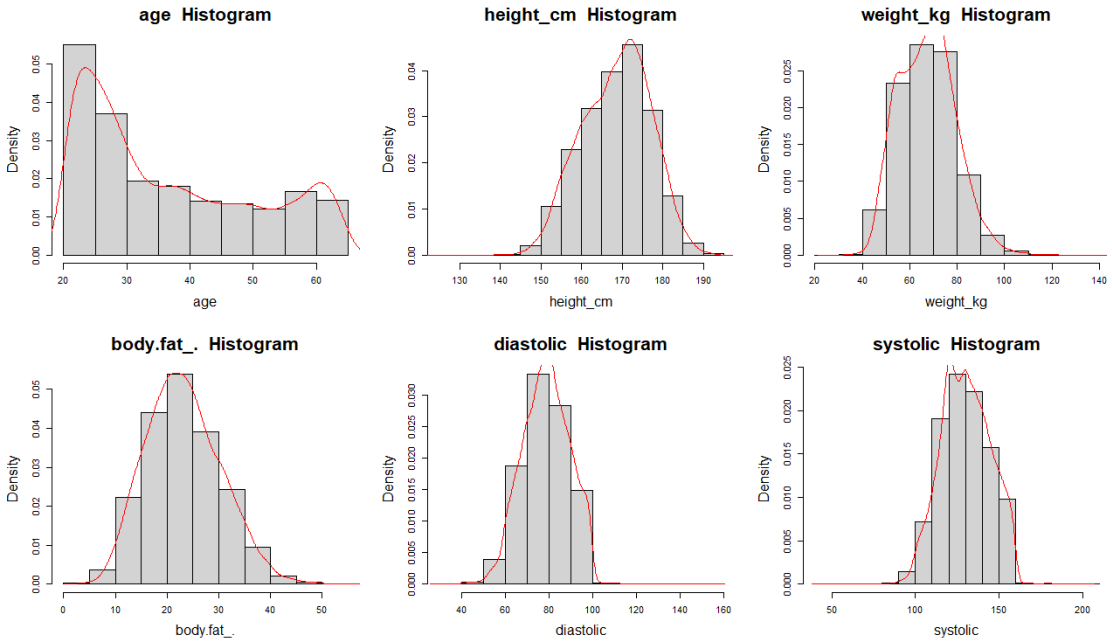
다음으로 앉아 윗몸 앞으로 굽히기 변수 또한 최댓값을 살펴보았다. 이 중 213과 185가 다른 값들과 비교하여 유독 크다. 앉아 윗몸 앞으로 굽히기는 앉아서 발 끝보다 손가락이 얼마나 더 멀리 떨어져 있는지를 측정하기 때문에, (팔 길이를 포함한 본인의 상체 길이 - 하체 길이)가 최댓값이다. 따라서 213cm, 185cm는 불가능한 값들이기 때문에 이를 이상치로 판단했다.

이완기 혈압과 수축기 혈압의 최솟값에 대해 살펴보았다. 가정의학과에서 정의한 저혈압의 기준이 60, 90 이하인 것을 생각하였을 때, 이완기 혈압 0,6,8, 수축기 혈압 0.0, 14.0은 불가능한 수치라고 생각하여, 이를 이상치로 판단했다.

같은 방법으로 악력과 제자리 멀리뛰기 변수의 경우도 악력기를 잡거나, 한걸음만 걸어도 0이 넘는 측정값이 나오기 때문에 0을 모두 이상치로 판단했다.

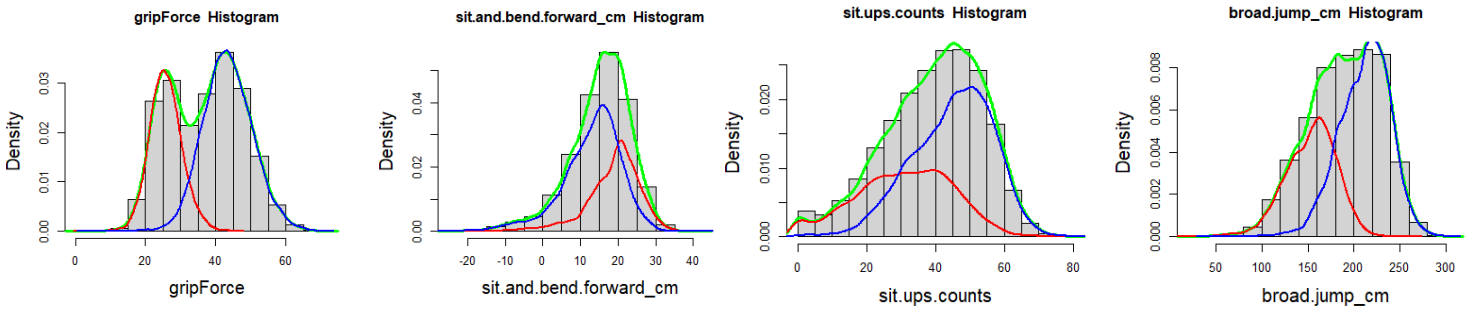
최종적으로 위에서 판단된 19개의 이상치들을 제거하여, 13,374개의 자료로 분석을 시작했다.

3) 자료 시각화



> <신체 변수 히스토그램, 그림 2.1>

<그림 2.1>에 의하면 신체 변수들 중에서 나이를 제외한 변수들은 끝부분에서 정규성을 벗어나는 모습을 보이지만, 큰 문제가 없다고 생각했다. 하지만 나이의 경우는 히스토그램과 Q-Q Plot(부록1)을 바탕으로 정규성을 만족하지 않음을 알 수 있다. 원래 주성분분석은 자료들이 정규분포를 따라야 주성분들의 설명력이 보장된다는 가정이 있지만, 본 연구에서는 설명력이 필요하지 않다고 판단하여 본 자료를 주성분분석에 사용하였다.



<운동 변수 히스토그램 - 남녀 구분, 그림 2.2>

<그림 2.2>는 악력, 앉아 윗몸 앞으로 굽히기, 윗몸 일으키기, 제자리 멀리뛰기의 히스토그램과 비율을 고려한 밀도 선 그래프이다. 녹색은 전체, 청색은 남성, 적색은 여성을 나타낸다.

이 중 악력은 남성과 여성의 정규분포가 결합된 쌍봉형태의 분포를 나타낸다. 또한, 여성들의 윗몸 일으키기 분포는 왼쪽벽이 존재한다는 사실을 알 수 있는데, 이는 아무리 못해도 0보다 작은 결과가 나올 수 없기 때문이다. 이러한 경향성을 제외하면, 남성과 여성이 오른쪽으로 치우친 모양을 나타낸다. 윗몸 일으키기에는 팔로 머리 당기기, 허리 반동 이용하기 등 올바르지 않은 자세로 측정하는 사람이 많기 때문에, 본인의 실제 성적보다 좋은 결과를 가진 사람들이 있어 오른쪽으로 치우쳤다고 생각한다. 이외의 변수들은 남녀 모두 정규분포에서 아주 조금씩 벗어난 모습을 보인다. 결론적으로 운동 수행 변수에서 남성과 여성들의 차이가 있음을 알 수 있다.

4) 주성분 분석(PCA)

운동 수행 능력의 차이를 구분할 새로운 척도를 찾기 위하여, 우리는 성별을 제외한 신체 특징들을 종합해줄 지표가 필요하다. 그래서 우리는 6개의 신체 특징 변수 age(나이), height(키), weight(몸무게), body.fat.(체지방률), diastolic(이완기 혈압), systolic(수축기 혈압)으로 주성분 분석을 진행하였다. 각 변수 별 단위가 다르기 때문에 상관계수행렬을 이용하였다.

주성분	PC1	PC2	PC3	…
주성분 설명력	37.986%	28.890%	13.721%	…
설명력 누적합	37.986%	66.876%	80.598%	…

<주성분 설명력(분산 비율), Table 2.7>

<Table 2.7>(부록 8)을 살펴보면 3개의 주성분의 설명력이 80.598 %로 일반적으로 생각하는 기준인 70%를 넘긴다. 표준화 변수의 선형결합인 주성분의 계수(주성분계수)는 다음과 같다.

신체 변수	PC1	PC2	PC3	…
age	0.106	0.515	0.121	…
height_cm	-0.556	-0.323	-0.105	…
weight_kg	-0.535	-0.052	-0.561	…
body.fat_.	0.276	0.419	-0.759	…
diastolic	-0.374	0.486	0.185	…
systolic	-0.421	0.464	0.221	…

<주성분 계수, Table 2.8>

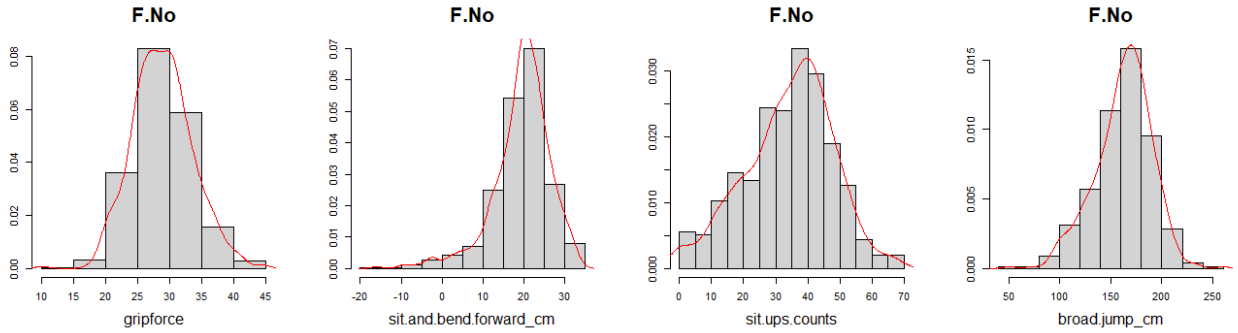
먼저 PC1의 경우 키, 몸무게, 수축기, 이완기 혈압의 유의미한 음의 방향성을, 체지방률은 약한 양의 방향성을 확인하였다. 이는 키, 몸무게, 혈압(수축기, 이완기)이 높을수록 PC1의 값이 낮아진다는 결과와 체지방이 높을수록 PC1의 값이 증가한다 라는 결과를 의미한다. 실제로 여성들의 키와 체중이 더 작은 편이며, 대한비만학회에서 발표한 비만의 기준에 의하면 여성들의 체지방 기준이 높다. 또한 미국의학협회 심장병학저널(JAMA Cardiology)에서 발표한 바에 의하면, 여성의 혈관 노화 속도가 호르몬에 의해 남성보다 빠르다고 한다. 즉, 여성들은 남성들에 비해 좁은 혈관과 높은 혈압을 가진다. 또한 <부록 2>에 의하면 여성들의 PC1 평균이 남성들의 PC1 평균보다 높았다. 이러한 근거들을 바탕으로 본 조는 PC1을 신체적 여성 성향으로 해석했다.

PC2는 나이, 체지방률, 혈압과 양의 방향성을, 키에 음의 방향성을 가진다. 인간은 노화가 진행되면서 키는 줄어들고, 체지방과 고혈압 위험성은 높아진다. 이를 바탕으로 PC2를 노화정도, 신체나이로 해석했다. 또한 PC3는 체지방률, 몸무게와 강한 음의 방향성을 보이기 때문에, 이를 반(反)비만도로 해석했다.

결론적으로 본 조는 성별 외에 추가적으로 운동 수행 능력을 기준으로 집단을 구분할 기준이 될 수 있는 PC1을 이어지는 분석에서 사용했다. 남성들의 PC1 평균과 여성들의 PC1 평균의 중간값을 기준으로 이보다 작으면 “신체적 남성”, 크면 “신체적 여성”으로 구분하여 이것이 실제 성별과 일치하면 “1”, 일치하지 않으면 “0”을 갖는 gender.test, “성향 일치”라는 범주형 변수를 생성하였다. 그리하여 성향 일치 변수를 기준으로 새로운 집단을 구성했다.

5) 다변량분산분석(MANOVA)

성별, 혹은 성향 일치 간의 차이가 있는지 확인하기 위하여, 본 조는 악력, 앉아 윗몸 앞으로 굽히기, 윗몸 일으키기, 제자리 멀리뛰기, 4개의 반응변수를 이용한 다변량분산분석을 진행하였다.



<신체적 성별이 남성인 여성의 운동 변수 히스토그램, 그림 2.3>

다변량분산분석을 사용하기 위해서는 각 모집단별 반응변수가 다변량 정규분포를 따라야 하며 공분산 행렬이 동일해야 한다. <그림 2.3>은 4개의 집단 중 하나인 성향이 일치하지 않는 여성들의 운동 수행 능력 변수 히스토그램이다. 모두 종 모양을 나타내고 있으나, 윗몸 일으키기는 본문 3에서 설명한 것처럼 왼쪽 벽에 의해 그래프 좌측의 빈도수가 우측에 비해 높게 나온다.

이어서 운동 수행 변수들의 공분산 동질성을 고려하였다. (부록 10, 11, 12, 13) 4개 집단의 공분산들이 큰 차이가 없이 비슷하다고 판단할 수 있다. 본 조는 이러한 점들을 고려하여 다변량

분산분석을 진행하였다.

다변량분산분석을 위해 자료를 일렬로 나열한 후, 고유 ID, 성별, 성향 일치, 반응 변수 종류, 그리고 ID를 제외한 변수들의 결합을 변수로 추가했다. 그리고 한 사람의 반응 변수들이 서로 독립임을 가정할 수 없기 때문에, GLS(Generalized Linear Squares) 모델을 사용하여 분석을 진행했다.

y	id	gi	gk	gj	gikj	gij	gkj	gik
34.1	1	F	No	1	F.No.1	F.1	No.1	F.No
19.0	1	F	No	2	F.No.2	F.2	No.2	F.No
30.0	1	F	No	3	F.No.3	F.3	No.3	F.No
155.0	1	F	No	4	F.No.4	F.4	No.4	F.No
32.1	2	F	No	1	F.No.1	F.1	No.1	F.No

<선형 변환 후 자료 상위 5개의 관측치, Table 2.9>

우선 성별 간 운동 수행 능력에 차이가 있는지 Likelihood Ratio Test를 통하여 유의수준, α = 0.05 하에서 검정해보았다.

H_{0a} : 성별 간의 운동 수행 능력에 차이가 없다.

Full Model : $Y_{ijd} = g_{ij} + \varepsilon_{ijd}$ Reduced Model : $Y_{jd} = g_j + \varepsilon_{jd}$

검정 통계량 값은 26367.8, 자유도가 4인 카이제곱 분포를 근사적으로 따르며, p-value는 <.0001로 유의수준 하에서 귀무가설을 기각하였다. 이는 성별 간의 운동 수행 능력에 차이가 있다는 주장에 반박할 근거가 부족하다는 것으로 해석된다. 그렇다면 성별에 추가적으로 성향 일치 변수를 독립 변수로 추가한다면, 유의미한 차이가 있을지 알아보고자 한다.

H_{0b} : 성별에 추가적으로 성향 일치 변수를 추가하는 것은 도움이 되지 않는다.

Full Model : $Y_{ikjd} = g_{ikj} + \varepsilon_{ikjd}$ Reduced Model : $Y_{ijd} = g_{ij} + \varepsilon_{ijd}$

검정 통계량 값은 1005.47, 자유도가 1인 카이제곱 분포를 근사적으로 따르며, p-value는 <.0001로 유의수준 하에서 귀무가설을 기각하였다. 이는 성향 일치 변수를 추가하는 것이 4개의 반응변수를 기준으로 집단을 구분함에 있어서 도움이 된다는 것을 기각할 근거가 부족하다는 의미이다.

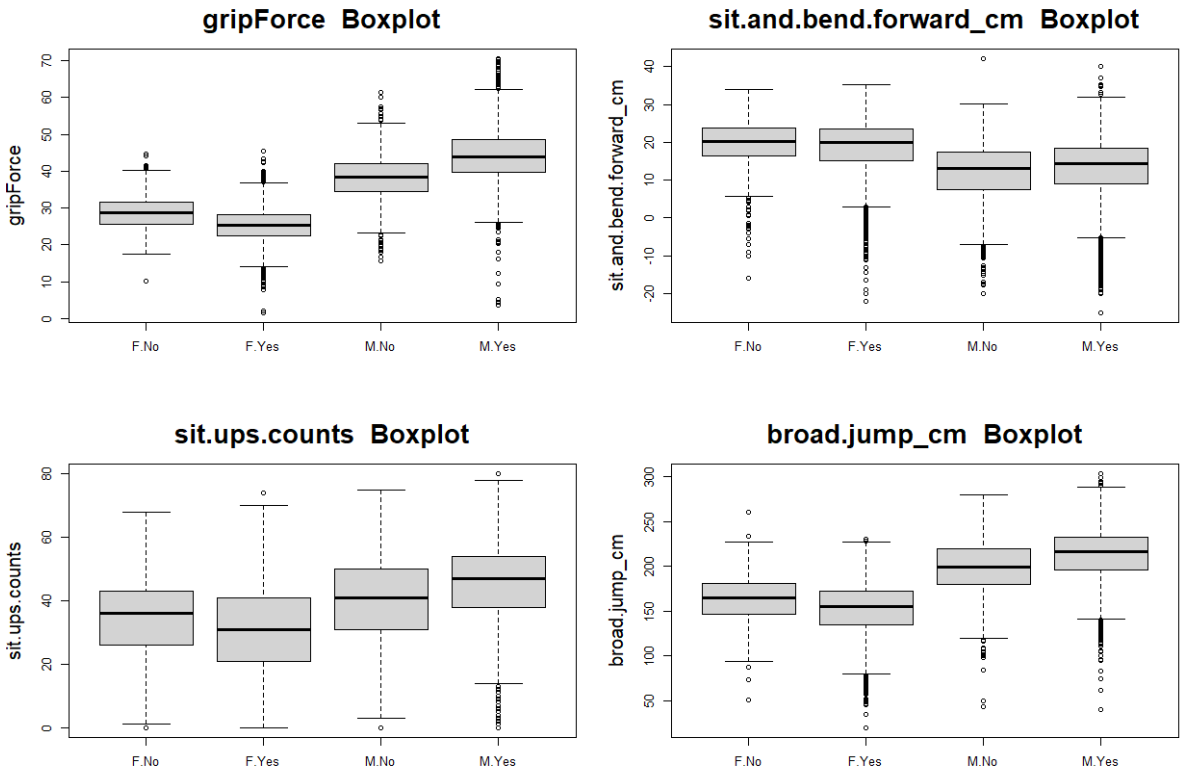
이번에는 성향 일치에 따라, 남성과 여성의 반응 변수 차이가 동일한지 검정하고자 한다. 이는 성향 일치와 성별 간의 교호작용이 없다는 의미로 해석할 수 있다.

H_{0c} : 성향 일치에 따라, 남성과 여성의 반응 변수 차이가 동일하다.

Full Model : $Y_{ikjd} = g_{ikj} + \varepsilon_{ikjd}$ Reduced Model : $Y_{ikjd} = (g_i + g_k) * g_j + \varepsilon_{ijd}$

검정 통계량 값은 750.77, 자유도가 1인 카이제곱 분포를 근사적으로 따르며, p-value는 <.0001로 유의수준 하에서 귀무가설을 기각하였다. 이는 성향 일치 여부가 성별 집단 간의 차이에 영향을 끼친다는 주장을 기각할 근거가 부족하다고 결론지을 수 있다.

세 번의 검정 결과를 종합한 결과, 본 조는 성향 일치와 성별에 따라 4개의 집단 [성향이 일치하는 여성, 성향이 일치하지 않는 여성, 성향이 일치하는 남성, 성향이 일치하지 않는 남성]으로 구분할 수 있다고 판단하였다.



<4개 집단의 4개 변수 Box-Plot, 그림 2.4>

최종적으로 <그림 2.4>를 보면 이 중 앉아 윗몸 앞으로 굽히기를 제외하면 유의미한 차이가 있는 것으로 보인다. 또한 <그림 2.4>와 최종 모델 추정량(부록 9)에 의하면 앉아 윗몸 앞으로 굽히기의 경우는 신체적 성향 일치 여부가 남녀의 차이에 큰 영향을 끼치지 않은 것으로 해석된다.

집단 별 운동 수행 변수의 분포도 살펴보았다. (부록 3, 4, 5, 6) 4개의 집단의 4개의 변수들의 분포는 모두 종모양을 보인다. 그러나 오른쪽 벽이 존재하는 앉아 윗몸 앞으로 굽히기, 왼쪽 벽이 존재하는 윗몸 일으키기 분포는 한쪽으로 치우친 분포의 형태를 띠고 있다는 점을 유의할 필요가 있다.

3. 결론 및 논의

1) 정리

지금까지 다변량 분석의 여러 기법을 통해 남녀 간의 운동 수행 능력을 구분하는 새로운 척도를 생성하고 그 척도가 성별과 함께 추가적으로 집단을 구분하는 것에 영향을 주는지 알아보았다. 본 연구에서 사용한 다변량 분석 기법은 다음과 같다.

① 기술적 통계 분석

② 자료 시각화 기법(히스토그램, Q-Q Plot)

③ 주성분분석(PCA)

④ 다변량분산분석(MANOVA)

우선, 기술적 통계 분석을 통해서 각 변수들의 이상치를 제거하고 히스토그램과 Q-Q Plot을 통해 자료들의 분포를 살펴보았다. 그 다음으로, 신체 변수로 주성분 분석을 진행하여 나온 주성분 중, 남성에 비해 여성에게서 두드러지는 특성과 유사한 PC1을 추가적인 분석의 대상으로 삼기로 하였다. PC1을 통해 분류한 성별과 실제 성별의 일치 여부를 “Yes”과 “No”로 변수화한 “성향 일치” 변수를 새로 만들었다. 그후 성별이나 성향 일치에 따라 운동 수행 변수가 차이 나는지 다변량분산분석(MANOVA)을 통해 분석하였고, 분석 결과 차이가 있다고 판단하였다. 성별과 성향 일치 간에도 관계가 있다고 판단하였다. 즉, 대부분 성향이 일치할 때 남녀의 운동 수행 능력 차이와 성향이 일치하지 않을 때 남녀의 운동 수행 능력 차이가 다름을 보였지만, 앉아 윗몸 앞으로 굽히기에서는 차이가 없음을 알 수 있었다.

2) 한계

다양한 다변량 분석 기법을 적절히 활용하여 본 연구의 목적에 근접한 결과를 얻어냈지만, 본 조의 분석에는 여전히 몇 가지 한계가 있었다. 우선, 신체 변수와 운동 수행 변수 모두 가짓수가 많지 않아 분석의 신뢰성을 충분히 확보하지 못했다. 신체 변수는 허리 둘레나 발 길이, 운동 수행 변수는 오래 달리기와 같은 변수들을 추가했다면 좋았을 것이다.

변수들이 동일한 조건에서 측정되었는지도 명시되지 않았다. 같은 날에 측정하였는지, 참가자들의 컨디션은 동일하였는지 등의 조건이 모두 같지 않았다면 자료의 신뢰성을 확보할 수 없다.

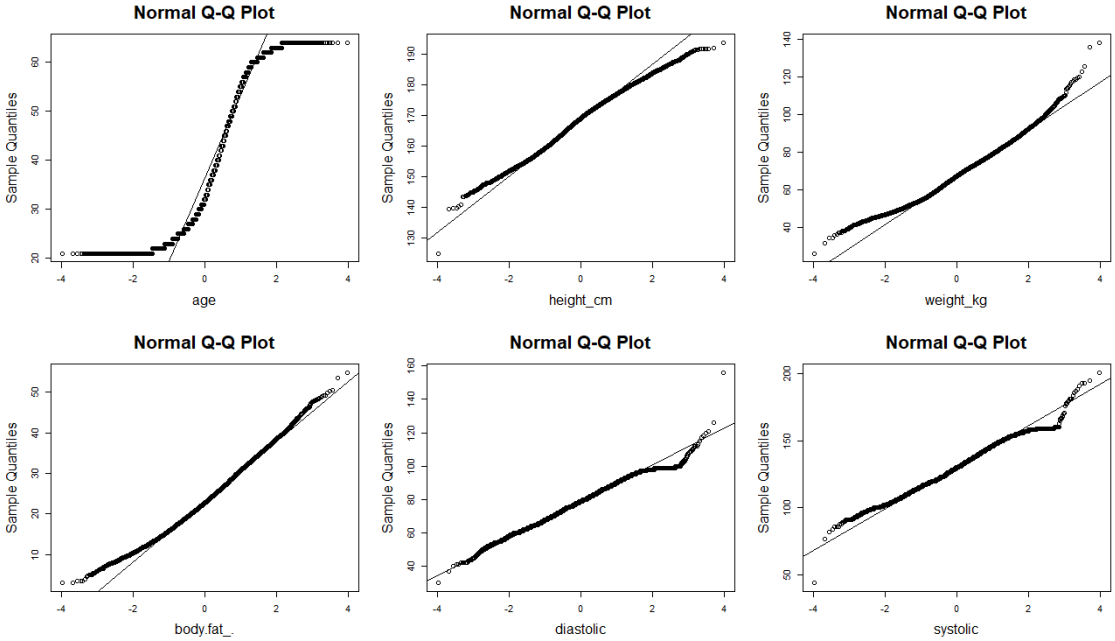
또한 터무니없는 이상치가 있었기 때문에 자료 측정의 신뢰성이 떨어진다. 예를 들어 제자리멀리뛰기가 0 cm로 나왔다면 한 발짝도 떼지 않았다는 뜻이므로 일반적으로는 재측정을 하였을 것이다. 하지만 0 cm로 기록되어 있다는 것은 참가자가 측정을 하지 않았거나 기록이 누락되어 0 cm로 기록되었다고 생각할 수 있고, 이는 측정과 기록이 제대로 이루어지지 않았음을 의미한다. 따라서 정상적인 것 같아 보이는 자료들도 잘못 측정되거나 기록되지 않았나 의심할 수 있다.

이상치를 제거할 때 통계적인 방법을 사용하지 않고 사람이 직접 자료를 나열하여 골라내는 방법을 사용했다. 수축기 혈압 43.9나 몸무게 26.3 kg과 같은 자료는 이상치인지 아닌지 판단하기 애매하고 어떤 방법을 사용하느냐에 따라 이상치로 분류될 수도, 그렇지 않을 수도 있다.

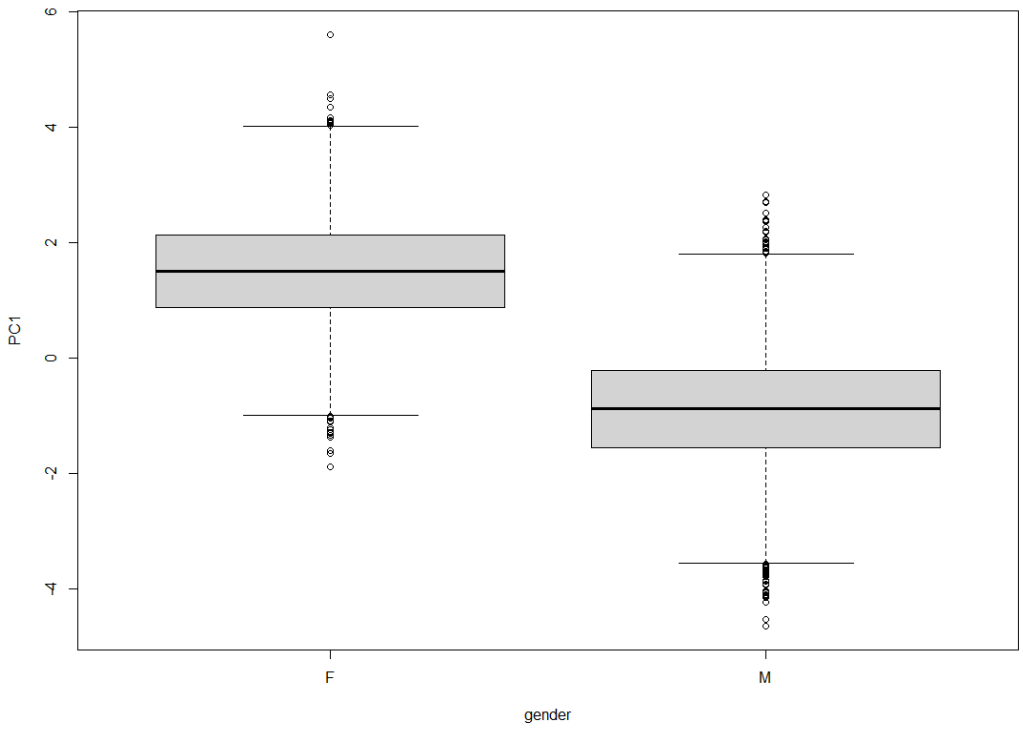
결론적으로 운동수행능력 데이터 분석에서 기본적으로 주어진 정보뿐만 아니라 추가적으로 수집할 수 있는 다양한 정보를 활용하여 분석을 하는 것이 좋은 방법이 될 것 같다. 또한 외적 요인을 고정시킨 상태에서 자료를 수집하고 분석하는 것도 고려해볼 만한 사항이다. 앞으로도 이러한 점들을 고려하여 과학적이고 체계적인 방법으로 운동수행능력에 대한 분석이 이뤄져야한다.

4. 부록

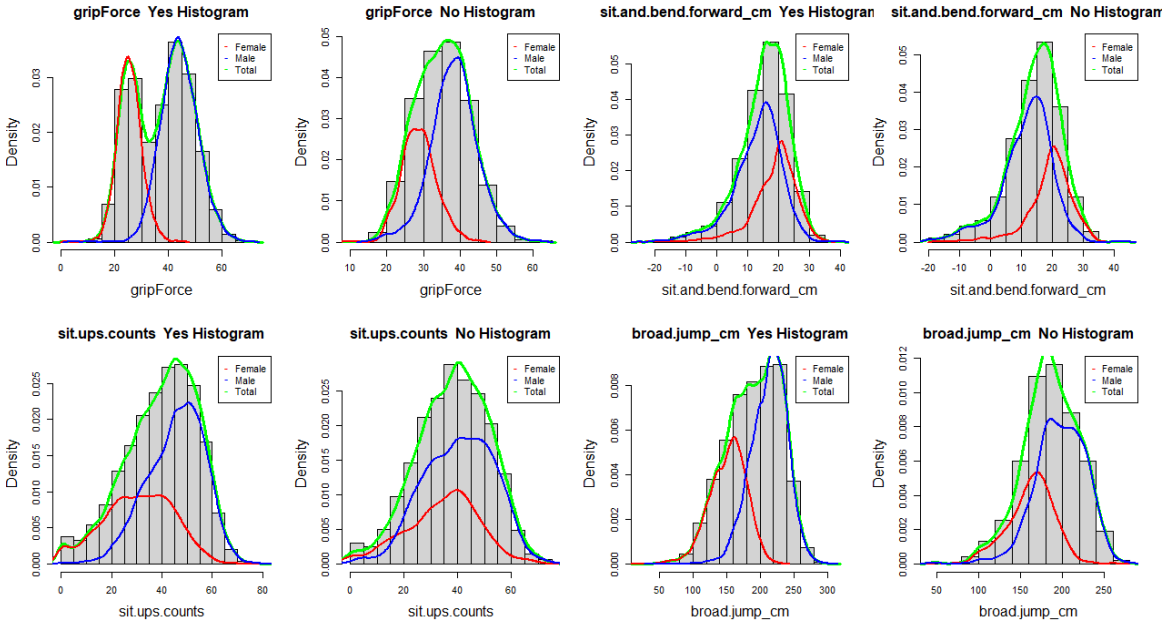
1) 부록



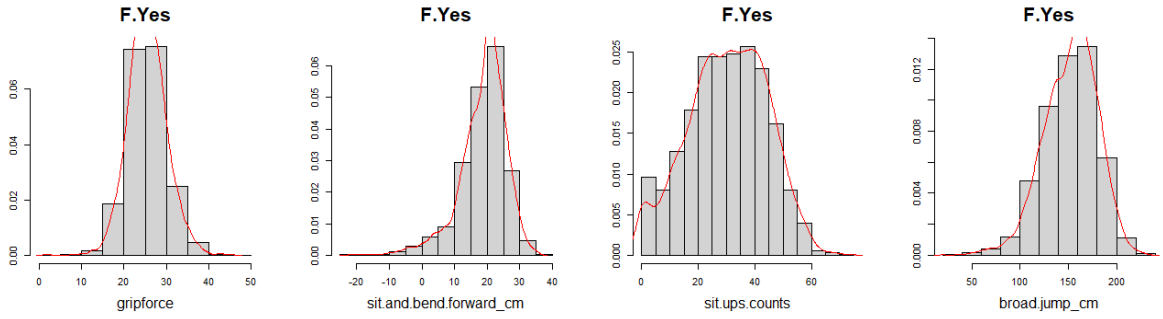
<신체 변수의 정규성 검정, Q-Q Plot, 부록 1>



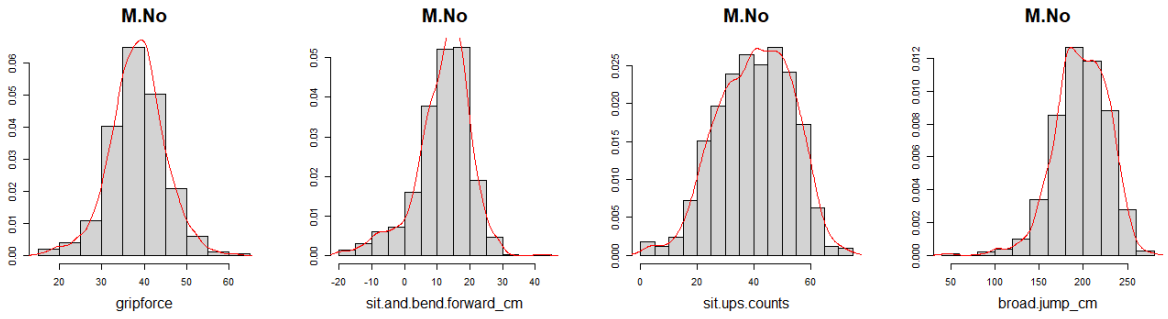
<남녀의 PC1 Box-Plot, 부록 2>



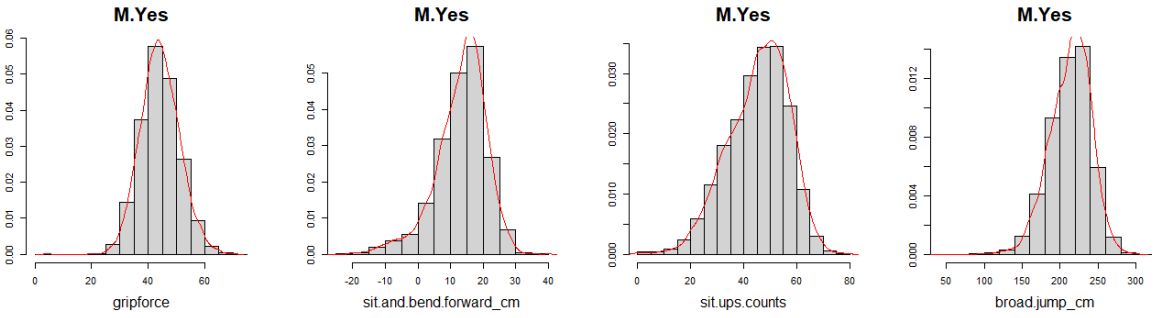
<성향 일치에 따른 히스토그램, 부록 3>



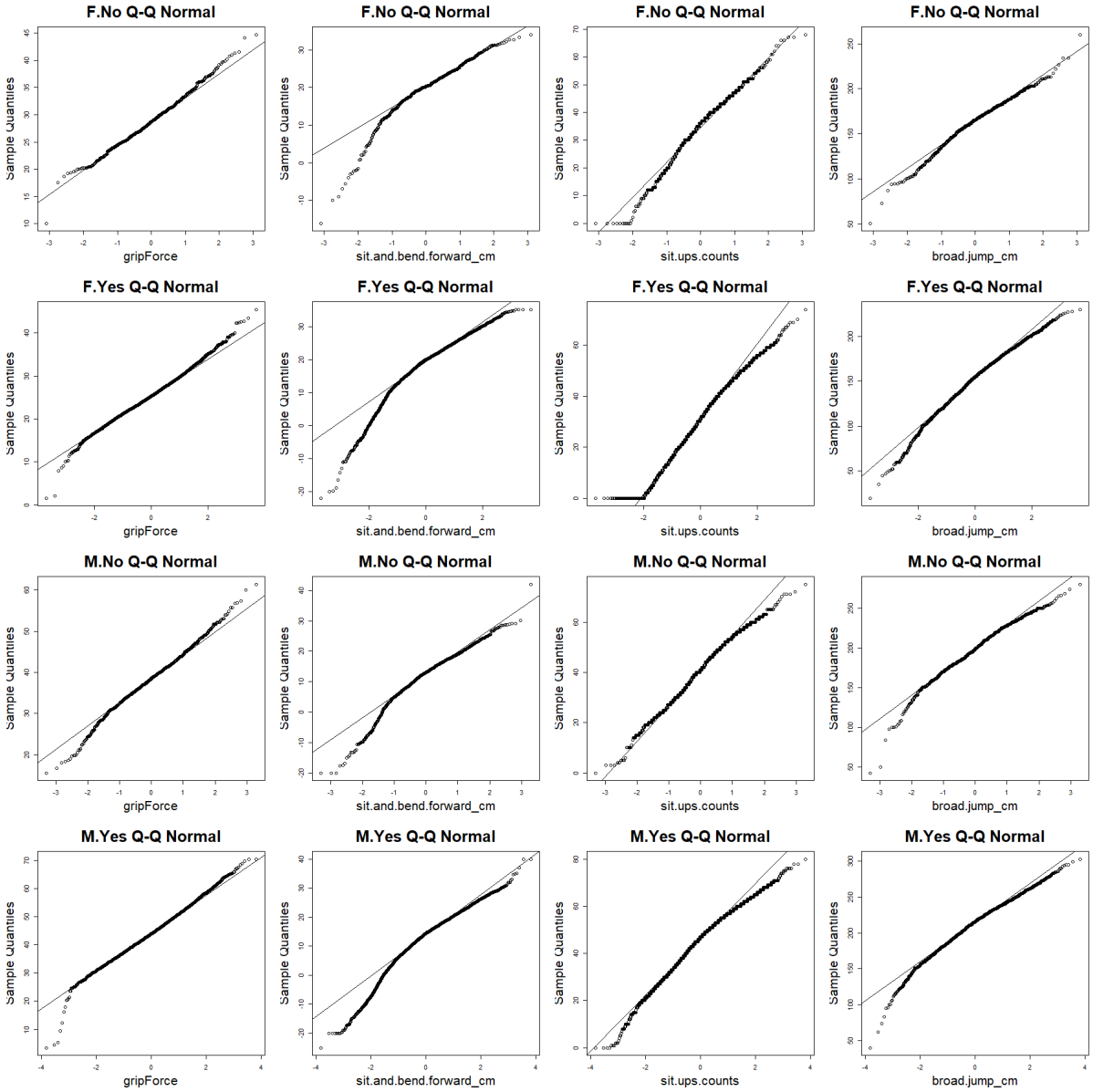
<신체적 성별이 여성인 여성의 운동 변수 분포, 부록 4>



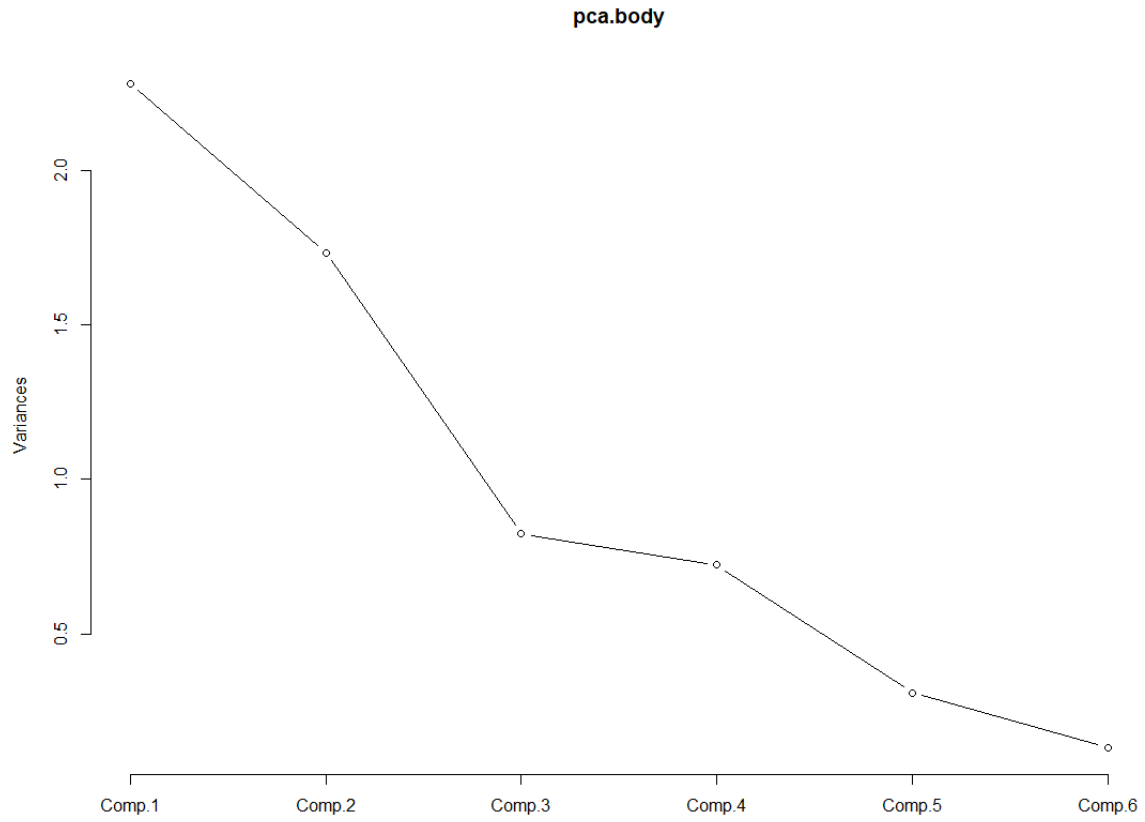
<신체적 성별이 여성인 남성의 운동 변수 분포, 부록 5>



<신체적 성별이 남성인 남성의 운동 변수 분포, 부록 6>



<4개의 집단 별 운동 수행 변수 Q-Q Plot, 부록 7>



<PCA Scree Plot, 부록 8>

Coefficients	Value	Std.Error	t-value	p-value
gikjF.No.1	28.87339	0.8685765	33.2422	0
gikjF.No.2	19.41319	0.8685765	22.3506	0
gikjF.No.3	34.21969	0.8685765.0	39.3974.9	0
gikjF.No.4	162.54724	0.8685765	187.1421	0
gikjF.Yes.1	25.48112	0.2948288	86.4268	0
gikjF.Yes.2	18.78281	0.2948288	63.7075	0
gikjF.Yes.3	30.53649	0.2948288	103.5736	0
gikjF.Yes.4	152.51009	0.2948288	517.2836	0
gikjM.No.1	38.39225	0.6129709	62.6331	0
gikjM.No.2	11.84089	0.6129709	19.3172	0
gikjM.No.3	40.53725	0.6129709	66.1324	0
gikjM.No.4	197.84127	0.6129709	322.7580	0
gikjM.Yes.1	44.15247	0.2270232	194.4844	0
gikjM.Yes.2	13.21695	0.2270232	58.2185	0
gikjM.Yes.3	45.54250	0.2270232	200.6072	0
gikjM.Yes.4	213.48790	0.2270232	940.3792	0

<최종 모델 추정량, 부록 9>

	gripForce	sit.and.bend.forward	sit.ups.counts	broad.jump
gripForce	22.025181	7.478549	24.01158	55.58584
sit.and.bend.forward	7.478549	50.612232	47.40818	88.45411
sit.ups.counts	24.011578	47.408182	184.94446	254.24128
broad.jump	55.585836	88.454109	254.24128	762.19697

<신체적 성별이 남성인 여성의 운동 변수 분포, 부록 10>

	gripForce	sit.and.bend.forward	sit.ups.counts	broad.jump
gripForce	20.617354	9.915044	25.84738	55.74165
sit.and.bend.forward	9.915044	51.047939	40.00303	76.42456
sit.ups.counts	25.847382	40.003033	191.78479	274.94692
broad.jump	55.741647	76.424562	274.94692	764.37110

<신체적 성별이 여성인 여성의 운동 변수 분포, 부록 11>

	gripForce	sit.and.bend.forward	sit.ups.counts	broad.jump
gripForce	41.39383	12.51143	34.65981	97.88991
sit.and.bend.forward	12.51143	68.83297	46.35013	98.40913
sit.ups.counts	34.65981	46.35013	166.83374	271.02304
broad.jump	97.88991	98.40913	271.02304	898.89471

<신체적 성별이 여성인 남성의 운동 변수 분포, 부록 12>

	gripForce	sit.and.bend.forward	sit.ups.counts	broad.jump
gripForce	48.30149	15.17152	27.72683	76.25209
sit.and.bend.forward	15.17152	62.07235	37.63402	86.36512
sit.ups.counts	27.72683	37.63402	130.41797	190.99446
broad.jump	76.25209	86.36512	190.99446	738.46132

<신체적 성별이 남성인 남성의 운동 변수 분포, 부록 13>

2) 참고문헌

1) ABC NEWS. (2021) Women's blood vessels age faster than men's: Study. [online] Available at: <https://abcnews.go.com/Health/womens-blood-vessels-age-faster-mens-study/story?id=68317205&cid=clicksource_4380645_2_heads_hero_live_headlines_heda>.

3) R Code

```
#install.packages(tidyverse)
library(tidyverse)
#gls
library(nlme)
#install.packages("multcomp")
library(multcomp)



##0.데이터 불러오기
body.df <- read.csv('C:/Users/lsj70/OneDrive - dongguk.edu/바탕 화면/학교 수업/3학년 1학기/다변량해석/프로젝트/body/bodyPerformance.csv')

#gender : 성별

#age : 나이
#weight_kg : 몸무게
#body.fat_. : 체지방
#diastolic : 이완기 혈압
#systolic : 수축기 혈압

#gripForc : 악력
#sit.and.bend.forward_cm : 앉아서 몸 앞으로 숙이기
#sit.ups.counts : 윗몸 일으키기
#broad.jump_cm : 제자리 멀리뛰기

#class : 등급



##1.EDA

#결측치 확인
str(body.df)
colSums(is.na(body.df))
#결측치 없음
```

```
#이상치 확인
summary(body.df)

#체지방률,수축기, 이완기 혈압, 몸 앞으로 숙이기, 악력, 제자리 멀리뛰기 이상치 존재

head(sort(body.df$body.fat_.,decreasing = T),10)
body.df %>%
  filter(body.fat_. > 50) %>%
  arrange(-body.fat_.)
#체지방 78이 말이 안되며 다른 수행 능력에 비해 이상함.

head(sort(body.df$diastolic),10)
head(sort(body.df$systolic),10)
#혈압 이상함

head(sort(body.df$sit.and.bend.forward_cm,decreasing = T),10)
#185cm 213cm는 말이 안되니까 제외


head(sort(body.df$gripForce),10)
body.df %>%
  filter(gripForce == 0)
#악력이 0인 것이 말이 안됨. 성인 여성 기준 나이 상관없이 최저 악력이 5.8이었음 -> KOSIS

head(sort(body.df$broad.jump_cm),10)
body.df %>%
  filter(broad.jump_cm == 0)
#제자리 멀리 뛰기가 0인 것도 이상

#이상치 제거
body.df2 <- body.df %>%
  filter(diastolic != 0, diastolic != 6, diastolic != 8,body.fat_. != 78.4,
         systolic != 0, systolic != 14,
         sit.and.bend.forward_cm != 185 , sit.and.bend.forward_cm != 213,
```

```
            gripForce != 0, broad.jump_cm!= 0)
summary(body.df2)

nrow(body.df)
nrow(body.df2)

#변수들 분포 그래프 시각화
#신체변수
par(mfrow=c(2,3), cex.lab = 1.5,cex.main = 2)
hist(body.df2[,1],freq=F, xlab = colnames(body.df2)[1],main = paste(colnames(body.df2)[1],
" Histogram"))
lines(density(body.df2[,1]),col='red')
hist(body.df2[,3],freq=F, xlab = colnames(body.df2)[3],main = paste(colnames(body.df2)[3],
" Histogram"))
lines(density(body.df2[,3]),col='red')
hist(body.df2[,4],freq=F, xlab = colnames(body.df2)[4],main = paste(colnames(body.df2)[4],
" Histogram"))
lines(density(body.df2[,4]),col='red')
hist(body.df2[,5],freq=F, xlab = colnames(body.df2)[5],main = paste(colnames(body.df2)[5],
" Histogram"))
lines(density(body.df2[,5]),col='red')
hist(body.df2[,6],freq=F, xlab = colnames(body.df2)[6],main = paste(colnames(body.df2)[6],
" Histogram"))
lines(density(body.df2[,6]),col='red')
hist(body.df2[,7],freq=F, xlab = colnames(body.df2)[7],main = paste(colnames(body.df2)[7],
" Histogram"))
lines(density(body.df2[,7]),col='red')


#PCA에서 정규성을 만족해야하는 이유는, 주성분들이 원 자료를 설명하는 설명력 때문
#우리는 설명력에 대한 내용이 필요가 없기 때문에 정규성 만족이 필수는 아니라고 판단.
par(mfrow=c(2,3), cex.lab = 1.5,cex.main = 2)
qqnorm(body.df2[,1], xlab = colnames(body.df2)[1])
qqline(body.df2[,1],distribution = qnorm, xlab = colnames(body.df2)[1])
qqnorm(body.df2[,3], xlab = colnames(body.df2)[3])
```

```
qqline(body.df2[,3],distribution = qnorm, xlab = colnames(body.df2)[3])
qqnorm(body.df2[,4], xlab = colnames(body.df2)[4])
qqline(body.df2[,4],distribution = qnorm, xlab = colnames(body.df2)[4])
qqnorm(body.df2[,5], xlab = colnames(body.df2)[5])
qqline(body.df2[,5],distribution = qnorm, xlab = colnames(body.df2)[5])
qqnorm(body.df2[,6], xlab = colnames(body.df2)[6])
qqline(body.df2[,6],distribution = qnorm, xlab = colnames(body.df2)[6])
qqnorm(body.df2[,7], xlab = colnames(body.df2)[7])
qqline(body.df2[,7],distribution = qnorm, xlab = colnames(body.df2)[7])


#운동변수들이 남녀에 대하여 어떠한 분포를 가지는지
N_f = sum(body.df2$gender == 'F')
N_m = sum(body.df2$gender == 'M')
N = nrow(body.df2)
par(mfrow=c(2,2))
hist(body.df2[,8],freq=F, xlab = colnames(body.df2)[8], main =
paste(colnames(body.df2)[8],' Histogram'))
lines(density(body.df2[,8]),col='green',lwd = 3)
lines(density(body.df2[body.df2$gender == 'F',8])$x,
      density(body.df2[body.df2$gender == 'F',8])$y * N_f/N ,col='red',lwd = 2)
lines(density(body.df2[body.df2$gender == 'M',8])$x,
      density(body.df2[body.df2$gender == 'M',8])$y * N_m/N ,col='blue',lwd = 2)

hist(body.df2[,9],freq=F, xlab = colnames(body.df2)[9], main =
paste(colnames(body.df2)[9],' Histogram'))
lines(density(body.df2[,9]),col='green',lwd = 3)
lines(density(body.df2[body.df2$gender == 'F',9])$x,
      density(body.df2[body.df2$gender == 'F',9])$y * N_f/N ,col='red',lwd = 2)
lines(density(body.df2[body.df2$gender == 'M',9])$x,
      density(body.df2[body.df2$gender == 'M',9])$y * N_m/N ,col='blue',lwd = 2)

hist(body.df2[,10],freq=F, xlab = colnames(body.df2)[10], main =
paste(colnames(body.df2)[10],' Histogram'))
lines(density(body.df2[,10]),col='green',lwd = 3)
```

```
lines(density(body.df2[body.df2$gender == 'F',10])$x,
      density(body.df2[body.df2$gender == 'F',10])$y * N_f/N ,col='red',lwd = 2)
lines(density(body.df2[body.df2$gender == 'M',10])$x,
      density(body.df2[body.df2$gender == 'M',10])$y * N_m/N ,col='blue',lwd = 2)

hist(body.df2[,11],freq=F, xlab = colnames(body.df2)[11], main =
paste(colnames(body.df2)[11],' Histogram'))
lines(density(body.df2[,11]),col='green',lwd = 3)
lines(density(body.df2[body.df2$gender == 'F',11])$x,
      density(body.df2[body.df2$gender == 'F',11])$y * N_f/N ,col='red',lwd = 2)
lines(density(body.df2[body.df2$gender == 'M',11])$x,
      density(body.df2[body.df2$gender == 'M',11])$y * N_m/N ,col='blue',lwd = 2)

#쌍봉 형태를 보이는 것들이 존재 -> 성별에 따라서 다르게 보이는 것으로 생각

#성별에 따른 운동 변수 시각화 - boxplot
par(mfrow = c(2,2))
boxplot(gripForce ~ gender, data = body.df2)
boxplot(sit.and.bend.forward_cm ~ gender, data = body.df2)
boxplot(sit.ups.counts ~ gender, data = body.df2)
boxplot(broad.jump_cm ~ gender, data = body.df2)
#outlier들을 아우를 수 있는 새로운 그룹 척도가 필요



#우선적으로 신체 변수들을 이용하기 위한 PCA 진행
##3.PCA
pca.body <- princomp(body.df2[,c(1,3,4,5,6,7)], cor = T)
pca.body$sdev^2 / sum(pca.body$sdev^2)
cumsum(pca.body$sdev^2 / sum(pca.body$sdev^2))
par(mfrow = c(1,1))
screeplot(pca.body, type = "lines")
#주성분 3개!

pca.body$loadings[,1:3]
```

```
#PC1 : 신체적 여성 성향
par(mfrow = c(1,3))
boxplot(diastolic ~ gender, data = body.df)
boxplot(systolic ~ gender, data = body.df)
boxplot(age ~ gender, data = body.df)
#주로 여성의 체지방이 더 높고, 키 몸무게가 낮다. 혈압도 여성이 더 낮고,
#이 데이터에서는 여성의 나이가 더 많았다.

#PC2 : 신체나이, 노화 정도
#나이가 늘어나면 체지방이 증가하고 혈압도 올라가기 때문에

#PC3 : 반(反)비만도

body.df2 %>%
  ggplot(aes(body.fat_. , diastolic)) +
  geom_point()
#생각보다 체지방률은 혈압과 연관이 없다.

par(mfrow = c(1,1))
PC1 <- pca.body$scores[,1];PC1
gender <- body.df2$gender;gender
boxplot(PC1 ~ gender, main = "")
#PC1의 boxplot 결과도 여성이 주로 더 높게 나옴!
#하지만, boxplot의 +-1.5*IQR 부분을 고려하였을 때 겹치는 부분이 생긴다.
#PC1: 여성 성향과 성별이 반드시 일치하는 것은 아니다.

pc1 <- pca.body$scores[,1]
body.df3 <- cbind(body.df2[,c(8:11,2)], pc1)
head(body.df3)

hist(body.df3$pc1,breaks = 50,freq = F)
lines(density(body.df3[body.df3$gender == 'M',]$pc1),col='red')
lines(density(body.df3[body.df3$gender == 'F',]$pc1),col='blue')
#PC1들이 남녀별로 정규분포를 따른다.
```

```
#남성 pc1 평균과, 여성 pc1 평균의 중간 값을 기준으로, 남성성(1), 여성성(0)으로 구분
cut.value = (mean(body.df3[body.df3$gender == 'M',]$pc1) +
mean(body.df3[body.df3$gender == 'F',]$pc1))/2
gender.power = ifelse(body.df3$pc1 >= cut.value, 0, 1)
data.lst <- cbind(body.df3[,1:5],gender.power)
#실제 gender와 gender.power가 일치하는지 여부를 gender.test로 생성
gender.test = ifelse(data.lst$gender == 'M',
                     ifelse(data.lst$gender.power == '1',"Yes","No"),
                     ifelse(data.lst$gender.power == '0',"Yes","No"))
data.lst <- cbind(data.lst,gender.test)
data.lst$gender <- as.factor(data.lst$gender)
data.lst$gender.power <- as.factor(data.lst$gender.power)
data.lst$gender.test <- as.factor(data.lst$gender.test)


##4.MANOVA
#PC1 변수가 과연 group을 나누는 데에 있어서 도움이 되는지 평가

#정규성 테스트
par(mfrow=c(1,4))
for(i in 1:4){
hist(data.lst[data.lst$gender == 'F' & data.lst$gender.test == 'No',i],freq=F, xlab =
colnames(data.lst)[i],main = 'F.No', ylab = '')
lines(density(data.lst[data.lst$gender == 'F' & data.lst$gender.test == 'No',i]),col = 'red')
}
for(i in 1:4){
  hist(data.lst[data.lst$gender == 'F' & data.lst$gender.test == 'Yes',i],freq=F, xlab =
colnames(data.lst)[i],main = 'F.Yes', ylab = '')
  lines(density(data.lst[data.lst$gender == 'F' & data.lst$gender.test == 'Yes',i]),col = 'red')
}
for(i in 1:4){
  hist(data.lst[data.lst$gender == 'M' & data.lst$gender.test == 'No',i],freq=F, xlab =
colnames(data.lst)[i],main = 'M.No', ylab = '')
  lines(density(data.lst[data.lst$gender == 'M' & data.lst$gender.test == 'No',i]),col = 'red')
```

```
}
for(i in 1:4){
  hist(data.lst[data.lst$gender == 'M' & data.lst$gender.test == 'Yes',i],freq=F, xlab =
colnames(data.lst)[i],main = 'M.Yes', ylab = '')
  lines(density(data.lst[data.lst$gender == 'M' & data.lst$gender.test == 'Yes',i]),col = 'red')
}
##Q-Q Plot
par(mfrow=c(4,4), cex.lab = 2,cex.main = 2.5)
for(i in 1:4){
  qqnorm(data.lst[data.lst$gender == 'F' & data.lst$gender.test == 'No',i], xlab =
colnames(data.lst)[i], main = "F.No Q-Q Normal")
  qqline(data.lst[data.lst$gender == 'F' & data.lst$gender.test == 'No',i],distribution = qnorm,
xlab = colnames(data.lst)[i])
}
for(i in 1:4){
  qqnorm(data.lst[data.lst$gender == 'F' & data.lst$gender.test == 'Yes',i], xlab =
colnames(data.lst)[i], main = "F.Yes Q-Q Normal")
  qqline(data.lst[data.lst$gender == 'F' & data.lst$gender.test == 'Yes',i],distribution = qnorm,
xlab = colnames(data.lst)[i])
}
for(i in 1:4){
  qqnorm(data.lst[data.lst$gender == 'M' & data.lst$gender.test == 'No',i], xlab =
colnames(data.lst)[i], main = "M.No Q-Q Normal")
  qqline(data.lst[data.lst$gender == 'M' & data.lst$gender.test == 'No',i],distribution = qnorm,
xlab = colnames(data.lst)[i])
}
for(i in 1:4){
  qqnorm(data.lst[data.lst$gender == 'M' & data.lst$gender.test == 'Yes',i], xlab =
colnames(data.lst)[i], main = "M.Yes Q-Q Normal")
  qqline(data.lst[data.lst$gender == 'M' & data.lst$gender.test == 'Yes',i],distribution = qnorm,
xlab = colnames(data.lst)[i])
}

cov(data.lst[data.lst$gender == 'F' & data.lst$gender.test == 'No',c(1:4)])
cov(data.lst[data.lst$gender == 'F' & data.lst$gender.test == 'Yes',c(1:4)])
```

```
cov(data.lst[data.lst$gender == 'M' & data.lst$gender.test == 'No',c(1:4)])
cov(data.lst[data.lst$gender == 'M' & data.lst$gender.test == 'Yes',c(1:4)])
```

```
#개수 확인
table(data.lst[,c('gender.test','gender')])
#비율적으로 생각했을 때 말이 된다고 생각.
```

```
#MANOVA를 위해서 1자로 데이터 나열.
data.lst <- arrange(data.lst,gender,gender.test)
head(data.lst)
nn = nrow(data.lst)
pp = 4
y <- as.numeric(t(as.matrix(data.lst[,1:4])))
id = rep(1:nn, each=pp)
gi=c(rep("F",4917*pp),rep("M",8456*pp))
gk=c(rep("No",508*pp),rep("Yes",4409*pp),rep("No",1020*pp),rep("Yes",7436*pp))
gj=rep(c("1","2","3","4"),nn)
gikj = paste(gi,gk,gj,sep=".")
gij = paste(gi,gj,sep=".")
gkj = paste(gk,gj,sep=".")
gik = paste(gi,gk,sep=".")

data.lst2 = data.frame(y,id,gi,gk,gj,gikj,gij,gkj,gik)
head(data.lst2)

##가설검정1
#H0 : 성별에 따른 운동 수행 능력의 차이가 없다.
#Full Model : y{ij} = g{ij}+ e{ij}
#Reduced Model : y{j} = mu{j} + e{j}

gF.1 = gls(y ~ gij, cor = corSymm(form = ~1|id),method = "ML", data = data.lst2)
```

```
gR.1 = gls(y ~ gj, cor = corSymm(form = ~1|id),method = "ML", data = data.lst2)
chi.test1 = anova(gR.1,gF.1);chi.test1

##가설검정2
#H0 : PC1에 따라 성질을 구분하는 것이 성별만으로 운동 수행 능력을 구분하는 것과 차이가 없다.
#Full Model : y{ijk} = gi{ij} + gk{kj} + gi{ij}:gk{kj} + e{ikj}
#Reduced Model : y{ij} = g{ij}+ e{ij}

gF.2 = gls(y ~ gikj, cor = corSymm(form = ~1|id),method = "ML", data = data.lst2)
gR.2 = gls(y ~ gij, cor = corSymm(form = ~1|id),method = "ML", data = data.lst2)
chi.test2 = anova(gR.2,gF.2); chi.test2

##가설검정3
#H0 : 본인의 성별과 PC1의 결과 일치 여부(gk)는, 성별(gi)이 운동 수행 능력(y)에 주는 영향과 상관이 없다.
#Full Model : y{ijk} = gi{ij} + gk{kj} + gi{ij}:gk{kj} + e{ikj}
#Reduced Model : y{ijk} = gi{ij} + gk{kj} + e{ikj}
gF.3 = gls(y ~ gikj, cor = corSymm(form = ~1|id),method = "ML", data = data.lst2)
gR.3 = gls(y ~ (gi+gk)*gj , cor = corSymm(form = ~1|id),method = "ML", data = data.lst2)
chi.test3 = anova(gR.3,gF.3); chi.test3



#셋 다 p-value : <.0001  ==>  H0 기각
#즉, 운동 수행 변수들에 대하여 성별 집단을 구분 하는데 있어서,
#PC1에 의해 구분한 성향과 실제 성별의 일치 여부도 고려해주어야 한다.
#최종 모델
gF = gls(y ~ gikj-1, cor = corSymm(form = ~1|id),method = "ML", data = data.lst2)
summary(gF)



#gender 와 test를 결합한 변수 추가
gender.lst = as.factor(paste(data.lst$gender, data.lst$gender.test, sep="."))
data.0 = cbind(data.lst,gender.lst)
str(data.0)
```

```
##분류에 의한 시각화 결과를 보여주며 마무리
data.0 %>%
  ggplot(aes(x=broad.jump_cm, y=sit.and.bend.forward_cm, col=gender, shape = gender.test))
+
  geom_point()

str(data.0)
N_fy = sum(data.0$gender.lst == "F.Yes")
N_fn = sum(data.0$gender.lst == "F.No")
N_my = sum(data.0$gender.lst == "M.Yes")
N_mn = sum(data.0$gender.lst == "M.No")
N = nrow(data.0)

par(mfrow=c(2,2))
for(i in 1:4){
  hist(data.0[,i],freq=F, xlab = colnames(data.0)[i], main = paste(colnames(data.0)[i],'
Histogram'))
  lines(density(data.0[,i]),col='green',lwd = 3)
  lines(density(data.0[data.0$gender.lst == "F.Yes",i])$x,
        density(data.0[data.0$gender.lst == "F.Yes",i])$y * N_fy/N ,col='red',lwd = 2)
  lines(density(data.0[data.0$gender.lst == "F.No",i])$x,
        density(data.0[data.0$gender.lst == "F.No",i])$y * N_fn/N ,col='magenta',lwd = 2)
  lines(density(data.0[data.0$gender.lst == "M.Yes",i])$x,
        density(data.0[data.0$gender.lst == "M.Yes",i])$y * N_my/N ,col='blue',lwd = 2)
  lines(density(data.0[data.0$gender.lst == "M.No",i])$x,
        density(data.0[data.0$gender.lst == "M.No",i])$y * N_mn/N ,col='cyan',lwd = 2)
}


par(mfrow=c(2,4),cex.lab = 1.5, cex.main = 1.5)
for(i in 1:4){
  hist(data.0[data.0$gender.test=='Yes',i],freq=F, xlab = colnames(data.0)[i], main =
paste(colnames(data.0)[i],' Yes Histogram'))
```

```
  legend("topright", c("Female", "Male", "Total"),
           col = c( "red", "blue", "green"), pch = c("-","-","-"))
  lines(density(data.0[data.0$gender.test=='Yes',i]),col='green',lwd = 3)
  lines(density(data.0[data.0$gender.lst == "F.Yes",i])$x,
        density(data.0[data.0$gender.lst == "F.Yes",i])$y * N_fy/(N_my+N_fy) ,col='red',lwd =
2)
  lines(density(data.0[data.0$gender.lst == "M.Yes",i])$x,
        density(data.0[data.0$gender.lst == "M.Yes",i])$y * N_my/(N_my+N_fy) ,col='blue',lwd
= 2)
  hist(data.0[data.0$gender.test=='No',i],freq=F, xlab = colnames(data.0)[i], main =
paste(colnames(data.0)[i],' No Histogram'))
  legend("topright", c("Female", "Male", "Total"),
           col = c( "red", "blue", "green"), pch = c("-","-","-"))
  lines(density(data.0[data.0$gender.test=='No',i]),col='green',lwd = 3)
  lines(density(data.0[data.0$gender.lst == "F.No",i])$x,
        density(data.0[data.0$gender.lst == "F.No",i])$y * N_fn/(N_mn+N_fn) ,col='red',lwd =
2)
  lines(density(data.0[data.0$gender.lst == "M.No",i])$x,
        density(data.0[data.0$gender.lst == "M.No",i])$y * N_mn/(N_mn+N_fn) ,col='blue',lwd
= 2)
}



par(mfrow=c(2,4))
for(i in 1:4){
  hist(data.0[data.0$gender=='F',i],freq=F, xlab = colnames(data.0)[i], main =
paste(colnames(data.0)[i],' Female Histogram'))
  legend("topright", c("Yes", "No", "Total"),
           col = c( "cyan", "magenta", "green"), pch = c("-","-","-"))
  lines(density(data.0[data.0$gender=='F',i]),col='green',lwd = 3)
  lines(density(data.0[data.0$gender.lst == "F.Yes",i])$x,
        density(data.0[data.0$gender.lst == "F.Yes",i])$y * N_fy/(N_fn+N_fy) ,fill='cyan',lwd =
2)
  lines(density(data.0[data.0$gender.lst == "F.No",i])$x,
```

```
          density(data.0[data.0$gender.lst == "F.No",i])$y * N_fn/(N_fn+N_fy) ,fill='magenta',lwd
= 2)

  hist(data.0[data.0$gender=='M',i],freq=F, xlab = colnames(data.0)[i], main =
paste(colnames(data.0)[i],' No Histogram'))
  legend("topright", c("Female", "Male", "Total"),
           col = c( "red", "blue", "green"), pch = c("-","-","-"))
  lines(density(data.0[data.0$gender=='M',i]),col='green',lwd = 3)
  lines(density(data.0[data.0$gender.lst == "M.Yes",i])$x,
          density(data.0[data.0$gender.lst == "M.Yes",i])$y *
N_my/(N_my+N_mn) ,fill='cyan',lwd = 2)
  lines(density(data.0[data.0$gender.lst == "M.No",i])$x,
          density(data.0[data.0$gender.lst == "M.No",i])$y *
N_mn/(N_my+N_mn) ,fill='magenta',lwd = 2)
}


par(mfrow=c(2,2))
for(i in 1:4){
  boxplot(data.0[,i] ~ data.0$gender.lst, xlab = "" ,ylab = colnames(data.0)[i], main =
paste(colnames(data.0)[i],' Boxplot'))
}
?boxplot
plot(density(data.0[data.0$gender=='F',1]),col='blue',lwd = 2)
lines(density(data.0[data.0$gender.lst=='F.No',1]),col='red',lwd = 2)


par(mfrow=c(2,4))
for(i in 1:4){
  plot(density(data.0[data.0$gender=='F',i]),col='blue',lwd = 2, xlab = colnames(data.0)[i],
main = paste(colnames(data.0)[i],' Female Density'))
  legend("topright", c("Male", "Male.No"),
           col = c( "blue", "red"), pch = c("-","-"))
  lines(density(data.0[data.0$gender.lst=='F.No',i]),col='red',lwd = 2)
```

```
  plot(density(data.0[data.0$gender=='M',i]),col='blue',lwd = 2, xlab = colnames(data.0)[i],
main = paste(colnames(data.0)[i],' Male Density'))
  legend("topright", c("Female", "Female.No"),
         col = c( "blue", "red"), pch = c("-","-"))
  lines(density(data.0[data.0$gender.lst=='M.No',i]),col='red',lwd = 2)
}
```