**1번.** (lonely.txt) Longley(1967) 데이터는 반응변수 S와 6개의 설명변수 X1, ⋯ , 𝑋6로 구성되어 있다. 최초 모형은 다음과 같다.

(model A) Y = β0 + 𝛽1𝑋1 + ⋯ + 𝛽6𝑋6 + 𝜖

원 변수들을 표준화변수로 변환하면 다음과 같이 쓸 수 있다.

(model B) 𝑌̃ = 𝜃1𝑋̃1 + ⋯ + 𝜃6𝑋̃6 + 𝜖′

1. 최소제곱법을 이용하여 model B 모형을 적합하여라. 어떠한 결론을 얻을 수 있는가?

결정계수와 수정된 결정계수 모두 0.99 이상이고 MAE와 RMSE이 0.05에 가까운 값으로 모델의 설명력이 매우 높음을 알 수 있다. ANOVA 분석에서는 p값이 0에 가까워 적어도 하나의 설명변수의 계수가 0이 아닌 대립가설을 채택하게 된다. 그러나 Y절편과 변수 X1,X2,X5의 p-value가 유의수준인 0.05보다 높기 때문에 이 변수들은 통계적으로 유의하지 않다는 결론을 내릴 수 있다.

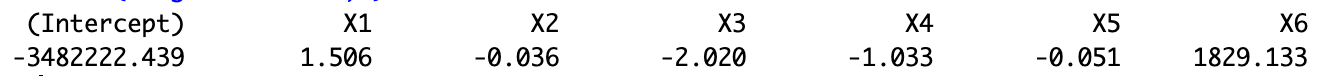
A paper with numbers and text

Description automatically generatedA close-up of a paper

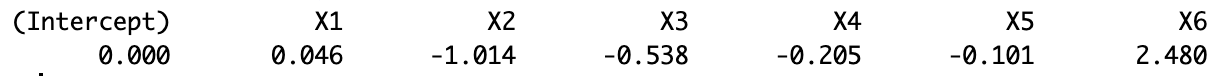
Description automatically generated

1. Model B를 이용하여 얻은 결과로부터 모형 A의 회귀계수에 대한 최소제곱추정치를 계산 하여라.

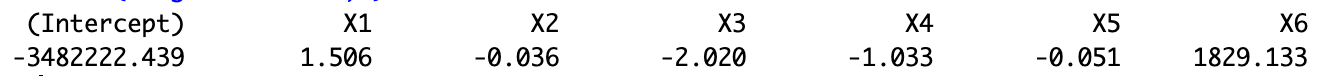
Model A 회귀계수:

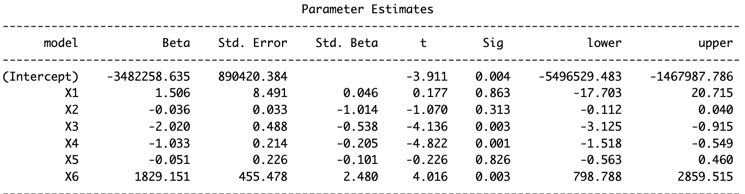


Model B 회귀계수:



1. 최소제곱법을 이용하여 model A를 적합시키고 model B에서 얻은 결과와 일치하는지 확 인하여라.





Y절편과 6개의 변수모두 일치한다.

1. 6개의 예측변수에 대한 상관행렬을 계산하고 산점도 행렬을 그려보아라. 다중공선성의 증 거가 있는지 확인하여라.

A number with numbers on it

Description automatically generated with medium confidence

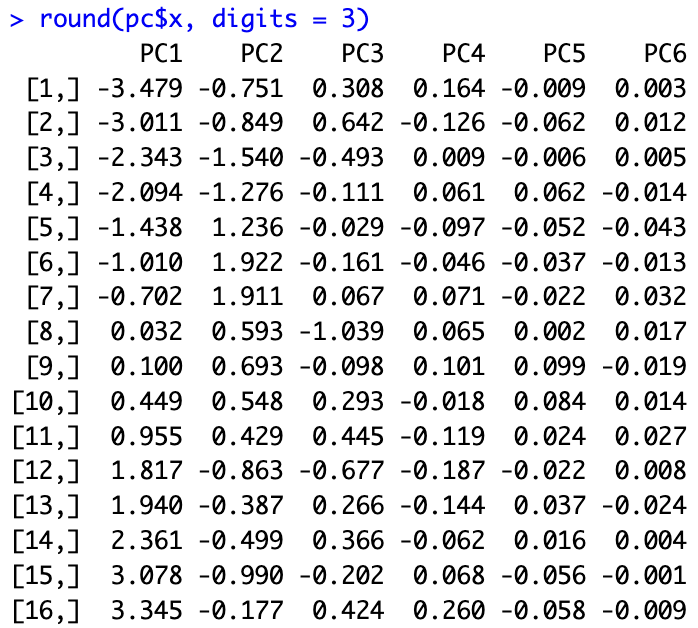
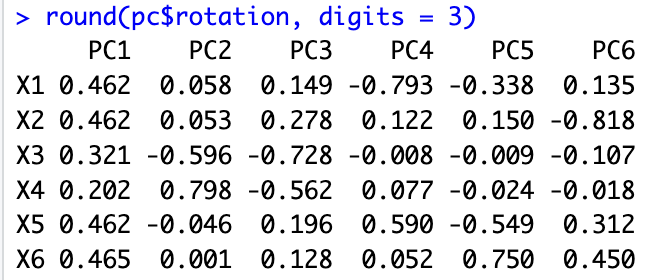
A screenshot of a graph

Description automatically generated

X1, X2, X5, X6 변수의 상관관계가 거의 1에 가까워 다중공선성의 문제가 있다고 판단된다.

1. 6개의 예측변수에 대한 주성분과 주성분의 분산, 상태지수를 계산하여라. 데이터에 몇 개 의 다중공선성 집합이 존재하는가? 각 다중공선성 집합에는 어떠한 변수가 포함되어 있는가?

주성분값과 고유벡터:

주성분 분산 (고유값):

A number and numbers on a white background

Description automatically generated

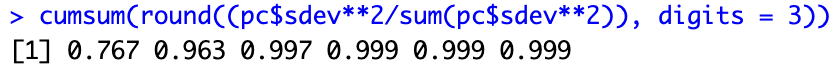
상태지수:

A close-up of a white background

Description automatically generated

고유값을 살펴보면 마지막 세개의 주성분 분산이 0에 가깝고 상태지수 값이 110.5로 매우 높으므로 다중공선성 문제가 존재한다. 고유벡터(eigenvector) 주성분PC1에서 X1,X2,X5,X6의 계수 값이 0.46으로 유사한 크기를 보인다. 이 때문에 다중공선성 집합에 X1,X2,X5,X6 4개의 변수가 포함된다.

1. 적절한 주성분의 수를 선택하여 model A와 model B의 주성분 추정치를 계산하여라.



누적분산비율이 0.9이상이 되는 지점이 2이기 때문에 주성분의 수를 2로 설정한다.

Model A:

A table of numbers with numbers

Description automatically generated

Model B:

A screenshot of a table of numbers

Description automatically generated

1. 능형방법을 이용하여 능형궤적도를 만들어 보아라. Model A와 model B에서의 회귀계수 추정을 위한 적절한 κ값은 무엇인가? 선택한 κ값을 사용하여 model A와 model B의 회귀 계수에 대한 능형추정치를 계산하여라.

Model A:

A graph of different colored lines

Description automatically generated

A close up of black text

Description automatically generated

K가 0.03과0.04 사이에 도달하는 순간 전체분산이 감소하는 폭이 줄어 k값을 0.04로 설정한다.

A close-up of numbers

Description automatically generated

Model B:

A graph of different colored lines

Description automatically generated

A close-up of a white background

Description automatically generated

Model A와 비슷하기 때문에 k값을 0.04로 설정한다.

A close-up of a number

Description automatically generated

1. 위에서 수행한 세가지 방법에서 각각 얻어진 추정치들을 비교하고 어떤 추정치를 추천할 지 설명하여라.

위 데이터셋은 다중공선성 문제에 직면하기 때문에 주성분회귀나 능형회귀가 더 적합하다.

주성분회귀의 결정계수는 0.972, MSE는 397201.2이다.

A table of numbers and text

Description automatically generated with medium confidence

능형회귀의 결정계수는0.996, MSE는 127585이다.

A close-up of a computer error

Description automatically generated

능형회귀가 주성분회귀보다 결정계수가 높고 MSE가 낮은 걸 확인할 수 있다. 이로서 능형회귀 모델이 주성분회귀 보다 데이터의 변동성을 효과적으로 설명하고 이를 바탕으로 예측 정확도가 높다는 결론을 내릴 수 있다.

**2번.** (Field.Goal.Kicking.txt) 1969년 아메리카 미식축구리그(AFL)와 국제미식축구리그(NFL) 전체에 대한 필드골(field goal kicking) 데이터를 분석한다. π(X)를 거리 X 야드(yard)에서 필드골의 확률이 라고 정의한다.

1. 각 리그에 대하여 다음의 모형을 적합하고 그 결과를 설명하여라.

π(X) = exp(β0 + β1X + β2X 2 ) /(1+ exp(β0 + β1X + β2X 2) )

AFL:

A white background with black numbers and symbols

Description automatically generated

Y절편과 거리 변수의 p값이 0.05보다 낮아 통계적으로 유의하다. 거리가 1야드 증가할때 성공 확률의 오즈값이 17.8% (1-e^(-0.197)) 만큼 감소하기 때문에 거리와 성공확률은 반비례하는 관계를 알 수 있다. 거리제곱 변수는 p값이 0.144로 유의하지는 않다. AIC는 28.443으로 낮을 수록 좋은 모형이다.

NFL:

A white background with black numbers and text

Description automatically generated

Y절편을 제외한 예측변수들은 p값이 0.05보다 유의하지 않다. AIC는 28.89로 AFL 모델과 비슷하다.

1. Z를 리그를 나타내는 지시변수라고 정의하자. 즉,

1, AFL

Z = {

0, NFL

지시변수 Z를 포함한 모형의 확장을 통하여 두 리그를 하나로 합친 모형을 적합하여라. 즉, 다음 모형을 적합하고 그 결과를 설명하여라.

exp (β0 + β1X + β2X2 + β3Z)

π(X, Z) = 1 + exp (β + β X + β X2 + β3Z)

0 1 2

A white background with black numbers and text

Description automatically generated

Y 절편과 거리 변수는 유의하지만 거리제곱과 지시변수은 큰 p값 때문에 통계적으로 유의하지 않다. 거리 또는 야드가 1증가할때마다 성공확률 오즈가 9%정도 감소한다. AIC는 기존 모형보다 큰 59.367로 높기때문에 모형을 단순화할 필요가 있다.

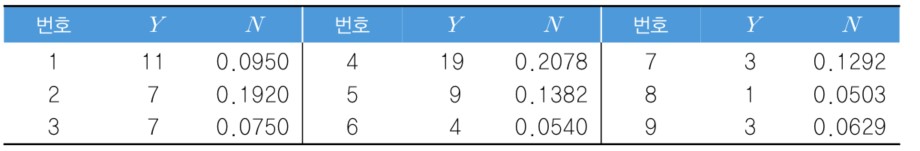
1. 모형에 2차항이 유의한 기여를 하는가? 설명하여라.

2차항 변수의 p값이 0.88로 1에 가까우며 계수는 0에 수렴하기 때문에 유의한 기여를 하지 않는다. 또한 2차항 변수를 포함한 모델의 AIC는 59.367인 반면에 제외한 모델은 57.389로 모델 성능에 기여하지 않는다.

1. 특정 거리에서의 필드골 득점 확률이 각 리그가 동일한가?

미식축구리그는 거리 야드가 1증가할떄마다 성공확률 오즈가 17.8% 감소하는 반면에 국제미식축구리그는 1.3%감소하기 때문에 특정 거리에서의 필드골 득점 확률이 확연히 차이가 난다.

**3번.** 6장에서 다뤘던 항공사 사상사고 데이터는 다음과 같다.



위 데이터는 뉴욕으로부터 미국 내 주요 9개 주로 운항하는 항공편에서 어느 한 해에 일어난 사 상사고 발생건수(Y)와 전체 항공편의 운항률(N)에 관한 데이터이다. 위 데이터를 최소제곱법과 포 아송회귀를 각각 적합하여 그 결과를 비교하여라. 어떤 방법론이 위 데이터를 잘 설명하는시 분석하여라.

선형회귀:

A screenshot of a computer error

Description automatically generated선형회귀에서는 절편이 -0.1402로 p-값이 매우 높아 통계적으로 유의하진 않지만 운항률의 계수는 64.9755로 p-값이 0.05보다 낮아 양의 상관관계를 파악할 수 있다. 하지만 결정계수와 수정된 결정계수 값이 각각 0.487, 0.4139로 다소 낮다. AIC는 55.1414이다.

포아송회귀:

포아송회귀에서는 절편A white paper with black text

Description automatically generated(0.8945)과 회귀계수(8.5)의 p-값이 모두 0.05이하이기 때문에 통계적으로 유의하다. AIC는 52.251로 선형회귀 모델보다 낮다.

포아송회귀가 더 적합한 모델인 첫번째 이유는 데이터의 특성 때문이다. 항공편에서 한 해 일어나는 사상사고는 희귀할 뿐만 아니라 종속변수가 카운트형 변수로 정규분포보다는 포아송 분포를 따른다고 볼 수 있다. 두번째는 선형회귀모델에서 잔차의 등분산성을 만족하지 않는다. 이분산성이 제거되지 않으면 표준오차 값이 커져 신뢰구간이 넓어지고 유의성 검정의 민감성이 감소된다. 세번째는 포아송회귀에서는 선형회귀모델과는 다르게 변수와 y절편이 유의한 동시에 AIC도 낮다.

A graph with a line drawn on it

Description automatically generated