Assignment #3

Due date: 4월 14일

잘 정리된 report를 작성해야 하며, 아래의 파일들을 제출해야 합니다.

- **R markdown 파일 (code, comment, 답안 해설 포함)**
- **생성된 html 파일**

1. Climate Change

ClimateChange.csv는 1983년 5월부터 2008년 12월까지의 지구의 평균적인 대기 질 및 기후와 관련된 월간 데이터를 포함한다. 변수에 대한 상세한 설명은 아래와 같다. 이를 활용하여 세계의 평균 기온을 예측하기 위한 모델을 만들어 보고자 한다. Temp 변수를 target으로, Year 및 Month를 제외한 나머지 8개의 변수를 feature로 사용하자.

- **Year** : 관측 년도
- **Month** : 관측 월
- **Temp** : 세계 평균 기온 (기준값 대비 차이)
- **CFC.11** : 대기 중 CFC-11 프레온가스 농도 (단위: ppbv)
- **CFC.12** : 대기 중 CFC-12 프레온가스 농도 (단위: ppbv)
- **CO2** : 대기 중 이산화탄소 농도 (단위: ppmv)
- **N2O** : 대기 중 아산화질소 농도 (단위: ppmv)
- **CH4** : 대기 중 메탄 농도 (단위: ppmv)
- **Aerosols** : The mean stratospheric aerosol optical depth - 성층권 에어로졸 깊이
- **TSI** : The total solar irradiance - 대기 중 단위 면적당 태양에너지
- **MEI** : Multivariate El Nino Southern Oscillation index - 태평양에서의 기후 효과의 강도에 대한 척도

1. Year 및 Month를 제외한 9개의 변수들 간의 상관 관계를 다양한 그래프를 활용하여 시각화해보고, 이로부터 데이터의 특성을 분석해보자.

2. 2004년 이후의 데이터를 test set으로 2003년까지의 데이터를 training set으로 분할하자. 그리고 training set을 활용하여 linear regression model을 수립하자. 이때 8개의 feature변수를 모두 포함시킨다.

   a) 어떠한 feature들이 Temp에 큰 영향을 미치는가?

   b) N2O와 CFC-11은 지구의 지표면에서 우주로 발산하는 적외선 복사열을 흡수하여 지구 표면의 온도를 상승시키는 역할을 하는 온실가스로 알려져 있다. 모델에서 N2O와 CFC-11 변수의 coefficient는 양수 값을 가지는가? 음수 값을 가지는가? 만약 음수값을 가진다면 N2O와 CFC-11의 양이 증가할수록 평균 기온이 감소한다는 것을 의미하므로 일반적인 지식과 모순된다. 이러한 모순된 결과가 도출되는 원인은 무엇일까?

3. MEI, TSI, Aerosols, N2O 4개의 feature만 사용하여 regression model을 만들어 보자.

   a) N2O 변수의 coefficient를 2번 모델과 비교해 보자.

   b) 두 모델의 R^2 값, Adusted R^2 값, test set error (test set에 대한 RMSE) 를 비교해 보자. 어떤 모델이 더 좋은 모델이라고 할 수 있는가?

4. 8개의 feature를 대상으로 cross validation을 활용한 stepwise variable selection을 수행해보자.

   a) Forward selection과 backward selection의 결과를 비교해보자.

   b) Prediction accuracy와 Model interpretability를 종합적으로 고려하여 best 모델을 하나 결정하자.

5. Prediction accuracy를 높이기 위해, 기존 8개의 feature들 외에 feature들 사이의 모든 interaction effect, 그리고 CO2, CFC.11, CFC.12의 제곱항들을 모두 추가한 모델을 대상으로 cross validation을 활용한 stepwise variable selection을 수행해보자.

   a) Forward selection과 backward selection의 결과를 비교해보자.

   b) Cross validated RMSE가 가장 낮은 best 모델을 결정하자. 어떠한 변수들이 best 모델에 포함되는가?

6. 2, 3, 4, 5번에서 수립된 4개의 모델에 대해서 test set (2004년 이후 데이터)에 대한 prediction accuracy(RMSE)를 비교해 보자. 예상한 대로 결과가 나오는가? 그렇지 않다면 그 원인은 무엇일지 분석해보자.

2. Regression on Simulated Data

먼저 아래와 같이 랜덤으로 데이터를 생성하자.

(i) **rnorm**() 함수를 활용해서 평균이 0, 표준편차가 1인 표준정규분포로부터 크기가 100인 vector X를 생성하고, 평균이 0, 표준편차가 4인 정규분포로부터 크기가 100인 오차 vector ϵ을 생성한다. X와 ϵ을 생성하기 위한 **rnorm**() 함수에 대해서 동일한 random seed 값을 사용하지 않도록 주의하자.

(ii) 크기가 100인 target vector Y 를 다음 식을 사용하여 생성한다.

$$Y = 1 - 2X + 3X^2 - 4X^3 + \epsilon$$

즉, i번째 관측치 Y_i 값은 세 가지 feature X, X^2, X^3에 대한 선형식에 오차 ϵ_i를 더한 것과 같다. 위의 선형 관계식을 모른 채 100개의 관측치만 주어졌을 때 이를 추정하기 위한 linear regression model을 아래의 순서대로 만들어보자. 즉, 실제 regression coefficient $\beta_0 = 1$, $\beta_1 = -2$, $\beta_2 = 3$, $\beta_3 = 4$를 데이터로부터 추정해야 한다.

1. $X, X^2, X^3 \ldots, X^{10}$의 10개 변수를 feature로, Y를 target으로 설정하자. 이때 feature 변수들과 target 변수 사이의 상관관계를 시각화해보자.

2. 10개의 feature를 모두 포함하는 linear regression model을 만들어보자. 통계적으로 유의한 변수가 있는가? regression coefficient $\hat{\beta}_j$ 값을 실제 β_j 값과 비교해보자.

3. X, X^2, X^3의 3개 변수를 feature로, Y 를 target으로 linear regression model을 만들어보자. 모든 feature들이 통계적으로 유의한가? regression coefficient $\hat{\beta}_j$ 값을 실제 β_j 값과 비교해보자.

4. $X, X^2, X^3 \ldots, X^{10}$의 10개 변수를 feature로, Y 를 target으로 Lasso regression model을 만들어 본다. Cross validation을 통해 합리적인 모델을 찾아보자. 이 모델에는 어떤 변수가 포함되었는가? regression coefficient 값을 실제 β 값과 비교해보자. 그리고 결과를 바탕으로 Lasso regression의 효과에 대해서 설명해보자.