규제된 선형 회귀

규제 선형 회귀

- 회귀 모형과 회귀 계수
  - 지금까지의 선형 회귀 모형은 비용 함수 RSS를 최소화 하는 데에 초점을 맞추었다.
  - 이에 따라서 훈련 데이터에는 지나치게 잘 맞고 회귀 계수가 커지게 되어, 검증 데이터에서는 올바르게 예측을 하지 못 할 수 있다. (과대적합)
  - 따라서 비용 함수는 RSS를 최소화 하면서도 회귀 계수 값이 너무 커지지 않도록 균형을 이룰 필요가 있다.
  - 즉, 비용 함수의 목표는 다음과 같이 수정된다.

$$\text{cost} = \min(\text{RSS}(w) + \alpha \times f(w))$$

(이 때, f(w)는 회귀 계수에 대한 함수, 즉 추가 제약 조건이고, α는 RSS(w)와 f(w)의 비중을 조정하는 매개변수이다.)

규제 선형 회귀

- 회귀 모형과 회귀 계수
  - 매개변수 α의 값에 따라서 잔차제곱합 RSS와 회귀 계수에 대한 추가 제약 조건 f(w)의 비중이 달라진다.

$$cost = \min(RSS(w) + \alpha \times f(w))$$

  - α가 0이라면, 비용 함수는 기존과 동일하게 RSS 값으로 적용된다. 이 때의 목표는 결국 RSS 값을 최소화 시키는 것이다.
  - α가 매우 큰 값이라면, 상대적으로 RSS 값은 의미가 없고 f(w)의 값이 비용 함수의 대부분을 차지한다. 이 때의 목표는 이 f(w)의 값을 최소화 시키는 것이다.

규제 선형 회귀

- 규제 (Regularization)
  - 결과적으로, α의 값을 0부터 차츰 증가시키면 회귀 계수 w 값의 크기를 감소시킬 수 있다.
  - 이렇게 회귀 계수 값의 크기를 감소시켜서 과대적합 문제를 개선하여 학습하는 것을 규제라고 하며, 크게 L1 규제와 L2 규제로 구분한다.
    - L1 규제 : f(w)가 회귀 계수 w들의 절대값들의 합이다.
    - L2 규제 : f(w)가 회귀 계수 w들의 제곱합이다.

릿지 회귀

- 릿지 회귀 (Ridge Regression)
  - L2 규제를 적용한 회귀를 릿지 회귀라고 한다.

$$\text{cost} = \min(\text{RSS}(w) + \alpha \times \frac{1}{2} \sum w_i^2)$$

  - α가 0이라면, 비용 함수는 기존과 동일하게 RSS 값으로 적용된다. 따라서 기본적인 선형 회귀와 같아진다.
  - α가 매우 큰 값이라면, 회귀 계수들의 제곱합이 최소화 되어야 하므로 회귀 계수들이 거의 0에 근접한다. 따라서 회귀 모형은 데이터의 평균을 지나가는 수평선이 된다.

릿지 회귀

- 사이킷런으로 릿지 회귀 수행
  - **linear_model** 모듈에 있는 **Ridge**를 이용하여 릿지 회귀를 수행한다.
    - 매개변수 alpha는 L2 규제 계수인 α 값이다.

```
1  import sklearn.linear_model as lm
2
3  ridge = lm.Ridge(alpha=0.01)
4  reg = ridge.fit(X_train, y_train)
```

    - alpha를 0으로 지정해서 수행하면, LinearRegression을 이용한 일반적인 선형 회귀와 동일한 결과가 나오는 것을 확인할 수 있다.
    - alpha를 증가시키면, 회귀 계수의 값들이 점차 작아져서 0에 가까워지는 것을 확인할 수 있다.

릿지 회귀

- 사이킷런으로 릿지 회귀 수행
  - 당뇨병 진단 데이터에 대해 alpha를 0.01로 설정한 경우

```
# 앞 부분 생략

diab = d.load_diabetes()
X_train, X_test, y_train, y_test = ₩
ms.train_test_split(diab.data, diab.target, ₩
                    test_size=0.3, random_state=78)

ridge = lm.Ridge(alpha=0.01)
reg = ridge.fit(X_train, y_train)

y_pred = reg.predict(X_test)

# 뒷 부분 생략
```

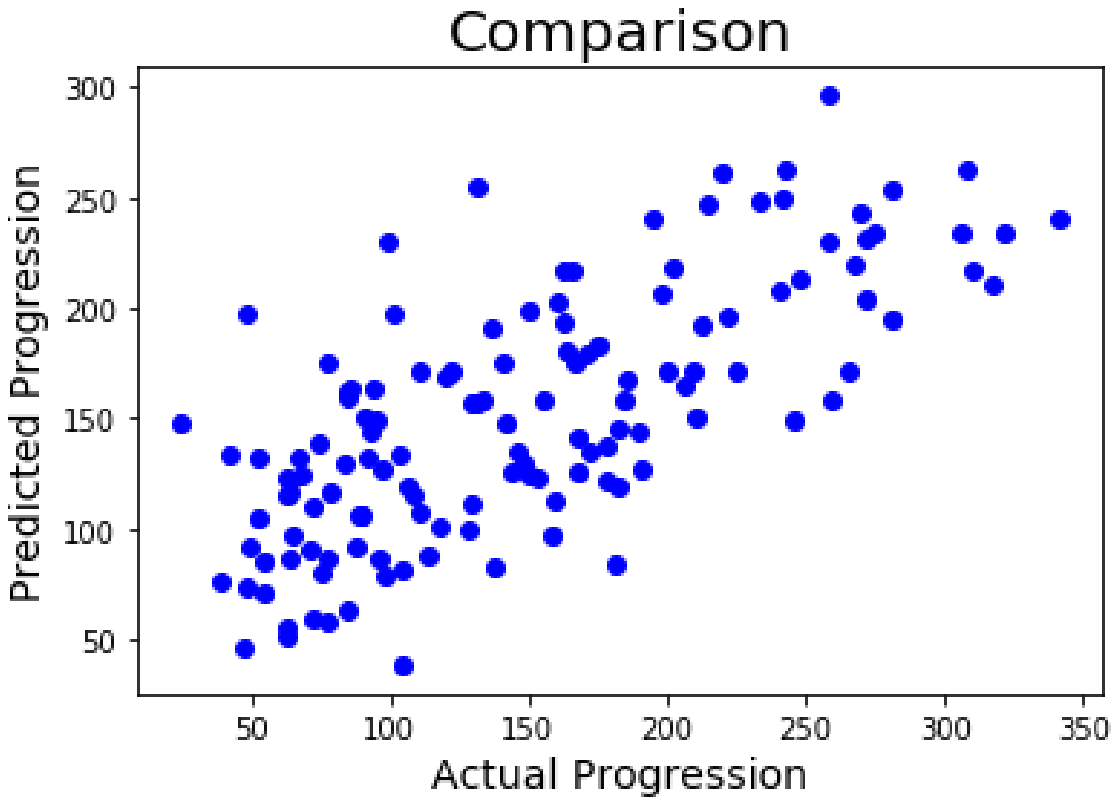
릿지 회귀

- 사이킷런으로 릿지 회귀 수행
  - 당뇨병 진단 데이터에 대해 alpha를 0.01로 설정한 결과

```
R2: 0.513
Adjusted R2: 0.496

회귀 계수:
s5      735.286
bmi     450.733
bp      367.440
s2      239.094
s6       60.548
s4      -23.619
age     -29.339
s3      -95.215
sex    -179.510
s1     -436.699
dtype: float64

절편: 153.422
```



릿지 회귀

- 사이킷런으로 릿지 회귀 수행
  - 당뇨병 진단 데이터에 대해 여러 alpha 값들로 수행한 결과 (훈련 및 검증 데이터를 분리하지 않고 원본 데이터 전체를 학습에 사용하였다.)

	alpha = 0	alpha = 0.01	alpha = 1	alpha = 10	alpha = 100
AGE	-10.012198	-7.199457	29.465746	19.812822	2.897090
SEX	-239.819089	-234.552930	-83.154885	-0.918458	0.585254
BMI	519.839787	520.583136	306.351627	75.416167	9.240719
BP	324.390428	320.523356	201.629434	55.025419	6.931321
S1	-792.184162	-380.607066	5.909369	19.924600	3.230957
S2	476.745838	150.483752	-29.515927	13.948686	2.616766
S3	101.044570	-78.591232	-152.040465	-47.553816	-6.174550
S4	177.064176	130.313059	117.311715	48.259420	6.678027
S5	751.279321	592.349587	262.944995	70.144068	8.876864
S6	67.625386	71.133768	111.878718	44.213876	5.955597
INTERCEPT	152.133484	152.133484	152.133484	152.133484	152.133484

라쏘 회귀

- 라쏘 회귀 (Lasso Regression)
  - L1 규제를 적용한 회귀를 라쏘 회귀라고 한다.

$$\text{cost} = \min(\text{RSS}(w) + \alpha \times \sum |w_i|)$$

  - L1 규제는 불필요한 회귀 계수를 가급적 0으로 만들어 제거하려는 특징이 있다. 따라서 결과적으로는 적절한 특성들만 회귀 모형에 포함시키게 된다.

라쏘 회귀

- 사이킷런으로 라쏘 회귀 수행
  - linear_model 모듈에 있는 Lasso를 이용하여 라쏘 회귀를 수행한다.
    - 매개변수 alpha는 L1 규제 계수인 α 값이다.

```
1  import sklearn.linear_model as lm
2
3  lasso = lm.Lasso(alpha=0.01)
4  reg = lasso.fit(X_train, y_train)
```

    - alpha를 0으로 지정해서 수행하면, LinearRegression을 이용한 일반적인 선형 회귀와 동일한 결과가 나오는 것을 확인할 수 있다.
    - alpha를 증가시키면, 회귀 계수의 값들이 급격하게 0에 수렴하는 것을 확인할 수 있다.

라쏘 회귀

- 사이킷런으로 라쏘 회귀 수행
  - 당뇨병 진단 데이터에 대해 alpha를 0.01로 설정한 경우

```
# 앞 부분 생략

diab = d.load_diabetes()
X_train, X_test, y_train, y_test = ₩
ms.train_test_split(diab.data, diab.target, ₩
                    test_size=0.3, random_state=78)

lasso = lm.Lasso(alpha=0.01)
reg = lasso.fit(X_train, y_train)

y_pred = reg.predict(X_test)

# 뒷 부분 생략
```

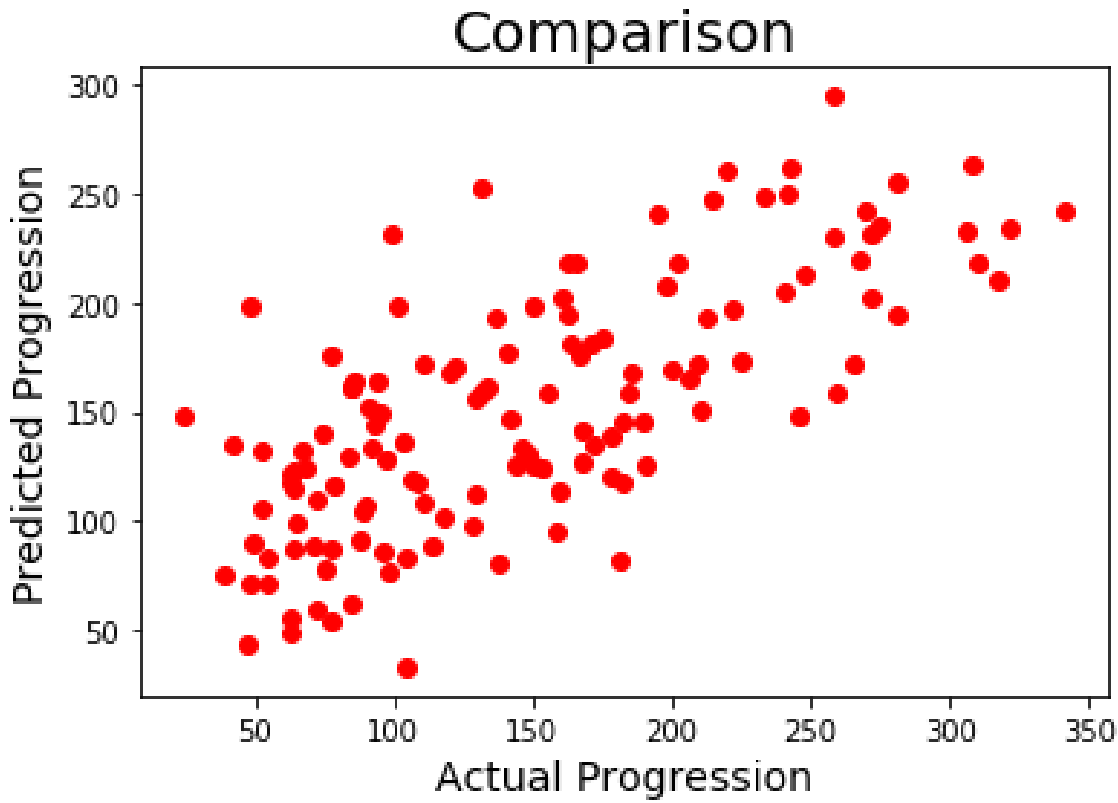
라쏘 회귀

- 사이킷런으로 라쏘 회귀 수행
  - 당뇨병 진단 데이터에 대해 alpha를 0.01로 설정한 결과

```
R2: 0.509
Adjusted R2: 0.492

회귀 계수:
s5      825.933
bmi     451.930
s2      400.500
bp      367.985
s6       50.191
s4        0.000
s3       -0.000
age     -27.913
sex    -177.585
s1     -643.630
dtype: float64

절편: 153.437
```



라쏘 회귀

- 사이킷런으로 라쏘 회귀 수행
  - 당뇨병 진단 데이터에 대해 여러 alpha 값들로 수행한 결과 (훈련 및 검증 데이터를 분리하지 않고 원본 데이터 전체를 학습에 사용하였다.)

	alpha = 0	alpha = 0.01	alpha = 1	alpha = 10	alpha = 100
AGE	-10.012198	-1.306575	0.000000	0.000000	0.000000
SEX	-239.819089	-228.822331	-0.000000	0.000000	0.000000
BMI	519.839787	525.560658	367.701852	0.000000	0.000000
BP	324.390428	316.175320	6.301904	0.000000	0.000000
S1	-792.184135	-307.013677	0.000000	0.000000	0.000000
S2	476.745817	89.321688	0.000000	0.000000	0.000000
S3	101.044558	-105.081398	-0.000000	-0.000000	-0.000000
S4	177.064173	119.597989	0.000000	0.000000	0.000000
S5	751.279311	571.330871	307.605700	0.000000	0.000000
S6	67.625386	65.007316	0.000000	0.000000	0.000000
INTERCEPT	152.133484	152.133484	152.133484	152.133484	152.133484

엘라스틱넷 회귀

- 엘라스틱넷 회귀 (Elastic Net Regression)
  - L1 규제와 L2 규제를 혼합한 회귀를 엘라스틱넷 회귀라고 한다.

$$\text{cost} = \min(\text{RSS}(w) + r\alpha\sum|w_i| + (1-r)\alpha\tfrac{1}{2}\sum w_i^2)$$

  - α는 규제 매개변수이고 r은 혼합 비율이다.
  - 두 종류의 규제를 혼합하여 릿지 회귀와 라쏘 회귀의 절충 형태를 학습 모형으로 도출할 수 있다.
  - r이 0이라면 L2 규제와 동일하므로 릿지 회귀가 된다.
  - r이 1이라면 L1 규제와 동일하므로 라쏘 회귀가 된다.

엘라스틱넷 회귀

- 사이킷런으로 엘라스틱넷 회귀 수행
  - linear_model 모듈에 있는 ElasticNet을 이용하여 엘라스틱넷 회귀를 수행한다.
    - 매개변수 alpha는 규제 계수인 α 값이다.
    - 매개변수 l1_ratio는 혼합 비율 r 이다.

```
1  import sklearn.linear_model as lm
2
3  elastic = lm.ElasticNet(alpha=0.01, l1_ratio=0.5)
4  reg = elastic.fit(X_train, y_train)
```

엘라스틱넷 회귀

- 사이킷런으로 엘라스틱넷 회귀 수행
  - 당뇨병 진단 데이터에 대해 alpha를 0.01로, l1_ratio를 0.5로 설정한 경우

```
# 앞 부분 생략

diab = d.load_diabetes()
X_train, X_test, y_train, y_test = ₩
ms.train_test_split(diab.data, diab.target, ₩
                    test_size=0.3, random_state=78)

elastic = lm.ElasticNet(alpha=0.01, l1_ratio=0.5)
reg = elastic.fit(X_train, y_train)

y_pred = reg.predict(X_test)

# 뒷 부분 생략
```

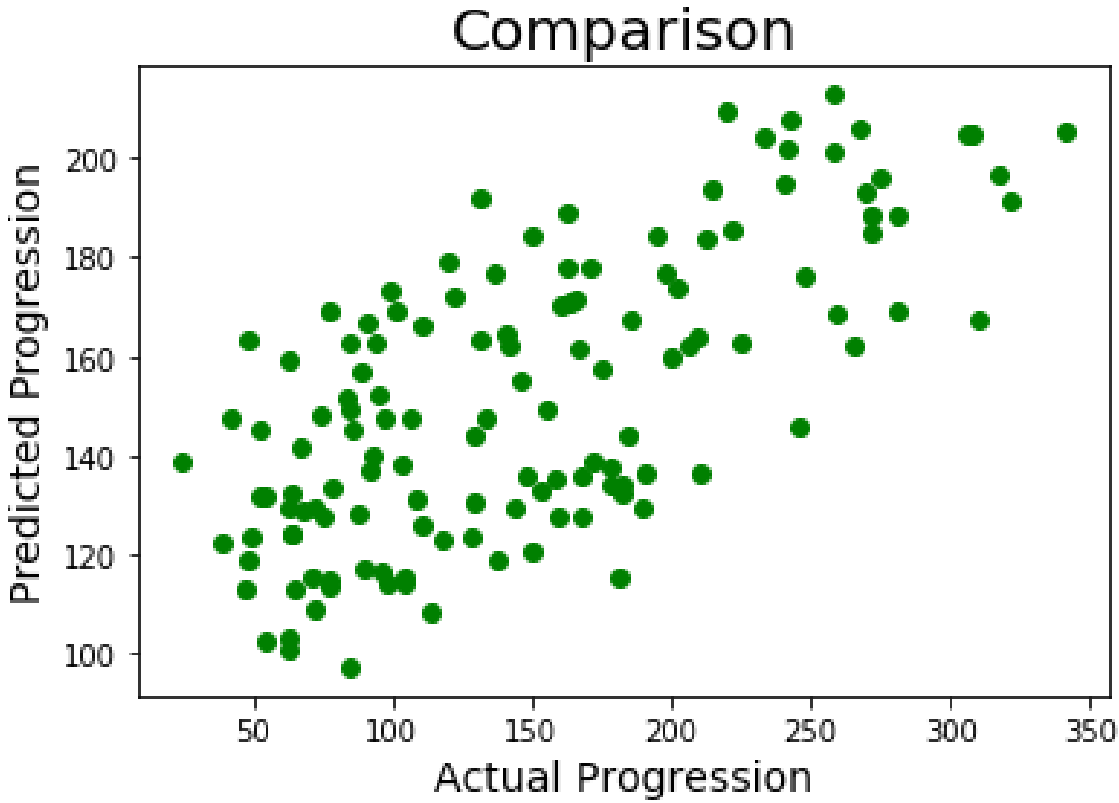
엘라스틱넷 회귀

- 사이킷런으로 엘라스틱넷 회귀 수행
  - 당뇨병 진단 데이터에 대해 alpha를 0.01로, l1_ratio를 0.5로 설정한 결과

```
R2: 0.383
Adjusted R2: 0.362

회귀 계수:
s5      193.035
bmi     188.291
bp      166.146
s6       97.612
s4       88.288
age      41.467
s1       17.336
s2       -6.645
sex     -15.595
s3     -113.204
dtype: float64

절편: 153.29
```



엘라스틱넷 회귀

- 사이킷런으로 엘라스틱넷 회귀 수행
  - 당뇨병 진단 데이터에 대해 l1_ratio 값은 0.5로 고정하고 alpha를 변경하면서 수행한 결과 (훈련 및 검증 데이터를 분리하지 않고 원본 데이터 전체를 학습에 사용하였다.)

	alpha = 0	alpha = 0.01	alpha = 1	alpha = 10	alpha = 100
AGE	-10.012198	33.147202	0.359018	0.000000	0.000000
SEX	-239.819089	-35.245609	0.000000	0.000000	0.000000
BMI	519.839787	211.023930	3.259767	0.000000	0.000000
BP	324.390428	144.560115	2.204356	0.000000	0.000000
S1	-792.184135	21.931533	0.528646	0.000000	0.000000
S2	476.745817	0.000000	0.250935	0.000000	0.000000
S3	101.044558	-115.620017	-1.861363	-0.000000	-0.000000
S4	177.064173	100.658838	2.114454	0.000000	0.000000
S5	751.279311	185.326334	3.105841	0.000000	0.000000
S6	67.625386	96.257214	1.769851	0.000000	0.000000
INTERCEPT	152.133484	152.133484	152.133484	152.133484	152.133484

엘라스틱넷 회귀

- 사이킷런으로 엘라스틱넷 회귀 수행
  - 당뇨병 진단 데이터에 대해 alpha 값은 0.01로 고정하고 l1_ratio를 변경하면서 수행한 결과 (훈련 및 검증 데이터를 분리하지 않고 원본 데이터 전체를 학습에 사용하였다.)

	l1_ratio = 0.01	l1_ratio = 0.1	l1_ratio = 0.5	l1_ratio = 0.75	l1_ratio = 1
AGE	29.664286	30.502682	33.147202	29.228680	-1.306575
SEX	-12.201647	-14.495972	-35.245609	-74.199262	-228.822331
BMI	139.301981	148.364700	211.023930	293.577325	525.560658
BP	98.762731	104.728595	144.560115	193.700364	316.175320
S1	25.795506	25.843804	21.931533	6.387102	-307.013677
S2	13.027108	11.984279	0.000000	-21.860869	89.321688
S3	-82.521697	-87.044388	-115.620017	-146.541969	-105.081398
S4	78.111935	81.552522	100.658838	114.735471	119.597989
S5	125.786550	133.448258	185.326334	252.764800	571.330871
S6	73.338215	76.793979	96.257214	110.043179	65.007316
INTERCEPT	152.133484	152.133484	152.133484	152.133484	152.133484

규제 선형 회귀

- 회귀 모형의 선택
  - 기본적인 선형 회귀, 릿지, 라쏘, 엘라스틱넷 중 어떤 것이 가장 좋은 모형을 도출하는지는 상황에 따라 달라진다.
  - 각 기법 별로 매개변수들을 변경해 가면서 학습하여 최적의 성능 지표를 나타내는 모형을 선택해야 한다.
  - 다만, 다음과 같은 가이드를 참고하면 도움이 될 수 있다.
    - 규제가 없는 기본 선형 회귀를 수행하면서 과대적합이 확실시 되면 릿지 회귀로 전환한다.
    - 전체 특성들 중 일부만 의미가 있다고 유추할 수 있거나 관찰되었다면 라쏘 또는 엘라스틱넷 회귀를 시도한다.
    - 특성들 간의 관련성이 높으면 (즉, 상관 계수가 높으면) 엘라스틱넷 회귀를 적용하는 것이 바람직하다.