Machine Learning & Deep Learning for Data Science

Assignment 1

2022-22047 데이터사이언스학과 신성구

Q1: Linear Regression with [Carseats.csv]

a) Fit a multiple linear regression model to predict Sales using Price, Urban, and US. Report the R2 of the model.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.239
Model:                            OLS   Adj. R-squared:                  0.234
Method:                 Least Squares   F-statistic:                     41.52
Date:                Sun, 16 Oct 2022   Prob (F-statistic):           2.39e-23
Time:                        01:20:33   Log-Likelihood:                -927.66
No. Observations:                 400   AIC:                             1863.
Df Residuals:                     396   BIC:                             1879.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         13.0435      0.651     20.036      0.000      11.764      14.323
Price         -0.0545      0.005    -10.389      0.000      -0.065      -0.044
Urban         -0.0219      0.272     -0.081      0.936      -0.556       0.512
US             1.2006      0.259      4.635      0.000       0.691       1.710
==============================================================================
Omnibus:                        0.676   Durbin-Watson:                   1.912
Prob(Omnibus):                  0.713   Jarque-Bera (JB):                0.758
Skew:                           0.093   Prob(JB):                        0.684
Kurtosis:                       2.897   Cond. No.                         628.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

- 해당 Linear Regression 모델의 R-squared 값은 0.239로 확인하였다.

b) Write out the model in equation form, being careful to handle the qualitative variables properly. Provide an interpretation of each coefficient in the model.

Model: $\widehat{Sales} = 13.0435 - 0.0545 * Price - 0.0219 * Urban + 1.2006 * US$

- Constant: y절편의 값이 13.0435로, predictor들이 0일 때 Sales의 평균 추정값이다.
- 다른 predictor는 고정된 상태에서, Price의 한 단위가 증가하면 Sales는 평균적으로 0.0545만큼 감소한다.
- 다른 predictor는 고정된 상태에서, Urban의 한 단위가 증가하면 Sales는 평균적으로 0.0219만큼 감소한다.
- 다른 predictor는 고정된 상태에서, US의 한 단위가 증가하면 Sales는 평균적으로 1.2006만큼 증가한다.

c) For which predictor variable j can you reject the null hypothesis H0 : βj = 0? for which there is evidence of association with the outcome.

- 귀무가설을 기각할 수 있는 변수: Price(X1), US(X2)
- P-value의 유의수준을 0.05로 가정했을 때, Price와 US의 p-value는 각각 0.000에 가까웠고, 또한 (coefficient의 추정치 ± std error) 범위에 0이 포함되지 않으므로 귀무가설을 기각한다고 볼 수 있다.

d) Obtain 95% confidence intervals for the coefficient(s).

- Code & Result

```
# 신뢰구간
print(result.conf_int(alpha=0.05))
```

```
                0          1
const  11.763597  14.323341
Price  -0.064764  -0.044154
Urban  -0.555973   0.512141
US      0.691304   1.709841
```

- 각 coefficient의 신뢰구간은 위와 같다.

Q2: Logistic Regression with Default.csv

a) Fit a logistic regression model that uses income and balance to predict default. Report the log-likelihood of the model.

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                default   No. Observations:                10000
Model:                          Logit   Df Residuals:                    9997
Method:                           MLE   Df Model:                           2
Date:                Thu, 13 Oct 2022   Pseudo R-squ.:                 0.4594
Time:                        22:54:41   Log-Likelihood:                -789.48
converged:                       True   LL-Null:                       -1460.3
Covariance Type:            nonrobust   LLR p-value:                4.541e-292
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const        -11.5405      0.435    -26.544      0.000     -12.393     -10.688
income      2.081e-05   4.99e-06      4.174      0.000     1.1e-05    3.06e-05
balance        0.0056      0.000     24.835      0.000       0.005       0.006
==============================================================================

Possibly complete quasi-separation: A fraction 0.14 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.
```

- 모델의 log-likelihood는 -789.48 로 확인되었다.

b) Write out the model in equation form and provide an interpretation of each coefficient in the trained model.

- Model: $\widehat{Y} = (\frac{p(X;\widehat{B})}{1-P(X;\widehat{B})}) = -11.5405 + 2.081e05 * income + 0.0056 * balance$
- Constant: 나머지 predictor가 0일 때, Y를 결정하는 절편값
- Income: 다른 predictor가 고정된 상태에서, Income이 한 단위 증가하면 log-odds가 2.081e-05만큼 증가한다.
- Balance: 다른 predictor가 고정된 상태에서, Balance가 한 단위 증가하면 log-odds가 0.0056만큼 증가한다.

c) Perform 5-fold cross-validation using the model in Part (a), and estimate the test error of this model.

- Code & Results

```
[CV_1] Validation Error: 0.029
[CV_2] Validation Error: 0.028
[CV_3] Validation Error: 0.028
[CV_4] Validation Error: 0.0365
[CV_5] Validation Error: 0.023
```

```
Mean Validation Error: 0.0289
```

- Test Error는 0.0289로 추정된다.

d) Now consider a logistic regression model that predicts the probability of default using income, balance, and a dummy variable for student. Estimate the test error for this model using the 5-fold cross-validation set approach. Comment on whether or not including a dummy variable for student would lead to a reduction in the test error rate.

- Code & Results

```
[CV_1] Validation Error: 0.03%
[CV_2] Validation Error: 0.027%
[CV_3] Validation Error: 0.0315%
[CV_4] Validation Error: 0.0365%
[CV_5] Validation Error: 0.0255%
```

```
Mean Validation Error: 0.0301%
```

- Predictor 'Student'를 추가한 모델의 test error가 0.0301로 추정되므로, 추가하기 이전 모델의 test error 추정값(0.0289)보다 오히려 증가하였다. 따라서 Student 변수의 추가는 test error rate을 증가시킬 것으로 추정된다.

Q3: LR, Ridge, Lasso with College.csv

a) Fit a linear model using least squares on the training set, and report the test error obtained.

- Linear Regression의 모델의 구성은 아래와 같다.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   Apps   R-squared:                       0.927
Model:                            OLS   Adj. R-squared:                  0.925
Method:                 Least Squares   F-statistic:                     505.8
Date:                Sun, 16 Oct 2022   Prob (F-statistic):               0.00
Time:                        02:55:42   Log-Likelihood:                -5856.1
No. Observations:                 699   AIC:                         1.175e+04
Df Residuals:                     681   BIC:                         1.183e+04
Df Model:                          17
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -566.4451    440.436     -1.286      0.199   -1431.221     298.331
Private     -548.4224    148.313     -3.698      0.000    -839.628    -257.217
Accept         1.5901      0.043     36.641      0.000       1.505       1.675
Enroll        -0.9100      0.209     -4.359      0.000      -1.320      -0.500
Top10perc     48.0672      6.065      7.926      0.000      36.160      59.975
Top25perc    -13.3022      4.906     -2.711      0.007     -22.935      -3.669
F.Undergrad    0.0614      0.036      1.684      0.093      -0.010       0.133
P.Undergrad    0.0429      0.035      1.238      0.216      -0.025       0.111
Outstate      -0.0874      0.021     -4.245      0.000      -0.128      -0.047
Room.Board     0.1607      0.052      3.091      0.002       0.059       0.263
Books          0.0392      0.252      0.156      0.876      -0.456       0.534
Personal       0.0427      0.067      0.641      0.522      -0.088       0.174
PhD          -10.0594      5.123     -1.963      0.050     -20.119    9.62e-05
Terminal      -3.2520      5.561     -0.585      0.559     -14.171       7.667
S.F.Ratio     19.0006     13.792      1.378      0.169      -8.079      46.080
perc.alumni    0.8089      4.407      0.184      0.854      -7.845       9.463
Expend         0.0847      0.014      5.928      0.000       0.057       0.113
Grad.Rate      9.9552      3.167      3.144      0.002       3.738      16.173
==============================================================================
Omnibus:                      429.701   Durbin-Watson:                   1.978
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7951.010
Skew:                           2.397   Prob(JB):                         0.00
Kurtosis:                      18.812   Cond. No.                     1.82e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.82e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

- 해당 Linear Regression 모델의 test-error는 아래와 같다.

```
[Linear Regression] Test MSE: 642554.0075
```

b) Fit a ridge regression model on the training set, with λ chosen by 10-fold cross-validation. Report the test error obtained.

- 10-fold Cross-Validation으로 구한 alpha(lambda)는 아래 코드의 결과와 같다.

```
from sklearn.linear_model import Ridge, RidgeCV, Lasso, LassoCV
alpha = np.logspace(-2, 3, 100) # alpha(lambda) 후보군 생성

# Cross Validation
cv_r = KFold(n_splits=10, shuffle=True, random_state=108)
ridge = RidgeCV(alphas = alpha, cv=cv_r, scoring="neg_mean_squared_error")
ridge_cv = ridge.fit(X_train, Y_train)
print(ridge_cv.alpha_)
```

```
 0.01
```

- 해당 Ridge 모델의 test-error는 아래와 같다.

```
[Ridge Regression] Test MSE: 642533.04046
```

c) Fit a lasso model on the training set, with λ chosen by 10-fold crossvalidation. Report the test error obtained, along with the number of non-zero coefficient estimates.

- 10-fold Cross-Validation으로 구한 alpha(lambda)는 아래 코드의 결과와 같다.

```
from sklearn.model_selection import GridSearchCV

# alpha(lambda) 후보군 생성
alpha = np.logspace(-2, 3, 100)
alpha = {'alpha':list(alpha)}

# Cross Validation of Lasso
cv_l = KFold(n_splits=10, shuffle=True, random_state=108)
lasso_cv = GridSearchCV(estimator=Lasso(), param_grid=alpha, scoring="neg_mean_squared_error", cv=cv_l)
lasso_cv.fit(X_train, Y_train)
print(lasso_cv.best_params_)
```

```
 {'alpha': 13.530477745798061}
```

- 해당 Lambda 모델의 test-error는 아래와 같다.

```
[Lasso Regression] Test MSE: 630397.65085
```

- Lambda 모델의 coefficient 중에서 0이 아닌 것의 개수는 15개로 확인되었다.

```
# coefficient == 0이 아닌 predictor의 개수
print(f"Number of Non-Zero Coefficients: {len(model_l.coef_) - np.sum(model_l.coef_ == 0)}")
```

```
 Number of Non-Zero Coefficients: 15
```

d) Comment on the results obtained. How accurately can you predict the number of college applications received? Is there much difference among the test errors resulting from these three

approaches? Which model would you use?

- Results

```
<<<Results>>>
[Linear] Test R2 Score: 0.95331, Test MSE: 642554.0075
[Ridge] Test R2 Score: 0.95331, Test MSE: 642533.04046
[Lasso] Test R2 Score: 0.95419, Test MSE: 630397.65085
```

- Linear Regression과 Ridge, Lasso의 R-squared 값은 모두 0.95보다 큰, 비교적 좋은 성능을 보이는 모델이다.
- Linear Regression과 Ridge Regression의 경우에는 Ridge 모델이 약간 낮은 test error를 보였다.
- 그러나 Lasso 모델의 경우 test error가 약 630397로, 앞선 두 모델보다 크게 낮은 test error를 보였다.
- 따라서 1) test error가 가장 낮고 2) 변수의 개수도 2개 감소한, 즉 simpler 모델인 Lasso 모델을 선택할 것이다.