과제1

모범답안

1. [**50 pts**] In this problem, you will use the `Carseats` data set attached in the assignment (`Carseats.csv`) for linear regression.

(a) [5 pts] Fit a multiple linear regression model to predict `Sales` using `Price`, `Urban`, and `US`. Report the R^2 of the model.

```
# Import libraries
import pandas as pd
import statsmodels.formula.api as smf

# Read the dataset
carseats = pd.read_csv('Carseats.csv')

# Create dummy variables for qualitative predictors 'Urban' and 'US'
carseats = pd.get_dummies(carseats)

# Fit the multiple linear regression model
mlr = smf.ols("Sales ~ Price + Urban_Yes + US_Yes", data = carseats).fit()

# Report summary
print(mlr.summary())
```

```
                            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:                Tue, 11 Oct 2022   Prob (F-statistic):           2.39e-23
Time:                        16:50:17   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]
------------------------------------------------------------------------------
Intercept     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_Yes     -0.0219      0.272     -0.081      0.936      -0.556       0.512
US_Yes         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.
```

```
# Report the R-squared of the model
print("R-squared:", mlr.rsquared.round(3))
```

```
R-squared: 0.239
```

$R^2 = 0.239$

(b) [10 pts] Provide an interpretation of each coefficient in the model. Be careful — some of the variables in the model are qualitative!

and

(c) [5 pts] Write out the model in equation form, being careful to handle the qualitative variables properly.

$$\text{Sales} = \beta_0 + \beta_1 \cdot \text{Price} + \beta_2 \cdot I(\text{Urban} = \text{Yes}) + \beta_3 \cdot I(\text{US} = \text{Yes}) + \epsilon,\ \epsilon \sim N(0, \sigma^2)$$
$$(\, I(\cdot) : \text{indicator function}\,)$$

(a) 에서의 Linear regression을 통하여 계산된 계수들을 통하여 모형을 나타내면 다음과 같다.

$$\text{Sales} = 13.0435 - 0.0545 \cdot \text{Price} - 0.0219 \cdot I(\text{Urban} = \text{Yes}) + 1.2006 \cdot I(\text{US} = \text{Yes}) + \epsilon$$

각 계수를 해석해보면 다음과 같다.

- β_0 : `Price` = 0, `Urban` = No, `US` = No일 때, Sales의 예측값이 13.0435이다.
- β_1 : 다른 설명 변수가 고정일 때, `Price` 가 1단위 증가하면, `Sales` 가 0.0545 감소한다.
- β_2 : 다른 설명 변수가 고정일 때, `Urban` 이 `No` 에서 `Yes` 일 때의 `Sales` 의 차이가 -0.0219이다.
- β_3 : 다른 설명 변수가 고정일 때, `US` 가 `No` 에서 `Yes` 일 때의 `Sales` 의 차이가 1.2006 이다.

(d) [5 pts] For which predictor variable j can you reject the null hypothesis $H_0 : \beta_j = 0$?

(a)에서의 OLS Regression Results table을 보면, β_2를 제외한 계수의 p-value가 0.000으로 아주 작고, β_2에 해당하는 p-value는 0.936으로 0.05보다 크므로, 유의수준 0.05하 귀무가설 $H_0 : \beta_j = 0$를 기각하는 $j = 0, 1, 3$이다.

참고자료 : 귀무가설 (영가설, null hypothesis)

출처: 위키피디아 https://ko.wikipedia.org/wiki/%EA%B7%80%EB%AC%B4_%EA%B0%80%EC%84%A4

귀무 가설(null hypothesis, 기호 H_0) 또는 **영 가설**은 통계학에서 처음부터 버릴 것을 예상하는 가설이다. **차이가 없거나 의미있는 차이가 없는 경우의 가설**이며 **이것이 맞거나 맞지 않다는 통계학적 증거를 통해 증명하려는 가설**이다.

예를 들어 범죄 사건에서 용의자가 있을 때 형사는 이 용의자가 범죄를 저질렀다는 추정인 대립가설을 세우게 된다. 이때 **귀무가설은 용의자는 무죄라는 가설**이다. 통계적인 방법으로 가설검정(hypothesis test)을 시도할 때 쓰인다.

예시

기본적으로는 참으로 추정되며 이를 거부하기 위해서는 증거가 꼭 필요하다. 예를 들어 남학생과 여학생들의 두 성적 샘플을 비교해 볼 때, 귀무가설은 남학생들의 평균이 여학생들의 평균과 같은 것이라는 것이다.

$H_0 : \mu_1 = \mu_2$

여기서:

H_0 = **귀무가설**

μ_1 = 집단1의 평균

μ_2 = 집단2의 평균

또한 귀무가설이 같은 집단으로부터 뽑힌 두 샘플들이라고 가정하고 그래서 평균과 더불어 분산과 분포는 같다고 가정한다. 이러한 귀무가설의 설정은 통계적 유의성을 시험하는 데 중요한 단계이다.

이러한 가설을 형성하고 얻어진 데이터에서 확률적 검정을 해봄으로써 귀무가설이 예측하는 것이 맞는지 아닌지를 알아 볼 수 있다. 또한 만약 이것이 참이라면 여기서 얻어진 확률은 결과의 유의수준으로 부른다.

(e) [5 pts] On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

OLS Regression Results

Dep. Variable:	Sales	R-squared:	0.239
Model:	OLS	Adj. R-squared:	0.235
Method:	Least Squares	F-statistic:	62.43
Date:	Mon, 27 Mar 2023	Prob (F-statistic):	2.66e-24
Time:	19:05:28	Log-Likelihood:	-927.66
No. Observations:	400	AIC:	1861.
Df Residuals:	397	BIC:	1873.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	13.0308	0.631	20.652	0.000	11.790	14.271
US[T.Yes]	1.1996	0.258	4.641	0.000	0.692	1.708
Price	-0.0545	0.005	-10.416	0.000	-0.065	-0.044

Omnibus:	0.666	Durbin-Watson:	1.912
Prob(Omnibus):	0.717	Jarque-Bera (JB):	0.749
Skew:	0.092	Prob(JB):	0.688
Kurtosis:	2.895	Cond. No.	607.

(f) [10 pts] How well do the models in Part (a) and Part (e) fit the data? Compare the two models. Can you say one model is better than the other based on R^2? Provide justification for your answer.

$$R_{adj}^{2} = 1 - \frac{(1-R^{r})(n-1)}{n-k-1}$$ n: 데이터 개수, k: 독립변수 개수

- R^2은 독립변수가 추가될수록 증가하여 overfitting이 발생 할 수 있다. R_{adj}^2은 유의미한 독립변수가 추가될 때 증가한다. a) 0.234 < e) 0.235 이므로 e)의 모델이 a)보다 적합하다.
- 두 모델의 R^2 는 0.239로 동일하고 위의 이유에 따라 R^2로의 비교는 적절하지 않다.

(g) [10 pts] Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

(a)에서의 OLS Regression Results table을 보면, 각각의 계수의 95% 신뢰구간은 다음과 같다.

β	신뢰 하한	신뢰 상한
β_0	11.764	14.323
β_1	-0.065	-0.044
β_2	-0.556	0.512
β_3	0.691	1.710

2. [**50 pts**] In class, we used the example of the logistic regression model to predict the probability of `default` using `income` and `balance` on the `Default` data set attached in the assignment (`Default.csv`). In this problem, we will estimate the test error of this logistic regression model using the validation set approach. Do not forget to set a random seed before beginning your analysis.

   (a) [10 pts] Fit a logistic regression model that uses `income` and `balance` to predict `default` using the validation set approach, and estimate the test error of this model. In order to do this, you must perform the following steps:

I. Split the sample set into a training set and a validation set.

- 7:3의 비율로 Training set과 Test set을 나누었다. 그 과정에서 결과 replication을 위해서 random_state는 123으로 설정했다.

```
X = defaults[["income","balance"]]
y = defaults["default_yes"]
y = np.array(y)  # for easy calculation


X_train, X_val, y_train, y_val = train_test_split(X,y,test_size=0.3, random_state=123)
```

II. Fit a multiple logistic regression model using only the training observations.

- sklearn에서 제공하는모듈을 사용해서 다중 로지스틱 회귀를 진행했다.

```
Q2_mlr = skl_lm.LogisticRegression()
Q2_mlr.fit(X_train,y_train)
```

III. Obtain a prediction of default status for each individual in the validation set by computing the posterior probability of default for that individual, and classifying the individual to the default category if the posterior probability is greater than 0.5.

```
y_val_pred = Q2_mlr.predict_proba(X_val)
threshold = 0.5
y_val_pred_threshold = np.where(y_val_pred[:,1] > threshold, 1, 0)
```

IV. Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.

- **Error rate은 0.033** 이다.

```
error_rate = sum(y_val != y_val_pred_threshold)/len(y_val)
error_rate
```

```
0.03266666666666666
```

(b) [10 pts] Provide an interpretation of each coefficient in the trained model.

We first obtain the intercept and the coefficients of the logistic regression model from part (a).

```
print('Intercept: ', logreg.intercept_)
print('Coefficients: ', logreg.coef_)
```

We get the following results:

```
Intercept: [-11.54047811]
Coefficients: [[2.08091984e-05 5.64710797e-03]]
```

The model in equation form would be written as:

$$\log\left(\frac{p(X;\beta)}{1-p(X;\beta)}\right) = \beta_0 + \beta_1 X_{\text{income}} + \beta_2 X_{\text{balance}}$$

In this equation,

- log(p(X;β)/(1-p(X;β))) : log odds of `default`
- X_{income}, X_{balance} : feature vectors (`income`, `balance`)
- β_0 : intercept
- β_1, β_2 : slope coefficients of the features `income` and `balance`

Plugging in the values into the model:

$$\log\left(\frac{p(X;\beta)}{1-p(X;\beta)}\right) = -11.5405 + (2.0809 \times 10^{-5})X_{\text{income}} + (5.6471 \times 10^{-3})X_{\text{balance}}$$

We may interpret the coefficient as follows:

For coefficient β_1 (`income`), a unit increase in `income` will cause an increase of 2.0809×10^{-5} in the log odds of `default`.

For coefficient β_2 (`balance`), a unit increase in `balance` will cause an increase of 5.6471×10^{-3} in the log odds of `default`.

(c) [15 pts] Perform 5-fold cross-validation using the same model as in Part (a). Write your own code for the K-fold cross-validation. How does the validation error of the cross-validation the differ from to the results in Part (a)?

2-(c). 5-fold cross-validation 평균 test error=0.0263

```
##### 5-fold cross validation의 test set index를 위한 난수 별도 생성 #####

idx=np.array(range(data2_y.shape[0])) #data의 index를 미리 생성한다.
idx_random=np.random.RandomState(seed=101).permutation(idx) #idx를 무작위로 섞는다.
#5-fold cross-validation이기 때문에, 각 validation에서 test index로 사용하기 위해 무작위로 섞은 idx를 5개의 그룹으로 나눈다.
idx_split=np.array_split(idx_random, 5)


### 5-fold cross validation, 나누어진 idx_split 갯수가 5개.

for iter in range(len(idx_split)):
  test_idx=idx_split[iter] #test index로 사용하기 위해 앞서 생성해둔 idx_split 이용
  train_idx=np.setdiff1d(idx_random,test_idx) #test index를 제외한 나머지 index를 train_idx로 정의
  x_train2, y_train2 = data2_x[train_idx,:], data2_y[train_idx]
  x_test2, y_test2 = data2_x[test_idx,:], data2_y[test_idx]

  print("5-fold cross validation %i 번째" %(iter+1))
  x_train_bias=sm.add_constant(x_train2) #bias를 주기 위함.
  x_test_bias=sm.add_constant(x_test2)
  model3=sm.Logit(y_train2,x_train_bias).fit()
  print("coefficient beta_0: ",model3.params[0], "beta_1: ", model3.params[1], "beta_2: ",model3.params[2],"\n")

  #Train accuracy
  y_pred_train=(model3.predict(x_train_bias)>=0.5).astype(int)
  y_miss_score_train=np.sum(y_pred_train!=y_train2)/len(y_train2)
  print("Training set error: %0.4f"%y_miss_score_train) #misclassified fraction in training set

  #Test accuracy
  y_pred=(model3.predict(x_test_bias)>=0.5).astype(int)
  y_miss_score=np.sum(y_pred!=y_test2)/len(y_test2)
  print("Validation set error: %0.4f"%y_miss_score) #misclassified fraction in test set
  print("-----------------------------------------------")
```

(d) [15 pts] 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.

2-(d). 5-fold cross-validation 평균 test error=0.0266

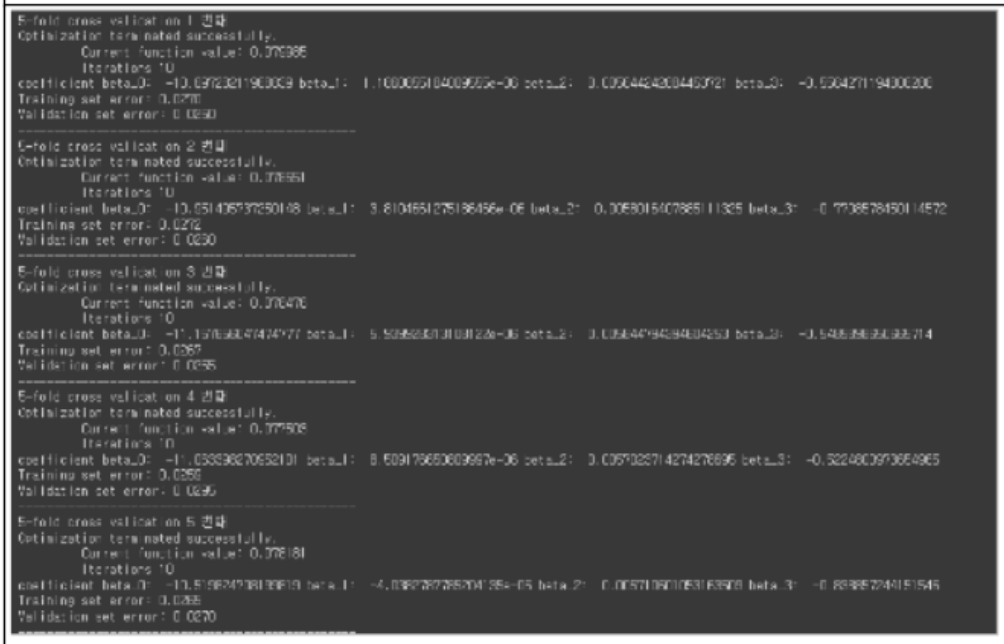
X_1은 Income, X_2는 Balance라 하고 default와 student를 아래와 같이 둘 경우,

$$\text{Default } Y = \begin{cases} 1 & \text{만약, default 항목이 yes일 경우} \\ 0 & \text{만약, default 항목이 No일 경우} \end{cases}, \quad \text{Student } X_3 = \begin{cases} 1 & \text{만약, studnet 항목이 yes일 경우} \\ 0 & \text{만약, student 항목이 No일 경우} \end{cases}$$

모델은 아래와 같습니다.

$$\mathrm{P}(Y=1|X) = p(\mathrm{x};\beta) = \frac{\exp(\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3)}{1+\exp(\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3)}$$

이때 2-(c)와 동일한 방법으로 5-fold cross validation을 수행하면 아래와 같은 결과가 나옵니다.



즉, Student variable을 dummy variable로 포함한 모델의 경우, 5-fold cross validation에서 평균 test error가 평균 0.0266이고, student variable을 dummy variable로 추가한다고 하여 2-(c)의 model에 비해 test error rate가 줄어들지 않습니다.