AI5213/EC4213: Machine Learning and Deep Learning **Student ID: 20195024**
Homework 2 (Linear Regression, Classification) **Name: Kim, Minjun**

Due: Sunday, Sep 24, 11:59PM

This homework comprises a set of conceptual problems and one coding exercise. Some problems are trivial, while others will require a lot of thought. Start this homework early!

Guideline for those new to data analysis using Python:
We recommend you to review the Lab section within each chapter (e.g., p. 116 of Ch. 3 or https://islp.readthedocs.io/en/latest/labs/Ch03-linreg-lab.html) prior to tackling the programming tasks. Additionally, find datasets and Jupyter notebooks at https://github.com/intro-stat-learning/ISLP_labs/ and https://islp.readthedocs.io/en/latest/.

Visit https://www.statlearning.com/forum — a dedicated forum created by and for the ISL community. Whether you have a question or encounter issues with ISLP labs, this platform is your go-to resource for assistance and collaborative discussions.

Deliverables:

1. Submit a PDF of your homework, with an appendix listing all your code, to the Gradescope assignment entitled "HW2 Write-Up". You may typeset your homework in LaTeX or Word or submit neatly handwritten and scanned solutions. Please start each question on a new page. If there are graphs, include those graphs in the correct sections. Do not put them in an appendix. We need each solution to be self-contained on pages of its own.

   • On the first page of your write-up, please sign your signature next to the following statement. (Mac Preview, PDF Expert, and Foxit PDF Reader, among others, have tools to let you sign a PDF file.) We want to make extra clear the consequences of cheating.

   "I certify that all solutions are entirely in my own words and that I have not looked at another student's solutions. I have given credit to all external sources I consulted."

   • On the first page of your write-up, please list students who helped you or whom you helped on the homework. (Note that sending each other code is not allowed.)

2. Submit all the code needed to reproduce your results to the Gradescope assignment entitled "HW2 Code". You must submit your code twice: once in your PDF write-up (above) so the readers can easily read it, and again in compilable/interpretable form so the readers can easily run it. Do NOT include any data files we provided. Please include a short file named README listing your name, student ID, and instructions on how to reproduce your results. Please take care that your code doesn't take up inordinate amounts of time or memory.

For staff use only

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Total
/ 4	/ 6	/ 8	/ 12	/ 14	/ 6	/ 6	/ 10	/ 14	/ 20	/ 100

Honor Code

Declare and sign the following statement:

"I certify that all solutions in this document are entirely my own and that I have not looked at anyone else's solution. I have given credit to all external sources I consulted."

Signature: I certify that all solutions in this document are entirely my own and that I have not look at anyone else's solution. I have given credit to all external sources I consulted.

We welcome group discussions, but the work you submit should be entirely your own. If you use any information or pictures not from our lectures or readings, make sure to say where they came from. Please note that breaking academic rules can lead to severe penalties.

(a) Did you receive any help whatsoever from anyone in solving this assignment? If your answer is 'yes', give full details (e.g. "Junho explained to me what is asked in Q2-a")

No.

(b) Did you give any help whatsoever to anyone in solving this assignment? If your answer is 'yes', give full details (e.g. "I pointed Josh to Ch. 2.3 since he didn't know how to proceed with Q2")

No.

(c) Did you find or come across code that implements any part of this assignment? If your answer is 'yes', give full details (book & page, URL & location within the page, etc.).

No.

Q1. Solve ISLP Ch.3, Exercise #1 [4 pts]

Describe the null hypothesis to which the p-values given in Table 1 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear model.

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Table 1: For the Advertising data, least squares coefficient estimates of the multiple linear regression of number of units sold (sales) on TV, radio, and newspaper advertising budgets. (Recall that the sales variable is in thousands of units, and the three predictor variables are in thousands of dollars.)

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 \tag{1}$$

Let's consider the equation (1) to apply this problem. The coefficient estimate $\hat{\beta}_i$ is estimate for β_i , and let TV, radio, newspaper is X_1, X_2, X_3 respectively.

For Intercept, the null hypothesis is $H_0 : \beta_0 = 0$. We know that $beta_0$ means the sales when there are no TV, radio, newspaper advertising. So H_0 means that there are no sales without $X_1, X_2, and X_3$. However, since $p - value < 0.0001$, it is clear that we reject this "null hypothesis", which means there would still exist sales without any advertisement.

For TV and radio, each null hypothesis is $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$. In this case, the null hypothesis means each corresponding X is not associated with Y when we hold all other predictors fixed. However, since both $p - value < 0.0001$, we can reject those "null hypothesis", which means Sales and TV advertising are highly correlated. Likewise, Sales and radio advertising are also highly correlated.

For newspaper, the null hypothesis is $H_0 : \beta_3 = 0$, which means X_3 is has no impact on the sales with fixed other predictors. Since $p - value$ is big enough, we cannot reject the null hypothesis. Therefore, Sales and newspaper advertising are not significantly correlated.

Q2. Solve ISLP Ch.3, Exercise #3 [6 pts]

Suppose we have a data set with five predictors, X_1 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School, X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Level. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = 0.01$, $\hat{\beta}_5 = -10$.

(a) Which answer is correct, and why? [2 pts]

(i) For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates.

(ii) For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.

(iii) For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.

(iv) For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates provided that the GPA is high enough.

We predict starting salary after graduation by using

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_5 \tag{2}$$

Since we know that $X_4 = X_1X_2$ and $X_5 = X_1X_3$, let's substitute given values to above model equation. Then we get $\hat{Y} = 50 + 20X_1 + 0.07X_2 + 35X_3 + 0.01X_1X_2 - 10X_1X_3$.
For high school graduates, $\hat{Y_{highschool}} = 50 + 20X_1 + 0.07X_2 + 0.01X_1X_2$.
For college graduates, $\hat{Y_{college}} = 85 + 10X_1 + 0.07X_2 + 0.01X_1X_2$.
If we subtract $\hat{Y_{highschool}}$ from $\hat{Y_{college}}$, then we get $\hat{Y_{college}} - \hat{Y_{highschool}} = 35 - 10X_1$. From this equation, we can realize that, for $X_1 > 3.5$, $\hat{Y_{college}} < \hat{Y_{highschool}}$, which means, for a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.
$\therefore$ (iii) is correct.

(b) Predict the salary of a college graduate with IQ of 105 and a GPA of 3.9. [2 pts]

By (2), $\hat{Y} = 50 + 20 \times 3.9 + 0.07 \times 105 + 35 \times 1 + 0.01 \times 3.9 \times 105 - 10 \times 3.9 \times 1 = 135.445$. The response is in units of thousands of dollars. Therefore, the predicted salary would be \$135,445.

(c) True or false: Since the coefficient for the GPA / IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer. [2 pts]

False, if GPA and IQ is big enough, there are some possibilities that GPA/IQ interaction term would significantly impact on $\hat{Y}$. So instead of coefficient, we have to compute p-value to determine whether each predictors are statistically meaningful.

Q3. Solve ISLP Ch.3, Exercise #4 [8 pts]

I collect a set of data ($n = 100$ observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$. [2 pts]

(a) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer. [2 pts]

> Since the cubic regression model is more flexible than a linear regression model, the cubic regression model would led to a better fit to the training data. This means that the training RSS for the cubic regression would be lower than the training RSS for the linear regression.

(b) Answer (a) using test rather than training RSS. [2 pts]

> The more flexible the model is, the greater the variance. Because of the condition that the true relationship between X and Y is linear, there are some possibilities of overfitting when we use the Cubic regression. Therefore, the linear regression will have lower test RSS.

(c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify the answer. [2 pts]

> Regardless of whether the true relationship between x and y is linear or not, more flexible model has a better fit to the training set. As the same case as (a), the training RSS for the cubic regression is lower than the training RSS for the linear regression, because the former is more flexible.

(d) Answer (c) using test rather than training RSS. [2 pts]

> There is not enough information to compare the test RSS for the two model. Since we don't know the exact true relationship between X and Y, we can't determine either model lead to the lower test RSS. We only know that the the true relationship is not linear, but we don't know how non-linear it is. So we need more information about the relationship to compare the test RSS of the two models.

Q4. Solve ISLP Ch.3, Exercise #10 [12 pts]

This question should be answered using the Carseats data set.

[Note] Your code for all of the programming exercises including this one should be submitted to the corresponding Programming submission slot on Gradescope.

(a) Fit a multiple regression model to predict Sales using Price, Urban, and US. [1.5 pts]

I fitted a multiple regression model to predict Sales using Price, Urban, and US. And R-squared I got from this model is 0.239.

(b) Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative! [1.5 pts]

	coef	std err	t	$P > \|t\|$
intercept	13.0435	0.651	20.036	0.000
Price	-0.0545	0.005	-10.389	0.000
Urban[Yes]	-0.0219	0.272	-0.081	0.936
US[Yes]	1.2006	0.259	4.635	0.000

I got the table above using "summarize()" function. Note that US[Yes] is 1 when US is Yes, and 0 when US is No. Urban[Yes] is also 1 when Urban is Yes, and 0 when Urban is No. Let's interpret above table. The p-values of Price and US are significantly small, which means that the Price and US are highly correlated with Sales. To be more specific, for the higher Price, the lower Sales follows, because of its negative coefficient. Likewise, there would be higher Sales in the US than those of other countries, because of the positive coefficient. On the other hand, the p-value of Urban is very big, which means that we should not reject null hypothesis corresponding to Urban. So we can conclude that Urban is not significantly correlated with Sales.

(c) Write out the model in equation form, being careful to handle the qualitative variables properly. [1.5 pts]

From the table in Q4(b), the equation form of the model is,
$Sales = 13.0435 - 0.0545 \times Price - 0.0219 \times Urban[Yes] + 1.2006 \times US[Yes]$.

(d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$? [1.5 pts]

As we discussed in the previous Q4(b), we can reject the null hypothesis of intercept, Price and US. This is because their p-values are significantly small, which indicates that they are highly correlated with Sales, the response. Whereas, the null hypothesis of Urban cannot be rejected, because of its big p-values.

(e) 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. [1.5 pts]

I picked Price and US[Yes] to fit a smaller model since it is found that they are highly correlated with Sales, while Urban[Yes] is not significant.

(f) How well do the models in (a) and (e) fit the data? [1.5 pts]

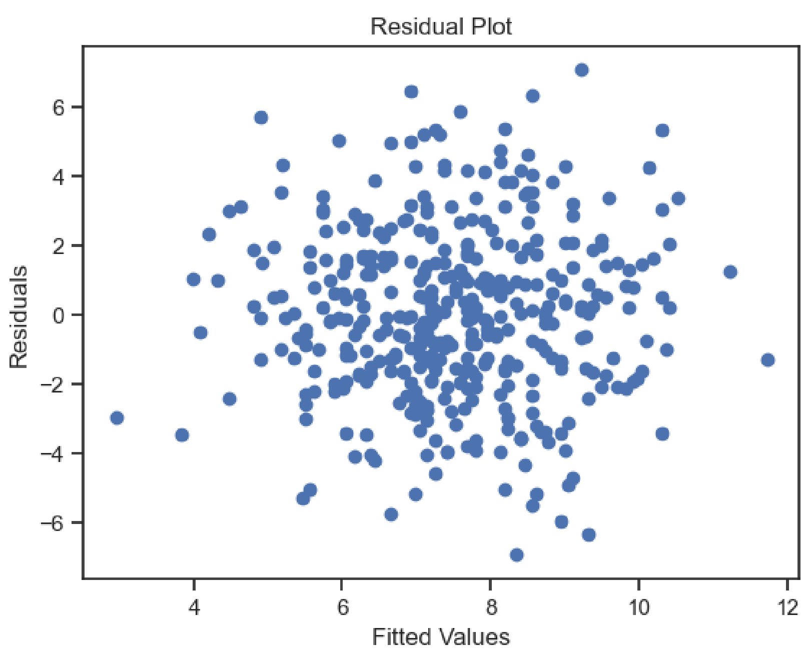
Since the R-squared for the models in (a) and (e) are almost identical, we can remove the terms of Urban[Yes] when we fit a multiple regression model.

(g) Using the model from (e), obtain 95% confidence intervals for the coefficients. [1.5 pts]

For linear regression, the 95% confidence interval for the coefficients approximately satisfies $\hat{\beta}_i \pm 2 \cdot SE(\hat{\beta}_i)$. Therefore, the 95% confidence intervals are the same as the below.

95% confidence intervals for intercept is [11.768799999999999, 14.2928].
95% confidence intervals for Price is [-0.0645, -0.0445].
95% confidence intervals for US[Yes] is [0.6836, 1.7156].

(h) Is there evidence of outliers or high leverage observations in the model from (e)? [1.5 pts]



We can use the residual plot to find outliers. Values that deviate significantly from the residual plot can be considered outliers. Therefore, if we see the plot above, we cannot find some outstanding outliers in there. But, I think there are some high leverage observations in the model.

Q5. Solve ISLP Ch.3, Exercise #14 [14 pts]

This problem focuses on the *collinearity* problem.

(a) Perform the following commands in Python:

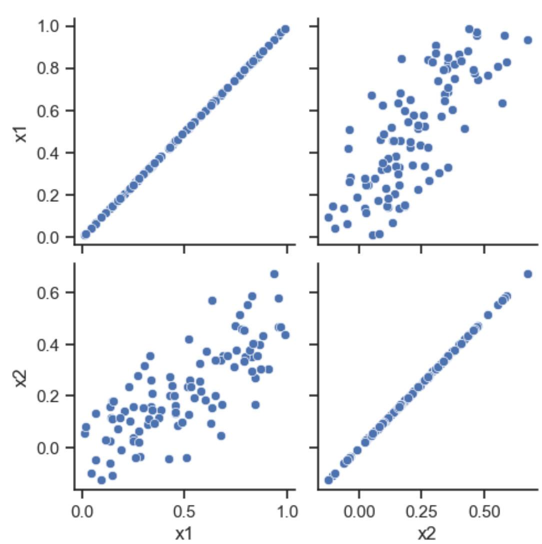
```
rng = np.random.default_rng(10)
x1 = rng.uniform(0, 1, size=100)
x2 = 0.5 * x1 + rng.normal(size=100) / 10
y = 2 + 2 * x1 + 0.3 * x2 + rng.normal(size=100)
```

The last line corresponds to creating a linear model in which y is a function of x1 and x2. Write out the form of the linear model. What are the regression coefficients? [2 pts]

The linear regression model takes the form $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, where the true regression coefficients are $\beta_0 = 2$, $\beta_1 = 2$, and $\beta_2 = 0.3$.

(b) What is the correlation between x1 and x2? Create a scatterplot displaying the relationship between the variables. [2 pts]

	x_1	x_2
x_1	1.000000	0.772324
x_2	0.772324	1.000000



The table above is correlation matrix between x1 and x2. Therefore, correlation coefficient between x1 and x2 is 0.772324. I also created a scatterplot showing the relationship between x1 and x2.

(c) Using this data, fit a least squares regression to predict y using x1 and x2. Describe the results obtained. What are $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$? How do these relate to the true β_0, β_1, and β_2? Can you reject the null hypothesis $H_0 : \beta_1 = 0$? How about the null hypothesis $H_0 : \beta_2 = 0$? [2 pts]

	coef	std err	t	$P > \|t\|$
intercept	1.9579	0.190	10.319	0.000
x_1	1.6154	0.527	3.065	0.003
x_2	0.9428	0.831	1.134	0.259

In this problem, I fitted a least squares regression model to predict y using x1 and x2.
If you see the table above, you can find that $\hat{\beta}_0 = 1.9579$, $\hat{\beta}_1 = 1.6164$, and $\hat{\beta}_2 = 0.9428$.
And the true regression coefficients are $\beta_0 = 2$, $\beta_1 = 2$, and $\beta_2 = 0.3$.
We found that $\hat{\beta}_2$ has a big difference with β_2.
Let's look at the p-values of each variable. Since the p-value of $\hat{\beta}_1$ is significantly small, which means x1 is highly correlated with y. So we can reject the null hypothesis $H_0 : \beta_1 = 0$.
On the other hand, we cannot reject the null hypothesis $H_0 : \beta_2 = 0$, because the p-value of $\hat{\beta}_2$ is significantly big.

(d) Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$? [2 pts]

	coef	std err	t	$P > \|t\|$
intercept	1.9371	0.189	10.242	0.0
x_1	2.0771	0.335	6.196	0.0

We can reject the null hypothesis $H_0 : \beta_1 = 0$ since the p-value corresponding to $\hat{\beta}_1$ is zero as you can see the table above.

(e) Now fit a least squares regression to predict y using only x2. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$? [2 pts]

	coef	std err	t	$P > \|t\|$
intercept	2.3239	0.154	15.124	0.0
x_2	2.9103	0.550	5.291	0.0

This least squares regression model takes the form of $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x2$.
And, we are able to reject the null hypothesis $H_0 : \beta_1 = 0$ since the p-value corresponding to $\hat{\beta}_1$ is zero as you can see the table above.

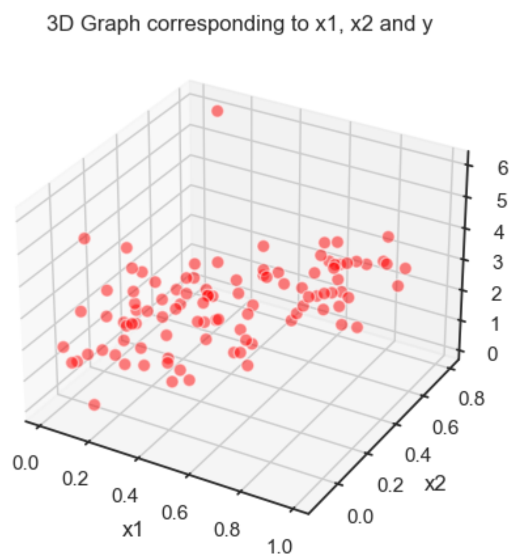
(f) Do the results obtained in (c)–(e) contradict each other? Explain your answer. [2 pts]

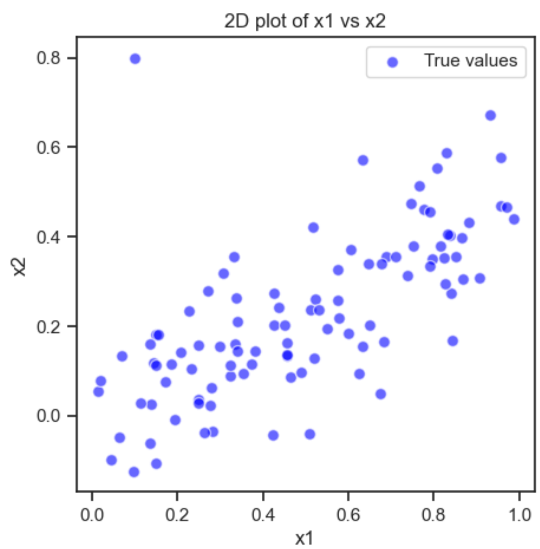
Yes, the results contradict each other. In the class, we learned that we should not use predictors that are highly correlated with each other. From Q5(b), we found that $x1$ and $x2$ are highly correlated. So, if we fit a least squares regression to predict y using both of $x1$ and $x2$, we consider $x2$ can be represented as $x1$ terms, since they are highly correlated.
In a nutshell, when we fit a least squares regression using only $x2$, it is significant. However, when we fit a least squares regression using both of $x1$ and $x2$, then $x2$ is not significant.
That's why the "Rejection availability" of null hypothesis of coefficient corresponding to $x2$ in the problem(c) and problem(e) were different.

(g) Suppose we obtain one additional observation, which was unfortunately mismeasured. We use the function np.concatenate() to add this additional observation to each of x1, x2, and y.

```
x1 = np.concatenate([x1, [0.1]])
x2 = np.concatenate([x2, [0.8]])
y = np.concatenate([y, [6]])
```

Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers. [2 pts]





On the previous problem (c), x1 was signnificantly correlated with y while x2 was not. However in problem(g), this reversed after adding one additional observation. Now, x2 is signnificantly correlated with y. On the other hand, x1 is less correlated with y, since the p-value corresponding to x1 is big. The R-squared slightly increased from 0.291 to 0.292, which means that this model is a little more predictive of data than previous model.
To find the outliers of the models, I made 3D plots above. Then, we can know that there are some outliers, outstanding points which is far away from dense points.
To find high-leverage of the models, I drew 2D plot of x1 vs x2. Then, we can find some strange points that far away from the dense points. Those are high-leverage points.

Q6. Solve ISLP Ch.4, Exercise #6 [6 pts]

Suppose we collect data for a group of students in a statistics class with variable X_1 = hours studied, X_2 = undergrad GPA, Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 0.9$.

(a) Estimate the probability that a student who studies for 30 hours and has an undergrad GPA of 3.6 gets an A in the class. [3 pts]

> We can get estimated probability of receiving an A by $\hat{p}(X) = \dfrac{e^{\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2 X_2}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2 X_2}}$.
> So, $\hat{p}(receive\, an\, A = Yes|X_1 = 30,\ X_2 = 3.6) = \dfrac{e^{-6+0.05\times30+0.9\times3.6}}{1+e^{-6+0.05\times30+0.9\times3.6}} = \frac{e^{-1.26}}{1+e^{-1.26}} \approx 0.221.$

(b) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class? [3 pts]

> We can get estimated probability of receiving an A by $\hat{p}(X) = \dfrac{e^{\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2 X_2}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2 X_2}}$.
> So, we need to find X_1 which satisfies $\hat{p}(X) = \dfrac{e^{-6+0.05X_1+0.9\times3.6}}{1+e^{-6+0.05X_1+0.9\times3.6}} = 0.5.$
> That is, $e^{0.05X_1-2.76} = 1. \quad \therefore X_1 = 55.2.$
> This means that student in part (a) need to study for 55 hours and 12 minutes to get A in the class.

Q7. Solve ISLP Ch.4, Exercise #7 [6 pts]

Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X, last year's percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was $\bar{X} = 10$, while the mean for those that didn't was $\bar{X} = 0$. In addition, the variance of X for these two sets of companies was $\hat{\sigma}^2 = 36$. Finally, 80% of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was $X = 4$ last year.

Hint: Recall that the density function for a normal random variable is $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$. You will need to use Bayes' theorem.

Let's remind Bayes theorem for discriminant analysis for this problem, which indicates

$$Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

where $f_k(x) = Pr(X = x|Y = k)$ is density for X in class k, and $\pi_k = Pr(Y = k)$ is prior probability for class k. There are given conditions that $\pi_{yes} = 0.8$, $\mu_{yes} = 10$, $\mu_{no} = 0$, $\hat{\sigma}^2_{yes} = \hat{\sigma}^2_{no} = 36$.

First, let's compute $f_{yes}(x)$, the density for X in class "yes".

$$f_{yes}(x) = \frac{1}{\sqrt{2\pi\sigma^2_{yes}}} e^{-(x-\mu_{yes})^2/2\sigma^2_{yes}}$$

$$\therefore\ f_{yes}(4) = \frac{1}{\sqrt{2\pi \times 36}} e^{-(4-10)^2/2\times 36} = \frac{1}{\sqrt{72\pi}} e^{-0.5}$$

Likewise, we can also get $f_{no}(x)$, the density for X in class "no".

$$f_{no}(4) = \frac{1}{\sqrt{2\pi \times 36}} e^{-(4-0)^2/2\times 36} = \frac{1}{\sqrt{72\pi}} e^{-2/9}$$

Finally, we can predict the probability $Pr(Y = yes|X = 4)$,

$$Pr(Y = yes|X = 4) = \frac{\pi_{yes} f_{yes}(4)}{\pi_{yes} f_{yes}(4) + \pi_{no} f_{no}(4)}$$

$$= \frac{0.8 \times e^{-0.5}}{0.8 \times e^{-0.5} + (1 - 0.8) \times e^{-2/9}} \approx 0.7519$$

Q8. Solve ISLP Ch.4, Exercise #12 [10 pts]

Suppose that you wish to classify an observation $X \in \mathbb{R}$ into apples and oranges. You fit a logistic regression model and find that

$$\widehat{Pr}(Y = \text{orange}|X = x) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)}.$$

Your friend fits a logistic regression model to the same data using the *softmax* formulation in (4.13), and find that

$$\widehat{Pr}(Y = \text{orange}|X = x) = \frac{\exp(\hat{\alpha}_{\text{orange}0} + \hat{\alpha}_{\text{orange}1} x)}{\exp(\hat{\alpha}_{\text{orange}0} + \hat{\alpha}_{\text{orange}1} x) + \exp(\hat{\alpha}_{\text{apple}0} + \hat{\alpha}_{\text{apple}1} x)}.$$

(a) What is the log odds of orange versus apple in your model? [2 pts]

The log odds of orange versus apple in my model is,

$$log\left[\frac{\hat{Pr}(Y = orange|X = x)}{1 - \hat{Pr}(Y = orange|X = x)}\right] = log\left[\frac{\frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)}}{1 - \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)}}\right]$$
$$= \hat{\beta}_0 + \hat{\beta}_1 x$$

(b) What is the log odds of orange versus apple in your friend's model? [2 pts]

The log odds ratio of orange versus apple in friend's model is,

$$log\left[\frac{\hat{Pr}(Y = orange|X = x)}{\hat{Pr}(Y = apple|X = x)}\right] = log\left[\frac{\frac{\exp(\hat{\alpha}_{\text{orange}0} + \hat{\alpha}_{\text{orange}1} x)}{\exp(\hat{\alpha}_{\text{orange}0} + \hat{\alpha}_{\text{orange}1} x) + \exp(\hat{\alpha}_{\text{apple}0} + \hat{\alpha}_{\text{apple}1} x)}}{\frac{\exp(\hat{\alpha}_{\text{apple}0} + \hat{\alpha}_{\text{apple}1} x)}{\exp(\hat{\alpha}_{\text{orange}0} + \hat{\alpha}_{\text{orange}1} x) + \exp(\hat{\alpha}_{\text{apple}0} + \hat{\alpha}_{\text{apple}1} x)}}\right]$$
$$= \hat{\alpha}_{orange0} + \hat{\alpha}_{orange1} x - \hat{\alpha}_{apple0} - \hat{\alpha}_{apple1} x$$
$$= (\hat{\alpha}_{orange0} - \hat{\alpha}_{apple0}) + (\hat{\alpha}_{orange1} - \hat{\alpha}_{apple1}) x$$

(c) Suppose that in your model, $\hat{\beta}_0 = 2$, $\hat{\beta}_1 = -1$. What are the coefficient estimates in your friend's model? Be as specific as possible. [2 pts]

$$\hat{\beta}_0 = \hat{\alpha}_{orange0} - \hat{\alpha}_{apple0} = 2$$
$$\hat{\beta}_1 = \hat{\alpha}_{orange1} - \hat{\alpha}_{apple1} = -1$$

However, we can not determine the exact value of estimated coefficients in friend's model. We only know that the estimated coefficients are satisfied with the above equation.

(d) Now suppose that you and your friend fit the same two models on a different data set. This time, your friend gets the coefficient estimates $\hat{\alpha}_{\text{orange0}} = 1.2$, $\hat{\alpha}_{\text{orange1}} = -2$, $\hat{\alpha}_{\text{apple0}} = 3$, $\hat{\alpha}_{\text{apple1}} = 0.6$. What are the coefficient estimates in your model? [2 pts]

$$\hat{\beta}_0 = \hat{\alpha}_{orange0} - \hat{\alpha}_{apple0} = 1.2 - 3 = -1.8$$
$$\hat{\beta}_1 = \hat{\alpha}_{orange1} - \hat{\alpha}_{apple1} = -2 - 0.6 = -2.6$$

(e) Finally, suppose that you apply both models from (d) to a data set with 2,000 test observations. What fraction of the time do you expect the predicted class labels from your model to agree with those from your friend's model? Explain your answer. [2 pts]

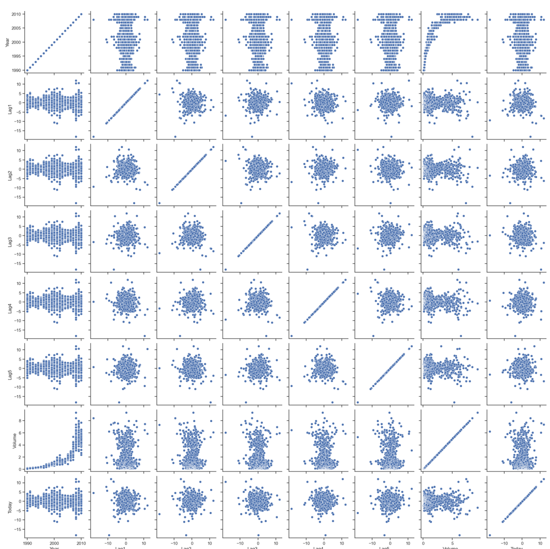
Their prediction would be identical. Softmax regression model is an extension by generalizing logistic regression model to apply to multi-class classification problems. In other words, logistic regression can be seen as a "case" where there are two classes specifically in soft max regression.

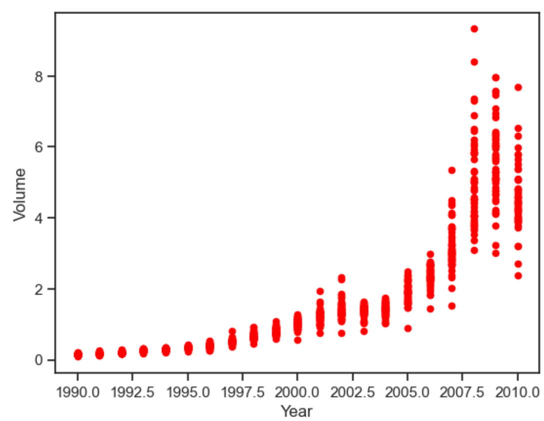
Q9. Solve ISLP Ch.4, Exercise #13 [14 pts]

This question should be answered using the Weekly data set, which is part of the ISLP package. This data is similar in nature to the Smarket data from this chapter's lab, except that it contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

(a) Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns? [1.5 pts]

	Year	Lag1	Lag2	Lag3	Lag4	Lag5	Volume	Today
Year	1.000000	-0.032289	-0.033390	-0.030006	-0.031128	-0.030519	0.841942	-0.032460
Lag1	-0.032289	1.000000	-0.074853	0.058636	-0.071274	-0.008183	-0.064951	-0.075032
Lag2	-0.033390	-0.074853	1.000000	-0.075721	0.058382	-0.072499	-0.085513	0.059167
Lag3	-0.030006	0.058636	-0.075721	1.000000	-0.075396	0.060657	-0.069288	-0.071244
Lag4	-0.031128	-0.071274	0.058382	-0.075396	1.000000	-0.075675	-0.061075	-0.007826
Lag5	-0.030519	-0.008183	-0.072499	0.060657	-0.075675	1.000000	-0.058517	0.011013
Volume	0.841942	-0.064951	-0.085513	-0.069288	-0.061075	-0.058517	1.000000	-0.033078
Today	-0.032460	-0.075032	0.059167	-0.071244	-0.007826	0.011013	-0.033078	1.000000





First, I made the correlation matrix corresponding to the Weekly data. As we can see the table above, Year and Volume are highly correlated. Therefore, I made a graph between them, then we can see that there are some correlations between 'Year' and 'Volume'.

(b) Use the full data to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones? [1.5 pts]

Lag2 appears to be statistically significant, because the p-value corresponding to Lag2 is significantly small. You can verify this, through my code.

(c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression. [1.5 pts]

	Truth	
Predicted	Down	Up
Down	54	48
Up	430	557

If we see the confusion matrix, our model correctly predicted that the Directions were ”Up” on 557 weeks and ”Down” on 54 weeks, for total 611 correct predictions. The overall fraction of correct prediction can also be calculated by $\frac{611}{1089} = 0.561$, while 1089 is the total number of observations.

(d) Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010). [1.5 pts]

	Truth	
Predicted	Down	Up
Down	9	5
Up	34	56

Overall fraction of correct prediction in the Logistic regression model : 0.625

(e) Repeat (d) using LDA. [1.5 pts]

	Truth	
Predicted	Down	Up
Down	9	5
Up	34	56

Overall fraction of correct prediction in the Logistic regression model : 0.625

(f) Repeat (d) using QDA. [1.5 pts]

Predicted	Truth Down	Up
Down	0	0
Up	43	61

Overall fraction of correct prediction in the QDA model : 0.5865384615384616

(g) Repeat (d) using KNN with $K = 1$. [1.5 pts]

Predicted	Truth Down	Up
Down	21	30
Up	22	31

Overall fraction of correct prediction in the KNN model with K=1 : 0.5

(h) Repeat (d) using naive Bayes. [1.5 pts]

Predicted	Truth Down	Up
Down	0	0
Up	43	61

Overall fraction of correct prediction in the Naive Bayes model : 0.5865384615384616

(i) Which of these methods appears to provide the best results on this data? [2 pts]

Through the process from (d) to (h), the logistic regression model and the LDA model showed the best results in this data. Their overall fractions of correct predictions were the same as 0.625.
In the case of QDA and naive Bayes models, they just predicted all the direction of the test data as "Up", so they couldn't predict any "Down". Therefore, I think it's not very good to use both models in this problem regardless of their fractions of correct predictions.

Q10. Exploratory Data Analysis with NYC Taxi Dataset [20 pts] ⌨

Please complete the exercises in the following Google Colab notebook: https://bit.ly/mldl23f-hw2-nyc-taxi and submit your .ipynb file.

You can see the solution through my code file.