1. (PCR and PLS, *10 pt* **Bonus**) Are the following sentences about principal component regression (PCR) and partial least square (PLS) True or False? Briefly justify your answer.

   (i) Both PCR and PLS come up with orthogonal features.

   (ii) Let $\boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)}, \cdots, \boldsymbol{Z}^{(p)}$ be the features obtained by PLS. For an intermediate $k < p$, we fit a regression model $\boldsymbol{Y}$ on $\boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)}, \cdots, \boldsymbol{Z}^{(k)}$, and obtain a predicted response $\hat{\boldsymbol{Y}}^{(k)}$ on the training set. Then $\hat{\boldsymbol{Y}}^{(k)}$ is orthogonal to the subsequent features $\boldsymbol{Z}^{(k+1)}, \boldsymbol{Z}^{(k+2)}, \cdots, \boldsymbol{Z}^{(p)}$. Therefore, to compute $\hat{\boldsymbol{Y}}^{(k+1)}$ (the predicted response based on $\boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)}, \cdots, \boldsymbol{Z}^{(k)}, \boldsymbol{Z}^{(k+1)}$), we can first regress $\boldsymbol{Y}$ on $\boldsymbol{Z}^{(k+1)}$ only and obtain $\hat{\boldsymbol{Y}}(\boldsymbol{Z}^{(k+1)})$ (the predicted response based on $\boldsymbol{Z}^{(k+1)}$ only), and then let

   $$\hat{\boldsymbol{Y}}^{(k+1)} \leftarrow \hat{\boldsymbol{Y}}^{(k)} + \hat{\boldsymbol{Y}}(\boldsymbol{Z}^{(k+1)}).$$

   (iii) The first feature from PLS is more predictive towards the response in the training set than that from PCR.

   (iv) In the procedures of constructing features from PCR and PLS, the earlier a feature is included in the regression model, the faster the training R^2 increases.

   (v) Using the features from PCR and PLS has lower training errors and test errors than those of the original linear model.

   Hint. Read more in Section 6.3 from the Textbook before getting started.

(i)
False.
Two principal components Z1 and Z2 of the Principal Component Analysis have the zero-correlation condition. In other words, the second principal component direction must be perpendicular, or perpendicular orthogonal, to the first principal component direction.
Partial Least Squares, use the annotated label to maximize inter-class variance. Principal components are pairwise orthogonal. Principal components are focus on maximize correlation.

2. (Logistic Regression, *10 pt*) Suppose we have observation pairs $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ where $\boldsymbol{X}_i \in \mathbb{R}^p$, $Y_i \in \{0, 1\}$.

   (i) Suppose $Y_i \sim$ **Binomial**(p_i) where $p_i \in [0, 1]$ is a parameter. Write down the probability mass function (PMF) of Y_i. What's the expectation of Y_i?

   (ii) In the terminology of generalized linear model (GLM), there is a link function $g : \mathbb{R} \to \mathbb{R}$ that establishes the relationship between the expectations of Y_i's and the linear combinations of $\boldsymbol{X}_i$'s

$$g(\mathbb{E}Y_i) = \boldsymbol{X}_i^T \boldsymbol{\beta} \quad (\forall 1 \leqslant i \leqslant n)$$

   where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unknown coefficient parameter for the linear part.

   Now for the Logistic regression problem, g is the **log-odds/logit** link

$$g(\mu) := \log\left(\frac{\mu}{1-\mu}\right). \quad (\mu \in [0, 1])$$

   First write down the log-likelihood function of Y_i in terms of parameter $\boldsymbol{\beta}$ given the observation $(\boldsymbol{X}_i, Y_i)$. Then write down the joint log-likelihood.

   Hint. In the PMF of Y_i, replace p_i by some quantities in terms of $\boldsymbol{X}_i^T\boldsymbol{\beta}$ through the link between $\mathbb{E}Y_i$ and $\boldsymbol{X}_i^T\boldsymbol{\beta}$.

(i)

$$P(x; p, n) = \binom{n}{x}(p)^x(1-p)^{n-x}$$
$$E(Y) = np$$

(ii)

(ii) $P(x_i) = P_Y(y_i = 1 | x_i) = \pi(x_i)$

$P_Y(y_i = 0 | x_i) = 1 - p(x_i)$

$\log \frac{p(x_i)}{1-p(x_i)} = \sum_{j=0}^{p} \beta_j x_{i,j}$

$\log \frac{p(x_i;\beta)}{1-p(x_i;\beta)} = x_i\beta$

$p(x_i;\beta) = \frac{e^{x_i\beta}}{1+e^{x_i\beta}}$

$1 - p(x_i;\beta) = \frac{1}{1+e^{x_i\beta}}$

log-likelihood

$l(\beta|x_i) = (1-y_i)\log[1-p(x_i;\beta)] + y_i \log p(x_i;\beta)$

$= \begin{cases} \log p(x_i;\beta) & \text{if } y_i = 1 \\ \log[1-p(x_i;\beta)] & \text{if } y_i = 0 \end{cases}$

Joint-log-likelihood

$L(\beta|x_1, \dots x_n) = \sum_{i=1}^{n} l_i(\beta|x_i)$

$= \sum_{i=1}^{n} (1-y_i)\log[1-p(x_i;\beta)] + y_i \log p(x_i;\beta)$

$= \sum_{i=1}^{n} y_i \log \frac{p(x_i;\beta)}{1-p(x_i;\beta)} + \log[1-p(x_i;\beta)]$

$= \sum_{i=1}^{n} y_i x_i\beta - \log(1+e^{x_i\beta})$

3. (Logistic Regression, Textbook 4.6, *15 pt*) Suppose we collect data for a group of students in a statistics class with variables X_1 = hours studied, X_2 = undergrad GPA and Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$.

   (a) Provide an interpretation of each coefficient in the model. Note that β_0 corresponds to an additional intercept in the model.

   (b) Estimate the probability that a student who studies for 40h and has an undergrad GPA of 3.5 gets an A in the class.

   (c) How many hours would the student in part (b) need to study to have a 50% chance of getting an A in the class.

(a)

β_1 is the coefficient of 'hours studied'. Exp(0.05) = 1.051271, meaning odds of earning an A are multiplied by 1.051271, with each 1 hour increase in 'hours studied'

β_2 is the coefficient of 'undergrad GPA'. Exp(1) = 2.718282, meaning odds of earning an A are multiplied by 2.718282, with each 1-unit increase in GPA.

(b)

$$p(x1 = 40, x2 = 3.5) = \frac{e^{-6+0.05*40+1*3.5}}{1+e^{-6+0.05*40+1*3.5}} = 0.3775407$$

(c)

$$p(x1 = X1, x2 = 3.5) = \frac{e^{-6+0.05*X1+1*3.5}}{1 + e^{-6+0.05*X1+1*3.5}} = 0.5$$

$$e^{-6+0.05*X1+1*3.5} = 1$$

$$e^{-6+0.05*X1+1*3.5} = 1$$

$$0.05 * X1 = 2.5$$

$$X1 = 50$$

He needs 50 hours study to reach the 50% chance of getting an A.

4. (LDA and QDA, Textbook 4.5, *20 pt*) We now examine the differences between linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA).

   (a) If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set?

   (b) If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set?

   (c) In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?

   (d) True or False: Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer.

(a)
If the Bayes decision boundary is linear, we may expect that LDA performs better on the training set and test set. Because the QDA will suffer from high variance without a corresponding decrease in bias.
(b)
If the Bays decision boundary is non-linear, we may expect the QDA will outperform the LDA on both training set and test set. Because QDA is more flexible and does not share a common covariance matrix we may expect a decrease in bias.
(c)
As sample size n increase, we may expect that QDA may outperform the LDA. Because as n gets large, reducing variance of classifier is not a major concern, or if the assumption of a common covariance matrix for the K classifiers clearly untenable. Besides, QDA is more flexible than LDA. Then, QDA may perform better than LDA.
(d)
False.
When the number of observations is small but with more predictors, QDA may lead to overfitting due to the flexibility of QDA. Therefore, when considering reducing the variance, it is better to use LDA.