Lee Jae Yeol
20308109
Assignment 1

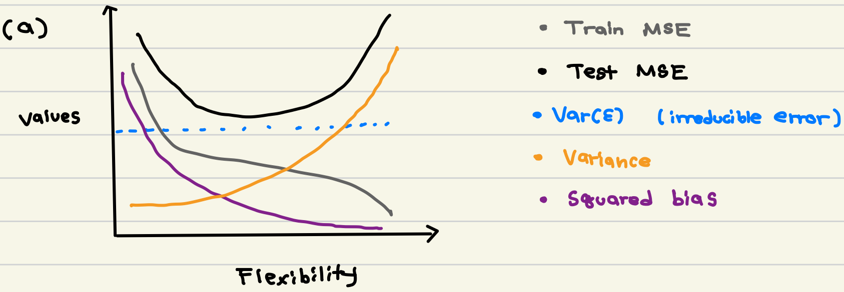
2-1) (a) Since the sample size is extremely large and the number of predictors is small, a more flexible model will have smaller residual errors and a lower chance of overfitting. Flexible model will have a better performance.

(b) Since the number of predictors is much bigger than the number of observations, a flexible model is prone to overfitting. Therefore, it is worse.

(c) If the predictors and response have a highly non-linear association, it is better to make the model flexible because inflexible model will have a very high bias, in other words, underfitting.
Due to strong non-linearity, overfitting won't happen easily.

(d) If the variance of the error terms is too high, making the model flexible will catch all the noises derived from the error term. So it is worse to use a flexible model.

2-3) (a)



(b) Squared bias will monotonically drop since it gets reduced as flexibility increases. It will catch all the noise from the training set. Similar for the Train MSE since the model will fit all the data.
In contrast, the Variance will increase monotonically with respect to flexibility because as flexibility of the model increases, using a different set of data will yield to a less accurate measurement. Test MSE is in the same manner with variance, except that it initially decreases to a certain level of flexibility, but as the model starts to overfit, it will increase. Test MSE is also always above the Var(ε) dotted line because Var(ε) is the minimum value for Test MSE.

2-8), 2-10) are done in .ipynb

3-1)

	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 3.4. *For the* Advertising *data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.*

The purpose of conducting hypothesis testing is to reject our null hypothesis Ho and ultimately concluding that our variables do affect our response variable which in this case is sales.

With null hypothesis $H_0^{(1)}: \beta_{TV} = 0$, $H_0^{(2)}: \beta_{radio} = 0$, $H_0^{(3)}: \beta_{newspaper} = 0$,

the p values < 0.0001 for variables TV and radio suggest that they are significant. On the other hand, pvalue of newspaper = 0.8599 means that the Ho can not be rejected. The variable newspaper is insignificant to our response variable.

3-5) $\hat{y}_i = x_i \hat{\beta}$, $\hat{\beta} = \dfrac{\Sigma x_i y_i}{\Sigma x_i^2}$

$$\hat{y}_i = x_i \cdot \frac{\Sigma x_j y_j}{\Sigma x_j^2} = \sum_{j=1}^{n} \frac{x_i x_j}{\sum_{j=1} x_j^2} \cdot y_j = \sum_{j=1}^{n} a_j y_j \qquad \therefore a_j = \frac{x_i x_j}{\sum_{j=1} x_j^2}$$

3-8)
3-9) are done in .ipynb

4-1) (4.2) $P(x) = \dfrac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$ (4.3) $\dfrac{P(x)}{1 - P(x)} = e^{\beta_0 + \beta_1 x}$

$$\frac{P(x)}{1-P(x)} = \frac{\dfrac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}{1 - \dfrac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}} = \frac{\dfrac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}{\dfrac{1}{1+e^{\beta_0+\beta_1 x}}} = e^{\beta_0+\beta_1 x}$$

and the equality holds.

$\therefore$ Logistic function representation and logit representation for logistic regression model are equivalent.

4-2) (4.12) $p_k(x) = \dfrac{\pi_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$

(4.13) $\delta_k(x) = x \cdot \dfrac{\mu_k}{\sigma^2} - \dfrac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$

We take log of $p_k(x)$ as log function increases monotonically.
Finding maximum of $\log p_k(x)$ will equal to finding max of $p_k(x)$.

$$\log p_k(x) = \log \pi_k + \log \frac{1}{\sqrt{2\pi\sigma^2}} + \left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right) - \log\left(\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)\right)$$

↓ ↓

These two terms are constant, independent of class k so do not regard.

$$\Rightarrow \log \pi_k - \frac{1}{2\sigma^2}\left(x^2 - 2x\mu_k + \mu_k^2\right) = -\frac{1}{2\sigma^2}x^2 + \frac{x \cdot \mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k$$

The first term is also independent of class k.
Therefore, we have the final result $x \cdot \dfrac{\mu_k}{\sigma^2} - \dfrac{\mu_k^2}{2\sigma^2} + \log \pi_k$
which is defined as the discriminant function $\delta_k(x)$.

4-3) QDA model, K classes, $X \sim N(\mu_k, \sigma_k^2)$, class specific covariance matrix.
The density function for one-dimensional normal distribution is given as

$$f_k(x) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \cdot \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)$$

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi\sigma_l^2}} \exp\left(-\frac{1}{2\sigma_l^2}(x-\mu_l)^2\right)}$$

Similar to the above question, considering the numerator of $\log p_k(x)$,

$$\Rightarrow \log \pi_k - \log \frac{1}{\sqrt{2\pi\sigma_k^2}} + \left(-\frac{1}{2\sigma_k^2}\left(x^2 - 2\mu_k x + \mu_k^2\right)\right)$$

This time, $-\dfrac{x^2}{2\sigma_k^2}$ can't be ignored since σ_k^2 is not a fixed term
but varies with respect to the specific class.
Therefore, the Bayes' classifier is not linear but quadratic.

4-10)
4-11) are done in .ipynb