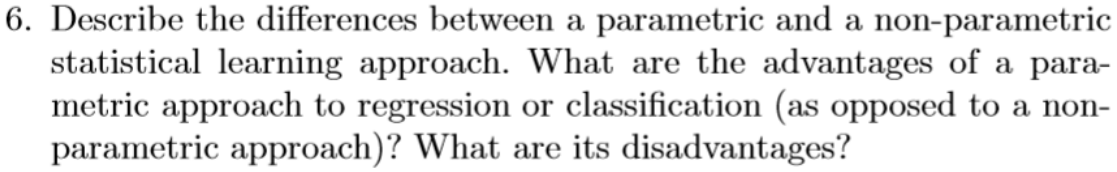
應數一 M072040019 梅瀚中



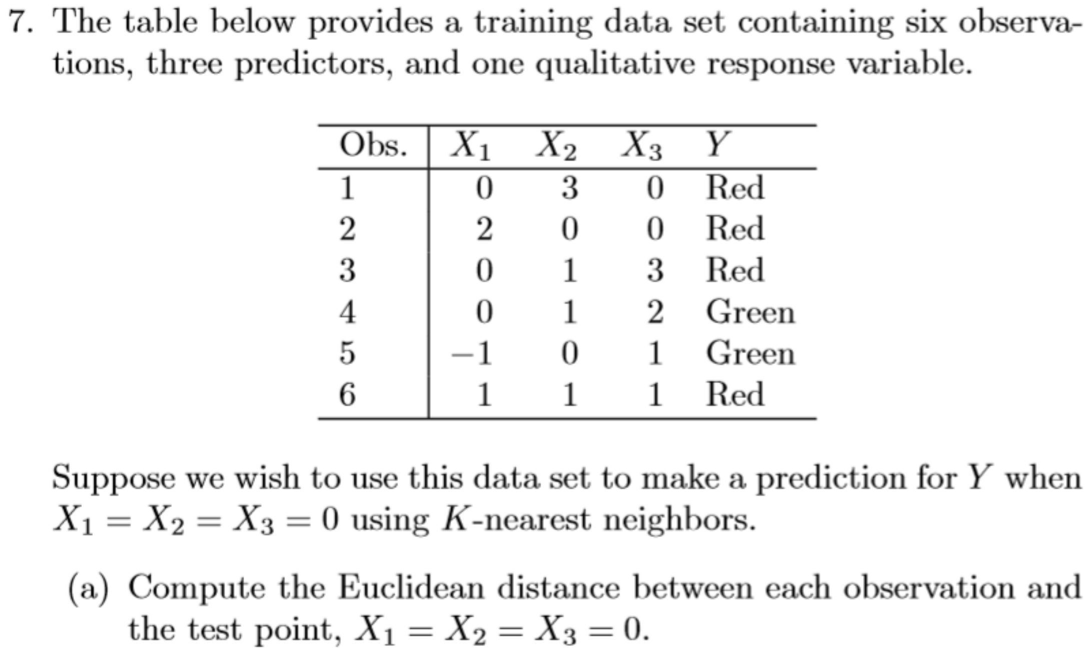
6. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a non-parametric approach)? What are its disadvantages?

有母數統計方法:假設母體為特定分配進行檢定。

無母數統計方法:所使用的統計量的抽樣分配通常與母體分配無關。

有母數統計方法大大簡化了估計 f(x)的問題。拿回歸跟分類問題來說,我們估計 β0,β1,...,βp,而不用估計整個 f(x)。而缺點是,假如我們選擇的模型如果與 f 真實的分配不符合,那麼我們的估計就會很差。

無母數統計方法沒有對 f(x)作任何的假設,根據 data 的形式,盡可能去接近數據點(例如:knn), 通過避免 f 的特定函數形式的假設，它有可能準確地為 f 提供更寬範圍的可能形狀。無母數方法存在一個主要缺點：由於它們不能減少將 f 估計為少量參數的問題，因此需要進行大量的觀測（遠遠超過參數方法通常需要的觀測值）。

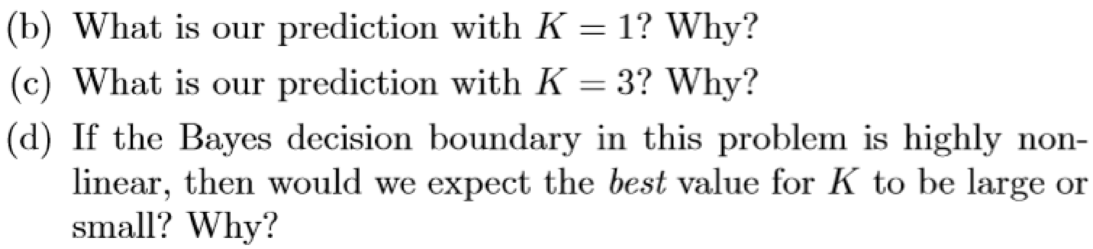


7. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

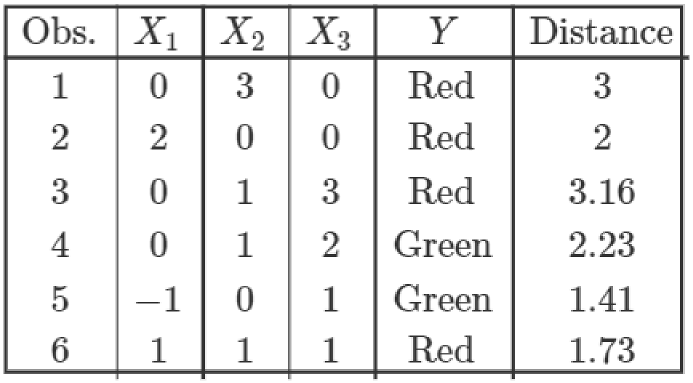


(b) What is our prediction with $K = 1$? Why?

(c) What is our prediction with $K = 3$? Why?

(d) If the Bayes decision boundary in this problem is highly non-linear, then would we expect the *best* value for K to be large or small? Why?

(a)

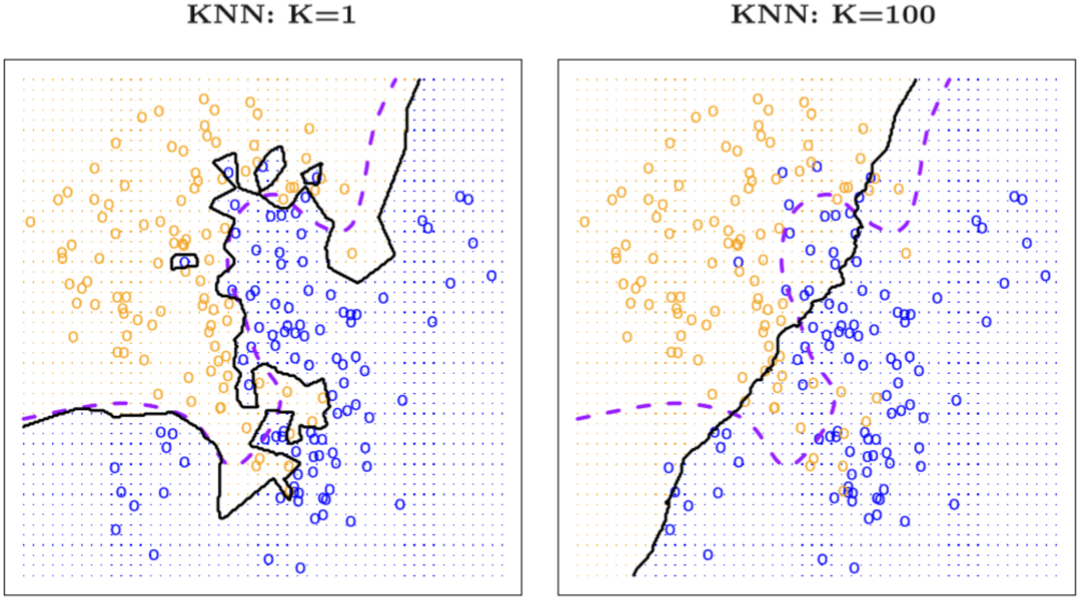


Obs.	X_1	X_2	X_3	Y	Distance
1	0	3	0	Red	3
2	2	0	0	Red	2
3	0	1	3	Red	3.16
4	0	1	2	Green	2.23
5	−1	0	1	Green	1.41
6	1	1	1	Red	1.73

(b) k=1 時,我們的 Y 是 Green,因為第 5 個觀察值離(0,0,0)最近,它是 green

(C)k=3 時,我們的 Y 是 Red,因為第 5,第 6,第 2 個觀察值離(0,0,0)最近,他們分別是 Green,red,red,然後取顏色多的,所以它是 y 是 red

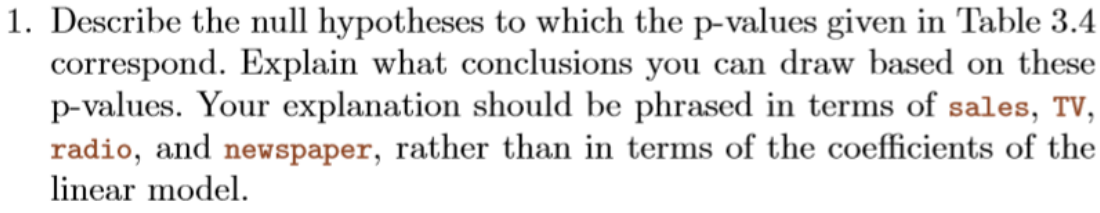
(d)



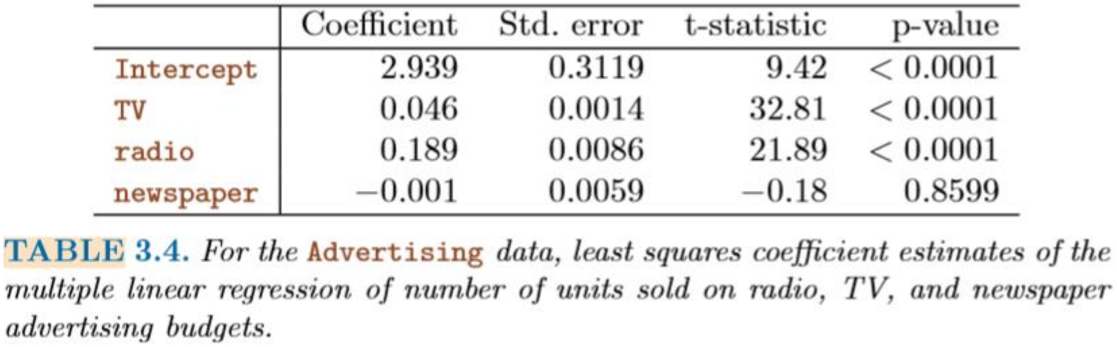
KNN: K=1

KNN: K=100

K 值取小的比較好,k 值取小的 bias 會比較小,variance 會比較大,適合在非線性的資料。



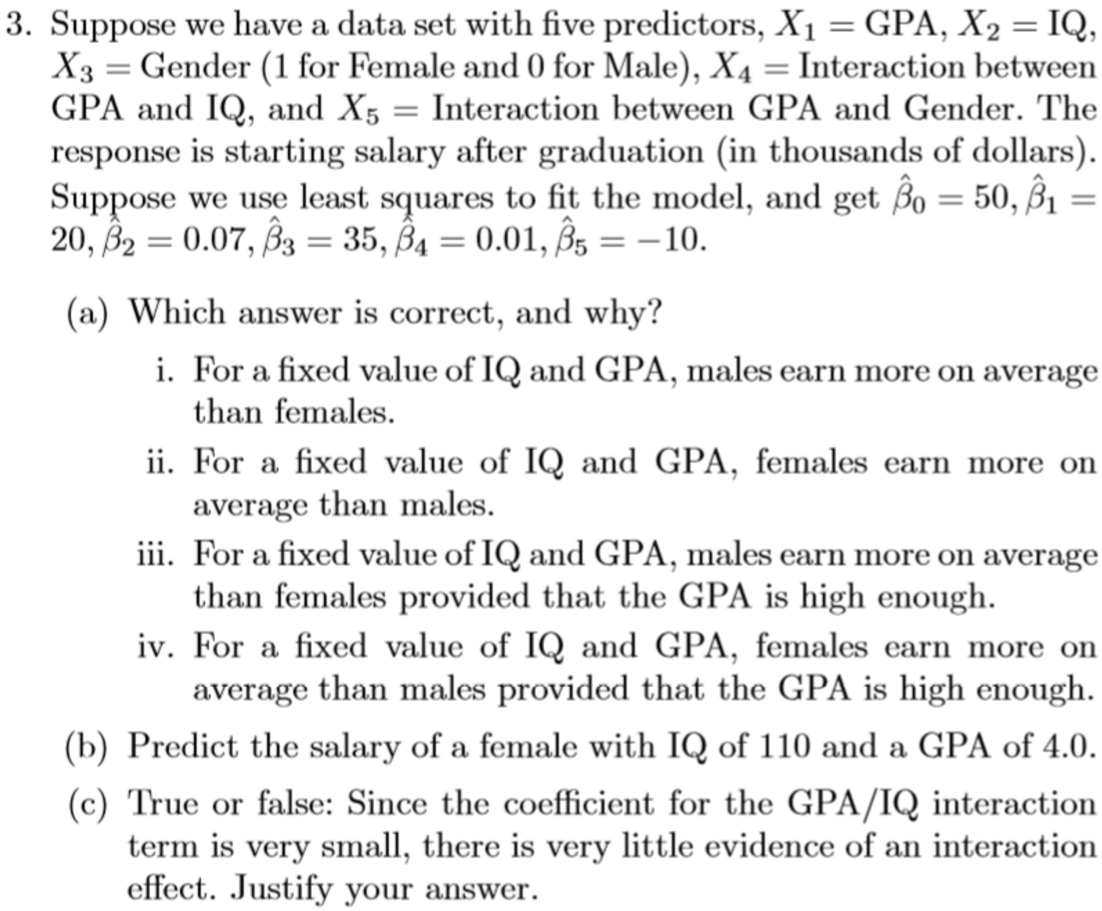
1. Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear model.



	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 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.*

由圖表可看出:TV,radio,的 p-value 顯著拒絕 H0:β_i=0,可認為 TV 跟 radio 對於 sales 是有影響。而 newspaper 的 p-value 不顯著拒絕 H0: β_i=0,所以我們沒有顯著證據證明 newspaper 對於 sales 有影響。



3. Suppose we have a data set with five predictors, $X_1 = \text{GPA}$, $X_2 = \text{IQ}$, $X_3 = \text{Gender}$ (1 for Female and 0 for Male), $X_4 = $ Interaction between GPA and IQ, and $X_5 = $ Interaction between GPA and Gender. 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?

      i. For a fixed value of IQ and GPA, males earn more on average than females.

      ii. For a fixed value of IQ and GPA, females earn more on average than males.

      iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

      iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

   (b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

   (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.

$$\hat{y} = 50 + 20GPA + 0.07IQ + 35Gender + 0.01GPA \times IQ - 10GPA \times Gender$$

(a)iii.是正確答案

(b) $\hat{y} = 85 + 40 + 7.7 + 4.4 = 137.1$

(c) false,我們應該去檢定 H0:β4=0 去看 P-value 才能決定。

8. This question involves the use of simple linear regression on the Auto data set.

(a) Use the lm() function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use the summary() function to print the results. Comment on the output. For example:

i. Is there a relationship between the predictor and the response?

```
library(ISLR)

## Warning: package 'ISLR' was built under R version 3.4.4

data(Auto)
lm.fit=lm(mpg ~ horsepower,data=Auto)
summary(lm.fit)

## 
## Call:
## lm(formula = mpg ~ horsepower, data = Auto)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -13.5710  -3.2592  -0.3435   2.7630  16.9240 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049 
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

p-value : <2e-16 ,我們可以顯著認為 horsepower 跟 mpg 有關係

ii. How strong is the relationship between the predictor and the response?

R-squared : 0.6059 代表用這個回歸模型 mpg 可以被 horsepower 解釋的變異有 60.59%

iii. Is the relationship between the predictor and the response positive or negative?

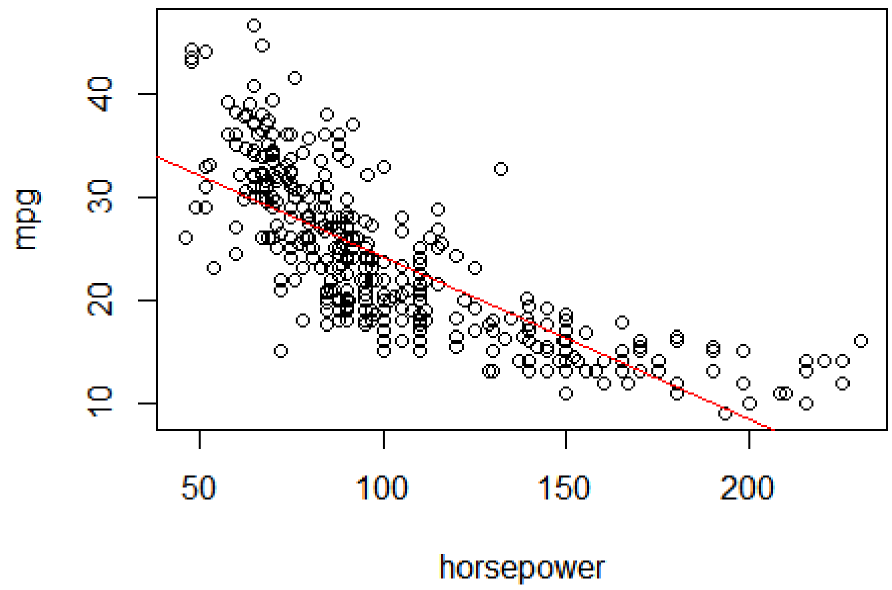
horsepower 的係數是-0.157845,所以 prefictor 和 response 的關係是負的

iv. What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals?

```
predict(lm.fit, data.frame(horsepower = 98), interval = "prediction")

##        fit     lwr      upr
## 1 24.46708 14.8094 34.12476

predict(lm.fit, data.frame(horsepower = 98), interval = "confidence")

##        fit      lwr      upr
## 1 24.46708 23.97308 24.96108
```

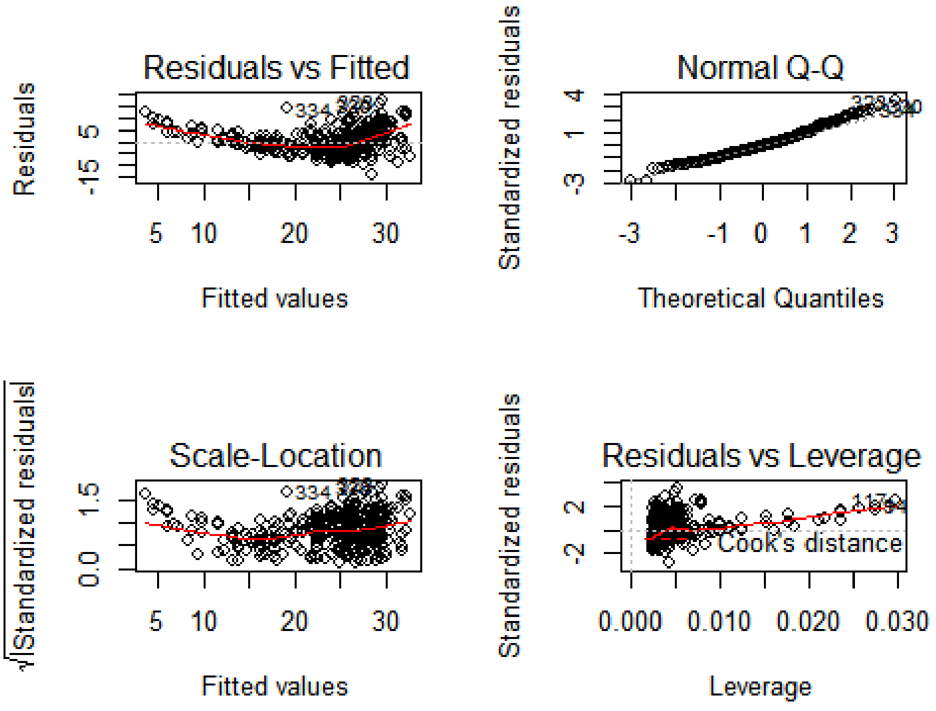
(b) Plot the response and the predictor. Use the abline() function to display the least squares regression line.

```
attach(Auto)
plot(horsepower,mpg)
abline(lm.fit,col="red")
```



(c) Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit

```
par(mfrow=c(2,2))
plot(lm.fit)
```



(1)從 Residuals vs Fitted 圖中可以看出 predictors 和 response 有一點非線性的趨勢 (2)從 Normal Q-Q 圖中得知殘差大致符合標準常態分佈 (3)從 Scale-Location 圖中大致得知取線周圍的點應該隨機分布,而圖中有一些點有微 outliers 的趨勢