2022 年春季

MIS 947/985實用業務分析

作業 1

· 滿分100分。請排版您的作業，另存為標題為“您的學生 ID_Homework_1.R”的 R 源代碼文件（例如 M024020001_Homework_1.R）。

· 請在04/21 11:59pm之前將您的代碼提交給 NSYSU Cyber University。不遲交。

· 不要在答案中使用任何循環。另請注意，您的代碼必須遵循課程中討論的建議編程和數據分析風格。

1. 請從“MASS”包中加載數據集“survey”。該數據集包含
阿德萊德大學 237 名統計學 I 學生的回答。考慮以下問題。

1.1。 [5 分]請從數據框中刪除任何 NA 的觀察結果，並創建學生練習頻率分佈的條形圖（Exer）。

1.2. [5 分]請將連續變量“Height”轉換成多組，分別為“150”、“160”、“170”、“180”和“>190”。然後創建一個交叉表，顯示學生練習 (Exer) 和新的分類變量“身高”之間的關係。

1.3. [10分]鍛煉頻率（Exer)與學生身高（Height)有關係嗎？使用任何統計方法來證明你的答案是正確的。

2. 請加載給定的“my_StudentsPerformance.csv”並考慮以下問題。

2.1。 [10 分]
請編寫一個 R 函數 my_summary(v)，它接受一個數字向量 v 併計算/返回匯總統計信息，包括平均值、標準差、最大值、最小值和中值。然後使用您的函數為數據集中的 3 個分數生成這些統計信息。不要使用內置的 summary()。

2.2. [10 分]
請添加一個新列“avg_score”，用於計算每個學生的數學、閱讀和寫作的平均分數。請在 R 和 SQL 中執行此操作。

2.3. [10 分]
使用方框和密度圖通過“avg_score”可視化“性別”。你認為性別與平均分有關嗎？使用任何描述性統計技術和/或統計方法（例如假設檢驗)來證明您的發現是正確的。你可以考慮

如果您願意，可以進行正態性測試。

2.4. [10 分]
使用任何統計方法來測試 parental_level_of_education 和 avg_score 之間是否存在任何關聯。

3. 請加載給定的數據集“diamonds.csv”並回答以下問題。

3.1 [10 分]考慮一系列關於“價格”與其他變量的雙變量分析。
具體來說，繪製數據並執行雙變量統計測試以了解變量之間的關係。 “克拉”和“切工”是否與“價格”相關聯？使用任何統計方法來證明你的答案是正確的。另請注意，您可能會考慮對“價格”進行任何有助於理解關聯或更好地預測“價格”的數據轉換。

3.2 [10 分]請使用 set.seed(1) 將數據集分為訓練集 (70%) 和測試集 (30%)。然後將連續變量重新縮放為從 0 到 1 的值而不進行集中。 （提示：您可以使用“caret”包，並且您應該使用相同的特徵轉換計劃重新調整訓練集和測試集。）

3.3 [5 分]編寫一個計算平均絕對誤差 (MAE) 的 R 函數，其定義為作為：

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

3.4 [5 分]用重新調整的訓練集建立一個通用線性模型，然後報告訓練和測試的 MAE（四捨五入到小數點後四位）。

3.5 [5 分]去除 p 值較高（> 0.05)的預測變量，然後建立一個新的通用線性模型。新模型在訓練和測試 MAE 方面的誤差是否更低？

3.6 [5 分]同樣，我們想要另一個新模型，它考慮所有雙向交互而不刪除任何預測變量。請報告培訓和測試 MAE。
新模型在訓練和測試 MAE 方面的誤差是否更低？這種具有更多參數的複雜模型能否改善預測？