從資料學習期中報告
第 1 組
組員：110304006 統計二 楊謹豪
108304016 統計四 江宏繹

資料集：Seoul Bike Sharing Demand Data(
https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand)

目次：

一、序言

本份報告關鍵模型結果及分析呈現於「三、模型預測表現和整體結論」，此階段的呈現基本為表格與敘述性分析，可快速理解本次研究之成果與結論。「二、資料簡述介紹與預計運用之模型」之部分是對於資料及整體模型的介紹，如希望在閱讀結論前對本研究的基本背景有更多了解，建議先閱讀此部份以對後續的研究結果有更多的了解。「四、各模型詳細參數及模型建立方法」之部分則是對於各模型建立的詳細介紹，如希望了解詳細模型運作過程可再行閱讀。

二、資料簡述介紹與預計運用之模型

A. 資料特色簡述

本份資料目標是希望藉由天氣資料及些許其他特徵，達成對於首爾內共享單車借用需求狀況的分析和預測。

此份資料內每筆資料皆為一個小時的資料，記錄時間為 2017/12/01 到 2018/11/30，共有此時間區段內 365 天全部 24 小時的資料，因此總計記錄了 8760 筆資料，且同時，本資料內通篇皆並無缺失值。

其中，可供於預測的變數有以下十三個：(1)日期 (2)小時 (3)氣溫 (4)濕度 (5)風速 (6)能見度 (7)露點溫度 (8)太陽輻射 (9)降雨量 (10)降雪量 (11)季節 (12)是否為節日 (13)該小時是否運作。

希望預測的變數為：首爾共享單車該小時借用數。

B. 使用模型特色與簡述

本次研究因為資料整體記錄是依據一個特定「時間軸」，同時，因期望能對於各變數的影響有更深入的研究，我們預計使用以下三類模型進行研究和預測：

1. Linear Regression：

此類模型目標是以直線做擬合以找出最佳的直線。

優勢：

I. 容易操作和建立
II. 在詮釋變數重要
III. 分析變數特色上清晰簡易。

劣勢：

I. 有較嚴謹的前提假設
II. 在最後模型預測度上較不精確。

詳細使用模型：

I. Linear Regression：

此為最基本且經典的線性迴歸模型，在使用與計算上非常簡單。

II. Lasso Regression：

此處在計算上加上了額外的修正，此可避免 Linear Regression 出現僅是符合此份資料而不具廣泛應用性的狀況，也就是我們所稱的過度擬合，同時，它還能夠幫忙讓某些不太重要的變數的係數變成 0，換言之，便是篩選掉不重要的變數，減少我們花時間篩選變數的時間，但因為具修正的效果，所以通常預測表現會較糟。

III. Ridge Regression：

此處與上方的 Lasso 相似，也是加入了額外的修正以避免上述曾言的過度擬合出現，然而，它並不具備篩選變數的功能，對於不重要之變數僅是降低其係數，也就是將其影響降低，相似的，Ridge Regression 的表現在大多時候也會因避免過度擬合而比較低一些。

2. GLM：

此類模型可被視為靈活的 Linear Regression，它將可讓原先的 Linear Regression 具備其他的隨機變數分布。

優勢：

I. 具有更高的彈性
II. 能一定程度的對模型做出解釋
III. 非常可能能做出比 Linear Regression 更精準的預測
IV. 能加入時間序列的考量。

劣勢：

I. 隨機變數選擇需注意
II. 對時間序列資料並非專長。

詳細使用模型：

I. Poisson GLM：

此處因目標是分析共享單車的借用「數量」，因此在此資料不可能出現非整數數值的情況底下，且目標為計算 Count 的概念下，在理論上我們能利用 GLM(Generalized Linear Model)中，以計算此類資料為主的 Poisson Regression 來幫助我們做出更好的擬合。

II. Poisson GLM with Lag：

此模型的特色便是在上方所稱的 Poisson GLM 的基礎上，配合加入前一筆資料的共享單車借用數，達成借用前方資料的歷史

對本筆資料的借用數做預測的目標，此作法可使得其同一時間兼備 Poisson GLM 原先的優勢，又能讓其展現我們本段開頭所言希望納入考量的「時間軸」問題。

3. Time Series：

此類模型專長是處理時間序列的資料，也就是對與時間有連續性相關的資料進行優良的預測。

優勢：

I. 對時間序列資料的預測效果的理論最佳。
II. Time Series 內的模型經常可綜合使用，具一定彈性。

劣勢：

I. 建立時需要考慮的參數眾多
II. 在解釋上較複雜一些。

詳細使用模型：

I. ARIMA：

ARIMA 事實上是整合了 AR(Autoregressive Model)和 MA(Moving Average Model)兩者，另外再藉由特定的差分來保持穩定性。

AR 模型是利用自身前方的數值做預測，換言之，在此例中只需要給予模型共享單車借用數就可以完成利用前方的借用數對後方的資料進行預測。MA 模型則也是僅僅需要以自身變數即可協助預測，但相較 AR 模型，其利用的並非是使用上幾筆資料，而是參考前方數據後依據特定算法所得的數值，兩者各有優勢，而融合後的 ARIMA 模型便是能夠盡力通併兩者優點的模型。

此模型的最大特色便是其參考前段資料以用於預測的特性，能使得我們對於有連續性時間性質資料有著優秀的預測。

II. ARMAX：

ARMAX 事實上便是建立在了 ARIMA 上，在加上了外生變數 (eXogenous Variables)，這樣的行為讓我們能夠對於原先僅利用自身迴歸所無法解釋的部份再做出進一步的更好解釋，讓我們能夠更精確對模型做出解釋以及預測。

三、模型預測表現和整體結論

A. 模型預測結果

在對各模型皆進行全面的優化及分析後，我們選擇藉由 MSE 對表現評估。在此簡介一下 MSE，MSE 是一種對於模型預測數值表現評估的優秀指標，全名為 Mean Square Error，代表的是模型預測值與和真實值相差的平均，因此如果值越低，就代表預測的結果跟真實的值差距越小，也就代表模型整體的表現更佳。

以下表格即為模型與其對應的 MSE 和依據 MSE 越小越前做成的表現排名：

Models	MSE	表現排名
Linear Regression	125435.829	5
Lasso Linear Regression	163059.694	7
Ridge Linear Regression	125522.115	6
Poisson GLM	91765.618	4
Poisson GLM with Lag	32904.325	1
ARIMA	67819.954	3
ARMAX	35586.125	2

B. 模型呈現之資訊與分析

首先，在 Linear Regression 系列的模型上，其預測的表現並未如其他兩種佳，但就如上述曾言的，若是要藉由模型本身來進一步對於資料內的各個變數進行解釋，他們能夠幫助我們對不同的變數所造成的影響做出更好的解釋，以下幾點即為我們可從 Linear Regression 系列的模型所挖掘出的資訊以及依據所得資訊可以對該現象做的推測：

1. 每日特定高峰及低峰：

七到十點以及十七點到二十二點的迴歸係數相較其他小時的的迴歸係數來的高上非常多，換言之，這幾個小時對於共享單車數量的有相當高的正面貢獻和影響，藉此我們可以確定此幾個區段的借用量會是每日高峰；相反地，三點至六點的迴歸係數不僅小於零，更是較其他小時來的低上非常多，因此我們可藉此得知此區段的共享單車借用量為明顯的低峰期。

此現象的出現可以從兩個面向推測，高峰的部份非常有可能是因為工作時段上下班人潮的緣故，因為出現高峰的二時間段便是大部分勞工的上下班時間，相對的，低峰也非常好的能發現該時段是絕大多數人的睡眠時段，也非常合理的鮮少友人會在該時段借用共

享單車。

2. 一點與十五點對預測的不穩定性：

一點與十五點在模型中的迴歸係數皆未通過不等於 0 的假設檢定，換言之，此兩個小時對於共享單車借用量的影響並無特定正向或負向的影響，此情形可以幫忙我們了解這兩個小時的借用狀況可能在一年內的狀況相當不一定，可能出現截然不同的高峰或是低峰，需要費更多心思多加注意這兩個時間的不穩定。

此現象的出現推測非常有可能是這兩個時間段有特殊的波動情形，舉例而言，如在六日的假日，人們凌晨一點仍在外的機會會比平日高上許多，因此一點的借用量很有可能會有不確定的浮動性，相似的，下午十五點也有可能會因為六日的休息在外出遊而使用共享單車，造成比平日更高的借用量，然而詳細的原因值得我們去進一步探討，甚者此不確定因素相當有可能幫助我們對整體的借用狀況有全新的認識。

3. 秋天對於借用量的影響明顯高於其他三個季節：

秋天的迴歸係數甚高呈現了相較於其餘三季，人們更加傾向在此季節使用共享單車，然而，此現象的出現難以輕易確定其原因，因如要以天氣狀況作為評估，但其餘天氣變數理應以能做出解釋，因此可能性出現在另外一個方面，也就是根據資料所記錄時間，可發現本篇資料是自冬天記錄至隔年的秋天，所以秋天更高的借用量非常有可能象徵者使用者隨著時間成長，因此較靠後的秋天表現出了相較其他三季更高的共享單車借用量。

4. 節日對於借用量的明顯負面影響：

根據模型可以發現，節日的迴歸係數明顯為負而且相當小，換言之，若該小時所在的日期屬於某個節日，則它會比預期狀況更加少人選擇借用共享單車。

此狀況背後原因推測必須先釐清此處所稱之節日不包括六日，僅有大型節日如春節等，因此此情形相當有可能是因為首爾住戶有不低的比例來自外地，於大型慶祝節日可能選擇回到故鄉，因此降低了使用者的數量。

5. 氣溫和風速對於借用量的高度影響：

從模型中可見，氣溫對借用量有不低的正向影響，相反地，風速則有著顯著的負面影響。

此現象的可能推測是共享單車因為在騎乘時是完全暴露在戶外空間

中，如果氣溫太低，將會讓使用者在騎乘時全身發冷，此概念也可以套用在風速上，因為若是風速過高，那在騎乘共享單車時必然會感受到一般移動更高的寒冷，且高風速亦會造成騎乘時的搖晃等問題，因此此現象的產生與常理有高度的符合。

上述便是我們可從 Linear Regression 系列的資料中發現的特色，從其提供的大量資訊便可發現使用此類模型所帶來的豐厚價值。

在對 Linear Regression 的部份分析完後，可將 GLM 與 Time Series 的部份整合在一起進行分析，從此兩類中的四個模型，我們能發現到以下幾點特色和我們能從中做的一些推測：

1. Poisson GLM 類的模型有明確優秀的效果：

   從純粹的 Poisson GLM 和有 Lag 的兩者都能發現到其相較原先的 Linear Regression 都有更低的 MSE，這點讓我們確定了以計數的想法對此份資料進行模型的擬合是一個正確的方法。

   此現象的發生並非不能想像，事實上 Poisson 分布本身就一定程度的表現出了在某一段時間某個事件發生的次數，在這次的模型做完後更是能確認其與我們資料完美契合。

2. Time Series 類的模型證明了此資料內時間序列的重要性：

   與上一點相對的，我們能夠發現到 Time Series 類型的模型也能表現出相當優異的表現，甚至純粹以 Time Series 概念建立的 ARIMA 的 MSE 更比 Poisson GLM 的 MSE 還低，由此可見「時間」這個概念在此份資料中明確的貫穿，將時間納入考量為必要之事。

   此情形的發生事實上也非常符合邏輯，因為確實前幾小時若有大量的借用數，那代表這幾個小時內必然具有什麼吸引使用者的特徵，因此若是能參考前幾小時的借用數必然能夠使我們對於此小時的借用數有良好的預測。

3. 此資料優良的預測模型需要同時兼備時間序列的概念和其他提供之變數：

   根據上面兩點以及我們整體的評估可以確認，Poisson GLM 利用其他變數的概念和 Time Series 中時間序列概念間某程度具有關連，才有可能在缺乏一者的情況下仍有不錯的表現。

   此階段將提出一個想法來推斷此二概念之間的關係。

   先從時間序列方看起，事實上，僅有時間序列的弱點便是因為它只參考了過去的借用量，因此難以對於該小時突然其來改變的天

氣因素，如我們前述在 Linear Regression 中確認其重要性的氣溫和風速，或是高峰時段的起始，如六點前幾個小時借用數皆較低，但六點開始卻有大量上班上課人潮帶來的借用量，這兩者因素造成了時間序列無法進行精確的察覺與預測，必須有另一者協助。

相對的，僅參考提供之變數會造就的問題來自於變數天生就不可能完美解釋全部所造就的局限性，也就是或許該連續時間段內有著來自其他變數的影響，這時時間序列可以讓我們藉由參考前幾小時的資料來補足這些原先變數難以觀測到的影響。

綜上所述，若是我們希望能夠完成一個能夠完美對於此資料進行預測者，唯有在兩者兼備，使其能互相補足另外一者所能解釋部份的情況下，才有可能達成目標。

4. Poisson GLM with Lag 和 ARMAX 有類似的概念且符合上述觀點：

在討論完何種類型的模型能達到最佳解後，我們必須將焦點放到 Poisson GLM with Lag 和 ARMAX 上，事實上，仔細思考就會發現到此二者在某程度上是高度類似的，因為 Poisson GLM with Lag 的邏輯是在 Poisson GLM 的模型之上，加入類似 AR 模型之中直接參考前項結果的概念，換言之便是加入在利用基本變數的前提下，加上時間序列的概念；相反地，ARMAX 便是在以 ARIMA 這個完全僅以時間序列為參考的模型上，加上了外生變數，也就其他可供參考的變數。

此情形在兩者的 MSE 上也出現了共通，從上方的表格可見，他們兩者最終產出的 MSE 相當接近，這進一步讓我們更有信心的說他們二者在邏輯上具有一定程度的重疊性，那個微小的差距非常有可能來自於 Poisson GLM with Lag 使用的是更加適合本份資料的 Poisson 分布而非是 ARMAX 內的常態分布，換言之就是，我們在此四模型中剖析所想的第一點之 Poisson GLM 類的模型有明確優秀效果的想法也可在此被以不同方法證明。

而更重要且特殊之處在於，他們兩者是在確實符合我們第三點的「需同時兼備時間序列和其他參考變數」要求之情況下，展現出了不相上下的第一及第二名的表現，這也變相的確定了我們的推測基本無誤，若要達成精確的目標，綜合此二概念將為關鍵。

5. Poisson GLM with Lag 和 ARMAX 有相同的劣勢：

然而，此二優秀的模型也並非毫無弱點，相較於直接性的 Poisson GLM 或是更加簡易的 Linear Regression 系列，此二者皆必須要取得前幾小時的資料才可以進行更加精確的預測，若是希望對某一小時進行突然的預測將無法達成，屬於是此資料在追求高準確預

測結果時必須面對到的限制。

因此，整合 Linear Regression, Poisson GLM 和 Time Series，我們能得到的結論便如上方統整，整體來說，我們成功的確認了對本資料來說最完美的預測模型架構與實際模型，同時的也對於可能影響的變數進行了進一步的探究，非常完美的對於這份資料做出精確的剖析。

四、各模型詳細參數及模型建立方法

A. 資料前處理

在開始應用模型前，我們對資料進行了以下處理：

1. 剔除該小時並未運作之資料：

根據資料介紹，我們可以發現當「該小時是否運作」的變數內值若為”No”，則該小時的共享單車借用數量就將是 0，此現象也相當合理因為當該小時系統不進行運作，那自然借用量就應該為 0，也因此，凡是該變數數值為”No”者，我們便會直接將其剔除，以避免其影響後續模型的擬合。

2. 剔除露點溫度（Dew Point Temperature）：

根據 Correlation Matrix，我們會發現 Temperature 和 Dew Point Temperature 的相關係數來到 0.91，換言之，此兩個預測變數有強烈的共線性問題，若是直接將他們兩者用於模型的擬合，將可能會產生兩者分別可對預測具貢獻，但共同存在將皆無法有顯著影響的問題。

因此，在此處在將二者分別加入線性迴歸模型以及評估分別與預測的共享單車借用數量的相關係數後，我們發現，溫度在模型中的表現會明顯優於露點溫度且溫度與共享單車借用數量的相關係數更高，所以最終我們選擇捨棄露點溫度這個變數。

3. 「季節」、「是否為節日」與「小時」轉換為虛擬變數：

根據資料探索的結果，我們會發現「季節」和「是否為節日兩者」是屬於類別類型的變數，無法直接將他們放入模型內協助整體模型的建立和評估，因此，我們將其轉換為虛擬變數，舉例而言，季節共有四季，我們便轉而創立三個新的變數，並限制其內只能填入 1 和 0，同時選擇將在資料探索階段所發現共享單車平均借用數最低的冬季作為基準，將季節內原為冬季的資料之轉而創建的三行

分別填入 0, 0, 0、春季填入 1, 0, 0、夏季填入 0, 1, 0 以及秋季填入 0, 0, 1，此行為便成功的將原來的類別變數轉換為數字型的虛擬變數，讓我們可以將其使用在模型當中，同時，因非節日時間較多，我們選擇將非節日設為基準進行轉換。

然而較為特別者為小時這個變數，此變數雖為數字呈現，然而因各小時之間實值代表了一個特定時間段，因此其也為類別變數，同時，此處最終我們決議不選擇特定小時為基準，而是將它直接轉換為虛擬變數，創造 24 個虛擬變數，原因是來自於小時實值上難以找出一個真實的基準點，並且 24 與 23 個變數的差異極小，若設立之基準不對還可能會產生模型上的不穩定，因此選擇了直接的全數轉換。

4. 捨棄日期及運作與否此二變數：

運作與否之捨棄理由如資料前處理一部分所言，我們已經刪除了值為”No”的資料，因此剩餘全為”Yes”的情況下，並無保留之必要性。

而對於日期來說，捨棄的理由有二，第一是在實作面上它必定屬於類型變數，因此如果要將它應用在模型內，必定是需要使它轉換成虛擬變數，然而若真的如此進行，那我們即便在有基準的情形下仍然會額外得到 364 個虛擬變數，此種狀況不僅僅會造成整體模型計算上的負擔，同時更有可能造成模型的偏誤。第二是在邏輯面上若是我們將他轉換做虛擬變數保留，那我們會難以解釋其存在的意義，因為每個虛擬變數都僅有 24 格為 1，此景會造成最終模型的難以解釋與複雜，更同時的在預測未來資料時，在邏輯上也相當難判定不同年但同月同日是否要在該虛擬變數上轉為 1。

因此在考量上述理由後，我們決議直接將日期這個變數直接捨去。

B. Linear Regression

我們在對於 Linear Regression 類型的模型正式開始研究前，提前做了兩件事：

1. 10 Folds Cross Validation：

為了保證本次研究的嚴謹性，我們選擇使用 k-folds cross validation 來保證過程中能減少因過度擬合或是 test data 選擇而出現的誤差，此次選擇 k 為 10，每次選擇 9 folds 作為 train data 來訓練模型，1 folds 作為 test data 來驗證模型效果，共進行十次，最終產

出的 MSE 為十次結果的平均。

2. Power Transformation：

為了盡力符合 Linear Regression 中 y 符合 normal distribution 的假設，我們選擇對 y 做 power transformation，並且使用 Shapiro-Wilk test 和 D'Agostino's K-squared test 來進行檢定，以確認是否轉換已使我們希望預測的共享單車借用數轉換為符合 Normal distribution 的分佈的狀況。

然而，在測試後我們發現，即便是使用最可能達成的 box-cox transformation 亦無法使其通過檢定，且在轉換回原值上會出現錯誤，因此，本階段最終選擇效果相當接近 box-cox 結果的 1/3 對 y 進行 power transformation，並在預測階段將其轉換回原先數值進行 MSE 計算，使其雖無法達致完美，但仍能相較原先狀況更加符合假設。

以下便是模型的詳細參數使用的進行過程補述：

1. Linear Regression：

此處在將前處理資料放入模型中進行擬合後會發現，「能見度」與「太陽輻射」此二變數之係數並未通過不為 0 的檢定，在將其分別以及同時移除後再次進行模型擬合後會發現，MSE 幾乎沒有發生改變，因此，我們選擇將此二變數直接移除。

2. Lasso Regression：

此處關鍵在於，Lasso Regression 在擬合的公式上為了達成修正的目標，在最小平方法的公式中除了原先的$(\sum_{i=1}^{n} y_i - \sum_j x_{ij}\beta_j)^2$外，還加上了$\lambda \sum_{j=1}^{p} |\beta_j|$，而其中的關鍵是，我們必須找出$\lambda$為多少的時候能夠擬合出最佳的模型，因此，此處使用了 Grid Search 來協助尋找，它能夠遍歷一個範圍內的結果，以此找出最佳的參數選擇，因此，在找完後，我們能發現到最適切的λ為 0.1，爾後，我們便以此為參數對整體模型進行了擬合，並得到了前段所展示的 MSE。

3. Ridge Regression：

此處的情形與 Lasso Regression 相似，它也是在最小平方法的公式上除原先的$(\sum_{i=1}^{n} y_i - \sum_j x_{ij}\beta_j)^2$外，加上了額外的修正，然而該修正的公式為$\lambda \sum_{j=1}^{p} {\beta_j}^2$，從絕對值改成了平方，因此$\lambda$也必須要重新尋找，此處採用與 Lasso 相同的方法，利用 Grid Search 來遍歷尋找，最

後我們能發現到最佳的解是 9.9，爾後，我們便以此為參數對整體模型進行了擬合，並得到了前段所展示的 MSE。

C. GLM

在開始使用，我們必須確認和釐清兩件事：

進行首先，因為本次我們 GLM 系列選用的模型是 Poisson GLM，所以為了盡力符合其假設的情況下，我們必須要驗證我們的 y，也就是共享單車借用數，符合 Poisson Distribution，然而在檢驗過後我們會發先，它事實上並不完全符合 Poisson Distribution 中平均數等於變異數的假設，因此我們實際上也曾經嘗試過對於假設更加寬鬆但也是為了此類計數問題而誕生之 GLM 類型模型的 Negative Binominal GLM，然而，其 MSE 相較 Poisson GLM 明顯高了非常巨大的數值，因此即便 Poisson GLM 在假設上無法完美符合，於是我們仍採取 Poisson GLM 為本次研究的在 GLM 領域上的模型。

次來，與 Linear Regression 相似的，在此我們也採用了與上方完全相同的 10 Folds Cross Validation，以保證我們所作出的結果能夠達成減少因過度擬合或是 test data 選擇而出現誤差的目標。

1. Poisson GLM

此處的 Poisson GLM 並沒有什麼特別需要篩選之變數，僅需要將在資料前處理處修正完成的資料代入模型內進行擬合，便會發現到所有變數在此處皆有通過其係數不為 0 的假設，因此在此處將全數保留，完成對 Poisson GLM 模型的建立。

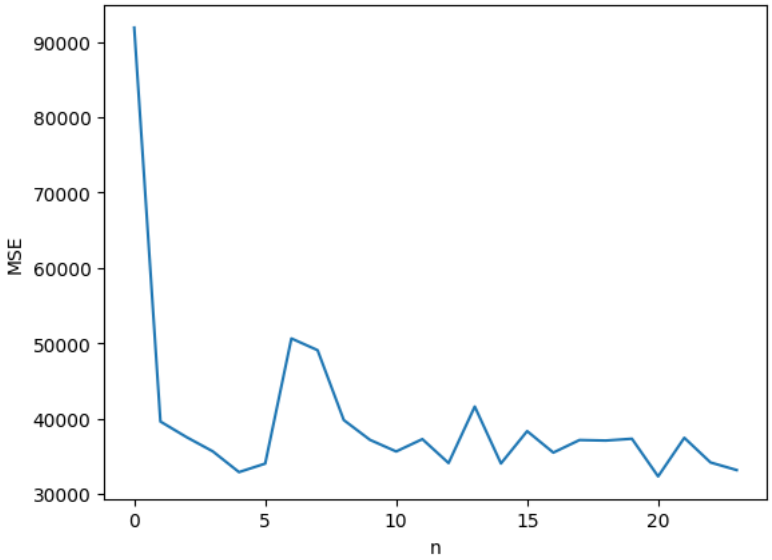
2. Poisson GLM with Lag

Poisson GLM with Lag 在本質上與 Poisson GLM 基本相同，差別之重點僅有在「Lag」上，此處創造 Lag 的方法與 AR(Autoregressive Model)相當相似，我們是藉由將 y，也就是共享單車借用數皆向後推移 n 格來產生第 n 個 Lag 值，因此，若是此出我們希望多出三個 Lag 產生的變數，那便是要產生第 1、第 2 和第 3 個 Lag 值，也就是分別讓共享單車的借用量都向後推移一格、二格和三格。

下一階段，我們要做的事情便是希望篩選要多出幾個 Lag 產生

的變數，理論上而言，越多的 Lag 變數通常能帶來更好的預測結果，也就是更低的 MSE，然而，這也象徵這他會更加複雜、難以解釋，而且更有可能出現過度擬合之類的問題，因此取得兩者之間的平衡成為了首要之務。

所以本階段對從未加 Lag 變數到加到 24 個 Lag 變數，換言之就是該小時前一天內的所有小時，進行 Lag 變數使用後 MSE 的結果進行計算，計算後的結果如下圖：



如圖可以發現前段在 n = 4 的時候達到了最佳的結果，而整體最佳則落在了 n = 20 的情況下。最終，因從圖中可以發現到在 n > 5 後 MSE 的不穩定性大幅提昇，同時較小的 n 象徵了更簡易解釋的情形下，為了達成穩定且好詮釋的目標，我們選擇了 n = 4。

最後，在參數徹底篩選結束後，我們將處理過的資料放入模型中進行擬合，便可以完成 Poisson GLM with Lag 模型的建立。

D. Time Series

在開始進行對於 Time Series 類型的模型進行建立前，我們必須確認和釐清兩件事：

1. 保持在資料前處理後所得之資料狀況：

   Time Series 的模型在基本上基本皆並沒有要求 y 的分布必須符合特定的分布，同時，對於 X，也就是用於預測的變數也並沒有特定的要求，因此在建立時僅需要直接拿起始已前處理完成的資料進行整體

的模型擬合即可。

2. 利用 Rolling Window 來進行預測

Times Series 模型的一大特色是它是以時間軸進行計算，因此它無法像是前面兩個模型類別一樣使用 10 Folds Cross Validation，因此，為了仍然保持一定程度的嚴謹性和避免過度擬合帶來的誤差，我們選擇使用 Rolling Window 來進行驗證，也就是每次都僅利用連續性的 7000 筆資料對謎行擬合，接著預測自最後一筆資料開始算起的後三筆資料，在記錄後再將原資料的前三筆刪去和將後三筆加入，然後再重複擬合以及其後面的動作。

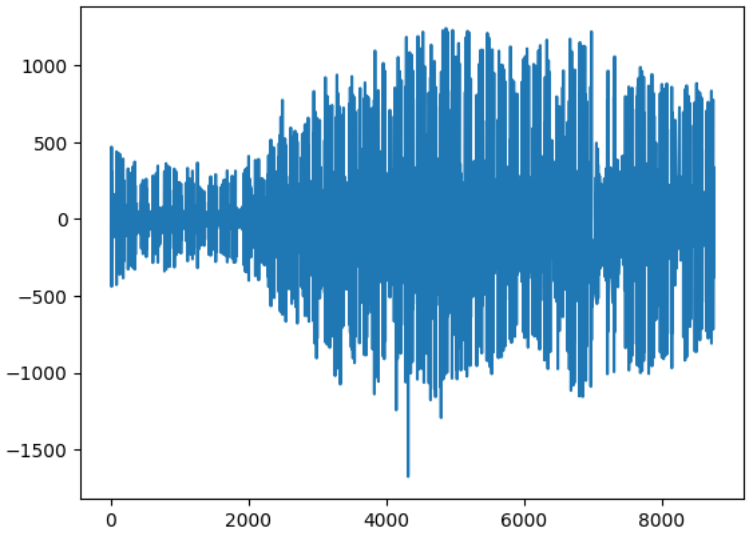
此種作法雖無法做到 K Folds Cross Validation 一樣的精準避免過度擬合和選擇到不適切樣本的狀況，但其仍能夠幫助我們達成對 Time Series 類型模型的檢驗。

以下便是模型的詳細參數使用的進行過程補述：

1. ARIMA

在開始進行模型建立前，我們必須要對於模型中的(p, d, q)參數進行選擇。

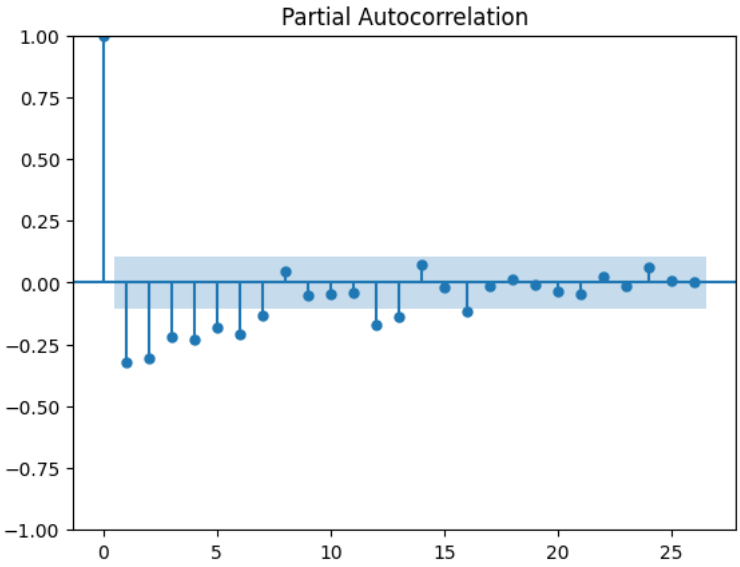
首先針對 d，d 的參數目的是為了讓我們能使 y，也就是希望預測的變數保持一定的穩定性，而 d 本身則代表了對 y 進行幾次的微分，因此，我們先看對 y 一次微分的情況下作圖以觀看其穩定性，我們會見到下圖：

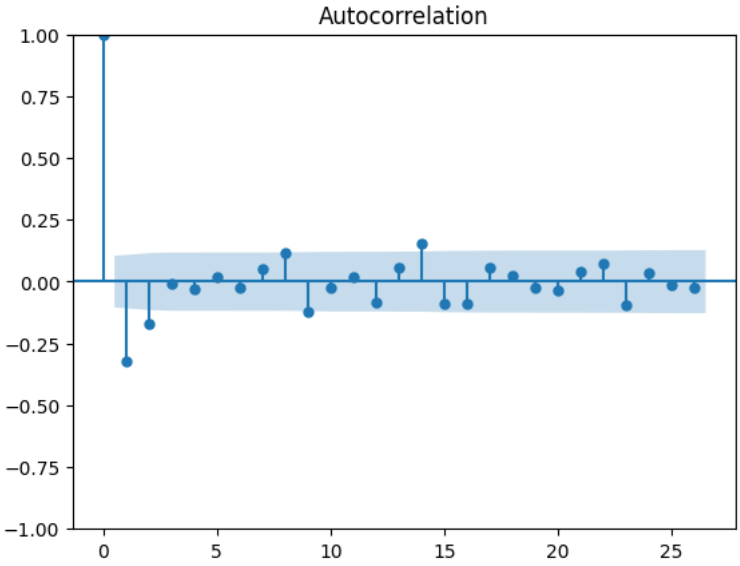


可以發現他的穩定性已經非常高，因此在此基礎下，我們對其進行 ADF(Augmented Dickey-Fuller) test，可以發現到它的 p-value 非常趨近於 0，在此假設檢定的虛無假設是此變數不穩定的情況下，我們可

以了解 d = 1 的時候確實拒絕了此虛無假設，換言之，d = 1 能夠完成我們所希望達成穩定的目標。

接下來是針對 p 和 q 的部份，此階段需要藉由對於 partial autocorrelation 和 autocorrelation 的圖來進一步分析此二參數可能的數字，二圖如下：





然而此參數難以直接確認，僅能藉由圖中明顯偏離中心之數額作為大概標準，搭配上嘗試與逐步調整以找出最佳參數組合，最後，在多方嘗試與確認後，p 和 q 之值將設定為 3 與 2。至此，我們便能確認(p, d, q)的參數選擇為(3, 1, 2)。

在完成設定後，便也是將前處理過後的資料放入其中，但 ARIMA 的特性是僅需將 y，也就是共享單車借用數的資料做擬合，因此便是

將去除掉非運作日的資料放入模型架構做擬合，即可得到上述的結果。

2. ARMAX

首先，ARMAX 在運作上與 ARIMA 只相差了外生變數(eXogenous Variable)，這使得他需要選擇(p, d, q)，但事實上會與在 ARIMA 的選擇相同，因此，我們在 ARMAX 的(p, d, q)選擇也為(3, 1, 2)。

次來，我們要解決外生變數的問題，此處外生變數的應用流程與使用線性迴歸時相當相似，我們一樣是將處理過的變數全數加入 ARMAX 模型內進行擬合，爾後我們便會發現到模型內「降雨量」、「降雪量」、「季節」和「是否為節日」皆未通過迴歸係數不為 0 的假設檢定，因此，在經過輪流的剔除和各式檢驗與討論後，我們確認此四個變數對於模型的建立與預測完全沒有幫助，將他們全數剔除並無使得 MSE 上升，因此最後選擇全數移除，剩餘變數加上前方所確認的(2, 1, 3)作為(p, d, q)的參數，便完成了對於 ARMAX 模型的建立。