Homework 2: R Practice

陳力行
r13627024

December 5, 2024

1 Read Data

Question 1.1

我感興趣的研究問題爲夜間車禍的死傷人數會不會相較於日間來得多，透過定義時間點來分類，夜間爲實驗組 (晚上 8 點 ~ 凌晨 3 點)，其他時點爲對照組，作爲我的 treatment，而我的 outcome 爲車禍事故的死傷人數。我使用以下的程式碼來讀取資料，其中資料來源是臺北市資料大平台，根據北市警察局交通大隊提供逐年搜集的所有車禍事故明細，包含事故類型、道路形態、視野距離、死傷人數以及駕駛飲酒情形等諸多變數，最終我選取民國 112 年 1 月到 12 月期間的車禍事故作爲主要的分析資料集。

```
# Q1 read data
install.packages("readxl")
library(readxl)
file_path <- "/Users/coco/Desktop/台大計量/R/raw/112.xlsx"
data <- read_excel(file_path)
```

2 Examine Data

Question 2.1

首先，我使用了 `str(data)`，查看我的所有欄位的，查看所有欄位的資料結構，確認了哪些欄位是字串，哪些欄位是數值資料；其次使用 `summary(data)` 查看原始資料中各個欄位四分位數以及平均數；最後透過 `colSums(is.na(data))` 計算每個欄位的缺失值爲多少。

```
# Q2.1 examine data
str(data)
summary(data)
colSums(is.na(data))
```

Question 2.2

首先，我使用了 `any(duplicated(data)` 指令確認是否存在重複值，接下來生成一個資料表來存放重複值，並使用 `print(duplicated_rows)` 將重複值列出來，可以看出原始資料中不存在重複值。

```r
# Q2.2 examine data
install.packages("dplyr")
library(dplyr)
any(duplicated(data))
duplicated_rows <- data[duplicated(data), ]
print(duplicated_rows)
```

3 Create Sample for Analysis

Question 3.1

首先，我透過 select 這個指令選擇我所需要的欄位，並計算了每個欄位的缺失值；接下來我建立了新的資料表並透過 drop_na 指令將缺失值移除，並重新確認一次是否存在缺失值。

再來針對區序欄位做處理，我提取資料值中的前兩個數字將其定義爲數值資料（例如：原始資料爲 04 大安區，我生成一個新的欄位叫做區序數字，其資料值爲 4)。

接下來針對天候欄位做處理，首先我發現資料中存在一些不合理的值（問卷中只定義了 1~8，但卻出現了 9)，我將其刪除，並透過 as.factor 確保其爲 factor 類型，並針對資料值與其對應的名稱去生成相應名稱的虛擬變數欄位（像是當資料值爲 1，則定義爲天候 _ 暴雨等以此類推)。

再來將死亡人數與受傷人數加總，並生成一個新的欄位叫做死傷人數。

接下來處理道路類別欄位，我透過 table(data$ 道路類別) 查看各種道路的分布數量，發現臺北市大部分車禍事故皆發生在市區道路，因此我將其分爲兩類，實驗組爲市區道路，對照組則爲其他種道路類型的累計加總。
針對路面狀態的處理類似於天候資料，同樣是確保其資料類型後將其針對不同數字做分類，並且生成相應名稱及資料值的虛擬變數。

接下來針對我的感興趣的 treatment 做處理，首先利用 as.numeric 確保欄位爲數值資料，並透過 ifeles 做時點分類，定義晚上八點到凌晨三點爲夜間，並生成夜間虛擬變數，並於後續將已處理完且不需做後續分析的欄位刪除。
最後將 data_clean 資料表的所有欄位名稱列出，確認是否有誤。

```r
# Q3 create sample for analysis
data_select <- select(data, "發生月", "發生日", "發生時-Hours", "區序", "死亡人數", "受傷人數",
                      "天候","道路類別", "路面狀況2")
colSums(is.na(data_select))
install.packages("tidyr")
library(tidyr)
data_clean <- data_select%>% drop_na()
sum(is.na(data_clean))
summary(data_clean)

data_clean$區序數字 <- as.numeric(gsub("\\D", "", data_clean$區序))
head(data_clean$區序數字)
str(data_clean$區序數字)

data_clean <- data_clean %>% filter(天候 != 9)
data_clean$天候 <- as.factor(data_clean$天候)
weather_labels <- c(
  "1" = "天候_暴雨",
```

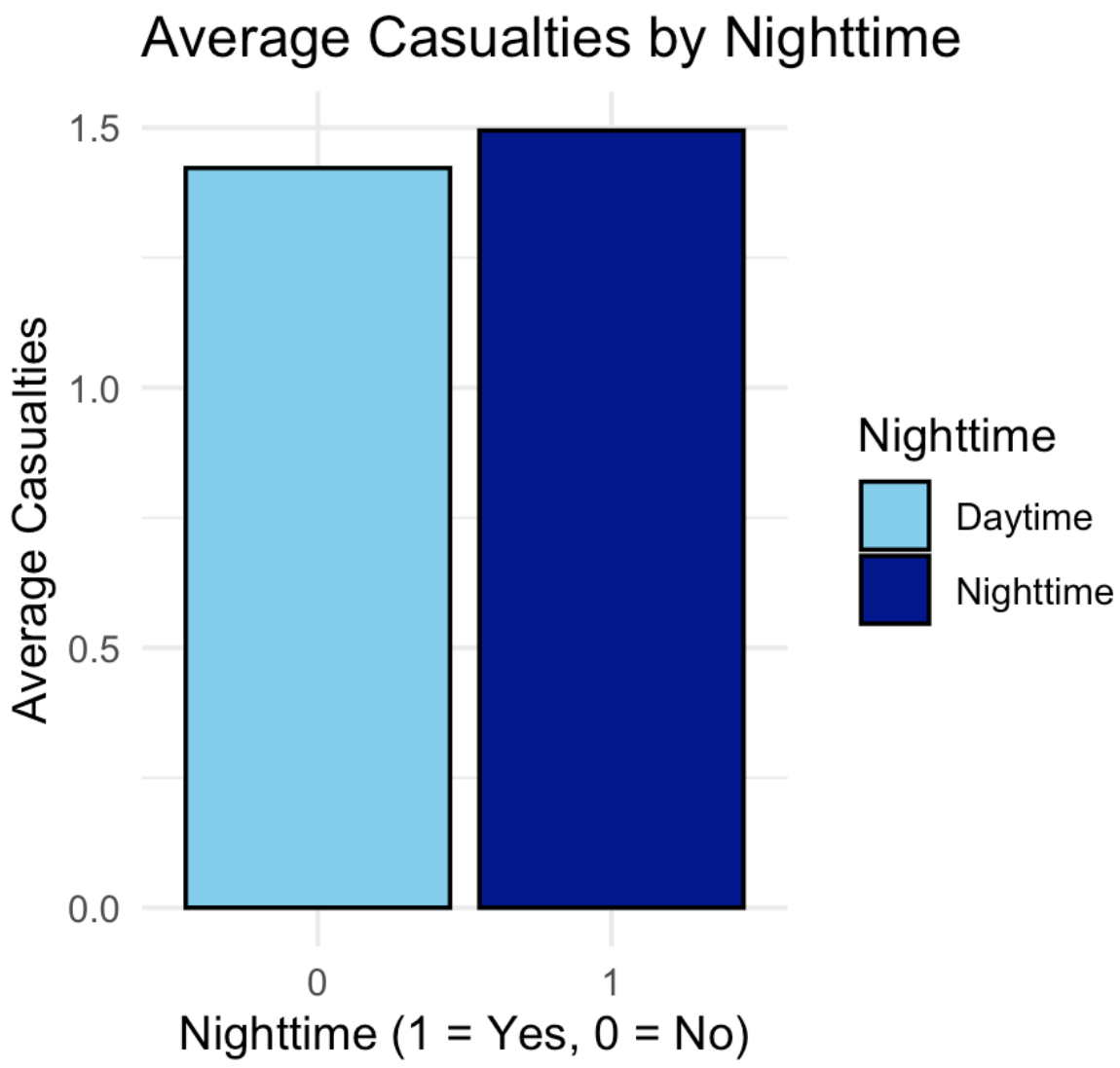
```
  "2" = "天候_強風",
  "3" = "天候_風沙",
  "4" = "天候_霧或煙",
  "5" = "天候_雪",
  "6" = "天候_雨",
  "7" = "天候_陰",
  "8" = "天候_晴"
)
for (val in names(weather_labels)) {
  col_name <- weather_labels[val]
  data_clean[[col_name]] <- ifelse(data_clean$天候 == val, 1, 0)
}

data_clean$死傷人數 <- data_clean$死亡人數 + data_clean$受傷人數 # 新增欄位 "死傷人數"

table(data$道路類別)
data_clean <- data_clean %>%
  mutate(市區道路 = ifelse(道路類別 == 5, 1, 0))
table(data_clean$市區道路)

data_clean$路面狀況2 <- as.factor(data_clean$路面狀況2)
surface_labels <- c(
  "1" = "路面狀態_冰雪",
  "2" = "路面狀態_油滑",
  "3" = "路面狀態_泥濘",
  "4" = "路面狀態_濕潤",
  "5" = "路面狀態_乾燥"
)
for (val in names(surface_labels)) {
  col_name <- surface_labels[val]
  data_clean[[col_name]] <- ifelse(data_clean$路面狀況2 == val, 1, 0)
}

data_clean$`發生時-Hours` <- as.numeric(data_clean$`發生時-Hours`)
data_clean$夜間 <- ifelse(data_clean$`發生時-Hours` >= 0 & data_clean$`發生時-Hours` <= 3
    |
                          data_clean$`發生時-Hours` >= 20 & data_clean$`發生時-Hours` <=
                              24, 1, 0)

data_clean <- data_clean %>%
  select(-c("天候", "道路類別", "路面狀況2"))

head(data_clean)
```

4 Visualize Data

Question 4.1

圖 1 顯示平均每次事故所造成死傷人數的直方圖，分別發生在夜間及非夜間的差距。

Figure 1:



```r
#Visualize Data
install.packages("ggplot2")
library(ggplot2)
data_clean$發生月 <- as.numeric(data_clean$發生月)
monthly_data <- data_clean %>%
  group_by(發生月) %>%
  summarise(total_casualties = sum(死傷人數, na.rm = TRUE))
ggplot(monthly_data, aes(x = 發生月, y = total_casualties)) +
  geom_line(color = "blue", linewidth = 1) +  # 折線
  geom_point(color = "red", size = 2) +
  labs(
    title = "Total Casualties by Month",
    x = "Month of Incident",
    y = "Total Casualties"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5)
  )+
  scale_x_continuous(breaks = 1:12)
```

Question 4.2

我們可以透過圖 1 直覺地觀察到夜間的平均死傷人數相較於非夜間時段略高。這反映了每起事故在夜間的嚴重程度可能更高，例如光線不足或高速駕駛等因素可能增加事故嚴重程度。

5 Term Paper Writing

Question 5.1

我的研究問題爲夜間車禍對於死傷人數的影響。
日常生活中車禍是常見的安全問題，而夜間車禍似乎更容易造成嚴重傷亡，其中夜晚駕駛視線不佳、疲勞影響及酒駕等因素，直覺上讓人認爲夜間車禍的風險較高。然而，這樣的假設是否成立仍然缺乏充分的實證研究。因此，我想探討夜間車禍是否比白天更嚴重，並分析影響死傷人數的關鍵因素。

Question 5.2

我選擇北市警察局交通大隊統計的民國 112 年 A1 及 A2 車禍事故明細表。
首先，針對數據做預處理，查看原始資料中有幾個欄位，利用 summary 查看各個欄位的狀態以及缺失值，後續利用 `any(duplicated())` 確認是否有重複值。

接下來篩選出我所需要的欄位（例如發生時、發生月、死亡人數、傷亡人數、區序以及天候等其他可能作爲控制變數的欄位），針對我感興趣的 treatment 做處理。首先定義夜間時段，我考量到四季的日出與太陽下山時間的不同，定義了最保守的時段作爲夜間時段（晚上八點 ~ 凌晨三點），並生成一個虛擬變數叫做夜間，若車禍發生在這個時段則資料值爲 1（夜間爲實驗組），其他爲 0（白天爲對照組）。再來針對我的 outcome 做處理，我生成一個新的欄位死傷人數 = 死亡人數 + 受傷人數，並確保其爲數值資料。

接下來針對控制變數做處理，首先，我查看了天候、道路類別以及路面狀態的資料值分布情況，發現了一些不在問卷定義範圍內的值，將其清理乾淨，後續我針對各個欄位去做虛擬變數，並設定了" 天候 _ 晴天"、" 市區道路以外的其他道路類型" 以及" 路面狀態 _ 乾燥" 作爲迴歸模型的對照組 (控制變數的資料處理細節在 Q3.1 已在詳細說明)。

Question 5.3

圖 2顯示了總死傷人數的平均數和標準差、實驗組 (Night) 與對照組 (Day) 的平均數及標準差，其中的計算方式以事故發生次數作爲權數，表示每次事故平均發生死傷人數爲 1.4345，標準差爲 0.7167。
可以透過夜間與白天的平均數看出車禍發生在夜間的平均死傷人數是較高的，表示夜間的車禍嚴重程度可能普遍較白天嚴重，這與我們一開始猜測的結論相同。再來我們發現夜間車禍的標準差也高於白天，表示夜間車禍的死傷情況變化較大，可能由於視線較差、駕駛疲勞或酒後駕駛等風險因素。儘管夜間車禍數量較少，但每起事故的影響可能較大，造成較嚴重的傷亡。

Figure 2:

	Group	平均數	標準差
1	Overall	1.434464	0.7166593
2	Day	1.422418	0.7123002
3	Night	1.494703	0.7351547

Question 5.4

以下是我估計的迴歸模型：

$$Y_i = \beta_0 + \beta_1 \text{Night}_i + \mathbf{X}_i\boldsymbol{\beta} + \epsilon_i$$

其中 Y_i 爲死傷人數，是我的主要 outcome，以事故爲單位做累加計算，透過原始資料的死亡人數與受傷人數欄位相加總計算得來，ϵ_i 爲誤差項。$Night_i$ 則是我主要的 treatment，若車禍發生在晚上 8 點 ~ 凌晨 3 點則定義爲 1 (屬於夜間時段)，其餘時段則爲 0。$\mathbf{X}_i$ 爲控制變數包含天候 (暴雨、風沙、霧或煙、雨、陰、晴天)、道路類別，以及路面狀態 (冰雪、油滑、泥濘、濕潤、乾燥) 的虛擬變數。

透過以上控制變數我可以更好控制影響死傷人數的外部環境因素，後續分析可以再加入地區及月份的固定效果控制季節性與區域性對 outcome 的影響。

在尚未看到估計結果的情況下，我們預期 treatment 與死傷人數之間存在正相關。夜間視線受限且駕駛反應時間可能較慢，更有可能增加事故的嚴重程度與死傷人數。

關鍵假設

1. **樣本存在限制性：**
   本研究只使用了台北市的交通事故資料，雖然可以深入了解這個地區的交通特性和事故情況，但結果可能無法適用於其他地區。

2. **誤差項的隨機性：**
   假設誤差項 ϵ 獨立且同質，且不與 treatment(夜間) 相關。

3. **控制變數具外生性：**
   控制變數中涵蓋了影響死傷人數的關鍵環境因素（如天候與路面狀態）以及結構性條件（如道路類型），並且與夜間具外生性，確保 treatment 的估計能隔離其他干擾因素，進一步支持我們因果推論的有效性。

Question 5.5

表 1顯示了加入控制變數前後的迴歸模型。
首先，可以看到 (1) 欄中發現，我們的 treatment 與死傷人數呈現正相關並且具統計顯著性，表示夜間的死傷人數相較於非夜間時段高。

接下來，在（2）欄中我加入了天候、道路以及路面狀態的控制變數後，可以發現天候變數中，霧與煙相較於晴天更容易造成車禍死傷人數增加且具統計顯著性。再來，針對市區道路變數可以看到也具統計顯著性，可被解讀爲市區道路相較於其他種道路型態對於死傷人數的影響爲負相關，因爲市區道路交通較爲雍塞且時速限制，可以有效降低車禍事故死傷人數。後續針對路面狀態變數可以看到每項都具統計顯著性，首先可以看到相較於乾燥路面來說，冰雪、油滑和泥濘的路面更容易提升車禍事故的死傷人數，然而針對濕潤的路面對死傷人數相較於乾燥路面呈負相關，其原因可能在於駕駛經過濕潤路面會降低速度並且小心行駛，進而導致事故死傷人數可以有效降低。最後，對於模型整體評估可以看到 F-value 皆具統計顯著性，而且透過判定係數的上升，可以看到加入更多控制變數後，模型對死傷人數的解釋能力有所提高。

我認爲後續可以針對不同行政區（如大安區及中正區等等）加入區域的固定效果去捕捉區域性且不隨時間變動而影響的效果，有助於消除區域間的潛在偏誤，使我們能夠更精確地分析其他變數（如夜間、天氣條件等）對車禍死傷的影響。還可以加入月份的時間固定效果，控制由時間或季節變化但不隨個體改變的固定效果，這些因素可能影響交通事故的發生率和死傷人數。例如，冬季的惡劣天氣或夏季的旅遊高峰期可能導致交通流量或道路狀況的變化，進而影響事故風險。因此，加入月份固定效果可以減少這些潛在偏誤，進而提高模型的準確性和解釋能力。

Table 1: Result

	Dependent variable:	
	死傷人數	
	(1)	(2)
夜間	0.072***	0.076***
	(0.008)	(0.008)
天候 _ 暴雨		−0.334
		(0.217)
天候 _ 風沙		−0.433
		(0.505)
天候 _ 霧或煙		0.570***
		(0.216)
天候 _ 雨		0.008
		(0.023)
天候 _ 陰		−0.009
		(0.008)
市區道路		−0.088***
		(0.018)
路面狀態 _ 冰雪		0.862***
		(0.207)
路面狀態 _ 油滑		0.355***
		(0.081)
路面狀態 _ 泥濘		0.471***
		(0.107)
路面狀態 _ 濕潤		−0.100***
		(0.021)
Constant	1.422***	1.522***
	(0.003)	(0.018)
Observations	56,076	56,076
R^2	0.001	0.005
Adjusted R^2	0.001	0.005
Residual Std. Error	0.716 (df = 56074)	0.715 (df = 56064)
F Statistic	79.339*** (df = 1; 56074)	26.881*** (df = 11; 56064)
Note:		*p<0.1; **p<0.05; ***p<0.01