統計期末報告 -

網購因素分析

組員: B084012001 李冠融、B084012010 戴子晴

分工表:

姓名	負責內容
李冠融	問卷設計、數據分析、報告製作、口頭報告、簡報製作
戴子晴	問卷設計、數據分析、報告製作、口頭報告

身處這個科技日新月異的世代，在日常生活中，人們更加提倡便利性與效率，使得網路線上購物逐漸成為現代人的採購模式，再加上近期疫情嚴峻，大幅降低出門購物的可行性，因此我們決定配合時事，將研究主題設定為「網路購物因素分析。」且因為此主題十分貼近日常生活，所以我們小組選擇自行設計網路問卷的方式收集統計資料，並已順利取得 135 個樣本數。(圖一為問卷內容)

(圖一)

模型建立

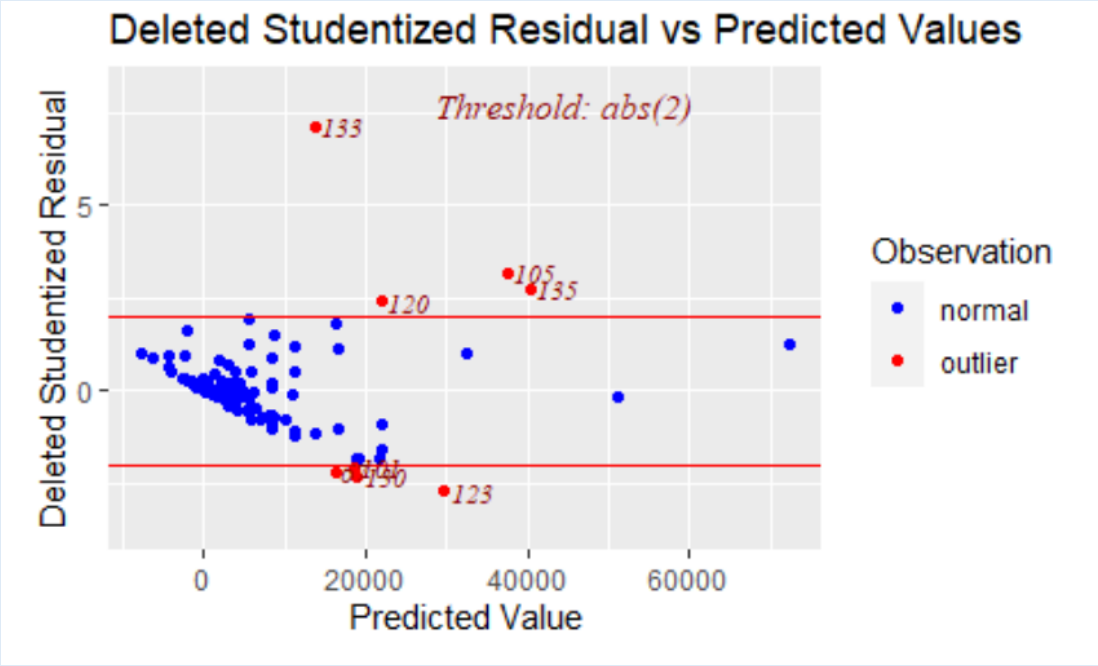
我們的問卷根據可能影響消費者進行網購行為的因素設計問題，我們假定以下五個因素作為原始模型的自變數，分別為:性別(gender)、年齡(年齡區間:0-24、25-49、50(含)以上，設 dummy 得自變數 agea 和 ageb)、收入(income)、節慶(holiday)、跟隨流行(trend)、天氣(weather)，及前一個月的網購金額(spend)則作為模型的依變數，並利用以上資訊建立網購因素分析的多元迴歸模型。

```r
data <- fread("C://Users//Gladice//Desktop//R 下//試做9.csv")

data$gender <- ifelse(data$gender=="male",1,0)
data$holiday <- ifelse(data$holiday=="yes",1,0)
data$trend <- ifelse(data$trend=="yes",1,0)
data$weather <- ifelse(data$weather=="yes",1,0)

agetable <- data.frame(data$age)
dummyage <-dummy.data.frame(agetable)
colnames(dummyage) <- c("agea","ageb","agec")
data <- cbind(data,dummyage)

datalm <- lm(spend~gender+income+agea+ageb+holiday+trend+weather,data=data)
summary(datalm)
```

(圖二)

殘餘項分析

我們使用 Deleted Studentized Residual 的方式進行殘餘項的分析，可以從(圖三)看出大部分資料的殘差值相當集中於左方中央部分，有越往右越擴散的分布狀況，且有些許離群值存在，需透過刪減自變數，找出配適度最高的自變數組合來修正模型。



(圖三)

檢定

透過 F 檢定(圖四)，我們可以發現 p-value 的數值極小，足以拒絕 H0:全部複迴歸係數皆為 0 的假設；判斷(圖四)中各項的 t 檢定數據可以發現，除了 trend 及 weather 二自變數以外，其餘自變數皆有相當的顯著程度；多重共線性的檢定部分(圖五)，所有自變數的數值皆<10，並沒有多重共線性的問題存在；Adj. R-squared 為 0.6342，初步判斷雖然目前模型準確率並不高，但可以經由修正模型得出更為準確的模型。

```
Residual standard error: 7849 on 127 degrees of freedom
Multiple R-squared:  0.6533,    Adjusted R-squared:  0.6342
F-statistic: 34.19 on 7 and 127 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Response: spend
           Df     Sum Sq    Mean Sq  F value  Pr(>F)
gender      1 1.8216e+08 1.8216e+08   2.9566 0.08796 .
income      1 1.3926e+10 1.3926e+10 226.0379 < 2e-16 ***
agea        1 2.8150e+08 2.8150e+08   4.5691 0.03447 *
ageb        1 1.6594e+08 1.6594e+08   2.6933 0.10324
holiday     1 1.8812e+08 1.8812e+08   3.0535 0.08298 .
trend       1 1.1534e+06 1.1534e+06   0.0187 0.89138
weather     1 7.5126e+04 7.5126e+04   0.0012 0.97220
Residuals 127 7.8245e+09 6.1610e+07
```

(圖四)

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  gender   income     agea     ageb  holiday    trend  weather
1.274881 1.642605 2.555288 1.898849 1.119814 1.091761 1.050130
```

(圖五)

增刪自變數

我們使用了四種選擇自變數的方式，首先用 Stepwise regression(圖六、七、八)進行初步篩選，藉由前面檢定所得到的顯著程度，依序檢驗修正模型的配適度，找出最合適的自變數組合；接著從 Forward selection 和 Backward elimination 兩種方法驗證得到同樣的結果(圖九、十)；最後以 Best subset selection 選出最合適的模型(圖十一、十二)，所得結果與前者相符合，於是最終我們選擇的自變數為:gender、income、agea、ageb、holiday，並使用以上自變數建立修正模型(圖十三)，得到修正後的 Adj. R-squared 為 0.6398，相對

原始模型準確率有提升。

```
data_1 <- lm(spend~income,data=data)
summary(data_1)
data_2 <- lm(spend~gender+income,data=data)
summary(data_2)
data_3 <- lm(spend~gender+income+agea+ageb,data=data)
summary(data_3)
data_4 <- lm(spend~gender+income+agea+ageb+holiday,data=data)
summary(data_4)
data_5 <- lm(spend~gender+income+agea+ageb+holiday+trend,data=data)
summary(data_5)
data_6 <- lm(spend~gender+income+agea+ageb+holiday+weather,data=data)
summary(data_6)
```

(圖六)

+ 收入(income)

```
income       2.312e-01  1.617e-02  14.298   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8178 on 133 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6029
```

+ 性別(gender)

```
gender      -3.745e+03  1.438e+03  -2.603   0.0103 *
income       2.435e-01  1.652e-02  14.740   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8006 on 132 degrees of freedom
Multiple R-squared:  0.6251,    Adjusted R-squared:  0.6194
```

R-squared 上升，表示適合此模型中。

+ 年齡(age)

```
gender      -2.442e+03  1.509e+03  -1.618  0.10801
income       2.660e-01  1.989e-02  13.376  < 2e-16 ***
agea         5.767e+03  2.142e+03   2.693  0.00802 **
ageb         3.313e+03  2.019e+03   1.641  0.10328
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7851 on 130 degrees of freedom
Multiple R-squared:  0.6449,    Adjusted R-squared:  0.634
```

R-squared 上升，表示適合此模型中。

(圖七)

+ 特殊節慶(holiday)

```
gender       -2.500e+03  1.497e+03  -1.670  0.09744 .
income        2.660e-01  1.973e-02  13.481  < 2e-16 ***
agea          6.194e+03  2.138e+03   2.897  0.00443 **
ageb          2.900e+03  2.017e+03   1.438  0.15291
holiday       2.751e+03  1.562e+03   1.761  0.08061 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7789 on 129 degrees of freedom
Multiple R-squared:  0.6533,    Adjusted R-squared:  0.6398
```

R-squared 上升，表示適合此模型中。

+ 跟隨流行(trend)

```
gender       -2530.2368  1519.2458  -1.665  0.09827 .
income           0.2659     0.0198  13.427  < 2e-16 ***
agea          6214.6701  2151.7129   2.888  0.00455 **
ageb          2899.9655  2024.7703   1.432  0.15451
holiday       2779.5499  1581.7695   1.757  0.08127 .
trend         -213.4564  1553.9392  -0.137  0.89096
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7819 on 128 degrees of freedom
Multiple R-squared:  0.6533,    Adjusted R-squared:  0.6371
```

R-squared 下降，因此不放入模型中。

+ 天氣(weather)

```
gender       -2.504e+03  1.505e+03  -1.664  0.09857 .
income        2.659e-01  1.981e-02  13.422  < 2e-16 ***
agea          6.194e+03  2.147e+03   2.886  0.00459 **
ageb          2.909e+03  2.031e+03   1.433  0.15443
holiday       2.757e+03  1.572e+03   1.754  0.08181 .
weather      -1.008e+02  1.750e+03  -0.058  0.95415
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7819 on 128 degrees of freedom
Multiple R-squared:  0.6533,    Adjusted R-squared:  0.637
```

R-squared 下降，因此不放入模型中。

(圖八)

```
ols_step_forward_p(datalm)
ols_step_backward_p(datalm)
ols_step_best_subset(datalm)
```

(圖九)

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	income	0.6059	0.6029	13.3880	2819.5942	8178.3327
2	gender	0.6251	0.6194	8.3358	2814.8341	8006.2658
3	agea	0.6376	0.6293	5.7667	2812.2663	7901.9466
4	holiday	0.6477	0.6369	4.0555	2810.4389	7820.6284
5	ageb	0.6533	0.6398	4.0199	2810.2926	7788.7219

Elimination Summary

Step	Variable Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	weather	0.6533	0.6371	6.0012	2812.2727	7818.5112
2	trend	0.6533	0.6398	4.0199	2810.2926	7788.7219

(圖十)

Best Subsets Regression

Model Index	Predictors
1	income
2	gender income
3	income agea ageb
4	gender income agea holiday
5	gender income agea ageb holiday
6	gender income agea ageb holiday trend
7	gender income agea ageb holiday trend weather

(圖十一)

```
--------------------------------------------------------------------
                     Adj.        Pred
Model    R-Square    R-Square    R-Square     C(p)          AIC          SBIC
SBC             MSEP                FPE              HSP          APC
--------------------------------------------------------------------------------------
--------------------------------------------------------------------------
  1        0.6059      0.6029      0.5795    13.3880    2819.5942    2436.2123
2828.3100    9029507046.1606    67876016.9259    506705.5013    0.4060
  2        0.6251      0.6194      0.5966     8.3358    2814.8341    2431.6181
2826.4552    8654050541.9978    65524743.3835    489315.2095    0.3919
  3        0.6378      0.6295      0.6017     5.6940    2812.1924    2429.2174
2826.7188    8425874691.8645    64255647.4040    480050.4925    0.3843
  4        0.6477      0.6369      0.6146     4.0555    2810.4389    2427.7819
2827.8706    8258358382.0866    63427496.5045    474125.8044    0.3794
  5        0.6533      0.6398      0.6097     4.0199    2810.2926    2427.9252
2830.6296    8191603408.3831    63360375.2518    473938.9771    0.3790
  6        0.6533      0.6371      0.6049     6.0012    2812.2727    2430.0331
2835.5149    8254887331.6327    64298774.6273    481331.6277    0.3846
  7        0.6533      0.6342      0.5993     8.0000    2814.2714    2432.1580
2840.4189    8320322423.8345    65260811.2886    488967.1175    0.3904
--------------------------------------------------------------------------------------
```

(圖十二)

修正後模型

```
datalm_new <- lm(spend~gender+income+agea+ageb+holiday,data=data)
summary(datalm_new)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.896e+03  2.502e+03  -3.157  0.00199 **
gender      -2.500e+03  1.497e+03  -1.670  0.09744 .
income       2.660e-01  1.973e-02  13.481  < 2e-16 ***
agea         6.194e+03  2.138e+03   2.897  0.00443 **
ageb         2.900e+03  2.017e+03   1.438  0.15291
holiday      2.751e+03  1.562e+03   1.761  0.08061 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7789 on 129 degrees of freedom
Multiple R-squared:  0.6533,    Adjusted R-squared:  0.6398
F-statistic: 48.61 on 5 and 129 DF,  p-value: < 2.2e-16
```

(圖十三)

雖然修正模型後 Adj. R-squared 有提升，但殘餘項的分布(圖十五)與原始模型的殘差圖(圖三)沒有顯著的變化，我們判斷造成此狀況可能的原因有以下幾種:

1. 樣本數不足，導致無法看出明顯分布，並針對問題進行模型修正。

2. 問卷問題設計不夠詳細明確，造成填答人無法根據實情填答導致數據誤差。

3. 沒有考量到所有可能對依變數造成影響的因素，而出現部分離群值。

以上三點根據外部原因進行考量，若探討內部因素，可以使用離群值的數據來討論模型中的自變數如何造成影響，進而導致誤差值的出現，如以下兩表所示，我們可以將離群值樣本分成兩類，分別為 T 化殘差值 < -2 以及 T 化殘差值 > +2，而這兩類受到收入正向影響與年齡負向影響的相互抵消程度是造成誤差值的關鍵。

樣本編號	性別 (gender)	月收入 (income)	節慶 (holiday)	年齡 (age)	網購金額 (spend)
67	男	100000	不會	50+	0
101	男	98000	會	50+	3000
105	女	130000	會	25~49	10000
120	女	100000	會	50+	2000

(表一)

➔ **T 化殘差值 < -2: 代表實際網購消費 < 預估網購消費**，此群樣本的網購金額占收入比重偏低，即使高年齡降低了預估值，但高收入大大提高了預估消費，使實際上的消費遠低於預估的金額，造成負的殘差出現。

樣本編號	性別 (gender)	月收入 (income)	節慶 (holiday)	年齡 (age)	網購金額 (spend)
105	女	150000	會	25~49	60000
120	女	100000	會	25~49	40000
133	女	80000	會	50+	60000
135	女	180000	會	50+	60000

(表二)

➔ **T 化殘差值 > +2: 表示實際網購消費 > 預估網購消費**，此群樣本的網購金額占收入比重偏高，雖然高收入預估他們有高消費，但必須考慮年齡層偏中高會減少預估消費的因素，相互抵銷的狀況下產生實際消費金額遠多過於預估金額的誤差，造成正的殘差出現。

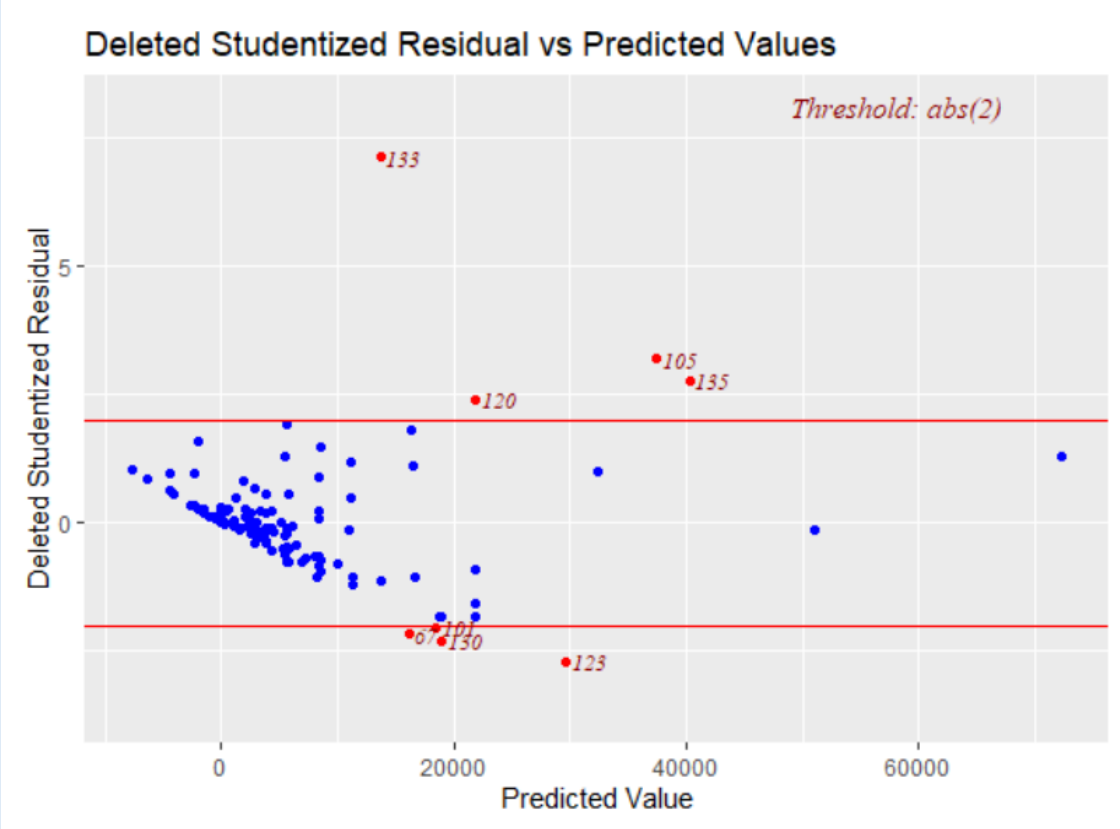
若考慮以上原因進行模型修正，可能會得到更為準確的模型，但從我們目前修正的模型中，還是能從數據得出影響網購行為的相關因素，並將其進行更實際的應用。

修正模型後的檢定

```
Residual standard error: 7789 on 129 degrees of freedom
Multiple R-squared:  0.6533,    Adjusted R-squared:  0.6398
F-statistic: 48.61 on 5 and 129 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Response: spend
            Df     Sum Sq    Mean Sq  F value  Pr(>F)
gender       1 1.8216e+08 1.8216e+08   3.0027 0.08551
income       1 1.3926e+10 1.3926e+10 229.5615 < 2e-16
agea         1 2.8150e+08 2.8150e+08   4.6403 0.03309
ageb         1 1.6594e+08 1.6594e+08   2.7353 0.10058
holiday      1 1.8812e+08 1.8812e+08   3.1011 0.08061
Residuals  129 7.8257e+09 6.0664e+07
```

(圖十四)

修正模型後的殘餘項分析



(圖十五)

自變數相關性

	gender	age	income	holiday	trend	weather	spend
gender	1.00000000	0.31820612	0.286217080	0.07052260	-0.18723149	-0.060740765	0.08983864
age	0.31820612	1.00000000	0.584223008	0.26166127	-0.15510713	-0.017241657	0.33696895
income	0.28621708	0.58422301	1.000000000	0.17546261	-0.10497096	0.002728857	0.77836565
holiday	0.07052260	0.26166127	0.175462605	1.00000000	0.08917149	0.086375124	0.20635152
trend	-0.18723149	-0.15510713	-0.104970960	0.08917149	1.00000000	0.177340439	-0.04677184
weather	-0.06074076	-0.01724166	0.002728857	0.08637512	0.17734044	1.000000000	0.01353639
spend	0.08983864	0.3369689	0.778365650	0.20635152	-0.04677184	0.013536389	1.00000000

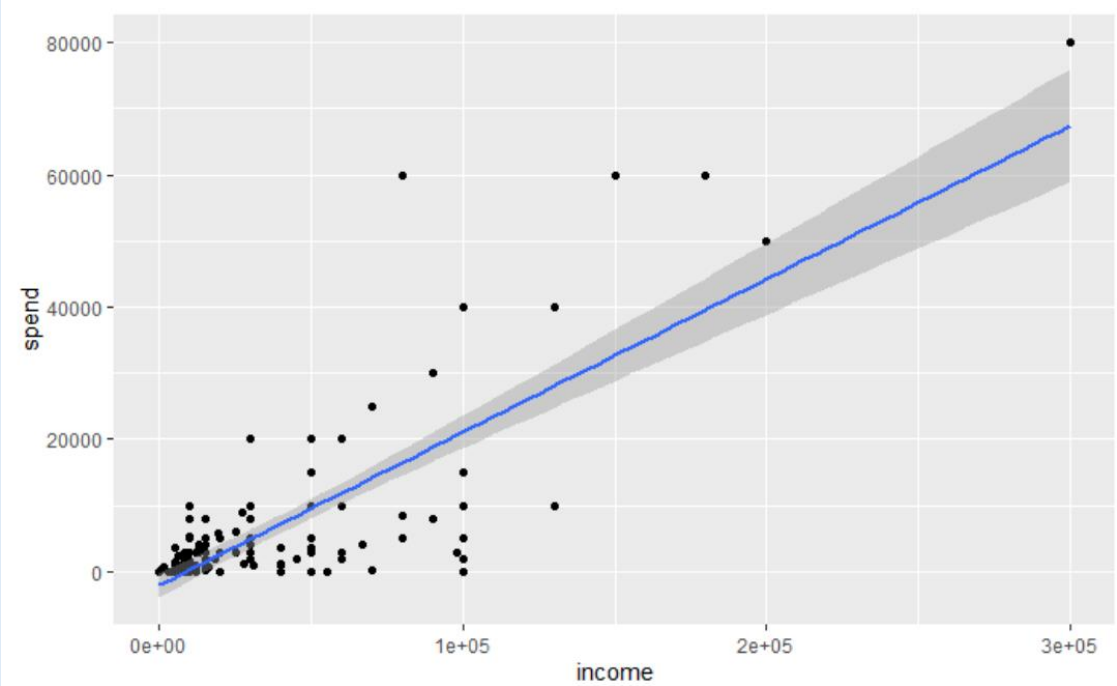
(圖十六)

管理意涵與實務應用

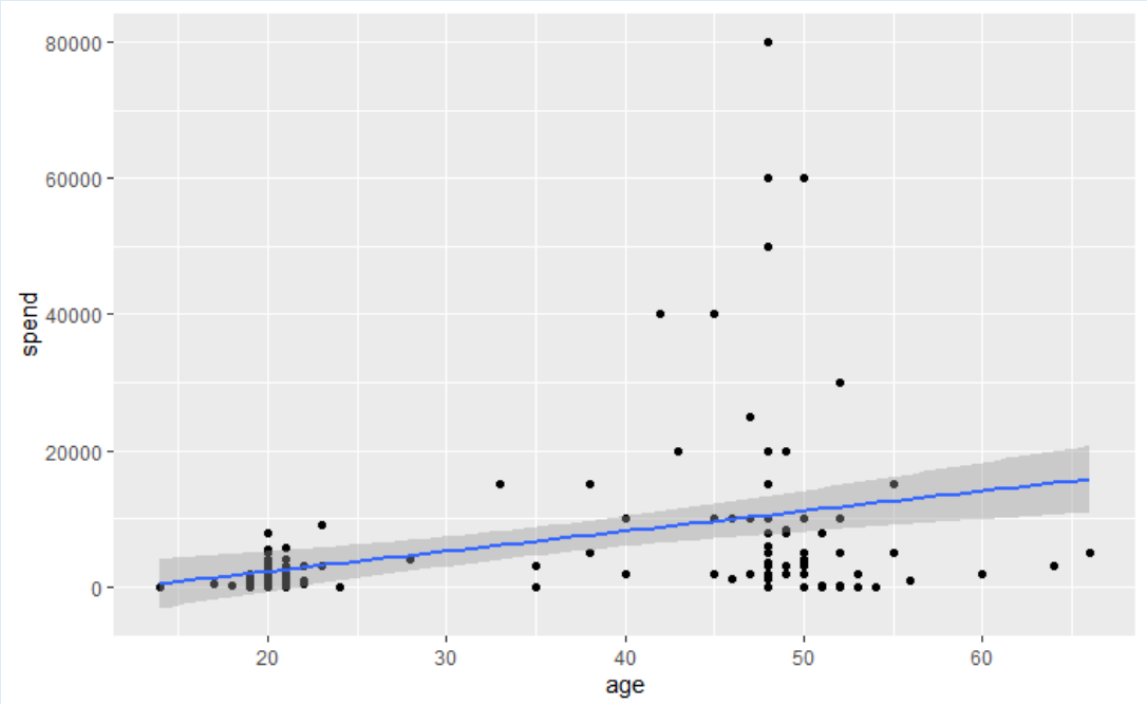
從修正模型(圖十三)中自變數 gender 及 holiday 的迴歸係數中可以了解，女性在網購上的花費較男性為多，以及在節慶假日時消費者較容易進行網購的行為，因此我們可以針對女性、節慶這兩點增加網購的因素提出行銷方案，例如母親節的行銷，可以在女性經常購買的商品進行各種折扣優惠，根據不同需求推出不同的促銷方案。

由(圖十六)可知自變數 income 與依變數 spend 間的相關係數高達 0.77837，我們可從中推得消費者的購物金額會隨著收入的上升而增加，此外，由(圖十六)可知自變數 income 與自變數 age 間的相關係數約為 0.5842，亦存在相關性，我們可從中推得以常態而言消費者的收入會隨著年齡增長而上升。不過透過分析我們也發現與中高齡族群相比，年輕族群較偏好以網路購物的方式進行採購，而這可能是由於年輕世代較能接受新時代的購物模式。

因此，網路商家在行銷產品時，針對不同收入、年齡層之消費者應採取不同的策略，明確區分目標客群。例如高價位之商品應聚焦於高收入的潛在顧客，並設計出專攻此族群的行銷手法，如實體店面的優惠活動，而另一方面，對於初入入職場、收入有限的年輕族群，推出網路購物免運費等的優惠方案。



(圖十七)



(圖十八)