回歸分析 作業二 多變數回歸分析 M10407519 林威佑

資料來源：<http://www.statsci.org/data/oz/bankbill.txt>

1. 找一個好的分析標的，能夠讓人非得要明白的例子的 dataset：

(你可以用 dataset for regression analysis 等 keywords，由 Google 找到很多提供 dataset 的來源。挑幾個資料量夠大，而且涉及的variables夠多，又能吸引你的候選例子，再從中選一個當作本次作業的分析對象。)

1. 請說明資料來源，並解釋此dataset之 response variable以及各 explanatory variables的物理意義，並列項分析幾個重要變數的大小、中位數等概括性的統計特性。(你知道 summary() 這個指令吧？)

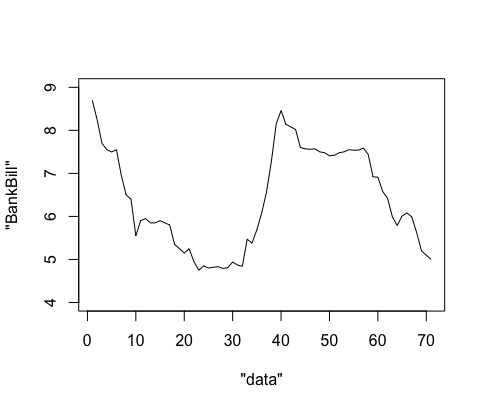
資料來源描述：

1991年10月至1997年8月記錄了各種股票價格和財務變量的月份觀察數據。

Francine Pritchard和Glen Dixon於1997年收集了MS305數據分析項目的數據。

總共有71筆資料與16個變數

整體利率與時間的變動如下：



response variable：

BankBill：未來90天銀行所發放執行的利率

利率指一定時期內利息與本金的比率，是決定利息多少的因素與衡量標準。利率作為資金的價格，決定和影響的因素很多、很複雜，利率水平最終是由各種因素的綜合影響所決定的。首先，利率分別受到產業的平均利潤水平、貨幣的供給與需求狀況、經濟發展的狀況的決定因素的影響，其次，又受到物價水平、利率管制、國際經濟狀況和貨幣政策的影響。

explanatory variables：

Bank：各銀行評比利率分享基數

AllOrds：all ordinaries普通個股指數

Develop：經濟發展指標

Mining：採礦生產指數

Gold：黃金換匯指數

Build:營建生產指數

Prop：Production：生產指數總額

Indust：Industrial Production：工業生產指數

Energy：能源指數

Finance：融資/資金融通（融資交易，投資人預期未來股價會上漲，但手中的資金不夠，於是繳交部份保證金向證券、金融公司等授信機構借錢買股票，不必繳足全數金額，而依融資融券標準繳交不同成數的自備款。）

Resource：原物料指數

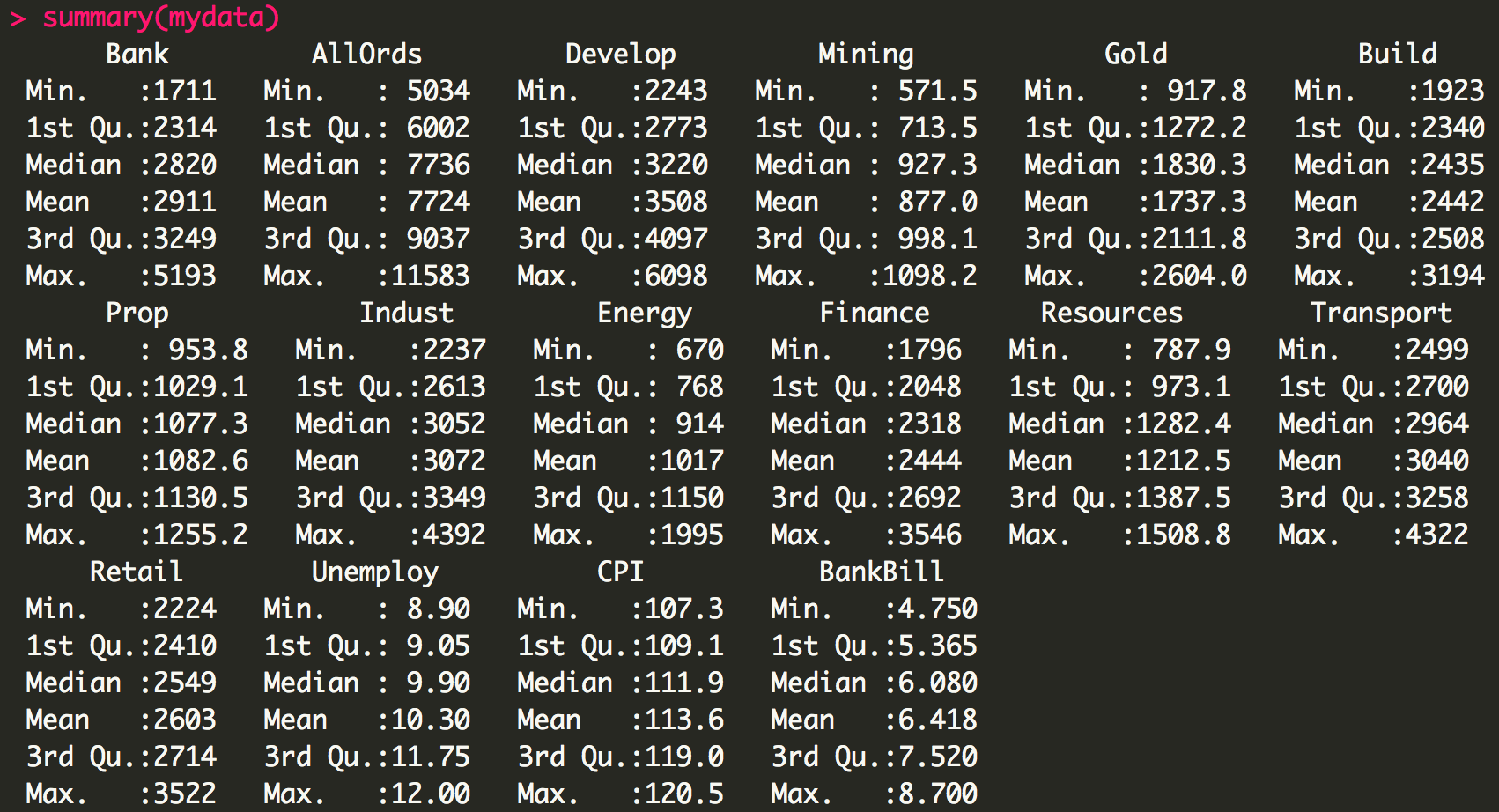
Transport：外匯轉移指數

Retail：零售指數

Unemploy：失業率

CPI：消費者物價指數

Summary():

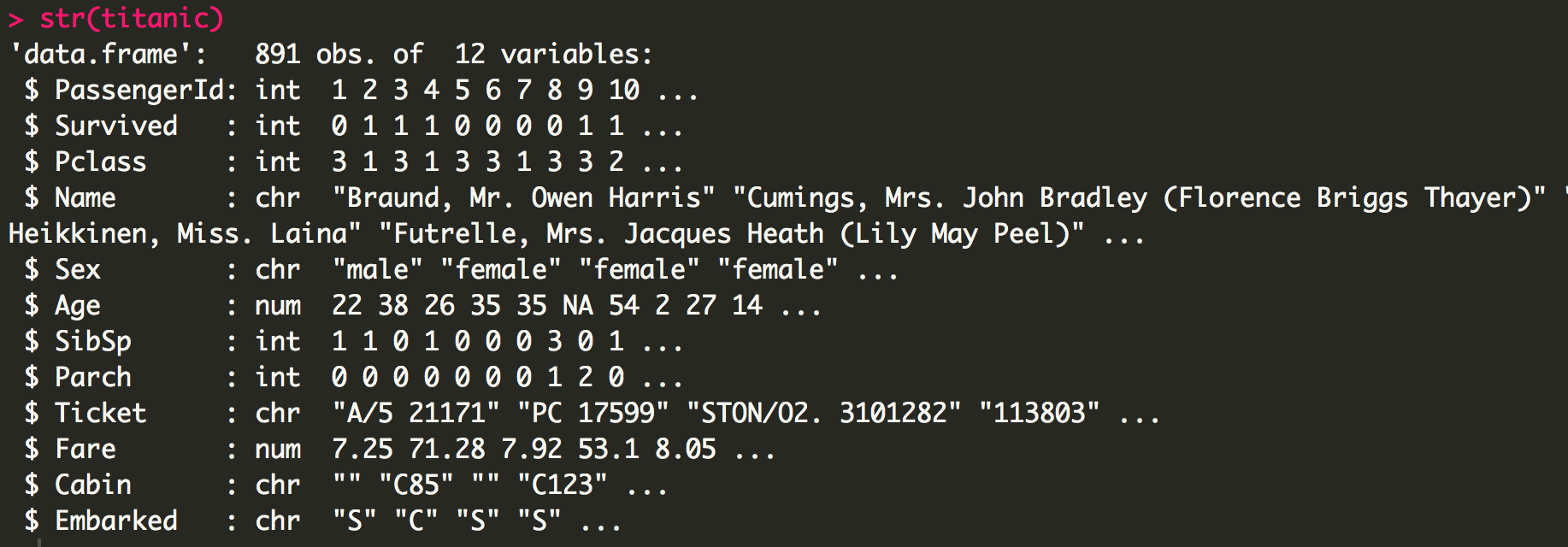


AllOrds：可以看出來在六年間的股票波動是相當大的，平均指數與中位數相當接近。

Unemploy：可以看出來六年間的失業率都高居不下。

CPI：可以由最小值到中位數與最大值到中位數做比較，發現有一半的CPI落在107~111之間，另外25%落在111~119之間，最後25%落在119~120之間。

1. 針對上述 (a)，請說明另一個你覺得也不錯的分析案例，然後說明最後你為什麼選擇 (a) 的那個例子。



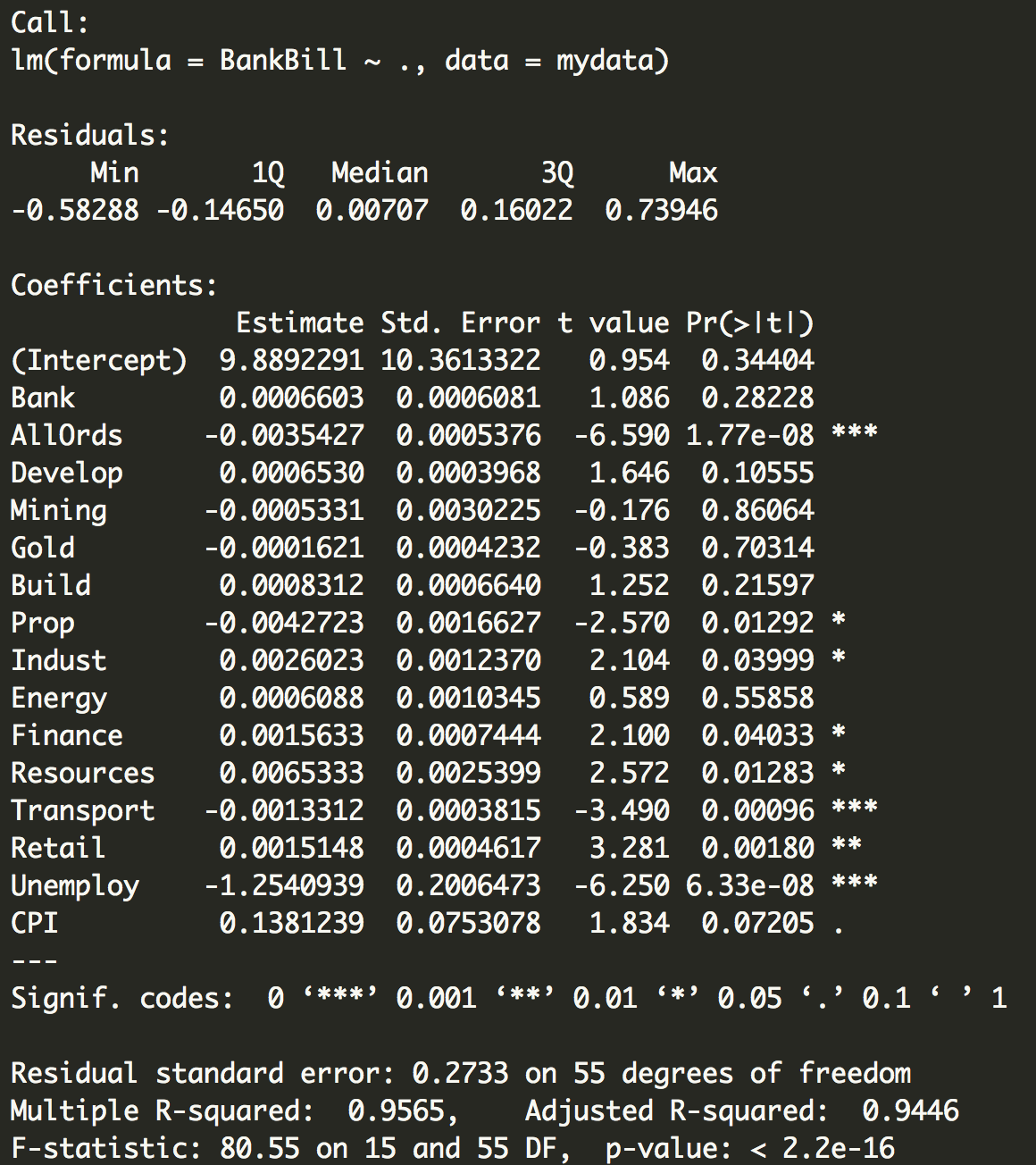
原先想要嘗試更有趣也大家知道的故事”鐵達尼號資料集”，不錯的原因是他的資料筆數891筆與12個變數，並且變數當中為電影中有的艙位、男女、年齡非常多種類，有類別的變數也有數值的變數是很好的預測標的，但是預測結果為是否生存，這種True/False問題應該由進階的邏輯回歸來做會比較適合。

1. 找出「最重要的」 explanatory variables

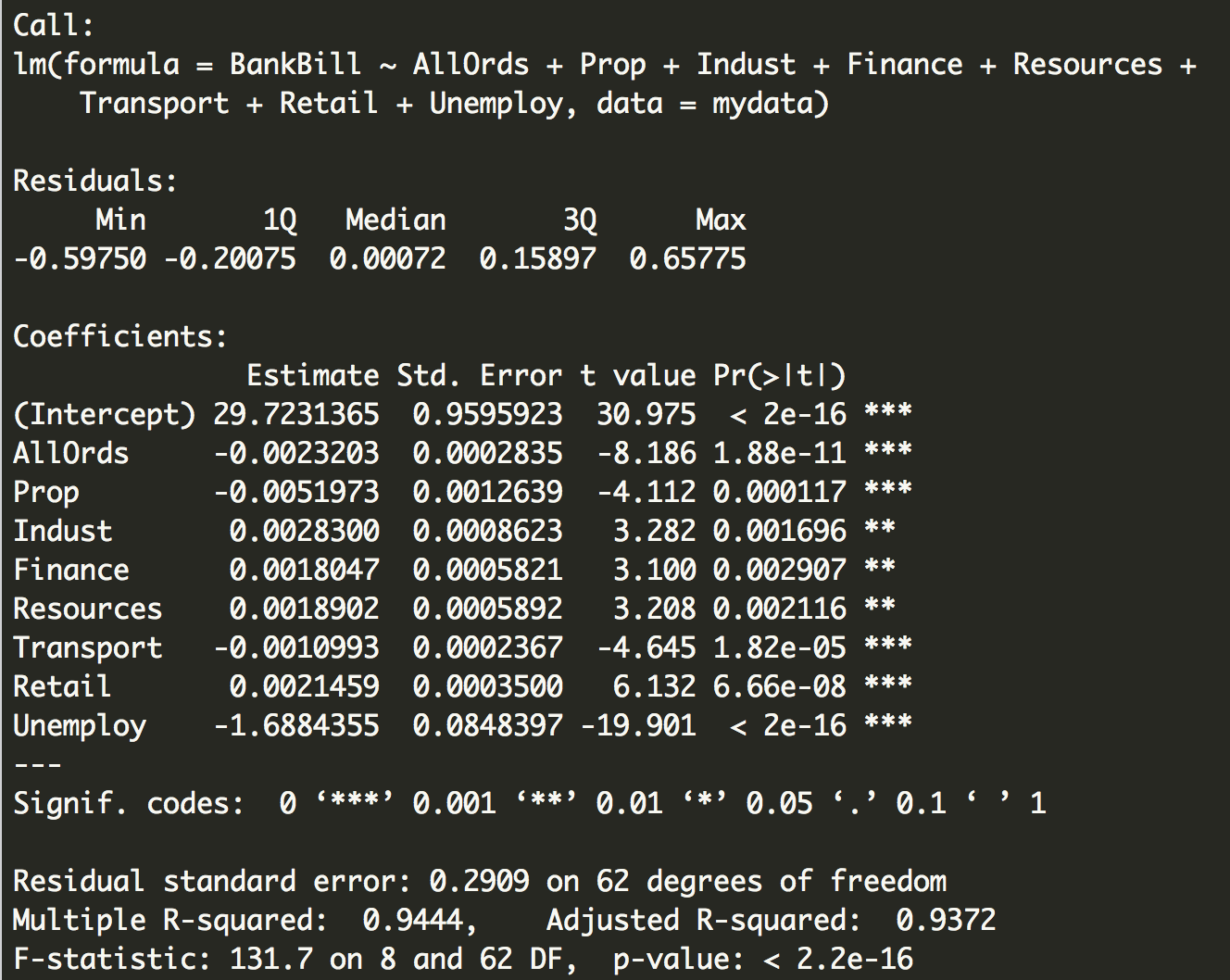
並非原 dataset中所有的 explanatory variables都具有關鍵作用：

1. 請找出「較關鍵」的一組explanatory variables。你可以先試用 lm()這個指令，推估選用不同組成的 explanatory variables 時，相對應 residual standard error 的大小，然後選一組你覺得代表「關鍵性」的 explanatory variables。

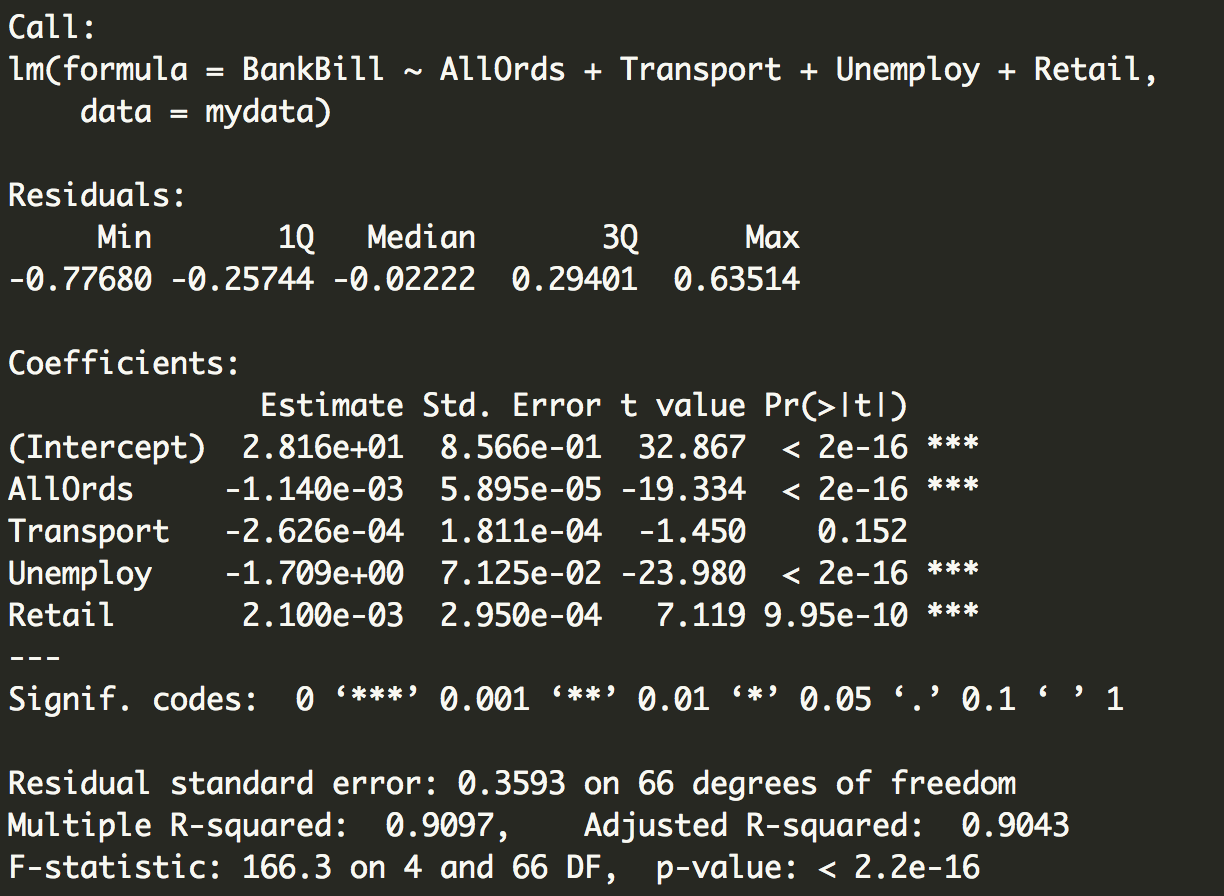
全部變數結果：



使用：以下變數所得到的residual standard error為0.2909



挑選最有相關的四個變數：residual standard error為0.3593

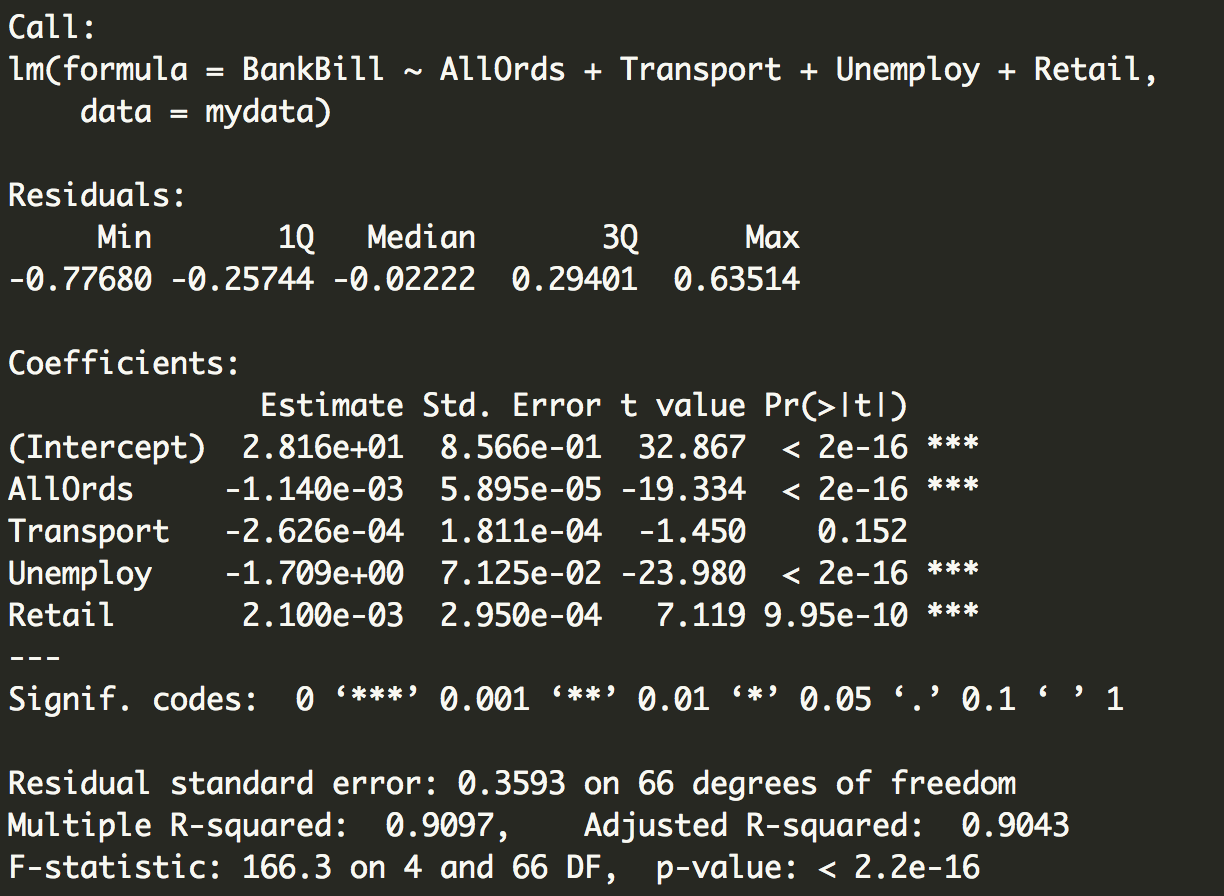


在以上結果中使用residual standard error來選我會使用全部變數，因為所得到的最小，變數越多誤差解釋能力應該要越好。

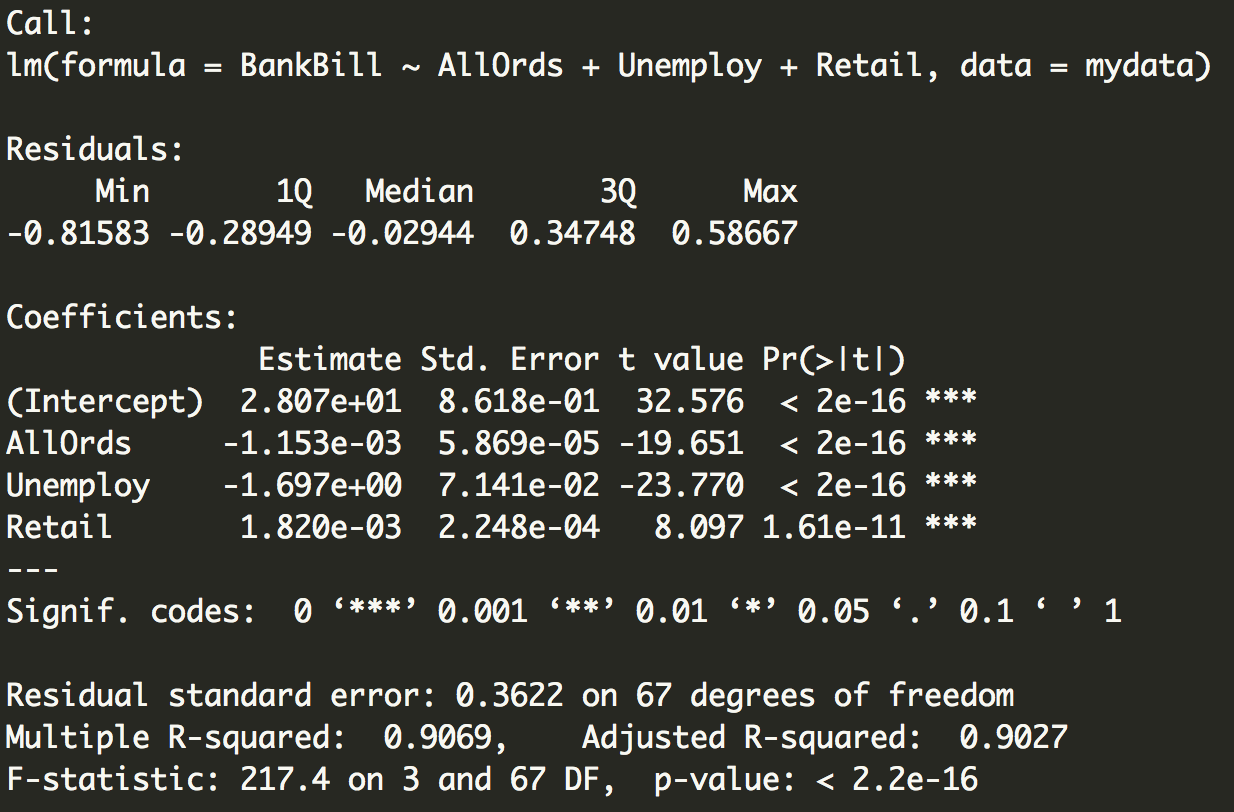
1. 針對同一組 dataset，請列出另一組你覺得也差不多好的 explanatory variables，解釋為什麼你最後不選這組？(你選取 explanatory variables 的準則是什麼？)

這是我最後所選擇最有相關的三個變數，我的選擇依據是相關性的星星指數與Pr(>|t|)的指標，星星指數越多代表這個變數是相關的，Pr(>|t|)為虛無假設中此變數與response variable不相關的比例，事實顯示Pr(>|t|)越低此變數與response variable是越相關的。

以下為全部變數中取前四名有相關性的：發現只與三個最有關係



以下三個變數是我要的！！！

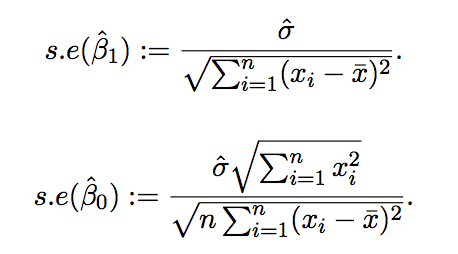


1. 解釋由 r code 的 lm所求出的模型
2. 按 2) 中所決定的那組 explanatory variables，代入 linear regression model，並解釋由 summary() 所得的各數字。

Residuals：提供估計後所剩下的殘差值，這個直我們希望越接近0越好，我所選擇的變數作出來的相當接近0這是我可以接受的。

Estimate：是估計所得的係數，股票指數與失業率是數值越大結果利率越低，零售指數則是數值越大利率越高。這裡我的B0是2.807e+01、AllOrds是-1.153e-03、Unemploy是-1.697e+00與Retail是1.820e-03。

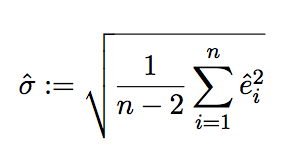
Std. Error：是估計係數所給的標準差區間，公式如下（第一次作業2(a)也有）



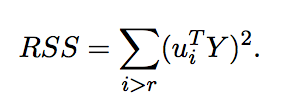
Pr(>|t|)：<0.05就可以代表非常顯著，其原理是透過上課所教導的虛無架設，分別有不同變數對response variable做假設情況下是沒有相關的，因此我們所得到的Pr(>|t|)越小代表沒有相關的機會越少，白話總結就是越小代表事實上非常有相關性，由此得知各個變數的值都<0.05，可以各個變數都是對模型貢獻上有相當重要性。

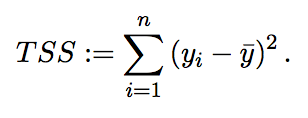
Residual standard error: 0.3622 on 67 degrees of freedom

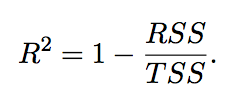
代表誤差的標準差公式如下，自由度是以71（資料筆數）-4（估計係數項目）=67所得的。



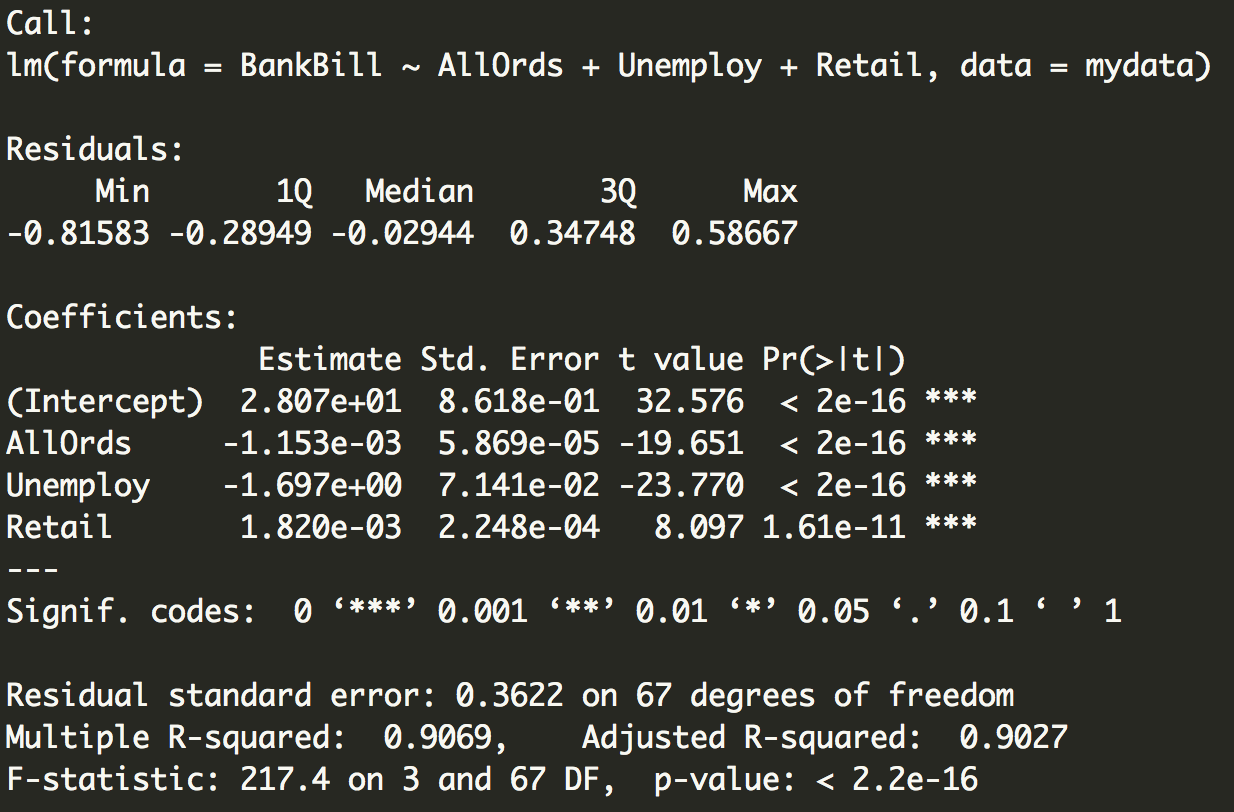
Multiple R-squared：公式如最下方，這個數值其實是會做落在0~1之間，如果越接近1代表RSS（殘差平方和）與TSS的比例比較小這是我們要的。結論是這邊所得到的Multiple R-squared: 0.9069與四個變數的Multiple R-squared: 0.9097，我會覺得三個變數是真正關鍵的。



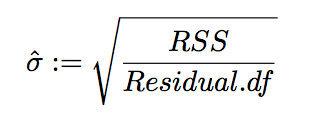




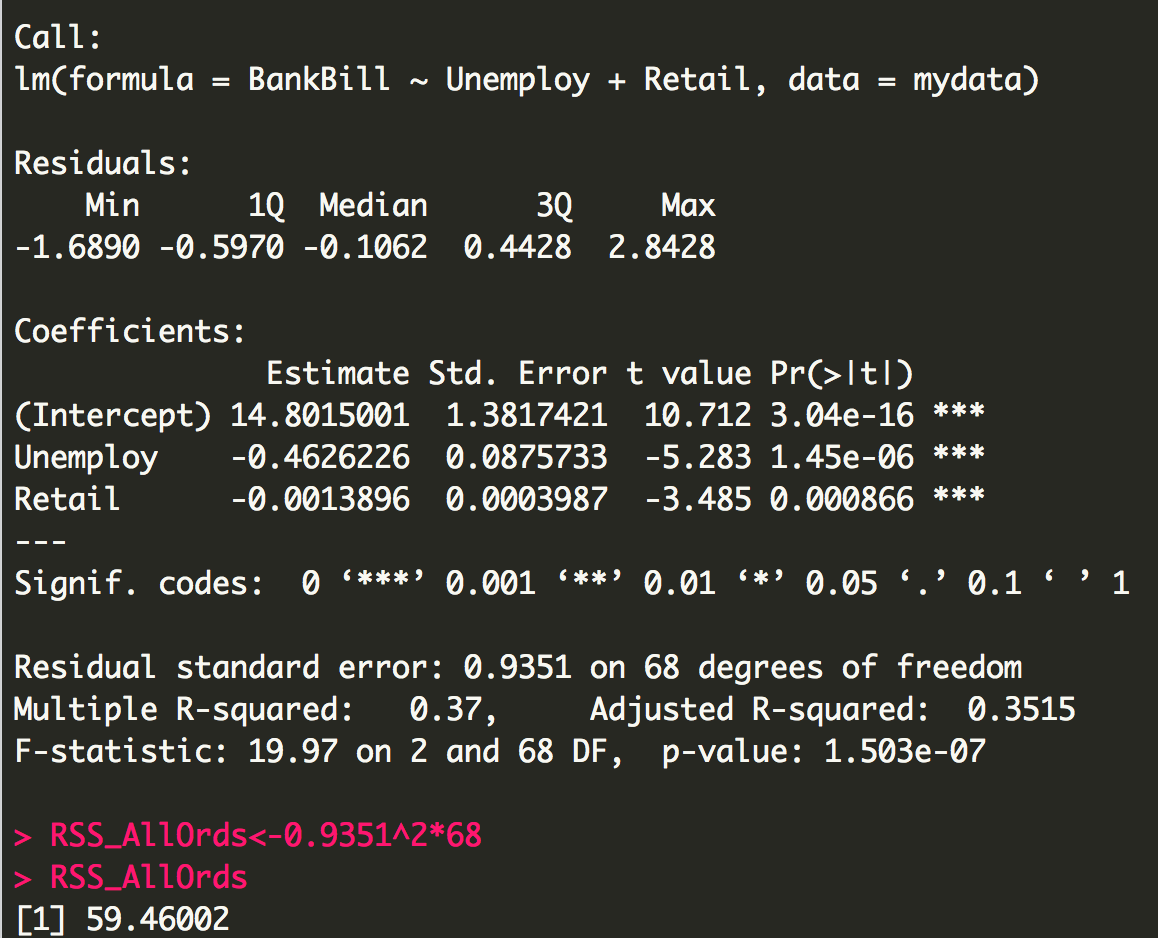
p-value: < 2.2e-16，其原理是透過上課所教導的虛無架設，對整個模型的變數對response variable做假設情況下是沒有相關的，因此我們所得到的p-value越小代表沒有相關的機會越少，白話總結就是越小代表事實上非常有相關性，從此得知這個變數組合是非常相關的。



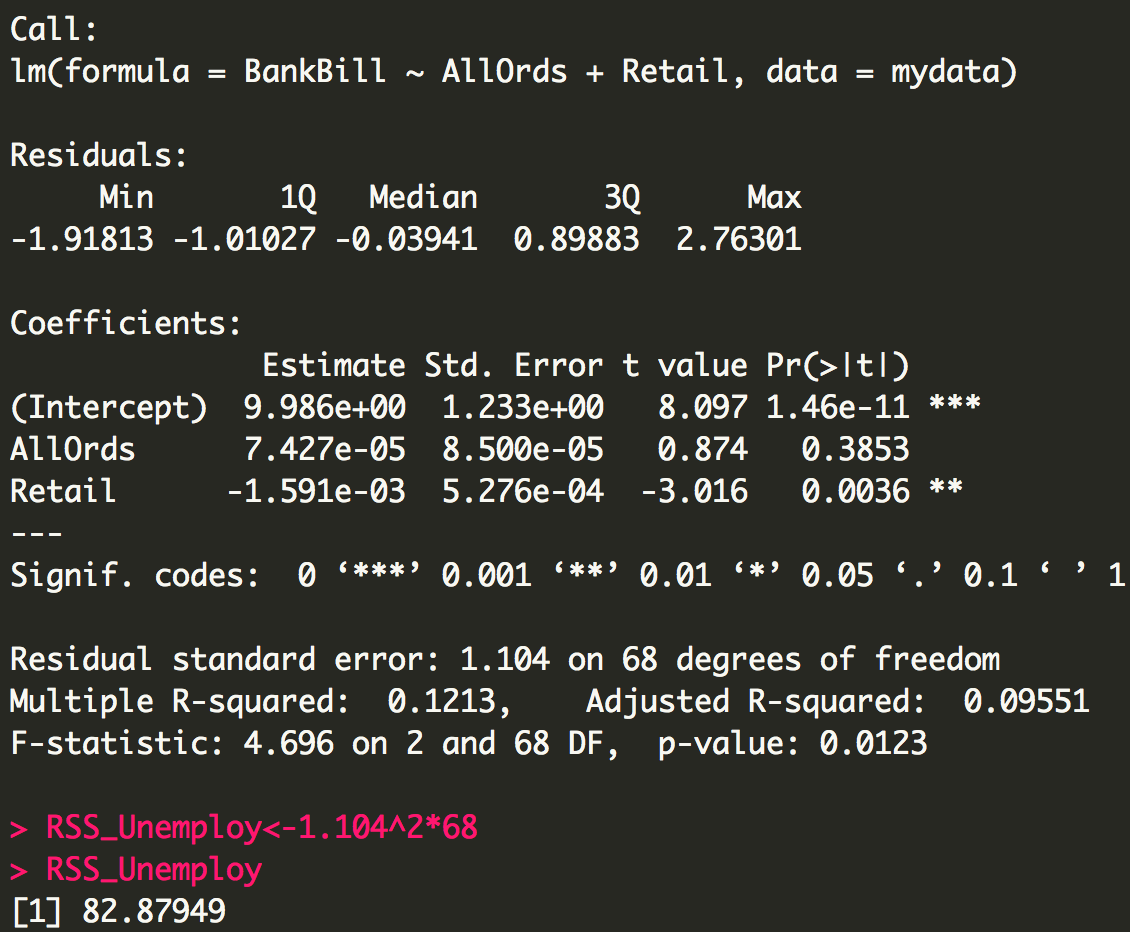
1. 請排列你所選取這組 explanatory variables，各別變數用來「解釋」response variable 的 power。(你可以試著拿掉其中某一項，看誰被拿掉之後，殘留的RSS較多。)



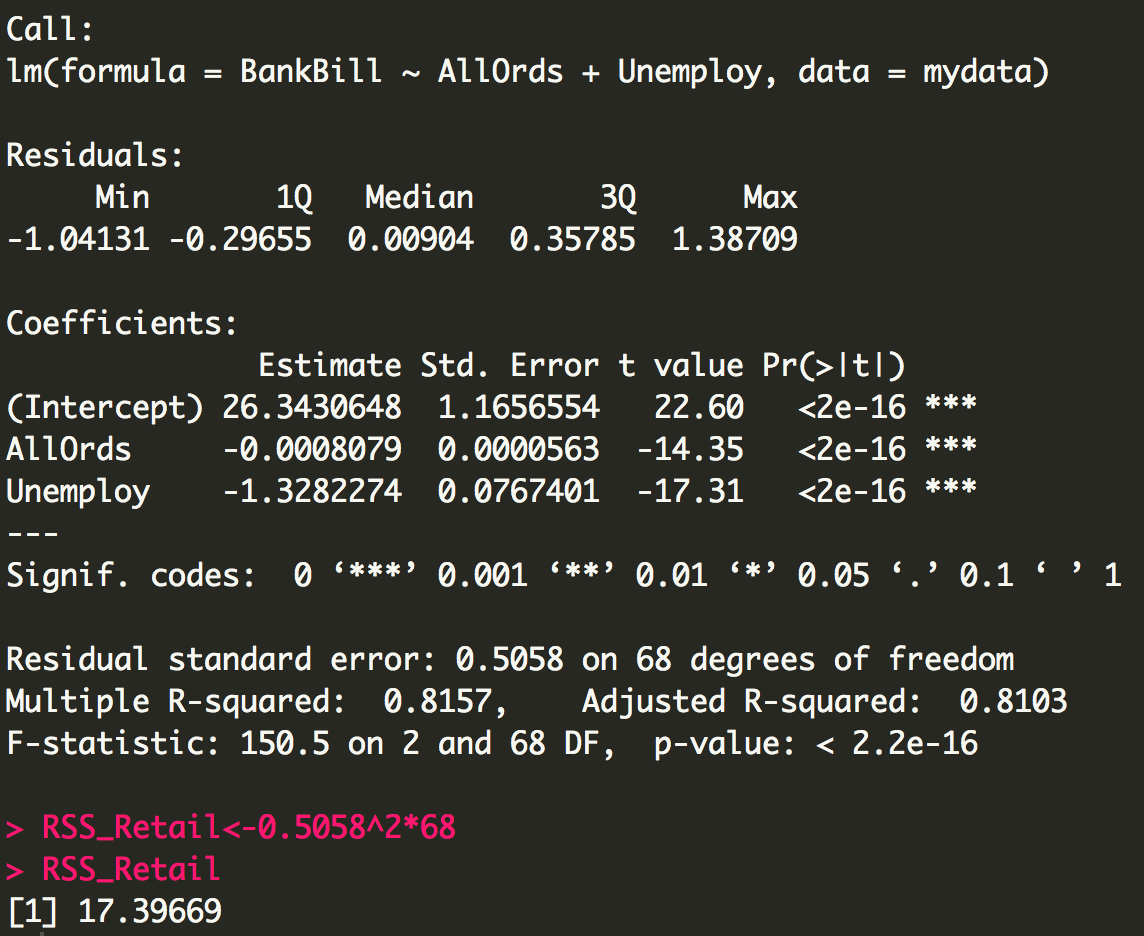
缺少：AllOrds-> RSS=59.46002



缺少：Unemploy-> RSS=82.87949



缺少：Retail-> RSS= 17.39669



結論可以知道單一變數來說Unemploy貢獻最多

1. 關於各explanatory variables 的 \beta\_i：
2. 計算出來的 \beta\hat 看起來合理嗎？(先看正負號，再看大小值，試著就其相對應的變數加減一，看相對應 y 的增減是否合理？)

(Intercept)：2.807e+01

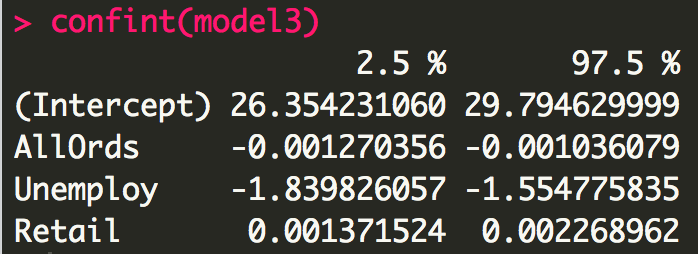
AllOrds：-1.153e-03，基本股指數越高利率越低，個人的財經看法是股市越好（報酬率）會有更多人想要融資，因此銀行一定會設法在股市好的時候放低利率使貸款又因增加，總結以上是合理的。（<https://www.cmoney.tw/notes/note-detail.aspx?nid=11867>）

Unemploy：-1.697e+00，失業率越高利率越低，利率上升，即貸款成本上升，造成企業減少投資，消費者減少消費，投資與消費的減少通過乘數效應使得社會總需求減少，從而社會總支出與總收入減少，進而導致失業率上升；反之，利率下降，投資增加，消費增加，總需求增加，若社會處于非充分就業狀態，則將導致就業增加，失業率下降，總結以上是合理的。

（<http://www.rocidea.com/roc-16758.aspx>）

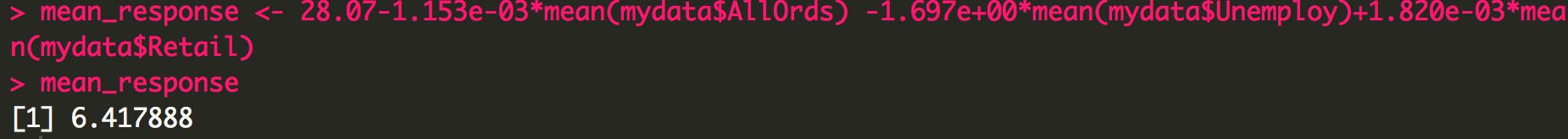
Retail：1.820e-03，零售指數越高利率越高，間接代表市場經濟變好，熱錢都在交易市場上流走，因此我推斷可能是利率政策不希望存款方的熱錢流走，也在經濟好的狀況下給貸款方的優惠變少，總結以上我認為是間接合理的。

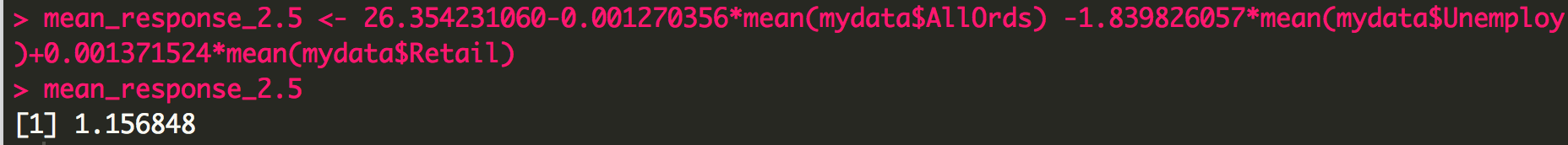
1. 求 \beta 的 95% 信心區間 (你可以用 confint()這個指令)



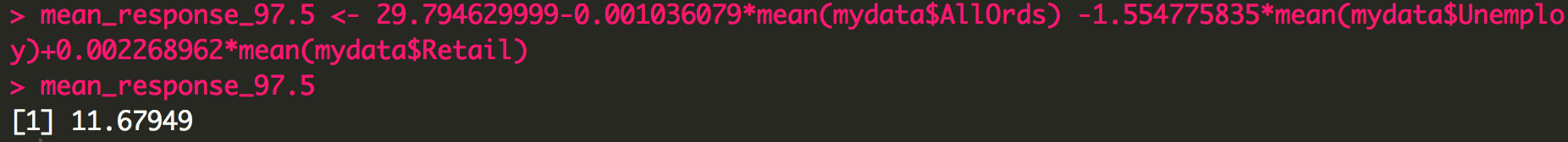
1. 由 dataset 中 explanatory variables 的合理範圍，取各 explanatory variable 的 mean的組合當作 x\_0，然後代入由 3) 求得的 model，求：
2. 對應 mean response，以及(包括 error) 之 response 的95% 的預測區間

mean response：



mean\_response\_2.5：

mean\_response\_97.5：



各個資料點的標準差區間：（如下圖）

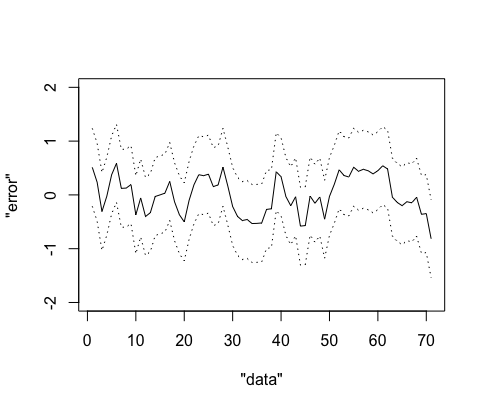
使用Residual standard error: 0.3622(由summary)

上限誤差為＋兩倍標準差up.bound=model3$residuals+2\*R.se

真實誤差為 error = model3$residuals

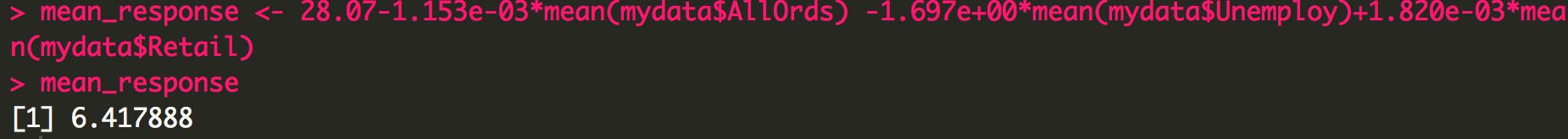
下限誤差為-兩倍標準差low.bound=model3$residuals-2\*R.se

真實的誤差應該要是越接近我們假設的0越好

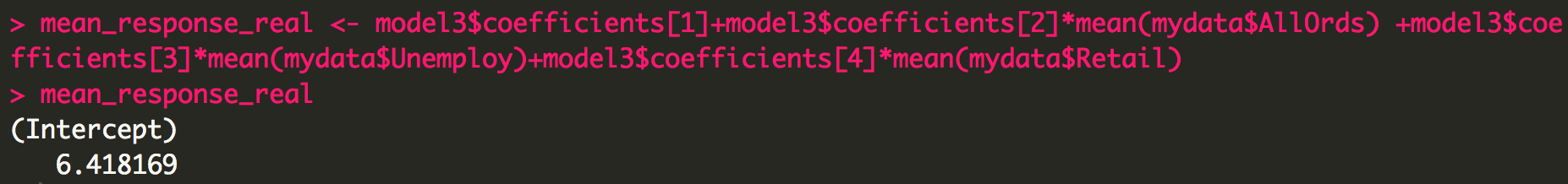


1. 請比較由 mean explanatory variables 組合的 x\_0所得的 y\_0與 dataset 中的 mean of y，你有觀察到什麼嗎？

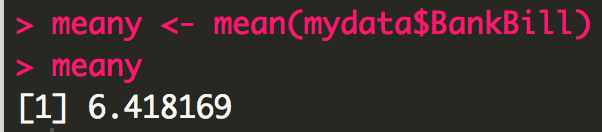
使用summary(model3)所得到的預測參數



使用model3$coefficients內的參數（小數點位數較準確），因此與真實的平均值超級像



由資料中的response variable所求得



結論：由summary所得到的參數與model3$coefficients所得到的參數差別在小數點的準確度上，因此數據model3$coefficients的與真實的完全相同，所以我可以把這三個變數所訓練的模型評斷為相當可信的。