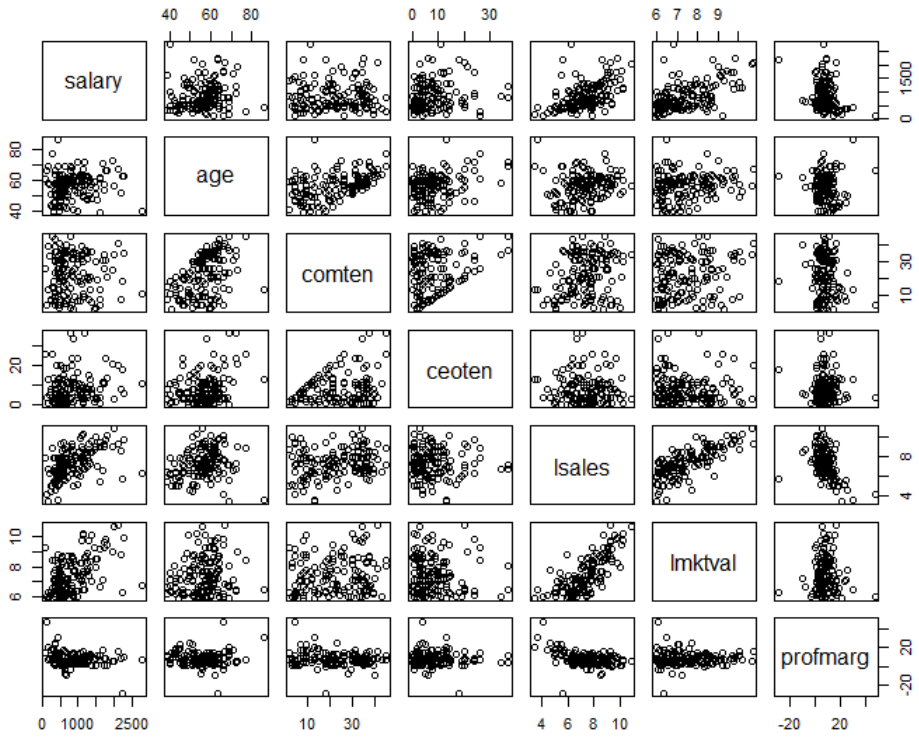
Những thông tin vê các giám đốc điều hành các tập đoàn Hoa Kỳ. Bộ dữ liệu gồm 177 quan trắc và 15 biến.

- Chọn mô hình phù hợp nhất giải thích biến phụ thuộc với từng bộ dữ liệu.
- Nêu rõ phương pháp chọn mô hình và lý do chọn phương pháp đó.
- Nói rõ ý nghĩa của mô hình đã chọn.

Tìm hiểu và tiền xử lý dữ liệu

Một số biến trong bộ dữ liệu kiểu số có đơn vị tính lớn như: *sales'*, *profits*, *mktval*. Nếu đưa những biến này vào phương trình hồi quy có thể dẫn tới hiện tượng bias do tác động của những biến này lên model lấn át những biến khác còn lại như *age*, *ceoten*.... Nên ta sẽ dùng phương pháp logarit cho 3 biến này trong model tương ứng với 3 biến mới là: *lsales"* , *lmktval* và *profmarg*. (1)

Từ biểu đồ dưới ta thấy ba biến định lượng *lsales*, *lmktval* và *profmarg* xảy ra hiện tượng đa cộng tuyến. Tuy nhiên có xảy ra hiện tượng đa cộng tuyến giữa 2 biến sales và profit luôn (Hình 1) Tính độ correlation của biến salary với lần lượt 2 biến trên ta có:



Hình 1: Mối tương quan giữa các biến

Xét bảng correlation giữa các biên độc lập với nhau và giữa các biến độc lập với biến phụ thuộc ta thấy: Giữa hai biến *lmktval* và biến *lsales* có mối tương quan rất cao (≈ 0.75). Tuy nhiên biến *lmktval* lại có mối tương quan cao hơn

```
> cor(train[c("salary", "lsales","lmktval","profmarg")])
             salary     lsales    lmktval    profmarg
salary    1.0000000  0.4912099 0.51978488 -0.24975911
lsales    0.4912099  1.0000000 0.75006264 -0.42949701
lmktval   0.5197849  0.7500626 1.00000000  0.04471558
profmarg -0.2497591 -0.4294970 0.04471558  1.00000000
```

Hình 2: Mức độ tương quan giữa biến lsales và promarg Correlation

với biến phụ thuộc *salary*. Mặt khác giữa biến *profmarg* và *lsales* cũng có mối tương quan cao (≈ -0.42). Nên ta loại bỏ biến *lsales* khỏi danh sách các biến được xét. (2)

Từ (1) và (2) ta có mô hình với đầy đủ các biến cần lựa chọn như sau:

$salary = \beta_0 + \beta_1 * age + \beta_2 * college + \beta_3 * grad + \beta_4 * comten + \beta_5 * ceoten + \beta_6 * lmktval + \beta_7 * profmarg$ (1.1)

Thực hiện phân rã hai biến phân loại gồm *college* và *grad* trước khi thực hiện phương pháp chọn biến StepWise tiến với tiêu chuẩn AIC.

Để đánh giá chất lượng mô hình ta chia tập dữ liệu thành hai phần training và testing với tỷ lệ 80:20 sau đó tiến hành phương pháp chọn biến.

Thực hiện chọn biến bằng phương pháp StepWise tiến và tiêu chuẩn AIC

```
[1] "salary"   "age"      "college"  "grad"     "comten"   "ceoten"   "lmktval"  "profmarg"
> l0 = lm(formula = train$salary ~ 1, data = train) # non independence variable
> l1 = lm(formula =  train$salary ~ ., data = train) # full independence variable
> modbest_Fow = step(l0, scope = list(lower = l0,
+                                     upper = l1), direction = 'forward', k =2)
Start:  AIC=1825.78
train$salary ~ 1

           Df Sum of Sq      RSS    AIC
+ lmktval   1  11242276 42481047 1794.4
+ profmarg  1    993901 52729423 1825.1
+ age       1    833601 52889723 1825.6
+ ceoten    1    816752 52906571 1825.6
+ comten    1    784116 52939207 1825.7
<none>                  53723323 1825.8
+ college   1    225711 53497612 1827.2
+ grad      1      1333 53721991 1827.8

Step:  AIC=1794.44
train$salary ~ lmktval

           Df Sum of Sq      RSS    AIC
+ profmarg  1   1319152 41161895 1792.0
+ ceoten    1   1069953 41411095 1792.8
<none>                  42481047 1794.4
+ grad      1    398593 42082454 1795.1
+ comten    1    199305 42281743 1795.8
+ age       1    177509 42303538 1795.8
+ college   1     90861 42390186 1796.1

Step:  AIC=1791.96
train$salary ~ lmktval + profmarg

           Df Sum of Sq      RSS    AIC
+ ceoten    1   1067048 40094847 1790.2
<none>                  41161895 1792.0
+ grad      1    215822 40946074 1793.2
+ age       1    170753 40991143 1793.4
+ college   1     92712 41069183 1793.6
+ comten    1     33866 41128029 1793.8

Step:  AIC=1790.23
train$salary ~ lmktval + profmarg + ceoten

           Df Sum of Sq      RSS    AIC
<none>                  40094847 1790.2
+ grad      1    142580 39952267 1791.7
+ college   1     38627 40056220 1792.1
+ comten    1     28636 40066211 1792.1
+ age       1         1 40094846 1792.2
```

Hình 3: Kết quả chọn biến theo phương pháp StepWise tiến với tiêu chuẩn AIC

Tông quan tieu chuẩn AIC thì mô hình tốt là mô hình có giá trị AIC nhỏ nhất. Ở mô hình 1 biến *lmktval* được chọn vào mô hình vì có AIC nhỏ nhất trong tất cả các kết hợp với các biến còn lại. Tương tự AIC được tính cho mô hình thêm biến thứ 2 biến *ceoten* và biến thứ 3 là *ceoten*.

```
> summary(modbest_Fow)

Call:
lm(formula = train$salary ~ lmktval + profmarg + ceoten, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-1339.1  -227.0   -72.8   163.7  4351.3

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -950.696    302.598  -3.142  0.00206 **
lmktval      248.204     38.909   6.379  2.5e-09 ***
profmarg     -13.929      6.544  -2.128  0.03508 *
ceoten        11.714      6.113   1.916  0.05738 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 539 on 138 degrees of freedom
Multiple R-squared:  0.2537,    Adjusted R-squared:  0.2375
F-statistic: 15.64 on 3 and 138 DF,  p-value: 8.262e-09
```

Hình 4: Kết quả hồi quy mô hình với các biên được chọn

Với ba biến trên được chọn mô hình (1.1) trở thành mô hình mới:

salary = -950.6 + 248.2 **lmktval* - -13.9 **profmarg* + 11.7 **ceoten* (1.2)

Tuy nhiên ta nhận thấy biến *ceoten* có $\rho_{value} \geq \alpha$ (0.05738 $\geq$ 0.05) nên không có ý nghĩa thống kê trong mô hình.

Ta tiến hành bỏ biến *ceoten* và hồi quy mô hình với hai biến còn lại kết quả thu được từ phần mềm R như bên dưới:

```
> new_train = train[c("salary", "lmktval", "profmarg")]
> newModel = lm(formula =new_train$salary ~ ., data = new_train )
> summary(newModel)

Call:
lm(formula = new_train$salary ~ ., data = new_train)

Residuals:
    Min      1Q  Median      3Q     Max
-1127.7  -256.6   -85.3   246.7  4404.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -830.739    298.886  -2.779   0.0062 **
lmktval      245.323     39.252   6.250 4.71e-09 ***
profmarg     -13.944      6.607  -2.111   0.0366 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 544.2 on 139 degrees of freedom
Multiple R-squared:  0.2338,    Adjusted R-squared:  0.2228
F-statistic: 21.21 on 2 and 139 DF,  p-value: 9.143e-09
```

Hình 5: Kết quả hồi quy mô hình với hai biến còn lại

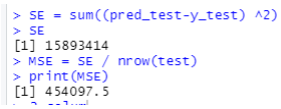
Mô hình thống kê mới:

salary = -830.7 + 245.3 **lmktval* -13.9 **profmarg* + 11.7 (1.3)

Trường hợp này hai biến còn lại có ý nghĩa thống kê. Tuy nhiên mô hình được tạo bởi hai biến này chỉ giải thích được 23 % kết quả biến phụ thuộc. Nguyên nhân dẫn tới kết quả thấp là do số lượng data ít, các biến giải thích ít không tạo nên mô hình đặc trưng được.

Test trên tập test và nhận xét kết quả

Thực hiện dự đoán trên tập dữ liệu test từ kết quả mô hình (1.3) và dùng chỉ số đánh MSE(trung bình sai số bình phương) ta có:



Hình 6: Chỉ số đo lường kết quả MSE