**B. Chọn mô hình với dữ liệu cho trước**

1. Chọn mô hình phù hợp nhát giải thích biến phụ thuộc với từng bộ dữ liệu sau. Phương pháp chọn và tiêu chuẩn chọn mô hình cho mỗi bộ dữ liệu là không trùng nhau

2. Nêu rõ phương pháp chọn mô hình và lý do chọn phương pháp đó

3. Nêu ý nghĩa của mô hình đã chọn

**Data 4: Tìm hiểu những yếu tố ảnh hưởng đến mức lương ($/giờ) của người lao động ở Anh năm 1976.**

**Nhập dữ liệu vào R**

library(readxl)

data4 <- read\_excel("data/data4.xls")

View(data4)

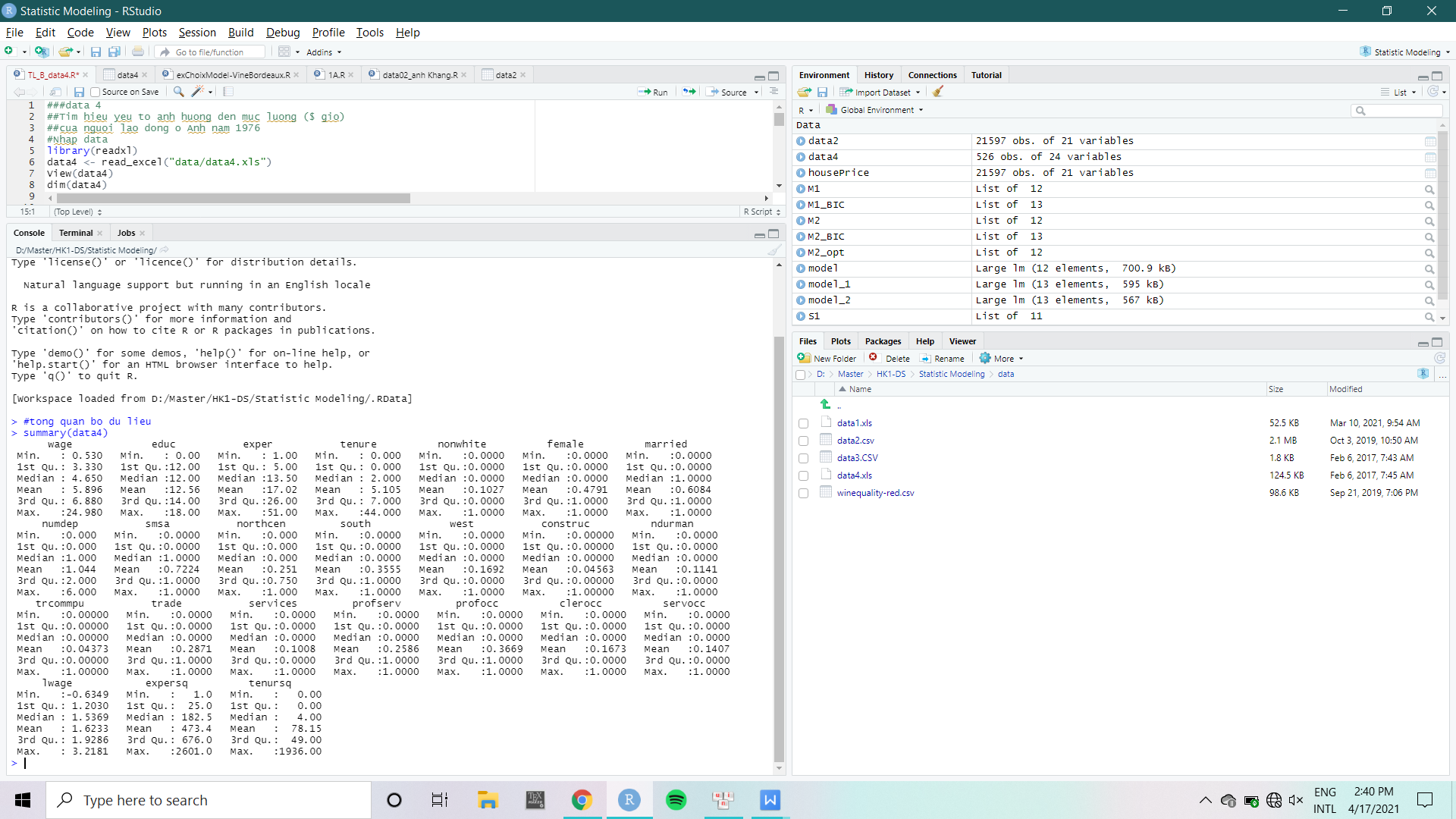
dim(data4)

#[1] 526 24

Dữ liệu có 526 quan trắc với 24 biến được mô tả như sau:

1. wage: tiền lương thu nhập mỗi giờ
2. educ: số năm giáo dục
3. exper: số năm kinh nghiệm
4. tenure: số năm làm việc với người chủ hiện tại
5. nonwhite: sắc tộc, 0 - người da trắng, 1 – không phải người da trắng
6. female: giới tính, 0 – nam, 1 - nữ
7. tình trạng hôn nhân: 0 – chưa kết hôn, 1 – đã kết hôn
8. numdep: số lượng người phụ thuộc
9. smsa: sống ở khu đô thị tiêu chuẩn (smsa), 0 – không, 1 - có
10. northcen: sống ở phía bắc trung tâm mỹ, 0 – không, 1 - có
11. south: sống ở khu vực phia nam, 0 – không, 1 - có
12. west: sống ở khu vực phía tây, 0 – không, 1 - có
13. construc: ngành xây dựng, 0 – không, 1 – có
14. ndurman: ngành sản xuất không bền, 0 – không, 1 – có
15. trcommpu: ngành giao thông vận tải, thông tin liên lạc, tiện ích công cộng, 0 – không, 1 – có
16. trade: ngành buôn bán, 0 – không, 1 – có
17. services: công nghiệp dịch vụ, 0 – không, 1 – có
18. profserv: ngành dịch vụ chuyên nghiệp, 0 – không, 1 – có
19. profocc: nghề nghiệp chuyên môn, 0 – không, 1 – có
20. clerocc: nghề văn thư , 0 – không, 1 – có
21. servocc: nghề dịch vụ, 0 – không, 1 – có
22. lwage: log(wage)
23. expersq: exper2
24. tenursq: tenure2

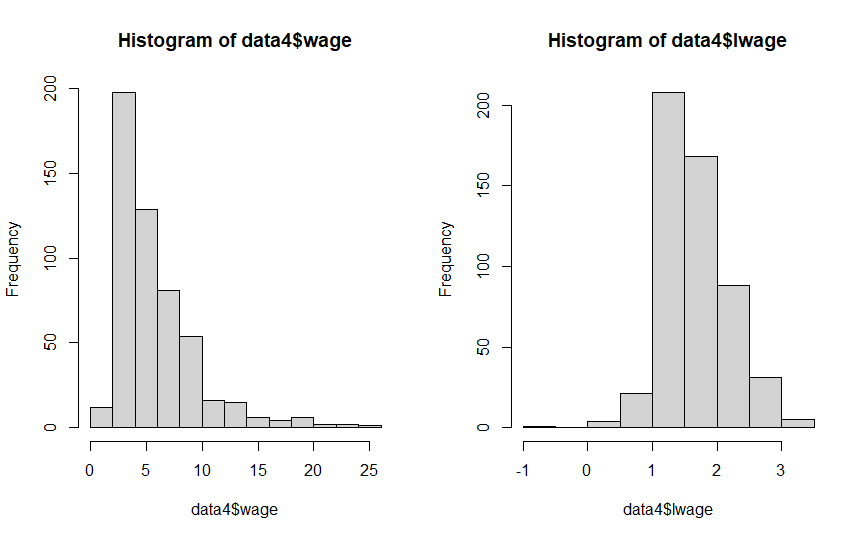
Tổng quan bộ dữ liệu:



Thống kê mô tả bằng đồ thị histogram của hai biến theo dõi wage và lwage cho thấy cả hai đều có dấu bị lệch.

hist(data4$wage)

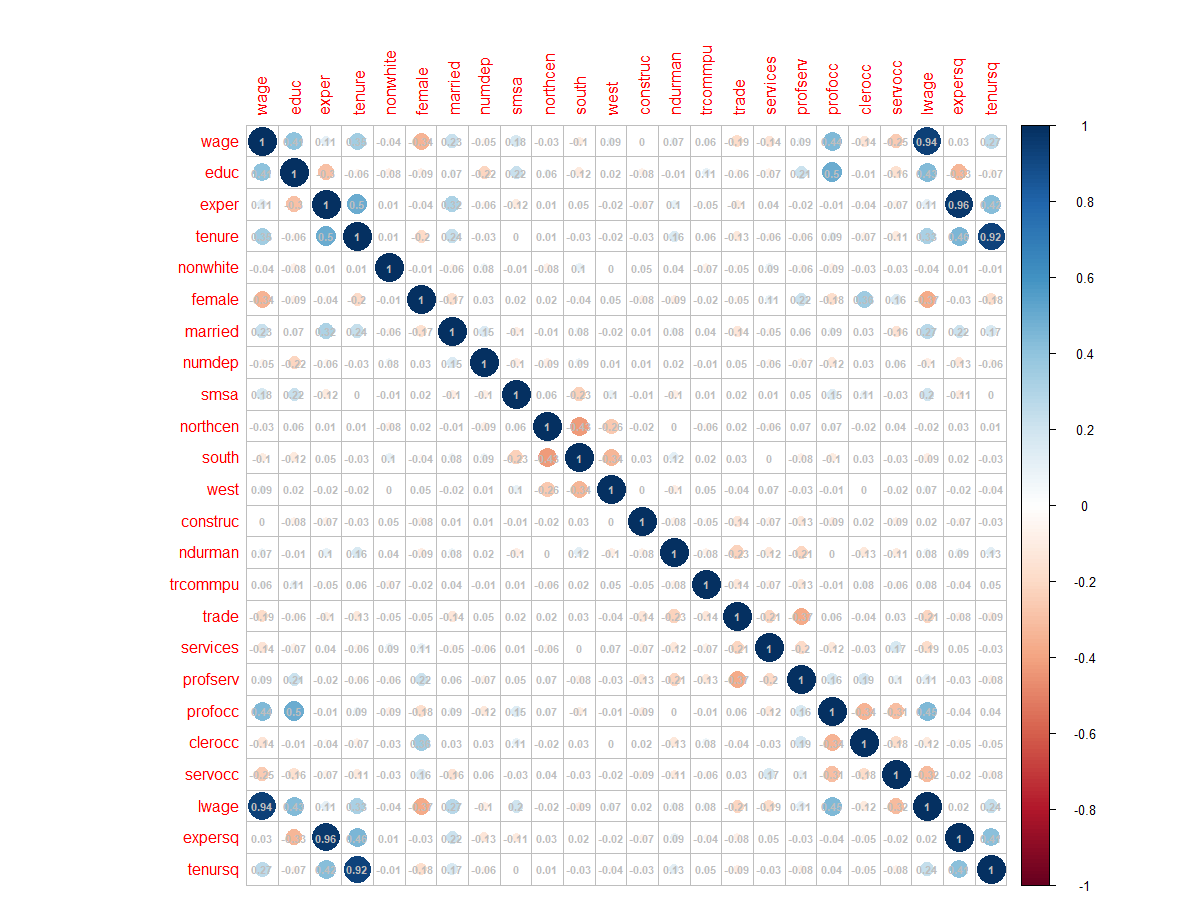
hist(data4$lwage)



Xét hệ số tương quan của các biến trong mô hình, ta có thể thấy hiện tượng đa cộng tuyến không xuất hiện quá nặng. Các biến có thể thấy mối liên quan ở lĩnh vực đào tạo là educ và profocc.

cor(data4)

corrplot::corrplot(cor(data4), addCoef.col = "grey", number.cex = 0.7)



**Mô hình hồi quy tuyến tính đầy đủ theo biến wage:**

> M1 <- lm(wage ~ . -lwage, data = data4)

> summary(M1)

Call:

lm(formula = wage ~ . - lwage, data = data4)

Residuals:

Min 1Q Median 3Q Max

-6.7816 -1.5113 -0.2980 0.9799 13.2006

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.8033646 0.8590539 0.935 0.350148

educ 0.3264568 0.0576139 5.666 2.46e-08 \*\*\*

exper 0.1650197 0.0374725 4.404 1.30e-05 \*\*\*

tenure 0.1568149 0.0471954 3.323 0.000956 \*\*\*

nonwhite -0.0846544 0.3985055 -0.212 0.831858

female -1.6295492 0.2738709 -5.950 5.03e-09 \*\*\*

married 0.1003878 0.2837037 0.354 0.723602

numdep -0.0022417 0.1045880 -0.021 0.982908

smsa 0.7094254 0.2826284 2.510 0.012383 \*

northcen -0.5668206 0.3437244 -1.649 0.099761 .

south -0.4482182 0.3297238 -1.359 0.174637

west 0.4380886 0.3825645 1.145 0.252697

construc -0.5050102 0.6365111 -0.793 0.427917

ndurman -0.8074112 0.4729921 -1.707 0.088434 .

trcommpu -1.0384439 0.6602717 -1.573 0.116405

trade -2.0302099 0.3973100 -5.110 4.59e-07 \*\*\*

services -1.7626229 0.5042524 -3.496 0.000515 \*\*\*

profserv -0.9333546 0.4357896 -2.142 0.032693 \*

profocc 1.8908140 0.3578277 5.284 1.88e-07 \*\*\*

clerocc 0.3351191 0.4221366 0.794 0.427649

servocc 0.0042979 0.4231809 0.010 0.991901

expersq -0.0034346 0.0007991 -4.298 2.07e-05 \*\*\*

tenursq -0.0017088 0.0016064 -1.064 0.287936

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.69 on 503 degrees of freedom

Multiple R-squared: 0.4917, Adjusted R-squared: 0.4695

F-statistic: 22.12 on 22 and 503 DF, p-value: < 2.2e-16

**Mô hình hồi quy tuyến tính đầy đủ theo biến lwage:**

> M2 <- lm(lwage ~ . -wage, data = data4)

> summary(M2)

Call:

lm(formula = lwage ~ . - wage, data = data4)

Residuals:

Min 1Q Median 3Q Max

-1.70463 -0.21176 -0.01842 0.21099 1.24324

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.8931247 0.1162617 7.682 8.25e-14 \*\*\*

educ 0.0467910 0.0077973 6.001 3.76e-09 \*\*\*

exper 0.0254056 0.0050714 5.010 7.56e-07 \*\*\*

tenure 0.0223215 0.0063873 3.495 0.000517 \*\*\*

nonwhite -0.0042677 0.0539325 -0.079 0.936960

female -0.2679739 0.0370648 -7.230 1.81e-12 \*\*\*

married 0.0562608 0.0383956 1.465 0.143465

numdep -0.0215152 0.0141546 -1.520 0.129136

smsa 0.1387299 0.0382500 3.627 0.000316 \*\*\*

northcen -0.0584407 0.0465186 -1.256 0.209595

south -0.0444269 0.0446238 -0.996 0.319929

west 0.0545441 0.0517751 1.053 0.292626

construc -0.0528536 0.0861434 -0.614 0.539787

ndurman -0.1074388 0.0640133 -1.678 0.093893 .

trcommpu -0.0961487 0.0893591 -1.076 0.282452

trade -0.3032698 0.0537707 -5.640 2.84e-08 \*\*\*

services -0.3091468 0.0682439 -4.530 7.37e-06 \*\*\*

profserv -0.0951315 0.0589784 -1.613 0.107374

profocc 0.2248381 0.0484273 4.643 4.39e-06 \*\*\*

clerocc 0.0383129 0.0571306 0.671 0.502771

servocc -0.0944223 0.0572720 -1.649 0.099841 .

expersq -0.0005294 0.0001081 -4.895 1.32e-06 \*\*\*

tenursq -0.0003734 0.0002174 -1.718 0.086475 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.364 on 503 degrees of freedom

Multiple R-squared: 0.5506, Adjusted R-squared: 0.5309

F-statistic: 28.01 on 22 and 503 DF, p-value: < 2.2e-16

So sánh hai mô hình mức lương wage và lwage - log(wage):

> summary(M1)$adj.r.squared

[1] 0.4694756

> summary(M2)$adj.r.squared

1. 0.5309194

Mô hình hồi quy tuyến tính của lwage có hệ số xác định R2 hiệu chỉnh là 53.09% cao hơn mô hình wage tương ứng là 46.95%, nên ta chọn mô hình theo biến lwage giải thích cho mức lương ($/giờ) của người lao động ở Anh năm 1976.

Theo kết quả xuất ra từ R, mô hình biến đầy đủ theo lwage còn khá nhiều biến không có ý nghĩa thống kê, nên ta tiến hành chọn biến mục đích tìm được mô hình đơn giản hơn nhưng vẫn có đầy đủ thông tin có ý nghĩa, với cách áp dụng phương pháp stepwise và direction là backward từ mô hình đầy đủ biến ban đầu.

Chọn mô hình theo biến lwage bằng phương pháp stepwise với tiêu chuẩn AIC:

> M2\_AIC <- MASS::stepAIC(M2, k=2, direction = "backward", trace = FALSE)

> summary(M2\_AIC)

Call:

lm(formula = lwage ~ educ + exper + tenure + female + smsa +

west + trade + services + profocc + servocc + expersq + tenursq,

data = data4)

Residuals:

Min 1Q Median 3Q Max

-1.69434 -0.21212 -0.01744 0.21211 1.27158

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.651e-01 1.016e-01 7.529 2.32e-13 \*\*\*

educ 4.962e-02 7.246e-03 6.848 2.15e-11 \*\*\*

exper 2.652e-02 4.596e-03 5.769 1.38e-08 \*\*\*

tenure 2.128e-02 6.326e-03 3.364 0.000825 \*\*\*

female -2.780e-01 3.325e-02 -8.362 5.85e-16 \*\*\*

smsa 1.501e-01 3.677e-02 4.081 5.19e-05 \*\*\*

west 9.743e-02 4.278e-02 2.277 0.023185 \*

trade -2.402e-01 3.699e-02 -6.493 1.99e-10 \*\*\*

services -2.372e-01 5.574e-02 -4.256 2.48e-05 \*\*\*

profocc 1.997e-01 4.054e-02 4.927 1.13e-06 \*\*\*

servocc -1.291e-01 4.952e-02 -2.606 0.009427 \*\*

expersq -5.357e-04 9.865e-05 -5.430 8.70e-08 \*\*\*

tenursq -3.260e-04 2.157e-04 -1.511 0.131292

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3637 on 513 degrees of freedom

Multiple R-squared: 0.5424, Adjusted R-squared: 0.5317

F-statistic: 50.67 on 12 and 513 DF, p-value: < 2.2e-16

Chọn mô hình theo biến lwage bằng phương pháp stepwise với tiêu chuẩn BIC:

> M2\_BIC <- MASS::stepAIC(M2, k=log(nrow(data4)), direction = "backward", trace = FALSE)

> summary(M2\_BIC)

Call:

lm(formula = lwage ~ educ + exper + tenure + female + smsa +

trade + services + profocc + servocc + expersq, data = data4)

Residuals:

Min 1Q Median 3Q Max

-1.6159 -0.2162 -0.0198 0.2135 1.2650

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.806e-01 1.021e-01 7.649 1.00e-13 \*\*\*

educ 4.951e-02 7.288e-03 6.793 3.03e-11 \*\*\*

exper 2.854e-02 4.460e-03 6.400 3.50e-10 \*\*\*

tenure 1.252e-02 2.651e-03 4.723 3.00e-06 \*\*\*

female -2.749e-01 3.341e-02 -8.229 1.56e-15 \*\*\*

smsa 1.582e-01 3.681e-02 4.298 2.06e-05 \*\*\*

trade -2.461e-01 3.710e-02 -6.635 8.22e-11 \*\*\*

services -2.376e-01 5.570e-02 -4.266 2.36e-05 \*\*\*

profocc 2.030e-01 4.066e-02 4.993 8.17e-07 \*\*\*

servocc -1.324e-01 4.976e-02 -2.660 0.00805 \*\*

expersq -5.762e-04 9.635e-05 -5.981 4.16e-09 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3658 on 515 degrees of freedom

Multiple R-squared: 0.5354, Adjusted R-squared: 0.5264

F-statistic: 59.34 on 10 and 515 DF, p-value: < 2.2e-16

Từ kết quả trên ta thấy mô hình được chọn theo tiêu chuẩn AIC giữ lại 12 biến có R2 hiệu chình là 53.17% cao hơn so với mô hình được chọn bằng tiêu chuẩn BIC có R2 hiệu chỉnh là 52.64% chỉ giữ lại 10 biến. Tuy nhiên, mô hình được chọn bằng phương pháp AIC có biến tenursq không có ý nghĩa thống kê, giá trị p-value > 0.1, chỉ số BIC loại biến với mức phạt nặng hơn, mô hình được chọn bằng tiêu chuẩn BIC loại biến tenuresq, các biến trong mô hình BIC có ý nghĩa thống kê, nên ta chọn mô hình của biến lwage theo tiêu chuẩn BIC.

So sánh mô hình xuất ra bằng phương pháp stepwise và mô hình đầy đủ ban đầu với kiểm định Fisher từng phần:

Giả thuyết H0: Mô hình đã giảm biến theo tiêu chuẩn BIC

Giả thuyết đối H1: Mô hình đầy đủ biến

Bảng anova so sánh giữa hai mô hình:

> anova(M2\_BIC,M2)

Analysis of Variance Table

Model 1: lwage ~ educ + exper + tenure + female + smsa + trade + services +

profocc + servocc + expersq

Model 2: lwage ~ (wage + educ + exper + tenure + nonwhite + female + married +

numdep + smsa + northcen + south + west + construc + ndurman +

trcommpu + trade + services + profserv + profocc + clerocc +

servocc + expersq + tenursq) - wage

Res.Df RSS Df Sum of Sq F Pr(>F)

1 515 68.918

2 503 66.663 12 2.2546 1.4176 0.1536

Giá trị p-value của kiểm định là 0.1536 > α = 0.05, với mức ý nghĩa 5%, không đủ cơ sở bác bỏ giả thuyết H0, nghĩa là chấp nhận mô hình hồi quy tuyến tính của lwage được rút gọn.

Xem xét hệ số phóng đại phương sai của mô hình hồi quy tuyến tính theo biến lwage đã chọn.

> car::vif(M2\_BIC)

educ exper tenure female smsa trade services profocc servocc expersq

1.597530 14.374639 1.438833 1.094659 1.068110 1.107096 1.104827 1.509386 1.176789 13.821237

Từ kết quả trên, ta thấy hệ số của biến exper và exersq lần lượt là 14.374639 và 13.821237 đều cao, nguyên nhân do biến expersq là bình phương của biến exper. Ta xem xét việc bỏ một trong hai biến exper hoặc expersq khỏi mô hình.

Mô hình lwage giữ biến exper loại biến expersq.

> M2\_exper <- lm(formula = lwage ~ educ + exper + tenure + female + smsa +

+ trade + services + profocc + servocc, data = data4)

> summary(M2\_exper)

Call:

lm(formula = lwage ~ educ + exper + tenure + female + smsa +

trade + services + profocc + servocc, data = data4)

Residuals:

Min 1Q Median 3Q Max

-1.65921 -0.23787 -0.01653 0.19786 1.38212

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.885035 0.103879 8.520 < 2e-16 \*\*\*

educ 0.054433 0.007481 7.276 1.28e-12 \*\*\*

exper 0.003334 0.001505 2.215 0.027200 \*

tenure 0.013414 0.002734 4.906 1.25e-06 \*\*\*

female -0.274321 0.034512 -7.948 1.20e-14 \*\*\*

smsa 0.148826 0.037998 3.917 0.000102 \*\*\*

trade -0.255611 0.038292 -6.675 6.38e-11 \*\*\*

services -0.239716 0.057542 -4.166 3.63e-05 \*\*\*

profocc 0.203975 0.042006 4.856 1.59e-06 \*\*\*

servocc -0.172110 0.050954 -3.378 0.000786 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3779 on 516 degrees of freedom

Multiple R-squared: 0.5031, Adjusted R-squared: 0.4944

1. statistic: 58.05 on 9 and 516 DF, p-value: < 2.2e-16

Thực hiện kiểm định Fisher từng phần cho mô hình loại biến expersq.

Giả thuyết H0: Mô hình lwage giữ biến exper loại biến expersq

Giả thuyết đối H1: Mô hình lwage chọn theo BIC

> anova(M2\_exper, M2\_BIC)

Analysis of Variance Table

Model 1: lwage ~ educ + exper + tenure + female + smsa + trade + services +

profocc + servocc

Model 2: lwage ~ educ + exper + tenure + female + smsa + trade + services +

profocc + servocc + expersq

Res.Df RSS Df Sum of Sq F Pr(>F)

1 516 73.704

2 515 68.918 1 4.7865 35.768 4.16e-09 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Giá trị p-value từ bảng anova là 4.16e-09 rất nhỏ, do đó ta bác bỏ H0, nghĩa là không thể loại bỏ biến expersq khỏi mô hình, thêm vào đó, khi loại biến expersq, mô hình có R2 hiệu chỉnh giảm còn 49.44%.

Mô hình lwage giữ biến expersq loại biến exper.

> M2\_expersq <- lm(formula = lwage ~ educ + tenure + female + smsa +

+ trade + services + profocc + servocc + expersq, data = data4)

> summary(M2\_expersq)

Call:

lm(formula = lwage ~ educ + tenure + female + smsa + trade +

services + profocc + servocc + expersq, data = data4)

Residuals:

Min 1Q Median 3Q Max

-1.69369 -0.22682 -0.01948 0.21069 1.39308

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 9.852e-01 1.006e-01 9.794 < 2e-16 \*\*\*

educ 4.965e-02 7.564e-03 6.563 1.28e-10 \*\*\*

tenure 1.611e-02 2.689e-03 5.992 3.88e-09 \*\*\*

female -2.705e-01 3.467e-02 -7.803 3.38e-14 \*\*\*

smsa 1.427e-01 3.813e-02 3.743 0.000202 \*\*\*

trade -2.607e-01 3.843e-02 -6.783 3.23e-11 \*\*\*

services -2.321e-01 5.781e-02 -4.016 6.81e-05 \*\*\*

profocc 2.147e-01 4.216e-02 5.093 4.95e-07 \*\*\*

servocc -1.785e-01 5.111e-02 -3.491 0.000522 \*\*\*

expersq 6.574e-06 3.267e-05 0.201 0.840626

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3797 on 516 degrees of freedom

Multiple R-squared: 0.4984, Adjusted R-squared: 0.4897

F-statistic: 56.97 on 9 and 516 DF, p-value: < 2.2e-16

Từ kết quả trên, nhận thấy khi loại biến exper ra khỏi mô hình, biến expersq không có ý nghĩa thống kê, R2 hiệu chỉnh giảm còn 48.97%.

Xét kiểm định Fisher từng phần cho mô hình giảm biến exper.

Giả thuyết H0: Mô hình lwage giữ biến expersq loại biến exper

Giả thuyết đối H1: Mô hình lwage chọn theo BIC

Analysis of Variance Table

Model 1: lwage ~ educ + tenure + female + smsa + trade + services + profocc +

servocc + expersq

Model 2: lwage ~ educ + exper + tenure + female + smsa + trade + services +

profocc + servocc + expersq

Res.Df RSS Df Sum of Sq F Pr(>F)

1 516 74.399

2 515 68.918 1 5.4814 40.961 3.496e-10 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Từ bảng anova so sánh giữa hai mô hình, giá trị p-value 3.496e-10 rất nhỏ, nên ta bác bỏ giả thuyết H0, nghĩa là không thể loại biến exper ra khỏi mô hình đã chọn bằng tiêu chuẩn BIC.

**Mô hình hồi quy tuyến tính được chọn theo biến lwage áp dụng phương pháp chọn biến stepwise theo tiêu chuẩn BIC.**

> coef(M2\_BIC)

(Intercept) educ exper tenure female smsa trade services

0.7806121337 0.0495069819 0.0285441179 0.0125198447 -0.2748832520 0.1582349049 -0.2461485439 -0.2376286536

profocc servocc expersq

0.2029898513 -0.1323915292 -0.0005762231

lwage = 0.7806121337 + 0.0495069819×educ + 0.0285441179×exper + 0.0125198447×tenure -0.2748832520×female + 0.1582349049×smsa - 0.2461485439×trade - 0.2376286536×services + 0.2029898513×profocc - 0.1323915292×servocc - 0.0005762231×expersq