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: làm việc ngành xây dựng, 0 – không, 1 – có
14. ndurman: làm việc ngành nondur.manuf.indus, 0 – không, 1 – có
15. trcommpu: làm việc ngành trans, commun, pub ut, 0 – không, 1 – có

16. trade: làm việc ngành buôn bán, 0 – không, 1 – có
17. services: làm việc ngành dịch vụ, 0 – không, 1 – có
18. profserv: làm prof. serv. indus, 0 – không, 1 – có
19. profocc: làm profess. occupation, 0 – không, 1 – có
20. clerocc: làm clerical occupation, 0 – không, 1 – có
21. servocc: làm service occupation, 0 – không, 1 – có
22. lwage: log(wage)
23. expersq: $exper^2$
24. tenursq: $tenure^2$

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

```
> M1 <- lm(wage ~ educ + exper + tenure + nonwhite + female + married + numdep + smsa + northcen + south + west + construc + ndurman + trcommpu + trade + services + profserv + profocc + clerocc + servocc + expersq + tenursq, data = data4)
> summary(M1)


Call:
lm(formula = wage ~ educ + exper + tenure + nonwhite + female +
    married + numdep + smsa + northcen + south + west + construc +
    ndurman + trcommpu + trade + services + profserv + profocc +
    clerocc + servocc + expersq + tenursq, 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
```

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

```
> M1_BIC <- MASS::stepAIC(M1, k=log(nrow(data4)), direction = "backward", trace = FALSE)
> summary(M1_BIC)
```

```
Call:
lm(formula = wage ~ educ + exper + tenure + female + smsa + trade +
    services + profocc + expersq, data = data4)

Residuals:
   Min      1Q  Median      3Q     Max
-6.9946 -1.4938 -0.3846  1.0135 13.1706

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0871112  0.7471923   0.117  0.90723
educ         0.3138293  0.0538335   5.830 9.80e-09 ***
exper        0.1867113  0.0326024   5.727 1.74e-08 ***
tenure       0.1134712  0.0195821   5.795 1.19e-08 ***
female      -1.6730322  0.2457156  -6.809 2.74e-11 ***
smsa         0.8904848  0.2719577   3.274  0.00113 **
trade       -1.4417386  0.2734359  -5.273 1.98e-07 ***
services    -1.1330508  0.4070236  -2.784  0.00557 **
profocc      1.7398764  0.2926449   5.945 5.09e-09 ***
expersq     -0.0038911  0.0007054  -5.516 5.49e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.703 on 516 degrees of freedom
Multiple R-squared:  0.4737,      Adjusted R-squared:  0.4645
F-statistic:  51.6 on 9 and 516 DF,  p-value: < 2.2e-16
```

Theo tiêu chuẩn BIC, ta chọn được mô hình hồi quy tuyến tính của biến wage theo 9 biến:

Wage = 0.0871112 + 0.3138293×educ + 0.1867113×exper + 0.1134712×tenure - 1.6730322×female + 0.8904848×smsa - 1.4417386×trade - 1.1330508×services + 1.7398764×profocc - 0.0038911×expersq

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

H1: Mô hình đầy đủ biến

Bảng anova so sánh giữa hai mô hình:

```
> anova(M1_BIC, M1)
Analysis of Variance Table


Model 1: wage ~ educ + exper + tenure + female + smsa + trade + services +
    profocc + expersq
Model 2: wage ~ educ + exper + tenure + nonwhite + female + married +
    numdep + smsa + northcen + south + west + construc + ndurman +
    trcommpu + trade + services + profserv + profocc + clerocc +
    servocc + expersq + tenursq
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    516 3768.7
2    503 3639.6 13    129.09 1.3724 0.1684
```

Giá trị p-value của kiểm định là 0.1684 > α = 0.05, nghĩa là với mức ý nghĩa 5%. không có cơ sở bác bỏ H0, nên ta chấp nhận mô hình được thu gọn biến.

Mô hình tuyến tính theo biến lwage:

```
> M2 <- lm(lwage ~ educ + exper + tenure + nonwhite + female + married + numdep + smsa + northcen
+ south + west + construc + ndurman + trcommpu + trade + services + profserv + profocc + clerocc +
servocc + expersq + tenursq, data = data4)
> summary(M2)


Call:
lm(formula = lwage ~ educ + exper + tenure + nonwhite + female +
```

```
   married + numdep + smsa + northcen + south + west + construc +
   ndurman + trcommpu + trade + services + profserv + profocc +
   clerocc + servocc + expersq + tenursq, 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
```

Chọn mô hình 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
```

Theo tiêu chuẩn BIC chọn mô hình, mô hình hồi quy tuyến tính của lwage theo 10 biến:

lwage = 0.7806 + 0.04951×educ + 0.02854×exper + 0.01252×tenure - 0.2749×female + 0.1582×smsa - 0.2461×trade - 0.2376×services + 0.2030×profocc - 0.1324×servocc - 0.0005762×expersq

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

H1: Mô hình đầy đủ biến

Bảng anova so sánh giữa hai mô hình:

```
Analysis of Variance Table


Model 1: lwage ~ educ + exper + tenure + female + smsa + trade + services +
    profocc + servocc + expersq
Model 2: lwage ~ educ + exper + tenure + nonwhite + female + married +
    numdep + smsa + northcen + south + west + construc + ndurman +
    trcommpu + trade + services + profserv + profocc + clerocc +
    servocc + expersq + tenursq
  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.

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

```
summary(M1_BIC)$adj.r.squared

#0.4644986

summary(M2_BIC)$adj.r.squared

#0.5263545
```

Mô hình hồi quy tuyến tính của lwage có hệ số xác định R2 hiệu chỉnh là 52.63% cao hơn mô hình wage tương ứng là 46.45%, nên ta chọn mô hình lwage giải thích cho mức lương ($/giờ) của người lao động ở Anh năm 1976. Mô hình hồi quy tuyến tính của lwage:

lwage = 0.7806 + 0.04951×educ + 0.02854×exper + 0.01252×tenure - 0.2749×female + 0.1582×smsa - 0.2461×trade - 0.2376×services + 0.2030×profocc - 0.1324×servocc - 0.0005762×expersq