Homework 5

Bùi Khánh Duy

2023-03-30

THỰC HÀNH HỒI QUY TUYẾN TÍNH ĐA BIẾN

Sử dụng bộ dữ liệu Boston trong gói lệnh MASS bao gồm 14 biến liên quan đến giá trị nhà ở vùng ngoại ô ở Boston và hàm `step`, phân tích hồi quy bội của biến `medv` (giá nhà trung bình – đơn vị: nghìn $) theo các biến còn lại.

a) Đưa ra mô hình hồi quy tuyến tính "forward" và " backward" tốt nhất.

```
# install.packages("MASS")
library(MASS)

only <- lm(medv ~ 1, data = Boston)
all <- lm(medv ~ ., data = Boston)
summary(all)
```

```
##
## Call:
## lm(formula = medv ~ ., data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.595  -2.730  -0.518   1.777  26.199
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
## crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
## zn           4.642e-02  1.373e-02   3.382 0.000778 ***
## indus        2.056e-02  6.150e-02   0.334 0.738288
## chas         2.687e+00  8.616e-01   3.118 0.001925 **
## nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
## rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
## age          6.922e-04  1.321e-02   0.052 0.958229
## dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
## rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
## tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
## ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
## black        9.312e-03  2.686e-03   3.467 0.000573 ***
## lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.745 on 492 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7338
## F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

Mô hình HQTT forward

```
library(stats)
# forward = only to all.
forward <- step(object = only, scope = formula(all), direction = "forward", trace = 0)
forward$coefficients
```

```
##   (Intercept)         lstat            rm       ptratio           dis
##  36.341145004  -0.522553457   3.801578840  -0.946524570  -1.492711460
##           nox          chas         black            zn          crim
## -17.376023429   2.718716303   0.009290845   0.045844929  -0.108413345
##           rad           tax
##   0.299608454  -0.011777973
```

Mô hình HQTT backward

```
backward <- step(object = all, scope = formula(only), direction = "backward", trace = 0)
backward$coefficients
```

```
##   (Intercept)          crim            zn          chas           nox
##  36.341145004  -0.108413345   0.045844929   2.718716303 -17.376023429
##            rm           dis           rad           tax       ptratio
##   3.801578840  -1.492711460   0.299608454  -0.011777973  -0.946524570
##         black         lstat
##   0.009290845  -0.522553457
```

=> PTHQTT tốt nhất:

$$\begin{aligned} medv = {} & 36.341145004 - 0.108413345 * crim + 0.045844929 * zn \\ & + 2.718716303 * chas - 17.376023429 * nox + 3.801578840 * rm \\ & - 1.492711460 * dis + 0.299608454 * rad - 0.011777973 * tax \\ & - 0.946524570 * ptratio + 0.009290845 * black - 0.522553457 * lstat \end{aligned}$$

b) Khi phân tích "forward", nếu biến medv được biểu diễn theo hai biến thì đó là những biến nào?

```
forward$anova
```

```
##         Step Df    Deviance Resid. Df Resid. Dev      AIC
## 1            NA          NA       505   42716.30 2246.514
## 2    + lstat -1 23243.91400       504   19472.38 1851.009
## 3       + rm -1  4033.07222       503   15439.31 1735.577
## 4  + ptratio -1  1711.32389       502   13727.99 1678.131
## 5      + dis -1   499.07761       501   13228.91 1661.393
## 6      + nox -1   759.56355       500   12469.34 1633.473
## 7     + chas -1   328.27141       499   12141.07 1621.973
```

```
## 8    + black -1   272.83713       498   11868.24 1612.473
## 9      + zn -1   189.93614       497   11678.30 1606.309
## 10    + crim -1    94.71193       496   11583.59 1604.189
## 11     + rad -1   228.60431       495   11354.98 1596.103
## 12     + tax -1   273.61928       494   11081.36 1585.761
```

```
two <-  lm(medv ~ lstat + rm, data = Boston)
two$coefficients
```

```
## (Intercept)       lstat          rm
##  -1.3582728  -0.6423583   5.0947880
```

$$medv = -1.3582728 - 0.6423583 * lstat + 5.0947880 * rm$$

PT: $medv = a0 + a1 * lstat + a2 * rm$

Kiểm định xem các hệ số a0, a1, a2 trong mô hình hồi quy có thực sự khác 0 hay không?

```
summary(two)
```

```
##
## Call:
## lm(formula = medv ~ lstat + rm, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.076  -3.516  -1.010   1.909  28.131
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.35827    3.17283  -0.428    0.669
## lstat       -0.64236    0.04373 -14.689   <2e-16 ***
## rm           5.09479    0.44447  11.463   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.54 on 503 degrees of freedom
## Multiple R-squared:  0.6386, Adjusted R-squared:  0.6371
## F-statistic: 444.3 on 2 and 503 DF,  p-value: < 2.2e-16
```

BT: H0: a0 = 0; H1: a0 != 0

Do p_value = 0.669 > 0.05 nên chấp nhận H0

=> KL: Với KTC 95%, có cơ sở để nói a0 = 0

BT: H0: a1 = 0; H1: a1 != 0

Do p_value < 2e-16 < 0.05 nên bác bỏ H0

=> KL: Với KTC 95%, có cơ sở để nói a1 != 0

BT: H0: a2 = 0; H1: a2 != 0

Do p_value < 2e-16 < 0.05 nên bác bỏ H0

=> KL: Với KTC 95%, có cơ sở để nói a2 != 0

Khi đó, ta viết lại mô hình HQTT của `medv` theo `lstat` và `rm` như sau:

```
two = lm(medv ~ lstat + rm + 0, data = Boston)
two$coefficients
```

```
##     lstat        rm
## -0.655740  4.906906
```

$$medv = -0.655740 * lstat + 4.906906 * rm$$

c) Khi phân tích "backward", kiểm định xem các hệ số trong mô hình hồi quy tuyến tính thu được có thực sự khác 0 không? Phần dư có tuân theo phân phối chuẩn với giá trị trung bình bằng 0 không?

$$\begin{aligned} medv = {} & 36.341145004 - 0.108413345 * crim + 0.045844929 * zn \\ & + 2.718716303 * chas - 17.376023429 * nox + 3.801578840 * rm \\ & - 1.492711460 * dis + 0.299608454 * rad - 0.011777973 * tax \\ & - 0.946524570 * ptratio + 0.009290845 * black - 0.522553457 * lstat \end{aligned}$$

```
# medv = a0 + a1*crim + a2*zn + a3*chas + a4*nox + a5*rm + a6*dis + a7*rad + a8*tax + a9*ptratio + a10*
```

Kiểm định xem các hệ số a_i $(i = \overline{0,11})$ trong mô hình hồi quy có thực sự khác 0 hay không?

BT: H0: $a_i = 0$; H1: $a_i != 0$ (i = 0, 1, ..., 11)

```
summary(backward)
```

```
##
## Call:
## lm(formula = medv ~ crim + zn + chas + nox + rm + dis + rad +
##     tax + ptratio + black + lstat, data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5984  -2.7386  -0.5046   1.7273  26.2373
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.341145   5.067492   7.171 2.73e-12 ***
## crim         -0.108413   0.032779  -3.307 0.001010 **
## zn            0.045845   0.013523   3.390 0.000754 ***
## chas          2.718716   0.854240   3.183 0.001551 **
## nox         -17.376023   3.535243  -4.915 1.21e-06 ***
## rm            3.801579   0.406316   9.356  < 2e-16 ***
## dis          -1.492711   0.185731  -8.037 6.84e-15 ***
## rad           0.299608   0.063402   4.726 3.00e-06 ***
## tax          -0.011778   0.003372  -3.493 0.000521 ***
## ptratio      -0.946525   0.129066  -7.334 9.24e-13 ***
## black         0.009291   0.002674   3.475 0.000557 ***
## lstat        -0.522553   0.047424 -11.019  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.736 on 494 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7348
## F-statistic: 128.2 on 11 and 494 DF,  p-value: < 2.2e-16
```

Do p-value < 2.2e-16 < 0.05 (p_value lấy từ phần F-statistic) nên bác bỏ H0 => với KTC 95%, có cơ sở để nói $\forall a_i \neq 0$

Kiểm định xem phần dư có tuân theo phân phối chuẩn với giá trị trung bình = 0 hay không?

```
shapiro.test(backward$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  backward$residuals
## W = 0.90131, p-value < 2.2e-16
```

Do p-value < 2.2e-16 < 0.05 nên bác bỏ H0 => với KTC 95%, có cơ sở để nói phần dư không tuân theo pp chuẩn

Kiểm định xem giá trị trung bình của phần dư khác 0 hay không?

```
wilcox.test(backward$residuals)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  backward$residuals
## V = 55447, p-value = 0.008285
## alternative hypothesis: true location is not equal to 0
```

BT: H0: mu_re = 0; H1: mu_re != 0

Do p-value = 0.008285 < 0.05 nên bác bỏ H0

KL: Với KTC 95%, có cơ sở để nói giá trị trung bình của phần dư khác 0.