TH5

Đinh Quang Dũng

2024-03-11

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

#a  
library("MASS")

## Warning: package 'MASS' was built under R version 4.3.2

data(Boston)  
  
only = lm(medv ~ 1,data = Boston)  
  
summary(only)

##   
## Call:  
## lm(formula = medv ~ 1, data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.533 -5.508 -1.333 2.467 27.467   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 22.5328 0.4089 55.11 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.197 on 505 degrees of freedom

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

forward = step(only,formula(all),direction = "forward",trace=0)  
  
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

summary(forward)

##   
## Call:  
## lm(formula = medv ~ lstat + rm + ptratio + dis + nox + chas +   
## black + zn + crim + rad + tax, 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 \*\*\*  
## lstat -0.522553 0.047424 -11.019 < 2e-16 \*\*\*  
## rm 3.801579 0.406316 9.356 < 2e-16 \*\*\*  
## ptratio -0.946525 0.129066 -7.334 9.24e-13 \*\*\*  
## dis -1.492711 0.185731 -8.037 6.84e-15 \*\*\*  
## nox -17.376023 3.535243 -4.915 1.21e-06 \*\*\*  
## chas 2.718716 0.854240 3.183 0.001551 \*\*   
## black 0.009291 0.002674 3.475 0.000557 \*\*\*  
## zn 0.045845 0.013523 3.390 0.000754 \*\*\*  
## crim -0.108413 0.032779 -3.307 0.001010 \*\*   
## rad 0.299608 0.063402 4.726 3.00e-06 \*\*\*  
## tax -0.011778 0.003372 -3.493 0.000521 \*\*\*  
## ---  
## 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

backward = step(all,formula(all),direction = "backward",trace = 0)  
backward$anova

## Step Df Deviance Resid. Df Resid. Dev AIC  
## 1 NA NA 492 11078.78 1589.643  
## 2 - age 1 0.06183435 493 11078.85 1587.646  
## 3 - indus 1 2.51754013 494 11081.36 1585.761

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

#b  
# khi bieu dien medv theo 2 bien tu forward thi 2 bien la age va indus  
# khi bien dien medv theo backward thi cac bien trong Boston tru 2 bien la age va indus   
  
  
#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?  
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

confint(backward)

## 2.5 % 97.5 %  
## (Intercept) 26.384649126 46.29764088  
## crim -0.172817670 -0.04400902  
## zn 0.019275889 0.07241397  
## chas 1.040324913 4.39710769  
## nox -24.321990312 -10.43005655  
## rm 3.003258393 4.59989929  
## dis -1.857631161 -1.12779176  
## rad 0.175037411 0.42417950  
## tax -0.018403857 -0.00515209  
## ptratio -1.200109823 -0.69293932  
## black 0.004037216 0.01454447  
## lstat -0.615731781 -0.42937513

shapiro.test(backward$residuals)

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

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