

Mô hình hồi qui tuyển tính đa biến

Tuan V. Nguyen

Garvan Institute of Medical Research

University of New South Wales (UNSW Sydney), Australia

University of Technology, Sydney (UTS), Australia

Ton Duc Thang University, Vietnam



Mô hình hồi qui tuyến tính đa biến

- Mô hình hồi qui tuyến tính như là t-test
- Mô hình với ≥ 2 biến độc lập
- Mô hình đa thức (polynomial regression)

**Biến tiên lượng thuộc loại biến
phân nhóm (categorical)**

Dữ liệu giá nhà ở thành phố Boston

- Dữ liệu thu thập 506 căn nhà được bán, 14 biến số
- Biến outcome: **MEDV**
- Biến tiên lượng:
 - **crime** - per capita crime rate by town
 - **zone** - proportion of residential land zoned for lots over 25,000 sq.ft.
 - **industry** - proportion of non-retail business acres per town.
 - **river** - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
 - **nox** - nitric oxides concentration (parts per 10 million)
 - **rooms** - average number of rooms per dwelling
 - **age** - proportion of owner-occupied units built prior to 1940
 - **distance** - weighted distances to five employment centres
 - **radial** - index of accessibility to radial highways
 - **tax** - full-value property-tax rate per \$10,000
 - **ptratio** - pupil-teacher ratio by town
 - **black** - $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
 - **Istat** - % lower status of the population
 - **MEDV** - median value of owner-occupied homes in \$1000's

Dữ liệu giá nhà: so sánh giá nhà gần sông và xa sông

```
house = read.csv("~/Dropbox/_Conferences and Workshops/Banking University/Data/Housing prices  
data.csv")  
head(house)  
  
  crime zone industry river nox rooms age distance radial tax ptratio black lstat MEDV  
1 0.00632   18     2.31      0 0.538 6.575 65.2    4.0900      1 296 15.3 396.90 4.98 24.0  
2 0.02731     0    7.07      0 0.469 6.421 78.9    4.9671      2 242 17.8 396.90 9.14 21.6  
3 0.02729     0    7.07      0 0.469 7.185 61.1    4.9671      2 242 17.8 392.83 4.03 34.7  
4 0.03237     0    2.18      0 0.458 6.998 45.8    6.0622      3 222 18.7 394.63 2.94 33.4  
5 0.06905     0    2.18      0 0.458 7.147 54.2    6.0622      3 222 18.7 396.90 5.33 36.2  
6 0.02985     0    2.18      0 0.458 6.430 58.7    6.0622      3 222 18.7 394.12 5.21 28.7  
  
> t.test(MEDV ~ river, data=house)  
  
Welch Two Sample t-test  
  
data: MEDV by river  
t = -3.1133, df = 36.876, p-value = 0.003567  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
-10.476831 -2.215483  
sample estimates:  
mean in group 0 mean in group 1  
22.09384      28.44000
```

Difference in price = 28.4 – 22.1 = 6.3

(nhà ở gần sông Charles đắt hơn nhà xa sông)

So sánh giá nhà gần sông và xa sông: mô hình HQTT

```
m = lm(MEDV ~ river, data=house)
summary(m)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	22.0938	0.4176	52.902	< 2e-16 ***
river	6.3462	1.5880	3.996	7.39e-05 ***

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

Residual standard error: 9.064 on 504 degrees of freedom
Multiple R-squared: 0.03072, Adjusted R-squared: 0.02879
F-statistic: 15.97 on 1 and 504 DF, p-value: 7.391e-05

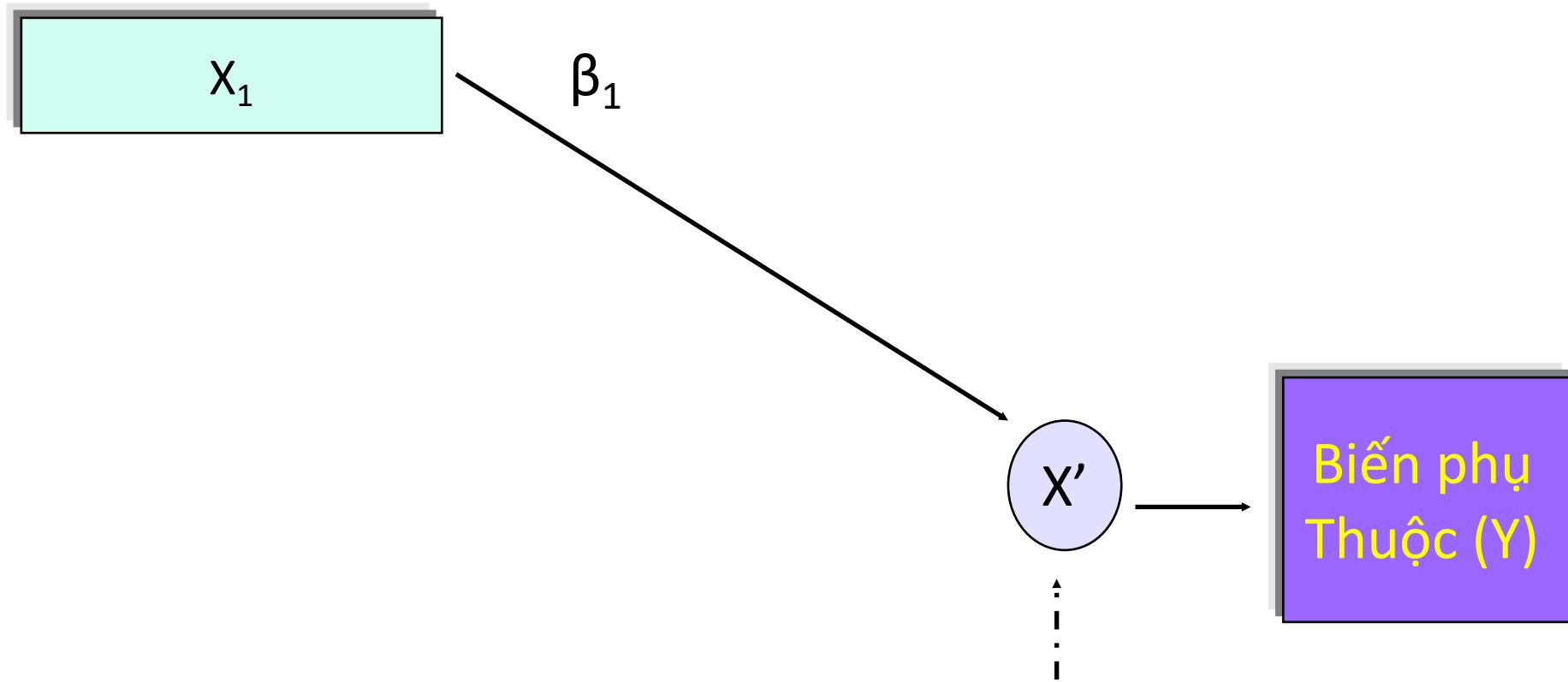
Mô hình hồi qui tuyến tính

$$\text{MEDV} = 22.1 + 6.34 * \text{river}$$

cho ra kết quả như t-test, nhưng có thêm thông tin

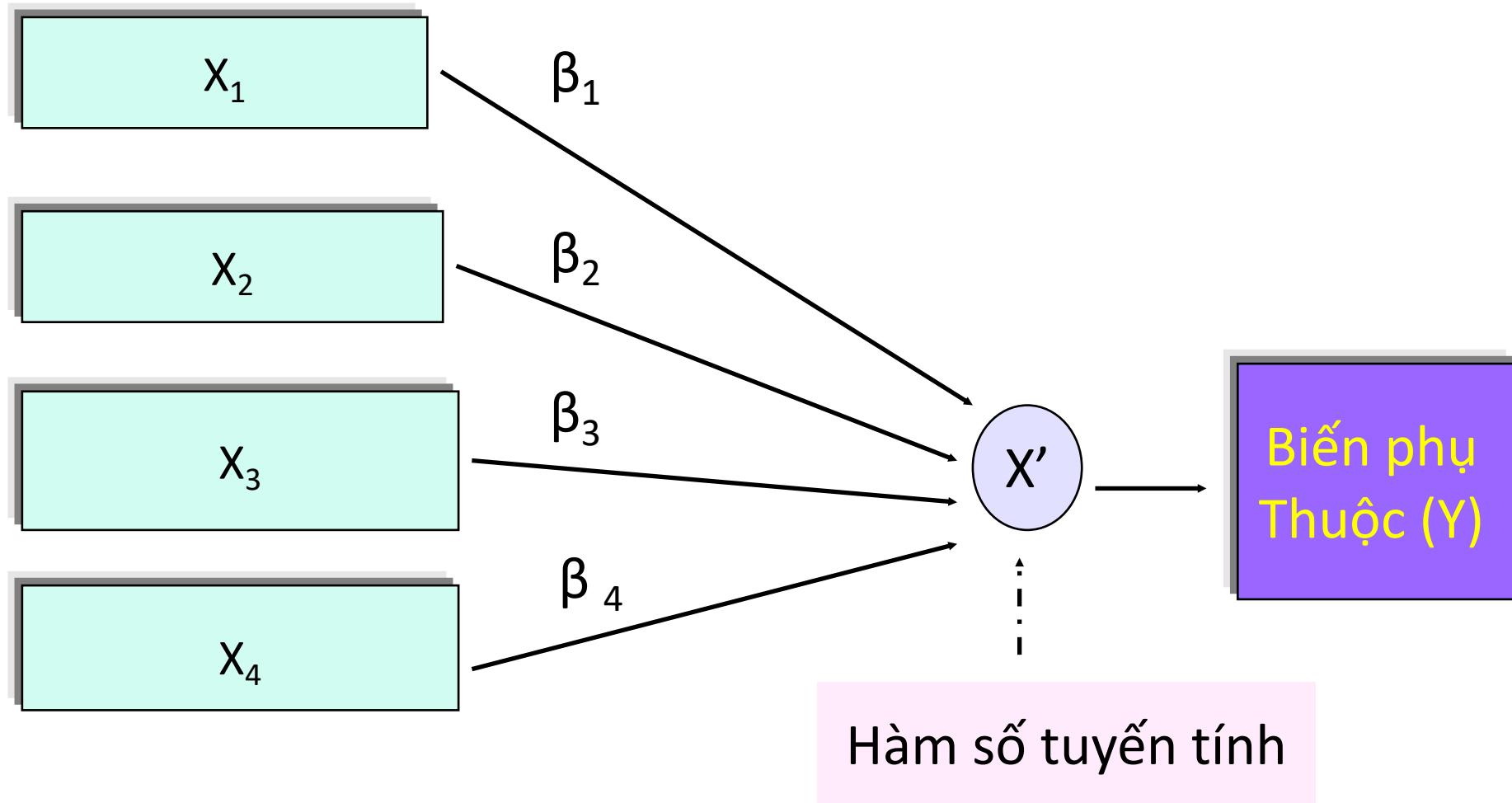
Mô hình với đa biến tiên lượng

Mô hình đơn biến



Hàm số tuyến tính

Khái niệm mô hình hồi qui đa biến



Mô hình hồi qui tuyến tính đa biến

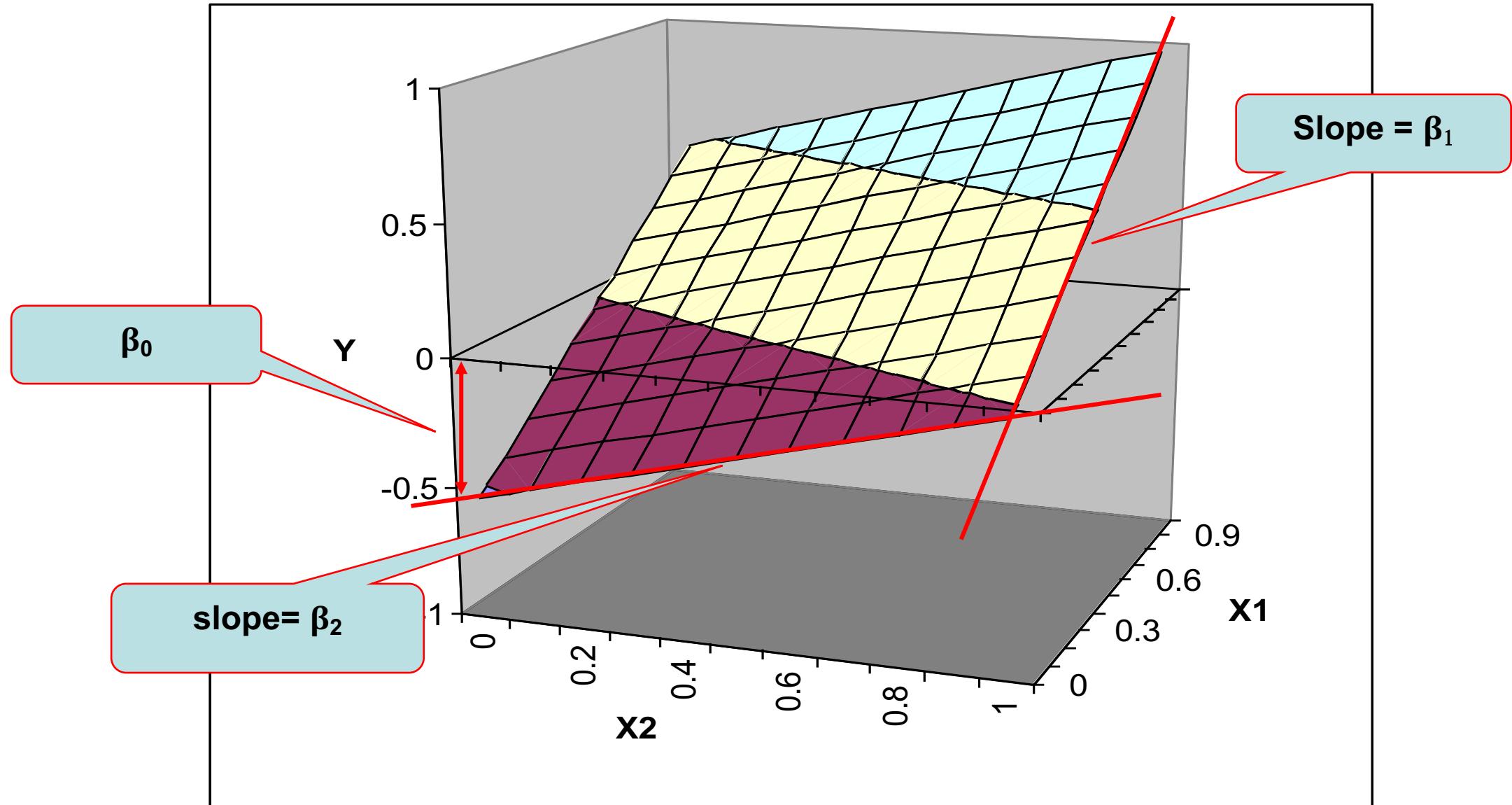
Mô hình căn bản

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

Y là biến phụ thuộc (dependent variable), **biến liên tục**

$X_1, X_2, X_3, \dots, X_p$: biến tiên lượng

$\beta_1, \beta_2, \beta_3, \dots, \beta_p$: regression coefficients (hệ số hồi qui)



Ước tính tham số

- Mô hình căn bản

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

- Dùng dữ liệu thực tế để ước tính beta:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p$$

$b_1, b_2, b_3, \dots, b_p$ là ước số của $\beta_1, \beta_2, \beta_3, \dots, \beta_p$

residuals $e = Y - \hat{Y}$

Yếu tố ảnh hưởng đến giá nhà Boston

- Giả thuyết: giá nhà chịu ảnh hưởng của số phòng (rooms), tuổi nhà (age), gần hay xa sông (river)

- Mô hình thống kê

$$\text{price} = b_0 + b_1 * \text{age} + b_2 * \text{rooms} + b_3 * \text{river}$$

- Dữ liệu: house prices

- Triển khai bằng R

```
m = lm(MEDV ~ age + rooms + river, data=house)
```

```
summary(m)
```

Yếu tố ảnh hưởng đến giá nhà

```
> m = lm(MEDV ~ age + rooms + river, data=house)
> summary(m)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-23.90897	2.81740	-8.486	2.42e-16 ***
age	-0.07800	0.01015	-7.684	8.18e-14 ***
rooms	8.18547	0.40686	20.119	< 2e-16 ***
river	5.02780	1.09662	4.585	5.74e-06 ***

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

Residual standard error: 6.194 on 502 degrees of freedom

Multiple R-squared: 0.5492, Adjusted R-squared: 0.5465

F-statistic: 203.8 on 3 and 502 DF, p-value: < 2.2e-16

Mô hình hồi qui tuyến tính

$\text{MEDV} = -23.9 - 0.08 * \text{age} + 8.2 * \text{rooms} + 5.0 * \text{river}$

Diễn giải

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-23.90897	2.81740	-8.486	2.42e-16	***
age	-0.07800	0.01015	-7.684	8.18e-14	***
rooms	8.18547	0.40686	20.119	< 2e-16	***
river	5.02780	1.09662	4.585	5.74e-06	***

Nhà cao tuổi hơn 1 năm giảm giá 0.078 (tức 78 USD). Nhà tăng 1 phòng tăng giá trị 8185 USD. Ở mỗi tuổi nhà và một số phòng, nhà ở gần sông tăng giá trị 5028 USD.

Đánh giá tầm quan trọng của mỗi biến tiên lượng

Câu hỏi quan trọng ...

- Trong các biến có liên quan, biến nào quan trọng nhất?
- Tiêu chuẩn nào để đánh giá?
 - Hệ số hồi qui trên mỗi SD (độ lệch chuẩn)
 - R^2 cho từng biến, nhưng tùy vào phân bố
- Phương pháp: "relative importance"

Câu hỏi quan trọng ...

Relative Importance for Linear Regression in R: The Package **relaimpo**

Ulrike Grömping
TFH Berlin – University of Applied Sciences

Abstract

Relative importance is a topic that has seen a lot of interest in recent years, particularly in applied work. The R package **relaimpo** implements six different metrics for assessing relative importance of regressors in the linear model, two of which are recommended - averaging over orderings of regressors and a newly proposed metric ([Feldman 2005](#)) called pmvd. Apart from delivering the metrics themselves, **relaimpo** also provides (exploratory) bootstrap confidence intervals. This paper offers a brief tutorial introduction to the package. The methods and **relaimpo**'s functionality are illustrated using the data set **swiss** that is generally available in R. The paper targets readers who have a basic understanding of multiple linear regression. For the background of more advanced aspects, references are provided.

Package “relaimpo” trong R

- **relaimpo** – có thể ước tính R^2 cho từng biến
- Phương pháp bootstrap
- **Img** = Lindermann, Merenda, Gold (một thước đo mới và tốt)
- Phương pháp Img "tách" R^2 cho từng biến tiên lượng

```
m = lm(MEDV ~ age + rooms + river, data=house)
library(relaimpo)
calc.relimp(m, type="lmg", rela=T, rank=T)
```

```
Response variable: MEDV
Total response variance: 84.58672
Analysis based on 506 observations

3 Regressors:
age rooms river
Proportion of variance explained by model: 54.92%
Metrics are normalized to sum to 100% (rela=TRUE).
```

Relative importance metrics:

```
      lmg
age  0.1796402
rooms 0.7732168
river 0.0471430
```

Average coefficients for different model sizes:

	1X	2Xs	3Xs
age	-0.1231627	-0.100930	-0.07799789
rooms	9.1021090	8.684510	8.18547207
river	6.3461571	5.833196	5.02780091

Tìm mô hình tối ưu

Model selection



Criteria for model selection

- Sometimes quantifiable
- Sometimes subjective
- Sometimes biased by pre-conceived ideas
- Sometimes pre-conceived ideas are truly important
- How well do they apply in future samples?

Model selection: the task of selecting a (mathematical) model from a set of potential models, given **evidence**.

Model selection



Bối cảnh và ý tưởng

- Trong một mô hình đa biến, câu hỏi quan trọng là biến X nào có liên quan đến Y **một cách độc lập**
- Trong hàng triệu markers, cái nào có liên quan đến bệnh loãng xương
- Làm sao để phát hiện những biến đó, marker đó?

Chọn biến / chọn mô hình

Một nghiên cứu thường có nhiều biến. Số mô hình có thể rất nhiều

$k = 2$, số mô hình: 3 (tối thiểu)

$k = 3$, số mô hình: 7 (tối thiểu)

$K = 10$, số mô hình: 1023 (tối thiểu)

Nói chung, số mô hình tối thiểu là $2^k - 1$

Overfitting

- Nếu chúng ta có quá nhiều biến tiên lượng trong mô hình, có thể dẫn đến tình trạng **overfitting**.
- Hệ quả:
 - Mô hình tiên lượng không chính xác
 - Quá phức tạp, nhiễu nhiều hơn tín hiệu

Underfitting

- Nếu mô hình có quá ít biến tiên lượng, và bỏ qua những biến quan trọng, mô hình ***underfitting***.
- Hệ quả:
 - Tiên lượng kém chính xác
 - Tham số bị biased (xa rời giá trị thật)
 - Phương sai tăng

Phương pháp chọn mô hình

- Stepwise regression
- All possible subsets
- AIC, BIC
- BMA

AIC và BIC: highly recommended

- AIC – Akaike Information Criterion
- BIC – Bayesian Information Criterion
- Những thước đo để cân đối tính phức tạp (số biến tiên lượng) và goodness-of-fit (qua RSS)
- $AIC = n \log(RSS_p) + 2p$
- Cũng có thể tính: $AIC = RSS/RMS_{Full} + 2p$, equivalent to Cp
- $BIC = n \log(RSS_p) + p \log n$
- **AIC và BIC càng thấp = mô hình càng “tốt”**

BMA – Bayesian Model Average

- Có lẽ là phương pháp hấp dẫn nhất
- Dùng BIC để chọn mô hình tốt nhất
- Dùng nhiều công suất máy tính

BMA – Bayesian Model Average

- Giả dụ chúng ta có nhiều mô hình khả dĩ $M_m, m = 1, \dots, M$ với tham số θ_m .
- Thông tin tiền định: $\Pr(\theta_m | M_m), m = 1, \dots, M$.
- Xác suất hậu định :
$$\Pr(M_m | Z) \propto \Pr(M_m) \cdot \Pr(Z | M_m).$$
- So sánh 2 mô hình qua xác suất hậu định

$$\frac{\Pr(M_m | Z)}{\Pr(M_l | Z)} = \frac{\Pr(M_m)}{\Pr(M_l)} \cdot \frac{\Pr(Z | M_m)}{\Pr(Z | M_l)}$$

Tìm yếu tố liên quan đến giá nhà Boston

Triển khai phương pháp bằng package ‘BMA’

```
library(BMA)
```

Định nghĩa biến phụ thuộc

```
yvar = house[, ("MEDV") ]
```

Định nghĩa biến độc lập

```
xvars = house[, c("crime", "zone", "industry", "river", "nox",
"rooms", "age", "distance", "radial", "tax", "ptratio", "black",
"lstat")]
```

Dùng hàm bicreg

```
bma = bicreg(xvars, yvar, strict=FALSE, OR=20)
```

Trình bày kết quả phân tích

```
summary(bma)
```

```
imageplot.bma(bma)
```

```
> summary(bma)
```

5 models were selected

Best 5 models (cumulative posterior probability = 1):

	p!=0	EV	SD	model 1	model 2	model 3
Intercept	100.0	36.520642	5.174370	3.634e+01	3.662e+01	3.523e+01
crime	93.4	-0.101775	0.041894	-1.084e-01	-1.141e-01	.
zone	95.0	0.043273	0.016564	4.584e-02	4.574e-02	4.173e-02
industry	0.0	0.000000	0.000000	.	.	.
river	90.1	2.465386	1.152744	2.719e+00	.	2.872e+00
nox	100.0	-17.325300	3.574303	-1.738e+01	-1.647e+01	-1.651e+01
rooms	100.0	3.813179	0.410411	3.802e+00	3.845e+00	3.832e+00
age	0.0	0.000000	0.000000	.	.	.
distance	100.0	-1.476279	0.198410	-1.493e+00	-1.526e+00	-1.420e+00
radial	100.0	0.295811	0.065488	2.996e-01	3.155e-01	2.389e-01
tax	100.0	-0.011751	0.003427	-1.178e-02	-1.267e-02	-1.143e-02
ptratio	100.0	-0.955902	0.133379	-9.465e-01	-9.784e-01	-9.355e-01
black	96.3	0.009058	0.003188	9.291e-03	9.730e-03	1.032e-02
lstat	100.0	-0.525639	0.048080	-5.226e-01	-5.281e-01	-5.479e-01
nVar				11	10	10
r2				0.741	0.735	0.735
BIC				-6.143e+02	-6.102e+02	-6.094e+02
post prob				0.747	0.099	0.066

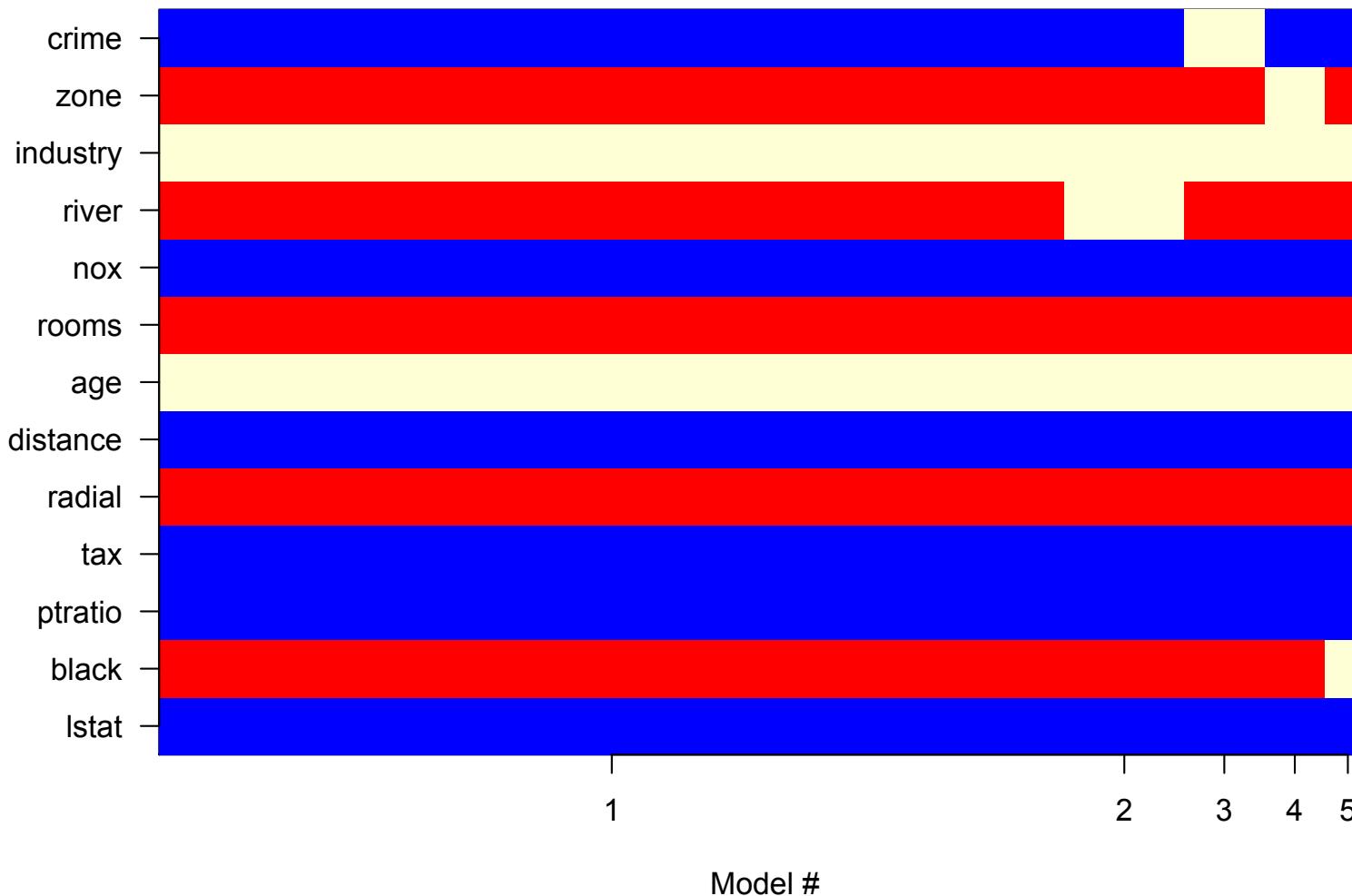
Điễn giải: Mô hình tốt nhất là “model 1” vì có BIC thấp nhất. Mô hình 1 gồm 11 predictors, với hệ số xác định 74.1%. Mô hình này có xác suất hậu định là 74.7%.

```
> summary(bma)
```

model 4	model 5
Intercept	3.703e+01 4.145e+01
crime	-9.819e-02 -1.217e-01
zone	. 4.619e-02
industry	. .
river	2.712e+00 2.872e+00
nox	-1.863e+01 -1.826e+01
rooms	4.003e+00 3.673e+00
age	. .
distance	-1.178e+00 -1.516e+00
radial	2.844e-01 2.839e-01
tax	-9.548e-03 -1.229e-02
ptratio	-1.097e+00 -9.310e-01
black	9.358e-03 .
lstat	-5.219e-01 -5.465e-01
nVar	10 10
r2	0.735 0.734
BIC	-6.089e+02 -6.083e+02
post prob	0.050 0.037

Kết quả tìm ‘mô hình tối ưu’

Models selected by BMA



Mô hình hồi qui tuyến tính đa biến

Mô hình hồi qui tuyến tính có thể dùng

- như là một t-test (biến tiên lượng là biến phân nhóm)
- để hiệu chỉnh trong mô hình đa biến

Cách tìm mô hình tối ưu: Bayesian Model Averaging