Diễn giải mô hình hồi qui tuyến tính

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

Diễn giải mô hình hồi qui tuyến tính

- Intercept và slope
- Predicted value – giá trị tiên lượng
- Confidence interval (CI)
- Prediction interval (PI)

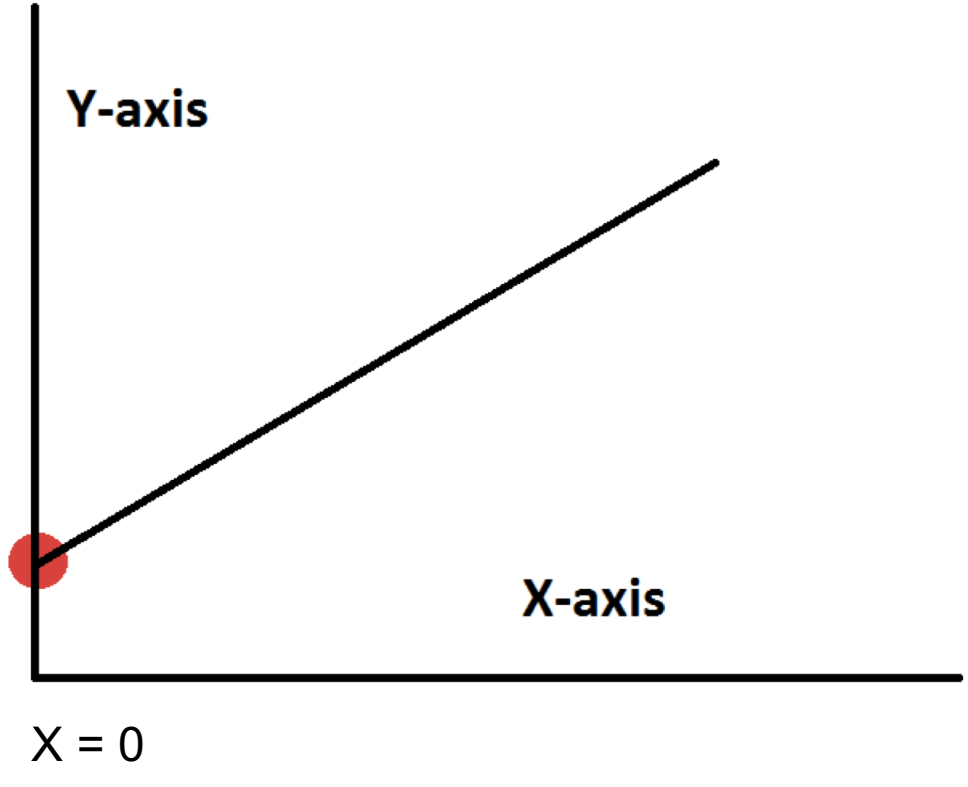
Ý nghĩa của intercept

Mô hình đơn giản

$$Y = \alpha + \beta X + \varepsilon$$

Do đó, α là giá trị của *Y* khi *X* = 0

Có khi … khó diễn giải

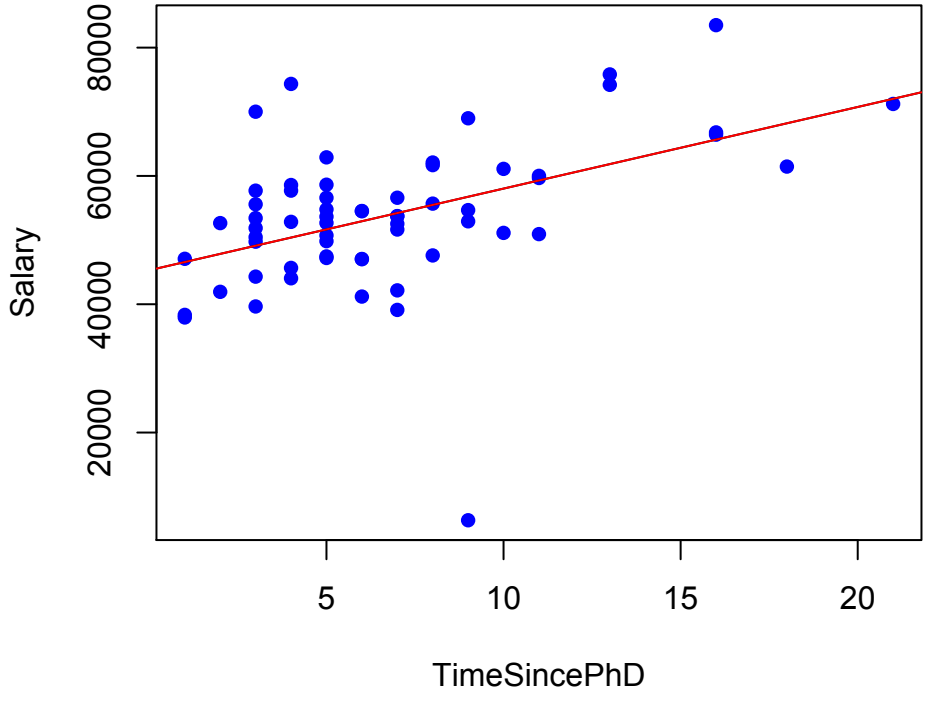


Ý nghĩa của intercept

- Mô hình tiên lượng đồng lương của các assistant professors (Mĩ)
- Y = lương ($); X = số năm sau khi tốt nghiệp PhD

$$\widehat{Y} = 45303 + 1272(\text{Years})$$

Khi Year = 0, $\widehat{Y} = 45303$



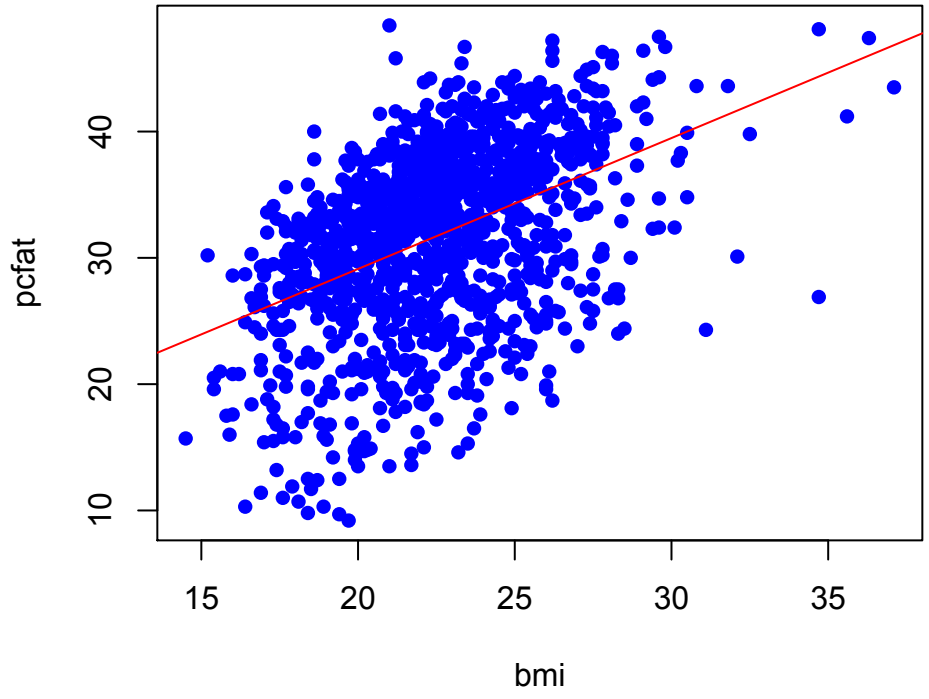
Ý nghĩa của intercept khi X không thể 0

- Mô hình tiên lượng tỉ trọng mỡ
- Y = tỉ trọng mỡ; X = body mass index (cân nặng chia cho chiều cao bình phương)

$$\hat{Y} = 8.4 + 1.04(\text{BMI})$$

Khi BMI = 0, $\hat{Y} = 8.4$

Không thể!



‘Chuẩn hoá’ X và ý nghĩa intercept

Chuẩn hoá bmi, c.bmi có trung bình 0 và độ lệch chuẩn 1

```
ob$c.bmi = scale(ob$bmi)
```

$$\hat{Y} = 31.6 + 3.2(\text{c.BMI})$$

c.BMI = 0 có nghĩa là trung bình BMI trong quần thể. Do đó, intercept = 31.6 có nghĩa là tỉ trọng mỡ của một người có BMI trung bình là 31.6%.

Ý nghĩa của intercept

- Giá trị trung bình của Y khi X = 0
- Nếu X không thể nào là 0, thì intercept cũng vô nghĩa
- Có thể chuẩn hoá X để giúp intercept có ý nghĩa

Ý nghĩa của slopes

- X là biến liên tục
- X là biến nhị phân
- X là biến phân loại

Diễn giải slope khi X là biến liên tục

- Mô hình tiên lượng lương của các assistant professors (Mĩ)
- Y = lương ($); X = số năm sau khi tốt nghiệp PhD

$$\hat{Y} = 45303 + 1272(\text{Years})$$

- Diễn giải: *mỗi năm sau khi tốt nghiệp PhD có liên quan đến 1272 USD*.

Diễn giải slope khi X là biến nhị phân

```
> summary(lm(pcfat ~ gender, data=ob))


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  34.6724     0.1826   189.9   <2e-16 ***
genderM     -10.5163     0.3381   -31.1   <2e-16 ***
```

- Mô hình

$$\widehat{Y} = 34.7 - 10.5(\text{Gender} = \text{M})$$

- Diễn giải: *nam (gender=M) có tỉ trọng mỡ thấp hơn nữ 10.5%*.

Diễn giải slope khi X là biến phân nhóm

```
> ob$obesity[ob$bmi<25.0] <- "Normal"
> ob$obesity[ob$bmi>=25.0 & ob$bmi<29.9] <- "Overweight"
> ob$obesity[ob$bmi>=30.0] <- "Obese"
> summary(lm(pcfat ~ obesity, data=ob))

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        30.6112     0.2214 138.232  < 2e-16 ***
obesityObese        7.4954     1.7963   4.173 3.23e-05 ***
obesityOverweight   4.7683     0.5062   9.419  < 2e-16 ***
```

- Mô hình

$$\hat{Y} = 30.6 + 7.5(\text{Obese})$$

$$\hat{Y} = 30.6 + 4.8(\text{Overweight})$$

- Diễn giải: *so sánh với nhóm bình thường, tỉ trọng mỡ ở người quá cân (overweight) và béo phì (obese) lần lược cao hơn 4.8% và 7.5%.*

Dự báo tương lai

Mô hình

$$\widehat{Y} = a + bX$$

Có thể dung để dự báo tương lai Y_h cho một giá trị X

Có thể tính khoảng dao động 95% của $Y_h \pm k \times SE$

- Confidence Interval (CI)
- Prediction Interval (PI)

Confidence Interval (CI)

- Khoảng tin cậy (CI) của giá trị dự báo: **khoảng giá trị trung bình** của Y_h

$$y_h \pm t_{\frac{\alpha}{2},n-2}\sqrt{MSE\left(\frac{1}{n}+\frac{(x_k-\bar{x})^2}{\sum(x_i-\bar{x}^2)}\right)}$$

```
> newvalue = data.frame(TimeSincePhD=30)
> predict(m, newdata=newvalue, interval="confidence")
       fit       lwr       upr
1 83468.83 69096.75 97840.92
```

Đối với người có 30 năm sau tiến sĩ, lương trung bình dự báo là 83468, nhưng có thể dao động trong khoảng 69097 đến 97841 USD với xác suất 95%.

Prediction Interval

- Khoảng dự báo (PI): giá trị khả dĩ của một quan sát mới (a range of values that is likely to contain the value of a *single new observation*)
- Thường rộng hơn CI

$$y_h \pm t_{\frac{\alpha}{2},n-2}\sqrt{MSE\left(1+\frac{1}{n}+\frac{(x_k-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$$

```
> newvalue = data.frame(TimeSincePhD=30)
> predict(m, newdata=newvalue, interval="prediction")
       fit       lwr       upr
1 83468.83 58554.15 108383.5
```

Xác suất 95% là một người với 30 năm sau tiến sĩ có lương dao động trong khoảng 58554 đến 108383 USD

Diễn giải mô hình hồi qui tuyến tính

- Intercept: là giá trị Y khi X = 0, nhưng có khi không có ý nghĩa!
- Slope: tham số phản ảnh mức độ ảnh hưởng của X đến Y
- Confidence Interval: giá trị dự báo trung bình
- Prediction Interval: giá trị dự báo của một cá nhân