BÀI THỰC HÀNH 4

TƯƠNG QUAN VÀ HỒI QUY

I. Thực hiện trên phần mềm R bài tập sau về tính hệ số tương quan và tính toán hàm hồi quy thực nghiệm:

Bài 1. *Một mẫu quan sát của đại lượng ngẫu nhiên hai chiều (*X; Y*) có giá trị như sau*

(2; 1; 4; 12); (2; 2; 4; 34); (2; 4; 4; 56); (2; 5; 4; 63)
(2; 25; 4; 38); (2; 45; 4; 75); (2; 16; 4; 4); (2; 34; 4; 62)

a) Hãy tính hệ số tương quan thực nghiệm của mẫu trên.
b) Hãy xây dựng hàm hồi quy tuyến tính của Y *theo* X.

Các thao tác cụ thể cần thực hiện với R:

1. Nhập biến X

```
> bienX <- c(2.1, 2.2, 2.4, 2.5, 2.25, 2.45, 2.16, 2.34)
```

2. Nhập biến Y

```
> bienY <- c(4.12, 4.34, 4.56, 4.63, 4.38, 4.75, 4.4, 4.62)
```

3. Tính hệ số tương quan theo cú pháp

```
> cor (bienX, bienY)
```

Xác nhận kết quả sau trên màn hình

```
[1] 0.9098077
>
```

4. Ước lượng các hệ số hồi quy theo cú pháp

```
> lm(bienY ~ bienX)
```

Xác nhận kết quả sau trên màn hình

```
lm(formula = bienY ~ bienX)

Coefficients:

(Intercept)       bienX

    1.544       1.274
```

>

Các thông tin cần nắm được khi thực hiện:

1. Cách nhập các biến X, Y.
2. Ký hiệu cor trong lệnh cor (bienX, bienY) nghĩa là *hệ số tương quan (coefficient of correlation)*. Công thức của hệ số này là
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$
3. Giá trị nhận được từ R là
[1] 0.9098077
là giá trị tính được của hệ số tương quan r.
4. Ký hiệu lm trong lệnh lm(bienY ~ bienX) nghĩa là *mô hình tuyến tính (**l**inear **m**odel).* Ký hiệu bienY ~ bienX có nghĩa là *mô tả bienY như một hàm số của bienX*. Công thức tính toán của mô hình là
$$y = \widehat{\beta_0} + \widehat{\beta_1}\, x$$
với $\widehat{\beta_0}, \widehat{\beta_1}$ là hai hệ số hồi quy thực nghiệm được ước lượng theo công thức
$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$
5. Kết quả được xác định từ R là

Coefficients:

(Intercept) bienX

1.544 1.274

có nghĩa là R tính ra được $\widehat{\beta_0} = 1.544$ *và* $\widehat{\beta_1} = 1.274$. Nói cách khác hàm hồi quy thực nghiệm được đưa ra là
$$y = 1{,}544 + 1{,}274\, x.$$

Trình bày lời giải của bài toán ra giấy sau các tính toán thực hành trên R (*áp dụng cho việc kiểm tra*):

Công thức tính hệ số tương quan là
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Kết quả tính toán thực nghiệm là r =0,9098077 (sử dụng R)

Công thức tính các hệ số hồi quy tuyến tính là

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

Kết quả tính toán thực nghiệm là $\widehat{\beta_0} = 1{,}544$ *và* $\widehat{\beta_1} = 1{,}274$ *(sử dụng R).*

Sinh viên có thể sử dụng các công thức tương đương để trình bày trong lời giải và cần giải thích được các ký hiệu $\bar{x}$, $\bar{y}$ trong lời giải trên.

Thực hành giải các bài tập sau bằng R

Bài 2. *Người ta lấy một mẫu thực nghiệm của đại lượng ngẫu nhiên hai chiều (X; Y) và thu được kết quả:*

X	*3, 6*	*3, 8*	*4, 3*	*4, 5*	*4, 9*	*5, 2*	*5, 4*
Y	*7, 1*	*7, 83*	*9, 62*	*10, 05*	*10, 7*	*11, 6*	*12,3*

a) Hãy tính hệ số tương quan thực nghiệm của mẫu trên.
b) Hãy xây dựng hàm hồi quy tuyến tính của Y theo X.

Kết quả đối chiếu

1. Kết quả tính hệ số tương quan thực nghiệm r
   [1] 0.9928191
2. Kết quả tính các hệ số hồi quy $\widehat{\beta_0}$, $\widehat{\beta_1}$
   lm(formula = bienY ~ bienX)
   Coefficients:
   (Intercept) bienX
   -2.590 2.755

Bài 3. *Để nghiên cứu về quan hệ giữa khối lượng bốc dỡ X (nghìn tấn) và thời gian bốc dỡ Y (giờ) người ta lấy một mẫu thực nghiệm và thu được kết quả:*

(10; 5, 5); (12; 6, 5); (11; 6, 3); (9; 4, 5);
(9, 5; 5, 3); (8; 4, 0); (12; 7, 0); (8, 5; 5, 0).

a) Hãy tính hệ số tương quan thực nghiệm của mẫu trên.
b) Hãy xây dựng hàm hồi quy tuyến tính của Y theo X.

Kết quả đối chiếu

1. Kết quả tính hệ số tương quan thực nghiệm r
   [1] 0.9619439
2. Kết quả tính các hệ số hồi quy $\widehat{\beta_0}$, $\widehat{\beta_1}$
   lm(formula = tgian ~ kluong)
   Coefficients:
   (Intercept) kluong
   -0.9420 0.6455

Bài 4. *Để nghiên cứu về quan hệ giữa khoảng cách X (km) từ nhà tới nơi làm việc và thời gian đi lại Y (phút), người ta lấy một mẫu thực nghiệm và có kết quả*

(10; 45); (12; 54); (11; 48); (9; 45);
(7; 30); (8; 32); (7, 5; 40); (8, 5; 42).

a) Hãy tính hệ số tương quan thực nghiệm của mẫu trên.
b) Hãy xây dựng hàm hồi quy tuyến tính của Y theo X.

Kết quả đối chiếu

1. Kết quả tính hệ số tương quan thực nghiệm r
   [1] 0.9012851
2. Kết quả tính các hệ số hồi quy $\widehat{\beta_0}$, $\widehat{\beta_1}$
   lm(formula = tgian1 ~ kcach)
   Coefficients:
   (Intercept) kcach
   4.433 4.117

II. Thực hiện trên phần mềm R bài tập về xác định hàm hồi quy và ước lượng giá trị dự báo:

Bài 5. *Số liệu về dân số (tính theo nghìn người) thành phố Hồ Chí Minh trong các năm gần đây được thống kê như sau:*

Năm	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*
Số dân	*7498, 4*	*7660,3*	*7820, 0*	*7981,9*	*8146, 3*	*8320,1*

a) Hãy tìm hàm xu thế tuyến tính biểu thị dân số của thành phố Hồ Chí Minh.
b) Vẽ hình mô tả dữ liệu (biểu đồ phân tán) và đồ thị hàm hồi quy tuyến tính thực nghiệm.
c) Xác định sai số của dữ liệu được cung cấp và hàm hồi quy thực nghiệm tại các điểm quan sát.
d) Dự báo số dân năm 2017 của thành phố này và tìm khoảng tin cậy 98% cho giá trị đó.

Các thao tác cụ thể cần thực hiện với R:

1. Nhập biến thời gian

```
> tgian <- c(2011, 2012, 2013, 2014, 2015, 2016)
```

2. Nhập biến dân số

```
> danso <- c(7498.4, 7660.4, 7820, 7981.9, 8146.3, 8320.1)
```

3. Ước lượng các hệ số hồi quy theo cú pháp

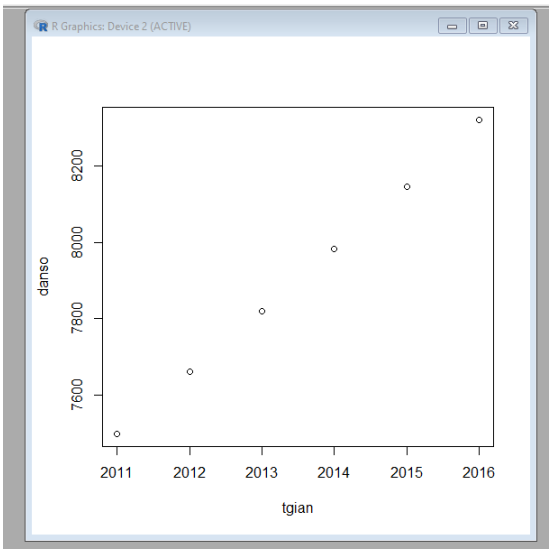
> lm(danso ~ tgian)

Xác nhận kết quả sau trên màn hình

Call:
lm(formula = danso ~ tgian)
Coefficients:
(Intercept) tgian
-321624.9 163.7

4. Vẽ biểu đồ miêu tả dữ liệu (biểu đồ phân tán) được cung cấp theo câu lệnh
> plot (tgian, danso)

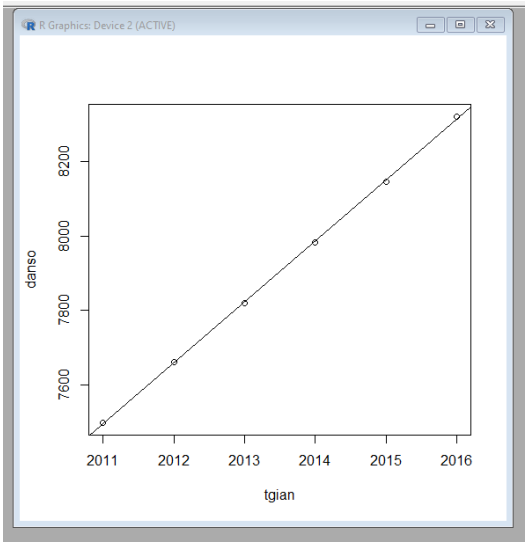
Xác nhận hình ảnh được R đưa ra



5. Tạo object chứa ***các thông tin về hồi quy*** trong R theo lệnh
> reg <- lm (danso ~ tgian)

6. Vẽ đường hồi quy thực nghiệm bằng R theo cú pháp
```
> abline(reg)
```

Xác nhận hình ảnh được R đưa ra



7. Tính sai số của dữ liệu được cung cấp và hàm hồi quy thực nghiệm tại các điểm quan sát theo lệnh

> residuals (reg)

hoặc câu lệnh thu gọn

> resid (reg)

Xác nhận kết quả R đưa ra

1	2	3	4	5	6
3.033333	1.373333	-2.686667	-4.446667	-3.706667	6.433333

8. Đưa ra công thức khoảng tin cậy sau để thực hiện tính toán theo yêu cầu d

$$\left(\widehat{y_0} - t_{(n-2,\frac{\alpha}{2})}\sqrt{s^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}\ ;\ \widehat{y_0} + t_{(n-2,\frac{\alpha}{2})}\sqrt{s^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}\right)$$

9. Nhập giá trị x_0 và tính $\widehat{y_0}$ theo các câu lệnh

> x0 <-2017

> beta0mu <- coef(reg)[1]

> beta1mu <- coef(reg)[2]

> y0mu <- beta0mu+beta1mu*x0

10. Đọc giá trị của $\widehat{y_0}$ từ R theo câu lệnh

>y0mu

Xác nhận kết quả được R đưa ra

(Intercept)

8477.327

11. Tính giá trị s^2 theo các câu lệnh

> n <- length(tgian)
> sbp <- sum(resid(reg)^2)/(n-2)

12. Đọc giá trị của s^2 từ R theo câu lệnh

>sbp

Xác nhận kết quả được R đưa ra

[1] **23.30133**

13. Tính $\bar{x}$ và S_{xx} theo các câu lệnh sau

> xtb <- mean(tgian)

> Sxx <-sum(tgian^2)-n*xtb^2

14. Đọc các giá trị $\bar{x}$ và S_{xx} và *xác nhận các kết quả từ R*

> xtb

[1] **2013.5**

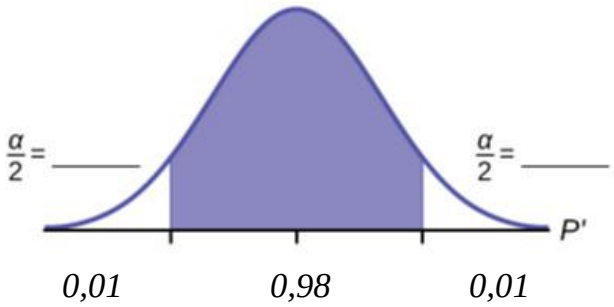
> Sxx

[1] **17.5**

15. Định nghĩa biến "phân vị" để ghi lại giá trị $t_{(n-2,\frac{\alpha}{2})}$

> phanvi <- qt(0.99,4)

Nhắc lại rằng *biến student với n-2 bậc tự do có đồ thị hàm mật độ là đường cong tạo với trục hoành một hình "quả chuông" có diện tích bằng 1.* ***Phân vị*** *(=vị trí phân chia) được xác định bởi hàm qt với 2 chỉ số. Chỉ số thứ nhất là diện tích mảnh chuông bên trái phân vị. Chỉ số thứ 2 là số bậc tự do.*



Do độ tin cậy là 98% nên ta lấy khoảng ước lượng theo mảnh chuông ở giữa có diện tích 0,98 và cắt bỏ 2 mẩu chuông cân đối 2 bên (mỗi mẩu diện tích 0,01). Điểm cắt bên phải là

$t_{(n-2,\frac{\alpha}{2})}$, điểm cắt bên trái là $-t_{(n-2,\frac{\alpha}{2})}$. Như vậy diện tích mẩu chuông bên phải phân vị $t_{(n-2,\frac{\alpha}{2})}$ là 0,01 và diện tích mảnh chuông bên trái phân vị $t_{(n-2,\frac{\alpha}{2})}$ là 1-0,01=0,99. ***Đây là lý do ta đưa 0.99 vào chỉ số thứ nhất của hàm qt.***

16. Đọc giá trị phân vị và *xác nhận các kết quả từ R*

```
> phanvi
[1] 3.746947
```

17. Tính bán kính khoảng ước lượng theo các câu lệnh

```
bkinh <- phanvi*sqrt(sbp*(1/n+(x0-xtb)^2/Sxx))
```

và *xác nhận kết quả từ R*

```
> bkinh
[1] 16.83814
```

18. Tính khoảng ước lượng theo câu lệnh

```
> y0mu+c(-1,1)*bkinh
```

và *xác nhận kết quả từ R*

```
[1] 8460.489 8494.165
```

Trình bày lời giải các ý a và d của bài toán ra giấy sau các tính toán thực hành trên R (*áp dụng cho việc kiểm tra*):

a) Công thức tính các hệ số hồi quy tuyến tính là

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

Kết quả tính toán thực nghiệm là $\widehat{\beta_0} = -321624{,}9$ và $\widehat{\beta_1} = 163{,}7$ (sử dụng R)

b) Thời điểm ước lượng dân số của TP Hồ Chí Minh là

$$x_0 = 2017$$

Điểm ước lượng cho dân số TP Hồ Chí Minh tại thời điểm được lựa chọn là

$$\widehat{y_0} = \widehat{\beta_0} + \widehat{\beta_1}\, x_0 = -321624{,}9 + 163{,}7 * 2017 = 8477{,}327$$

Công thức khoảng tin cậy cho dân số thành phố tại thời điểm được lựa chọn là

$$\left(\widehat{y_0} - t_{(n-2,\frac{\alpha}{2})}\sqrt{s^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}\ ;\ \widehat{y_0} + t_{(n-2,\frac{\alpha}{2})}\sqrt{s^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}\right)$$

Sử dụng R ta tính được các giá trị

$$\bar{x} = 2013{,}5;\ \ S_{xx} = 17{,}5;\ \ \ s^2 = 23{,}30133.$$

Phân vị của biến student là $t_{(n-2,\frac{\alpha}{2})} = \ 3{,}74694.$

Bán kính ước lượng là $\varepsilon = t_{(n-2,\frac{\alpha}{2})}\sqrt{s^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)} = 16{,}83814$

Kết quả tính toán khoảng tin cậy cho dân số thành phố tại thời điểm được lựa chọn là

(8460,489; 8494,165).

Thực hành giải các bài tập sau bằng R

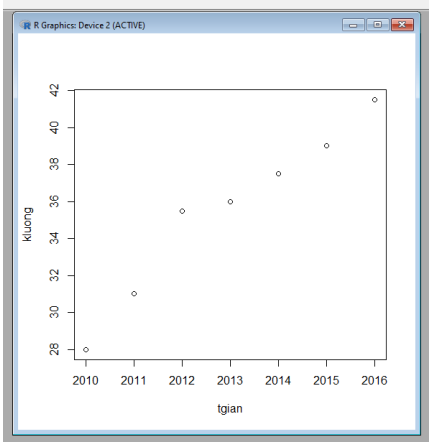
Bài 6. *Số liệu về lượng vận chuyển của một công ty vận tải trong các năm qua (tính theo triệu tấn) là như sau:*

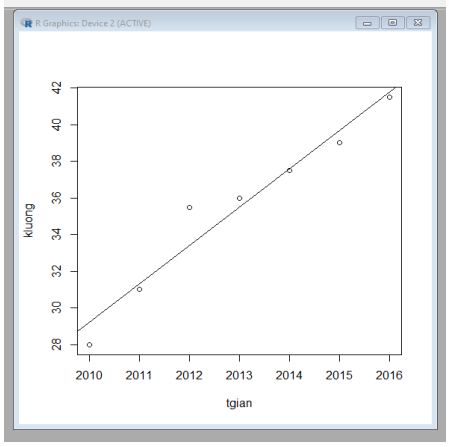
Năm	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*
Khối lượng	*28*	*31*	*35, 5*	*36*	*37, 5*	*39*	*41, 5*

a) Hãy tìm hàm xu thế tuyến tính biểu thị năng lực vận chuyển của công ty đó.
b) Vẽ hình mô tả dữ liệu và đồ thị hàm hồi quy tuyến tính thực nghiệm.
c) Xác định sai số của dữ liệu được cung cấp và hàm hồi quy thực nghiệm tại các điểm quan sát.
d) Dự báo khối lượng vận chuyển năm 2017 và tìm khoảng tin cậy 95% cho giá trị đó.

Kết quả đối chiếu

1. Kết quả tính các hệ số hồi quy $\widehat{\beta_0}, \widehat{\beta_1}$
   Call:
   lm(formula = kluong ~ tgian)
   Coefficients:
   (Intercept) tgian
   -4170.232 **2.089**
2. Hình ảnh của các mô tả trực quan





Sai lệch giữa giá trị quan sát và hàm hồi quy

```
> resid (reg)
        1         2         3         4
-1.23214286 -0.32142857  2.08928571  0.50000000
        5         6         7
-0.08928571 -0.67857143 -0.26785714
```

3. Điểm ước lượng cho lượng vận chuyển của công ty vận tải tại thời điểm được lựa chọn là

```
> y0mu
(Intercept)
   43.85714
```

4. Kết quả tính các biến trong công thức độ tin cậy

```
> xtb
[1] 2013
> Sxx
[1] 28
> sbp
[1] 1.355357
```

5. Phân vị và bán kính khoảng ước lượng

```
> phanvi
[1] 2.570582
> bkinh
[1] 2.529265
```

6. Khoảng tin cậy 95% cho lượng vận chuyển của công ty vận tải tại thời điểm được lựa chọn là

```
> y0mu+c(-1,1)*bkinh
[1] 41.32788 46.38641
```

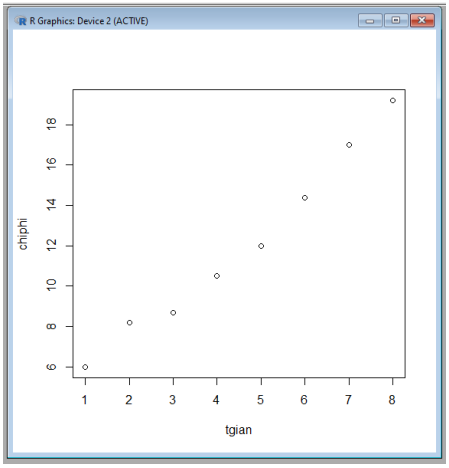
Bài 7. *Phân tích chi phí bảo dưỡng cho xe tải trong 8 năm sử dụng đầu tiên (tính theo triệu đồng) ta có kết quả:*

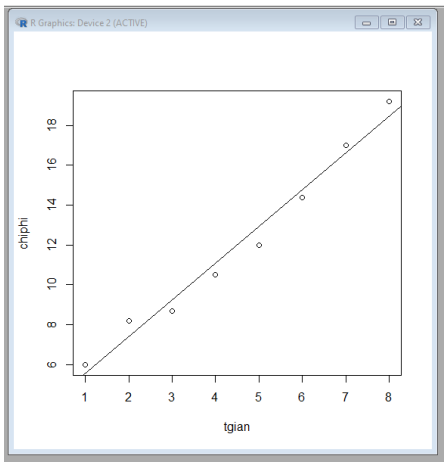
Năm thứ	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Chi phí TB	*6*	*8, 2*	*8, 7*	*10, 5*	*12*	*14, 4*	*17*	*19, 2*

a) Hãy tìm hàm xu thế tuyến tính biểu thị chi phí bảo dưỡng xe.
b) Vẽ hình mô tả dữ liệu và đồ thị hàm hồi quy tuyến tính thực nghiệm.
c) Xác định sai số của dữ liệu được cung cấp và hàm hồi quy thực nghiệm tại các điểm quan sát.
d) Dự báo chi phí bảo dưỡng trung bình cho xe trong năm sử dụng thứ 10 và tìm khoảng tin cậy 90% cho giá trị đó.

Kết quả đối chiếu

1. Kết quả tính các hệ số hồi quy $\widehat{\beta_0}, \widehat{\beta_1}$
   Call:
   lm(formula = chiphi ~ tgian)
   Coefficients:
   (Intercept) tgian
   3.696 **1.845**
2. Hình ảnh của các mô tả trực quan





   Sai lệch giữa giá trị quan sát và hàm hồi quy
   > resid (reg)

1	2	3	4	5
0.4583333	0.8130952	-0.5321429	-0.5773810	-0.9226190
6	7	8		
-0.3678571	0.3869048	0.7416667		

3. Điểm ước lượng cho lượng vận chuyển của công ty vận tải tại thời điểm được lựa chọn là
   > y0mu
   (Intercept)
   22.14881
4. Kết quả tính các biến trong công thức độ tin cậy
   > xtb
   [1] **4.5**
   > Sxx
   [1] **42**
   > sbp
   [1] **0.5290079**
5. Phân vị và bán kính khoảng ước lượng
   > phanvi
   [1] **1.94318**

> bkinh
[1] **1.299373**

6. Khoảng tin cậy 95% cho lượng vận chuyển của công ty vận tải tại thời điểm được lựa chọn là

> y0mu+c(-1,1)*bkinh

[1] **20.84944 23.44818**

III. Bài tập làm thêm

Bài 8. Tốc độ xói mòn đất tại một công trường xây dựng được xem là hàm của độ dốc của khu vực địa hình đó. Dữ liệu về tốc độ xói mòn đất và độ dốc của một số điểm khảo sát được cho dưới đây:

Độ dốc (%)	1,2	1,6	2,4	3,2	3,6	4,1	4,9
Tốc độ xói mòn (tấn/ha/năm)	38	78	55	84	52	111	94

a) Vẽ biểu đồ phân tán của dữ liệu.
b) Hãy xác định đường hồi quy tuyến tính thực nghiệm biểu diễn tốc độ xói mòn theo độ dốc.

Bài 9. Tại một trường đại học, môn giải tích là điều kiện tiên quyết để sinh viên có thể học môn thống kê. Người ta lấy mẫu ngẫu nhiên 10 sinh viên đã hoàn thành cả hai môn học và ghi lại điểm của các sinh viên đó. Dữ liệu được cho dưới đây:

Giải tích	6,5	5,8	9,3	6,8	7,4	8,1	5,8	8,5	8,8	7,5
Thống kê	7,4	7,2	8,4	7,1	6,8	8,5	6,3	7,3	7,9	8,5

a) Tìm đường hồi quy tuyến tính biểu diễn điểm thống kê theo điểm giải tích.
b) Vẽ biểu đồ phân tán và đường hồi quy tuyến tính thực nghiệm. Dựa vào đồ thị để nhận xét về quan hệ giữa điểm của hai môn học.

Bài 10. Quảng cáo được xem là chìa khóa dẫn đến thành công. Để đánh giá hiệu quả của quảng cáo đến doanh thu, nhà quản lý của một chuỗi cửa hàng bán lẻ thu thập dữ liệu về doanh thu và chi phí dành cho quảng cáo (đơn vị: triệu đồng) từ các cửa hàng trong $n = 8$ tuần gần nhất. Dữ liệu được ghi lại trong bảng dưới đây.

Chi phí QC	3,0	7,0	6,5	3,5	4,5	7,0	7,5	8,5
Doanh thu	50	200	150	75	100	180	190	210

a) Hãy tính hệ số tương quan mẫu.
b) Hãy tìm hồi quy tuyến tính biểu diễn doanh thu qua chi phí quảng cáo.
c) Vẽ hình biểu đồ phân tán và đồ thị hàm hồi quy tuyến tính thực nghiệm.

d) Xác định sai số của dữ liệu được cung cấp và hàm hồi quy thực nghiệm tại các điểm quan sát.

e) Dự báo doanh thu đạt được trung bình ứng với chi phí quảng cáo 11 triệu và tìm khoảng tin cậy 95% cho giá trị đó.