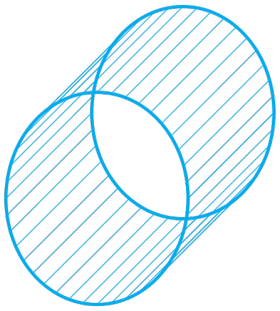
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN TP.HCM
CAO HỌC KHÓA 30

_____*_____



Khoa Toán - Tin học
Fac. of Math. & Computer Science

Bài tập lần 1

MÔ HÌNH HÓA THỐNG KÊ

Giảng viên hướng dẫn: **TS. Nguyễn Thị Mộng Ngọc**
Nhóm thực hiện: **Nhóm 4**

TP. Hồ Chí Minh – Tháng 01, 2021

BẢNG PHÂN CÔNG CÔNG VIỆC

Thành viên	**Công việc**	**Mã số học viên**
1. Đặng Khánh Thi	– Code R bài 1 – Kiểm tra code R bài 2, bài 3 – Thảo luận ghi nhận xét 3 bài	20C29038
2. Đinh Thị Nữ	– Code R bài 2 – Kiểm tra code R bài 1, bài 3 – Thảo luận ghi nhận xét 3 bài	20C29013
3. Lý Phi Long	– Code R bài 3 – Kiểm tra code R bài 1, bài 2 – Thảo luận ghi nhận xét 3 bài	20C29028
4. Phan Thị Thùy An Nhóm trưởng	– Tổng hợp, kiểm tra code R bài 1,2,3 – Thảo luận ghi nhận xét 3 bài – Trình bày file nhận xét bằng Latex	20C29002

Exercise 1 (2.8.1 page 38)

$Y = \beta_0 + \beta_1 x + e$ where Y is the gross box office results for the current week (in \$) and x is the gross box office results for the previous week (in \$).

(a) Find a 95% confidence interval for the slope of the regression model, β_1. Is 1 a plausible value for β_1?

```
> coef(lm_fit)
 (Intercept)     LastWeek
6804.8860355    0.9820815
>
> ### a - Confidence interval for B1
> confint(lm_fit, level = 0.95)[2,]
    2.5 %    97.5 %
0.9514971 1.0126658
```

Dựa vào kết quả mô hình, ta tìm được 95% khoảng tin cậy cho giá trị β_1 là $[0.9515; 1.0127]$ Nhận thấy giá trị 1 thuộc khoảng tin cậy của β_1, vì vậy có thể xem $\beta_1 = 1$.

(b) Test the null hypothesis $H_0 : \beta_0 = 10000$ against a two-sided alternative.

```
> t_obs = (B0_hat - B0) / B0_se; t_obs
[1] -0.3217858
> p_value = 2 * pt(t_obs, nrow(playbill) - 2); p_value
[1] 0.7517807
```

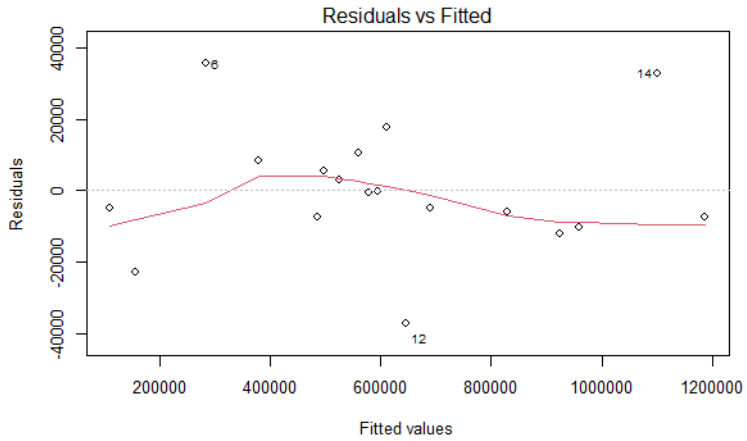
Với $p_{value} = 0.75 \not< \alpha_{default} = 0.05$ chúng ta không đủ cơ sở để bác bỏ giả thuyết H_0.

(c) Find a 95% prediction interval for the gross box office results for the current week (in \$) for a production with \$400,000 in gross box office the previous week. Is \$450,000 a feasible value for the gross box office results in the current week, for a production with \$400,000 in gross box office the previous week?

```
> ### c
> predict(lm_fit, data.frame(LastWeek = 400000), interval = "prediction")
       fit      lwr      upr
1 399637.5 359832.8 439442.2
```

Dễ nhận thấy 450000 không thuộc khoảng dự đoán $[359832.8; 439442.2]$ của mô hình, nên nó không là giá trị phù hợp.

(d) Some promoters of Broadway plays use the prediction rule that next week's gross box office results will be equal to this week's gross box office results.



Dựa vào biểu đồ Residuals vs Fitted, ta thấy rằng chiến thuật dự đoán doanh số tuần hiện tại so với tuần trước là một chiến lược khá hợp lý.

Tuy nhiên, có ít nhất 3 điểm bị lệch khỏi đường hồi quy.

Vì vậy, họ cần thêm dữ liệu để có cái nhìn khách quan hơn hoặc dùng một mô hình khác có thể tạo ra quy tắc dự đoán tốt hơn, giảm thiểu rủi ro.

Exercise 2 (2.8.2 page 38)

$Y = \beta_0 + \beta_1 x + e$ where Y = Percentage change in average price from July 2006 to July 2007; and x = Percentage of mortgage loans 30 days or more overdue in latest quarter.

(a) Find a 95% confidence interval for the slope of the regression model, β_1. On the basis of this confidence interval decide whether there is evidence of a significant negative linear association.

```
> ### a - Find 95% confident interval of B1
> confint(linearModel2, level = 0.95)
                 2.5 %     97.5 %
(Intercept) -2.532112 11.5611000
loanPayment -4.163454 -0.3335853
```

Dựa vào kết quả mô hình, ta tìm được 95% khoảng tin cậy cho giá trị β_1 là $[-4.16; -0.33]$.

Ta có dạng tổng quát của mô hình hồi quy tuyến tính được cho là:

$$Y = 4.514 - 2.249x$$

nên Y và x có mối tương quan nghịch hay x tăng thì Y giảm và ngược lại.

(b) Use the fitted regression model to estimate $E(Y|X = 4)$. Find a 95% confidence interval for $E(Y|X = 4)$. Is 0% a feasible value for $E(Y|X = 4)$?

```
> ### b - Predict Y with X = 4
> result <- predict(linearModel2,data.frame(loanPayment = 4),interval = 'confidence')
> result
        fit       lwr       upr
1 -4.479585 -6.648849 -2.310322
```

Dựa vào mô hình tổng quát mối liên hệ giữa 2 biến x và Y ta tính được giá trị ước lượng của Y tại $x = 4$ là $Y = -4.48$

Tương tự ta tính được 95% độ tin cậy của Y khi $x = 4$ nằm trong khoảng $[-6.6; -2.3]$

Vì 95% độ tin cậy của y khi $x = 4$ nằm trong khoảng $[-6.6; -2.3]$ nên 0% không phải là giá trị mong đợi phù hợp khi $x = 4$.

Exercise 3 (2.8.3 page 38)

$Y = \beta_0 + \beta_1 x + e$ where Y is the processing time and x is the number of invoices.

(a) Find a 95% confidence interval for the start-up time, i.e., β_0.

```
> ### a - Find a 95% confidence interval for the start-up time
> confint(fit, "(Intercept)", level=0.95)
                2.5 %    97.5 %
(Intercept) 0.3912496 0.8921701
```

Dựa vào kết quả mô hình, ta tìm được 95% khoảng tin cậy cho giá trị β_0 là $[0.3912; 0.8922]$

(b) Test the null hypothesis $H_0 : \beta_1 = 0.01$ against a two-sided alternative.

```
> ### b - Test the null hypothesis H0: B1 = 0.01 against a two-sided alternative
> confint(fit, "Invoices", level=0.95)
                2.5 %     97.5 %
Invoices 0.009615224 0.01296806
```

$\beta_1 = 0.01$ nằm trong khoảng tin cậy vừa tìm được, do đó chưa đủ cơ sở để bác bỏ H_0 với mức ý nghĩa $\alpha = 5\%$

(c) Find a point estimate and a 95% prediction interval for the time taken to process 130 invoices.

```
> ### c - Find a point estimate and a 95% prediction interval for the time taken to process
 130 invoices
> predict(fit, data.frame(Invoices = 130), interval="prediction")
       fit      lwr    upr
1 2.109624 1.422947 2.7963
```

Dựa vào mô hình dự đoán, thời gian cần để xử lý 130 hóa đơn là 2.1 giờ với khoảng tin cậy 95% là $[1.4; 2.8]$ hours.