Лабораторна робота №4
з асимптотичної статистики
Варіант №4

Горбунова Даніела Денисовича
4 курс бакалаврату
група "комп'ютерна статистика"

16 травня 2021 р.

1 Вступ.

У даній роботі для нелінійно залежних даних була коректно підібрана регресійна формула та оцінені невідомі коефіцієнти у ній. Додатково наводяться довірчі інтервали коефіцієнтів та короткі відомості про модель.

2 Досліджувані дані.

Досліджується вибірка з $N = 500$ елементів, форма якої визначає певну функцію з коливаннями. Покажемо діаграму розсіювання початкових даних.

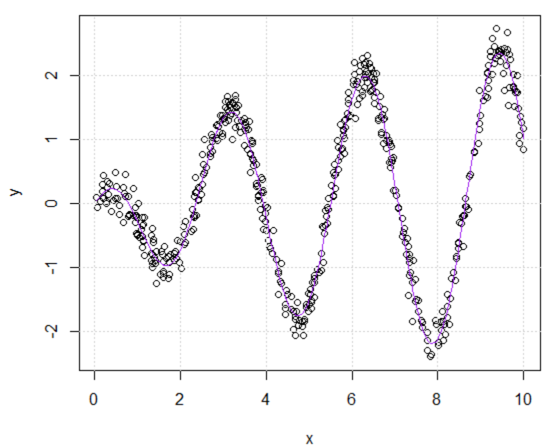

Рис. 1: Графічна інтерпретація даних.

Визначимо формулу регресії з нескладних емпіричних міркувань, що наведемо далі.

3 Визначення формули та підгонка.

З діаграми (2) ми зауважимо, що коливання повторюються із сталим періодом, який приблизно дорівнює $T \approx 3.15$.

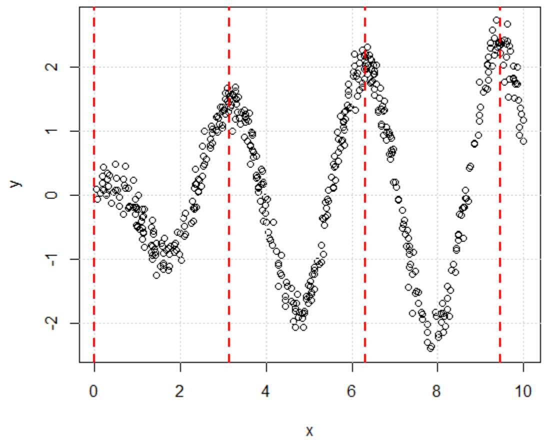

Рис. 2: Та сама інтерпретація, лише провели червоні вертикальні прямі в точках $T \cdot k$, $k = \overline{0, 3}$.

Будемо припускати, що дані поточного завдання описуються двома можливими формулами:

$$f(x, a)\sin(bx) \text{ або } f(x, a)\cos(bx),\ a, b > 0$$

З рисунку видно, що екстремальні значення зростають досить повільно, як логарифм. Тому покладемо $f(x, a) = \ln(a + x)$. Саме значення $a > 0$ обчислимо, наприклад, в $x_0 = T$:

$$\ln(a + T) \approx \frac{3}{2} \Leftrightarrow a \approx \exp\left(\frac{3}{2}\right) - T \approx 1.331689$$

Оскільки $a > 1$, тому логарифм $\ln(a + x)$ затухає, спрямовуючи $x - a + 1$.

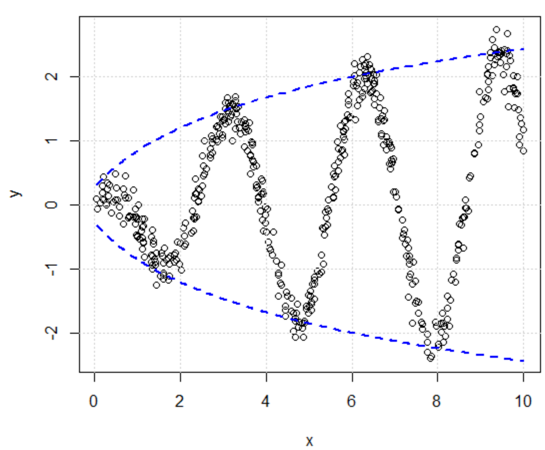

Рис. 3: Та сама інтерпретація, лише додали гілки $\pm \ln(a + x)$.

В якості кандидата на тригонометричну функцію беремо $\cos(bx)$, оскільки в нулі у нас значення на початку координат скупчуються трохи вище нуля – приблизно біля $\ln(a+0) \approx 0.2864481$ (хоча висновок про косинус можна було б раніше зробити з простіших міркувань, що базуються на точках максимуму при значеннях $T \cdot k, k \geqslant 1$). Оскільки нам відомий період T, то $b > 0$ знаходимо за один крок:

$$\cos(b(x+T)) = \cos(b \cdot x) \Rightarrow b \cdot T \approx 2\pi \Rightarrow b \approx \frac{2\pi}{T} \approx 1.994662$$

Отримані приблизні значення параметрів нам знадобляться для підгонки моделі за допомогою nls() в стандартному пакеті R:

```
# Параметри, отримані з ідейних міркувань
T <- 3.15
a.0 <- exp(3/2)-T
b.0 <- 2*pi/T
# Підгонка моделі
nonlin.model <- nls(y ~ log(a+x)*cos(b*x), start=list(a=a.0, b=b.0))
print(summary(nonlin.model))
```

Результати підгонки у формі звіту:

```
Parameters:
  Estimate Std. Error t value Pr(>|t|)
a  1.00345    0.04202   23.88   <2e-16 ***
b  1.99886    0.00106 1885.61   <2e-16 ***
Residual standard error: 0.1944 on 498 degrees of freedom
```

Начебто все гаразд із оціненими коефіцієнтами, вони відрізняються від нуля значущо для будь-якого адекватного рівня значущості α (надалі ми покладемо $\alpha = 0.05$). Лише трохи прогадали з a, коли виводили вигляд рівняння. Подивимося поведінку залишків моделі на QQ-діаграмі та гістограмі відносних частот.

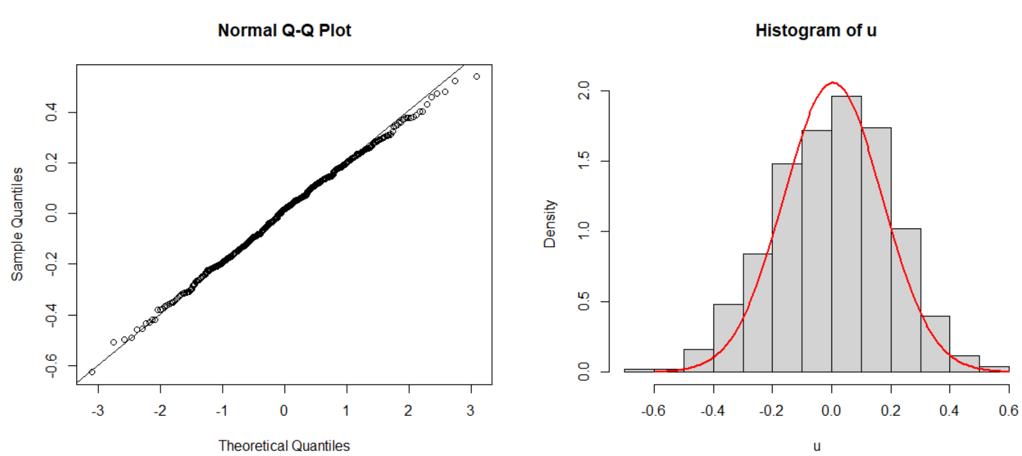

Рис. 4: QQ-діаграма (зліва) та гістограма відносних частот (справа, з підігнаною щільністю нормального розподілу) залишків моделі.

Форма розподілу нагадує гауссовий, значущих відхилень на квантильній діаграмі не помічаємо. Із залишками моделі все гаразд.

Залишається побудувати довірчі інтервали для оцінених параметрів. Це ми зробимо для фіксованого вірогідного рівня $1 - \alpha = 0.95$, підрахувавши:

$$\hat{\theta}_N^{\pm} = \hat{\theta}_N \pm Q^{N(0,1)}(1 - \alpha/2)\hat{\sigma}_N \sqrt{\frac{\overline{a}_N^{ii}}{N}},$$

де $\overline{a}_N^{ii}$ - ii-тий діагональний елемент матриці $(A_N)^{-1}$, де A_N визначена за формулою:

$$A_N = \frac{1}{N} \sum_{j=1}^{N} \nabla g(X_j; \hat{\theta}_N)(\nabla g(X_j; \hat{\theta}_N))^{\top}$$

$$\nabla g(X_j; \theta) = \nabla g(X_j; \theta_1, \ldots, \theta_d) = \left(\frac{\partial g}{\partial \theta_1}, \ldots, \frac{\partial g}{\partial \theta_d} \right)^{\top}$$

Побудовані інтервали такі:

	2.5%	97.5%
a	0.9225204	1.088126
b	1.9967682	2.000947

4 Висновки.

Вдалося визначити нелінійну залежність початкових даних та побудувати адекватну регресійну модель. Остаточна модель вигляд:

$$Y_j = \ln(1.00345 + X_j) \cdot \cos(1.99886 \cdot X_j),\ j = \overline{1, N} + \varepsilon_j,\ \varepsilon \sim N(0, (0.1944)^2)$$