Київський національний університет імені Тараса Шевченка
Кафедра теорії ймовірностей, статистики та актуарної математики

Р. Майборода

Самостійна робота по курсу

"Асимптотична статистика"

Індивідуальні завдання
та рекомендації по виконанню

Київ — 2017

УДК 519.22.35
ББК 22.172я73

Майборода Р.Є. Самостійна робота по курсу “Асимптотична статистика” (індивідуальні завдання та рекомендації по виконанню). Навчальний посібник.

Посібник містить 10 варіантів завдань для самостійного виконання, що охоплюють курс “Асимптотична статистика” для студентів механіко-математичного факультету спеціальності “статистика”. Завдання нелінійної регресії, перевірки узгодженості розподілу та незалежності тестом χ^2 і класифікації спостережень з використанням логістичної регресії. Для кожного завдання наведено рекомендації по виконанню та приклади виконання аналогічних завдань. Основна увага приділена змістовній інтерпретації результатів статистичного аналізу.

Вступ

Цей посібник призначений для використання при вивченні курсу "Асимптотична статистика", який читають студентам четвертого курсу механіко-математичного факультету спеціальності математика, що спеціалізуються у галузі теорії ймовірності та математичної статистики. Метою курсу є поглиблене вивчення математичної статистики та її практичних застосувань, порівняно з загальним курсом теорії ймовірностей, статистики та випадкових процесів, який читають на третьому курсі. Виконуючи самостійну роботу студенти можуть познайомитись з тим, як теоретичні знання, отримані ними на лекціях, використовуються у прикладній статистиці. В усіх завданнях самостійної роботи передбачається статистичний аналіз деяких даних: модельованих, реальних наданих викладачем, або таких, які студент сам збирає у цікавій йому області.

Для проведення статистичної обробки рекомендовано застосовувати систему програмування `R`. (Студенти, знайомі з іншими статистичними системами, наприклад, STATISTICA або SPSS, можуть виконувати завдання у цих системах). Останню версію `R` для операційної системи Windows можна отримати за адресою:

`http://cran.r-project.org/bin/windows/base/`

На цьому ж сайті є версії `R` і для інших операційних систем. Мінімальні необхідні знання з `R` можна знайти у [3]. Встановивши `R` на своєму комп'ютері студенти можуть виконувати завдання цілком самостійно і оформлювати результати у вигляді звітів, які подаються викладачу. У звітах потрібно вказувати постановку задачі, результати статистичної обробки комп'ютерною програмою та змістовну інтерпретацію отриманих результатів. Приклади, наведені у посібнику, показують, як це можна робити.[1]

Основні теоретичні відомості, потрібні для виконання завдань, можна знайти у підручниках [1] і [5]. Для додаткового вивчення лінійної регресії можна рекомендувати [4], а уявлення про логістичну регресію з бінарним відгуком дає [6].

Дані для виконання самостійної роботи а також використані у прикладах знаходяться у архівованому файлі разом з текстом цього посіб-

[1] Детальний опис того, які саме функції були використані для отримання результатів у звіті не потрібний. Але, звичайно, бажано, щоб студент вмів відтворити на комп'ютері ту статистичну обробку, яка реалізована у звіті.

ника. Його можна отримати на сайті кафедри теорії ймовірностей, статистики та актуарної математики за адресою:

`http://probability.univ.kiev.ua/userfiles/mre/asympt.rar`

Робота 1. Перевірка узгодженості розподілу.

Завдання роботи.

Варіанти 1-3: Поїзди у метро ходять регулярно за розкладом, кожні 10 хвилин. У файлі metro.txt містяться дані про тривалість чекання пасажирів на станції до приходу поїзда (у хвилинах). Перевірте за змінною, вказаною у вашому варіанті, чи враховують пасажири розклад? (Основна гіпотеза – не враховують).

1. Var1;

2. Var2;

3. Var3.

Варіанти 4-6: У файлі haker.txt містяться дані про спроби несанкціонованого доступу до інформації в інтернеті. Перевірте за змінною, вказаною у вашому варіанті, чи мала місце хакерська атака, чи спроби були результатом випадкових помилок.(Основна гіпотеза – атаки немає).

4. Var1 – інтервали між послідовними спробами несанкціонованого доступу;

5. Var2 – кількість спроб протягом чергового 30-хвилинного інтервалу

6. Var3 – час від початку спостережень, в який відбулась чергова спроба.

Варіанти 7-9: У файлі fish.txt знаходяться дані про довжину трирічних риб одного виду, виловлених у певному озері. Перевірте за за змінною, вказаною у вашому варіанті, чи можна стверджувати, що риби належать до кількох різних популяцій. (Основна гіпотеза – одна популяція).

7. Var1;

8. Var2;

9. Var3.

Варіант 10. У таблиці

48:12	40:24	64:28	110:38	129:37	50:24	74:31
19:7	74:31	94:18	75:21	45:26	102:33	84:16

вміщено дані про розщеплення рослин гороху за кольором квітки у досліді з генетики. Перевірте, чи відповідають ці дані закону Менделя, за яким співвідношення має має коливатись навколо теоретичного значення 3:1.

У наступному прикладі показано, як застосовувати тест χ^2 для перевірки того, що дані мають розподіл, заданий з точністю до невідомих параметрів.

Рекомендації по виконанню. У варіантах 1-9 потрібно використати техніку перевірки узгодженості теоретичного розподілу з розподілом даних на основі тесту χ^2. (Про теорію тестів χ^2, див., наприклад, [5], п. 18.2, або [1] п. 3.15.4). Для цього слід визначити, який теоретичний розподіл відповідає гіпотезі, що перевіряється у вашому варіанті. Якщо потрібно, дані слід перетворити так, щоб утворилась вибірка з незалежних однаково розподілених випадкових величин. Якщо теоретичний розподіл даних є неперервним, їх потрібно групувати, щоб отримати дані з дискретним розподілом.

Далі можна оцінити невідомі параметри і застосувати тест χ^2.

При роботі з даними, що мають непервний розподіл, потрібно вивести гістограму разом із теоретичною щільністю розподілу. При роботі з дискретними даними — графік теоретичних та емпіричних частот.

У 10-му варіанті потрібно скласти статистику χ^2 виходячи з того, що дані у різних клітинках таблиці є незалежними між собою і є результатами спостережень незалежних випробувань з відношенням ймовірності успіху до ймовірності невдачі 3:1.

Приклад 1. У наборі даних `airquality` містяться дані щоденних вимірювань метеорологічної станції у Нью-Йорку з травня по вересень 1973р. Зокрема, змінна `airquality$Wind` вказує силу вітру у відповідний день.. Досить природним є припущення про нормальність розподілу сили вітру. Нарисуємо гістограму даних разом із теоретичною щільністю нормального розподілу (рис. 1).

```
> g = airquality$Wind
> m<-mean(g)
> std<-sqrt(var(g))
> #  гістограма абсолютних частот
> #
> hi<-hist(g, density=20, breaks=10,
+ xlab="x-variable", ylim=c(0, 45),
+ main="absolute frequencies")
> curve(dnorm(x, mean=m, sd=std)
+ *length(g)*(hi$breaks[2]-hi$breaks[1]),
+ col="darkblue", lwd=2, add=TRUE, yaxt="n")
```

На гістограмі помітні відхилення від нормальності. Перевіримо, наскільки значущими є ці відхилення, використовуючи тест χ^2.

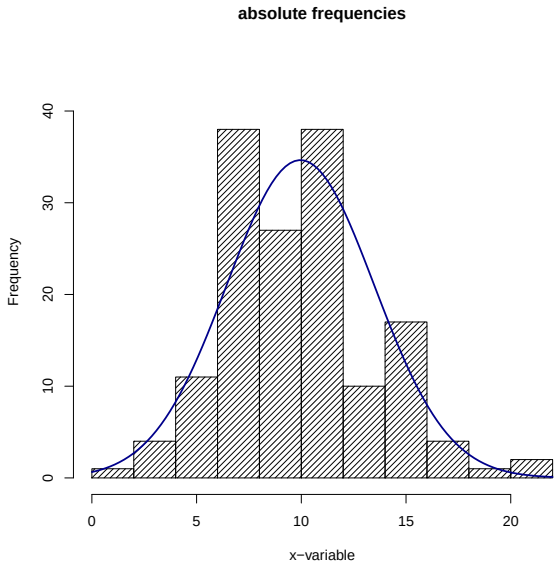

Рис. 1: Гістограма з графіком щільності

Для застосування тесту χ^2 ми оцінимо математичне сподівання `m` та середньоквадратичне відхилення `sd` за групованими даними, причому при оцінці `sd` використаємо поправку Шеппарда. Такий підхід є стандартним при перевірці нормальності.

```
> g = airquality$Wind # дані для аналізу
> # виконуємо групування:
> r<-hist(g,  breaks=10,plot = FALSE)
> nn<-r$counts   # емпіричні частоти
> tt<-r$breaks   # межі комірок
> h<-tt[2]-tt[1] # ширина комірки
> x<-tt[-length(tt)]+h/2 # середини комірок
> m<-sum(x*nn)/sum(nn)   # оцінка математичного сподівання
> # оцінка сер.кв.відх. з поправкою Шеппарда:
> s<-sqrt(sum((x-m)^2*nn)/sum(nn)+h^2/12)
> pp<-pnorm(tt,mean=m,sd=s) # теоретичні ймовірності
> pp[c(1,length(tt))]<-c(0,1) # розширюємо крайні комірки
> nth<-length(g)*(pp[-1]-pp[-length(pp)]) # теоретичні частоти
> chi2emp<-sum((nn-nth)^2/nth) # статистика тесту хі-квадрат
```

```
> 1-pchisq(chi2emp,df=length(tt)-4) # досягн. рівень значущості

[1] 0.0007148321
```

Ми отримали досягнутий рівень значущості $p = 0.0007148$, отже основну гіпотезу треба відхилити: розподіл даних значущо відрізняється від нормального.

Робота 2. Перевірка залежності двох змінних.

Завдання роботи.

Варіанти 1-5: Відомі вірші російського поета-декабриста Г.Батенькова (1793—1863) розділяються на дві частини: (Б, безсумнівні) ті, які були надруковані до 70-х років XXст. і (С, сумнівні) ті, що опублікував філолог А.А. Ілюшин у 70-ті роки. Порівнявши ці дві групи за різними характеристиками віршового метру та ритму, М.І.Шапір дійшов висновку, що вірші групи С написані самим А.А. Ілюшиним. Використовуючи наступні дані визначте, чи є значущими відмінності, виявлені М.І.Шапіром.

Варіанти:

1. Метричний репертуар ямбічних віршей (у відсотках)

	4 стопний	5 стопний	6 стопний	Вільний	Всього ямбу	Всього рядків
Б	87,8	2,2	0	1,8	91,8	1212
С	89,5	0	2,8	0	92,3	716

2. Ритм наголосів у 4-стопному ямбі. Ритмічні форми (у відсотках)

	I	II	III	IV	V	VI	VII	Всього рядків
Б	25.8	5.2	13.2	44.3	0.1	8.8	0.7	1080
С	29.3	5.5	11.2	45.4	0	8.4	0.2	641

3. Ритм наголосів у першому рядочку строфи. Ритмічні форми (у відсотках)

	I	II	III	IV	V	VI	VII	Всього рядків
Б	24.1	0	10.8	56.6	0	7.2	1.2	83
С	36.4	6.8	13.6	36.4	0	6.8	0	44

4. Ритм наголосів у десятому рядочку строфи Ритмічні форми (у відсотках)

	I	II	III	IV	V	VI	VII	Всього рядків
Б	32.5	6.6	15.8	36.8	0	7.9	0	76
С	27.3	11.4	11.4	43.2	0	4.5	2.2	44

5. Ритм наголосів у другому рядочку строфи. Ритмічні форми (у відсотках)

	I	II	III	IV	V	VI	VII	Всього рядків
Б	20.5	3.6	13.3	47.0	0	15.7	0	83
С	25.0	9.1	11.4	40.9	0	13.6	0	44

Варіанти 6-10: У файлі Values.xls знаходяться дані соціологічного опитування про людські цінності, проведеного в Україні у 2006р. За цими даними перевірте, чи є залежність між змінними, вказаними для вашого варіанту у заданому регіоні:

6. Income і Religimportance у Чернігівській області
7. Workimp і Happyness у м. Києві
8. Income і Workimp у Чернігівській області
9. Workimp і Religimportance у м. Києві
10. Income і Religimportance у Харківській області

(Врахуйте, що від'ємні числа у файлі відповідають пропущеним значенням, відповідні спостереження потрібно вилучити з аналізу).

Рекомендації по виконанню роботи. Для перевірки незалежності потрібно застосувати відповідний тест ξ^2. (Про тести χ^2 для перевірки незалежності, див., наприклад, [5], п. 18.3 та [1], п. 417). При цьому у звіті повинні бути таблиця спряженості ознак, значення статистики χ^2-емпіричне і досягнутий рівень значущості тесту.

Приклад 2. У фреймі даних `survey` з бібліотеки `MASS` містяться дані опитування 237 студентів університету Аделаїди. Зокрема у змінній `Smoke` знаходяться відповіді цих студентів на запитання про те, як багато вони палять сигарет (`Heavy` — багато, `Regul` — регулярно, `Occas` — зрідка і `Never` — ніколи) а у змінній `Exer` — відповіді про те, як часто вони роблять фізичні вправи (`Freq` — часто, `Some` — інколи, `None` — ніколи). Нас цікавить, чи є залежність між звичками до паління і до фізичних вправ?

Щоб з'ясувати це побудуємо табличку, де вказуються частоти всіх можливих пар відповідей на ці запитання. Це робиться за допомогою функції `table()`:

```
> library(MASS)
> tbl = table(survey$Smoke, survey$Exer)
> tbl
```

```
       Freq None Some
Heavy     7    1    3
Never    87   18   84
Occas    12    3    4
Regul     9    1    7
```

Тепер у змінній `tbl` знаходиться табличка (матриця) емпіричних частот пар відповідей. Наприклад, варіант відповідей `(Never,Some)` обрало 84 студенти серед всіх 237 що взяли участь у опитуванні. Такі таблиці прийнято називати **таблицями спряженості** змінних/ознак (англ. contingency table). Якщо змінні незалежні, то розподіли студентів по

змінній `Exer` при фіксованій змінній `Smoke` мають бути приблизно однаковими для всіх значень `Smoke`. Тобто всі рядочки таблиці повинні бути приблизно пропорційними (відрізнятись лише сталими множниками). Те ж має виконуватись і для стовпчиків.

Проглядаючи табличку бачимо, що строгої пропорційності немає, але і надзвичайно сильних відхилень не помітно. Застосуємо тест χ^2.

```
> chisq.test(tbl)


        Pearson's Chi-squared test

data:  tbl
X-squared = 5.4885, df = 6, p-value = 0.4828
```

Функція підрахувала статистику $\chi^2_{emp} = 5.4885$, визначила кількість ступенів вільності `df`=6 і знайшла досягнутий рівень значущості `p-value` =0.4828. Отже треба прийняти основну гіпотезу про незалежність між звичками до паління та до виконання фізичних вправ.

Можна було можна передати функції як параметр не таблицю спряженості а безпосередньо змінні, що нас цікавлять:

```
chisq.test(survey$Smoke, survey$Exer)
```

функція сама підрахує емпіричні частоти і видасть ті ж значення, що і у попередньому випадку.

Робота 3. Асимптотичний аналіз тестів відношення вірогідності.

Завдання роботи.

1. Побудуйте тест для перевірки простої основної гіпотези H_0 проти альтернативи H_1 для вибірки обсягу n. Поріг тесту спочатку визначте використовуючи нормальну апроксимацію відношення вірогідності, а потім, при потребі, уточніть, застосовуючи імітаційне моделювання.

2. Оцініть потужність отриманого тесту використовуючи наближення для фіксованих альтернатив та для альтернатив, що зближуються. Перевірте точність цих наближень, використовуючи імітаційне моделювання.

3. Визначте мінімальний обсяг вибірки, при якому ймовірність помилки мінімаксного тесту не перевищує 0.05. (Далі Exp(λ) – експоненційний розподіл з інтенсивністю λ; Beta(α, β) – бета-розподіл з параметрами α, β; Poisson(λ) – Пуассонів розподіл з параметром λ; Binom(p, m) – біноміальний розподіл з ймовірністю успіху p та кількістю випробувань m)

Варіанти:

1. H0 : Exp(7), H1 : Exp(9), n=80.
2. H0 : Beta(2,5), H1 : Beta(2,6), n=65.
3. H0 : Poisson(0.5), H1 : Poisson(0.3), n=75.
4. H0 : Binom(0.5,2), H1 : Binom(0.6,2), n=65.
5. H0 : Beta(4,3), H1 Beta(5,3) : n=50.
6. H0 : Poisson(0.3), H1 : Poisson(0.5), n=45.
7. H0 : Exp(6), H1 : Exp(5), n=55.
8. H0 : Beta(2,5), H1 : Beta(2,4), n=50.
9. H0 : Binom(0.7,2), H1 : Binom(0.5,2), n=25.
10. H0 : Poisson(3), H1 : Poisson(4), n=45.

Робота 4. Нелінійна регресія.

Завдання роботи.

У файлах D<варіант>.txt містяться дані про значення змінних x та y для 500 спостережень. Залежність між ними описується формулою, вигляд якої є різним у різних варіантах. Можливі такі залежності:

$y = \exp(ax)\sin(bx) + \varepsilon$,
$y = \exp(ax)\cos(bx) + \varepsilon$,
$y = \ln(a + x)\sin(bx) + \varepsilon$,
$y = \ln(a + x)\cos(bx) + \varepsilon$,
$y = \sin(bx)/\ln(a + x) + \varepsilon$,
$y = \cos(bx)/\ln(a + x) + \varepsilon$.

Підберіть правильну регересійну формулу та оцініть невідомі параметри a та b. Вкажіть довірчі інтервали для невідомих параметрів, перевірте, чи значущо вони відрізняються від 0. Проведіть аналіз залишків, перевірте чи мають похибки регресії нормальний розподіл.

Рекомендації по виконанню роботи.

Для підгонки формули потрібно застосувати нелінійний метод найменших квадратів (див. [3] п.7). Щоб вибрати формулу та задати початкові значення для пошуку оцінок параметрів скористайтесь діаграмою розсіювання даних як описано у наступному прикладі.

Приклад 3. Для підгонки нелінійних регресійних моделей використовується функція `nlm()`. Продемонструємо її роботу на модельованих даних `x`, `y`[2]. Наша мета — вибрати одну з регресійних функцій, що вказані у завданні і оцінити її параметри за даними.

Виведемо діаграму розсіювання даних (див. рис. 2):

```
> plot(x,y,cex=0.5)
```

На рисунку помітно, що при $x \approx 0$ значення змінної y розкидані поблизу 0, отже на роль періодичної складової моделі природно взяти $\sin(bx)$ (а не cos). Якщо провести лінію через точки максимумів коливань, отримаємо криву, схожу на графік експоненти, тому обираємо модель

$y = \exp(ax)\sin(bx) + \varepsilon$.

Якщо спробувати підігнати цю модель за допомогою `nls()` не вказуючи початкові значення параметрів, хорошого результату не отримуємо. Виберемо початкові значення виходячи з діаграми розсіювання. Помітимо, що період T коливань на діаграмі складає приблизно 3.5. Оскільки $bT = 2\pi$, отримуємо $b \approx 2 \cdot 3.14/3.5 \approx 1.79$.

[2] Як саме дані моделювались тут не описую, щоб витримати інтригу.

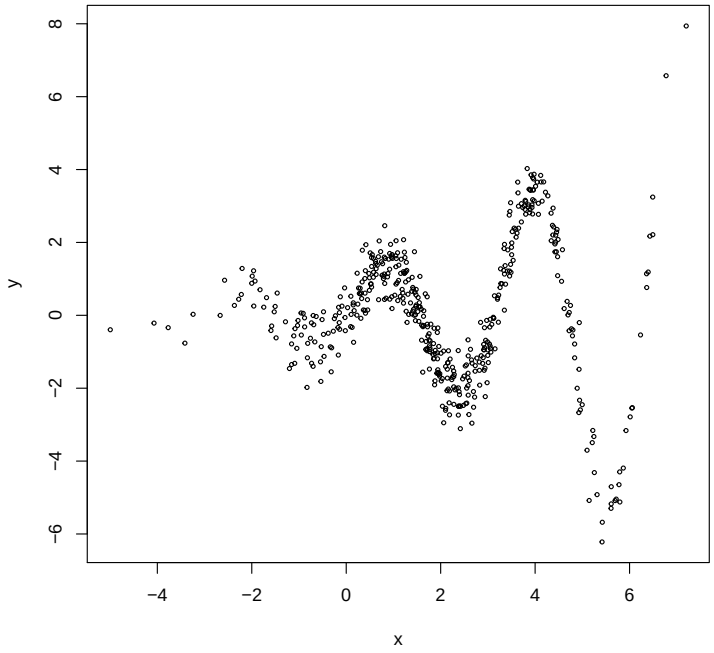

Рис. 2: Дані для підгонки за нелінійним МНК

Тепер помітимо, що кожен максимум на діаграмі приблизно вдвічі вищий ніж попередній. Отже

$$\frac{\exp(a(x+T))}{\exp(aT)} \approx 2,$$

звідки $a \approx \log(2)/T \approx 0.38$.

Тепер можна скористатись нелінійним методом найменших квадратів:

```
> fi<-nls(y~exp(a*x)*sin(b*x),start=list(a=0.38,b=1.79))
> summary(fi)


Formula: y ~ exp(a * x) * sin(b * x)

Parameters:
  Estimate Std. Error t value Pr(>|t|)
a 0.302728   0.002874   105.3   <2e-16 ***
b 2.001034   0.002640   758.0   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5096 on 498 degrees of freedom

Number of iterations to convergence: 5
Achieved convergence tolerance: 5.106e-06

> plot(x,y,cex=0.5)
> a<-coef(fi)["a"]
> b<-coef(fi)["b"]
> curve(exp(a*x)*sin(b*x),col="red",add=T)


> confint(fi,level=0.95)

       2.5%     97.5%
a 0.2970189 0.3082708
b 1.9958269 2.0062371
```

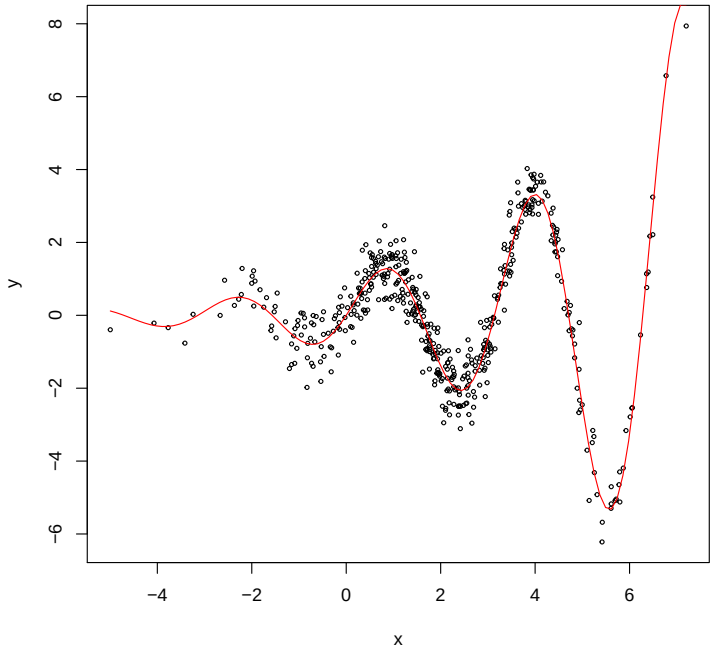

Рис. 3: Результати підгонки за нелінійним МНК

Отримали оцінки параметрів $a = 0.3027$, $b = 2.0010$. Довірчі інтервали з рівнем $\alpha = 0.05$ для a — $[0.2970189, 0.3082708]$, для b — $[1.9958269, 2.0062371]$.

За результатами можна стверджувати, що a і b значущо відрізняються від 0.

Середньоквадратична похибка залишків дорівнює 0.5096.

Проведемо графічний аналіз залишків:

```
> u<-residuals(fi)
> pr_y<-fitted(fi)
> plot(y,pr_y,cex=0.5,xlab="Response",ylab="Prediction")
> abline(0,1,col="red")
> plot(pr_y,u,xlab="Prediction",ylab="Residuals")
> qqnorm(u,col="blue",cex=0.2,main="Residuals normal QQ-diagram")
> qqline(u,col="red")
```

На рис. 4 — результати аналізу. Діаграма розсіювання відгук-прогноз показує, що прогноз добре відтворює поведінку відгуку. На діаграмі прогноз-залишки не помітно яких-небудь закономірностей, що дозволяли б поліпшити прогноз. QQ діаграма підтверджує нормальний розподіл похибок.

Таким чином, наша остаточна модель для зв'язку між змінними y та x має вигляд:

$$y = \exp(0.3027x)\sin(2.001x) + \varepsilon,$$

де ε — гауссова похибка з нульовим математичним сподіванням і дисперсією 0.2597.

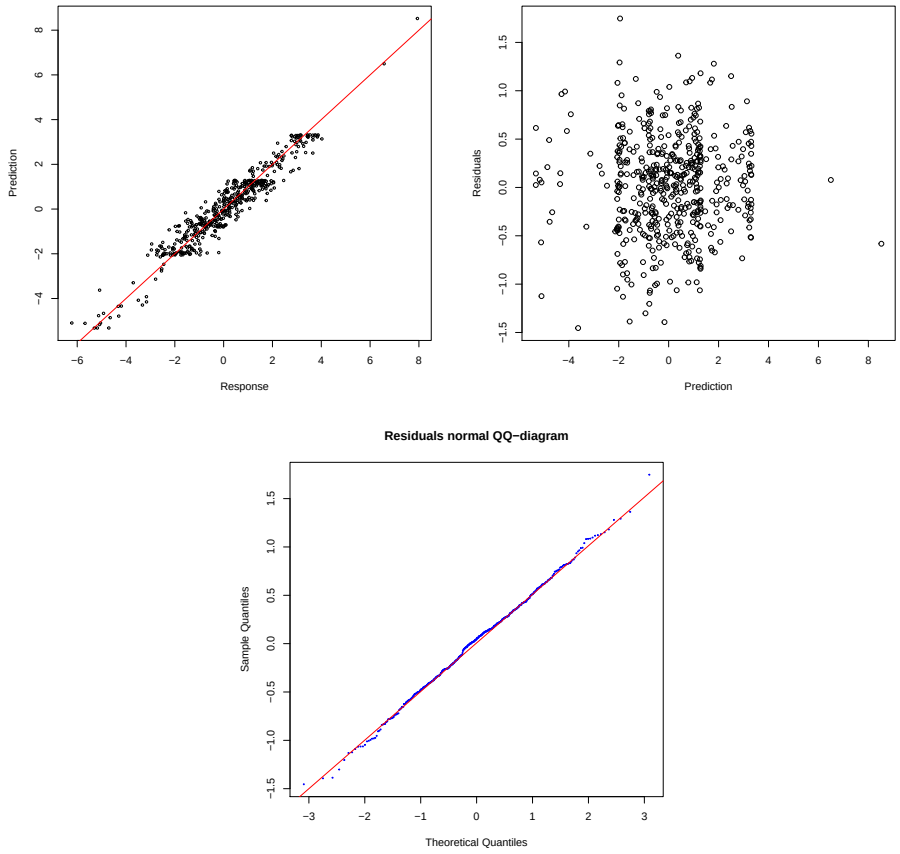

Рис. 4: Аналіз залишків

Робота 5. Логістична регресія.

Завдання роботи.

У файлі wine.csv знаходяться результати аналізу зразків вина, вибраного з трьох різних виноградників. Номер винограднику міститься у змінній site. Використовуючи техніку логістичної регресії, визначте апостеріорну ймовірність того, що зразок із заданими характеристиками належить даному винограднику з двох вказаних у завданні. Перевірте якість класифікації. Які з вказаних змінних можна не використовувати для класифікації?

Варіанти:

N	Виноградники	Змінні
1	1 2	Alcohol; Malic_acid; Ash; Alcalinity_of_ash
2	2 3	Magnesium; Phenols; Flavanoids;NF
3	3 1	Proanthocyanins;Color intensity; Hue;OD
4	1 2	Proline; Alcohol; Malic acid; Ash
5	2 3	Alcalinity_of_ash;Magnesium; Phenols; Flavanoids
6	3 1	NF; Proanthocyanins;Color_intensity; Hue
7	1 2	OD; Proline; Alcohol; Malic_acid
8	2 3	Ash; Alcalinity_of_ash;Magnesium; Phenols
9	3 1	Flavanoids; NF; Proanthocyanins;Color_intensity
10	1 2	Magnesium; Phenols; Flavanoids;NF

Рекомендації по виконанню роботи.

У роботі потрібно провести підгонку моделі логістичної регресії, використовуючи дані про виноградник, де виготовлено зразок, як бінарний відгук (один з виноградників відповідає 0, інший — 1). На основі підігнаної моделі провести класифікацію зразків вина і порівняти результат із справжніми даними. Спробувати вилучити деякі зі змінних-регресорів і повторити підгонку та класифікацію. Вибрати найменший набір змінних, який, на вашу думку, дає хорошу класифікацію. (Див. [6], п. 4.3).

Приклад 4. Розглянемо підгонку моделі логістичної регресії за даними `wine.csv` для опису апостеріорної ймовірності того, що вино вироблено на 2-му або 3-му винограднику за змінними `Alcogol`, `Ash`, `Flavanoids`, `Proline`. Для цього використаємо функцію `glm()` (узагальнена лінійна модель) з опцією `family = binomial()`, яка задає модель логістичної регресії з бінарним відгуком.

Дані з першого виноградинка, непотрібні для підгонки, видаляємо. На роль відгуку використовуємо змінну `Site`, у якій 2 замінено на 0, 3 — на 1, оскільки відгук має бути бінарним.

```
> w2<-read.csv2("c:/rem/term/wine.csv",header=T) # читаємо дані з файлу
> w2<-w2[w2$Site!=1,] # вилучаємо перший виноградник
> w2$Site<-w2$Site-2  # позначаємо 2->0, 3->1
> attach(w2)          # приєднуємо дані
> # проводимо підгонку моделі:
> res<-glm(Site~Alcogol+Ash+ Flavanoids+Proline,family = binomial())
> res


Call:  glm(formula = Site ~ Alcogol + Ash + Flavanoids + Proline, family = binomia

Coefficients:
(Intercept)      Alcogol          Ash   Flavanoids      Proline
 -95.822206     7.650032     9.837367   -15.678226    -0.008844

Degrees of Freedom: 118 Total (i.e. Null);  114 Residual
Null Deviance:            160.5
Residual Deviance: 13.94         AIC: 23.94

> confint(res,level=0.95)

                    2.5 %        97.5 %
(Intercept) -229.40316923 -35.919561629
Alcogol        2.76291328  18.728951274
Ash            3.74228242  22.846608650
Flavanoids   -40.12120714  -6.541050025
Proline       -0.02830901   0.003999477

> # робимо класифікацію виноградників:
> clasif<-as.numeric(fitted.values(res)>0.5)
> # будуємо таблицю помилок класифікації:
> table(clasif,Site)

      Site
clasif  0  1
     0 69  1
     1  2 47
```

Отримали модель для ймовірності того, що вино вироблено на 3-му винограднику:

$$\mathsf{P}\{\texttt{Site} = 3\} = \text{Logist}(-95.822206 + 7.650032\texttt{Alcogol} + 9.837367\texttt{Ach}$$

$$-15.678226\texttt{Flavanoids} - 0.008844\texttt{Proline})$$

При цьому довірчі інтервали для коефіцієнтів не включають 0, тому у цій формулі залежності від всіх змінних слід вважати значущими.

За таблицею помилок класифікації бачимо, що якість класифікатора на основі цієї моделі є задовільною — частоти помилок для 2-го виноградинка $2/71 = 0.028$, для третього — $1/48 = 0.0208$.

Спробуємо вилучити деякі змінні і перевірити, наскільки при цьому погіршиться якість класифікації[3]:

```
> res<-glm(Site~Alcogol+ Flavanoids,family = binomial())
> confint(res,level=0.95)

                 2.5 %     97.5 %
(Intercept) -92.630078 -18.880314
Alcogol       2.043395   8.399333
Flavanoids  -13.357278  -4.325866

> clasif<-as.numeric(fitted.values(res)>0.5)
> table(clasif,Site)

      Site
clasif  0  1
     0 69  2
     1  2 46
```

Бачимо, що, незважаючи на зменшення кількості регресорів, частота помилок класифікації збільшилась несуттєво. Цю модель обираємо як остаточну:

$$\mathsf{P}\{\texttt{Site} = 3\} = \text{Logist}(-92.630078 + 2.043395\texttt{Alcogol} - 13.357278\texttt{Flavanoids})$$

[3]Потрібно перевірити різні варіанти і обрати найбільш вдалий на ваш погляд.

Литература

[1] Карташов М.В. "Імовірність, процеси, статистика". Київ, Видавничо-поліграфічний центр “Київський університет”, 2007, 494 с.

[2] Майборода Р.Є. Регресія: Лінійні моделі.- К. ВПЦ "Київський університет 2007, 296с.

[3] Майборода Р.Є., Сугакова О.В. "Аналіз даних за допомогою пакета R". , 2015 65 с.

[4] Себер Дж. Линейный регрессионный анализ.— М.: Мир, 1980.— 456с.

[5] Турчин В.М. Теорія ймовірностей і математична статистика.- Дніпропетровськ, ІМА-пресс, 2014 - 566 с.

[6] James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning with Applications in R.— Springer NY 2013.— 440p.