Практическая работа №5. Линейная регрессия. Оценка адекватности модели, оценка доверительных интервалов параметров.

Юрченков Иван Александрович, ассистент кафедры ПМ

2022-10-10

```
##
##           : 'dplyr'
##                  'package:stats':
##
##     filter, lag
##                  'package:base':
##
##     intersect, setdiff, setequal, union
```

Постановка задачи для выполнения практической работы

Для выполнения практического задания необходимо:

1. Открыть папку, соотвествующую своей группе.
2. Открыть папку с вариантом, совпадающим с вашим номером в списке.

В папке 3 файла с данными.

1. 1-ый файл содержит 2 ряда данных. Первый стоблец $ x$ содержит факторную переменную, второй столбец $ y$ результирующую. Для первого файла необходимо:

- Оценить коэффициент корреляции Пирсона $ r(x, y)$ между двумя переменными в первом и втором столбце.
- По шкале Чеддока оценить хакактеристику корреляционной связи между величинами.
- Проверить статистическую значимость коэффициента корреляции Пирсона с помощью t-статистики.
- Построить доверительный интервал для $ r(x, y)$ с надежностью $\gamma = 0.95$.
- Построить линейную регрессию между столбцами, оценить значение коэффициентов линейной зависимости.
- Оценить значимость полученных коэффициентов прямой.
- Построить доверительные интервалы для полученных коэффициентов.
- Оценить адекватность модели *по коэффициенту детерминации*.
- Оценить интервал прогноза для линейной модели на $ ft = 3$ значения вперед.

2. 2-ой файл содержит 4 ряда данных. Первый ряд (столбец) содержит количественную факторную переменную, следующие два - качественную факторную переменную, последний - результирующую переменную. Для второго файла данных необходимо:

- Необходимо с помощью теста Чоу обосновать необходимость деления выборки по одной из качественных факторных переменных.
- Произвести разбиение и построить две линейных регрессии, оценить коэффициенты моделей.

3. 3-ий файл содержит 2 ряда данных. Для третьего файла данных необходимо:

- Необходимо двумя способами (тест Спирмена и тест Гольдфельда-Квандта) определить, присутствует ли в данных гетероскедастичность.
- Построить линейную регрессию, оценить значения коэффициентов модели.
- Оценить значимость полученных коэффициентов и адекватность модели.
- Все расчеты проводить для уровня значимости $\alpha = 0.05$.

Пример проведения регрессионного анализа для ряда данных

Исследуемый ряд данных

Рассмотрим таблицу переменных парных данных (x, y) одинаковой длины без пропущенных значений для данных о цене алмазов (diamonds) с категориальными параметрами: $cut = Ideal$ (огранка), $color = J$ (цвет), $clarity = SI2$ (чистота).

```
## [1] "x"
```

```
## 1 -1.171
##  2 0.02
##  3 0
##  4 0
##  5 0.077
##  6 0.049
##  7 0.01
##  8 0.039
##  9 0.01
##  10 0.058
##  11 0.01
##  12 0.01
##  13 0.03
##  14 0.03
##  15 0.104
##  16 0.104
##  17 0.01
##  18 0.104
##  19 0.049
##  20 0.095
##  21 0.01
##  22 0.131
##  23 0.095
##  24 0.113
##  25 0.03
##  26 0.049
##  27 0.122
##  28 0.174
```

```
##  29 0.182
##  30 0.182
##  31 0.095
##  32 0.231
##  33 0.182
##  34 0.215
##  35 0.199
##  36 0.239
##  37 0.231
##  38 0.182
##  39 0.182
##  40 0.239
##  41 0.215
##  42 0.231
##  43 0.239
##  44 0.239
##  45 0.086
##  46 0.207
##  47 0.207
##  48 0.293
##  49 0.293
##  50 0.322
##  51 0.445
##  52 0.322
##  53 0.419
##  54 0.315
##  55 0.507
##  56 0.438
##  57 0.438
##  58 0.464
##  59 0.531
##  60 0.536
##  61 0.571
##  62 0.531
##  63 0.698
##  64 0.723
##  65 0.703
##  66 0.723
##  67 0.708
##  68 0.732
##  69 0.698
##  70 0.718
##  71 0.708
##  72 0.703
##  73 0.728
##  74 0.698
##  75 0.829
##  76 0.698
##  77 0.798
##  78 0.829
##  79 0.829
##  80 0.916
##  81 0.916
##  82 0.904
```

```
##  83 0.916
##  84 0.928
##  85 1.102
##  86 0.9
##  87 0.967
##  88 0.959
##  89 1.001
##  90 0.956
##  91 -1.109
##  92 -0.892
##  93 -0.942
##  94 -0.635
##  95 -0.673
##  96 -0.654
##  97 -0.616
##  98 -0.635
##  99 -0.462
##  100 -0.357
##  101 -0.357
##  102 -0.357
##  103 -0.274
##  104 -0.342
##  105 -0.329
##  106 -0.288
##  107 -0.357
##  108 -0.211
##  109 0.02
##  110 0
```

```
## [1] "y"
```

```
## 1 5.841
##  2 7.965
##  3 8.15
##  4 8.168
##  5 8.171
##  6 8.193
##  7 8.225
##  8 8.233
##  9 8.243
##  10 8.277
##  11 8.29
##  12 8.293
##  13 8.296
##  14 8.307
##  15 8.312
##  16 8.317
##  17 8.318
##  18 8.319
##  19 8.325
##  20 8.333
##  21 8.337
##  22 8.346
##  23 8.349
##  24 8.372
```

```
##  25 8.377
##  26 8.381
##  27 8.383
##  28 8.414
##  29 8.414
##  30 8.415
##  31 8.439
##  32 8.446
##  33 8.447
##  34 8.448
##  35 8.45
##  36 8.454
##  37 8.464
##  38 8.465
##  39 8.465
##  40 8.472
##  41 8.473
##  42 8.489
##  43 8.503
##  44 8.521
##  45 8.524
##  46 8.534
##  47 8.548
##  48 8.57
##  49 8.586
##  50 8.66
##  51 8.66
##  52 8.694
##  53 8.714
##  54 8.715
##  55 8.76
##  56 8.825
##  57 8.849
##  58 8.876
##  59 8.918
##  60 8.948
##  61 8.958
##  62 9.048
##  63 9.307
##  64 9.327
##  65 9.334
##  66 9.336
##  67 9.389
##  68 9.407
##  69 9.439
##  70 9.446
##  71 9.451
##  72 9.452
##  73 9.455
##  74 9.458
##  75 9.488
##  76 9.492
##  77 9.525
##  78 9.527
```

```
##  79 9.582
##  80 9.582
##  81 9.632
##  82 9.644
##  83 9.68
##  84 9.68
##  85 9.683
##  86 9.709
##  87 9.736
##  88 9.753
##  89 9.787
##  90 9.818
##  91 5.903
##  92 6.594
##  93 6.111
##  94 6.752
##  95 6.786
##  96 6.829
##  97 6.886
##  98 6.91
##  99 6.971
## 100 7.477
## 101 7.51
## 102 7.513
## 103 7.514
## 104 7.55
## 105 7.563
## 106 7.573
## 107 7.624
## 108 7.65
## 109 7.867
## 110 7.885
```

Table 1: Таблица данных

n	x	y	n	x	y	n	x	y	n	x	y
1	-1.171	5.841	31	0.095	8.439	61	0.571	8.958	91	-1.109	5.903
2	0.020		32	0.231		62	0.531		92		
3	0.000		33	0.182		63	0.698		93		
4	0.000		34	0.215		64	0.723		94		
5	0.077		35	0.199		65	0.703		95		
6	0.049		36	0.239		66	0.723		96		
7	0.010		37	0.231		67	0.708		97		
8	0.039		38	0.182		68	0.732		98		
9	0.010		39	0.182		69	0.698		99		
10	0.058		40	0.239		70	0.718		100		
11	0.010		41	0.215		71	0.708		101		
12	0.010		42	0.231		72	0.703		102		
13	0.030		43	0.239		73	0.728		103		
14	0.030		44	0.239		74	0.698		104		
15	0.104		45	0.086		75	0.829		105		
16	0.104		46	0.207		76	0.698		106		
17	0.010		47	0.207		77	0.798		107		
18	0.104		48	0.293		78	0.829		108		

n	x	y	n	x	y	n	x	y	n	x	y
19	0.049		49	0.293		79	0.829		109		
20	0.095		50	0.322		80	0.916		110		
21	0.010		51	0.445		81	0.916				
22	0.131		52	0.322		82	0.904				
23	0.095		53	0.419		83	0.916				
24	0.113		54	0.315		84	0.928				
25	0.030		55	0.507		85	1.102				
26	0.049		56	0.438		86	0.900				
27	0.122		57	0.438		87	0.967				
28	0.174		58	0.464		88	0.959				
29	0.182		59	0.531		89	1.001				
30	0.182		60	0.536		90	0.956				

В рассматриваемой таблице данных присутствует $n = 110$ наблюдений для каждой из рассматриваемых переменных.

При условии нормальности данных наши описательные статистики для каждой переменной выглядят следующим образом:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i =$$

Корреляционный анализ числовых данных

Тест гетероскедастичности для ряда данных

Построение линейной модели регрессии

Оценка статистической значимости коэффициентов линейной модели регрессии

Оценка адекватности линейной модели регрессии

Оценка прогнозного интервала для линейной модели регрессии

Темы вопросов на защиту практической работы

1. Задачи корреляционного анализа. Выборочный коэффициент линейной корреляции (Пирсона) и его свойства. Шкала Чеддока.
2. Выборочный коэффициент линейной корреляции (Пирсона) и его свойства. Оценка значимости коэффициента корреляции.
3. Корреляция и причинная связь. Проблемы корреляционного анализа.
4. Ранговая корреляция. Коэффициент ранговой корреляции Спирмена.
5. Задачи регрессионного анализа. Функциональная и статистическая связь. Аппроксимационные модели. Параметрическое множество функций.
6. Линейная регрессия. Определение коэффициентов линейной модели методом наименьших квадратов.
7. Проверка значимости полученных коэффициентов модели. Проверка адекватности модели с помощью критерия Фишера.
8. Доверительный интервал прогноза. Проверка адекватности модели с помощью критерия Фишера.