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

Юрченков Иван Александрович, ассистент кафедры ПМ

2022-10-17

```
##
##           : '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

Для демонстрации проведения регрессионного анализа над рядом данных выбран набор данных цен на алмазы (diamonds), являющийся классическим набором данных для проверки регрессионных моделей и алгоритмов идентификации, очистки или корректировки выбросов. Всего в наборе данных 10 переменных. В рассмотрение возьмем только две из них:

1. carat — караты алмазов,
2. price — цена алмазов.

Предварительный анализ данных рядов показывает их нелинейную зависимость, похожую на параболическую, и чтобы избежать её в линейном регрессионном анализе, принято решение **прологарифмировать** оба ряда данных для спрямления зависимости в декартовых координатах.

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

Table 1: Таблица данных

n	ln(x)	ln(y)	n	ln(x)	ln(y)	n	ln(x)	ln(y)	n	ln(x)	ln(y)
1	-1.171	5.841	**31**	0.095	8.439	**61**	0.571	8.958	**91**	-1.109	5.903
2	0.020	7.965	**32**	0.231	8.446	**62**	0.531	9.048	**92**	-0.892	6.594
3	0.000	8.150	**33**	0.182	8.447	**63**	0.698	9.307	**93**	-0.942	6.111
4	0.000	8.168	**34**	0.215	8.448	**64**	0.723	9.327	**94**	-0.635	6.752
5	0.077	8.171	**35**	0.199	8.450	**65**	0.703	9.334	**95**	-0.673	6.786
6	0.049	8.193	**36**	0.239	8.454	**66**	0.723	9.336	**96**	-0.654	6.829
7	0.010	8.225	**37**	0.231	8.464	**67**	0.708	9.389	**97**	-0.616	6.886
8	0.039	8.223	**38**	0.182	8.465	**68**	0.732	9.407	**98**	-0.635	6.910
9	0.010	8.243	**39**	0.182	8.465	**69**	0.698	9.439	**99**	-0.462	6.971
10	0.058	8.227	**40**	0.239	8.472	**70**	0.718	9.446	**100**	-0.357	7.477
11	0.010	8.290	**41**	0.215	8.473	**71**	0.708	9.451	**101**	-0.357	7.510
12	0.010	8.293	**42**	0.231	8.489	**72**	0.703	9.452	**102**	-0.357	7.513
13	0.030	8.296	**43**	0.239	8.503	**73**	0.728	9.455	**103**	-0.274	7.514
14	0.030	8.307	**44**	0.239	8.521	**74**	0.698	9.458	**104**	-0.342	7.550

n	ln(x)	ln(y)	n	ln(x)	ln(y)	n	ln(x)	ln(y)	n	ln(x)	ln(y)
15	0.104	8.312	**45**	0.086	8.524	**75**	0.829	9.488	**105**	-0.329	7.563
16	0.104	8.317	**46**	0.207	8.534	**76**	0.698	9.492	**106**	-0.288	7.573
17	0.010	8.318	**47**	0.207	8.548	**77**	0.798	9.525	**107**	-0.357	7.624
18	0.104	8.319	**48**	0.293	8.570	**78**	0.829	9.527	**108**	-0.211	7.650
19	0.049	8.325	**49**	0.293	8.586	**79**	0.829	9.582	**109**	0.020	7.867
20	0.095	8.333	**50**	0.322	8.660	**80**	0.916	9.582	**110**	0.000	7.885
21	0.010	8.337	**51**	0.445	8.660	**81**	0.916	9.632			
22	0.131	8.346	**52**	0.322	8.694	**82**	0.904	9.644			
23	0.095	8.349	**53**	0.419	8.714	**83**	0.916	9.680			
24	0.113	8.372	**54**	0.315	8.715	**84**	0.928	9.680			
25	0.030	8.377	**55**	0.507	8.760	**85**	1.102	9.683			
26	0.049	8.381	**56**	0.438	8.825	**86**	0.900	9.709			
27	0.122	8.383	**57**	0.438	8.849	**87**	0.967	9.736			
28	0.174	8.414	**58**	0.464	8.876	**88**	0.959	9.753			
29	0.182	8.414	**59**	0.531	8.918	**89**	1.001	9.787			
30	0.182	8.416	**60**	0.536	8.948	**90**	0.956	9.818			

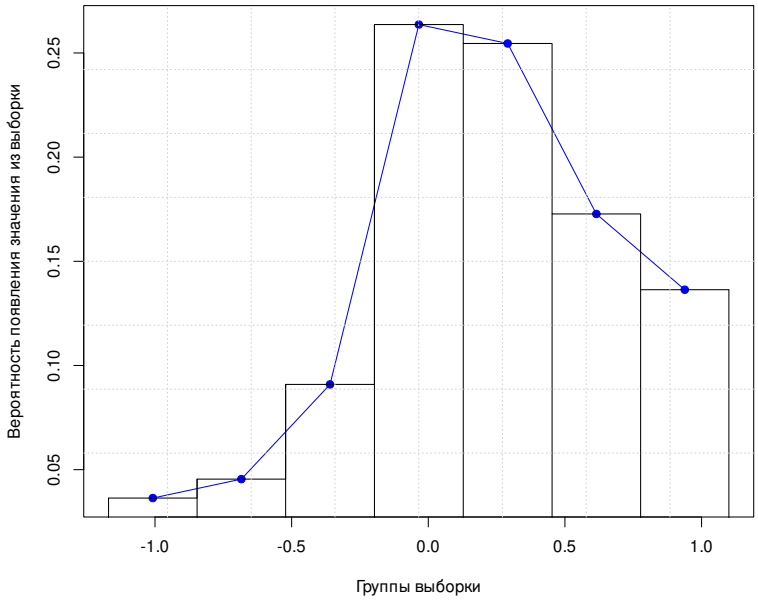
Далее наши логарифмированные данные обозначим как $x := ln(x), \quad y := ln(y)$, и примем данные переменные как рассматриваемые в нашем регрессионном анализе факторные и результирующие соответственно.

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

Построим гистограммы распределений наших данных в каждой из переменных:

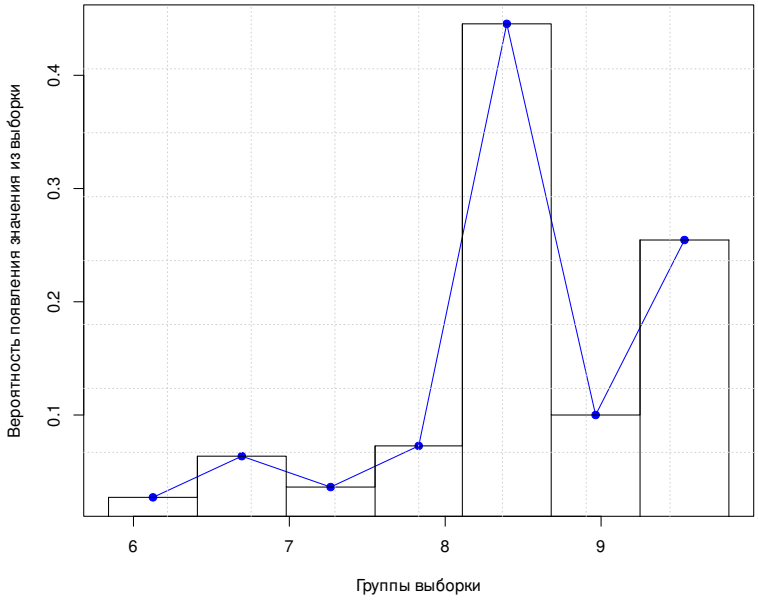
```
##        groupnames abs_freq   rel_freq    low   high     med     h
## 1  (-1.17,-0.846]        4 0.03636364 -1.170 -0.846 -1.0080 0.324
## 2 (-0.846,-0.522]        5 0.04545455 -0.846 -0.522 -0.6840 0.324
## 3 (-0.522,-0.197]       10 0.09090909 -0.522 -0.197 -0.3595 0.325
## 4  (-0.197,0.128]       29 0.26363636 -0.197  0.128 -0.0345 0.325
## 5   (0.128,0.453]       28 0.25454545  0.128  0.453  0.2905 0.325
## 6   (0.453,0.777]       19 0.17272727  0.453  0.777  0.6150 0.324
## 7     (0.777,1.1]       15 0.13636364  0.777  1.100  0.9385 0.323
```

Гистограмма для логарифмов карат




```
##    groupnames abs_freq   rel_freq  low high   med    h
## 1 (5.84,6.41]        3 0.02727273 5.84 6.41 6.125 0.57
## 2 (6.41,6.98]        7 0.06363636 6.41 6.98 6.695 0.57
## 3 (6.98,7.55]        4 0.03636364 6.98 7.55 7.265 0.57
## 4 (7.55,8.11]        8 0.07272727 7.55 8.11 7.830 0.56
## 5 (8.11,8.68]       49 0.44545455 8.11 8.68 8.395 0.57
## 6 (8.68,9.25]       11 0.10000000 8.68 9.25 8.965 0.57
## 7 (9.25,9.82]       28 0.25454545 9.25 9.82 9.535 0.57
```

Гистограмма для логарифмов цены



Для полученных выборок $X = (x_1, x_2, \dots, x_n), \quad Y = (y_1, y_2, \dots, y_n)$ наши описательные статистики рассчитываем следующим образом:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \approx 0.219, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \approx 8.481$$

$$g = 1 + \lfloor log_2(n) \rfloor, \quad h_x = \frac{max(X) - min(X)}{g}, \quad h_y = \frac{max(Y) - min(Y)}{g},$$

$$Z_j = min(X) + j \cdot h_x, \quad L_j = min(Y) + j \cdot h_y, \quad j = 0, 1, \dots, g$$

$$\zeta_i = Z_i - Z_{i-1}, \quad \theta_i = L_i - Li - 1, \quad i = 1, 2, \dots, g,$$

$$\sigma_x = \sqrt{\sum_{i=1}^{g} (\zeta_i - \overline{x})^2 \cdot p_i^x} \approx 0.488, \quad \sigma_y = \sqrt{\sum_{i=1}^{g} (\theta_i - \overline{x})^2 \cdot p_i^y} \approx 0.865$$

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.