СТАТИСТИКА БОЛЬШИХ ДАННЫХ - ИДЗ 1

Студент: Эспинола Ривера, Хольгер Элиас

Тема: Категориальный Анализ Данных

Вариант 9.

Описание эксперимента:

4.5. Table 4.6 refers to a sample of subjects randomly selected for an Italian study on the relation between income and whether one possesses a travel credit card (such as American Express or Diners Club). At each level of annual income in millions of lira, the table indicates the number of subjects sampled and the number of them possessing at least one travel credit card. Analyze these data.

Table 4.6

Income	Number Cases	Credit Cards	Income	Number Cases	Credit Cards
24	1	0	48	1	0
27	1	0	49	1	0
28	5	2	50	10	2
29	3	0	52	1	0
30	9	1	59	1	0
31	5	1	60	5	2
32	8	0	65	6	6
33	1	0	68	3	3
34	7	1	70	5	3
35	1	1	79	1	0
38	3	1	80	1	0
39	2	0	84	1	0
40	5	0	94	1	0
41	2	0	120	6	6
42	2	0	130	1	1
45	1	1			

Source: *Categorical Data Analysis*, Quaderni del Corso Estivo di Statistica e Calcolo delle Probabilità, n. 4., Istituto di Metodi Quantitativi, Università Luigi Bocconi, a cura di R. Piccarreta.

Файл данных: "Credit.csv"

Задания:

1) Разбить уровень доходов на 3 группы и изобразить графически частоты наличия хоть одной карты при различных уровнях дохода (barplot).
2) Проверить зависимость наличия хоть одной карты от уровня дохода (группировка из п (1)) методами классического категориального анализа (Хи-квадрат)
3) Проверить зависимость наличия хоть одной карты от уровня дохода без использования группировки с использованием GLM (модель логистической регрессии). Изобразить графически полученную зависимость вероятности наличия хоть одной карты от уровня дохода.
4) Проверить зависимость наличия хоть одной карты от уровня дохода с группировкой из п. (1) с использованием GLM. Сделать это двумя способами: с использованием модели логистической регрессии и с использованием пуассоновской модели.
5) Интерпретировать результаты анализа.

01. Разбить уровень доходов на 3 группы и изобразить графически частоты наличия хоть одной карты при различных уровнях дохода (barplot).

Шаг 01:

Мы делим уровень дохода на эти 3 группы:

- низкий уровень: менее 40 миллионов лир

- средний уровень: от 40 до 70 миллионов лир

- высокий уровень: более 70 миллионов лир

Шаг 02:

Мы определяем таблица частота:

		уровень дохода		
	есть карта	0	1	2
нет	0	39	24	6
Да	1	7	14	10

Шаг 03:

Мы определяем частотный график, который содержит всех показанных людей и людей, у которых есть хотя бы одна карта, разделенных на уровни дохода.

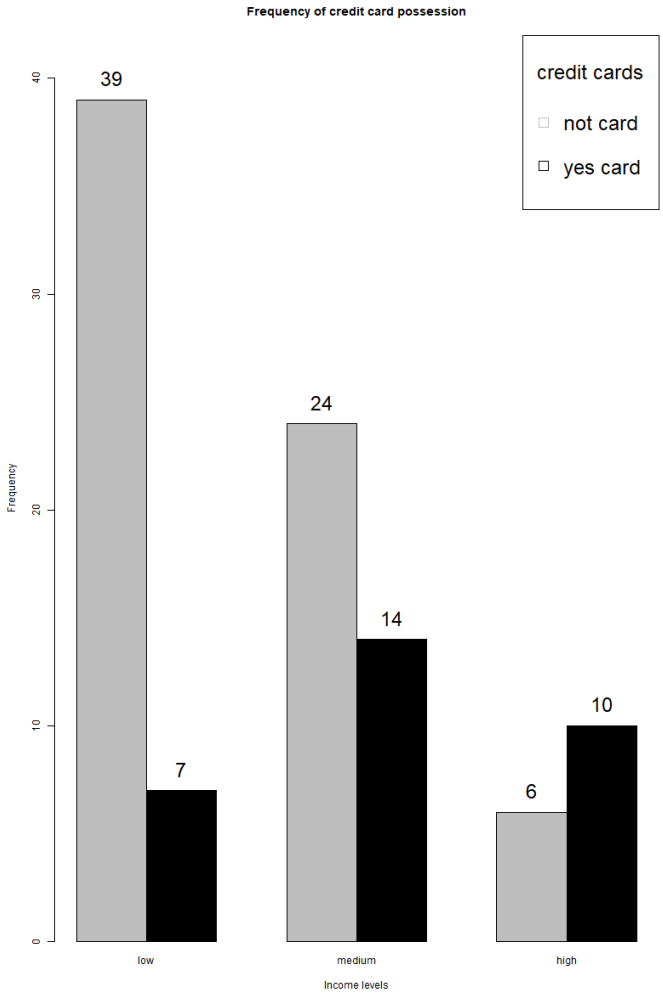


рисунок 1. частота наличия хоть одной карты при различных уровнях дохода

02. Проверить зависимость наличия хоть одной карты от уровня дохода (группировка из п (1)) методами классического категориального анализа (Хи-квадрат)

Шаг 01: Мы определяем таблица частота:

	низкий	середина	высокий
нет карта	39	24	6
есть карта	7	14	10

Всего людей в выборке: 100

Шаг 02: Определить нулевую и альтернативную гипотезы

H_0: наличия хотя бы одной карты независима от переменная уровня дохода

H_a: переменные наличия хотя бы одной картой и уровень дохода имеют уровень зависимости

Шаг 03: проверка гипотезы Xi - квадрат

Pearson's Chi-squared test

data: tablitza

X-squared = 13.385, df = 2, p-value = 0.00124

[1] "p-value = 0.00124000270145794"

[1] "1 ===> Reject H0, because 0.00124000270145794 <= 0.05"

[1] "variables contribute significantly in variability of data"

Шаг 04: Заключение

Нулевая гипотеза H0 отвергается,

Следовательно, переменная уровня дохода вносит значительный вклад в дисперсию переменной, которая измеряет владение хотя бы одной картой.

03. Проверить зависимость наличия хоть одной карты от уровня дохода без использования группировки с использованием GLM (модель логистической регрессии). Изобразить графически полученную зависимость вероятности наличия хоть одной карты от уровня дохода.

Шаг 01: Реконфигурация базы данных, содержащей все категориальные экземпляры (без использования группировки).

	id	income	yescard
1	1	24	0
2	2	27	0
3	3	28	1
4	4	28	1
5	5	28	0
6	6	28	0
7	7	28	0
8	8	29	0
9	9	29	0
10	10	29	0
11	11	30	1
12	12	30	0
13	13	30	0
14	14	30	0
15	15	30	0
16	16	30	0
17	17	30	0
18	18	30	0
19	19	30	0
20	20	31	1
21	21	31	0
22	22	31	0
23	23	31	0
24	24	31	0
25	25	32	0
26	26	32	0
27	27	32	0
28	28	32	0
29	29	32	0
30	30	32	0
31	31	32	0
32	32	32	0
33	33	33	0
34	34	34	1
35	35	34	0
36	36	34	0
37	37	34	0
38	38	34	0
39	39	34	0
40	40	34	0
41	41	35	1
42	42	38	1
43	43	38	0
44	44	38	0
45	45	39	0
46	46	39	0
47	47	40	0
48	48	40	0
49	49	40	0
50	50	40	0
51	51	40	0
52	52	41	0
53	53	41	0
54	54	42	0
55	55	42	0
56	56	45	1
57	57	48	0
58	58	49	0
59	59	50	1
60	60	50	1
61	61	50	0
62	62	50	0
63	63	50	0
64	64	50	0
65	65	50	0
66	66	50	0
67	67	50	0
68	68	50	0
69	69	52	0
70	70	59	0
70	70	59	0
71	71	60	1
72	72	60	1
73	73	60	0
74	74	60	0
75	75	60	0
76	76	65	1
77	77	65	1
78	78	65	1
79	79	65	1
80	80	65	1
81	81	65	1
82	82	68	1
83	83	68	1
84	84	68	1
85	85	70	1
86	86	70	1
87	87	70	1
88	88	70	0
89	89	70	0
90	90	79	0
91	91	80	0
92	92	84	0
93	93	94	0
94	94	120	1
95	95	120	1
96	96	120	1
97	97	120	1
98	98	120	1
99	99	120	1
100	100	130	1

Шаг 02: обобщенная линейная модель

$\log it(y) = b_0 + b_1 \cdot x$, где:

x: уровень дохода

y: наличие хотя бы одной карты

Получена модель логистической регрессии:

> lr_model

Call: glm(formula = yescard ~ income, family = "binomial", data = categ_table)

Coefficients:

(Intercept) income

-3.55611 0.05318

Degrees of Freedom: 99 Total (i.e. Null); 98 Residual

Null Deviance: 123.8

Residual Deviance: 96.96 AIC: 101

Шаг 03: Мы проводим проверку гипотез с использованием GLM (модель логистической регрессии).

$H_0 : b_1 = 0$

$H_a : b_1 \neq 0$

критерий отношения правдоподобия:

$G = 2(LL_s - LL_a)$

LL_S: максимум логарифма правдоподобия в общей моделей

LL_a: максимум логарифма правдоподобия в общей в адитивной моделей

предельное распределение: χ^2_{d-1}

P – значение: $pv = 1 - K_{d-1}(G)$

$$\phi(x) = \begin{cases} 0 \ ; \ p - value > \alpha \\ 1 \ ; \ p - value < \alpha \end{cases}$$

если ϕ(x) = 0 ➔ принимает H_0

если ϕ(x) = 1 ➔ отвергает H_0

Шаг 04: получить результаты

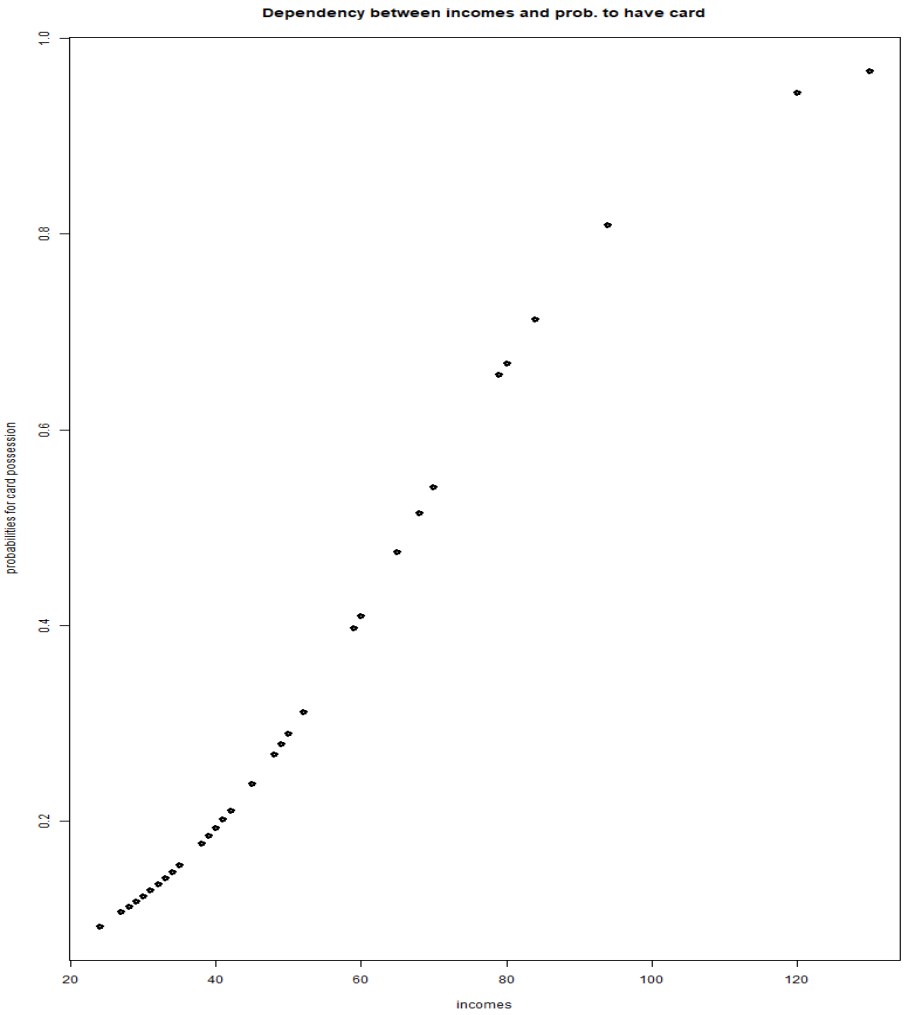
```
> # compute the pvalue
> llr_pvalue <- pchisq(q = llr, df = 1, lower.tail = FALSE)
> print(paste("p-value = ", llr_pvalue))
[1] "p-value =  2.19088847218211e-07"

> # inference with relative maximum likelihood and Chi-square
> hypothesis_proof(phi_llr, llr_pvalue)
[1] "1  ===> Reject H0, because  2.19088847218211e-07  <=  0.05"
[1] "variables contribute significantly in variability of data"

> # inference with anova
> glm_anova <- anova(lr_ind, lr_model, test = "LRT")
> print(glm_anova)

Analysis of Deviance Table

Model 1: yescard ~ 1
Model 2: yescard ~ income
  Resid.      Df      Resid. Dev    Df      Deviance      Pr(>Chi)
1             99    123.820
2             98     96.963         1       26.857        2.191e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Шаг 05: Заключение

Нулевая гипотеза H0 отвергается.

Переменная уровня дохода (без использования группировки) вносит значительный вклад в дисперсию данных при наличии хотя бы 1 карты для модели логистической регрессии.

Шаг 06: Построить зависимость вероятности получения хотя бы одной карты от уровня дохода (без группировки)



04. Проверить зависимость наличия хоть одной карты от уровня дохода с группировкой из п. (1) с использованием GLM. Сделать это двумя способами: с использованием модели логистической регрессии и с использованием пуассоновской модели.

МОДЕЛЬ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ

Шаг 01: Мы создаем соответствующую структуру данных для применения модели логистической регрессии с учетом 3 групп по уровням дохода.

> group_table <- group_glm(freq_table)

> print(group_table)

	id	possession	levels
1	1	0	0
2	2	0	0
3	3	0	0
4	4	0	0
5	5	0	0
6	6	0	0
7	7	0	0
8	8	0	0
9	9	0	0
10	10	0	0
11	11	0	0
12	12	0	0
13	13	0	0
14	14	0	0
15	15	0	0
16	16	0	0
17	17	0	0
18	18	0	0
19	19	0	0
20	20	0	0
21	21	0	0
22	22	0	0
23	23	0	0
24	24	0	0
25	25	0	0
26	26	0	0
27	27	0	0
28	28	0	0
29	29	0	0
30	30	0	0
31	31	0	0
32	32	0	0
33	33	0	0
34	34	0	0
35	35	0	0
35	35	0	0
36	36	0	0
37	37	0	0
38	38	0	0
39	39	0	0
40	40	0	1
41	41	0	1
42	42	0	1
43	43	0	1
44	44	0	1
45	45	0	1
46	46	0	1
47	47	0	1
48	48	0	1
49	49	0	1
50	50	0	1
51	51	0	1
52	52	0	1
53	53	0	1
54	54	0	1
55	55	0	1
56	56	0	1
57	57	0	1
58	58	0	1
59	59	0	1
60	60	0	1
61	61	0	1
62	62	0	1
63	63	0	1
64	64	0	2
65	65	0	2
66	66	0	2
67	67	0	2
68	68	0	2
69	69	0	2
70	70	1	0
70	70	1	0
71	71	1	0
72	72	1	0
73	73	1	0
74	74	1	0
75	75	1	0
76	76	1	0
77	77	1	1
78	78	1	1
79	79	1	1
80	80	1	1
81	81	1	1
82	82	1	1
83	83	1	1
84	84	1	1
85	85	1	1
86	86	1	1
87	87	1	1
88	88	1	1
89	89	1	1
90	90	1	1
91	91	1	2
92	92	1	2
93	93	1	2
94	94	1	2
95	95	1	2
96	96	1	2
97	97	1	2
98	98	1	2
99	99	1	2
100	100	1	2

Шаг 02: Определите модель логистической регрессии, связывающую переменные владения картами с уровнями доходов.

```
> lr_group <- glm(possession ~ levels, data = group_table, family = binomial)
> lr_group

Call:  glm(formula = possession ~ levels, family = binomial, data = group_table)

Coefficients:
(Intercept)       levels
     -1.693        1.122

Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
Null Deviance:      123.8
Residual Deviance: 110.4        AIC: 114.4
```

Шаг 03: Определите модель независимо от переменных

```
> lr_gind <- glm(possession ~ 1, data = group_table, family = binomial)
> lr_gind

Call:  glm(formula = possession ~ 1, family = binomial, data = group_table)

Coefficients:
(Intercept)
    -0.8001

Degrees of Freedom: 99 Total (i.e. Null);  99 Residual
Null Deviance:      123.8
Residual Deviance: 123.8        AIC: 125.8
```

Шаг 04: Мы проводим проверку гипотез с использованием GLM (модель логистической регрессии).

H_0: линейно независимое сравнение ($H_0 : b_1 = b_2 = 0$)

H_a: линейно зависимое сравнение ($H_a : b_1 \neq b_2 \neq 0$)

критерий отношения правдоподобия:

$G = 2(LL_s - LL_a)$

LL_s: максимум логарифма правдоподобия в общей моделей

LL_a: максимум логарифма правдоподобия в общей в адитивной моделей

предельное распределение: χ^2_{d-1} , где:

possession = {0, 1}; поэтому: d = 2

P – значение: $pv = 1 - K_{d_1-1}(G)$

$$\phi(x) = \begin{cases} 0 \ ; \ p-value > \alpha \\ 1 \ ; \ p-value < \alpha \end{cases}$$

если ϕ(x) = 0 ➔ принимает H_0

если ϕ(x) = 1 ➔ отвергает H_0

Шаг 05: получить результаты

- вычисление p-значения

```
> llr2_pvalue <- pchisq(llr2, 1, lower.tail = FALSE)
> print(paste("logistic reg. grouping pvalue = ", llr2_pvalue))
[1] "logistic reg. grouping pvalue =  0.000254237967368209"
```

- Дисперсионный анализ ANOVA

```
> # analysis anova
> llr2_anova <- anova(lr_gind, lr_group, test = "LRT")
> print(llr2_anova)
Analysis of Deviance Table
```

Model 1: possession ~ 1

Model 2: possession ~ levels

Resid.	Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	99	123.82			
2	98	110.44	1	13.381	0.0002542 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- проверка гипотезы

> # make inference

> hypothesis_proof(phi_llrgroup, llr2_pvalue)

[1] "1 ===> Reject H0, because 0.000254237967368209 <= 0.05"

[1] "variables contribute significantly in variability of data"

Шаг 06: Заключение

Нулевая гипотеза H0 отвергается.

Переменная уровня дохода (с группами) вносит значительный вклад в дисперсию данных о наличии хотя бы 1 карты для модель логистической регрессии.

МОДЕЛЬ ПУАССОНА

Шаг 01: Чтобы использовать модель Пуассона для случая, когда данные сгруппированы по трем уровням дохода, необходимо использовать следующую структуру данных:

```
> freq_table <- poiss_table(tablitza)
> print(freq_table)
      id possession levels frequency
1      1          0      0        39
2      2          0      1        24
3      3          0      2         6
4      4          1      0         7
5      5          1      1        14
6      6          1      2        10
```

Шаг 02: Мы проверили размерность уровней для каждой из категориальных переменных.

```
> nl1 <- length(levels(as.factor(freq_table$possession)))
```

possession = {0, 1}

```
> nl2 <- length(levels(as.factor(freq_table$levels)))
```

Levels = {0, 1, 2}

Степени свободы определяются:

df = (nl1 - 1) * (nl2 - 1)

Шаг 03: Мы определяем модель Пуассона для взаимодействия переменных уровня владения и дохода (сгруппированных) по частотам

```
> poiss_model <- glm(frequency ~ as.factor(possession) * as.factor(levels),
+                    data = freq_table, family = "poisson")
> poiss_model
```

```
Call:  glm(formula = frequency ~ as.factor(possession) * as.factor(levels),
    family = "poisson", data = freq_table)

Coefficients:
                          (Intercept)
                               3.6636
                    as.factor(possession)1
                              -1.7177
                       as.factor(levels)1
                              -0.4855
                       as.factor(levels)2
                              -1.8718
as.factor(possession)1:as.factor(levels)1
                               1.1787
as.factor(possession)1:as.factor(levels)2
                               2.2285

Degrees of Freedom: 5 Total (i.e. Null);  0 Residual
Null Deviance:      44.31
Residual Deviance: 2.22e-15     AIC: 38.64
```

Шаг 04: Определим аддитивную модель

```
> poiss_ind <- glm(frequency ~ as.factor(possession) + as.factor(levels),
+                  data = freq_table, family = "poisson")
> poiss_ind

Call:  glm(formula = frequency ~ as.factor(possession) + as.factor(levels),
    family = "poisson", data = freq_table)

Coefficients:
```

(Intercept)	as.factor(possession)1	as.factor(levels)1
3.4576	-0.8001	-0.1911

as.factor(levels)2

-1.0561

Degrees of Freedom: 5 Total (i.e. Null); 2 Residual

Null Deviance: 44.31

Residual Deviance: 13.4 AIC: 48.04

Шаг 05: Мы проводим проверку гипотез с использованием GLM (модель Пуассона).

H_0: линейно независимое сравнение ($H_0 : b_1 = b_2 = 0$)

H_a: линейно зависимое сравнение ($H_a : b_1 \neq b_2 \neq 0$)

критерий отношения правдоподобия:

$$G = 2(LL_s - LL_a)$$

LL_s: максимум логарифма правдоподобия в общей моделей

LL_a: максимум логарифма правдоподобия в общей в адитивной моделей

предельное распределение: $\chi^2_{(d_1-1)\cdot(d_2-1)}$

P – значение: $pv = 1 - K_{(d_1-1)(d_2-1)}(G)$

$$\phi(x) = \begin{cases} 0 \ ; \ p-value > \alpha \\ 1 \ ; \ p-value < \alpha \end{cases}$$

если ϕ(x) = 0 ➔ принимает H_0

если ϕ(x) = 1 ➔ отвергает H_0

Шаг 06: получить результаты

- вычисление p-значения

> # compute p-value

```
> llrpoiss_pvalue <- pchisq(q = llr_poiss, df = (nl1 - 1) * (nl2 - 1),
+                           lower.tail = FALSE)
> print(paste("Poisson p-value = ", llrpoiss_pvalue))
[1] "Poisson p-value =  0.00123134758798991"
```

- Дисперсионный анализ ANOVA

```
> poiss_anova <- anova(poiss_ind, poiss_model, test = "LRT")
> print(poiss_anova)

Analysis of Deviance Table

Model 1: frequency ~ as.factor(possession) + as.factor(levels)
Model 2: frequency ~ as.factor(possession) * as.factor(levels)
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1         2     13.399
2         0      0.000  2   13.399  0.001231 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- проверка гипотезы

```
> phi_poisson <- as.numeric(llr_pvalue < alpha)
> hypothesis_proof(phi_poisson, llrpoiss_pvalue)
[1] "1  ===> Reject H0, because  0.00123134758798991  <=  0.05"
[1] "variables contribute significantly in variability of data"
```

Шаг 07: Заключение

Нулевая гипотеза H0 отвергается.

Переменная уровня дохода (с группами) вносит значительный вклад в дисперсию данных о наличии хотя бы 1 карты для модели Пуассона.