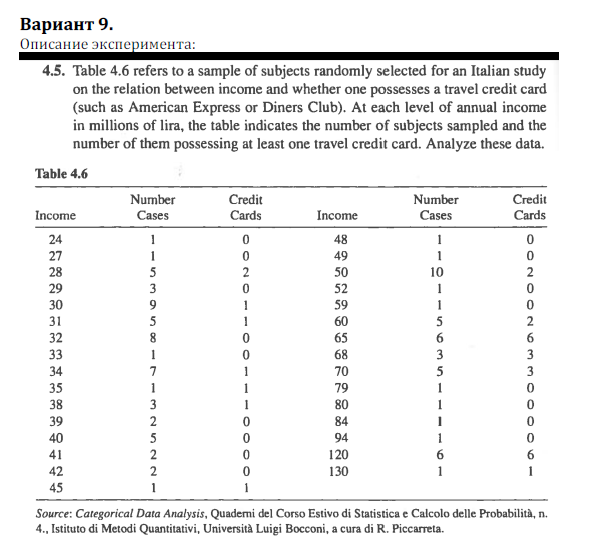
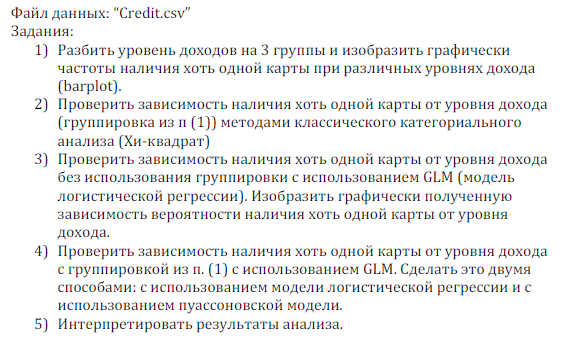
**СТАТИСТИКА БОЛЬШИХ ДАННЫХ - ИДЗ 1**

Студент: Эспинола Ривера, Хольгер Элиас

Тема: Категориальный Анализ Данных





**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.

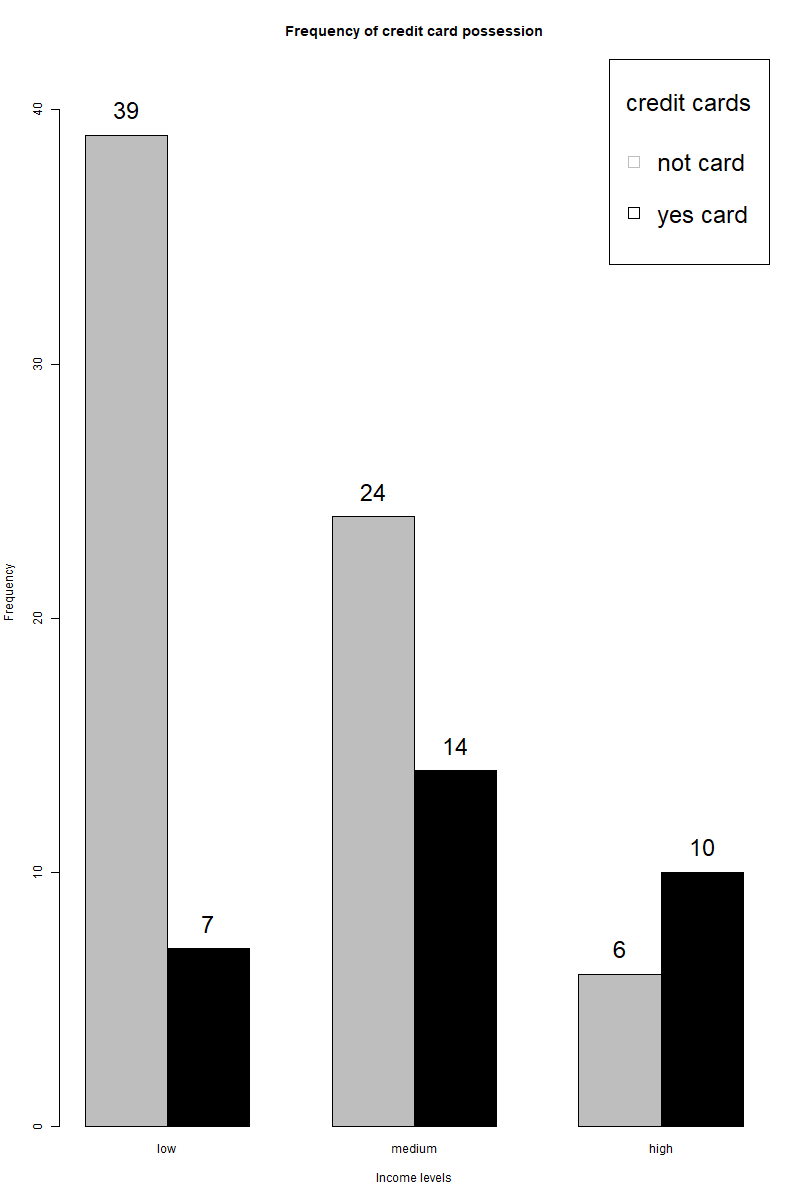


рисунок 1. частота наличия хоть одной карты при различных уровнях дохода

**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

H0: наличия хотя бы одной карты независима от переменная уровня дохода

Ha: переменные наличия хотя бы одной картой и уровень дохода зависимы

Шаг 03: проверка гипотезы Xi - квадрат

Pearson's Chi-squared test

-data: tablitza

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

Статистическая критерия:



[1] "p-value = 0.00124000270145794"

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

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

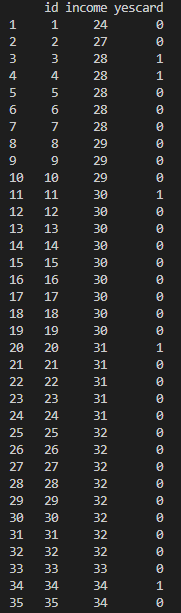
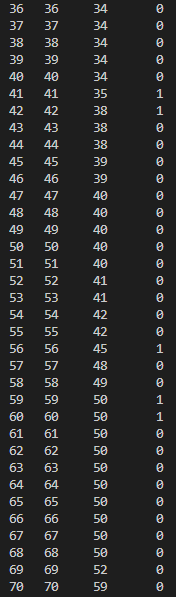
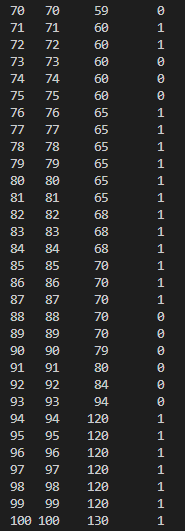
Шаг 04: Заключение

Нулевая гипотеза **H0 отвергается**,

Следовательно, переменная уровня дохода вносит значимый 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).



Шаг 02: обобщенная линейная модель

, 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).





критерий отношения правдоподобия:



LLs: максимум логарифма правдоподобия в общей моделей

LLa: максимум логарифма правдоподобия в общей в адитивной моделей

предельное распределение: 

P – значение: 

Функция для проверки гипотеза:



если ϕ(x) = 0 🡺 принимает H0

если ϕ(x) = 1 🡺 отвергает H0

Шаг 04: Получить результаты

# statistic criteria logarithm relative max-likelihood

> llr <- lr\_ind$deviance - lr\_model$deviance

> print(paste("G = ", llr))

[1] "G = 26.8569165934696"

> # 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)



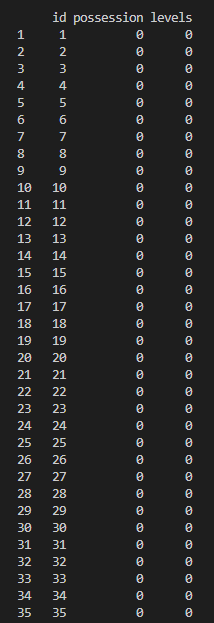
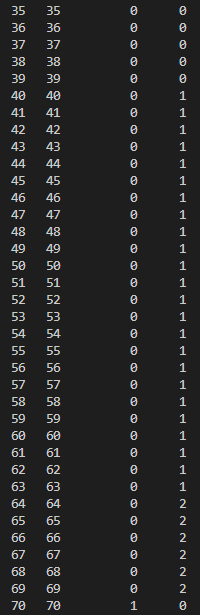
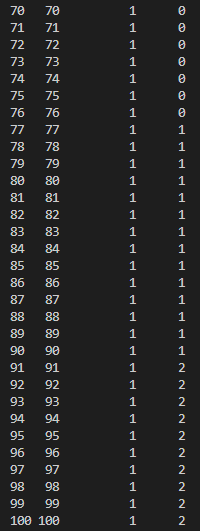
**04. Проверить зависимость наличия хоть одной карты от уровня дохода с группировкой из п. (1) с использованием GLM. Сделать это двумя способами: с использованием модели логистической регрессии и с использованием пуассоновской модели.**

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)



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

> lr\_group

Call: glm(formula = possession ~ as.factor(levels), family = binomial,

data = group\_table)

Coefficients:

(Intercept) as.factor(levels)1 as.factor(levels)2

-1.718 1.179 2.228

Degrees of Freedom: 99 Total (i.e. Null); 97 Residual

Null Deviance: 123.8

Residual Deviance: 110.4 AIC: 116.4

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).

H0: линейно независимое сравнение ()

Ha: линейно зависимое сравнение ()

критерий отношения правдоподобия:



LLs: максимум логарифма правдоподобия в общей моделей

LLa: максимум логарифма правдоподобия в общей в адитивной моделей

предельное распределение:  , где:

levels = {0, 1, 2}; поэтому: d = 3

P – значение: 

Функция для проверки гипотеза:



если ϕ(x) = 0 🡺 принимает H0

если ϕ(x) = 1 🡺 отвергает H0

Шаг 05: получить результаты

* Статистическая критерия отношения правдоподобия

> # statistic criteria logaritm relative max-likelihood

> llr2 <- lr\_gind$deviance - lr\_group$deviance

> print(paste("G-logistic = ", llr2))

[1] "G-logistic = 13.3992922186754"

* вычисление p-значения

> # calculate p-value

> llr2\_pvalue <- pchisq(q = llr2, df = nl3 - 1, lower.tail = FALSE)

> print(paste("logistic reg. grouping pvalue = ", llr2\_pvalue))

[1] "logistic reg. grouping pvalue = 0.00123134758798991"

* 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 ~ as.factor(levels)

Resid. Df Resid. Dev Df Deviance Pr(>Chi)

1 99 123.82

2 97 110.42 2 13.399 0.001231 \*\*

---

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

* проверка гипотезы

> # define criteria function

> phi\_llrgroup <- as.numeric(llr2\_pvalue < alpha)

>

> # make inference

> hypothesis\_proof(phi\_llrgroup, llr2\_pvalue)

[1] "1 ===> Reject H0, because 0.00123134758798991 <= 0.05"

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

Шаг 06: Заключение

Нулевая **гипотеза H0 отвергается**.

Переменная уровня дохода (с группами) вносит значимый 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).

H0: линейно независимое сравнение ()

Ha: линейно зависимое сравнение ()

критерий отношения правдоподобия:



LLs: максимум логарифма правдоподобия в общей моделей

LLa: максимум логарифма правдоподобия в общей в адитивной моделей

предельное распределение: 

P – значение: 

Функция дла проверки гипотеза:



если ϕ(x) = 0 🡺 принимает H0

если ϕ(x) = 1 🡺 отвергает H0

Шаг 06: получить результаты

* Статистическая критерия отношения правдоподобия

> # statistic criteria logarithm max-likelihood

> llr\_poiss <- poiss\_ind$deviance - poiss\_model$deviance

> print(paste("G = ", llr\_poiss))

[1] "G = 13.3992922186754"

* 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 значимый вклад в дисперсию данных о наличии хотя бы 1 карты для модели Пуассона.