Семён Алексеевич Илясов , 4 вариант, 20930гр.

Расчетное задание по математической статистике. Часть 3.

3.5. Задание

По заданному набору данных двух переменных (приложение 2):

3.1. Проверить гипотезу о независимости переменных по критерию Хи-квадрат (2 балла)

3.2. Вычислить оценку ковариации, коэффициента корреляции (2 балла). Проверить гипотезу о незначимости коэффициента корреляции (2 балла).

3.3. Оценить параметры линейной регрессии (1 балл), вычислить коэффициент детерминации (1 балл), проверить значимость модели по критерию Фишера (2 балла).

Замечание: программе необходимы файлы iris.data , alpha_table.csv, student_table.csv , приложенные в письме. В этих файлах содержатся таблицы с данными, необходимыми для решения.
Их необходимо поместить в ту же директорию, из которой запускается сама программа.

Данные из таблицы вариантов для 4 варианта:

- `X = "petal_length"`
- `Y = "sepal_width"`
- `TYPE = "Iris-setosa"`

3.1

В этом пункте необходимо проверить гипотезу независимости двух признаков друг от друга.

Алгоритм:

1) Вытащить данные из таблицы iris.data, отсортировать их по возрастанию значения в столбце X.
2) Преобразовать табличные данные в таблицу сопряженности. (функция create_partition)
3) Вычислить суммы в столбцах и в строках получившейся таблицы сопряжённости:

Таблица сопряженности X и Y имеет вид:

X	Y			Всего
	B_1	$\dots$	B_k	
C_1	ν_{11}	$\dots$	ν_{1k}	$\nu_{1\bullet}$
$\dots$	$\dots$		$\dots$	$\dots$
C_s	ν_{s1}	$\dots$	ν_{sk}	$\nu_{s\bullet}$
Всего	$\nu_{\bullet 1}$	$\dots$	$\nu_{\bullet k}$	n

где $\nu_{\bullet k}$ - сумма в k столбце, $\nu_{s\bullet}$ - сумма в s строке, n - сумма всех Vij элементов таблицы.

4) Вычислить статистику критерия Пирсона:

$$\chi_n^2 = n\sum_{i=1}^{s}\sum_{j=1}^{k}\frac{\left(\nu_{ij} - \nu_{i\bullet}\nu_{\bullet j} / n\right)^2}{\nu_{i\bullet}\nu_{\bullet j}} = n\left(\sum_{i=1}^{s}\sum_{j=1}^{k}\frac{\nu_{ij}^{\ 2}}{\nu_{i\bullet}\nu_{\bullet j}} - 1\right)$$

5) Получить критическое значение статистики Пирсона по таблице alpha_table.csv и заданному параметру alpha. (alpha по умолчанию 0.01)

6) Сравнить полученное значение статистики критерия Пирсона с критическим значением, если значение больше критического, гипотеза отвергается.

$$\chi_n^2 > F^{-1}_{\chi^2_{(s-1)(k-1)}}(1-\alpha),$$

Результат работы программы:
Статистика критерия Пирсона: 23.061324197687828
Критическое значение для alpha == 0.05: 26.3
Значение статистики меньше критического, значит величины независимы, гипотеза принимается.

```
result is 23.061324197687828
function is 26.3
the values are independent
```

3.2

В этом пункте нужно вычислить оценку ковариации, коэффициент корреляции. Проверить гипотезу о незначимости коэффициента корреляции.

Алгоритм:

1) вычислить выборочные средние и среднее квадратичное для X и Y:

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i)^2,\ \overline{y^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i)^2,$$

$\overline{y}, \overline{x}$ - выборочные средние.

2) вычислить выборочное стандартное отклонение:

$$s_x^2 = \overline{x^2} - (\overline{x})^2,\ s_y^2 = \overline{y^2} - (\overline{y})^2$$

$s_x = \sqrt{s_x^2}$ и s_y - выборочное стандартное отклонение,

3) вычислить оценку ковариации:

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

4) вычислить выборочный коэффициент линейной корреляции:

$$r_{xy} = \frac{s_{xy}}{s_x s_y},$$

Результат:

оценка ковариации: 0.011448000000000014

коэффициент корреляции: 0.06143539453137004

```
correlation  0.06143539453137004
covariance  0.011448000000000014
```

Проверка гипотезы о незначимости коэффициента корреляции.

Алгоритм:

1)

Распределение статистики

$$T = \frac{r}{\sqrt{1 - r^2}} \sqrt{n - 2}$$

r - коэффициент корреляции

n - количество элементов выборки

2) вычислить t = $t_{cr}(\alpha, n - 2)$. по таблице Стьюдента.

Если T > t, то гипотеза отвергается.

Результат:

T: 0.4264424220961077

t: 1.96

T < t -> гипотеза принимается

```
T ==  0.4264424220961077
t ==  1.96
the hypothesis of the insignificance
 of the linear correlation coefficient is ACCEPTED
```

3.3

В этом пункте нужно оценить коэффициенты линейной регрессии, вычислить коэффициент детерминации, проверить значимость модели по критерию Фишера.

Алгоритм:

1) Оценим коэффициенты линейной регрессии:

$$\hat{\beta}_0 = \boxed{\overline{y} - \beta_1 \overline{x}}, \ \hat{\beta}_1 = \boxed{\frac{\sum_i y_i x_i - n\,\overline{y}\,\overline{x}}{\sum_i (x_i)^2 - n(\overline{x})^2}}$$

$\overline{y}, \overline{x}$ - выборочные средние.

$\widehat{\boldsymbol{\beta}1}$ == $\boldsymbol{\beta}$1

2) Найдём остаточную вариацию, стандартную ошибку, общую вариацию и вариацию, объяснённую регрессией:

Остаточная вариация (residual sum of squares)

$$RSS = \sum_{i=1}^{n} (e_i)^2 ;$$

Стандартная ошибка (несмещенная оценка дисперсии ошибки):

$$s^2 = RSS \,/\, (n - m - 1),$$

где $m = 1$ - число независимых переменных.

Общая вариация

$$TSS = \sum_{i=1}^{n} (y_i - \overline{y})^2 ;$$

Вариация, объясненная регрессией

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 .$$

3) Найдём коэффициент детерминации:

$$R^2 = 1 - RSS/TSS = ESS/TSS; \quad R^2 \in [0,1]$$

4) Рассмотрим гипотезу о значимости регрессии:

Статистика F-критерия:

$$F = \frac{R^2}{1 - R^2}\frac{n - m - 1}{m},$$

где R^2 - коэффициент детерминации, $m = 1$ - число независимых переменных.

H_0 отвергается на уровне значимости α, если

$$F > F(\alpha; m, n - m - 1),$$

где $F(\alpha; m, n - m - 1)$ определяется из таблицы F-распределения.

Результат:

$\widehat{\boldsymbol{\beta}1}$ == $\boldsymbol{\beta}$1 == 2.3352464949710456

$\boldsymbol{\beta}$0 == -0.0008008686376106411

R^2 == 0.9351015413611502

F == 691.616949411893

F(alpha, m, n - m - 1) == 4.042652128566653

```
---------------------------------------------------------------------------
Beta0 =  -0.0008008686376106411 , Beta1 =  2.3352464949710456
R^2 ==  0.9351015413611502
F ==  691.616949411893
Fisher's criterion ==  4.042652128566653
The hypothesis is REJECTED at the significance level alpha =  95.0 %
---------------------------------------------------------------------------
```