**Задание. Линейная регрессия – зарплата и температура**

Для выполнения задания были загружены данные из файла “temp\_wage.tsv” и пропущены первые три строки поскольку они не содержат данных.

setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

df <- read.delim("temp\_wage.tsv", skip = 3)

Далее был построен график зарплат от температуры.

plot(y = df$wage, x = df$temp, xlab = "Temp", ylab = "Wage", type = "p",

main = "График зарплаты от температуры", pch = 19, col = "#000000")

Chart, scatter chart

Description automatically generated

Рис. 1. График зарплаты от температуры

После был найден коэффициент корреляции и его значение равно -0,509.

cor\_xy <- cor(df$temp, df$wage)

Далее было подобрано две линейных модели и посчитаны все основные статистические характеристики.

cor\_xy <- cor(df$temp, df$wage) # считаем корреляцию температуры и зарплаты

reg\_xy <- lm(df$wage ~ df$temp) # строим регрессию

reg\_yx <- lm(df$temp ~ df$wage) # снова строим регрессию

summary\_ <- summary(reg\_xy)

Проверяем факт что доля дисперсии в переменной отклика регрессионной модели, которая может быть объяснена предсказанными переменными равен квадрату корреляции.

is\_true <- round(summary\_$r.squared, digits = 7) ==

           round(cor\_xy \* cor\_xy, digits = 7)

В результате они равны.

Добавили на график линию регрессии к графику зарплат от температуры.

abline(reg\_xy, col = "red")

Chart, scatter chart

Description automatically generated

Рис. 2. График зарплаты от температуры с линией регрессии

После было необходимо построить диаграмму фактической зарплаты от расчётных значений зарплаты и добавить линию с наклоном 1 из начала координат.

plot(x = df$wage, y = reg\_xy$fitted.values,

xlab = "Wage", ylab = "Regression", type = "p",

main = "Фактическая зарплата от расчётных значений зарплаты",

pch = 19, col = "#000000")

abline(c(0, 1), col = "red")

Chart, scatter chart

Description automatically generated

Рис.3. График фактической зарплаты от расчётных значений зарплаты с линией с наклоном 1

Поскольку мы построили регрессию температуры от зарплаты раньше и также рассчитали все основные статистические характеристики, то нам нет необходимости это делать снова. В коде сделали небольшие переобозначения. Сравнили доли дисперсии в переменных отклика регрессионной модели, которая могут быть объяснены предсказанными переменными.

summary\_1 <- summary(reg\_xy) # коэф r^2 будет сейм для них

summary\_2 <- summary(reg\_yx)

is\_true <- round(summary\_1$r.squared, digits = 7) ==

           round(summary\_2$r.squared, digits = 7)

В результате они равны.

Проверяем что коэффициенты другие перевернув уравнение обратной регрессии и сравнив с исходной

a1 <- summary\_1$coefficients[1]

b1 <- summary\_1$coefficients[2]

b2 <- summary\_2$coefficients[2]

a2 <- -summary\_2$coefficients[1] / b2

b2 <- 1 / b2

Добавляем на точечную диаграмму зарплаты от температуры две линии регрессии и уровни среднего значения температуры и зарплаты.

Chart, scatter chart

Description automatically generated

Рис. 4. График зарплаты от температуры с регрессиями и средними уровнями