ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ

ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«БАШКИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Отчет по лабораторной работе №5

«Меры связи. Регрессионный анализ»

Выполнил: студент гр. 1ФКЦ

Ахметов Руслан Олегович

Проверил: кандидат технических наук, доцент

Прудников Вадим Борисович

Уфа – 2023

1. Оценим с помощью R линейные коэффициенты корреляции Пирсона между признаками «wage» и «age», «wage» и «hours», а также проведем характер и силу связи. (рис.1)

Var3$wage <- as.double(var3$wage)

Var3$age <- as.double(var3$age)

double – двойной

Здесь переменные были преобразованы к типу «double» для расчета корреляционных коэффициентов.



Рисунок 1.

Корреляционные коэффициенты были рассчитаны равными -0.05064909 и 0.1325229 соответственно (рис.2):

r1 <- cor(var3$wage, var3$age, method = 'pearson'); r1

r2 <- cor(var3$wage, var3$hours, method = 'pearson'); r2

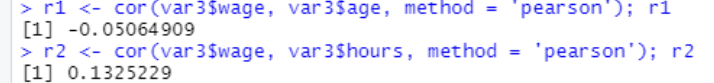


Рисунок 2.

method - символьная строка, указывающая, какой тип теста был выполнен.

Таким образом, заработная плата в среднем с увеличением возраста падает и, наоборот, растет при увеличении продолжительности рабочей недели.

Тестирование коэффициентов корреляции на значимость (рис.3):

cor.test(var3$wage, var3$age, alternative = 'two.sided', method = 'pearson', conf.level = 0.95); cor.test(var3$wage, var3$hours, alternative = 'two.sided', method = 'pearson', conf.level = 0.95);

alternative - указывает на альтернативную гипотезу и должен быть одним из «two.sided» (по умолчанию), «less» или «greater».

conf.level (level of confidence) – доверительный интервал

sample estimates – выборочная оценка/ выборочный коэффициент корреляции

В статистике любая сумма квадратов отклонений связана с числом степеней свободы df (degress of freedom), которое показывает, сколько независимых отклонений из n возможных участвует в образовании данной суммы квадратов.

p-значение (англ. p-value), p-уровень значимости, p-критерий — вероятность получить для данной вероятностной модели распределения значений случайной величины такое же или более экстремальное значение статистики (среднего арифметического, медианы и др.), по сравнению с ранее наблюдаемым, при условии, что нулевая гипотеза верна.

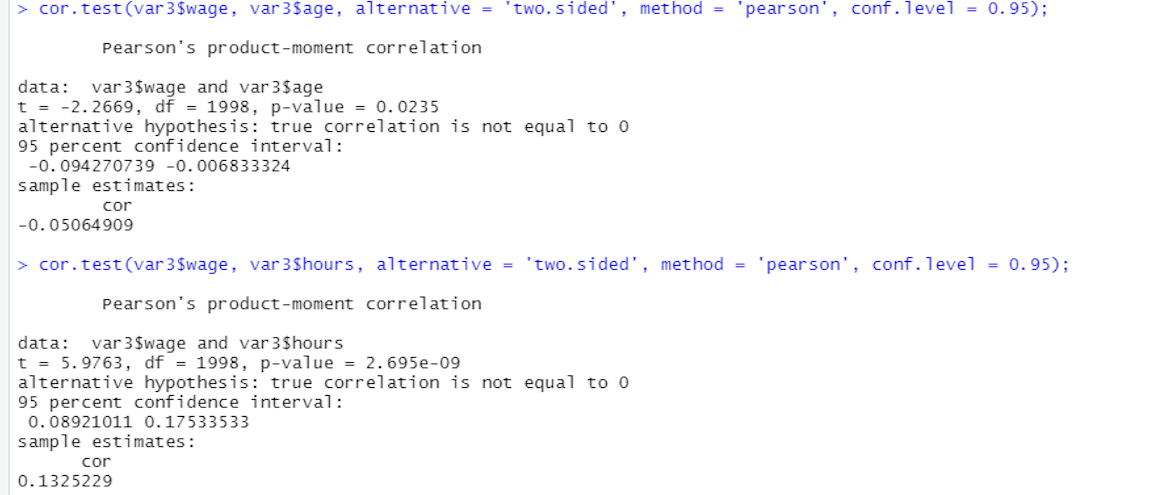


Рисунок 3 - Результаты тестирования на значимость коэффициентов корреляции Пирсона

В обоих случаях (рис. 3) нулевая гипотеза (о равенстве коэффициентов корреляции нулю) отвергается на очень малых уровнях значимости. Следовательно, между рассматриваемыми показателями в обоих случаях есть значимая линейная корреляционная связь.

Значимость линейного коэффициента корреляции проверяется на основе t-критерия Стьюдента. Если t > tкр (P, n-2), то линейный **коэффициент** **корреляции** **значим**, а следовательно, **значима** и статистическая связь X и Y.

Если **гипотеза Н0 о равенстве нулю коэффициента корреляции**будет отвергнута, то выборочный коэффициент значим, а соответствующие величины связаны линейным соотношением. Если **гипотеза Н0** будет принята, то оценка коэффициента не значима, и величины линейно не связаны друг с другом

2.Рассчитаем теперь коэффициенты корреляции Спирмена и проверим гипотезы о значимости для них.

cor.test(var3$wage, var3$age, alternative = 'two.sided', method = 'spearman', conf.level = 0.95); cor.test(var3$wage, var3$hours, alternative = 'two.sided', method = 'spearman', conf.level = 0.95);

Рассчитанные значения соответственно равны -0.09745393 и 0.1624396. Направление связи практически совпадает с результатами для коэффициентов линейной корреляции Пирсона. Значимость коэффициентов также подтверждается (малые значения p-value – 1.266е-05 и 2.7е-13). (рис.4):

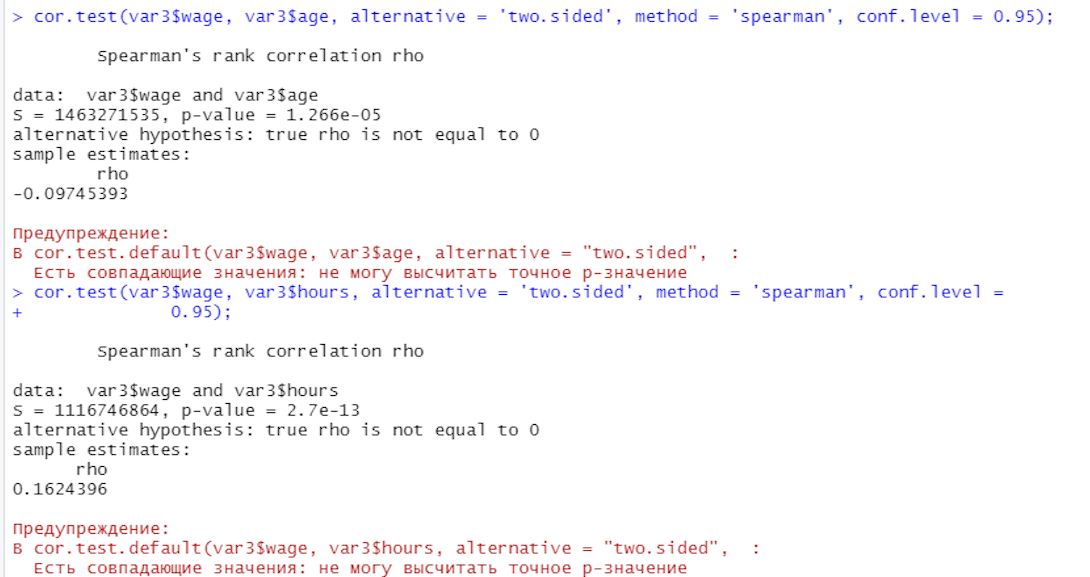


Рисунок 4 – Результаты тестирования на значимость коэффициентов корреляции Спирмена

3.Рассчитаем коэффициенты корреляции (тау) Кендалла и проверим соответствующие гипотезы об их значимости.

cor.test(var3$wage, var3$age, alternative = 'two.sided', method = 'kendall', conf.level = 0.95); cor.test(var3$wage, var3$hours, alternative = 'two.sided', method = 'kendall', conf.level = 0.95).

Рассчитанные значения соответственно равны -0.06693111 и 0.1206498. Можно сделать вывод о том, что направления связи совпадают для всех трех коэффициентов. Степень тесноты связи также очень близка во всех трех случаях. Значимость коэффициентов корреляции Кендалла также подтверждается (малые значения p-value – 1.318е-05 и 2.391е-13). (рис.5):

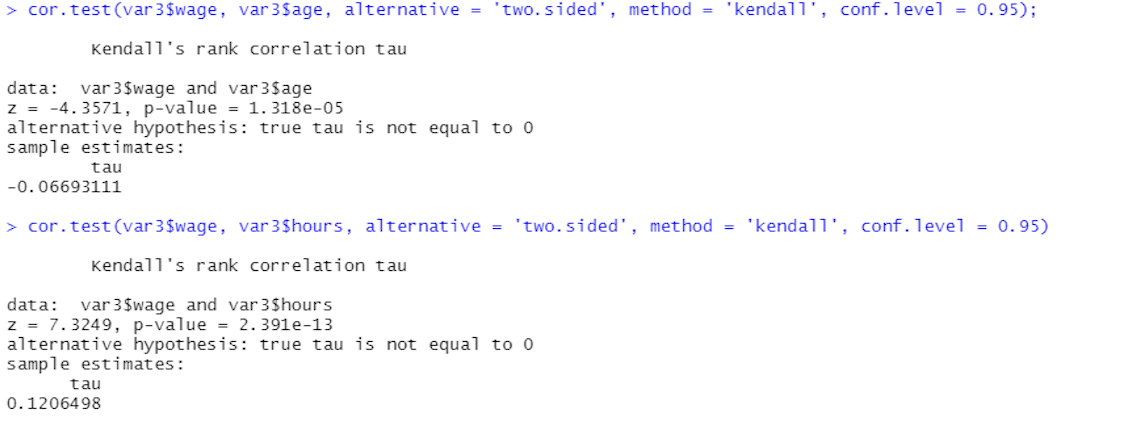


Рисунок 5 – Результаты тестирования на значимость коэффициентов корреляции Кендалла (тау)

4.Сформируем случайную выборку из 200 наблюдений среди 2000 имеющихся:

используем функцию sample:

forlm <- var3[sample(nrow(var3), size = 200, replace=FALSE),]. (рис.6)



Рисунок 6.

replace – замена

sample – образец

Функция nrow в R возвращает количество строк, присутствующих во фрейме данных или матрице.

Для полученной выборки построим точечный график зависимости среднемесячной заработной платы о продолжительности рабочей недели, используя функцию qplot пакета ggplot2 (рис.7):

qplot(forlm$hours, forlm$wage, xlab = 'Продолжительность рабочей недели', ylab='Среднемесячная заработная плата', main = 'Зависимость ЗП от продолжительности рабочей недели').

Пакет ggplot2 предназначен для построения двумерных графиков и диаграмм. В ggplot2 есть альтернативный более простой вариант для создания диаграмм - функция qplot, которая во многом похожа на обычную функцию plot. Она не обладает большим разнообразием возможностей и подходит для быстрого построения не очень сложных графиков.

Параметры main, xlab, ylab – задают соответствующие подписи.

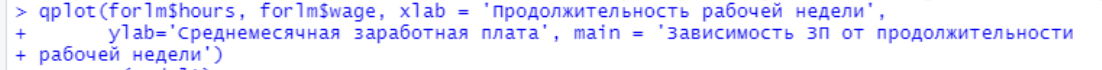


Рисунок 7

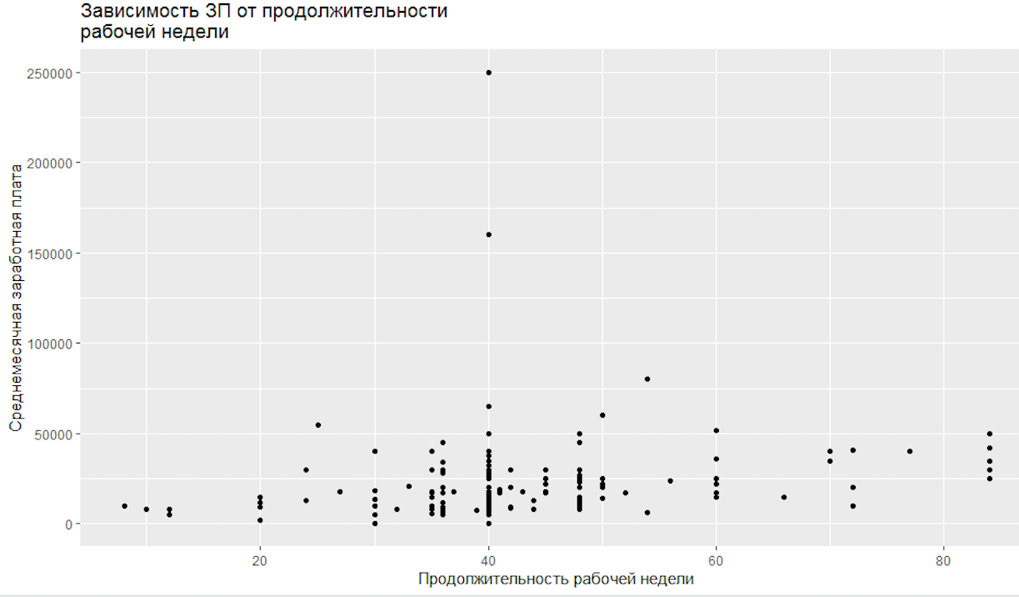


Рисунок 8 – Точечный график зависимости ЗП от продолжительности рабочей недели

По рис. 8 можно предположить, что корреляционная связь прямая, как и было выявлено выше для всей выборки.

5. Оценим теперь по выборке forlm линейную модель зависимости заработной платы от продолжительности рабочей недели. Для этого необходимо воспользоваться такой функцией:

model = lm(wage ~ hours, data = forlm)

data.name - символьная строка, дающая имя (имена) данных.

Далее:

summary (model1)

Результаты оценивания представлены на рис. 9.

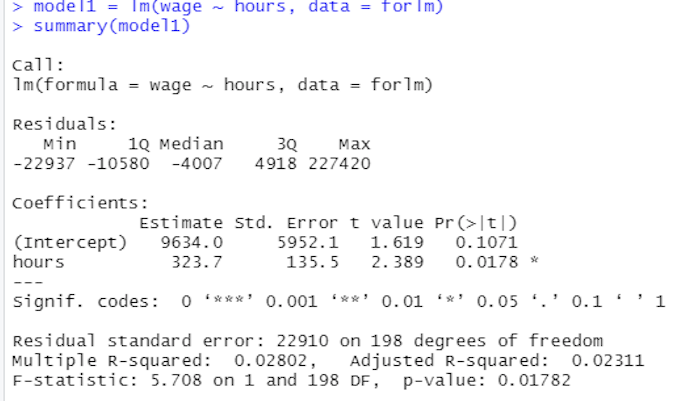


Рисунок 9 – Результаты оценивания линейной модели

Residuals – остаточное количество

Estimate – оценка

Significant codes – значимые коды

Multiple R-squared - кратный r-квадрат

Отметим, что значимым является только свободный член (Intercept), переменная hours является незначимой на уровне значимости 5% (p-value равно 0.0178). Оцененное уравнение также не является значимым (p-value равно 0.01782).

Вывод: в результате проверки Гипотезы о равенстве соответствующих выборочных коэффициентов корреляции нулю сделали вывод о том, что заработная плата в среднем с увеличением возраста падает и, наоборот, растёт при увеличении продолжительности рабочей недели, и в обоих случаях в тестировании на значимость коэффициентов корреляции Пирсона нулевая Гипотеза отвергается, получается что между показателями в обоих случаях есть значимая линейная корреляционная связь.

В результате тестирования на значимость коэффициентов корреляции Спирмена сделали вывод о том, что направление связи практически совпадает с результатами для коэффициентов линейной корреляции Пирсона, значимость коэффициентов также подтверждается.

В результате оценивания линейной модели зависимости заработной платы от продолжительности рабочей недели отметили, что значимым является только свободный член Intercept, а переменная hours и оцененное уравнение не являются значимыми.