

Маголего «Анализ данных»

Гаврилина Александра,
магистратура «Суперкомпьютерное моделирование в науке и инженерии», группа
МСКМ191

aagavrilina@edu.hse.ru

gavrilinasanya@gmail.com

Ответьте, пожалуйста, на вопросы:

Задание 1. Проверка отличия распределения от «нормального»

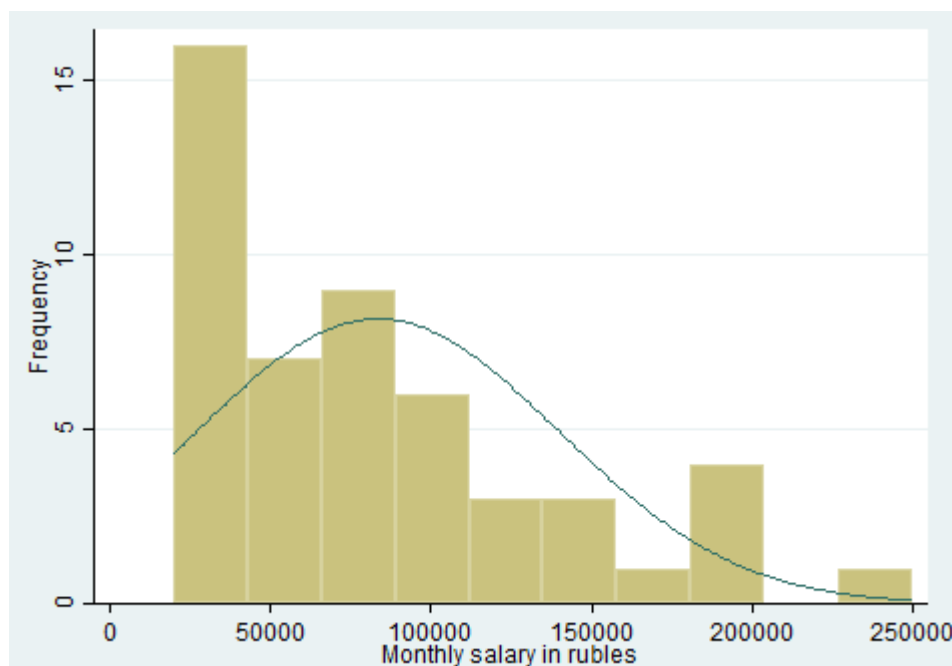
*Откройте файл **salary.dta***

Проанализируйте распределение значений переменной **salary**. Проведите тест для проверки того отличается ли распределение значений этой переменной от нормального распределения. Сформулируйте гипотезы и выводы по результатам анализа.

```
swilk salary
```

```
p-value = 0.00005
```

```
histogram salary, bin(10) frequency normal
```



По имеющимся данным, распределение зарплаты значительно отличается от нормального.

Задание 2. Тест Хи-квадрат

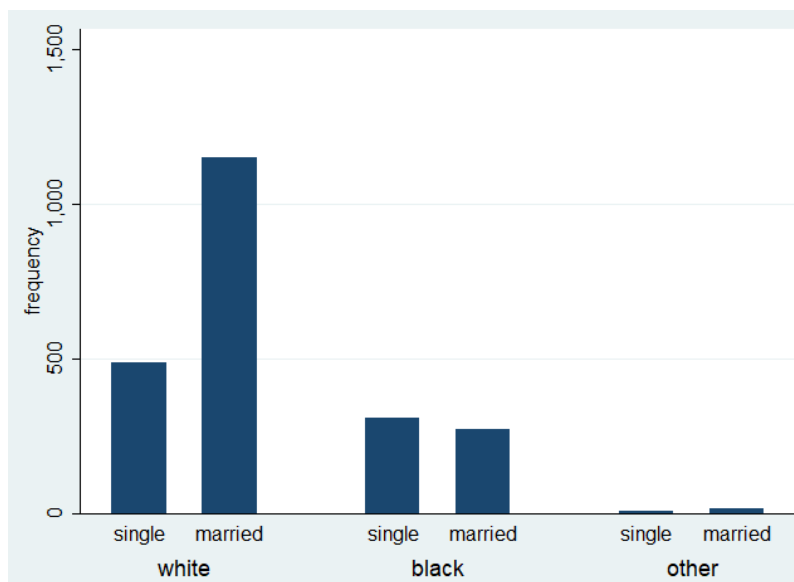
Откройте файл *nlsw88.dta* (example datasets)

Проведите тест Хи-квадрат с парами переменных. Сформулируйте гипотезы. Интерпретируйте результаты анализа. Обоснуйте свои выводы.

a) married и race

tabulate race married, chi2 expected lrchi2

p-value = 0.000

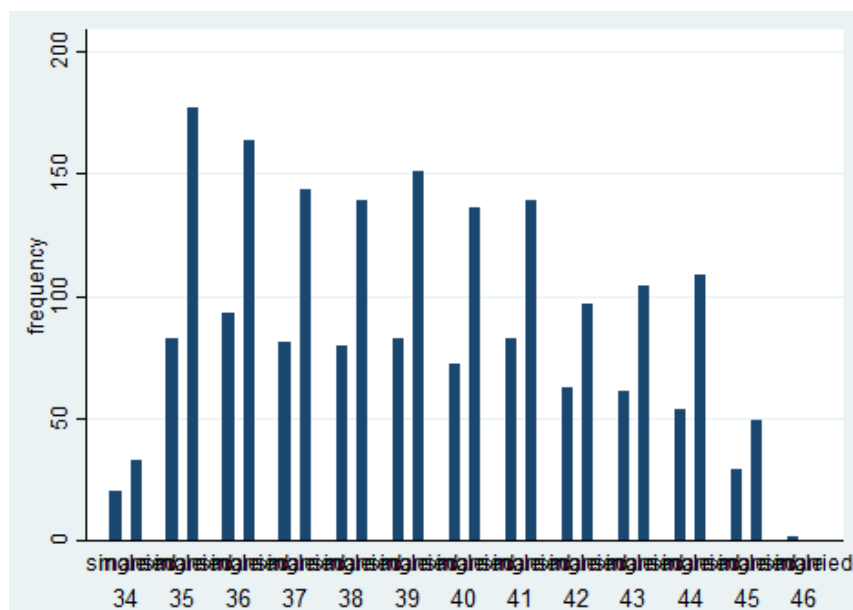


p-value < 0.05 => переменные зависимы

b) age и married

tabulate age married, chi2 expected lrchi2

p-value = 0.831



p-value > 0.05 => переменные независимы

Задание 3. Парная корреляция

Откройте файл *auto.dta* (example datasets)

Подчитайте парный коэффициент корреляции между переменными. Обоснуйте выбор коэффициента. Сформулируйте гипотезы. Оцените значимость коэффициента. Сделайте выводы относительно силы и направления взаимосвязи.

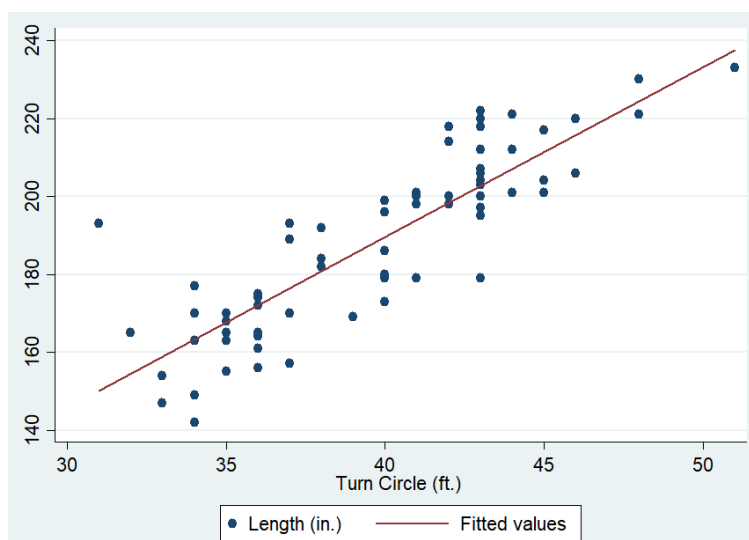
a) length и turn

ktau length turn

corr = 0.7201

length, turn – интервальные переменные, большое число связанных рангов.

Сильная, прямая корреляция



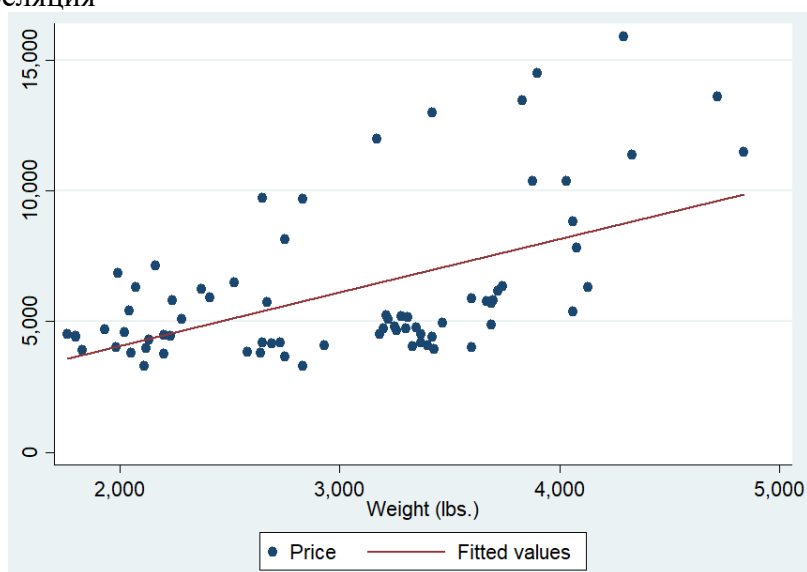
b) price и weight

pwcrr price weight

corr = 0.3416

price, weight – интервальные переменные, большое число связанных рангов

Слабая, прямая корреляция



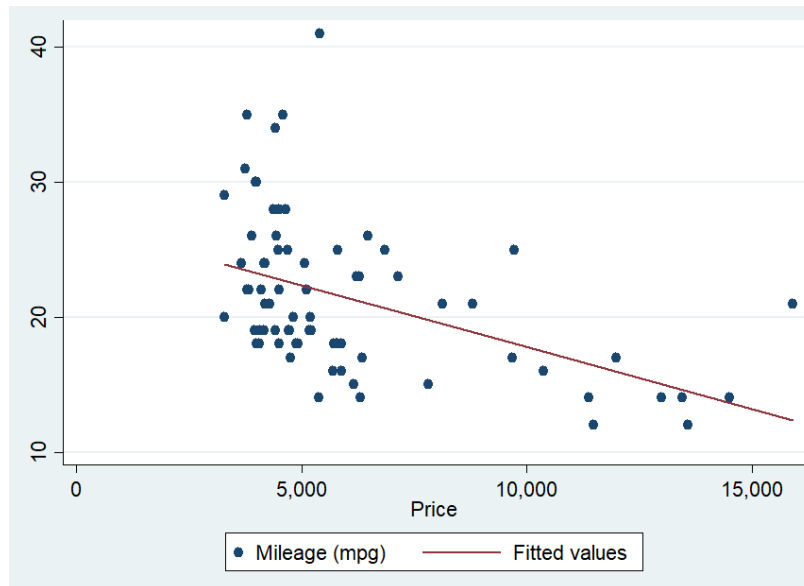
c) mpg и price

`pwcorr mpg price`

`corr = -0.4002`

mpg, price – интервальные переменные, большое число связанных рангов

Слабая, обратная корреляция



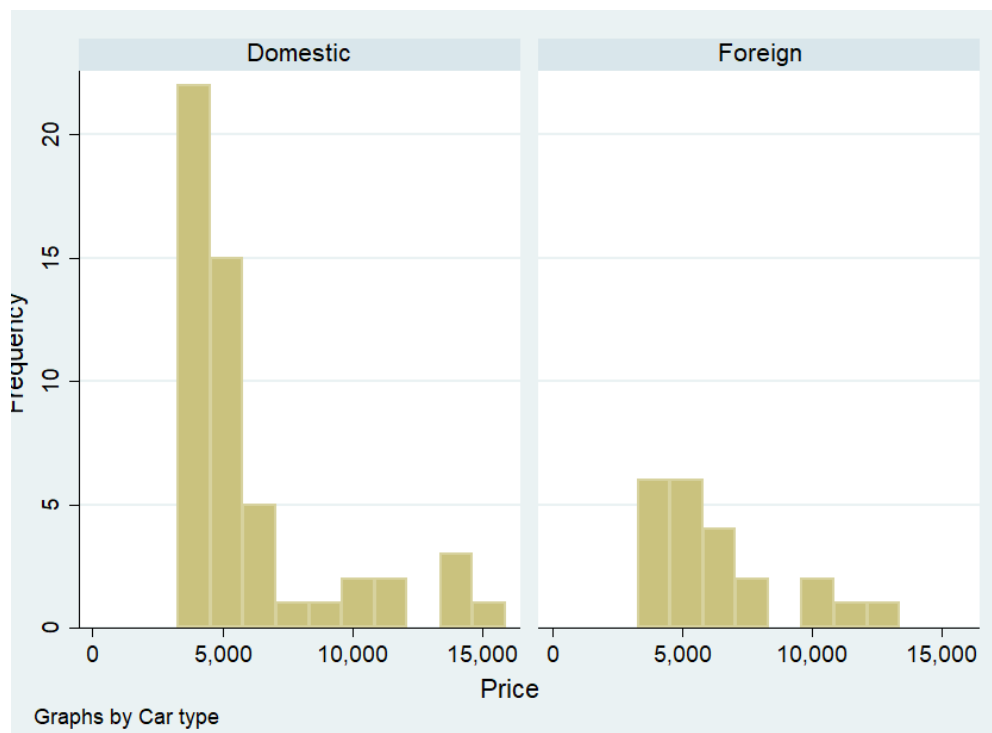
d) foreign и price

`pwcorr price foreign`

`corr = 0.0487`

price, foreign – интервальная и дихотомная переменные

Очень слабая, прямая корреляция



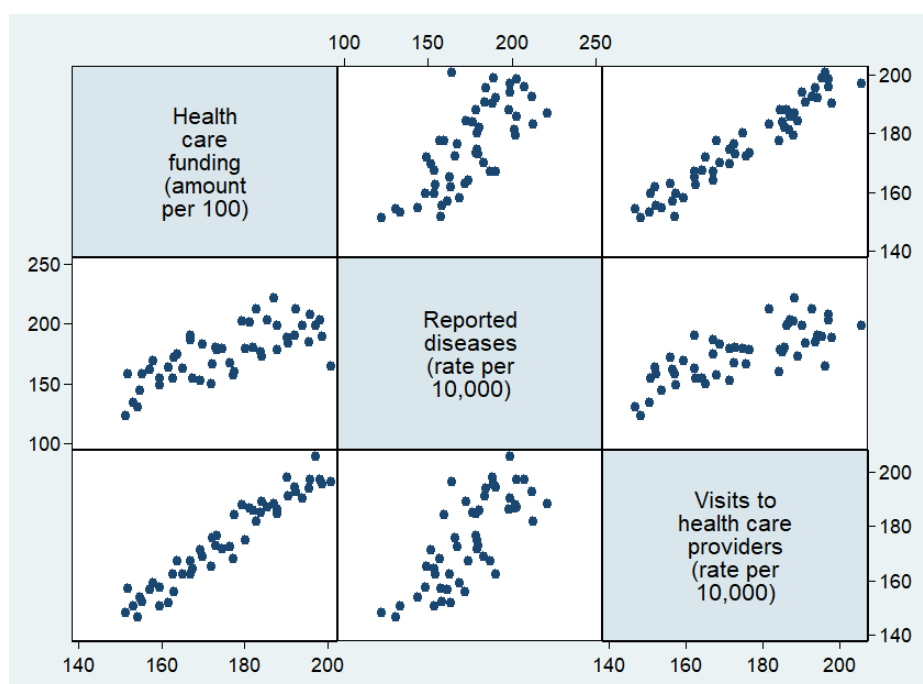
Задание 4. Частная корреляция

Откройте файл *health_funding.dta*

Посчитайте парный коэффициент корреляции между переменными *funding* и *disease*. Посчитайте частный коэффициент корреляции между этими переменными, исключив влияние переменной *visits*. Сделайте выводы.

```
ktau funding disease  
corr = 0.5472
```

```
pcorr funding disease visits  
partial corr = 0.0133  
sign value = 0.9275
```



Простая корреляция: средняя зависимость переменных *funding* и *disease*.

Частная корреляция: эти переменные очень слабо коррелируют, при исключении влияния переменной *visits*.