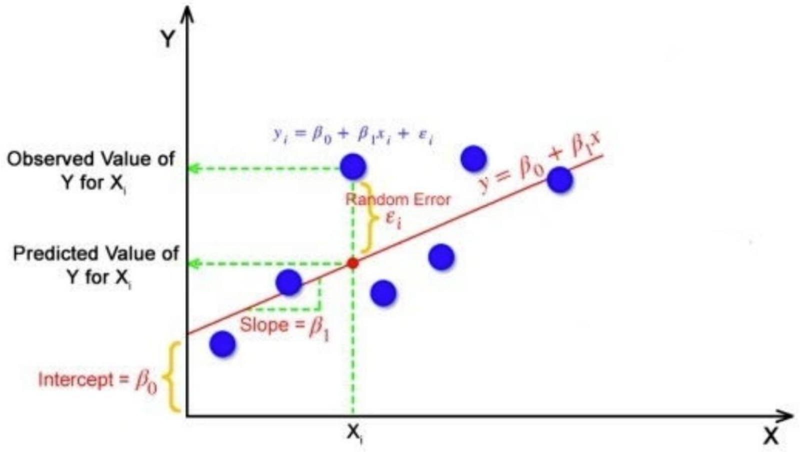
PUBH 501 Biostatistics


Y
X
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
$y = \beta_0 + \beta_1 x$
Observed Value of Y for X_i
Predicted Value of Y for X_i
Random Error ε_i
Slope = β_1
Intercept = β_0
X_i

STATA: LINEAR REGRESSION AND CORRELATION

Overview

- Linear regression
  - Predicting one variable from another
  - Understanding the regression line (slope & intercept)
- Correlation
  - Measuring strength and direction of a relationship
  - Interpreting r vs R squared

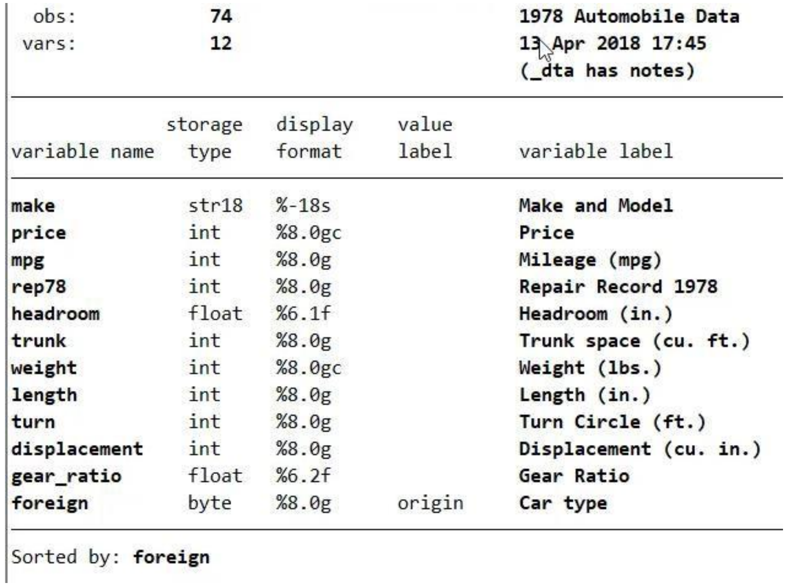
Data: Preinstalled in Stata

- Using Stata dataset: load data using the following code

```
sysuse auto
```

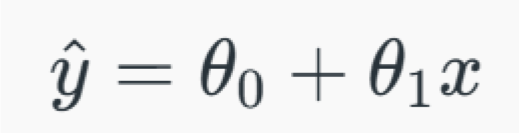
- Data on automobiles from 1978. Mixture of variable types

- `describe`

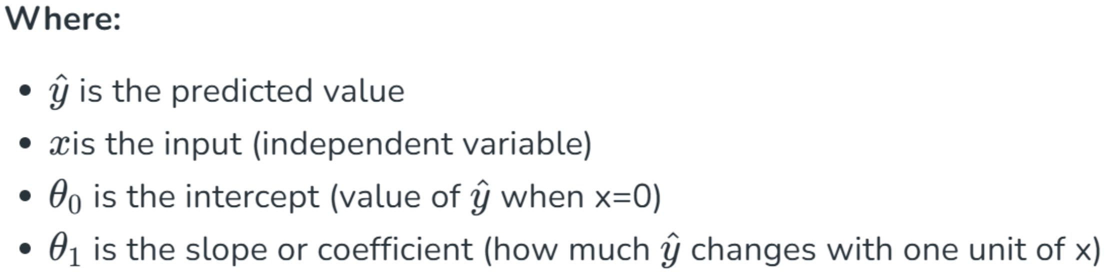


```
  obs:            74                         1978 Automobile Data
 vars:            12                         13 Apr 2018 17:45
                                             (_dta has notes)
------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------
make            str18   %-18s                 Make and Model
price           int     %8.0gc                Price
mpg             int     %8.0g                 Mileage (mpg)
rep78           int     %8.0g                 Repair Record 1978
headroom        float   %6.1f                 Headroom (in.)
trunk           int     %8.0g                 Trunk space (cu. ft.)
weight          int     %8.0gc                Weight (lbs.)
length          int     %8.0g                 Length (in.)
turn            int     %8.0g                 Turn Circle (ft.)
displacement    int     %8.0g                 Displacement (cu. in.)
gear_ratio      float   %6.2f                 Gear Ratio
foreign         byte    %8.0g      origin     Car type
------------------------------------------------------------------
Sorted by: foreign
```

Linear regression



$$\hat{y} = \theta_0 + \theta_1 x$$



Where:

- $\hat{y}$ is the predicted value
- $x$is the input (independent variable)
- θ_0 is the intercept (value of $\hat{y}$ when x=0)
- θ_1 is the slope or coefficient (how much $\hat{y}$ changes with one unit of x)

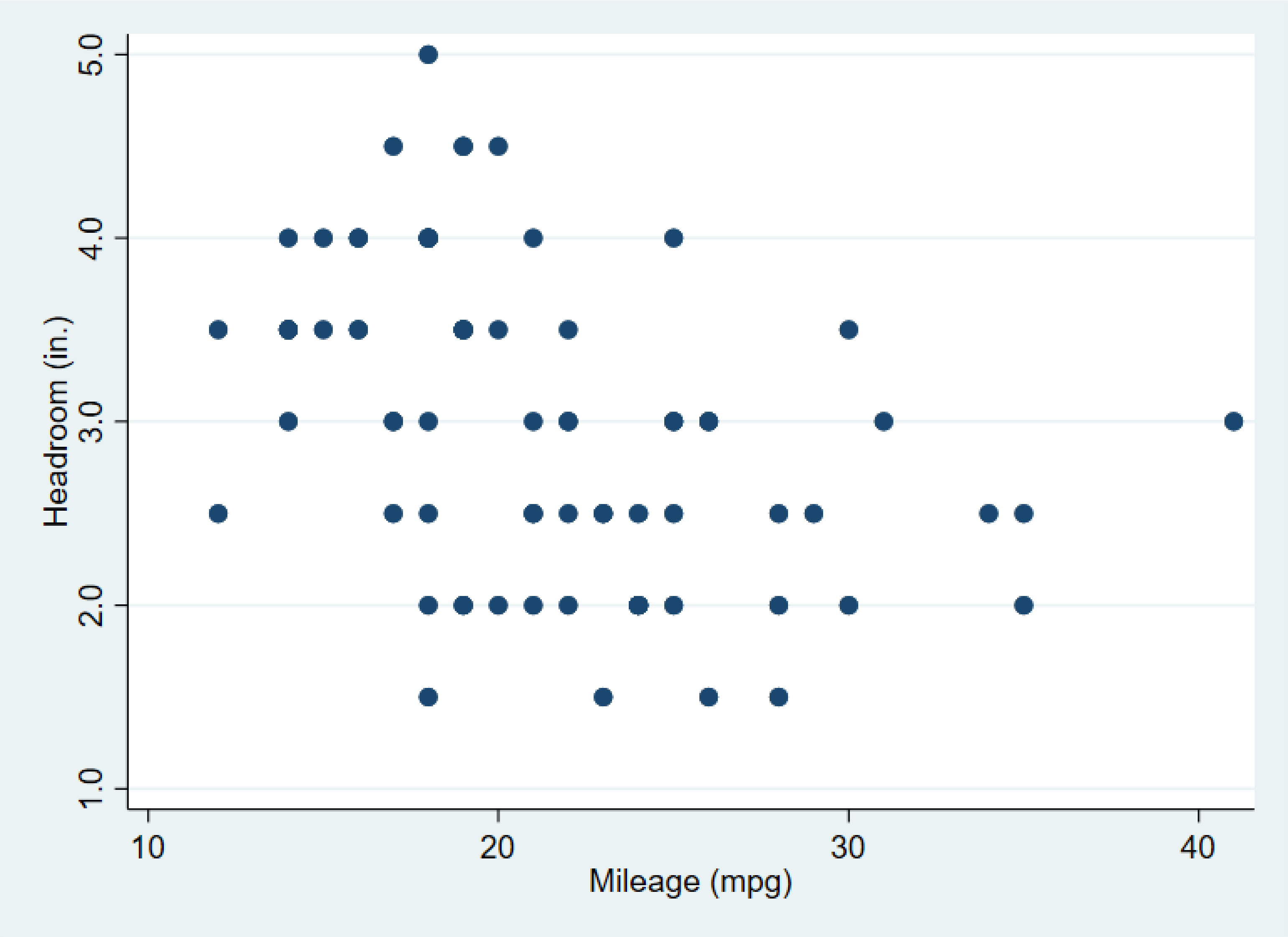
Linear regression

- Look for a linear relationship between an exposure (IV) and a continuous outcome (DV).
- Is car head room associated with gas mileage (mpg)?
- Assumptions:
  - Normality of residuals (can start with check of normality of variable)
  - Relationship between variables is linear (Is there a directional trend?)
  - Homoscedasticity (error terms don’t change based on variable value)

Linear regression

- H_0: slope = 0
- H_1: slope !=0

```
scatter headroom mpg
```



- The scatterplot suggests a negative linear relationship between headroom and mpg

-regress- command

- Linear regression is run using the –regress- command
- `regress outcome exposure`
- `regress headroom mpg`

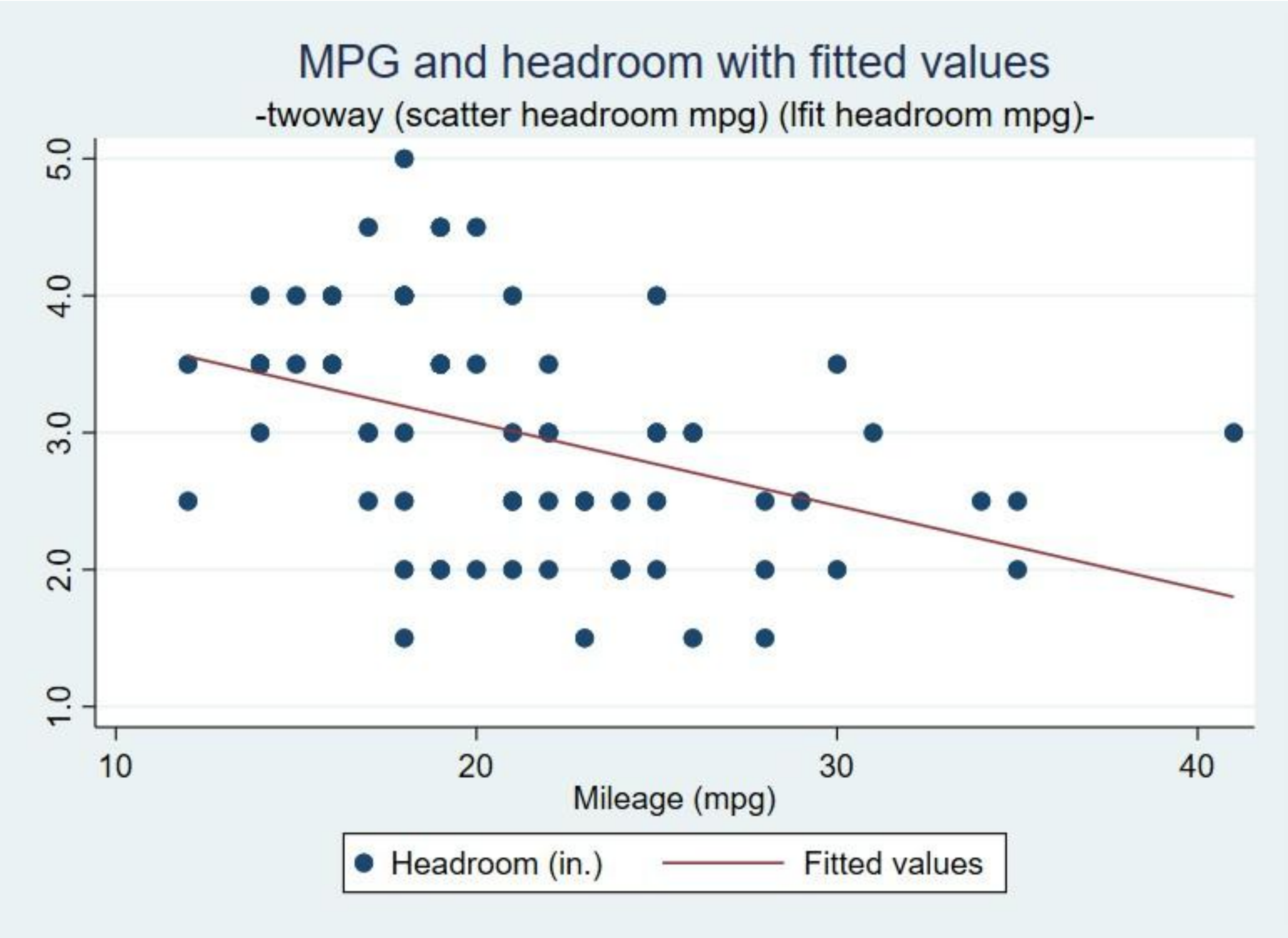
- Where the first variable is the outcome, or the dependent variable
- The second variable is the independent variable

regress headroom mpg

```
      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(1, 72)        =     14.88
       Model |  8.94634123         1  8.94634123   Prob > F        =    0.0002
    Residual |  43.3002804        72  .601392783   R-squared       =    0.1712
-------------+----------------------------------   Adj R-squared   =    0.1597
       Total |  52.2466216        73  .715707146   Root MSE        =     .7755

------------------------------------------------------------------------------
    headroom |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |   -.060509   .0156883    -3.86   0.000    -.0917831   -.0292349
       _cons |   4.281922    .346067    12.37   0.000      3.59205    4.971794
------------------------------------------------------------------------------
```

Reading the output

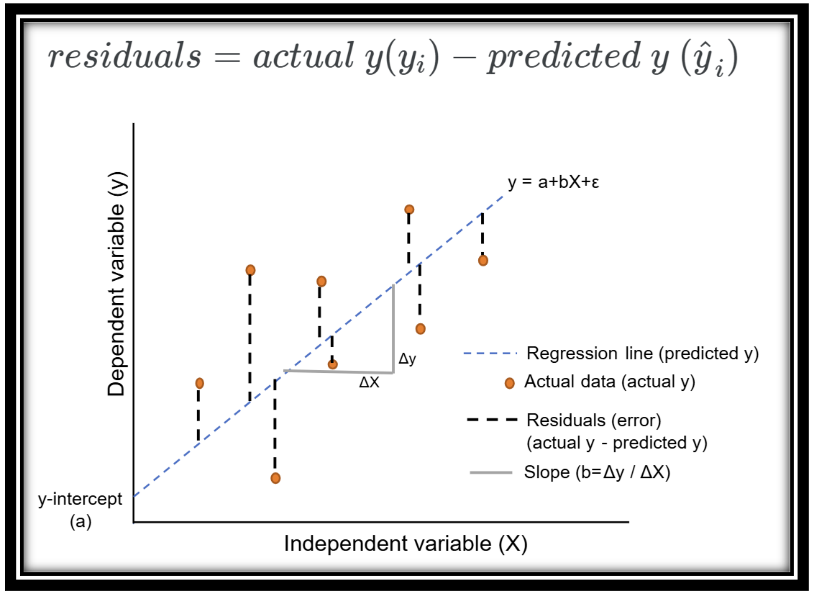
- The coef for mpg is -0.06 (95% CI -.09, -.03). This is the slope of the line
- Coef for _cons is the intercept of the line = 4.3
- R-squared = 0.17 ➔ overall model fit, 17% of the variation in headroom is explained by mpg
- Overall F p-value is <0.001 --> overall model significance


MPG and headroom with fitted values
-twoway (scatter headroom mpg) (lfit headroom mpg)-
5.0
4.0
3.0
2.0
1.0
10
20
30
40
Mileage (mpg)
● Headroom (in.)
Fitted values

Examining residuals

- Assumption of linear regression ➔ residuals are normally distributed
- We looked at the distribution of the variable instead
- Can assess the residuals using the –predict- command *after* you run the regression command

```
regress headroom mpg
predict head_r, residual
hist head_r
```

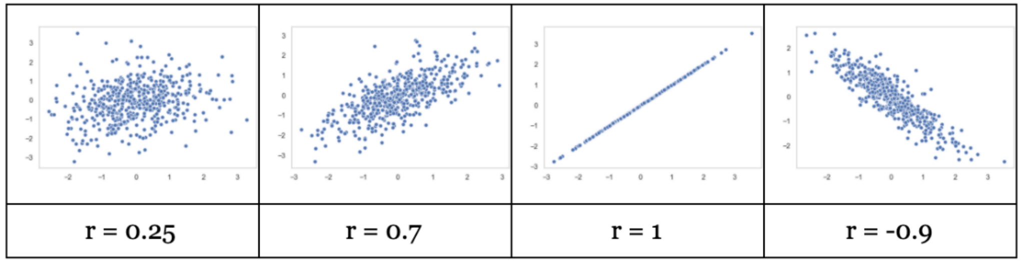


$$residuals = actual\ y(y_i) - predicted\ y\ (\hat{y}_i)$$

Linear regression results summary

- There is a significant association between mpg and headroom in cars
- With each unit increase in mpg, there is a 0.06 unit decrease in head room
- 17% of the variation in headroom can be explained by car mpg

Correlation



Correlation

- We will look at two types of correlation
  - Pearson (parametric)
    - Measures linear relationships between continuous, normally distributed variables
    - Based on actual values
  - Spearman (non-parametric)
    - Monotonic, not necessarily linear relationships
    - Based on ranks of values

Pearson correlation

- Linear relationship
- Continuous and normal distributed variables
- H_0: there is no correlation
- H_1: the variables are correlated

```
pwcorr mpg headroom, sig
             |      mpg headroom
-------------+------------------
         mpg |   1.0000
             |
             |
    headroom |  -0.4138   1.0000
             |   0.0002
             |
```

Significant p-value, correlation coefficient = -0.4138, $p<0.05$

Spearman correlation

- Nonparametric ➔ doesn't assume normality
- Based on ranks
- Again, H_0: there is no correlation

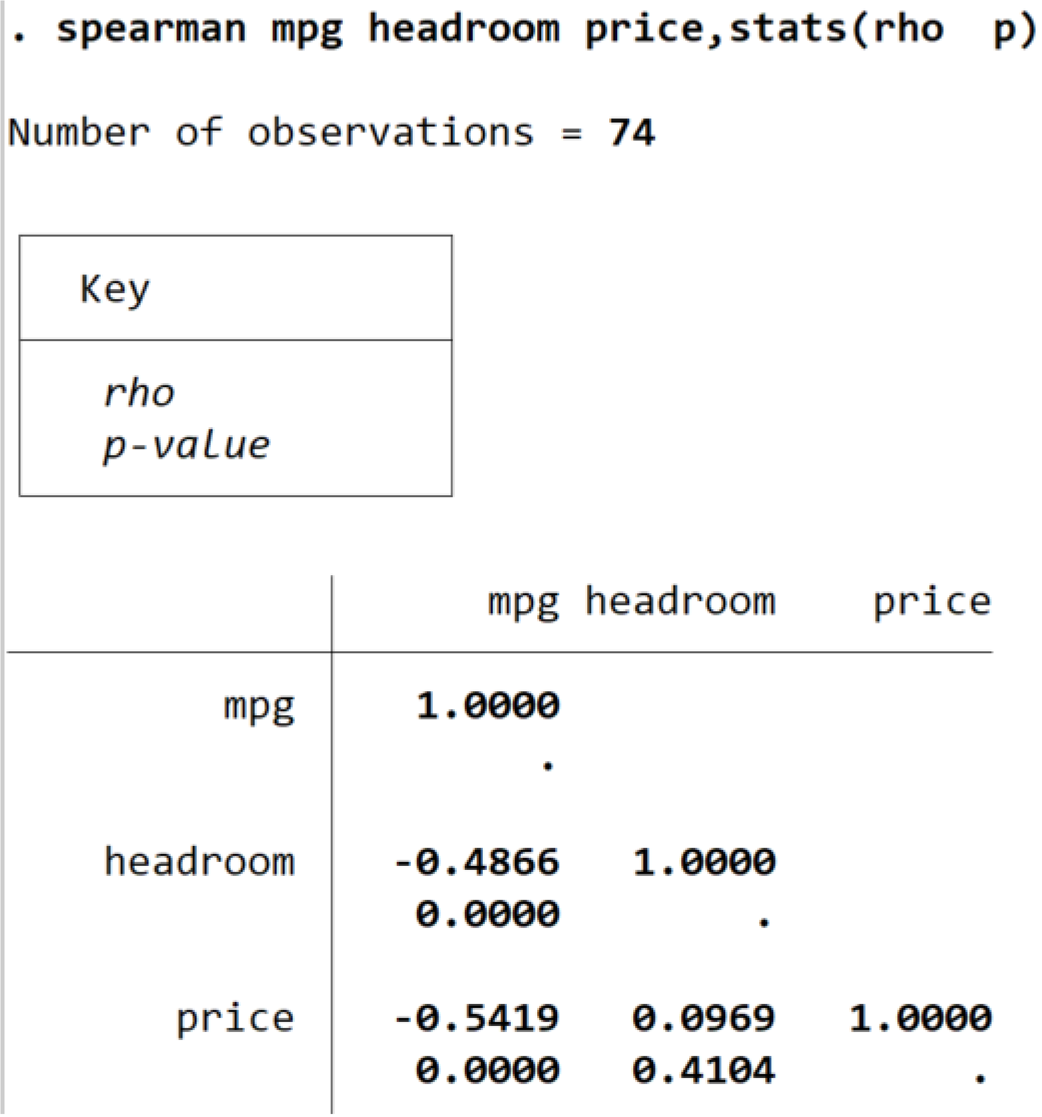
spearman mpg headroom

```
. spearman mpg headroom

 Number of obs =       74
Spearman's rho =       -0.4866

Test of Ho: mpg and headroom are independent
    Prob > |t| =        0.0000
```

Correlation coefficient (r) of -0.4866 is significant p<0.001. However we should use the Pearson correlation coefficient in this case, because we meet assumptions for parametric test.

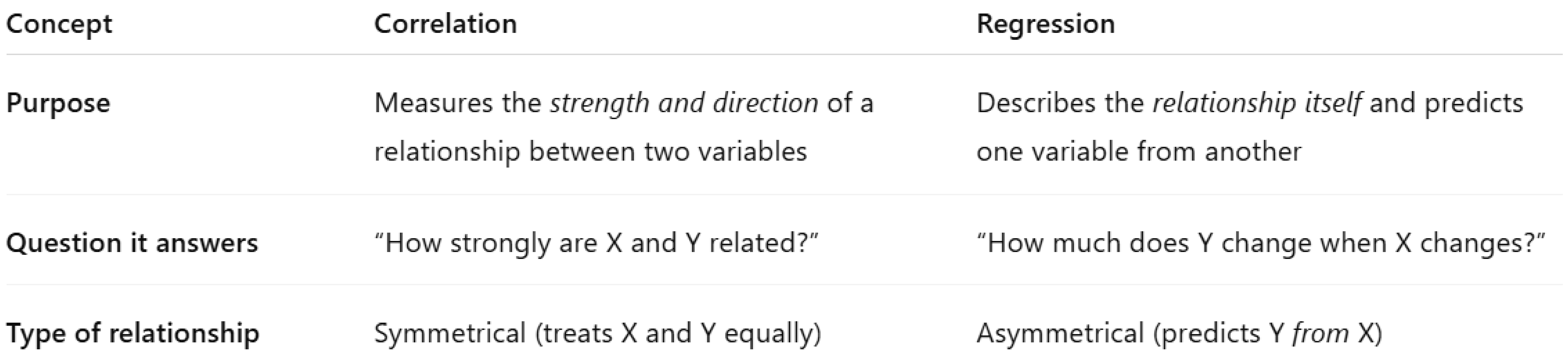
You can look at more than one variable in Spearman, but must include the –stats- option

```
spearman mpg headroom price, stats(rho p)
```



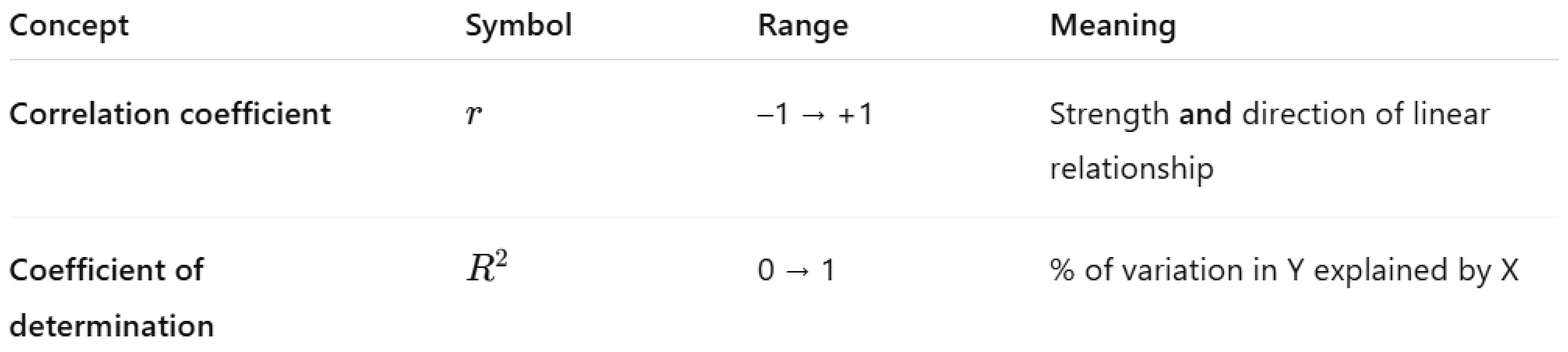
```
. spearman mpg headroom price,stats(rho  p)

Number of observations = 74

  Key

   rho
   p-value

             |      mpg headroom    price
-------------+---------------------------
         mpg |   1.0000 
             |        .
             |
    headroom |  -0.4866   1.0000 
             |   0.0000        .
             |
       price |  -0.5419   0.0969   1.0000 
             |   0.0000   0.4104        .
```

Review



Concept	Correlation	Regression
Purpose	Measures the *strength and direction* of a relationship between two variables	Describes the *relationship itself* and predicts one variable from another
Question it answers	“How strongly are X and Y related?”	“How much does Y change when X changes?”
Type of relationship	Symmetrical (treats X and Y equally)	Asymmetrical (predicts Y *from* X)

Review



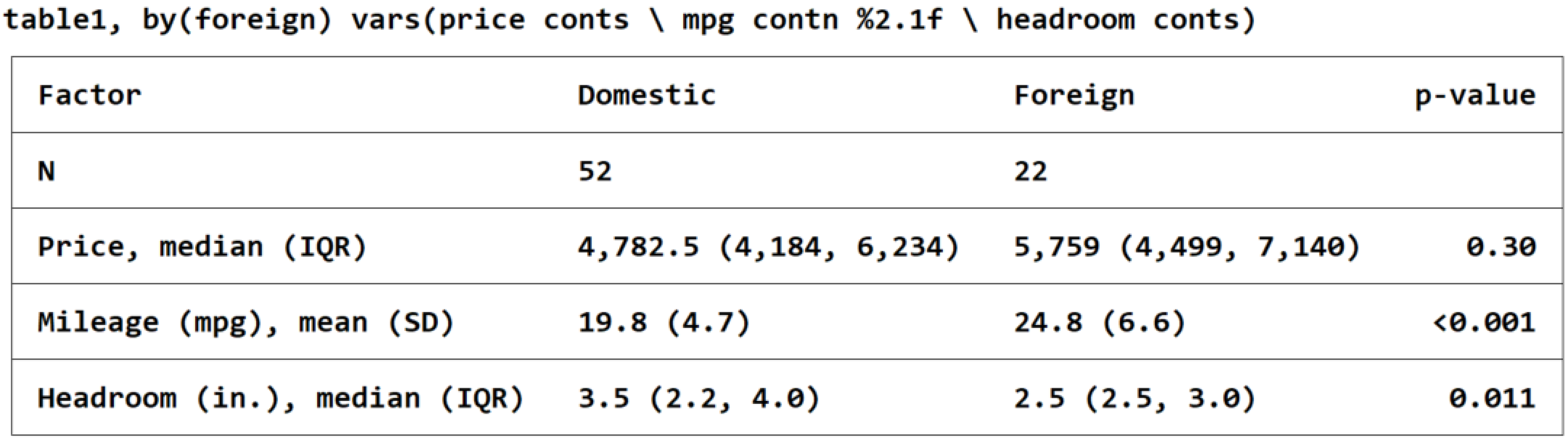
Concept	Symbol	Range	Meaning
Correlation coefficient	r	–1 → +1	Strength **and** direction of linear relationship
Coefficient of determination	R^2	0 → 1	% of variation in Y explained by X

Tip of the day: tables

- There are a lot of commands, user-written and Stata native, that produce publication ready table 1
- Table 1 is usually the first table in the manuscript, which describes your sample
- Some examples include
  - table1
  - tabout

Tables

	Mean	Mean	Mean
	price	mpg	headroom
Car type			
Domestic	6,072.40	19.8	3.2
Foreign	6,384.70	24.8	2.6
Total	6,165.30	21.3	3
Repair Record 1978			
1	4,564.50	21	1.8
2	5,967.60	19.1	3.4
3	6,429.20	19.4	3.2
4	6,071.50	21.7	3
5	5,913.00	27.4	2.5
Total	6,146.00	21.3	3



```
table1, by(foreign) vars(price conts \ mpg contn %2.1f \ headroom conts)
```

Factor	Domestic	Foreign	p-value
N	52	22	
Price, median (IQR)	4,782.5 (4,184, 6,234)	5,759 (4,499, 7,140)	0.30
Mileage (mpg), mean (SD)	19.8 (4.7)	24.8 (6.6)	<0.001
Headroom (in.), median (IQR)	3.5 (2.2, 4.0)	2.5 (2.5, 3.0)	0.011