Pearson Correlation

QAC 201

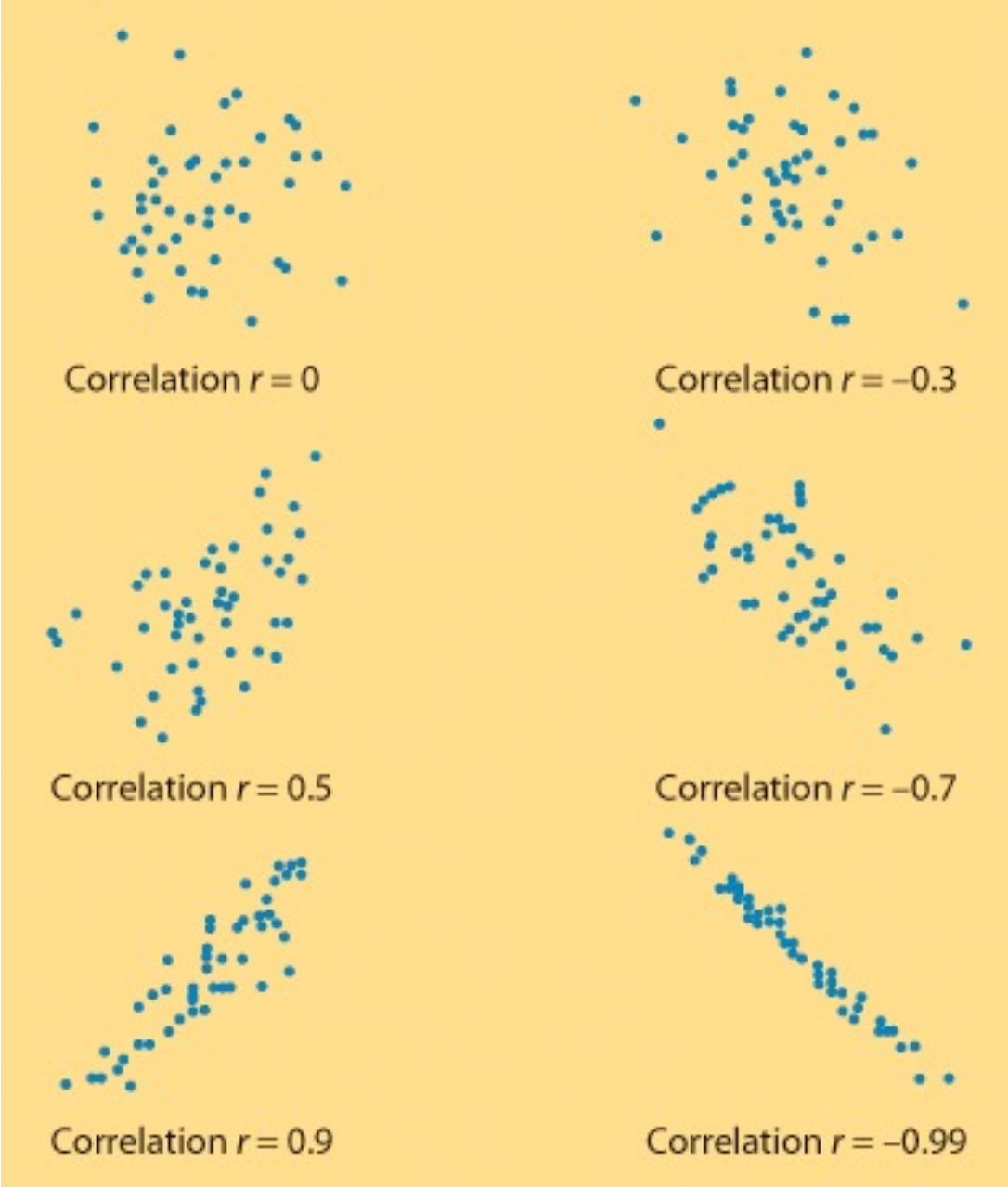
"*r*" ranges from −1 to +1

"r" quantifies the strength and direction of a linear relationship between two quantitative variables.

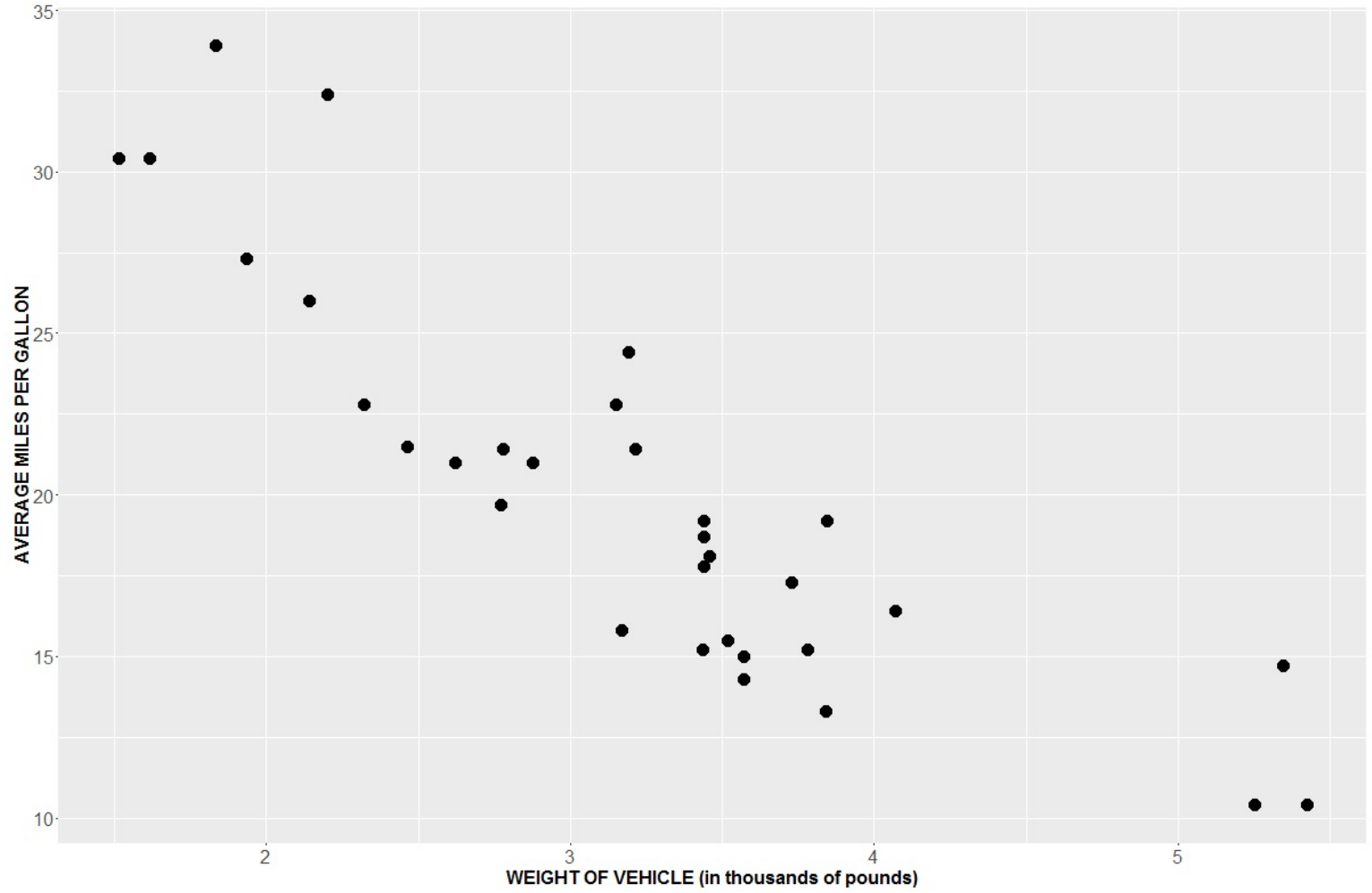
Strength: *How closely the points follow a straight line.*

Direction *is positive when individuals with higher x values tend to have higher values of y.*

Significance: *Is the correlation coefficient significantly different from 0? This requires a statistical test to determine.*



Suppose we want to look at the relationship between mpg of a vehicle and the corresponding weight of a vehicle. Ultimately we would like to see whether weight can help us predict mpg of a car. Weight is in thousands of pounds.



Is there a significant linear relationship between weight and mpg?

One way to test this is with the Pearson Correlation test.

R OUTPUT:

```
        Pearson's product-moment correlation

data:  mtcars$wt and mtcars$mpg
t = -9.559, df = 30, p-value = 1.294e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9338264 -0.7440872
sample estimates:
       cor
-0.8676594
```

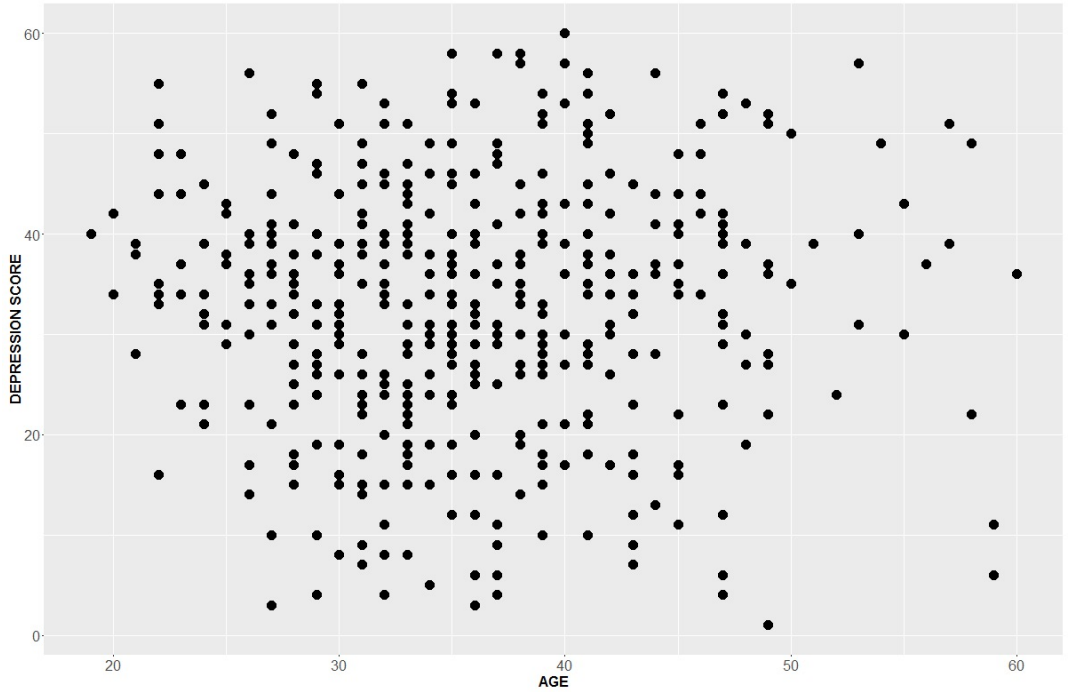
H0: There is no linear association between weight of vehicle and MPG of vehicle

HA: There is a linear association between weight of vehicle and MPG of vehicle

Stata OUTPUT:

	mpg	wt
mpg	1.0000	
wt	-0.8677	1.0000
	0.0000	

Suppose we were interested in the linear relationship between age and depression scores.



```
        Pearson's product-moment correlation

data:  HELPrct$age and HELPrct$cesd
t = 0.17778, df = 451, p-value = 0.859
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.08382526  0.10042509
sample estimates:
        cor
0.008370962
```

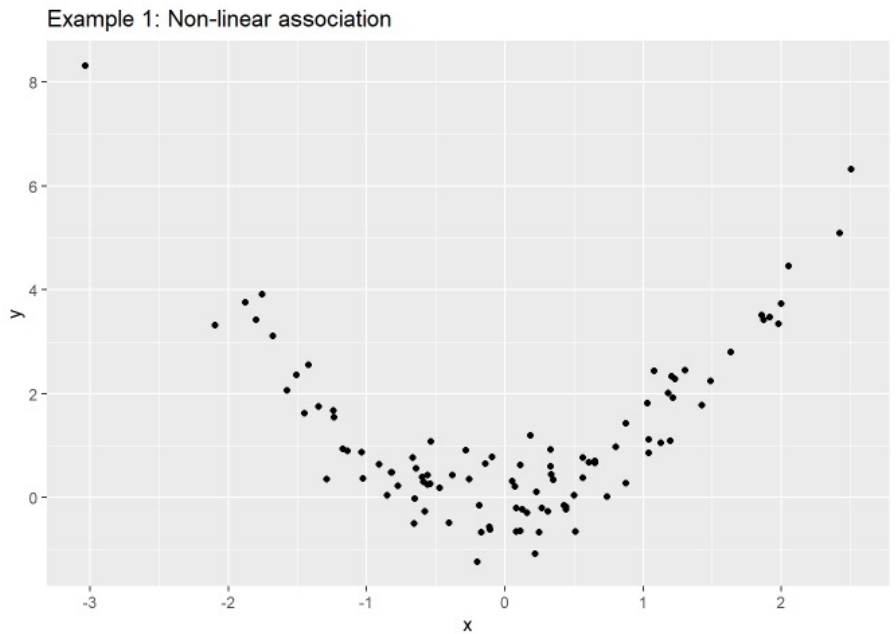
	cesd	age
cesd	1.0000	
age	0.0084 0.8590	1.0000

H0: There is no linear association between age and depression

HA: There is a linear association between age and depression

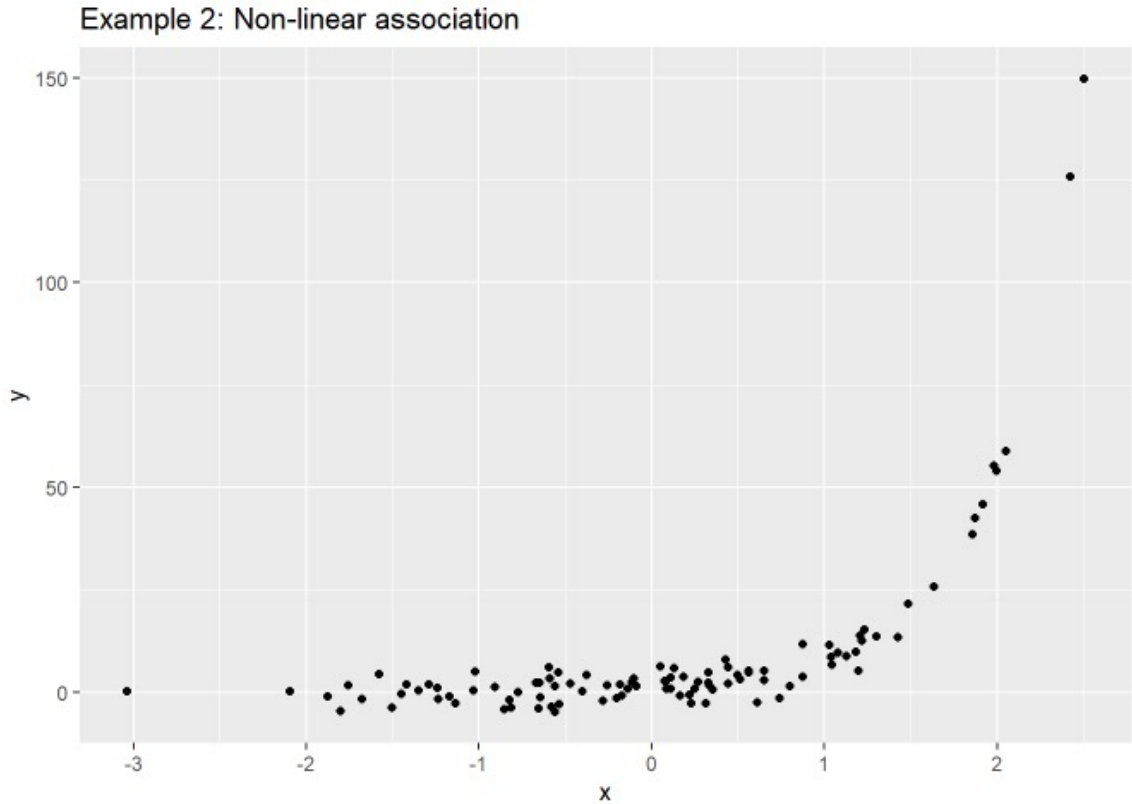
Pearson correlation only tests the degree of the linear association between two variables.

If two variables have a non-linear association, Pearson correlation will not adequately describe the strength of the relationship.




```
cor.test(x, y)
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 1.0636, df = 98, p-value = 0.2901
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.09151605  0.29700858
## sample estimates:
##       cor
## 0.1068222
```




```
cor.test(x, y)
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 7.7489, df = 98, p-value = 8.676e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4778092 0.7250210
## sample estimates:
##       cor
## 0.6163798
```

Sample Questions

- Suppose I want to study the relationship between SAT preparation methods (Kaplan, Princeton Review, Khan Academy, and Self-Study) and improvements in SAT score.
  - What type of graph would be appropriate?
  - What type of statistical test would be appropriate?
  - What are the corresponding statistical hypotheses?
  - Suppose you run the appropriate test and discover the p-value of that test is 0.7862. What is your next step?
  - Suppose you run the appropriate test and discover the p-value of that test is 0.0001. What is your next step?

- Suppose I want to study the relationship between SAT preparation methods (Kaplan, Princeton Review, Khan Academy, and Self-Study) and whether a student feels prepared (Yes/No) to take the test.

    - What type of graph would be appropriate?
    - What type of statistical test would be appropriate?
    - Suppose you run the appropriate test and discover the p-value of that test is 0.0001. What is your next step?

Would it be more useful to report conditional row or conditional column percentages?

Prepared?	Kaplan	Princeton Review	Khan Academy	Self-study
No	20	25	50	150
Yes	10	40	50	200

Conditional Column Percentages:

Prepared?	Kaplan	Princeton Review	Khan Academy	Self-study
No	20/30 =67%	25/65=38%	50/100=50%	150/350=43%
Yes	10/30=33%	40/65=62%	50/100=50%	200/350=57%

Conditional Row Percentages:

Prepared?	Kaplan	Princeton Review	Khan Academy	Self-study
No	20/245 =8%	25/245=10%	50/245=20%	150/245=61%
Yes	10/300=3%	40/300=13%	50/300=17%	200/300=67%