Statistics for CSAI II

4 – Correlation (continued.)

Dr. Travis J. Wiltshire |

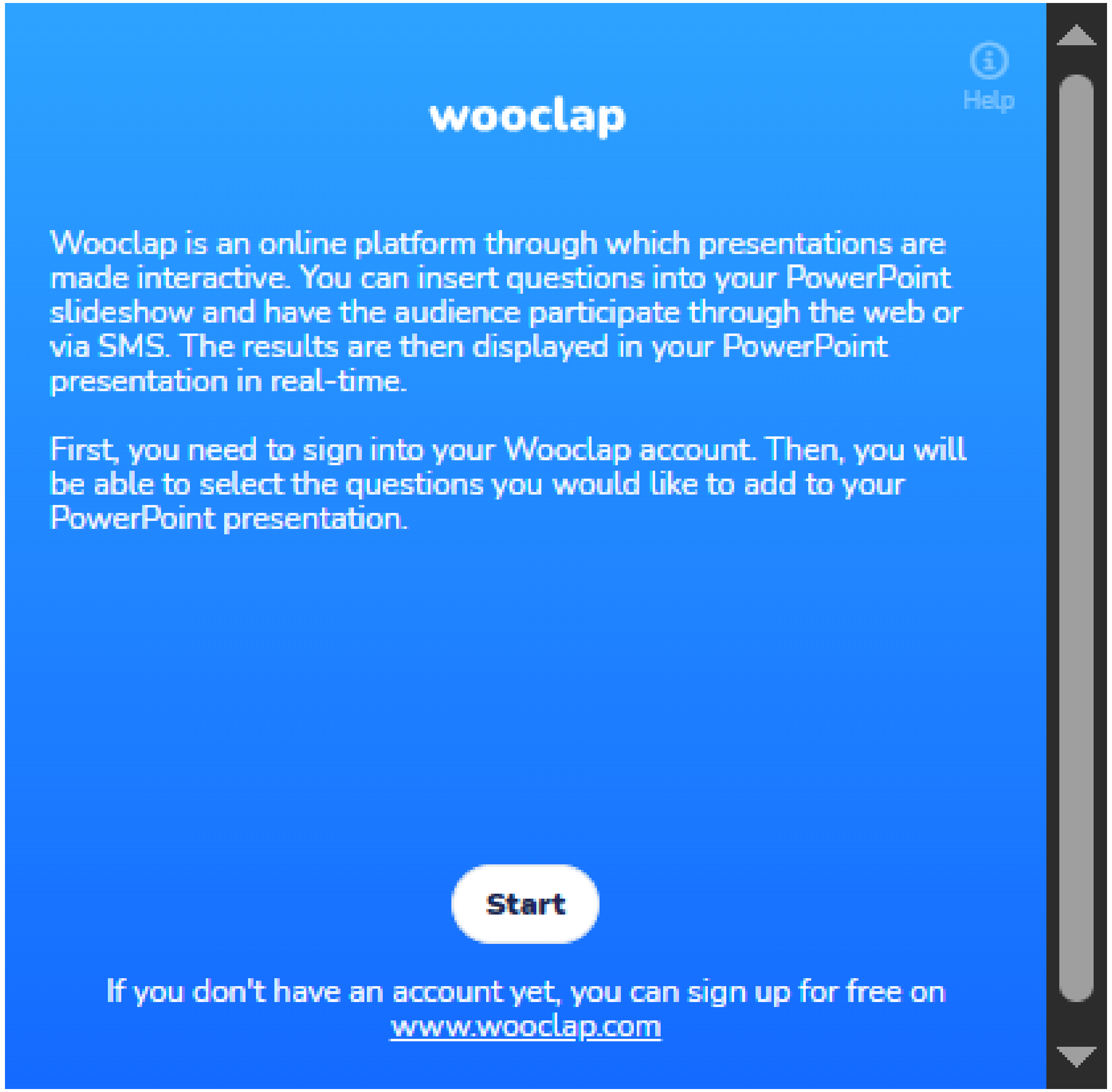
TILBURG UNIVERSITY

Modules

1. Introduction and Probability
2. Sampling Theory
3. Revisiting Hypothesis Testing & Intro to Correlation
4. *Correlation*
5. Intro to Regression
6. More Regression Centering and Checking Assumptions
7. Multiple Regression and Assumptions
8. Interactions
9. Multiple Regression with Categories
10. Multiple Regression with Polynomials
11. Mixed Models
12. Growth Curve Analysis

Outline

1. Correlation (revisited)
2. Nonparametric measures
    - Spearman’s rho
    - Kendall’s tau
3. Interpreting correlations
    - Causality
4. Partial correlations
5. Syllabus/Open Stats Lab
6. PE updates



wooclap

Wooclap is an online platform through which presentations are made interactive. You can insert questions into your PowerPoint slideshow and have the audience participate through the web or via SMS. The results are then displayed in your PowerPoint presentation in real-time.

First, you need to sign into your Wooclap account. Then, you will be able to select the questions you would like to add to your PowerPoint presentation.

If you don't have an account yet, you can sign up for free on www.wooclap.com

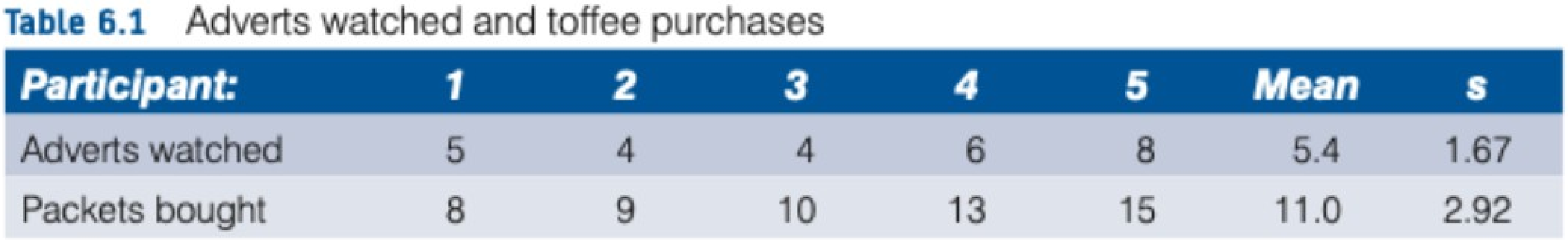
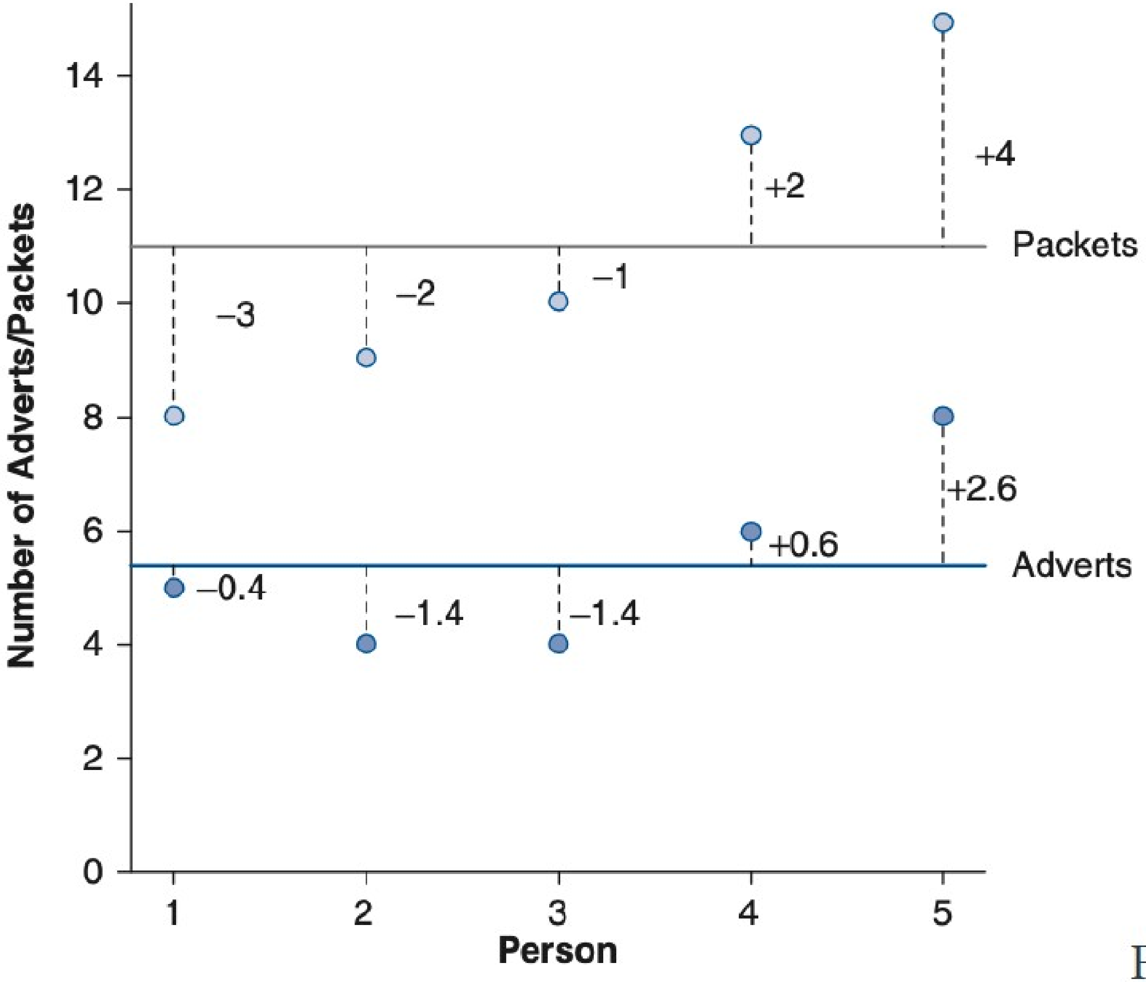


Table 6.1 Adverts watched and toffee purchases

Participant:	1	2	3	4	5	Mean	s
Adverts watched	5	4	4	6	8	5.4	1.67
Packets bought	8	9	10	13	15	11.0	2.92



Field, Miles & Field (2012). *Chapter 6*

Revisiting the Variance

- The variance tells us by how much scores deviate from the mean for a single variable.
- It is closely linked to the sum of squares.
- Covariance is similar – it tells is by how much scores on two variables differ from their respective means.

Sum of squared errors (deviations from the mean) Square to avoid values canceling each other out when summing.

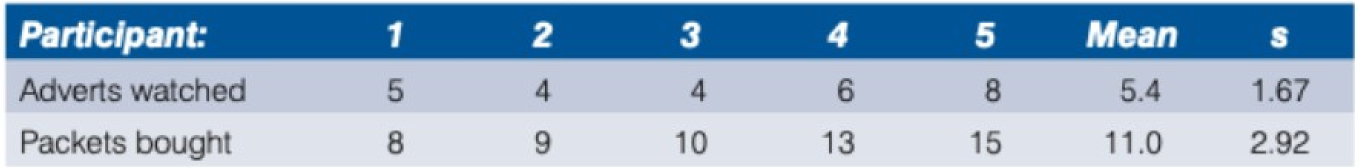
$$\text{variance} = \frac{\boxed{\sum (x_i - \bar{x})^2}}{N - 1}$$

Covariance

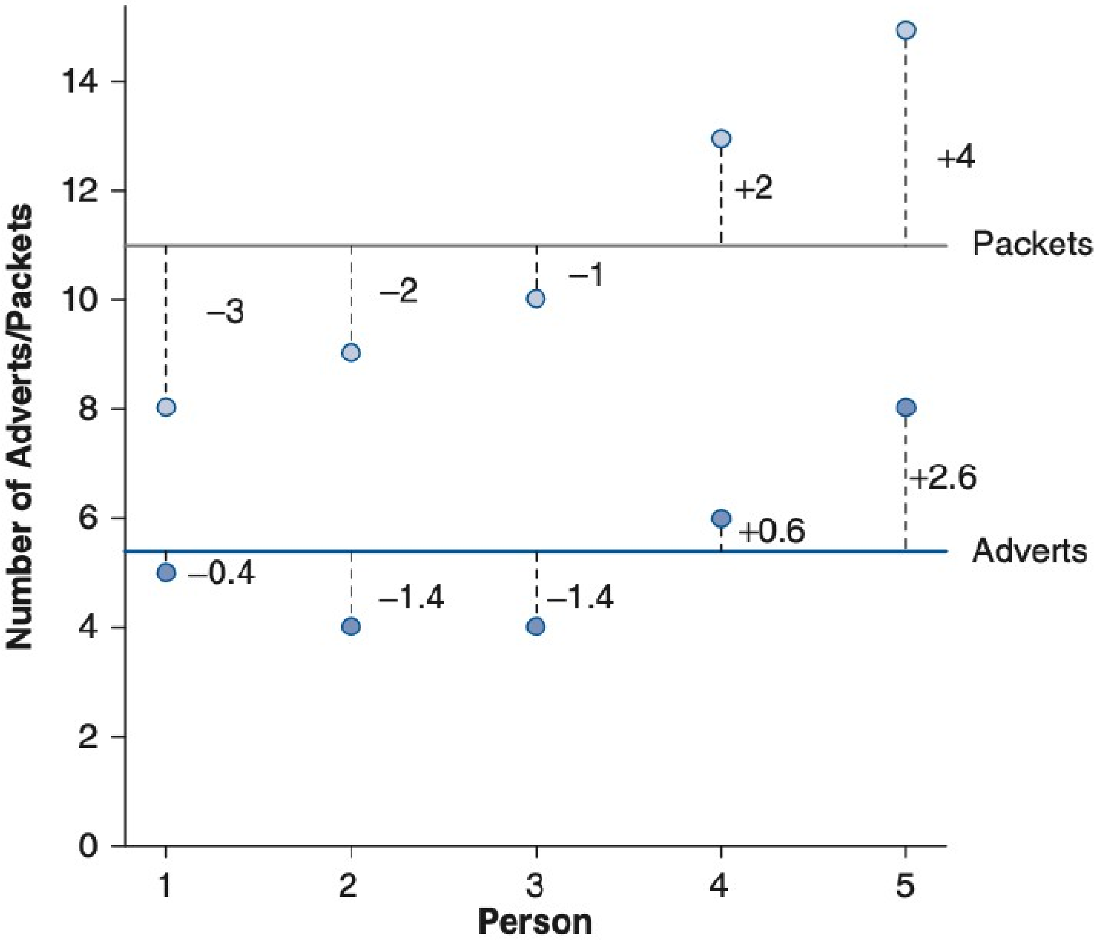
- Calculate the error between the mean and each subject’s score for the first variable (x).
- Calculate the error between the mean and their score for the second variable (y).
- Multiply these error values.
- Add these values and you get the cross product deviations.
- The covariance is the average cross-product deviations:

$$\mathrm{cov}(x \quad y = \sum (\quad \frac{-\qquad\qquad -}{N-}$$

Table 6.1 Adverts watched and toffee purchases



Participant:	*1*	*2*	*3*	*4*	*5*	*Mean*	*s*
Adverts watched	5	4	4	6	8	5.4	1.67
Packets bought	8	9	10	13	15	11.0	2.92



$$
\begin{aligned}
\operatorname{cov}(x,y) &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N-1} \\
&= \frac{(-0.4)(-3) + (-1.4)(-2) + (-1.4)(-1) + (0.6)(2) + (2.6)(4)}{4} \\
&= \frac{1.2 + 2.8 + 1.4 + 1.2 + 10.4}{4} \\
&= \frac{17}{4} \\
&= 4.25
\end{aligned}
$$

Covariance in R

- Create two variables in R:
  - Ads with values 5,4,4,6,8
  - Packets with values 8, 9, 10, 13, 15
- Use the cov() function to calculate the covariance of ads and packets

Problems with Covariance

- It depends upon the **units of measurement**.
  - E.g. the covariance of two variables measured in miles might be 4.25, but if the same scores are converted to kilometres, the covariance is 11.
- One solution: **standardize it!**
  - Divide by the standard deviations of both variables.
- The standardized version of covariance is known as the **correlation coefficient.**
  - It is relatively unaffected by units of measurement.
- BONUS: We can manually standardize variables to put them on the same scale by subtracting the mean from each value and dividing by the standard deviation (turning them into z-scores).

Pearson’s Correlation Coefficient

$$r = \frac{\mathrm{cov}_{xy}}{s_x s_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(N-1) s_x s_y}$$

- Standardized covariance based on standard deviation of the two variables
  - Pearsons product moment correlation coeffecient
  - Ranged between -1 and +1 (where 0 means no correlation)

$$r =$$

$$= \frac{4.25}{1.67 \times \qquad 2}$$

$$=$$

Standardizing and Correlation in R

- Create new variables by standardizing both variables below using the scale() function:

  - Ads

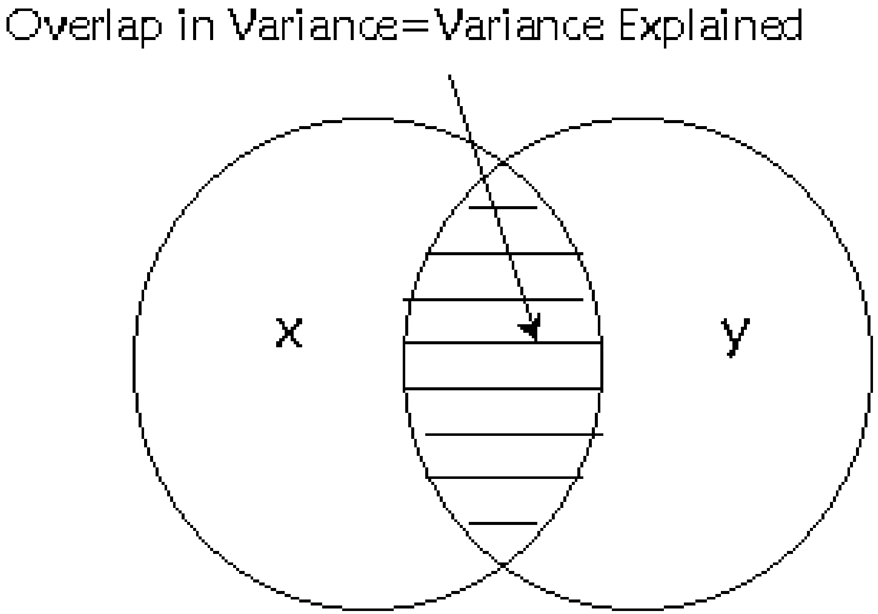
  - Packets

- Use the cov() function to calculate the covariance of the **standardized versions** of ads and packets

- Use the cor() function to calculate the correlation of ads and packets

Things to Know about the Correlation

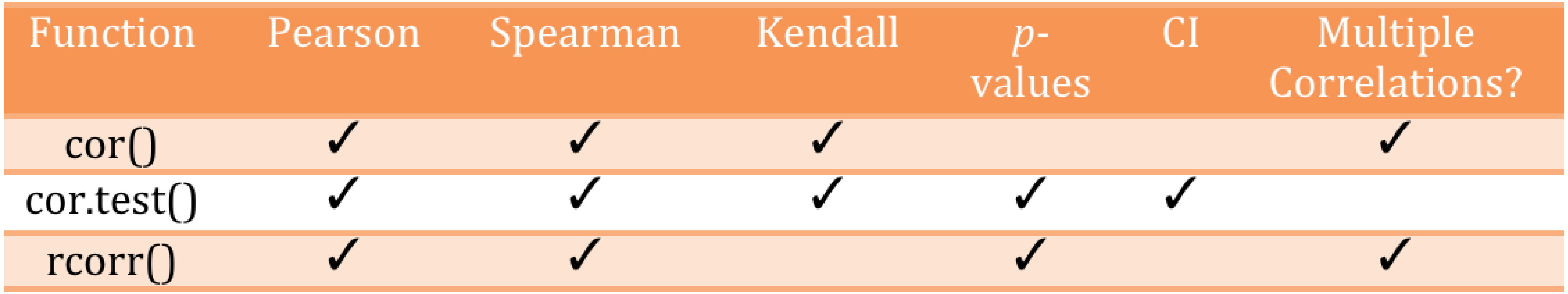
- It varies between -1 and +1
  - 0 = no relationship
- It is an effect size
  - ±.1 = small effect
  - ±.3 = medium effect
  - ±.5 = large effect
- Coefficient of determination, r^2
  - By squaring the value of r you get the proportion of variance in one variable shared by the other.



General Procedure for Correlations Using R

- To compute basic correlation coefficients there are three main functions that can be used:

  cor(), *cor.test()* and *rcorr()*.



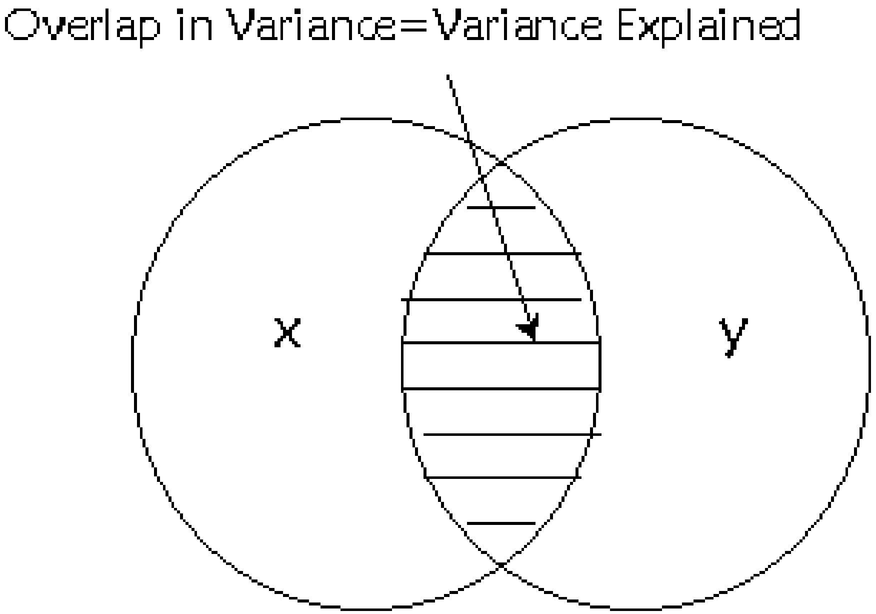
Function	Pearson	Spearman	Kendall	*p*-values	CI	Multiple Correlations?
cor()	✓	✓	✓			✓
cor.test()	✓	✓	✓	✓	✓	
rcorr()	✓	✓		✓		✓

Hypothesis testing with correlations

- We can use cor.test() to test the null hypothesis that the correlation in the population is 0.
- And we can also specify our alternative for whether there should be a negative relationship (alternative =‘less’) or positive association (alternative=‘greater’)
- Also provides p-values, and CIs

Things to Know about the Correlation

- It varies between -1 and +1
  - 0 = no relationship
- It is an effect size
  - ±.1 = small effect
  - ±.3 = medium effect
  - ±.5 = large effect
- Coefficient of determination, r^2
  - By squaring the value of r you get the proportion of variance in one variable shared by the other. *Variance accounted for by…*

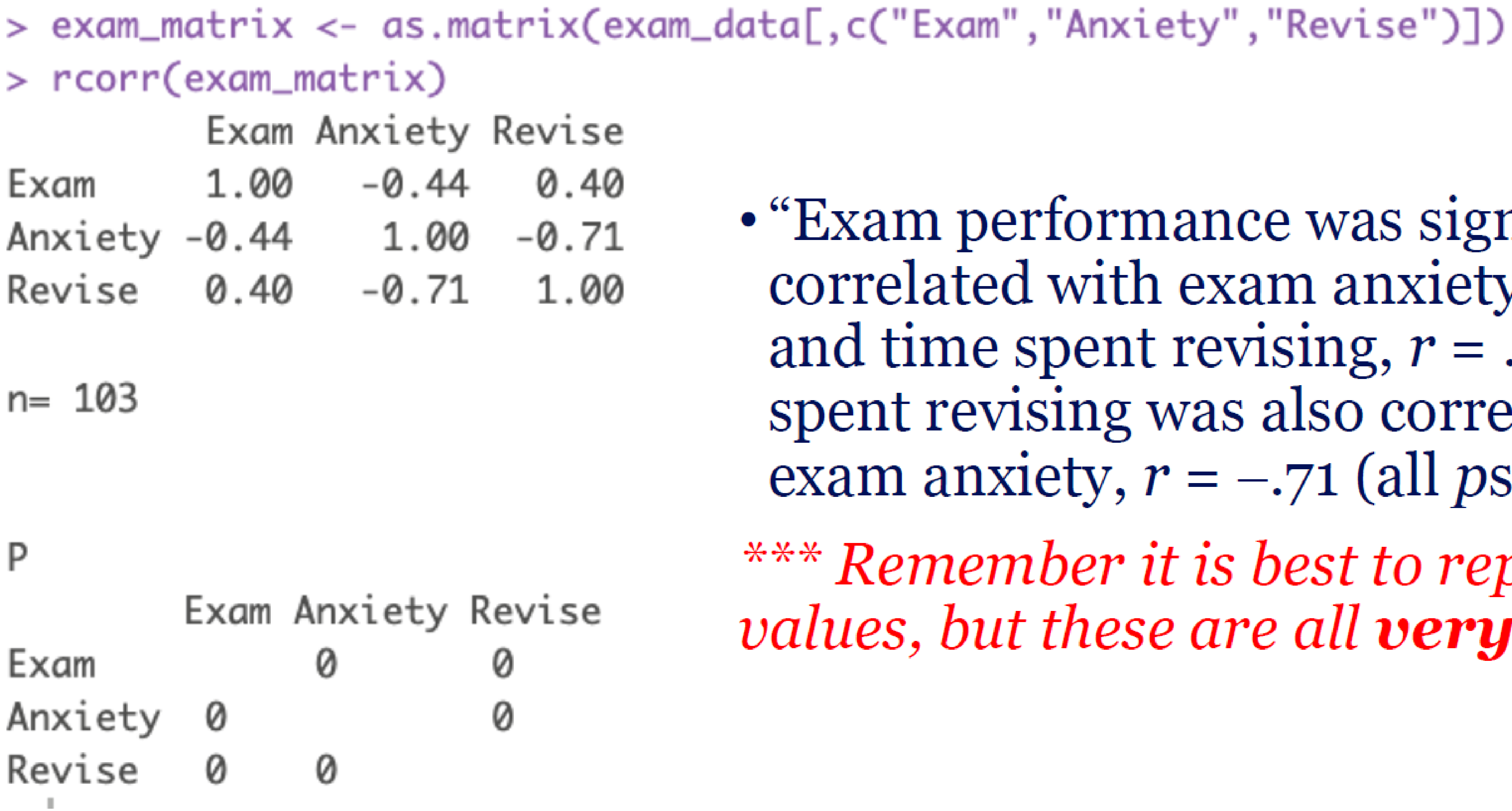


Correlation Testing in R

- Use the cor.test() function to calculate the correlation of ads and packets
- Use the cor.test() function to calculate the correlation of ads and packets predict a negative association
- Use the cor.test() function to calculate the correlation of ads and packets predict a positive association

- What do these results suggest to you? Is the correlation between these variables significant?

Reporting results of a correlation



```
> exam_matrix <- as.matrix(exam_data[,c("Exam","Anxiety","Revise")])
> rcorr(exam_matrix)
         Exam Anxiety Revise
Exam     1.00   -0.44   0.40
Anxiety -0.44    1.00  -0.71
Revise   0.40   -0.71   1.00

n= 103


P
        Exam Anxiety Revise
Exam          0       0
Anxiety  0            0
Revise   0    0
```

- “Exam performance was significantly correlated with exam anxiety, $r = -.44$, and time spent revising, $r = .40$; the time spent revising was also correlated with exam anxiety, $r = -.71$ (all ps < .001)”

**** Remember it is best to report exact p-values, but these are all **very** small.*

Non-parametric Correlations

(some) assumptions of Pearson’s Correlation

- Interval or ratio scale data
- Normally distributed

We can then use:

- Spearman’s rho
  - Pearson’s correlation on the ranked data
  - With outliers, ordinal data
- Kendall’s tau
  - Better than Spearman’s for small samples
  - Based on defining concordant and discordant pairs

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

ρ = Spearman's rank correlation coefficient

d_i = difference between the two ranks of each observation

n = number of observations

Non-Parametric Correlation Testing in R

- Load the The Biggest Liar.dat file into R and inspect it.
  - 68 participants, variables indicate where they were placed in the competition (first, second, third, etc.), Creativity questionnaire (maximum score 60)
  - Make a prediction about the relationship between position in the competition and creativity and test it using the cor.test() function use method spearman and kendall

    -What did you observe? Was there a significant correlation? What is the size (small, medium, large)? Do both analyses give you similar conclusions?

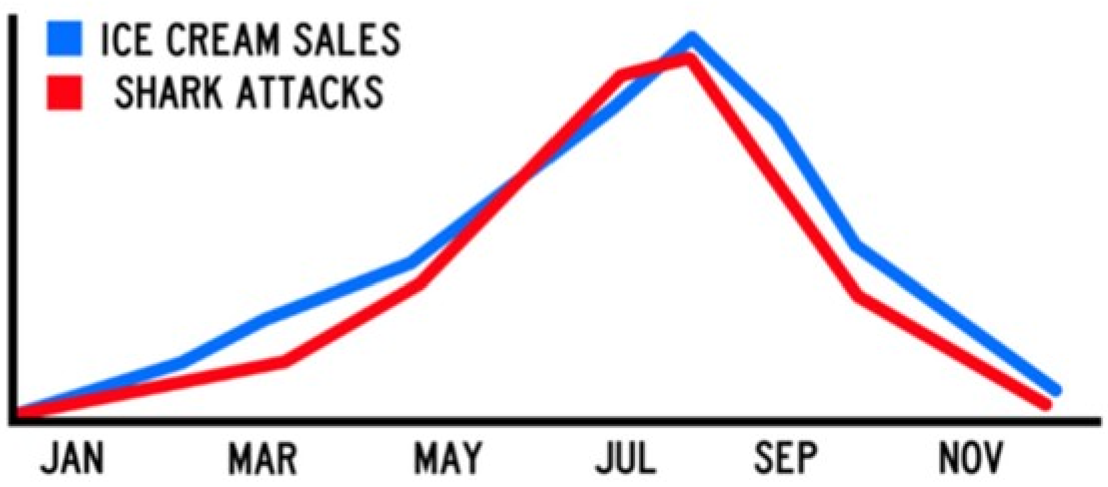
    -How would you report your findings? Try to summarize your results in a write up.

Correlation and Causality

- The third-variable problem:
  - In any correlation, causality between two variables cannot be assumed because there may be other measured or unmeasured variables affecting the results.
- Direction of causality:
  - Correlation coefficients say nothing about which variable causes the other to change.

Check this out for some interesting spurious correlations

CORRELATION IS NOT CAUSATION!



Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)

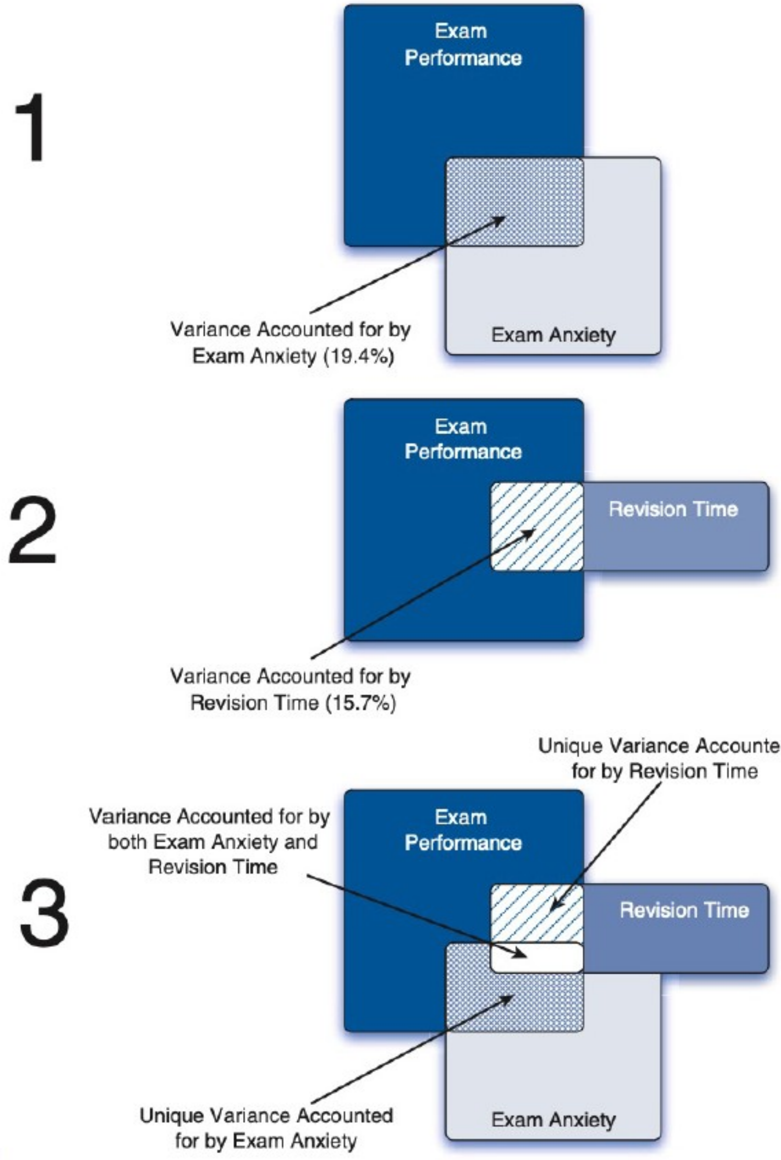
Partial and Semi-partial Correlations

- *Partial correlation:*
  - Measures the relationship between two variables, controlling for the effect that a third variable has **on them both**.

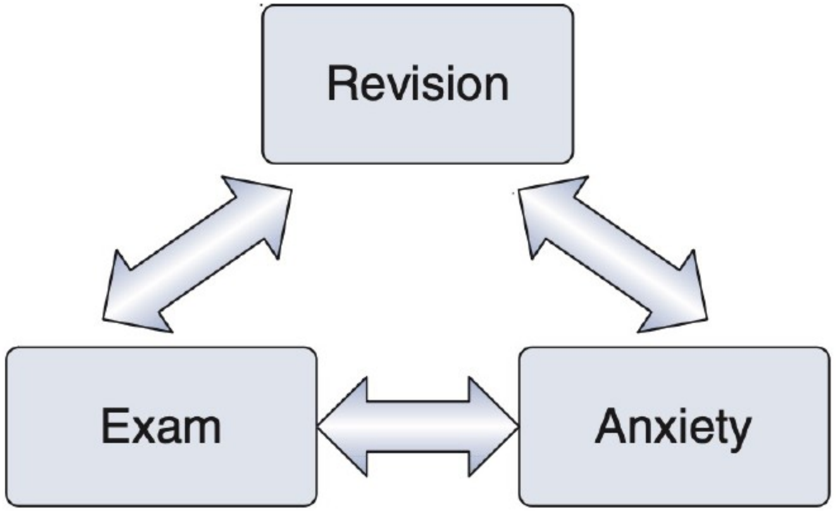
$$pr_1 = \frac{r_{Y1} - r_{Y2}r_{12}}{\sqrt{1-r_{Y2}^2}\sqrt{1-r_{12}^2}}$$

- *Semi-partial correlation:*
  - Measures the relationship between two variables controlling for the effect that a third variable has **on only one of the others**.

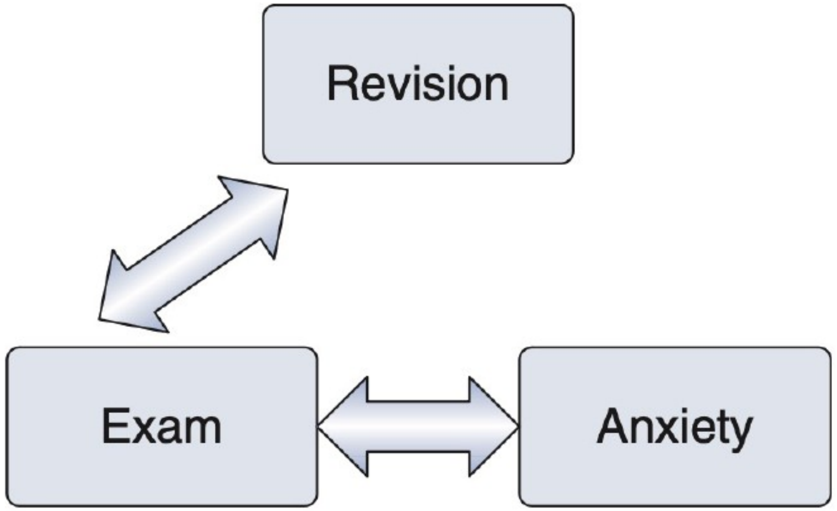
$$sr_1 = \frac{r_{Y1} - r_{Y2}r_{12}}{\sqrt{1-r_{12}^2}}$$


1
Exam Performance
Exam Anxiety
Variance Accounted for by Exam Anxiety (19.4%)
2
Exam Performance
Revision Time
Variance Accounted for by Revision Time (15.7%)
3
Unique Variance Accounted for by Revision Time
Variance Accounted for by both Exam Anxiety and Revision Time
Exam Performance
Revision Time
Unique Variance Accounted for by Exam Anxiety
Exam Anxiety

Partial and Semi-partial Correlations



Partial Correlation



Semi-Partial Correlation

Used for when you think that the confounding variable does not influence both variables.

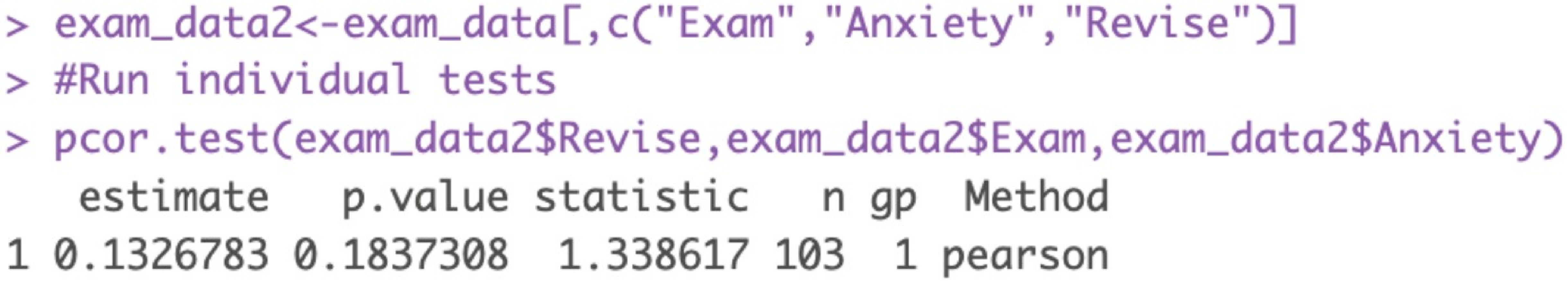
Need more details. Watch this video.

Doing Partial Correlation using R

- Need ‘ppcor’ package
- The general form of *pcor.test()* is:

  pcor.test(x,y,z)

- X and Y are the main variables of interest
- Z is the one you want to control for



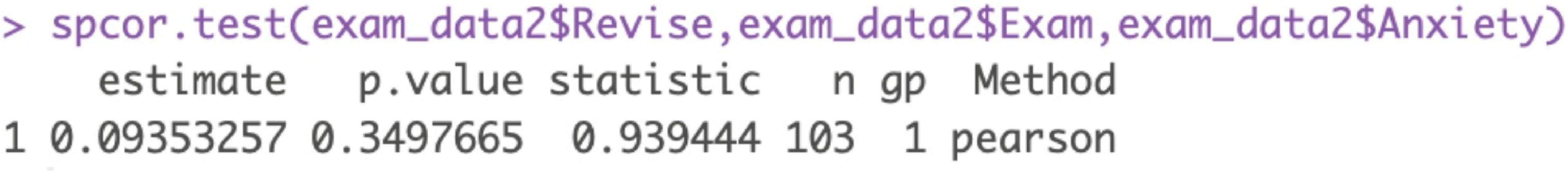
```
> exam_data2<-exam_data[,c("Exam","Anxiety","Revise")]
> #Run individual tests
> pcor.test(exam_data2$Revise,exam_data2$Exam,exam_data2$Anxiety)
   estimate   p.value statistic   n gp  Method
1 0.1326783 0.1837308  1.338617 103  1 pearson
```

Doing Semi-Partial Correlation using R

- The general form of *spcor.test()* is:

  spcor.test(x,y,z)

- X and y are the main variables of interest
- Z is the one you want to control for, and it’s relationship with Y (the second variable entered)

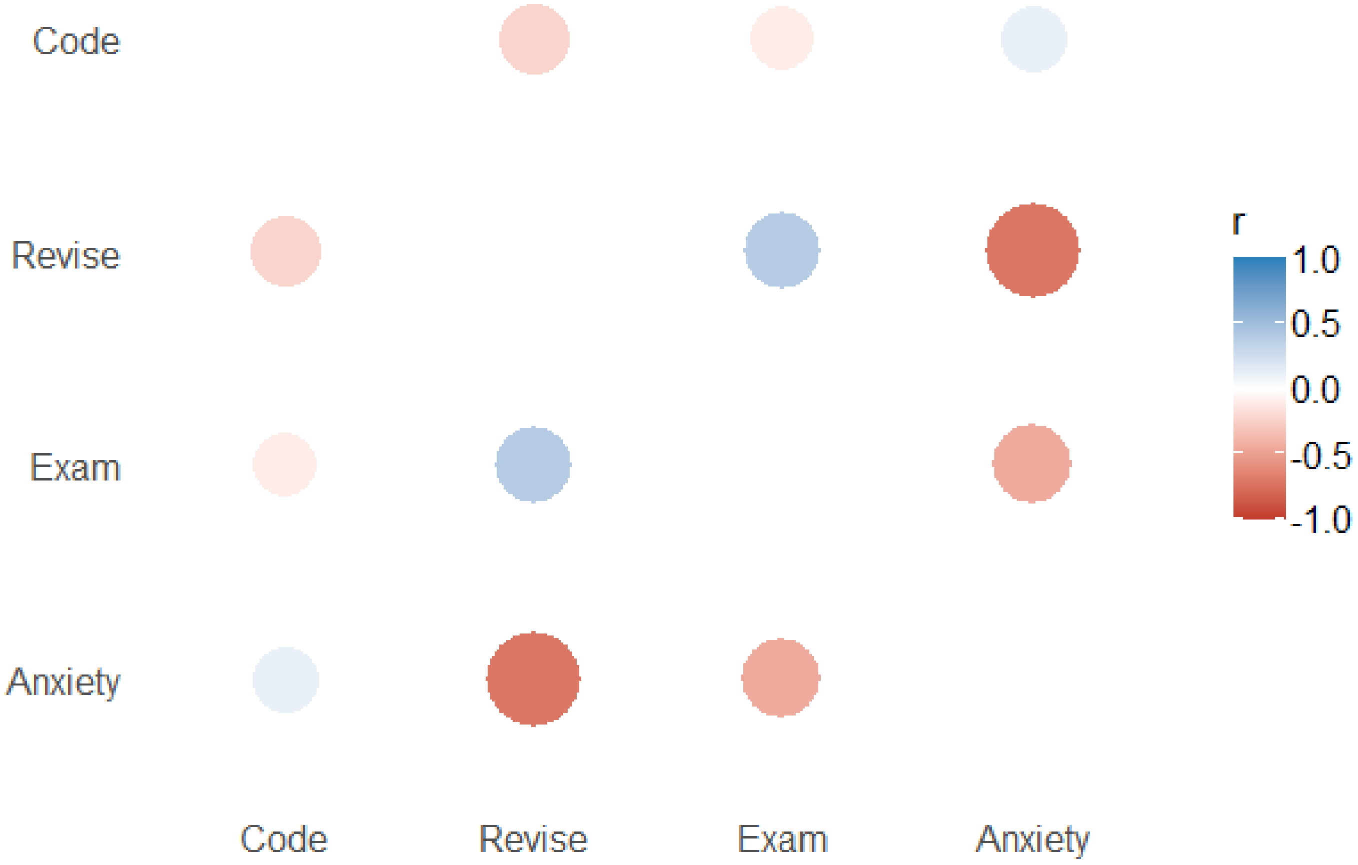


```
> spcor.test(exam_data2$Revise,exam_data2$Exam,exam_data2$Anxiety)
    estimate   p.value statistic   n gp  Method
1 0.09353257 0.3497665  0.939444 103  1 pearson
```

Partial and Semi-Partial Correlation Testing in R

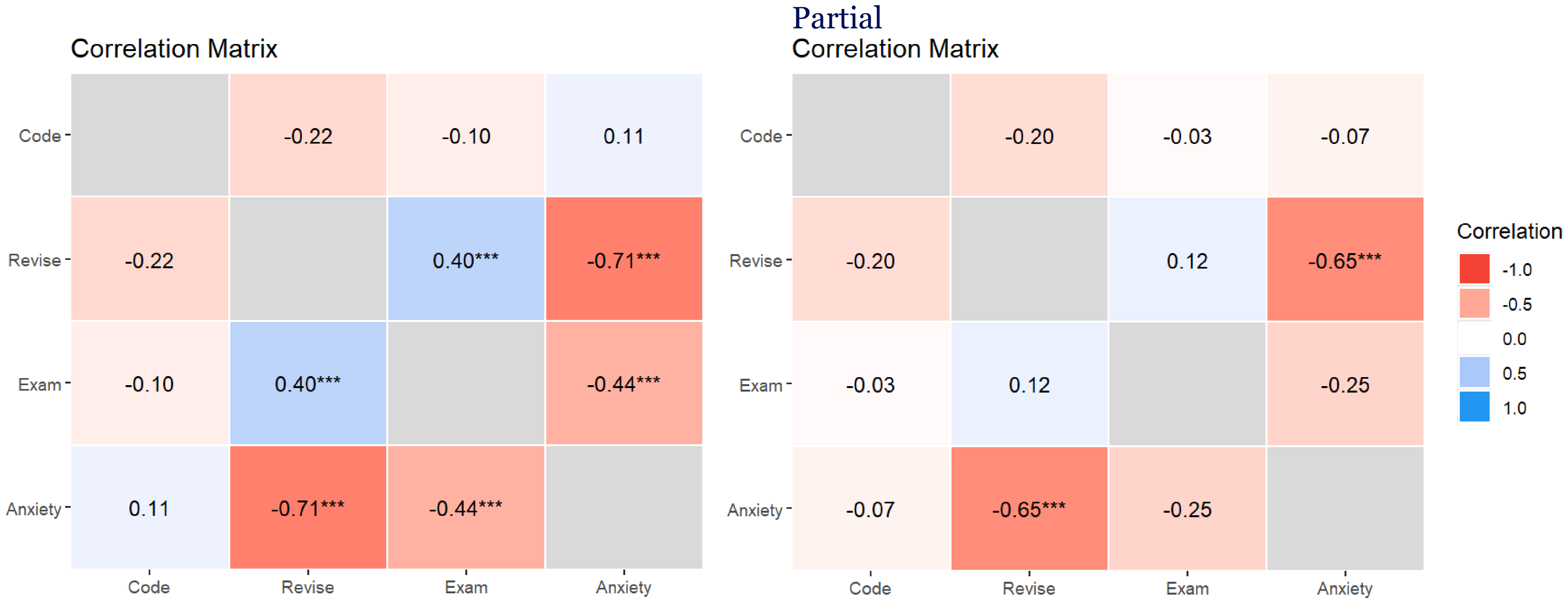
- Load the Exam.Anxiety.csv file into R.
- Install and load the ‘ppcor’ package
- Optional: Create a new data frame that only contains the variables "Exam","Anxiety","Revise“
  - Run partial correlation test to see the relationship between exam score and revision time, while controlling for anxiety (use pcor.test())
  - Run a semi-partial correlation test to see the relationship between exam score and revision time, while controlling for the effect of anxiety on exam score only (use spcor.test())

Easystats: correlation package



Comparing Correlation to Partial Correlation Matrices

- Compare these correlation matrices. What differences do you observe and why do you think that is?



Correlation Matrix

	Code	Revise	Exam	Anxiety
Code		-0.22	-0.10	0.11
Revise	-0.22		0.40***	-0.71***
Exam	-0.10	0.40***		-0.44***
Anxiety	0.11	-0.71***	-0.44***	

Partial
Correlation Matrix

	Code	Revise	Exam	Anxiety
Code		-0.20	-0.03	-0.07
Revise	-0.20		0.12	-0.65***
Exam	-0.03	0.12		-0.25
Anxiety	-0.07	-0.65***	-0.25	

Summing Up

- Correlations
  - Positive, negative and range from -1 to 1
  - Small, medium, large effects
  - Not causal
- Non parametric
  - If data are non-interval/ratio or non normal
- Partial/semi-partial correlations

Preparing for Module 5: Intro to Regression

- Navarro – Ch 15
- Attend Practical Sessions/Complete Exercise

Thanks! See you next week! Questions?

TILBURG UNIVERSITY