Exploring Relationship between Variables - Quiz

Solutions

Question 1

What does the R command *str()* do?

a. It generates a string of random numbers
b. It strips any existing labels from a plot in R
c. It provides information on the structure of an R object
d. It stretches the y-axis on a graph to fit the data displayed

SOLUTION: c.

Question 2

How can you explore relationships between two numerical variables?
(Note: Choose as many answers as you need)

a. Correlation coefficent
b. Scatter plot
c. Box plot
d. Shapiro Wilk test

SOLUTION: a. and b.

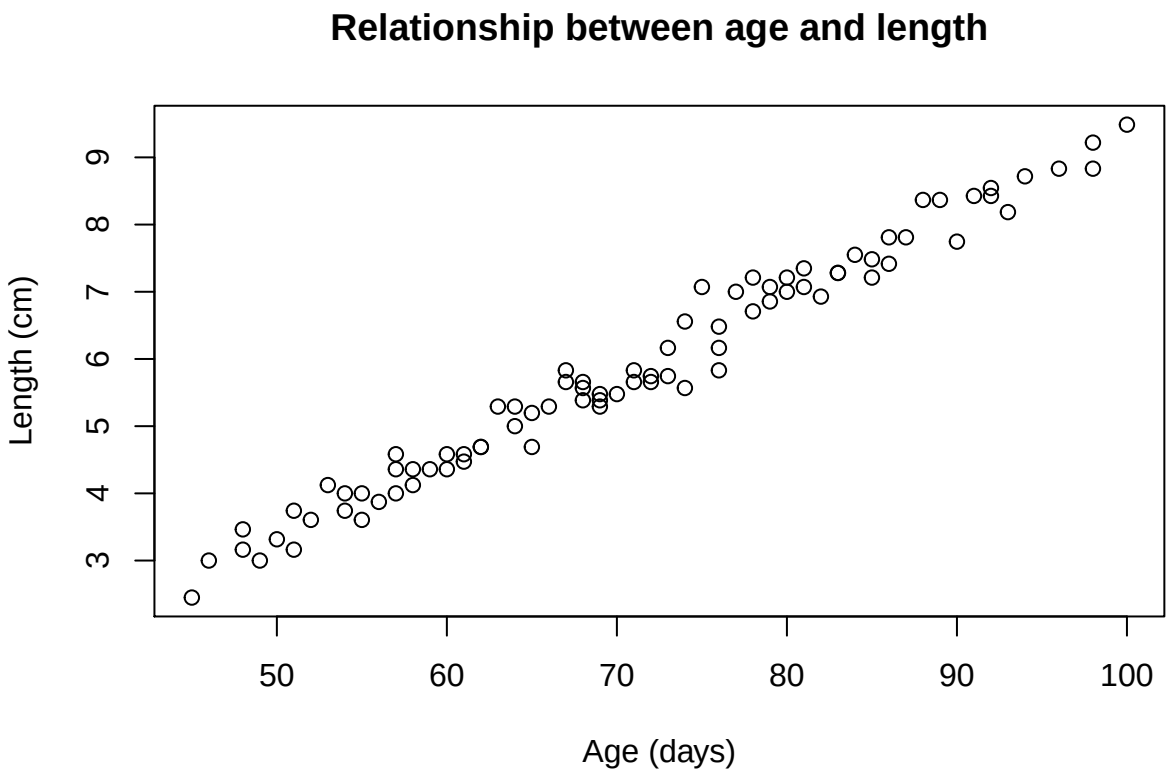
Question 3

What is the difference between the *cor()* and *cor.test()* commands in R?

SOLUTION: The *cor()* command computes only the correlation coefficient, while the *cor.test()* command also computes a hypothesis test for that coefficient. This means that the latter also includes a p-value you can use to judge how signigicant the correlation is.

Question 4

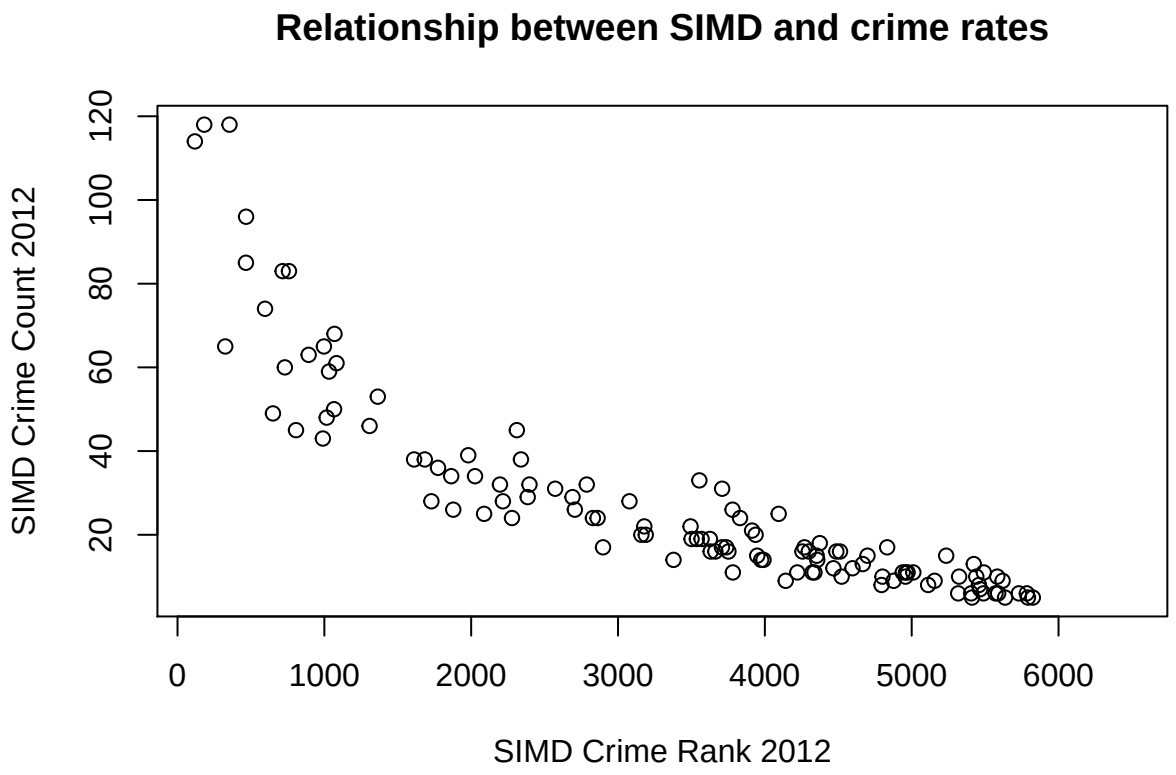
Describe the relationship between the two variable in the following plot:



SOLUTION: From this scatter plot, it looks like there is a strong, positive relationship between the age of a foetus and its length. In other words, as the age of a foetus increases, so does its length. The relationship looks very linear.

Question 5

Describe the relationship between the two variable in the following plot:



SOLUTION: From this scatter plot, it looks like there is a fairly strong, negative relationship between the SIMD crime rank and the SIMD crime count in 2012. In other words, as the crime rank increases, the crime count descreases. However, the relationship does not look entirely linear.

Question 6

Below is the R output for a correlation between precipitation (mm) and cucumber yield (kg/m2). Interpret the correlation coefficient.

```
## [1] 0.8708738
```

SOLUTION: The correlation coefficient between precipitation and cucumber yield is 0.87. This indicates that there is a strong, positive linear relationship between the two variables. As the precipitation increases, so does the cucumber yield.

Question 7

Below is the R output for a correlation between the weight and pulse rate of patients. Interpret the correlation coefficient.

```
## [1] -0.2029782
```

SOLUTION: The correlation coefficient is -0.20. There thus seems to be a fairly weak linear relationship between a patient's weight and their pulse rate (i.e. as a patient's weight increases, their pulse rate decreases or vice versa). However, such a weak relationship might not be of much interest for further research.

Question 8

You have collected some data on cinema ticket sales and car accidents. You run a correlation in R and your correlation coefficient is 0.95. You conclude that watching movies causes car accidents. Is this a correct interpretation of a correlation coefficient? Please also provide a brief explanation for your chosen answer.

- Yes
- No

SOLUTION: The correct answer is no. Correlation does not equal causation and a correlation coefficient does not allow for any conculsions as to the causal relationship between two variables. A better interpretation would thus have been to say that there is a very strong, positive linear relationship between the variables. Another point to note is that this relationship might be due to another variables you have not considered in your study (such as bad weather).

Question 9

After running a correlation in R, you found that there is no linear relationship between the two numerical variables. Should you go ahead and perform a linear regression? Please also provide a brief explanation for your chosen answer.

- Yes
- No

SOLUTION: The correct answer is no. A linear regression assumes that there is a linear relationship between your variables. Therefore, if you already know that this is no the case, you should not perform a linear regression and instead choose a different test that is more suited to your data.

Question 10

What is the code to compute a linear regression in R?
(Note: You may assume that the data set had been attached.)

a. *reg(dependent_variable ~ independent_variable)*
b. *lm(dependent_variable ~ independent_variable)*
c. *linear.reg(dependent_variable ~ independent_variable)*
d. *l.regression(dependent_variable ~ independent_variable)*

SOLUTION: b.

Question 11

Interpret the intercept and coefficient of the following R output from a linear regression of the healing time (in days) of a wound in terms of the wound dimension (in mm):

```
##
## Call:
## lm(formula = wound$time ~ wound$dim)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.340  -6.941  -1.724   6.159  18.416
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.0837     5.5046   1.287    0.207
## wound$dim     0.7063     0.0705  10.017 1.12e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.059 on 34 degrees of freedom
## Multiple R-squared:  0.7469, Adjusted R-squared:  0.7395
## F-statistic: 100.3 on 1 and 34 DF,  p-value: 1.116e-11
```

SOLUTION: The slope is 0.71, so very every one unit increase in wound dimension (in this case, for every mm), the healing time for a wound increases by 0.71 days. The intercept is 7.08, which means that at a wound dimension of 0 mm, the healing time would be 7.08 days. Clearly, this interpretation does not make sense in the given example, as a non-existent would does not require any healing time. This likely due to the fact that the smallest wound in the sample was 30 mm and thus the predication for 0 mm in this particular regression is outwith the range of the data and thus very inaccurate. All in all, this means that here the interpretation of the intercept is not very meaningful.

Question 12

Interpret the intercept and coefficient of the following R output from a linear regression of the length of a foetus (in cm) in terms of age (in days):

```
##
## Call:
## lm(formula = foetus$length ~ foetus$age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65506 -0.18972  0.01104  0.15364  0.72780
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.690881   0.148621  -18.11   <2e-16 ***
## foetus$age   0.120455   0.002055   58.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2655 on 82 degrees of freedom
## Multiple R-squared:  0.9767, Adjusted R-squared:  0.9764
## F-statistic:  3436 on 1 and 82 DF,  p-value: < 2.2e-16
```

SOLUTION: The coefficient for age is 0.12, so for every one unit increase in age (in this case for every day) the lenghth of a foetus increases by 0.12 cm. The intercept is -2.69, which means that a foetus at age 0 would be -2.69 cm long. However, in this particular case interpreting the intercept does not make sense.

Question 13

Interpret the intercept and coefficient of the following R output from a linear regression of the SIMD crime count in terms of the SIMD crime rank:

```
##
## Call:
## lm(formula = crimes$SIMD.Crime.2012.count ~ crimes$SIMD.Crime.2012.rank)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -20.79  -7.75  -0.33   5.51  51.82
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  70.6854678  2.5285067   27.95   <2e-16 ***
## crimes$SIMD.Crime.2012.rank -0.0127240  0.0006698  -19.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.1 on 117 degrees of freedom
##   (23 observations deleted due to missingness)
## Multiple R-squared:  0.7552, Adjusted R-squared:  0.7531
## F-statistic: 360.9 on 1 and 117 DF,  p-value: < 2.2e-16
```

SOLUTION: The coefficient for the SIMD crime rank is -0.013. This means that for every one unit increase in crime rank, the crime count decreases by -0.013. The intercept is 70.69, which means that at crime rank 0 the crime count is 70.69. Interpreting this intercept only makes sense if there is a crime rank 0 on the SIMD.

Question 14

Why does it make sense to use *summary(lm())* rather than just *lm()* when running a linear regression in R?

SOLUTION: Because *lm()* only computes the coefficients, while *summary(lm())* provides you with a lot more information on the regression (including p-values and the coefficent of determination).

Question 15

Amend the following R code to make predictions not only for 100 but also for 200 and 500 without adding an additional line of code:

predict(lm(dependent_variable~independent_variable), newdata=data.frame(independent_variable=100), interval="pred")

SOLUTION:

predict(lm(dependent_variable~independent_variable), newdata=data.frame(independent_variable=c(100, 200, 500)), interval="pred")

Note: This code only works when the data set you are working with is attached.

Question 16

Interpret the following predictions of exam scores (in %) for the average amount of sleep for 7 and 9.5 hours respectively:

```
##        fit      lwr      upr
## 1 51.34277 36.83852 65.84701
## 2 69.66099 55.09294 84.22905
```

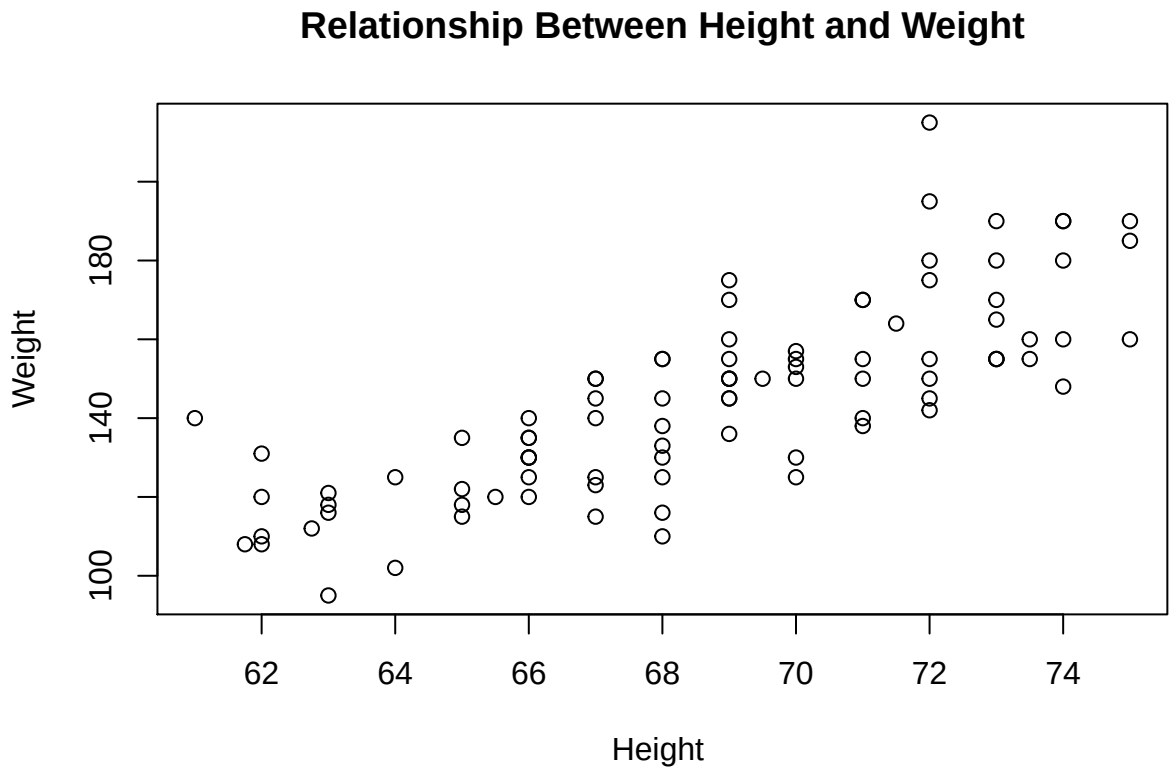
SOLUTION:
The expected exam score for student who slept 7 hours, on average, is 51.34%. The expected exam score for student who slept 9.5 hours, on average, is 69.66%.

Question 17

What R code can be used to add a regression line (in the colour red) to the following plot:

```r
activity <- read.table(file.choose(), header = TRUE, sep = ",")

plot(activity$Height, activity$Weight, main = "Relationship Between Height and Weight",
    xlab = "Height ", ylab = "Weight")
```



SOLUTION:

```r
abline(lm(activity$Weight ~ activity$Height), col="red")
```

Question 18

Interpret the coefficent of determination below for a regression of the lenth of a foetus in terms of age:

```
## [1] 0.976694
```

SOLUTION: The coefficient of determination is 0.977, which means that 97.7% of the variability in the length of a foetus is explained by its age.