大數據統計分析與預測 第十七章作業 (Correlation)

Created by Weber YC, Huang (m946108006) 2019-12-27

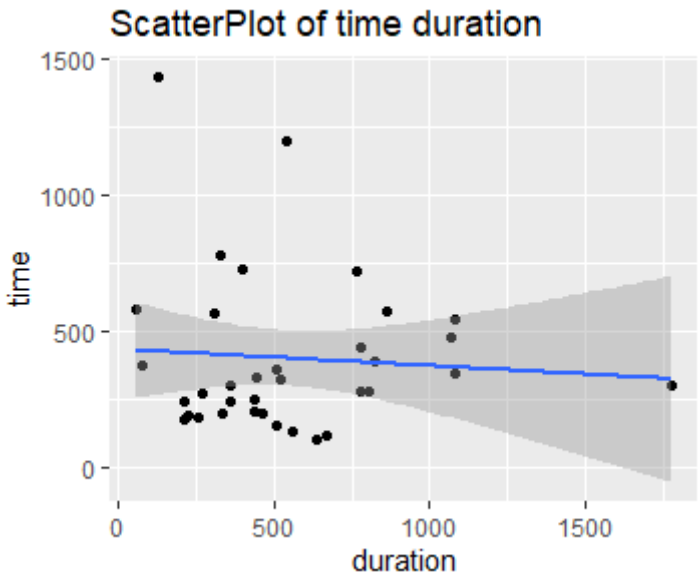
Q6. Thirty-five patients with ischemic heart disease, a suppression of blood flow to the heart, took part in a series of tests designed to evaluate the perception of pain. In one part of the study, the patients exercised until they experienced angina, or chest pain；time until the onset of angina and duration of the attack were recorded. The data are saved on your disk in the file ischemic. Time to angina in seconds is saved under the variable name time； the duration of angina, also in seconds, is saved under the name duration.

(a) Create a two-way scatter plot for these data.

> R-code :

```
library(ggplot2)
# Basic scatter plot
ggplot(df, aes(x=duration, y=time)) + geom_point() + ggtitle('ScatterPlot of time
duration') + stat_smooth(method=lm)
```

> Run code ...



(b) In the population of patients with ischemic heart disease, does there appear to be any evidence of a linear relationship between time to angina and duration of the attack ?

According to the trend line, it doesn't has enough evidences to tell there is any relationship between time to angina and duration of the attack. But still, it slightly shows negative relationship...

(c) Calculate Pearson's correlation coefficient.

> R-code :

```
co <- function(x ,y){
  method <- as.character(readline(prompt="What correlation method do ya wanna
perform? (pearson, kendall, spearman): "))
  cat(' === ', method, ' === ', '\n')
  cor(x, y, method = method)
}
co(df$time, df$duration)
```

> Run code ...

```
What correlation method do ya wanna perform? (pearson, kendall, spearman):
pearson

===  pearson  ===
   -0.07372094
```

(d) Does the duration of angina tend to increase or decrease as time to angina increases?

According to the Pearson's correlation coefficient, the time to angina increase, the duration of angina tend to decrease slightly.

(e) Test the null hypothesis ($H_0 : \rho = 0$) what do you conclude?

> R-code :

```
co2 <- function(x ,y){
  method <- as.character(readline(prompt="What correlation method do ya wanna
perform? (pearson, kendall, spearman): "))
  cat(' === ', method, ' === ', '\n')
  cor.test(x ,y , method = method)
}
co2(df$time, df$duration)
```

> Run code ...

```
What correlation method do ya wanna perform? (pearson, kendall, spearman):
pearson

===  pearson  ===

    Pearson product-moment correlation

data:  df$duration and df$time
t = -0.42465, df = 33, p-value = 0.6738
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3972091  0.2660620
sample estimates:
        cor
-0.07372094
```

According to the P-value, we do not reject the H_0, and take $\rho = 0$.

(f) Compute spearman's rank correlation.

> R-code:

```
co(df$time, df$duration)
```

> Run code ...

```
What correlation method do ya wanna perform? (pearson, kendall, spearman):
spearman

 ===  spearman  ===
     0.07578092
```

(g) Using r_s again test the null hypothesis that the population correlation is equal to zero. What do you conclude?

> R-code :

```
co2(df$time, df$duration)
```

> Run code ...

```
 ===  spearman  ===

    Spearman rank correlation rho

data:  df$duration and df$time
S = 6598.9, p-value = 0.6653
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.07578092

Warning message:
In cor.test.default(df$duration, df$time, method = method) :
  Cannot compute exact p-value with ties
```

Since the P-value is 0.6653, we do not reject the null hypothesis and take rho is equal to 0.

Q9. One of the functions of the Federation of State Medical Boards is to collect data summarizing disciplinary actions taken against non-federal physicians by medical licensing boards. Serious actions include license revocations, suspensions and probations for each of the years 1991 through 1995, the number of serious actions per 1000 s doctors was ranked by state from highest to lowest. The ranks are contained in a data set called actions. The ranks for 1991 are saved under the variable named rank91 those for 1992 under rank92, and so on.

(a) Which states have the highest rates of serious actions in which of five years 1991 through 1995? Which states have a lowest rates?

R-code :

```
max_min_row <- function(df) {
  cat('===', 'highest', '===', '\n')
  for (i in 2:ncol(df)) {
    print(df[which.min(df[,i]),])
  }
  cat('\n', '===', 'lowest', '===', '\n')
  for (i in 2:ncol(df)) {
    print(df[which.max(df[,i]),])
  }
}
max_min_row(df)
```

Run code ...

```
=== highest ===
   state rank91 rank92 rank93 rank94 rank95
2 Alaska      1      7      8      2      8
      state rank91 rank92 rank93 rank94 rank95
37 Oklahoma      2      1      2      5     12
           state rank91 rank92 rank93 rank94 rank95
49 West_Virginia      8      3      1      6      7
     state rank91 rank92 rank93 rank94 rank95
51 Wyoming      9      4     21      1      3
         state rank91 rank92 rank93 rank94 rank95
25 Mississippi      6      6      9      9      1

 === lowest ===
          state rank91 rank92 rank93 rank94 rank95
40 Rhode_Island     51     41     42     26     26
     state rank91 rank92 rank93 rank94 rank95
8 Delaware     16     51     43     48     48
                state rank91 rank92 rank93 rank94 rank95
9 District_of_Columbia     45     45     51     51     50
                state rank91 rank92 rank93 rank94 rank95
9 District_of_Columbia     45     45     51     51     50
    state rank91 rank92 rank93 rank94 rank95
12 Hawaii     41     50     46     50     51
```

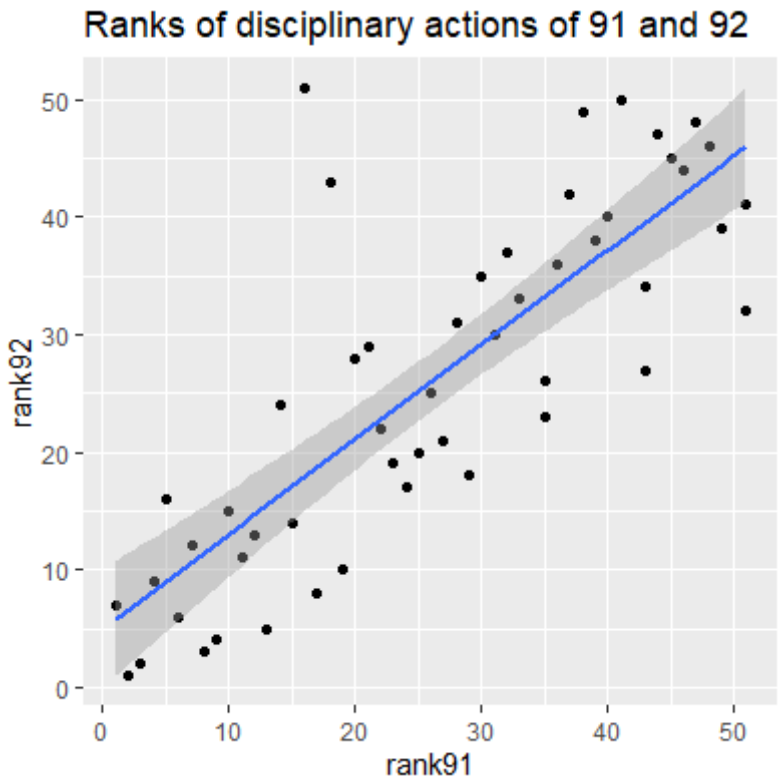
91-95	highest	lowest
91	Alaska	Rhode_Island
92	Oklahoma	Delaware
93	West_Virginia	District_of_Columbia
94	Wyoming	District_of_Columbia
95	Mississippi	Hawaii

(b) Construct a two-way scatter plot for the ranks of disciplinary actions in 1992 versus the ranks in 1991.

R-code :

```
library(ggplot2)
# Basic scatter plot
ggplot(df, aes(x=rank91, y=rank92)) + geom_point() + ggtitle('Ranks of
disciplinary actions of 91 and 92') + stat_smooth(method=lm)
```

Run code ...



(c) Does there appear to be a relationship between these two quantities?

According to the scatter plot of rank91 and rank92, it shows that there is a strong positive relationship between two quantities. The value of rank92 increase as rank91 increase.

(d) Calculate the correlation of the two sets of ranks.

R-code :

```
co(df$rank92, df$rank91)
```

Run code ...

```
What correlation method do ya wanna perform? (pearson, kendall, spearman):
spearman
 ===  spearman  ===
     0.810236
```

(e) Is this correlation significantly different from zero? What do you conclude?

R-code :

```
co2(df$rank92, df$rank91)
```

> Run code ...

```
What correlation method do ya wanna perform? (pearson, kendall, spearman):
spearman
 ===  spearman  ===

    Spearman rank correlation rho

data:  x and y
S = 4193.8, p-value = 5.921e-13
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.810236

Warning message:
In cor.test.default(x, y, method = method) :
  Cannot compute exact p-value with ties
```

As the $P\,value < 0.05$, the correlation is significantly different from zero. We conclude that the state's ranking in serious disciplinary actions in 1992 tends to be higher if it also has a higher ranking in 1991.

(f) Calculate the correlations of the ranks in 1991 and those in 1993； those in 1991 and 1994； and those in 1991 and 1995. What happens to the magnitude of the correlation as years being compared get further apart?

> R-code :

```
co(df$rank91, df$rank93)
co(df$rank91, df$rank94)
co(df$rank91, df$rank95)
```

> Run code ...

```
=== 91 & 93 ===
What correlation method do ya wanna perform? (pearson, kendall, spearman):
spearman
 ===  spearman  ===
     0.764272

=== 91 & 94 ===
What correlation method do ya wanna perform? (pearson, kendall, spearman):
spearman
 ===  spearman  ===
     0.6292114

=== 91 & 95 ===
What correlation method do ya wanna perform? (pearson, kendall, spearman):
spearman
 ===  spearman  ===
     0.5735813
```

As we can see the ρ decrease as the year get further apart.

(g) Is each of these three correlations significantly different from zero?

> R-code :

```
co2(df$rank91, df$rank93)
co2(df$rank91, df$rank94)
co2(df$rank91, df$rank95)
```

> Run code ...

```
=== 91 & 93 ===
What correlation method do ya wanna perform? (pearson, kendall, spearman):
spearman
 ===  spearman  ===

    Spearman rank correlation rho

data:  x and y
S = 5209.6, p-value = 6.768e-11
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.764272

=== 91 & 94 ===
What correlation method do ya wanna perform? (pearson, kendall, spearman):
spearman
===  spearman  ===

    Spearman rank correlation rho

data:  x and y
S = 8194.4, p-value = 7.598e-07
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.6292114

=== 91 & 95 ===
What correlation method do ya wanna perform? (pearson, kendall, spearman):
spearman
===  spearman  ===

    Spearman rank correlation rho

data:  x and y
S = 9423.9, p-value = 1.084e-05
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.5735813
```

Since all $P\ value < 0.05$, so each of these three correlations significantly different from zero.

(h) Do you believe that all states are equally strict in taking disciplinary action against physicians?

These data suggest all states are not strict in taking disciplinary actions against physicians；some are more strict than others.

Extra1. Discussion the association between waist and SBP in by correlation coefficient and linear regression model

R-code :

```
# Pearson's correlation coefficient
co2(df_1$Waist, df_1$SBP)
# Simple linear regression
linearMod <- lm(SBP ~ Waist, data=df_1)
summary(linearMod)
```

> Run code ...

```
What correlation method do ya wanna perform? (pearson, kendall, spearman):
pearson
 ===  pearson  ===

    Pearson product-moment correlation

data:  x and y
t = 117.66, df = 62381, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4197215 0.4325658
sample estimates:
      cor
0.4261651

 === linear model ===
Call:
lm(formula = SBP ~ Waist, data = df_1)

Residuals:
     Min       1Q   Median       3Q      Max
-112.092  -12.824   -2.279   10.344  147.215

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 58.140635   0.558520   104.1   <2e-16 ***
Waist        0.831113   0.007064   117.7   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.82 on 62381 degrees of freedom
```

```
  (2106 observations deleted due to missingness)
Multiple R-squared:  0.1816,    Adjusted R-squared:  0.1816
F-statistic: 1.384e+04 on 1 and 62381 DF,  p-value: < 2.2e-16
```

Extra2. propose your opinions about the similarities and differences between the models of the Pearson correlation coefficient and the simple linear regression model.

From the models we may see that both two p-value are in the same. Theoretically,

similarities	differences
the standardized regression coefficient is the same as Pearson's correlation coefficient.	The regression equation (i.e., a+bXa+bX) can be used to make predictions on YY based on values of XX
The square of Pearson's correlation coefficient is the same as the R2R2 in simple linear regression.	While correlation typically refers to the linear relationship, it can refer to other forms of dependence, such as polynomial or truly nonlinear relationships
Neither simple linear regression nor correlation answer questions of causality directly. This point is important, because I've met people that think that simple regression can magically allow an inference that XX causes YY.	While correlation typically refers to Pearson's correlation coefficient, there are other types of correlation, such as Spearman's.