M.Sc. in ‚Transportation Systems'

Applied Statistics in Transport

Exercises: Correlation, Regression, ANOVA

1. The university in A-city has collected the following data on the intelligence quotient (IQ, X_i) and the weekly hours of watching TV (Y_i) for a sample of 10 persons. Please analyse the relationship between those two variables. Give reasons for the measure you have chosen and interpret your results.

IQ, X_i	Hours of TV per Week, Y_i
106	7
86	0
100	27
101	50
99	28
103	29
97	20
113	12
112	6
110	17

Solution:

This relationship can be analysed with the help of a correlation coefficient.

We can try to apply Pearson (analysing the linear relationship) and Spearman (analysing the relationship between the rankings). Spearmen seems to be more suitable (more robust, needs less assumption about the distributions of x and y etc.)

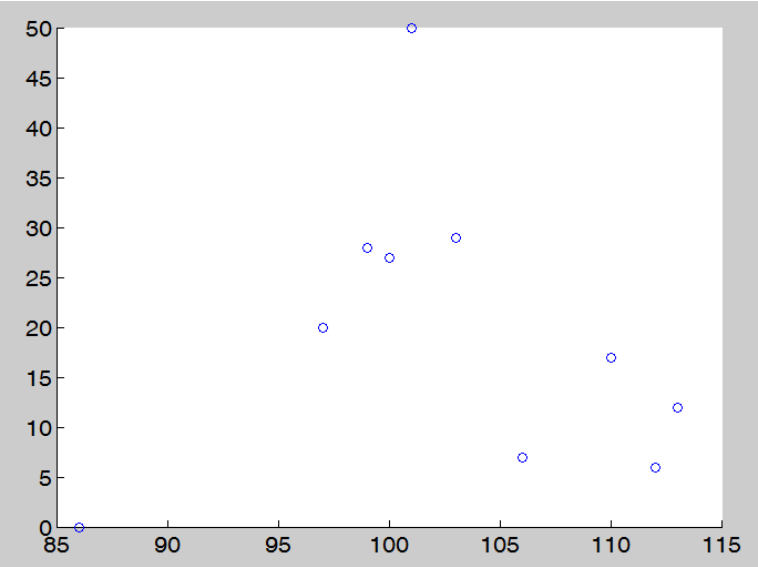
IQ, X_i	Hours of TV per Week, Y_i	Rank X_i	Rank Y_i	d_i	D_i^2
86	0	1	1	0	0
97	20	2	6	-4	16
99	28	3	8	-5	25
100	27	4	7	-3	9
101	50	5	10	-5	25
103	29	6	9	-3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

There are no ties, we can apply the simplified formula:

$$r_{SP} = 1 - \frac{6\sum d_i^2}{(n^2-1)n} = 1 - \frac{6*194}{(100-1)10} = -0.1758$$

R:

```
iq<-c(86,97,99,100,101,103,106,110,112,113)
tv<-c(0,20,28,27,50,29,7,17,6,12)
cor(iq,tv,method="spearman") #gives -0.1757576
cor(iq,tv,method="pearson") #gives -0.03760147
plot(iq,tv)
```



There seems to be no linear relationship but some relationship between the rankings (Spearman). However, both values are close to zero, so the relationship is very weak.

2. The originating traffic of a region depends on several variables. Please check whether the originating traffic can be explained with the help of simple linear regression as a function of the number of registered passenger cars.

   X number of registered passenger cars [1,000 cars]

   Y originating traffic [1,000 veh./16h]

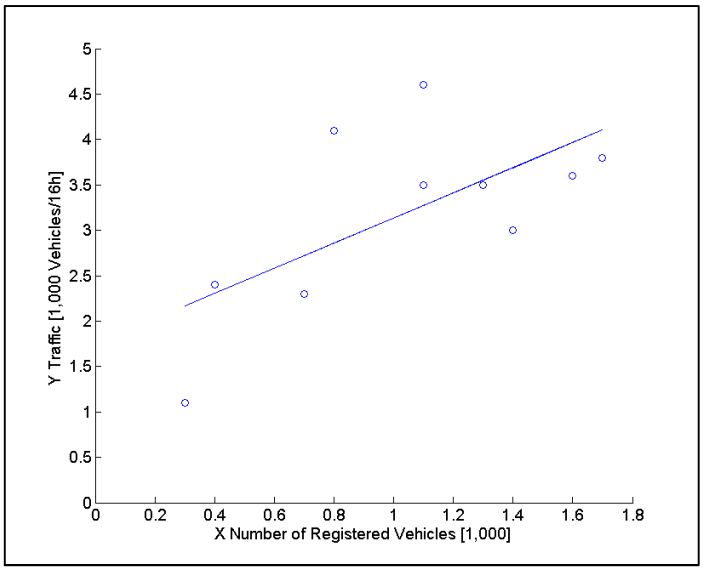
   The following table shows the observations for 10 regions:

i	x_i	y_i
1	0.3	1.1
2	0.4	2.4
3	1.1	3.5
4	0.8	4.1
5	0.7	2.3
6	1.6	3.6
7	1.1	4.6
8	1.3	3.5
9	1.7	3.8
10	1.4	3
Total	10.4	31.9

Determine the correlation coefficient; compute the coefficients of the regression line and the coefficient of determination.

Solution:

i	x_i	y_i	x_i^2	y_i^2	x_i*y_i
1	0.3	1.1	0.09	1.21	0.33
2	0.4	2.4	0.16	5.76	0.96
3	1.1	3.5	1.21	12.25	3.85
4	0.8	4.1	0.64	16.81	3.28
5	0.7	2.3	0.49	5.29	1.61
6	1.6	3.6	2.56	12.96	5.76
7	1.1	4.6	1.21	21.16	5.06
8	1.3	3.5	1.69	12.25	4.55
9	1.7	3.8	2.89	14.44	6.46
10	1.4	3	1.96	9	4.2
Total	10.4	31.9	12.9	111.13	36.06



Computing the coefficients:

$$S_{XX} = \sum x_i^2 f_i - \frac{(\sum x_i f_i)^2}{n} = 12.9 - \frac{10.4^2}{10} = 2.0840,$$

$$S_{YY} = \sum y_j^2 f_j - \frac{(\sum y_j f_j)^2}{n} = 111.13 - \frac{31.9^2}{10} = 9.3690,$$

$$S_{XY} = \sum x_i y_j f_{ij} - \frac{(\sum x_i f_i)(\sum y_j f_j)}{n} = 36.06 - \frac{10.4*31.9}{10} = 2.8840$$

$$s_X{}^2 = \frac{1}{n-1} * S_{XX} = \frac{2.0840}{9} = 0.2316, s_X = 0.4812\ ,$$

$$s_Y{}^2 = \frac{1}{n-1} * S_{YY} = \frac{9.3690}{9} = 1.0410, s_Y = 1.0203,$$

$$s_{XY} = \frac{1}{n-1} * S_{XY} = \frac{2.8840}{9} = 0.3204$$

Correlation Coefficient

$$r = \frac{S_{XY}}{\sqrt{S_{XX} * S_{YY}}} = \frac{2.8840}{\sqrt{2.0840*9.3690}} = 0.6527,\ r = \frac{s_{xy}}{s_x * s_y} = \frac{0.3204}{0.4812*1.0203} = 0.6526$$

Since r>0 there is a positive linear relationship between x and y, that means with increasing number of registered passenger cars also the originating traffic increases.

However the relationship is not too strong, as r is only 0.65 (it could be 1).

Computing the coefficients:

$$\hat{y} = a_y + b_y * x$$

$$\overline{x} = \frac{10.4}{10} = 1.0400,\ \overline{y} = \frac{31.9}{10} = 3.1900$$

$$b_y = \frac{S_{XY}}{S_{XX}} = \frac{2.8840}{2.0840} = 1.3839,\ a_y = \overline{y} - b_y * \overline{x} = 3.1900\ - 1.3839 * 1.0400 = 1.7507$$

$$\hat{x} = a_x + b_x * y$$

$$b_x = \frac{S_{XY}}{S_{YY}} = \frac{2.8840}{9.3690} = 0.3078,\ a_x = \overline{x} - b_x * \overline{y} = 1.0400 - 0.3078 * 3.1900 = 0.0581$$

We get the following regression lines:

$$\hat{y} = a_y + b_y * x = 1.7507 + 1.3839 * x$$

$$\hat{x} = a_x + b_x * y = 0.0581 + 0.3078 * y$$

The coefficient of determination

for simple linear regression: $R^2 = b_X * b_Y = 0.3078 * 1.3839 = 0.4260$

$R^2 = r^2 = \frac{S_{XY}{}^2}{S_{XX} * S_{YY}} = \frac{2.8840^{\,2}}{2.0840 * 9.3690} = 0.4260$

$R^2 = r^2 = 0.6527^2 = 0.4260$

42.6 percent of the (squared) deviations of the observed values from the mean can be explained by the regression line.

The standard error (not relevant for the exam!)

What is the standard error for the deviations of the observed values and the regression line: $\hat{y} = a_Y + b_Y * x$?

$S(y - \hat{y})^2 = S_{YY} - b_Y * S_{XY} = 9.3690 - 1.3839 * 2.8840 = 5.3778,$

$s^2_{Y.X} = \frac{1}{n-2} S(y - \hat{y})^2 = \frac{5.3778}{8} = 0.6722,\ s_{Y.X} = \sqrt{0.6722} = 0.8199$

In R: (see Ex_AppliedStatisticsInferential20101115.r)

```
x<-c(0.3,0.4,1.1,0.8,0.7,1.6,1.1,1.3,1.7,1.4)
sum(x) #gives 10.4
y<-c(1.1,2.4,3.5,4.1,2.3,3.6,4.6,3.5,3.8,3)
sum(y) #gives 31.9
cor(x,y) #gives 0.6526792
(cor(y,x))^2 #coefficient of determination: 0.4259902

plot(y~x)
model<-lm(y~x)
summary(model) #y=1.7507+1.3839*x

plot(x~y)
model<-lm(x~y)
summary(model) #x=0.0581+0.3078*y
```