Lab week 7

Correlation coefficient

In a laboratory containing polarographic equipment six samples of dust were taken at various distances from the polarograph and the mercury content of each sample was determined. The following results were obtained.

Distance from polarograph,m	1.4	3.8	7.5	10.2	11.7	15.0
Mercury concentration, ng/g	2.4	2.5	1.3	1.3	0.7	1.2

Examine the possibility that the mercury contamination arose from the polarograph.

First produce a graph of the data representing the dependence of mercury concentration(Y) on distance from polarograph (X). Create the plot of the mercury concentration against the distance.

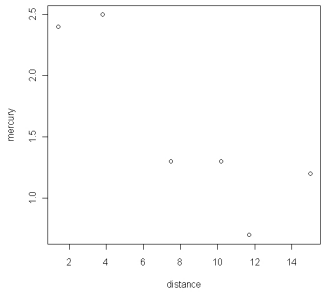
> distance <− c(1.4,3.8,7.5,10.2,11.7,15.0)

> mercury <− c(2.4,2.5,1.3,1.3,0.7,1.2)

> plot(distance, mercury)

As expected the plot indicates a negative relationship between the distance and and the mercury concentration. When the distance increases the mercury content from the machine decreases.

Examining the strength of linear dependence by calculating the correlation coefficient r:



> cor(distance, mercury)

> [1] -0.8569411

The sign of the correlation coefficient is negative indicating a negative relationship and the value of -0.8569411 is close to -1 suggesting a strong negative relationship. It it necessary to use a proper statistical test to see whether the correlation coefficient is statistically significant. Test the hypothesis that

H_0: the correlation coefficient=0

H_1: the correlation coefficient $\neq$ 0. The method of doing this involves calculating the test statistic $t = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}$, where r is the correlation coefficient and n is the number of points used in the calculation of r.

>n < − length(distance) calculates the size of the sample

>r < − cor(distance, mercury) stores the correlation coefficient

> $test.statistic < -abs(r) \star sqrt(n-2)/sqrt(1-r \star r)$

[1] 3.325252

We can compare the test statistic with the critical value obtained using the t-distribution with $n-2$ degrees of freedom

> $qt(0.975, n-2)$

[1] 2.776445

The test statistic 3.325252 is greater than the critical value of 2.776445 hence we reject the null hypothesis that states $r = 0$ and implies there is no correlation between distance and mercury concentration. We accept the alternative hypothesis that r$\neq$ 0, hence the relationship between distance and mercury concentration is significant. Testing r for significance can be done with

> $cor.test(distance, mercury)$

Pearson's product-moment correlation

data: distance and mercury

t = -3.3253, df = 4, p-value = 0.02923

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.9841000 -0.1490113

sample estimates:

cor

-0.8569411

The same conclusion can be obtained by comparing the p-value 0.02923 with 0.05. The p-value is less than the significance level of 0.05, hence we reject the null hypothesis and accept the alternative. There is a significant linear relation between the two variables.

Calibration: correlation and regression

The response of a colorimetric test for glucose was checked with the aid of standard glucose solutions. Determine the relationship between glucose concentration and absorbance from the following data.

Concentration	0	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30
Absorbance	0.03	0.16	0.23	0.43	0.59	0.73	0.87	1.05	1.14	1.22	1.46	1.55	1.73	1.81	1.92	2.11

First produce the calibration plot representing the dependence of absorbance(Y) on the glucose concentration(X). Create the plot of absorbance against the glucose concentration.

> x <− c(0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30)

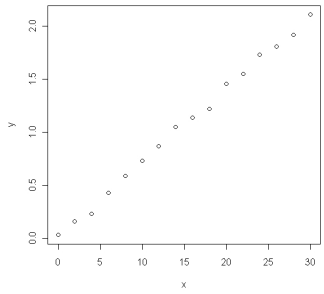
> y <− c(0.03, 0.16, 0.23, 0.43, 0.59, 0.73, 0.87, 1.05, 1.14, 1.22, 1.46, 1.55, 1.73, 1.81, 1.92, 2.11)

> plot(x, y)

As expected the plot indicates a negative relationship between the distance and and the mercury

concentration. When the glucose concentration increases so does the absorbency.

The plotted points are very close to a straight line suggesting a near-perfect linear relationship



between the data. The relationship is strong positive linear. This is confirmed by the correlation coefficient.

> cor(x, y)

> [1] 0.998638

The sign of the correlation coefficient is positive indicating a positive relationship and the value of 0.998638 is close to +1 suggesting a very strong positive linear relationship. Testing r for significance, i.e testing:

H_0: the correlation coefficient=0

H_1: the correlation coefficient $\neq 0$

leads to rejecting the null hypothesis.

> cor.test(x,y)

Pearson's product-moment correlation

data: x and y

t = 71.6185, df = 14, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.9959660 0.9995406

sample estimates:

cor

0.998638

The p-value of 2.2e-16 is very small, less than the significance level of 0.05, hence we reject the null hypothesis and accept the alternative. There is a significant linear relation between the two variables.

The next step is to determine the slope and intercept of the calibration plot, and their confidence limits. Using the *R* function *summary* on results of linear regression analysis we obtain all the necessary information to get the estimates and their confidence limits.

$> model < -lm(y \sim x)$

$> summary(model)$

Call:

lm(formula = y ∼ x)

Residuals:

Min	1Q	Median	3Q	Max
-0.068662	-0.010934	0.003743	0.013871	0.055235

Coefficients:

	Estimate	Std. Error	t value	Pr(> |t|)	
(Intercept)	0.020221	0.017113	1.182	0.257	
x	0.069610	0.000972	71.618	<2e-16	***

We read the slope b = 0.07 and intercept a = 0.02 and we can express the relationship between absorbance and glucose concentration as:

y= 0.02+0.07 · x

absorbance = 0.02+0.07 concentration

We can extract the confidence intervals for the two estimates using the **confint** function.

```
> confint(model)
                2.5 %       97.5 %
(Intercept) -0.01648341  0.05692458
          x  0.06752565  0.07169494
```

The 95% CI for the intercept is $[-0.01648341, 0.05692458]$ and includes zero. This makes the intercept statistically not significant.

The 95% CI for the slope is $[0.06752565, 0.07169494]$ and it does not include zero. This makes the slope statistically significant.