Example on p. 336

The objective is to identify the correlation between mechanic performance in the morning (x) and mechanic performance in the late afternoon.

x	y	x^2	y^2	xy
11.1	10.9	123.21	118.81	120.99
10.3	14.2	106.09	201.64	146.26
12.0	13.8	144.0	190.44	165.60
15.1	21.5	228.01	462.25	324.65
13.7	13.2	187.69	174.24	180.84
18.5	21.1	342.25	445.21	390.35
17.3	16.4	299.29	268.96	283.72
14.2	19.3	201.64	372.49	274.06
14.8	17.4	219.04	302.76	257.52
15.3	19.0	234.09	361.00	290.70

$\Sigma x = 142.3$ $\Sigma y = 166.8$ $\Sigma x^2 = 2085.31$ $\Sigma xy = 2434.69$

$\bar{x} = 14.23$ $\bar{y} = 16.68$ $\Sigma y^2 = 2897.80$

$n = 10$

From the Pearson formula, $r = \dfrac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$

$$= \frac{10(2434.69) - (142.3)(166.8)}{\sqrt{10(2085.31) - (142.3)^2}\sqrt{10(2897.80) - (166.8)^2}} = 0.732$$

So, YES! we have a reasonably strong correlation!

Example on p. 336

We earlier showed a strong (reasonably) correlation between x and y. Let us now characterize the relationship using linear regression.

So wish to establish a linear relationship (regression line)

$$\hat{y} = a + bx$$

Using the Least Squares estimators;

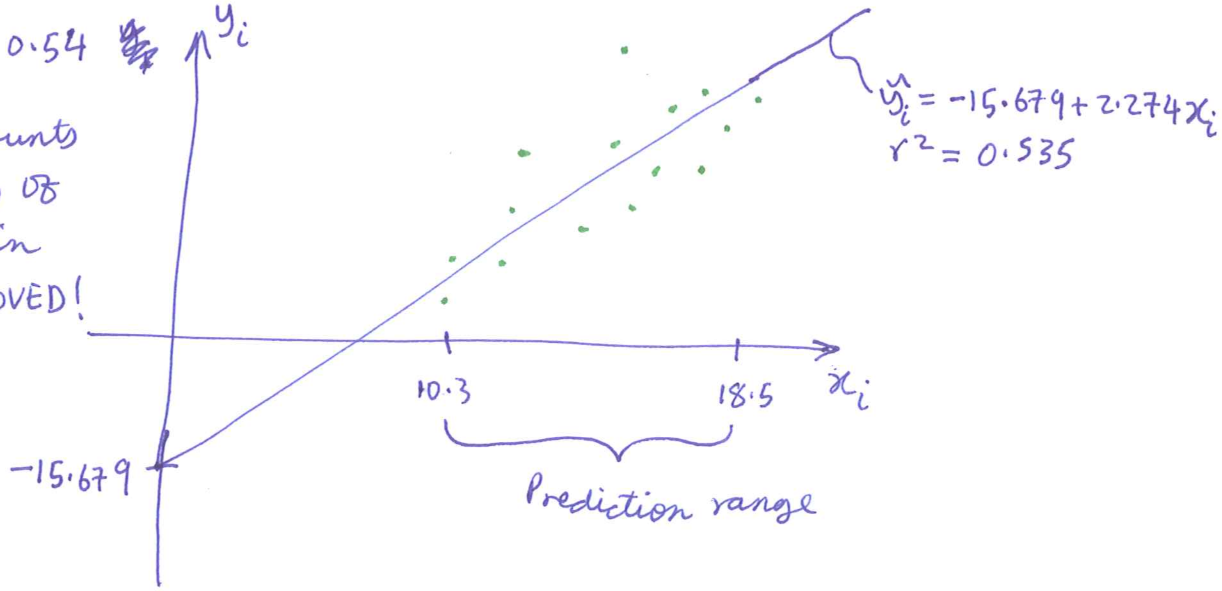
$$b = \frac{(\sum y)(\sum x^2)^{\cancel{2}} - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2} = \frac{(166.8)\overset{(2085.31)}{\cancel{(142.3)^2}} - (142.3)(2434.69)}{10(2085.31) - (142.3)^2} = 2.274$$

$$a = \bar{y} - b\bar{x} = 16.68 - 2.274(14.23) = -15.679$$

So $\hat{y}_i = -15.679 + 2.274x_i$

$r^2 = 0.732^2 = 0.54$ ~~$\%$~~

Our model accounts for the majority of the variability in our data. APPROVED!



Example on p. 336 – Computer

Repeat the problem, this time use software (Excel). In addition compute and tabulate the residuals. Check the adequacy of your linear model by plotting e_i versus $\hat{y}_i$ [Start by drawing a scatter plot of your data]

Interpretation of r^2

Some references use R^2 or R-squared. This value indicates the proportion of the variability in your data that is accounted for by the regression model. By simple majority $r^2 \geq 50\%$ means the model passes (subject to model adequacy check), otherwise it fails. However many professions and fields of specialty go beyond this minimum requirement. In my field we go by 75%. So the problem we just worked on with $R^2 = 0.535$, meets theoretical muster, but if it was a real problem in my job, it will be rejected, plain and simple.

Predictions

Your regression model (if it passes all checks and requirements) Can be used to make predictions within the range of your data. Predictions beyond your data range are NOT VALID !!