Math 207
Introduction to Statistics

Regression Example

x	y	z_x	z_y	$z_x\, z_y$	y values predicted by regression	regression error
1	10	$-3/2$	$1/2$	$-3/4$	7.85	2.15
3	9	$-1/2$	$1/4$	$-1/8$	7.95	1.05
4	8	0	0	0	8.00	0.00
4	7	0	$-1/4$	0	8.00	-1.00
5	1	$1/2$	$-7/4$	$-7/8$	8.05	-7.05
7	13	$3/2$	$5/4$	$15/8$	8.15	4.85

- Compute the necessary statistics.

$$\text{mean}(x) = \frac{1}{6}\,(1+3+4+4+5+7) = 4$$

$$\text{sd}(x) = \sqrt{\frac{(1-4)^2+(3-4)^2+(4-4)^2+(4-4)^2+(5-4)^2+(7-4)^2}{5}} = \sqrt{\frac{9+1+0+0+1+9}{5}}$$

$$= \sqrt{\frac{20}{5}} = \sqrt{4} = 2$$

$$\text{mean}(y) = \frac{1}{6}\,(10+7+9+8+1+13) = 8$$

$$\text{sd}(y) = \sqrt{\frac{(10-8)^2+(7-8)^2+(9-8)^2+(8-8)^2+(1-8)^2+(13-8)^2}{5}} = \sqrt{\frac{4+1+1+0+49+25}{5}}$$

$$= \sqrt{\frac{80}{5}} = 4$$

- Convert the x values to standard units. For example,

$$x=1 \quad \text{becomes} \quad z_x \frac{1-4}{2} = -3/2 \qquad \text{and} \qquad x=3 \quad \text{becomes} \quad z_x \frac{3-4}{2} = -1/2$$

- Convert the y values to standard units: For example,

$$y=10 \quad \text{becomes} \quad z_y = \frac{10-8}{4} = 1/2 \qquad \text{and} \qquad y=7 \quad \text{becomes} \quad z_y \frac{7-8}{4} = -1/4$$

- Compute the products $z_x\, z_y$.
- Compute the correlation coefficient.

$$r = \frac{1}{5}\left(-\frac{3}{4} - \frac{1}{8} + 0 + 0 - \frac{7}{8} + \frac{15}{8}\right) = \frac{1}{5}\cdot\frac{1}{8} = \frac{1}{40} = 0.025$$

- Find the regression line.

$$\begin{aligned}
\big(y - \text{mean}(y)\big) &= r\,\frac{\text{sd}_y}{\text{sd}_x}\,\big(x - \text{mean}(x)\big) \\
(y-8) &= \frac{1}{40}\cdot\frac{4}{2}\,(x-4) \\
(y-8) &= \frac{1}{20}\,(x-4) \\
20\,(y-8) &= x-4 \\
20\,y - 160 &= x-4 \\
20\,y &= x+156 \\
y &= \frac{x+156}{20} \\
&= 0.05\,x + 7.8
\end{aligned}$$

- Compute the y values predicted by the regression line. If $x = 1$, then

$$y = (1 + 156)/20 = 157/20 = 7.85$$

and if $x = 3$, then

$$y = (3 + 156)/20 = 159/20 = 7.95.$$

- Compute the regression errors (error $= y -$ predicted y). For example,

if $x = 1$, then error $= 10 - 7.85 = 2.15$ and if $x = 3$, then error $= 9 - 7.95 = 1.05$.

- Compute the RMS size of the errors.

$$\text{RMS}_{\text{reg}} = \sqrt{(2.15^2 + 1.05^2 + 0^2 + (-1)^2 + (-7.05)^2 + 4.85^2)/6} = \sqrt{(79.95)/6} = \sqrt{13.325} \approx 3.65.$$

Here is how to get `R` to do all this:

`x <- c(1, 3, 4, 4, 5, 7)`	define the x values
`y <- c(10, 9, 8, 7, 1, 13)`	define the y values
`mean(x)`	We can actually skip from here ...
`sd(x)`	
`mean(y)`	
`sd(y)`	
`zx <- (x - mean(x)) / sd(x)`	
`zy <- (y - mean(y)) / sd(y)`	
`r <- mean(zx * zy)`	
`slope <- r * sd(y) / sd(x)`	
`yIntercept <- mean(y) - slope * mean(x)`	... to the next line
`linearModel <- lm(y ~ x)`	
`summary(linearModel)`	read the slope and y intersecpt from this
`fitted.values(linearModel)`	this gives the y values predicted by regression
`RegressionErrors <- residuals(linearModel)`	
`SDrms <- sqrt(mean(RegessionErrors**2))`