Regression and correlation

We consider bivariate data: i.e. data for two variables, $x$ and $y$, and we seek to investigate the relationship between these variables.

Regression is used for prediction: using $x$ to predict $y$. Correlation is used to measure association between $x$ and $y$.

The data consists of $n$ pairs of data points: $\{(x_i, y_i), \ i = 1, 2, \ldots, n\}$, which can be plotted on a “scatter diagram” or “scatter plot”, the bivariate analogue of a dotplot.

Regression

The regression of $Y$ on $x$ is $E(Y \mid x)$, the expectation of $Y$ given the value of $x$.

For example, $Y$ may be the measured pressure of a gas in a given volume $x$, the measurement being subject to error. Here, we might expect that $E(Y \mid x) = \frac{c}{x}$. There is no need for the regression to be of a specific form.

However, we will consider the case:

$$E(Y \mid x) = \alpha + \beta x \quad \text{and} \quad \text{var}(Y \mid x) = \sigma^2$$

so that the regression of $Y$ on $x$ is a straight line.

straight line vs linear

Why “regression”? 

Note that it is possible to transform data to produce a linear regression. For example:

1. In the above example on pressure and volume, if we write $x^* = \frac{1}{x}$, then $E(Y \mid x^*) = c x^*$, which is a linear model with zero intercept.

2. If $y \approx \alpha x^3$, then it might be appropriate to take logs and consider the model $E(Y^* \mid x^*) = \alpha^* + \beta x^*$, where $Y^* = \ln y$, $x^* = \ln x$, $\alpha^* = \ln \alpha$. But . . .
Estimation of $\alpha$, $\beta$: the method of least squares

We select the straight line $y = a + bx$ for which:

$$\Delta = \Delta(a, b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

is a minimum.

This line we denote by

$$\hat{\mu}(x) = \hat{\alpha} + \hat{\beta}x$$

$\hat{\alpha}$ and $\hat{\beta}$ denote the least squares estimates of $\alpha$ and $\beta$.

To find where $\Delta$ is a minimum, we find $\frac{\partial \Delta}{\partial a}$, $\frac{\partial \Delta}{\partial b}$ and equate them to zero. This gives:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

so that

$$\hat{\mu}(x) = \bar{y} + \hat{\beta}(x - \bar{x})$$

Note: this derivation is simpler in terms of $\alpha_0$ and $\beta$ (see PB=p126) as this produces an orthogonal parameterisation (of which more in Chapter 8).

model: $\alpha + \beta x = \alpha_0 + \beta(x - \bar{x}) \quad \alpha_0 = \alpha + \beta \bar{x}$

estimates: $\hat{\alpha} + \hat{\beta}x = \bar{y} + \hat{\beta}(x - \bar{x}) \quad \hat{\alpha}_0 = \bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}$

example Consider the really simple data set given by:

<table>
<thead>
<tr>
<th>$x$</th>
<th>73</th>
<th>62</th>
<th>81</th>
<th>82</th>
<th>65</th>
<th>49</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>57</td>
<td>49</td>
<td>84</td>
<td>78</td>
<td>62</td>
<td>48</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
n &= 6 & \Sigma xy &= 26781 & \Sigma (x - \bar{x})(y - \bar{y}) &= 825 \\
\Sigma x &= 412 & \Sigma x^2 &= 29084 & \Sigma (x - \bar{x})^2 &= 793.33 \\
\Sigma y &= 378 & \Sigma y^2 &= 24398 & \Sigma (y - \bar{y})^2 &= 584
\end{align*}
\]

$$\hat{\beta} = \frac{825}{793.33} = 1.0399, \quad \hat{\alpha} = 63 - 1.0399 \times 68.67 = -8.4065.$$ 

Therefore $\hat{\mu}(x) = -8.41 + 1.04x$,

so that, for example, the mean of $y$ when $x = 62$ is estimated by $\hat{\mu}(62) = -8.41 + 1.04 \times 62 = 56.1$. 
But what about standard errors?

Results:

- $\bar{Y} = \frac{1}{n} \sum Y_i$
- $\hat{B} = \sum k_i Y_i$
- $E(\bar{Y}) = \alpha + \beta \bar{x}$
- $E(\hat{B}) = \beta$
- $\text{var}(\bar{Y}) = \frac{\sigma^2}{n}$
- $\text{var}(\hat{B}) = \frac{\sigma^2}{K}$
- $\text{cov}(\bar{Y}, \hat{B}) = 0$

where $K = \sum (x_i - \bar{x})^2$ and $k_i = (x_i - \bar{x})/K$.

The variance and covariance results are proved using

$$\text{cov}(\sum \ell_i Y_i, \sum k_i Y_i) = (\sum \ell_i k_i) \sigma^2,$$

since the $Y_i$ are independent, each with variance $\sigma^2$.

So for example: $\text{var}(\hat{A}) = \frac{\sigma^2}{n} + \frac{\bar{x}^2}{K}$. [Note: $\bar{Y} = \hat{A}_0$.]

Further, if $\hat{M}(x)$ denotes the estimator of $\mu(x)$, that is, $\hat{M}(x) = \bar{y} + (x - \bar{x})\hat{B}$ then we have

$$\text{var}(\hat{M}(x)) = \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{K}\right) \sigma^2.$$ [Note: $\hat{A} = \hat{M}(0)$.]

**Example** For the above data, we have:

- $\text{var}(\bar{Y}) = 0.1667 \sigma^2$,  $\text{var}(\hat{B}) = 0.001261 \sigma^2$
- $\text{var}(\hat{A}) = 6.1101 \sigma^2$,  $\text{var}(\hat{M}(62)) = 0.2227 \sigma^2$.

But we don’t know $\sigma^2$ — and so we need an $s^2$.

We define the residual sum of squares (after fitting the regression line) as

$$d^2 = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}x_i)^2.$$

To estimate $\sigma^2$, we use: $s^2 = \frac{d^2}{n - 2}$, as $S^2$ is unbiased for $\sigma^2$.

To prove this:

$$\sum (Y_i - \hat{\alpha} - \hat{\beta}u_i)^2$$

$$= \sum ((Y_i - \hat{\alpha} - \hat{\beta}u_i) + (\hat{\alpha}_0 - \alpha_0) + (\hat{\beta} - \beta)u_i)^2$$

$$= \sum (Y_i - \hat{\alpha}_0 - \hat{\beta}u_i)^2 + n(\hat{\alpha}_0 - \alpha_0)^2 + (\hat{\beta} - \beta)^2 \sum u_i^2$$

since the cross-product terms are zero. Taking expectations gives:

$$n\sigma^2 = E(D^2) + \sigma^2 + \sigma^2,$$

so that $E(D^2) = (n - 2)\sigma^2$; and hence $E(S^2) = \sigma^2$.

For computation (if you have to):

$$d^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \left( \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)^2$$

**Example** For the figures above we find:

$$d^2 = 266.07 \quad \text{so that} \quad s^2 = \frac{1}{4} d^2 = 66.52.$$
Straight line regression assuming normally distributed errors: i.e. $Y_i \overset{d}{=} N(\alpha + \beta x_i, \sigma^2)$

In this case:

$$Y \overset{d}{=} N(\alpha + \beta \bar{x}, \sigma^2/n) \quad \text{and} \quad \hat{B} \overset{d}{=} N(\beta, \sigma^2/K)$$

and further, $\bar{Y}$ and $\hat{B}$ are independent.

Also $D^2 \overset{d}{=} \sigma^2 \chi^2_{n-2}$, so that $\frac{(n-2)S^2}{\sigma^2} \overset{d}{=} \chi^2_{n-2}$.

These results can be used for inference on $\alpha$ and $\beta$; and, more generally, on $\mu(x)$.

For example:

$$\text{se}(\hat{B}) = \frac{s}{\sqrt{K}} \quad \text{and} \quad \frac{\hat{B} - \beta}{S/\sqrt{K}} \overset{d}{=} t_{n-2}$$

$$\text{se}(\hat{M}) = \frac{cs}{S} \quad \text{and} \quad \frac{\hat{M} - \mu}{cS} \overset{d}{=} t_{n-2}$$

**example** For the data given above, a 95% confidence interval for $\beta$ is given by:

$$1.0399 \pm 2.776 \times 8.156 \times \sqrt{0.001261}$$

i.e. $0.236 < \beta < 1.844$

Similarly, a 95% confidence interval for $\mu(62)$ is given by

$$56.07 \pm 2.776 \times 8.156 \times \sqrt{0.2227}$$

i.e. $45.39 < \mu(62) < 66.75$

This interval gives a confidence interval for the unknown mean of $Y$ when $x = 62$ — it says nothing about the actual observations for $x = 62$.

To do that, a *prediction interval* for $Y$ is required: an interval within which we are 95% sure that a future observation will lie. To obtain such an interval, we use:

$$Y^* - \hat{M}(x) \overset{d}{=} N \left( 0, \sigma^2 \left( 1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{K} \right) \right).$$

95% prediction interval for $Y$:

$$\hat{\mu}(x) \pm c_{0.975}(t_{n-2}) s \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{K}}.$$

**example** A 95% prediction interval for $Y$ at $x = 62$ is given by:

$$56.07 \pm 2.776 \times 8.156 \times \sqrt{1.2227}$$

i.e. $31.03 < Y(62) < 81.11$.

And, as usual, the prediction interval is much wider than the confidence interval.
Analysis of variance approach

\[
data = \text{model} + \text{error} = \begin{bmatrix} y - \alpha_0 \\ \beta(x - \bar{x}) \end{bmatrix} + \begin{bmatrix} e \end{bmatrix}
\]

\[
\Sigma(y - \bar{y})^2 = \beta^2 \Sigma(x - \bar{x})^2 + \Sigma(y - \hat{\alpha} - \hat{\beta}x)^2 = K\beta^2 + d^2
\]

\[
\text{total variation} = \text{variation due to the straight line} + \text{variation around to the straight line}
\]

\[
\text{total SS} = \text{regression SS} + \text{residual SS}
\]

\[
\sigma^2 \chi^2_{n-1} \quad \sigma^2 \chi^2_1 \quad \sigma^2 \chi^2_{n-2}
\]

On Minitab a straight line regression is fitted using the `regr` command: apart from the parameter estimates and standard errors, this produces an analysis of variance as a matter of course. Confidence intervals and prediction intervals are obtained using the subcommand `predict`.

MTB > print c1 c2
ROW x y
1 73 57
2 62 49
3 81 84
4 82 78
5 65 62
6 49 48

MTB > regr c2 1 c1;
SUBC> predict 62.

The regression equation is \( y = -8.4 + 1.04 \times \)

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Stdev</th>
<th>t-ratio</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-8.41</td>
<td>20.16</td>
<td>-0.42</td>
<td>0.698</td>
</tr>
<tr>
<td>x</td>
<td>1.0399</td>
<td>0.2896</td>
<td>3.59</td>
<td>0.023</td>
</tr>
</tbody>
</table>

\( s = 8.156 \quad R^2 = 76.3\% \quad R^2(\text{adj}) = 70.4\% \)

Analysis of Variance

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>857.93</td>
<td>857.93</td>
<td>12.90</td>
<td>0.023</td>
</tr>
<tr>
<td>Error</td>
<td>4</td>
<td>266.07</td>
<td>66.52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>5</td>
<td>1124.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fit Stdev.Fit 95% C.I. 95% P.I.
56.07 3.85 (45.38, 66.76) (31.02, 81.11)

Occasionally, we want to fit the model:

\[ E(Y_i) = \beta x_i, \quad \text{var}(Y_i) = \sigma^2. \]

This requires a different fit (see Problem 7.2 and Homework 9.2) and can be simply obtained on Minitab using the `noconst` subcommand:

MTB > regr c2 1 c1;
SUBC> noconst.

The other subcommands (e.g. `predict`), or menu choices, still operate.
Method of Least Squares

The method of least squares is quite general: Suppose that \( Y_1, Y_2, \ldots, Y_n \) are independent random variables whose expectations depend on a given (vector) variable taking values \( x_1, x_2, \ldots, x_n \) and unknown parameters \( \theta_1, \theta_2, \ldots, \theta_p \) such that:

1. \( \text{E}(Y_i) = \mu(x_i; \theta) \)
2. \( \text{var}(Y_i) = \sigma^2(\theta) \)

Then, the least squares estimate \( \hat{\theta} \) is that value of \( t \) for which the sum of squares:

\[
\Delta(t) = \sum_{i=1}^{n} (y_i - \mu(x_i; t))^2
\]

is a minimum over all possible vectors \( t \).

If instead of assumption (2) above, we have:

(2a) \( \text{var}(Y_i) = w(x_i)\kappa(\theta) \)

then, the (weighted) least squares estimate \( \hat{\theta} \) is that value of \( t \) such that

\[
\Delta(t) = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; t))^2}{w(x_i)}
\]

is a minimum over all possible vectors \( t \).

This is derived by applying the method of least squares applied to \( Y_i/\sqrt{w(x_i)} \) which has constant variance.

**example** If \( Y_i \overset{d}{=} \text{Pn}(\beta x_i) \), then to estimate \( \alpha \) by the method of least squares we minimise

\[
\Delta = \sum_{i=1}^{n} \frac{(y_i - bx_i)^2}{x_i}
\]

To do this, we take a derivative with respect to \( b \) and equate to zero:

\[
\frac{\partial \Delta}{\partial b} = \sum_{i=1}^{n} \left( -2x_i \right) \frac{y_i - bx_i}{x_i} = 0 \quad \Rightarrow \quad \hat{\beta} = \frac{\Sigma y_i}{\Sigma x_i}.
\]

Unweighted least squares gives \( \hat{\beta} = \frac{\Sigma x_i y_i}{\Sigma x_i^2} \) (regression through the origin: see above).

Both \( \hat{\beta} \) and \( \hat{\beta} \) are unbiased but \( \hat{\beta} \) is more efficient in this case.

If \( Y_i \overset{d}{=} \text{Pn}(\beta x_i) \), then it is found that the maximum likelihood estimate of \( \beta \) is the same as the weighted least squares estimate:

\[
L(\beta) = K \prod e^{-\beta x_i} (\beta x_i)^{y_i}
\]

\[
\ln L = -\beta \sum x_i + \ln \beta \sum y_i + K
\]

\[
\frac{\partial \ln L}{\partial \beta} = -\sum x_i + \frac{1}{\beta} \sum y_i
\]

\[
\frac{\partial^2 \ln L}{\partial \beta^2} = -\frac{1}{\beta^2} \sum y_i
\]
Correlation

We define the sample correlation coefficient as:

\[ r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{s_{XY}}{s_X s_Y} \]

This is an estimator of the population correlation coefficient, \( \rho \).

Note that \( |r| \leq 1 \) (as is the case for \( \rho \)).

On MINITAB, to obtain correlation, use the command corr: thus corr c1 c2 gives the correlation between the data in c1 and c2. A command such as corr c1-c5 gives all the pairwise correlations between the data in c1, c2, ..., c5.

Analysis of variance approach:

\[ \Sigma(y - \hat{y})^2 = \hat{\beta}^2 \Sigma(x - \bar{x})^2 + \Sigma(y - \hat{\alpha} - \hat{\beta}x)^2 \]
\[ \Sigma(y - \bar{y})^2 = r^2 \Sigma(y - \bar{y})^2 + (1 - r^2) \Sigma(y - \bar{y})^2 \]

So, \( r^2 \) is the proportion of the variation in \( y \) explained by \( x \). And, as we have seen, the MINITAB output includes the value of \( r^2 \).

Distribution of \( R \) when \( \rho = 0 \) (for bivariate normal data)

\[ F = \frac{\text{regression MS}}{\text{residual MS}} = \frac{r^2 \Sigma(y - \bar{y})^2/1}{(1 - r^2) \Sigma(y - \bar{y})^2/(n - 2)} \]

i.e. \[ F = \frac{(n - 2)r^2}{1 - r^2} \]

and if \( \beta = 0 \) (\( \rho = 0 \)) then \[ \frac{(n - 2)R^2}{1 - R^2} \geq F_{1,n-2}. \]

It follows that if \( \rho = 0 \) then \[ R \sqrt{\frac{n - 2}{n - 2}} \geq t_{n-2}. \]
This can be used to test $H_0: \rho = 0$. For example to test $H_0: \rho = 0$ against $H_1: \rho > 0$, we would reject $H_0$ if
\[
\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} > c_{0.05}(t_{n-2}).
\]
In the case of normally distributed data, testing $\rho = 0$ is really equivalent to testing $\beta = 0$.

To test $H_0: \rho = \rho_0$ (for $\rho_0 \neq 0$) and to set confidence intervals for $\rho$, we must use another approach:

For $\rho \neq 0$, we have the following approximate result:
\[
z = \frac{1}{2} \ln \frac{1 + r}{1 - r} \approx N \left( \frac{1}{2} \ln \frac{1 + \rho}{1 - \rho}, \frac{1}{n - 3} \right).
\]
Note that $z = \text{artanh} r$, or $r = \tanh z$, which is on most calculators and is also tabulated.

**Example** Suppose that, for a sample of $n = 25$ observations on $(X, Y)$, we obtain $r = 0.5$.

Using the above result, an approximate 95% confidence interval for $\text{artanh} \rho$ is given by $0.5493 \pm 1.96 \times 0.2132$, that is, $0.1314 < \text{artanh} \rho < 0.9679$.

Thus, since tanh is an increasing function, an approximate 95% confidence interval for $\rho$ is given by $0.13 < \rho < 0.75$.

Alternatively, and more simply, an approximate 95% confidence interval for $\rho$ based on $R$ can be read off the SP diagram in the tables (PB=p258).

**Rank correlation**

correlation measures linear relationship;

rank correlation measures monotonic relationship

The rank correlation coefficient, $r'$ is just what it says:

the (sample) correlation between the $x$-ranks and the $y$-ranks.

Let $u_i = \text{rank}(x_i), \ i = 1, 2, \ldots, n$;

and $v_i = \text{rank}(y_i), \ i = 1, 2, \ldots, n$; then

\[
r' = r'_{XY} = r_{UV} = \frac{s_{uv}}{s_u s_v} = \frac{\Sigma(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\Sigma(u_i - \bar{u})^2 \Sigma(v_i - \bar{v})^2}}
\]

Since the $u$s and $v$s are ranks,

$\Sigma u = \Sigma v = \frac{1}{2}n(n + 1)$;

and (provided there are no ties)

$\Sigma u^2 = \Sigma v^2 = \frac{1}{6}n(n + 1)(2n + 1)$.

This leads to the relatively simple expression

\[
r' = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \quad \text{where } d_i = u_i - v_i
\]

(which is only approximately correct if there are ties: the more ties the worse the approximation).

The rank correlation gives a useful distribution-free test of independence, since we can derive the distribution of $r'$ assuming only that $X$ and $Y$ are independent.
If $X$ and $Y$ are independent, then all possible rank orderings are equally likely. Thus all that is required is to compute $r'$ for each ordering and to count the number for which $r' = t\;,-1 \leq t \leq 1$. Critical values for $r'$ are tabulated: see PB=p256.

**example**  The following data give the scores of eight individuals on two tests $A$ and $B$. Are the test scores positively related? i.e. test $H_0$: $X$ & $Y$ independent vs $H_1$: $X$ & $Y$ positively related.

<table>
<thead>
<tr>
<th></th>
<th>$x$</th>
<th>68</th>
<th>43</th>
<th>87</th>
<th>67</th>
<th>20</th>
<th>50</th>
<th>52</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>53</td>
<td>52</td>
<td>67</td>
<td>72</td>
<td>43</td>
<td>51</td>
<td>63</td>
<td>84</td>
<td></td>
</tr>
</tbody>
</table>

Ranking the $x$s and $y$s separately, we obtain

<table>
<thead>
<tr>
<th></th>
<th>$u$</th>
<th>6</th>
<th>2</th>
<th>8</th>
<th>5</th>
<th>1</th>
<th>3</th>
<th>4</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v$</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

$\sum d^2 = 16 \Rightarrow r' = 1 - \frac{6 \times 16}{8 \times 43} = 0.8095$

Tables: for $n = 8$ reject $H_0$ if $r' \geq 0.643$ (size 0.05)

Thus we reject $H_0$ and say they are positively related.

On MINITAB, the rank correlation of $X$ (in C1) and $Y$ (in C2) is obtained as follows:

```
MTB > RANK C1 C11
MTB > RANK C2 C12
MTB > CORR C11 C12
```

For large values of $n$, we have the approximate result:

$$H_0 \text{ (independence)} \Rightarrow r' \approx N\left(0, \frac{1}{n-1}\right)$$

For example, if $n = 250$, then we would reject independence (in favour of non-independence) if $|r'| > 0.124 = 1.96 \sqrt{\frac{1}{249}}$. 
