Linear Regression and Correlation Notes

Suppose there is a data set of \( n \) data points \((x_i, y_i)\) where you have plotted these using a scatter plot and it appears that a linear relationship between them is reasonable. Then the least-squares line (regression line) that best fits these data,

\[
y = \hat{m} x + \hat{b}
\]

has the regression coefficients \( \hat{m} \) and \( \hat{b} \) chosen so as to minimize the sum of the square errors

\[
\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\hat{m} x_i + \hat{b}))^2
\]

This says that the regression line that "best fits" the data is the line chosen so as to provide the smallest average difference between the data points \((y_i)\) and the \( y \)-values predicted by the regression line \((\hat{y}_i)\).

The values of the regression coefficients are calculated from

\[
\hat{m} = \frac{S_{xy}}{S_{xx}}
\]

where

\[
S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left( \sum_{i=1}^{n} x_i \right)^2}{n} = \sum_{i=1}^{n} (x_i - \bar{x})^2
\]

and

\[
S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left( \sum_{i=1}^{n} x_i \right) \left( \sum_{i=1}^{n} y_i \right)}{n} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})
\]

and

\[
\hat{b} = \bar{y} - \hat{m} \bar{x}
\]

and \( \bar{x} \) and \( \bar{y} \) are the means defined by

\[
\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{and} \quad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}
\]

The correlation coefficient is defined to be

\[
\hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}
\]

where

\[
S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left( \sum_{i=1}^{n} y_i \right)^2}{n} = \sum_{i=1}^{n} (y_i - \bar{y})^2
\]
Note that \(-1 \leq \rho \leq 1\).

A way to interpret this is to define the Total Sum of Squares (TSS) of the data set as

\[
TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2
\]

(note that \(TSS = S_{yy}\)) and the Sum of Squares of the Regression (SSR) as

\[
RSS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2
\]

and note that since \(y_i\) will not be exactly on the regression line, \(TSS > RSS\) (unless the points are exactly on a line in which case \(TSS = RSS\)). Then the closer the points are to the regression line, the closer \(TSS\) is to \(RSS\). The Coefficient of Determination is defined to be \(r^2 = RSS/TSS\). So as the data points get close to being exactly on a line, \(RSS\) gets close to \(TSS\) and so \(r^2\) gets close to 1. When \(r^2\) is close to 1, the points are said to be highly correlated which means that a very large proportion of the Total Sum of Squares is accounted for by the regression (SSR). Thus the Coefficient of Determination is a measure of the strength of the straight-line relationship.

It is possible to show that

\[
RSS = \frac{S_{xy}^2}{S_{xx}}
\]

and so that

\[
r^2 = \frac{RSS}{TSS} = \frac{S_{xy}^2}{S_{xx} S_{yy}} = \rho^2
\]

so that the correlation coefficient can be thought of as measuring how well as regression line fits a data set.