Análise de Regressão e Correlação

Foi já estudado a forma de descrever um conjunto de observações de uma só variável. Quando se consideram observações de duas ou mais variáveis surge um novo ponto.

"O estudo das relações porventura existentes entre as variáveis."

A Análise de regressão e correlação, compreende a análise de dados amostrais para saber se e como as duas ou mais variáveis estão relacionadas uma com a outra numa população.

A **análise de regressão** estuda o relacionamento entre uma variável chamada a variável dependente e outras variáveis chamadas variáveis independentes. Este relacionamento é representado por um modelo matemático, i.e., por uma equação que associa a variável dependente com as variáveis independentes. Este modelo é designado por **modelo de regressão linear simples** se define uma relação linear entre a variável dependente e uma variável independente. Se em vez de uma, forem incorporadas várias variáveis independentes, o modelo passa a denominar-se **modelo de regressão linear múltipla**.

A análise de correlação dedica-se a inferências estatísticas das medidas de associação linear que se seguem:

- **coeficiente de correlação simples**: mede a "força" ou "grau" de relacionamento linear entre duas variáveis"
- **coeficiente de correlação múltiplo**: mede a "força" ou "grau" de relacionamento entre uma variável dependente e um conjunto de outras variáveis.

As técnicas de análise de correlação e regressão estão intimamente ligadas.

Correlação e Regressão Simples

Só vamos falar de correlação e regressão linear simples, i.e., no caso de uma variável dependente (Y) e uma variável independente (X).

Exemplos:

1. Relação entre o peso e a altura de um homem adulto. A variável dependente é o peso e a variável independente a altura.
2. A relação entre o preço do vinho e o montante da colheita em cada ano. Aqui a variável dependente é o preço do vinho e a variável independente o montante da colheita.

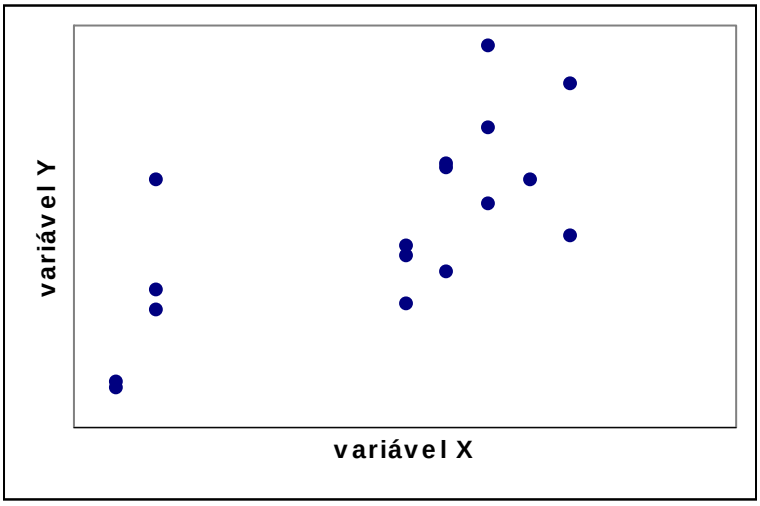
Para estudar estas relações recorre-se a uma amostra e utiliza-se a análise de correlação e regressão simples.

Note, que para os exemplos anteriores pode suceder que dois homens adultos tenham a mesma altura e pesos diferentes e vice-versa, no entanto **em média** quanto maior for a altura maior será o peso; do mesmo modo a colheitas iguais podem corresponder preços diferentes e vice-versa, no entanto **em média** quanto maior for a colheita menor será o preço do vinho.

É essa variação em média que vai ser estudada. A correlação (entre X e Y) é positiva quando os fenómenos variam no mesmo sentido (primeiro caso apresentado no exemplo 1), a correlação (entre X e Y) é negativa quando os fenómenos variam em sentido inverso (segundo caso apresentado no exemplo 1).

Diagramas de dispersão

Os dados para a análise de regressão e correlação provém de observações de variáveis emparelhadas, isto significa que cada observação origina dois valores, um para cada variável, com estes valores constrói-se o digrama de dispersão.



A regressão linear simples constitui uma tentativa de estabelecer uma equação matemática linear (linha recta) que descreva o relacionamento entre duas variáveis.

Note-se que nem todas as situações são bem aproximadas por uma equação linear. Através dos diagramas de dispersão pode-se ver se uma relação linear parece razoável ou não. Recorrendo à análise do diagrama de dispersão pode-se também concluir se o grau de correlação é forte ou fraca, conforme o modo com se situem os pontos em redor de uma linha recta imaginária que passa

através de um “enxame” pontos. A correlação é tanto maior quanto mais os pontos se concentram, com pequenos desvios, em relação a essa recta.

Determinação da Recta de Regressão

Consideremos uma recta arbitrária, $y=\beta_0+\beta_1 x$, desenhada no diagrama. A x_i chamamos valor da variável explicativa ou independente e à imagem de x_i pela recta $y=\beta_0+\beta_1 x$ chamamos valor predito, que denotamos por $\hat{y}_i$, y_i é o valor da variável resposta ou dependente.

A diferença entre y_i e $\hat{y}_i$, i.e., $d_i = y_i - \hat{y}_i$ é a distância vertical do ponto à linha recta. Se consideramos a soma dos quadrados dos desvios anteriores, i.e.,

$$D = \sum_{i=1}^{n} d_i^2$$

obtemos uma medida do desvio total dos pontos observados à recta estimada.

A medida anterior depende da recta considerada, ou seja depende de β_0 e β_1. Assim, podemos escrever

$$D(\beta_0, \beta_1) = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

ou ainda,

$$D(\beta_0, \beta_1) = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x))^2 \ .$$

Pretendemos então os valores de β_0 e β_1 que minimizem $D(\beta_0, \beta_1)$, i.e., pretendemos o valor mínimo de $D(\beta_0, \beta_1)$.

Um modo de estimar os coeficientes β_0 e β_1 é determinar o mínimo da função $D(\beta_0,\beta_1)$ em relação a β_0 e β_1 e resolver as equações normais.

Temos então que:

$$D(\beta_0, \beta_1) = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x)^2$$

donde

$$\frac{\partial D(\beta_0, \beta_1)}{\partial \beta_0} = \sum_{i=1}^{n} -2(y_i - \beta_0 - \beta_1 x_1)$$

$$\frac{\partial D(\beta_0, \beta_1)}{\partial \beta_1} = \sum_{i=1}^{n} -2x_i(y_i - \beta_0 - \beta_1 x_1)$$

Os valores de b_0 e b_1 para os quais a função $D(\beta_0,\beta_1)$ apresenta um valor mínimo são obtidos igualando as equações anteriores a zero, i.e., resolvendo as equações normais. Assim,

$$\begin{cases} -2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_1)=0 \\ -2\sum_{i=1}^{n}(y_i x_i-\beta_0 x_i-\beta_1 x_1^{\,2})=0 \end{cases} \Leftrightarrow \begin{cases} \sum_{i=1}^{n} y_i - nb_0 - b_1\sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} y_i x_i - b_0\sum_{i\sim=1}^{n} x_{\ i} - b_1\sum_{i=1}^{n} x_1^{\,2} = 0 \end{cases} \Leftrightarrow \qquad (5.1)$$

$$\Leftrightarrow \begin{cases} b_0 = \dfrac{\sum_{i=2}^{n} y_i - b_1\sum_{i=1}^{n} x_i}{n} \\ b_1\sum_{i=1}^{n} x_i^2 - \dfrac{\sum_{i=1}^{n} x_i\sum_{i=1}^{n} y_i - b_1\sum_{i=1}^{n} x_i\sum_{i=1}^{n} x_i}{n} = -\sum_{i=1}^{n} x_i y_i \end{cases} \Leftrightarrow$$

$$\Leftrightarrow \begin{cases} \\ b_1\left(\sum_{i=1}^{n} x_i^2 - \dfrac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right) = -\sum_{i=1}^{n} x_i y_i + \dfrac{1}{n}\sum_{i=1}^{n} x_i\sum_{i=1}^{n} y_i \end{cases} \Leftrightarrow$$

$$\Leftrightarrow \begin{cases} \\ b_1 = \dfrac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i\sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \end{cases} \Leftrightarrow$$

Temos então que

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1\sum_{i=1}^{n} x_i}{n} \text{ e } b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i\sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

ou

$$b_0 = \bar{y} - b_1\bar{x} \text{ e } b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - n^2\overline{xy}}{n\sum_{i=1}^{n} x_i^2 - n^2\bar{x}^2},$$

são as soluções dos sistema inicial sendo além disso os valores de β_0 e β_1 que minimizam $D(\beta_0,\beta_1)$.

Este método é conhecido pelo método dos mínimos quadrados, uma vez que estamos a minimizar uma função quadrática.

A melhor recta, no sentido dos mínimos quadrados, que melhor se ajusta aos dados do diagrama de dispersão é dada por: $y=b_0 + b_1x$.

Qualidade do ajustamento

Uma medida útil associada à recta de regressão, é o grau com que as predições baseadas na equação de regressão, superam as predições baseadas em $\bar{y}$. Isto é, se as predições baseadas na recta não são melhores que as baseadas no valor médio de Y ($\bar{y}$), então não adianta dispormos de uma equação de regressão.

Para a observação y_i a diferença em relação ao valor médio $\bar{y}$ é conhecida por desvio total e pode decompor-se numa soma de parcelas:

$$\underbrace{(y_i - \bar{y})^2}_{\text{Desvio Total}} = \underbrace{(\hat{y}_i - \bar{y})^2}_{\substack{\textit{Desvio}\,\text{exp}\,\textit{licado} \\ \text{pelo modelo}}} + \underbrace{(y_i - \hat{y}_i)^2}_{\substack{\textit{Desvio não}\,\text{exp}\,\textit{licado} \\ \text{ou resíduo}}}$$

Considerando todas as observações (x_i, y_i), i01,...,n, obtemos a variação total:

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{Variação Total}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\substack{\textit{Variação}\,\text{exp}\,\textit{licado} \\ \text{pelo modelo}}} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\textit{Variação não}\,\text{exp}\,\textit{licado}}$$

O **coeficiente de determinação** — $\mathbf{R^2}$— é uma medida do poder explicativo do modelo utilizado. Dá a proporção da variação da variável dependente, Y, que é explicada em termos lineares pela variável independente, X, i.e., a proporção da variação de Y explicada pelo modelo.

$$R^2 = \frac{\text{variação explicada}}{\text{variação total}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Na prática,

$$R^2 = \frac{a\sum_{i=1}^{n} y_i + b\sum_{i=1}^{n} x_i y_i - n\bar{y}^2}{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}$$

Tem-se que

- $0 \leq R^2 \leq 1$ a proporção da variação de Y explicada pelo modelo é no máximo 1 e no mínimo 0.
- Se $R^2 \cong 1$ significa que grande parte da variação de Y é explicada linearmente por X (modelo adequado).

- Se $R^2 \cong 0$ o modelo não é adequado aos dados.
- 1- R^2 é a proporção de variação de Y não explicada pela variável X, resultante de factores não incluídos no modelo.

O coeficiente de determinação pode ser utilizado como uma medida da qualidade do ajustamento ou como medida da qualidade de confiança depositada na equação de regressão como instrumento de precisão.

A $R = \sqrt{R^2}$ dá-se o nome de **coeficiente de correlação simples**. É uma medida do grau de associação linear entre as variáveis X e Y.

Tendo-se que

- $-1 \leq R \leq 1$
- Se $R > 0$ então as duas variáveis tendem a variar no mesmo sentido; em média uma aumento da variável X provoca um aumento da variável Y;
- Se $R < 0$ então as duas variáveis tendem a variar em sentido negativo; em média um aumento da variável X provoca uma diminuição da variável Y;
- $R = 1$ ou $R = -1$ indicam a existência de uma relação linear perfeita entre X e Y, positiva ou negativa, respectivamente;
- $R = 0$ indica a inexistência de uma relação linear entre X e Y, podendo, no entanto, existir uma relação não linear entre elas.

<u>Observações:</u>

1. Um modelo de regressão linear não dá respostas exactas; assim, para um determinado valor de x da variável X espera-se, em média, que $\hat{y} = b_1 x + b_0$;
2. A estimação, ou previsão, de uma variável com base em valores conhecidos da outra deve ser cautelosa! Não deve ser feita qualquer extrapolação dessa recta para valores fora do âmbito dados. O perigo de extrapolar para fora do âmbito dos dados amostrais é que a mesma relação possa não mais se verificar.
3. A existência de correlação nada diz sobre a natureza da relação causal que porventura exista entre as variáveis. Ao interpretar um coeficiente de correlação deve ter-se presente, que uma valor elevado de R não significa que X seja causa de Y ou Y seja causa de X. A análise de regressão apenas indica qual o relacionamento matemático pode existir, se existir algum; a lógica de uma relação causal deve provir de teorias externas ao âmbito da Estatística.

Note-se que a n pontos observados é teoricamente possível ajustar uma infinidade de curvas. No estudo feito, apenas foi possível abordar o modelo de regressão linear simples. No entanto, como já vimos, o modelo o modelo linear nem sempre é o mais adequado; a representação gráfica dos dados por vezes sugere que estes são melhor ajustados por outras curvas do que por uma recta. É portanto necessário, em primeiro lugar, fixar o modelo que melhor se adapta às observações.

Outros exemplos possíveis, além do modelo dado $Y=b_0+b_1X$:

$$Y = b_0 + b_1 X + b_2 X^2$$

$$Y = ab^x, \ldots.$$

Além do tipo de curva, outro factor importante na análise de regressão, é o número de variáveis envolvidas. Em muitos problemas práticos, em vez de ser considerada apenas uma variável independente, é do interesse estudar a relação entre uma variável e um conjunto de variáveis — Análise de Regressão Múltipla.

$$Y = b_0 + b_1 X_1 + b_2 X_2 p + b_3 X_3$$

Trata-se de uma análise mais complexa e que caí fora do programa da disciplina.