Correlação e Regressão Linear Simples

Introdução

A análise de regressão estuda o relacionamento entre uma variável, designada por variável dependente, e outras variáveis, designadas por variáveis independentes.

Este relacionamento é representado por um modelo matemático, isto é, por uma equação que associa a variável dependente com as variáveis independentes.

O modelo é designado por **modelo de regressão linear simples** se define uma relação linear entre a variável dependente e uma única variável independente.

Se em vez de uma, forem incorporadas no modelo várias variáveis independentes, o modelo designa-se por
modelo de regressão linear múltipla.

Diagrama de Dispersão

O Diagrama de dispersão é uma representação gráfica para dados bivariados, onde cada par (x_i, y_i) é representado pelo ponto de coordenadas (x_i, y_i), num sistema de eixos coordenados.

As técnicas de análise de correlação e regressão estão intimamente ligadas, e para esse fim, este tipo de representação gráfica é muito útil, pois permite realçar se existe ou não uma relação entre as variáveis e, em caso afirmativo, que tipo de relação.

Em particular, permite decidir empiricamente se se pode assumir que existe um relacionamento linear entre 2 variáveis; situação esta que ocorre se a nuvem de pontos marcados exibir uma tendência linear.

Diagrama de Dispersão

Pela análise deste gráfico pode-se ainda concluir, empiricamente, se o grau de relacionamento linear é forte ou fraco, consoante se situam os pontos em redor da "recta imaginária"que atravessa o conjunto de dados.

A correlação linear é tanto maior quanto mais os pontos se aproximam, com pequenos desvios, dessa recta.

Se o declive da recta for positivo, concluímos que a correlação linear entre as variáveis é positiva, isto é as variávies variam no mesmo sentido, ou seja, quando uma aumenta a outra também aumenta e quando uma diminui a outra também diminui.
Por outro lado, se o declive da recta for negativo, então concluímos que a correlação linear entre as variáveis é negativa, isto é as variávies variam em sentidos contrários, quando uma aumenta a outra diminui e vice-versa, quando uma diminui a outra aumenta.

Coeficiente de correlação linear de Pearson

Um gráfico de dispersão permite pôr em evidência a forma, a direcção, e a intensidade da relação entre duas variáveis quantitativas.
A relação linear é, pela sua simplicidade, particularmente importante.

Apesar de bastante intuitiva, é por vezes dificil dizer quando é que um par de variáveis revela uma maior associação que outro. Por outro lado, esta análise é bastante subjectiva, dependendo, inclusivé, da escala usada no gráfico de dispersão.

Torna-se assim importante quantificar a relação entre as variáveis em estudo. Uma forma simples de quantificar a associação linear entre duas variáveis quantitativas, é através do coeficiente de correlação linear de Pearson.

Coeficiente de correlação linear de Pearson

O coeficiente de correlação linear de Pearson (representado por ρ) mede o grau de associação linear entre duas variáveis quantitativas e o seu valor varia entre -1 e 1.

$$\rho = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}$$

Usualmente desconhecido, o seu valor é estimado utilizando o coeficiente de correlação amostral de Pearson, que se representa por r, e é definido da seguinte forma:

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left[n\sum x_i^2 - (\sum x_i)^2\right]\left[n\sum y_i^2 - (\sum y_i)^2\right]}}$$

O coeficiente de correlação linear de Pearson permite avaliar o sentido e intensidade da relação linear entre as variáveis.

Coeficiente de correlação linear de Pearson

Como interpretar o coeficiente de correlação de Pearson?

- O sinal do coeficiente está directamente relacionado com o tipo de correlação (positiva ou negativa) existente entre as duas variáveis. Assim:
  - se o $r > 0$ então existe uma correlação linear positiva entre as variáveis
  - se o $r < 0$ então existe uma correlação linear negativa entre as variávies
  - se $r = 0$ então não existe relação linear entre as duas variáveis.
- o valor do coeficiente indica a intensidade da relação linear entre as duas variáveis:
  quanto mais próximo de 1 for o valor absoluto do coeficiente, mais intensa é a relação linear entre as duas variáveis (por outro lado, um valor próximo de zero indica uma fraca associação linear).

Coeficiente de correlação linear de Pearson

Para facilitar a interpretação do coeficiente de correlação, podemos caracterizar a intensidade da relação de acordo com:

- se $0 \leq |r| \leq 0.25$ dizemos que a associação linear é residual ou inexistente
- se $0.25 < |r| \leq 0.5$ dizemos que a associação linear é fraca
- se $0.5 < |r| \leq 0.75$ dizemos que a associação linear é moderada
- se $0.75 < |r| \leq 0.90$ dizemos que a associação linear é forte.
- se $0.90 < |r| \leq 1$ dizemos que a associação linear é muito forte.

observações:

- o r é uma medida de associação linear entre duas variávis quantitativas; não descreve outro tipo de associações não-lineares.
- Como o cálculo de r se baseia em médias e desvio-padrões, r é sensível a observações discordantes ("outliers") e deve ser por isso usado com cuidado quando o gráfico de dispersão sugerir a presença de tais observações.

Coeficiente de correlação linear de Pearson

Muitas vezes, o que interessa é saber se, na população, existe correlação (não nula) entre as duas variáveis, isto é, saber se as variáveis estão ou não (linearmente) relacionadas.
Um coeficiente de correlação amostral não nulo não garante que seja não-nulo o coeficiente de correlação populacional, mas através dos testes de significância sobre correlações, podemos avaliar se o coeficiente de correlação obtido na amostra é estatisticamente diferente de zero, ou seja, se a correlação observada não se deve a flutuações amostrais mas é sim algo de real na população de onde a amostra foi retirada.

No SPSS, podemos determinar o valor de coeficiente de correlação de Pearson e avaliar a sua signficância através do menu: Analisar $\longrightarrow$ Correlacionar $\longrightarrow$ Bivariável.

Regressão Linear Simples

Uma vez estabelecido que existe uma associação entre as variáveis, é regra geral fundamental estabelecer a função matemática que melhor descreve essa associação.

Neste capítulo apenas se abordará a relação linear entre estas duas variáveis, pelo que o modelo traduz a equação de uma recta. No entanto a relação poderá não ser linear e assumir outras formas, como sejam a quadrática, exponencial, logarítmica entre outras.

Objectivos de uma análise de regressão linear simples:

- Descrever a relação linear existente entre duas variáveis: X e Y
- Estimar(predizer) o valor de uma das variáveis (Y), conhecido o valor da outra (X).
- Medir o impacto(efeito)que uma das variáveis (X) tem na outra (Y).

O Modelo de Regressão Linear Simples

O modelo de regressão linear simples assume a forma,

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

onde:
Y - designa a variável dependente ou explicada (aleatória)
X - represenata a variável independente ou explicativa (não aleatória)
ε - variável aleatória residual que inclui todas as influências no comportamenrto da variável Y que não podem ser explicadas linearmente pelo comportamento da variável X
β_0 e β_1 - são os parâmetros desconhecidos do modelo (a estimar)

O Modelo de Regressão Linear Simples

Numa análise de regressão e correlação simples, os dados são da forma (x_i, y_i), onde x_i é o valor observado da variável X par o indivíduo i e y_i a observação correspondente da variável aleatória Y_i.

Num modelo de regressão linear simples assume-se que cada observação satisfaz a relação:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

admite-se ainda que os $\varepsilon_i (i = 1, ..n)$ são variáveis aleatórias independentes de média zero e variância constante.

A partir dos dados estimamos β_0 e β_1 e substituímos estes parâmetros pelas suas estimativas (b_0, b_1) para obter a equação de regressão estimada:

$$\widehat{y} = b_0 + b_1 x$$

Esta equação estima o valor médio de Y para um dado valor x de X; no entanto é usada para estimar o próprio valor de Y.

O Modelo de Regressão Linear Simples

Os Parâmetros são estimados utilizando o método dos mínimos quadrados. O objectivo é escolher b_0 e b_1 de modo a minimizar a soma dos quadrados dos resíduos:

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

Assim, para determinar b_0 e b_1, resolve-se o seguinte sistema de equações lineares:

$$\begin{cases} \frac{\partial}{\partial b_1} \sum \varepsilon_i^2 = \frac{\partial}{\partial b_1} \sum (y_i - b_0 - b_1 x_i)^2 = 0 \\ \frac{\partial}{\partial b_2} \sum \varepsilon_i^2 = \frac{\partial}{\partial b_2} \sum (y_i - b_0 - b_1 x_i)^2 = 0 \end{cases}$$

O Modelo de Regressão Linear Simples

A resolução deste sistema permite obter as equações que são utilizadas para estimar os coeficientes b_0 e b_1.

$$\begin{cases} b_1 = \dfrac{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2} \\ b_0 = \overline{Y} - b_1\overline{X} \end{cases}$$

Observações: (a desenvolver nas aulas)

- o problema de extrapolar para fora do âmbito dos dados amostrais.....
- a relação causa-efeito .....
- a existência de outliers ....

Coeficiente de determinação

O Coeficiente de determinação: R^2 ([1])

Este coeficiente pode ser usado como uma medida da qualidade do ajustamento, ou seja, como uma medida da confiança depositada na equação de regressão como instrumento de previsão.
Quando $R^2 = 0$ o modelo claramente não se ajusta aos dados, por outro lado quando $R^2 = 1$ o ajustamento é perfeito.

O valor de R^2 que se considera produzir um ajustamento adequado é algo subjectivo. Em geral considera-se que:

- se $R^2 > 0.9$ temos um bom ajustamento, grande parte da variação de Y é explicada linearmente pela variável independente e portanto o modelo "é bastante capaz" de prever Y (temos estimativas de muito boa qualidade)
- se $R^2 < 0.5$ o ajustamento é mau; o modelo linear é muito pouco adequado e consequentemente as estimativas são de fraca qualidade.

[1] No caso da RL simples, o valor de R^2 é igual ao quadrado do coeficiente de correlação linear de pearson entre as variáveis X e Y. Tem-se que $0 \leq R^2 \leq 1$.