Modelos Lineares
Covariância e Correlação

Susana Faria

Notas Iniciais

- **O uso destas notas como único material de estudo é fortemente desaconselhado.**
- **Neste capítulo começa-se por estudar a relação entre uma variável resposta Y e uma variável independente X.**
  De seguida, apresentam-se a covariância e o coeficiente de correlação como medidas do grau de associação entre as duas variáveis.

Covariância e Correlação

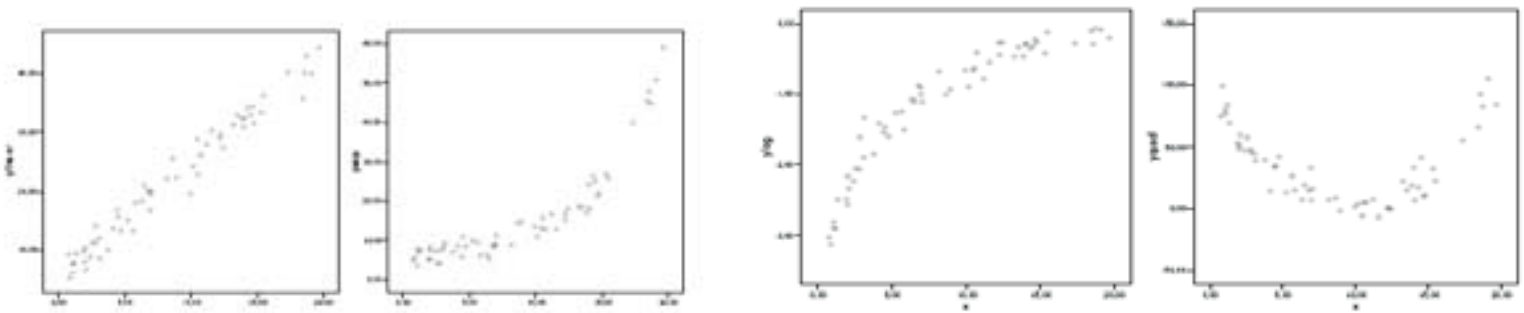
Considere que temos n observações:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_n, y_n)$$

provenientes de duas variáveis quantitativas, X e Y, e pretende-se medir o grau de associação entre essas variáveis.

O método mais simples para averiguar a relação entre a variável independente X e a variável resposta Y é construir o **diagrama de dispersão**. Este gráfico sugere-nos qual a função mais adequada a ajustar aos dados.

Formas de associação entre variáveis numéricas: lineares, exponenciais, logarítmicas ou quadráticas.



Revisão

A variância duma v.a. define-se como:

$$V[X] = E[(X - E[X])^2] = E[X^2] - E^2[X]$$

Sejam X e Y variáveis aleatórias e a e b constantes. Então:

- $E[X + a] = E[X] + a$.
- $E[bX] = bE[X]$.
- $E[X \pm Y] = E[X] \pm E[Y]$
- $V[X + a] = V[X]$
- $V[bX] = b^2 V[X]$.
- $V[X \pm Y] = V[X] + V[Y] \pm 2Cov[X, Y]$, onde Cov[X,Y] é a covariância de X e Y.
- Se X e Y forem v.a. independentes, então $V[X \pm Y] = V[X] + V[Y]$.

Covariância

A covariância entre Y e X:

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

indica a direcção da relação linear entre X e Y.

- $Cov(Y, X) > 0$ associação linear positiva;
- $Cov(Y, X) < 0$ associação linear negativa;
- $Cov(Y, X) = 0$ ausência de associação linear.

Propriedades:

- $Cov(X, Y + Z) =$
- $Cov(X, \alpha) =$
- $Cov(X, \alpha + \beta Y)$
- $Cov(X, X) =$

Desvantagem: Depende das unidades em que as variáveis são expressas.

Covariância amostral

Considere uma amostra bivariada $(x_1, y_1), \ldots, (x_n, y_n)$, a covariância amostral é:

$$cov(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\overline{x}\ \ \overline{y}\right)$$

Exercicio: Determine a covariância entre a altura e o peso dos individuos na seguintes situações:

altura(m)	peso (kg)
1.74	68
1.83	80
1.68	62
1.89	89
1.78	70
1.83	78

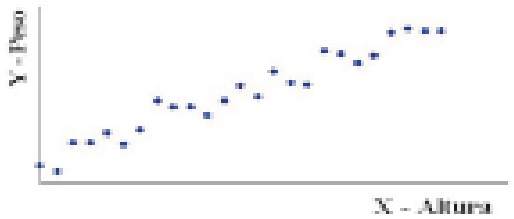
altura(polegadas)	peso (libras)
68.5	149.9
72	176.4
74.4	196.2
72	172
70.1	154.3
66.1	136.7

Covariância: Indicador da Associação Linear

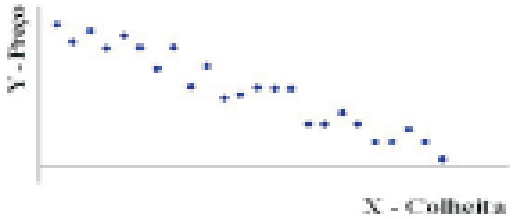
- Associação Linear Crescente: é de esperar que indivíduos com altura abaixo da média tenham peso abaixo da média e que indivíduos com altura superior à média tenham peso superior à média.

- Associação Linear Decrescente: por exemplo, preço de um quilo de morangos e quantidade transaccionada no mercado abastecedor nesse dia, é de esperar que nos dias em que a oferta está acima da média o preço desça abaixo do preço médio, e vice-versa.

Sugerem uma regressão linear
(i.e., a relação entre as duas variáveis poderá ser descrita por uma equação linear)



Existência de correlação positiva (em média, quanto maior for a altura maior será o peso)



Existência de correlação negativa (em média, quanto maior for a colheita menor será o preço)

Coeficiente de Correlação

O coeficiente de correlação de Pearson:

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X)\ Var(Y)}}$$

permite avaliar o grau de associação linear entre duas variáveis.

- O sinal de ρ indica a direcção da associação;
- O valor absoluto de ρ mede a intensidade da associação;
- $-1 \leq \rho \leq 1$:
  - $\rho > 0$ relação linear positiva;
  - $\rho = 0$ ausência de relação linear;
  - $\rho < 0$ relação linear negativa;
  - $\rho = 1$ relação linear positiva perfeita;
  - $\rho = -1$ relação linear negativa perfeita.

Coeficiente de Correlação

O coeficiente de correlação amostral de Pearson:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

permite avaliar o grau de associação linear entre duas variáveis.

Notações Alternativas:

O coeficiente de correlação amostral de Pearson:

$$r = \frac{cov(X, Y)}{s_x \; s_y}$$

onde s_x e s_y são o desvio padrão amostral de X e Y, respectivamente.

Designando por:

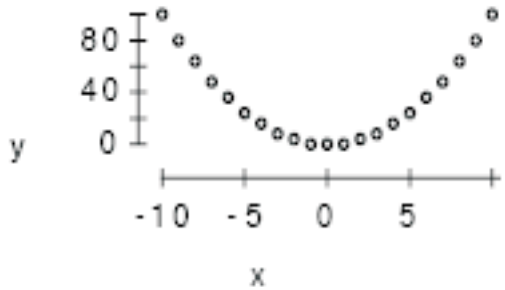
$$S_{XX} = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad S_{YY} = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

e $$S_{XY} = \sum_{i=1}^{n} x_i \; y_i - n\overline{x} \; \overline{y}$$

$$r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}$$

Más Interpretações do Coeficiente de Correlação

- O coeficiente de correlação igual a zero não significa que as variáveis não estejam associadas.



- O coeficiente de correlação pode ser influenciado por um ou mais outliers
- ver site: *http* : *//www.tylervigen.com/spurious − correlations*

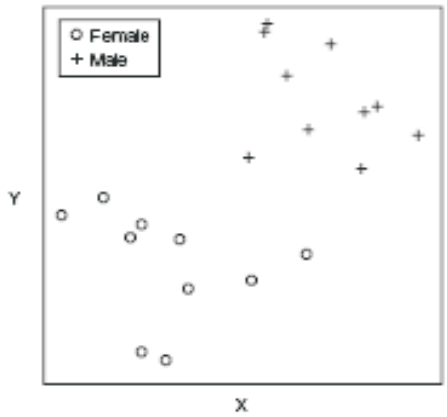
Coeficiente de Correlação

- Uma correlação forte não significa necessariamente uma relação de causa e efeito entre as variáveis.

  Exemplos:

  - Uma publicação anticlerical que ficou célebre mostrava claramente que o número de crimes nas cidades inglesas tinha crescido com o aumento dos pastores anglicanos, durante o século XIX.
  - O número de prédios destruídos em cada fogo urbano e o número de viaturas de bombeiros utilizados no combate ao mesmo.

- Calcular o coeficiente de correlação numa população não homogénea (por exemplo, constituída por homens e mulheres) pode conduzir a más interpretações.



Coeficiente de Correlação

- Se X e Y são independentes então $Cov(X, Y) = 0$ e consequentemente $\rho = 0$. Mas, atenção, o recíproco é falso!

- No entanto, se X e Y são independentes e seguem uma distribuição normal então verifica-se a equivalência, ou seja, X e Y independentes e normais $\iff \rho = 0$

Exercicio: Determine a correlação nas seguintes situações:

x_i	$y_i = 2x_i + 3$	$z_i = x_i^2$
-3		
-2		
-1		
0		
1		
2		
3		

Testar a Correlação

Dado X e Y duas variáveis aleatórias de distribuição normal, testar:

$$H_0 : \rho = 0 \quad \text{vs} \quad H_1 : \rho \neq 0$$

A estatística-teste é:

$$T = \frac{\sqrt{n-2}r}{\sqrt{1-r^2}} \quad \sim \quad t_{n-2}$$

A região de rejeição é:

$$RC = \{t : |t| > t_{\frac{\alpha}{2};n-2}\} \text{ em que } \alpha \text{ é o nível de significância.}$$

Exercicio: Testar a hipótese de não existir correlação entre a altura e o peso dos individuos.

Testar a Correlação

- Se estivermos perante duas variáveis medidas apenas numa escala ordinal, ou que apresentam uma relação não linear mas monótona, o coeficiente de Pearson não pode ser aplicado.
- Se as variáveis X e Y não seguem uma distribuição normal

utiliza-se o Coeficiente de Correlação de Spearman:

$$r_S = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

em que d_i são as diferenças entre as ordens de x_i e y_i.

Nota: No cálculo deste coeficiente, começa-se por ordenar para cada variável as observações e atribui-se, a cada observação, um número indicando a sua posição relativa na ordenação.

$H_0 : X e Y$ são independentes vs $H_1 : X e Y$ não são independentes

A estatística-teste é:

$$z = \sqrt{n-1}|r_S| \quad \sim \quad N(0,1)$$

A região de rejeição é:

$RC = \{z : |z| > z_{\frac{\alpha}{2}}\}$ em que α é o nível de significância.

Testar a Correlação

Exercício: Um individuo atribuiu uma nota de qualidade a 10 perfumes. Na tabela seguinte apresentam-se os índices de qualidade(X) definidos de 1 a 10 (sendo 10 o melhor perfume) e o preço (Y) dos 10 perfumes:

10	1	2	5	4	3	6	7	9	8
95	60	52,5	51,5	49,5	47,5	55	48	56	53

Podemos afirmar que o preço dos perfumes depende da qualidade?

O coeficiente de correlacção de Kendall

- Uma alternativa ao coeficiente de Spearman é o coeficiente de Kendall que se aplica nas mesmas condições;
- Se as amostras tiverem dimensão muito reduzida e valores repetidos, os resultados do teste ao coeficiente de correlação de Kendall são mais precisos;
- O coeficiente de Kendall pode ser generalizado para correlações parciais que são correlações medidas entre duas variáveis após remoção do efeito de uma possível terceira variável sobre ambas.
- Uma diferença muito importante entre os dois coeficientes (Kendall e Spearman) reside na sua interpretação e na impossibilidade de comparar directamente valores provenientes de ambos.
- O coeficiente de Kendall é muitas vezes descrito como uma medida de concordância entre dois conjuntos de classicações relativas a um conjunto de objectos ou experiências.

$$T = \frac{\text{número concordâncias} - \text{número discordâncias}}{\text{número total de pares}}$$

- Tal como para os coeficientes de Pearson e Spearman é possível efectuar um teste de hipóteses para averiguar se a associação é significativa.