INSTITUTO POLITÉCNICO DE TOMAR
Escola Superior de Tecnologia de Tomar
Unidade Departamental de Matemática e Física

Curso: **Mestrado em Analítica e Inteligência Artificial**
Unidade Curricular: **Modelos Matemáticos de Análise e de Apoio à Decisão**

Sessão: 7 (duração: 2 horas)
Ano lectivo: 2016/2017 (2.º Semestre)
Docente: Luís Miguel Grilo

Assunto: Correlação e Regressão Linear Simples

Sumário: Correlação linear de Pearson, regressão Linear Simples (diagrama de dispersão, método dos mínimos quadrados, pressupostos do modelo, coeficiente de determinação, inferência sobre os parâmetros).

Em Estatística estamos, por vezes, interessados em estudar a natureza da relação entre variáveis, sendo que esta pode ser determinística ou probabilística. Assim, designamos uma relação entre duas variáveis X e Y por determinística se existir uma relação funcional entre elas, isto é, o valor de Y é completamente determinado para um dado valor de X. A relação entre as duas variáveis é probabilística se a relação é imprecisa, ou seja, o valor y, da variável aleatória de interesse Y, não é unicamente determinado para um dado valor x de X. Com a correlação linear é possível analisar relações de linearidade entre duas variáveis e a regressão linear é uma metodologia estatística que permite estabelecer relações entre a variável de interesse (variável dependente ou explicada) e uma ou mais variáveis (independentes ou explicativas), sendo possível prever o comportamento da variável dependente à custa da(s) independente(s), sob determinadas condições.

Pretendemos estudar a relação entre duas variáveis quantitativas, usando diferentes métodos de medição do grau de associação entre as duas variáveis. Podemos percorrer as seguintes fases (em posse dos dados):

1.ª fase – Verificar se existe uma relação de causa-efeito entre as duas variáveis que pretendemos estudar (quando a variação de uma variável pode ser atribuída à variação de outra variável, caso contrário a aplicação de um método de regressão ou correlação não tem qualquer utilidade), com o objectivo de evitar a falácia do *post-hoc* (atribuição da existência de um nexo de causalidade entre duas variáveis apenas contemporâneas) e identificar a variável dependente/explicada/resposta/endógena (cujo comportamento se desconhece, Y_i) e a independente/explicativa/estímulo/exógena (cujo comportamento é conhecido/controlado, X_i). Esta fase depende do conhecimento de outras áreas, como a Economia, a Gestão, a Psicologia, a Sociologia, etc.;

2.ª fase – Analisar o tipo de relação que existe entre as variáveis, nomeadamente, esboçando o diagrama de dispersão dos dados observados, que nos dá uma ideia da relação (positiva ou negativa, linear ou não) entre as duas variáveis;

3.ª fase – Utilizar medidas numéricas (nomeadamente o Coeficiente de Correlação Linear de Pearson e o Coeficiente de Determinação) para averiguar o grau de associação ou correlação entre as variáveis, sendo que nenhum dos dois indica se existe uma relação de causa-efeito entre as variáveis;

4.ª fase – Usar um método objectivo (no nosso caso, o Método dos Mínimos Quadrados - MMQ) para ajustar um modelo linear (recta) aos dados, sendo que podemos procurar avaliar a qualidade do ajustamento linear com recurso ao Coeficiente de Determinação. Nesta fase, podemos verificar que a soma dos resíduos é nula;

5.ª fase – Inferência sobre os parâmetros do modelo (intervalos de confiança e testes de hipóteses aos parâmetros);

6.ª fase – Utilizar o modelo estimado pelo MMQ para fazer previsões, caso o modelo "passe" nos testes da fase anterior.

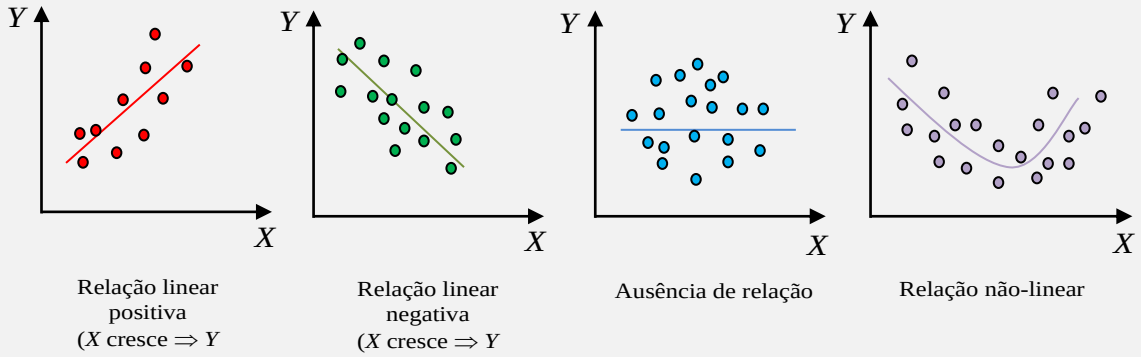
7.1 DIAGRAMA DE DISPERSÃO

Diagrama de Dispersão – gráfico onde representamos os pares ordenados $(x_i; y_i)$ no eixo cartesiano (considerando a variável independente em abcissas e a dependente em ordenadas), com intuito de:

- averiguar se existe alguma relação entre as variáveis;
- identificar que tipo de relação existe entre as variáveis – linear (recta) ou não linear (exponencial, logarítmica, potência, logística, etc.) – ou seja, identificar a equação mais apropriada para descrever a relação entre as variáveis. A relação mais simples é a linear, sendo possível linearizar muitas das relações não lineares descritas, pelo que reduziremos o nosso estudo ao caso linear;
- identificar o sentido da relação entre as variáveis, o qual pode ser positivo (X e Y variam no mesmo sentido, o que significa que se X cresce/decresce, Y cresce/decresce, respetivamente), negativo (X e Y variam em sentido contrário, isto é se X cresce/decresce então Y decresce/cresce, respectivamente) ou nulo (quando X cresce ou decresce, Y não se altera).

De notar que a cada valor de *X* podem corresponder um ou mais valores de *Y* diferentes.

A título de exemplo, podemos ter as seguintes relações diferentes entre as variáveis (as relações são perfeitas quando todos os pontos/observações "caem" sobre as "curvas de ajustamento" lineares ou não-lineares, e têm uma associação tanto mais forte, quanto mais próximos estiverem os pontos das "curvas"):



Relação linear positiva (*X* cresce ⇒ *Y*

Relação linear negativa (*X* cresce ⇒ *Y*

Ausência de relação

Relação não-linear

7.2 MODELO DE REGRESSÃO LINEAR SIMPLES. MÉTODO DOS MÍNIMOS QUADRADOS

Modelo de Regressão Linear Simples (**MRLS**) – é um modelo matemático de regressão (porque relaciona uma variável de interesse com uma ou com um conjunto de variáveis explicativas) que é linear (porque ajusta uma recta às observações, sendo que existem modelos não-lineares) e é simples (porque considera uma única variável explicativa – caso considerasse mais do que uma teríamos um modelo de regressão linear múltipla).

Especificação do Modelo de Regressão Linear Simples – admitindo que a teoria (Economia, Gestão, Sociologia, Psicologia, Medicina, Engenharia, etc.) considera que uma determinada variável *Y* é influenciada de uma certa forma pela variável *X*, temos analiticamente,

$$Y = f(X),$$

onde $f(\cdot)$ pode ser uma função linear, exponencial, logarítmica, de potência, logística, etc.. Deste modo, após identificarmos uma relação de causa-efeito entre as duas variáveis, há que especificar a função $f(\cdot)$. Vamos admitir que a relação é linear, ou seja, analiticamente, temos a equação (da recta):

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

(pois podemos melhor descrever a relação entre ambas as variáveis se, em vez de considerarmos apenas o diagrama de dispersão ajustarmos uma recta aos dados observados – ver últimos gráficos), mas como sabemos que os fenómenos em estudo são

sempre muito complexos, certamente não existe uma relação perfeita entre *X* e *Y* (ou seja, *Y* não depende exclusivamente de *X*, pois, na prática, há outras variáveis que influenciam *Y* para além de *X*), pelo que esta equação não estará inteiramente correcta, sendo necessário acrescentar o chamado termo **erro** ao modelo (*e*),

$$Y = \beta_0 + \beta_1 X + e\,,$$

onde:

- *Y* representa a variável dependente, explicada, endógena, regressando, de interesse ou de resposta;
- *X* representa a variável independente, explicativa, exógena, regressor ou de estímulo;
- *e* representa o termo erro que é não observável (inclui os erros de medição e todas as variáveis que explicam a variação de *Y*, para além de *X*, que não são incluídas no modelo ou porque não as conseguimos medir ou porque as consideramos pouco importantes na explicação da variável dependente – variáveis observáveis omitidas – já que pretendemos modelos o mais parcimoniosos possível);
- β_0 e β_1 são parâmetros do modelo (ou coeficientes da regressão) os quais são sempre desconhecidos (daí que o principal objectivo deste capítulo seja estimar os seus valores com a maior precisão possível, isto é, calcular um valor para estes parâmetros tão próximo quanto possível dos verdadeiros valores).

  β_0 representa o parâmetro associado à constante/regressor (1) que nos dá o valor esperado para a variável dependente quando a variável independente é nula: $X = 0 \Rightarrow Y = \beta_0$ (ou seja, é a ordenada na origem). Por vezes o seu valor tem pouco interesse, mas a sua presença é extremamente importante para a obtenção de determinados resultados matemáticos úteis.

  β_1 é o parâmetro associado à variável explicativa (*X*) e dá-nos o efeito provocado em *Y* dada uma variação unitária em *X*, isto é, o acréscimo da variável dependente sempre que a variável independente varia uma unidade (em termos discretos): $\frac{\Delta Y}{\Delta X} = \beta_1 \Leftrightarrow \Delta Y = \beta_1 \cdot \Delta X$, logo se $\Delta X = 1 \Rightarrow \Delta Y = \beta_1$ (trata-se de um efeito *cœteris paribus* pois assume-se que todas as outras variáveis que influenciam *Y* se mantêm constantes, ou seja, $\Delta e = 0$). Deste modo, indica-nos o tipo de relação (positiva ou negativa) e a magnitude dessa relação, daí o principal objectivo ser a sua estimação. Analiticamente, dá-nos o declive da recta, pelo que se $\beta_1 > 0$ (declive é positivo) a recta é crescente, se $\beta_1 < 0$ (declive negativo) a

recta é decrescente e se $\beta_1 = 0$, o declive da recta é nulo, significando que *Y* e *X* são independentes).

A estimação deste modelo só conduz a resultados correctos se se verificar um determinado conjunto de pressupostos (que resultam da formulação do modelo). Um dos pressupostos mais importantes é o de que $E(e|X) = 0$, o que significa que o termo erro tem de ser independente da variável independente *X* (não pode existir qualquer relação entre *e* e *X*, caso contrário não será possível manter *e* constante quando variamos *X*, pois $\Delta Y = \beta_1 \Delta X + \Delta e$, admitindo os parâmetros constantes).

Após a especificação do modelo, passamos à fase da recolha dos dados que, posteriormente, permitem estimar os parâmetros do modelo. Vamos admitir que é recolhida uma amostra aleatória de dimensão *n*, isto é, para *n* indivíduos escolhidos ao acaso, são recolhidos dados relativos às variáveis *Y* e *X*. O modelo que queremos estimar pode, então ser escrito da seguinte forma,

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, 2, ..., n.$$

A questão consiste em saber como podemos estimar os parâmetros do modelo (β_0 e β_1) a partir das *n* observações que recolhemos para *Y* e *X*, sabendo que não é possível (ou não considerámos necessário) recolher dados relativamente a *e*.

Nunca será possível conhecermos com exactidão os valores dos parâmetros β_0 e β_1, o que é possível é tentar encontrar **estimadores** para esses parâmetros que se situem tão perto quanto possível dos verdadeiros valores (designaremos $\hat{\beta}_0$ e $\hat{\beta}_1$ os estimadores de β_0 e β_1) e uma vez obtidos $\hat{\beta}_0$ e $\hat{\beta}_1$, é possível prever o valor de *Y*, para cada indivíduo, através da recta da regressão (modelo estimado):

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, 2, ..., n.$$

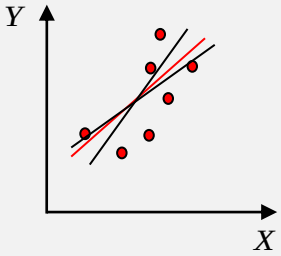
Ignoramos o termo erro ajustado (resíduo, $\hat{e}_i$), porque em média ele é nulo, se o modelo estiver bem especificado.

Deste modo, para cada valor da variável explicativa (x_i), temos dois valores da variável explicada (um real, y_i, e outro observado, $\hat{y}_i$). À diferença entre o valor observado para a variável dependente (*Y*) e o valor estimado da variável dependente ($\hat{y}_i$) chamamos **resíduo** ($\hat{e}_i$) que, basicamente corresponde ao erro cometido na estimação do valor de *Y*, para cada indivíduo (diferente do termo erro que é não observável),

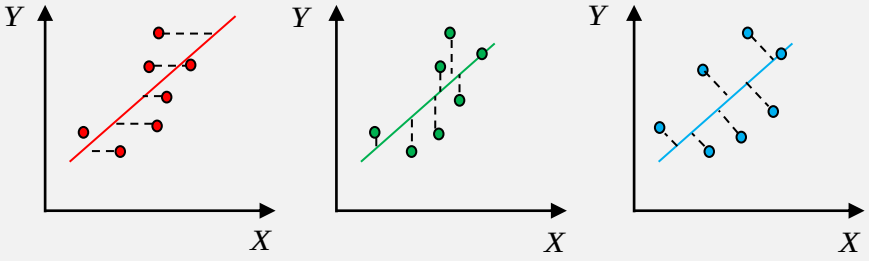
$$y_i = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_i}_{\hat{y}_i} + \hat{e}_i \Leftrightarrow \hat{e}_i = y_i - \hat{y}_i, \quad i = 1, 2, ..., n.$$

Logicamente que qualquer método de estimação tentará fazer com que os erros cometidos sejam tão pequenos quanto possível, para tal existem vários métodos, mais ou menos objectivos:

▪ Métodos subjectivos – consistem em traçar uma recta ajustada com uma régua aos dados. É um método subjectivo, uma vez que cada um de nós pode considerar uma recta, ainda que ligeiramente diferente, que nos parece o melhor ajustamento:



▪ Métodos objectivos – consistem em minimizar a soma das distâncias dos pontos observados à recta, as quais podem ser medidas na horizontal, na vertical ou na perpendicular:



Distâncias medidas na horizontal | Distâncias medidas na vertical | Distâncias medidas na perpendicular

Estimação dos Parâmetros da Regressão pelo Método dos Mínimos Quadrados (MMQ) – é um dos métodos objectivos mais simples e mais usado para ajustar uma recta aos dados ($\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$), o qual consiste em minimizar a soma do quadrado dos resíduos, ou seja, em minimizar a soma do quadrado das distâncias verticais entre os valores observados e a recta ajustada (não se minimiza apenas a soma dos resíduos, porque este valor é sempre zero), que, como qualquer outro problema de optimização, implica o cálculo das condições de 1.ª e de 2.ª ordem:

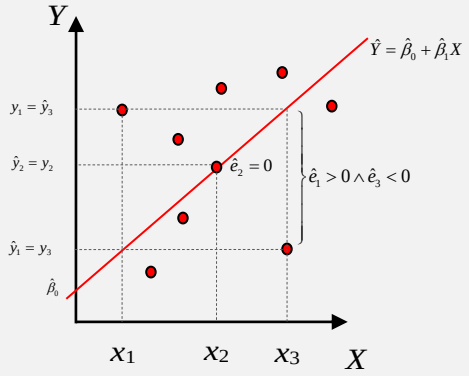
$$\min\left[\sum_{i=1}^{n}\hat{e}_i^2 = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2\right] \Rightarrow \begin{cases} \bullet \text{ Condições 1.ª ordem} \begin{cases} \dfrac{\partial \sum_{i=1}^{n}\hat{e}_i^2}{\partial \beta_0} = 0 \\ \dfrac{\partial \sum_{i=1}^{n}\hat{e}_i^2}{\partial \beta_1} = 0 \end{cases} \Leftrightarrow \begin{cases} \hat{\beta}_1 = \dfrac{s_{xy}}{s_x^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x} \end{cases} \\ \bullet \text{ Condições 2.ª ordem} \\ \text{(determinante Hessiano com as derivadas de 2.ª ordem)} \end{cases}$$

De notar que o declive da recta é dado pelo rácio entre a covariância das duas variáveis e a variância da variável independente $\left(\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{\frac{\sum_i x_i y_i}{n} - \bar{x} \cdot \bar{y}}{\frac{\sum_i x_i^2}{n} - \bar{x}^2} \right)$ que, para ser definido, obriga a que $s_x^2 \neq 0$.

Todavia, existem outras expressões equivalentes para o cálculo deste parâmetro, por exemplo:

$$\hat{\beta}_1 = \frac{SQ_{xy}}{SQ_{xx}}, \text{ onde } SQ_{xy} = \sum_i x_i y_i - \frac{1}{n} \sum_i x_i \sum_i y_i \quad \text{e} \quad SQ_{xx} = \sum_i x_i^2 - \frac{1}{n} \left(\sum_i x_i \right)^2.$$

Com a recta da regressão, a relação linear entre as duas variáveis torna-se perfeita, porque se anulam todos os efeitos da variável residual. Sabendo que, para cada valor da variável explicativa (x_i), temos dois valores da variável explicada (um real/observado, y_i, e outro estimado, $\hat{y}_i$, dado pelo modelo teórico), a recta da regressão sobrestima os valores de *Y* para pontos abaixo da recta da regressão, com resíduos negativos ($y_3 < \hat{y}_3 \Rightarrow \hat{e}_3 < 0$) e subestima os valores de *Y* correspondentes a valores observados acima da recta da regressão, com resíduos positivos ($y_1 > \hat{y}_1 \Rightarrow \hat{e}_1 > 0$), enquanto para pontos sobre a recta da regressão se tem $y_2 = \hat{y}_2 \Rightarrow \hat{e}_2 = 0$ (o valor estimado "acerta" no valor efectivamente observado para *Y*):



Propriedades Algébricas dos Estimadores do MMQ

- A soma dos resíduos é nula (logo o erro cometido na estimação é, em média, nulo): $\sum_{i=1}^{n} \hat{e}_i = 0$.
- Os resíduos são independentes dos regressores: $\sum_{i=1}^{n} x_i \hat{e}_i = 0$.
- O ponto $(\bar{x}; \bar{y})$ situa-se sempre sobre a recta da regressão.
- A média do valor observado para *Y* é igual à respectiva média do valor estimado: $\bar{\hat{y}} = \bar{y}$.

- $\sum_{i=1}^{n} \hat{y}_i \hat{e}_i = 0$.
- $\sum_{i=1}^{n} \hat{y}_i^2 = \sum_{i=1}^{n} y_i \hat{y}_i$.

Pressupostos do Modelo

- Linearidade nos parâmetros (diferente de linearidade nos regressores 1 e *X*) – a verdadeira relação entre *X* e *Y* será dada por $Y = \beta_0 + \beta_1 X + e$ e os coeficientes da regressão (β_0 e β_1) não estão sujeitos a restrições, ou seja, não existe informação a priori sobre eles (sob a forma de igualdades ou desigualdades).
- Ausência de autocorrelação – a amostra tem que ser aleatória (recolhida ao acaso) por forma a que as observações sejam i.i.d.; assim, o MMQ não deve ser aplicado a dados temporais do modo anteriormente descrito (por não cumprirem este pressuposto).
- Exogeneidade – o valor esperado da variável residual condicionado pela matriz dos regressores é nulo, $E(e|X) = 0$. Nestas condições, dizemos que os regressores são exógenos o que significa que não há associação (linear ou outra) entre cada X_i e e_i, pelo que a relação funcional entre *Y* e os regressores está correctamente especificada. De notar que, no caso do MRL simples temos, apenas, dois regressores (1 e *X*).

- Ausência de multicolinearidade exacta – a variável independente não pode ser constante (caso contrário será obtida por combinação linear da constante), isto é, a característica da matriz **X** é igual ao número de coeficientes da regressão (*p*) e inferior à dimensão da amostra (*n*), logo os regressores são linearmente independentes (caso contrário há multicolinearidade) – a diferença entre *n* e *p* (*n* − *p*) dá-nos o número de graus de liberdade (que nos indica que não podemos ter um número de regressores maior que a dimensão da amostra, sendo que neste caso de RLS temos *p* = 2).

Se estes quatro pressupostos se verificarem, os estimadores do MMQ são centrados (ou não enviesados), ou seja, qualquer que seja a dimensão da amostra, o estimador obtido será, em média, igual ao verdadeiro parâmetro: $E(\hat{\beta}_0) = \beta_0$ e $E(\hat{\beta}_1) = \beta_1$. Basta que um dos pressupostos não se verifique, para que os estimadores não sejam centrados (sendo o terceiro o mais problemático).

Portanto, se recolhêssemos uma série de amostras, os estimadores obtidos seriam diferentes em cada amostra, mas seriam distribuídos em torno do verdadeiro valor dos parâmetros, pelo que também é importante saber qual a dispersão da distribuição dos estimadores de modo a saber o quão distante estes podem estar da sua média. A medida

de dispersão que se usa para tal é a variância, que se pode calcular com base nos quatro pressupostos anteriores (obtendo uma expressão complicada) ou acrescentando um quinto pressuposto:

- Homocedasticidade – indica que, com dados seccionais e amostra casual, a variância do erro é constante $V(e|X)=\sigma^2$ (caso contrário existe heterocedasticidade). Neste caso, as variâncias dos estimadores do MMQ são dadas por:

$$V(\hat{\beta}_0)=\sigma^2_{\hat{\beta}_0}=\frac{\sigma^2\sum_{i=1}^{n}X_i^2}{n\sum_{i=1}^{n}(X_i-\bar{X})^2}=\frac{\sigma^2\sum_{i=1}^{n}X_i^2}{n\,SQ_{xx}} \quad \text{e} \quad V(\hat{\beta}_1)=\sigma^2_{\hat{\beta}_1}=\frac{\sigma^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}=\frac{\sigma^2}{SQ_{xx}},$$

Donde concluímos que:

- quanto maior a variabilidade das variáveis não incluídas no modelo (σ^2), maior $V(\hat{\beta}_1)$, porque é mais difícil estimar β_1 com precisão;
- quanto maior $V(X)$, menor $V(\hat{\beta}_1)$, isto é, quanto maior a amplitude dos valores de *X* incluídos na amostra, maior a probabilidade de estimarmos correctamente o seu efeito sobre *X*;
- os desvios padrão dos coeficientes estimados são obtidos extraindo a raiz quadrada da variância.

Juntamente com os quatro pressupostos anteriores, se este pressuposto de homocedasticidade se verificar, garante que os estimadores sejam eficientes, isto é, os de menor variância entre todos os estimadores centrados, que seria possível obter aplicando outros métodos (como, por exemplo, o Método da Máxima Verosimilhança).

Para podermos estimar as variâncias (e desvios padrão) dos coeficientes da regressão, necessitamos de um estimador para a variância do erro (σ^2).

Como $\sigma^2=V(e|X)=E(e^2|X)$, o estimador natural seria $\hat{\sigma}^2=\frac{\sum_{i=1}^{n}\hat{e}_i^2}{n}$, mas porque se prova que este estimador não é centrado, utilizamos

$$\sigma^2=\frac{\sum_{i=1}^{n}\hat{e}_i^2}{n-p},\ \text{com } p \text{ regressores (neste caso, } p=2).$$

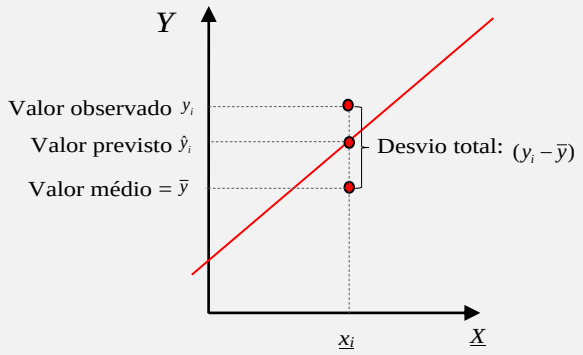
O **erro padrão da regressão** é $\hat{\sigma}=\sqrt{\hat{\sigma}^2}$.

Para a distribuição dos estimadores, obtidos pelo MMQ, com a suposição adicional $e_i\sim N(0,\sigma^2)$, temos, então:

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^{n} X_i^2}{n\, SQ_{xx}}\right) \quad \text{e} \quad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{SQ_{x\,x}}\right).$$

7.3 ANÁLISE DE VARIÂNCIA: ANOVA

A variação total da variável dependente *Y* pode ser divida em componentes com significado estatístico e tratados de forma sistemática. Designamos este método de análise da qualidade do modelo de regressão, em inglês, por "**an**alysis **o**f **va**riance" (ANOVA, abreviadamente).



Decomposição dos desvios: $(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$, ou seja;

desvio total = desvio não explicado + desvio explicado.

Elevando ao quadrado cada uma das diferenças: $\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$, donde,

variação total de *Y* = variação não explicada (erro) + variação explicada (regressão),

que se representa por: *SQT* = *SQE* + *SQR*, onde *SQT* é a soma de quadrados dos desvios totais, e mede a variabilidade total do conjunto de observações de *Y*, com decomposição em duas partes: *SQE* é a soma de quadrados dos resíduos (erros ajustados) e representa a variabilidade do conjunto de observações de *Y* à volta da recta de regressão; *SQR* é soma de quadrados dos desvios explicados pela regressão. Temos, então,

SQT – variabilidade total do conjunto de observações de *Y*;

SQR – parte da variabilidade do conjunto de observações de *Y* que é eliminada quando se usa o conhecimento da variável explicativa *x* para prever *Y*;

SQE – parte da variabilidade do conjunto de observações de *Y* que permanece mesmo quando o conhecimento de *x* é considerado.

Ao dividirmos uma soma de quadrados de desvios pelos correspondentes graus de liberdade temos a média de quadrados de desvios:

$MQR = \frac{SQR}{1} = SQR$ que é a média dos quadrados dos desvios explicados pela regressão;

$MQE = \frac{SQE}{n-2}$ que é a média dos quadrados dos resíduos (estimador da variância do erro, σ^2).

Nota: Ao contrário das somas de quadrados de desvios, as médias de quadrados de desvios não são aditivas, ou seja, $MQR + MQE \neq \frac{SQT}{n-1}$.

Para sintetizarmos a informação anterior utilizamos, habitualmente, a designada tabela da ANOVA:

Origem/fonte de Variação	Soma de Quadrados ***SQ***	Graus de liberdade ***gl***	Média de Quadrados ***MQ***	Estatística de teste ***F***
Regressão	$SQR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$	1	$MQR = SQR / 1 = SQR$	MQR / MQE
Erro	$SQE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	$n - 2$	$MQE = SQE / n - 2$	
Total	$SQT = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$	$n - 1$		

As somas de quadrados podem ser obtidas pelas fórmulas:

$SQR = \hat{\beta}_1 \times SQ_{xy} = \frac{(SQ_{xy})^2}{SQ_{xx}}$; $SQE = SQ_{yy} - \frac{(SQ_{xy})^2}{SQ_{xx}} = SQ_{yy} - \hat{\beta}_1 SQ_{xy}$,

com $SQ_{yy} = \sum_i y_i^2 - \frac{1}{n}\left(\sum_i y_i\right)^2$ e $SQT = SQR + SQE$.

7.4 COEFICIENTES DE DETERMINAÇÃO E DE CORRELAÇÃO

Coeficiente de Correlação Linear de Pearson (*r*) – medida do grau de associação ou correlação linear entre duas variáveis quantitativas (*X* e *Y*), dada pelo rácio da covariância entre *X* e *Y* (s_{xy}) e o produto dos respectivos desvios-padrão (s_x e s_y),

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\frac{\sum_i x_i y_i}{n} - \bar{x} \cdot \bar{y}}{\sqrt{\left(\frac{\sum_i x_i^2}{n} - \bar{x}^2\right) \cdot \left(\frac{\sum_i y_i^2}{n} - \bar{y}^2\right)}} = \frac{SQ_{xy}}{\sqrt{SQ_{xx} SQ_{yy}}}, -1 \leq r \leq 1.$$

De notar que também a covariância (média do produto das diferenças entre os valores de cada variável e a respectiva média) é uma medida do grau de associação linear entre duas variáveis, todavia é preterida relativamente às restantes por ser de difícil interpretação (dado que é expressa nas unidades de medida do produto das duas variáveis, e porque não conseguimos avaliar se se trata de uma covariância reduzida, média ou elevada). Pelo

contrário, o coeficiente de correlação é uma medida relativa, logo não depende das unidades de medida das variáveis (é adimensional) e, dado que varia entre −1 e 1, permite distinguir graus de associação, entre as variáveis, elevados e reduzidos.

Deste modo:

- Se $r=-1 \vee r=1 \Rightarrow$ relação linear perfeita (negativa ou positiva) entre as variáveis;
- Se $0{,}8<r<1 \vee -1<r<-0{,}8 \Rightarrow$ relação linear forte (positiva ou negativa) entre as variáveis (tanto mais forte quanto mais próxima de 1 ou de −1);
- Se $0{,}5<r<0{,}8 \vee -0{,}8<r<-0{,}5 \Rightarrow$ relação linear moderada (positiva ou negativa) entre as variáveis (tanto mais forte quanto mais próxima de 0,8);
- Se $0{,}1<r<0{,}5 \vee -0{,}5<r<-0{,}1 \Rightarrow$ relação linear fraca (positiva ou negativa) entre as variáveis (tanto mais fraca quanto mais próxima de 0,1);
- Se $0<r<0{,}1 \vee -0{,}1<r<0 \Rightarrow$ relação linear ínfima (positiva ou negativa) entre as variáveis (tanto mais fraca quanto mais próxima de 0);
- Se $r=0 \Rightarrow$ ausência de relação linear entre as variáveis.

Podemos, então, dizer que o coeficiente de correlação linear de Pearson, nos indica:

- a intensidade (mais próxima de 0 ou dos valores limite: 1 ou −1) de associação entre as variáveis;
- o sentido (positivo, se $r>0$, ou negativo, se $r<0$) do grau de associação entre as variáveis.

Coeficiente de Determinação (r^2) – medida do poder explicativo da equação de regressão, que nos dá a proporção da variação de *Y* explicada pela variável *X* (ou pela recta da regressão/modelo), logo a restante variação $(1-r^2)\times 100\%$, dá-nos a percentagem da variação de *Y* que não é explicada por *X* (depende de causas aleatórias desconhecidas, isto é, poderá dever-se a variáveis não incluídas no modelo e/ou a erros de medição):

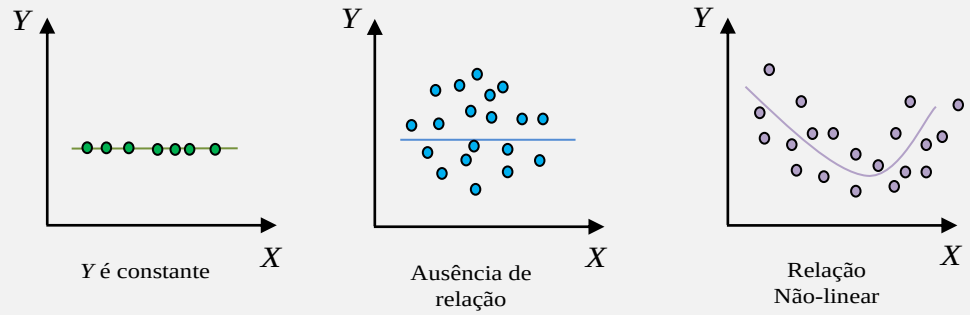
$$r^2=\frac{s_{xy}^2}{s_x^2\cdot s_y^2}=\frac{\left(\frac{\sum_i x_i y_i}{n}-\bar{x}\cdot\bar{y}\right)^2}{\left(\frac{\sum_i x_i^2}{n}-\bar{x}^2\right)\cdot\left(\frac{\sum_i y_i^2}{n}-\bar{y}^2\right)}=\frac{SQR}{SQT}=\frac{SQT-SQE}{SQT}=1-\frac{SQE}{SQT},\quad 0\le r^2\le 1.$$

Este coeficiente também é usado para medir o grau de ajustamento do modelo pelo MMQ e, o facto de ser dado pelo quadrado do coeficiente de correlação, só toma valores positivos, o que significa que nos indica apenas a intensidade de associação entre as variáveis (deixando de indicar o sentido positivo ou negativo da respectiva associação) –

o grau de associação (e a qualidade do ajustamento) é tanto mais forte (melhor) quanto mais próximo estiver r^2 de 1 e tanto mais fraco quanto mais próximo de 0 (quanto mais próximo estiver este coeficiente de 1, maior é a capacidade predictiva do modelo).

De notar que, quando o coeficiente de determinação toma o valor zero, tal pode ficar a dever-se a três causas distintas:

- a variável dependente, qualquer que seja o valor da variável independente, toma sempre o mesmo valor;
- não existe uma relação linear entre as duas variáveis;
- existe uma relação não linear entre as duas variáveis.



De salientar que, um baixo valor do coeficiente de determinação não significa que não exista uma relação importante entre *X* e *Y*, nem que o modelo não é adequado para estimar a relação *cœteris paribus* entre *X* e *Y*, tal só se pode concluir com testes apropriados. Temos, no entanto, a certeza de que, seja *X* importante ou não, há outras variáveis muito mais importantes para explicar o comportamento de *Y*, pelo que com coeficientes de determinação baixos é difícil prever comportamentos individuais.

7.5 INFERÊNCIAS SOBRE OS PARÂMETROS DO MODELO

Para além das estimativas pontuais, obtidas pelo MMQ, para os parâmetros do modelo (β_0 e β_1) podemos fazer inferência sobre os mesmos, através da construção de intervalos de confiança ou de testes hipóteses.

Intervalo de Confiança para β_1

Um intervalo aleatório de confiança de $(1-\alpha)\times 100\%$, para o declive da recta β_1, é dado por

$$IAC_{(1-\alpha)\times 100\%}(\beta_1) = \left[\hat{\beta}_1 \pm t_{(n-2;1-\frac{\alpha}{2})} \times \sqrt{\frac{SQE}{(n-2)SQ_{xx}}} \right],$$

onde $t_{(n-2;\,1-\frac{\alpha}{2})}$ é o percentil $100\times(1-\alpha/2)$ da distribuição $t_{(n-2)}$.

Teste de Hipóteses sobre β_1 (teste à significância do modelo)

O teste de hipóteses adequado, que permite verificar se de facto existe uma relação estatística entre *X* e *Y* (ou seja, se de facto faz sentido efectuar a regressão linear), consiste em testar:

$$H_0: \beta_1 = 0 \quad \text{contra} \quad H_1: \beta_1 \neq 0 \text{ (teste bilateral).}$$

A estatística de teste a usar é: $T = \sqrt{\frac{(n-2)SQ_{xx}}{SQE}} \times \hat{\beta}_1 \underset{\text{Sob } H_0}{\sim} t_{(n-2)}$.

A região crítica é do tipo: $RC =]-\infty, -c] \cup [c, +\infty[$, onde $c = t_{\left(n-2;\ 1-\frac{\alpha}{2}\right)}$.

Se rejeitarmos a hipótese nula estamos, então, convictos que a variação do *Y* não é completamente aleatória, mas que pode, pelo menos parcialmente, ser explicada pelo valor que toma a variável independente.

Teste de Hipóteses sobre β_0 e β_1 (teste à adequabilidade do modelo)

De forma análoga, às inferências sobre o parâmetro β_1, pode haver interesse em realizar inferências sobre β_0 (ordenada na origem da recta de regressão), dado que é, também, um componente da relação probabilística. Todavia, neste ponto, vamos cingir-nos apenas ao teste global do modelo:

$H_0: \beta_0 = \beta_1 = 0$ (o modelo não é adequado) *versus* $H_1: \exists \beta_i \neq 0$ ($i = 0,1$) (o modelo é adequado).

A estatística de teste a usar é: $T = \frac{MQR}{MQE} \underset{\text{Sob } H_0}{\sim} F_{(1;\ n-2)}$.

A região crítica é do tipo unilateral à direita: $RC = [c, +\infty[$ onde $c = F_{(1;\ n-2;\ 1-\alpha)}$.

Se rejeitarmos a hipótese nula significa que há motivo para considerarmos que existe uma relação probabilística entre *X* e *Y*, dependendo a distribuição de *Y* de, pelo menos, uma das componentes desta.

7.6 PREVISÃO DA RESPOSTA

Previsão com a Recta da Regressão – o principal objectivo da recta da regressão é prever o comportamento futuro da variável dependente/resposta com base em valores conhecidos da variável independente. Assim, é possível determinar intervalos de confiança para o valor esperado da resposta e para o valor de uma nova e única resposta, sendo a incerteza, associada à previsão desta, superior à previsão da resposta média.

Exercício 17: A seguinte tabela fornece os dados de uma amostra referente ao número de horas de estudo fora da sala de aula para um curso de estatística com duração de 3 semanas, bem como as classificações obtidas no final do curso.

Estudante	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Horas de estudo, *X*	20	16	34	23	27	32	18	22
Classificação no exame, *Y*	64	61	84	70	88	92	72	77

(a) Calcule a equação de regressão, o coeficiente de correlação e o de determinação. Interprete.

(b) Utilize a equação de regressão para estimar a classificação obtida por um estudante que dedicou 30h de estudo fora da sala de aula.

(c) Construa o quadro ANOVA e calcule $s^2 = \hat{\sigma}^2$ (estimativa da variância do erro, σ^2).

(d) Construa um intervalo de confiança, a 95%, para β_1.

(e) Teste, com α = 5%, a significância do modelo (i.e., teste: H_0: $\beta_1 = 0$ contra H_1: $\beta_1 \neq 0$).

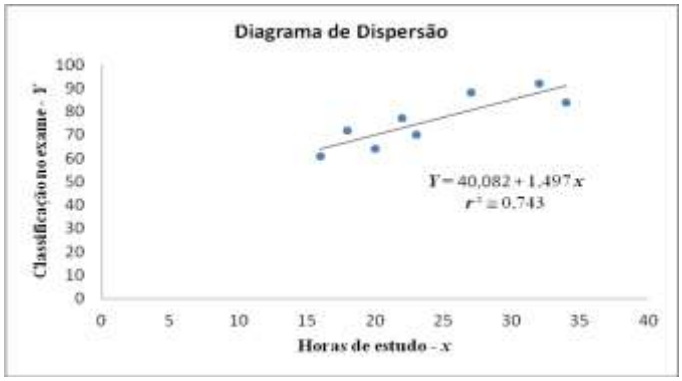
(f) Teste, com um coeficiente de confiança de 95%, a adequabilidade do modelo (ou seja, teste: H_0: $\beta_0 = \beta_1 = 0$ contra H_1: $\exists\, \beta_i \neq 0$, $i = 0, 1$).

Resolução:

(a) Resumimos na tabela seguinte alguns dos cálculos auxiliares,

Horas de estudo - x_i	Classif. no exame - Y_i	x_i^2	y_i^2	$x_i y_i$	$\hat{y}_i$	$\hat{e}_i = y_i - \hat{y}_i$	$\hat{e}_i^2 = (y_i - \hat{y}_i)^2$
20	64	400	4096	1280	70,014	-6,014	36,1682
16	61	256	3721	976	64,0276	-3,0276	…
34	84	1156	7056	2856	90,9664	-6,9664	
23	70	529	4900	1610	74,5038	-4,5038	
27	88	729	7744	2376	80,4902	7,5098	
32	92	1024	8464	2944	87,9732	4,0268	
18	72	324	5184	1296	67,0208	4,9792	
22	77	484	5929	1694	73,0072	3,9928	15,94245
$\sum_i x_i = 192$	$\sum_i y_i = 608$	$\sum_i x_i^2 = 4902$	$\sum_i y_i^2 = 47094$	$\sum_i x_i y_i = 15032$		-0,0032	227,4966

No diagrama de dispersão das observações verificamos que, como se esperava, a classificação tende a aumentar com as horas de estudo. O "alinhamento" do conjunto de pontos da amostra sugere que, na população, a relação é probabilística e linear. Estamos, assim, em presença de uma relação linear positiva, de relativa intensidade, dada a proximidade a uma recta da nuvem de dispersão:



O **Modelo de Regressão Linear Simples** estimado pelo **MMQ** tem a equação: $\hat{y} = 40,082 + 1,497x$.

Com $\hat{\beta}_1 = \dfrac{SQ_{xy}}{SQ_{xx}} = \dfrac{440}{294} \cong 1,497$ e $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 76 - (1,497 \times 24) \cong 40,082$,

em que,

$SQ_{xy} = \sum_i x_i y_i - \frac{1}{n}\sum_i x_i \sum_i y_i = 15032 - \frac{1}{8}(192 \times 608) = 440$, $SQ_{xx} = \sum_i x_i^2 - \frac{1}{n}\left(\sum_i x_i\right)^2 = 4902 - \frac{1}{8}(192)^2 = 294$,

$\bar{x} = \frac{1}{n}\sum_{i=1}^{8} x_i = \frac{192}{8} = 24$ horas e $\bar{y} = \frac{1}{n}\sum_{i=1}^{8} y_i = \frac{608}{8} = 76$ valores.

Interpretação dos parâmetros estimados:

$\hat{\beta}_0 \cong 40,082$ corresponde ao valor da "classificação no exame", quando o valor da variável "horas de estudo" é nulo;

$\hat{\beta}_1 \cong 1,497$ representa a variação esperada da variável "classificação no exame", para cada unidade de variação da variável "horas de estudo".

Coeficiente de correlação linear de Pearson

$r = \frac{SQ_{xy}}{\sqrt{SQ_{xx}}.\sqrt{SQ_{yy}}} = \frac{440}{\sqrt{294 \times 886}} \cong 0,862$ (confirma-se que o sinal de *r* é o sinal de $\hat{\beta}_1$),

onde $SQ_{yy} = \sum_i y_i^2 - \frac{1}{n}\left(\sum_i y_i\right)^2 = 47094 - \frac{1}{8}(608)^2 = 886$.

O valor de $r = 0,862$ confirma as ilações retiradas com o diagrama de dispersão, ou seja, a relação entre as variáveis é positiva (quando *x* cresce, *Y* cresce) e com intensidade relativamente forte, dada a proximidade a 1.

Coeficiente de determinação

$r^2 = \frac{SQR}{SQT} = \frac{SQT - SQE}{SQT} = 1 - \frac{SQE}{SQT} = 1 - \frac{227,497}{886} \cong 0,74$

onde,

$SQE = SQ_{yy} - \frac{(SQ_{xy})^2}{SQ_{xx}} = SQ_{yy} - \hat{\beta}_1 \times SQ_{xy} \cong 227,497$; $SQR = \frac{(SQ_{xy})^2}{SQ_{xx}} = \hat{\beta}_1 \times SQ_{xy} = 1,497 \times 440 \cong 658,503$;

$SQT = SQ_{yy} = 886$ e, em alternativa, $SQT = SQR + SQE \Leftrightarrow SQR = SQT - SQE = 658,503$.

Da variação total de *Y* (de 0 a 100%), 74% é explicada pela presença da variável *x* e pelo ajustamento realizado. Fica por explicar $1 - r^2 = 0,26$ (26%), que se deve a causas desconhecidas ou aleatórias. Podemos, então, considerar que o modelo obtido pelo MMQ apresenta uma qualidade de ajustamento relativamente "boa".

(b) **Previsão**: $x_0 = 30$ horas $\Rightarrow \hat{y} = 40,082 + (1,497 \times 30) \cong 85$ valores

(c) **Tabela da ANOVA**

Origem de Variação	Soma de Quadrados - ***SQ***	Graus de liberdade - ***gl***	Média de Quadrados - ***MQ***	Valor da Estatística - ***F***
Regressão	$SQR \cong 658{,}503$	1	$MQR = SQR / 1 \cong 658{,}503$	$\boldsymbol{F} = MQR / MQE \cong \cong 17{,}3674$
Erro	$SQE \cong 227{,}497$	$n - 2 = 6$	$MQE = SQE / n - 2 \cong 37{,}916$	
Total	$SQT = 886$	$n - 1 = 7$		

Estimativa da variância do erro: $\hat{\sigma}^2 = s^2 = MQE = \frac{1}{n-2}\sum_i e_i^2 \cong 37{,}916$.

(d) Um **intervalo aleatório de confiança** de $(1 - \alpha) \times 100\%$, para o declive da recta β_1, é dado por

$$IAC_{(1-\alpha)\times 100\%}(\beta_1) = \left[\hat{\beta}_1 \pm t_{(n-2;1-\frac{\alpha}{2})} \times \sqrt{\frac{SQE}{(n-2)SQ_{xx}}}\right],$$

onde $t_{(n-2;1-\frac{\alpha}{2})}$ é o percentil $100\times(1 - \alpha/2)$ da distribuição $t_{(n-2)}$. Obtemos, então, o **intervalo determinístico a 95% confiança**, para β_1, onde $t_{(6;0{,}975)} = 2{,}447$,

$$IC_{95\%}(\beta_1) = \left[1{,}497 \pm 2{,}447 \times \sqrt{\frac{227{,}497}{6\times 294}}\right] = [0{,}618; 2{,}375].$$

(e) **Teste à significância do modelo**

- H_0: $\beta_1 = 0$ contra H_1: $\beta_1 \neq 0$ (teste bilateral)
- $\alpha = 5\% \Rightarrow 1 - \alpha = 0{,}95$
- $T = \sqrt{\frac{(n-2)SQ_{xx}}{SQE}} \times \hat{\beta}_1 \underset{\text{Sob } H_0}{\sim} t_{(n-2)}$
- $RC =]-\infty; -2{,}447] \cup [2{,}447; +\infty[$ (reunião de abas)

De notar que: $t_{(n-2;1-\alpha/2)} = t_{(6;0{,}975)} = 2{,}447$.

- $t_{obs} = \sqrt{\frac{6\times 294}{227{,}497}} \times 1{,}497 \cong 4{,}167$. Como $t_{obs} \in RC$ rejeita-se H_0, para α = 5%. Logo, estamos convictos que a variação da classificação no exame (*Y*) não é completamente aleatória, mas que pode, pelo menos parcialmente, ser explicada pelo número de horas de estudo (variável independente, *X*).

(f) **Teste à adequabilidade do modelo**

- H_0: $\beta_0 = \beta_1 = 0$ *versus* H_1: $\exists \beta_i \neq 0$ (i = 0,1) [ou, H_0: modelo não é adequado *versus* H_1: ~ H_0]
- $\alpha = 5\% \Rightarrow 1 - \alpha = 0{,}95$
- $T = \frac{MQR}{MQE} \underset{\text{Sob } H_0}{\sim} F_{(1;\, n-2)}$
- $RC = [5{,}99; +\infty[$ (teste unilateral à direita)

De notar que: $F_{(1;n-2;1-\alpha)} = F_{(1;6;0{,}95)} = 5{,}99$.

- $t_{obs} = \frac{658,503}{37,916} \cong 17,3674$. Como $t_{obs} \in RC$ rejeita-se H_0, para α = 5%. Logo, há motivo para considerarmos que existe uma relação probabilística entre o número de horas de estudo (*X*) e a classificação obtida no exame (*Y*), onde a distribuição da classificação obtida no exame (*Y*) depende de, pelo menos, uma das componentes desta.

Modelo de Regressão Linear Simples (MRLS)

Medidas de Associação Linear

Coeficiente de correlação

$$r = \frac{SQ_{xy}}{\sqrt{SQ_{xx}} \cdot \sqrt{SQ_{yy}}}, \quad -1 \le r \le 1$$

Coeficiente de Determinação

$$r^2 = 1 - \frac{SQE}{SQ_{yy}} = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}, \quad 0 \le r \le 1$$

Covariância entre *x* e *y*

$$s_{xy} = \left(\frac{\sum_i x_i y_i}{n} - \bar{x} \cdot \bar{y} \right)$$

Modelo de Regressão Linear Simples

Especificação do Modelo

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

Modelo Estimado pelo MMQ (Recta da regressão)

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Parâmetros estimados pelo MMQ

$$\hat{\beta}_1 = \frac{SQ_{xy}}{SQ_{xx}} \quad \text{e} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x},$$

em que: $SQ_{xy} = \sum_i x_i y_i - \frac{1}{n} \sum_i x_i \sum_i y_i$ e $SQ_{xx} = \sum_i x_i^2 - \frac{1}{n} \left(\sum_i x_i \right)^2$

Variância dos Parâmetros estimados

$$\sigma^2_{\hat{\beta}_0} = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \, SQ_{x\,x}} \quad \text{e} \quad \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{SQ_{xx}}$$

Resíduo

$$\hat{e} = Y - \hat{Y}$$

Estimador da Variância do Erro

$$S^2 = \hat{\sigma}^2 = MQE = \frac{SQE}{n-2}$$

Distribuição dos estimadores de Mínimos Quadrados

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \cdot SQ_{xx}} \right), \quad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{SQ_{xx}} \right) \quad \text{e} \quad e_i \sim N(0, \sigma^2)$$

Tabela da ANOVA (Analysis of Variance)

Origem de Variação	Soma de Quadrados **(*SQ*)**	Graus de Liberdade **(*gl*)**	Média de Quadrados **(*MQ*)**	Valor da Estatística **F**
Regressão	$SQR = \hat{\beta}_1 \times SQ_{xy} = \frac{(SQ_{xy})^2}{SQ_{xx}}$	1	$MQR = \frac{SQR}{1}$	$F = \frac{MQR}{MQE}$
Erro	$SQE = SQ_{yy} - \frac{(SQ_{xy})^2}{SQ_{xx}} = SQ_{yy} - \hat{\beta}_1 SQ_{xy}$	$n-2$	$MQE = \frac{SQE}{n-2}$	
Total	$SQT = SQ_{yy} = SQR + SQE = \sum_i y_i^2 - \frac{1}{n} \left(\sum_i y_i \right)^2$	$n-1$	---	

	Intervalo Aleatório de Confiança (1-α)100%	$IAC_{(1-\alpha)100\%}(\beta_1) = \left[\hat{\beta}_1 \pm t_{(n-2;1-\frac{\alpha}{2})}\sqrt{\frac{SQE}{(n-2)SQ_{xx}}}\right]$	
	Testes de hipóteses	**Significância do MRLS**	H_0: $\beta_1 = 0$ *versus* H_1: $\beta_1 \neq 0$ $T = \sqrt{\frac{(n-2)SQ_{xx}}{SQE}} \times \hat{\beta}_1 \underset{H_0}{\sim} t_{(n-2)}$ $RC =]-\infty, -c] \cup [c, +\infty[$, onde $c = t_{(n-2;\,1-\frac{\alpha}{2})}$
		Adequação do MRLS	H_0: O modelo de RLS não é adequado *versus* H_1: $\sim H_0$ H_0: $\beta_0 = \beta_1 = 0$ *versus* H_1: $\exists\, \beta_i \neq 0$ (i = 0, 1) $T = \frac{MQR}{MQE} \underset{H_0}{\sim} F_{(1,\,n-2)}$ $RC = [c, +\infty[$ onde $c = F_{(1,\,n-2;\,1-\alpha)}$