Modelos Lineares
Regressão Linear Simples

Susana Faria

Notas Iniciais

- **O uso destas notas como único material de estudo é fortemente desaconselhado.**
- **Neste capítulo estuda-se o modelo de regressão linear simples.**

Modelo Regressão Linear Simples (MRLS)

Considere a relação linear entre a variável resposta Y e a variável explicativa X, representada por:

$$Y = \beta_0 + \beta_1 x + \epsilon$$

onde β_0 e β_1 são designados por parâmetros ou coeficientes de regressão desconhecidos do modelo.
Ao termo ϵ designamos por erro aleatório e assumimos que tem distribuição normal com média nula e variância σ^2.

A este modelo designamos por Modelo de Regressão Linear Simples (MRLS).

Exemplo:

- Estudar a relação entre o peso ao nascer e o número de semanas de gestação;

Regressão Linear Simples (RLS)

Exemplo:
Pretende-se estudar a relação entre a pressão sistólica e a idade em indivíduos adultos.

Antes de qualquer tentativa de construção de um modelo é preciso explorar os dados. Nomeadamente:

- Conhecer o tipo de variáveis de que dispomos;
- Descrever os dados relativos a cada uma das variáveis através de representações gráficas e estatísticas sumárias;
- avaliar o comportamento conjunto das variáveis, calculando medidas de associação e através de representações gráficas.

Regressão Linear Simples (RLS)

Dada uma amostra bivariada $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ de n observações independentes onde x_i e y_i são, respectivamente, os valores da variável X e Y para o individuo i, tem-se:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \ldots, n$$

em que:
$Y_i -$ resposta aleatória do indivíduo i (variável dependente aleatória)
$x_i -$ $i-$ésima observação da variável independente
$\beta_0 -$ ordenada na origem (parâmetro desconhecido do modelo)
$\beta_1 -$ declive (parâmetro desconhecido do modelo)
$\epsilon_i -$ erro aleatório associado à observação da resposta do indivíduo i.

Nota: Os valores x_i são considerados determinísticos (pré-determinados à partida). Os valores Y_i representam a variàvel dependente e estes sim são considerados variáveis aleatórias.

Regressão Linear Simples (RLS)

Pressupostos usuais do modelo RLS :

- $E[\epsilon_i] = 0$, o que implica que, dado um valor de x,

$$E[Y|x] = \beta_0 + \beta_1 x$$

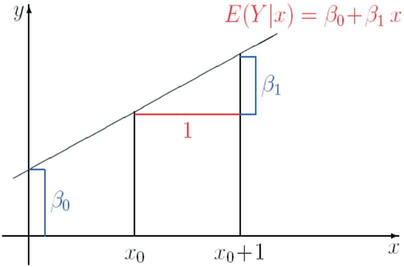
  conhecida por equação ou recta de regressão do modelo.
- $Var[\epsilon_i] = \sigma^2 \ \ \forall i$ (variância constante desconhecida).
- ϵ_i's são variáveis aleatórias independentes.
- ϵ_i segue uma distribuição Normal.

Regressão Linear Simples (RLS)

Interpretação dos coeficientes:

β_0- ordenada na origem. Representa o valor esperado de Y para um valor nulo da variável explicativa.

β_1- declive. Representa a variação do valor esperado de Y por cada incremento unitário na variável explicativa.



Estimação dos Parâmetros de um MRLS

- Estamos interessados em determinar estimadores $\hat{\beta}_0$ de β_0 e $\hat{\beta}_1$ de β_1 de forma a obter a variável resposta estimada $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1\, x_i$ para cada valor observado de x_i.
- Um método de estimação dos coeficientes de regressão é o Método de Mínimos Quadrados que consiste em minimizar a soma de quadrados dos erros aleatórios. Ou seja, o valor que minimiza a função:

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_i)]^2$$

- Para a determinação da estimativa associada a $\hat{\beta}_0$ e $\hat{\beta}_1$, deve-se encontrar as derivadas parciais da função $Q(\beta_0, \beta_1)$ avaliada em (y_i, x_i) em relação aos parâmetros β_0 e β_1.

$$\begin{cases} \frac{\partial Q(\beta_0,\beta_1)}{\partial \beta_0} = 0 \\ \frac{\partial Q(\beta_0,\beta_1)}{\partial \beta_1} = 0 \end{cases} \Leftrightarrow \cdots \Leftrightarrow \begin{cases} \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x} \\ \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i \; Y_i - n\overline{x}\;\overline{Y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \end{cases}$$

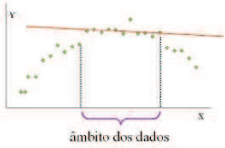
Nota: Pode-se provar que este é ponto de mínimo, visto que a matriz hessiana avaliada neste ponto é definida positiva.

Estimação dos Parâmetros de um MRLS

- A **equação ou recta de regressão é estimada por:**

$$\hat{Y} = \hat{E}[Y|x] = \hat{\beta}_0 + \hat{\beta}_1 x$$

- **A estimação pontual de $E(Y|X)$ deve restringir-se ao domínio dos valores observados na amostra da variável explicativa X.**



- **Os valores $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ designam-se por valores estimados ou valores preditosde Y_i , em inglês, "fitted values"ou "predicted values".**

- **As quantidades $e_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1\ x_i)$ são designados por resíduos.**

Nota: O ponto $(\overline{x}, \overline{y})$ pertence à recta de regressão.

Propriedades dos Estimadores

- $E[\hat{\beta}_1] = \beta_1$ $\qquad VAR[\hat{\beta}_1] = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}$
- $E[\hat{\beta}_0] = \beta_0$ $\qquad VAR[\hat{\beta}_0] = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \right)$

Os estimadores dos Mínimos Quadrados dos parâmetros :

- são combinações lineares de Y_i.
- são **centrados** ou **não enviesados**.
- têm variância mínima.
- são, de entre os centrados, os de menor variância. (**BLUES**)

Estimador centrado de σ^2:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{SSE}{n-2}$$

onde SSE é a soma dos quadrados dos resíduos.
Notação alternativa: $\hat{\sigma}^2 = \frac{1}{n-2}\left(S_{YY} - \frac{S_{XY}^2}{S_{XX}}\right)$

Nota: $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$

Inferências no Modelo de Regressão Linear Simples

Inferências sobre β_1

- Pretende-se testar a hipótese:

$$H_0 : \beta_1 = b_1 \quad \text{vs} \quad H_1 : \beta_1 \neq b_1$$

A estatística-teste é:

$$T = \frac{\hat{\beta}_1 - b_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n x_i^2 - n\overline{x}^2}}} \quad \sim \quad t_{n-2}$$

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

- Pretende-se calcular o intervalo de confiança para β_1:

$$\left(\hat{\beta}_1 - t_{\frac{\alpha}{2};n-2}\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n x_i^2 - n\overline{x}^2}},\ \ \hat{\beta}_1 + t_{\frac{\alpha}{2};n-2}\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n x_i^2 - n\overline{x}^2}}\right)$$

Inferências no Modelo de Regressão Linear Simples

Inferências sobre β_0

- Pretende-se testar a hipótese:

$$H_0 : \beta_0 = b_0 \quad \text{vs} \quad H_1 : \beta_0 \neq b_0$$

A estatística-teste é:

$$T = \frac{\hat{\beta}_0 - b_0}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n x_i^2 - n\overline{x}^2} \right)}} \quad \sim \quad t_{n-2}$$

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

- Pretende-se calcular o intervalo de confiança para β_0:

$$\left(\hat{\beta}_0 - t_{\frac{\alpha}{2};n-2} \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n x_i^2 - n\overline{x}^2} \right)}; \hat{\beta}_0 + t_{\frac{\alpha}{2};n-2} \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n x_i^2 - n\overline{x}^2} \right)} \right)$$

Inferências no Modelo de Regressão Linear Simples

Estimação do valor esperado de Y quando a variável explicativa toma o valor x_0: $E[Y_0] = E[Y|x = x_0] = \beta_0 + \beta_1 x_0$

- Estimador pontual:

$$\hat{E}[Y_0] = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

- Intervalo de confiança a $100(1-\alpha)\%$ para $E[Y_0]$:

$$\left(\hat{E}[Y_0] - t_{\frac{\alpha}{2};n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(\overline{x} - x_0)^2}{\sum_{i=1}^n x_i^2 - n\overline{x}^2}\right)};\right.$$

$$\left.\hat{E}[Y_0] + t_{\frac{\alpha}{2};n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(\overline{x} - x_0)^2}{\sum_{i=1}^n x_i^2 - n\overline{x}^2}\right)}\right)$$

Nota: As inferências podem não ser válidas fora do intervalo de valores de x considerado.

Inferências no Modelo de Regressão Linear Simples

Previsão do valor de Y quando a variável explicativa toma o valor x_0:
$Y_0 = Y|x = x_0 = \beta_0 + \beta_1 x_0$

- Estimador pontual:

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

- Intervalo de confiança a $100(1-\alpha)\%$ para Y_0:

$$\left(\hat{Y}_0 - t_{\frac{\alpha}{2};n-2} \sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(\overline{x} - x_0)^2}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\right)} ; \right.$$

$$\left. \hat{Y}_0 + t_{\frac{\alpha}{2};n-2} \sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(\overline{x} - x_0)^2}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\right)} \right)$$

Nota: As inferências podem não ser válidas fora do intervalo de valores de x considerado.

ANOVA

Para a observação y_i, tem-se:

$$(y_i - \overline{y}) = (\widehat{y}_i - \overline{y}) + (y_i - \widehat{y}_i)$$

Considerando todas as observações:

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

$$SST = SSR + SSE$$

em que:

- SST: soma de quadrados total
- SSE: soma de quadrados dos resíduos
- SSR: soma de quadrados de regressão

ANOVA

Para testar:

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

A tabela da ANOVA correspondente é:

Fonte de Variação	SS	gl	MS	F_0	$p - value$
Regressão	SSR	1	MSR	$\frac{MSR}{MSE}$	
Erros	SSE	n-2	MSE		
Total	SST	n-1			

Rejeita-se a hipótese H_0 ao nível de significância α se o valor da estatística de teste, F_0 for maior do que o valor de F com (1,n-2) graus de liberdade.

Avaliação da qualidade e significado da regressão

Para avaliar a qualidade e significado da regressão vamos considerar vários métodos.

- Métodos Gráficos
- Teste ao Declive
- Coeficiente de Determinação

Métodos Gráficos

- O método mais intuitivo para avaliar a qualidade e signicado de uma regressão baseia-se na observação do gráfico de dispersão quando traçado com a recta de regressão sobreposta.
- Uma alternativa gráfica que pode detectar eventuais desvios à linearidade não detectàveis no gráfico de dispersão dos dados é um outro gráfico de dispersão que apresente os valores observados Y_i versus os valores preditos $\hat{Y}_i$.

Teste ao declive

Será que Y depende mesmo de x? Podemos responder a esta questãoo através do teste:

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

Coeficiente de Determinação

Definição

O coeficiente de determinação é uma medida relativa da qualidade de ajustamento do modelo de regressão linear, dada por:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \left(\frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \right)^2 = \frac{S_{XY}^2}{S_{XX} S_{YY}}$$

É interpretada como a percentagem da variabilidade de Y que é explicada pelo modelo de regressão linear.

O coeficiente de determinação é tal que $0 \leq R^2 \leq 1$ onde:

- $R^2 \simeq 1$ indica bom ajustamento do modelo;
- $R^2 \simeq 0$ indica mau ajustamento do modelo.

Nota: $1 - R^2$: proporção da variação de Y não explicada pela variável X, resultante de factores não incluídos no modelo.

Coeficiente de Determinação

- O coeficiente de determinação R^2 apresenta um viés positivo em relação ao seu valor na população, o que pode induzir em erro na análise dos dados. Uma forma de compensar este viés consiste em considerar um coeficiente de determinação ajustado definido a partir de R^2 e ajustado com base na dimensão da amostra.

$$R_a^2 = 1 - \frac{\frac{SSE}{(n-2)}}{\frac{SST}{(n-1)}}$$

Relação entre o Coeficiente de Correlação e o Coeficiente de Determinação

- No caso do modelo RLS, o coeficiente de determinação (R^2) é o quadrado do coeficiente de correlação entre x e y, (r_{xy}).

Relação entre o Coeficiente de Correlação e Análise de Regressão

- O sinal da correlação indica a direcção da relação.

Análise dos Resíduos

Relembremos que de acordo com o modelo de regressão linear simples os erros das observações satisfazem os seguintes pressupostos:

- seguem uma distribuição normal;
- têm media zero;
- têm variância constante (homocedasticidade);
- são independentes.

- A verificação das hipóteses é fundamental, visto que toda a inferência estatística no modelo de regressão linear (testes de hipóteses) se baseia nesses pressupostos.
- Nesse sentido, se houver violação dos mesmos, a utilização do modelo deve ser posta em causa.
- A Análise dos Resíduos é uma ferramenta usada para detectar violações dos pressupostos.

Análise dos resíduos

Recorde-se que **resíduo** é:

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 \, x_i \quad i = 1, \ldots, n$$

Os resíduos padronizados são:

$$e_i^s = \frac{e_i}{\hat{\sigma}}$$

Normalidade dos erros

- O pressuposto da normalidade pode ser testado traçando um Normal QQ-plot ou um Normal PP-plot para os resíduos. **Se os erros possuírem distribuição Normal, todos os pontos dos gráficos devem posicionarem-se mais ou menos sobre uma recta.**
- Também se pode proceder a testes de ajustamento dos resíduos a uma distribuição Normal: Teste Kolmogorov- Smirnov e Teste de Shapiro.

Análise dos Resíduos

Média Nula, Variância constante e independência dos erros

- Estes pressupostos podem ser verificados graficamente representando os resíduos versus valores estimados da variável dependente $\hat{Y}_i$ (ou versus valores da variável independente).
- **Os pontos do gráfico devem distribuir-se de forma aleatória em torno da recta que corresponde ao resíduo zero, formando uma mancha de largura uniforme.** Dessa forma será de esperar que os erros sejam independentes, de média nula e de variância constante.
- Se a dispersão dos resíduos aumentar ou diminuir com os valores da variável independentes x_i, ou com os valores estimados da variável dependente $\hat{Y}_i$, deve ser posta em causa a hipótese de variâncias constante dos erros.

Independência dos erros

- Para verificar o pressuposto da independência dos erros, representam-se os resíduos padronizados *versus* a ordem pela qual os dados foram recolhidos. É de esperar que a nuvem de pontos não apresente padrão, o que significará que as observações foram recolhidas de forma independente.
- A verificação da independência pode ser feita através do teste de Durbin-Watson à correlação entre resíduos sucessivos.

Análise dos Resíduos

Variáveis explicativas - adequabilidade

- É importante analisar a relação existente entre os resíduos do modelo estimado e as variáveis explicativas. O que se espera, de acordo com os pressupostos do modelo, é que tal relação seja inexistente. Isto é, quando os resíduos são representados *versus* os valores de cada uma das variáveis explicativas, a nuvem de pontos não deverá apresentar qualquer padrão.

- Quando as variáveis explicativas são de natureza quantitativa contínua, representam-se os pontos (x_i, e_i). Na presença de variáveis categóricas, a representação (x_i, e_i) não faz sentido. Como alternativa, poderemos optar por qualquer representação que permita averiguar se os valores dos resíduos para cada classe apresentam distribuição semelhante - por exemplo, box-plot paralelos.

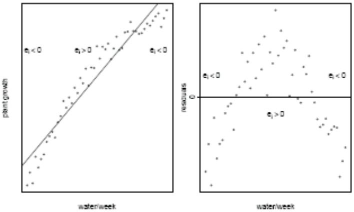
Transformações

O uso de transformações da variável resposta ou das variáveis explicativas é frequentemente suficiente para garantir os pressupostos do modelo de regressão quando aplicado a dados transformados.

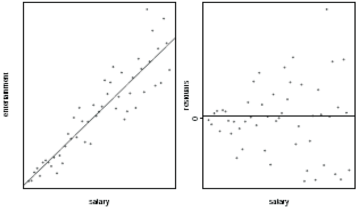
- Transformações de X podem ser úteis para linearizar a relação de regressão não linear sem afetar a distribuição de Y;
- Transformações $\sqrt(Y)$ e $log(Y)$ são recomendadas quando a variÂncia dos erros aleatórios cresce proporcionalmente a x_i e a x_i^2, respetivamente, $i = 1, \ldots, n$;
- Transformações de Box-Cox.

Exemplos

Example:
X_i = amount of water/week
Y_i = plant growth in first 2 months



Example:
X_i = salary
Y_i = money spent on entertainment



Points on a straight line: Errors are normal (left)
Points on a curve: Errors are not normal (right)