Universidade Federal do Paraná

Laboratório de Estatística e Geoinformação - LEG

Regressão Linear - Revisão

Prof.: Eduardo Vargas Ferreira

Introdução

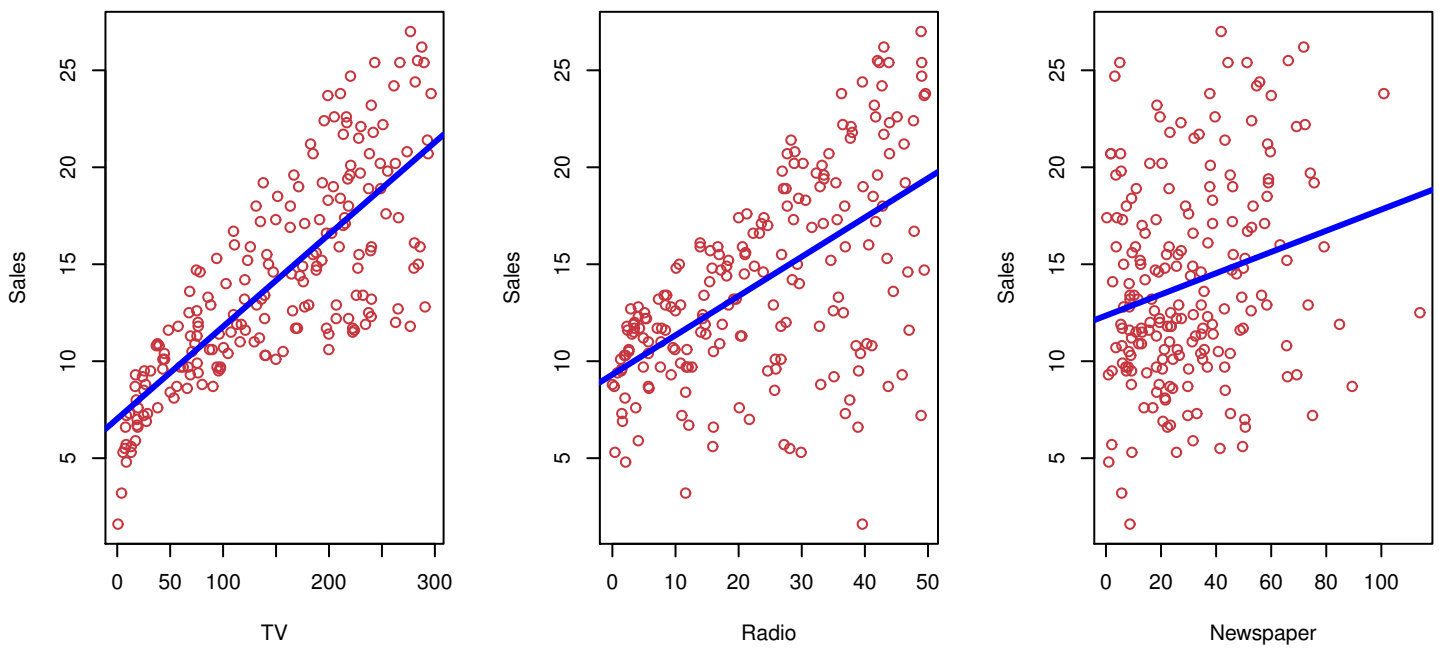
- Regressão linear é uma das mais simples ferramentas de aprendizagem supervisionada;
- Em particular, é útil na predição para resposta quantitativa;
- Além de ser uma abordagem inicial para métodos mais elaborados (veremos adiante);
- Matematicamente, podemos escrever a relação linear entre X e Y como

$$Y \approx \beta_0 + \beta_1 X \tag{1}$$

- β_0 e β_1 serão estimados.

Introdução

- Os dados utilizados referem-se ao `Advertising` data set;
  - ⋆ Y = `Sales` de um particular produto em 200 lojas;
  - ⋆ X = investimento em publicidade na `TV`, `Radio` e `newspaper` de cada loja.



Introdução

- Em regressão queremos responder questões como:
  - ⋆ Existe relação entre investimento com publicidade e vendas de determinado produto?
  - ⋆ Quão forte é essa relação?
  - ⋆ É uma relação linear, quadrática etc?
  - ⋆ Qual mídia mais contribui para o aumento nas vendas?
  - ⋆ Existe uma sinergia entre elas?
  - ⋆ Quão precisa são as estimativas?

Introdução

- Descrevemos (1) como a regressão de Y em X; Por exemplo, podemos estudar a relação entre `TV` e `Sales`

$$\texttt{Sales} \approx \beta_0 + \beta_1 \times \texttt{TV}$$

- β_0 e β_1 representam o intercepto e o coeficiente angular do modelo;

- Na prática, β_0 e β_1 são desconhecidos;

- Então, antes de utilizar a Equação (1) para fazer predição, precisamos estimar os coeficientes;

- Utilizamos os dados de treinamento para encontrar $\hat{\beta}_0$ e $\hat{\beta}_1$.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{2}$$

Introdução

- De forma geral, o modelo de regressão linear pode ser escrito como

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

com

$$\underset{n\times 1}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}, \quad \underset{n\times(p+1)}{\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}}, \quad \underset{(p+1)\times 1}{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}} \quad e \quad \underset{n\times 1}{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}},$$

em que

  - ⋆ **Y** é a **variável dependente**. Representa as n respostas;
  - ⋆ **X** as **variáveis independentes** (em colunas);
  - ⋆ $\boldsymbol{\beta}$ vetor de parâmetros (desconhecidos);
  - ⋆ $\boldsymbol{\varepsilon}$ vetor de erro.

Interpretação geométrica da regressão linear

- A variação de um específico y_i em torno da média é dada por $(y_i - \bar{y})$.

- Somando e subtraindo $\hat{y}$ temos

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y})\,.$$

- Elevando ao quadrado e somando em i, obtém-se

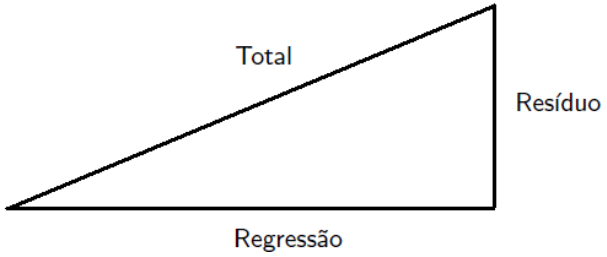
$$\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{SQT} = \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}_{SQR} + \underbrace{\sum_{i=1}^{n} (y_i - \hat{y})^2}_{SQE}\,.$$

- Reescrevendo em termos da norma dos vetores temos

$$\left\|\mathbf{Y} - \bar{\mathbf{Y}}\right\|_2^2 = \left\|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\right\|_2^2 + \left\|\mathbf{Y} - \hat{\mathbf{Y}}\right\|_2^2$$

Interpretação geométrica da regressão linear

- Assim, pelo Teorema de Pitágoras, temos a seguinte situação



- Tais subespaços são definidos como

1. **Espaço coluna de X:** é o subconjunto de R^n definido por:

$$\mathcal{R}(\mathbf{X}) = \left\{ y \in \mathrm{R}^n / y = \mathbf{X}\boldsymbol{\beta} \ ; \ \boldsymbol{\beta} \in \mathrm{R}^{p+1} \right\}.$$

2. **Espaço nulo de X^t:** é o subconjunto de R^n definido por:

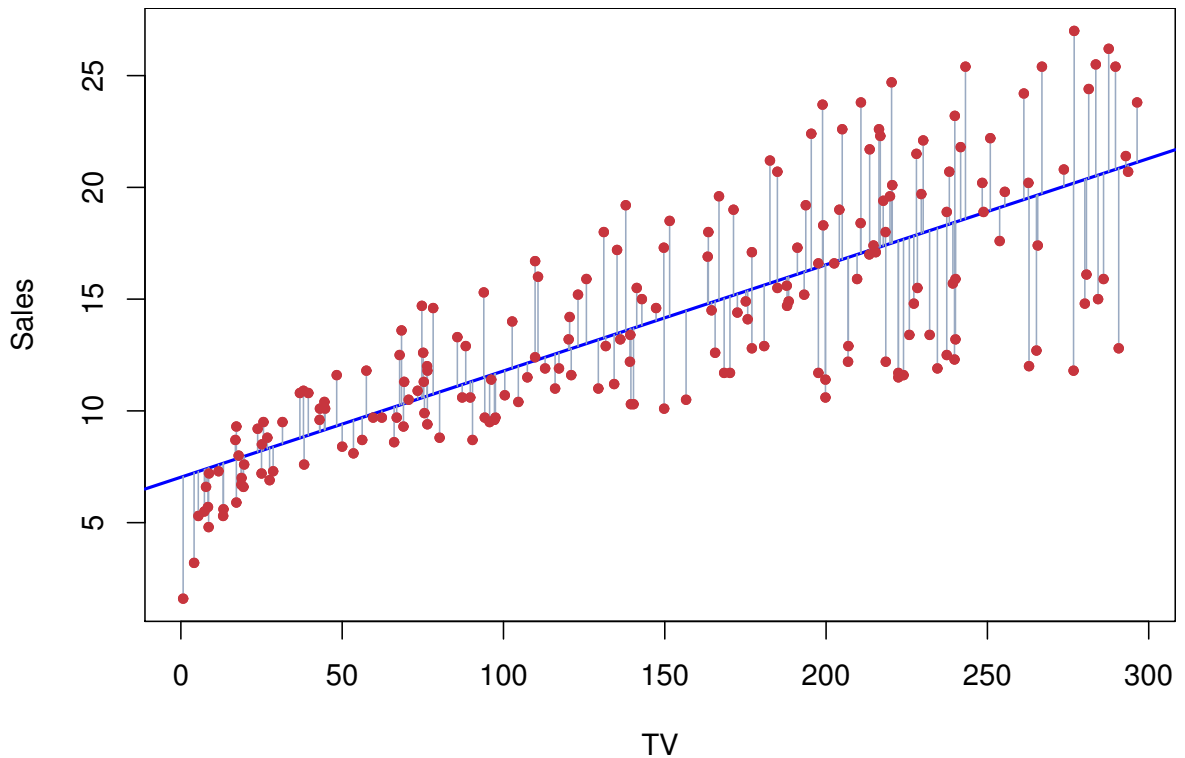
$$\mathcal{N}(\mathbf{X}^t) = \left\{ y \in \mathrm{R}^n / \mathbf{X}^t y = 0_{\mathrm{R}^{p+1}} \right\}.$$

- Seja $z \in \mathcal{N}(\mathbf{X}^t)$, isto é, $\mathbf{X}^t z = 0$. Para todo $\boldsymbol{\beta} \in \mathrm{R}^{p+1}$ tem-se

$$0 = \left\langle \mathbf{X}^t z, \boldsymbol{\beta} \right\rangle = \left\langle z, \mathbf{X}\boldsymbol{\beta} \right\rangle$$

Estimando os coeficientes

- Sejam $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ os n pares de observações;
- Queremos encontrar os valores de $\hat{\beta}_0$ e $\hat{\beta}_1$, tq, a reta resultante seja a mais próxima possível dos pontos;



Estimando os coeficientes

- Existem várias formas de medir a proximidade;
- A mais comum envolve minimizar a Soma de quadrado dos desvios

$$
\begin{aligned}
min\left\{J(y_i, h(\mathbf{x}))\right\} &\approx min\left\{\frac{1}{n}\sum_{i=1}^{n}[y_i - h(x_i)]^2\right\} \\
&= min\left\{\frac{1}{n}\sum_{i=1}^{n}[y_i - \beta_0 + \beta_1 x_i]^2\right\}
\end{aligned}
$$

- Assim, o parâmetro estimado é obtido da forma

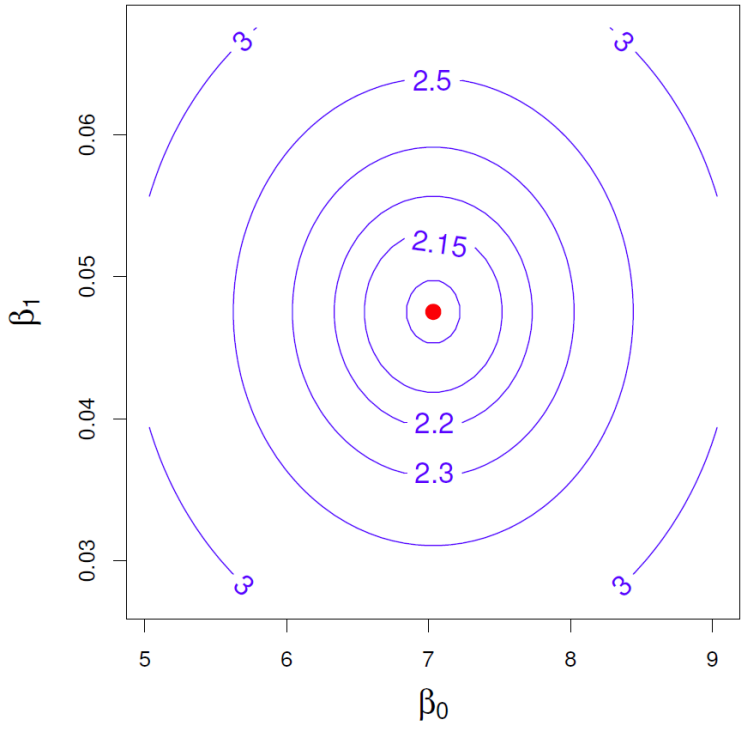
$$\frac{\partial J(y_i, h(\mathbf{x}))}{\partial \beta_j} = 0$$

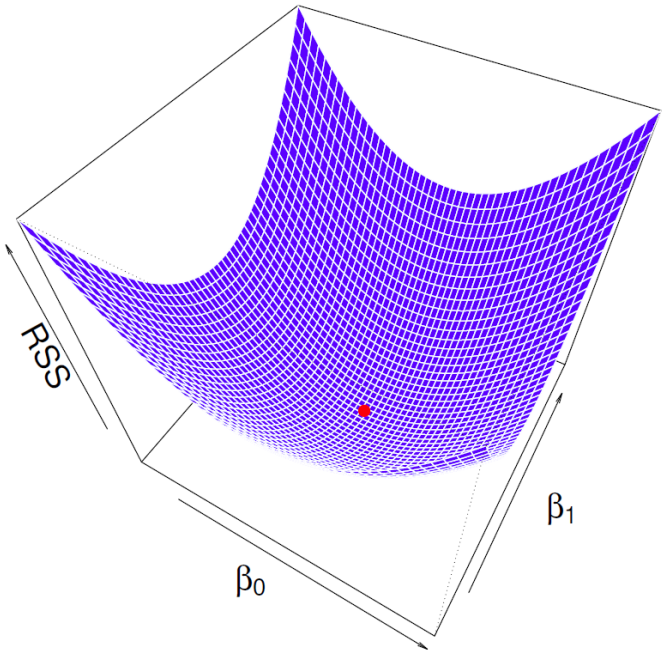
- Chegando em

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}\mathbf{y}$$

Estimando os coeficientes

- Abaixo o gráfico de contorno da SQD para os dados `Advertising`;
- O ponto vermelho representa as estimativas dos parâmetros.





Avaliando a precisão dos β's

- Para testar hipóteses sobre os parâmetros, sabemos que

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \sim t_{(n-p-1)}$$

- C_{jj} é o j-ésimo elemento da diagonal de $(\mathbf{X}^t\mathbf{X})^{-1}$;

- Para os dados `Advertising`, temos os seguintes parâmetros estimados

Variável	Coeficiente	Erro padrão	Estatística t	p-valor
`Intercepto`	7,0325	0,4578	15,36	$< 0,0001$
`TV`	0,0475	0,0027	17,67	$< 0,0001$

- Assim, um investimento adicional de \$1000 está associado a um aumento nas vendas de 47 unidades (lembrando que `Sales` está em mil unidades).

Avaliando a precisão dos β's

- O intervalo de confiança de $100(1-\alpha)\%$ de β_j é

$$\hat{\beta}_j - t_{\alpha/2,(n-p-1)}\sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2,(n-p-1)}\sqrt{\hat{\sigma}^2 C_{jj}}$$

- C_{jj} é o j-ésimo elemento da diagonal de $(\mathbf{X}^t\mathbf{X})^{-1}$;

- Para os dados `Advertising`, o intervalo de confiança de 95% para β_1 é $[0,042; 0,053]$.

- Note que σ^2 era desconhecido e foi estimado na seguinte forma

$$\hat{\sigma}^2 = \frac{SQE}{n-p+1} = \frac{\|\mathbf{y}-\hat{\mathbf{y}}\|_2^2}{dim(\mathcal{N}(\mathbf{X}^t))}$$

Interpretando os coeficientes da regressão

- Considere o modelo de regressão linear múltipla

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

- Interpretamos β_j como o **efeito médio** em Y quando aumentamos X_j em uma unidade, **mantendo fixo as outras preditoras**;

- Idealmente, gostaríamos que as preditoras fossem não correlacionadas;

- Assim, além das interpretações como acima serem possíveis, os coeficientes poderiam ser estimados (e testados) separadamente;

- Mas na prática isso (geralmente) não acontece, e um dos problemas causados por essas correlações é a colinearidade.

Colinearidade

- Duas variáveis X_1 e X_2 são ditas exatamente colineares se $c_1X_1 + c_2X_2 = c_0$ para algumas constantes c_i.

- Entretanto, na análise de dados é mais comum haver variáveis aproximadamente colineares do que exatamente colineares.

- Assim, duas variáveis X_1 e X_2 são ditas aproximadamente colineares se $c_1X_1 + c_2X_2 \approx c_0$.

- A colinearidade é medida através do quadrado do coeficiente de correlação entre as variáveis, r_{12}^2.

  - ⋆ Colinearidade exata ocorre quando $r_{12}^2 = 1$;
  - ⋆ Mas, se $r_{12}^2 \approx 1$, devemos nos preocupar.

Colinearidade

- Quando temos p variáveis explanatórias, $X_1, \ldots, X_p$, elas são ditas (aproximadamente) colineares se $c_1 X_1 + \ldots + c_p X_p \approx c_0$.

- Para um j específico,

$$X_j \approx \frac{c_0}{c_j} - \sum_{i \neq j} \frac{c_i}{c_j} X_i$$

- Esta expressão corresponde (aproximadamente) a uma regressão com intercepto $\frac{c_0}{c_j}$.

- Portanto, as variáveis $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p$ explicam X_j se o coeficiente de determinação R_j^2 for próximo de um.

- Se as variáveis são exatamente colineares, a solução das equações normais não é única.

Colinearidade

- A colinearidade (aproximada) tem um efeito na precisão dos coeficientes de regressão.
- Com efeito, numa regressão com duas variáveis explanatórias e intercepto,

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{1 - r_{12}^2} \frac{1}{S_{xx_j}}, \quad j = 1, 2.$$

- Então obtemos a menor variância quando $r_{12}^2 = 0$ e a variância aumenta (inflaciona) na medida em que r_{12}^2 se aproxima de um.
- Uma expressão equivalente para p variáveis explanatórias é,

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{1 - R_j^2} \frac{1}{S_{xx_j}}, \quad j = 1, 2.$$

em que $\frac{1}{(1-R_j^2)}$ é conhecido como *Fator de inflação da variância (FIV)*.

- De forma que o *FIV* representa o aumento na variância devido à correlação entre as variáveis explanatórias.

Modelo de regressão com preditor qualitativo

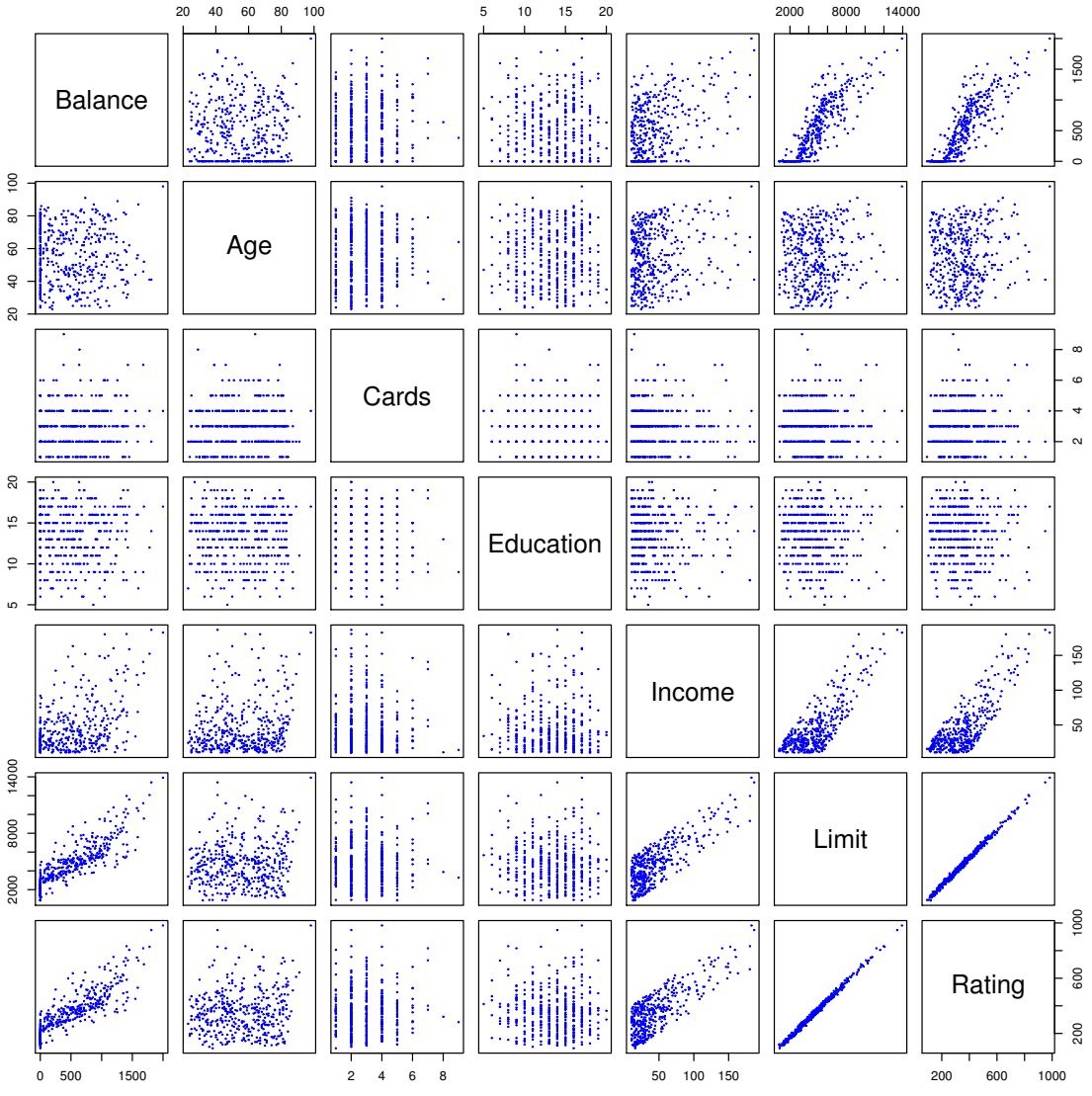
- Em muitas situações, algumas variáveis independentes não são **quantitativas**, mas sim **qualitativas** (assumindo valores discretos);

- Essas são chamadas **preditoras categóricas** ou **fatores**;

- Veja por exemplo o *scatterplot* referente aos dados `Credit`. Além das 7 variáveis quantitativas mostradas, temos também:

  - ⋆ `gender`;
  - ⋆ `student`(é ou não estudante);
  - ⋆ `status` (marital status);
  - ⋆ `ethnicity` (Caucasian, African American (AA) or Asian).

Modelo de regressão com preditor qualitativo



Modelo de regressão com preditor qualitativo

- Suponha que queremos estudar a diferença no saldo do cartão de crédito entre homens e mulheres, ignorando as outras variáveis.

- Para tanto, criamos uma varável da forma

$$x_i = \begin{cases} 1, & \text{se a } i\text{-ésima pessoa é mulher,} \\ 0, & \text{se a } i\text{-ésima pessoa é homem,} \end{cases}$$

- Resultando no modelo

$$y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 x_1 + \varepsilon_i, & \text{se a } i\text{-ésima pessoa é mulher,} \\ \beta_0 + \varepsilon_i, & \text{se a } i\text{-ésima pessoa é homem,} \end{cases}$$

Variável	Coeficiente	Erro padrão	Estatística t	p-valor
`Intercepto`	509,80	33,13	15,389	$< 0,0001$
`gender[Female]`	19,73	46,05	0,429	0,6690

Modelo de regressão com preditor qualitativo

- Se tivermos mais de dois níveis, criamos uma variável *dummy* adicional;
- Por exemplo, para variável `ethnicity` criamos duas variáveis *dummy*. A primeira pode ser

$$x_{i1} = \begin{cases} 1, & \text{se a } i\text{-ésima pessoa é asiática,} \\ 0, & \text{caso contrário,} \end{cases}$$

- E a segunda pode ser

$$x_{i2} = \begin{cases} 1, & \text{se a } i\text{-ésima pessoa é caucasiana,} \\ 0, & \text{caso contrário,} \end{cases}$$

- Resultando no modelo

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i, & \text{se a } i\text{-ésima pessoa é asiática,} \\ \beta_0 + \beta_2 + \varepsilon_i, & \text{se a } i\text{-ésima pessoa é caucasiana,} \\ \beta_0 + \varepsilon_i, & \text{se a } i\text{-ésima pessoa é AA,} \end{cases}$$

Modelo de regressão com preditor qualitativo

- Note que sempre teremos menos variáveis *dummy* do que o número de níveis;

- Assim, o nível não codificado (no exemplo `African American`) é conhecido como *baseline*;

Variável	Coeficiente	Erro padrão	Estatística t	p-valor
`Intercepto`	531,00	46,32	11,464	$< 0,0001$
`ethnicity[Asian]`	-18,69	65,02	-0,287	0,7740
`ethnicity[Caucasian]`	-12,50	56,68	-0,221	0,8260

Inclusão de interação

- Nos dados `Advertising` assumimos que o efeito médio em `sales` é explicado por determinado meio de publicidade sem levar em conta os demais investimentos;

- Ou seja, não consideramos as interações entre `TV`, `radio` e `newspaper`:

$$\widehat{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper}$$

- Assim, o efeito médio em `sales` com o aumento de uma unidade em `TV` é sempre β_1, independente do investimento em `radio`;

- Mas suponha que a propaganda em `radio` aumente a eficácia da propaganda na `TV`;

- Então o coeficiente angular de `TV` deve aumentar com o aumento de `radio`;

Inclusão de interação

- O modelo fica da forma

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{TV} \times \texttt{radio}) + \varepsilon$$

- Resultando em

Variável	Coeficiente	Erro padrão	Estatística t	p-valor
Intercepto	6,7502	0,248	27,23	$< 0,0001$
TV	0,0191	0,002	12,70	$< 0,0001$
radio	0,0289	0,009	3,24	0,0014
TV×radio	0,0011	0,000	20,73	$< 0,0001$

- Os resultados sugerem que a interação é significativa;

- Com efeito, o p-valor para TV×radio é extremamente baixo, indicando uma forte evidência para $\beta_3 \neq 0$;

- O R^2 para o modelo com interação é 96,8%. Já o modelo com somente TV e radio (sem interação) foi de 89,7%.

Inclusão de interação

- Veremos agora como fica a inclusão de interação quando temos variável quantitativa e qualitativa;

- Considere os dados `Credit`, e suponha que desejamos predizer `balance` pelas variáveis `income` (quantitativa) e `student` (qualitativa);

- Sem a interação o modelo fica

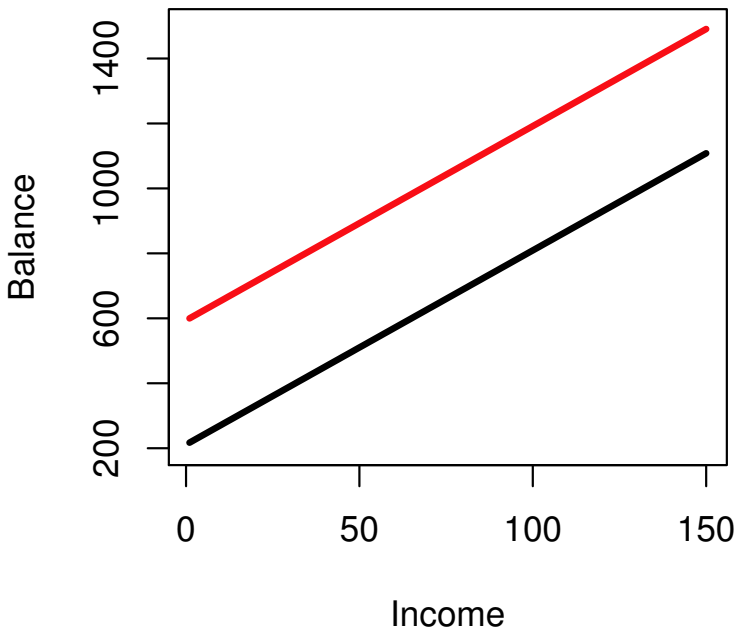
$$\begin{aligned}
\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2, & \text{se a } i\text{-ésima pessoa é estudante,} \\ 0, & \text{caso contrário,} \end{cases} \\
&= \beta_1 \times \texttt{income}_i + \begin{cases} \beta_0 + \beta_2, & \text{se a } i\text{-ésima pessoa é estudante,} \\ \beta_0, & \text{caso contrário,} \end{cases}
\end{aligned}$$

- E com a interação

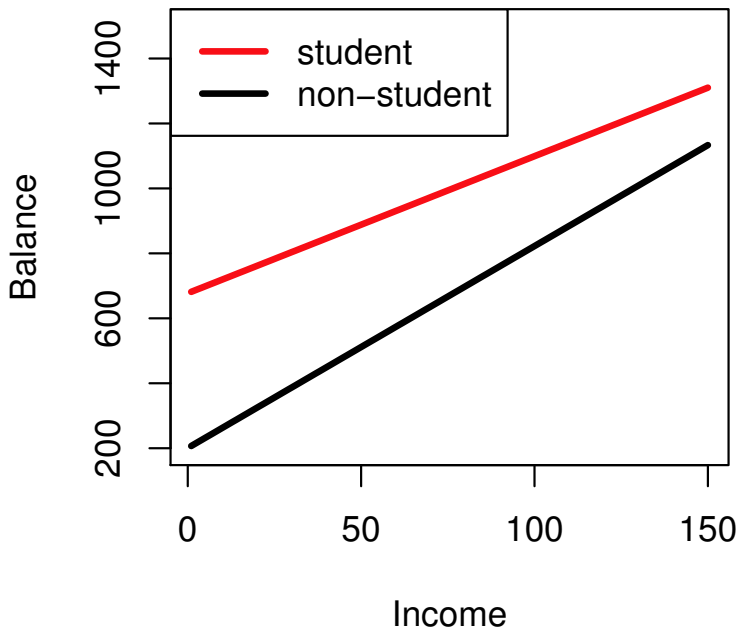
$$\begin{aligned}
\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{income}_i, & \text{para estudante,} \\ 0, & \text{caso contrário,} \end{cases} \\
&= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{income}_i, & \text{para estudante,} \\ \beta_0 + \beta_1 \times \texttt{income}_i, & \text{caso contrário,} \end{cases}
\end{aligned}$$

Inclusão de interação

Sem interação



Com interação



Inclusão de interação

Observações:

- Em alguns casos o termos da interação tem um p-valor baixo, mas os efeitos principais associados a ele não;
- O **Princípio da hierarquia** diz que se incluirmos a interação no modelo, devemos permanecer com seus fatores principais (ainda que suas significâncias não sejam altas);
- Sem os fatores principais, as interações mudam de significado (além de serem mais difíceis de interpretar);
- Especificamente, na interação conterá os fatores principais.