ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS – PUC-RIO
MACHINE LEARNING

AULA 3: PROBLEMAS DE REGRESSÃO

Tatiana Escovedo, PhD.
tatiana@inf.puc-rio.br

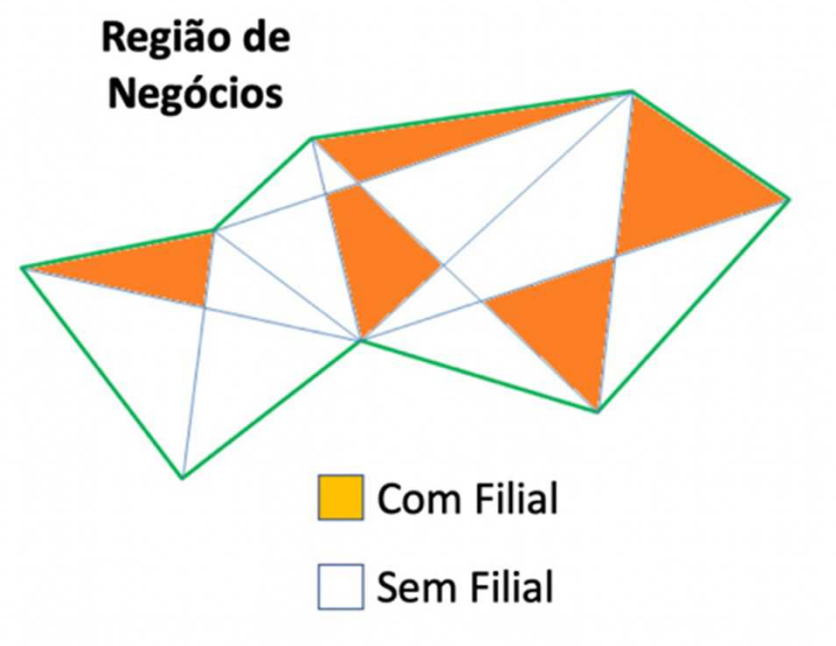
AGENDA

- Problemas de Regressão
- Regressão Linear
- Regularização
- Regressão Logística
- KNN (para Regressão)
- Árvore de Regressão
- SVM (para Regressão)

PROBLEMAS DE REGRESSÃO

PROBLEMA DE REGRESSÃO: EXEMPLO

- Considere um grupo varejista com esta região de negócios:



- **Problema:** onde abrir uma nova filial?
- **Métrica de qualidade de uma boa filial:** Faturamento Médio Annual
  - Métrica conhecida nas regiões onde há filiais e desconhecida nas demais.

PROBLEMA DE REGRESSÃO: EXEMPLO

- Como podemos inferir os valores para os bairros não disponíveis (NA)?

Bairro	Faturamento
A	105.000
B	NA
C	NA
D	NA
...	...
K	NA
L	150.000
M	NA

- 1º passo: levantar variáveis que estão tanto presente nos bairros com/sem Faturamento:
  - Renda per Capita
  - IDH
  - Número de Concorrentes
  - Número de Habitantes
  - Preço do m^2
  - Etc.

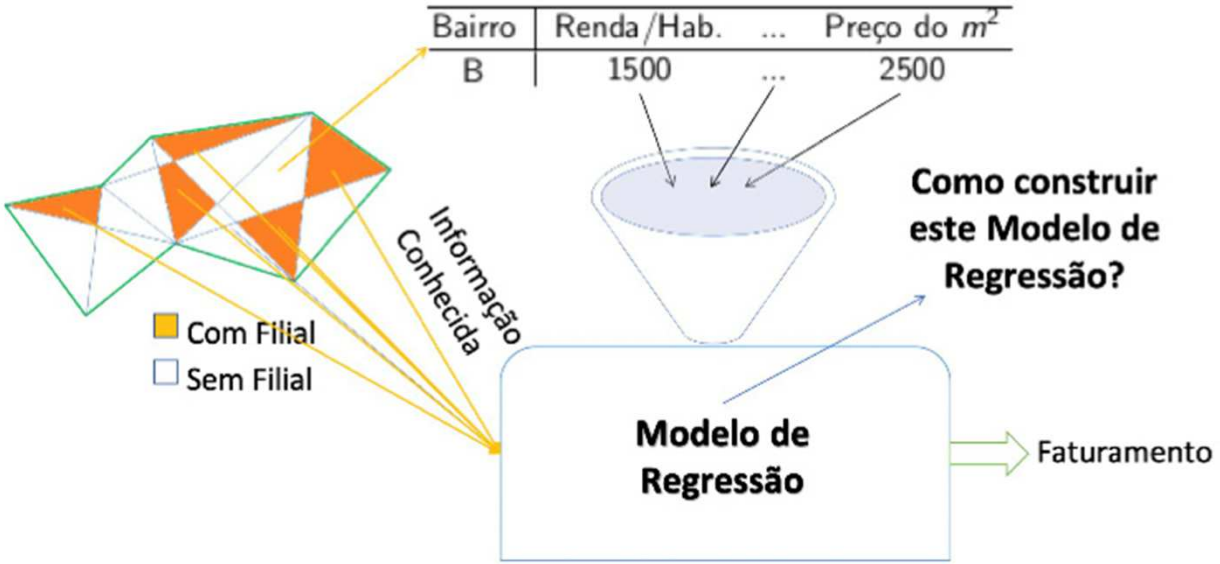
PROBLEMA DE REGRESSÃO: EXEMPLO

- 2º passo: Separar as bases de dados – Com Faturamento e Sem Faturamento

	Bairro	Renda/Hab.	...	Preço do m^2	Faturamento
Com	A	1500	...	2500	105.000
	H	2400	...	5300	180.000
	...	...	...	...	...
	L	3400	...	2750	150.000
Sem	B	2500	...	1780	NA
	C	1000	...	3500	NA
	D	4300	...	6500	NA
	...	...	...	...	...
	K	7000	...	8900	NA
	M	2800	...	3900	NA

PROBLEMA DE REGRESSÃO: EXEMPLO

- 3º passo: Elaborar um modelo de Regressão



PROBLEMA DE REGRESSÃO: EXEMPLO

- **Finalmente:** a partir do modelo, seleciona-se o bairro com **Maior Faturamento Esperado**

	Bairro	Renda/Hab.	...	Preço do m^2	Faturamento
Observado	A	1500	...	2500	105.000
	H	2400	...	5300	180.000
	...	...	...	...	...
	L	3400	...	2750	150.000
Esperado	B	2500	...	1780	125.000
	C	**1000**	...	**3500**	**200.000**
	D	4300	...	6500	
	...	...	...	...	...
	K	7000	...	8900	180.000
	M	2800	...	3900	120.000

PROBLEMA DE REGRESSÃO: DEFINIÇÃO

- Dado um conjunto de *n* **padrões**, em que cada padrão é composto por informação de **variáveis explicativas** (X) e de uma **variável resposta contínua/discreta** (y).

- **Objetivo:** construir um modelo de regressão que dado um novo padrão estime o **valor mais esperado** para a variável resposta.

Padrão	Explicativas			Resposta
$\mathbf{x}_1$	x_{11}	…	x_{1J}	y_1
$\mathbf{x}_2$	x_{21}	…	x_{2J}	y_2
$\mathbf{x}_3$	x_{31}	…	x_{3J}	y_3
…	…	…	…	…
$\mathbf{x}_i$	x_{i1}	…	x_{iJ}	y_i
…	…	…	…	…
$\mathbf{x}_n$	x_{n1}	…	x_{nJ}	y_n

CLASSIFICAÇÃO X REGRESSÃO

- **Classificação:** Resultado categórico
  - Conceder **ou não** crédito para um cliente?

- **Regressão:** Resultado numérico (contínuo ou discreto)
  - Conceder **qual valor** de crédito para um cliente?

Tarefas como:
- Preparação da base de dados;
- Separação em conjuntos de treino e teste;
- Definição dos critérios de parada do algoritmo;
- Treinamento e teste;

são feitas de forma **equivalente**.

REGRESSÃO

- Também chamada de **Estimação**
- Aprendizagem **Supervisionada** a partir de **dados históricos**
- Diferença na **avaliação de saída**: em vez da acurácia, estima-se a **distância** ou o **erro** entre a saída do estimador e a saída desejada: o processo de treinamento do estimador tem por objetivo corrigir o erro observado, buscando minimizar um critério de maneira que os valores estimados estejam próximos dos valores reais no sentido estatístico.
- A Classificação pode ser vista como um **caso particular** da Regressão.

MÉTRICAS DE DESEMPENHO PARA REGRESSÃO

- **RMSE** (*root mean squared error*, ou raiz do erro quadrático médio): quanto **menor**, melhor o modelo.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n} e_j^2}$$

- **MSE** (mean squared error, ou erro quadrático médio): quanto **mais próximo de 0**, melhor o modelo. Fornece uma ideia da magnitude do erro, mas nenhuma ideia da direção.

$$MSE = \frac{1}{n}\sum_{j=1}^{n} e_j^2$$

- **Coeficiente de Determinação (R^2)**: quanto mais próximo de 1, melhor o ajuste do modelo (o quanto y é explicado por x):

$$R^2 = 1 - \frac{SQE}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

- Sendo **SQE** a soma dos erros quadráticos (*sum of squared errors*):

$$SQE = \sum_{j=1}^{n} e_j^2$$

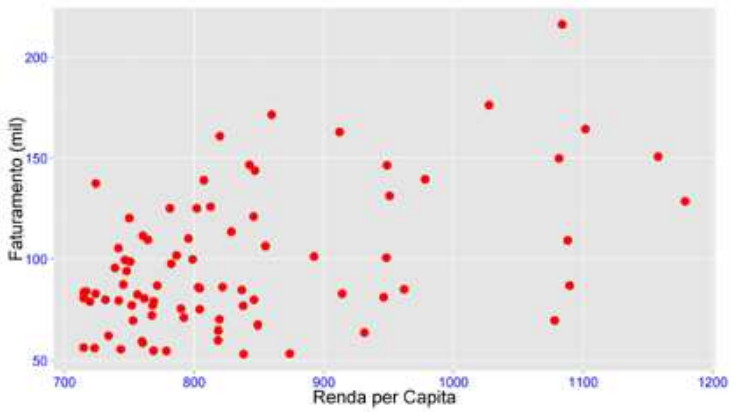
REGRESSÃO LINEAR

REGRESSÃO LINEAR

- No exemplo anterior, vamos somente considerar:
  - os dados que possuem valores para a Variável Resposta
  - a relação entre Variável Explicativa Renda/Hab e o Faturamento

- **Observação trivial:** Quanto maior a Renda per Capita do Bairro, maior o Faturamento no Bairro. Mas como formular esta relação matematicamente?

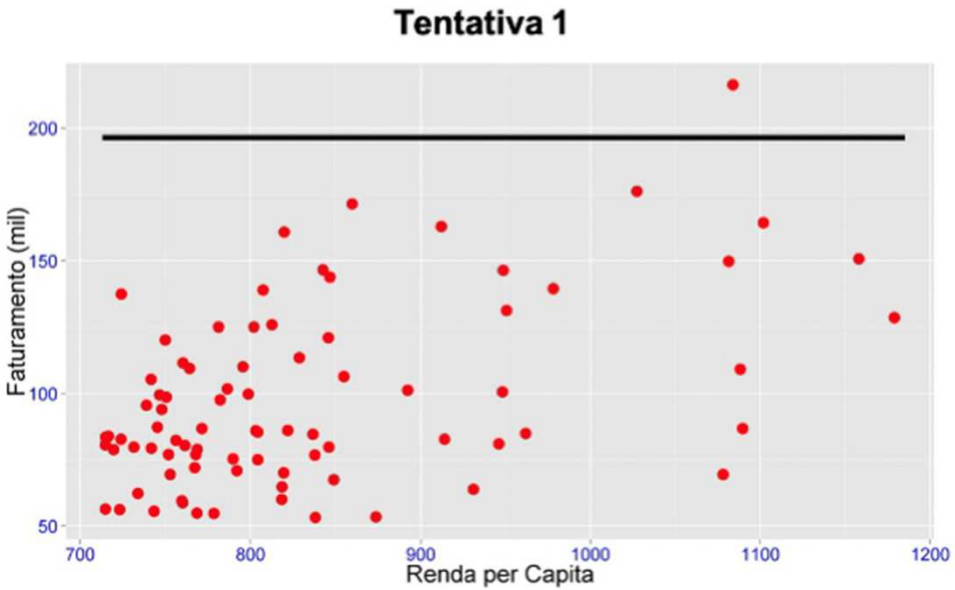
Bairro	Renda/Hab.	Faturamento
A	1500	105.000
H	2400	180.000
...	...	...
L	3400	150.000



REGRESSÃO LINEAR

- **Solução Regressão Linear:** ajustar uma reta que melhor passe pelos pontos.

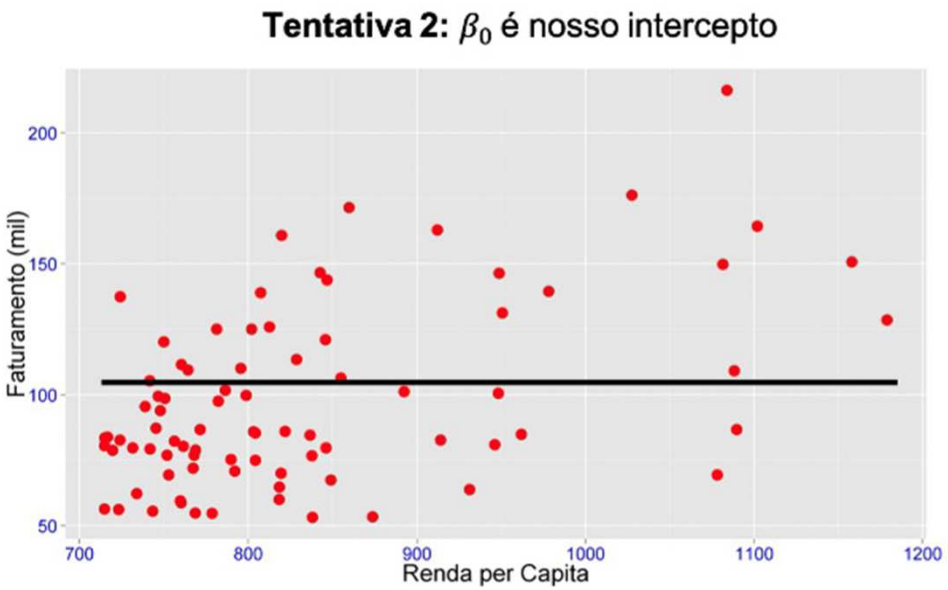
$$\widehat{Fat}_{Bairro} = \beta_0 + \beta_1 \times RendaPerCapita_{Bairro}$$



Modelo: $\widehat{Fat}_{Bairro} = 190$

Matematicamente, o faturamento estimado para todos os bairros é de 190, o que não parece um valor muito real. Vamos tentar novamente, alterando o valor do intecepto (onde a linha corta o eixo y) :

REGRESSÃO LINEAR

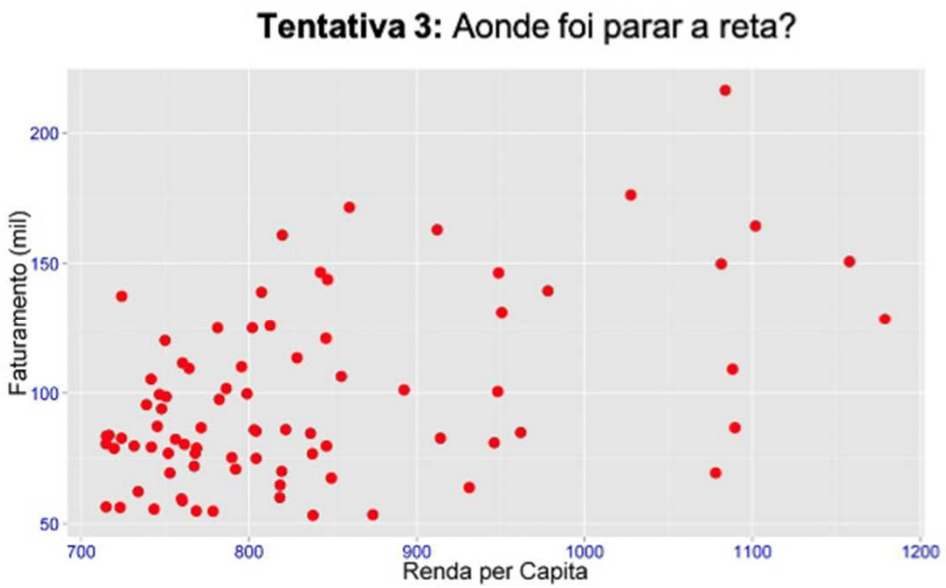


Modelo: $\widehat{Fat}_{Bairro} = 120$

- Agora, o faturamento estimado para todos os bairros é de 120. Apesar de a reta estar um pouco mais próxima dos pontos, este modelo também não parece muito útil.

- Vamos tentar usar a informação de renda per capita, diminuindo, por exemplo, 2 vezes o seu valor pelo faturamento estimado que temos até o momento:

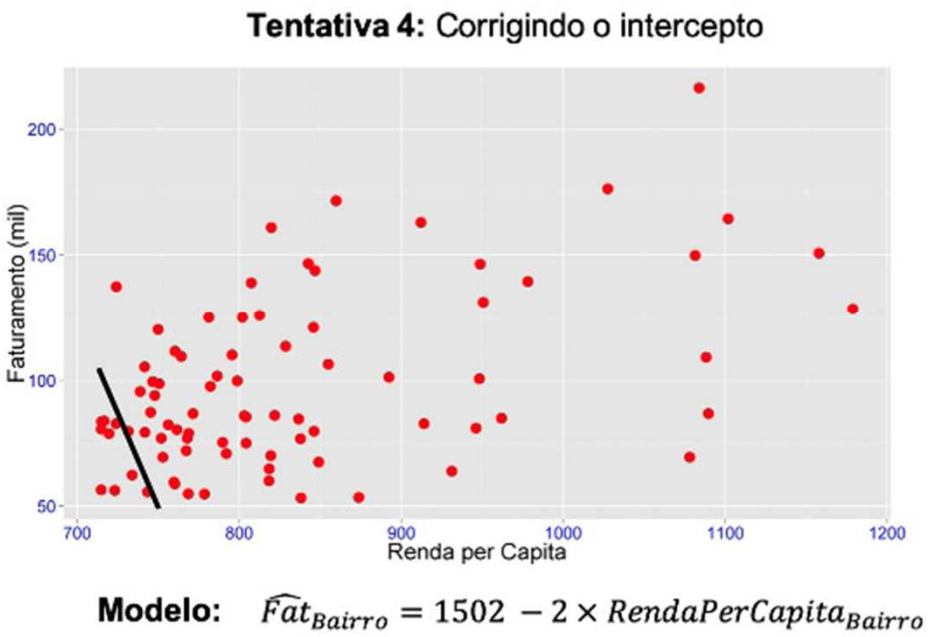
REGRESSÃO LINEAR



Modelo: $\widehat{Fat}_{Bairro} = 120 - 2 \times RendaPerCapita_{Bairro}$

- Agora a reta sumiu do gráfico. Vamos corrigir o intercepto para ver se obtemos melhores resultados:

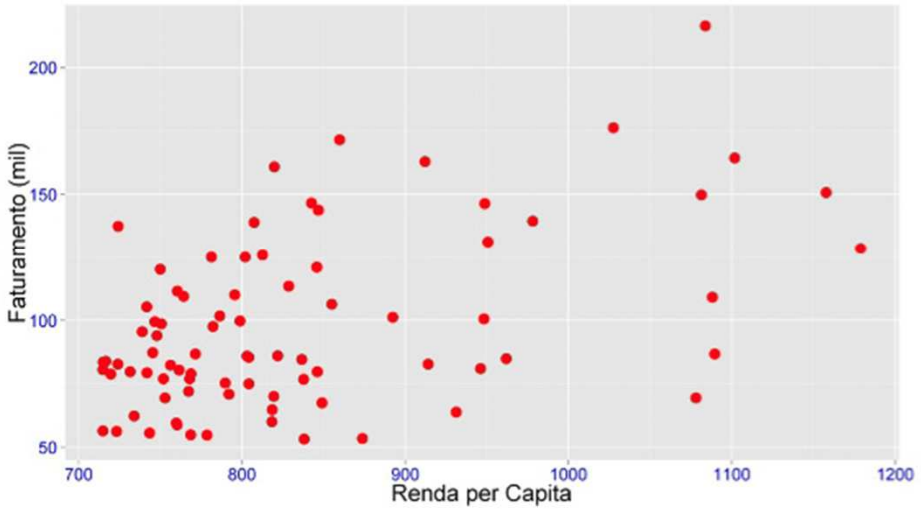
REGRESSÃO LINEAR



- Mas espere... tínhamos observado anteriormente que a relação estre as variáveis renda per capita e faturamento era positiva.
- Vamos trocar o sinal da nossa equação para exprimir essa relação:

REGRESSÃO LINEAR

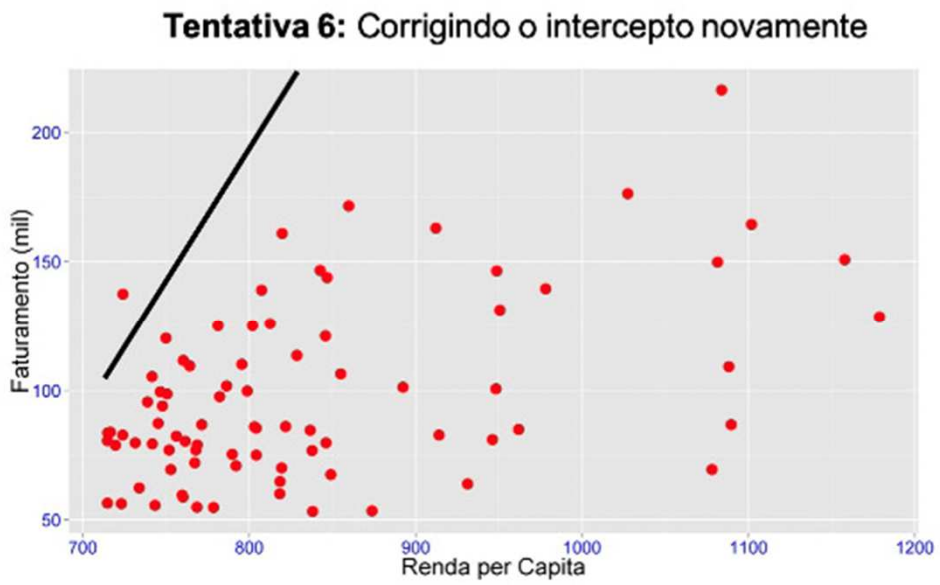
Tentativa 5: Aonde foi parar a reta novamente?



Modelo: $\widehat{Fat}_{Bairro} = 1502 + 2 \times RendaPerCapita_{Bairro}$

- A reta sumiu novamente! Vamos, outra vez, corrigir o intercepto:

REGRESSÃO LINEAR

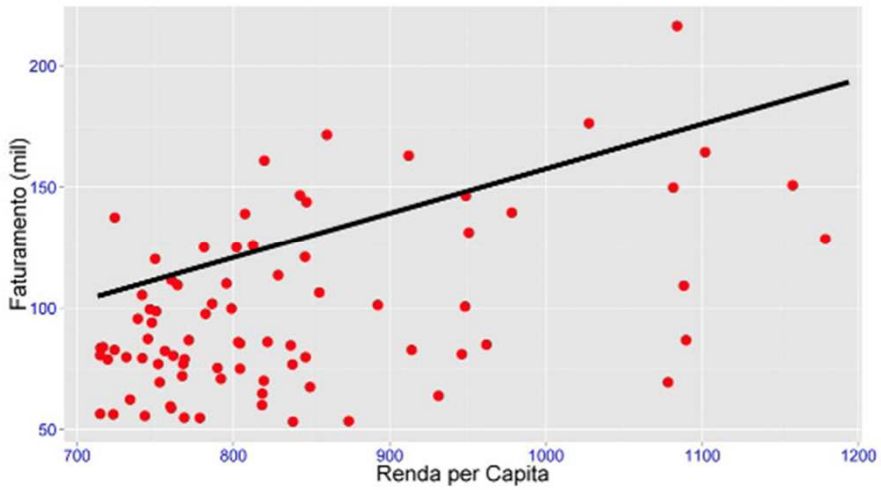


Modelo: $\widehat{Fat}_{Bairro} = -1400 + 2 \times RendaPerCapita_{Bairro}$

- Estamos chegando perto, bastando apenas corrigir um pouco a inclinação da reta:

REGRESSÃO LINEAR

Tentativa 7: Corrigindo a inclinação

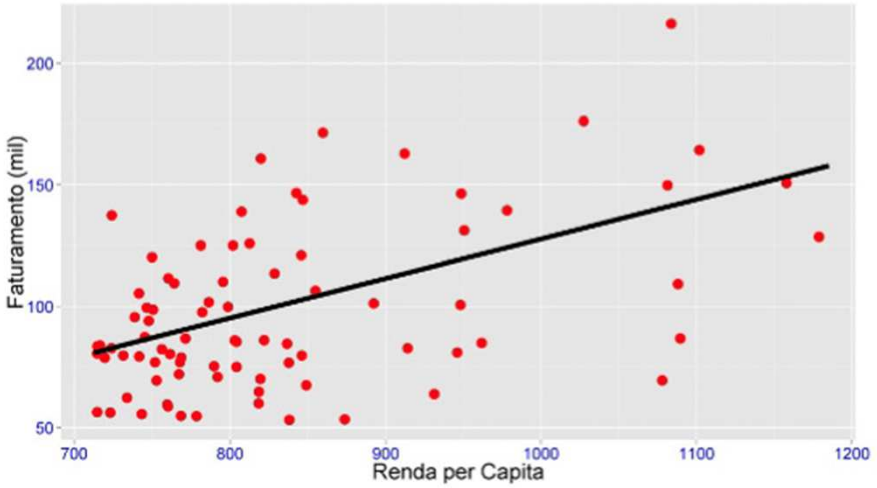


Modelo: $\widehat{Fat}_{Bairro} = -1400 + 1{,}35 \times RendaPerCapita_{Bairro}$

- Se continuarmos neste processo manual (e cansativo!), vamos em algum momento chegar à solução ótima para o problema:

REGRESSÃO LINEAR

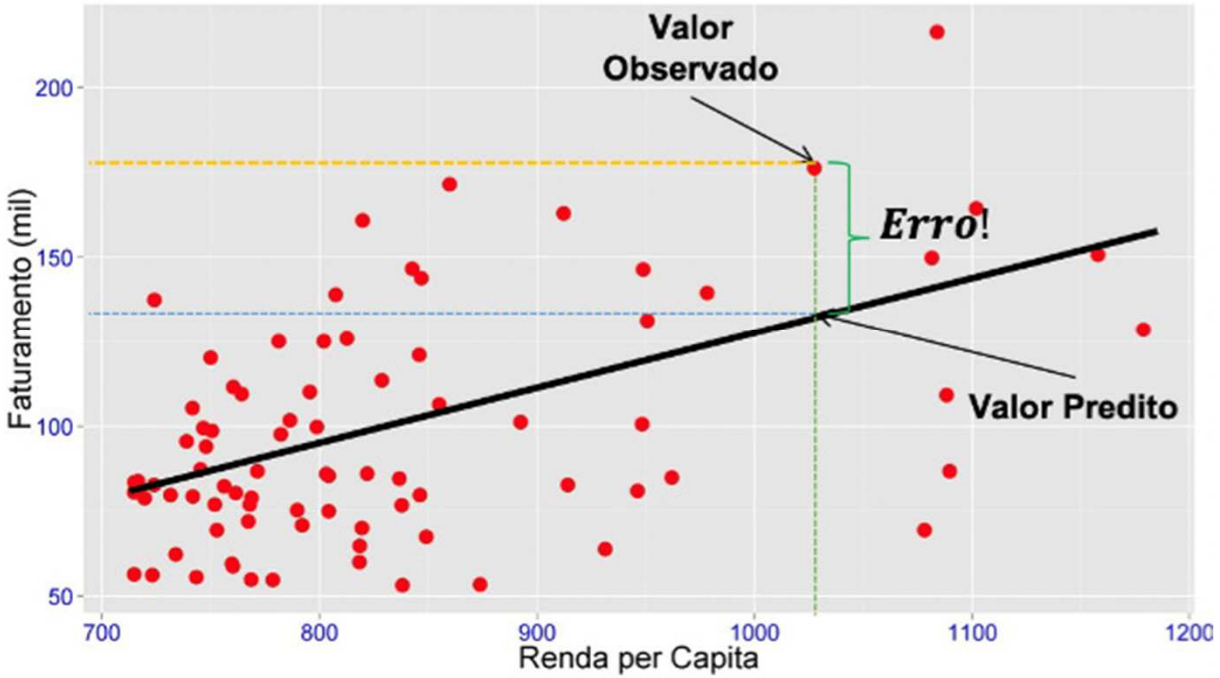
Melhor solução: β_0 e β_1 ótimos



Modelo: $\widehat{Fat}_{Bairro} = -24{,}49 + 0{,}15 \times RendaPerCapita_{Bairro}$

- Este modelo significa que, a cada aumento de R$100 na renda per capita do bairro, espera-se que isso reflita em $0{,}15 * 100 = 15$ mil de faturamento para a filial.

REGRESSÃO LINEAR



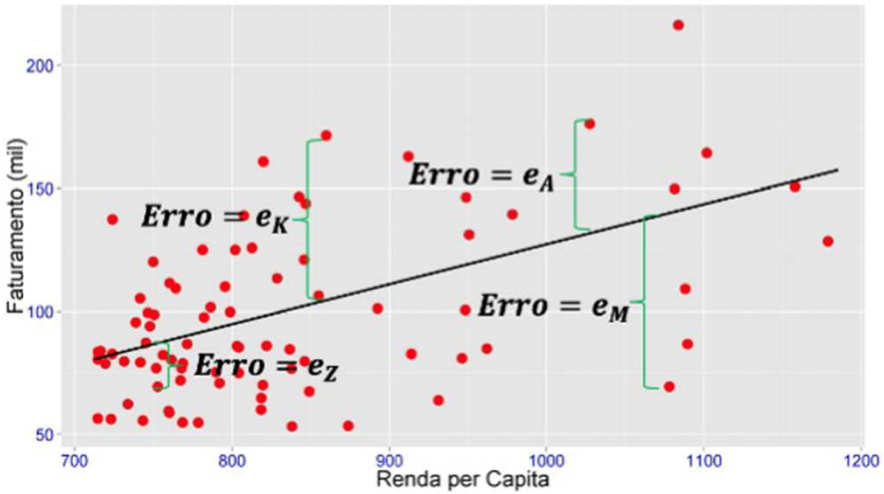
Esta solução é dita ótima porque passa mais perto dos pontos (considerando a distância euclidiana).

REGRESSÃO LINEAR

- Logo, para cada escolha dos parâmetros β_o e β_1 na equação

$$\widehat{Fat}_{Bairro} = \beta_0 + \beta_1 \times RendaPerCapita_{Bairro}$$

- é possível calcular os erros (ou desvios) dessa escolha.



$$\widehat{Fat}_{Bairro} = -24{,}49 + 0{,}15 \times RendaPerCapita_{Bairro}$$

REGRESSÃO LINEAR

- Porém, observe que se somarmos os erros para calcular o erro total do modelo, pelos erros individuais serem positivos e negativos, eles se anulam.

Bairro	Fat	$\hat{Fat}$	$e = Fat - \hat{Fat}$
A	180	130	50
K	175	115	60
M	65	120	-55
...	...	...	...
Z	70	82	-12

- Assim, a melhor prática é trabalhar com a **magnitude do erro** (erro ao quadrado, por e

$$\sum_{i \in Bairro} e_i^2 = (Fat_i - \hat{Fat}_i)^2 = 50^2 + 60^2 + (-55)^2 + ... + (-12)^2 = 230$$

REGRESSÃO LINEAR

- **Moral da história:** a Regressão Linear consiste em escolher β_o e β_1 para construir uma reta que minimize a soma dos quadrados dos erros (SQE):

$$SQE = \sum_{i \in Bairro}^{n} e_i^2$$

Na primeira tentativa, o SQE foi 3224, na segunda, 1224, e assim por diante, até que chegamos na melhor solução, com SQE 230.

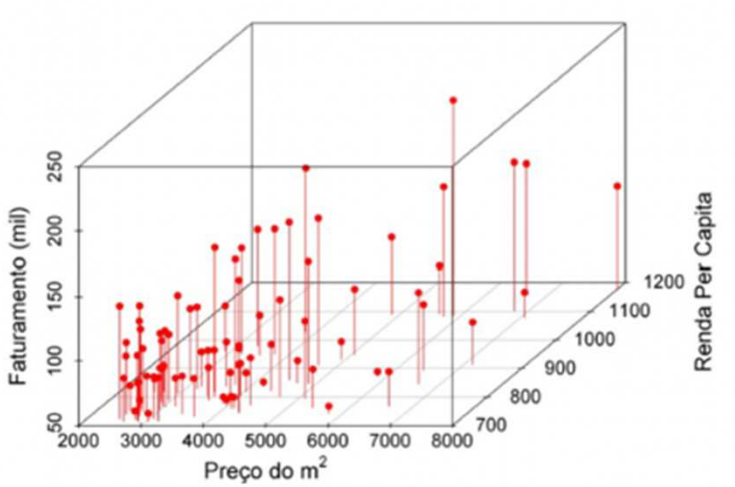
- Neste exemplo consideramos apenas a relação entre Faturamento e Renda Per Capita. **Em problemas reais, dificilmente haverá uma única variável x capaz de prever o y!**

REGRESSÃO LINEAR MÚLTIPLA

- Se quiséssemos adicionar uma nova variável (como por exemplo preço do m^2), teríamos o modelo de **regressão linear múltipla**:

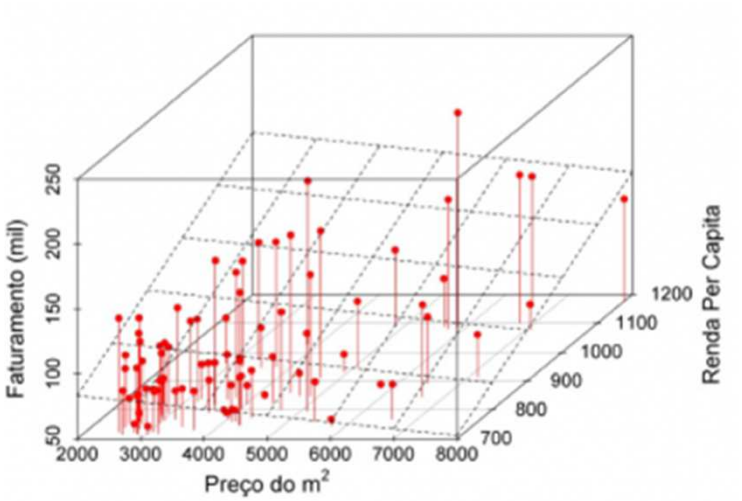
$$\widehat{Fat}_i = \beta_0 + \beta_1 \times RendaPerCapita_i + \beta_2 \times PrecoM^2$$

Basta adicionar as demais variáveis preditoras (x_1, x_2, ..., x_n) e seus coeficientes correspondentes!



REGRESSÃO LINEAR MÚLTIPLA

- A solução seria construir um **plano** (em vez de uma reta) que melhor se ajuste aos pontos:



- Neste caso, os coeficientes podem ser estimados pelo **Método dos Mínimos Quadrados*** (*Ordinary Least Squares*).
- Este método busca encontrar o **melhor ajuste** para um conjunto de dados tentando **minimizar a soma dos quadrados das diferenças** entre o valor estimado e os dados observados.

*

https://pt.wikipedia.org/wiki/M%C3%A9todo_dos_m%C3%ADnimos_quadrados
https://towardsdatascience.com/linear-regression-understanding-the-theory-

REGRESSÃO LINEAR: RESUMO

- A Regressão Linear modela a relação entre uma variável de **resposta** (y) e os **preditores** (X).
- Assume-se um relacionamento **linear** entre X e y, ou seja, que y pode ser calculado através de uma **combinação linear** de X.
  - Apenas um x: regressão linear simples
  - Mais de um x: regressão linear múltipla

REGRESSÃO LINEAR: RESUMO

- Corresponde ao problema de estimar uma **função** a partir de pares entrada-saída.
  - **Simples**: $y = \beta_o + \beta_1 x$, sendo β_0 e β_1 os coeficientes de regressão (especificam o intercepto do eixo y e a inclinação da reta)
  - **Múltipla**: a equação deve ser estendida para equação de plano/hiperplano: $y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \ldots$
- A solução da tarefa de regressão consiste em **encontrar valores** para os **coeficientes de regressão** de forma que a reta (ou plano/hiperplano) se **ajuste** aos valores assumidos pelas variáveis no conjunto de dados.

REGRESSÃO LINEAR: AVALIAÇÃO

- A saída de um estimador é um valor numérico contínuo que deve ser **o mais próximo possível** do valor desejado, e a diferença entre esses valores fornece uma **medida de erro** de estimação do algoritmo.
  - Seja $d_j = 1, ..., n$ a resposta desejada para o objeto j e y_j a resposta predita do algoritmo, obtida a partir da entrada x_j. Então, $e_j = d_j - y_j$ é o **erro** observado na saída do sistema para o objeto j.

REGRESSÃO LINEAR: AVALIAÇÃO

- O processo de treinamento do estimador tem por objetivo **corrigir** este erro observado e, para tal, busca **minimizar** um critério (função objetivo) baseado em e_j, de maneira que os valores de y_j estejam próximos dos de d_j no sentido estatístico.
- Se a equação de regressão aproxima suficiente bem os dados de treinamento, então ela pode ser usada para **estimar** o valor de uma variável (y) a partir do valor da outra variável (x).
- **Resumindo:** a regressão linear procura pelos coeficientes da reta que minimizam a distância dos objetos à reta.
- *OBS: apesar da sua simplicidade, modelos lineares são surpreendentemente competitivos em relação a modelos não lineares.*

MÉTODOS DE REGULARIZAÇÃO

- A regressão linear usa o método dos mínimos quadrados para estimar os seus coeficientes, buscando minimizar a soma dos quadrados (RSS – *residual sum of squares*).
- No caso de um coeficiente ser zero, a influência da variável de entrada no modelo é removida (0 * x = 0).
- O modelo se torna menos complexo e com melhor interpretabilidade, pois as variáveis não-relevantes (que não estão realmente associadas à resposta) são eliminadas.

MÉTODOS DE REGULARIZAÇÃO

- Os métodos de regularização são extensões de treinamento do modelo linear que, além de buscar minimizar a soma do erro quadrático, buscam reduzir a complexidade do modelo através de uma função de penalidade:
  - **Ridge:** busca minimizar também o quadrado da soma absoluta dos coeficientes (regularização L2).
  - **Lasso:** busca minimizar também o valor absoluto da soma dos coeficientes (regularização L1).
- Esses métodos são eficazes quando há correlação entre os atributos de entrada, e os mínimos quadrados comuns superestimam os dados de treinamento, ocorrendo o *overfitting*.

Ridge: $RSS + \lambda \sum_{j=1}^{p} \beta_j^2$ Lasso: $RSS + \lambda \sum_{j=1}^{p} |\beta_j|$

MÉTODOS DE REGULARIZAÇÃO

- O parâmetro de ajuste λ serve para controlar o impacto da penalidade.
  - Quando λ = 0, o termo da penalidade não tem efeito e o resultado é similar ao método dos mínimos quadrados.
  - Quanto maior é λ, maior é o impacto da penalidade e maior a diminuição dos coeficientes.
- **Ridge:** A penalidade poderá diminuir todos os coeficientes para próximo de zero, mas nunca exatamente a zero. O modelo gerado sempre terá todas as variáveis preditoras e não é robusto a *outliers*, podendo prejudicar a interpretabilidade do modelo, mas sendo capaz de aprender padrões mais complexos.
- **Lasso:** A penalidade pode levar alguns coeficientes a exatamente zero (quando λ é suficientemente grande), realizando a seleção de variáveis preditoras e facilitando a interpretabilidade do modelo, que é mais simples e robusto a *outliers*, mas não é capaz de aprender padrões mais complexos.

Para saber mais: *https://towardsdatascience.com/intro-to-linear-model-selection-and-regularization-d47bd2c5d54*
https://medium.com/datadriveninvestor/l1-l2-regularization-7f1b4fe948f2

REGRESSÃO LOGÍSTICA

REGRESSÃO LOGÍSTICA

- Não se confunda! É um algoritmo de **classificação** e não de regressão.
- Usado para estimar valores discretos (valores binários como 0/1, sim / não, verdadeiro / falso) com base em um conjunto de variáveis independentes.
- Calcula a probabilidade de ocorrência de um evento, ajustando os dados a uma função *logit* (por isso, também é conhecido como regressão logit).

$$\operatorname{logit}(p) = \log\left(\frac{p}{1-p}\right) = \log(p) - \log(1-p).$$

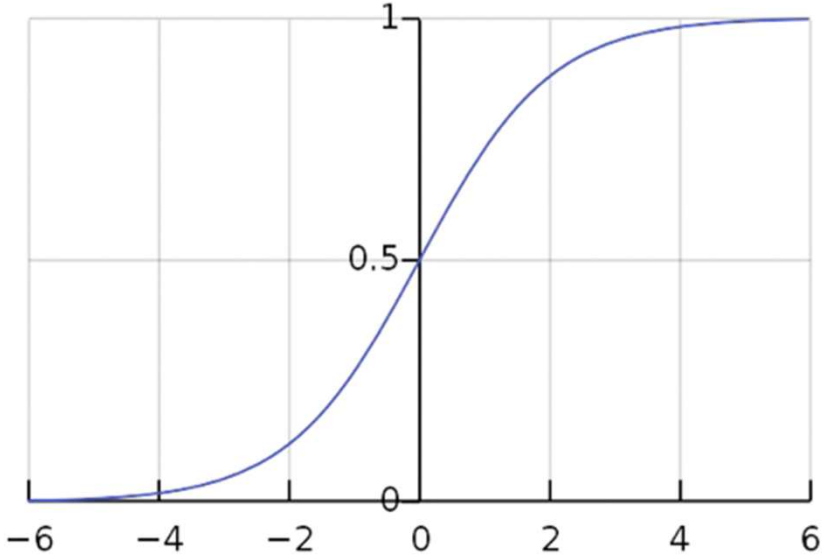
- Como prevê a probabilidade, seus valores de saída estão entre 0 e 1 (como esperado).

REGRESSÃO LOGÍSTICA

- Utiliza função **logística**, também chamada de **sigmóide**: uma curva em forma de S que pode mapear qualquer número em um intervalo entre 0 e 1 (mas nunca exatamente nestes limites).

$$f(x) = \frac{1}{1 + e^{-x}}$$

- onde f(x) é a saída prevista.



Transformação logística entre -6 e 6 (Fonte: Wikipedia)

REGRESSÃO LOGÍSTICA

- De forma similar à regressão linear, usa uma equação como representação:
  - os valores de entrada (x) são combinados linearmente usando coeficientes para prever um valor de saída (y).

- O valor de saída é modelado em valor binário (0 ou 1) em vez de um valor numérico.

REGRESSÃO LOGÍSTICA - PROBABILIDADES

- A regressão logística modela a probabilidade da classe padrão. Por exemplo, se estivermos modelando o sexo de uma pessoa dada sua altura, o modelo de regressão logística pode ser escrito como:

$$P(sexo = masculino \mid altura)$$

- Escrito de outra forma, estamos modelando a probabilidade de uma entrada X pertencer à classe padrão (Y = 1):

$$P(X) = P(Y=1 \mid X)$$

- Note que a predição da probabilidade pode ser transformada em um valor binário (0 ou 1) para fazer a classificação.

REGRESSÃO LOGÍSTICA – ESTIMANDO OS COEFICIENTES

- Os coeficientes da Regressão Logística podem ser estimados usando os dados de treinamento, através do método de estimação de máxima verossimilhança*, um método matemático que busca valores para os coeficientes de forma a minimizar o erro nas probabilidades preditas pelo modelo para os dados.
  - Os melhores coeficientes resultarão em um modelo que vai prever um valor muito próximo de 1 para a classe padrão e um valor muito próximo de 0 para a outra classe.

- Após determinados os coeficientes, para fazer predições com a Regressão Logística, basta calcular os coeficientes e aplicar a equação resultante.

*
https://pt.wikipedia.org/wiki/M%C3%A1xima_verossimilhan%

KNN PARA REGRESSÃO

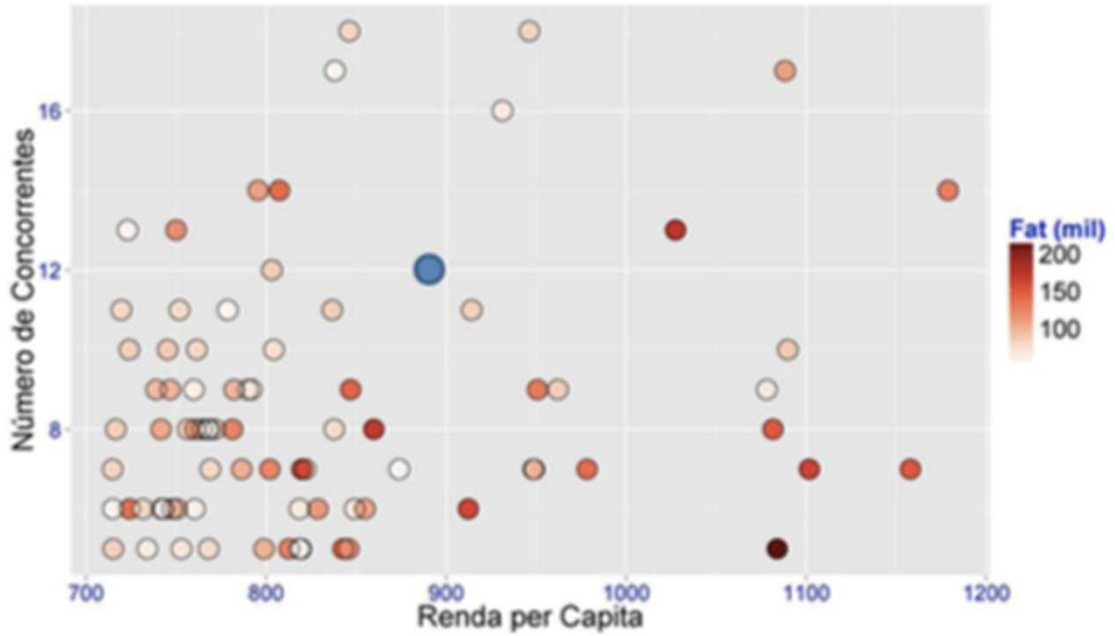
KNN PARA REGRESSÃO

- Lembrando o problema anterior:

	Bairro	Renda/Hab.	# Concorrentes	Faturamento
Com	A	1500	14	105.000
	H	2400	8	180.000
	...	...	...	...
	L	3400	10	150.000
Sem	B	20	1780	NA
	C	1000	5	NA
	D	4300	15	NA
	...	...	...	...
	K	7000	9	NA
	M	2800	7	NA

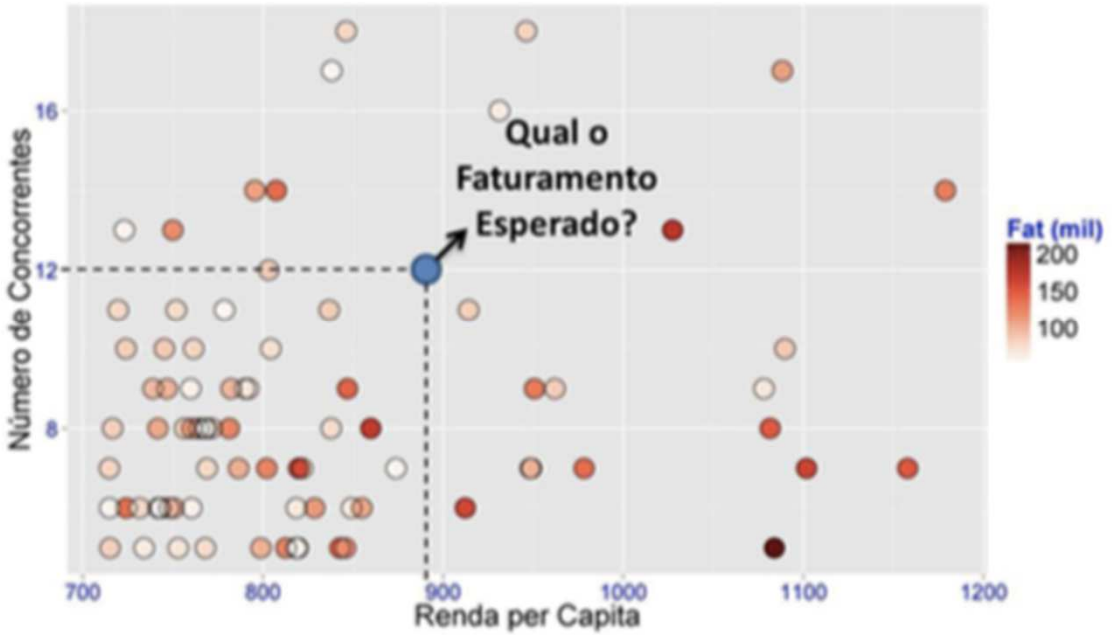
KNN PARA REGRESSÃO

- Ao organizar esses dados em um gráfico, tem-se:



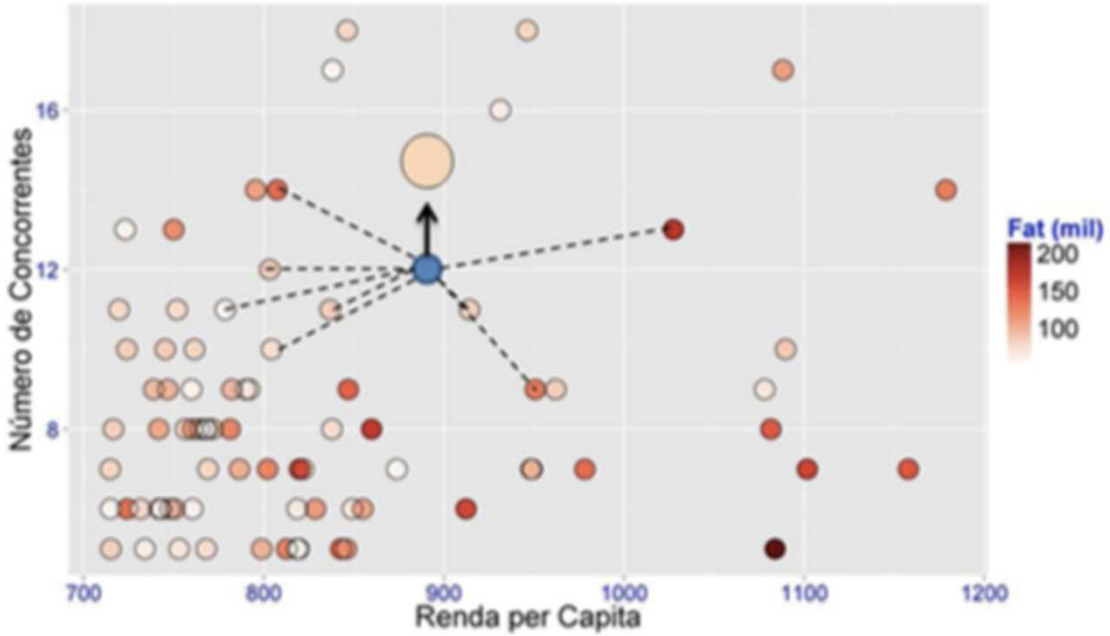
KNN PARA REGRESSÃO

- Pergunta:



KNN PARA REGRESSÃO

- **KNN para Regressão:** O Faturamento Esperado é a média aritmética dos k vizinhos mais próximos



ÁRVORE DE REGRESSÃO

ÁRVORE DE REGRESSÃO

- É construída de forma **similar** à árvore de classificação: a partir do nó raiz, os dados são particionados usando uma estratégia de divisão e conquista de acordo com a característica que resultará no resultado mais homogêneo após a separação ser realizada.
- Nas árvores de classificação, a homogeneidade é medida pela entropia; nas árvores de regressão, a homogeneidade é medida por **estatísticas** como variância, desvio padrão ou desvio absoluto da média.
- A predição é a **média** dos valores dos exemplos de cada folha.

ÁRVORE DE REGRESSÃO: CRITÉRIO DE DIVISÃO

- Um critério de divisão comum para árvores de regressão é a redução de desvio padrão (*SDR – Standard Deviation Reduction*), que mede a **redução no desvio padrão**, comparando o desvio padrão antes da divisão com o desvio padrão ponderado após a divisão.

$$\mathrm{SDR} = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i)$$

- A função *sd(T)* refere-se ao desvio padrão dos valores no conjunto *T*, enquanto T_1, T_2, ..., T_n são os conjuntos de valores resultantes de uma divisão em uma característica. |T| refere-se ao número de observações no conjunto T.

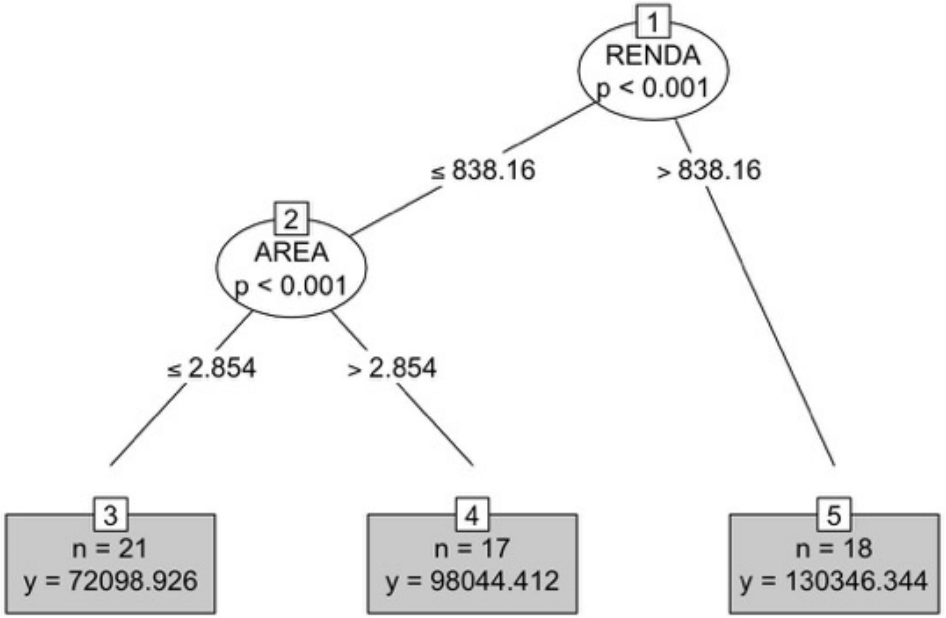
ÁRVORE DE REGRESSÃO: PREDIÇÃO - EXEMPLO

- Se o desvio padrão é **mais reduzido** com a divisão em uma característica B do que em A, deve-se realizar a divisão **primeiro em A**, resultando em uma árvore mais homogênea.

original data	1	1	1	2	2	3	4	5	5	6	6	7	7	7	7
split on feature A	1	1	1	2	2	3	4	5	5	6	6	7	7	7	7
split on feature B	1	1	1	2	2	3	4	5	5	6	6	7	7	7	7
				T_1								T_2			

- Supondo que a árvore de regressão está pronta com apenas esta divisão em B, a predição pode ser feita considerando se o novo exemplo caiu no conjunto T_1 ou no conjunto T_2, calculando a media dos valores deste conjunto.

ÁRVORE DE REGRESSÃO



SVM PARA REGRESSÃO

SVM PARA REGRESSÃO

- O SVM também pode ser usado como método de regressão, mantendo todos os principais recursos que caracterizam o algoritmo, como a margem máxima), sendo chamado **SVR** (Support Vector Regression).
- O SVR foi proposto em 1997, mas é **pouco utilizado**, pois existem modelos mais simples para regressão com resultados semelhantes ou melhores.
- Assim como no SVM para Classificação, o modelo produzido pelo SVR depende apenas de um subconjunto dos dados de treinamento.

SVM PARA REGRESSÃO

- A ideia básica do SVR é mapear um conjunto de dados X em um espaço multidimensional, através um mapeamento não-linear (usando funções kernel) e realizar uma regressão linear neste espaço transformado, considerando apenas os pontos que estão dentro da margem. O melhor modelo é o hiperplano que possui o número máximo de pontos.
- Comparado ao modelo de Regressão Linear, o SVR tem a vantagem de utilizar uma grande variedade de funções que se adequa aos modelos.
- Na regressão linear, tentamos minimizar a taxa de erro, enquanto que no SVR, tentamos ajustar o erro **dentro de um determinado limite**, definido pela margem.

SVM PARA REGRESSÃO