Modelos de Regressão Analítica

Modelos de Regressão

Como Começar:

Modelos de machine learning:

- classificação,
- **regressão**,
- clusterização,
- detecção de anomalias

Modelos de Regressão

Introdução a Regressão:

- É uma das técnicas **mais utilizadas** na academia e no mercado;
- A finalidade é **estimar valores e a relação entre variáveis**, com base em **valores conhecidos**;
- Preferencialmente o modelo de regressão deve ser **definido com base na teoria e na experiência** do pesquisador;
- **Evitando relações absurdas** dentro de um contexto de pesquisa específico;

PREDIÇÃO

Predição de demanda, prever itens que serão comprados.

SUPORTE A DECISÕES

Ferramenta no auxílio a tomada de decisão, fornecendo modelos de previsão.

INSIGHTS

Encontra relações entre variáveis ainda não percebidas.

OTIMIZAR PROCESSOS

Análise da relação entre espera ao telefone e reclamações.

CORREÇÃO DE ERROS

Fornece suporte quantitativo para decisões e evita erros baseados em intuições.

Modelos de Regressão

Introdução a Regressão:

- ✓ Existem diversos tipos de modelos de regressão.
- ✓ Conhecidos como G.L.M (*Modelos Generalizados*).

Modelos de Regressão	Variável Dependente	Distribuição
Regressão Linear	Quantitativa	Normal
Logística Binária	Binária(Sucesso/Fracasso)	Bernoulli
Logística Multinominal	Categórica(>2)	Binomial
Logística Ordinal	Valores Ordenados	Binomial
Regressão de Poisson	Contagem/Taxa (+)	Poisson
Regressão Binomial Negativa	Contagem(-/+)	Binomial Negativa
Regressão de Cox	Tempo (Sobrevivência)	Exponencial Weibull Log-Normal

Modelos de Regressão

Introdução a Regressão:

[**RStudio*: 15 Tipos de Modelos de Regressão:]

- Regressão Linear
  - Regressão Polinomial
  - Regressão Ridge
  - Regressão Lasso
  - Regressão ElasticNet
  - Regressão Quantílica
- Regressão Logística
  - Regressão Ordinal
  - Regressão Multinominal

- Regressão de Componentes Principais
- Regressão por Mínimos Quadrados Parciais
- Regressão Vetorial de Suporte
- Regressão de Poisson
- Regressão Binomail Negativa
- Regressão Quasi-Poisson
- Regressão de Cox

Modelos de Regressão

Introdução a Regressão:

Modelos de Regressão	Função
Regressão Linear	$\hat{Y} = \alpha + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \mu_i$
Logística Binária (0,1)	$f(\hat{Y}) = \left(\frac{1}{1+e^{-\hat{Y}}}\right)$ $ln\left(\frac{p}{1-p}\right) = \hat{Y} = \alpha + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$
Logística Multinominal (0,1,2)	$(1)\ ln\left(\frac{P(1\|X)}{P(0\|X)}\right) = \hat{Y} = \alpha + \sum \beta_{1i} X_{1i}$ $(2)\ ln\left(\frac{P(2\|X)}{P(0\|X)}\right) = \hat{Y} = \alpha + \sum \beta_{2i} X_{2i}$
Logística Ordinal	$f(\hat{Y}) = \left(\frac{e^{\hat{Y}}}{1+e^{\hat{Y}}}\right)$
Regressão de Poisson	$ln(\lambda_i) = \hat{Y} = \alpha + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$
Regressão Binomial Negativa	$ln(\lambda_i) = \hat{Y} = \alpha + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \mu_i$
Regressão de Cox	$\lambda\,(t) = \hat{Y} = \lambda_0(t) + e^{\beta_1 X_{1i} + \cdots + \beta_k X_{ki}}$

Modelos de Regressão

Introdução a Regressão:

- **Correlação:** mede a ***força da relação*** entre duas variáveis quantitativas.
- **Regressão:** mede a ***relação*** entre duas variáveis quantitativas.

Modelo de Regressão	Função (variável dependente)
Regressão Linear ***Múltipla***	$\hat{Y} = \alpha + \beta_1 . X_{1i}$ ***Simples***
Logística Binária	$\ln\left(\frac{p}{1-p}\right) = \hat{Y} = \alpha + \beta_1 . X_{1i} + \beta_2 . X_{2i} + \cdots + \beta_k . X_{ki}$

Modelos de Regressão

Regressão Linear Simples:

1. Determinar como duas variáveis se relacionam.

2. Estimar a **função que determina a relação** entre as variáveis.

3. Usar a equação ajustada para **prever valores da variável dependente**:

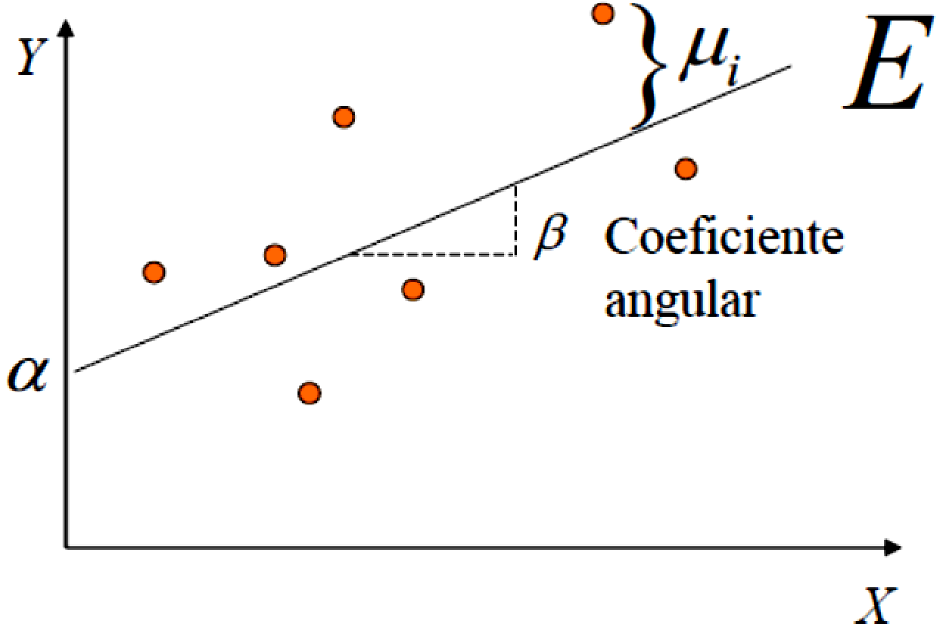
Modelo de Regressão Linear Simples

$$Y_i = \alpha + \beta X_i + \mu_i$$

onde, $\mu_i = (Y_i - \hat{Y}_i)$

Modelos de Regressão

Regressão Linear Simples:



$$E(Y) = \alpha + \beta x_i$$

Inclinação populacional

Intercepto populacional

Erro Aleatório

$$Y_i = \alpha + \beta X_i + \mu_i$$

Variável Dependente

Variável Independente

Modelos de Regressão

Regressão Linear Simples:

- **Estimadores α e β**

$$\boxed{\beta = \frac{\sum_{i=1}^{n}(X_i - \bar{X}).(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

$$\boxed{\alpha = \bar{Y} - \beta.\bar{X}}$$

Modelos de Regressão

Regressão Linear Simples: • *Exemplo*

- ✓ Deseja-se saber, para uma turma de 10 alunos, qual a influência da distância percorrida para se chegar à escola em relação ao tempo de percurso;

- ✓ Elaborou-se um questionário e aplicou para os 10 alunos da turma;

- ✓ Com as variáveis: (*y)Tempo para se chegar a escola* e (x) *Distância percorrida até a escola*.

$$tempo_i = \alpha + \beta . dist_i + \mu_i$$

Modelos de Regressão

Regressão Linear Simples: • *Exemplo*

✓ A tabela abaixo mostra os resultados declarados pelos alunos:

Estudante	Tempo para chegar à escola (min)	Distância percorrida até a escola (km)
Gabriela	15	8
Dalila	20	6
Gustavo	20	15
Letícia	40	20
Luiz Otávio	50	25
Leonor	25	11
Ana	10	5
Antônio	55	32
Júlia	35	28
Mariana	30	20

Modelos de Regressão

Regressão Linear Simples: • *Exemplo*

✓ Constrói-se a tabela de valores:

Estudante	**Tempo (Y_i)**	**Distância (X_i)**	$(Y_i - \bar{Y}_i)$	$(X_i - \bar{X}_i)$	$(Y_i - \bar{Y}_i)*(X_i - \bar{X}_i)$	$(X_i - \bar{X}_i)^2$
Gabriela	15	8	-15	-9	135	81
Dalila	20	6	-10	-11	110	121
Gustavo	20	15	-10	-2	20	4
Letícia	40	20	10	3	30	9
Luiz Otávio	50	25	20	8	160	64
Leonor	25	11	-5	-6	30	36
Ana	10	5	-20	-12	240	144
Antônio	55	32	25	15	375	225
Júlia	35	28	5	11	55	121
Mariana	30	20	0	3	0	9
Soma	**300**	**170**			**1155**	**814**
Média	***30***	***17***				

Modelos de Regressão

Regressão Linear Simples: • *Exemplo*

✓ Por meio da planilha construída podemos calcular os parâmetros αe β:

$$\beta = \frac{\sum_{i=1}^{n}(X_i - \bar{X}).(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \Rightarrow \beta = \frac{1155}{814} \Rightarrow \boxed{\beta = 1{,}4189}$$

$$\alpha = \bar{Y} - \beta.\bar{X} \Rightarrow \alpha = 30 - 1{,}4189 * 17 \Rightarrow \boxed{\alpha = 5{,}8784}$$

$$\boxed{tempo_i = 5{,}8784 + 1{,}4189 * dist_i}$$

Distância (Xi)	Tempo (Yi)	
0	5,88	
1	7,30	1,42
2	8,72	1,42
3	10,14	1,42

Modelos de Regressão

Regressão Linear Simples:

- ✓ O **coeficiente de determinação** é uma medida da proporção da variabilidade em uma variável que é explicada pela variabilidade da outra.
- ✓ O coeficiente está entre $0 \leq R^2 \leq 1$.
- ✓ O coeficiente de determinação é definido por:

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}(u_i)^2}$$

$$SQ_{\text{tot}} = SQ_{\text{exp}} + SQ_{\text{res}}.$$

$$SQ_{\text{tot}} = \sum_{i=1}^{n}(y_i - \bar{y})^2,$$

$$SQ_{\text{res}} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad SQ_{\text{exp}} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$R^2 = \frac{SQ_{\text{exp}}}{SQ_{\text{tot}}} = 1 - \frac{SQ_{\text{res}}}{SQ_{\text{tot}}}$$

Modelos de Regressão

Regressão Linear Simples: • *Exemplo*

✓ Construindo a tabela, temos...:

Estudante	Tempo (Y_i)	Distância (X_i)	$\hat{Y}_i$	$\mu_i = (Y_i - \hat{Y}_i)$	$(\hat{Y}_i - \bar{Y}_i)$	μ_i^2
Gabriela	15	8	17,23	-2,23	163,08	4,97
Dalila	20	6	14,39	5,61	243,61	31,45
Gustavo	20	15	27,16	-7,16	8,05	51,3
Letícia	40	20	34,26	5,74	18,12	32,98
Luiz Otávio	50	25	41,35	8,65	128,85	74,8
Leonor	25	11	21,49	3,51	72,48	12,34
Ana	10	5	12,97	-2,97	289,92	8,84
Antônio	55	32	51,28	3,72	453,00	13,81
Júlia	35	28	45,61	-10,61	243,61	112,53
Mariana	30	20	34,26	-4,26	18,12	18,12
Soma	300	170			**1638,85**	**361,15**
Média	***30***	***17***				

Modelos de Regressão

Regressão Linear Simples: • *Exemplo*

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(u_i)^2} \Rightarrow R^2 = \frac{1638{,}85}{1638{,}85 + 361{,}15} \Rightarrow \boxed{R^2 = 0{,}8194}$$

- Podemos afirmar que, para a amostra estudada, 81,94% da variabilidade do tempo para se chegar a escola pode ser explicado pela variável distância.

Modelos de Regressão

Regressão Linear Simples:

- ✓ O **coeficiente de determinação** é uma medida da proporção da variabilidade em uma variável que é explicada pela variabilidade da outra.

- ✓ O coeficiente está entre $0 \leq R^2 \leq 1$.

- ✓ O coeficiente de determinação é definido por:

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}(u_i)^2}$$

$$SQ_{\text{tot}} = SQ_{\text{exp}} + SQ_{\text{res}}.$$

$$SQ_{\text{tot}} = \sum_{i=1}^{n}(y_i - \bar{y})^2,$$

$$SQ_{\text{res}} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad SQ_{\text{exp}} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$R^2 = \frac{SQ_{\text{exp}}}{SQ_{\text{tot}}} = 1 - \frac{SQ_{\text{res}}}{SQ_{\text{tot}}}$$

Modelos de Regressão

Regressão Linear Simples: • *Exemplo*

- **Teste de significância dos parâmetros:**

Existe realmente alguma relação linear entre X e Y ?

✓ A estatística F é uma estatística para testar: $\begin{cases} H_0: \beta = 0 \\ H_1: \beta \neq 0 \end{cases}$

$$F = \frac{\text{QMReg}}{\text{QMRes}} \sim F_{1;n-2}$$

se H_0 **verdadeiro** (Não existe relação linear)
se H_0 **falso** (existe relação linear)

se $p > 0.05$, β=0 não existe relação linear explicando Y em função de X.

Modelos de Regressão

Regressão Linear Simples:

Pressupostos do modelo de Regressão Linear

- A relação entre *X* e *Y* é Linear.
- Os valores de X são fixos, isto é, X não é uma variável aleatória.
- A média dos erros é nula, isto é:

$$E(\mu_i) = 0 \qquad i = 1, 2, \ldots, n$$

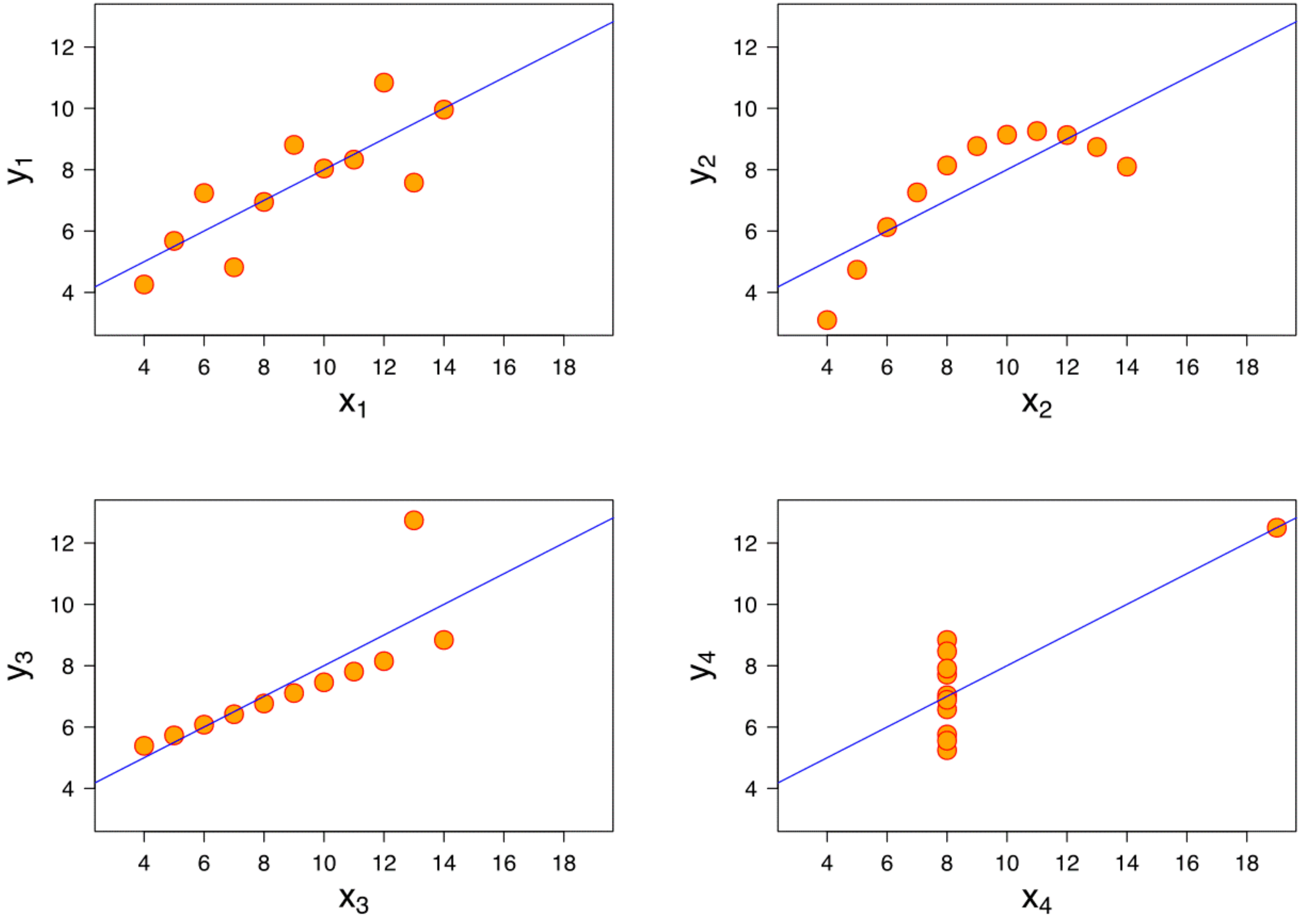
- O erro em uma observação é **não correlacionado** com o erro em qualquer outra observação.

$$Var(\mu_i) = \sigma^2$$

- Os erros têm **distribuição normal.** $\mu_i \sim N(0, \sigma^2)$

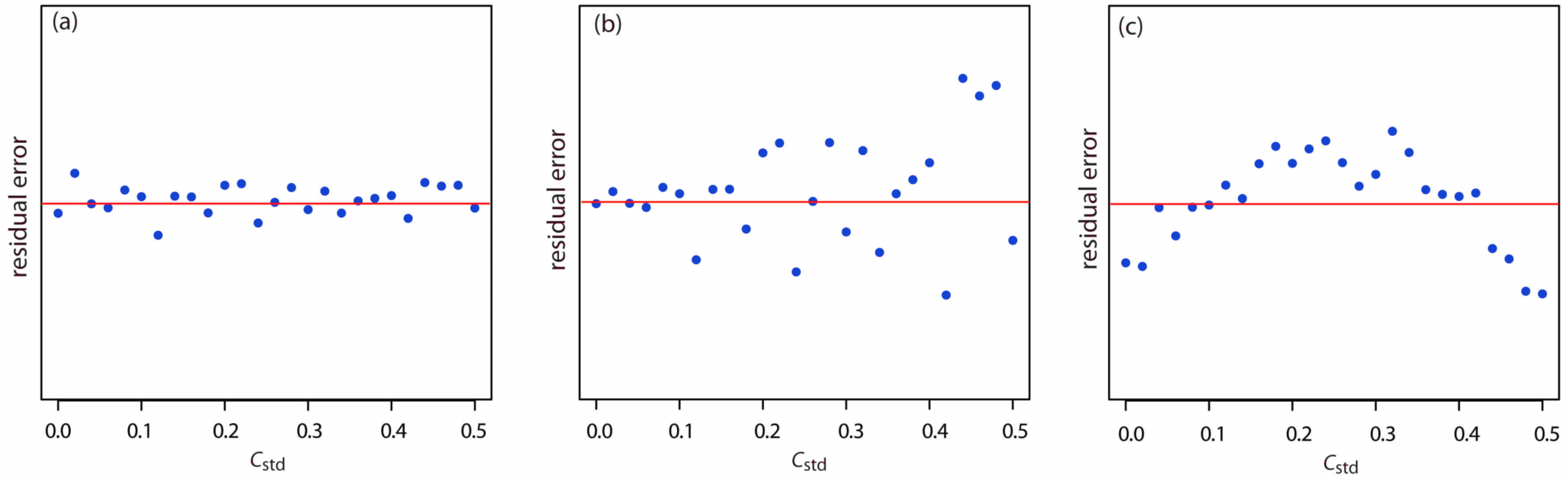
Modelos de Regressão

Regressão Linear Simples:

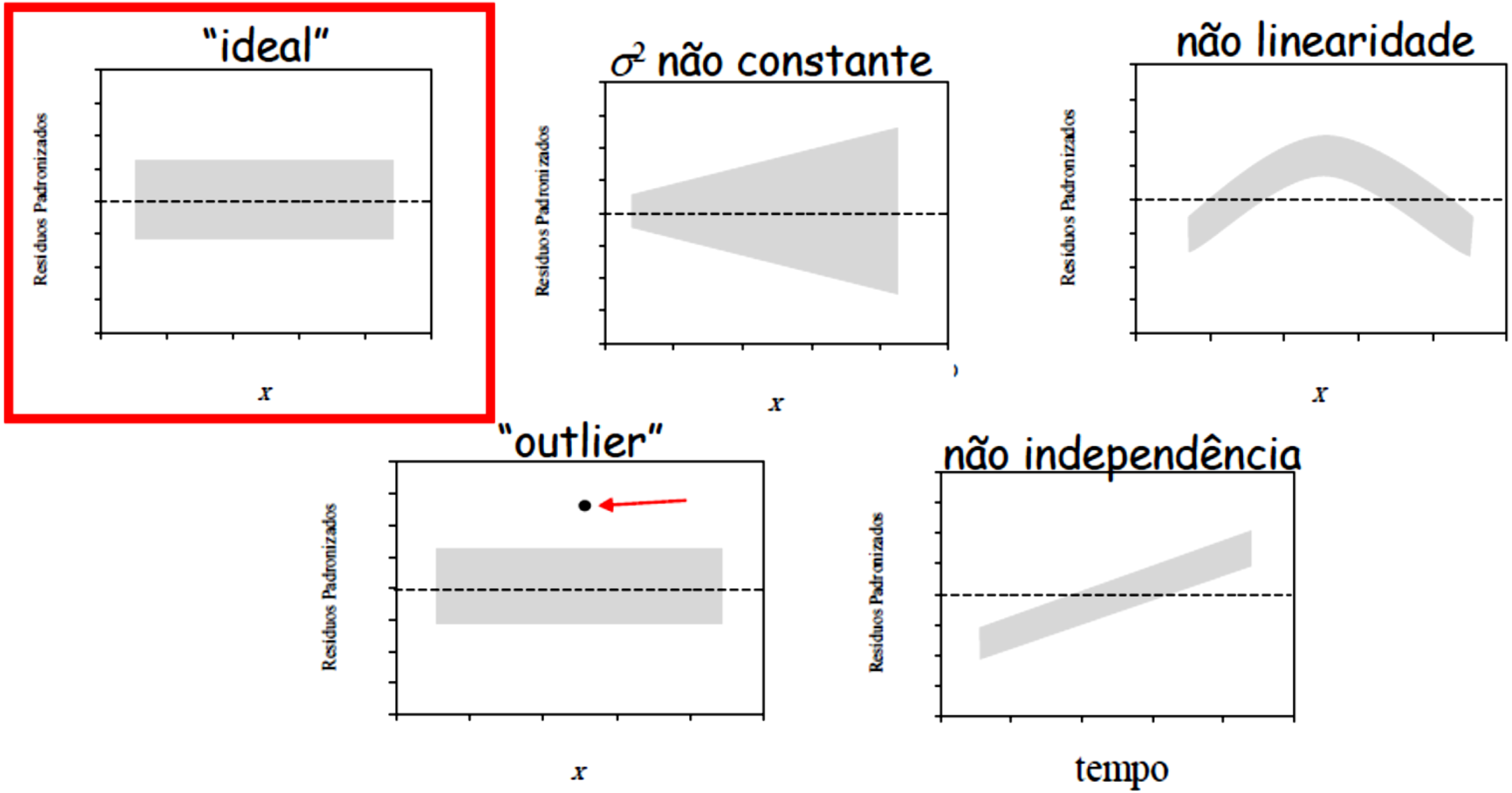


Modelos de Regressão

Regressão Linear Simples:



Modelos de Regressão



Modelos de Regressão

Selecionar o modelo:

Essas técnicas de regressão devem ser aplicadas **conforme as condições dos dados**.

Um dos melhores truques para descobrir qual técnica usar é **verificar a família de variáveis**, ou seja, discretas ou contínuas.

SIAD – Caso de Estudo

Prever real dimensionamento de estoque

FHEMIG
FUNDAÇÃO HOSPITALAR DO ESTADO DE MINAS GERAIS

2017 | 4504 registros
Dados fictícios

jupyter notebook
MACHINE LEARNING
REGRESSÃO

jupyter SIAD_RegressionTests (unsaved changes)   Logout

File Edit View Insert Cell Kernel Widgets Help   Not Trusted Python 3

Run Code

```
from sklearn.metrics import r2_score
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
from math import sqrt
from scipy import stats
#from datetime import datetime
from sklearn import preprocessing
from sklearn.model_selection import KFold
import matplotlib.pyplot as plt
%matplotlib inline
```

In [2]:

```
#Reading data and basic data check - Média de 2017
df = pd.read_csv('2271_ItemMaterial_Media_2017.csv',usecols=[1,2,3,4,5,6,7,8,9,10,11,12])#, index_col=0)
print(df.shape)
df.describe()

#original = pd.read_csv('2271_ItemMaterial_Media_2018.csv')
```

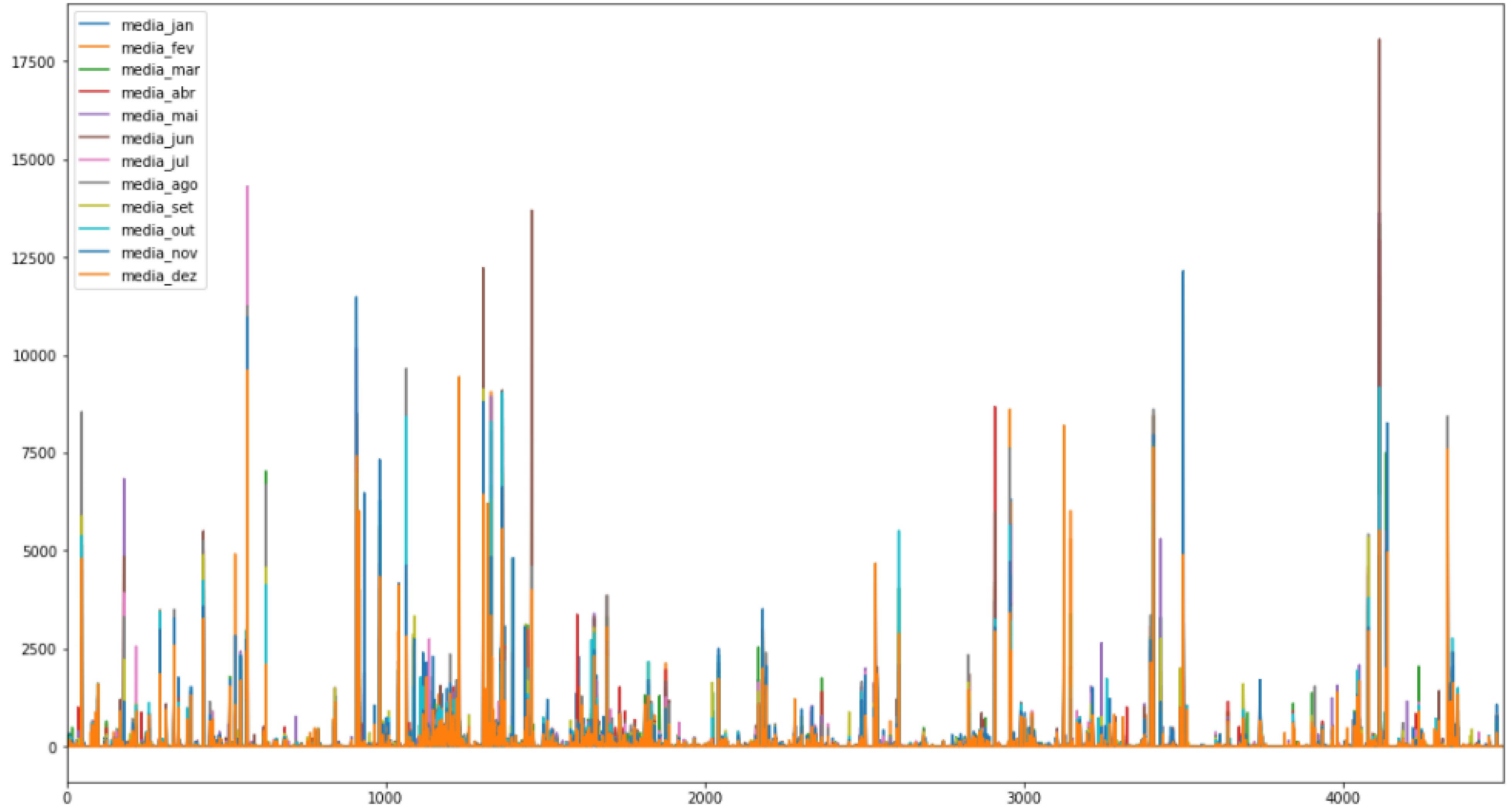
(4504, 12)

Out[2]:

	media_jan	media_fev	media_mar	media_abr	media_mai	media_jun	media_jul	media_ago	media_set	media_out	media_n
count	4504.000000	4504.000000	4504.000000	4504.000000	4504.000000	4504.000000	4504.000000	4504.000000	4504.000000	4504.000000	4504.0000
mean	92.612294	86.852202	103.401767	92.837977	100.260167	103.147980	87.402476	98.949338	89.583066	94.635337	99.2256
std	533.370349	486.854327	546.439197	515.136448	573.417497	614.486556	495.856920	543.423385	456.854068	500.711842	542.4130
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
50%	0.960000	0.670000	1.240000	0.670000	1.000000	1.000000	0.690000	0.670000	1.000000	0.670000	0.5000
75%	13.270000	13.032500	16.915000	12.207500	15.000000	15.000000	13.250000	12.762500	14.260000	12.500000	12.9625
max	12150.000000	12025.000000	13358.570000	12941.130000	13640.000000	18066.250000	14312.270000	11260.230000	9133.530000	9175.710000	11484.3100

In [3]:
```
#Plot - Verifying all data
df.plot(figsize=(18,10))
```

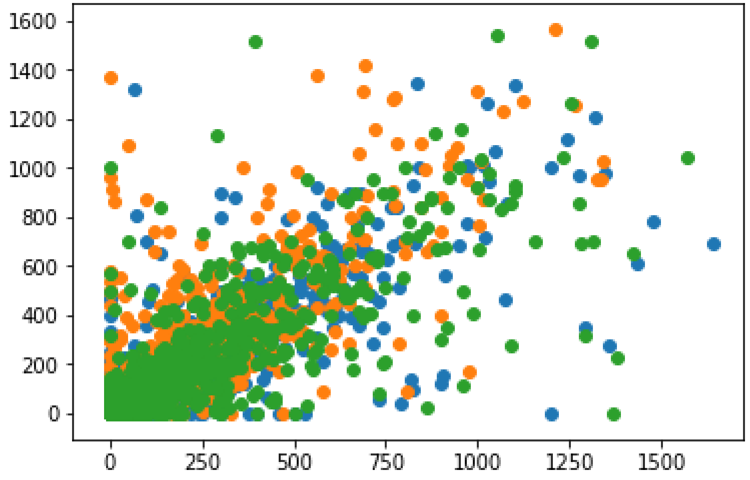
Out[3]: `<matplotlib.axes._subplots.AxesSubplot at 0xc159588>`



In [5]:
```python
#Validate linear relationship
plt.scatter(df['media_jan'], df['media_fev'])
plt.scatter(df['media_fev'], df['media_mar'])
plt.scatter(df['media_mar'], df['media_abr'])

#Possível verificar que não há uma relação muito forte entre as médias que possam corroborar para a predição de demanda.
```

Out[5]: <matplotlib.collections.PathCollection at 0x75dc149e8>



In [6]:
```python
#Creating arrays for features and response variable
target_column = ['media_dez']
predictors = list(set(list(df.columns))-set(target_column))
df[predictors] = df[predictors]/df[predictors].max()
df.describe()
```

In [7]:
```
#Creating training and test datasets
X = df[predictors].values
y = df[target_column].values

X_train, X_test, y_train, y_test = train_test_split(X, y, shuffle = True, test_size=0.30, random_state=40)
print(X_train.shape);
print(X_test.shape)
```

```
(3079, 11)
(1320, 11)
```

In [8]:
```
#Test cross_validation
from sklearn.model_selection import cross_val_score, cross_val_predict
lr = LinearRegression()
scores = cross_val_score(lr, X, y, cv = 5)
predictions = cross_val_predict(lr, X, y, cv = 5)
print (predictions)
```

```
[[2.7894928 ]
 [2.3995233 ]
 [3.05234729]
 ...
 [2.46050903]
 [2.68749046]
 [3.12208764]]
```

```
In [9]: #Build, Predict and Evaluate the Regression Model
        #Linear Regression
        lr = LinearRegression()
        lr.fit(X_train, y_train)

Out[9]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None,
                 normalize=False)
```

```
In [10]: pred_train_lr= lr.predict(X_train)
         print(np.sqrt(mean_squared_error(y_train,pred_train_lr)))
         print(r2_score(y_train, pred_train_lr))

         pred_test_lr= lr.predict(X_test)
         print(np.sqrt(mean_squared_error(y_test,pred_test_lr)))
         print(r2_score(y_test, pred_test_lr))

         67.09018025439444
         0.7179598247244705
         49.96318051966589
         0.8270858688286777
```

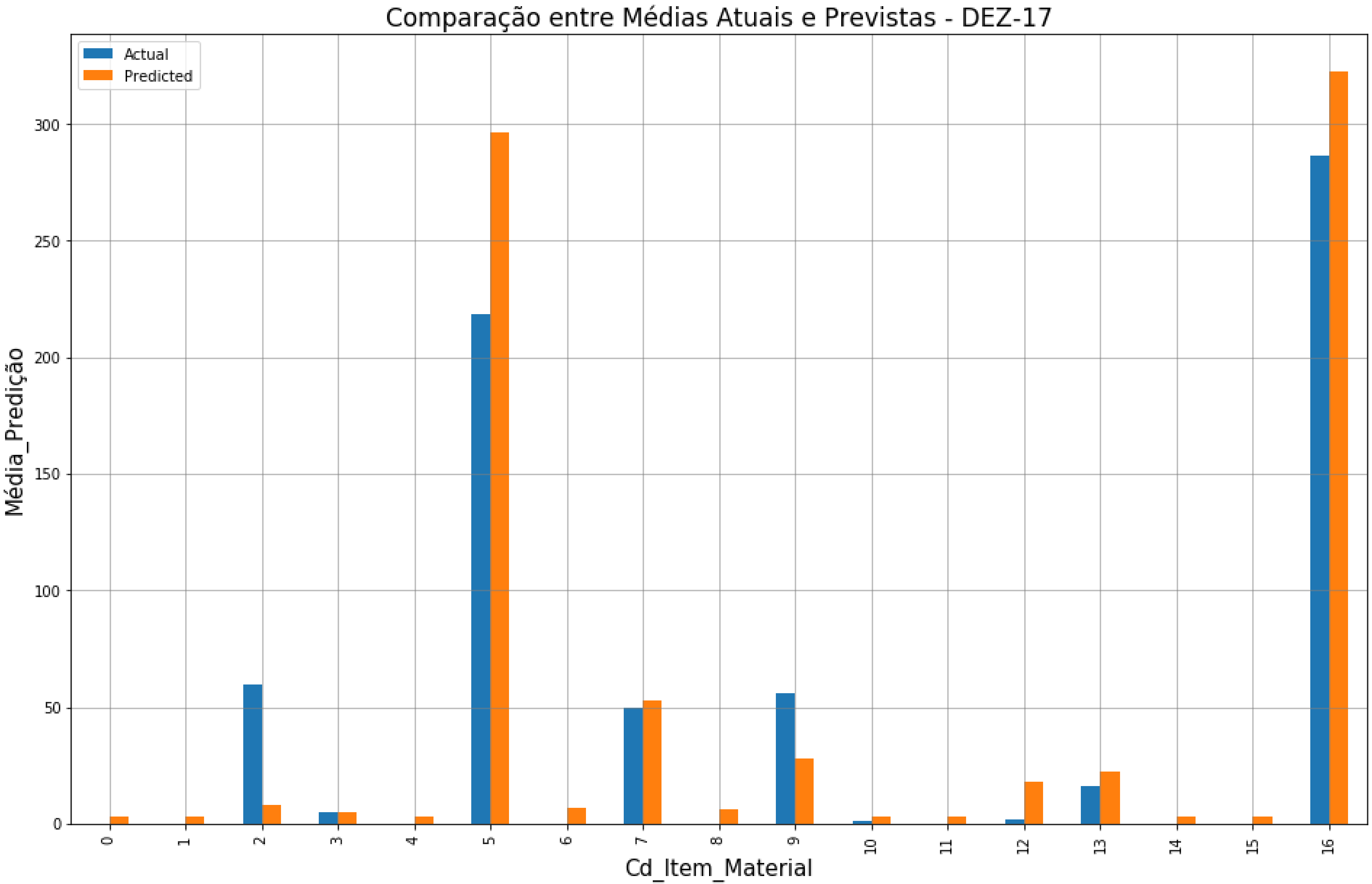
```
In [11]: #MSE:67.09
         #R-square:71.79%
```

	Item Material	Atual	Prevista
0	TERLIPRESSINA - FORMA FARMACEUTICA: INJETAVEL	0.00	2.95
1	VITAMINAS: PLURIVITAMIN	0.00	3.14
2	ETOXIPO-LIPROPILENOGLICOL + ASSOSSIACOES	60.00	7.85
3	MEDICAMENTO MANDADO JUDICIAL - PLURIMINERAL	5.00	4.60
4	FORMULA MANIPULADA - FORMA FARMACEUTICA: CAPSULA	0.00	2.93
5	FORMULA MANIPULADA - SOLUCAO OFTALMICA	218.46	296.35
6	CICLOBENZAPRINA	0.00	6.86
7	FERRO POLIMALTOSO - SOLUCAO INJETAVEL	50.00	52.58
8	CETAPHIL	0.00	6.30
9	TRICLOSAN - SOLUCAO TOPICA	56.00	27.85
10	DACARBAZINA - FORMA FARMACEUTICA	1.00	2.95
11	FEXOFENADINA	0.00	3.00
12	MEDICAMENTO MANDADO JUDICIAL – ENVID	2.00	17.69
13	PERINDOPRIL	16.00	22.45
14	CLORIDRATO DE LERCANIDIPINO	0.00	2.94
15	CLORIDRATO DE DULOXETINA	0.00	2.94
16	PROGESTERONA	286.25	322.48

In [13]:
```
df = pd.DataFrame({'Actual': y_test.flatten(), 'Predicted': pred_test_lr.flatten()})
df
```

Out[13]:

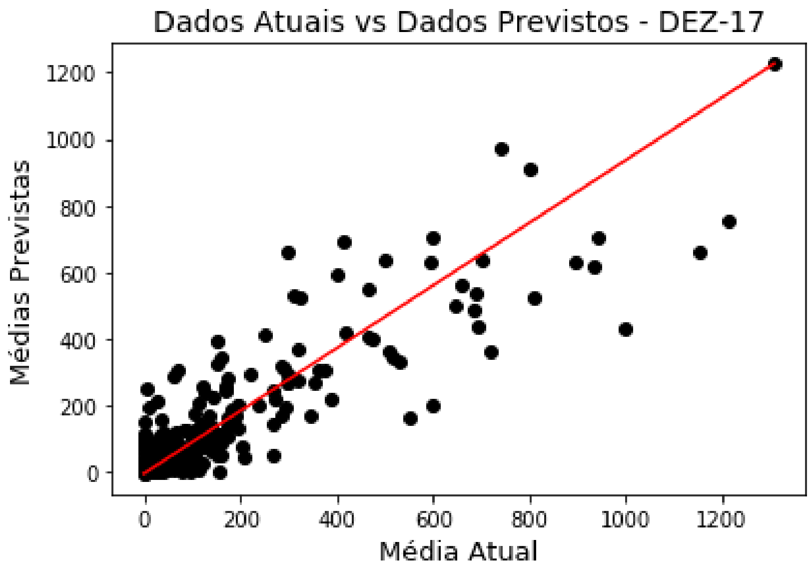
	Actual	Predicted
0	0.00	2.947065
1	0.00	3.141084
2	60.00	7.844464
3	5.00	4.591030
4	0.00	2.929617
5	218.46	296.349959
6	0.00	6.862481
7	50.00	52.583423
8	0.00	6.303412
9	56.00	27.847497
10	1.00	2.943754
11	0.00	2.994982
12	2.00	17.690702
13	16.00	22.451556
14	0.00	2.947065
15	0.00	2.947065
16	286.25	322.483907


Comparação entre Médias Atuais e Previstas - DEZ-17
Actual
Predicted
Média_Predição
Cd_Item_Material
0
50
100
150
200
250
300
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

In [17]:
```
plt.scatter(y_test,pred_test_lr, color ='black')
x0=min(y_test)
x1=max(y_test)
y0=min(pred_test_lr)
y1=max(pred_test_lr)
plt.plot([x0,x1],[y0,y1],color='red')

plt.xlabel('Média Atual', fontsize=13)
plt.ylabel('Médias Previstas', fontsize=13)
plt.title('Dados Atuais vs Dados Previstos - DEZ-17', fontsize=14 )

fig2 = plt.gcf()
plt.show()
```



Modelos de Regressão

Obrigado a todos!!!

Nathália Santiago – GPG/PRODEMGE | Neander Ferreira – GAC/PRODEMGE

Modelos de Regressão

Fontes:

https://analyticsindiamag.com/top-6-regression-algorithms-used-data-mining-applications-industry/

https://www.newgenapps.com/blog/business-applications-uses-regression-analysis-advantages

https://igti.com.br/blog/machine-learning-na-pratica-como-escolher-seus-algoritmos/

https://www.newgenapps.com/blog/business-applications-uses-regression-analysis-advantages

https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/

https://mineracaodedados.wordpress.com/2014/06/05/passos-para-a-criacao-de-um-projeto-de-modelagem-preditiva/