

Variáveis categóricas (“dummy variables”)

Thomas Peron

thomas.peron@usp.br

Sala 3-250 B

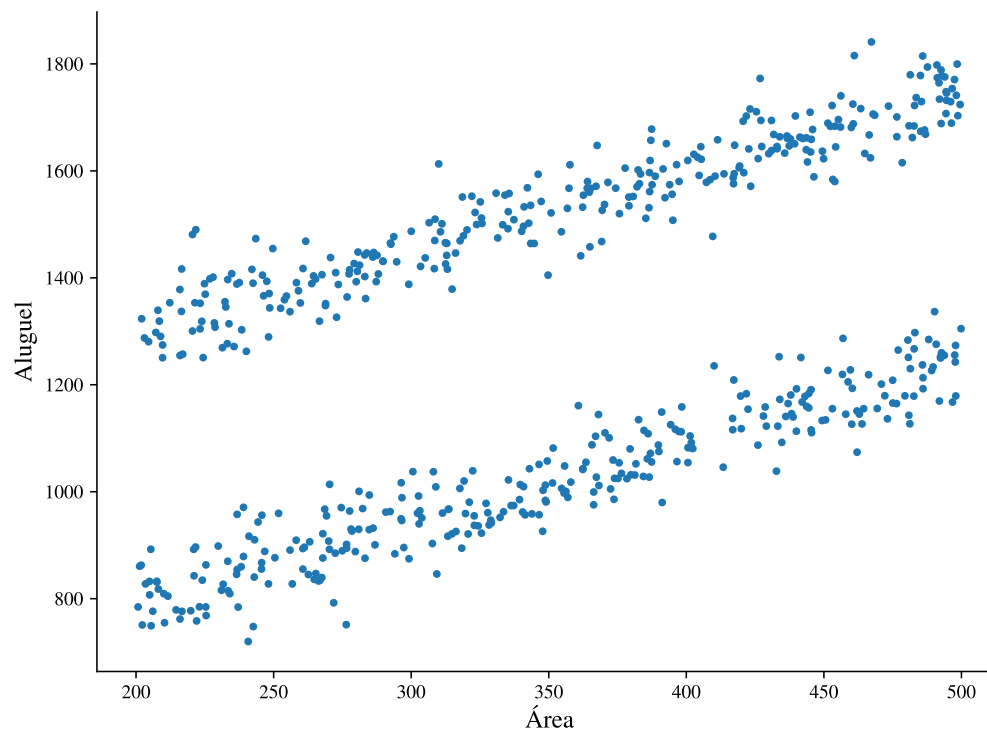
Problema: Explicar o aluguel, Y , em termos da metragem, X .

Problema: Explicar o aluguel, Y , em termos da metragem, X .

Aluguel (Y)	Metragem (X)
5618	70
7973	320
3841	80
1421	51
836	25
8955	376
2750	72
7253	213
16440	152
2955	35
2747	26
1181	46
2556	36
5268	55

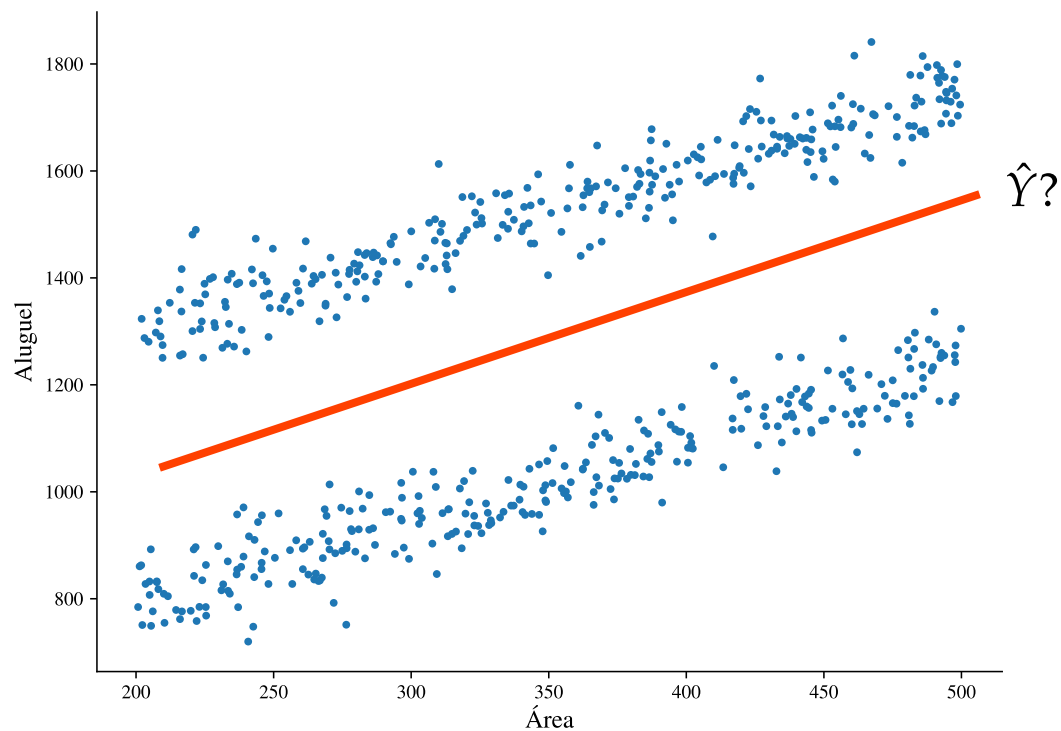
Problema: Explicar o aluguel, Y , em termos da metragem, X .

Aluguel (Y)	Metragem (X)
5618	70
7973	320
3841	80
1421	51
836	25
8955	376
2750	72
7253	213
16440	152
2955	35
2747	26
1181	46
2556	36
5268	55



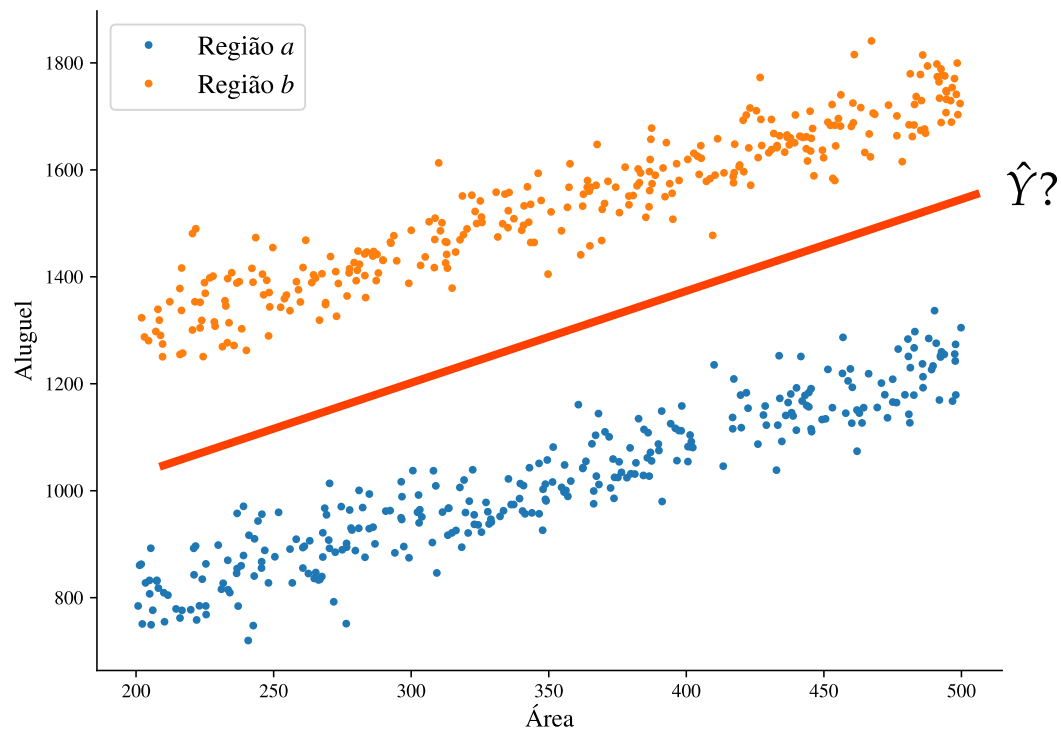
Problema: Explicar o aluguel, Y , em termos da metragem, X .

Aluguel (Y)	Metragem (X)
5618	70
7973	320
3841	80
1421	51
836	25
8955	376
2750	72
7253	213
16440	152
2955	35
2747	26
1181	46
2556	36
5268	55



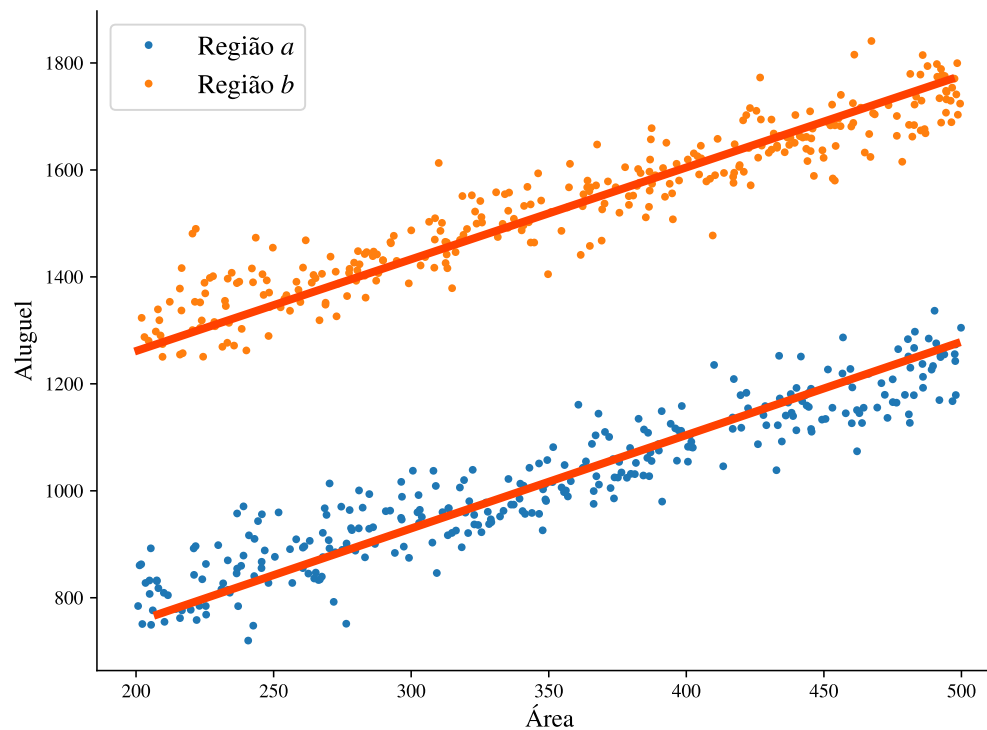
Problema: Explicar o aluguel, Y , em termos da metragem, X .

Aluguel (Y)	Metragem (X)	Região
5618	70	a
7973	320	b
3841	80	a
1421	51	a
836	25	a
8955	376	b
2750	72	a
7253	213	a
16440	152	a
2955	35	a
2747	26	b
1181	46	a
2556	36	b
5268	55	b



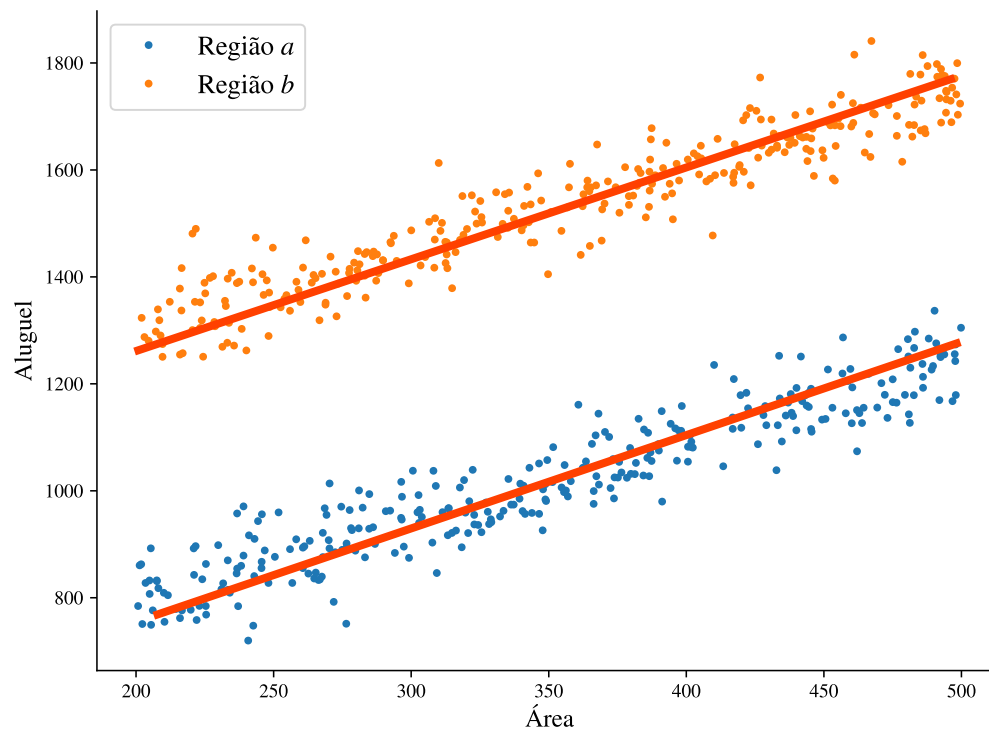
Problema: Explicar o aluguel, Y , em termos da metragem, X .

Aluguel (Y)	Metragem (X)	Região
5618	70	a
7973	320	b
3841	80	a
1421	51	a
836	25	a
8955	376	b
2750	72	a
7253	213	a
16440	152	a
2955	35	a
2747	26	b
1181	46	a
2556	36	b
5268	55	b



Problema: Explicar o aluguel, Y , em termos da metragem, X .

Aluguel (Y)	Metragem (X)	Região
5618	70	a
7973	320	b
3841	80	a
1421	51	a
836	25	a
8955	376	b
2750	72	a
7253	213	a
16440	152	a
2955	35	a
2747	26	b
1181	46	a
2556	36	b
5268	55	b



Como escrever o modelo?

Modelos com variáveis indicadoras

Podemos escrever:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon.$$

$$X_1 = \text{Área}, \quad X_2 = \begin{cases} 1 & \text{se o imóvel for da Região } a; \\ 0 & \text{c.c.} \end{cases} \quad X_3 = \begin{cases} 1 & \text{se o imóvel for da Região } b; \\ 0 & \text{c.c.} \end{cases}$$

Matriz $\mathbf{X}$ para um exemplo de $n = 4$ observações

$$\mathbf{X} = \begin{array}{ccccc} & X_0 & X_1 & X_2 & X_3 \\ \begin{bmatrix} 1 & X_{11} & 1 & 0 \\ 1 & X_{21} & 1 & 0 \\ 1 & X_{31} & 0 & 1 \\ 1 & X_{41} & 0 & 1 \end{bmatrix} & \rightarrow & \mathbf{X}^T \mathbf{X} = & \begin{bmatrix} 4 & \sum_i^4 X_{i1} & 2 & 2 \\ \sum_i^4 X_{i1} & \sum_i^4 X_{i1}^2 & \sum_i^2 X_{i1} & \sum_i^4 X_{i1} \\ 2 & \sum_i^2 X_{i1} & 2 & 0 \\ 2 & \sum_i^4 X_{i1} & 0 & 2 \end{bmatrix} \end{array}$$

A variável categórica com c classes será representada por $c - 1$ variáveis indicadoras.

O modelo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

Interpretação dos parâmetros

O modelo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

$$X_1 = \text{Área}, X_2 = \begin{cases} 1 & \text{se o imóvel for da Região } b; \\ 0 & \text{c.c.} \end{cases}$$

Resposta média do modelo:

$$\mathbb{E}[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

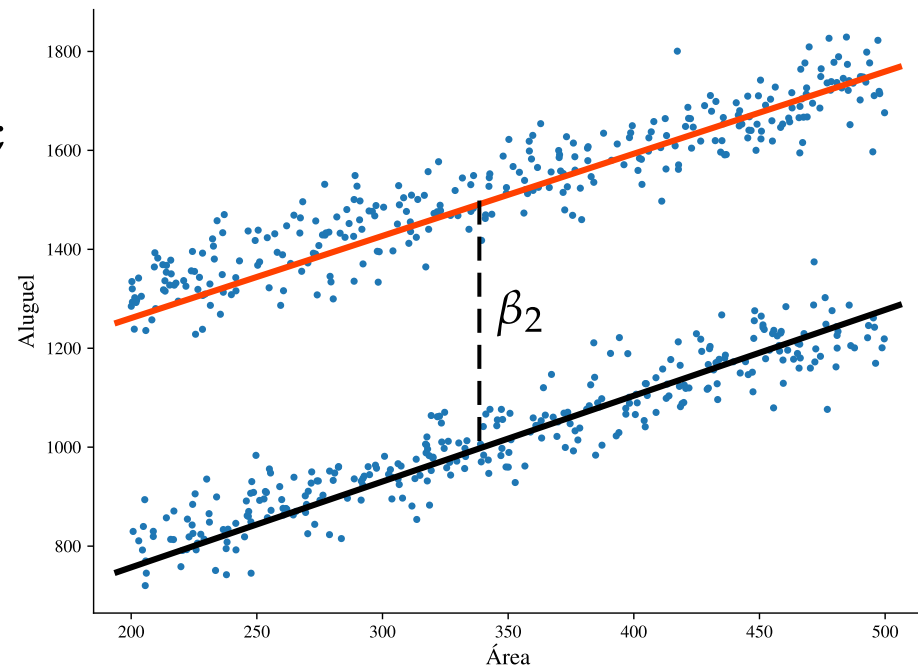
Se o imóvel for da região a ($X_2 = 0$):

$$\mathbb{E}[Y|X_1, X_2 = 0] = \beta_0 + \beta_1 X_1$$

Se o imóvel for da região b ($X_2 = 1$):

$$\mathbb{E}[Y|X_1, X_2 = 1] = (\beta_0 + \beta_2) + \beta_1 X_1$$

$\beta_2 = \text{diferença entre as respostas médias}$



Interpretação dos parâmetros

Para $p - 1$ covariáveis:

$$Y = \beta_0 + \beta_B X_B + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon.$$

onde $X_B \in \{0, 1\}$.

β_B é chamado de **contraste** entre as duas classes.

Mais de uma variável *indicadora*

Quatro regiões: $\{a, b, c, d\}$.

Região	Aluguel (X_1)	X_2	X_3	X_4
a	X_{i1}	0	0	0
b	X_{i1}	1	0	0
c	X_{i1}	0	1	0
d	X_{i1}	0	0	1

Modelo amostral:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i$$

Valores esperados:

$\mathbb{E}[Y] = \beta_0 + \beta_1 X_1$ (Região a)

$\mathbb{E}[Y] = (\beta_0 + \beta_3) + \beta_1 X_1$ (Região c)

$\mathbb{E}[Y] = (\beta_0 + \beta_2) + \beta_1 X_1$ (Região b)

$\mathbb{E}[Y] = (\beta_0 + \beta_4) + \beta_1 X_1$ (Região d)

Modelo com interações

Adicionando o termo de interação $X_1 X_2$:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{i1} X_{i2} + \varepsilon_i.$$

Resposta média:

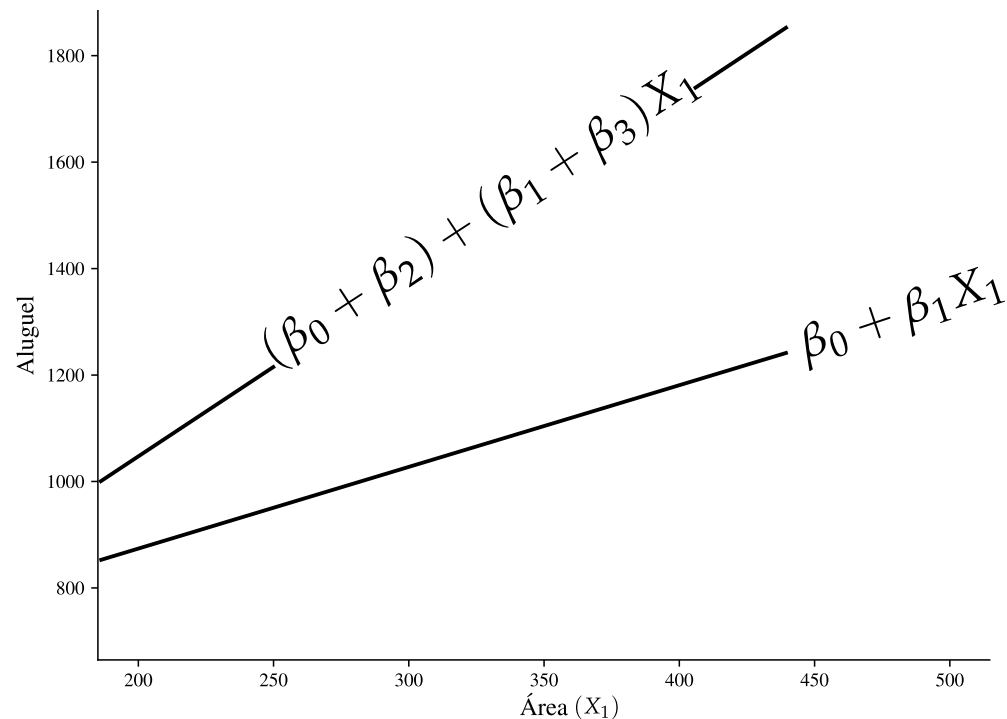
$$\mathbb{E}[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

Se o imóvel for da região a ($X_2, X_1 X_2 = 0$) :

$$\mathbb{E}[Y|X_1, X_2] = \beta_0 + \beta_1 X_1$$

Se o imóvel for da região b ($X_2, X_1 X_2 \neq 0$) :

$$\mathbb{E}[Y|X_1, X_2] = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1$$



Termos de interação: mudanças no intercepto e nos coeficientes.

Os dados

Aluguel (Y)	Metragem (X)	Região	Variável indicadora X_2	$X_1 X_2$
5618	70	<i>a</i>	0	0
7973	320	<i>a</i>	0	0
3841	80	<i>a</i>	0	0
1421	51	<i>a</i>	0	0
836	25	<i>a</i>	0	0
8955	376	<i>a</i>	0	0
2750	72	<i>b</i>	1	72
7253	213	<i>b</i>	1	213
16440	152	<i>b</i>	1	152
2955	35	<i>b</i>	1	35
2747	26	<i>b</i>	1	26
1181	46	<i>b</i>	1	46
2556	36	<i>b</i>	1	36
5268	55	<i>b</i>	1	55

Mais de uma variável *categórica*

Aluguel (Y)	Metragem (X)	Região	Tipo do imóvel
5618	70	<i>a</i>	casa
7973	320	<i>a</i>	apto
3841	80	<i>a</i>	apto
1421	51	<i>a</i>	apto
836	25	<i>a</i>	casa
8955	376	<i>a</i>	casa
2750	72	<i>b</i>	casa
7253	213	<i>b</i>	apto
16440	152	<i>b</i>	apto
2955	35	<i>b</i>	apto
2747	26	<i>b</i>	apto
1181	46	<i>b</i>	casa
2556	36	<i>b</i>	casa
5268	55	<i>b</i>	apto

Duas classes em cada var. categórica.

Possível modelo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon.$$

X_1 : aluguel

$$X_2 = \begin{cases} 1 & \text{se o imóvel for da Região } b; \\ 0 & \text{c.c.} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{se o imóvel for um apto.} \\ 0 & \text{c.c.} \end{cases}$$

$\beta_2, \beta_3 \rightarrow$ efeito aditivo em β_0 .

Mais de uma variável *categórica*

Mais complicado:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 +$$
$$\underbrace{\beta_4 X_1 X_2}_{\text{interação } m^2 \text{ e bairro}} + \underbrace{\beta_5 X_1 X_3}_{\text{interação } m^2 \text{ e tipo}} +$$
$$\underbrace{\beta_6 X_2 X_3}_{\text{interação bairro e tipo}}.$$

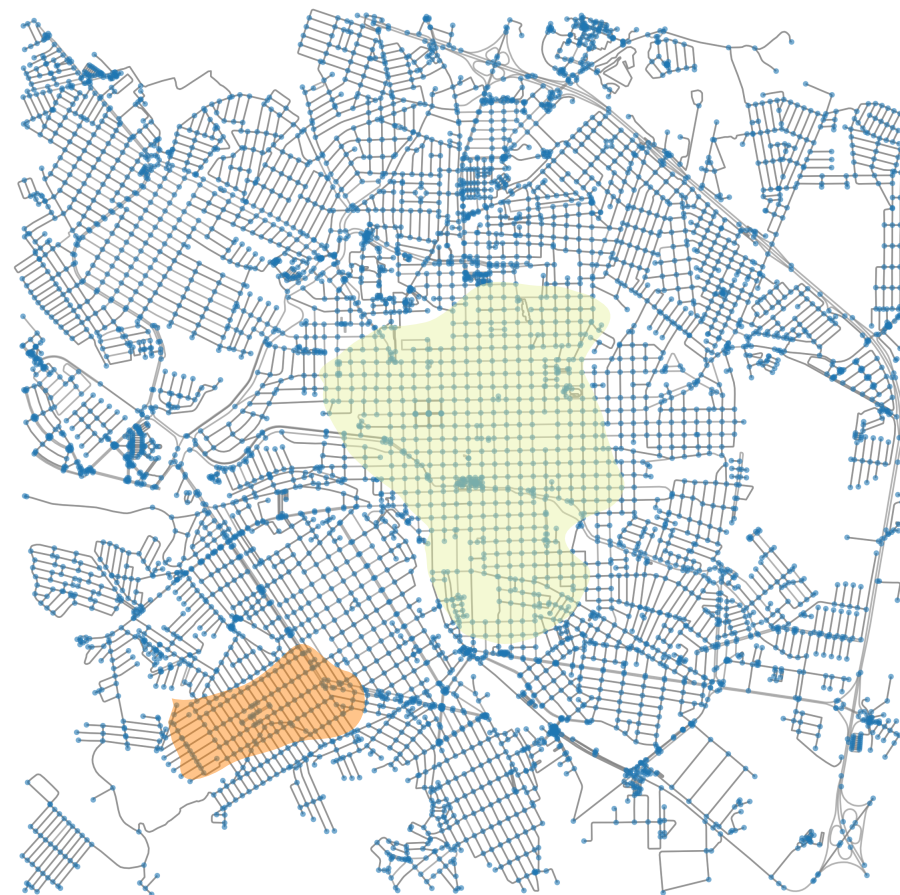
Valores esperados:

Casas no centro: $\mathbb{E}[Y] = \beta_0 + \beta_1 X_1$

Casas no Parque Faber: $\mathbb{E}[Y] = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) X_1$

Aptos no Parque Faber:

$$\mathbb{E}[Y] = (\beta_0 + \beta_2 + \beta_3 + \beta_6) + (\beta_1 + \beta_4 + \beta_5) X_1$$



Feito com osmnx.

Região a: Centro

Região b: Parque Faber

Comparação de modelos

No modelo

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon,$$

X_2 é realmente útil para descrever a variância de Y ?

Teste de hipótese:

$$H_0 : \beta_2 = \beta_3 = 0 \quad (\mathbb{E}[Y] = \beta_0 + \beta_1 X_1)$$

$$H_1 : \beta_2 \text{ e/ou } \beta_3 \neq 0.$$

$$f^* = \frac{\text{SSR}(X_2, X_1 X_2 | X_1)}{2} \bigg/ \frac{\text{SSE}(X_1, X_2, X_1 X_2)}{n - 4}$$

Testando apenas a mudança de inclinação:

$$H_0 : \beta_3 = 0 \quad (\mathbb{E}[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2)$$

$$H_1 : \beta_3 \neq 0.$$

$$f^* = \frac{\text{SSR}(X_1 X_2 | X_1, X_2)}{1} \bigg/ \frac{\text{SSE}(X_1, X_2, X_1 X_2)}{n - 4}$$

Mais sobre regressão e variáveis categóricas

