Regressão linear múltipla

Universidade Estadual de Santa Cruz

Ivan Bezerra Allaman

Introdução

- A regressão múltipla é uma generalização da regressão simples, visto que, há mais de uma variável explicativa no modelo. Logo, o que determina se uma regressão é do tipo simples ou múltipla é o número de variáveis explicativas no modelo.
- É comum encontrar alguns materiais se referir a regressão do tipo polinomial como múltipla simplesmente por haver no modelo o termo de segundo grau, terceiro grau, etc. Se há uma regressão com termos acima do de primeiro grau, mas a variável explicativa continua sendo apenas uma, então temos uma regressão simples.
- Existem também alguns materiais referindo a regressão múltipla como regressão multivariada. O termo "multivariado", é utilizado em análises estatísticas quando ocorre mais de uma **variável resposta**. Aliás, é o número de variáveis respostas que determina se a análise é do tipo **univariado** ou **multivariado**.

Objetivo

- Estudar a relação funcional entre variáveis, sendo uma resposta e duas ou mais explicativas.
- Estabelecer um modelo para entender a relação funcional entre as variáveis.
- Fazer predições como o modelo ajustado principalmente para valores que não foram observados na amostra.

O modelo

- O modelo matemático que estabele a relação funcional entre as variáveis é definido como:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

em que:
y= é a variável dependente
$\beta_0, \beta_1, \cdots, \beta_k$ = são parâmetros a serem estimados
$x_1, x_2, \cdots, x_k$ = são as variáveis independentes
ε= é o erro aleatório referente a variabilidade em y quem não pode ser explicada pelas variáveis $x's$.

Estimação dos parâmetros (Mínimos quadrados)

- Utiliza-se o método dos mínimos quadrados para estimação dos parâmetros.
- A idéia é exatamente a mesma que foi apresentada para regressão simples. No entanto, por se tratar de "k" variáveis explicativas, é inviável termos uma equação para estimarmos os parâmetros. Logo, lançamos mão da álgebra matricial para tal feito. Portanto, podemos reescrever o modelo de regressão já apresentado na forma matricial como:

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

em que:
$\mathbf{Y} =$ é um vetor coluna das observações com dimensões $n \times 1$.
$\mathbf{X} =$ é uma matrix $n \times (k+1)$ das variáveis explicativas.
$\beta =$ é um vetor coluna dos parâmetros que se quer estimar com dimensões $(k+1) \times 1$.
$\varepsilon =$é um vetor coluna dos resíduos com dimensões $n \times 1$.

- Logo, a equação de quadrados mínimos para estimar os parâmetros de um modelo de regressão é:

$$\hat{\beta} = (\mathbf{X^T X})^{-1} \mathbf{X^T Y}$$

Aplicação

1. Gerente da Butler Trucking Company querem avaliar se o tempo de viagem para entregar uma carga está em função das milhas percorridas e do número de entregas deliveries. Um amostra aleatória simples de dez tarefas de entrega forneceram os seguintes dados:

Tempo de viagem (horas)	Milhas percorridas	Número de entregas deliveries
9.3	100	4
4.8	50	3
8.9	100	4
6.5	100	2
4.2	50	2
6.2	80	2
7.4	75	3
6.0	65	4
7.6	90	3
6.1	90	2

Tem-se as seguintes matrizes:

$$Y = \begin{bmatrix} 9.3 \\ 4.8 \\ 8.9 \\ 6.5 \\ 4.2 \\ 6.2 \\ 7.4 \\ 6 \\ 7.6 \\ 6.1 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 100 & 4 \\ 1 & 50 & 3 \\ 1 & 100 & 4 \\ 1 & 100 & 2 \\ 1 & 50 & 2 \\ 1 & 80 & 2 \\ 1 & 75 & 3 \\ 1 & 65 & 4 \\ 1 & 90 & 3 \\ 1 & 90 & 2 \end{bmatrix}$$

Os valores 1 na matriz **X** é uma constante utilizada para estimar o β_0.

Logo, os valores estimados dos parâmetros da regressão são:

$$X^T X = \begin{bmatrix} 10 & 800 & 29 \\ 800 & 67450 & 2345 \\ 29 & 2345 & 91 \end{bmatrix}$$

$$(X^T X)^{-1} = \begin{bmatrix} 2.756 & -0.0207 & -0.3453 \\ -0.0207 & 0.000298 & -0.0010785 \\ -0.3453 & -0.00108 & 0.148835 \end{bmatrix}$$

$$X^T Y = \begin{bmatrix} 67 \\ 5594 \\ 202,2 \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} -0,8687 \\ 0,0611 \\ 0.9234 \end{bmatrix}$$

Logo, o modelo de regressão ajustado é:

$$\hat{y} = -0,8687 + 0,06114 \cdot x_1 + 0,9234 \cdot x_2$$

ou de uma maneira mais "coloquial",

$$\hat{y} = -0,8687 + 0,06114 \cdot milhas + 0,9234 \cdot entregas$$

Interpretação do modelo ajustado

- No caso de uma regressão múltipla, os coeficientes que acompanham as variáveis independentes são interpretados de modo diferente comparado à regressão linear simples.
- No exemplo anterior a interpretação a equação ajustada é a seguinte:
  - O aumento em uma unidade nas milhas percorridas, aumenta em média o tempo de viagem em 0,06114 horas quando todas as outras variáveis independentes permanecem constantes.
  - Entretanto, o aumento em uma unidade no número de entregas deliveries aumenta em média o tempo de viagem em 0,9234 horas quando todas as demais variáveis independentes permanecem constantes.

Coeficiente de determinação múltiplo

- Tem a mesma interpretação e cálculo como no caso da regressão simples.
- No entanto, utilizamos o R^2 maiúsculo para representar tal eficiência, uma vez que, não tem relação direta com o coeficiente de correlação.
- Logo, tem-se:

$$R^2 = \frac{SSR}{SST}$$

em que:
$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ a soma de quadrados de regressão
$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$ a soma de quadrados total

Coeficiente de determinação múltiplo ajustado

- Quanto maior o número de variáveis independentes no modelo, maior será o R^2. Logo, uma medida que corrija este inconveniente se faz necessário. Portanto, o coeficiente de determinação múltiplo **ajustado** corrige este problema, pois leva em consideração o número de variáveis no modelo de regressão.
- O cálulo é o seguinte:

$$R^2_{ajust} = 1 - \left(\frac{n-1}{n-p-1} \cdot (1 - R^2) \right)$$

Aplicação

2. Considere o exemplo 1. Calcule o coeficiente de determinação múltiplo ajustado.

   Calculando primeiro o coeficiente de determinação múltiplo sem ajuste.

$$R^2 = \frac{SSR}{SST} = \frac{21,60}{23,9} = 0,9038$$

   Calculando o coeficiente de determinação múltiplo ajustado.

$$\begin{aligned} R^2_{ajust} &= 1 - \left(\frac{10-1}{10-2-1} \cdot (1 - 0,9038) \right) \\ &= 1 - 0,1237 \\ &= 0,8763 \end{aligned}$$

Inferência sobre os β_i

Pressupostos do modelo

- São os mesmos já abordados para regressão simples.

Teste de hipótese

- Também tem a mesma abordagem para regressão simples com uma pequena diferença. No caso formulação das hipóteses temos:

$$\begin{cases} H_0 : \beta_i = 0 \\ H_a : \beta_i \neq 0 \end{cases}$$

- Na estimação do desvio padrão dos b_i (que são os estimadores dos β_i) tem-se:

$$s_{b_i} = \frac{s_{erro}}{\sqrt{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}}$$

- Lembre-se de que x_i é a variável independente relacionada ao b_i.
- Logo, tem-se a seguinte estatística de teste:

$$t = \frac{b_i}{s_{bi}}$$

em que t tem distribuição t de student com $n - k - 1$ graus de liberdade, sendo k o número de variáveis independentes.

Aplicação

3. Considere ainda o exemplo 1. Avalie se as variáveis independentes são significativas.

   As hipóteses são:

$$\begin{cases} H_0 : \beta_{milhas} = 0 \\ H_a : \beta_{milhas} \neq 0 \end{cases}$$

   e

$$\begin{cases} H_0 : \beta_{entregas} = 0 \\ H_a : \beta_{entregas} \neq 0 \end{cases}$$

Testando primero o coeficiente para a variável "milhas".

$$s_{erro} = \sqrt{\frac{2,299}{10-2-1}} = 0,5731$$

$$s_{b_{milhas}} = \frac{0,5731}{58,7367} = 0,0099$$

$$t = \frac{0,0611}{0,0099} = 6,172$$

$$p_{valor} = (1 - pt(6.172, 7)) \cdot 2 = 0,0004576$$

Testando o coeficiente para a variável "entregas".

$$s_{b_{entregas}} = \frac{0,5731}{2,6267} = 0,2182$$
$$t = \frac{0,9234}{0,2182} = 4,232$$
$$p_{valor} = (1 - pt(4.232, 7)) \cdot 2 = 0,0039$$

Método de seleção de variáveis

- Quando existem muitas variáveis independentes, a intepretação do modelo ajustado se torna muitas vezes difícil e de pouca aplicabilidade prática.
- Logo, um método que nos retorne um modelo mais parcimonioso possível se faz necessário.
- Dentre os métodos de seleção iremos citar apenas o método stepwise com o critério de Akaike (AIC) para a escolha do modelo mais parcimonioso.
- O método stepwise será abordado computacionalmente, uma vez que, manualmente é trabalhoso e a medida que aumenta o número de variáveis se torna impossível.
- Portanto, iremos utilizar a função **stepAIC** do pacote **MASS**.
- O exemplo abaixo será executado diretamente com o programa R.

Aplicação

4. O artigo "Response surface methodology for protein extraction optimization of red pepper seed" forneceu os seguintes dados a respeito da variável reposta ppro = produção proteica (%) e das variáveis independentes temp = temperatura ($^{\circ}C$), ph = pH, text = tempo de extração (min) e solv = razão solvente/refeição.

temp	ph	text	solv	ppro	temp	ph	text	solv	ppro
35	7.5	30	15	9.74	45	8.5	50	25	12.25
45	7.5	30	15	9.91	30	8.0	40	20	11.84
35	8.5	30	15	11.80	50	8.0	40	20	11.84
45	8.5	30	15	11.69	40	7.0	40	20	8.32
35	7.5	50	15	10.68	40	9.0	40	20	12.22
45	7.5	50	15	10.71	40	8.0	20	20	11.28
35	8.5	50	15	10.91	40	8.0	60	20	12.72
45	8.5	50	15	11.77	40	8.0	40	10	9.63
35	7.5	30	25	9.84	40	8.0	40	30	11.17
45	7.5	30	25	9.82	40	8.0	40	20	12.08
35	8.5	30	25	11.78	40	8.0	40	20	11.95
45	8.5	30	25	12.31	40	8.0	40	20	11.77
35	7.5	50	25	11.06	40	8.0	40	20	11.71
45	7.5	50	25	11.24	40	8.0	40	20	12.02
35	8.5	50	25	12.31					

a. Ajuste um modelo de segunda ordem completo e utilize o método stepwise para obter o melhor conjuto de variáveis preditoras.

```
```
##
Call:
lm(formula = ppro ~ temp * ph + temp * text + temp * solv + ph *
text + ph * solv + text * solv + I(temp^2) + I(ph^2) + I(text^2) +
I(solv^2), data = dados)
##
Residuals:
Min   1Q   Median   3Q   Max
-0.34833 -0.13417 0.05083 0.15250 0.32417
##
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.195e+02 1.853e+01 -6.449 1.53e-05 ***
temp   -1.047e-01 2.839e-01 -0.369 0.717803
ph   2.868e+01 3.625e+00   7.910 1.56e-06 ***
text   4.074e-01 1.303e-01   3.127 0.007426 **
solv   2.711e-01 2.606e-01   1.040 0.315830
I(temp^2)   -7.517e-04 2.110e-03 -0.356 0.726982
I(ph^2)   -1.645e+00 2.110e-01 -7.797 1.85e-06 ***
I(text^2)   2.121e-04 5.275e-04   0.402 0.693726
I(solv^2)   -1.515e-02 2.110e-03 -7.181 4.70e-06 ***
temp:ph   2.150e-02 2.687e-02   0.800 0.436998
temp:text   5.500e-04 1.344e-03   0.409 0.688460
temp:solv   -8.000e-04 2.687e-03 -0.298 0.770285
ph:text   -5.900e-02 1.344e-02 -4.391 0.000615 ***
ph:solv   3.900e-02 2.687e-02   1.451 0.168703
text:solv   2.725e-03 1.344e-03   2.028 0.062003 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
Residual standard error: 0.2687 on 14 degrees of freedom
Multiple R-squared: 0.9668, Adjusted R-squared: 0.9336
F-statistic: 29.11 on 14 and 14 DF, p-value: 6.404e-08
```
```

a. … segue o resultado do método stepwise.

```
## [1] "Step:  AIC=-74.86"
## [2] "ppro ~ temp + ph + text + solv + I(ph^2) + I(solv^2) + ph:text + "
## [3] "    ph:solv + text:solv"
## [4] ""
## [5] "            Df Sum of Sq    RSS     AIC"
## [6] "                         1.1012 -74.856"
## [7] "- temp       1    0.1040 1.2052 -74.238"
## [8] "+ temp:ph    1    0.0462 1.0550 -74.099"
## [9] "+ I(text^2)  1    0.0165 1.0847 -73.294"
## [10] "+ I(temp^2)  1    0.0140 1.0872 -73.227"
## [11] "+ temp:text  1    0.0121 1.0891 -73.176"
## [12] "- ph:solv    1    0.1521 1.2533 -73.104"
## [13] "+ temp:solv  1    0.0064 1.0948 -73.025"
## [14] "- text:solv  1    0.2970 1.3982 -69.930"
## [15] "- ph:text    1    1.3924 2.4936 -53.153"
## [16] "- I(solv^2)  1    3.9693 5.0705 -32.572"
## [17] "- I(ph^2)    1    4.6789 5.7801 -28.773"
## [18] ""
## [19] "Call:"
## [20] "lm(formula = ppro ~ temp + ph + text + solv + I(ph^2) + I(solv^2) + "
## [21] "    ph:text + ph:solv + text:solv, data = dados)"
## [22] ""
## [23] "Coefficients:"
## [24] "(Intercept)         temp           ph         text         solv  "
## [25] " -1.258e+02    1.317e-02    2.956e+01    4.463e-01    2.397e-01  "
## [26] "    I(ph^2)    I(solv^2)      ph:text      ph:solv    text:solv  "
## [27] " -1.647e+00   -1.517e-02   -5.900e-02    3.900e-02    2.725e-03  "
## [28] ""
```

a. Portanto o melhor conjunto de variáveis foi:

$$temp + ph + text + solv + pH^2 + solv^2 + pH : text + pH : solv + text : solv$$

b. Ajuste um novo modelo a partir do conjunto de variáveis selecionadas na alternativa anterior.

```
##
## Call:
## lm(formula = ppro ~ temp + ph + text + solv + I(ph^2) + I(solv^2) +
##     ph:text + ph:solv + text:solv, data = dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44903 -0.13864  0.04136  0.15513  0.32307
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.258e+02  1.307e+01  -9.628 9.66e-09 ***
## temp         1.317e-02  9.828e-03   1.340  0.19616
## ph           2.956e+01  3.012e+00   9.815 7.11e-09 ***
## text         4.463e-01  9.938e-02   4.491  0.00025 ***
## solv         2.397e-01  2.119e-01   1.131  0.27199
## I(ph^2)     -1.647e+00  1.833e-01  -8.985 2.87e-08 ***
## I(solv^2)   -1.517e-02  1.833e-03  -8.276 1.01e-07 ***
## ph:text     -5.900e-02  1.204e-02  -4.901 9.91e-05 ***
## ph:solv      3.900e-02  2.407e-02   1.620  0.12172
## text:solv    2.725e-03  1.204e-03   2.264  0.03548 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2407 on 19 degrees of freedom
## Multiple R-squared:  0.9638, Adjusted R-squared:  0.9467
## F-statistic: 56.24 on 9 and 19 DF,  p-value: 8.489e-12
```

c. Escreva a esquação ajustada. O modelo explica bem a variabilidade dos dados?

$$\hat{ppro} = -125.85 + 0.013temp + 29.56pH + 0.446text + 0.2397solv -$$
$$- 1.647pH^2 - 0.0152solv^2 - 0.059pH : text + 0.039pH : solv +$$
$$+ 0.0027text : solv$$

Sim, pois o coeficiente de determinação ajustado foi muito satisfatório, ou seja, $R^2_{ajust} = 0.9467$

Multicolinearidade

- A multicolinearidade é um problema que ocorre quando duas ou mais variáveis independentes são altamente correlacionadas.
- Dependendo do grau de multicolinearidade, as estimativas dos parâmetros podem ter sinais contrários ao que se poderia esperar.

- Uma das técnicas para verificar a multicolinearidade é por meio do VIF (Fator de inflação da variância) que é calculado da seguinte maneira:

$$VIF_j = \frac{1}{1 - R_j^2}$$

em que R_j^2 é o coeficiente de determinação da variável explicativa x que foi considerada como variável dependente em detrimento a todas as demais variáveis explicativas.

- Aquela variável que apresentar um VIF maior que 5, deve ser descartada do modelo.
- Uma outra forma de resolver a multicolinearidade é fazer um rescalonamento na variável ou utilizar a regressão **ridge**. No entanto, este assunto ficará para uma disciplina mais avançada em regressão.

Aplicação

5. Um pesquisador está interessado em predizer a quantidade do meio bacteriano (L/semana) necessário para executar biorreatores que suportam o crescimento microbiano contínuo ao longo de um mês. Os pesquisadores utilizaram as seguintes variáveis independentes: temperatura do biorreator (x1), log10 da população microbiana por mm^2 (x2), concentração do meio proteico (x3), razão cálcio/fósforo (x4), nível de nitrogênio (x5) e nível de metal pesado (x6). Seguem os dados.

x1	x2	x3	x4	x5	x6	y
20	2.1	1.0	1.00	56	4.1	56
21	2.0	1.0	0.98	53	4.0	61
27	2.4	1.0	1.10	66	4.0	65
26	2.0	1.8	1.20	45	5.1	78
27	2.1	2.0	1.30	46	5.8	81
29	2.8	2.1	1.40	48	5.9	86
37	5.1	3.7	1.80	75	3.0	110
37	2.0	1.0	0.30	23	5.0	62
45	1.0	0.5	0.25	30	5.2	50
20	3.7	2.0	2.00	43	1.5	41
20	4.1	3.0	3.00	79	0.0	70
25	3.0	2.8	1.40	57	3.0	85
35	6.3	4.0	3.00	75	0.3	115
26	2.1	0.6	1.00	65	0.0	55
40	6.0	3.8	2.90	70	0.0	120

a. Calcule o VIF para cada variável independente e verifique se há problema de multicolinearidade.

Apresentando os cálculos detalhados apenas para a variável x1 em função das outras variáveis explicativas. Seja o modelo $x1 = x2 + x3 + x4 + x5 + x6$. Então o $R^2_{x1} = 0.4833$. Logo:

$$VIF_{x1} = \frac{1}{1 - 0.4833} = 1.935475$$

Para as demais variáveis tem-se os seguintes valores de VIF: $x2 = 15.5987, x3 = 11.5051, x4 = 11.0841, x5 = 2.7915, x6 = 3.9961$.

Portanto, as variáveis $x2, x3$ e $x4$ apresentaram um VIF maior que 5 ,indicando problemas de multicolinearidade.

b. Ajuste um modelo de regressão apenas com as variáveis que não apresentaram um VIF maior que 5 e apresente o coeficiente de determinação ajustado.

b. …

Antes de removermos todas a variáveis com alto VIF, vamos obter a matriz de correlação entre as variáveis. É provável que se as três variáveis com alto VIF estiverem altamente correlacionadas, basta remover duas delas e não as três, melhorando deste modo a explicabilidade do modelo. Segue a matriz de correlação:

	x1	x2	x3	x4	x5	x6
x1	1.000	0.214	0.187	-0.083	-0.175	0.079
x2	0.214	1.000	0.920	0.895	0.695	-0.685
x3	0.187	0.920	1.000	0.860	0.624	-0.489
x4	-0.083	0.895	0.860	1.000	0.748	-0.734
x5	-0.175	0.695	0.624	0.748	1.000	-0.697
x6	0.079	-0.685	-0.489	-0.734	-0.697	1.000

Podemos observar que as variáveis x2 e x4 possui valores mais alto de correlação com as demais variáveis quando comparada a x3. Logo, é provável que a remoção apenas destas duas variáveis resolva o problema da multicolinearidade.

b. …

Ajustando um modelo com as variáveis dependentes x1, x3, x5 e x6 tem-se os seguintes VIF's: $x1 = 1.2122, x3 = 1.9353, x5 = 2.7453, x6 = 1.9624$. Logo, não há mais o problema da multicolinearidade. Segue o modelo ajustado juntamento com o coeficiente de determinação.

$$\hat{y} = -28.25 + 1.0403x1 + 14.2475x3 + 0.5964x5 + 3.8167x6$$

$$R^2_{ajust} = 0.8437$$

Análise de superfície resposta

- O **método de superfície resposta** é assunto que certamente se encaixa em uma disciplina devido a sua extenção e complexidade.
- Logo, será abordado apenas uma particularidade da superfície resposta que ocorre quando uma regressão múltipla tem apenas duas variáveis independentes.
- A abardagem será mais gráfica do que da própria metodologia da superfície resposta per si.
- Neste caso para elaboração dos gráficos será utilizado o software R.

Aplicação

6. O artigo "The undrained strength of some thawed permafrost soils" contém os dados a seguir sobre a resistência ao corte de solos arenosos sem drenagem (y,em KPa), profundidade (x1, em m) e conteúdo de água (x2, em %).

y	x1	x2
14.7	8.9	31.5
48.0	36.6	27.0
25.6	36.8	25.9
10.0	6.1	39.1
16.0	6.9	39.2
16.8	6.9	38.3
20.7	7.3	33.9
38.8	8.4	33.8
16.9	6.5	27.9
27.0	8.0	33.1
16.0	4.5	26.3
24.9	9.9	37.8
7.3	2.9	34.6
12.8	2.0	36.4

a. Ajuste um modelo completo do segundo grau. Todos os termos são significativos? Escreva o modelo final ajustado.

```
##
## Call:
## lm(formula = y ~ x1 * x2 + I(x1^2) + I(x2^2), data = dad6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6478 -2.3660  0.4494  1.8326 13.3732
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -151.35551  134.06915  -1.129   0.2916
## x1           -16.21614    8.83147  -1.836   0.1037
## x2            13.47558    8.18686   1.646   0.1384
## I(x1^2)        0.09353    0.07093   1.319   0.2238
## I(x2^2)       -0.25282    0.12706  -1.990   0.0818 .
## x1:x2          0.49224    0.22813   2.158   0.0630 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.988 on 8 degrees of freedom
## Multiple R-squared:  0.7586, Adjusted R-squared:  0.6077
## F-statistic: 5.028 on 5 and 8 DF,  p-value: 0.02223
```

a. …Falarmos de significância quando se trata de parâmetros em regressão sempre é delicado. Acredito que $\alpha = 0.10$ seja um bom nível de significância. Sempre é bom ir removendo variável por variável e ir testando e observando a significância dos betas. No caso acima, considerando o α já exposto, vamos abandonar primeiro o termo quadrático da variável x1 e fazer um novo ajuste.

```
##
## Call:
## lm(formula = y ~ x1 * x2 + I(x2^2), data = dad6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.6245  -2.9795  -0.5391   2.7644  12.4675
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -157.1633   139.3883  -1.128   0.2887
## x1            -5.0988     2.7353  -1.864   0.0952 .
## x2            11.6020     8.3870   1.383   0.1999
## I(x2^2)       -0.1998     0.1254  -1.594   0.1455
## x1:x2          0.2200     0.1010   2.179   0.0573 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.269 on 9 degrees of freedom
## Multiple R-squared:  0.7061, Adjusted R-squared:  0.5755
## F-statistic: 5.406 on 4 and 9 DF,  p-value: 0.01689
```

a. …Agora vamos remover o termo quadrático da variável x2 já que não foi significativa e ajustar novamente o modelo.

```
##
## Call:
## lm(formula = y ~ x1 * x2, data = dad6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.256 -3.518 -2.324  4.113 16.084
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  60.6963    29.2462   2.075   0.0647 .
## x1           -5.6230     2.9170  -1.928   0.0828 .
## x2           -1.6896     0.9500  -1.779   0.1057
## x1:x2         0.2337     0.1081   2.163   0.0558 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.809 on 10 degrees of freedom
## Multiple R-squared:  0.6232, Adjusted R-squared:  0.5102
## F-statistic: 5.513 on 3 and 10 DF,  p-value: 0.01702
```

Parece que chegamos em um modelo razoável. Portanto a equação estimada ficou:

$$\hat{y} = 60.696 - 5.623x1 - 1.6896x2 + 0.2337x1x2$$

b. Elabore um gráfico com o modelo ajustado.

Como há duas variáveis explicativas no modelo, vamos elaborar um gráfico de superfície de resposta. Segue o gráfico: