Regressão para tratamentos com níveis quantitativos

Universidade Estadual de Santa Cruz

Ivan Bezerra Allaman

INTRODUÇÃO

- Quando os níveis de um fator são do tipo quantitativo, é prudente que se faça o uso de regressão ao invés das técnicas de comparações múltiplas de médias.
- Exemplo de tratamentos com níveis quantitativos são:
  - Efeito de 5 espaçamentos na produção de sementes de soja (46, 61, 76, 91, 106);
  - Efeito de 4 níveis de nitrogênio na produção de cana de açúcar (0, 60, 120, 240);
  - Efeito de 5 níveis de proteína bruta sobre a conversão alimentar de suínos (10%,12%,14%,16%,18%);
  - Efeito da adição de sorgo no ganho de peso de pintainhos (sem sorgo (0%), 10%, 20%, 30%);

- Dentre as diversas técnicas de regressão existentes na literatura podemos destacar:
    - Regressão por polinômios ortogonais;
    - Regressão segmentada;
    - Regressão não-linear;
- Antes de aplicar qualquer técnica de regressão, é aconselhável que se faça um gráfico da variável resposta em função dos níveis quantitativos. Este passo é extremamente importante, pois evitará perda de tempo de ajustes de regressão que não faz sentido aos dados.

REGRESSÃO POR POLINÔMIOS ORTOGONAIS

Conceituação

- É a técnica mais comum em experimentos cujo o fator (tratamento) possui níveis quantitativos.
- Um polinômio é uma regressão de ordem **k** que pode ser expressa pela seguinte equação:

$$y_{ij} = \beta_0 + \beta_1 * x + \beta_2 * x^2 + \beta_3 * x^3 + \cdots + \beta_k * x^k + \varepsilon_{ij}$$

- Na maioria dos casos, um modelo só é interpretável biologicamente até o grau 2 do polinômio, ou seja, até uma regressão quadrática.
- Quando ocorre significância de um grau maior ou igual a 3 do polinômio, caso não haja interpretação biológica, deve-se buscar outras técnicas de regressão. Veremos de um modo bem sucinto nesta disciplina a regressão segmentada e a regressão não-linear.

- Iremos trabalhar especificamente com polinômios **ortogonais**, ou seja, para decompormos a soma de quadrados de tratamentos em termos independentes de cada grau do polinômio que desejamos estudar, os contrastes devem ser **ortogonais** conforme visto no assunto *contrastes ortogonais*.
- Os níveis do fator em estudo não precisam ser necessariamente igualmente espaçados. Este tipo de exigência foi de uma época em que só haviam disponíveis as tabelas de Fisher e Yates e mais tarde Anderson e Houseman, em que, mostravam coeficientes e divisores para uma certa quantia de tratamentos e para um certo grau do polinômio desejado.
- A abordagem será exclusivamente computacional, uma vez que, não iremos requerer um conhecimento prévio de álgebra de matrizes.

Análise de variância

- Só é possível estudar um polinômio de ordem **k-1**, em que, *k* é o número de níveis do tratamento em questão.
- Por exemplo, ao estudarmos 4 níveis de um dado tratamento, poderemos então decompor a soma de quadrados de tratamentos até a ordem 3 do polinômio. Logo, a tabela da anova é:

Fontes de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	F calculado
Tratamentos	4 - 1	SQ_{trat}	$QM_{trat} = SQ_{trat}/(k-1)$	$\frac{QM_{trat}}{QM_{erro}}$
Regressão linear	1	SQ_{linear}	$QM_{linear} = SQ_{linear}/1$	$\frac{QM_{linear}}{QM_{erro}}$
Regressão quadrático	1	$SQ_{quadrático}$	$QM_{quadrático} = SQ_{quadrático}/1$	$\frac{QM_{quadrático}}{QM_{erro}}$
Regressão cúbica	1	$SQ_{cúbica}$	$QM_{cúbica} = SQ_{cúbica}/1$	$\frac{QM_{cúbica}}{QM_{erro}}$
Erro	N - 4	SQ_{erro}	$QM_{erro} = SQ_{erro}/(N-k)$	
Total	N - 1	SQ_{total}		

- Todos os pressupostos devem ser avaliados antes de qualquer conclusão da anova, ou seja, normalidade e homocedasticidade dos resíduos.
- As técnicas de avaliação dos pressupostos são as mesmas já abordados nos assuntos anteriores.

Coeficiente de determinação (R^2)

- Tem a mesma interpretação do assunto *regressão linear simples* abordado na disciplina *nivelamento em estatística*.

- No caso do ajuste de uma regressão linear, o R^2 é calculado como:

$$R^2 = \frac{SQ_{linear}}{SQ_{tratamento}}$$

- No caso do ajuste de uma regressão quadrática tem-se:

$$R^2 = \frac{SQ_{linear} + SQ_{quadrático}}{SQ_{tratamento}}$$

- No caso do ajuste de uma regressão cúbica tem-se:

$$R^2 = \frac{SQ_{linear} + SQ_{quadrático} + SQ_{cúbica}}{SQ_{tratamento}}$$

- E o mesmo raciocínio vale para as demais ordens do polinômio.

Ponto de mínimo/máximo de uma regressão quadrática

- Quando há um ajuste quadrático, é possível determinarmos o ponto de mínimo ou de máximo da equação de regressão.
- Este ponto é calculado como a primeira derivada da equação.
- Supondo que a equação ajutada foi $\hat{y} = 2 + 20 * x + 5 * x^2$.
- Derivando a equação e igualando a zero, encontraremos o valor de **x** que nos fornecerá um valor de mínimo ou de máximo. Então,

$$\frac{d\hat{y}}{dx} = 20 + 2 * 5 * x$$
$$0 = 20 + 10 * x$$
$$x = \frac{-20}{10} = -2$$

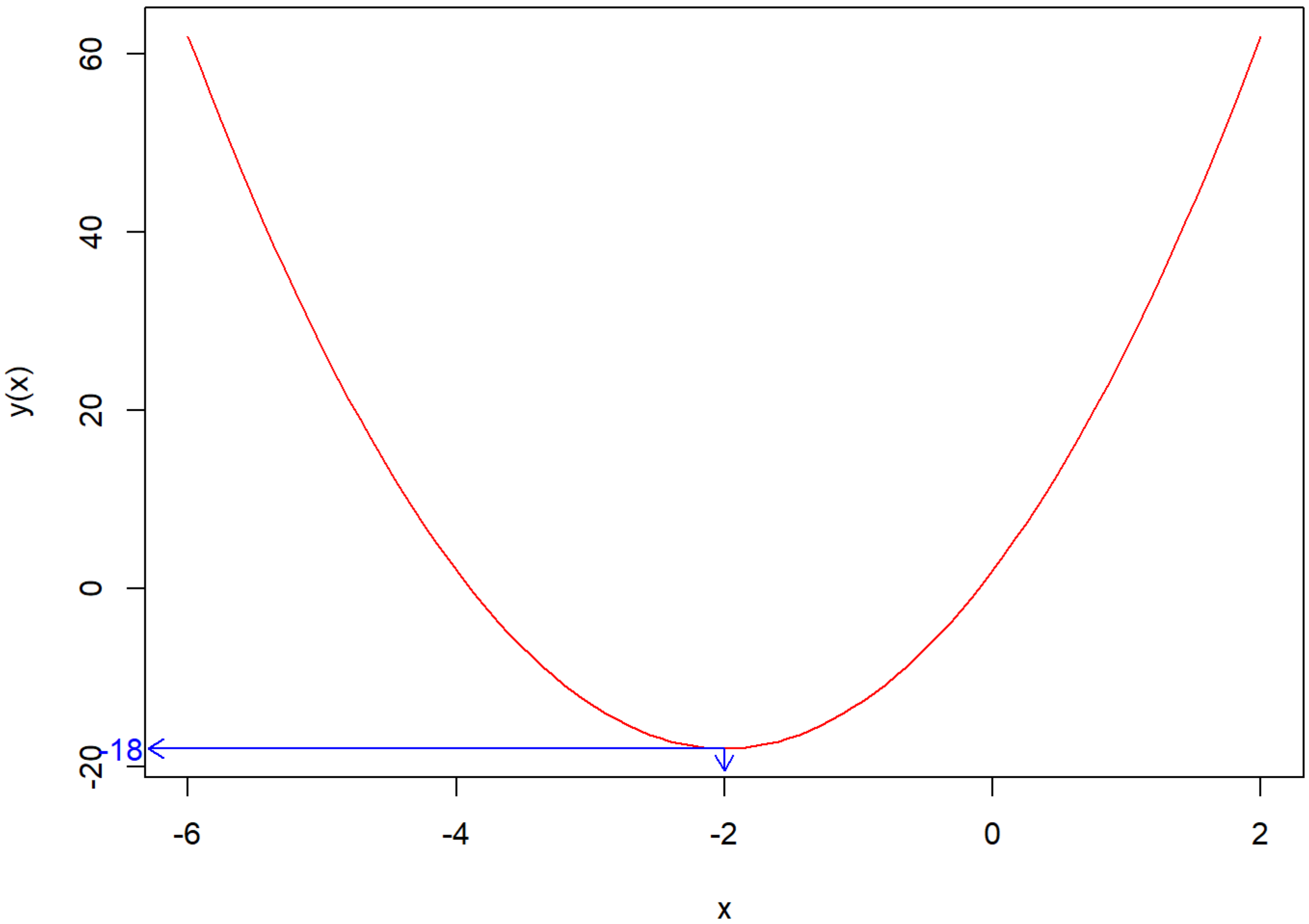
- Para sabermos se o ponto encontrado retorna um valor de $\hat{y}$ mínimo ou máximo, basta derivarmos novamente (segunda derivada) e observarmos o sinal. Se for positivo, o $\hat{y}$ será um mínimo, caso contrário, será um máximo.
- Na exemplo anterior temos:

$$\frac{d^2\hat{y}}{dx^2} = 2 * 5 = 10$$

- Portanto, para encontrarmos qual é o valor de mínimo, basta substituirmos o valor de $x = -2$ na equação ajustada. Logo,

$$\begin{aligned}\hat{y}_{mínimo} &= 2 + 20 * (-2) + 5 * (-2)^2 \\ &= 2 - 40 + 20 = -18\end{aligned}$$

- Graficamente, temos:



Aplicação

1. The goal of this experiment was to analyze the effect of protein level in a pig ration on feed conversion. The experiments started at an approximate weight of 39 kg and finished at 60 kg. There were five litters with five pigs randomly chosen from each litter. One of the five protein levels (10, 12, 14, 16 and 18%) was randomly assigned to each pig from each litter. The data can be to access in the link http://nbcgib.uesc.br/lec/download/R/dados/KapsPg_386.txt

2. The following data show the measurements of hemoglobin (grams per 100 ml) in the blood of brown trout. The trout were placed at random in four different troughs. The fish food added to the troughs contained, respectively, 0, 5, 10, and 15 grams of sulfamerazine per 100 pounds of fish. The measurements were made on ten randomly selected fish from each trough after 35 days. The data can be to access in the link http://nbcgib.uesc.br/lec/download/R/dados/VossPg_275.txt

3. Um ensaio conduzido por Torres e Pimentel-Gomes (1959) em que se usaram 4 rações (sem sorgo, e com 10%, 20% e 30% de sorgo) com duas repetições para pintainhos machos e outras duas para pintainhos fêmeas, com um delineamento inteiramente casualizado em um esquema fatorial do tipo 4 x 2 (4 níveis de sorgo e 2 níveis de sexo). Cada parcela continha inicialmente 13 aves. O peso total das parcelas em decagramas pode ser acessado pelo seguinte link http://nbcgib.uesc.br/lec/download/R/dados/PGomesPg_267.txt

4. Foi realizado um estudo com o intuito de avaliar o impacto de diferentes densidades e tempo de transporte em uma determinada espécie de peixe. Uma das variáveis analisadas foi o hematócrito. Os dados estão disponíveis no seguinte link: http://nbcgib.uesc.br/lec/download/R/dados/march.txt

REGRESSÃO SEGMENTADA

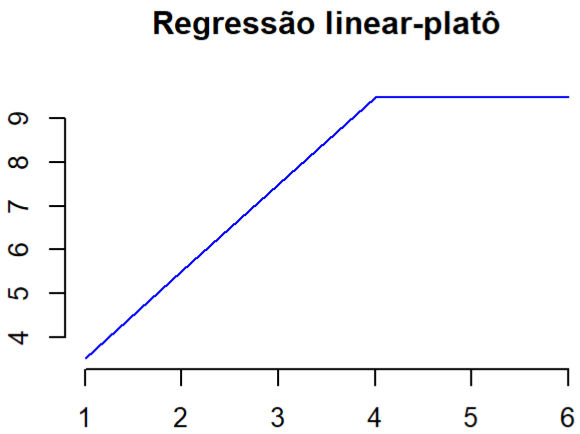
Conceituação

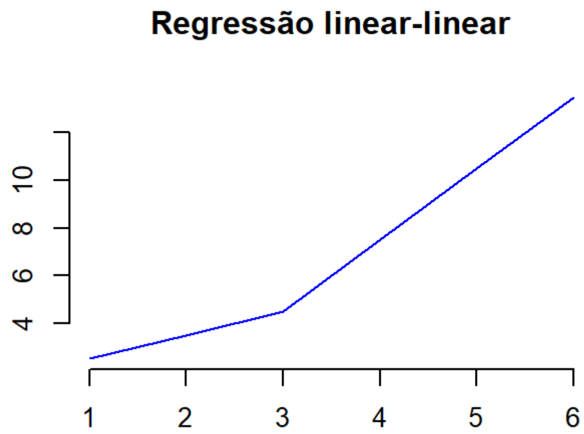
- Uma regressão segmentada, é um tipo de regressão ajustado por segmentos, ou seja, em cada segmento do eixo x, é ajustado um tipo de regressão aos dados.
- Dê um modo bem simples, podemos representar uma regressão segmentada por polinômios como:

$$f(y) = \begin{cases} \beta_{10} + \beta_{11} * x + \beta_{12} * x^2 + \cdots + \beta_{nk} * x^k + \varepsilon, & x \leq S_1 \\ \beta_{20} + \beta_{21} * x + \beta_{22} * x^2 + \cdots + \beta_{nk} * x^k + \varepsilon, & S_1 \leq x \leq S_2 \\ \vdots \qquad\qquad\qquad \vdots \qquad\qquad\qquad \vdots & \\ \beta_{n0} + \beta_{n1} * x + \beta_{n2} * x^2 + \cdots + \beta_{nk} * x^k + \varepsilon, & S_n \leq x \end{cases}$$

- em que:
  - B_{nk} são os parâmetros da regressão e S_n é o n-ésimo ponto de quebra da regressão.

- Na regressão segmentada, existem usualmente dois objetivos:
  - Estimar os parâmetros da regressão nos dados;
  - Determinar realmente se os dados tem padrões separados ou se os dados devem ser modelados sem um ponto de quebra;
- Como exemplo de regressão segmentada podemos citar: linear-platô, linear-linear, quadrático-platô entre outros.
- As figuras abaixo ajudam a visualizar os tipos de regressões citadas.


Regressão linear-platô
4
5
6
7
8
9
1
2
3
4
5
6


Regressão linear-linear
4
6
8
10
1
2
3
4
5
6


Regressão quadrática-platô
-17
-15
-13
-11
1
2
3
4
5
6

- Existe vários métodos de estimação de parâmetros para modelos de regressão segmentada como: método da máxima verossimilhança, busca na grade, bootstrap, regressão piecewise e métodos não-paramétricos.
- Como será utilizado uma abordagem prática, lançaremos mão de recursos computacionais.
- Segue alguns modelos de regressão segmentada considerando apenas um ponto de quebra.

Linear - Platô

$$f(y) = \begin{cases} \beta_0 + \beta_1 * x + \varepsilon, & x \leq S_1 \\ \beta_0 + \beta_1 * S_1 + \varepsilon, & x > S_1 \end{cases}$$

Quadrático - Platô

$$f(y) = \begin{cases} \beta_0 + \beta_1 * x + \beta_2 * x^2 + \varepsilon, & x \leq -0,5 * \beta_1/\beta_2 \\ \beta_0 - \frac{\beta_1^2}{4 * \beta_2} + \varepsilon, & x > -0,5 * \beta_1/\beta_2 \end{cases}$$

Linear - Linear

$$f(y) = \begin{cases} \beta_0 + \beta_1 * x + \varepsilon, & x \leq S_1 \\ \beta_0 + S_1 * (\beta_1 - \beta_2) + \beta_2 * x + \varepsilon, & x > S_1 \end{cases}$$

5. The bean-soaking experiment was run by Gordon Keeler in 1984 to study how long mung bean seeds ought to be soaked prior to planting in order to promote early growth of the bean sprouts. The experiment was run using a completely randomized design. The data can be to access in the link http://nbcgib.uesc.br/lec/download/R/dados/VossPg_264.txt