Análise de variância (ANOVA)

Universidade Estadual de Santa Cruz

Ivan Bezerra Allaman

CRONOGRAMA

1. História
2. Concepção da ideia
3. Formalização da ideia e o surgimento da distribuição F
4. Hipóteses e Organizando as ideias em uma tabela
5. Pressupostos
6. Exemplos

História

- A análise de variância foi proposta pelo gênio Ronald Aylmer Fisher.

- O primeiro relato da técnica foi em 1923.

32

STUDIES IN CROP VARIATION.

II. THE MANURIAL RESPONSE OF DIFFERENT POTATO VARIETIES.

By R. A. FISHER, M.A. and W. A. MACKENZIE, B.Sc.

Rothamsted Experimental Station, Harpenden.

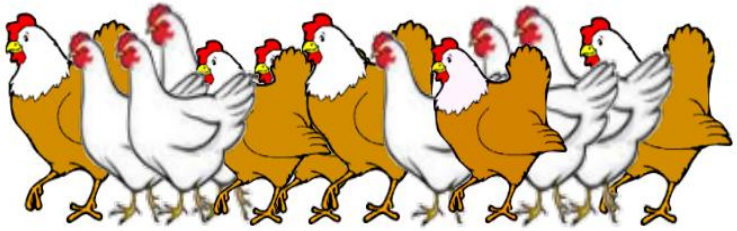
- A anova revolucionou a ciência do século XX e ainda continua soberana.
- Um pouco diferente de como a conhecemos hoje, eis a primeira tabela da anova.

Table III.

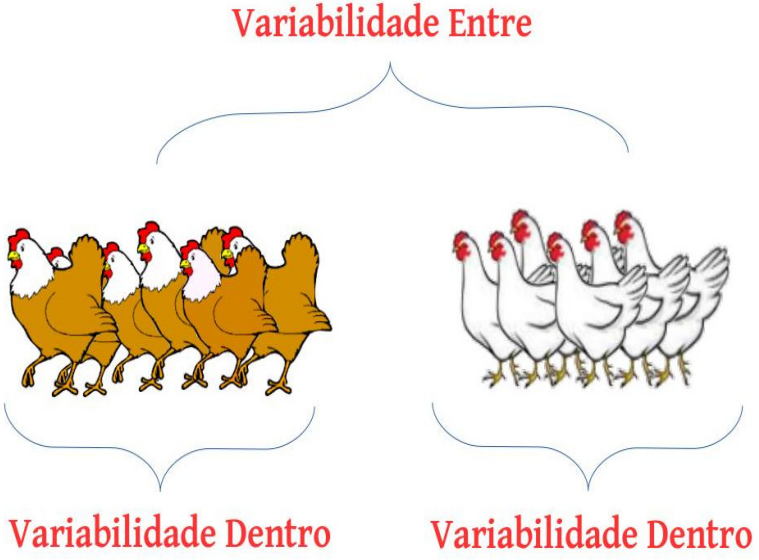
Variation due to	Degrees of freedom	Sum of squares	Mean square	Standard deviation
Manuring	5	6,158	1231·6	35·09
Variety	11	2,843	258·5	16·07
Deviations from summation formula	55	981	17·84	4·22
Variation between parallel plots ...	141	1,758	12·47	3·53
Total	212	11,740	—	—

Concepção da ideia

Você consegue distinguir as causas de variabilidade entre os animais?



- Basicamente, é possível distinguirmos as seguintes variabilidades:

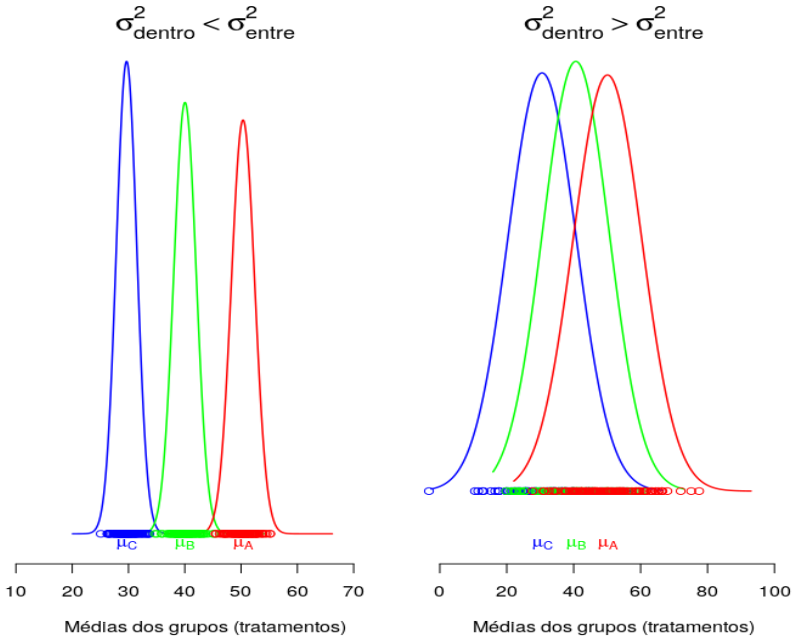


- No entanto, como podemos julgar se as duas linhagens da figura anterior diferen quanto a uma determinada característica quantitativa?
- Quando pensamos em comparar grupos quanto a características quantitativas nos vem à mente a média. Mas então qual é a relação entre a variabilidade exposta na figura anterior com a média? De que maneira isto nos levará a conclusão de que os grupos (linhagens no nosso exemplo) são diferentes?
- Vamos tentar chegar na resposta manipulando a simulação abaixo.

Formalização da ideia e o surgimento da distribuição F

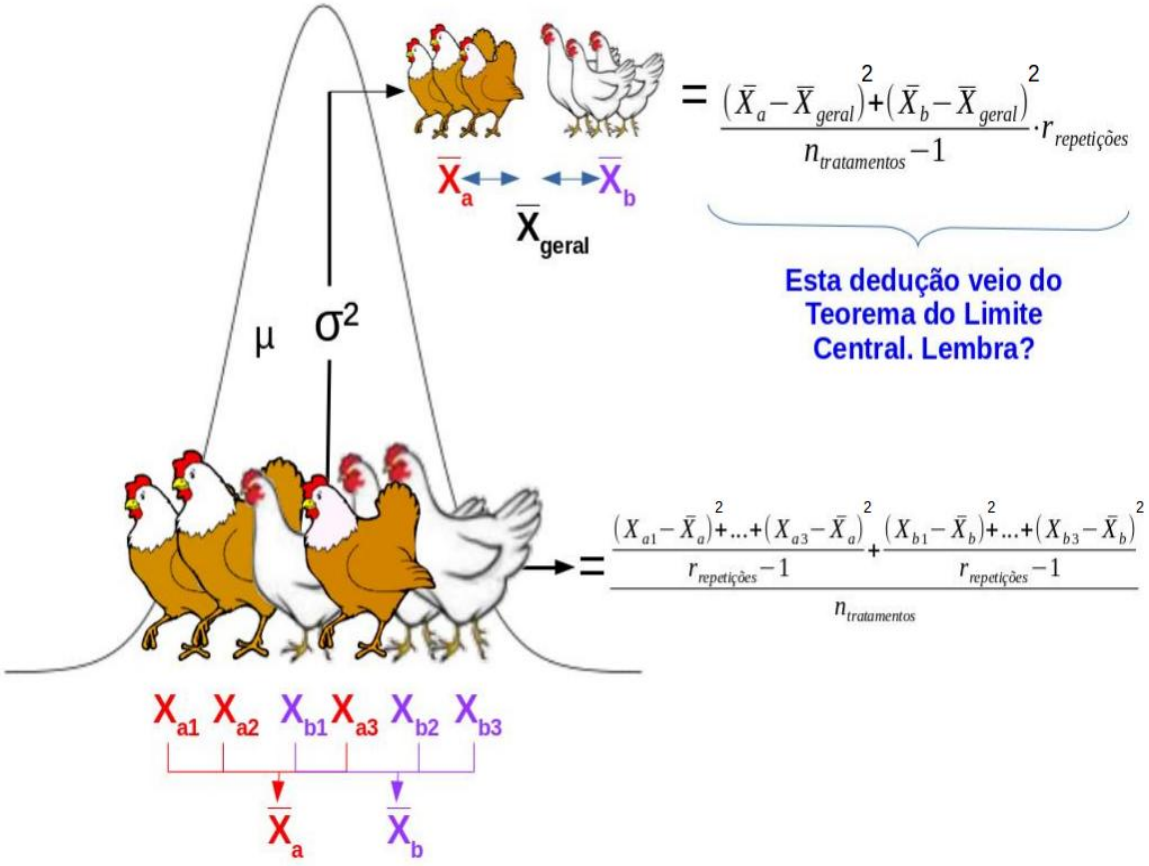
Formalização da ideia

- Vimos na ferramenta virtual que para detectarmos diferenças médias entre vários grupos, teríamos que contrastar a variabilidade existente entre os grupos com a variabilidade dentro dos grupos.



- Como a variância é uma técnica que visa avaliar a distância média entre um valor observado com a tendência central, Fisher vislumbrou em tal técnica um meio para comparar se mais de dois grupos eram em média distintos ou não.
- Então Fisher pensou: "A variância entre os grupos de fato me fornece uma medida do quanto as médias estão distantes entre si, pois avalia a soma dos desvios quadráticos em relação a uma média geral entre os grupos".
- Logo Fisher concluiu: "Deste modo, basta avaliar a razão entre a variabilidade entre os grupos e a variabilidade dentro dos grupos (do acaso). Se a variabilidade do acaso superar a variabilidade entre os grupos, há uma grande chance dos grupos pertencerem a uma mesma população".

- Fisher partiu da hipótese de que todos os tratamentos (grupos) eram provenientes de uma mesma população com distribuição normal de probabilidade com média μ e variância σ^2. Com isto em mente, enchergou como determinar as variabilidades entre e dentro de tratamentos.


$\bar{X}_a$
$\bar{X}_b$
$\bar{X}_{geral}$
$= \frac{(\bar{X}_a - \bar{X}_{geral})^2 + (\bar{X}_b - \bar{X}_{geral})^2}{n_{tratamentos} - 1} \cdot r_{repetições}$
Esta dedução veio do Teorema do Limite Central. Lembra?
μ
σ²
$= \frac{\frac{(X_{a1} - \bar{X}_a)^2 + \ldots + (X_{a3} - \bar{X}_a)^2}{r_{repetições} - 1} + \frac{(X_{b1} - \bar{X}_b)^2 + \ldots + (X_{b3} - \bar{X}_b)^2}{r_{repetições} - 1}}{n_{tratamentos}}$
X_{a1} X_{a2} X_{b1} X_{a3} X_{b2} X_{b3}
$\bar{X}_a$
$\bar{X}_b$

Surgimento da distribuição F

- Tal chance (probabilidade) foi descrita formalmente por Fisher como distribuição de probabilidade em 1924 em um congresso de matemática em Toronto,e publicado como anais. Naquela ocasião, Fisher apresentou a estatística Z com a seguinte expressão:

$$z = \frac{1}{2} \ln \frac{s_1}{s_2}$$

- Em 1932, Mahalanobis tabulou os quantis da razão entre as duas variâncias da quantidade anterior para 1% e 5%.
- Em 1934, Snedecor citou a razão entre duas variâncias como a estatística F.

- Logo, a razão entre as duas variâncias deu origem a distribuição F de Snedecor, sendo F em homenagem ao Fisher.

$$F = \frac{s_1^2/gl_1}{s_2^2/gl_2}$$

Nota: gl_1 se refere aos graus de liberdade associado a variância 1 e gl_2 o graus de liberdade associado a variância 2.

Modelo estatístico, Hipóteses e Organizando as ideias em uma tabela

Modelo estatístico

- Podemos representar os efeitos sobre uma variável resposta da seguinte forma:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

em que:

- y_{ij} = é a observação da repetição **j** no tratamento **i**;
- μ = é a média geral associada a todas as observações;
- τ_i = é o efeito do tratamento **i**;
  - tal efeito é medido como a substração da média do tratamento **i** pela média geral ($\bar{x}_i - \mu$).
- ε_{ij} = é o erro associado a observação da repetição **j** no tratamento **i**;

Hipóteses em uma ANOVA

- Como o interesse é comparar médias de grupos, as hipóteses são formuladas da seguinte maneira:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_1 : \mu_i \neq \mu_j$$

com $i \neq j$.

Organizando as ideias em uma tabela

- A soma de quadrados total é particionado em dois: uma parte devido a fonte de variação conhecida e a outra parte a fonte de variação desconhecida.
- Matematicamente tem-se a seguinte relação da soma de quadrados:

$$SQ_{total} = SQ_{entre} + SQ_{dentro}$$

$$\sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 = \sum_i n_i (\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2$$

- Agora podemos organizar os cálculos na seguinte tabela:

Fontes de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	F calculado
Entre	k - 1	SQ_{entre}	$QM_{entre} = SQ_{entre}/gl$	$\frac{QM_{entre}}{QM_{dentro}}$
Dentro	N - k	SQ_{dentro}	$QM_{dentro} = SQ_{dentro}/gl$	
Total	N - 1	SQ_{total}		

Pressupostos

O protocolo

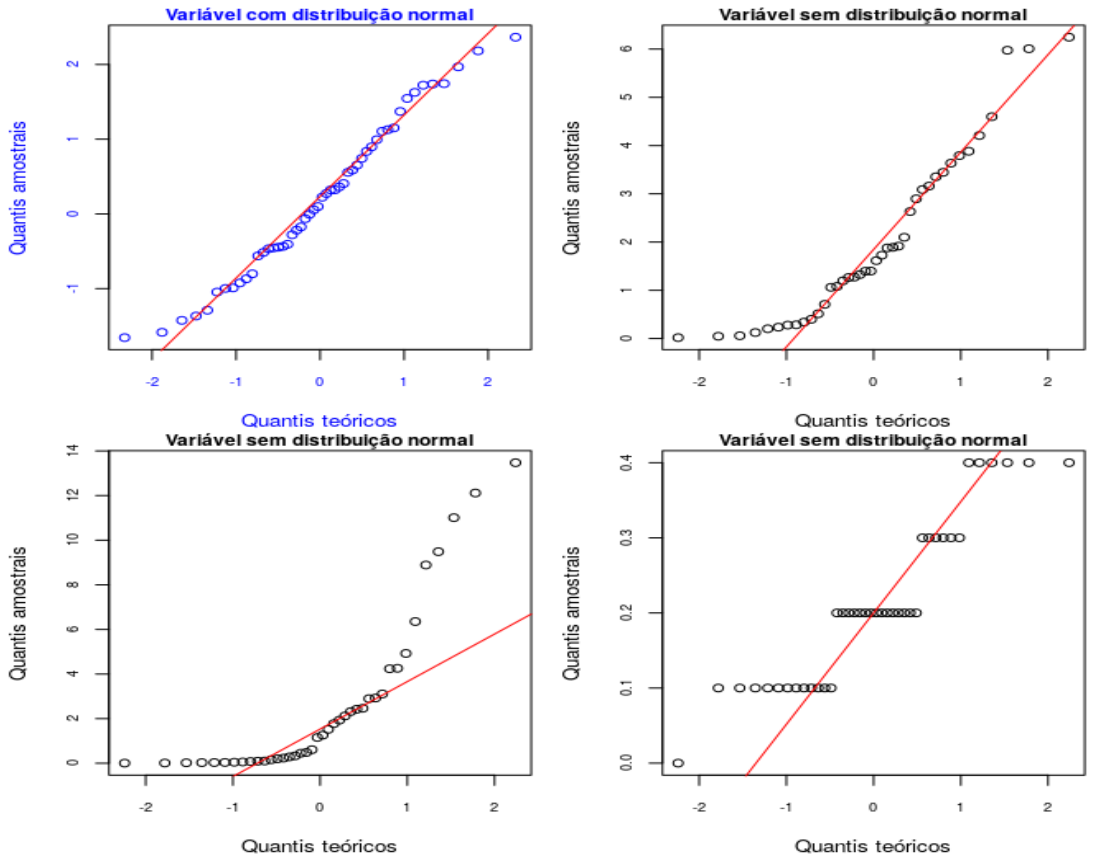
- São basicamente três os pressupostos que devem ser seguidos para que a análise de variância seja válida:
  - Os erros devem ser independentes e identicamente distribuídos.
  - Os erros devem ter distribuição normal com média 0 e variância σ^2.
  - Os erros devem ser homocedásticos, ou seja, a dispersão dos erros deve ser semelhante entre os tratamentos. Isto é equivalente ao dizer que a variância dos tratamentos deve ser semelhante entre si.
- O erros são estimados a partir do modelo estatístico como:

$$\hat{\varepsilon_{ij}} = y_{ij} - \underbrace{(\mu + \tau_i)}_{\hat{y}_{ij}}$$

Avaliando a normalidade

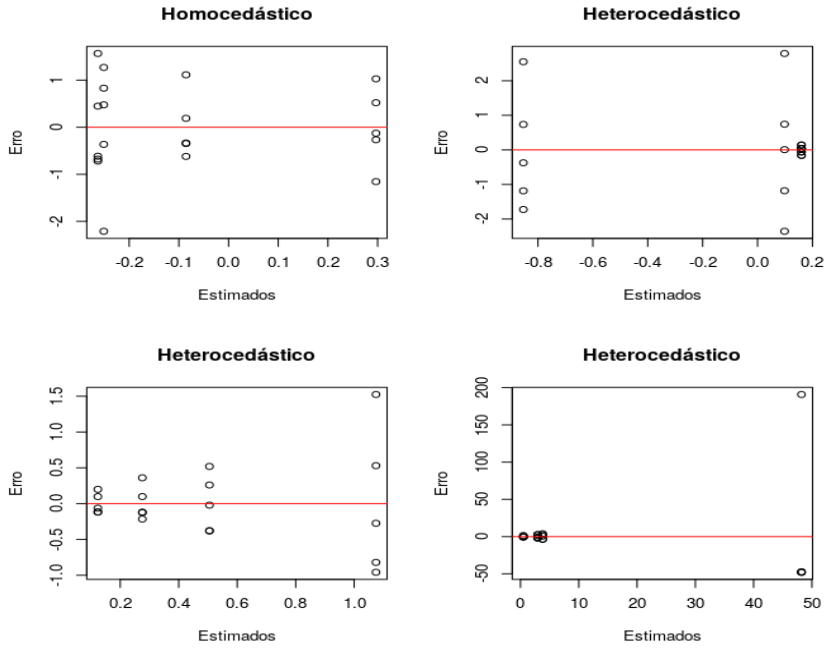
- No assunto **regressão linear simples** na disciplina de **Nivelamento em estatísca**, foi ministrado com detalhes como fazer tal avaliação de modo manual e utilizando o software `r`. Portanto, tal cálculo será feito utilizando diretamente a função do software `r`.

- Exemplos do qqplot.



Avaliando a homocedasticidade

- Também foi ministrado com detalhes no assunto de **regressão linear simples** na disciplina de **Nivelamento em estatísca**.



- Quando há dúvidas quanto a homocedasticidade de variância dos tratamentos, podemos lançar mão do teste F-máximo de Hartley. Tal teste avalia a razão entre a variância máxima de um dado tratamento e a variância mínima de outro tratamento. Tal razão é testada por meio da estatística F com graus de liberdade no numerador e denominador iguais ao número de repetições menos 1 (r-1).
- Portanto tem-se a seguinte estatística F.

$$F_{\text{máximo}} = \frac{S^2_{máximo}/(r-1)}{S^2_{mínimo}/(r-1)}$$

Aplicação

1. Quatro linhagens de galinhas (A,B,C e D) foram cruzadas para obter outras quatro linhagens AB, AC, BC e BD. O peso dos ovos (g) das linhagens cruzadas seguem abaixo. Podemos afirmar que as linhagens se diferem significativamente quanto ao peso médio dos ovos? Considere uma $\alpha = 0.03$.

AB	AC	BC	BD
58	59	56	59
51	62	57	55
56	64	56	50
52	60	55	64
54	62	54	57

As hipóteses a serem testadas são:

$$H_0 : \mu_{AB} = \mu_{AC} = \mu_{BC} = \mu_{BD}$$
$$H_a : \mu_i \neq \mu_j, \text{com } i \neq j$$

Vamos determinar a soma de quadrados entre e dentro de tratamentos.

$$\begin{aligned} SQ_{entre} &= 5 \cdot \{(54.2 - 57.05)^2 + (61.4 - 57.05)^2 + \cdots + \\ &+(57 - 57.05)^2\} \\ &= 145,75 \\ SQ_{dentro} &= (58 - 54.2)^2 + (51 - 54.2)^2 + \cdots + (57 - 57)^2 \\ &= 159,2 \end{aligned}$$

- Segue então a tabela da análise de variância.

Fontes de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	F calculado
Entre	4 - 1	$145,75$	$QM_{entre} = 145,75/3$	$\frac{48,58}{9,95} = 4,88$
Dentro	20 - 4	$159,2$	$QM_{dentro} = 159,2/16$	
Total	20 - 1	$304,95$		

- Calculando o p-valor tem-se:

$$\text{pf}(4.88, 3, 16, lower.tail = FALSE) = 0,0135.$$

- Logo, considerando o nível de significância adotado na pesquisa, rejeita-se H0 com 97% de confiança.

2. Considerando a aplicação 1, verifique se os pressupostos da análise de variância são válidos.

   As médias dos tratamentos foram: AB = 54,2; AC = 61,4; BC = 55,6; BD = 57.

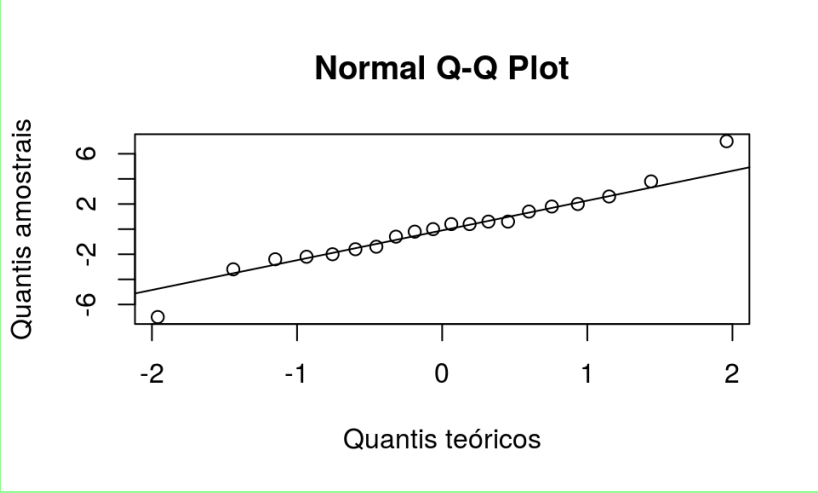
   A média geral foi de 57,05.

   Logo o erro estimado da observação 1 no tratamento 1 foi:

$$\hat{\varepsilon}_{11} = 58 - (57,05 + (54,2 - 57,05)) = 3,8$$

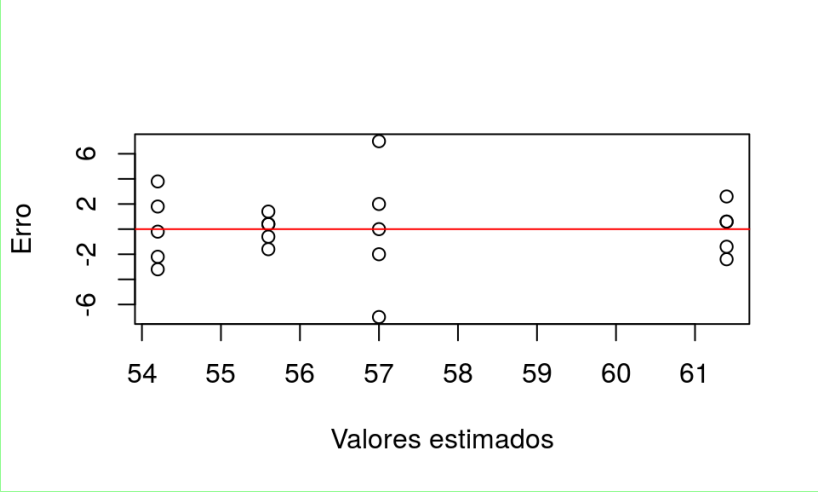
   Os demais erros são estimados seguindo o mesmo raciocínio anterior.

Vamos verificar primeiro a normalidade dos erros, por meio do gráfico qqplot e verificar se os pontos estão próximos da reta ao longo do eixo x.



Podemos concluir que o erro tem distribuição normal.

Para avaliar a homocedasticidade, basta elaborar um gráfico do erro em função dos valores estimados.



Parece que há uma discrepância significativa entre a dispersão dos erros de dois tratamentos. Nestes casos em que há dúvidas, podemos lançar mão de um teste inferencial como o teste F-máximo de Hartley.

A variância dos tratamentos foram:

AB	AC	BC	BD
8.2	3.8	1.3	26.5

Calculando o F-máximo de Hartley tem-se:

$$\text{F-máximo} = \frac{26,5}{1,3} = 20,38$$

Tem-se 5 repetições. Logo o graus de liberdade é igual a 4. Portanto, calculando o p-valor tem-se:

$$\text{pf}(20.38, 4, 4, \text{lower.tail}{=}\text{FALSE}) = 0.00636$$

Logo, adotando um $\alpha = 0,01$ rejeita-se H_0, ou seja, as variâncias são heterocedástica.