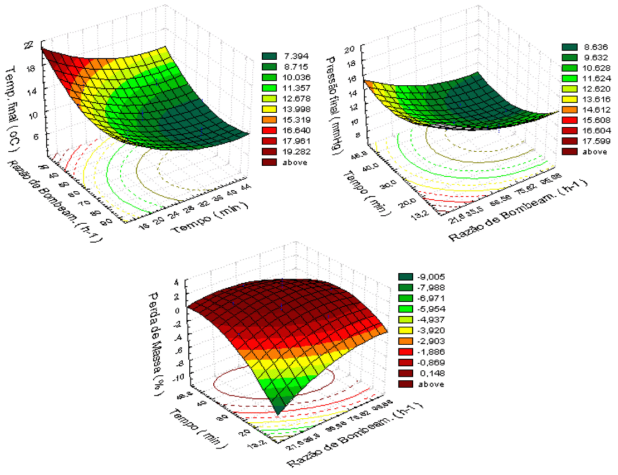
UFMG

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de ciências agrárias

2. Regressão múltipla: Testes de hipótese e avaliadores da qualidade do ajuste



Alcinei Mistico Azevedo
Professor Adjunto de Estatística e experimentação agrícola.

2.5 Análise de variância (para regressão sem delineamento)

- A análise de variância (ANOVA) é um dos testes estatísticos mais utilizados para a avaliação de experimentos em todas as áreas do conhecimento.
- Esta análise e vários fundamentos para a estatística moderna foi desenvolvido por **Ronald Fisher**.

Fonte: *Science Photo Library* - www.hamhigh.co.uk

Figura1: Fotografia de Ronald Fisher.

2.5 Análise de variância (para regressão sem delineamento)

- A análise de variância tem como principal finalidade a **decomposição da variação total em parte conhecida e desconhecida**.

  - A **variação total** corresponde a variância dos dados experimentais.
  - A **parte conhecida** corresponde a variância dos efeitos controlados pelo pesquisador:
    - **Neste contexto é a variação explicada pelo modelo de regressão**.
    - Controle local (blocos, linhas e colunas) em casos onde há delineamento estatístico.
  - **A parte desconhecida** refere-se à variação não controlada pelo pesquisador (resíduo). No contexto da análise de regressão **é toda variação que não é explicada pelo modelo de regressão**.

2.5 Análise de variância (para regressão sem delineamento)

- A análise de variância para regressão (sem uso de delineamento)

FV	GL	SQ	QM	F_c
Regressão	p	SQ_{Reg}	$QM_{Reg}=SQ_{Reg}/p$	QM_{Reg}/QM_{Res}
Resíduo	$N-p-1$	SQ_{Res}	$QM_{Res=}SQ_{Res}/(N-P-1)$	
Total	$N-1$	SQ_{Total}		

- Onde:
  - p é o número de parâmetros do modelo (o intercepto não é contado como parâmetro)
  - N é o número de observações (tamanho amostral) utilizado no ajuste da regressão

2.5 Análise de variância (para regressão sem delineamento)

2.5.1 Raciocínio lógico para a obtenção das somas de quadrado:

- Sendo a soma de quadrados total (*SQtotal*) a variação dos dados experimentais (dados observados, basta fazermos:

$$SQtotal = \sum_{i=1}^{N} (Y_i - \bar{Y})^2$$

$$SQtotal = \sum_{i=1}^{N} Y_i^2 - \frac{\left(\sum_{i=1}^{N} Y_i\right)^2}{N}$$

C (correção)

$$SQtotal = \underset{\sim}{Y}' \underset{\sim}{Y} - C$$

2.5 Análise de variância (para regressão sem delineamento)

2.5.1 Raciocínio lógico para a obtenção das somas de quadrado:

- Sendo a soma de quadrados da Regressao (*SQreg*) a variação explicada pelo modelo de regressão, basta obter a SQ com os valores preditos pelo modelo (Y_p):

$$SQreg = \sum_{i=1}^{N} (\widehat{Y}_{pi} - \overline{Y})^2$$

$$\underset{\sim}{\hat{Y}}_p = X\underset{\sim}{\hat{B}}$$

$$SQreg = \underset{\sim}{\hat{Y}}_p{}' \underset{\sim}{\hat{Y}}_p - C$$

$$SQreg = \underset{\sim}{B}' X' X\underset{\sim}{B} - C = \underset{\sim}{B}' X' Y - C$$

2.5 Análise de variância (para regressão sem delineamento)

2.5.1 Raciocínio lógico para a obtenção das somas de quadrado:

- A soma de quadrados do resíduo (*SQres*) é a parte da variação total que não pode ser explicada pela regressão. Logo pode ser obtido por diferença (SQres=SQtotal-SQreg).

$$SQres = SQtotal - SQreg$$

$$SQres = \underset{\sim}{Y}'\underset{\sim}{Y} - C - (\underset{\sim}{B}'X'Y - C)$$

$$SQres = \underset{\sim}{Y}'\underset{\sim}{Y} - \underset{\sim}{B}'X'Y$$

2.5 Análise de variância (para regressão sem delineamento)

2.5.1 Raciocínio lógico para a obtenção das somas de quadrado:

FV	GL	SQ	QM	F_c
Regressão	*p*	$\hat{\underset{\sim}{Y}}_p'\hat{\underset{\sim}{Y}}_p - C$	QM_{Reg}	QM_{Reg}/QM_{Res}
Resíduo	*N-P-1*	$\underset{\sim}{Y}'\underset{\sim}{Y} - \underset{\sim}{B}'X'Y$	QM_{Res}	
Total	*N-1*	$\underset{\sim}{Y}'\underset{\sim}{Y} - C$		

- Teste de hipótese (teste F):
  - Ho: A regressão é não significativa (A variação explicada pelo modelo de regressão deve-se ao acaso)
  - Ha: A regressão é significativa (A variação explicada pelo modelo de regressão não se deve ao acaso)

2.5 Análise de variância (para regressão sem delineamento)

Exemplo com os dados da Acerola: Modelo 9 ($z_i=a+bx_i+cy_i+dx_iy_i+e_i$)

$$\underset{\sim}{Y} = X \underset{\sim}{B} + \underset{\sim}{e}$$

$$\begin{bmatrix} 13,39 \\ 12,52 \\ 8,78 \\ 9,60 \\ \vdots \\ 8,23 \end{bmatrix} = \begin{bmatrix} 1 & 6,00 & 3,20 & 6,00*3,20 \\ 1 & 5,50 & 3,20 & 5,50*3,20 \\ 1 & 5,00 & 2,50 & 5,00*2,50 \\ 1 & 5,00 & 2,70 & 5,00*2,70 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 4,30 & 2,20 & 4,30*2,20 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ \vdots \\ e_{300} \end{bmatrix}$$

$$\hat{\underset{\sim}{B}} = (X'X)^{-1}X'Y = \begin{bmatrix} 0,2811 \\ -0.3494 \\ 0,2092 \\ 0,6350 \end{bmatrix}$$

$$SQreg = \underset{\sim}{B}'X'Y - C = 16722.05\text{-}14162.73 = 4582.56$$

$$SQres = \underset{\sim}{Y}'\underset{\sim}{Y} - \underset{\sim}{B}'X'Y = 16861.02\text{-}16722.05 = 138.98$$

$$SQtotal = \underset{\sim}{Y}'\underset{\sim}{Y} - C = 16861.02\text{-}14162.73 = 4721.54$$

2.5 Análise de variância (para regressão sem delineamento)

- Análise de variância para a regressão

FV	GL	SQ	QM	F_c	pValor
Regressão	*p =3*	4582.56	1527.52	3802.86	$2.2x10^{-16}$
Resíduo	*N-p-1=346*	138.98	*0,40*		
Total	*N-1=349*	4721.54			

pValor=1-pf(3802.863,3,346)

- Teste de hipótese (teste F):
  - Ho: A regressão é não significativa (A variação explicada pelo modelo de regressão deve-se ao acaso)
  - Ha: A regressão é significativa (A variação explicada pelo modelo de regressão não se deve ao acaso)

O teste para significância da regressão é um teste para determinar se há uma relação linear entre a variável resposta Y e algumas das variáveis regressora $x_1, x_2, \ldots, x_p$. Consideremos as hipóteses

$$\begin{cases} H_0: & \beta_1 = \beta_2 = \ldots = \beta_p = 0 \\ H_1: & \beta_j \neq 0 \quad \text{para qualquer } j = 1, \cdots, p \end{cases}$$

Se rejeitamos H_0, temos que ao menos uma variável explicativa $x_1, x_2, \ldots, x_p$ contribui significativamente para o modelo.

- Observação: Esta ANAVA é para um modelo global. Porém informações adicionais podem ser obtidas decompondo os GL e SQ da regressão por meio de um procedimento sequencial.

2.5 Análise de variância (para regressão sem delineamento)

- Os graus de liberdade e a SQ da regressão pode ser decompostos, obtendo-se uma fonte de variação referente à cada parâmetro do modelo estatístico. Para isso deve-se:
  - Obter a soma de regressão considerando apenas o primeiro parâmetro considerado do modelo de regressão.
  - Posteriormente, deve-se ir adicionando um parâmetro de regressão por vez, e a sua soma de quadrado será a subtração da SQreg atual com a SQreg obtida anteriormente.
  - A fonte de variação de cada parâmetro do modelo de regressão sempre terá apenas 1 grau de liberdade na ANAVA.

Metodologia para a decomposição dos graus de liberdade e soma de quadrado da regressão em cada um dos efeitos dos modelos de regressão. **Modelo 9 ($z_i=a+bx_i+cy_i+dx_iy_i+e_i$)**

- Obter SQ da regressão considerando no modelo apenas o primeiro coeficiente de regressão ($Z_i=a+bx_i+e_i$). $SQreg_{1coef} = \tilde{B}' X' Y - C = 4180.69$

- Obter SQ da regressão considerando no modelo apenas os dois primeiros coeficientes de regressão ($Z_i=a+bx_i+cy_i+e_i$). $SQreg_{2coef} = \tilde{B}' X' Y - C = 4423.777$

- Obter SQ da regressão considerando no modelo apenas os três primeiros coeficientes de regressão (neste caso, o modelo completo -> $Z_i=a+bx_i+cy_i+dx_iy_i+e_i$). $SQreg_{3coef} = \tilde{B}' X' Y - C = 16722.05\text{-}14162.73 = 4582.56$

GL dos efeitos é sempre 1

$SQreg_{1coef}$

$SQreg_{2coef}$- $SQreg_{1coef}$

$SQreg_{3coef}$- $SQreg_{2coef}$- $SQreg_{1coef}$

FV	GL	SQ	QM	F_c	pValor
Regressão	3	4582.56	1527.52	3802.86	
x	1	4180.69	4180.7	10408.21	$<2.2x10^{-16}$
y	1	243.09	243.1	605.16	$<2.2x10^{-16}$
xy	1	158.78	158.8	395.31	$<2.2x10^{-16}$
Resíduo	346	138.98	0,40		
Total	N-1=349	4721.54			

Modelo 9 ($z_i=a+bx_i+cy_i+dx_iy_i+e_i$)

FV	GL	SQ	QM	F_c	pValor
Regressão	3	4582.56	1527.52	3802.86	
x	1	4180.69	4180.7	10408.21	$<2.2x10^{-16}$
y	1	243.09	243.1	605.16	$<2.2x10^{-16}$
xy	1	158.78	158.8	395.31	$<2.2x10^{-16}$
Resíduo	346	138.98	0,40		
Total	N-1=349	4721.54			

Curiosidade: Vamos fazer a mesma análise de variância, porém com os parâmetros com outra ordem -> $z_i=a+by_i+cx_i+dx_iy_i+e_i$)

FV	GL	SQ	QM	F_c	pValor
Regressão	3	4582.56	1527.52	3802.86	
y	1	4240.9	4240.9	10558.00	$<2.2x10^{-16}$
x	1	182.9	182.9	455.37	$<2.2x10^{-16}$
xy	1	158.78	158.8	395.31	$<2.2x10^{-16}$
Resíduo	346	138.98	0,40		
Total	N-1=349	4721.54			

- Observe que o valor de F mudou, e que consequentemente a significância para cada fonte de variação muda também.

Isso não é muito estranho???

- A ANOVA anterior é obtida por um procedimento sequencial, quando parâmetros vão sendo adicionados nos modelos um a um.

- Nesta tabela avalia-se o ganho na predição pela inclusão de variáveis independentes.

- Logo após a adição de cada parâmetro pode-se responder:
  - y (largura) sozinha prediz a z (área foliar)?
  - A inclusão da variável x (comprimento) contribui significativamente para a predição de z (área foliar) após incluir y (Largura)?
  - A inclusão da interação xy (comprimento*Largura) contribui significativamente para a predição de z (área foliar) após incluir y (Largura) e x (comprimento)?

- Essas perguntas são importantes e permite que se obtenham modelos contendo apenas parâmetros que realmente contribuam para a predição.

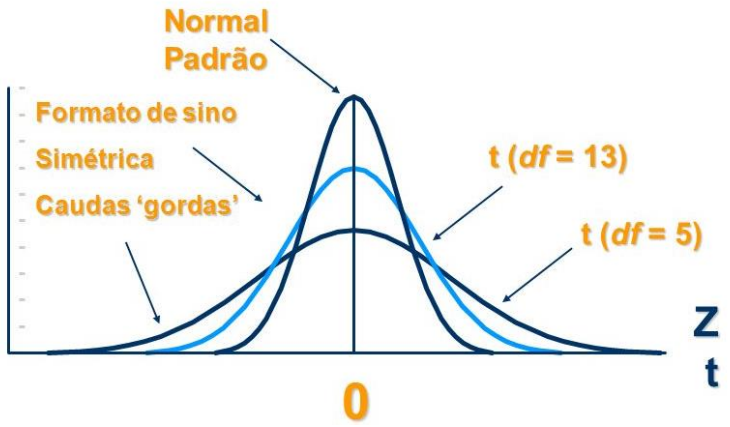
- Isso é importante para que se tenha modelos parcimoniosos. Para isso, pode-se também estudar a significância de cada parâmetro do modelo de regressão pelo teste t.

2.7 Teste t para significância dos coeficientes de regressão

- Por pressuposição do modelo matemático, admitimos que os coeficientes de regressão seguem distribuição normal.

- Logo, podemos testar a significância dos coeficientes por meio do teste t, para situações em que o tamanho amostral não é muito grande.

- É válido lembrar que o formato distribuição t depende do número de graus de liberdade, e que para a despadronização é preciso obter as variâncias associadas aos coeficientes.



- Como o número de graus de liberdade já é obtido na ANAVA (Glerro), é necessário obter as variâncias associadas à cada coeficiente de regressão, para posteriormente realizar-se o teste de hipótese (t).

2.7 Teste t para significância dos coeficientes de regressão

- Obtendo as variâncias dos estimadores dos coeficientes de regressão:
  - Sabe-se que $\widehat{B}$ é um estimador de B (Vetor com os coeficientes dos modelos de regressão. Logo, a estimativa destes coeficiente é influenciado por erros oriundos da variável dependente. Portanto, pode dizer que:

$$\hat{B} = (X'X)^{-1}X'Y$$

$\hat{B} = B + (X'X)^{-1}X'e$ Logo: $\hat{B} - B = (X'X)^{-1}X'e$

  - Podemos obter a variância associada a qualquer estimador por meio das propriedades da esperança matemática. $Var(\hat{\theta}) = E[(\theta - \hat{\theta})^2]$
    - $Var(\hat{B}) = E\left[(\hat{B} - B)'(\hat{B} - B)\right] = E[(X'X)^{-1}X'ee'X(X'X)^{-1}]$

$ee' = \sigma^2$ Logo:

$$Var(\hat{B}) = (X'X)^{-1}X'\sigma^2X(X'X)^{-1}$$

$$Var(\hat{B}) = (X'X)^{-1}\sigma^2$$

    - Onde σ^2=QMR obtido na ANAVA.

2.6 Teste t para significância dos coeficientes de regressão

- Logo, para o modelo em questão ($z_i=a+bx_i+cy_i+dx_iy_i+e_i$), se fizermos:

$$Var\left(\hat{B}\right) = (X'X)^{-1}\sigma^2$$

Onde:

σ^2=QMresiduo

$$X = \begin{bmatrix} 1 & 6,00 & 3,20 & 6,00*3,20 \\ 1 & 5,50 & 3,20 & 5,50*3,20 \\ 1 & 5,00 & 2,50 & 5,00*2,50 \\ 1 & 5,00 & 2,70 & 5,00*2,70 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 4,30 & 2,20 & 4,30*2,20 \end{bmatrix}$$

- teremos:

$$Var\left(\hat{B}\right) = (X'X)^{-1}QMres$$

$$= \begin{bmatrix} \hat{V}(\hat{a}) & C\hat{o}v(\hat{a},\hat{b}) & C\hat{o}v(\hat{a},\hat{c}) & C\hat{o}v(\hat{a},\hat{d}) \\ C\hat{o}v(\hat{a},\hat{b}) & \hat{V}(\hat{b}) & C\hat{o}v(\hat{b},\hat{c}) & C\hat{o}v(\hat{b},\hat{d}) \\ C\hat{o}v(\hat{a},\hat{c}) & C\hat{o}v(\hat{b},\hat{c}) & \hat{V}(\hat{c}) & C\hat{o}v(\hat{c},\hat{d}) \\ C\hat{o}v(\hat{a},\hat{d}) & C\hat{o}v(\hat{b},\hat{d}) & C\hat{o}v(\hat{c},\hat{d}) & \hat{V}(\hat{d}) \end{bmatrix}$$

2.7 Teste t para significância dos coeficientes de regressão

- De posse das variâncias associadas a cada parâmetro (diagonal principal da matriz $Var(\hat{B})$) podemos fazer o teste t:

$$t_{calc(a)} = \frac{\hat{a} - 0}{\sqrt{\hat{V}(\hat{a})}} \qquad t_{calc(b)} = \frac{\hat{b} - 0}{\sqrt{\hat{V}(\hat{b})}}$$

$$t_{calc(c)} = \frac{\hat{c} - 0}{\sqrt{\hat{V}(\hat{c})}} \qquad t_{calc(d)} = \frac{\hat{d} - 0}{\sqrt{\hat{V}(\hat{d})}}$$

- Após obter as estimativas de t_{calc}, pode-se fazer o teste de hipótese comparando esta estimativa com o valor de t crítico (tabelado). Para isso, precisa-se apenas considerar o nível de significância desejado e o número de graus de liberdade dos resíduos.

2.8 Avaliadores da qualidade do ajuste de regressão

- Coeficiente de determinação:

$$R^2 = \frac{SQregressão}{SQtotal}$$

  - Representa a proporção da variação total que é explicada pelo modelo de regressão.
  - Pode variar de 0 até 1, sendo que quanto mais próximo de 1 melhor é o ajuste.
  - Quanto maior é o número de parâmetros no modelo maior tende a ser o R^2.
  - Logo, não é indicado como critério para a comparação de modelos, pois geralmente leva à seleção de modelos mais complexos.

- Coeficiente de determinação ajustado

$$R^2{}_{aj} = \frac{R^2(n-p)-p}{n-p-1}$$

  - Este coeficiente leva em consideração o número de parâmetros do modelo estatístico.
  - É mais indicado para a seleção de modelos parcimoniosos

2.8 Avaliadores da qualidade do ajuste de regressão

- Critério de informatividade de Akaike (AIC):

$$AIC = -2 \log L\left(\widehat{\boldsymbol{\theta}}\right) + 2\,(p)$$

  - Onde:
    - L é o estimador de máxima verossimilhança do modelo de regressão.
    - p é o número de parâmetros no modelo de regressão.
  - Quanto menor sua estimativa, melhor é o modelo de regressão.
  - Leva em consideração o número de parâmetros do modelo.
  - É indicado para a seleção de modelos parcimoniosos.

- Critério de informatividade bayesiano (BIC)

$$BIC = -2 \log f\,(x_n|\boldsymbol{\theta}) + p \log n$$

  - Onde:
    - L é o estimador de máxima verossimilhança do modelo de regressão.
    - p é o número de parâmetros no modelo de regressão.
    - n é o número de observações (tamanho amostral)
  - Quanto menor sua estimativa, melhor é o modelo de regressão.
  - Leva em consideração o número de parâmetros do modelo.
  - É indicado para a seleção de modelos parcimoniosos

Stepwise

Obrigado