MAE 327
Planejamento e Pesquisa II

Profa. Júlia Maria Pavan Soler

pavan@ime.usp.br

IME/USP – 2º Semestre/2020

Modelos mais Gerais

Temos considerado modelos ANOVA para os seguintes delineamentos:

- **Estrutura de Tratamentos** (Fatores FIXOS sob Estudo):
  - Um único Fator em J níveis
  - Fatorial Cruzado, Fatorial 2^K (sem réplicas, com K elevado)
  - Fatorial Hierárquico
- **Estrutura das Unidades Experimentais**:
  - Delineamento Completamente Aleatorizado (DCA)
  - Delineamento Aleatorizado em Blocos Completos (DABC)

Fatores de efeitos Fixos, Dados balanceados, matrix de planejamento X com colunas ortogonais!

- Delineamentos Desbalanceados
- Modelos de Análise de Covariância (ANCOVA)
- Modelos mais Gerais: diferentes ajustes via Modelos de Regressão

Delineamentos Desbalanceados e Análise de Covariância (ANCOVA)

1. O que caracteriza o desbalanceamento em um Delineamento? Cite algumas fontes que geram dados desse tipo.
2. Quais dificuldades analíticas ocorrem na análise de dados desbalanceados?
3. Defina o modelo ANOVA e o modelo de regressão para um DCA Fatorial 3x2.
4. Como estão definidas as Somas de Quadrados Sequenciais (SQ Tipo I) em uma tabela de ANOVA para um DCA Fatorial 3x2?
5. Em um DCA com 3 fatores (A, B e C, ex., Fatorial 3x2x3), como estão definidas as Somas de Quadrados (SQ) Tipo I, II e III?
6. Qual é o modelo estrutural e distribucional da ANCOVA?
7. Por que na ANCOVA retas paralelas são adotadas aos tratamentos?
8. Compare o ajuste do efeito de um fator via blocagem e via covariável?
9. Como alternativa à ANCOVA poderia ser adotada a variável resposta corrigida como Y_{ij}-X_{ij}. Justifique esta afirmação.
10. Na ANCOVA, se a covariável X é influenciada pelos tratamentos, o efeito do tratamento sobre Y deve ser estimado a partir do modelo que usa como covariável $\tilde{X}$ e não X, em que $\tilde{X}$ é o resíduo do ajuste que prediz X dos tratamentos. Justifique essa recomendação.

Delineamentos com Um Único Fator

Dados desbalanceados: No caso de um único Fator as Somas de Quadrados satisfazem a ortogonalidade

SQTotal = SQTrat + SQRes

$$y_{ij} = \bar{y} + \left(\bar{y}_j - \bar{y}\right) + \left(y_{ij} - \bar{y}_j\right) \Rightarrow \sum_{ij}\left(y_{ij} - \bar{y}\right)^2 = \sum_{j} n_j \left(\bar{y}_j - \bar{y}\right)^2 + \sum_{ij}\left(y_{ij} - \bar{y}_j\right)^2$$

Dados "mtcars" do R:

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
…											
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

Variável resposta de desempenho do carro

Câmbio automático ou mecânico

am	0	1	Total
	19	13	32

Há interesse em avaliar o efeito de am em mpg (dados desbalanceados).

Dados Desbalanceados

As boas práticas no planejamento de experimentos recomendam delineamentos balanceados (número igual de réplicas em todos os grupos sob comparação).

Mas, por que Dados Desbalanceados ocorrem?

- Desbalanceamento não planejado: ocorrência de observações faltantes no banco de dados devido, por exemplo, à desistência de pacientes em participar do estudo, morte do animal, etc.

- Desbalanceamento planejado: obedecer na amostra a mesma distribuição dos níveis dos fatores presente na população. Por exemplo, na avaliação do efeito do tipo sanguíneo do paciente na resposta à vacina, a amostra pode guardar a mesma estrutura dos diferentes tipos sanguíneos presente na população.
- Desbalanceamento planejado: o grupo controle, composto de pacientes com a doença e tratados com Placebo, pode ter tamanho amostral menor que os grupos tratados devido, por exemplo, a condutas éticas.

Delineamentos com Um Único Fator

Dados "mtcar":

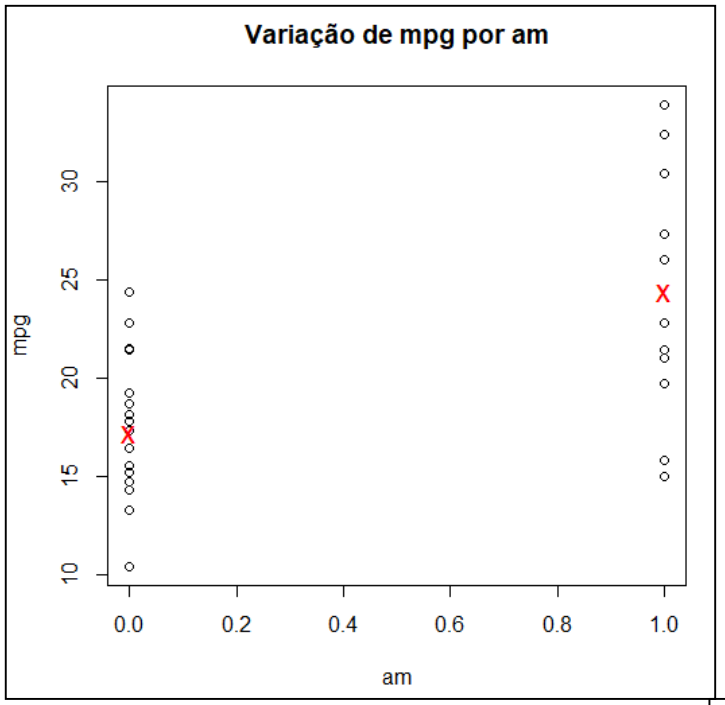


Tabela de ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
am	1	405.2	405.2	16.86	0.000285
Residuals	30	720.9	24.0		

Coeficientes (saída do "lm")

	Estimate	St.Error	tvalue	Pr(>|t|)
(Intercept)	17.147	1.125	15.247	1.13e-15
am	7.245	1.764	4.106	0.000285

$$\hat{\mu}_0 = 17.147 \quad \hat{\mu}_1 = 24.392$$

Propriedades da ortogonalidade do modelo (as colunas da matriz X são ortogonais)

$$t_{30}^2 = (4.106)^2 = 16.86 = F_{1;30}$$

$$\Rightarrow \sum_{ij} (y_{ij} - \bar{y})^2 = \sum_{j} n_j (\bar{y}_j - \bar{y})^2 + \sum_{ij} (y_{ij} - \bar{y}_j)^2$$

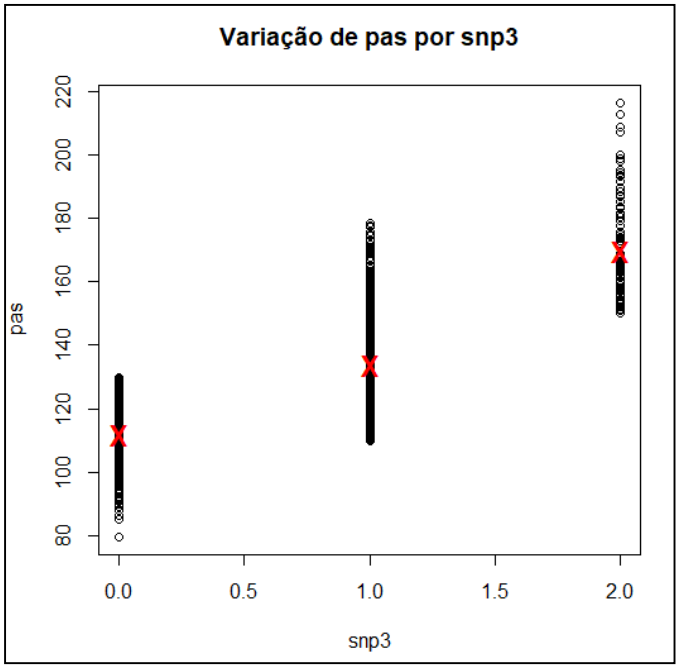
$$SQTotal = 405.2 + 720.9$$

Delineamentos com Um Único Fator

Dados de Pressão Arterial Sistólica (pas) de acordo com o efeito do Marcador Molecular snp3 (0=aa, 1=Aa, 2=AA).

Banco de Dados pas

	sex	idade	pas	snp1	snp2	snp3
1	1	69	117.0	0	2	1
2	2	63	137.7	1	2	1
3	1	62	135.7	1	1	1
4	2	59	144.7	2	2	1
5	2	45	136.7	1	1	1
6	1	34	119.0	0	2	1
7	2	28	107.0	0	0	0
8	2	43	120.3	1	1	0
…						
1679	2	43	123.0	1	1	0
1680	2	22	103.7	0	0	0



Snp3	0	1	2
	695	884	101

	Df	Sum Sq	MeanSq	F value	Pr(>F)
factor(snp3)	**2**	388138	194069	1327.1	<2.2e-16
Residuals	1677	245239	146		

DCA com Um Fator em 3 níveis, dados desbalanceados.

Há efeito significante do fator snp3 no desempenho do carro?

Análise com 2 graus de Liberdade.

	Estimate	SError	tvalue	Pr(>|t|)
(Intercept)	111.7458	0.4587	243.61	<2e-16
factor(snp3)1	22.1723	0.6131	36.17	<2e-16
factor(snp3)2	57.9513	1.2877	45.00	<2e-16

Delineamentos com Um Único Fator

Dados de Pressão Arterial Sistólica (pas) de acordo com o efeito do Marcador Molecular snp3 (0=aa, 1=Aa, 2=AA).

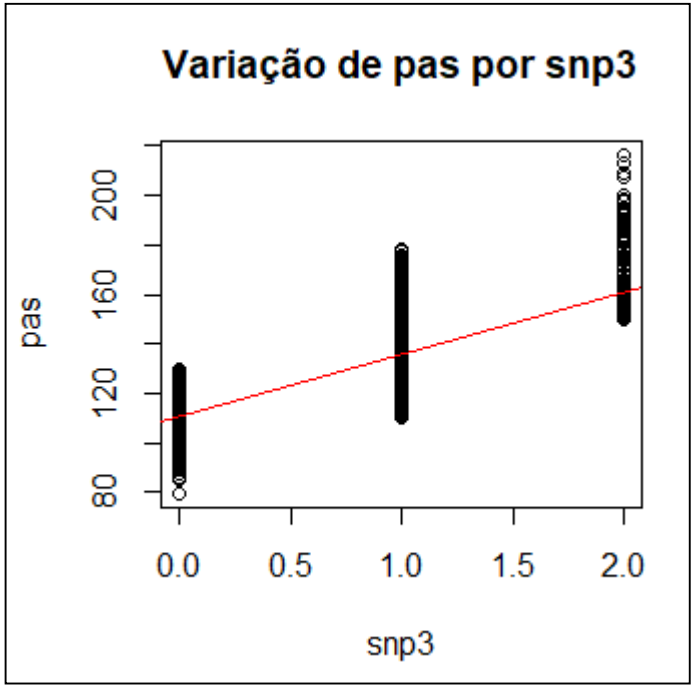
Banco de Dados pas

	sex	idade	pas	snp1	snp2	snp3
1	1	69	117.0	0	2	1
2	2	63	137.7	1	2	1
3	1	62	135.7	1	1	1
4	2	59	144.7	2	2	1
5	2	45	136.7	1	1	1
6	1	34	119.0	0	2	1
7	2	28	107.0	0	0	0
8	2	43	120.3	1	1	0
…						
1679	2	43	123.0	1	1	0
1680	2	22	103.7	0	0	0

Snp3	0	1	2
	695	884	101

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
snp3	**1**	376468	376468	2459	<2e-16
Resid	1678	256909	153		

	Estimate	Std. Error	t value	Pr(>|t|)
Mi	**110.5117**	0.4476	246.92	<2e-16
snp3	**25.3468**	0.5112	49.59	<2e-16



DCA com Um Fator em 3 níveis, dados desbalanceados.

Níveis do fator: número de alelos A

Há efeito linear significante do fator snp3?

Delineamento Fatorial Desbalanceado

Delineamento Fatorial 3x2: dados de Pressão Arterial Sistólica (pas) de acordo com o efeito do Marcador Molecular snp3 (0=aa, 1=Aa, 2=AA) e do Sexo do paciente.

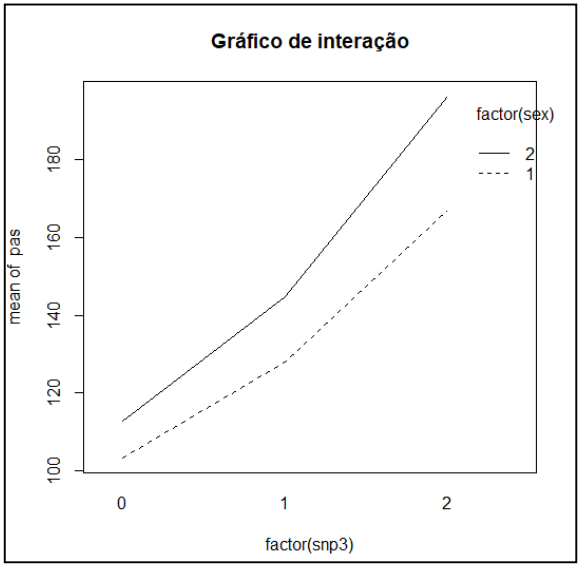


Tabela de ANOVA:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sex	1	14911	14911	**143.095**	< 2.2e-16
factor(snp3)	2	439466	219733	2108.656	< 2.2e-16
sex:factor(snp3)	2	4560	2280	21.882	4.158e-10
Residuals	1674	174439	104		

	snp3		
sex	0	1	2
1	73	567	91
2	622	317	10

Coeficientes:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	94.086	2.424	38.809	< 2e-16
sex	9.319	1.263	**7.379**	2.50e-13
factor(snp3)1	17.009	2.635	6.456	1.40e-10
factor(snp3)2	43.380	4.569	9.494	< 2e-16
sex:factor(snp3)1	7.479	1.452	5.152	2.88e-07
sex:factor(snp3)2	20.007	3.628	5.515	4.03e-08

Note que, para o fator sexo, a estatística $t^2 \neq F$
Ainda, no caso desbalanceado, note que ao mudar a ordem de entrada dos fatores as SQ na tabela de ANOVA mudam, veja a seguir:

Delineamento Fatorial

Fatorial 3x2: Dados pas de acordo com sn3 (0=aa, 1=Aa, 2=AA) e Sexo.

SQ Sequencial (TipoI): adotada no R

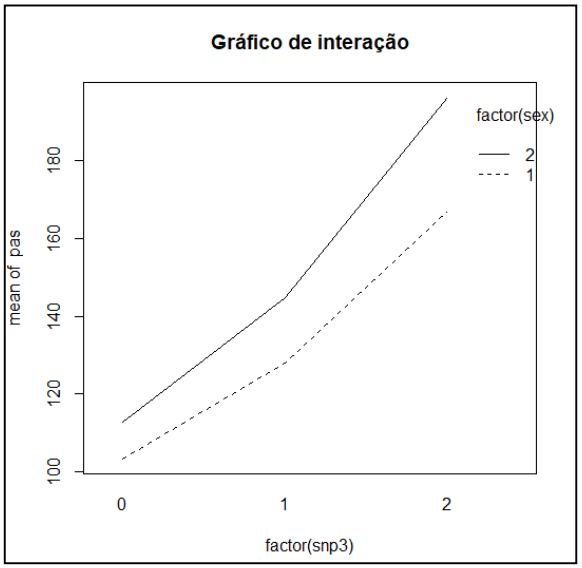


Tabela de ANOVA:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(snp3)	2	388138	194069	1862.373	< 2.2e-16
sex	1	66239	66239	635.659	< 2.2e-16
factor(snp3):sex	2	4560	2280	21.882	4.158e-10
Residuals	1674	174439	104		

Coeficientes:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	94.086	2.424	38.809	< 2e-16
sex	9.319	1.263	7.379	2.50e-13
factor(snp3)1	17.009	2.635	6.456	1.40e-10
factor(snp3)2	43.380	4.569	9.494	< 2e-16
sex:factor(snp3)1	7.479	1.452	5.152	2.88e-07
sex:factor(snp3)2	20.007	3.628	5.515	4.03e-08

	snp3		
sex	0	1	2
1	73	567	91
2	622	317	10

Nos delineamentos desbalanceados a ordem da entrada dos fatores na Tabela de ANOVA muda o valor das correspondentes SQ. Os delineamentos fatoriais desbalanceados não são ortogonais e há diferentes formulações das SQ. Note que, SQ(snp3) no modelo (1,sex, snp3,sex:snp3) é ≠ da do modelo (1,snp3,sex,sex:snp3)

Delineamento Fatorial Desbalanceado
Soma de Quadrados Sequencial

Tabela de ANOVA:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sex	1	**14911**	14911	143.095	< 2.2e-16
factor(snp3)	2	**439466**	219733	2108.656	< 2.2e-16
sex:factor(snp3)	2	4560	2280	21.882	4.158e-10
Residuals	1674	174439	104		

Soma de Quadrados Sequencial (SQ Tipo I)

SQ(X2|X1): é a SQ de X2 dado que X1 está no modelo. É obtida da comparação de dois modelos, o reduzido (só com X1) e o complete (com X1 e X2)

Modelo Completo:

SQMod(X1,X2) = SQMod(X1)+ SQMod(X2|X1)

= 14911 + 439466 = 454377

Modelo Reduzido (sem X2):

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sex	1	**14911**	14911.2	40.456	2.586e-10
Residuals	1678	618466	368.6		

SQMod(X2|X1) = SQRes(X1) - SQRes(X1,X2)

= 618466 - (174439+4560) = **439466**

É a redução na SQRes devido ao efeito do fator snp3

Análise de Covariância

Variáveis Preditoras Qualitativas e Quantitativas

$$y_{ij} = \mu + \tau_j + \varepsilon_{ij}; \quad \sum_{j=1}^{J} \tau_j = 0; \quad \varepsilon_{ij} \overset{iid}{\sim} N\left(0; \sigma^2\right)$$

Efeito de Tratamento ajustado pela covariável X:

$$y_{ij} = \mu + \beta\, X_{ij} + \tau_j + \varepsilon_{ij}; \quad \sum_{j=1}^{J} \tau_j = 0; \quad \varepsilon_{ij} \overset{iid}{\sim} N\left(0; \sigma^2\right)$$

$$y_{ij} = \mu + \beta\left(X_{ij} - \bar{X}\right) + \tau_j + \varepsilon_{ij}$$

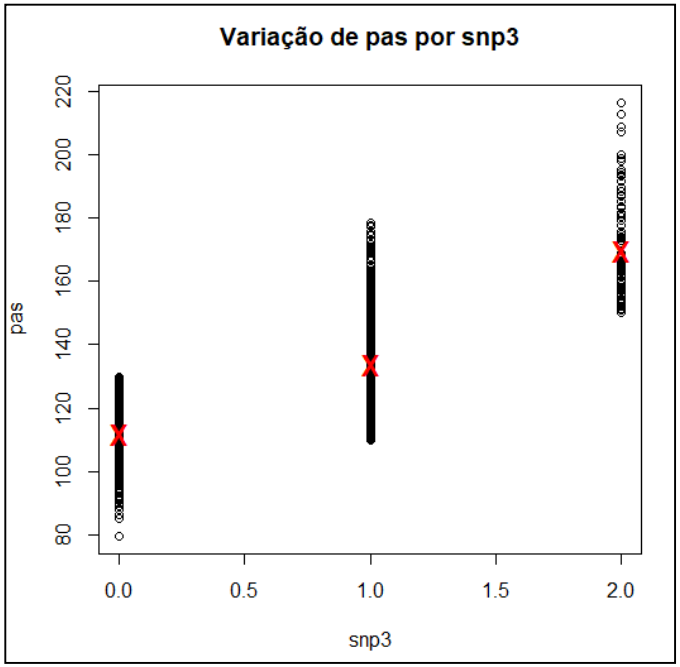
Suposições: Normalidade, homocedasticidade, independência, relação linear entre Y e X, retas paralelas (não há efeito de tratamento sobre X)

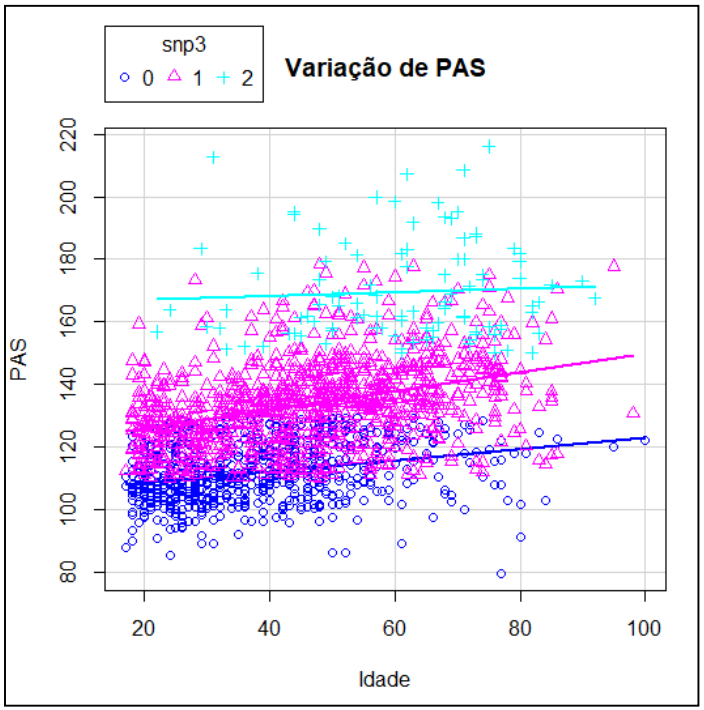
$\Rightarrow$ Ajuste via Moodelos de Regressão com covariáveis e variáveis indicadoras do tratamento (“dummy”)

Análise de Covariância

Efeito de Tratamento Ajustado por Covariáveis

Motivação: **Efeito do snp3 em pas ajustado pela idade dos pacientes**





Delineamentos em Blocos (DABC): controlam o efeito de uma fonte conhecida de erro
Análise de Covariância: usada para melhorar a precisão de um experimento. A covariável afeta Y mas não afeta o fator. É uma fonte de erro que não pôde ser controlada via blocagem mas que foi observada durante o experimento e pode ser usada para ajustar o efeito do fator.

Análise de Covariância

Efeito de Tratamento Ajustado por Covariáveis

Motivação: **Efeito do snp3 em pas ajustado pela idade dos pacientes**

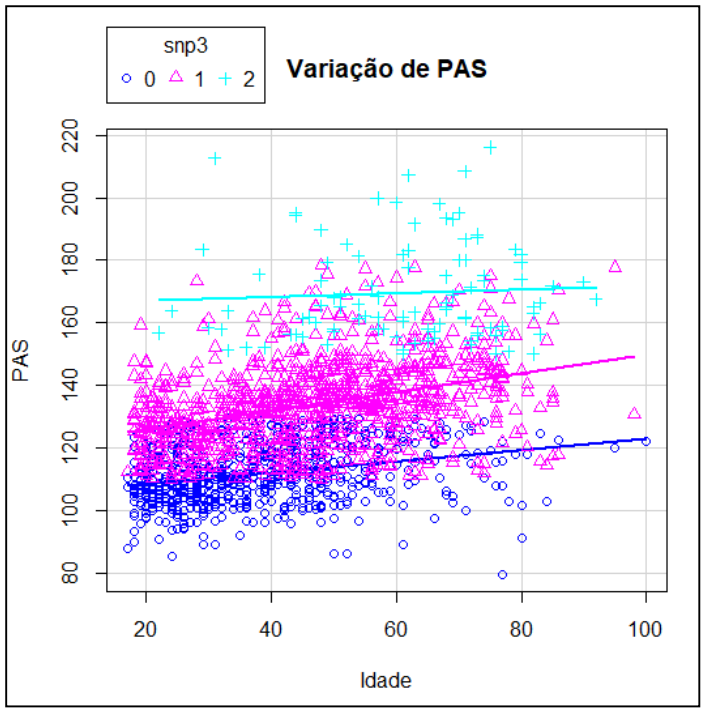


Tabela de ANOVA

	Df	Sum Sq	MeanSq	F value	Pr(>F)
idade	1	131061	131061	997.3	< 2.2e-16
factor(snp3)	2	282062	141031	1073.2	< 2.2e-16
Residuals	1676	220254	131		

Coeficientes:

	Estimate	Std.Error	t value	Pr(>|t|)
Intercept	102.42527	0.80375	127.44	<2e-16
idade	0.24212	0.01756	13.79	<2e-16
fact(snp3)1	20.18813	0.59871	33.72	<2e-16
fact(snp3)2	52.48792	1.28345	40.90	<2e-16

Sem o ajuste por covariável ⇒

	Estimate	SError	tvalue	Pr(>|t|)
(Intercept)	111.7458	0.4587	243.61	<2e-16
factor(snp3)1	22.1723	0.6131	36.17	<2e-16
factor(snp3)2	57.9513	1.2877	45.00	<2e-16

Análise de Covariância

Efeito de Tratamento Ajustado por Covariáveis

Motivação: **Efeito do snp3 em pas ajustado pela idade dos pacientes**

Tabela de ANOVA

Modelo(X1,X2)

	Df	Sum Sq	MeanSq	F value	Pr(>F)
idade	1	131061	131061	997.3	< 2.2e-16
factor(snp3)	2	**282062**	141031	1073.2	< 2.2e-16
Residuals	1676	220254	131		

Modelo(X1)

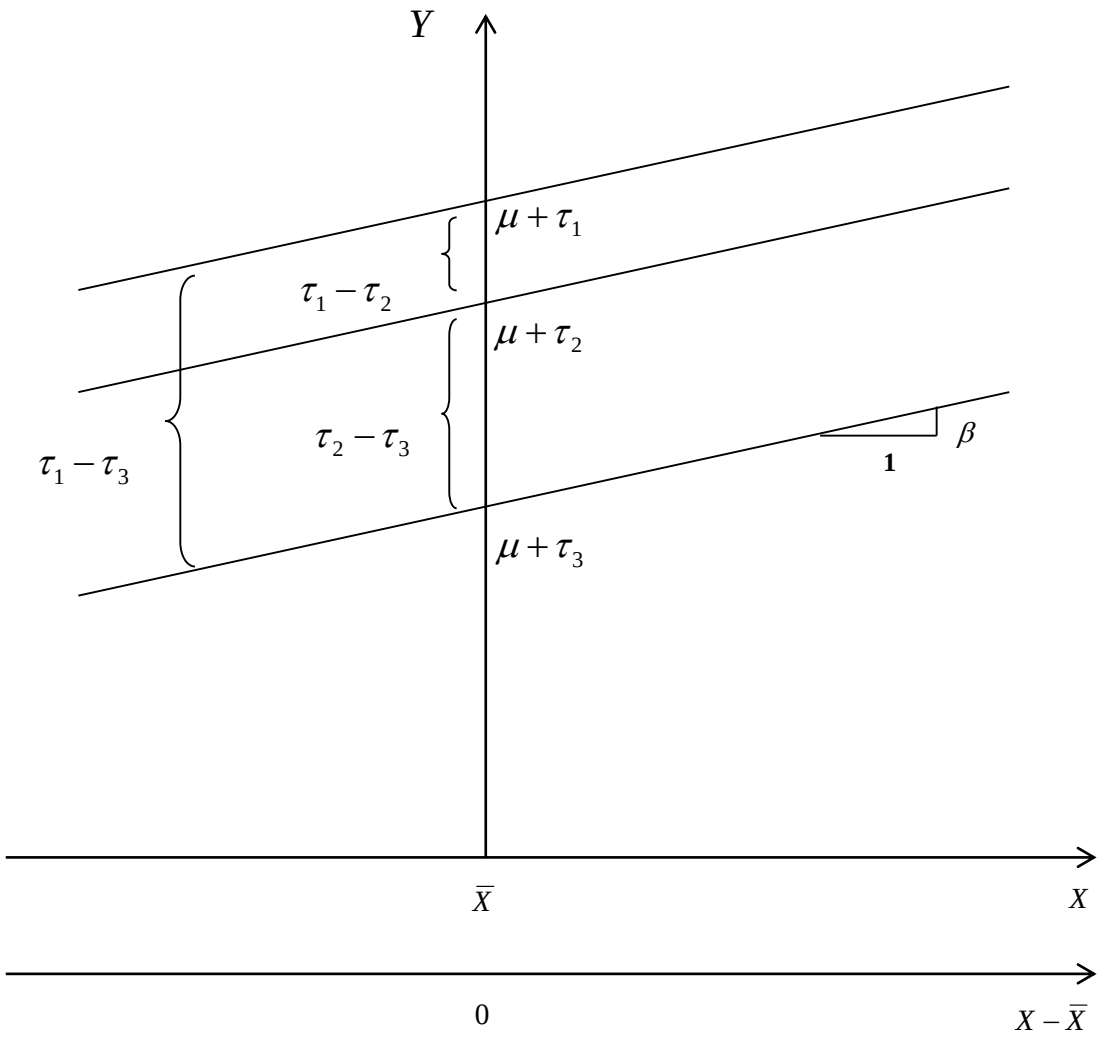
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
idade	1	131061	131061	437.81	< 2.2e-16
Residuals	1678	502316	299		

SQMod(X1,X2) = SQMod(X1)+ SQMod(X2|X1)
= 131061 + 282062 = 413123

SQMod(X2|X1) = SQRes(X1) - SQRes(X1,X2)
= 502316 - 220254 = **282062**

É a redução na SQRes devido ao efeito do fator

ANCOVA – 1Fator (3 níveis) e 1 Covariável



T1

T2

T3

Análise de covariância: supõe retas paralelas aos tratamentos.

Efeito de Tratamento não depende da covariável. A diferença na resposta esperada ao tratamento j e j' para indivíduos com o mesmo valor da covariável é:

$$\left(\mu + \tau_j + \beta X_{ij}\right) - \left(\mu + \tau_{j'} + \beta X_{ij}\right) = \tau_j - \tau_{j'}$$

Análise de Covariância

$$y_{ij} = \mu + \tau_j + \beta\, X_{ij} + \varepsilon_{ij}; \quad \sum_{j=1}^{J} \tau_j = 0$$

Defina as seguintes estatísticas:

$$SQTot_Y = \sum_{j=1}^{J}\sum_{i=1}^{r}\left(y_{ij} - \bar{y}\right)^2$$

$$SQTrat_Y = \sum_{j=1}^{J} r\left(\bar{y}_{.j} - \bar{y}\right)^2$$

$$SQ\,\mathrm{Re}\,s_Y = \sum_{j=1}^{J}\sum_{i=1}^{r}\left(y_{ij} - \bar{y}_{.j}\right)^2$$

$$SQTot_X = \sum_{j=1}^{J}\sum_{i=1}^{r}\left(x_{ij} - \bar{x}\right)^2$$

$$SQTrat_X = \sum_{j=1}^{J} r\left(\bar{x}_{.j} - \bar{x}\right)^2$$

$$SQ\,\mathrm{Re}\,s_X = \sum_{j=1}^{J}\sum_{i=1}^{r}\left(x_{ij} - \bar{x}_{.j}\right)^2$$

$$SQTotal_{XY} = \sum_{j=1}^{J}\sum_{i=1}^{r}\left(x_{ij} - \bar{x}\right)\left(y_{ij} - \bar{y}\right)$$

$$SQTrat_{XY} = \sum_{j=1}^{J} r\left(\bar{x}_{.j} - \bar{x}\right)\left(\bar{y}_{.j} - \bar{y}\right)$$

$$SQ\,\mathrm{Re}\,s_{XY} = \sum_{j=1}^{J}\sum_{i=1}^{r}\left(x_{ij} - \bar{x}_{.j}\right)\left(y_{ij} - \bar{y}_{.j}\right)$$

SQ para o efeito do Trat em Y

SQ para o efeito do Trat em X

Análise de Covariância

Modelo completo

$$y_{ij} = \mu + \tau_j + \beta\, X_{ij} + \varepsilon_{ij}; \quad \sum_{j=1}^{J} \tau_j = 0$$

$$\hat{\mu} = \bar{y}; \quad \hat{\tau}_j = \left(\bar{y}_{.j} - \bar{y}\right) - \hat{\beta}\, X_{ij}; \quad \hat{\beta} = SQ\,\mathrm{Res}_{XY} / SQ\,\mathrm{Res}_X$$

$$SQ\,\mathrm{Res}(X1, X2) = SQ\,\mathrm{Res}_Y - \left(SQ\,\mathrm{Res}_{XY}\right)^2 / SQ\,\mathrm{Res}_X \qquad QM\,\mathrm{Res} = SQ\,\mathrm{Res} / \left[J(r-1)-1\right]$$

Modelo reduzido

$$y_{ij} = \mu + \beta\, X_{ij} + \varepsilon_{ij}$$

$$\hat{\mu} = \bar{y}; \qquad \hat{\beta} = SQTrat_{XY} / SQTrat_X$$

$$SQ\,\mathrm{Res}(X1) = SQTrat_Y - \left(SQTrat_{XY}\right)^2 / SQTrat_X$$

$$SQ\,\mathrm{Res}(X1) - SQ\,\mathrm{Res}(X1, X2) = SQ(X2 \mid X1)$$

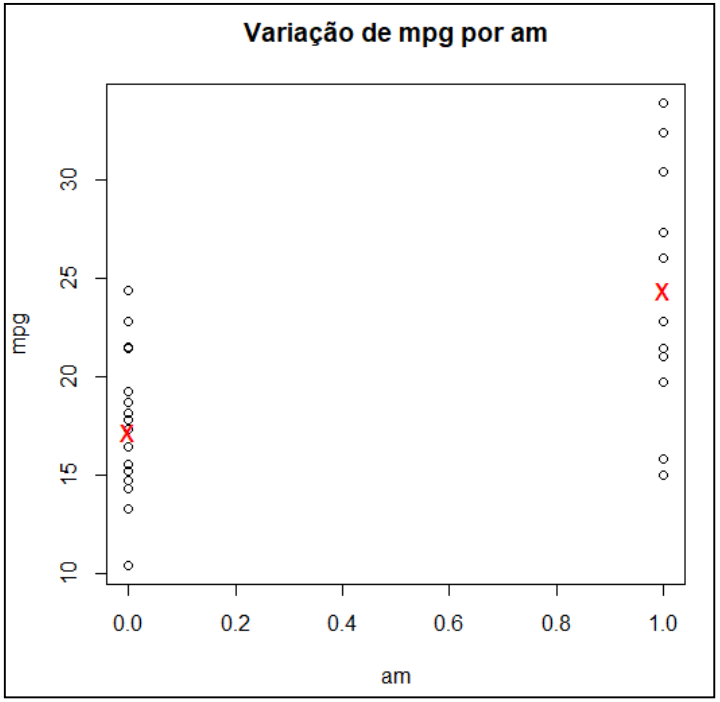
é a redução na SQResíduo devido ao efeito de tratamento

Testar o Efeito de Tratamento Ajustado pela Covariável:

$$H_0 : \tau_j = 0 \qquad F_{X2|X1} = \frac{\left(SQ\,\mathrm{Res}(X1) - SQ\,\mathrm{Res}(X1, X2)\right)/(J-1)}{QM\,\mathrm{Res}(X1, X2)} \sim F_{(J-1);[J(r-1)-1]}$$

Análise de Covariância

Efeito de am em mpg



Efeito de am em mpg ajustado pela covariável hp

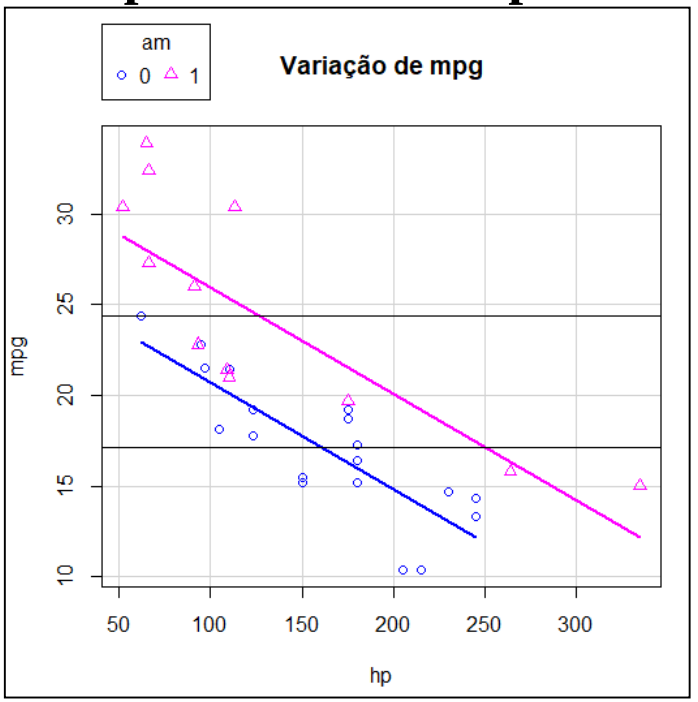


Tabela de ANOVA

	Df	SumSq	MeanSq	Fvalue	Pr(>F)
am	1	405.2	405.2	16.86	0.000285
Residuals	30	720.9	24.0		

Coeficientes (saída do "lm")

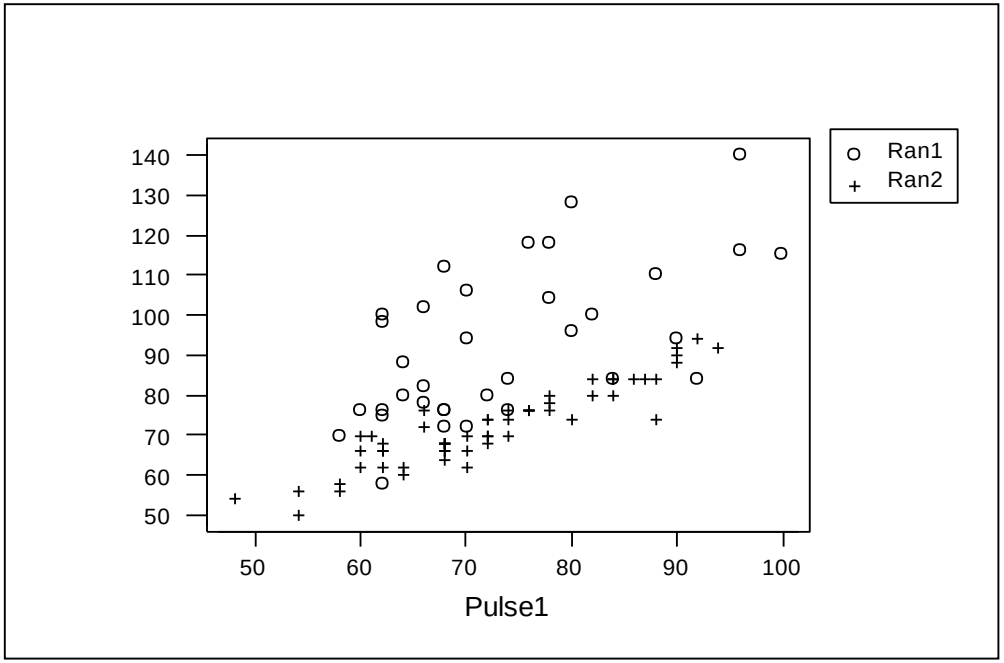
	Estimate	SError	tvalue	Pr(>|t|)
Intercept	17.147	1.125	15.247	1.13e-15
am	7.245	1.764	4.106	0.000285

	Df	SumSq	MeanSq	Fvalue	Pr(>F)
hp	1	678.37	678.37	80.153	7.627e-10
am	1	202.24	202.24	23.895	3.460e-05
Resid	29	245.44	8.46		

	Estimate	SError	tvalue	Pr(>|t|)
Interc	26.5849	1.4250	18.655	< 2e-16
hp	-0.0588	0.0078	-7.495	2.92e-08
am	5.2770	1.0795	4.888	3.46e-05

Análise de Covariância

Motivação: **Verifique se há efeito da corrida na Pulsação dos estudantes (Pulse 2), ajustando os dados pela Pulsação Inicial dos estudantes antes da corrida (Pulse 1)**

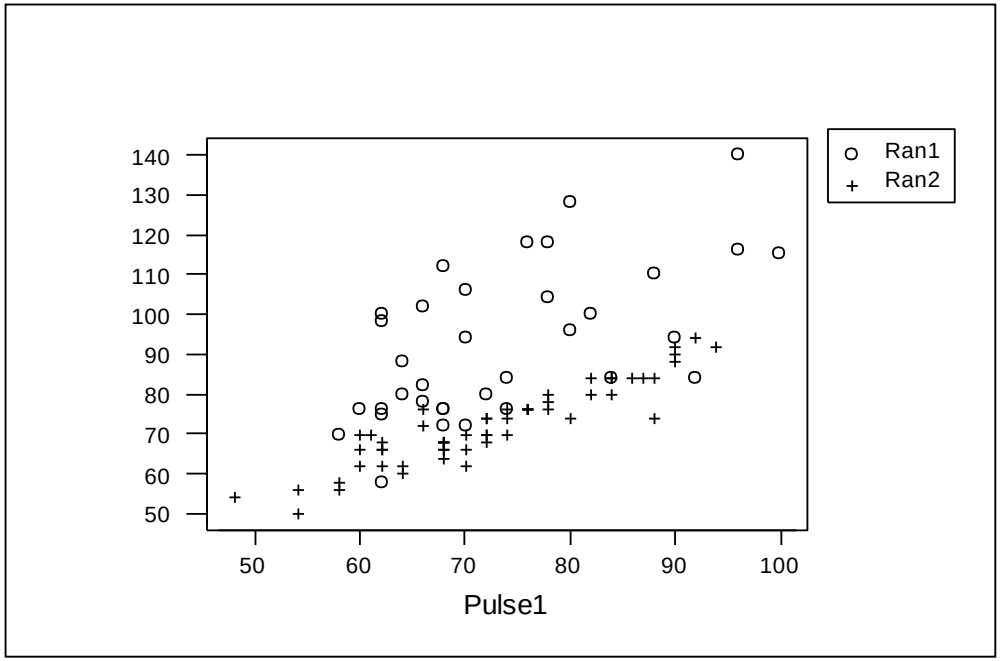


Delineamentos em Blocos: usados para controlar o efeito de uma fonte conhecida de erro
Análise de Covariância: usada para melhorar a precisão de um experimento. A covariável é uma fonte de erro que não pode ser controlada mas que pode ser observada.

Análise de Covariância

Arquivo Pulse

	P2	Ran	P1
[1,]	88	1	64
[2,]	70	1	58
[3,]	76	1	62
[4,]	78	1	66
[5,]	80	1	64
[6,]	84	1	74
[7,]	84	1	84
[8,]	72	1	68
[9,]	75	1	62
[10,]	118	1	76
[11,]	94	1	90
[12,]	96	1	80
[13,]	84	1	92
[14,]	76	1	68
…			
[91,]	84	2	86
[92,]	76	2	76



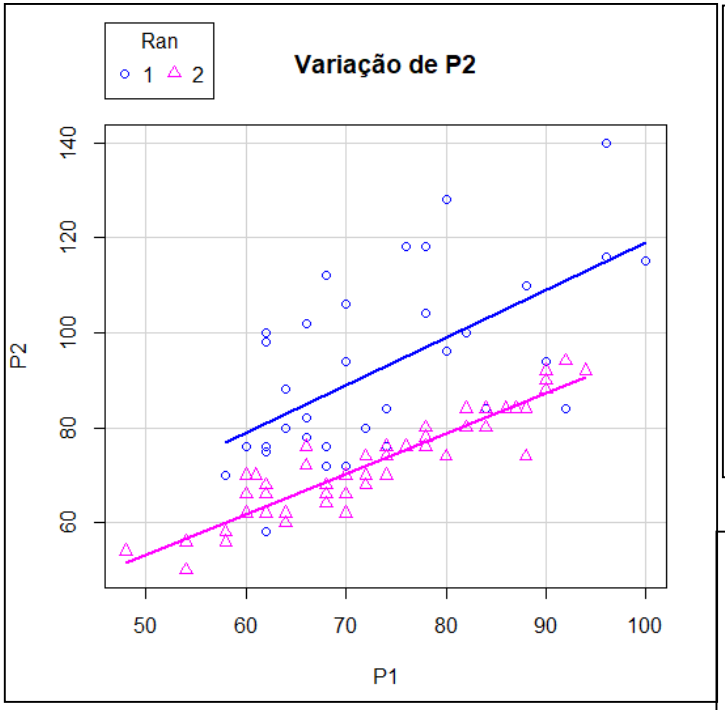
Estatísticas descritivas de P1 (covariável)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
48.00	64.00	71.00	72.87	80.00	100.00	11.01

Análise de Covariância: usada para melhorar a precisão do experimento. A covariável é uma fonte de erro que não foi controlada por planejamento (blocagem) mas que foi observada e usada no ajuste de P2.

Delineamentos em Blocos: usados para controlar o efeito de uma fonte conhecida de erro

ANCOVA-Incorporando Informação Adicional



Efeito de Ran sobre P2

Tabela de ANOVA:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Ran	1	8846.9	8846.9	44.875	1.768e-09
Residuals	90	17743.1	197.1		

Coeficientes:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	112.713	5.098	22.109	< 2e-16
Ran	-20.198	3.015	-6.699	1.77e-09

Efeito de Ran sobre P2 ajustado por P1

Tabela de ANOVA:

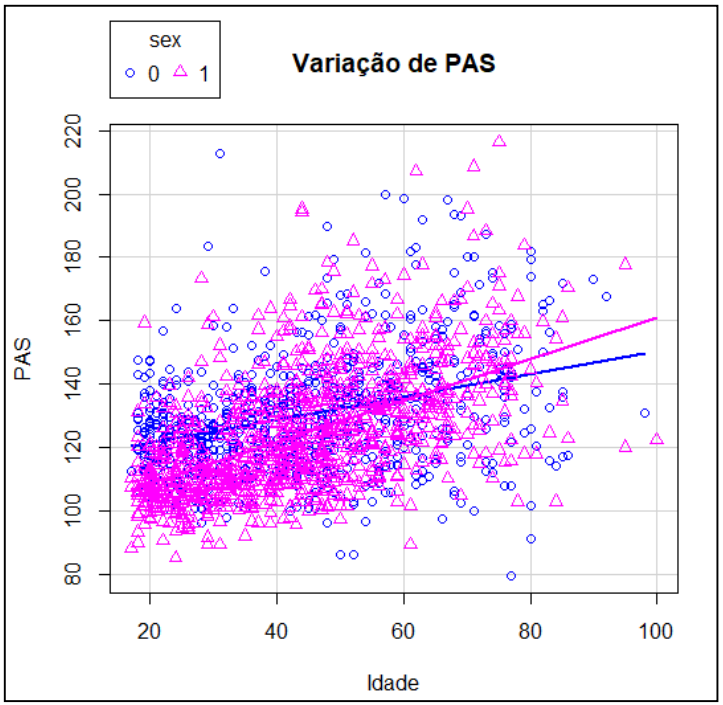
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
P1	1	10096.1	10096.1	104.656	< 2.2e-16
Ran	1	7908.0	7908.0	81.974	2.905e-14
Residuals	89	8585.8	96.5		

Coeficientes:

	Estimate	Std.Error	tvalue	Pr(>|t|)
(Intercept)	44.47919	7.85915	5.660	1.83e-07
P1	0.91247	0.09366	9.743	1.09e-15
Ran	-19.12274	2.11209	-9.054	2.90e-14

ANCOVA-Incorporando Informação Adicional

Dados de Pressão (pas):




```
Coefficients:Modelo sem ajuste
            Estimate Std.Error tvalue Pr(>|t|)
Intercept 136.3003   1.5509  87.886  < 2e-16
sex        -6.0092   0.9448  -6.361 2.59e-10

Tabela de ANOVA _ ANCOVA
          Df Sum Sq Mean Sq F value    Pr(>F)
idade      1 131061  131061 458.037 < 2.2e-16
sex        1  12629   12629  44.135 4.131e-11
idade:sex  1  10122   10122  35.374 3.305e-09
Residuals 1676 479565     286

Coeficientes:
 Estimate Std.Error tvalue  Pr(>|t|)
Mi       114.17259   1.71636 66.520  < 2e-16
idade      0.36071   0.03577 10.085  < 2e-16
sex      -18.38481   2.31587 -7.939 3.71e-15
idade:sex  0.29041   0.04883  5.948 3.30e-09
```

$$\hat{E}(y|_{Idade,Sex}) = 114.17 + 0.36X_{idade} - 18.38X_{sex} + 0.29\left(X_{idade} * X_{sex}\right)$$

$$\hat{E}(y|_{M,Idade}) = 114.17 + 0.36X_{idade}$$

$$\hat{E}(y|_{F,Idade}) = \left(114.17 - 18.38\right) + \left(0.36 + 0.29\right)X_{idade} = 95.79 + 0.65X_{idade}$$

Modelos Mais Gerais

Variáveis Preditoras Quantitativas e Qualitativas

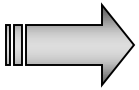
X_1: ***var. quantitativa***
(var. concomitante, covariável)

X_2: ***var. qualitativa***
(var. categorizada, fator)

$$y_i = \mu + \beta_2 X_{2i} + \varepsilon_i$$

$$y_i = \mu + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$
$$+ \beta \left(X_{ij} - \bar{X} \right)$$

$$y_i = \mu + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_{12} X_{1i} * X_{2i} + \varepsilon_i$$



"Análise de Regressão com Variável Categorizada

Modelos Mais Gerais

Variáveis Preditoras Quantitativas e Qualitativas

Dados de Pressão (pas):

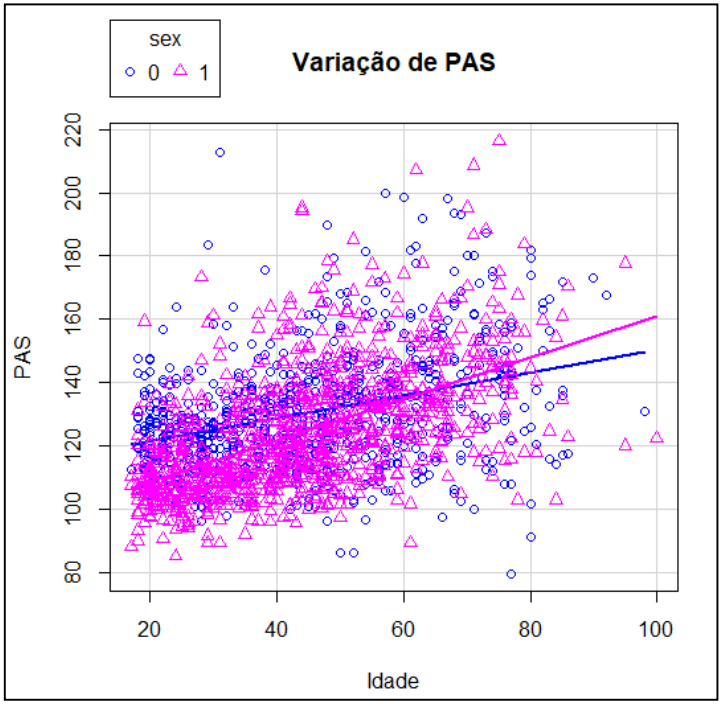


Tabela de ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
idade	1	131061	131061	458.037	< 2.2e-16
sex	1	12629	12629	44.135	4.131e-11
idade:sex	1	10122	10122	35.374	3.305e-09
Residuals	1676	479565	286		

Coeficientes:

	Estimate	Std.Error	tvalue	Pr(>|t|)
Mi	114.17259	1.71636	66.520	< 2e-16
idade	0.36071	0.03577	10.085	< 2e-16
sex	-18.38481	2.31587	-7.939	3.71e-15
idade:sex	0.29041	0.04883	5.948	3.30e-09

Ajuste também a idade ao quadrado.

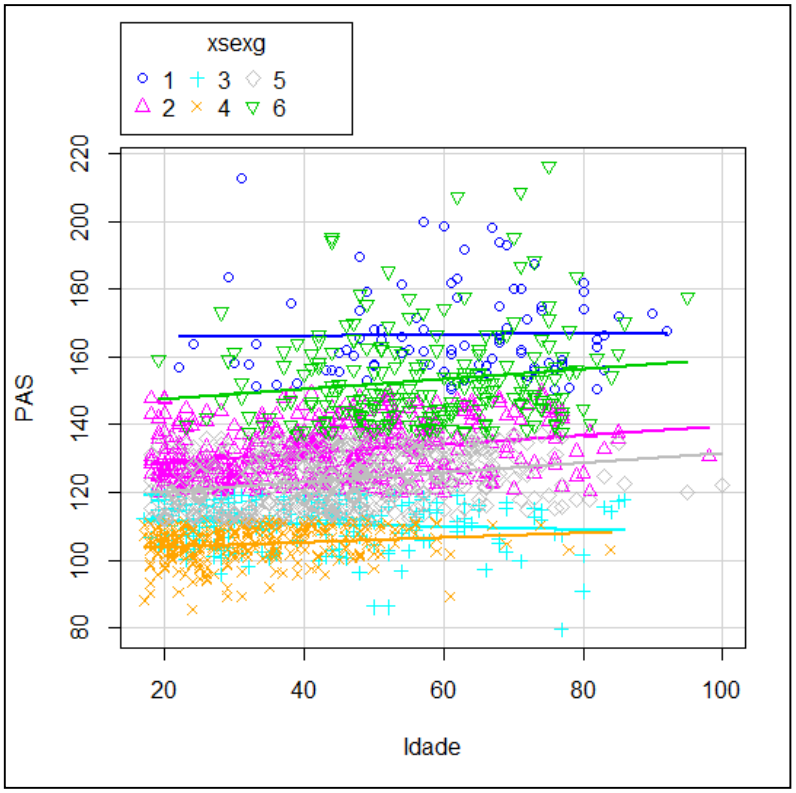
$$\hat{E}(y|_{Idade,Sex}) = 114.17 + 0.36X_{idade} - 18.38X_{sex} + 0.29\left(X_{idade} * X_{sex}\right)$$

$$\hat{E}(y|_{M,Idade}) = 114.17 + 0.36X_{idade}$$

$$\hat{E}(y|_{F,Idade}) = \left(114.17 - 18.38\right) + \left(0.36 + 0.29\right)X_{idade} = 95.79 + 0.65X_{idade}$$

Modelos Mais Gerais

Variáveis Preditoras Quantitativas e Qualitativas



xsexg			
Sex | SNP2	0	1	2
M=0	1	2	3
F=1	4	5	6

	Estimate	Std.Error	t value	Pr(>|t|)
(Intercept)	185.98498	4.37728	42.489	< 2e-16
idade	0.61814	0.08188	7.549	7.17e-14
sex	-40.83467	2.44400	-16.708	< 2e-16
snp2	-49.07538	3.25761	-15.065	< 2e-16
idade:sex	-0.32073	0.04976	-6.445	1.51e-10
idade:snp2	-0.50121	0.06260	-8.006	2.19e-15
sex:snp2	32.23153	1.94018	16.613	< 2e-16
idade:sex:snp2	0.32727	0.03831	8.542	< 2e-16

$$\hat{E}(y|_{Idade,Sex,G2}) = 185.98 + 0.61X_{idade} - 40.83X_{sex} - 49.07X_{G2}$$

$$-0.32\left(X_{idade} * X_{sex}\right) - 0.50\left(X_{idade} * X_{G2}\right) + 32\left(X_{sex} * X_{G2}\right) + 0.32\left(X_{idade} * X_{sex} * X_{G2}\right)$$

Tabela de ANOVA – Notação Matricial

Fonte de Variação	Número de graus de liberdade	Soma de Quadrados Sequencial (SQ Tipo I)
Tratamento	*J-1*	$Y'\left[P-\frac{1}{n}1_{n\times n}\right]Y=\hat{\beta}'X'Y-n\bar{Y}^2$
Resíduo	*n-J*	$Y'\left[I_n-P\right]Y=Y'Y-\hat{\beta}'X'Y$
TOTAL	*n-1*	$Y'\left[I_n-\frac{1}{n}1_{n\times n}\right]Y=Y'Y-n\bar{Y}^2$

$P=X\left(X'X\right)^{-1}X'$: projetor

$1_{n\times n}=1_n1_n'$: matriz de 1's

SQ Sequencial (Tipo I) acompanha a ordem de partição do modelo.

Fatorial 3x2 - Desbalanceado

Codificação dos dados para a parametrização casela de Referência
Fatorial 3x2: Fator A (j=1,2,3) Fator B (k=1,2)

Cada combinação dos níveis dos fatores há um número específico de replicas.

j	k	Y	Mi	X1	X2	X3	X1*X3	X2*X3
1	1	-	1	0	0	0	0	0
2	1	-	1	1	0	0	0	0
3	1	-	1	0	1	0	0	0
1	2	-	1	0	0	1	0	0
2	2	-	1	1	0	1	1	0
3	2	-	1	0	1	1	0	1

Matriz X

Modelo estrutural da ANOVA

$$y_{ijk} = \begin{cases} \mu_0 & \text{se} \quad j = k = 1 \\ \mu_0 + \tau_j + \beta_k + \gamma_{jk} + e_{ijk} & \text{se} \quad j = 2,3; k = 2 \end{cases}$$

Modelo estrutural de Regressão

$$y_{ijk} = \mu_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 (X_1 * X_3) + \beta_5 (X_2 * X_3) + e_{ijk}$$

Fatorial 3x2 - Desbalanceado

Codificação dos dados para a parametrização casela de Referência
Fatorial 3x2: Fator A (j=1,2,3) Fator B (k=1,2)

Cada combinação dos níveis dos fatores há um número específico de replicas.

j	k	Y	Mi	X1	X2	X3	X1*X3	X2*X3
1	1	-	1	0	0	0	0	0
2	1	-	1	1	0	0	0	0
3	1	-	1	0	1	0	0	0
1	2	-	1	0	0	1	0	0
2	2	-	1	1	0	1	1	0
3	2	-	1	0	1	1	0	1

Matriz X

Coeficientes de regressão associados às variáveis em X:

Média de referência j=1=k

respectivos desvios para efeitos principais e de interação dos fatores A (em j=2 e j=3) e B (em k=2), fixadas as demais preditoras

Tabela de ANOVA – Notação Matricial

Fonte de Variação	Número de graus de liberdade	Soma de Quadrados Sequencial (SQ Tipo I)
A (X1,X2)	2	SQ(A)
B|A	1	SQ(B|A) = SQMod(A,B) – SQ(A)
A*B|A,B	2	SQMod(A*B|A,B) = SQ(A,B,A*B) – SQ(A,B)
Resíduo	*n-ab=n-6*	SQRes = SQRes(Mi) - SQRes(A,B,A*B)
TOTAL	*n-1*	$Y'\left[I_n - \frac{1}{n}1_{n\times n}\right]Y = Y'Y - n\bar{Y}^2$

Tabela de ANOVA – Notação Matricial

Fonte de Variação	Número de graus de liberdade	Soma de Quadrados Sequencial (SQ Tipo I)
X1	1	SQ(A) = SQMod(X1,X2) = SQMod(X1) + SQMod(X2|X1)
X2|X1	1	
X3|X1,X2	1	SQ(B|A)=SQ(X3|X1,X2) = SQMod(X1,X2,X3) – SQMod(X1,X2)
X1*X3|X1,X2,X3	1	SQ(A2*B|A,B) = SQMod(X1*X3|X1,X2,X3) = SQMod(X1,X2,X3,X1*X3) - SQMod(X1,X2,X3)
X2*X3|X1,X2,X3,X1*X3	1	SQ(A3*B) = SQMod(X2*X3|X1,X2,X3,X1*X3) = SQMod(X1,X2,X3,X1*X3,X2*X3) – SQMod(X1,X2,X3,X1*X3)
Resíduo	$n-ab=n-6$	
TOTAL	$n-1$	$Y'\left[I_n - \frac{1}{n}1_{n\times n}\right]Y = Y'Y - n\bar{Y}^2$

Partição do Modelo

$y \doteq X\beta$ **Modelo Completo (Full)**

$y \doteq X_1\beta_1 + X_2\beta_2$

$y \doteq X_2\beta_2 + X_1\beta_1$

$\hat{\beta} = \left(X'X\right)^{-1} X'y$ **Estimativas no Modelo Completo**

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \left(X_1'X_1\right)^{-1} & \left(X_1'X_2\right)^{-1} \\ \left(X_2'X_1\right)^{-1} & \left(X_2'X_2\right)^{-1} \end{pmatrix} \begin{pmatrix} X_1'y \\ X_2'y \end{pmatrix}$$

$$\hat{\beta}_1 = \left(X_1'Q_2X_1\right)^{-1} X_1'y; \quad Q_2 = \left(I - X_2\left(X_2'X_2\right)^{-1} X_2'\right)$$

$$\hat{\beta}_2 = \left(X_2'Q_1X_2\right)^{-1} X_2'y; \quad Q_1 = \left(I - X_1\left(X_1'X_1\right)^{-1} X_1'\right)$$

Assim, no caso geral, as estimativas dos coeficientes de X1 dependem de X2 e vice-versa. Quando X1 e X2 são ortogonais as estimativas são independentes e a ordem de entrada não importa, facilitando a interpretação.

Fatorial 3x2x3 - Desbalanceado

Fatorial 3x2x3: Fator A (3 níveis) Fator B (2 níveis) Fator C (3 níveis)

A SQ para um efeito de interesse (digamos o efeito de C) é calculada como a diferença entre SQResíduo de dois modelos diferindo somente no termo de interesse.
Considere as seguintes possibilidades:

Pares de modelos	SQ(C)
(1), (1, C)	SQ(C|1)=SQRes(1)-SQRes(1,C)
(1, A), (1, A, C)	SQ(C|1,A)
(1, B), (1, B, C)	SQ(C|1,B)
(1, A, B), (1, A, B, C)	SQ(C|1,A,B)
(1, A, B, AB), (1, A, B, AB, C)	SQ(C|1,A,B,AB)

Em geral, estas SQ diferem. Qual SQ(C) deve ser adotada na avaliação do efeito de C?
SQ Tipo I (SQ Sequencial), SQ Tipo II e SQ Tipo III (SQ Ajustada)

Fatorial 3x2x3 - Desbalanceado

Fatorial 3x2x3: Fator A (3 níveis) Fator B (2 níveis) Fator C (3 níveis)

SQ Tipo I (SQ Sequencial, adotada no R): depende da ordem de entrada dos fatores na tabela de ANOVA

SQ Tipo II: na SQ de um efeito, devem precede-lo os modelos de mais alta hierárquica não incluindo o termo de interesse. Exemplo, os seguintes modelos devem precede o efeito desejado:

Na SQ(ABC): SQ(ABC|1,A,B,C,AB,AC,BC)
Na SQ(BC): SQ(BC|1, A, B, C, AC, AB)
Na SQ(C): SQ(C|1, A, B, AB)

SQ Tipo III (SQ completamente Ajustada, adotada no SAS, Minitab):

Na SQ(A): SQ(A|1, B, C, AB, AC, BC, ABC)

Em todos os casos, o Quadrado Médio Residual (QMRes) usado para testar o efeito é sempre definido pelo QMREs do modelo completo, que inclui todos os fatores da tabela de ANOVA sob análise.