O princípio da parcimónia na RLM

Recordemos o princípio da parcimónia na modelação: queremos um modelo que descreva adequadamente a relação entre as variáveis, mas que seja o mais simples (parcimonioso) possível.

Caso se disponha de um modelo de Regressão Linear Múltipla com um ajustamento considerado adequado, a aplicação deste princípio traduz-se em saber se será possível obter um modelo com menos variáveis preditoras, sem perder significativamente em termos de qualidade de ajustamento.

Modelo e Submodelos

Se dispomos de um modelo de Regressão Linear Múltipla, com relação de base

$$Y \;=\; \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \;,$$

chamamos submodelo a um modelo de regressão linear múltipla contendo apenas algumas das variáveis preditoras, e.g.,

$$Y \;=\; \beta_0 \qquad + \beta_2 x_2 \qquad\qquad + \beta_5 x_5 \;,$$

Podemos identificar o submodelo pelo conjunto $\mathscr{S}$ das variáveis preditoras que pertencem ao submodelo. No exemplo, $\mathscr{S} = \{2, 5\}$.

O modelo e o submodelo são idênticos se $\beta_j = 0$ para qualquer variável x_j cujo índice não pertença a $\mathscr{S}$.

Comparando modelo e submodelos

Para comparar um modelo e um seu submodelo (identificado pelo conjunto $\mathscr{S}$ dos índices das suas variáveis), precisamos de optar entre as hipóteses:

$$H_0 : \beta_j = 0, \quad \forall j \notin \mathscr{S} \qquad \text{vs.} \qquad H_1 : \exists j \notin \mathscr{S} \quad \text{tal que} \;\; \beta_j \neq 0.$$

[SUBMODELO = MODELO] [SUBMODELO $\neq$ MODELO]

NOTA: Esta discussão só envolve coeficientes β_j de variáveis preditoras ($j > 0$). O coeficiente β_0 faz sempre parte dos submodelos e não é relevante do ponto de vista da parcimónia.

Caso não se rejeite H_0, opta-se pelo submodelo (mais parcimonioso).

Caso se rejeite H_0, opta-se pelo modelo completo (ajusta-se significativamente melhor).

Este coeficiente β_0 não é relevante do ponto de vista da parcimónia: a sua presença não implica trabalho adicional de recolha de dados, nem de interpretação do modelo. Apenas permite um melhor ajustamento.

Estatística de teste para comparar modelo/submodelo

A estatística de teste compara as Somas de Quadrados Residuais do:

- modelo completo (referenciado pelo índice C); e do
- submodelo (referenciado pelo índice S)

Seja k o número de preditores do submodelo ($k+1$ parâmetros). Tem-se, sob H_0 ($\beta_j = 0$, para todas as variáveis x_j que não estão no submodelo):

$$F = \frac{\frac{SQRE_S - SQRE_C}{p-k}}{\frac{SQRE_C}{n-(p+1)}} \frown F_{[p-k\,,\,n-(p+1)]}\,,$$

Nota: Necessariamente $SQRE_s \geq SQRE_c$.

São os valores grandes da estatística que levantam dúvidas sobre H_0.

O teste a um submodelo (teste F parcial)

Teste F de comparação dum modelo com um seu submodelo

Dado o Modelo de Regressão Linear Múltipla,

Hipóteses:

$$H_0 : \beta_j = 0, \quad \forall j \notin \mathscr{S} \quad \text{vs.} \quad H_1 : \exists j \notin \mathscr{S} \quad \text{tal que} \quad \beta_j \neq 0.$$

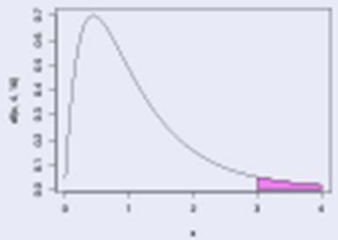
Estatística do Teste:

$$F = \frac{\frac{SQRE_S - SQRE_C}{p-k}}{\frac{SQRE_C}{n-(p+1)}} \frown F_{[p-k,\, n-(p+1)]}, \text{ sob } H_0.$$

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se $F_{calc} > f_{\alpha[p-k,\, n-(p+1)]}$



Expressão alternativa para a estatística do teste

A estatística do teste F parcial pode ser escrita na forma alternativa:

$$F = \frac{n-(p+1)}{p-k} \cdot \frac{R_C^2 - R_S^2}{1-R_C^2} \, .$$

NOTA: A Soma de Quadrados Total apenas depende dos valores observados da variável resposta Y e não do modelo ajustado. Assim, SQT é igual no modelo completo e no submodelo.

As hipóteses do teste também se podem escrever como

$$H_0 : \mathscr{R}_C^2 = \mathscr{R}_S^2 \qquad \text{vs.} \qquad H_1 : \mathscr{R}_C^2 > \mathscr{R}_S^2 \, ,$$

A hipótese H_0 indica que o grau de relacionamento linear entre Y e o conjunto dos preditores é idêntico no modelo e no submodelo.

Caso não se rejeite H_0, opta-se pelo submodelo (mais parcimonioso).
Caso se rejeite H_0, opta-se pelo modelo completo (ajusta-se significativamente melhor).

Teste F parcial: formulação alternativa

Teste F de comparação dum modelo com um seu submodelo

Dado o Modelo de Regressão Linear Múltipla,

Hipóteses:

$$H_0 : \mathscr{R}^2_C = \mathscr{R}^2_S \quad \text{vs.} \quad H_1 : \mathscr{R}^2_C > \mathscr{R}^2_S \, .$$

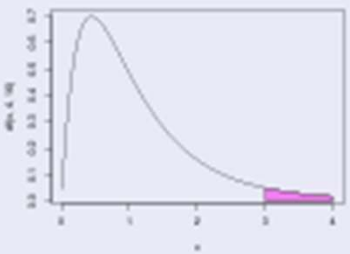
Estatística do Teste:

$$F = \frac{n-(p+1)}{p-k} \cdot \frac{R^2_C - R^2_S}{1-R^2_C} \quad \frown \quad F_{[p-k\,,\,n-(p+1)]}, \text{ sob } H_0.$$

Nível de significância do teste: α

Região Crítica (Região de Rejeição): Unilateral direita

Rejeitar H_0 se $F_{calc} > f_{\alpha[p-k\,,\,n-(p+1)]}$



Relações dos testes F parcial

O teste de ajustamento global é equivalente a um teste F parcial comparando um modelo linear e o submodelo nulo (sem preditores).

Caso o modelo e submodelo difiram num único preditor, X_i, o teste F parcial é equivalente ao teste t com as hipóteses $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$.

Nesse caso, não apenas as hipóteses dos dois testes são iguais, como a estatística do teste F parcial é o quadrado da estatística do teste t referido.

- as hipóteses dos dois testes são iguais ($H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$);
- a estatística do teste F parcial é o quadrado da estatística do teste t referido:

$$F_{calc} = T^2_{calc}$$

Tem-se $p - k = 1$, e como é sabido, se uma variável aleatória T tem distribuição t_ν, então o seu quadrado, T^2 tem distribuição $F_{1,\nu}$.

Numa regressão linear simples, o teste t ao declive da recta ser nulo é equivalente ao teste F de ajustamento global. A segunda destas estatística de teste é o quadrado da primeira.

Ainda o exemplo dos lírios

Teste F Parcial de comparação de um modelo (com p preditores) com um seu submodelo (com k preditores)

Submodelo: PetalWidth = β_0 + β_3 PetalLength

Modelo completo: PetalWidth = β_0 +β_1 SepalLength+ β_2 SepalWidth+ β_3 PetalLength

$$H_0: \beta_1 = \beta_2 = 0 \qquad vs. \qquad H_1: \exists_{j=1,2}: \beta_j \neq 0$$

```
proc reg data=iris;
  model PetalWidth = SepalLength SepalWidth PetalLength;
  test SepalLength, SepalWidth;
run;
```

The SAS System

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable PetalWidth				
Source	DF	Mean Square	F Value	Pr > F
Numerator	2	0.46490	12.62	<.0001
Denominator	146	0.03685		

$p-k$

$\frac{SQRE_S - SQRE_C}{p-k}$

$n-(p+1)$ $\quad \frac{SQRE_C}{n-(p+1)}$

$$F_{calc} = \frac{\frac{SQRE_S - SQRE_C}{p-k}}{\frac{SQRE_C}{n-(p+1)}}$$

Rejeita-se a hipótese nula (para qualquer nível de significância usual), portanto, a qualidade do ajustamento dos dois modelos difere significativativamente.

Ainda o exemplo dos lírios

Teste F Parcial de comparação de um modelo (com p preditores) com um seu submodelo (com k preditores)

Submodelo: PetalWidth = β_0 + β_3 PetalLength

Modelo completo: PetalWidth = β_0 +β_1 SepalLength+ β_2 SepalWidth+ β_3 PetalLength

De forma equivalente: $H_0 : \mathscr{R}^2_C = \mathscr{R}^2_S \quad \text{vs.} \quad H_1 : \mathscr{R}^2_C > \mathscr{R}^2_S$

Os valores dos coeficientes de determinação amostrais ($R_S^2 = 0.9271 \; e \; R_C^2 = 0.9379$) são significativamente diferentes.

```
proc reg data=iris;
   model PetalWidth = PetalLength/clb covb xpx;
run;
```

Root MSE	0.20648	R-Square	0.9271
Dependent Mean	1.19933	Adj R-Sq	0.9266
Coeff Var	17.21659		

```
proc reg data=iris;
   model PetalWidth = SepalLength SepalWidth
PetalLength/clb covb xpx R CLI CLM ;
run;
```

Root MSE	0.19197	R-Square	0.9379
Dependent Mean	1.19933	Adj R-Sq	0.9366
Coeff Var	16.00615		

Exercícios:

1) Usando os valores dos coeficientes de determinação dos dois modelos ajustados verifique que o valor do $F_{ParcialCalc} = 12.62$ (slide anterior);
2) Usando os valores das somas dos quadrados dos resíduos dos dois modelos ajustados verifique que o valor do $F_{ParcialCalc} = 12.62$ (slide anterior).

Ainda o exemplo dos lírios

Exercícios (continuação):

Relembrando alguns resultados do ajustamento do modelo completo:

```
proc reg data=iris;
   model PetalWidth = SepalLength SepalWidth PetalLength/clb covb xpx R CLI CLM ;
run;
```

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-0.24031	0.17837	-1.35	0.1800	-0.59283	0.11221
SepalLength	1	-0.20727	0.04751	-4.36	<.0001	-0.30115	-0.11338
SepalWidth	1	0.22283	0.04894	4.55	<.0001	0.12611	0.31955
PetalLength	1	0.52408	0.02449	21.40	<.0001	0.47568	0.57249

3) a) Utilize um teste F parcial para ver se é possível concluir que os modelos com e sem o preditor SepalLength têm ajustamento significativamente diferente (utilize $\alpha = 0.05$).

b) Qual o coeficiente de determinação do submodelo resultante da exclusão dessa variável?

Como escolher um submodelo?

O teste F parcial (teste aos modelos encaixados) permite-nos optar entre um modelo e um seu submodelo. Por vezes, um submodelo pode ser sugerido por:

- razões de índole teórica, sugerindo que determinadas variáveis preditoras não sejam, na realidade, importantes para influenciar os valores de Y.
- razões de índole prática, como a dificuldade, custo ou volume de trabalho associado à recolha de observações para determinadas variáveis preditoras.

Nestes casos, pode ser claro que submodelo(s) se deseja testar.

Como escolher um submodelo? (cont.)

Mas em muitas situações não é evidente qual o subconjunto de variáveis preditoras que se deseja considerar no submodelo. Pretende-se apenas ver se o modelo é simplificável. Nestes casos, a opção por um submodelo não é um problema fácil.

Dadas p variáveis preditoras, o número de subconjuntos, de qualquer cardinalidade, excepto 0 (modelo nulo) e p (o modelo completo) que é possível escolher é dado por $2^p - 2$. A tabela seguinte indica o número desses subconjuntos para $p = 5, 10, 15, 20, 30$.

p	$2^p - 2$
5	30
10	1 022
15	32 766
20	1 048 574
30	1 073 741 822

Para valores de p pequenos, é possível analisar todos os possíveis subconjuntos. Com algoritmos e rotinas informáticas adequadas, a pesquisa completa de todos os possíveis subconjuntos ainda é possível para valores grandes de p (até $p \approx 35$). Mas para p muito grande, uma pesquisa completa é computacionalmente inviável.

Não é legítimo optar pela exclusão de várias variáveis preditoras em simultâneo, com base nos testes t à significância de cada coeficiente β_j no modelo completo.

De facto, os testes t aos coeficientes β_j admitem que todas as restantes variáveis pertencem ao modelo. A exclusão de um qualquer preditor altera o ajustamento: altera os valores estimados b_j e os respectivos erros padrão das variáveis que permanecem no submodelo. Pode acontecer que um preditor seja dispensável num modelo completo, mas deixe de o ser num submodelo, ou viceversa.

Algoritmos de pesquisa sequenciais

Caso não esteja disponível *software* apropriado, ou se o número p de preditores for demasiado grande, pode recorrer-se a algoritmos de pesquisa que simplificam uma regressão linear múltipla sem analisar todo os possíveis submodelos e sem a garantia de obter os melhores subconjuntos.

Vamos considerar um algoritmo que, em cada passo, exclui uma variável preditora, até alcançar uma condição de paragem considerada adequada, ou seja, um algoritmo de exclusão sequencial (*backward elimination*).

Existem variantes deste algoritmo, não estudadas aqui:

- algoritmo de inclusão sequencial (*forward selection*).
- algoritmos de exclusão/inclusão alternada (*stepwise selection*).

O algoritmo de exclusão sequencial com testes aos β_j

1. ajustar o modelo completo, com os p preditores;
2. definir um nível de significância α para os testes de hipóteses a $\beta_j = 0$;
3. para todas as variáveis rejeita-se $H_0 : \beta_j = 0$?
   - Se sim: não é possível simplificar o modelo (passar ao ponto 4).
   - Se não: variáveis em que não se rejeita H_0 são dispensáveis (candidatas à exclusão).
     - ⋆ se apenas existe uma candidata a sair, excluir essa variável;
     - ⋆ se existir mais do que uma variável candidata a sair, excluir a variável associada ao maior *p-value* (isto é, ao valor da estatística t mais próxima de zero)

   Reajustar o modelo após a exclusão da variável e repetir este ponto 3
4. Quando não existirem variáveis candidatas a sair, ou quando sobrar um único preditor, o algoritmo pára. Tem-se então o submodelo final.

Critério de Informação de Akaike

O R disponibiliza funções para automatizar pesquisas sequenciais de submodelos, semelhantes à que aqui foi enunciada, mas em que o critério de exclusão duma variável em cada passo se baseia no Critério de Informação de Akaike (AIC).

Critério de Informação de Akaike (AIC)

O AIC é uma medida geral da qualidade de ajustamento de modelos. No contexto duma Regressão Linear Múltipla com k variáveis preditoras, define-se como

$$AIC \; = \; n \cdot \ln\left(\frac{SQRE_k}{n}\right) + 2(k+1) \; .$$

Nota: O AIC pode tomar valores negativos.

Interpretando o AIC

$$AIC = n \cdot \ln\left(\frac{SQRE_k}{n}\right) + 2(k+1)$$

- a primeira parcela é função crescente de $SQRE_k$, i.e., quanto melhor o ajustamento, mais pequena a primeira parcela;
- a segunda parcela mede a complexidade do modelo ($k+1$ é o número de parâmetros), pelo que quanto mais parcimonioso o modelo, mais pequena a segunda parcela.

Assim, o AIC depende simultaneamente da qualidade do ajustamento e da simplicidade do modelo.

Um modelo para a variável resposta Y é considerado melhor que outro se tiver um AIC menor (quando ajustados com os mesmos dados).

Algoritmo de exclusão sequencial com base no AIC

Pode definir-se um algoritmo de exclusão sequencial, com base no critério AIC:

- ajustar o modelo completo e calcular o respectivo AIC.
- ajustar cada submodelo com menos uma variável e calcular o respectivo AIC.
- Se nenhum dos AICs dos submodelos considerados for inferior ao AIC do modelo anterior, o algoritmo termina sendo o modelo anterior o modelo final.
  Caso alguma das exclusões reduza o AIC, efectua-se a exclusão que mais reduz o AIC e regressa-se ao ponto anterior.

As duas variantes dos algoritmos

Os algoritmos de exclusão sequencial baseados nos testes *t* ou no AIC coincidem nas variáveis a excluir, podendo diferir apenas no momento de paragem.

Em geral, um algoritmo de exclusão sequencial baseado no AIC é mais cauteloso na exclusão, sobretudo se o valor de α usado nos testes *t* for baixo. Nos algoritmos baseados nos testes *t*, é aconselhável usar valores mais elevados de α, como $\alpha = 0.10$.

Um algoritmo de exclusão sequencial não garante a identificação do "*melhor* submodelo" com um dado número de preditores. Apenas identifica, de forma computacionalmente ligeira, submodelos "bons".

Deve ser usado com bom senso e o submodelo obtido cruzado com outras considerações (e.g., o custo ou dificuldade de obtenção de cada variável, ou o papel que a teoria relativa ao problema em questão reserva a cada preditor).

Exemplo: prever a percentagem de músculo em carcaça de porcos a partir de 7 preditores

```
proc reg data=porcos;
model Musculo = Area Gordurasubcut Peso Rendimento Gordurarenalpel Comprimento LarguraAnca /clb
covb xpx R CLI CLM ;
output out=out_reg p=predicted_value;
test Area, Gordurasubcut, Peso, Rendimento;
RUN;
quit;
```

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	54.11487	9.27061	5.84	<.0001	33.91594	74.31380
Area	1	0.06200	0.70162	0.09	0.9310	-1.46670	1.59070
Gordurasubcut	1	-0.93861	0.36030	-2.61	0.0230	-1.72363	-0.15359
Peso	1	0.24489	0.26196	0.93	0.3683	-0.32587	0.81565
Rendimento	1	0.00623	0.08323	0.07	0.9416	-0.17511	0.18756
Gordurarenalpel	1	-0.01436	0.00714	-2.01	0.0673	-0.02991	0.00119
Comprimento	1	0.01774	0.04832	0.37	0.7199	-0.08755	0.12302
LarguraAnca	1	0.11974	0.06255	1.91	0.0797	-0.01654	0.25602

- Há 6 preditores cuja exclusão individual é admissível (com $\alpha = 0.05$).

- **Mas não é legítimo concluir que Area, Peso, Rendimento, Gordurarenalpel, Comprimento e LarguraAnca são dispensváveis em conjunto.**

O algoritmo de exclusão sequencial com testes aos β_j

(com $\alpha = 0.10$)

ajustar o modelo completo, com os p preditores

Modelo inicial, Completo, $p = 7$

```
model Musculo = Area Gordurasubcut Peso Rendimento
Gordurarenalpel Comprimento LarguraAnca
```

The REG Procedure
Model: MODEL1
Dependent Variable: Musculo

Number of Observations Read	20
Number of Observations Used	20

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	77.98252	11.14036	21.50	<.0001
Error	12	6.21748	0.51812		
Corrected Total	19	84.20000			

Root MSE	0.71981	R-Square	0.9262
Dependent Mean	54.30000	Adj R-Sq	0.8831
Coeff Var	1.32561		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	54.11487	9.27061	5.84	<.0001	33.91594	74.31380
Area	1	0.06200	0.70162	0.09	0.9310	-1.46670	1.59070
Gordurasubcut	1	-0.93861	0.36030	-2.61	0.0230	-1.72363	-0.15359
Peso	1	0.24489	0.26196	0.93	0.3683	-0.32587	0.81565
Rendimento	1	0.00623	0.08323	0.07	0.9416	-0.17511	0.18756
Gordurarenalpel	1	-0.01436	0.00714	-2.01	0.0673	-0.02991	0.00119
Comprimento	1	0.01774	0.04832	0.37	0.7199	-0.08755	0.12302
LarguraAnca	1	0.11974	0.06255	1.91	0.0797	-0.01654	0.25602

para todas as variáveis rejeita-se $H_0 : \beta_j = 0$?
$(\alpha = 0.10)$

se existir mais do que uma variável candidata a sair, excluir a variável associada ao maior *p-value* (isto é, ao valor da estatística *t* mais próxima de zero)

Exemplo (continuação)

Reajustar o modelo após a exclusão da variável rendimento

```
model Musculo = Area Gordurasubcut Peso Gordurarenalpel Comprimento LarguraAnca
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	54.50052	7.40466	7.36	<.0001	38.50373	70.49732
Area	1	0.05151	0.66065	0.08	0.9390	-1.37575	1.47877
Gordurasubcut	1	-0.94880	0.32059	-2.96	0.0111	-1.64138	-0.25621
Peso	1	0.25275	0.23057	1.10	0.2929	-0.24536	0.75087
Gordurarenalpel	1	-0.01427	0.00677	-2.11	0.0549	-0.02889	0.00034730
Comprimento	1	0.01672	0.04456	0.38	0.7135	-0.07955	0.11300
LarguraAnca	1	0.11927	0.05981	1.99	0.0675	-0.00994	0.24849

para todas as variáveis rejeita-se $H_0 : \beta_j = 0$? $(\alpha = 0.10)$

se existir mais do que uma variável candidata a sair, excluir a variável associada ao maior *p-value* (isto é, ao valor da estatística *t* mais próxima de zero)

Exemplo (continuação)

Reajustar o modelo após a exclusão da variável Area

```
model Musculo = Gordurasubcut Peso Gordurarenalpel Comprimento LarguraAnca
```

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	54.95081	4.46691	12.30	<.0001	45.37024	64.53139
Gordurasubcut	1	-0.96660	0.21686	-4.46	0.0005	-1.43173	-0.50147
Peso	1	0.25637	0.21768	1.18	0.2585	-0.21050	0.72325
Gordurarenalpel	1	-0.01457	0.00534	-2.73	0.0162	-0.02602	-0.00313
Comprimento	1	0.01747	0.04195	0.42	0.6834	-0.07251	0.10745
LarguraAnca	1	0.11937	0.05764	2.07	0.0573	-0.00425	0.24298

para todas as variáveis rejeita-se $H_0 : \beta_j = 0$?
$(\alpha = 0.10)$

se existir mais do que uma variável candidata a sair, excluir a variável associada ao maior *p-value* (isto é, ao valor da estatística *t* mais próxima de zero)

Exemplo (continuação)

Reajustar o modelo após a exclusão da variável Comprimento

```
model Musculo = Gordurasubcut Peso Gordurarenalpel LarguraAnca
```

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	55.95724	3.65149	15.32	<.0001	48.17427	63.74020
Gordurasubcut	1	-0.98841	0.20456	-4.83	0.0002	-1.42443	-0.55240
Peso	1	0.26808	0.20982	1.28	0.2208	-0.17915	0.71531
Gordurarenalpel	1	-0.01423	0.00512	-2.78	0.0141	-0.02515	-0.00331
LarguraAnca	1	0.12268	0.05549	2.21	0.0430	0.00441	0.24095

para todas as variáveis rejeita-se $H_0 : \beta_j = 0$?

$(\alpha = 0.10)$

se apenas existe uma candidata a sair, excluir essa variável;

Exemplo (continuação)

Reajustar o modelo após a exclusão da variável Peso

```
model Musculo = Gordurasubcut Gordurarenalpel LarguraAnca
```

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	60.21488	1.52200	39.56	<.0001	56.98839	63.44138
Gordurasubcut	1	-0.92545	0.20242	-4.57	0.0003	-1.35456	-0.49633
Gordurarenalpel	1	-0.01721	0.00465	-3.70	0.0019	-0.02707	-0.00734
LarguraAnca	1	0.11380	0.05613	2.03	0.0596	-0.00519	0.23279

para todas as variáveis rejeita-se $H_0 : \beta_j = 0$?

$(\alpha = 0.10)$

Quando não existirem variáveis candidatas a sair, ou quando sobrar um único preditor, o algoritmo pára. Tem-se então o submodelo final.

Submodelo final:

```
model Musculo = Gordurasubcut Gordurarenalpel LarguraAnca
```

Root MSE	0.66078	R-Square	0.9170
Dependent Mean	54.30000	Adj R-Sq	0.9015
Coeff Var	1.21690		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	60.21488	1.52200	39.56	<.0001	56.98839	63.44138
Gordurasubcut	1	-0.92545	0.20242	-4.57	0.0003	-1.35456	-0.49633
Gordurarenalpel	1	-0.01721	0.00465	-3.70	0.0019	-0.02707	-0.00734
LarguraAnca	1	0.11380	0.05613	2.03	0.0596	-0.00519	0.23279

Algoritmo de exclusão sequencial com base no AIC

Um modelo para a variável resposta *Y* é considerado melhor que outro se tiver um AIC menor (quando ajustados com os mesmos dados).

```
proc reg data=porcos;
model Musculo = Area Gordurasubcut Peso
Rendimento Gordurarenalpel Comprimento
LarguraAnca /clb covb xpx R CLI CLM
selection=adjrsq aic;
RUN;
```

Number in Model	Adjusted R-Square	R-Square	AIC	Variables in Model
4	0.9052	0.9252	-13.1025	Gordurasubcut Peso Gordurarenalpel LarguraAnca
3	0.9015	0.9170	-13.0365	Gordurasubcut Gordurarenalpel LarguraAnca
5	0.8997	0.9261	-11.3487	Gordurasubcut Peso Gordurarenalpel Comprimento LarguraAnca
5	0.8987	0.9253	-11.1425	Area Gordurasubcut Peso Gordurarenalpel LarguraAnca
5	0.8985	0.9252	-11.1101	Gordurasubcut Peso Rendimento Gordurarenalpel LarguraAnca
4	0.8971	0.9188	-11.4592	Gordurasubcut Gordurarenalpel Comprimento LarguraAnca
4	0.8962	0.9181	-11.2878	Area Gordurasubcut Gordurarenalpel LarguraAnca
4	0.8956	0.9176	-11.1738	Gordurasubcut Rendimento Gordurarenalpel LarguraAnca
6	0.8920	0.9261	-9.3580	Area Gordurasubcut Peso Gordurarenalpel Comprimento LarguraAnca
6	0.8920	0.9261	-9.3543	Gordurasubcut Peso Rendimento Gordurarenalpel Comprimento LarguraAnca
5	0.8915	0.9201	-9.7781	Gordurasubcut Rendimento Gordurarenalpel Comprimento LarguraAnca
6	0.8909	0.9253	-9.1440	Area Gordurasubcut Peso Rendimento Gordurarenalpel LarguraAnca
5	0.8905	0.9193	-9.5898	Area Gordurasubcut Gordurarenalpel Comprimento LarguraAnca
5	0.8900	0.9189	-9.5004	Area Gordurasubcut Rendimento Gordurarenalpel LarguraAnca
6	0.8842	0.9208	-7.9614	Area Gordurasubcut Rendimento Gordurarenalpel Comprimento LarguraAnca
2	0.8834	0.8957	-10.4631	Gordurasubcut Gordurarenalpel

...