ESTUDOS I

FACULDADE de ECONOMIA da UNIVERSIDADE do ALGARVE

E S T U D O S I

Cidadania, Instituição e Património

Economia e Desenvolvimento Regional

Finanças e Contabilidade

Gestão e Apoio à Decisão

Modelos Aplicados à Economia e à Gestão

A Faculdade de Economia da Universidade do Algarve

Faculdade de Economia da Universidade do Algarve

2004

FICHA TÉCNICA

Faculdade de Economia da Universidade do Algarve
Campus de Gambelas, 8005-139 Faro
Tel. 289817571 Fax. 289815937
E-mail: ccfeua@ualg.pt
Website: www.ualg.pt/feua

Título
Estudos I - Faculdade de Economia da Universidade do Algarve

Autor
Vários

Editor
Faculdade de Economia da Universidade do Algarve
Morada: Campus de Gambelas
Localidade: FARO
Código Postal: 8005-139

Compilação e Design Gráfico
Susy A. Rodrigues

Revisão de Formatação e Paginação
Lídia Rodrigues

Fotolitos e Impressão
Serviços Gráficos da Universidade do Algarve

ISBN
972-99397-0-5 - Data: 26.10.2004

Depósito Legal
218279/04

Tiragem
500 exemplares

Data
Novembro 2004

Semelhanças entre análise de variância com classificação simples e análise de regressão com variáveis dummy

Patrícia Oom do Valle
Faculdade de Economia, Universidade do Algarve

Efigénio Rebelo
Faculdade de Economia, Universidade do Algarve

Resumo

Na sua abordagem aos procedimentos de análise da variância e covariância, os manuais de estatística limitam-se a proporcionar uma descrição mais ou menos extensiva destas técnicas, sem referir que os seus objectivos podem ser alcançados mediante a especificação de modelos de regressão linear com uma ou mais variáveis dummy. Do mesmo modo, alguns manuais de econometria referem, quanto muito, que um modelo de regressão linear que apenas considere regressores dummy pode ser encarado como um modelo de análise de variância enquanto que um modelo de regressão que combine regressores dummy com regressores quantitativos pode ser entendido como um modelo de análise de covariância. Todavia, estas afirmações não são fundamentadas pela apresentação de quaisquer demonstrações matemáticas, o que dificulta a sua compreensão e, até, aceitação. Assim, em estudos anteriores, demonstrou-se a relação estreita entre os procedimentos estatísticos de análise de variância com classificação dupla e análise de covariância e a especificação de determinados modelos de regressão linear com regressores dummy (Valle e Rebelo, 2002a; 2002b). O presente estudo completa estas abordagens ao desenvolver o suporte matemático necessário à compreensão da semelhança existente entre a análise de variância com classificação simples e a análise de regressão apenas com variáveis dummy.

Palavras chave: Regressão linear, Variáveis dummy, Análise de variância com classificação simples

Abstract

In the approach to analysis of variance and covariance, the statistics literature provides a more or less detailed description of these techniques, without making reference that their purposes can be accomplished by specifying linear regression models with dummy variables. Similarly, some econometricians point out that a linear regression model with all regressors being dummy variables can be referred to as an "analysis of variance model", and that a linear regression model combining both dummy and quantitative regressors can be classified as an "analysis of covariance model". However, these statements are not accompanied by the corresponding mathematical support, which compromises their comprehension and even their acceptability. Previous studies have demonstrated the narrow relationship among the statistical procedures of two way analysis of variance, and covariance analysis and the specification of particular linear regression models with dummy variables (Valle and Rebelo, 2002a; 2002b). The current study completes these approaches by providing

the needed mathematical support in order to understand the similarities between one way analysis of variance and regression with dummy variables only.

Keywords: Linear regression, Dummy variables, One way analysis of variance

1. Introdução

A análise de variância é uma técnica quantitativa que se insere no âmbito de uma área da estatística bastante importante que se designa por "desenho de experiências". Como refere Martins (2002, p. 229), "trata-se de um método estatístico, desenvolvido por Fisher, que por meio de um teste de igualdade de médias, verifica se factores (variáveis independentes) produzem mudanças sistemáticas em alguma variável de interesse (variável dependente)". A análise de variância não é mais do que uma técnica de análise de dados que comporta um vasto leque de situações distintas consoante o número de factores que se pretende testar e a ausência ou presença de interacções entre esses factores (Mendenhall, Beaver and Beaver, 2003).

O objectivo deste estudo é mostrar que determinadas experiências habitualmente realizáveis mediante a aplicação da análise de variância com classificação simples e efeitos fixos podem ser efectuadas através da estimação de um modelo de regressão linear apenas com variáveis dummy. Como se demonstrará, os objectivos subjacentes à aplicação da técnica de análise de variância com classificação simples e efeitos fixos são semelhantes ao que suportam a especificação de um modelo de regressão com apenas uma variável explicativa nominal. A par de objectivos similares demonstrar-se-á também que ambas as técnicas conduzem a testes estatísticos exactamente iguais.

2. A análise de variância com classificação simples

A análise de variância com classificação simples constitui a forma mais elementar de análise de variância uma vez que toda a atenção do investigador se centra na influência que um único factor explicativo (com dois ou mais níveis ou "tratamentos") pode ter na variável de interesse. Utilizando a classificação de Milton e Arnold (1990), esta forma de análise de variância comporta dois tipos de modelos com as características particulares que passamos a descrever:

i) Modelos de Efeitos Fixos: são aqueles em que se seleccionam os k níveis do factor explicativo de acordo com o interesse que possam ter na realização da experiência. Cada um destes k níveis define uma população da qual será retirada uma amostra aleatória de observações. O objectivo de um modelo deste tipo é testar se as diferenças encontradas entre as médias das k amostras (de dimensão $n_1, n_2, ..., n_k$, respectivamente) se devem somente ao acaso (e, portanto, todas as populações têm média idêntica) ou se, pelo contrário, resultam das amostras pertencerem, efectivamente, a populações com médias distintas. Quer dizer, com base nas médias das k amostras faz-se inferência acerca das médias das populações de onde essas amostras foram retiradas;

ii) Modelos de Efeitos Aleatórios: são aqueles em que as k categorias que serão tidas em consideração na realização da experiência são escolhidas de forma aleatória a partir de um vasto conjunto de categorias possíveis. Num modelo com esta particularidade, as k populações de onde se retiram as amostras não esgotam todo o conjunto relevante de populações. Neste caso, não se pretende ensaiar se existem diferenças entre as médias das k populações de onde se extraíram as amostras, mas sim, se existem diferenças entre as médias de todo o conjunto de populações.

A fim de comparar este tipo de análise de variância com a análise de regressão apenas com uma variável explicativa nominal (que poderá ser incluída na equação de regressão através de um ou mais regressores dummy), ambas as técnicas devem reportar-se ao mesmo conjunto de dados. Considere-se então, como exemplo, que se pretende testar o efeito da variável nominal "escalão etário"–com três níveis ou categorias (k = 3), "menos de 35 anos", "entre 36 e 50 anos" e "mais de 50 anos"–na variável de interesse "número de unidade produzidas por hora de um dado bem por um trabalhador". Para tal, admita-se três populações hipotéticas, a primeira constituída pelos valores de produtividade dos indivíduos com menos de 35 anos, a segunda composta pelos respectivos valores dos trabalhadores com idade compreendida entre os 36 e os 50 anos e, por fim, uma terceira constituída pelos valores de produtividade dos indivíduos com idade superior a 50 anos. A questão que se coloca é a seguinte: fará sentido falar em três populações com médias distintas ou, pelo contrário, a diferença de produtividade média entre os três grupos de trabalhadores não é estatisticamente significativa e, como tal, a média das três populações é a mesma? Formalmente, a hipótese nula e a hipótese alternativa são:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$H_a : \text{pelo menos duas médias são diferentes.}$$

É conveniente salientar que se está perante um modelo de análise de variância com classificação simples e efeitos fixos. Classificação simples, porque se incide apenas no efeito de um único factor explicativo: o escalão etário. Efeitos fixos, porque os níveis seleccionados do referido factor constituem todos os níveis que interessa analisar.

Como se referiu, no exemplo em análise, k = 3. Contudo, de modo a permitir a adaptação das fórmulas que serão apresentadas a um número superior de populações, optou-se por representá-las na sua forma mais genérica sem concretizar, portanto, o valor de k.

Para desenvolver um teste à hipótese de que as médias de k populações são iguais no contexto da análise de variância, considere-se k amostras aleatórias, uma de cada população, com $n_1, n_2, ..., n_k$ observações, respectivamente. Seja $\bar{y}_i$ o valor da média da amostra extraída da i-ésima população de tal forma que $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} \Big/ n_i$ (i = 1,2,...,k).

Perante uma diferença entre os valores obtidos para as médias amostrais, pode levantar-se a seguinte questão: será a diferença encontrada entre as médias amostrais

de dimensão suficientemente grande para se poder rejeitar a hipótese nula $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$?

Antes de se avançar para a apresentação do teste estatístico que permitirá avaliar a magnitude da diferença registada entre as várias médias, é importante referir que as observações obtidas das k populações podem ser representadas através do seguinte modelo

$$y_{ij} = \mu_i + e_{ij} \tag{2.1}$$

em que y_{ij} é a j-ésima observação da população i (i = 1,2,...,k; j = 1,2,..., n_i), μ_i a média dessa população e e_{ij} a variável residual associada à j-ésima observação da população i (normalmente distribuída com variância constante e igual a σ^2).

Uma forma alternativa de representar o modelo (2.1) consiste em definir $\alpha_i = \mu_i - \mu$ onde μ é a média global de todas as populações, dada portanto, por:

$$\mu = \sum_{i=1}^{k} n_i \mu_i \Big/ \sum_{i=1}^{k} n_i \, . \tag{2.2}$$

Representando α_i a diferença entre a média de uma determinada população e a média global, o seu valor não é mais do que uma medida do efeito do i-ésimo nível ou tratamento do factor explicativo na variável de interesse. Ora, de $\alpha_i = \mu_i - \mu$ vem $\mu_i = \mu + \alpha_i$ o que significa que o modelo (2.1) pode expressar-se de forma equivalente como:

$$y_{ij} = \mu + \alpha_i + e_{ij} \, . \tag{2.3}$$

É também importante notar que se $\mu_1 = \mu_2 = \cdots = \mu_k$ então,

$$\mu = \frac{\sum_{i=1}^{k} n_i \mu_i}{\sum_{i=1}^{k} n_i} = \frac{n\mu_i}{n} = \mu_i \quad (i = 1,2,...,k;\ n = n_1 + n_2 + ... + n_k) \tag{2.4}$$

e, consequentemente, $\alpha_i = \mu_i - \mu = 0, \forall i$. Portanto, testar $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ é equivalente a testar $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$, ou seja, que todos os níveis do factor explicativo têm o mesmo efeito na variável de interesse.

A realização da análise de variância com classificação simples requer a verificação simultânea de três pressupostos:

i) A distribuição de probabilidade das k populações envolvidas deverá ser normal;

ii) Todas as populações deverão possuir a mesma variância;

iii) As k amostras deverão ser independentes.

A credibilidade da hipótese nula será testada através da seguinte estatística com distribuição F

$$F = \frac{\dfrac{\sum_{i=1}^{k} n_i \left(\bar{y}_i - \bar{\bar{y}}\right)^2}{k-1}}{\dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2}{\sum_{i=1}^{k} (n_i - 1)}} \sim F\left[k-1, \left(\sum_{i=1}^{k} (n_i - 1)\right)\right] \Bigg| \; H_0 : \mu_1 = \mu_2 = \mu_3 \tag{2.5}$$

em que $\bar{\bar{y}}$, a média global das n observações, se define como

$$\bar{\bar{y}} = \sum_{i=1}^{k}\sum_{i=1}^{n_i} y_{ij} \Big/ \sum_{i=1}^{k} n_i \tag{2.6}$$

ou, na forma equivalente:

$$\bar{\bar{y}} = \sum_{i=1}^{k} n_i \bar{y}_i \Big/ \sum_{i=1}^{k} n_i \,. \tag{2.7}$$

A expressão que se encontra no numerador do rácio F em (2.5) é geralmente conhecida como "Quadrado Médio Explicado" (QME) ou "Variância Explicada" enquanto que a expressão que aparece no denominador deste rácio é habitualmente designada por "Quadrado Médio Residual" (QMR) ou "Variância Residual ou Não Explicada". Os numeradores de cada um destes quocientes, designados por "Soma de Quadrados Explicada" (SQE) e "Soma dos Quadrados Residual" (SQR), respectivamente, relacionam-se da seguinte forma:

$$\text{SQE} + \text{SQR} = \text{SQT}$$

em que a SQT ("Soma dos Quadrados Totais") é determinada a partir da totalidade das observações, ignorando, portanto, a sua divisão em amostras distintas. Note-se que, em extensão, esta igualdade pode escrever-se como:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(y_{ij} - \bar{\bar{y}}\right)^2 = \sum_{i=1}^{k} n_i \left(\bar{y}_i - \bar{\bar{y}}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2 . \tag{2.8}$$

Os rácios que se encontram no numerador e no denominador da estatística F constituem dois estimadores para a variância comum de todas as populações, σ^2. Contudo, enquanto que o primeiro só é centrado se a hipótese nula for verdadeira, pois, na situação contrária, sobrestima o valor da variância, o segundo é centrado independentemente da hipótese nula ser verdadeira ou falsa. Quer dizer, o rácio (2.5) caracteriza-se por flutuar em torno de 1 quando a hipótese de igualdade das médias é verdadeira. Quando esta hipótese é falsa, o rácio F tende a exceder o valor 1, dado que,

nesta circunstância, o QME será grande relativamente ao QMR. Portanto, quanto mais elevado for o rácio F menos credível será H_0. A regra de decisão para o teste é:

$$\text{Rejeitar } H_0 \text{ se } F > F_\alpha[k-1, n-k]$$
$$\text{Não rejeitar } H_0 \text{ se } F \leq F_\alpha[k-1, n-k].$$

Os cálculos envolvidos na análise de variância são normalmente organizados num quadro que se designa por Quadro ANOVA, cuja estrutura se apresenta seguidamente. A representação tabular da análise de variância, para além de permitir uma apresentação organizada dos cálculos, possibilita a sua confirmação. Assim, o primeiro elemento da última linha do quadro, SQT, determina-se a partir de todas as observações adicionando o quadrado do desvio de cada observação em relação à média global $\bar{\bar{y}}$. Como se indicou em (2.8), o valor resultante para a SQT deverá coincidir com a soma dos valores obtidos para a SQE e para a SQR, isto é, os restantes valores que se encontram nessa coluna. Da mesma forma, o número de graus de liberdade da SQT deverá ser igual à soma dos graus de liberdade da SQE e da SQR.

Quadro 2.1 - Quadro ANOVA

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Quadrados Médios	Rácio F
Factor Explicativo	$SQE = \sum_{i=1}^{k} n_i \left(\bar{y}_i - \bar{\bar{y}}\right)^2$	$k-1$	$QME = \frac{SQE}{k-1}$	$F = \frac{QME}{QMR}$
Resíduos	$SQR = \sum_{I=1}^{K} \sum_{J=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2$	$\sum_{i=1}^{k} (n_i - 1)$	$QMR = \frac{SQR}{\sum_{i=1}^{k} (n_i - 1)}$	----------
Total	$SQT = \sum_{I=1}^{k} \sum_{J=1}^{n_i} \left(y_{ij} - \bar{\bar{y}}\right)^2$	$\left(\sum_{i=1}^{k} n_i\right) - 1$	----------	----------

3. Regressão com uma variável nominal e análise de variância com classificação simples

A equivalência entre a análise de variância com classificação simples e a análise de regressão com uma variável explicativa nominal verifica-se independentemente do número de categorias identificáveis nesta última.

Uma vez que a comparação das duas técnicas só é possível se ambas incidirem sobre os mesmos dados, a exposição que seguidamente será efectuada incidirá também na ilustração que serviu de base ao desenvolvimento da secção anterior. Assim, a questão a que se pretende dar resposta através da aplicação da técnica da análise de variância com classificação simples pode ser formulada da seguinte forma: pertencerão os valores de produtividade dos indivíduos que pertencem a cada um dos escalões etários a populações com médias distintas (μ_1, μ_2 e μ_3, respectivamente) ou, pelo contrário, não existem diferenças de produtividade significativas entre os três grupos etários, situação em que é legítimo afirmar que as médias das três populações são

iguais? Como se viu, a hipótese nula a testar e a hipótese alternativa podem, portanto, ser definidas como:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$H_a : \text{pelo menos duas médias são diferentes.}$$

Como se definiu em (2.8), fazendo k = 3, a SQT decompõe-se da seguinte forma

$$\sum_{i=1}^{3}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{\bar{y}}\right)^2 = \sum_{i=1}^{3}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2 + \sum_{i=1}^{3} n_i\left(\bar{y}_i-\bar{\bar{y}}\right)^2 \qquad (3.1)$$

e o rácio F (2.5) concretiza-se em:

$$F = \frac{(n_1+n_2+n_3-3)}{2}\,\frac{\sum_{i=1}^{3} n_i\left(\bar{y}-\bar{\bar{y}}\right)^2}{\sum_{i=1}^{3}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2} \sim F(2, n_1+n_2+n_3-3)\,. \qquad (3.2)$$

Numa perspectiva de análise de regressão, considere-se a divisão dos trabalhadores do processo produtivo em apreciação de acordo com o escalão etário a que pertencem, por exemplo, "menos de 35 anos", "de 36 a 50 anos" e "mais de 50 anos". Admita-se que o output produzido por hora segue uma distribuição normal com variância σ^2 e valor esperado μ_1, μ_2 e μ_3, respectivamente, para cada um dos grupos etários. Com a finalidade de verificar se existem diferenças de produtividade em resultado dos trabalhadores pertencerem a grupos etários distintos, pode representar-se a variável explicativa nominal "escalão etário" definindo os seguintes regressores dummy

$$I_{1i} = \begin{cases} 1, \text{ se o i-ésimo trabalhador tem menos de 35 anos} \\ 0, \text{ caso contrário} \end{cases}$$

$$I_{2i} = \begin{cases} 1, \text{se o i - ésimo trabalhador tem entre 36 e 50 anos} \\ 0, \text{caso contrário} \end{cases}$$

ou, de forma equivalente,

$$I_{1i} = \begin{cases} 1, \ i=1,2,...,n_1 \\ 0, \text{ caso contrário} \end{cases}$$

$$I_{2i} = \begin{cases} 1, \ i=n_1+1,n_1+2,...,n_1+n_2 \\ 0, \text{ caso contrário} \end{cases}$$

É importante notar que se está a assumir que quando I_1 e I_2 assumem simultaneamente o valor 0 a observação diz respeito a um indivíduo com idade superior a 50 anos.

A relação entre a variável dependente, "número de unidades produzidas pelo i-ésimo trabalhador", e a variável explicativa nominal "escalão etário" pode agora ser expressa da seguinte forma

$$y_i = \beta_0 + \beta_1 I_{1i} + \beta_2 I_{2i} + u_i \; ; i = 1,..., n \; ; u_i \sim \text{IIN}(0, \sigma^2) \tag{3.3}$$

em que I_1 e I_2 são as variáveis dummy definidas acima. Em termos matriciais, o modelo (3.3) pode escrever-se como

$$y = X\beta + u \;\; ; \;\; u \sim N(0, \sigma^2 I_n) \, . \tag{3.4}$$

Como habitualmente, o vector de estimadores dos mínimos quadrados ordinários (EMQO) é dado por

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (X'X)^{-1} X'y$$

sendo, neste caso particular,

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \bar{y}_3 \\ \bar{y}_1 - \bar{y}_3 \\ \bar{y}_2 - \bar{y}_3 \end{bmatrix} \tag{3.5}$$

como se demonstra em apêndice.

É importante referir que o vector (3.5) é perfeitamente coerente com o respectivo vector de parâmetros. Na verdade, sob esta codificação específica, a hipótese do modelo de regressão linear $E(u_i) = 0 \;\; (i = 1,2,.., n)$ conduz às seguintes esperanças matemáticas para cada combinação de valores dos regressores dummy

$$E(y_i \mid I_{1i} = 1, I_{2i} = 0) = \beta_0 + \beta_1 \tag{3.6}$$

$$E(y_i \mid I_{1i} = 0, I_{2i} = 1) = \beta_0 + \beta_2 \tag{3.7}$$

$$E(y_i \mid I_{1i} = 0, I_{2i} = 0) = \beta_0 \tag{3.8}$$

ou, considerando a definição apresentada no início desta secção para μ_1, μ_2 e μ_3,

$$\beta_0 + \beta_1 = \mu_1$$

$$\beta_0 + \beta_2 = \mu_2$$
$$\beta_0 = \mu_3$$

donde provem:

$$\beta_0 = \mu_3 \quad (3.9)$$
$$\beta_1 = \mu_1 - \mu_3 \quad (3.10)$$
$$\beta_2 = \mu_2 - \mu_3 . \quad (3.11)$$

Consequentemente, β_0 representa o valor médio da variável dependente para o grupo de referência, isto é, o grupo identificado com o valor 0 por ambas as variáveis dummy. Por sua vez, β_1 e β_2 indicam a diferença entre o valor esperado condicional da variável explicada para a categoria que os respectivos regressores identificam com o valor 1 e o valor esperado condicional de y para a categoria de referência. As estimativas desse valores esperados são, obviamente, dadas por $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_2$. Especificando com o exemplo apresentado, conclui-se que:

i) o coeficiente $\hat{\beta}_0$ estima a produtividade esperada de um trabalhador da população que possua mais de 50 anos;

ii) o coeficiente $\hat{\beta}_1$ estima a diferença entre a produtividade média de um trabalhador da população com menos de 35 anos e a produtividade média de um trabalhador que pertença ao escalão etário mais avançado;

iii) o coeficiente $\hat{\beta}_2$ estima a diferença entre a produtividade média de um trabalhador da população com idade compreendida entre os 36 e os 50 anos e a produtividade média de um trabalhador com mais de 50 anos.

No modelo (3.3), as diferenças de produtividade entre os vários grupos etários são captadas através das duas variáveis dummy, I_1 e I_2, o que significa que o teste apropriado ao efeito na produtividade que resulta da pertença a grupos etários diferentes é o teste F de significância global do modelo. Na verdade, testar a hipótese de que "a produtividade não varia com o nível etário em que se insere o trabalhador", equivale a ensaiar a hipótese nula $H_0 : \beta_1 = \beta_2 = 0$ contra a hipótese alternativa $H_a : \beta_1 \neq 0 \vee \beta_2 \neq 0$. Com efeito, se $\beta_1 = \beta_2 = 0$, $E(y_i \mid I_{1i} = 1, I_{2i} = 0) = E(y_i \mid I_{1i} = 0, I_{2i} = 1) = E(y_i \mid I_{1i} = 0, I_{2i} = 0) = \beta_0$. Uma vez que este teste pode ser entendido como um teste de significância a R^2, não rejeitar a hipótese nula significa que a variabilidade da produtividade não é explicada pelo facto de existirem no processo de fabrico indivíduos com idades diferentes. Se esta hipótese for rejeitada, pode testar-se a existência de diferenças significativas de produtividade esperada em resultado do indivíduo possuir menos de 35 anos ou de estar inserido no nível etário "36 a 50 anos" em vez de pertencer ao grupo de trabalhadores mais velhos, através da realização de testes de significância individual aos coeficientes β_1 e β_2, respectivamente, no âmbito do modelo (3.3). O resultado destes testes é imediatamente conhecido, através da análise do valor da estatística t associado à estimativa de cada um destes coeficientes.

A observação das igualdades (3.10) e (3.11) permite concluir que a hipótese a testar no âmbito da análise de regressão, isto é, $H_0 : \beta_1 = \beta_2 = 0$, equivale a testar $H_0 : \mu_1 = \mu_2 = \mu_3$, ou seja, à hipótese a ensaiar no contexto da Análise de Variância com classificação simples.

Pretende-se agora demonstrar que a estatística do teste a aplicar na Análise de Regressão é equivalente à estatística (3.2) apresentada a propósito da Análise de Variância com classificação simples.

Assim, no modelo (3.3), a hipótese a testar e a estatística que permite realizar o teste são, respectivamente,

$$H_0 : \beta_1 = \beta_2 = 0$$

$$F = \frac{\left(\hat{u}_r' \hat{u}_r - \hat{u}' \hat{u}\right)/g}{\hat{u}' \hat{u}/(n-k)} \sim F_{(g,n-k)} \Big|_{H_0: \beta_1 = \beta_2 = 0} \tag{3.12}$$

onde $\hat{u}' \hat{u}$ é a SQR do modelo (3.3) e $\hat{u}_r' \hat{u}_r$ a SQR do modelo restrito (isto é, do modelo (3.3) após a imposição da restrição $\beta_1 = \beta_2 = 0$). O valor de g representa o número de parâmetros cuja significância estatística conjunta se pretende a testar, isto é, g = 2 no caso em análise.

Autores como Stewart (1991) demonstram a relação de equivalência entre o teste de significância agora descrito e o teste de significância a R^2, isto é, se R^2 é do ponto de vista estatístico diferente de zero. Na verdade, se a hipótese $H_0 : \beta_1 = \beta_2 = 0$ não for rejeitada, pode conclui-se que os regressores utilizados na equação de regressão não explicam a variação da variável dependente o que equivale a afirmar que R^2 é zero. Portanto, a estatística (3.12) aparece por vezes apresentada como

$$F = \frac{n-k}{k-1} \cdot \frac{R^2}{1-R^2} \sim F \Big|_{(k-1,n-k)} \Big|_{H_0 : R^2 = 0} \tag{3.13}$$

o que facilita a realização do teste uma vez que o valor de R^2 é calculado de forma rotineira pela generalidade dos programas informáticos de regressão.

Sendo

$$R^2 = 1 - \frac{\hat{u}' \hat{u}}{y' M_{X_1} y} \tag{3.14}$$

com $M_{X_1} = I_n - x_1 (x_1' x_1)^{-1} x_1'$ e $x_{1i} = 1, (i = 1,2,...,n)$, importa agora converter $\hat{u}' \hat{u}$ e $y' M_{X_1} y$ em somatórios por forma a demonstrar que a expressão (3.13) é idêntica à apresentada em (3.2). Observe-se que

$$SQR = \hat{u}'\hat{u} = \begin{bmatrix} (y_1 - \bar{y}_1)' & (y_2 - \bar{y}_2)' & (y_3 - \bar{y}_3)' \end{bmatrix} \begin{bmatrix} (y_1 - \bar{y}_1) \\ (y_2 - \bar{y}_2) \\ (y_3 - \bar{y}_3) \end{bmatrix} = \sum_{i=1}^{3} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 . \qquad (3.15)$$

Para encontrar a expressão que define $y'M_{X_1}y$, é necessário começar por explicitar o conteúdo da matriz M_{X_1}. Tendo em atenção a definição apresentada acima para esta matriz e o facto do vector x_1 ser constituído apenas por 1's, M_{X_1} pode escrever-se da seguinte forma:

$$M_{X_1} = \begin{bmatrix} 1 - 1/n & -1/n & \cdots & -1/n \\ -1/n & 1 - 1/n & \cdots & -1/n \\ \cdots & \cdots & \cdots & \cdots \\ -1/n & -1/n & \cdots & 1 - 1/n \end{bmatrix}. \qquad (3.16)$$

Consequentemente, e dada a idempotência da matrix M_{X_1},

$$y'M_{X_1}y = y'M_{X_1}M_{X_1}y =$$

$$= \begin{bmatrix} y_1' & y_2' \end{bmatrix} \begin{bmatrix} 1 - 1/n & -1/n & \cdots & -1/n \\ -1/n & 1 - 1/n & \cdots & -1/n \\ \cdots & \cdots & \cdots & \cdots \\ -1/n & -1/n & \cdots & 1 - 1/n \end{bmatrix} \begin{bmatrix} 1 - 1/n & -1/n & \cdots & -1/n \\ -1/n & 1 - 1/n & \cdots & -1/n \\ \cdots & \cdots & \cdots & \cdots \\ -1/n & -1/n & \cdots & 1 - 1/n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} =$$

$$= \sum_{i=1}^{3} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{\bar{y}} \right)^2 . \qquad (3.17)$$

Substituindo (3.15) e (3.17) na fórmula de R^2 em (3.14), vem

$$R^2 = 1 - \left(\sum_{i=1}^{3} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \Big/ \sum_{i=1}^{3} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{\bar{y}} \right)^2 \right)$$

pelo que

$$R^2 / (1 - R^2) = \left(\sum_{i=1}^{3} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{\bar{y}} \right)^2 - \sum_{i=1}^{3} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \right) \Big/ \sum_{i=1}^{3} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{\bar{y}} \right)^2 . \qquad 3.18)$$

Como se mostra em apêndice, o numerador de (3.18) pode ser simplificado da seguinte forma

$$\sum_{i=1}^{3} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{\bar{y}} \right)^2 - \sum_{i=1}^{3} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = \sum_{i=1}^{3} n_i \left(\bar{y}_i - \bar{\bar{y}} \right)^2 \qquad (3.19)$$

pelo que a expressão (3.18) é equivalente a

$$\frac{R^2}{1-R^2} = \frac{\sum_{i=1}^{3} n_i \left(\bar{y}_i - \bar{\bar{y}}\right)^2}{\sum_{i=1}^{3}\sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2}. \qquad (3.20)$$

Em resultado, o rácio F apresentado em (3.13) pode escrever-se como

$$F = \frac{(n_1 + n_2 + n_3 - 3)}{2} \frac{\sum_{i=1}^{3} n_i \left(\bar{y} - \bar{\bar{y}}\right)^2}{\sum_{i=1}^{3}\sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2} \sim F(2, n_1 + n_2 + n_3 - 3) \Big| \; H_0: \mu_1 = \mu_2 = \mu_3$$

ou seja, coincide com (3.2), tornando evidente a equivalência entre a estatística do teste usada no contexto da análise de variância com classificação simples e a estatística do teste à significância global de um modelo de regressão linear com uma variável explicativa nominal. Fica assim demonstrada a equivalência entre estas duas técnicas que, para além de objectivos similares, conduzem a testes estatísticos idênticos.

4. Considerações finais

Este estudo mostra que determinadas experiências tradicionalmente realizáveis através da técnica estatística de análise de variância com classificação simples e efeitos fixos podem, com relativa facilidade, ser conduzidas mediante a especificação de um modelo de regressão com variáveis dummy.

Importa também salientar que o ajustamento de uma equação de regressão com termo independente apresenta vantagens relativamente à aplicação da técnica de análise de variância uma vez que permite prontamente a realização de um número superior de ensaios de hipóteses. Com efeito, se no contexto do modelo de regressão (3.3) a hipótese $H_0 : \beta_1 = \beta_2 = 0$ for rejeitada, consegue-se saber que médias diferem entre si através da realização de testes de significância individual a estes dois parâmetros e subsequente eliminação da variável que tem associada ao parâmetro uma estatística t de valor inferior. Através da aplicação da análise de variância à comparação das médias de k populações, ou se rejeita ou não se rejeita H_0. Neste primeiro caso, conclui-se que pelo menos duas populações têm médias diferentes, mas não é possível identificar de imediato que populações são essas.

Referências

Martins, G. A. (2002) *Estatística Geral e Aplicada*, São Paulo, Editora Atlas.

Mendenhall, W., R. J. Beaver and B. M. Beaver (2003) *Probability and Statistics*, London, Thomson Brooks/Cole.

Milton, J. S. e J. C. Arnold (1990) *Introduction to Probability and Statistics, Principles and Applications for Engineering and the Computing Sciences*, New York, McGraw-Hill.

Stewart, J. (1991), *Econometrics*, Cambridge, Philip Allan.

Valle, P. e E. Rebelo (2002a) Análise de Variância e Análise de Regressão com Variáveis Dummy: Mais Semelhanças do que Diferenças, *Revista de Estatística*, Vol. I, 47-86.

---------- (2002b) Dualidades entre Análise de Covariância e Análise de Regressão com Variáveis Dummy, *Revista de Estatística*, Vol. II, 65-86.