# ESTUDOS II

FACULDADE de ECONOMIA da UNIVERSIDADE do ALGARVE

# ESTUDOS II

**Cidadania, Instituições e Património**

**Economia e Desenvolvimento Regional**

**Finanças e Contabilidade**

**Gestão e Apoio à Decisão**

**Modelos Aplicados à Economia e à Gestão**

Faculdade de Economia da Universidade do Algarve

2005

---

**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 II - 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

***Capa e Design Gráfico***
Susy A. Rodrigues

***Compilação, Revisão de Formatação e Paginação***
Lídia Rodrigues

***Fotolitos e Impressão***
Grafica Comercial – Loulé

***ISBN***
972-99397-1-3 Data: 26-08-2005

***Depósito Legal***
218279/04

***Tiragem***
250 exemplares

***Data***
Novembro 2005

# Previsão com regressores *dummy*

**Patrícia Oom do Valle**
*Faculdade de Economia, Universidade do Algarve*

**Efigénio Rebelo**
*Faculdade de Economia, Universidade do Algarve*

**Resumo**

O presente estudo pretende descrever uma importante mas pouco conhecida aplicação dos regressores *dummy* na análise econométrica: a previsão ex-ante. Em concreto, mostra-se em que medida a estimação de uma única equação de regressão com variáveis *dummy* possibilita que se obtenha toda a informação necessária para a construção de intervalos de confiança para previsão: o valor das próprias previsões pontuais e as variâncias estimadas dos erros de previsão.

A aplicação potencial dos regressores *dummy* na previsão ex-ante foi proposta por Salkever em 1976. Os contributos do presente estudo residem na comparação deste método como o processo *standard* de previsão ex-ante, habitualmente seguido na análise de regressão, bem como na apresentação detalhada do suporte matemático que permite perceber a correspondência entre as duas abordagens.

**Palavras chave**: Regressão linear, Variáveis *dummy*, Previsão ex-ante.

**Abstract**

This study describes an important but almost unknown application of dummy variables in Econometrics: the ex-ante prediction. In particular, this work shows how the estimation of a unique regression model with dummy variables produces all information needed to define confidence intervals for prediction: the punctual predictions and the estimated variances of the prediction errors.

The potential application of dummy variables in prediction was first introduced by Salkever in 1976. The contributions of the current study rely on the comparison of this method with the standard procedures of ex-ante prediction in regression analysis, as well as on the detailed presentation of the mathematical background which allows the understanding of the correspondence between the two approaches.

**Keywords**: Linear Regression, Dummy variables, ex-ante forecasting.

## 1. Introdução

Muito embora os regressores dummy tenham sido originalmente criados com o objectivo de possibilitar a incorporação de factores explicativos de natureza qualitativa em modelos de regressão linear, a sua utilidade estende-se a domínios mais vastos e complexos, entre os quais se pode destacar a determinação de previsões pontuais e o cálculo de erros de previsão. Esta aplicação dos regressores dummy, proposta por Salkever (1976), é, todavia, pouco conhecida e, consequentemente, pouco utilizada.

Este trabalho pretende demonstrar o paralelismo entre o uso de regressores dummy em contextos de previsão e o processo standard de previsão ex-ante. Desta abordagem resultará a compreensão dos pontos de consonância entre as duas abordagens mas também a identificação das situações em que uma ou outra se mostra mais vantajosa.

## 2. O Processo Standard de Previsão

Uma vez estimada a relação de comportamento entre uma determinada variável dependente e um conjunto de variáveis explicativas e testada a sua adequação, quer do ponto de vista estatístico quer do ponto de vista da sua aderência ao enquadramento teórico de suporte, a etapa que se segue em muitos estudos econométricos consubstancia-se na realização de previsão ex-ante, isto é, na determinação do valor que a variável dependente irá assumir em função de valores fixados para as variáveis explicativas. Tipicamente, o processo de previsão ex-ante desenrola-se em três fases:

**(1)** Especificação de um modelo de regressão e estimação dos seus parâmetros para um conjunto de n observações originais. Seja esse modelo dado por

$$y = X\beta + u \; ; \; u \sim N(0, \sigma^2 I_n) \tag{2.1}$$

onde:

$y$ : vector $(n \times 1)$ de observações da variável dependente

X: matriz $(n \times k)$ de observações de k regressores (com $X_{1i} = 1, \; \forall i \;\; i = 1,2,\ldots,n$)

$\beta$ : vector de $(k \times 1)$ parâmetros

u : vector $(n \times 1)$ de desvios

Os E.M.Q.O. são dados, como habitualmente, pelas fórmulas $\hat{\beta} = (X'X)^{-1}X'y$ e $\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{n-k}$ em que $\hat{\sigma}^2 (X'X)^{-1}$ representa a matriz estimada de variâncias e covariâncias de $\hat{\beta}$.

**(2)** Utilização do vector $\hat{\beta}$, estimado com base nas n observações originalmente consideradas para y e para as k variáveis independentes, no cálculo de previsões e erros de previsão para g ( $g \geq 1$) observações adicionais dos regressores. Assim, se definirmos

$X_p$: matriz $(g \times k)$ de g observações adicionais nos k regressores originais

e

$y_p$: vector $(g \times 1)$ de observações previstas para a variável dependente

cujo conteúdo se define, respectivamente, como

$$X_p = \begin{bmatrix} X_{1,n+1} & X_{2,n+1} & \mathbf{L} & X_{k,n+1} \\ X_{1,n+2} & X_{2,n+2} & \mathbf{L} & X_{k,n+2} \\ \mathbf{M} & \mathbf{M} & \mathbf{L} & \mathbf{M} \\ X_{1,n+g} & X_{2,n+g} & \mathbf{L} & X_{k,n+1} \end{bmatrix} \qquad y_p = \begin{bmatrix} y_{n+1} \\ y_{n+2} \\ \mathbf{M} \\ y_{n+g} \end{bmatrix}$$

então $y_p = X_P\beta + u_p$.

No caso em que as hipóteses assumidas para o modelo estimado a partir dos dados originais permanecem válidas para os g períodos de previsão (designadamente, as que dizem respeito à variável residual), então

$$\hat{y}_p = X_P \hat{\beta} \qquad (2.2)$$

apresenta-se como a melhor previsão pontual para $y_p$.

Os erros de previsão são dados por

$$y_p - \hat{y}_p = \hat{u}_p = y_p - X_p\hat{\beta} = X_p\beta + u_p - X_p\hat{\beta} = X_p(\beta - \hat{\beta}) + u_p \qquad (2.3)$$

e devem-se, portanto, a dois factores: por um lado, ao erro de estimação associado à diferença entre $\beta_j$ e $\hat{\beta}_j$ e, por outro lado, ao erro aleatório inerente à previsão em si ($u_p$);

**(3)** Determinação da variância estimada dos erros de previsão estimados e sua utilização na construção de intervalos de confiança para a previsão. Após a realização de algumas manipulações algébricas, Stewart (1991, p. 81), por exemplo, demonstra que

$$\mathrm{Var}(\hat{y}_p) = \sigma^2\left[I_g + X_p(X'X)^{-1}X_p'\right] \qquad (2.4)$$

cujo valor pode ser estimado utilizando $\hat{\sigma}^2 = \hat{u}'\hat{u}/(n-k)$ no lugar de $\sigma^2$.

Consequentemente, o intervalo de confiança para $y_j$ $(j = n+1, n+2, \ldots, n+g)$ será dado por

$$\hat{y}_j \pm t_{\alpha/2}\,\hat{\sigma}_{(\hat{y}_j)} \qquad (2.5)$$

em que $\hat{\sigma}_{(\hat{y}_j)}$ representa o erro padrão do erro de previsão para a j-ésima observação adicional $(j = n+1, n+2, \ldots, n+g)$ (Greene, 1993, p. 195).

.

## 3. Regressores dummy e previsão

Salkever (1976) desenvolveu um método alternativo, assente na definição de regressores dummy, que pode ser utilizado para gerar previsões e avaliar a sua precisão. A importância da técnica apresentada por este autor é ressaltada nomeadamente por Greene (1993) e Maddala (1992) que não hesitam em classifica-la como "...um método conveniente de realizar previsões" (Greene,1993, p.196) ou como uma "...forma mais fácil de gerar o erro de *previsão*" (Maddala,1992, p.154).

Salkever inicia a sua análise explicitando a relação entre a variável dependente y e k variáveis explicativas, para um primeiro conjunto de n observações, através de uma equação de regressão como a apresentada em (2.1), muito embora usando notações diferentes. Com o objectivo de prever o valor de y para g observações adicionais de cada uma das variáveis independentes, este autor propõe que se proceda à especificação de um modelo mais amplo, o qual inclui, para além dos k regressores originais, g variáveis dummy de tal forma que a j-ésima variável dummy assume o valor 1 para o observação $n+j$ ( $j = n+1, n+2, ..., n+g$ ) e o valor zero para as demais observações. Isto é,

$$\begin{bmatrix} y \\ T \end{bmatrix} = \begin{bmatrix} X & 0 \\ X_p & I_g \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \end{bmatrix} + \begin{bmatrix} u \\ u_p \end{bmatrix} \tag{3.1}$$

onde $X_p$ é a matriz acima apresentada, T representa um vector $(g \times 1)$ de valores arbitrariamente atribuídos às observações da variável dependente (cujo conteúdo será discutido posteriormente), $I_g$ é uma matriz identidade $(g \times g)$ constituída pelos valores assumidos pelos regressores dummy e 0 uma matriz nula $(n \times g)$. Se designarmos por

$$C = \begin{bmatrix} \beta \\ \gamma \end{bmatrix} \qquad Q = \begin{bmatrix} X & 0 \\ X_p & I_g \end{bmatrix} \quad \text{e} \qquad w = \begin{bmatrix} y \\ T \end{bmatrix}$$

então, os E.M.Q.O. da regressão (2.1) serão dados por:

$$\hat{C} = (Q'Q)^{-1}Q'w = \begin{bmatrix} X'X + X_p'X_p & X_p' \\ X_p & I_g \end{bmatrix}^{-1} \begin{bmatrix} X'y + X_p'T \\ T \end{bmatrix}$$

$$= \begin{bmatrix} (X'X)^{-1} & -(X'X)^{-1}X_p' \\ -X_p(X'X)^{-1} & I_g + X_p(X'X)^{-1}X_p' \end{bmatrix} \begin{bmatrix} X'y + X_p'T \\ T \end{bmatrix}$$

$$= \begin{bmatrix} (X'X)^{-1}X'y \\ T - X_p\hat{\beta} \end{bmatrix}. \tag{3.2}$$

Pretende-se agora mostrar, nos (8) pontos que se seguem, de que forma a técnica das variáveis dummy substitui de forma bastante eficiente o processo standard de previsão ex-ante. No ponto (9) identifica-se um potencial inconveniente da técnica das variáveis dummy comparativamente ao método standard de previsão mas descreve-se, também,

uma forma de o contornar. Neste contexto, várias considerações ressaltam da especificação (3.2):

**(1)** $\hat{C}_i = \hat{\beta}_i$ (i = 1,2,...,k), isto é, os E.M.Q.O. dos coeficientes das k variáveis explicativas quantitativas originais são idênticos aos que obtêm ao estimar o modelo sem as g observações adicionais;

**(2)** A partir de (3.2) é fácil demonstrar que os resíduos associados às primeiras n observações são iguais aos que resultam de uma regressão que não considere as g observações extra (nem, consequentemente, as g variáveis dummy) e que os resíduos correspondentes às g restantes observações são nulos. De facto,

$$\begin{bmatrix} \hat{u} \\ \hat{u}_p \end{bmatrix} = \begin{bmatrix} y \\ T \end{bmatrix} - \begin{bmatrix} X & 0 \\ X_p & I_g \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} y - X\hat{\beta} \\ T - X_p\hat{\beta} - \hat{\gamma} \end{bmatrix}$$

ou, tendo em atenção a expressão que define $\hat{\gamma}$ em (3.2),

$$\begin{bmatrix} \hat{u} \\ \hat{u}_P \end{bmatrix} = \begin{bmatrix} y - X\hat{\beta} \\ 0 \end{bmatrix}; \qquad (3.3)$$

**(3)** Sendo nulo o vector dos resíduos associados às g observações extra, a Soma do Quadrado dos Resíduos (SQR) do modelo de regressão maior (3.1) é igual à SQR do modelo (2.1) que omite essas observações. Em resultado, o estimador da variância da variável residual é o mesmo em qualquer um dos modelos, uma vez que o número de graus de liberdade de cada um deles também é idêntico. De facto, A equação de regressão maior (3.1) tem simultaneamente mais g parâmetros e mais g observações do que a equação de regressão original (2.1).

Assim, se se definir o estimador da variância da variável residual do modelo expandido (3.1) como $\hat{\sigma}_c^2$ e $\hat{\sigma}_s^2$ como sendo o respectivo estimador no modelo (2.1), tem-se que:

$$\hat{\sigma}_c^2 = \frac{SQR}{(n+g)-(k+g)} = \frac{SQR}{n-k} \qquad (3.4)$$

e

$$\hat{\sigma}_s^2 = \frac{SQR}{n-k}; \qquad (3.5)$$

**(4)** Em consequência das expressões (3.4) e (3.5) produzirem o mesmo valor, a estimativa da variância dos E.M.Q.O. dos parâmetros associados às k variáveis explicativas originais é igual quer o modelo inclua ou não os g regressores dummy e as g observações adicionais. Em qualquer dos casos,

$$\hat{Var}(\hat{\beta}_j) = \hat{\sigma}^2 (X'X)^{-1}_{jj} \qquad j = 1,2,...,k \qquad (3.6)$$

em que $(X'X)^{-1}_{jj}$ representa o j-ésimo elemento da diagonal principal da matriz $(X'X)^{-1}$ (j = 1,2,...k). Assim, pode concluir-se que

$$\hat{Var}(\hat{C}_j) = \hat{Var}(\hat{\beta}_j) \qquad j = 1,2,...,k \qquad (3.7)$$

**(5)** Se T representar os verdadeiros valores da variável dependente relativos às g observações adicionais, então $C_{k+1}, C_{k+2}, ..., C_{k+g}$ traduzem os erros de previsão estimados inerentes a essas observações. Com efeito, se $T = y_p$ em que $y_p = X_p\beta + u_p$, tal como se definiu no início desta secção, o bloco inferior do vector (3.2) transforma-se em:

$$y_p - X_p \hat{\beta} = \hat{u}_p$$

o que corresponde ao vector de erros de previsão apresentado em (2.3);

**(6)** Para encontrar a matriz de variâncias estimadas dos erros de previsão estimados basta utilizar o segundo elemento (bloco) da diagonal principal da matriz $(Q'Q)^{-1}$ e pré-multiplicá-lo pelo valor estimado da variância da variável residual. De facto, a multiplicação destes dois elementos resulta em $\hat{\sigma}^2\left[I_g + X_p(X'X)^{-1}X_p'\right]$, o que corresponde à versão estimada da variância que se apresentou em (2.4). Este resultado pode posteriormente ser utilizado na construção de intervalos de confiança para previsão, tal como foram definidos em (2.5);

**(7)** Como (3.2) permite facilmente concluir, se se atribuir a todos os elementos do vector T o valor zero, o segundo bloco do vector de E.M.Q.O. devolverá o valor simétrico da previsão pontual para $y_p$, isto é, $\hat{\gamma} = -X_p\hat{\beta} = -\hat{y}_p$. O mesmo é dizer que, nesta circunstância, as previsões pontuais são indicadas através do valor simétrico dos E.M.Q.O. dos coeficientes das variáveis dummy;

**(8)** Para que os E.M.Q.O. dos coeficientes das variáveis dummy indiquem o próprio valor da previsão pontual, basta impor T = 0 no modelo (3.1), em que 0 é um vector nulo com g elementos, e definir as g variáveis dummy da seguinte forma:

$$D_j = \begin{cases} -1, & \text{para a } n+j \text{ é-sima observação } (j = 1,2,\ldots,g) \\ 0, & \text{caso contrário} \end{cases}$$

Neste caso particular, o modelo (3.1) converte-se em

$$\begin{bmatrix} y \\ 0 \end{bmatrix} = \begin{bmatrix} X & 0 \\ X_p & -I_g \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \end{bmatrix} + \begin{bmatrix} u \\ u_p \end{bmatrix} \tag{3.8}$$

de onde resulta, após algumas simplificações matemáticas,

$$\begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \boxed{\phantom{X}} = \begin{bmatrix} \left(X'X\right)^{-1}X'y \\ X_p\hat{\beta} \end{bmatrix}; \tag{3.9}$$

**(9)** A utilização do método das variáveis dummy ao invés do método standard de previsão ex-ante tem, no entanto, o inconveniente de alterar o valor do coeficiente de determinação da regressão. Este problema pode ser evitado se se afectar a T um conjunto de valores convenientes. Para um fácil entendimento desta afirmação, considere-se a expressão que define $R^2$ em cada um dos modelos, amplo e original, em que, mais uma

vez, os índices **c** e **s** são usados para distinguir o modelo com as g observações originais e g dummies do modelo sem essas observações e variáveis:

$$R_s^2 = 1 - \frac{\sum_{i=1}^{n+g} \hat{u}_i^2}{\sum_{i=1}^{n+g} (y_i - \bar{y})^2} \tag{3.10}$$

$$R_s^2 = 1 - \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sum_{i=1}^{n} (y_i - \bar{\bar{y}})^2}. \tag{3.11}$$

Em (3.10) e (3.11) $\bar{y}$ e $\bar{\bar{y}}$ representam a média da variável dependente em cada um dos modelos (3.1) e (2.1), respectivamente. Se se assumir, por uma questão de simplificação, que $g = 1$, o denominador da expressão (3.10) pode escrever-se em função do denominador da expressão (3.11) como se segue:

$$\sum_{i=1}^{n+1} \left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n} \left(y_i - \bar{\bar{y}}\right)^2 + \left(y_{n+1} - \bar{y}\right)^2 + n\left(\bar{\bar{y}} - \bar{y}\right)^2. \tag{3.12}$$

Esta igualdade encontra-se demonstrada em apêndice.

Uma vez que a SQR é idêntica nos dois modelos, tal como se mostrou no ponto **(3)**, as fórmulas (3.10) e (3.11) apenas diferem na expressão que se encontra no seu denominador, isto é na Soma dos Quadrados Total (SQT). Dada a relação (3.12), conclui-se que o denominador de (3.10) é maior do que o denominador de (3.11) a menos que $y_{n+1} = \bar{y}$. De facto, neste caso, (3.12) pode escrever-se como $\sum_{i=1}^{n+1} \left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n} \left(y_i - \bar{\bar{y}}\right)^2 + n\left(\bar{\bar{y}} - \bar{y}\right)^2$.

Em apêndice, demonstra-se também que

$$\bar{\bar{y}} = \frac{\bar{y}(n+1) - y_{n+1}}{n},$$

o que, no caso particular em que $y_{n+1} = \bar{y}$, se pode expressar como:

$$\bar{\bar{y}} = \frac{\bar{y}(n+1) - \bar{y}}{n} = \frac{\bar{y}(n+1-1)}{n} = \bar{y}.$$

Logo, nesta circunstância, $\sum_{i=1}^{n+1}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(y_i - \bar{\bar{y}}\right)^2$.

Consequentemente, o coeficiente de determinação de um modelo amplo do tipo (3.1) é, em geral, superior ao coeficiente de determinação do modelo (2.1) que não considera as g observações adicionais. Este efeito será ainda maior se vários regressores dummy forem incluídos no modelo uma vez que a relação (3.12) será aplicada sucessivamente a cada variável dummy adicional.

Portanto, para encontrar o valor do coeficiente de determinação do modelo original (2.1), a partir da estimação do modelo mais geral (3.1), basta impor a cada elemento que compõe o vector T o valor $\bar{y}$, como aliás Greene (1993) também refere.

## 4. Considerações finais

O método de Salkever, clarificado neste trabalho, possibilita que mediante a estimação de um único modelo de regressão que englobe todas as observações (as originais e as g extra) e g regressores dummy, seja possível encontrar de imediato as variâncias estimadas dos erros de previsão (através do produto da estimativa da variância da variável residual pelos elementos que se encontram no segundo bloco da diagonal principal da matriz $(Q'Q)^{-1}$) bem como o valor das próprias previsões pontuais (se se definir T = 0 e os regressores dummy usarem a codificação -1/0).

Paralelamente, deixou-se evidente que a utilização da técnica das variáveis dummy num contexto de previsão não altera a grande generalidade dos resultados standard da regressão nas observações originais, nomeadamente, o vector de E.M.Q.O. dos coeficientes das k variáveis explicativas quantitativas, a SQR do modelo, o erro padrão da variável residual, as variâncias estimadas dos E.M.Q.O. dos coeficientes das k variáveis independentes e a respectiva estatística t. No entanto, demonstrou-se também que a utilização deste método apresenta a propriedade indesejável de alterar o valor do coeficiente de determinação da regressão inicial restrita, problema este que pode, no entanto, ser evitado se se atribuir a cada elemento do vector T a média da variável dependente do modelo mais restrito.

## Referências

Greene, W. H. (1993), *Econometric Analysis*, 2ª edição, Nova York, Macmillan Publishing Company.

Maddala, G. S. (1992), *Introduction to Econometrics*, 2ª edição, Nova York, Macmillan Publishing Company.

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

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

Salkever, D. S. (1976), "The Use of Dummy Variables to Compute Predictions, Prediction Errors, and Confidence Intervals", *Journal of Econometrics*, Vol. 4, n. 4, pp 393-397.

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

# Apêndice

## Demonstração da Igualdade (3.12)

Isolando do somatório do primeiro membro da igualdade (3.12), o seu último elemento, vem

$$\sum_{i=1}^{n+1}\left(y_i-\bar{y}\right)^2=\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2+\left(y_{n+1}-\bar{y}\right)^2. \tag{a.1}$$

Observando (a.1), é fácil perceber que para demonstrar a igualdade (3.12) basta mostrar que

$$\sum_{i=1}^{n}\left(y_i-\bar{\bar{y}}\right)^2+n\left(\bar{\bar{y}}-\bar{y}\right)^2=\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2. \tag{a.2}$$

Assim, e a fim de provar a igualdade (a.2), pode agregar-se a sua segunda parcela no primeiro somatório da seguinte forma:

$$\sum_{i=1}^{n}\left(y_i-\bar{\bar{y}}\right)^2+n\left(\bar{\bar{y}}-\bar{y}\right)^2=\sum_{i=1}^{n}\left[\left(y_i-\bar{\bar{y}}\right)^2+\left(\bar{\bar{y}}-\bar{y}\right)^2\right]=$$

$$=\sum_{i=1}^{n}\left(y_i^2-2y_i\bar{\bar{y}}+2\bar{\bar{y}}^2-2\bar{\bar{y}}\bar{y}+\bar{y}^2\right). \tag{a.3}$$

Observe-se também que

$$\bar{\bar{y}}=\sum_{i=1}^{n}y_i\Big/n\Leftrightarrow\sum_{i=1}^{n}y_i=n\bar{\bar{y}} \tag{a.4}$$

e que

$$\bar{y}=\sum_{i=1}^{n+1}y_i\Big/n+1\Leftrightarrow\sum_{i=1}^{n}y_i=\bar{y}(n+1)-y_{n+1} \tag{a.5}$$

o que resulta em

$$n\bar{\bar{y}} = \bar{y}(n+1) - y_{n+1} \Leftrightarrow \bar{\bar{y}} = \frac{\bar{y}(n+1) - y_{n+1}}{n}. \qquad (a.6)$$

Substituindo (a.6) em (a.3), (a.3) transforma-se em

$$\sum_{i=1}^{n}\left[y_i^2 - 2y_i\left(\frac{\bar{y}(n+1) - y_{n+1}}{n}\right) + 2\left(\frac{\bar{y}(n+1) - y_{n+1}}{n}\right)^2 - 2\left(\frac{\bar{y}(n+1) - y_{n+1}}{n}\right)\bar{y} + \bar{y}^2\right]$$

ou, colocando $\left(\frac{\bar{y}(n+1) - y_{n+1}}{n}\right)$ em evidência,

$$\sum_{i=1}^{n}\left\{y_i^2 - \left(\frac{\bar{y}(n+1) - y_{n+1}}{n}\right)\left[2y_i - 2\left(\frac{\bar{y}(n+1) - y_{n+1}}{n}\right) + 2\bar{y}\right] + \bar{y}^2\right\}. \qquad (a.7)$$

Algumas manipulações algébricas simples permitem que expressemos (a.7) como

$$\sum_{i=1}^{n}\left\{y_i^2 - \frac{\bar{y}(n+1) - y_{n+1}}{n}\left(\frac{2y_i n - 2\bar{y}n - 2\bar{y} + 2y_{n+1} + 2\bar{y}n}{n}\right) + \bar{y}^2\right\}$$

ou ainda

$$\sum_{i=1}^{n}\left\{y_i^2 - 2\frac{\left(\bar{y}n + \bar{y} - y_{n+1}\right)\left(y_i n - \bar{y} + y_{n+1}\right)}{n^2} + \bar{y}^2\right\}. \qquad (a.8)$$

Para simplificar a expressão (a.8), é conveniente mostrar a seguinte igualdade:

$$\frac{\sum_{i=1}^{n}\left(\bar{y}n + \bar{y} - y_{n+1}\right)\left(y_i n - \bar{y} + y_{n+1}\right)}{n^2} = \sum_{i=1}^{n} y_i \bar{y}. \qquad (a.9)$$

Com efeito,

$$\frac{\sum_{i=1}^{n}\left(\bar{y}n+\bar{y}-y_{n+1}\right)\left(y_i n-\bar{y}+y_{n+1}\right)}{n^2}=\frac{\sum_{i=1}^{n}\left(\bar{y}(n+1)-y_{n+1}\right)\left(y_i(n+1)-y_i+y_{n+1}-\bar{y}\right)}{n^2}=$$

$$=\frac{\sum_{i=1}^{n}\left[\bar{y}y_i(n+1)^2-\bar{y}(n+1)y_i+\bar{y}(n+1)y_{n+1}-\bar{y}^2(n+1)-y_{n+1}y_i(n+1)+y_{n+1}y_i-y_{n+1}^2+y_{n+1}-\bar{y}\right]}{n^2}=$$

$$=\frac{\bar{y}(n+1)^2}{n^2}\sum_{i=1}^{n}y_i-\frac{\bar{y}(n+1)}{n^2}\sum_{i=1}^{n}y_i+\frac{\bar{y}(n+1)ny_{n+1}}{n^2}-\frac{\bar{y}^2(n+1)n}{n^2}-\frac{y_{n+1}(n+1)}{n^2}\sum_{i=1}^{n}y_i+$$
$$+\frac{y_{n+1}}{n^2}\sum_{i=1}^{n}y_i-\frac{y_{n+1}^2 n}{n^2}+\frac{y_{n+1}\bar{y}n}{n^2}=$$

$$=\left[\frac{\bar{y}(n+1)^2}{n^2}-\frac{\bar{y}(n+1)}{n^2}-\frac{y_{n+1}(n+1)}{n^2}+\frac{y_{n+1}}{n^2}\right]\sum_{i=1}^{n}y_i+$$
$$+\frac{1}{n}\left[\bar{y}(n+1)y_{n+1}-\bar{y}^2(n+1)-y_{n+1}^2+y_{n+1}\bar{y}\right] \qquad \text{(a.10)}$$

Substituindo $\bar{y}$ por $\sum_{i=1}^{n+1}y_i \Big/ n+1$, a expressão (a.10) é equivalente a

$$\left[\frac{\bar{y}(n+1)^2}{n^2}-\frac{\bar{y}(n+1)}{n^2}-\frac{y_{n+1}(n+1)}{n^2}+\frac{y_{n+1}}{n^2}\right]\sum_{i=1}^{n}y_i+$$
$$+\frac{1}{n}\left[(n+1)y_{n+1}\frac{\sum_{i=1}^{n+1}y_i}{n+1}-(n+1)\frac{\left(\sum_{i=1}^{n+1}y_i\right)^2}{(n+1)^2}-y_{n+1}^2+y_{n+1}\frac{\sum_{i=1}^{n+1}y_i}{n+1}\right]$$

ou, atendendo a que $\sum_{i=1}^{n+1} y_i = \sum_{i=1}^{n} y_i + y_{n+1}$

$$\left[\frac{\bar{y}(n+1)^2}{n^2} - \frac{\bar{y}(n+1)}{n^2} - \frac{y_{n+1}(n+1)}{n^2} + \frac{y_{n+1}}{n^2}\right]\sum_{i=1}^{n} y_i +$$
$$+\frac{1}{n}\left[y_{n+1}\left(\sum_{i=1}^{n} y_i + y_{n+1}\right) - \bar{y}\left(\sum_{i=1}^{n} y_i + y_{n+1}\right) - y_{n+1}^2 + y_{n+1}\frac{\sum_{i=1}^{n} y_i + y_{n+1}}{n+1}\right]$$

que pode ainda simplificar-se da seguinte forma

$$\left[\frac{\bar{y}(n+1)^2}{n^2} - \frac{\bar{y}(n+1)}{n^2} - \frac{y_{n+1}(n+1)}{n^2} + \frac{y_{n+1}}{n^2}\right]\sum_{i=1}^{n} y_i +$$
$$+\frac{1}{n}\left[y_{n+1}\sum_{i=1}^{n} y_i + y_{n+1}^2 - \bar{y}\sum_{i=1}^{n} y_i - \bar{y}y_{n+1} - y_{n+1}^2 + y_{n+1}\frac{\sum_{i=1}^{n} y_i}{n+1} + \frac{y_{n+1}^2}{n+1}\right]. \qquad \text{(a.11)}$$

Atendendo, mais uma vez, a que $\bar{y} = \left(\sum_{i=1}^{n} y_i + y_{n+1}\right) \Big/ n+1$, a expressão (a.11) pode expressar-se como:

$$\left[\frac{\bar{y}(n+1)^2}{n^2} - \frac{\bar{y}(n+1)}{n^2} - \frac{y_{n+1}(n+1)}{n^2} + \frac{y_{n+1}}{n^2}\right]\sum_{i=1}^{n} y_i +$$
$$+\frac{1}{n}\left[y_{n+1}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} y_i - \frac{y_{n+1}\sum_{i=1}^{n} y_i}{n+1} - \frac{y_{n+1}^2}{n+1} + \frac{y_{n+1}\sum_{i=1}^{n} y_i}{n+1} + \frac{y_{n+1}^2}{n+1}\right]$$

ou, colocando $\sum_{i=1}^{n} y_i$ em evidência,

$$\sum_{i=1}^{n} y_i\left[\frac{\bar{y}(n+1)^2}{n^2} - \frac{\bar{y}(n+1)}{n^2} - \frac{y_{n+1}(n+1)}{n^2} + \frac{y_{n+1}}{n^2} + \frac{y_{n+1}}{n} - \frac{\bar{y}}{n}\right]$$

e, reduzindo ao mesmo denominador e retirando parêntesis,

$$\sum_{i=1}^{n} y_i \left[ \frac{\bar{y}n^2 + \bar{y} + 2\bar{y}n - \bar{y}n - \bar{y} - y_{n+1}n - y_{n+1} + y_{n+1} + ny_{n+1} - \bar{y}n}{n^2} \right]. \qquad \text{(a.12)}$$

Efectuando as simplificações necessárias, (a.12) é equivalente a

$$\sum_{i=1}^{n} y_i \left( \frac{\bar{y}n^2}{n^2} \right)$$

e, finalmente,

$$\sum_{i=1}^{n} y_i \bar{y} \qquad \text{(a.13)}$$

o que comprova a igualdade (a.9). Consequentemente, (a.8) pode escrever-se como

$$\sum_{i=1}^{n} \left( y_i^2 - 2y_i \bar{y} + \bar{y}^2 \right)$$

que, por sua vez é equivalente a

$$\sum_{i=1}^{n} \left( y_i - \bar{y} \right)^2 .$$

Deste modo, a igualdade (a.2) fica provada e, consequentemente, a igualdade (3.12) também.