# Variáveis dummy: especificações de modelos com parâmetros variáveis

Fabrício Missio[1], Luciane Flores Jacobi[2]

*[1]Curso de Ciências Econômicas/Universidade Estadual de Mato Grosso do Sul*
*E-mail: fabriciomissio@gmail.com*
*[2]Departamento de Estatística - CCNE/UFSM*
*E-mail: lfjacobi@ccne.ufsm.br*

**Resumo**

O presente trabalho busca estudar a regressão sobre variáveis dummy, mais especificamente, revisar a teoria e os casos em que elas podem ser utilizadas, a fim de elaborar, de forma sucinta, um material simples e abrangente sobre o assunto, capaz de auxiliar em pesquisas e trabalhos. Após a formalização, apresentou-se os procedimentos operacionais para executá-los, com ajuda de recursos computacionais, utilizando-se o software Statistica versão 5.1.
Palavras-chave: Regressão, Variável Dummy, Software Estatístico.

**Abstract**

The present work search to study the regression on variables dummies. More specifically, to revise the theory and the cases where they can be used, in order to elaborate, in a succinct form, a simple and including material on the subject, capable to assist in research and works. After the formalization, will present the operational procedures to execute them, with aid of computational resources, using in such a way, the Statistica software version 5.1.
Key-Words: Regression, Variable Dummy, Statistica Software.

## 1. Introdução

Na análise de regressão, a variável dependente pode ser influenciada por variáveis quantitativas e qualitativas. As variáveis quantitativas são facilmente mensuradas em alguma escala o que não ocorre com as variáveis qualitativas, uma vez que essas indicam a presença ou a ausência de uma qualidade ou atributo.

Dessa forma, um método para "quantificar" esses atributos é construir variáveis artificiais que assumam valores de 1 ou 0 ( indicando ausência de um atributo e indicando a sua presença) que são conhecidas pela literatura existente de "variáveis dummy". A rigor, não é essencial que as variáveis dummy assumam os valores de 0 e 1. O par (0,1) pode ser transformado em qualquer outro par por uma função linear tal que $Z = a + bD$ (b ≠ 0) em que a e b são constantes e em que $D = 1$ ou 0. Quando $D = 1$ , tem-se $Z = a + b$; e quando $D = 0$, tem-se $Z = a$. Assim, o par (0,1) se torna (a, a + b). Observa-se que a atribuição de valores é puramente arbitrária, exigindo cuidado na hora de interpretar os resultados.

A introdução de variáveis qualitativas (dummy) torna o modelo de regressão linear uma ferramenta extremamente flexível capaz de lidar com muitos problemas encontrados, principalmente, em estudos empíricos. Os modelos que incluem como variáveis explicativas somente variáveis qualitativas são chamados de modelos de análise de variância (ANOVA), enquanto que os que incluem também variáveis quantitativas são chamados de modelos de análise de covariância (ANCOVA). Do ponto de vista econômico, as variáveis dicotômicas dummy são introduzidas no modelo para representar adequadamente os efeitos diferenciais produzidos pelo comportamento dos agentes (econômicos) devido, principalmente, a diferentes causas, dentre as quais se destacam as de tipo temporal (estacionárias, etc), de caráter espacial (estado, país, etc), de caráter puramente qualitativo (sexo, etc).

Quanto à sua aplicação, este tipo de variável pode ser usado em modelos simples, em que a única variável explicativa é a própria dummy, e em modelos mais complexos, em que uma variável categórica é desdobrada em duas ou mais variáveis dummies. Atenção especial requer a especificação de modelos que combinam dummies para diferentes categorias e para modelos que combinam dummies e variáveis quantitativas. Neste último caso, duas análises são possíveis: incorporar mudanças no intercepto e/ou na declividade de uma função; possibilitar a identificação de mudanças estruturais.

A literatura especializada referente à abordagem da análise de regressão sobre variáveis dummy desenvolveu-se, principalmente, a partir das décadas de 70 e 80 do século passado, embora já tenha sido objeto de estu-

dos há muitos anos. Do ponto de vista de uma ordem cronológica, tem-se como referência os estudos de SUITS (1957, 1984), CHOW (1960), GUJARATI (1970 a, b) KOOYMAN (1976), ERLAT (1978, 1985), DUFOUR (1980, 1981, 1982), KENNEDY (1986) e, mais recentemente, STEWART (1991), MADDALA (1992), GREENE (1993), HARDY (1993), dentre outros.

Observa-se que existe uma série de textos relacionados à utilização das variáveis dummy na análise de regressão. Entretanto, este tópico ressente de um maior número de publicações, no sentido de que as contribuições individuais (explorada em cada texto) passem a ser incorporadas em uma teoria mais completa do que as apresentadas nos livros textos de econometria principalmente, porque as demonstrações e análises ficam a posteriori prejudicadas pela inacessibilidade à grande parte destes materiais.

O objetivo do presente trabalho é desenvolver um estudo teórico-prático sobre a utilização de variáveis dummy e suas principais aplicações. De forma sucinta, elaborar um material simples e abrangente a fim de destacar os casos em que as variáveis dummy são aplicadas e apresentar alguns resultados dessas aplicações usando o programa computacional Statistica versão 5.1. A metodologia a ser utilizada para apresentação dos resultados corresponde à dos livros texto de econometria, tais como, GUJARATI (2000), MADDALA (2003) e HILL (1999).

Este trabalho constará, além desta introdução, de cinco seções onde se apresenta: na segunda o método de estimação sob variáveis dummy em modelos com variações descontínuas nos parâmetros; na terceira, a estimação em modelos com variações contínuas; na quarta apresenta as regressões com variáveis dummy sob modelos de análise de variância (ANOVA) e covariância (ANCOVA); e na quinta, como exemplo, o desenvolvimento destes modelos no programa computacional Statistica versão 5.1. As considerações finais estarão na última seção.

## 2. Utilização de variáveis dummy: o caso de variações descontínuas nos parâmetros

Nesta sessão busca-se demonstrar, baseado em REBELO & VALLE (2002), quando a utilização de variáveis dummy torna-se importante na análise econométrica. Para tanto, considera-se como exemplo um estudo (de caráter espacial, temporal ou puramente qualitativo) em que, num primeiro momento, se assume que a relação entre a variável dependente e a variável explicativa é estável para todas as observações de uma amostra. Ou seja;

$$Y_i = \alpha + \beta X_i + u_i;\ i = 1, 2, \ldots n;,\ u_i \sim NID(0, \sigma^2) \qquad (2.1)$$

onde $X_i$ é uma variável quantitativa.

Supõe-se que a estimação desta equação de regressão apresenta um valor significativo para a estatística $t$ associada ao coeficiente da variável $X$ e, simultaneamente, um valor para o coeficiente de determinação $(R^2)$ relativamente baixo e/ou um valor do Durbin Watson ($D.W.$) longe de 2. A análise destes resultados leva à conclusão de que, embora $X$ constitua uma variável importante na determinação do comportamento da variável $Y$, existe uma parcela relativamente alta do comportamento desta variável que não é explicada pelo modelo. Em outras palavras, o modelo descrito anteriormente pode encontrar-se mal especificado por incorreta omissão de variáveis explicativas.

O gráfico a seguir expressa a relação entre a variável dependente e o regressor $(X)$ retratado pela Figura 1:

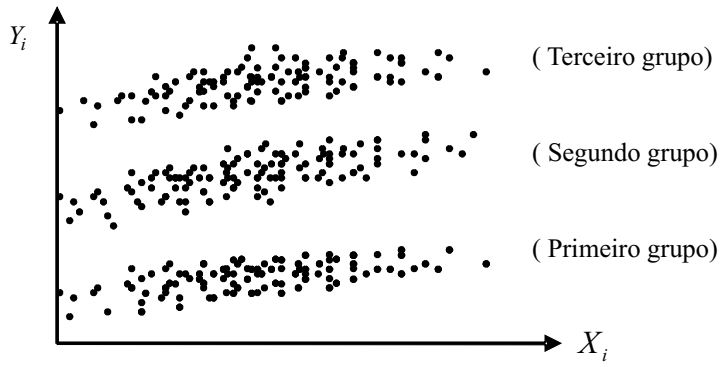



**Figura 1**: Formato das observações do estudo econométrico.

Observa-se, pela Figura 1, que o comportamento das variáveis está relacionado positivamente. Entretanto, parece existir uma relação distinta entre as duas variáveis para as observações que pertencem a cada um dos grupos. Isso explicaria os resultados obtidos quando do ajustamento de uma única reta de regressão para o conjunto de dados, ou seja, o valor relativamente baixo de $R^2$. Neste caso, o ajustamento de única reta traduz-se em uma estimativa de valor elevado para a variância da variável resi-

dual o que pode produzir uma estatística $t$ associada ao coeficiente $X$ não significativa.

Além disso, cabe explicitar que o ajuste do modelo utilizando apenas a variável $X$ como independente significa a omissão de uma informação conhecida, ou seja, a não utilização de uma variável que indica os diferentes grupos onde foram tomadas as observações. A inclusão desta variável, neste caso, significa a incorporação de duas variáveis dummy no modelo.

Para resolver este problema, considera-se, em separado, cada um dos grupos de observações e utiliza-os em três modelos distintos, pois, como mostra a Figura 1, as retas de regressão que melhor se ajustam aos dados parecem diferir apenas no termo intercepto $(\alpha)$ e não na inclinação $(\beta)$. Em termos formais;

$$Y_i = \alpha_1 + \beta X_i + u_i \qquad \text{para o primeiro grupo} \qquad (2.2)$$

$$Y_i = \alpha_2 + \beta X_i + u_i \qquad \text{para o segundo grupo} \qquad (2.3)$$

$$Y_i = \alpha_3 + \beta X_i + u_i \qquad \text{para o terceiro grupo} \qquad (2.4)$$

Contudo, a estimação dos três diferentes modelos certamente não produzirá o mesmo valor para o parâmetro $\beta$ que, para efeito de análise, foi considerado comum a ambas as especificações, pois se os três grupos reagem de forma similar a uma variação em $X$, deve-se reunir todas as observações para ajustar um modelo de regressão que produza três termos independentes, mas uma estimativa única para o coeficiente de inclinação. Dessa forma, a definição de regressores dummy apresenta-se como o procedimento adequado para este caso.

Em termos formais, a definição de variáveis seria a seguinte:

$$D_{2i} = \begin{cases} 1, & \text{se a observação verifica a característica que define o segundo grupo;} \\ 0, & \text{caso contrário} \end{cases}$$

$$D_{3i} = \begin{cases} 1, & \text{se a observação verifica a característica que define o terceiro grupo;} \\ 0, & \text{caso contrário;} \end{cases}$$

onde a introdução da variável dummy $D_2$ tem por objetivo captar (e o valor dela representa) a diferença entre os termos independentes das equações de regressão relativas aos dois primeiros grupos. De forma análoga, a

dummy $D_3$ refere-se às diferenças existentes entre o terceiro e primeiro grupo[1].

Logo, com a introdução de regressores dummy pode-se ajustar a equação de regressão da seguinte forma;

$$Y_i = \alpha_1 + (\alpha_2 - \alpha_1)D_{2i} + (\alpha_3 - \alpha_1)D_{3i} + \beta X_i + u_i ; \quad i = 1,2,....,n;$$

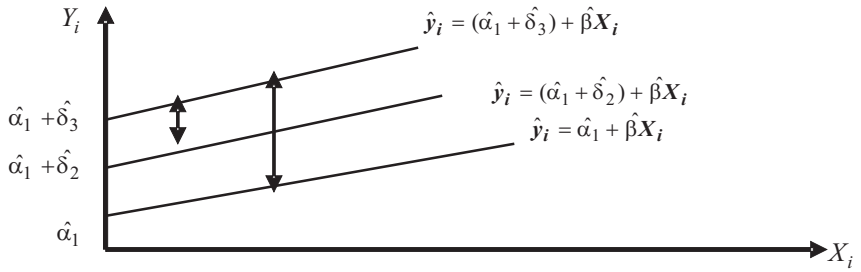
$$u_i \sim NID(0,\sigma^2) \qquad (2.5)$$

ou, de forma equivalente;

$$Y_i = \alpha_1 + \delta_2 D_{2i} + \delta_3 D_{3i} + \beta X_i + u_i ; \; i = 1,2,....,n; \; u_i \sim NID(0,\sigma^2) \qquad (2.6)$$

onde: $\delta_2 = (\alpha_2 - \alpha_1)$ e $\delta_3 = (\alpha_3 - \alpha_1)$.

Pela equação anterior é possível se obter uma única estimativa para o parâmetro $\beta$ e, simultaneamente, três ordenadas, na origem, distintas. A tradução geométrica da estrutura estimada neste modelo pode ser representada conforme a Figura 2:




Fonte: REBELO & VALLE (2002)
**Figura 2:** Estrutura geométrica do modelo (2.6).

Observa-se que na Figura 2 tem-se que $\delta_2 > 0, \delta_3 > 0, \delta_3 > \delta_2$. Neste caso, para cada grupo o modelo de regressão seria dado por:

$$Y_i = \alpha_1 + \beta X_i + u_i \text{ se } D_{2i} = D_{3i} = 0 \text{ (primeiro grupo)} \quad (2.7)$$

$$Y_i = (\alpha_1 + \delta_2) + \beta X_i + u_i \text{ se } D_{2i} = 1 \text{ e } D_{3i} = 0 \text{ (segundo grupo)} \quad (2.8)$$

[1]O grupo, categoria ou classificação designado pelo valor 0 é freqüentemente referido como categoria-base. É "base" no sentido de que as comparações são feitas em relação a esta categoria.

$$Y_i = (\alpha_1 + \delta_3) + \beta X_i + u_i \text{ se } D_{2i} = 0 \text{ e } D_{3i} = 1 \text{ (terceiro grupo)} \quad (2.9)$$

Admitiu-se que o coeficiente de inclinação é semelhante a todos os modelos. Neste caso, vale ressaltar que se considera como hipótese implícita, que as variáveis de intercepto sejam aditivas. O efeito de cada fator qualitativo é somado ao intercepto de regressão, e o efeito de qualquer variável binária é independente de qualquer outro fator qualitativo. Às vezes, é possível que os efeitos de fatores qualitativos não sejam independentes, isto é, o modelo pode ser multiplicativo. Em outras palavras, pode haver interação entre as variáveis qualitativas.

A situação oposta também pode ocorrer. As retas de regressão podem ter o mesmo intercepto com coeficientes de inclinação distintos. Dessa forma, as retas de regressão que representam essas especificações, para cada um dos diferentes grupos, devem ser especificadas novamente, como a seguir:

$$Y_i = \alpha + \beta_1 X_i + u_i \quad \text{(para o primeiro grupo)} \quad (2.10)$$

$$Y_i = \alpha + \beta_2 X_i + u_i \quad \text{(para o segundo grupo)} \quad (2.11)$$

$$Y_i = \alpha + \beta_3 X_i + u_i \quad \text{(para o terceiro grupo)} \quad (2.12)$$

Segundo Hill et all. (2003) o produto de uma variável dummy por uma variável contínua resulta no que se pode chamar de "variável dummy de inclinação e/ou variável de interação", e é recomendável, a fim de que se possa, em um único modelo, produzir uma estimativa para o termo independente e três coeficientes de inclinação distintos. O modelo que deve ser estimado é:

$$Y_i = \alpha + \beta_1 X_i + (\beta_2 - \beta_1)(D_{2i} X_i) + (\beta_3 - \beta_1)(D_{3i} X_i) + u_i \ ;$$

$$i = 1, 2 \ldots, n \quad u_i \sim NID(0, \sigma^2) \quad (2.13),$$

onde as variáveis $D_2$ e $D_3$ são as variáveis dummy definidas anteriormente e medem, portanto, a diferença entre os declives de dois modelos de regressão.

Ou, de forma equivalente, o modelo anterior pode ser representado por:

$$Y_i = \alpha + \beta_1 X_i + \gamma_2 (D_{2i} X_i) + \gamma_3 (D_{3i} X_i) + u_i \qquad i = 1, 2, \ldots, n;$$

$$u_i \sim NID(0, \sigma^2); \quad (2.14),$$

onde: $\gamma_2 = (\beta_2 - \beta_1)$ e $\gamma_3 = (\beta_3 - \beta_1)$.

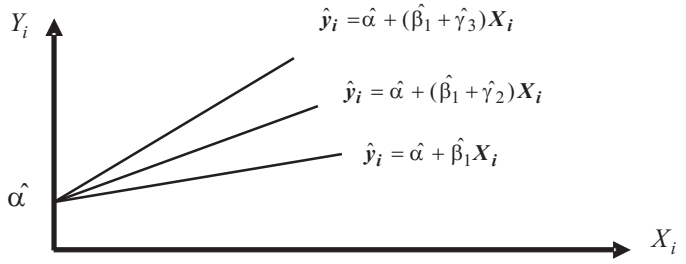
Assim, a utilização de variáveis dummy permite determinar a equação relativa a cada grupo[2] ;

$$Y_i = \alpha + \beta_1 X_i + u_i \text{ se } D_{2i} = D_{3i} = 0 \quad \text{(primeiro grupo)} \qquad (2.15)$$

$$Y_i = \alpha + \beta_2 X_i + u_i \text{ se } D_{2i} = 1 \text{ e } D_{3i} = 0 \quad \text{(segundo grupo)} \qquad (2.16)$$

$$Y_i = \alpha + \beta_3 X_i + u_i \text{ se } D_{2i} = 0 \text{ e } D_{3i} = 1 \quad \text{(terceiro grupo)} \qquad (2.17)$$

A representação gráfica desta situação é mostrada pela Figura 3, onde $\gamma_2 > 0, \gamma_3 > 0, \gamma_3 > \gamma_2$ :



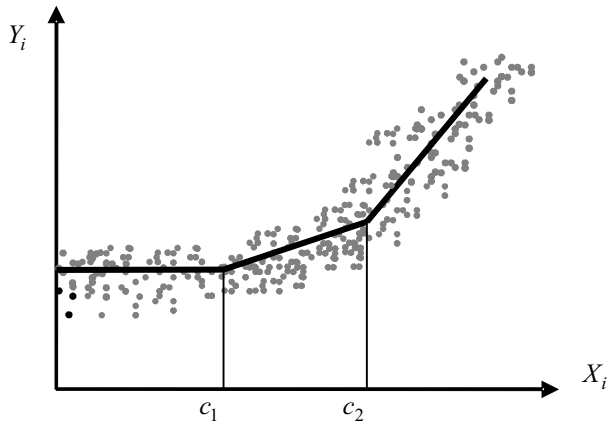

Fonte: REBELO & VALLE (2002)
**Figura 3:** Estrutura geométrica do modelo (2.14).

Observa-se que esta análise pode ser combinada. Pode-se estimar uma equação de regressão com diferentes interceptos e diferentes coeficientes de inclinação.

## 3. Utilização de variáveis dummy: o caso de variações contínuas nos parâmetros

Nesta classe de modelos a variação no declive da reta de regressão não é descontínua, embora seja bastante acentuada. Por exemplo, pretende-se ajustar um modelo de regressão para os dados ilustrados na Figura 4;

[2]Outra forma de se obter esses resultados é calcular as derivadas parciais (ver Maddala 2003).



Fonte: REBELO & VALLE (2002)
**Figura 4:** Estudo Econométrico Hipotético.

Efetivamente, pela simples ilustração da Figura 4 é possível se observar que a estimação do modelo do tipo $Y_i = \alpha + \beta X_i + u_i$ não é correta. A melhor forma de se obter uma estimativa que represente os dados é ajustar os seguintes modelos de regressão [esse método é apresentado em alguns livros textos como regressão linear por partes (Gujarati, 2000)]:

$$Y_i = \alpha_1 + \beta_1 X_i + u_i \quad \text{para} \quad X_i \leq c_1 \tag{3.1}$$

$$Y_i = \alpha_2 + \beta_2 X_i + u_i \quad \text{para} \quad c_1 \leq X_i \leq c_2 \tag{3.2}$$

$$Y_i = \alpha_3 + \beta_3 X_i + u_i \quad \text{para} \quad X_i \geq c_2 \tag{3.3}$$

Desse modo podem ser definidas duas variáveis dummy, a saber:

$$D_{2i} = \begin{cases} 1, & \text{Se } c_1 \leq X_i \leq c_2 \\ 0, & \text{Caso contrário} \end{cases}$$

$$D_{3i} = \begin{cases} 1, & \text{Se } X_i \geq c_2\ ; \\ 0, & \text{Caso contrário;} \end{cases}$$

Logo, a especificação do modelo correta seria semelhante a que se segue, permitindo uma aproximação adequada do problema em questão:

$$Y_i = \alpha_1 + (\alpha_2 - \alpha_1)D_{2i} + (\alpha_3 - \alpha_1)D_{3i} + \beta_1 X_1 + (\beta_2 - \beta_1)(D_{2i}X_i) + (\beta_3 - \beta_1)(D_{3i}X_i) + u_i$$
$$i = 1,2, ..., n; u_i \sim NID(0, \sigma^2) \quad (3.4)$$

Entretanto, o problema resultante da estimação deste modelo é que para $E(Y_i / X_i = c_1)$ e $E(Y_i / X_i = c_2)$ obtêm-se dois valores diferentes. Em síntese, isso significa que o modelo de regressão não é contínuo, para torná-lo contínuo, é necessário definir as seguintes restrições lineares:

$$\alpha_1 + \beta_1 c_1 = \alpha_2 + \beta_2 c_1$$

$$\alpha_2 + \beta_2 c_2 = \alpha_3 + \beta_3 c_2$$

que, por simples manipulações algébricas, tornam-se:

$$\alpha_2 - \alpha_1 = -c_1(\beta_2 - \beta_1)$$

$$\alpha_3 - \alpha_2 = -c_2(\beta_3 - \beta_2)$$

Impondo-se essas restrições ao modelo (3.4), ele se transforma em;

$$Y_i = \alpha_1 + \beta_1 X_1 + (\beta_2 - \beta_1)D_{2i}(X_i - c_1) + (\beta_3 - \beta_1)D_{3i}(X_i - c_2) + u_i \quad (3.5)$$

ou, de forma equivalente;

$$Y_i = \alpha_1 + \beta_1 X_1 + \gamma_2 D_{2i}(X_i - c_1) + \gamma_3 D_{3i}(X_i - c_2) + u_i \quad (3.6)$$

Logo, para cada segmento da reta de regressão, têm-se as seguintes combinações de valores das variáveis dummy, observando-se que para o primeiro caso a reta de regressão vale para o intervalo $X_i \le c_1$, a segunda para o intervalo $c_1 \le X_i \le c_2$ e a terceira para o intervalo $X_i \ge c_2$.

$$Y_i = \begin{cases} \alpha_1 + \beta_1 X_1 + u_i \\ \alpha_1 - c_1\gamma_2 + (\beta_1 + \gamma_2)X_1 + u_i \\ \alpha_1 - c_2\gamma_3 + (\beta_1 + \gamma_3)X_i + u_i \end{cases} \quad (3.7)$$

Observe que neste caso, o processo de estimação produzirá um coeficiente inclinação distinto para cada uma das categorias. A significância das mudanças estimadas nos declives pode ser testada por um teste $F$.

Na classe de modelos em que a variação no declive da reta de re-

gressão é descontínua, e sendo esta bastante acentuada, a especificação de um único modelo, tal como proposto em (3.6), pode ser uma aproximação inadequada para o problema em questão. Neste caso, a especificação de um modelo como o apresentado em (3.4) solucionaria este problema, muito embora a resultante fosse três equações de regressão totalmente distintas e dois valores totalmente diferentes para $E(Y_i/X_i = c_1)$ e $E(Y_i/X_i = c_2)$. Observe, portanto, que a técnica apresentada anteriormente, de impor restrições ao modelo de tal forma a torná-lo continuo, busca solucionar através da especificação de um modelo aproximado o problema da obtenção de diferentes retas de regressão, uma vez que sua aplicação proporciona como resultado retas de regressão em que apenas o coeficiente de inclinação difere entre elas[3].

## 4. Regressão com variáveis dummy sob modelos de análise de variância (ANOVA) e covariância (ANCOVA)

Antes de serem analisados os modelos de variância e covariância separadamente, admite-se que as retas de regressão para os distintos grupos diferem apenas no termo de intercepto, mantendo-se os mesmos coeficientes angulares, conforme pode ser observado na Figura 1 apresentado anteriormente. Neste caso, a variável dummy é incorporada ao modelo de regressão para captar o efeito do deslocamento do intercepto como resultado de algum fator qualitativo.

Para exemplificar a primeira classe de modelos em que as variáveis explicativas são exclusivamente dummies, apresenta-se o seguinte modelo de regressão (4.1) (Gujarati, 2000); onde através do uso de variáveis dummy busca-se identificar se existe diferença entre os salários médios recebidos por professores e professoras universitários. A hipótese implícita deste modelo é de que os professores universitários receberiam um salário maior. Neste caso, mantidos constantes todos os demais fatores, caso a diferença se confirme, pode-se especular sobre a possibilidade de haver discriminação com relação ao salário pago às professoras.

[3]Observe, contudo, que se nos distintos pontos de mudança a descontinuidade da reta for muita acentuada, a estimação através de um modelo aproximado pode gerar uma estimativa incorreta. Neste caso, solucionar-se-iam os problemas com a obtenção de distintas retas de regressão, muito embora as estimativas obtidas não representem fielmente os dados que estão sendo analisados.

O modelo de regressão é dado por;

$$y_i = \alpha + \beta D_i + u_i \qquad (4.1)$$

onde: $y =$ salário anual de um professor universitário; $D_i = 1$ se do sexo masculino, 0 caso contrário.

Admitindo-se que as perturbações satisfaçam as hipóteses usuais do modelo clássico de regressão linear $\left[u_i \sim NID\left(0,\sigma^2\right); E\left(u_i\right)=0\right]$, tem-se, de (4.1);

- Salário-médio de uma professora universitária: $E\left(Y_i/D_i=0\right)=\alpha$

- Salário-médio de um professor universitário: $E\left(Y_i/D_i=1\right)=\alpha+\beta$

onde o coeficiente de inclinação $\beta$ informa em quanto o salário médio de um professor universitário difere do salário-médio de uma professora.

Caso os resultados obtidos mostrem que $\beta$ é estatisticamente significativo, conclui-se que, o salário de um professor, de fato, é superior ao de uma professora. Graficamente, este resultado pode ser apresentado como na Figura 5.

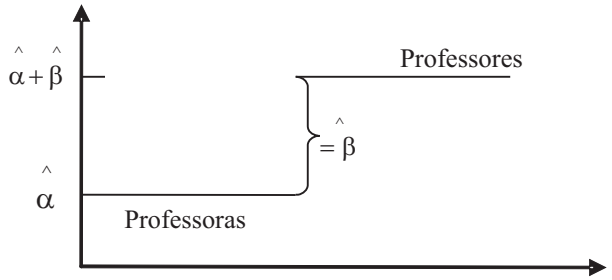



**Figura 5:** Funções salários (médios).

*4.1 Estimação e teste de parâmetros do modelo*

A estimação do modelo na forma matricial requer que os seguintes cálculos sejam efetuados[4]:

[4]A matriz $X$ representa a matriz dos coeficientes do modelo. Neste caso, $X = D$. Observa-se que, especificando-se adequadamente a matriz $X$, os resultados apresentados valem para todos os modelos abordados ao longo deste trabalho.

a) Calcular $X'X$ (onde $X'$ = Matriz transposta de X, matriz original);

b) Obter $\left(X'X\right)^{-1}$;

c) Calcular $X'Y$;

De posse desses cálculos pode-se obter as estimativas dos parâmetros do modelo de regressão através da fórmula:

$$\hat{B} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \left(X'X\right)^{-1} X'Y$$

Contudo, falta avaliar se o valor estimado para o parâmetro $\beta$ é estatisticamente significativo, isto é, testar em nível de significância de $\alpha\%$, a hipótese $H_0 : \beta = 0$ contra a hipótese alternativa $H_1 : \beta > 0$[5]. Para tanto, calcula-se a análise de variância de regressão, apresentada genericamente na Tabela 1;

**Tabela 1:** Análise de Variância.

| CV | GL | SQ | QM |
|---|---|---|---|
| Regressão | $K-1$ | $\hat{b}' X'y - n\bar{Y}^2$ | $\hat{b}' X'y - n\bar{Y}^2 \Big/ K-1$ |
| Resíduo | $n-K$ | $y'y - \hat{b}' X'y$ | $y'y - \hat{b}' X'y \Big/ n-K$ |
| Total | $n-1$ | $y'y - n\bar{Y}^2$ | |

OBS: $K$ é o número de parâmetros estimados. Para o caso do modelo (4.1) $K = 2$.

A partir da análise de variância pode-se determinar o grau de ajuste do modelo, bem como realizar o teste de significância dos parâmetros. Logo, o coeficiente de determinação (ou grau de ajuste do modelo) é dado por $R^2 = \dfrac{SQE/K-1}{SQT/n-1}$.

[5]Observa-se que a hipótese alternativa pode ser especificada de forma diferente, como por exemplo, $H_1 : \beta \neq 0$. Neste caso, ela foi definida como sendo maior que zero dado a hipótese implícita de que poderia haver um diferencial positivo entre os salários médios recebidos pelos professores universitários em comparação aos salários recebidos pelas professoras universitárias.

A realização do teste de significância dos parâmetros exige que se calcule a variância do modelo e, posteriormente, a matriz de variância-covariância de onde se obtém, para cada parâmetro individualmente, a sua respectiva variância (diagonal principal).

O teste $F$ de significância global para os parâmetros de regressão pode ser calculado por: $F = \dfrac{\hat{b}' X' y - n\bar{y}^2 / K-1}{y' y - \hat{b}' X' y / n-K}$. Entretanto, neste caso devem-se especificar novamente as hipóteses, ou seja, este teste é utilizado para testar a hipótese de que todos os coeficientes de inclinação são simultaneamente iguais a zero contra a hipótese alternativa que pelo menos um dos coeficientes é diferente de zero.

Cálculo da variância: $\hat{\sigma}^2 = \dfrac{SQR}{n-K}$. Logo, a matriz de variância-covariância para $\hat{B}$ pode ser mostrada como:

$$\text{var}-\text{cov}\left(\hat{b}\right) = \hat{\sigma}^2 \left(X' X\right)^{-1} = \begin{bmatrix} \text{var}(\alpha) & \text{cov}(\alpha, \beta) \\ \text{cov}(\alpha, \beta) & \text{var}(\beta) \end{bmatrix}$$

Sabe-se que os elementos da diagonal principal representam as variâncias de $\hat{\alpha}$ e $\hat{\beta}$, respectivamente, e suas raízes quadradas fornecem os correspondentes erros padrões. De posse destes dados pode-se, utilizar o teste $t$ para testar a significância dos parâmetros.

Para $\hat{\beta}$, tem-se que:

$$t = \frac{\hat{\beta} - \beta}{ep\left(\hat{\beta}\right)} \qquad (4.2)$$

Logo, pela regra de decisão sabe-se que, $t_{cal} > t_{crítico}$ $\left(t_{\alpha/2,(n-K)}\right)$, rejeita-se a hipótese nula. Neste caso, se o valor observado da estatística calculada superar o seu valor crítico, não se pode afirmar que o coeficiente $\hat{\beta}$ é estatisticamente igual a zero, ou seja, isso significa que os resultados indicam que os salários-médios das duas categorias são diferentes.

Para a segunda classe de modelos, onde as variáveis explicativas são tanto de ordem qualitativas (dummy) como quantitativas, inicia-se o

estudo da regressão sobre uma variável quantitativa e uma variável qualitativa com apenas duas classes[6]. Para exemplificar, apresenta-se o seguinte modelo de regressão:

$$Y_i = \alpha_i + \alpha_2 D_1 + \beta X_i + u_1 \qquad (4.3)$$

onde: $X_i =$ anos de experiência de ensino e as demais variáveis seguem como antes definidas.

Admitindo-se que as perturbações satisfaçam as hipóteses usuais $[E(u_i) = 0]$, tem-se que:

- Salário médio de uma professora universitária:

$E(Y_i / X_i, D_i = 0) = \alpha_1 + \beta X_i$

- Salário médio de um professor universitário:

$E(Y_i / X_i, D_i = 1) = (\alpha_1 + \alpha_2) + \beta X_i$

Isso significa que as funções salários, em relação aos anos de experiência de ensino, têm a mesma inclinação $(\beta)$, mas diferentes interceptos, ou seja, o salário dos professores difere do salário das professoras, mas a taxa de variação média anual, dada pelos anos de experiência, é igual para ambos os sexos. Graficamente, esses resultados podem ser representados na Figura 6:

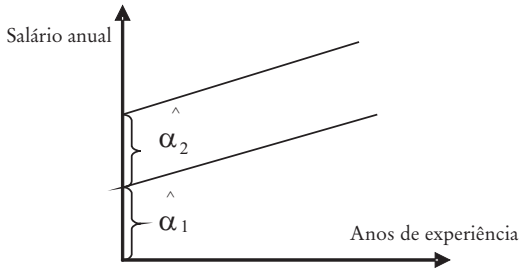



**Figura 6:** Funções salários em relação aos anos de experiência.

Logo, a partir do modelo (4.3) para estimar o salário dos professores a relação fica;

$$Y_i = (\alpha_1 + \alpha_2) + \beta X + u_i$$

[6]Variável qualitativa com duas classes, a saber, homem e mulher.

e para estimar o salário das professoras;

$$Y_i = \alpha_1 + \beta X_i + u_i$$

Os cálculos de "a" a "c", citados anteriormente, devem ser realizados. O vetor das estimativas dos parâmetros neste caso é dado por:

$$\hat{B} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{bmatrix} = \left(X' X\right)^{-1} X' Y$$

Falta-se avaliar se o valor estimado para o parâmetro $\hat{\beta}$ é estatisticamente significativo, para tanto, calcula-se a análise de variância de regressão como apresentado na Tabela 1. Os mesmos cálculos podem ser realizados para se obter o grau de ajuste do modelo e/ou para testar a significância global dos parâmetros da regressão.

Assim, a matriz de variância-covariância para $\hat{\beta}$ é definida como:

$$\text{var}-\text{cov}\left(\hat{b}\right) = \hat{\sigma}^2 \left(X' X\right)^{-1} = \begin{bmatrix} \text{var}(\alpha_1) & \text{cov}(\alpha_1, \alpha_2) & \text{cov}(\alpha_1, \beta) \\ \text{cov}(\alpha_1, \alpha_2) & \text{var}(\alpha_2) & \text{cov}(\alpha_2, \beta) \\ \text{cov}(\alpha_1, \beta) & \text{cov}(\alpha_2, \beta) & \text{var}(\beta) \end{bmatrix}$$

Pelos elementos da diagonal principal, que representam as variâncias de $\hat{\alpha}_1$, $\hat{\alpha}_2$ e $\hat{\beta}$, respectivamente, pode-se obter seus correspondentes erros padrões. De posse destes dados utiliza-se o teste $t$, como mostrado em (4.2), para testar a significância dos parâmetros. Ressalta-se a necessidade de se fazer uma análise dos resíduos do modelo, a fim de que, em conjunto com o valor de $R^2$, possa ser estabelecido a qualidade do ajuste, em qualquer um dos casos estudados.

## 5. Aplicações com saídas do software statistica, versão 5.1

Inicia-se com o exemplo do modelo (4.1) apresentado na sessão anterior em que a variável explicativa é uma variável dummy. Para a realização deste exemplo, considera-se os seguintes dados hipotéticos representados na Tabela 2.

**Tabela 2:** Dados sobre os salários médio de professores (as) universitários (as).

| Coluna | I | II | III | IV | V |
|---|---|---|---|---|---|
| Professor | Salário Inicial (Y) | Sexo ($D_1$) | Xi (anos de experiência) | D2 | D3 |
| 1º | 22 | 1 | 2 | 1 | 0 |
| 2º | 19 | 0 | 2 | 0 | 1 |
| 3º | 18 | 0 | 3 | 0 | 1 |
| 4º | 21,7 | 1 | 4 | 1 | 0 |
| 5º | 18,5 | 0 | 3 | 0 | 1 |
| 6º | 21 | 1 | 2 | 1 | 0 |
| 7º | 20,5 | 1 | 4 | 1 | 0 |
| 8º | 17 | 0 | 1 | 0 | 1 |
| 9º | 17,5 | 0 | 3 | 0 | 1 |
| 10º | 21,2 | 1 | 2 | 1 | 0 |

Fonte: Adaptado de Gujarati (2000)

OBS: As variáveis D2 e D3 são variáveis dummy definidas como se segue: no primeiro caso, a variável dummy admite valor zero quando o elemento da amostra for uma professora e 1 quando, professor. A variável dummy D2 inverte esta relação.

Utilizando-se o programa Statistica 5.1 para realizar este exemplo, os seguintes passos devem ser efetuados:

(1) Ao se iniciar o programa, escolhe-se a opção Multiple Regression e em seguida cliqua-se em Switch to Figura 7; em seguida, digita-se os dados conforme mostra Figura 8;

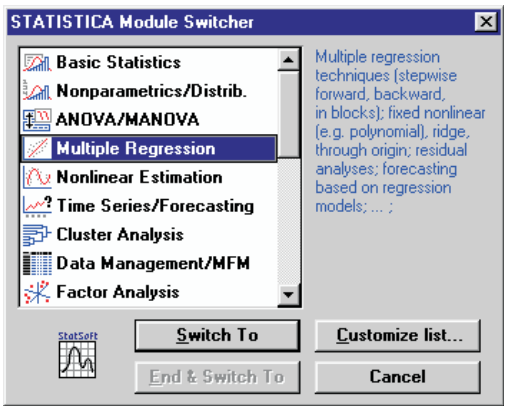



**Figura 7:** Iniciando o programa.

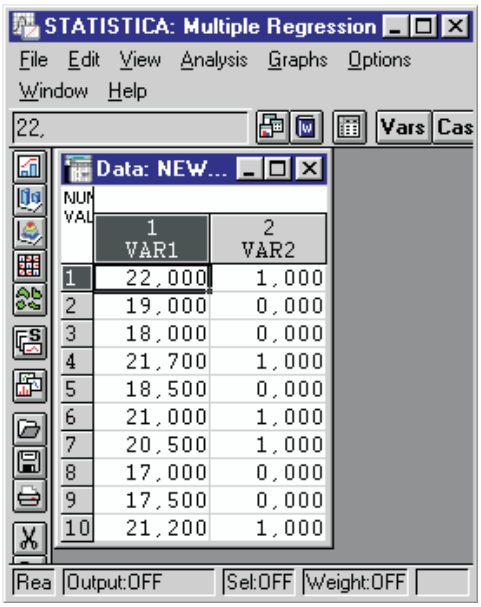



**Figura 8:** Da disposição dos dados.

(2) A seguir, clica-se com o botão esquerdo do mouse sobre a barra de ferramentas em Analysis - Startup panel e uma janela como mostrado na Figura 9 abaixo aparecerá; em seguida, clica-se em Variables e seleciona-se a variável dependente (neste caso a variável 1) e a independente (variável 2) e depois clica-se em OK  Figura 10.

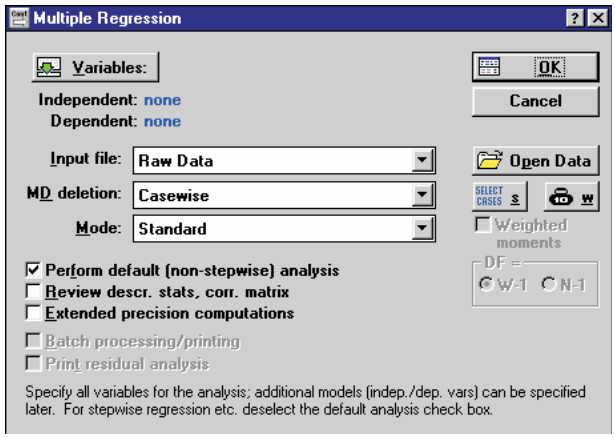



**Figura 9:** Definição das variáveis.

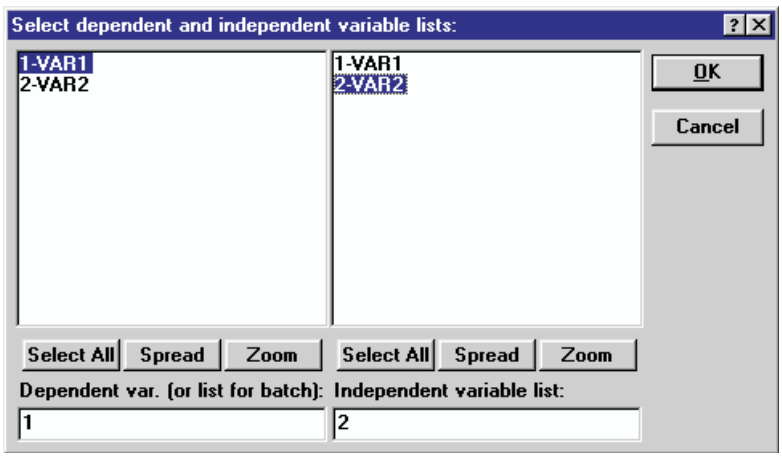



**Figura 10:** Variável dependente e independente.

(3) A janela anteriormente apresentada na Figura 7 aparecerá de novo com a definição das variáveis. Clica-se em OK. A janela a seguir aparecerá com os resultados da regressão Figura 11.

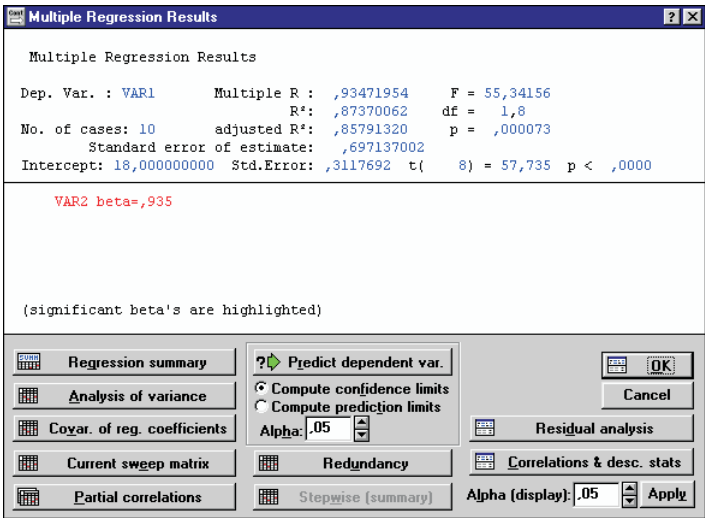



**Figura 11:** Caixa de seleção dos resultados da regressão.

(4) Clica-se em Regression Summary para obter um resumo das estatísticas. A seguinte janela será mostrada pelo programa.

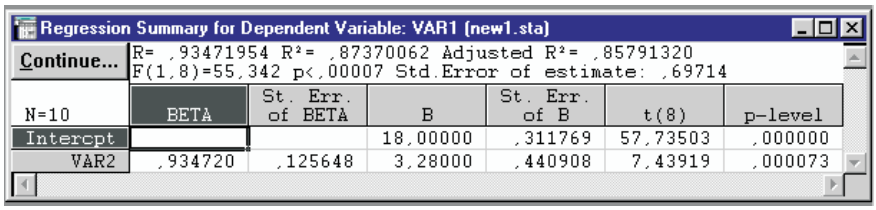



**Figura 12:** Tela com o resultado das estimativas para o modelo e suas significâncias.

(5) Para voltar à janela apresentada em (11) clica-se em Continue. A seguir obtém-se a análise de variância clicando em Analysis of variance. A seguinte janela aparecerá;

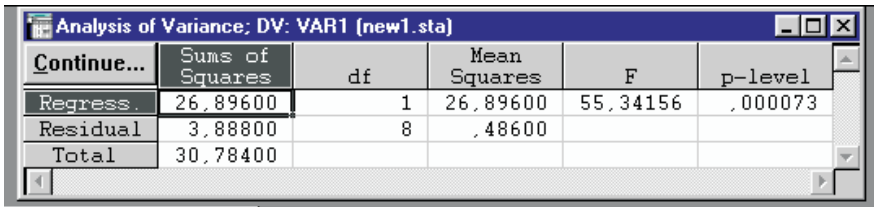



**Figura 13:** Tela com o resultado da Análise de variância.

(6) Observa-se que o programa oferece uma série de opções que podem ser testadas a fim de melhorar a análise. Clica-se em Continue e, por exemplo, obtém-se a análise dos resíduos clicando em Residual Analysis.

Para o segundo exemplo, onde o modelo apresenta uma variável quantitativa e uma variável qualitativa com duas classes (modelo 4.3), tem-se que, pelo programa computacional os seguintes passos devem ser seguidos:

Repete-se os passos apresentados anteriormente observando, contudo, que a disposição dos dados deve ser feita conforme a Figura 14, assim como a definição das variáveis independentes, tais como ilustrado na Figura 15.

(7) A janela anteriormente apresentada em (9) aparecerá de novo com a definição das variáveis. Clica-se em OK. A janela a seguir aparecerá com os resultados da regressão.Todos os demais passos, para este caso, podem ser repetidos como mostrado anteriormente

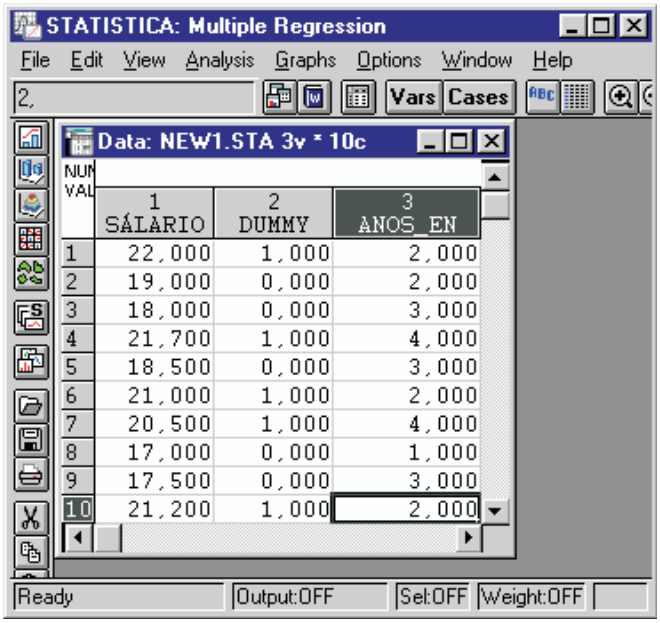



**Figura 14:** Da nova disposição dos dados.

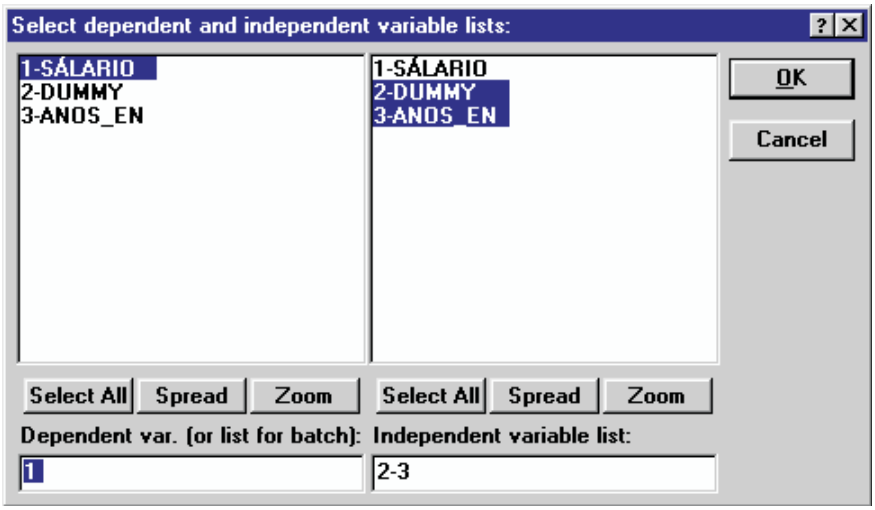



**Figura 15:** Definição da variável dependente e independente.

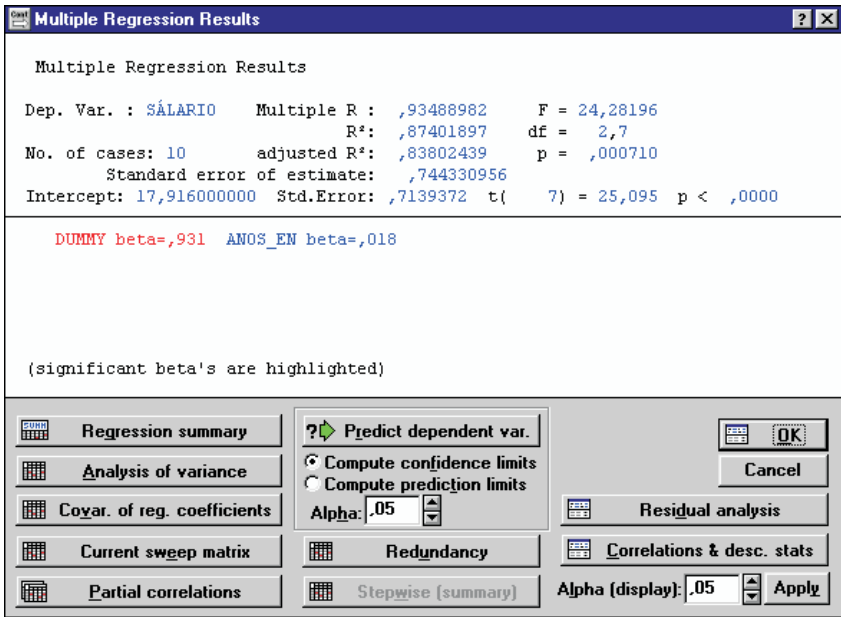



**Figura 16:** Caixa de seleção dos resultados da regressão do novo modelo.

Observa-se, portanto, com base nos procedimentos computacionais descritos acima, que o software estatístico Statistica versão 5.1 é uma ferramenta útil na estimativa da equação de regressão com variáveis independentes classificadas como dummies, permitindo desta forma, uma maior agilidade na estimação desta classe de modelos, a destacar-se, principalmente, pela simplicidade de operacionalização requerida pelo programa.

## 6. Considerações finais

A introdução das variáveis dummy na análise de regressão, como mostrado no presente trabalho, constitui-se em importante instrumento que amplia, de certa forma, o poder de análise dos modelos. Isso se deve ao fato de que este instrumental permite incorporar nos modelos variáveis importantes no contexto que se pretende analisar, e que não podem ser medidas quantitativamente.

Nesse sentido, o presente trabalho teve por objetivo apresentar, resumidamente, situações em que as variáveis dummy podem ser inseridas na análise, em especial, no caso em que estas são consideradas variáveis independentes. Observou-se, neste caso, que o software estatístico Statistica versão 5.1 é uma ferramenta computacional útil nas estimativas de equações que levam em consideração este tipo de variáveis.

Ressalta-se, no entanto, que o trabalho limitou-se estudar a inclusão da variável dummy na análise de regressão como uma variável independente. Neste caso, como sugestão para trabalhos futuros, recomenda-se o estudo de modelos de regressão que utilizem a variável dummy como uma variável dependente, especificamente, o estudo dos modelos logit, probit, tobit e/ou o modelo de probabilidade linear.

Observa-se, ainda, que o mesmo limitou-se também à situação em que o ajuste do modelo de regressão leva em consideração uma única variável independente quantitativa e que o uso de variáveis dummy pode ser estendido, com as devidas adaptações, para o caso da presença de duas ou mais variáveis independentes quantitativas no modelo.

## 7. Referências bibliográficas


CHOW, G. C. "Tests of Equality Between sets of coefficients in two Linear Regressions". **Econometrica**, V. 28, n. 3, pp. 591-605, 1960.

DUFOUR, J. M. Dummy Variables and Predictive Tests for Structural Changes: A coordinate Free Approach. **International Economic Review**, Vol. 23, pp. 565-575, 1980.

___. Dummy Variables and Predictive Tests for Structural Change. **Economic Letters**, Vol. 6, pp. 241-247, 1981.

___. Generalized Chow Tests for Structural Change: A Coordinate- Free Approach. **International Economic Review**, Vol. 23, n. 3, pp.565-575, 1982.

ERLAT, H. "On the Chow Test when the Degrees of Freedom are Inadequate", **METU Studies in Development**, n. 21, pp. 17-48, 1978.

___. Testing for Structural Change at More than One Switch Point: Inadequate Degrees of Freedom and Dummy Variables. **Oxford Bulletin of Economics and Statistics**, Vol. 47, n. 3, pp 293-302, 1985.

GUJARATI, D. Use of Dummy Variables in Testing for Equality between Sets of Coefficients in Two Linear Regressions: A Note. **The American Statistician**, Vol. 24, n. 1, pp. 50-52, 1970a.

___.Use of Dummy Variables in Testing for Equality between Sets of Coefficients in Two Linear Regressions: A Generalisation, **The American Statistician**, Vol. 24, n. 5, pp 18-21, 1970b.

___. **Econometria básica**. São Paulo: Makron Boooks, 2000.

GREENE, W. H. **Econometric Analysis**. New York: Macmillan Publishing Company, 1993.

HARDY, M. A. **Regression With Dummy Variables**. Newbury Park: Sage Publications, 1993.

HILL, C. et al. **Econometria**. São Paulo: Saraiva, 1999.

HOFFMAN, R. **Estatística para Economistas**. São Paulo, Pioneira, 1998.

KENNEDY, P. Interpreting Dummy Variables. **The Review of Economics and Statistics**, Vol. 68(1), pp. 174-175, 1986.

KOOYMAN, M. A. **Dummy Variables in Econometrics.** Netherlands: Tilburg University Press, 1976.

MADDALA, G.S. **Introdução à Econometria**. Rio de Janeiro, LTC, 2003.

REBELO, E; VALLE, P.O. O uso de regressores dummy na especificação de modelos com parâmetros Variáveis. **Revista de Estatística,** 3º quadrimestre de 2002, pp. 17-40..

___. Testes à Estabilidade dos Parâmetros de um Modelo de Regressão: Uma Aplicação Especial dos Regressores Dummy. **Revista de Estatística**. 3º quadrimestre de 2002, pp. 41-70.

___. Análise de Variância e Análise de Regressão com variáveis Dummy: Mais Semelhanças do que Diferenças. **Revista de Estatística**, Vol. I, pp. 49-86, 2002.

___. Dualidades entre Análise de Covariância e Análise de Regressão com variáveis dummy. **Revista de Estatística**. 2º quadrimestre de 2002, pp. 65-86.

STEWART, J. **Econometrics**. Cambridge: Philip Allan, 1991

SUITS, D. B.; Use of Dummy Variables in Regression Equations. **Journal of the American Statistical Association**, Vol. 52(280), pp. 548-551, 1957.

___. Dummy Variables: Mechanics V. Interpretation. **The Review of Economics & Statistics**, Vol. 66, pp. 177-180, 1984.