Notas de aula para o curso de Econometria I

Nota 4: regressão múltipla, motivação, estimação, interpretação, explicativas binárias, coeficiente de determinação ajustado e notação matricial

Thiago Fonseca Morello

fonseca.morello@ufabc.edu.br

sala 301, Bloco Delta, SBC

1 Motivação[1]

É raro que a variável que se deseja explicar, Y, seja concebida como função de apenas uma variável independente. Cabe tomar dois exemplos.

A literatura sugere que há outros fatores, além da renda familiar per capita, que atuam para determinar a ocorrência de desnutrição infantil. Entre eles, a capacidade dos pais para tomar decisões adequadas, medida por seu nível educacional (Kassouf, 1994[2]) e também pela quantidade de informação que possuem acerca de serviços de saúde localmente disponíveis (Agee, 2010[3]). Além disso, o acesso a serviços públicos, tais como o saneamento básico e o abastecimento público de água, tem influência no status nutricional das crianças. Crianças que vivem em lares sem acesso a tais serviços têm menor probabilidade de contração de doenças que podem comprometer a saúde e, portanto, seu desenvolvimento.

Há uma maneira particularmente esclarecedora de compreender como a influência de múltiplos fatores sobre uma variável explicada, severidade de desnutrição infantil, por exemplo, se manifesta em dados transversais ou em cross-section. Segundo a teoria, quanto maior a renda familiar per capita, menor a severidade (e probabilidade) de desnutrição. Mas e se houver, na população, duas famílias com níveis suficientemente próximos de renda per capita e que diferem, mesmo assim, consideravelmente em termos da severidade de desnutrição infantil? Como explicar esta discrepância? Claramente não se pode apelar para o princípio teórico que relaciona desnutrição apenas à renda. Daí o papel desempenhado por variáveis explicativas adicionais, tais como as mencionadas no final do parágrafo anterior. A educação dos pais, por exemplo, pode diferir de maneira relevante entre duas famílias com níveis suficientemente próximos de renda, mas que diferem em função da severidade de desnutrição de suas crianças. Neste caso, pois, a variável "educação dos pais" explica pelo menos parte da diferença em termos de Y que permanece não explicada após considerar diferenças de renda.

[1] É recomendada a leitura em detalhe desta motivação, uma vez que ela procura deixar mais clara a motivação apresentada por Wooldridge no início do capítulo 3 (segunda edição em inglês).

[2] Kassouf, A. L. A demanda de saúde infantil no Brasil por região e setor. Pesquisa e Planejamento Econômico, v. 24, n. 2, p. 235-260, ago. Disponível em http://www.memoria.nemesis.org.br/index.php/ppe/article/view/806/745.

[3] Agee, M. Reducing child malnutrition in Nigeria: Combined effects of income growth and provision of information about mothers' access to health care services. Social Science & Medicine 71 (2010) 1973-1980. Disponível em http://www.sciencedirect.com/science/article/pii/S0277953610006696.

Dando mais um passo a frente, a discrepância de educação pode se mostrar, mesmo relevante, ainda muito pequena quando comparada à discrepância em desnutrição. Passando-se a considerar como as famílias diferem quanto ao acesso a serviços de utilidade pública, tem-se uma razão adicional pela qual a discrepância em desnutrição existe.

Outro exemplo interessante de problema de pesquisa empírica é o de explicar porque dois estudantes do ensino médio diferem em função de seu desempenho no Exame Nacional do Ensino Médio (ENEM). Trata-se da tarefa levada a cabo por Andreia Curi e Naércio Menezes Filho em um artigo publicado em 2013[4]. A base de dados utilizada combina dados do ENEM 2006, incluindo o questionário socioeconômico, com um levantamento de valores de mensalidades de escolas particulares do estado de São Paulo. Apenas estudantes paulistas de escolas particulares são considerados.

O objetivo dos autores é medir a influência da qualidade da escola em que foi cursado o ensino médio sobre o desempenho no ENEM. A medida de qualidade empregada é o valor da mensalidade cobrada pela escola.

É possível que, na população, existam dois estudantes entre os quais não haja diferença relevante quanto à mensalidade paga ao cursar o ensino médio mas que, mesmo assim, difiram de maneira relevante em função de seu desempenho no ENEM. O que pode explicar isso? Quais outras características destes estudantes, além da qualidade da escola em que estudaram, podem explicar o desempenho no exame? Os autores consideram dois conjuntos de características adicionais. Em primeiro lugar, características da família, tais como educação dos pais e renda familiar. Em segundo lugar, características dos colegas, como renda familiar dos colegas, status dos colegas quanto ao acesso a internet (possuem/não possuem), etc. Deste modo é esperado que, dois alunos que não difiram pela qualidade do ensino médio, mas apresentem desempenho no ENEM distinto, façam parte de famílias e de grupos de colegas com características distintas.

A base fundamental do modelo de regressão linear, a função de expectativa condicional (FEC), é geral o bastante para acomodar múltiplas variáveis explicativas, conforme requerido pelos dois exemplos acima e por diversos outros problemas de pesquisa empírica.

Seja assumido que há interesse em obter a expectativa condicional da desnutrição infantil, Y, em função de três variáveis independentes, renda familiar, X_1, educação dos pais, X_2 e acesso a serviços públicos, X_3. Formalmente, o objetivo da análise consiste em conhecer melhor $E[Y|X_1,X_2,X_3] = g(X_1,X_2,X_3)$. O que pode ser feito a partir de uma aproximação linear para a função g(.), tal como segue.

$$g(X_1,X_2,X_3) \approx \beta_0 + \beta_1 X_1 + \beta_1 X_2 + \beta_3 X_3 \quad (1)$$

[4] Curi, A., Z., Menezes Filho, N.A., Mensalidade escolar, background familiar e resultados do exame nacional do ensino médio (ENEM). Revista Pesquisa e Planejamento Econômico, v.43,n.2. Disponível em http://ppe.ipea.gov.br/index.php/ppe/article/view/1469/1131.

A variável aleatória Y pode sempre ser decomposta em uma parcela correlacionada com as explicativas X_1, X_2 e X_3 e em uma parcela não correlacionada, como segue:

$$Y = E[Y|X_1,X_2,X_3] + u \quad (2)$$

Em que "u" é a parcela de Y não correlacionada com as três explicativas consideradas, parcela essa com valor esperado supostamente nulo.

Combinando (1) e (2) chega-se à função de regressão populacional (FRP) em que Y é tomada como função de múltiplas explicativas, X_1, X_2 e X_3, ou seja:

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_1 X_2 + \beta_3 X_3 + u$$

Também neste modelo o termo de perturbação, "u", capta a influência combinada de todos os outros fatores que afetam Y mas não são observados.

Um modelo de regressão linear com mais de uma variável independente é denominado "modelo de regressão linear múltipla". Estabelecer as características e propriedades deste modelo é o objetivo desta seção. No que segue, as variáveis independentes serão também referidas como "variáveis explicativas" ou "covariadas".

2 Vantagens da regressão múltipla: controlando por explicativas observáveis

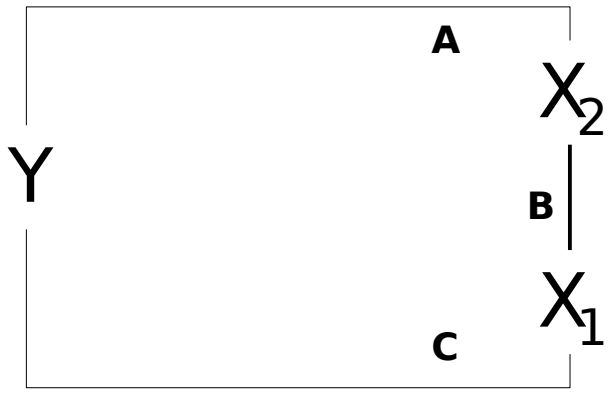
Antes de prosseguir, cabe apresentar mais algumas respostas para duas perguntas cruciais. Porque incorporar mais de uma variável explicativa à regressão linear? O que se ganha como isso?

Há, pelo menos, três respostas possíveis, esta seção trata da primeira delas.

A inclusão de mais variáveis permite obter uma estimativa mais precisa para o efeito individual, específico, de uma explicativa sobre a expectativa condicional de Y. Para explicar, será utilizado um exemplo.

Seja assumido que a variável que se deseja explicar, Y, está relacionada com duas explicativas, X_1 e X_2. Há, contudo, uma particularidade crucial acerca de X_1 e X_2: elas também são relacionadas entre si, ou seja, a covariância entre elas é não-nula, i.e., $cov(X_1,X_2) \neq 0$. É possível, mesmo assim, especificar uma regressão simples que omita uma das explicativas, X_2. Porém, o coeficiente da explicativa incluída, X_1, no caso, não capta o efeito isolado de um incremento nesta variável sobre Y. A razão para isso está exatamente na relação entre X_1 e X_2. Para compreender porque, devem ser considerados dois dos estágios do processo por meio do qual um incremento em X_1 produz efeito sobre Y. A figura abaixo representa os dois estágios.

Figura 1 Dois estágios do efeito de X_1 sobre Y



No primeiro estágio, quando X_1 é alterada, X_2 também é, automaticamente, alterada, exatamente porque há covariância entre ambas: se X_1 e X_2 co-variam, elas têm de se mover juntas. A seta B da figura 1 indica este primeiro estágio.

No segundo estágio, como X_1 está relacionada com Y, a alteração da primeira se reverte em alteração da segunda. Isso também é verdade para X_2, ou seja, a alteração dela como "efeito colateral" do incremento em X_1 também repercute em alteração de Y. Estes dois efeitos de "segundo estágio" estão representados pelas setas A e C da figura 1.

Ao final, portanto, a alteração de X_1 gera um efeito composto, o qual compreende tanto o efeito "direto" de X_1 sobre Y como o efeito "colateral" de X_1 sobre X_2 e de X_2 sobre Y. É este efeito composto que o coeficiente de X_1 na regressão simples contra Y capta. O que quer dizer que ele contém também o efeito de X_2 sobre Y. O problema está em que, na maioria dos estudos empíricos, se deseja ter uma medida do efeito individual de cada variável e não uma amálgama de efeitos, pois esta tem pouca utilidade prática.

A análise empírica do problema de desnutrição infantil, por exemplo, pode ter por objetivo recomendar políticas públicas. Há diversas razões pelas quais a renda familiar e o nível educacional dos pais podem estar relacionados. O coeficiente de uma regressão simples da renda familiar contra a desnutrição infantil capta o efeito composto desta variável, mas também do nível educacional. Isso quer dizer que tal coeficiente superestima o efeito de uma política de transferência de renda, elaborada pelo governo para combater a desnutrição infantil, caso tal política não tenha efeito relevante sobre o nível educacional dos pais. Há diversas razões pelas quais, na prática, mesmo estando renda familiar e nível educacional relacionados, um aumento na primeira pode não render efeito relevante no segundo[5]. Por exemplo, alguns pais e mães podem não ter interesse em aumentar seu nível educacional, preferindo, pois, destinar o valor que recebem do governo, por exemplo, ao investimento em educação dos filhos.

Caso seja especificada uma regressão múltipla, incorporando tanto X_1 como X_2, tal como na equação abaixo, o coeficiente da primeira variável captará apenas o efeito direto dela sobre Y, ou seja, apenas a seta C da figura 1 acima. Esta "purificação

[5] É o que ocorre quando, por trás da relação entre renda dos pais e escolaridade há a relação causal em que escolaridade dos pais causa a renda familiar, mas não há a relação causal reversa, em que renda familiar causa a escolaridade dos pais.

automática" decorre da estrutura do modelo de regressão simples. Nele, os coeficientes assumem a forma abaixo[6].

$$\beta_k = \frac{cov(\widetilde{X}_k, \widetilde{y})}{V(\widetilde{X}_k)}, k = 1,2$$

Em que $\widetilde{X}_k$ é a parte de X_k que está livre da influência das outras variáveis explicativas incorporadas no modelo e $\widetilde{y}$ é a parte de y que está livre da influência de todas variáveis explicativas exceto X_k. O coeficiente estimado capta, pois, a correlação entre a variável explicativa correspondente, X_k, e a variável dependente, eliminando, para isso, outras fontes de influência que podem modificar tal correlação. O que significa que, quanto mais completo for o modelo, no sentido de incluir o maior número de fatores que influenciam, segundo a teoria, a variável dependente e para os quais há dados disponíveis, mais preciso será o coeficiente estimado enquanto medida da influência de X_k sobre Y. Ou, alternativamente, mais "pura" será esta medida.

De fato, os coeficientes de uma regressão múltipla podem ser obtidos a partir de um processo de três estágios. Seja considerada a regressão abaixo.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \quad (1)$$

Concentrando a explicação no coeficiente de X_1, o primeiro estágio consiste em rodar a regressão de X_1 contra X_2. Ou seja, estimar o modelo:

$$X_1 = \alpha_0 + \alpha_1 X_2 + e \quad (2)$$

Em que o termo de erro "e" capta todos os fatores que influenciam X_1 e são não-correlacionados com X_2. Ou seja, tal termo corresponde à parte de X_1 livre da influência de X_2. Isso significa que os resíduos da regressão, $\hat{e}_i = x_{1i} - \hat{\alpha}_0 - \hat{\alpha}_1 x_{2i}$ podem ser utilizados como medida para calcular o efeito puro de X_1 sobre Y – dado que captam a parcela da variação de X_1 não explicada por X_2. Eles devem, portanto, ser armazenados como resultado do primeiro estágio.

No segundo estágio, estima-se a regressão a seguir.

$$Y = \gamma_0 + \gamma_1 X_2 + \epsilon \quad (3)$$

O termo de erro "ϵ" representa a parte de Y livre da influência de X_2. Desta maneira, os resíduos dessa regressão também estão "purificados" de X_2.

No terceiro estágio, roda-se a regressão da parcela de Y livre de X_2 contra a parcela de X_1 livre de X_2, ou seja, trata-se de:

$$\hat{\epsilon} = \gamma_0 + \gamma_1 \hat{e} + \xi \quad (4)$$

[6] Esta expressão é denominada por Angrist & Pischke (no capítulo 3 de "Mostly harmless econometrics") de "anatomia da regressão" e resulta do estudo seminal de Frisch e Waugh, Frisch, R., and F. Waugh. "Partial Time Regressions as Compared with Individual Trends." Econometrica, 1, 1933, pp. 387–401.

O coeficiente γ_1 capta apenas o efeito direto de X_1 sobre Y, sendo equivalente ao coeficiente β_1 na equação (1). O que quer dizer que o efeito "puro" ou parcial de X_1 sobre Y pode ser obtido diretamente a partir da estimação do coeficiente β_1 com base em (1). Este resultado é conhecido como teorema de Frisch-Waugh, tendo sido originalmente demonstrado por Ragnar Frisch e Frederick Waugh em um artigo de 1933[7].

A regressão múltipla proporciona, pois, a obtenção do efeito líquido de uma variável explicativa com um procedimento de apenas um estágio. Esta é uma vantagem relevante, dada a economia de tempo e de cálculos.

Em termos rigorosos, em uma regressão múltipla, o coeficiente de uma explicativa capta a influência dela líquida, ou expurgada, da influência das demais explicativas incluídas no modelo. Variáveis que explicam Y e permanecem omitidas continuam a ser captadas pelos coeficientes das explicativas incluídas. Por hora, cabe concentrar a atenção no esclarecimento de um termo recorrentemente utilizado em econometria. Entende-se por "controlar por um conjunto de variáveis" o ato de incorpora-las no modelo com o intuito de obter uma medida mais pura quanto possível para o efeito de uma explicativa definida como principal.

Por exemplo, Curi e Menezes Filho (2013) "controlam" por características socioeconômicas dos alunos e dos colegas, procurando, pois, ter uma medida do efeito individual da escola de ensino médio tão "pura" quanto possível. O que significa expurgar, do efeito individual da escola, a influência de outros dois conjuntos de fatores, a família e o meio social em que o aluno vive.

3 Vantagens da regressão múltipla: previsão e otimização de informação

Há, ainda, duas vantagens adicionais da regressão múltipla. A análise pode ter como objetivo prever Y. Ou seja, quer-se saber qual valor de Y tende a prevalecer dados os valores das explicativas. Neste caso, quanto maior é a proporção da variação de Y explicada pela amostra (maior o coeficiente de determinação), menor será o erro de previsão. Quanto mais variáveis correlacionadas com Y forem, pois, introduzidas, maior será a proporção total da variação explicada e, portanto, mais preciso o modelo de previsão.

A última vantagem é estabelecida por um dos princípios fundamentais da econometria. Trata-se do princípio de aproveitamento completo da informação disponível, segundo o qual não se deve desperdiçar dados, incorporando toda a informação útil ao modelo. O que, para um modelo que procura explicar Y, significa incluir todas as explicativas sugeridas pela teoria.

[7] Frisch, Ragnar, and Frederick V. Waugh (1933): "Partial Time Regression as Compared with Individual Trends,"Econometrica, 1, 387–401.

4 Estimação

O modelo de regressão múltipla com duas variáveis explicativas pode ser representado pela função de regressão populacional (FRP) abaixo.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \ (1)$$

O método de mínimos quadrados é aplicado de maneira equivalente à regressão simples, basta recordar que ele procura minimizar o quadrado do erro de aproximação linear dado por:

$$E\left[\left(Y-\widetilde{Y}\right)^2 \middle| X_1, X_2\right](2)$$

Em Y ≡ valor observado de Y e $\widetilde{Y}$ ≡ valor de Y previsto pela aproximação linear. Aplicando a definição de FRP linear em (1) à expressão (2), tem-se:

$$E\left[\left(Y-\beta_0-\beta_1 X_1-\beta_2 X_2\right)^2 \middle| X_1, X_2\right](2')$$

O problema a ser resolvido é, pois:

$$min_{[\beta_0,\beta_1,\beta_2]} E\left[\left(Y-\beta_0-\beta_1 X_1-\beta_2 X_2\right)^2 \middle| X_1, X_2\right]$$

As condições de primeira ordem são:

$$E\left[\left(Y-\beta_0-\beta_1 X_1-\beta_2 X_2\right) \middle| X_1, X_2\right]=0(3a)$$

$$E\left[X_1\left(Y-\beta_0-\beta_1 X_1-\beta_2 X_2\right) \middle| X_1, X_2\right]=0(3b)$$

$$E\left[X_2\left(Y-\beta_0-\beta_1 X_1-\beta_2 X_2\right) \middle| X_1, X_2\right]=0(3c)$$

Pode-se chegar às mesmas condições a partir do método dos momentos. Para isso é preciso assumir as três condições abaixo

$$E[u_i] = 0 \ (MM1), \ i=1,...,N \ (4a)$$

$$E[x_{i1}\, u_i] = 0 \ (MM2), \ i=1,...,N \ (4b)$$

$$E[x_{i2}\, u_i] = 0 \ (MM3), \ i=1,...,N \ (4c)$$

Aplicando o princípio da analogia às condições 3a-3c e também a definição do termo de perturbação ($u_i = y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2}$), tem-se as equações abaixo, a partir das quais os estimadores para os parâmetros β_0, β_1 e β_2 podem ser obtidos.

$$\sum_{i=1}^{N}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_{i1}-\hat{\beta}_2 x_{i2}\right)(5a)$$

$$\sum_{i=1}^{N} x_{i1}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_{i1}-\hat{\beta}_2 x_{i2}\right)(5b)$$

$$\sum_{i=1}^{N} x_{i2}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_{i1}-\hat{\beta}_2 x_{i2}\right)(5c)$$

As condições de primeira ordem acima podem ser generalizadas para o caso de K variáveis. Neste caso, a FRP é $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 +...+ \beta_KX_K +u$ e as condições 5a-5c passam às K+1 condições abaixo:

$$\sum_{i=1}^{N}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_{i1}-\hat{\beta}_2 x_{i2}-\ldots-\hat{\beta}_K x_{iK}\right)$$

$$\sum_{i=1}^{N} x_{i1}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_{i1}-\hat{\beta}_2 x_{i2}-\ldots-\hat{\beta}_K x_{iK}\right)$$

$$\sum_{i=1}^{N} x_{i2}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_{i1}-\hat{\beta}_2 x_{i2}-\ldots-\hat{\beta}_K x_{iK}\right)$$

...

$$\sum_{i=1}^{N} x_{iK}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_{i1}-\hat{\beta}_2 x_{i2}-\ldots-\hat{\beta}_K x_{iK}\right)$$

Com base nessas K+1 condições é possível obter estimadores para os K+1 parâmetros (um intercepto e K coeficientes).

5 Interpretação dos coeficientes

A interpretação dos coeficientes como medidas do efeito “puro” ou “direto” das explicativas já foi apresentada na seção 1 acima. Uma interpretação que vai na mesma direção toma por base o efeito de um incremento pequeno em apenas uma das explicativas. Seja considerado um modelo com duas explicativas, X_1 e X_2 e assumido que X_2 sofre um incremento equivalente a Δ a partir de um valor inicial x_2^0.

O valor da aproximação linear da expectativa condicional de Y para $X_2 = x_2^0 + \Delta$ é $E[Y|X_1, X_2 = x_2^0 + \Delta] = \beta_0 + \beta_1X_1 + \beta_2(x_2^0 + \Delta)$. Já, para o valor inicial de X_2, a expectativa condicional de Y é equivalente a $E[Y|X_1, X_2 = x_2^0 + \Delta] = \beta_0 + \beta_1X_1 + \beta_2x_2^0$. O efeito do pequeno incremento Δ sobre a expectativa condicional é, pois, dado por:

$$E[Y|X_1, X_2 = x_2^0 + \Delta] - E[Y|X_1, X_2 = x_2^0] =$$

$$\beta_0 + \beta_1X_1 + \beta_2(x_2^0 + \Delta) – (\beta_0 + \beta_1X_1 + \beta_2x_2^0) = \beta_2\Delta.$$

De modo que:

$$\beta_2=\frac{E\left[Y\middle|X_1, X_2=x_2^0+\Delta\right]-E\left[Y\middle|X_1, X_2=x_2^0\right]}{\Delta}$$

O coeficiente de X_2, β_2, capta o efeito de um pequeno incremento em X_2 sobre o valor da expectativa condicional mantendo-se X_1 inalterada. Este princípio pode ser generalizado para uma regressão múltipla com K explicativas. Neste caso, a expectativa condicional de Y em relação a K explicativas, $X_1,...,X_K$, pode ser aproximada segundo a forma linear:

$$E[Y| X_1,X_2,...,X_K] = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_KX_K$$

O efeito parcial da k-ésima explicativa, sobre a expectativa condicional, é dado por:

$$\beta_k = \frac{E\left[Y \middle| X_1,\ldots,X_{k-1},\ldots,X_K,X_k = x_k^0 + \Delta\right] - E\left[Y \middle| X_1,\ldots,X_{k-1},\ldots,X_K,X_k = x_k^0 + \Delta\right]}{\Delta}$$

Tomando-se um valor desprezível, infinitesimal para Δ, a expressão acima converge, no limite, para a derivada parcial de $E[Y| X_1,X_2,...,X_K]$ em relação a X_k.

$$\frac{\partial}{\partial X_k}\{E[Y| X_1,X_2,...,X_K]\} = \frac{\partial}{\partial X_k}\{\beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_KX_K\} = \beta_k$$

O coeficiente de cada explicativa capta, portanto, o efeito sobre Y de um incremento infinitesimal na explicativa, mantendo-se todas as demais explicativas inalteradas. O uso da cláusula "ceteris paribus"[8] não poderia ser mais adequado, uma vez que, segundo o teorema de Frisch-Waugh, cada coeficiente capta apenas e tão-somente o efeito da alteração da variável correspondente[9].

Daqui em diante no curso, pois, os termos "efeito puro", "efeito parcial" e "efeito marginal" serão usados intercambiavelmente.

Há um detalhe a assinalar quanto ao efeito parcial. Ele sempre pode ser calculado mesmo quando não existam, na amostra, observações que assumam valores equivalentes para todas as explicativas, exceto por uma delas. No exemplo da equação de desnutrição infantil, podem não existir duas famílias com níveis equivalentes para a educação dos pais e para o acesso a serviços públicos e que difiram apenas pela renda familiar. Mesmo assim, é possível calcular o efeito parcial pois ele corresponde ao valor da estimativa pontual do coeficiente da renda familiar.

De fato, o modelo de regressão amostral, uma vez estimado, pode ser utilizado para realizar simulações que extrapolam a amostra. Ou seja, pode-se prever o valor de Y para indivíduos com características distintas daqueles que compõem a amostra. Para exemplificar, pode ser que não exista, na amostra da POF 2008/2009, nenhuma família com renda zero e com nível acima da média para a educação dos pais. Mesmo assim, o valor do indicador de desnutrição infantil desta "família artificial" pode ser previsto atribuindo-se valor zero para a renda familiar e um valor acima da média para a educação dos pais.

[8] Ceteris paribus pode ser traduzido como "mantendo todas as demais condições (ou variáveis) inalteradas".

[9] É claro que, de fato, o coeficiente acaba o efeito de X_k expurgado da correlação que esta explicativa possui com as demais explicativas incluídas na regressão.

6 Variáveis qualitativas binárias e seus coeficientes[10]

Boa parte das variáveis que se procura explicar em estudos empíricos de ciências sociais são influenciadas não apenas por variáveis quantitativas, mas também por variáveis qualitativas. Uma maneira de diferenciar essas duas classes de variáveis está em observar que, enquanto as primeiras captam a intensidade ou magnitude em que uma característica ou atributo está presente em um indivíduo, firma ou região, as segundas indicam simplesmente, de maneira binária, se o atributo é ou não possuído.

Uma medida quantitativa para o acesso à energia elétrica, uma explicativa para o grau de desnutrição infantil, é a duração média diária de suprimento ininterrupto de eletricidade. Uma medida qualitativa para a mesma explicativa é uma variável binária indicando simplesmente se o domicílio tem ou não acesso à energia elétrica.

Outros atributos cuja detenção pode ser indicada a partir de variáveis binárias, dicotômicas, são o gênero, se um indivíduo está empregado ou não, posse de um bem durável específico, como o carro, cumprimento de uma norma ambiental por uma empresa, posse de título de propriedade sobre uma área fundiária, etc.

A incorporação de variáveis qualitativas binárias no modelo de regressão múltipla é simples. Basta criar uma variável que assume valor unitário quando está presente o atributo e valor zero caso contrário.

No estudo da desnutrição infantil, domicílios com acesso a saneamento básico, por exemplo, recebem valor unitário e domicílios sem acesso a saneamento básico, valor zero.

Em um dos exemplos mais recorrentes no livro de Wooldridge, o salário de um trabalhador pode ser explicado em função de algumas características quantitativas, como nível educacional e experiência, e em função do gênero, masculino ou feminino, uma variável qualitativa binária. Neste caso, pode-se definir uma variável binária que assume valor unitário para indivíduos do gênero feminino e valor zero para indivíduos do sexo masculino.

Focando no estudo da desnutrição infantil, o acesso a serviços públicos pode ser medido a partir da incorporação da variável binária indicando, com valor unitário, o acesso a saneamento básico. Esta é denotada por d_saneamento. A FEC de Y neste caso pode ser aproximada pela forma linear abaixo, em que X representa um vetor de características socioeconômicas.

$$E[Y|\ X, d\_saneamento] = \beta_0 + \beta_1 X + \delta d\_saneamento$$

[10] Este tópico é coberto no capítulo 7 do livro-texto de Wooldridge (segunda edição em inglês) e o capítulo 9 do livro-texto de Gujarati (quarta edição em inglês).

Um aspecto importante quanto aos modelos com variáveis binárias diz respeito à interpretação do coeficiente destas variáveis. O teorema de Frisch-Waugh é válido com explicativas binárias ou não, portanto, o coeficiente β_3 acima capta o efeito do acesso à saneamento básico livre da influência das demais explicativas. Mas, contudo, o efeito parcial não pode ser interpretado com a medida de uma variação infinitesimal, uma vez que esta interpretação se aplica apenas a variáveis contínuas o que não é, claramente, o caso de uma variável binária. De fato, trata-se do caso mais extremo de uma variável discreta, pois há apenas dois valores possíveis.

Para as crianças residentes em domicílios com acesso a saneamento, a FRP acima passa a:

$$E[Y| X, d\_saneamento = 1] = \beta_0 + \beta_1 X + \delta \quad (a)$$

Já, para crianças residentes em domicílios sem acesso a saneamento prevalece a versão da FRP a seguir:

$$E[Y| X, d\_saneamento = 0] = \beta_0 + \beta_1 X \quad (b)$$

A diferença entre os dois grupos de crianças, no que tange à expectativa condicional de Y é, pois:

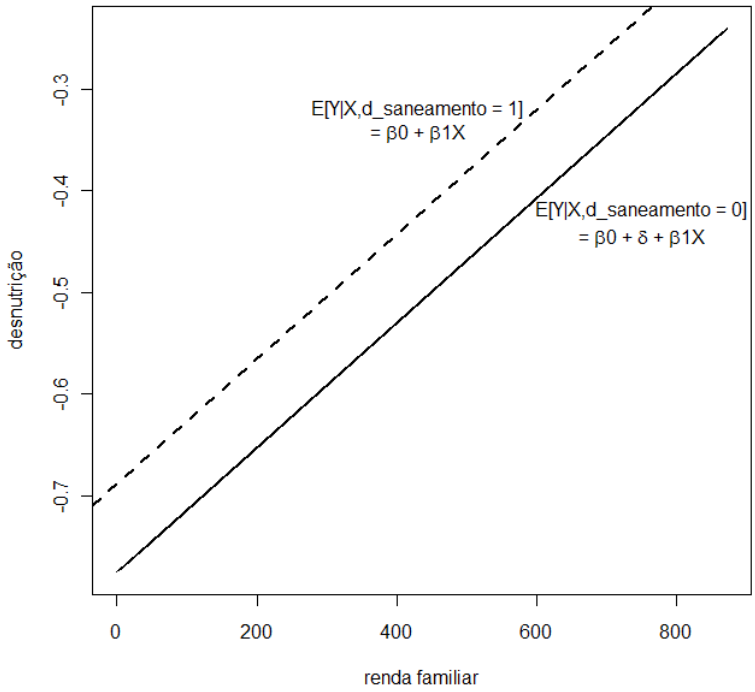
$$E[Y| X, d\_saneamento = 1] - E[Y| X, d\_saneamento = 0] = \delta$$

O coeficiente da binária indicando acesso a saneamento, pois, capta a magnitude em que a expectativa condicional de Y para o grupo com acesso a saneamento difere da expectativa condicional de Y para o grupo sem acesso a saneamento. Trata-se, portanto, do efeito parcial, ou específico, do acesso a saneamento, mantendo-se inalteradas todas as demais explicativas que, no caso, são incorporadas por meio do vetor X.

O coeficiente δ, portanto, dá uma resposta para a seguinte pergunta: em que medida a expectativa condicional de Y difere entre indivíduos (domicílios ou empresas) equivalentes para todas as características socioeconômicas consideradas, com exceção do acesso a saneamento básico?

O coeficiente de uma variável binária também pode ser entendido como a magnitude em que difere o intercepto do grupo-base, que não porta o atributo, i.e., todos os indivíduos i tais que $d_i = 0$, do intercepto do grupo que possui o atributo, i.e., todos os indivíduos i tais que $d_i = 1$. Basta ver que, no exemplo acima, para d_saneamento = 0, o modelo assume a forma $E[Y| X, d\_saneamento = 0] = \beta_0 + \beta_1 X$, de maneira que o intercepto é equivalente a β_0. E, para d_saneamento = 1, tem-se $E[Y| X, d\_saneamento = 1] = \beta_0 + \beta_1 X + \delta$, de modo que o intercepto é equivalente a $\beta_0 + \delta$. A figura abaixo ilustra esta interpretação. A distância entre as duas curvas é exatamente equivalente a δ.

Figura 2 Variável binária d_saneamento compreendida como deslocadora de intercepto



7 Variáveis qualitativas com múltiplas categorias[11]

Algumas variáveis qualitativas não representam atributos dicotômicos, não sendo binárias. São exemplos a unidade da federação e região brasileira em que um domicílio está localizado, o meio de transporte escolhido por um pessoa ao comutar diariamente para o trabalho, etnia e até a situação perante o mercado de trabalho (empregado formalmente, empregado informalmente, desempregado e procurando emprego, desempregado e não procurando empregado). Como incorporar variáveis qualitativas que captam múltiplas categorias? A partir de múltiplas variáveis binárias, de maneira a que cada uma assuma valor unitário quando a unidade observacional se enquadra em uma dada categoria e valor nulo caso contrário. Cabe detalhar, tomando um exemplo.

No caso da característica qualitativa referente à região brasileira de localização, há cinco categorias possíveis, Norte, Nordeste, Centro-Oeste, Sudeste e Sul. Uma variável binária assumindo valor unitário para observações localizadas na região Norte e zero para as demais, d_N permitiria distinguir apenas as observações que pertencem à região Norte das demais. Caso seja adicionalmente considerada uma variável binária d_NE, com valor unitário apenas para unidades localizadas na região Nordeste, seria possível identificar observações pertencentes às regiões Norte e Nordeste e também observações

[11] Este tópico é coberto no capítulo 7 do livro-texto de Wooldridge (segunda edição em inglês) e o capítulo 9 do livro-texto de Gujarati (quarta edição em inglês).

pertencentes às demais regiões. A identificação se tornaria mais precisa com a inclusão de d_CO, d_SE e d_S para a região Sul.

Não é difícil ver que uma das cinco variáveis binárias é redundante. Basta ter quatro variáveis binárias para identificar as cinco regiões, excluindo, por exemplo, d_Sul, uma vez que unidades que não são identificadas com valor unitário para nenhuma das quatro demais binárias (d_N, d_NE, d_CO e d_SE) claramente só podem pertencer, exclusivamente, à região remanescente, Sul. Se, portanto, a binária indicando essa região não for incluída no modelo, não haverá nenhuma perda de precisão no que tange à identificação completa da região a que pertencem as observações.

Seja considerado o modelo genérico abaixo com X representando um vetor de explicativas socioeconômicas.

$$Y = \beta_0 + \beta_1 X_1 + \delta_1 d\_N + \delta_2 d\_NE + \delta_3 d\_CO + \delta_4 d\_SE + u$$

Este modelo também pode ser representado a partir da função de expectativa condicional (FEC), i.e.:

$$E[Y| X, d\_N, d\_NE, d\_CO, d\_SE] = \beta_0 + \beta_1 X + \delta_1 d\_N + \delta_2 d\_NE + \delta_3 d\_CO + \delta_4 d\_SE$$

O que exatamente representam os coeficientes das variáveis binárias? Para compreender, cabe focar em um coeficiente em específico, por exemplo, o de d_N. Em primeiro lugar, cabe considerar que, para as observações localizadas no Norte, o modelo passa a:

$$E[Y|X, d\_N=1, d\_NE=0, d\_CO=0, d\_SE=0] = \beta_0 + \beta_1 X + \delta_1 \quad (1)$$

E, para as observações localizadas no Sul, o modelo assume a seguinte forma:

$$E[Y| X, d\_N=0, d\_NE=0, d\_CO=0, d\_SE=0] = \beta_0 + \beta_1 X \quad (2)$$

As expectativas condicionais acima podem ser sinteticamente representadas por E[Y| d_N=1, X] e E[Y| d_N = d_NE= d_CO= d_SE=0, X], respectivamente. A diferença, em termos da expectativa condicional de Y em relação a X, entre os grupos localizados no Norte e no Sul, i.e., a diferença entre (1) e (2), é equivalente a:

$$E[Y|X, d\_N=1] - E[Y| X, d\_N = d\_NE= d\_CO= d\_SE=0] = \delta_1$$

Exatamente, portanto, o coeficiente da binária que indica com valor unitário se a observação pertence à região Norte. Este coeficiente, pois, capta o efeito parcial da localização na região Norte sobre a expectativa condicional da variável explicada em relação a X, tomando-se como base de comparação o valor desta expectativa condicional na região Sul.

É preciso assinalar que o diferencial de expectativa condicional em questão é definido com base em valores fixos para todas as explicativas exceto a região de localização. Ou seja, há um experimento mental subjacente em que são comparados dois grupos de unidades (indivíduos, domicílios, etc). O primeiro é definido por valores para as

características socioeconômicas captadas pelo vetor X, os quais são captados por um vetor específico x, e também por residir na região Norte, i.e., d_N = 1. O segundo grupo também se define por características socioeconômicas X = x, mas reside na região Sul, i.e., d_N = d_NE= d_CO= d_SE=0. Este é a única maneira de captar, com a FRP, o efeito exclusivo, puro, de residir em uma dada região. Trata-se, conforme visto na nota de aula 7, de um efeito específico, puro, ou "ceteris paribus".

Concretamente, a hipótese de que existem dois grupos de indivíduos (ou domicílios ou firmas) que diferem apenas em função da região de localização é muito forte. Tal característica tende a estar correlacionada com outras características socioeconômicas, ainda mais em um País tão regionalmente desigual como o Brasil.

O exercício pode ser repetido, agora para o coeficiente da binária indicando a região Nordeste. A expectativa condicional de Y para observações localizadas na região Nordeste é:

$$E[Y|X, d\_N=0, d\_NE=1, d\_CO=0, d\_SE=0] = \beta_0 + \beta_1 X + \delta_2 \quad (3)$$

Ou, sinteticamente, $E[Y|X, d\_NE=1] = \beta_0 + \beta_1 X + \delta_2$.

E, analogamente, o coeficiente δ_2 capta a diferença das expectativas condicionais dos grupos populacionais residentes na região Nordeste e na região Sul, como segue:

$$E[Y|X, d\_NE=1] - E[Y|X, d\_N = d\_NE= d\_CO= d\_SE=0] = \delta_2$$

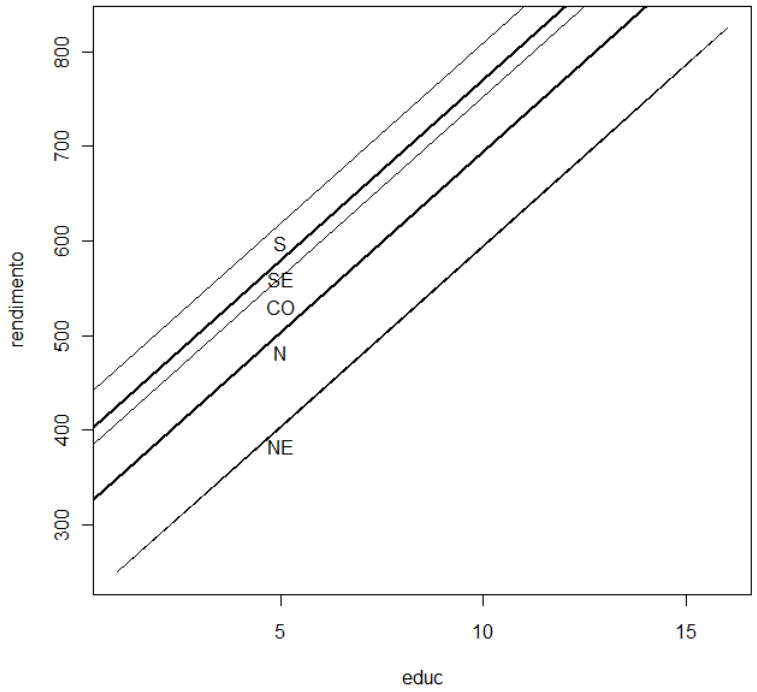
O mesmo vale para as regiões Centro Oeste (CO) e Sudeste (SE). Uma implicação crucial é a de que a região Sul é a categoria base de localização regional, contra a qual as expectativas condicionais de Y correspondentes às demais regiões são comparadas. De fato, a categoria base é aquela à qual não corresponde nenhuma das binárias que constam na regressão. A exclusão, automaticamente, especifica a categoria-base e não há nenhuma regra restritiva neste sentido; a escolha da categoria-base é livre, trata-se de uma decisão do analista.

Mas é claro que uma escolha aleatória não é a melhor maneira de proceder. O mais adequado é selecionar a categoria da variável qualitativa (região brasileira de localização) que se mostra uma base de comparação natural, ou, pelo menos, adequada aos propósitos do estudo. Por exemplo, resultados prévios de pesquisa apontam para um maior nível médio de desnutrição infantil na região Nordeste. Se esta for tomada como região-base na FRP que explica severidade de desnutrição, então o coeficiente da região Norte capta o grau em que Norte e Nordeste diferem em termos da expectativa condicional da medida de severidade. Ou seja, tem-se ai uma medida para o efeito, sobre a severidade de desnutrição, de residir no Norte comparado ao efeito de residir no Nordeste.

Os coeficientes das variáveis binárias são sempre, pois, medidas relativas, comparações, nunca captando o efeito absoluto de uma categoria qualitativa.

Uma interpretação alternativa (e complementar) de tais coeficientes é a de que eles representam deslocamentos de intercepto, conforme a figura abaixo indica.

Figura 1 Relação rendimento do trabalho (remuneração) vs educação para as UFs brasileiras, PNAD 2009, S ≡ E[Y|X,d_S= 1], SE ≡ E[Y|X,d_SE= 1], CO ≡ E[Y|X,d_CO= 1], NE ≡ E[Y|X,d_NE= 1] e N ≡ E[Y|X,d_N= 1].



8 Coeficiente de determinação ajustado (e inclusão de variáveis irrelevantes)

O coeficiente de determinação, R^2, é, tal como no caso da regressão simples, uma medida para a proporção da variação de Y explicada pelas variáveis independentes que compõem o modelo de regressão linear múltipla. A definição da estatística é equivalente à empregada no caso da regressão simples, tal como segue.

$$R^2=\frac{SQE}{SQT}=1-\frac{SQR}{SQT}=1-\frac{\left(\sum_{i=1}^{N} y_i-\hat{\beta}_0-\hat{\beta}_1 x_{i1}-\hat{\beta}_2 x_{i2}-\ldots-\hat{\beta}_K x_{iK}\right)^2}{\left(\sum_{i=1}^{N} y_i-\overline{y}\right)^2}$$

Uma característica desta fórmula padrão está em que o R^2 nunca cai com a inclusão de explicativas adicionais[12]. Cabe discutir melhor este fato antes de passar à demonstração, a qual fica relegada à próxima sub-seção.

Uma implicação direta é a de que a prática de incluir explicativas adicionais com o exclusivo intuito de aumentar o R^2, e, pois, inflar, a medida de qualidade do ajuste do modelo aos dados, não encontra barreiras à sua perseguição. O que quer dizer que mesmo explicativas não previstas pela teoria e/ou com correlação baixa ou nula com a variável explicada (Y) podem ser introduzidas em prol da consecução deste objetivo. Um alto valor do R^2, portanto, pode ter fundamento um modelo repleto de explicativas com relação pouco ou nada relevante com Y. Daí uma razão para não medir a qualidade de um modelo de regressão linear exclusivamente com base no R^2.

É possível, contudo, alterar a fórmula do R^2 de maneira a desestimular a introdução de explicativas espúrias, no sentido estatístico, estritamente (e não em relação à teoria), penalizando o aumento do número de variáveis. Trata-se de uma maneira de disciplinar a análise de regressão múltipla, reforçando os princípios (i) de escolha criteriosa de explicativas, preferencialmente com base na teoria, e (ii) de parcimônia. Este último estabelece que o conjunto de explicativas não deve ser exaustivo, mas sim captar os determinantes essenciais.

A alteração, ou ajuste, consiste em incorporar os graus de liberdade (ver Box ao final desta sub-seção) da SQR e da SQT à expressão do R^2, obtendo o que se denomina por R^2 ajustado ou $\overline{R^2}$, conforme a fórmula abaixo esclarece.

$$\overline{R^2}=1-\frac{SQR/(N-K-1)}{SQT/(N-1)}=1-\frac{SQR}{SQT}\frac{(N-1)}{(N-K-1)}$$

O correto entendimento do porque tal ajuste gera uma nova fórmula cujo valor não pode ser inflado artificialmente pode ser alcançado a partir da inspeção da fórmula anterior. Seja a atenção concentrada, em um primeiro passo, na SQR. É preciso recordar que o R^2 é ajustado com o objetivo de punir a inclusão de explicativas que pouco ou nada acrescentam ao poder explicativo do modelo.

Seja considerada, por simplicidade, a inclusão de uma explicativa adicional, denotada por x_{K+1}. Se ela pouco ou nada acrescenta em poder explicativo, isso quer dizer que a inclusão dela resulta em um aumento praticamente nulo da SQE, uma que esta última é a medida de poder explicativo do modelo que dá fundamento ao R^2. Uma variação praticamente nula da SQE sempre vem acompanhada de uma variação praticamente

[12] Ver Greene, W., Econometric Analysis (2002), seção 3.5.1, especialmente o teorema 3.6 e a demonstração no apêndice desta nota.

nula da SQR. A razão para isso está em que SQT = SQE + SQR e a SQT é nada mais do que a soma dos quadrados dos desvios de Y em relação a sua média, algo que não varia em função de alterações no conjunto de explicativas que compõem o modelo de regressão linear. Desta maneira, pois, sempre que a SQE varia em uma determinada magnitude, a SQR tem de variar na mesma magnitude, mas, contudo, na direção contrária, uma vez que as duas são as únicas parcelas de uma soma de valor constante.

Independente da magnitude da SQR provocada pela inclusão de x_K, tal inclusão resulta em um aumento do número de variáveis explicativas, inevitavelmente. Este é o "custo" pago pela introdução de x_k. Caso o benefício, medido pela queda na SQR, seja desprezível, tem-se que a inclusão de x_k não compensa e é exatamente isso que o R^2 ajustado tem que indicar. E é isso, de fato, o que ocorre, uma vez que o R^2 ajustado é uma função positiva da SQR e negativa do número de explicativas, conforme a expressão abaixo destaca. Nela, os sinais subscritos aos argumentos de f(.) representam a direção (sinal) da influência de cada argumento no valor do R^2 ajustado.

$$\overline{R^2} = 1 - \frac{SQR}{SQT}\frac{(N-1)}{(N-K-1)} = f_{?}$$

Desta maneira, pois se com a inclusão de x_k a SQR se mantém inalterada mas o número de explicativas, K, aumenta, o valor do R^2 tem de cair, o que aponta para uma redução da qualidade de ajuste do modelo aos dados. Desta maneira, variáveis explicativas que proporcionam um benefício, medido em termos de aumento do poder explicativo do modelo, inferior ao custo de perda de graus de liberdade, acabam por implicar na redução da qualidade do ajuste do modelo aos dados, caso a última seja medida com base no R^2 ajustado. É exatamente por ser capaz de sinalizar a existência de explicativas "desvantajosas", no sentido estatístico, que o R^2 ajustado é uma medida mais adequada do que R^2 para a qualidade de uma regressão múltipla.

Antes de avançar, cabe esclarecer um ponto. O R^2 ajustado, assim como qualquer outra estatística, não pode penalizar exercícios empíricos com parco ou insuficiente embasamento teórico. De fato, é impossível desenhar uma fórmula para o R^2 que desestimule a seleção de explicativas sem embasamento na teoria. O que o R^2 ajustado pune é, estritamente, um mau desempenho estatístico das explicativas, medido, por exemplo (mas não unicamente), em função de uma baixa redução da parcela de Y não explicada pelo modelo (SQR).

Outra característica do R^2 ajustado que cabe mencionar é a de que seu valor não está restrito ao intervalo [0;1], ele pode ser negativo, é o que ocorre quando o poder explicativo de um modelo com um número relevante de variáveis explicativas é muito baixo, ou seja, trata-se do caso em que a maioria das explicativas não tem significância estatística. Como se verá na seção 2.3, é o que ocorre quando é consumido um número relevante de graus de liberdade por explicativas que pouco (ou nada) contribuem para a

redução da parcela de Y não explicada pela regressão (ou, de maneira equivalente, para o aumento da parcela de Y explicada pela regressão).

Box: calculando os graus de liberdade da SQT e da SQR

Este Box retoma a nota de aula 5 e a nota suplementar 1. Entende-se por "graus de liberdade" o número de partículas informacionais contidas na amostra cujo valor não é fixado pelas estatísticas. A SQT é equivalente a $\sum_{i=1}^{N}\left(y_i-\overline{y}\right)^2$. Seu cálculo, pois, depende do cálculo prévio da média de Y, o que elimina uma partícula de informação livre. Sobram, portanto, N – 1 partículas livres e este é o conteúdo informacional com base no qual a SQT é calculada. A SQR é dada por $\left(\sum_{i=1}^{N} y_i-\hat{\beta}_0-\hat{\beta}_1 x_{i1}-\hat{\beta}_2 x_{i2}-\ldots-\hat{\beta}_K x_{iK}\right)^2$. Seu cálculo pressupõe a obtenção das estimativas pontuais para o intercepto e para todos os coeficientes, K+1 estatísticas, de modo que restam, N – (K + 1) partículas informacionais livres para, com base nelas, obter a SQR. É também possível calcular os graus de liberdade associados à SQE. Leva-se, para isso, em conta o fato de que SQE = SQT – SQR, ou seja, a SQE pode ser calculada diretamente a partir desta diferença. Os graus de liberdade associados correspondem, analogamente, à diferença dos graus de liberdade correspondentes à SQT e à SQR, i.e., N – 1 – (N – (K +1)) = K.

9 Notação matricial

As propriedades estatísticas do modelo de regressão múltipla podem ser demonstradas mais facilmente com o recurso a uma notação sintética, em que os valores que as variáveis assumem para cada observação são subsumidos a matrizes.

O conjunto de dados disponível, para uma análise de regressão múltipla, pode ser dividido em duas categorias de variáveis, como segue.

(1) A variável dependente, ou explicada, será denotada por:

$$Y=\left[y_i\right]_{Nx1}=\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \ldots \\ Y_N \end{bmatrix}$$

Trata-se, portanto, de um vetor de dimensão N x 1[13].

(2) As variáveis explicativas serão denotadas por:

[13] A letra "x" minúscula é emprega aqui para denotar as dimensões da matriz. Não confundir com as variáveis explicativas.

$$X = \left[x_{ik} \right]_{NxK} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1K} \\ x_{21} & x_{22} & \dots & x_{2K} \\ x_{31} & x_{32} & \dots & x_{3K} \\ \dots & \dots & \dots & \dots \\ x_{N1} & x_{N2} & \dots & x_{NK} \end{bmatrix}$$

A inclusão, na matriz X, de uma primeira coluna com valor unitário para todos os componentes (uma "coluna de uns"), representada sinteticamente por 1_N, permite simplificar as manipulações algébricas. Neste curso, portanto, a matriz X será sempre dada por:

$$X = \left[1_N \, x_{ik} \right]_{Nx(K+1)} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1K} \\ 1 & x_{21} & x_{22} & \dots & x_{2K} \\ 1 & x_{31} & x_{32} & \dots & x_{3K} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{N1} & x_{N2} & \dots & x_{NK} \end{bmatrix}$$

Notar que a dimensão da matriz X é agora N x K+1.

Há duas maneiras de representar a matriz X a partir de vetores. A primeira delas apresenta a matriz X linha por linha:

$$\left[x_i' \right]_{1x(K+1)} = \left[1 \, x_{i1} \dots x_{iK} \right], i = 1, \dots, N$$

A apóstrofe em x_i' indica que se trata de um vetor transposto. Aqui é preciso esclarecer que alguns autores indicam uma linha da matriz X a partir deste vetor transposto, x_i', de modo que por x_i, se represente, pois, a transposição da i-ésima linha de X:

$$\left[x_i \right]_{(K+1)x1} = \begin{bmatrix} 1 \\ x_{i1} \\ x_{i2} \\ x_{i3} \\ \dots \\ x_{iK} \end{bmatrix}, i = 1, \dots, N$$

Em segundo lugar, é possível apresentar X coluna por coluna:

$$\left[x_k \right]_{Nx1} = \begin{bmatrix} x_{1k} \\ x_{2k} \\ x_{3k} \\ \dots \\ x_{Nk} \end{bmatrix}, k = 1, \dots, K$$

O modelo de regressão linear, na notação matricial, assume a forma:

$$Y = X\beta + u$$

Em que u $=[u_i]_{Nx1}=\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \dots \\ u_K \end{bmatrix}$, um vetor de termos de perturbação.

O vetor β, de dimensão (K + 1) X 1, é tal que:

$$\beta=[\beta_k]_{(K+1)x1}=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \dots \\ \beta_K \end{bmatrix}$$

As duas outras maneiras de apresentar o modelo de regressão linear, cada uma delas adotando uma representação vetorial diferente para X, são:

(1) Representação linha a linha (dá destaque às observações):

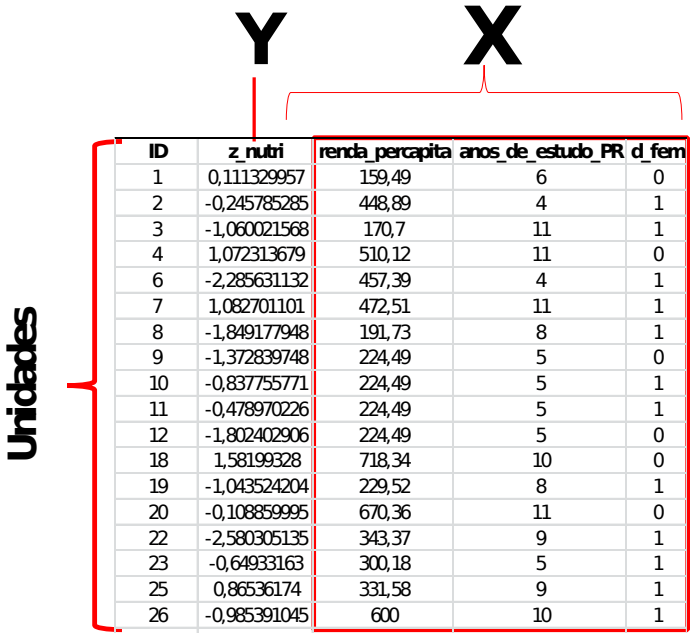
$$y_i = x_i'\beta + u_i,\ i=1,...,N$$

(2) Representação coluna a coluna (dá destaque às variáveis explicativas):

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K + u$$

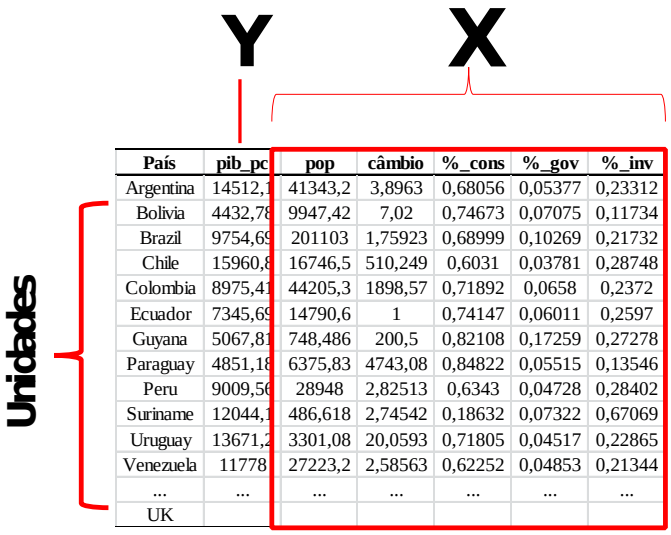
A notação matricial é sintética pois permite representar todas as variáveis explicativas a partir de apenas um símbolo, X. Este é o principal benefício. O custo está associado à dificuldade de compreender exatamente o que a notação representa. Os dois exemplos abaixo procuram mostrar a relação entre a notação matricial e a organização de dados em planilhas.

Figura 2 Visualizando a notação matricial em dados microeconômicos



ID	z_nutri	renda_percapita	anos_de_estudo_PR	d_fem
1	0,111329957	159,49	6	0
2	-0,245785285	448,89	4	1
3	-1,060021568	170,7	11	1
4	1,072313679	510,12	11	0
6	-2,285631132	457,39	4	1
7	1,082701101	472,51	11	1
8	-1,849177948	191,73	8	1
9	-1,372839748	224,49	5	0
10	-0,837755771	224,49	5	1
11	-0,478970226	224,49	5	1
12	-1,802402906	224,49	5	0
18	1,58199328	718,34	10	0
19	-1,043524204	229,52	8	1
20	-0,108859995	670,36	11	0
22	-2,580305135	343,37	9	1
23	-0,64933163	300,18	5	1
25	0,86536174	331,58	9	1
26	-0,985391045	600	10	1

Figura 3 Visualizando a notação matricial em dados macroeconômicos



País	pib_pc	pop	câmbio	%_cons	%_gov	%_inv
Argentina	14512,1	41343,2	3,8963	0,68056	0,05377	0,23312
Bolivia	4432,78	9947,42	7,02	0,74673	0,07075	0,11734
Brazil	9754,69	201103	1,75923	0,68999	0,10269	0,21732
Chile	15960,8	16746,5	510,249	0,6031	0,03781	0,28748
Colombia	8975,41	44205,3	1898,57	0,71892	0,0658	0,2372
Ecuador	7345,69	14790,6	1	0,74147	0,06011	0,2597
Guyana	5067,81	748,486	200,5	0,82108	0,17259	0,27278
Paraguay	4851,18	6375,83	4743,08	0,84822	0,05515	0,13546
Peru	9009,56	28948	2,82513	0,6343	0,04728	0,28402
Suriname	12044,1	486,618	2,74542	0,18632	0,07322	0,67069
Uruguay	13671,2	3301,08	20,0593	0,71805	0,04517	0,22865
Venezuela	11778	27223,2	2,58563	0,62252	0,04853	0,21344
...	...	...	...	...	...	...
UK						

10 Estimação com notação matricial

O problema de obtenção do intercepto e dos coeficientes que correspondem à melhor aproximação linear da FEC populacional, o problema de MQO, pode ser enunciado, com base na notação matricial, da seguinte maneira.

$$min_{\{\beta\}} E\left[\left(y_i - x_i{}'\beta\right)^2\right] (1)$$

Recordando que β é o vetor K+1 x N com todos os parâmetros da FRP.

A solução é alcançada em três passos. No primeiro, é obtida a condição de primeira ordem. Para isso, é necessário recorrer ao cálculo matricial, um tópico geralmente estudado na pós-graduação. O resultado, de qualquer maneira, é:

$$E\left[x_i\left(y_i - x_i'\beta\right)\right] = 0 \quad (2)$$

No segundo passo, manipula-se a equação acima visando isolar o vetor de coeficientes β.

$$E\left[x_i y_i - x_i x_i'\beta\right] = 0$$
$$E\left[x_i y_i\right] = E\left[x_i x_i'\beta\right]$$

Como o vetor β contém apenas parâmetros, i.e., constantes, ele é não-aleatório e pode, pois, ser retirado do operador expectativa:

$$E\left[x_i y_i\right] = E\left[x_i x_i'\right]\beta \quad (2')$$

Neste ponto, cabe abrir um parêntesis com o intuito de recordar alguns conceitos de álgebra linear.

Seja A uma matriz quadrada - i.e., com número de colunas igual ao número de linhas – de dimensão M x M. A matriz inversa de A, caso ela exista, é denotada por A^{-1} tal que $AA^{-1} = I_M$, em que I_M é uma matriz identidade de dimensão M x M; i.e., uma matriz N x N que contém valores unitários em sua diagonal principal e zeros nas demais células. A expressão $E\left[x_i x_i'\right]$ é uma matriz de dimensão (K+1) x (K+1), quadrada, portanto. Por hora, será assumido que a inversa desta matriz existe, o que nem sempre é verdade. Essa discussão será retomada na próxima nota de aula, por enquanto, será assumido que $E\left[x_i x_i'\right]^{-1}$ existe. Segundo a definição de matriz inversa, pois, $E\left[x_i x_i'\right]E\left[x_i x_i'\right]^{-1} = I_{K+1}$.

A matriz identidade é o ente matemático análogo, no "mundo matricial" ao número 1 do mundo escalar. Isso quer dizer que o resultado da multiplicação de um vetor ou matriz pela matriz identidade é equivalente ao próprio vetor ou matriz. Em particular, $I_{K+1}\beta = \beta$.

À luz dos conceitos retomados, a equação (2') acima pode ser manipulada de maneira a isolar o vetor β do lado direito da equação. Basta, para isso, pré-multiplicar os dois lados da equação por $E\left[x_i x_i'\right]$. O termo "pré-multiplicar" estabelece que o elemento $E\left[x_i x_i'\right]$ estará do lado esquerdo nas multiplicações. No "mundo matricial", a posição dos elementos em uma multiplicação altera o resultado, por isso ao invés de simplesmente dizer "multiplicar a equação por A", é preciso dizer "pré-multiplicar a equação por A" ou "pós-multiplicar a equação por A".

Procedendo com a pré-multiplicação de (2') por $E\left[x_i x_i'\right]^{-1}$, obtém-se:

$$E\left[x_i x_i'\right]^{-1} E\left[x_i y_i\right] = E\left[x_i x_i'\right]^{-1} E\left[x_i x_i'\right]\beta$$

Empregando o conceito de matriz inversa:

$$E\left[x_i x_i'\right]^{-1} E\left[x_i y_i\right] = I_{K+1}\beta$$

E, de acordo com a definição de matriz identidade:

$$E\left[x_i x_i'\right]^{-1} E\left[x_i y_i\right] = \beta$$

Finalmente, portanto, tem-se uma expressão para o vetor de parâmetros:

$$\beta = E\left[x_i x_i'\right]^{-1} E\left[X_i y_i\right]$$

Essa expressão, contudo, não pode ser empregada para obter os estimadores de MQO a partir da amostra, uma vez que o operador esperança é um conceito que se aplica apenas à população e, na prática, dispõe-se apenas da amostra. O terceiro passo soluciona este

impasse aplicando o princípio da analogia. Os momentos populacionais são então substituídos pelos momentos amostrais correspondentes, quais sejam, $\left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i,\right)^{-1}$ para $E[x_i x_i']^{-1}$ e $\frac{1}{N}\sum_{i=1}^{N} x_i y_i$, para $E[x_i y_i]$. O estimador para o vetor de coeficientes, obtido (exclusivamente) a partir da informação contida na amostra disponível é, portanto:

$$\hat{\beta}_{MQO}=\left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i,\right)^{-1}\frac{1}{N}\sum_{i=1}^{N} x_i y_i$$

Esta expressão é a fórmula geral do estimador de MQO para o vetor de parâmetros. O recurso a ela permite simplificar consideravelmente o estudo das propriedades estatísticas do estimador de MQO, simplesmente porque é possível concentrar a atenção no vetor completo de parâmetros ao invés de procurar obter as propriedades de cada um dos parâmetros.

Mesmo assim, cabe apresentar a fórmula geral para o estimador do k-ésimo coeficiente, qual seja:

$$\hat{\beta}_{MQOk}=\frac{\sum_{i=1}^{N}\left(y_i-\overline{y}\right)\left(\tilde{x}_{ik}-\overline{\tilde{x}_k}\right)}{\sum_{i=1}^{N}\left(\tilde{x}_{ik}-\overline{\tilde{x}_k}\right)^2}$$

Em que $\tilde{x}_{ik}$ é o resíduo da regressão da k-ésima explicativa contra as demais explicativas, i.e., trata-se da parcela da k-ésima explicativa não correlacionada com as (livre da influência das) demais explicativas. Uma vez que se trata de um resíduo, sua soma e, portanto, média, é nula. Usando unicamente este fato chega-se à fórmula abaixo.

$$\hat{\beta}_{MQOk}=\frac{\sum_{i=1}^{N}\tilde{x}_{ik}\,y_i}{\sum_{i=1}^{N}\tilde{x}_{ik}^{\,2}}(1)$$

A fórmula para o k-ésimo coeficiente é importante pois retoma o conceito de efeito parcial ou "puro", i.e., o efeito da k-ésima explicativa sobre a expectativa condicional de Y, livre da influência das demais explicativas. Ela pode ser escrita de uma maneira alternativa considerando-se que a expressão $\sum_{i=1}^{N}\tilde{x}_{ik}^{\,2}$ é a SQR da regressão da k-ésima explicativa contra as demais explicativas. O R^2 desta regressão é dado por:

$$R^2_k=1-\frac{SQR_k}{SQT_k}=1-\frac{\sum_{i=1}^{N}\tilde{x}_{ik}^{\,2}}{\sum_{i=1}^{N}\left(x_{ik}-\overline{x_k}\right)^2}$$

Manipulando, chega-se a:

$$\sum_{i=1}^{N}\tilde{x}_{ik}^{\,2}=\left(1-R^2_k\right)\sum_{i=1}^{N}\left(x_{ik}-\overline{x_k}\right)^2(2)$$

Incorporando (2) a (1) obtém-se:

$$\hat{\beta}_{MQOk} = \frac{\sum_{i=1}^{N} \tilde{x}_{ik} y_i}{\left(1 - R^2{}_k\right) \sum_{i=1}^{N} \left(x_{ik} - \overline{x_k}\right)^2}$$

E vale a pena recordar que R_k^2é o coeficiente de determinação da regressão da k-ésima explicativa contra as demais.

Apêndice Coeficiente de determinação e penalização da inclusão de explicativas estatisticamente insignificativas

A.1 Variações no R^2 ordinário (não ajustado)

Seja considerada a FRP a seguir.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_K x_{iK} + u_i$$

A partir da estimação dos parâmetros, podem-se obter resíduos da forma a seguir.

$$\hat{u}_i^0 = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_K x_{iK} \quad (1)$$

Uma FRP alternativa, que inclui uma explicativa a mais, x_{K+1}, gera, caso estimada, resíduos dados por:

$$\hat{u}_i^1 = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_K x_{iK} - \hat{\beta}_{K+1} x_{iK+1} \quad (2)$$

Combinando (1) e (2) chega-se a:

$$\hat{u}_i^1 = \hat{u}_i^0 - \hat{\beta}_{K+1} x_{iK+1} \quad (3)$$

De (3) decorre uma relação entre as somas dos quadrados dos resíduos dos dois modelos, conforme demonstrado a seguir.

$$SQR_1 = \sum_{i=1}^{N} \left(\hat{u}_i^1\right)^2 = \sum_{i=1}^{N} \left(\hat{u}_i^0 - \hat{\beta}_{K+1} x_{iK+1}\right)^2 \quad (4)$$

Definindo A ≡ $\hat{u}_i^0$ e B ≡ $\hat{\beta}_K x_{iK+1}$, pode-se reconhecer um quadrado perfeito no interior do somatório da forma $(A-B)^2 = A^2 + B^2 - 2AB$. Deste modo:

$$SQR_1 = \sum_{i=1}^{N} \left(\hat{u}_i^0 - \hat{\beta}_{K+1} x_{iK+1}\right)^2 = \sum_{i=1}^{N} \left(\hat{u}_i^0\right)^2 + \sum_{i=1}^{N} \left(\hat{\beta}_{K+1} x_{iK+1}\right)^2 - 2\sum_{i=1}^{N} \hat{u}_i^0 \hat{\beta}_{K+1} x_{iK+1}$$

A primeira expressão à direita, $\sum_{i=1}^{N} \left(\hat{u}_i^0\right)^2$ é equivalente a SQR_0. A terceira expressão pode ser transformada recordando que, de (3), $\hat{u}_i^0 = \hat{u}_i^1 + \hat{\beta}_{K+1} x_{iK+1}$. Assim:

$$SQR_1 = SQR_0 + \sum_{i=1}^{N} \left(\hat{\beta}_K x_{iK+1}\right)^2 - 2\sum_{i=1}^{N} \left(\hat{u}_i^1 + \hat{\beta}_{K+1} x_{iK+1}\right)\hat{\beta}_{K+1} x_{iK+1} = SQR_0 + \sum_{i=1}^{N} \left(\hat{\beta}_{K+1} x_{iK+1}\right)^2 - 2\hat{\beta}_{K+1}\sum_{i=1}^{N} \hat{u}_i^1 x_{iK+1} + ¿-$$

Uma das propriedades algébricas do estimador de MQO, a qual decorre de uma das condições de primeira ordem para a estimação do modelo (2), garante que $\sum_{i=1}^{N} \hat{u}_i^1 x_{iK+1} = 0$. Desta maneira, chega-se, finalmente, a:

$$SQR_1 = SQR_0 - \sum_{i=1}^{N} \hat{\beta}_{K+1}{}^2 x_{iK+1}{}^2$$

Ou, de maneira equivalente:

$$SQR_0 - SQR_1 = \sum_{i=1}^{N} \hat{\beta}_{K+1}{}^2 x_{iK+1}{}^2 \quad (4')$$

Esta expressão garante que a inclusão de uma explicativa adicional nunca pode aumentar a SQR. Basta notar que a alteração da SQR provocada pela introdução de uma explicativa adicional é dada por $SQR_0 - SQR_1$ a qual é equivalente, por (4') a uma soma de quadrados, operação algébrica que sempre retorna resultado não-negativo. Esta conclusão é intuitiva: a inclusão de uma explicativa adicional não pode reduzir a parcela da variação de Y já explicada pelo modelo. Isso pois, do contrário, a inclusão da explicativa adicional minaria a contribuição das explicativas já incluídas para a explicação da variável dependente, o que não faz sentido. A explicativa adicional, na pior das hipóteses, não acrescenta nada ao poder explicativo do modelo e, na melhor das hipóteses, acrescenta consideravelmente.

É deste princípio fundamental que decorre o fato de que o R^2 nunca cai com a introdução de uma explicativa adicional. Basta considerar que tal estatística tem apenas dois componentes, SQR e SQT. Uma vez que última representa a variabilidade amostral de Y, a qual é completamente independente da composição do modelo linear (FRP) que busca explicá-la, a única via por meio da qual a inclusão de uma explicativa adicional pode afetar o R^2 é por meio do impacto de tal alteração sobre o SQR. As passagens a seguir formalizam este raciocínio.

De (4'), tem-se:

$$SQR_0 = SQR_1 + \sum_{i=1}^{N} \hat{\beta}_{K+1}{}^2 x_{iK+1}{}^2$$

Dividindo ambos os lados da expressão por SQT:

$$\frac{SQR_0}{SQT} = \frac{SQR_1}{SQT} + \frac{\sum_{i=1}^{N} \hat{\beta}_{K+1}{}^2 x_{iK+1}{}^2}{SQT} \quad (4'')$$

Recordando a fórmula do R^2, é possível definir $R_0^2 = 1 - \frac{SQR_0}{SQT}$ e $R_1^2 = 1 - \frac{SQR_1}{SQT}$, de modo que:

$$1 - R_0^2 = 1 - R_1^2 + \frac{\sum_{i=1}^{N} \hat{\beta}_{K+1}{}^2 x_{iK+1}{}^2}{SQT}$$

O que é equivalente a

$$R_0^2 = R_1^2 - \frac{\sum_{i=1}^{N} \hat{\beta}_{K+1}{}^2 x_{iK+1}{}^2}{SQT}$$

Ou:

$$R_1^2 - R_0^2 = \frac{\sum_{i=1}^{N} \hat{\beta}_{K+1}{}^2 x_{iK+1}{}^2}{SQT}$$

A diferença $R_1^2 - R_0^2$ capta o impacto da inclusão de uma explicativa adicional sobre o coeficiente de determinação. Este é claramente não-negativo, conforme a expressão do lado direito deixa claro. Fica, pois, estabelecido que o R^2 não pode cair com a inclusão de uma explicativa adicional. Este resultado pode ser estendido para a inclusão de um conjunto de explicativas adicionais.

A.2 Variações do R^2 ajustado

O R^2 ajustado foi desenhado para não aumentar com a inclusão de explicativas que não tenham relação estatística relevante com Y. É possível verificar que o R^2 ajustado funciona exatamente desta maneira, atentando, por hora, apenas para dois componentes do R^2, a SQR e o número de explicativas. Mais a frente no curso serão discutidas, em complemento, relações entre o R^2 ajustado e os valores de estatísticas empregadas em testes de hipóteses. Desta maneira, pois, por hora, por "relação estatística" entre Y e a explicativa adicional se entende a contribuição dada pela última para a redução da fração da primeira não explicada pelo modelo (e, consequentemente, para o aumento da fração explicada).

O R^2 ajustado para a FRP com K explicativas será representado como

$$\overline{R_0^2} = 1 - \frac{SQR_0}{SQT}\left(\frac{N-1}{N-K-1}\right) \quad (1)$$

E, analogamente, para o modelo com K+1 explicativas:

$$\overline{R_1^2} = 1 - \frac{SQR_1}{SQT}\left(\frac{N-1}{N-K-2}\right) \quad (2)$$

De (1) e (2) tem-se:

$$\overline{R_1^2} - \overline{R_0^2} = \frac{N-1}{SQT}\left[\left(\frac{SQR_0}{N-K-1}\right) - \left(\frac{SQR_1}{N-K-2}\right)\right] (3)$$

Conclusivamente, $S\left(\overline{R_1^2} - \overline{R_0^2}\right) = S\left(\left(\frac{SQR_0}{N-K-1}\right) - \left(\frac{SQR_1}{N-K-2}\right)\right)$ (4).

Em que S(.) é a função que dá o sinal de seu argumento. A última passagem decorre do fato de que (N - 1/SQT) é sempre um valor positivo, não interferindo no sinal de $\overline{R_1^2} - \overline{R_0^2}$.

A equação (4) diz que o R^2 ajustado aumenta com a inclusão de uma explicativa adicional, i.e., $\overline{R_1^2} - \overline{R_0^2} > 0$, sempre que esta explicativa proporciona uma redução na razão entre a SQR e os graus de liberdade a ela associados. Enquanto, portanto, basta, para aumentar o R^2 ordinário (não-ajustado), que a inclusão da explicativa adicional reduza a SQR, este critério se mostra insuficiente para aumentar o R^2 ajustado. É preciso que não apenas a SQR seja reduzida, mas também que isso se dê em uma magnitude superior ao decréscimo nos graus de liberdade imposto pela explicativa adicional.

Colocando de outra maneira, se a inclusão da explicativa adicional não permite reduzir a porção da variação de Y não explicada pelo modelo (SQR) em uma magnitude superior àquela em que o conteúdo informacional efetivo (graus de liberdade) é reduzido, então a explicativa adicional não contribui para o aumento do R^2 ajustado. Neste caso, o beneficio da introdução da explicativa, medido em termos do aumento da variação explicada de Y, se mostra inferior ao custo, medido este em termos da informação bruta sacrificada.

A medida precisa em que a redução da SQR tem de compensar a redução dos graus de liberdade para que o R^2 ajustado aumente pode ser calculada da seguinte maneira. A princípio, é preciso considerar que ter $\overline{R_1^2} > \overline{R_0^2}$ é equivalente a ter 1 - $\overline{R_1^2}$ < 1 - $\overline{R_0^2}$ ou $\frac{1-\overline{R_0^2}}{1-\overline{R_1^2}} > 1$. Esta última inequação é equivalente a $\frac{SQR_0/SQT\,(N-1/N-K-1)}{SQR_1/SQT\,(N-1/N-K-2)} > 1$ e, manipulando, $\frac{SQR_1}{SQR_0} < \frac{N-K-2}{N-K-1}$. Conclui-se, pois, que para que haja um aumento no R^2 ajustado com a inclusão de uma explicativa, a razão entre a SQR pós-inclusão e a SQR pré-inclusão tem de ser menor do que a razão entre os graus de liberdade pós-inclusão e os graus de liberdade pré-inclusão.