O Coeficiente de Determinação R^2 como Instrumento Didático para Avaliar a Utilidade de um Modelo de Regressão Linear Múltipla

Roberto C. Quinino
Edna A. Reis
Departamento de Estatística – ICEx – UFMG - Brasil

Lupércio F. Bessegato
Departamento de Estatística – ICE – UFJF - Brasil

RESUMO

Este artigo apresenta uma proposta de uso do coeficiente de determinação como estatística para um atrativo teste de hipóteses, ignorado pela maioria dos livros didáticos, baseado na distribuição por amostragem Beta. Adicionalmente, mostramos que o valor amostral do r-quadrado múltiplo pode ser obtido com uso de sucessivas regressões lineares simples, viabilizando o seu cálculo em sala de aula por meio de máquinas de calcular básicas.

Palavras-chave: Regressão linear, coeficiente de determinação, r-quadrado, distribuição Beta.

1. Introdução

Quando introduzimos o conceito de regressão linear para os alunos dos primeiros anos de graduação, é comum que já tenham ouvido de professores de outras disciplinas que um bom modelo deveria ter o "famoso" r^2 elevado. Os alunos entendem que o valor de r^2 constitui-se em um grau percentual da qualidade de ajuste de um modelo. A tentativa de mostrar-lhes uma análise mais precisa e que a interpretação do valor impresso pode ser equivocada geralmente não é compreendida, podendo entrar em conflito com explicações dos professores das disciplinas específicas do seu curso. Corroborando a nossa experiência, Goldberg (1991), cita que não é raro ler em relatórios de pesquisa empírica declarações como "eu tenho um r^2 elevado, por isso a minha teoria é boa" ou "o meu r^2 é maior do que o seu, por isso a minha teoria é melhor que a sua". Além disso, é importante salientar que a avaliação tradicional da utilidade do modelo pelo teste F não apresenta uma ordem de grandeza intuitiva como a proporção, utilizada para quantificar r^2, que está amplamente incorporada na sociedade e é de fácil entendimento.

Quando a regressão é múltipla, a dificuldade de explicá-la aos alunos é ainda maior, pois surge o agravante de não existir uma figura introdutória simples, análoga ao gráfico de dispersão, para indicar se um determinado modelo de regressão múltipla será considerado útil. O valor de r^2 certamente transmite uma mensagem preliminar, mas esse valor pode ser ilusório, devido ao pequeno tamanho amostral e, por isso, precisa ser melhor trabalhado para aproveitar a motivação inicial dos alunos.

Entretanto, o exagero das críticas em relação ao uso do r^2 é mais desmotivante aos alunos do que propriamente útil no processo de aprendizado. Goldberg (1991), por exemplo, argumenta que "o mais importante do r^2 é que ele não tem importância no modelo de regressão clássico. Este trata de parâmetros da população, não da qualidade do ajustamento da amostra...". Já Cameron (1993) ressalta que "o r^2 não é um teste estatístico e parece não haver qualquer justificativa intuitiva para seu emprego como estatística descritiva", sugerindo que o valor de r^2 não deveria sequer ser reportado.

Entendemos, entretanto, que o coeficiente de determinação deva ser aproveitado no processo de aprendizado dos alunos. Ele pode ser usado como uma estatística de teste para avaliação da existência de uma relação útil entre a variável resposta e pelo menos uma das variáveis regressoras em um modelo de regressão linear. Em nossa opinião, apesar dos resultados serem equivalentes ao teste F, o entendimento e apelo didático é superior.

Para tanto, consideraremos, neste artigo, o fato de que o R^2 possui uma distribuição amostral Beta (Wheatherburn, 1962). Esta distribuição está disponível em planilhas eletrônicas e na maioria dos softwares estatísticos e pode, assim, ser facilmente utilizada para gerar tabelas similares às da distribuição F, com o objetivo de calcular valores críticos para comparação com o valor observado de r^2. Além disso, o cálculo do r^2 pode ser obtido, mesmo em casos de regressão múltipla, com o uso sucessivo de regressões lineares simples, podendo, assim, ser facilmente trabalhado em sala de aula, uma vez que a maioria das calculadoras básicas realizam os cálculos de regressão linear simples.

2. Conhecimento Teórico e Notação

Em geral, nas aulas de regressão linear, começamos o curso motivando os alunos a citarem situações práticas da vida cotidiana que poderiam ser considerados exemplos de fenômenos que podem ser explicados por um conjunto de variáveis. Tal atividade conta com a contribuição e interesse de vários alunos. Em seguida, é comum expressarmos o modelo matematicamente como:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon ,$$

em que *Y* indica a variável resposta, *X*´s são as *k* variáveis regressoras e β_j, j = 0,1,...,*k*, são chamados de coeficientes de regressão. O termo ε indica o erro aleatório e usualmente é suposto ter distribuição Normal com média zero e variância constante σ^2. Neste artigo, presumimos que tal suposição é aceitável e, assim, não discutiremos a avaliação dos resíduos (Maiores detalhes em Montgomery et al., 2006).

O processo de estimação dos parâmetros β_j, j=0,1,...,*k*, é inicialmente explicado pelo método dos mínimos quadrados e exemplos numéricos podem ser trabalhados com a participação ativa dos alunos, resultando nas estimativas $\hat{\beta}_j$, $j = 0,1,...,k$ dos coeficientes e na equação de regressão estimada, dada por

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_k X_k$$

Em termos gerais, os alunos sentem dificuldades com estes cálculos, pois a maioria possuem máquinas de calcular apenas com a função regressão linear simples e correlação linear de Pearson disponível. Veremos adiante na Seção 6 como aproveitar este recurso dos alunos para o cálculo múltiplo.

Nesta fase, um gráfico de dispersão entre Y observados e $\hat{Y}$ estimados constitui-se em uma ferramenta eficaz para motivar os alunos na construção de indicadores da qualidade do modelo. Motivados pela regressão linear simples, é muito comum que os alunos sugiram o coeficiente de correlação de Pearson (*r*) entre Y e $\hat{Y}$ como uma medida do grau de ajuste. A tentativa de motivá-los a trabalhar com a medida ao quadrado (r^2), denominada coeficiente de determinação, não apresenta uma boa compreensão[1].

[1] Uma curiosidade que pode ser citada é que se usarmos os estimadores de mínimos quadrados de y sobre x e de x sobre y, respectivamente, y=a+bx e x=c+dy, então $r^2 = bd$.

Uma alternativa é construir histogramas suavizados e sobrepostos dos valores observados Y e ajustados $\hat{Y}$ com os dados utilizados na exemplificação numérica. A Figura 1 ilustra esta alternativa, na qual percebe-se que a variância dos $\hat{Y}_i$ na amostra

$$V(\hat{Y}) = \sum_{i=1}^{n} \frac{\left(\hat{Y}_i - \overline{Y}\right)^2}{n-1},$$

(já usando o fato de que a média dos $\hat{Y}_i$ é igual a $\overline{Y}$, a média dos Y_i) é menor (ou igual) à variância dos Y_i na amostra

$$V(Y) = \frac{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}{n-1}.$$

É intuitivo pensar que, quanto menor a variância dos valores estimados em relação à variância dos valores observados, pior é o ajuste do modelo, pois indica que as variáveis regressoras contém "pouca informação" sobre a variável resposta. Assim, é razoável utilizarmos a razão

$$\frac{V(\hat{Y})}{V(Y)} = \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}$$

como uma medida do grau de ajuste do modelo. Sua interpretação é naturalmente percebida como a proporção da variação total de Y explicada pela variação nas variáveis X´s através do modelo de regressão. Denotamos a razão por r^2 (mostrando que, de fato, ela é o coeficiente r elevado ao quadrado) e a intitulamos como coeficiente de determinação ou r-quadrado.

Uma discussão mais avançada que pode ser realizada seria chamar a atenção dos alunos para o fato de que o coeficiente de determinação pode ser escrito como

$$r^2 = \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2 / (n-1)}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 / (n-1)},$$

em que $e_i = Y_i - \hat{Y}_i$ é o *resíduo* da i-ésima observação. Este desenvolvimento permite informar aos alunos que o termo

$$\sum_{i=1}^{n} e_i^2 /(n-1)$$

é um estimador viciado[2] para σ^2, e que o estimador não viciado é dado por

$$\sum_{i=1}^{n} e_i^2 /(n-k-1).$$

Se usarmos esta estimativa não viciada no cálculo do r^2, obteremos uma nova medida, denominada r^2 *ajustado* e representada por $\bar{r}^2$. Esta nova medida possui a propriedade de penalizar o r^2 tradicional pelo número de variáveis explicativas. Ou seja, ao contrário do r^2 tradicional, que sempre aumenta com a entrada de variáveis explicativas, o r^2 ajustado poderá aumentar ou diminuir com a entrada de novas variáveis independentes no modelo. Um problema é que $\bar{r}^2$ pode ser negativo e, assim, dificultar ainda mais a interpretação. Maiores detalhes podem ser obtidos em Gujarati (2009).

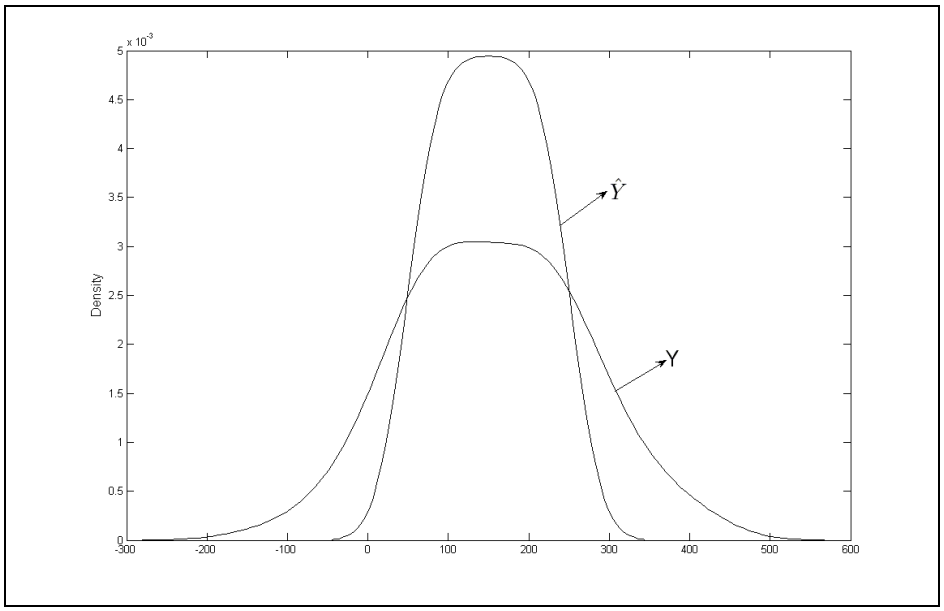

Figura 1: Histogramas suavizados de Y_i e $\hat{Y}_i$.

[2] Um ponto interessante que pode ser citado, é que o denominador da fórmula usual da variância, para uma amostra de uma população normal com média e desvio desconhecidos, pode apresentar diferentes alternativas, por exemplo, *n-1, n*, *n-3*, *n+1*, *n-5/3*, dependendo da propriedade que se deseja: não viciado, erro médio quadrático mínimo, median unbiasedness, modal unbiasedness, etc. O leitor interessado pode consultar maiores detalhes em Sahai e Misra (1992).

3. Um Teste de Hipóteses para o ρ^2

Considerando que os alunos já possuem a percepção de que valores baixos de r^2 constituem-se em indicativo de que o modelo pode não ser útil, a questão adicional e fundamental é explicar que eventuais repetições independentes do experimento provavelmente implicariam (na distribuição por amostragem) em diferentes valores de r^2. Este conjunto de possíveis valores de r^2 é representado pela variável aleatória R^2. É fundamental que entendam a diferença da variável aleatória R^2 e o particular valor r^2 obtido de uma amostra.

Além disso, é também fundamental que entendam que existe um coeficiente de determinação populacional (ρ^2), mas que seu valor é desconhecido. Como trabalhamos com amostras, o que conseguimos é uma estimativa (r^2). O objetivo é saber se um valor de $\rho^2 = 0$ na população poderia facilmente ter originado uma amostra com o particular r^2 observado. Se a resposta for sim, então mesmo que r^2 fosse alto, o modelo não deveria ser considerado útil, pois neste caso, tem-se que $(\beta_1 = \beta_2 = ... = \beta_k = 0)$. Portanto, a questão principal é saber qual deve ser o valor mínimo de r^2 (denotado por *L*) a ser observado para que possamos considerar o modelo útil, ou seja, para concluirmos que muito provavelmente a amostra foi gerada de uma população com $\rho^2 > 0$.

Os alunos precisam entender que, para $\rho^2 = 0$, diferentes amostras obtidas da população poderiam originar diferentes valores de r^2, representados aqui pela variável aleatória R^2. Como os valores possíveis de R^2 variam entre zero e um, alguns alunos sugerem corretamente que a variável aleatória R^2 possui uma distribuição amostral Beta, sem, no entanto, indicarem os parâmetros adequadamente. Mais precisamente, R^2 possui distribuição amostral Beta com parâmetros igual a *k/2* e *(n-k-1)/2*, em que *k* é o número de regressores e *n* o tamanho amostral (Weatherburn, 1962). Considerando que a média da distribuição Beta é data por *k/(n-1)* e representa o valor médio que se obteria para R^2 quando $(\beta_1 = \beta_2 = ... = \beta_k = 0)$, uma possível sugestão para um novo $\bar{r}^2$ ajustado seria $r^2 - k/(n-1)$. Outra possibilidade, procurando manter a nova medida entre zero e um, seria utilizar $r^2[1 - k/(n-1)]$. Estas duas medidas são motivadas pela média da distribuição Beta e possuem propriedades similares ao $\bar{r}^2$ tradicional. A segunda proposta é praticamente igual à sugestão de Goldberger

(1991, p. 178), que propôs $r^2(1-k/n)$, não apresentando, entretanto, a motivação da sugestão.

A Figura 2 ilustra uma distribuição Beta com a indicação de um possível valor mínimo L para r^2 para considerar útil o modelo em análise. O valor L é definido em função do nível de significância desejado para o teste.

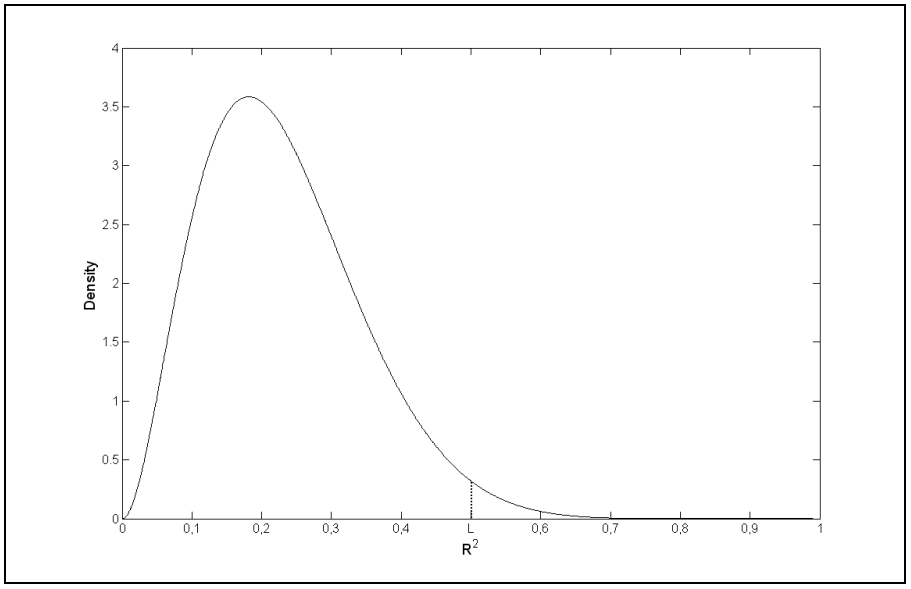

Figura 2: Exemplo da distribuição Beta.

Um exemplo numérico pode consolidar toda a explicação para os alunos: basta mostrar a obtenção de L e comparação com o valor observado r^2. Também se pode obter facilmente a probabilidade de significância (valor p) pelo cálculo de $P(R^2 \geq r^2)$. Observe que estamos destacando a seguinte equivalência entre os testes de hipóteses:

$$\begin{cases} H_0: \beta_1 = \beta_2 = ... = \beta_k = 0 \\ H_1: \text{ Caso Contrário} \end{cases} \Leftrightarrow \begin{cases} H_0: \rho^2 = 0 \\ H_1: \rho^2 > 0 \end{cases}$$

Apesar de não ser objetivo do artigo, também é possível realizar teste de hipóteses da forma $H_0: \beta_j = 0 \, versus \, H_1: \beta_j \neq 0$, para $j = 1, ..., k$, com a utilização do coeficiente de correlação parcial ao quadrado. Por exemplo, se desejamos testar $H_0: \beta_1 = 0$ versus $H_1: \beta_1 \neq 0$, seria necessário calcular o coeficiente de correlação parcial ao quadrado entre Y e X_1, [$r^2_{Y;X_1|X_2,...,X_k}$] desconsiderando o efeito das demais variáveis e considerar que a variável aleatória $R^2_{Y;X_1|X_2,...,X_k}$ possui distribuição por amostragem Beta com parâmetros igual a *1/2* e *(n-k-1)/2*. Para o cálculo do coeficiente de correlação parcial entre Y e X_1, por exemplo, basta calcular a correlação

de Pearson entre os resíduos de, Y em função $X_2,...X_k$, e os resíduos de X_1 em função $X_2,...X_k$. Em geral tal abordagem é bem compreendida pelos alunos, principalmente chamando a atenção de que os resíduos contém a parte não explicada da variável dependente em função das variáveis explicativas. De maneira geral, a explicação anterior pode ser utilizada para todos os testes do tipo $H_0 : \beta_j = 0 \text{ versus } H_1 : \beta_j \neq 0$. Além disso, podemos utilizar os coeficientes de correlação parcial para realizar testes de hipóteses para quaisquer submodelos com a hipótese nula formulada como $H_0 : \beta_1 = \beta_2 = ...\beta_j = 0$ em que *j<k*. Neste caso, o coeficiente de correlação parcial ao quadrado $R^2_{Y;X_1,...,X_j|X_{j+1},...,X_k}$ possui distribuição por amostragem Beta com parâmetros iguais a *j/2* e *(n-k-1)/2*. Quanto ao j>1, devemos entender a correlação no sentido múltiplo. Maiores detalhes sobre esta seção podem ser obtidos em Weatherburn (1962).

4. Exemplo Numérico

O exemplo descrito nesta seção foi retirado de Anderson *et al.* (2002) e será resolvido por meio da planilha eletrônica Microsoft Excel2010©. Poderíamos utilizar softwares como o Minitab, R, Matlab, SPSS, SAS, etc., mas constatamos que muitas vezes o Excel é a única opção disponível. A Figura 3 mostra a planilha com os dados da *potência* (em HP), do *peso* (em libras) e da *velocidade* (em milhas por hora), após percorrida ¼ de milha, para 16 carros GT e esporte (1998 Road & Track Sports & GT Cars). O objetivo é avaliar se a velocidade (*V*) do carro está relacionada ao seu peso e à sua potência. Para isso, podemos inicialmente considerar o modelo

$$V_i = \beta_0 + \beta_1 Peso_i + \beta_2 Pot_i + \varepsilon ,$$

para o qual desejamos realizar o teste das hipóteses $H_0 : \beta_1 = \beta_2 = 0$ *versus* H_1*:* caso contrário. Vamos assumir que as suposições clássicas do modelo de regressão estão satisfeitas.

Primeiramente, obtemos os estimadores de mínimos quadrados $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_2$. Na célula E4 inserimos a função =PROJ.LIN(D2:D17;B2:C17;1;1). Selecionamos um número de células igual ao número de parâmetros a serem estimados, E4:G4. Em

seguida, apertamos F2, seguido por CTRL+SHIFT+ENTER, e obtemos as estimativas de mínimos quadrados. A *velocidade estimada* (VE) é obtida como $VE = \hat{\beta}_0 + \hat{\beta}_1 Peso + \hat{\beta}_2 Pot$. No Excel inserimos na célula H2 a fórmula =$G$2+$F$2*B2+$E$2*C2 e arrastamos até a célula H17. O valor de r^2 é obtido dividindo-se a variância da velocidade estimada pela variância da velocidade observada. No Excel, inserimos na célula I2 a fórmula =VAR(H2:H17)/VAR(D2:D17) obtendo 0.880367. Adotando que R^2 possui distribuição Beta com parâmetros 1 [k/2] e 6.5 [(n-k-1)/2] e considerando um nível de significância de 5%, podemos obter L com ajuda da função BETA.ACUM.INV(0,95;2;6.5). Considerando que r^2=0,8804 e L=0,3693, rejeitamos H_0 e consideramos que o modelo é útil para explicar velocidade por, ao menos, uma das variáveis explicativas ao nível de significância de 5%.

	A	B	C	D
1	**Carros Esporte e GT**	**Peso (lsb)**	**Potência (HP)**	**Velocidade (mph)**
2	Acura Integra Type R	2,577	195	90,7
3	Acura NSX-T	3,066	290	108
4	BMW Z3 2,8	2,844	189	93,2
5	Chevrolet Camaro Z28	3,439	305	103,2
6	Chevrolet Corvette Convertible	3,246	345	102,1
7	Dodge Viper RT/10	3,319	450	116,2
8	Ford Mustang GT	3,227	225	91,7
9	Honda Prelude Type SH	3,042	195	89,7
10	Mercedes-Benz CLK320	3,24	215	93
11	Mercedes-Benz SLK230	3,025	185	92,3
12	Mitsubishi 3000GT VR-4	3,737	320	99
13	Nissan 240SX SE	2,862	155	84,6
14	Pontiac Firebird Trans Am	3,455	305	103,2
15	Porsche Boxster	2,822	201	93,2
16	Toyota Supra Turbo	3,505	320	105
17	Volvo C70	3,285	236	97

Figura 3: Planilha com dados de potência, peso e velocidade para 16 carros GT e esporte.

A Figura 4 ilustra a região de rejeição em função do L. A probabilidade de significância (valor p) pode ser obtida como $P(R^2 \geq 0.8804)$ e no Excel por meio da fórmula =1-DISTBETA(I2;1;6,5).

A Figura 5 apresenta a planilha com os resultados numéricos obtidos para avaliação da utilidade do modelo de regressão linear múltipla, explicados nesta seção.

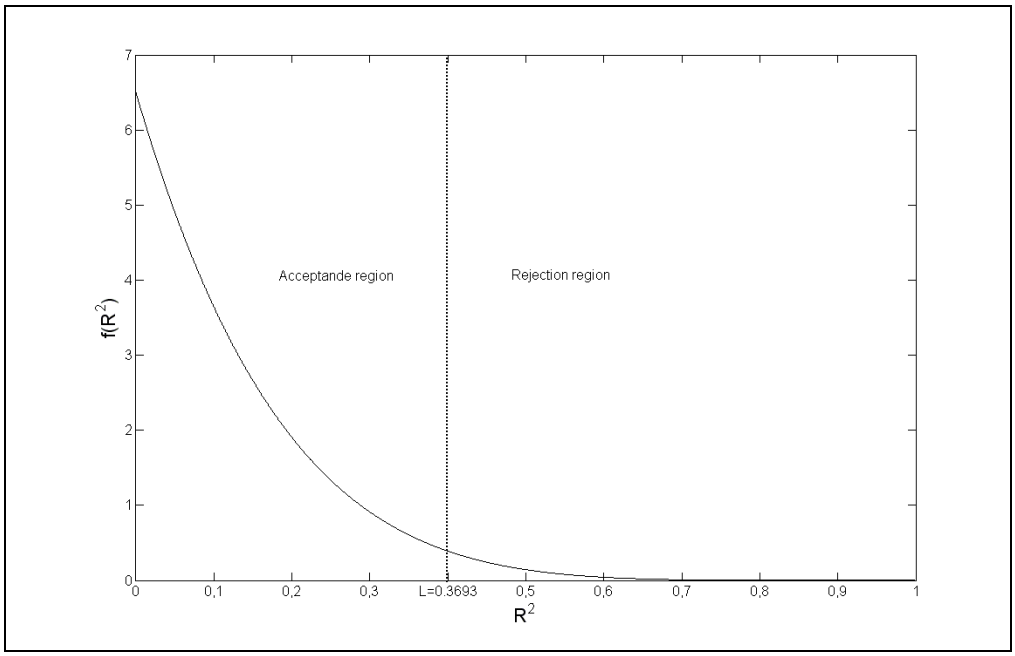

Figura 4: Região de rejeição do teste do exemplo.

E	F	G	H	I	J	K
Beta2	**Beta1**	Beta0	**VE**	**r2**	L	p-value
0,104706	-3,12156	80,48727	92,86059291	0,880367	0,369273	1,01401E-06
			101,2811827			
			91,39890144			
			101,6874236			
			106,4781107			
			117,2443276			
			93,97274492			
			91,40906546			
			92,88510829			
			90,41507576			
			102,3277819			
			87,78272188			
			101,6374786			
			92,72404341			
			103,0519848			
			94,94345611			

Figura 5: Cálculos realizados para avaliação da utilidade do modelo.

Destacamos que todos os resultados obtidos nesta seção seriam iguais se fosse utilizado a estatística F para análise.

5. Tabela para Avaliação do r^2

Esta seção apresenta a Tabela 1, com valores críticos para comparação com com r^2, com objetivo de avaliar a utilidade de um modelo geral $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$ considerando o número de regressores k variando entre 1 e 10 e o tamanho amostral n variando entre 3 e 25. Tal escolha de *n* e *k* contempla a maioria dos exemplos e exercícios contidos em livros didáticos. O nível de significância adotado será o 5% no sentido de que é o mais usual. Assim, pretendemos permitir aos alunos uma rápida tomada de decisão sob a utilidade do modelo ao nível de 5% de significância. Fixados o número de regressores (k) e o tamanho amostral (n), temos o menor r^2 para consideramos o modelo útil (pelo menos um dos parâmetros $(\beta_1, \beta_2, ..., \beta_k)$ é diferente de zero ao nível de significância 5%). Por exemplo, na aplicação da Seção 3, temos n=16 e k=2, resultando que o mínimo valor de r^2, para considerar o modelo útil ao nível de significância 5%, é 0,3693.

Para outros níveis de significância, número de regressores e tamanho amostral, tabelas adicionais podem ser facilmente construídas utilizando a função BETA.ACUM.INV(1-alfa;k/2;(n-k-1)/2) do Excel ou função equivalente de softwares estatísticos. Observe que os valores críticos da distribuição Beta (L) podem também ser obtidos por meio da relação $L = F^* / (1 + F^*)$, em que $F^* = Fk / (n - k - 1)$ e F é o valor crítico da distribuição F de Fisher. Assim, caso seja necessário, o professor poderá utilizar as tradicionais tabelas F para obter os valores críticos da distribuição Beta. Entretanto, recomendamos o uso da tabela Beta uma vez que esta opção mostrou-se mais compreensível e interessante aos alunos.

Tabela 1: Valor mínimo de r^2 para considerar o modelo útil, ao nível de significância 5%.

n \ k	1	2	3	4	5	6	7	8	9	10
3	0,9938	-	-	-	-	-	-	-	-	-
4	0,9025	0,9975	-	-	-	-	-	-	-	-
5	0,7715	0,9500	0,9985	-	-	-	-	-	-	-
6	0,6584	0,8643	0,9664	0,9989	-	-	-	-	-	-
7	0,5693	0,7764	0,9027	0,9747	0,9991	-	-	-	-	-
8	0,4995	0,6983	0,8318	0,9240	0,9797	0,9993	-	-	-	-
9	0,4441	0,6316	0,7645	0,8647	0,9376	0,9830	0,9994	-	-	-
10	0,3993	0,5751	0,7040	0,8060	0,8866	0,9470	0,9855	0,9995	-	-
11	0,3625	0,5271	0,6507	0,7514	0,8347	0,9024	0,9540	0,9873	0,9995	-
12	0,3318	0,4861	0,6039	0,7019	0,7852	0,8559	0,9143	0,9593	0,9887	0,9996
13	0,3058	0,4507	0,5629	0,6574	0,7394	0,8107	0,8722	0,9236	0,9636	0,9898
14	0,2835	0,4200	0,5266	0,6176	0,6974	0,7682	0,8307	0,8852	0,9310	0,9670
15	0,2642	0,3930	0,4945	0,5818	0,6592	0,7287	0,7911	0,8468	0,8957	0,9372
16	0,2473	0,3693	0,4660	0,5497	0,6245	0,6922	0,7539	0,8098	0,8601	0,9045
17	0,2325	0,3482	0,4404	0,5207	0,5929	0,6587	0,7192	0,7747	0,8254	0,8712
18	0,2193	0,3293	0,4174	0,4945	0,5641	0,6280	0,6870	0,7417	0,7922	0,8386
19	0,2075	0,3123	0,3967	0,4707	0,5378	0,5997	0,6572	0,7108	0,7607	0,8071
20	0,1969	0,2970	0,3778	0,4490	0,5137	0,5737	0,6296	0,6819	0,7311	0,7771
21	0,1874	0,2831	0,3607	0,4291	0,4916	0,5496	0,6040	0,6551	0,7032	0,7486
22	0,1787	0,2705	0,3450	0,4109	0,4713	0,5275	0,5802	0,6300	0,6771	0,7218
23	0,1708	0,2589	0,3306	0,3942	0,4525	0,5069	0,5581	0,6066	0,6527	0,6965
24	0,1635	0,2482	0,3173	0,3787	0,4351	0,4878	0,5376	0,5848	0,6297	0,6726
25	0,1569	0,2384	0,3050	0,3644	0,4190	0,4701	0,5184	0,5644	0,6082	0,6502

6. Obtendo o r^2 com uma Calculadora com o Módulo Regressão Linear Simples

Muitas vezes não dispomos de um laboratório com computadores para serem usados nos cálculos desenvolvidos na Seção 4. Apesar de podemos fornecer tabelas da distribuição Beta como a ilustrada na seção anterior ainda temos o problema do cálculo do valor amostral de r^2 que não é praticável de ser realizado manualmente durante o tempo de uma aula usual.

Em geral a maioria dos alunos dos primeiros anos de graduação possuem calculadoras que realizam apenas regressão linear simples e correlação linear de Pearson. Esta seção objetiva mostrar como utilizar este recurso para o cálculo de r^2. A capacidade do método de regressão linear simples ser usado em problemas que demandam regressão linear múltipla com variáveis Dummy's foi tratado por Levin et. al. (1989) para o caso de análise de variância. Entretanto, o artigo não tratou da regressão linear múltipla no caso geral.

Se as variáveis explicativas são não correlacionadas então $r^2 = r^2_{yx_1} + r^2_{yx_2} + ... + r^2_{yx_k}$ e o cálculo demandaria simplesmente calcular o coeficiente de correlação linear entre Y e cada uma das variáveis explicativas $X_1, X_2, ..., X_k$.

Entretanto, na prática as variáveis independentes quase sempre possuem algum grau de correlação, o que complica o cálculo.

Assim, o objetivo seria gerar variáveis não correlacionadas $X_1^*, X_2^*, ..., X_k^*$ a partir de $X_1, X_2, ..., X_k$ e consequentemente utilizarmos $r^2 = r^2_{yx_1^*} + r^2_{yx_2^*} + ... + r^2_{yx_k^*}$ Sem perda de generalidade usaremos o caso k=4 para uma melhor explicação. Numa primeira etapa calculamos os resíduos da regressão linear simples entre X_1 e X_4 e denotamos por R_1; calculamos os resíduos entre X_2 e X_4 e denotamos por R_2; calculamos os resíduos entre X_3 e X_4 e denotamos por R_3. Agora calculamos os resíduos entre R_1 e R_3 e denotamos por R_4 e calculamos os resíduos entre R_2 e R_3 e denotamos por R_5. A variável X_1^* será igual aos resíduos entre R_4 e R_5. A variável X_2^* será igual ao resíduo entre R_2 e R_3; a variável X_3^* será igual a R_3 e $X_4^* = X_4$. Para o exemplo descrito na Seção 4, temos que

$$r^2 = r^2_{yx_1^*} + r^2_{yx_2^*} = 0.007401 + 0.872966 = 0.880367.$$

Nossa sugestão é que o professor solicite um exercício em sala de aula com duas ou três variáveis explicativas, com *n* aproximadamente igual a cinco. Para o caso de três variáveis explicativas, o tempo para resolução, incluindo o teste de hipóteses discutido neste artigo, variou entre quinze e vinte minutos.

7. Conclusões

O objetivo deste trabalho foi expor a importância de apresentar o teste de hipóteses baseado no coeficiente de determinação R^2 e em sua distribuição amostral Beta, como alternativa para testar a significância de um modelo de regressão linear múltipla. Este procedimento se mostrou, pela nossa experiência em sala de aula, mais compreensível e intuitivo aos alunos em relação ao equivalente e tradicional teste F. No nosso entender, a principal explicação é que o valor de r^2 pode ser interpretado como um índice percentual de uso comum aos alunos diferentemente do valor F.

O uso da distribuição Beta também não apresenta problemas, estando disponível inclusive em planilhas como o Excel. A Tabela 3 apresentou grande aceitabilidade e compreensão dos alunos, permitindo-lhes uma rápida tomada de

decisão em relação à utilidade do modelo. Além disso, o valor de r^2 pode ser obtido utilizando o módulo de regressão linear simples rotineiramente presentes em calculadoras básicas e acessíveis.

Finalmente, enfatizamos que é necessário que os alunos entendam que o teste realizado só será aceitável se as hipóteses clássicas relativas ao componente erro sejam julgadas satisfatórias.

8. Referências

Anderson, D. R.; Sweeney, D. J. & Williams, T. A. *Essentials of statistics for business and economics*. Thomson Learning, 2002.

Anscombe, F.J. Graphs in Statistical Analysis. *The American Statistician*, n.27, p.17-21, 1973.

Cameron, S. Why is the R square Adjusted Reposted?. *Journal of Quantitative Economics*, v.9, n.1, p.183-186, 1993.

Foster, F. D; Smith, T. & Whaley, Robert E. Assessing Goodness-of-Fit of Asset Pricing Models: The Distribution of the maximal R2. *The Journal of Finance*, V. LII, n.2, p.591-607, 1997.

Goldberg, A. S. *A Course en Econometrics*. Cambridge, Mass: Havard University, Press, 1991.

Gujarati, D. N. *Basic Econometrics*, 4th ed. McGraw-Hill Companies, Inc, 2003.

Levin, J. R.; Serlin, R. C. & Webne-Berman, L. Analysis of variance though simple correlation. *The American Statistician*, n.43, p.32-34, 1989.

Montgomery, D. C.; Peck, E. A & Vining, G. G. *Introduction to linear regression analysis*, 3nd ed, Wiley-Interscience, 2006.

Sahai, H. & Misra, S. Definitions of sample variance: some teaching problems to be overcome. *The Statistician*, n.41, p.55-64, 1992.

Weatherburn, C. E. A *First Course in Mathematical Statistics*. Cambridge at The University Press, 1962.