Um coeficiente de determinação para Modelos Lineares Generalizados

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

15 de junho de 2021

O coeficiente de determinação, também conhecido como R^2, é bem definido em modelos de regressão linear e mede a proporção de variação na variável dependente explicada pelos preditores incluídos no modelo.

Para estendê-lo para modelos lineares generalizados, usamos a função de variância para definir a variação total da variável dependente, bem como a variação restante da variável dependente após modelar os efeitos preditivos das variáveis independentes. Ao contrário de outras definições que exigem a especificação completa da função de verossimilhança, esta nova definição do R^2 precisa apenas conhecer as funções de média e variância, portanto aplicáveis a quase-modelos mais gerais.

Referência:
A Coefficient of Determination for Generalized Linear Models
The American Statistician · Dabao Zhang (2016).

Considere o modelo de regressão linear,

$$y_i = X_i\beta + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2),$$

independentes e $i = 1, \cdots, n$ onde y_i é o i-ésimo componente do vetor de coluna n-dimensional Y, X_i é a i-ésima linha da matriz $n \times p$ de projeto X e β é um vetor de coluna p-dimensional de coeficientes de regressão desconhecidos.

O estimador $\widehat{\beta}$ de β fornece valores ajustados $\widehat{y}_i(X) = X_i\widehat{\beta}$ e

$$SSE(X) = \sum_{i=1}^{n} \left(y_i - \widehat{y}_i(X)\right)^2,$$

explica a variação nas respostas não explicada pelos preditores disponíveis.

Por outro lado, quando $p = 1$ e $X_i = 1$, ou seja, nenhum preditor é considerado, a estimativa $\widehat{\beta} = \bar{y}$ implica valores ajustados $\widehat{y}_i(\mathbf{1}_n) = \bar{y}$ e

$$SSE(\mathbf{1}_n) = \sum_{i=1}^{n} \left(y_i - \widehat{y}_i(\mathbf{1}_n)\right)^2 = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 = SST,$$

responde pela variação total nas respostas.

Supondo que cada modelo inclui um termo de intercepto, o coeficiente de determinação

$$R^2 = 1 - \frac{SSE(X)}{SSE(\mathbf{1}_n)},$$

mede a proporção de variação nas respostas explicadas pelos preditores disponíveis.

Este coeficiente de determinação é bem definido para modelos de regressão linear e é popularmente usado na prática como uma medida de adequação dos modelos subjacentes.

No entanto, sua extensão para modelos lineares generalizados (GLMs) e outros modelos mais gerais não é direta. Diferentes perspectivas levaram a várias generalizações para o coeficiente de determinação.

Seja $\ell(Y, \mu(X))$ a log-verossimilhança do modelo $\mathrm{E}(Y|X) = \mu(X)$ para os dados observados (Y, X) e $\ell(Y, \mu(\mathbf{1}_n))$, a log-verossimilhança do modelo $\mathrm{E}(Y) = \mu(\mathbf{1}_n) = \mu\mathbf{1}_n$, este chamado de modelo restrito.

Magee (1990) observou a relação entre o R^2 e a estatística da razão de verossimilhança nos modelos de regressão linear,

$$R^2_{LR} = 1 - \exp\left(\frac{2}{n}\ell(Y, \widehat{\mu}(\mathbf{1}_n)) - \frac{2}{n}\ell(Y, \widehat{\mu}(X))\right),$$

onde $\widehat{\mu}(\mathbf{1}_n)$ e $\widehat{\mu}(X)$ são obtidos através da maximização das funções de verossimilhança correspondentes.

Propõe-se, portanto, generalizar R^2 usando R^2_{LR} para modelos mais gerais com função de verossimilhança bem definida. Esta generalização coincide com Maddala (1983) e Cox e Snell (1989)

Para um modelo de regressão logística, os valores perfeitamente ajustados resultam em $\ell(Y, \widehat{\mu}(X)) = 0$, portanto,

$$\max\{R^2_{LR}\} = 1 - \exp\left(\frac{2}{n}\ell(Y, \widehat{\mu}(\mathbf{1}_n))\right).$$

Acontece que, por exemplo, com dados de caso-controle balanceados, $\max\{R^2_{LR}\} = 0.75$. Então o R^2_{LR} é limitado de cima por $\ell(Y, \widehat{\mu}(\mathbf{1}_n))$ e nunca atingirá o valor um.

Para resolver este problema, Nagelkerke (1991) sugeriu a seguinte correção,

$$R^2_N = R^2_{LR} / \max\{R^2_{LR}\}.$$

No entanto, tal correção torna R^2_N inconsistente com a definição clássica de coeficiente de determinação.

Cameron and Windmeijer (1997) propuseram usar a divergência de Kullback-Leibler para quantificar a incerteza remanescente na resposta após a contabilização dos preditores.

Ou seja, a variação na resposta não explicada por X é quantificada pela divergência de Kullback-Leibler estimada entre Y e $\widehat{\mu}(X)$, com $\boldsymbol{I}_n$ a matriz de identidade n-dimensional,

$$\widehat{KL}(Y, \widehat{\mu}(X)) = 2\ell(Y, \widehat{\mu}(\boldsymbol{I}_n)) - 2\ell(Y, \widehat{\mu}(X)).$$

Aqui, a matriz identidade $\boldsymbol{I}_n$ implica que $\mu(\boldsymbol{I}_n)$ é a média de Y seguindo o modelo saturado e $\widehat{\mu}(\boldsymbol{I}_n)$ é seu estimador de máxima verossimilhança (MLE).

Então o coeficiente de determinação é generalizado como,

$$R^2_{KL} = 1 - \widehat{KL}(Y, \widehat{\mu}(X))/\widehat{KL}(Y, \widehat{\mu}(\mathbf{1}_n)).$$

Como ambos $\widehat{KL}(Y, \widehat{\mu}(X)$ e $\widehat{KL}(Y, \widehat{\mu}(\mathbf{1}_n)$ são desvios, R^2_{KL} pode ser interpretado como a razão de redução de desvio devido aos preditores em X.

Todas as generalizações do coeficiente de determinação acima mencionadas são dadas com base na função de verossimilhança completamente especificada, mas não aplicável a GLMs mais gerais, como quase-modelos, que especificam apenas as funções de média e variância.

O R^2 clássico é bem definido para modelos lineares gerais, desde que os termos de erro sejam homocedásticos.

Definir um coeficiente de determinação para modelos lineares generalizados que reconheça a relação entre as funções de média e variância.

Com esse objetivo vamos medir as mudanças de variação ao longo da função de variância.

Todas as generalizações anteriores do coeficiente de determinação abordam assintoticamente as medidas envolvendo a entropia, que mede tanto a incerteza quanto a informação (Shannon, 1948). Em vez disso, seguimos a prática estatística popular para considerar uma medida mais simples de incerteza, ou seja, a variância, que pode ser especificada por meio de um parâmetro de dispersão ϕ e uma função de variância conhecida $V(\cdot)$ em modelos lineares generalizados.

Ou seja, com $\mathrm{E}(y_i|X_i) = \mu(X_i), i = 1, \cdots, n$

$$\mathrm{Var}(y_i|X_i) = \phi V(\mu(X_i))\cdot$$

Em geral, desde que a média $\mu(X_i)$ possa ser bem modelada e ligada de forma adequada a um conjunto de preditores, um modelo linear generalizado com função de variância conhecida $V(\cdot)$ pode ser investigado para a utilidade desses preditores.

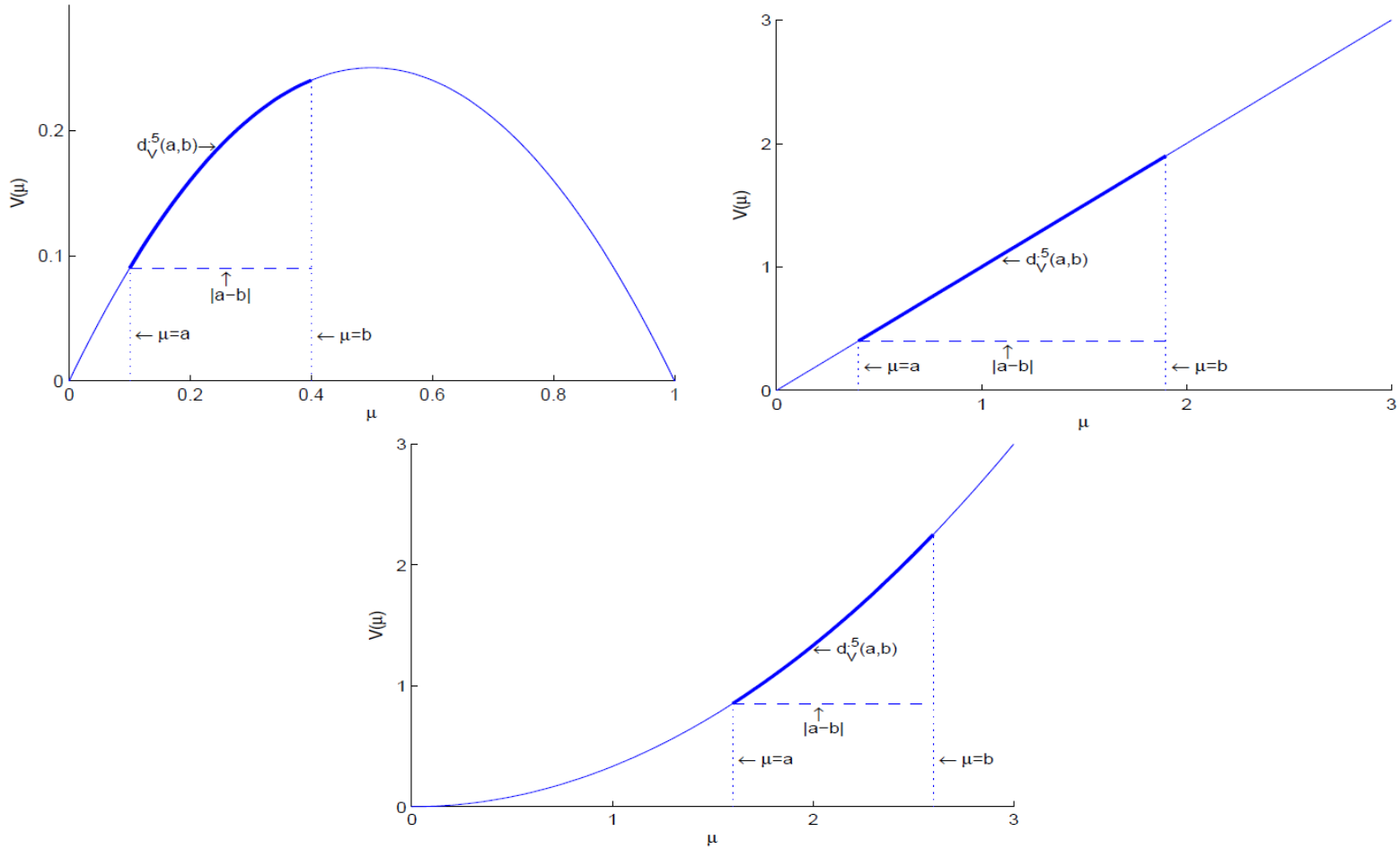
Enquanto a função de variância descreve o efeito da média na variação da variável de resposta além do parâmetro de dispersão, Jorgensen (1987) mostrou que a função de variância $V(\cdot)$ de fato caracteriza a distribuição da família exponencial subjacente.

Para uma variável de resposta com sua média mudando de a para b, sua variação se move de acordo com a função de variância de $\phi V(a)$ para $\phi V(b)$.

Portanto, a mudança de variação da variável de resposta deve ser medida usando, em vez de $(a-b)^2$, o comprimento ao quadrado da função de variância $V(\cdot)$ entre $V(a)$ a $V(b)$, ou seja,

$$d_V(a,b) = \left(\int_a^b \sqrt{1 + \left(V'(t)\right)^2} \mathrm{d}t \right)^2 .$$

Conforme mostrado na Figura a seguir, $d_V(a,b)$ pode diferir dramaticamente da distância euclidiana $(a-b)^2$ quando a função de variância subjacente é não linear.



As funções de variância das distribuições binomial, Poisson e gama. O comprimento da linha grossa é $\sqrt{d_V(a,b)}$ e o comprimento da linha tracejada é $|a-b|$. Esses dois comprimentos podem diferir dramaticamente nas distribuições binomial e gama.

Conforme mostrado em Morris (1982, 1983), muitas distribuições familiares popularmente consideradas exponenciais, como binomial, binomial negativa e distribuições gama, têm funções de variância quadrática.

Assumimos um caso geral, $\nu_2 \neq 0$

$$V(\mu) = \nu_2\mu^2 + \nu_1\mu + \nu_0.$$

Então, $d_V(a, b)$ pode ser calculado como

$$d_V(a,b) = \frac{1}{16\nu_2^2}\left(V'(b)\sqrt{1+\left(V'(b)\right)^2} - V'(a)\sqrt{1+\left(V'(a)\right)^2} + \log\left(V'(b)+\sqrt{1+\left(V'(b)\right)^2}\right) - \log\left(V'(a)+\sqrt{1+\left(V'(b)\right)^2}\right)\right)^2$$

Quando $\nu_2 = 0$, isto é, quando a função de variância é linear ou constante, como no caso da distribuição de Poisson ou distribuição normal, temos

$$d_V(a, b) = (1 + \nu_1^2)(b - a)^2.$$

Enquanto a variação total em Y é $\sum_{i=1}^{n} = d_V(y_i, \widehat{\mu}_i(\mathbf{1}_n))$, o modelo com preditores X reduz a variação inexplicada em Y para $\sum_{i=1}^{n} d_V(y_i, \widehat{\mu}_i(X))$.

Portanto, definimos o coeficiente de determinação como

$$R_V^2 = 1 - \frac{\sum_{i=1}^{n} d_V(y_i, \widehat{\mu}_i(X))}{\sum_{i=1}^{n} d_V(y_i, \widehat{\mu}_i(\mathbf{1}_n))}.$$

Esse coeficiente de determinação é bem definido, desde que as funções de média e variância sejam especificadas, como os quase-modelos. Portanto, $\widehat{\mu}_i(X)$ e $\widehat{\mu}_i(\mathbf{1}_n)$ podem ser derivados de estimadores de quase-verossimilhança, além do MLE.

Como $V'(\cdot)$ é constante para distribuições normais e de Poisson, R^2_V é consistente com a definição clássica do coeficiente de determinação, no caso de modelos de regressão linear seguindo distribuições normais e modelos de regressão log-linear seguindo distribuições de Poisson, ou seja, $R^2_V = R^2$.

Semelhante ao coeficiente de determinação, o coeficiente de determinação parcial é bem definido para modelos lineares, medindo a proporção de variação na variável de resposta não explicada por um conjunto de preditores que podem ser explicados por um conjunto adicional de preditores.

Por exemplo, considerando dois conjuntos de preditores X_1 e X_2 em um modelo de regressão linear, temos

$$R^2(X_2|X_1) = 1 - \frac{SSE(X_1, X_2)}{SSE(X_1)} = \frac{R^2(X_1, X_2) - R^2(X_1)}{1 - R^2(X_1)},$$

medindo a proporção da variação remanescente na resposta, ao incluir X_1, explicada por X_2.

Com a definição de R^2_V, podemos facilmente estendê-lo a um coeficiente de determinação para modelos mais gerais

$$R^2_V(X_2|X_1) = \frac{R^2_V(X_1, X_2) - R^2_V(X_1)}{1 - R^2_V(X_1)}.$$

De fato, quando X_2 é uma variável univariada, também podemos definir uma medida de correlação parcial baseada em modelo entre a resposta e X_2, dado X_1 no modelo, como segue

$$r_V(X_2|X_1) = \text{sign}(\widehat{\beta}_2)\sqrt{R^2_V(X_2|X_1)},$$

onde $\widehat{\beta}_2$ é o coeficiente de regressão de X_2 ao regredir contra X_1 e X_2 com uma função de ligação monotonicamente crescente.

Embora seja bem definido quando as funções de média e variância são conhecidas, R^2_V também sofre com o aumento do número de preditores como o R^2 clássico e pode aumentar mesmo se preditores irrelevantes forem adicionados ao modelo subjacente.

Portanto, as medidas médias da mudança de variação ao longo da função de variância podem ser usadas para levar em consideração os efeitos causados por diferentes números de preditores.

Ou seja, definimos uma versão ajustada de R^2_V da seguinte forma,

$$R^2_{V_{adj}} = 1 - \frac{\sum_{i=1}^{n} d_V(y_i, \widehat{\mu}_i(X))/(n-p)}{\sum_{i=1}^{n} d_V(y_i, \widehat{\mu}_i(\mathbf{1}_n))/(n-1)}.$$

Como o R^2_V, o $R^2_{V_{adj}}$ é bem definido desde que o modelo subjacente esteja bem definido, como quase-modelos, especifique as funções de média e variância.

Conclusão

O coeficiente de determinação, também conhecido como R^2, é uma estatística chave que indica quão bem um modelo incluindo um conjunto de preditores é responsável pela variação na variável de resposta. Embora mostre a utilidade desses preditores no ajuste do modelo, também fornece uma medida de previsibilidade da variável de resposta usando o conjunto de preditores. R^2 pode ser usado para escolher o conjunto ideal de preditores quando o tamanho do modelo, ou seja, o número de preditores, é fixo.

A versão ajustada pode ser usada para comparar modelos incluindo diferentes números de preditores. Por este motivo, o R^2_{adj} também pode ser usado para ajudar na seleção do modelo, seleção de parâmetro de ajuste, etc.

Conclusão

A extensão proposta do R^2_{adj} torna tudo isso possível quando qualquer modelo estatístico com uma função de variância bem definida, como modelos lineares generalizados ou mesmo quase modelos, é considerado.

Como é bem definido para quase-modelos, o R^2_V pode ser usado para avaliar a previsibilidade de modelos construídos em aprendizado de máquina ou aprendizado profundo (Krizhevsky et al., 2012). Por exemplo, ao identificar fatores de risco para medicamentos personalizados ou estudos de câncer (Sirinukunwattana et al., 2016), podemos usar o R^2_V para medir a previsibilidade com fatores de risco disponíveis e investigar a importância de um conjunto de fatores de risco potenciais usando o coeficiente correspondente de determinação parcial.

Conclusão

Algumas referências:

- Dabao Zhang (2016). A Coefficient of Determination for Generalized Linear Models. The American Statistician · December 2016.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25.
- Sirinukunwattana, K., Raza, S., Tsang, Y. W., Snead, D., Cree, I., Rajpoot, N. (2016). Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE Transactions on Medical Imaging, 35, 1196-1206.