

**University of Illinois at Chicago
School of Public Health
Division of Epidemiology and Biostatistics**

*Technical report#:2006-006
May 2006*

Title: R-squared for general regression models in the
presence of sampling weights

Authors: Sally A. Freels and Karabi Sinha

**Affiliation(s): University of Illinois at Chicago, Division of Epidemiology and
Biostatistics**

R-squared for general regression models in the presence of sampling weights

S. Freels¹ and K. Sinha
University of Illinois at Chicago, School of Public Health,
Division of Epidemiology and Biostatistics

Abstract

The coefficient of determination for general regression models is extended to incorporate sampling weights.

Key words: coefficient of determination; maximum likelihood; sampling weights.

¹Address reprint orders to: Sally Freels; University of Illinois at Chicago, School of Public Health, Division of Epidemiology and Biostatistics; 1601 W. Taylor #953 (M/C 923), Chicago, Illinois 60612, USA.

The measure $R^2 = 1 - \{L(0)/L(\hat{\beta})\}^{2/n}$ has been proposed for general regression models using maximum likelihood for parameter estimation, where $L(0)$ is the restricted maximized likelihood under the intercept-only model and $L(\hat{\beta})$ is the unrestricted maximized likelihood under the full model (Maddala, 1983; Magee, 1990; Nagelkerke, 1991). An adjusted version of this measure for discrete models which can reach a maximum value of 1.0, $\bar{R}^2 = R^2 / \{1 - L(0)^{2/n}\}$, has also been proposed (Maddala, 1983; Nagelkerke, 1991). In the presence of sampling weights, where w_i is the number in the population represented by individual i , the appropriate formulas can be obtained by substituting weighted for unweighted maximized likelihoods and substituting $\sum_{i=1}^n w_i$ for n ,

or $R_w^2 = 1 - \{L_w(0)/L_w(\hat{\beta}_w)\}^{2/\sum_{i=1}^n w_i}$ and $\bar{R}_w^2 = R_w^2 / \{1 - L_w(0)^{2/\sum_{i=1}^n w_i}\}$, where $L_w(0)$ and $L_w(\hat{\beta}_w)$ are the restricted and unrestricted maximized weighted likelihoods.

In the case of linear regression, R^2 and R_w^2 are consistent with classical measures of proportion of variation explained expressed as a ratio of sums of squares. Without sampling weights, the restricted and unrestricted log-likelihoods are

$$l(0) = \frac{n}{2} \log(n) - \frac{n}{2} \log(2\pi) - \frac{n}{2} - \frac{n}{2} \log(SST),$$

$$l(\hat{\beta}) = \frac{n}{2} \log(n) - \frac{n}{2} \log(2\pi) - \frac{n}{2} - \frac{n}{2} \log(SSE);$$

therefore $R^2 = 1 - \{L(0)/L(\hat{\beta})\}^{2/n} = 1 - SSE/SST$ (Magee, 1990). In the case of

weighted linear regression, the restricted and unrestricted log-likelihoods are

$$l_w(0) = \frac{\sum_{i=1}^n w_i}{2} \log \left(\sum_{i=1}^n w_i \right) - \frac{\sum_{i=1}^n w_i}{2} \log(2\pi) - \frac{\sum_{i=1}^n w_i}{2} - \frac{\sum_{i=1}^n w_i}{2} \log(SST_w),$$

$$l_w(\hat{\beta}_w) = \frac{\sum_{i=1}^n w_i}{2} \log \left(\sum_{i=1}^n w_i \right) - \frac{\sum_{i=1}^n w_i}{2} \log(2\pi) - \frac{\sum_{i=1}^n w_i}{2} - \frac{\sum_{i=1}^n w_i}{2} \log(SSE_w);$$

therefore $R_w^2 = 1 - \left\{ L_w(0) / L_w(\hat{\beta}_w) \right\}^{2/\sum_{i=1}^n w_i} = 1 - SSE_w / SST_w$.

For discrete models which are products of probabilities rather than densities,

R^2 has a maximum value of $1 - L(0)^{2/n}$ and the adjusted measure $\bar{R}^2 = R^2 / \left\{ 1 - L(0)^{2/n} \right\}$

has a maximum value of 1.0 (Maddala, 1983; Nagelkerke, 1991). In the weighted case,

R_w^2 has a maximum value of $1 - L_w(0)^{2/\sum_{i=1}^n w_i}$ and the adjusted measure

$\bar{R}_w^2 = R_w^2 / \left\{ 1 - L_w(0)^{2/\sum_{i=1}^n w_i} \right\}$ has a maximum value of 1.0.

References

Maddala, G.S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics* (Cambridge University Press).

Magee, L. (1990), R^2 measures based on Wald and likelihood ratio joint significance tests, *Am. Statistician*, **44**, 250-3.

Nagelkerke, N.J.D. (1991), A note on a general definition of the coefficient of determination, *Biometrika*, **78**, 3, 691-2.