FAQ: WHAT ARE PSEUDO R-SQUAREDS?

FAQ: What are pseudo R-squareds?

As a starting point, recall that a non-pseudo R-squared is a statistic generated in ordinary least squares (OLS) regression that is often used as a goodness-of-fit measure. In OLS,

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

where *N* is the number of observations in the model, *y* is the dependent variable, *y*-bar is the mean of the *y* values, and *y*-hat is the value predicted by the model. The numerator of the ratio is the sum of the squared differences between the actual *y* values and the predicted *y* values. The denominator of the ratio is the sum of squared differences between the actual *y* values and their mean.

There are several approaches to thinking about R-squared in OLS. These different approaches lead to various calculations of pseudo R-squareds with regressions of categorical outcome variables.

1. **R-squared as explained variability** – The denominator of the ratio can be thought of as the total variability in the dependent variable, or how much *y* varies from its mean. The numerator of the ratio can be thought of as the variability in the dependent variable that is not predicted by the model. Thus, this ratio is the proportion of the total variability unexplained by the model. Subtracting this ratio from one results in the proportion of the total variability explained by the model. The more variability explained, the better the model.
2. **R-squared as improvement from null model to fitted model** – The denominator of the ratio can be thought of as the sum of squared errors from the null model–a model predicting the dependent variable without any independent variables. In the null model, each *y* value is predicted to be the mean of the *y* values. Consider being asked to predict a *y* value without having any additional information about what you are predicting. The mean of the *y* values would be your best guess if your aim is to minimize the squared difference between your prediction and the actual *y* value. The

numerator of the ratio would then be the sum of squared errors of the fitted model. The ratio is indicative of the degree to which the model parameters improve upon the prediction of the null model. The smaller this ratio, the greater the improvement and the higher the R-squared.

3. **R-squared as the square of the correlation** – The term “R-squared” is derived from this definition. R-squared is the square of the correlation between the model’s predicted values and the actual values. This correlation can range from -1 to 1, and so the square of the correlation then ranges from 0 to 1. The greater the magnitude of the correlation between the predicted values and the actual values, the greater the R-squared, regardless of whether the correlation is positive or negative.

When analyzing data with a logistic regression, an equivalent statistic to R-squared does not exist. The model estimates from a logistic regression are maximum likelihood estimates arrived at through an iterative process. They are not calculated to minimize variance, so the OLS approach to goodness-of-fit does not apply. However, to evaluate the goodness-of-fit of logistic models, several pseudo R-squareds have been developed. These are “pseudo” R-squareds because they look like R-squared in the sense that they are on a similar scale, ranging from 0 to 1 (though some pseudo R-squareds never achieve 0 or 1) with higher values indicating better model fit, but they cannot be interpreted as one would interpret an OLS R-squared and different pseudo R-squareds can arrive at very different values. Note that most software packages report the natural logarithm of the likelihood due to floating point precision problems that more commonly arise with raw likelihoods.

Commonly Encountered Pseudo R-Squareds

Pseudo R-Squared	Formula	Description
Efron's	$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{\pi}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$ $\hat{\pi}$ = model predicted probabilities	Efron's mirrors approaches 1 and 3 from the list above–the model residuals are squared, summed, and divided by the total variability in the dependent variable, and this R-squared is also equal to the squared correlation between the predicted values and actual values. When considering Efron's, remember that model residuals from a logistic regression are not comparable to those in OLS. The dependent variable in a logistic regression is not continuous and the predicted value (a probability) is. In OLS, the predicted values and the actual values are both continuous and on the same scale, so their differences are easily interpreted.

McFadden's	$$R^2 = 1 - \frac{\ln \hat{L}(M_{Full})}{\ln \hat{L}(M_{Intercept})}$$ M_{full} = Model with predictors $M_{intercept}$ = Model without predictors $\hat{L}$ = Estimated likelihood	McFadden's mirrors approaches 1 and 2 from the list above. The log likelihood of the intercept model is treated as a total sum of squares, and the log likelihood of the full model is treated as the sum of squared errors (like in approach 1).The ratio of the likelihoods suggests the level of improvement over the intercept model offered by the full model (like in approach 2).A likelihood falls between 0 and 1, so the log of a likelihood is less than or equal to zero. If a model has a very low likelihood, then the log of the likelihood will have a larger magnitude than the log of a more likely model. Thus, a small ratio of log likelihoods indicates that the full model is a far better fit than the intercept model. If comparing two models on the same data, McFadden's would be higher for the model with the greater likelihood.
McFadden's (adjusted)	$$R_{adj}^2 = 1 - \frac{\ln \hat{L}(M_{Full}) - K}{\ln \hat{L}(M_{Intercept})}$$ $\hat{L}$ = Estimated likelihood	McFadden's adjusted mirrors the adjusted R-squared in OLS by penalizing a model for including too many predictors.If the predictors in the model are effective, then the penalty will be small relative to the added information of the predictors. However, if a model contains predictors that do not add sufficiently to the model, then the penalty becomes noticeable and the adjusted R-squared can *decrease*with the addition of a predictor, even if the R-squared increases slightly.Note that negative McFadden's adjusted R-squared are possible.

Cox & Snell	$$R^2 = 1 - \left\{ \frac{L(M_{Intercept})}{L(M_{Full})} \right\}^{2/N}$$	Cox & Snell's mirrors approach 2 from the list above. The ratio of the likelihoods reflects the improvement of the full model over the intercept model (the smaller the ratio, the greater the improvement).Consider the definition of *L(M)*. *L(M)* is the conditional probability of the dependent variable given the independent variables. If there are N observations in the dataset, then *L(M)* is the product of N such probabilities. Thus, taking the n^{th} root of the product *L(M)* provides an estimate of the likelihood of each Y value. Cox & Snell's presents the R-squared as a transformation of the – $2ln[L(M_{Intercept})/L(M_{Full})]$ statistic that is used to determine the convergence of a logistic regression.Note that Cox & Snell's pseudo R-squared has a maximum value that is not 1: if the full model predicts the outcome perfectly and has a likelihood of 1, Cox & Snell's is then $1-L(M_{Intercept})^{2/N}$, which is less than one.
Nagelkerke / Cragg & Uhler's	$$R^2 = \frac{1 - \left\{ \frac{L(M_{Intercept})}{L(M_{Full})} \right\}^{2/N}}{1 - L(M_{Intercept})^{2/N}}$$	Nagelkerke/Cragg & Uhler's mirrors approach 2 from the list above. It adjusts Cox & Snell's so that the range of possible values extends to 1.To achieve this, the Cox & Snell R-squared is divided by its maximum possible value, $1-L(M_{Intercept})^{2/N}$. Then, if the full model perfectly predicts the outcome and has a likelihood of 1, Nagelkerke/Cragg & Uhler's R-squared = 1.When $L(M_{full}) = 1$, then $R^2 = 1$; When $L(M_{full}) = L(M_{intercept})$, then $R^2 = 0$.

McKelvey & Zavoina	$R^2 = \frac{\hat{V}ar(\hat{y}^*)}{\hat{V}ar(\hat{y}^*) + Var(\varepsilon)}$	McKelvey & Zavoina's mirrors approach 1 from the list above, but its calculations are based on predicting a continuous latent variable underlying the observed 0-1 outcomes in the data. The model predictions of the latent variable can be calculated using the model coefficients (NOT the log-odds) and the predictor variables.McKelvey & Zavoina's also mirrors approach 3. Because of the parallel structure between McKelvey & Zavoina's and OLS R-squareds, we can examine the square root of McKelvey & Zavoina's to arrive at the correlation between the latent continuous variable and the predicted probabilities.Note that, because *y** is not observed, we cannot calculate the variance of the error (the second term in the denominator). It is assumed to be $\pi^2/3$ in logistic models.