

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/287333683>

Pseudo-R² and related measures. Online Supplement 5 to Serious stats: A guide to advanced statistics for the...

Chapter · January 2012

CITATIONS

0

READS

315

1 author:



[Thom S Baguley](#)

Nottingham Trent University

64 PUBLICATIONS 889 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



WORKAGE project on workplace policies that can support engagement and delayed retirement (EU-funded, 2013-2016) [View project](#)



Serious stats [View project](#)

Online Supplement 4

Pseudo- R^2 and related measures

This supplement draws primarily on Chapters 7, 12 and 17.

OS4.1 Variance explained measures for generalized linear models

OS4.1.1 Pseudo- R^2

The deviance for the observed model, null model and saturated model are useful quantities for exploring the fit of a logistic regression. One slightly controversial application of the deviance is to derive a pseudo- R^2 measure from it, known as the loglikelihood or Hosmer and Lemeshow R^2 (Hosmer and Lemeshow, 1989).¹ This is done by expressing the deviance of the model as a proportion of deviance for the null model. If the deviance for the model in question is D_M , loglikelihood pseudo- R^2 is:

$$R_L^2 = \frac{\ln(\ell_M) - \ln(\ell_0)}{\ln(\ell_S) - \ln(\ell_0)} = \frac{D_0 - D_M}{D_0} = 1 - \frac{D_M}{D_0} \quad \text{Equation OS4.1}$$

This is termed a pseudo- R^2 measure because there is no agreed equivalent to R^2 in logistic regression (or other generalized linear models). The problem stems mostly from the problem that R^2 can be defined in several ways. One definition is the improvement in fit from adding predictors to a null model (which R_L^2 attempts to tackle). Another definition is in terms of the square of the correlation between predicted and observed values (see Section OS4.1.3 below). A further definition is in terms of the proportion of explained variation in the data (e.g., R^2 can be calculated by subtracting the unexplained variance from one). For a normal generalized linear model with an identity link, these definitions coincide and lead to the same quantity, but they

will not coincide for a generalized linear model. Applying the logic of the explained variance measure leads to the Cox and Snell pseudo- R^2 :

$$R_{CS}^2 = 1 - \left(\frac{\ell_0}{\ell_M} \right)^{2/N} = 1 - e^{-2/N[\ln(\ell_M) - \ln(\ell_0)]} \quad \text{Equation OS4.2}$$

For a normal generalized linear model this formula has a maximum of one, but for logistic regression its maximum is .75 or lower. A correction known as the Nagelkerke pseudo- R^2 (Nagelkerke, 1991) adjusts it to range between zero and one (by the simple expedient of dividing it by its maximum possible value).² The corrected formula is:

$$R_N^2 = \frac{R_{CS}^2}{1 - \ell_0^{2/N}} = \frac{R_{CS}^2}{1 - e^{2/N \ln(\ell_0)}} \quad \text{Equation OS4.3}$$

The first two measures are often similar (but rarely identical) in value. The Nagelkerke R^2 measure will typically be substantially larger than the other two by virtue of the correction. On the other hand, all pseudo- R^2 measures produce low R^2 values compared to those associated with good fits in least squares regression. For this reason it is inappropriate to compare R^2 with pseudo- R^2 measures (or to compare different pseudo- R^2 variants). Comparisons between pseudo- R^2 must be restricted to the same measure within the same data set to be at all meaningful. Menard (2000) argues that R_L^2 is the preferred measure for such comparisons for two reasons: it is the most ‘intuitively reasonable’ interpretation and is insensitive to base rates (a problem for the other two measures). However, Estrella noted that R_L^2 doesn’t necessarily increase monotonically with the OR in a single predictor logistic regression model (Estrella, 1998; Zheng and Agresti, 2000).

R_L^2 may be useful within a given data set, though an alternative measure calculated from the correlation or squared correlation between observed and predicted responses is also attractive (Agresti, 1996; Zheng and Agresti, 2000). This measure is described in Section OS4.1.3. Its main advantages are that it is readily adapted to other types of generalized linear model and that it has a straightforward interpretation in terms of prediction within the sample. Remember, however, there are serious problems with all these measures. A particular issue is that the homogeneity of variance assumption implicit in any variance explained metric is highly implausible for generalized linear models with a random component other than normal (and is implausible for many normal models). Standardized effect size measures such as R^2 are popular ways to assess the fit of a least square regression model, but have important limitations (see Chapter 7). A pseudo- R^2 measure not only shares these limitations, but introduces new problems that restrict its utility when assessing model fit or comparing two models. They should be used with extreme caution (if at all).

OS4.1.2 Percentage correct classification

For logistic regression it is also possible to classify predictive power in terms of the *percentage correct classification*. This is a crude measure, but one with strong intuitive appeal (and is often reported by logistic regression software). To assess correct classification, a cut-off or threshold is applied to the predictive probability \hat{P}_i for each observation. A common choice is .5. If $\hat{P}_i < .5$ then the outcome is classified as zero (a failure) and if $\hat{P}_i \geq .5$ then the outcome is classified as one

(a success). The main drawbacks are that the proportion of correctly classified responses depends on the chosen cut-off and that much of the potential information in the predictive probabilities is ignored. Moreover, it is trivial to demonstrate that obviously incorrect models can obtain high percentage classification probabilities. Thus, if average $\hat{P}_i = .76$, a model that predicts that all outcomes are successes (and is therefore untrue) will have 76% correct classification. Further discussion of classification approaches can be found in Cohen *et al.* (2003).

OS4.1.3 Assessing the predictive power of a generalized linear model

The multiple correlation coefficient R can be considered as the correlation between the predicted and observed values in a linear regression model. For a normal generalized linear model the squared multiple correlation R^2 will also equal the proportion of variance explained, but this is not true of models with a different random component.

To generalize the predictive power to a wider range of models Zheng and Agresti (2000) argue that it makes sense to define the predictive power as

$$R_{PP} = r_{Y_i \hat{\mu}_i} \quad \text{Equation OS4.4}$$

where Y_i represents the observed outcomes on the untransformed scale and $\hat{\mu}_i$ is the predicted mean of Y_i (conditional on other predictors in the model). Thus $\hat{\mu}_i$ represents $E(Y|X)$, the expectation or mean of Y for a particular set of predictor values. An important feature of this quantity is that both the predictions and observed responses are on the same, untransformed scale. For logistic regression, R_{PP} is the correlation between the observed outcomes (zero or one) and the predictive probability P_i for all N observations. Agresti (1996) prefers the statistic here labeled R_{PP} , but the squared correlation R_{PP}^2 can also be calculated and interpreted as a measure of predictive power.

Agresti (*ibid.*) notes that, like all correlation measures, the value of R_{PP} depends on the range of values in the model (i.e., it will be distorted by range restriction). Furthermore, it has similar drawbacks to pseudo- R^2 measures such as R_L^2 . It might even decrease as model complexity increases, though it usually behaves well (Zheng and Agresti, 2000). Likewise, it should only be used to compare models of the same data (and can't be used to compare models fitted with different random components). Its chief advantage is its ease of interpretation in terms of the predicted outcomes, and it is more likely to match what a researcher is interested in than some competing measures. While using the observed outcomes in the calculation is a strong point in its favor, it does mean that the measure can be sensitive to outliers (though in this respect it is no different from R). Zheng and Agresti (2000) also explore methods for bootstrapping CIs for predictive power.

Example OS4.1 The simple Pearson correlation between *majority* and *problem* for the expenses data (introduced in Example 17.2) is $r = .146$, 95% CI [.070, .221]. In this normal, linear model the size of the majority accounts for about .0215 or 2.15% of the sample variance. The loglikelihood pseudo- R^2 for the logistic regression of the expenses data is:

$$R_L^2 = 1 - \frac{D_M}{D_0} = 1 - \frac{841.6}{855.5} = \frac{13.9}{855.5} = .0162$$

Note that 13.9 is the value of the G^2 statistic in the likelihood ratio test of the model. The Cox and Snell pseudo- R^2 is:

$$R_{CS}^2 = 1 - e^{-2/N[\ln(\ell_M) - \ln(\ell_0)]} = 1 - e^{-2/646[(-420.8) - (-427.7511)]} = .0213$$

To scale it with a maximum of one requires the Nagelkerke version:

$$R_N^2 = \frac{R_{CS}^2}{1 - e^{2/N \ln(\ell_0)}} = \frac{.0213}{1 - e^{2/646 \times (-427.7511)}} = .0290$$

What about predictive power? R_{pp}^2 is obtained by squaring $r_{Y_i \hat{\mu}_i}$, the correlation between the observed outcome (problem) and the predictive probabilities from the model. $r_{Y_i \hat{\mu}_i} = .146$ and so $R_{pp}^2 = .0214$. These values are very similar to the squared simple correlation between majority and problem (but this won't always be true).

Given the similarity in values, R_{pp}^2 is preferable because it is simple to interpret and easy to generalize, though Zheng and Agresti (2000) advocate reporting it in raw correlation form (e.g., $R_{pp} = .146$). A major difficulty is that all these measures 'underplay' the value of the logistic regression model. Proportion of variance explained measures make more sense for continuous outcomes than discrete ones. A much better way of dealing with explanatory power is through graphical summaries. Even relatively simple ones such as those in Figure 17.4 are useful, but it is important to use scales that your audience will understand (e.g., odds and predictive probabilities are better than log odds).

OS4.2 R code for Online Supplement 4

OS4.2.1 Calculating Pseudo- R^2 (Example OS4.1)

The Pearson correlation between majority and expenses problem for the expenses data can be used to give a crude, but potentially misleading, R^2 estimate for the logistic regression of the expenses data.

```
expenses <- read.csv('expenses.csv')
with(expenses, cor(majority, problem)^2)
```

Attempts to get R^2 from the logistic regression produce one of several pseudo- R^2 indices. The loglikelihood R^2 is based on the reduction of residual deviance (obtained after refitting the models for the analysis of the expenses data):

```
majority.10k <- expenses$majority/10000
model.10k <- glm(problem ~ majority.10k, family='binomial',
  data = expenses)
```

```

model.null <- glm(problem ~ 1, family='binomial', data =
  expenses)

m.dev <- model.10k$deviance
n.dev <- model.null$deviance
ll.R2 <- 1 - (m.dev/n.dev)
ll.R2

```

The Cox and Snell pseudo- R^2 uses N and the loglikelihood directly (not deviance):

```

N <- length(expenses$majority)
m.ll <- logLik(model.10k)[1]
n.ll <- logLik(model.null)[1]
cs.R2 <- 1 - exp(-2/N * (m.ll - n.ll))
cs.R2

```

The Nagelkerke R^2 rescales this to have a maximum of one.

```

n.R2 <- cs.R2 / (1 - exp(2/N * n.ll))
n.R2

```

An alternative with a very simple interpretation is the predictive power measure of Agresti and Zheng (which also generalizes very easily to other models).

```

pp.R2 <- cor(model.10k$fitted, expenses$problem)^2
pp.R2

```

These measures tend to underestimate the impact of predictors on discrete outcomes (relative to the R^2 for continuous outcomes using normal linear models). None of these measures are particularly appealing, but the predictive power measure is probably easiest to interpret.

OS4.3 Notes on SPSS syntax for Online Supplement 4

OS4.3.1 Pseudo- R^2 measures

The logistic regression commands in SPSS provide Cox and Snell pseudo- R^2 and Nagelkerke pseudo- R^2 .

```

SPSS data file: expenses.sav

LOGISTIC REGRESSION VARIABLES problem
  /METHOD=ENTER majority
  /SAVE=PRED
  /PRINT=CI(95).

```

Note that the percentage correctly classified is also reported.

Predictive power can be obtained as R or R^2 from a logistic regression of the saved predictions from the previous analysis and the outcome:

```
REGRESSION
  /STATISTICS R
  /DEPENDENT problem
  /METHOD=ENTER PRE_1.
```

OS4.4 Notes

1. Although often referred to as Hosmer and Lemeshow R^2 , the measure appears to have been derived earlier, possibly independently, by several authors (see Zheng and Agresti, 2000; Menard, 2000). It is also sometimes referred to as the McFadden R^2 . This appears to be another instance of Stigler' law.
2. This measure appears to have been first derived by Cragg and Uhler (1970). Likewise, Cox and Snell's measure appears in earlier work by Maddala (see Menard, 2000).

OS4.5 References

- Agresti, A. (1996) *An Introduction to Categorical Data Analysis*. New York: Wiley.
- Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2003) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Mahwah, NJ: Erlbaum.
- Cragg, J. G., and Uhler, R. (1970) The Demand for Automobiles. *Canadian Journal of Economics*, 3, 386–406.
- Estrella, A. (1998) A New Measure of Fit for Equations with Dichotomous Dependent Variables. *Journal of Business and Economic Statistics*, 16, 198–205.
- Hosmer, D. W., and Lemeshow, S. (1989) *Applied Logistic Regression*. New York: Wiley.
- Menard, S. (2000) Coefficients of Determination for Multiple Logistic Regression Analysis. *The American Statistician*, 54, 17–24.
- Nagelkerke, N. J. D. (1991) A Note on a General Definition of the Coefficient of Determination. *Biometrika*, 78, 691–2.
- Zheng, B., and Agresti, A. (2000) Summarizing the Predictive Power of a Generalized Linear Model. *Statistics in Medicine*, 19, 1771–81.