1 R^2 definitions

1.1 SSE based

Classical definition (as in linear model):

$$R^2 = 1 - \frac{SSE(X)}{SST} \tag{1}$$

$$SSE(X) = \sum_{n=1}^{n} (y_i - \hat{y}(x))^2 \tag{2}$$

$$SST = \sum_{n=1}^{n} (y_i - \hat{y})^2 \tag{3}$$

Advantge:

1. clear interpretation, although only in mathematical terms
2. free in used model fitting methods

Disadvantage:

1. is not optimized by by the fitting process (maxiumum likelihood) (MittelbÃ¶ck 1996), therfore
2. it is possible that R^2 decreases when more predictors are added.
3. can be negative when predictor has no explanatory power
4. predictors on poison with log link are not additivly added. E.g. when a model is completly specified by two predictors(Treatment (0 or 1) + Habitat(0 or 1)), including only one predictor explaines the R^2 only by $\frac{1}{3}$ and not $\frac{1}{2}$. Because for person with both Treatment = 1 and Habitat = 1, the expected value will be multiplied on the data scale.

1.2 Likelihood based measures:

Likelihood based R^2 measures do not decrease when predictors are added when using maximum likelihood estimation.

- Likelihood ratio:

$$R^2_{LR} = 1 - \left(\frac{L_{intercept}}{L_{full}}\right)^{\frac{2}{n}} \tag{4}$$

, where L = Likelihood

Advantage:

- measures gemoetric mean improvement per observation
- in linear models equals R^2_{OLS}
- for linear mixed model this definition is interesting. It was also recommended by Kramer(2005). Nakagawa(2013) argue against likelihood based statistic by 1. REML can not be used, 2. it is not clear which is the Null model and 3. it can decrease when adding predictors. To point 1: i think the simple R^2 in simple linear regression is also based on maxiumum likelihood and not on REML estimation when specified this way, otherwise it wouln't be guaranteed

that R^2 increases when predictors are added. To point 3: I'm not sure if this is true when using maximum likelihood on mixed models instead of REML. I think the loglikelhood should always increase or stay the same when adding fixed predictors. I think the same also holds when adding random effects when fitted by ML. To point 2: The parameter estimates of the fixed effects can change when adding random effects. When comparing against a Null-model including the intercept and the random effects the $R-2$ tells us how much of the remaining variance is explained by fixed effects. E.g. a random intercept (only) model has an R^2 of 0.94 comparing against an intercept only model. Including time as fixed effects changes the R^2 to 0.97. Comparing this model against the random intercept model the R^2 would be 0.5. Dummy coded predictors do not contribuite as much because they also can be compensated by random effects. However, the question is also whether we divide by N or by the number of independent observations. Depending on the denominator similar R^2 are obtained as in the wald test approach with different degrees of freedom possibilities by Eduards(2008). When deviding by the number of independent observations for dummy coded fixed effects, it is then similar to comparing the model without random intercept to intercept only model. The question is also wether we use the marginal or the conditional likelihood. R^2 on the conditional likelihood has an easier interpretation and is then very similar to the classical R^2. However, in the mixed model approach the marginal likelihood is optimized.

The non decreasing R^2 property should then only hold with the marginal likelihood (for glmms for sure, but i think also for linear mixed models). The marginal likelihood R^2 is also better suited for comparing more than one random effect. A further question is whether we assume the same covariance structure for the intercept model. Compound symmetry may only hold when all predictors are included. Otherwise an unstructured covariance matrix may be prefered. The likelihood ratio test results may differ depending on this.

In the linear mixed model the regression coefficients are the same, no matter, whether we use the marginal or person specific model. The marginal model only includes the population average, wheras the mixed model in addtion includes the person specific effects. I'm not sure if we can interpret the R^2 increase when adding a random intercept to an intercept only model on the marginal scale. I think we can interpret it somehow as a change of pattern. The conditional likelihood really is about the subjects we observed. In general, the R^2 is based on the observations we observed

I'm not sure if for glmms when we use the marginal likelihood R^2 if this can then be interpreted as the population average based R^2.

Disadvantage

- in logistic regression upper bound is 0.75, even if the model predicts perfectly (p=0 and p=1). Because of this Nagelkerke proposes the correction factor $R^2_N = R^2_{LR}/U$, where $U = 1 - L(0)^{\frac{2}{n}}$ is the maximum value that can be obtained by R^2_{LR} .

- Eduards (2008) focus on REML estimation and propose for fixed effects $R^2_k = \frac{(q-1)v^-1F(\hat{B},\hat{\sum})}{1+(q-1)v^-1F(\hat{B},\hat{\sum})}$, where v is denominator d.f.

- $F(\hat{B},\hat{\sum}) = \frac{(C\hat{B}^T)\left[C(X^T\hat{\sum}^{-1}C^T)\right]^{-1}(C\hat{B})}{rank(C)}$

- Approximations for v include Kenward-rodger Satterhwaithe and residuals methods. The choise

of residual d.f. can substantiall affect the value of R_k^2 They recommend to use the Kenward-Rodger approach.

- The R^2 can decrease when adding predictors

- This formula is only for fixed effects. They mention in the limitation section, that R^2 on a marginal statistics means it cannot be used to determine person-specific goodnes-of-fit. The statistic is for fixed effects given the covariance structure. It assumes, that the covariance structure holds for both the model of interest and the implied null model. The R^2 is somehow similar to comparing a model with random intercept to a model with random intercept and predictors, with adjusted dedgrees of freedom depending on the chosen approach. For dummy coded predictors the R^2 is similar to comparing the model to a model with only an intercept. It does not get eaten up by the random effects. What i do not like so much, is that it increases, when random slopes are added. The model then fits better, but the $R2$ only is for the fitted effects which then usually have a larger p-value and are less "significant". The R^2 can increase alot even when the random slope model is not preffered over the random intercept model. I think it makes then more sense to say that the whole model has a bigger R^2, but the fixed effects contribution is lower (R_{LI} changes this way). It also depends on if we really believe the random slopes (or unstructed covariance matrix) is inherent in the data structure or the result of missing information.

- The authors do not compare their R^2 to the maximum likelihood approach, which i think would also be interesting. They only say that they focus on REML estimation.

- This approach is extended to GLMMs in jaeger (2016). The formula is the same. GLMMs can be expressed by Penalized quasi likelihood(PQL) as an approximate LMM.

- PQL needs to be used. The authors note that for small sample binomial distribution PQL can give biased estimates.

- again only for fixed effects

- They compare their R^2 only to the one proposed by Nakagawa (2013).

- McFadden's:

$$R^2_{LRI} = 1 - \frac{l_{full}}{l_{intercept}} \tag{5}$$

  , where l=loglikelihood

- From Paul Allison: What is the Best R^2 for Logistic Regression? "So, with some reluctance, I ve decided to cross over to the McFadden camp. As Menard (2000) argued, it satisfies almost all of Kvalseth s (1985) eight criteria for a good R2. When the marginal proportion is around .5, the McFadden R2 tends to be a little smaller than the uncorrected Cox-Snell R2. When the marginal proportion is nearer to 0 or 1, the McFadden R2 tends to be larger".

- Kullback-Leibler-divergence:

$$R^2_{KL} = 1 - \frac{D_{full}}{D_{intercept}} \tag{6}$$

  , where D = Deviance = $2[l(y,y) - l(\hat{y},y)] = K(y,\hat{u})$ Cameron (1997).

  \+ measures the proportionate reduction in uncertainty due to the inclusion of regressors

+ for bernoulli reduces to McFadden's R^2
+ R^2_{KL} equals the likelihood ratio index $(1 - l_{full}/l_{intercept})$ only if $l(y, y) = 0$.
+ R^2_{LRI} is a scalar multiple of R^2_{KL} for poisson.
+ For models with canonical link function, R^2 can additionally be interpreted as the fraction of uncertainty explained by the fitted model.

$$R^2_{KL} = \frac{K(\hat{u}, \hat{u}_0}{K(y, \hat{u}_0)}) \tag{7}$$

+ for other links it can be interpreted as measuring the fraction of empirical uncertainty explained by the model.
+ For quasipoisson models assuming Negbin 1 variance function $var(y_i|X_i) = u_i + \alpha * u_i$ the same formula can be used du to cancellation of the common factor $(1+\hat{\alpha})$ (Cameron 1995).
+ The same should then also hold for quasibinomial models?
+ another expression using maximum likelihood estimation $R^2_{KL} = 1 - \frac{l_{max} - l_{fit}}{l_{max} - l_0} = \frac{l_{fit} - l_0}{l_{max} - l_0}$.
+ The measure can be used for fitted means obtained by any estimation method. However, most of the nice properties do not hold then any longer.

- In mixed models the marginal likelihood is otpimized. The deviance function in R returns the deviance of the conditional estimates for GLMMs, which is not optimized by the fitting process. I think for fixed effects the R^2 based on likelihood should not decrease when using $1 - \frac{D_{full}}{D_{intercept}}$. But it is possible that the conditional deviance with less random effects is smaller than with more random effects. I think the main importance is in fixed effects anyway. However, defined this way: $R^2_{KL} = 1 - \frac{l_{max} - l_{fit}}{l_{max} - l_0} = \frac{l_{fit} - l_0}{l_{max} - l_0}$ the non decreasing property should also work for random effects. l_{max} would then just be the estimate with a predictor per observation, but no random effects.

- R^2 for mixed models

- $R^2_{Nakagawa}$

- $R^2_{Ng} = \frac{\sigma^2_f}{\sigma^2_f * \sum_{l=1}^{u} \sigma^2_l * \sigma^2_e * \sigma^2_d}$,
where σ^2_f is the fixed effects variance, σ^2_l are the random effects variance component, σ^2_e is the additive overdispersion variance term and σ^2_d is the distibution spefific variance (e.g. $\frac{pi^2}{3}$ for latent scale bernoulli. In LME σ^2_E replaces the last two variance components. The observation level variance $\sigma^2_E = \sigma^2_d + \sigma^2_e$ can be derived by the Delta method.

+ They wrote 2018 (in their rptR package). We have previously adviced to estimate the distribution specific variance for the link scale approximation as $\frac{pi^2}{3}$, While this gives a reasonable approximation, it is preferable to estimate the link scale variance as$1/(p * (1 - p))$, where p is the expected probability of success. For poisson they recommend 2018 to use the mean of the observations in the sample.

1. advantage

+ simple formula for lots of GLMs

+ when for binomials the distribution variance $\frac{\pi^2}{3}$ is used we assume a latent scale variable y, but we observe only discrete values. I am not sure how useful this latent scale is. Even if the process is in reality interval scaled and we only see discrete values, we do not know the interval range.

+ on the link sale for poisson the predictors are additivly added

+ free in fitting process of variance components

2. Disadvantage

+ When not assuming a latent scale variable the measure from other R^2 measures differs substantially for the bernoulli distribution. The R^2_{Ng} can be much higher or lower. We are interested in the variance that is still left in the model after predictors are included and not in the variance of the intercept only model, which for binomial makes a big difference. E.g a model that estimates p = 1 for y=1 and p = 0 for y= 0 has no variance. On the logit link scale the variance would explode.

+ can decrease when adding predictors.

3. a similar formula is proposed for baysian models

+ Baysian $R^2_s = \frac{V(\sum_{n=1}^{N} \hat{y}_n^s)}{V(\sum_{n=1}^{N} \hat{y}_n^s + V(\sum_{n=1}^{N} e_n^s}$, where $\theta^s, s = 1, ..., S$ are draws from the posterior paramter distribution. For each θ^s we can compute the vector of predicted values $\hat{y}_n^s = E\left(y \mid X_n, \theta^s\right)$ and the vector of errors $e_n^s = y_n - \hat{y}_n^s$.

+ This R^2 can be interpreted as a data-based estimate of the proportion of variance explained for new data.

+ The argue, a new issue then arises,when fitting a set of a models to a single dataset. Now that the denominator of R^2 is no longer fixed, we can no longer interpret an increase in R^2 as a improved fit to a fixed target.

+ By default the random terms Zb are included when computing $\hat{y}$, because we condition on Z when fitting the model.

- the variance components are not added additivly in poisson model.