Analytical computation of the Cox–Snell R^2_{CS} from a reported C statistic

Jérôme Pasquier

7 April 2026

Context

Riley et al.[1] describe a five-step simulation procedure to estimate the Cox–Snell R^2_{CS} from a reported C statistic C and outcome proportion ϕ. The bottleneck of that procedure is fitting a logistic regression model on a large simulated dataset ($n = 10^6$). Because the data-generating distributions are fully specified, the population-level regression coefficients and the resulting R^2_{CS} can be obtained analytically, replacing the simulation with two one-dimensional integrals.

Five-step procedure (Riley et al.)

Following the notation of Riley et al.[1], the steps are:

1. Simulate n participants (e.g. $n = 10^6$).
2. Draw outcomes $Y_i \sim \mathrm{Bernoulli}(\phi)$, where ϕ is the outcome proportion.
3. Simulate linear predictor values LP_i with $\mathrm{LP}_i \mid Y_i = 0 \sim \mathcal{N}(0, 1)$ and $\mathrm{LP}_i \mid Y_i = 1 \sim \mathcal{N}(\mu, 1)$, where
$$\mu = \sqrt{2}\,\Phi^{-1}(C), \tag{1}$$
and $\Phi^{-1}(\cdot)$ is the inverse standard normal CDF.
4. Fit a logistic regression $\mathrm{logit}(p_i) = \alpha + \beta\,\mathrm{LP}_i$.
5. Compute
$$R^2_{\mathrm{CS}} = 1 - \exp\left(-\frac{\mathrm{LR}}{n}\right), \tag{2}$$
where LR is the likelihood-ratio statistic of the fitted model.

Analytical approach

Population-level logistic regression coefficients

Since the conditional distributions of LP_i given Y_i are fully specified, the true conditional probability $\Pr(Y_i = 1 \mid \mathrm{LP}_i)$ can be derived by Bayes' theorem. Let $f_m(\cdot)$ denote the normal PDF with mean m and unit variance. Then

$$\begin{aligned}\Pr(Y_i = 1 \mid \mathrm{LP}_i) &= \frac{f_\mu(\mathrm{LP}_i)\,\phi}{f_0(\mathrm{LP}_i)\,(1-\phi) + f_\mu(\mathrm{LP}_i)\,\phi} \\ &= \sigma\left(\log\frac{\phi}{1-\phi} + \mu\,\mathrm{LP}_i - \frac{\mu^2}{2}\right),\end{aligned}$$

where $\sigma(x) = (1+e^{-x})^{-1}$ is the logistic function. Since the logistic model is correctly specified, the MLE is consistent and the population-level coefficients towards which $(\hat{\alpha}, \hat{\beta})$ converge as $n \to \infty$ are therefore

$$\alpha = \log \frac{\phi}{1-\phi} - \frac{\mu^2}{2}, \qquad \beta = \mu. \tag{3}$$

Expected log-likelihoods

By definition of the LR statistic

$$\mathrm{LR} = 2\bigl(\ell_{\text{fitted}} - \ell_{\text{null}}\bigr),$$

where ℓ denotes the total log-likelihood (sum over all n observations). Dividing both sides by n:

$$\frac{\mathrm{LR}}{n} = 2\left(\frac{\ell_{\text{fitted}}}{n} - \frac{\ell_{\text{null}}}{n}\right).$$

Each term is a sample average

$$\frac{\ell_{\text{fitted}}}{n} = \frac{1}{n}\sum_{i=1}^{n} \log p(Y_i \mid \mathrm{LP}_i, \hat{\alpha}, \hat{\beta}), \qquad \frac{\ell_{\text{null}}}{n} = \frac{1}{n}\sum_{i=1}^{n} \log p(Y_i, \phi).$$

As $n \to \infty$, $(\hat{\alpha}, \hat{\beta}) \to (\alpha, \beta)$ (MLE consistency), so each term is an average of i.i.d. bounded random variables evaluated at their probability-limit parameters. By the law of large numbers, both averages converge to their expectations:

$$\frac{\ell_{\text{fitted}}}{n} \xrightarrow{n\to\infty} \bar{\ell}, \qquad \frac{\ell_{\text{null}}}{n} \xrightarrow{n\to\infty} \bar{\ell}_0,$$

giving $\mathrm{LR}/n \to 2(\bar{\ell} - \bar{\ell}_0)$, where

$$\bar{\ell} = \mathbb{E}\bigl[\log p(Y_i \mid \mathrm{LP}_i, \alpha, \beta)\bigr], \quad \bar{\ell}_0 = \mathbb{E}\bigl[\log p(Y_i, \phi)\bigr].$$

Null model. The null model estimates $\Pr(Y_i = 1) = \phi$, giving

$$\bar{\ell}_0 = \phi \log \phi + (1-\phi)\log(1-\phi).$$

Fitted model. Using $\mathrm{LP}_i \mid Y_i = 1 \sim \mathcal{N}(\mu, 1)$ and $\mathrm{LP}_i \mid Y_i = 0 \sim \mathcal{N}(0, 1)$,

$$\bar{\ell} = \phi \int_{-\infty}^{\infty} \log \sigma(\alpha + \beta\,\mathrm{LP})\, f_\mu(\mathrm{LP})\,\mathrm{dLP} + (1-\phi)\int_{-\infty}^{\infty} \log\bigl(1 - \sigma(\alpha + \beta\,\mathrm{LP})\bigr)\, f_0(\mathrm{LP})\,\mathrm{dLP}, \tag{4}$$

with α and β given by (3). Each integral in (4) is a one-dimensional Gaussian integral evaluated numerically (e.g. with `integrate()` in R).

Final formula

Combining with (2), the asymptotic value of the Cox–Snell R^2_{CS} is

$$\boxed{R^2_{\mathrm{CS}} = 1 - \exp\Bigl(-2\,(\bar{\ell} - \bar{\ell}_0)\Bigr).}$$

R Implementation

```
r2cs_from_auc <- function(auc, prevalence) {
    mu <- sqrt(2) * qnorm(auc)

    # True MLE coefficients for logistic(y ~ lp) as n -> Inf:
    a <- qlogis(prevalence) - mu^2 / 2
    b <- mu

    # Expected log-likelihood under fitted model (asymptotic)
    # Guard against -Inf * 0 = NaN at tails (true limit is 0)
    ll_y1 <- integrate(function(lp) {
        v <- plogis(a + b * lp, log.p = TRUE) * dnorm(lp, mu)
        ifelse(is.finite(v), v, 0)
    }, -Inf, Inf)$value
    ll_y0 <- integrate(function(lp) {
        v <- plogis(a + b * lp, lower.tail = FALSE, log.p = TRUE) * dnorm(lp)
        ifelse(is.finite(v), v, 0)
    }, -Inf, Inf)$value
    ll <- prevalence * ll_y1 + (1 - prevalence) * ll_y0

    # Expected log-likelihood under null model
    ll_null <- prevalence * log(prevalence) +
        (1 - prevalence) * log(1 - prevalence)

    1 - exp(-2 * (ll - ll_null))
}
```

Inverse problem: recovering C from R^2_{CS}

Existence and uniqueness

Fix the outcome proportion ϕ and define the forward map

$$h(C;\phi) \;=\; 1-\exp\Big(-2\left(\bar{\ell}(C;\phi)-\bar{\ell}_0(\phi)\right)\Big), \qquad C\in(\tfrac{1}{2},1),$$

where $\bar{\ell}$ depends on C through $\mu=\sqrt{2}\,\Phi^{-1}(C)$ and the coefficients in (3). The map $C\mapsto\mu$ is strictly increasing, $\mu\mapsto\bar{\ell}$ is strictly increasing (a larger μ separates the two class distributions more, increasing the expected log-likelihood), and $r\mapsto 1-e^{-2r}$ is strictly increasing. Hence $h(\,\cdot\,;\phi)$ is *strictly increasing* on $(\frac{1}{2},1)$, with $h(\frac{1}{2};\phi)=0$ and $h(1;\phi)=R^2_{\mathrm{CS,max}}(\phi)<1$. By the intermediate-value theorem, for every target $\rho\in\big(0,R^2_{\mathrm{CS,max}}(\phi)\big)$ there exists a *unique* $C^*\in(\frac{1}{2},1)$ such that $h(C^*;\phi)=\rho$.

Numerical inversion

Because the integrals in (4) have no closed form in C, the inverse cannot be expressed analytically. It is obtained numerically by solving

$$h(C;\phi)-\rho=0$$

for C via a one-dimensional root-finding algorithm (e.g. Brent's method). Each function evaluation calls the same two Gaussian integrals used in the forward direction, so the inversion is equally fast and deterministic. In R, the implementation reduces to a single `uniroot()` call with the forward function as its objective:

```
auc_from_r2cs <- function(r2cs, phi) {
  uniroot(
    f      = function(C) r2cs_from_auc(C, phi) - r2cs,
    lower  = 0.5 + .Machine$double.eps,
    upper  = 1   - .Machine$double.eps
  )$root
}
```

Benefits over the simulation approach

- **Speed**: two scalar integrals replace fitting a GLM on 10^6 rows; roughly 10,000× faster in practice.
- **Determinism**: no random seed required; the result is the exact asymptotic ($n \to \infty$) value.
- **Inversion** is straightforward.

References

[1] Riley, R. D., Calster, B. V., and Collins, G. S. "A Note on Estimating the Cox-Snell R2 from a Reported C Statistic (AUROC) to Inform Sample Size Calculations for Developing a Prediction Model with a Binary Outcome". In: *Statistics in Medicine* 40.4 (2021), pp. 859–864. DOI: 10.1002/sim.8806.