Goodness of fit for series systems

Alex Towell

Contents

In order to facilitate estimating the performance of the MLE when we do not know $\boldsymbol{\theta}$, assuming the following condition simplifies our task.

Condition 0.1. *The distribution of C_i given K_i is independent of T_i,*

$$\Pr\{C_i = c_i | K_i = j\} = \Pr\{C_i = c_i | K_i = j, T_i = t_i\}. \tag{0.1}$$

In this paper, we do not normally need to assume Condition 0.1 is met, but it is not an unreasonable condition to assume in many cases. For instance, suppose a series system is repaired by replacing a circuit board, where each circuit board consists of one or more components. If component indexed by j fails, then *all* the components that reside on that circuit board are in the candidate set.

Adding Condition 0.1 to the set of conditions allows us to derive the conditional distribution of T_i given C_i without knowing any more details about how C_i is generated.

The primary motivation of knowing the conditional distribution of T_i given C_i is that it alows us to assess the goodness of fit of our model with respect to a masked sample without needing to know $\boldsymbol{\theta}$.[1]

Theorem 0.1. *Assuming Conditions* **??**, **??**, *and 0.1, the conditional pdf of T_i given $C_i = c_i$ is given by*

$$f_{T_i|C_i}(t_i; \boldsymbol{\theta}|c_i) = \frac{R(t_i; \boldsymbol{\theta}) \sum_{j \in c_i} h_j(t_i; \boldsymbol{\theta_j})}{E_{\boldsymbol{\theta}}\left[\sum_{l \in c_i} h_l(T_i; \boldsymbol{\theta_l}) / h(T_i; \boldsymbol{\theta})\right]}. \tag{0.2}$$

Proof. The condition distribution of T_i given $C_i = c_i$ may be rewritten as

$$f_{T_i|C_i}(t_i; \boldsymbol{\theta}|c_i) = \frac{\sum_{j=1}^{m} f_{T_i,K_i,C_i}(t_i, j, c_i; \boldsymbol{\theta})}{\int_0^\infty \sum_{j=1}^{m} f_{T_i,K_i,C_i}(t, j, c_i; \boldsymbol{\theta}) dt}. \tag{0.3}$$

We may rewrite $f_{T_i,K_i,C_i}(t_i, j, c_i; \boldsymbol{\theta})$ as

$$\Pr\{C_i = c_i | T_i = t_i, K_i = j\} f_{T_i,K_i}(t_i, j; \boldsymbol{\theta}).$$

Making this substitution in Equation (0.3) and applying Condition **??** yields

$$f_{T_i|C_i}(t_i; \boldsymbol{\theta}|c_i) = \frac{\sum_{j \in c_i} \Pr\{C_i = c_i | T_i = t_i, K_i = j\} f_{T_i,K_i}(t_i, j; \boldsymbol{\theta})}{\int_0^\infty \sum_{j \in c_i} \Pr\{C_i = c_i | T_i = t, K_i = j\} f_{T_i,K_i}(t, j; \boldsymbol{\theta}) dt}.$$

We may further simplify by applying Condition **??**, yielding

$$f_{T_i|C_i}(t_i; \boldsymbol{\theta}|c_i) = \frac{\Pr\{C_i = c_i | T_i = t_i, K_i = j'\} \sum_{j \in c_i} f_{T_i,K_i}(t_i, j; \boldsymbol{\theta})}{\int_0^\infty \Pr\{C_i = c_i | T_i = t, K_i = j'\} \sum_{j \in c_i} f_{T_i,K_i}(t, j; \boldsymbol{\theta}) dt}.$$

[1] There are many other approaches to this problem, but they often venture into the subject of model selection, which is beyond the scope of this paper.

where $j' \in c_i$. Applying Condition 0.1, we may rewrite the above as

$$f_{T_i|C_i}(t_i;\boldsymbol{\theta}|c_i) = \frac{\Pr\{C_i = c_i|K_i = j'\}\sum_{j\in c_i} f_{T_i,K_i}(t_i,j;\boldsymbol{\theta})}{\Pr\{C_i = c_i|K_i = j'\}\int_0^\infty \sum_{j\in c_i} f_{T_i,K_i}(t_i,j;\boldsymbol{\theta})dt}.$$

where $j' \in c_i$. Canceling the common factors in the above yields the simplification

$$f_{T_i|C_i}(t_i;\boldsymbol{\theta}|c_i) = \frac{\sum_{j\in c_i} f_{T_i,K_i}(t_i,j;\boldsymbol{\theta})}{\int_0^\infty \sum_{j\in c_i} f_{T_i,K_i}(t,j;\boldsymbol{\theta})dt}.$$

The joint pdf $f_{T_i,K_i}(t,j|\boldsymbol{\theta})$ may be written as $h_j(t;\boldsymbol{\theta_j})R(t;\boldsymbol{\theta})$. Performing this substitution yields

$$f_{T_i|C_i}(t_i;\boldsymbol{\theta}|c_i) = \frac{R(t;\boldsymbol{\theta})\sum_{j\in c_i} h_j(t;\boldsymbol{\theta_j})}{\int_0^\infty R(t;\boldsymbol{\theta})\sum_{j\in c_i} h_j(t;\boldsymbol{\theta_j})dt}.$$

Recall that $h(x) = f(x)/R(x)$, and thus $R(x) = f(x)/h(x)$. Making this substitution yields the result

$$f_{T_i|C_i}(t_i;\boldsymbol{\theta}|c_i) = \frac{R(t;\boldsymbol{\theta})\sum_{j\in c_i} h_j(t;\boldsymbol{\theta_j})}{\int_0^\infty f(t;\boldsymbol{\theta})\sum_{j\in c_i} \frac{h_j(t;\boldsymbol{\theta_j})}{h(t;\boldsymbol{\theta})}dt},$$

where we note the denominator is an expectation, and thus

$$f_{T_i|C_i}(t_i;\boldsymbol{\theta}|c_i) = \frac{R(t;\boldsymbol{\theta})\sum_{j\in c_i} h_j(t;\boldsymbol{\theta_j})}{E_{\boldsymbol{\theta}}\left\{\sum_{j\in c_i} \frac{h_j(T_i;\boldsymbol{\theta_j})}{h(T_i;\boldsymbol{\theta})}\right\}}.$$

□

An open real interval is a set of real numbers that contains all real numbers lying strictly between two particular real numbers, e.g., $a < x < b$ defines an interval denoted by (a, b).

The conditional distribution T_i given C_i may be used to assess the goodness-of-fit of MLE $\hat{\theta}$ for an observed masked data sample. In particular, we are interested in finding the set of minimum length intervals satisfying

$$\left\{(l_i,u_i) : \big(b-a \geq u_i - l_i\big) \text{ and } \big(\Pr_\theta\{a < T_i < b\} = 1-\alpha\big) \text{ for all } (a,b)\right\},$$

which we call *prediction intervals*.

For most conditional distributions of T_i given C_i, there is only one prediction interval, e.g., when the component lifetimes are exponentially distributed, the prediction interval is given by

$$(l,u) = \big(0, F_{T_i}^{-1}(1-\alpha;\boldsymbol{\theta})\big)$$

for $i = 1,\ldots,n$ where $F_{T_i}^{-1}$ is the quantile function. (Observe that in the case of exponentially distributed component lifetimes, (l, u) is a constant, which we prove later.)

To assess the goodness-of-fit of $\hat{\boldsymbol{\theta}}$, we partition the masked data set into a *training set* and *test set*. We estimate $\boldsymbol{\theta}$ using the MLE method on the training set, and then we compute (l_i, u_i) for each observation in the test set with respect to $\hat{\boldsymbol{\theta}}$ to see how many of the observations fall outside the $(1-\alpha)\times 100\%$ prediction intervals.

If too many observations fall outside their respective prediction intervals, we have evidence that the estimated series system is not a good fit to the data. There are many reasons this could be the case, e.g., the candidate set conditions may not be sufficiently satisfied or the lifetime distribution of the components may not be a good fit. Of course, for any particular observation, there is a $\alpha \times 100\%$ chance that it will fall outside its prediction interval.[2]

[2]We could construct a $(1-\alpha)\%100$ prediction band where we look to see if any observation falls outside the band.

We may do a similar analysis for right-censoring, e.g., if the i^{th} observation is right-censored, then if

$$\Pr_{\boldsymbol{\theta}}(T_i > \tau_i | C_i = c_i) < \alpha$$

for some sufficiently small α, we may consider this, too, to be evidence of a lack of goodness of fit for the model.