QUALITY ENGINEERING
2025, VOL. 37, NO. 3, 347–358
https://doi.org/10.1080/08982112.2024.2396930

Taylor & Francis
Taylor & Francis Group

Modeling the output from a commercial chemical process using regression models from survival analysis

R. L. J. Coetzer[a,b], D. de Waal[c,d], M. Smuts[e], and I. J. H. Visagie[e]

[a]School of Industrial Engineering, North-West University, Potchefstroom, South Africa; [b]Focus Area for Pure and Applied Analytics, North-West University, Potchefstroom, South Africa; [c]Department of Mathematical Statistics and Actuarial Sciences, University of the Free State, Bloemfontein, South Africa; [d]Department of Statistics, University of Pretoria, Pretoria, South Africa; [e]School of Mathematical and Statistical Sciences, North-West University, Potchefstroom, South Africa

ABSTRACT

This article is concerned with the modeling of the gas output of a commercial chemical plant using the coal sources as predictor variables. We consider the use of two models to incorporate these predictors; the Cox proportional hazards and accelerated failure time regression models. These models are chosen for their simplicity and for the ease with which the effects of explanatory variables can be interpreted. The contribution of this article lies therein that these models are used in the current context for the first time. We show, using both graphical and formal hypothesis testing procedures that these models (with a Weibull baseline distribution) fit observed gas production data well. We provide a discussion of the interpretation of the estimated model parameters and we comment on how these estimates can be of substantial practical value. The large scale of production from the chemical plant in question ensures that potential cost savings and increases in production associated with more accurate models are of great practical importance.

KEYWORDS

Accelerated failure time model; Cox proportional hazards model; gas production; survival analysis

1. Introduction

Due to the internet of things and the drive for digitalization, large amounts of data are generated every second across value chains of complex processing and manufacturing plants; i.e., more online sensors, measurements, scanners etc. are being employed. In a commercial process, equipment is subjected to circumstances and extraneous variables which were not factored into the design, making the real-time and historical data across value chains critical assets for decision making in safe and efficient operations. Qin (2014) motivated that process operations can benefit a great deal from statistical and data-driven methods due to the complexity of process operations. The large data sets from operations becoming available must be utilized for real-time process monitoring and predictive analytics to ensure longer term performance sustainability, process health and conformance to increasing environmental compliance. The coal to gas operating facility at the Sasol South Africa operations is a very complex facility, where the amount and quality of the gas produced depend crucially on the quality of the coal used in the process, see Coetzer, Rossouw, and Lin (2008). The coal used is a blend of six different coal sources which are delivered to the factory *via* overland conveyors. The quality of each coal source is different, and blending is done by stacking and reclaiming to improve the homogeneity of the coal used in the factory. Rossouw, Coetzer, and Le Roux (2018) provided a detailed description of the coal stacking-reclaiming process.

Since the mentioned plant provides roughly 29% of the fuel and gas demand in South Africa, it is critical to develop accurate predictive models for the raw gas produced to conform to specified product quality and volumes. In addition, accurate predictive models are required for real time performance monitoring, detecting performance deviations and diagnostic analysis to ensure process improvement and longer term sustainability. Previous efforts to model the gas

CONTACT R. L. J. Coetzer roelof.coetzer@nwu.ac.za School of Industrial Engineering, North-West University, 11 Hoffman street, Potchefstroom, 2531, South Africa.

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

production were mainly focused on applying response surface modeling and classification approaches.

Coetzer and Keyser (2003) discussed the factorial experimental design employed in 1998 to test the effect of various coal parameters on the performance of the gasification plant. They showed that the stone content in the coal (higher stone content typically reflects in higher ash yield in the coal) has a significant effect on the gas production from the commercial plant. The ash content negatively effects the gasification process through limiting the transfer of heat and mass in the reactor. Coetzer and Keyser (2004) applied response surface modeling and robustness analysis to show that the operability region for higher pure gas production is significantly expanded by changing and controlling the particle size distribution of the coal.

Coetzer, Rossouw, and Lin (2008) developed response surface models for the gas produced as a function of the coal feed and other process variables, and applied dual response optimization to specify optimum operating conditions for maximum production at sustainable rates. Specifically, the authors showed that the effect of the varying coal quality can be minimized by optimizing and controlling the particle size distribution of the coal. de Waal, Coetzer, and van der Merwe (2016b) used a multivariate gamma model for the particle size distribution of the coal feed to classify the gas produced into three distinct classes, while de Waal, Coetzer, and van der Merwe (2016a) presented a new classification scheme using the Dirichlet distribution for a similar classification of the gas produced as a function of the size distribution and chemical analysis of the coal. Recently, the application of machine learning models was communicated in the literature. Ceylan and Ceylan (2021) discussed the application of various machine learning models for predicting the gas yield and product gas heating value as a function of three coal parameters, including mineral matter, and three process variables. They found that a random forest model yielded the best predictive performance for the responses. However, they only focused on prediction and did not provide interpretations of the effect of the input variables on the responses. Chavan et al. (2012) found that an Artificial Neural Network provided better predictions compared to a multivariate regression approach for predicting the gas yield from fluidized-bed coal gasification. They used three coal parameters, including ash yield, and three process variables for prediction, but did not provide any interpretations of the variables' effects.

In this article, we model gas production as a function of the coal sources using parametric proportional hazard regression and accelerated failure time models. The application of these models to commercial observational data is new in the process modeling and engineering literature. The proportional hazard regression and accelerated failure time models provide a probability of achieving a certain performance (e.g. gas production) or higher given explanatory variables, which is different from just predicting the response variable itself. We specifically consider the sources of the coal for various reasons. First, the various coal sources have different coal quality, such as ash yield, which varies widely between the different coal sources. As alluded to above, previous studies have shown that coal quality has a significant effect on the gas production (Coetzer and Keyser 2004; Coetzer, Rossouw, and Lin 2008), which can be modeled by considering the coal sources as predictor variables in the model. Second, coal sources are considered as controllable variables since the sources are blended to specific ratios through stacking and reclaiming (Rossouw, Coetzer, and Le Roux 2018). Therefore, gas production can be improved through manipulating the coal sources blends, which is illustrated in Section 3. Thirdly, since the application of survival analysis models to modeling gas production from a commercial process is new in the literature, we only include the coal sources as predictor variables in order to simplify the models and to demonstrate the interpretation of the variables' effects for process improvement. The importance of coal source as a predictor is underscored by the numerical results shown in Section 3. Note that there are many process variables and extraneous factors that may affect the amount of gas produced, which this article does not aim to investigate extensively. However, a more detailed discussion regarding these factors is provided in the conclusions.

The Cox proportional hazards (CPH) model (Cox 1972) and the accelerated failure time (AFT) model are prominent models in survival analysis which have received considerable attention in the literature for modeling time-to-event data. Smuts and Allison (2020) provided an overview of survival analysis with specific application to credit risk. Majeed (2020) gave an introduction to AFT models with specific application to insurance attrition. Survival analysis is well established in the medical sciences (see *e.g.,* Wei (1992), Ma and Krings (2008), Tolley, Barnes, and Freeman (2016)) and more recently gained traction in reliability studies in engineering; Equeter et al. (2020) used the CPH model to estimate the lifespan of cutting tools whereas Thijssens and Verhagen (2020) investigated factors that affect airline component reliability in order to optimize their fleet utilization. Furthermore, Chen et al. (2020) considered the use of the CPH model in the context of predictive maintenance. However, the use of survival analysis in the modeling of a fully observed continuous response, as a function of predictors, has not received

much attention in the literature. The relevant models are almost exclusively used to predict or explain the time to a specified event. In this article, we use the CPH and AFT as models for the output of a commercial chemical plant using coal sources as predictor variables. The CPH and AFT models were chosen due to their simple interpretations. The CPH model affords the modeler the opportunity to compare the hazard rates associated with different covariate settings, while the AFT model considers the "rate at which time passes" as a function of the covariates (both of these interpretations are discussed in detail later in the article). Importantly, the chosen models allow us to interpret the individual parameter estimates separately, meaning that specific courses of action can be recommended in practical situations. Therefore, this article fills a gap in the literature by providing an alternative approach for modeling process output variables together with interpretations of variable contributions.

In this setup, censoring is not present and the output is fully observed (since the amount of gas produced is fully observed and not censored). We illustrate how the estimated distribution functions (obtained from fitting either the CPH or AFT regression models) can be used to quantify the effect of different coal sources blends on the predicted output.

Being able to accurately model and predict the output from a coal gasification process could be beneficial to the production company. For example, improved coal blends can be determined and used to produce the same or greater volumes of gas at lower feed rates, which will reduce the company's dependence on coal. In addition, improved blends can lead to greater environmental compliance in terms of carbon dioxide production and particulate matter emissions. In Section 3.3 we demonstrate and discuss how the models can be used in order to realize these objectives and potential benefits.

The remainder of the article is structured as follows. Section 2 introduces the survival models used. Section 3 shows the numerical results obtained when fitting the proposed models to observed production data; here we demonstrate that the proposed models provide an acceptable fit to the observed data and we provide interpretations of the fitted regression coefficients. Section 4 provides some concluding remarks as well as directions for further research.

2. Survival models

Survival analysis is commonly employed to model data for which the time until a specified event occurs is of interest. In this article, the response is a fully observed continuous random variable; the amount of gas produced by a commercial facility. Typically, in survival analysis, some censoring mechanism is present. This means that we only observe the exact value of the response in the case where this value is uncensored. In the current context, no censoring is present. As a result, all of the observations are treated as uncensored. Although extending the results to the censored case would be a simple matter, it is not considered here as it is unlikely that the observed gasification output would be censored in practice.

Throughout this article, we limit our analyses to fully parametric versions of the specified models as this substantially simplifies both the implementation and interpretation when compared to their semi and non-parametric counterparts. Although the semi and non-parametric alternatives are more general in the sense that they are not limited to specified parametric forms, our analyses indicate that the parametric models used fit the observed data well and are therefore fit for use. In fact, both graphical and formal goodness-of-fit techniques indicate that the observed data very closely resemble what we would expect if these data originated from the specified models. As a result, we believe that the assumptions made are justified, at least for the data under consideration. Furthermore, as our aim is to provide a method which will be of practical use, we endeavor to provide practitioners with as simple a model as possible. However, since the models advocated for in this article can be used in a much wider context, it is possible that situations exist where these assumptions would not serve the modeler well. In these cases, we would advise using the goodness-of-fit tests outlined below in order to test the parametric assumptions. If these assumptions prove detrimental, then the use of semi or non-parametric models will be required.

We now introduce some notation. Let $Y > 0$ denote the response variable, in our case this is the total gas production, and let $\mathbf{X} = (X_1, \ldots, X_p)^\top$ denote a p-variate column vector of predictors or explanatory variables; specifically in this case each row of $\mathbf{X}$ represents the log-transformed ratios of the coal sources used, see (3) below. The models used below specify both the unconditional and conditional distributions of Y. Consider first the unconditional distribution function denoted by $F_{\boldsymbol{\theta}}$, indexed by a, typically vector valued, parameter $\boldsymbol{\theta}$. This function is known as the baseline distribution and does not depend on the values of the predictors, $\mathbf{X}$. When the values of the predictors are taken into account, we are able to specify the conditional distribution function of Y given $\mathbf{X}$;

$\tilde{F}_{\theta,\phi}$, where ϕ is a p-variate vector of coefficients corresponding to the values of $\mathbf{X}$. We use the notation $\tilde{\cdot}$ throughout to indicate a function conditional on the predictors. Denote the conditional density function of Y by $\tilde{f}_{\theta,\phi}(y|\mathbf{X})$. In the analyses to follow, the survival function of Y plays an important role since it represents the probability of exceeding a given production level, say y. Denote the unconditional and conditional survival functions by $S_\theta(y) = 1 - F_\theta(y)$ and $\tilde{S}_{\theta,\phi}(y|\mathbf{X}) = 1 - \tilde{F}_{\theta,\phi}(y|\mathbf{X})$, respectively. Let $h_\theta(y) = f_\theta(y)/S_\theta(y)$ and $\tilde{h}(y|\mathbf{X}) = \tilde{f}_{\theta,\phi}(y|\mathbf{X})/\tilde{S}_{\theta,\phi}(y|\mathbf{X})$ denote the unconditional and conditional hazard rates, respectively. These hazards are very useful when specifying the survival models used below. Note that $\tilde{h}(y|\mathbf{X})$ provides an intuitive interpretation; it is the instantaneous rate at which the increase in production level will cease, given the predictors $\mathbf{X}$, when y units have already been produced.

Below we discuss two survival models; the CPH and the AFT models. Both of the models considered specify the conditional hazard rate as a function of the baseline hazard as well as some parametric regression function.

2.1. The Cox proportional hazards (CPH) model

The CPH model specifies the conditional hazard rate of Y given the observed predictors $\mathbf{X}$ to be

$$\tilde{h}_{\theta,\beta}(y|\mathbf{X}) = h_\theta(y)\mathrm{e}^{\beta^\top \mathbf{X}},$$

with $\beta = (\beta_1, \ldots, \beta_p)^\top$ indicating the regression coefficients associated with the CPH model. As the name suggests, given two realizations of the predictors, say $\mathbf{X}_1$ and $\mathbf{X}_2$, the resulting conditional hazard functions are proportional to one another;

$$\frac{\tilde{h}_{\theta,\beta}(y|\mathbf{X_1})}{\tilde{h}_{\theta,\beta}(y|\mathbf{X_2})} = \frac{h_\theta(y)\mathrm{e}^{\beta^\top \mathbf{X}_1}}{h_\theta(y)\mathrm{e}^{\beta^\top \mathbf{X}_2}} = \mathrm{e}^{\beta^\top(\mathbf{X}_1 - \mathbf{X}_2)},$$

which is constant for all $y > 0$. The survival function of the CPH model can be expressed as

$$\tilde{S}_{\theta,\beta}(y|\mathbf{X}) = S_\theta(y)^{\mathrm{e}^{\beta^\top \mathbf{X}}}.$$

2.2. The accelerated failure time (AFT) model

The hazard function of the AFT model can, once again, be expressed as a function of the baseline hazard and a regression component;

$$\tilde{h}_{\theta,\gamma}(y|\mathbf{X}) = \mathrm{e}^{-\gamma^\top \mathbf{X}} h_\theta(y\mathrm{e}^{-\gamma^\top \mathbf{X}}),$$

where we use $\gamma = (\gamma_1, \ldots, \gamma_p)^\top$ to indicate the regression coefficients of the model. The resulting survival function can be expressed as

$$\tilde{S}_{\theta,\gamma}(y|\mathbf{X}) = S_\theta(y\mathrm{e}^{-\gamma^\top \mathbf{X}}). \quad (1)$$

The AFT model lends itself to simple interpretations. As can be observed from (1), if $\gamma^\top \mathbf{X} > 0$, then the inclusion of the regression component increases the value of the conditional survival function for every value of y. The reverse implication holds if $\gamma^\top \mathbf{X} < 0$. In fact, we may think of the factor $\mathrm{e}^{-\gamma^\top \mathbf{X}}$ as a constant rate at which "time is accelerated" in the model. As a result, values of $\mathbf{X}$ resulting in larger values of $\mathrm{e}^{-\gamma^\top \mathbf{X}}$ tend to be associated with smaller values of Y.

We now turn our attention to three distributions commonly used as baseline distributions for the CPH and AFT models.

2.3. Baseline distributions

Below, we provide the details of the Weibull, lognormal and extreme value distributions in the context of the CPH and AFT models. Note that additional details related to the Weibull distribution are provided; this is done since the Weibull baseline will be used extensively in the next section.

2.3.1. The Weibull distribution

The *Weibull*(k, λ) distribution has hazard and survival functions

$$h_\theta(y) = ky^{k-1}\lambda^{-k}, \quad S_\theta(y) = \mathrm{e}^{-(y/\lambda)^k},$$

for $y > 0$, $\theta = (k, \lambda)^\top$.

For the CPH model, the hazard and survival functions are

$$\tilde{h}_{\theta,\beta}(y|\mathbf{X}) = \mathrm{e}^{\beta^T \mathbf{X}} ky^{k-1}\lambda^{-k},$$

$$\tilde{S}_{\theta,\beta}(y|\mathbf{X}) = \mathrm{e}^{-(y/\lambda)^k \exp(\beta^\top \mathbf{X})}.$$

In the case of the AFT model, the hazard and survival functions can be expressed as

$$\tilde{h}_{\theta,\gamma}(y|\mathbf{X}) = \mathrm{e}^{-k\gamma^T \mathbf{X}} ky^{k-1}\lambda^{-k},$$

$$\tilde{S}_{\theta,\gamma}(y|\mathbf{X}) = \mathrm{e}^{-(y\lambda^{-1}\mathrm{e}^{-\gamma^\top \mathbf{x}})^k}.$$

Note that, in the case of the Weibull distribution, if we set $\gamma = -\beta/k$, the CPH and AFT models coincide. In general, the CPH and AFT models do not coincide; this property is unique to the Weibull distribution.

2.3.2. The extreme value distribution

Denote the extreme value distribution with parameters $\theta = (\mu, \sigma^2)^\top$, such that $\mu \in \mathbf{R}$ and $\sigma^2 > 0$, by $EV(\mu, \sigma^2)$. The corresponding hazard and survival, for $y > 0$, are

$$h_{\boldsymbol{\theta}}(y) = \frac{t(z)\mathrm{e}^{-t(z)}}{\sigma(1-\mathrm{e}^{-t(z)})}, \quad S_{\boldsymbol{\theta}}(y) = 1-\mathrm{e}^{-t(z)},$$

with $z = (x-\mu)/\sigma$ and $t(z) = \mathrm{e}^{-z}$.

2.3.3. The log-normal distribution

The hazard and survival functions of the $lognormal(\mu, \sigma^2)$ distribution, for $y > 0$ and $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$, are

$$h_{\boldsymbol{\theta}}(y) = (y\sigma)^{-1}\phi(z)\Phi^{-1}(-z), \quad S_{\boldsymbol{\theta}}(y) = 1-\Phi(z),$$

where $z = \sigma^{-1}\log(y-\mu)$, and ϕ and Φ are the density and distribution functions of the standard normal distribution, respectively.

2.3.4. Parameter estimation

The numerical results in Section 3 are obtained by fitting the proposed models to observed data using the method of maximum likelihood. All calculations are performed in the statistical computer software R; see R Core Team (2019).

Given a data set, $(Y_1, \mathbf{X_1}), ..., (Y_n, \mathbf{X_n})$, consisting of n observations on the response and predictors, we may estimate the parameters of the models by maximizing the likelihood functions. In general, the likelihood function of the models considered can be expressed as

$$L(\boldsymbol{\theta}, \phi | (Y_1, \mathbf{X_1}), ..., (Y_n, \mathbf{X_n})) = \prod_{j=1}^{n} \tilde{S}_{\boldsymbol{\theta},\phi}(Y_j|\mathbf{X}_j)\tilde{h}_{\boldsymbol{\theta},\phi}(Y_j|\mathbf{X}_j).$$

The maximum likelihood estimators of $(\boldsymbol{\theta}, \phi)$ are

$$(\hat{\boldsymbol{\theta}}, \hat{\phi}) = \text{argmax } L(\boldsymbol{\theta}, \phi | (Y_1, \mathbf{X_1}), ..., (Y_n, \mathbf{X_n})). \quad (2)$$

Due to the unavailability of explicit expressions for $(\hat{\boldsymbol{\theta}}, \hat{\phi})$ in (2), we use a non-linear optimization routine in order to arrive at parameter estimates. When fitting the AFT model using the various baseline distributions described above, we use the *aftreg* function in the *eha* package in R; see Broström (2020) and Broström (2012). However, the mentioned package is not used when fitting the CPH model; the parameter estimation procedure employed in this case is explained below.

When fitting the CPH model to the observed data, we use the *optim* function and we specify the optimization method to be "BFGS"; see Broyden (1970), Fletcher (1970), Goldfarb (1970) and Shanno (1970). This optimization algorithm is a quasi-Newton method which uses gradients in order to perform the required optimization. The procedure requires the specification of starting values for the optimizer. In order to arrive at parameter estimates we use techniques similar to those described in Visagie (2018). The procedure is sketched below; for more details, see Visagie (2018).

We begin by fitting the baseline distribution using maximum likelihood. In order to provide starting values for the algorithm, we specify a range of values that we believe to be likely to include the optimal parameter values. Admittedly, this specification is subjective. However, we emphasize that this step is merely to obtain starting values for the optimizer. Given the chosen range for the parameters, we simulate a large number of possible values for each of the parameters to be estimated from independent uniform distributions. Thereafter, we evaluate the likelihood function for each of the resulting parameter combinations and simply choose the parameter set resulting in the largest likelihood as starting values for the optimization procedure.

After fitting the baseline distribution to the data, we are required to estimate the parameters of the regression component of the CPH model (note that the parameter values which maximize the likelihood of the baseline models will not necessarily maximize the likelihood of the CPH model). Again, we require starting values for the parameters associated with the baseline as well as those associated with the regression coefficients. For the parameters associated with the baseline distribution, we use the estimates obtained when fitting the baseline distribution. For the regression coefficients we randomly simulate possible values from independent uniform distributions as before. The parameter combination resulting in the largest likelihood function value is used as starting values in the optimization process.

Although this article is not concerned with inferential techniques for variable selection, we realize that this is an important aspect of modeling in the current context, especially in the case where additional explanatory variables are included in the model. For the sake of completeness, we include an algorithm for the construction of confidence intervals for the model parameters (which can be used for model selection) in Appendix B.

3. Application

In this section, we discuss the operational data available. We then proceed to fit the various models discussed above and we consider the goodness-of-fit of each. Finally, interpretations of the results are

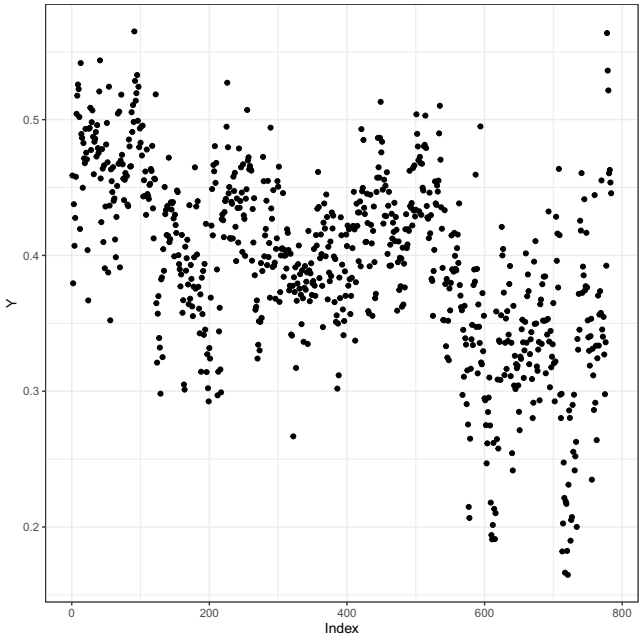

Figure 1. Production data, scaled between 0 and 1, plotted over time.

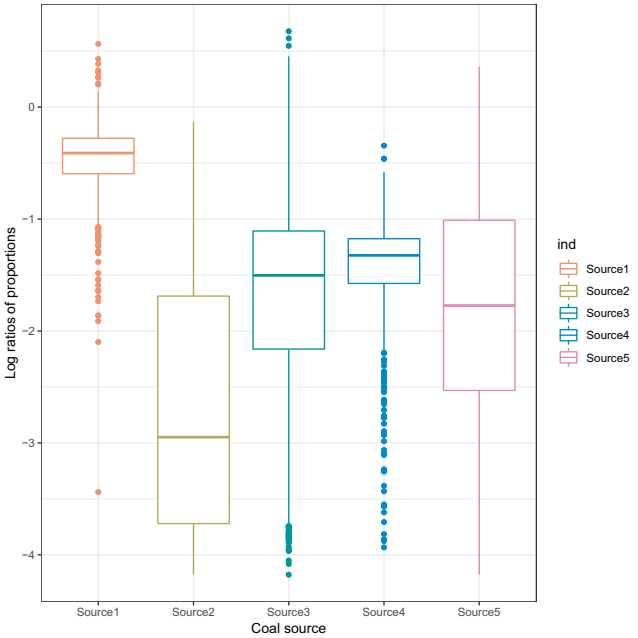

Figure 2. Boxplots of the transformed coal sources data.

provided together with comments on the practical implementation of the proposed models.

3.1. Operational data

We apply the CPH and AFT survival functions to model the raw gas production from a commercial chemical facility as a function of the blend from six coal sources. Note that the production data are scaled to between 0 and 1 due to confidentiality restrictions. Furthermore, the actual names of the coal sources cannot be mentioned. The coal quality of each source is different, and therefore quantifying and understanding the effect of the coal blend on the gas produced is critically important for continuous improvement and sustainable operation. The operational data are 12-h averages for just over one year, giving a total of $n = 784$ observations of production data as well as the corresponding coal sources blends. The production data are plotted in Figure 1. We will illustrate that the CPH and AFT models, incorporating the coal sources blends as predictors, provide an excellent fit to the data.

The proportions associated with the various coal sources constitute a compositional variable (also referred to as a mixture variable in the current context), the components of which sum to 1. Denote the proportion from source j by Z_j; the predictors are $(Z_1, \ldots, Z_6)$, with $\sum_{j=1}^{6} Z_j = 1$. Cornell (1981) provided a detailed treatment of how to specify regression models in the presence of compositional variables. However, in this article we use the additive log ratio transformation for the compositional variables (Aitchison 1986). That is, the compositional variables are transformed to log ratios relative to one specific component. We use coal source 6 as the denominator in the log ratio transform because it is the source with the highest average proportion in the blends. In order to avoid numerical difficulties, we replace all values in $Z_j, j = 1, \ldots, 6$ which are smaller than 0.01 by 0.01. The resulting transformed predictors are

$$X_j = \log(Z_j/Z_6), j = 1, \ldots, 5. \quad (3)$$

Using log ratios is advantageous since the effect of each variable on the output is interpreted relative to one compositional component, see Aitchison (1986). Note that it is not really necessary to use a log-transformation of the coal sources for specifying the model since Scheffé or K-polynomials may also be used for the mixture variables (Cornell 2002). However, using log-ratios is advantageous since the effect of each variable on the output is interpreted relative to a fixed compositional component (Aitchison 1986). In addition, the recommendation of improved coal sources blends is simplified using log-ratios, as will be illustrated below. Figure 2 shows boxplots of the empirical distributions of the different sources on the transformed scale. It is clear that there are substantial differences in both the averages and ranges of the proportions associated with the various coal sources.

3.2. Model fitting and validation

The models discussed in Section 2 are fitted to the production data, using the proportions associated with

Table 1. Log-likelihood results of models.

	CPH	AFT
Weibull	1064.226	1064.819
EV	960.594	1017.674
Log-normal	784.718	1033.740

the various coal sources as predictors. In order to compare the fit of the various models, we use the calculated log-likelihood functions (together with the estimated parameters). Since all of the models considered contain the same number of parameters, no correction terms, such as those employed in the Akaike information criterion are required in order to compare the levels of fit of the various models. The achieved log-likelihood function values are reported for each of the fitted models in Table 1. The Weibull baseline distribution yielded the highest log-likelihood for both the CPH and AFT models. Furthermore, both the log-likelihood values reported for the CPH and AFT models with this baseline distribution are approximately equal in this case, confirming that the CPH and AFT models are equivalent when using the Weibull distribution and that one can be obtained from the other. The small observed differences can be ascribed to the difference in the numerical optimizers used.

A comparison of the calculated values of the log-likelihoods indicates that the models with the Weibull baseline provide the best fit of the models considered. Additionally, the light tails of the Weibull distribution, compared to the lognormal and extreme value distributions, makes for a conservative model for gas production. In the context of predicting gas production, we believe that under-predicting gas production would likely be a preferable outcome compared to over-predicting. This provides a second motivation for choosing the Weibull distribution as baseline. A third reason supporting this choice is the high level of mathematical tractibility which this choice affords the modeler; see the simple closed form expressions provided in Section 2.3.1.

Following the discussion above, we select the models with the Weibull baseline (we specifically use the fitted CPH model to arrive at the numerical results below). In order to assess the fit of this model, we consider a graphical test before turning our attention to a formal goodness-of-fit test. Figure 3 illustrates the fitted Weibull baseline distribution function, obtained using the averaged log-ratios as predictors, as well as the empirical distribution function of the data. The figure indicates that the proposed model fits the observed data well. This claim is made more precise below using formal techniques.

In order to proceed with more formal techniques, we introduce the composite goodness-of-fit hypothesis that

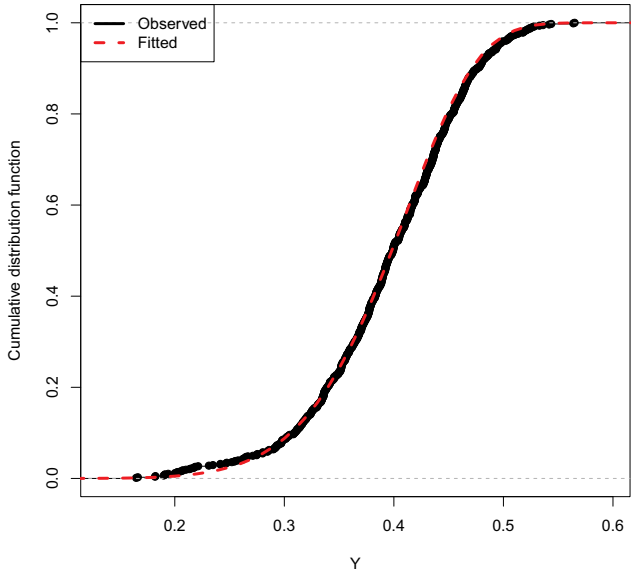

Figure 3. Empirical and fitted distribution functions.

we wish to test; given the predictors $\mathbf{X}$, Y has the distribution specified by $\tilde{F}_{\theta,\phi}$. This hypothesis is to be tested against general alternatives. An application of the probability integral transform specifies that, if the null hypothesis is true, then $\tilde{S}_{\theta,\phi}(Y|\mathbf{X})$ follows the standard uniform distribution. In order to obtain a graphical goodness-of-fit test, we compare the quantiles of

$$\hat{U}_j = \tilde{S}_{\hat{\theta},\hat{\phi}}(Y_j|\mathbf{X}_j), j = 1, \ldots, n,$$

to those of the standard uniform distribution. From the consistency of the maximum likelihood estimators used and the fact that the sample under consideration is relatively large, we infer that $\hat{U}_1, \ldots, \hat{U}_n$ should approximately follow a standard uniform distribution if the null hypothesis holds. Figure 4 shows a quantile-quantile plot for $\hat{U}_1, \ldots, \hat{U}_n$ against the uniform quantiles (the line of perfect fit is superimposed as a dashed line). The figure indicates that the empirical quantiles correspond closely to those of the standard uniform distribution. Furthermore, the mean and standard deviation of $\hat{U}_1, \ldots, \hat{U}_n$ are calculated to be 0.502 and 0.277, compared to the corresponding 0.5 and $1/\sqrt{12} \approx 0.289$ of the uniform distribution. Taking all of the above into account, the graphical goodness-of-fit test indicates that the CPH model with Weibull baseline fits the data well (and the same conclusion holds for the AFT model, since these models are the same).

We now turn our attention to a formal goodness-of-fit test. Again, our test will be based on the idea that the realized values of $\hat{U}_1, \ldots, \hat{U}_n$ should approximately follow a uniform distribution under the null

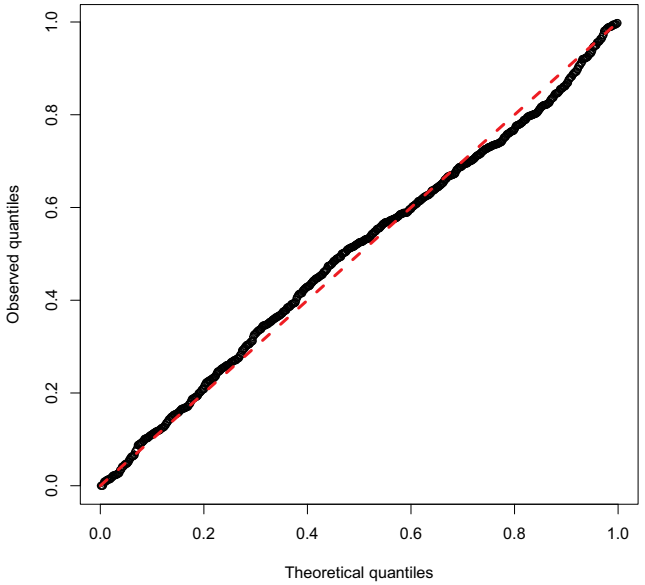

Figure 4. Comparison of the quantiles of $\hat{U}_1, \ldots, \hat{U}_n$ and the quantiles of the standard uniform distribution.

hypothesis. As above, we present only the results for the CPH model with Weibull baseline, since the results for the AFT model are nearly identical.

We calculate the Kolmogorov-Smirnov test statistic, measuring the supremum distance between the empirical distribution function of $\hat{U}_1, \ldots, \hat{U}_n$ and the distribution function of the standard uniform distribution;

$$KS_n = \sup_{0 \le u \le 1} \{|F_n(u) - u|\}, \tag{4}$$

where $F_n(u) = \frac{1}{n}\sum_{j=1}^{n} \mathrm{I}(\hat{U}_j < u)$ with $I(\cdot)$ the indicator function. The test statistic in (4) can be expressed in the following easy calculable form:

$$\begin{aligned} KS_n &= \max(\Delta_n^+, \Delta_n^-), \\ \Delta_n^+ &= \max_{j \in \{1,\ldots,n\}} \left(\frac{j}{n} - \tilde{F}_{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}}(Y_j | \mathbf{X}_j) \right), \\ \Delta_n^- &= \max_{j \in \{1,\ldots,n\}} \left(\tilde{F}_{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}}(Y_j | \mathbf{X}_j) - \frac{j-1}{n} \right). \end{aligned}$$

In order to test the goodness-of-fit hypothesis, we are required to approximate the null distribution of KS_n. To this end, we simulate a single observation from the fitted conditional distribution of Y given $\mathbf{X}_j$, for each of $j = 1, \ldots, n$. Denote the resulting values by $Y_1^*, \ldots, Y_n^*$. Based on these data, fit the model once more and calculate the values of $\hat{U}_1^*, \ldots, \hat{U}_n^*$ as well as the corresponding value of KS_n^*. Repeating this process a large number of times, we simulate from the null distribution of KS_n. The algorithm required for this procedure is included in more detail in Appendix A. For the CPH model with Weibull baseline, the *p*value is calculated to be approximately 0.035 and we do not reject the hypothesis that the CPH model with Weibull baseline is appropriate at a significance level of 1%. It should be noted that, given the large sample, using a significance level of 1% is quite conservative as a small deviation from the null model is likely to result in the rejection of the goodness-of-fit hypothesis.

3.3. Interpretation and implementation

In this section we consider the fitted CPH and AFT models with the Weibull distribution as baseline. Recall that the CPH and AFT models are equivalent when used with this baseline distribution. As a result, the predicted values associated with these models are nearly identical and can be ascribed to small differences due to variations in the optimization procedures used. We report the estimated parameters obtained for both models and we comment on both sets of estimated parameters. However, in an attempt to avoid repetition, we arbitrarily use the CPH model when constructing graphs and when predicting outcomes.

In the remainder of this section we:

- discuss the interpretation of the regression parameters and the impact of the variables on the level of production,
- illustrate how the estimated conditional cumulative distribution function is used to show the impact of different blends on the level of production, and for recommending blends for process improvement.

Consider the CPH model with Weibull baseline. Using maximum likelihood, the estimated parameters of the Weibull distribution are $\hat{k} = 7.147$ and $\hat{\lambda} = 0.392$ for the shape and scale parameters respectively, and the estimated regression parameters are

$$\hat{\boldsymbol{\beta}} = (0.394, 0.090, -0.141, -0.136, 0.268),$$

based on the log-ratios of the proportions of the coal sources relative to the last component (source 6). For the CPH model, the exponentiated coefficients can be interpreted as multiplicative effects on the hazard. The exponentiated coefficients are:

$$e^{\hat{\boldsymbol{\beta}}} = (1.483, 1.094, 0.869, 0.873, 1.307).$$

Therefore, for example, increasing the proportion of coal source 2 relative to coal source 6 in the blend increases the hazard rate by a factor of approximately 1.09 or 9%, which is substantial given the large volumes of gas produced (Coetzer, Rossouw, and Lin 2008). Similarly, increasing the proportion of coal source 3 relative to coal source 6 in the blend reduces

the hazard rate by a factor of approximately 13%. Therefore, the composition of the blend has an important effect on the production of the factory. This indicates that the CPH model can be used to perform diagnostic analysis on the coal sources for understanding the reasons of an increase or decrease in production. The interpretations afforded by the coefficients provide the modeler with valuable information regarding the effect of the coal source used. This information can be used in the decision making process when preparing blends for optimal production. More specifically, improved coal blends can be specified for higher production at sustainable rates. This will reduce the company's dependence on coal, and should improve environmental compliance.

When turning our attention to the AFT model, the fitted baseline distribution is $Weibull(7.127, 0.393)$ while the estimated regression coefficients are

$$\hat{\boldsymbol{\phi}} = (-0.055, -0.013, 0.02, 0.02, -0.037).$$

For interpreting the coefficients above, it is simplest to consider the effect of a given coefficient on the survival function. The effect of a unit increase in the log-ratio of source $j = 1, \ldots, 5$, relative to source 6 results in an estimated acceleration of $e^{-\phi_j}$ on the scale on which the survival function is measured. As a result, we are again interested in the exponentiated coefficients; however, this time around, we consider

$$e^{-\hat{\phi}} = (1.057, 1.013, 0.98, 0.98, 1.038).$$

A unit increase in the log-ratio of source 1 relative to source 6, for example, results in an acceleration of time (as measured by the fitted survival function) of 5.7%. This means that an increase in the proportion of coal source 1 is likely to substantially reduce production as compared to the source 6. Again, the fitted regression coefficients provide the modeler with directly interpretable results, which may influence the choice of coal source as well as the production achieved.

To illustrate the effect of the blends on production, we selected two blends from historical data which yielded a low and a high production, respectively. We selected actual blends from historical data to ensure that the blends are realistic. The two blends are referred to as a low production scenario and a high production scenario, respectively. The proportions of the two blends from the 6 sources are:

$$\text{Low scenario} : (0.163, 0.047, 0.010, 0.117, 0.151, 0.513),$$
$$\text{High scenario} : (0.241, 0.010, 0.010, 0.174, 0.002, 0.563).$$

Note that the proportion of coal source 6 in the blend is increased by about 6% and the proportion of coal source 2 in the blend is reduced by about 4.8% from the low to the high production scenario, which is in agreement with the effect of the variables corresponding to the estimated coefficients. The log-ratios used in the models are

$$\text{Low scenario} : (-1.149, -2.381, -3.947, -1.479, -1.225),$$
$$\text{High scenario} : (-0.847, -4.050, -4.050, -1.172, -5.812).$$

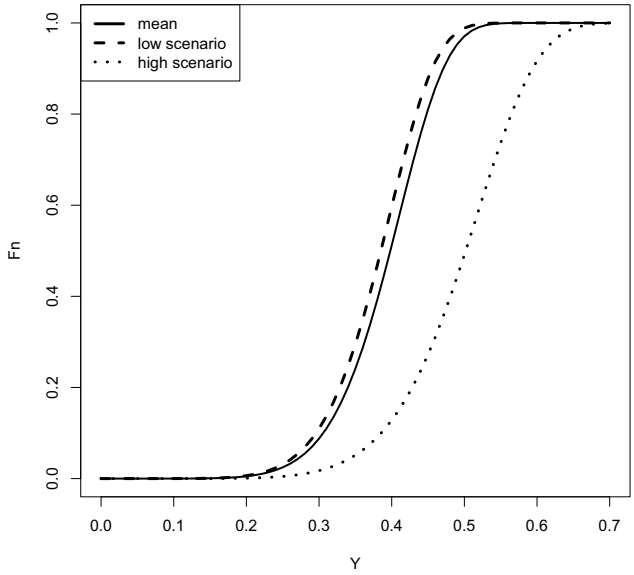

Figure 5. Fitted distribution functions for the mean coal sources, the low scenario and the high scenario.

The fitted cumulative distribution functions can now be calculated for the two scenarios to illustrate the effect of the blends on the production. Figure 5 shows the fitted cumulative distribution functions for the two scenarios, together with the fitted distribution achieved when averaging over the log ratios. Figure 5 shows that the probability of obtaining greater production is appreciably higher for the high scenario compared to the mean and low scenarios. The estimated cumulative distribution function from the fitted CPH model can be used to determine the predicted probability of obtaining at least a targeted production for any specified coal sources blend.

Predicted quantiles can be obtained for a specified probability; inverting the survival function, the $(1-\delta)\%$ quantile of the fitted distribution is

$$q(1-\delta) = \hat{\lambda}\left(-\log\left(e^{\log(1-\delta)e^{-\hat{\boldsymbol{\beta}}^{\top}\mathbf{x}}}\right)\right)^{1/\hat{k}}. \qquad (5)$$

Specifically, if we are interested in the quantile such that there is a 5% probability of exceedence of production; i.e., a 95% chance of observing a level of production less than or equal to this quantile. This is

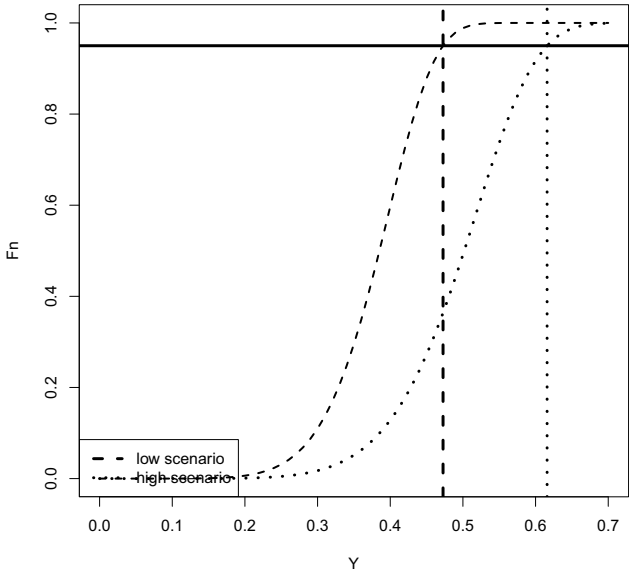

Figure 6. Fitted distribution functions for the low and high scenarios together with the corresponding quantiles at 95%.

obtained by setting $\delta = 0.95$ in (5). The predicted quantiles for the two blend scenarios are:

- Low scenario: 0.47,
- High scenario: 0.62.

The difference in the quantiles for the two scenarios is illustrated in Figure 6. For the high production scenario, a production of at most 0.62 (scaled units) can be expected, compared to a production of only 0.42 (scaled units) for the low production scenario. This is an improvement in production of roughly 32% for the high blend scenario compared to the low blend scenario. Given that the plant delivers approximately 29% of the fuel demand in South Africa, there is a strong incentive for identifying and specifying improved coal sources blends. Furthermore, given sufficient supply of coal from the different coal sources, the CPH and AFT models can be used to specify optimum coal sources blends subject to constraints such as the available coal supply, coal quality requirements of the blends and factory demand.

4. Conclusions and directions for future research

The results discussed in Section 3 show that gas production can be accurately modeled using parametric CPH and AFT models using the Weibull as the baseline distribution. These models were developed in the survival analysis literature and, to date, we have been unable to find other applications of these models in the context considered in this article.

At present, gas production is typically modeled without considering the presence of heterogeneity due to the coal sources blend used. We demonstrated that the CPH and AFT models can substantially increase the accuracy with which this process is modeled by taking the dependence on the blend used into account. Specifically, these models allow us to make predictions regarding the achievable levels of production given the source of the coal. The potential value added to the business is substantial in terms of volumes produced at sustainable rates, as well as savings on coal and improvements in carbon efficiency.

The current study is concerned with a single set of compositional variables as predictors. However, in practice, coal gasification is a very complex process. Process performance can be affected by many other operational variables and extraneous factors, such as utility consumption, feed availability, unplanned shutdowns and downstream upsets. Furthermore, in addition to coal sources blends, variables such as the quality of the coal sources, particle size distribution and feed availability have substantial effects on gas production. Therefore, future research is necessary in order to expand the models presented to those with additional predictor variables. The aim is to improve the prediction accuracy of the models in the hope of adding value to the business.

At present, we model only a single output. It is envisaged to extend the proposed models to the joint modeling of several output variables simultaneously, such as gas composition, gas outlet temperature and the variability thereof. Predictions from the CPH and AFT models can also be used to implement a statistical process monitoring methodology, which is another avenue for future research.

About the authors

Dr Roelof Coetzer obtained his PhD in mathematical Statistics in 2004. He worked in various industrial environments i.e., business, agricultural, physical and engineering sciences. He is currently Professor in the Faculty of Engineering, North-West University. His research interests include design of experiments, response surface methodology, statistical learning and multivariate process monitoring.

Dr Daan de Waal is an Extraordinary Professor at the University of the Free State and the University of Pretoria. He has extensive experience in teaching, research and consulting to industry. His research interests include Bayesian statistics, multivariate statistics, and probability distributions.

Dr Marius Smuts obtained his PhD in 2020 in Business Mathematics. His research interests include optimization methods, credit risk modeling, goodness-of-fit testing, and

survival analysis. He is an Extraordinary Senior Lecturer in the School of Mathematical and Statistical Sciences, North-West University.

Dr Jaco Izak Visagie obtained his PhD in Risk Analysis in 2015. His research interests include mathematical and applied statistics, finance, and hypothesis testing. He has been involved in many industry projects. He is currently Professor in Statistics in the School of Mathematical and Statistical Sciences, North-West University.

Acknowledgements

The authors would like to thank Sasol South Africa for providing the data and approval to publish this research. The work of RLJ Coetzer and IJH Visagie are based on research supported by the National Research Foundation (NRF). Any opinion, finding and conclusion or recommendation expressed in this material is that of the authors and the NRF does not accept any liability in this regard. The authors thank Professor James Allison for many insightful comments and discussions.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Aitchison, J. 1986. *The statistical analysis of compositional data*. London: Chapman and Hall.

Broström, G. 2012. *Event history analysis with R*. Boca Raton: Chapman and Hall.

Broström, G. 2020. *Event history analysis*. R package version 2.8.3, https://cran.r-project.org/package=eha.

Broyden, C. G. 1970. The convergence of a class of double-rank minimization algorithms. *Journal of the Institute of Mathematics and Its Applications* 6:76–90.

Ceylan, Z., and C. Ceylan. 2021. Applications of machine learning algorithms to predict the performance of coal gasification process. In *Applications of artificial intelligence in process systems engineering*, 165–86. Amsterdam: Elsevier.

Chavan, P. D., T. Sharma, B. Mall, B. Rajurkar, S. Tambe, B. Sharma, and B. Kulkarni. 2012. Development of data-driven models for fluidized-bed coal gasification process. *Fuel* 93:44–51. doi:10.1016/j.fuel.2011.11.039.

Chen, C., Y. Liu, S. Wang, X. Sun, C. Di Cairano-Gilfedder, S. Titmus, and A. A. Syntetos. 2020. Predictive maintenance using Cox proportional hazard deep learning. *Advanced Engineering Informatics* 44:101054. doi:10.1016/j.aei.2020.101054.

Coetzer, R. L. J., and M. J. Keyser. 2003. Experimental design and statistical evaluation of a full-scale gasification project. *Fuel Processing Technology* 80 (3):263–78. doi:10.1016/S0378-3820(02)00251-5.

Coetzer, R. L. J., and M. J. Keyser. 2004. Robustness studies on coal gasification process variables. *ORiON* 20 (2):89–108. doi:10.5784/20-2-9.

Coetzer, R. L. J., R. F. Rossouw, and D. K. J. Lin. 2008. Dual response surface optimization with hard-to-control variables for sustainable gasifier performance. *Journal of the Royal Statistical Society Series C: Applied Statistics* 57 (5):567–87. doi:10.1111/j.1467-9876.2008.00631.x.

Cornell, J. 1981. *Experiments with mixtures: Designs, models and the analysis of mixture data*. New York: Wiley.

Cornell, J. A. 2002. *Experiments with mixtures: Designs, models, and the analysis of mixture data*. 3rd ed. New York: John Wiley & Sons.

Cox, D. R. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 34 (2):187–202. doi:10.1111/j.2517-6161.1972.tb00899.x.

de Waal, D., R. Coetzer, and S. van der Merwe. 2016a. Classification of multiple Dirichlet observations under a multinomial model. *Chemometrics and Intelligent Laboratory Systems* 150:51–7. doi:10.1016/j.chemolab.2015.10.012.

de Waal, D., R. Coetzer, and S. van der Merwe. 2016b. A multivariate gamma distribution applied to compositional data. *South African Statistical Journal* 50 (02):273–83. doi:10.37920/sasj.2016.50.2.6.

Equeter, L., F. Ducobu, E. Rivière-Lorphèvre, R. Serra, and P. Dehombreux. 2020. An analytic approach to the cox proportional hazards model for estimating the lifespan of cutting tools. *Journal of Manufacturing and Materials Processing* 4 (1):27. doi:10.3390/jmmp4010027.

Fletcher, R. 1970. A new approach to variable metric algorithms. *The Computer Journal* 13 (3):317–22. doi:10.1093/comjnl/13.3.317.

Goldfarb, D. 1970. A family of variable metric updates derived by variational means. *Mathematics of Computation* 24 (109):23–6. doi:10.1090/S0025-5718-1970-0258249-6.

Ma, Z., and A. W. Krings. 2008. Survival analysis approach to reliability, survivability and prognostics and health management (PHM). In *2008 IEEE Aerospace Conference*, 1–20. IEEE. doi:10.1109/AERO.2008.4526634.

Majeed, A. 2020. Accelerated failure time models: An application in insurance attrition. *The Journal of Risk Management and Insurance, Bangkok, Thailand: The University* 2020:1–18.

Qin, S. J. 2014. Process data analytics in the era of big data. *AIChE Journal* 60 (9):3092–100. doi:10.1002/aic.14523.

R Core Team. 2019. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rossouw, R. F., R. L. J. Coetzer, and N. J. Le Roux. 2018. Simulation of a coal stacking process using an online X-ray fluorescence analyser. *ORiON* 34 (1):65–81. doi:10.5784/34-1-575.

Shanno, D. F. 1970. Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation* 24 (111):647–56. doi:10.1090/S0025-5718-1970-0274029-X.

Smuts, M., and J. S. Allison. 2020. An overview of survival analysis with an application in the credit risk environment. *ORiON* 36 (2):89–110.

Thijssens, O. W. M., and W. J. C. Verhagen. 2020. Application of extended Cox regression model to time-on-wing data of aircraft repairables. *Reliability Engineering & System Safety* 204:107136. doi:10.1016/j.ress.2020.107136.

Tolley, H. D., J. M. Barnes, and M. D. Freeman. 2016. Survival analysis. In *Forensic epidemiology*, 261–84. Amsterdam: Academic Press.

Visagie, I. J. H. 2018. On parameter estimation in multi-parameter distributions. *Statistics, Optimization & Information Computing* 6 (3):452–67. doi:10.19139/soic.v6i3.583.

Wei, L.-J. 1992. The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine* 11 (14-15):1871–9. doi:10.1002/sim.4780111409.

Appendix A. Bootstrap algorithm

The algorithm below can be used to approximate the critical value of the Kolmogorov-Smirnov test statistic used in Section 3.2.

1. Estimate the parameters of the model based on $(Y_1, ..., Y_n)$. Denote the resulting vector of estimates by $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\beta}})$.
2. For each value of $\boldsymbol{X}_j$ and using $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\beta}})$, calculate the conditional distribution function $\tilde{F}_j$.
3. From every $\tilde{F}_j$, simulate a single observation. Denote the result by Y_j^*.
4. Estimate the parameters of the model based on $(Y_1^*, ..., Y_n^*)$. Denote the resulting vector of estimates by $(\hat{\boldsymbol{\phi}}^*, \hat{\boldsymbol{\beta}}^*)$.
5. Calculate $\hat{U}_j^* = \tilde{S}_{\hat{\boldsymbol{\phi}}^*, \hat{\boldsymbol{\beta}}^*}(Y_j^*), j = 1, ..., n$.
6. Based on $\hat{U}_j^*, j = 1, ..., n$, calculate the Kolmogorov-Smirnov test statistic; KS^*.
7. Repeat steps 3 to 5 B times.
8. Denote the order statistic of the calculated test statistics by $KS_{(1)}, ..., KS_{(B)}$. The empirical critical value at significance level α is $KS_{(\gamma)}$, where $\gamma = (\lfloor B(1-\alpha) \rfloor)$.

Appendix B. Algorithm for the construction of confidence intervals

In this appendix, we provide an algorithm which can be used for the construction of bootstrap confidence intervals for the parameters in the model. This is of interest since, if the confidence interval for a specific regression coefficient includes 0, then the associated explanatory variable may be removed from the model. Below we use the notation that was introduced in the body of the article. The required algorithm is as follows.

1. Based on the responses, $\mathbf{Y}$, and predictors, $\mathbf{X}$, fit the model using maximum likelihood. That is, calculate
$$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}) = \arg\max\ L(\boldsymbol{\theta}, \boldsymbol{\beta} | (Y_1, \mathbf{X_1}), ..., (Y_n, \mathbf{X_n})).$$
2. Generate n independent standard uniform random variables. Denote the resulting random variates by U_j^*, for $j \in \{1, ..., n\}$.
3. For $j \in \{1, ..., n\}$, calculate the fitted distribution function:
$$\tilde{F}_{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}}(y | \mathbf{X}_j) = 1 - \tilde{S}_{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}}(y | \mathbf{X}_j).$$
4. For $j \in \{1, ..., n\}$, calculate of the fitted quantile function:
$$\tilde{F}^{-1}_{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}}(y | \mathbf{X}_j).$$
5. For $j \in \{1, ..., n\}$, calculate
$$Y_j^* = \tilde{F}^{-1}_{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}}(U_j | \mathbf{X}_j).$$
6. Based on the simulated responses, $(Y_1^*, ..., Y_n^*)^\top$, and observed predictors, $\mathbf{X}$, fit the model using maximum likelihood. That is, calculate
$$(\hat{\boldsymbol{\theta}}^*, \hat{\boldsymbol{\beta}}^*) = \arg\max\ L(\boldsymbol{\theta}, \boldsymbol{\beta} | (Y_1^*, \mathbf{X_1}), ..., (Y_n^*, \mathbf{X_n})).$$
7. Repeat steps (2) to (6) B times and denote the j^{th} realization of β_k^* by $\beta_{k,j}^*$.

Let $\beta^*_{k,(j)}$ denote the j^{th} order statistic of $\beta^*_{k,j}$. A $100(1-\alpha)\%$ confidence interval for β_j is given by
$$\left[\beta^*_{k,(l)}, \beta^*_{k,(u)}\right],$$
with $l = \lfloor B\alpha/2 \rfloor$ and $u = \lfloor B(1-\alpha/2) \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor function.