ELECTRICAL SUBMERSIBLE PUMP SURVIVAL ANALYSIS

MICHELLE PFLUEGER

Petroleum Engineer, Chevron Corp. & Masters Degree Candidate

Advisor Dr. Jianhua Huang

With help from PHD Candidate Sophia Chen

Department of Statistics, Texas A&M, College Station

MARCH 2011

ABSTRACT

A common metric in Petroleum Engineering is “Mean Time Between Failures” or “Average Run Life”. It is used to characterize wells and artificial lift types, as a metric to compare production conditions, as well as a measure of the performance of a given surveillance & monitoring program. Although survival curve analysis has been in existence for many years, the more rigorous analyses are relatively new in the area of Petroleum Engineering. This paper describes the basic theory behind survival analysis and the application of those techniques to the particular problem of Electrical Submersible Pump (ESP) Run Life. In addition to the general application of these techniques to an ESP data set, this paper also attempts to answer: Is there a significant difference between the survival curves of an ESP system with and without emulsion present in the well? Of the variables collected, which variables best describe the survival function? Do the variables collected in the dataset capture the variation in the survival function?

TABLE OF CONTENTS

Survival Analysis in Petroleum Engineering

A common metric in Petroleum Engineering is "Mean Time Between Failure" or "Average Run Life". It is used to characterize the average "life span" of wells and artificial lift types, as a metric to compare production conditions (which ones give better run life), as well as a measure of the performance of a given surveillance and monitoring program (improved surveillance and monitoring should extend run life). Although survival curve analysis has been in existence for many years, the more rigorous analyses are relatively new in the area of Petroleum Engineering. As an example of the growth of these analysis techniques in the petroleum industry, Electrical Submersible Pump (ESP) survival analysis has been sparsely documented in technical journals for the last 20 years:

- First papers on the fitting of Weibull & Exponential curves to ESP run life data in 1990 (Upchurch) & 1993 (Patterson)
- Papers discussing the inclusion of censored data in 1996 (Brookbank) & 1999 (Sawaryn)
- Paper discussing the use of Cox Regression in 2005 (Bailey)

Unfortunately, the papers applying these techniques did little to transfer the knowledge to the practicing Petroleum Engineers. They shared the technical concepts and equations, but not the practical knowledge of how to apply them to real life problems or why these analyses improved upon the "take the average of the run life of failed wells" technique most commonly used.

Theory of Survival Analysis

Survival analysis models the time it takes for events to occur and focuses on the distribution of the survival times. It can be used in many fields of study where survival time can indicate anything from time to death (medical studies) to time to equipment failure (reliability metrics). This paper will present three methodologies for estimating survival distributions as well as a technique for modeling the relationship between the survival distribution and one or more predictor variables (both covariates and factors). Appendix A has a list of important definitions relevant to survival analysis.

Kaplan Meier (Non-Parametric)

Non-parametric survival analysis characterizes survival functions without assuming an underlying distribution. The analysis is limited to reliability estimates for the failure times included in the data set (not prediction outside the range of data values) and comparison of survival curves one factor at a time (not multiple explanatory variables).

A common non-parametric analysis is Kaplan Meier (KM). KM is characterized by a decreasing step function with jumps at the observed event times. The size of the jump depends on the number of events at that time t and the number of survivors prior to time t. The KM estimator provides the ability to estimate survival functions for right censored data.

$$\hat{S}(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i}$$

t_i is the time at which a "death" occurs. d_i is the number of deaths that occur at time t_i. When there is no censoring, n_i is the number of survivors just prior to time t_i. With censoring, n_i is the number of survivors minus the number of censored units. The resulting curve, as noted, is a decreasing step function with jumps at the times of "death" t_i. The MTBF is the area under the resulting curve; the P50 (median) time to failure is $\hat{S}$(t) $\leq$0.5.

Upper and lower confidence intervals can be calculated for the KM curve using statistical software. A back-of-the-envelope calculation for the confidence interval is the KM estimator +/- 2 standard deviations. Greenwood's formula can be used to estimate the variance for non-parametric data (Cran.R-project):

$$\widehat{Var}\left(\hat{S}(t)\right) = \hat{S}(t)^2 \sum_{ti<t} \frac{d_i}{n_i(n_i - d_i)}$$

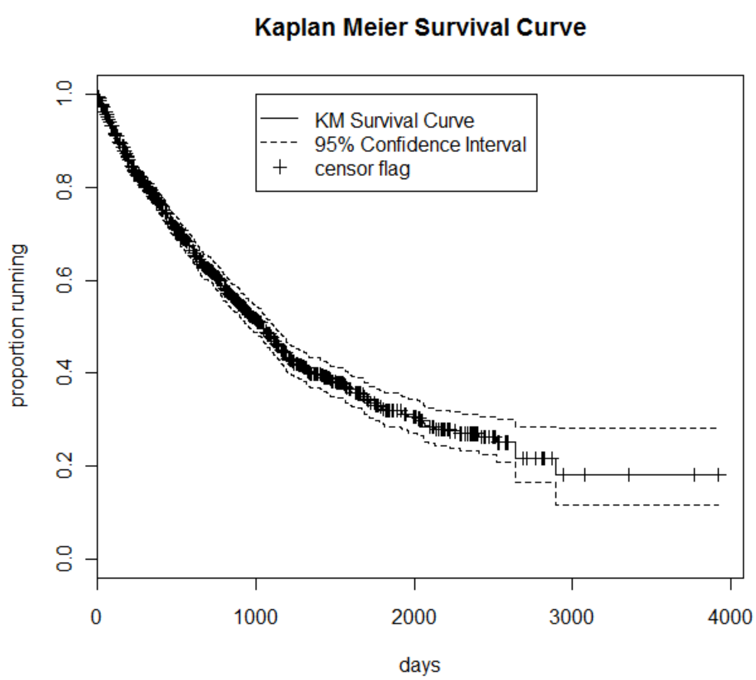


Figure 1: Example Kaplan Meier survival curve showing estimate, 95% confidence interval, and censored data points

When comparing two survival curves differing by a factor, a visual inspection of the null hypothesis Ho: survival curves are equal, can be conducted by plotting two survival curves and their confidence intervals. If the confidence intervals do not overlap, there is significant evidence that the survival curves are different (with alpha < 0.05%)

Cox Proportional Hazard (Semi-Parametric)

Semi-Parametric analysis enables more insight than the Non-Parametric method. It can estimate the survival curve from a set of data as well as account for right censoring, but it also conducts regression based on multiple factors/covariates as well a judge the contribution of a given factor/covariate to a survival curve. CPH is not as efficient as a parametric model (Weibull or Exponential), but the proportional hazards assumption is less restrictive than the parametric assumptions (Fox).

Instead of assuming a distribution, the proportional hazards model assumes that the failure rate (hazard rate) of a unit is the product of:

- a baseline failure rate (which doesn't need to be specified and is only a function of time)

$$h_0(t)$$

- and a positive function which incorporates the effects of factors & covariates $x_{i1} - x_{ik}$ (independent of time)

$$\exp\,(\beta_1 x_{i1} + \beta_2 x_{ik} + \cdots + \beta_k x_{ik})$$

This model is called semi-parametric because while the baseline hazard can take any form, the covariates enter the model linearly. Given two observations *i* & *i'* with the same baseline failure rate function, but that differ in their *x* values (ie two wells with different operating parameters x_k), the hazard ratio for these two observations are independent of time:

$$\frac{h_i(t)}{h_{i'}(t)} = \frac{h_0(t) * \exp\,(\beta_1 x_{i1} + \beta_2 x_{ik} + \cdots + \beta_k x_{ik})}{h_0(t) * \exp\,(\beta_1 x_{i'1} + \beta_2 x_{i'k} + \cdots + \beta_k x_{i'k})} = \frac{\exp\,(\beta_1 x_{i1} + \beta_2 x_{ik} + \cdots + \beta_k x_{ik})}{\exp\,(\beta_1 x_{i'1} + \beta_2 x_{i'k} + \cdots + \beta_k x_{i'k})}$$

The above ratio is why the Cox model is a proportional-hazards model; even though the baseline failure rate $h_0(t)$ is unspecified, the β parameters in the Cox model can still be estimated by the method of partial likelihood. After fitting the Cox model, it is possible to get an estimate of the baseline failure rate and survival function (Fox).

A result of the regression is an estimate for the various β coefficients and an R-square value describing the amount of variability explained in the hazard function by fitting this model. Relative contributions of factors/covariates can be interpreted as:

- $\beta > 0$, covariate decreases the survival time as value increases, by factor of $\exp(\beta)$
- $\beta < 0$, covariate increase the survival time as the value increases, by factor of $\exp(\beta)$
- For factors, the contribution is relative to the first level of the factor

As with any regression model, there are assumptions that need to be tested prior to CPH (Fox):

- Proportional Hazards Assumption – Conducted with a statistical test. Rejecting this assumption results in the need to stratify factors with time.
- Influential Observations – Conducted visually using index plots. These plots estimate the changes in the regression coefficients upon deleting each observation in turn divided

by the coefficient's standard error. Comparing the magnitudes of the largest values in the plot to the regression coefficients illustrates if any of the observations is influential.

- Non-Linearity – Conducted visually using residual plots. These plots illustrate if the factors are non-linear with time. If so, that factor would need to be treated as time series data or with a different model (quadratic, etc).

There are special considerations and treatments for:

- Time series data (factors not consistent over time, an example is the number and timing of ESP restarts prior to failure)
- Factors/covariates that do not meet the proportional hazards assumption

The treatment of time series data or data sets that do not meet the proportional hazards assumption are not discussed in this paper, however these methodologies are documented in Fox's paper "Cox Proportional-Hazards Regression for Survival Data".

WEIBULL (PARAMETRIC)

Parametric survival modeling assumes a distributional form. As previously noted, for ESP run life data, papers have been written on the use of both Weibull and Exponential distributions for the modeling of survival data. This paper will focus on the use of the Weibull distribution as it is the more widely used distribution in reliability and life/failure data analysis due to its versatility.

Parametric survival analysis is more efficient than Semi or Non-Parametric survival curve estimation but it has more restrictive assumptions. Parametric analysis is important in situations where extrapolation of results is necessary to enable prediction of failure under different conditions to those present in the original study. All other functionality provided by Parametric modeling is also available in the more flexible Semi-Parametric approach, but with lower efficiency (Fox).

TABLE OF WEIBULL DISTRIBUTION EQUATIONS:

WEIBULL EQUATION	FORMULA
PARAMETERS	λ>0 scale; k>0 shape
PDF	$f(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} exp-\left(\frac{t}{\lambda}\right)^{k} & x \geq 0 \\ 0 & x < 0 \end{cases}$
CDF	F(t;k,λ)=1- $exp-\left(\frac{t}{\lambda}\right)^{k}$
SURVIVAL	$exp-\left(\frac{t}{\lambda}\right)^{k}$
HAZARD	$h(t;k,\lambda) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1}$
WEIBULL REGRESSION (SIMILAR TO COX REGRESSION)	$h(t;k,\lambda) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} * \exp\,(\beta_1 x_{i1} + \beta_2 x_{ik} + \cdots + \beta_k x_{ik})$
MEDIAN	$\lambda(\ln(2))^{1/k}$

The Weibull shape parameter, *k*, is also known as the Weibull slope. Values of *k* less than 1 indicate that the failure rate is decreasing with time (infant failures). Values of *k* equal to 1 indicate a failure rate that does not vary over time (random failures). Values of *k* greater than 1 indicate that the failure rate is increasing with time (mechanical wear out) (Weibull).

A change in the scale parameter, *λ*, has the same effect on the distribution as a change of the *X* axis scale. Increasing the value of the scale parameter, while holding the shape parameter constant, has the effect of stretching out the PDF and survival curve (Weibull).

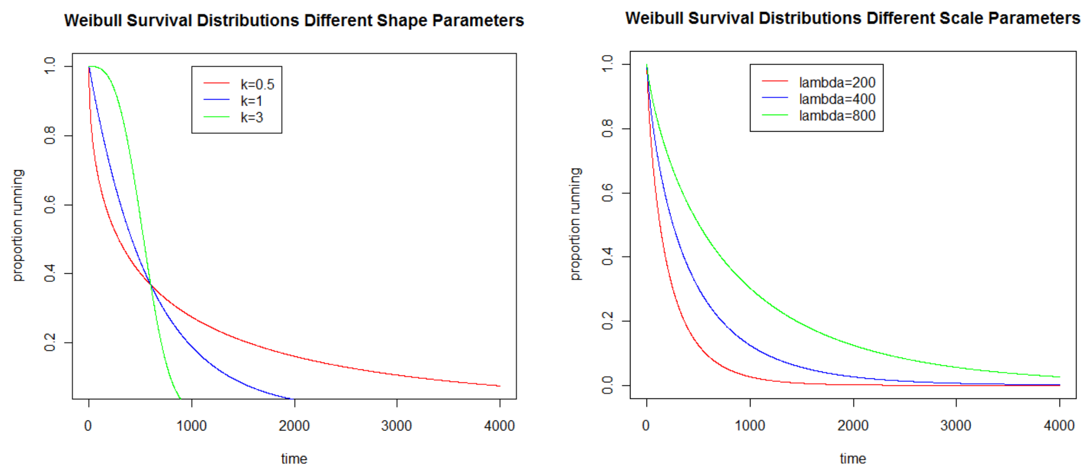


Figure 2: Example Weibull curves with varying shape & scale parameters

The Weibull regression model is the same as the Cox regression model with the Weibull distribution as the baseline hazard. The proportional hazards assumption used by the CPH method, when applied to a survival curve with a Weibull function baseline hazard, only holds if two survival curves vary by a difference in the scale parameter (*λ*) not by a difference in the shape parameter (*k*).

If goodness of fit to the Weibull distribution can be achieved, a confidence interval can be calculated for the curve, the median value and its confidence interval can be calculated, and a comparison of the differences in two survival curves can be conducted. Goodness of fit can be tested in R using an Anderson Darling calculation and verified with a Weibull probability plot. Poor fit in the tails of the Weibull distribution is a common occurrence for reliability data due to infant mortality and longer than expected wear out time.

Stepwise Cox & Weibull Regression

Given a large number of explanatory variables and the larger number of potential interactions, not all of those variables may be necessary to develop a model that characterizes the survival curve.   One way of determining a model is by using Stepwise model selection through minimization of AIC (Akaike Information Criterion). This model selection technique allows variables to enter/exit the model using their impact on the AIC calculated at that step. AIC is an improvement over maximizing the R-Square in that it's a criterion that rewards goodness of fit while penalizing for model complexity.

APPLICATION TO AN ESP DATA SET

As stated previously, these survival analysis techniques can be applied to many types of data in many industries ranging from survival data for people in a medical study to survival data for equipment in a reliability study. These methodologies have many uses in the petroleum industry; from surface equipment system and component reliability used by facility and reliability engineers, to well and downhole system and component reliability used by petroleum and production engineers.

As an example, this paper illustrates the use of these techniques on the run life of Electrical Submersible Pumps (ESP). ESPs are a type of artificial lift for bringing produced liquids to the surface from within a wellbore. Appendix B includes a diagram of an ESP. For this paper, the run life will refer to the run life of an ESP system, not the individual components within the ESP system. While this paper focuses on ESP systems, these same techniques could be applied to other areas of Petroleum Engineer interests including run life of individual ESP components, other types of artificial lift, entire well systems, etc.

DATA DESCRIPTION

ESP-RIFTS JIP (Electrical Submersible Pump Reliability Information and Failure Tracking System Joint Industry Project) is a group of 14 international oilfield operators who have joined efforts to gain a better understanding of circumstances that lead to a success or failure in a specific ESP application. The JIP includes access to a data set of 566 oil fields, 27861 wells, 89232 ESP installations, and 182 explanatory factors/covariates related to either the description of the ESP application or the description of the ESP failure.

For the analysis described in this paper, a subset of the data has been used, restricted to:

- Observations related to Chevron operated fields
- observations with no conflicting information (as defined by the JIP's data validation techniques)
- factors that were related to the description of the ESP application (excluded 27)
- factors not confounded with or multiples of other factors (excluded 30)
- factors with a large number (>90%) of non-missing data points (excluded 78)
- factors that were not free-form comment fields (excluded 27)

Appendix C has a list of the original 182 variables with comments on why they were removed from the analyzed data set, below is a table of the 20 remaining explanatory variables included in this analysis.

Summary Table of Data Included in the CPH/Regression Analysis:

Observations: 1588

Description	Covariate/Factor & # of Levels	Description
RunLife	Response	Time between date put on production and date stopped or censored
Censor	Censor Flag (0, 1)	1 if ESP failure 0 if still running or stopped for a different reason
Country	Factor (7 levels)	Country & Field in which the ESP is operated
Offshore	Factor (2 levels)	Indication of whether the ESP was an onshore or offshore installation
Oil	Covariate	Estimated average surface oil rate (m3/day)
Water	Covariate	Estimated average surface water rate (m3/day)
Gas	Covariate	Estimated average surface gas rate (1000m3/day)
Scale	Factor (5 levels)	Qualitative level of scaling present in the well
CO2	Covariate	% of CO2 present in the well
Emulsion	Factor (3 levels)	Qualitative level of emulsion present in the well
CtrlPanelType	Factor (2 levels)	Type of surface control panel used
NoPumpHouse	Covariate	Number of pump housings
PumpVendor	Factor (2levels)	Pump Vendor
NoPumpStage	Covariate	Number of pump stages
NoSealHouse	Covariate	Number of seal housings
NoMotorHouse	Covariate	Number of motor housings
MotorPowerRating	Covariate	Motor rated power at 60Hz (HP)
NoIntakes	Covariate	Number of intakes
NoCableSys	Covariate	Number of cable systems
CableSize	Covariate	Size of cable
DHMonitorInstalled	Factor (2 levels)	Flag for installation of a downhole monitor
DeployMethod	Factor (2 levels)	Method of ESP deployment into the well

Finding the P50 Time to Failure for a Dataset

Example 1: Using the entire data set, what is the P50 estimate for the runtime of a Chevron ESP? The answers differ considerably for the 4 calculation types:

Methodology	Includes Censored?	P50 Estimate (days)	Assumption	Assumptions Met?
Mean or Median	No	Mean: 563 Median: 439	None	N/A
Kaplan Meier Median	Yes	1044	None	N/A
CPH Median	Yes	1043	None (as no comparison of levels/covariates, essentially same results as KM)	N/A
Weibull Median	Yes	1067	Anderson Darling GOF for Weibull Distribution	NO (rejected the null hypothesis of good fit, due to poor fit in the tails)

In this example, the biggest impact on the difference between the methods is the inclusion of censored data. A large number of the ESPs in this data set have been running for >3000 days without a failure and were excluded in the often used calculation of the average run life of all failed ESPs. Given that the Weibull distribution did not pass the Anderson Darling goodness of fit test, the most appropriate calculation would have been the KM or CMH. Appendix E has the output from the various methodologies.

The interpretation of these results is that the P50 estimate of run life for an ESP installation in Chevron is ~ 1044 days. Additional, output from the KM analysis sets the 95% confidence interval at 952 to 1113 days.

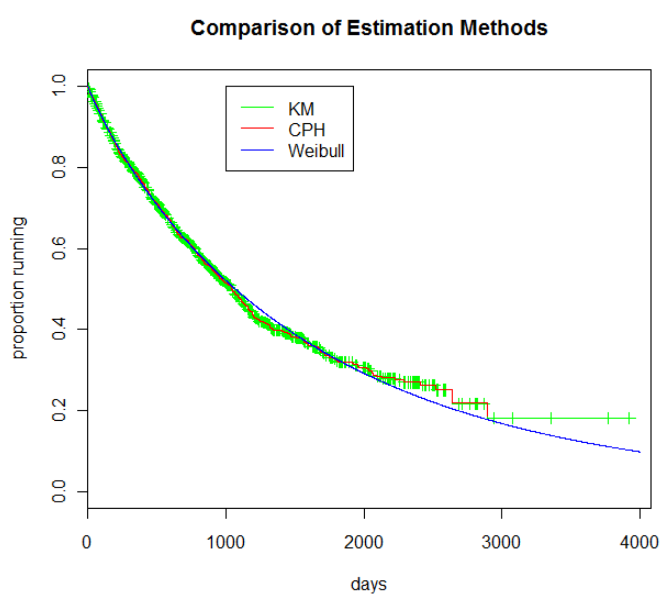


Figure 3: Comparison of estimation methods for full data survival curve. Note the deviation of the Weibull in the tails of the data.

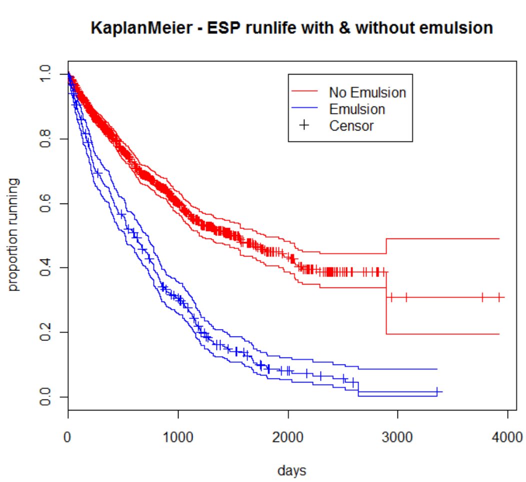
Comparing two survival curves differing by a factor

Example 2: Using the 2 level factor emulsion, does the presence of emulsion in the well make a significant difference in the P50 run life of an ESP system?

Methodology	Include Censor?	Emulsion P50 (Days)	No Emulsion P50 (days)	Significant Difference?	Interpretation	Assumptions Met?
Mean or Median	No	Mean 600 Median 458	Mean 536 Median 424	Don't know	Well performance is about the same	
Kaplan Meier Median	Yes	606	1508	Yes (visual Inspection of CI)	Wells without emulsion perform much better	
CPH Median	Yes	533	1408	Yes, with a Likelihood ratio test and a pvalue of 0, reject that B's are the same.	Wells without emulsion survive longer. Exp(B) indicates 2.5 times increased survival time for no emulsion.	No. (Reject null hypothesis of prop. hazards with a p value of 0.01.)
Weibull Median	Yes	531	1463	Yes, with a z test statistic and a pvalue of 0, reject that the scale values are the same.	Wells without emulsion survive longer. Scale parameter value indicates 2.75 times increased survival time for no emulsion.	No. (Reject null hypothesis of good of fit due to poor fit in the tails)

The more complex the methodology used, the more information is available to interpret the results. Again, the addition of censored data resulted in a very different interpretation of the data than just using the mean/median value of all failed ESPs; not just in the order of magnitude of the results, but also determination of which condition resulted in a longer run life. The results of both the CPH & Weibull methodologies are suspect due to their failure to meet the pre-requisite assumptions. Looking at the plots, it is apparent that the fit is poor in the tails. Appendix F has the output from the various methodologies

The interpretation of these results is that wells without emulsion have > a 2x increase in their P50 run life than wells with emulsion. It should be noted that given the other factors that differ in the operation of these ESPs, this difference may not be fully attributed only to the difference in emulsion, but this interpretation should lead to further investigation.



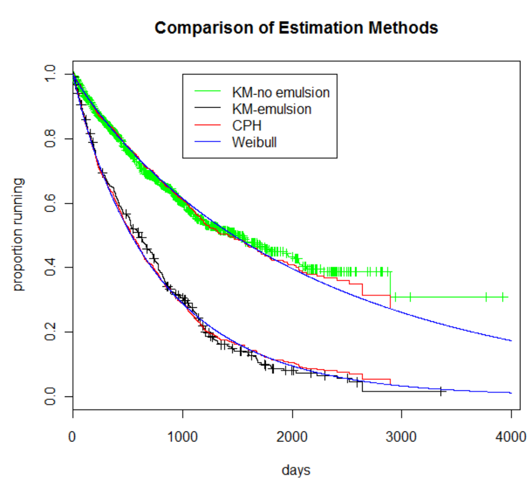


Figure 4: KM estimated survival curves for ESPs with and without emulsion with confidence interval
Figure 5: Comparison of estimation methods (KM, CPH, Weibull) for ESPs with and without emulsion

CHOOSING THE VARIABLES THAT CHARACTERIZE A SURVIVAL CURVE

Example 3: Of the variables collected by the JIP, which most describe the survival function? Do the variables collected in the dataset capture the variation in the survival function?

As stated previously, both Weibull & Cox regression fit a model using explanatory variables. The introduction of Stepwise variable selection to that regression allows the preferential fitting of the model by minimizing the AIC. As Weibull regression is a special case of Cox regression with a Weibull baseline hazard function, and as Cox regression has less restrictive assumptions than parametric regression, this example will focus solely on Cox regression using Stepwise variable selection.

Starting with an initial model consisting of all 20 explanatory variables and assuming no interactions, Stepwise variable selection resulted in the following model and interpretations: (Appendix G contains the output)

VARIABLE	FACTOR LEVEL COMPARED TO	COEFF	PVALUE	INTERPRETATION
Country Argentina_El Trapial	Country Angola_Banzala	2.208	0.0022	Angola is a significant improvement over Argentina
Country Thailand_Benchamas	Country Angola_Banzala	2.178	<0.0001	Angola is a significant improvement over Thailand
Country Thailand_Benchamas NW	Country Angola_Banzala	2.455	0.0034	Angola is a significant improvement over Thailand
Country United Kingdom _Captain	Country Angola_Banzala	-1.864	0.071	Given large SE of coefficient, UK is not significantly different than Angola
Country United States_Rangely	Country Angola_Banzala	1.917	0.13	Given large SE of coefficient, US is not significantly different than Angola
Country Venezuela_Boscan	Country Angola_Banzala	6.428	<0.0001	Angola is a significant improvement over Venezuela
Oil	Covariate	0.002	<0.0001	For a one unit (m3/day) increase in oil rate, there is a small decrease in survival time
Water	Covariate	-0.002	<0.0001	For a one unit (m3/day) increase in water rate, there is a small increase in survival time
Gas	Covariate	0.033	<0.0001	For a one unit (1000 m3/day) increase in gas rate, there is a small decrease in survival time
Scale Light	Scale None	0.305	0.62	Given large SE of coefficient, is not significantly different than no scale
Scale Moderate	Scale None	-4.587	<0.0001	Scale Moderate is a significant improvement over no scale
Scale Severe	Scale None	0.439	0.45	Given large SE of coefficient, is not significantly different than no scale
Scale Yes-Present	Scale None	0.717	0.23	Given large SE of coefficient, is not significantly different than no scale
CtrlPanelType Variable Speed Drive	CtrlPanelType Switchboard	-0.442	0.00034	Having a variable speed drive control panel is a significant improvement over a switchboard
NoPumpHouse	Covariate	0.259	0.00042	For a one unit increase in the number of pump houses, there is a decrease in survival time
PumpVendor REDA	PumpVendor Centrilift	-0.406	0.034	REDA is a significant improvement over Centrilift
NoPumpStage	Covariate	-0.002	0.001	For a one unit increase in the number of pump stages, there is an increase in survival time
NoCableSys	Covariate	0.965	<0.0001	For a one unit increase in the number of cable systems, there is a decrease in survival time
CableSize	Covariate	-0.202	<0.0001	For a one unit increase in the cable size, there is an increase in survival time

Writing out the final model:

$h_i(t) = h_0(t) * \exp$ (2.208 * Argentina_El Trapial + 2.178 * Thailand_Benchamas + 2.455 * Thailand_Benchamas NW + -1.864 * United Kingdom _Captain + 1.917 * United States_Rangely + 6.428 * Venezuela_Boscan + 0.001952 * Oil + -0.00159 * Water + 0.03317 * Gas + 0.3045 * ScaleLight + -4.587 * ScaleModerate + 0.4391 * ScaleSevere + 0.7174 * ScaleYes-Present + -0.4421 * CtrlPanelTypeVariableSpeedDrive + 0.2593 * NoPumpHouse + -0.4055 * PumpVendorREDA + -0.002249 * NoPumpStage + 0.9645 * NoCableSys + -0.202 * CableSize

To use the model, a 0 or 1 is used for the factor explanatory variables and the actual value for the covariates.

As stated previously, there are assumptions that need to be tested to use CPH or Cox regression. Appendix H has the output and plots necessary to test the following assumptions:

- Proportional Hazards Assumption – The global p-value for the chi-square test for Ho: proportional hazards assumption holds is 0.0628, therefore we fail to reject the null hypothesis. The proportional hazards assumption holds.
- Influential Observations – Conducted visually using an index plot. Comparing the magnitudes of the largest values in the plot to the regression coefficients shows there aren't any significant influencers other than in the levels of Scale. Reviewing the plots leads me to believe there are issues with the comparative level of Scale-None (possible data-miscoding). No action taken, however given those plots coupled with the fact that the response led to a non-intuitive interpretation, a recommendation to the JIP could be made that they put more stringent controls on scale data and re-evaluate.
- Non-Linearity – Conducted visually using a residual plot. Comparing the distribution of the residuals to a line with mean zero, there are a few points for both oil and gas that may indicate quadratic behavior in the upper ranges. If this model was going to be used rigorously, further investigation would need to go into those points.

The R-square value for this model is calculated to be 0.393. This means only 39.3% of the variation was captured by this model and that there are potentially other explanatory variables that would need to be measured to more fully explain the variation.

As a Petroleum Engineering practitioner, there are a few things that surprised me about these results:

- Downhole monitor flag was removed from the model: This is a technology that is believed to extend the runlife of ESPs because a quicker reaction to downhole conditions can be achieved.
- Emulsion was removed from the model: Given what we saw previously that no emulsion was a significant improvement on run life, the fact that it does not show up in the final model indicates that its effect can be explained using other factors in the model.

- Scale Moderate improved run life and the others were not significantly different from no scale: Scale in a wellbore fouls an ESP system and is thought to be one of the key reasons an ESP system might fail. As noted after looking at the influential residual plots, there may be underlying issues with the data collection for the Scale data.
- No support for difficulty operating in less developed countries: ESPs in Angola did not have a significantly different run life than ESPs in the US & UK, and were significantly better than Thailand & Venezuela. It should be noted that country is confounded by a significant amount of reservoir characterization data so it cannot be determined with this data set if differences between countries is due to the skill of the workforce or the difficulties associated with the reservoir.
- Only 39.3% of the variation in the survival curve was characterized by this model: As previously stated, of the 182 variables tracked by the JIP, only 20 could be used due to confounding and missing data. A more concerted study should be kicked off to determine if there is other information that should be collected and/or why the participating companies do not record all 182 variables.

Knowing that this is an observational data set and not a controlled experiment, it is difficult to draw more conclusions from the data than already stated. However, these conclusions alone should be enough to spur significant discussion within the JIP and additional data capture to understand some of the surprising results.

CONCLUSIONS

Understanding and applying different levels of survival analysis methodologies to petroleum industry run life data could lead to new insights and interpretations of old data sets. Historical use of straight means and/or medians of failed systems and interpreting them as P50 run life can be very misleading when there is a significant amount of censored data that is not taken into account. Although the use of more complex methodologies, such as Weibull regression and CPH or Cox regression, brings more information to the analysis of run life data, even the expansion of the usage of the Kaplan Meier technique would improve upon the current practice of calculating MTTF and could be managed in an existing toolset (Microsoft Excel). The use of R-square alone as a goodness of fit measure may not point out lack of fit in the tails. JIPs and other groups investigating explanatory effects on run life could benefit from these tools by using them not only to analyze their data, but also to better target the range of explanatory variables they capture to make their data collection efforts more effective.

APPENDICES

APPENDIX A: IMPORTANT DEFINITIONS

SURVIVAL FUNCTION: The Survival function is the probability that an event occurs at or after time *t*. It is also 1-Cumulative Distribution Function (CDF) for a given distribution. For reliability data, this is often modeled after an Exponential or Weibull distribution.

$$R(t) = P(\{T > t\}) = \int_t^{\infty} f(u)du = 1 - F(t)$$

HAZARD FUNCTION/FAILURE RATE: The Hazard function is also known as the failure rate or frequency with which a system or component fails. It is also the Probability Distribution Function (PDF)/Survival Function.

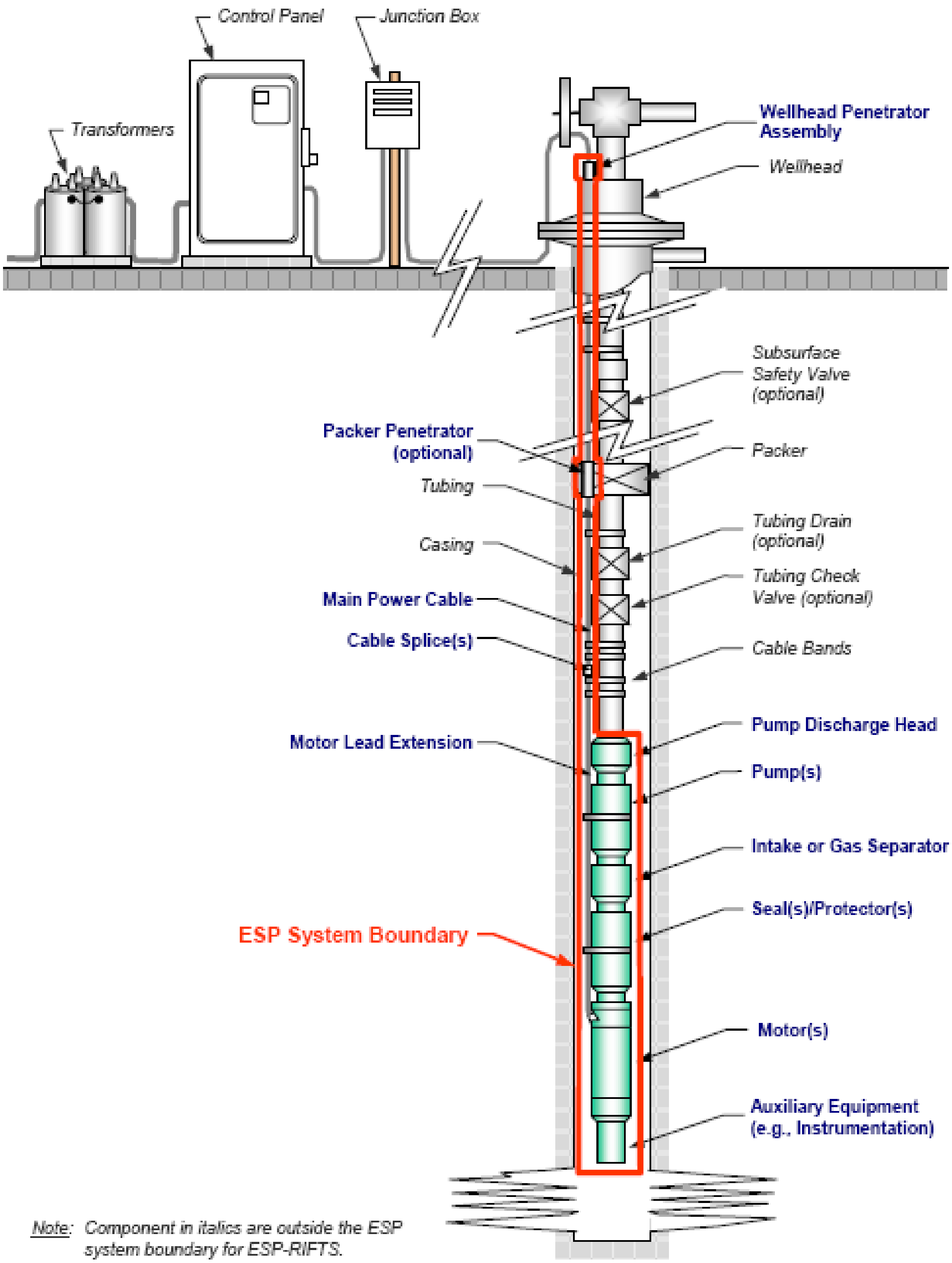
$$h(t) = \frac{f(t)}{R(t)}$$

MTBF (MEAN TIME BETWEEN FAILURE) OR MTTF (MEAN TIME TO FAILURE): Although it contains the word “Mean”, that is not always how it is used in practice. The MTBF can be a curve calculated as 1/failure rate. It can also be the mean value of that curve. In Petroleum Engineering, this same term is often used to indicate the mean or P50 survival time (often used interchangeably) for a system or component and is often calculated as the estimated average run-life based on a population of failed systems or components.

RIGHT CENSORED DATA: Units (people, systems, etc) that have not “died” before the end of a study or at the time of analysis. Excluding these data points will change the estimated MTBF. Note that there also exists left and random censored data.

GOODNESS OF FIT: A test for determining how well a given curve can be modeled by a distribution. In Petroleum Engineering, this is often conducted using R-square values (the closer to 1, the better the fit). In statistics, there are other goodness of fit measures including Anderson Darling (used in this paper) which may give a more powerful test of fit.

APPENDIX B: ESP SCHEMATIC
(ESPRIFTS)



Note: Component in italics are outside the ESP system boundary for ESP-RIFTS.

Appendix C: Data Description Summary

Variables that characterized ESP failure (Removed):

Reason For Pull: General
Reason For Pull: Specific
Primary Failed Item
Primary Failed Item - Major Component
Primary Failure Descriptor
Secondary Failure Descriptor
Primary Contaminant
Failure Cause: General
Failure Cause: Specific
Failure Comments
Pump Pull Condition
Pump Primary Failure Descriptor
Seal Pull Condition
Seal Primary Failure Descriptor
Motor Pull Condition
Motor Primary Failure Descriptor
Pump Intake Pull Condition
Pump Intake Primary Failure Descriptor
Cable Pull Condition
Cable Primary Failure Descriptor
MLE Pull Condition
Cable Splice Pull Condition
Pothead Pull Condition
Wellhead Penetrator Pull Condition
Packer Penetrator Pull Condition
DH Monitoring System Pull Condition
DH Monitoring System Primary Failure Descriptor

Explanatory variables confounded with other explanatory variables (Removed):

Field Name
Pad or Platform Name
Location of ESP Supply Centre (Country)
Location of ESP Teardown Facility (Country)
Oil Density at STP (°API)
Dead Oil Viscosity at STP (cp)
Live Oil Viscosity at Reservoir Conditions (cp)
Oil Bubble Point Pressure (kPa,abs)
Reservoir/Zone Name
Reservoir Type
Reservoir Consolidation
Reservoir Recovery Mechanism
Reservoir Temperature (°C)
Well Type (Geometry)
Well Location (Onshore/Offshore)
Wellhead Location
Production Casing Outer Diameter (mm)
Production Casing Weight (kg/m)
Completion Type
Sand Control Type
Free Gas at Pump Intake (%)
Company Installing ESP System
Seal Vendor
Motor Vendor
Pump Intake Vendor
Cable Vendor

Variables with >10% missing data (Removed):

Pump Intake Pressure (kPa,g)
Producing Fluid Level (mKB to Fluid)
Reservior Static Pressure (Latest) (kPa,g)
Productivity Index (PI)
Pump Intake Temperature (°C)
Wellhead Pressure (kPa,g)
Wellhead Temperature (°C)
Casing Head Pressure (kPa,g)
Surface Sand Rate (m^3/d)
Sand Cut (%)
Other Solids Concentration
Total Solids Concentration
Solids?
Paraffin/Wax?
Asphaltenes?
H2S (% by Volume) (%)
Water pH
Water Salinity (Cl- ppm)
Sand?
Corrosion?
Motor Frequency (Hz)
Motor Voltage (Volts)
Motor Current (Amps)
Percent Deviation from Motor Rated Current
Number of Restarts During Period
Control Panel Vendor
Control Panel Model
Control Panel Power Rating (kVA)
Power Source
Power Quality
Power supply frequency (Hz)
Pump Series
Pump Type/Model
Pump Operator Option Suffix
Pump Housing
Pump Trim
Pump Impeller Type
Pump New/Used
Seal Series
Seal Type/Model
Seal Trim
Seal New/Used
Motor Series
Motor Type/Model
Motor Rated Voltage @ 60 Hz
Motor Rated Current @ 60 Hz
Motor Trim
Motor New/Used
Pump Intake Type
Pump Intake Series
Pump Intake Type/Model
Pump Intake Trim
Pump Intake New/Used
Cable Type (Round/Flat)
Cable Type/Model
Cable Power Rating
Cable Armour
Cable Material Type
Cable New/Used
MLE Type/Model
Wellhead Penetrator Type/Model
Packer Penetrator Type/Model
DH Monitoring System Vendor
DH Monitoring System New/Used
Shroud Installed?
Shroud Casing Outer Diameter (mm)
Tubing Outer Diameter (mm)
Tubing Weight (kg/m)
Packer Installed?
Packer Depth
Y-Tool Installed?
Pump Seating Depth MD (mKB)
Pump Seating Depth TVD (mKB)
Inclination at PSD (°)
Maximum Dogleg (°/30m)
Number of ESP Systems In Well
ESP System Configuration (Single/Parallel/Series)
First ESP System Installed in Well?

Non-informative or comment fields (Removed):

ProductionPeriodID
Company Name
Division Name
Field Comments
Number of Reservoirs
Reservoir Comments
Well Name
Production Period Number
Production Role
Well Comments
Period Comments
Control Panel Serial Number
Control Panel Part Number
Control Panel Comments
Pump Serial Number
Pump Comments
Seal Serial Number
Seal Comments
Motor Serial Number
Motor Comments
Pump Intake Serial Number
Pump Intake Comments
Cable Serial Number
Cable Comments
DH Monitoring System Comments
Artificial Lift Type
Qualification Status

Variables that were multiples of other variables (Removed):

Total Flow Rate (m^3/d)
Water Cut (%)
GOR & GLR (m^3/m^3)

APPENDIX D: REFERENCES

Bailey, W.J, *et al*.: "Survival Analysis: The Statistically Rigorous Method for Analyzing Electrical Submersible Pump System Performance," paper presented at the 2005 SPE Annual Technical Conference and Exhibition, Dallas, Texas October 9 – 12, SPE96772

Brookbank, B.: "How Do you Measure Run Life?", presented at the ESP Workshop May 1 – 3 1996 North Side Branch of Gulf Coast Chapter, SPE

Cran.R-project.org: "CRAN Task View: Survival Analysis", Version 2010-07-30

ESPRIFTS.com: Electric Submersible Pump Joint Industry Project data access

Fox, John "Cox Proportional-Hazards Regression for Survival Data", February 2002

Longnecker, M: STAT 641 Lecture Notes, Fall 2009

Patterson, M.M.: "A Model for Estimating the Life of Electrical Submersible Pumps", *SPE Production & Facilities*, November 1993, 247

Sawaryn, S.J, *et al*.: "The Analysis and Prediction of Electric-Submersible-Pump Failures in the Milne Point Field, Alaska", paper presented at the 1999 SPE Annual Technical Conference and Exhibition, Houston Texas October 3 – 6, SPE56663

Upchurch, E.R.: "Analyzing Electrical Submersible Pump Failures in the East Wilmington Field of California," paper presented at the 1991 SPE Electrical Submersible Pump Workshop, Houston, Texas April 30-May2, SPE20675

Weibull.com: "Characteristics of the Weibull Distribution", Reliability Hot Wire, Issue 14, April 2002

APPENDIX E – H: OUTPUT

APPENDIX E: OUTPUT: FINDING THE P50 TIME TO FAILURE FOR A DATASET

KM:

```
Call: survfit(formula = Surv(RunLife, Censor) ~ 1, data = espdata)

records   n.max n.start  events  median 0.95LCL 0.95UCL
   1588    1588    1588     745    1044     952    1113
```

CPH:

```
Call:  coxph(formula = Surv(RunLife, Censor) ~ 1, data = espdata)

Null model
  log likelihood= -4928.446
  n= 1588


Call: survfit.coxph.null(formula = CPHALLesp)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1   1588       3    0.998 0.00109        0.996        1.000
    3   1585       2    0.997 0.00141        0.994        1.000
    4   1582       2    0.996 0.00166        0.992        0.999
    5   1580       2    0.994 0.00188        0.991        0.998
    7   1577       2    0.993 0.00208        0.989        0.997
```

… (data removed)…

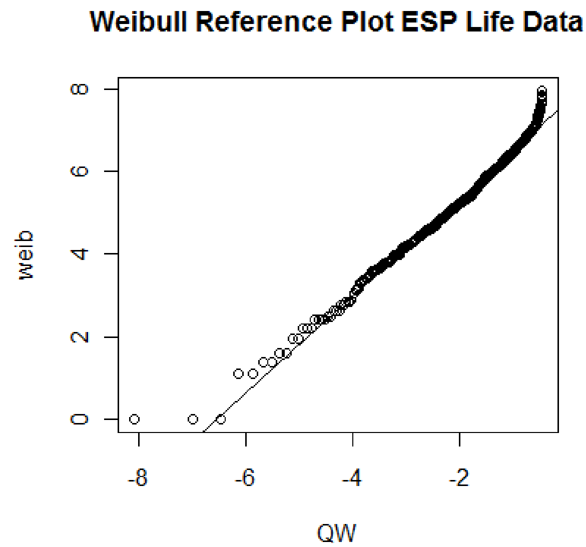
```
 1042    392       1    0.503 0.01498        0.475        0.533
 1043    391       2    0.501 0.01502        0.472        0.531
 1044    388       2    0.498 0.01505        0.469        0.528
 1047    386       1    0.497 0.01507        0.468        0.527
 1049    385       1    0.495 0.01508        0.467        0.526
 1052    382       1    0.494 0.01510        0.465        0.525
```

WEIBULL:

```
Call:
survreg(formula = Surv(espdata$RunLife, espdata$Censor) ~ 1,
    dist = "weibull")

Coefficients:
(Intercept)
   7.373513

Scale= 1.093867

Loglik(model)= -6204.2   Loglik(intercept only)= -6204.2
n= 1588
>
> # survreg's scale  =    1/(rweibull shape)
> shape=1/1.093837
> # survreg's intercept = log(rweibull scale)
> scale=exp(7.3729)
> shape
[1] 0.914213
> scale
[1] 1592.245
> median = scale*log(2)^(1/shape)
> median
[1] 1066.348
```

GOODNESS OF FIT:

```
> #Anderson Darling Fit
> exvalphi=-log(scale)
> exvaltheta=1/shape
> x=AD_test1$RunLife
> x2=espdata$RunLife
> n=length(x)
> n2=length(x2)
> y=-log(x)
> y=sort(y)
> i=seq(1:n)
> a = exvalphi
> b = exvaltheta
> z=exp(-exp(-(y-a)/b))
> A1i=(2*i-1)*log(z)
> A2i=(2*n2+1-2*i)*log(1-z)
> s1=sum(A1i)
> s2=sum(A2i)
> AD=-n2-(1/n2)*(s1+s2)
> AD=AD*(1+.2/sqrt(n2))
> AD  #Reject fit at a=.05 if is AD >.757
[1] 143.2291
```



APPENDIX F: OUTPUT: COMPARING TWO SURVIVAL CURVES DIFFERING BY A FACTOR

KM:

```
Call: survfit(formula = Surv(RunLife, Censor) ~ Emulsion, data = espdata)

              records n.max n.start events median 0.95LCL 0.95UCL
Emulsion=No      1217  1217    1217    434   1508    1194    1783
Emulsion=Yes      371   371     371    311    606     519     726
```

CPH:

```
Call:
coxph(formula = Surv(RunLife, Censor) ~ Emulsion, data = espdata)

  n= 1588

               coef exp(coef) se(coef)     z Pr(>|z|)
EmulsionYes 0.92705   2.52705  0.07439 12.46   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

            exp(coef) exp(-coef) lower .95 upper .95
EmulsionYes     2.527     0.3957     2.184     2.924

Rsquare= 0.086   (max possible= 0.998 )
Likelihood ratio test= 143.2  on 1 df,   p=0
Wald test            = 155.3  on 1 df,   p=0
Score (logrank) test = 166.7  on 1 df,   p=0
```

```
Call: survfit(formula = CPHALLesp2, newdata = esp2dftest)

 time n.risk n.event survival1 survival2
    1   1588       3     0.999    0.9965
    3   1585       2     0.998    0.9941
    4   1582       2     0.997    0.9918
    5   1580       2     0.996    0.9895
    7   1577       2     0.995    0.9871
```

…(data removed)…

```
  532    818       1     0.762    0.5033
  533    817       2     0.761    0.5010
  537    813       1     0.760    0.4999
  542    808       1     0.759    0.4987
  544    806       1     0.759    0.4976
  547    804       1     0.758    0.4964
```

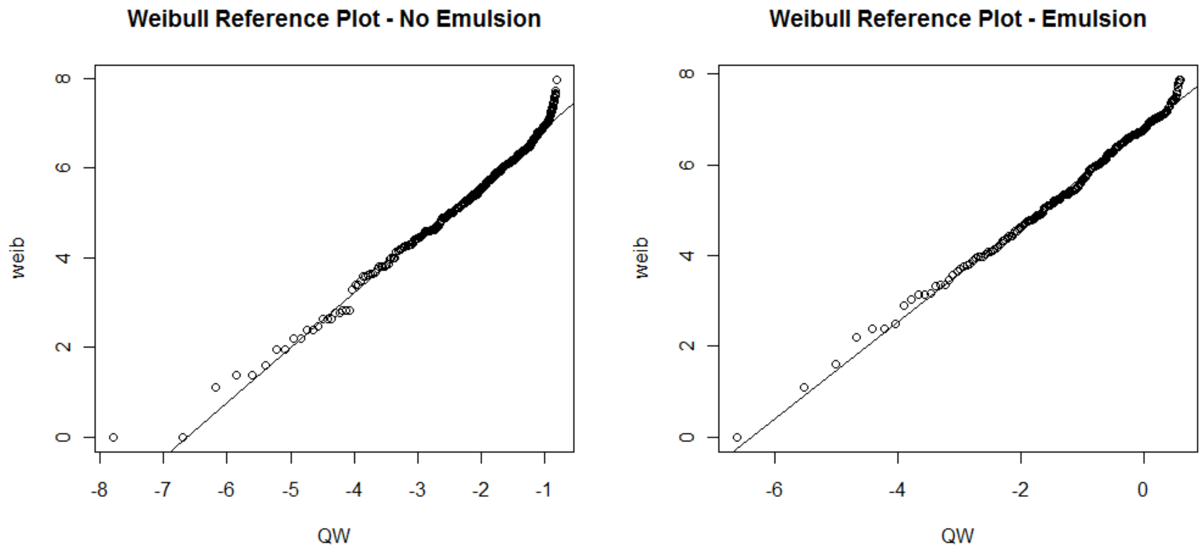
…(data removed)…

```
 1334    229       1     0.506    0.1785
 1342    226       1     0.504    0.1769
 1403    209       1     0.502    0.1753
 1408    206       1     0.500    0.1736
 1423    201       1     0.498    0.1720
 1425    200       1     0.496    0.1703
 1449    196       1     0.494    0.1686
 > cox.zph(CPHALLesp2)
                rho chisq      p
 EmulsionYes 0.0903  6.11 0.0134
```

WEIBULL:

```
Call:
survreg(formula = Surv(espdata$RunLife, espdata$Censor) ~ espdata$Emulsion,
    dist = "weibull")

Coefficients:
        (Intercept) espdata$EmulsionYes
           7.686692           -1.014834

Scale= 1.084455

Loglik(model)= -6131.2   Loglik(intercept only)= -6204.2
        Chisq= 145.98 on 1 degrees of freedom, p= 0
n= 1588
> # survreg's scale  =    1/(rweibull shape)
> shape2=1/1.084547
> # survreg's intercept = log(rweibull scale)
> scale2=exp(7.685359)
> shape3=1/1.084547
> scale3=exp(7.685359-1.013514)
> median_noemulsion=scale2*log(2)^(1/shape2)
> median_emulsion=scale3*log(2)^(1/shape3)
> effect=1/(exp(-1.013514))
> median_noemulsion
[1] 1462.436
> median_emulsion
[1] 530.7783
> effect
[1] 2.755266
```

GOODNESS OF FIT (NOTE: GOODNESS OF FIT MET FOR "EMULSION" PLOT)



APPENDIX G: OUTPUT: CHOOSING THE VARIABLES THAT CHARACTERIZE A SURVIVAL CURVE

FIRST STEP:

```
Start:  AIC=9117.55
Surv(RunLife, Censor) ~ Country + Offshore + Oil + Water + Gas +
    Scale + CO2 + Emulsion + CtrlPanelType + NoPumpHouse + PumpVendor +
    NoPumpStage + NoSealHouse + NoMotorHouse + MotorPowerRating +
    NoIntakes + NoCableSys + CableSize + DHMonitorInstalled +
    DeployMethod

                     Df    AIC
- Offshore            1 9115.6
- DeployMethod        1 9115.6
- NoIntakes           1 9115.6
- NoMotorHouse        1 9115.6
- DHMonitorInstalled  1 9115.6
- NoSealHouse         1 9116.0
- MotorPowerRating    1 9116.1
- Emulsion            1 9116.4
- CO2                 1 9116.8
<none>                  9117.6
- PumpVendor          1 9119.3
- CtrlPanelType       1 9126.3
- NoPumpStage         1 9126.9
- Country             6 9128.1
- NoPumpHouse         1 9128.3
- NoCableSys          1 9128.6
- Oil                 1 9130.6
- Water               1 9131.0
- CableSize           1 9133.0
- Gas                 1 9134.5
- Scale               4 9425.2
```

Final Step (10 steps):

```
Step:  AIC=9102.92
Surv(RunLife, Censor) ~ Country + Oil + Water + Gas + Scale +
    CtrlPanelType + NoPumpHouse + PumpVendor + NoPumpStage +
    NoCableSys + CableSize

                Df    AIC
<none>             9102.9
- PumpVendor     1 9105.2
- NoPumpStage    1 9112.0
- NoPumpHouse    1 9112.9
- NoCableSys     1 9113.4
- Oil            1 9113.9
- CtrlPanelType  1 9114.3
- CableSize      1 9118.2
- Gas            1 9119.2
- Water          1 9121.6
- Country        6 9296.6
- Scale          4 9478.4
```

Summary:

```
Call:
coxph(formula = Surv(RunLife, Censor) ~ Country + Oil + Water +
    Gas + Scale + CtrlPanelType + NoPumpHouse + PumpVendor +
    NoPumpStage + NoCableSys + CableSize, data = espdata)

  n= 1588

                                               coef exp(coef)  se(coef)      z Pr(>|z|)
CountryArgentina _El Trapial              2.208e+00 9.094e+00 7.218e-01  3.059 0.002224 **
CountryThailand _Benchamas                2.178e+00 8.827e+00 5.287e-01  4.119 3.80e-05 ***
CountryThailand _Benchamas North West     2.455e+00 1.165e+01 8.370e-01  2.934 0.003350 **
CountryUnited Kingdom _Captain           -1.864e+00 1.551e-01 1.032e+00 -1.806 0.070851 .
CountryUnited States of America _Rangely  1.917e+00 6.800e+00 1.265e+00  1.515 0.129786
CountryVenezuela _Boscan                  6.428e+00 6.188e+02 9.327e-01  6.892 5.51e-12 ***
Oil                                       1.952e-03 1.002e+00 4.581e-04  4.260 2.04e-05 ***
Water                                    -1.590e-03 9.984e-01 3.590e-04 -4.429 9.46e-06 ***
Gas                                       3.317e-02 1.034e+00 7.245e-03  4.579 4.68e-06 ***
ScaleLight                                3.045e-01 1.356e+00 6.186e-01  0.492 0.622514
ScaleModerate                            -4.587e+00 1.019e-02 8.109e-01 -5.657 1.54e-08 ***
ScaleSevere                               4.391e-01 1.551e+00 5.752e-01  0.763 0.445241
ScaleYes-Present                          7.174e-01 2.049e+00 6.008e-01  1.194 0.232389
CtrlPanelTypeVariable Speed Drive        -4.421e-01 6.427e-01 1.234e-01 -3.583 0.000340 ***
NoPumpHouse                               2.593e-01 1.296e+00 7.356e-02  3.524 0.000424 ***
PumpVendorREDA                           -4.055e-01 6.667e-01 1.915e-01 -2.117 0.034237 *
NoPumpStage                              -2.249e-03 9.978e-01 6.823e-04 -3.296 0.000981 ***
NoCableSys                                9.645e-01 2.624e+00 2.269e-01  4.250 2.14e-05 ***
CableSize                                -2.020e-01 8.171e-01 4.819e-02 -4.192 2.77e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Rsquare= 0.393   (max possible= 0.998 )
Likelihood ratio test= 792  on 19 df,   p=0
Wald test            = 258.5  on 19 df,   p=0
Score (logrank) test = 589.4  on 19 df,   p=0
```

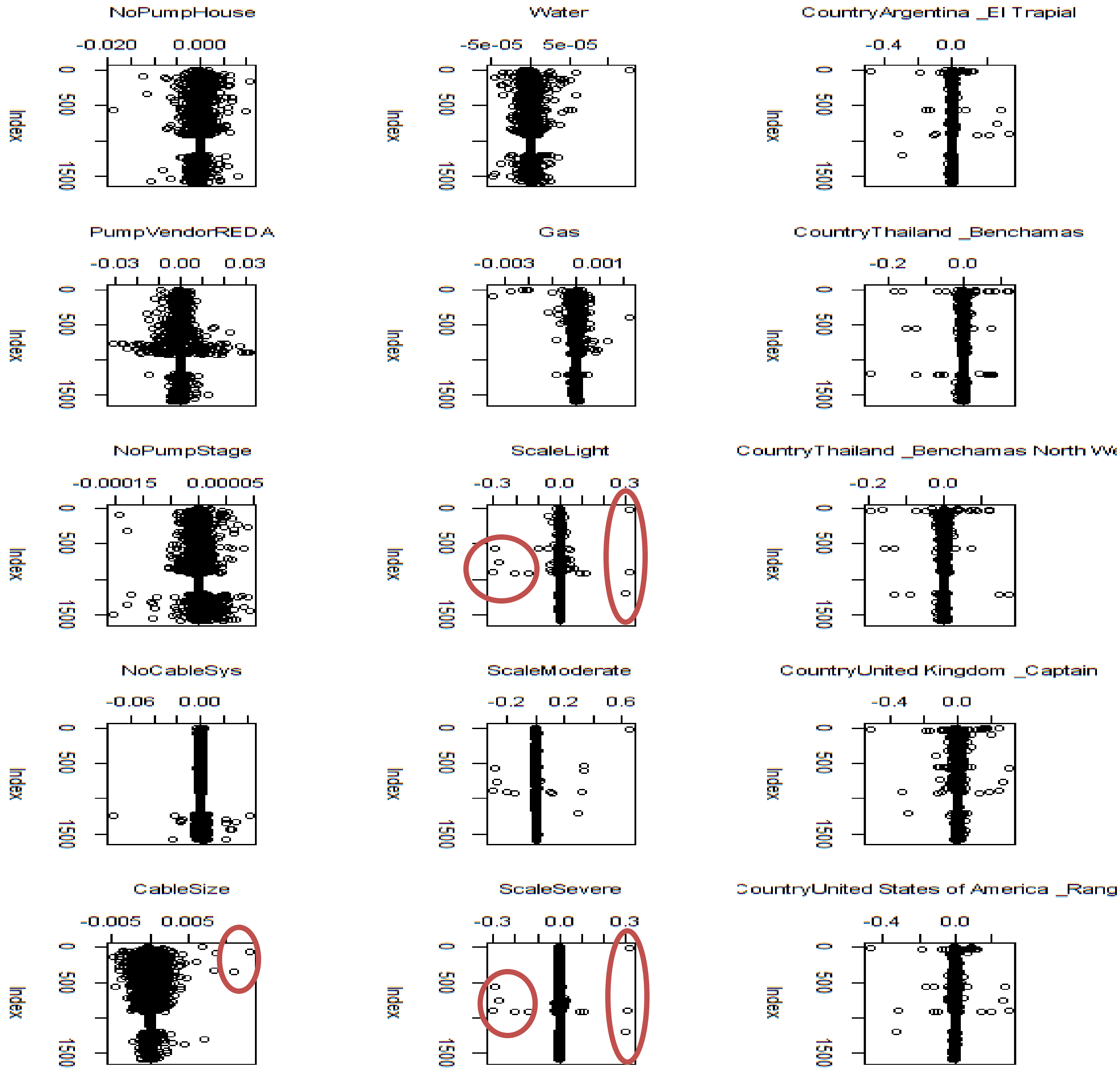
APPENDIX H: OUTPUT: TESTING CPH ASSUMPTIONS

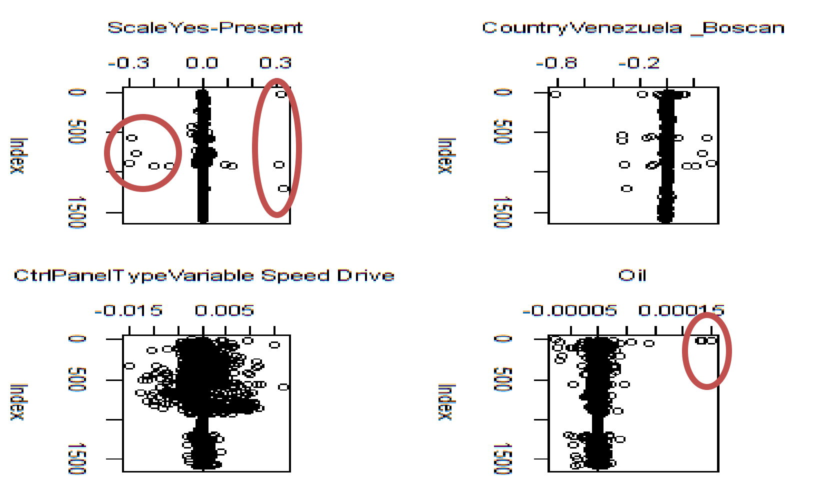
PROPORTIONAL HAZARDS:

```
                                                rho   chisq      p
CountryArgentina _El Trapial                -0.03327  1.6532 0.1985
CountryThailand _Benchamas                  -0.04316  1.3901 0.2384
CountryThailand _Benchamas North West        0.02147  0.3527 0.5526
CountryUnited Kingdom _Captain              -0.04893  2.4512 0.1174
CountryUnited States of America _Rangely -0.01999  0.3958 0.5293
CountryVenezuela _Boscan                    -0.02439  0.8619 0.3532
Oil                                          0.00555  0.0361 0.8492
Water                                        0.02479  0.5322 0.4657
Gas                                          0.00646  0.0402 0.8411
ScaleLight                                   0.03762  2.1458 0.1430
ScaleModerate                                0.02460  0.8710 0.3507
ScaleSevere                                  0.02826  1.3144 0.2516
ScaleYes-Present                             0.04502  3.2330 0.0722
CtrlPanelTypeVariable Speed Drive            0.03320  0.8185 0.3656
NoPumpHouse                                  0.00848  0.0576 0.8103
PumpVendorREDA                               0.01189  0.1126 0.7372
NoPumpStage                                 -0.04296  1.7375 0.1875
NoCableSys                                   0.05657  2.3420 0.1259
CableSize                                    0.07870  4.8095 0.0283
GLOBAL                                            NA 29.2059 0.0628
```

Need a global pvalue of > 0.05 to fail to reject null hypothesis of proportional hazards

INFLUENTIAL OBSERVATIONS:





- Oil and cable size have a few influential observations on the order of magnitude of the coefficients, but I do not have enough knowledge about the dataset to remove them or understand the implications
- All the levels of scale have influential observations I'd be concerned about. As the residuals look the same, this leads me to think there are issues with the comparative level of Scale-None (possible data-miscoding). That coupled with the fact that the response led to a non-intuitive interpretation, I would recommend to the JIP that they put more stringent controls around the scale data

NON-LINEARITY:

- NoCableSys, CableSize, & NoPumpHouse could have been treated as factors given their low number of levels (doing this did not change the R-square of the model)
- Oil shows a potentially quadratic trend however this is mostly caused by the few data points for oil >750m3/day. Given additional data collection, this is a relationship that should be investigated
- Gas is linear but for the 2 points above 800,000 M3/day. Given additional data collection of wells in the higher ranges, this is a relationship that should be investigated.