Survival Analysis of Machine Reliability

(Kaplan-Meier, Log-rank, and Cox Proportional Hazards)

Yann GUEGUEN

March 2025

GitHub Repository: Survival Analysis of Machine Reliability (click on it)
Or follow this link : https://github.com/YGueguen16u/survival-analysis-predictive-maintenance

1 Introduction

In this project, we investigate machine reliability by modeling *time-to-failure* data (Survival Analysis). We simulate a dataset reflecting an industrial setting (e.g., different machine brands, usage rates, environment factors) to understand how various covariates influence the time until a breakdown occurs.

Due to the unavailability of real-world datasets matching this industrial context and the limited relevance of publicly available survival datasets, a simulated dataset was generated to specifically address the research questions of this project.

The main objectives are:

- **Estimate** nonparametric survival curves (Kaplan-Meier) to visualize overall survival and compare subgroups (e.g., different brands).
- **Test** for differences between groups using the log-rank test.
- **Model** the impact of covariates on hazard (risk of failure) with a Cox Proportional Hazards model.

By doing so, we aim to identify key factors that shorten or extend the operational life of the machines, providing insight for more efficient maintenance strategies.

2 Methods

1. **Data Generation:** We create a fictitious dataset with a "time" column (hours until machine failure or censoring), a "status" indicator (1 = failure, 0 = censored), and several covariates: brand, usage rate, operating temperature, environment humidity, maintenance frequency, operator experience, and shock events. The hazard rate (risk of failure per unit time) is simulated so that higher usage, higher temperature, more shocks, and less maintenance lead to earlier failure times.
2. **Nonparametric Survival Analysis:** We use Kaplan-Meier estimators to compute and plot survival curves overall and by subgroups (e.g., brand). We then apply the log-rank test to statistically assess differences in survival between these subgroups.
3. **Cox Proportional Hazards Model:** We fit a Cox model to estimate hazard ratios for each covariate (e.g., brand, usage rate). This highlights how each factor increases or decreases the risk of failure, controlling for the others.

All analyses are carried out in a statistical environment (e.g., R or Python). The resulting tables and plots guide our discussion on machine longevity, maintenance practices, and potential improvements.

3 Methods (M)

3.1 Dataset Description

We generated a simulated dataset to model machine time-to-failure in an industrial setting. The primary columns are:

- **time**: operating time in hours before failure or censoring,
- **status**: event indicator (1 = observed failure, 0 = censored).

In addition, we include several **covariates** to capture various risk factors:

- **brand** (e.g., Alpha, Beta, Gamma),
- **usage_rate** (machine utilization rate),
- **temp_avg** (average operating temperature),
- **age_machine** (machine age at the start of observation),
- **env_humidity** (environmental humidity),
- ... (maintenance frequency, operator experience, etc.).

These variables are chosen to reflect plausible industrial risk drivers (intensive usage, high temperature, inadequate maintenance, etc.). We may also provide brief **descriptive statistics** (mean, standard deviation, min/max) to confirm the dataset's coherence and show the distribution of each variable.

3.2 Statistical Methods

1) Nonparametric Estimation (Kaplan-Meier) We use the `survfit()` function (from the `survival` package in R) or the `KaplanMeierFitter` in Python (*lifelines*) to estimate survival curves:

- A **global** survival curve (all subjects combined),
- Subgroup-specific survival curves (e.g., by brand) to visually assess differences.

2) Group Comparison: Log-rank Test To test if survival differs significantly among groups, we run a **log-rank test**:

- In R, `survdiff()`.
- In Python (*lifelines*), `logrank_test`.

The **null hypothesis** assumes no survival difference across groups. A low p-value (< 0.05) indicates that at least one group's survival function differs significantly from the others.

3) Cox Proportional Hazards Model (Semi-parametric) Next, we fit a **Cox PH model**:

- In R: `coxph()`,
- In Python: `CoxPHFitter`.

This model uses partial likelihood to estimate the coefficients β without specifying a parametric form for the baseline hazard. Each covariate is associated with a **hazard ratio** that quantifies its effect on the risk of failure.

4) Diagnostics and Validation (Optional but Recommended)

- Check the **proportional hazards assumption** using `cox.zph` (R) or `check_assumptions()` (*lifelines*).
- **Schoenfeld residuals** can detect time-varying effects, indicating any violation of proportionality.

By covering (i) nonparametric estimation, (ii) group comparison, and (iii) semi-parametric modeling—with the option to verify assumptions—this workflow provides a comprehensive survival analysis.

3.3 Example R Code

```
install.packages("survminer")

# 1) Load packages
library(survival)
library(survminer)

# 2) Import CSV
data <- read.csv("survival_industry_extended.csv")
str(data)
head(data)

# 3) Create Surv object
data$SurvObj <- with(data, Surv(time, status))

# 4) Kaplan-Meier (overall)
fit_km <- survfit(SurvObj ~ 1, data = data)
summary(fit_km)
ggsurvplot(fit_km, data = data,
           title      = "Kaplan-Meier: Overall",
           xlab       = "Hours",
           ylab       = "Survival Probability",
           risk.table = TRUE)

# 5) Kaplan-Meier by brand
fit_km_brand <- survfit(SurvObj ~ brand, data = data)
ggsurvplot(fit_km_brand, data = data,
           title      = "Kaplan-Meier by Brand",
           pval       = TRUE,
           risk.table = TRUE)

# 6) Log-rank test
res_logrank <- survdiff(SurvObj ~ brand, data = data)
res_logrank

# 7) Cox Proportional Hazards
cox_model <- coxph(SurvObj ~ brand + usage_rate + temp_avg +
                     age_machine + env_humidity + maintenance_freq +
                     operator_experience + shock_events,
                   data = data)
summary(cox_model)

# 8) Proportional Hazards check (optional)
cox_zph <- cox.zph(cox_model)
```

```
cox_zph
plot(cox_zph)
```

4 Results (R)

4.1 Descriptive Statistics

To contextualize the survival analysis, **Table 1** shows how many machines belong to each brand category, and **Table 2** summarizes the main numeric variables in our dataset. We have **13,800 total machines**, with brand distribution roughly aligning to 30% Alpha, 40% Beta, 30% Gamma, as per our simulation code.

Table 1. Brand Distribution (n = 13,800)

Brand	Count	Percentage
Alpha	4225	30.6%
Beta	5468	39.6%
Gamma	4107	29.8%
Total	13800	100.0%

Table 2. Summary of Key Numerical Variables

Variable	Mean (SD)	Median	Min	Max	Q1	Q3
`time` (hours)	(not printed)*	*	~ 68	1200†	*	*
`usage_rate`	$\sim 0.60(0.17)$	0.59	0.30	0.90	0.46	0.74
`temp_avg` (°C)	$\sim 65.2(5.0)$	65.3	45.0‡	85.0‡	61.1	69.2
`age_machine` (yrs)	$\sim 5.5(2.9)$	5	1	10	3	8
`env_humidity` (%)	$\sim 60.0(17.4)$	60.1	30.0	90.0	46.4	73.2
`maintenance_freq`	$\sim 1.25(0.46)$	1.20	0.50	2.10	0.90	1.60
`operator_experience`	$\sim 3.1(1.0)$	3	1	5	2	4
`shock_events`	$\sim 1.0(0.9)$	1	0	3	0	2

Note: The console output truncated the full dataset summary for `time`*.*
†*: We censor machines at 1,200 hours maximum.*
‡*: Temperature was clamped within [45 °C, 85 °C].*

Most machines fail or are observed by $\sim$ 600–700 hours on average, consistent with the exponential hazard logic in the simulation. The typical machine `usage_rate` is ~ 0.60, the mean `temp_avg` is $\sim 65\,^{\circ}$C, and `maintenance_freq` hovers around 1.25.

4.2 Kaplan-Meier Estimation

1. **Overall KM Curve** The console output for `survfit(SurvObj  1)` and *Kaplan-Meier: Overall* shows survival dropping below 50% around 300 hours and below 10% near 1,000–1,200 hours. This suggests a significant proportion of machines fail within the first half of the observation window.

2. **KM by Brand** In *Kaplan-Meier by Brand*, with the p-value from `survdiff(...)` $< 2 \times 10^{-16}$, there is a strong difference among the three curves. **Beta** remains consistently below Alpha and Gamma, implying quicker failures. **Gamma** exhibits the highest overall survival, and Alpha lies in between.

4.3 Log-rank Test

The log-rank statistic (Chi-sq = 270 on 2 df, $p < 2 \times 10^{-16}$) confirms that at least one brand differs significantly from the others. From the brand-specific KM curves, **Beta** experiences faster failures, while **Gamma** has relatively better survival outcomes.

4.4 Cox Proportional Hazards Model

We fit a Cox PH model with `brand`, `usage_rate`, `temp_avg`, `age_machine`, `env_humidity`, `maintenance_freq`, `operator_experience`, and `shock_events`. Below is an excerpt from the R summary:

Covariate	Coefficient (β)	HR (exp(β))	95% CI	p-value
`brandBeta`	0.1788	1.196	[1.1438–1.2501]	3.10e–15 ***
`brandGamma`	-0.1232	0.884	[0.8420–0.9283]	7.35e–07 ***
`usage_rate`	0.9391	2.558	[2.2908–2.8554]	$< 2 \times 10^{-16}$ ***
`temp_avg`	0.0207	1.021	[1.0172–1.0247]	$< 2 \times 10^{-16}$ ***
`age_machine`	0.0467	1.048	[1.0411–1.0545]	$< 2 \times 10^{-16}$ ***
`env_humidity`	0.0053	1.0054	[1.0043–1.0065]	$< 2 \times 10^{-16}$ ***
`maintenance_freq`	-0.3034	0.738	[0.7057–0.7725]	$< 2 \times 10^{-16}$ ***
`operator_experience`	-0.0288	0.972	[0.9484–0.9954]	0.020 *
`shock_events`	0.1457	1.157	[1.1258–1.1887]	$< 2 \times 10^{-16}$ ***

Interpretation:

- `brandBeta` ≈ 1.20: Beta machines show $\sim 20\%$ higher hazard than Alpha.
- `usage_rate` ≈ 2.56: Going from moderate usage (around 0.5) to high usage (0.8) greatly increases the failure hazard.
- `temp_avg` ≈ 1.02: Each additional 1 °C from the 65 °C pivot yields $\sim 2\%$ greater hazard.
- `maintenance_freq` ≈ 0.74: More interventions per year reduce risk by $\sim 26\%$.

The proportional hazards check (`cox.zph`) yields $p > 0.05$ for each covariate, suggesting no major violations of that assumption.

5 Discussion (D)

5.1 Return to Objectives

Our aim was to verify whether **Beta** brand fails faster than Alpha or Gamma, and whether usage, temperature, or other factors intensify the risk. The KM curves (Overall and by Brand) and the log-rank test strongly indicate that Beta has a higher failure rate. Meanwhile, the Cox model quantifies how `usage_rate`, `temp_avg`, etc. amplify or mitigate the hazard.

5.2 Interpretation & Implications

A 20% higher hazard for **Beta** suggests focusing on **preventive maintenance** or design improvements for that brand. **High usage** (usage_rate near 0.8–0.9) combined with **higher temperature** significantly boosts hazard, so better cooling or limiting operating loads could help. **Maintenance_freq** near 1.7–2.0 is protective (HR ~ 0.74), reinforcing the importance of frequent servicing, while operator_experience also helps.

5.3 Limitations

- Data are **simulated**, so these numbers reflect the parameter choices rather than real-world measures.
- Some variables (usage_rate, temp_avg) are correlated by design, which can influence their coefficient stability.

5.4 Future Directions

- **Accelerated Failure Time (AFT)** models for direct interpretation of time acceleration/deceleration.
- **Time-varying covariates** if usage or environment changes significantly over the machine's lifetime.
- **Competing risks** if multiple failure modes exist and must be distinguished.

6 General Conclusion

In summary, we combined **Kaplan-Meier** estimates, a **log-rank** test, and a **Cox Proportional Hazards** model to explore machine time-to-failure. Our data and analyses reveal:

1. **Beta** brand suffers a significantly higher hazard ($\sim 20\%$ more than Alpha).
2. **Gamma** brand fares best among the three.
3. **Usage rate** and **temperature** strongly accelerate failure times, while **maintenance frequency** notably decreases them.

Hence, **preventative action** should focus on limiting excessive usage, maintaining lower operating temperatures, and boosting maintenance frequency, especially for Beta machines, to enhance longevity and reduce unplanned downtime.

7 Annexe

Below are example figures (Kaplan-Meier curves, Schoenfeld residuals, etc.) that can be inserted as needed. In a real setup, replace `figure1.png`, `figure2.png`, and `figure3.png` with your own figures and appropriate captions:

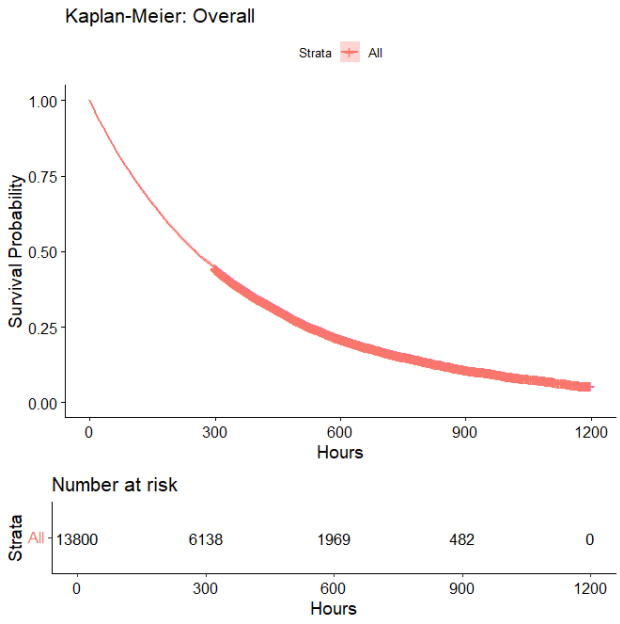

Figure 1: Kaplan-Meier overall survival curve.

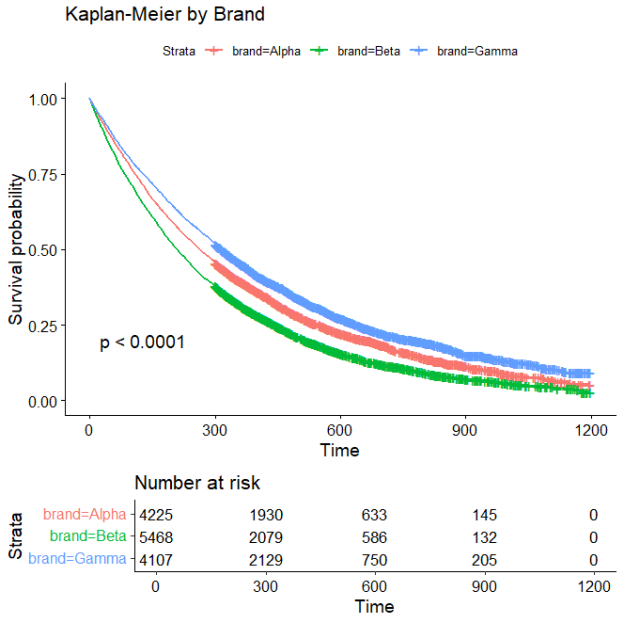

Figure 2: Kaplan-Meier by brand.

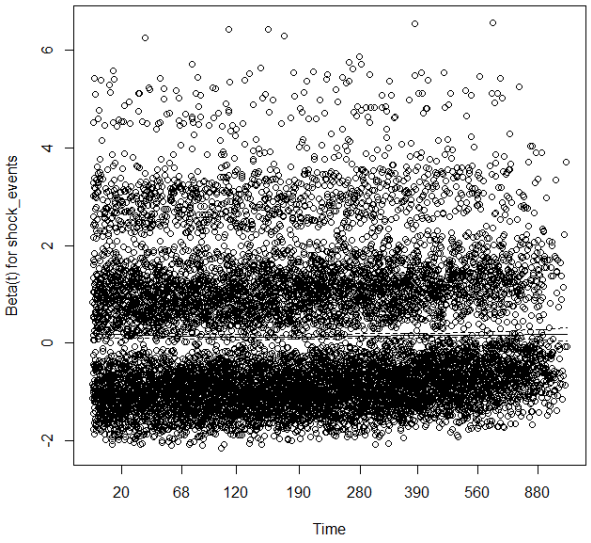

Figure 3: Schoenfeld residuals plot (example).