Survivability Using Cox Regression

In patients Diagnosed with Cancer

By Nathan Butler

Purpose and Application

- Purpose: Investigating Time to Event
  - Used for analyzing the impact of variables on event occurrence (e.g., death, disease recurrence, machine failure).
  - Measures time until a specific event happens.
- Applications:
  - Medical Research: Predicting time until cancer relapse or death.
  - Engineering: Assessing the reliability of machine components.
  - Social Sciences: Analyzing event timelines in sociological studies.

Advantages & Disadvantages

- Advantages:
  - Flexibility: No assumption about hazard function shape.
  - Censoring: Considers subjects lost to follow-up or not experiencing the event.
  - Proportional Hazards: Tests & adjusts for constant predictor effects over time.
- Disadvantages
  - Relies on the proportional hazards assumption, which may not always hold.
  - Does not provide survival probabilities or median survival times directly.
  - Sensitive to outliers or model assumptions, requiring careful data checking and diagnostics.

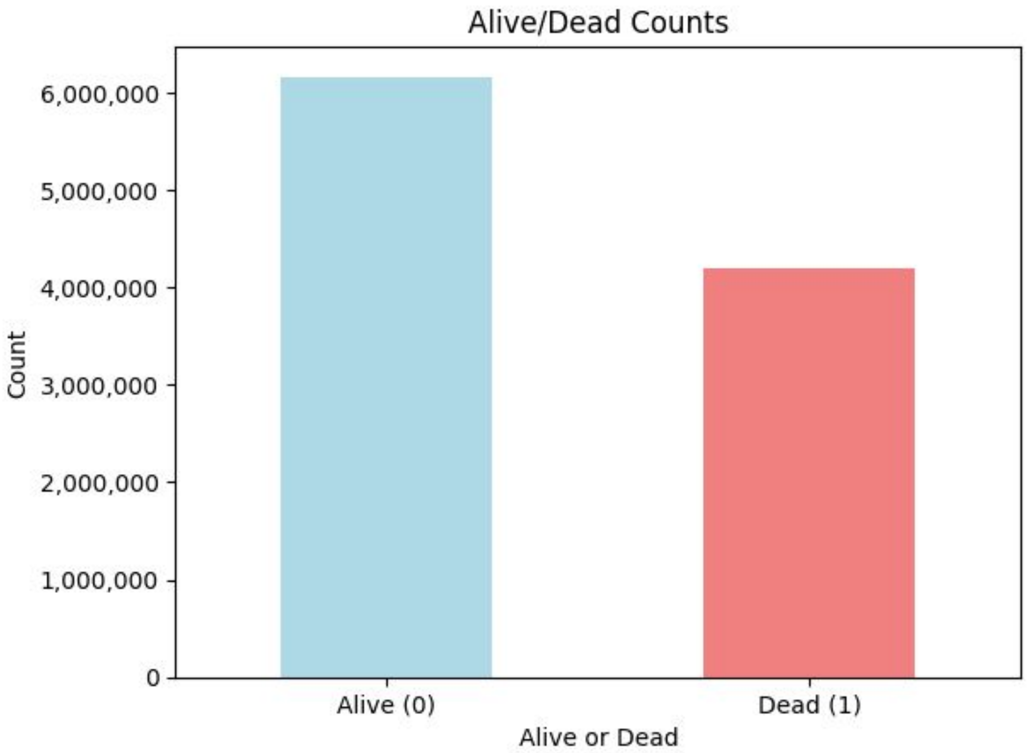
How Cox Regression Works

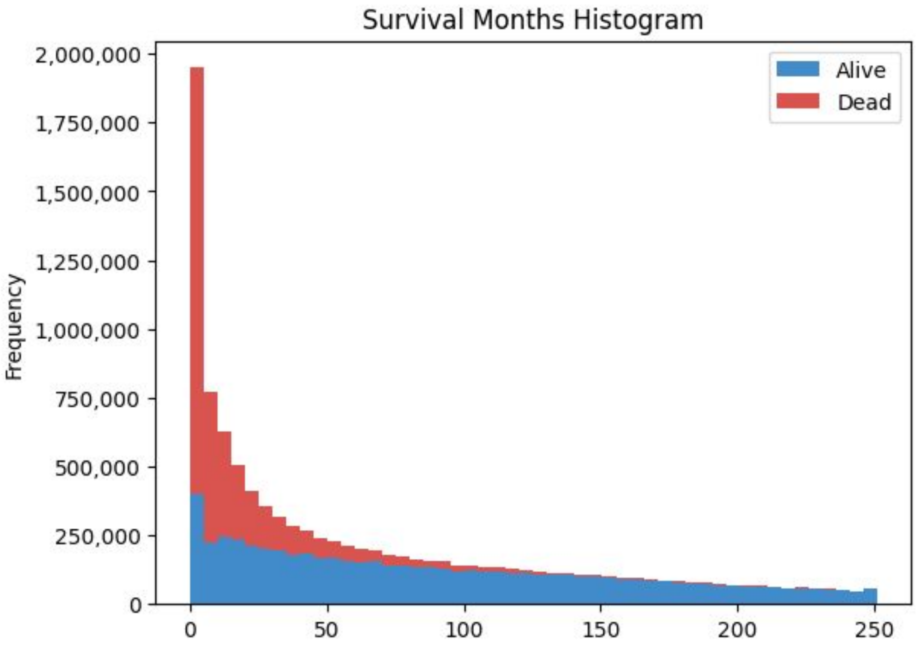
Mathematics Behind the Model:

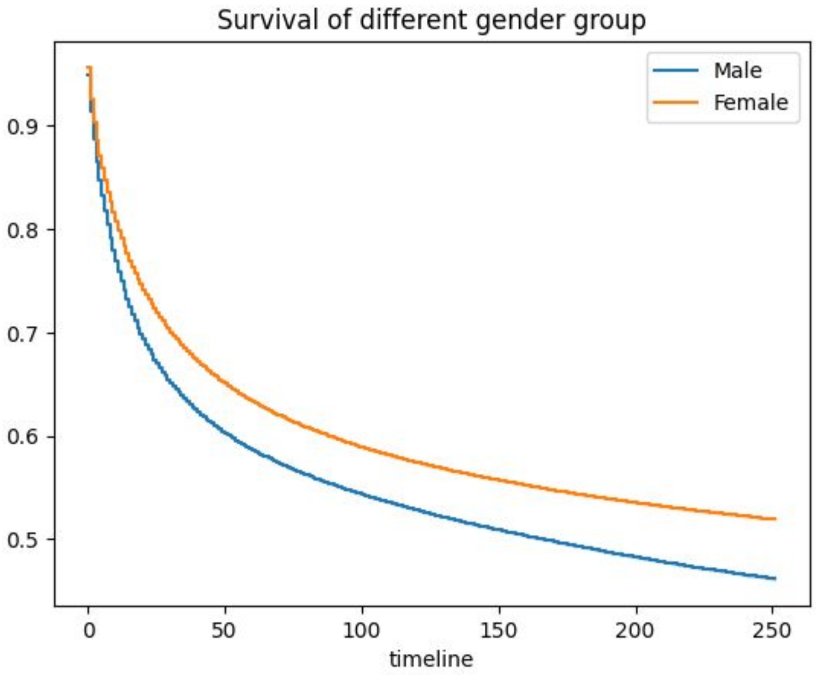
- The Cox model hazard function:
    - $$h(t)=h_0(t)\times\exp(b_1X_1+b_2X_2+\ldots+b_{\square}X_{\square})$$
- $h(t)$ is the hazard function, $h_0(t)$ is the baseline hazard, and $b_1, b_2, \ldots, b_{\square}$ are regression coefficients.
- Model Assumptions:
    - Assumes the hazard ratio is constant over time, known as the proportional hazards assumption.
    - The model can include time-varying covariates, interactions, stratification, and frailty.
- Estimation Method:
    - Utilizes the partial likelihood method, maximizing the likelihood of observing event order without specifying the baseline hazard function.

Cancer Survival Analysis

- Primary event is alive or dead
- Starting dataframe is roughly 12 million through filtering we bring it down to 10
- 6 million alive and 4 million dead







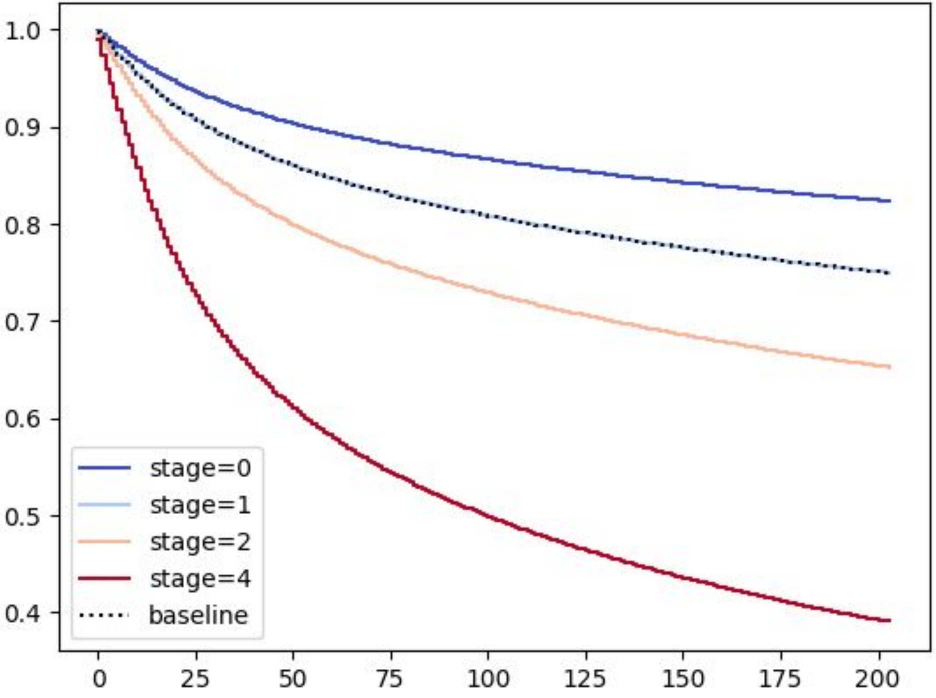
- Many more died in the first months in this df.
- Steady drop off of patient counts at longer time periods
- KMF(Kaplan-Meier curve)
- Male survivability drops off faster

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
age	0.02	1.02	0.00	0.02	0.02	1.02	1.02	0.00	461.97	<0.005	inf
Sex	-0.13	0.88	0.00	-0.13	-0.13	0.88	0.88	0.00	-90.58	<0.005	inf
Months from diagnosis to treatment	-0.05	0.95	0.00	-0.05	-0.05	0.95	0.95	0.00	-104.70	<0.005	inf
chemo	0.17	1.19	0.00	0.17	0.18	1.19	1.19	0.00	111.94	<0.005	inf
Income	-0.00	1.00	0.00	-0.00	-0.00	1.00	1.00	0.00	-89.48	<0.005	inf
stage	0.39	1.48	0.00	0.39	0.40	1.48	1.49	0.00	577.48	<0.005	inf
Late or early	-0.51	0.60	0.00	-0.52	-0.51	0.60	0.60	0.00	-282.38	<0.005	inf
Total number of in situ/malignant tumors for patient	-0.01	0.99	0.00	-0.01	-0.00	0.99	1.00	0.00	-6.37	<0.005	32.32

Concordance	0.85
Partial AIC	49603193.23
log-likelihood ratio test	2366392.69 on 100 df
-log2(p) of ll-ratio test	inf

- Positive coef means adds to death rate every integer and negative means takes away every integer.
- Concordance 85% means the model can positively identify 85% of deaths

Cancer Stage Predictions



- One of the most important factors
- What are the different stages?
- Stage 3?