Distribution-Free Inference Methods for Threshold Regression

G. A. Whitmore

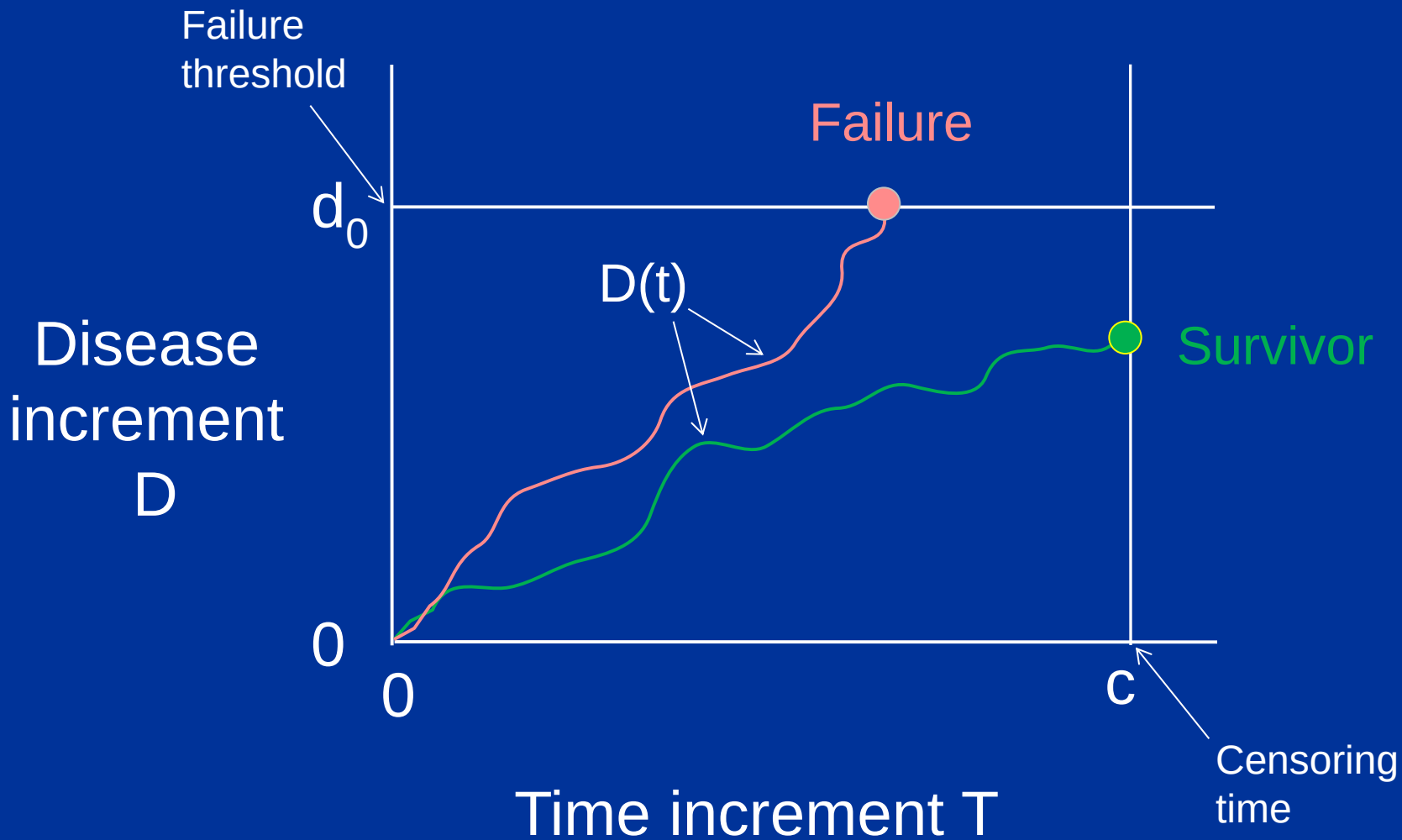
McGill University, Montreal

george.whitmore@mcgill.ca

Joint work with:

Mei-Ling T. Lee
University of Maryland – College Park

A picture worth 1000 words!



Introduction

Disease advances until it triggers a failure event when a critical threshold is reached for the first time.

Complete data consist of survival times of failures and current disease levels of survivors.

Our model and methodology require few distributional assumptions.

Estimation and predictive inference methods are developed.

Covariates are incorporated as in threshold regression

Computational aspects of the approach are straightforward.

Overview of Threshold Regression

Threshold regression (TR) applies to first hitting times of a threshold or boundary by sample paths of a stochastic process.

Regression functions are used for (1) the threshold level, (2) process parameters, or (3) the time scale of the process.

Previous applications of the TR model have been fully parametric. For example, event analysis based on first hitting times in a Wiener diffusion process.

Our approach here takes TR toward a distribution-free and, hence, more robust analysis.

The Model

A disease process denoted by D(t), with initial value D(0)=0.

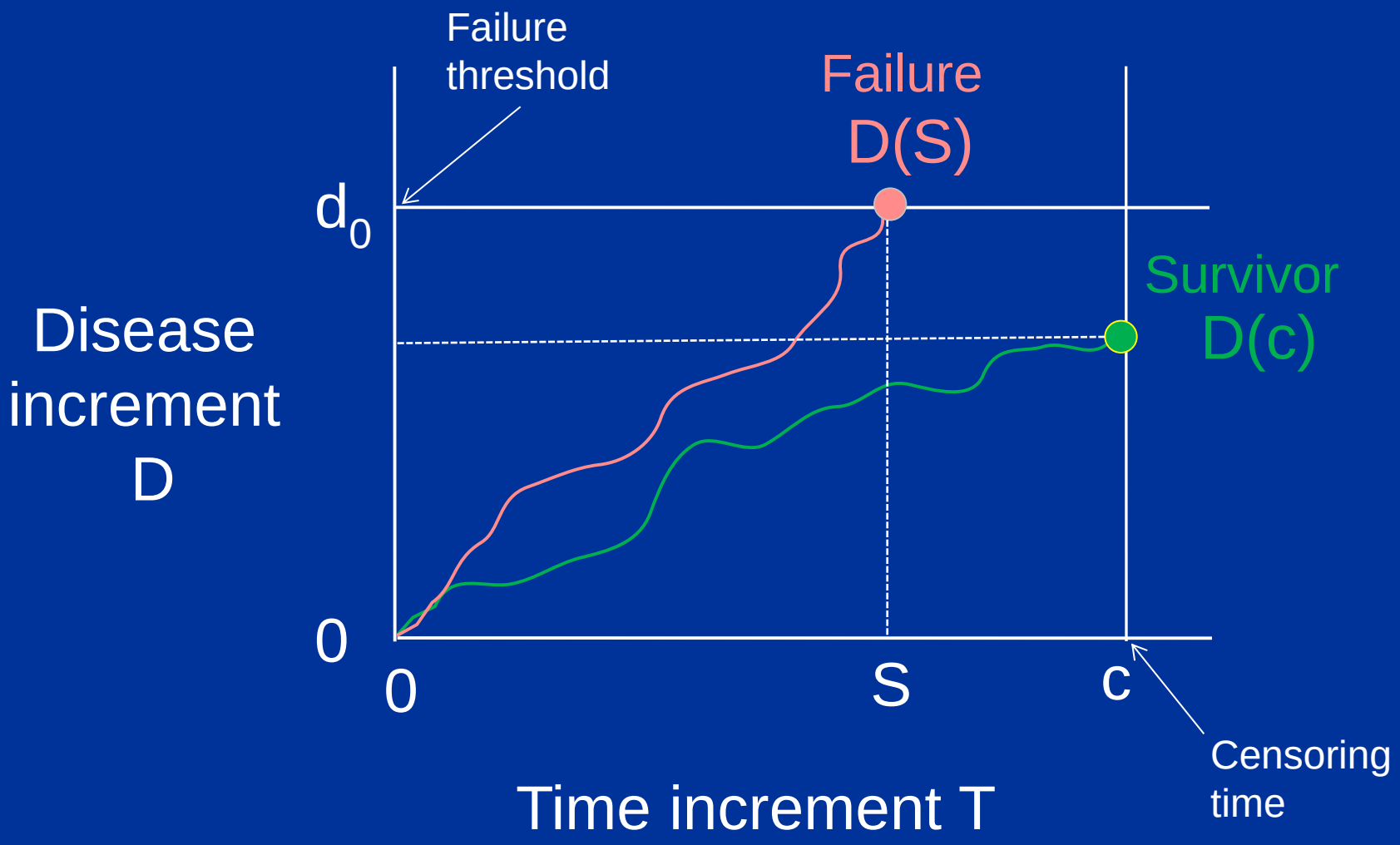
Failure occurs when D(t) first reaches a threshold at d_0.

First hitting time S denotes the survival time.

Survival time may be censored, say, at time c.

Disease level of a survivor at censoring time is D(c).

Revisiting our picture!



Dual Data Structure

The outcome for any subject is a random pair (D, T), where D denotes a disease increment and T denotes the matching time increment. The preceding figure gives a visual representation. Pair (D, T) possesses the following dual data structure:

$$(D,T) = \begin{cases} [D(S), S] & \text{if } S \leq c \\ [D(c), c] & \text{if } S > c \end{cases}$$

Our main assumption: Disease process D(t) has stationary independent increments.

The Fundamental Identity

Assume the cumulant generating function of process D(t) exists in an open interval:

$$\kappa(\zeta) = \ln E\{\exp[\zeta D(1)]\} \text{ exists for all } \zeta \in \mathcal{Z}$$

Our fundamental identity: The following variant of Wald's identity holds for the pair (D, T):

$$E\{\exp[\zeta D - T\kappa(\zeta)]\} = 1 \quad \text{for all } \zeta \in \mathcal{Z}$$

Estimating Equations

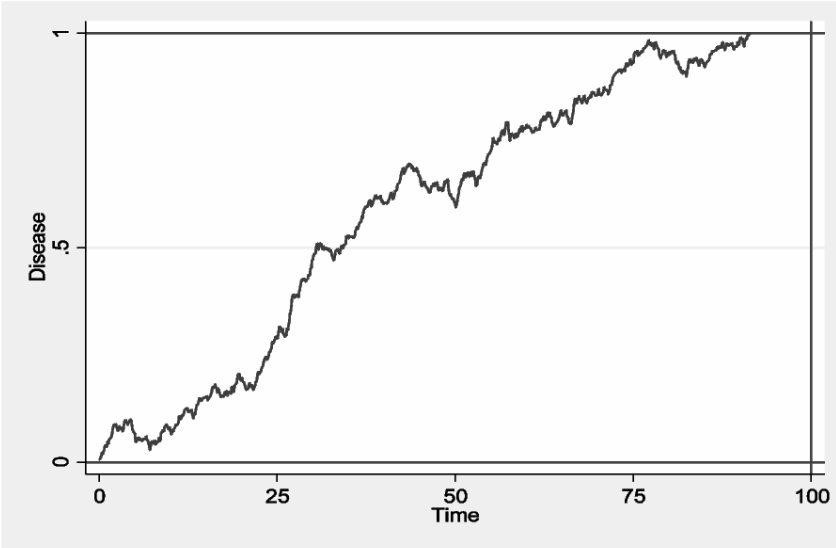
The fundamental identity yields the following two estimating equations for the mean δ and variance σ^2 of disease process D(t):

$$(a) \quad E(D - \delta T) = 0$$
$$(b) \quad E\left[(D - \delta T)^2 - \sigma^2 T\right] = 0$$

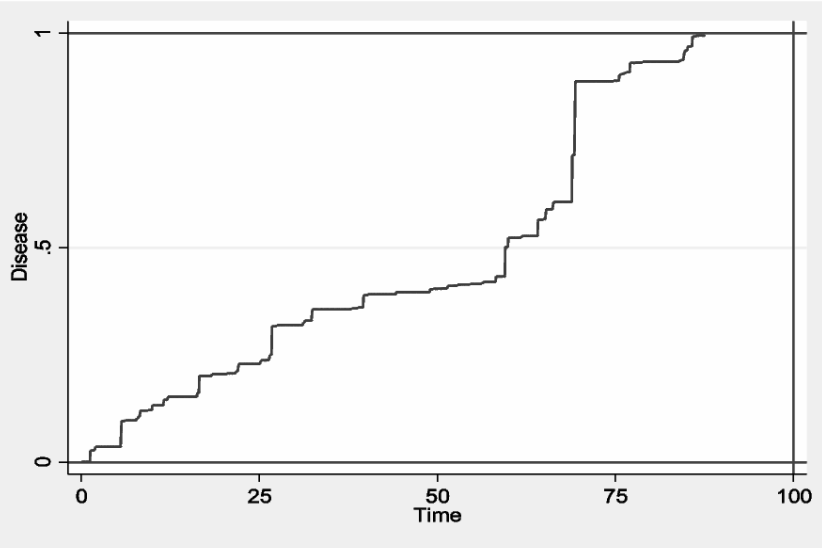
A Variety of Sample Paths

Wiener process
(continuous bi-directional sample path)

Gamma process
(discontinuous monotonic sample path)



(a)



(b)

Illustrative Processes

The preceding graphs illustrate sample paths for two families of processes: (1) a Wiener diffusion process and (2) a gamma process.

Wiener processes have continuous bi-directional sample paths (normally distributed increments)

Gamma processes have discontinuous monotonic sample paths comprised of steps of random size at random moments (gamma-distributed increments).

Wide Class of Candidate Processes

Stochastic processes possessing stationary independent increments define the class of Lévy processes. Families in the class include, for example:

Wiener processes
Gamma processes
Inverse Gaussian processes
Poisson and compound Poisson processes
Negative binomial processes

The family of Lévy processes is closed under addition.

Continuous/Discontinuous Sample Paths

Only Wiener processes in the Lévy class have continuous sample paths; all others have discontinuous sample paths.

Discontinuous sample paths overshoot the failure threshold so $D(S)>d_0$.

Time scale transformations can sometimes convert a non-stationary process into one that is stationary.

Outcome Data and Covariates

Consider n subjects having independent disease processes but possibly different censoring times and thresholds

Parameters may depend on baseline covariates $\mathbf{z}$.

A pair of disease and time increments (d_i, t_i) is observed for every subject i, $i=1,2,\ldots,n$. Some of these pairs correspond to failures and others to survivors.

Regression Link Functions

For covariate vector **z** and regression coefficient vectors **α** and **β**, we have the following link functions:

$$\delta = A(\mathbf{z}\boldsymbol{\alpha}), \quad \sigma = \sigma_0 \lambda \quad \text{where } \lambda = B(\mathbf{z}\boldsymbol{\beta})$$

Note that the process standard deviation is given a product form.

Estimation

Our estimation equations give us the following simple quadratic function to minimize with respect to the regression coefficient vectors α and β:

$$\min \sum_{i=1}^{n} \frac{(d_i - \delta_i t_i)^2}{\lambda_i^2 t_i}$$

Prediction

Predictive inferences can be made using the following pivotal sample residuals w_i for subjects $i=1,2,\ldots,n$:

$$w_i = \frac{d_i - \hat{\delta}_i t_i}{\hat{\sigma}_i t_i^{1/2}}$$

Surrogate Disease Process

Where disease process D(t) is latent (unobservable) then we might construct a surrogate disease process from a vector **Z**(t) of disease-related covariate processes as follows:

$$D(t) = Y(0) - Y(t) \quad \text{where } \ln Y(t) = \gamma_0 + \mathbf{Z}(t)\boldsymbol{\gamma}$$

The same quadratic objective function presented earlier can be minimized for estimates of coefficients γ_0 and **γ**.

Process Y(t) denotes a health process with a sample path that is a mirror image of disease process D(t), reflected about initial health level y_0=Y(0).

Surrogate Disease Process (cont.)

Mean values of the disease process δ and health process μ have opposite signs ($\mu = -\delta$) but the same variance σ^2.

A subsequent case example demonstrates the construction of a surrogate disease process.

Simulation Experiment

We have simulated sample paths from Wiener and gamma processes and estimated the mean and variance parameters using exact maximum likelihood methods and the distribution-free method.

The distribution-free estimates closely match their maximum likelihood counterparts for both small (n=100) and large (n=1000) samples.

See the following table for details.

Estimation Method	Estimated Parameter δ	$\ln(\sigma^2)$
(a)		
Wiener process	Simulation of $n = 100$ sample paths	
Exact Maximum Likelihood	0.010138	-7.780
Distribution-free	0.010158	-7.762
Wiener process	Simulation of $n = 1000$ sample paths	
Exact Maximum Likelihood	0.010009	-7.819
Distribution-free	0.010031	-7.791
(b)		
Gamma process	Simulation of $n = 100$ sample paths	
Exact Maximum Likelihood	0.010082	-7.838
Distribution-free	0.010072	-7.826
Gamma process	Simulation of $n = 1000$ sample paths	
Exact Maximum Likelihood	0.009988	-7.870
Distribution-free	0.010006	-7.847

Table 1: Comparison of estimates from an exact maximum likelihood approach and the proposed distribution-free method, based on $n = 100$ and $n = 1000$ simulated sample paths. Panel (a) has simulated sample paths from a Wiener process. Panel (b) has simulated sample paths for a gamma process. Exact parameter estimates are $\delta = .010$ and $\ln(\sigma^2) = \ln(0.020^2) = -7.824$ in all cases. The threshold d_0 is set at 1.

Cystic Fibrosis: Two-part Case Example

Cystic fibrosis (CF) is an inherited disorder that causes severe damage to the lungs, digestive system and other organs in the body.

Cystic fibrosis affects the cells that produce mucus, sweat and digestive juices.

A defective gene causes the secretions to become sticky and thick. The secretions plug up tubes, ducts and passageways, especially in the lungs and pancreas.[1]

CF prevalence (live births per case)[2]:
Norway 1:4500, Canada 1:2500, Ireland 1:1400

Sources:
1. https://www.mayoclinic.org/diseases-conditions/cystic-fibrosis/symptoms-causes/syc-20353700
2. https://www.cftrscience.com/?q=epidemiology

Case I: Severely Compromised Lung Function

We use hypothetical CF patient registry data to protect confidentiality.

Event of interest: First occurrence of severely compromised lung function in an adult CF patient:

Percent predicted FEV1 < 30 percent

850 adult patients
Initial visit plus 5088 follow-up (roughly annual) visits
235 longitudinal records end with severely compromised lung function (failures)

Records censored by end of record, death or lung/organ transplant

The Variables

Administrative variables:
patient number
visit number

Covariates:
age
body mass index
pancreatic insufficiency (yes=1, no=0)
percent predicted FEV1

Outcome variable (severely compromised lung function):
low (yes=1, no=0)

A fragment of the data and a few summary statistics appear in the next two slides

Fragment of the Data Set

(a)										
patient	visit	age0	bmi0	panc0	fev0	agefwd	bmifwd	pancfwd	fevfwd	low
...	...	...	...	...	...	...	...	...	...	...
34	18	36.6	21.3	1	39.7	37.5	21.3	1	41.7	0
34	19	37.5	21.3	1	41.7	38.4	21.8	1	43.3	0
35	1	20.3	21.2	1	90.7	22.5	19.8	1	62.7	0
35	2	22.5	19.8	1	62.7	23.6	20.6	1	87.2	0
35	3	23.6	20.6	1	87.2	24.3	19.4	1	53.6	0
35	4	24.3	19.4	1	53.6	25.4	17.3	1	27.2	1
36	1	20.9	18.5	1	77.9	21.9	19.6	1	53.5	0
36	2	21.9	19.6	1	53.5	23.0	19.4	1	80.6	0
36	3	23.0	19.4	1	80.6	23.8	18.3	1	73.2	0
36	4	23.8	18.3	1	73.2	25.0	20.0	1	73.6	0
...	...	...	...	...	...	...	...	...	...	...

d = fev0 - fevfwd, t = agefwd - age0, d0 = fev0 - 30

Some Summary Statistics

(b)					
Variable	Sample Size	Mean	Std. Dev.	Min.	Max.
age0	5088	25.4	4.31	20.0	39.5
bmi0	5088	21.8	3.09	12.6	39.1
panc0	5088	.924	.266	0	1
fev0	5088	65.3	20.55	7.0	143.7

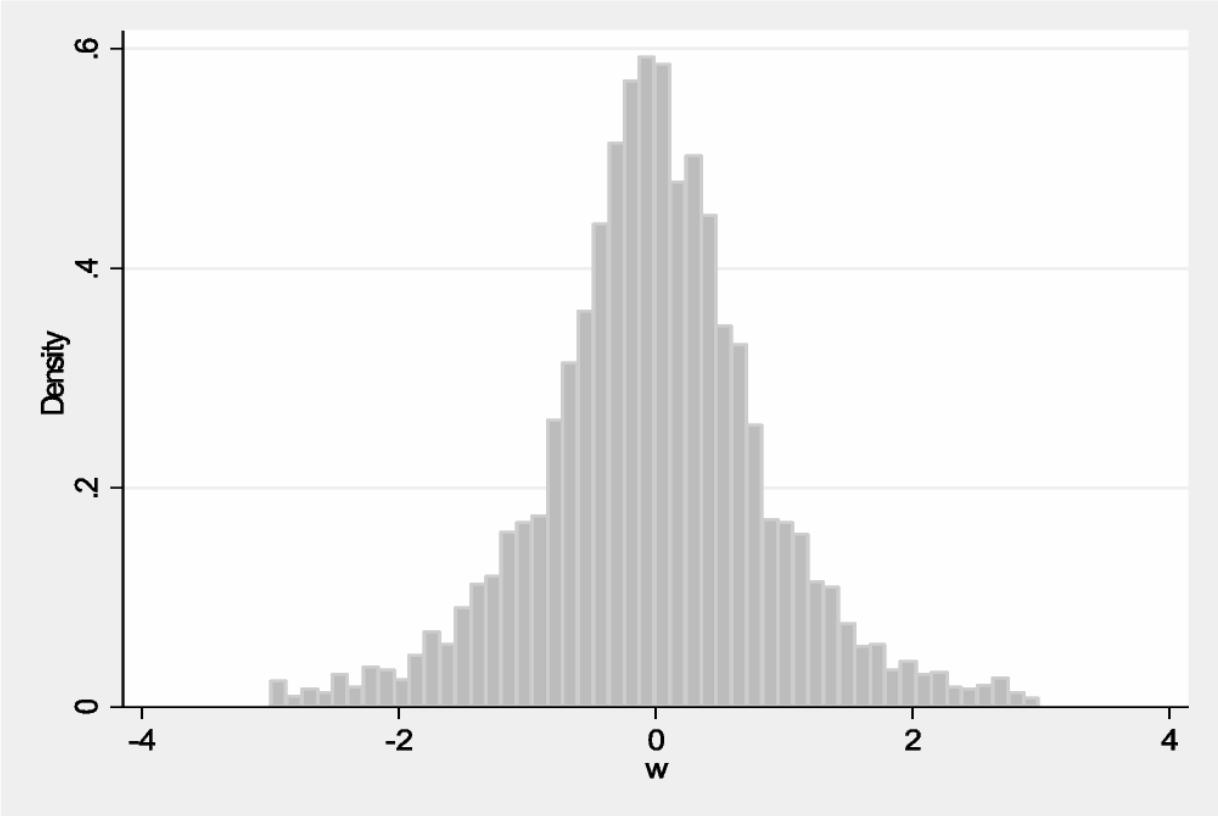
Threshold Regression Results

Threshold regression results, with standard errors and confidence intervals obtained by jackknife estimation.

Parameter	Variable	Mean Est.	Std. Error	90% Conf. Limits	
δ	Mean rate of disease progression				
α_0	intercept	3.2048	1.4005	.9008	5.5088
α_1	age0	-.1226	.0343	-.1790	-.0663
α_2	bmi0	.0478	.0502	-.0347	.1304
α_3	panc0	1.0700	.5021	.2440	1.8960
$\ln(\sigma^2)$	Log-variance of disease progression				
	intercept	4.7272	.0392	4.6627	4.7917

Pivotal Sample Residuals

Histogram of pivotal sample residuals (graph truncates 2% of residuals at abs(w)<3)



Prediction Illustration

Prediction of time to severely compromised lung function for patient #34 at visit 19:

age=38.4, bmi=21.8, panc=1, fev=43.3

Simulation of sample paths for the patient using bootstrapped pivotal sample residuals w_k and time increments h:

$$\begin{aligned} \text{sum}_k &= \text{sum}_{k-1} + \hat{\delta} h + w_k \hat{\sigma} \sqrt{h} \\ \tau_k &= \tau_{k-1} + h \end{aligned}$$

Prediction Illustration (cont.)

Estimated time (in years) to severely compromised lung function for patient #34 who is currently 38 years old with a percent predicted FEV1 of 43.3%:

25^{th} percentile:	1.3
50^{th} percentile:	3.7
75^{th} percentile:	13.5

The estimated residual event time distribution is extremely right skewed here.

The 75^{th} percentile of 13.5 years extrapolates a decade beyond the age range of patients in the data set so the prediction must be considered with caution.

Case II: Cystic fibrosis surrogate health process and risk of death

Construct a surrogate health process $Y(t)$ using covariate processes $\mathbf{Z}(t)$ consisting of: age, bmi, panc and fev.

850 adult patients
Initial visit plus 5637 follow-up (roughly annual) visits
109 longitudinal records end with death (failures)

Records censored by end of record or lung/organ transplant

Estimated Surrogate Health Process

Coefficient estimates for the surrogate log-health process ln Y(t). Standard errors and confidence limits obtained by jackknife estimation.

Parameter	Variable	Mean Est.	Std. Error	90% Conf. Limits	
$\ln Y(t)$	Log-health process				
γ_0	intercept	6.5220	.3588	5.9311	7.1128
γ_1	age0	-.2929	.0192	-.3245	-.2613
γ_2	bmi0	.0210	.0084	.0073	.0348
γ_3	panc0	-.0152	.1582	-.2757	.2453
γ_4	fev0	.0056	.0012	.0036	.0075
$\ln(\sigma^2)$	Log-variance of disease progression				
	intercept	-.4363	.0578	-.5315	-.3410