Exploratory subgroup identification in the heterogeneous Cox model

A relatively simple approach

Larry F. León
Merck, Statistical Methodology Research

Introduction

- Goal of identifying an existing subgroup H consisting of subjects who derive the least benefit from treatment
  - If treatment is detrimental $\rightarrow$ possible action; "Lack of benefit, or mild benefit" may not be reason enough to "recommend NOT-to-treat" or to exclude from inclusion in future program development
- In case of detrimental H, the complementary subgroup H^c may be considered to derive benefit with a "higher degree of confidence" relative to the ITT population
- Our approach is based on the idea of all-possible subsets regression from the area of model selection:
  - For K factors, $X_1, X_2, \ldots, X_k$, one fits all possible model combinations and chooses the model which minimizes a fit/penalty criteria (e.g., AIC/BIC)

Overview of Forest Search

We extend to all-possible subgroups:

- X_1=Sex (M,F), X_2=Age ($A \leq 50$, $A > 50$), and X_3=Age ($A \leq 35$, $A > 35$)
- 8 SG combinations from X_1 and X_2: {M}, {F}, $\{A \leq 50\}$, $\{A > 50\}$, {M $\times$ A$\leq$ 50}, {M $\times$ A$>$ 50}, {F $\times$ A$\leq$ 50}, {F $\times$ A$>$ 50}
- Another 6 from X_1 and X_3
- And from Age intervals such as $\{Age > 35 \times Age \leq 50\}$
- Some are null, $\{A > 50\} \times \{A \leq 35\} = \emptyset$
- This is not an exhaustive list for this example (For L binary factors: $2^{2L} - 1$ possible SG's)
- The number of combinations grows large; We restrict to a minimum SG size (e.g., 60 subjects) and a minimum number of events per treatment arm (e.g., 10 events)

Forest Search identification criteria

For identifying H we define (screen) candidates as SG's with Cox (HR) estimates ≥ 1.5 and employ a "splitting consistency criteria":

- Suppose there are SG's with estimates ≥ 1.5 and for each SG we randomly split (e.g, 500 times) the SG 50/50
- Consider a split "consistent with harm" if **both** estimated HRs are ≥ 1.25 for the two SG splits
- We define H-candidates as those with consistency rates at least 90% (across the 500 splits);
- Define the SG with the highest consistency rate as exhibiting "maximal harm"
- If no SG achieves a consistency rate of at least 90% then consider $H = \emptyset$ (and H^c is the ITT population)
- The consistency criterion heuristically represents – "no matter how you split the SG H, those splits are consistent with harm"

Random splitting: Choice of 1.5 and 1.25 thresholds

- Random splitting 50/50 via random strata $\sim Bin(0.5)$
- Cox score $L(d; strata) = L(d1, strata = 1) + L(d2, strata = 2)$
- $\hat{\beta}_s$, $\hat{\beta}_{s1}$, $\hat{\beta}_{s2}$ denote the corresponding estimators
- Un-stratified $\hat{\beta} \approx \hat{\beta}_s$ since strata are purely random
- Use approximation (Jennison and Turnbull (1984)):

$$\hat{\beta} \approx 4L(d)/d \approx N(\beta, 4/d)$$

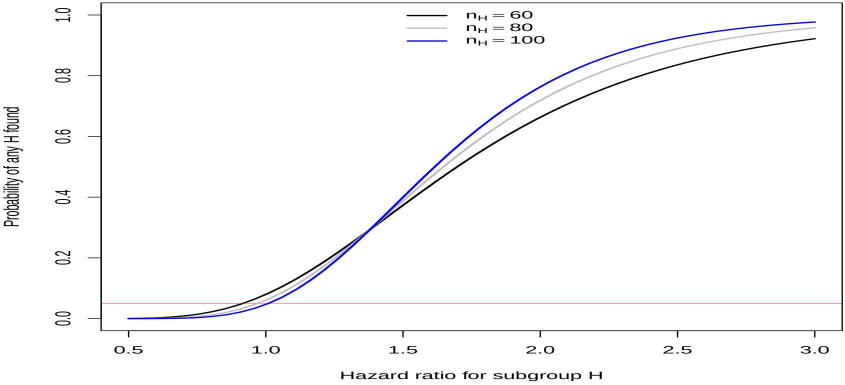
- $\hat{\beta}_{s1} \perp\!\!\!\perp \hat{\beta}_{s2} \approx 8L(d_1, s = 1)/d \approx N(\beta, 8/d)$
- $L(d) \simeq L(d; strata) = L(d1, s = 1) + L(d2, s = 2)$:

$$\hat{\beta} \geq log(1.5) \iff \hat{\beta}_{s1} + \hat{\beta}_{s2} \geq 2 * log(1.5)$$

- Numerical integration: $(\{W_1, W_2\} \sim N(\beta, 8/d)$, independently) $a(\beta) = P(W_1 + W_2 \geq 2 * \log(1.5), \min(W_1, W_2) \geq \log(1.25)) =$
-

$$\int I(w_1 + w_2 \geq a)I(w_1 \geq b)I(w_2 \geq b)\varphi(w_1; \beta, 8/d)\varphi(w_2; \beta, 8/d)dw_1dw_2$$

Approximate Prob of finding H via Forest Search: Subgroup H of size n_H exists with hr from 0.5 to 3.0 $(n_H = 60, 80, 100;\ pC = 45\%;\ d \approx (1 - pC) * n)$



Simulations: GBSG data, Schumacher et al. (1994), $\approx$ 56% censoring [$\bar{C} = 46\%$ in simulations]

- Simulations based on German Breast Cancer Study (n=686, 7 baseline factors)
- $\log(T) = \mu + \beta_0 \text{Treat} + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + \beta_5 Z_5 + \gamma \text{Treat} Z_1 Z_4$,
- Z_1=Estrogen, Z_2=Age, Z_3=Progesterone, Z_4=Menopausal, Z_5=Positive nodes
- Size of harm subgroup H= $\{Z_1 = 1\} \cap \{Z_4 = 1\}$ is $n = 84$
- Size of non-harm subgroup H^c= $\{Z_1 = 0\} \cup \{Z_4 = 0\}$ is $n = 602$
- The choice of β_0 and γ determining the treatment effects; $\gamma = 0$ generates no subgroup ($H = \emptyset$)
- Analyst has available factors Z_1-Z_5 plus 2 additional "noise" factors Z_6=Size and Z_7=Grade
- Null model: $H = \emptyset$, H^c is ITT population, hr =0.63
- Alt model: hr(H) =2.5, hr(H^c) =0.57

VT (Foster et al., 2011); GRF (Athey et al., 2019; Athey and Wager, 2021)

Restrict to SG's with at least 60 subjects; VT and GRF maximum tree depth of 3 baseline factors; For VT, employ "censoring unbiased transformation" of Fan and Gijbels, 2018

- GRF targets RMST and we denote GRF as RMST based on the truncation point $\tau = \min(\tau_0, \tau_1)$
- An RMST benefit of 6 months for control is required for selection of an H
- GRF employs a double-robust (IPCW) approach for RMST $\rightarrow$ sensitive to τ. GRF.70: $\tau_{70} := 0.7 \min(\tau_0, \tau_1)$
- VT(24) targeting survival rates at $t = 24$ months. A treatment effect of $\delta = 0.25$, in favor of control, is required for selection of an H
- $VT^{\#}(24)$ survival rates at $t = 24$ months based on non-censored (latent) outcomes. Remove the challenge of censoring and base analyses on the (ideal) latent outcomes
- VT(36) same as VT(24) but with $t = 36$
- $VT^{\#}(36)$ same as $VT^{\#}(24)$ but with $t = 36$

% any H: $\bar{n}_H = 56[51, 61]; 78[73, 84]; 112[105, 119]$

$hr(H)$	FS	GRF	GRF.70	VT(24)	$VT^{\#}(24)$	VT(36)	$VT^{\#}(36)$
N=500							
$\emptyset$	0.03	0.26	0.10	0.05	0.04	0.06	0.05
1	0.07	0.26	0.12	0.09	0.08	0.09	0.09
1.5	0.20	0.40	0.24	0.19	0.20	0.19	0.21
2	0.37	0.51	0.35	0.32	0.35	0.27	0.34
2.5	0.51	0.60	0.46	0.46	0.49	0.36	0.42
3	0.62	0.64	0.54	0.55	0.59	0.42	0.49
N=700							
$\emptyset$	0.04	0.23	0.08	0.05	0.04	0.05	0.04
1	0.10	0.28	0.13	0.10	0.08	0.09	0.10
1.5	0.39	0.55	0.36	0.26	0.28	0.25	0.31
2	0.71	0.75	0.59	0.49	0.52	0.42	0.54
2.5	0.89	0.89	0.77	0.68	0.72	0.58	0.69
3	0.95	0.94	0.87	0.81	0.86	0.67	0.79
N=1000							
$\emptyset$	0.05	0.20	0.07	0.03	0.03	0.03	0.03
1	0.16	0.29	0.15	0.07	0.07	0.09	0.09
1.5	0.59	0.63	0.45	0.27	0.30	0.29	0.38
2	0.88	0.87	0.74	0.53	0.57	0.54	0.66
2.5	0.98	0.95	0.89	0.73	0.78	0.72	0.83
3	1.00	0.98	0.96	0.85	0.89	0.82	0.90

Classification metrics: H$= \{Z_1 = 1\} \cap \{Z_4 = 1\}$, $H^c = \{Z_1 = 0\} \cup \{Z_4 = 0\}$

- Note that there always exists $\hat{H}^c$; $\hat{H} = \emptyset$ implies $\hat{H}^c = \Omega$
- $ppv(\hat{H})$:
$$\#\{i \in \hat{H} \cap H\} / \#\{i \in H\}$$
- $ppv(\hat{H}^C)$:
$$\#\{i \in \hat{H}^c \cap H^c\} / \#\{i \in H^c\}$$
- $sens(\hat{H})$:
$$\#\{i \in \hat{H} \cap H\} / \#\{i \in \hat{H}\}$$
- $sens(\hat{H}^C)$:
$$\#\{i \in \hat{H}^c \cap H^c\} / \#\{i \in \hat{H}^c\}$$
- $sens(\hat{H}^C | \hat{H} \neq \emptyset)$:
$$\#\{i \in \hat{H}^c \cap H^c, \hat{H} \neq \emptyset\} / \#\{i \in \hat{H}^c, \hat{H} \neq \emptyset\}$$

Classification rates under the null $N = 700$

	FS	GRF	GRF.70	VT(24)	$VT^{\#}(24)$	VT(36)	$VT^{\#}(36)$
Finding H							
any(H)	0.04	0.23	0.08	0.05	0.04	0.05	0.04
$ppv(\hat{H})$	.	.	.	.	.	.	.
$ppv(\hat{H}^C)$	1	0.97	0.99	0.99	1	0.99	1
$sens(\hat{H})$	0	0	0	0	0	0	0
$sens(\hat{H}^C)$	1	1	1	1	1	1	1
$sens(\hat{H}^C \mid \hat{H} \neq \emptyset)$	0.89	0.88	0.89	0.89	0.89	0.89	0.89
Size of H and Hc							
$avg\lvert\hat{H}\rvert$	76	85	79	75	74	78	77
$min\lvert\hat{H}\rvert$	61	60	60	60	60	60	60
$max\lvert\hat{H}\rvert$	206	222	197	166	136	161	170
$avg\lvert\hat{H}^C\rvert$	697	681	694	696	697	696	697
$min\lvert\hat{H}^C\rvert$	494	478	503	534	564	539	530
$max\lvert\hat{H}^C\rvert$	700	700	700	700	700	700	700

Classification rates under $hr(H) = 2.5$ $(N = 700)$; $\bar{n}_H = 78[0.25 = 73, 0.75 = 84, min = 52, max = 106]$

	FS	GRF	GRF.70	VT(24)	$VT^{\#}(24)$	VT(36)	$VT^{\#}(36)$
Finding H							
any(H)	0.89	0.89	0.77	0.68	0.72	0.58	0.69
$ppv(\hat{H})$	0.74	0.71	0.63	0.61	0.67	0.48	0.61
$ppv(\hat{H}^C)$	0.99	0.97	0.98	0.98	0.98	0.98	0.98
$sens(\hat{H})$	0.77	0.69	0.62	0.57	0.61	0.47	0.58
$sens(\hat{H}^C)$	0.97	0.97	0.96	0.96	0.96	0.94	0.96
$sens(\hat{H}^C \mid \hat{H} \neq \emptyset)$	0.99	0.97	0.97	0.97	0.97	0.97	0.97
Size of H and Hc							
$avg\lvert\hat{H}\rvert$	74	83	83	87	90	83	86
$min\lvert\hat{H}\rvert$	61	60	60	60	60	60	60
$max\lvert\hat{H}\rvert$	138	185	233	175	180	185	174
$avg\lvert\hat{H}^C\rvert$	634	626	636	641	635	652	640
$min\lvert\hat{H}^C\rvert$	562	515	467	525	520	515	526
$max\lvert\hat{H}^C\rvert$	700	700	700	700	700	700	700

Estimation and Bootstrap Bias Corrected Estimates

For the observed data, with estimated SG $\hat{H}$, and boostrap data with estimated $\hat{H}^*$ we have:

1. $\hat{H} \rightarrow \hat{\theta}(\hat{H})$
2. $\hat{H}_b^* \rightarrow \hat{\theta}_b^*(\hat{H}_b^*)$
3. $\hat{H}_b^* \rightarrow \hat{\theta}(\hat{H}_b^*)$
4. $b_1^*(\hat{H}_b^*) = \hat{\theta}_b^*(\hat{H}_b^*) - \hat{\theta}(\hat{H}_b^*)$
5. $\hat{\theta}_1^*(\hat{H}) = \hat{\theta}(\hat{H}) - (1/B)\sum_{b=1}^{B} b_1^*(\hat{H}_b^*)$
6. $b_2^*(\hat{H}) = \hat{\theta}_b^*(\hat{H}) - \hat{\theta}(\hat{H})$
7. $\hat{\theta}_2^*(\hat{H}) = \hat{\theta}(\hat{H}) - (1/B)\sum_{b=1}^{B} \left\{ b_1^*(\hat{H}_b^*) + b_2^*(\hat{H}) \right\}$

The bias corrected estimator $\hat{\theta}_1^*(\hat{H})$ is motivated by Harrell Jr. et al. (1996).

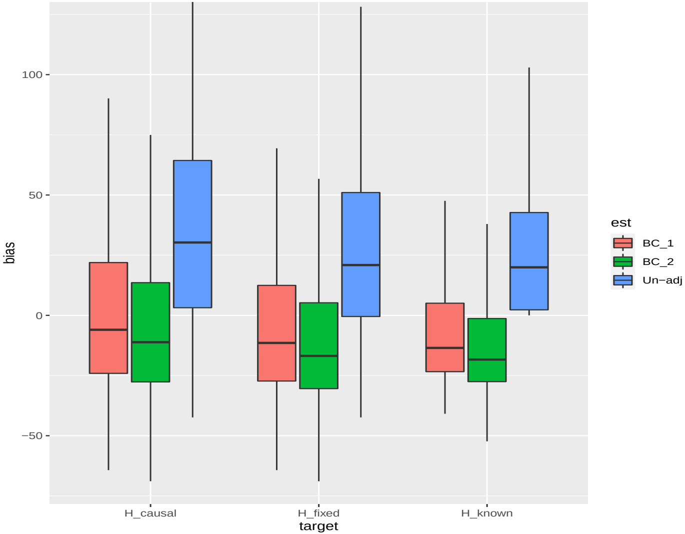
Variance estimates generally require double-bootstrapping and are approximated by the (Infinitesimal) Jacknife (Wager et al., 2014; Rosenkranz, 2016).

Bias-correction $\theta^{\dagger}(H) = 3.5$; $\theta^{\dagger}(H^c) = 0.57$ $(N = 1000)$

	Avg	median	SD	est(SD)	min	max	$C\{\hat{\theta}(H)\}$	$C\{\theta^{\dagger}(\hat{H})\}$	$C\{\theta^{\dagger}(H)\}$
$\hat{\theta}(H)$	3.58	3.39	1.00	.	1.51	8.66	.	.	.
$\theta^{\dagger}(\hat{H})$	3.33	3.50	0.41	.	1.31	3.50	.	.	.
$\hat{\theta}(\hat{H})$	4.59	4.23	1.69	1.46	2.02	23.91	95.49	85.97	90.88
$\hat{\theta}_1^*(\hat{H})$	3.38	3.10	1.33	0.85	1.25	18.90	89.38	78.26	78.26
$\hat{\theta}_2^*(\hat{H})$	3.16	2.91	1.15	1.64	1.09	14.08	99.90	98.00	97.90

	Avg	median	SD	est(SD)	min	max	$C\{\hat{\theta}(H^c)\}$	$C\{\theta^{\dagger}(\hat{H}^c)\}$	$C\{\theta^{\dagger}(H^c)\}$
$\hat{\theta}(H^c)$	0.69	0.69	0.06	.	0.52	0.99	.	.	.
$\theta^{\dagger}(\hat{H}^c)$	0.71	0.68	0.11	.	0.57	0.98	.	.	.
$\hat{\theta}(\hat{H}^c)$	0.71	0.71	0.07	0.06	0.52	0.99	100	67.64	36.77
$\hat{\theta}_1^*(\hat{H}^c)$	0.72	0.71	0.07	0.06	0.53	1.00	100	64.23	30.16
$\hat{\theta}_2^*(\hat{H}^c)$	0.72	0.71	0.07	0.13	0.53	1.00	100	97.09	88.68

Relative bias for boostrap bias-corrected estimators

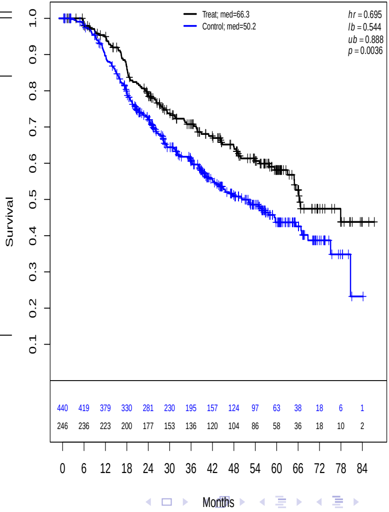


GBSG Dataset

Table: GBSG Baseline Summary

	0	1
	$N = 440$	$N = 246$
age	45 50 59	50 58 63
meno	48% (209)	76% (187)
grade : 1	11% (48)	13% (33)
2	64% (281)	66% (163)
3	25% (111)	20% (50)
size	20 25 35	20 25 35
nodes	1 3 7	1 3 7
pgr	7.0 32.0 130.0	7.2 35.0 133.0
er	8 32 92	9 46 182

a b c represent the lower quartile a, the median b, and the upper quartile c for continuous variables. Numbers after percents are frequencies.

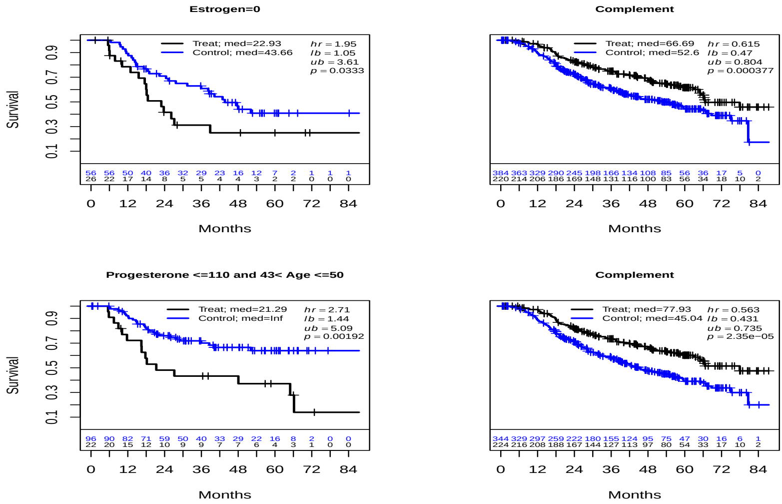


GBSG Analysis

1. GRF trees (depths 1,2, and 3) split on: Age at 33, 43, 48, and 50; Progesterone at 8, and 74; Estrogen at 0, and 103
2. (V_1) 3 binary factors for Estrogen based on cuts ($\leq$) at 0, 103, and median; (V_2) Progesterone at 8, 74, 110, and 132; (V_3) Age at 50, 33, and 43; (V_4) Menopausal, (V_5) Nodes, and (V_6) Size at medians; (V_7) Grade 1 vs 2/3 and Grade 3 vs 1/2

p.consistency	M.1	M.2	M.3	M.4	K	n	E	d1	m1	m0	HR
0.981	v3a.1	v3c.0	v2c.1		3	118	43	15	21.290	Inf	2.711
0.974	v3a.1	v2b.1	v3c.0		3	102	39	13	17.741	Inf	2.899
0.965	v1a.0	v3a.1	v3c.0	v2c.1	4	91	29	12	21.290	Inf	2.867
0.961	v3a.1	v7b.0	v3c.0	v2c.1	4	86	30	11	27.170	Inf	2.847
0.96	v3a.1	v2b.1	v1b.1	v3c.0	4	94	37	12	27.170	Inf	2.613
0.956	v3a.1	v1b.1	v3c.0	v2d.1	4	113	43	15	27.170	Inf	2.466
0.953	v1a.0	v3a.1	v3c.0	v2d.1	4	98	31	13	48.066	Inf	2.593
0.948	v3a.1	v3c.0	v2d.1		3	125	45	16	37.651	Inf	2.436
0.944	v1a.0	v3a.1	v2b.1	v3c.0	4	76	26	10	17.741	Inf	2.949
0.938	v3a.1	v1b.1	v3c.0	v2c.1	4	109	41	14	27.170	Inf	2.433
0.919	v3a.1	v7b.0	v3c.0	v2d.1	4	93	32	12	37.651	Inf	2.514

GBSG: $\hat{\theta}_2^*(\hat{H})$: 1.47[0.68, 3.19]; $\hat{\theta}_2^*(\hat{H}^c)$: 0.6[0.37, 0.99]



References I

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.

Fan, J. and Gijbels, I. (2018). *Local polynomial modelling and its applications*.

Foster, J. C., Taylor, J. M., and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30(24):2867–2880.

Harrell Jr., F. E., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387.

Jennison, C. and Turnbull, B. W. (1984). Repeated confidence intervals for group sequential clinical trials. *Controlled Clinical Trials*, 5(1):33–45.

References II

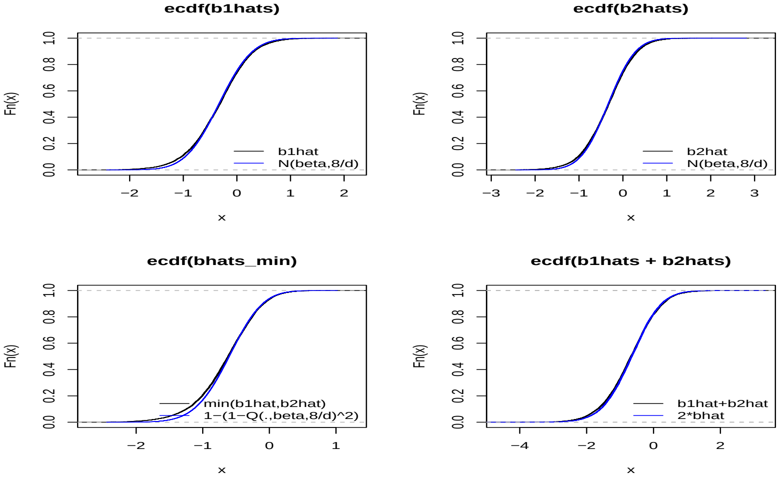
Rosenkranz, G. K. (2016). Exploratory subgroup analysis in clinical trials by model selection. *Biometrical Journal*, 58(5):1217–1228.

Schumacher, M., Bastert, G., Bojar, H., Hübner, K., Olschewski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., Neumann, R. L., and Rauschecker, H. F. (1994). Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group. *Journal of Clinical Oncology*, 12(10):2086–2093.

Wager, S., Hastie, T., and Efron, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. 15:1625–1651.

Back-up Slides

Approximations of Cox estimates via $N(\beta, 4/d)$



Fan and Gijbel's Censoring Unbiased Transformation (CUT)

Transforming the observed survival times:

$$E(\varphi(Y_i)|L_i, Z_i) = E(T_i|L_i, Z_i)$$

$$\varphi(Y_i) = \begin{cases} \varphi_1(T_i), & \text{if uncensored,} \\ \varphi_2(C_i), & \text{if censored.} \end{cases}$$

$$\varphi_1(T_i) = (1+\theta)\int_0^{T_i} \frac{dt}{G(t)} - \theta \frac{T_i}{G(T_i)}$$

$$\varphi_2(C_i) = (1+\theta)\int_0^{C_i} \frac{dt}{G(t)}$$

$$\theta = \min_{\{i:\delta_i=1\}} \frac{\int_0^{Y_i} \frac{dt}{G(t)} - Y_i}{\frac{Y_i}{G(Y_i)} - \int_0^{Y_i} \frac{dt}{G(t)}}$$