Results

2020-03-23

Model setting

Cox PH model

Suppose the event time T and the censoring time C are both following cox models, who are sharing the same $S_0(t)$ function, i.e.

$$\text{event time: } S_T(t|X=x) = P(T > t|X=x) = S_0(t)^{\exp(\beta' x)}$$

$$\text{censoring time: } S_C(t|X=x) = P(C > t|X=x) = S_0(t)^{\exp(\gamma' x)}$$

where X is the covariates vector and $\beta = (\beta_1, \beta_2, ..., \beta_p)$ are the coefficients for cox PH model in terms event time, $\gamma = (\gamma_1, \gamma_2, ..., \gamma_p)$ are the coefficients for cox PH model in terms of censoring time.

Therefore, the associated hazard functions are

$$\text{event time: } \lambda_T(t|x) = \lambda_0(t)\exp(\beta' x),\ \Lambda_T(t|x) = \Lambda_0(t)\exp(\beta' x)$$

$$\text{censoring time: } \lambda_C(t|x) = \lambda_0(t)\exp(\gamma' x),\ \Lambda_C(t|x) = \Lambda_0(t)\exp(\gamma' x)$$

The associated $m()$ function can be

$$m(t,x) = \frac{\lambda_T(t|x)}{\lambda_T(t|x) + \lambda_C(t|x)} = \frac{\lambda_0(t)\exp(\beta' x)}{\lambda_0(t)\exp(\beta' x) + \lambda_0(t)\exp(\gamma' x)} = \frac{1}{1 + \exp(-(\beta - \gamma)' x)}$$

which follows a logistic distribution.

Let's just consider a two dimension simple scenario, where $\lambda_0(t) = 1,\ \Lambda_0(t) = t, S_0(t) = \exp(-t)$

- Event time:

$$\text{hazard function: } \lambda_T(t|x) = \lambda_0(t)\exp(\beta_1 x_1 + \beta_2 x_2) = \exp(\beta_1 x_1 + \beta_2 x_2)$$

$$\text{cumulative hazard function: } \Lambda_T(t|x) = \Lambda_0(t)\exp(\beta_1 x_1 + \beta_2 x_2) = t\exp(\beta_1 x_1 + \beta_2 x_2)$$

$$\text{survival function: } S(t|x) = S_0(t)^{\exp(\beta_1 x_1 + \beta_2 x_2)} = \exp(-t \times (\beta_1 x_1 + \beta_2 x_2))$$

- Censoring time

$$\text{hazard function: } \lambda_C(t|x) = \lambda_0(t)\exp(\gamma_1 x_1 + \gamma_2 x_2) = \exp(\gamma_1 x_1 + \gamma_2 x_2)$$

$$\text{cumulative hazard function: } \Lambda_C(t|x) = \Lambda_0(t)\exp(\gamma_1 x_1 + \gamma_2 x_2) = t\exp(\gamma_1 x_1 + \gamma_2 x_2)$$

$$\text{survival function: } S(t|x) = S_0(t)^{\exp(\gamma_1 x_1 + \gamma_2 x_2)} = \exp(-t \times (\gamma_1 x_1 + \gamma_2 x_2))$$

We will check whether:

1. X_1, X_2 have effect on the survival time $S(t)$, i.e. the coefficients are significant in the Cox PH model fitted with X_1, X_2
2. X_1, X_2 have effect on the censoring time, i.e. the coefficients are significant by fitting the logistic regression with status $\sim X_1, X_2$
3. When the Cox PH model is mis-specified, with only X_1 or X_2, how well the survival time is estimated?
4. The estimation methods:
    - Cox PH model with X_1, X_2
    - Cox PH model with only X_1
    - Cox PH model with only X_2
    - by using the true $m(t, x)$ function (true $\beta_1 - \gamma_1, \beta_2 - \gamma_2$)
    - by using the estimated $\hat{m}(t, x)$ function (estimated from the logistic regression with X_1, X_2)

Simulation steps

Data set generation

- 1. Simulate the covariates for n subjects: $X = (x1, x2) \sim MVN(\mu, \Sigma)$, where X is n by 2 matrix and $x1, x2$ are n by 1 vectors. For each subject, the covariates is x_{i1}, x_{i2}.
    - $\beta = (\beta_1, \beta_2) = (0.2, 0.1), \gamma = (\gamma_1, \gamma_2) = (0.1, 0.2)$
    - $X = (x1, x2) \sim MVN(\mu, \Sigma)$, $\mu = (1, 2)$, $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$.
- 2. For subject i, calculate the βX_i. Simulate the survival time T_i from below CDF by using the inverse probablity sampling.

$$S_{T,i}(t) = S_{0,i}(t)^{\exp(\beta X_i)} = \exp(-t \times \exp(\beta X_i))$$

- 3. Simulate the censoring time C_i from below CDF by using the inverse probablity sampling.

$$S_{C,i}(t) = S_{0,i}(t)^{\exp(\gamma X_i)} = \exp(-t \times \exp(\gamma X_i))$$

- 4. Calculate observed time $Z_i = T_i \wedge C_i$ and indicator $\delta_i = I(T_i < C_i)$.
- Then we get the dataset with death time, censoring time, observed time, censoring indicator, and covariates. The times are calculated one by one based on the subject's covariates values.

We generate a training dataset, with $n = 500$, following the above steps. A test dataset is generated with $n = 100$, following the above procedures.

100 repetitions are conducted.

Models

- Cox PH model with both x1, x2
- Cox PH model with only x1 (mis-specified)
- Cox PH model with only x2 (mis-specified)
- Dikta's method (true m() function)
- Dikta's method (estimated m() function)
- Dikta's method, new formula, follows Stute's paper (true m() function)
- Dikta's method, new formula, follows Stute's paper (estimated m() function)

When there is no covariates are considered in the dataset, the most commonly applied Kaplan Meier prodcut limit estimator is defined as

$$S^{KM}(t) = \prod_{Z_i \leq t} \Big(1 - \frac{\delta_i}{n - R_{i,n} + 1}\Big) \tag{1}$$

where Z_i is the ith ordered observed time, $R_{i,n}$ is the rank of Z_i.

Dikta (1998) proposed another product limited estimator defined as

$$S^{D1} = \prod_{Z_i \leq t} \Big(1 - \frac{1}{n - R_{i,n} + 1}\Big)^{m(Z_{k:n})} \tag{2}$$

He argued that the semiparameter S^{D1} is unbiased and has less variance than Kaplan Meier estimator.

if we consider covariates, could we just replace $m(Z_{k:n})$ with $m(Z_{k:n}, X_{k:n})$ is the above formula ?

The Stute estimator:

$$F_0(x,t) = P(X \leq x, T \leq t) = \sum_{i=1}^{n} W_i \phi(x,t) \tag{3}$$

If we set $\phi(x,t) = I_{(X \leq x, T \leq t)}$, it leads to

$$F_0(x,t) = \sum_{i=1}^{n} W_i I_{(X \leq x, T \leq t)} = \sum_{i=1}^{n} \Big\{ \frac{\delta_i}{n-i+1} \Big[\prod_{j=1}^{i-1} (1 - \frac{1}{n-j+1})^{\delta_i} \Big] \Big\} I_{(X \leq x, T \leq t)} \tag{4}$$

where $W_i = \frac{\delta_i}{n-i+1} \prod_{j=1}^{i-1} (1 - \frac{1}{n-j+1})^{\delta_i}$.

Dikta replaced δ_i with $m(t,x)$ and got

$$S^{D2} = \sum_{i=1}^{n} \Big\{ \frac{m(Z_i, X_i)}{n-i+1} \Big[\prod_{j=1}^{i-1} (1 - \frac{m(Z_i, X_i)}{n-j+1}) \Big] \Big\} I_{(X \le x, T \le t)} \tag{5}$$

Comparsion methods

- bias of quantile times
  - quantile time at $t = 10\%, 25\%, 50\%, 75\%, 90\%$
  - for each subject in the dataset, calculate the true survival function $S_i(t, x_i)$
  - for each subject in the dataset, calculate the estimated survival functions by above methods, $\hat{S}_i(t, x_i)$.

  - calculate the bias: $\hat{S}_i(t, x_i) - S_i(t, x_i)$
  - then calculate the mean bias for the dataset $\frac{1}{n} \sum_{i=1}^{n} (\hat{S}_i(t, x_i) - S_i(t, x_i))$, which is marked as the bias for the dataset, bias_{data}.
  - for the 100 repetitions, get the mean value and standard deviation of the bias for the dataset $\frac{1}{100} \sum_{i=1}^{100} \text{bias}_{data,i}$
- absolute bias of quantile times
  - similar with the above processes but use the absolute bias: $|\hat{S}_i(t, x_i) - S_i(t, x_i)|$
- Time dependent AUC
  - quantile time at $t = 10\%, 25\%, 50\%, 75\%, 90\%$
  - for each subject in the dataset, calculate the estimated survival functions by above methods, $\hat{S}_i(t, x_i)$.

  - lables: whether the subjects is alive at the quantile time $P(T > t)$

Results

The Cox PH model with both x1, x2

```
fitcox = coxph(Surv(time, status) ~ x1 + x2, data=data)
summary(fitcox)$coefficient
```

```
##          coef  exp(coef)   se(coef)        z     Pr(>|z|)
## x1 0.23333395   1.262803 0.06426176 3.630992 0.0002823335
## x2 0.06550981   1.067703 0.06474762 1.011772 0.3116472222
```

The Cox PH model with only x1

```
fit1 = coxph(Surv(time, status) ~ x1, data=data)
summary(fit1)$coefficient
```

```
##         coef  exp(coef)   se(coef)        z     Pr(>|z|)
## x1 0.2412302   1.272814 0.06378269 3.782063 0.0001555339
```

The Cox PH model with only x2

```r
fit2 = coxph(Surv(time, status) ~ x2, data=data)
summary(fit2)$coefficient
```

```
##          coef exp(coef)   se(coef)        z  Pr(>|z|)
## x2 0.0945726  1.099189 0.06441263 1.468231 0.1420416
```

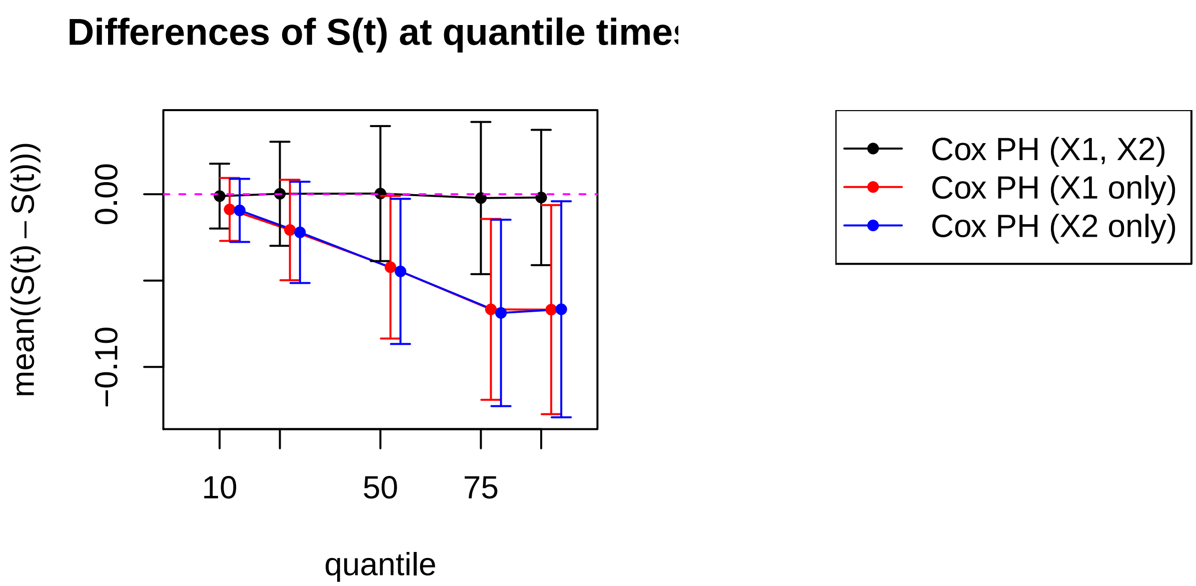
The logistic regression for $m(t,x) = E(\delta|X = x)$ function calculation.

```r
logfit = glm(status ~ x1 + x2 - 1, data = data, family = 'binomial')
summary(logfit)$coefficient
```

```
##      Estimate Std. Error   z value   Pr(>|z|)
## x1  0.1911909 0.08356338  2.287975 0.02213899
## x2 -0.1102284 0.05541257 -1.989231 0.04667566
```

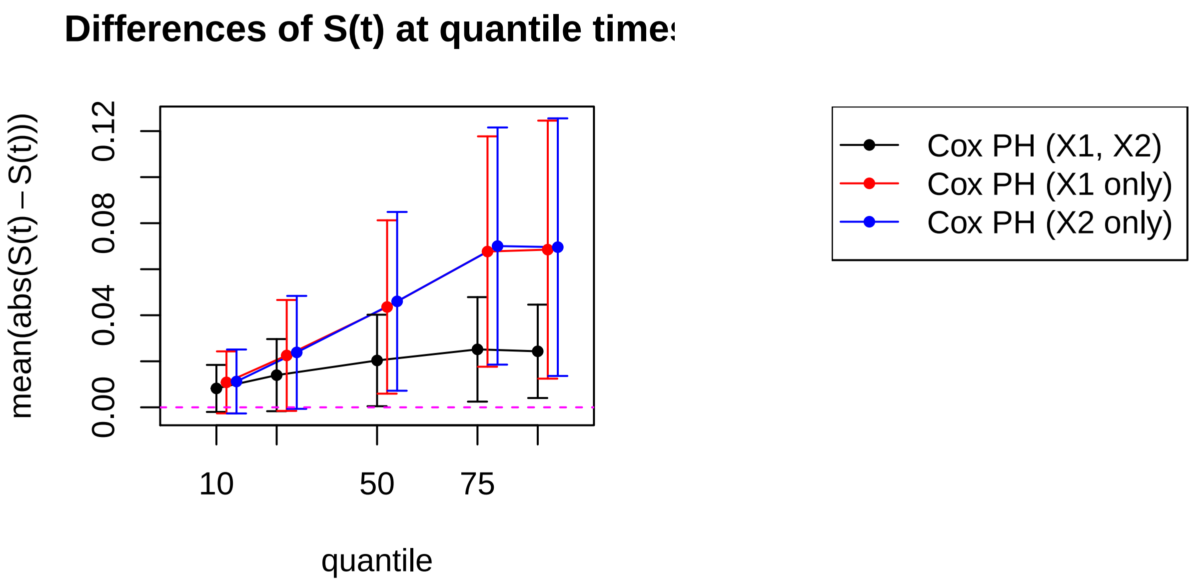
Bias $(S(t) - \hat{S}(t))$

quantile time	Cox (x1,x2)		Cox (x1 only)		Cox (x2 only)	
	mean	sd	mean	sd	mean	sd
10	-0.0011	0.0096	-0.0088	0.0093	-0.0094	0.0093
25	0.0003	0.0154	-0.0207	0.0148	-0.0221	0.0150
50	0.0004	0.0199	-0.0423	0.0211	-0.0447	0.0214
75	-0.0022	0.0225	-0.0667	0.0267	-0.0687	0.0275
90	-0.0019	0.0200	-0.0668	0.0309	-0.0666	0.0319



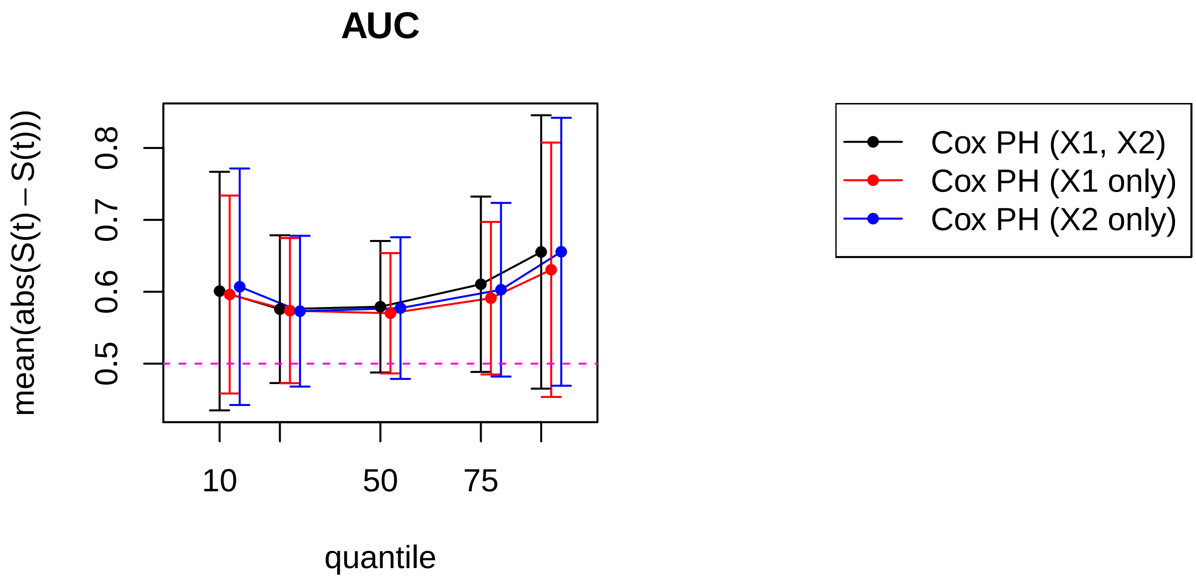
Absolute bias ($|S(t) - \hat{S}(t)|$)

quantile time	Cox (x1,x2)		Cox (x1 only)		Cox (x2 only)	
	mean	sd	mean	sd	mean	sd
10	0.0082	0.0052	0.0108	0.0069	0.0112	0.0071
25	0.0140	0.0080	0.0225	0.0123	0.0239	0.0125
50	0.0204	0.0101	0.0436	0.0192	0.0460	0.0198
75	0.0252	0.0116	0.0677	0.0255	0.0701	0.0263
90	0.0243	0.0104	0.0685	0.0286	0.0696	0.0286



Time depedent AUC

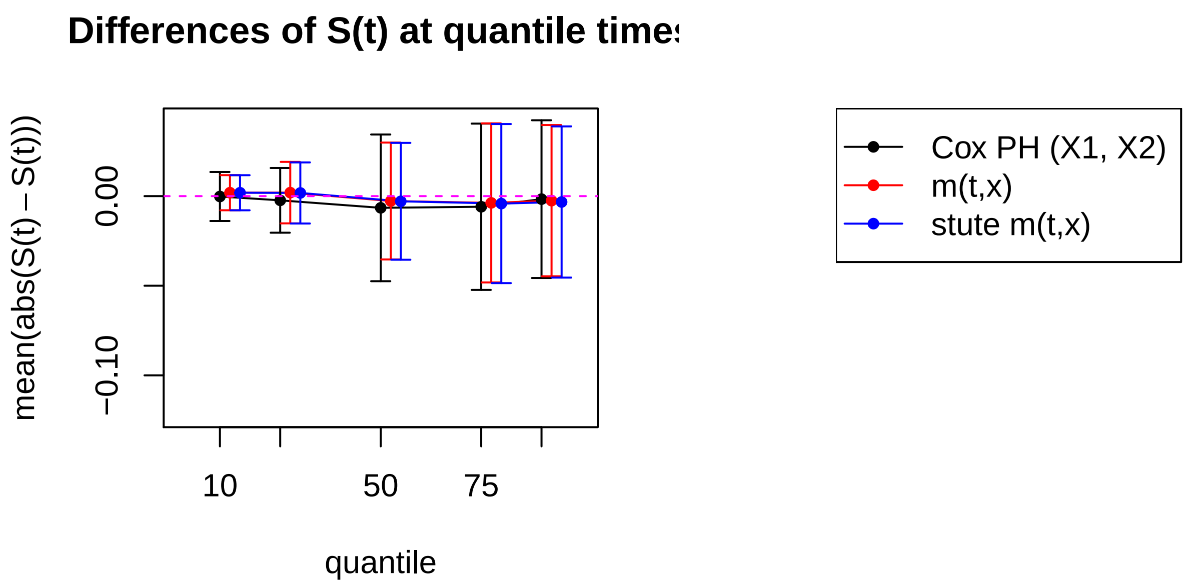
quantile time	Cox (x1,x2)		Cox (x1 only)		Cox (x2 only)	
	mean	sd	mean	sd	mean	sd
10	0.601	0.085	0.596	0.070	0.607	0.084
25	0.576	0.052	0.574	0.051	0.573	0.054
50	0.579	0.047	0.570	0.043	0.577	0.050
75	0.610	0.062	0.591	0.054	0.603	0.062
90	0.655	0.097	0.631	0.090	0.656	0.095

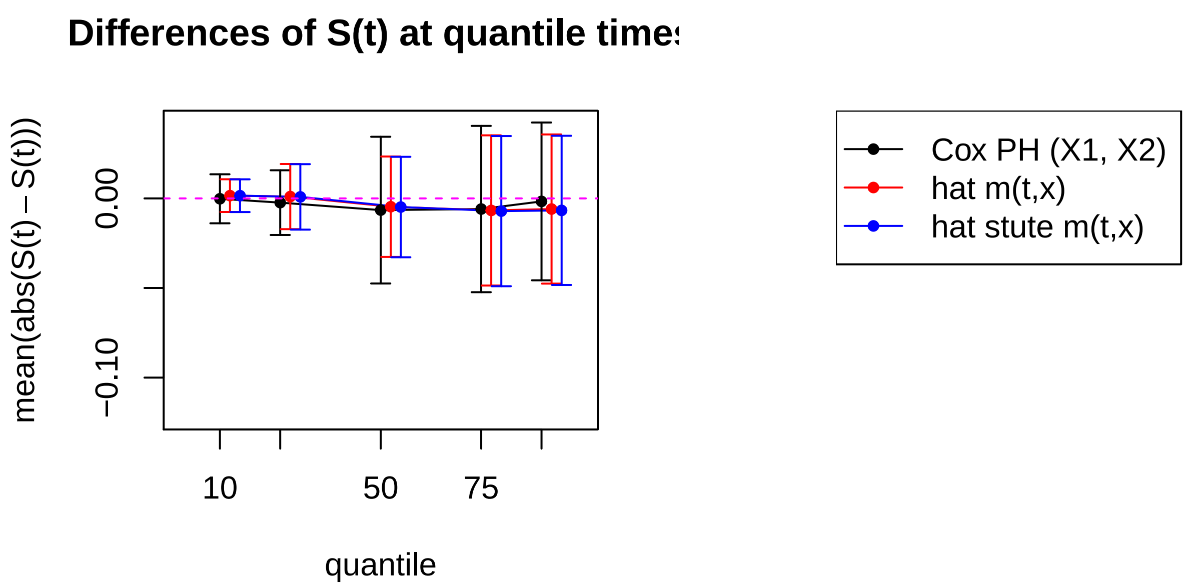


With Dikta's method (bias)

quantile time	Cox (x1,x2)		Cox (x1 only)		Cox (x2 only)	
	mean	sd	mean.1	sd.1	mean.2	sd.2
10	-0.0002	0.0070	-0.0071	0.0071	-0.0077	0.0071
25	-0.0023	0.0092	-0.0231	0.0082	-0.0245	0.0083
50	-0.0065	0.0209	-0.0480	0.0206	-0.0505	0.0210
75	-0.0059	0.0237	-0.0713	0.0242	-0.0738	0.0242
90	-0.0017	0.0225	-0.0645	0.0295	-0.0645	0.0288

quantile time	true m(t,x)		est m(t,x)		true stute m(t,x)		est stute m(t,x)	
	mean	sd	mean.1	sd.1	mean.2	sd.2	mean.3	sd.3
10	-0.0002	0.0050	0.0016	0.0047	0.0019	0.0050	0.0015	0.0047
25	-0.0023	0.0088	0.0010	0.0093	0.0018	0.0087	0.0008	0.0093
50	-0.0065	0.0166	-0.0046	0.0143	-0.0029	0.0166	-0.0048	0.0143
75	-0.0059	0.0226	-0.0067	0.0214	-0.0042	0.0226	-0.0071	0.0214
90	-0.0017	0.0215	-0.0059	0.0212	-0.0033	0.0215	-0.0067	0.0212





With Dikta's method (absolute bias)

quantile time	Cox (x1,x2)		Cox (x1 only)		Cox (x2 only)	
	mean	sd	mean.1	sd.1	mean.2	sd.2
10	0.0064	0.0028	0.0085	0.0053	0.0091	0.0053
25	0.0093	0.0053	0.0232	0.0081	0.0247	0.0080
50	0.0207	0.0116	0.0484	0.0202	0.0513	0.0202
75	0.0262	0.0119	0.0718	0.0235	0.0748	0.0229
90	0.0250	0.0117	0.0660	0.0276	0.0680	0.0251

quantile time	true m(t,x)		est m(t,x)		true stute m(t,x)		est stute m(t,x)	
	mean	sd	mean.1	sd.1	mean.2	sd.2	mean.3	sd.3
10	0.0064	0.0017	0.0081	0.0017	0.0082	0.0017	0.0081	0.0017
25	0.0093	0.0039	0.0193	0.0039	0.0187	0.0039	0.0193	0.0038
50	0.0207	0.0078	0.0379	0.0086	0.0376	0.0078	0.0379	0.0086
75	0.0262	0.0103	0.0546	0.0114	0.0534	0.0103	0.0546	0.0114
90	0.0250	0.0112	0.0577	0.0120	0.0558	0.0112	0.0578	0.0120