Survival Analysis with Stan

Mick Cooney mickcooney@gmail.com

Setting the Scene

Irish Life Insurance

12 months for expenses

12 months for commission

Credit Crisis 2008–2011

Lapses EXPLODED

Needed to understand this

Wanted a Bayesian approach

Very messy data

policy_id	sa_id	cluster_id	prem_ape	policy_startdate	policy_duration	gender_life1	smoker_life1	isjointlife	sum_assured	mortgage_status	policy_status	policy_lifetime	lapsed
C031552724	A267029034	n6_c3	506.95	1999-02-10	20	M	S	FALSE	€ 100,000	MORTDECR	inforce	829.14	FALSE
C056432445	A187069008	n6_c0	4361.25	2003-06-16	30	M	S	FALSE	€ 500,000	MORTFULL	inforce	602.43	FALSE
C090642429	A267065025	n6_c5	947.04	2011-05-20	20	M	N	FALSE	€ 500,000	MORTDECR	inforce	188.86	FALSE
C021175270	A267120015	n6_c4	5230.08	1996-12-11	20	F	Q	TRUE	€ 450,000	TERM	lapsed	460.86	TRUE
C086610431	A267022005	n6_c5	4815.34	2009-03-19	35	M	N	TRUE	€ 400,000	TERM	lapsed	95.86	TRUE
C013351780	A267007015	n6_c5	35893.10	1993-07-09	10	M	N	TRUE	€ 450,000	MORTDECR	lapsed	4.43	TRUE
C077078315	A267134010/267134011	n6_c5	71.29	2006-12-20	5	M	N	FALSE	€ 100,000	MORTFULL	lapsed	8.86	TRUE
C045495528	A067014004	n6_c1	2141.31	2001-08-01	20	M	N	FALSE	€ 500,000	MORTFULL	lapsed	726.00	TRUE
C088994518	A267158004	n6_c5	427.06	2010-05-19	10	F	Q	FALSE	€ 300,000	MORTDECR	inforce	241.14	FALSE
C057468093	A167038023	n6_c5	5232.56	2003-09-01	27	M	S	TRUE	€ 200,000	MORTDECR	inforce	591.43	FALSE
C018278160	A087002010	n6_c3	288.96	1996-02-20	20	M	N	FALSE	€ 100,000	MORTFULL	lapsed	591.14	TRUE
C089762507	A147042008	n6_c0	930.30	2010-11-03	15	M	N	FALSE	€ 200,000	MORTDECR	lapsed	104.43	TRUE
C059659062	A087071029	n6_c5	1070.70	2004-02-10	20	F	Q	FALSE	€ 100,000	TERM	lapsed	304.29	TRUE
C045833535	A267021009	n6_c4	6144.22	2001-09-05	20	F	Q	TRUE	€ 450,000	MORTFULL	inforce	695.14	FALSE
C088717225	A207001001	n6_c2	10826.43	2010-04-13	20	M	S	TRUE	€ 300,000	MORTDECR	inforce	246.29	FALSE

Logistic regression

Conditional probability of lapse occurring


VAT
69

“Computer science research in the corporate environment is not a good idea”

Nothing to lose...

Try using survival analysis instead

Basic Concepts

Time-to-Event modelling

Origins in medical statistics

Cancer research

Censoring and Truncation

Censoring: Incomplete observation of data used

Truncation: Data not observed as a result of the observation size

Survival and Hazard Functions

Aim of modelling is to estimate distribution of lifetime of subjects.

Data often heavily censored

Right Censoring

Event not occurred for most subjects at point of observation

Only know lower-bound for the time-to-event

Models need to account for this

Survival function, $S(t)$:

$$S(t) = P(T > t),\, 0 \leq t \leq \infty$$

Cumulative Hazard function, $\Lambda(t)$:

$$S(t) = e^{-\Lambda(t)}$$

Hazard function, $\lambda(t)$:

$$\lambda(t) = \lim_{\delta t \to 0^+} \frac{P(t \leq T < t + \delta t \mid T \geq t)}{\delta t}$$

Key Relationship

$$S(t) = e^{-\Lambda(t)} = e^{-\int_0^t \lambda(\tau)\, d\tau}$$

Often (but not always) most intuitive to model hazard function.

Semi-parametric Models

Non-parametric baseline, parametric hazards etc.

Proportional hazards

Cox PH Models

$$\lambda(t \mid \mathbf{x}) = \lambda_0(t)\ \exp\left(\mathbf{x}^\top \beta\right)$$

- **Baseline hazard**: $\lambda_0(t)$ non-parametric (unspecified)
- **Relative risk**: multiplicative effect per covariate

Assumptions and interpretation:

- **Proportional hazards**: ratios of hazards are constant over time
- **Semi-parametric**: baseline is free; covariate effects parametric
- **Coefficients**: log hazard ratios; $\exp(\beta)$ is hazard ratio

Estimation (partial likelihood):

$$L(\beta) = \prod_{i \in \mathcal{D}} \frac{\exp(\mathbf{x}_i^\top \beta)}{\sum_{j \in R(t_i)} \exp(\mathbf{x}_j^\top \beta)}$$

- $\mathcal{D}$: set of observed events
- $R(t_i)$: risk set just before time t_i

Bayesian Cox via brms/rstanarm:

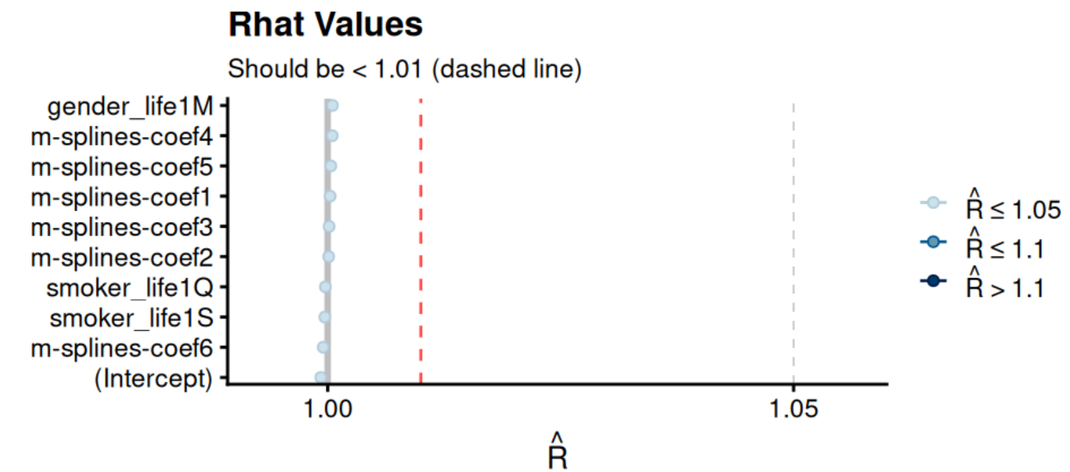
- **brms**: `family = cox(bhaz = list(df = 8))` for M-spline baseline
- **rstanarm**: `stan_surv()` in branch ‘feature/survival’
- **Outputs**: posterior hazard ratios, baseline hazard/survival via stored basis

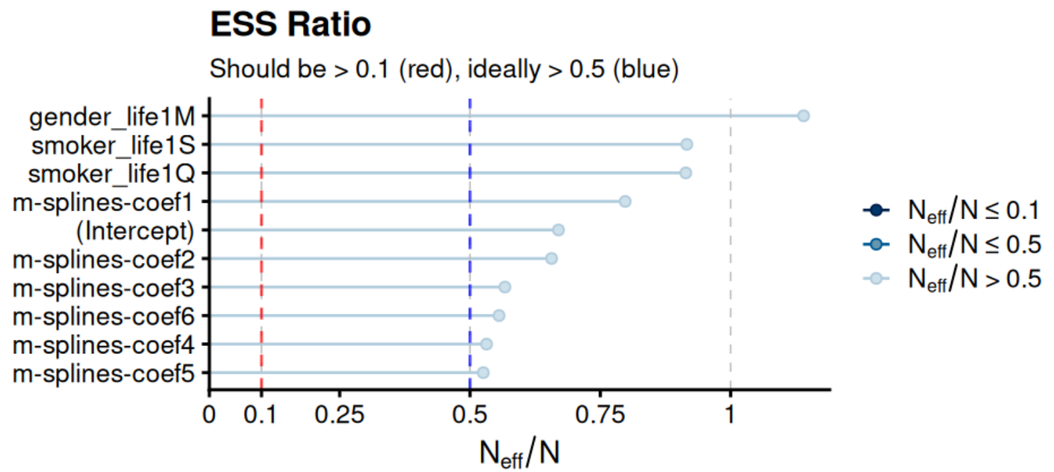
Bayesian Survival Analysis

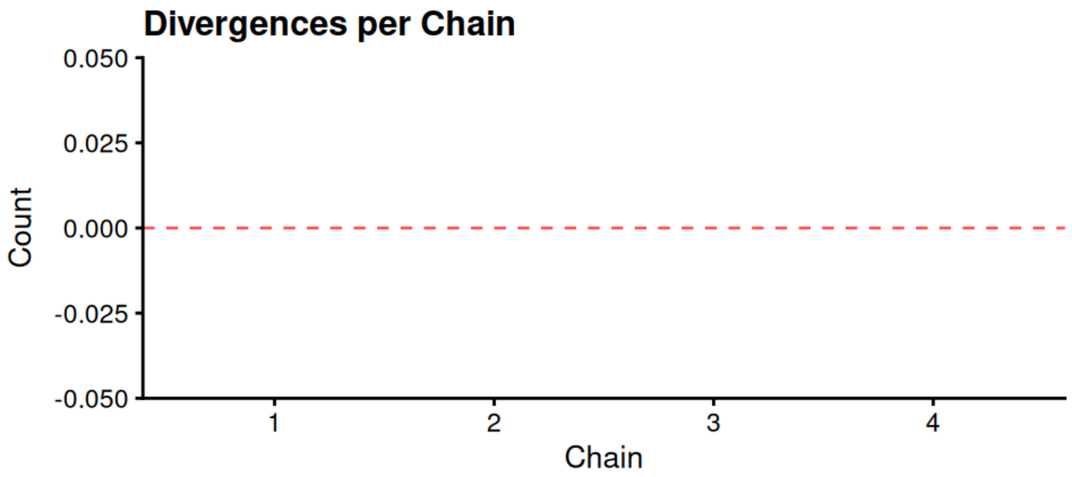
Fit a simple model first

```
lapse1_coxph_stansurv <- stan_surv(
  Surv(policy_lifetime, lapsed) ~ gender_life1 + smoker_life1,
  data   = model_training_tbl,

  # MCMC sampling parameters
  ...

  # Baseline hazard
  basehaz     = "ms",          # M-splines (flexible, default)
  basehaz_ops = list(df = 6),  # Degrees of freedom for baseline hazard

  # Prior specifications (explicit is better than implicit)
  prior           = rstanarm::normal(location = 0, scale = 2.5), # Weakly informative for coefficients
  prior_intercept = rstanarm::normal(location = 0, scale = 10),  # For baseline hazard
  )
```

Model 1 Convergence Diagnostics: Gender + Smoker







Sampling Diagnostics

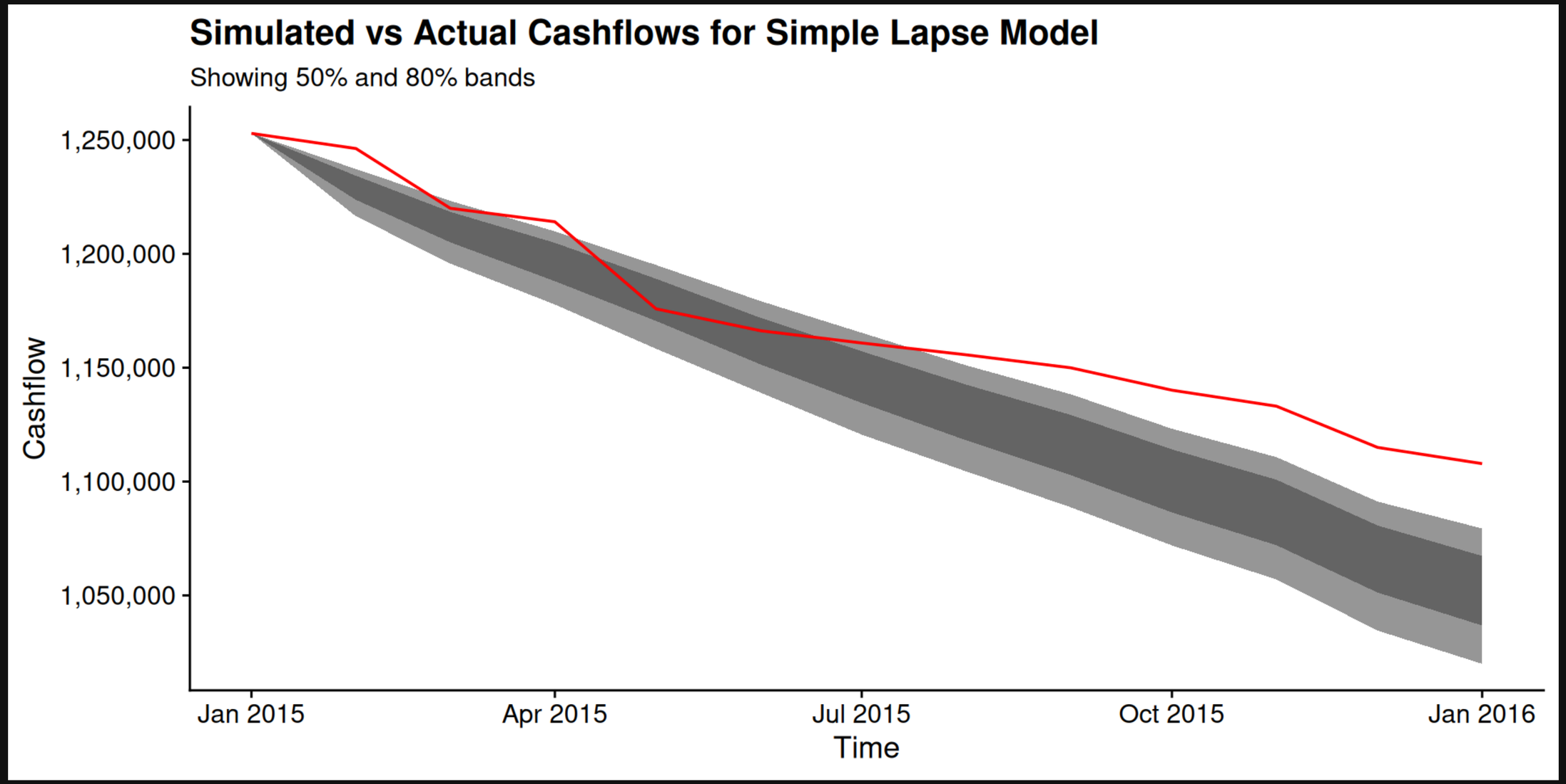
```
MCMC Diagnostics Summary
========================

Chains: 4
Iterations: 2000
Warmup: 1000

Divergent transitions: 0
Max treedepth hits: 0

Parameters monitored: 10
```

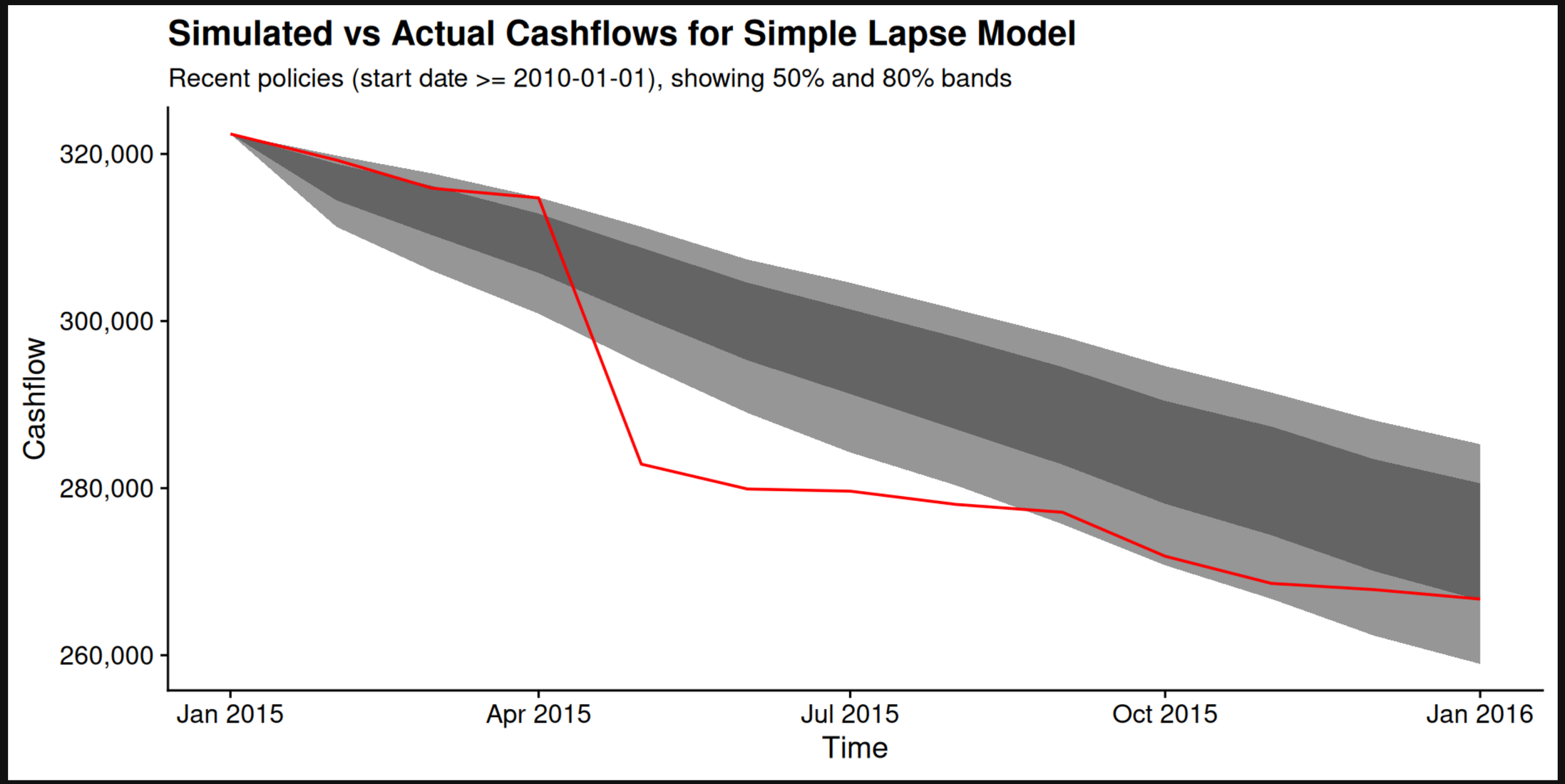
Assess the Model

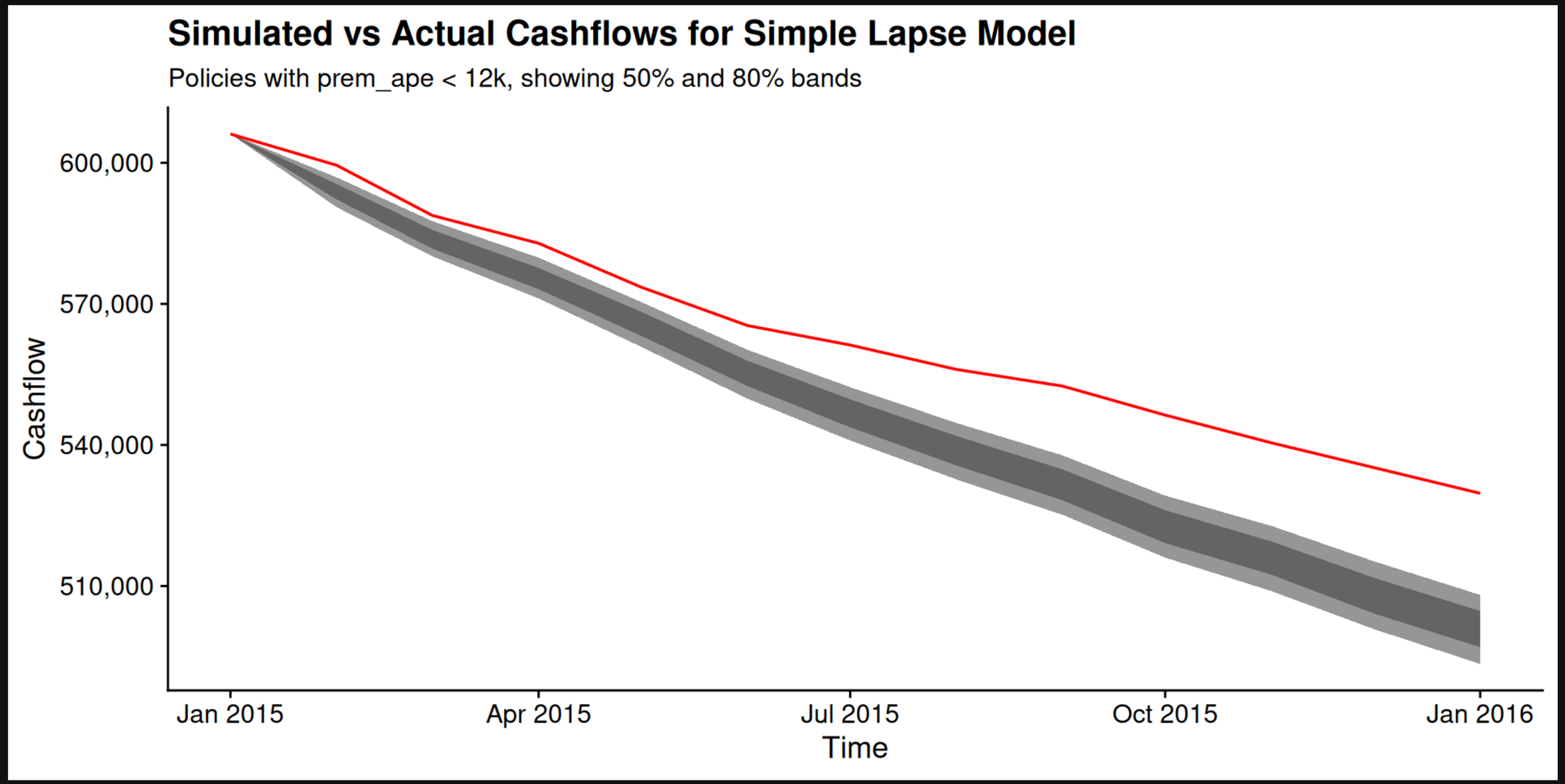
Simulate lapses over a year

Calculate corresponding cashflows

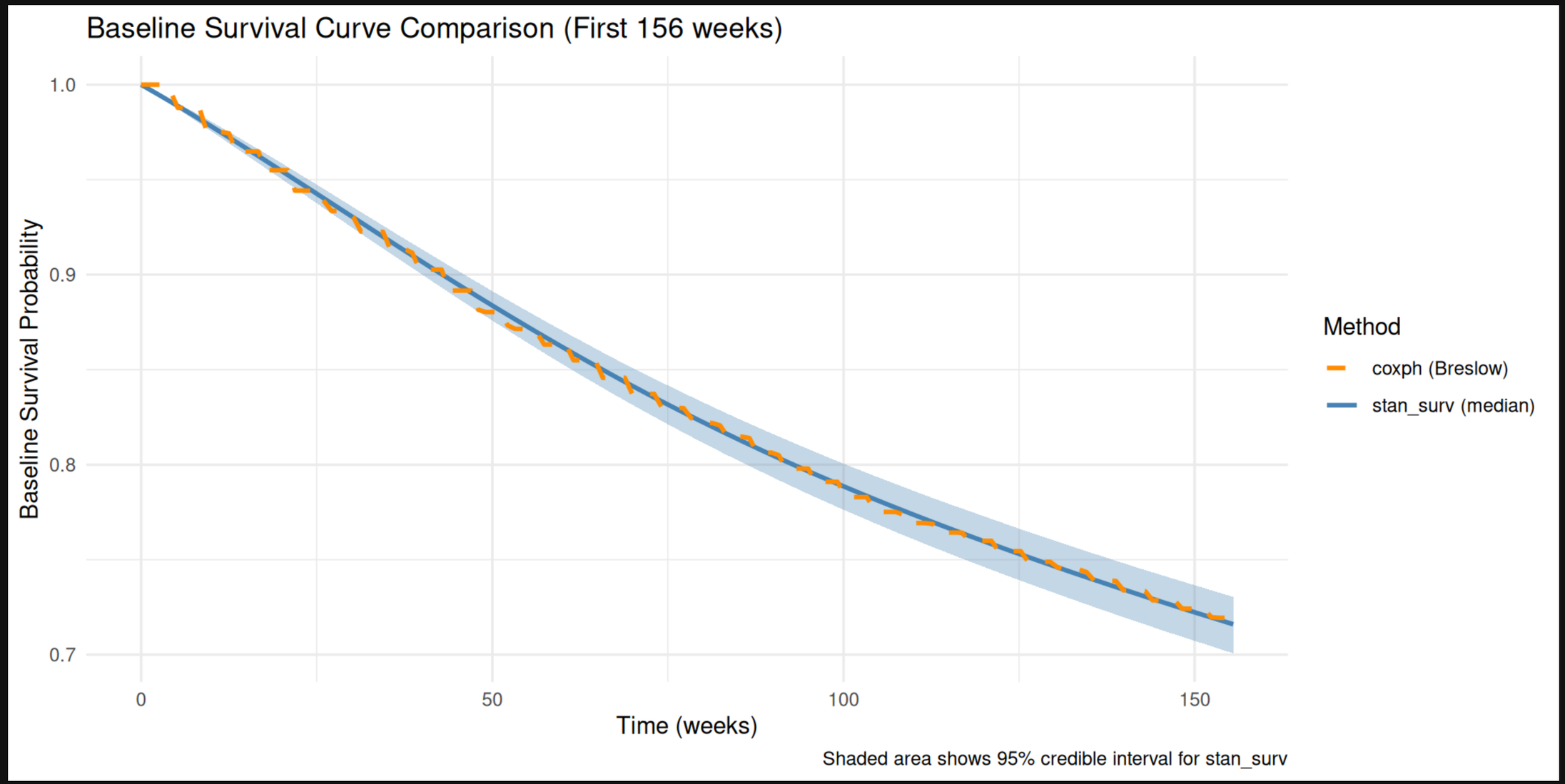
Compare to actual


Simulated vs Actual Cashflows for Simple Lapse Model
Showing 50% and 80% bands
Cashflow
1,250,000
1,200,000
1,150,000
1,100,000
1,050,000
Jan 2015
Apr 2015
Jul 2015
Oct 2015
Jan 2016
Time


Simulated vs Actual Cashflows for Simple Lapse Model
Recent policies (start date >= 2010-01-01), showing 50% and 80% bands
Cashflow
320,000
300,000
280,000
260,000
Jan 2015
Apr 2015
Jul 2015
Oct 2015
Jan 2016
Time


Simulated vs Actual Cashflows for Simple Lapse Model
Policies with prem_ape < 12k, showing 50% and 80% bands
Cashflow
600,000
570,000
540,000
510,000
Jan 2015
Apr 2015
Jul 2015
Oct 2015
Jan 2016
Time

Baseline hazard too high?


Baseline Survival Curve Comparison (First 156 weeks)
Baseline Survival Probability
1.0
0.9
0.8
0.7
0
50
100
150
Time (weeks)
Method
coxph (Breslow)
stan_surv (median)
Shaded area shows 95% credible interval for stan_surv

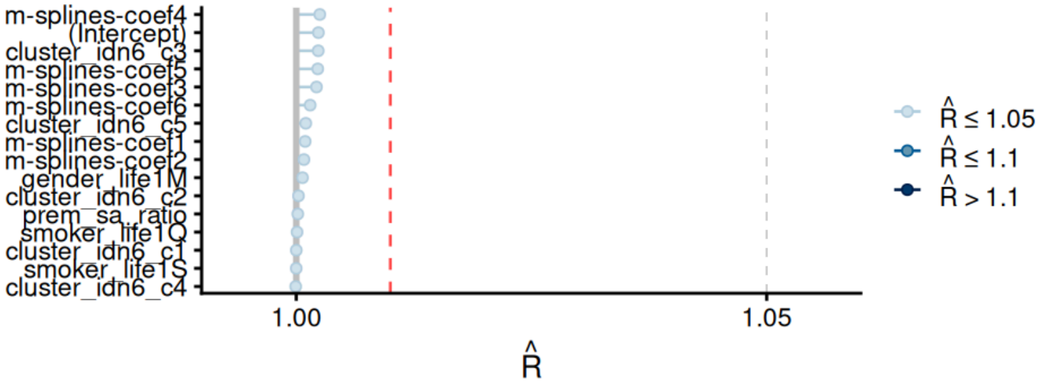
Improving the Model

```
lapse2_coxph_stansurv <- stan_surv(
  Surv(policy_lifetime, lapsed) ~ gender_life1 + smoker_life1 + cluster_id + prem_sa_ratio,
  data   = model_training_tbl,

  # MCMC sampling parameters
  ...

# Baseline hazard specification
basehaz     = "ms",  # M-splines (flexible, default)
basehaz_ops = list(df = 6),  # Degrees of freedom for baseline hazard

# Prior specifications (explicit is better than implicit)
prior           = rstanarm::normal(location = 0, scale = 2.5), # Weakly informative for coefficients
prior_intercept = rstanarm::normal(location = 0, scale = 10),  # For baseline hazard
)
```

Model 2 Convergence Diagnostics

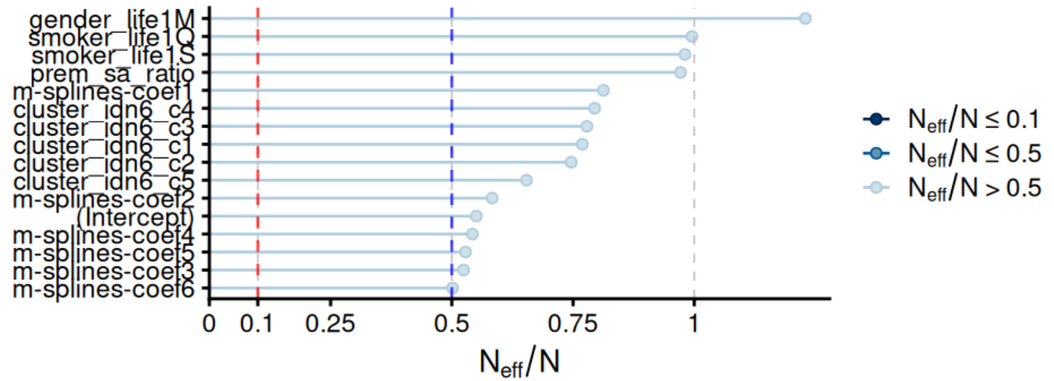
Rhat Values

Should be < 1.01 (dashed line)

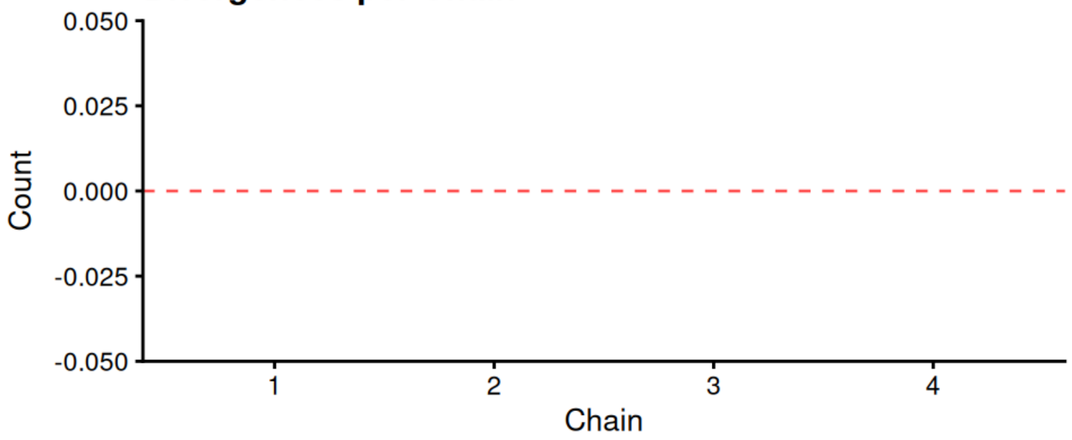


ESS Ratio

Should be > 0.1 (red), ideally > 0.5 (blue)

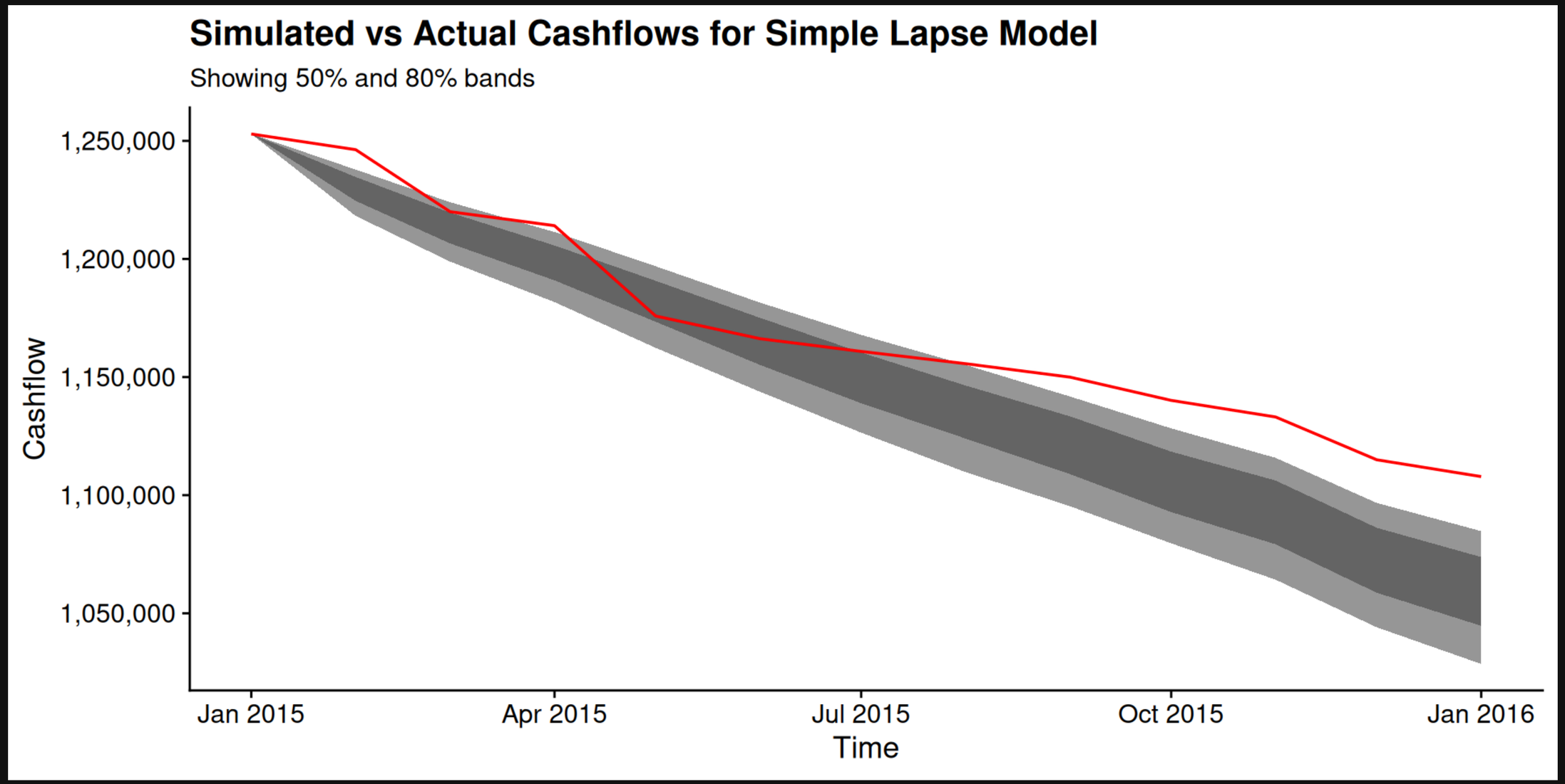


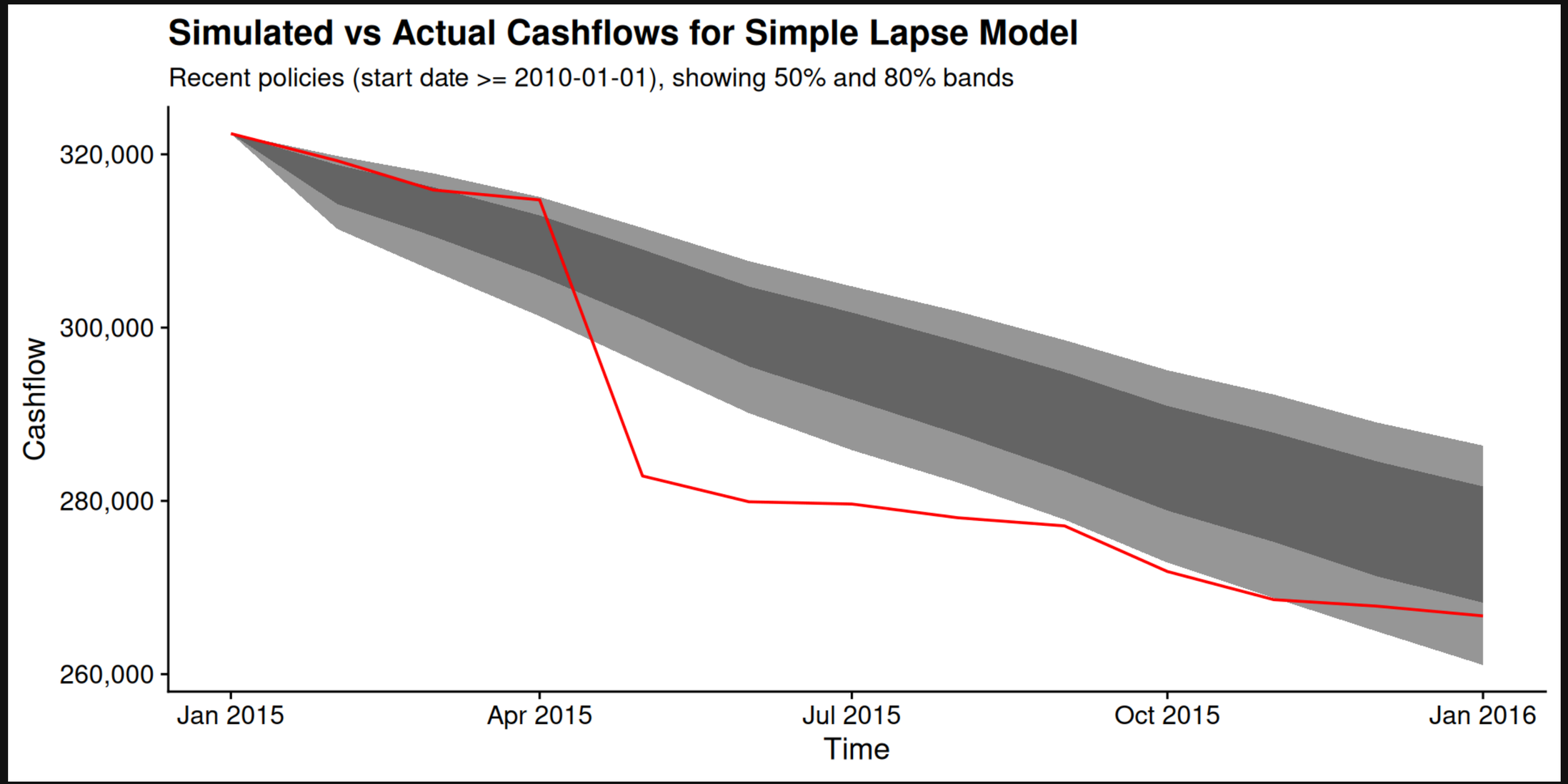
Divergences per Chain

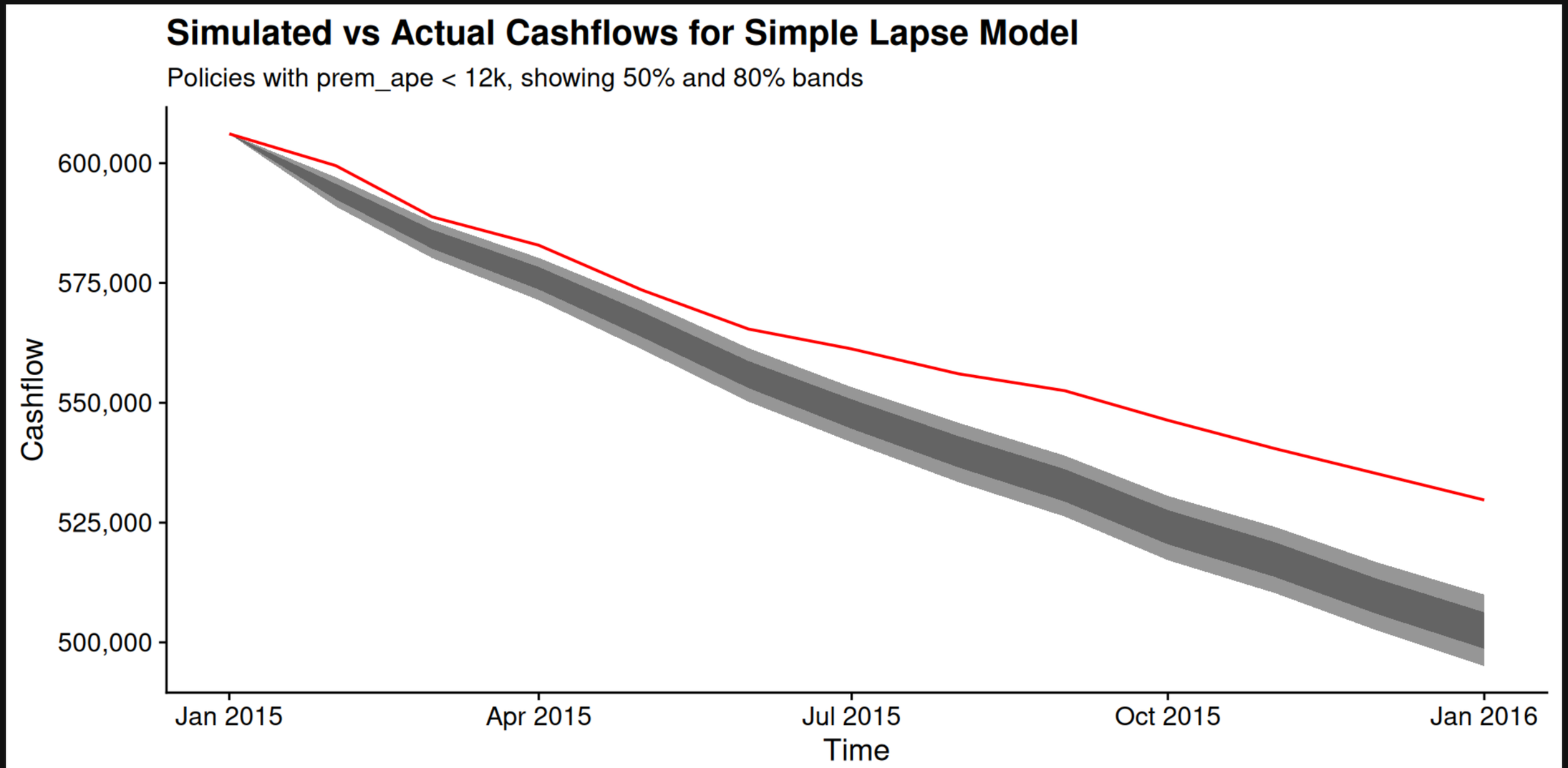


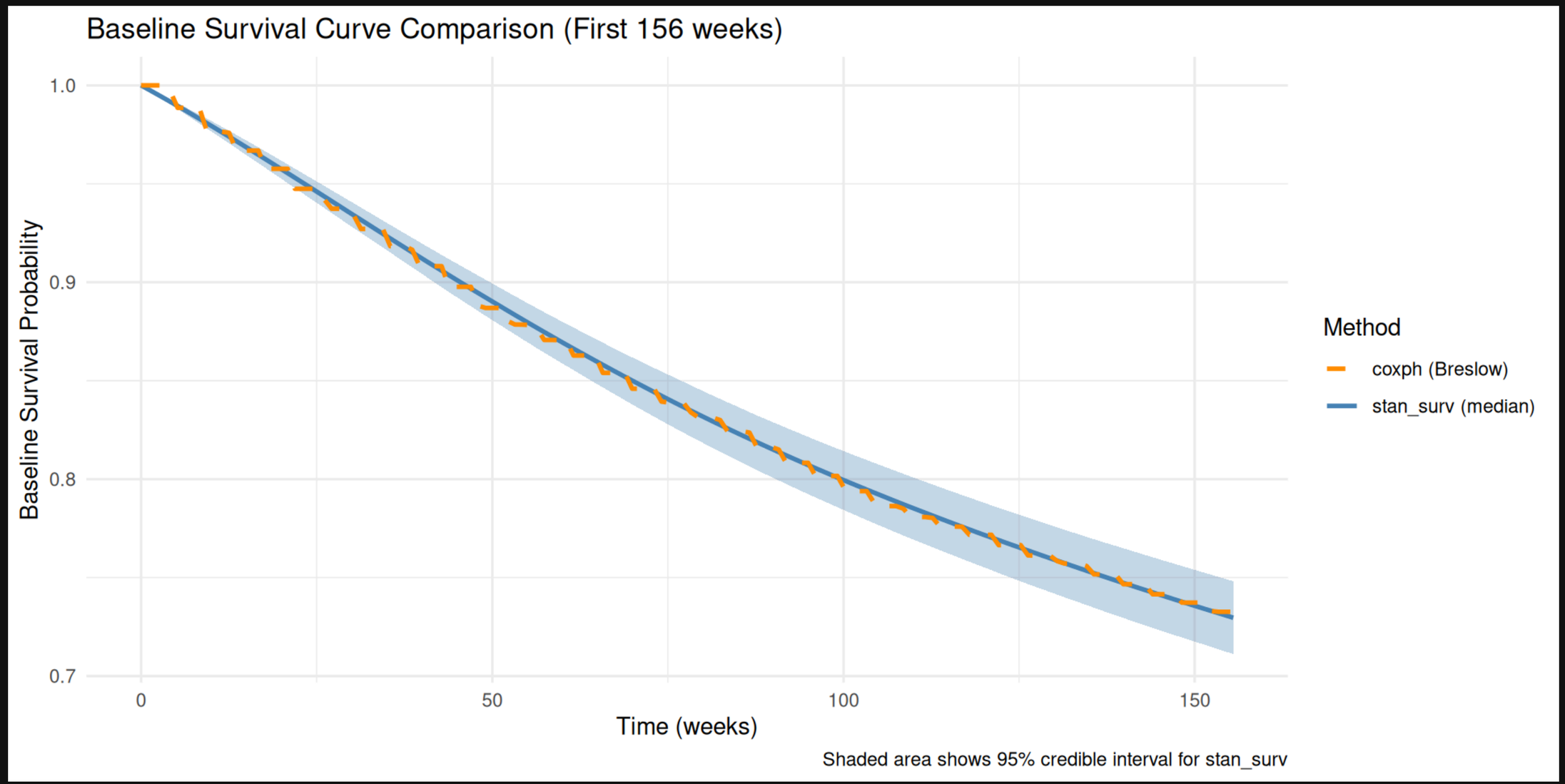
Sampling Diagnostics

```
MCMC Diagnostics Summary
========================

Chains: 4
Iterations: 2000
Warmup: 1000

Divergent transitions: 0
Max treedepth hits: 0

Parameters monitored: 16
```


Simulated vs Actual Cashflows for Simple Lapse Model
Showing 50% and 80% bands
Cashflow
1,250,000
1,200,000
1,150,000
1,100,000
1,050,000
Jan 2015
Apr 2015
Jul 2015
Oct 2015
Jan 2016
Time


Simulated vs Actual Cashflows for Simple Lapse Model
Recent policies (start date >= 2010-01-01), showing 50% and 80% bands
Cashflow
320,000
300,000
280,000
260,000
Jan 2015
Apr 2015
Jul 2015
Oct 2015
Jan 2016
Time


Simulated vs Actual Cashflows for Simple Lapse Model
Policies with prem_ape < 12k, showing 50% and 80% bands
Cashflow
600,000
575,000
550,000
525,000
500,000
Jan 2015
Apr 2015
Jul 2015
Oct 2015
Jan 2016
Time


Baseline Survival Curve Comparison (First 156 weeks)
Baseline Survival Probability
1.0
0.9
0.8
0.7
0
50
100
150
Time (weeks)
Method
coxph (Breslow)
stan_surv (median)
Shaded area shows 95% credible interval for stan_surv

Conclusions and Next Steps

Survival analysis feasible in Stan

Need to try more in `brms`

Lapse rates over-estimated by the model

Needs a lot more work to be usable

Thank You!

https://github.com/kaybenleroll/data_workshops/talk_idsc_survivalbayes_202606

Main project

https://github.com/kaybenleroll/bayesian_survival_analysis

Old Workshop (non-Bayes)

https://kaybenleroll.github.io/data_workshops/oldertalks/ws_survival_201809/worksheet