Федеральное государственное автономное образовательное учреждение высшего профессионального образования «Национальный исследовательский университет «Высшая школа экономики»

Факультет экономических наук

Образовательная программа «Экономика»

**Домашнее задание №2**

по курсу

«Анализ панельных данных и данных о длительности состояний»

Выполнил:

Пешков Максим БЭК181

Москва, 2021

# Задание №1

Создайте выборку из 1000 случайных чисел , распределённых по закону Вейбулла с с параметрами . Это будут изучаемые длительности.

Ниже представлен код, чтобы задать количество наблюдений, зафиксировать сид для генерации случайных величин, а также код для генерации случайных чисел.

**set obs 1000**

**set seed 777**

**gen t\_star = (-ln((1-uniform()))/0.07)^(1/0.7)**

Эта формула взята из учебника, которая на самом деле легко выводится:

# Задание №2

Теперь сгенерируйте 1000 значений равномерно распределённых на отрезке [0:80] - это будут моменты цензурирования.

Ниже представлен код для генерации. Сначала мы генерируем равномерные случайные величины на [0,1), после чего при умножении на 80 мы получаем случайные величина распределенные равномерно на отрезке [0,80).

**gen t\_c = uniform()\*80**

# Задание №3

Рассчитайте величины . Это будут наблюдаемые длительности. Если , то состояние i наблюдалось только в течение единиц времени и наблюдение за ним оказалось цензурированным. Если же , то наблюдаемая длительность совпадает с полной, наблюдение не цензурировано.

Ниже представлен код для генерации минимума между

**gen t = min(t\_star, t\_c)**

# Задание №4

Создайте переменную

То есть для не цензурированных наблюдений, для цензурированных.

Ниже представлен код для генерации .

**gen delta = 1**

**replace delta=0 if t\_c <= t\_star**

# Задание №5

По данным о наблюдаемых длительностях и индикаторе цензурирования/завершения состояний рассчитайте оценку Каплана–Майера для функции дожития и оценку Нельсона– Аалена для интегральной функции риска.

Ниже представлен код, который позволяет передать STATA формат данных через stset, посчитать на каждый момент времени оценку функции дожития Каплана-Майера и оценку интегральной функции риска Нельсона-Аалена, а также вывести их графическое представление на Рисунке 1 и Рисунке 2 соответственно.

**stset t, fail(delta)**

**sts gen s\_km = s**

**sts gen cumhaz = na**

**sts graph //KM**

**sts graph, na //NA**

Рисунок 1: Оценка Каплана-Майера для функции дожития

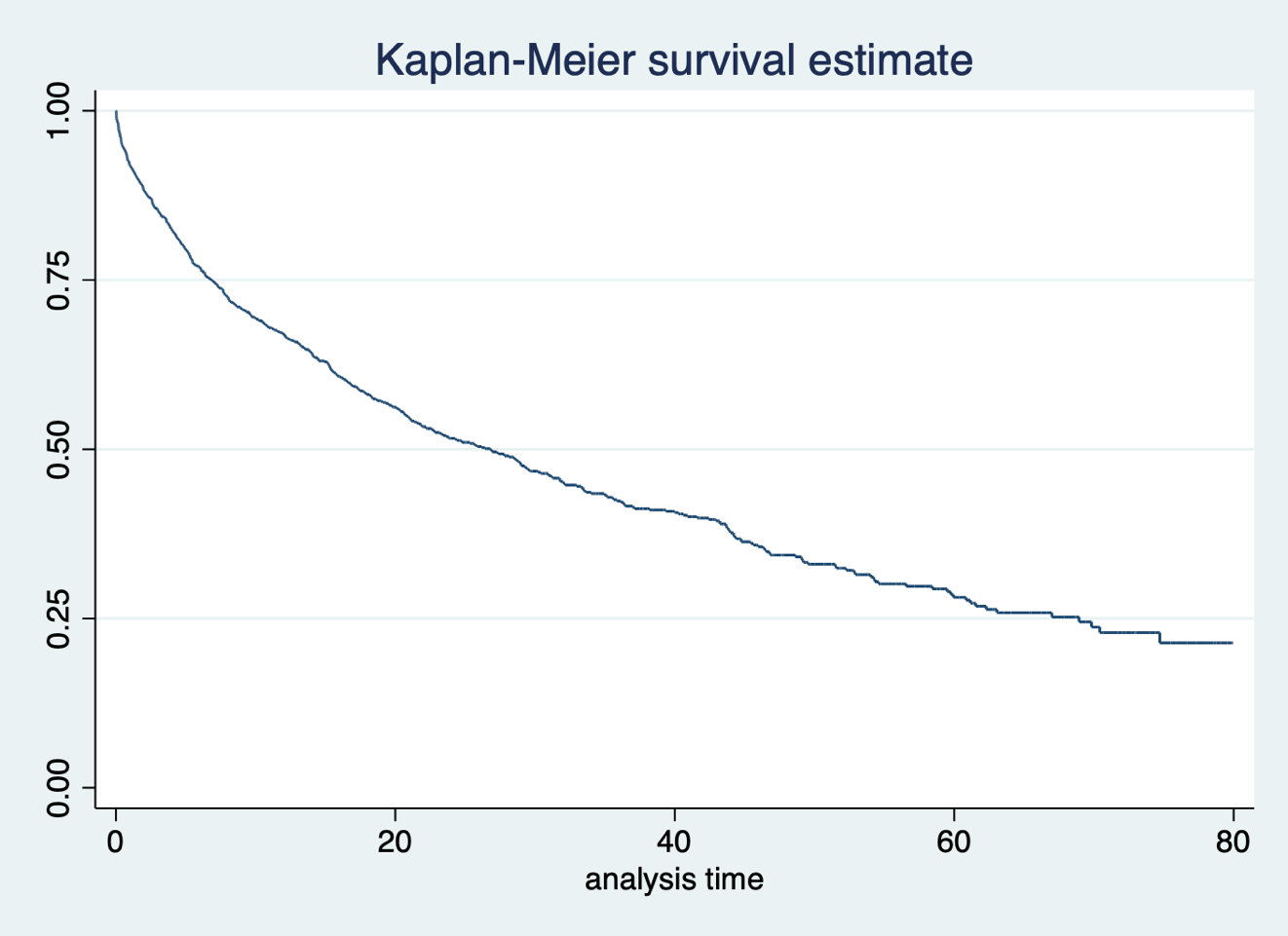
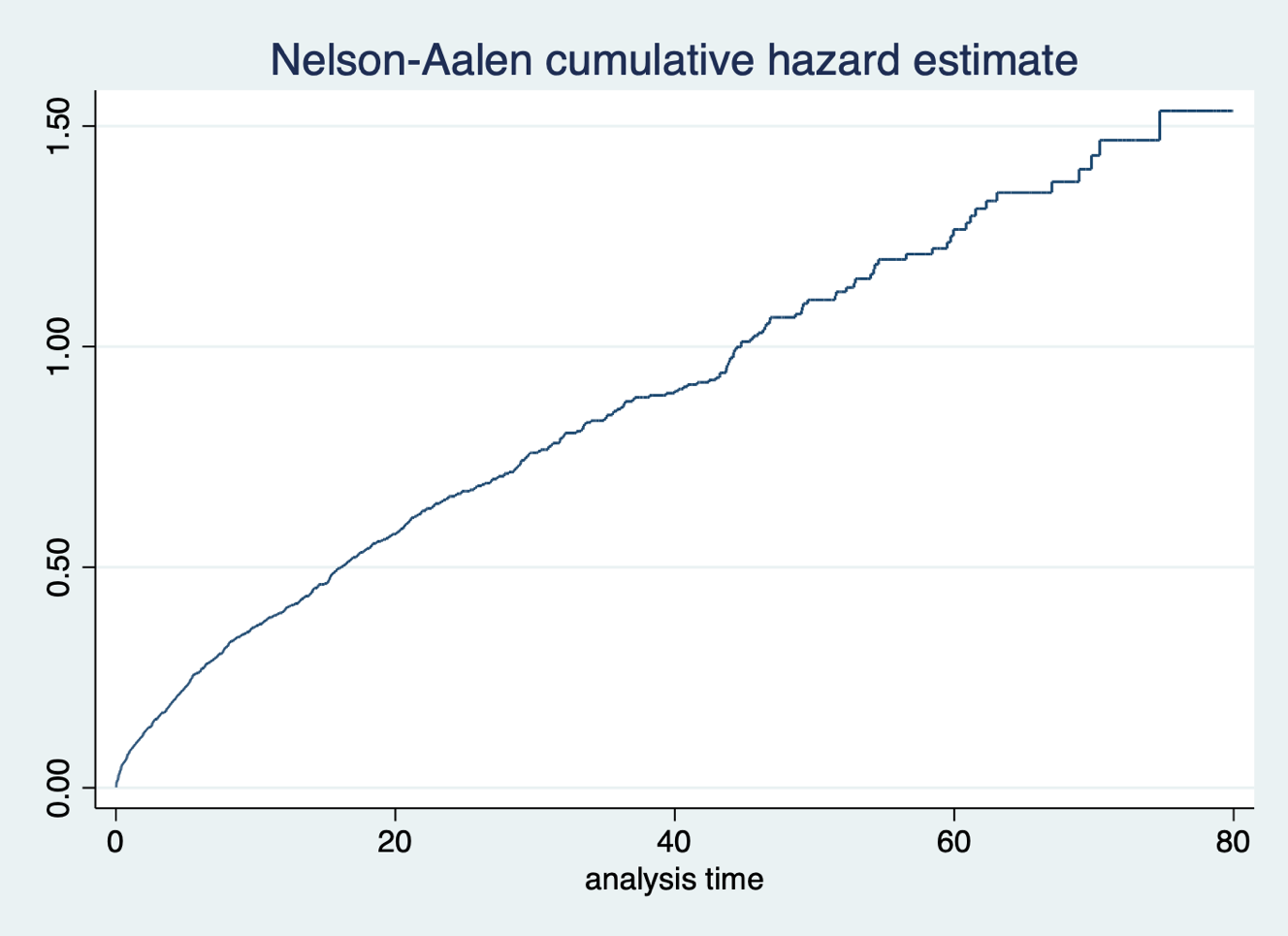


Рисунок 2: Оценка Нельсона-Аалена для интегральной функции риска



Как видно из рисунков обе оценки находятся на промежутке [0,80], что ожидаемо, так как нет наблюдений, которые мы могли наблюдали больше 80, ведь выборка цензурирована справа (а моменты цензурирования ограничены 80).

Также видно, что функция дожития в момент времени 80 чуть меньше 0.25 и не равно 0, что в целом ожидаемо, так как в момент цензурирования есть наблюдения, которые еще не завершились и находятся под риском.

Более того из рисунка интегральной функции риска видно, что риск уменьшается по мере увеличения длительности, так как функция выпукла вверх (вогнута).

# Задание №6

Изобразите на одном графике:

* Настоящую функцию дожития для распределения Вейбулла с параметрами
* Оценку Каплана-Майера

То же самое проделайте и для интегральной функции риска.

Если вы можете как-нибудь прокомментировать получившиеся графики, сделайте это, пожалуйста.

Для начала напишем какие настоящие функции дожития и функции интегрального риска для распределения Вейбулла

То есть с заданными параметрами они будут равны

Ниже представлен код для вывода нужных графиков, где через True\_S и True\_H обозначены истинные функции дожития и интегрального риска (сами графики представлены на Рисунке 3 и Рисунке 4).

**twoway (line s\_km \_t, connect(stairstep) sort lcolor(black))||function True\_S=exp(-0.07\*(x)^(0.7)), range(0 80)**

**twoway (line cumhaz \_t, connect(stairstep) sort lcolor(black))||function True\_H=0.07\*(x)^(0.7), range(0 80)**

Рисунок 3: Сопоставление оценки функции дожития Каплана-Майера и истинной функций дожития

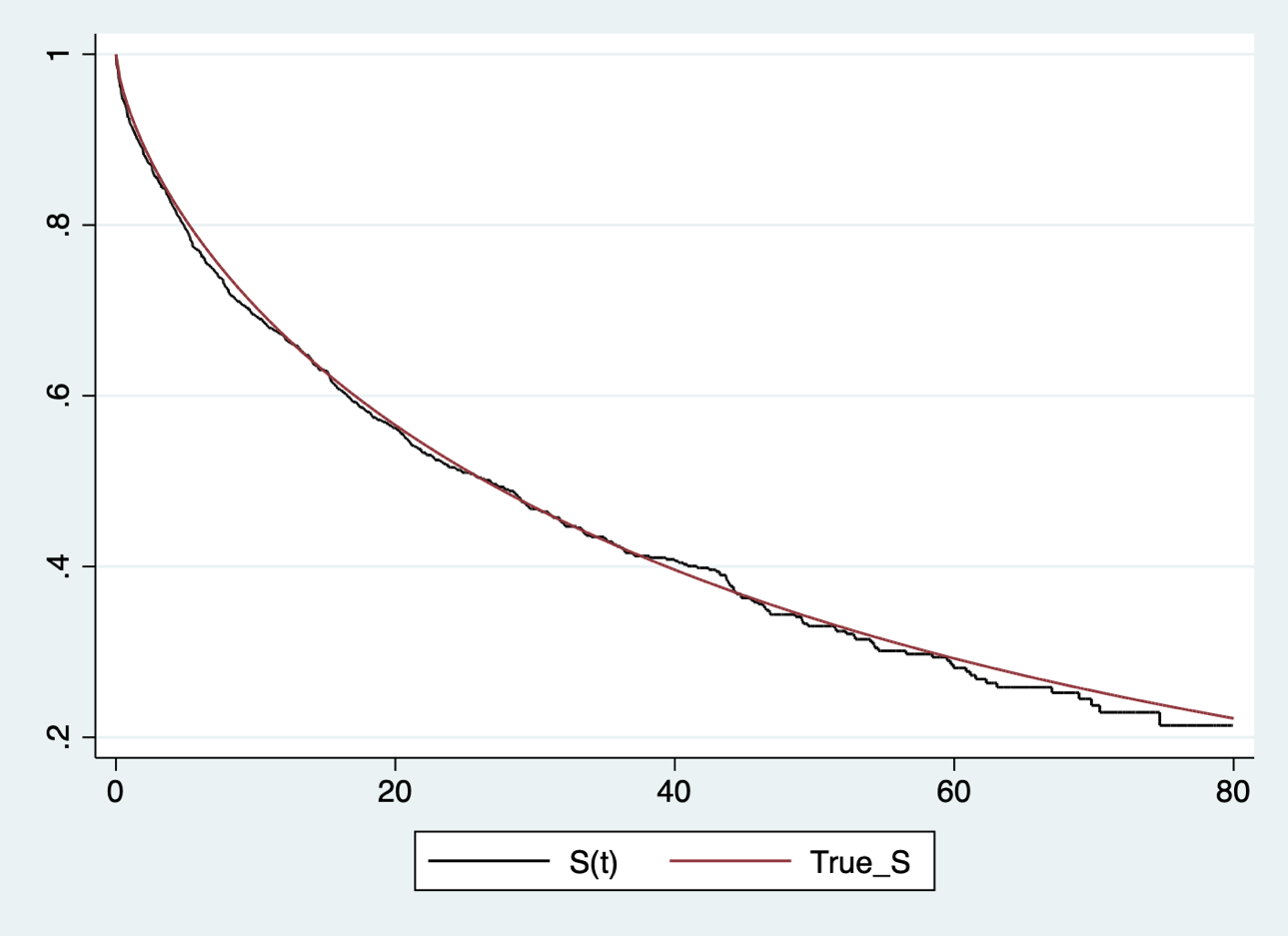
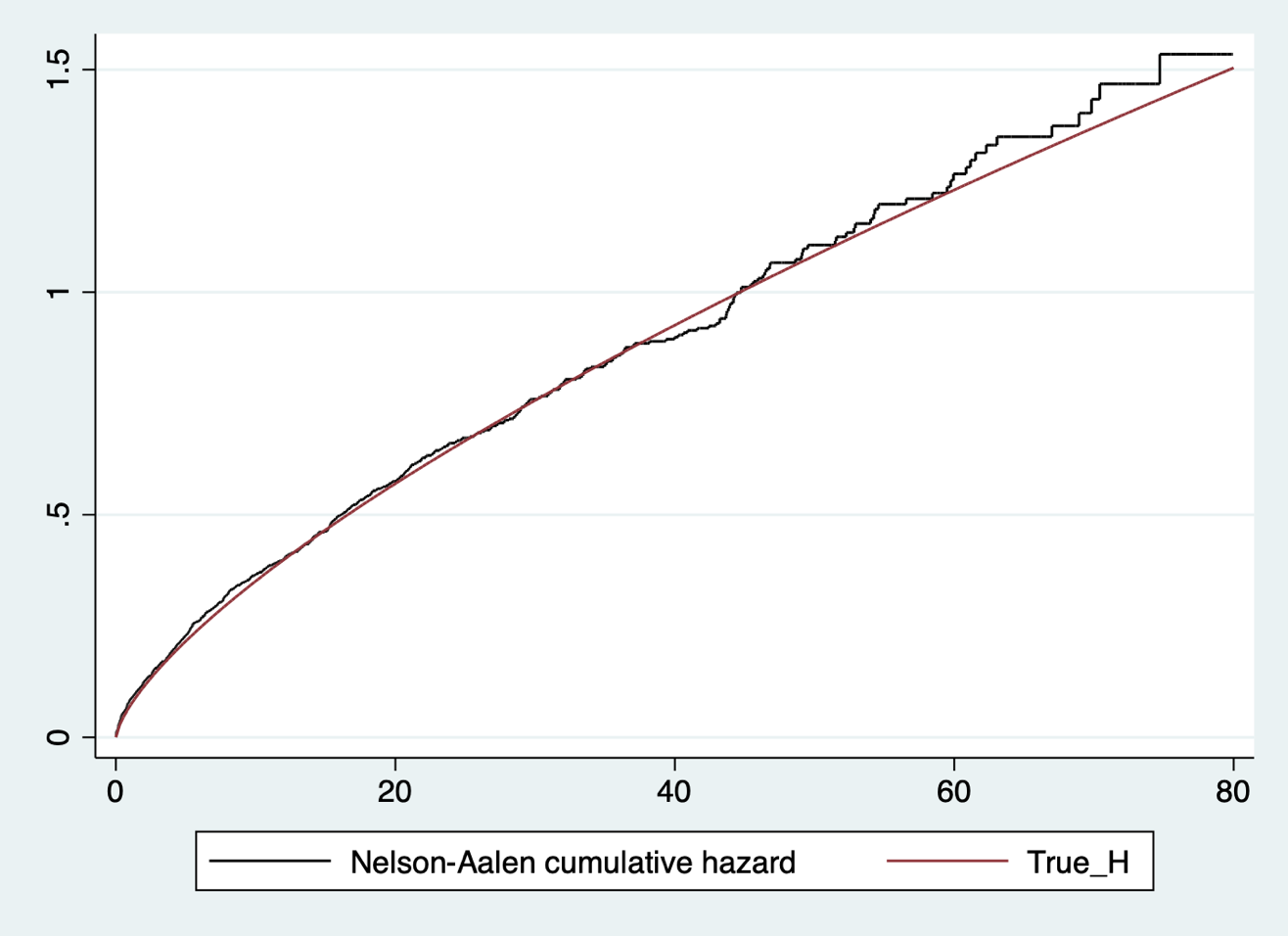


Рисунок 4: Сопоставление оценки интегральной функции риска Нельсона-Аалена и истинной функции интегрального риска



Как можно заметить на обоих рисунках оценки практически совпадают с истинными функциями, хотя и видно, что на правом хвосте есть небольшие расхождения, особенно у оценки интегральной функции риска Нельсона-Аалена, это может быть обосновано тем, что в конце остается мало наблюдений и дисперсия оценок увеличивается.

Также видно, что функция дожития в момент времени 80 чуть меньше 0.25 и не равно 0, что в целом ожидаемо, так как в момент цензурирования есть наблюдения, которые еще не завершились и находятся под риском.

Более того из рисунка интегральной функции риска видно, что риск уменьшается по мере увеличения длительности, так как функция выпукла вверх (вогнута), что также свойственно и для истинной функции интегрального риска, ведь для распределения Вейбулла риск уменьшается по мере увеличения длительности наблюдения при заданных параметрах.

Таким образом, найденные оценки позволяют найти по выборке близкие к истинным значения.

# Задание №7

Сравните выборочную медиану с истинной медианой распределения Вейбулла с заданными параметрами.

Для начала выпишем функцию квантилей для распределения Вейбулла, в которую потом подставим заданные параметры и найдем медиану.

В нашем случае это будет равно

Ниже представлен код для подсчета медианы и выборочной медианы, а также записаны какие значения медианы получаются.

**stsum //26.78968**

**display (-(ln(0.5)/0.07))^(1/0.7) //26.452562**

Значение по выборки: 26.78968, Истинное значение: 26.452562, которые означают, что половина наблюдений в выборке закончились до 26.78968, а другая половина после.

Как видно значения практически равны, что было ожидаемо, так как функции дожития и интегральные функции риска практически совпадают, то есть оценки являются достаточными точными, что подтверждается медианами.

# Задание №8

Теперь рассмотрите случай, когда данные подвержены усечению справа вместо цензурирования, так что все наблюдения, в которых исключаются из выборки. По полученной выборке оцените функцию дожития и интегральную функцию риска теми же способами, рассчитайте выборочную медиану. Сравните полученные оценки с истинным распределением.

Ниже представлен код для исключения наблюдений из выборки и подсчета оценок функции дожития Каплана-Майера, функции интегрального риска Нельсона-Аалена, выборочной медианы и построения графиков (смотри Рисунок 5 и Рисунок 6).

**drop if delta < 1**

**stset t, fail(delta)**

**sts gen surv = s**

**sts gen haz\_cum\_2 = na**

**twoway (line surv \_t, connect(stairstep) sort lcolor(black))||function True\_S=exp(-0.07\*(x)^(0.7)), range(0 80)**

**twoway (line haz\_cum\_2 \_t, connect(stairstep) sort lcolor(black))||function True\_H=0.07\*(x)^(0.7), range(0 80)**

**stsum //8.687847**

**display (-(ln(0.5)/0.07))^(1/0.7) //26.452562**

Рисунок 5: Сопоставление оценки функции дожития Каплана-Майера в случае усечения справа и истинной функции дожития

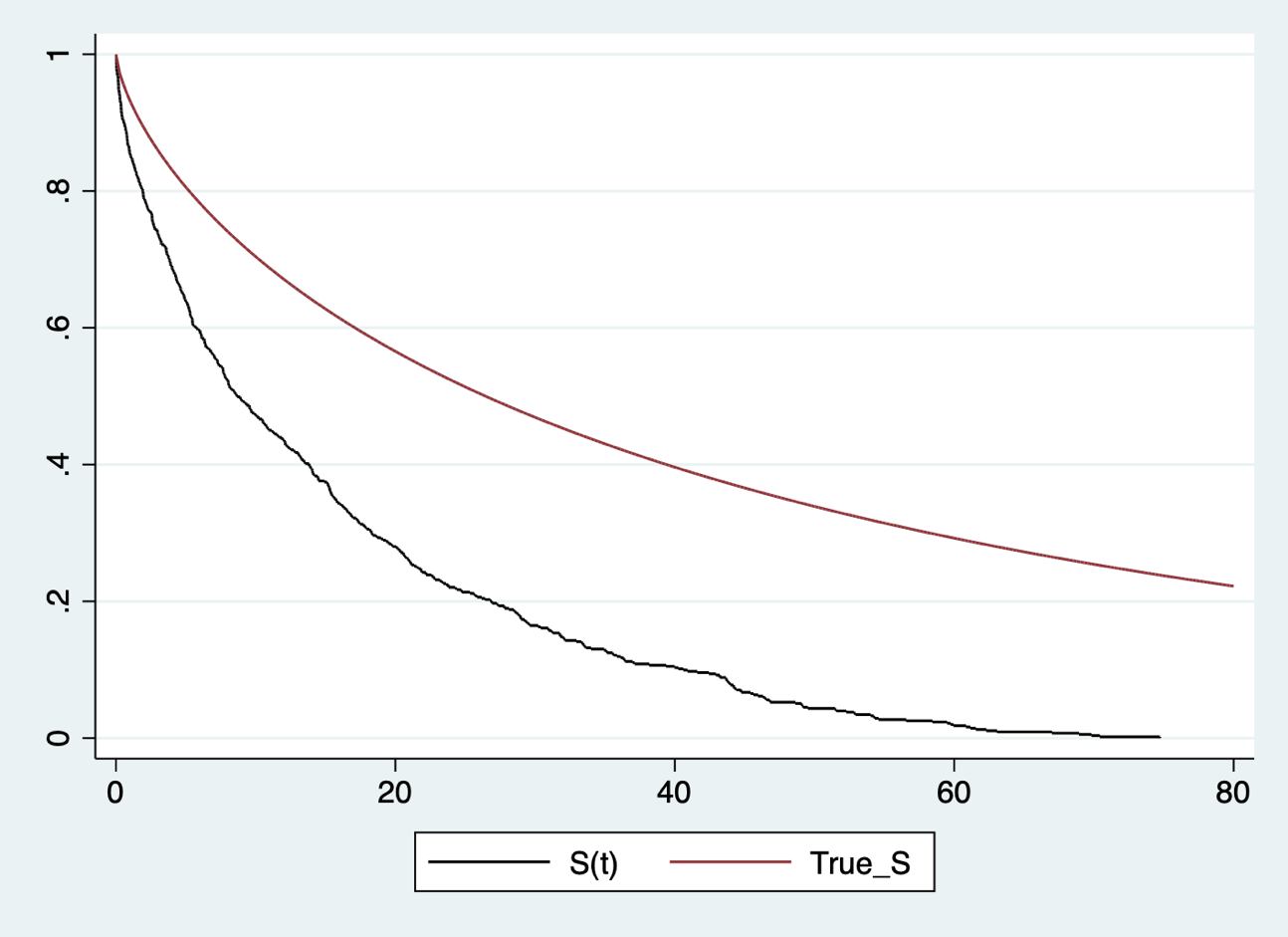
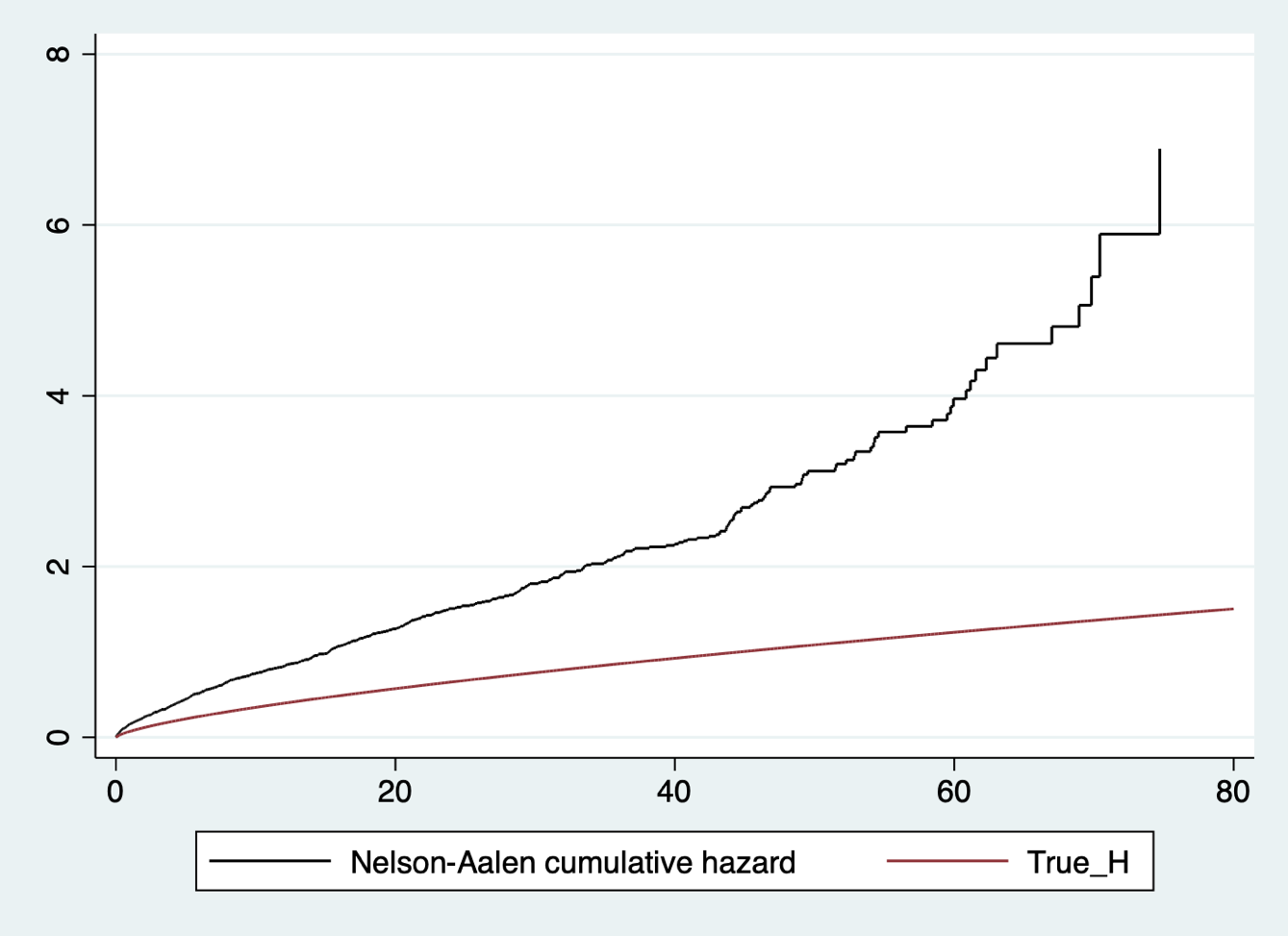


Рисунок 6: Сопоставление оценки интегральной функции риска Нельсона-Аалена в случае усечения справа и истинной функции интегрального риска



Как видно из Рисунков 5 и 6 распределения сильно отличаются как по функции дожития, так и по функции интегрального риска. Причем оценка функции дожития Каплана-Майера значительно ниже истинной и заканчивается до 80 значением в нуле (то есть все наблюдения закончились немного ранее 80). При этом и медиана длительностей тоже значительно ниже истинного (меньше в 3 раза и равна **8.687847**). Также оценка Нельсона-Аалена интегральной функции риска значительно выше истинной.

Таким образом, оценки Каплана-Майера и Нельсона-Аалена для усеченной справа выборки дают неточные оценки, значительно отличающиеся от истинного распределения.