МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Київський національний університет імені Тараса Шевченка
Механіко-математичний факультет
Кафедра теорії ймовірностей, статистики та актуарної математики

На правах рукопису

Використання узагальненої моделі виживання Кокса із залежним часом

Напрям підготовки: Статистика
Спеціалізація: Актуарна математика

Кваліфікаційна робота бакалавра
студента 4 курсу,
Сікорського Дениса Андрійовича

Науковий керівник:
доктор фіз.-мат. наук, професор
Ростислав Євгенович Ямненко

Робота заслухана на засіданні кафедри
та рекомендована до захисту на ЕК, протокол № від 2022р.

Зав. кафедрою проф. Мішура

Київ 2022

ВИТЯГ

з протоколу №_____

засідання екзаменаційної комісії №_____

Визнати, що студент Сікорський Денис Андрійович виконав та захистив кваліфікаційну роботу бакалавра з оцінкою_____.

Голова ЕК _____

«__»_____2022

Курсова робота Сікорський Денис

Мета: використання узагальненої моделі Кокса із залежним часом.

1. Теорія

Аналіз виживання

$$S(t)=P(T>t)$$

Іншими словами, $S(t)$ відповідає ймовірності виживання через час t. Тут, T відповідає випадковому часу життя, взятому з популяції. Зауважте, що $S(t)$ находиться від нуля до одиниці (включно), і $S(t)$ є спадною функцією t.

Функція небезпеки

$$h(t)=\lim_{\delta t \to 0} \frac{Pr(t \le T \le t+\delta t \vee T>t)}{\delta t}$$

Ліміт ΔT наближається до нуля, що означає, що наша мета — виміряти ризик події, що відбудеться в певний момент часу. Отже, взявши ліміт ΔT, маємо що наближення до нуля дає нескінченно малий проміжок часу. Тут слід зазначити, що небезпека не є ймовірністю. Це тому, що, навіть якщо ми маємо ймовірність у чисельнику, але ΔT у знаменнику може призвести до значення, більшого за одиницю.

Цензуровані дані

Узагальнені моделі виживання Кокса із залежним часом

Регресія виживання

Модель Кокса

$$h(t \vee x)=b_0(t)\exp\sum_{i=1}^{n} b_i(x_i)$$

- t представляє час виживання, небезпека може змінюватися з часом. - $h(t)$ функція небезпеки визначена набором з n коваріатів $(x_1,x_2,\ldots,x_n)$. - $b_0(t)$ є базовою функцією небезпеки, і вона визначається як ймовірність настання події, що цікавить, коли всі інші коваріати дорівнюють нулю. Це єдиний залежний від часу компонент у моделі. Модель не робить припущення щодо базової функції небезпеки і приймає параметричну

форму для впливу коваріатів на небезпеку. - $\exp\sum_{i=1}^{n} b_i\left(x_i\right)$ часткова небезпека – це незмінний у часі скалярний коефіцієнт, який лише збільшує або зменшує базову небезпеку. Він подібний до скаляру у звичайній регресії. Коваріати або коефіцієнти регресії x дають пропорційну зміну, яку можна очікувати в небезпеці. - коефіцієнти $\left(b_1, b_2, \ldots, b_n\right)$ визначають вплив (тобто розмір ефекту) коваріатів. Знак коефіцієнту регресії, b_i, відіграє певну роль у небезпеці суб'єкта. Зміна цих коефіцієнтів регресії або коваріатів призведе до збільшення або зменшення базової небезпеки. Позитивний знак для b_i означає, що ризик події вищий, а отже, ймовірність події, що представляє інтерес для цього конкретного суб'єкта, вища. Так само негативний знак означає, що ризик події нижчий. Також треба звернути увагу, що величина, тобто її розмір, також грає роль. Наприклад, якщо значення змінної дорівнює одиниці, це означатиме, що вона не матиме впливу на небезпеку. Якщо значення менше одиниці, це зменшить небезпеку, а значення більше одиниці — збільшить небезпеку. Ці коефіцієнти регресії, b, оцінюються шляхом максимізації часткової ймовірності. Узагальнені моделі виживання Кокса із залежним часом є напівпараметричною моделю у тому сенсі, що базову функцію небезпеки не потрібно вказувати, тобто вона може змінюватися, дозволяючи використовувати інший параметр для кожного унікального часу виживання. Але передбачається, що коефіцієнт залишається пропорційним протягом досліджуваного періоду. Це призводить до підвищення гнучкості моделі. Повністю параметрична пропорційна модель ризиків також передбачає, що базова функція небезпеки може бути параметризована відповідно до конкретної моделі для розподілу часу виживання. Модель Кокса може обробляти дані з правою цензурою, але не може обробляти дані з лівою або інтервальною цензурою безпосередньо.

2. Завантаження даних

Для цього завдання я вибрав дані які вже є в стандартній бібліотеці, а саме дані Rossi recidivism, який містить інформацію про повторні випадки скоєння злочинів. Спочатку завантажимо всі потрібні бібліотеки.

```
import numpy as np
import pandas as pd
import pandas_profiling
import matplotlib.pyplot as plt
import seaborn as sns

from lifelines import CoxPHFitter
from lifelines import KaplanMeierFitter
from lifelines.datasets import load_rossi

%matplotlib inline
```

Тепер можемо завантажити дані і зробити поверхневий огляд.

```
df = load_rossi()
df.head(10)
```

```
   week  arrest  fin  age  race  wexp  mar  paro  prio
0    20       1    0   27     1     0    0     1     3
1    17       1    0   18     1     0    0     1     8
2    25       1    0   19     0     1    0     1    13
3    52       0    1   23     1     1    1     1     1
4    52       0    0   19     0     1    0     1     3
5    52       0    0   24     1     1    0     0     2
6    23       1    0   25     1     1    1     1     0
7    52       0    1   21     1     1    0     1     4
8    52       0    0   22     1     0    0     0     6
9    52       0    0   20     1     1    0     0     0
```

Перша колонка відповідає за повторне ув'язнення після звільнення, або про припинення нагляду якщо значення рівне 52. Відповідно у другій колонці 1 - людина була заарештована до 52 тижня, 0 - не була. fin - чи була людині надана фінансова допомога, 0 - ні, 1 - так. Наступна колонка показує вік у якому людину звільнили з під варти. race показує була людина афроамериканцем чи ні. wexp - чи була у людини робота на повну ставку до ув'язнення. mar - чи була людина одружена у час звільнення, paro - чи була людина звільнена достроково. prio - кількість судимостей до останнього ув'язнення.

```
df.describe()
```

```
             week      arrest         fin         age        race
wexp  \
count  432.000000  432.000000  432.00000  432.000000  432.000000
432.000000
mean    45.854167    0.263889    0.50000   24.597222    0.877315
0.571759
std     12.662293    0.441251    0.50058    6.113375    0.328456
0.495398
min      1.000000    0.000000    0.00000   17.000000    0.000000
0.000000
25%     50.000000    0.000000    0.00000   20.000000    1.000000
0.000000
50%     52.000000    0.000000    0.50000   23.000000    1.000000
1.000000
75%     52.000000    1.000000    1.00000   27.000000    1.000000
1.000000
max     52.000000    1.000000    1.00000   44.000000    1.000000
1.000000

              mar        paro        prio
count  432.000000  432.000000  432.000000
```

```
mean     0.122685    0.618056    2.983796
std      0.328456    0.486426    2.896068
min      0.000000    0.000000    0.000000
25%      0.000000    0.000000    1.000000
50%      0.000000    1.000000    2.000000
75%      0.000000    1.000000    4.000000
max      1.000000    1.000000   18.000000
```

Як ми бачимо, у нашому датасеті більшість людей не скоює ніяких злочинів після звільнення, що видно і з колонки week, і з колонки arrest. 50 відсотків людей отримують фінансову допомогу, а вік людей є рівномірно розподіленим з середнім у районі 25 років. Лише 13 відсотків людей не є афроамериканцями та більшість людей мала повноцінну роботу до в'язниці. Більша кількість людей вийшла достроково, але дуже мало людей були одруженими після в'язниці. Більше 75 відсотків людей вже скоювали злочини до випадку нашого досліду.

3. Створення та аналіз моделі

Завантажимо модель і перевіримо результати, перед цим подівливші дані на train і test.

```
train_df = df.iloc[:400]
test_df = df.iloc[400:]

model = CoxPHFitter()
model.fit(train_df, duration_col='week', event_col='arrest')
print(model.score(train_df))
print(model.score(test_df))
model.print_summary()
```

```
-1.4689456583579346
-1.076857310695416

<lifelines.CoxPHFitter: fitted with 400 total observations, 297 right-
censored observations>
             duration col = 'week'
                event col = 'arrest'
      baseline estimation = breslow
   number of observations = 400
number of events observed = 103
   partial log-likelihood = -587.58
         time fit was run = 2022-05-23 06:45:52 UTC

---
            coef  exp(coef)   se(coef)   coef lower 95%   coef upper
95%  exp(coef) lower 95%  exp(coef) upper 95%
covariate

fin        -0.41       0.67       0.20            -0.80            -
```

```
0.01                 0.45                 0.99
age        -0.05       0.95       0.02           -0.09            -
0.00                 0.91                 1.00
race        0.32       1.38       0.32           -0.31
0.95                 0.73                 2.59
wexp       -0.20       0.82       0.22           -0.64
0.24                 0.53                 1.27
mar        -0.45       0.64       0.41           -1.26
0.35                 0.28                 1.42
paro       -0.04       0.96       0.21           -0.44
0.36                 0.64                 1.44
prio        0.09       1.10       0.03            0.04
0.15                 1.04                 1.16

             cmp to     z      p   -log2(p)
covariate
fin            0.00 -2.02   0.04       4.51
age            0.00 -2.11   0.03       4.85
race           0.00  0.99   0.32       1.63
wexp           0.00 -0.91   0.36       1.46
mar            0.00 -1.10   0.27       1.88
paro           0.00 -0.19   0.85       0.24
prio           0.00  3.23 <0.005       9.67
---
Concordance = 0.64
Partial AIC = 1189.16
log-likelihood ratio test = 30.24 on 7 df
-log2(p) of ll-ratio test = 13.51
```

Як ми бачимо, всі коефіцієнти прийняли логічне значення, дійсно при отриманні фінансової допомоги зменшується потреба в незаконому отримані грошей, чим старша людина, тем менше їй потрапити знову до в'язниці. Також наявність роботи та особливо другої половинки сильно зменшує вірогідність знову потрапити до в'язниці. Логічно, що чим більше людина була у в'язниці тим більше може туди потрапити ще раз, оскільки вона вже звикла скоювати злочини і просто живе цим. Також модель каже, що афроамериканці частіше будуть повторно скоювати злочини, ніж інші раси. Тепер спробуємо покращити модель і додати регуляризацію.

```
model = CoxPHFitter(penalizer=0.1, l1_ratio=0.005)
model.fit(train_df, duration_col='week', event_col='arrest')
print(model.score(train_df))
print(model.score(test_df))
model.print_summary()
```

```
-1.4713620138252017
-1.087131735553136
```

```
<lifelines.CoxPHFitter: fitted with 400 total observations, 297 right-
censored observations>
             duration col = 'week'
                event col = 'arrest'
                penalizer = 0.1
                  l1 ratio = 0.005
      baseline estimation = breslow
   number of observations = 400
number of events observed = 103
   partial log-likelihood = -591.31
         time fit was run = 2022-05-23 06:46:16 UTC

---
            coef  exp(coef)   se(coef)   coef lower 95%   coef upper
95%  exp(coef) lower 95%  exp(coef) upper 95%
covariate

fin        -0.28       0.75       0.17            -0.62
0.05                 0.54                    1.05
age        -0.03       0.97       0.02            -0.06
0.00                 0.94                    1.00
race        0.20       1.22       0.26            -0.31
0.71                 0.73                    2.04
wexp       -0.23       0.79       0.18            -0.58
0.12                 0.56                    1.13
mar        -0.32       0.73       0.30            -0.90
0.27                 0.41                    1.31
paro       -0.03       0.97       0.17            -0.37
0.31                 0.69                    1.37
prio        0.07       1.08       0.03             0.02
0.12                 1.02                    1.13

            cmp to     z    p   -log2(p)
covariate
fin           0.00 -1.68 0.09       3.44
age           0.00 -1.92 0.06       4.18
race          0.00  0.77 0.44       1.17
wexp          0.00 -1.29 0.20       2.34
mar           0.00 -1.06 0.29       1.80
paro          0.00 -0.16 0.87       0.19
prio          0.00  2.78 0.01       7.54
---
Concordance = 0.65
Partial AIC = 1196.63
log-likelihood ratio test = 22.76 on 7 df
-log2(p) of ll-ratio test = 9.06
```

Як ми бачимо, регуляризація не бала бажаного ефекту, проте коефіцієнти одні й ті самі. Спробуємо змінити формулу і прибрати колонки з незначними коефіцієнтами, а також додати нові.

```
model.fit(train_df, duration_col='week', event_col='arrest',
formula="fin + wexp + mar + age * prio")
print(model.score(train_df))
print(model.score(test_df))
model.print_summary()
```

```
-1.4727230702086451
-1.0886336493330575
```

```
<lifelines.CoxPHFitter: fitted with 400 total observations, 297 right-
censored observations>
             duration col = 'week'
                event col = 'arrest'
                penalizer = 0.1
                  l1 ratio = 0.005
      baseline estimation = breslow
   number of observations = 400
number of events observed = 103
   partial log-likelihood = -591.56
         time fit was run = 2022-05-23 06:46:29 UTC

---
            coef  exp(coef)   se(coef)   coef lower 95%   coef upper
95%  exp(coef) lower 95%  exp(coef) upper 95%
covariate

age        -0.03       0.97       0.02            -0.06
0.00                 0.94                  1.00
fin        -0.28       0.76       0.17            -0.61
0.05                 0.54                  1.06
mar        -0.34       0.72       0.29            -0.91
0.24                 0.40                  1.28
prio        0.06       1.06       0.04            -0.01
0.14                 0.99                  1.15
wexp       -0.23       0.80       0.18            -0.58
0.12                 0.56                  1.13
age:prio    0.00       1.00       0.00            -0.00
0.00                 1.00                  1.00

            cmp to     z    p   -log2(p)
covariate
age           0.00 -1.93 0.05       4.21
fin           0.00 -1.64 0.10       3.31
mar           0.00 -1.14 0.26       1.97
prio          0.00  1.61 0.11       3.21
wexp          0.00 -1.28 0.20       2.31
age:prio      0.00  0.37 0.71       0.49
---
Concordance = 0.64
Partial AIC = 1195.12
```

```
log-likelihood ratio test = 22.27 on 6 df
-log2(p) of ll-ratio test = 9.85
```

Як ми бачимо, новий коефіцієнт не дав бажаного ефекту. Найкращою була перша модель.

4. Аналіз базових факторів

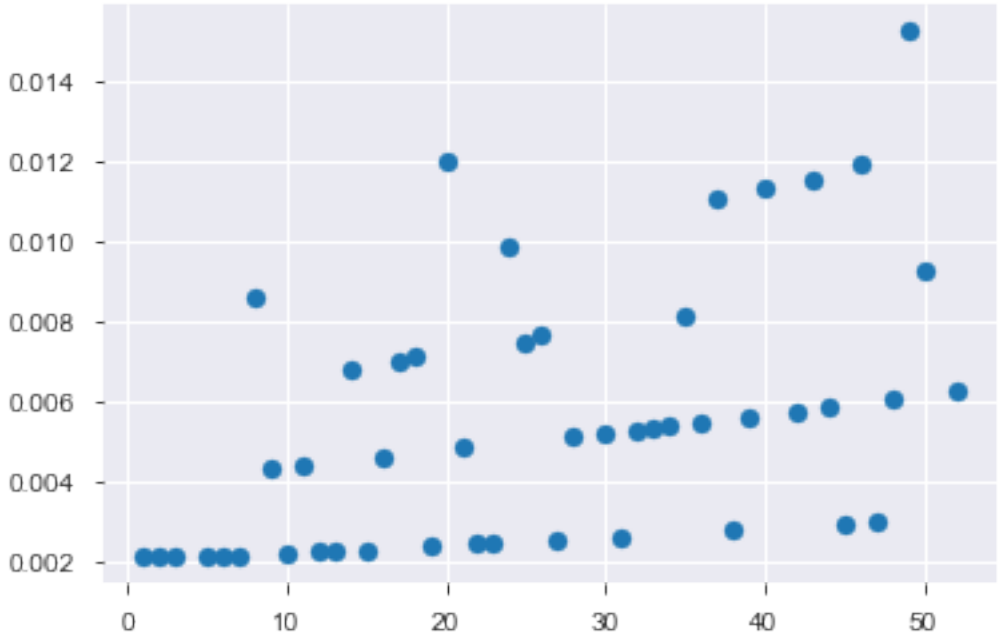
Спочатку подивимось на базову функцію небезпеки.

```
model = CoxPHFitter()
model.fit(train_df, duration_col='week', event_col='arrest')
baseline_hazard = model.baseline_hazard_
baseline_hazard
```

```
      baseline hazard
1.0          0.002118
2.0          0.002124
3.0          0.002126
5.0          0.002131
6.0          0.002140
7.0          0.002153
8.0          0.008626
9.0          0.004370
10.0         0.002192
11.0         0.004432
12.0         0.002264
13.0         0.002267
14.0         0.006831
15.0         0.002307
16.0         0.004630
17.0         0.006975
18.0         0.007110
19.0         0.002398
20.0         0.012027
21.0         0.004868
22.0         0.002446
23.0         0.002461
24.0         0.009858
25.0         0.007456
26.0         0.007647
27.0         0.002566
28.0         0.005151
30.0         0.005205
31.0         0.002620
32.0         0.005267
33.0         0.005308
34.0         0.005389
35.0         0.008159
36.0         0.005484
37.0         0.011045
```

```
38.0         0.002800
39.0         0.005626
40.0         0.011309
42.0         0.005758
43.0         0.011555
44.0         0.005865
45.0         0.002955
46.0         0.011899
47.0         0.003009
48.0         0.006048
49.0         0.015276
50.0         0.009300
52.0         0.006302
```

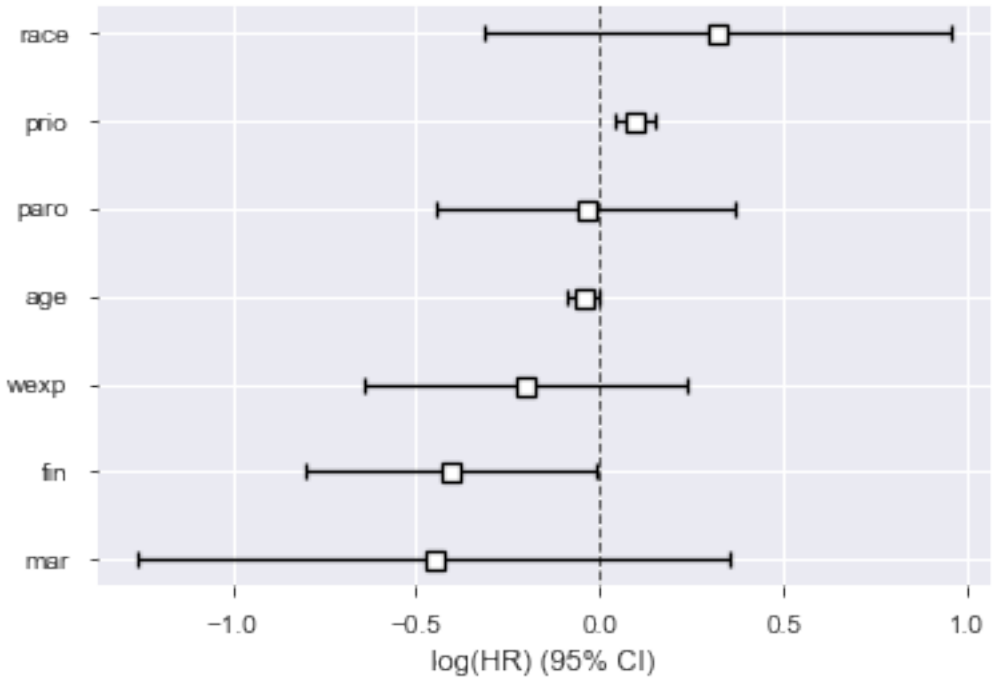
```
plt.scatter(x = baseline_hazard.index, y = baseline_hazard['baseline
hazard']);
```



Як ми бачимо, у перші декілька тижнів ймовірність скоєння злочину дуже низька, проте чим далі, тим більше такий шанс (що логічно).

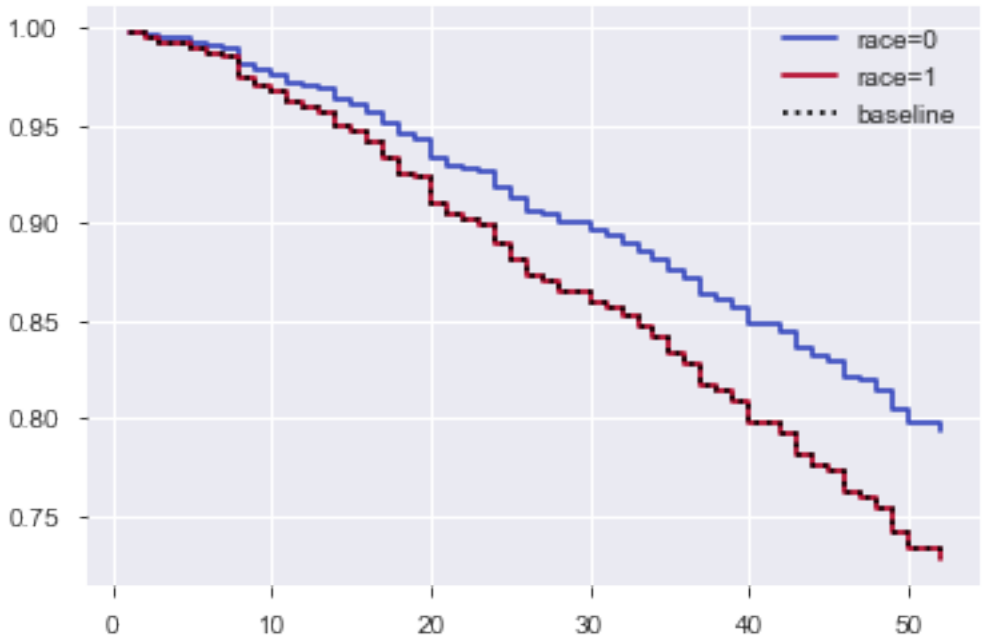
Тепер проаналізуємо коефіцієнти які ми отримали для кожної змінної.

```
model.plot();
```



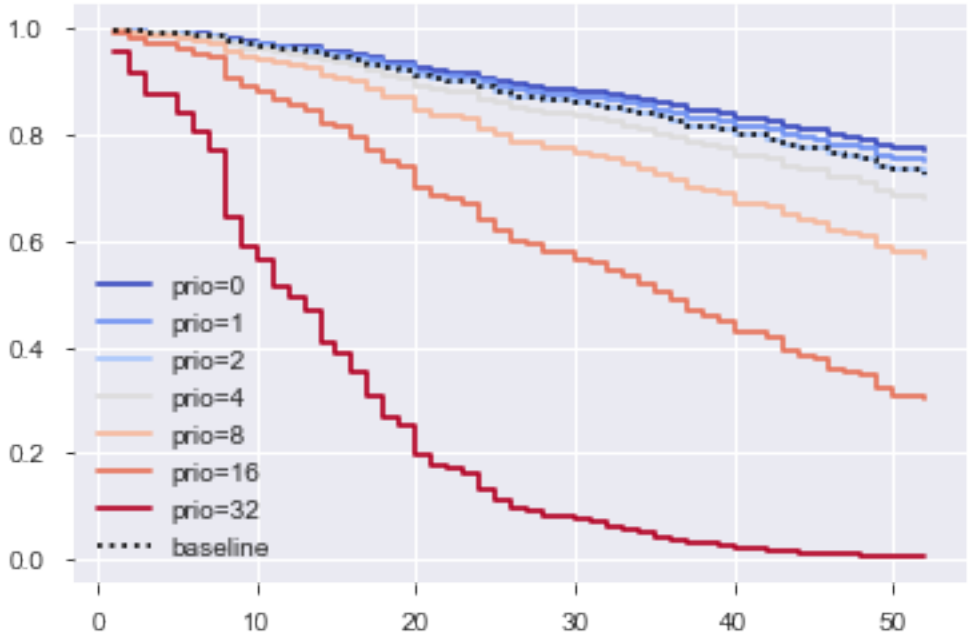
З коефіцієнтів однозначно можна зробити висновки, що кількість минулих злочинів дійсно впливає на ймовірність рецидиву, зрілість людей зменшує шанс рецидиву. На 100 відсотків можна стверджувати, що наявність фінансової допомоги знижиує ризик рецидиву. Наявність роботи та сім'ї також скоріш за все зменшує вірогідність рецидиву. Проте приналежність до афроамериканців скоріш за все підвищує шанс повторного скоєння злочинів. І в кінці дострокове звільнення майже не впливає на ймовність повторного скоєння злочинів. Тепер подивимось на функцію небезпеки в залежності від різних показників.

```
model.plot_partial_effects_on_outcome(covariates='race', values=[0,
1], cmap='coolwarm');
```



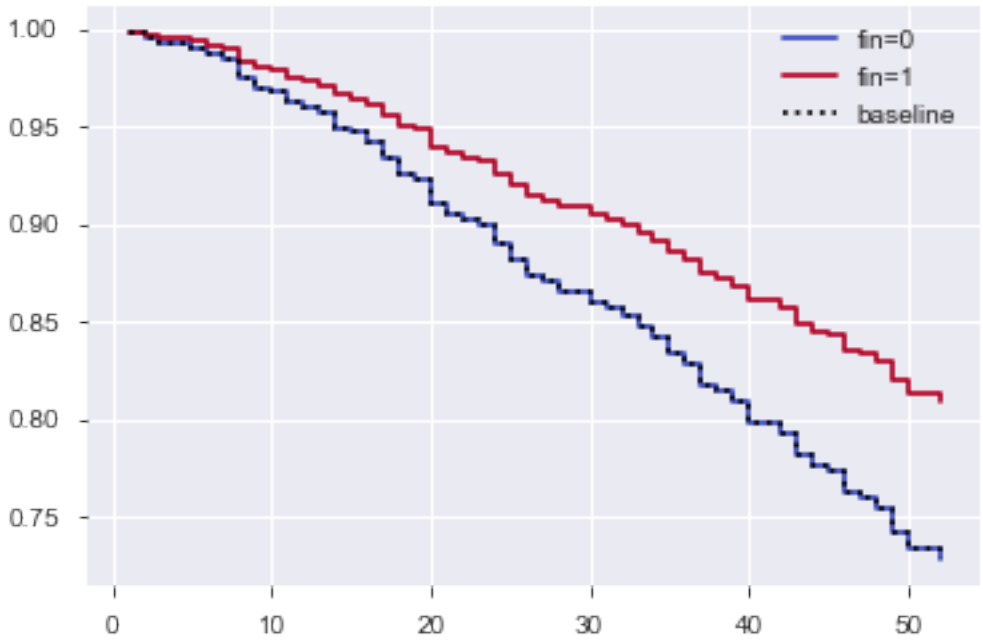
З графіку видно, що раса має деякий впилив, проте він є таким великим, як здавалось.

```
model.plot_partial_effects_on_outcome(covariates='prio', values=[0, 1,
2, 4, 8, 16, 32], cmap='coolwarm');
```



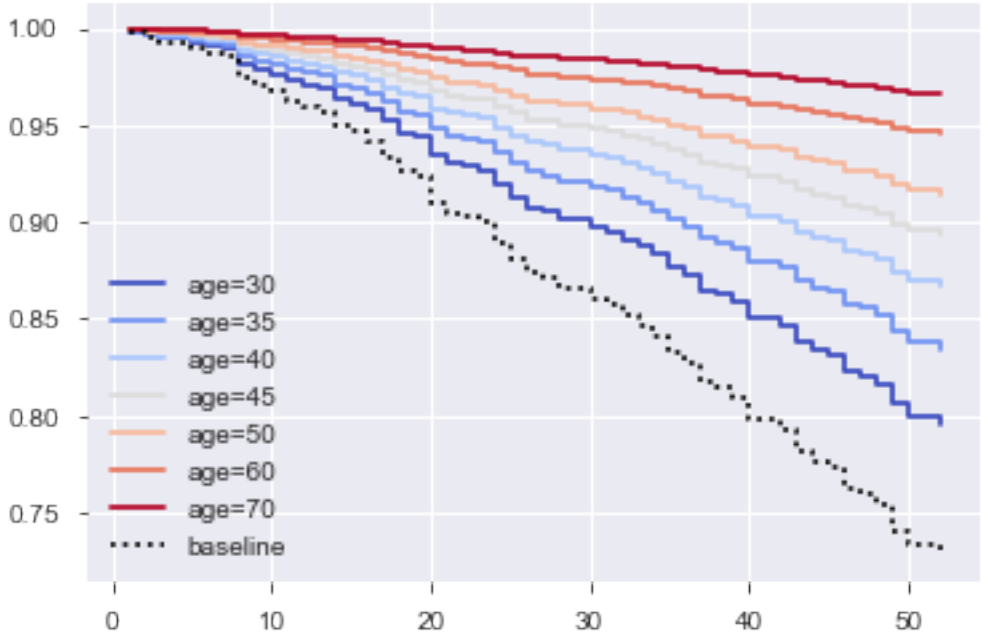
Кількість минулих ув'язнень має величезний вплив на майбутні правопорушення, бачимо люди з великою кількістю ув'язнень майже напевно порушать закон ще один раз впродовж 52 тижнів.

```
model.plot_partial_effects_on_outcome(covariates='fin', values=[0, 1],
cmap='coolwarm');
```



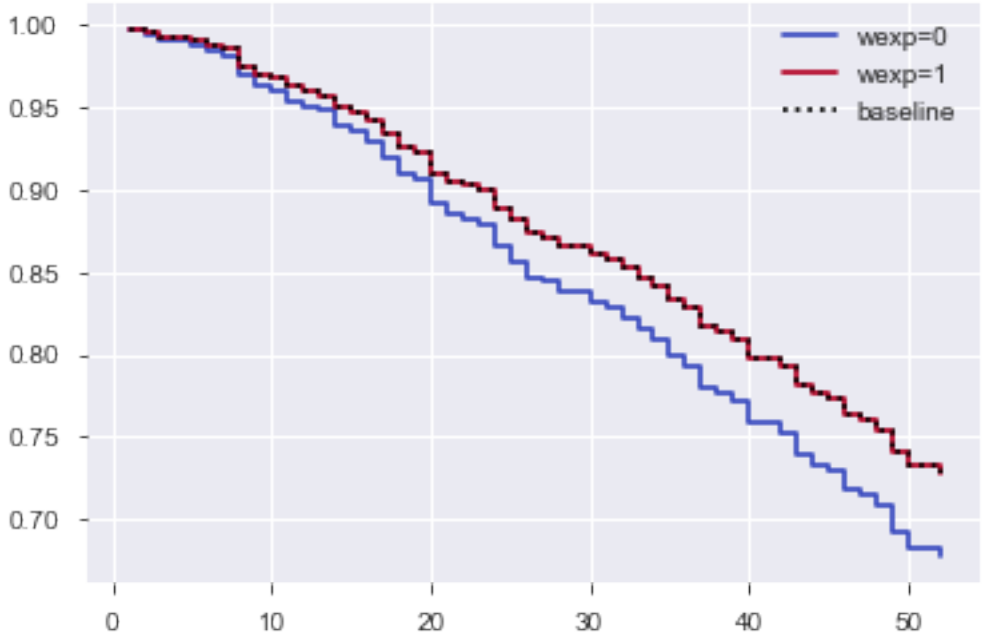
З графіку дуже добре видно, що фінансова допомога дає результат, проте не дуже великий, але це дає змогу розрахувати фінансову правильність цього експерименту.

```
model.plot_partial_effects_on_outcome(covariates='age', values=[30,
35, 40, 45, 50, 60, 70], cmap='coolwarm');
```



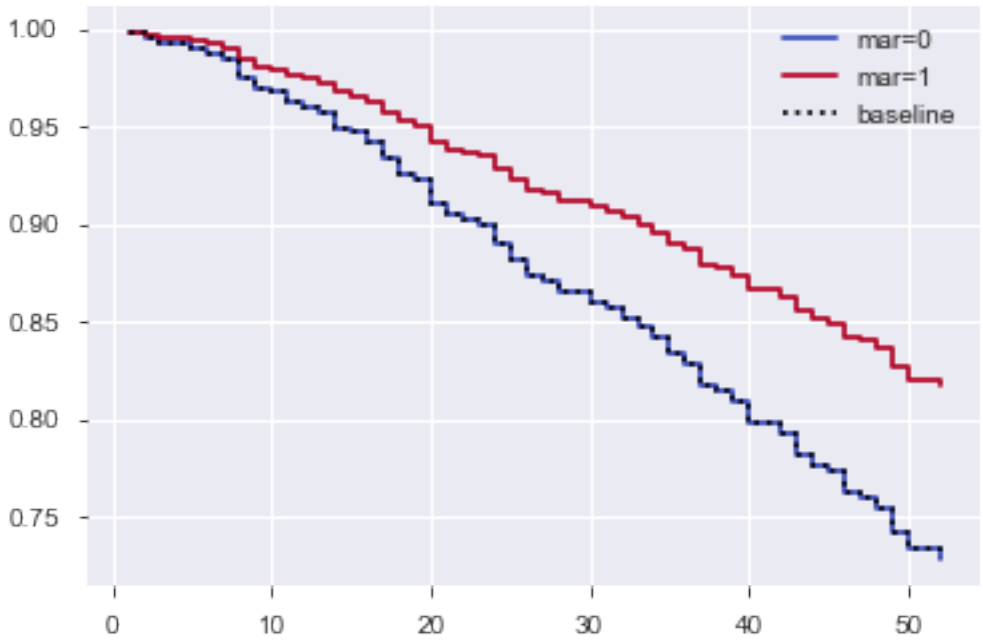
З цього графіку добре видно підтвердження, що чим старша людина, тим менший шанс повторного арешту.

```
model.plot_partial_effects_on_outcome(covariates='wexp', values=[0,
1], cmap='coolwarm');
```



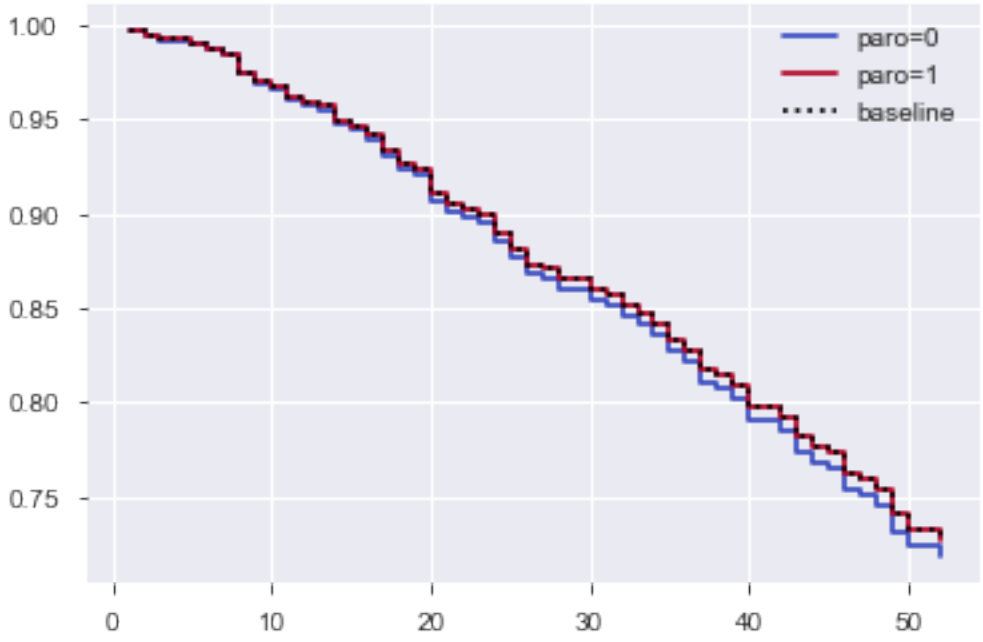
Тут також видно незначний вплив наявності роботи на момент виходу з в'язниці.

```
model.plot_partial_effects_on_outcome(covariates='mar', values=[0, 1],
cmap='coolwarm');
```



Вже більше впливу показує наявність другою половинки на момент звільнення.

```
model.plot_partial_effects_on_outcome(covariates='paro', values=[0,
1], cmap='coolwarm');
```



Єдина змінна, яка не впливає, що випуск в'язня достроково.

5. Висновок

З нашого дослідження видно, що ймовірність повторного правопорушення з боку однієї людини може бути зменшеною за допомогою зовнішніх факторів, на які органи влади мають вплив. Основим фактором є фінансова допомога, яка дала зменшення майже на 10 відсотків. Також можна зробити висновок, що наявність роботи має значний вплив, тому можливо треба робити умовою виходу влаштування на постійну роботу. Варто зазначити, що дострокове ув'язнення не має жодного впливу на ймовірність рецидиву, тому його слід поєднати з працевлаштуванням. Інші фактори, які мають вплив це факт одруження, вік, кількість правопорушень та раса. Можливо, треба зменшувати кількість років ув'язнення одруженим та старим людям, бо вони з меншою ймовірністю скоюють повторні злочини. Також, я би радив давати пожиттєве ув'язнення рецидивістам, оскільки в'язниця не має жодного впливу на них. Дуже цікаве питання стоїть з афроамериканцями, проте різниця не дуже велика, тому тут треба зробити більш детальний аналіз з більшою кількістю даних.

6. Джерела

Матеріали: NCSS User's Guide V Автор: Dr. Jerry L. Hintze Видавництво: NCSS Посилання: https://www.ncss.com/wp-content/uploads/2012/09/NCSSUG5.pdf
Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model Автори: Terry Therneau, Cynthia Crowson, Elizabeth Atkinson Посилання: https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf Дані були взяті з сайту: https://rdrr.io/cran/RcmdrPlugin.survival/man/Rossi.html