Звіт

до лабораторної роботи №3

з дисципліни “Аналіз даних”

на тему “Регресійний аналіз”

Роботу виконали студенти групи КМ-01:

Бабич Ірина

Іваник Юрій

Романецький Микита

Суховій Ігор

Шолоп Любомир

Київ – 2023

Основна частина

1. Мотивація проведення дослідження

Дослідити вплив факторів на зміну Total.Charges, зокрема довжини перебування, віку, раси, статі, важкості захворювання

Гіпотези, щодо знаків коефіцієнтів структурної моделі:

1) Коефіцієнт біля кількості днів перебування додатній (чим більше тим дорожче)

2) Коефіцієнт біля Gender додатній (якщо чоловік, то йому більше прийдеться заплатити) або дуже малий і незначний (подумати треба)

3) Коефіцієнт біля важкості хвороби додатній

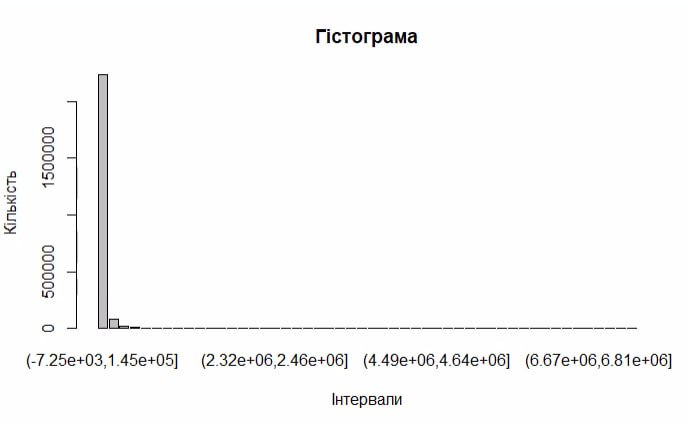


Рис. 1 – Гістограма ціни

Гістограма є «скошеною», отже потрібно використовувати логартим для цієї змінної.

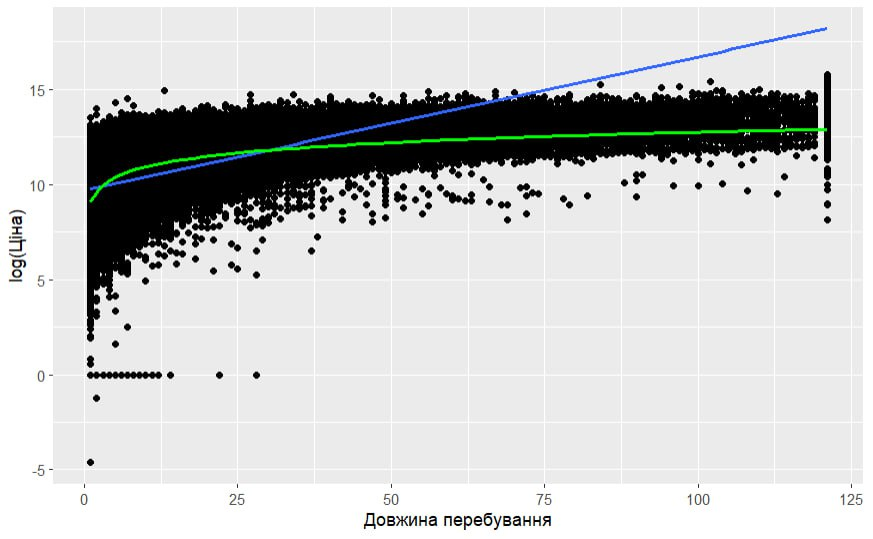


Рис. 2 – Графік залежністі ціни довжини перебування

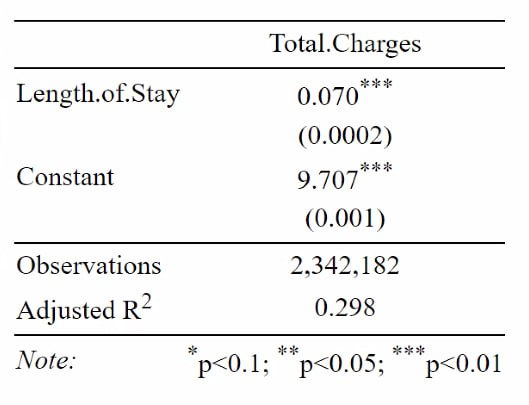


Рис. – Базова модель

Довірчі інтервали для Total.Charges

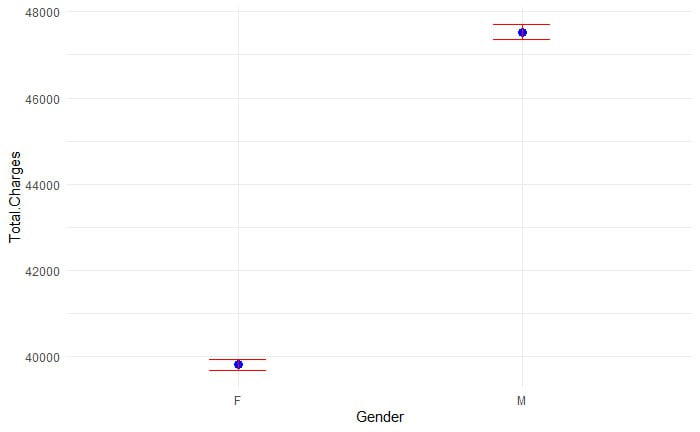


Рис. – Довірчі інтервали для ціни в залежності від статі

В нас Gender це 1 - жінки, 0 чоловіки. Середнє для жінок менше ніж середнє для чоловіків. Отже, якщо наш регресор Gender буде 1, то це буде негативно впливати на модель, тобто коефіцієнт буде від'ємний.

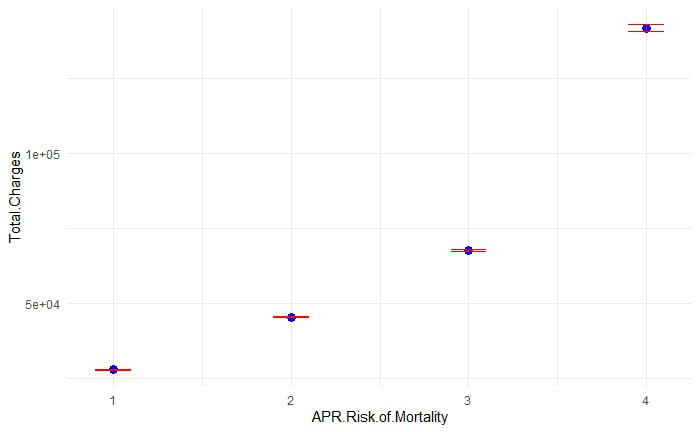


Рис. – Довірчі інтервали для ціни в залежності від ризику смертності

У порівняння з ризиком 1, значення коефіцієнтів 2, 3, 4 будуть більші (чим більший ризик, тим більше тобі прийдеться платити). Щодо знаку коефіцієнта - вона буде у всіх додатня

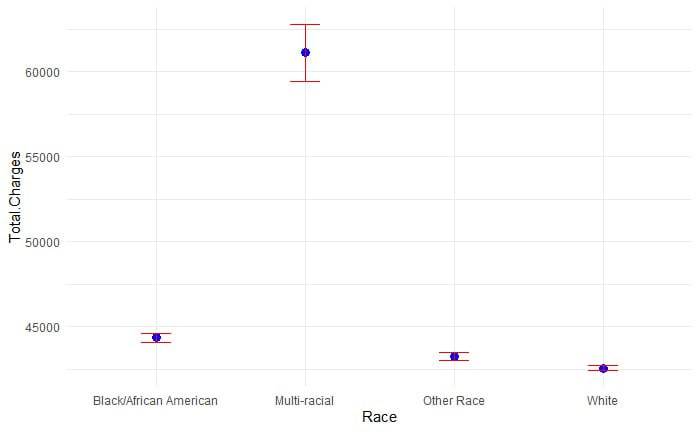


Рис. – Довірчі інтервали для ціни в залежності від раси

Оскільки людей білої раси найбільше в лікарні, вони базові (тобто від них відштовхуємося у висновках). Виходить що білі платять найменше, отже знак коефіцієнта в регресії для інших раз буде додатнім. Щодо абсолютного значення, то найбільше воно буде для мультирас.

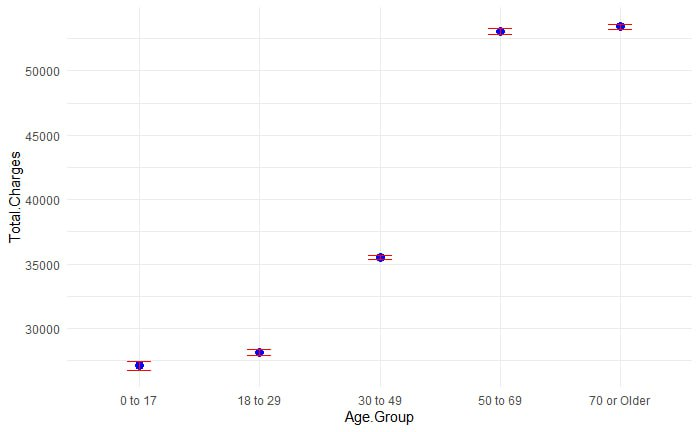


Рис. – Довірчі інтервали для ціни в залежності від вікової групи

За основу тут візьмемо 50-69 і 70+. Для інших вікових категорій тоді значення коефіцієнта буде від'ємне (тобто ранній вік негативно впливає на величину оплати)

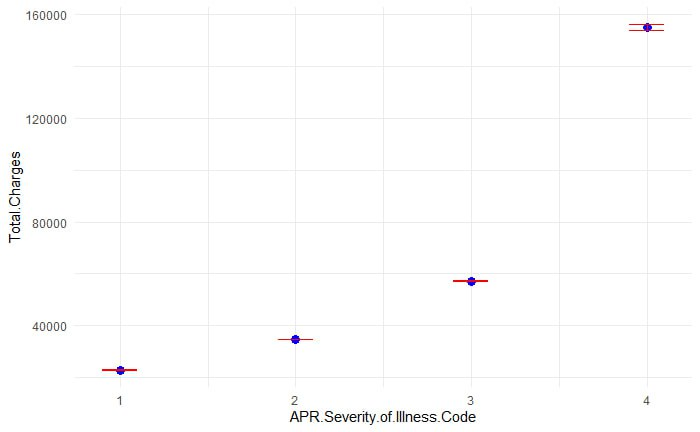


Рис. – Довірчі інтервали для ціни в залежності від хвороби

Базова буде друга важкість, оскільки її найбільше. Отже перша важкість буде мати від'ємний коефіцієнт, 3 і 4 – додатній.

Кореляції

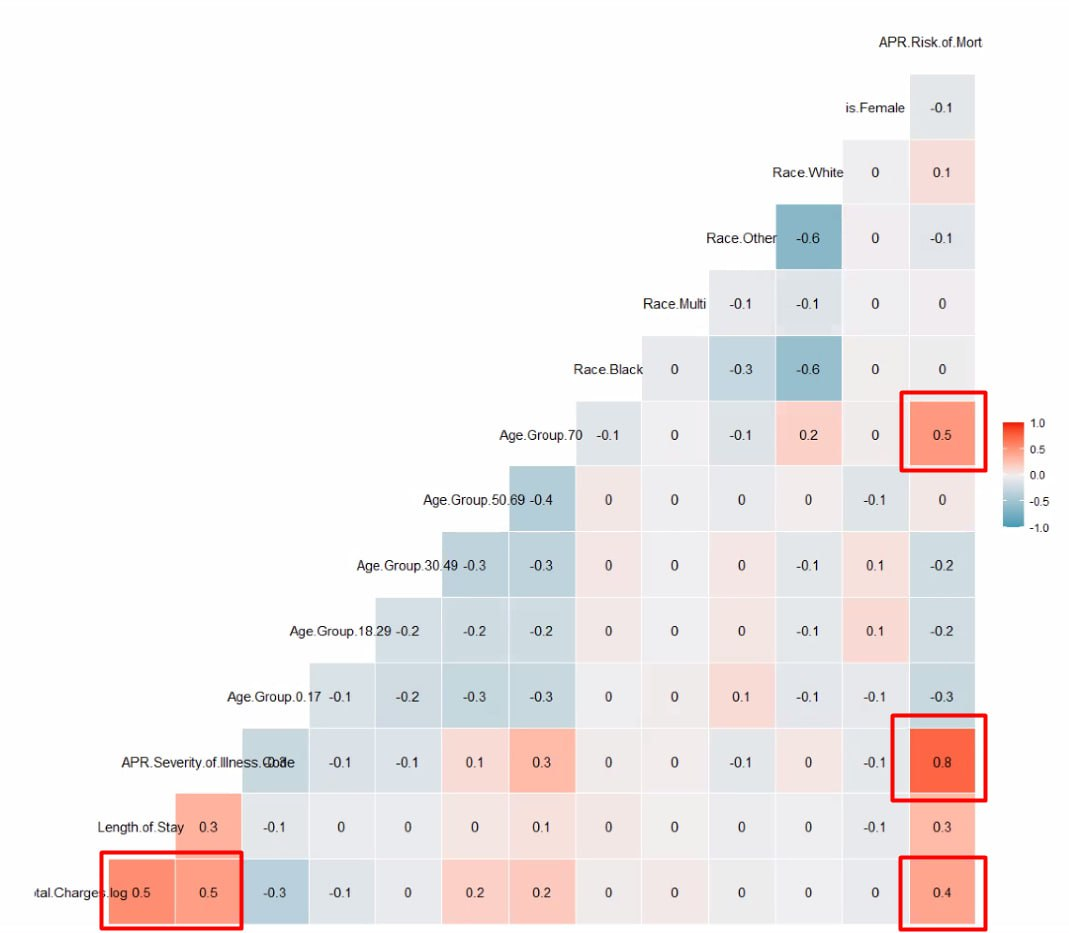


Рис. – Корелограма

Рис. – Базова модель

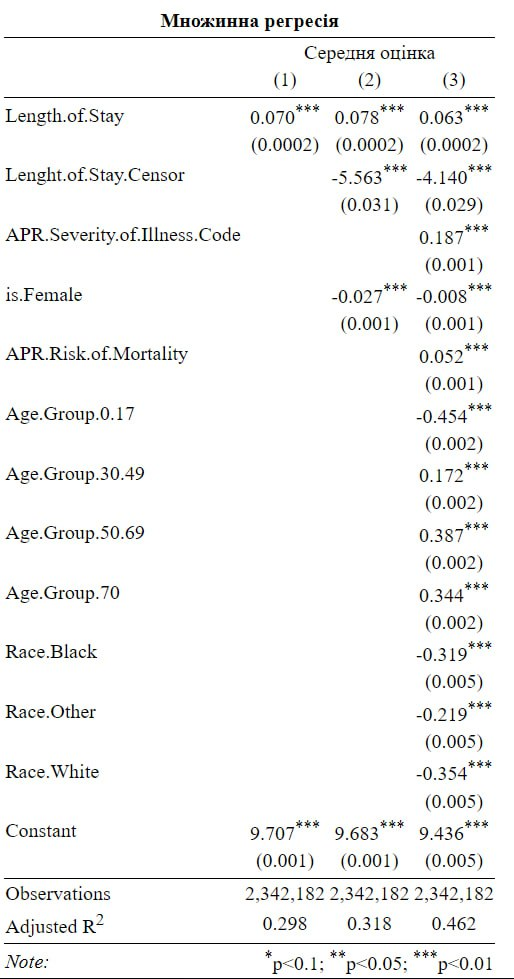


Рис. – картинка токена

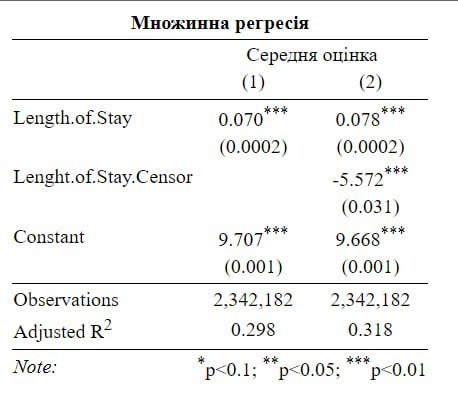


Рис. - модель з довжиною перебування

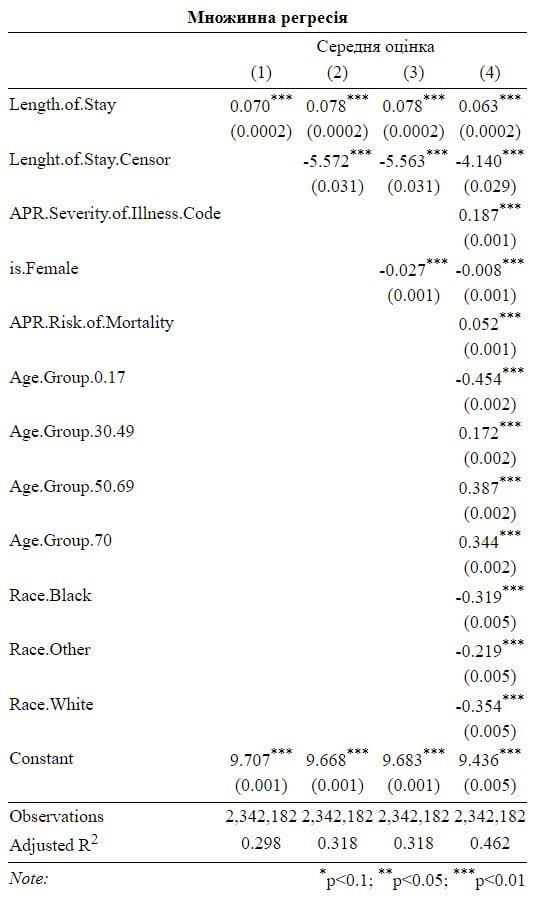


Рис. – Модель з контрольними змінними.

Цензуровані дані впливають, стать – ні.

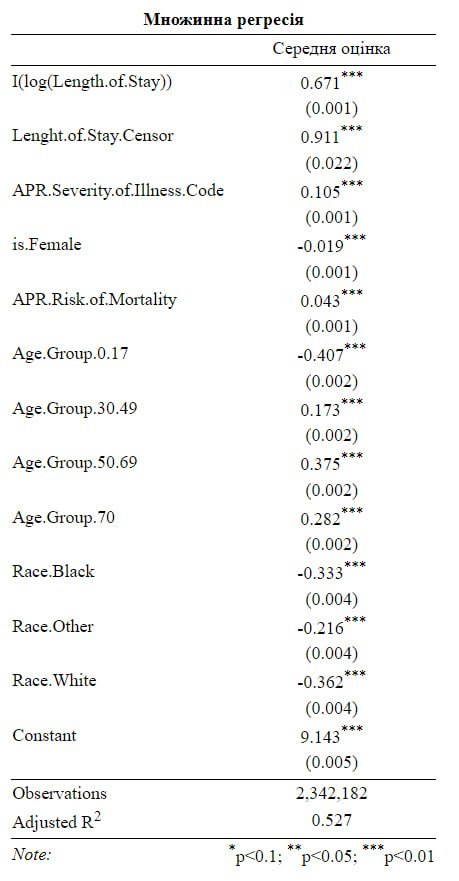


Рис. - модель з логарифмом довжини перебування

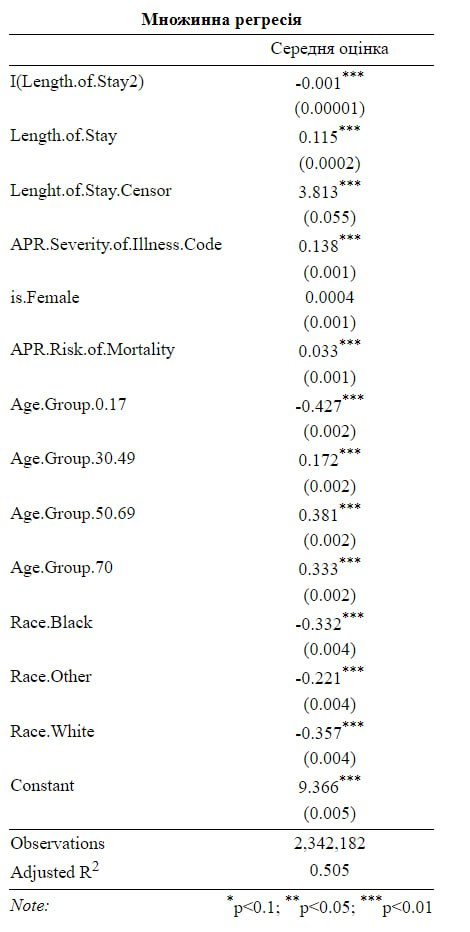
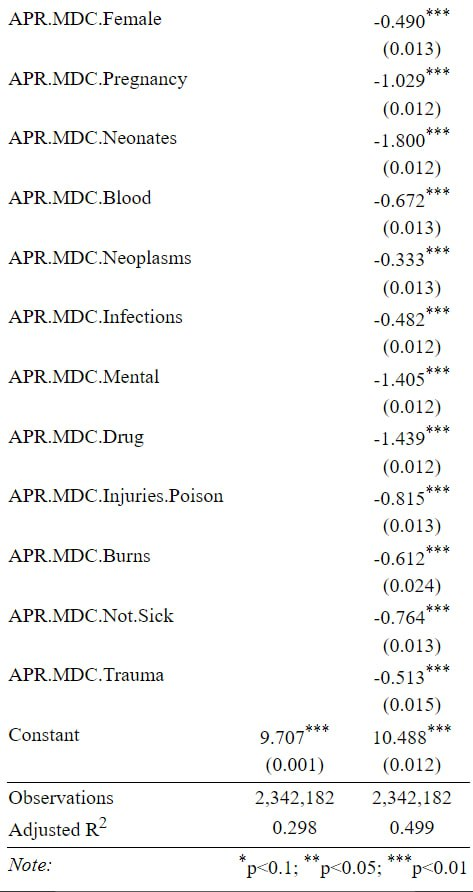
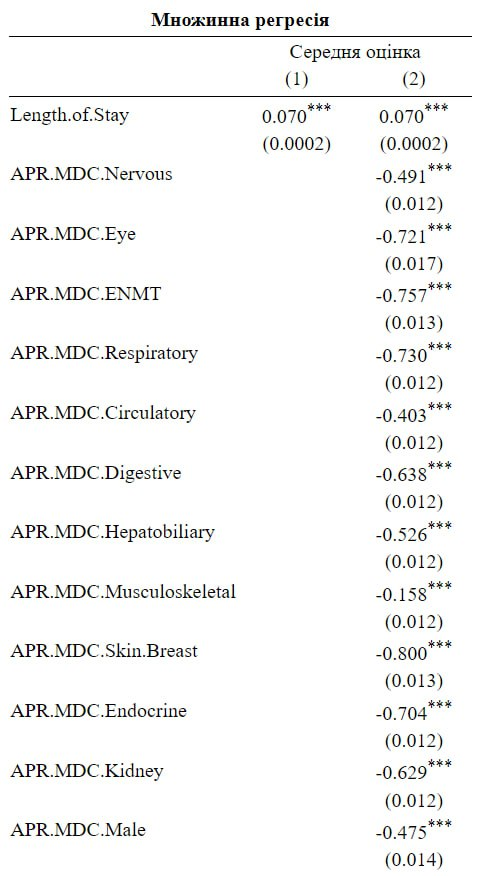
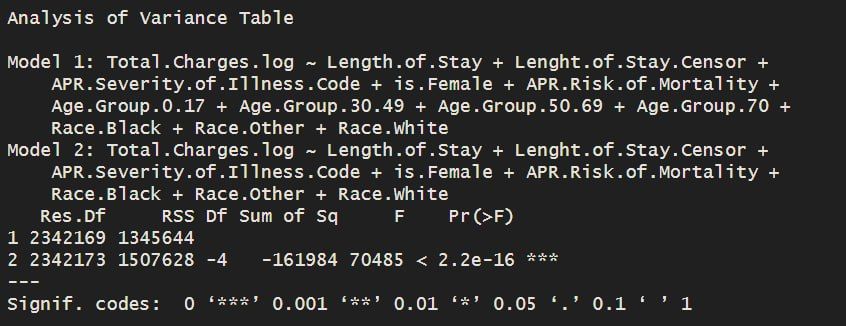


Рис. - Модель з квадратом логарифму

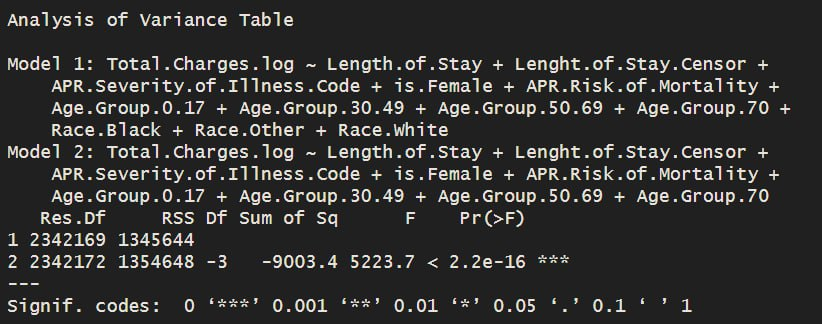


Дослідження значущісті групи коефіцієнтів. 

Вік.

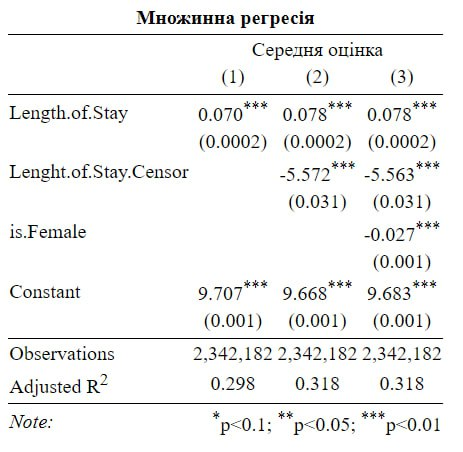
Існує статистично значуща різниця між Model 1 та Model 2, оскільки p-рівень менше заданого рівня значущості (у цьому випадку, менше за 0.05). Це означає, що група коефіцієнтів Age.Group (які відсутні в Model 2) є статистично значущою для пояснення варіації залежної змінної Total.Charges.log

Раса.



Результат ідентичний до попереднього дослідження.

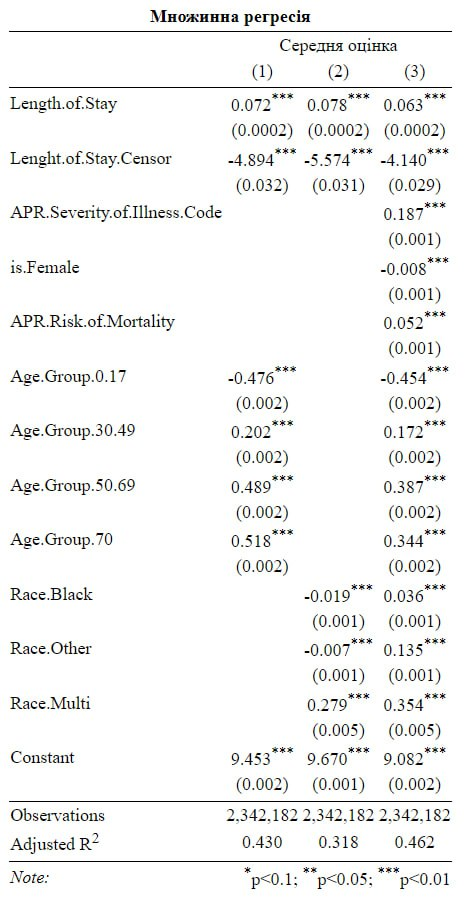
Регресійна модель



1. Спочатку просто довжина перебування

2. Додали наше значення яке відповідає за цензуровану змінну. Як бачимо результат помінявся, отже все ок

3. Додали стать. це не вплинуло на результат, тому дальше її юзати не будемо



4. Досліджуємо вікові групи (базова змінна вік 18-29). Бачимо що він впливає, тому його залишаємо

5. Раса. Вона також впливає, і її залишаємо. (увага! я поміняв базову расу на білих. Раніше було інакше)

6. Додаємо ще ризик смертності і важкість хвороби, які, очевидно, також впливають на ціну. І отримали такі кінцеву модель