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

Зміст роботи

Основна частина

I. Мотивація проведення дослідження

Дослідити вплив наступних змінних вартість оплати (Total.Charges):

- довжина перебування
- стать
- раса
- вік
- важкість захворювання
- ризик смертності

Дослідження Total.Charges

Дослідимо розподіл ціни та результати покажемо на гістограмі (Рис. 1)

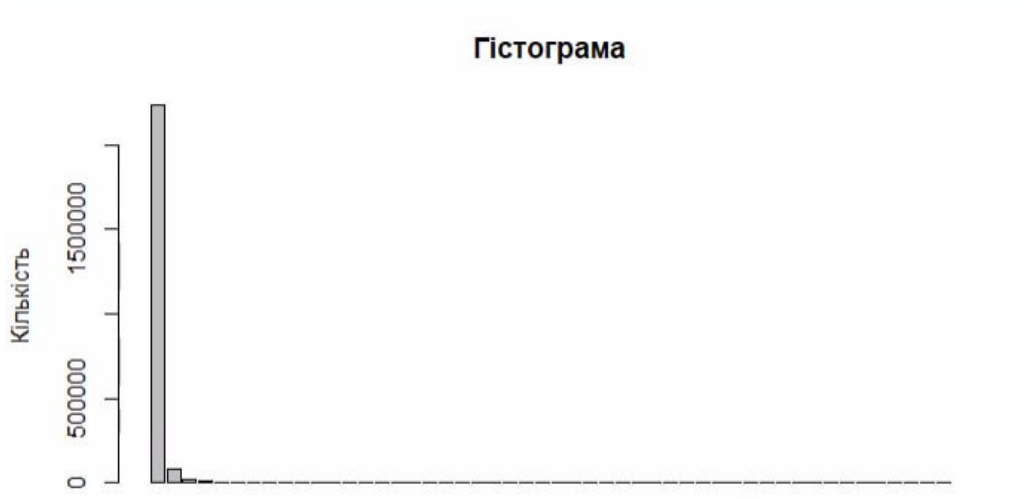

Рис. 1 – Гістограма ціни

Як можемо побачити, гістограма є “скошеною”, тобто має дуже багато значень при малих цінах та дуже мало при великих. Тому варто прологарифмувати нашу змінну та в подальшому досліджувати значення ln(ціна)

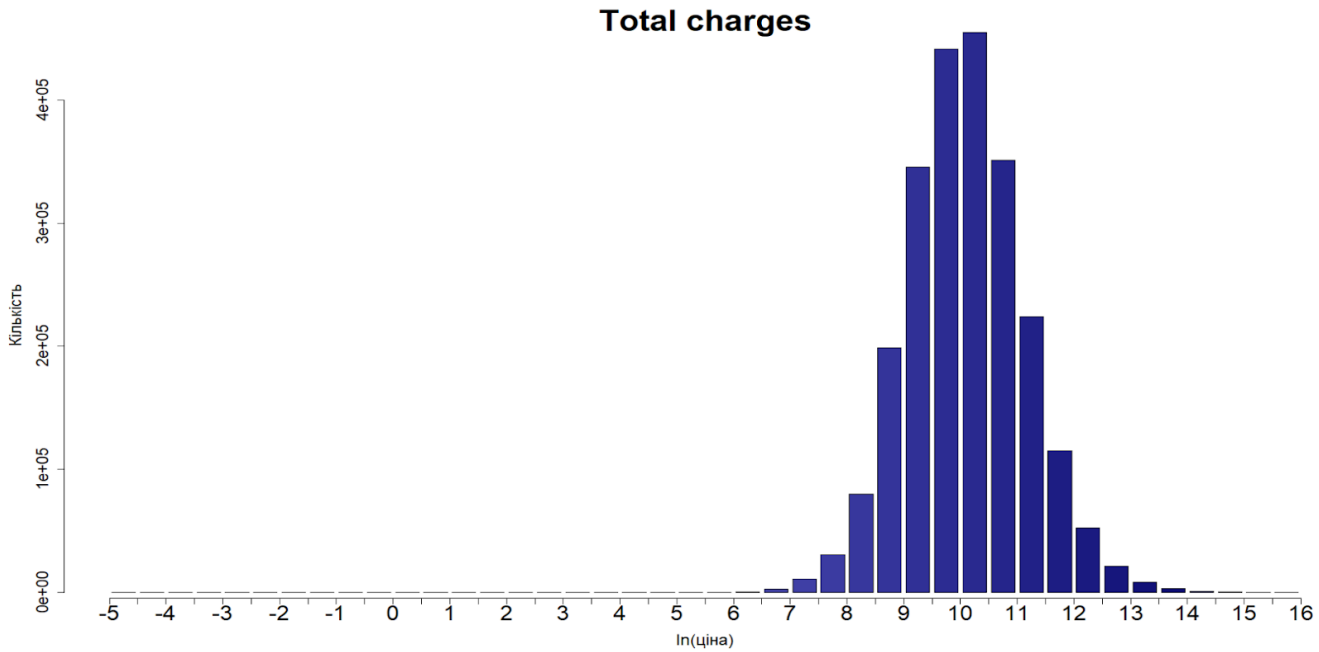

Рис. 2 – Гістограма ln(ціни)

Базова регресійна модель

Розглянемо регресійну модель, де залежною змінною будемо вважати log(Total.Charges), а незалежною - Length.of.Stay (довжина перебування в лікарні). Побудуємо її та отримаємо наступне:

	Total.Charges
Length.of.Stay	0.070***
	(0.0002)
Constant	9.707***
	(0.001)
Observations	2,342,182
Adjusted R^2	0.298
Note:	*p<0.1; **p<0.05; ***p<0.01

Рис. 3 – Базова модель

Коефіцієнт Length.of.Stay має значення 0.07, його стандартна похибка - 0.0002. Бачимо, що коефіцієнт є значущий, отже можна зробити наступні висновки з цієї моделі: збільшення кількості днів перебування на 1 призводить до збільшення суми оплати на 7%.

Проте, на основі цієї моделі не можна робити загальних висновків, оскільки вона складається тільки з однієї змінної і ми можемо не враховувати дуже багато інших факторів, які впливають на ціну.

Проста регресія. Довірчі інтервали

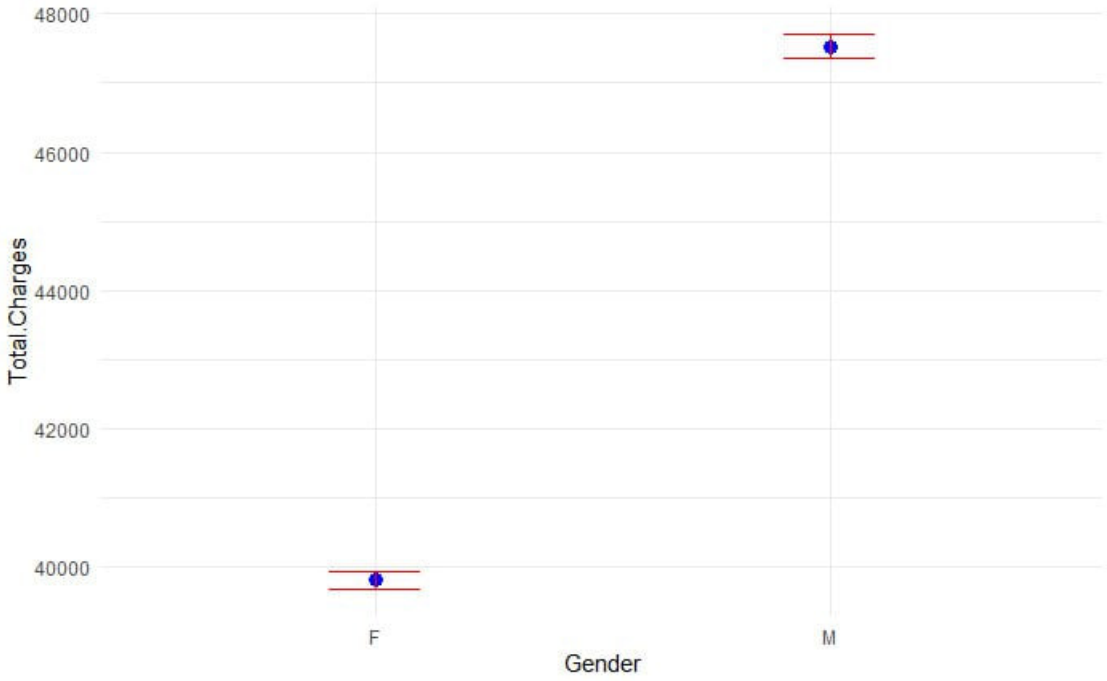

Рис. – Довірчі інтервали для ціни в залежності від статі

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

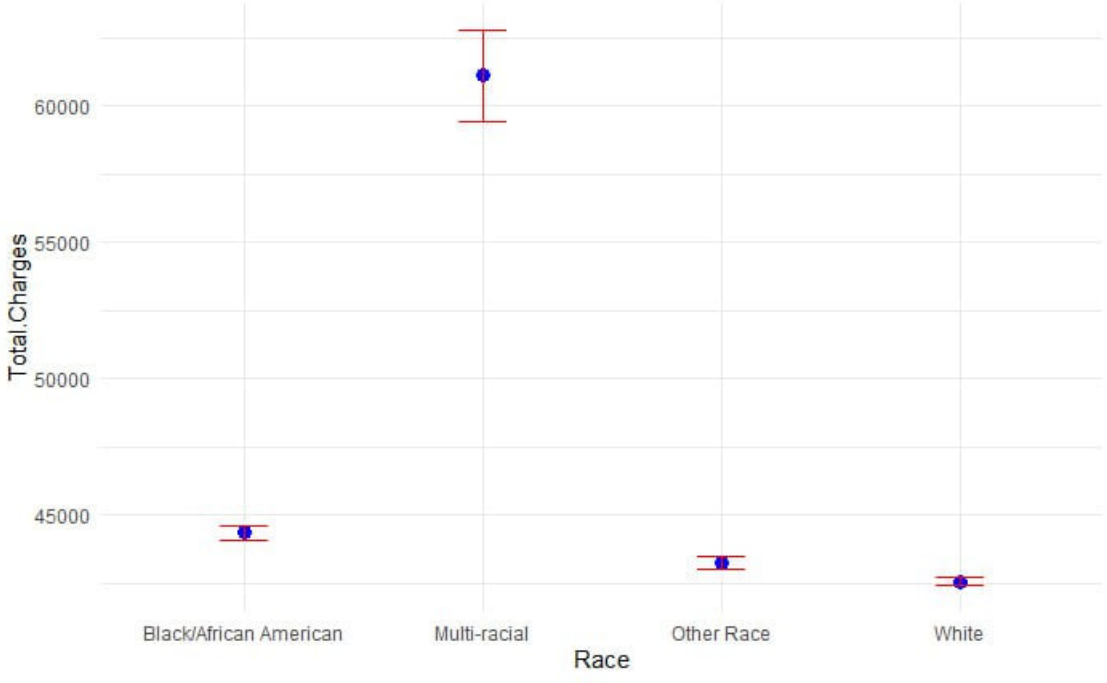

Рис. – Довірчі інтервали для ціни в залежності від раси

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 рас буде додатнім. Щодо абсолютного значення, то найбільше воно буде для мультирас.

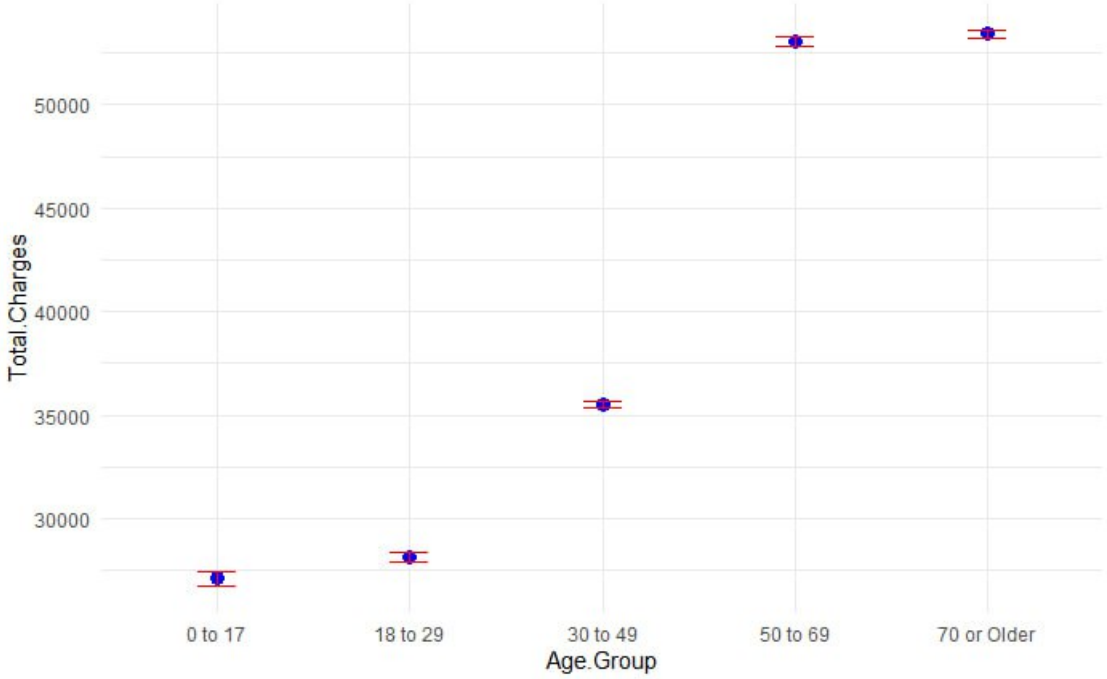

Рис. – Довірчі інтервали для ціни в залежності від вікової групи

За основу у даному випадку візьмемо категорію 18-29 років. Це означає, що вікова група 0-17 буде мати від’ємний коефіцієнт, а для решти груп коефіцієнт буде додатнім.

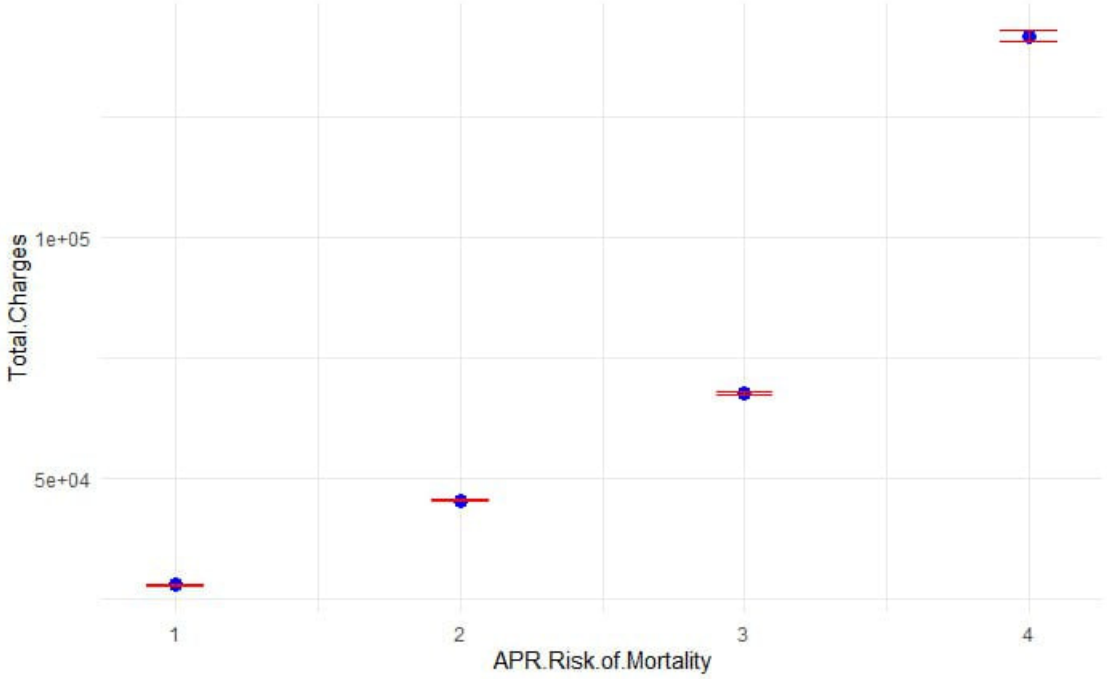

Рис. – Довірчі інтервали для ціни в залежності від ризику смертності

У порівняння з ризиком 1, значення коефіцієнтів 2, 3, 4 будуть більші (чим більший ризик, тим більше тобі прийдеться платити). Отже, можемо зробити висновок, що коефіцієнт при цій змінній буде додатнім.

Рис. – Довірчі інтервали для ціни в залежності від важкості хвороби

Якщо розглядати категорії в порядку зростання важкості, то з її збільшенням зростає і ціна. Згідно з цим можемо зробити висновок, що коефіцієнт при цій змінній теж буде додатнім.

Таким чином, можемо сформулювати остаточні гіпотези щодо знаків коефіцієнтів структурної моделі:

1) Коефіцієнт біля Gender додатній (якщо чоловік, то йому більше прийдеться заплатити) або дуже малий і незначний (подумати треба)
2) Коефіцієнт біля важкості хвороби додатній
3) Коефіцієнт біля ризику смерті буде додатнім
4) Коефіцієнт біля раси буде додатнім
5) Коефіцієнт біля вікової групи буде додатнім

Вплив контрольних змінних

Раніше ми уже побачили регресійну модель з довжиною перебування, а також побудували певні гіпотези щодо знаку коефіцієнту інших можливих регресорів. Спробуємо спочатку дослідити вплив категорії захворювання на ціну. Оскільки у нас ця змінна є категоріальною, розділимо її на множину бінарних та поглянемо на результат моделі:

Множинна регресія

	Середня оцінка	
	(1)	(2)
Length.of.Stay	0.070***	0.070***
	(0.0002)	(0.0002)
APR.MDC.Nervous		-0.491***
		(0.012)
APR.MDC.Eye		-0.721***
		(0.017)
APR.MDC.ENMT		-0.757***
		(0.013)
APR.MDC.Respiratory		-0.730***
		(0.012)
APR.MDC.Circulatory		-0.403***
		(0.012)
APR.MDC.Digestive		-0.638***
		(0.012)
APR.MDC.Hepatobiliary		-0.526***
		(0.012)
APR.MDC.Musculoskeletal		-0.158***
		(0.012)
APR.MDC.Skin.Breast		-0.800***
		(0.013)
APR.MDC.Endocrine		-0.704***
		(0.012)
APR.MDC.Kidney		-0.629***
		(0.012)
APR.MDC.Male		-0.475***
		(0.014)
APR.MDC.Female		-0.490***
		(0.013)
APR.MDC.Pregnancy		-1.029***
		(0.012)
APR.MDC.Neonates		-1.800***
		(0.012)
APR.MDC.Blood		-0.672***
		(0.013)
APR.MDC.Neoplasms		-0.333***
		(0.013)
APR.MDC.Infections		-0.482***
		(0.012)
APR.MDC.Mental		-1.405***
		(0.012)
APR.MDC.Drug		-1.439***
		(0.012)
APR.MDC.Injuries.Poison		-0.815***
		(0.013)
APR.MDC.Burns		-0.612***
		(0.024)
APR.MDC.Not.Sick		-0.764***
		(0.013)
APR.MDC.Trauma		-0.513***
		(0.015)
Constant	9.707***	10.488***
	(0.001)	(0.012)
Observations	2,342,182	2,342,182
Adjusted R^2	0.298	0.499
Note:		*p<0.1; **p<0.05; ***p<0.01

Як бачимо, у порівнянні з базовою моделлю, коефіцієнт при довжині перебування не змінився. З цього ми можемо зробити висновок, що категорія захворювання не впливає у даній моделі на похибку і коефіцієнти моделі.

Побудуємо ще різні регресійні моделі, щоб перевірити вплив інших факторів:

Множинна регресія

	Середня оцінка		
	(1)	(2)	(3)
Length.of.Stay	0.070***	0.078***	0.078***
	(0.0002)	(0.0002)	(0.0002)
Lenght.of.Stay.Censor		-5.572***	-5.563***
		(0.031)	(0.031)
is.Female			-0.027***
			(0.001)
Constant	9.707***	9.668***	9.683***
	(0.001)	(0.001)	(0.001)
Observations	2,342,182	2,342,182	2,342,182
Adjusted R^2	0.298	0.318	0.318
Note:	*p<0.1; **p<0.05; ***p<0.01		

Першою є наша базова модель, побудована раніше. Оскільки сама змінна Length.of.Stay є цензурованою (значення 121 відповідає за час перебування в лікарні більше 120 днів), нами було прийнято рішення створити окрему бінарну змінну, яка демонструє чи є значення цензурованим. Як бачимо, це впливає на коефіцієнти моделі, тому в подальшому будемо використовувати цей регресор. Отримали, що довжина перебування ще більше впливає на ціну, ніж було спочатку.

У цій самій таблиці наведено третю модель, у якій ми вирішили перевірити вплив статі та помітили, що це ніяк не впливає на значення коефіцієнту при довжині перебування. Можемо зробити висновок, що стать не впливає на похибку в даній моделі та ніяким чином не покращує її, отже в подальших моделях її використовувати не будемо.

Дослідимо ще кілька змінних, а саме вік, расу та важкість хвороби із ризиком смертності:

Множинна регресія

	Середня оцінка		
	(1)	(2)	(3)
Length.of.Stay	0.072***	0.078***	0.063***
	(0.0002)	(0.0002)	(0.0002)
Lenght.of.Stay.Censor	-4.894***	-5.574***	-4.141***
	(0.032)	(0.031)	(0.029)
APR.Severity.of.Illness.Code			0.187***
			(0.001)
APR.Risk.of.Mortality			0.052***
			(0.001)
Age.Group.0.17	-0.476***		-0.452***
	(0.002)		(0.002)
Age.Group.30.49	0.202***		0.173***
	(0.002)		(0.002)
Age.Group.50.69	0.489***		0.388***
	(0.002)		(0.002)
Age.Group.70	0.518***		0.345***
	(0.002)		(0.002)
Race.Black		-0.019***	0.036***
		(0.001)	(0.001)
Race.Other		-0.007***	0.135***
		(0.001)	(0.001)
Race.Multi		0.279***	0.354***
		(0.005)	(0.005)
Constant	9.453***	9.670***	9.076***
	(0.002)	(0.001)	(0.002)
Observations	2,342,182	2,342,182	2,342,182
Adjusted R^2	0.430	0.318	0.462
Note:			*p<0.1; **p<0.05; ***p<0.01

Спочатку ми спробували дослідити вплив віку. Оскільки це категоріальна змінна, ми розбили її на бінарні, де за базисну взяли вік 18-29. Коефіцієнти в моделі змінились, на цей раз уже зменшились, отже, можемо вважати, що вік має певний вплив. Тому його залишаємо для подальших моделей.

Після цього було досліджено вплив раси, аналогічно до вікової групи та були отримані аналогічні результати - отже, змінні залишаємо.

Останні контрольні змінні, які ми вирішили додати до нашої моделі - це ризик смертності та важкість хвороби. Вони є категоріальними, але впорядкованими, тому просто додаємо їх у модель та порівнюємо результати. Як бачимо, вони

зменшують залежність днів перебування від ціни, тож впливають на якість моделі.

Всі незалежні змінні, які було розглянуто, є статистично значущими, проте оскільки певні категоріальні змінні були розбиті на бінарні, ми вирішили перевірити статистичну значущість групи змінних віку та раси. Для цього були перевірені гіпотези на рівність їх нулю, результати цієї перевірки наведено в таблицях нижче.

Linear hypothesis test

Hypothesis:

Age.Group.0.17 = 0

Age.Group.30.49 = 0

Age.Group.50.69 = 0

Age.Group.70 = 0

Model 1: restricted model

Model 2: Total.Charges.log ~ Length.of.Stay + Lenght.of.Stay.Censor +
APR.Severity.of.Illness.Code + APR.Risk.of.Mortality +
Age.Group.0.17 + Age.Group.30.49 + Age.Group.50.69 + Age.Group.70 +
Race.Black + Race.Other + Race.Multi

Note: Coefficient covariance matrix supplied.

	Res.Df	Df	F	Pr(>F)
1	2342170			
2	2342168	4	56176	< 2.2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Табл. - Перевірка значущості впливу коефіцієнту при змінних, що відповідають за вік

Linear hypothesis test

Hypothesis:

Race.Black = 0

Race.Other = 0

Race.Multi = 0

Model 1: restricted model

Model 2: Total.Charges.log ~ Length.of.Stay + Lenght.of.Stay.Censor +
APR.Severity.of.Illness.Code + APR.Risk.of.Mortality +
Age.Group.0.17 + Age.Group.30.49 + Age.Group.50.69 + Age.Group.70 +
Race.Black + Race.Other + Race.Multi

Note: Coefficient covariance matrix supplied.

	Res.Df	Df	F	Pr(>F)
1	2342170			
2	2342168	3	5576.9	< 2.2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Табл. - Перевірка значущості впливу коефіцієнту при змінних, що відповідають за расу

Результати в двох випадках ідентичні, тому можемо зробити наступний висновок: існує статистично значуща різниця між Model 1 (з віком чи расою) та Model 2 (без віку чи раси відповідно), оскільки p-рівень менше заданого рівня значущості (у цьому випадку, менше за 0.05). Це означає, що група коефіцієнтів Age.Group або Race (які відсутні в Model 2) є статистично значущою для пояснення варіації залежної змінної Total.Charges.log.

Узагальнюючи отримані нами результати, можемо зробити наступний висновок: кожен додатковий день в лікарні збільшує ціну на 6.3%

Мультиколінеарність

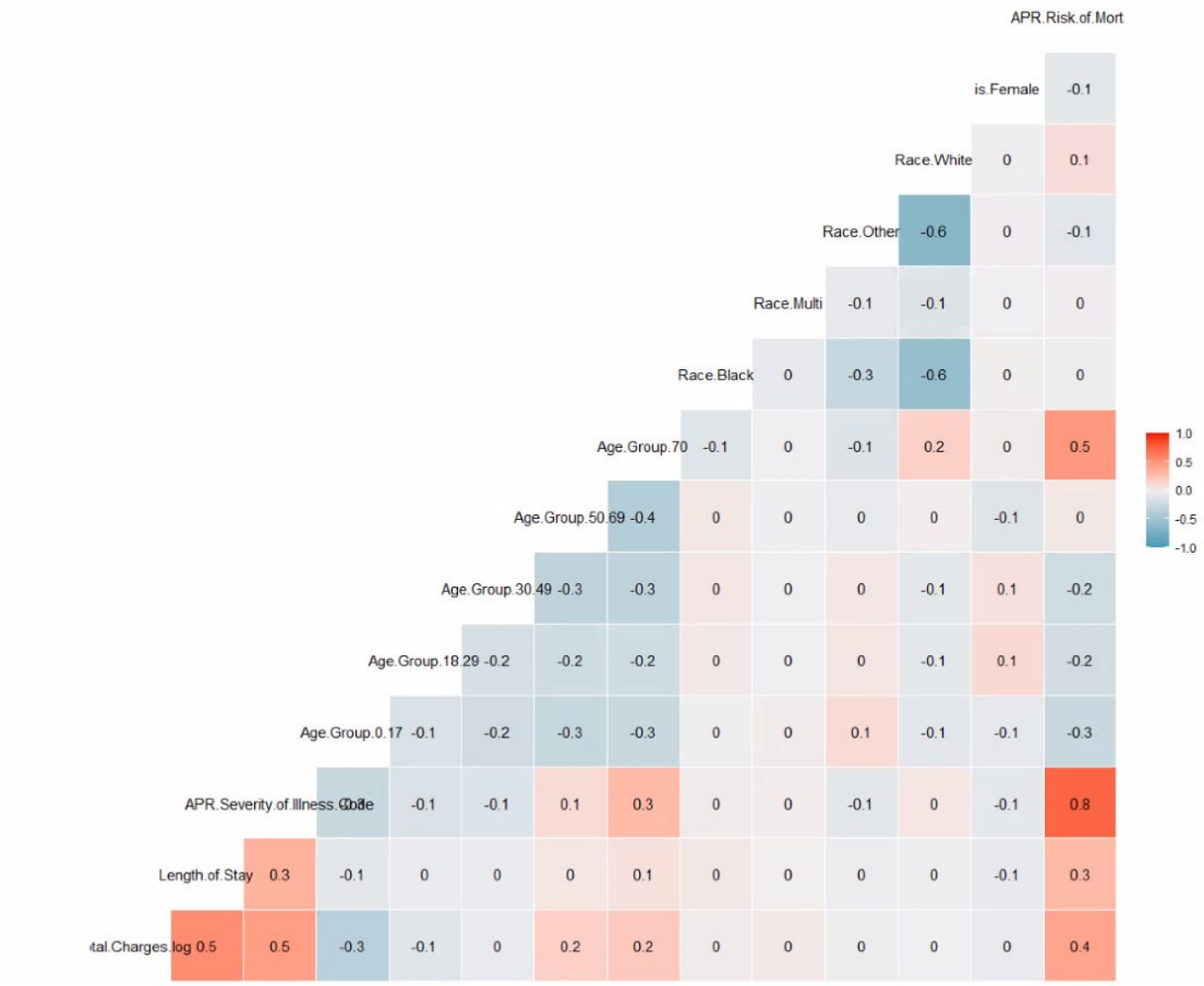

Рис. – Корелограма

Перевіримо чи є мультиколінеарність. Як можемо побачити на графіку, у нас є зв’язок між деякими змінними, особливо сильний між важкістю захворювання та ризиком смертності. Проте, оскільки ці дві змінні у нас мають достатньо сильну кореляцію з Total.Charges.log, ми не можемо їх просто так викинути.

Додавання поліномів вищих порядків та логарифмів відносно регресорів

Для перевірки наявності нелінійних ефектів у моделі, було досліджено кілька варіантів їх наявності: квадратичний поліном, логаритм від змінної і квадратичний поліном від логаритм змінної. Це привело до таких результатів:

Множинна регресія

	Середня оцінка
I(log(Length.of.Stay))	0.671***
	(0.001)
Lenght.of.Stay.Censor	0.912***
	(0.022)
APR.Severity.of.Illness.Code	0.105***
	(0.001)
APR.Risk.of.Mortality	0.043***
	(0.001)
Age.Group.0.17	-0.402***
	(0.002)
Age.Group.30.49	0.174***
	(0.002)
Age.Group.50.69	0.379***
	(0.002)
Age.Group.70	0.284***
	(0.002)
Race.Black	0.029***
	(0.001)
Race.Other	0.145***
	(0.001)
Race.Multi	0.362***
	(0.004)
Constant	8.766***
	(0.002)
Observations	2,342,182
Adjusted R^2	0.527
Note:	*p<0.1; **p<0.05; ***p<0.01

Рис. – Модель з логарифмом довжини перебування

Для цієї моделі, звісно ж, було перевірено статистичну значущість коефіцієнту при логаритмі змінної, і результати наведено у таблиці нижче.

Linear hypothesis test

Hypothesis:
I(log(Length.of.Stay)) = 0

Model 1: restricted model

Model 2: Total.Charges.log ~ I(log(Length.of.Stay)) + Lenght.of.Stay.Censor + APR.Severity.of.Illness.Code + APR.Risk.of.Mortality + Age.Group.0.17 + Age.Group.30.49 + Age.Group.50.69 + Age.Group.70 + Race.Black + Race.Other + Race.Multi

Note: Coefficient covariance matrix supplied.

	Res.Df	Df	F	Pr(>F)
1	2342171			
2	2342170	1	985466	< 2.2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Згідно з тестом в таблиці, коефіцієнт є статистично значущим, тобто ненульовим.

Множинна регресія

	Середня оцінка
I(Length.of.Stay2)	-0.001***
	(0.00001)
Length.of.Stay	0.115***
	(0.0002)
Lenght.of.Stay.Censor	3.813***
	(0.055)
APR.Severity.of.Illness.Code	0.138***
	(0.001)
APR.Risk.of.Mortality	0.033***
	(0.001)
Age.Group.0.17	-0.427***
	(0.002)
Age.Group.30.49	0.172***
	(0.002)
Age.Group.50.69	0.380***
	(0.002)
Age.Group.70	0.333***
	(0.002)
Race.Black	0.025***
	(0.001)
Race.Other	0.137***
	(0.001)
Race.Multi	0.357***
	(0.004)
Constant	9.009***
	(0.002)
Observations	2,342,182
Adjusted R^2	0.505
Note:	*p<0.1; **p<0.05; ***p<0.01

Рис. – Модель з квадратом довжини перебування

Для цієї моделі, звісно ж, було перевірено статистичну значущість коефіцієнту при обидвох регресорах, які відповідають змінній, і результати наведено у таблиці нижче.

Linear hypothesis test

Hypothesis:

Length.of.Stay = 0

I(Length.of.Stay^2) = 0

Model 1: restricted model

Model 2: Total.Charges.log ~ I(Length.of.Stay^2) + Length.of.Stay + Lenght.of.Stay.Censor + APR.Severity.of.Illness.Code + APR.Risk.of.Mortality + Age.Group.0.17 + Age.Group.30.49 + Age.Group.50.69 + Age.Group.70 + Race.Black + Race.Other + Race.Multi

Note: Coefficient covariance matrix supplied.

	Res.Df	Df	F	Pr(>F)
1	2342171			
2	2342169	2	357941	< 2.2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Згідно з результатами, наведеними в таблиці, обидва коефіцієнти є ненульовими, хоч і величина того, який стоїть при квадраті змінної, доволі мала, щоб достатньо впливати на результат.

Множинна регресія

	Середня оцінка
I(log(Length.of.Stay)2)	0.088***
	(0.001)
I(log(Length.of.Stay))	0.409***
	(0.002)
Lenght.of.Stay.Censor	0.039*
	(0.023)
APR.Severity.of.Illness.Code	0.103***
	(0.001)
APR.Risk.of.Mortality	0.034***
	(0.001)
Age.Group.0.17	-0.408***
	(0.002)
Age.Group.30.49	0.173***
	(0.002)
Age.Group.50.69	0.377***
	(0.002)
Age.Group.70	0.301***
	(0.002)
Race.Black	0.024***
	(0.001)
Race.Other	0.142***
	(0.001)
Race.Multi	0.361***
	(0.004)
Constant	8.912***
	(0.002)
Observations	2,342,182
Adjusted R^2	0.534
Note:	*p<0.1; **p<0.05; ***p<0.01

Рис. – Модель з квадрат логарифму

Для цієї моделі, було перевірено статистичну значущість коефіцієнту при обидвох регресорах, які відповідають змінній, і результати наведено у таблиці нижче.

Linear hypothesis test

Hypothesis:

I(log(Length.of.Stay)) = 0

I(log(Length.of.Stay)^2) = 0

Model 1: restricted model

Model 2: Total.Charges.log ~ I(log(Length.of.Stay)^2) + I(log(Length.of.Stay)) + Lenght.of.Stay.Censor + APR.Severity.of.Illness.Code + APR.Risk.of.Mortality + Age.Group.0.17 + Age.Group.30.49 + Age.Group.50.69 + Age.Group.70 + Race.Black + Race.Other + Race.Multi

Note: Coefficient covariance matrix supplied.

	Res.Df	Df	F	Pr(>F)
1	2342171			
2	2342169	2	510419	< 2.2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

З результатів можна зробити висновок, що обидва коефіцієнти є ненульовими.

Для порівняння усіх чотирьох моделей було використано значення R_{adj}^2 та графіки залежності ціни від різних функцій від тривалості перебування. Згідно з цими даними, найкраще описує дані модель із квадратним поліномом від логаритму тривалості перебування.

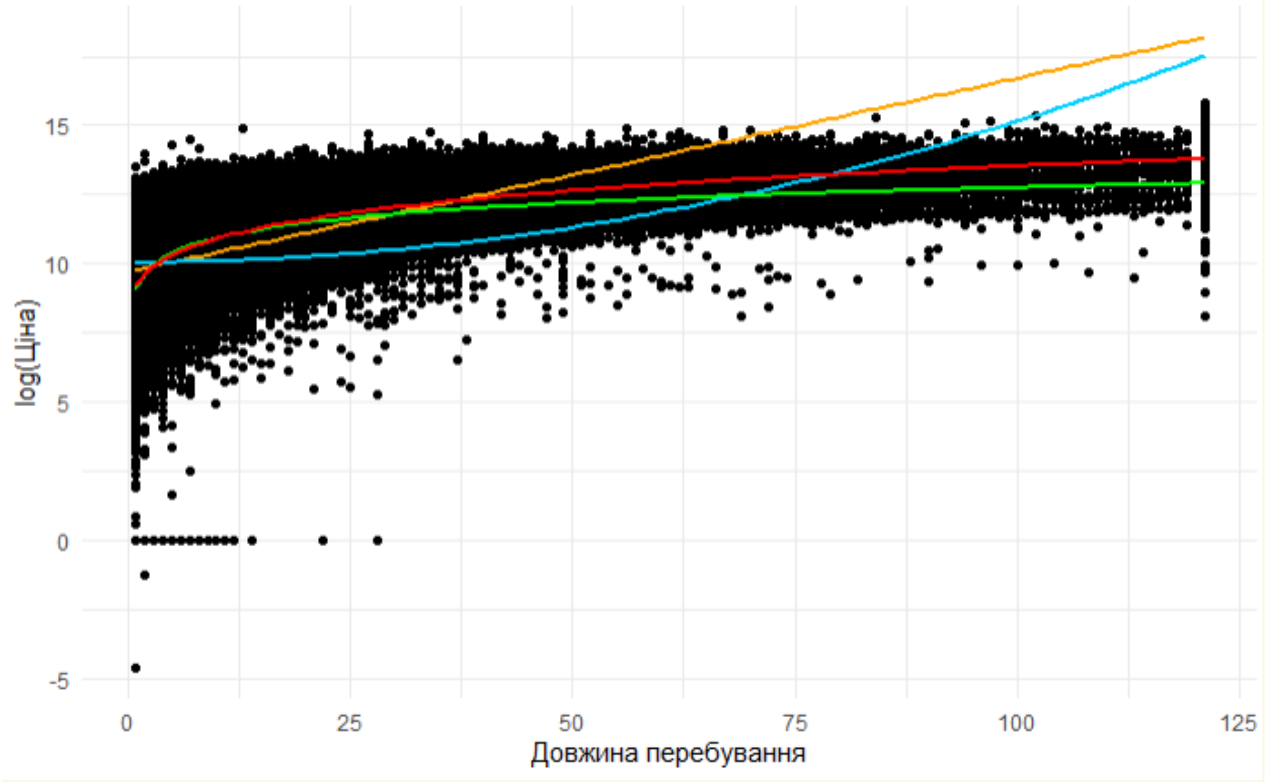

Рис. – порівняння моделей з різними нелінійними ефектами

(жовтий - просто тривалість перебування, синій – квадрат тривалості перебування, зелений – з логарифмом, червоний - квадрат логаритму тривалості перебування)

Висновки

- Ура, ми встигли!
- Було досліджено чи існує вплив факторів на змінну Total.Charges
- Була розроблена модель з контрольними змінними
- Було застосовано логарифмування та використані поліноми
- Було визначено, що ціна пов’язана з:
    - Довжиною перебування
    - Віком
    - Расою
    - Важкістю захворювання
    - Ризиком смертності