РЕГРЕСІЙНИЙ АНАЛІЗ

Лабораторна робота №3

Виконали студенти КМ-01:
Бабич Ірина
Іваник Юрій
Романецький Микита
Суховій Ігор
Шолоп Любомир

Мотивація проведення дослідження

Дослідити наступні змінні на вплив вартості оплати (Total.Charges):

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

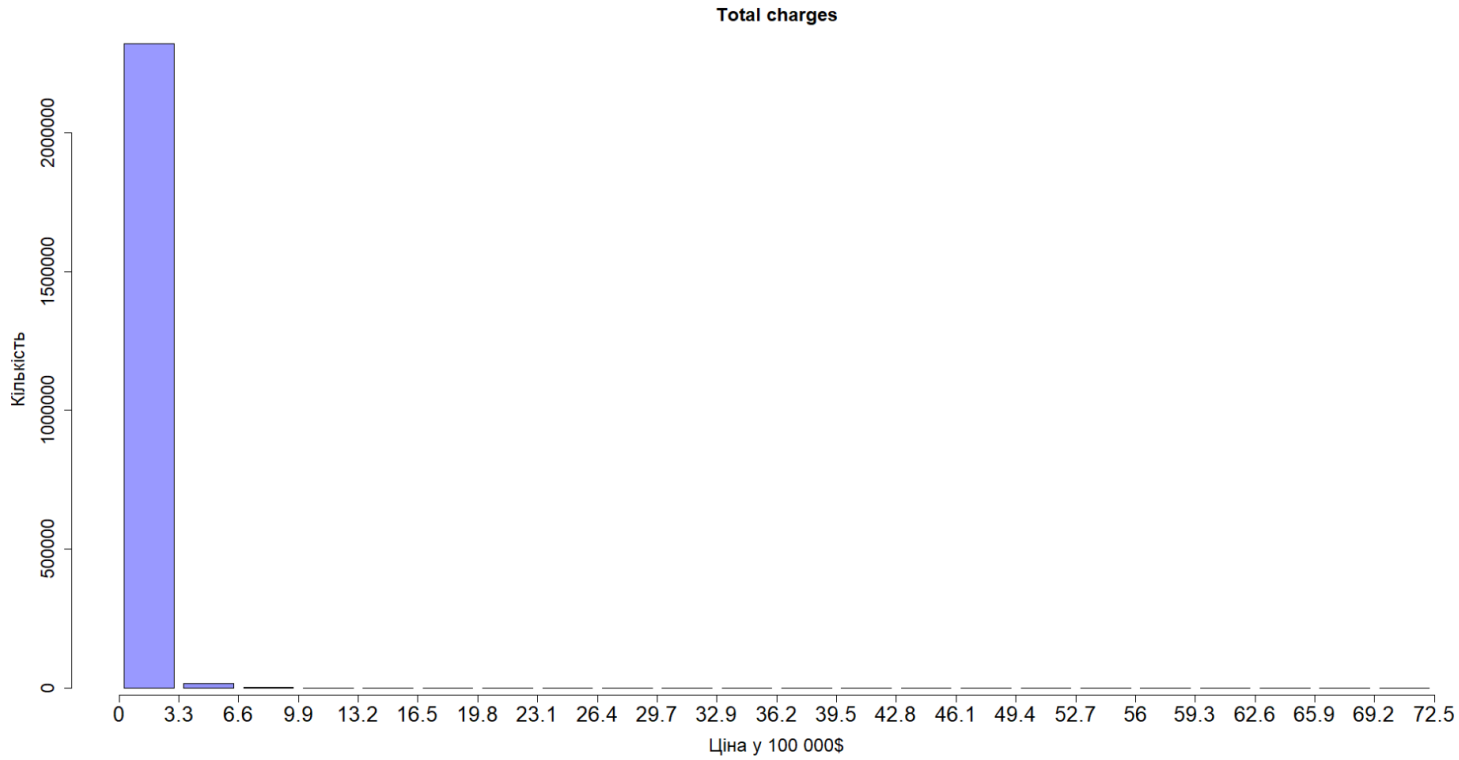
Опис даних

- Обраний датасет містить дані щодо надання лікарських послуг пацієнтам у штаті Нью-Йорк за 2015 рік.
- Він не містить даних, які би могли вказувати на причетність до них окремих осіб. Також, з цього файлу було виключено вторинні діагнози та процедури, а також коди оплати послуг. Один рядок відповідає інформації про рівно 1 клінічний випадок.
- Датасет після очистки містить 2342182 записів та 33 змінних.
- Вік пацієнтів представлений у таких вікових групах:
  - від 0 до 17 років
  - від 18 до 29 років
  - 30 до 49 років
  - від 50 до 69 років
  - і від 70 років і старше

Змінні датасету

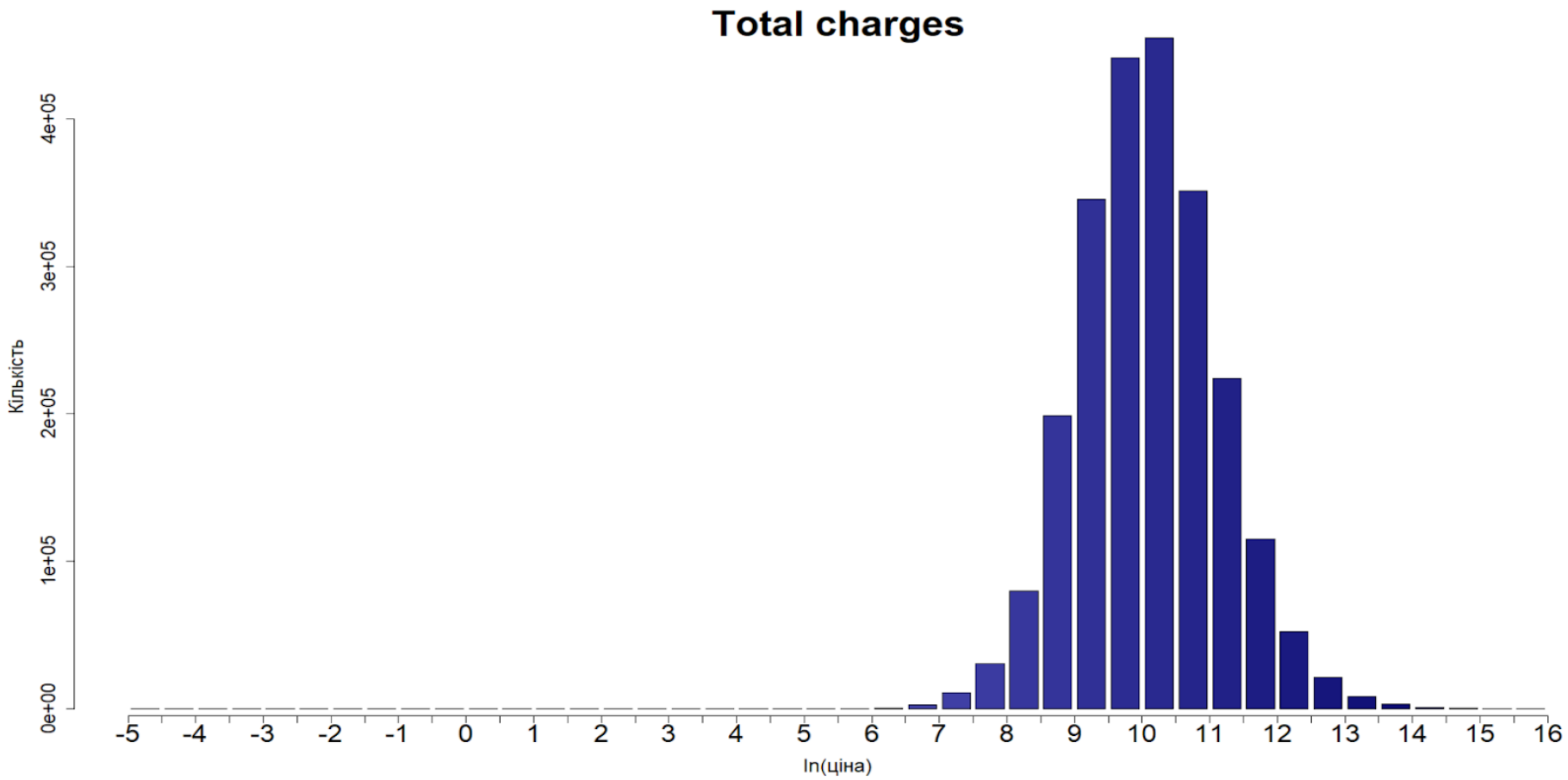
- Кожна колонка датасету відповідає одній змінній.
- Більшість змінних в датасеті є категорійними
- Числовими і впорядкованими змінними в датасеті є:
  - *Length.of.Stay*
  - *Birth.Weight (через природу даних містить найбільшу кількість NA значень)*
  - *Total.Charges*
  - *Total.Costs*

Дослідження Total.Charges



Як можемо побачити, гістограма є “скошеною”, тобто має дуже багато значень при малих цінах та дуже мало при великих.

Дослідження Total.Charges



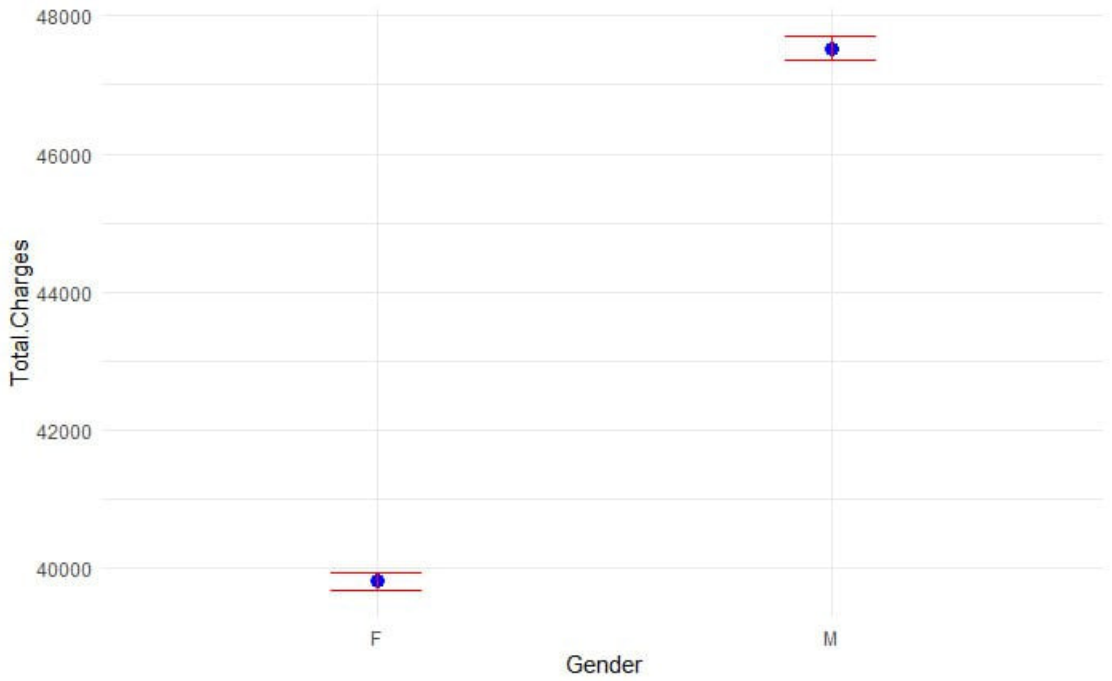
Базова регресійна модель

- Розглянемо регресійну модель, де залежною змінною будемо вважати ln(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

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%.

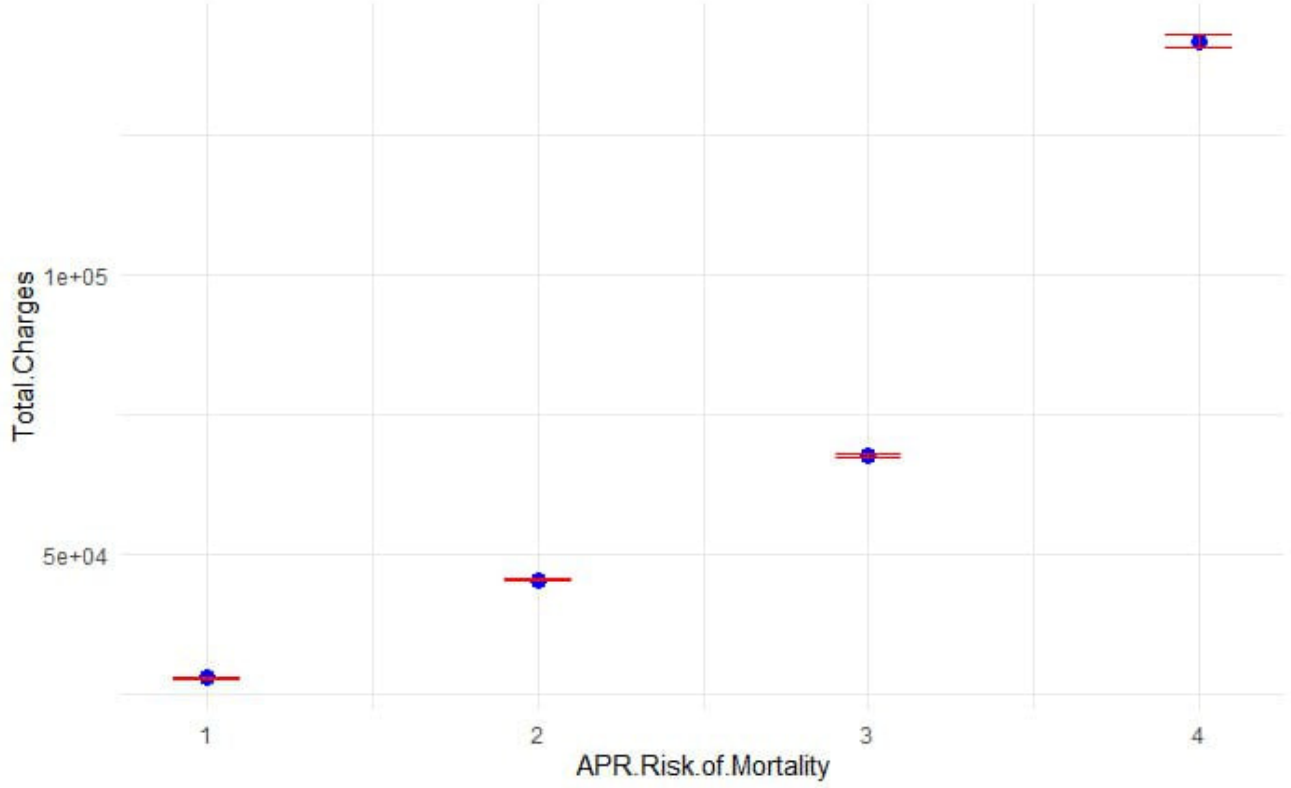
Проста регресія. Довірчі інтервали



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 зменшувати значення, повернені моделлю, тобто коефіцієнт буде від'ємний.

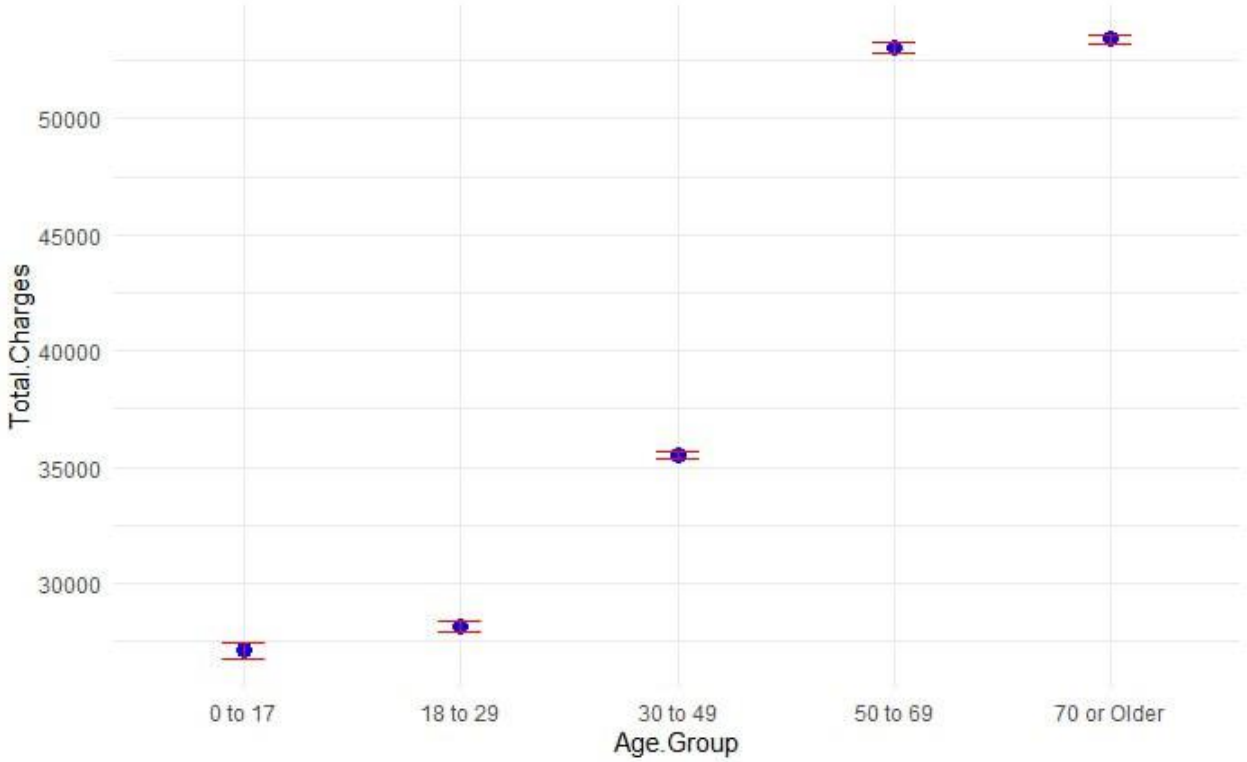
Проста регресія. Довірчі інтервали



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

Проста регресія. Довірчі інтервали



Довірчі інтервали для ціни в залежності від вікової групи

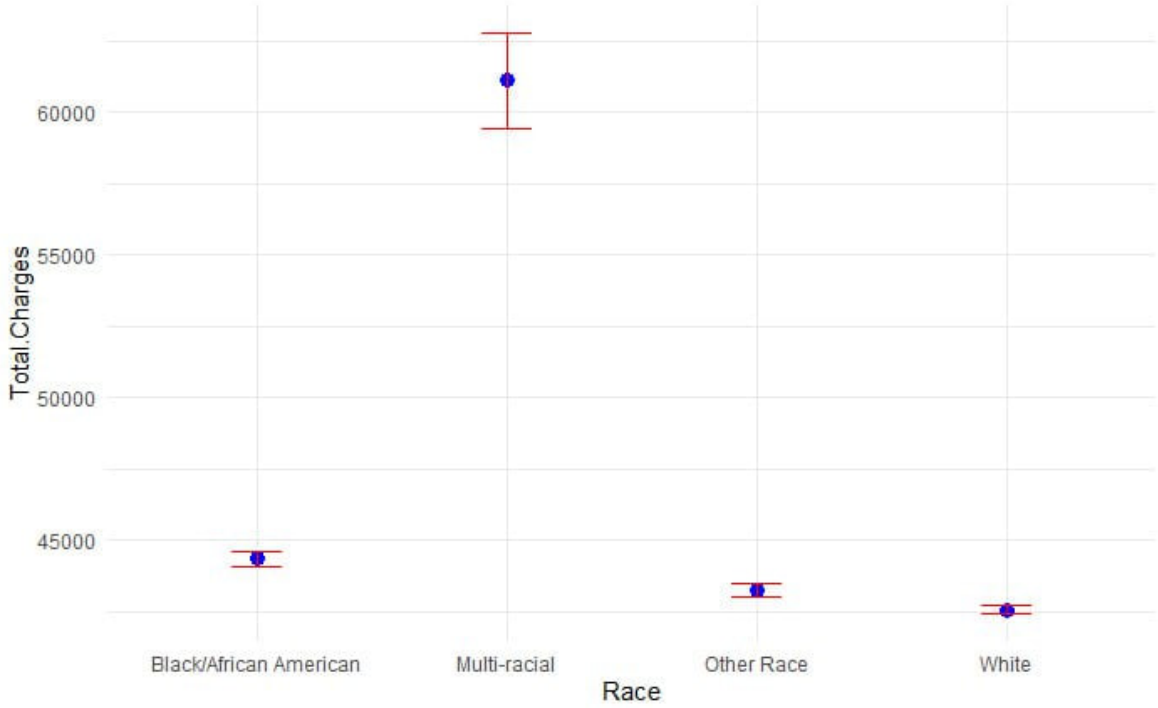
За основу тут візьмемо 18-29. Це означає, що вікова група 0-17 буде мати від'ємний коефіцієнт, а для решти груп коефіцієнт буде додатнім.

Проста регресія. Довірчі інтервали

Довірчі інтервали для ціни в залежності від важкості хвороби

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

Проста регресія. Довірчі інтервали



Довірчі інтервали для ціни в залежності від раси

Оскільки людей білої раси найбільше в лікарні, вони базові. Виходить, що білі платять найменше, отже знак коефіцієнта в регресії для інших рас буде додатнім. Щодо абсолютного значення, то найбільше воно буде для мультирас.

Гіпотези

- Гіпотези, щодо знаків коефіцієнтів структурної моделі:
  - Коефіцієнт біля Gender додатній (якщо чоловік, то йому більше прийдеться заплатити) або дуже малий і незначний (подумати треба)
  - Коефіцієнт біля важкості хвороби додатній
  - Коефіцієнт біля ризику смерті буде додатнім
  - Коефіцієнт біля раси буде додатнім
  - Коефіцієнт біля вікової групи буде додатнім

Вплив контрольних змінних

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	

Вплив контрольних змінних

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		

Отримали, що довжина перебування ще більше впливає на ціну, ніж було спочатку

Після цього 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 ціну, отже в подальших моделях її використовувати не будемо

Вплив контрольних змінних

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

- У першій моделі додаємо вікові групи
- У другій – раси
- У третій – додаємо ризик смертності та важкість хвороби

Дослідження значущості груп коефіцієнтів віку та раси

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

Linear hypothesis test

Hypothesis:

Race.Black = 0

Race.Other = 0

Race.Multi = 0

Model 1: restricted model

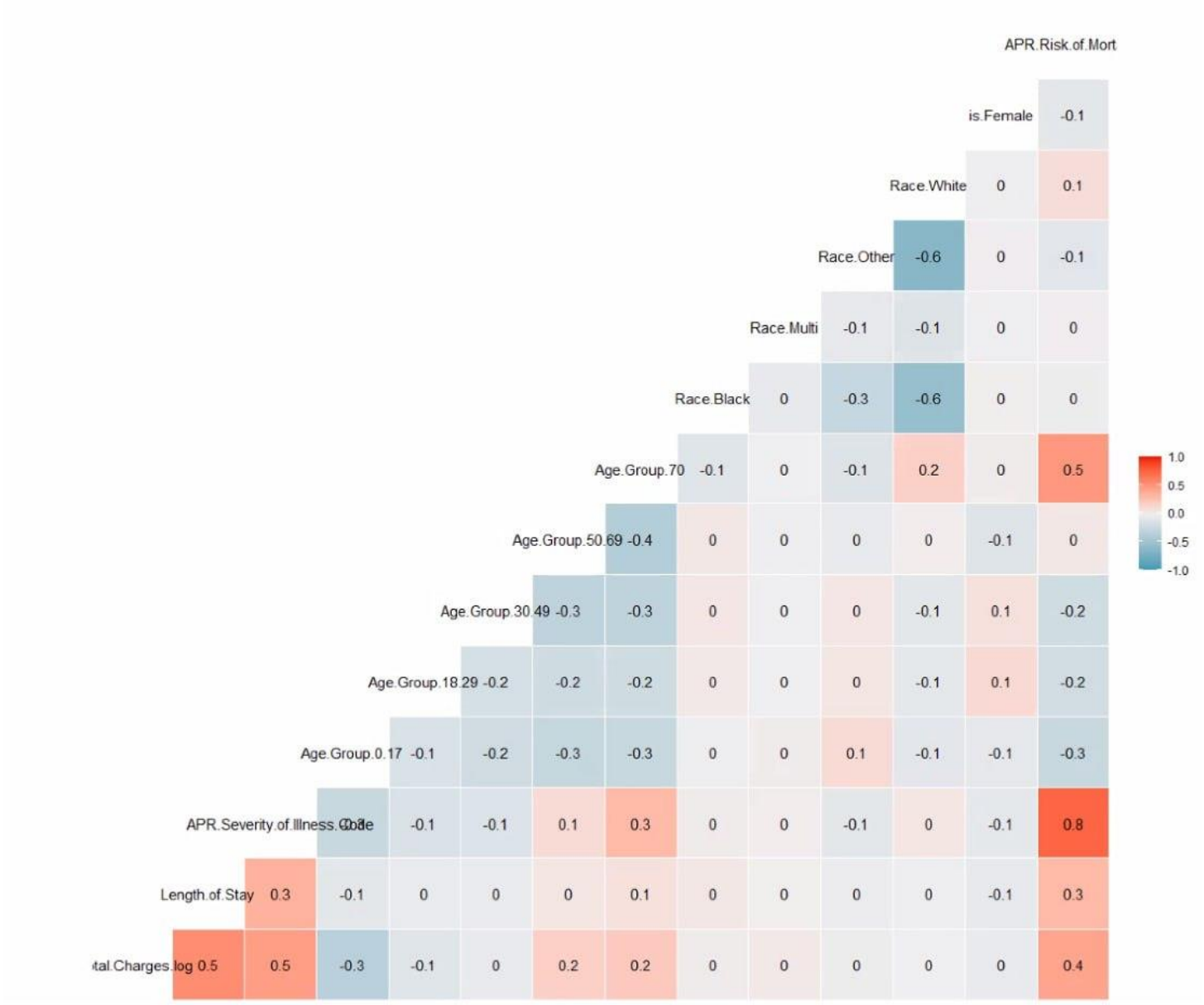
Model 2: Total.Charges.log ~ Length.of.Stay + Lenght.of.Stay.Censor +
APR.Severity.of.Illness.Code + APR.Risk.of.Mortality +
Age.Group.0.17 + Age.Group.30.49 + Age.Group.50.69 + Age.Group.70 +
Race.Black + Race.Other + Race.Multi

Note: Coefficient covariance matrix supplied.

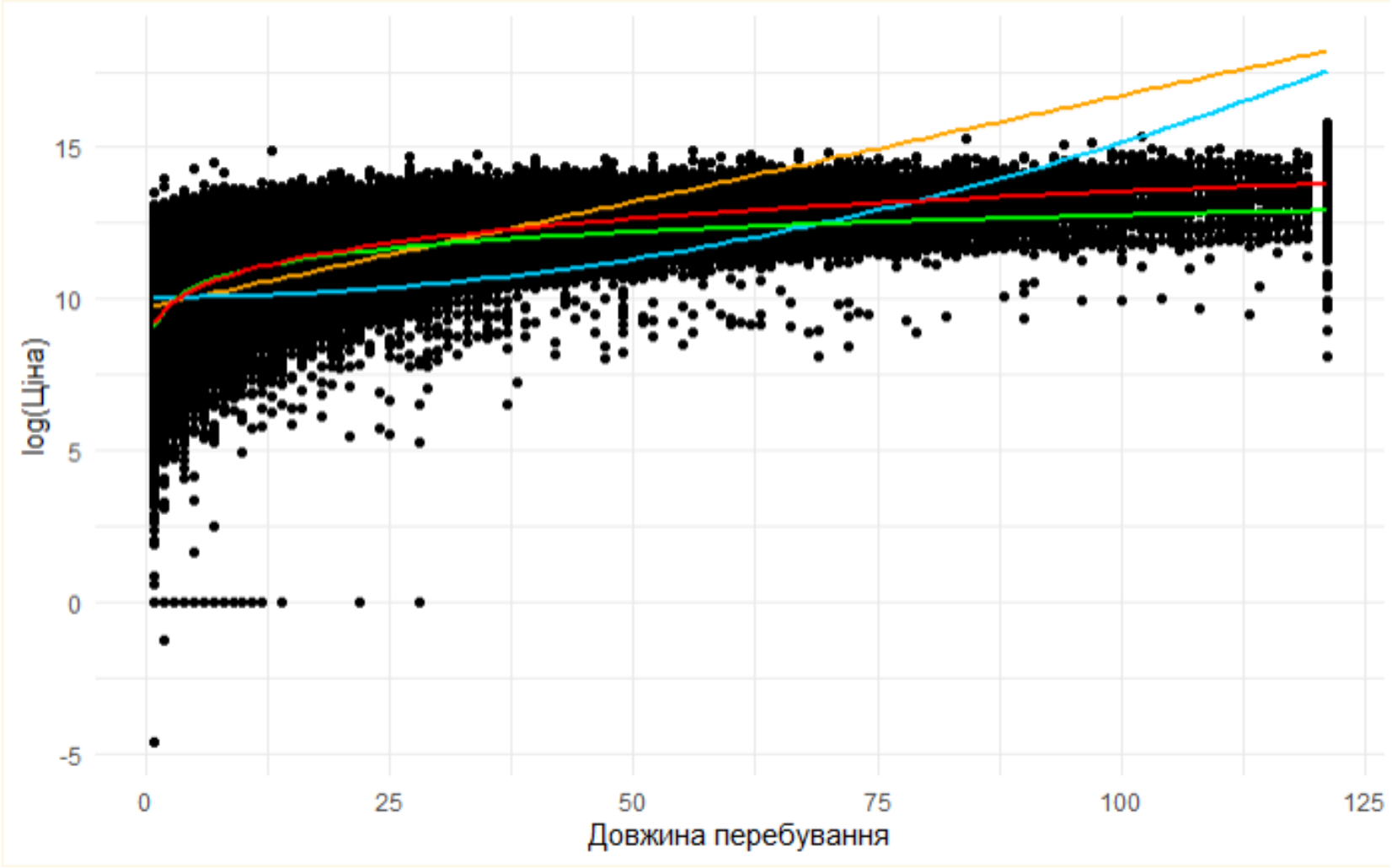
	Res.Df	Df	F	Pr(>F)
1	2342170			
2	2342168	3	5576.9	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Дослідимо мультиколінеарність



Length.of.Stay vs ln(Length.of.Stay) vs Length.of.Stay^2 vs ln(Length.of.Stay)^2



ln(Length.of.Stay)

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

Length.of.Stay^2

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

ln(Length.of.Stay)^2

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

ln(Length.of.Stay)

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

Length.of.Stay^2

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

ln(Length.of.Stay)^2

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

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