lab3

2024-05-30

Назва команди - Команда №3

Перелік учасників колективу виконавців:

- Пономаренко Олександр (КМ-12)
- Земляний Даниїл (КМ-12)
- Борисенко Данило (КМ-11)
- Заіченко Дамир (КМ-13)
- Лук'яненко Василь (КМ-13)

ПИТАННЯ НА РЕГРЕСІЮ

1. Чим зумовлено скасування/нескасування резервації? (з "Які характерні риси скасованих записів?")
2. Що впливає на появу особливих побажань? (з "Чи є істотною наявність особливих побажань?")
3. Чи є вплив необхідності в паркувальному місці на середню ціну за кімнату?
4. Чи є вплив повторного гостя на середню ціну за кімнату?
5. Чи є вплив кількості людей на середню ціну за кімнату?

1. Чим зумовлено скасування/нескасування резервації?

- Побудуємо базову логит-модель, де в ролі залежної бінарної змінної виступатиме атрибут booking_status ('Canceled', 'Not Canceled').
- Взагалі, можливість з'ясувати чи буде конкретний запис скасований чи нескасований видається досить "профітною". Подумки можна швидко прикинути, що можливо повторні гості відмінятимуть записи з меншою ймовірністю. В той же час статус бронювання мав би бути пов'язаний з часом до прибуття (бо інутїтивно очікується, що чим більше lead_time, то тим менша ймовірність того, що гість все ж прибуде) та, наприклад, кількістю дітей. Хоч записів з дітьми відносно загальної кількості записів не так вже й багато (2559 з 36275), проте захворювання напередодні поїздки або поведінкові проблеми (чи будь-які інші непередбачувані обставини пов'язані з дітьми) мали б зробити свій внесок у ймовірність скасування.
- Окрім того, оскільки lead_time має досить високий ренж значень (від 0 до 443), то було б логічно дослідити зміну даної фічі на 1%, а не на одну одиницю, аби наглядніше було видно її вплив.
- Отже маємо наступні результати:

```
denylogit <- glm(booking_status ~ log(lead_time+1) + no_of_adults + no_of_children
                 + required_car_parking_space + no_of_nights + repeated_guest,
                 family = binomial(link = "logit"),
                 data = hotel)

modelsummary(list("basic" = denylogit),
             gof_omit = "^(?!Num.Obs.|R2 Adj.)",
             stars = TRUE)
```

	basic
(Intercept)	3.982***
	(0.071)
log(lead_time + 1)	-0.750***
	(0.012)
no_of_adults	-0.108***
	(0.025)
no_of_children	-0.315***
	(0.031)
required_car_parking_space1	1.272***
	(0.105)
no_of_nights	-0.017*
	(0.007)
repeated_guest1	1.803***
	(0.260)

Num.Obs.	35674

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

- (додатній коефіцієнт - в сторону НЕскасування, тобто прибуття)
- (від'ємний коефіцієнт - в сторону скасування бронювання)
- Як бачимо, поки що усі коефіцієнти є статистично значущі. Одразу в очі кидаються коефіцієнти при логаритмі "часу до прибуття", "потребі у паркувальному місці" та "належністю до повторних гостей", проте робити якісь висновки ще зарано.
- Підключимо у гру контрольні змінні. Нажаль ми все ще не знаємо у чому полягає фактична різниця між типами кімнат, єдине що ми з'ясували напевно, ще на етапі EDA, це те що в різні типи кімнат заселяються різні кількість дітей. Тому щоб зменшити кореляцію похибки з відповідним регресором вважається необхідним додати типи кімнат до розгляду.

```
denylogit_with_controls <- glm(booking_status ~ log(lead_time+1) + no_of_adults
                               + no_of_children + required_car_parking_space
                               + no_of_nights + repeated_guest
                               + R2 + R3 + R4 + R5 + R6 + R7,
                family = binomial(link = "logit"),
                data = hotel_room)

modelsummary(list("basic" = denylogit, "with_control_vars" = denylogit_with_controls),
             stars = TRUE,
             gof_omit = "^(?!Num.Obs.)")
```

	basic	with_control_vars
(Intercept)	3.982***	3.938***
	(0.071)	(0.071)
log(lead_time + 1)	-0.750***	-0.760***
	(0.012)	(0.012)
no_of_adults	-0.108***	-0.052+
	(0.025)	(0.027)
no_of_children	-0.315***	-0.200***
	(0.031)	(0.044)
required_car_parking_space1	1.272***	1.283***
	(0.105)	(0.106)
no_of_nights	-0.017*	-0.014*
	(0.007)	(0.007)
repeated_guest1	1.803***	1.817***
	(0.260)	(0.260)
R2TRUE		0.195*
		(0.092)
R3TRUE		-0.070
		(0.955)
R4TRUE		-0.163***
		(0.035)
R5TRUE		-0.111
		(0.156)
R6TRUE		-0.495***
		(0.105)

R7TRUE		-0.070
		(0.221)
Num.Obs.	35674	35674

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

- Як можемо спостерігати, ситуація для усіх коефіцієнтів в цілому лишилася незмінною, за виключенням кількості дорослих. Коефіцієнт все ще лишається статистично значущим, проте тепер є ймовірність того, що справжнє його значення майже не відрізняється від нуля.
- Аби детальніше розглянути вплив кількості дорослих на статус бронювання, розглянемо фактори взаємодії кількості дорослих з кількістю дітей (аби взяти до розгляду сім'ї) та належністю гостя до повторних гостей:

```
denylogit_factors <- glm(booking_status ~ log(lead_time+1)
                       + no_of_adults*(no_of_children + repeated_guest)
                       + required_car_parking_space + no_of_nights
                       + R2 + R3 + R4 + R5 + R6 + R7,
                 family = binomial(link = "logit"),
                 data = hotel_room)

modelsummary(list("basic" = denylogit, "with_control_vars" = denylogit_with_controls,
                  "all_in_one_with_factors" = denylogit_factors),
             stars = TRUE,
             gof_omit = "^(?!Num.Obs.)")
```

	basic	with_control_vars	all_in_one_with_factors
(Intercept)	3.982***	3.938***	3.924***
	(0.071)	(0.071)	(0.072)
log(lead_time + 1)	-0.750***	-0.760***	-0.761***
	(0.012)	(0.012)	(0.012)
no_of_adults	-0.108***	-0.052+	-0.043
	(0.025)	(0.027)	(0.027)
no_of_children	-0.315***	-0.200***	-0.097
	(0.031)	(0.044)	(0.104)
required_car_parking_space1	1.272***	1.283***	1.284***
	(0.105)	(0.106)	(0.106)
no_of_nights	-0.017*	-0.014*	-0.014*
	(0.007)	(0.007)	(0.007)
repeated_guest1	1.803***	1.817***	2.567**
	(0.260)	(0.260)	(0.783)
R2TRUE		0.195*	0.159
		(0.092)	(0.097)
R3TRUE		-0.070	-0.071
		(0.955)	(0.954)
R4TRUE		-0.163***	-0.167***
		(0.035)	(0.035)
R5TRUE		-0.111	-0.107
		(0.156)	(0.156)
R6TRUE		-0.495***	-0.465***
		(0.105)	(0.108)

R7TRUE		-0.070	-0.049
		(0.221)	(0.222)
no_of_adults × no_of_children			-0.062
			(0.057)
no_of_adults × repeated_guest1			-0.547
			(0.518)
Num.Obs.	35674	35674	35674

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

- Як можемо помітити, коефіцієнти при “кількості дорослих” та факторах взаємодії “кількості дорослих” з “повторним гостем” та “кількістю дітей” стали статистично незначущі.
- Є необхідність протестувати на статистичну значущість групу цих коефіцієнтів:

```
names(coef(denylogit_factors))
```

```
##  [1] "(Intercept)"                  "log(lead_time + 1)"
##  [3] "no_of_adults"                 "no_of_children"
##  [5] "repeated_guest1"              "required_car_parking_space1"
##  [7] "no_of_nights"                 "R2TRUE"
##  [9] "R3TRUE"                       "R4TRUE"
## [11] "R5TRUE"                       "R6TRUE"
## [13] "R7TRUE"                       "no_of_adults:no_of_children"
## [15] "no_of_adults:repeated_guest1"
```

```
coef_names <- names(coef(denylogit_factors))
print(coef_names)
```

```
##  [1] "(Intercept)"                  "log(lead_time + 1)"
##  [3] "no_of_adults"                 "no_of_children"
##  [5] "repeated_guest1"              "required_car_parking_space1"
##  [7] "no_of_nights"                 "R2TRUE"
##  [9] "R3TRUE"                       "R4TRUE"
## [11] "R5TRUE"                       "R6TRUE"
## [13] "R7TRUE"                       "no_of_adults:no_of_children"
## [15] "no_of_adults:repeated_guest1"
```

```
# Example updated based on hypothetical correct names
linearHypothesis(denylogit_factors,
                 c("no_of_adults", "no_of_adults:no_of_children", "no_of_adults:repeated_guest1"),
                 vcov. = vcovHC(denylogit_factors, "HC1"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## no_of_adults = 0
## no_of_adults:no_of_children = 0
## no_of_adults:repeated_guest1 = 0
##
## Model 1: restricted model
## Model 2: booking_status ~ log(lead_time + 1) + no_of_adults * (no_of_children +
##     repeated_guest) + required_car_parking_space + no_of_nights +
##     R2 + R3 + R4 + R5 + R6 + R7
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1  35662
## 2  35659  3 6.3427    0.09608 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Відтак бачимо, що кількість дорослих разом зі своїми факторами взаємодії є статистично значуща, попри те що як окремий коефіцієнт при кількості дорослих, так і коефіцієнти суто на факторах взаємодії, є статистично незначущими

- Оскільки наша початкова модель досить гарно витримала тест на стійкість, то розглянемо таблицю зі значеннями середнього маржинального ефекту кожної змінної на появу особливих побажань.

```
summary(margins(denylogit))
```

```
##                       factor     AME     SE         z      p   lower   upper
##                    lead_time -0.0041 0.0000 -113.9018 0.0000 -0.0042 -0.0040
##                 no_of_adults -0.0201 0.0046   -4.3220 0.0000 -0.0292 -0.0110
##               no_of_children -0.0587 0.0057  -10.3769 0.0000 -0.0698 -0.0476
##                 no_of_nights -0.0032 0.0013   -2.4306 0.0151 -0.0058 -0.0006
##              repeated_guest1  0.2441 0.0204   11.9735 0.0000  0.2041  0.2840
##  required_car_parking_space1  0.1954 0.0120   16.2815 0.0000  0.1719  0.2189
```

- (варто взяти до уваги, що в дійсності належність до класу повторних гостей ще має ще більший середній маржинальний ефект, ніж вказано у таблиці)

2. Що впливає на появу особливих побажань?

Першу базову модель побудуємо з досить інтуїтивно очікуваними регресорами.

- В першу чергу нам цікаво дізнатися чи враховується таки потреба у паркувальному місці як особливе побажання. Окрім цього, можна очікувати, що на наявність особливих побажань впливатиме ціна, що заплачена за кімнату.
- Окрім цього, достовірно невідомо що саме являють собою особливі побажання, бо це може бути як ранкова корзинка фруктів під дверима так і лебідь з рушників, що очікуватиме гостей на двуспальному ліжку, тому досить важливо було б врахувати ціну, яка заплачена за заброньваний номер. При побудові моделі варто врахувати, що для avg_price_per_room для адекватнішого аналізу його впливу необхідно взяти логаритм цього регресора, адже межі змінної досить широкі (від 9 до 540), через що вплив збільшення значення ціни на 1 одиницю буде майже непомітним. Тож було б логічно розглянути вплив збільшення ціни на 1 відсоток відносно наявності особливих побажань.
- Для початкової моделі братимемо дорослих і дітей поокремо. У нашому датасеті більшість записів просто з дорослими (26,5 тисячі записів мають двох дорослих без дітей), тому першочергово було б цікаво розглянути модель без фактору взаємодії дорослих та дітей.
- Останнім регресором в початкову модель додамо бінарну змінну repeated_guest, бо є підозра, що у повторних гостей можуть бути певні бонуси, або ж у них за попередні відвідування з'явились вподобання, які теоретично можна віднести до особливих побажань

```
requests_logit <- glm(no_of_special_requests ~ required_car_parking_space
                      + log(avg_price_per_room) + no_of_adults + no_of_children
                      + repeated_guest,
                        data = hotel,
                        family = binomial(link = "logit"))

modelsummary(list("basic" = requests_logit),
             gof_omit = "^(?!Num.Obs.|R2 Adj.)",
             stars = TRUE)
```

	basic
(Intercept)	-5.182***
	(0.185)
required_car_parking_space1	1.054***
	(0.071)
log(avg_price_per_room)	0.811***
	(0.041)
no_of_adults	0.641***
	(0.023)
no_of_children	0.310***
	(0.031)
repeated_guest1	0.194*
	(0.079)
Num.Obs.	35674

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

- Як можемо бачити, принаймні поки що, усі коефіцієнти є статистично значущими, мають додатній знак і жоден не обертається в нуль
- Додамо тепер до базової моделі контрольні змінні: можна підозрювати, що наявність особливих побажань може бути пов'язана з тривалістю зупинки у готелі (тобто кількістю ночей). Окрім цього є ще такий фактор як час до прибуття. Оскільки час до прибуття аналогічно до avg_price_per_room варіюється на досить великому проміжку, то було б логічно взяти від lead_time логаритм, аби подивитись на вплив збільшення часу до прибуття на 1%. Також варто врахувати те, що типи кімнат відрізняються між собою як мінімум по кількості дітей, тож у нас є ґрунтовні підстави вважати, що через невідому нам різницю між цими типами кімнат може з'являтися більше чи менше причин для появи особливих побажань.

```
requests_logit_controls <- glm(no_of_special_requests ~ log(lead_time+1) + required_car_parking_space
                               + log(avg_price_per_room) + no_of_adults + no_of_children
                               + repeated_guest
                               + no_of_nights
                               + R2 + R3 + R4 + R5 + R6 + R7,
                           data = hotel_room,
                           family = binomial(link = "logit"))

modelsummary(list("basic" = requests_logit, "with_control_vars" = requests_logit_controls),
             gof_omit = "^(?!Num.Obs.)",
             stars = TRUE)
```

	basic	with_control_vars
(Intercept)	-5.182***	-5.325***
	(0.185)	(0.210)
required_car_parking_space1	1.054***	1.025***
	(0.071)	(0.071)
log(avg_price_per_room)	0.811***	0.876***
	(0.041)	(0.045)
no_of_adults	0.641***	0.660***
	(0.023)	(0.025)
no_of_children	0.310***	0.563***
	(0.031)	(0.042)
repeated_guest1	0.194*	0.118
	(0.079)	(0.081)
log(lead_time + 1)		-0.122***
		(0.008)
no_of_nights		0.081***
		(0.007)
R2TRUE		0.467***
		(0.084)
R3TRUE		-1.100
		(1.138)
R4TRUE		0.137***
		(0.033)
R5TRUE		-1.515***
		(0.166)
R6TRUE		-1.079***
		(0.100)

R7TRUE		-0.345
		(0.222)
Num.Obs.	35674	35674

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

- Як бачимо, додавання контрольних змінних лиш незначним чином вплинуло на коефіцієнти перед “потребою у паркувальному місці”, “логарифм ціни”, “кількість дорослих” та “кількість дітей”. Цікавим є те, що коефіцієнт перед повторним гостем перестав бути статистично значущим, і тепер він теоретично може набувати значення 0.
- Додамо фактори взаємодії між змінними. Першочергово, для детальнішого аналізу впливу змінної “повторний гість” спробуємо врахувати фактори взаємодії з потребою у паркувальному місці (це зможе нам підказати, чи зберігається такий статистично значущий зв’язок між наявністю особливих побажань та потребою у паркувальному місці при бронюванні із разу в раз, чи здебільшого він притаманний новим гостям), та фактор взаємодії з кількістю дорослих (знову таки, оскільки більшість записів у датасеті містять виключно дорослих, то цікаво з’ясувати чи пов’язана кількість особливих побажань з кількістю людей у класі повторних гостей).
- Крім того, оскільки в датасеті є записи як з виключно дорослими так і з виключно дітьми, то варто врахувати фактор взаємодії між ними аби включити до розгляду сім’ї з дорослих та дітей.

```
requests_logit_factor <- glm(no_of_special_requests ~ log(lead_time+1) + required_car_parking_space
                                          + log(avg_price_per_room) + no_of_adults + no_of_children
                                          + no_of_adults:no_of_children + no_of_nights + repeated_guest
                                          + repeated_guest:required_car_parking_space
                                          + repeated_guest:no_of_adults
                                          + R2 + R3 + R4 + R5 + R6 + R7,
                        data = hotel_room,
                        family = binomial(link = "logit"))

modelsummary(list("basic" = requests_logit, "with_control_vars" = requests_logit_controls,
                  "all_in_one_with_factors" = requests_logit_factor),
             gof_omit = "^(?!Num.Obs.)",
             stars = TRUE)
```

	basic	with_control_vars	all_in_one_with_factors
(Intercept)	-5.182***	-5.325***	-5.412***
	(0.185)	(0.210)	(0.211)
required_car_parking_space1	1.054***	1.025***	1.081***
	(0.071)	(0.071)	(0.078)
log(avg_price_per_room)	0.811***	0.876***	0.880***
	(0.041)	(0.045)	(0.045)
no_of_adults	0.641***	0.660***	0.702***
	(0.023)	(0.025)	(0.026)
no_of_children	0.310***	0.563***	0.974***
	(0.031)	(0.042)	(0.099)
repeated_guest1	0.194*	0.118	1.168***
	(0.079)	(0.081)	(0.246)
log(lead_time + 1)		-0.122***	-0.123***
		(0.008)	(0.008)
no_of_nights		0.081***	0.081***
		(0.007)	(0.007)
R2TRUE		0.467***	0.329***
		(0.084)	(0.089)
R3TRUE		-1.100	-1.105

		(1.138)	(1.139)
R4TRUE		0.137***	0.122***
		(0.033)	(0.033)
R5TRUE		-1.515***	-1.497***
		(0.166)	(0.166)
R6TRUE		-1.079***	-0.960***
		(0.100)	(0.102)
R7TRUE		-0.345	-0.278
		(0.222)	(0.221)
no_of_adults × no_of_children			-0.248***
			(0.054)
required_car_parking_space1 × repeated_guest1			-0.505*
			(0.210)
no_of_adults × repeated_guest1			-0.777***
			(0.185)
Num.Obs.	35674	35674	35674

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

- Ситуація з коефіцієнтами для “потреби в паркувальному місці”, “логаритму ціни” та “кількості дорослих” в цілому залишається незмінною, а для “кількості дітей” коефіцієнт зріс майже вдвічі. Як можемо бачити, коефіцієнти перед факторами взаємодії між “дорослими та дітьми” та “дорослими та належністю до повторних гостей” є статистично значущими. Загалом картина така, що наявність дітей у сім’ї значним чином зменшує ймовірність того, що сім’я матиме особливі побажання. Самі по собі діти мають додатний статистично значущий зв’язок, тобто дійсно, коли в діти самі бронюють номер в готелі, то вони в такій ситуації частіше супроводжуються певними “спеціальними умовами”. Також варто звернути увагу на те, що серед людей, які потребують паркувальне місце, “повторність гостя” та зростання кількості дорослих у бронюванні значним чином нівелюють додатний статистично значущий зв’язок паркувального місця з наявністю особливих побажань.
- Оскільки наша початкова модель досить гарно витримала тест на стійкість, то розглянемо таблицю зі значеннями середнього маржинального ефекту кожної змінної на появу особливих побажань.

```
summary(margins(requests_logit))
```

```
##                      factor    AME     SE       z      p  lower  upper
##          avg_price_per_room 0.0020 0.0001 20.2236 0.0000 0.0018 0.0021
##                no_of_adults 0.1492 0.0053 28.4040 0.0000 0.1389 0.1595
##            no_of_children 0.0722 0.0072 10.0307 0.0000 0.0581 0.0863
##             repeated_guest1 0.0453 0.0186  2.4440 0.0145 0.0090 0.0817
##  required_car_parking_space1 0.2405 0.0146 16.4739 0.0000 0.2119 0.2691
```

- Як бачимо, є досить природнім, що на появу особливих побажань значним чином впливає кількість дорослих. При цьому можемо бачити підтвердження гіпотези про те, що потреба у паркувальному місці дійсно певним чином пов’язана з появою особливих побажань.

Питання №3

3. Чи є вплив необхідності в паркувальному місці на середню ціну за кімнату?

- Під час минулої лабораторної роботи було побудовано довірчі інтервали для середньої ціни для різних категорій людей - тих, кому потрібне, і кому не потрібне паркувальне місце. Було помічено статистично значущу різницю в цінах. У нас закралася думка, що паркувальне місце, можливо, коштує додаткових грошей окремо від ціни за кімнату. Тим не менш, у нас було надто мало інструментів, щоб показати причинно-наслідковий зв’язок. Зараз же, можемо спробувати пояснити це за допомогою регресії
- Отже, побудуємо базову модель (1) для залежності логарифму ціни за кімнату від необхідності в паркувальному місці. За цією моделлю видно статистично значущий додатний вплив необхідності в паркувальному місці, проте існує багато сумнівів щодо істинності даної моделі.
- Тому, розширимо дану модель додавши декілька логічних контрольних змінних (2). Контрольні змінні - це змінні, які при минулому дослідженні показали зв’язок з необхідністю в паркувальному місці, а саме: кількість людей, наявність особливих побажань, кількість ночей, логарифмований час до прибуття, повторний гість і ринковий сегмент. Бачимо, що майже вдвічі знизився коефіцієнт біля основного регресора, тобто дійсно існувала похибка від неврахованих змінних для першої моделі.

- Дослідження попереднього питання показало статистично значущий зв'язок між необхідністю в паркувальному місці і спеціальному запиті. Тому, необхідно додати фактор взаємодії паркувального місця і наявністю особливих побажань (3), бо вони можуть бути взаємопов'язані між собою. Наприклад, особливе побажання може бути пов'язано з автомобілем, це може бути спеціальне паркувальне місце (криті стоянки), додаткова мийка тощо. Бачимо, що даний фактор робить коефіцієнт при паркувальному місці незначущим, при цьому для тих, кому необхідне паркувальне місце і при цьому є особливі побажання коефіцієнт є значущим, і досить немалим. Тобто, якщо необхідні ці обидві складові ціна збільшиться на 8%. Це немало, вважаючи що середня ціна знаходиться в межах 100 євро. Але цього ще недостатньо, щоб говорити про причинно-наслідковий зв'язок.
- Насправді, наші дані мають панельну природу, дійсно, маємо часовий проміжок, з яким, очевидно, змінюється і ціна, а, також, існують особливості для сегментів ринку і типів кімнат, такі як кількість людей в них, ціна тощо. Тому доцільно буде зафіксувати ефекти кімнат (4) і часові ефекти(5).
- Бачимо і в 4, і в 5 моделі, ключовий коефіцієнт стає значущим, хоча і невеликим. Тобто, можемо сказати, що ті, кому необхідно лише паркувальне місце платять дещо більше, це може бути символічна ціна за паркувальне місце. Але, якщо при цьому є необхідність в особливому побажанні клієнти платять більше в середньому на 9%. Як вже було зазначено вище, можливо, це пов'язано з додатковими послугами щодо автомобіля.
- Також нам відомо, що повторні гості частіше приїжджають автомобілем, тому буде розумно перевірити фактор взаємодії повторного гостя і необхідності в паркувальному місці. Це змінило коефіцієнт при паркувальному місці і дещо збільшило його, тобто можемо сказати що, напевно, існує ціна за паркувальне місце. Але більш цікаво буде подивитися на коефіцієнти при факторах взаємодії: ті, кому потібрне особливе побажання (ймовірно пов'язане з автомобілем), платять ще трохи більше, а ті, хто ще й є повторним гостем платять дещо менше, це спонукає нас перевірити вплив повторного гостя на ціну в наступному запитанні.

Винести на перезентацію (В презентації)

```
hotel_lead <- hotel_market %>%
  mutate(lead_time = if_else(lead_time == 0, lead_time + 1, lead_time))

model_car <- feols(log_price ~ required_car_parking_space, data = hotel, vcov = "HC1")
model_car_ext <- feols(log_price ~ required_car_parking_space + no_of_special_requests + no_of_people + no_of_nig
hts + log(lead_time) + repeated_guest + Online + Corporate + Complementary + Aviation, data = hotel_lead, vcov =
"HC1")
model_car_extvz <- feols(log_price ~ required_car_parking_space*no_of_special_requests + no_of_people + no_of_nig
hts + log(lead_time) + repeated_guest + Online + Corporate + Complementary + Aviation , data = hotel_lead, vcov =
"HC1")
model_car_room <- feols(log_price ~ required_car_parking_space*no_of_special_requests + no_of_people + no_of_nigh
ts + log(lead_time) + repeated_guest + Online + Corporate + Complementary + Aviation  | room_type_reserved , data
= hotel_lead)
model_car_room_time <- feols(log_price ~ required_car_parking_space*no_of_special_requests + no_of_people + no_of
_nights + log(lead_time) + repeated_guest + Online + Corporate + Complementary + Aviation  | room_type_reserved +
arrival_year_and_month, data = hotel_lead)
model_car_st <- feols(log_price ~ required_car_parking_space*no_of_special_requests + no_of_people + no_of_nights
+ log(lead_time) + repeated_guest + required_car_parking_space:repeated_guest + Online + Corporate + Complementar
y + Aviation  | room_type_reserved + arrival_year_and_month, data = hotel_lead)

desc_row <- tibble(title = c("Фіксовані ефекти кімнат", "Часові фіксовані ефекти"),
  model_car = c("Hi", "Hi"),
  model_ext = c("Hi", "Hi"),
  model_car_extvz = c("Hi", "Hi"),
  model_car_room = c("Так", "Hi"),
  model_car_room_time = c("Так", "Так"),
  model_car_st =c("Так", "Так"))


modelsummary(list(model_car, model_car_ext, model_car_extvz, model_car_room, model_car_room_time, model_car_st),
             stars = TRUE,
             gof_omit = "^(?!Num.Obs.|R2 Adj.)",
             add_row = desc_row)
```

	(1)	(2)	(3)	(4)	(5)	(6)
(Intercept)	4.607***	4.283***	4.284***			
	(0.002)	(0.007)	(0.007)			
required_car_parking_space1	0.112***	0.070***	0.016	0.029**	0.034**	0.061***
	(0.011)	(0.007)	(0.011)	(0.007)	(0.009)	(0.008)
no_of_special_requests		0.006+	0.004	0.011	0.000	0.000
		(0.003)	(0.003)	(0.011)	(0.008)	(0.008)
no_of_people		0.163***	0.163***	0.083**	0.082**	0.082**
		(0.002)	(0.002)	(0.015)	(0.014)	(0.014)

no_of_nights		-0.017***	-0.017***	-0.020***	-0.014***	-0.014***
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
log(lead_time)		-0.011***	-0.011***	-0.005**	-0.030***	-0.029***
		(0.001)	(0.001)	(0.001)	(0.003)	(0.003)
repeated_guest1		-0.159***	-0.157***	-0.161***	-0.164***	-0.145***
		(0.011)	(0.011)	(0.005)	(0.006)	(0.010)
OnlineTRUE		0.150***	0.150***	0.127*	0.113**	0.112**
		(0.004)	(0.004)	(0.039)	(0.025)	(0.025)
CorporateTRUE		-0.008	-0.007	-0.053	-0.052*	-0.051*
		(0.008)	(0.008)	(0.031)	(0.014)	(0.015)
ComplementaryTRUE		-0.491***	-0.489***	-0.543+	-0.543+	-0.531+
		(0.146)	(0.146)	(0.247)	(0.240)	(0.241)
AviationTRUE		0.258***	0.260***	0.131***	0.002	0.000
		(0.013)	(0.012)	(0.008)	(0.021)	(0.022)
required_car_parking_space1 × no_of_special_requests			0.076***	0.060***	0.053***	0.037**
			(0.014)	(0.007)	(0.007)	(0.006)
required_car_parking_space1 × repeated_guest1						-0.136**
						(0.029)
Num.Obs.	35674	35674	35674	35674	35674	35674
R2 Adj.	0.004	0.253	0.253	0.351	0.509	0.510
Фіксовані ефекти кімнат	Ні	Ні	Ні	Так	Так	Так
Часові фіксовані ефекти	Ні	Ні	Ні	Ні	Так	Так

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Питання №4

4. Чи є вплив повторного гостя на середню ціну за кімнату?

- При дослідженні, проведеному в минулій роботі було показано, що повторний гість має декілька цікавих особливостей, зокрема, як ми побачили по довірчих інтервалах, він платить менше, причому різниця скаладає приблизно 30% в меншу сторону для повторного гостя. Дуже важко повірити, що різниця настільки велика для повторного гостя. Тому необхідно провести регресійний аналіз.
- Спочатку побудуємо базову модель в якій зазначимо лише залежність логарифмованої ціни від повторного гостя, і отримаємо схожий результат. Але у нас є всі сумніви щодо цієї моделі.
- Далі побудуємо модель вказавши контрольні змінні, такі як: кількість ночей, кількість людей, логаритмований час до прибуття, необхідність в паркувальному місці, ринковий сегмент, - це основні змінні, які можуть корелювати з повторним гостем. Тепер видно вже кращий результат - вплив повторного гостя на ціну вже не такий великий, проте значущий. Але і дана модель не викликає повної довіри, бо ми маємо дані панельної природи і, відповідно, потрібно врахувати вплив ефектів кімнат і часу.
- По черзі побудуємо відповідні моделі (3 і 4), бачимо, що вплив повторного гостя залишається значущим, але при врахуванні ефектів став дещо більший. Тобто, можемо сказати, що повторний гість дійсно платить менше, але не так багато, як ми вважали спочатку, а всього приблизно на 16%, в це віриться значно більше ніж в попередній результат. Пояснити це можна тим, що повторні гості мають якіть привілеї, наприклад знижки чи особливі пропозиції, або просто на власному досвіді вже знають як можна заплатити менше.
- Також, ми знаємо про існування деякого зв'язку між необхідністю в паркувальному місці і повторним гостем - повторні гості частіше приїжджають власним авто, аніж нові гості. Тому, доцільно буде додати фактор взаємодії в кінцеву модель. Бачимо, що вплив повторного гостя на ціну незначно впав, але якщо подивитися на коефіцієнт при факторі взаємодії побачимо дуже цікаву картину. Виявляється, що повторні гості з власним автомобілем платять ЗНАЧНО менше - ще на 15%. Дуже важко пояснити такий вплив - єдина думка, що, можливо це якісь далекобійники, які платять лише символічну плату "за кімнату" і сплять в автомобілі.

Винести на перезентацію (В презентації)

```
model_guest <-  feols(log_price ~ repeated_guest, data = hotel, vcov = "HC1")

model_guest_ext <- feols(log_price ~ repeated_guest + no_of_nights + no_of_people + log(lead_time) + required_car
_parking_space + Online + Corporate + Complementary + Aviation, data = hotel_lead, vcov = "HC1")

model_guest_room <- feols(log_price ~ repeated_guest + no_of_nights + no_of_people + log(lead_time) + required_ca
r_parking_space + Online + Corporate + Complementary + Aviation  | room_type_reserved, data = hotel_lead)

model_guest_room_time <- feols(log_price ~ repeated_guest + no_of_nights + no_of_people + log(lead_time) + requir
ed_car_parking_space + Online + Corporate + Complementary + Aviation  | room_type_reserved + arrival_year_and_mon
th, data = hotel_lead)

model_guest_nons <- feols(log_price ~ repeated_guest*required_car_parking_space + no_of_nights + no_of_people + l
og(lead_time) + Online + Corporate + Complementary + Aviation | room_type_reserved + arrival_year_and_month, data
= hotel_lead)

desc_row <- tibble(title = c("Фіксовані ефекти кімнат", "Часові фіксовані ефекти"),
  model_guest = c("Ні", "Ні"),
  model_guest_ext = c("Ні", "Ні"),
  model_guest_room = c("Так", "Ні"),
  model_guest_room_time = c("Так", "Так"),
  model_guest_nons = c("Так", "Так"))


modelsummary(list(model_guest, model_guest_ext, model_guest_room, model_guest_room_time, model_guest_nons),
             stars = TRUE,
             gof_omit = "^(?!Num.Obs.|R2 Adj.)",
             add_row = desc_row)
```

	(1)	(2)	(3)	(4)	(5)
(Intercept)	4.617***	4.283***			
	(0.002)	(0.007)			
repeated_guest1	-0.315***	-0.158***	-0.161***	-0.165***	-0.144***
	(0.009)	(0.011)	(0.004)	(0.005)	(0.008)
no_of_nights		-0.017***	-0.020***	-0.014***	-0.014***
		(0.001)	(0.001)	(0.001)	(0.001)
no_of_people		0.164***	0.085***	0.082***	0.082***
		(0.002)	(0.014)	(0.013)	(0.013)
log(lead_time)		-0.011***	-0.006**	-0.030***	-0.029***
		(0.001)	(0.001)	(0.003)	(0.003)
required_car_parking_space1		0.071***	0.074***	0.072***	0.088***
		(0.007)	(0.003)	(0.004)	(0.004)
OnlineTRUE		0.152***	0.132**	0.113**	0.112**
		(0.003)	(0.035)	(0.022)	(0.021)
CorporateTRUE		-0.008	-0.054	-0.053*	-0.051*
		(0.008)	(0.030)	(0.014)	(0.015)
ComplementaryTRUE		-0.490***	-0.543+	-0.544+	-0.531+
		(0.145)	(0.247)	(0.241)	(0.243)
AviationTRUE		0.257***	0.127***	0.001	-0.001
		(0.013)	(0.011)	(0.018)	(0.020)
repeated_guest1 × required_car_parking_space1					-0.145**
					(0.028)
Num.Obs.	35674	35674	35674	35674	35674

R2 Adj.	0.024	0.253	0.351	0.509	0.510
Фіксовані ефекти кімнат	Ні	Ні	Так	Так	Так
Часові фіксовані ефекти	Ні	Ні	Ні	Так	Так

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Питання №5

(Залишається тільки в звіті)

5. Чи є вплив кількості людей на середню ціну за кімнату?

- Ми вже неодноразово бачили що збільшення кількості дорослих і дітей має зв'язок зі збільшенням ціни за кімнату, але чи можна вважати, що він причинно-наслідковий?
- Спочатку побудуємо примітивну модель (1) і на ній бачимо, що збільшення дорослого на 1 призведе до підвищення ціни на 16%, а збільшення дитини на 1 аж на 25%. У нас немає приводу довіряти цій моделі.
- Розширимо дану модель додавши контрольні змінні (2), які можуть корелювати з кількістю людей - це кількість ночей і сегмент ринку, також додамо фактор взаємодії між дітьми і дорослими, бо логічно припустити, що існують зміни для кожного з цих регресорів в залежності від іншого. Бачимо, що значення коефіцієнтів зменшилися, особливо для кількості дітей. Але знову є всі здогадки, що модель не є ідеальною.
- Моделі 3 і 4 відповідно додають фіксацію щодо впливу типу кімнати і часового проміжку. Бачимо, що коефіцієнти знову змінилися - дещо зросли для дітей і впали для дорослих, тут ми вже врахували всі доступні нам ефекти і можемо вважати, що збільшення кількості дітей і дорослих на одиницю має зв'язок зі збільшенняи ціни за кімнату на 6 і 8 відсотків відповідно.
- Протестуємо дану модель на стійкість (5, 6): спочатку (5) перетворимо кількість дітей на бінарну змінну там де їх багато(більше ніж 1) і мало, бачимо, що коефіцієнт при кількості дорослих не сильно змінився, тобто модель є стійкою відносно кількості дорослих. Далі (6) зробимо схоже перетворення кількості дорослих на бінарну змінну - там де їх багато(більше ніж 2) і мало, коефіцієнт біля кількості дітей значно змінився, тобто модель не стійка щодо цієї змінної і могли б існувати інші змінні для зменшення OVB.

```
model_people <- feols(log_price ~ no_of_adults + no_of_children, data = hotel, vcov = "HC1")
model_people_ext <- feols(log_price ~ no_of_adults*no_of_children + no_of_nights + Online + Corporate + Complemen
tary + Aviation  , data = hotel_market, vcov = "HC1")
model_people_room <- feols(log_price ~ no_of_adults*no_of_children + no_of_nights + Online + Corporate + Compleme
ntary + Aviation |room_type_reserved   , data = hotel_market)
model_people_time <- feols(log_price ~ no_of_adults*no_of_children + no_of_nights + Online + Corporate + Compleme
ntary + Aviation |room_type_reserved + arrival_year_and_month , data = hotel_market)
#стійкість
model_people_st <- feols(log_price ~ no_of_adults*no_of_children + no_of_nights + Online + Corporate + Complement
ary + Aviation |room_type_reserved + arrival_year_and_month , data = hotel_market %>% mutate (no_of_children = if
else(no_of_children > 1, 1, 0)))
model_people_st_2 <- feols(log_price ~ no_of_adults*no_of_children + no_of_nights + Online + Corporate + Compleme
ntary + Aviation |room_type_reserved + arrival_year_and_month , data = hotel_market %>% mutate (no_of_adults = if
else(no_of_adults > 2, 1, 0)))

modelsummary(list(model_people, model_people_ext, model_people_room, model_people_time, model_people_st, model_pe
ople_st_2),
             stars = TRUE,
             gof_omit = "^(?!Num.Obs.|R2 Adj.)"
             )
```

	(1)	(2)	(3)	(4)	(5)	(6)
(Intercept)	4.281***	4.316***				
	(0.006)	(0.006)				
no_of_adults	0.164***	0.119***	0.072+	0.061+	0.062+	0.215***
	(0.003)	(0.003)	(0.030)	(0.029)	(0.026)	(0.012)
no_of_children	0.246***	0.051***	0.078*	0.048	0.102+	0.121*
	(0.004)	(0.014)	(0.030)	(0.027)	(0.043)	(0.037)
no_of_nights		-0.017***	-0.021***	-0.018***	-0.018***	-0.017***
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
OnlineTRUE		0.163***	0.137**	0.139**	0.144***	0.143***
		(0.003)	(0.037)	(0.024)	(0.021)	(0.023)

CorporateTRUE		-0.049***	-0.090+	-0.050	-0.052	-0.078**
		(0.007)	(0.042)	(0.030)	(0.028)	(0.015)
ComplementaryTRUE		-0.510***	-0.558+	-0.518+	-0.503+	-0.526+
		(0.151)	(0.234)	(0.232)	(0.224)	(0.223)
AviationTRUE		0.241***	0.116***	0.063***	0.064**	0.036
		(0.011)	(0.016)	(0.009)	(0.011)	(0.027)
no_of_adults × no_of_children		0.092***	0.026	0.040	0.031	-0.061
		(0.007)	(0.034)	(0.031)	(0.032)	(0.058)
Num.Obs.	35674	35674	35674	35674	35674	35674
R2 Adj.	0.179	0.257	0.346	0.490	0.481	0.504

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

```
linearHypothesis(model_people_st, c("no_of_children = 0", "no_of_adults:no_of_children = 0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## no_of_children = 0
## no_of_adults:no_of_children = 0
##
## Model 1: restricted model
## Model 2: log_price ~ no_of_adults * no_of_children + no_of_nights + Online +
##     Corporate + Complementary + Aviation | room_type_reserved +
##     arrival_year_and_month
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1  35644
## 2  35642  2 19.339  6.319e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(model_people_st_2, c("no_of_adults = 0", "no_of_adults:no_of_children = 0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## no_of_adults = 0
## no_of_adults:no_of_children = 0
##
## Model 1: restricted model
## Model 2: log_price ~ no_of_adults * no_of_children + no_of_nights + Online +
##     Corporate + Complementary + Aviation | room_type_reserved +
##     arrival_year_and_month
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1  35644
## 2  35642  2 431.46  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Додаток

дослідимо наскільки більша ймовірність відмови бронювання для повторних гостей порівняно з неповторними

```
numeric_means <- hotel %>%
  dplyr::select(lead_time, avg_price_per_room, no_of_adults, no_of_children, no_of_nights) %>%
  summarize(across(everything(), ~ mean(.)))

new_data <- cbind(numeric_means,
                  data.frame(factorno_of_special_requests = c(0, 1),
                             required_car_parking_space = factor(0, levels = levels(hotel$required_car_parking_sp
ace)),
                             repeated_guest = factor(0, levels = levels(hotel$repeated_guest))))

new_data <- rbind(new_data, new_data)
new_data[2, "factorno_of_special_requests"] <- 1

prediction_logit <- predict(requests_logit, new_data, type = "response")

difference <- prediction_logit[2] - prediction_logit[1]
difference
```

```
## 2
## 0
```

кореляції (пріколи)