

Лабораторна робота №3

Регресійний аналіз

Команда

Долинний Денис (КМ-01)

Ганушкévич Євгеній (КМ-02)

Рижкóва Дар'я (КМ-02)

Грінів Юрій (КМ-02)

Голінський Денис (КМ-02)

Про датасет

Датасет створений шляхом SQL-запитів до бази даних Hotel property management systems. Дані відображають готелі у Португалії за 2015-2017 роки. Датасет має 119390 спостережень і 32 змінні.



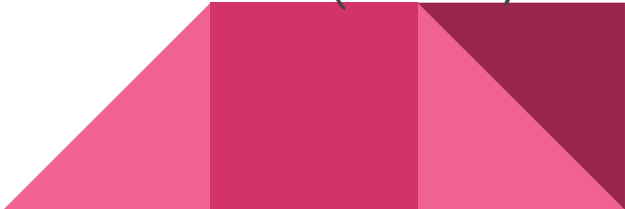
Додаткова робота з даними

Була створена змінна `market_segment_b`, яка показує сегмент ринку, до якого належить гість (онлайн – 1, інше – 0).

Була створена змінна `distribution_channel_b`, як було зроблено бронювання (через TA/TO – 1, інше – 0).

Була створена змінна `deposit_type_b`, яка показує, чи був зроблений депозит (так – 1, ні – 0).

Була створена змінна `with_meal`, яка показує, чи була замовлена їжа (так – 1, ні – 0).



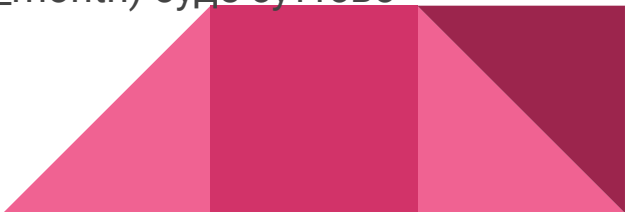
Наші гіпотези

У якості залежної змінної було взято adr (прибуток готелю за 1 номер за 1 ніч).

Висуваємо такі гіпотези щодо коефіцієнтів:

- 1) lead-time<0
- 2) all-nights<0
- 3) booking-changes>0
- 4) all-guests<0
- 5) total-of-special-request>0
- 6) market-segment-b<0
- 7) distribution-channel-b<0
- 8) deposit-type-b<0
- 9) with-meal>0

Також ми очікуємо, що місяць прибуття (arrival_date_month) буде суттєво впливати на adr.



Побудова моделей

Наша основна модель:

Table 1: Regression

<i>Dependent variable:</i>	
Середнє adr	
lead_time	−0.041*** (−0.043, −0.039)
Constant	107.979*** (107.591, 108.367)
Observations	116,920
Adjusted R ²	0.009
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Які висновки можна зробити

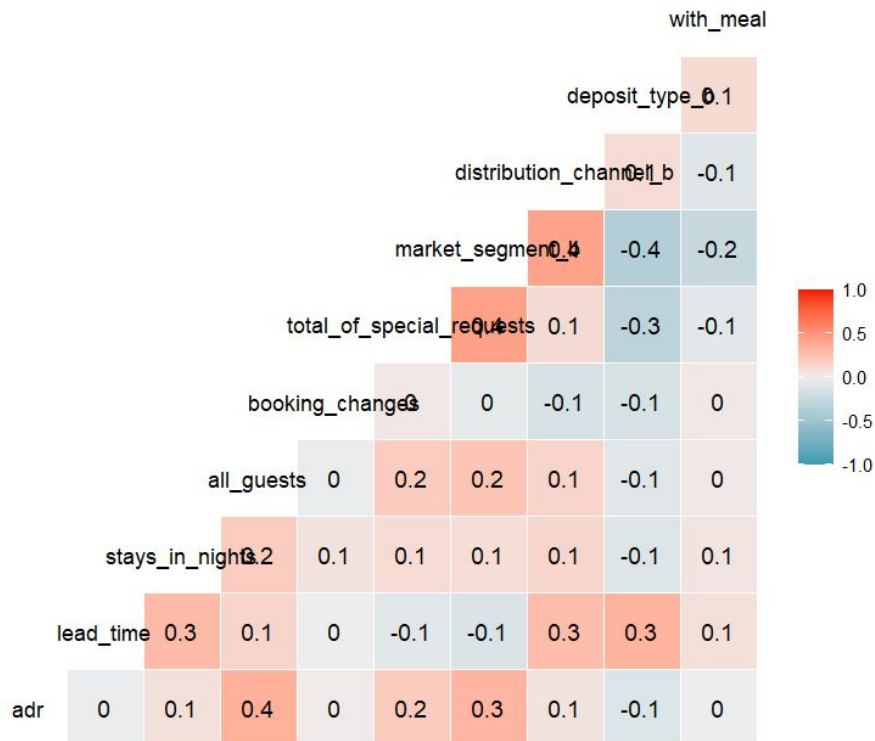
Можемо побачити, що збільшення `lead_time` на 1 змінюється `adr` на 0,041 (що дуже мало) та що коефіцієнт статистично значущий.

Проаналізувавши відповідні середньоквадратичні відхилення, ми побачимо, що збільшення середньоквадратичного відхилення `lead_time` на 1 зменшує одиницю середньоквадратичного відхилення `adr` на 9%.

Також, дивлячись на R^2 можна сказати, що частка `lead_time` у варіації `adr` **дуже** мала.



Проаналізуємо кореляційну матрицю:



Що ми побачили?

У нас немає змінних, що сильно корелюють між собою тому ми можемо включити усіх їх у нашу модель. Побудуємо 2 нові моделі. Перша включає в себе усі фактори + поліном другого порядку `all_guests`, що розглядаються, друга - ті, у яких кореляція 0.



Результат побудови моделей(1)

Table 1: Multiple regression

	Середнє adr		
	(1)	(2)	(3)
lead_time	−0.041*** (−0.043, −0.039)	−0.022*** (−0.024, −0.020)	
stays_in_nights		0.237*** (0.133, 0.341)	−0.005 (−0.107, 0.096)
all_guests		−13.319*** (−14.972, −11.666)	27.927*** (27.449, 28.405)
I(all_guests^2)		9.077*** (8.685, 9.468)	
booking_changes		1.346*** (0.911, 1.781)	
total_of_special_requests		3.029*** (2.667, 3.391)	2.440*** (2.076, 2.805)

Результат побудови моделей(2)

market_segment_b		23.852*** (23.256, 24.447)	24.236*** (23.676, 24.797)
distribution_channel_b		-11.326*** (-12.098, -10.553)	-16.316*** (-17.063, -15.568)
deposit_type_b		9.033*** (8.360, 9.706)	5.611*** (4.930, 6.292)
with_meal		8.019*** (7.362, 8.676)	
Constant	107.979*** (107.591, 108.367)	79.863*** (79.475, 80.251)	48.278*** (47.890, 48.667)
Observations	116,920	116,920	116,920
Adjusted R ²	0.009	0.274	0.244

Note:

*p<0.1; **p<0.05; ***p<0.01

Висновки, щодо нових моделей

Можемо побачити, що для моделі(2) усі коефіцієнти виявилися статистично значущими, але прибравши змінні, кореляція яких 0, коефіцієнт R^2 зменшиться всього на 0,03, тобто вилучення незначущих регресорів було виправдано. Також ми бачимо, що змінна `stays_in_nights` стала статистично незначущою.



Додатковий аналіз stays_in_nights

Протестувавши явно гіпотезу, що коефіцієнт при `stays_in_nights` = 0, отримаємо, що p-value дуже велике, тому немає підстав стверджувати, що `stays_in_nights` справді є статистично значущою. Змінюємо відповідно модель(3).



Отримаємо(1)

Table 1: Multiple regression

	Середнє adr		
	(1)	(2)	(3)
lead_time	−0.041*** (−0.043, −0.039)	−0.022*** (−0.024, −0.020)	
stays_in_nights		0.237*** (0.133, 0.341)	
all_guests		−13.319*** (−14.972, −11.666)	27.925*** (27.449, 28.400)
I(all_guests^2)		9.077*** (8.685, 9.468)	
booking_changes		1.346*** (0.911, 1.781)	
total_of_special_requests		3.029*** (2.667, 3.391)	2.440*** (2.075, 2.804)
market_segment_b		23.852*** (23.256, 24.447)	24.239*** (23.681, 24.797)
distribution_channel_b		−11.326*** (−12.098, −10.553)	−16.320*** (−17.062, −15.578)

Отримаємо(2)

deposit_type_b		9.033*** (8.360, 9.706)	5.616*** (4.944, 6.289)
with_meal		8.019*** (7.362, 8.676)	
Constant	107.979*** (107.591, 108.367)	79.863*** (79.475, 80.251)	48.267*** (47.878, 48.655)
Observations	116,920	116,920	116,920
Adjusted R ²	0.009	0.274	0.244

Note:

*p<0.1; **p<0.05; ***p<0.01

Далі...

Створюємо моделі, де, крім зазначених змінних, додається ще місяць (базовим було обрано квітень).



Остаточна модель (1)

Table 1: Multiple regression

	Середнє adr		
	(1)	(2)	(3)
factor(arrival_date_month)August	39.924	32.841	35.723
factor(arrival_date_month)December	-17.876	-17.630	-18.606
factor(arrival_date_month)February	-26.624	-24.786	-27.834
factor(arrival_date_month)January	-29.598	-26.345	-29.476
factor(arrival_date_month)July	26.733	21.094	25.528
factor(arrival_date_month)June	16.214	17.117	19.969
factor(arrival_date_month)March	-20.158	-18.117	-19.525
factor(arrival_date_month)May	8.602	11.044	13.035
factor(arrival_date_month)November	-26.125	-20.186	-21.176
factor(arrival_date_month)October	-11.827	-6.874	-4.958
factor(arrival_date_month)September	4.916	9.992	12.576
lead_time			-0.087*** (-0.089, -0.085)

Остаточна модель (2)

stays_in_nights			−0.129** (−0.233, −0.025)
all_guests		23.082*** (21.430, 24.735)	23.740*** (22.088, 25.393)
booking_changes			2.701*** (2.266, 3.136)
total_of_special_requests		1.633*** (1.271, 1.996)	1.709*** (1.347, 2.071)
market_segment_b		25.248*** (24.652, 25.843)	22.487*** (21.891, 23.083)
distribution_channel_b		−19.592*** (−20.364, −18.820)	−12.839*** (−13.611, −12.066)
deposit_type_b		6.736*** (6.063, 7.408)	14.860*** (14.187, 15.533)
with_meal			9.342*** (8.684, 9.999)
Constant	101.798*** (101.111, 102.486)	57.933*** (57.245, 58.621)	50.802*** (50.114, 51.489)
Observations	116,920	116,920	116,920
Adjusted R ²	0.224	0.406	0.438


Note:

*p<0.1; **p<0.05; ***p<0.01

Що можемо сказати?

З моделі(1) бачимо, що літні місяці суттєво збільшують adr , а зимові + березень і листопад суттєво зменшують, хоча ці коефіцієнти можуть не бути статистично значущими. Протестувавши гіпотези, що коефіцієнти для всіх місяців з $\text{factor}(\text{arrival_date_month}) = 0$ для всіх 3-х моделей, ми отримали, що $p\text{-value} < 2.2e-16$, тобто ці коефіцієнти все ж статистично значущі.


Також ми можемо побачити, що використання всіх коефіцієнтів не є виправданим, R^2 для моделі(2) і моделі(3) відрізняється всього лише на 0,03.



Висновки

- 1) lead-time<0 – підтверджено, хоча коефіцієнт впливає не сильно.
- 2) all-nights<0 – підтверджено, хоча коефіцієнт впливає не сильно.
- 3) booking-changes>0 – підтверджено, хоча статистично незначущий.
- 4) all-guests<0 – спростовано, коефіцієнт суттєво впливає на adr.
- 5) total-of-special-request>0 – підтверджено, хоча коефіцієнт впливає не сильно.
- 6) market-segment-b<0 – спростовано, коефіцієнт суттєво впливає на adr.
- 7) distribution-channel-b<0 – підтверджено, коефіцієнт суттєво впливає на adr.
- 8) deposit-type-b<0 – спростовано, коефіцієнт суттєво впливає на adr.
- 9) with-meal>0 підтверджено, хоча статистично незначущий.

Також, можемо підтвердити, що місяць сильно впливає на adr, причому літні суттєвого його збільшують, а зимові + березень і листопад суттєво зменшують.





Дякуємо за увагу!