

буду в 15.00, а пока:

- пожалуйста, включите камеру,
- назовитесь понятным именем
- и располагайтесь поудобнее



регрессия

Елена Эльзесер,
Наставник DA, Яндекс.Практикум

Яндекс Практикум

план

1. Зачем оно нам надо
2. Пример с разбором
3. Интерпретации
4. Типичные проблемы и их решение
5. Перерыв
6. Практика: Jupyter Notebook



зачем оно нам надо

1. Описательный анализ
2. Является ли изменение переменной X причиной изменения переменной Y ?
3. Прогнозирование

пример

Дано:

Price - цена
однокомнатной
квартиры (тыс. руб.)

TotalArea - размер
квартиры (кв. м).

	TotalArea	Price
count	121.000000	121.000000
mean	35.131405	5540.334537
std	3.518310	792.712745
min	24.400000	4218.000000
25%	32.600000	5101.430000
50%	35.000000	5389.700000
75%	37.900000	5877.027000
max	45.900000	8636.886000

пример. вопросы

Влияет ли площадь квартиры на цену? Как?
На сколько точен результат? На сколько можно ему
доверять?

Оценить параметры уравнения (1). R^2 (2)
Коэффициент при переменной площадь статистически
значим, при каком уровне значимости (3)?
Дайте содержательную интерпретацию коэффициента (4).
Найдите 95%-й доверительный интервал (5).

пример. расчеты

1. Подготовка данных:

```
 $\Sigma x = df['TotalArea'].sum()$ 
```

пример. расчеты

1. Подготовка данных:

$\Sigma x = df['TotalArea'].sum()$

$\Sigma x^2 = df['TotalArea'].pow(2).sum()$

пример. расчеты

1. Подготовка данных:

$\Sigma x = \text{df['TotalArea'].sum()}$

$\Sigma x^2 = \text{df['TotalArea'].pow(2).sum()}$

$\Sigma y = \text{df['Price'].sum()}$

$\Sigma y^2 = \text{df['Price'].pow(2).sum()}$

$\text{df['xy']} = \text{df['TotalArea'] * df['Price']}$

пример. расчеты

1. Подготовка данных:

$\Sigma x = \text{df['TotalArea'].sum()}$

$\Sigma x^2 = \text{df['TotalArea'].pow(2).sum()}$

$\Sigma y = \text{df['Price'].sum()}$

$\Sigma y^2 = \text{df['Price'].pow(2).sum()}$

$\text{df['xy']} = \text{df['TotalArea'] * df['Price']}$

$\Sigma x * y = \text{df['xy'].sum()}$

$n = \text{df['Price'].count()}$

пример. расчеты

1. Подготовка данных.
2. Расчет коэффициентов:

$$y = w_0 + w_1 * x$$

$$w_1 = 135; w_0 = 786$$

$$y = 786 + 135 * x$$

пример. расчеты

1. Подготовка данных.
2. Расчет коэффициентов.
3. Расчет R^2 :

$$\frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2}$$

$$\sum(y - \bar{y})^2$$


пример. расчеты

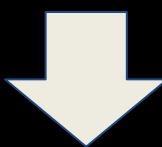
1. Подготовка данных.
2. Расчет коэффициентов.
3. Расчет R^2

$$\frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2}$$

$$\sum(y - \bar{y})^2$$

Но! Adj. R^2

1 - 

0 - 

пример. ответы

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.361
Model:	OLS	Adj. R-squared:	0.355

	coef	std err	t	P> t	[0.025	0.975]
Intercept	786.4562	583.051	1.349	0.180	-368.043	1940.956
TotalArea	135.3171	16.514	8.194	0.000	102.617	168.017

пример. ответы

$$\text{Price}_i = 786 + 135 * \text{TotalArea}_i$$

	coef	std err	t	P> t	[0.025	0.975]
¹ Intercept	786.4562	583.051	1.349	0.180	-368.043	1940.956
TotalArea	135.3171	16.514	8.194	0.000	102.617	168.017

пример. ответы

$$\text{Price} = 786 + 135 * \text{TotalArea}$$

при увеличении площади квартиры на 1 м² ее цена в среднем увеличивается на 135 тыс. руб.

	coef	std err	t	P> t	[0.025	0.975]
Intercept	786.4562	583.051	1.349	0.180	-368.043	1940.956
TotalArea	135.3171	16.514	8.194	0.000	102.617	168.017

пример. попробуем предсказать

$$x_{122} = 45$$

пример. попробуем предсказать

$$x_{122} = 45$$

$$y_{122} = 786 + 135 * 45 = 6861$$

$$y_{122} - s * t_{119}, y_{122} + s * t_{119}$$

5569

8181

интерпретация

0. Бессмысленна, если коэффициенты стат не значимы

интерпретация

0. Бессмысленна, если коэффициенты стат не значимы

1. Линейная модель:

$$\text{Price} = 786 + 135 * \text{TotalArea}$$

при увеличении площади на 1 кв.м. стоимость
увеличивается на 135 тыс. рублей

интерпретация

0. Бессмысленна, если коэффициенты стат не значимы

1. Линейная модель:

Застройщики: Тик, Вертолет, Ишимстрой

$$\text{Price} = 22 * \text{TotalArea} + 33 * \text{Тик} + 11 * \text{Вертолет}$$

при прочих равных

квартира Тика в среднем на 33 тыс дороже Ишимстроя

интерпретация

- 0. Бессмысленна, если коэффициенты стат не значимы
- 1. Линейная модель.
- 2. Линейно-логарифмическая:

$$\text{Price} = 2 + 4000 * \ln(\text{TotalArea})$$

при увеличении площади на 1% стоимость увеличивается
на 40 тыс. рублей

интерпретация

- 0. Бессмысленна, если коэффициенты стат не значимы
- 1. Линейная модель.
- 2. Линейно-логарифмическая.
- 3. Логарифмическая:

$$\ln(\text{Price}) = 6 + 0.8 * \ln(\text{TotalArea})$$

при увеличении площади на 1% стоимость увеличивается
на 0.8%

интерпретация

- 0. Бессмысленна, если коэффициенты стат не значимы
- 1. Линейная модель.
- 2. Линейно-логарифмическая.
- 3. Логарифмическая.
- 4. Логарифмически-линейная:

$$\ln(\text{Price}) = 8 + 0.02 * \text{TotalArea}$$

при увеличении площади на 1 кв.м. стоимость
увеличивается на 2%

интерпретация

- 0. Бессмысленна, если коэффициенты стат не значимы
- 1. Линейная модель.
- 2. Линейно-логарифмическая.
- 3. Логарифмическая.
- 4. Логарифмически-линейная:

$$\ln(y) = 8 + 15 * x$$

при увеличении x на 1 y увеличивается на $(e^{15}-1)*100\%$

проблемы

Где границы генеральной совокупности?

Чем плохо: нет внешней обоснованности - нет переноса на другие генсовокупности

Как детектить: ?

Как чинить: гарантировать отсутствие различий в генсовокупности и условиях

проблемы

Выбросы

Чем плохо: неоднородные данные

Как детектить: `describe()`, здравый смысл

Как чинить: удалять

проблемы

Самоотбор

Чем плохо: оценка будет смещена, а, может быть, даже несостоятельна

Как детектить: здравый смысл, коэффициенты

Как чинить: продумывать дизайн исследования, ...

проблемы

Пропуск существенной переменной

Чем плохо: оценка будет смещена, а, может быть, даже несостоятельна

Как детектить: здравый смысл, коэффициенты, тесты (Саргана, Хаусмана)

Как чинить №1: добавить :) и посмотреть на $\text{adj } R^2$

проблемы

Пропуск существенной переменной

Как чинить №2:

1. Переменная-заменитель
2. Инструментальные переменные
3. Панель
4. Контролируемый эксперимент

проблемы

Добавление ненужной переменной

Чем плохо: снижается точность модели

Как детектить: здравый смысл, статзначимость коэффициентов

Как чинить: удалять

проблемы

Ошибки измерения

Чем плохо: несостоятельная оценка коэффициента

Как детектить: здравый смысл, исследования данных

Как чинить: инструментальные переменные*

проблемы

Выбор неверного уравнения

Чем плохо: оценка будет смещена, а, может быть, даже несостоятельна

Как детектить: здравый смысл, графический анализ данных, тесты

Как чинить: см. как детектить

проблемы

Мультиколлинеарность

Чем плохо: неустойчивость, неправдоподобность, незначимость

Как детектить: $VIF > 10$, незначимость

Как чинить: удалить, скомбинировать, см. тренажер
“Регуляризация”

проблемы

Гетероскедастичность

Чем плохо: дисперсия случайной ошибки не является постоянной => стандартными ошибками оценок коэффициентов нельзя пользоваться

Как детектить: тесты Бройша-Пагана, Голдфелда-Квандта

Как чинить: **НС0, НС1, НС2, НС3**; прологорифмировать

перерыв



10 минут



Спасибо!