Практическое занятие №2 «**Множественный регрессионный анализ**»

по дисциплине «Многомерный статистический анализ в социологических исследованиях»

План занятия

1. Множественный регрессионный анализ.
2. Решение задач.

Про корреляцию & регрессию

- Задача корреляционного анализа – определение тесноты и направления связи между изучаемыми величинами.
- В ходе ***регрессионного анализа*** определяется аналитическое выражение связи зависимой случайной величины Y (результативный признак) с независимыми случайными величинами X_1, X_2, …X_m (факторами).

Зачем?

- Регрессия используется для анализа воздействия на отдельную зависимую переменную значений одной или нескольких независимых переменных.
- Например, на спортивные качества атлета влияют несколько факторов, включая возраст, рост и вес.
- Можно вычислить степень влияния каждого из этих трех факторов по результатам выступления спортсмена, а затем использовать полученные данные для предсказания выступления другого спортсмена.

Задачи регрессионного анализа

- При помощи регрессионного анализа возможно решение ***задачи прогнозирования***. Прогнозные значения вычисляются путем подстановки в уравнение регрессии параметров значений объясняющих переменных.

Основные задачи регрессионного анализа

1. установление формы зависимости,
2. определение функции регрессии,
3. оценка неизвестных значений зависимой переменной.

1 задача - Установление формы зависимости.

Характер и форма зависимости между переменными могут образовывать следующие разновидности регрессии:

- положительная линейная регрессия (выражается в равномерном росте функции);
- положительная равноускоренно возрастающая регрессия;
- положительная равнозамедленно возрастающая регрессия;
- отрицательная линейная регрессия (выражается в равномерном падении функции);
- отрицательная равноускоренно убывающая регрессия;
- отрицательная равнозамедленно убывающая регрессия.

Однако описанные разновидности обычно встречаются не в чистом виде, а в сочетании друг с другом. В таком случае говорят о комбинированных формах регрессии.

Задачи регрессионного анализа

2 задача - Определение функции регрессии.

- Вторая задача сводится к выяснению действия на зависимую переменную главных факторов или причин, при неизменных прочих равных условиях, и при условии исключения воздействия на зависимую переменную случайных элементов. *Функция регрессии* определяется в виде математического уравнения того или иного типа.

3 задача - Оценка неизвестных значений зависимой переменной.

- Оценка значений зависимой переменной внутри рассматриваемого интервала исходных данных, т.е. пропущенных значений; при этом решается задача интерполяции.
- Оценка будущих значений зависимой переменной, т.е. нахождение значений вне заданного интервала исходных данных; при этом решается задача экстраполяции.
- Обе задачи решаются путем подстановки в уравнение регрессии найденных оценок параметров значений независимых переменных. Результат решения уравнения представляет собой оценку значения целевой (зависимой) переменной.

Уравнение регрессии -

это форма связи результативного признака Y с факторами X1, X2, …Xm.

В зависимости от типа выбранного уравнения различают линейную и нелинейную (квадратичную, экспоненциальную, логарифмическую и т. д.) регрессию.

Парная и множественная регрессия

- В зависимости от числа взаимосвязанных признаков различают парную и множественную регрессию.
- **Парная** – исследуется связь между двумя признаками (результативным и факторным).
- **Множественная** (многофакторная) – между тремя признаками (результативным и несколькими факторными).

Уравнение регрессии

- **Уравнение регрессии выглядит следующим образом: Y=a+b*X**
- При помощи этого уравнения переменная Y выражается через константу a и угол наклона прямой (или угловой коэффициент) b, умноженный на значение переменной X. Константу a также называют свободным членом, а угловой коэффициент - коэффициентом регрессии или B-коэффициентом.
- В большинстве случав (если не всегда) наблюдается определенный разброс наблюдений относительно регрессионной прямой.
- ***Остаток*** - это отклонение отдельной точки (наблюдения) от линии регрессии (предсказанного значения).
- Для решения задачи регрессионного анализа в MS Excel выбираем в меню **Сервис "Пакет анализа"** и инструмент анализа "Регрессия". Задаем входные интервалы X и Y. Входной интервал Y - это диапазон зависимых анализируемых данных, он должен включать один столбец. Входной интервал X - это диапазон независимых данных, которые необходимо проанализировать. Число входных диапазонов должно быть не больше 16.

Этапы регрессионного анализа

1. Задание аналитической формы уравнения регрессии и определение параметров регрессии.
2. Определение в регрессии степени стохастической взаимосвязи результативного признака и факторов, проверка общего качества уравнения регрессии.
3. Проверка статистической значимости каждого коэффициента уравнения регрессии и определение их доверительных интервалов.

Предположения, на которые опирается РА

- Предположение линейности, т.е. предполагается, что связь между рассматриваемыми переменными является линейной. Так, в рассматриваемом примере мы построили диаграмму рассеивания и смогли увидеть явную линейную связь. Если же на диаграмме рассеивания переменных мы видим явное отсутствие линейной связи, т.е. присутствует нелинейная связь, следует использовать нелинейные методы анализа.
- Предположение о нормальности *остатков*. Оно допускает, что распределение разницы предсказанных и наблюдаемых значений является нормальным. Для визуального определения характера распределения можно воспользоваться гистограммами *остатков*.
- При использовании регрессионного анализа следует учитывать его ***основное ограничение***. Оно состоит в том, что регрессионный анализ позволяет обнаружить лишь зависимости, а не связи, лежащие в основе этих зависимостей.
- Регрессионный анализ дает возможность оценить степень связи между переменными путем вычисления предполагаемого значения переменной на основании нескольких известных значений.

Таким образом,

- Регрессионный анализ позволяет установить степень влияния независимых величин на зависимую переменную.
- При помощи регрессионного анализа возможно решение задачи прогнозирования.
- **Уравнение регрессии выглядит следующим образом: Y=a+b*X**
- Прогнозные значения вычисляются путем подстановки в уравнение регрессии параметров значений объясняющих переменных
- Используем пакет «Регрессия».

2 вопрос занятия – решение задач с помощью методов линейной регрессии

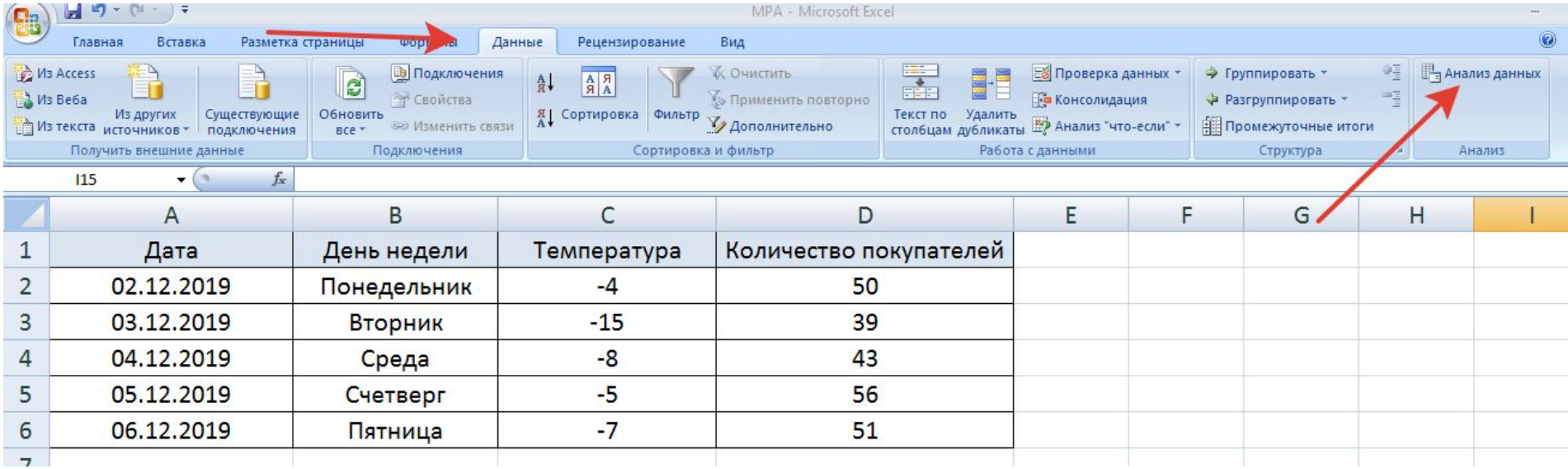
Подключение пакета анализа

- Анализ данных в Microsoft Excel Microsoft Excel имеет большое число статистических функций. Некоторые являются встроенными, некоторые доступны после установки **пакета анализа**.
- Средства, включенные в пакет анализа данных, доступны через команду С*ервис == Анализ данных.* Если эта команда отсутствует в меню, в меню *Сервис/Надстройки* необходимо активировать пункт "Пакет анализа".
- Пошаговый алгоритм есть здесь https://lumpics.ru/regression-analysis-in-excel/

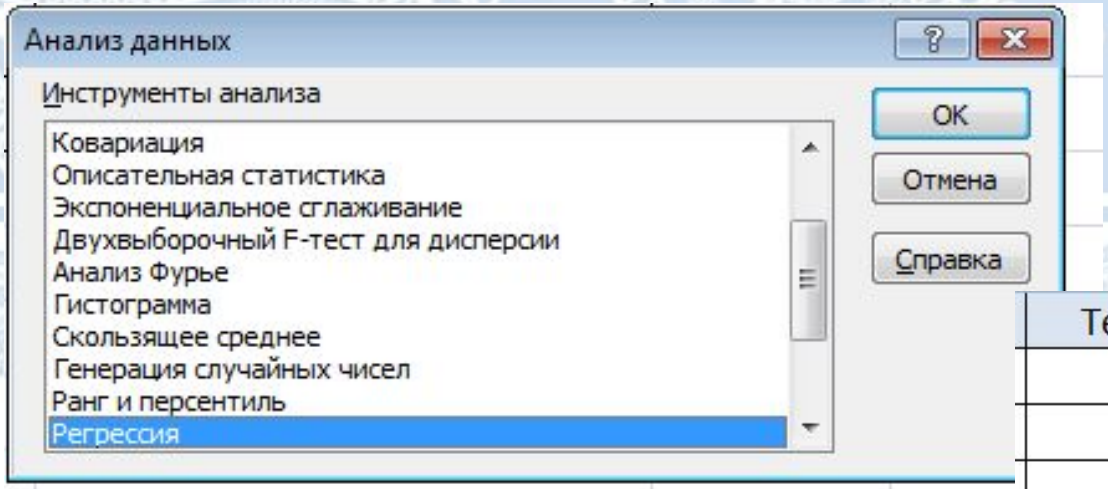
Создаем базу данных

	A	B	C	D
1	Дата	День недели	Температура	Количество покупателей
2	02.12.2019	Понедельник	-4	50
3	03.12.2019	Вторник	-15	39
4	04.12.2019	Среда	-8	43
5	05.12.2019	Счетверг	-5	56
6	06.12.2019	Пятница	-7	51
7				

В новой верхней вкладке «Данные» выбираем меню «Анализ данных»



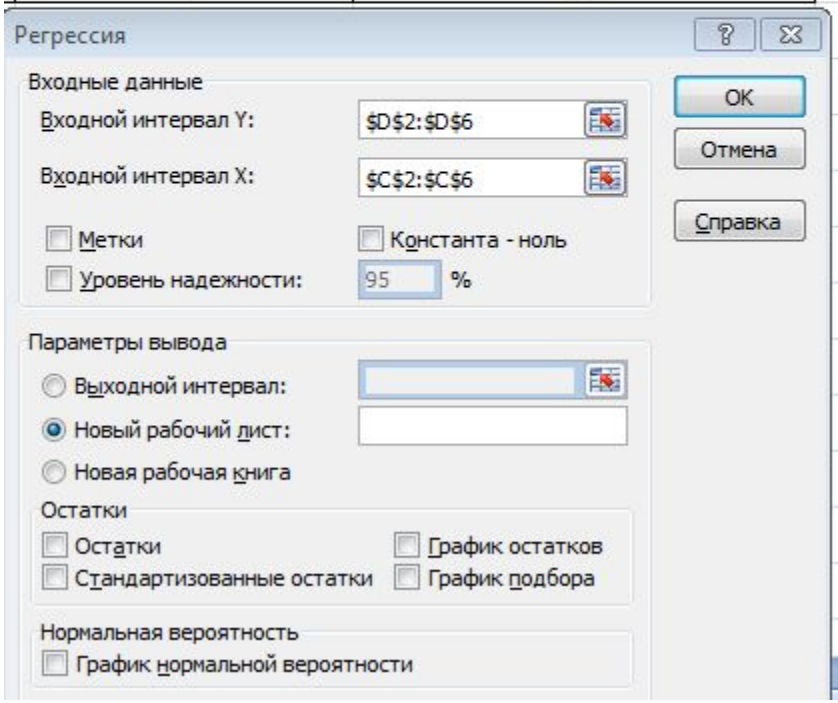
	A	B	C	D	E	F	G	H	I
1	Дата	День недели	Температура	Количество покупателей					
2	02.12.2019	Понедельник	-4	50					
3	03.12.2019	Вторник	-15	39					
4	04.12.2019	Среда	-8	43					
5	05.12.2019	Счетверг	-5	56					
6	06.12.2019	Пятница	-7	51					



Температура	Количество покупателей
-4	50
-15	39
-8	43
-5	56
-7	51

Количество покупателей – входной интервал Y

Температура – входной интервал X



OUTPUT (вывод итогов)

ВЫВОД ИТОГОВ									
Регрессионная статистика									
Множественный R	0,839793663								
R-квадрат	0,705253396								
Нормированный R-квад	0,607004528								
Стандартная ошибка	4,237911402								
Наблюдения	5								
Дисперсионный анализ									
	df	*SS*	*MS*	*F*	*Значимость F*				
Регрессия	1	128,9203209	128,920321	7,17823433	0,075098537				
Остаток	3	53,87967914	17,959893						
Итого	4	182,8							
	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*	
Y-пересечение	58,04010695	4,266145634	13,6048114	0,00085903	44,46332754	71,61688636	44,46332754	71,61688636	
Переменная X 1	1,312834225	0,490005635	2,67922271	0,07509854	-0,246582397	2,872250846	-0,246582397	2,872250846	

Разбор результатов анализа

1 шаг – установить наличие статистически значимой линейной связи между переменными

Одним из основных показателей является **R-квадрат**. В нем указывается качество модели.

В нашем случае данный коэффициент равен 0,705 или около 70,5%. Это приемлемый уровень качества. Следовательно, можно построить уравнение регрессии

Зависимость менее 0,5 является плохой. В этом случае уравнение регрессии построить нельзя. Анализ на этом заканчивается.

ВЫВОД ИТОГОВ	
Регрессионная статистика	
Множественный R	0,839793663
R-квадрат	0,705253396
Нормированный R-квад	0,607004528
Стандартная ошибка	4,237911402
Наблюдения	5

Разбор результатов анализа

2 шаг – доказать значимость линейной модели (дисперсионный анализ)

- В данном шаге нужно указать вероятность, с которой независимая переменная (время) влияет на зависимую (успеваемость).
- Оценка значимости уравнения регрессии в целом производится на основе F -критерия Фишера.
- В данном примере F=7,18, которому соответствует уровень значимости 0,07. Это фразу следует расшифровывать следующим образом: с вероятностью 93% можно утверждать, что температура воздуха влияет на количество покупателей

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	128,9203209	128,920321	7,17823433	0,075098537
Остаток	3	53,87967914	17,959893		
Итого	4	182,8			

Разбор результатов анализа

3 шаг – составить уравнение регрессии, доказать значимость коэффициента и свободного члена построенного уравнения.

- Для построения модели линейной регрессии из данной таблицы используется коэффициент Y-пересечения.
- Оценка его значимости проводится по t-критерию Стьюдента. В данном случае уровень значимости t-критерия Стьюдента меньше 0,001 (равен 0,0008), следовательно, можно говорить о статистической значимости коэффициента Y-пересечения.
- В случае, если уровень значимости t-критерия Стьюдента (p-значение) меньше, чем 0,05, уравнение регрессии

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	58,04010695	4,266145634	13,60481145	0,000859033
Переменная X 1	1,312834225	0,490005635	2,67922271	0,075098537

Разбор результатов анализа

3 шаг – составить уравнение регрессии, доказать значимость коэффициента и свободного члена построенного уравнения.

Математическое уравнение, которое оценивает линию простой (парной) линейной регрессии:

Y=a+bX, где

X - независимая переменная,

Y – зависимая переменная (или переменная отклика). Это значение, которое мы ожидаем для y (в среднем), если мы знаем величину x, т.е. это «предсказанное значение y»

a – свободный член (пересечение) линии оценки; это значение Y, когда X=0,

b – угловой коэффициент или градиент оценённой линии; она представляет собой величину, на которую Y увеличивается в среднем, если мы увеличиваем X на одну единицу.

Уравнение регрессии в данном случае выглядит как:

Y (количество посетителей) = 58 + 1,3 * X.

Построение предсказательной модели

- Регрессионный анализ позволяет предсказать - на основе уравнения регрессии – вероятностный прогноз изменения исследуемых переменных.
- К примеру, мы хотим узнать, каково будет количество покупателей на следующей неделе.

Дата	День недели	Температура	Количество покупателей
02.12.2019	Понедельник	-4	50
03.12.2019	Вторник	-15	39
04.12.2019	Среда	-8	43
05.12.2019	Счетверг	-5	56
06.12.2019	Пятница	-7	51
09.12.2019	Понедельник	-8	
10.12.2019	Вторник	-10	

Построение предсказательной модели

- заложим вычисленное уравнение регрессии **= 58 + 1,3 * X** в строку формул, где X – показатели температуры из прогноза погоды на следующую неделю.

D7 f_x =58+1,3*(C7)

A	B	C	D
Дата	День недели	Температура	Количество покупателей
02.12.2019	Понедельник	-4	50
03.12.2019	Вторник	-15	39
04.12.2019	Среда	-8	43
05.12.2019	Счетверг	-5	56
06.12.2019	Пятница	-7	51
09.12.2019	Понедельник	-8	47,6
10.12.2019	Вторник	-10	45

Ответ задачи

1. Уравнение линейной регрессионной зависимости числа покупателей от температуры воздуха **Y = 58 + 1,3* X.**
2. Прогноз числа покупателей для температуры -8 С равен 47,6 чел; для температуры -10 С равен 45 чел.
3. В целом можно говорить о температуры окружающей среды на количество покупателей в торговой точке.

Решение задач

Для каждой задачи необходимо выполнить 4 шага и записать ответ

1. установить наличие статистически значимой линейной связи между переменными
2. доказать значимость линейной модели (дисперсионный анализ)
3. составить уравнение регрессии, доказать значимость коэффициента и свободного члена построенного уравнения
4. рассчитать прогнозные показатели

Задача 1

- Исследователь пытается выявить взаимосвязь между количеством времени X, бесполезно потраченного студентами, и средним баллом Y их академической успеваемости, который варьируется в пределах от 2,0 до 5,0. Под потраченным без пользы временем понимается количество часов определенного соответствующего времяпровождения в неделю (например, занятого просмотром телесериалов). Данные для выборки студентов приведены в таблице.
- Требуется построить линейную регрессионную зависимость среднего балла успеваемости от показателя бесполезно потраченного времени, а также выполнить прогноз успеваемости для значений X, равных 20, 30 и 40 часов.

№	X	Y
1	42	2,8
2	23	4
3	31	3,2
4	35	3,9
5	16	4,7
6	26	4
7	39	3,4
8	19	4,4
9	29	3,8

Задача 2

- Исследователями были изучены данные о расходах потребителей на питание за 1959-1983 годы (данные на следующем слайде).
- Требуется вычислить уравнение регрессии между расходами потребителя на питание (Y) и располагаемым личным доходом (X) по данным, приведенным для США за период с 1959 по 1983 год.
- Исследователю хотелось бы предсказать расход на питание в 1984 году при личном доходе потребителя 1 239,3.

Личные потребительские расходы на питание населения с 1959 по 1983 год

Год	1959	1960	1961	1962	1963	1964	1965	1966	1967	1968
x	479,7	489,7	503,8	524,9	542,3	580,8	616,3	646,8	673,5	701,3
y	99,7	100,9	102,5	103,5	104,6	108,8	113,7	116,6	118,6	123,4

Год	1969	1970	1971	1972	1973	1974	1975	1976	1977	1978
x	722,5	751,6	779,2	810,3	865,3	858,4	875,8	906,8	942,9	988,8
y	125,9	129,4	130,0	132,4	129,4	128,1	132,3	139,7	145,2	146,1

Год	1979	1980	1981	1982	1983	среднее
x	1015,5	1021,6	1049,3	1058,3	1095,4	780,032
y	149,3	153,2	153,0	154,6	161,2	128,084

Задача 3

- Проведено исследование, направленное на выявление взаимосвязи когнитивных и ценностно-мотивационных характеристик и показателя успешности учебной деятельности студентов-экономистов по изучению компьютерных технологий. Использовались следующие психологические показатели, измеренные в баллах по шкале от 1 до 7. Показатель успешности учебной деятельности рассчитывался по специальной методике в шкале 20-80 (данные в отдельной таблице).
- Требуется построить для успешности рассматриваемой деятельности оптимальную линейную регрессионную зависимость от психологических показателей.
- Предсказать, насколько будет успешен Иван Иванович Иванов (испытуемый 19).

Задача 4 (не обязательно, возможно для зачета)

- Проведены измерения черт характера и адаптивных способностей у солдат срочной службы - новобранцев в космических войсках (данные в отдельной таблице).
- Определить, какие черты характера соответствуют высоким адаптивным способностям (8 баллов), а какие – низким (6 баллов).

Выводы

Таким образом, в результате использования регрессионного анализа в пакете Microsoft Excel мы:

- построили уравнение регрессии;
- установили форму зависимости и направление связи между переменными - положительная линейная регрессия, которая выражается в равномерном росте функции;
- установили направление связи между переменными;
- оценили качество полученной регрессионной прямой;
- смогли увидеть отклонения расчетных данных от данных исходного набора;
- предсказали будущие значения зависимой переменной.

Если *функция регрессии* определена, интерпретирована и обоснована, и оценка точности регрессионного анализа соответствует требованиям, можно считать, что построенная модель и прогнозные значения обладают достаточной надежностью.

Прогнозные значения, полученные таким способом, являются средними значениями, которые можно ожидать.

Задание к следующему занятию

1. Кластерный анализ: понятие и назначение процедуры.
2. Виды кластерного анализа.