Анализ и представление данных психологического исследования

Лекция 7. Регрессионный анализ

Задача регрессионного анализа

- Задача регрессионного анализ (РА) состоит в построении модели, позволяющей по значениям независимых переменных получать оценки значений зависимой переменной.

- Используется также для выявления связи переменных.

Типы переменных регрессионной модели

- **Зависимая** (результирующая) – в модели играет роль функции, значение которой определяется значениями объясняющих переменных.

- **Независимые** (объясняющие) – в модели играют роль аргументов, определяют значения результирующей переменной. Их называют предикторами, или факторными признаками.

Типы переменных регрессионной модели

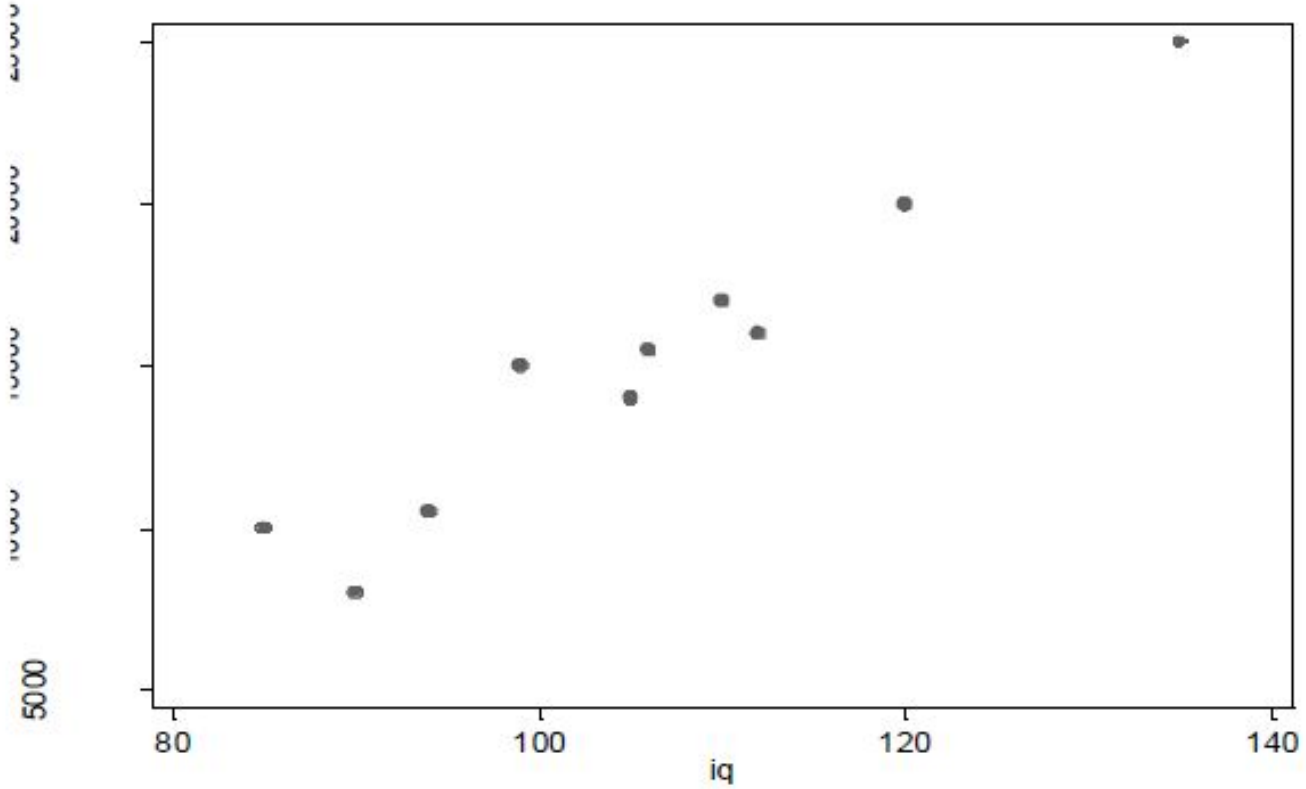
- **Зависимая** переменная:
  - непрерывная
- **Независимые** переменные
  - непрерывные,
  - дискретные,
  - категориальные

Интеллект и зарплата (простой пример)

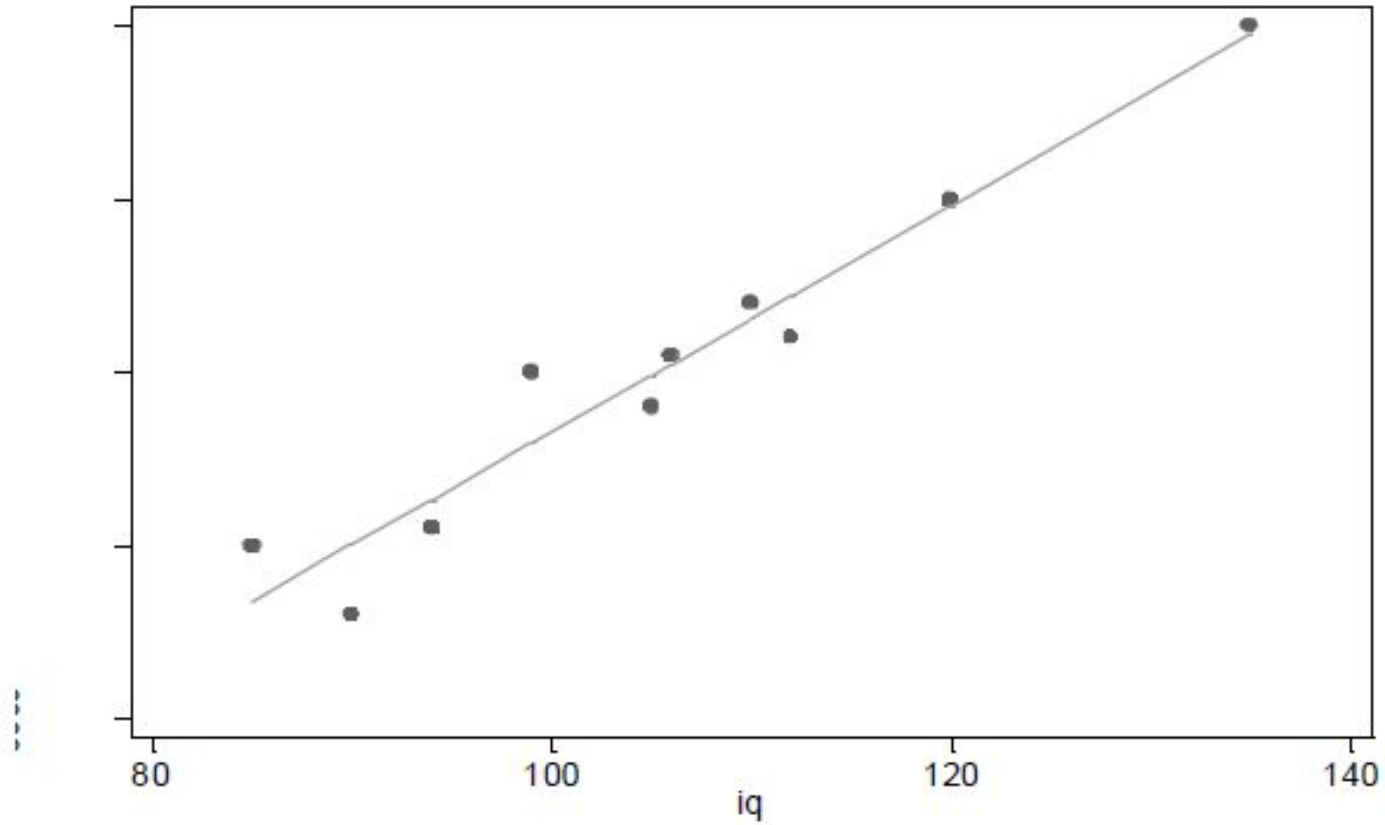
	IQ	зарплата
Гриша	85	10000
Люба	90	8000
Маша	94	10500
Марина	99	15000
Антон	105	14000
Вера	106	15500
Сергей	110	17000
Петя	112	16000
Андрей	120	20000
Оля	135	25000

- Согласно этим данным, люди с более высоким интеллектом больше зарабатывают. (Все данные я выдумал; мы знаем, что в жизни все сложнее).
- Мы хотим узнать точнее, как связаны зарплата и IQ.
- Зарплата – зависимая переменная. IQ – независимая переменная (фактор, регрессор).
- Построим диаграмму рассеивания для двух переменных.

Диаграмма рассеивания



Регрессионная линия



Формула для прямой линии

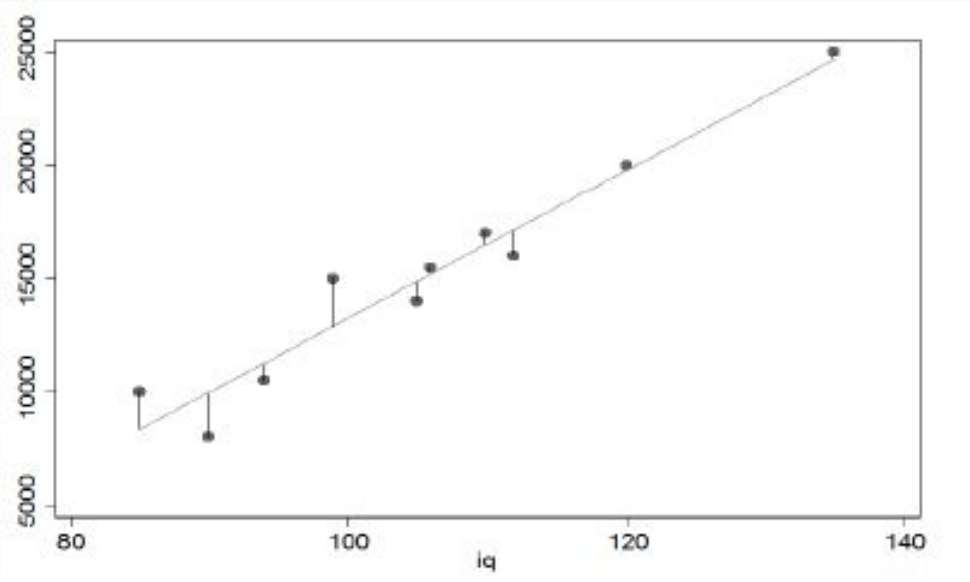
- Прямая линия описывается линейной функцией $y = a + bx$,
  - Где a – это точка, в которой прямая пересекает ось y, b – тангенс угла наклона прямой.
- Если $b>0$, то прямая идет «вверх», т.е. с увеличением x увеличивается y. Связь положительна.
- Если $b<0$, то прямая идет «вниз», т.е. с увеличением x уменьшается y. Связь отрицательна.
- Если $b=0$, то прямая идет параллельно оси x ($y=a$). Изменения x не влияют на значение y. Связь отсутствует.

Как интерпретировать коэффициент *b*?

- Знак *b* говорит о направлении связи.
- Значение *b* показывает, насколько измениться *y* если x изменить на единицу.
- Например, $y = 2 + 5x$.
  Если x1 = 5, то y1 = 27. Если x2 = 6, то н2 = 32. y2 – y1 = 32 – 27 = 5 = b.
- В данном случае $b>0$, поэтому *y* растет с увеличением *x*.

Проведение регрессионной линии

- Задача: провести через множество точек линию, которая наилучшим образом описывала бы это множество.
- Наилучшей линией будет такая, для которой расстояние от линии до точек будет минимальным.



Метод наименьших квадратов (МНК)

Сумма расстояний будет близка к нулю, т.к. часть из них являются положительными, часть – отрицательными величинами. Поэтому:

- Можно сложить модули расстояний.
- Можно сложить квадраты расстояний. По ряду статистических причин в качестве наилучшего метода для построения регрессионной линии используется метод наименьших квадратов.

Вернемся к примеры об IQ и доходе

- Зарплата = -19403 + 327 * IQ
- a = -19403, b = 327.
- Соответственно, согласно этой модели, с увеличением IQ на один пункт зарплата увеличивается на 327 руб. в месяц.

Ожидаемые значения

- Регрессионная формула позволяет определить ожидаемые (предсказанные) значения *y* для определённого уровня x.
- В примере для IQ = 90 ожидаемое значение зарплаты будет = -19403 + 327*90 = 10027 руб.
- Ожидаемые значения y отличаются от реальных значений в базе (у Любы IQ = 90, а зарплата = 8000).
- Разница между ожидаемыми и реальными значениями называется остатками (residuals), или ошибками. В случае Любы остаток равен 10027 – 8000 = 2027. Люба получает меньше, чем предсказывает регрессионное уравнение.

Источники ошибок

- Ошибки (остатки) являются неотъемлемой частью регрессионных уравнений. Случаи, когда одна переменная идеально предсказывает другую (все точки находятся на регрессионной прямой), являются исключением (и не интересны).
- Ошибки состоят из двух компонентов:
  - Ошибки измерения.
  - Вероятностный компонент, неизменно присущий отношению между двумя переменными.

Регрессия – вероятностная модель

- Формула регрессионной функции:

  $E(y) = a + \beta x$, где $E(y)$ – ожидаемое значение (среднее) y на определенном уровне x.

- Иначе эту же формулу можно записать так:

  $y = a + \beta x + \varepsilon$, где ε – ошибка.

Значимость и сила связи

- Значимость и сила статистической связи – два разных понятия.
  - Значимость: действительно ли *x* и *y* связаны?
  - Сила: как сильно связаны *x* и *y*?
- Размер коэффициента *b* говорит о силе связи. Однако его интерпретация сильно зависит от единиц измерения *x*. Например, если в нашем примере IQ/10, то коэффициент *b* уменьшится в 10 раз.
- Следовательно, коэффициенты при разных переменных НЕЛЬЗЯ непосредственно сравнивать (но можно сравнивать, если переменные измерены по одной и той же шкале).

Коэффициент детерминации *R*-квадрат

- *R*-квадрат является квадратом коэффициента корреляции
- *R*-квадрат принимает значения от 0 до 1.
  1 указывает на идеальную связь, 0 – на отсутствие связи. Чем больше значение, тем сильнее связь.
- *R*-квадрат можно интерпретировать как долю дисперсии зависимой переменной, которую «объясняет» независимая переменная.
- *R*-квадрат имеет смысл, только когда речь идет о линейной связи.

Статистическая связь и причинность

- Наличие статистической связи не означает наличия причинной связи между переменными и не позволяет утверждать, что зависимая переменная влияет на независимую.

Условия использования РА

- Все переменные должны быть случайными, подчиняться нормальному распределению.
- Линейная регрессия используется тогда(и только тогда!), когда зависимая переменная является интервальной(метрической).
- Зависимость между переменными является линейной.
- Математическое ожидание остатков должно приближаться к нулю, т.е. они должны быть нормально распределены.
- Отсутствие связи между зависимыми переменными (*отсутствие мультиколлинеарности*).

Методы устранения или уменьшения мультиколлинеарности

- Исключение одного из двух сильно связанных факторов.
- Переход от первоначальных факторов к их главным компонентам.
- Использование стратегии шагового отбора факторов.

Построение модели

- Качество результатов регрессионного анализа определяется качеством теоретического обоснования спецификации модели.
- Как выбирать переменные для включения в модель?
  - *Теоретическая логика*
  - *Эксплораторная логика*
- Следует избегать стратегии «мусорной корзины».

Этапы построения модели регрессии

1. Проверка распределения всех переменных на нормальность
2. Проверка объясняющих переменных на наличие мультиколлинеарности
3. Построение линейного уравнения регрессии
4. Оценка качества модели
5. Построение прогноза по модели регрессии

1. Проверка распределения всех переменных на нормальность

- Критерий Колмогорова-Смирнова
- Переменные, не являющиеся нормально распределенными, не могут использоваться в модели

2. Проверка объясняющих переменных на наличие мультиколлинеарности

- Анализ матрицы коэффициентов парной корреляции

  Если коэффициент парной корреляции между двумя переменными больше 0.8, то явление мультиколлинеарности можно считать установленным

Методы устранения или уменьшения мультиколлинеарности

- Исключение одного из двух сильно связанных факторов.
- Переход от первоначальных факторов к их главным компонентам.
- Использование стратегии шагового отбора факторов.

3. Построение линейного уравнения регрессии

Analyze → Regression → Linear…

4. Оценка качества модели

- **Коэффициент детерминации R^2** *(Доля вариации результативного признака под воздействием изучаемых факторов)*
- **Коэффициент множественной корреляции *R*** *(теснота связи зависимой переменной со всеми включенными в модель объясняющими факторами)*
- ***F*-критерий Фишера** *(Проверка значимости уравнения регрессии)*

Model Summary[d]				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,416[a]	,173	,154	9,335
2	,508[b]	,258	,224	8,941
3	,571[c]	,326	,278	8,622

a. Predictors: (Constant), симпатия

b. Predictors: (Constant), симпатия, польза

c. Predictors: (Constant), симпатия, польза, агрессия

d. Dependent Variable: помощь

ANOVA[d]						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	800,289	1	800,289	9,183	,004[a]
	Residual	3834,515	44	87,148		
	Total	4634,804	45			
2	Regression	1197,635	2	598,817	7,491	,002[b]
	Residual	3437,170	43	79,934		
	Total	4634,804	45			
3	Regression	1512,813	3	504,271	6,784	,001[c]
	Residual	3121,991	42	74,333		
	Total	4634,804	45			

a. Predictors: (Constant), симпатия

b. Predictors: (Constant), симпатия, польза

c. Predictors: (Constant), симпатия, польза, агрессия

d. Dependent Variable: помощь

4. Оценка качества модели-2

- Проверка распределение остатков

  Критерий Колмогорова-Смирнова

5. Построение прогноза по модели регрессии

- Необходимо построить регрессионное уравнение

Coefficients[a]						
Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	14,739	5,200		2,834	,007
	симпатия	1,547	,510	,416	3,030	,004
2	(Constant)	2,886	7,284		,396	,694
	симпатия	1,230	,509	,331	2,418	,020
	польза	1,387	,622	,305	2,230	,031
3	(Constant)	-5,315	8,075		-,658	,514
	симпатия	1,033	,500	,278	2,065	,045
	польза	1,257	,603	,276	2,083	,043
	агрессия	1,168	,567	,269	2,059	,046

a. Dependent Variable: помощь

Помощь = 1,033*симпатия+1,257*польза+1,168*агрессия