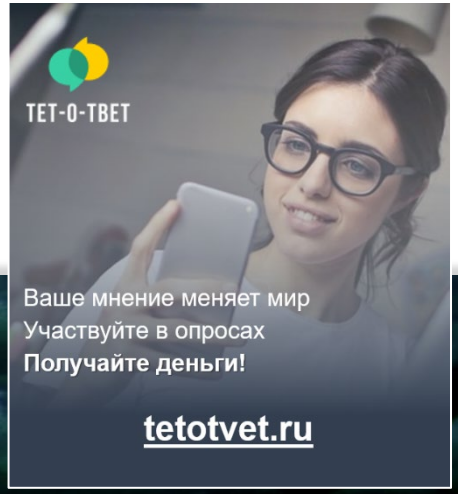
НАФИ



ТЕМА 7

РЕГРЕССИОННЫЙ АНАЛИЗ

Для свободного использования в образовательных целях
Copyright © Академия НАФИ. Москва
Все права защищены
nafi.ru

ОГЛАВЛЕНИЕ

1. **Применение регрессионного анализа**

2. **Регрессионный анализ: основные положения**

3. **Построение регрессионных моделей в SPSS**
   - 3.1 Парная регрессия
   - 3.2 Множественная регрессия
   - 3.3 Другие виды регрессий

1. ПРИМЕНЕНИЕ РЕГРЕССИОННОГО АНАЛИЗА

Цели регрессионного анализа

1. Предсказание значения зависимой переменной с помощью независимых переменных.

2. Определение вклада отдельных независимых переменных в вариацию зависимой переменной.

3. Регрессионный анализ нельзя использовать для определения наличия связи между переменными, поскольку наличие такой связи и есть предпосылка для применения этого вида анализа.

ПРИМЕР: Регрессионный анализ помогает оценить, какие факторы сильнее связаны с лояльностью, удовлетворенностью, готовностью к покупке или интересом к продукту. В клиентских проектах это позволяет перейти от описания различий к приоритизации управленческих решений: на какие факторы стоит воздействовать в первую очередь.

Допущения (assumptions) регрессионного анализа

1. Переменные модели должны иметь распределение, близкое к **нормальному**.
2. Зависимая и независимые переменные должны быть **метрическими*.**
3. Для построения линейных регрессий, зависима и независимые переменные должны иметь **линейную** связь.

** Для линейной регрессии зависимая переменная обычно количественная; независимые могут быть количественными или категориальными, если они закодированы через фиктивные переменные.*

Допущения (assumptions) регрессионного анализа

4. **Отсутствие мультиколлинеарности** – независимость между собой переменных-предикторов, отсутствие высокой корреляции (для множественной регрессии). Решение: удаление высоко коррелируемых переменных из анализа или центрирование данных (вычитание средних значений из каждого наблюдения по необходимым переменным).

5. **Отсутствие автокорреляции** – отсутствие независимости остатков. Выявляется с помощью теста Дурбина-Уотсона (обнаруживает автокорреляцию первого порядка).
   - Если d=0 – полная положительная автокорреляция
   - Если d=4 – полная отрицательная автокорреляция
   - Если d=2 – отсутствие автокорреляции

6. **Гомоскедастичность** - дисперсия остатков одинакова для каждого значения. Определяется с помощью диаграммы рассеяния.

2. РЕГРЕССИОННЫЙ АНАЛИЗ: ОСНОВНЫЕ ПОЛОЖЕНИЯ

2. Регрессионный анализ: основные положения

Регрессионный анализ – это инструмент для количественного определения значения одной переменной на основании другой.

Парная (простая) линейная регрессия даёт нам правила, определяющие линию регрессии, которая лучше других предсказывает наиболее вероятные значения одной переменной на основании другой (переменных всего две).

Множественная регрессия является расширением простой линейной регрессии.

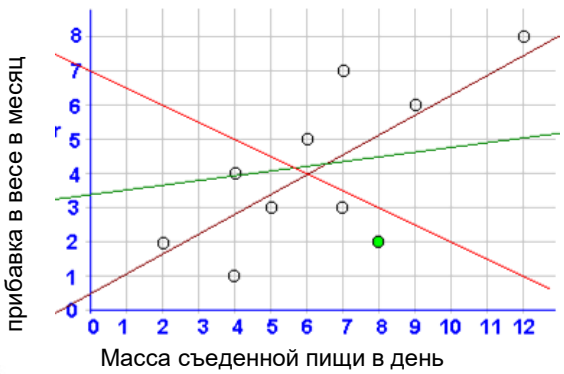
По оси Y располагают переменную, которую необходимо предсказать (зависимую), а по оси X – переменную, на основе которой будет осуществляться предсказание (независимую).

Зависимая переменная – это переменная в регрессии, которую нельзя изменять, её изменение является следствием влияния независимой переменной (переменных).
Независимая переменная – это та переменная в регрессии, которую можно изменять.
Коэффициенты регрессии (β) — это коэффициенты, которые рассчитываются в результате выполнения регрессионного анализа. Вычисляются величины для каждой независимой переменной, которые представляют силу и тип взаимосвязи независимой переменной по отношению к зависимой.

2. Регрессионный анализ: основные положения

Пример: Чем больше еды съедает каждый день детёныш бегемота (x), тем больше у него будет прибавка в весе за месяц (y).

Определяем прямую, которая наилучшим образом будет предсказывать значения Y на основании значений X.



Парная (простая) линейная регрессия (Linear Regression)

$$Y_i = a + bX_i$$

- Y_i – зависимая переменная
- X_i – независимая переменная
- a – константа, определяет **точку пересечения** прямой с осью Y. Экономически не интерпретируется.
- b – угловой коэффициент, характеризует **наклон** прямой (slope). Коэффициент регрессии b показывает, на какую величину в среднем изменится результативный признак Y_i, если переменная X_i увеличится на единицу своего измерения.

- **Коэффициент эластичности (Э)** показывает, на сколько процентов в среднем изменится Y_i при изменении X_i на 1%. Для простой линейной регрессии: $Э = b\frac{\bar{x}}{\bar{y}}$

2. Регрессионный анализ: основные положения

Задача регрессионного анализа сводится к поиску коэффициентов a и b.

$$b = r\frac{s_X}{s_Y}$$

коэффициент корреляции Пирсона

стандартные отклонения для X и Y

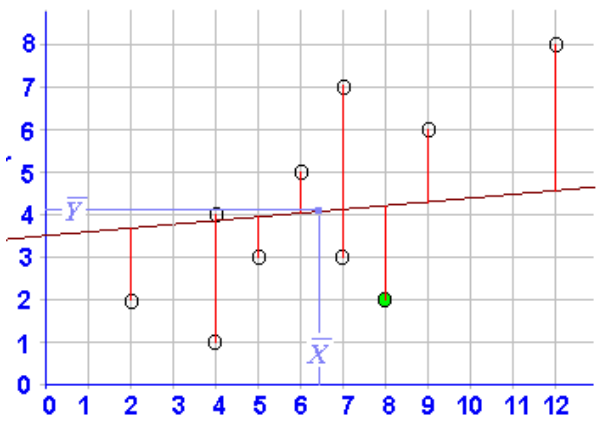
$$\overline{Y} = a + b\overline{X} \longrightarrow a = \overline{Y} - b\overline{X}$$

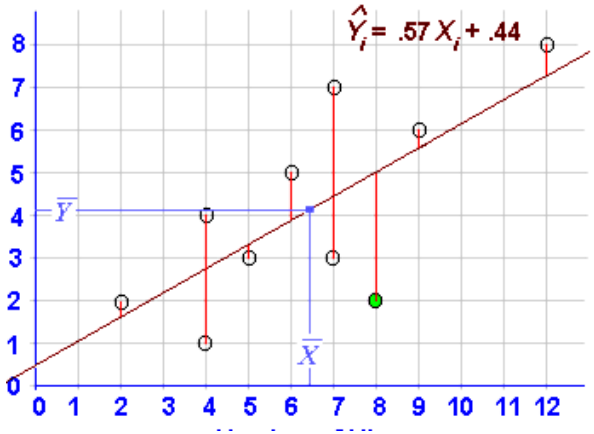
Как определить наилучшую линию регрессии?

Используют **метод наименьших квадратов** – подбирают такую линию регрессии чтобы общая сумма квадратов отклонений (Residuals) значений зависимой переменной была наименьшей.

$$\sum e_i = 0$$

$$\sum e_i^2 \text{ - минимальна}$$





$$\sum e_i^2$$ - residual sum of squares = residual SS

Суть метода наименьших квадратов

- Пусть имеются n наблюдений признаков x и y. Причем известен вид уравнения регрессии - f(x), например, прямолинейная зависимость: $f(x_i)=a + b\cdot x_i$

- Необходимо подобрать такие значения параметров (a и b), которые смогут минимизировать сумму квадратов отклонений фактических значений признака-результата y_i от расчетных (теоретических) значений $f(x_i)$ для всех наблюдений i=1:n

$$S = \sum_{i=1}^{n} (y_i - (a + b \cdot x_i))^2 \Rightarrow \min_{a,b}$$

Оценка качества уравнения регрессии и коэффициент детерминации

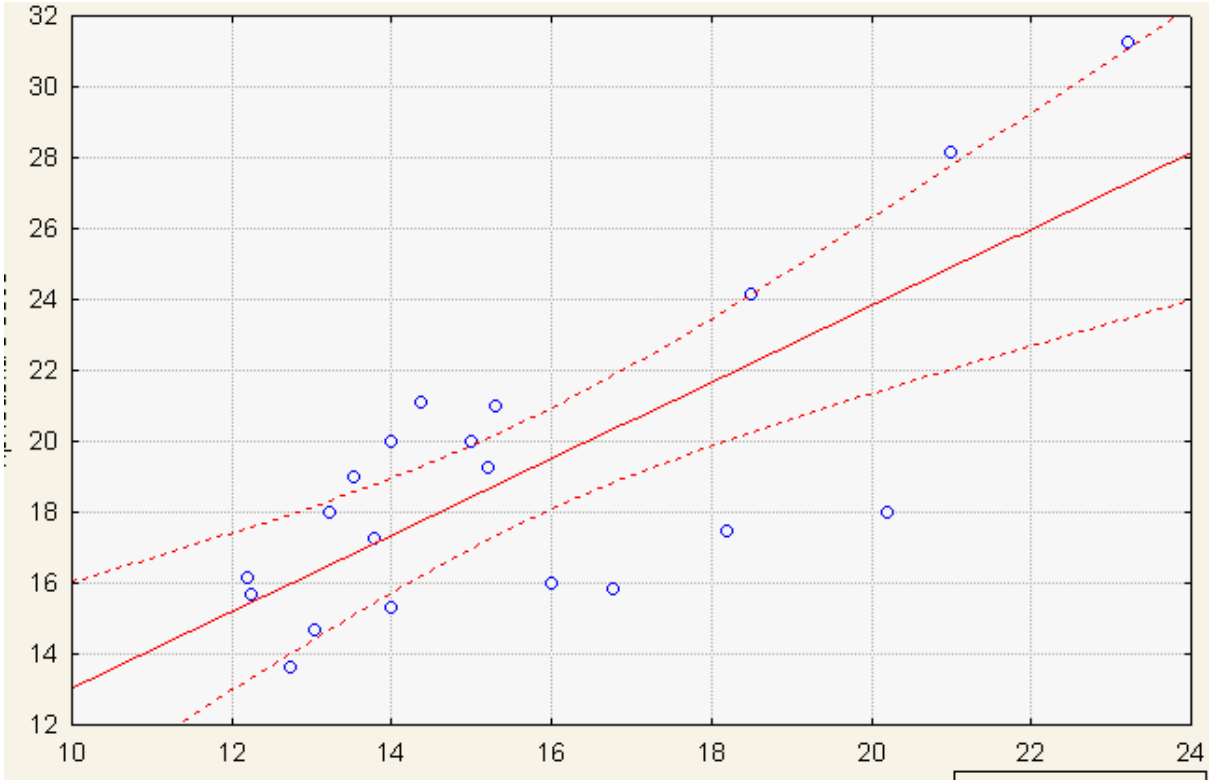
$$R^2 = \frac{SS_{regression}}{SS_{total}} = \frac{\sum_i (Y_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2}$$

Коэффициент множественной детерминации R-квадрат показывает, какую долю изменчивости (можно выразить в процентах) зависимой переменной (Y) объясняет независимая переменная (регрессионная модель).

- Под **качеством уравнения регрессии** понимается степень близости (соответствия) рассчитанных по данному уравнению значений признака-результата f(x) фактическим (наблюдаемым) значениям y.
- Чем ближе R-квадрат к 1,тем выше качество регрессионной модели.

2. Регрессионный анализ: основные положения

Обязателен расчет **доверительного интервала** для значений зависимой переменной: строится для каждого значения X, причём наименьшая ошибка получается для среднего Y.

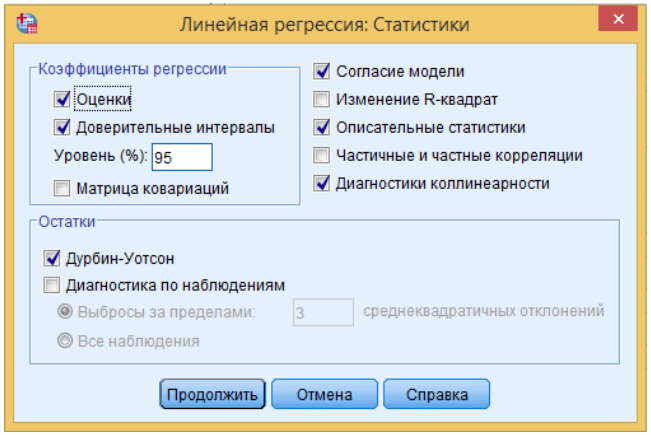


2. Регрессионный анализ: основные положения

«Линейная регрессия»

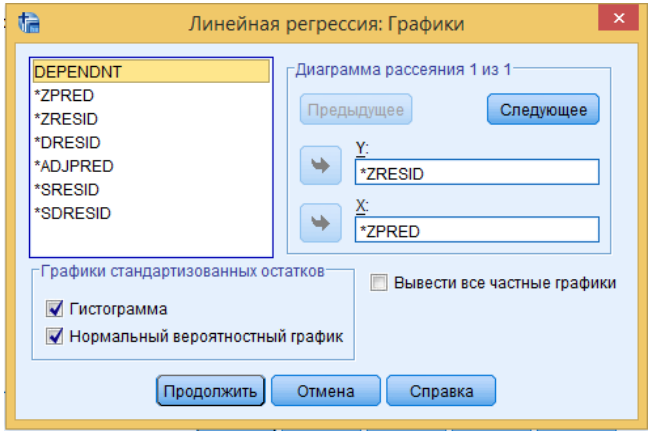
команда «Статистика» в SPSS

- **«Оценки коэффициента регрессии»** – оценки значимости коэффициентов регрессионной модели.
- **«Доверительные интервалы»** – доверительные интервалы для коэффициентов регрессионной модели.
- **«Согласие модели»** – параметры соответствия модели эмпирическим данным (коэффициенты множественной корреляции, множественной детерминации и др.).
- **«Описательные статистики»** – описательная статистика по эмпирическим данным (среднее арифметическое, стандартное отклонение и объем выборки).
- **«Диагностики коллинеарности»** – параметры для оценки мультиколлинеарности (связанность независимых переменных).
- **«Дурбин Уотсон»** – проверка на автокорреляцию остатков.



команда «Графики» в SPSS

- Проверка на **гомоскедастичность**: вставить ***ZRESID** в поле Y, а ***ZPRED** в поле X. Отметить «Гистограмма» и «Нормальный вероятностный график».



ВАЖНЫЕ ЗАМЕЧАНИЯ

- Любая регрессионная модель позволяет обнаружить только **количественные зависимости**, которые не обязательно отражают причинные зависимости, т.е. влияние одного фактора на другой.
- Гипотезы о причинной связи признаков должны дополнительно **обосновываться с помощью теоретического анализа**, содержательно объясняющего изучаемое явление или процесс.

3. ПОСТРОЕНИЕ РЕГРЕССИОННЫХ МОДЕЛЕЙ В SPSS

3.1 ПАРНАЯ РЕГРЕССИЯ

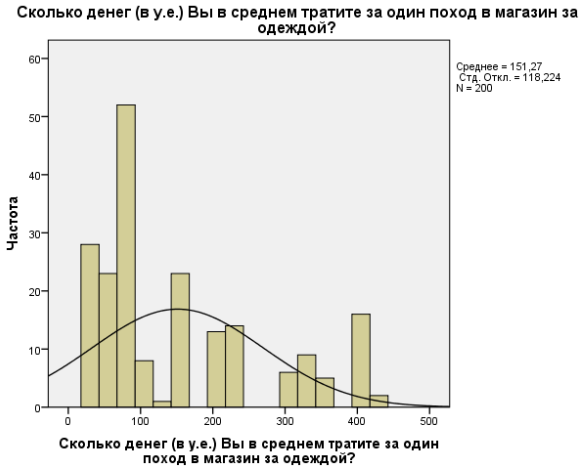
Проверка допущений (assumptions)

Задача: Узнать, как изменится стоимость покупки, если интерес к моде увеличится вдвое. (Массив данных –fashion.sav)

1. Нормальность распределения

Статистика

		Сколько денег (в у.е.) Вы в среднем тратите за один поход в магазин за одеждой?	Насколько для Вас в целом важна мода в одежде, обуви, аксессуарах?
N	Допустимо	200	200
	Пропущенные	0	0
Асимметрия		,996	-,214
Стандартная Ошибка асимметрии		,172	,172
Эксцесс		-,303	-,901
Стандартная ошибка эксцесса		,342	,342

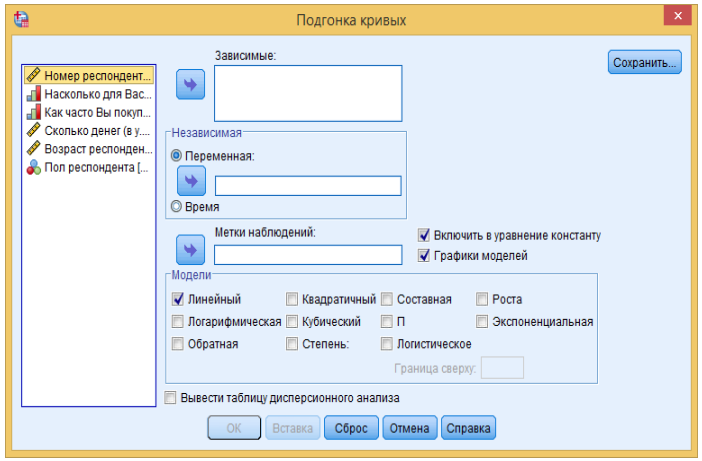


Исходя из графиков и значений асимметрии и эксцесса (в пределах от -1 до 1), можно говорить о нормальности распределения переменных.

Проверка допущений (assumptions)
Оценка вида парной регрессионной зависимости

2. Проверка на определенные виды модели

- Инструмент **«Подгонка кривых»** (меню **«Анализ»** → **«Регрессия»**).
- Вводится зависимая переменная (Dependent(s)) и одна из независимых переменных (Variable)
- Отмечаются виды регрессионных моделей, которые должны оцениваться на соответствие эмпирическим данным и статистическую значимость: линейная (Linear), квадратичная (Quadratic), кубическая (Cubic), логарифмическая (Logarithmic), экспоненциальная (Exponential), S-кривая (S), обратная (Inverse), логистическая (Logistic), показательная (Compound) и др.



Linear. Model whose equation is Y = b0 + (b1 * t). The series values are modeled as a linear function of time.

Logarithmic. Model whose equation is Y = b0 + (b1 * ln(t)).

Inverse. Model whose equation is Y = b0 + (b1 / t).

Quadratic. Model whose equation is Y = b0 + (b1 * t) + (b2 * t**2). The quadratic model can be used to model a series that "takes off" or a series that dampens.

Cubic. Model that is defined by the equation Y = b0 + (b1 * t) + (b2 * t**2) + (b3 * t**3).

Power. Model whose equation is Y = b0 * (t**b1) or ln(Y) = ln(b0) + (b1 * ln(t)).

Compound. Model whose equation is Y = b0 * (b1**t) or ln(Y) = ln(b0) + (ln(b1) * t).

S-curve. Model whose equation is Y = e**(b0 + (b1/t)) or ln(Y) = b0 + (b1/t).

Logistic. Model whose equation is Y = 1 / (1/u + (b0 * (b1**t))) or ln(1/y-1/u) = ln (b0) + (ln(b1) * t) where u is the upper boundary value. After selecting Logistic, specify the upper boundary value to use in the regression equation. The value must be a positive number that is greater than the largest dependent variable value.

Growth. Model whose equation is Y = e**(b0 + (b1 * t)) or ln(Y) = b0 + (b1 * t).

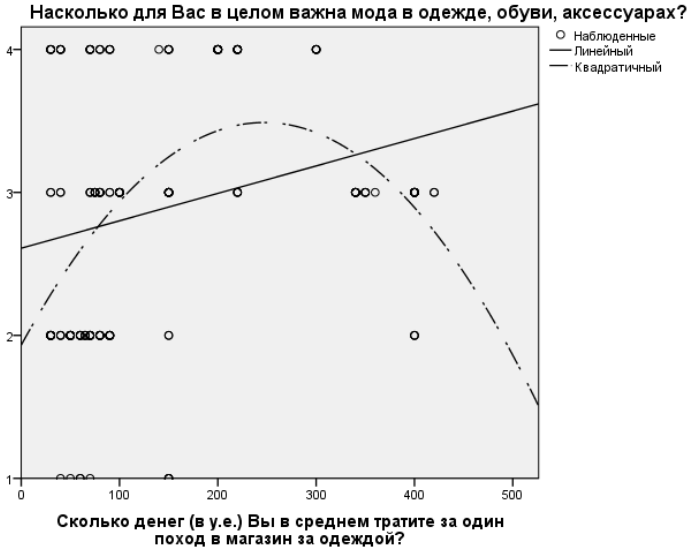
Exponential. Model whose equation is Y = b0 * (e**(b1 * t)) or ln(Y) = ln(b0) + (b1 * t).

Проверка допущений (assumptions)

3. Линейная или квадратичная модель?

Исходя из графика и значений R-квадрата, видно, что лучше было бы использовать квадратичную регрессию, однако можно построить и линейную.

Ниже будет расписано, как делать нелинейные регрессии.



Сводка для модели и оценки параметров

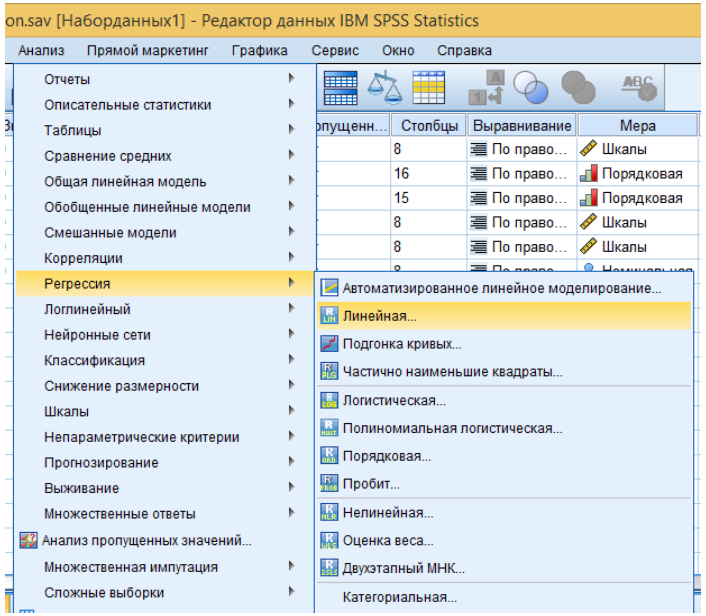
Зависимая переменная: Насколько для Вас в целом важна мода в одежде, обуви, аксессуарах?

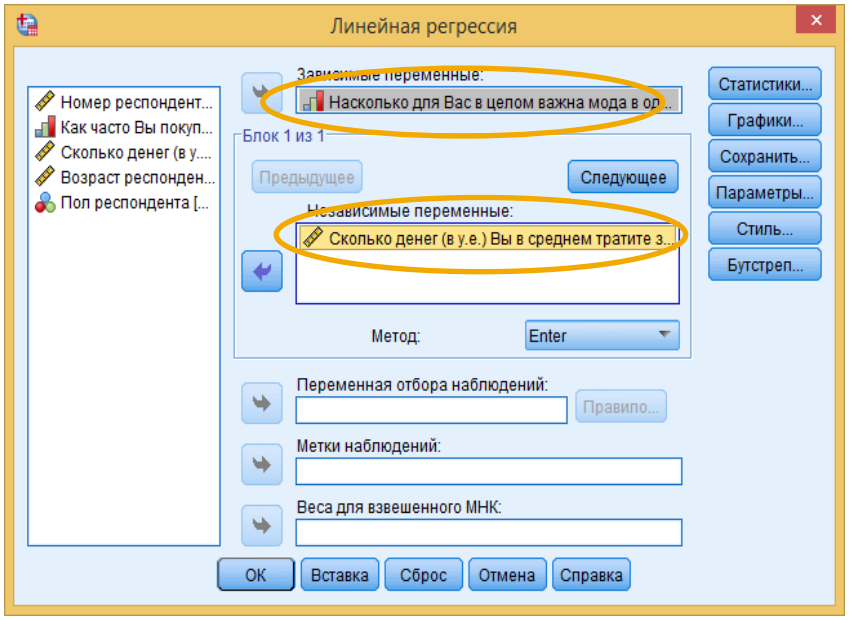
Уравнение	Сводка для модели					Оценки параметров		
	R-квадрат	F	ст.св.1	ст.св.2	Знач.	Константа	b1	b2
Линейная	,067	14,295	1	198	,000	2,610	,002	
Квадратичная	,182	21,984	2	197	,000	1,932	,013	-2,547E-5

Независимая переменная - это Сколько денег (в у.е.) Вы в среднем тратите за один поход в магазин за одеждой?.

Построение простой регрессионной модели

1. Открыть массив данных fashion.sav.
2. Команды **«Анализ» → «Регрессия» → «Линейная».**
3. Зависимая переменная переносится в поле **«Зависимые переменные».**
4. Независимые переменные (факторные признаки) – в поле **«Независимые переменные».**





3.1 Построение регрессионных моделей в SPSS

5. В команде **«Статистики»** выбрать:
   - «Оценки коэффициента регрессии»
   - «Доверительные интервалы»
   - «Согласие модели»
   - «Описательные статистики»
   - «Дурбин Уотсон»

6. В команде «Графики» вставить *ZRESID в оба свободных окна и отметить «Гистограмма» и «Нормальный вероятностный график».

7. Раздел «**Сводка для модели»** содержит статистику соответствия модели эмпирическим данным:

Сводка для модели[b]

Модель	R	R-квадрат	Скорректированный R-квадрат	Стандартная ошибка оценки	Дурбин-Уотсон
1	,259[a]	,067	,063	,846	1,683

коэффициент множественной корреляции R

коэффициент множественной детерминации

скорректированный R-квадрат (не брать в расчет)

стандартная ошибка оценки зависимой переменной

d в пределах [1,5;2,5] отсутствует автокорреляция

6,7% дисперсии зависимой переменной объясняется влиянием независимой переменной

3.1 Построение регрессионных моделей в SPSS

8. Раздел **«ANOVA»** показывает суммы квадратов отклонений, F-критерий Фишера, уровень значимости модели (Sig), по которому можно судить о достоверности построенной связи переменных.

ANOVA[a]

Модель		Сумма квадратов	ст.св.	Средний квадрат	F	Знач.
1	Регрессия	10,235	1	10,235	14,295	,000[b]
	Остаток	141,765	198	,716		
	Всего	152,000	199			

a. Зависимая переменная: Насколько для Вас в целом важна мода в одежде, обуви, аксессуарах?

b. Предикторы: (константа), Сколько денег (в у.е.) Вы в среднем тратите за один поход в магазин за одеждой?

Статистики, оценивающие долю дисперсии зависимой переменной, обусловленную влиянием независимых переменных

Статистики, оценивающие долю дисперсии зависимой переменной, НЕ обусловленную влиянием независимых переменных

3.1 Построение регрессионных моделей в SPSS

9. В разделе **«Коэффициенты»** приводятся значения параметров регрессионной модели и показатели их статистической значимости:
   - **B** – значения коэффициентов регрессионного уравнения (Unstandardized Coefficients B)
   - **Std. Error** – стандартная ошибка коэффициентов
   - **Standardized Coefficients Beta** – стандартные β-коэффициенты регрессионной модели (фактически – коэффициент корреляции Пирсона)
   - **t** – эмпирическое значение t-критерия для проверки статистической значимости соответствующего коэффициента
   - **Sig** – p-уровень значимости коэффициентов (вероятность ошибочного принятия гипотезы о существовании ненулевых коэффициентов регрессии)

Коэффициенты[a]

	Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	т	Знач.
			B	Стандартная Ошибка	Бета		
a	1	(Константа)	2,610	,097		26,817	,000
b		Сколько денег (в у.е.) Вы в среднем тратите за один поход в магазин за одеждой?	,002	,001	,259	3,781	,000

a. Зависимая переменная: Насколько для Вас в целом важна мода в одежде, обуви, аксессуарах?

y = a + b*x ⇨ Тест статистика покупки = 2,610 + 0,002*x удвоенный интерес к моде

3.1 Построение регрессионных моделей в SPSS

10. Последняя выведенная диаграмма говорит о гомо- или гетероскедастичности. Диаграмма показывает однородную вариативность значений наблюдений, выражающаяся в относительной стабильности, гомогенности дисперсии случайной ошибки.
Полученный график показывает скорее гетероскедастичность.

Построение нестандартных нелинейных регрессионных моделей

Меню **«Анализ»** → **«Регрессия»** → **«Нелинейная».**

- В окне **«Нелинейная Регрессия»** зависимую переменную нужно перенести в соответствующее поле.
- В поле **«Выражение»,** задающее модель вводится формула предполагаемой связи зависимой переменной и одной или нескольких независимых переменных, используя соответствующие символы и функции.
- В формулу связи кроме имен независимых переменных должны быть включены коэффициенты – параметры регрессионной модели, которые будут оцениваться с помощью итерационной процедуры.
- Задать начальные значения параметров регрессии, щелкнув на кнопке **«Параметры».** В появившемся диалоговом окне укажите в поле имен имя первого параметра, **«Начальное значение»**, затем щелкните на **«Добавить»**, и так для каждого параметра регрессионной модели.

Построение линейных моделей методом пошаговой регрессии

- Построение регрессионных моделей на основе пошаговой регрессии в SPSS практически не отличается от процедуры построения множественной линейной регрессии.

- Выбор режима пошаговой регрессии осуществляется в окне **«Линейная Регрессия»** в поле **«Метод»,** все остальные действия аналогичны рассмотренным выше.

Пошаговая регрессия (Stepwise Regression)

- **Обратная пошаговая регрессия** заключается в том, что последовательно исключаются наименее значимые факторы.
- На нулевом шаге проводится регрессионный анализ для всех факторов. Каждый фактор проверяется на значимость.
- Если статистический показатель значимости меньше критического значения, называемого величиной F-удаления (F-to remove), то фактор исключается из анализа и строится новое уравнение регрессии по оставшимся факторам (по умолчанию критический p-уровень значимости для величины F-удаления задается на уровне 0,1).

Пошаговая регрессия (Stepwise Regression)

- **Прямая пошаговая регрессия** организована в противоположном направлении: на первом шаге в уравнение регрессии включается фактор, имеющий наибольший коэффициент корреляции с y и проверяется адекватность и значимость модели.
- Если модель значима, включается следующий фактор и вычисляется F-статистика для каждой переменной модели.
- Если статистический показатель значимости какой-либо переменной меньше величины F-удаления, то фактор исключается, если больше – сохраняется, и в уравнение включается следующая переменная.
- Поскольку проверка всех выбранных переменных осуществляется на каждом шаге, может оказаться, что переменная, включенная в уравнение на предыдущем шаге, может быть исключена на следующих шагах.
- Процедура пошаговой регрессии позволяет значительно сократить объем работы при конструировании адекватной и значимой регрессионной модели.

3.2 МНОЖЕСТВЕННАЯ РЕГРЕССИЯ

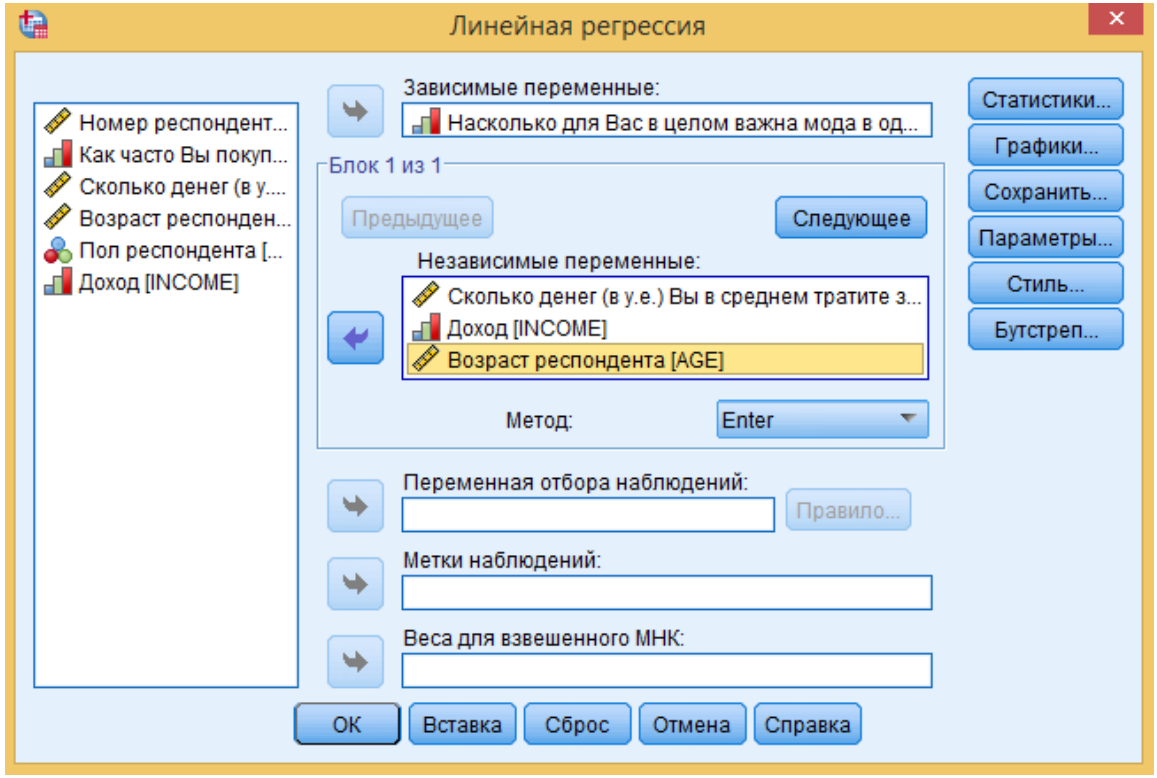
Построение множественной регрессионной модели

- **Множественная регрессия** является расширением простой линейной регрессии. С помощью простой регрессии оценивалась степень влияния одной независимой переменной (предиктора) на зависимую переменную (критерий). В отличие от простой регрессии (Y=B*X+A), множественная регрессия исследует влияние двух и более предикторов на критерий (Y=B1*X1+B2*X2+B3*X3+…+A).

- Применение множественной регрессии позволяет исследователю ответить на вопрос, **насколько хорошо оцененное уравнение аппроксимирует данные**, есть ли значимая линейная связь, а также каковы оцененные значения коэффициентов для уравнения наилучшего предсказания. Кроме того, может быть определена относительная важность независимых переменных в предсказании зависимой переменной.

3.2 Множественная регрессия

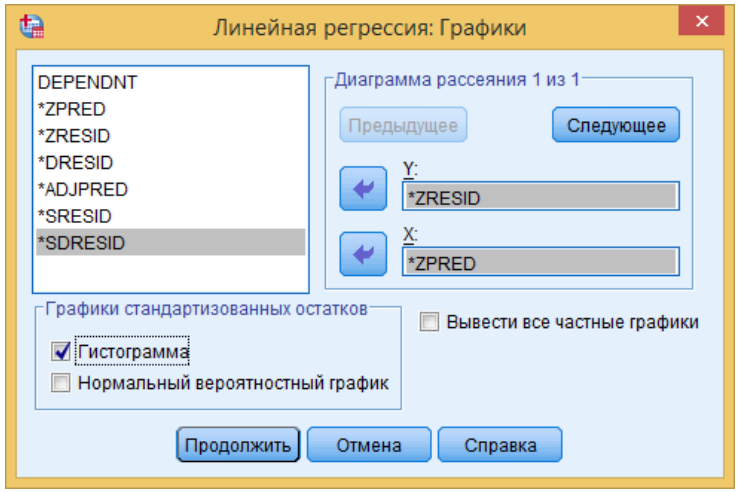
Задача: Расширим рассмотренную ранее модель предсказания важности моды для респондентов, включив в неё следующие переменные: «Возраст респондента» (AGE), «Доход респондента» (INCOME).

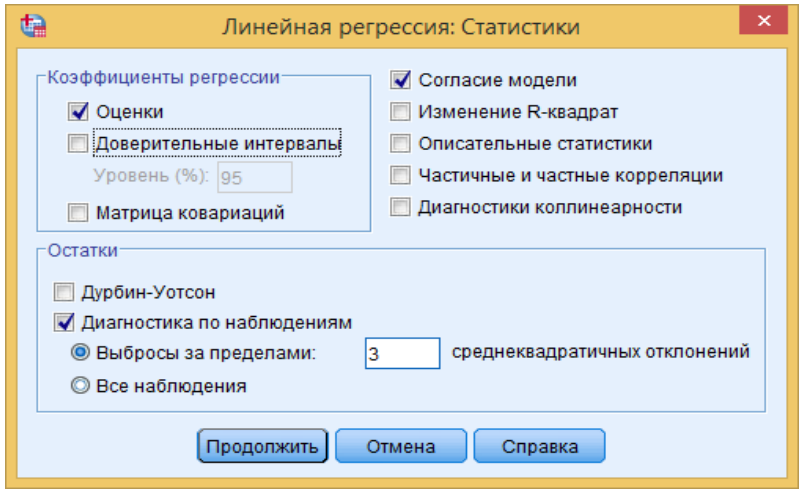


Графики остатков

При необходимости можно запросить вывод некоторых диагностических графиков, включающих остатки и информацию о выбросах. По умолчанию графики остатков не выводятся. Щелкните по кнопке «Графики». Пометьте элемент «Гистограмма» в группе «Графики стандартизированных остатков».

- Переместите *ZRESID в поле Y;
- Переместите *ZPRED в поле X.





Результаты множественной регрессии

Сводка для модели[b]

Модель	R	R-квадрат	Скорректированный R-квадрат	Стандартная ошибка оценки	Дурбин-Уотсон
1	,635[a]	,403	,394	,681	1,809

a. Предикторы: (константа), Возраст респондента, Сколько денег (в у.е.) Вы в среднем тратите за один поход в магазин за одеждой?, Доход

b. Зависимая переменная: Насколько для Вас в целом важна мода в одежде, обуви, аксессуарах?

Мера R-квадрат составляет 0,403, что говорит, о том, что с помощью предикторных переменных можно объяснить около 40% вариации частоты покупки одежды.

Значение Дурбин-Уотсон не выходит за границы [1,5;2,5], поэтому можно говорить о том, что автокорреляции нет.

ANOVA

ANOVA[a]

Модель		Сумма квадратов	ст.св.	Средний квадрат	F	Знач.
1	Регрессия	61,219	3	20,406	44,058	,000[b]
	Остаток	90,781	196	,463		
	Всего	152,000	199			

a. Зависимая переменная: Насколько для Вас в целом важна мода в одежде, обуви, аксессуарах?

b. Предикторы: (константа), Возраст респондента, Сколько денег (в у.е.) Вы в среднем тратите за один поход в магазин за одеждой?, Доход

Поскольку имеются три независимых переменных, F-критерий проверяет, имеет ли какая-либо из этих переменных линейную взаимосвязь с частотой покупки одежды.

Неудивительно, что критерий показывает уровень значимости, поскольку известно, что между затратами денег на одежду за один поход и уровнем важности моды имеется значимая взаимосвязь.

Корреляции

Корреляции

		Насколько для Вас в целом важна мода в одежде, обуви, аксессуарах?	Сколько денег (в у.е.) Вы в среднем тратите за один поход в магазин за одеждой?	Доход	Возраст респондента
Корреляция Пирсона	Насколько для Вас в целом важна мода в одежде, обуви, аксессуарах?	1,000	,259	,251	-,553
	Сколько денег (в у.е.) Вы в среднем тратите за один поход в магазин за одеждой?	,259	1,000	,944	,081
	Доход	,251	,944	1,000	,100
	Возраст респондента	-,553	,081	,100	1,000
Знач. (односторонняя)	Насколько для Вас в целом важна мода в одежде, обуви, аксессуарах?	.	,000	,000	,000
	Сколько денег (в у.е.) Вы в среднем тратите за один поход в магазин за одеждой?	,000	.	,000	,126
	Доход	,000	,000	.	,080
	Возраст респондента	,000	,126	,080	.
N	Насколько для Вас в целом важна мода в одежде, обуви, аксессуарах?	200	200	200	200
	Сколько денег (в у.е.) Вы в среднем тратите за один поход в магазин за одеждой?	200	200	200	200
	Доход	200	200	200	200
	Возраст респондента	200	200	200	200

Однако, посмотрев на таблицу с корреляциями, можно заметить высокую прямую корреляцию между переменными «Доход» и «Сколько денег (в у.е.) Вы в среднем тратите за один поход в магазин за одеждой?».

Стоило бы убрать из модели переменную «Доход» и **построить новую модель**, поскольку есть мультиколлинеарность.

B и Бета коэффициенты множественной регрессии и статистика остатков

y = a + b1*x1 + b2*x2 + b3*x3 ⇨ статистика покупки = 4,857 + 0,001*x1 + 0,154*x2 – 0,105*x3

Коэффициенты[a]

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	т	Знач.	95,0% Доверительный интервал для B		Статистика коллинеарности	
		B	Стандартная Ошибка	Бета			Нижняя граница	Верхняя граница	Допуск	VIF
1	(Константа)	4,857	,243		19,967	,000	4,377	5,337		
	Сколько денег (в у.е.) Вы в среднем тратите за один поход в магазин за одеждой?	,001	,001	,138	,823	,411	-,001	,003	,109	9,188
	Доход	,154	,144	,179	1,069	,287	-,130	,437	,108	9,219
	Возраст респондента	-,105	,010	-,582	-10,487	,000	-,125	-,086	,988	1,012

a. Зависимая переменная: Насколько для Вас в целом важна мода в одежде, обуви, аксессуарах?

Высокое значение VIF также говорит о мультиколлинеарности

3.3 ДРУГИЕ ВИДЫ РЕГРЕССИЙ

Регрессия с фиктивными переменными

- **Фиктивная переменная** – сконструированная количественная переменная, описывающая **качественные факторы** (например, пол, профессия, образование, принадлежность к какой-либо группе).
- На практике количество фиктивных переменных в модели на 1 меньше чем число градаций признака.

Пример:

Пусть Y – поквартальные наблюдения ВВП. Реальный ВВП зависит от реальных государственных расходов. В первом квартале ситуация всегда лучше (это связано с началом нового финансового года и т. п.)

$$D_1 = \begin{cases} \mathbf{1}, \text{I квартал} \\ \mathbf{0}, \text{II} - \text{IV кварталы} \end{cases} \quad D_2 = \begin{cases} \mathbf{1}, \text{II квартал} \\ \mathbf{0}, \text{остальные кварталы} \end{cases} \quad D_3 = \begin{cases} \mathbf{1}, \text{III квартал} \\ \mathbf{0}, \text{остальные кварталы} \end{cases}$$

D4 уже не нужно, т.к. четвертый квартал будет служить базовой категорией, с которой будут сравниваться все остальные кварталы.

Итоговое уравнение с константой и тремя фиктивными переменными:

$$\boxed{Y = a + bX + z_1 D_1 + z_2 D_2 + z_3 D_3}$$

Бинарная логистическая регрессия

- Зависимая переменная – **дихотомическая**.
- Цель – построение модели прогноза вероятности события {Y=1} в зависимости от независимых переменных X1,…,Xp путём подгонки данных к **логистической кривой.**
- Отношение вероятности того, что событие произойдет к вероятности того, что оно не произойдет $\frac{P}{1-P}$ называется отношением шансов.

Уравнение логистической регрессии:

$$\boxed{Z = B_0 + B_1X_1 + \ldots + B_p}$$

- В связи с этим отношение шансов может быть записано в следующем виде:

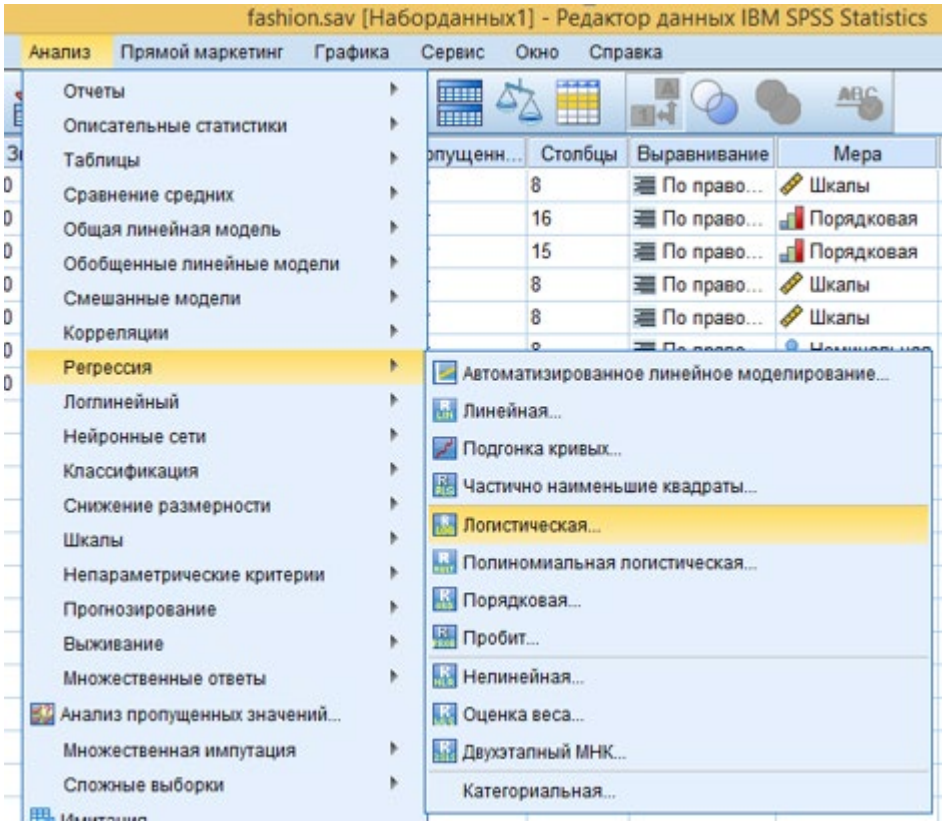
$$\frac{P}{1-P} = e^{B_0+B_1X^1+B_2X^2+\ldots+B_pX^p} = e^{B_0}e^{B_1X^1}\cdot\ldots\cdot e^{B_pX^p} = e^{B_0}\left(e^{B_1}\right)^{X^1}\cdot\ldots\cdot\left(e^{B_p}\right)^{X^p}$$

- Отсюда получается, что, если модель верна, при независимых $x^1, \ldots, x^p$ изменение x^k на единицу вызывает изменение отношения шансов в e^{b_k} раз.

Пробит-регрессия

- Зависимая переменная – **дихотомическая**.

- Метод, похожий на **логистическую регрессию**, но основанный не на моделировании логарифма отношения вероятностей интересующих категорий зависимой переменной, а на моделировании аргумента функции **нормального распределения**, через которую и рассчитывается вероятность интересующей категории зависимой переменной.



Литература по Теме 7

1. **Бююль А., Цеффель П. SPSS: искусство обработки информации. – М., 2005**
   - Глава 16. Регрессионный анализ
2. **Наследов А. IBM SPSS Statistics 20 и AMOS: профессиональный статистический анализ данных. – СПб., 2013**
   - Глава 17. Простая линейная регрессия
3. **Осипов Г.В. Рабочая книга социолога. – М., 2006**
   - Глава 5. Методы статистики в социологическом исследовании

Чек-лист аналитика

- зависимая переменная определена корректно
- предикторы выбраны на основе гипотезы, а не случайного перебора
- проверена линейность
- проверены остатки
- проверена мультиколлинеарность
- коэффициенты интерпретированы без причинных преувеличений
- вывод переведен на язык управленческого решения

НАФИ

О том, как на практике применяется регрессионный анализ, вы можете увидеть, изучив, например, страницу решения «Ценовые исследования (Price Research)» на нашем сайте.

Хотите узнать, как проводятся онлайн-опросы, и попробовать себя в роли респондента? Регистрируйтесь на нашей платформе https://tetotvet.ru или скачивайте наше приложение «Тет-о-твет» в любом из магазинов приложений.

ДОСТУПНО В Google Play

ЗАГРУЗИТЕ В App Store

ОТКРОЙТЕ В AppGallery

Для свободного использования в образовательных целях
Copyright © Академия НАФИ. Москва
Все права защищены
nafi.ru