1.1.1.1 Введение в регрессионный анализ

Регрессионный анализ остается одной из наиболее востребованных и популярных количественных методов в социальных науках. Возможность одновременного изучения неограниченного количества объектов, а также «прозрачность» техники создали ему репутацию надежного инструмента анализа. Сильная сторона метода состоит в том, что он направлен не просто на изучение изменений, но на сведение причины и следствия. Иначе говоря, регрессионный анализ отвечает на вопрос: «Влияет ли одна или несколько переменных (потенциальных причин) на другую переменную (результат) и, если да, то в какой степени?» В данном случае мы ограничимся введением в регрессионное моделирование и рассмотрим наиболее простую модель регрессии – линейную. По ходу дальнейшего изложения читатель также узнает о других видах регрессии.

Для демонстрации возможностей регрессионного анализа нам нужен эмпирический пример. Рассмотрим такую проблему, как распространение ВИЧ-инфекции в Центральной и Южной Африке. Несмотря на то, что сам вирус устойчиво ассоциируется с Африканским континентом, сравнение показателей распространенности заболевания указывает на значительную разницу между странами, располагающимися недалеко друг от друга [Levi-Faur]. Постараемся оценить различные социальные факторы, влияющие на распространение вируса.

В отличие от Европы и Северной Америки, ВИЧ в Центральной и Южной Африке передается большей частью через гетеросексуальные контакты. Соответственно, в первую очередь социологу следует обратить внимание на факторы, определяющие сексуальное поведение: риск заражения увеличивается по мере расширения круга сексуальных контактов без контрацепции. Тогда релевантными окажутся, например, такие вопросы: «Являются ли до- и внебрачные сексуальные отношения социально приемлемыми в странах?» или «Наложено ли в данном обществе табу на обсуждение ВИЧ?»

Перед исследователем встает проблема измерения этих показателей. Поскольку прямому количественному подсчету они не поддаются, приходится выбирать индикаторы, которые могут отражать искомые переменные – такие индикаторы называются **прокси-переменными**. В целях простоты изложения воспользуемся следующей переменной - религиозный состав населения. Разумеется, такая генерализация неизбежно ведет к ошибкам измерения. В то же время, очевидно, что религия определяет широкий круг форм поведения, включая и сексуальное. Мы можем ожидать, что в обществе, где

доминирующей религией является ислам, существует большое количество предписаний (в первую очередь для женщин) относительно взаимодействия с другими людьми. Мы также предполагаем, что в (условно) католических странах количество внебрачных сексуальных контактов будет меньше, чем в протестантских.

В поисках нужных показателей обратимся к базам данных. Сейчас нас интересует две переменных:

- количество носителей вич / СПИД в возрасте 15-49 лет по странам (здесь мы воспользуемся базой данных UNAIDS - http://www.unaids.org/en/dataanalysis/datatools/aidsinfo/);
- процент населения, причисляющий себя к одной из религий (католичество, протестантизм, ислам или традиционные религиозные верования - AfricaResearchProgram - http://africa.gov.harvard.edu/).

Перед нами выборка из 47 стран Центральной и Южной Африки. Для начала оценим тесноту связи между религиозной принадлежностью и количеством носителей вируса. Для этого воспользуемся парным коэффициентом корреляции. В SPSS это можно сделать следующим образом: **Анализ – Корреляции – Парные...**

Получаем следующие результаты:

Корреляции

		% Protestant	HIV
% Protestant	Корреляция Пирсона	1	,425**
	Знч.(1-сторон)		,001
	N	47	47
HIV	Корреляция Пирсона	,425**	1
	Знч.(1-сторон)	,001	
	N	47	47

**. Корреляция значима на уровне 0.01 (1-сторон.).

Таблица 1.1. Коэффициент корреляции переменных «заболеваемость ВИЧ» и «количество протестантов».

		HIV	% Adhering to traditional Religions
HIV	Корреляция Пирсона	1	-,203
	Знч.(1-сторон)		,085

% Adhering to traditional Religions	N	47	47
	Корреляция Пирсона	-,203	1
	Знч.(1-сторон)	,085	
	N	47	47

Таблица 1.2. Коэффициент корреляции переменных «заболеваемость ВИЧ» и «количество приверженцев традиционных религий".

		HIV	% Catholic
HIV	Корреляция Пирсона	1	-,163
	Знч.(1-сторон)		,137
	N	47	47
% Catholic	Корреляция Пирсона	-,163	1
	Знч.(1-сторон)	,137	
	N	47	47

Таблица 1.3. Коэффициент корреляции переменных «заболеваемость ВИЧ» и «количество католиков».

		HIV	% Muslim
HIV	Корреляция Пирсона	1	-,147
	Знч.(1-сторон)		,161
	N	47	47
% Muslim	Корреляция Пирсона	-,147	1
	Знч.(1-сторон)	,161	
	N	47	47

Таблица 1.4. Коэффициент корреляции переменных «заболеваемость ВИЧ» и «количество мусульман».

Как мы видим, самая тесная связь между уровнем заболеваемости и религиозной принадлежностью наблюдается у жителей протестантских стран. Кроме того, только здесь коэффициент корреляции принимает положительное значение (в остальных случаях наблюдается обратная связь). Значит ли это, что протестантские церкви, секты и деноминации в своей этике содержат предпосылки, «благоприятные» для распространения ВИЧ-инфекции? Не исключено. Однако не будем также забывать, что ареал распространения различных религий Старого света неразрывно связан и с колониальным прошлым африканских стран. Возможно, дело состоит не только (или не столько) в самой религии, сколько в институциональной структуре, которую привезли с собой ее изначальные носители: экономическая система британского колониализма основывалась на труде мигрантов, когда мужчины в течение длительного времени жили и работали вдали от семей [Hunt 1996; 1284].

Так или иначе, определенные выше коэффициенты корреляции фактически фиксируют то, как *тесно* взаимосвязаны две переменные между собой, а не то, насколько *сильно* они

взаимосвязаны. Разницу между теснотой и силой связи можно пояснить на следующем примере. Мы знаем, что если солить еду, то она становится солонее - другими словами, эти характеристики тесно взаимосвязаны (при наличии одной будет иметь место другая). Но на практике крайне важно знать, *насколько* блюдо становится солонее при добавлении определенного количества соли – это и есть сила связи. Без этого знания нельзя составлять рецепты, то есть формировать обобщения [Крыштановский 2007].

Чтобы модель давала нам полезную информацию, которую можно использовать при нахождении причинно-следственных связей, общих для сравниваемых случаев, необходимо иметь представление о силе соответствующих связей, то есть понимать, какие из показателей влияют на результат сильнее, а какие слабее, а также насколько велико результирующее влияние всех факторов.

С этой задачей и призваны справляться *регрессионные модели*. Общая цель работы с регрессиями – определить, существует ли статистически значимые отношения между зависимой и независимыми переменными, и как они себя проявляют. Регрессия, с одной стороны, дает возможность исследователю «ухватить» общую закономерность, а с другой – оставляет пространство для возможных исключений из правила (случаев, которые в закономерность не вписываются).

Рассмотрим диаграмму рассеивания по двум переменным: количеству протестантов в стране и числу носителей вируса иммунодефицита.

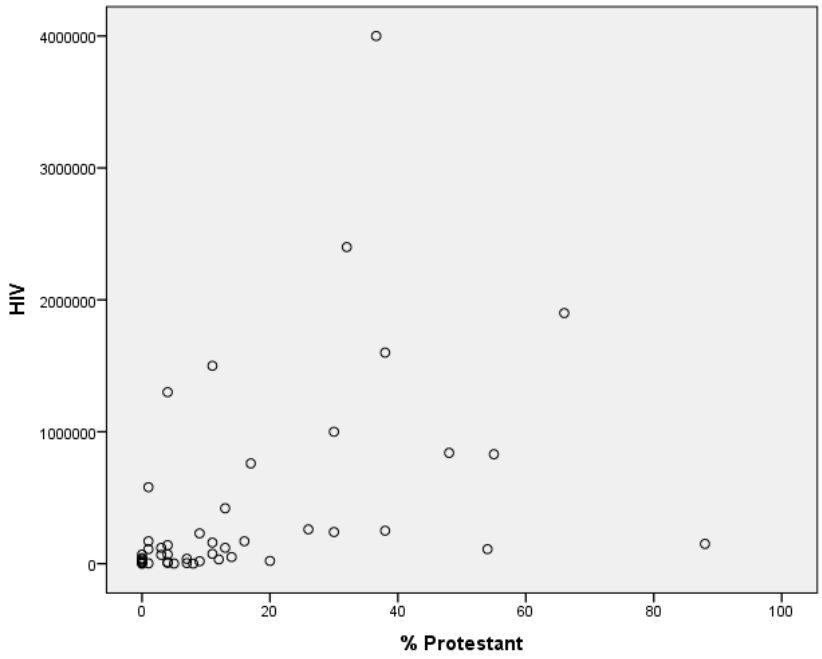


Рисунок 1.1. Диаграмма рассеивания переменных "Количество протестантов" и "Заболеваемость ВИЧ".

Определенная зависимость здесь просматривается: чем меньше сторонников протестантизма, тем меньше заболеваемость, или, наоборот, с ростом значений независимой переменной наблюдается тенденция возрастания значений зависимой

переменной. Это подтверждается значением коэффициента корреляции Пирсона 0,425 при уровне значимости a = 0,01. Формально предложенную модель зависимости можно записать в виде следующей математической формулы:

$$y = f(x) + e,$$

где *y* — показатель «уровень заболеваемости» (зависимая переменная); *x* — показатель «процент протестантов» (независимая переменная); *f(x)*— функция, описывающая силу и форму влияния *x* на *y; e*— все остальные факторы, влияющие на *y*.Задача построения модели превращается в подбор функции, которая наилучшим образом опишет зависимость *x* и *y.* Рассмотрим, как это можно сделать.

Наиболее удобным и простым вариантом выступает приложение линейной модели зависимости. График линейной функции представляет собой прямую линию и описывается формулой:

$$y = kx + b,$$

где *y* – наблюдаемые значения зависимой переменной, *x* – наблюдаемые значения независимой переменной, *b* – точка, в которой наш график пересекает ось ординат (в нашем примере, количество ВИЧ-инфицированных при полном отсутствии протестантов в стране), *k* – наклон линии. При идеальной линейной зависимости достаточно двух значений сравниваемых переменных для того, чтобы вычислить наклон и точку пересечения и прочертить прямую линию. Для калькуляции k и b используются следующие формулы:

$$k = \frac{(Y_2 - Y_1)}{(X_2 - X_1)}$$

$$b = Y_1 - kX_1$$

Но правомерно ли ожидать, что распределение, которое мы видим на диаграмме, можно описать прямой линией? Ведь диаграмма рассеивания подсказывает нам лишь то, что это должна быть какая-то возрастающая функция, а в этом качестве могут выступать и показательная функция, и логарифмическая, и иные. При этом также видно, что какую бы функцию мы ни взяли, она не сможет описать все точки (пройти через все точки) – в социальных науках мы редко сталкиваемся с существованием линейной зависимости.

Однако этого и не требуется. Ведь в выражении общей формулы регрессии значения зависимой переменной описываются не просто как функция от независимой, а как сумма f(x) и *e.* Таким образом, можно принять, что несовпадения положения точек с графиком некоторой функции объясняются наличием именно «добавки» в виде *e* (факторов, которые мы в уравнение не включаем). Зависимость не является полностью линейной:

регрессионный анализ оценивает, *в какой степени* прямая линия соответствует рассеиванию. Стандартное уравнение для линии принимает следующую форму:

$$\hat{Y} = a + \beta X + \varepsilon ,$$

где Y – предсказанное значение зависимой переменной, X – наблюдаемое значение независимой переменной, a – точка пересечения оси ординат (соответствует b в формуле линейной функции), b – наклон линии регрессии (соответствует k в формуле линейной функции), e – ошибка, возникающая вследствие несовпадения предсказанных и реальных значений. Несовпадение происходит оттого, что на зависимую переменную, помимо исследуемых нами факторов, влияют и другие условия.

В данном случае нам необходима несколько иная формула для вычисления *a* и *b*. Однако логика остается вычислений простой. Прямая должна пролегать *максимально близко ко всем точкам графика*, то есть сумма расстояний от всех точек до искомой прямой должна быть как можно меньше. Метод решения задачи вычисления параметров регрессии путем минимизации выражения называется *методом наименьших квадрато в*(МНК / least sum of squares / ordinary least squares). Формула для наклона и точки пересечения в таком случае примет вид:

$$b = \frac{\Sigma(X_1 - \bar{X})(Y_1 - \bar{Y})}{\Sigma(X_1 - \bar{X})^2}$$

$$a = \bar{Y} - b\bar{X}$$

Эти формулы вычисляют разницу между фактическим значением Хи Y и средними значениями этих переменных.

Как это выглядит в нашем примере? Чтобы запустить линейную регрессию, в SPSS необходимо выбрать **Анализ** – **Регрессия** – **Линейная**. В качестве независимой переменной выставляем заболеваемость ВИЧ, а независимой – процент протестантов в стране. В результате программа выдает следующие значения:

Сводка для модели[b]

Модель	R	R-квадрат	Скорректированный R-квадрат	Стд. ошибка оценки
1	,425[a]	,180	,162	705715,654

a. Предикторы: (конст) % Protestant

b. Зависимая переменная: HIV

Таблица 1.5. Вывод коэффициента детерминации для модели.

Что означает данная таблица? Напомним, что уравнение регрессии описывается несколькими параметрами. Выражение $a + \beta X$ представляет собой ту часть значения *Y* для случая, которая объясняется линейным влиянием *X*. Что же касается ε, то это —

результат воздействия всех остальных факторов на *Y* для каждого объекта наблюдения. Другими словами, первое выражение — закономерная, объясняемая линейной моделью часть значения *Y*, а второе — часть, объясняемая всеми другими, возможно, случайными и малопонятными, причинами.

Регрессионная модель тем лучше, чем большая часть рассеивания объясняется изменением закономерной составляющей. Это соображение подталкивает к определению показателя, который может выступать как характеристика качества регрессионной модели. Традиционно таким показателем выступает отношение дисперсии объясняемой части к дисперсии необъясняемой части. Данный показатель называется коэффициентом детерминации и обозначается как **R-квадрат**.

В демонстрационном случае R-квадрат равен 0,180. Это означает, что 18% вариации зависимой переменной объясняется вариацией независимой переменной. Тот факт, что количество носителей ВИЧ-инфекции всего на 18% определяется доминированием протестантизма в стране, свидетельствует не в пользу тестируемой модели. Однако учитывая, какое количество условий *в действительности* определяет значение нашей зависимой переменной, это не такое уж маленькое число.

Перейдем к коэффициентам *a* и *b*, которые высчитала программа. Здесь мы видим следующую картину:

Коэффициенты[a]

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знч.
		B	Стд. Ошибка	Бета		
1	(Константа)	161640,575	132573,936		1,219	,229
	% Protestant	16365,350	5203,423	,425	3,145	,003

a. Зависимая переменная: HIV

Таблица 1.6. Вывод регрессионных коэффициентов.

Какие образом интерпретировать эту таблицу? Разберем ее, начиная с левого столбца.

Константа соответствует a, % Protestant (наша независимая переменная) – b.

Модель линейной регрессии выглядит следующим образом:

$$\hat{Y}(\text{количество ВИЧ} - \text{инфицированных}) = 161640{,}575 + 16365{,}350 * (\text{процент протестантов в}$$

На практике это означает следующее. Модель предсказывает, что при отсутствии (0%) протестантов количество ВИЧ-инфецированных составило бы 161641 человек. А увеличение количества протестантов в стране на 1% ведет к росту заболеваемости ВИЧ-инфекцией на 16365 человек.

В этом примере используется выборка, включающая практически все страны Африки южнее Сахары. В том случае, когда наша генеральная совокупность значительно отличается от выборки (если мы хотим обобщить результаты на все страны мира), нам необходимо оценить устойчивость коэффициентов регрессионной модели. Одним из показателей устойчивости выступает **стандартная ошибка коэффициентов**, которая демонстрирует дисперсию независимой переменной.

Стандартные ошибки коэффициентов (в примере они равняются 132573,9 и 5203,4, см. таблицу) дают оценку точности для этих коэффициентов при переносе результатов модели с выборки на генеральную совокупности. Говорить о том, что зависимость между заболеваемостью и количеством протестантов в стране описывается полученным уравнением, мы не имеем права, не указав, с каким уровнем точности можно переносить результаты выборки на генеральную совокупность. Вычисленные стандартные ошибки коэффициентов дают возможность с определенной, задаваемой нами вероятностью определить доверительные интервалы для характеристик регрессионной прямой в генеральной совокупности. В данном случае мы можем сказать, что с вероятностью 0,95 (z=1,96) значения коэффициентов a и b для модели в генеральной совокупности будут иметь вид:

$$a = 161540{,}75 \pm 1{,}96 * 132573{,}936$$

$$b = 16365{,}350 \pm 1{,}96 * 5203{,}423$$

Рассмотрим следующую колонку – стандартизированные коэффициенты. Зачем они нужны? Эти коэффициенты позволяют привести значения полученных множителей к единой системе измерения. В данном случае это не так важно, поскольку независимая переменная одна. Однако в уравнении с несколькими предикторами мы сталкиваемся с проблемой разных единиц измерения, что типично для социологических данных (если в уравнение включены, например, проценты с интервалом в сто единиц, годы с интервалом в тысячи и доходы с интервалом в сотни тысяч).

Поскольку размерности используемых переменных могут быть очень разные, оказывается, что регрессионные коэффициенты часто не дают нам возможности сказать, какая же из переменных сильнее влияет на результат. Для решения задачи сопоставления влияния независимых переменных на Y и используют стандартизованную форму регрессионного уравнения. При этом подходе все переменные в уравнении регрессии приводят к z-показателям: вместо Y и всех предикторов используют их стандартизованные значения.

Как изменится регрессионное уравнение, если вместо нестандартизированных значений мы используем z-значения? Во-первых, поскольку в результате преобразования не

изменятся коэффициенты корреляции между всеми переменными, показатель качества модели (коэффициент детерминации) останется тем же. Во-вторых, значение *a* (константа) в регрессионном уравнении станет равным нулю. Таким образом, для самой модели не меняется ничего – сила коэффициентов остается той же. Меняется только форма записи, которая становится гораздо более удобной. Поскольку в отличие от использовавшихся в основном уравнении предикторов все z в уравнении имеют одинаковый масштаб измерений, регрессионные коэффициенты в этом уравнении сравнимы между собой. Таким образом, сопоставляя эти коэффициенты, мы можем понять, какой из предикторов оказывает на зависимую переменнуюнаиболее сильное влияние.

Подчеркнем, что стандартизованные коэффициенты регрессии *не заменяют* нестандартизованных: у них иное назначение. Если нестандартизованные коэффициенты показывают, как меняется зависимая переменная при изменении соответствующей независимой на условную единицу (километры, килограммы, доллары), стандартизованные коэффициенты позволяют сопоставить между собой общую степень воздействия каждого из предикторов на зависимую переменную.

Еще один важный показатель, который автоматически выводит SPSS–**t.** Он показывает, во сколько раз полученное значение коэффициента превосходит его стандартную ошибку.

16365,350/ 5203,423 = 3,145

161640,575 / 132573,936 = 1,219

Эти значения необходимо сравнить с критическими точками t-распределения с тем, чтобы выяснить, можно ли принять нулевую гипотезу. При «ручном» способе обработки нам понадобятся статистические таблицы уровня значимости для t-статистики. Однако в SPSS этого не требуется – здесь автоматически выводится колонка **Знч.**, которая показывать уровень значимости. В ней выводится вероятность, с которой вычисленный для произвольной выборки регрессионный коэффициент будет больше или равен найденному нами значению (при условии равенства нулю коэффициента в генеральной совокупности). Это тот уровень значимости, на котором может быть отвергнута нулевая гипотеза об отсутствии связи. В нашем примере коэффициент b значим на уровне 0,003, коэффициент a незначим (0,229> 0,05).

Другой важный показатель, который также следует учитывать при обобщении полученных результатов, называется **F-статистика**. Что это такое? Мы установили доверительные интервалы для коэффициентов регрессии. Сделать то же самое для коэффициента детерминации не представляется возможным. Однако мы можем указать

вероятность, с которой независимая переменная влияет на зависимую. Для оценки значимости коэффициента детерминации и используется F-статистика, которая вычисляется как отношение объясненной суммы квадратов (в расчете на одну переменную) к необъясненной сумме квадратов (в расчете на одну степень свободы).

Дисперсионный анализ[a]

Модель		Сумма квадратов	ст.св.	Средний квадрат	F	Знч.
1	Регрессия	4926424957397,332	1	4926424957397,332	9,892	,003[b]
	Остаток	22411556299198,414	45	498034584426,631		
	Всего	27337981256595,746	46			

Таблица 1.7. F-статистика для регрессионной модели.

В примере F=9,892, которому соответствует уровень значимости 0,003<0,005. Эту фразу следует расшифровывать следующим образом: с вероятностью в 97% мы можем утверждать, что количество протестантов влияет на уровень заболеваемости ВИЧ в стране.

В графическом виде линия регрессия для примера выглядит следующим образом.

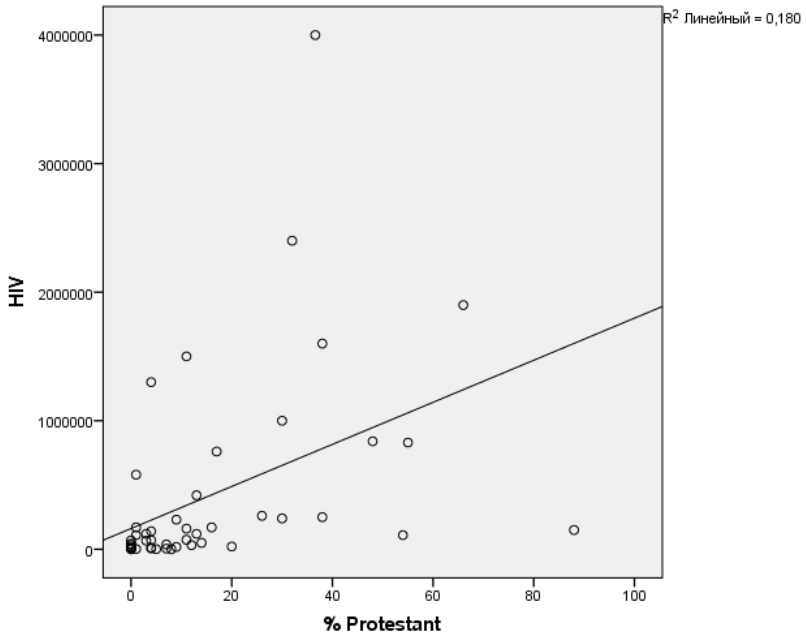


Рисунок 1.2. Диаграмма рассеивания с добавлением линии регрессии.

Для сравнения приведем вычисленные значения коэффициентов для независимой переменной «процент мусульман в стране». Попробуйте проинтерпретировать эти таблицы самостоятельно.

Сводка для модели[b]

Модель	R	R-квадрат	Скорректированный R-квадрат	Стд. ошибка оценки
1	,147[a]	,022	,000	770909,794

a. Предикторы: (конст) % Muslim

b. Зависимая переменная: HIV

Таблица 1.8. . Коэффициент детерминации для регрессии с независимой переменной "Число мусульман".

Коэффициенты[a]

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знч.
		B	Стд. Ошибка	Бета		
1	(Константа)	520477,198	147904,476		3,519	,001
	% Muslim	-3353,901	3353,638	-,147	-1,000	,323

a. Зависимая переменная: HIV

Таблица 1.9. Регрессионные коэффициенты уравнения с независимой переменной "Число мусульман".

Каким образом проверить надежность получившейся модели? Существует несколько предпосылок, определяющих надежность модели и возможность получения на ее основе достоверных обобщений. Первая предпосылка – необходимость наличия *нормального распределения остатков*. Остатки – результат деятельности большого числа различных факторов, поэтому логично ожидать, что ни один из этих факторов не должен оказывать большего влияния, чем остальные. Остатки должны представлять собой случайные величины, а значит подчиняться закону нормального распределения. Это означает следующее: основная масса точек должна лежать близко к регрессионной прямой, а чем дальше от прямой, тем точек должно быть меньше.

Каким образом проверить нашу модель на наличие нормального распределения остатков? В SPSS мы можем посчитать остатки по каждому наблюдению и построить гистограмму их распределения. Для этого в меню «**Линейная регрессия**» следует нажать кнопку «**Сохранить**» и в колонке «**Остатки**» выбрать «**Нестандартизированные значения**». После этого программа создаст новую переменную под названием *Residuals*, в которой вычислит остатки по наблюдениям.

Получившиеся остатки можно представить визуально в виде гистограммы (**Графика - Гистограмма**).

Получившаяся гистограмма напоминает нормальную кривую распределения, однако не полностью соответствует ей: малых остатков на кривой подавляющее количество, и чем крупнее остаток, тем меньше случаев. При этом значение остатков сильно отличается от кривой нормального распределения. Поскольку большая часть данных располагается у линии регрессии, можно говорить о том, что наши наблюдения представляет собой единый массив, подчиняющийся схожим закономерностям, влияющим на зависимую переменную. Из общей массы особо выступает случай ЮАР, демонстрирующий наибольший положительный остаток

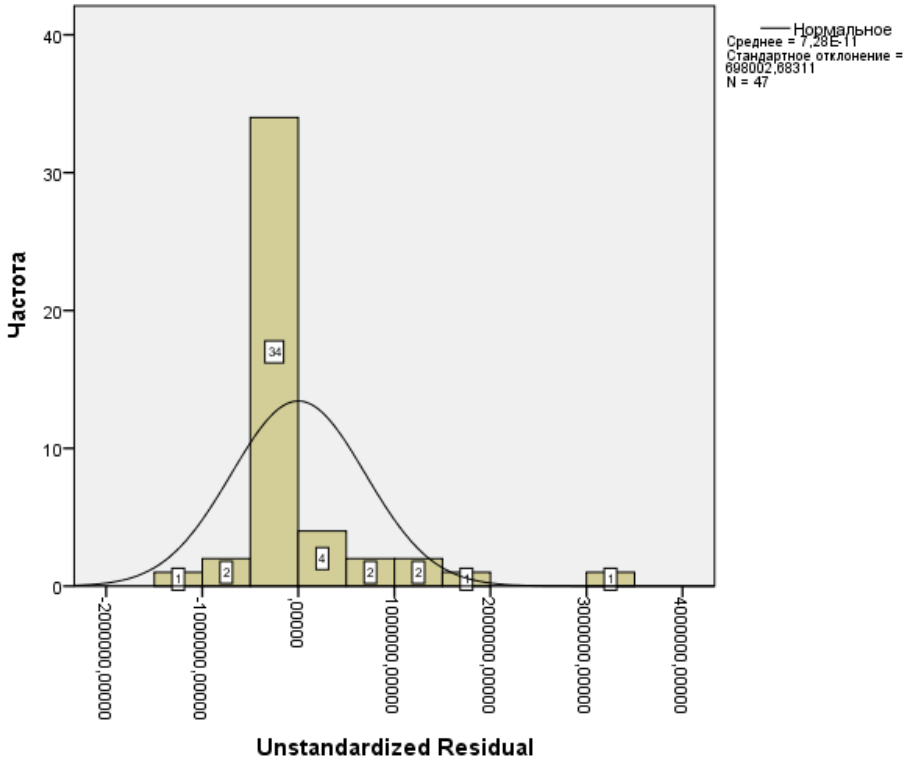


Рисунок 1.3. Диаграмма нестандартизированных остатков.

Второй важной предпосылкой надежности модели выступает равенство дисперсии распределения остатков, или **гомоскедастичность.** Если разброс наблюдений вокруг линии регрессии окажется неодинаковым, это будет означать, что стандартные ошибки регрессионных коэффициентов также будут разными, что чревато смещением результатов анализа как минимум относительно части массива данных. Если изменение остатков представляет собой не набор случайных величин, а функцию от независимой переменной (например, с увеличением количества протестантов начинает увеличиваться и разброс остатков), модель оказывается ненадежной; в таком случае следует строить несколько моделей регрессии для разных массивов данных.

Как можно проверить остатки на гомоскедастичность? Вновь прибегнем к помощи графиков (**Графики→Линия→Значения по группам наблюдений**). Если на оси абсцисс

отложить независимую переменную, а ординат – нестандартизированные значения регрессионных остатков, график примет такой вид:

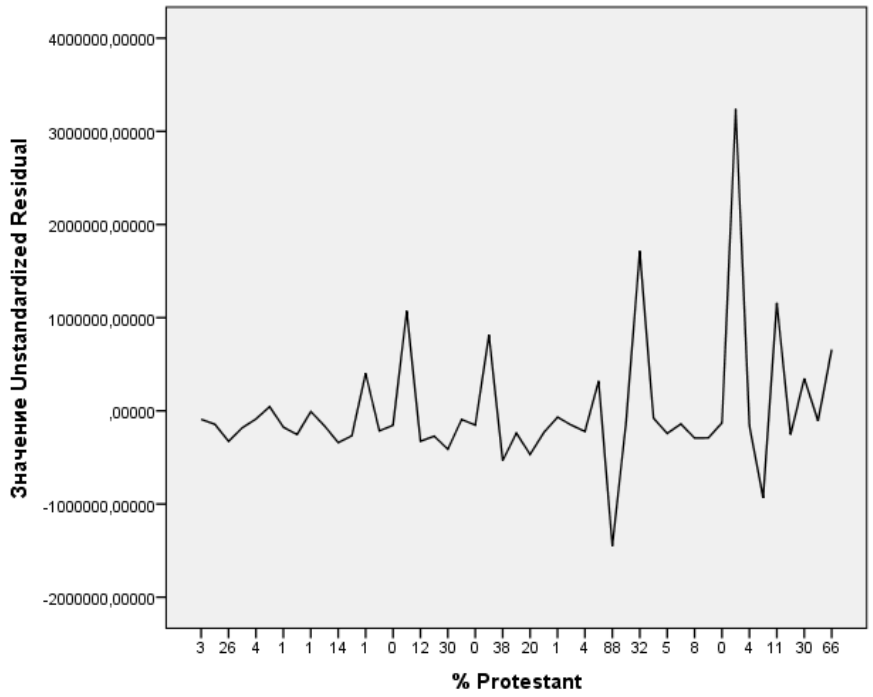


Рисунок 1.4. Проверка модели на гомоскедастичность.

Мы можем видеть, что по мере увеличения количества протестантов в стране растут и регрессионные остатки. Это указывает на неслучайный характер дисперсии распределения остатков - **гетероскедастичность**. Каким образом трактовать такое распределение? Судя по всему, протестантизм в данном случае не является переменной, которая самостоятельно влияет на уровень заболеваемости ВИЧ. В священных текстах протестантских сект, деноминаций и церквей не содержится ни слова о доблести распространения заболеваний, передающихся половым путем. Скорее количество протестантов в стране является косвенным индикатором наличия переменных, которые в гораздо большей степени влияют на интересующий нас результат.

Значит, необходимо подняться в теоретическом плане на ступень выше и подумать, например, каким образом в страну мог попасть протестантизм? Что принесли с собой протестантские страны-гегемоны после колонизации африканских стран? Может быть, особый характер труда, вынуждающий рабочих мигрировать и подолгу находиться далеко от семьи? Может быть, развитие производства и уровень индустриализации, заставляющий африканцев мигрировать из сельской местности в города с высокой плотностью населения? Так или иначе, наличие гетероскедастичности показывает, что, вероятнее всего, **существует некая пропущенная переменная, одновременно связанная как с независимой, так и зависимой переменными** – она и определяет функцию распределения регрессионных остатков.

В качестве метода борьбы с гетероскедастичностью рекомендуется искать переменные, которые сильно связаны как с Y, так и с X. Найдя такую переменную, можно разделить на нее X и Y и затем искать регрессию уже для этих новых переменных.

Зависимость уровня распространения ВИЧ от протестантской «ориентированности» страны является примером простой, или парной регрессии. Однако техники линейной регрессии могут быть приложены к любому количеству переменных. Количество включенных условий, как утверждалось в начале, ограничивается числом наблюдений.
Для построения математической модели одновременного влияния нескольких факторов (независимых переменных, предикторов) на зависимую переменную используют усложненный вариант простой линейной регрессии — модель **множественной линейной регрессии**, которая является расширением уже знакомой модели

$$\hat{Y} = a + \beta X_1 + \beta X_2 + \cdots + \beta X_n + \varepsilon$$

Продолжим исследование причин распространения ВИЧ-инфекции в странах Африки, добавив несколько новых условий.

Во-первых, мы вправе ожидать, что на распространение вируса влияет уровень грамотности: неграмотный человек не знает про опасность болезни и не осознает риска. Соответственно, чем выше грамотность, тем уровень распространения ВИЧ должен быть меньше.

Обратимся к базам данных в поисках эмпирических индикаторов. В демонстрационном примере в регрессионную модель включаются две переменные:

- Процент протестантов (как имеющий наибольшую корреляцию среди всех «религиозных» переменных);
- Грамотность (процент жителей 15-49 лет, имеющий читать и писать) (UNDP 2000)

Приступая к построению множественной регрессионной модели, прежде всего необходимо ответить на вопрос: существует ли вообще хоть какая-то зависимость между y и предикторами? Быть может, никакой зависимости нет, и наши усилия по построению модели заведомо обречены на неудачу? Поэтому выбор переменных следует делать исходя из следующих условий:

- В анализ следует включать теоретически важные переменные (регрессионный анализ всегда опирается на исследовательский вопрос: «Что определяет изменение зависимой переменной?» и не может мыслиться вне предварительной теоретической работы);

- В анализ следует включать переменные, которые имеют строгую ассоциацию с объясняемой переменной.

Как и в ситуации простой регрессионной модели, индикатором наличия зависимости выступает коэффициент корреляции Пирсона. При выборе независимых переменных для модели целесообразно вычислить корреляции между независимой переменной и предикторами.

Корреляции

		1995	Illiteracy rate, adult total (% of people 15+)
1995	Корреляция Пирсона	1	,287
	Знч.(2-сторон)		,051
	N	47	47
Illiteracy rate, adult total (% of people 15+)	Корреляция Пирсона	,287	1
	Знч.(2-сторон)	,051	
	N	47	47

Таблица 1.10. Коэффициенты корреляции для двух переменных.

Корреляция уровня грамотности и зависимой переменной есть, и она положительная, хотя и не слишком сильная.

Посмотрим значения коэффициентов, которые вычислила программа:

Сводка для модели

Модель	R	R-квадрат	Скорректированный R-квадрат	Стд. ошибка оценки
1	,478[a]	,229	,194	481098,012

a. Предикторы: (конст) Illiteracy rate, adult total (% of people 15+) , % Protestant

Коэффициенты[a]

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знч.
		B	Стд. Ошибка	Бета		
1	(Константа)	82914,853	230749,188		,359	,721
	% Protestant	12240,966	4138,694	,457	2,958	,005
	Illiteracy rate, adult total (% of people 15+)	1060,850	4125,389	,040	,257	,798

a. Зависимая переменная: 1998

Таблица 1.11. R-квадрат и регрессионные коэффициенты для модели многомерной регрессии.

Дисперсионный анализ[a]

Модель		Сумма квадратов	ст.св.	Средний квадрат	F	Знч.
1	Регрессия	3023945378235,342	2	1511972689117,671	6,532	,003[b]
	Остаток	1018403305878 5,936	44	231455296790,589		
	Всего	1320797843702 1,277	46			

a. Зависимая переменная: 1998

b. Предикторы: (конст) Illiteracy rate, adult total (% of people 15+) , % Protestant

Таблица 1.12. Результаты дисперсионного анализа для четырех переменных.

Что мы видим в данной таблице? Коэффициент детерминации незначительно, но возрос. Его значимость 0,003<0,005. В то же время коэффициент независимой переменной «количество протестантов» увеличился. Уровень грамотности не имеет статистической значимости. Это означает, что наша гипотеза о влиянии количества людей, умеющих читать и писать, на распространенность заболевания, не подтверждается.

Важное правило, которое необходимо соблюдать при работе с множественной регрессией, заключается в необходимости проверки независимых переменных на **мультиколлинеарность.** Проблема мультиколлинеарности – это проблема связанности независимых переменных (предикторов) друг с другом. Хотя классический регрессионный анализ предполагает, что X независимы между собой, на практике оказывается, что так бывает достаточно редко. Как правило, между иксами есть корреляция, и порой достаточно высокая.

В каких случаях нам следует задуматься о мультиколлинеарности? Здесь нас могут подтолкнуть следующие свидетельства:

- Регрессионные коэффициенты значительно изменяются по мере удаления или добавления новых предикторов;
- Регрессионный коэффициент отрицательный, хотя, исходя из теории, значения зависимой переменной должны расти пропорционально изменению предиктора (или наоборот);
- Ни один из коэффициентов не обладает статистической значимостью, однако F-статистика показывает значимость коэффициента детерминации.
- Регрессионный коэффициент не является значимым, хотя теоретически связь между ним и зависимой переменной должна быть существенной.

- При изменении данных (увеличении или уменьшении выборки) оценки коэффициентов значительно изменяются

Как получается, что коэффициенты могут не иметь значимости при подтвержденной надежности R-квадрат? Дело в том, что когда переменные связаны друг с другом, они в сущности являются носителями одной и той же информации (как два человека, которые одновременно кричат – сложно выделить отдельные голоса, поскольку они перекрывают друг друга), при этом их стандартные ошибки накладываются друг на друга, что искажает результаты. Мультиколлинеарность повышает стандартную ошибку коэффициентов, тем самым искусственно снижая значимость.

Продемонстрируем это на простом примере: добавим еще одну переменную «процент людей, которые не исповедуют ни католичество, ни ислам, ни традиционные религии стран». Мы понимаем, что данная переменная должна значительно коррелировать с переменной «процент протестантов» (хотя и не полностью совпадает с ней). Проверим регрессионную модель по этим двум переменным.

Сводка для модели

Модель	R	R-квадрат	Скорректированный R-квадрат	Стд. ошибка оценки
1	,523[a]	,273	,240	467094,755

a. Предикторы: (конст) relig, % Protestant

Таблица 1.13. Сводка для модели с присутствием мультиколлинеарности.

Дисперсионный анализ[a]

Модель		Сумма квадратов	ст.св.	Средний квадрат	F	Знч.
1	Регрессия	3608167998066,469	2	1804083999033,234	8,269	,001[b]
	Остаток	9599810438954,809	44	218177509976,246		
	Всего	13207978437021,277	46			

a. Зависимая переменная: 1998

b. Предикторы: (конст) relig, % Protestant

Таблица 1.14. Результаты дисперсионного анализа для модели с наличием мультиколлинеарности.

Коэффициенты[a]

Модель	Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знч.
	B	Стд. Ошибка	Бета		

	(Константа)	107193,841	89632,999		1,196	,238
1	% Protestant	3274,486	6693,774	,122	,489	,627
	relig	9501,725	5731,954	,414	1,658	,104

a. Зависимая переменная: 1998

Таблица 1.15. Коэффициенты регрессии в модели с наличием мультиколлинеарности.

Коэффициент детерминации показывает, что данная модель объясняет 27% вариации, и значимость R-квадрат высокая. Однако ни один из коэффициентов теперь не является статистически значимым – верный признак присутствия мультиколлинеарности.

Как определить, что переменные не являются независимыми друг от друга? Наиболее простой способ – проверить их на наличие корреляции. В данном случае корреляционная матрица выглядит следующим образом:

Корреляции

		% Protestant	relig	1998
% Protestant	Корреляция Пирсона	1	,857**	,477**
	Знч.(1-сторон)		,000	,000
	N	47	47	47
relig	Корреляция Пирсона	,857**	1	,519**
	Знч.(1-сторон)	,000		,000
	N	47	47	47
1998	Корреляция Пирсона	,477**	,519**	1
	Знч.(1-сторон)	,000	,000	
	N	47	47	47

**. Корреляция значима на уровне 0.01 (1-сторон.).

Таблица 1.16. Наличие высокой корреляции независимых переменных.

При каких значениях взаимосвязи между предикторами можно сказать, что мы сталкиваемся с проблемой мультиколлинеарности? В некоторых работах можно встретить рекомендации, указывающие пороговое значение как 0,7. Тем не менее, однозначных рекомендаций в данном случае не существует, и нам следует полагаться и на другие способы проверки на мультиколлинеарность.

Вернемся к нашему примеру с протестантизмом и уровнем грамотности и рассмотрим некоторые показатели. Для этого в **Вывод** необходимо включить показатели коллинеарности (**Анализ→Регрессия→Линейная→Статистики→Диагностика коллинеарности**).

Коэффициенты[a]

Модель	Статистики коллинеарности	
	Толерантность	КРД

1	% Protestant	,735	1,361
	Illiteracy rate, adult total (% of people 15+)	,735	1,361

a. Зависимая переменная: 1998

Таблица 1.17. Коэффициенты вздутия и толерантности.

Диагностики коллинеарности[a]

Модель	Измерение	Собственное значение	Показатель обусловленности	Доли дисперсии		
				(Константа)	% Protestant	Illiteracy rate, adult total (% of people 15+)
1	1	2,544	1,000	,01	,05	,01
	2	,415	2,477	,05	,75	,01
	3	,042	7,808	,94	,20	,98

a. Зависимая переменная: 1998

Таблица 1.18. Анализ коллинеарности переменных.

Один из часто используемых показателей коллинеарности – **толерантность** (допуск переменной). Толерантность определяется по формуле 1-R-квадрат. Проверяя уровень толерантности, фактически мы запускаем новую регрессию, где в качестве зависимой переменной выступает одна из независимых - тогда толерантность показывает, какое количество вариации не объясняется другими переменными. Если толерантность мала (<0,2), модель вызывает серьезные подозрения, поскольку в этом случае переменная представляет собой близкую к линейной комбинацию других независимых переменных.

С толерантностью связан и другой фактор – вздутие вариации (КРД), который является обратным по отношению к толерантности (1 / (1-R-квадрат). С возрастанием фактора влияния на дисперсию возрастает и дисперсия регрессионного коэффициента, поэтому высокие значения КРД (>5)свидетельствуют о наличии мультколлинеарности.

Другой характеристикой коллинеарности выступают **собственные значения**. Если собственные значения превышают 13, это свидетельствует о возможном наличии мультиколлинеарности, если они выходят за 80, это является очень серьезным сигналом. Другой признак – высокое значение показателей **обусловленности**. Чем выше обусловленность, тем в большей степени введение данной переменной способствует появлению мультиколлинеарности. Здесь интервалы следующие – показатель обусловленности >15 идентифицирует возможную проблему, если же он превышает 30, это уже является свидетельством наличия устойчивой мультиколлинеарности

После проведения регрессии необходимо реализовать еще одну важную процедуру - проанализировать модель на наличие выбросов. Выбросы – это наблюдения, которые далеко отстоят от линии регрессии и не вписываются в общую тенденцию. Очевидно, что выбросы увеличивают стандартную ошибку коэффициента, снижают значение R-квадрат. Кроме того, выбросы могут не вписываться в общую модель в силу иного характера воздействия предикторов на результат: зависимость для них может быть совершенно иной, и попытка построить единую модель здесь не увенчается успехом.

Наличие выбросов играет и положительную роль для исследования. Само по себе их присутствие показывает существование резко отклоняющихся из общей совокупности групп или отдельных случаев. Чтобы решить эту проблему, массив придется разбивать на однородные группы и строить для них собственные модели. Таким образом, наличие выбросов в уравнении регрессии может способствовать более точной классификации при проведении сравнительного исследования.

Чтобы исключить выбросы из модели, необходимо провести диагностику остатков (**Анализ→Регрессия→Линейная→Статистики→Остатки→Диагностика по наблюдениям**). Здесь необходимо установить границу – что считать выбросом (**Выбросы за пределами…**) Изначально SPSS руководствуется правилом трех стандартных отклонений. Однако значение можно задать любое, и руководствоваться здесь следует величиной дисперсии остатков – если стандартное отклонение остатков большое, тогда установка небольшого порога выбросов исключит слишком большое число случаев, что нежелательно с точки зрения репрезентативности.

Споры вокруг регрессионного анализа

Построение регрессионных уравнений, как упоминалось в самом начале, считается весьма популярным, а также одним из «наиболее строгих, дисциплинированных и научных аналитических методов» установления причинно-следственных связей среди исследователей, отдающих предпочтение количественным методам сравнения. Однако это не означает, что данный метод лишен принципиальных недостатков. Здесь мы приведем некоторые аргументы, указывающие на уязвимые места регрессионного анализа в сравнительных исследованиях.

Главной задачей социального ученого является построение модели и проверка причинно-следственных связей. В идеально-типическом случае теории оперируют разными переменными, предлагая ясные объяснения того, *каким образом* эти переменные связаны с наблюдаемым результатом. Однако в действительности значительная часть теорий не

объясняет механизмы связи причины и следствия (в частности, не уточняет условий, при которых искомые переменные становятся причинами). Часто исследователь ограничивается формированием общего списка потенциально важных переменных, основываясь на приблизительных представлениях о социальном феномене. Тогда ключевой задачей эмпирического исследования выступает оценка относительной важности каждой переменной из списка: если предикторы показывают самые высокие регрессионные коэффициенты, они принимаются в качестве наилучшего объяснения.

С критикой такого подхода выступает Ч. Рагин, называя его «результирующим мышлением» (net-effect thinking). Данное мышление, пишет он, основывается на представлении о том, что каждая переменная сама по себе способна влиять на вероятность наступления того или иного результата. Это означает, что предиктор действует *независимо* от других переменных (и их сочетаний). Кроме того, использование регрессии накладывает на нашу модель свойство **аддитивности** – то есть общий вклад всех переменных в значения целевой функции определяется как простая сумма вкладов каждой отдельной переменной. Это значительно упрощает интерпретацию причинно-следственных связей, однако ведет к тому, что влияние каждой переменной на результат (результирующий эффект) рассматривается как неизменное *при любых значениях* других переменных. Более того, связь предикторов друг с другом рассматривается как нежелательное и вредоносное явление, о чем мы говорили при обсуждении мультиколлинеарности. Такой подход, в сущности, отрывает всякое социальное явление от того *контекста*, в котором оно существует. Эндрю Эббот резюмировал данное положение словами: «Слишком часто общие модели ведут к линейному представлению о реальности и ограниченному видению социальных процессов» [Abbott; 183].

По существу, «результирующее мышление» выгодно именно «слабым теориям», в общем виде характеризующим социальный феномен и не претендующим на выявление **комплексной, или конъюнктурной каузальности**. Что такое комплексная каузальность? Один и тот же результат может быть следствием влияния переменных, которые действуют не раздельно, но только в сочетании друг с другом. Предикторы также могут образовывать кластеры, в некоторых обстоятельствах заменять друг друга, а также давать значимый результат только при определенных условиях. Ситуации, когда различные условия могут приводить к одному результату, называется **множественной каузальностью**[1]. Помимо этого, условие может приводить и к противоречивым эффектам

[1] Проблема сложности каузальных моделей будет рассматриваться на протяжении всей методической части.

(способствовать наступлению результата или, наоборот, отдалять его) в различных контекстах. При изучении таких сложных объектов, как социальные явления, нельзя быть абсолютно уверенным, что открытая каузальная связь является законом – в разных случаях могут «работать» различные условия и каузальные цепочки. Представим себе исследование, посвященное связи высшего образования и материального благополучия. Влияет ли его наличие на положение «белого православного россиянина из семьи представителей среднего класса»? Безусловно. Однако высшее образование гипотетически играет еще большую роль для будущего представителей этнических / религиозных меньшинств или мигрантов из семей с низким доходом, и связано это с другими факторами. В данном случае одна и та же переменная может иметь совершенно разное влияние в разных контекстах.

Если же мы обратим внимание на исследования, опирающиеся на регрессионную модель, большинство из них, хотя и представляют статистически достоверные результаты, однако мало говорят о том, *почему* те или иные переменные приводят к интересующему результату, ограничиваясь констатацией общей связи, присущей выборке в целом (аллегория «средней температуры по больнице» здесь как нельзя кстати). Регрессионный анализ не дает возможности выявления комплексной каузальности через простое добавление новых независимых переменных – статистический анализ не делает различий между аддитивными переменными, специфическими обусловленностями и конъюнктурной каузальностью. По образному выражению Ч. Рагина, регрессия (как, впрочем, и другие методы многомерного анализа) делает случаи *невидимыми* и использует их только как источник эмпирических наблюдений по абстрактным переменным [Ragin 1987].

Нельзя утверждать, что регрессия совсем не работает с комплексной каузальностью. Представим такой абстрактный набор данных:

A	B	Y
0	1	-1
0	1	-1
0	1	-1
1	0	-1
1	0	-1
1	0	-1
1	1	1

1	1	1
0	0	0
0	0	0

Таблица 1.19. Пример набора данных с эффектом взаимодействия переменных.

Сводка не показывает значительных результатов влияния переменных A и B на Y (скорректированный R-квадрат принимает отрицательное значение).

Сводка для модели

Модель	R	R-квадрат	Скорректированный R-квадрат	Стд. ошибка оценки
1	,395[a]	,156	-,085	,878

a. Предикторы: (конст) Var2, Var1

Таблица 1.20. Коэффициент детерминации для переменных с эффектом взаимодействия.

Однако если мы добавим еще один предиктор, показывающий взаимодействие переменных (A*B), результаты получатся совершенно иными:

Сводка для модели

Модель	R	R-квадрат	Скорректированный R-квадрат	Стд. ошибка оценки
1	1,000[a]	1,000	1,000	,000

a. Предикторы: (конст) Var3, Var2, Var1

Таблица 1.21. Коэффициент детерминации для модели с выраженным эффектом взаимодействия.

Коэффициенты[a]

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знч.
		B	Стд. Ошибка	Бета		
1	(Константа)	-5,551E-016	,000		,000	1,000
	Var1	-1,000	,000	-,625	-47453132,812	,000
	Var2	-1,000	,000	-,625	-47453132,812	,000
	Var3	3,000	,000	1,500	100663296,000	,000

a. Зависимая переменная: Outcome

Таблица 1.22. Регрессионные коэффициенты модели с выраженным эффектом взаимодействия.

По отдельности оба предиктора оказывают отрицательный эффект, однако, взаимодействуя друг с другом, образуют положительный регрессионный коэффициент. В данном случае каузальная способность переменных обусловлена именно их сочетанием.

Другой практической проблемой регрессионного анализа выступает необходимость изоляции переменных. Многие переменные в силу своей природы сильно связаны друг с другом, и воспринимать это исключительно как проблему не следует. Наоборот, величайший интерес представляет сам факт их совместной работы: накладываясь друг на друга, различные переменные (например, семейный бэкграунд, тип школы, место проживания в случае исследования образовательного неравенства) усиливают общий эффект – именно их пересекающийся характер обеспечивает силу и устойчивость конечного явления.

Существует несколько проблем, которые вытекают из представления о причинно-следственных связях, свойственных регрессионному анализу. Фундаментальным практическим недостатком является зависимость наших умозаключений от количества переменных, включенных в модель. Ограничьтесь небольшим набором переменных, и отдельный предиктор покажет значительное воздействие на результат; включите большее количество переменных – его влияние может сойти на нет. Знатоки регрессионного анализа, понимая это свойство, уделяют большое внимание спецификации модели, однако такая спецификация требует сильной теории и глубоких знаний о случаях, чем может похвастаться не каждый исследователь, работающий в количественной стратегии.

Переменные в регрессионном анализе изолируются не только друг от друга, но и от временных характеристик, в рамках которых они развиваются. Всякий результат является следствием стечения множества исторических обстоятельств, которые условно можно разделить на две категории: временной контекст (эффекты, характерные для определенного периода) и временная последовательность (эффекты, связанные с уникальным историческим развитием объекта). Это не означает, что регрессионный анализ не имеет инструментов работы с темпоральным измерением (впоследствии мы расскажем об одном из них – анализе временных рядов). Однако глубокое изучение исторического развития в количественном анализе сталкивается с недостатком информации и проблемой малой генеральной совокупности.

Констатацией раскола, связанного с оценкой методологической корректности метода регрессионного анализа, можно считать дискуссию, вошедшую в двадцать четвертый том серии Comparative Social Research (2009), вторая часть которого полностью посвящена обсуждению резонансной работы Майкла Шалева «Ограничения и альтернативы множественной регрессии в сравнительном исследовании» [Shalev].

М. Шалев дополняет уже озвученные аргументы несколькими новыми [Shalev]. Во-первых, регрессионный анализ дает слишком грубые результаты при наличии проблемы малой генеральной совокупности, характерной для сравнительной макросоциологии. В отличие от исследований с крупными выборками (например, национальных опросов), где регрессионная модель вычисляется по большому числу конфигураций значений переменных, в исследованиях крупных объектов таких конфигураций обычно не много. Более того, если определенная конфигурация не встречается в выборке (которая равна генеральной совокупности), ее, вероятно, не существует и в природе.

Представим простой пример: мы хотим построить регрессионную модель зависимости уровня публикационной активности факультетов Санкт-Петербургского государственного университета. Возьмем всего три предиктора – средний уровень зарплат преподавателей, количество преподавателей с ученой степенью, объем грантов, полученных преподавателями, - и предельно грубое измерение (переведем показатели в ранговые переменные по пятизначной шкале). Даже при таком поверхностном измерении таблица возможных сочетаний условий и результата будет включать в себя 625 клеток (5*5*5*5). Однако в СПБГУ на 2013 год всего 24 факультета. Это означает, что даже если все факультеты продемонстрируют разные конфигурации значений и результата, пустым останется подавляющее большинство клеток. Регрессионная модель в поисках заветной линии автоматически заполнит их несуществующими в действительности значениями, что, в свою очередь, может привести к сильным искажениям коэффициентов итогового уравнения.

Джон Голдторп писал, что «проблема малой генеральной совокупности относится не к методу, но скорее к данным» [Goldthorpe 2009; 8], рекомендуя обращаться к объединенной информации по нескольким временным отрезкам (см. параграф 3.2..). Традиционно количественные исследования включают либо кросс-секционные (синхронные) измерения по нескольким объектам, либо измерения одного объекта во времени. Объединенные (pooled) наборы данных совмещают эти перспективы, давая возможность исследовать как вариацию переменных между объектами, так и динамическую вариацию. Количество наблюдений – и, соответственно, степеней свободы, - в таком случае увеличивается, что повышает надежность результатов. М. Шалев, однако, уточняет, что использование объединенных массивов данных при построении регрессии может привести к принципиальным ошибкам в том случае, если темпоральная и кросс-секционная вариации имеют разные каузальные логики. Концентрация на одной гипотезе в таком случае приведет к неверному толкованию результатов. Более конструктивной

стратегией может выступить параллельное исследование временного ряда для каждого объекта и последующее сравнение результатов. Но в этом случае статистический вывод скорее всего окажется гораздо более разнородным и сложным для корректной интерпретации. Так или иначе, если между линиями регрессии, полученными в результате темпорального или кросс-секционного исследования, находится мало общего, это будет являться веским доводом против построения и интерпретации результатов единой регрессионной модели.

О малой эффективности регрессионного анализа для установления каузальных связей в сравнительных исследованиях с малой генеральной совокупностью критически высказывается и Госта Эспинг-Андерсон [Esping-Anderson]. По его мнению, регрессионные коэффициенты, полученные на таком массиве данных, редко отражают реальные каузальные связи. Тем не менее, от регрессии отказываться не следует, поскольку она остается эффективным механизмом корректировки гипотезы. В этом отношении, пишет он, наиболее ценную информацию несут не коэффициенты, а *регрессионные остатки* – в исследовании небольшого количества объектов остатки легко соотнести с конкретными случаями для дальнейшего исследования. Тогда регрессионный анализ оказывается незаменимым инструментом развития теории.

Данная позиция перекликается с аргументом «Логика социального исследования», в которой Адам Пржеворски и Генри Тьюн предлагают относить всякие отклонения, наблюдаемые в процессе измерения силы связи переменных, к внутрисистемным свойствам. Таким образом, полученные остатки следует использовать как повод для проведения более тщательного изучения с использованием кейс-стади или сравнения нескольких случаев.

Так или иначе, регрессионный анализ остается в наборе методов, которые необходимо знать и уметь применять любому социальному ученому, проводящему сравнительные исследования. Регрессионный анализ не всегда может быть использован для получения надежных результатов. Однако трудно переоценить его значимость как инструмента для определения наличия связи между переменными и для формирования каузальных гипотез.

Методы выбора случаев

Одной из первых задач, с которыми сталкивается определившийся с проблематикой своег о исследования ученый, является организация поля. Всякое социологическое поле можно представить как совокупность единиц анализа (случаев), их характеристик и числа наблюд

ений.

Хотя качественная стратегия по определению своему должна работать с целостными объе ктами (случаями), в каждой из этих целостностей исследователя привлекают только опред еленные ее стороны – существенные характеристикам, по которым он производит сравнен ие. Поэтому, хотя для удобства мы поместили параграфы, посвященные случаям и характе ристикам (переменным), друг за другом, это не означает, что сначала ученый решает, каки е случаи он будет исследовать, а уже затем – что именно его интересует (на практике така я последовательность вообще вряд ли возможна). Важно придерживаться сформулирован ной проблемы и общей гипотезы, которые определяют протекание единовременных проце ссов выбора и тех, и других.

В начале работы необходимо определить однородную область исследования - установить границы популяции, внутри которых будет производиться отбор случаев. Случаи должны быть подобны друг другу и сравнимы в рамках заданного измерения. В этом смысле, праг матичным будет с самого начала определить интересующие нас объект и проблему, а уже потом приступать к выбору случаев.

Проблема, которую сформулирована в методологической части исследования, должна быт ь «переведена» в методическое измерение и стать предельно конкретной. Поскольку сравн ительное исследование сконцентрировано на поиске причинно-следственных связей, логи чно будет развернуть сформулированную проблему в суждение, представленное как резул ьтат, для которого нужно найти причины. Подобным образом мы конкретизируем нашу пр облему и трансформируем ее в конкретную исследовательскую задачу. После этого, уже и сходя из сформулированного результата, нам будет гораздо проще выбрать случаи и опре делить интересующие нас переменные. Кроме того, формулировка следствия – это уже ва жный шаг на пути редуцирования многообразия социальных явлений к теоретической мод ели, поскольку она предполагает формирование определенной аксиоматики в рамках иссл едования. Другими словами, мы наделяем исследуемые случаи некоторыми общими фоно выми характеристиками, которые получают статус констант для нашего исследования.

Итак, с помощью формулировки проблемы и предварительного результата мы задаем неко торые границы, популяция случаев внутри которых отвечает критерию однородности и де монстрирует общие фоновые особенности («общий бэкграунд», как выражаются в повсед невной речи специалисты). Поскольку большинство качественных исследований строится на небольшом количестве случаев, перед ученым возникает проблема корректного выделе ния *границ популяции* (ведь мы не можем прибегнуть к статистическим процедурам опред

еления генеральной совокупности). Особенно это касается так называемых «пограничных случаев», которые могут относиться или не относиться к сформулированной проблеме.

Например, французская энвиронменталистская партия “Generation Ecologie” не является «зеленой» партией в общеразделяемом смысле этого слова. Другим примером служит Цыганский вопрос, имеет свои особенности на территории Венгрии, Франции, Чехии и т.д., более того – с трудом сопоставим с традиционными «конфликтами между сообществами» (в силу отсутствия четких территориальных контуров, размытых границ сообщества и пр.). В этом и проявляется особенность качественной стратегии: *сравнению случаев всегда предшествует получение информации о каждом изучаемом случае*.

Дж. Сирайт и Дж. Герринг, обобщая опыт других исследователей (Милль, Экштейн, Лийп харт, Пржеворски и Тьюн) выделяют несколько техник выбора случаев, которые «работаю т» для проведения сравнительных кейс-стади (см. параграф о кейс-стади), а также констру ирования более широких выборок.

Выбор типичных случаев (Typical Cases).

«Типичный случай» представляет собой кейс, демонстрирующий стабильно схожие (типи чные) для данной популяции отношения.

«Типичность» понимать различным образом. В первом значении «типичность» - это совок упность характеристик, которые должны быть присущи определенным объектам («типичн ый средневековый город», «среднестатистическая семья из двух человек с одним ребенко м и доходом на уровне среднего», «типичная компания по производству фаст-фуда»). В да нном значении типичный случай является синонимом понятия «репрезентативный случай », а целью исследования является поиск условий существования искомого объекта. Он пре дназначен для иллюстрации устойчивых черт некоторых общезначимых ситуаций и проце ссов. Наблюдения, полученные на материале такого случая, представляют ценную инфор мацию о «средних» индивидах, институтах, организациях.

В количественном исследовании для определения типичного случая можно воспользовать ся дескриптивными статистиками (например, модой или медианой). В качественном иссле довании поиск «типичного» обычно происходит через определение набора фоновых харак теристик, представленных в терминах «наличия» или «отсутствия». Например, Роберт и Х елен Линд ставят задачу найти «город, представляющий современную американскую жиз нь настолько репрезентативно, насколько это возможно». Они формулируют несколько кр итериев [Lynd]:

- умеренный климат;
- высокие темпы роста;
- индустриальная культура и машинное производство;
- разнообразие индустрий;
- насыщенная культурная жизнь;
- отсутствие «выдающихся особенностей» и острых социальных проблем

Пример из книги *"Middletown"*показывает, каким образом можно выделять типичные слу чаи при помощи качественного анализа. Линд не использовали статистических методов – проверка условий проводилась исходя из собственных знаний авторов. Так они пришли к выводу, что город Манси, штат Индиана, подходит для изучения среднеамериканского об раза жизни. Такой подход характерен для качественного исследования – авторы сформули ровали достаточно широкие критерии выбора случая, а затем приступили к полевой работ е.

Стоит, однако, отметить, что выделение объектов исследования по качественным критери ям требует возврата к вопросу об их «типичности» по мере появления новых исследовател ьских вопросов. Например, уже в ходе исследования Линд поставили вопрос об уровне до верия представителей различных классов друг к другу. Как мы понимаем, на данное следс твие влияет множество условий, лишь небольшая часть которых была включена в первона чальные критерии отбора «типичного случая». Поэтому вопрос о том, насколько классовы е отношения в Манси репрезентируют ситуацию в средних городах США того времени, ос тается открытым.

Второе значение слова «типичность» относится уже не столько к случаю как таковому, но к тем связям, в которые этот случай включен. Такая трактовка привязана к исследовательс кой гипотезе, которую исследователь предварительно сформулировал и теперь намерен пр оверить. Исходя из этого понимания типичный случай – это кейс, который по своим харак теристикам хорошо «вписывается» в исследовательскую гипотезу. Поскольку типичный с лучай хорошо объясняется существующей моделью (или как минимум представляет ее), з агадки, представляющие интерес для исследователя, лежат в самой природе этого случае. В частности, исследователь ищет типичные случаи того или иного явления, чтобы лучше изучить общие причинно-следственные механизмы для большинства случаев из своей поп уляции.

Обычным использованием стратегии выбора типичных случаев являются поддержка или р азъяснение существующей причинно-следственной гипотезы. Тогда отбору случаев должн а предшествовать идентификация релевантных переменных / условий.

Изучение гипотетических причинно-следственных связей может привести к нескольким р азным результатам. Если налицо расхождение между гипотезой и тем, что мы наблюдаем при исследовании типичного случая, исследователь может заявить, что каузальные механ измы не соответствуют определенным ранее, или вообще не существует никаких правдоп одобных механизмов сочетания независимой переменной с данным конкретным результат ом. При подобном результате исследование типичного случая может способствовать *фаль сификации гипотезы*.

Каким образом можно выделять типичные случаи при помощи количественных методов? Здесь нам поможет индуктивная статистика и любая программа для ее анализа, например, **SPSS**.

Для демонстрации воспользуемся простым примером. Предположим, что нам необходимо проверить гипотезу о связи общественного богатства и уровня развития демократических институтов. Примем (для наглядности), что эта зависимость линейная, и других факторов, которые могли бы помешать определению причинно-следственных связей, здесь нет (разу меется, на самом деле это не так).

Для количественного измерения необходимо представить исчислимые эквиваленты этих п онятий, например, ВВП по паритету покупательной способности на душу населения и инд екс развития демократии. Возьмем показатели двух баз данных по 2010 году:

- Penn World Table – для определения ВВП (показатель PPP converted GDP per capita) ,
- Polity IV – для измерения уровня развития демократии по разным странам (показат ель polity2).

В результате объединения мы получаем выборку (n=144). Рассмотрим диаграмму рассеян ия (в SPSS это можно сделать)

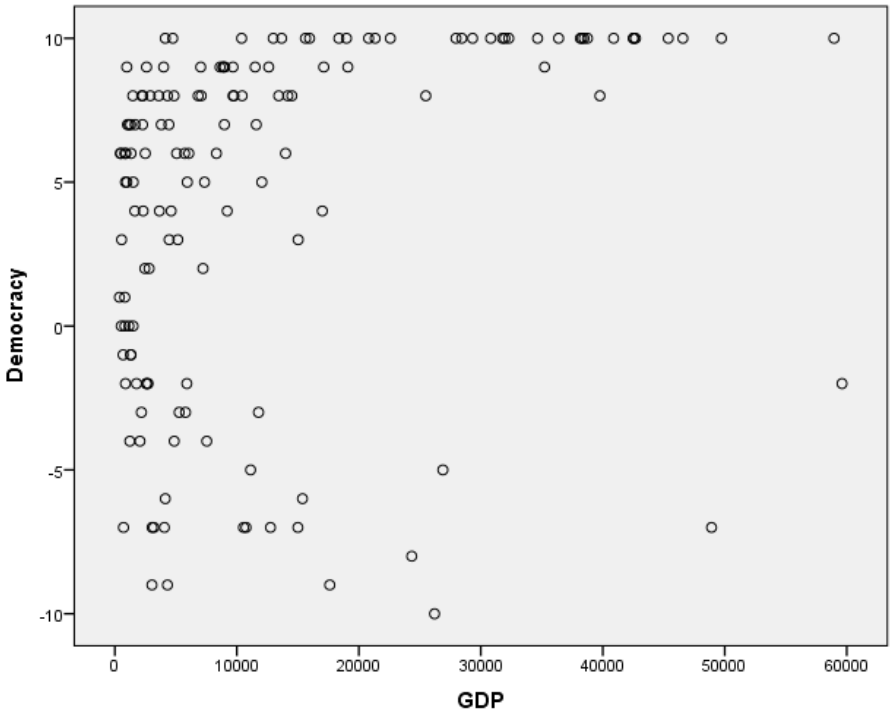


Рисунок 1.5. Диаграмма рассевания по ВВП и демократическому развитию.

Диаграмма показывает, что в целом гипотеза о связи ВВП на душу населения и «уровня де мократии» подтверждается. Например, мы видим, что богатые страны данной выборки пр актически целиком являются демократическими. При этом, однако, существует большое к оличество стран, которые демонстрируют высокий уровень демократии и низкий уровень ВВП, что указывает на недостаточную проработанность модели.

Рассмотрим теперь, как из всего многообразия случаев можно выбирать необходимые для конкретных исследовательских целей. Начнем с выбора типичного случая.

Поскольку речь идет о количественном исследовании, опирающемся на статистические ме тоды, для определения места случаев в общей картине сведения причины и следствия мы вновь можем обратиться к уже знакомому регрессионному анализу. Напомним, что уравне ние регрессии описывается следующим образом:

$$E(y_i) = b_0 + b_1x_{1,I} + \ldots + b_kx_{k,i}$$

Что же будет означать «типичность» при создании регрессионной модели? Очевидно, это будут случаи, располагающиеся как можно ближе к линии регрессии. Тогда идентифицир овать типичные случаи в общей популяции можно через **регрессионные остатки (**разниц у между наблюдаемыми значениями объясняемой переменной и предсказанными).

Регрессионные остатки можно оценить визуально, если при проведении линейной регресс ии **(Анализ → Регрессия → Линейная → Сохранить… → Стандартизированные оста**

тки). После этого можно построить гистограмму стандартизированных остатков и выбрать те из них, которые максимально близки к нулю.

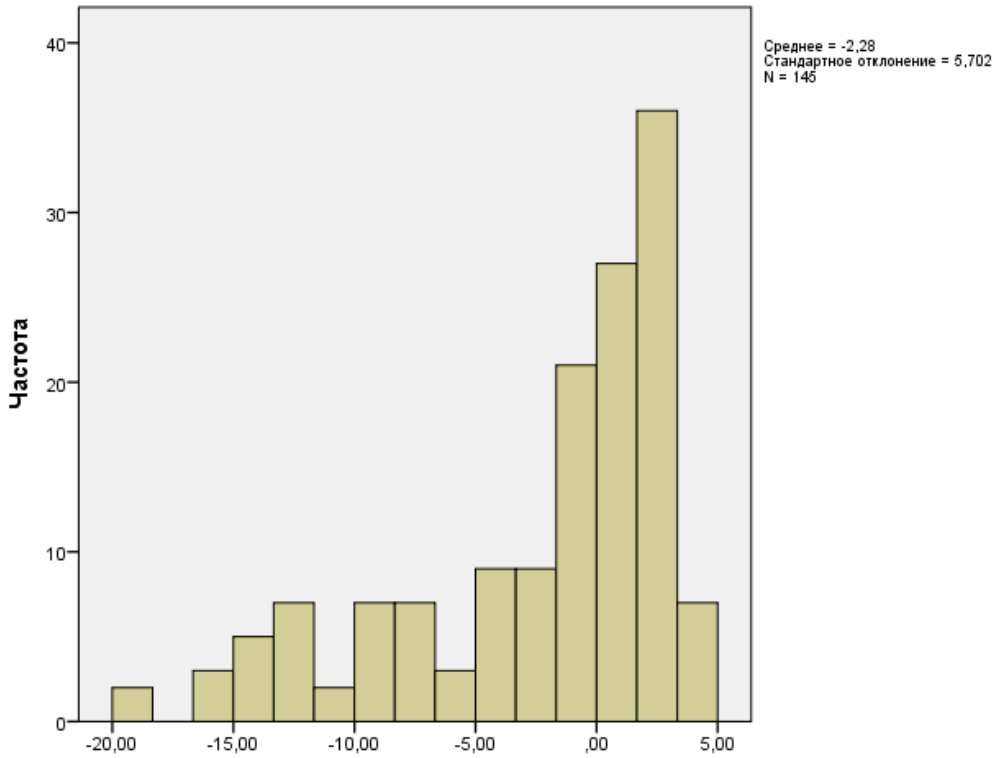


Рисунок 1.6. Определение типичных случаев с помощью диаграммы остатков.

Выбор разнообразных случаев (Diverse cases)

Выбор разнообразных случаев.

В отличие от первой стратегии, выбор разнообразных случаев направлен на получение максимально большого количества информации о самых разных случаях. Так, в примере для регрессионного анализа, исследователь может прибегнуть к сравнению протестантских, католических исламских стран – это будут разнообразные случаи. Для непрерывных данных, где нет очевидного разделения на категории, (человеческий рост, урожайность кукурузы, температура воздуха) выбор не столь очевиден. В таких ситуациях обычно рекомендуется выбирать случаи с высокими и низкими значениями, а также, при необходимости, медиану.

Если мы работаем с теорией, включающей несколько условий, удобнее всего выбирать случаи с различными сочетаниями условий. Например, если гипотеза включает три условия (пол, семейное положение, наличие детей), можно отбирать случаи, включающие в себя различные сочетания этих условий.

Случай	Условия

	Пол	Семейное положение	Наличие детей
1	мужской	женат	есть
2	мужской	женат	нет
3	мужской	не женат	нет
4	мужской	не женат	есть
5	женский	замужем	есть
6	женский	замужем	нет
7	женский	не замужем	нет
8	женский	не замужем	есть

Теперь исследователь может сравнить восемь разнообразных случаев. Если переменные я вляются интервальными, их можно попытаться дихотомизировать.

Выделение разнообразных случаев из большой популяции не представляет особых трудно стей при использовании случайной стратифицированной выборки. Стратификация нужна для достижения внутренней *гомогенности* каждой подгруппы: не следует выбирать случа и, которые являются атипичными и не представляют свое подмножество – это затруднит п роцесс создания обобщений. В данном случае процесс выбора разнообразных случаев вкл ючает в себя и предварительное определение типичных случаев.

В исследовании глобализации и государств всеобщего благосостояния Дюан Свонк приме няет именно такую методику. Сначала он определяет три типа систем благостостояния: «с оциально-демократические», «консервативно-корпоративистские» и «либеральные». Посл е этого внутри каждой группы исследователь выбирает типичные случаи, которые более в сего соответствуют теоретическим моделям систем, и проводит ряд монографических исс ледований.

Выбор экстремальных случаев (Extreme Case)

При использовании данной стратегии мы выбираем случаи, которые представляют собой э кстремум (наибольшее или наименьшее значение) для какого-либо условия. Например, пр и исследовании насилия в семье можно сосредоточиться на крайних случаях его проявлен ия (смерть, изнасилования и пр.). В изучении альтруистического поведения может сосредо точиться на случаях, когда кто-либо рисковал своей жизнью, чтобы помочь другим (напри мер, деятельность О. Шиндлера или Р. Валленберга во время Второй мировой войны). Исс ледования в области межэтнических отношений можно сфокусировать на анализе стран с максимальным количеством этносов (Россия, США, Индия), а изучение индустриального развития сосредоточить вокруг наиболее успешных образцов («Азиатских тигров», БРИК С, «Группы одиннадцати» и т.д.).

Если определить понятие «экстремальный» в данном контексте более точно, то речь идет о случае, которые находится далеко от предполагаемого распределения какого-либо атриб ута большей части популяции. Выражаясь простым языком, этот случай сразу покажется с оциологу «необычным». Если большинство случаев демонстрируют присутствие условия, тогда негативный случай может считаться экстремальным, и наоборот. Ценность кейса в т акой перспективе заключается не в конкретном значении, но в уникальности, которую он демонстрирует. Например, при исследовании табу на инцест более информативной может оказаться работа с культурами, в которых данного табу нет. По этой же причине при иссле довании редких макрообъектов можно сосредоточиться на наиболее «необычных» случая х. Почему Сида Скочпол исследует именно Францию, но не Австро-Венгрию? Ответ прос т: первый случай является более «необычным». Социальных революций мало, и необходи мые автору свидетельства, раскрывающие причинно-следственные связи в наиболее явно м виде, можно начинать искать именно там, с тем чтобы потом сравнить их с широкой поп уляцией случаев «не-революции».

Отметим, что «экстремальность» зависит от исследовательского вопроса. Например, при и зучении межнациональных отношений на постсоветском пространстве возможно включит ь случай «Первая война в Чечне» как экстремальный, поскольку он демонстрирует наибол ее «горячую» их форму – открытый вооруженный конфликт. В то же самое время, исследо ватель вооруженных столкновений в границах бывшего СССР, вероятно, посмотрит на че ченскую кампанию как на типичный случай и поставит ее в один ряд с Карабахом, Абхази ей и Приднестровьем (для него экстремальным может оказаться, например,).

С указанным замечанием связана проблема соотнесения случаев. Экстремальный кейс мо жет рассматриваться прототип или даже архетип: многие понятия зачастую определяются своими крайними проявлениями, которые могут в дальнейшем приниматься в качестве эта лонов. **Архетипическим случаем** в социальных науках называются кейсы, определившие категорию, в рамках которой они рассматриваются как родоначальники. Вспомним Велик ую Французскую революцию: этот эпизод изменил само понятие революции, которая тепе рь определяется как прогрессивная модернизационная сила. Таким образом, данное событ ие сделало возможным «появление» последующих революций (отнесение последующих с обытий к «революциям»). Соответственно, определение данного случая как «типичного» по отношению к широкой категории «революция» не совсем верно: здесь речь идет не про сто о прототипе («ранней версии» случая), но об архетипе (случае, определяющем понятие).

Об экстремальных случаях говорится, что они качественным образом отличаются от друг

их случаев. Тем не менее, сравнение с типичными случаями зачастую показывает, что меж ду «экстремальным» и «обычным» гораздо больше общего, чем может показаться на перв ый взгляд. Многие атрибуты революции присутствуют и в менее радикальных формах соц иальных конфликтов. Революция, как это демонстрируется в работах исторических социо логов, случается не в силу присутствия некоторых экстремальных компонентов, но благод аря *особому взаимодействию* этих компонентов – изучение таких комбинаций и представ ляет наиболее важный аргумент в пользу исследования экстремальных случаев [Ragin 198 7].

Связь экстремальности и архетипа также является удобным исследовательским инструмен том: экстремальный случай ценен тем, что демонстрирует рафинированные примеры соци альных явлений. По этой причине, например, в «Элементарных формах религиозной жизн и» Дюркгейм изучает не западное общество, а австралийских аборигенов, а Эмиль Дюмон пишет книгу *"Homo Hierarchicus"*, в которой обращается к кастовому строю Индии как чи стому образцу социальной стратификации.

Кроме того, включение экстремального случая позволяет достичь более широкой вариаци и искомого условия и, в частности, служит мощным аргументом доказательства и развити я теории. В частности, если случай, который кажется экстремальным, в действительности может быть в действительности описан выбранными категориями, это укрепит существую щую теорию.

В параграфе о систематической ошибке отбора мы говорили, что не следует строить выбо рку, опираясь исключительно на результат (зависимую переменную). Включение *только* э кстремальных случаев чревато сильным смещением и неверной оценкой веса факторов[2], п оэтому использовать экстремальные случаи следует *в постоянном соотнесении со всей по пуляцией*, а после их изучения переходить к другим методам в поисках более устойчивых обобщений.

В статистическом смысле значение переменной для экстремального случая располагается далеко от среднего по выборке. При работе с большими популяциями экстремальность (E) для случая *i*может быть определена в терминах среднего значения и стандартного отклоне ния по переменной:

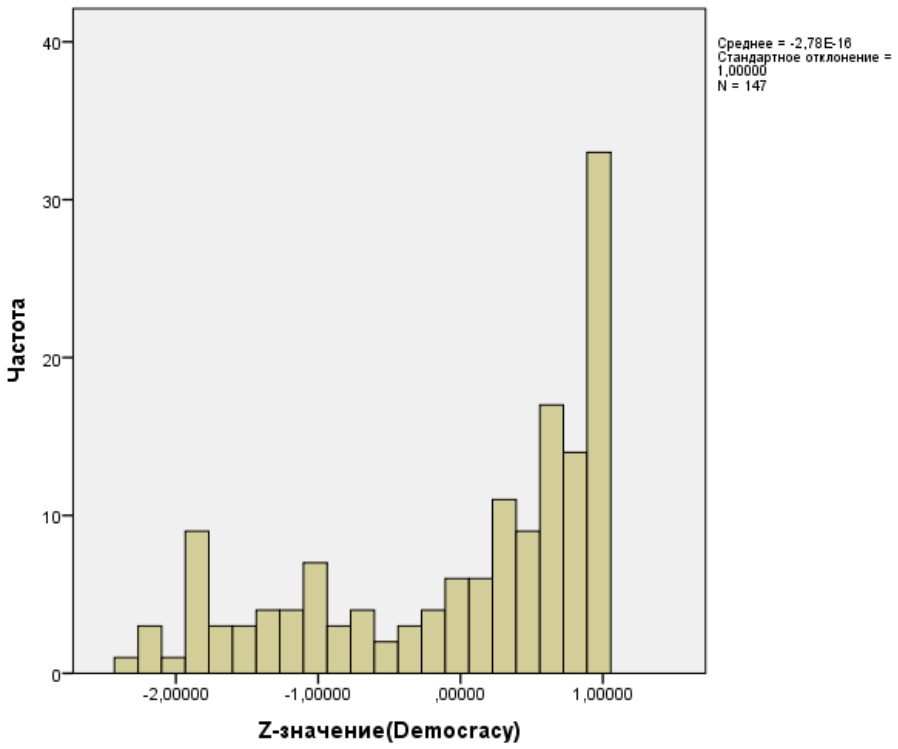
$$E_i = \left| \frac{X_i - \bar{X}}{s} \right|$$

[2] Об этом говорилось в параграфе об ошибках отбора случаев.

где X_i - значение переменной для случая i, $\bar{X}$ – среднее значение, s – стандартное отклонен ие.

Мы получаем стандартный (z) показатель. Случаи с высоким значением *E* можно признать экстремальными. Зачастую определить уровень *E*, необходимый для отнесения случая к эк стремальным, не так просто. Тем не менее, здесь существуют некоторые рекомендации. В соответствии со статистической традицией, случаи с показателем E≤2 не следует признава ть экстремальными, даже если это максимальные значения в выборке. Для достижения ма ксимальной достоверности можно выставить порог на 3.

В примере с демократией и национальным богатством среднее для переменной «демократ ия» равно 4,14, стандартное отклонение (s) = 6,002. Если построить гистограмму «экстрем альности» (**Анализ→Описательные статистики→Сохранить стандартизированные зн ачения в переменных→Графика→Гистограмма**), она примет следующий вид:



Максимальное значение, которое принимает *E*, равно -2,4. Соответственно, слишком экст ремальных случаев в данной выборке не наблюдается. Поскольку среднее значение показа теля равно 4,14, страны выборки можно назвать «условно демократичными». Четыре стра ны демонстрируют наименьшие результаты по z- показателю (Оман, Свазиленд, Узбекист ан, Туркменистан, Саудовская Аравия). Саудовская Аравия имеет наиболее высокое значе ние *E* и обладают показателем «демократичности» -10. Изучение этого случая прольет све т на то, при каких условиях богатая страна может не являться демократичной. Определив новые переменные, мы сможем проверить их на следующем наборе случаев.

Выбор отклоняющихся случаев (Deviant Cases)

При использовании данной стратегии мы выбираем случаи, которые демонстрируют неож иданное значение. Поскольку речь идет об исследовании эмпирических аномалий, в качес тве альтернативного названия стратегии можно предложить «выбор аномальных (девиант ных) случаев».

Классическим примером сравнительного исследования с включением аномального случая можно назвать работы, посвященные отсутствию в США социалистических настроений. Г ипотезы, сформулированные на материале европейских стран, показывали, что в Соедине нных Штатах присутствуют все необходимые компоненты для изменения общественно-по литических установок, однако страна настойчиво придерживалась идей капиталистическо го развития. Внимательно изучив отклоняющийся случай, Вернер Зомбарт в книге «Почем у в Соединенных Штатах нет социализма?» предположил, что данный строй со временем обретет поддержку и в Америке [Zombart]. Когда время показало, что гипотеза не оправда ла себя, Джон Ласлет и Сеймур Липсет в сборнике *«Падение мечты»* вновь обратились к аномалии и на этот раз попытались объяснить недостатки теории немецкого социолога [La slet and Lipset]. Таким образом, использование отклоняющегося случая применялось, во-п ервых, для подтверждения каузального аргумента, а во-вторых, для дальнейшего развития теории.

Цель стратегии выбора девиантных случаев, как правило, состоит в апробации новых - по ка еще не до конца определенных, - объяснений. Исследователь надеется, что каузальные процессы в рамках девиантного случая помогут выявить некоторые факторы, которые ока жутся применимы к другим (девиантным) случаям. Это означает, что в большинстве случа ев исследование девиации завершается конструированием общей гипотезы, которую можн о экстраполировать на другие случаи. Как следствие, исследование девиантного случая мо жет привести к построению новой модели, которая определит уже совершенно другой наб ор девиантных случаев. Отбирая отклоняющиеся случаи, уместно задавать вопрос: *«По о тношению к какой общей модели данный кейс демонстрирует отклонения?»*

Отметим: в то время как выбор экстремальных случаев осуществляется на основе распред еления значений определенного атрибута, девиантные случаи отбираются *исходя из некот орой модели причинно-следственных связей*. Наличие девиации присутствует только в рам ках целостной модели. Это означает, что с изменением теоретической модели представлен ие о девиации изменится, и нам придется подбирать иные случаи. Например, США предст авляет собой девиантный случай государства всеобщего благосостояния, когда мы связыв аем последнее с уровнем общественного богатства. Но если сместить акцент на другие по

литические и социальные факторы (как это сделали, например, Алесина и Э. Гэйзер [2004]), от девиации не останется и следа.

Каким образом можно определить отклоняющиеся случаи при работе с большой генераль ной совокупностью? Девиантный случай представляет собой противоположность случаю типичному: в то время как типичный случай максимально близко репрезентирует теорети ческую модель, отклоняющийся случай в наибольшей степени ей не соответствует. Возвра щаясь к уравнению, предложенному для типичного случая, и просто поменяем знак:

$$\text{Deviantness}(i) = \text{abs}[y_i - E(y_i \mid x_{1,i,} \ldots x_{k,i})] = \text{abs}[y_i - b_0 + b_{1x1,i} + \ldots + b_K x_{K,i}]$$

Уровень отклонения варьируется от 0 (для случаев, располагающихся вдоль линии регресс ии) до бесконечности. Исследователь, разумеется, наиболее заинтересован в случаях, демо нстрирующих самое высокое значение отклонений.

В сквозном примере наиболее отклоняющиеся случаи располагаются ниже линии регрессии.

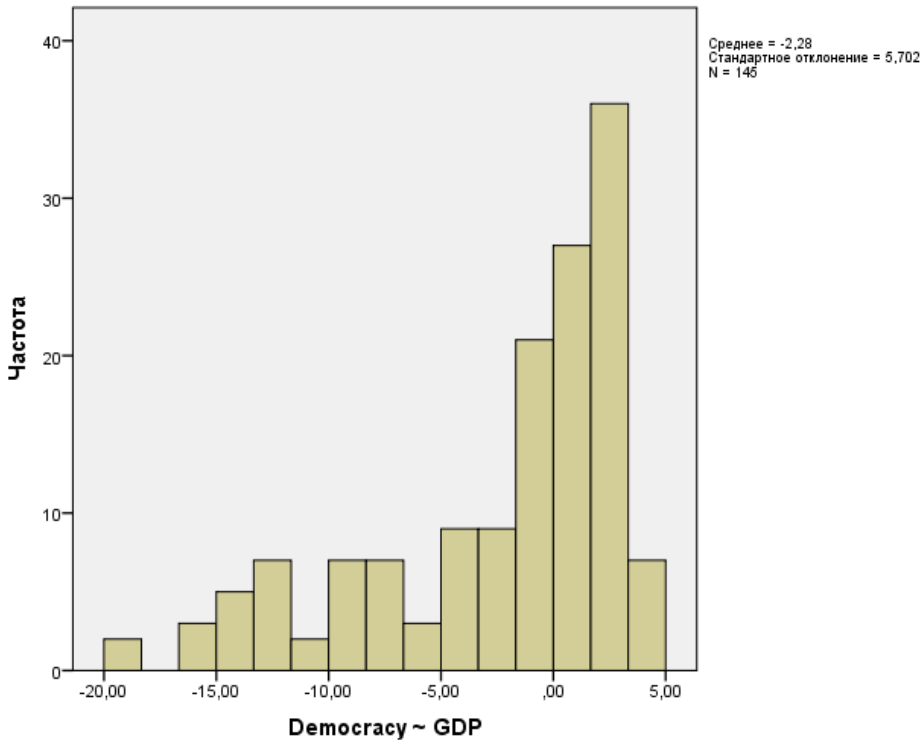


Рисунок 1.7. Идентификация отклоняющихся случаев.

Девятнадцать случаев имеют показатель отклонения, уходящий за-10. Для каждого и этих кейсов можно сформулировать новый исследовательский вопрос, например, такой: «Что, помимо экономических показателей, определяет социально-политическую ситуацию в эти х странах?» Возможность постановки подобных вопросов и определяет важность включен ия отклоняющихся случаев в исследовательскую практику.

В нашем примере наиболее отклоняющимися от тестируемой гипотезы случаями выступа ют:

Страна	Уровень ВВП	Уровень демократии
Кувейт	48915	-7
Саудовская Аравия	26205	-10
Турменистан	17619	-9
Оман	24337	-8

Таблица 1.23. Результат определения отклоняющихся случаев.

Выбор похожих случаев

Выбор похожих случаев предполагает исследование сходных объектов, при рассмотрении различий между которыми общие для них признаки принимаются за константы. Тогда ост авшиеся факторы, которые отличают эти случаи, и будут рассматриваться как теоретическ и важные. Хотя на практике мы редко можем оставить только один (решающий) фактор, э та стратегия позволяет, по крайней мере, исключить некоторое число других и сосредоточ иться на качественном анализе оставшихся.

В наиболее простом виде сущность метода наиболее похожих систем можно выразить сле дующим образом:

Здесь представлены два случая, сравниваемые по трем дихотомизированным переменным. Показатели переменной X_2 идентичны, в то время как значения X_1 различаются – как и рез ультат. Предполагается, что присутствие или отсутствие этого условия является причиной наступления результата.

Случаи	Переменные		
	X_1	X_2	X_3
I	+	+	+
II	-	+	-

Выбор сходных случаев происходит следующим образом:

1. определение соответствующего универсума случаев;
1. определение ключевых переменных, представляющих интерес, значения которых являются аналогичными для всех случаев;

2. определение переменной или переменных, которые гипотетически сильно варьируются от случая к случаю;

3. выбор нужного количества случаев (часто двух, но иногда и больше), которые демонстрируют указанные сходства и различия.

Часто аналитик начинает этот процесс, уже имея один из случаев «в голове». Тогда исслед ователь проделывает описанные выше действия, но уже с прицелом на поиск других случа ев, которые похожи на первый.

Примером может служить уже упоминавшаяся работа С. Скочпол «Государство и социальная революция» (“State and Social Revolution”). Основными объектами сравнения в ней выступают Россия, Франция и Китай. Во всех этих случаях (системах) произошли социальные революции, несмотря на наличие очень разных политико-экономических и социальных условий. Соответственно, Скочпол задается вопросом: «Какие общие особенности этих систем привели их к похожим политическим событиям?».

Метод мэтчингадля определения похожих случаев

Каким образом можно выделить наиболее похожие случаи из имеющейся в наличии совокупности? Наиболее полезной статистической методикой здесь выступает мэтчинг (matching), или поиск соответствий.

Мэтчинг (поиск соответствий) представляет собой метод целенаправленного отбора с целью нахождения сопоставимых единиц в наборе данных.

В социальном мире тяжело найти ситуации, при которых случаи с высокими показателями независимой переменной демонстрируют абсолютную схожесть с теми случаями, где наблюдаются низкие показатели независимой переменной относительно «контрольной группы».

Например, мы исследуем влияние гражданского общества на рост заработной платы. Логичным действием исследователя будет сравнить зарплаты в странах с большим и малым уровнем развития гражданского общества. Проблема возникает вследствие того, что в демократических странах (где гражданское общество должно быть развито) уровень общественного богатства в среднем выше, чем в недемократических – сквозной пример это подтверждает. Рост благосостояния сказывается и на изменении форм экономического распределения, поэтому экономические системы в странах с разным уровнем гражданского общества отличаются – это создает трудности при определении ключевых переменных.

Распространенным подходом в таких случаях является введение дополнительных переменных для каждого искажающего фактора в общем анализе причинно-следственных связей[3]. Однако можно поступить иначе. Альтернативный подход состоит в том, чтобы сначала определить набор переменных (помимо уже выделенных зависимой и независимой), по которым мы сопоставим случаи - ковариаты. Затем для каждого случая из экспериментальной группы исследователь подбирает контрольные случаи с таким же результатом ковариат. Наконец, ученый смотрит на разницу между значениями зависимой переменной между сопоставленными случаями в контрольной и экспериментальной группах. В отличие от регрессионного анализа, который направлен на контроль переменной через включение ее в регрессионную модель, процедура мэтчинга позволяет сбалансировать выборки за счет включение в сравнение объектов с одинаковыми или похожими значениями переменных.

Задача мэтчинга состоит в том, чтобы обеспечить максимальное соответствие фоновых характеристик сравниваемых случаев друг другу.

При условии включения достаточного количества ковариат сравнение сопоставленных случаев может помочь в определении причинно-следственных связей. Даже если в процессе мэтчинга мы задействуем не все посторонние факторы, его результаты помогают создавать более надежные каузальные обобщения, чем при использовании регрессионного анализа, поскольку случаи, подходящие друг другу по выделенным переменным, вероятно, будут схожи и в отношении невыделенных факторов.

Данная методика называется **поиск точных соответствий** (exact matching). К сожалению, в большинстве случаев ее применение оказывается невозможным. Таким образом нельзя сопоставлять непрерывные переменные, поскольку в мире, вероятно, не существует двух объектов с абсолютно одинаковыми показателями. Например, мы не сможем подобрать «демократическую» и «недемократическую» страны с одинаковыми значениями ВВП. По мере увеличения количества возможных условий или переменных шанс встретить одинаковые объекты стремится к нулю, что делает применение поиска точных сочетаний невозможным в статистическом анализе, ограничивая сферу его применения малыми выборками.

В ситуации невозможности точного сопоставления исследователи прибегают к другим процедурам. Если ковариаты являются непрерывными, их можно попытаться дихотомизировать или разделить на отрезки в соответствии с теоретическими или

[3] Ковариационный анализ будет представлен в четвертой главе.

прагматическими соображениями. Такой метод называется **поиск приближенно-точных соответствий** (Coarsened Exact Matching).

Процедура начинается с определения укрупненных значений для каждой переменной. Происходит это на основе теоретических, личных знаний или здравого смысла. К примеру, если в качестве ковариаты выступает уровень образования, мы можем свести значения согласно ступеням образования: начальное, среднее, высшее. Такой подход позволит нам выбрать в качестве соответствующих друг другу студента-первокурсника и старшекурсника, но совершенно точно отделит обоих от выпускника школы.

В данном случае сравнение происходит не на уровне значений, а на уровне страт. Единицы анализа объявляются похожими, если они попадают в одну группу. Случаи, для которых не находится пары в «контрольной» или «экспериментальной» группах, исключаются из исследования.

Вернемся к нашему примеру и попытаемся найти наиболее похожие случаи для кейса «Россия». Допустим, мы продолжаем проверку гипотезы о связи уровня ВВП на душу населения по паритету покупательной способности и развития демократии. Чтобы яснее показать эту причинно-следственную связь, мы должны изолировать остальные переменные (ковариаты), которые также могут оказать влияние на результат (уровень демократии). Как это сделать? Разумным представляется выбрать случаи таким образом, чтобы значения всех переменных, кроме нашей искомой (ВВП), были как можно более похожи – тогда мы сможем указать, что варьируется только независимая переменная, а все остальные являются константами.

Добавим к нашей несложной модели несколько ковариат. Допустим, на развитие демократии могут влиять также:

- уровень безработицы (источником здесь может выступить **WorldCIAFactBook** - UnemploymentRate);
- политическая ориентация правящей партии (**Database of Political Institutions: Changes and Variable Definitions** (World Bank) – переменная **execrlc**).

Уровень безработицы, равно как и ВВП, являются непрерывными переменными, поэтому поиск точных соответствий в данном случае вряд ли возможен. Воспользуемся поиском приближенных соответствий.

В первую очередь нам необходимо разделить «контрольную» и «экспериментальную» группы. Для этого придется дихотомизировать независимую переменную, чтобы разбить все случаи на две группы. Группа с низким ВВП на душу населения будет отнесена к «контрольной», а с «высоким» - в экспериментальную – так обеспечивается необходимая вариация признака.

Примем следующее условие:

- если ВВП меньше $7000 на душу, его относим к «контрольной» группе;
- если ВВП больше $7000 – к «экспериментальной».

Таким образом, мы перекодируем переменную GDP (**Преобразование – перекодировать**

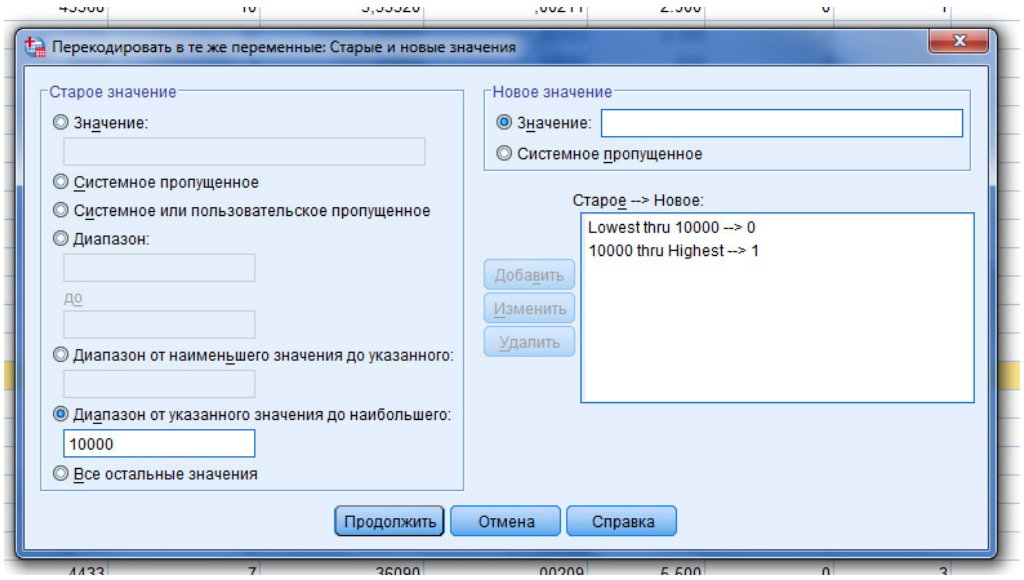


Рисунок 1.8. Перекодирование переменной в бинарную.

в те же переменные).

Не забудьте определить независимую переменную как «номинальную» (вкладка «**Переменные**», столбец «**Шкала**»).

Таким же образом следует поступить и с переменными ковариатами. В нашем случае переменная «политическая ориентация» является номинальной и перекодируется следующим образом:

- 1 – условно «правая» ориентация;
- 2 – условно «центристы»;
- 3 – условно «левая» ориентация.

Переменная «уровень безработицы» для простоты перекодируется по такому правилу:

- Уровень безработицы меньше 10% - 1;
- Уровень безработицы больше 10% - 0.

Подобную кодировку следует читать так: присутствует ли в данном случае (стране) высокий уровень безработицы.

Следует отметить, что исследователь волен выделять любое число градаций переменной для уточнения модели. Наш пример представляет лишь грубую модель.

Когда модель определена, можно приступать к процедуре мэтчинга[4]. Анализ→Coarsened Exact Matching. Здесь нам предлагают определить группирующую переменную и ковариаты.

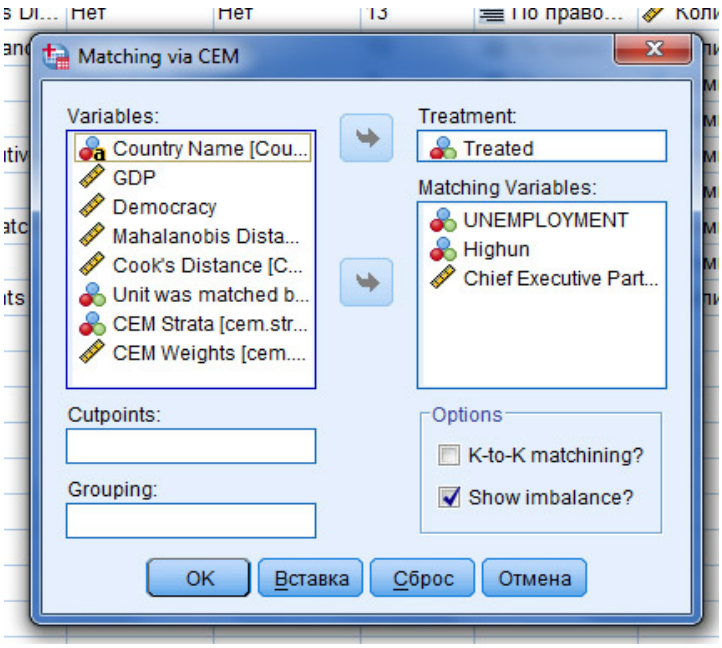


Рисунок 1.9. Проведение поиска приближенных соответствий с помощью

В **Выводе** мы получаем такую картину:

Matching Summary

[4] На момент выпуска учебника в SPSS не предусмотрена функция мэтчинга. Однако существуют специальные расширения, позволяющие внедрять некоторые функции других языков в SPSS. Одним из них является созданная Стефано Лакусом, Гарри Кингом и Джузеппе Порро программа CEM. Данный пакет включает в себя компоненты R (R Essentials for SPSS) и Python (Python Essentials for SPSS). Преимуществами расширения являются простота установки и наличие всех необходимых компонентов, что устраняет проблему совместимости различных версий программных оболочек SPSS, R и Python.
Скачать CEM можно по адресу: http://projects.iq.harvard.edu/cem-spss

	G0	G1
All	71	75
Matched	66	74
Unmatched	5	1

Рисунок 1.10.Результаты поиска приближенных соответствий.

В результате дихотомизации независимой переменной «ВВП на душу населения» 71 случай оказалось возможным отнести к «контрольной» группе и 75 – к «экспериментальной». После реализации мэтчинга приближенно-точных соответствий для 60 случаев из «контрольной» группы нашлась пара в «экспериментальной».

Россия попала в «экспериментальную группу» (ВВП на душу населения = $17005). Для дальнейшего изучения связи общественного богатства и развития демократии можно предложить к сравнению следующие случаи:

Страна	ВВП	Уровень демократии
Армения	5936	5
ЦАР	675	-1
Киргизстан	2318	4
Папуа Новая Гвинея	3650	4
Марокко	4134	6
Филиппины	3596	8
Фиджи	4864	-4

Таблица 1.24. Страны, похожие на случай России согласно выбранным фоновым факторам.

Как мы понимаем, точность мэтчинга напрямую зависит от количества ковариат в модели и уровня детализации значений по каждой из них: чем подробнее модель, тем меньший объем страт окажется на выходе. Кроме того, сопоставляться таким образом могут не только макрообъекты, но и любые случаи (например, поиск приближенных соответствий активно используется в психологических исследованиях).

Еще одной, более сложной, методикой мэтчинга является **поиск соответствий по предрасположенности** (propensity score matching). Эта методика оперирует иным пониманием схожести: вместо прямого сравнения значений ковариат исследователь определяет *вероятность* отнесения случаев к первой группе. Таким образом, в поисках пары для случая из одной группы исследователь пытается найти в контрольной группе случаи, которые также могли бы попасть в экспериментальную.

Предрасположенность определяется как *вероятность отнесения случая к «экспериментальной группе» при заданных значениях ковариат*:

$$e(x)=P(Z=1|X)$$

где $e(x)$ – propensityscore, P – вероятность, Z – индикатор отнесения к экспериментальной группе (1 – отнесен, 0 - нет), X–набор ковариат.

Иными словами, предрасположенность показывает, насколько вероятно, что случай (человек, страна и пр.) окажется отнесенным к «экспериментальной» группе при данных фоновых характеристиках.

Техника поиска соответствий состоит из двух этапов. На первом этапе ученый превращает независимую переменную из гипотезы в зависимую, а сопоставляемые переменные – в независимые. На втором этапе каждому случаю даются предсказанные значения, которые показывают вероятность его попадания в «экспериментальную» группу. После этого исследователь может выбрать из «контрольной» группы случаи, имеющие схожие значения со случаями из «экспериментальной» группы.

Шаги выполнения поиска соответствий по предрасположенности:

1. Исследователь отбирает ковариаты, которые могут оказывать влияние на результат (зависимую переменную). Как и всегда, данный шаг очень важен, поскольку надежность propensity score напрямую зависит от заданных ковариат. Рекомендуется не ограничиваться общеизвестными переменными (пол, возраст, доход и пр.), но включить как можно большее количество искажающих факторов.

2. Оценивается предрасположенность случаев. Обычно для этого используется логистическая регрессия, когда дихотомизированная переменная «контрольная-экспериментальная группа» используется в качестве зависимой переменной, а ковариаты – предикторов (независимых переменных).

3. После оценки предрасположенности происходит сам мэтчинг. Наиболее часто для этого применяется попарное сопоставление по методу ближайшего соседства (nearest neughbor): одиночные наблюдения из «экспериментальной» группы сопоставляются с наблюдениями из «контрольной» по схожести значений предрасположенности. Если размеры «контрольной» и «экспериментальной» групп значительно отличаются друг от друга, можно проводить последовательный мэтчинг, когда для одного случая подбирается несколько соответствий. Чтобы избежать «плохих» соответствий, исследователь может назначить **калипер** – максимальную разницу (дистанцию) между показателями предрасположенности.

Установка малого калипера (небольшая разница между двумя значениями) дает более надежные результаты, однако сокращает количество совпадений.

Перед тем, как приступить к мэтчингу, необходимо разобраться с тем, что из себя представляет **логистическая регрессия.**

Кратко смысл регрессионного анализа можно свести к нахождению аналитического выражения, наиболее адекватно отражающего связь между зависимой переменной и множеством независимых переменных. Главным отличием логистической регрессии от множественной является толкование уравнения регрессии. Если множественная регрессия позволяет прогнозировать количественное значение зависимой переменной (критерия) на основе известных значений независимых переменных (предикторов), то логистическая регрессия прогнозирует *вероятность* некоторого события, находящуюся в пределах от 0 до 1. Кроме того, при помощи индикаторной схемы кодирования допускается использование в качестве предикторов категориальных (номинативных) переменных. Более детально особенности логистической регрессии рассмотрены в разделе «Представление результатов», сейчас же достаточно знать, что категориальный предиктор может быть представлен серией бинарных переменных — по одной на каждую категорию предиктора. Этим бинарным переменным присваивается значение 1 или 0 в зависимости от того, к какой категории относится объект.

С логистической регрессией связаны такие математические понятия, как вероятность, шанс и натуральный логарифм шанса. *Вероятность* — это ожидаемая относительная частота некоторого события. *Шанс* представляет собой отношение вероятности того, что событие произойдет, к вероятности того, что событие не произойдет. Так, если вероятность дождя равна 0,2, то вероятность отсутствия дождя равна 0,8, следовательно, шанс, что дождь все-таки прольется, равен 0,2/0,8 = 0,25. Обратите внимание на то, что шанс, в отличие от вероятности, не ограничен максимальным единичным значением: если, к примеру, вероятность дождя составляет не 0,2, а 0,8, то получаем шанс 0,8/0,2 = 4.

Единичное значение шанса соответствует ситуации, когда вероятности события и его отсутствия равны.

Непосредственно включить в регрессионную модель дихотомическую переменную нельзя. Однако это можно сделать, если вместо нее использовать некоторую производную функцию — логит (logit).Логит равен натуральному логарифму шанса. Например, логит вероятности 20 % равен -1,386.

Отношение вероятности того, что событие произойдет, к вероятности того, что оно не произойдет P / (1 - P), называется отношением шансов (или отношением предпочтения). Модель логистической регрессии определяется формулой:

$$Z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

Посмотрим, как это выглядит на практике. Мы продолжаем исследование влияния общественного богатства на уровень демократичности страны, и нам необходимо выбрать случаи, которые, в соответствии с принципом наиболее похожих систем, были максимально похожи по всем параметрам, кроме двух – ВВП (независимая переменная) и уровня демократии (зависимая переменная). Допустим также, что у нас есть экспериментальный случай – Россия, и наша задача – вновь подобрать для него контрольные случаи.

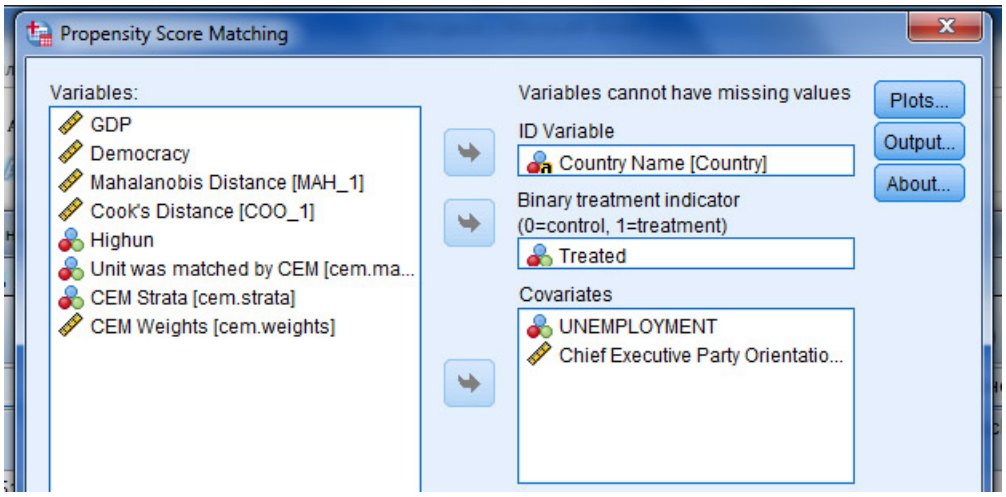


Рисунок 1.11. Спецификация поиска по предрасположенности.

Преимуществом поиска соответствий по предрасположенности является возможность работы с «сырыми» данными – ведь нас интересует не точные совпадения, а вероятность попадания в «экспериментальную» группу. Для нашего примера мы воспользуемся теми же ковариатами, однако на этот раз не будем их категоризировать. Исключение составляет переменная, отвечающая за причисление случая к «контрольной» или «экспериментальной» группам: ее необходимо дихотомизировать (таково правило логистической регрессии)[5].

[5] В демонстрационном примере используется расширение под названием Propensity Score Matching For SPSS. Приложение создано Феликсом Томмесом и представляет собой адаптацию пакета MatchIt для языка R.

Запускаем **PS Matching**. Перед нами появляется окно, в котором необходимо выбрать переменные. В примере это выглядит следующим образом:

Установим калипер (caliper), равный 0,2 – это повысит точность соотнесения. В результате мы получаем новый набор данных с переменными:

- ps – показатель предрасположенности для случая;
- psweight – вес случая.

В **Выводе** появляется следующий результат:

Sample Sizes

	Control	Treated
All	73	75
Matched	57	57
Unmatched	16	18
Discarded	0	0

Таблица 1.25. Результаты поиска по предрасположенности.

Для 16 случаев из «контрольной» группы и 18 случаев из «экспериментальной» не нашлось пары (использовался поиск 1:1). В результате получилась новая выборка из 114 случаев. Результат по России в ней – 0,597. Близкими можно признать результаты Таджикистана (0,587), Лаоса (0,587), Монголии (0,606). Таким образом, если бы мы отбирали случаи согласно принципу наиболее похожих систем, основываясь на этих двух весьма сомнительных ковариатах, наш выбор мог бы быть следующим:

Страна	ВВП	Уровень демократии
Россия	17005	4
Лаос	3048	-7
Монголия	4747	10
Таджикистан	2175	-3

Таблица 1.26. Похожие на Россию случаи по результатам поиска по предрасположенности.

Если на компьютере уже установлен плагин CEM, R устанавливать не нужно. В противном случае следует установить оболочку R с сайта http://cran.r-project.org/ (версия 2.12.0 для SPSS 20, 2.8.1 – для SPSS 19), а затем пакет SPSS R Essentials (https://www14.software.ibm.com/)
После этого необходимо скачать файл для установки настраиваемого диалогового окна (http://sourceforge.net/projects/psmspss/) и установить его в SPSS (**Сервис → Настраиваемые диалоговые окна → Установить настраиваемое диалоговое окно**). Если все завершится успешно, в меню «Анализ» появится новый пункт "**PS Matching**".

Подобный набор случаев может показаться странным и даже комичным, однако с точки зрения теоретической модели он точен: здесь реализуется правило «идти не от случаев, а от гипотезы исследования».

Качество мэтчинга напрямую зависит от качества статистической модели, на основе которой определяются **показатели предрасположенности**. Поверхностная модель (как в примере) ведет к поверхностным результатам.

Поиск соответствий по предрасположенности является спорной процедурой. Тем не менее, все больше ученых прибегает к нему в условиях невозможности построения случайной экспериментальной выборки. Обратимся к наукометрической базе данных Web of Science. В период 1995-2000 гг. термин “propensity score matching” базе журналов по общественным наукам встречается всего дважды, а на интервале 2001-2008 гг. – уже 189 раз, причем в совершенно разных исследованиях. Так, К. Грин и Энсмингер используют эту методику при проведении исследования о потреблении марихуаны афроамериканцами, Глик, Гуо и Хатчинсон с ее помощью проверяют гипотезу о связи национального регулирования потоков капиталов и устойчивости к валютным кризисам, а Митас, Альмирал и Кришман выделяют взаимоотношения между стратегиями отношений с покупателями и маркетинговой эффективностью фирм.