**Правительство Российской Федерации**

**Федеральное государственное автономное образовательное**

**учреждение** **высшего профессионального образования**

**"Национальный исследовательский университет"**

**"Высшая школа экономики"**



**Проект**

**по дисциплине**

“Количественные методы в экономике”

**Работу выполнили**

**студенты 3 курса**

Щекотихин Евгений Анатольевич

Стельмашенко Григорий Сергеевич

Нуруллин Булат Ильсурович

Москва 2021

Содержание

[**Введение**](#_124nz98n8iaz)4

[**Базовая статья**](#_uwgozxrz87ao)4

[**Предположения модели**](#_4k5ilqdqwtkd)4

[**Методология сбора данных**](#_5vlz3ol71t7h)6

[**Построение модели**](#_oohi271hkqie)7

[Черновик, R notebook 7](#_9duw3a7g3cry)

[HTML версия 7](#_iki18qligll9)

[Датасет 7](#_5evwo3d4hfs)

[Модель 1 8](#_oooq6t8q76s6)

[Модель 2 8](#_u0uwvztdh7yz)

[Модель 3 8](#_jl6izty25aw2)

[Модель 4 10](#_5bh2rfr104kq)

[Модель 5 10](#_bwx7gtqr664e)

[Модель 6](#_ipahpvml6drx) 10

[Модель 7](#_szaz9h9jk4l) 10

[**Итоги**](#_k7n0k1avms0x)18

[**Интерпретация результатов:**](#_x9wmavd9u9k0)19

[**Источники:**](#_yuwn5mmcu0ey)19

# Введение

В данной статье исследуется влияние расходов на образование на ВВП на душу населения. Обозначены основные предпосылки моделирования, приведен анализ методологии сбора данных, а также построение модели.  
За сбор данных и обоснование выбора их были ответственны: Нуруллин Булат, Евгений Щекотихин. Моделирование зависимости в рамках классической регрессионной модели было проведено Стельмашенко Григорием с использованием RStudio.

# Базовая статья

За базовую статью была взята статья про влияние расходов образования на ВВП на душу населения:

Muktdair-Al-Mukit Dewan (2012), Public Expenditure on Education and Economic Growth: The Case of Bangladesh. IJAR-BAE 1(4): p. 10-18.

# Предположения модели

1. Модель линейна по параметрам (множественная регрессия)

**2. Объясняющие переменные** ( **не являются** случайными (предполагаются детерминированными - известными, заданными)

**3. Среднее значение каждой ошибки равно нулю** (несистематичная ошибка)

**4. Гомоскедастичность** (дисперсия ошибок одинакова для всех наблюдений)

**5. Нулевая ковариация между ошибками разных наблюдений** (отсутствие связи между случайными составляющими различных наблюдений). Или же, независимость ошибок регрессии (в случае, если они нормально распределены)

**6. Нулевая ковариация между**  и (отсутствие связи между случайной составляющей и объясняющей переменной)

**7. Количество наблюдений превышает количество параметров, которые требуется определить**

**8. Остатки одинаково распределены во всех наблюдениях**

**9. Модель правильно специфицирована** (нет пропущенных или лишних переменных, а также выбрана правильная функциональная форма)

**10. Отсутствие мультиколлинеарности** (ни один объясняющий фактор модели не находится в точно аппроксимируемой линейной зависимости от другого)

**11. Каждый регрессор рассчитан в одной своей единице измерения**

# Методология сбора данных

Все показатели относятся к 2017 году. Данные для исследования были взяты из нескольких источников.  
  
 С сайта **статистического отдела ООН** (<http://data.un.org/Search.aspx?q=GDP>) были взяты значения для:

* ВВП (GDP)
* ВВП на душу населения (GDPPC)
* Государственных расходов (GOVEXP)
* Экспорта (EXPORTS)

Использовалась база Всемирного Банка (<https://data.worldbank.org/>) для таких значений как:

* Доля расходов на образование в общих государственных расходах (EDUPER)
* Доля выплаченных налогов в общем ВВП стран (TAXPER)  
  + Далее использовались для пересчета показателей расходов на образование (EDUEXP) и выплаченных налогов (TAXEXP) в долларах США
* Значение дефлятора ВВП
* Налоги (Taxes)
* Общий объем инвестиций, %ВВП (IPER)
* Валовые национальные сбережения, %ВВП (SPER)
* Инфляция, средние потребительские цены, индекс (ACP)
* Население (POP)
* Объем импорта, изменение в% (IMPCHNGPER)
* Объем экспорта, изменение в% (EXPCHNGPER)
* Инфляция, средние потребительские цены, изменение в % (ACPCHNGPER)
* Уровень безработицы, % от общей рабочей силы (UR)
* Валовая добавленная стоимость в базовых цена, в US$ (GVA)
* Прирост населения, годовой% (PG)

Данные о производительности труда (PPW) были взяты из портала Организации экономического сотрудничества и развития (<https://data.oecd.org/>)

Взяты данные только за один год для того, чтобы не было возникновения потенциальной автокорреляции и так как они перекрёстные.

Также, было взято большое количество наблюдений для того (в том числе), чтобы была возможность тестирования гипотез в смысле t критерия, поскольку если в модели [гетероскедастичност](https://ru.wikipedia.org/wiki/%D0%93%D0%B5%D1%82%D0%B5%D1%80%D0%BE%D1%81%D0%BA%D0%B5%D0%B4%D0%B0%D1%81%D1%82%D0%B8%D1%87%D0%BD%D0%BE%D1%81%D1%82%D1%8C)ь и автокорреляция, то только много наблюдений могут спасти ситуацию

# Построение модели

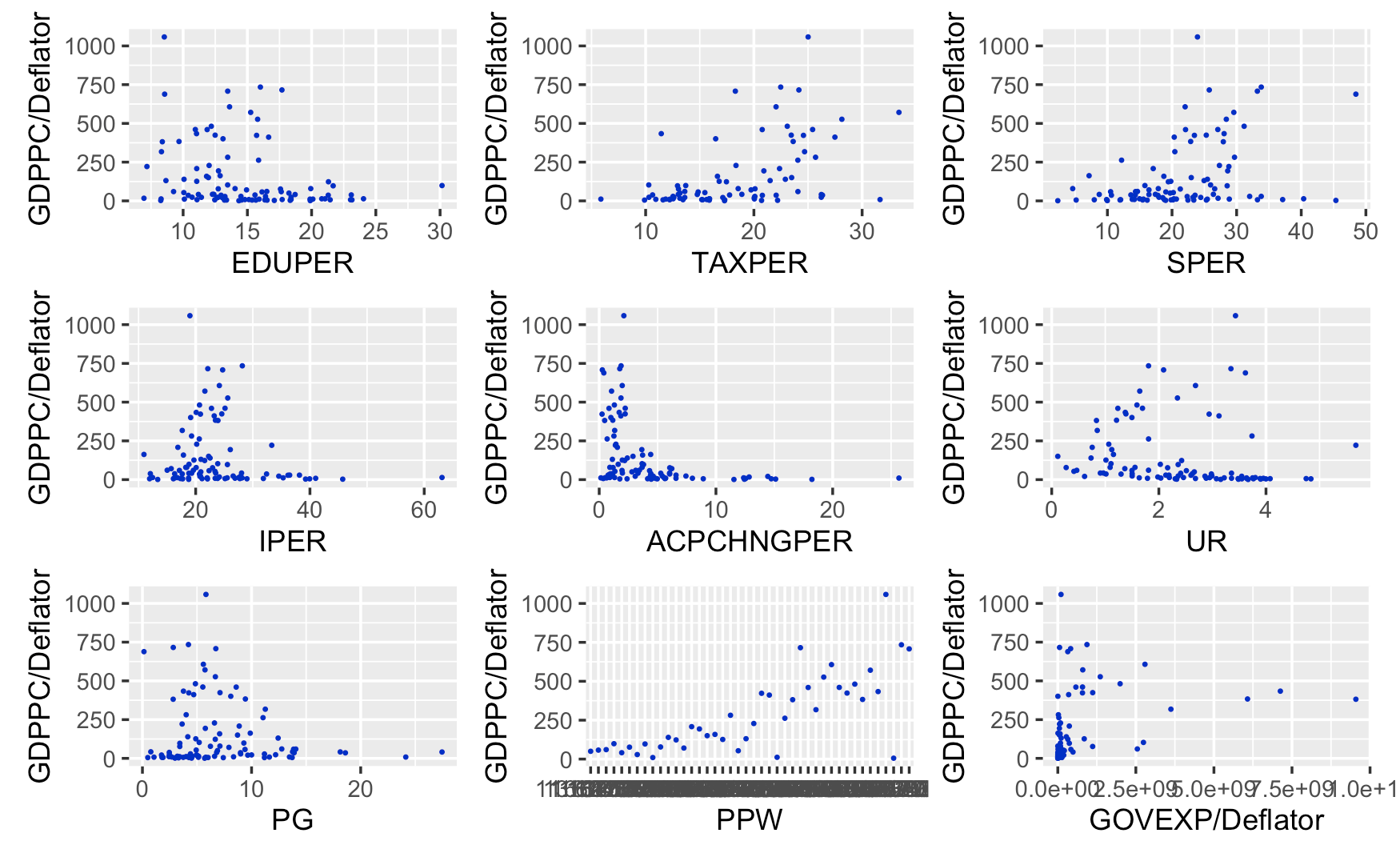
## [R notebook](https://github.com/GrihanS/Classical-regression/blob/main/econometric%20model.Rmd). Представлен код с построением моделей.

## [HTML версия](https://htmlpreview.github.io/?https://raw.githubusercontent.com/GrihanS/Classical-regression/main/econometric-model.html). Читаемая версия кода.

## [Датасет](https://github.com/GrihanS/Classical-regression/blob/fb9293bc8e99b9575b30a841002b2ac4c5ef3e40/Final%201.3.xlsx), то есть данные, собранные для исследования.

Базовое моделирование проходило в рамках классической линейной регрессии, предпосылки которой были обозначены ранее. Итак, для начала нам необходимо понять визуально как зависят регрессоры и зависимая переменная. Было сделано предположение в рамках которых строилась модель, согласно которой зависит линейно от ряда регрессоров. Для этого понадобился сет графиков диаграмм рассеяния, где по оси абсцисс откладывается регрессор, а по оси ординат зависимая переменная

могут оказать воздействие на



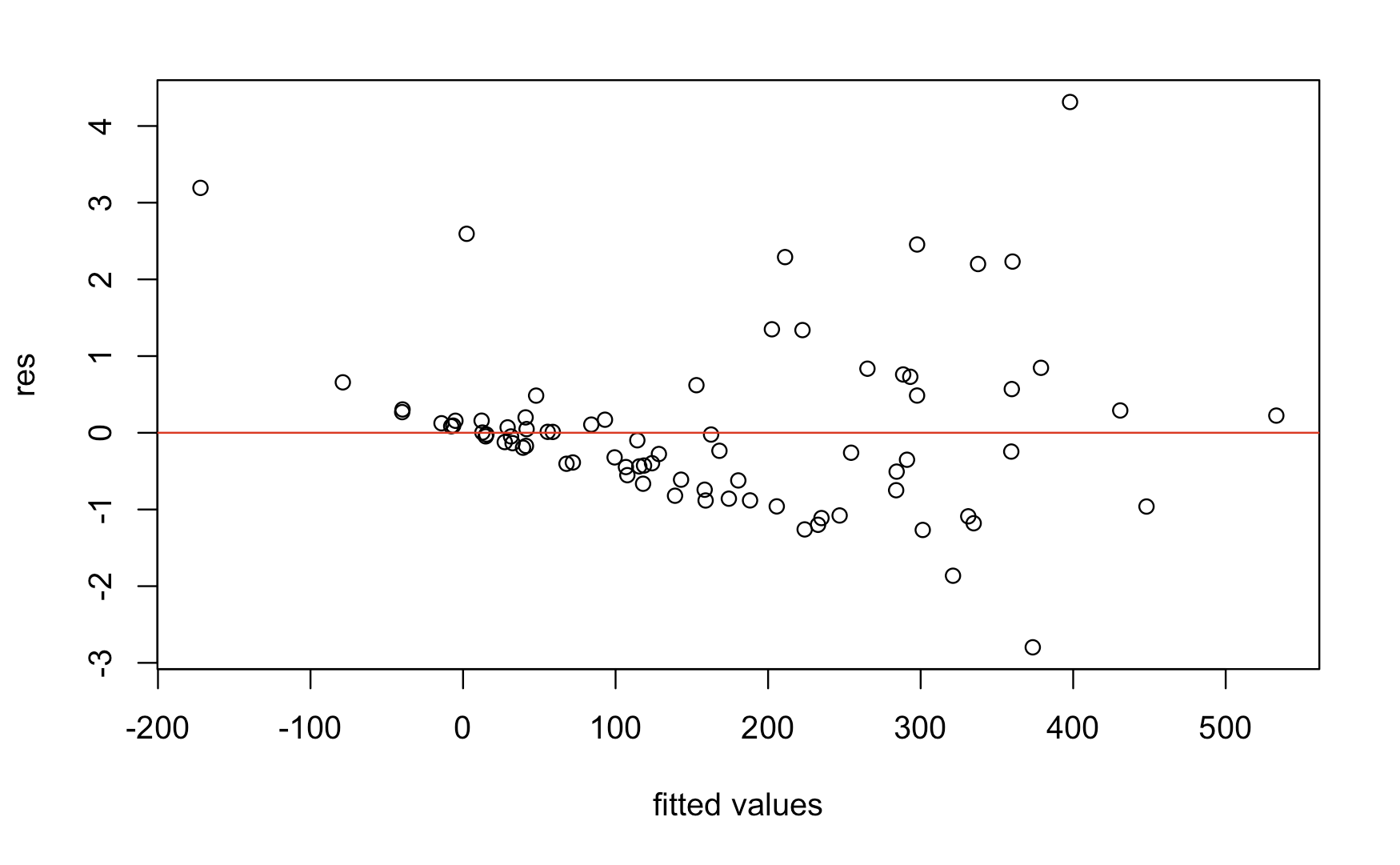
Согласно этому сету графиков можно сделать такой вывод: между и такими регрессорами как:

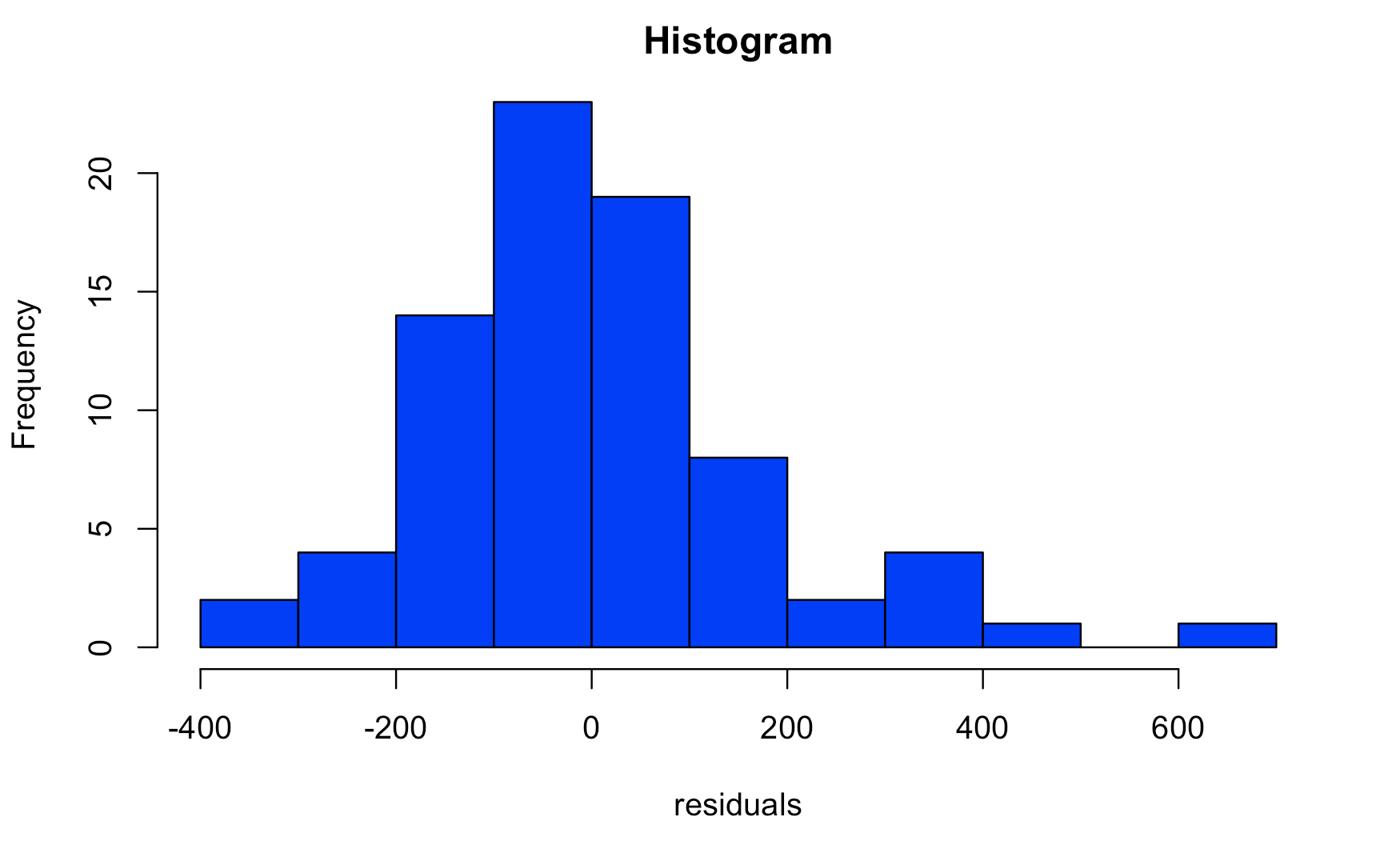
Поэтому будем моделировать в рамках этих регрессоров.

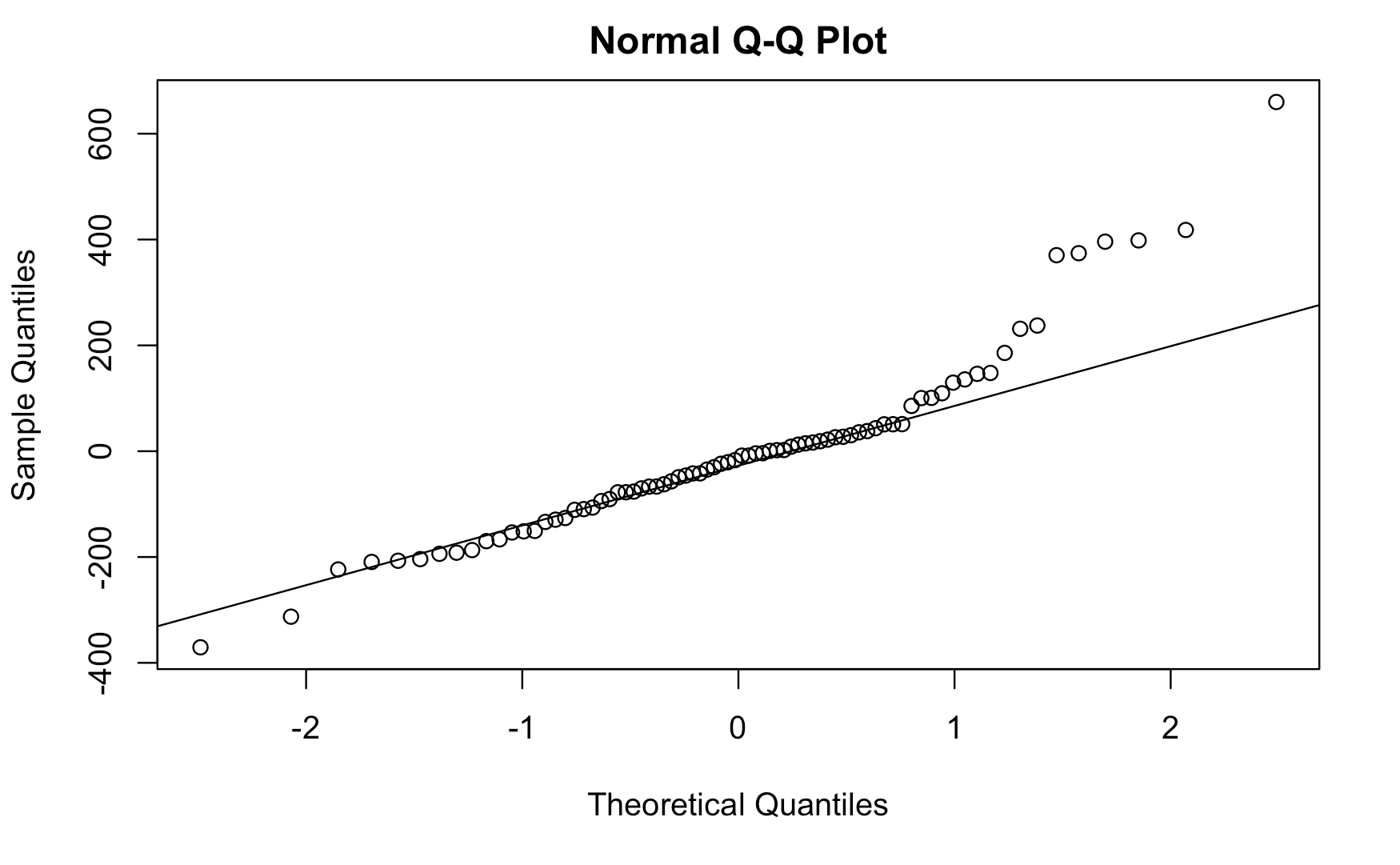
Далее, было оценено несколько возможных зависимостей с помощью МНК. Представлены некоторые из этих моделей ниже.

|  |  |
| --- | --- |
| Вид модели | Некоторые её характеристики |
|  | JB = 37.253,  p-value = 8.139e-09 |
| + | JB = 52.247,  p-value = 4.515e-12 |
|  | JB = 54.115,  p-value = 1.774e-12 |

Краткий вывод по этим трем моделям и схожим с ними: видно, что все три модели страдают очень серьезной проблемой, они некорректно специфицированны. Например, это может быть какой-то пропущенный регрессор или же вообще неверна функциональная форма выбранной модели.[[1],стр 268] [[2], стр 161-163] Ниже приводится один график для первой модели, однако, для других паттерн тот же.





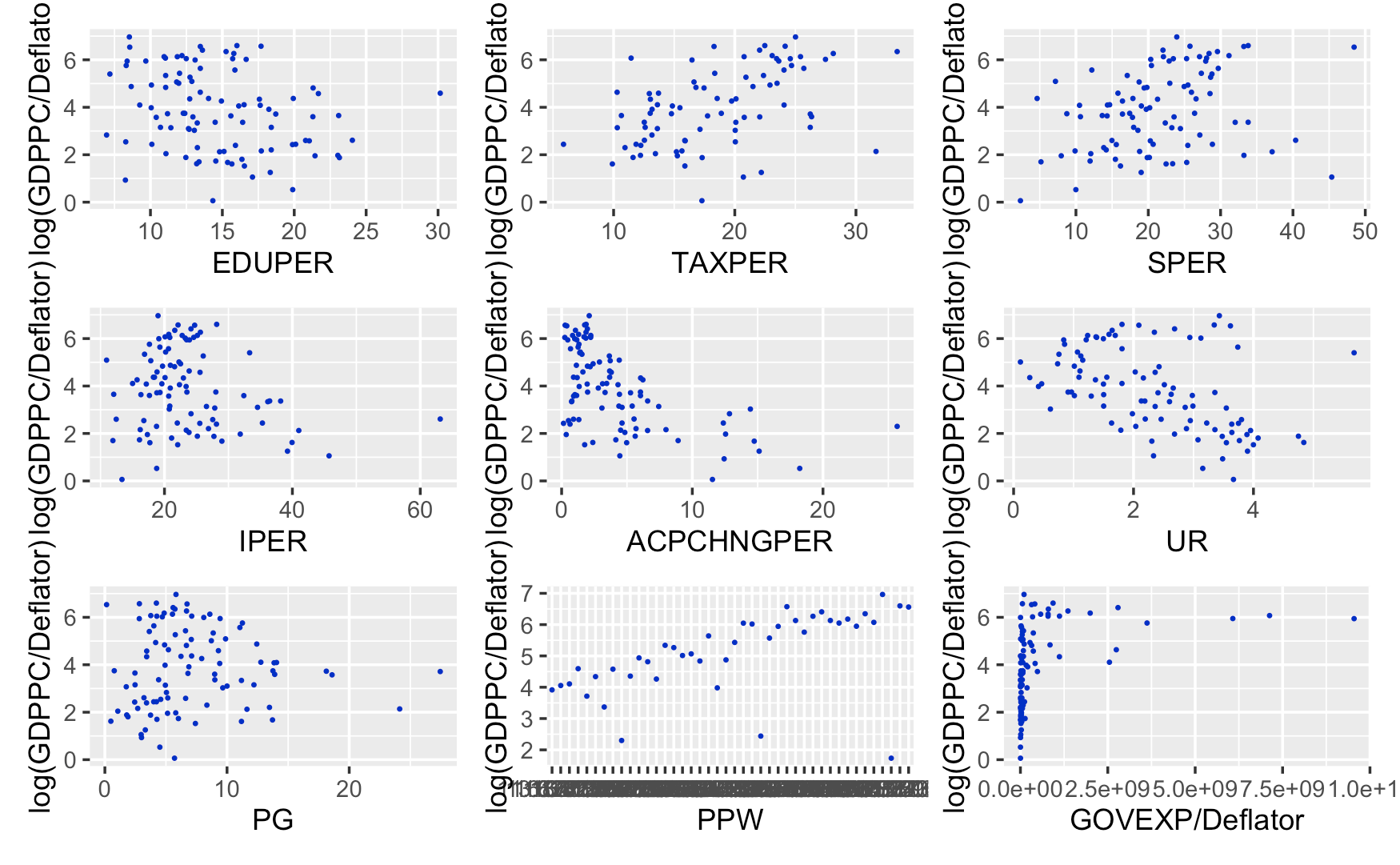


Так, на графике “прогнозы - остатки” мы видим некоторую корреляцию между подогнанным значением и стьюдентизированными остатками, а также смещенность остатков, чего не должно быть в корректной модели.

Графики

предпосылке модели о нормальности ошибок, а тест Jarque&Bera отвергает гипотезу о нормальности остатков на крайне высоком уровне. Может быть пропущена переменная PPW? Полученные результаты по 4й модели также свидетельствуют о некорректности функциональной формы и/или о пропущенных регрессорах (была добавлена просто переменная PPW в степени 1, поскольку, график зависимости GGDPPC / Deflator от PPW демонстрирует линейную зависимость).

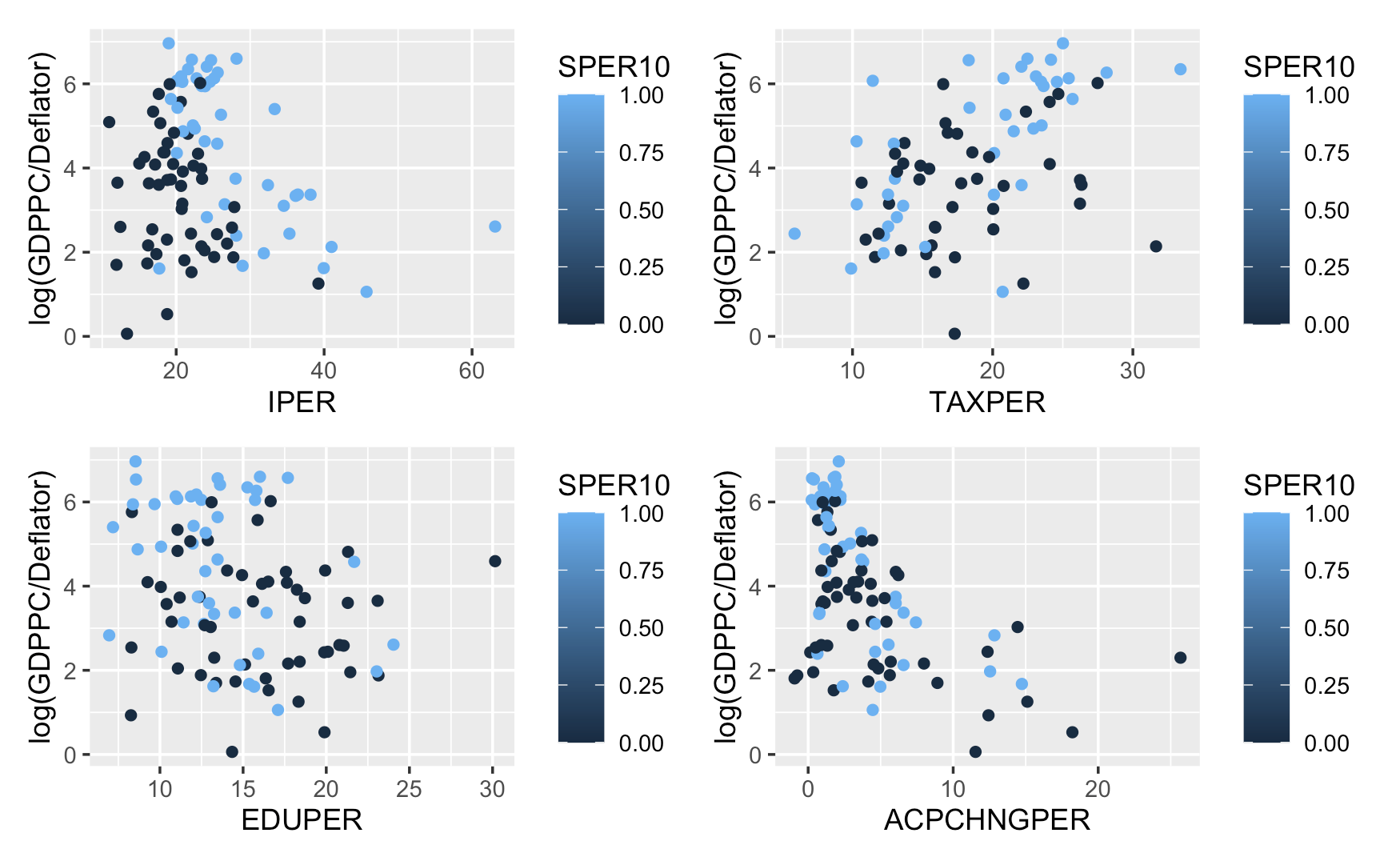
После этого анализа было решено исследовать зависимость вида: . Ниже представлены те же 12 графиков диаграмм рассеивания, но на оси ординат уже представлена переменная



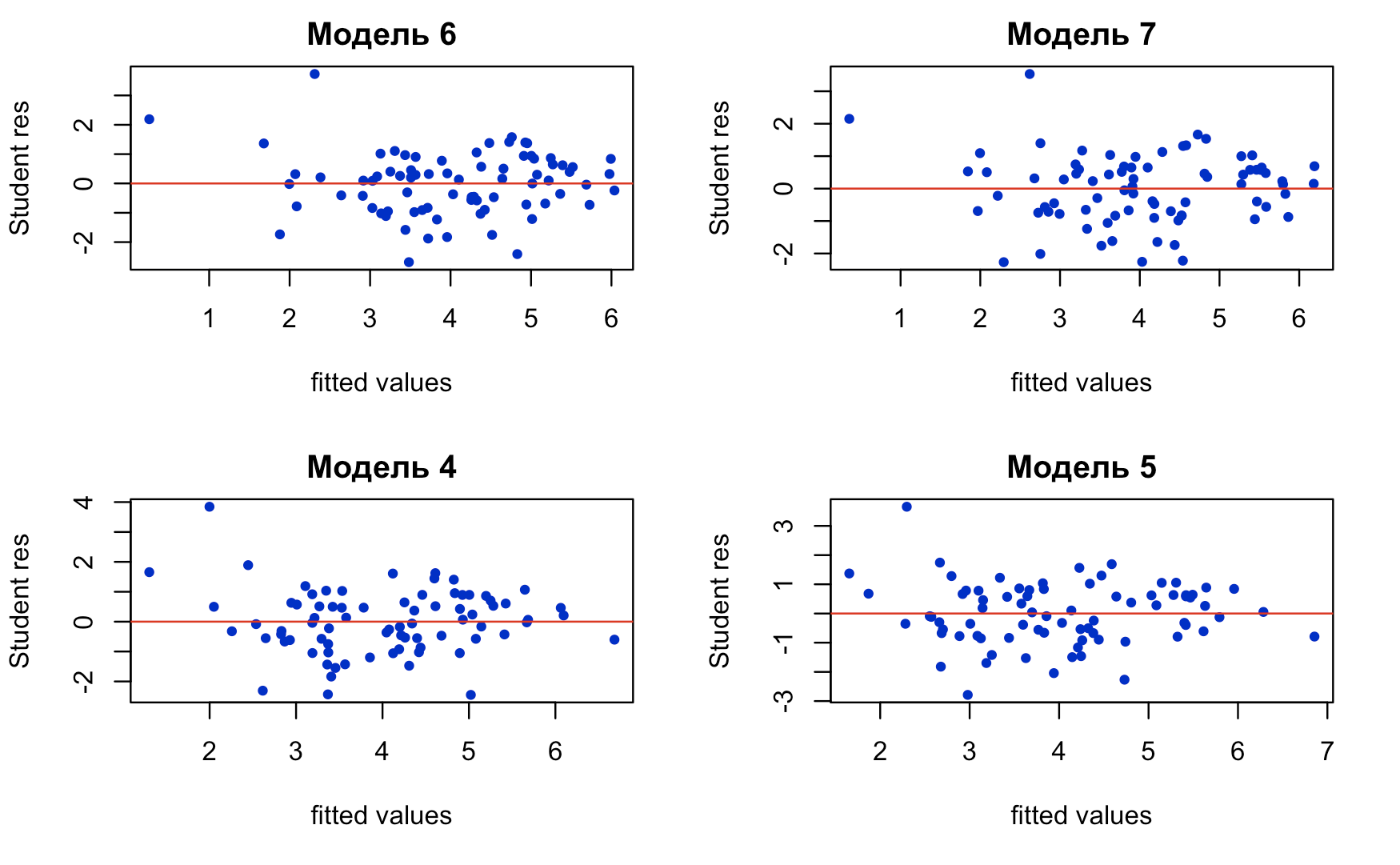
Было отобрано несколько типовых моделей с разными вариациями. Наиболее релевантные модели представлены ниже в таблице.

|  |  |
| --- | --- |
| Вид модели | Некоторые ее характеристики |
|  | JB = 0.84248,  p-v = 0.6562  RESET p-v: 0.0014 |
| Модель 5: | JB = 0.16733,  p-v = 0.9197  RESET p-v: 0.017 |
| Модель 6: | JB = 0.21448,  p-v = 0.8983  RESET p-v: 0.011 |
| Модель 7: | JB = 0.0024281,  p-v = 0.9988  RESET p-v: 0.023 |

Дадим некоторые комментарии по моделям: отличие моделей 4-7 от моделей 1-3 в том, что поменялась зависимая переменная и в целом сохранена структура зависимостей, а также для моделей 5 и 7 введена дамми переменная, обозначающая дополнительный прирост ВВП на душу населения от доли инвестиций в ВВП (IPER) для тех стран у которых норма сбережений (SPER) выше медианного (21.31%). График представлен ниже. Как можно заметить, видна ярко выраженная зависимость.

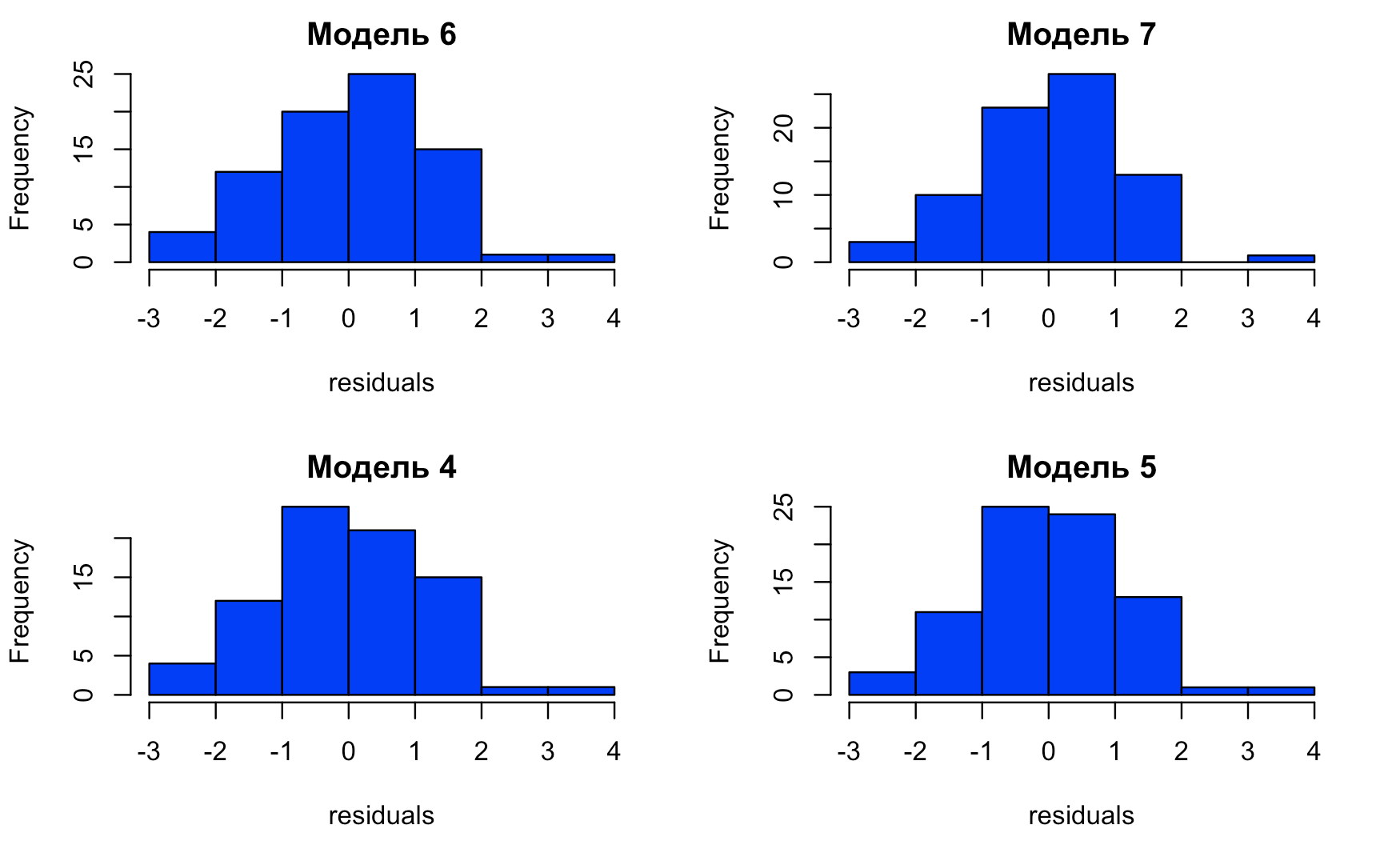


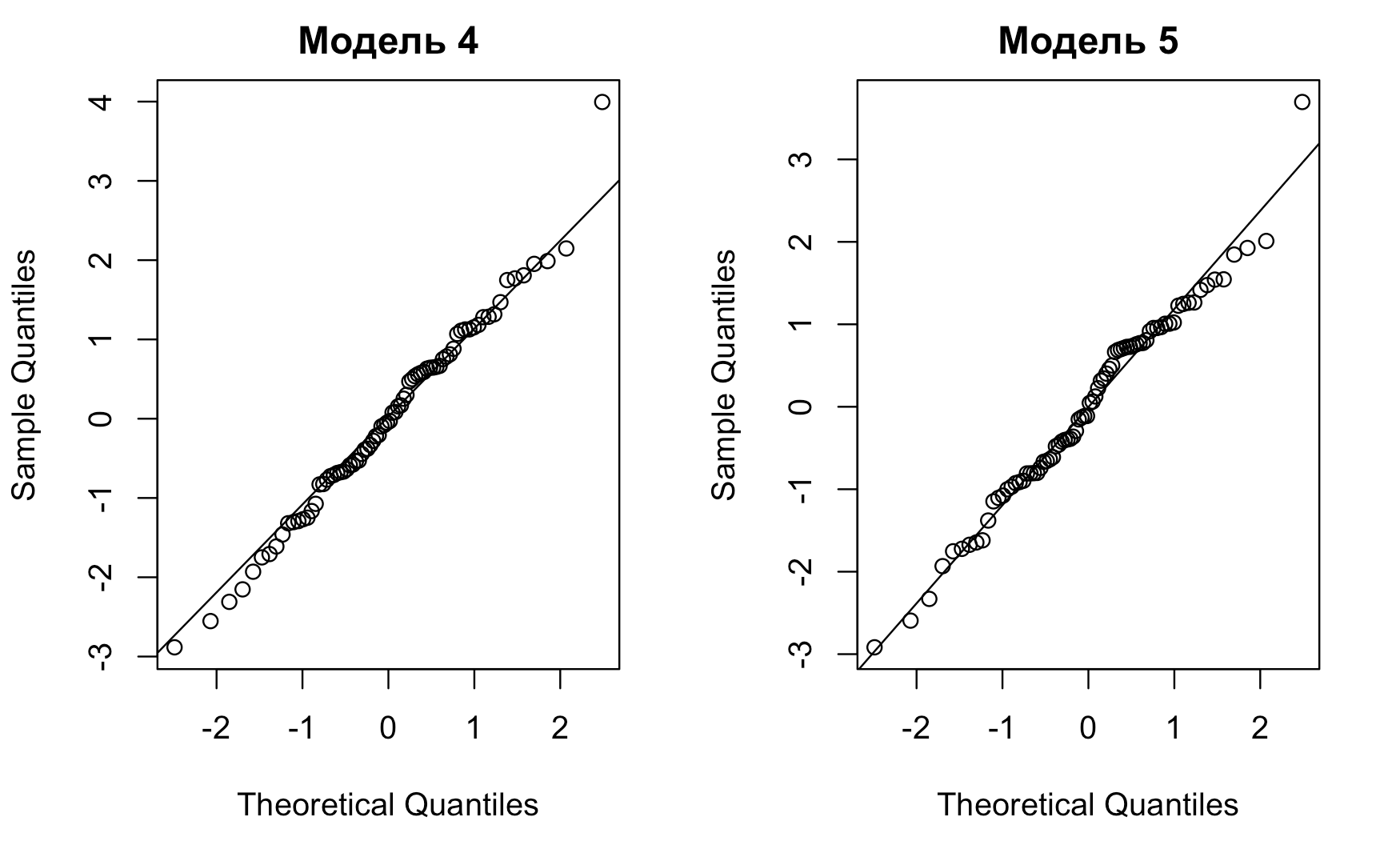
Ниже представлена диагностика этих моделей с использованием графиков “остатки-прогнозы”

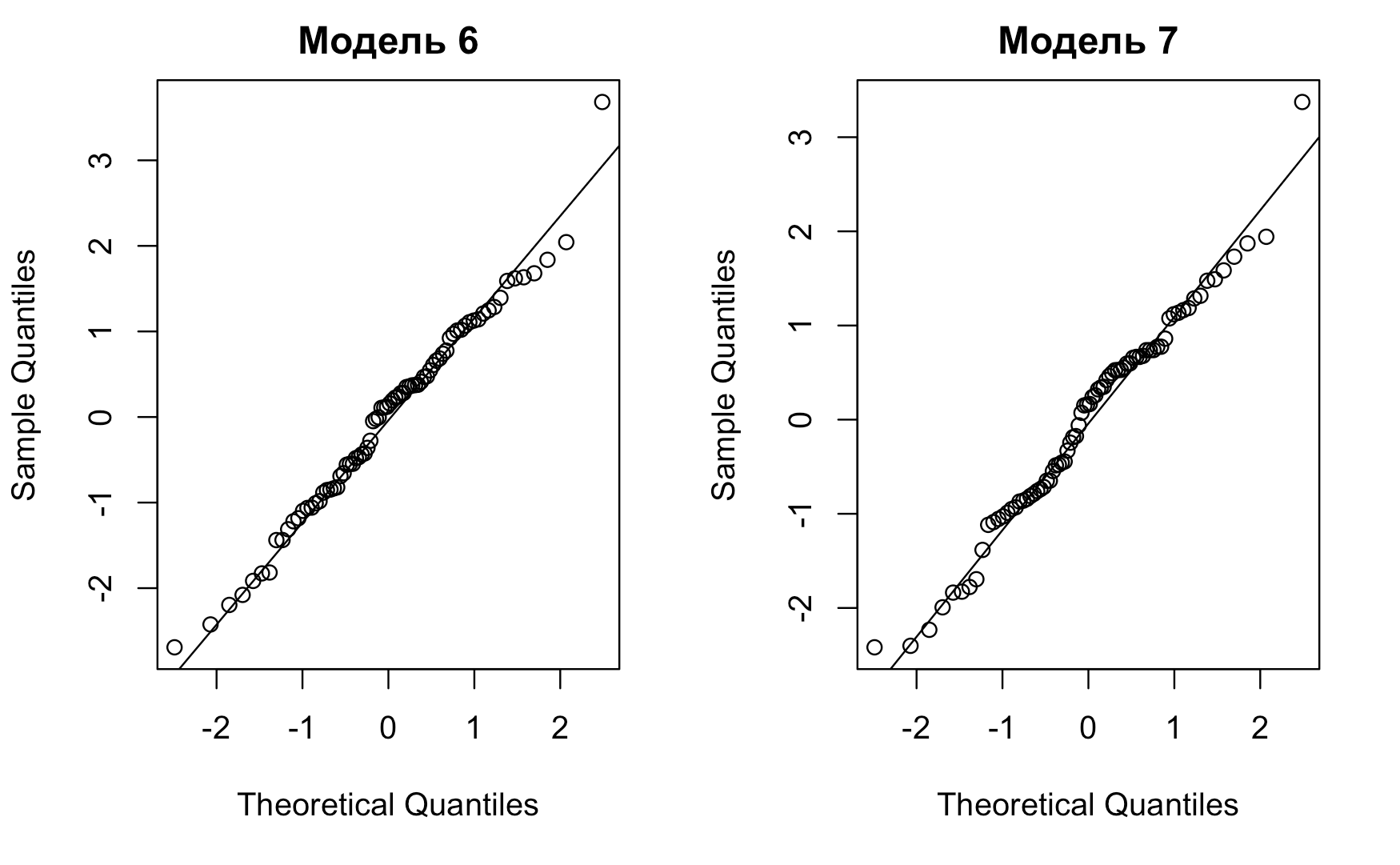


Комментарий: на всех 4х графиках виден очень хороший паттерн, говорящий о том, что модели могут быть правильно специфицированы.

Теперь посмотрим нормальность остатков по qq plot и гистограммам.







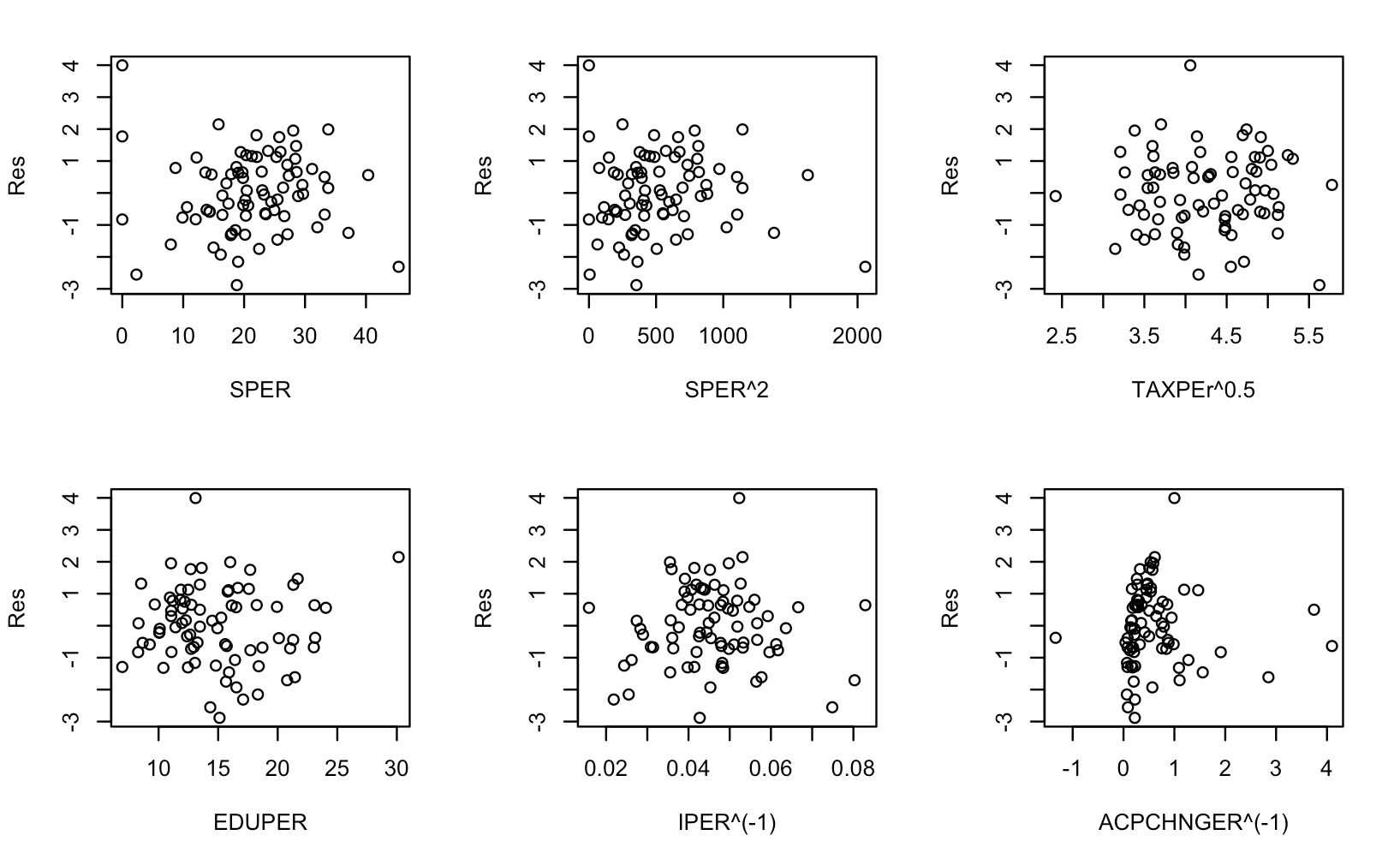
Комментарий: остатки (или, асимптотически, ошибки, при правильной спецификации модели), могут быть распределены нормально, вполне вероятно как

7, в которых присутствуют дамми переменные,

хотя, все модели успешно прошли тест на нормальность ошибок (стоит отметить, что результат, полученный по JB тесту, несколько искажен, поскольку у нас лишь 78 наблюдений, а тест асимптотический, тем не менее, p-value все равно очень высокие).

Далее, проанализируем по графикам потенциальную гетероскедастичность в моделях. Вновь смотрим по графикам, в данной ситуации, зависимость остатков от регрессоров.

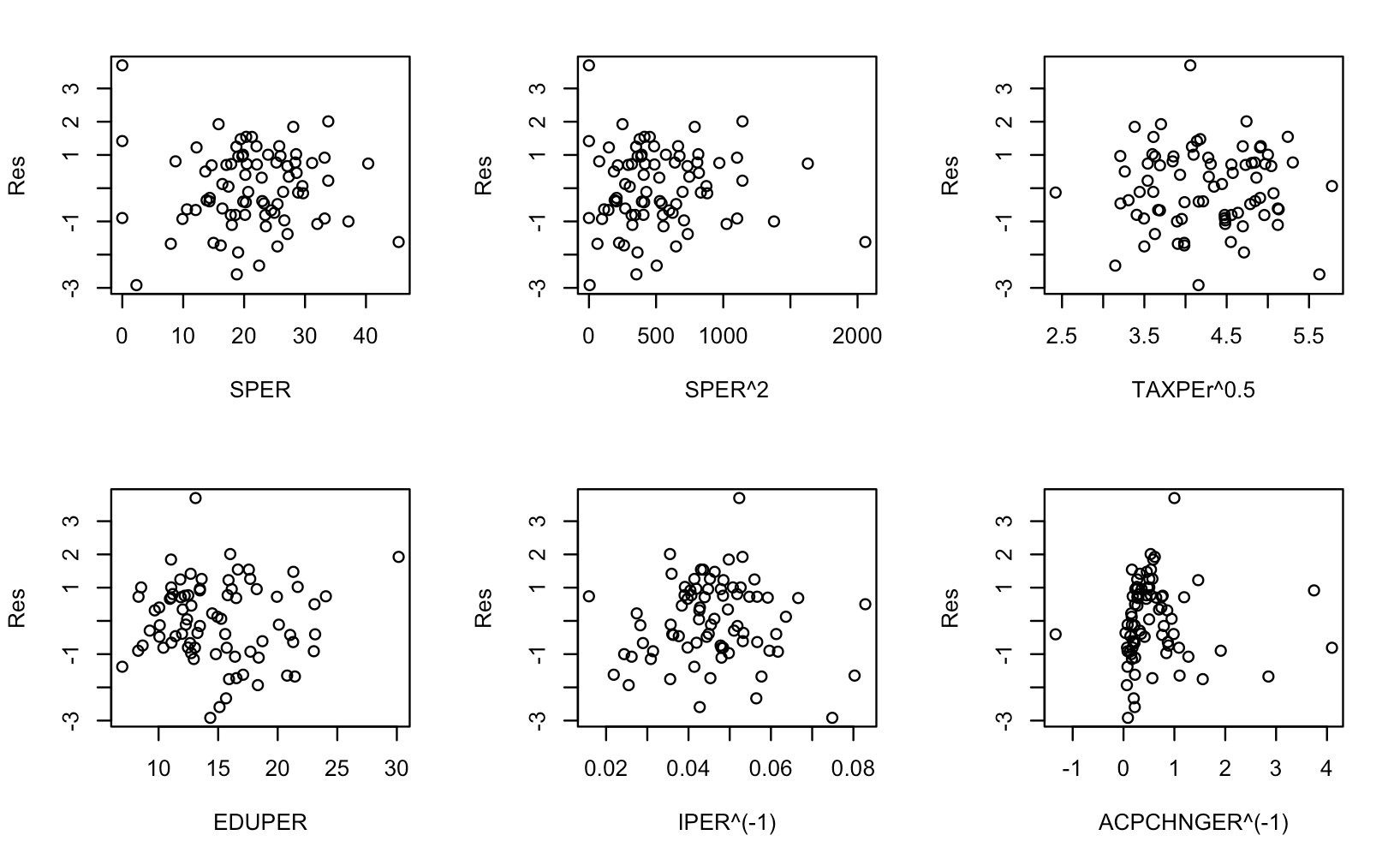
Для модели 4:



Комментарий: потенциальной зависимости остатков от регрессоров с виду не обнаружено, разве что есть подозрения на зависимость между Проведем проверку по тесту Голдфельда-Квандта:

1. По переменной
2. По переменной

Для модели 5:

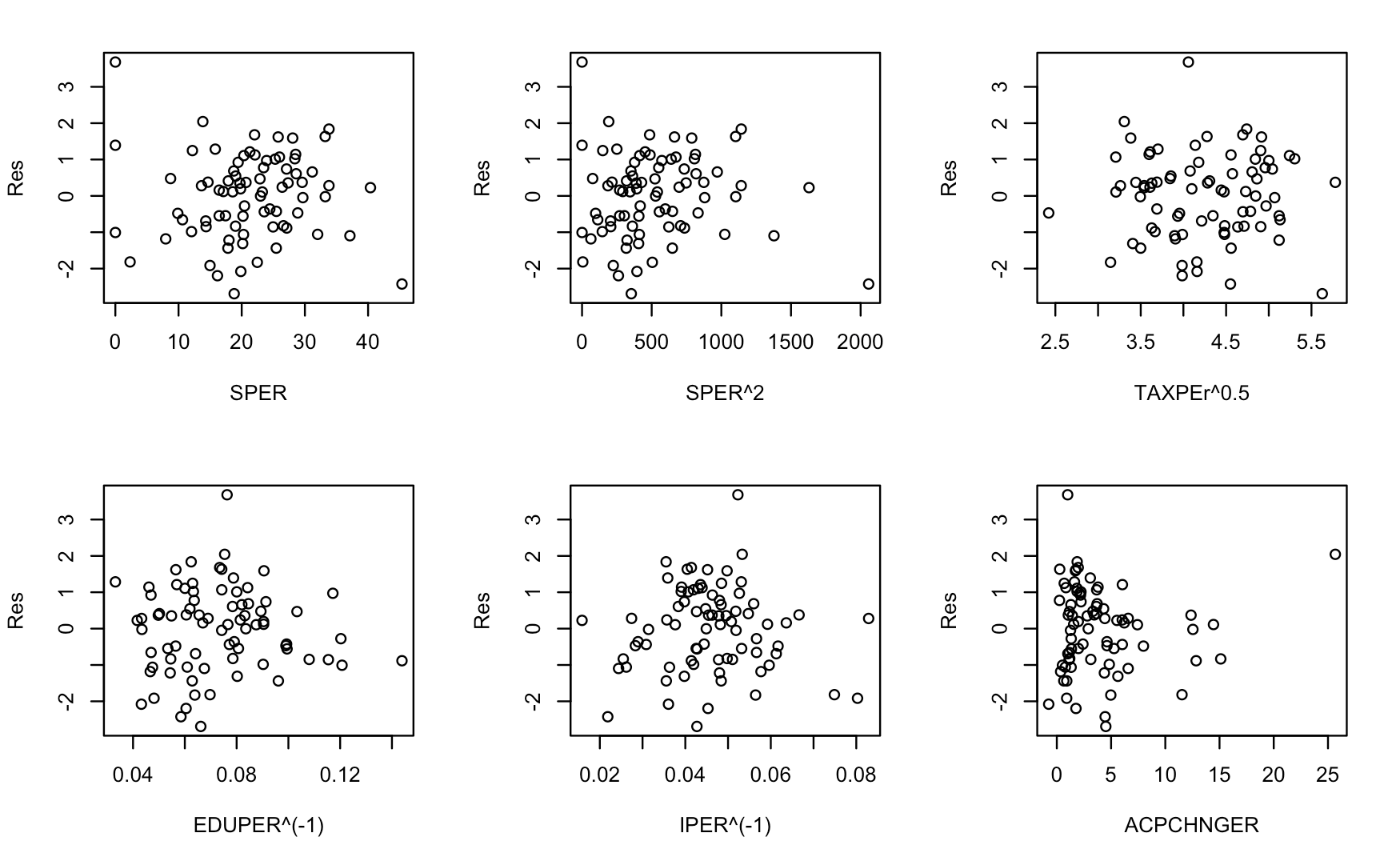


Комментарий: та же ситуация, что и с моделью 4. Не было приведено графика для дамми переменной, т к SPER10 т к это скорее качественная переменная, нежели численная.

Проведем тест Голдфельда-Квандта:

1. По переменной
2. По переменной

Для модели 6:

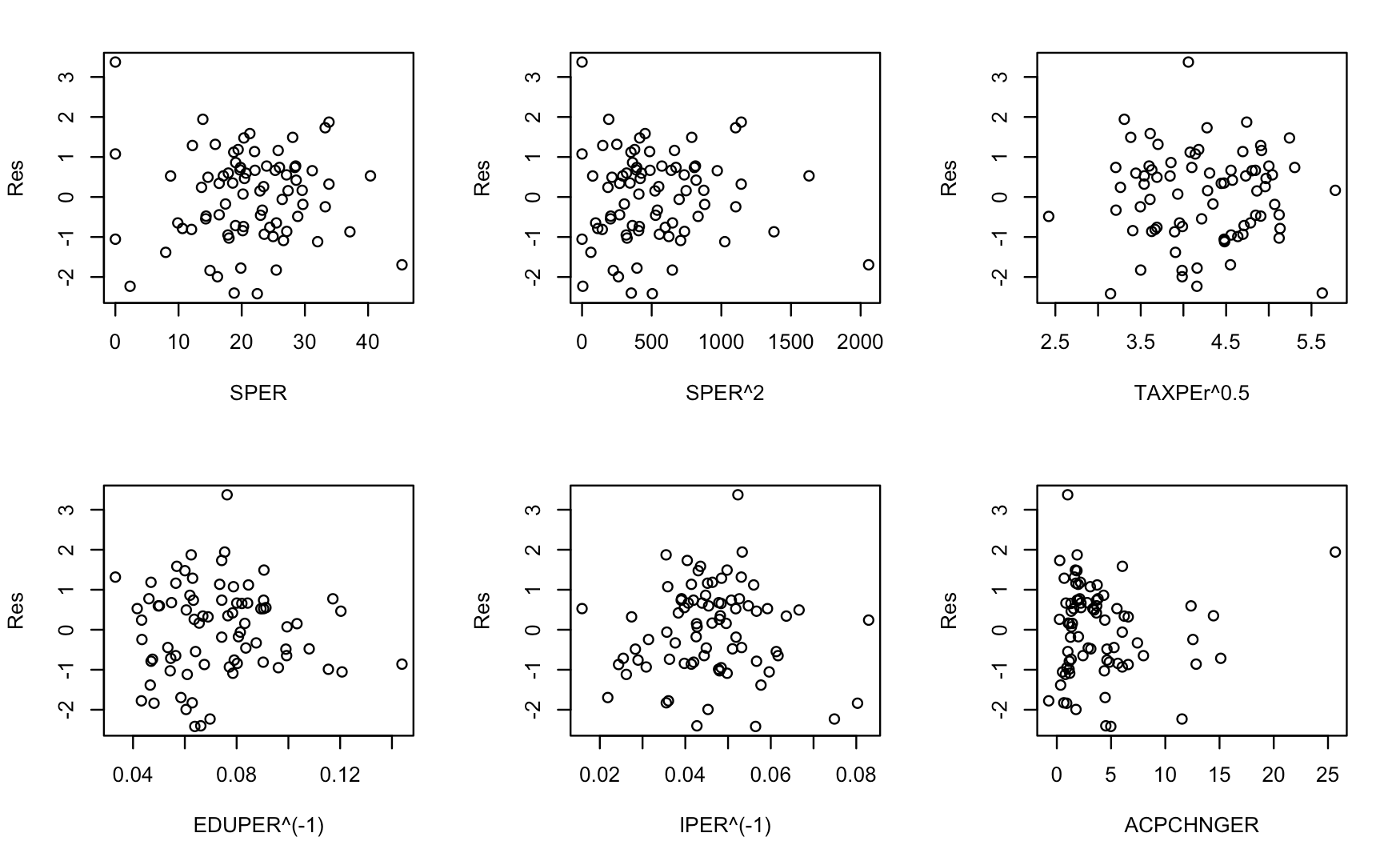


Комментарий: та же ситуация, что и с моделью 5 и 4, наблюдается такой же паттерн, что и в предыдущих моделях. Хотя, не исключено, что теперь есть несколько неоднородная дисперсия из-за неоднородности остатков в зависимости от

Проведем тест Голдфельда-Квандта:

1. По переменной
2. По переменной
3. По переменной 0.9877

Модель 7:



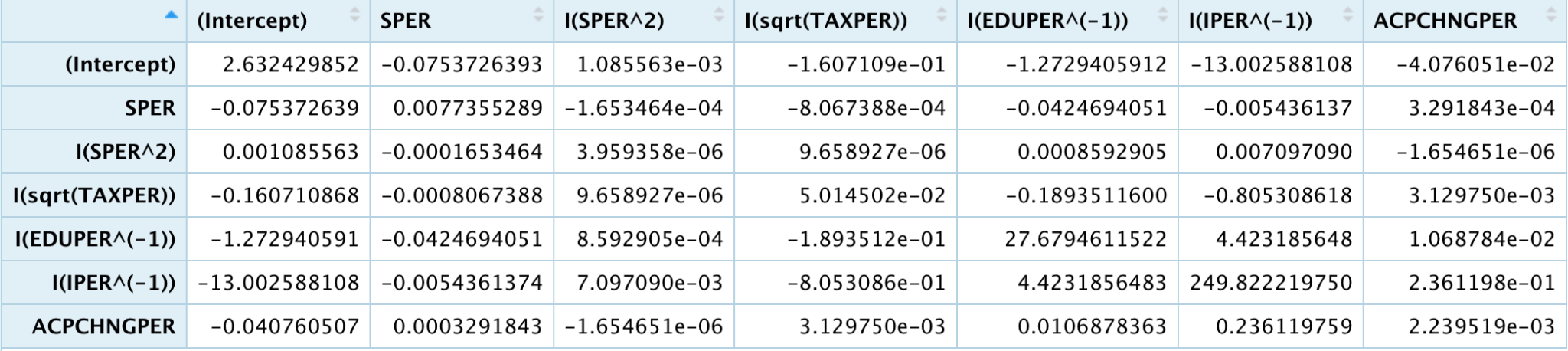
Комментарий: та же ситуация, что и с моделью 6.

Проведем тест Голдфельда-Квандта:

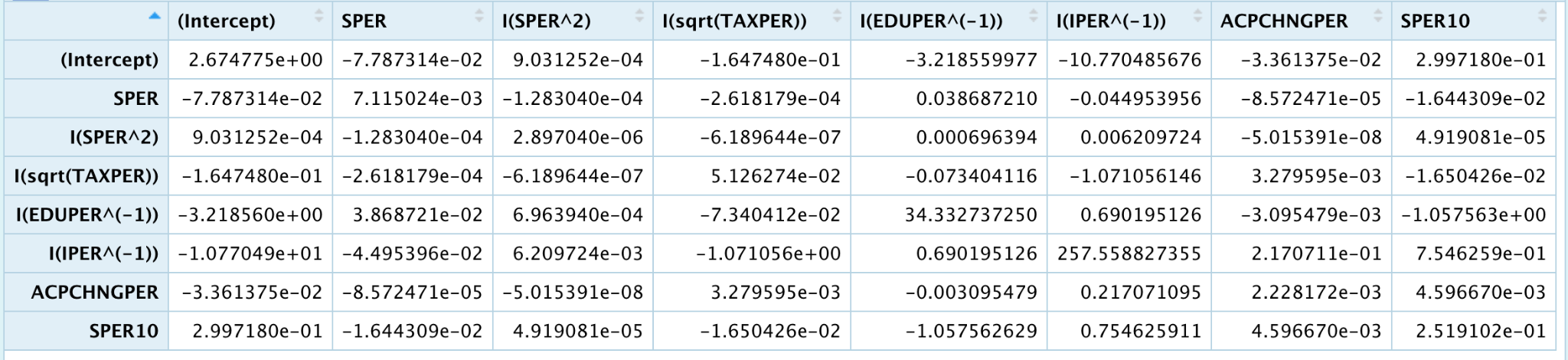
1. По переменной
2. По переменной
3. По переменной 0.9907

Теперь посмотрим на матрицу ковариаций между остатками (ошибки в форме Ньюи-Вэста):

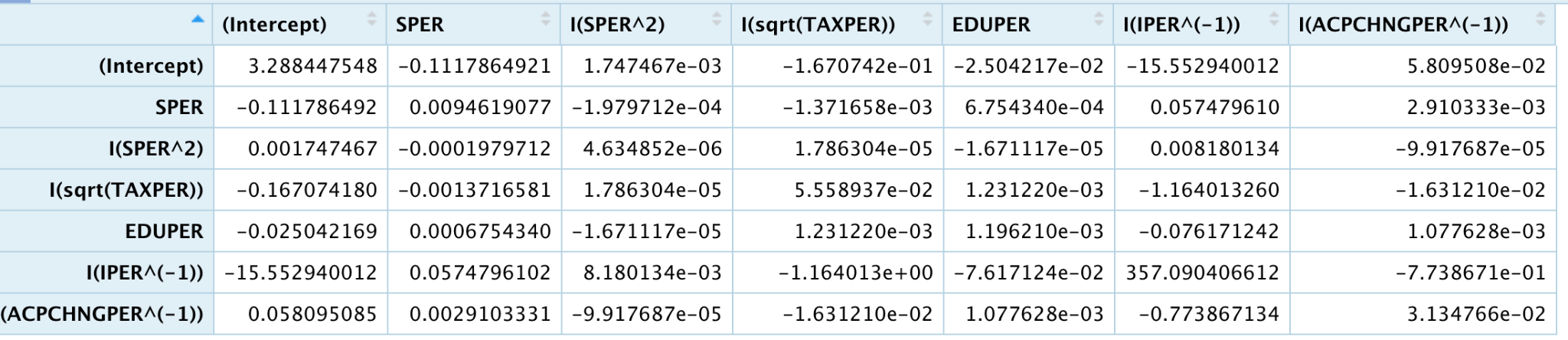
1. Для модели 6



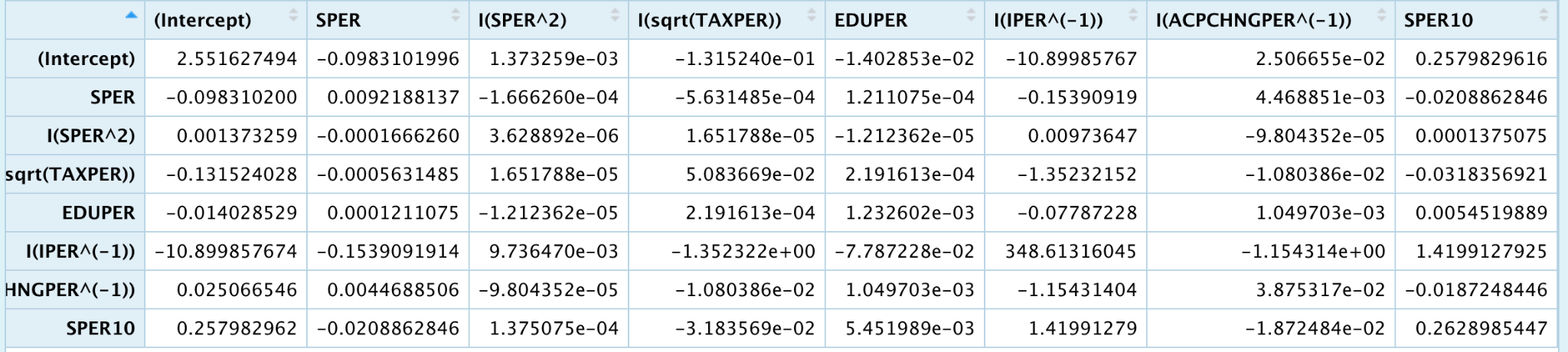
1. Для модели 7:



1. Для модели 4:



1. Для модели 5:



Как видим, по этим данным, на диагональных элементах стоят несколько разные значения, особо резко выделяется показатель , хотя, согласно тесту Голдфельда Квандта, тест на однородность нулевая гипотеза проходит. То есть в этих четырех моделях гипотеза об однородности принимается, хотя и лучше тест прошли модели 6, 7.

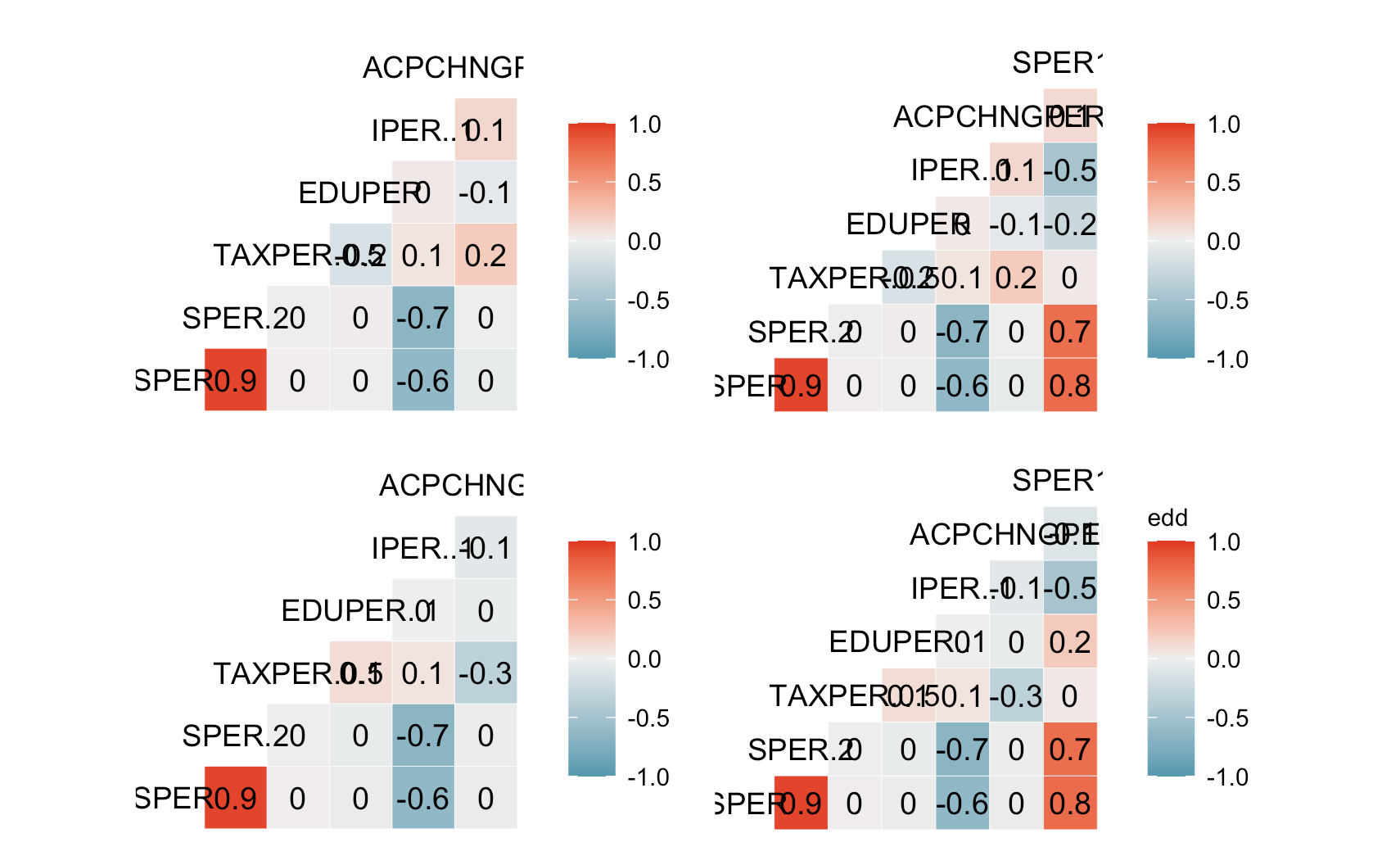
Также, стоит обратить внимание на тот факт, что на не диагональных элементах стоят почти нули, что еще раз подтверждает гипотезу о правильной спецификации модели и об отсутствии пропуска значимых переменных (например в такой ситуации автокорреляция может возникнуть в перекрестных выборках).

Также, стоит заметить, что как минимум модели 6, 7 прошли тест Рэмси на отсутствие пропущенных переменных (на уровне значимости 1%, смотрите таблицу выше). Однако, все модели ( и в особенности 5-6) имеют относительно малые p-value, что может говорить о пропуске какой-то переменной. Ранее была приведена одна важная переменная, а именно производительность труда, ее можно было бы включить в модель, однако, это не гарантирует того, что проблемы уйдут из-за этого, хотя бы потому что в нашем распоряжении оказалось лишь 43 таких наблюдения и выборка может быть не репрезентативна. Также, эта переменная может коррелировать с другими (например, ожидаемая продолжительность жизни населения, образованность, уровень демократии и тп). Поэтому, в идеале, модель может быть доработана в будущем, однако, уже сейчас мы видим, что потенциальных острых проблем не обнаружено. То есть если и есть смещение и несостоятельность коэффициентов модели, то они явно незначительны. Также, предпосылка 3 о равенстве Поскольку это следует из доказанной корректности модели, а также об отсутствии автокорреляции. Например, нарушение этой предпосылки возникает тогда когда есть пропущенная переменная. Мы отвергли эту гипотезу по RESET тесту. Однако, даже если истинное

регрессиях, а коэффициенты не будут затронуты и останутся несмещенными.

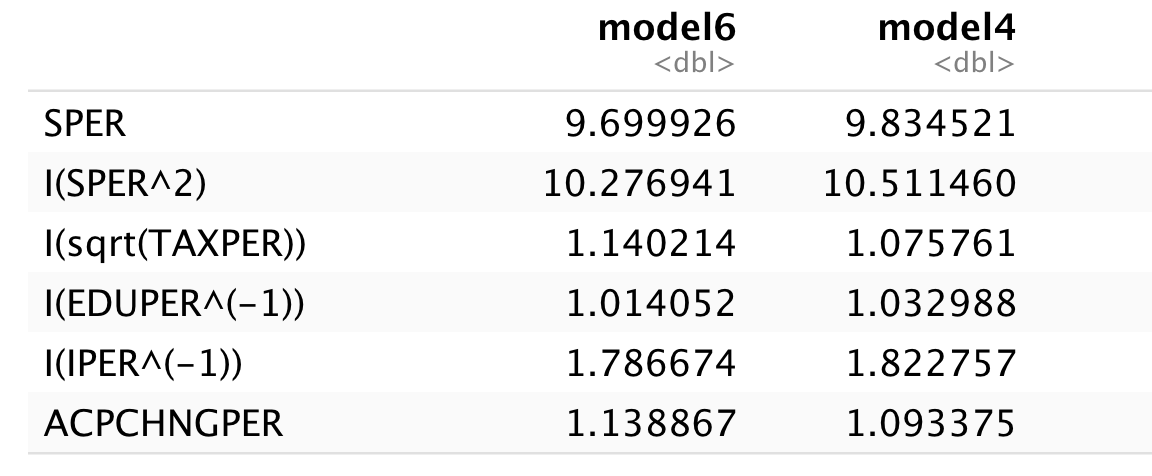
Теперь остается взглянуть на проблему мультиколлинеарности в этих 4х моделях. Ожидается, что SPER и

Взглянем на график корреляции между каждым регрессором в этих четырех моделях.



Действительно, существует проблема высокой корреляции (0,94) между этими двумя регрессорами. Также, ожидаемо, между дамми переменной и SPER.

Проверим по показателю VIF есть ли проблема мультиколлинеарности



Комментарий: как видим, высокой степени мультиколлинеарности по показателю вздутия дисперсии, не обнаружено.

Также и не наблюдается мультиколлинеарности по t критерию, почти все коэффициенты в этих четырех моделях значимы, а по F тесту каждая регрессия также значима. То есть нет противоречия.

Посчитаем теперь число обусловленности для каждой модели.

|  |  |  |  |
| --- | --- | --- | --- |
| Модель 4 | Модель 5 | Модель 6 | Модель 7 |
| 72227.85 | 72233.92 | 71533.84 | 71534.25 |

Итак, мы получили, что число обусловленности (чувствительность изменения значений объясняемых переменных к значению регрессоров) достаточно высоко для всех четырех моделей. Однако, в целом, вывод такой: сильной мультиколлинеарности в этих четырех моделях не наблюдается (например об этом говорят t критерий, коэффициенты VIF, отсутствие коллинеарности между другими регрессорами). Стоит вновь обратить внимание на тот факт, что в моделях 6 и 7 число обусловленности оказалось меньше чем в 4й и 5й моделях. Это вновь нам говорит о том, что модели 6 и 7 качественнее 4й и 5й.

# Итоги

Подводя итоги, выше был проведен анализ состоятельности, несмещенности и эффективности модели и ее коэффициентов. Доказано, что полученная модель удовлетворяет предпосылкам в том числе о гомоскедастичности и об отсутствии автокорреляции, правильно специфицирована и ошибки регрессии имеют нормальное распределение. Значит, мы можем применять t и F критерии для анализа коэффициентов модели. Согласно F критерию, все 4 регрессии значимы, однако, в моделях с дамми переменной выделяется особый паттерн:

важный для нас показатель, а именно влияние расходов на образования значим на 5% уровне в обеих регрессиях, где нет дамми переменных. Но незначим в регрессиях с дамми переменной. Также, интересно, что в модели 6 регрессоры все значимы на 5% уровне, за исключением IPER. В остальных моделях меньше значимых коэффициентов.

Итого, говоря о предпочтении и выборе какой-то одной модели из 4х найденных, можно попробовать применить критерии AIC и BIC. Приведем их ниже в таблице:

AIC:

|  |  |  |  |
| --- | --- | --- | --- |
| Модель 4 | Модель 5 | Модель 6 | Модель 7 |
| 268.3985 | 264.9784 | 260.343 | 256.1476 |

BIC:

|  |  |  |  |
| --- | --- | --- | --- |
| Модель 4 | Модель 5 | Модель 6 | Модель 7 |
| 287.2522 | 286.1888 | 279.1966 | 277.3579 |

И мы снова видим, что модели 6 и 7 лучше по в том числе этим параметрам.

Подводя главные итоги этого исследования стоит указать на следующие факты:

1. Модели 6, 7 лучше, чем 4 и 5 по многим критериям
2. Отдать предпочтение какой-то одной нельзя, т к нужно дополнительное исследование (выше об этом было сказано)
3. Найденные 4 модели удовлетворяют предпосылкам классической линейной регрессии.

# Интерпретация результатов:

Нас интересовало прежде всего влияние расходов на образование на ВВП на душу населения (GDPPC / Deflator). Итак, если отдавать предпочтение модели 6, то выйдет, что зависимость расходов на образование может быть представлена в виде:

У него получился такой результат:

допустим, увеличить расходы на образование на 1%, то ВВП на душу увеличится на 0.46%. У нас же ситуация другая, моделировалась доля ВВП, уходящая на расходы на образование. То есть, если увеличить долю расходов ВВП на образование, то *в среднем,* реальный ВВП на душу будет падать). Казалось бы, что это некоторый парадокс, который сложно объяснить, но, как мы уже выше установили, модель может претендовать на верно специфицированную. Также, виден общий паттерн по всем странам, что при росте доли расходов на образование, ВВП на душу в среднем падает. Понятное дело, что это может быть связано с какими-то структурными проблемами (например, в разнице мультипликаторов расходов). Также, еще один парадокс, следующий из модели: увеличивая долю налогов в ВВП вы увеличите ВВП на душу населения. Интересно, что наблюдается четкая взаимосвязь между группой бедных стран и группой богатых.



Так, по графику видно как распределены страны для которых ВВП на душу выше медианного, они показаны голубым цветом. Это не просто совпадение, что для них ВВП на душу выше. Возможно, это может быть связано с большей прозрачностью статистики по налогам. А также, в выборке отсутствовали страны, у которых высокие доли налогов в ВВП (выше 35%). Поэтому делать какие-либо выводы можно лишь только в рамках похожих данных, а не экстремальных. Для выяснения причин выше названных парадоксов, а также улучшение качества модели (включение дополнительных переменных, влияющих на экономический рост, например учет технологического прогресса как в модели Ромера[3]), необходимо провести дополнительное исследование (хотя, возможно, такое уже проводилось).

# Источники:

[1] Jan Kmenta, Elements of Econometrics, Macmillan, New York, 1986

[2] Себер Дж. Линейный регрессионный анализ, M 1980

[3] Wendy Carlin, David W. Soskice Macroeconomics: Institutions, Instability and the Finansial System, Oxford University Press, 2014