Часть 1

1. Наша команда выбрала датасет «Computers» из пакета «Ecdat». Он представляет собой цены и характеристики персональных компьютеров 486 серии (с микропроцессором Intel 486) в США за 1993–1995 гг., n = 6259 наблюдений.

В датасете имеются следующие переменные:

- price – цена в долларах (зависимая переменная);
- speed – тактовая частота (МГц);
- hd – размер жесткого диска (МБ);
- ram – объем оперативной памяти (МБ);
- screen – диагональ экрана (дюймы);
- cd – наличие CD-дисковода (yes/no);
- multi – мультимедийный комплект (yes/no);
- ads – количество рекламных объявлений в каждом месяце;
- premium – бренд IBM/Compaq (yes/no);
- trend – временной тренд (пронумерованные месяцы начиная с января 1993).

Мы хотим построить регрессию, которая бы показывала зависимость цены компьютера от его технических характеристик.

Для начала мы решили включить все регрессоры со следующим предполагаемым влиянием на цену компьютера:

- speed, hd, ram, screen – тактовая частота процессора, размер жесткого диска, объем оперативной памяти, размер экрана: чем больше, тем дороже компьютер (**положительная** связь);
- cd, multi – наличие дисковода и дополнительных аксессуаров увеличивает стоимость (**положительная** связь);
- premium – бренды IBM, COMPAQ имеют премиальную наценку (**положительная** связь);
- trend – временной тренд, чем больше, тем доступнее и относительно дешевле стоит компьютер, так как технологии развиваются (**отрицательная** связь).

Первоначальная модель такова: price ~ speed + hd + ram + screen + cd + multi + premium + ads + trend

Однако анализ показателя VIF убеждает нас в том, что, например, регрессор hd излишен, поскольку для него VIF равен почти 4. Уберем из регрессоров hd, а также ads, так как коэффициент перед ads близок к 0. Заметим, что избавление от ads не повлияло на значение оценок коэффициентов перед другими регрессорами:

	Зависимая переменная: цена компьютера		
Переменные \ Модель	Mod	Mod1	Mod2
Частота	9.320^{***}	9.827^{***}	9.700^{***}
	(0.185)	(0.196)	(0.196)
Размер жесткого диска	0.782^{***}		

	(0.028)		
Оперативная память	48.256***	70.389***	69.661***
	(1.066)	(0.770)	(0.768)
Размер экрана	123.089***	125.911***	125.450***
	(3.999)	(4.247)	(4.267)
Наличие CD-дисковода (да/нет)	60.917***	93.771***	106.805***
	(9.516)	(10.031)	(9.944)
Наличие мультимедийного комплекта (да/нет)	104.324***	74.793***	73.088***
	(11.413)	(12.071)	(12.128)
Премиум-бренд (да/нет)	-509.225***	-488.137***	-503.674***
	(12.342)	(13.085)	(13.001)
Количество рекламных объявлений	0.657***	0.426***	
	(0.051)	(0.054)	
Месяц	-51.850***	-43.400***	-44.725***
	(0.629)	(0.588)	(0.566)
Константа	307.988***	269.103***	412.076***
	(60.353)	(64.085)	(61.791)
Количество наблюдений	6,259	6,259	6,259
R^2	0.776	0.747	0.744
Скорректированный R^2	0.775	0.746	0.744
F-статистика	2,399.397***	2,304.032***	2,598.578***

*Примечание: в скобках указаны стандартные ошибки. Уровни значимости: *p<0.1; **p<0.05; ***p<0.01.*

Таблица 1. Модели множественной регрессии для цены компьютера

Итак, регрессоры мы отобрали, все показатели VIF имеют значения, близкие к 1. Все оценки коэффициентов перед регрессорами в данной модели значимы на 1% уровне значимости. Предполагаемые зависимости подтвердились за исключением одного регрессора – premium. Оказывается, что принадлежность компьютера к брендам IBM, COMPAQ при прочих равных снижали его цену! Действительно, это соотносится, например, с информацией в Википедии: «В ноябре 1982 года Compaq анонсировала свой первый продукт, переносной IBM PC совместимый персональный компьютер Compaq Portable. Он был выпущен в марте 1983 года и стоил 2995 американских долларов, значительно дешевле предложений конкурентов в то время» (https://ru.wikipedia.org/wiki/Compaq).

Итак, наша модель имеет вид: price ~ speed + ram + screen + cd + multi + premium + trend

Теперь необходимо разобраться со спецификацией модели. Начнем с зависимой переменной: необходимо ли брать ln(price)? Для ответа на этот вопрос проведем тест Бокса-Кокса:

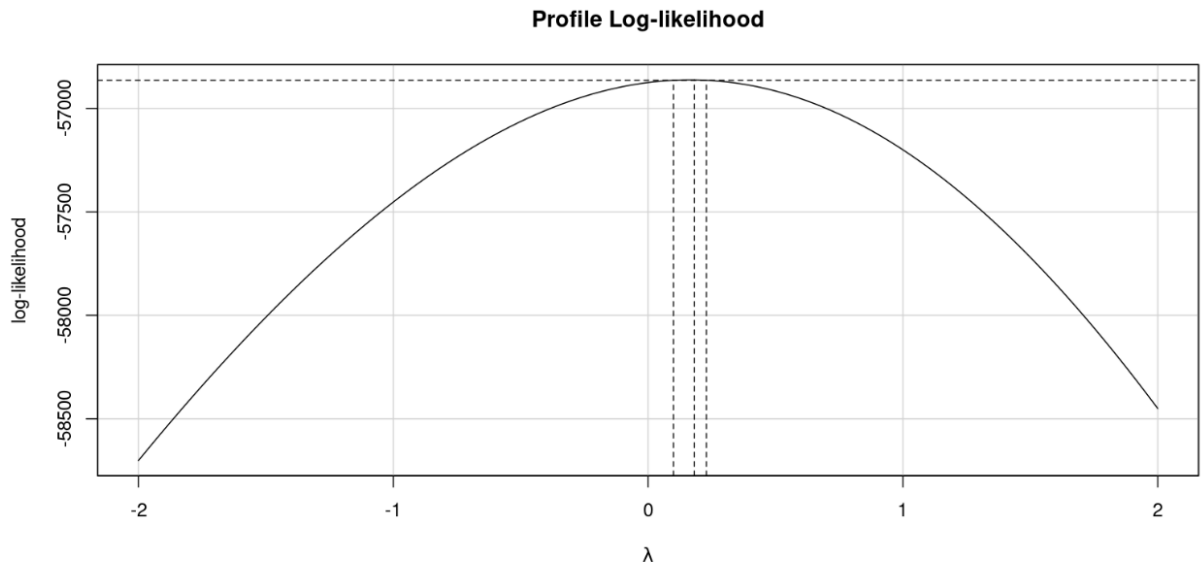

Рисунок 1. Функция максимума правдоподобия для λ в модели Mod2

Видим, что λ близок к 0, значит, надо логарифмировать. Об этом же нам говорит анализ остатков QQplot, мы видим «тяжелый» правый хвост:

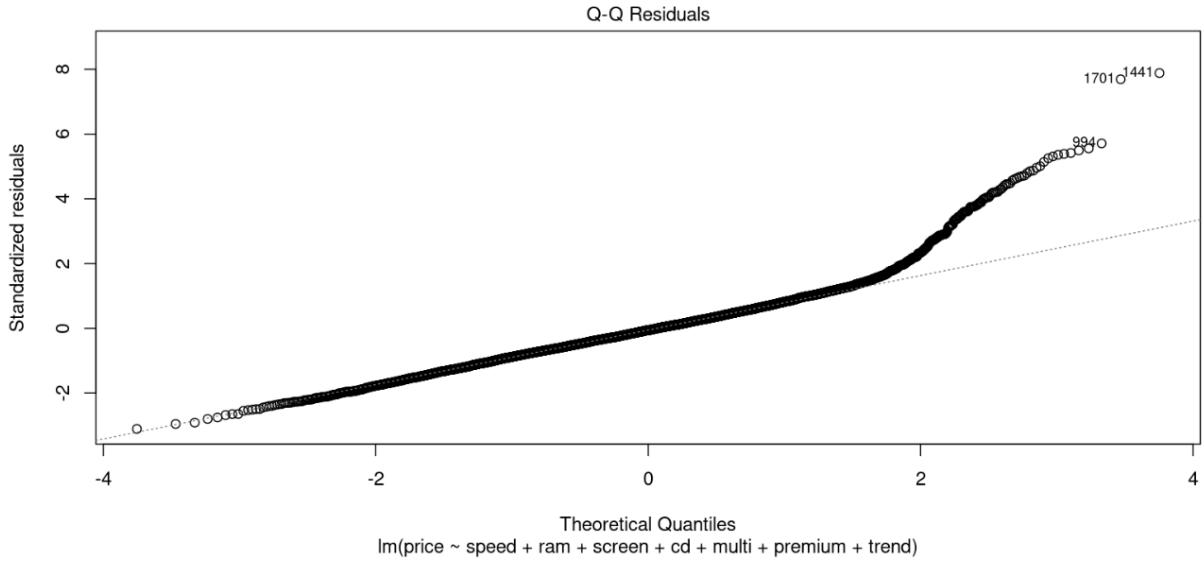

График 2. Отклонение остатков в модели Mod2 от нормального распределения

Значит, преобразуем модель и возьмем в качестве зависимой переменной log (price): тогда λ станет ближе к 1, чем к 0, а также правый хвост остатков станет меньше:

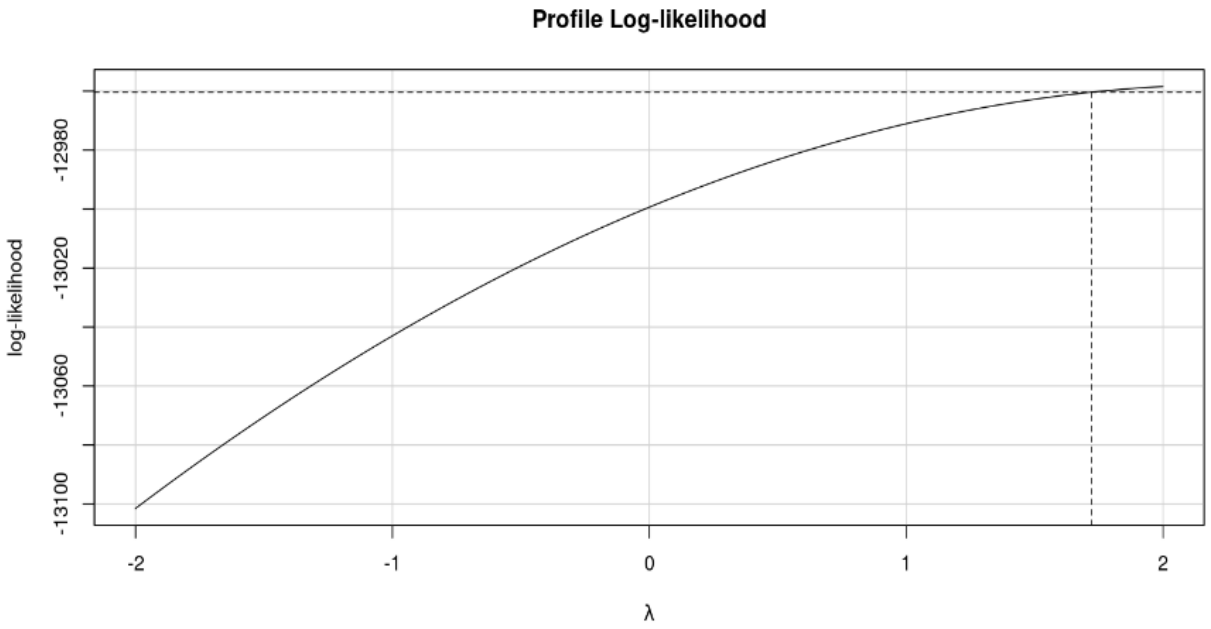

Рисунок 3. Функция максимума правдоподобия для λ в модели с log(price)

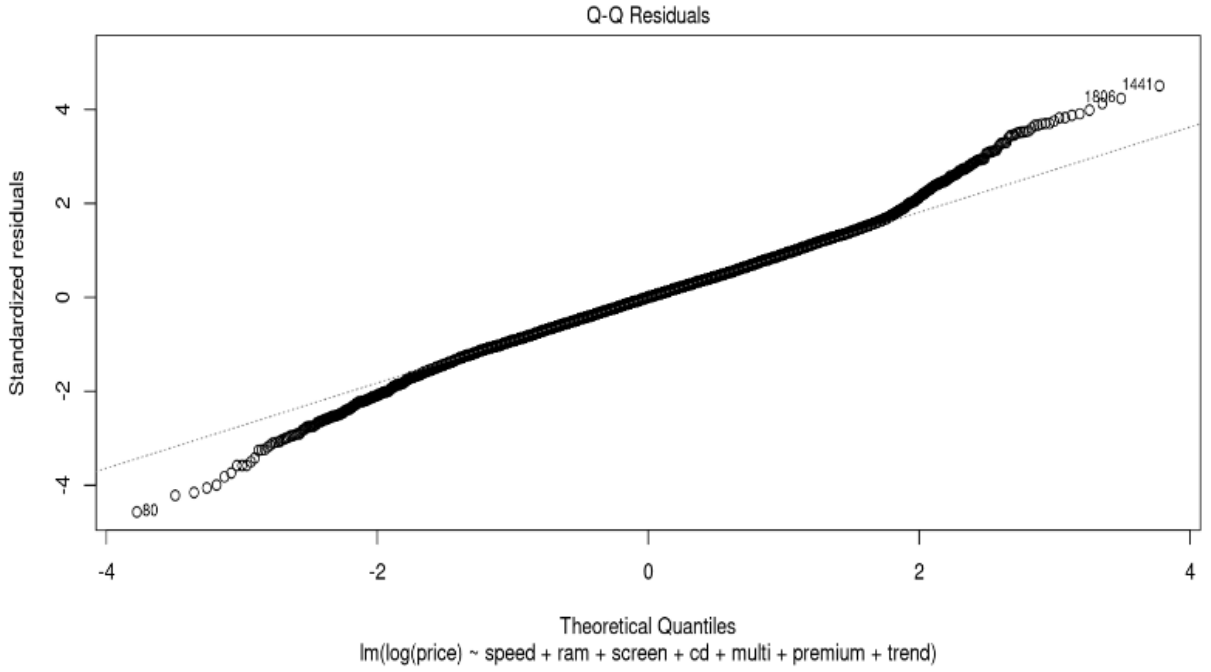

Рисунок 4. Отклонение остатков в модели с log(price) от нормального распределения

Однако же хвосты остатков все равно имеются, это связано в том числе с тем, что наши данные устроены специфически.

Теперь мы задались вопросом, а нет ли в нашей модели пропущенных квадратов, кубов и других степеней регрессоров?

Тест Рамсея показал, что в модели пропущены квадраты переменных ($p - value < 2.2 * 10^{-16}$). По графикам crPlots не видно явных нелинейных зависимостей:

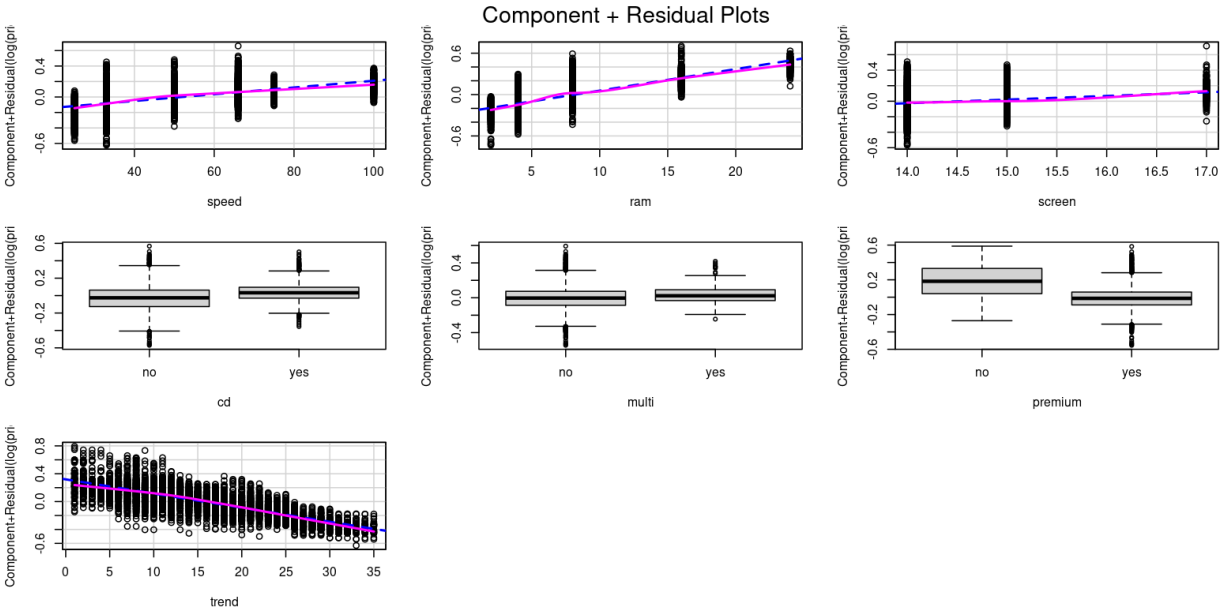

Рисунок 5. Характер зависимости между регрессором и зависимой переменной

Безусловно сложности анализу добавляет «скученность» распределения признаков. Например, у диагонали экрана только три различных показателя, у скорости процессора – шесть, у объема оперативной памяти – тоже шесть. Поэтому мы сознательно не стали добавлять квадраты в модель, поскольку:

1. Нет визуально обнаруживаемого отклонения остатков
2. Данные распределены очень неравномерно, что создает «эффект гармошки»
3. Интерпретация квадратов регрессоров затруднена. Сложно объяснить, что означает «квадрат скорости процессора» или «квадрат объема памяти»
4. Модель может быть усложнена без содержательного выигрыша в интерпретации

Поэтому для данной модели мы решили оставить линейную спецификацию как более прозрачную и экономически интерпретируемую, при этом, признавая ее ограничения.

Теперь посмотрим, есть ли в наших данных точки высокой напряженности, выбросы и влиятельные наблюдения. За высокую напряженность отвечает значение hatvalue, если оно превышает 2*k/n, то наблюдение считается точкой высокой напряженности, и если оно сочетается вместе с большим расстоянием Кука, то становится влиятельным и сильно влияет на оценки в нашей модели. На рисунке ниже изображены hatvalues:

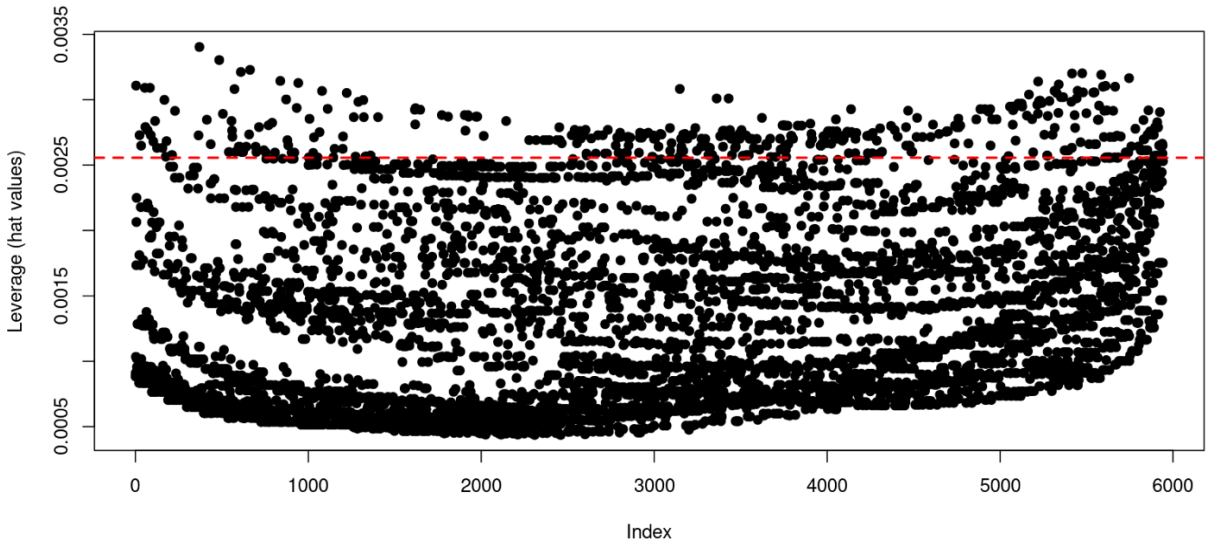

Рисунок 6. Hatvalues в модели с log(price)

На самом деле, 315 наблюдений имеют hatvalue выше порогового значения, соответственно мы их можем убрать и посмотреть, изменились ли оценки в нашей модели.

Таков график выбросов, точек высокой напряженности и влиятельных наблюдений:

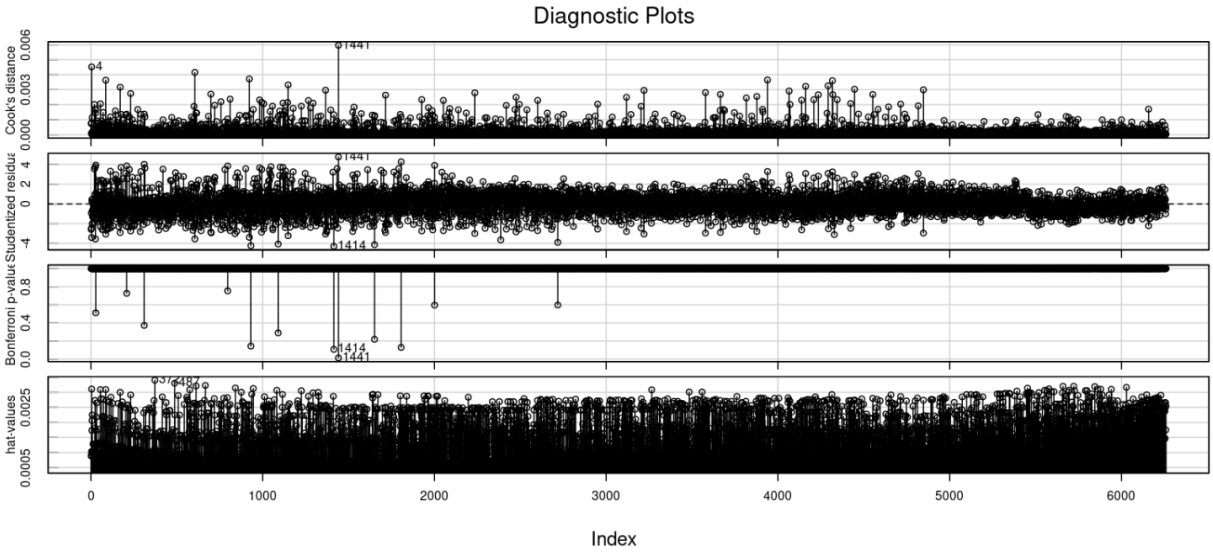

Рисунок 7. Выбросы, точки высокой напряженности и влиятельные наблюдения в модели с log(price)

Мы решили убрать из модели наблюдения с высоким hatvalues и расстоянием Кука и посмотреть, как изменились оценки коэффициентов в модели:

	Зависимая переменная: log(цена компьютера)	
Переменные \ Модель	Модель с log(price)	Модель с log(price) без влиятельных наблюдений и точек высокой напряженности
Частота	0.004^{***}	0.004^{***}

	(0.0001)	(0.0001)
Оперативная память	0.030***	0.031***
	(0.0003)	(0.0004)
Размер экрана	0.055***	0.047***
	(0.002)	(0.002)
Наличие CD-дисковода (да/нет)	0.069***	0.066***
	(0.004)	(0.005)
Наличие мультимедийного комплекта (да/нет)	0.034***	0.039***
	(0.005)	(0.006)
Премиум-бренд (да/нет)	-0.224***	-0.189***
	(0.006)	(0.007)
Месяц	-0.020***	-0.020***
	(0.0002)	(0.0003)
Константа	6.875***	6.956***
	(0.027)	(0.029)
Количество наблюдений	6,259	5,942
R^2	0.754	0.753
Скорректированный R^2	0.753	0.753
F-статистика	2,730.222***	2,588.824***

*Примечание: в скобках указаны стандартные ошибки. Уровни значимости: *p<0.1; **p<0.05; ***p<0.01*

Таблица 2. Устойчивость модели с log(price)

Мы видим, что оценки практически не изменились, значимость тоже везде сохранилась. Модель оказалась устойчива к нестандартным наблюдениям. Проведем тест Уайта на сравнение короткой регрессии против длинной. В результате получим, что F-статистика = 2730.2, p-value < $2{,}2 \times 10^{-16}$. Значит, длинная регрессия действительно лучше, и мы ее оставляем.

Осталось посмотреть, нет ли в наших ошибках гетероскедастичности. Тест Бреуша-Пагана показывает, что есть: p-value < $2{,}2 \times 10^{-16}$, а значит необходимо сделать модель с робастными стандартными ошибками. Возьмем их в форме Уайта. Так и поступим, выведем итоговую модель:

Переменные \ Модель	**log(цена компьютера)**
Частота	0.004***
	(0.0001)
Оперативная память	0.030***
	(0.0003)
Размер экрана	0.055***
	(0.002)
Наличие CD-дисковода (да/нет)	0.069***
	(0.004)
Наличие мультимедийного комплекта (да/нет)	0.034***
	(0.004)
Премиум-бренд (да/нет)	-0.224***
	(0.008)
Месяц	-0.020***

	(0.0002)
Константа	6.875***
	(0.025)
Количество наблюдений	6,259
R^2	0.754
Скорректированный R^2	0.753
F-статистика	2,730.222***

*Примечание: в скобках указаны стандартные ошибки. Уровни значимости: *p<0.1; **p<0.05; ***p<0.01*

Таблица 3. Модель с log(price) и робастными стандартными ошибками

$$\log(price) = 6.86 + 0.004 \times speed + 0.03 \times ram + 0.06 \times screen + 0.07 \times cd + 0.03 \times multi - 0.22 \times premium - 0.02 \times trend$$

Теперь, наконец, интерпретируем результаты модели. Поскольку зависимая переменная логарифмируется, а регрессоры нет, то изменение регрессора на 1 единицу меняет зависимую переменную на β%, где β – оценка коэффициента перед регрессором.

Интерпретация в % изменения цены:

– speed: Увеличение частоты на 1 МГц (скорости работы CPU) → цена растет на 0.4%;

– ram: Увеличение памяти на 1 МБ → цена растет на 3.0%;

– screen: Увеличение экрана на 1 дюйм → цена растет на 5.5%;

– cdyes: Наличие CD-привода → цена выше на 6.9%;

– multiyes: Наличие мультимедийного комплекта → цена выше на 3.4%;

– premiumyes: Премиальный бренд IBM, COMPAQ → цена ниже на 22.4%;

– trend: С каждым месяцем цена падает на 2.0% (технологическое устаревание).

Отметим, что все коэффициенты в модели значимы ($p < 0.01$), R-квадрат составляет около 75%, что отражает долю объясненной дисперсии в данной модели.

При прочих равных: чем выше технические характеристики компьютера, тем он дороже; чем новее компьютер, тем он дешевле (видимо меняются технологии производства). И крупные фирмы IBM, COMPAQ имеют более дешевые аналоги по сравнению со своими конкурентами.

2. В качестве второго датасета был выбран «DoctorVisits» из пакета «AER». Он представляет собой пространственные данные, полученные в ходе обследования состояния здоровья населения Австралии за 1977-1978 годы, 5190 наблюдений.

В датасете имеются следующие переменные:

– visits – количество посещений врача за последние 2 недели;

– gender – фактор, указывающий на пол;

– age – возраст в годах, деленный на 100;

– income – годовой доход в десятках тысяч долларов;

– illness – количество заболеваний за последние 2 недели;

– reduced – количество дней пониженной активности за последние 2 недели из-за болезни или травмы;

– health – оценка по опроснику общего состояния здоровья по методу Голдберга;

– private – наличие у этого человека частной медицинской страховки;

– freepoor – наличие бесплатной государственной медицинской страховки из-за низкого дохода;

– freerepat – наличие бесплатной государственной медицинской страховки по старости, инвалидности или статусу ветерана войны;

– nchronic – наличие хронических заболеваний, не ограничивающих активность;

– lchronic – наличие хронических заболеваний, ограничивающих активность.

Мы хотим построить регрессию, которая бы показывала зависимость количества посещений человеком врача за последние 2 недели от его различных характеристик.

Для начала мы решили включить все регрессоры:

$$visits \sim gender + age + income + illness + reduced + health + private + freepoor + freerepat + nchronic + lchronic$$

Предполагаемое влияние на количество визитов к врачу:

– gender – положительное (женщина=1) в связи с комплексом биологических, социальных и поведенческих факторов;

– age – положительное, чем старше, тем больше проблем со здоровьем;

– income – отрицательное, более высокий доход может означать лучшее здоровье и меньшую потребность в услугах, либо больший доступ к частным услугам;

– illness – положительное, большое количество заболеваний заставляет обследоваться;

– reduced – положительное, показывает степень серьезности заболевания на основе количества дней без активностей;

– health – положительное, чем выше данный показатель, тем хуже здоровье;

– private, freepoor, freerepat – положительное, наличие страховки - больше доступ к врачам;

– nchronic, lchronic – положительное, хронические заболевания требуют регулярного медицинского наблюдения.

Анализ показателя VIF не показал сильной мультиколлинеарности между регрессорами, для всех регрессоров он меньше 4. Уберем из регрессоров freerepat с максимальным из представленных VIF=2,45 из-за возможной мультиколлинеарности с age, а также nchronic, оставив другой показатель наличия хронических заболеваний. Наконец, рассмотрим модель без незначимых регрессоров private и lchronic:

	Зависимая переменная: количество визитов		
	(1)	(2)	(3)
Пол (женщина)	0.034	0.036*	0.038*
	(0.022)	(0.021)	(0.021)
Возраст	0.148**	0.183***	0.186***
	(0.067)	(0.054)	(0.053)
Доход	-0.056*	-0.061**	-0.052*
	(0.031)	(0.031)	(0.029)
Количество заболеваний	0.060***	0.061***	0.062***
	(0.008)	(0.008)	(0.008)
Дни пониженной активности	0.103***	0.103***	0.104***
	(0.004)	(0.004)	(0.004)
Оценка здоровья	0.017***	0.017***	0.018***
	(0.005)	(0.005)	(0.005)
Частная страховка	0.035	0.024	
	(0.025)	(0.021)	

Бесплатная страховка из-за низкого дохода	-0.103**	-0.112**	-0.119**
	(0.052)	(0.052)	(0.051)
Бесплатная страховка по статусу здоровья	0.033		
	(0.038)		
Хронические заболевания, не ограничивающие активность	0.005		
	(0.024)		
Хронические заболевания, ограничивающие активность	0.042	0.042	
	(0.036)	(0.032)	
Константа	0.036	0.037	0.042
	(0.036)	(0.036)	(0.035)
Количество наблюдений	5,190	5,190	5,190
R^2	0.202	0.202	0.201
Скорректированный R^2	0.200	0.200	0.200
F Статистика	119.029***	145.425***	186.519**

*Примечание: в скобках указаны стандартные ошибки. Уровни значимости: *p<0.1; **p<0.05; ***p<0.01.*

Таблица 4. Модели множественной регрессии для количества визитов врача (модель 1)

Все предполагаемые влияния подтвердились, за исключением freepoor: наличие бесплатной государственной медицинской страховки из-за низкого дохода, как оказалось, имеет отрицательное влияние на количество походов к врачу. Причины для этого могут быть разными, например, люди, получившие такие страховки, могут игнорировать симптомы болезни или жить в отдаленных районах.

Таким образом, вид модели:

$$visits \sim gender + age + income + illness + reduced + health + freepoor$$

При проведении теста Бокса-Кокса заметим, что λ находится близко к (-2), значит, стоит попробовать выполнить преобразование и представить зависимую переменную в виде $\frac{1}{y^2}$.

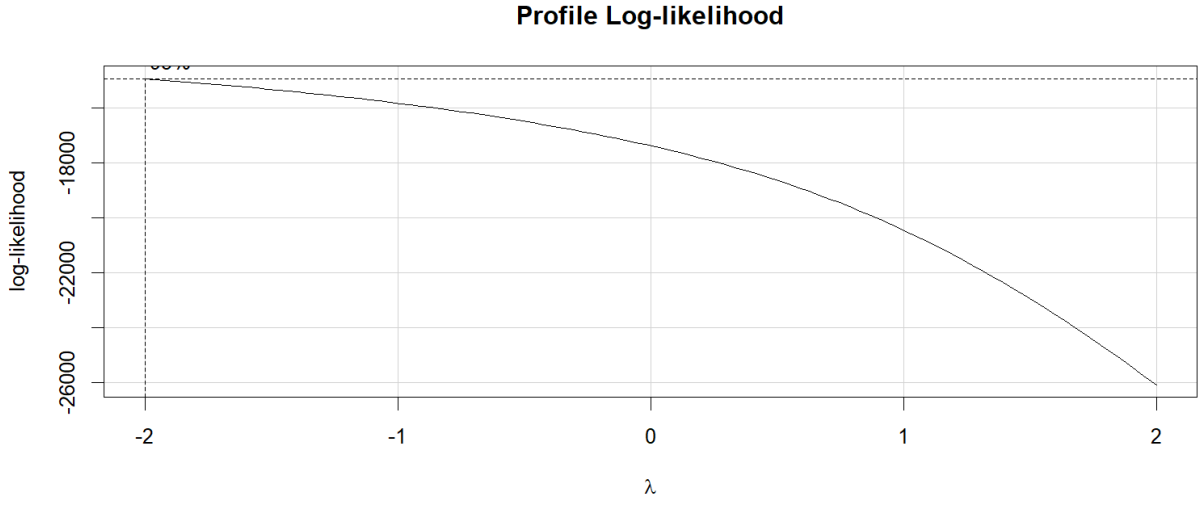

Рисунок 8. Функция максимума правдоподобия для λ (модель 1)

Анализ остатков показывает «тяжелые» хвосты:

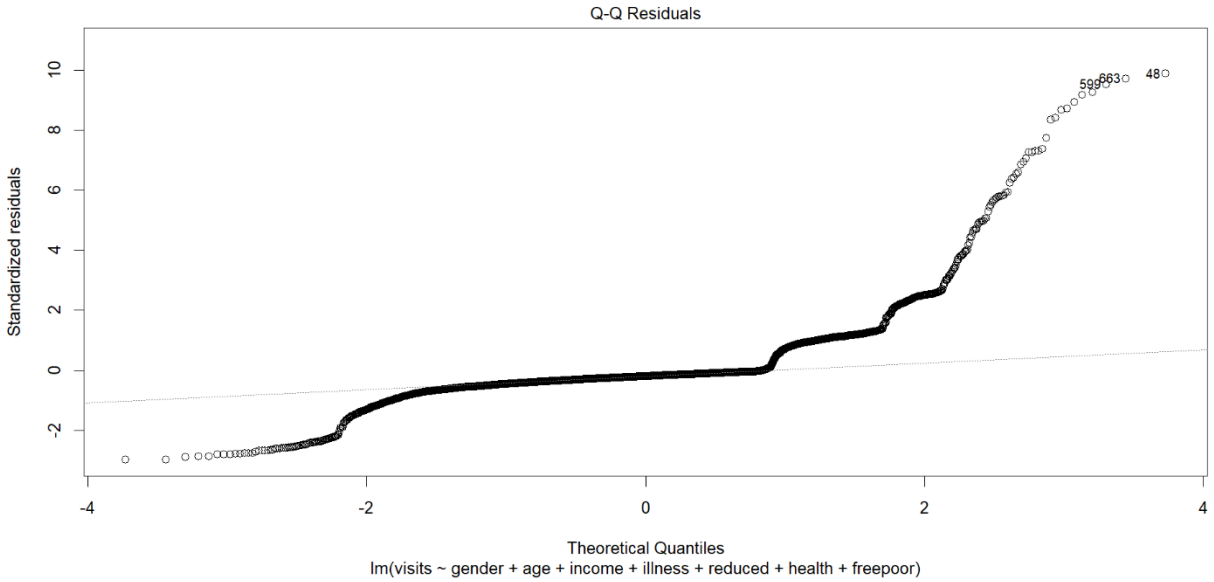

Рисунок 9. Отклонение остатков в модели 1 от нормального распределения

По тесту Бокса-Кокса λ в модели с преобразованием стала ближе к 1, чем к 0:

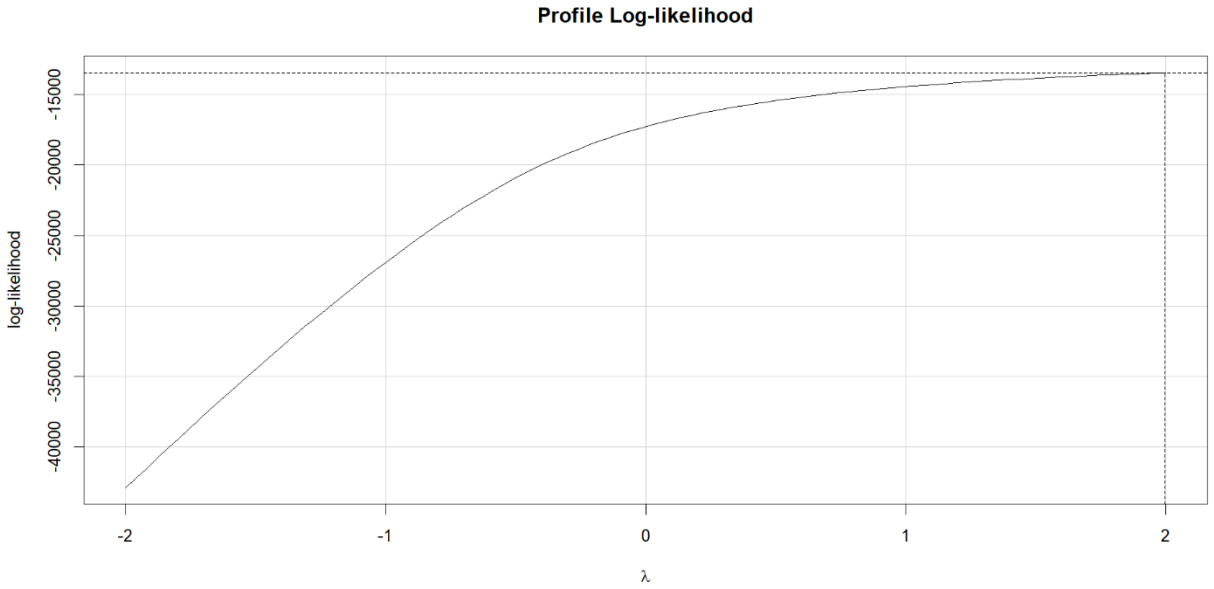

Рисунок 3. Функция максимума правдоподобия для λ (модель 2)

Однако по анализу остатков ситуация не улучшилась, к тому же новая модель с зависимой переменной в виде $\frac{1}{y^2}$ тяжело интерпретируема, так что было принято решение оставить зависимую переменную без преобразований.

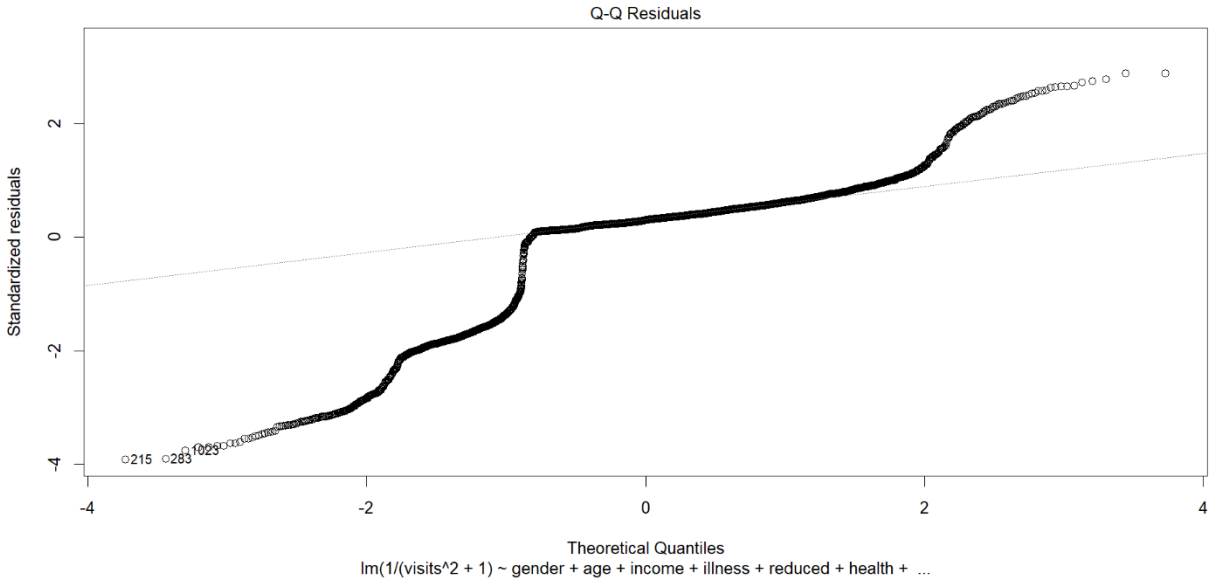

График 4. Отклонение остатков в модели 2 от нормального распределения

Тест Рамсея показал, что в модели пропущены квадраты переменных. По графикам crPlots зависимость скорее линейная, но не идеальная, возможны улучшения.

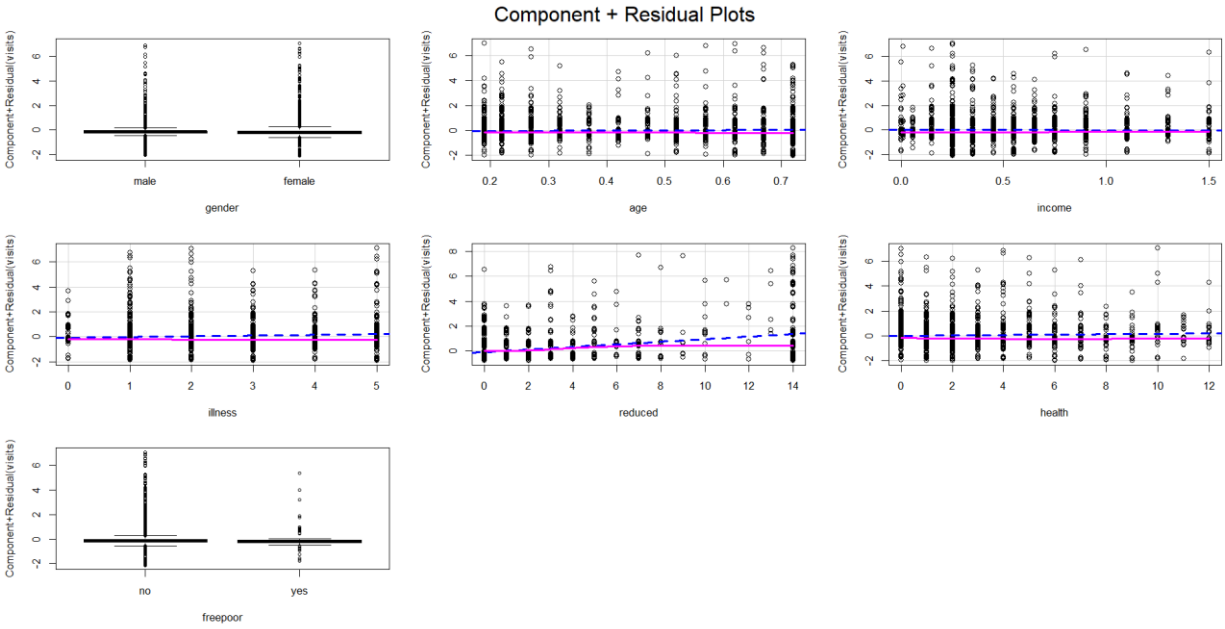

Рисунок 10. Характер зависимости между регрессором и зависимой переменной (модель 1)

Добавление квадрата reduced ситуацию немного улучшило, и он статистически значим. Вершина параболы ветвями вниз лежит между значениями 11 и 12. Это имеет теоретический смысл: до 11-12 дня нетрудоспособности каждый последующий день увеличивает визиты, а потом эффект ослабевает, так как люди могут перестать обращаться к врачам в связи с, например, отсутствием наблюдаемых улучшений по состоянию.

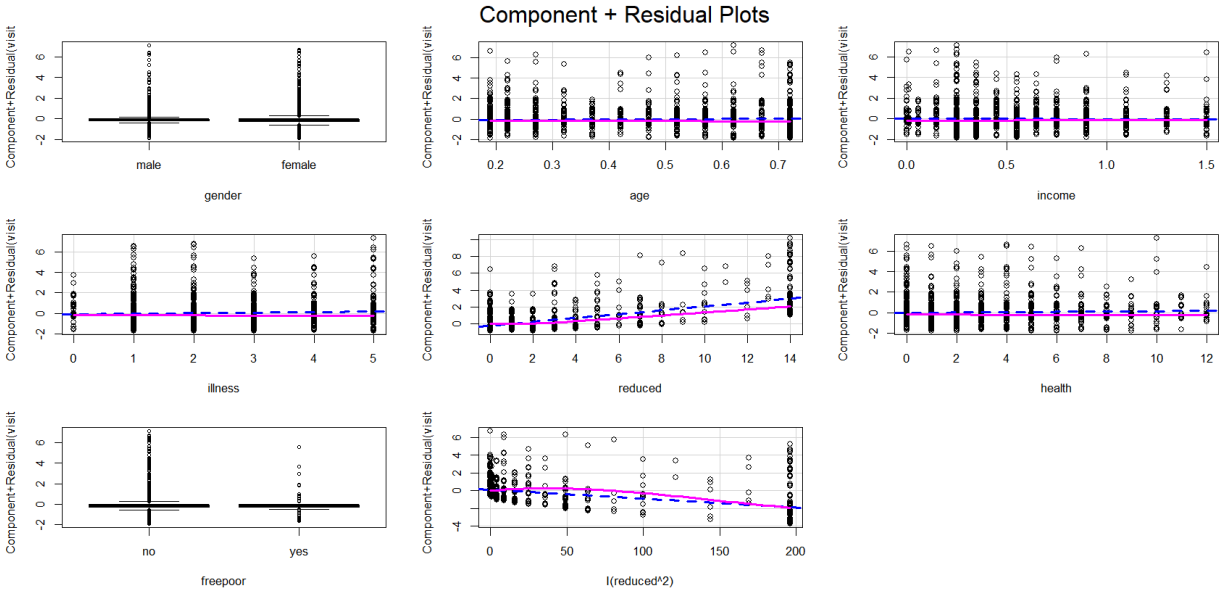

Рисунок 6. Характер зависимости между регрессором и зависимой переменной с новой переменной

Из-за того, что наши регрессоры могут принимать небольшое количество различных значений, мы можем наблюдать их распределение в виде «гармошки». Значит, имеющиеся сдвиги могут быть обоснованы не неверно подобранной спецификацией модели, а особенностью рассматриваемых данных.

Перейдем к проверке устойчивости к точкам высокой напряженности, выбросам и влиятельным наблюдениям. На рисунке ниже можем видеть hatvalues, находящиеся выше порогового значения $\frac{2k}{n}$, они являются точками высокой напряженности.

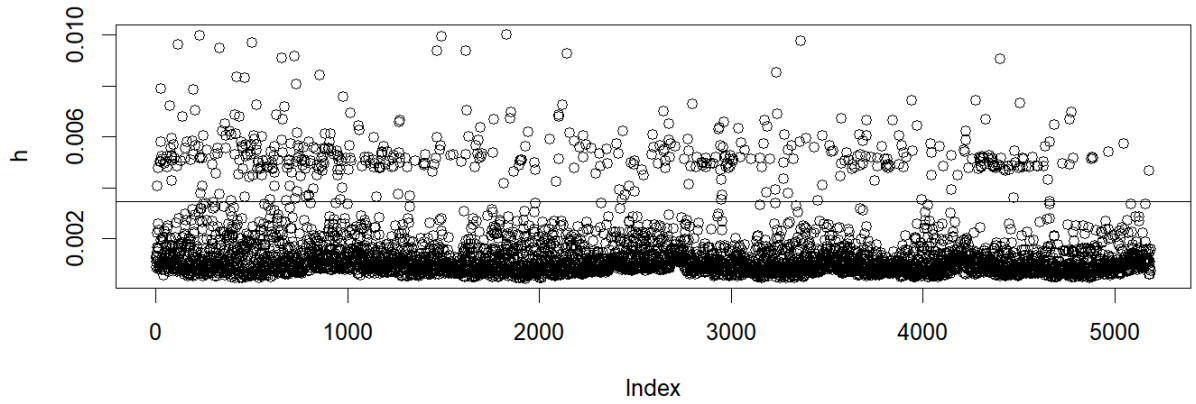

Рисунок 7. Hatvalues в модели 1

498 наблюдений имеют hatvalue выше порогового значения, уберем их и посмотрим, сохранятся ли полученные ранее результаты. После удаления точек высокой напряженности пропали все наблюдения с freepoor=’yes’ (наличие медицинского страхования из-за низкого дохода). Тогда уберем этот регрессор из модели на очищенных данных и сравним с такой же моделью на данных с точками высокой напряженности.

	Зависимая переменная: количество визитов	
	(1)	(2)
Пол (женщина)	0.036*	0.038**
	(0.021)	(0.018)

Возраст	0.261***	0.226***
	(0.052)	(0.046)
Доход	-0.042	-0.041
	(0.029)	(0.025)
Количество заболеваний	0.052***	0.049***
	(0.008)	(0.007)
Дни пониженной активности	0.229***	0.189***
	(0.015)	(0.020)
Оценка здоровья	0.018***	0.014**
	(0.005)	(0.006)
Дни пониженной активности^2	-0.010***	-0.002
	(0.001)	(0.003)
Константа	-0.004	0.021
	(0.034)	(0.030)
Количество наблюдений	5,190	4,692
R^2	0.212	0.158
Скорректированный R^2	0.211	0.157
F Статистика	199.431***	125.859***

Примечание: в скобках указаны стандартные ошибки. Уровни значимости: *p<0.1; **p<0.05; ***p<0.01

Таблица 5. Устойчивость модели

Из приведенной выше таблицы видно, что значимость некоторых коэффициентов изменилась. Для gender она повысилась, для health понизилась, а I(reduced^2) и вовсе стал незначимым. Таким образом, можем сказать, что только age, income, illness и reduced оказались устойчивы к точкам высокой напряженности и не подстраиваются под них.

Далее проведем тест Вальда на сравнение короткой регрессии против длинной. Он показал, что F-статистика=175.35, p-value < $2{,}2 \times 10^{-16}$, значит, набор выбранных переменных совместно объясняет значительную часть вариации в количестве визитов к врачу, и длинная регрессия действительно лучше. Проверим также наличие гетероскедастичности с помощью теста Бреуша-Пагана. Он показал, что p-value < $2{,}2 \times 10^{-16}$, значит, нам необходимо использовать робастные стандартные ошибки. Тогда итоговая модель выглядит так:

	Зависимая переменная: количество визитов
Пол (женщина)	0.033
	(0.022)
Возраст	0.237***
	(0.052)
Доход	-0.057*
	(0.030)

Количество заболеваний	0.051^{***}
	(0.010)
Дни пониженной активности	0.229^{***}
	(0.030)
Оценка здоровья	0.019^{***}
	(0.007)
freepooryes	-0.119^{***}
	(0.046)
Дни пониженной активности^2	-0.010^{***}
	(0.002)
Константа	0.022
	(0.033)
Количество наблюдений	5,190
R^2	0.213
Скорректированный R^2	0.212
F Статистика	175.348^{***}

Примечание: в скобках указаны стандартные ошибки. Уровни значимости: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Таблица 6. Итоговая модель с робастными стандартными ошибками

Перейдем к интерпретации результатов:

1. Gender: женщины совершают в среднем на 0.033 больше визитов к врачу, чем мужчины, при прочих равных, однако эффект стал статистически не значим после модификации модели (добавления I(reduced^2))
2. Age: при увеличении возраста на 1 год в среднем при прочих равных число посещений врача вырастает на 0.00237 (так как age – это возраст, деленный на 100), эффект значим на 1% уровне.
3. Income: с ростом дохода на 10000 долларов (income в десятках тысяч) количество визитов уменьшается на 0.057, эффект значим на 10% уровне.
4. Illness: каждое дополнительное заболевание увеличивает количество визитов на 0.051, эффект значим на 1% уровне.
5. Reduced, I(reduced^2): каждый дополнительный день сниженной активности увеличивает визиты на 0.229, однако такой эффект наблюдается до 11.73544 дня, далее он идет на спад, значим на 1% уровне.
6. Health: ухудшение показателя здоровья на 1 балл увеличивает визиты на 0.019, эффект значим на 1% уровне.
7. Freepoor: люди с бесплатной государственной страховкой для малоимущих совершают на 0.119 меньше визитов, чем люди без такой страховки. Парадоксальный, но значимый эффект на 1% уровне.

Часть 2

Из данных Exam мы отобрали случайным образом половину школ. Нам нужно построить регрессию, где:

- Зависимая переменная normexam — стандартизированный экзаменационный балл каждого ученика. Этот показатель рассчитывается как z-оценка: исходный балл по экзамену минус средний балл по выборке, делённый на стандартное отклонение. То есть, normexam показывает на сколько баллов (в стандартных отклонениях) результат ученика выше или ниже среднего результата в группе;

- Регрессор standLRT — стандартизированный балл по тесту способностей. Также переводится в z-оценку, отражая насколько тестовый результат ученика отличается от среднего по выборке;
- Регрессор sex – бинарная переменная пола.

Мы построим несколько моделей со случайными свободными членами и/или наклонами (объект для случайных эффектов – школа), запишем уравнение для каждой оцененной модели со случайными эффектами и выберем лучшую спецификацию.

Модель 1 – это модель только со случайной «константой». Мы оцениваем следующую регрессию: $normexam \sim standLRT + sex + (1 \mid school)$. Последнее слагаемое означает, что наша «константа» может быть для каждого наблюдения своей (т.е. случайной). Суть — определить, как экзаменационная успеваемость (normexam) объясняется результатами теста способностей и полом ученика с помощью уравнения:

$$normexam = \beta_0 + \beta_1 \times standLRT + \beta_2 \times sex + \varepsilon$$

Здесь β_0 — свободный коэффициент (intercept), β_1 и β_2 — значения эффекта каждого признака. Получились такие результаты:

	Экзаменационный балл		
Регрессоры	*Оценки*	*CI*	*p-value*
Константа	0.08	-0.04 – 0.20	0.196
Балл по тесту способностей	0.57	0.53 – 0.60	<0.001
Пол [M]	-0.19	-0.29 – -0.10	<0.001
Random Effects			
Внутригрупповая дисперсия σ^2	0.58		
Межгрупповая дисперсия $\tau_{00\ school}$	0.10		
Внутриклассовый коэффициент корреляции ICC	0.14		
Количество школ	32		
Количество наблюдений	2139		
Marginal R^2 / Conditional R^2	0.323 / 0.420		

Таблица 7. Модель 1 со случайной константой

И сразу проверим, стал ли результат лучше, когда мы включаем случайные «константы» в регрессию или нет. Когда из регрессии убрали «случайную» константу, хуже стал критерий AIC (было 4986, стало 5165), а $p - value < 2.2 * 10^{-16}$ для LRT оказался меньше 0.05. Значит принимаем гипотезу о том, что случайная «константа» нужна.

Тогда регрессия принимает следующий вид: $normexami = \beta i + 0.57 * standLRTi - 0.19 * sexi$, причем β_i распределены нормально с параметрами (0, 0.09).

Предоставим визуализацию этих случайных «констант» β_i:

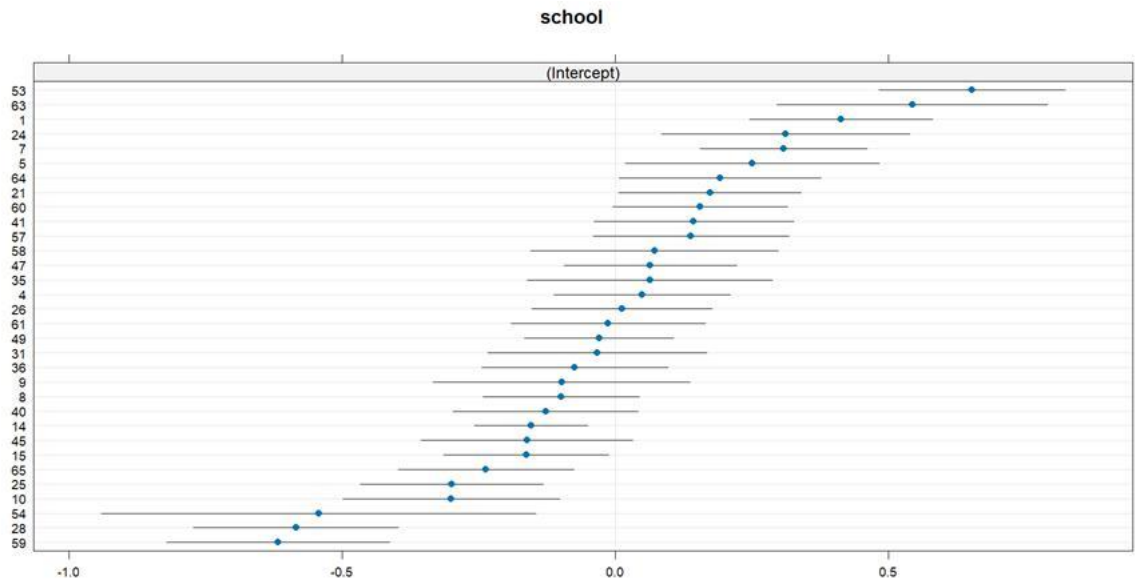

Рисунок 1. Визуализация случайных констант для Модели 1

Мы поняли, что в модели должны присутствовать случайные константы. Осталось понять, будет ли там случайный наклон или нет. Один из регрессоров – это пол человека, бинарная переменная, поэтому случайный наклон к ней не построить. Значит, будем проверять, становится ли модель лучше, если добавить случайный наклон к переменной standLRT.

Для этого посмотрим сначала, как ведут себя остатки модели 1. Как видно, тяжелые хвосты отсутствуют и ситуация с остатками хорошая, они распределены близко к нормальному распределению.

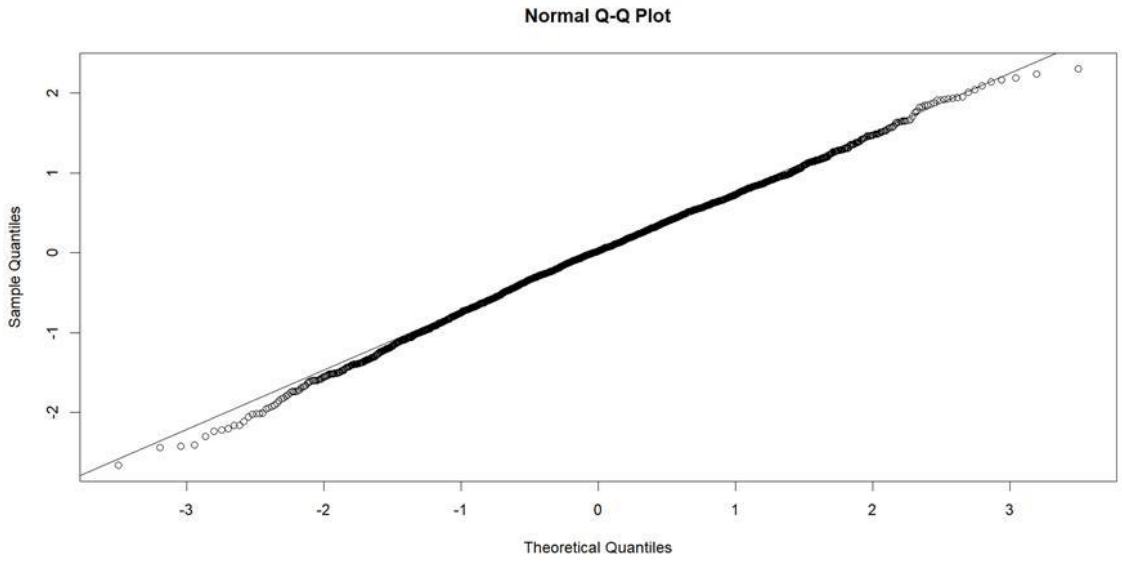

Рисунок 2. QQ-Plot для остатков Модели 1

Проверим, нет ли тренда в самих остатках? Тренда нигде не наблюдается (точки рассеяны как облако везде).

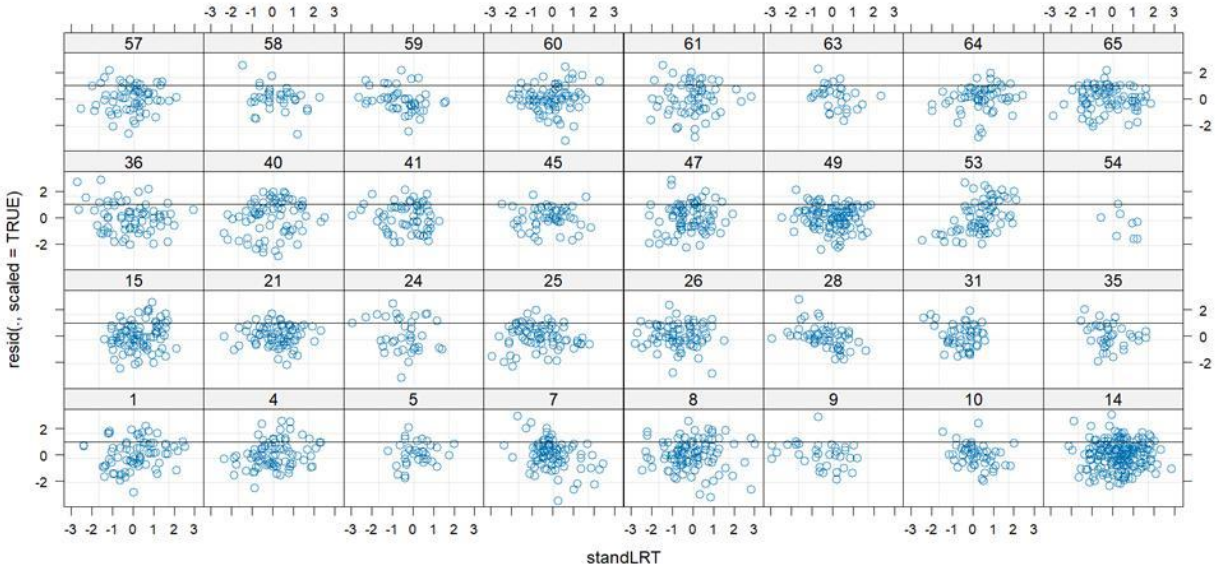

Рисунок 3. Наличие трендов для остатков Модели 1

Из этого можно сделать вывод, что случайные наклоны не нужны. И тем не менее, проверим спецификацию модели со случайными наклонами сначала без корреляции, а потом с корреляцией.

Делаем **Модель 2** со случайными наклонами без корреляции:

$$normexami = \beta_{0i} + \beta_{1i} \times standLRT + \beta_2 \times sex + \varepsilon$$

	Экзаменационный балл		
Регрессоры	*Оценка*	*CI*	*p-value*
Константа	0.07	-0.05 – 0.19	0.275
Балл по тесту способностей	0.55	0.49 – 0.61	<0.001
Пол [M]	-0.20	-0.29 – -0.10	<0.001
Random Effects			
Внутригрупповая дисперсия σ^2	0.56		
Дисперсия случайной константы τ_{00} school	0.09		
Дисперсия случайного наклона τ_{11} school.standLRT	0.02		
Внутриклассовый коэффициент корреляции ICC	0.14		
Количество школ	32		
Количество наблюдений	2139		
Marginal R^2 / Conditional R^2	0.316 / 0.414		

Таблица 8. Модель 2 со случайной константами и наклонами без корреляции

Сразу проверим, стала ли модель лучше после включения случайных наклонов. Из результатов теста видно, что игнорирование как случайной константы, так и случайного наклона ухудшает нашу регрессию (потому что в обоих случаях p-value меньше уровня значимости). Значит, случайные наклоны нашей модели нужны.

Посмотрим, как поменялись остатки модели 2. Хуже точно не стало, хвостов тяжелых нет, распределение остатков близко к нормальному. Трендов в самих остатках нет.

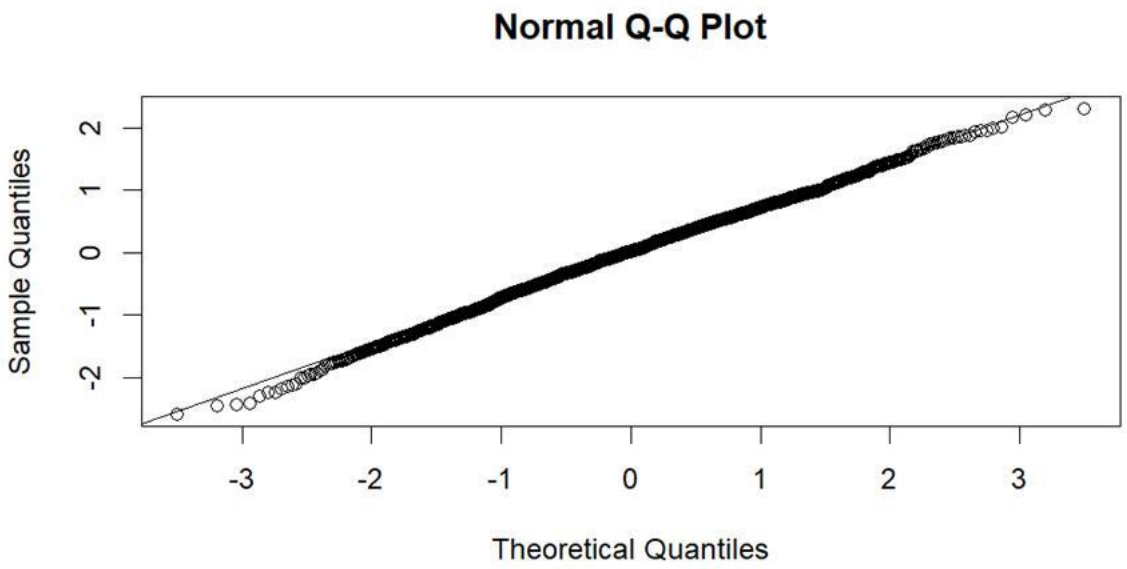

Рисунок 4. QQ-Plot для остатков Модели 2

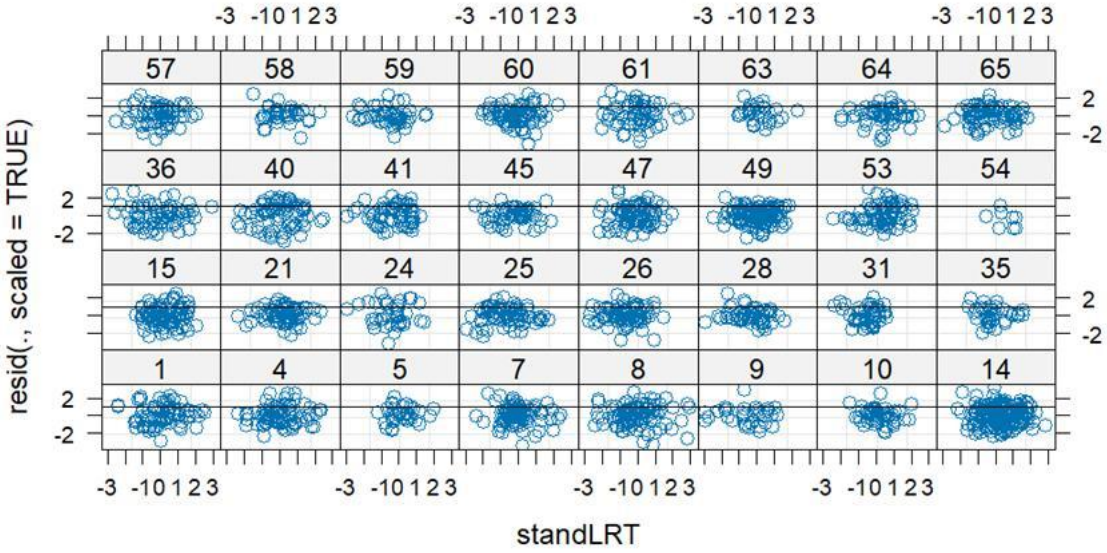

Рисунок 5. Наличие трендов для остатков Модели 2

Также покажем, как выглядят случайные эффекты для Модели 2:

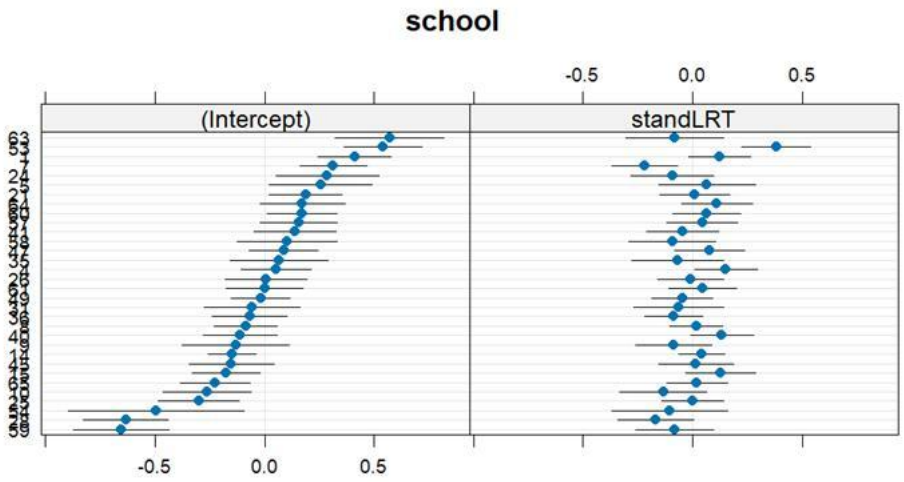

Рисунок 6. Визуализация случайных констант и случайных наклонов для Модели 2

Теперь осталось проверить **Модель 3**, которая включает в себя случайные наклоны с корреляцией:

$$normexami = \beta_{0i} + \beta_{1i} \times standLRT + \beta_2 \times sex + \varepsilon$$

	Экзаменационный балл		
Регрессоры	*Оценки*	*CI*	*p-value*
Константа	0.06	-0.06 – 0.18	0.296
Балл по тесту способностей	0.55	0.49 – 0.61	<0.001
Пол [M]	-0.20	-0.30 – -0.11	<0.001
Random Effects			
Внутригрупповая дисперсия σ^2	0.56		
Дисперсия случайной константы $\tau_{00\ school}$	0.10		
Дисперсия случайного налокна $\tau_{11\ school.standLRT}$	0.02		
Корреляция слцчайных эффектов $\rho_{01\ school}$	0.54		
Внутриклассовый коэффициент корреляции ICC	0.17		
Количество школ	32		
Количество наблюдений	2139		
Marginal R^2 / Conditional R^2	0.310 / 0.427		

Таблица 9. Модель 3 со случайными константами и наклонами без корреляции

При этом корреляция случайных эффектов равняется 0.54. Проверим, стала ли модель с коррелированными случайными наклонами лучше, чем предыдущий вариант. Видим, что $p - value$ для $LRT = 0.019 < 0.05$. Значит отвергаем гипотезу об отсутствии корреляции и принимаем Модель 3.

Вот визуализация эффектов для Модели 3:

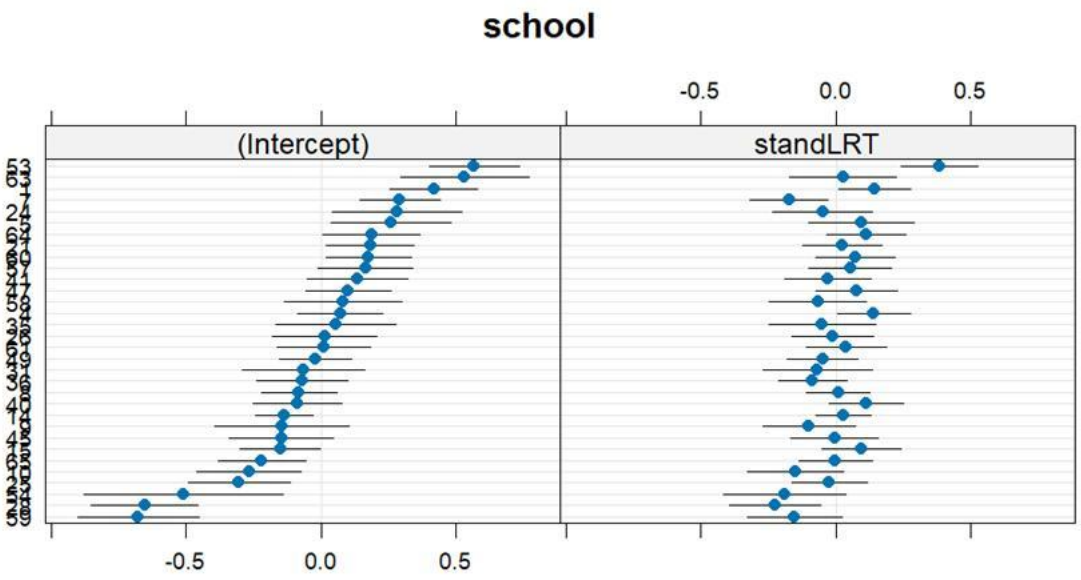

Рисунок 7. Визуализация случайных эффектов для Модели 3

Итого, мы получаем следующую модель:

$$normexami = \beta 1i + \beta 2i * standLRTi + \beta 3 * sexi + epsi, \text{где}$$

- $eps \sim N(0, 0.559)$;
- $\beta 1i \sim N(0, 0.095)$;
- $\beta 2i = 0.55 + \theta i, \text{где } \theta i \sim N(0, 0.021)$;
- $\beta 3 = -0.2$.

Приведем ее интепретацию:

1. Матожидание случайной константы $\beta 1i$ в модели равно 0, что ожидаемо, поскольку переменная теста способностей standLRT стандартизирована. Поскольку переменные в модели стандартизированы, коэффициенты регрессии следует интерпретировать как изменение зависимой переменной в единицах стандартного отклонения при изменении регрессора на одно стандартное отклонение;

2. $\beta 2i$ перед standLRT (0.55) означает, что при увеличении стандартизированного балла теста способностей на 1 стандартного отклонения, экзаменационный балл увеличивается в среднем на 0.55 стандартного отклонения. То есть начальные способности сильно предсказывают будущие академические успехи;

3. $\beta 3$ перед переменной пола (-0.2) означает, что мальчики в среднем получают на 0.2 стандартного отклонения меньше баллов, чем девочки, при прочих равных условиях;

4. Корреляция случайных эффектов ρ = 0.54: в школах с более высокими средними связь между начальными способностями и экзаменационными результатами обычно больше (более крутые наклоны). Это означает, что «сильные» школы не только дают лучшие результаты в среднем, но и лучше развивают потенциал способных учеников;

5. Качество модели. Модель с фиксированный эффектами имеет $Marginal\ R^2 = 0.31$. Другими словами, 31% вариации результатов объясняется фиксированными эффектами. Модель со случайными константами и наклоном имеет $Conditional\ R^2 = 0.427$. Другими словами, 42.7% вариации объясняется всей моделью (фиксированные + случайные эффекты). Получается, что случайные эффекты школ добавляют примерно 11.7% объясненной вариации.