НИУ ВШЭ, ОП «Политология», 2022
Курс «Теория вероятностей и математическая статистика»

Домашнее задание
Рубанов Владислав БПТ 201

Задача 1.

```
library(haven)

lab <- read_dta('hwdata.dta')

# почистим наши данные
labhw <- dplyr::select(lab, -c(province,uezd,serfperc1,redist))
labhw <- na.omit(labhw)

m1 <- lm(ch_schools_pc ~ afreq + nozemstvo + distance_moscow + goodsoil + lnurban + lnpopn + p
summary(m1)

##
## Call:
## lm(formula = ch_schools_pc ~ afreq + nozemstvo + distance_moscow +
##     goodsoil + lnurban + lnpopn + province_capital, data = labhw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33848 -0.09634 -0.03551  0.04695  1.45921
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.683238   0.218271   3.130 0.001853 **
## afreq            -0.181470   0.054400  -3.336 0.000916 ***
## nozemstvo         0.080091   0.021793   3.675 0.000264 ***
## distance_moscow  -0.012096   0.031894  -0.379 0.704660
## goodsoil         -0.008286   0.023997  -0.345 0.730028
## lnurban           0.013287   0.007274   1.827 0.068371 .
## lnpopn           -0.042304   0.019890  -2.127 0.033937 *
## province_capital  0.040362   0.030172   1.338 0.181621
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1782 on 481 degrees of freedom
## Multiple R-squared:  0.1018,Adjusted R-squared:  0.08873
## F-statistic: 7.788 on 7 and 481 DF,  p-value: 6.081e-09
```

Прокомментируем полученные результаты:

1. Общее описание полученных оценок:

   (a) $\hat{\beta}_0 = 0.683238 \Rightarrow$ Среднее значение зависимой переменной («Изменение в количестве сельских школ с 1860 до 1880 гг. на душу сельского населения уезда») при условии равенства всех предикторов 0 в среднем равно 0.683238.

(b) $\hat{\beta}_1 = -0.181470 \Rightarrow$ при увеличении значения предиктора («Доля лет между 1851 и 1863 гг., в которые были зафиксированы крестьянские выступления») на 1 ед. изм., значение зависимой переменной (изменение числа сельских школ на душу сельского населения уезда) в среднем уменьшится на 0.181470, при прочих равных условиях.

(c) $\hat{\beta}_2 = 0.080091 \Rightarrow$ при увеличении значения предиктора («Бинарная переменная: Единицей закодированы уезды тех губерний, в которых в результате реформы 1864 года земства созданы не были, 0 – в противном случае») на 1 ед. изм., значение зависимой переменной (изменение числа сельских школ на душу сельского населения уезда) в среднем увеличится на 0.080091, при прочих равных условиях.

(d) $\hat{\beta}_3 = -0.012096 \Rightarrow$ при увеличении значения предиктора («Расстояние от Москвы до центра уезда») на 1 ед. изм., значение зависимой переменной (изменение числа сельских школ на душу сельского населения уезда) в среднем уменьшится на 0.012096, при прочих равных условиях.

(e) $\hat{\beta}_4 = -0.008286 \Rightarrow$ при увеличении значения предиктора («Показатель плодородности почвы») на 1 ед. изм., значение зависимой переменной (изменение числа сельских школ на душу сельского населения уезда) в среднем уменьшится на 0.008286, при прочих равных условиях.

(f) $\hat{\beta}_5 = 0.013287 \Rightarrow$ при увеличении значения предиктора («Логарифм городского населения уезда на 1863 г.») на 1 ед. изм., значение зависимой переменной (изменение числа сельских школ на душу сельского населения уезда) в среднем увеличится на 0.013287, при прочих равных условиях.

(g) $\hat{\beta}_6 = -0.042304 \Rightarrow$ при увеличении значения предиктора («Логарифм населения уезда на 1863 г.») на 1 ед. изм., значение зависимой переменной (изменение числа сельских школ на душу сельского населения уезда) в среднем уменьшится на 0.042304, при прочих равных условиях.

(h) $\hat{\beta}_7 = 0.040362 \Rightarrow$ при увеличении значения предиктора («Бинарная переменная: принимает значение 1, если в уезде находился «столичный» город губернии, 0 – в противном случае.») на 1 ед. изм., значение зависимой переменной (изменение числа сельских школ на душу сельского населения уезда) в среднем увеличится на 0.040362, при прочих равных условиях.

2. Более частные выводы:

(a) Как можно заметить из выдачи, три переменных (nozemstvo, lnurban, province_capital) были оценены как имеющие **положительную связь с откликом**, а четыре других (afreq, distance_moscow, goodsoil, lnpopn) — как имеющие **отрицательную связь**.
Это означает, что с создание земства в уезде, увеличение доли городского населения в уезде и тот факт, находился ли в уезде «столичный» город губернии в среднем положительно влияли на перераспределение благ в пользу народа (в виде финансирования публичных благ), а частота крестьянских выступлений, увеличение расстояния от Москвы до центра уезда, увеличение показателя плодородности почвы и увеличение числа населения уезда в среднем негативно влияли на перераспределение благ в пользу народа.

(b) Если говорить о значимости полученных коэффициентов, нужно отметить, что мы строим следующие гипотезы относительно них:

$$H_0 : \beta_0 = 0$$
$$H_1 : \beta_0 \neq 0$$

и так для каждого из предикторов.

Так, можно утверждать, что **лишь 3 из 7 оценок** коэффициентов при предикторах (а также оценка константы ($\hat{\beta}_0$)) **статистически значимы** на конвенциональном уровне значимости 0.05: а именно частота крестьянских выступлений (p-value = 0.000916), создание земства в уезде (p-value = 0.000264) и логарифм населения уезда на 1863 г. (p-value = 0.033937)

(c) В целом, информация из предыдущего пункта логична, т. к., по замыслу исследователей, именно переменная о частоте крестьянских выступлений является ключевой (по ней строится основная содержательная гипотеза), а остальные переменные являются контрольными — это заметно, учитывая тот факт, что у двух из них p-value > 0.7. Хотя, в теории, можно содержательно предположить, что переменной «создание земства в уезде» также можно объяснить относительно много вариации зависимой переменной, что может подтверждать ее статистическая значимость на довольно большом уровне (p-value ≈ 0.0003)
При этом, оценка коэф. при предикторе логарифма населения уезда на 1863 г. имеет пограничное значение p-value ≈ 0.034, что говорит о возможной необходимости дополнительных проверок на устойчивость (у логарифма городского населения p-value $= 0.068$ также находится на пограничном уровне, хотя и отвергается на уровне 0.05).

(d) Также направления связей из п. 1 интуитивно понятны, учитывая контекст исследования: Российская империя сер. XIX века с абсолютной монархией в качестве политического строя. Политические элиты того времени (в особенности, дворянство) точно не стремились перераспределять блага в пользу крестьян, в особенности, в тех уездах, где происходили крестьянские восстания.
Можно предположить, что таким образом власти желали «наказать» наиболее «буйные» уезды и «поощрить» наиболее «спокойные».

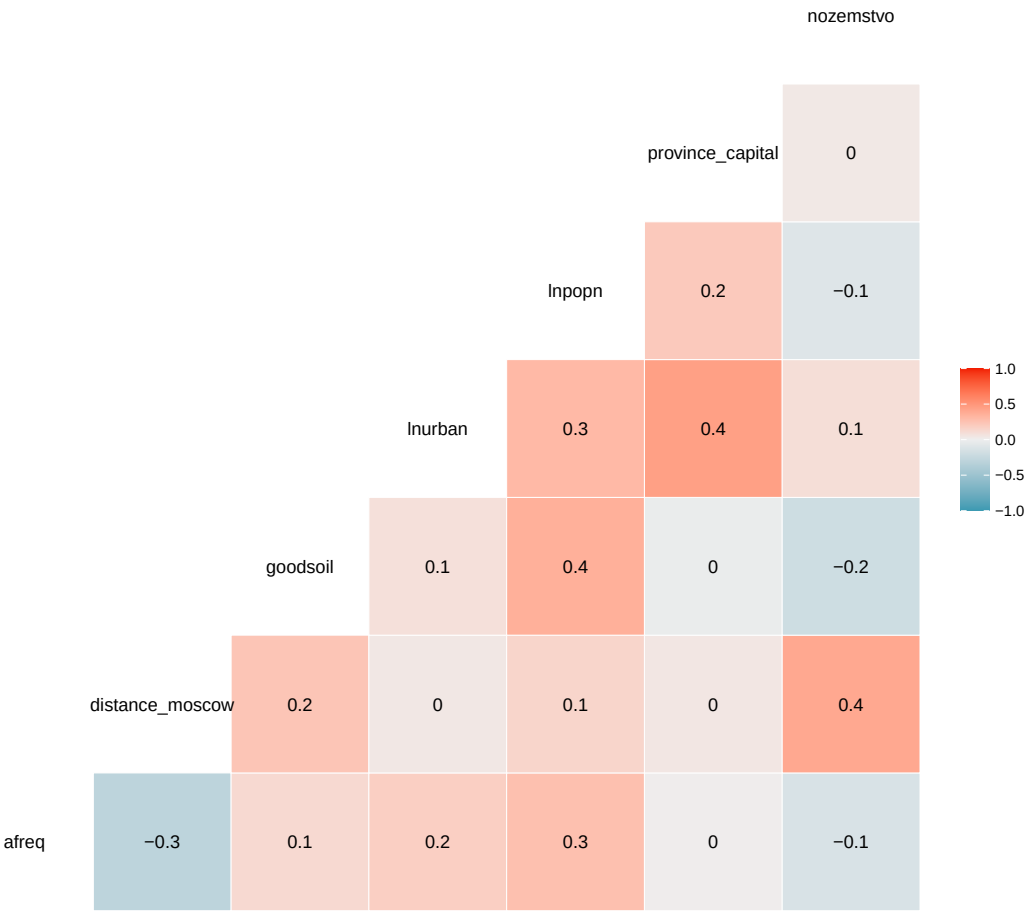
(e) Помимо этого, можно отметить, что R^2 нашей модели $= 0.1018$, а также он статистически значим (p-value: 6.081e-09 при альтер. гипотезе $H_1 : R^2 > 0$).
Это говорит о том, что **доля объясненной вариации зависимой переменной** нашей моделью составила около **0.1**. Модель смогла объяснить лишь 10% информации (вариации), что является довольно маленьким показателем.

Задача 2.

1. Визуальные диагностики и корреляционная матрица

```
library("ggplot2")
library("GGally")

## Error in library("GGally"): there is no package called 'GGally'

ggpairs(labhw[,-c(1, 8)])

## Error in ggpairs(labhw[, -c(1, 8)]): could not find function "ggpairs"

ggcorr(labhw[,-c(1, 8)], label = TRUE)

## Error in ggcorr(labhw[, -c(1, 8)], label = TRUE): could not find function "ggcorr"
```

**Вставлю рисунки вручную, т. к. LaTeXне хочет считывать некоторые пакеты и компилировать их сам.*




```
corrmatrix <- cor(labhw[,-c(1, 8)])
corrmatrix
```

```
##                      afreq distance_moscow    goodsoil    lnurban     lnpopn
## afreq             1.00000000     -0.30349155  0.11876758 0.17853805  0.2628479
## distance_moscow  -0.30349155      1.00000000  0.23544149 0.04206744  0.1467256
## goodsoil          0.11876758      0.23544149  1.00000000 0.08360277  0.3532937
## lnurban           0.17853805      0.04206744  0.08360277 1.00000000  0.3051746
## lnpopn            0.26284789      0.14672561  0.35329373 0.30517463  1.0000000
## province_capital  0.01009203      0.04557890 -0.02597862 0.44572143  0.2107133
## nozemstvo        -0.13738560      0.39978045 -0.20422244 0.09441004 -0.1006247
##                  province_capital   nozemstvo
## afreq                 0.01009203 -0.13738560
## distance_moscow       0.04557890  0.39978045
## goodsoil             -0.02597862 -0.20422244
## lnurban               0.44572143  0.09441004
## lnpopn                0.21071332 -0.10062471
## province_capital      1.00000000  0.03600985
## nozemstvo             0.03600985  1.00000000
```

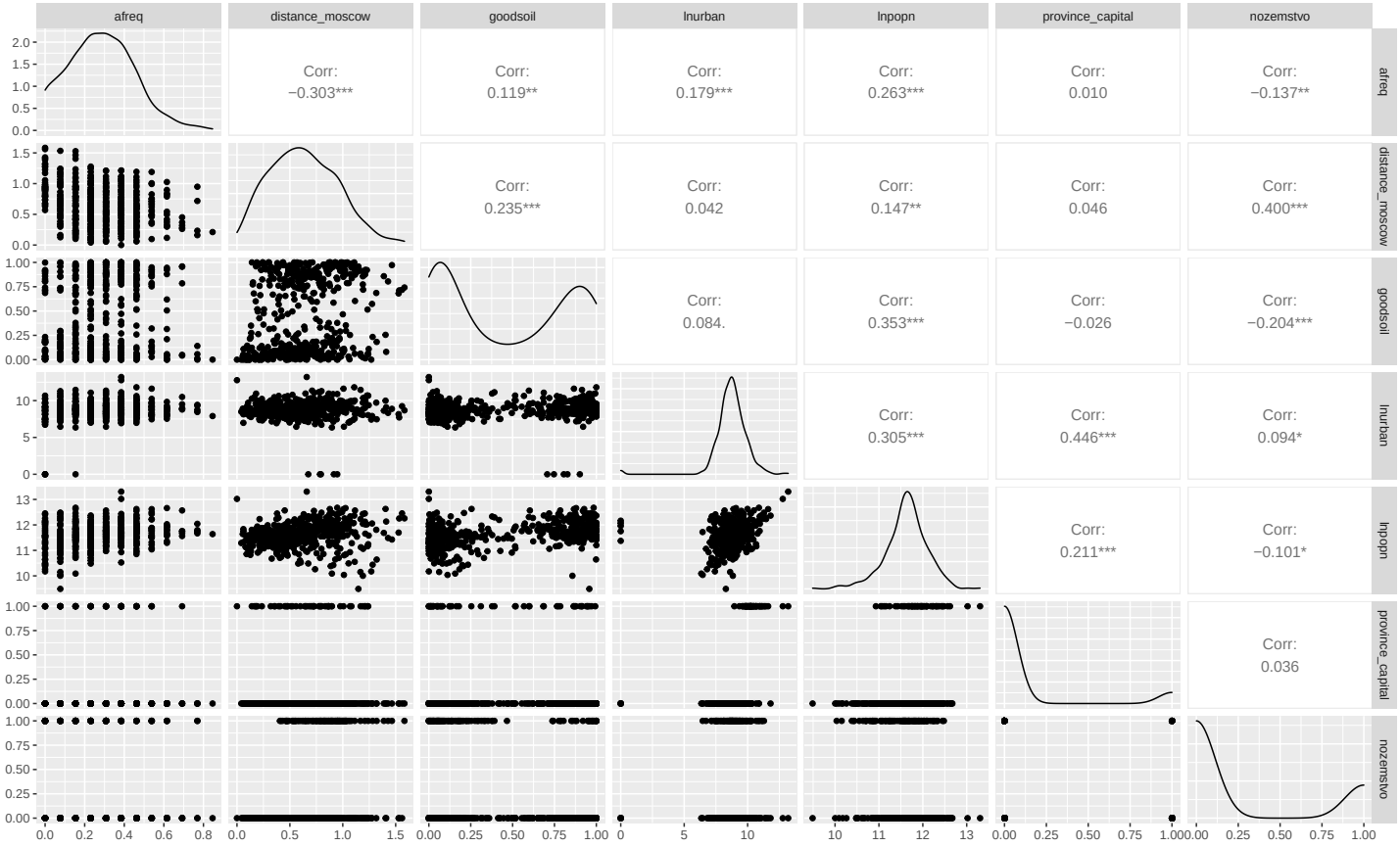
Как наглядно видно из рисунка (и не так наглядно из выдачи — корреляционной матрицы предикторов), максимальная корреляция (≈ 0.4) между предикторами была обнаружена между:

- плодородностью почвы (goodsoil) и логарифмом населения уезда (npopn);

- расстоянием от Москвы (distance_moscow) и фактом создания земств (nozemstvo); прим.: земства были созданы только в европейской части Российской империи;
- нахождением в уезде столичного города (province_capital) и логарифмом городского населения уезда (lnurban); без округления до десятых корреляция для этих предикторов даже выше, чем 0.4: она равна 0.45.

В целом, все эти корреляции логически объяснимы, их даже можно обосновать исторически.

Таким образом, формально можно утверждать, что угрозы мультиколлинеарности в нашей модели нет, т. к. конвенциональным порогом для значения корреляции двух предикторов между собой является $Cor \approx 0.6$. Однако данное значение чаще определяется экспертным образом, поэтому посмотрим не только на сухие цифры, но и на графики распределения наших предикторов.



По главной диагонали мы можем увидеть графики распределения предикторов (вместо единиц в корреляционной матрице). В верхней части таблицы находятся те же коэффициенты корреляции, которые мы обсуждали ранее, а в нижней части — диаграммы рассеяния предикторов.

Некоторые распределения двух предикторов выглядят достаточно странно из-за того, что один из них (или оба) могут быть закодированы бинарно, либо (как в случае с долей восстаний), принимать фиксированные значения на определенном интервале.

Тем не менее, для всех предикторов можно сказать, что на их диаграммах рассеяния действительно **нельзя проложить некоторую линию, вокруг которой они были бы плотно сосредоточены**, что говорило бы об угрозе сильной мультиколлинеарности.

Однако из этого тезиса выбивается лишь диаграмма рассеяния логарифма городского населения уезда и логарифма населения уезда. Формальное значение корреляции для них = 0.305,

но, возможно, такое низкое значение, визуально не соответствующее распределению, связано с несколькими группами нетипичных наблюдений слева, внизу и в правом верхнем углу. Но, даже без них вряд ли бы значение корреляции оставшегося «облака» было бы выше 0.6

2. Формальный тест (VIF)

   Расчитаем в R значения VIF («фактора вздутия дисперсии»).

```
library(car)

## Loading required package: carData

vif <- vif(m1)
vif

##            afreq        nozemstvo  distance_moscow         goodsoil
##         1.279557         1.383755         1.583747         1.342736
##          lnurban           lnpopn province_capital
##         1.380490         1.377546         1.286367
```

   Напомню, что значение VIF считается как $\frac{1}{1-R_j^2}$ и означает то, во сколько раз увеличится значение дисперсии оценки коэф. при предикторе при условии того, что предикторы в модели являются скоррелированными (мультиколлинеарными), по сравнению с тем, если бы было столько же предикторов, но они не были бы скоррелированы и при прочих равных условиях.

   Таким образом, помня, что пороговым значением VIF является 10 и выше, можно утверждать, что **наша модель не подвержена сильной мультиколлинеарности**: ни одно значение VIF не превышает даже «2». Само по себе (небольшое, слабое) наличие мультиколлинеарности не является проблемой, т. к. иначе предикторы не смогут объяснить зависимую переменную, это неизбежно.

Итак, благодаря визуальным диагностикам, корреляционной матрице предикторов и фактору «вздутия» дисперсии (VIF) мы выяснили, что наша **модель не подвержена сильной мультиколлинеарности**.

Задача 3.

Действительно, существуют теоретические основания полагать, что данная модель страдает от гетероскедастичности.

Перечислю свои аргументы на теоретическом уровне:

1. По сути, из-за разных распределений предикторов, мы имеем не единую выборку, а несколько выборок с разными распределениями. Об этом свидетельствует тот факт, что некоторые переменные закодированы бинарно (nozemstvo, province_capital); другие представляют собой логарифмы (lnurban, lnpopn), хотя именно они наиболее должны быть приближены к нормальному распределнию; другие являются долей лет; некоторые (как показатель плодородности) должны иметь наибольшую плотность для наибольших и наименьших значений, в отличие от близких к нормальному (предположительно) распределению логарифмов.

   Естественным образом, при совмещении в одном датасете, они будут представлять собой **неоднородные данные**, что является **одной из основных причин** гетероскедастичности.

2. **На содержательном уровне**, вполне логично ожидать гетероскедастичности, учитывая контекст исследования. Так, очевидно, что некоторые переменные качественно связаны с географическими и социально-экономическими показателями, такими как плодородность почвы, расстояние от Москвы, логарифмы городского населения и населения в целом. Очевидно, что подобные показатели не будут представлять собой относительно равномерное распределение. К примеру, равно отдаленные уезды из северной части страны (нынешние Архангельская область, Республика Карелия, Новгородская область и Ленинградская область) и уезды из южной части страны (территории современных Волгоградской области, Ростовской области, Ставропольского края, Краснодарского края) будут диаметрально отличаться с т. з. плодородности почвы, однако находиться примерно на одном расстоянии от Москвы. То же самое можно сказать и про логарифм городского населения: в северных частях, очевидно, оно будет больше, а в южных меньше — в силу распределения труда в разных регионах. Это, опять же, порождает большую неоднородность данных

3. Также от представленной спецификации можно ожидать такое нарушение, как **значимые пропущенные переменные**. Поскольку большая часть предикторов несут довольно общий характер, можно ожидать, что такие факторы могут не до конца качественно разделить представленные данные и создать некоторые «скопления», возникающие по причине того, что авторы не учли какой-либо из важных факторов, обуславливающих зависимую переменную. Такие пропущенные переменные будут не только вызывать гетероскедастичность, но и, к тому же, являться причиной смещенности в оценках (*omitted variable bias*).

4. Также можно ожидать **нелинейную форму взаимосвязи** в представленных данных. Так, сложно ожидать того, что представленные переменные (такие, как логарифмы городского населения, плодородность почвы, расстояние от Москвы, доля лет, в которые были зафиксированы крестьянские выступления) будут описываться линейной формой взаимосвязи. По моему мнению, вероятность того, что их влияние на зависимую переменную будет линейным крайне мало (это также следует из предыдущих тезисов). Таким образом, попытка оценивания нелинейной формы взаимосвязи линейной множественной регрессионной моделью может привести к гетероскедастичности.

5. Помимо этого, учитывая тот факт, что данные описывают сложный социальный мир (к тому же, более 150 лет назад), я считаю, что в данных можно ожидать довольно большое число **нетипичных наблюдений**, которые могут быть либо точечными, либо формировать некоторые группы нетипичных значений. Наверняка в Российской империи были регионы, кардинально отличающиеся своим социально-экономическим развитием, которое «не вписывалось» бы в контекст соседних регионов: к примеру, особенности заложенные традицией, историей, расположением или личностным фактором.
   К тому же, нельзя исключать *технические ошибки*, ведь данные действительно имеют большой возраст, часть из них могла быть утеряна, либо некорректно записана, не говоря уже об умышленном искажении данных российской бюрократией XIX века.

6. Последним тезисом я хотел бы выделить угрозу **различных методик сбора данных**. Так, можно заметить, что представленные переменные относятся к разным временным периодам: крестьянские выступления — между 1851 и 1863 гг., логарифмы городского населения и населения уездов в целом — за 1863 г., изменение в количестве сельских школ — с 1860 по 1880 гг., а также они взяты из разных источников (архивов). Следовательно, при совмещении, они могут некорректно налагаться друг на друга, что также может представлять одну из угроз для появления гетероскедастичности.

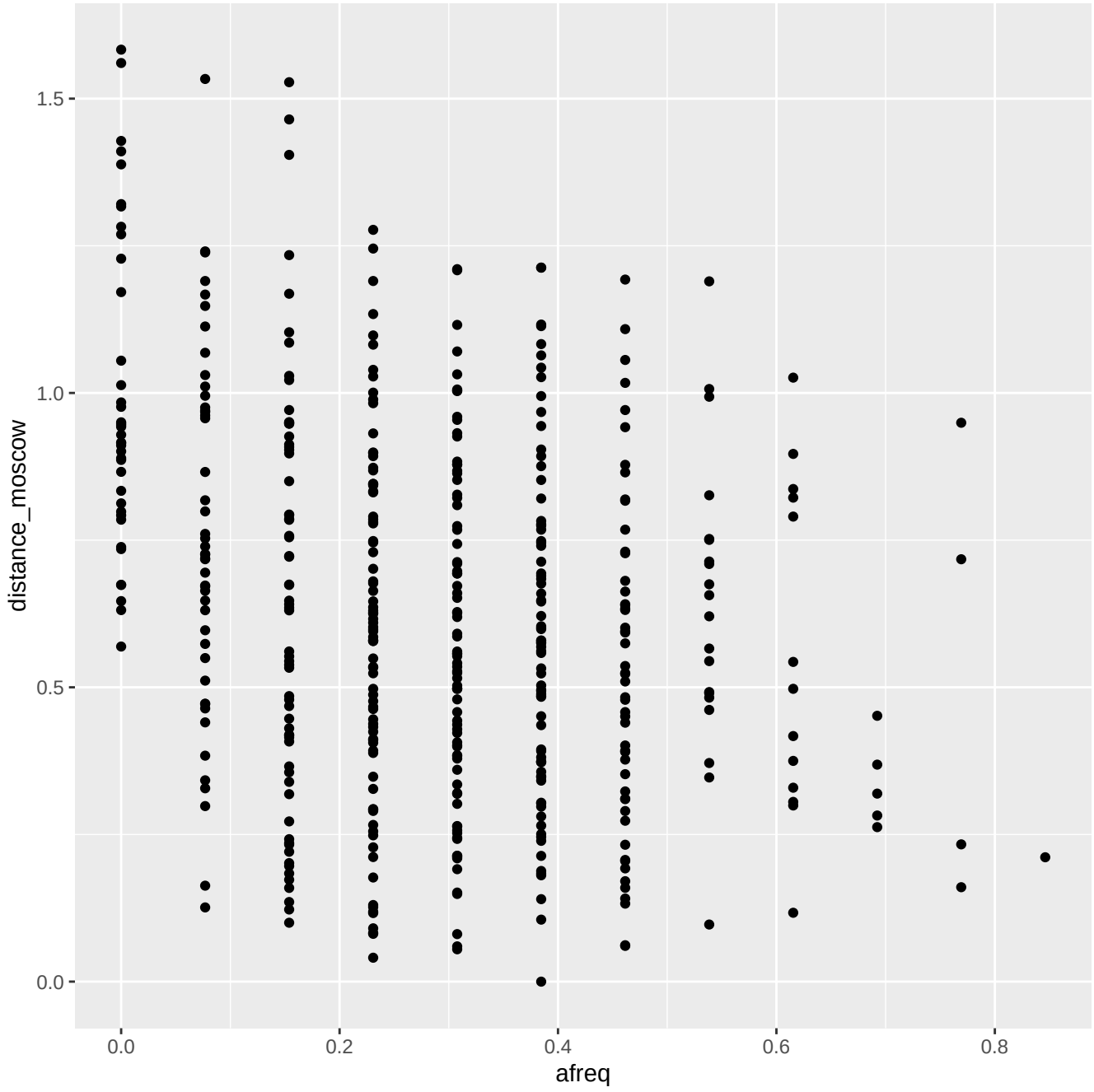
Задача 4.

1. **Визуальные способы определения гетероскедастичности.**

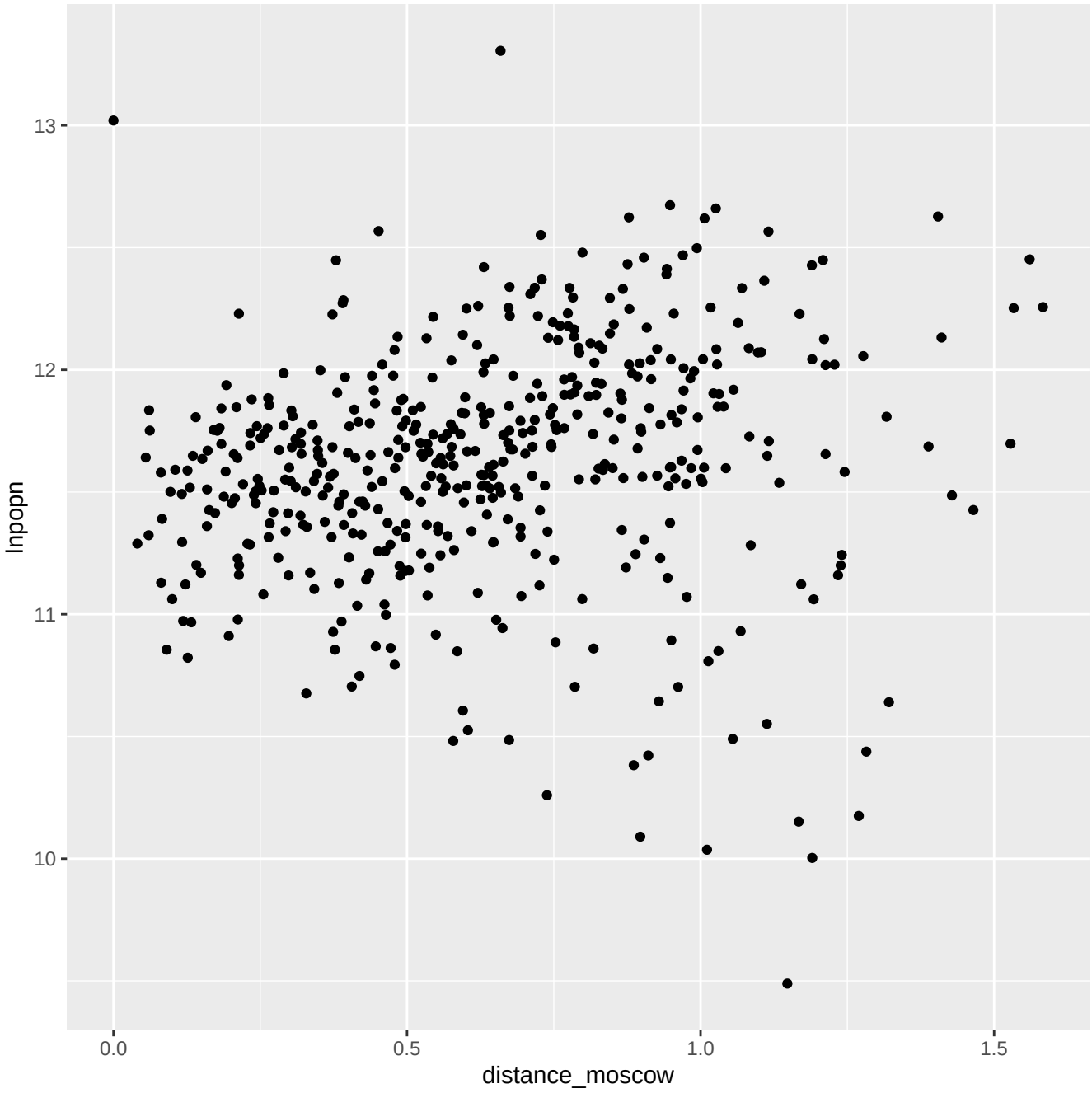
   В начале построим диаграммы рассеяния (*scatterplots*) для разных регрессоров, от которых (теоретически) можно ожидать гетероскедастичность при «взаимодействии» (подробнее о теор. предпосылках в предыдущем задании).

   Сделаем это для большинства «подозрительных предикторов», хотя, если гетероскедастичность будет подтверждена хотя бы для одной переменной (в особенности ключевой), уже можно говорить говорить о гетероскедастичности относительно всей модели.

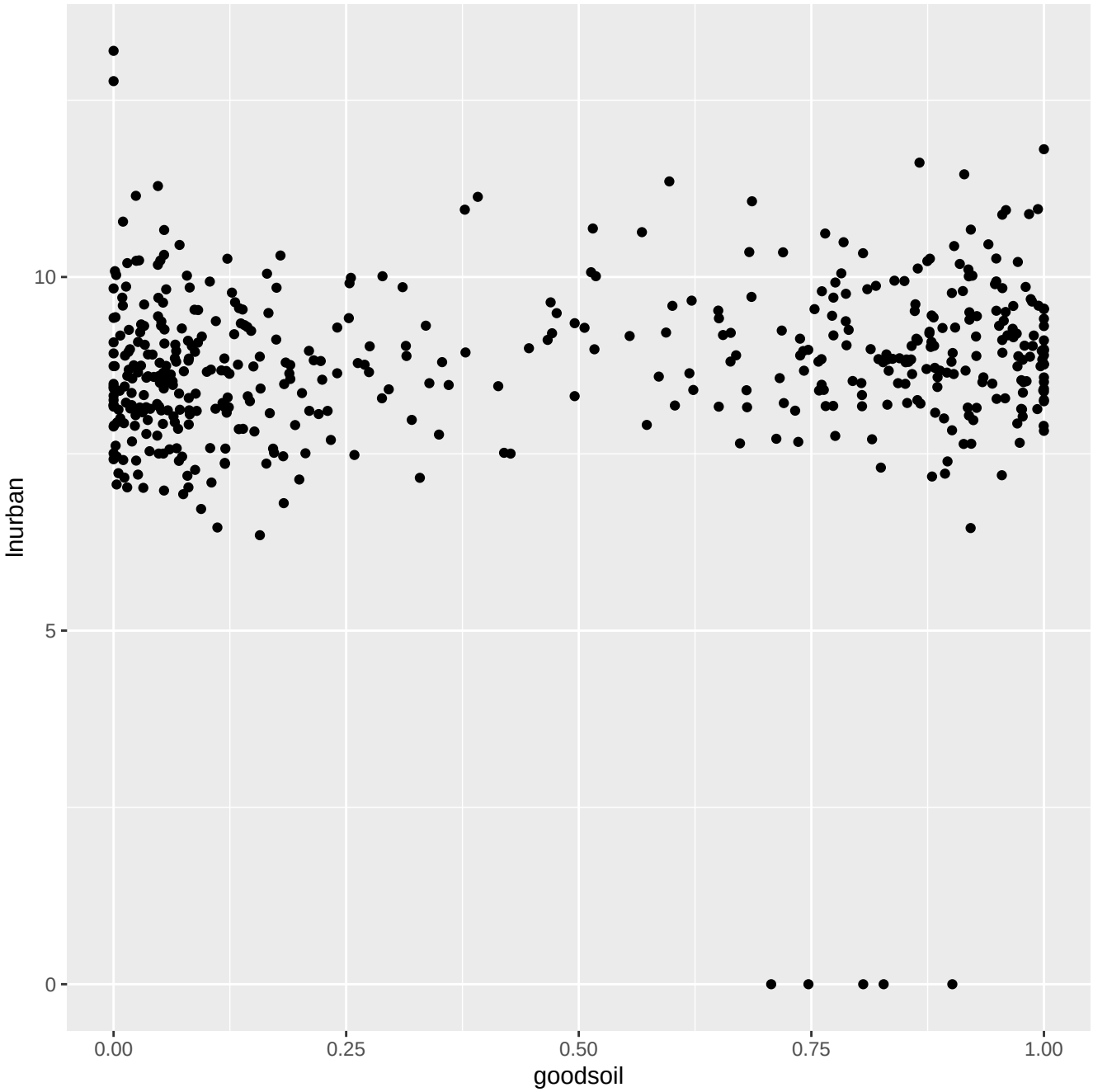
```
library(ggplot2)

ggplot(labhw, aes(afreq, distance_moscow)) + geom_point()
```



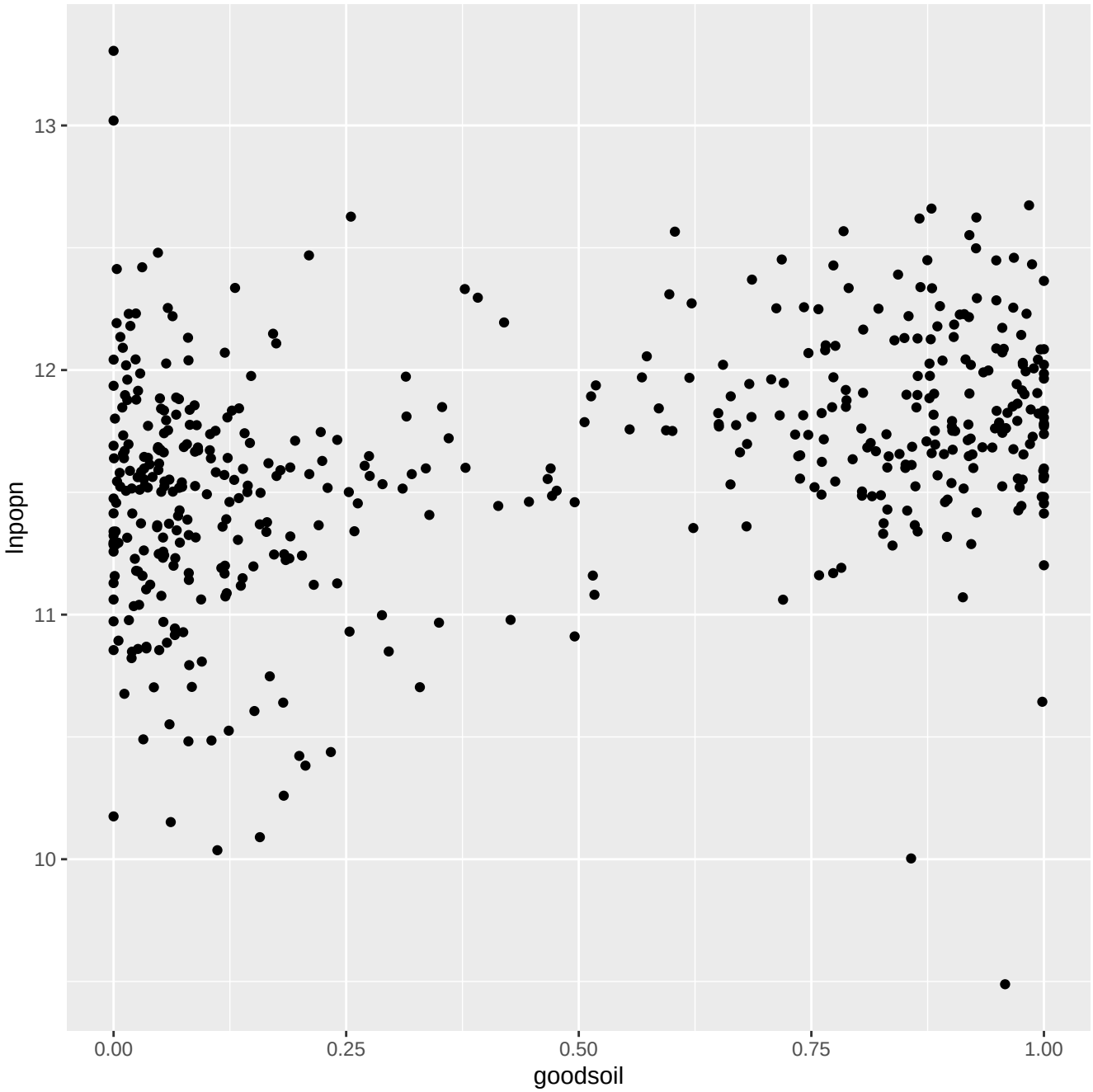

```
ggplot(labhw, aes(distance_moscow, lnpopn)) + geom_point()
```




```
ggplot(labhw, aes(goodsoil, lnurban)) + geom_point()
```



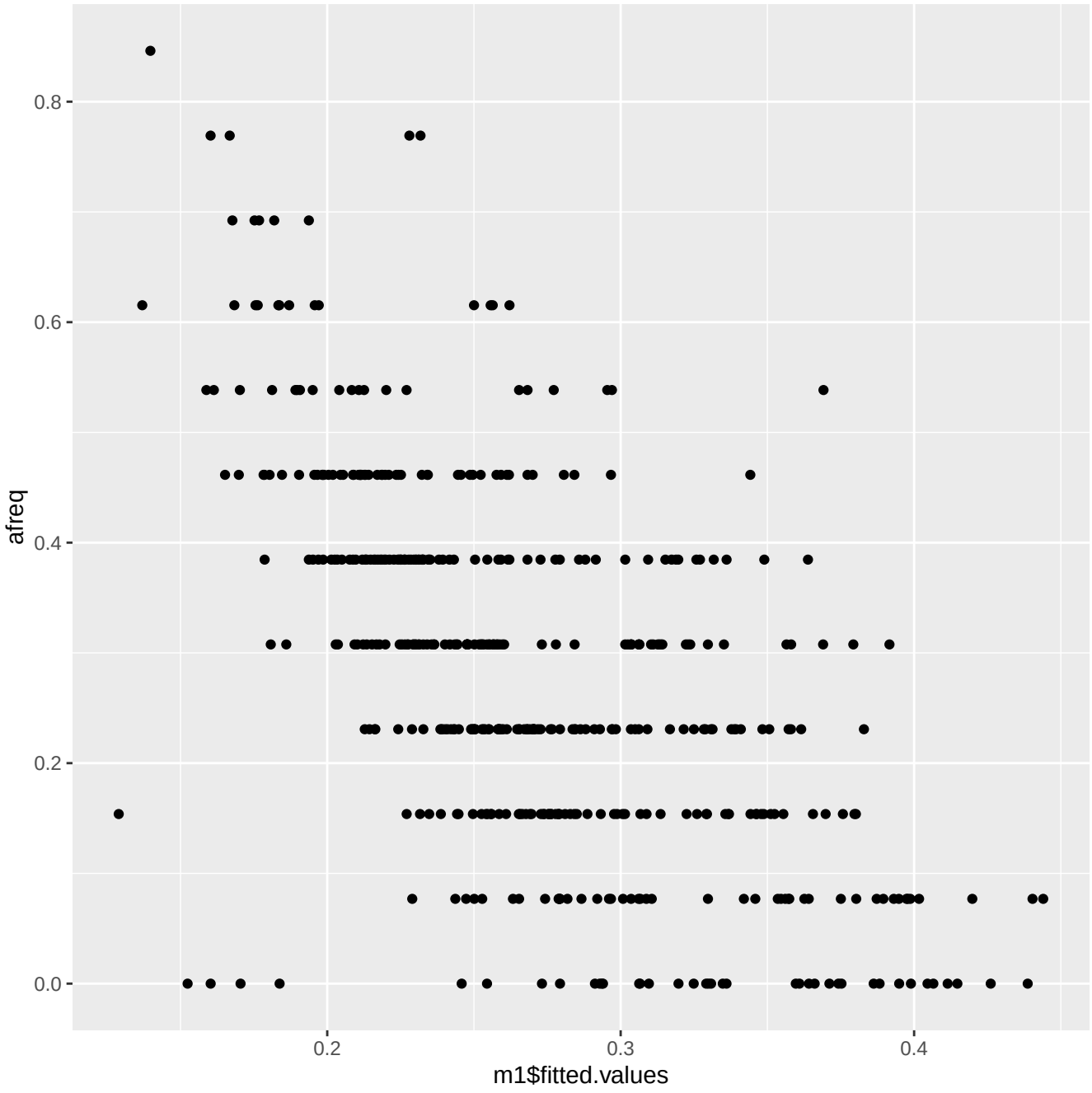

```
ggplot(labhw, aes(goodsoil, lnpopn)) + geom_point()
```



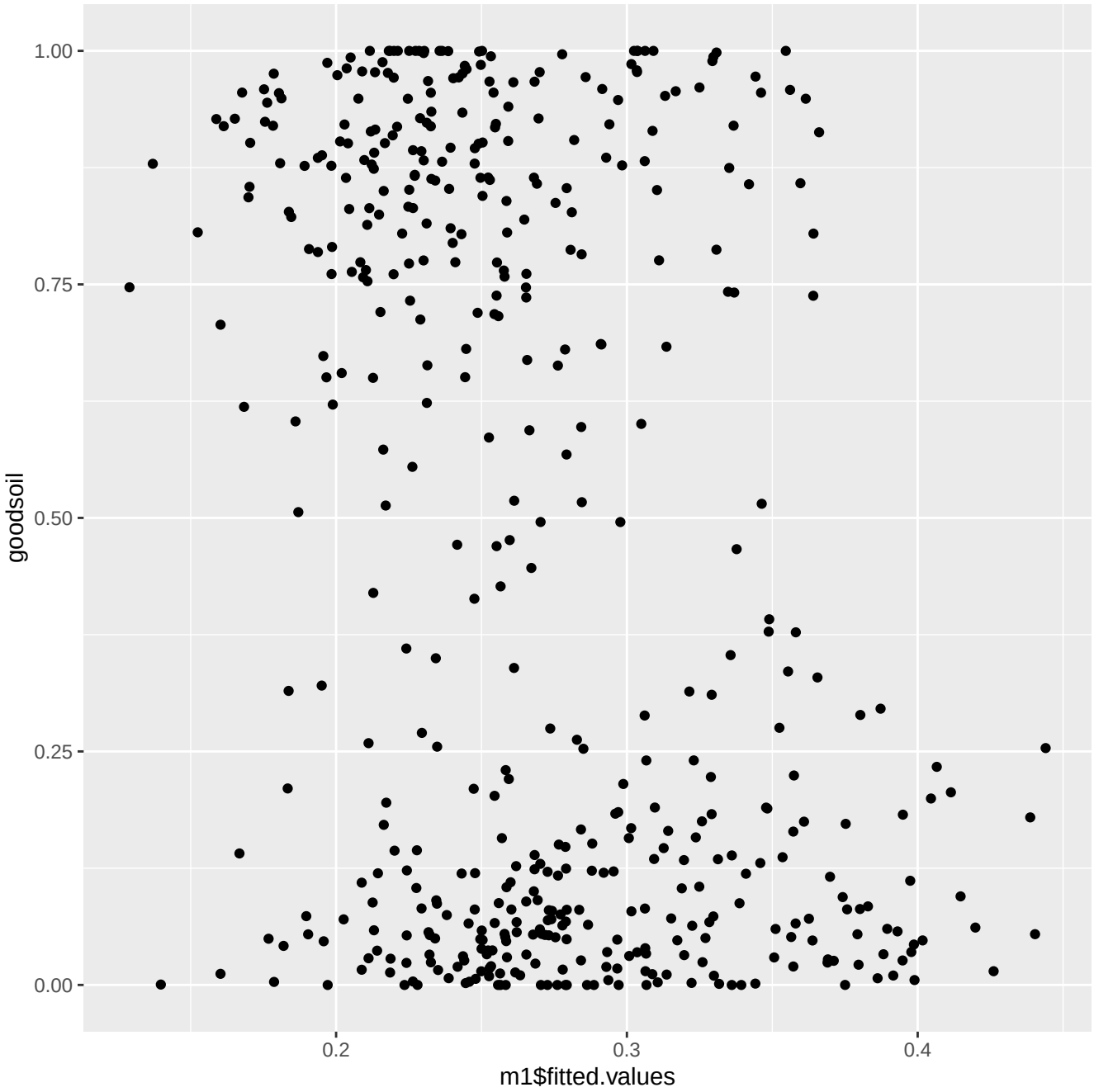
Итак, уже невооруженным взглядом заметно, что на приведенных графиках **имеется гетероскедастичность** (где-то *монотонная*, где-то возможно заметить подобие *нелинейной связи*).

Смотреть на гетероскедастичность отдельно для каждого регрессора *неэффективно*, поэтому лучше посмотреть на **связь с предсказанными значениями** ($\hat{y}_i$), чтобы увидеть некоторый *общий показатель*.

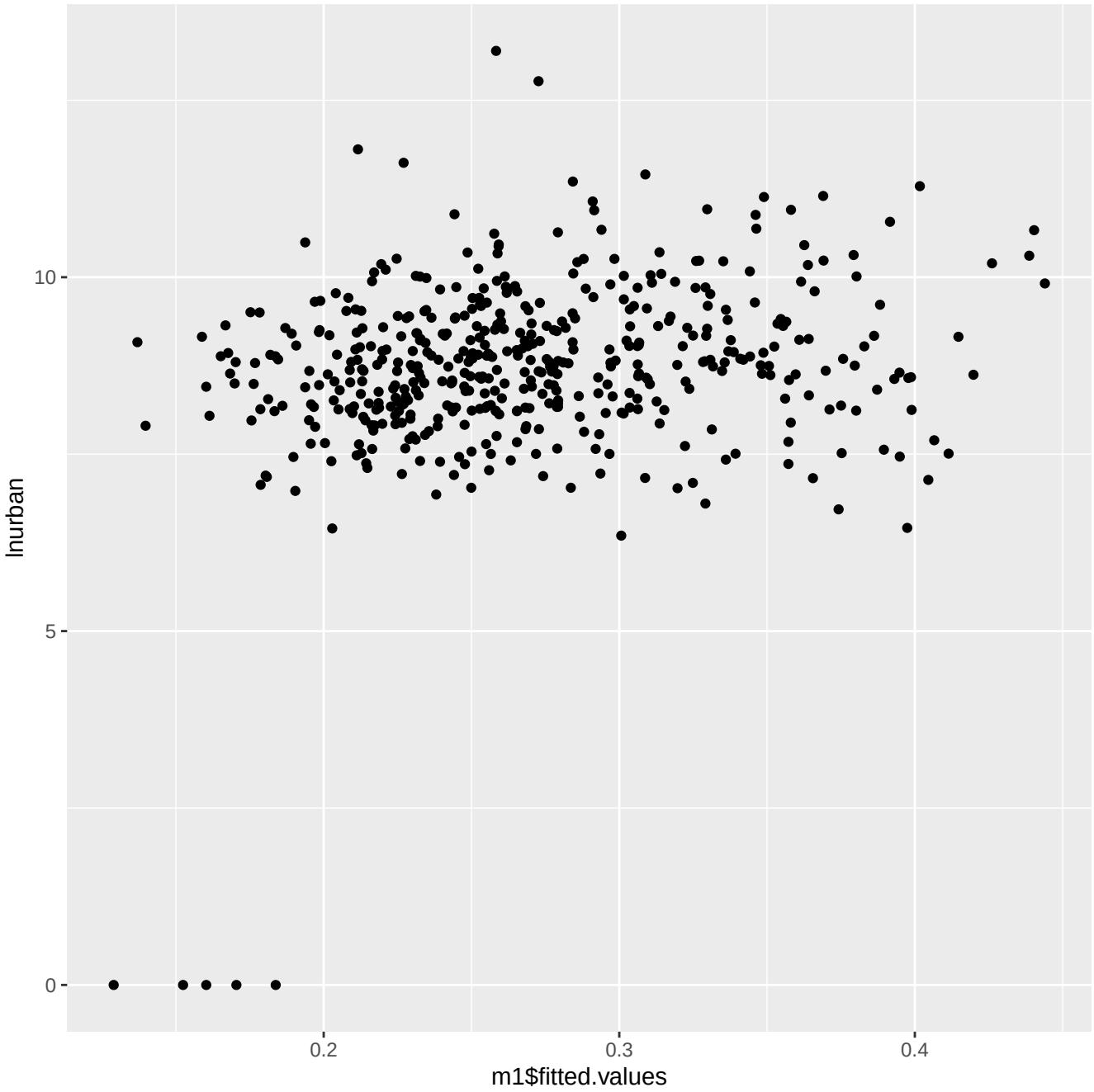
```
library(ggplot2)

ggplot(labhw, aes(m1$fitted.values, afreq)) + geom_point()
```



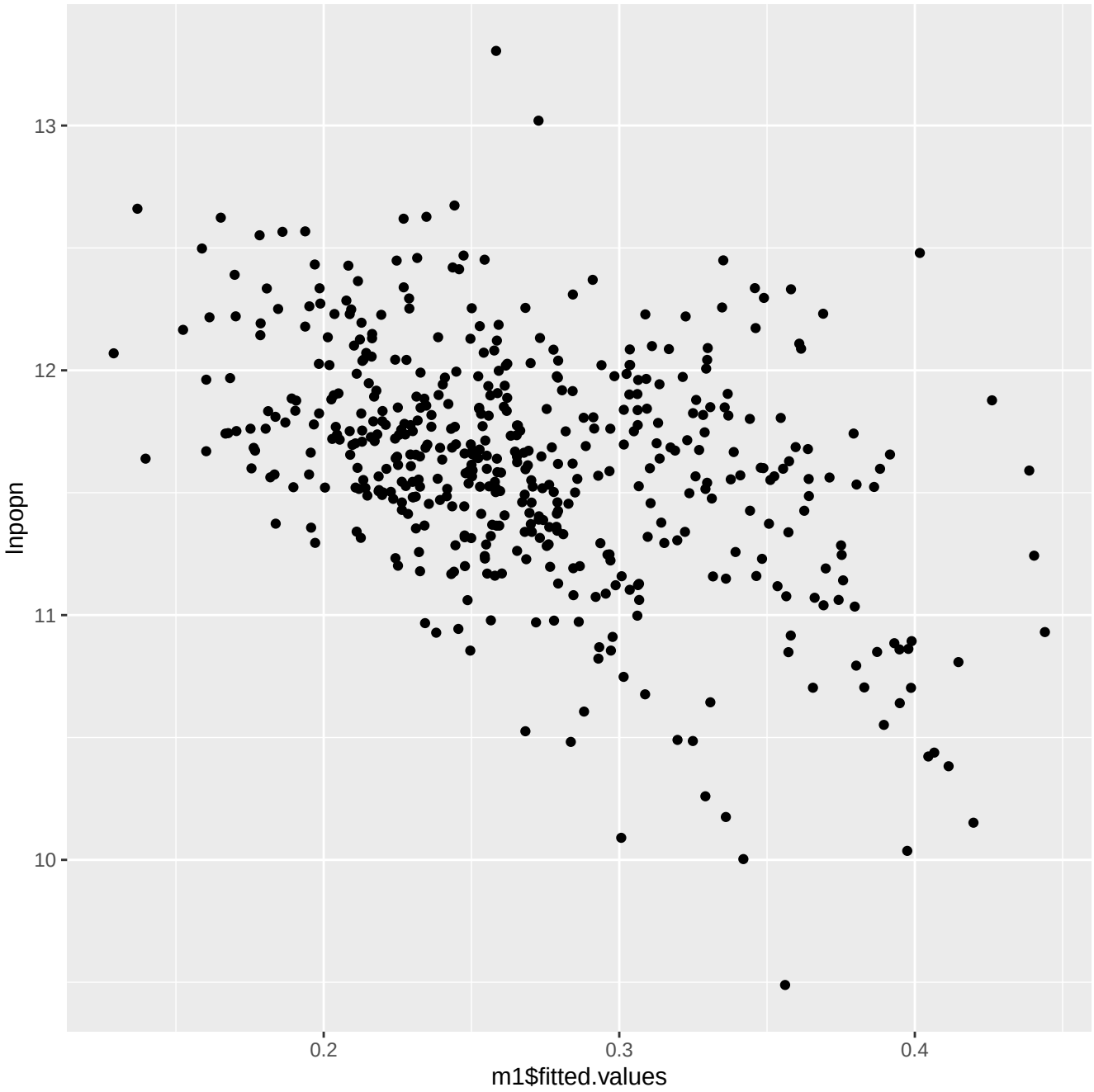

```
ggplot(labhw, aes(m1$fitted.values, goodsoil)) + geom_point()
```




```
ggplot(labhw, aes(m1$fitted.values, lnurban)) + geom_point()
```

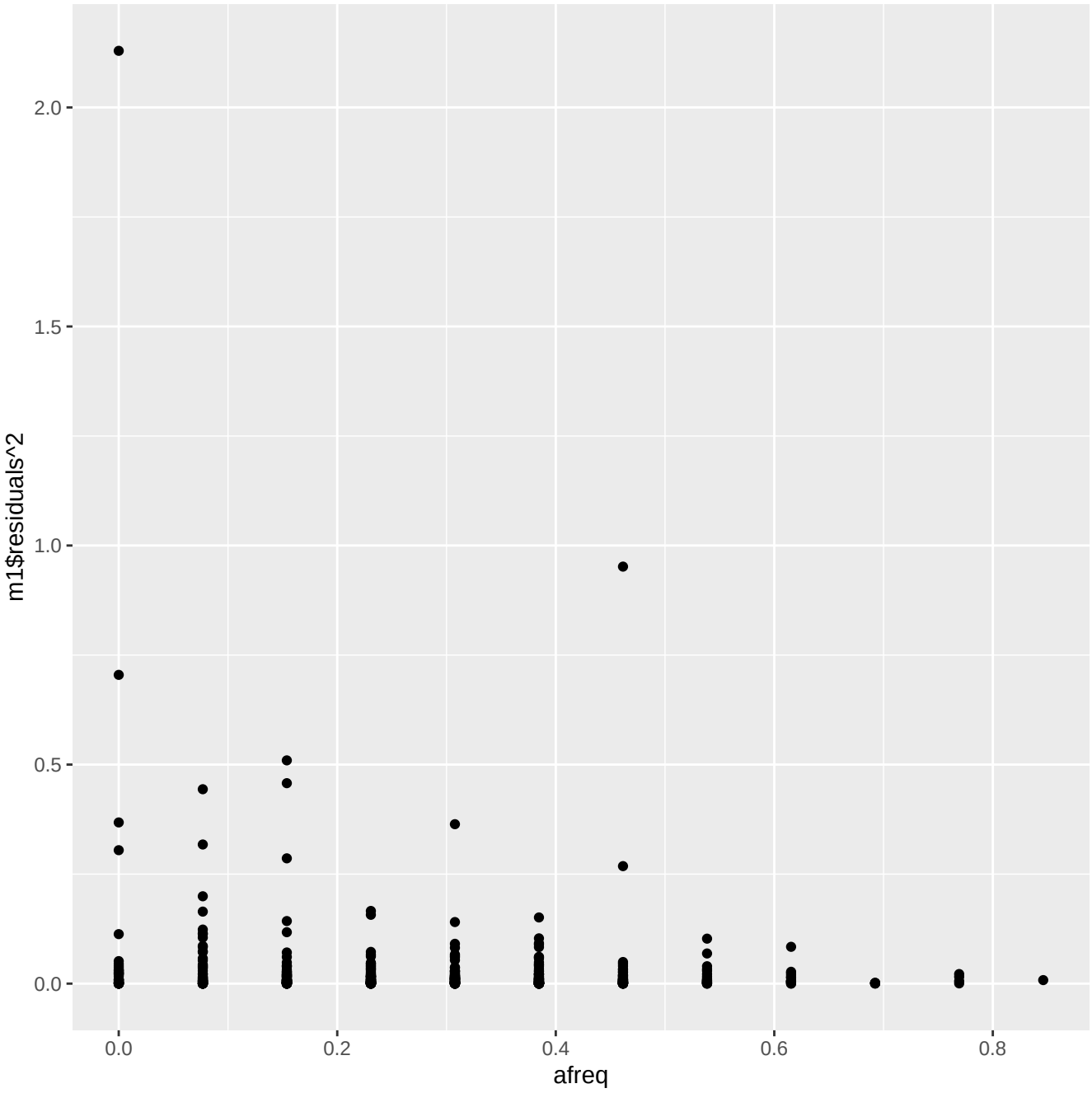



```
ggplot(labhw, aes(m1$fitted.values, lnpopn)) + geom_point()
```

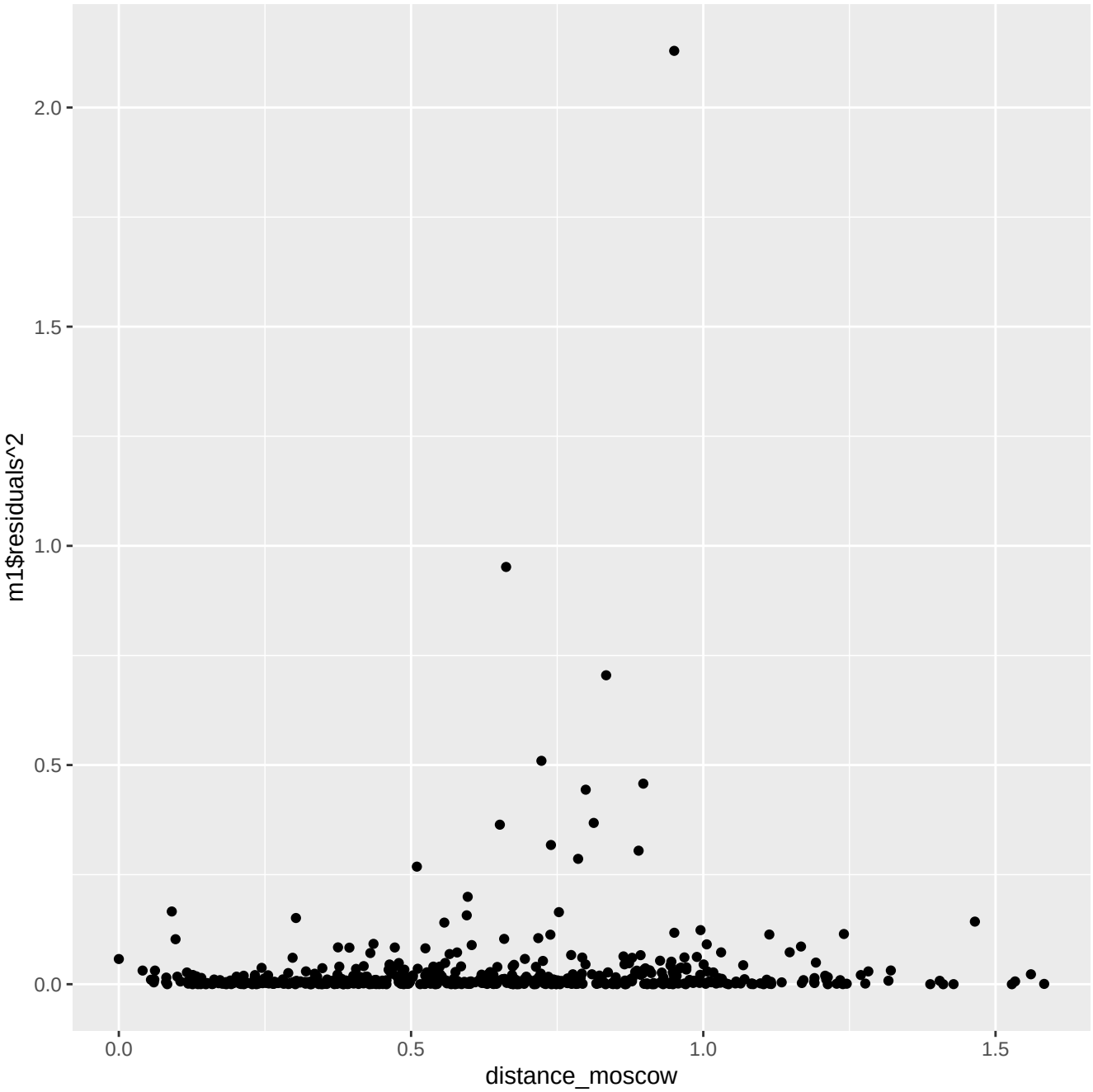


Возведем остатки в квадрат для наглядности, чтобы посмотреть на более выраженную взаимосвязь и посмотрим на их *связъ с предиктором*, а также *с предсказанными значениями*.

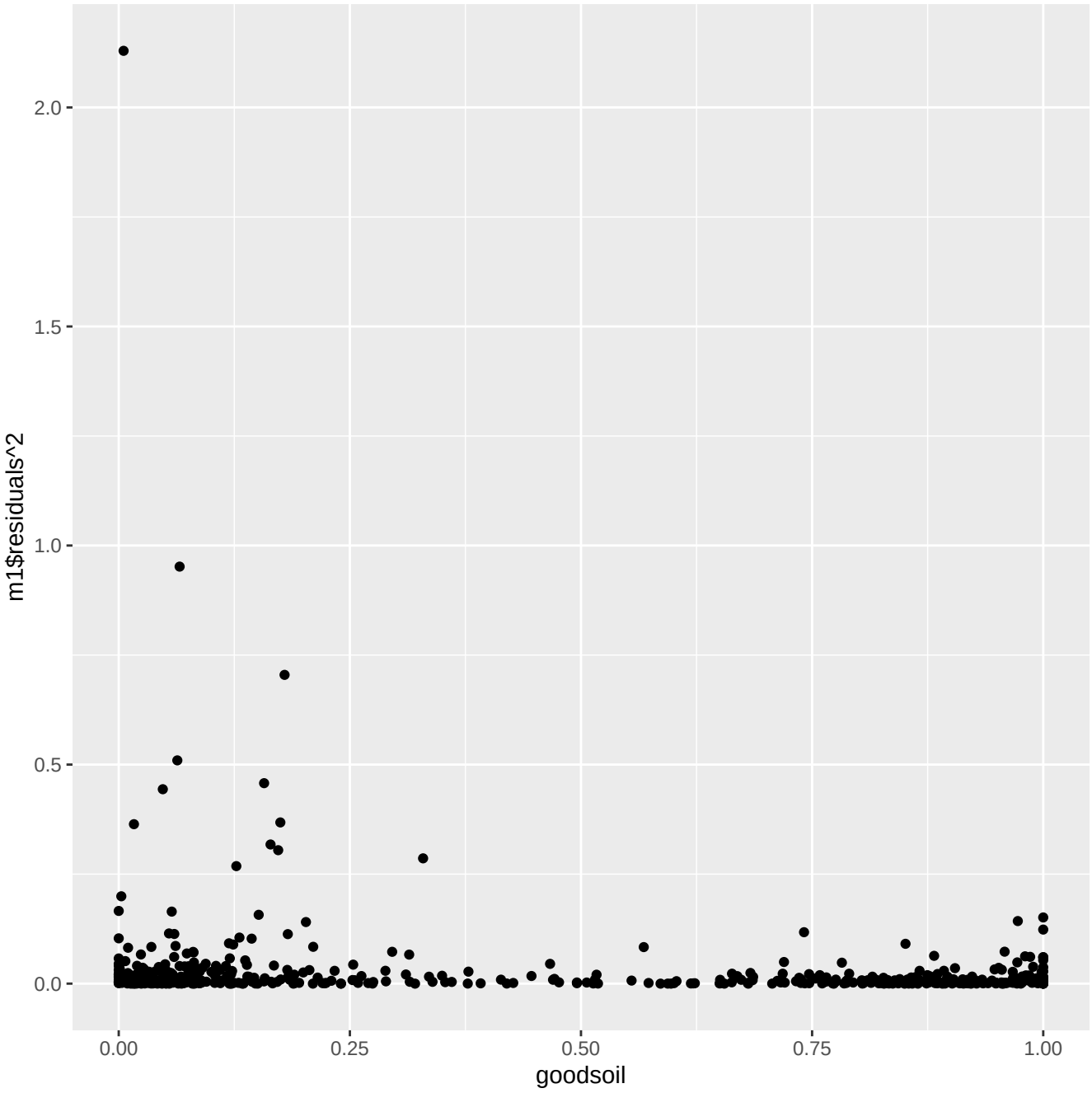
```
library(ggplot2)

ggplot(labhw, aes(afreq, m1$residuals^2)) + geom_point()
```



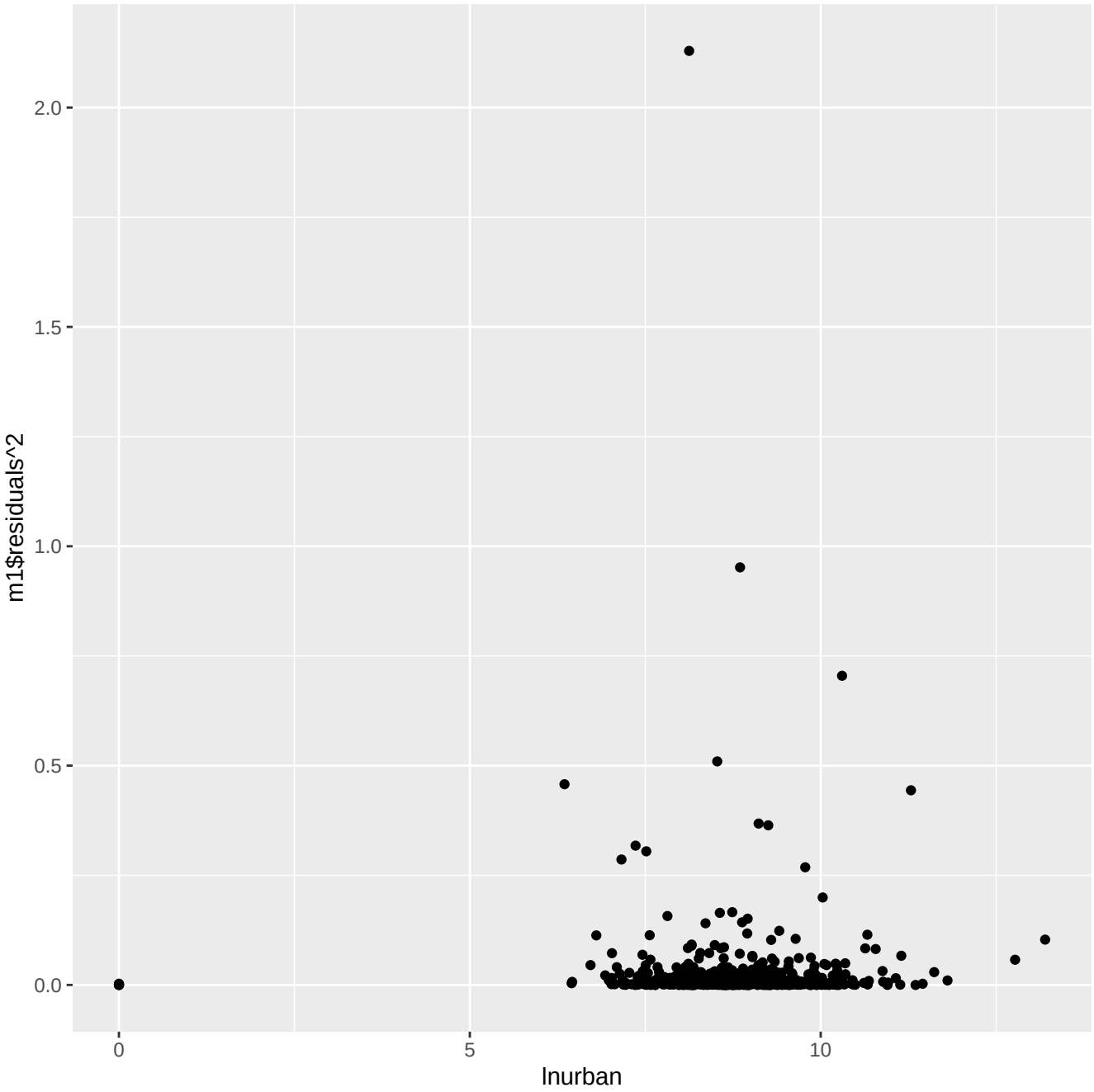

```
ggplot(labhw, aes(distance_moscow, m1$residuals^2)) + geom_point()
```



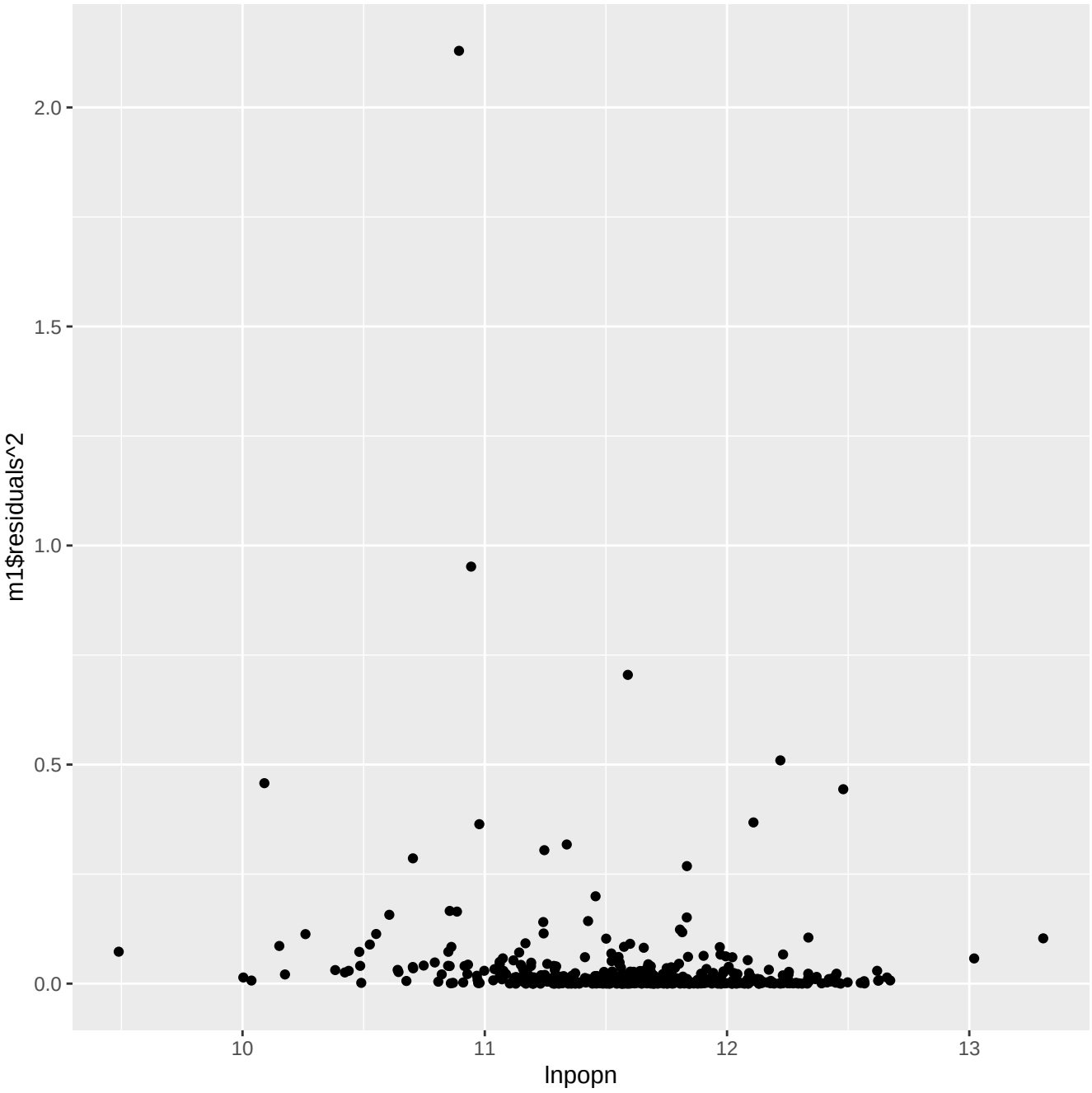

```
ggplot(labhw, aes(goodsoil, m1$residuals^2)) + geom_point()
```



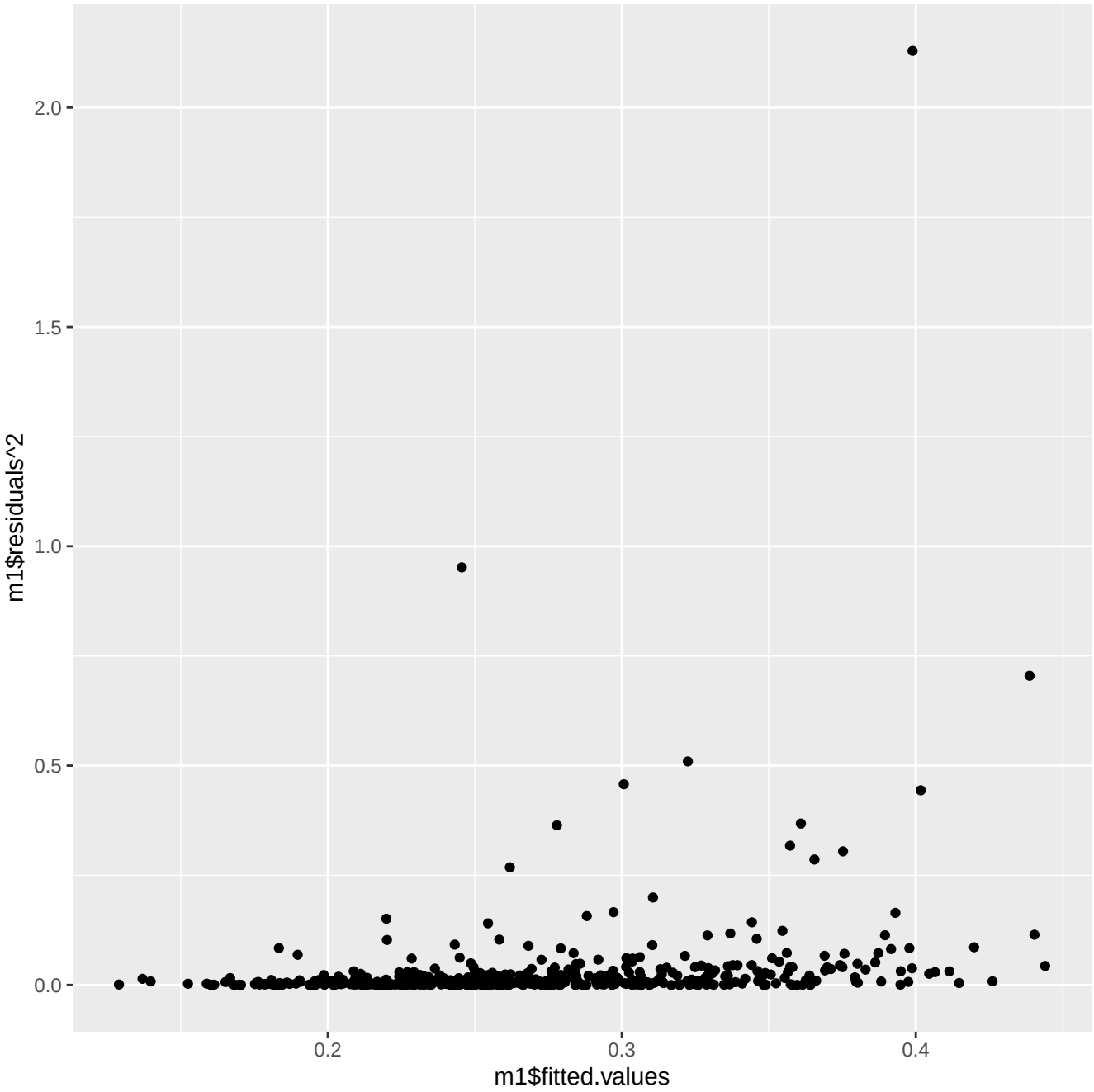

```
ggplot(labhw, aes(lnurban, m1$residuals^2)) + geom_point()
```




```
ggplot(labhw, aes(lnpopn, m1$residuals^2)) + geom_point()
```




```
ggplot(labhw, aes(m1$fitted.values, m1$residuals^2)) + geom_point()
```



Теперь мы довольно явно можем увидеть, что наша модель подвержена гетероскедастичности.

Итак, действительно, для многих предикторов, о которых я выдвигал теоретические и содержательные предпосылки, можно **заметить довольно сильную гетероскедастичность**.

Убедимся же теперь в этом формально.

2. **Формальный тест Бреуша—Пагана.**

   Немного о самом тесте:

   Тест Бреуша—Пагана очень похож на тест Уайта, однако вместо более расширенной вспомогательной модели в нем используются только исходные переменные (в нашем случае: afreq, distance_moscow, goodsoil, lnurban, lnpopn, province_capital, nozemstvo).

   Также он является более модернизированным тестом, т. к. он одинаково хорошо действует как на больших выборках, так и на маленьких («стьюдентизированный» тест). Однако в нашем случае это не столь принципиально, поскольку наша выборка большая, она поддавалась бы ассимптотике, заложенную в тест Уайта.

Помимо этого, не всегда рационально предполагать, что у нас есть взаимосвязь ошибок с квадратичными значениями и попарными взаимодействиями, это нуждается в теоретическом обосновании. К тому же лишнее нагромождение будет лишним: все попарные произведения точно не внесут какой-либо значимый вклад в объяснение.

```
library(lmtest)

## Loading required package: zoo
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

bptest(m1)

##
##  studentized Breusch-Pagan test
##
## data:  m1
## BP = 33.11, df = 7, p-value = 2.526e-05
```

В начале запишем нулевую и альтернативную гипотезы для данного теста:

$$H_0 : Var(\varepsilon|X) = \sigma^2$$
$$H_1 : Var(\varepsilon|X) \neq \sigma^2$$

Как видно, мы получили довольно маленькое значение p-value = 2.526e-05. Следовательно, у нас **есть основания отвергнуть нулевую гипотезу** на конвенциональном уровне значимости 0.05.

Таким образом, на уровне значимости 0.05 мы можем утверждать, что наша ***модель подвержена гетероскедастичности***. Однако по тесту Бреуша—Пагана мы не можем сказать, каким образом выглядит эта связь — мы не предполагали никакую направленность связи из-за особенности теста.

3. **Формальный тест Goldfeld-Quandt.**

Действительно, **есть основания полагать, что что вариация ошибок зависит от одной из объясняющих переменных**: теоретические обоснования подобной опасности были высказаны в ответе на задание 3, они визуально подтвердились в 1 пункте данного задания.
Поэтому проверим наши опасения о гетероскедастичности через формальный тест Goldfeld-Quandt на основе содержательных предпосылок.

Прежде чем сделать это, нужно сказать о двух **допущениях** данного теста:

(a) тест предполагает, что связь дисперсии остатков и предиктора **монотонна**.

(b) тест предполагает, что ошибки (и остатки как их оценка) имеют нормальное распределение $\varepsilon_i \sim N(0; \sigma^2)$

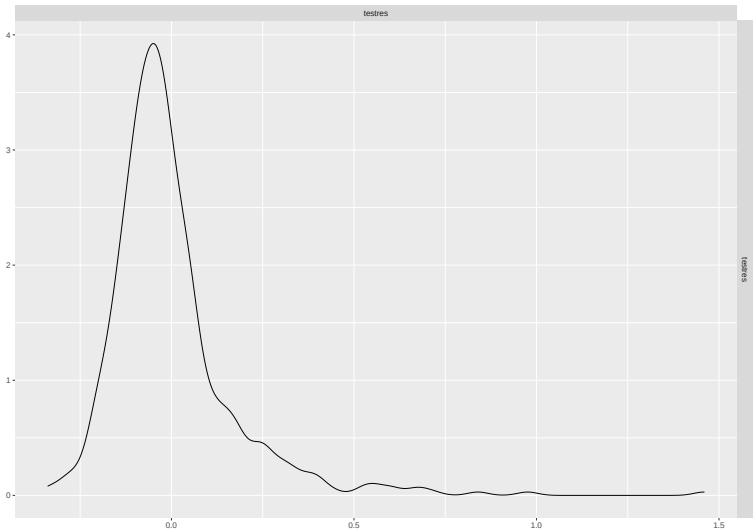
Во-первых, можно высказать опасение о том, что не везде связь между предиктором и дисперсией остатков может убывать или возрастать монотонно в явном виде (было видно, что иногда

она «проседает в середине»). Однако, по моему мнению, это не столь критично для теста: в большинстве случаев она скорее монотонна, чем нет. К тому же, это скорее касается взаимодействия переменных.

Таким образом, можно утверждать, что **тест Goldfeld-Quandt лучше подходит для наших данных**, поскольку благодаря нему мы сможем более явно высказаться о направлении связи и о том, какие именно предикторы приводят к появлению гетероскедастичности, а также поскольку мы могли заметить монотонную связь между изменением предикторов и квадратами остатков на графиках.

Во-вторых, визуально проверим нормальность распределения остатков для того, чтобы иметь право пользоваться данным тестом. Для этого добавим вектор остатков в новый (тестовый) датасет и выведем его распределение:

```
testdata = labhw
testdata$testres = m1$residuals

ggpairs(testdata[,-c(1:9)])

## Error in ggpairs(testdata[, -c(1:9)]): could not find function "ggpairs"
```



Из графика можно заметить, что **распределение остатков действительно близко к нормальному** (с небольшим сдвигом влево от 0).

Итак, проведем тест Goldfeld-Quandt для интересующих нас переменных (мы выяснили, что наиболее подвержены гетероскедастичности переменные *afreq, distance_moscow, goodsoil, lnpopn, province_capital, nozemstvo*).

Как и в случае с визуальными диагностиками, приведу тесты сразу для нескольких переменных, однако даже подтвержденной гетероскедастичности для одной из них было бы достаточно для распространения этого тезиса относительно всей модели.

```
library(lmtest)
gqtest(m1, order.by = ~afreq, data = labhw, fraction = 0.2, alternative = "less")

##
##  Goldfeld-Quandt test
##
## data:  m1
```

```
## GQ = 0.37354, df1 = 188, df2 = 187, p-value = 1.963e-11
## alternative hypothesis: variance decreases from segment 1 to 2

gqtest(m1, order.by = ~distance_moscow, data = labhw, fraction = 0.2, alternative = "grea

##
##  Goldfeld-Quandt test
##
## data:  m1
## GQ = 3.9175, df1 = 188, df2 = 187, p-value < 2.2e-16
## alternative hypothesis: variance increases from segment 1 to 2

gqtest(m1, order.by = ~goodsoil, data = labhw, fraction = 0.2, alternative = "less")

##
##  Goldfeld-Quandt test
##
## data:  m1
## GQ = 0.2353, df1 = 188, df2 = 187, p-value < 2.2e-16
## alternative hypothesis: variance decreases from segment 1 to 2

gqtest(m1, order.by = ~lnpopn, data = labhw, fraction = 0.2, alternative = "less")

##
##  Goldfeld-Quandt test
##
## data:  m1
## GQ = 0.40706, df1 = 188, df2 = 187, p-value = 7.289e-10
## alternative hypothesis: variance decreases from segment 1 to 2

gqtest(m1, order.by = ~province_capital, data = labhw, fraction = 0.2, alternative = "les

##
##  Goldfeld-Quandt test
##
## data:  m1
## GQ = 0.62137, df1 = 188, df2 = 187, p-value = 0.0005996
## alternative hypothesis: variance decreases from segment 1 to 2

gqtest(m1, order.by = ~nozemstvo, data = labhw, fraction = 0.2, alternative = "greater")

##
##  Goldfeld-Quandt test
##
## data:  m1
## GQ = 3.0884, df1 = 188, df2 = 187, p-value = 2.825e-14
## alternative hypothesis: variance increases from segment 1 to 2
```

Напомню нулевую и альтернативную гипотезу для теста Goldfeld-Quandt:

$$H_0 : Var(\varepsilon|X) = \sigma^2$$
$$H_1 : Var(\varepsilon|X) \neq \sigma^2$$

Комментарий к коду:
Поскольку R по умолчанию подставляет в числитель значение RSS для подвыборки с наибольшими значениями X, некоторые альтернативы были левосторонними (когда числитель был меньше, т. е. ближе к 0) и правосторонними (когда числитель был больше).

Напомню статистику критерия для теста:

$$\frac{RSS_1/(n_1 - k)}{RSS_2/(n_2 - k)} \sim F(n_1 - k, n_2 - k)$$

, где (при расчете в R) в числителе значение RSS для подвыборки с наибольшими значениями X.

Поскольку наши переменные сильно не скоррелированы, это даст нам увидеть действительный «вклад» каждой из них в гетероскедастичность.

Как мы можем видеть, наши опасения подтвердились **для всех переменных**.

Для 6 выбранных из 7 предикторов (afreq, distance_moscow, goodsoil, lnpopn, province_capital, nozemstvo) на конвенциональном уровне значимости 0.05 мы **можем утверждать о наличии гетероскедастичности из-за них** (условной гетероскедастичности). При этом, кроме province_capital, для остальных 5 предикторов p-value имеет 10 и более нолей после запятой, **чего и следовало ожидать, исходя из обозначенных выше теоретических предпосылок**.

Благодаря тесту Goldfeld-Quandt мы также можем сказать о подтвержденной монотонной направленности связи:

$X \uparrow Var(\varepsilon) \downarrow$	$X \uparrow Var(\varepsilon) \uparrow$
afreq, goodsoil, lnpopn, province_capital	distance_moscow, nozemstvo

То есть с увеличением значений переменных из первого столбца значение условной дисперсии ошибок уменьшается, а для переменных из второго столбца таблицы — при увеличении значений условная дисперсия также увеличивается.

4. **Переоцените стандартные ошибки (используйте тип ошибок HC3).**

Поскольку мы обнаружили и подтвердили гетероскедастичность в модели, необходимо переоценить стандартные ошибки с учетом этого.

Воспользуемся для этого типом ошибок *HC3* и рассчитаем скорректированные значения.

В начале выведем новую ковариационную матрицу с новыми значениями дисперсии, а затем — старые стандартные ошибки оценок коэффициентов при предикторах вместе с посчитанными p-value, а также новые — для сравнения.

```
library(sandwich)

vcovHC(m1)
```

```
##                   (Intercept)         afreq     nozemstvo distance_moscow
## (Intercept)       0.1025468385  7.092250e-03 -2.748507e-03    6.058804e-03
## afreq             0.0070922505  3.853950e-03 -1.001303e-03    1.026836e-03
## nozemstvo        -0.0027485072 -1.001303e-03  7.718827e-04   -4.690936e-04
## distance_moscow   0.0060588043  1.026836e-03 -4.690936e-04    9.824750e-04
## goodsoil          0.0004554261 -9.917193e-05  1.017407e-04   -1.995505e-04
## lnurban          -0.0002231290 -2.627990e-05 -2.406308e-05    1.154785e-05
## lnpopn           -0.0090601252 -7.224821e-04  2.892127e-04   -5.868134e-04
## province_capital  0.0011164659 -1.180201e-04  6.820253e-05   -4.366848e-06
##                       goodsoil       lnurban        lnpopn province_capital
## (Intercept)       4.554261e-04 -2.231290e-04 -9.060125e-03     1.116466e-03
## afreq            -9.917193e-05 -2.627990e-05 -7.224821e-04    -1.180201e-04
## nozemstvo         1.017407e-04 -2.406308e-05  2.892127e-04     6.820253e-05
## distance_moscow  -1.995505e-04  1.154785e-05 -5.868134e-04    -4.366848e-06
## goodsoil          4.614287e-04 -1.949157e-05 -3.604594e-05    -1.948549e-05
## lnurban          -1.949157e-05  2.305612e-05  3.918778e-06    -3.092961e-05
## lnpopn           -3.604594e-05  3.918778e-06  8.164268e-04    -7.598238e-05
## province_capital -1.948549e-05 -3.092961e-05 -7.598238e-05     1.107928e-03
```

```
summary(m1) # старые значения стандартных ошибок и значимости
```

```
##
## Call:
## lm(formula = ch_schools_pc ~ afreq + nozemstvo + distance_moscow +
##     goodsoil + lnurban + lnpopn + province_capital, data = labhw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33848 -0.09634 -0.03551  0.04695  1.45921
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.683238   0.218271   3.130 0.001853 **
## afreq            -0.181470   0.054400  -3.336 0.000916 ***
## nozemstvo         0.080091   0.021793   3.675 0.000264 ***
## distance_moscow  -0.012096   0.031894  -0.379 0.704660
## goodsoil         -0.008286   0.023997  -0.345 0.730028
## lnurban           0.013287   0.007274   1.827 0.068371 .
## lnpopn           -0.042304   0.019890  -2.127 0.033937 *
## province_capital  0.040362   0.030172   1.338 0.181621
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1782 on 481 degrees of freedom
## Multiple R-squared:  0.1018,Adjusted R-squared:  0.08873
## F-statistic: 7.788 on 7 and 481 DF,  p-value: 6.081e-09
```

```
coeftest(m1, vcov=vcovHC(m1)) # новые значения стандартных ошибок и значимости
```

```
##
```

```
## t test of coefficients:
##
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.6832379  0.3202294  2.1336 0.033382 *
## afreq            -0.1814696  0.0620802 -2.9231 0.003628 **
## nozemstvo         0.0800914  0.0277828  2.8828 0.004118 **
## distance_moscow  -0.0120961  0.0313445 -0.3859 0.699735
## goodsoil         -0.0082858  0.0214809 -0.3857 0.699867
## lnurban           0.0132868  0.0048017  2.7671 0.005873 **
## lnpopn           -0.0423038  0.0285732 -1.4805 0.139383
## province_capital  0.0403619  0.0332855  1.2126 0.225880
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Как мы можем видеть, результаты действительно значительно изменились.

Во-первых, в целом, значимость 5 оценок (включая константу) снизилась, в то время, как у 3 оценок — увеличилась (2 из них — distance_moscow и goodsoil, которые до этого имели p-value $>$ 0.7).

Во-вторых, подробнее:

- дисперсия оценки константы увеличилась, и значимость оценки приняла пограничное значение (≈ 0.033 вместо ≈ 0.002);
- дисперсия оценки коэф. при предикторе lnpopn увеличилась, и оценка стала статистически незначимой (≈ 0.139 вместо ≈ 0.034);
- дисперсия оценки коэф. при предикторе lnurban уменьшилась, и оценка стала статистически значимой (≈ 0.006 вместо ≈ 0.068).

Таким образом, значимость большинства предикторов, посчитанная с учетом гетероскедастичности, снизилась (в то время, как один из коэффициентов, наоборот, стал значимым на высоком уровне значимости).

Можно сказать, что после переоценки стандартных ошибок **результаты изменились значимо**.

Итак, мы содержательно объяснили, почему мы можем ожидать гетероскедастичность в предложенной модели, а также подтвердили это на основе визуализации и формальных тестов и рассчитали новые значения значимости оценок коэффициентов параметров модели с учетом гетероскедастичности.

Задача 5. Бонусное задание

Покажем, что

$$se(\hat{\beta}) = \frac{\sigma \cdot \sqrt{VIF}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

В начале воспользуемся матричной формулой для ковариационной матрицы.

$$\Omega = \sigma^2 (X^T X)^{-1}$$

где $\hat{\sigma}^2 = \dfrac{RSS}{n-k}$

Поскольку нам нужны элементы главной диагонали ковариационной матрицы (т. е. дисперсии оценок коэффициентов), представим формулу для дисперсии β_i:

$$\sigma^2(\beta_i) = \sigma^2(X_i^T M_a X_i)^{-1} \tag{1}$$

где X_i — вектор N на 1, представляющий из себя вектор наблюдений для необходимого предиктора;

$$M_a = I - X_a(X_a^T X_a)^{-1} X_a^T$$

где X_a — матрица N на k-1, где первый вектор-столбец — вектор единиц (для константы), а остальные вектор-столбцы — наблюдения для всех предикторов, кроме β_i, для оценки которого мы ищем дисперсию;
I — единичная матрица соответствующего размера.
Так, M_a является «создателем остатка» («residual maker») для i-й независимой переменной; то есть, когда M_a, умножается на вектор X_i (или предварительно умножается на X_i^T), она создает вектор из n остатков i-й независимой переменной.

Далее отцентрируем векторы X_i и представим их как x_i. Представим, что xx_{ii} представляет собой i-й элемент главной диагонали обратной матрицы $(X_i^T M_a X_i)^{-1}$. Тогда:

$$xx_{ii} = (x_i^T x_i - x_i^T X_a(X_a^T X_a)^{-1} X_a^T x_i)^{-1}$$

где $x_i^T x_i = \sum x_i^2$, т.е. $\sum(x_i - \bar{x})^2$ или TSS для i-го оцененного предиктора, а $x_i^T X_a(X_a^T X_a)^{-1} X_a^T x_i$ — та часть информации i-го предиктора, которую мы смогли объяснить засчет других переменных (или ESS для i-го предиктора).

Теперь перезапишем последнюю формулу:

$$xx_{ii} = \left(x_i^T x_i \left(1 - \frac{x_i^T X_a(X_a^T X_a)^{-1} X_a^T x_i}{x_i^T x_i} \right) \right)^{-1}$$

Во внутренних скобках мы получили выражение: 1 минус доля дисперсии i-й независимой переменной, объясненной другими предикторами (или $1 - R_i^2$, т. к. $R^2 = \frac{ESS}{TSS}$).

Далее можно записать последнее выражение в более привычном виде, избавившись от -1 степени:

$$xx_{ii} = (x_i^T x_i(1 - R_i^2))^{-1} = \frac{1}{x_i^T x_i \cdot (1 - R_i^2)} = \frac{1}{\sum x_i^2 \cdot (1 - R_i^2)} = \frac{1}{\sum x_i^2} \cdot \frac{1}{(1 - R_i^2)}$$

Помня, что $VIF = \dfrac{1}{1 - R_i^2}$, можно сказать, что:

$$xx_{ii} = \frac{1}{\sum x_i^2} \cdot \frac{1}{(1 - R_i^2)} = \frac{VIF}{\sum x_i^2} = \frac{VIF}{\sum(x_i - \bar{x})^2}$$

Теперь подставим получившееся уравнение в выражение 1:

$$\sigma^2(\hat{\beta}) = \sigma^2 \cdot \frac{VIF}{\sum(x_i - \bar{x})^2} = \frac{\sigma^2 \cdot VIF}{\sum(x_i - \bar{x})^2}$$

Следовательно,

$$se(\hat{\beta}) = \sqrt{\frac{\sigma^2 \cdot VIF}{\sum(x_i - \bar{x})^2}} = \frac{\sigma \cdot \sqrt{VIF}}{\sqrt{\sum(x_i - \bar{x})^2}}$$

, где $\hat{\sigma} = \sqrt{\frac{RSS}{n-k}}$

Итак, выражение $se(\hat{\beta}) = \frac{\sigma \cdot \sqrt{VIF}}{\sqrt{\sum (x_i - \bar{x})^2}}$ доказано.

Также теперь можно заметить, **почему в формуле расчета $se(\hat{\beta})$ для парной регрессии нет множителя в виде VIF**: поскольку в случае с парной регрессией у нас попросту больше нет фактора, которым можно объяснить i-й регрессор (он всего один), R^2 вспомогательной модели для парной регрессии всегда будет равен 0, т. к. это будет регрессия на константу. Следовательно, в таком случае, **VIF всегда будет равен 1**, и его бессмысленно включать в формулу.

А для множественной линейной регрессии, где всегда есть, как минимум два регрессора, всегда будет другой фактор, через который можно попытаться объяснить ту или иную переменную. Следовательно, R^2 вспомогательной модели всегда будет ненулевым, и VIF вспомогательной модели всегда будет больше 1 (только если предикторы не совершенно не связаны, но это скорее гипотетический случай).