ДЗ №3

Агалян Роберт

2023-03-29

Подгрузка библиотек:

```
library(haven)
library(ggplot2)
library(sandwich)
library(lmtest)
```

```
## Загрузка требуемого пакета: zoo
```

```
##
## Присоединяю пакет: 'zoo'
```

```
## Следующие объекты скрыты от 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(skedastic)
library(car)
```

```
## Загрузка требуемого пакета: carData
```

```
library(memisc)
```

```
## Загрузка требуемого пакета: lattice
```

```
## Загрузка требуемого пакета: MASS
```

```
##
## Присоединяю пакет: 'memisc'
```

```
## Следующий объект скрыт от 'package:car':
##
##     recode
```

```
## Следующий объект скрыт от 'package:ggplot2':
##
##     syms
```

```
## Следующие объекты скрыты от 'package:stats':
##
##     contr.sum, contr.treatment, contrasts
```

```
## Следующий объект скрыт от 'package:base':
##
##     as.array
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

1

Загрузка датасета:

```
hw <- read_dta('hwdata.dta')
head(hw)
```

```
## # A tibble: 6 × 13
##   masterid   afreq provi…¹ uezd  serfpe…² dista…³ goods…⁴ lnurban lnpopn provi…⁵
##      <dbl>   <dbl> <chr>   <chr>   <dbl>   <dbl>   <dbl>   <dbl>  <dbl>   <dbl>
## 1        1  0.0769 Arkhan… Arkh…  3.50e-4   1.07    0.254    9.91   10.9       1
## 2        2  0      Arkhan… Kems…  0         1.28    0.234    7.70   10.4       0
## 3        3 NA      Arkhan… Kol'… NA        NA      NA       NA      NA        NA
## 4        4  0      Arkhan… Meze…  0         1.32    0.182    7.47   10.6       0
## 5        5  0      Arkhan… Onez…  3.22e-5   0.886   0.206    7.51   10.4       0
## 6        6  0      Arkhan… Pech…  0         1.70    0.148   NA      NA         0
## # … with 3 more variables: ch_schools_pc <dbl>, nozemstvo <dbl>, redist <dbl>,
## #   and abbreviated variable names ¹province, ²serfperc1, ³distance_moscow,
## #   ⁴goodsoil, ⁵province_capital
```

Очистка пропусков и запуск модели регрессии:

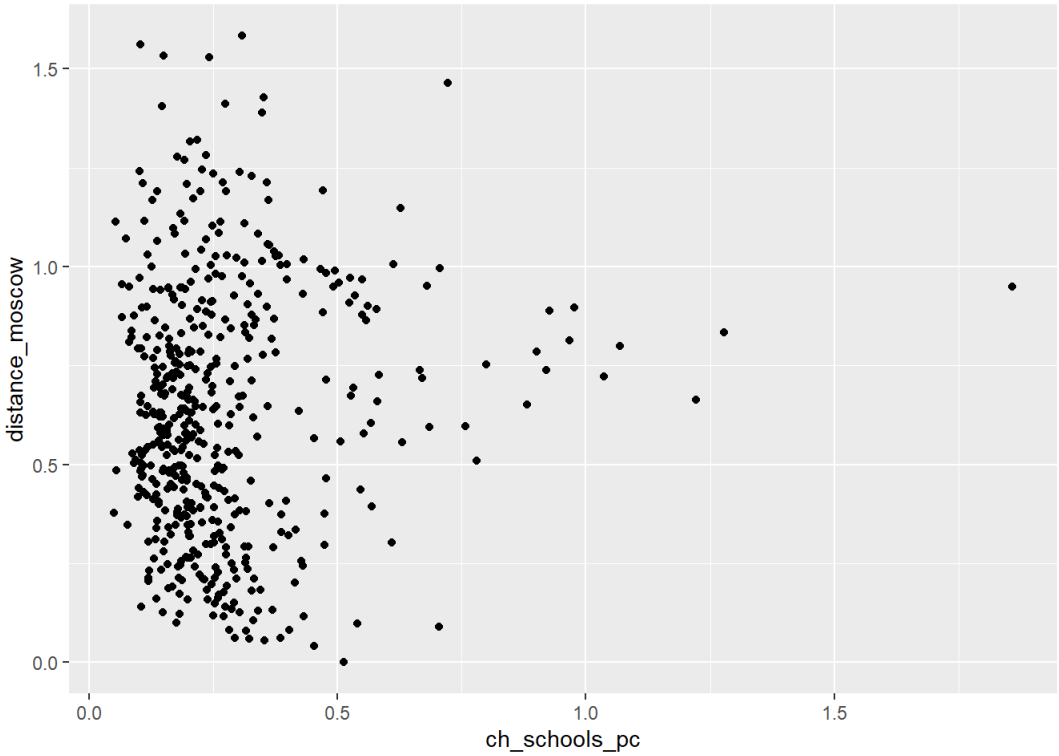
```
hw <- na.omit(hw)
mod <- lm(ch_schools_pc ~ afreq + nozemstvo + distance_moscow + goodsoil + lnurban + lnpopn + province_capital, hw)
summary(mod)
```

```
##
## Call:
## lm(formula = ch_schools_pc ~ afreq + nozemstvo + distance_moscow +
##     goodsoil + lnurban + lnpopn + province_capital, data = hw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33950 -0.09568 -0.03539  0.04702  1.45789
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.676390   0.218253   3.099 0.002055 **
## afreq            -0.179940   0.054391  -3.308 0.001009 **
## nozemstvo         0.081681   0.021824   3.743 0.000204 ***
## distance_moscow  -0.012284   0.031880  -0.385 0.700184
## goodsoil         -0.009406   0.024005  -0.392 0.695347
## lnurban           0.013754   0.007281   1.889 0.059494 .
## lnpopn           -0.042032   0.019883  -2.114 0.035030 *
## province_capital  0.038771   0.030189   1.284 0.199664
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1782 on 480 degrees of freedom
## Multiple R-squared:  0.1032, Adjusted R-squared:  0.09014
## F-statistic: 7.892 on 7 and 480 DF,  p-value: 4.526e-09
```

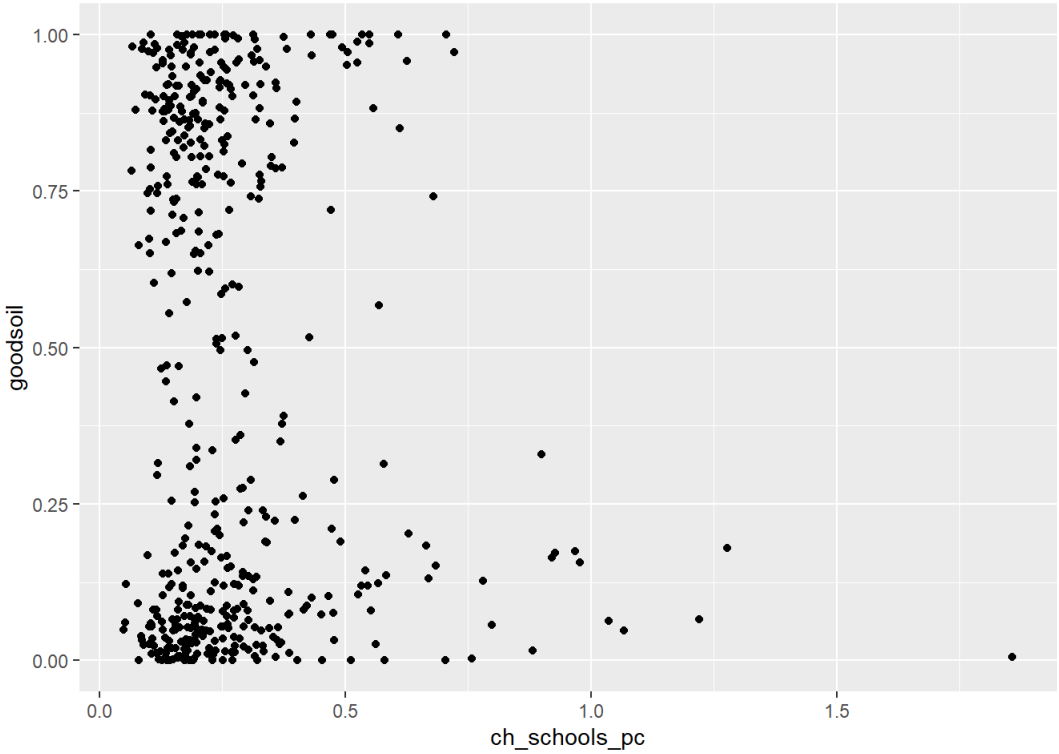
Статистически значимыми являются переменные *afreq*(0.1% уровень значимости), *nozemstvo*(~0% уровень значимости), *lnurban*(5% уровень значимости) *lnpopn* (1% уровень значимости)

Контрольные переменные 1. Требования к ним контрольные переменные должны влиять одновременно на предиктор и на отклик, однако не должно происходить обратного, предиктор не должен влиять на контрольную перемнную. 2. Соблюдаются ли они Переменная *province_capital*, обозначающая, находится ли в данном уезде “столичный” город губернии, вполне вероятно может зависеть от переменной *lnpopn*, содержащей логарифм населения в уезде и от переменной *lnurban*, содержащей логарифм городского населения в уезда.

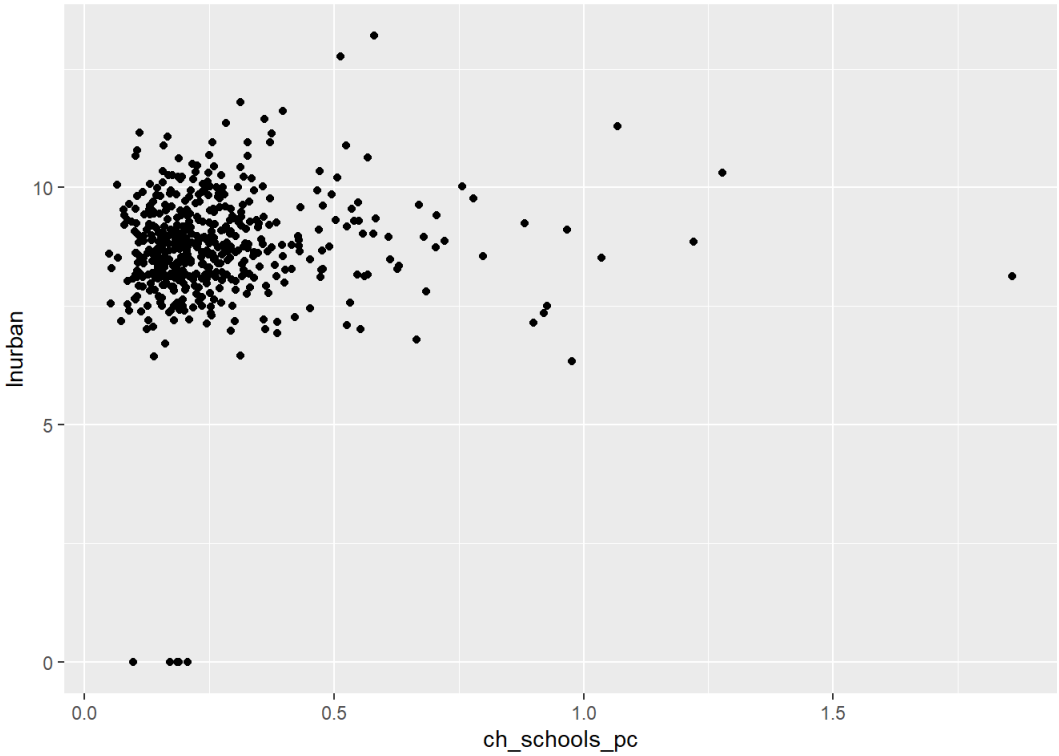
```
ggplot(hw, aes(ch_schools_pc, distance_moscow))+
  geom_point()
```



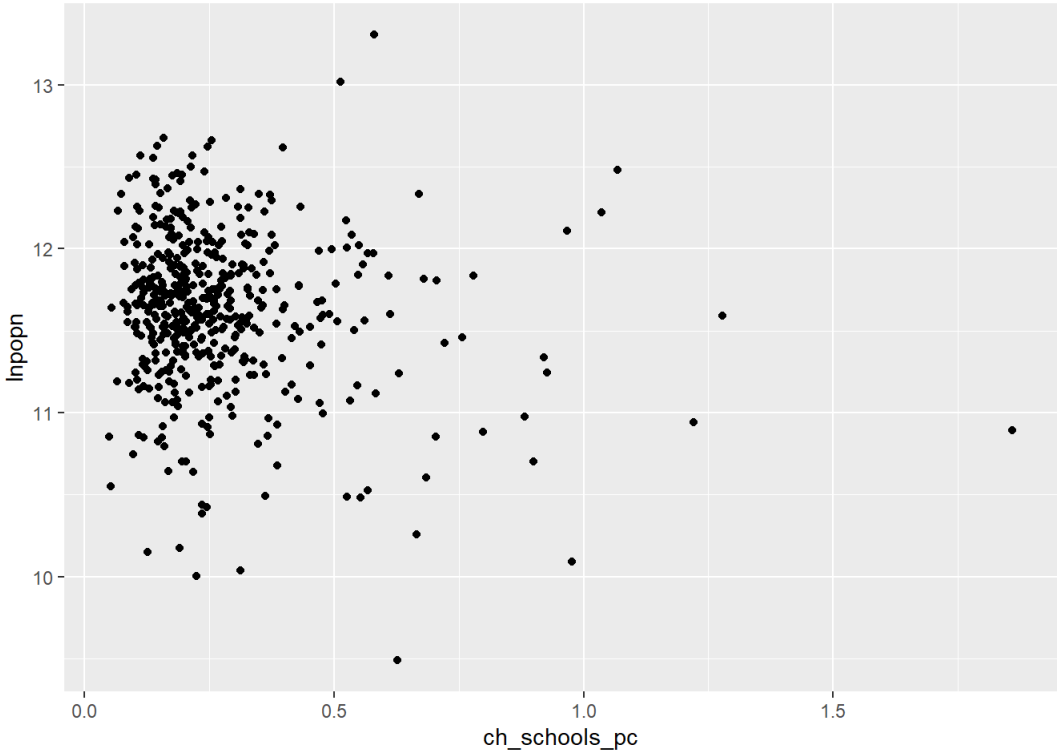

```
ggplot(hw, aes(ch_schools_pc, goodsoil))+
  geom_point()
```




```
ggplot(hw, aes(ch_schools_pc, lnurban))+
  geom_point()
```




```
ggplot(hw, aes(ch_schools_pc, lnpopn))+
  geom_point()
```



Связи между предикторами и зависимой переменной нелинейны. Возможно стоило использовать предположение об ином типе связи между переменными.

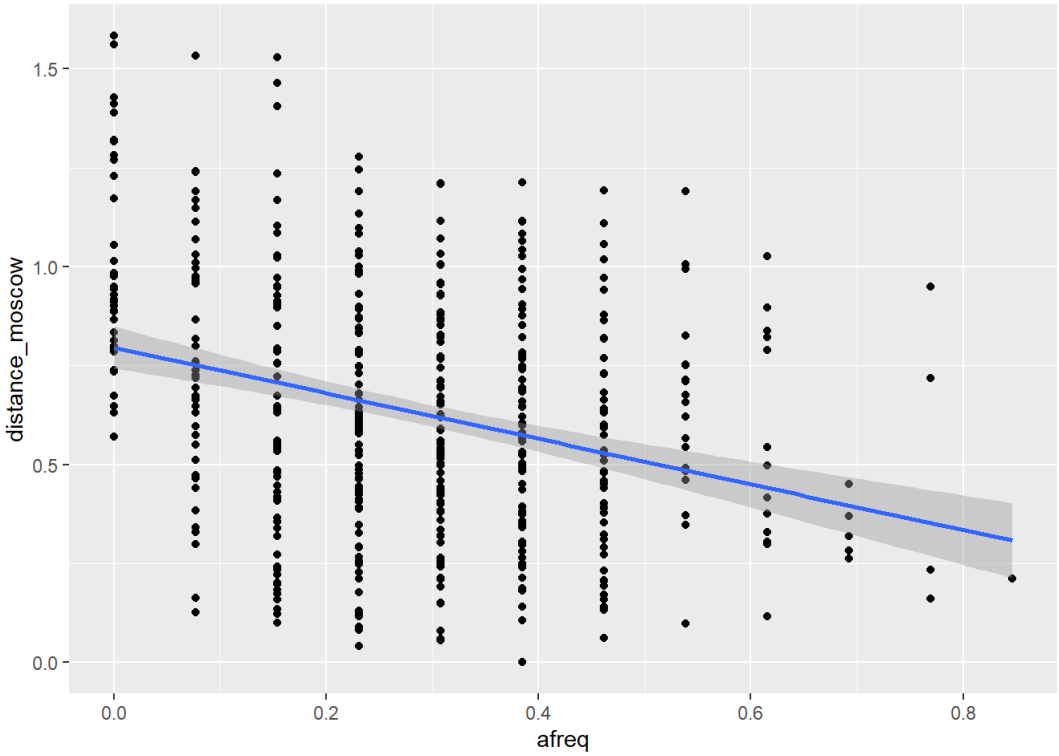
2

Мультиколлинеарность

1. Визуальная диагностика

```
ggplot(hw, aes(afreq, distance_moscow))+
  geom_point()+
  geom_smooth(method='lm')
```

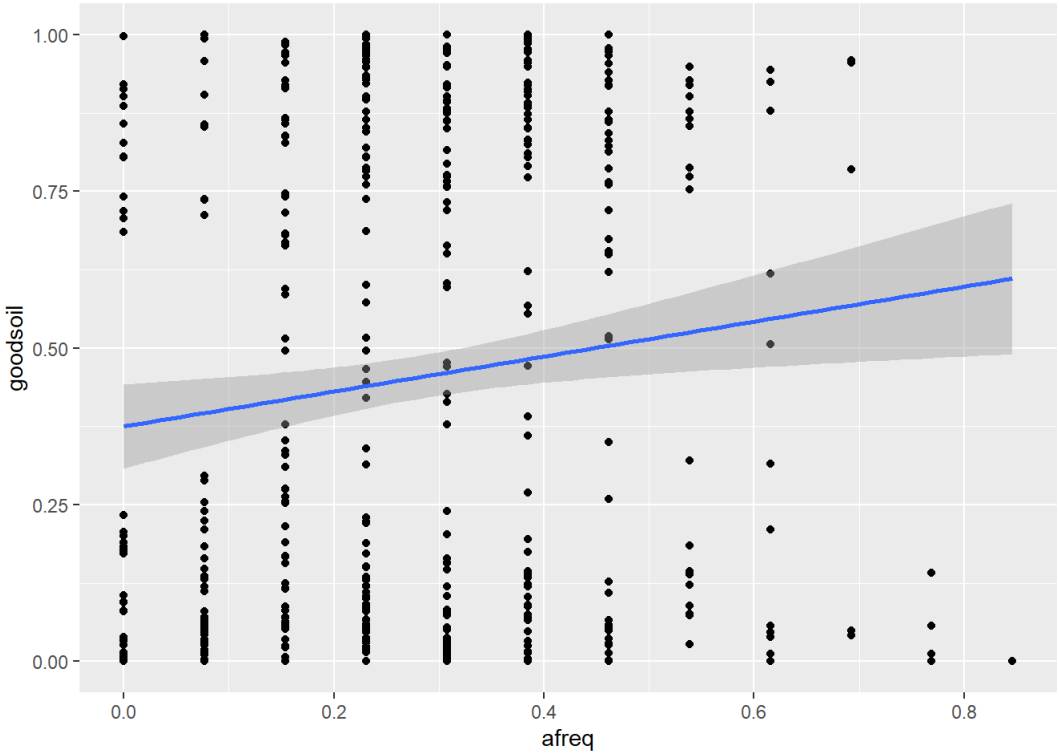
```
## `geom_smooth()` using formula = 'y ~ x'
```



Некоторая взаимосвязь присутствует между переменной *afreq* и *distance_moscow*.

```
ggplot(hw, aes(afreq, goodsoil))+
  geom_point()+
  geom_smooth(method='lm')
```

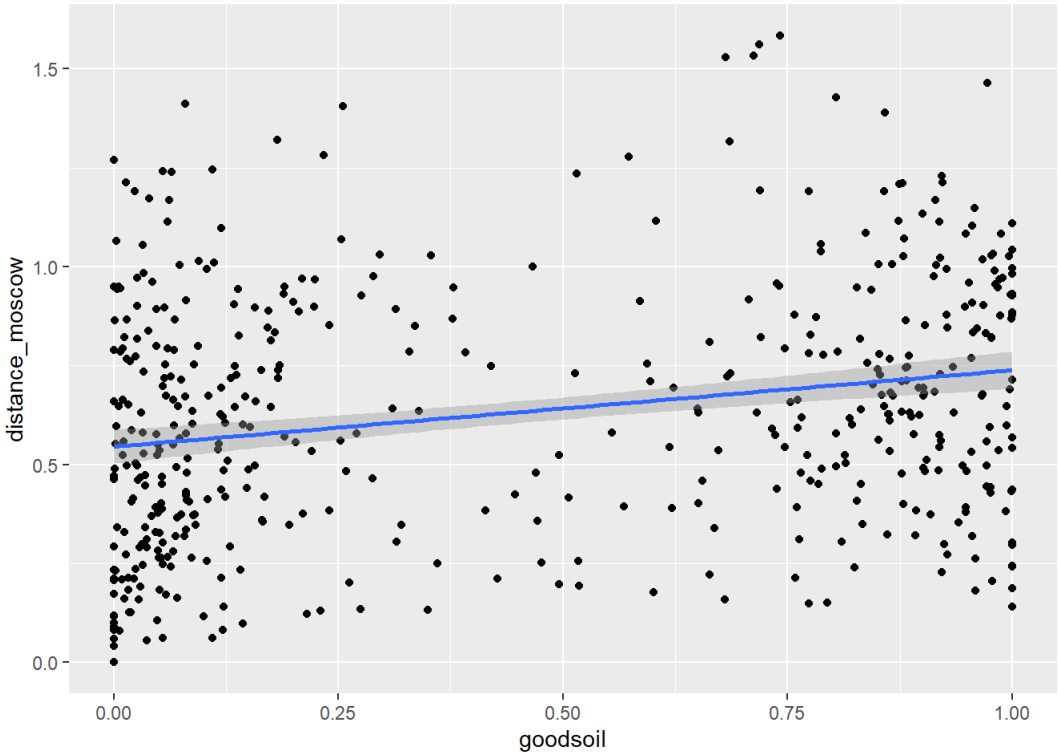
```
## `geom_smooth()` using formula = 'y ~ x'
```



Между переменными *afreq* и *goodsoil* есть предположительно слабая взаимосвязь.

```
ggplot(hw, aes(goodsoil, distance_moscow))+
  geom_point()+
  geom_smooth(method='lm')
```

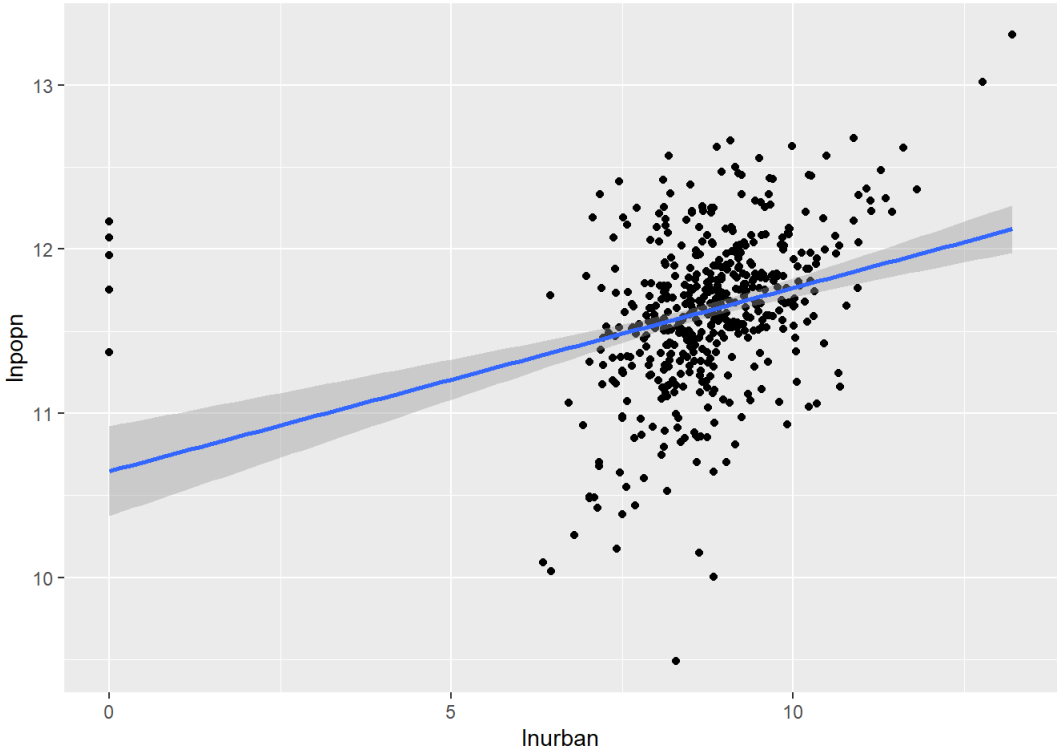
```
## `geom_smooth()` using formula = 'y ~ x'
```



Видимо, присутствует слабая взаимосвязь между расстоянием уезда до Москвы и показателем плодородности почвы.

```
ggplot(hw, aes(lnurban, lnpopn))+
  geom_point()+
  geom_smooth(method='lm')
```

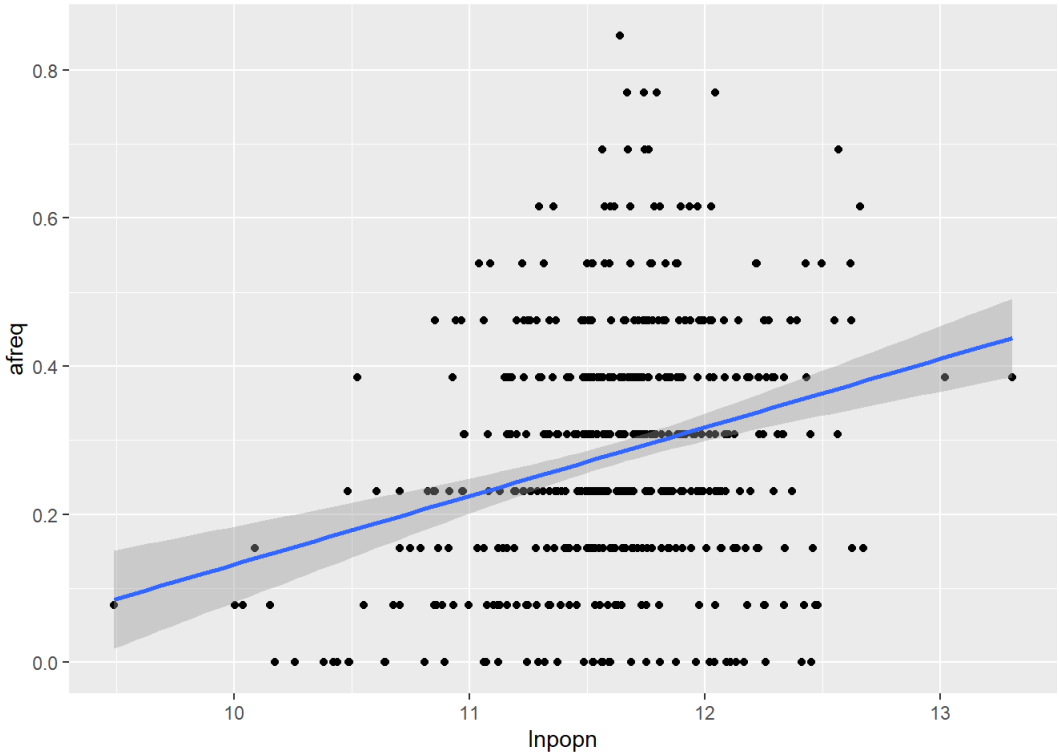
```
## `geom_smooth()` using formula = 'y ~ x'
```



У переменных логарифма населения уезда и логарифма городского населения уезда умеренная взаимосвязь.

```
ggplot(hw, aes(lnpopn, afreq))+
  geom_point()+
  geom_smooth(method='lm')
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

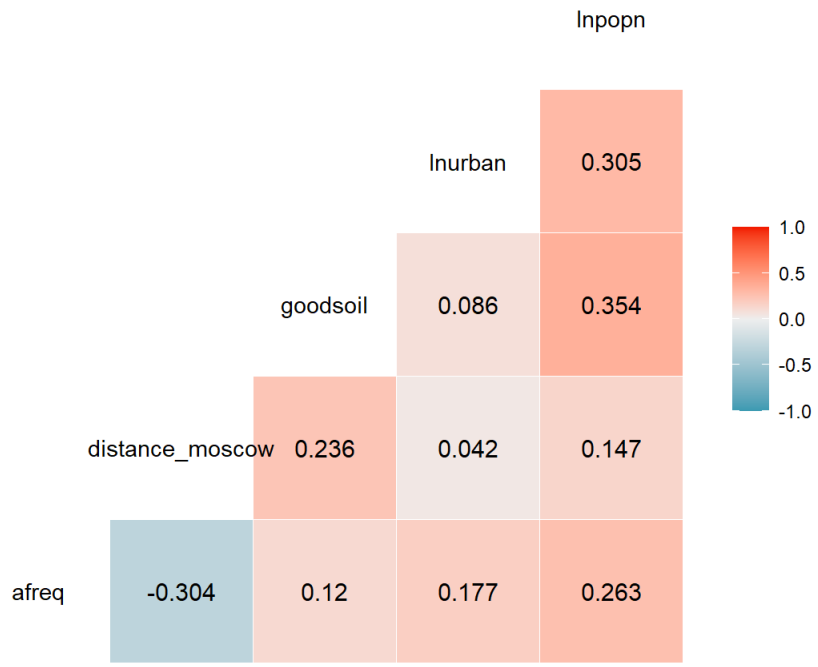


Визуально между переменны просматрвиается взаимосвязь.

На первом этапе диагностики можно предположить, что мультиколлинеарность в нашей модели присутствует.

2. Корреляция между предикторами

```
X_new <- dplyr::select(hw, c(afreq, distance_moscow, goodsoil, lnurban, lnpopn))
ggcorr(X_new, label = T, label_round = 3)
```




```
corrmatrix <- cor(X_new)
round(corrmatrix, 3)
```

```
##                afreq distance_moscow goodsoil lnurban lnpopn
## afreq          1.000         -0.304    0.120   0.177  0.263
## distance_moscow -0.304         1.000    0.236   0.042  0.147
## goodsoil        0.120         0.236    1.000   0.086  0.354
## lnurban         0.177         0.042    0.086   1.000  0.305
## lnpopn          0.263         0.147    0.354   0.305  1.000
```

Присутствует слабая корреляция.

3. VIF-диагностика

Воспомогательные модели:

```
m1v <- lm(afreq ~ nozemstvo + distance_moscow + goodsoil + lnurban + lnpopn + province_capital, hw)
m2v <- lm(nozemstvo ~ afreq + distance_moscow + goodsoil + lnurban + lnpopn + province_capital, hw)
m3v <- lm(distance_moscow ~ afreq + nozemstvo + goodsoil + lnurban + lnpopn + province_capital, hw)
m4v <- lm(goodsoil ~ afreq + nozemstvo + distance_moscow + lnurban + lnpopn + province_capital, hw)
m5v <- lm(lnurban ~ afreq + distance_moscow + goodsoil + nozemstvo + lnpopn + province_capital, hw)
m6v <- lm(province_capital ~ afreq + distance_moscow + goodsoil + nozemstvo + lnurban + lnpopn, hw)
m7v <- lm(lnpopn ~ afreq + distance_moscow + goodsoil + nozemstvo + lnurban + province_capital, hw)

mtable(m1v, m2v, m3v, m4v, m5v, m6v)
```

```
##
## Calls:
## m1v: lm(formula = afreq ~ nozemstvo + distance_moscow + goodsoil +
##     lnurban + lnpopn + province_capital, data = hw)
## m2v: lm(formula = nozemstvo ~ afreq + distance_moscow + goodsoil +
##     lnurban + lnpopn + province_capital, data = hw)
## m3v: lm(formula = distance_moscow ~ afreq + nozemstvo + goodsoil +
##     lnurban + lnpopn + province_capital, data = hw)
## m4v: lm(formula = goodsoil ~ afreq + nozemstvo + distance_moscow +
##     lnurban + lnpopn + province_capital, data = hw)
## m5v: lm(formula = lnurban ~ afreq + distance_moscow + goodsoil + nozemstvo +
##     lnpopn + province_capital, data = hw)
## m6v: lm(formula = province_capital ~ afreq + distance_moscow + goodsoil +
##     nozemstvo + lnurban + lnpopn, data = hw)
##
## ============================================================================================================
##                      m1v         m2v           m3v          m4v         m5v           m6v
##                   ----------  ----------  ---------------  ----------  ----------  ---------------
##                    afreq     nozemstvo   distance_moscow   goodsoil    lnurban    province_capital
## ------------------------------------------------------------------------------------------------------------
##   (Intercept)       -0.833***    0.807       -0.670*       -2.302***    2.220        -1.700***
##                    (0.179)     (0.454)      (0.311)       (0.401)     (1.363)       (0.320)
##   nozemstvo           0.023                   0.320***     -0.273***    0.405**      -0.022
##                    (0.018)                 (0.028)       (0.040)     (0.135)       (0.033)
##   distance_moscow    -0.209***    0.683***                  0.439***   -0.128         0.020
##                    (0.025)     (0.059)                   (0.057)     (0.200)       (0.048)
##   goodsoil            0.051*     -0.331***      0.249***                0.168        -0.088*
##                    (0.020)     (0.048)      (0.032)                   (0.150)       (0.036)
##   lnurban             0.018**     0.045**      -0.007         0.015                  0.101***
##                    (0.006)     (0.015)      (0.010)       (0.014)                   (0.010)
##   lnpopn              0.092***   -0.109**      0.115***      0.202***    0.512***      0.086**
##                    (0.016)     (0.041)      (0.028)       (0.037)     (0.122)       (0.030)
##   province_capital   -0.048      -0.042         0.018        -0.139*     1.734***
##                    (0.025)     (0.063)      (0.043)       (0.057)     (0.172)
##   afreq                          0.143       -0.607***      0.264*      0.988**      -0.156
##                                (0.113)      (0.073)       (0.103)     (0.338)       (0.082)
## ------------------------------------------------------------------------------------------------------------
##   R-squared           0.218       0.276        0.369         0.255       0.276         0.224
##   N                 488         488          488           488         488           488
## ============================================================================================================
##   Significance: *** = p < 0.001; ** = p < 0.01; * = p < 0.05
```

Значения R-квадрат не очень велики.

VIF

```
vif(mod)
```

```
##           afreq       nozemstvo  distance_moscow         goodsoil
##        1.279303        1.380589        1.583726        1.341854
##         lnurban          lnpopn province_capital
##        1.380721        1.377693        1.288597
```

Значения VIF не превышают 10, следовательно, можем сказать, что в нашей модели нет серьёзных проблем связанных с присутствием мультиколлиенарности.

3

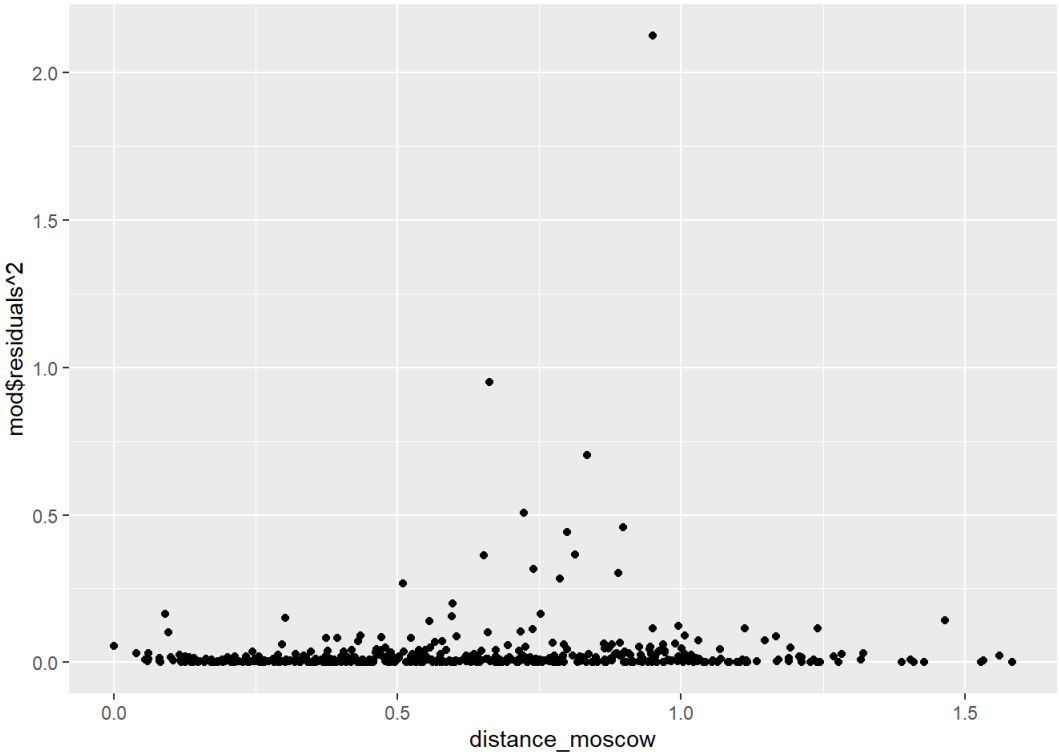
В ходе анализа нашей модели мы уже определили, что наши данные, не имеют линейных связей. Это уже представляет собой значимую предпосылку гетероскедастичности.

4

Гетероскедастичность

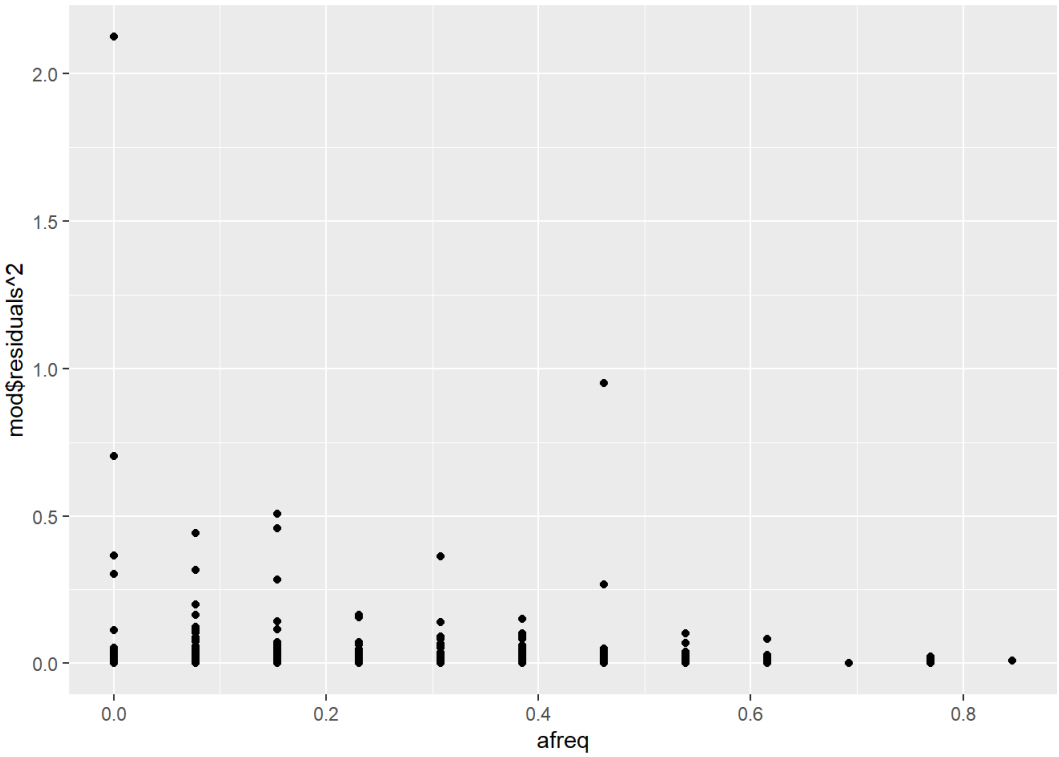
1. Визуальная диагностика

```
ggplot(hw, aes(distance_moscow, mod$residuals^2))+
 geom_point()
```



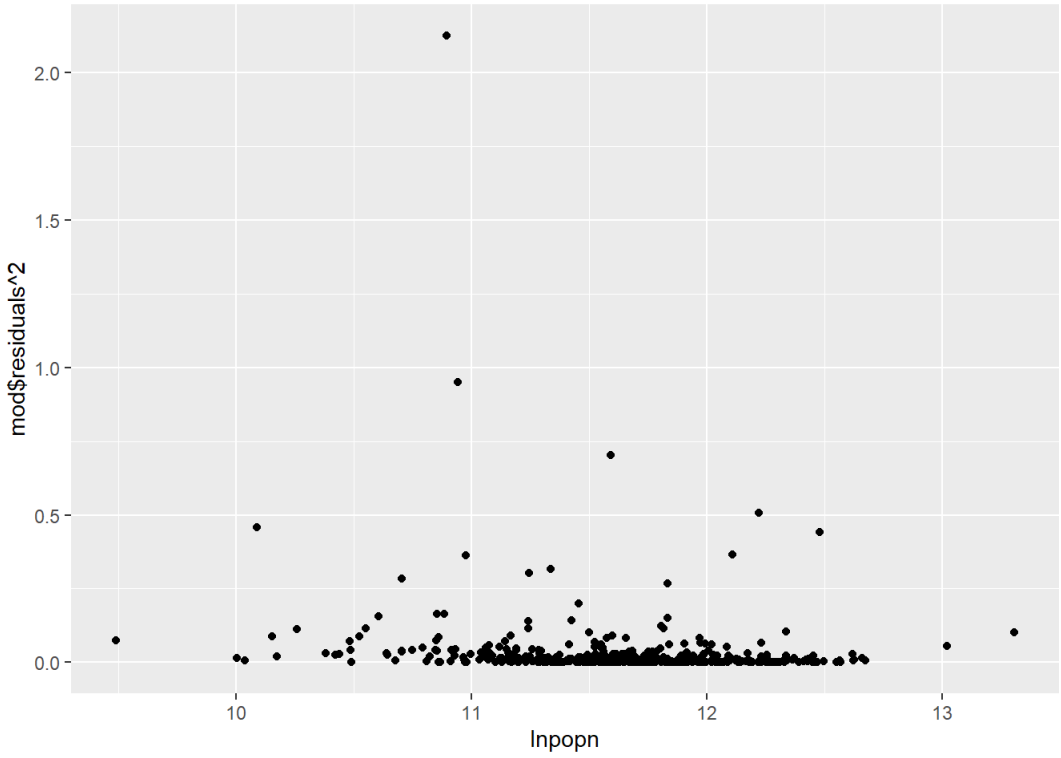
Очень слабовыраженная гетероскедастичность.

```
ggplot(hw, aes(afreq, mod$residuals^2))+
 geom_point()
```



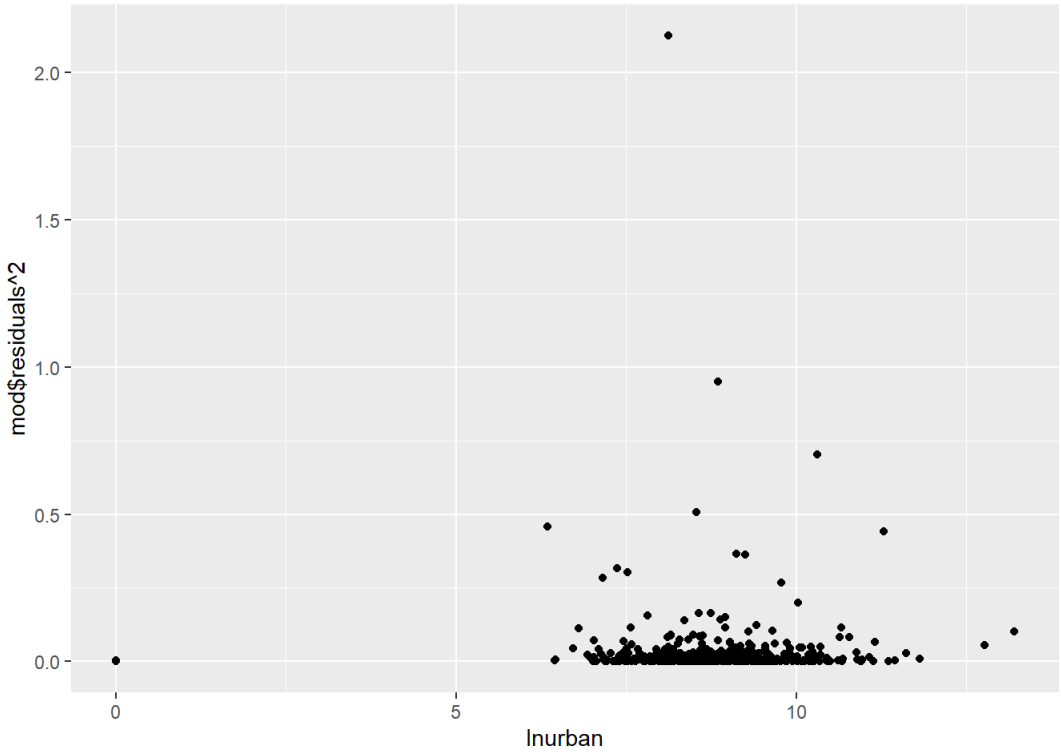
Похоже на явную гетероскедастичность.

```
ggplot(hw, aes(lnpopn, mod$residuals^2))+
 geom_point()
```



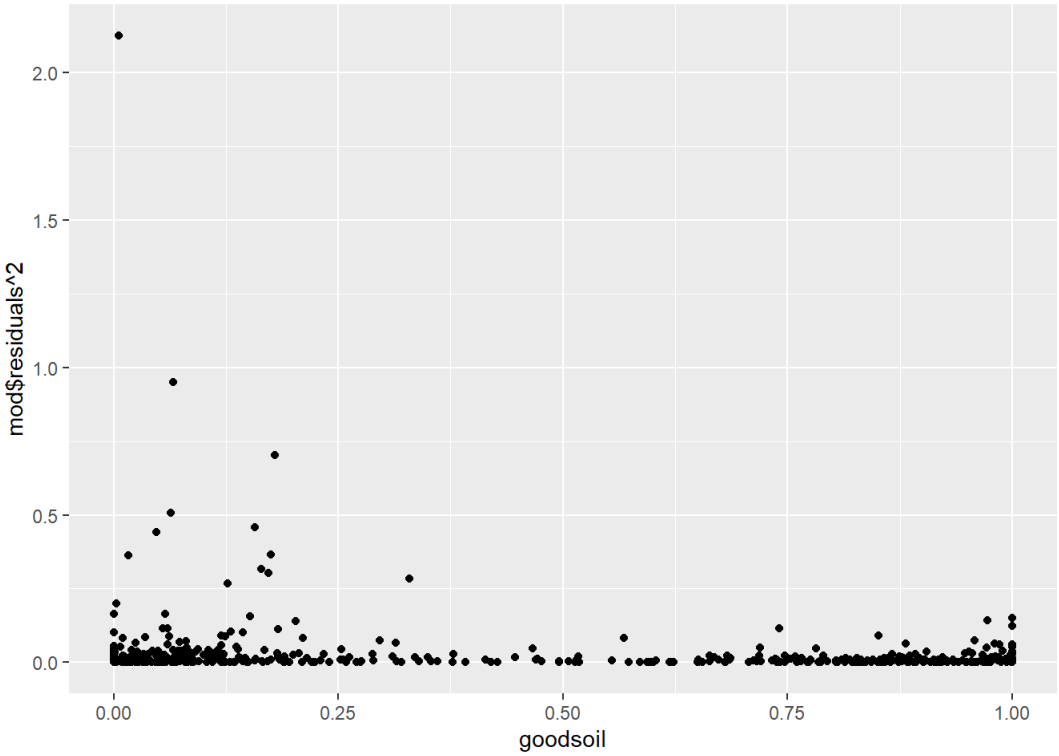
Слабо похоже на гетероскедастичность.

```
ggplot(hw, aes(lnurban, mod$residuals^2))+
  geom_point()
```

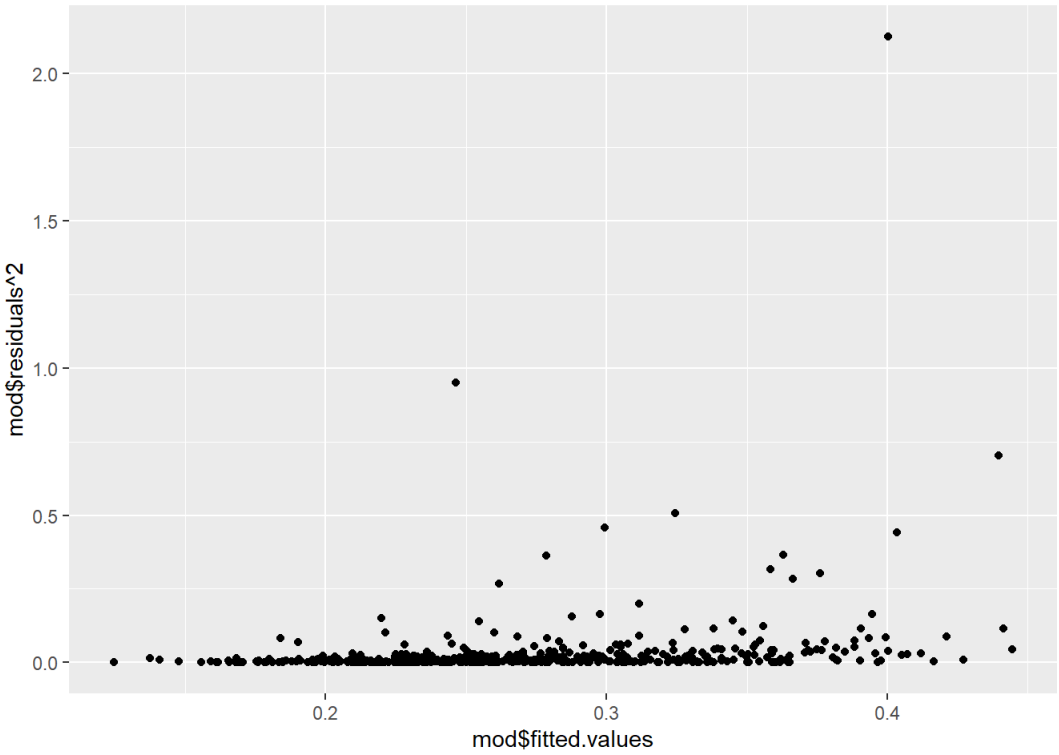


С этой переменной так же слабо похоже на гетероскедастичность.

```
ggplot(hw, aes(goodsoil, mod$residuals^2))+
  geom_point()
```




```
ggplot(hw, aes(mod$fitted.values, mod$residuals^2))+
  geom_point()
```



2. Тест Бреуша-Пагана

```
bptest(mod)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod
## BP = 33.289, df = 7, p-value = 2.338e-05
```

Нулевая гипотеза отвергается. В модели присутствует гетероскедастичность.

3. Тест Goldfeld-Quandt

Применим в этом тесте переменную *afreq*.

```
gqtest(mod, order.by = ~afreq, data = hw, fraction = 0.2, alternative = "two.sided")
```

```
##
##  Goldfeld-Quandt test
##
## data:  mod
## GQ = 0.37299, df1 = 188, df2 = 187, p-value = 3.679e-11
## alternative hypothesis: variance changes from segment 1 to 2
```

Нулевая гипотеза отвергается. Вариация ошибок изменяется. В модели присутствует гетероскедастичность.

4. Использование робастных стандартных ошибок

```
vcovHC(mod)
```

```
##                   (Intercept)         afreq     nozemstvo distance_moscow
## (Intercept)       0.1026071865  7.080502e-03 -2.753538e-03    6.059348e-03
## afreq             0.0070805015  3.845109e-03 -9.991129e-04    1.026828e-03
## nozemstvo        -0.0027535378 -9.991129e-04  7.761460e-04   -4.695218e-04
## distance_moscow   0.0060593484  1.026828e-03 -4.695218e-04    9.831934e-04
## goodsoil          0.0004582821 -9.897354e-05  9.828419e-05   -1.995171e-04
## lnurban          -0.0002225546 -2.762131e-05 -2.357578e-05    1.138073e-05
## lnpopn           -0.0090654903 -7.203197e-04  2.894121e-04   -5.867766e-04
## province_capital  0.0011264548 -1.115038e-04  6.745428e-05   -3.481565e-06
##                       goodsoil       lnurban        lnpopn province_capital
## (Intercept)       4.582821e-04 -2.225546e-04 -9.065490e-03     1.126455e-03
## afreq            -9.897354e-05 -2.762131e-05 -7.203197e-04    -1.115038e-04
## nozemstvo         9.828419e-05 -2.357578e-05  2.894121e-04     6.745428e-05
## distance_moscow  -1.995171e-04  1.138073e-05 -5.867766e-04    -3.481565e-06
## goodsoil          4.615631e-04 -1.951928e-05 -3.628306e-05    -1.899528e-05
## lnurban          -1.951928e-05  2.310414e-05  3.872756e-06    -3.088491e-05
## lnpopn           -3.628306e-05  3.872756e-06  8.168718e-04    -7.710112e-05
## province_capital -1.899528e-05 -3.088491e-05 -7.710112e-05     1.106944e-03
```

```
coeftest(mod, vcov = vcovHC, type = "HC3")
```

```
##
## t test of coefficients:
##
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.6763896  0.3203236  2.1116  0.03524 *
## afreq            -0.1799397  0.0620089 -2.9018  0.00388 **
## nozemstvo         0.0816815  0.0278594  2.9319  0.00353 **
## distance_moscow  -0.0122836  0.0313559 -0.3917  0.69542
## goodsoil         -0.0094062  0.0214840 -0.4378  0.66171
## lnurban           0.0137544  0.0048067  2.8615  0.00440 **
## lnpopn           -0.0420321  0.0285810 -1.4706  0.14205
## province_capital  0.0387709  0.0332708  1.1653  0.24447
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(mod)
```

```
##
## Call:
## lm(formula = ch_schools_pc ~ afreq + nozemstvo + distance_moscow +
##     goodsoil + lnurban + lnpopn + province_capital, data = hw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33950 -0.09568 -0.03539  0.04702  1.45789
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.676390   0.218253   3.099 0.002055 **
## afreq            -0.179940   0.054391  -3.308 0.001009 **
## nozemstvo         0.081681   0.021824   3.743 0.000204 ***
## distance_moscow  -0.012284   0.031880  -0.385 0.700184
## goodsoil         -0.009406   0.024005  -0.392 0.695347
## lnurban           0.013754   0.007281   1.889 0.059494 .
## lnpopn           -0.042032   0.019883  -2.114 0.035030 *
## province_capital  0.038771   0.030189   1.284 0.199664
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1782 on 480 degrees of freedom
## Multiple R-squared:  0.1032, Adjusted R-squared:  0.09014
## F-statistic: 7.892 on 7 and 480 DF,  p-value: 4.526e-09
```

Коэффициент при предикторе *lnurban* перестал быть статистически значимым на 0.1% уровне значимости. (Теперь он значим на 0.01% уровне значимости) Коэффициент при предикторе *lnpopn* перестал быть статистически значимым. Замена стандартных ошибок заметно изменила нашу модель.