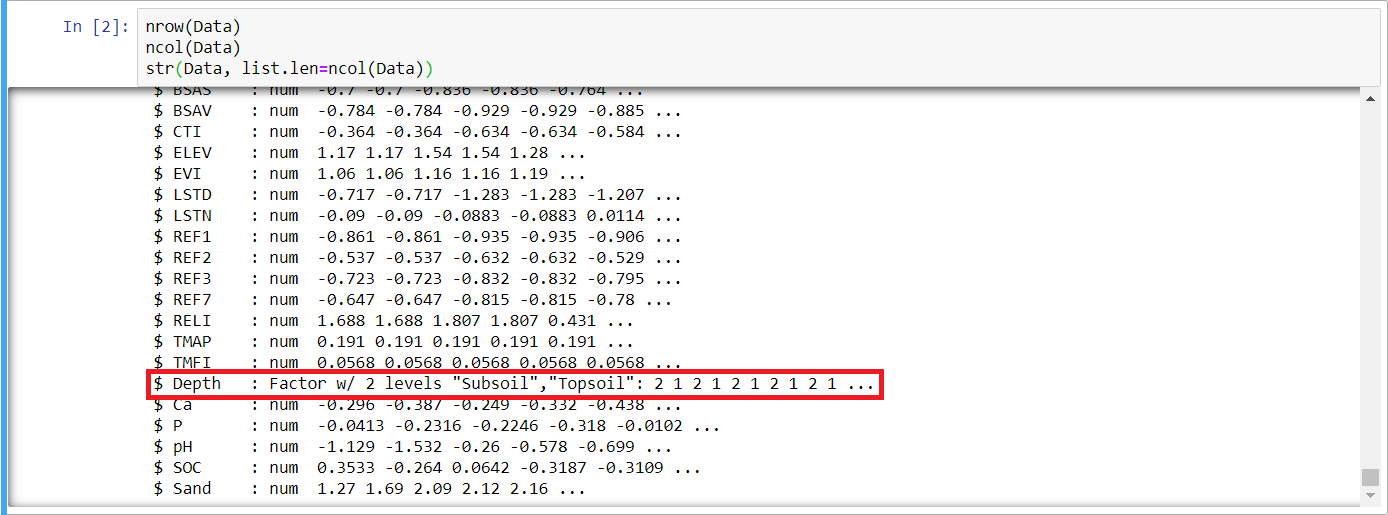
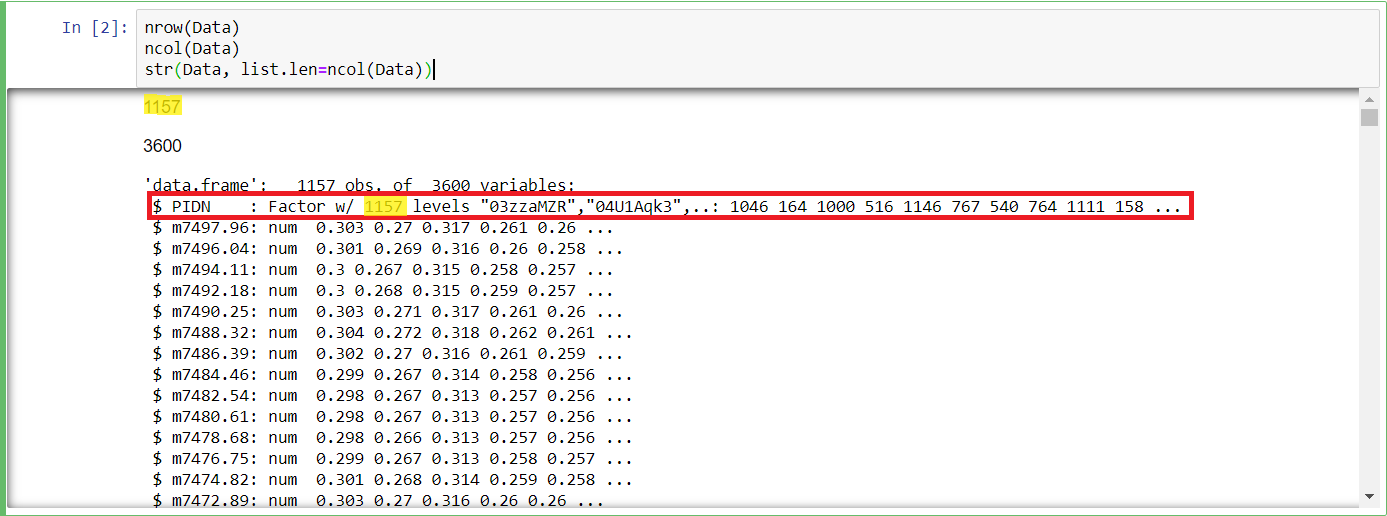
**Отчет по анализу измерений параметров участков почвы**

***Разведочный анализ данных:***

Данные составлены из 1157 измерений. Измерения проводились с 3600 различными параметрами для каждого участка земли. Среди этих параметров, двое имеют категориальные значения: PIND – идентификатор измерения и Depth – глубина почвы (Topsoil – ближе к поверхности и Subsoil – дальше от поверхности). Очевидно, что идентификатор измерения никак не влияет на содержание тех или иных исследуемых элементов, так что его можно сразу исключить и не рассматривать как предиктор.

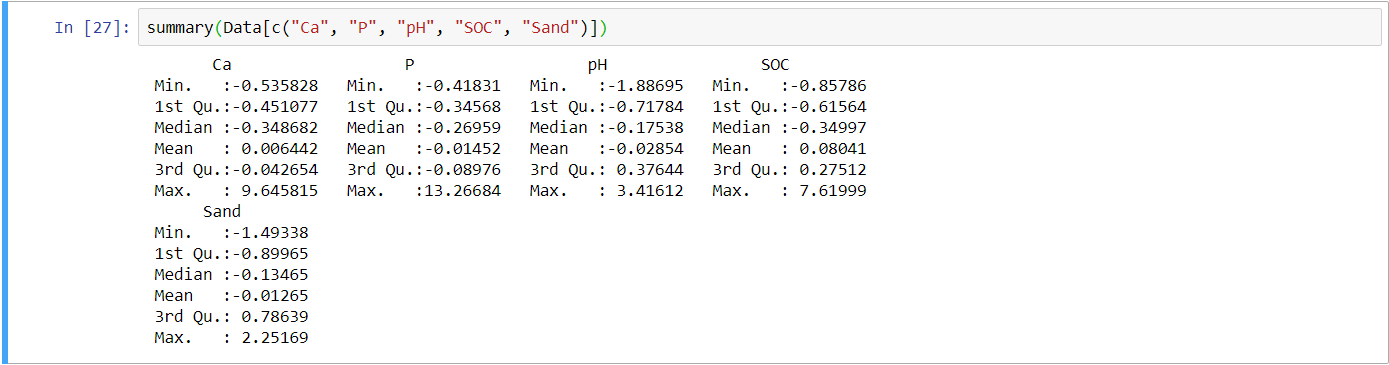


Проверим данные на пропущенные значения (равные null)



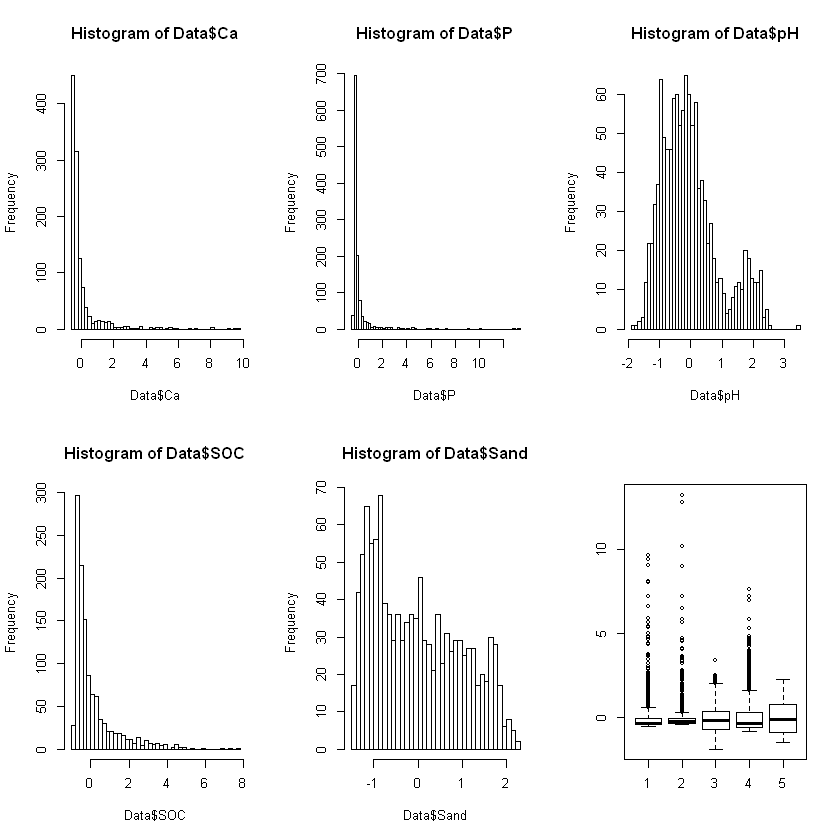
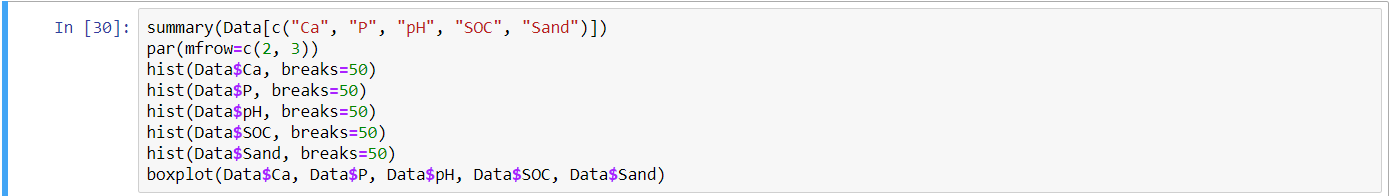
Можно заметить, что среди данных нет пустых элементов.

Применим метод summary к зависимым переменным:



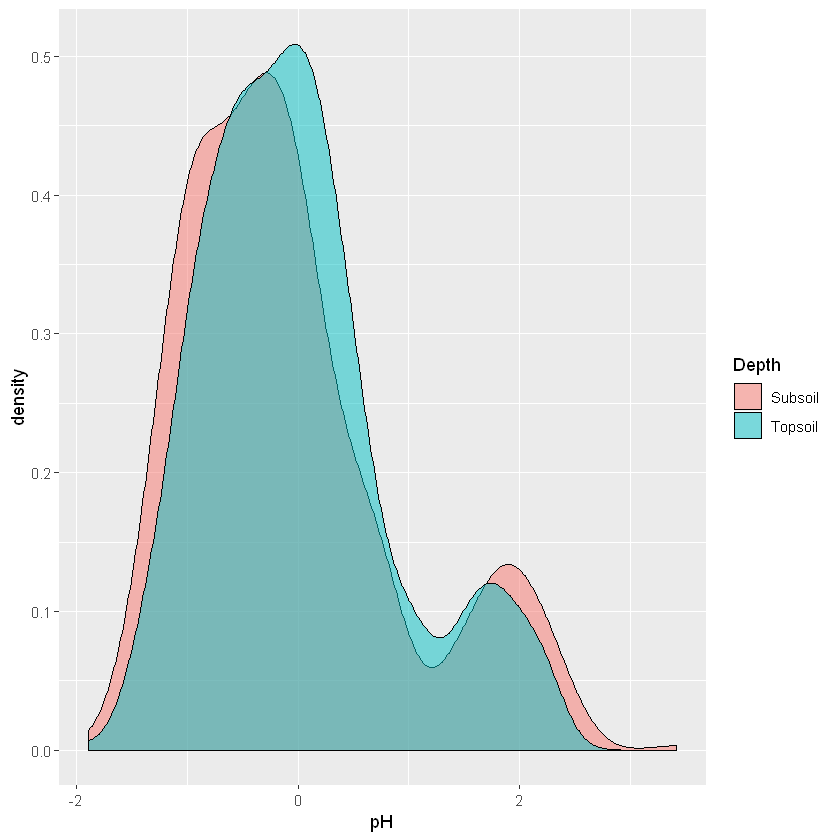
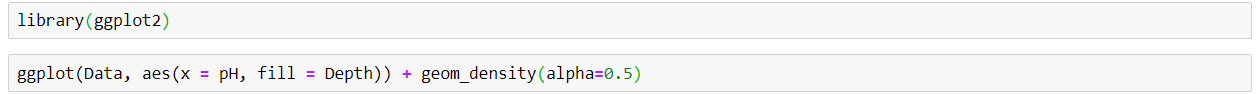
Из полученных данных можно выделить, что значения измерений Ca по большей части расположены ближе к минимальному элементу, т.к. медиана очень близка к первой квантили, причем и 3-я квантиль расположена не далеко, в то время как максимальное значение расположено очень далеко от этих значений. Такое же поведение прослеживается для P и SOC. То есть эти факторы ассиметричны. В то время, как для полей pH и Sand дела обстоят по-другому: расстояние между квантилями приблизительно равномерно распределена, особенно для песка.

Построим гистограммы и ящики с усами для визуализации такого распределения значений:



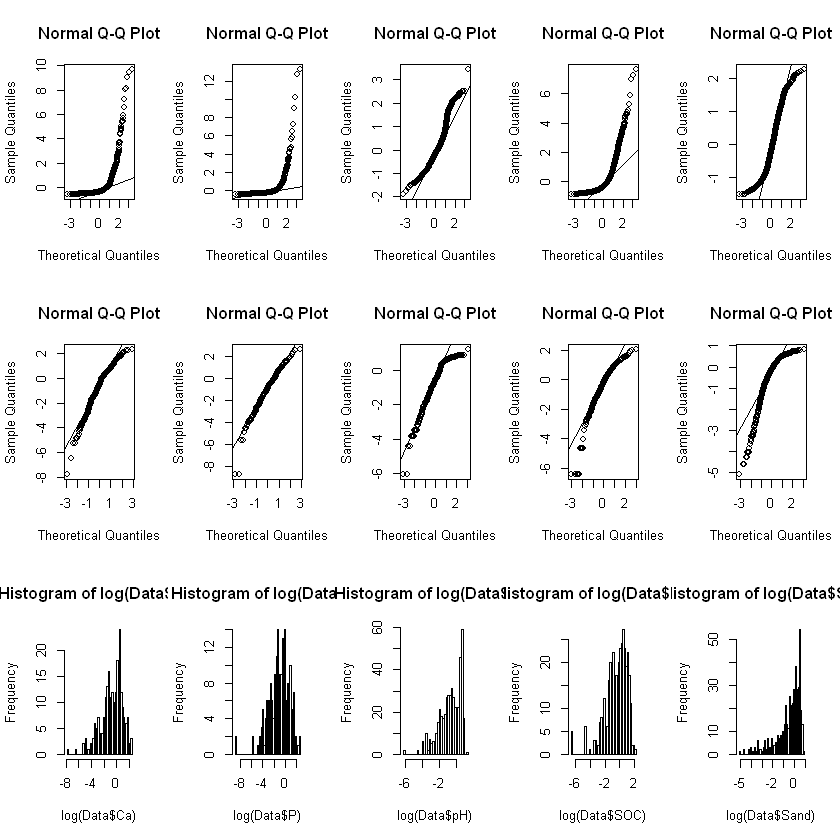
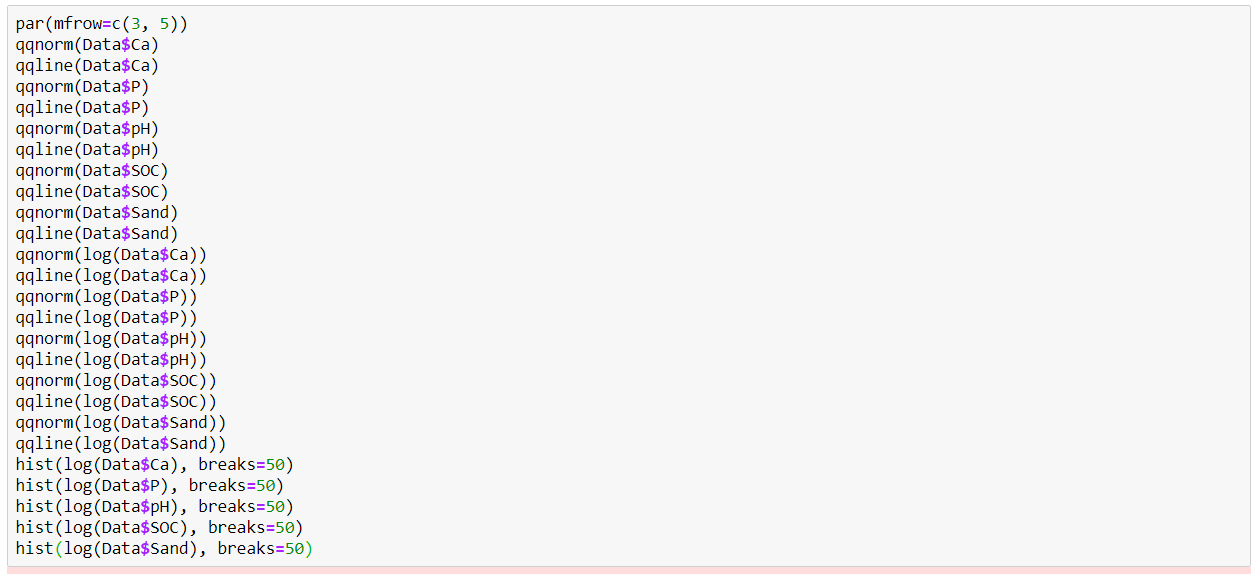
Собственно, гистограммы визуализируют описанное распределение данных и из-за такого поведения, на графике ящика с усами можно увидеть большое количество выбросов. Причем по pH можно предположить, что есть какие-то два вида исследований, дающие разные значения (до 1 и после), то есть прослеживается бимодальность распределения.

Первая мысль по этому поводу, что значение категориального фактора Depth и делит выборку на 2 части, что и приводит к бимодальности. Давайте проверим эту теорию.

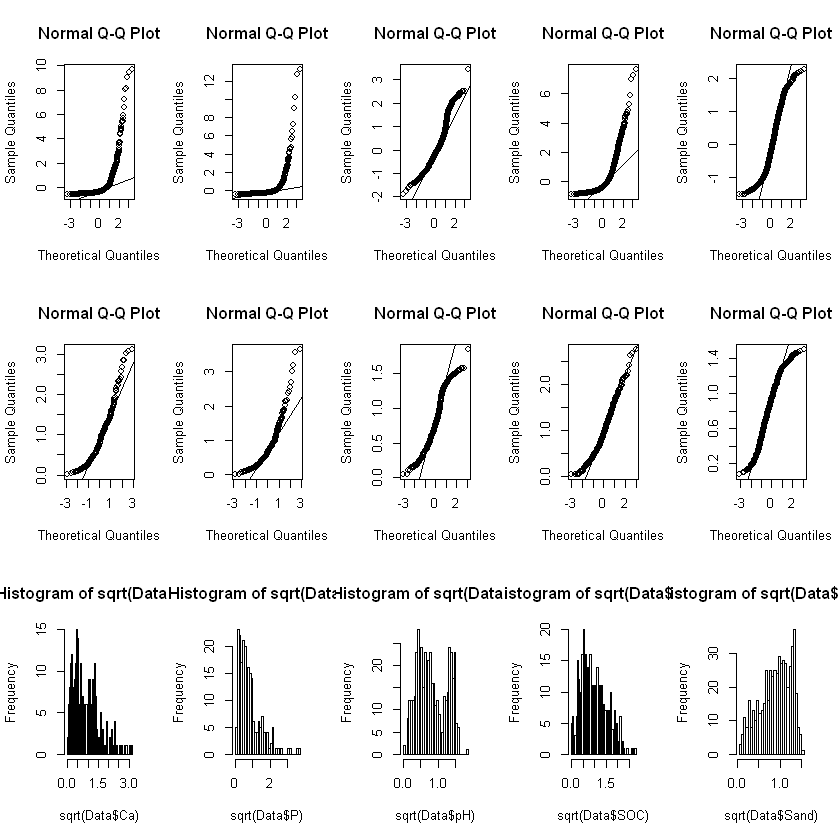


По данному графику мы видим, что нет, глубина измеряемой почвы не является причиной бимодальности распределения водорода (pH).

Теперь на будущее попробуем нормализовать эти факторы и посмотрим, что из этого получится. Сравним нормализацию по получению логарифма и квадратного корня этих значений. Это может быть полезно для приведения выборочных данных к глобальной выборке, при условии, что она нормально распределена в свою очередь. (Но мы не можем 100% знать о распределении факторов почвы по всему миру, так что сравним модели для нормализованных значений и ненормализованных).

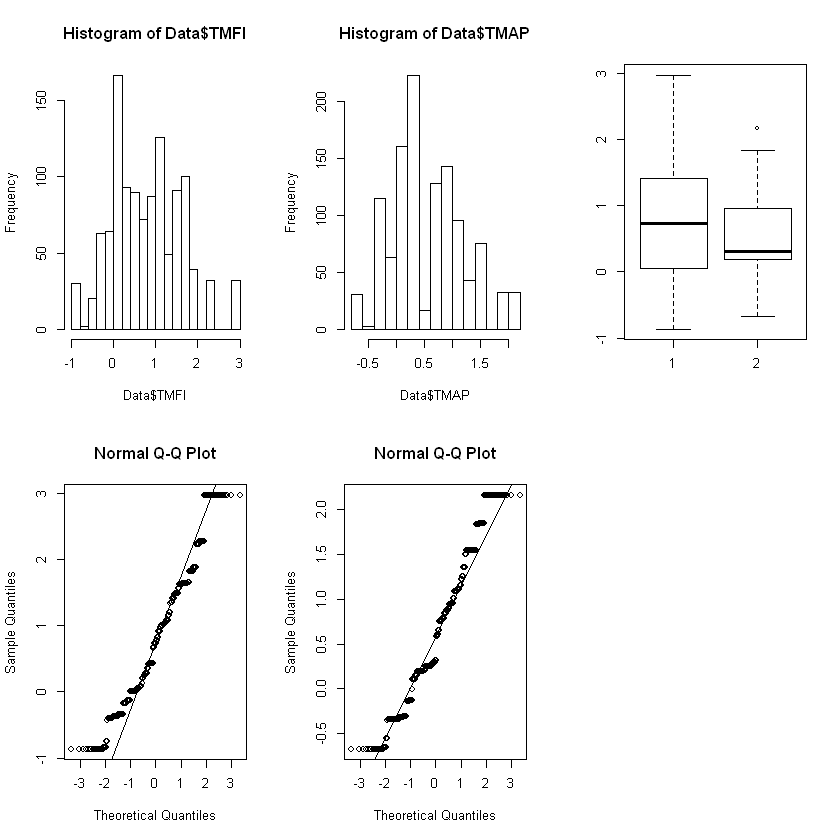
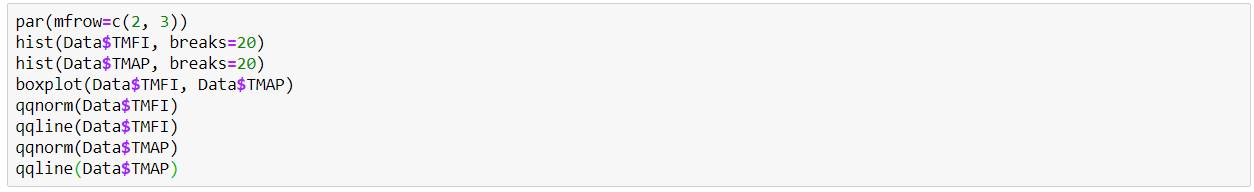


Для квадратного корня просто заменяем команду log на sqrt.

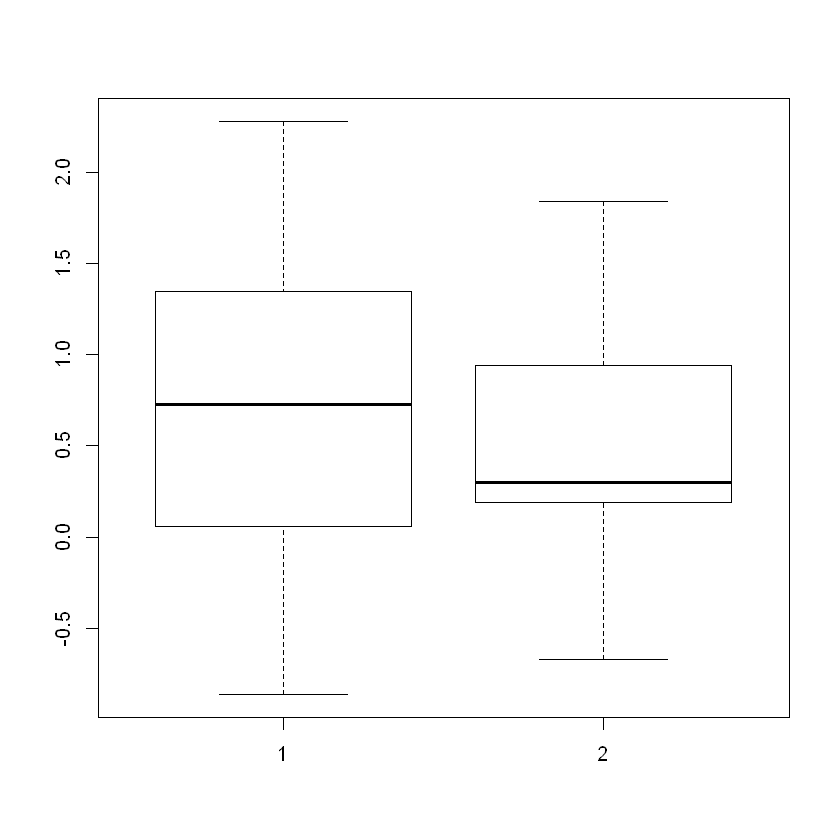
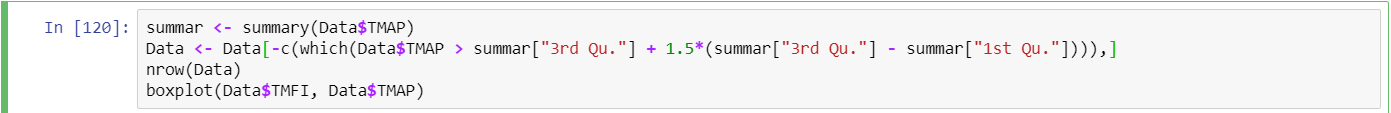


Какой вывод можно сделать по этим графикам? Ну, для кальция (Ca) и фосфора (P) логарифм неплохо справляется с нормализацией в отличии от квадратного корня, который неплохо себя проявил для углерода (SOC). Для песка (Sand) и водорода (pH) лучше себя проювило взятие в квадрат, хоть существенной разницы и не видно особо.

Теперь рассмотрим измерения влажности почвы (TMAP и TMFI):

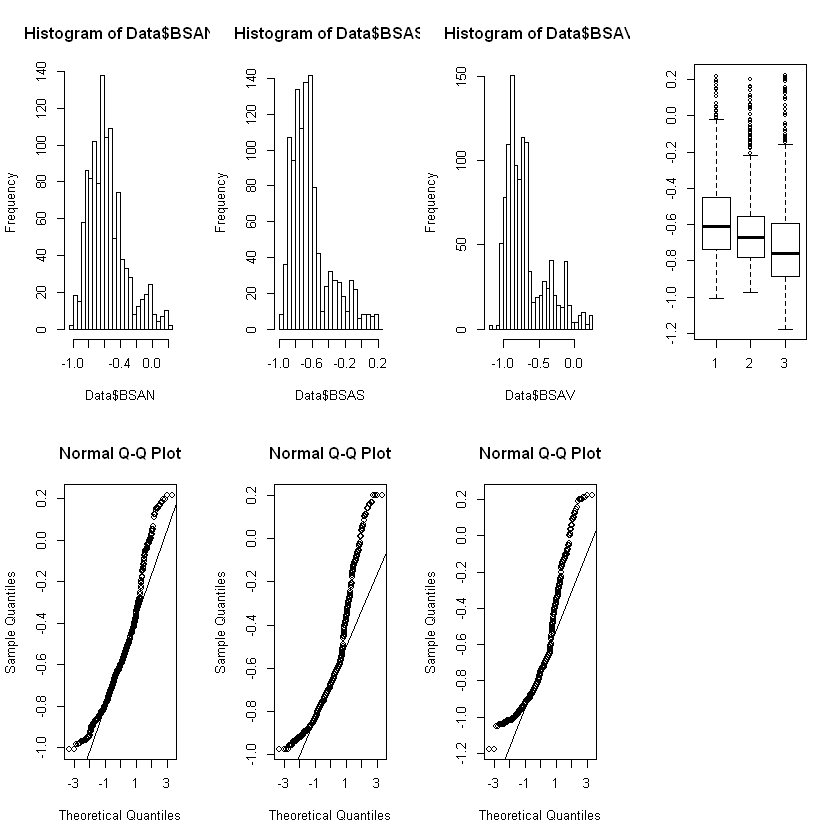


Здесь, мы можем выделить 2 факта: 1) в графике QQPlot точки не расположены непрерывно, это может значить, что здесь многие значения перескакивают возможные значения; 2) TMAP имеет некоторые выбросы, что может плохо сказаться на итоговую модель. Так что, избавимся от выбросов:



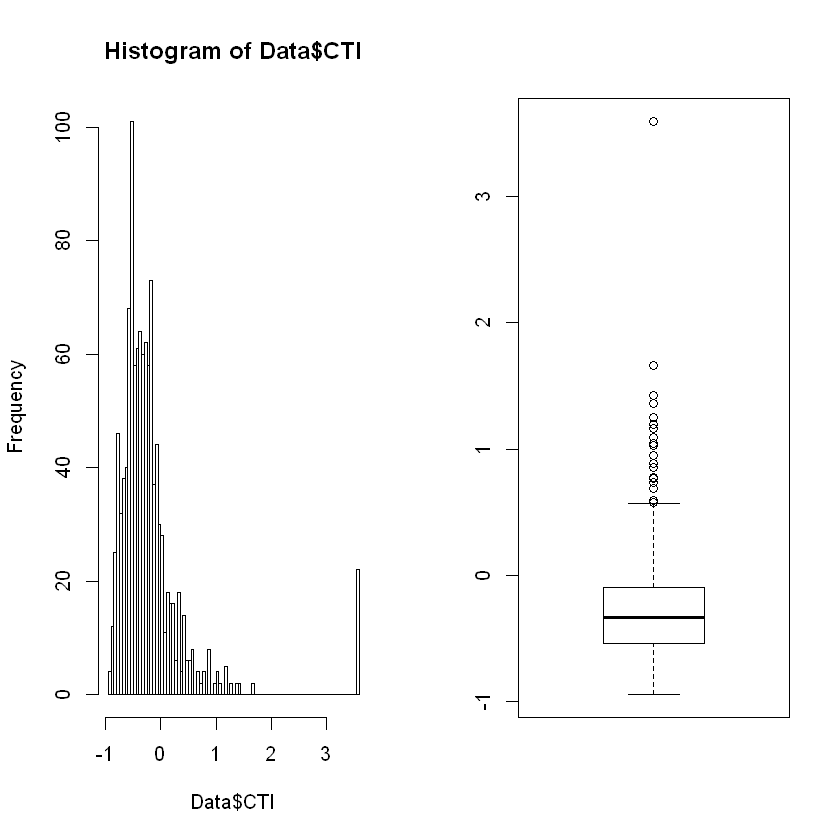
Данной командой было удалено 32 записи, имеющих выбросы по показателям влажности.

Проведем аналогичные наблюдения для показаний отражения различных лучей от поверхности (BSAN, BSAS, BSAV):

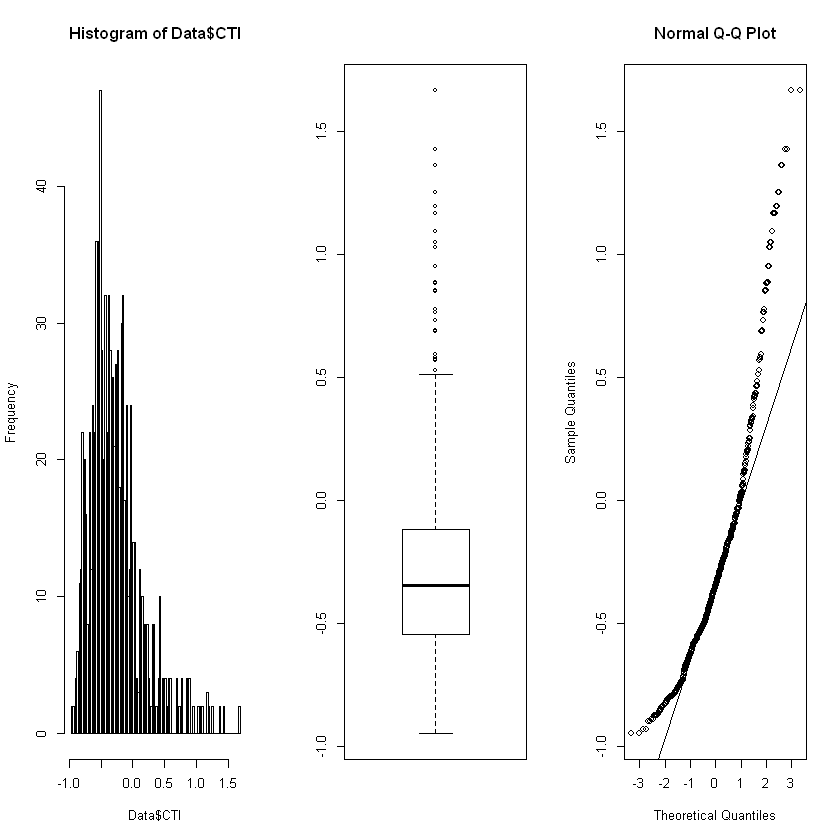


Здесь, также можно увидеть большое количество выбросов в исследованиях, но тут уже появляется подозрение что выборка делится на 2 разные категории исследования из чего и следуют эти выбросы, так что избавляться от них мы пока не будем. Ну и распределение по данным факторам ассиметричны.

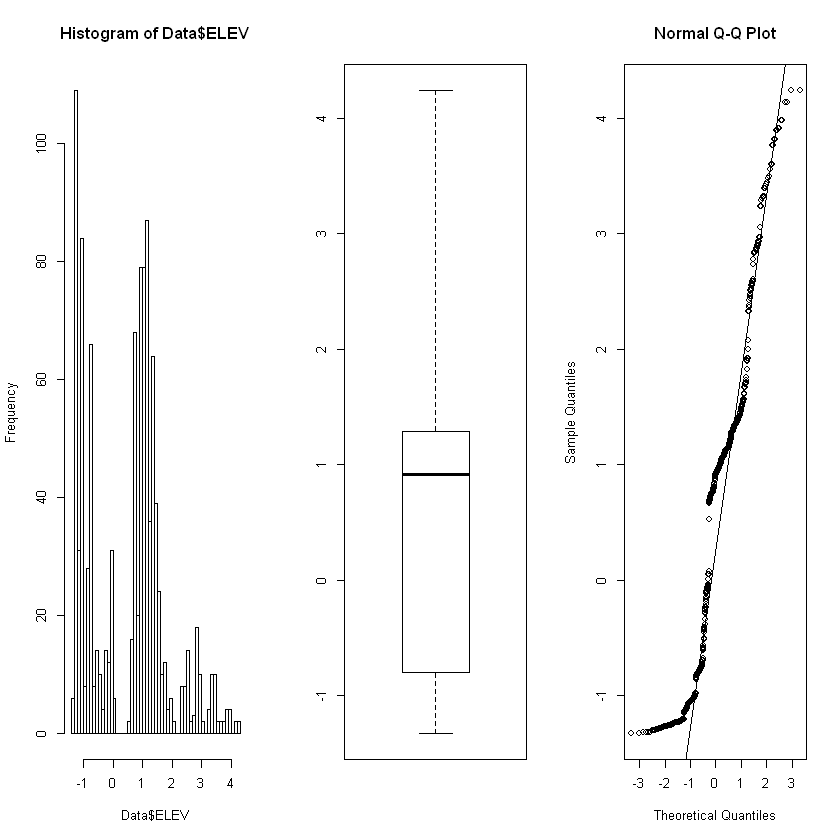
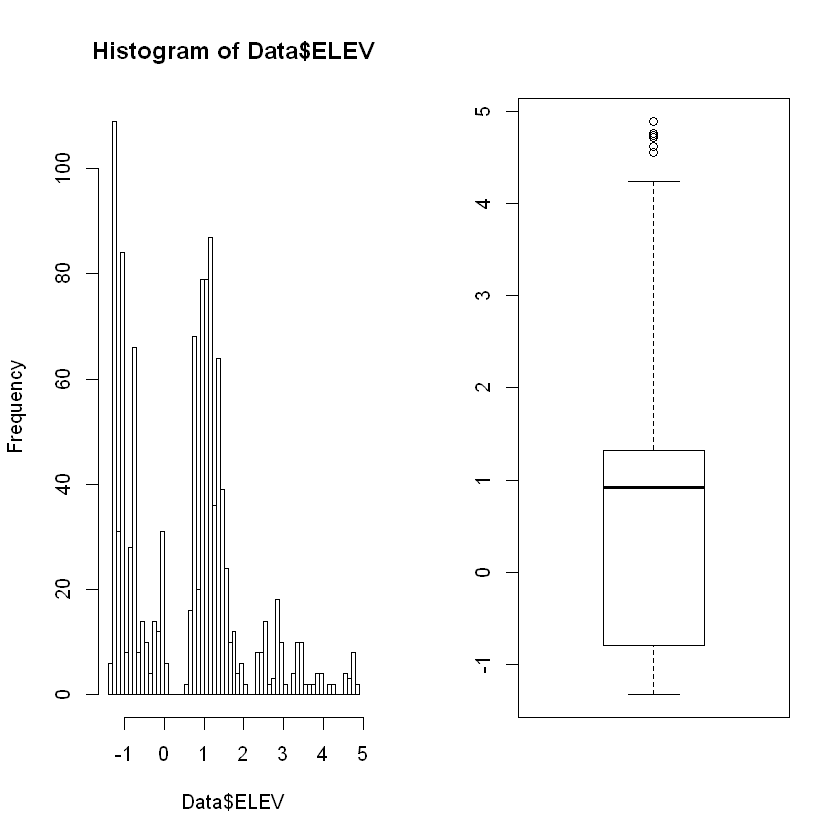
Для показателей влажности CTI можно увидеть, что есть показатели очень сильно отличные от остальных, возможно измерения были сделаны в период длинной засухи или наоборот в период обильных дождей. В любом случае эти исключительные показания стоит исключить (- 22 строки).



После удаления получаем следующие результаты:

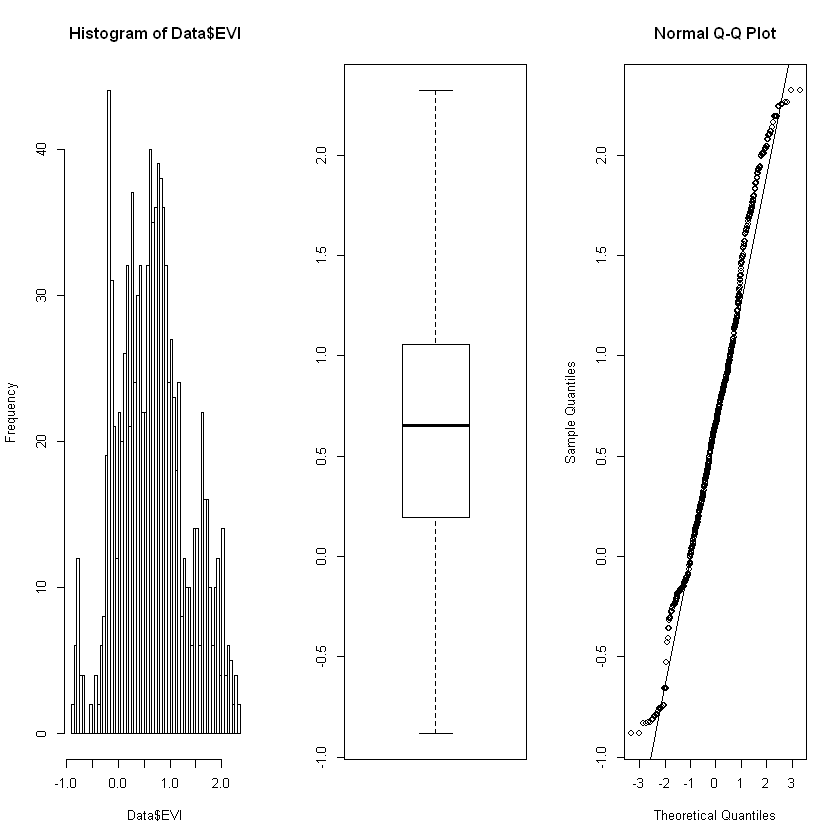
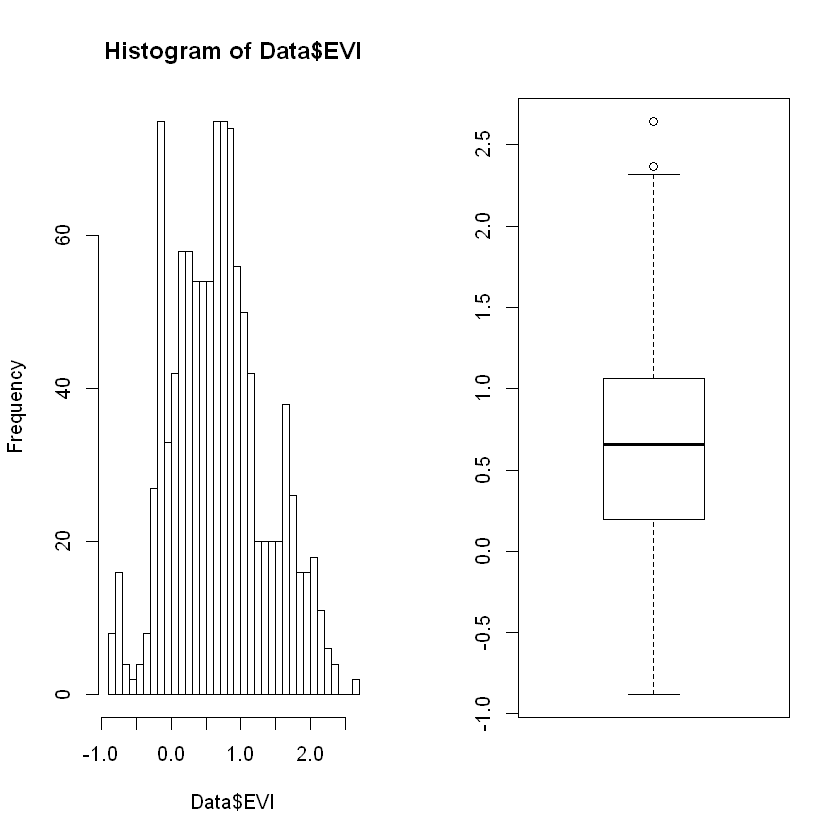


Такую же процедуру проведем для измерений высоты поверхности ELEV (- 17 строк) в исследуемой области:



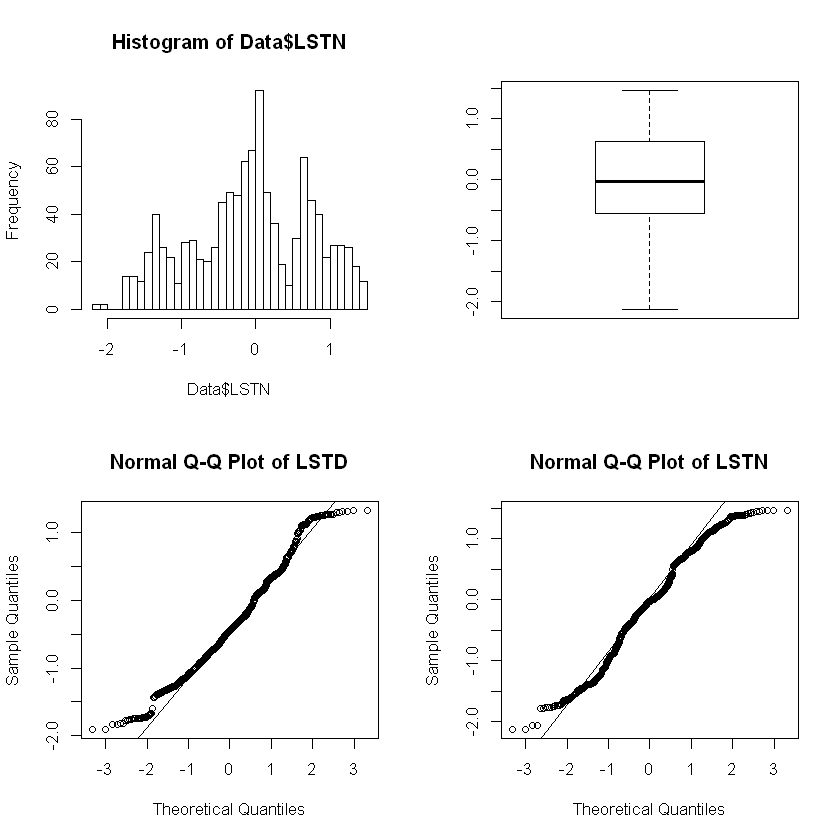
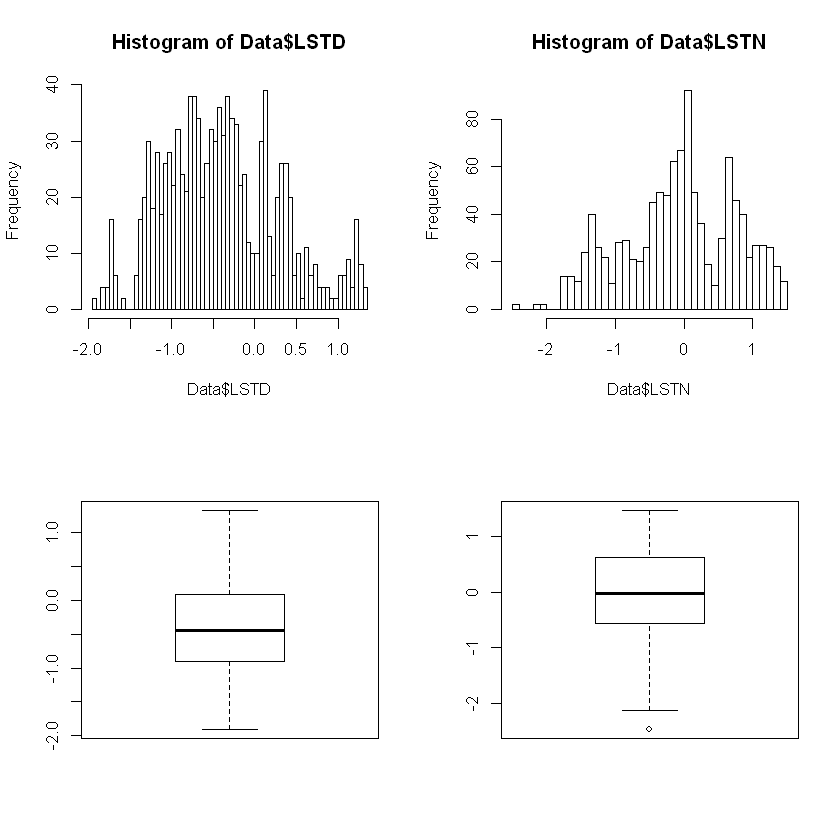
Из-за бимодальности не удается нормализавать данное значение по средствам взятия квадратного корня или логарифма.

Также и для показателей растительности – избавимся от выбросов (- 4 строки):



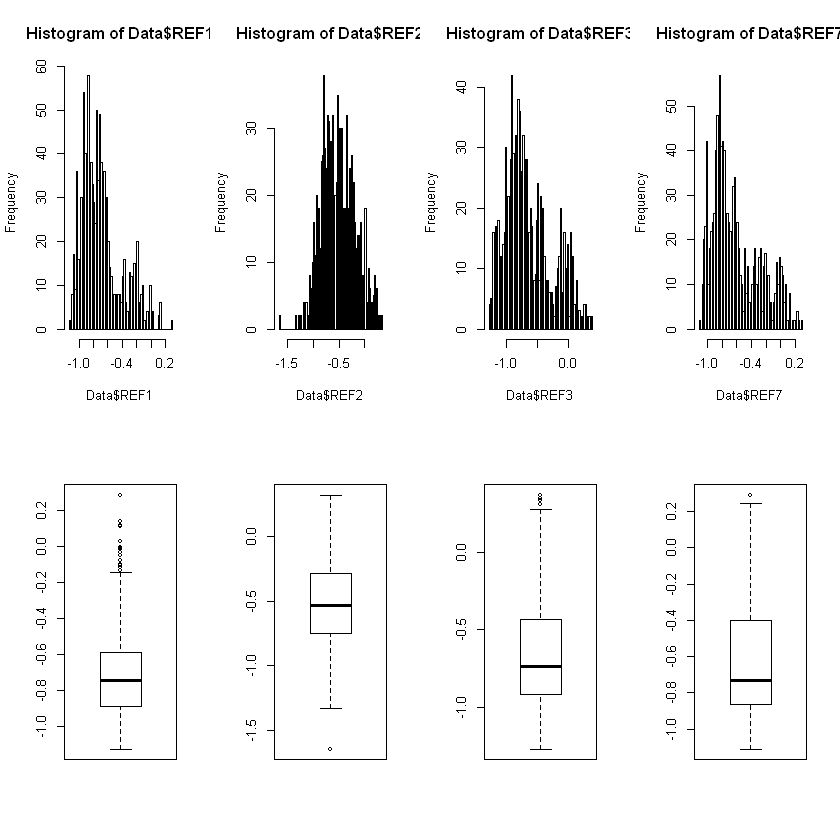
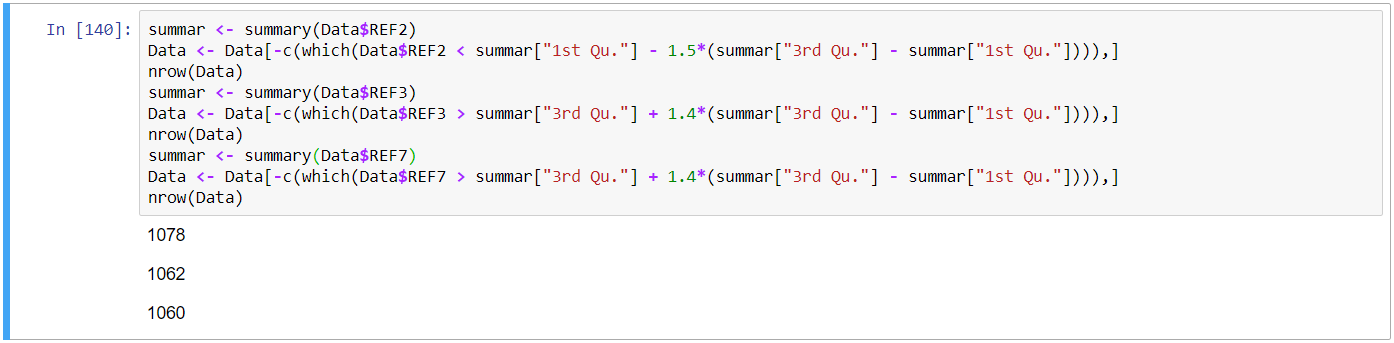
И кстати здесь мы видим более или менее похожее распределение на нормальное.

Далее продолжаем избавление от выбросов для температуры днем и вечером в области исследования почвы (-2 строки):

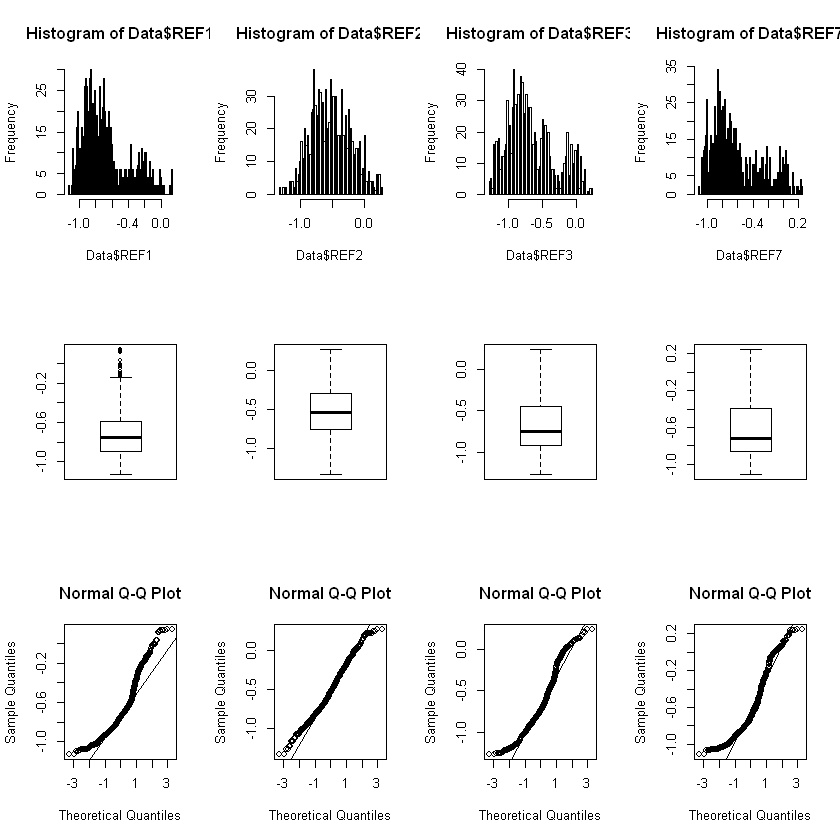


Здесь, также не сказать, что распределение нормально распределена, но она схожа на нормальную и логарифмирование и взятие квадратного корня, только усугубляют положение.

Далее, параметр отражения электромагнитных волн. Также избавимся от выбросов для REF2, REF3 и REF7 (- 20 строк):

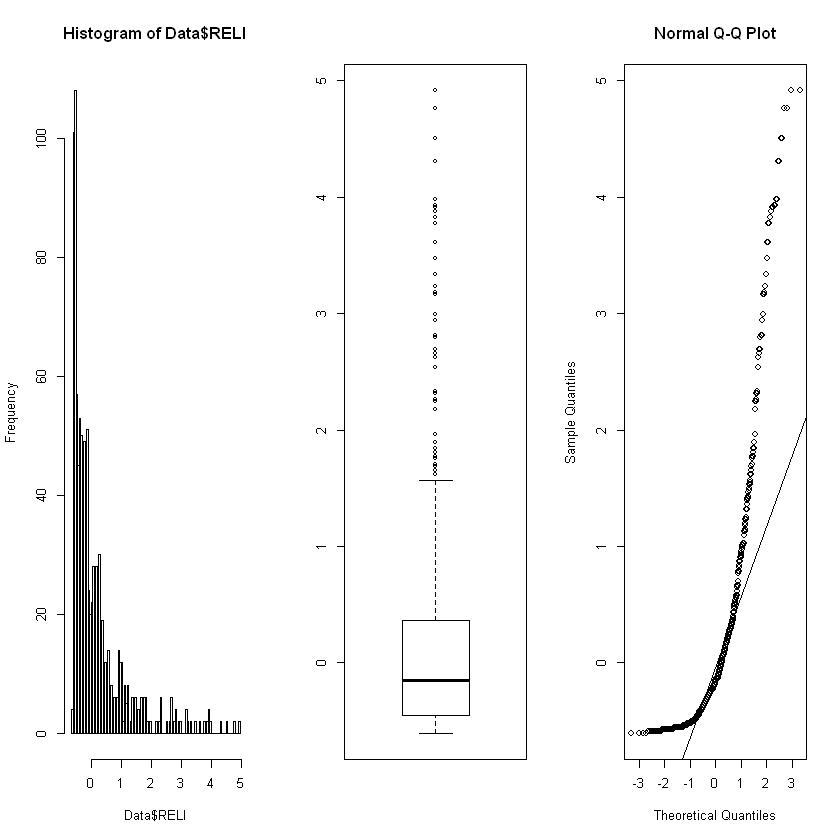


После преобразований:

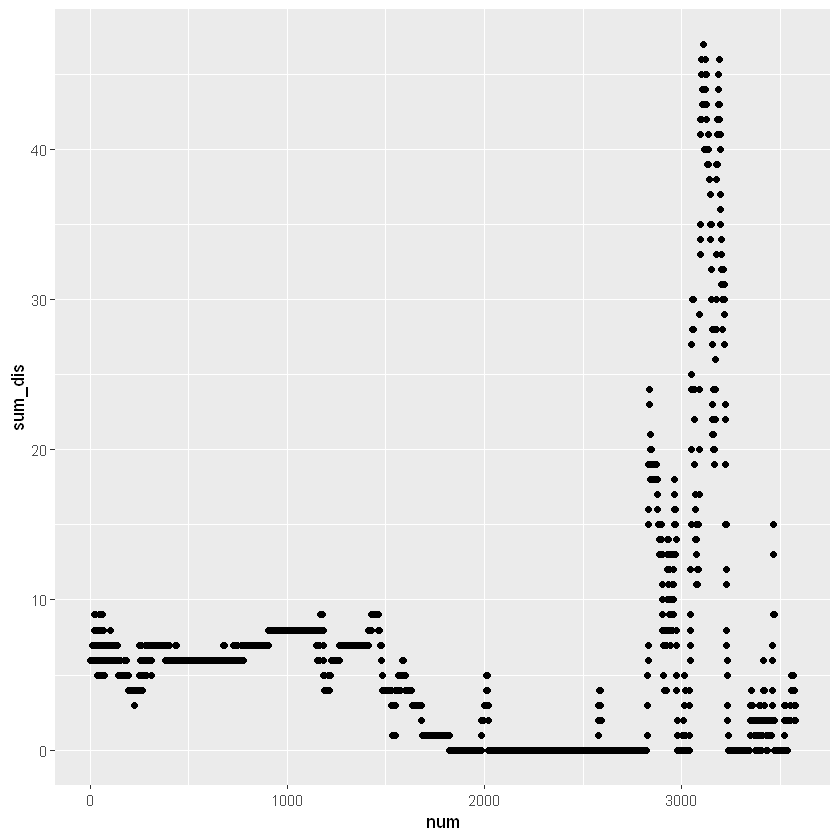


Также по графикам QQPlot видно, что распределение значений REF2 схож с нормальным распределением, у REF3 подозрение на бимодальность, а REF1 и REF7 ассинхронны.

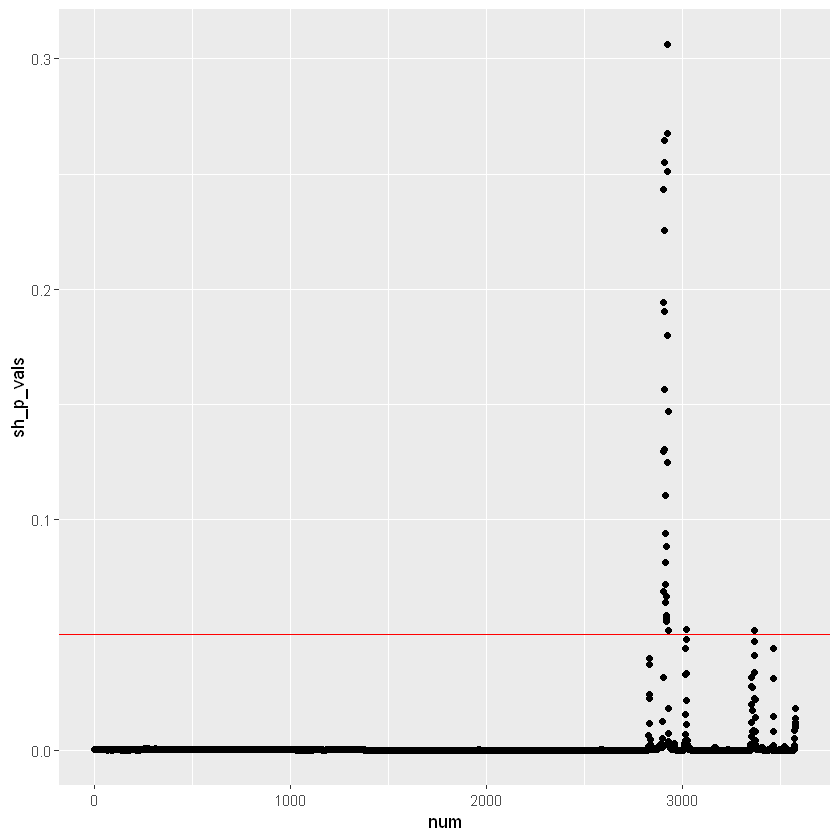
У рельефа области исследования (RELI) также прослеживается сильная ассинхронность распределения значений и большое количество выбросов из-за этого, но избавляться от них я считаю не стоит, ведь возможно получиться пронормализовать это распределение и квартили передвинуться вправо и в таком случае выбросы таковыми уже не будут являться.



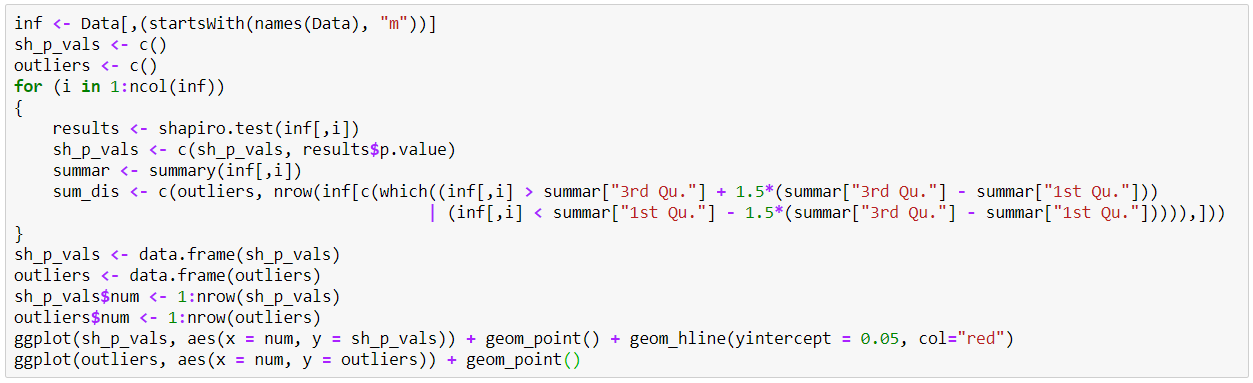
Что же касается инфракрасной спектроскопии по 3 578 различным частотам, сложно оценить такое количество данных, но небольшой анализ все же можно провести. Посчитаем количество выбросов для каждой имеющейся частоты:



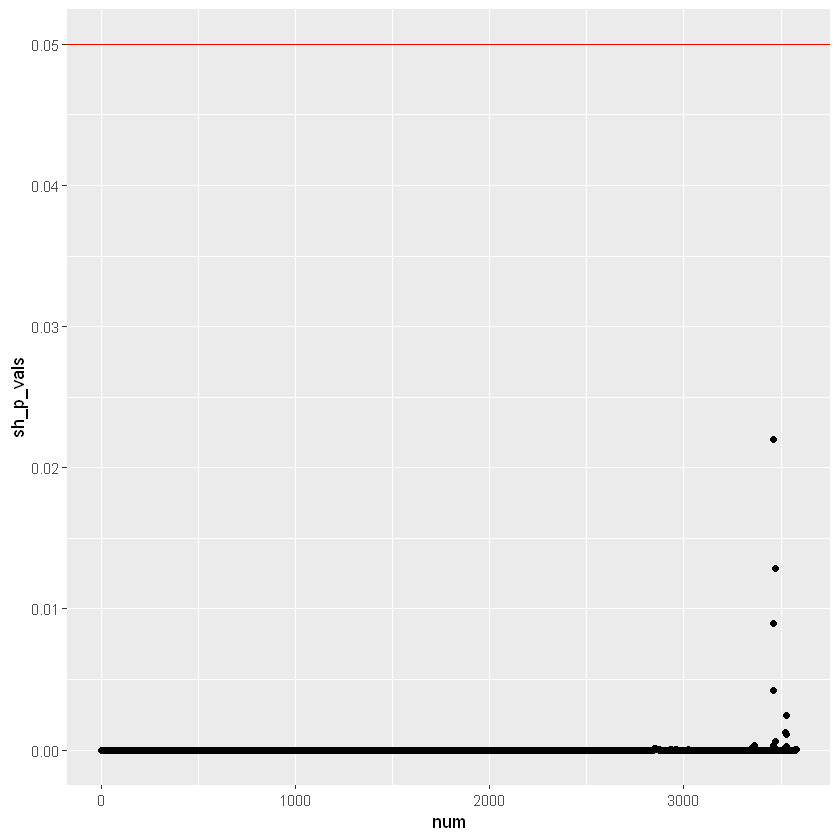
Оценить же нормальность распределения по графикам QQPlot для каждой частоты уже не получиться, так как таких факторов слишком много. Мы видим, что при удалении исследований с выбросами по частотам приведет к большой потере данных. Можно провести автоматический тест Шапиро-Уилка по каждой частоте и посмотреть, есть ли значения p больше 0,05, что будет означать, что мы не можем отвергнуть гипотезу о том, что рассматриваемое распределение нормально.



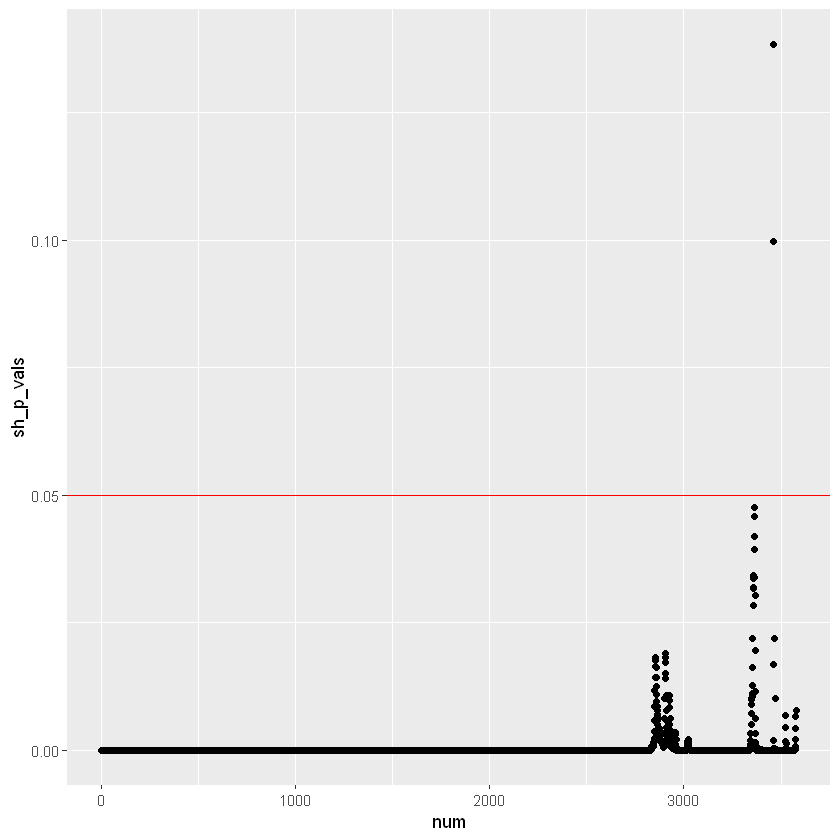
Как мы видим, большинство частот имеет ненормальное распределение. Скрип для получения таких результатов:



Если логарифмировать эти значения с целью нормализации значений, то ничего хорошего из этого не получиться:



Также и с квадратным корнем:

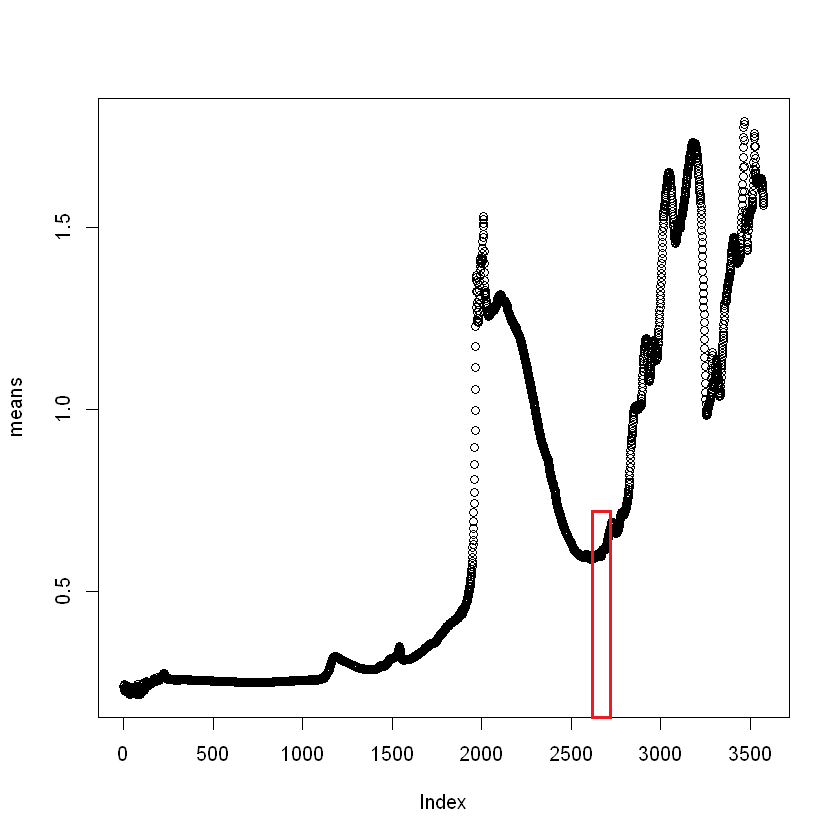


Следовательно, мы наверно не будем нормализовывать эти данные во время построения моделей.

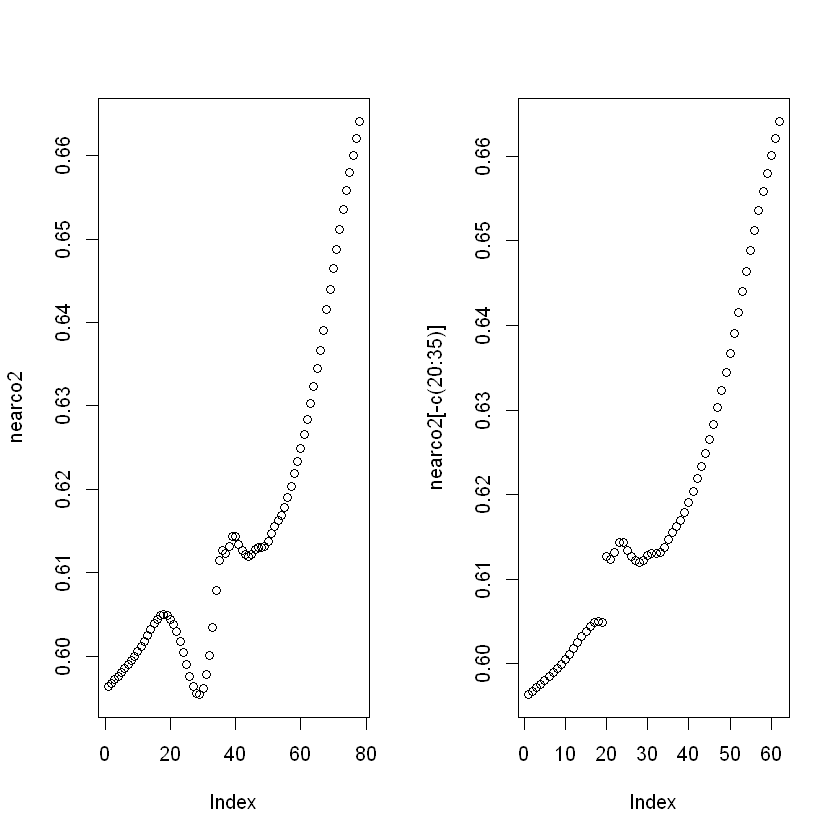
Стоит отметить, что по данным частотам есть замечание в описании структуры данных:

«We suggest you to remove spectra CO2 bands which are in the region m2379.76 to m2352.76, but you do not have to»

Нам предлагается удалить спектры, связанные с диоксидом углерода CO2. Но что же здесь ни так? Будем характеризовывать спектры через их среднее. Построим график распределения этих средних:

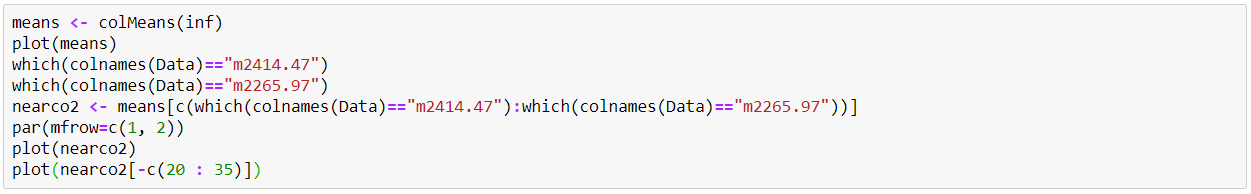


В отмеченной области и предполагается, что находятся эти значения. Давайте приблизимся к данной области графика:



Здесь, мы видим, как изменится график при удалении указанных спектров. На мой взгляд этот просадок не критичен для нашей модели, так как этот просадок не значителен в сравнении с общим графиком средних.

Скрип для получения таких результатов:



И заключающие преобразования, которые здесь стоит применить, это “Label encoding”: перекодируем в числовом виде (в виде 0 и 1) параметр глубины измерения (Depth) (если “Topsoil”, то 1, иначе 0) и избавимся от идентификаторов измерений, так как они не несут в себе никакой значимости:



***Построение регрессионных моделей.***

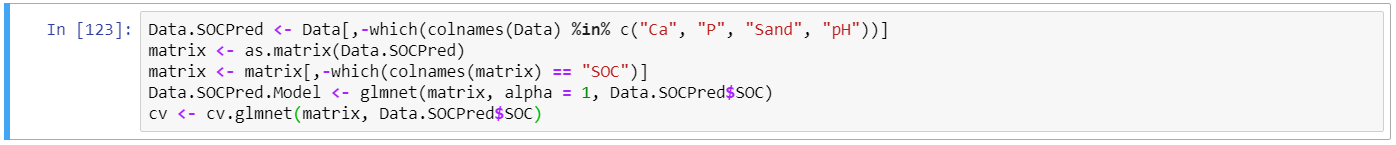
Очевидно, что из-за большого количества факторов в нашей выборке, мы столкнемся с проблемой мультиколлинеарности. Тогда, используем штрафы для среднего квадратов остатков, то есть будем штрафовать модель за слишком большие коэффициенты. Используем 2 такие модели:

Ридж-регрессия:

LASSO регрессия:

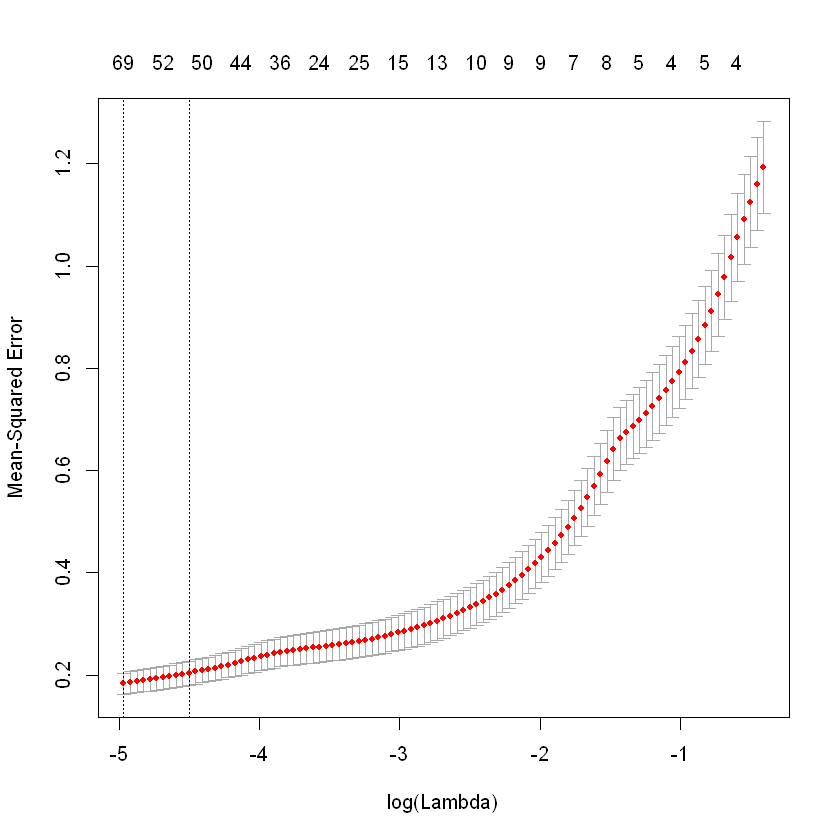
Начнем с модели для зависимой переменной SOC (углерод):

Используем модель Лассо:



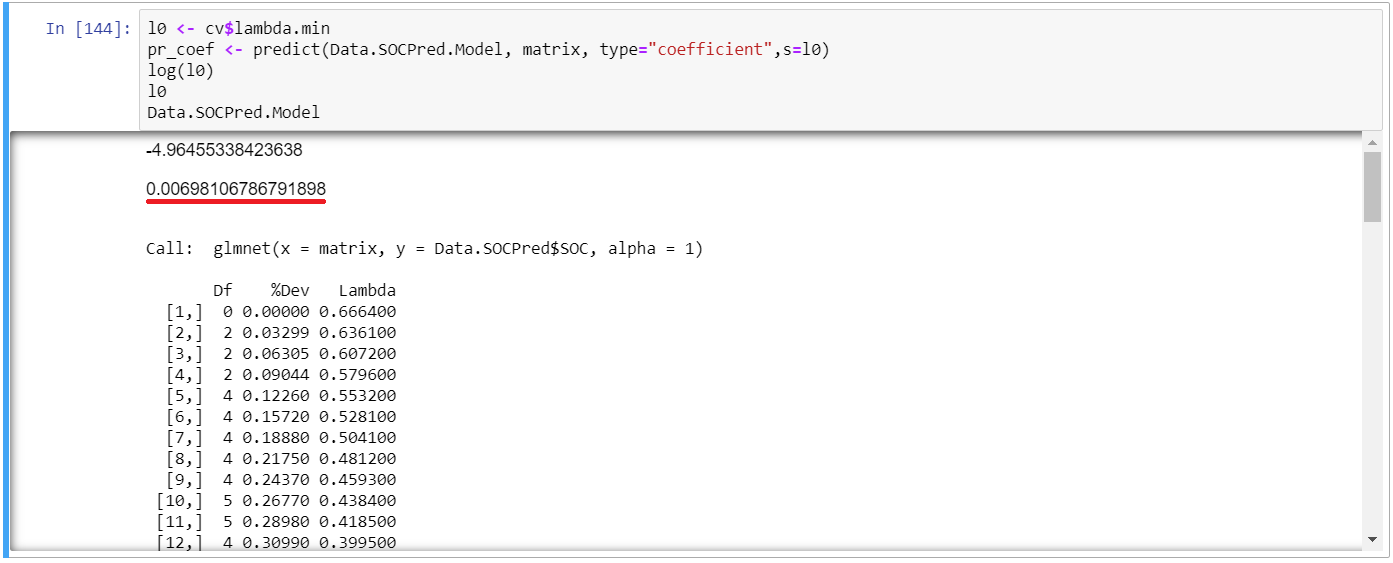
Убираем остальные будущие зависимые переменные из выборки. Построим матрицу предикторов. Через glmnet образуем модель Лассо. При помощи кросвалидации получает вектор значений .

Теперь посмотрим на график этих лямбд:



Отсюда видно, что есть 2 значения , которые могут быть эффективно использованы в нашей модели lambda.min и lambda.1se. Первая лямбда будет иметь минимальную среднюю сумм остатков, но из-за того, что лямбда сама по семе мала, коэффициенты предикторов будут большими. Вторая лямбда берется такая, что MSE будет больше, но коэффициенты будут меньше.

Я предпочту воспользоваться первой лямбдой, так как, используя вторую лямбду мы не получим уменьшения предикторов по отношению к первой , но значение при первой лямбде будет больше.

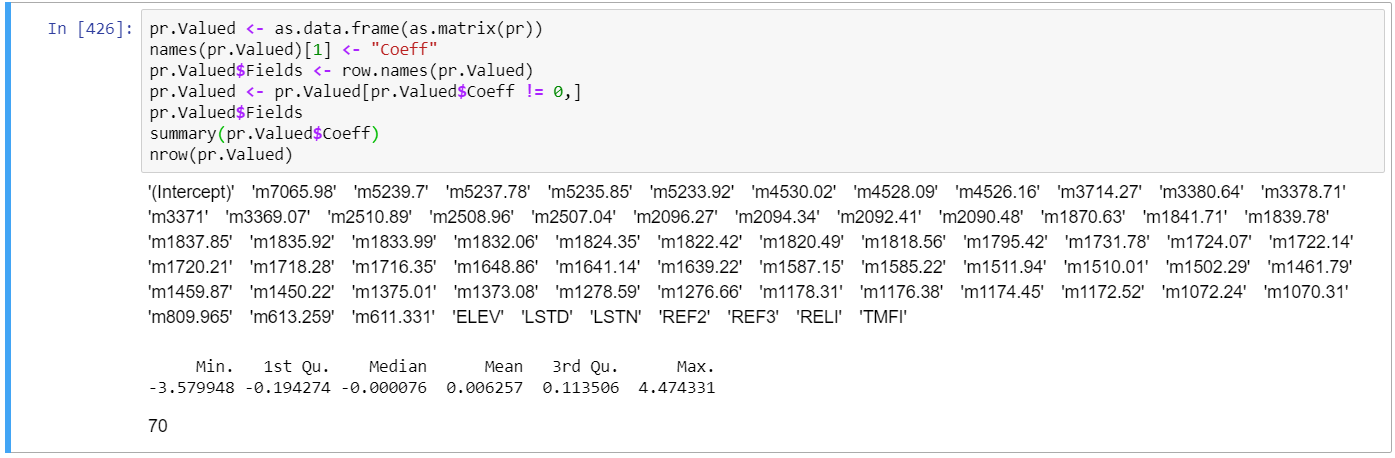


Красным выделено само значение лямбды.



Отсюда видно какова доля дисперсии зависимой переменной, объясняемая полученной моделью (). Она составляет ~ 86,2%, что весьма-таки не плохо.

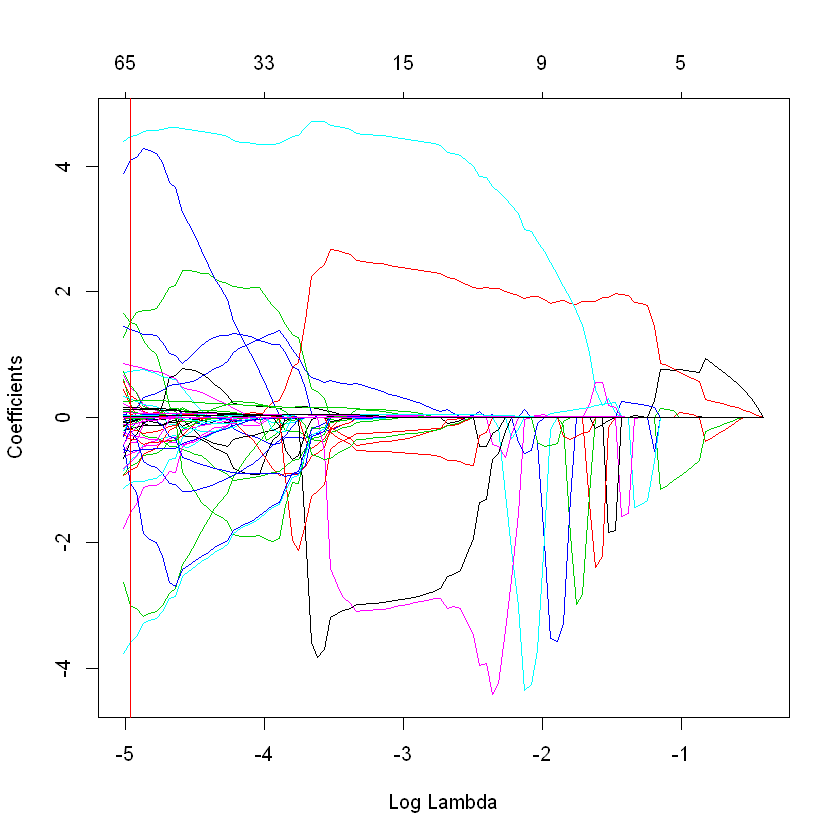
Теперь отбросим предикторы с нулевыми коэффициентами и посмотрим какие предикторы остались.



Получили модель с 70 ю зависимыми переменными.

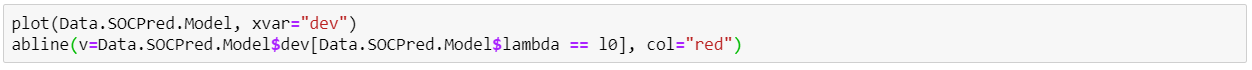
Проанализируем несколько графиков полученных по данной модели:

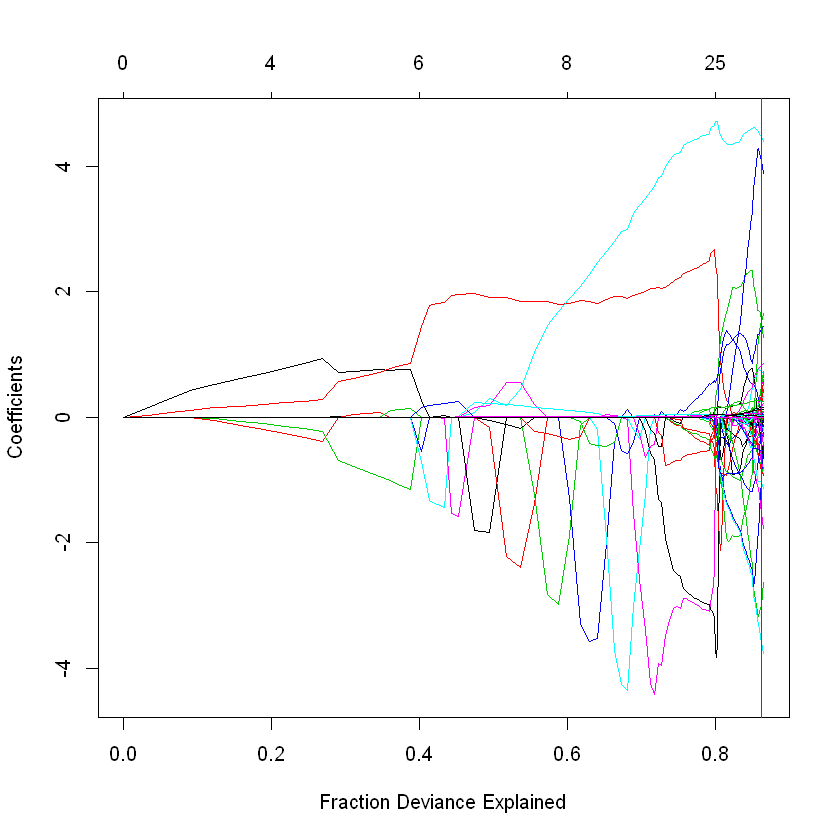




Здесь представлена зависимость значения лямбда и значений коэффициентов. Красная вертикальная линия – это выбранное нами значение .

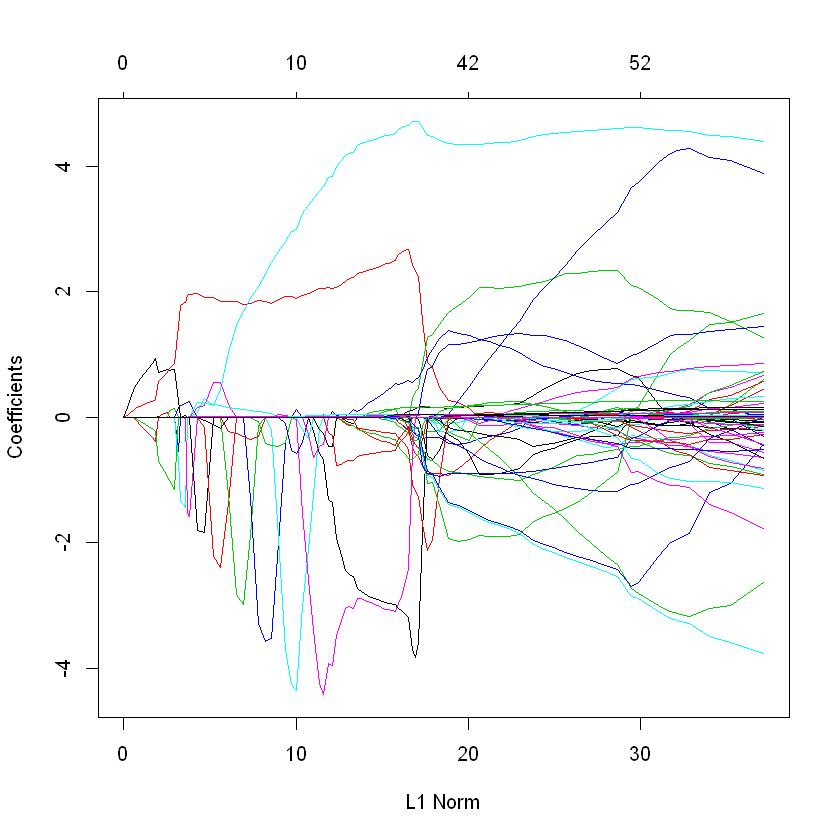
Далее рассмотрим зависимость между и значениями коэффициентов предикторов. Красная линия – наше полученное значение .





И последний график – зависимость между суммой абсолютных величин коэффициентов и значениями этих коэффициентов.





Теперь при помощи кросвалидации можно сравнить модель Лассо с Ридж моделью:



Оценка двух моделей осуществляется по трем параметрам:

- доля дисперсии зависимой переменной, объясняемая полученной моделью.

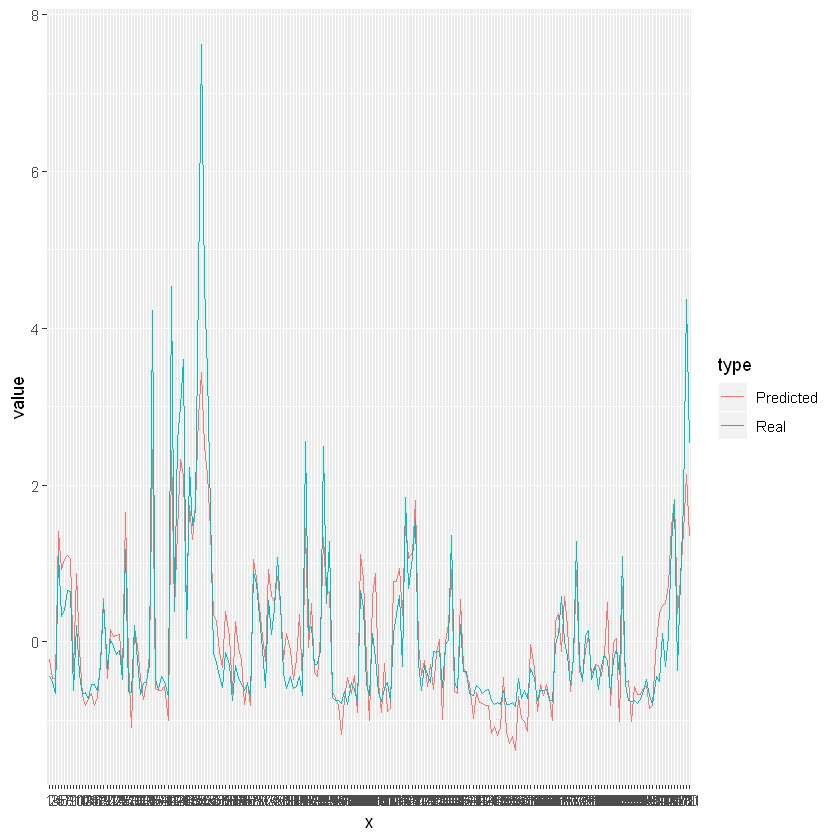
Чем оно больше, тем лучше модель.

– среднеквадратическая ошибка под корнем.

Чем оно меньше, тем лучше (предсказанные значения не сильно отличаются от настоящих).

– среднее по абсолютной ошибке.

Следовательно, по всем параметрам у Лассо модели преимущество по сравнению с Ридж.



Все бы хорошо, но хотелось бы избавиться от лишних предикторов в нашей модели, так как уменьшение их числа даст нам возможность определить базовые предикторы и избавиться от мультиколлинеарности. Воспользуемся функцией step для пошагового увеличения и оценки линейной модели (это уже 3я модель). Я использовал параметр BIC и прямое добавление предикторов.

step(lm(SOC ~ 1,data=Data.SOCPred), scope=SOC ~ m7065.98 + m5239.7 + m5237.78 + m5235.85 + m5233.92 + m4530.02 + m4528.09 + m4526.16 + m3714.27 + m3380.64 + m3378.71 + m3371 + m3369.07 + m2510.89 + m2508.96 + m2507.04 + m2096.27 + m2094.34 + m2092.41 + m2090.48 + m1870.63 + m1841.71 + m1839.78 + m1837.85 + m1835.92 + m1833.99 + m1832.06 + m1824.35 + m1822.42 + m1820.49 + m1818.56 + m1795.42 + m1731.78 + m1724.07 + m1722.14 + m1720.21 + m1718.28 + m1716.35 + m1648.86 + m1641.14 + m1639.22 + m1587.15 + m1585.22 + m1511.94 + m1510.01 + m1502.29 + m1461.79 + m1459.87 + m1450.22 + m1375.01 + m1373.08 + m1278.59 + m1276.66 + m1178.31 + m1176.38 + m1174.45 + m1172.52 + m1072.24 + m1070.31 + m809.965 + m613.259 + m611.331 + ELEV + LSTD + LSTN + REF2 + REF3 + RELI + TMFI, direction="forward", k=log(nrow(Data.SOCPred)))

После чего была получена следующая модель:

Call:

lm(formula = SOC ~ m2510.89 + m1832.06 + m1585.22 + m2096.27 +

m1511.94 + m1731.78 + m1639.22 + m1172.52 + m613.259 + LSTD +

m1716.35 + m3714.27 + m1795.42 + REF3 + m1818.56 + ELEV +

m1724.07 + m1070.31 + m1174.45 + m1178.31 + m1375.01 + m1820.49 +

m2507.04 + m1722.14 + m1276.66 + m1072.24 + m1822.42 + m1278.59,

data = Data.SOCPred)

Residuals:

Min 1Q Median 3Q Max

-2.3307 -0.1921 -0.0190 0.1599 3.6826

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.09437 0.23645 4.628 4.15e-06 \*\*\*

m2510.89 -62.07956 14.61168 -4.249 2.35e-05 \*\*\*

m1832.06 -6.40278 9.06059 -0.707 0.479936

m1585.22 3.02267 0.63719 4.744 2.39e-06 \*\*\*

m2096.27 -5.61876 0.76764 -7.320 5.00e-13 \*\*\*

m1511.94 -1.10520 0.37282 -2.964 0.003103 \*\*

m1731.78 117.28653 22.27287 5.266 1.70e-07 \*\*\*

m1639.22 -7.29432 0.50934 -14.321 < 2e-16 \*\*\*

m1172.52 -212.05239 39.49918 -5.369 9.81e-08 \*\*\*

m613.259 -1.43128 0.21126 -6.775 2.09e-11 \*\*\*

LSTD 0.11182 0.02864 3.904 0.000101 \*\*\*

m1716.35 35.05571 18.01344 1.946 0.051916 .

m3714.27 0.98326 0.16373 6.005 2.64e-09 \*\*\*

m1795.42 6.79187 1.04095 6.525 1.07e-10 \*\*\*

REF3 0.29413 0.05073 5.798 8.91e-09 \*\*\*

m1818.56 -637.35556 123.60464 -5.156 3.02e-07 \*\*\*

ELEV 0.06947 0.01211 5.737 1.27e-08 \*\*\*

m1724.07 -539.66312 133.02067 -4.057 5.35e-05 \*\*\*

m1070.31 18.48931 4.83770 3.822 0.000140 \*\*\*

m1174.45 310.11869 58.97258 5.259 1.76e-07 \*\*\*

m1178.31 -96.82277 19.86665 -4.874 1.27e-06 \*\*\*

m1375.01 0.31670 0.34689 0.913 0.361475

m1820.49 1191.59449 272.21885 4.377 1.32e-05 \*\*\*

m2507.04 65.89147 14.68675 4.486 8.06e-06 \*\*\*

m1722.14 402.39588 127.80870 3.148 0.001689 \*\*

m1276.66 -31.69126 11.22208 -2.824 0.004834 \*\*

m1072.24 -17.09196 4.88996 -3.495 0.000494 \*\*\*

m1822.42 -564.53473 157.41222 -3.586 0.000351 \*\*\*

m1278.59 30.87719 11.33786 2.723 0.006571 \*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

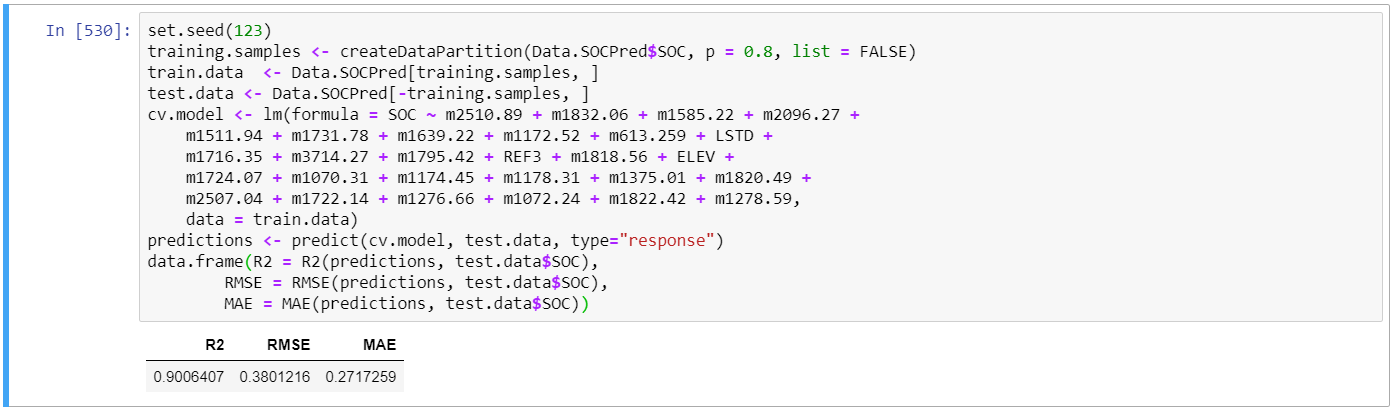
Residual standard error: 0.3635 on 1031 degrees of freedom

Multiple R-squared: 0.8925, Adjusted R-squared: **0.8896**

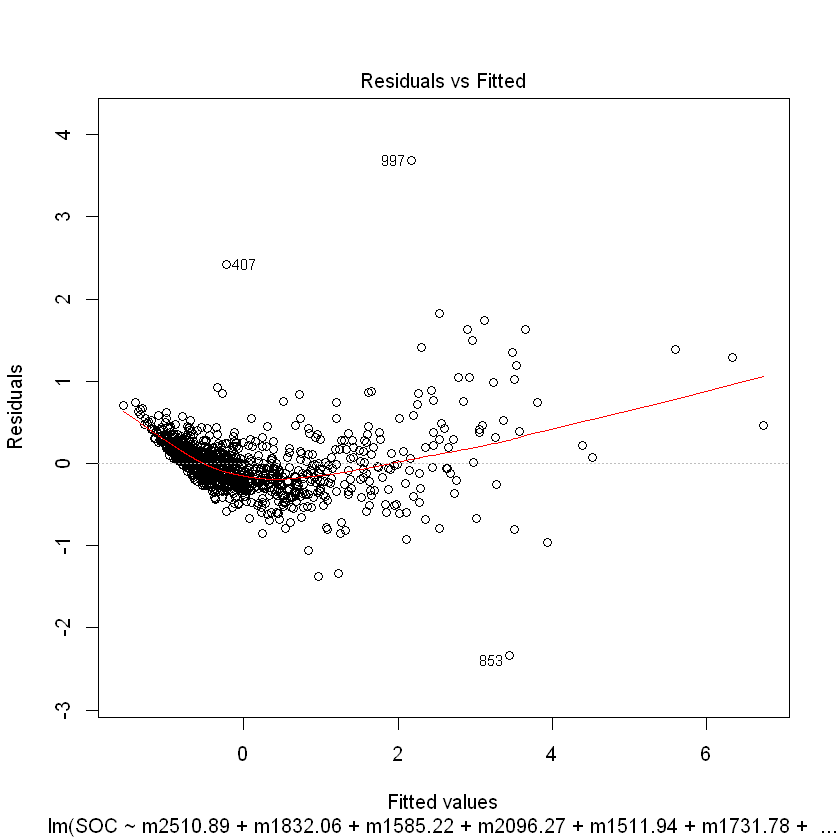
F-statistic: 305.8 on 28 and 1031 DF, p-value: < 2.2e-16

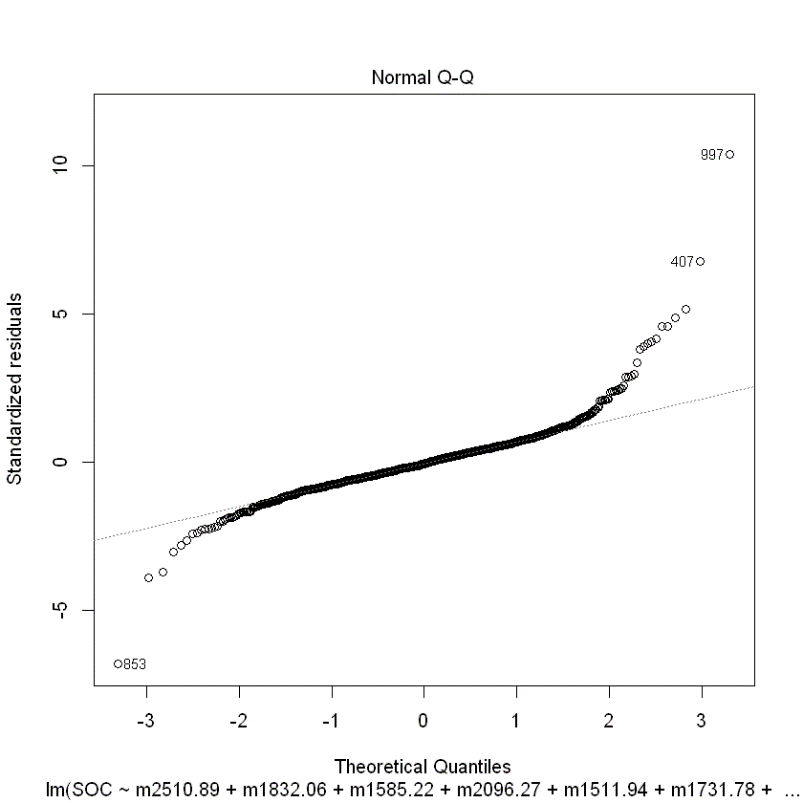
Видно, что удалось избавиться от 42 предикторов из 70, а процент объясняемой дисперсии только увеличился (0.87 –> 0.89).

Проведем кроссвалидацию уже этой модели:

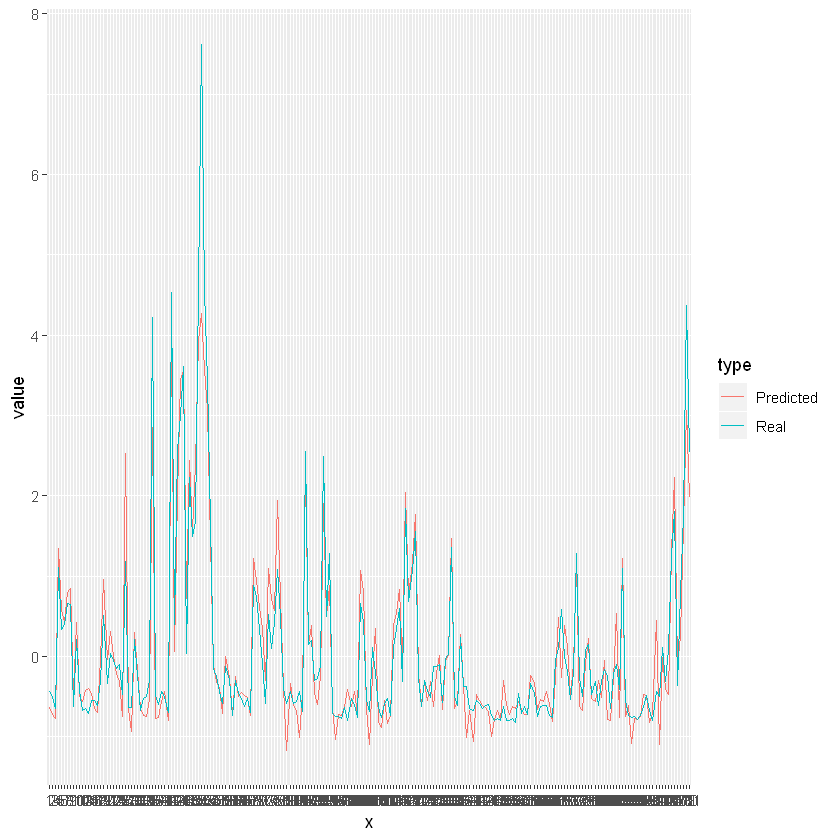


Посмотрим на распределение остатков нашей линейной модели:





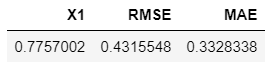
К сожалению, нельзя сказать, что остатки распределены нормально и что у нас выполняется гомоскедостичность. Таким образом это не есть идеальная модель, описывающая зависимость параметра SOC от других измерений, но как мы уже увидели, для нашей выборки удалось более или менее предсказать значение SOC по этой модели.



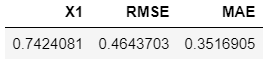
Повторим эти действия построения регрессионных моделей и для других зависимых переменных. Выделю лишь основные моменты рассмотрения остальных моделей:

***Модель для pH:***

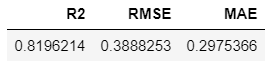
Модель Лассо:



Модель Ридж:

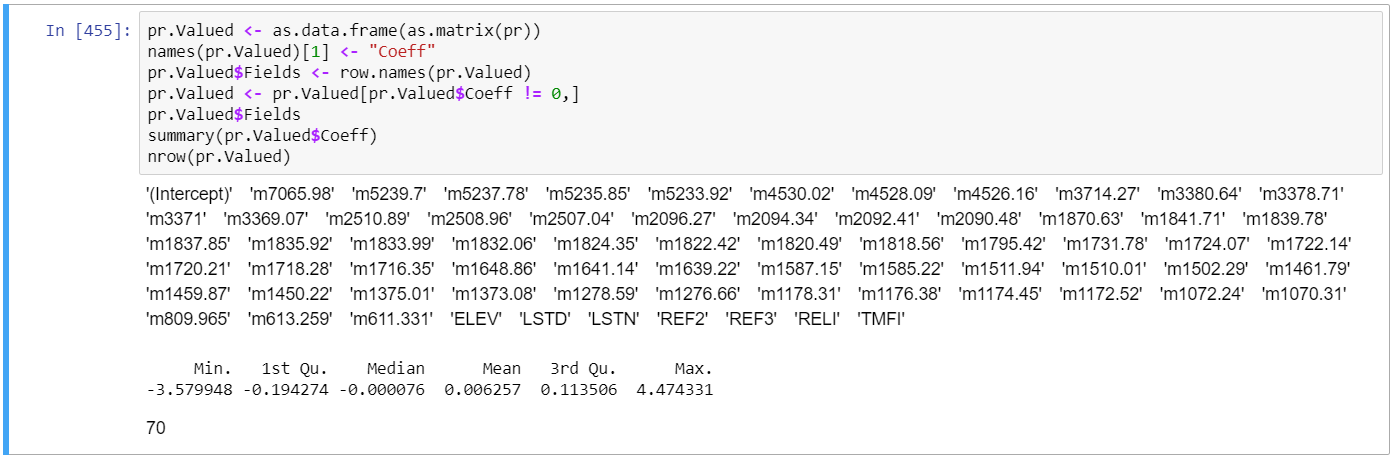


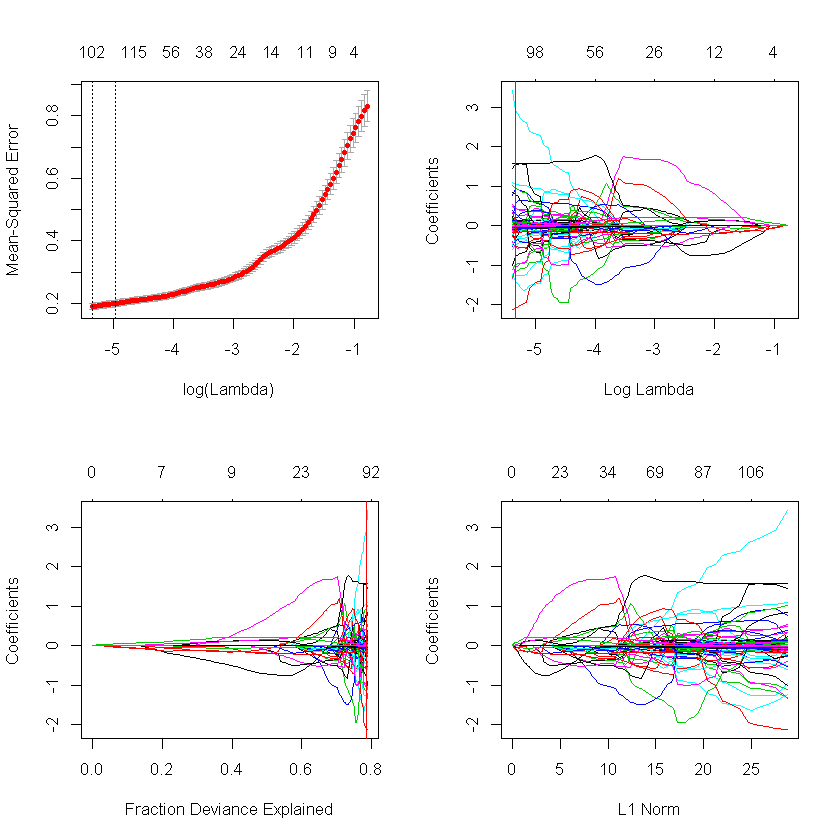
Линейная модель:

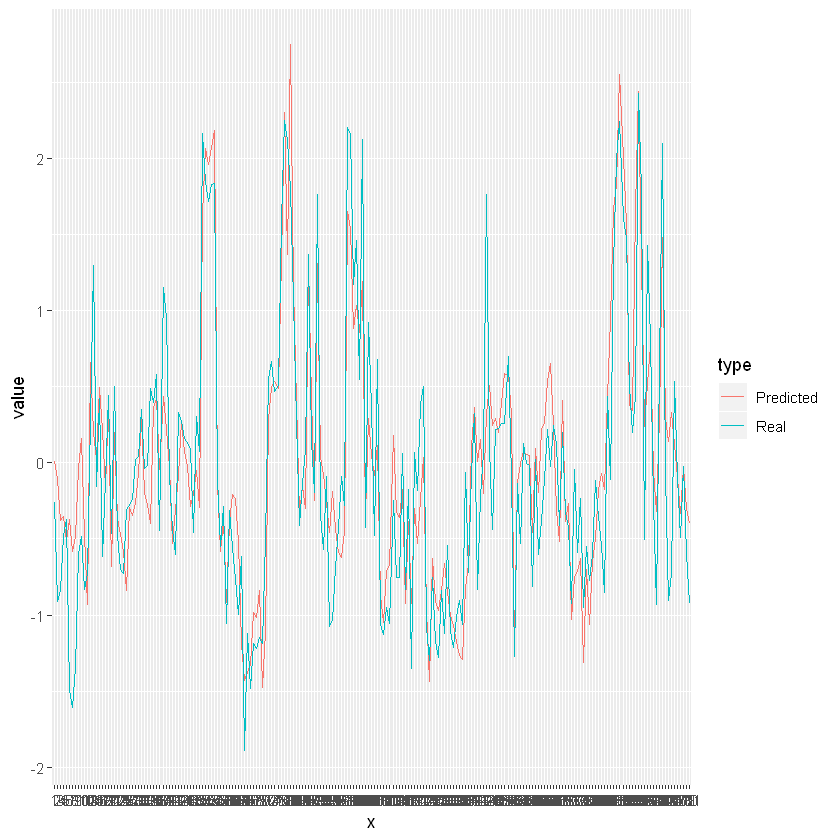


Опять же модель Лассо проявила себя лучше по сравнению с Ридж моделью, но она же в свою очередь проявила себя хуже, чем линейная модель.

Ниже приведены фрагменты кода и графики по аналогии с SOC:





****

step(lm(pH ~ 1,data=Data.pHPred), scope=pH ~ m7065.98 + m5239.7 + m5237.78 + m5235.85 + m5233.92 + m4530.02 + m4528.09 + m4526.16 + m3714.27 + m3380.64 + m3378.71 + m3371 + m3369.07 + m2510.89 + m2508.96 + m2507.04 + m2096.27 + m2094.34 + m2092.41 + m2090.48 + m1870.63 + m1841.71 + m1839.78 + m1837.85 + m1835.92 + m1833.99 + m1832.06 + m1824.35 + m1822.42 + m1820.49 + m1818.56 + m1795.42 + m1731.78 + m1724.07 + m1722.14 + m1720.21 + m1718.28 + m1716.35 + m1648.86 + m1641.14 + m1639.22 + m1587.15 + m1585.22 + m1511.94 + m1510.01 + m1502.29 + m1461.79 + m1459.87 + m1450.22 + m1375.01 + m1373.08 + m1278.59 + m1276.66 + m1178.31 + m1176.38 + m1174.45 + m1172.52 + m1072.24 + m1070.31 + m809.965 + m613.259 + m611.331 + ELEV + LSTD + LSTN + REF2 + REF3 + RELI + TMFI, direction="forward", k=log(nrow(Data.pHPred)))

Call:

lm(formula = pH ~ LSTD + m1373.08 + TMFI + m1459.87 + m3714.27 +

m5233.92 + m2507.04 + m1648.86 + m2508.96 + m2096.27 + m3369.07 +

m1461.79 + m2510.89 + m4530.02 + m1731.78 + m1639.22 + m1587.15 +

m1585.22 + m2092.41 + ELEV + LSTN + m5239.7 + m1795.42 +

m1870.63 + m1450.22 + m7065.98 + m1841.71 + m3380.64 + m1824.35 +

m1178.31 + m5237.78 + m1375.01 + m4526.16, data = Data.pHPred)

Residuals:

Min 1Q Median 3Q Max

-1.30932 -0.24128 -0.01706 0.20936 3.08575

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.21820 0.25649 0.851 0.395129

LSTD 0.11148 0.03278 3.401 0.000696 \*\*\*

m1373.08 -52.94940 15.33218 -3.453 0.000576 \*\*\*

TMFI -0.03263 0.02492 -1.309 0.190821

m1459.87 194.38380 29.56018 6.576 7.69e-11 \*\*\*

m3714.27 -0.80031 0.24520 -3.264 0.001135 \*\*

m5233.92 89.03412 92.53480 0.962 0.336191

m2507.04 -64.44534 100.29574 -0.643 0.520658

m1648.86 27.79150 3.96617 7.007 4.40e-12 \*\*\*

m2508.96 242.90472 208.33887 1.166 0.243921

m2096.27 204.96644 29.14984 7.031 3.73e-12 \*\*\*

m3369.07 11.04035 2.85486 3.867 0.000117 \*\*\*

m1461.79 -162.77036 25.84980 -6.297 4.50e-10 \*\*\*

m2510.89 -190.61710 112.57592 -1.693 0.090715 .

m4530.02 36.63035 11.89170 3.080 0.002123 \*\*

m1731.78 -9.35199 1.11437 -8.392 < 2e-16 \*\*\*

m1639.22 -26.78474 4.17959 -6.408 2.24e-10 \*\*\*

m1587.15 73.58044 19.49302 3.775 0.000169 \*\*\*

m1585.22 -69.22536 19.29625 -3.588 0.000350 \*\*\*

m2092.41 -191.56338 29.34735 -6.527 1.05e-10 \*\*\*

ELEV -0.11889 0.02428 -4.896 1.14e-06 \*\*\*

LSTN -0.17476 0.03683 -4.744 2.39e-06 \*\*\*

m5239.7 605.20141 113.54724 5.330 1.21e-07 \*\*\*

m1795.42 9.92627 1.41255 7.027 3.84e-12 \*\*\*

m1870.63 -3.56343 1.07683 -3.309 0.000968 \*\*\*

m1450.22 -31.03488 5.09479 -6.091 1.58e-09 \*\*\*

m7065.98 3.93581 1.09064 3.609 0.000323 \*\*\*

m1841.71 -24.14361 4.69599 -5.141 3.27e-07 \*\*\*

m3380.64 -10.88169 2.58525 -4.209 2.79e-05 \*\*\*

m1824.35 19.42179 4.52991 4.287 1.98e-05 \*\*\*

m1178.31 -0.47245 0.12784 -3.696 0.000231 \*\*\*

m5237.78 -699.32401 197.19849 -3.546 0.000408 \*\*\*

m1375.01 52.15715 15.67251 3.328 0.000906 \*\*\*

m4526.16 -33.50173 11.56829 -2.896 0.003860 \*\*

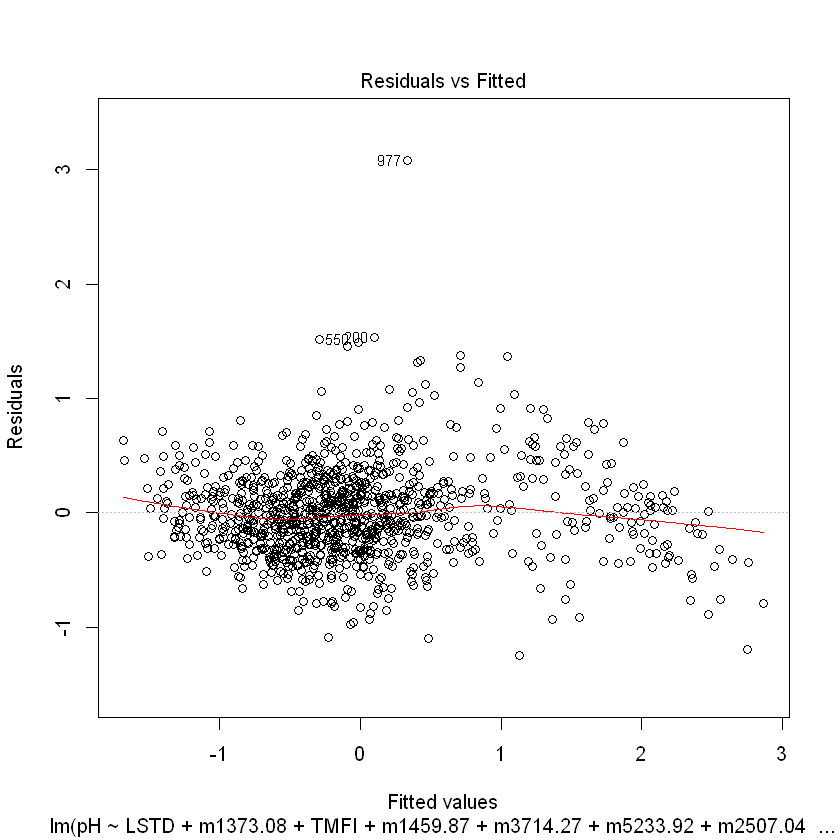
---

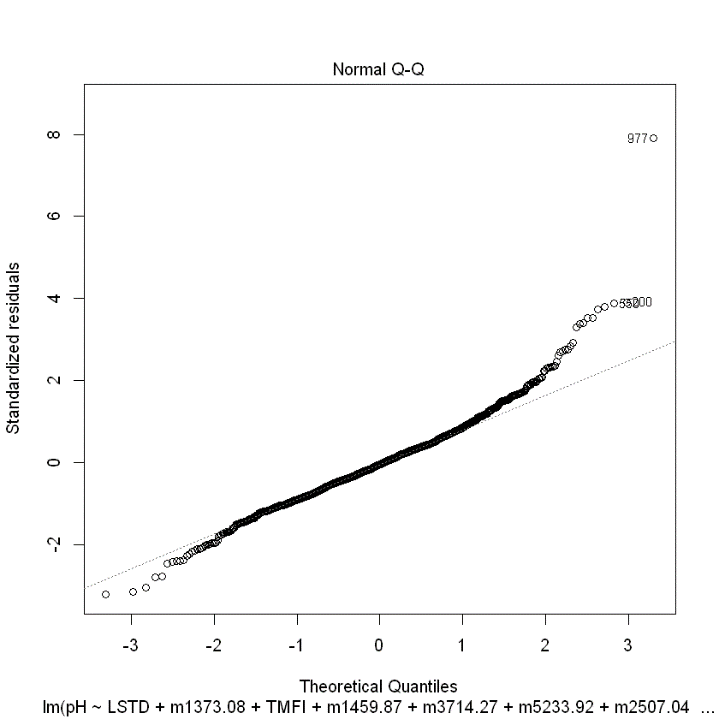
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4004 on 1026 degrees of freedom

Multiple R-squared: 0.8136, Adjusted R-squared: **0.8076**

F-statistic: 135.7 on 33 and 1026 DF, p-value: < 2.2e-16

****

****

А по этим двум графикам распределения остатков можно судить что модель подобрана весьма недурно, так как на первом графике остатки распределены одинаково по обоим сторонам линии и их изменчивость ни так сильно выражена. Да и QQPlot говорит о схожести с нормальным распределением. Таким образом у данной модели лишь один остаток, это количество предикторов, а в целом модель проявила себя не плохо.

***Модель для Ca:***

Модель Лассо:



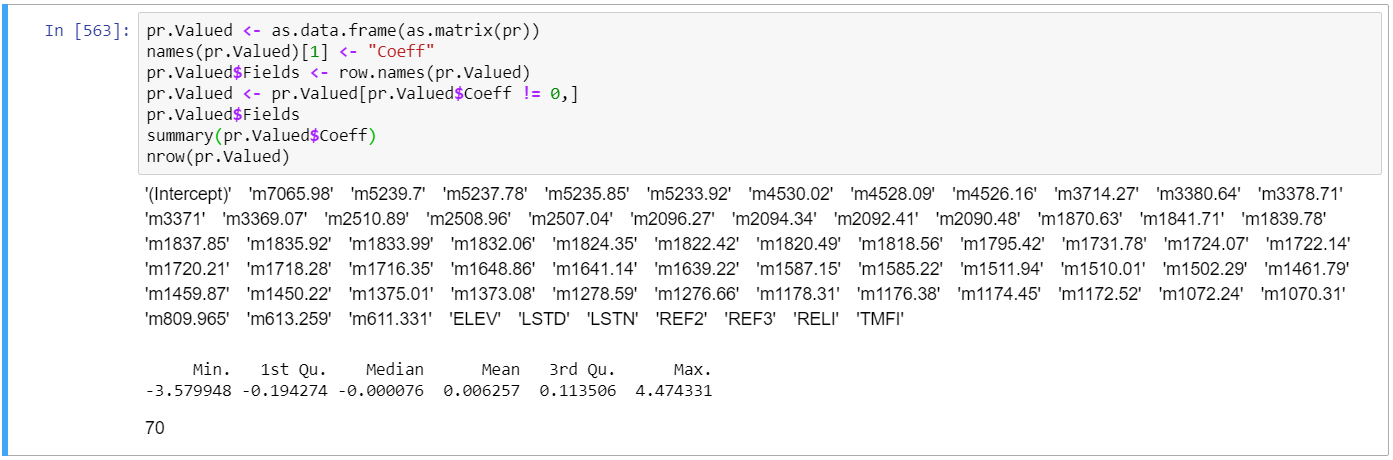
Модель Ридж:

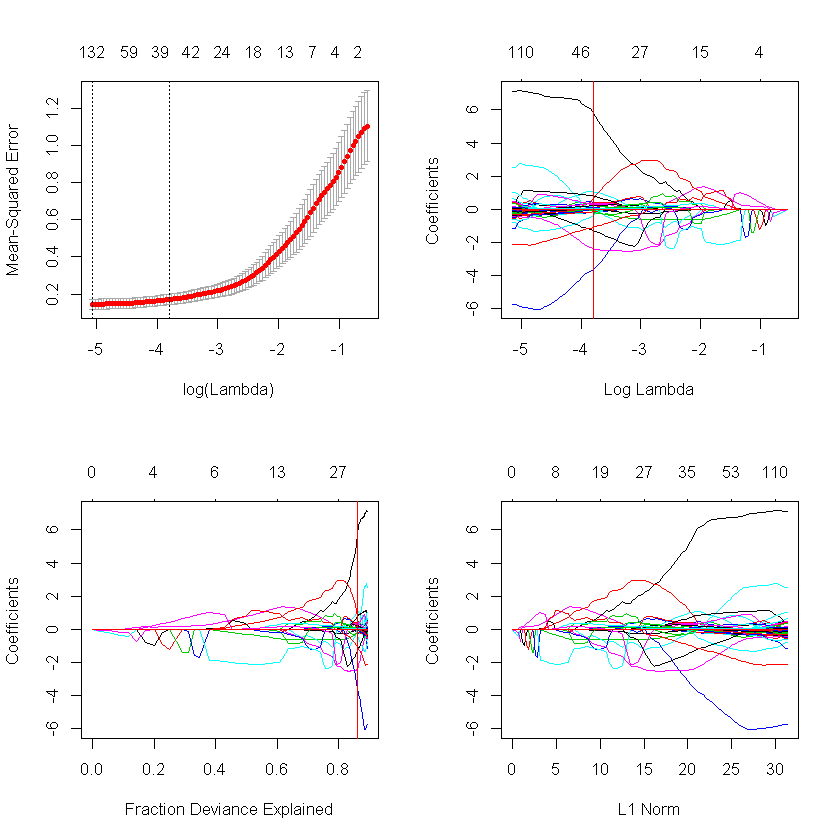


Линейная модель:



Ниже приведены фрагменты кода и графики по аналогии с SOC:

****

****

И здесь уже можно видеть, что разница между предлагаемыми лямбдами на первом графике существенна, следовательно, лучше взять ту, что по-больше для уменьшения коэффициентов предикторов.

Call:

lm(formula = Ca ~ m1373.08 + m1461.79 + m1841.71 + m1795.42 +

m1178.31 + m1070.31 + m1648.86 + m1072.24 + m5239.7 + m1716.35 +

m1833.99 + m1278.59 + ELEV + REF2 + m2507.04 + m1450.22 +

m2508.96 + m2096.27 + m3714.27 + m613.259 + TMFI + LSTN,

data = Data.CaPred)

Residuals:

Min 1Q Median 3Q Max

-2.45965 -0.11435 -0.00838 0.10447 2.98134

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.23889 0.20008 -6.192 8.55e-10 \*\*\*

m1373.08 -0.81274 0.21453 -3.788 0.000160 \*\*\*

m1461.79 -11.06533 2.27505 -4.864 1.33e-06 \*\*\*

m1841.71 -28.25829 4.57625 -6.175 9.48e-10 \*\*\*

m1795.42 9.19492 0.71031 12.945 < 2e-16 \*\*\*

m1178.31 1.23288 0.20596 5.986 2.96e-09 \*\*\*

m1070.31 -9.42085 4.17044 -2.259 0.024093 \*

m1648.86 1.00964 0.33554 3.009 0.002684 \*\*

m1072.24 7.78372 4.21232 1.848 0.064910 .

m5239.7 1.99620 0.22050 9.053 < 2e-16 \*\*\*

m1716.35 -2.19738 0.54451 -4.036 5.85e-05 \*\*\*

m1833.99 18.06184 4.39154 4.113 4.22e-05 \*\*\*

m1278.59 -1.77273 0.23754 -7.463 1.79e-13 \*\*\*

ELEV 0.03255 0.01959 1.661 0.096917 .

REF2 0.22390 0.04596 4.872 1.28e-06 \*\*\*

m2507.04 -148.32666 29.74967 -4.986 7.23e-07 \*\*\*

m1450.22 13.79977 2.35427 5.862 6.16e-09 \*\*\*

m2508.96 144.04375 29.58768 4.868 1.30e-06 \*\*\*

m2096.27 4.13729 0.60960 6.787 1.92e-11 \*\*\*

m3714.27 0.50539 0.13892 3.638 0.000288 \*\*\*

m613.259 0.76909 0.18916 4.066 5.15e-05 \*\*\*

TMFI 0.07236 0.02049 3.532 0.000430 \*\*\*

LSTN -0.08074 0.02923 -2.762 0.005839 \*\*

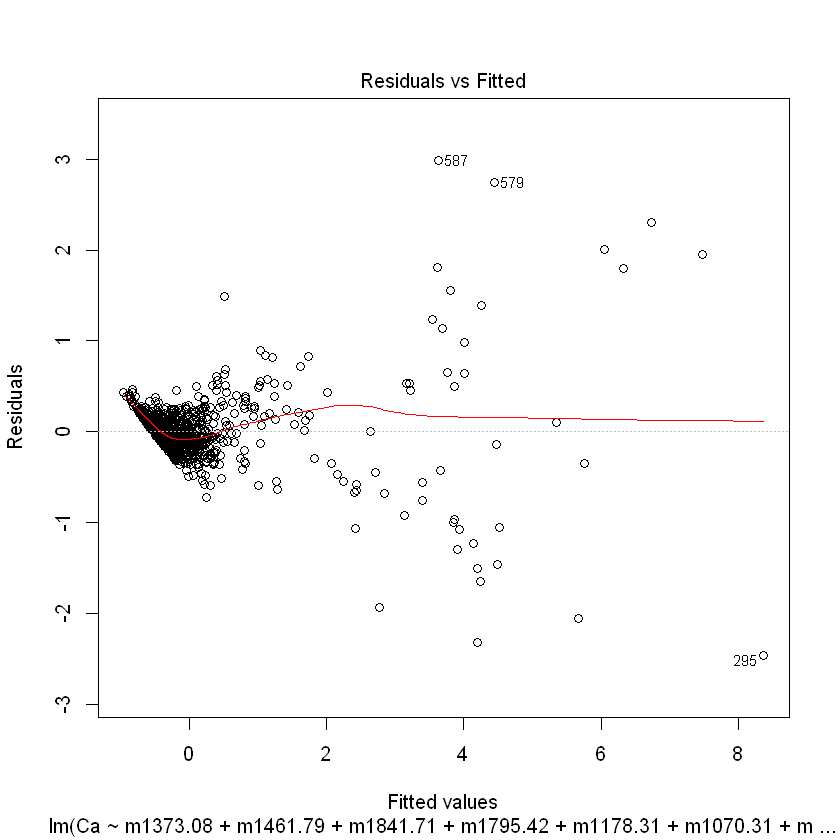
---

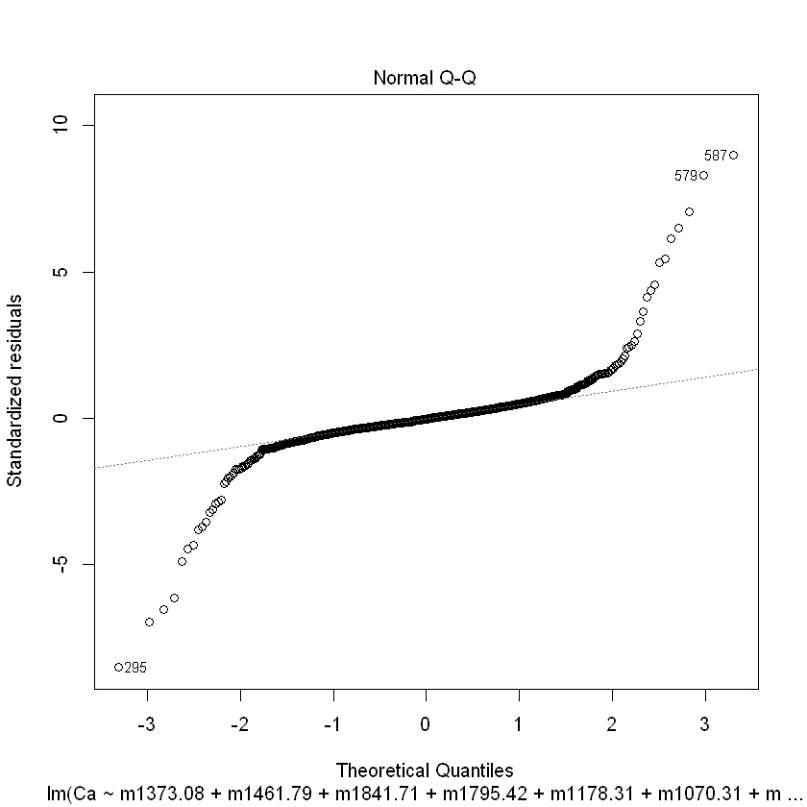
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3443 on 1037 degrees of freedom

Multiple R-squared: 0.8954, Adjusted R-squared: **0.8932**

F-statistic: 403.5 on 22 and 1037 DF, p-value: < 2.2e-16





Опять же по итогу оказалось, что линейная модель наилучшим образом описывает зависимую переменную Ca. Но по графикам распределения остатков мы видим, что не так все хорошо: остатки распределены не нормально и их разброс весьма таки изменчив.