Регрессионный анализ

Evgeny Karpushkin

15.09.2021

Введение

Возможность прогнозировать объем речного стока существенно повышает эффективность работы хозяйственных и природоохранных объектов. Одной из возможностей предсказать речной сток является анализ информации об осадках в зимний период. Интуитивно очевидно, что такая связь должна быть, но не менее важно (помимо подтверждения ее наличия) оценить силу этой связи и ее прогностические возможности.
Нулевая гипотеза, рассматриваемая в этом анализе, сформулирована так: объем речного стока не зависит от величины осадков в той же местности в зимний период. В случае отказа от нее, то есть подтверждения наличия связи, предполагается построить и оптимизировать линейную модель этой связи на основе экспериментальных данных, определить ее статистические характеристики и использовать ее для предсказаний.
Учебной задачей проекта является сравнительное изучение качества нескольких линейных моделей, полученных с помощью инструмента `lm`, а также ознакомление с инструментами высокого уровня, позволяющими решать те же задачи.

Материалы и методы

Данные для анализа были загружены из пакета `alr4` 1.0.6 [1] (датасет `water`). В ходе анализа были использованы пакеты `tidyverse` 1.3.1 [2] (набор пакетов для основных операций по преобразованию и визуализации данных, который включает, в частности, пакеты `ggplot2` 3.3.3 [3], `dplyr` 1.0.6 и `tidyr` 1.1.3 [5]), `cowplot` 1.1.1 [6] (организация нескольких графиков `ggplot2` на рисунке), `qqplotr` 0.0.5 [7] (расширение `ggplot2`; для построения квантильных графиков), `DAAG` 1.24 [8] (для проведения кросс-валидации), `GGally` 2.1.1 [9] (расширение `ggplot2`; для попарного сравнения наборов данных) и `knitr` 1.33 [10] (оформление таблицы в отчете).

Анализ состоял из следующих шагов (их реализация прокомментирована в файле скрипта `runoff.R`):

- загрузка данных
- разведочный анализ данных: ознакомление со структурой, проверка на пропуски, визуализация, проверка на нормальность, выявление выбросов и корреляций
- построение полной линейной модели, предварительный анализ ее корректности
- удаление коррелированных переменных, построение сокращенной модели, анализ ее корректности
- проверка предсказательной способности моделей методом кросс-валидации

Результаты и обсуждение

Описание данных

Согласно описанию датасета `water` [11, с. 54], переменная `Year` содержит год наблюдения, переменная `BSAAM` - величину стока (в акро-футах), а остальные шесть переменных (`APMAM`, `APSAB`, `APSLAKE`, `OPBPC`, `OPRC`, `OPSLAKE`) - величины зимних осадков в различных точках местности. О географической близости точек наблюдения осадков и стока информации нет.

```
str(water, vec.len=3)
```

```
## 'data.frame':    43 obs. of  8 variables:
##  $ Year   : int  1948 1949 1950 1951 1952 1953 1954 1955 ...
##  $ APMAM  : num  9.13 5.28 4.2 4.6 7.15 9.7 5.02 6.7 ...
##  $ APSAB  : num  3.58 4.82 3.77 4.46 4.99 5.65 1.45 7.44 ...
##  $ APSLAKE: num  3.91 5.2 3.67 3.93 4.88 4.91 1.77 6.51 ...
##  $ OPBPC  : num  4.1 7.55 9.52 11.14 ...
##  $ OPRC   : num  7.43 11.11 12.2 15.15 ...
##  $ OPSLAKE: num  6.47 10.26 11.35 11.13 ...
##  $ BSAAM  : int  54235 67567 66161 68094 107080 67594 65356 67909 ...
```

```
head(water)
```

```
##   Year APMAM APSAB APSLAKE OPBPC  OPRC OPSLAKE  BSAAM
## 1 1948  9.13  3.58    3.91  4.10  7.43    6.47  54235
## 2 1949  5.28  4.82    5.20  7.55 11.11   10.26  67567
## 3 1950  4.20  3.77    3.67  9.52 12.20   11.35  66161
## 4 1951  4.60  4.46    3.93 11.14 15.15   11.13  68094
## 5 1952  7.15  4.99    4.88 16.34 20.05   22.81 107080
## 6 1953  9.70  5.65    4.91  8.88  8.15    7.41  67594
```

Загруженные данные имеют ожидаемую структуру: это таблица, содержащая 8 числовых переменных (столбцы) для 43 наблюдений (строки).

```
colSums(is.na(summary(water)))
```

```
##       Year      APMAM      APSAB    APSLAKE      OPBPC       OPRC    OPSLAKE
##          0          0          0          0          0          0          0
##      BSAAM
##          0
```

```
(max(water$Year) - min(water$Year)) == nrow(water) - 1
```

```
## [1] TRUE
```

```
is.element(FALSE, seq(min(water$Year), max(water$Year)) %in% water$Year)
```

```
## [1] FALSE
```

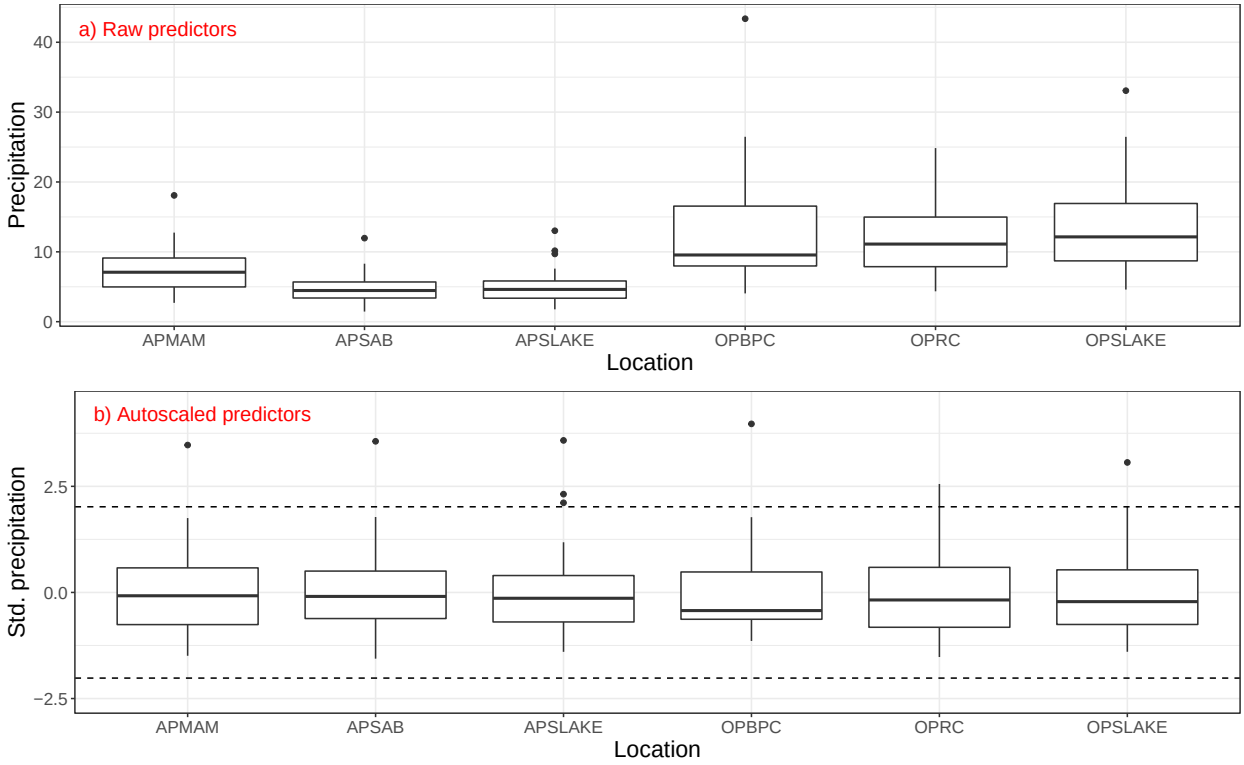
Пропущенных значений в датасете нет. Размах данных в столбце `Year` (разница между максимальным и минимальным значениями) ровно на единицу меньше количества строк в датасете, причем каждое целое число из последовательности между первым и последним годами выборки встречается в столбце данных `Year`. Можно заключить поэтому, что все переменные измерены для каждого года из анализируемого диапазона, причем каждому году соответствует строго одно наблюдение.

Разведочный анализ

Цели разведочного анализа данных:

- проанализировать распределение переменных, выявить возможные ошибки (аномальные значения)
- проанализировать корреляции между переменными
- сделать выводы о применимости линейных моделей к анализу данных
  - следует ли отвергнуть или принять нулевую гипотезу о корреляции между откликом и предикторами?
  - выполняются ли условия применимости линейных моделей для имеющихся данных?

Fig. 1 Distribution of values of predictors.



Распределение значений предикторов (рис. 1a) показывает, что для первых трех из них значения в среднем ниже, чем для последних трех. Первые три предиктора находятся в переменных, имена которых начинаются с A, а имена последних трех начинаются с O. Возможно, это говорит о географическом положении точек сбора данных об осадках, и эти переменные нужно проверить на пространственную автокорреляцию.
Визуализированные точками отскоки расположены в верхней части диаграммы, а в нижней их нет, то есть данные содержат несколько экстремально высоких значений, а выбросы вниз отсутствуют. Это подтверждает и анализ стандартизированных значений предикторов (рис. 1b), где видно, что значения, по модулю превышающие соответствующее критическое значение t-статистики (показана горизонтальной штриховой линией) имеются только в положительной части диаграммы. Наличие такого правого хвоста у данных хорошо видно на рис. 2.

Fig. 2. Pairwise comparison of predictors.

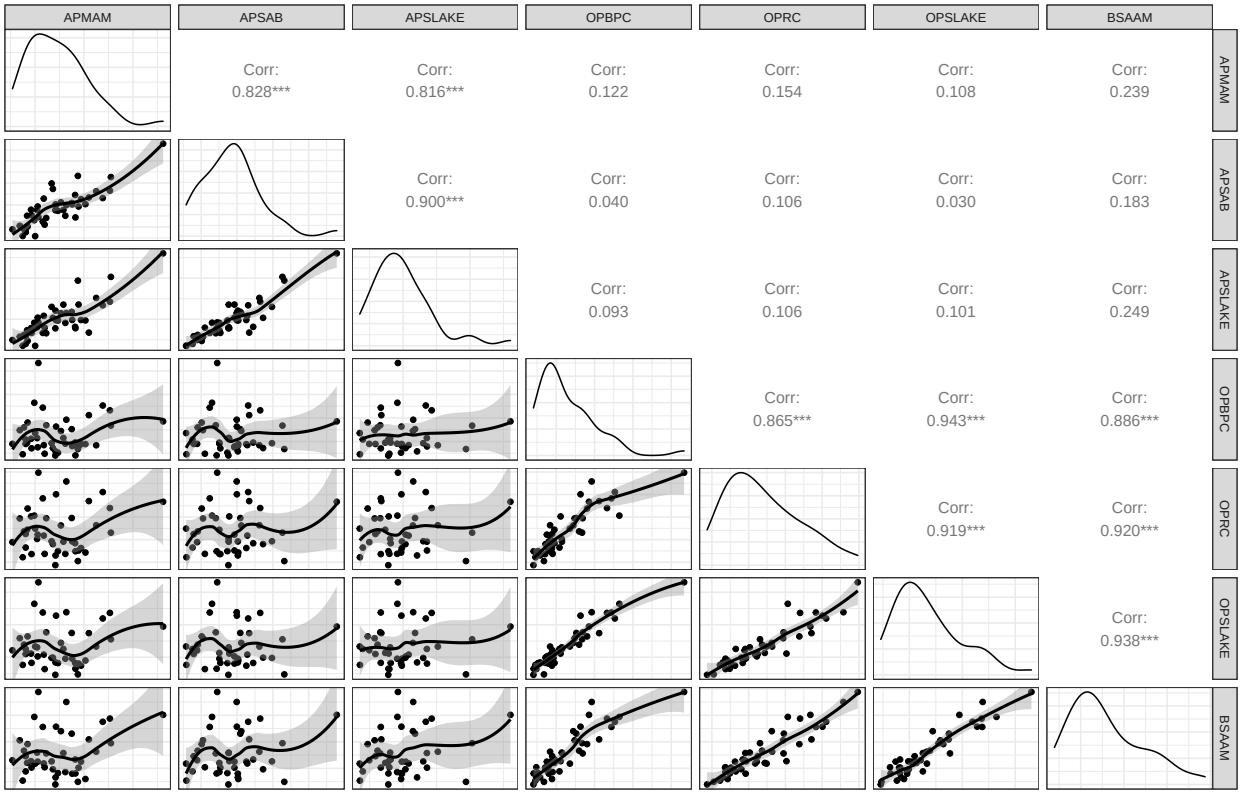


Рис. 2, компактно визуализирует распределения и связи между переменных. Каждая строка и каждый столбец таблицы соответствуют одной из переменных (переменная-отклик помещена в крайний правый столбец и крайнюю нижнюю строку). В правом верхнем углу таблицы приведены значения коэффициентов корреляции между соответствующими парами переменных [12, с. 22], в нижней части таблицы даны диаграммы рассеяния для соответствующих пар переменных и линия тренда, полученная по алгоритму LOESS [12, с. 276], а по диагонали таблицы изображены плотности распределения каждой переменной. Анализ рис. 2 позволяет сделать следующие выводы:

- Для всех переменных (и предикторов, и отклика) на графиках плотности распределения наблюдается длинный правый хвост.
- Переменные с именами A… попарно сильно скоррелированы (коэффициент корреляции > 0.81), аналогично, сильно скоррелированы переменные O… ($r > 0.85$), а корреляция между переменными A… и O… практически отсутствует ($r < 0.16$).
- Переменная отклика сильнее скоррелирована с переменными O… (r 0.88-0.94), чем с переменными A… (r 0.18-0.25).

Так как представленные данные являются ежегодными, маловероятно, что наличие аномально высоких значений связано с ошибками сбора данных. Скорее всего, они представляют собой редкие наблюдения (в данном случае - аномально снежные зимы). Разумно оценить, как вели себя остальные переменные в годы, для которых некоторые переменные принимали аномальные значения.

```
outliers_y <- water_long$Year[water_long$scale_value > t_crit & water_long$key != "B
table(outliers_y)
```

```
## outliers_y
## 1963 1969 1976 1982 1983 1986
##    1    3    1    3    1    1
```

```
data.frame(water_long[water_long$Year %in% outliers_y, ] %>% select(-value) %>% spre
```

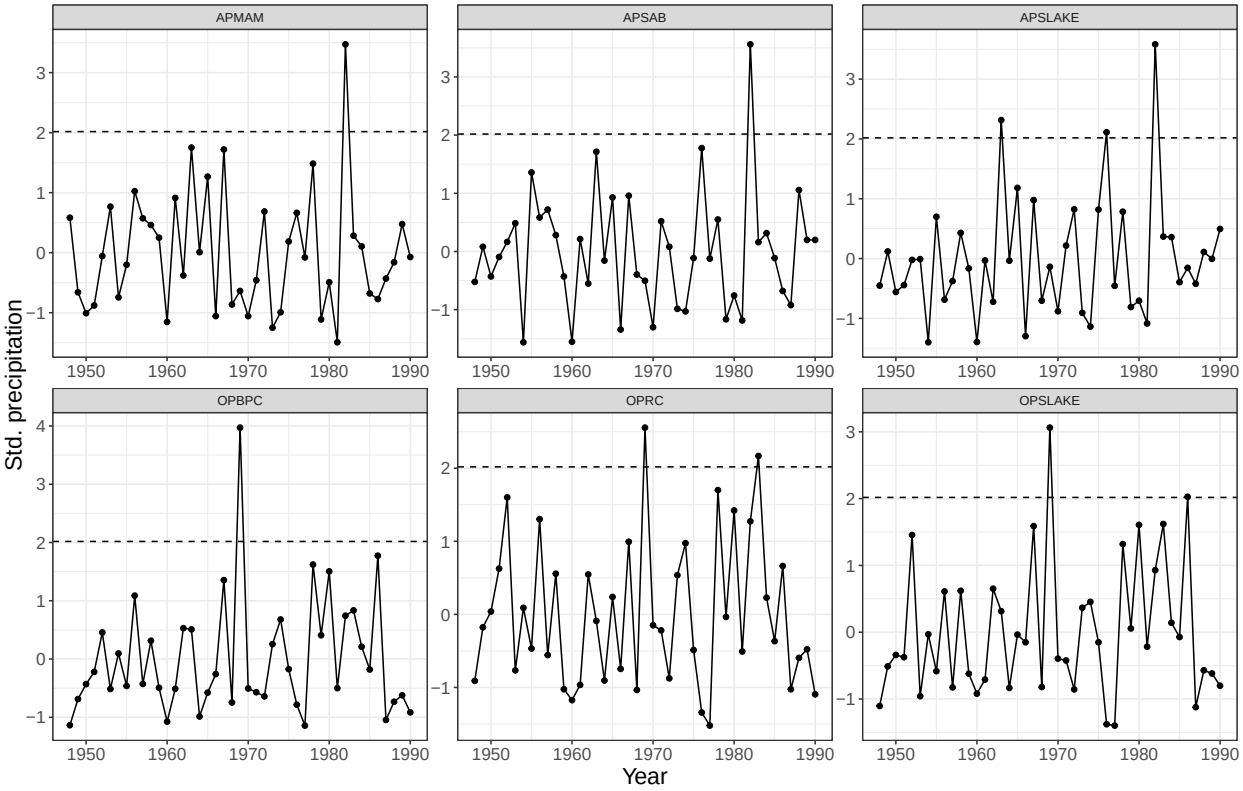
```
##   Year APMAM APSAB APSLAKE OPBPC  OPRC OPSLAKE BSAAM
## 1 1963  1.75  1.71    2.32  0.51 -0.09    0.31  0.42
## 2 1969 -0.64 -0.50   -0.14  3.97  2.56    3.06  2.69
## 3 1976  0.66  1.78    2.11 -0.79 -1.34   -1.38 -1.29
## 4 1982  3.47  3.56    3.58  0.74  1.27    0.93  1.67
## 5 1983  0.28  0.16    0.37  0.83  2.17    1.62  2.24
## 6 1986 -0.77 -0.68   -0.16  1.77  0.66    2.03  1.58
```

Видно, что наиболее "подозрительные" годы - 1969 и 1982 (аномальны значения 3 из 6 переменных), тогда как в остальных случаях отскок только один. Единичные отскоки (в 1963, 1976, 1983 и 1986 гг.) не очень значительны - стандартизированное значение отскакивающей переменной от 2.03 до 2.32 при t_{crit} = 2.02 для α = 0.05, df = 42). С другой стороны, в 1969 и 1983 годах, когда было зафиксировано несколько аномальных значений, они согласованно наблюдались для переменных, между которыми была ранее обнаружена значимая корреляция; таким образом, их следует считать природными аномалиями, а не ошибками сбора данных. Анализ значений отклика (последний столбец) показывает, что именно в 1969 и 1983 годах аномальные величины осадков привели к величине годового стока, выходящей за пределы доверительного интервала.

С точки зрения будущего построения предсказательной модели включение аномальных данных в анализ может привести к некоторой потере точности предсказаний для обычных уровней осадков, но, с другой стороны, позволит провести обоснованную оценку в аномально снежные годы. Исключение аномальных данных по осадкам из модели будет иметь противоположные эффекты: вероятно, точность предсказания "обычных" значений повысится, но предсказание аномальных значений будет в этом случае представлять собой экстраполяцию, что может привести к потере точности. Выгоднее всего было бы построить отдельную модель для аномально снежных зим, но имеющихся данных для этого недостаточно.

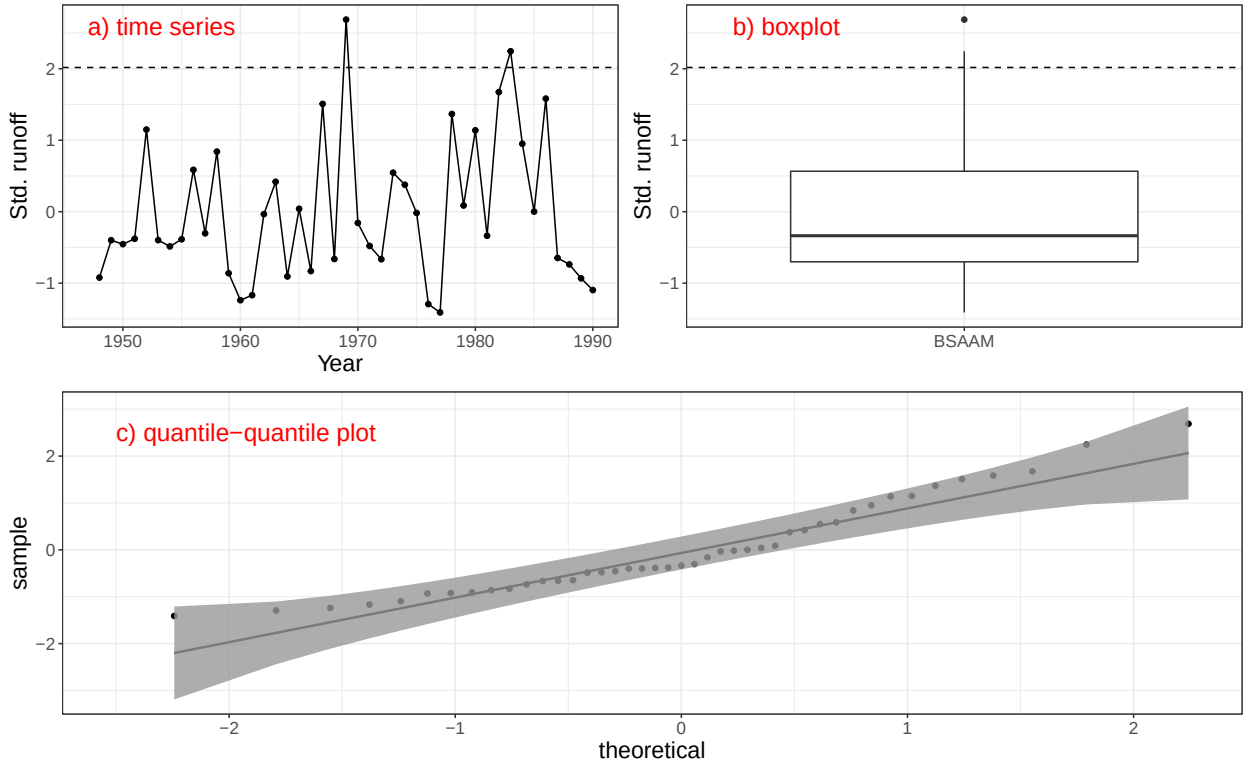
Последний вопрос, который осталось рассмотреть при разведочном анализе предикторов - их возможная временная автокорреляция, которой естественно ожидать для временных серий данных. Однако наши данные представляют собой годовые наблюдения, причем в начале каждого цикла сбора данных состояние системы "обнуляется" (снежный покров полностью тает к концу сезона). Заметная временная автокорреляция, с учетом природы данных, могла бы наблюдаться, если бы рассматривался регион с круглогодичным снежным покровом, и величина стока могла бы определяться осадками не только в последнюю, но и более ранние зимы. Прямых данных об этом для региона Owens Valley, где были собраны данные [11, с. 54], найти не удалось, но по имеющимся данным о среднемесячной температуре в ряде городов Южной Калифорнии [13] можно косвенно заключить, что такая ситуация маловероятна. Это подтверждают и временные серии изменения предикторов (рис. 3). Отметим, что приведенные графики фактически представляют собой точечные диаграммы Кливленда с переменой горизонтальной и вертикальной осей. Исходя из интерпретации этих диаграмм можно дополнительно подтвердить данные анализа рис. 2, что наиболее подозрительными отскоками являются точки для 1969 и 1982 годов.

Fig. 3 Time series for autoscaled predictors



Наконец, рассмотрим распределение отклика (переменная BSAAM в исходных данных) с точки зрения применимости в построении линейной модели. Видно что, заметной временной автокорреляции в отклике не обнаруживается (рис. 4a), в данных имеется лишь пара значений, стандартизированная величина которых немного превышает t_{crit} (2.01) (рис. 4a) и, несмотря на выраженную асимметрию данных (несовпадение медианы и среднего значения, рис. 4b), отклонения от нормальности находятся в пределах доверительного интервала для $\alpha = 0.95$ (рис. 4c).

Fig. 4 Distributions for response variable.



Дополнительно проанализируем значимость выявленных на рис. 2 корреляций переменной отклика с каждым из предикторов (табл. 1). Как видно из представленных в таблице данных, корреляция между тремя из предикторов и переменной отклика значимая ($p < 0.0001$). Таким образом, исходную нулевую гипотезу о независимости стока от осадков следует отвергнуть.

Таблица 1: Table 1. Correlations of response with predictors

	corr	t	p
APMAM	0.239	1.57	0.123
APSAB	0.183	1.19	0.239
APSLAKE	0.249	1.65	0.107
OPBPC	0.886	12.22	0.000
OPRC	0.920	14.99	0.000
OPSLAKE	0.938	17.39	0.000

Обобщая результаты разведочного анализа, заключаем, что данные пригодны для построения линейной модели, так как удовлетворяют основным условиям ее применимости (статистически значимая линейная связь между некоторыми предикторами и откликом; независимость наблюдений; отсутствие коллинеарности некоторых из предикторов).

Построение линейных моделей

Множественные линейные модели можно строить различными способами [14], причем последовательный перебор моделей можно делать как вручную, так и автоматизированно:

- начиная с модели без предикторов, добавлять их последовательно по одному и анализировать, насколько улучшается качество модели

- начиная с полной модели, удалять последовательно наименее значимые или наиболее скоррелированные предикторы, пока при удалении очередного предиктора качество модели не начнет ухудшаться
- провести полный перебор всех возможных моделей и сравнение их параметров

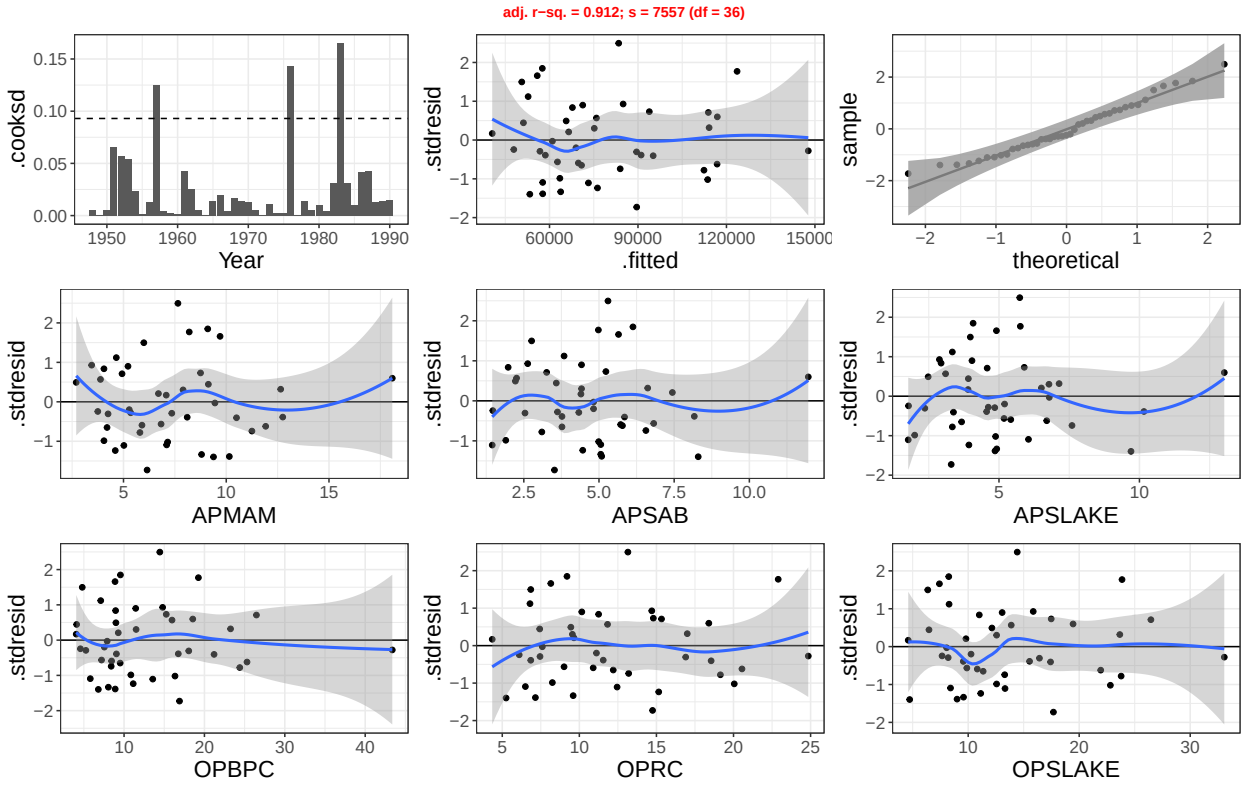
В нашем случае выбор подхода является во многом делом вкуса, так как объем данных не очень велик, а для полного перебора всех комбинаций 6 предикторов необходимо 64 модели, что легко осуществимо на современном компьютере. Тем не менее, для иллюстрации принципов выбора модели будем конструировать их вручную. Для проверки качества построенной модели можно использовать широкий спектр диагностик, например, коэффициент детерминации, статистика значимости модели Фишера, статистика значимости параметров модели Стьюдента, анализ распределения остатков и т. д. [12]. В данном анализе были использованы: скорректированный коэффициент детерминации, среднеквадратичное отклонение модели, диаграмма расстояний Кука, квантильное распределение стандартных остатков модели и графики стандартных остатков от предсказанных значений и от значений каждой переменной. Это позволило выводить результаты тестирования модели в достаточно компактной форме (см., например, рис. 5, где показаны результаты для полной модели, включающей все предикторы).

```
model_full <- lm(BSAAM ~ APMAM + APSAB + APSLAKE + OPBPC + OPRC + OPSLAKE, data = wa
model.diag(model_full, fig_n = 5, printing = TRUE)
```

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

```
## [1] "Table of Cook's distance outliers"
##  Year .cooksd
##  1957   0.125
##  1976   0.144
##  1983   0.165
## [1] "Table of model coefficients"
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15944.7       4100   3.889    0.000
## APMAM          -12.8        709  -0.018    0.986
## APSAB         -664.4       1523  -0.436    0.665
## APSLAKE       2270.7       1341   1.693    0.099
## OPBPC           69.7        462   0.151    0.881
## OPRC          1916.5        641   2.988    0.005
## OPSLAKE       2211.6        753   2.938    0.006
```

Fig. 5 Model: BSAAM ~ APMAM + APSAB + APSLAKE + OPBPC + OPRC + OPSLAKE



Из рис. 5 и диагностических таблиц видно, что полная модель:

- показывает высокий коэффициент детерминации (r^2 0.912) и небольшую стандартную ошибку моделирования (7556 при медиане отклика 69177)
- содержит лишь три наблюдения, которые, исходя из расстояния Кука, избыточно влияют на модель (левый верхний график на рис. 5)
  - эти отскоки - данные для 1957, 1976 и 1983 года; интересно, что наиболее аномальные наборы (1969 и 1982 годы) полную модель не переобучают, а набор переменных для 1957 года вообще не содержит ни одного аномального значения предикторов
- характеризуется близким к нормальному распределением остатков (правый верхний график на рис. 5), а их зависимость от предсказанных значений не выявляет заметной гетероскедастичности или необъясненных паттернов (центральный верхний график на рис. 5)

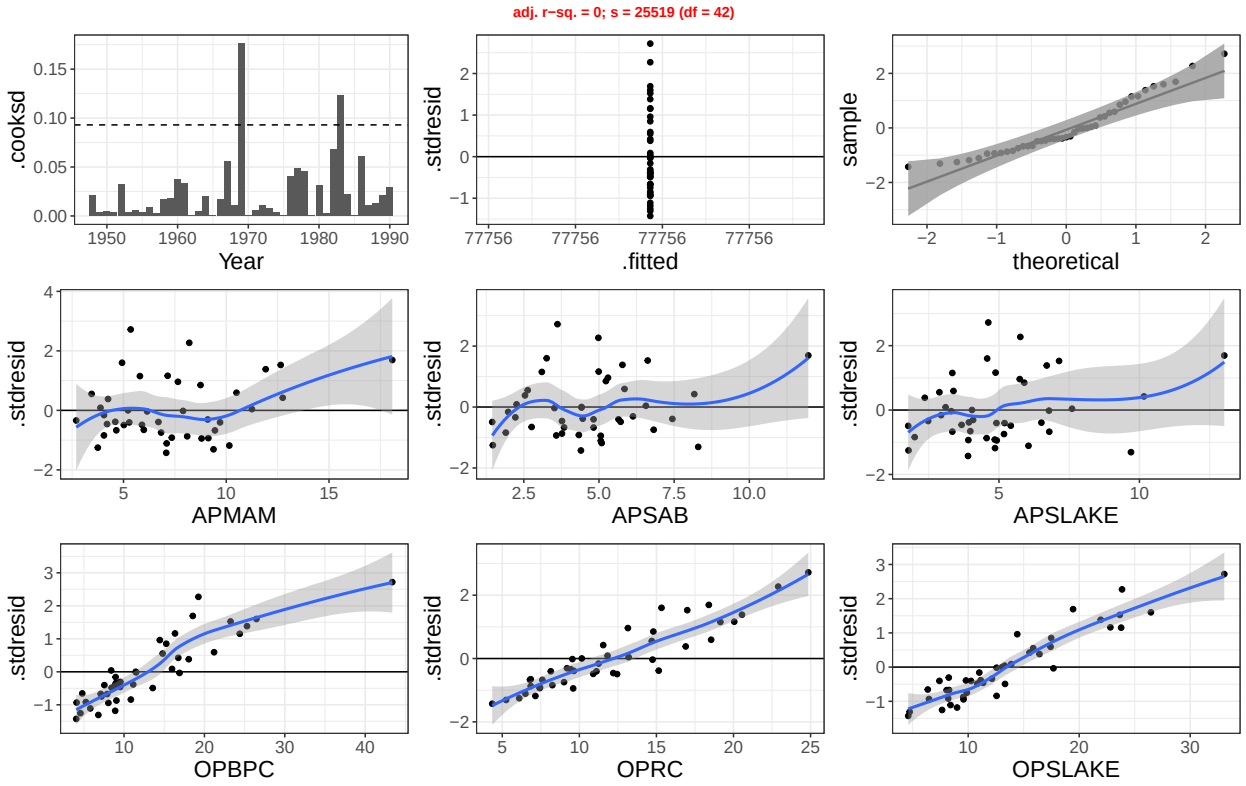
- практически не показывает гетероскедастичности остатков или закономерных паттернов в зависимости от значений предикторов (средний и нижний ряды на рис. 5)
- включает три значимых коэффициента (при OPRC и OPSLAKE, а также свободный член), один коэффициент, значимость которого сомнительна (при APSLAKE) и три незначимых предиктора

В то же время, выявленные ранее мультиколлинеарность предикторов, слабая корреляция отклика с некоторыми из них (рис. 2) и сравнительно небольшое количество наблюдений (примерно 6 на параметр полной модели) указывают, что построенная модель может не быть оптимальной. Проанализируем, какие предикторы модели могут быть избыточными, исходя из фактора инфляции дисперсии [12, с. 216]. Но перед этим в качестве второго крайнего случая построим модель вообще без предикторов (рис. 6).

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

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

Fig. 6 Model: BSAAM ~ 1



Из рис. 6 видно, как проявляется сильная недоопределенность модели для анализируемых данных:

- диаграмма расстояний Кука принципиально не изменяется (максимальные значения соответствуют другим годам, но величины остаются примерно такими же)
- распределение остатков модели лишь немного сильнее отклоняется от нормального, но остается в пределах доверительного интервала (это не удивительно, так как исходная переменная отклика была распределена практически нормально, см. рис. 4, а в модели без предикторов остатки должны быть распределены так же как исходная переменная отклика)
- коэффициент детерминации модели падает до нуля, стандартная ошибка резко возрастает (25519 против 7556 у полной модели)
- на графиках зависимости предсказания от предикторов (исключенных из модели) появляется сильная зависимость, в двух случаях близкая к линейной.

Таким образом, при первичном грубом сравнении моделей следует ориентироваться прежде всего на графики остатков и значения скорректированного коэффициента детерминации и стандартной ошибки. Теперь построим частичную модель, исключая из полной модели предикторы на основе значений их фактора инфляции дисперсии.

```
vif(model_full)
```

```
##    APMAM    APSAB APSLAKE    OPBPC     OPRC OPSLAKE
##     3.55     7.18     6.75     9.27     7.65    16.97
```

```
model_A <- update(model_full, . ~ . -OPSLAKE)
vif(model_A)
```

```
##    APMAM    APSAB APSLAKE    OPBPC     OPRC
##     3.53     6.51     6.00     4.22     4.18
```

```
model_A <- update(model_A, . ~ . -APSAB)
vif(model_A)
```

```
##   APMAM APSLAKE   OPBPC    OPRC
##    3.04    3.00    3.97    4.01
```

```
model_A <- update(model_A, . ~ . -OPRC)
vif(model_A)
```

```
##   APMAM APSLAKE   OPBPC
##    3.01    2.99    1.02
```

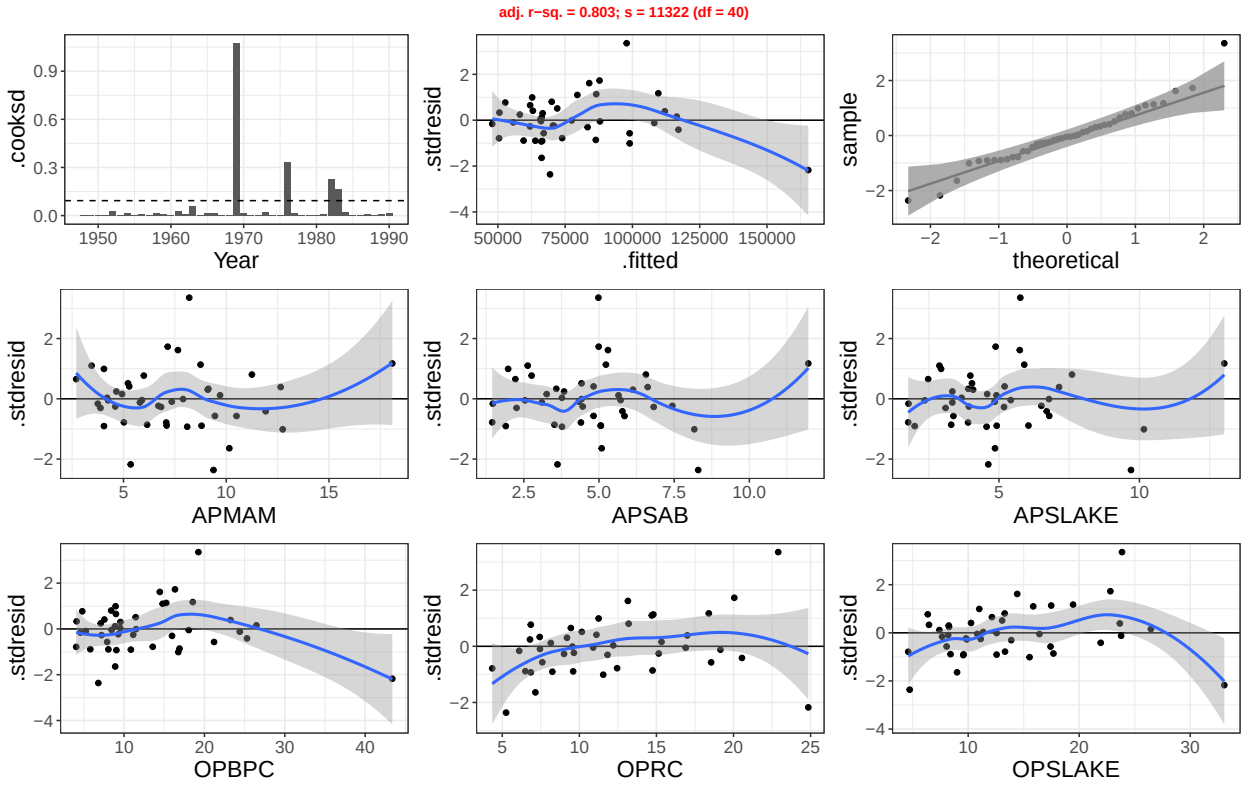
```
model_A <- update(model_A, . ~ . -APMAM)
vif(model_A)
```

```
## APSLAKE   OPBPC
##    1.01    1.01
```

Как видно из консольного вывода, последовательное применение функции vif и удаление предиктора с наибольшим значением фактора инфляции дисперсии приводит к приемлемому уровню VIF (<3) лишь после удаления всех предикторов кроме двух. Заметим, что первой мы были вынуждены удалить из модели переменную OPSLAKE, корреляция которой с откликом была максимальной (рис. 2). Кроме того, начиная с третьего шага алгоритма нам приходилось делать выбор между переменными с очень близкими значениями VIF (4.01 и 3.97; 4.01 и 3.97; 3.01 и 2.99). Это связано именно с тем, что имеются две группы предикторов, сильно скоррелированных внутри группы и практически не скоррелированных между группами.
Диагностика полученной частичной модели A с двумя предикторами (рис. 7) показывает уменьшение коэффициента детерминации и увеличение стандартной ошибки по сравнению с полной моделью (рис. 5). Еще более важным является появление закономерного тренда на графиках зависимости остатка от значения переменных, не включенных в модель (заметнее всего для переменной OPSLAKE, но проявляется и для других переменных). Кроме того, на параметры этой модели наиболее сильное влияние оказывают данные для 1969 года (при этом максимальное расстояние Кука возрастает по сравнению с полной моделью приблизительно в 6 раз), которые ранее были сочтены наиболее аномальными. Таким образом, построенную модель (A) следует признать недообученной.

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

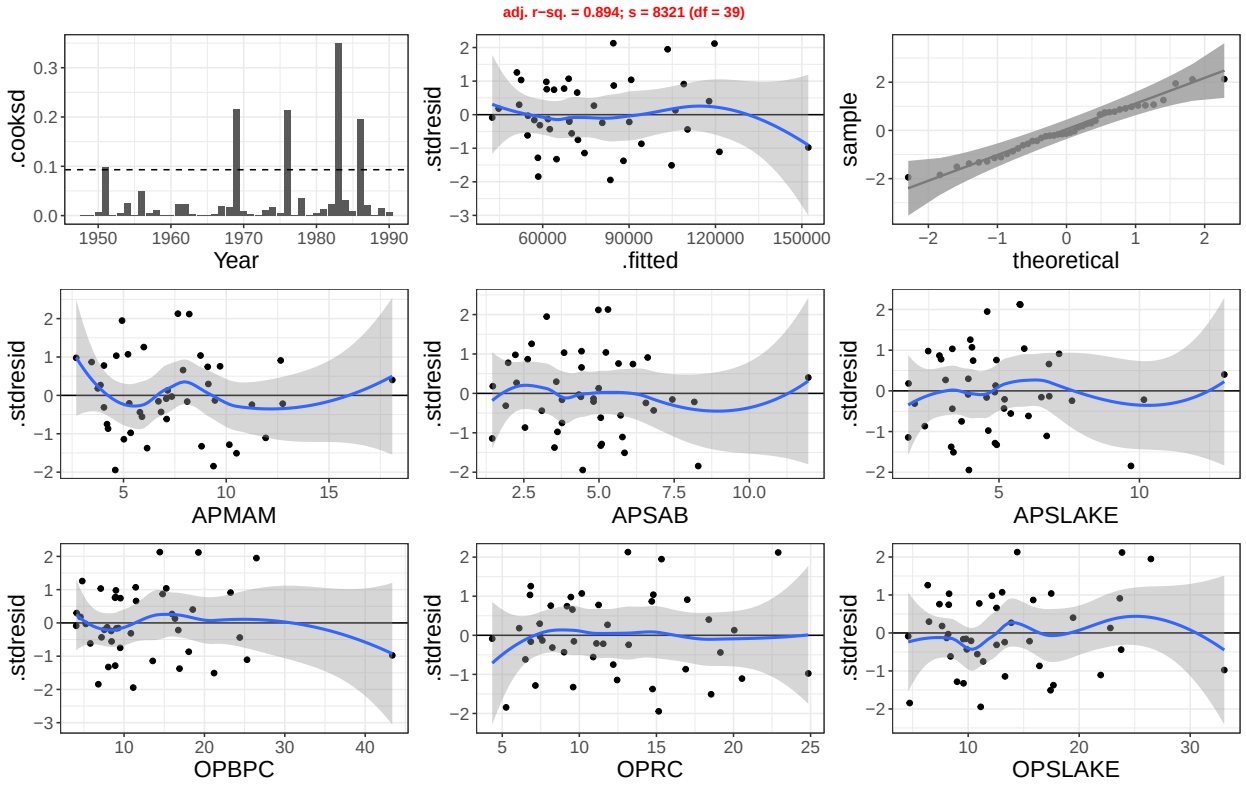
Fig. 7 Model: BSAAM ~ APSLAKE + OPBPC



Разумным шагом является дополнение модели A одним из предикторов, исключенных на основании фактора инфляции дисперсии. Из соображений наличия наиболее выраженного закономерного тренда на графиках остатков (рис. 7) выбор следует делать между предикторами OPRC и OPSLAKE (были построены обе модели, названные, соответственно, B и C). Диагностические графики для этих моделей приведены на рис. 8 и 9, соответственно.

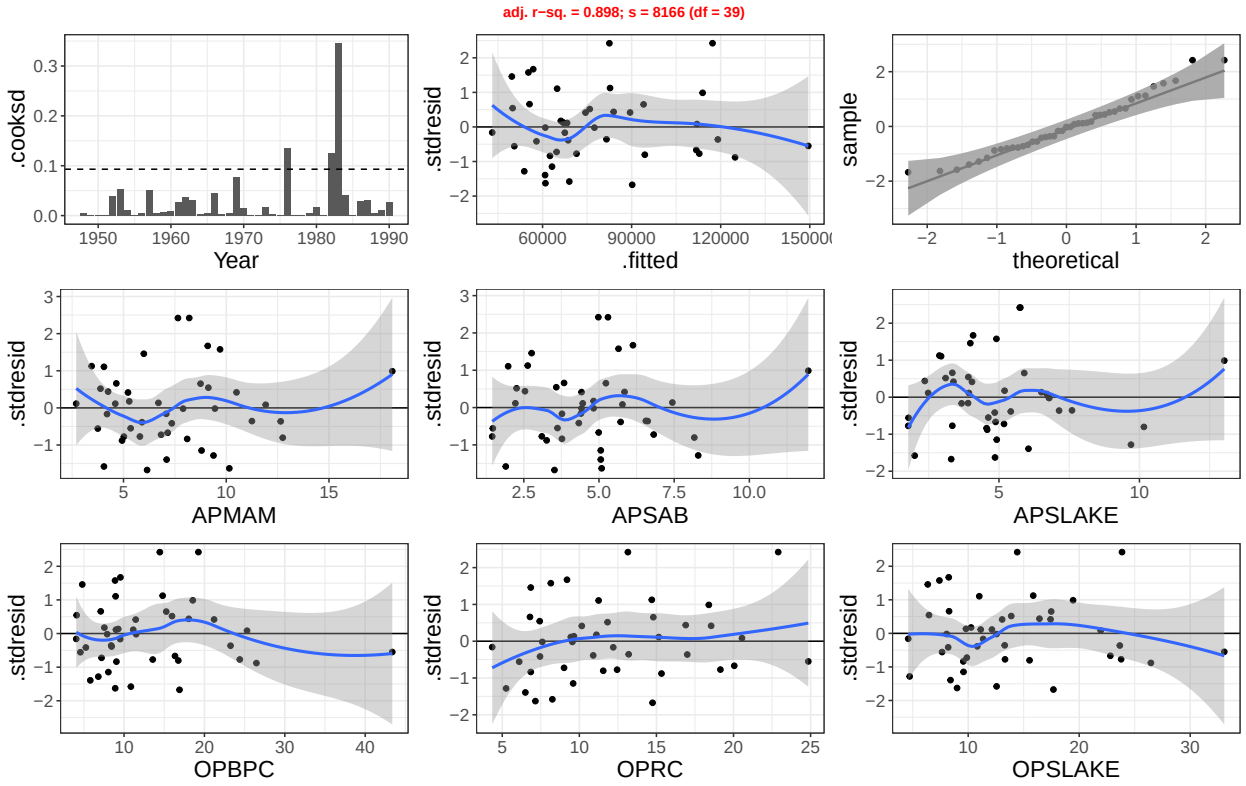
```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

Fig. 8 Model: BSAAM ~ APSLAKE + OPBPC + OPRC




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

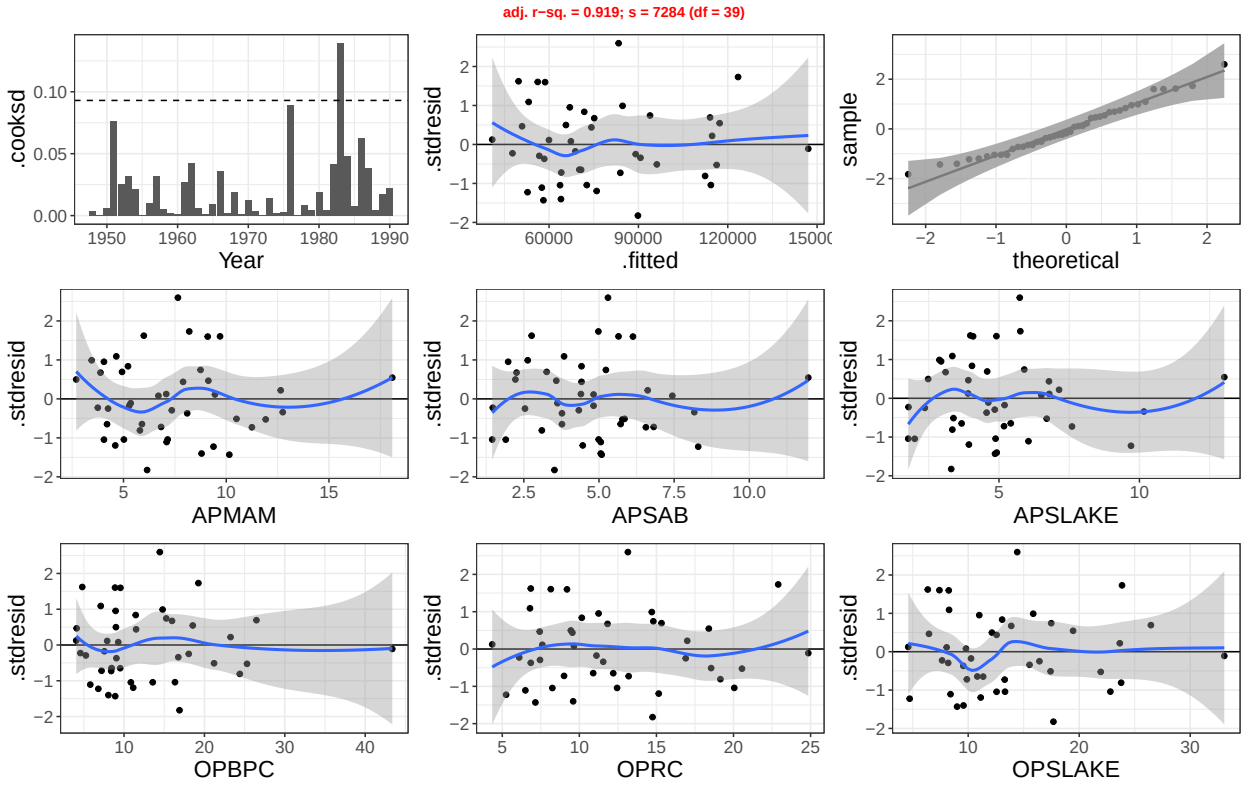
Fig. 9 Model: BSAAM ~ APSLAKE + OPBPC + OPSLAKE



Базовые статистические показатели моделей B и C близки, и графики остатков в зависимости от предикторов, не включенных в модель, практически не отличаются. В целом диагностические графики остатков не показывают выраженной гетероскедастичности или закономерных паттернов, что позволяет считать их кандидатами для проверки предсказательной способности. Так как величины r^2 и s для этих моделей очень близки к таковым для полной модели, дальнейшее усложнение полученных трехпредикторных моделей не представляется разумным.
Для сравнения также была построена оптимальная модель по критерию Акаике [12, с. 217]. Этот алгоритм осуществляет автоматический перебор линейных моделей, минимизируя некоторый информационный критерий. В целом, оптимизация должна приводить к одновременной минимизации стандартного отклонения и количества использованных предикторов. AIC-оптимальная модель также включает три предиктора, но их набор отличается от моделей B и C. Диагностические графики для этой модели приведены на рис. 10. Отметим, что первичные статистические показатели этой модели ($adj.r^2 = 0.919$ и $s = 7284$) оказываются даже лучше, чем у полной модели.

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

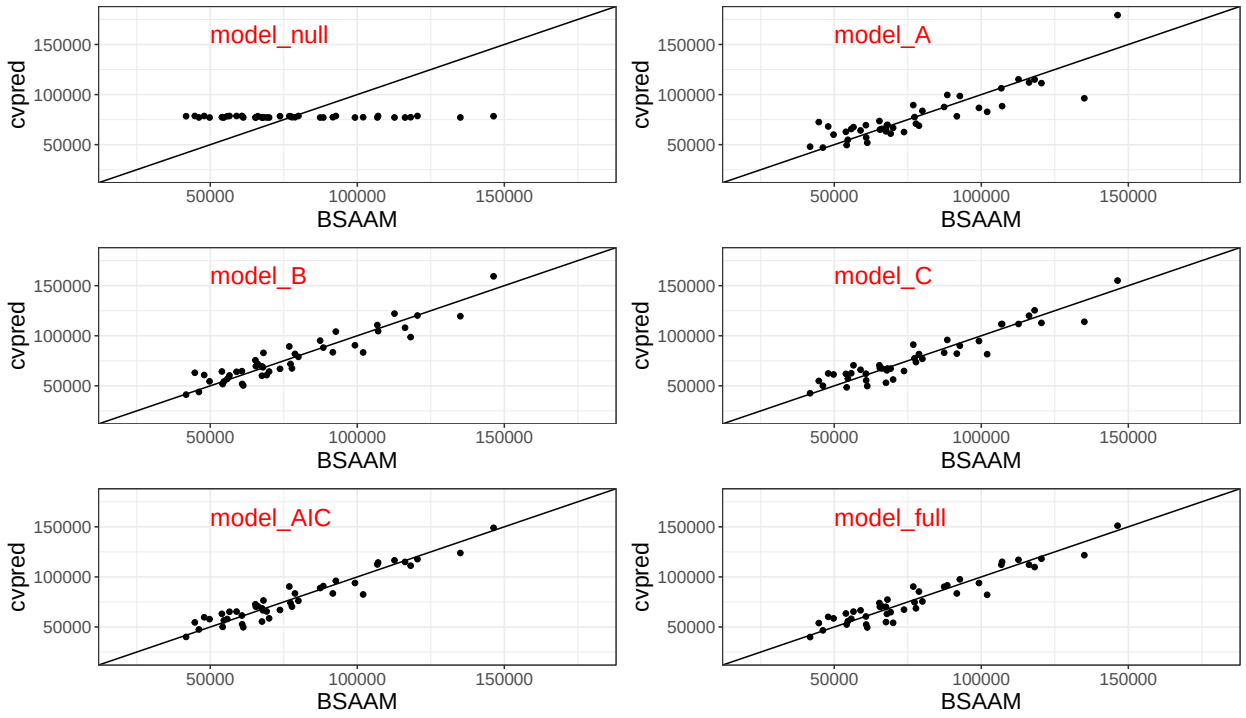
Fig. 10 Model: BSAAM ~ OPSLAKE + APSLAKE + OPRC



Проверка предсказательной способности моделей

Окончательное решение о качестве построенных моделей в свете поставленной задачи (предсказание летнего речного стока по данным о зимних осадках) может дать только независимая проверка. С учетом специфики данных, единственной возможностью провести такую проверку является кросс-валидация [12, с. 220]. При этом из исходного набора данных случайным образом выделяют тестовое подмножество, линейная модель строится на основе оставшихся данных, а затем для тестового набора предсказываются значения с помощью построенной модели и сравниваются с реальными значениями. В данном исследовании была использована пятикратная полная кросс-валидация: из исходного набора в 43 наблюдения в тестовый набор были выделены примерно 1/5 (8 или 9 наблюдений), проведена кросс-валидация, а затем процедура повторялась еще четыре раза с иным выбором тестового набора, причем в пяти проходах каждое из наблюдений ровно один раз играло роль тестового. Результаты кросс-валидации приведены на рис. 11 в форме зависимостей предсказанных значений отклика в тестовом наборе от истинных значений.

Fig. 11 Cross−validation of the models.



Результаты на рис. 11 подтверждают ранее сделанные предположения о том, что модель A, построенная по двум предикторам, является недообученной, так как разброс точек вокруг теоретической прямой визуально больше, чем в случае B, C и AIC (по три предиктора). Особенно хорошо это заметно в области аномально высоких значений отклика (две крайние правые точки). Описание основного облака данных моделями B, C, AIC визуально не хуже, чем для полной модели. Аномальные точки лучше всего предсказывают модели AIC и полная.

Чтобы количественно сопоставить эти модели, рассмотрим сводную таблицу их характеристик (табл. 2). В столбце “Prediction MSe” этой таблицы приведены среднеквадратичные отклонения кросс-валидации моделей. Отметим, что они незначительно выше, чем приведенные там же значения среднеквадратичной ошибки моделирования (Model MSe), так как при построении моделей были использованы лишь 4/5 имеющихся данных (исключен тестовый набор).

При переходе от модели A к выбранным вручную моделям B и C ошибка предсказания за счет учета одного дополнительного предиктора уменьшается примерно на 30%, а дальнейшее добавление еще трех предикторов (до полной модели) улучшает точность предсказания лишь на 6-11%. Важно отметить, что оптимизированная автоматически модель AIC показывает характеристики предсказания и моделирования лучше, чем полная модель, то есть полная модель является переобученной. Это отражается и в параметрах моделирования - коэффициент детерминации для полной модели лишь немного выше, чем для модели AIC, но за счет большего количества параметров значения скорректированного коэффициента детерминации и среднеквадратичной ошибки моделирования оказываются выше.

Таблица 2: Table 2. Parameters of the built models.

	r-squared	Adj. r-squared	Model MSe	Model df	Prediction MSe
Null	0.000	0.000	25519	42	25295
A	0.813	0.803	11322	40	12177
B	0.901	0.894	8321	39	8935
C	0.905	0.898	8166	39	8466
AIC	0.924	0.919	7284	39	7329

	r-squared	Adj. r-squared	Model MSe	Model df	Prediction MSe
Full	0.925	0.912	7557	36	7975

Выводы

Проведенный анализ показал, что величина годового стока значимо связана с величиной зимних осадков, причем эта связь может быть описана линейной регрессионной моделью. Приемлемое качество моделирования (для лучшей модели AIC среднеквадратичная ошибка предсказания 10.6% от медианного значения выборки) было достигнуто с использованием моделей с тремя предикторами. Модель с двумя предикторами является недообученной (ошибка предсказания 17.6%), а увеличение количества предикторов в модели даже до шести делает ее переобученной (ошибка предсказания 11.5%).

Список литературы

[1] https://cran.r-project.org/web/packages/alr4/index.html [2] https://cran.r-project.org/web/packages/tidyverse/index.html
[3] https://cran.r-project.org/web/packages/ggplot2/index.html
[4] https://cran.r-project.org/web/packages/dplyr/index.html
[5] https://cran.r-project.org/web/packages/tidyr/index.html
[6] https://cran.r-project.org/web/packages/cowplot/index.html
[7] https://cran.r-project.org/web/packages/qqplotr/index.html
[8] https://cran.r-project.org/web/packages/DAAG/index.html
[9] https://cran.r-project.org/web/packages/GGally/index.html
[10] https://cran.r-project.org/web/packages/knitr/index.html
[11] https://cran.r-project.org/web/packages/alr4/alr4.pdf
[12] S. Weisberg, Applied linear regression, 3rd ed., Wiley-Interscience, 2005.
[13] https://en.wikipedia.org/w/index.php?title=Climate_of_California&section=1
[14] https://en.wikipedia.org/wiki/Stepwise_regression