Липецкий государственный технический университет

Кафедра прикладной математики

КОМПЬЮТЕРНЫЕ ТЕХНОЛОГИИ МАТЕМАТИЧЕСКИХ ИССЛЕДОВАНИЙ

Лекция 5

5. Регрессионные модели

Составитель - Сысоев А.С., к.т.н.

Липецк – 2017

Outline

5.3. Необычные наблюдения

Необычные наблюдения требуют отдельного изучения: либо потому, что они каким-то образом отличаются от прочих, либо потому, что они значительно влияют на общие результаты.

ВЫБРОСЫ

Характеризуются большими положительными или отрицательными остатками $\hat{Y}_i - Y_i$. Положительные остатки свидетельствуют о том, что модель *недооценивает* зависимую переменную, отрицательные остатки – признак *переоценки*.

В пакете `car` также реализован статистический тест на выбросы. Функция `outlierTest()` вычисляет *значение вероятности статистической ошибки первого рода с поправкой Бонферрони для наибольшего остатка Стьюдента*

```
> library(car)
> outlierTest(fit)
       rstudent unadjusted p-value Bonferonni p
Nevada      3.5            0.00095         0.048
```

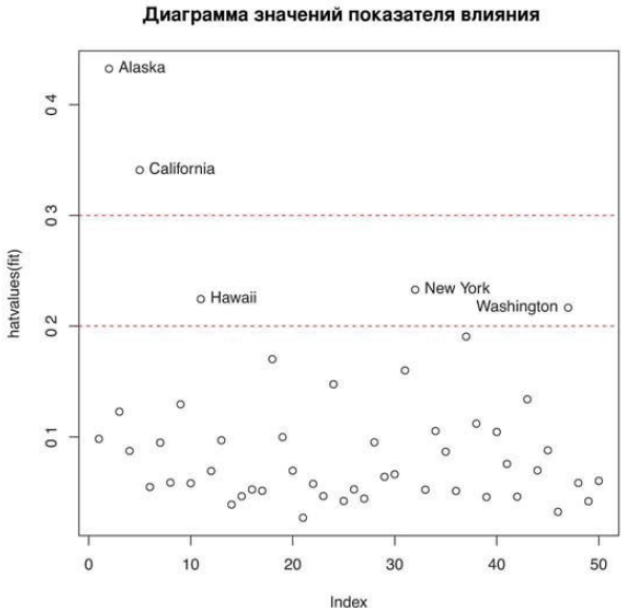
5.3. Необычные наблюдения

ТОЧКИ ВЫСОКОЙ НАПРЯЖЕННОСТИ

Точки с высокой напряженностью – это выбросы в отношении других независимых переменных. Они характеризуются необычным сочетанием значений независимых переменных. Значение зависимой переменной не используется при вычислении напряженности.

Идентифицируются при помощи показателя влияния (`hat statisctic`). Для определенного набора данных среднее значение этой статистики вычисляется как *p*/*n*, где *p* – это число параметров в модели (включая свободный член), *n* – размер выборки. Наблюдения, для которых значение этой статистики превышает среднее в два или три раза, должны быть проанализированы.

```
hat.plot <- function(fit) {
              p <- length(coefficients(fit))
              n <- length(fitted(fit))
              plot(hatvalues(fit), main="Диаграмма значений
➥показателя влияния")
              abline(h=c(2,3)*p/n, col="red", lty=2)
              identify(1:n, hatvalues(fit), names(hatvalues(fit)))
            }
hat.plot(fit)
```



5.3. Необычные наблюдения

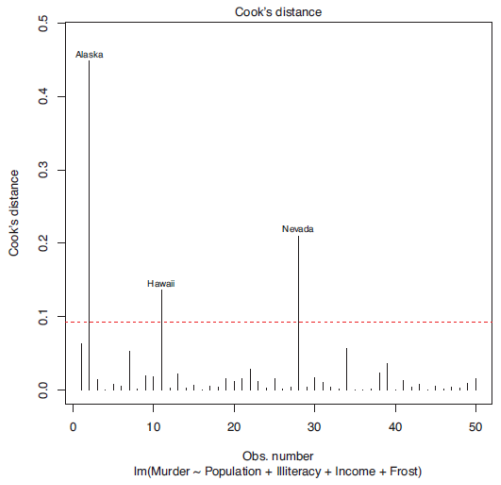
ВЛИЯТЕЛЬНЫЕ НАБЛЮДЕНИЯ

Влиятельные наблюдения – это наблюдения, которые оказывают непропорционально большое влияние на значения параметров модели.

Существуют два метода обнаружения влиятельных наблюдений: *расстояние Кука (или D-статистика)* и *диаграммы добавленных переменных*.

Значения расстояния Кука, превышающие 4/(n – k – 1), где n – объем выборки, а k – число независимых переменных, свидетельствуют о влиятельных наблюдениях.

```
cutoff <- 4/(nrow(states)-length(fit$coefficients)-2)
plot(fit, which=4, cook.levels=cutoff)
abline(h=cutoff, lty=2, col="red")
```



Диаграммы дистанции Кука помогают обнаружить влиятельные наблюдения, но они не позволяют понять, как эти наблюдения влияют на модель.

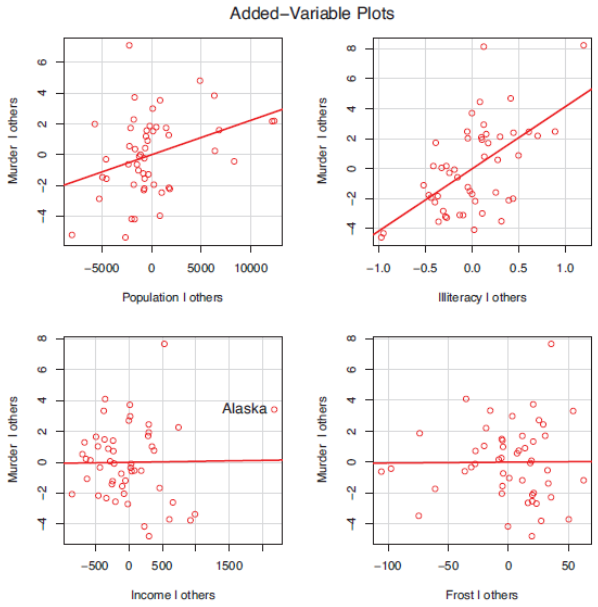
5.3. Необычные наблюдения

ВЛИЯТЕЛЬНЫЕ НАБЛЮДЕНИЯ (ПРОДОЛЖЕНИЕ)

Для одной зависимой и k независимых переменных создается k диаграмм добавленных переменных.

Для каждой независимой переменной X_k отображаются остатки от регрессии зависимой переменной по остальным k – 1 независимым переменным. Диаграммы добавленных переменных строятся при помощи функции `avPlots()` из пакета `car`

```
library(car)
avPlots(fit, ask=FALSE, onepage=TRUE, id.method="identify")
```



Прямая на каждой диаграмме – это регрессионный коэффициент для данной независимой переменной. Вклад влиятельных наблюдений можно оценить, если представить, как изменится линия, если удалить точку, соответствующую данному наблюдению.

Можно свести информацию о выбросах, точках с высокой напряженностью и влиятельных наблюдениях на одну диаграмму при помощи функции `influencePlot()` из пакета `car`

```
library(car)
influencePlot(fit, id.method="identify", main="Диаграмма влияния",
              sub="Размер круга пропорционален расстоянию Кука")
```

5.4. Способы корректировки регрессионной модели

УДАЛЕНИЕ НАБЛЮДЕНИЙ

Удаление выбросов часто может улучшить соответствие набора данных требованию нормальности. Влиятельные наблюдения также часто удаляют, поскольку они слишком сильно влияют на результаты. После удаления наибольшего выброса или влиятельного наблюдения модель подбирается заново. Если после этого все равно остаются выбросы или влиятельные наблюдения, процесс повторяется, пока не будет достигнуто допустимое соответствие модели данным.

ПРЕОБРАЗОВАНИЕ ПЕРЕМЕННЫХ

Когда модели не отвечают требованию нормальности, линейности или гомоскедастичности, трансформация одной или более переменных может улучшить или исправить ситуацию. Преобразования обычно заключаются в замене переменной Y на переменную Y^{λ}.

	−2	−1	−0.5	0	0.5	1	2
Преобразование	$1/Y^2$	$1/Y$	$1/\sqrt{Y}$	$\log(Y)$	$\sqrt{Y}$	отсутствует	Y^2

Если модель не соответствует требованиям нормальности, обычно пытаются преобразовать зависимую переменную. При помощи функции `powerTransform()` из пакета `car` можно оценить по *методу максимального правдоподобия* величину λ, возведение в которую, скорее всего, нормализует переменную X^{λ}.

```
> library(car)
> summary(powerTransform(states$Murder))

bcPower Transformation to Normality
             Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
states$Murder       0.6     0.26            0.088               1.1

Likelihood ratio tests about transformation parameters
                      LRT df  pval
LR test, lambda=(0) 5.7   1 0.017
LR test, lambda=(1) 2.1   1 0.145
```

Из результата применения функции следует, что переменную `Murder` можно нормализовать, заменив ее на $\texttt{Murder}^{0.5}$.

ПРЕОБРАЗОВАНИЕ ПЕРЕМЕННЫХ (ПРОДОЛЖЕНИЕ)

В том случае, когда требования линейности не выполняются, обычно помогает преобразование независимых переменных. Для оценки методом наибольшего правдоподобия той степени, в которую нужно возвести независимые переменные для большего соответствия модели требованию линейности, можно использовать функцию `boxTidwell()` из пакета `car`.

Пример *преобразований Бокса-Тидвелла (Box-Tidwell)*

```
> library(car)
> boxTidwell(Murder~Population+Illiteracy,data=states)

            Score Statistic p-value MLE of lambda
Population              -0.32    0.75          0.87
Illiteracy               0.62    0.54          1.36
```

Из полученного результата следует, что для получения большей линейности стоит попробовать преобразования $\texttt{Population}^{0.87}$ и $\texttt{Illiteracy}^{1.35}$. Однако результаты теста для переменных `Population (p = 0.75)` и `Illiteracy (p = 0.54)` свидетельствуют, что ни одну из них не нужно преобразовывать.

Преобразования зависимой переменной могут помочь в случае гетероскедастичности (непостоянной дисперсии остатков). Функция `spreadLevelPlot()` из пакета `car` позволяет понять, в какую степень нужно возвести зависимую переменную, чтобы увеличить гомоскедастичность.

ДОБАВЛЕНИЕ ИЛИ УДАЛЕНИЕ ПЕРЕМЕННЫХ

Изменение числа переменных, входящих в модель, будет влиять на степень ее соответствия данным. Иногда добавление важной переменной может исправить многие проблемы. Удаление причиняющих беспокойство переменных может привести к аналогичному эффекту.

ПРИМЕНЕНИЕ ДРУГОГО ПОДХОДА

- При наличии выбросов и/или влиятельных наблюдений можно использовать устойчивую регрессионную модель, а не МНК-регрессию.
- Если не выполняется требование нормальности, можно подобрать нелинейную регрессионную модель.
- В случае отклонения от независимости ошибок можно применить модели, которые учитывают структуру остатков, – такие как модели временных рядов или многоуровневые регрессионные модели.
- Если требования, лежащие в основе МНК-регрессии, не выполняются, вы можете обратиться к обобщенным линейным моделям.

5.5. Сравнение моделей и выбор лучшей

СРАВНЕНИЕ МОДЕЛЕЙ

Информационный критерий Акаике (Akaike Information Criterion, AIC). При расчете этого критерия учитывается статистическое соответствие модели данным и число необходимых для достижения этого соответствия параметров. Предпочтение нужно отдавать моделям с меньшими значениями AIC, указывающими на хорошее соответствие данным при использовании меньшего числа параметров. Этот критерий вычисляется при помощи функции `AIC()`

```
> fit1 <- lm(Murder ~ Population + Illiteracy + Income + Frost,
            data=states)
> fit2 <- lm(Murder ~ Population + Illiteracy, data=states)
> AIC(fit1,fit2)
     df      AIC
fit1  6 241.6429
fit2  4 237.6565
```

ПОШАГОВАЯ РЕГРЕССИЯ

При пошаговом выборе переменные добавляются в модель или удаляются из нее по одной, пока не будет достигнуто заданное значение критерия для остановки процесса.

При методе пошагового включения (forward stepwise) переменные по одной добавляются в модель, пока добавление новых переменных не перестанет ее улучшать.

При пошаговом исключении (backward stepwise) начинают с модели, включающей все независимые переменные, а потом удаляют их по одной до тех пор, пока модель не начнет ухудшаться.

При комбинированном методе (stepwise stepwise) совмещены оба подхода. Переменные добавляются по одной, однако на каждом шаге происходит переоценка модели, и те переменные, которые не вносят значительного вклада, удаляются.

ПОШАГОВАЯ РЕГРЕССИЯ (ПРОДОЛЖЕНИЕ)

Результат применения метода пошаговой регрессии зависит от критериев включения или удаления переменных. При помощи функции `stepAIC()` из пакета `MASS` можно провести все три типа пошаговой регрессии с использованием точного критерия `AIC`.

Пошаговая регрессия – спорный подход. Хотя с его помощью можно найти хорошую модель, нет гарантии, что она будет лучшей, поскольку не рассмотрены все возможные модели. Попытка обойти это ограничение делается при использовании *регрессии по всем подмножествам*.

РЕГРЕССИЯ ПО ВСЕМ ПОДМНОЖЕСТВАМ

В ходе регрессии по всем подмножествам исследуются все возможные модели. Регрессия по всем подмножествам проводится при помощи функции `regsubsets()` из пакета `leaps`. В качестве критерия «лучшей» модели можно выбрать *коэффициент детерминации, скорректированный коэффициент детерминации* или *Ср-статистику Мэллоуса* (`Mallows Cp statistic`).

Коэффициент детерминации – это доля дисперсии зависимой переменной, объясненная независимыми переменными.

Скорректированный коэффициент детерминации учитывает число параметров модели.

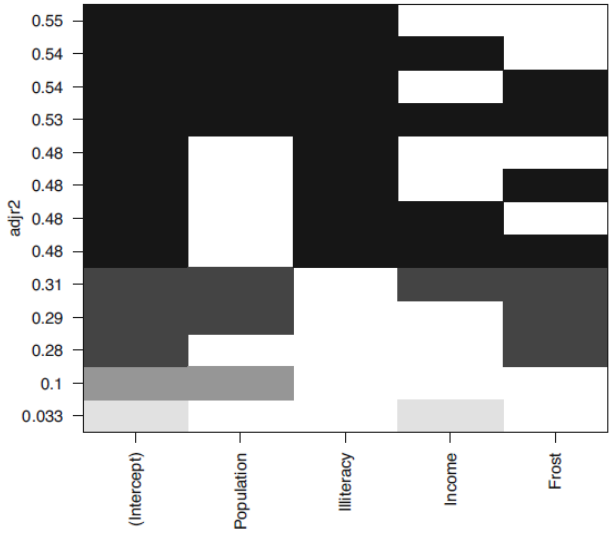
Статистика Мэллоуса также используется в качестве критерия «лучшей» модели. Считается, что для хорошей модели эта статистика должна принимать значения, близкие к числу параметров модели (включая свободный член).

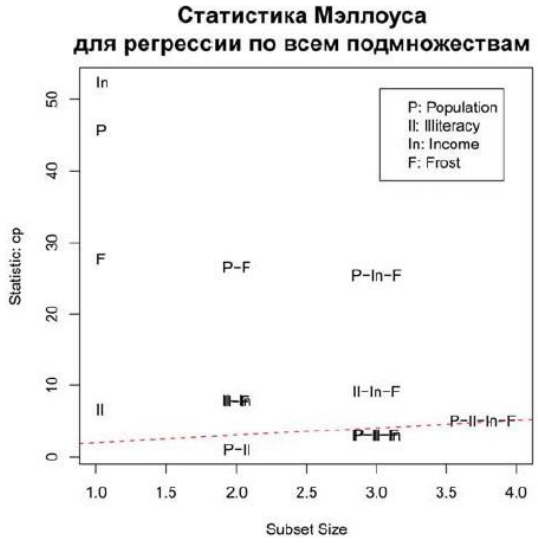
5.5. Сравнение моделей и выбор лучшей

РЕГРЕССИЯ ПО ВСЕМ ПОДМНОЖЕСТВАМ (ПРОДОЛЖЕНИЕ)

```
library(leaps)
leaps <-regsubsets(Murder ~ Population + Illiteracy + Income +
  Frost, data=states, nbest=4)
plot(leaps, scale="adjr2")

library(car)
subsets(leaps, statistic="cp",
  main="Статистика Мэллоуса для регрессии по всем подмножествам")
abline(1,1,lty=2,col="red")
```





Лучшие четыре модели для подмножеств всех размерностей, определенные на основании скорректированного коэффициента детерминации

Лучшие четыре модели для подмножеств всех размерностей, определенные на основании Ср-статистики Мэллоуса

5.5. Сравнение моделей и выбор лучшей

КРОСС-ВАЛИДАЦИЯ

Насколько хорошо полученное уравнение работает в реальном мире?

При кросс-валидации часть данных используется как обучающая выборка, а часть – как тестовая. Регрессионное уравнение подгоняется для обучающей выборки, а затем применяется для проверочной.

При *k*-кратной кросс-валидации выборка разделяется на *k* подвыборок. Каждая из них играет роль тестовой выборки, а объединенные данные оставшихся *k* – 1 подвыборок используются как обучающая группа. Применимость *k* уравнений к *k* тестовым выборкам фиксируется и затем усредняется. Если *k* = *n* (общему числу наблюдений), то такой подход называется оценкой по *методу «складного ножа»* (последовательного исключения значений выборки с возвратом – jackknifing).

Выполнить *k*-кратную кросс-валидацию можно при помощи функции `crossval()` из пакета `bootstrap`.

5.5. Обобщенные линейные модели

Не делается никаких предположений относительно распределения значений независимых переменных X_j. Они, в отличие от Y, не обязательно должны иметь нормальное распределение. Они часто бывают категориальными. Кроме того, допускаются нелинейные комбинации независимых переменных.

При создании обобщенных линейных моделей подбирают модели в виде

$$g(\mu_Y) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j ,$$

где g() - это функция условного среднего (называемая *связующей функцией*).

Обобщенные линейные модели обычно подгоняются в R при помощи функции `glm()`

```
glm(formula, family=family(link=function), data=)
```

Семейство распределений	Связующая функция по умолчанию
binomial	(link = "logit")
gaussian	(link = "identity")
gamma	(link = "inverse")
inverse.gaussian	(link = "1/mu^2")
poisson	(link = "log")
quasi	(link = "identity", variance = "constant")
quasibinomial	(link = "logit")
quasipoisson	(link = "log")

Функция `glm()` позволяет подгонять разные распространенные модели, включая логистическую и пуассоновскую регрессии, а также модели для анализа выживания

5.5. Обобщенные линейные модели

5.5.1. Соответствие модели данным и регрессионная диагностика

Проверка адекватности модели также важна для обобщенных линейных моделей, как и для стандартных (МНК) линейных моделей.

Вычисляемые в R стандартизованные по Стьюденту остатки, значения показателя влияния наблюдения и D-статистика Кука будут лишь примерными.

Кроме того, не существует общего согласия по поводу пороговых значений для обнаружения проблемных наблюдений. Следует принимать решения по поводу значений этих статистик методом сравнения их друг с другом.

Один подход заключается в том, чтобы графически изобразить значения статистики для всех наблюдений и поискать необычно большие значения.

Диагностические диаграммы обычно наиболее полезны, когда зависимая переменная имеет много значений. Когда зависимая переменная принимает ограниченное число значений (например, логистическая регрессия), польза от диагностических диаграмм не так велика.

Библиографический список

Кабаков Р. К. (2014) R в действии. Анализ и визуализация данных на языке R Издательство: ДМК Пресс, 580 с.

Шипунов А. Б., Балдин Е. М., Волкова П. А., Коробейников А. И., Назарова С. А., Петров С. В., Суфиянов В. Г. (2012) Наглядная статистика. Используем R! - М.: ДМК Пресс, 298 с.