Методы современной прикладной статистики
9. Регрессионный анализ II.

Родионов Игорь Владимирович
vecsell@gmail.com

Весна, 2018

Проверка предположений Гаусса-Маркова

Напомним, мы предполагали, что в рамках гауссовской линейной модели выполнены предположения

1. $E\varepsilon = 0$,
2. $D\varepsilon = \sigma^2 I_n$,
3. $\{\varepsilon_i\}_{i=1}^n$ независимы в совокупности,
4. ε_i имеют нормальное распределение,
5. $X^T X$ обратима.

Если хотя бы одно из этих предположений нарушено, то оценки коэффициентов регрессии $\{\beta_j\}$ могут оказаться смещенными и неустойчивыми, а доверительные интервалы для откликов на новых объектах будут иметь заниженные уровни доверия. Обсудим методы проверки этих предположений.

Кросс-валидация

Предположение $E\varepsilon = 0$ относится к выбору оптимальной модели в линейной регрессии, что мы обсуждали ранее.

Проверить предсказательную силу модели можно с помощью LOO-метода (leave-one-out): пусть $\hat{Y}_i$ – предсказание отклика на объекте x_i, полученное с помощью обучения модели по всем объектам, кроме x_i. Качество оценивается с помощью функции

$$LOO = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Но при большом количестве объектов LOO работает крайне медленно, в этом случае используется метод k-fold CV.

Кросс-валидация по блокам отличается от LOO тем, что мы разбиваем данные на k блоков (обычно $k \approx 10$) и получаем предсказание на объектах i-того блока с помощью модели, обученной по всем блокам, кроме i-того.

Проверка гомоскедастичности

Перейдем к проверке **условия гомоскедастичности** $D\varepsilon = \sigma^2 I_n$. Если условие не выполнено, то такая ситуация называется **гетероскедастичностью**.

Оценивать ошибки ε_i мы будем с помощью **остатков** $e_i = Y_i - \widehat{Y}_i$, где $\widehat{Y}_i = (X(X^T X)^{-1} X^T Y)_i$ – i-тая координата оценки отклика.

Заметим, что если $E\varepsilon = 0$, то

$$Ee = E(Y - \widehat{Y}) = E(Y - X\widehat{\beta}) = X\beta - X\beta = 0.$$

Анализ остатков

Далее, можем представить вектор остатков в виде $e = (I - H)Y$, где $H = X(X^TX)^{-1}X^T$. Поскольку $H^2 = H$ и $H^T = H$, то

$$Var[e] = Var[(I-H)Y] = (I-H)Var[Y](I-H)^T = \sigma^2(I-H).$$

Т.е. $De_i = \sigma^2(1 - h_{ii})$, где $H = \|h_{ij}\|$. Чтобы унифицировать все остатки по распределению, перейдем к нормированным остаткам $\tilde{e}_i = \frac{e_i}{\sqrt{De_i}} = \frac{e_i}{\sigma\sqrt{1-h_{ii}}}$.

Но σ^2 неизвестно, поэтому заменим σ^2 на её оценку $\frac{1}{n-k}RSS$, тем самым переходя к рассмотрению **стьюдентизированных остатков**

$$d_i = \frac{e_i}{\sqrt{\frac{RSS}{n-k}}\sqrt{1 - h_{ii}}}.$$

Анализ остатков

Заметим, что $\sum_i h_{ii} = tr(H) = k$, действительно,

$$tr(H) = tr(X(X^TX)^{-1}X^T) = tr((X^TX)^{-1}X^TX) = tr(I_k) = k.$$

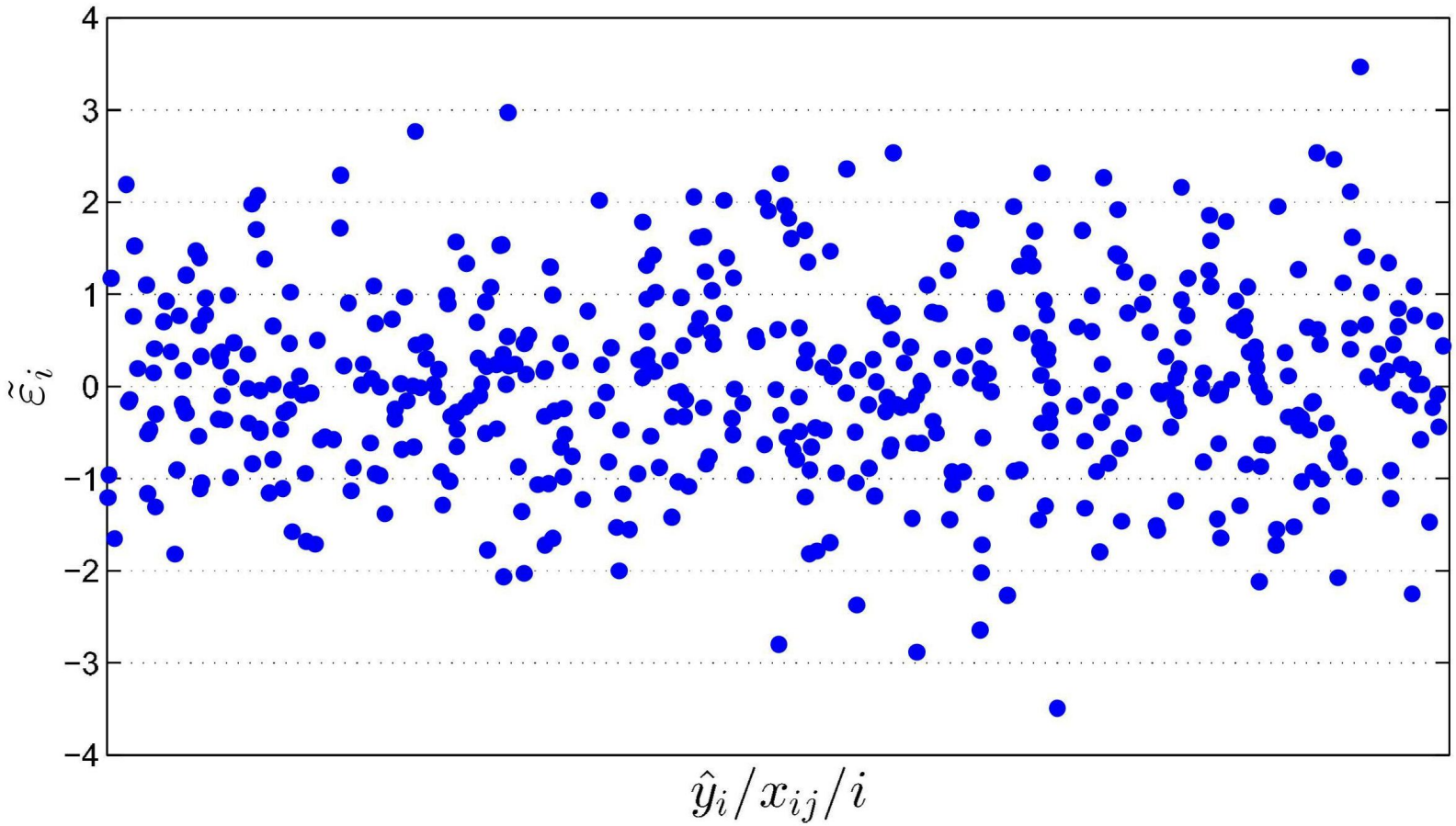
Поэтому если k много меньше n, то можно считать $h_{ii} \approx 0$ и перейти к рассмотрению **стандартизированных остатков**

$$\tilde{\varepsilon}_i = \frac{e_i}{\sqrt{\frac{RSS}{n-k}}}.$$

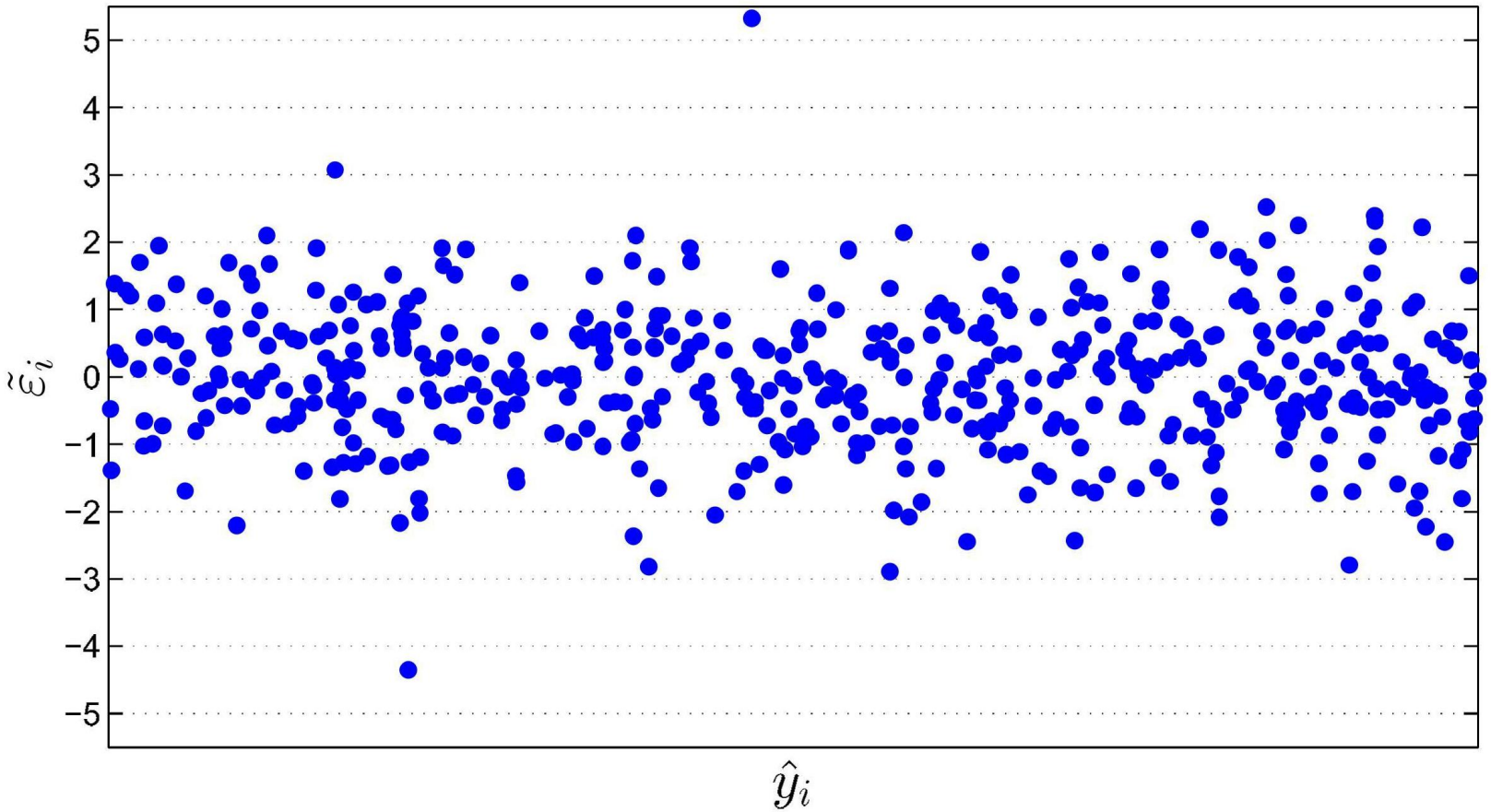
Хотя стандартизированные остатки являются зависимыми, при больших n их поведение схоже с поведением вектора независимых $N(0,1)$ случайных величин (при условии, что $\varepsilon \sim N(0, \sigma^2 I_n)$).

Визуальный анализ остатков

Строится график $\tilde{\varepsilon}_i$ от $\widehat{y}_i$, x_{ij} при фиксированном j или просто от i.

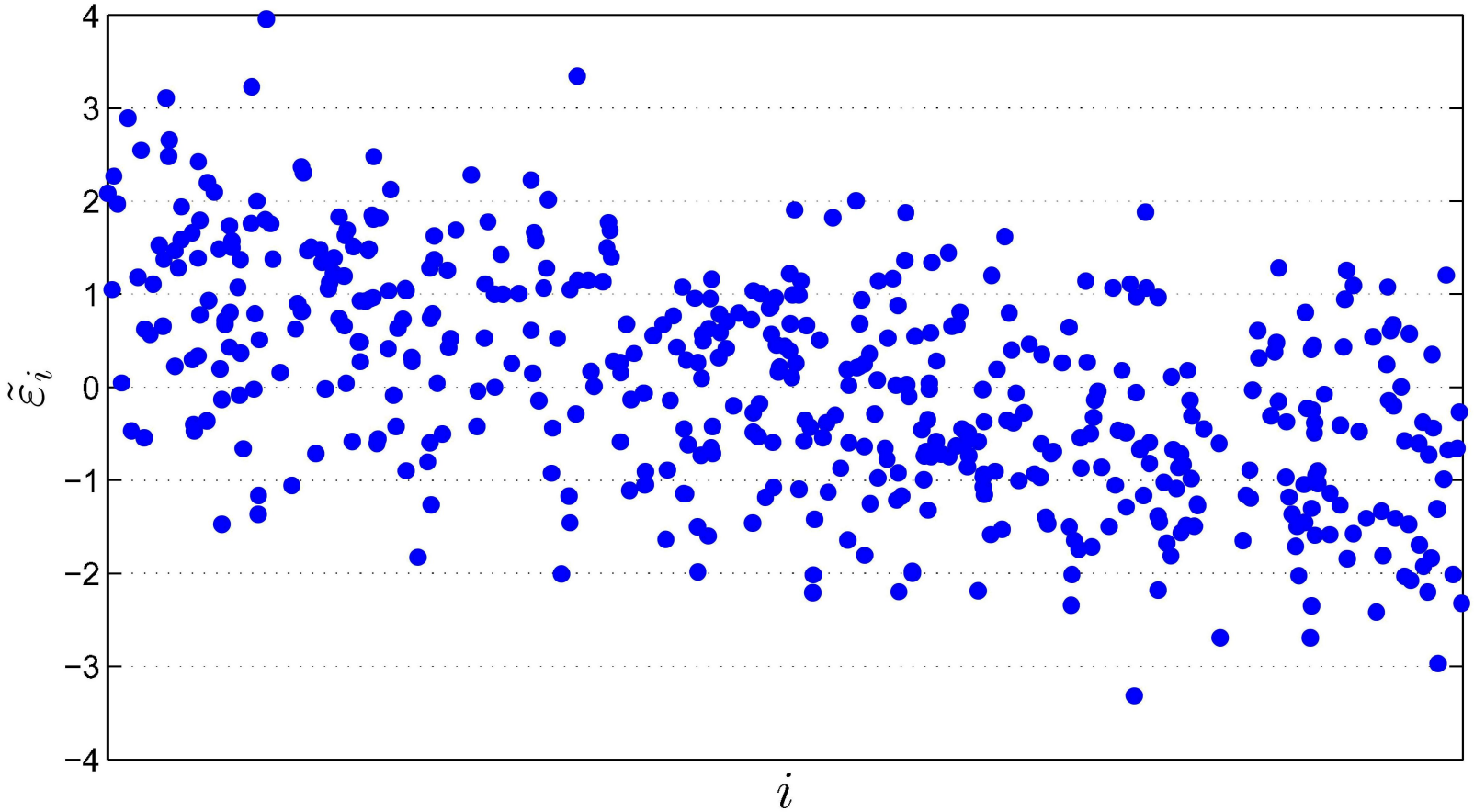


Визуальный анализ остатков



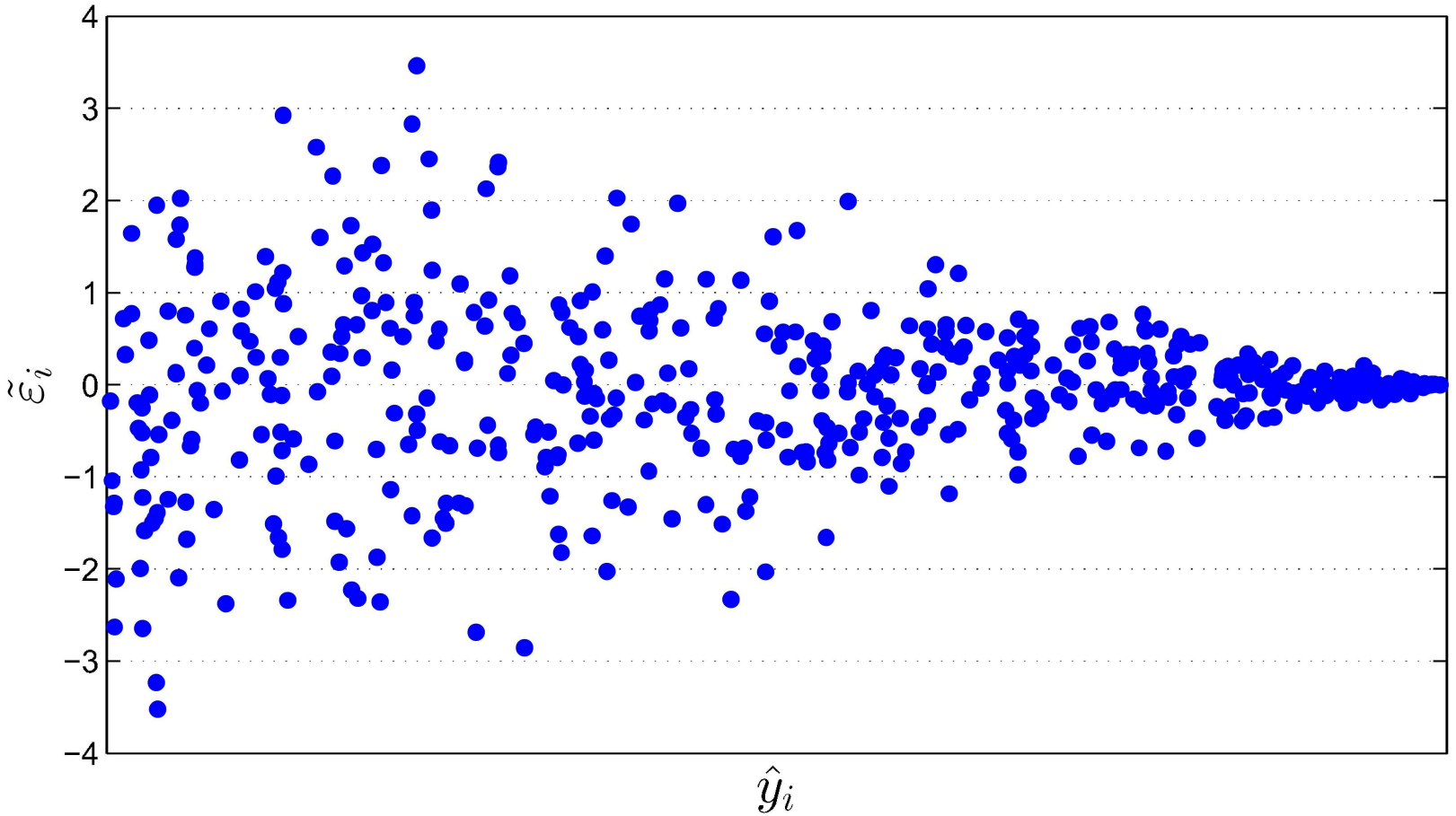
Возможно, в данных имеются выбросы.

Визуальный анализ остатков



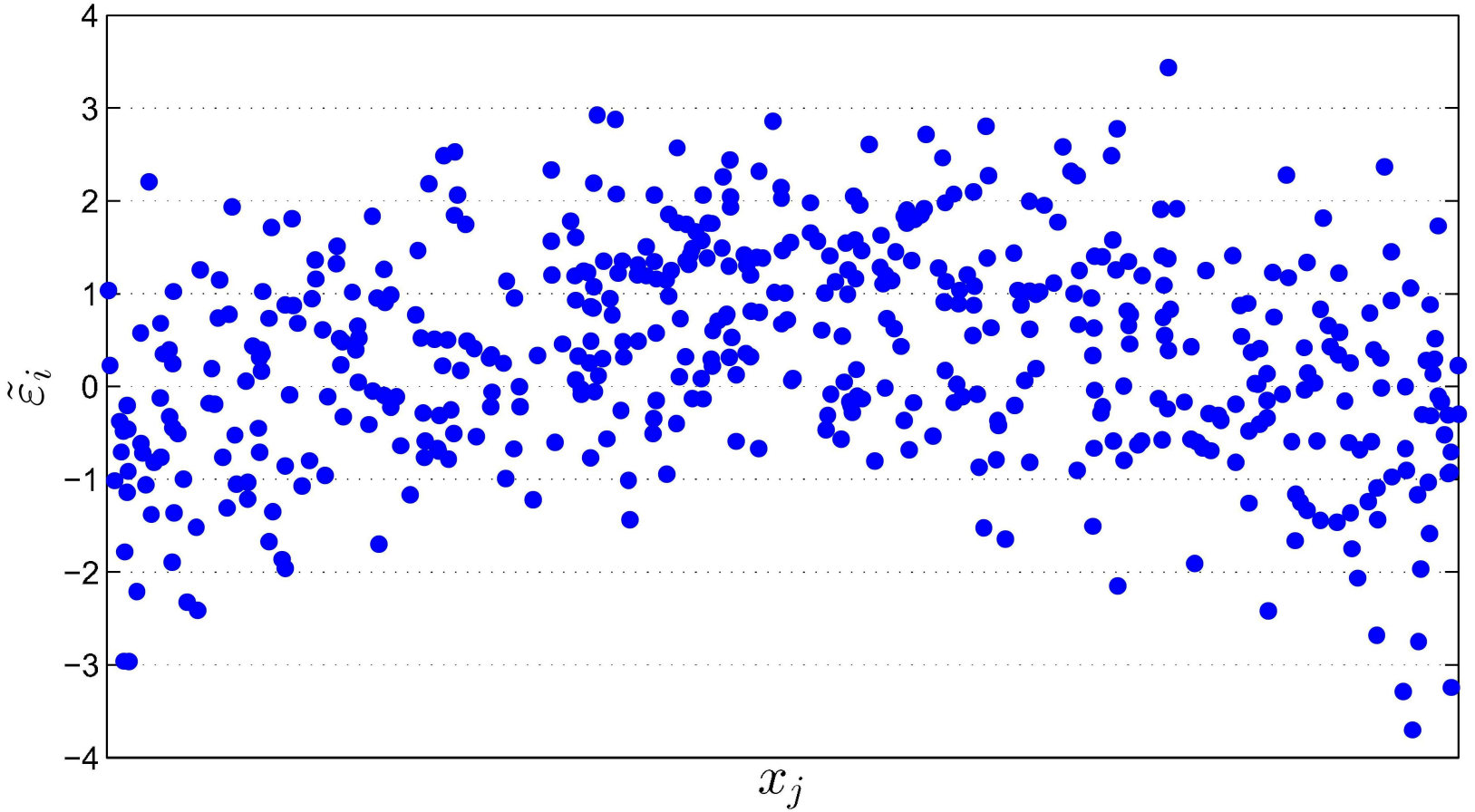
В данных имеется тренд (не учтен признак времени).

Визуальный анализ остатков



Гетероскедастичность данных.

Визуальный анализ остатков



В пространство признаков стоит добавить признак $(x_j)^2$.

Критерий Голдфелда-Квандта

Итак, проверим гипотезу гомоскедастичности $H_0 : \varepsilon \sim N(0, \sigma^2 I_n)$ (но критерий устойчив и к небольшому отклонению от нормальности распределения ε).

Упорядочим наблюдения по возможному возрастанию дисперсий ошибок и отбросим r центральных наблюдений. Построим 2 регрессионных модели: по первым $(n-r)/2$ наблюдениям и по последним $(n-r)/2$, вычислим по ним RSS_1 и RSS_2 соответственно. Если H_0 верна, то

$$F = \frac{RSS_2}{RSS_1} \sim F_{\frac{n-k}{2}-k, \frac{n-k}{2}-k}$$

Гипотезу H_0 следует отвергать, если $F > u_{1-\alpha}$, где $u_{1-\alpha}$ – квантиль данного распределения Фишера.

Критерий Уайта

Рассмотрим вспомогательную модель

$$e_i^2 = \alpha_0 + \sum_{j=1}^{k} \alpha_j x_{ij} + \sum_{j=1}^{k} \gamma_j x_{ij}^2 + \nu_i,$$

где e_i – обычные остатки в линейной регрессии. Проверим гипотезу $H_0 : \alpha_j = \gamma_j = 0 \ \forall j$.

Обозначим через $\widetilde{R}^2$ коэффициент детерминации вспомогательной модели. При верной гипотезе H_0 и достаточно большом n

$$n\widetilde{R}^2 \approx \chi^2_{2k}.$$

Если во введенной модели сразу полагать $\gamma_j = 0 \ \forall j$, то мы получим **критерий Бройша-Пагана**, и тогда при верной гипотезе $H_0' : \alpha_j = 0 \ \forall j$ выполнено $n\widetilde{R}^2 \approx \chi^2_k$.

Гетероскедастичность

Гетероскедастичность может быть следствием недоопределенности модели, тогда следует её доопределить с помощью рассмотренных методов.

От гетероскедастичности можно избавиться с помощью взвешенного МНК, давая разным объектам подходящие веса, если соотношения между дисперсиями известны заранее (или они явно определяются по графикам остатков).

Последствия гетероскедастичности:
1) нарушаются предположения критериев Стьюдента и Фишера и методов построения доверительных интервалов для σ^2, β и отклика на новом объекте (независимо от объёма выборки);
2) однако МНК-оценки β и R^2 остаются несмещёнными и состоятельными.

Устойчивая оценка дисперсии Уайта

Если от гетероскедастичности избавиться не удаётся, для оценки значимости признаков (и не только) можно использовать критерии, основанные на оценке HCE (White heteroscedasticity-consistent estimator, HC0) матрицы ковариаций вектора $\widehat{\beta}$

$$\widehat{\Omega} = (X^T X)^{-1}(X^T\{e_1^2, \ldots, e_n^2\}X)(X^T X)^{-1}.$$

Если $\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} N(0, \Omega)$ при $n \to +\infty$, то $n\widehat{\Omega}$ является состоятельной оценкой Ω.

Также вместо e_i^2 в выражение для оценки $\widehat{\Omega}$ можно подставлять $\frac{ne_i^2}{n-k}$, $\frac{e_i^2}{1-h_{ii}}$ или $\frac{e_i^2}{(1-h_{ii})^2}$ (поправки Маккиннона-Уайта), что будет увеличивать мощность наших критериев.

Проверка остатков на нормальность и независимость

Проверка остатков на нормальность, как правило, проводится не очень точными методами, такими, как визуальный анализ гистограммы и QQ-plot, и критериями типа Харке-Бера, потому что стандартизированные остатки зависимы.

Проверить остатки на независимость, по сути, невозможно. Если вдруг кажется, что остатки образуют временной ряд, то можно проверить остатки на наличие тренда (критериями случайности) и автокоррелированность (например, критерием Дарбина-Уотсона), что мы обсудим в лекции про временные ряды.

Преобразование Бокса-Кокса

Если всё же $\{\tilde{\varepsilon}_i\}$ не очень похожи на выборку из $N(0,1)$, то стоит задуматься о применении преобразования Бокса-Кокса к откликам модели.

Пусть все $Y_i > 0$, иначе возьмем $Y_i' = Y_i - \min_i Y_i$. Применим к вектору Y преобразование

$$V(y,\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda \dot{y}}, & \lambda \neq 0, \\ \dot{y} \ln y, & \lambda = 0, \end{cases}$$

где $\dot{y} = (Y_1 \dots Y_n)^{1/n}$ – среднее геометрическое отклика.

Преобразование Бокса-Кокса

Как выбирать оптимальное значени λ?

1) с помощью критериев проверки нормальности;

2) методом максимального правдоподобия для нормального распределения;

3) методом Бокса-Кокса: пусть $\lambda \in (-a, a)$, для каждого λ из этого интервала (по сетке) строим регрессию отклика $V(Y, \lambda)$ по признакам X и получаем $RSS(\lambda)$, по нему строим график $R^2(\lambda)$ и по графику выбираем оптимальное λ. Если значение $R^2(\lambda)$ не сильно отличается от $R^2(1)$, то выполнять преобразование не имеет смысла.

Мультиколлинеарность

Ситуация, когда матрица $X^T X$ является плохо обратимой, называется мультиколлинеарностью. Эта ситуация возникает, когда столбцы матрицы X близки к линейно зависимым, т.е. есть сильная совместная корреляция между признаками.

Критерием плохой обратимости, или плохой обусловленности, матрицы является ситуация, когда $\lambda_{max}/\lambda_{min} > 100$, где λ_{max} и λ_{min} – максимальное и минимальное собственное значение матрицы $X^T X$ соответственно.

Мультиколлинеарность чревата тем, что оценки откликов на новых объектах будут крайне неустойчивыми, поскольку $\widehat{\beta} \sim N(\beta, \sigma^2(X^T X)^{-1})$ и у некоторых коэффициентов будет очень большая дисперсия.

Мультиколлинеарность

Методы борьбы с мультиколлинеарностью:

1. Нормировка данных и их преобразование (например, методом Грама-Шмидта);
2. Variance inflation factor (VIF);
3. Методы понижения размерности (например, метод главных компонент);
4. Регуляризация.

Преобразования данных

Бывает так, что вектор-признак X_j имеет небольшой разброс значений и потому будет коррелировать с константным признаком X_0. В этой ситуации стоит перейти к нормированным данным

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{\sqrt{s_j^2}}$$

и центрировать отклик.

Метод ортогонализации Грама-Шмидта

Пусть X представимо в виде $X = GR$, где R – верхнетреугольная матрица размера $k \times k$, а G – ортогональная $n \times k$ матрица, т.е. $G^T G = I_k$. Таких разложений бесконечно много. Имея одно из них, можем найти псевдообратную матрицу в виде

$$X^+ = (R^T G^T GR)^{-1} R^T G^T = R^{-1}(R^{-1})^T R^T G^T = R^{-1} G^T.$$

Обозначим $\overline{v} = v/\|v\|$ для любого вектора v.

Процесс ортогонализации Грама-Шмидта строит ортогональные векторы $(g_1, \ldots, g_k)$, линейная оболочка которых совпадает с $(X_1, \ldots, X_k)$.

Метод ортогонализации Грама-Шмидта

Процесс устроен следующим образом:

$g_1 = X_1,$

$g_2 = X_2 - (\overline{g}_1, X_2)\overline{g}_1$

$\ldots$

$g_k = X_k - (\overline{g}_1, X_k)\overline{g}_1 - \ldots - (\overline{g}_{k-1}, X_k)\overline{g}_{k-1}.$

Тогда матрица $G = (\overline{g}_1, \ldots, \overline{g}_k)$ является ортонормированной, а R – верхнетреугольная матрица, для $i \leq j$

$$r_{ij} = \begin{cases} \overline{g}_i^T X_j, & i < j, \\ \|g_j\|, & i = j. \end{cases}$$

Существует также модификация алгоритма (см. лекции Воронцова по линейной регрессии), которая позволяет не только избавиться от мультиколлинеарности, но и произвести отбор признаков.

Variance Inflation Factor

Рассмотрим величину

$$VIF_j = \frac{1}{1 - R_j^2},$$

где R_j^2 – коэффициент детерминации модели

$$X_j = c_0 + c_1 X_1 + c_{j-1} X_{j-1} + \ldots + c_{j+1} X_{j+1} + \ldots + c_k X_k.$$

Логично, что если R_j^2 высок, то признак X_j хорошо объясняется остальными признаками, и его надо удалить. Рекомендуется удалять признак, если $VIF_j > 5$.

Гребневая регрессия

Всё это время мы искали оценки коэффициентов регрессии как точку минимума функционала

$$\widehat{\beta} = arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{k} \beta_j x_{ij} \right)^2 = arg\min_{\beta} \|Y - X\beta\|^2.$$

Решение $\widehat{\beta} = (X^T X)^{-1} X^T Y$ может иметь очень большую норму из-за плохой обустовленности матрицы $X^T X$, поэтому для большей устойчивости решения можно штрафовать большие значения нормы вектора весов β. Итак, определим для $\tau > 0$

$$\widehat{\beta}_{ridge} = arg\min_{\beta} \left(\|Y - X\beta\|^2 + \tau \|\beta\|^2 \right).$$

Гребневая регрессия

Оценку весов $\widehat{\beta}_{ridge}$ легко найти в явном виде,

$$\widehat{\beta}_{ridge} = (X^T X + \tau I_k)^{-1} X^T Y.$$

Добавление "гребня" τI_k к матрице $X^T X$ увеличивает все её собственные значения на τ, но не изменяет её собственных векторов. В результате матрица становится хорошо обусловленной, оставаясь в то же время "похожей" на исходную.

Гребневая регрессия

Недостаток метода в том, что $E\widehat{\beta}_{ridge} = (I_k - \tau W)\beta$, где $W = (X^T X + \tau I_k)^{-1}$, т.е. чем больше λ, тем больше смещение оценки вектора весов. Но зато норма матрицы ковариаций $Var\widehat{\beta}_{ridge} = \sigma^2 W X^T X W$ с ростом τ уменьшается.

Как выбирать оптимальное значение τ?

Теорема. $\exists\, \tau : E\|\widehat{\beta}_{ridge} - \beta\|^2 < E\|\widehat{\beta} - \beta\|^2$.

Единственная проблема, что на практике такое значение τ находить пока не научились, поэтому τ выбирается по следующему правилу: фиксируют некое не очень большое M и полагают τ таким, что $\frac{\lambda_{max}+\tau}{\lambda_{min}+\tau} = M$. Ну или берут τ от 0.1 до 0.4, если матрица X стандартизирована.

Least absolute shrinkage and selection operator

Помимо L_2-, существует ещё масса регуляризаций, наиболее популярной из которых является L_1-регуляризация, или LASSO Тибширани.

$$\widehat{\beta}_{lasso} = arg\min_{\beta}(\|Y - X\beta\|^2), \quad \sum_{j=1}^{k} |\beta_j| < \kappa.$$

К сожалению, явной формы оценки $\widehat{\beta}_{lasso}$ не существует.

При больших значениях κ решение совпадает с решением по методу наименьших квадратов. Чем меньше κ, тем больше коэффициентов β_j обнулятся (по теореме Куна-Такера) и тем меньше признаков войдут в уравнение регрессии. Т.е. по сути происходит отбор (селекция) признаков, что выгодно отличает LASSO от остальных видов регуляризации.

Проблема выбросов

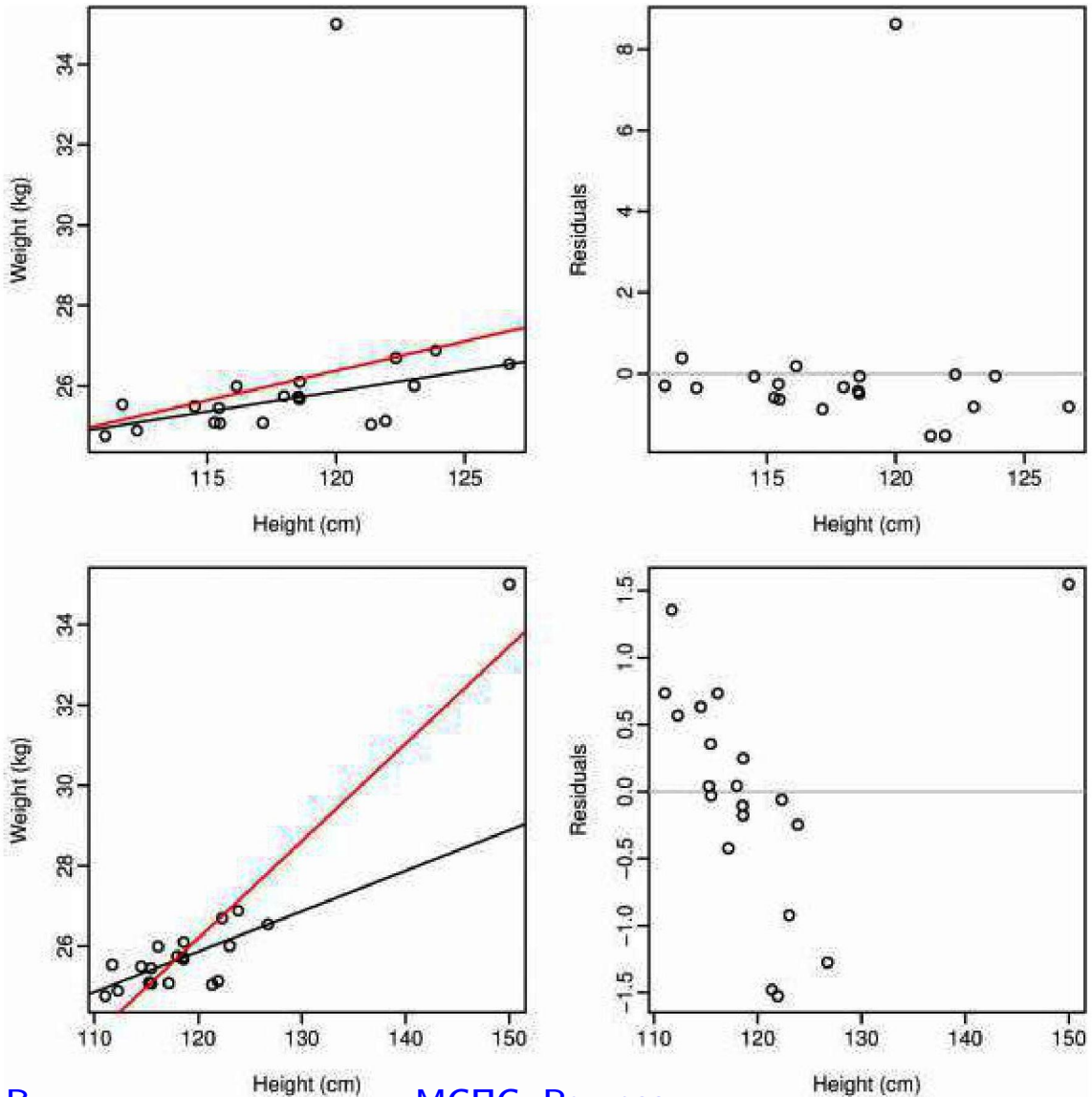
Не секрет, что при наличии выбросов в данных качество приближения в линейной регрессии может сильно ухудшиться (см. тот же квартет Энскомба). Методы решения этой проблемы:

1) удалить выбросы (например, с помощью расстояния Кука);

2) воспользоваться методами построения регрессии, которые устойчивы к выбросам (например, робастными регрессионными моделями или методом Тейла).

Проблема выбросов

Искажение регрессионной модели при наличии выбросов.



Расстояние Кука

Грубо говоря, это мера воздействия j-того объекта на уравнение регрессии:

$$D_j = \frac{\sum_{i=1}^{n}(\widehat{Y}_i - \widehat{Y}_{i\setminus j})^2}{RSS(k+1)} = \frac{e_j^2}{RSS(k+1)}\frac{h_{jj}}{(1-h_{jj})^2},$$

где предсказание $\widehat{Y}_{i\setminus j}$ получено по модели, обученной по всем объектам, кроме j-того, а h_{jj} – элемент матрицы $H = X(X^TX)^{-1}X^T$.

Выбирать, какие объекты выбрасывать, можно по методу крутого склона (отсекать те объекты, D_j которых значимо больше остальных) или по тому, превысил ли D_j заранее заданный порог. Варианты порогов: $D_j > 1$, $D_j > 4/n$, $D_j > 3\overline{D}$. Или по графику зависимости D_j от $\widehat{Y}_j$.

Метод Тейла

В рамках регрессионной модели

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i,\ 1 \le i \le n,$$

выбираем

$$\widehat{\beta}_j = med\left\{\frac{y_i - y_k}{x_{ij} - x_{kj}},\ i \neq k\right\},\ \widehat{\beta}_0 = med\left\{y_i - \sum_{j=1}^{k} \widehat{\beta}_j x_{ij}\right\}.$$

Поскольку используется медиана, данный метод является устойчивым к выбросам. Однако он плохо реагирует на мультиколлинеарность.

Робастные регрессионные модели

Ранее для поиска оценки вектора параметров в линейной модели мы использовали метод наименьших квадратов

$$\widehat{\beta} = arg \min_{\beta} \mathcal{R}(Y - X\beta), \text{ где } \mathcal{R}(x) = \|x\|^2.$$

Почему бы не использовать другие функции $\mathcal{R}$, для которых оценка вектора параметров будет не так сильно зависеть от выбросов, как в случае оценки методом наименьших квадратов? Будем считать, что

$$\mathcal{R}(x) = \sum_{i=1}^{n} \rho(x_i).$$

Требуем, чтобы такие функции $\rho(x)$ были: 1) симметричными, 2) неотрицательными, 3) монотонно неубывающими при $x > 0$.

Робастные регрессионные модели

Если ρ дифференцируема, то $\psi = \rho'$ называют функцией влияния. С помощью неё поиск оценки можно проводить, решая систему уравнений

$$\sum_{i=1}^{n} \psi\left(y_i - \sum_{j=1}^{k} x_{ij}\beta_j\right) x_{ij} = 0,\ 1 \le j \le k.$$

Чтобы метод был устойчивым к выбросам, выбирают модели с ограниченной функцией влияния (в МНК функция влияния равна $2x$ и не ограничена). Проблема в том, что для решения такой системы методы линейной алгебры, как правило, не помогают, и приходится обращаться к численным методам оптимизации, например, методу сопряженных градиентов.

Робастные регрессионные модели

Примеры функций ρ :

1) L_1 : $\rho(x) = |x|$;

2) Хьюбер: $\rho(x) = \frac{x^2}{2} I(|x| \leq k) + k(|x| - \frac{k}{2}) I(|x| > k)$;

3) L_p : $\rho(x) = |x|^p / p$;

4) Коши: $\rho(x) = \frac{c^2}{2} \ln(1 + (\frac{x}{c})^2)$;

5) Мешалкин: $\rho(x) = 1 - \exp(-\frac{x^2}{2\sigma})$;

6) German-Macclure: $\rho(x) = \frac{x^2}{2(1+x^2)}$;

В моделях L_1, Коши и Макклура решение не обязательно единственно. Параметр в модели Хьюбера обычно выбирают равным $k = 1.345$, кроме того, оценки методом Хьюбера близки к асимптотически эффективным.

Finita!