Тестирование регрессионных моделей

Даниленко Елена Владимировна, гр. 522

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель: д.ф.-м.н., проф. Мелас В.Б.
Рецензент: к.ф.-м.н. Шпилев П.В.

Санкт-Петербург
2008г.

Постановка задачи

Параметрические модели имеют важное практическое значение, поэтому ставится задача проверки адекватности используемой модели.

Рассмотрим следующую модель:

$$Y_i = m(\mathbf{x}_i) + \varepsilon_i, \quad i = 1, \ldots, n$$

где $\mathbf{x_i} \in \mathbb{R}^d$ — предиктор, $\varepsilon_i \sim N(0, \sigma^2(\mathbf{x}_i))$ – случайная ошибка, $E\varepsilon_i\varepsilon_j = 0,\ i \neq j,\ m$ – гладкая (неизвестная) регрессионная функция. Пусть $\mathcal{M} = \{m(\cdot, \theta) | \theta \in \Theta\}$ – некоторое параметрическое функциональное пространство.

Проверяем гипотезу

$$H_0: \quad m \in \mathcal{M} \qquad\qquad H_1: \quad m \notin \mathcal{M}$$

Проверка адекватности модели

Применяем следующие методы:

- Тест, основанный на разности параметрической и непараметрической оценок дисперсии, Dette (1998)
- Адаптивный тест Неймана, Fan, Huang (2001)
- F-тест

$$\mathcal{M} = \{g^T(x)\theta | \theta \in \Theta\},$$

где $g(x) = (g_1(x), \ldots, g_p(x))^T$ — ЛНЗ регрессионные функции.
Введем оценки дисперсии (параметрическую и непараметрическую):

$$\hat{\sigma}^2_{LSE} = \frac{1}{n-p} \sum_{j=1}^{n} (y_j - g^T(x_j)\hat{\theta}_n)^2$$

$$\hat{\sigma}^2_{HM} = \frac{1}{v} \sum_{j=1}^{n} \left(y_j - \frac{\sum_{i=1}^{n} y_i K\left(\frac{x_j - x_i}{h}\right)}{\sum_{l=1}^{n} K\left(\frac{x_j - x_l}{h}\right)} \right)^2$$

Тестовая статистика:

$T_n = \hat{\sigma}^2_{LSE} - \hat{\sigma}^2_{HM}$

Теорема [Dette,1998]

$$H_0 : \quad n\sqrt{h}(T_n + C_2 h^{2r}) \xrightarrow{D} N(0, \mu_0^2) \tag{1}$$

$$H_1 : \quad \sqrt{n}(T_n - M^2) \xrightarrow{D} N(0, \mu_1^2) \tag{2}$$

отвергаем гипотезу H_0, если $n\sqrt{h}T_n > u_{1-\alpha}\hat{\mu}_0$, $\quad h = o(n^{-2/(4r+1)})$

применяем преобразование Фурье к остаткам $\hat{\varepsilon}_i = Y_i - \mathbf{x}_i^T \hat{\theta}_n$
$\hat{\theta}_n \xrightarrow{p} \theta_0$ – состоятельность оценки МНК
$\eta_i = m(\mathbf{x}_i) - \mathbf{x}_i^T \theta_0$

$$\begin{aligned} \hat{\sigma}_1^2 &= \frac{1}{n - I_n} \sum_{i=I_n+1}^{n} \hat{\varepsilon}_i^{*2} - \left(\frac{1}{n - I_n} \sum_{i=I_n+1}^{n} \hat{\varepsilon}_i^{*} \right)^2 \\ \hat{\sigma}_2^2 &= \frac{1}{n - I_n} \sum_{i=I_n+1}^{n} \hat{\varepsilon}_i^{*4} - \left(\frac{1}{n - I_n} \sum_{i=I_n+1}^{n} \hat{\varepsilon}_i^{*2} \right)^2 \end{aligned}$$

$$\begin{aligned} T_{AN,1}^{*} &= \max_{1 \le m \le n} \frac{1}{\sqrt{2m\hat{\sigma}_1^4}} \sum_{i=1}^{m} (\hat{\varepsilon}_i^{*2} - \hat{\sigma}_1^2) \\ T_{AN,2}^{*} &= \max_{1 \le m \le n} \frac{1}{\sqrt{m\hat{\sigma}_2^2}} \sum_{i=1}^{m} (\hat{\varepsilon}_i^{*2} - \hat{\sigma}_1^2) \end{aligned}$$

Тестовая статистика:

$$T_{AN,j} = \sqrt{2 \ln \ln n} T_{AN,j}^{*} - (2 \ln \ln n + 0.5 \ln \ln \ln n - 0.5 \ln(4\pi))$$

Теорема 1 [Fan, 1996]

$$P(T_{AN,j} < x) \underset{n\to\infty}{\longrightarrow} \exp(-\exp(-x)), \qquad j = 1, 2.$$

Критическая область

$$T_{AN,j} > -\ln(\ln(1-\alpha)) \qquad (j = 1, 2) \tag{3}$$

имеет асимптотический уровень значимости α

Теорема 2 [Fan, 1996]

Если

$$(\ln\ln n)^{-1/2} \max_{1\le m\le n} m^{-1/2} \sum_{i=1}^{m} \eta_i^* \longrightarrow \infty$$

Тогда критическая область (3) имеет асимптотическую мощность 1.

Сходимость медленная, используем точные квантили распределения

отвергаем гипотезу H_0, если $T_{AN,j} > T_{\alpha,j}$

На рисунках изображена мощность тестов против данной параметризованной альтернативы в зависимости от значения параметра с уровнем значимости 5%.
Рассмотрены следующие виды альтернатив:

Название	**Вид**
квадратичная	$m(x) = 5x + \theta x^2$
тригонометрическая	$m(x) = 1 + \cos(\theta x \pi)$
логистическая	$m(x) = \dfrac{10}{1 + \theta \exp(-2x)}$
F-тест построен в модели	$Y = a_0 + a_1 x + a_2 x^2 + \varepsilon$

Таблица: Виды альтернатив в одномерном случае

Название	**Вид**
квадратичная	$m(x) = x_1 + \theta x_2^2 + 2x_3$
смешанная	$m(x) = x_1 + \theta x_2 \cdot x_3$
тригонометрическая	$m(x) = x_1 + \cos(\theta x_2 \pi) + 2x_3$
F-тест построен в модели	$Y = a_0 + \sum_i a_i x_i + \sum_{i \leq j} a_{i,j} x_i x_j + \varepsilon$

Таблица: Виды альтернатив в многомерном случае

Пример 1

$$Y = 5x + \theta x^2 + \varepsilon, \qquad \varepsilon \sim N(0,1)$$

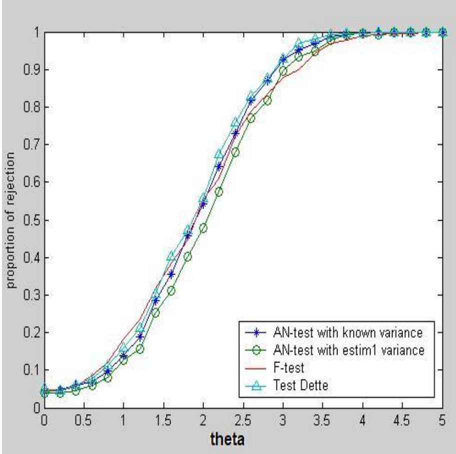


Рис.: Пример 1

- $x \sim p.p.(0,1)$
- фиксированный равномерный план — лучше всего для теста Dette
- гетероскедастическая модель $\sigma^2(x) = 3(1+x^2)/4$ — наблюдается незначительное снижение мощности
- все тесты ведут себя примерно одинаково, тест Dette несколько лучше

Пример 1

$$Y = 5x + \theta x^2 + \varepsilon, \qquad \varepsilon \sim N(0,1)$$

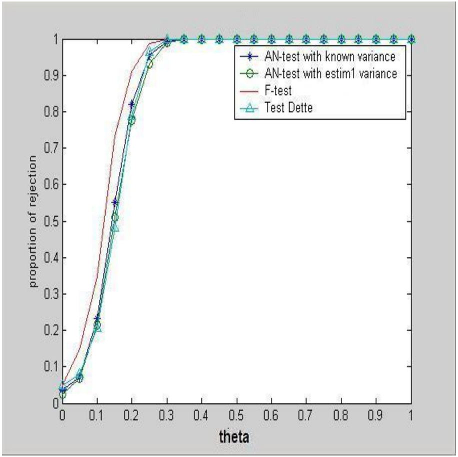


Рис.: Пример 1

- $x \sim p.p.(-2, 2)$
- при увеличении длины интервала мощности всех тестов уже при небольшом отклонении от нулевой гипотезы приближаются к единице

Пример 2

$$Y = 1 + \cos(\theta x \pi) + \varepsilon, \qquad \varepsilon \sim N(0,1)$$

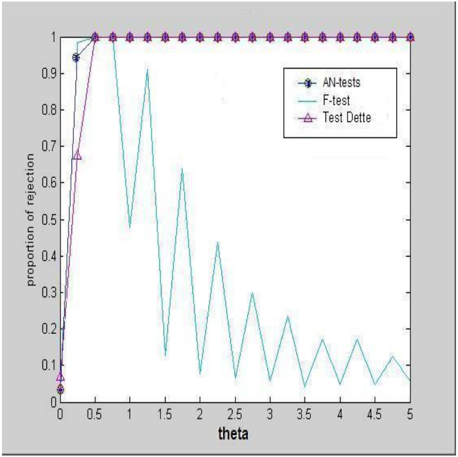


Рис.: Пример 2

- $x \sim p.p.(-2, 2)$
- F-тест значительно проигрывает
- При большом или очень малом значении параметра мощности тестов низкие, нужно корректировать длину промежутка, на котором берется план

Пример 2

$$Y = 1 + \cos(\theta x \pi) + \varepsilon, \qquad \varepsilon \sim N(0,1)$$

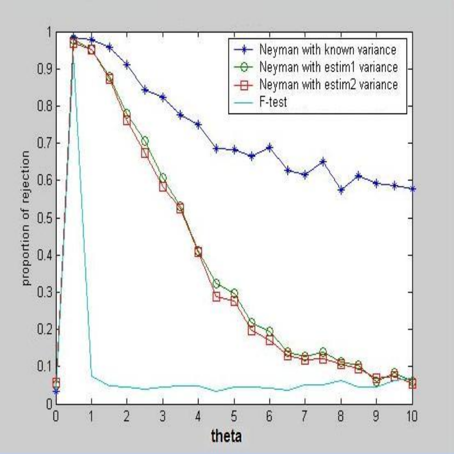
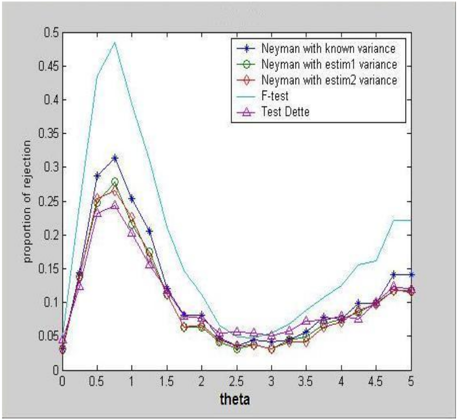


Рис.: Пример 2

- $x \sim N(0,1)$
- F-тест значительно проигрывает
- при увеличении параметра мощность АН-теста с оцененной дисперсией значительно снижается, нужно скорректировать оценку дисперсии, увеличив порог отсечения I_n

Пример 3

$$Y = \frac{10}{1+\theta\exp(-2x)} + \varepsilon, \qquad \varepsilon \sim N(0,1)$$



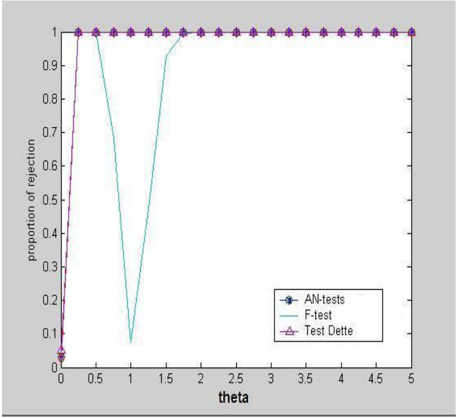


Рис.: Пример 3 ($x \sim p.p.(0,1)$ слева) ($x \sim p.p.(-2,2)$ справа)

тест Dette:

- $K(x_1, \ldots, x_d) = \prod_{j=1}^{d} K_j(x_j)$
- нормализующий фактор в условиях H_0: $n\sqrt{h_1 \cdots h_d}$

AH-тест:

- упорядочивать по первой ГК
- частично-линейная альтернатива $H1: \quad m(\mathbf{x}) = F(x_1) + \mathbf{x}_2^T \beta_2$
  упорядочить по x_1, оценить β_2
- аддитивная альтернатива $H1: \quad m(\mathbf{x}) = f_1(x_1) + f_2(x_2) + \ldots + f_d(x_d)$
  упорядочить по каждому из предикторов
  $\hat{T} = \max\limits_{1 \le j \le d} \hat{T}_j$

Пример 4

$$Y = X_1 + \theta X_2^2 + 2X_3 + \varepsilon, \qquad X_1, X_2, X_3 \sim p.p.(-2, 2)$$

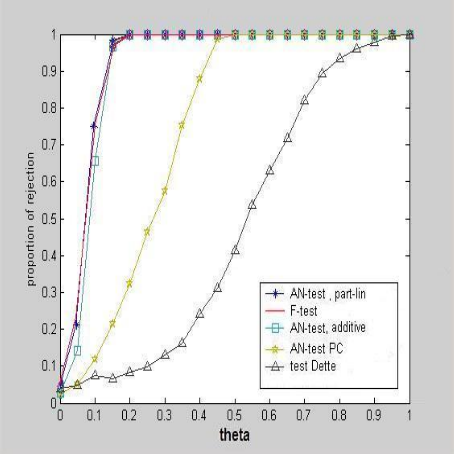


Рис.: Пример 4

- АН-тесты для аддитивной и частично линейной альтернативы, а также F-тест лидируют
- АН-тест с упорядочиванием по первой ГК немного проигрывает
- мощность теста Dette растет значительно медленнее

Пример 5

$$Y = X_1 + \theta X_2 \cdot X_3 + \varepsilon, \qquad X_1, X_2, X_3 \sim p.p.(-2, 2)$$

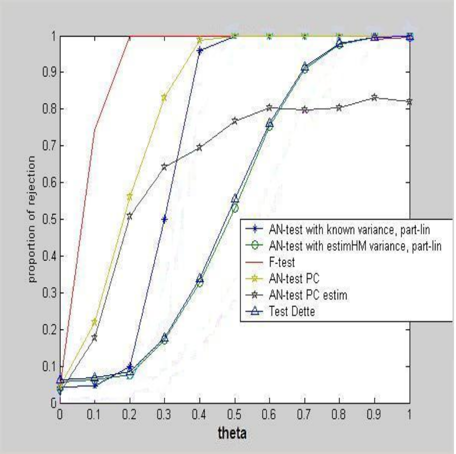


Рис.: Пример 5

- F-тест лидирует
- среди всех модификаций АН-тестов, лидирует АН-тест с упорядочиванием по первой ГК
- в АН-тестах для аддитивной и частично линейной альтернативы применили другую непараметрическую оценку дисперсии
- тест Dette практически не изменился

Пример 6

$$Y = X_1 + \cos(\theta X_2 \pi) + 2X_3 + \varepsilon, \qquad X_1, X_2, X_3 \sim p.p.(-2,2)$$

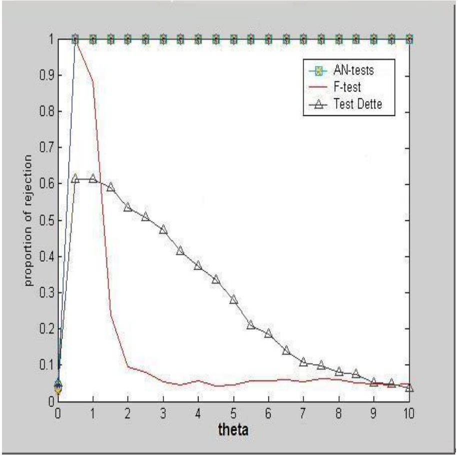


Рис.: Пример 6

- АН-тесты уже при небольшом значении параметра имеют высокую мощность
- F-тест провалился

Выводы

- Выполнены реализация и ранее не проводившееся сравнение методов
- Тест Dette обобщен на произвольный отрезок
- В AH-тесте использованы оценки дисперсии, значительно увеличивающие его мощность
- Оба теста, хоть и основаны на различных идеях, показывают примерно одинаковые результаты в одномерном случае. В многомерном тест Fan'a значительно лучше
- В случае, когда находимся в модели F-теста:
  – в одномерном случае тесты имеют незначительно отличающуюся мощность
  – в многомерной регрессии мощность AH-тест путем различных модификаций можно приблизить к мощности F-теста
  В остальных примерах:
  мощность обоих тестов значительно превосходит мощность F-теста