Seminar 10: Проверка линейных гипотез в гауссовской регрессионной модели

Carlos Buitrago

Table of contents

1 1. Гауссовская линейная модель 1

2 2. Распределение Фишера 2

3 3. Точный F-критерий линейных ограничений 2

4 4. Критерий Вальда 3

5 5. Одновыборочный t-критерий Стьюдента 4

6 6. Двухвыборочный t-критерий (равные дисперсии) 5

1 1. Гауссовская линейная модель

Пусть $X = Z\theta + \varepsilon$, где $X \in \mathbb{R}^n$, $Z \in \mathbb{R}^{n\times k}$ — известная матрица полного столбцового ранга $k < n$, $\theta \in \mathbb{R}^k$, а $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$, $\sigma^2 > 0$.

Оценка МНК и остаточная сумма квадратов:

$$\hat\theta = (Z^T Z)^{-1} Z^T X, \qquad RSS = \|X - Z\hat\theta\|^2.$$

Причём

$$\hat\theta \sim \mathcal{N}_k(\theta, \sigma^2 (Z^T Z)^{-1}), \quad \frac{RSS}{\sigma^2} \sim \chi^2_{n-k}, \quad \hat\theta \perp RSS.$$

> ❗ Линейная гипотеза
>
> $$H_0 : T\theta = \tau, \qquad H_1 : T\theta \neq \tau,$$
>
> где $T \in \mathbb{R}^{m\times k}$, $\tau \in \mathbb{R}^m$, $\operatorname{rank}(T) = m \le k$. При H_0 параметр лежит в аффинном подпространстве размерности $k - m$ (не обязательно линейном, если $\tau \neq 0$).

Утверждение. $\hat t = T\hat\theta$ — оптимальная (в смысле минимальной ковариационной матрицы среди несмещённых линейных оценок; при нормальности также UMVU) оценка $T\theta$.

$$\mathbb{E}\hat t = T\theta, \qquad \operatorname{Cov}(\hat t) = \sigma^2 T (Z^T Z)^{-1} T^T.$$

Обоснование: следует из теоремы Гаусса—Маркова; в нормальной модели достаточность и полнота соответствующей статистики дают UMVU.

2 2. Распределение Фишера

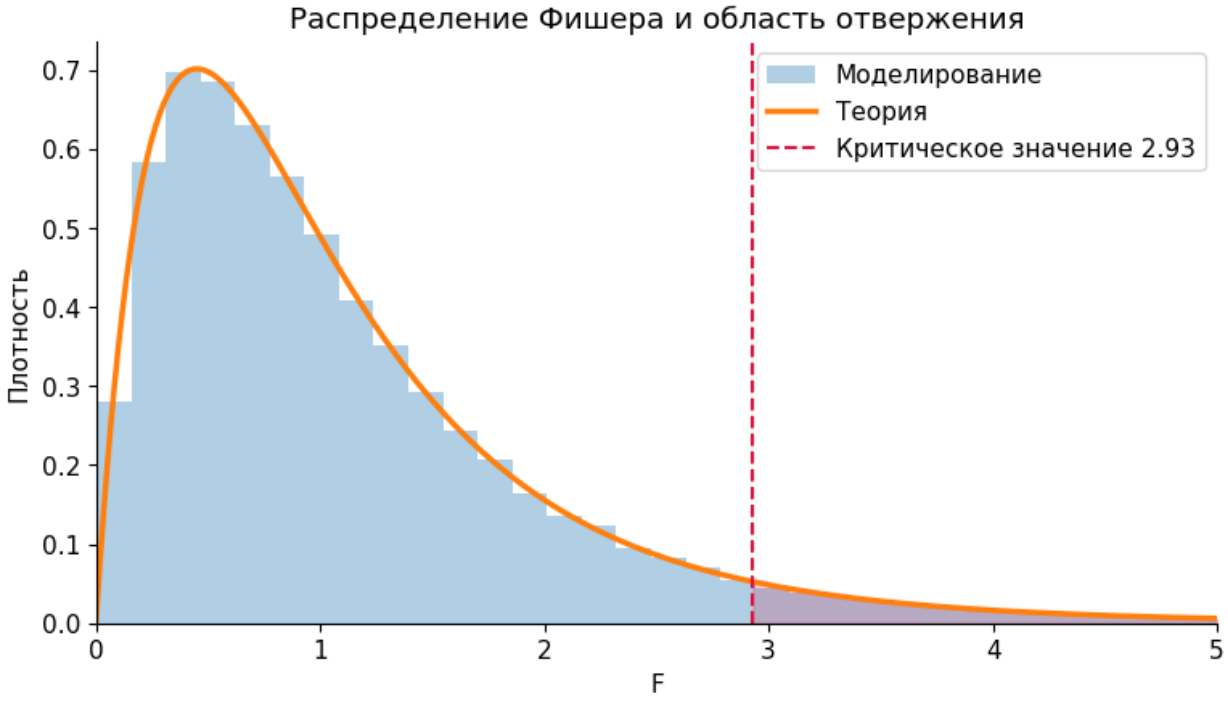
Определение

Если $U \sim \chi^2_a$, $V \sim \chi^2_b$ независимы, то

$$F = \frac{U/a}{V/b} \sim F_{a,b}.$$

Числа a, b называются степенями свободы числителя и знаменателя.

Критическая область правостороннего критерия: $F > F_{a,b;1-\alpha}$, где $F_{a,b;1-\alpha}$ — квантиль уровня $1-\alpha$.




```
Доля в критической области: 0.0486
```

3 3. Точный F-критерий линейных ограничений

Положим $B = T(Z^T Z)^{-1} T^T$. Из условия $\operatorname{rank}(T) = m$ следует, что B обратима.

F-тест (критерий Фишера)

При верной H_0 статистика

$$F_{\text{obs}} = \frac{(T\hat{\theta} - \tau)^T B^{-1} (T\hat{\theta} - \tau)/m}{RSS/(n-k)} \sim F_{m,n-k}.$$

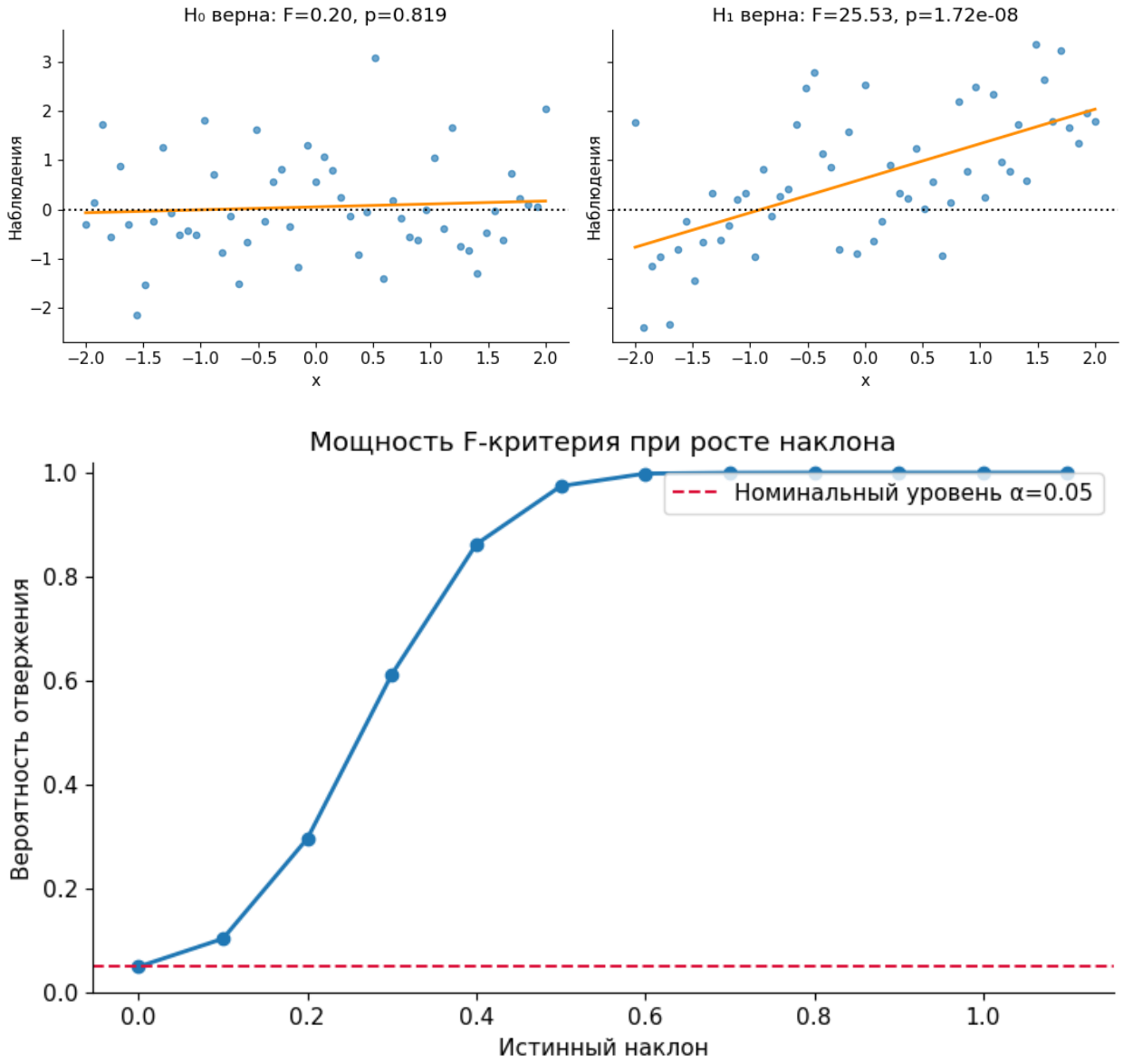
Гипотеза отвергается на уровне α, если

$$F_{\text{obs}} > F_{m,n-k;1-\alpha}.$$

p-значение равно $1 - F^{\text{CDF}}_{m,n-k}(F_{\text{obs}})$.

Почему это верно? При H_0 вектор $B^{-1/2}(T\hat{\theta} - \tau)/\sigma$ стандартный гауссовский размерности m, поэтому квадрат его нормы имеет распределение χ^2_m. Независимая остаточная сумма квадратов, делённая на σ^2, имеет распределение χ^2_{n-k}. Их нормированное отношение — распределение Фишера.

Эквивалентная форма: $F = ((RSS_0 - RSS_1)/m)/(RSS_1/(n-k))$, где RSS_0 — минимум при ограничениях, а RSS_1 — без ограничений.




```
Оценка ошибки I рода: 0.048857142857142856
```

4 4. Критерий Вальда

Рассмотрим скалярную гипотезу $H_0 : \theta = \theta_0$ против $H_1 : \theta \neq \theta_0$ и предположим, что

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta)), \qquad \frac{\hat{\sigma}}{\sigma(\theta)} \xrightarrow{P} 1.$$

Критерий Вальда

По теореме Слуцкого

$$W_n = \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\hat{\sigma}} \xrightarrow[H_0]{d} \mathcal{N}(0,1).$$

Приближённая критическая область: $|W_n| > z_{1-\alpha/2}$.

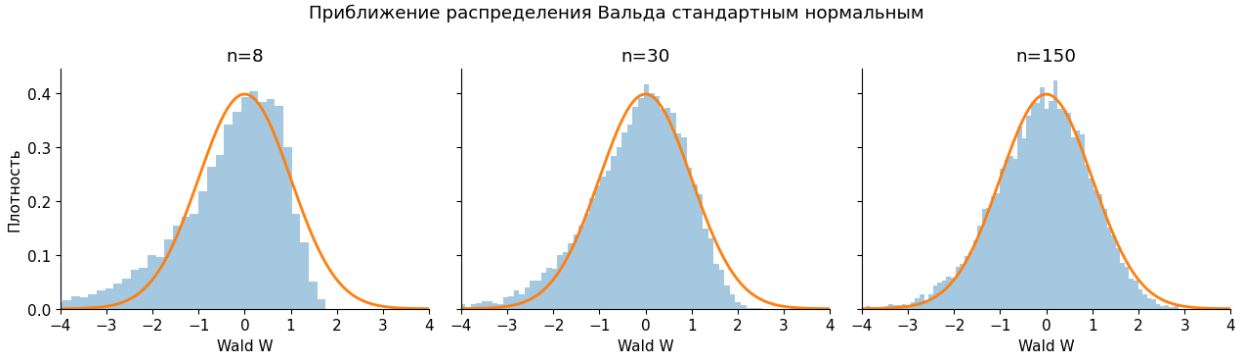
Для векторного параметра $\theta \in \mathbb{R}^k$ и m ограничений при подходящих условиях

$$W_n = n\,(T\hat{\theta} - \tau)^T (T\widehat{V}T^T)^{-1}(T\hat{\theta} - \tau) \xrightarrow[H_0]{d} \chi^2_m,$$

где $\widehat{V}$ состоятельно оценивает асимптотическую ковариационную матрицу $\sqrt{n}(\hat{\theta} - \theta)$.

Важно

Критерий Вальда обычно **асимптотический**, тогда как F-критерий выше **точный** при нормальных гомоскедастичных ошибках. Состоятельности $\hat{\sigma}$ самой по себе недостаточно: нужна корректная асимптотическая нормальность и положительная предельная дисперсия.



5 5. Одновыборочный t-критерий Стьюдента

Пусть $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(a, \sigma^2)$, дисперсия неизвестна. Проверяем $H_0 : a = a_0$ против $H_1 : a \neq a_0$.

Определим $\overline{X} = n^{-1}\sum_i X_i$, $S^2 = n^{-1}\sum_i (X_i - \overline{X})^2$ и $s^2 = (n-1)^{-1}\sum_i (X_i - \overline{X})^2$.

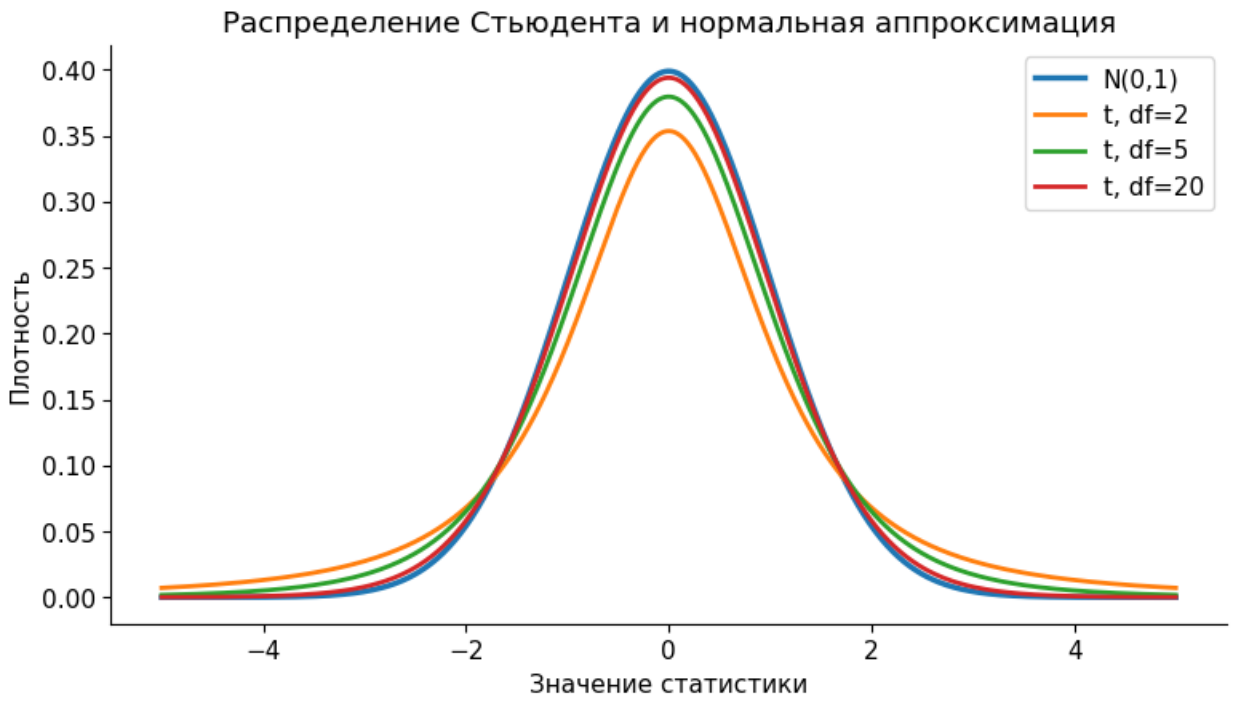
Точная статистика

$$t = \frac{\sqrt{n}(\overline{X} - a_0)}{s} = \sqrt{n-1}\frac{\overline{X} - a_0}{\sqrt{S^2}} \sim_{H_0} t_{n-1}.$$

Отвергаем H_0, если $|t| > t_{n-1;1-\alpha/2}$.

Доказательство: $\sqrt{n}(\overline{X} - a_0)/\sigma \sim \mathcal{N}(0,1)$ и $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$ независимы. Следовательно, их отношение имеет распределение Стьюдента.

Доверительный интервал уровня $1-\alpha$: $\overline{X} \pm t_{n-1;1-\alpha/2}\, s/\sqrt{n}$.




```
Доля отвержений при верной H0: 0.0507
```

6 6. Двухвыборочный t-критерий (равные дисперсии)

Пусть независимы $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(a_1, \sigma^2)$ и $Y_1, \dots, Y_m \overset{iid}{\sim} \mathcal{N}(a_2, \sigma^2)$. Проверяем $H_0 : a_1 = a_2$ против $H_1 : a_1 \neq a_2$.

Обозначим $s_X^2 = \frac{1}{n-1}\sum_i (X_i - \overline{X})^2$, $s_Y^2 = \frac{1}{m-1}\sum_j (Y_j - \overline{Y})^2$ и объединённую дисперсию

$$s_p^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}.$$

! Двухвыборочный t-критерий

$$t = \frac{\overline{X} - \overline{Y}}{s_p\sqrt{1/n + 1/m}} = \sqrt{\frac{nm}{n+m}}\frac{\overline{X} - \overline{Y}}{s_p} \sim_{H_0} t_{n+m-2}.$$

Отвергаем H_0, если $|t| > t_{n+m-2;1-\alpha/2}$.

Примечание к исходному конспекту: степень свободы — именно $n+m-2$ (не $n-m-2$). Если в конспекте S^2 обозначает объединённую *несмещённую* оценку дисперсии, то $\sqrt{S^2} = s_p$.

⚠ Когда нельзя применять эту формулу?

Если дисперсии различаются, предпочтителен **t-критерий Уэлча**, использующий $\sqrt{s_X^2/n + s_Y^2/m}$ и приближённое число степеней свободы Уэлча—Саттертуэйта.