

1. Винни-Пух поделил выборку на обучающую и тестовую, (y, X) и (y_B, X_B) . Предположим, что все наблюдения независимы, одна и та же зависимость выполнена на обучающей и тестовой выборке $y = X\beta + u$, $y_B = X_B\beta + u_B$. Кроме того, что $\mathbb{E}(u|X, X_B) = \mathbb{E}(u_B|X, X_B) = 0$, совместный закон распределения предикторов и ошибки одинаков для тестовой и обучающей выборке, в частности, $\text{Var}(u_i|x_i) = h(x_i)$, $\text{Var}(u_i^B|x_i^B) = h(x_i^B)$.

Для краткости обозначим $H = X(X^T X)^{-1} X^T$, оценки коэффициентов по обучающей выборке, $\hat{\beta}$, прогнозы на обучающую и тестовую выборку, $\hat{y} = X\hat{\beta}$, $\hat{y}_B = X_B\hat{\beta}$, ошибки прогнозов, $\hat{u} = y - \hat{y}$, $\hat{u}_B = y_B - \hat{y}_B$.

Найдите ковариационные матрицы $\text{Var}(\hat{y}|X, X_B)$, $\text{Cov}(\hat{y}_B, \hat{u}_B|X, X_B)$.

2. Рассмотрим регрессию с L^2 -регуляризацией, где штраф накладывается на все коэффициенты, кроме первого, который соответствует первому столбцу из единиц в матрице X размера $[n \times k]$. Целевая функция имеет вид:

$$Q(\hat{\beta}) = (y - X\hat{\beta})^T (y - X\hat{\beta}) + \lambda \sum_{j=2}^k \hat{\beta}_j^2.$$

Какие наблюдения надо добавить в обычную регрессию, чтобы результат обычной регрессии идеально совпал с регрессией с данной регуляризацией? Можно ли обойтись добавлением одного наблюдения?

3. Винни-Пух и Кролик пытаются оценить эффект воздействия бинарной переменной $w_i \in \{0, 1\}$ на целевую переменную y_i .

Винни-Пух использует множественную регрессию $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_w w_i + \hat{\beta}_x x_i$, классическую стандартную ошибку $se_{\text{class}}(\hat{\beta}_w)$ для построения доверительного интервала.

Кролик использует CUPED в следующей вариации. На первом шаге строит ровно ту же регрессию, что и Винни-Пух, получает почти-остатки $r_i = y_i - \hat{\beta}_x x_i$. На втором шаге оценивает парную регрессию $\hat{r}_i = \hat{\gamma}_1 + \hat{\gamma}_w w_i$ и также использует классическую стандартную ошибку $se_{\text{class}}(\hat{\gamma}_w)$ для построения доверительного интервала.

- Верно ли, что совпадают точечные оценки эффекта воздействия $\hat{\beta}_w$ и $\hat{\gamma}_w$?
- Во сколько раз отличаются классические стандартные ошибки $\hat{\beta}_w$ и $\hat{\gamma}_w$?

4. Рассмотрим модель линейной регрессии

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + \hat{\beta}_3 x_i^2$$

для набора данных https://github.com/bdemeshev/psmo_2022-23/raw/main/ha_02/psmo-ha_02.csv.

- а) Найдите классическую и *НСЗ*-оценку ковариационной матрицы коэффициентов.
- б) Постройте классический и *НСЗ*-робастный доверительный интервал для $\hat{\beta}_2$.
- в) Используя *НСЗ*-оценку ковариационной матрицы и предполагая, что оценки коэффициентов имеют распределение, похожее на многомерное нормальное, постройте с помощью генератора случайных чисел из нормального распределения 95%-й доверительный интервал для вершины параболы.

5. Рассмотрим модель логистической регрессии для набора данных

$$\Lambda(\mathbb{P}(\text{admit}_i = 1 | X)) = \beta_1 + \beta_2 gre_i + \beta_3 gpa_i + \beta_4 rank_i$$

для набора данных <https://stats.idre.ucla.edu/stat/data/binary.csv>. Переменная admit_i равна 1 для попавших на обучение, gre_i — результат GRE экзамена, gpa_i — средний балл, $rank_i$ — рейтинг студента.

- а) Найдите оценку ковариационной матрицы коэффициентов.
- б) Постройте 95%-й доверительный интервал для коэффициента при gpa .
- в) Постройте точечную и 95%-ю интервальную оценку предельного эффекта $\partial \mathbb{P}(y_i = 1 | X) / \partial gpa_i$ для медианных значений предикторов.
- г) С помощью подходящего статистического теста сделайте выбор между предложенной моделью и моделью, в которой переменная $rank$ считается категориальной, то есть вводится дополнительная бинарная переменная индикатор для каждого значения $rank$ кроме базового.