# Лекция 13.

## Нормализация регрессионного анализа

Пусть имеется уравнение общей линейной регрессии $\vec{X} = Z^T\vec{\beta} + \vec{\varepsilon}$, где $n$ - число экспериментов, $\vec{X}$ - столбец результатов экспериментов, $Z$ - матрица плана, $\vec{\beta}$ - столбец коэффициентов регрессии, $\vec{\varepsilon}$ - вектор теоретических ошибок

При этом ранее предполагали, что выполнены условия:

1. Cond. 1: Строки $Z$ - независимы
2. Cond. 2: $\varepsilon_i \in N(0, \sigma^2)$ и независимы

Условие 2 часто нарушается

### Взвешенный метод наименьших квадратов

Пусть $\varepsilon_i \in N(0, v_i\sigma^2)$ и независимы (то есть дисперсия ошибки зависит от номера наблюдения). Другими словами, $D\vec{\varepsilon} = \sigma^2 V$, где $V = \operatorname{diag}(v_1, \dots, v_n)$

Логично наблюдениям с меньшей дисперсии ошибки предать больший вес. Пусть вес $w_i = \dfrac{1}{v_i}$.

Домножим обе части уравнения регрессии на $\sqrt{w_i}$, тогда получим $\tilde{\vec{X}} = \tilde{Z}^T\vec{\beta} + \tilde{\vec{\varepsilon}}$, где $\tilde{x}_i = \sqrt{w_i}x_i$, $\tilde{Z}_i^{(j)} = \sqrt{w_i}Z_i^{(j)}$, $\tilde{\varepsilon}_i = \sqrt{w_i}\varepsilon_i$

$D\tilde{\vec{\varepsilon}} = D(\sqrt{w_i}\varepsilon_i) = w_i D\varepsilon_i = \dfrac{1}{v_i}v_i\sigma^2 = \sigma^2$ - получаем, что $D\tilde{\vec{\varepsilon}} = \sigma^2 E_n$, то есть стандартную ситуацию

Тогда оценки $\vec{b}$ будут несмещенными и эффективными

Недостаток у этого метода: нужно знать коэффициенты $v_i$

*Ex.* Рассмотрим модель линейной парной регрессии без свободного члена $X = \beta_0 Z + \varepsilon$

Теоретическое уравнение $A\vec{B} = Z\vec{X}$, где $Z = (Z_1, \dots, Z_n)$ - матрица плана, $\vec{B} = \hat{\beta}_0$, $A = ZZ^T = z_1^2 + \dots + z_n^2$, $Z\vec{X} = z_1x_1 + \dots + z_nx_n$

$\sum\limits_{i=1}^{n} z_i^2\hat{\beta}_0 = \sum\limits_{i=1}^{n} z_ix_i \Longrightarrow \hat{\beta}_0 = \dfrac{\sum z_ix_i}{\sum z_i^2}$ - оценка МНК

По взвешенному методу наименьших квадратов $\tilde{\beta}_0 = \dfrac{\sum w_iz_ix_i}{\sum w_iz_i^2}$ - оценка взвешенного МНК

*Ex. а.* Взвешенное среднее

Допустим, что проводим серию измерений «скоропортящимся» инструментом. При $Z \equiv 1$: $X = \beta_0 + \varepsilon$, $\varepsilon_i \in N(0, v_i\sigma^2)$, $w_i = \dfrac{1}{v_i}$

Тогда $\hat{\beta}_0 = \dfrac{\sum_{i=1}^{n} w_ix_i}{\sum_{i=1}^{n} w_i}$ - взвешенное среднее

*Ex. б.* Повторное наблюдение

Пусть было $n$ серий по $k_i$ наблюдений ($1 \le i \le n$). В каждой серии вычислили выборочное среднее $\overline{x_i}$. Если $\varepsilon \in N(0, \sigma^2)$, то дисперсия ошибки для каждого выборочного среднего $D\varepsilon_i = \dfrac{\sigma^2}{k_i}$.

Оценка по результатам всех наблюдений будет $\hat{\beta}_0 = \dfrac{\sum_{i=1}^n k_i \overline{x_i}}{\sum_{i=1}^n k_i}$

*Ex. в.* Пропорция

Пусть $X$ - потери тепла в квартире. Основной фактор $Z$ - разница температур снаружи и внутри. Так как при $Z = 0$ $X = 0$, то уравнением регрессии будет $X = \beta_0 Z + \varepsilon$

Логично предположить, что дисперсия ошибки зависит от величины $Z$. Рассмотрим две гипотезы:

1. Дисперсия ошибки прямо пропорциональна $Z$: $D\varepsilon_i = CZ_i = \sigma^2 \dfrac{CZ_i}{\sigma^2}$

   Тогда $w_i = \dfrac{\sigma^2}{CZ_i}$ и $\hat{\beta}_0 = \dfrac{\sum_{i=1}^n \frac{\sigma^2}{CZ_i} Z_i X_i}{\sum_{i=1}^n \frac{\sigma^2}{CZ_i} Z_i^2} = \dfrac{\sum X_i}{\sum Z_i} = \dfrac{\overline{x}}{\overline{z}}$

2. Дисперсия ошибки квадратично зависит от $Z$: $D\varepsilon_i = CZ_i^2 = \sigma^2 \dfrac{CZ_i^2}{\sigma^2}$

   Тогда $w_i = \dfrac{\sigma^2}{CZ_i^2}$ и $\hat{\beta}_0 = \dfrac{\sum_{i=1}^n \frac{\sigma^2}{CZ_i^2} Z_i X_i}{\sum_{i=1}^n \frac{\sigma^2}{CZ_i^2} Z_i^2} = \dfrac{\sum X_i}{\sum Z_i} = \dfrac{\sum \frac{X_i}{Z_i}}{n} = \overline{\left(\dfrac{x}{z}\right)}$

## Коррелированные наблюдения

Пусть ошибки не только имеют различные дисперсии, но и коррелированны между собой: $\operatorname{cov}(\varepsilon_i, \varepsilon_j) = v_{ij}$

Тогда $D\vec{\varepsilon} = \sigma^2 V$, где $V = (v_{ij})$

Так как матрица ковариаций симметричная и положительно определенная, то существует $\sqrt{V}$.

Домножим обе части уравнения регрессии на $\sqrt{V^{-1}}$:

$\vec{X} = Z^T \vec{\beta} + \vec{\varepsilon} \;\Big|\; \cdot \sqrt{V^{-1}}$

$\tilde{\vec{X}} = \tilde{Z}^T \vec{\beta} + \varepsilon$, где $\tilde{\vec{X}} = \sqrt{V^{-1}}\vec{X}$, $\tilde{Z} = \sqrt{V^{-1}} Z$, $\tilde{\vec{\varepsilon}} = \sqrt{V^{-1}}\vec{\varepsilon}$

Тогда матрица ковариаций нового вектора ошибок будет $D\tilde{\vec{\varepsilon}} = D(\sqrt{V^{-1}}\vec{\varepsilon}) = \sqrt{V^{-1}} D\vec{\varepsilon} \left(\sqrt{V^{-1}}\right)^T = \sigma^2 I_n$

То есть получили классическую ситуацию, когда выполнено Cond. 2 и вектор оценок $\hat{\beta}_0$ будет несмещенным и эффективным

## Составление матрицы плана при управляемом эксперименте

Если строки матрицы плана взять ортогональными, то дисперсии оценки коэффициентов $b_i$ регрессии будут минимальными. Поэтому лучше матрицу плана составлять таким образом:

Дисперсии оценок при этом $Db_i = \sigma^2 A_{ii}^{-1}$. Если $Z$ - ортогональная (не обязательно нормированная), то $A = ZZ^T = E_n$, а $Db_i = \dfrac{\sigma^2}{Z_i^2}$. Несложно доказать, что во всех других случаях дисперсия будет больше

## Метод главных осей

Помимо метода наименьших квадратов существует метод «главных осей». Идея следующая: матрицу ковариаций приводим к диагональной форме
В МНК мы минимизируем расстояние отрезков, параллельных оси Oy, а в методе главных осей - перпендикуляр от точки до возможной прямой
Результатом метода главных осей получаем прямую, являющуюся главной осью эллипса, появляющегося из корреляционного облака

## Нелинейные регрессии

Помимо общего МНК многие нелинейные зависимости могут быть сведены к линейным при помощи простых приемов

*Ex. а.* $X = \alpha + \beta f(Z) + \varepsilon$, где $f(x)$ - известная функция
Тогда можно взять новый фактор $Z' = f(Z)$, свели задачу к стандартной, получаем уравнение $X = \alpha + \beta Z' + \varepsilon$
Пример: $X = \alpha + \beta \ln Z + \varepsilon$, то $Z' = \ln Z$

*Ex. б.* $X = \alpha Z^\beta \varepsilon$
Логарифмируем: $\ln X = \ln \alpha + \beta \ln Z + \ln \varepsilon \Longleftrightarrow X' = \alpha' + \beta Z' + \varepsilon'$

*Ex. в.* $X = \alpha e^{\beta Z} \varepsilon$
Логарифмируем: $\ln X = \ln \alpha + \beta Z + \ln \varepsilon \Longleftrightarrow X' = \alpha' + \beta Z + \varepsilon'$

*Ex. г.* Зависимость в виде полинома: $X = \beta_0 + \beta_1 Z + \beta_2 Z^2 + \cdots + \beta_k Z^k$
Введем новые факторы $Z_1 = Z, Z_2 = Z^2, \ldots, Z_k = Z^k$
$X = \beta_0 + \beta_1 Z + \beta_2 Z_2 + \cdots + \beta_k Z_k$
При этом, чтобы избежать мультиколлинеарность, лучше брать $k < 4$. При больших $k$ получить многочлен большой степени, который сможет гарантировано пройти через все точки - это будет статистически незначимо

*Nota.* Если из теории мы знаем вид зависимости и подбираем ее под данные, то желательно строить модель как можно проще
*Nota.* Из построенных моделей предпочтительней та, где коэффициент детерминации больше

Построение даже удачной регрессионной модели не означает появление причинно-следственной связи. Исторический пример: исследовалась точность бомбометания от различных условий. Пусть $X$ - точность, $Z_1$ - высота, $Z_2$ - ветер, $Z_3$ - количество истребителей противника
Построили модель $X = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3$ и получили $\hat{\beta}_3 < 0$ - то есть при большем числе техники противника точность увеличивалась. Оказалось, что не был учтен фактор облачности - при нем $\hat{\beta}_3 > 0$ и коэффициент детерминации улучшился
Или другой пример: корреляция численности аистов и рождения детей в Голландии XX века оказалась прямой