$f_1(x), \dots, f_n(x)$ — числовые признаки;
Модель многомерной линейной регрессии:

$$f(x, \alpha) = \sum_{j=1}^{n} \alpha_j f_j(x), \qquad \alpha \in \mathbb{R}^n.$$

Матричные обозначения:

$$\underset{\ell \times n}{F} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}, \quad \underset{\ell \times 1}{y} = \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix}, \quad \underset{n \times 1}{\alpha} = \begin{pmatrix} \alpha_1 \\ \dots \\ \alpha_n \end{pmatrix}.$$

Функционал квадрата ошибки:

$$Q(\alpha, X^\ell) = \sum_{i=1}^{\ell} \bigl(f(x_i, \alpha) - y_i\bigr)^2 = \|F\alpha - y\|^2 \to \min_{\alpha}.$$

Необходимое условие минимума в матричном виде:

$$\frac{\partial Q}{\partial \alpha}(\alpha) = 2F^{\mathsf{т}}(F\alpha - y) = 0,$$

откуда следует *нормальная система* задачи МНК:

$$F^{\mathsf{т}}F\alpha = F^{\mathsf{т}}y,$$

где $F^{\mathsf{т}}F$ — $n \times n$-матрица.

**Решение системы:** $\alpha^* = (F^{\mathsf{т}}F)^{-1}F^{\mathsf{т}}y = F^{+}y$.

Значение функционала: $Q(\alpha^*) = \|P_F y - y\|^2$,
где $P_F = FF^{+} = F(F^{\mathsf{т}}F)^{-1}F^{\mathsf{т}}$ — *проекционная матрица*.

Произвольная $\ell\times n$-матрица представима в виде *сингулярного разложения* (singular value decomposition, SVD):

$$F = VDU^{\mathsf{T}}.$$

**Основные свойства** сингулярного разложения:

1. $\ell\times n$-матрица $V = (v_1, \dots, v_n)$ ортогональна, $V^{\mathsf{T}}V = I_n$, столбцы $v_j$ — собственные векторы матрицы $FF^{\mathsf{T}}$;
2. $n\times n$-матрица $U = (u_1, \dots, u_n)$ ортогональна, $U^{\mathsf{T}}U = I_n$, столбцы $u_j$ — собственные векторы матрицы $F^{\mathsf{T}}F$;
3. $n\times n$-матрица $D$ диагональна, $D = \mathrm{diag}\bigl(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n}\bigr)$, $\lambda_j \geqslant 0$ — собственные значения матриц $F^{\mathsf{T}}F$ и $FF^{\mathsf{T}}$, $\sqrt{\lambda_j}$ — сингулярные числа матрицы $F$.

Псевдообратная $F^+$, вектор МНК-решения $\alpha^*$, МНК-аппроксимация целевого вектора $F\alpha^*$:

$$F^+ = (UDV^\intercal VDU^\intercal)^{-1}UDV^\intercal = UD^{-1}V^\intercal = \sum_{j=1}^{n} \frac{1}{\sqrt{\lambda_j}} u_j v_j^\intercal;$$

$$\alpha^* = F^+ y = UD^{-1}V^\intercal y = \sum_{j=1}^{n} \frac{1}{\sqrt{\lambda_j}} u_j (v_j^\intercal y);$$

$$F\alpha^* = P_F y = (VDU^\intercal)UD^{-1}V^\intercal y = VV^\intercal y = \sum_{j=1}^{n} v_j (v_j^\intercal y);$$

$$\|\alpha^*\|^2 = \|D^{-1}V^\intercal y\|^2 = \sum_{j=1}^{n} \frac{1}{\lambda_j} (v_j^\intercal y)^2.$$

**Проблема:** мультиколлинеарность при $\lambda_j \to 0$.

# Проблема мультиколлинеарности и переобучения

Если имеются сингулярные числа, близкие к нулю, то:

- матрица $\Sigma = F^{\intercal} F$ плохо обусловлена;
- решение становится неустойчивым и неинтерпретируемым, слишком большие коэффициенты $\|\alpha_j^*\|$ разных знаков;
- возникает переобучение:
  на обучении $Q(\alpha^*, X^\ell) = \|F\alpha^* - y\|^2$ малó;
  на контроле $Q(\alpha^*, X^k) = \|F'\alpha^* - y'\|^2$ велико;

Стратегии устранения мультиколлинеарности и переобучения:

- отбор признаков: $f_1, \ldots, f_n \to f_{j_1}, \ldots, f_{j_m}, \quad m \ll n$.
- регуляризация: $\|\alpha\| \to \min$;
- преобразование признаков: $f_1, \ldots, f_n \to g_1, \ldots, g_m, \quad m \ll n$;

# Резюме

- Задача многомерной линейной регрессии может быть решена через сингулярное разложение
- Мультиколлинеарность приводит к плохой обусловленности, неустойчивости и переобучению
- Методы устранения мультиколлинеарности (гребневая регрессия, метод главных компонент) также связаны с сингулярным разложением (об этом в следующих лекциях)