Семинар - 9.

Линейная регрессия. Доверительные интервалы

Гауссовская линейная регрессия.

$$y = \sum_{j=1}^{p} \psi_j(x)\theta_j + \varepsilon_i = \boldsymbol{\psi}(x)^\intercal \boldsymbol{\theta} + \varepsilon_i, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Пусть имеются данные $\{(X_i, Y_i) : 1 \le i \le n\}$, причем наблюдения $Y_1, \dots, Y_n \in \mathbb{R}$ получены из модели

$$\mathbf{Y} = \Psi^\intercal \boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n),$$

где $\theta \in \mathbb{R}^p$, $\Psi = (\boldsymbol{\psi}(X_1), \dots, \boldsymbol{\psi}(X_n)) \in \mathbb{R}^{p \times n}$– *матрица плана* полного ранга, то есть $\operatorname{rank}(\Psi) = p$, $X_1, \dots, X_n$ детерминированные, а $\varepsilon_1, \dots, \varepsilon_n$ – независимые центрированные случайные величины с дисперсией σ^2. Мы предполагаем, что матрица $\Psi \in \mathbb{R}^{p \times n}$ имеет полный ранг,

Оценка в методе наименьших квадратов (ОМНК):

$$\widehat{\boldsymbol{\theta}}_{LSE} = \arg\min_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^{n} (Y_i - \boldsymbol{\psi}(X_i)^\intercal \boldsymbol{\theta})^2 = \arg\min_{\boldsymbol{\theta} \in \Theta} \|\mathbf{Y} - \Psi^\intercal \boldsymbol{\theta}\|^2 = (\Psi\Psi^\intercal)^{-1}\Psi\mathbf{Y}.$$

Введем ортопроектор на линейное подпространство размерности p, натянутое на строки матрицы Ψ: $(\psi_j(X_1), \dots, \psi_j(X_n))^\intercal \in \mathbb{R}^n$, $1 \le j \le n$

$$\Pi = \Psi^\intercal(\Psi\Psi^\intercal)^{-1}\Psi \in \mathbb{R}^{n \times n}.$$

Тогда оценка регрессионной функции $\boldsymbol{\psi}(x)^\intercal \boldsymbol{\theta}$ в точках $X_1, \dots, X_n$:

$$\widehat{f} = \Psi^\intercal \widehat{\boldsymbol{\theta}}_{LSE} \equiv \Pi\mathbf{Y}.$$

Утверждение 1 (Геометрическая интерпретация $\widehat{f}$)**.** *$\widehat{f}$ является проекцией вектора наблюдений на линейную оболочку строк матрицы плана Ψ, то есть на* $\operatorname{Im}(\Psi^\intercal)$.

Лемма (Свойство гауссовских векторов)**.** *Пусть вектор X имеет нормальное распределение $\mathcal{N}(\mu, \Sigma_n)$. Тогда для произвольной матрицы $B \in \mathbb{R}^{m \times n}$ вектор BX имеет нормальное распределение $\mathcal{N}(B\mu, B\Sigma_n B^\intercal)$.*

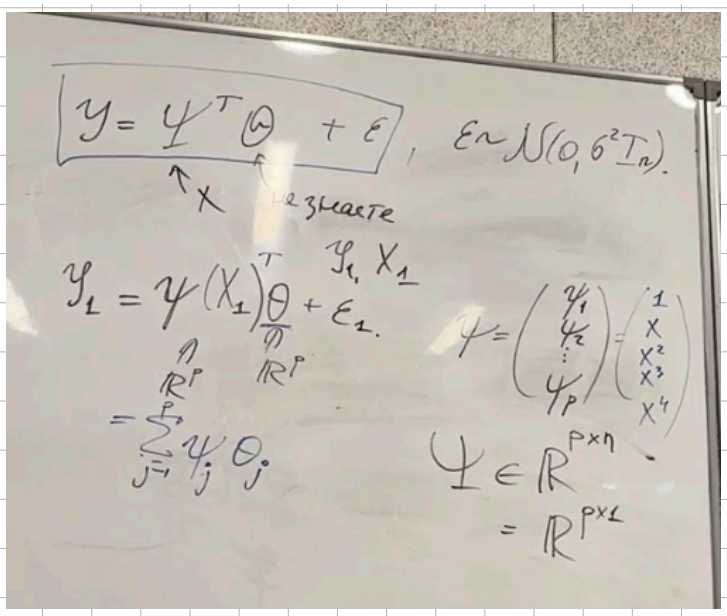
Лемма. *Пусть вектор X имеет стандартное нормальное распределение $\mathcal{N}(0, I_n)$ и пусть B – ортопроектор. Тогда $Q = X^\intercal BX$ имеет распределение хи-квадрат $\chi^2(\operatorname{tr} B)$.*

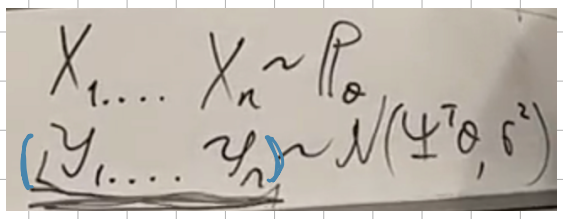
Задачи

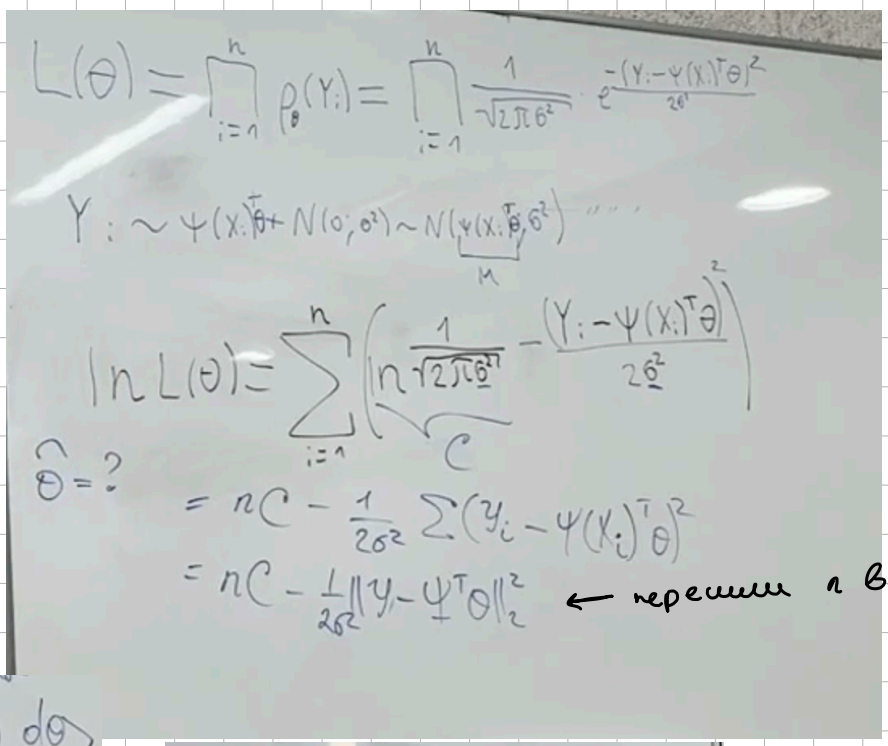
1. Найти оценку неизвестного параметра $\boldsymbol{\theta}$ методом максимального правдоподобия. Будет ли она несмещенной? Будет ли она оптимальна? Сравнить с оценкой наименьших квадратов в модели линейной регрессии (то есть когда вектор ошибок не обязательно гауссовский).
2. Найти оценку максимального правдоподобия для σ^2 в гауссовской линейной регрессии. Будет ли она несмещенной?
3. Доказать, что $\frac{1}{\sigma^2}\|\Pi\boldsymbol{\varepsilon}\|^2$ имеет распределение $\chi^2(p)$. - не успели
4. Доказать, что $\widehat{\boldsymbol{\theta}}_{LSE}$ и $\mathbf{Y} - \Psi^\intercal \widehat{\boldsymbol{\theta}}_{LSE}$ независимы. Доказать, что $\frac{1}{\sigma^2}\|\mathbf{Y} - \Psi^\intercal \widehat{\boldsymbol{\theta}}\|^2$ имеет распределение хи-квадрат $\chi^2(n-p)$. – не успели.

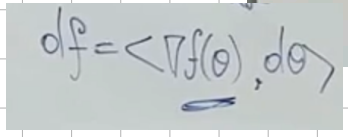
5. Постройте точный $1-\alpha$-доверительный интервал для $\boldsymbol{\theta}_j$, где $j = \{1, \dots, p\}$, если
   (a) σ^2 известен;
   (b) σ^2 неизвестена.
6. Постройте точное $1-\alpha$-доверительное множество для $\boldsymbol{\theta}$, если σ^2 неизвестена.
7. Постройте точный $1-\alpha$-доверительный интервал для σ^2.

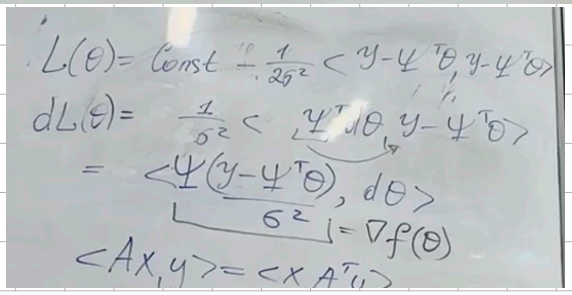
1. Найти оценку неизвестного параметра $\boldsymbol{\theta}$ методом максимального правдоподобия. Будет ли она несмещенной? Будет ли она оптимальна? Сравнить с оценкой наименьших квадратов в модели линейной регрессии (то есть когда вектор ошибок не обязательно гауссовский).

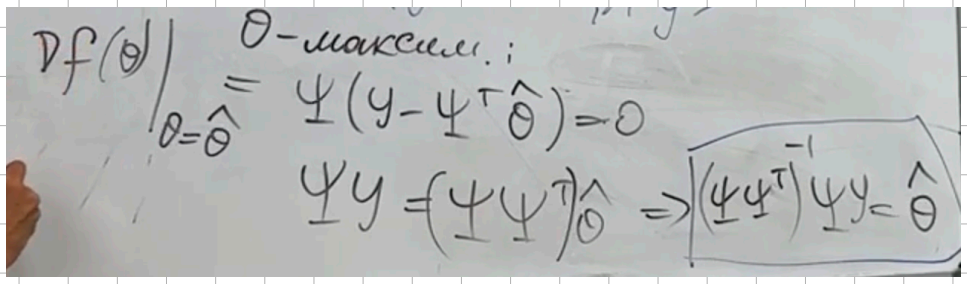




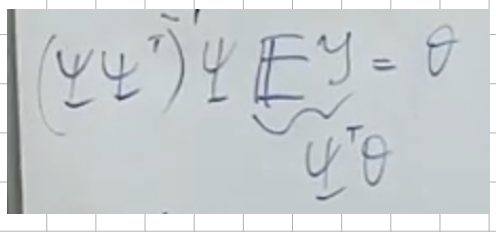






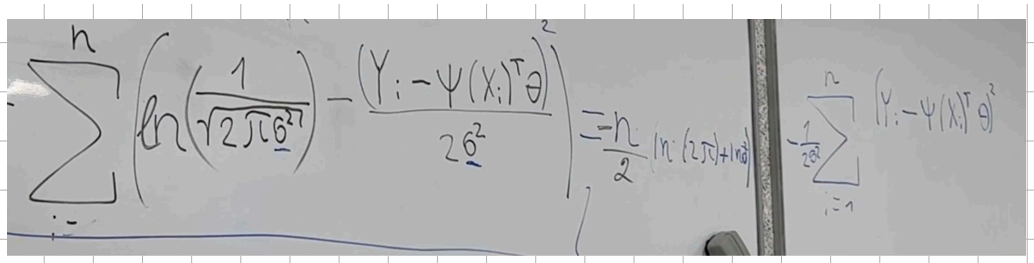


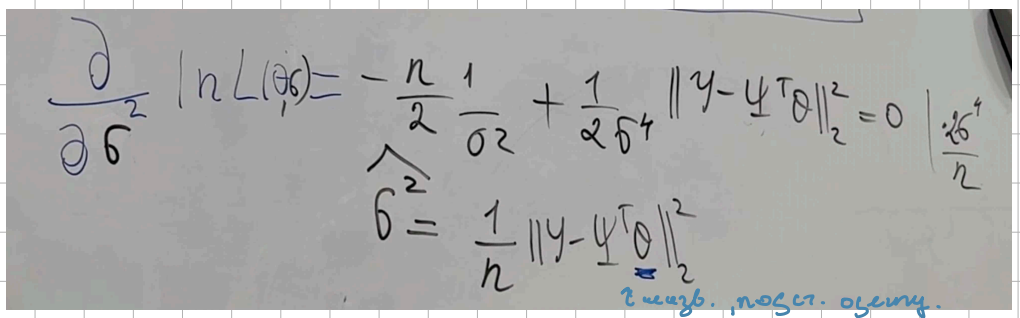
Несмещенность:

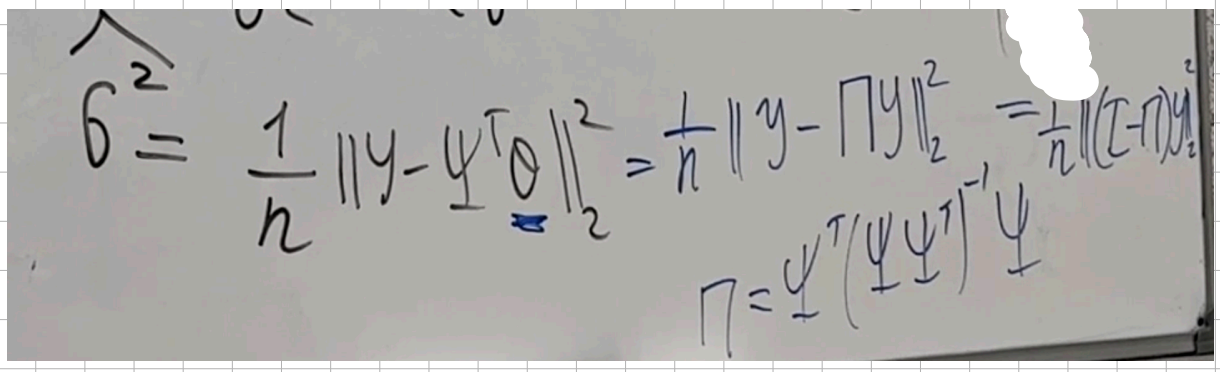


2. Найти оценку максимального правдоподобия для σ^2 в гауссовской линейной регрессии. Будет ли она несмещенной?

$\ln L(\theta)=$







5. Постройте точный $1-\alpha$-доверительный интервал для $\boldsymbol{\theta}_j$, где $j=\{1,\dots,p\}$, если

   (a) σ^2 известен;

   (b) σ^2 неизвестена.