Statistical Optimization: Lecture 2

Optimization for Statistics

Zijian Guo

Zhejiang University
Center for data science

March 17, 2026

Outline

Optimization for Statistics: Mean Problem

- Maximum Likelihood
- Uncertainty Quantification

Supervised and Unsupervised Learning

Least Squares Principle

- Ordinary Least Squares
- Principal Component Analysis

Risk Function and Optimization

- Negative Likelihood
- Risk Optimization: Semiparametric View

Statistical Inference

- The goal of statistical inference is to use the observed sample data $X_1, X_2, \cdots, X_n$ to **draw inferences** about the true population distribution or the unknown parameters.
- Since the underlying distribution that generates the data is often unknown, we use statistical methodologies to make inference about the true distribution P_{θ^*}, where θ^* denotes the true parameter(s).
- Therefore, the three main goals of statistical inference are:
  - Point Estimation
  - Confidence Interval
  - Hypothesis Testing (A dual problem of Confidence Interval, not covered here.)

Point Estimation

- Point estimation is the method of finding a single-value estimate of the true parameter θ^* based on the observed sample.
- In other words, a point estimator is a function of the random sample $\{X_1, \cdots, X_n\}$ that produces a best-guess value for the unobserved population parameter.
- Mathematically speaking, we are looking for $\widehat{\theta}(X_1, \cdots, X_n)$ such that

$$|\widehat{\theta}(X_1, \cdots, X_n) - \theta^*| \quad \text{is as small as possible.}$$

Outline

Optimization for Statistics: Mean Problem

- Maximum Likelihood
- Uncertainty Quantification

Supervised and Unsupervised Learning

Least Squares Principle

- Ordinary Least Squares
- Principal Component Analysis

Risk Function and Optimization

- Negative Likelihood
- Risk Optimization: Semiparametric View

Maximum Likelihood

Assume we have a probabilistic model $p_\theta(x)$ with θ denoting the unknown parameter. Given data $\{x_i\}_{1 \le i \le n}$, we define the likelihood function as

$$L(\theta) = \prod_{i=1}^{n} p_\theta(x_i).$$

MLE chooses the parameter that makes the observed data most likely:

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_\theta L(\theta).$$

Usually we maximize the log-likelihood instead:

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_\theta \ell_n(\theta), \qquad \ell_n(\theta) = \sum_{i=1}^{n} \log p_\theta(x_i).$$

Bernoulli (Success Probability Estimation)

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p^*)$, i.e.,

$$\mathbb{P}(X_i = 1) = p^*, \qquad \mathbb{P}(X_i = 0) = 1 - p^*.$$

Likelihood.

$$L(p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}.$$

Log-likelihood.

$$\ell_n(p) = \log L(p) = \sum_{i=1}^{n} \Big(X_i \log p + (1 - X_i) \log(1-p) \Big).$$

MLE. Differentiate and set to zero:

$$\frac{d}{dp}\ell_n(p) = \sum_{i=1}^{n} \left(\frac{X_i}{p} - \frac{1 - X_i}{1-p} \right) = 0 \quad \Longrightarrow \quad \hat{p}_{\text{MLE}} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Gaussian Mean Estimation

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu^*, \sigma^2)$ with known σ^2.

Likelihood.

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right).$$

Log-likelihood.

$$\ell_n(\mu) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 = C - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2.$$

MLE.

$$\frac{d}{d\mu}\ell_n(\mu) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu) = 0 \quad \Longrightarrow \quad \hat{\mu}_{\text{MLE}} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Least Squares

If we consider $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu^*, \sigma^2)$, then MLE leads to the Least Squares estimator

$$\hat{\mu}_{\text{MLE}} = \arg\min_{\mu} \sum_{i=1}^{n} (X_i - \mu)^2. \tag{1}$$

If we consider

$$Y_i = X_i\beta + \epsilon_i, \qquad X_i \perp \epsilon_i, \quad \text{and} \quad \epsilon_i \sim N(0, \sigma^2),$$

then we generalize the Least Squares principle and define

$$\widehat{\beta}^{\text{LS}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2.$$

Law of Large Numbers (LLN)

Idea. If we repeat the same random experiment many times, the *sample average* stabilizes near the true mean.

Let $X_1, X_2, \ldots$ be i.i.d. with $E[X_1] = \mu$ and $\mathrm{Var}(X_1) < \infty$. Define

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then, for any $\varepsilon > 0$,

$$\mathbb{P}\big(|\bar{X}_n - \mu| > \varepsilon\big) \to 0 \qquad \text{as } n \to \infty.$$

Interpretation. For large n, it is *unlikely* that $\bar{X}_n$ differs from μ by more than a small margin ε. (So $\bar{X}_n$ is a reliable estimator of μ when n is large.)

Outline

Optimization for Statistics: Mean Problem

- Maximum Likelihood
- Uncertainty Quantification

Supervised and Unsupervised Learning

Least Squares Principle

- Ordinary Least Squares
- Principal Component Analysis

Risk Function and Optimization

- Negative Likelihood
- Risk Optimization: Semiparametric View

Uncertainty Quantification: Confidence Interval

- We construct a confidence interval (CI) from the observed sample, which gives a set of plausible values for the true parameter θ^*.
- Suppose $X_1, X_2, \cdots, X_n$ are iid samples from P_{θ^*}. For a confidence level of $1-\alpha$ (e.g., $\alpha = 0.05$), a $(1-\alpha)$ confidence interval $CI(X_1, \cdots, X_n)$ is defined to satisfy

$$\mathbb{P}_{\theta^*}\left(\theta^* \in CI(X_1, X_2, \cdots, X_n)\right) = 1-\alpha.$$

  Thus, the probability that the confidence interval contains the true parameter θ^* is $1-\alpha$.
- In words, if we were to repeat the data collection many times and construct a CI each time using the same procedure, then about a fraction $1-\alpha$ of those intervals would contain the true value θ^*. Since the CI is constructed from random variables $X_1, \cdots, X_n$, the CI itself is random.

Central Limit Theorem (CLT)

Idea. The LLN says the average gets close to μ; the CLT tells us *how it fluctuates around* μ for large n.

Let $X_1, \ldots, X_n$ be i.i.d. with $E[X_1] = \mu$ and $\mathrm{Var}(X_1) = \sigma^2 \in (0, \infty)$. Then

$$\frac{\sqrt{n}\,(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0,1), \qquad \text{as } n \to \infty,$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Interpretation. For large n,

$$\bar{X}_n \approx \mathcal{N}\Big(\mu^*,\ \frac{\sigma^2}{n}\Big),$$

so the typical error size of $\bar{X}_n$ is about $\sigma/\sqrt{n}$.

Practical takeaway. This approximation is the main reason we can build confidence intervals, e.g.

$$\bar{X}_n \pm 1.96\,\frac{\sigma}{\sqrt{n}} \quad \text{(approximately 95\% confidence, if } \sigma \text{ is known).}$$

Confidence Interval for Mean

Let $X_1, \ldots, X_n$ be i.i.d. with $E[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. The sample mean is

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Known variance. If σ is known and $X_i \sim \mathcal{N}(\mu^*, \sigma^2)$, then

$$\frac{\bar{X} - \mu^*}{\sigma / \sqrt{n}} \sim \mathcal{N}(0, 1).$$

A $(1 - \alpha)$ confidence interval for μ^* is

$$\bar{X} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}},$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ quantile of $\mathcal{N}(0, 1)$.

Confidence Interval for Mean

Unknown variance. If $X_i \sim \mathcal{N}(\mu^*, \sigma^2)$ with unknown σ^2, let

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

Then

$$\frac{\bar{X} - \mu^*}{S/\sqrt{n}} \sim t_{n-1}.$$

A $(1-\alpha)$ confidence interval for μ^* is

$$\bar{X} \pm t_{n-1,\,1-\alpha/2} \frac{S}{\sqrt{n}},$$

where $t_{n-1,\,1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of t_{n-1}.

Finite Sample Confidence Interval

We use $b_{n,1/2,\alpha/2}$ and $b_{n,1/2,1-\alpha/2}$ to denote the $\alpha/2$ and $1-\alpha/2$ quantiles of $\mathrm{Bin}(n, 1/2)$ Random Variable.
Then we have

$$\mathbb{P}\left(b_{n,1/2,\alpha/2} \leq \sum_{i=1}^{n} \mathbf{1}(X_i \leq \mu) \leq b_{n,1/2,1-\alpha/2}\right) = 1-\alpha.$$

A valid confidence interval for μ^* is

$$\left\{\mu : b_{n,1/2,\alpha/2} \leq \sum_{i=1}^{n} \mathbf{1}(X_i \leq \mu) \leq b_{n,1/2,1-\alpha/2}\right\}. \tag{2}$$

Outline

Optimization for Statistics: Mean Problem

- Maximum Likelihood
- Uncertainty Quantification

Supervised and Unsupervised Learning

Least Squares Principle

- Ordinary Least Squares
- Principal Component Analysis

Risk Function and Optimization

- Negative Likelihood
- Risk Optimization: Semiparametric View

Supervised and Unsupervised Learning

Supervised learning (with labels). We observe *labeled* data: each input X_i comes with an output/label Y_i. The goal is to learn a rule that predicts Y from X.

Unsupervised learning (no labels). We observe *unlabeled* data: only inputs X_i are available. The goal is to discover structure in the data distribution, such as low-dimensional representations, clusters, or patterns.

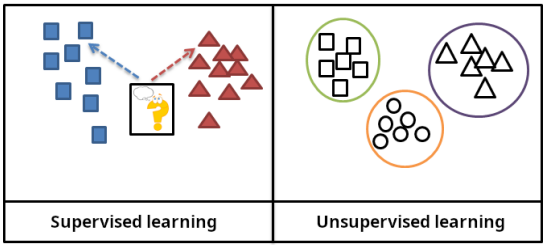


Figure: In supervised learning, the training data is labeled with the expected answers, while in unsupervised learning, the model identifies patterns or structures in unlabeled data.

Regression and Classification

Prediction: use observed features (inputs) to guess an unknown outcome (output) for a new case.

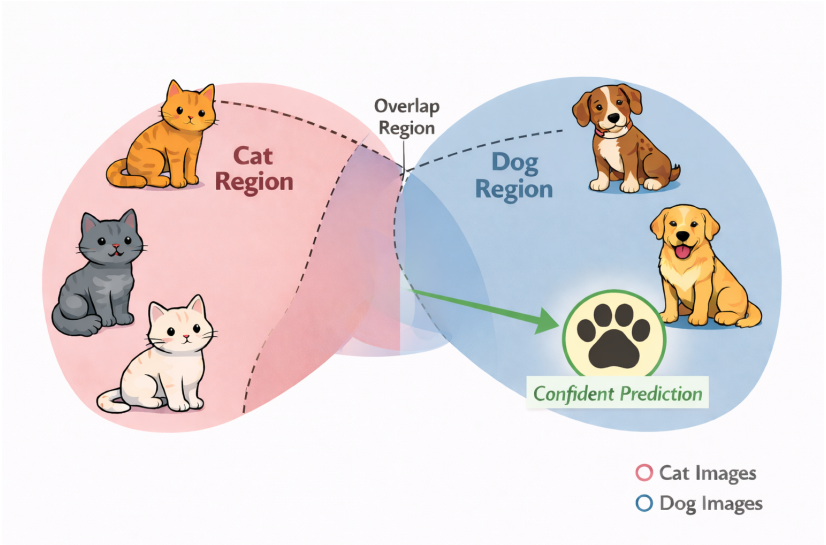
Regression (continuous output): predict a *number*.

- Outputs can take any value on a scale.
- Examples: temperature, height/weight.

Classification (discrete output): predict a *label/category*.

- Outputs must be chosen from a finite set.
- Examples: cat vs dog, digit 0–9.

Classification Example



Outline

Optimization for Statistics: Mean Problem

- Maximum Likelihood
- Uncertainty Quantification

Supervised and Unsupervised Learning

Least Squares Principle

- Ordinary Least Squares
- Principal Component Analysis

Risk Function and Optimization

- Negative Likelihood
- Risk Optimization: Semiparametric View

Math Review: Vectors

For $x, y \in \mathbb{R}^d$, the inner product and Euclidean norm are

$$\langle x, y \rangle = x^\top y = \sum_{i=1}^{d} x_i y_i, \qquad \|x\|_2 = \sqrt{x^\top x}.$$

Orthogonality. $x \perp y$ if $x^\top y = 0$.

Common norms.

$$\|x\|_1 = \sum_{i=1}^{d} |x_i|, \qquad \|x\|_2 = \Big(\sum_{i=1}^{d} x_i^2 \Big)^{1/2}, \qquad \|x\|_\infty = \max_{1 \le i \le d} |x_i|.$$

Math Review: Matrices

A matrix $A \in \mathbb{R}^{m \times n}$ is a rectangular array; multiplying A by $x \in \mathbb{R}^n$ gives $Ax \in \mathbb{R}^m$.

Matrix multiplication. If $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times r}$,

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}.$$

Transpose. $(A^\top)_{ij} = A_{ji}$ and $(AB)^\top = B^\top A^\top$.

Identity and inverse. The identity matrix satisfies $I_n x = x$. If A is invertible, then there exists A^{-1} such that $AA^{-1} = A^{-1}A = I_n$, and the linear system $Ax = b$ has the unique solution $x = A^{-1}b$.

Symmetry. A is symmetric if $A = A^\top$.

Outline

Optimization for Statistics: Mean Problem

- Maximum Likelihood
- Uncertainty Quantification

Supervised and Unsupervised Learning

Least Squares Principle

- Ordinary Least Squares
- Principal Component Analysis

Risk Function and Optimization

- Negative Likelihood
- Risk Optimization: Semiparametric View

Least Squares Regression

Assume a linear regression model

$$Y_i = X_i^\top \beta + \varepsilon_i, \qquad i = 1, \dots, n,$$

where $X_i \in \mathbb{R}^d$ are covariates and $\beta \in \mathbb{R}^d$ is unknown.

Least squares estimator.

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^d} \sum_{i=1}^{n} (Y_i - X_i^\top \beta)^2.$$

Let $X \in \mathbb{R}^{n \times d}$ be the design matrix with rows $X_i^\top$, and $Y = (Y_1, \dots, Y_n)^\top$. Then the objective is

$$\sum_{i=1}^{n} (Y_i - X_i^\top \beta)^2 = \|Y - X\beta\|_2^2 = (Y - X\beta)^\top (Y - X\beta).$$

Least Squares Regression: Closed-form Solution

Consider

$$Q(\beta) = \|Y - X\beta\|_2^2 = (Y - X\beta)^\top (Y - X\beta).$$

Expand:

$$Q(\beta) = Y^\top Y - 2\beta^\top X^\top Y + \beta^\top X^\top X \beta.$$

Differentiate with respect to β and set to zero:

$$\nabla_\beta Q(\beta) = -2X^\top Y + 2X^\top X \beta = 0 \quad \Longrightarrow \quad X^\top X \hat{\beta} = X^\top Y.$$

These are the **normal equations**.

If $X^\top X$ is invertible, then

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y.$$

Outline

Principal Component Analysis (PCA): Reconstruction Error

Goal. The central goal of PCA is to identify a low-rank linear subspace that captures the dominant variation in the data. We present two equivalent formulations: minimizing reconstruction error and maximizing explained variance.

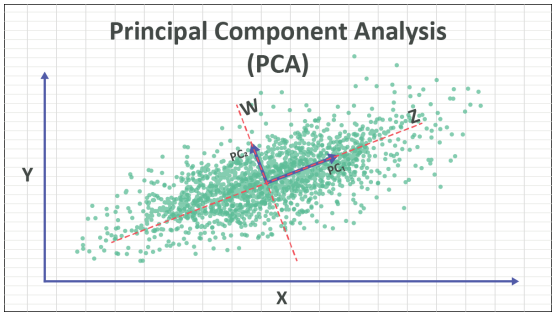


Figure: PCA

Principal Component Analysis (PCA)

Goal. PCA finds a k-dimensional linear subspace ($k < d$) so that data points are *close* to this subspace.

Projection. Choose a rank-k projection matrix P (this represents a k-dimensional subspace). For any $X \in \mathbb{R}^d$, the projected point is

$$PX \quad \text{(the closest point to } X \text{ on the subspace)}.$$

Projection distance. The distance from X to the subspace is

$$\|X - PX\|_2.$$

So $\|X - PX\|_2^2$ is the squared distance to the subspace.

Least squares principle. PCA chooses P to minimize the average squared distance:

$$P^* \in \arg\min_{P \in \mathcal{P}^k} \|X - PX\|_2^2.$$

Outline

Optimization for Statistics: Mean Problem

- Maximum Likelihood
- Uncertainty Quantification

Supervised and Unsupervised Learning

Least Squares Principle

- Ordinary Least Squares
- Principal Component Analysis

Risk Function and Optimization

- Negative Likelihood
- Risk Optimization: Semiparametric View

Outline

Optimization for Statistics: Mean Problem

- Maximum Likelihood
- Uncertainty Quantification

Supervised and Unsupervised Learning

Least Squares Principle

- Ordinary Least Squares
- Principal Component Analysis

Risk Function and Optimization

- Negative Likelihood
- Risk Optimization: Semiparametric View

Negative Conditional Likelihood

1. Modeling choice: specify a conditional model for Y given X,

$$p_\theta(y \mid x), \qquad \theta \in \Theta,$$

   while leaving the marginal distribution of X unspecified.
2. Conditional likelihood: given data $(X_i, Y_i)_{i=1}^n$,

$$L_c(\theta) = \prod_{i=1}^{n} p_\theta(Y_i \mid X_i).$$

3. Conditional MLE:

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta \in \Theta} \log L_c(\theta) = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell_\theta(Y_i, X_i),$$

   where the per-sample loss is the negative log-likelihood

$$\ell_\theta(y, x) := -\log p_\theta(y \mid x).$$

Least Squares Loss

Assume Gaussian conditional model:

$$Y_i \mid X_i \sim \mathcal{N}(X_i^\mathsf{T}\beta, \sigma^2).$$

Thus the conditional density function is

$$p_\beta(Y_i \mid X_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - X_i^\mathsf{T}\beta)^2}{2\sigma^2}\right).$$

Per-sample negative log-likelihood:

$$\ell_\beta(Y_i, X_i) := -\log p_\beta(Y_i \mid X_i) = \frac{(Y_i - X_i^\mathsf{T}\beta)^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2).$$

Conditional MLE:

$$\hat{\beta}_{\mathrm{MLE}} = \arg\min_\beta \sum_{i=1}^n \ell_\beta(Y_i, X_i) = \arg\min_\beta \sum_{i=1}^n (Y_i - X_i^\mathsf{T}\beta)^2.$$

Logistic Loss

Assume the logistic model:

$$Y_i \in \{0,1\}, \qquad \mathbb{P}(Y_i = 1 \mid X_i) = \sigma(X_i^\mathsf{T}\beta), \qquad \sigma(t) = \frac{1}{1+e^{-t}}.$$

Thus the conditional density function is

$$p_\beta(Y_i \mid X_i) = \sigma(X_i^\mathsf{T}\beta)^{Y_i}\Big(1-\sigma(X_i^\mathsf{T}\beta)\Big)^{1-Y_i}.$$

Per-sample negative log-likelihood:

$$\ell_\beta(Y_i, X_i) := -\log p_\beta(Y_i \mid X_i) = -\Big[Y_i \log\sigma(X_i^\mathsf{T}\beta) + (1-Y_i)\log\big(1-\sigma(X_i^\mathsf{T}\beta)\big)\Big].$$

Conditional MLE:

$$\hat\beta_{\mathrm{MLE}} = \arg\min_\beta \sum_{i=1}^n \ell_\beta(Y_i, X_i) = \arg\min_\beta \sum_{i=1}^n \Big[\log\big(1+e^{X_i^\mathsf{T}\beta}\big) - Y_i X_i^\mathsf{T}\beta\Big].$$

Cross-Entropy Loss

$$Y_i \in \{1, \dots, K\}, \qquad \mathbb{P}_\beta(Y_i = k \mid X_i) =: p_{ik}(\beta) = \frac{\exp(X_i^\top \beta_k)}{\sum_{j=1}^{K} \exp(X_i^\top \beta_j)},$$

where $\beta = (\beta_1, \dots, \beta_K)$ with $\beta_k \in \mathbb{R}^d$.

Per-sample negative log-likelihood (cross-entropy): with one-hot labels $Y_{ik} = \mathbf{1}\{Y_i = k\}$,

$$\ell_\beta(Y_i, X_i) = -\sum_{k=1}^{K} Y_{ik} \log p_{ik}(\beta).$$

Conditional MLE:

$$\hat{\beta}_{\mathrm{MLE}} = \arg\min_\beta \sum_{i=1}^{n} \ell_\beta(Y_i, X_i) = \arg\min_\beta \sum_{i=1}^{n} \left[\log\Big(\sum_{j=1}^{K} e^{X_i^\top \beta_j} \Big) - \sum_{k=1}^{K} Y_{ik} X_i^\top \beta_k \right].$$

Outline

Risk Function View

We consider the following general loss function (e.g., the negative log-likelihood)

$$\ell_\theta(y, x).$$

We shall define the corresponding risk function

$$r_\theta(y, x) = \mathbb{E}_{Y_i, X_i}\left[\ell_\theta(Y_i, X_i)\right].$$

Then we define

$$\theta^* = \arg\min_\theta \mathbb{E}_{Y_i, X_i}\left[\ell_\theta(Y_i, X_i)\right]. \tag{3}$$

Semi-parametric view: no model assumption on the generative distribution of Y_i and X_i.

Two Interpretation of Least Squares

1. Classical (well-specified model) view: the model parameters in a well-specified probability model.
2. Modern (semi-parametric) view: a low-dimensional objective defined with respect to the data generative model P_θ.

Linear Regression Example: We consider the following generative model

$$Y_i = X_i^\mathsf{T}\beta^* + \epsilon_i, \qquad X_i \perp \epsilon_i.$$

Semi-parametric view: for a random data Y_i, X_i, we may simply define

$$\beta^* = \arg\min \mathbb{E}(Y_i - X_i^\mathsf{T}\beta)^2. \tag{4}$$

We refer to this as the best linear population approximation.