# Modern Statistics

Xiangyu Chang

April 15, 2026

**Abstract**

To be undated.

## 1 Lecture 13: Simple Linear Regression

The preceding lectures developed the theory of parametric inference: maximum likelihood estimation (MLE), the score function and Fisher information, and the Cramér–Rao lower bound. These results show that the MLE $\widehat{\theta}_n$ is consistent, asymptotically normal with variance $[nI(\theta^*)]^{-1}$, and asymptotically efficient. Combined with CLT-based confidence intervals and hypothesis tests, we now have a complete toolkit for inference about a finite-dimensional parameter $\theta$.

In many applications, however, we observe *pairs* $(X_i, Y_i)$ and want to understand how the response $Y$ depends on the covariate $X$. This is the domain of **regression analysis**. This lecture introduces the simplest and most fundamental setting—**simple linear regression** (SLR)—where a single covariate $X$ is related to the response $Y$ through a linear function. We begin with a brief recap of MLE-based inference, motivate the regression problem through mean squared error minimization, derive the **ordinary least squares** (OLS) estimators, and establish their statistical properties.

### 1.1 Recap: MLE-Based Inference

We briefly recall the key results from parametric inference that will guide our analysis of regression estimators. Given i.i.d. observations $X_1, \ldots, X_n$ from a parametric family $\{f_\theta : \theta \in \Theta\}$, the MLE $\widehat{\theta}_n = \operatorname{argmax}_\theta \ell_n(\theta)$ satisfies, under regularity conditions:

$$\frac{\widehat{\theta}_n - \theta^*}{\widehat{\mathrm{se}}} \xrightarrow{d} N(0,1), \quad \text{where } \widehat{\mathrm{se}} = \frac{1}{\sqrt{n\, I(\widehat{\theta}_n)}}.$$

Here $I(\theta) = \mathrm{Var}\big(S_\theta(X)\big) = -\,\mathbb{E}\big[\nabla^2 \log f_\theta(X)\big]$ is the Fisher information for a single observation, and $S_\theta(x) = \nabla \log f_\theta(x)$ is the score function. The total Fisher information for $n$ observations is $I_n(\theta) = nI(\theta)$.

**Example 1.1** (MLE for Ber$(p)$: Confidence Interval)**.** For $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$, the MLE is $\widehat{p}_n = \bar{X}_n$, with score $S_p(x) = x/p - (1-x)/(1-p)$ and Fisher information $I(p) = 1/\big(p(1-p)\big)$. The estimated standard error is

$$\widehat{\text{se}} = \frac{1}{\sqrt{n\, I(\widehat{p}_n)}} = \sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}},$$

yielding the approximate $1-\alpha$ confidence interval $\widehat{p}_n \pm z_{\alpha/2}\, \widehat{\text{se}}$.

This machinery extends immediately to hypothesis testing. The duality between confidence intervals and tests (reject $H_0\colon \theta = \theta_0$ at level $\alpha$ if and only if $\theta_0 \notin C_n$) provides a unified framework.

**Example 1.2** (A/B Testing)**.** Suppose we run an A/B test comparing two webpage designs, observing click-through rates $\widehat{p}_A = 7.30\%$ and $\widehat{p}_B = 7.29\%$. To determine whether the difference is statistically significant, we test

$$H_0\colon p_A = p_B \qquad \text{vs.} \qquad H_1\colon p_A \neq p_B.$$

Under $H_0$, the test statistic $T_n = |\widehat{p}_A - \widehat{p}_B| \,/\, \widehat{\text{se}}_{\text{diff}}$ is approximately $N(0,1)$ by the CLT. We reject $H_0$ at significance level $\alpha$ if $T_n > z_{\alpha/2}$, or equivalently, if the confidence interval $(\widehat{p}_A - \widehat{p}_B) \pm z_{\alpha/2}\, \widehat{\text{se}}_{\text{diff}}$ does not contain zero.

## 1.2 The Regression Problem

We now turn from estimating a single parameter to modeling the relationship between two variables. Suppose we observe data pairs $\{(x_i, y_i)\}_{i=1}^n$ drawn from a joint distribution $F_{X,Y}$, and wish to predict $Y$ from $X$. A natural prediction framework is:

$$\textbf{Model:} \quad Y = r(X) + \varepsilon, \quad \text{where } \mathbb{E}[\varepsilon \mid X] = 0.$$

Here $r$ is the **regression function** mapping covariates to predictions, and $\varepsilon$ is a mean-zero noise term.

**Terminology:**

- $X$: covariate (also called feature, predictor, or independent variable).
- $Y$: response (also called outcome or dependent variable).

A natural criterion for choosing $r$ is to minimize the **mean squared prediction error**:

$$\min_r \; \mathbb{E}\big[(Y - r(X))^2\big].$$

The following result identifies the optimal predictor under this criterion.

**Theorem 1.3** (Optimality of Conditional Expectation)**.** *Among all measurable functions $r$, the mean squared error $\mathbb{E}\big[(Y - r(X))^2\big]$ is minimized uniquely by the* ***regression function***

$$r^*(x) = \mathbb{E}[Y \mid X = x].$$

*Proof.* Add and subtract $\mathbb{E}[Y \mid X]$ inside the squared loss:

$$\begin{aligned}\mathbb{E}\big[(Y - r(X))^2\big] &= \mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid X] + \mathbb{E}[Y \mid X] - r(X)\big)^2\Big] \\ &= \underbrace{\mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid X]\big)^2\Big]}_{\text{(I): irreducible error}} + \underbrace{\mathbb{E}\Big[\big(\mathbb{E}[Y \mid X] - r(X)\big)^2\Big]}_{\text{(II): approximation error}} \\ &\quad + 2\,\mathbb{E}\big[\big(Y - \mathbb{E}[Y \mid X]\big)\big(\mathbb{E}[Y \mid X] - r(X)\big)\big].\end{aligned}$$

We show that the cross term vanishes. By the law of iterated expectations, conditioning on $X$:

$$\begin{aligned}&\mathbb{E}\big[\big(Y - \mathbb{E}[Y \mid X]\big)\big(\mathbb{E}[Y \mid X] - r(X)\big)\big] \\ &\quad = \mathbb{E}\Big[\mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid X]\big)\big(\mathbb{E}[Y \mid X] - r(X)\big) \;\Big|\; X\Big]\Big] \\ &\quad = \mathbb{E}\Big[\big(\mathbb{E}[Y \mid X] - r(X)\big) \cdot \mathbb{E}\Big[Y - \mathbb{E}[Y \mid X] \;\Big|\; X\Big]\Big] = 0,\end{aligned}$$

where the factor $\mathbb{E}[Y \mid X] - r(X)$ is pulled out of the inner expectation because it is $X$-measurable, and $\mathbb{E}[Y - \mathbb{E}[Y \mid X] \mid X] = \mathbb{E}[Y \mid X] - \mathbb{E}[Y \mid X] = 0$. Therefore,

$$\mathbb{E}\big[(Y - r(X))^2\big] = \mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid X]\big)^2\Big] + \mathbb{E}\Big[\big(\mathbb{E}[Y \mid X] - r(X)\big)^2\Big].$$

Term (I) does not depend on $r$. Term (II) is non-negative and equals zero if and only if $r(X) = \mathbb{E}[Y \mid X]$ almost surely. Hence the unique minimizer is $r^*(X) = \mathbb{E}[Y \mid X]$. ■

### 1.3 The Simple Linear Regression Model

Computing the full conditional expectation $\mathbb{E}[Y \mid X = x]$ is a nonparametric estimation problem that can be challenging, especially with limited data. The simplest parametric approach assumes that the regression function is *linear* in $x$: $r(x) = \beta_0 + \beta_1 x$.

**Definition 1.4** (Simple Linear Regression Model)**.** The **simple linear regression** (SLR) model assumes

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\beta_0$ (intercept) and $\beta_1$ (slope) are unknown parameters, $x_1, \ldots, x_n$ are fixed covariates, and $\varepsilon_1, \ldots, \varepsilon_n$ are independent error terms satisfying

$$\mathbb{E}[\varepsilon_i] = 0, \qquad \mathrm{Var}(\varepsilon_i) = \sigma^2.$$

The quantity $\sigma^2$ is the (unknown) error variance.

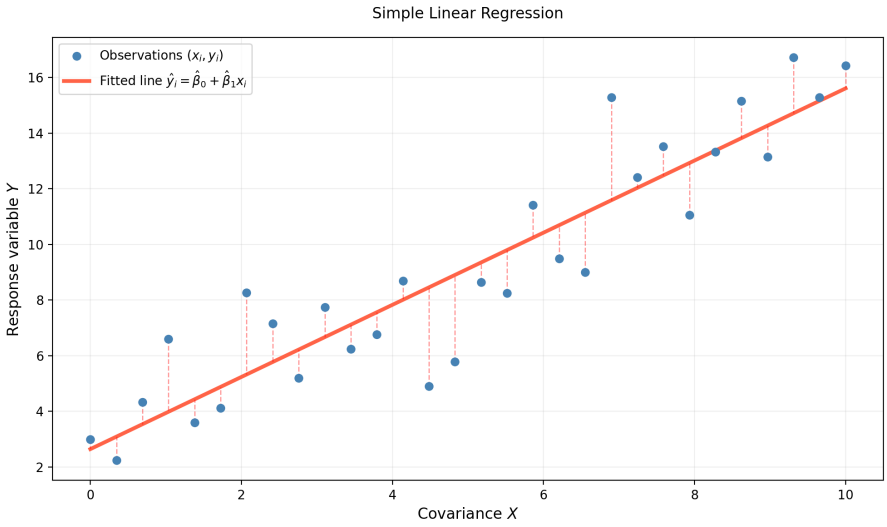


Figure 1: Simple linear regression: the data points $(x_i, y_i)$ scatter around the true line $\beta_0 + \beta_1 x$. The vertical distances are the errors $\varepsilon_i$. The estimated line $\widehat{\beta}_0 + \widehat{\beta}_1 x$ is chosen to minimize the sum of squared vertical distances.

The data generating process produces observed values $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, and our goal is to estimate $\beta_0$ and $\beta_1$ from $\{(x_i, y_i)\}_{i=1}^n$. The fitted values are $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$, and the residuals are $\widehat{\varepsilon}_i = y_i - \widehat{y}_i$.

## 1.4 Ordinary Least Squares Estimation

The **ordinary least squares** (OLS) method estimates $\beta_0$ and $\beta_1$ by minimizing the sum of squared residuals:

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 .$$

**Remark 1.5** (Comparison with Least Absolute Deviations)**.** *An alternative is the **least absolute deviations** (LAD) estimator, which minimizes $\sum_{i=1}^n |y_i - \beta_0 - \beta_1 x_i|$. While LAD is more robust to outliers, the OLS objective is differentiable everywhere, leading to closed-form estimators and simpler theoretical analysis. We focus on OLS throughout.*

### 1.4.1 Derivation of OLS Estimators

Define the objective function $S(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$. Setting the partial derivatives to zero yields the **normal equations**:

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \qquad \Rightarrow \quad \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i,$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i \, (y_i - \beta_0 - \beta_1 x_i) = 0 \qquad \Rightarrow \quad \sum_{i=1}^{n} x_i y_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2.$$

Dividing the first equation by $n$ gives

$$\bar{y}_n = \beta_0 + \beta_1 \bar{x}_n \quad \Rightarrow \quad \widehat{\beta}_0 = \bar{y}_n - \widehat{\beta}_1 \bar{x}_n.$$

Substituting this expression for $\beta_0$ into the second equation and simplifying:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}.$$

### 1.4.2 Expression in Terms of Sample Moments

Define the following sample quantities:

$$S_{xx} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)^2 \quad \text{(sample variance of } x\text{)}, \qquad S_{xy} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n) \quad \text{(sample covariance)}.$$

The OLS estimators then take the compact form:

$$\boxed{\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \qquad \widehat{\beta}_0 = \bar{y}_n - \widehat{\beta}_1 \bar{x}_n.}$$

Note that $\widehat{\beta}_1$ is the ratio of the sample covariance of $(x, y)$ to the sample variance of $x$, mirroring the population relationship $\beta_1 = \mathrm{Cov}(X, Y)/\mathrm{Var}(X)$ that holds when the linear model is correctly specified. In this sense, the OLS estimator can also be viewed as a *method of moments* estimator.

## 1.5 Properties of OLS Estimators

We now establish the fundamental statistical properties of the OLS estimators: unbiasedness and variance formulas. Throughout, we treat the covariates $x_1, \ldots, x_n$ as fixed and compute expectations and variances with respect to the randomness in $\varepsilon_1, \ldots, \varepsilon_n$.

**Theorem 1.6** (Unbiasedness and Variance of OLS Estimators)**.** *Under the SLR model (Definition 1.4):*

$$\mathbb{E}[\widehat{\beta}_0] = \beta_0, \qquad \mathrm{Var}(\widehat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}_n^2}{nS_{xx}}\right),$$

$$\mathbb{E}[\widehat{\beta}_1] = \beta_1, \qquad \mathrm{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{nS_{xx}}.$$

*Proof.* We prove each claim in turn. A key identity used repeatedly is: for any constant $c$,

$$\sum_{i=1}^n (x_i - \bar{x}_n) \cdot c = c\sum_{i=1}^n (x_i - \bar{x}_n) = 0.$$

**Step 1: $\mathbb{E}[\widehat{\beta}_1] = \beta_1$.** Since $S_{xx}$ depends only on the fixed covariates, $\mathbb{E}[\widehat{\beta}_1] = \mathbb{E}[S_{xy}]\,/\,S_{xx}$. We compute $\mathbb{E}[S_{xy}]$ by first noting that $\sum_i (x_i - \bar{x}_n)\bar{y}_n = 0$, so

$$S_{xy} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n) = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)\, y_i.$$

Taking expectations:

$$\begin{aligned}\mathbb{E}[S_{xy}] &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}_n)\ \mathbb{E}[y_i] = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}_n)(\beta_0+\beta_1 x_i)\\ &= \frac{\beta_1}{n}\sum_{i=1}^{n}(x_i-\bar{x}_n)\,x_i \qquad \big(\text{since } \textstyle\sum_i (x_i-\bar{x}_n)\,\beta_0 = 0\big)\\ &= \frac{\beta_1}{n}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2 \qquad \big(\text{since } \textstyle\sum_i (x_i-\bar{x}_n)\,\bar{x}_n = 0\big) = \beta_1\, S_{xx}.\end{aligned}$$

Therefore, $\mathbb{E}[\widehat{\beta}_1] = \beta_1 S_{xx} \,/\, S_{xx} = \beta_1$.

**Step 2:** $\mathrm{Var}(\widehat{\beta}_1) = \sigma^2/(nS_{xx})$**.** Using $S_{xy} = \frac{1}{n}\sum_i (x_i-\bar{x}_n)\,y_i$ and the independence of the $y_i$ (each with variance $\sigma^2$):

$$\mathrm{Var}(S_{xy}) = \frac{1}{n^2}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2\ \mathrm{Var}(y_i) = \frac{\sigma^2}{n^2}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2 = \frac{\sigma^2\, S_{xx}}{n}.$$

Therefore,

$$\mathrm{Var}(\widehat{\beta}_1) = \frac{\mathrm{Var}(S_{xy})}{S_{xx}^2} = \frac{\sigma^2\, S_{xx}}{n\, S_{xx}^2} = \frac{\sigma^2}{nS_{xx}}.$$

**Step 3:** $\mathbb{E}[\widehat{\beta}_0] = \beta_0$**.** From $\widehat{\beta}_0 = \bar{y}_n - \widehat{\beta}_1\bar{x}_n$:

$$\mathbb{E}[\widehat{\beta}_0] = \mathbb{E}[\bar{y}_n] - \bar{x}_n\ \mathbb{E}[\widehat{\beta}_1] = (\beta_0+\beta_1\bar{x}_n) - \bar{x}_n\cdot\beta_1 = \beta_0.$$

**Step 4:** $\mathrm{Var}(\widehat{\beta}_0) = \sigma^2\big(\frac{1}{n} + \frac{\bar{x}_n^2}{nS_{xx}}\big)$**.** Since $\widehat{\beta}_0 = \bar{y}_n - \widehat{\beta}_1\,\bar{x}_n$,

$$\mathrm{Var}(\widehat{\beta}_0) = \mathrm{Var}(\bar{y}_n) + \bar{x}_n^2\ \mathrm{Var}(\widehat{\beta}_1) - 2\bar{x}_n\ \mathrm{Cov}(\bar{y}_n,\,\widehat{\beta}_1).$$

We claim that $\mathrm{Cov}(\bar{y}_n, \widehat{\beta}_1) = 0$. To verify, recall that $\bar{y}_n = \frac{1}{n}\sum_j y_j$ and $\widehat{\beta}_1 = \frac{1}{nS_{xx}}\sum_i (x_i-\bar{x}_n)\,y_i$. Since the $y_i$ are independent with common variance $\sigma^2$:

$$\mathrm{Cov}(\bar{y}_n,\,\widehat{\beta}_1) = \frac{1}{n^2 S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x}_n)\ \mathrm{Var}(y_i) = \frac{\sigma^2}{n^2 S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x}_n) = 0.$$

The covariance vanishes because the centered covariates sum to zero. Substituting:

$$\mathrm{Var}(\widehat{\beta}_0) = \frac{\sigma^2}{n} + \bar{x}_n^2\cdot\frac{\sigma^2}{nS_{xx}} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}_n^2}{nS_{xx}}\right).$$ ■

**Remark 1.7** (Interpreting the Variance Formulas)**.** *The variance of $\widehat{\beta}_1$ is inversely proportional to $nS_{xx}$: more data (large $n$) and greater spread in the covariates (large $S_{xx}$) both improve estimation precision. The variance of $\widehat{\beta}_0$ has an additional term $\bar{x}_n^2/(nS_{xx})$ reflecting the uncertainty in extrapolating the fitted line to $x = 0$ when the data are centered far from the origin.*

### 1.6 Estimation of the Error Variance

The variance formulas in Theorem 1.6 involve the unknown error variance $\sigma^2$. To make them operational—for instance, to construct confidence intervals for $\beta_0$ and $\beta_1$—we must estimate $\sigma^2$ from the data.

The natural approach uses the residuals $\widehat{\varepsilon}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$. However, the naïve plug-in estimator $\frac{1}{n}\sum_{i=1}^n \widehat{\varepsilon}_i^2$ is biased downward: since the OLS line was *fitted* to the data, the residuals tend to be smaller than the true errors. The correct normalizing constant is $n-2$ rather than $n$, accounting for the two degrees of freedom consumed by estimating $\beta_0$ and $\beta_1$.

**Theorem 1.8** (Unbiased Estimation of $\sigma^2$)**.** *Define the* ***residual sum of squares*** $\mathrm{RSS} = \sum_{i=1}^n (y_i - \widehat{y}_i)^2$. *The estimator*

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n (y_i - \widehat{y}_i)^2 = \frac{\mathrm{RSS}}{n-2}$$

*is unbiased:* $\mathbb{E}[\widehat{\sigma}^2] = \sigma^2$.

The proof proceeds by expanding RSS as a quadratic form in the observations, computing $\mathbb{E}[\mathrm{RSS}]$ term by term, and showing that $\mathbb{E}[\mathrm{RSS}] = (n-2)\sigma^2$. The complete derivation is given in the next lecture, where the same argument generalizes to $n-p$ in the multiple regression setting.

### 1.7 Summary and Outlook

This lecture introduced simple linear regression and derived the ordinary least squares estimators:

1. **The regression problem**: the optimal predictor of $Y$ given $X$ under squared loss is the conditional expectation $r^*(x) = \mathbb{E}[Y \mid X = x]$. Simple linear regression parametrizes this as $r(x) = \beta_0 + \beta_1 x$.
2. **OLS estimators**: $\widehat{\beta}_1 = S_{xy}/S_{xx}$ and $\widehat{\beta}_0 = \bar{y}_n - \widehat{\beta}_1 \bar{x}_n$, obtained by minimizing the sum of squared residuals.
3. **Statistical properties**: both estimators are unbiased, with variances $\mathrm{Var}(\widehat{\beta}_1) = \sigma^2/(nS_{xx})$ and $\mathrm{Var}(\widehat{\beta}_0) = \sigma^2\big(1/n + \bar{x}_n^2/(nS_{xx})\big)$.
4. **Error variance**: the unbiased estimator is $\widehat{\sigma}^2 = \mathrm{RSS}/(n-2)$.

**What comes next.** The next lecture extends these results in two directions. First, we show that under Gaussian errors $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, the OLS estimators coincide with the MLEs and follow exact normal distributions, enabling $t$-based confidence intervals and hypothesis tests for $\beta_0$ and $\beta_1$. Second, we generalize to **multiple linear regression**, where the response depends on $p$ covariates: $Y = X\beta + \varepsilon$. The OLS estimator becomes $\widehat{\beta} = (X^\top X)^{-1} X^\top Y$, and the theory of projection matrices provides an elegant geometric framework for analyzing its properties.

## References