Yale University

Linear Regression I

S&DS 265 · Introductory Machine Learning · Lecture 3

Zhuoran Yang

Fall 2026

AI usage: figures were generated, and these slides polished, with the help of AI tools.

Outline

Learning Objectives

In this lecture you will learn:

1. How a prediction problem is set up, and why we start with a linear model;
2. How to fit a linear model to data by least squares;
3. How to interpret the fitted coefficients, and what they do not tell us;
4. Why the objective is convex, and what convexity does and does not guarantee;
5. When linear regression breaks down, and what each failure calls for.

Outline

Motivating Example: Advertising Budgets

A firm sells one product in 200 markets. In each market it knows what it spent on TV, radio, and newspaper advertising, and how many units it sold.

It now has $100,000 to spend in a market where it has never advertised.

How many units should it expect to sell, and how should it split the budget across the three channels?

Example and data: ISLP §2.1 and §3.1 (James, Witten, Hastie, Tibshirani, and Taylor, 2023).

The Advertising Data

	advertising budget ($1000s)			sales
market	TV	radio	newspaper	(1000s)
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
⋮	⋮	⋮	⋮	⋮
200	232.1	8.6	8.7	13.4

One row per market, a distinct sales region. The budget columns are what was spent advertising this product through each medium in that market.

Source: The `Advertising` data set, from ISLP §2.1 (James et al., 2023); available at `statlearning.com`.

Reading One Row

market	TV	radio	newspaper	sales
1	230.1	37.8	69.2	22.1

In market 1, the firm bought $230,100 of television advertising, $37,800 of radio advertising, and $69,200 of newspaper advertising for this product. It then sold 22,100 units there.

Problem Setup

Training data. n observed pairs, $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$, assumed drawn independently from some distribution $\mathcal{P}$.

Features (input). $x_i = (x_{i1}, \dots, x_{id})^\top \in \mathbb{R}^d$, where x_{ij} is feature j of observation i.

Response (output). $y_i \in \mathbb{R}$, a single number per observation.

Goal. Use $\mathcal{D}_n$ to build $\hat{f}$ so that $\hat{f}(x^*)$ predicts the response at an unseen x^*.

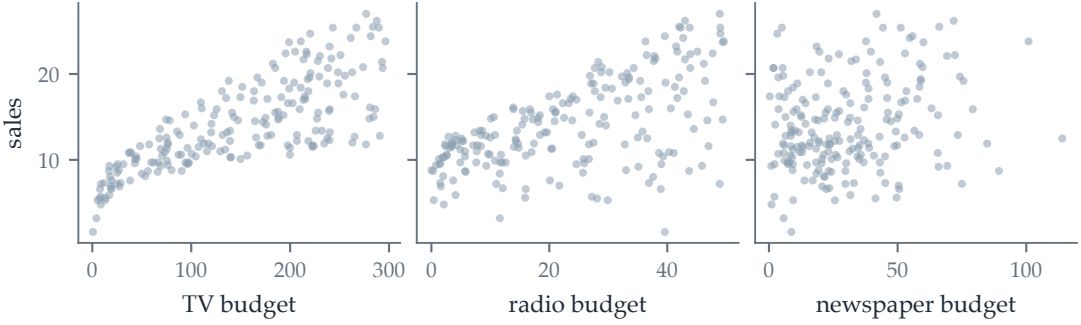
For `Advertising`: $n = 200$, $d = 3$, $x_i = (\text{TV}, \text{radio}, \text{newspaper})^\top$, $y_i = \text{sales}$.

Regression versus Classification

Recall from Lecture 1: this setup is fixed for the whole course. Only the type of the response changes.

	response y_i	the question about a market
regression	a number, $y_i \in \mathbb{R}$	how many units will it sell?
classification	a label, $y_i \in \{1, \dots, K\}$	will it sell more than 15,000?

Sales Against Each Channel



Sales rise with all three budgets, but the relationships are noisy and none is exact. We need a way to summarize a trend like this, and a way to decide which summary is best.

`Advertising` data; compare ISLP Fig. 2.1.

Model, Loss, and Optimizer

We want a function f that turns a market's budgets into a predicted number of units:

$$f : \underbrace{\mathbb{R}^d}_{\text{budgets } x} \longrightarrow \underbrace{\mathbb{R}}_{\text{predicted sales}} .$$

Three choices have to be made, and they are genuinely separate:

1. Which functions f we are willing to consider (model)
2. What it costs to be wrong by a given amount (loss)
3. How we actually find the best f we allowed (optimizer)

Keeping these three apart is the single most useful habit in this course. Today we make the simplest choice for each; even so, the optimizer takes up half the lecture.

Outline

The Statistical Model (Our Assumption)

Assume the response is generated as

$$Y = m(X) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid X] = 0, \qquad \mathrm{Var}(\varepsilon \mid X) = \sigma^2,$$

where m is a fixed but unknown function and ε collects everything about sales that the three budgets do not determine: local competition, weather, the store manager, luck.

m	the systematic part, the same for every market with budget x
ε	the unsystematic part, unpredictable from x by definition

Nothing here says m is a line. We first identify the prediction target, then choose a model class that can approximate it.

The Optimal Predictor under Squared Loss

Suppose (X, Y) is a random market, and score a predictor f by its expected squared error,

$$R(f) = \mathbb{E}[(Y - f(X))^2].$$

Write $m(x) = \mathbb{E}[Y \mid X = x]$ for the conditional mean. We will show that at every x,

$$\mathbb{E}[(Y - f(X))^2 \mid X = x] = \mathrm{Var}(Y \mid X = x) \; + \; (m(x) - f(x))^2,$$

so that only the second term depends on our choice of f.

The first term is the variability of sales among markets with identical budgets. No predictor can touch it.

Derivation: Decomposing the Risk

Step 1. Fix x and split the error by adding and subtracting $m(x)$:

$$Y - f(x) = \underbrace{(Y - m(x))}_{\text{random}} + \underbrace{(m(x) - f(x))}_{\text{a fixed number, once } x \text{ is given}}.$$

Step 2. Square both sides:

$$(Y - f(x))^2 = (Y - m(x))^2 + 2(Y - m(x))(m(x) - f(x)) + (m(x) - f(x))^2.$$

Derivation: Taking the Expectation

Step 3. Take $\mathbb{E}[\,\cdot \mid X = x]$ of each of the three terms:

$$\mathbb{E}\big[(Y - m(x))^2 \mid X = x\big] = \operatorname{Var}(Y \mid X = x),$$

$$\mathbb{E}\big[(Y - m(x))(m(x) - f(x)) \mid X = x\big] = (m(x) - f(x))\,\mathbb{E}[Y - m(x) \mid X = x] = 0,$$

$$\mathbb{E}\big[(m(x) - f(x))^2 \mid X = x\big] = (m(x) - f(x))^2.$$

The first holds because $\mathbb{E}[Y \mid X = x] = m(x)$. In the second, $m(x) - f(x)$ is a constant given x, so it factors out of the expectation, and what remains is zero. The third is the expectation of a constant. **Step 4.** Add the three, then average over X:

$$R(f) = \underbrace{\mathbb{E}\operatorname{Var}(Y \mid X)}_{\text{does not involve } f} + \mathbb{E}(m(X) - f(X))^2.$$

The Regression Function

Only the second term involves f, and it is smallest when f matches m exactly. The best possible predictor under squared loss is therefore the conditional mean,

$$\arg\min_{a \in \mathbb{R}} \; \mathbb{E}\big[(Y - a)^2 \mid X = x\big] \;=\; m(x) = \mathbb{E}[Y \mid X = x],$$

and no predictor can beat it. The leftover $\mathbb{E}\,\mathrm{Var}(Y \mid X)$ is irreducible: it is the part of sales that the budgets simply do not determine.

Everything else in this course is an attempt to approximate m from a finite sample.

Why We Need a Model

We know what we are looking for: the regression function m. But m is an unknown function, and the set of all functions $f : \mathbb{R}^d \to \mathbb{R}$ is vastly too large to search—for any finite sample, infinitely many of its members fit the data perfectly and disagree everywhere else.

Why We Need a Model

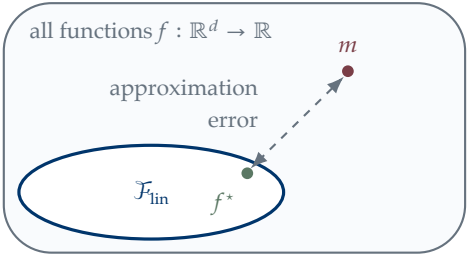
We know what we are looking for: the regression function m. But m is an unknown function, and the set of all functions $f : \mathbb{R}^d \to \mathbb{R}$ is vastly too large to search—for any finite sample, infinitely many of its members fit the data perfectly and disagree everywhere else.

So we must assume something. We fix in advance a **function class** $\mathcal{F}$—a hypothesis about what m looks like—and search only inside it:

$$f^\star = \arg\min_{f \in \mathcal{F}} \mathbb{E}[(Y - f(X))^2].$$

This is what modeling means, and it is what makes the problem tractable. Every method in this course is a choice of $\mathcal{F}$.

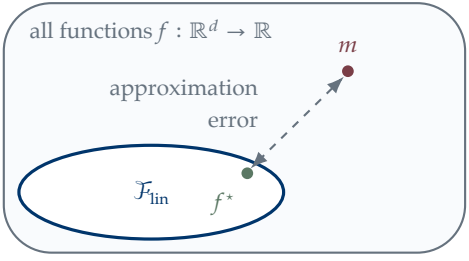
The Model Class Inside the Space of Functions



m is the best predictor of all, but it need not lie in the class we chose.

$f^\star$ is the best our class can do, and no amount of data closes the gap.

The Model Class Inside the Space of Functions



m is the best predictor of all, but it need not lie in the class we chose.

$f^\star$ is the best our class can do, and no amount of data closes the gap.

How large should $\mathcal{F}$ be? Two forces pull against each other:

- A **small** class is easy to fit well from n observations, but $f^\star$ may be far from m;
- A **large** class puts $f^\star$ close to m, but finding it from only n observations is hard.

Lecture 6 makes this trade-off precise.

The Linear Model Assumption

The simplest useful choice is the class of linear functions,

$$\mathcal{F}_{\text{lin}} = \{ f_\beta(x) = x^\top \beta \ : \ \beta \in \mathbb{R}^p \},$$

which reduces the search from a space of functions to a search over p numbers.

what we gain	few parameters, each estimated from all n rows; coefficients that mean something
what we assume	that m is close to linear over the range we care about

If m is far from $\mathcal{F}_{\text{lin}}$, then $f^\star$ is a poor approximation. More data and better optimization cannot repair this modeling error.

The Linear Model

We restrict attention to functions of the form

$$f_\beta(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d,$$

so the statistical model becomes

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_d x_{id} + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i] = 0, \quad \mathrm{Var}(\varepsilon_i) = \sigma^2.$$

Absorbing the intercept by appending a constant 1 to every feature vector, $x_i \leftarrow (1, x_{i1}, \dots, x_{id})^\top \in \mathbb{R}^p$ with $p = d + 1$,

$$y_i = x_i^\top \beta + \varepsilon_i.$$

From here on $p = d + 1$ counts coefficients. The augmentation is bookkeeping, not an assumption.

Residuals and the Empirical Risk

For market i, a candidate β predicts $\hat{y}_i = x_i^\top \beta$ and leaves the **residual**

$$r_i(\beta) = y_i - x_i^\top \beta.$$

We need one number summarizing all n of them. Square each residual and take the average:

$$\widehat{R}_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} r_i(\beta)^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2.$$

The average is a typical squared error per market and remains comparable across data sets of different sizes.

$\mathrm{RSS}(\beta) = n\,\widehat{R}_n(\beta)$ is the classical residual sum of squares; it has the same minimizer.

Why the Average Is the Right Thing to Minimize

derive

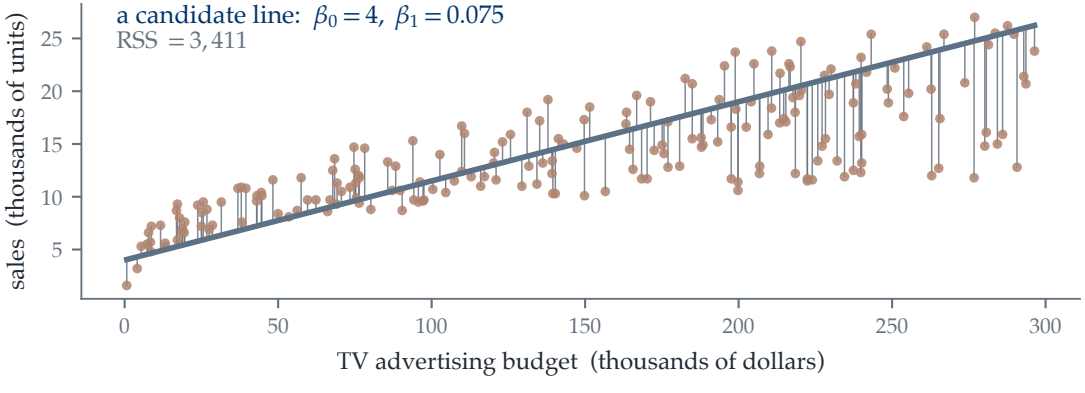
Population risk $R(\beta) = \mathbb{E}[(Y - X^\top \beta)^2]$ is unavailable. Under the i.i.d. assumption, the sample average converges to it:

$$\widehat{R}_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2 \xrightarrow{n \to \infty} R(\beta).$$

Therefore we solve the empirical problem

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \widehat{R}_n(\beta).$$

Residuals Depend on Which Line You Pick



Here is one candidate line and its residuals. A different line gives different residuals and a different empirical risk. Least squares picks the line making that risk as small as possible.

`Advertising` data; the line shown is a guess, not the fit.

The Optimization Problem

Model, loss, and the i.i.d. assumption together turn the original question into a concrete optimization problem: choose the coefficient vector minimizing the empirical squared error,

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad \widehat{R}_n(\beta) \;=\; \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2 \;=\; \frac{1}{n}\|y - X\beta\|_2^2.$$

There are no constraints on β, so this is an unconstrained problem in p real variables. Setting the gradient to zero solves it.

Outline

Simple Regression: First-Order Conditions

$$\widehat{R}_n(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

Differentiate, using the chain rule on each term:

$$\frac{\partial \widehat{R}_n}{\partial \beta_0} = \frac{1}{n}\sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)\cdot(-1) = -\frac{2}{n}\sum_{i=1}^{n} r_i,$$

$$\frac{\partial \widehat{R}_n}{\partial \beta_1} = \frac{1}{n}\sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)\cdot(-x_i) = -\frac{2}{n}\sum_{i=1}^{n} x_i r_i.$$

Both conditions can be read directly: at the optimum the residuals sum to zero and are uncorrelated with the feature.

Simple Regression: The Solution

Setting $\partial \widehat{R}_n / \partial \beta_0 = 0$ gives $\sum_i (y_i - \beta_0 - \beta_1 x_i) = 0$, so

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}, \qquad \bar{x} = \tfrac{1}{n}\sum_i x_i, \;\; \bar{y} = \tfrac{1}{n}\sum_i y_i.$$

Substitute this into $\sum_i x_i (y_i - \beta_0 - \beta_1 x_i) = 0$:

$$0 = \sum_i x_i (y_i - \bar{y} + \beta_1 \bar{x} - \beta_1 x_i) = \sum_i x_i (y_i - \bar{y}) - \beta_1 \sum_i x_i (x_i - \bar{x}).$$

Since $\sum_i \bar{x}(y_i - \bar{y}) = 0$ and $\sum_i \bar{x}(x_i - \bar{x}) = 0$, we may centre both factors:

$$\boxed{\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.}$$

Worked Example: Sales on TV

From the `Advertising` table, with x = TV budget and y = sales:

$$\bar{x} = 147.04, \qquad \bar{y} = 14.02, \qquad S_{xy} = 69\,727.6, \qquad S_{xx} = 1\,466\,819.$$

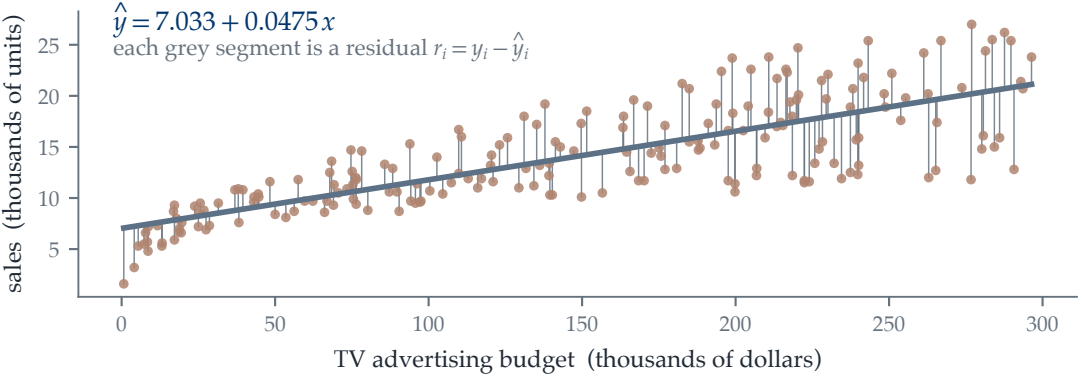
Therefore

$$\hat{\beta}_1 = \frac{69\,727.6}{1\,466\,819} = 0.04754, \qquad \hat{\beta}_0 = 14.02 - 0.04754 \times 147.04 = 7.033.$$

Each extra \$1000 of TV budget goes with $0.0475 \times 1000 \approx 48$ **more units sold, so 1000 more units needs about** $1000/0.0475 \approx 21{,}000$ **dollars of extra TV budget.**

$\hat{\beta}_0 = 7.03$ is predicted sales at zero TV budget. Here that is near the observed range, so it is interpretable; often it is not.

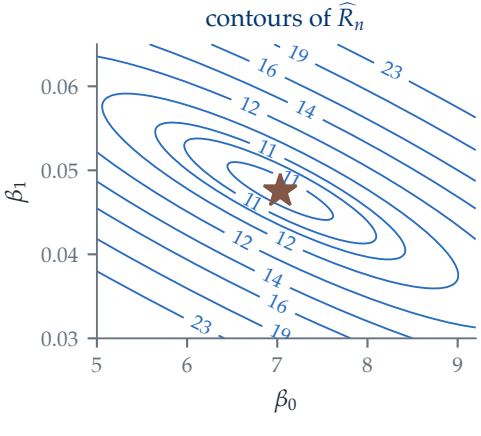
Least Squares on the Advertising Data

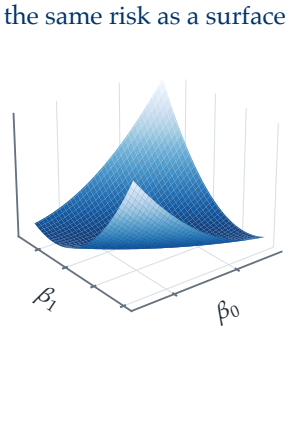


Those are the coefficients we just computed. The fitted line says about **47.5 extra units sold per additional $1000** of TV budget.

Recreation of ISLP Fig. 3.1 from the `Advertising` data.

The Least-Squares Objective





There is exactly one flat point, and it is the global minimum. Few of the models later in this course are this well behaved.

Recreation of ISLP Fig. 3.2 from the `Advertising` data.

Multiple Regression in Matrix Form

Put a column of ones in front, so the intercept becomes just another coefficient:

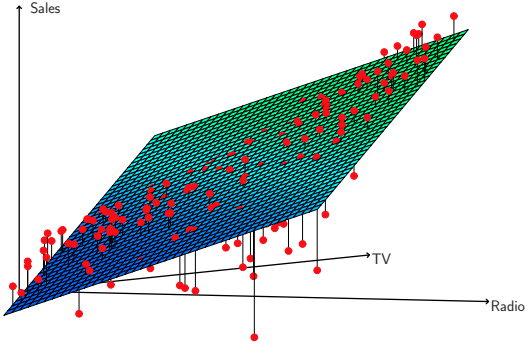
$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1d} \\ 1 & x_{21} & \cdots & x_{2d} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nd} \end{pmatrix} \in \mathbb{R}^{n \times p}, \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{pmatrix} \in \mathbb{R}^p.$$

Then $X\beta \in \mathbb{R}^n$ holds all n predictions at once, and

$$\widehat{R}_n(\beta) = \frac{1}{n}\|y - X\beta\|_2^2.$$

Row i of X is the augmented feature vector $x_i^\top$. For `Advertising`, $n = 200$ markets, $d = 3$ channels, and $p = d + 1 = 4$ coefficients.

With Two Features the Line Becomes a Plane



With one feature the fit was a line; with two it is a plane, and with d it is a d-dimensional hyperplane. The residuals are still measured vertically, and least squares still minimizes their average square.

Source: ISLP Fig. 3.5 (James et al., 2023), reproduced unmodified.

The Gradient of the Empirical Risk

Since $\|v\|_2^2 = v^\top v$,

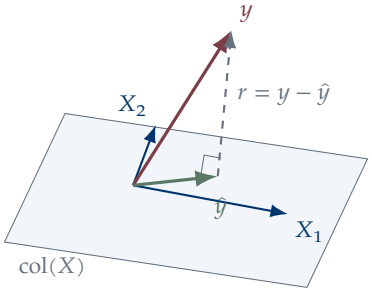
$$\begin{aligned} n\,\widehat{R}_n(\beta) &= (y - X\beta)^\top (y - X\beta) \\ &= y^\top y - \underbrace{y^\top X\beta - \beta^\top X^\top y}_{\text{both scalars, hence equal}} + \beta^\top X^\top X\beta \\ &= y^\top y - 2\beta^\top X^\top y + \beta^\top X^\top X\beta. \end{aligned}$$

Using $\nabla_\beta(\beta^\top a) = a$ and $\nabla_\beta(\beta^\top A\beta) = 2A\beta$ for symmetric $A = X^\top X$,

$$\nabla_\beta \widehat{R}_n(\beta) = \frac{1}{n}(-2X^\top y + 2X^\top X\beta) = -\frac{2}{n}X^\top (y - X\beta).$$

Shape check: $(p \times n)(n \times 1) = p \times 1$, one gradient entry per coefficient.

The Normal Equations



Setting the gradient to zero,

$$X^\top X\hat{\beta} = X^\top y \quad \Longleftrightarrow \quad X^\top (y - X\hat{\beta}) = 0.$$

Every feature column is orthogonal to the residual. Equivalently, the fitted vector $\hat{y} = X\hat{\beta}$ is the point of col(X) closest to y: least squares is a **projection**.

The fit has extracted everything the columns of X can say about y; whatever remains in the residual is invisible to them.

The Closed-Form Solution

If $X^\top X$ is invertible, the normal equations have the unique solution

$$\hat{\beta} = (X^\top X)^{-1} X^\top y,$$

and the fitted values are $\hat{y} = X\hat{\beta}$.

Two questions are still open, and convexity answers both:

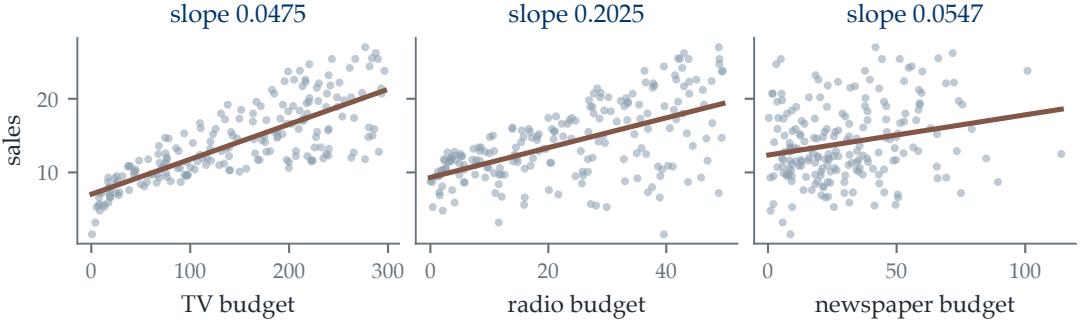
- Why a stationary point of $\widehat{R}_n$ has to be a minimum at all;
- What happens when $X^\top X$ is not invertible.

Overview: Fitting a Linear Model

Data assumption	(x_i, y_i) i.i.d. from $\mathcal{D}$, with $Y = m(X) + \varepsilon$, $\mathbb{E}[\varepsilon \mid X] = 0$
Model assumption	m is close to linear: we search $\mathcal{F}_{\text{lin}} = \{x \mapsto x^\top \beta\}$
Loss	squared error—justified as maximum likelihood below
Objective	the empirical risk $\widehat{R}_n(\beta) = \frac{1}{n}\|y - X\beta\|_2^2$
Computation	a closed form, obtained by setting the gradient to zero

Every row is an assumption that can fail, and each fails differently. We take them apart one at a time.

One Simple Regression per Channel



Now that we can fit a line, here is the least-squares fit of sales on each budget separately. All three slopes are positive, and all three are clearly nonzero.

`Advertising` data; compare ISLP Fig. 2.1 and Table 3.3.

Multiple Regression on the Advertising Data

Compare separate fits with one joint fit.

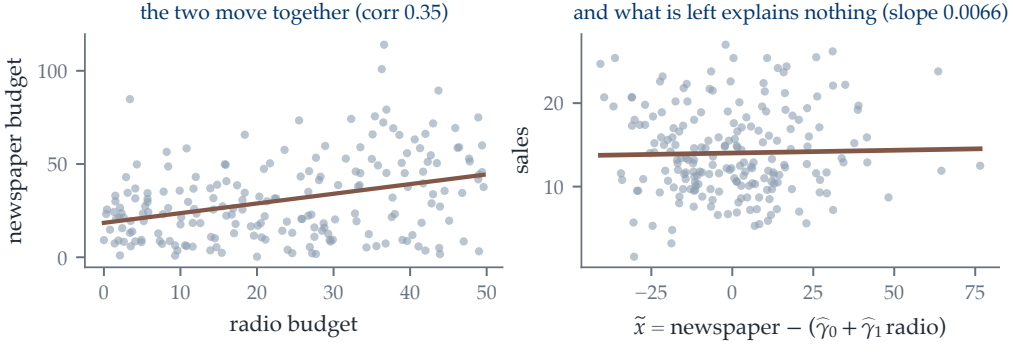
channel	slope, fitted alone	slope, fitted together
TV	0.0475	0.0458
radio	0.2025	0.1885
newspaper	0.0547	−0.0010

$\widehat{R}_n$ falls from 10.51 to 2.78: a typical miss of 1,670 rather than 3,240 units.

TV and radio barely move. Newspaper collapses to zero. Why?

Source: ISLP `Advertising` data; compare Tables 3.3–3.4.

Newspaper Was Standing In for Radio



Markets that buy radio also buy newspaper. So regress newspaper on radio and keep what is left, $\tilde{x}$ = newspaper − ($\hat{\gamma}_0 + \hat{\gamma}_1$ radio): the newspaper spending radio does not account for. Against sales it is flat—slope 0.0066, against 0.0547 for newspaper itself—and that slope is exactly the newspaper coefficient once radio is in the model.

`Advertising` data; the comparison follows ISLP §3.2.

Interpreting a Regression Coefficient

The newspaper slope answers a different question in each fit.

- **Alone**: across markets, how does sales differ as newspaper budget differs? Markets with big newspaper budgets also have big radio budgets, so newspaper takes credit for radio.
- **Together**: among markets with the same TV and radio budgets, how does sales differ as newspaper budget differs? Hardly at all.

A coefficient is always conditional on which other columns are present. When one feature is largely predictable from the others, the design is collinear, which inflates the variance of the estimates.

Aside: Squared Loss Is Maximum Likelihood

If $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ independently, then $y_i \mid x_i \sim \mathcal{N}(x_i^\top \beta, \sigma^2)$ and the log-likelihood is

$$\log L(\beta) = \text{constant} \; - \; \frac{n}{2\sigma^2} \underbrace{\frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2}_{\widehat{R}_n(\beta)}.$$

Therefore

$$\arg\max_{\beta} \log L(\beta) \;=\; \arg\min_{\beta} \widehat{R}_n(\beta).$$

Squared loss corresponds to a Gaussian noise model.

Outline

Why We Trusted the Stationary Point

We solved $\nabla \widehat{R}_n(\beta) = 0$ and called the answer the minimizer. That step needs justification: a stationary point could be a maximum, a saddle, or one of many local minima.

For $\widehat{R}_n$ it is none of these, and the reason is a property of the function itself rather than anything special about our data.

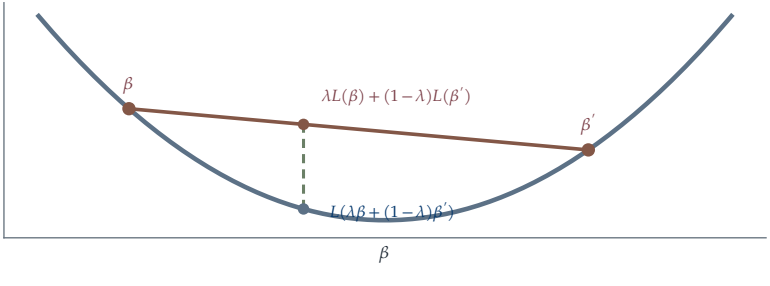
We now make that property precise, because it is exactly the property that fails for the models later in the course.

Convex Functions

Take $L : \mathbb{R}^p \to \mathbb{R}$ differentiable. L is **convex** if for all β, β' and all $\lambda \in [0,1]$,

$$L(\lambda\beta + (1-\lambda)\beta') \leq \lambda L(\beta) + (1-\lambda)L(\beta'),$$

and **strictly convex** if the inequality is strict for $\beta \neq \beta', \lambda \in (0,1)$.

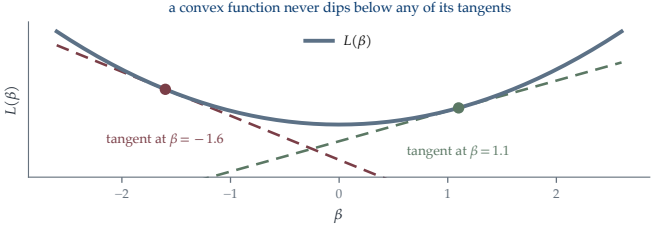


The chord joining any two points of the graph never dips below the graph.

First-Order Condition

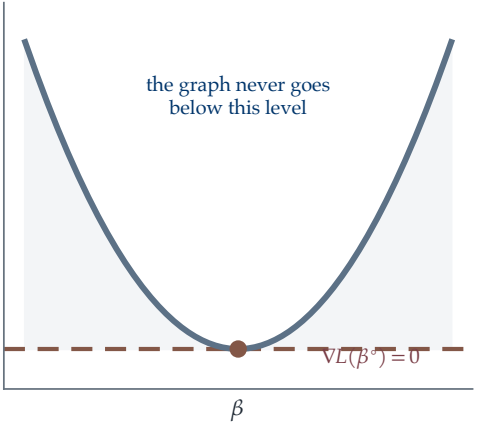
For differentiable L, convexity is exactly the statement that the graph never dips below any of its tangent lines:

$$L(\beta') \;\geq\; L(\beta) + \nabla L(\beta)^\top (\beta' - \beta) \qquad \text{for all } \beta, \beta'.$$



Every tangent is a global underestimate of a differentiable convex function.

Every Stationary Point Is a Global Minimum



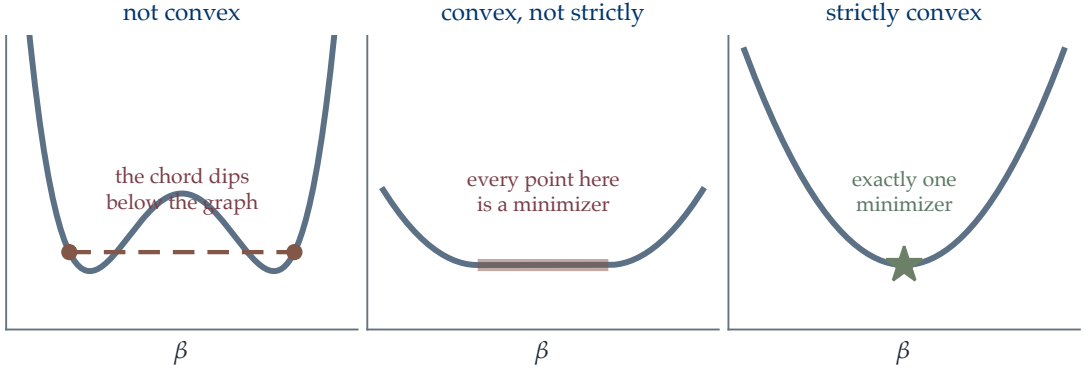
Suppose $\nabla L(\beta^\circ) = 0$. Putting that into the tangent inequality, for every β',

$$L(\beta') \geq L(\beta^\circ).$$

The tangent at β° is horizontal, and the graph stays above it, so β° is a global minimum.

No local minima to get trapped in, and no saddle points.

Convex, Strictly Convex, and the Difference



A convex function has a unique minimum value, but it may be attained on a whole flat set. Strict convexity is what makes the minimizer unique.

Second-Order Conditions

For twice-differentiable L, these are equivalent:

$\nabla^2 L(\beta) \succeq 0$ for all β	L is convex
$\nabla^2 L(\beta) \succ 0$ for all β	L is strictly convex
$\nabla^2 L(\beta) \succeq \mu I$, $\mu > 0$	L is **μ-strongly convex**

Strong convexity is the quantitative version: the function curves upward at least as fast as $\frac{\mu}{2}\|\beta\|^2$ in every direction, so the minimizer is well separated from near-minimizers.

Applying This to Least Squares

derive

Differentiating the gradient once more, $\nabla^2 \widehat{R}_n(\beta) = \frac{2}{n} X^\top X$. For any $v \in \mathbb{R}^p$,

$$v^\top \Big(\frac{2}{n} X^\top X \Big) v = \frac{2}{n} \|Xv\|_2^2 \;\geq\; 0,$$

so $\widehat{R}_n$ is always convex, and the $\hat{\beta}$ we found is a global minimum.

It is strictly convex exactly when $\|Xv\|_2^2 > 0$ for every $v \neq 0$, that is when the columns of X are linearly independent. In that case $\nabla^2 \widehat{R}_n \succeq \frac{2}{n} \lambda_{\min}(X^\top X)\, I$ with $\lambda_{\min} > 0$, so $\widehat{R}_n$ is even strongly convex.

If some $v \neq 0$ has $Xv = 0$, the objective is flat along v: convex, with a whole line of minimizers. That is what we take up next.

Outline

Four Ways the Fit Can Fail

The estimator is well defined and the objective is convex. That does not mean the answer is trustworthy. Four distinct failures, each with its own remedy:

collinearity	the minimizer is not unique, or barely so
high dimension	with $p > n$ collinearity is automatic
outliers	a few observations dominate the fit
wrong flexibility	the model is too rigid, or too free

The first two are the same failure: high dimension is the case where collinearity is unavoidable.

Collinearity: The Minimizer Stops Being Unique

Say $x_2 = 2x_1$. Take $v = (2, -1)^\top$, so $Xv = 0$ and hence

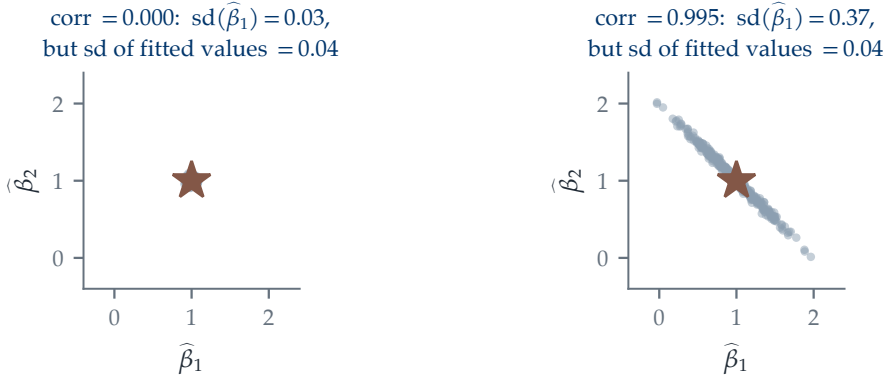
$$v^\top (X^\top X) v = \|Xv\|_2^2 = 0 \quad \Longrightarrow \quad X^\top X \text{ is not invertible.}$$

Then $X(\hat{\beta} + cv) = X\hat{\beta}$ for every c: a whole line of coefficient vectors fits equally well.

High dimension is the extreme case. If $p > n$ then $\operatorname{rank}(X) \leq n < p$, so such a v always exists and the solution is never unique.

In practice columns are rarely exactly dependent. What happens instead is that $\lambda_{\min}(X^\top X)$ is very small: $X^\top X$ is invertible, but only barely, and the same instability appears in degree. Ridge regression replaces $X^\top X$ by $X^\top X + \lambda I$, lifting every eigenvalue by λ; Lecture 4 develops this.

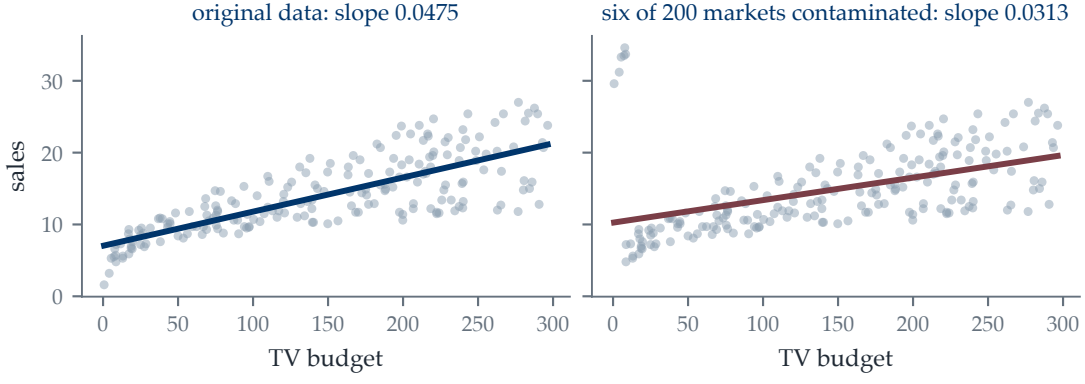
Near Dependence Makes Coefficients Unstable



Refitting on 300 fresh draws of the noise. With correlated features the estimates scatter along a ridge—their spread grows tenfold—while the fitted values stay just as stable. The data pins down the sum of the two effects, not the split between them.

Simulation, $n = 200$, $\beta^\star = (1, 1)$, noise sd 0.5, seed 7.

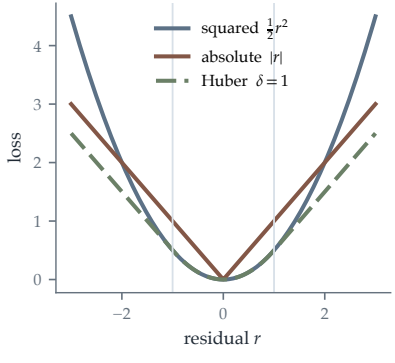
Outliers: A Few Points Take Over



Shifting the response in six of the 200 markets swings the least-squares slope from 0.0475 to 0.0313. Squaring makes one large residual enormously expensive, so the line moves to accommodate a handful of points.

`Advertising` data, six responses shifted by +28 thousand units.

Outliers and the Huber Loss



Squared loss charges r^2: a residual ten times the typical size costs a hundred times as much, so least squares tilts the line toward an outlier.

The **Huber loss** keeps the squared cost near zero but charges only linearly beyond a threshold δ:

$$\ell_\delta(r) = \begin{cases} \frac{1}{2}r^2, & |r| \le \delta, \\ \delta(|r| - \frac{1}{2}\delta), & |r| > \delta. \end{cases}$$

Least absolute deviation, $\sum_i |r_i|$, is another standard choice; the homework looks at both.

The Price of Each Parameter

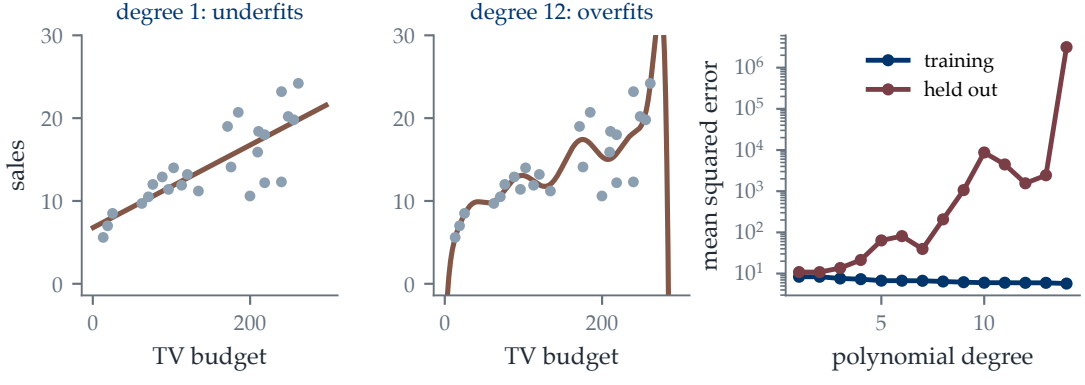
With $y = X\beta^\star + \varepsilon$, $\mathrm{Var}(\varepsilon) = \sigma^2 I$, and X of full column rank d, the in-sample prediction error of least squares is exactly

$$\frac{1}{n}\,\mathbb{E}\|X(\hat{\beta} - \beta^\star)\|_2^2 \;=\; \sigma^2\,\frac{d}{n}.$$

Every parameter costs σ^2/n. The error falls as n grows and rises in proportion to the number of features—so a richer model is not free.

With $d \geq n$ the bound is vacuous, and indeed $\hat{\beta}$ is then not unique. The derivation is in the appendix.

Wrong Flexibility: Underfitting and Overfitting



A straight line cannot bend to the data; a degree-12 polynomial passes near every training point and behaves wildly between them. Training error falls with degree throughout, while held-out error turns around.

`Advertising`: sales on TV, 25 training markets, 175 held out, seed 1.

What Each One Calls For

	what it means	what helps
underfitting	$\mathcal{F}$ is too small to contain anything close to m	enrich $\mathcal{F}$; more data does not help
overfitting	$\mathcal{F}$ is rich enough to chase the noise in these n points	shrink $\mathcal{F}$, or collect more data

These are two sides of one ratio: how rich the class is relative to how much data there is. Shrinking $\mathcal{F}$ and enlarging n move it the same way.

"Held out" means fresh observations not used for fitting. They are the only way to tell the two apart, since training error falls in both cases.

Back to the Advertising Example

What have we learned about the firm's question?

- TV and radio carry real predictive value; newspaper does not, once the other two are present.
- Using all three cut the typical miss from about 3,240 units to about 1,670.
- Per \$1000, the coefficients are 0.0458 for TV and 0.1885 for radio—as associations across markets, radio looks about four times as productive.

But the firm asked how to split a new budget: a question about intervening. These coefficients describe markets as they were.

What the Model Does Not Support

1. **Causal claims.** Budgets were not randomly assigned. Markets with larger budgets may differ systematically.
2. **Extrapolation.** TV budgets in the data run from \$700 to \$296,400. The line says nothing outside that range.
3. **Additivity.** Adding the product $x_{\text{TV}} \cdot x_{\text{radio}}$ cuts the empirical risk from 2.78 to 0.87: the channels reinforce each other.
4. **Constant effect.** One slope per channel assumes the same increment everywhere.

ISLP calls the interaction synergy. Lecture 6 introduces model classes that can express it.

Summary: Target, Model, and Objective

Setup. $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$. **Model**, **loss**, **optimizer** are three separate choices.

Target. The best predictor under squared loss is $m(x) = \mathbb{E}[Y \mid X = x]$; the rest is irreducible noise.

Model. We cannot search all functions, so we fix a class. Taking $\mathcal{F}_{\text{lin}}$ is an assumption.

Objective. Minimize $\widehat{R}_n(\beta) = \frac{1}{n}\|y - X\beta\|_2^2$; its stationary point solves $X^\top X\hat{\beta} = X^\top y$, so the residual is orthogonal to every feature column.

Summary: Geometry and Limitations

Projection. $X\hat{\beta}$ is the orthogonal projection of y onto $\mathrm{col}(X)$, so $X^\top(y - X\hat{\beta}) = 0$.

Convexity. $\widehat{R}_n$ is convex, so a stationary point is a global minimum. A unique minimizer also needs X of full column rank.

Collinearity. Nearly dependent columns leave coefficients unstable even when predictions are similar; when $p > n$, the minimizer is not unique without another criterion.

Claim boundary. Prediction associations do not establish causal effects, extrapolation behavior, or nonlinear interactions.

Next Lecture

Several coefficient vectors can fit nearly the same predictions when features are strongly correlated or numerous.

Lecture 4 — Regularized Regression and Model Selection adds ridge and lasso penalties, chooses their strength by cross-validation, and distinguishes predictive stability, sparsity, and coefficient stability.

Reading for this lecture: ISLP Chapter 3; D2L Chapter 3.

Appendix: Where $\sigma^2 d/n$ Comes From

derive

Take $y = X\beta^\star + \varepsilon$ with $\mathbb{E}\varepsilon = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2 I$, and X of full column rank d. Then $\hat{\beta} = (X^\top X)^{-1} X^\top y$ gives

$$X(\hat{\beta} - \beta^\star) = H\varepsilon, \qquad H = X(X^\top X)^{-1} X^\top,$$

the projection onto $\mathrm{col}(X)$. Because H is a projection, $H^\top H = H^2 = H$, so

$$\mathbb{E}\|H\varepsilon\|_2^2 = \mathbb{E}[\varepsilon^\top H^\top H \varepsilon] = \sigma^2 \operatorname{tr}(H) = \sigma^2 d,$$

the last step because the trace of a rank-d projection is d. Dividing by n gives $\sigma^2 d/n$.

Two facts carry the argument: a projection satisfies $H^2 = H = H^\top$, and $\mathbb{E}[\varepsilon^\top A \varepsilon] = \sigma^2 \operatorname{tr}(A)$ when $\mathrm{Var}(\varepsilon) = \sigma^2 I$.