Predictive Modeling

Review The best function of X to predict Y in terms of the MSE is $f(X) = \mathbb{E}(Y|X)$.

Strategies

(a) Model the joint distribution of $(X, Y) \Rightarrow \mathbb{E}(Y|X)$.

(b) Model the conditional distribution of $Y|X$ directly, often

$$Y|X \sim \mathrm{N}\left(\mathbb{E}(Y|X), \sigma^2\right),$$

so the "only component left" to model is $\mathbb{E}(Y|X)$ itself.

Notation Will begin to use lower-case letters for RVs as well; need upper-case letters for matrices, etc.

Natural Starting Point

For $y \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^p$, model

$$\mathbb{E}(y|\boldsymbol{x}) = \alpha + \boldsymbol{\beta}^\top \boldsymbol{x}$$

as a linear function of $\boldsymbol{x}$.

Fact [Example 3.5 (p. 37)] If $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{p+q}$ is modelled jointly as a (multivariate) normal distribution with

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{bmatrix},$$

then it can be derived that

$$\mathbb{E}(\boldsymbol{y}|\boldsymbol{x}) = \boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_x)$$

is linear in $\boldsymbol{x}$, and $\mathbb{V}\text{ar}(\boldsymbol{y}|\boldsymbol{x}) = \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ "regardless of" $\boldsymbol{x}$.

Random Vectors

For $\boldsymbol{z} \in \mathbb{R}^m$,

$$\mathbb{E}(\boldsymbol{z}) = \begin{bmatrix} \mathbb{E}(z_1) \\ \mathbb{E}(z_2) \\ \vdots \\ \mathbb{E}(z_m) \end{bmatrix}_{m \times 1}$$

and

$$\mathbb{V}\mathrm{ar}(\boldsymbol{z}) = \begin{bmatrix} \mathbb{V}\mathrm{ar}(z_1) & \mathbb{C}\mathrm{ov}(z_1, z_2) & \cdots & \mathbb{C}\mathrm{ov}(z_1, z_m) \\ \mathbb{C}\mathrm{ov}(z_2, z_1) & \mathbb{V}\mathrm{ar}(z_2) & \cdots & \mathbb{C}\mathrm{ov}(z_2, z_m) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}\mathrm{ov}(z_m, z_1) & \mathbb{C}\mathrm{ov}(z_m, z_2) & \cdots & \mathbb{V}\mathrm{ar}(z_m) \end{bmatrix}_{m \times m}.$$

Linear Regression

$$y_i = \overset{\beta_0}{\underset{\downarrow}{\alpha}} + \underset{\substack{\uparrow \\ \mathbb{R}^{p-1}}}{\boldsymbol{\beta}^\top \boldsymbol{x}_i} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2), \quad i = 1, 2, ..., n$$

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\substack{n \times 1 \\ \downarrow \\ \boldsymbol{y}}} = \underbrace{\begin{bmatrix} 1 & \boldsymbol{x}_1^\top \\ 1 & \boldsymbol{x}_2^\top \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_n^\top \end{bmatrix}}_{\substack{n \times p \\ \downarrow \\ \boldsymbol{X}}} \underbrace{\begin{bmatrix} \beta_0 \\ \boldsymbol{\beta} \end{bmatrix}}_{\substack{p \times 1 \\ \downarrow \\ \boldsymbol{\beta}}} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{\substack{n \times 1 \\ \downarrow \\ \boldsymbol{\varepsilon}}}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$$

$$\Rightarrow \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Remarks (i) Notice that $\boldsymbol{\beta}^\top \boldsymbol{x}_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$. (ii) Often write compactly where each $\boldsymbol{x}_i$ implicitly contains a "1".

Two Initial Questions

(a) How to estimate the (vector) parameter $\boldsymbol{\beta}$?

- Answer: lease squares; maximum likelihood.

(b) How good is the estimator $\widehat{\boldsymbol{\beta}}$?

- Answer: mean-squared error; bias-variance analysis.

Least Squares (LSQ)

Calculus	Geometry
$\min_{\boldsymbol{\beta}} \quad \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$ $\Rightarrow \frac{d}{d\boldsymbol{\beta}}(\cdots) = 0$ $\Rightarrow \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = 0$	$\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \perp \boldsymbol{X}$ $\Rightarrow \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = 0$

$$\boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{X}^\top \boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$$

Bias-Variance Analysis

Tools

$$\mathbb{E}(aX+b) = a\mathbb{E}(X)+b \quad \Rightarrow \quad \mathbb{E}(\boldsymbol{A}\boldsymbol{y}+\boldsymbol{b}) = \boldsymbol{A}\mathbb{E}(\boldsymbol{y})+\boldsymbol{b}$$

$$\mathbb{V}\mathrm{ar}(aX+b) = a^2\mathbb{V}\mathrm{ar}(X) \quad \Rightarrow \quad \mathbb{V}\mathrm{ar}(\boldsymbol{A}\boldsymbol{y}+\boldsymbol{b}) = \boldsymbol{A}\mathbb{V}\mathrm{ar}(\boldsymbol{y})\boldsymbol{A}^\top$$

Analysis

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

$$\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top \underbrace{\mathbb{E}(\boldsymbol{y})}_{\boldsymbol{X}\boldsymbol{\beta}} = \boldsymbol{\beta}$$

$$\mathbb{V}\mathrm{ar}(\widehat{\boldsymbol{\beta}}) = \underbrace{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top}_{\boldsymbol{A}}\underbrace{\mathbb{V}\mathrm{ar}(\boldsymbol{y})}_{\sigma^2\boldsymbol{I}}\underbrace{\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}}_{\boldsymbol{A}^\top} = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$$

Gauss-Markov Theorem

Theorem Suppose $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where

$$\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0} \quad \text{and} \quad \mathbb{V}\text{ar}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}.$$

Let

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

be the least-squares estimator, and let $\widetilde{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{y}$ be another linear and unbiased estimator of $\boldsymbol{\beta}$ s.t. $\mathbb{E}(\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{\beta}$. Then

$$\mathbb{V}\text{ar}(\boldsymbol{\ell}^\top \widehat{\boldsymbol{\beta}}) \leq \mathbb{V}\text{ar}(\boldsymbol{\ell}^\top \widetilde{\boldsymbol{\beta}})$$

for all $\boldsymbol{\ell} \in \mathbb{R}^p$.

Translation LSQ estimator is best (in terms of MSE) among all linear and unbiased estimators. [Will prove but later.]

A Look at Assumptions

Assumptions	LSQ	B-V Analysis
$\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$	$\times$	$\surd$
$\mathbb{V}\text{ar}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$	$\times$	$\surd$
$\boldsymbol{\varepsilon} \sim \text{N}(.,.)$	$\times$	$\times$

Remark Don't automatically trust the software! All regression packages/libraries implement these assumptions and they can be wrong in practice.

Exercise Consider the trivial model $y_i = \beta + \varepsilon_i$, where $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{V}\text{ar}(\epsilon_i) = \sigma^2$ for all i but $\mathbb{C}\text{orr}(\varepsilon_i, \varepsilon_j) = \rho \neq 0$. Find $\mathbb{V}\text{ar}(\widehat{\beta})$ and compare with what the software would produce under the standard assumptions. [For example, try $\sigma^2 = 1$, $n = 1000$ and $\rho = 0.1$.]

LSQ and MLE

Note

$$\|\boldsymbol{z}\|^2 = \sum_i z_i^2 \quad \Rightarrow \quad \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = \sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2,$$

so LSQ $\Leftrightarrow$ MLE under model $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$, $\varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$:

$$\max_{\boldsymbol{\beta}} \quad \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2}{2\sigma^2}\right] =$$

$$\left[\frac{1}{\sqrt{2\pi\sigma^2}}\right]^n \exp\left[-\frac{\sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2}{2\sigma^2}\right]$$

$$\Leftrightarrow \quad \min_{\boldsymbol{\beta}} \quad \sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2.$$

But the LSQ principle itself does NOT depend on the modeling assumption $\varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$.

Simple Linear Regression

Exercise Consider the special case of $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

(a) Show that

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}, \quad \widehat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}.$$

(b) Show that

$$\mathbb{V}\text{ar}(\widehat{\beta}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}.$$

(c) Show that $\mathbb{C}\text{ov}(\bar{y}, \widehat{\beta}_1) = 0$, and that

$$\mathbb{V}\text{ar}(\widehat{\beta}_0) = \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{\sum (x_i - \bar{x})^2}, \quad \mathbb{C}\text{ov}(\widehat{\beta}_0, \widehat{\beta}_1) = -\frac{\sigma^2 \bar{x}}{\sum (x_i - \bar{x})^2}.$$

Remark These formulae contain many interesting insights.

(a)

- the fitted regression line $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta} x$ goes through $(\bar{x}, \bar{y})$
- if data are centered in the sense that

$$x_i \leftarrow x_i - \bar{x} \quad \text{and} \quad y_i \leftarrow y_i - \bar{y},$$

  then $\bar{x} = \bar{y} = 0$ and no intercept term β_0 is needed

- if data are not only centered but also scaled in the sense that

$$\sum x_i^2 = 1 \quad \text{and} \quad \sum y_i^2 = 1,$$

  then

$$|\widehat{\beta}_1| = \left| \sum x_i y_i \right| \le \sqrt{\sum x_i^2 \sum y_i^2} = 1 \quad \text{(why)},$$

  aka "regression to the mean" [Fun Box 1, p. 74]

(b)

Exercise Two researchers are investigating the effect of `age` on `blood pressure` with simple linear regression. Amy collected her data by setting up a desk in a supermarket, and her subjects' ages are

$$\{21, 27, 32, 35, 39, 50, 58, 65, 73, 79\}.$$

Bob collected his data set by visiting a university classroom, and his subjects' ages are

$$\{19, 19, 20, 20, 20, 20, 21, 21, 21, 22\}.$$

Whose estimate of the "age effect on blood pressure" will be more accurate? Why?

(c)

Prediction Given a new observation x_0, predict $\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$. How accurate is the prediction?

$$\begin{aligned}
\mathbb{V}\text{ar}(\widehat{y}_0) &= \mathbb{V}\text{ar}(\widehat{\beta}_0 + \widehat{\beta}_1 x_0) \\
&= \mathbb{V}\text{ar}(\widehat{\beta}_0) + x_0^2 \mathbb{V}\text{ar}(\widehat{\beta}_1) + 2x_0 \mathbb{C}\text{ov}(\widehat{\beta}_0, \widehat{\beta}_1) \\
&= \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{\sum (x_i - \bar{x})^2} + \frac{\sigma^2 x_0^2}{\sum (x_i - \bar{x})^2} - \frac{2x_0 \bar{x} \sigma^2}{\sum (x_i - \bar{x})^2} \\
&= \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]
\end{aligned}$$

Lesson Prediction most accurate when $x_0 \approx \bar{x}$, and progressively less accurate as x_0 moves away from center of training data.

Important Take-Home Messages

- You can always use your model — any model — to make a prediction, but you shouldn't always trust it.
- Not all predictions made by the same model are equal; some are more reliable than others.
- Extrapolation is dangerous in general.

$$\Downarrow$$

applicable in general; NOT just for simple linear regression