Prediction

MIT 18.655

Dr. Kempthorne

Spring 2016

Prediction Problems

Targets of Prediction

- Change in value of portfolio over fixed holding period.
- Long-term interest rate in 3 months
- Survival time of patients being treated for cancer
- Liability exposures of a drug company
- Sales of a new prescription drug
- Landfall zone of developing hurricane
- Total snowfall for next winter season
- First-year college grade point average given SAT test scores

General Setup

- Random Variable Y: response variable (target of prediction).
- Random Vector $Z = (Z_1, Z_2, \ldots, Z_p)$: explanatory variables
- Joint distribution: $(Z, Y) \sim P_\theta, \theta \in \Theta$.

Prediction Problem

General Setup (continued)

- Predictor function: $g(Z) \in \{g(\cdot) : \mathcal{Z} \to \mathcal{R}\}$

  $\mathcal{Z}$ = sample space of explanatory-variables vector Z

  $\mathcal{R}$ = sample space of response variable Y.

- Performance Measures

  - Mean Squared Prediction Error

    $$MSPE(g(Z)) = E[(Y - g(Z))^2]$$

  - Mean Absolute Prediction Error

    $$MAPE(g(Z)) = E[|Y - g(Z)|]$$

  where $E[\cdot]$ is expectation under joint distribution of (Z, Y).

- Classes of possible predictor functions

  - Non-parametric class $\mathcal{G}_{NP} = \{g : \mathcal{R}^p \to \mathcal{R}\}$
  - Linear-predictor class

    $\mathcal{G}_L = \{g : g(z) = a + \sum_{j=1}^{p} b_j Z_j,$ for fixed $a, b_1, \ldots, b_p \in \mathcal{R}\}$

Optimal Predictors

Case 1: No Covariates

- With no covariates, $g(Z) = c$, a constant

Lemma 1.4.1 Suppose $EY^2 < \infty$. Then
(a) $E(Y-c)^2 < \infty$ for all c
(b) $E(Y-c)^2$is minimized uniquely by $c = \mu = E(Y)$.
(c) $E(Y-c)^2) = Var(Y) + (\mu - c)^2$

Proof

(a): See Exercise 1.4.25. Hint: Whatever Y and c:
$$\frac{1}{2}Y^2 - c^2 \le (Y-c)^2 \le 2(Y^2 + c^2)$$

(b): $E(Y^2) < \infty \Longrightarrow \mu$ exists.
$$E[(Y-c)^2] = E[Y^2] - 2cE[Y] + c^2 = f(c)$$
$f(c)$ is a concave-up parabola in c
with minimum at $c = E[Y]$

(c): $E[(Y-\mu)^2] = E[Y^2] - \mu^2 = Var(Y)$

Optimal Predictors

Case 2: Covariates Z

- Find the function g that minimizes $E[(Y-g(Z))^2]$

Theorem 1.4.1 If Z is any random vector and Y is any random variable and $\mu(Z)=E[Y \mid Z]$, then either

(a). $E(Y-g(Z))^2) = \infty$ for every function g or

(b). $E(Y-\mu(Z))^2 \leq E(Y-g(Z))^2$ for every g and

- Strict inequalty holds unless $g(Z)=\mu(Z)$
- $\mu(Z)=E[Y \mid Z]$ is unique best MSPE predictor.
- $E(Y-g(Z))^2) = E(Y-\mu(Z))^2 + E(g(Z)-\mu(Z))^2$

Proof By substitution theorem for cond. expectations $(B.1.16)$

$$E[(Y-g(Z))^2 \mid Z=z] = E[(Y-g(z))^2 \mid Z=z]$$

for any function $g(\cdot)$. By Lemma 1.4.1, since $g(z)$ is a constant

$$E(Y-g(z))^2 \mid Z=z) = E((Y-\mu(z))^2 \mid Z=z) + (g(z)-\mu(z))^2$$

Result (b) follows by $B.1.20$ taking expectations of both sides

Optimal Predictors

By Theorem 1.4.1 If $E(Y^2) < \infty$ then

$$E(Y - g(Z))^2 = E(Y - \mu(Z))^2 + E(g(Z) - \mu(Z))^2$$

where $\mu(Z) = E[Y \mid Z]$

Special Case: $g(z) \equiv \mu = E(Y)$ (no dependence on z)

$$E(Y - \mu)^2 = E(Y - \mu(Z))^2 + E(\mu - \mu(Z))^2$$

i.e.,

$$Var(Y) = E(Var(Y \mid Z)) + Var(E(Y \mid Z))$$

Definition: Random variables U and V with $E[UV] < \infty$ are **uncorrelated** if $E\left([V - E(V)][U - E(U)]\right) = 0$

General Prediction Problem

- Predict Y given $Z = z$ using the joint distribution of (Z, Y).
- Let $\mu(Z) = E(Y \mid Z)$ be predictor of Y
- Let $\epsilon = Y - \mu(Z)$ be random prediction error

  $$Y = \mu(Z) + \epsilon$$

Prediction

General Prediction Problem (again)

- Predict Y given $Z = z$ using the joint distribution of (Z, Y).
- Let $\mu(Z) = E(Y \mid Z)$ be predictor of Y
- Let $\epsilon = Y - \mu(Z)$ be random prediction error
$$Y = \mu(Z) + \epsilon$$

Proposition 1.4.1 Suppose that $Var(Y) < \infty$, then

(a) ϵ is uncorrelated with every function of Z

(b) $\mu(Z)$ and ϵ are uncorrelated

(c) $Var(Y) = Var(\mu(Z)) + Var(\epsilon)$

Proof (a). Let $h(Z)$ be any function of Z, then

$$\begin{aligned} E\{h(Z)\epsilon\} &= E\{E[h(Z)\epsilon \mid Z]\} \\ &= E\{h(Z)E[Y - \mu(Z) \mid Z]\} = 0 \end{aligned}$$

(b) follows from (a), and (c) follows from (a) given $Y = \mu(Z) + \epsilon$

Prediction

Theorem 1.4.2 If $E(|Y|) < \infty$ but Z and Y are arbitrary random variables, then

$$Var(E(Y \mid Z)) \leq Var(Y).$$

If $Var(Y) < \infty$ then strict inequality holds unless

$$Y = E(Y \mid Z), \text{ i.e., } Y \text{ is a function of } Z.$$

Proof Recall the special case of Theorem 1.4.1

Special Case: $g(z) \equiv \mu = E(Y)$ (no dependence on z)

$$E(Y - \mu)^2 = E(Y - \mu(Z))^2 + E(\mu - \mu(Z))^2$$

i.e.,

$$Var(Y) = E(Var(Y \mid Z)) + Var(E(Y \mid Z))$$

The first part follows immediately. The second part follows iff

$$E(Var(Y \mid Z)) = E(Y - E(Y \mid Z))^2 = 0.$$

Prediction Example

Example 1.4.1 Assembly line operating at varying capacity, month-by-month. Every day, the assembly line is susceptible to shutdowns due to mechanical failure.

- Z = capacity state, $Z \in \{\frac{1}{4}, \frac{1}{2}, 1\}$ (fraction of full capacity)
- Y: number of shutdowns on a given day
  sample space $\mathcal{Y} = \{0, 1, 2, 3]$
- Joint distribution of (Z, Y) given by the pmf function:

$z \backslash y$	$p(z,y) = P(Z = z, Y = y)$ 0	1	2	3	$p_Z(z)$
$\frac{1}{4}$	0.10	0.05	0.05	0.05	0.25
$\frac{1}{2}$	0.025	0.025	0.10	0.10	0.25
1	0.025	0.025	0.15	0.30	0.50
$p_Y(y)$	0.15	0.10	0.30	0.45	1.00

Note: marginal pmf of Z (Y) given by row (col) sums

Prediction Example

- $p_Z(z)$ gives marginal distribution of capacity states
- $p_Y(y)$ gives marginal distribution of the number of failures/shutdowns per day.

Goal: Predict the number of failures per day given the capacity state of the assembly line for the month.

Solution: The best MSPE predictor function is $E[Y \mid Z]$
Using the joint distribution for (Z, Y) we can compute:

$$\mu(z) = E[Y \mid Z = z] = \sum_{y=0}^{3} y p(z,y) / \sum_{y=0}^{3} p(z,y)$$
$$= \begin{cases} 1.20, & if \quad Z = \frac{1}{4}, \\ 2.10, & if \quad Z = \frac{1}{2}, \\ 2.45, & if \quad Z = 1 \end{cases}$$

Prediction Example

Two ways to compute the MSPE of $\mu(z)$:

$$E[Y - E(Y \mid Z)]^2 = \sum_x \sum_{y=0}^{3} (y - \mu(z))^2 p(z, y) = 0.088625$$

or

$$\begin{aligned} E[Y - E(Y \mid Z)]^2 &= Var(Y) - Var(E(Y \mid Z)) \\ &= E(Y^2) - E[(E(Y \mid Z))^2] \\ &= \sum_y y^2 p_Y(y) - \sum_z E[(Y \mid Z = z)]^2 p_Z(z) \\ &= 0.088625 \end{aligned}$$

Regression Toward the Mean

Bivariate Normal Distribution (See Section B.4)

- $\begin{bmatrix} Z \\ Y \end{bmatrix} \sim N_2\left(\begin{bmatrix} \mu_Z \\ \mu_Y \end{bmatrix}, \Sigma\right)$

  where $E\begin{bmatrix} Z \\ Y \end{bmatrix} = \begin{bmatrix} \mu_Z \\ \mu_Y \end{bmatrix}$

  and $\Sigma = \begin{bmatrix} Cov(Z,Z) & Cov(Z,Y) \\ Cov(Y,Z) & Cov(Y,Y) \end{bmatrix} = \begin{bmatrix} \sigma_Z^2 & \rho\sigma_Z\sigma_Y \\ \rho\sigma_Z\sigma_Y & \sigma_Y^2 \end{bmatrix}$
- Conditional Distribution
  $$Y \mid Z = x \sim N(\mu_Y + \rho(\sigma_Y/\sigma_Z)(z - \mu_Z), \sigma_Y^2(1-\rho^2).$$
- Best Predictor of Y given Z: $\mu(z) = E[Y \mid Z = z]$
  $$\mu(z) = \mu_Y + \rho(\sigma_Y/\sigma_Z)(z - \mu_Z)$$
  "Regression toward the mean"
- MSPE of $\mu(z)$:
  $$\text{MSPE} = E[(Y - \mu(Z))^2] = \sigma_Y^2(1-\rho^2)$$

Bivariate Normal: Bivariate Regression

- Special Cases:
  - $\rho = 1$: Y is perfectly predicted given Z:
    $$\mu(Z) = \mu_Y + \rho(\sigma_Y/\sigma_Z)(z - \mu_z).$$
  - $\rho = 0$: Best predictor of Y is its mean:
    $$\mu(Z) = \mu_Y \text{ (constant, independent of } Z\text{)}$$
- Measure of dependence of Y on Z:
  $$\rho^2 = 1 - \frac{MSPE}{\sigma_Y^2}$$
  Ranges from 0 (no dependence) to 1 (if $\rho = +1$ or -1)
- Galton: studied distributions of heights for fathers and sons.
  Will taller parents have taller children?

Multivariate Normal Distribution

Joint Distribution of (Z, Y) is

$$\begin{bmatrix} Z \\ Y \end{bmatrix} \sim N_{d+1}\left(\begin{bmatrix} \mu_Z \\ \mu_Y \end{bmatrix}, \Sigma \right) \text{ where}$$

- Z is now d-variate
  $$Z = (Z_1, Z_2, \ldots, Z_d)^T$$
- Scalar μ_Z is now a vector: $\mu_Z = (\mu_1, \mu_2, \ldots, \mu_d)^T$
- The covariance matrix Σ is now of dimension $(d+1) \times (d+1)$:
  $$\Sigma = \begin{pmatrix} \Sigma_{Z,Z} & \Sigma_{ZY} \\ \Sigma_{YZ} & \sigma_{YY} \end{pmatrix}, \text{ where } \sigma_{YY} = \sigma_Y^2 \text{ and}$$
  Σ_{ZZ} is $d \times d$ matrix with $||\Sigma_{ZZ}||_{i,j} = Cov(Z_i, Z_j)$
  $\Sigma_{Z,Y} = \Sigma_{Y,Z}^T = (Cov(Z_1, Y), Cov(Z_2, Y), \ldots, Cov(Z_d, Y))^T$
  See Section B.6 for derivation of density function.

Multivariate Normal Distribution

Conditional Distribution: $[Y \mid Z = z]$. By Theorem B.6.5:

$$Y \mid Z = z \sim N(\mu(z), \sigma_{YY|z})$$

where

- $\mu(Z) = \mu_Y + (Z - \mu_Z)^T \boldsymbol{\beta}$
  with $\boldsymbol{\beta} = \Sigma_{ZZ}^{-1}\Sigma_{ZY}$
- $\sigma_{YY|z} = \sigma_{YY} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}$.

Note:

- $\mu(Z) = E[Y \mid Z]$ is the best predictor of Y
- The MSPE of $\mu(Z)$ is
$$\begin{aligned} MSPE &= E\left\{E[Y - \mu(Z)]^2 \mid Z]\right\} = E(\sigma_{YY|z}) \\ &= \sigma_{YY} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY} \end{aligned}$$
- Measure of dependence of Y on Z (analagous to ρ^2)
$$\rho^2_{ZY} = 1 - \frac{MSPE}{\sigma^2_Y}$$
- Terms for ρ^2_{ZY}: "coefficient of determination", "squared multiple-correlation coefficient"

Linear Prediction

Objective: Predict Y Given Z

- Joint distribution of (Z, Y) may be complex
  $\mu(Z) = E[Y \mid Z]$ may be hard to compute
- Alternative: consider class of simple predictors

Linear Predictors: 1-Dimensional Case

- Linear predictor: $g(Z) = a + bZ$, with constants a (intercept) and b (slope).
- Zero-Intercept linear predictor: $g(Z) = a + bZ$ with $a \equiv 0$
- Identify best linear predictors based on MSPE

Linear Prediction

Theorem 1.4.3 Suppose that $E(Z^2)$ and $E(Y^2)$ are finite and Z and Y are not constant. Then

(a). The unique best zero-intercept linear predictor is obtained by taking

$$b = b_0 = \frac{E(ZY)}{E(Z^2}$$

(b). The unique best linear predictor is

$$\mu_L(Z) = a_1 + b_1 Z, \text{ where}$$

$$b_1 = \frac{Cov(Z,Y)}{Var(Z)}, \text{ and}$$

$$a_1 = E(Y) - b_1 E(Z).$$

Proof (a). $E[(Y - bZ)^2] = E[Y^2] - 2bE[ZY] + b^2 E[Z^2] = h(b)$.
$h(b)$ is a parabola in b: achieves minimum when $h'(b) = 0$, i.e.,

$$-2E[ZY] + 2bE[Z^2] = 0 \Longrightarrow b = \frac{E(ZY)}{E(Z^2)}$$

In this case: $MSPE = E(Y - b_0 Z)^2 = E(Y^2) - \frac{[E(ZY)]^2}{E(Z^2)}$

Proof (b). By Lemma 1.4.1

$$E(Y-a-bZ)^2 = Var(Y-bZ) + [E(Y) - bE(Z) - a]^2$$

For any fixed value of b, this is minimized by taking

$$a = E(Y) - bE(Z).$$

Substituting for a, we find b minimizing

$$\begin{aligned} E(Y-a-bZ)^2 &= E([Y-E(Y)] - b[Z-E(Z)])^2 \\ &= E[Y-E(Y)]^2 + b^2E[Z-E(Z)]^2 \\ &\quad -2bE([Z-E(Z)][Y-E(Y)]) \\ &= Var(Y) - 2bCov(Z,Y) + b^2Var(Z) = h_*(b) \end{aligned}$$

$h_*(b)$ is a parabola in b which is minimized when $h_*'(b) = 0$

$$-2bCov(Z,Y) + 2bVar(Z) = 0 \Longrightarrow b = b_1 = \frac{Cov(Z,Y)}{Var(Z)}$$

In this case: $MSPE = E[Y - a_1 - b_1Z]^2 = Var(Y) - \frac{[Cov(ZY)]^2}{Var(Z)}$

Linear Prediction

Notes

- If the best predictor is linear ($E(Y \mid Z)$ is linear in Z) it must coincide with the best linear predictor.
- If the best predictor is non-linear ($E(Y \mid Z)$ is not linear in Z) then the best linear predictor will not have optimal MSPE.
  See Example 1.4.1

Multivariate Linear Predictor For (Z, Y), where $Z = (Z_1, \ldots, Z_d)^T$ is d-dimensional covariate vector, linear predictors of Y are given by

$$\mu_L(Z) = a + \sum_{j=1}^{d} b_j Z_j = a + Z^T \mathbf{b}$$

where $\mathbf{b} = (b_1, b_2, \ldots, b_d)^T$

Linear Prediction

Definition/Notation:

- $E(Y) = \mu_Y$, (scalar) $\mu_Z = E(Z)$ (column d-vector)
- $\Sigma_{ZZ} = E([Z - E(Z)][Z - E(Z)]^T)$ ($d \times d$ matrix)
- $\Sigma_{ZY} = E([Z - E(Z)][Y - E(Y)]$ (column d-vector)

Theorem 1.4.4 If $EY^2 < \infty$ and Σ_{ZZ}^{-1} exists, then the unique best linear MSPE predictor is

$$\mu_L(Z) = \mu_Y + (Z - \mu_Z)^T \boldsymbol{\beta} \text{ where } \boldsymbol{\beta} = \Sigma_{ZZ}^{-1}\Sigma_{ZY}.$$

Proof The MSPE of the linear predictor μ_L is

$MSPE = E_P[Y - \mu_L(Z)]^2$, where P is the joint distribution of $X = (Z^T, Y)^T$. This expression depends only on the first and second moments of X, equivalently $\mu = E[X]$, and $\Sigma = Cov(X)$.

If the distribution P were P_0, the multivariate normal distribution with this expectation and covariance, then $MSPE$ is minimized by $E_{P_0}[Y \mid Z] = \mu_Y + (Z - \mu_Z)^T \boldsymbol{\beta} = \mu_L(Z)$. Since P and P_0 have the same μ and Σ, if μ_L is best MSPE for P_0 it is also best for P.

Linear Prediction

- Defining the *multiple correlation coefficient* or *coefficient of determination*
  $$\rho^2_{ZY} = Corr^2(Y, \mu_L(Z))$$
- **Remark 1.4.4** Suppose the model for $\mu(Z)$ is linear:
  $$\mu(Z) = E(Y \mid Z) = \alpha + Z^T\boldsymbol{\beta}$$
  for unknown $\alpha \in R$, and $\boldsymbol{\beta} \in R^d$.
  Solving for α and $\boldsymbol{\beta}$ minimizing
  $$MSPE = E[Y - \mu(Z)]^2$$
  is solving for parameters minimizing a quadratic form in first/second moments of (Z, Y). These yield the same solution as Theorem 1.4.4.

- **Remark 1.4.5** Consider a Bayesian estimation problem where $X \sim P_\theta$ and $\theta \sim \pi$, and the loss function is squared-error loss:
  $$L(\theta, a) = (a - \theta)^2.$$
  Identify Y with θ, and X with Z, then the Bayes risk of an estimator $\delta(X)$ of θ is:
  $$r(\delta) = E[(\theta - \delta(X))^2] = MSPE(\delta) \text{ which is}$$
  minimized by $\delta(X) = E[\theta \mid X]$.
- **Remark 1.4.6** Connections to Hilbert Spaces (Sectin B.10)
  - Space $\mathcal{H}$ with inner product $< \cdot, \cdot >: \mathcal{H} \times \mathcal{H} \rightarrow R$.
    (bilinear, symmetric, and $< h, h >= 0$ iff $h = 0$)
  - $||h||^2 =< h, h >$ is a norm
    $||ch|| = |c| \cdot ||h||$ for scalar c, and
    $||h_1 + h_2|| \leq ||h_1|| + ||h_2||$ (triangle inequality)
  - $\mathcal{H}$ is *complete*: (contains limits)
    If $\{h_m, m \geq 1\}$: $||h_m - h_n|| \rightarrow 0$, as $m, n \rightarrow \infty$ then there exists $h \in \mathcal{H}$: $||h_n - h|| \rightarrow 0$.

Connections to Hilbert Spaces (continued)

- Projections on Linear Spaces
  - $\mathcal{L} \subset \mathcal{H}$, a closed linear subspace of $\mathcal{H}$.
  - Project operator $\Pi(\cdot \mid \mathcal{L}) : \mathcal{H} \to \mathcal{L}$:
    $\Pi(h \mid \mathcal{L}) = h' \in \mathcal{L}$: achieves $min\{||h - h'||, h' \in \mathcal{L}\}$
    which has the property
    $$h - \Pi(h \mid \mathcal{L}) \perp h', \text{ for all } h' \in \mathcal{L}.$$
  - Π is idempotent ($\Pi^2 = \Pi$).
  - Π is norm-reducing: $||\Pi(h)|| \leq ||h||$
  - From Pythagoras' Theorem:
    $||h||^2 = ||\Pi(h \mid \mathcal{L})||^2 + ||h - \Pi(h \mid \mathcal{L})||^2$

- Hilbert Space Example:
  - $L_2(P) = \{$ All r.v.'s X on a probability space:$EX^2 < \infty\}$
  - $< Z, Y >= E(XY)$
  - If $E(Z) = E(Y) = 0$ and $E(ZY) = 0$, then
    $Var(X + Y) = Var(X) + Var(Y)$ (Pythagoras' Therem)
  - $\mathcal{L}$ is the linear span of $1, Z_1, \ldots, Z_d$
    $$\Pi(Y \mid \mathcal{L}) = E(Y) + (\Sigma_{ZZ}^{-1}\Sigma_{ZY})^T(Z - E(Z)).$$
    See 1.4.14.
  - $\mathcal{L}$ is the space of all $X = g(Z)$ for some g (measurable). This is a linear space that can be shown to be closed and
    $$\Pi(Y \mid \mathcal{L}) = E(Y \mid Z).$$
    See 1.4.6.

Problems

Problem 1.4.4 Determining dependence between random variables.

Problem 1.4.7 Minimizing mean-absolute prediction error – the role of the median.

Problem 1.4.11 Best estimators of Y given Z when (Y, Z) are bivariate normal considering MSPE vs considering mean abolute prediction error.

Problem 1.4.19 Minimizing a convex risk function $R(a, b)$ by solving for (a, b)

Problem 1.4.20 Binomial mixture model.

Problem 1.4.25 Mutual bounding of $E[Y^2]$ and $E(Y - c)^2$.

18.655 Mathematical Statistics
Spring 2016