Math 6266. Fall 2017. Course summary and homework problems.

(Updated on October 15, 2017)

1 Least Squares Estimation

1.1 Basic notions and Gauss-Markov Theorem

Consider the following linear regression model with uncorrelated mean zero noise with equal variance:

$$Y_i = X_i^\top \theta + \varepsilon_i, \text{ for } i = 1 \ldots, n, \quad \mathbf{E}\varepsilon_i = 0, \quad \operatorname{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2 > 0, & i = j, \\ 0, & i \neq j. \end{cases} \tag{1}$$

The regressors or vectors of explanatory variables $X_i \in \mathbf{R}^p$ are determinate (i.e. non-random). $\theta \in \mathbf{R}^p$ is unknown parameter. The noise level $\sigma^2 > 0$ is unknown as well. The matrix-vector form of this model is

$$Y = X^\top \theta + \varepsilon, \tag{2}$$

where $Y \overset{\text{def}}{=} (Y_1, \ldots, Y_n)^\top$, $\varepsilon \overset{\text{def}}{=} (\varepsilon_1, \ldots, \varepsilon_n)^\top$, $\mathbf{E}\varepsilon = (0, \ldots, 0)^\top$, $\operatorname{Var}\varepsilon = \sigma^2 \boldsymbol{I}_n$, and $X \overset{\text{def}}{=} (X_1, \ldots, X_n)$. We assume that matrix $XX^\top = \sum_{i=1}^n X_i X_i^\top$ is invertible.

The *Least Squares Estimator (LSE)* of parameter θ is defined as follows:

$$\begin{aligned} \hat{\theta} &\overset{\text{def}}{=} \operatorname*{argmin}_{\theta} \sum_{i=1}^n (Y_i - X_i^\top \theta)^2 \\ &= \operatorname*{argmin}_{\theta} \|Y - X^\top \theta\|^2 \\ &= (XX^\top)^{-1} XY. \end{aligned}$$

where $\|x\|^2 = x^\top x = \langle x, x \rangle$ for $x \in \mathbf{R}^p$. The LSE $\hat{\theta}$ is unbiased estimator of θ in the model given above. The following theorem proves optimality of the LSE in terms of variance among all linear unbiased estimators of θ

Let $\hat{\beta}$ be some estimator of parameter θ, we will call $\hat{\beta}$ *linear* iff $\hat{\beta} = AY$ for some matrix $A \in \mathbf{R}^{p \times n}$. According to this definition, the LSE $\hat{\theta} = (XX^\top)^{-1} XY$ is a linear estimator.

Introduce the following notation: let $A, B \in \mathbf{R}^{p \times p}$, $A \succcurlyeq B$ denotes $x^\top A x \geq x^\top B x$ for all $x \in \mathbf{R}$, which is equivalent to $A - B$ being positive semidefinite or nonnegative definite.

The theorem below justifies that LSE is BLUE (Best Linear Unbiased Estimator).

Theorem 1.1 (Gauss-Markov)**.** *Assume that matrix $XX^\top$ is invertible (full-rank), then for any unbiased linear estimator $\hat{\beta}$*

$$\operatorname{Var}\hat{\beta} \succcurlyeq \operatorname{Var}\hat{\theta}.$$

Now we consider an estimator of σ^2. Denote the vector of residuals

$$\hat{\varepsilon}^2 \overset{\text{def}}{=} Y - X^\top \hat{\theta},$$

and

$$\hat{\sigma}^2 \overset{\text{def}}{=} \frac{\|Y - X^\top \hat{\theta}\|^2}{n - p} = \frac{\|\hat{\varepsilon}\|^2}{n - p}.$$

$\hat{\sigma}^2$ is unbiased estimator of σ^2.

1.2 LSE under linear constraints

1.3 Geometric interpretation of the LSE

1.3.1 Properties of the projection matrix Π

Matrix $A \in \mathbf{R}^{n\times n}$ corresponds to an orthogonal projection in $\mathbf{R}^n$ iff $A = A^\top$ and $AA = A$. Consider $\Pi \stackrel{\text{def}}{=} X^\top (XX^\top)^{-1} X$.

Lemma 1.1. *1.* $\Pi = \Pi^\top$.

2. $\Pi\Pi = \Pi$ *and* $\Pi(\boldsymbol{I}_n - \Pi) = \mathbf{0}$.

3. $\forall h \in \mathbf{R}^n$ $\|h\|^2 = \|\Pi h\|^2 + \|h - \Pi h\|^2$.

4. $\operatorname{trace}(\Pi) = \dim(\operatorname{image}(\Pi))$, *where* $\operatorname{image}(\Pi) = \{\Pi x : x \in \mathbf{R}^n\}$.

5. Let $\psi_1, \ldots, \psi_p \in \mathbf{R}^n$ *denote the columns of matrix* $X^\top = (\psi_1, \ldots, \psi_p)$. Π *projects* $\mathbf{R}^n$ *on the linear subspace* $L_p \stackrel{\text{def}}{=} \operatorname{span}(\psi_1, \ldots, \psi_p)$*:*

$$\|y - \Pi y\| = \inf_{g \in L_p} \|y - g\| \ \forall y \in \mathbf{R}^n.$$

6. Matrix Π *can be represented as follows* $\Pi = U^\top \Lambda_p U$, *where* $U \in \mathbf{R}^{n\times n}$ *is orthogonal matrix and* $\Lambda_p = \operatorname{diag}\{1, \ldots, 1, 0, \ldots, 0\}$ *with* p *"1"-s on the main diagonal.*

1.4 Orthogonal design

[Optimality of the orthogonal design]

1.5 LSE for the simple linear regression model

Here we consider model (1) with $p = 2$, $X_i = (1, x_i)^\top$ and $\theta = (\theta_1, \theta_2)^\top$:

$$Y_i = \theta_1 + \theta_2 x_i + \varepsilon_i, \ i = 1, \ldots, n.$$

In this case

$$X = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \end{pmatrix}.$$

We assume that $\operatorname{rank}(X) = 2$, i.e. there exist at least two points $x_{i_1} \neq x_{i_2}$. Below we find explicit expressions for the (MLE) estimators in this model. Denote

$$\bar{x} \stackrel{\text{def}}{=} n^{-1} \sum\nolimits_{i=1}^n x_i, \quad \overline{Y} \stackrel{\text{def}}{=} n^{-1} \sum\nolimits_{i=1}^n Y_i,$$

It holds

$$XX^\top = n \begin{pmatrix} 1 & \bar{x} \\ \bar{x} & n^{-1} \sum_{i=1}^n x_i^2 \end{pmatrix}, \quad (XX^\top)^{-1} = \frac{n}{\det(XX^\top)} \begin{pmatrix} n^{-1} \sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix},$$

where

$$\det(XX^\top) = n \sum\nolimits_{i=1}^n x_i^2 - \left(\sum\nolimits_{i=1}^n x_i \right)^2 = n \sum\nolimits_{i=1}^n (x_i - \bar{x}^2) > 0.$$

$$XY = n(\overline{Y}, n^{-1}\sum_{i=1}^{n} x_i Y_i)^\top$$

$$\hat{\theta} = (XX^\top)^{-1}XY = \frac{n^2}{\det(XX^\top)} \begin{pmatrix} n^{-1}\sum_{i=1}^{n} x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} \overline{Y} \\ n^{-1}\sum_{i=1}^{n} x_i Y_i \end{pmatrix} = \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix}.$$

This implies the expressions for $\hat{\theta}_1, \hat{\theta}_2$:

$$\hat{\theta}_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\theta}_1 = \overline{Y} - \bar{x}\hat{\theta}_2.$$

Since $\operatorname{Var}(\hat{\theta}) = \sigma^2(XX^\top)^{-1}$, it holds

$$\begin{aligned} \operatorname{Var}(\hat{\theta}_2) &= \frac{n\sigma^2}{\det(XX^\top)} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x}^2)}, \\ \operatorname{Var}(\hat{\theta}_1) &= \frac{\sigma^2\sum_{i=1}^{n} x_i^2}{\det(XX^\top)} = \frac{\sigma^2\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x}^2)}, \\ \operatorname{Cov}(\hat{\theta}_1, \hat{\theta}_2) &= -\frac{n\sigma^2\bar{x}}{\det(XX^\top)} = -\frac{\sigma^2\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n}(x_i - \bar{x}^2)}. \end{aligned}$$

Now we calculate $\hat{\sigma}^2$:

$$\begin{aligned} (n-p)\hat{\sigma}^2 = \|\hat{\varepsilon}\|^2 &= \sum_{i=1}^{n}(Y_i - \hat{\theta}_1 - \hat{\theta}_2 x_i)^2 \\ &= \sum_{i=1}^{n}(Y_i - \overline{Y} - \hat{\theta}_2(x_i - \bar{x}))^2 \\ &= \sum_{i=1}^{n}(Y_i - \overline{Y})^2 - \hat{\theta}_2^2\sum_{i=1}^{n}(x_i - \bar{x})^2, \end{aligned}$$

in the last equality we used

$$\sum_{i=1}^{n}(Y_i - \overline{Y})(x_i - \bar{x}) = \hat{\theta}_2\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

1.6 Polynomial approximations

[Chebyshev polynomials]

1.7 Other methods of estimation in linear regression model

[GLM, LAD, quantile regression, regression MLE, weighted LSE.]

2 Gaussian Linear Model

(We started this section on September 20, lecture 8.)

Here we consider model (1) with the additional assumption: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

$$Y_i = X_i^\top\theta^* + \varepsilon_i, \text{ for } i = 1\ldots, n, \quad \mathbf{E}\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \text{ i.i.d.,} \tag{3}$$

where $\theta^* \in \mathbf{R}^p$ is unknown parameter of interest. An equivalent definition:

$$Y = X^\top\theta^* + \varepsilon, \; \varepsilon \sim \mathcal{N}(0, \sigma^2\boldsymbol{I}_n).$$

We use the following notation throughout this section:

$$
\begin{aligned}
\hat{\theta} &\stackrel{\text{def}}{=} (XX^\top)^{-1}XY, \\
S_1 &\stackrel{\text{def}}{=} \|Y - X^\top\hat{\theta}\|^2 \\
&= \min_{\theta} \|Y - X^\top\theta\|^2, \\
S_2 &\stackrel{\text{def}}{=} \|Y - X^\top\theta^*\|^2 - \|Y - X^\top\hat{\theta}\|^2 \\
&= \|(XX^\top)^{1/2}(\hat{\theta} - \theta^*)\|^2. \qquad (4)
\end{aligned}
$$

Equality (4) was checked in the proof of Theorem 2.1.

Theorem 2.1. *Assume that matrix $XX^\top$ is invertible, then*

(a) $\hat{\theta}$ and S_1 are independent of each other;

(b) S_1 and S_2 are independent of each other;

(c) $\hat{\theta} - \theta^* \sim \mathcal{N}(0, \sigma^2(XX^\top)^{-1})$;

(d) $S_1/\sigma^2 \sim \chi^2(n-p)$;

(e) $S_2/\sigma^2 \sim \chi^2(p)$.

Corollary 2.1.1. *Let $\hat{\theta}_j$ and θ^*_j denote the j-th coordinates of vectors $\hat{\theta}$ and θ^*; let also a_{jj} correspond to the j-th diagonal element of matrix $(XX^\top)^{-1}$. It holds for all $j = 1, \ldots, p$:*

$$
\frac{\hat{\theta}_j - \theta^*_j}{\sigma\sqrt{a_{jj}}} \sim \mathcal{N}(0,1), \qquad (5)
$$

$$
\sqrt{\frac{n-p}{a_{jj}S_1}}\left(\hat{\theta}_j - \theta^*_j\right) \sim t(n-p). \qquad (6)
$$

$t(n-p)$ denotes Student's t-distribution with $n-p$ degrees of freedom.

Corollary 2.1.2.

$$
\frac{n-p}{p}\frac{S_2}{S_1} = \frac{n-p}{p}\frac{\|(XX^\top)^{1/2}(\hat{\theta} - \theta^*)\|^2}{\|Y - X^\top\hat{\theta}\|^2} \sim F(p, n-p). \qquad (7)
$$

$F(p, n-p)$ denotes Fisher-Snedecor or F-distribution with $p, n-p$ degrees of freedom.

Property (d) from Theorem 2.1 implies that the unbiased estimator of σ^2 can be expressed through χ^2 distribution:

$$
\hat{\sigma}^2 = \frac{\|Y - X^\top\theta^*\|^2}{n-p}, \quad (n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p), \qquad (8)
$$

which allow to construct confidence sets for σ^2.

2.1 Properties of quadratic forms in Gaussian random variables

Consider i.i.d random variables $X_1, \ldots, X_n \sim \mathcal{N}(0,1)$. Let $A = \{a_{i,j}\}_{i,j=1}^n$ be a determinate matrix in $\mathbf{R}^{n \times n}$. Let also A be symmetric, i.e. $A = A^\top$ or $a_{i,j} = a_{j,i}$ $\forall i,j = 1, \ldots, n$. Let Q denote the following quadratic form:

$$Q = Q(X) \stackrel{\text{def}}{=} \sum_{i,j=1}^{n} a_{i,j} X_i X_j = X^\top A X,$$

where $X \stackrel{\text{def}}{=} (X_1, \ldots, X_n)^\top$. Introduce matrix $B = \begin{pmatrix} b_1^\top \\ b_2^\top \\ \ldots \\ b_m^\top \end{pmatrix} \in \mathbf{R}^{m \times n}$, where $b_1, \ldots, b_m \in \mathbf{R}^n$ are some fixed vectors; let also

$$t = t(X) \stackrel{\text{def}}{=} BX \in \mathbf{R}^m,$$

we can write also $t = (t_1, \ldots, t_m)^\top$ for $t_k \stackrel{\text{def}}{=} b_k^\top X$, $k = 1, \ldots, m$.

Lemma 2.1. *If $BA = \mathbf{0}$, then B and t are independent of each other.*

Lemma 2.2. *Let $Q_1 = Q_1(X) \stackrel{\text{def}}{=} X^\top A_1 X$, $Q_2 = Q_2(X) \stackrel{\text{def}}{=} X^\top A_2 X$ for some symmetric matrices $A_1, A_2 \in \mathbf{R}^{n \times n}$. If $A_1 A_2 = A_2 A_1 = \mathbf{0}$, then Q_1 and Q_2 are independent of each other.*

2.2 Confidence sets

Theorem 2.1 and properties (6), (7), (8) allow to construct confidence sets for θ^*, θ_j^* and $\hat{\sigma}^2$.

2.2.1 Simultaneous confidence sets

Theorem 2.2. *Let $q_F(1-\alpha)$ denote the $(1-\alpha)$-quantile of the distribution $F(p, n-p)$, i.e. if a random variable $z_2 \sim F(p, n-p)$, then $\mathbf{P}(z_2 \le q_F(1-\alpha)) = 1-\alpha$. It holds for the linear Gaussian model*

$$\mathbf{P}\left(\forall \gamma \in \mathbf{R}^p \; |\gamma^\top(\hat{\theta} - \theta^*)| \le \sqrt{u(q_F(1-\alpha), \gamma, Y)}\right) = 1 - \alpha, \tag{9}$$

where $u(x, \gamma, Y) \stackrel{\text{def}}{=} x S_1 p (n-p)^{-1} \gamma^\top (XX^\top)^{-1} \gamma$. (Notice that $u(x, \gamma, Y)$ is computable, hence we can find explicit quantiles for $|\gamma^\top(\hat{\theta} - \theta^*)|$.) The result (9) is equivalent to the following simultaneous confidence statement:

$$\mathbf{P}\left(\forall \gamma \in \mathbf{R}^p \; \gamma^\top \hat{\theta} - \sqrt{u(q_F(1-\alpha), \gamma, Y)} \le \gamma^\top \theta^* \le \gamma^\top \hat{\theta} + \sqrt{u(q_F(1-\alpha), \gamma, Y)}\right) = 1 - \alpha.$$

Proof. At first we will derive a simpler (but non-pivotal) version of the statement, and then we will extend it to the final result. We need to employ lemma, which is given after the theorem. The lemma states that for a symmetric positive-definite matrix $B \in \mathbf{R}^{p \times p}$

$$t^\top B t = \max_{\gamma \in \mathbf{R}^p} \frac{(\gamma^\top t)^2}{\gamma^\top B^{-1} \gamma}. \tag{10}$$

Take $B \coloneqq XX^\top$, $t \coloneqq \hat{\theta} - \theta^*$, then (10) and (e) imply for all $x > 0$:

$$\begin{aligned}
&\mathbf{P}\left(\forall\gamma \in \mathbf{R}^p\ \{\gamma^\top(\hat{\theta} - \theta^*)\}^2 \le x\gamma^\top(XX^\top)^{-1}\gamma\right) \\
&= \mathbf{P}\left(\max_{\gamma\in\mathbf{R}^p} \frac{\{\gamma^\top(\hat{\theta} - \theta^*)\}^2}{\gamma^\top(XX^\top)^{-1}\gamma} \le x\right) \\
&= \mathbf{P}\left((\hat{\theta} - \theta^*)^\top XX^\top(\hat{\theta} - \theta^*) \le x\right) \\
&= \mathbf{P}\left(S_2 \le x\right) \\
&= \mathbf{P}\left(z_1 \le \sigma^2 x\right),
\end{aligned}$$

where z_1 is a random variable with distribution $\chi^2(p)$. The last expression depends on σ^2.

Now we derive a *pivotal* result using similar arguments. Let z_2 denote a random variable with distribution $F(p, n-p)$. Properties (10) and (7) imply

$$\begin{aligned}
&\mathbf{P}\left(z_2 \le x\right) \\
&= \mathbf{P}\left(\frac{n-p}{p}\frac{S_2}{S_1} \le x\right) \\
&= \mathbf{P}\left((\hat{\theta} - \theta^*)^\top XX^\top(\hat{\theta} - \theta^*) \le xS_1 p/(n-p)\right) \\
&= \mathbf{P}\left(\max_{\gamma\in\mathbf{R}^p} \frac{\{\gamma^\top(\hat{\theta} - \theta^*)\}^2}{\gamma^\top(XX^\top)^{-1}\gamma} \le xS_1 p/(n-p)\right) \\
&= \mathbf{P}\left(\forall\gamma \in \mathbf{R}^p\ \{\gamma^\top(\hat{\theta} - \theta^*)\}^2 \le xS_1 p(n-p)^{-1}\gamma^\top(XX^\top)^{-1}\gamma\right) \\
&= \mathbf{P}\left(\forall\gamma \in \mathbf{R}^p\ \{\gamma^\top(\hat{\theta} - \theta^*)\}^2 \le u(x,\gamma,Y)\right) \\
&= \mathbf{P}\left(\forall\gamma \in \mathbf{R}^p\ |\gamma^\top(\hat{\theta} - \theta^*)| \le \sqrt{u(x,\gamma,Y)}\right),
\end{aligned}$$

where $u(x,\gamma,Y) = u_{n,p}(x,\gamma,Y) \stackrel{\text{def}}{=} xS_1 p(n-p)^{-1}\gamma^\top(XX^\top)^{-1}\gamma$. Since $\mathbf{P}\left(z_2 \le q_F(1-\alpha)\right) = 1-\alpha$, it holds

$$\mathbf{P}\left(\forall\gamma \in \mathbf{R}^p\ |\gamma^\top(\hat{\theta} - \theta^*)| \le \sqrt{u(q_F(1-\alpha),\gamma,Y)}\right) = \mathbf{P}\left(z_2 \le q_F(1-\alpha)\right) = 1-\alpha.$$

$\square$

Corollary 2.2.1. *Let* $\Gamma \stackrel{\text{def}}{=} \{\gamma_1, \ldots, \gamma_m\} \subset \mathbf{R}^p$ *be a finite set of* p*-dimensional vectors. For example, if* $m \coloneqq p$*, and for all* $j = 1, \ldots, p$ $\gamma_j = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$ *with "1" on* j*-th coordinate only, then* $\gamma_j^\top\theta^* = \theta_j^*$*. Since* $\Gamma \subset \mathbf{R}^p$*, the theorem above implies*

$$\mathbf{P}\left(\forall\gamma \in \Gamma\ |\gamma^\top(\hat{\theta} - \theta^*)| \le \sqrt{u(q_F(1-\alpha),\gamma,Y)}\right) \ge 1-\alpha.$$

The above bound is equivalent to the following simultaneous confidence statement:

$$\mathbf{P}\left(\forall\gamma \in \Gamma\ \gamma^\top\hat{\theta} - \sqrt{u(q_F(1-\alpha),\gamma,Y)} \le \gamma^\top\theta^* \le \gamma^\top\hat{\theta} + \sqrt{u(q_F(1-\alpha),\gamma,Y)}\right) \ge 1-\alpha.$$

Lemma 2.3. *Let* $B \in \mathbf{R}^{p\times p}$ *be a symmetric, positive-definite matrix. Then it holds for any vector* $t \in \mathbf{R}^p$

$$t^\top Bt = \max_{h\in\mathbf{R}^p} \frac{(h^\top t)^2}{h^\top B^{-1}h}.$$

Proof of lemma. Let U be the orthogonal matrix, which diagonalizes B: $B = U\Lambda U^\top$ for $\Lambda = \operatorname{diag}\{\lambda_1, \ldots, \lambda_p\}$. Moreover, all eigenvalues $\lambda_1, \ldots, \lambda_p$ are positive by Lemma's conditions. Denote

$$\begin{aligned}
\Lambda^{1/2} &\stackrel{\text{def}}{=} \operatorname{diag}\{\lambda_1^{1/2}, \ldots, \lambda_p^{1/2}\}, \\
H &\stackrel{\text{def}}{=} U\Lambda^{1/2}, \\
x &\stackrel{\text{def}}{=} H^\top t, \ y \stackrel{\text{def}}{=} H^{-1}h.
\end{aligned}$$

Then it holds:

$$\begin{aligned}
&B = HH^\top, \\
&y^\top x = h^\top H^{\top -1} H^\top t = h^\top t, \\
&x^\top x = t^\top H H^\top t = t^\top B t, \\
&y^\top y = h^\top H^{\top -1} H^{-1} h = h^\top (HH^\top)^{-1} h = h^\top B^{-1} h.
\end{aligned}$$

Furthermore, by Cauchy-Schwarz inequality

$$(y^\top x)^2 \le (y^\top y)(x^\top x), \tag{11}$$

which is equivalent to

$$(h^\top t)^2 \le (h^\top B^{-1} h)(t^\top B t) \tag{12}$$

for any $t, h \in \mathbf{R}^p$. (12) implies:

$$t^\top B t \ge \frac{(h^\top t)^2}{h^\top B^{-1} h} \tag{13}$$

for all $h \in \mathbf{R}^p, h \ne 0$.

Cauchy-Schwarz inequality (11) becomes an equality iff vectors x and y are linearly dependent (or collinear), i.e. $x = cy$ for some constant $c \in \mathbf{R}$, which means that (12) becomes an equality iff $H^\top t = cH^{-1}h$ for some constant $c \in \mathbf{R}$. This implies that bound (12) becomes an equality when $h = c^{-1} H H^\top t = c^{-1} B t$. In this case the right side of bound (13) reads as:

$$\frac{(h^\top t)^2}{h^\top B^{-1} h} = \frac{c^{-2}}{c^{-2}} \frac{(t^\top B t)^2}{t^\top B B^{-1} B t} = \frac{(t^\top B t)^2}{t^\top B t} = t^\top B t,$$

the last expression proves that maximum is attained, and equals to the value from the lemma's statement. This finishes the proof. □

2.3 Sequential testing procedure for selection of a number of covariates

2.4 Hypothesis Testing in linear Gaussian regression

(to be discussed on Monday, October 16).

Further topics

include bootstrap procedures for the linear regression model; ridge regression, projection, and shrinkage, advanced topics in the theory of Gaussian models, topics on high-dimensional statistics.

3 Exercises

(Please note that the numbers of the exercises may change with new versions of the file, since I'll add more exercises.)

Section	Exercises
1.1	1, 2, 3
1.3	4, 5, 6
2.1	7, 8
2.2	9, 10, 11, 12

Exercise 1. Consider the linear regression model with mean zero, uncorrelated, heteroscedastic noise:

$$Y_i = X_i^\top \theta + \varepsilon_i, \text{ for } i = 1 \ldots, n, \quad \mathbf{E}\varepsilon_i = 0, \quad \operatorname{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma_i^2 > 0, & i = j, \\ 0, & i \neq j. \end{cases}$$

Find expressions for the LSE $\hat{\theta}$ and response estimator $\hat{Y}$ in this model.

Exercise 2. Assume that $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ in the previous problem. What is known about the distribution of $\hat{\theta}$ and $\hat{Y}$? Now suppose additionally that $\sigma_i^2 \equiv \sigma^2 > 0$. What can be said about distribution of the estimator $\hat{\sigma}^2$?

Exercise 3. Consider the linear regression model from exercise 1. Suppose, that the target of estimation is $h^\top \theta$ for some determinate non-zero vector $h \in \mathbf{R}^p$. Find expression for the LSE of $h^\top \theta$. Is this estimate optimal in sense of Gauss-Markov theorem, i.e. does it have the smallest variance among all linear unbiased estimators? Study the same issue for the target $\eta \stackrel{\text{def}}{=} H^\top \theta$, where $H \in \mathbf{R}^{q \times p}$ is some non-zero matrix with $q \leq p$.

Exercise 4. Let $A \in \mathbf{R}^{n \times n}$ be a matrix (corresponding to a linear map in $\mathbf{R}^n$). Show that A preserves length for all $x \in \mathbf{R}^n$ iff it preserves the inner product. I.e. one needs to show the following:

$$\|Ax\| = \|x\| \ \forall x \in \mathbf{R}^n \iff (Ax)^\top (Ay) \ \forall x, y \in \mathbf{R}^n.$$

Exercise 5. (a) Let $x_0 \in \mathbf{R}^n$ be some fixed vector, find a projection map on the subspace $\operatorname{span}(x_0)$. Compare your result with matrix Π (from section 1.3) for the case of $p = 1$.
(b) Prove part 3) of Lemma 1.1 for an arbitrary orthogonal projection in $\mathbf{R}^n$.

Exercise 6. Let L_1, L_2 be some subspaces in $\mathbf{R}^n$, and $L_2 \subseteq L_1 \subseteq \mathbf{R}^n$. Let P_{L_1}, P_{L_2} denote orthogonal projections on these subspaces. Prove the following properties:
(a) $P_{L_2} - P_{L_1}$ is an orthogonal projection,
(b) $\|P_{L_2} x\| \leq \|P_{L_1} x\|$ for all $x \in \mathbf{R}^n$,
(c) $P_{L_2} P_{L_1} = P_{L_2}$.

Exercise 7. (a) Using the notation from section 2.1, consider $X \sim \mathcal{N}(\mu, \boldsymbol{I}_n)$ for some $\mu \in \mathbf{R}^n$. Find $\mathbf{E}\{Q(X)\}$ and $\operatorname{Var}\{Q(X)\}$.
(b) Generalize the results from part (a) to the case $X \sim \mathcal{N}(\mu, \Sigma)$ for some positive-definite covariance matrix $\Sigma \in \mathbf{R}^{n \times n}$.

Exercise 8. Let $X \sim \mathcal{N}(0, \boldsymbol{I}_n)$, $Q \stackrel{\text{def}}{=} X^\top X$. Suppose that Q is decomposed into the sum of two quadratic forms: $Q = Q_1 + Q_2$, where $Q_i = X^\top A_i X$, $i = 1, 2$ for some symmetric matrices A_1, A_2 with $\operatorname{rank}(A_1) = n_1$ and $\operatorname{rank}(A_2) = n_2$. Show that if $n_1 + n_2 = n$, then Q_1 and Q_2 are independent and $Q_i \sim \chi^2(n_i)$ for $i = 1, 2$.

Exercise 9. In the Gaussian linear regression model 3, consider the target of estimation $\eta^* \stackrel{\text{def}}{=} H^\top \theta^*$, where $H \in \mathbf{R}^{q \times p}$ is some non-zero matrix with $q \le p$. Find an analogue of the quadratic form S_2 (from (4)) for the new target η^*, and prove for the new quadratic form statements similar to *(e)* from Theorem 2.1, and Corollary 2.1.2.

Exercise 10. (a) Consider model (3) for $p = 2$, $X_i = (1, x_i)^\top$, $\theta^* = (\theta_1^*, \theta_2^*)^\top$ (similarly to section 1.5). Write explicit expressions for the confidence sets for θ^*, θ_1^*, θ_2^*.
(b) Find a confidence interval for the expected response $\mathbf{E}(Y_i)$ in the model in part (a).

Exercise 11. Find an elliptical confidence set for the expected response $\mathbf{E}(Y)$ in model (3).

Exercise 12. Construct simultaneous confidence intervals (e.g., as in Corollary 2.2.1) for the expected responses $\mathbf{E}(Y_1), \ldots, \mathbf{E}(Y_n)$ in model (3).