Regression Analysis

MIT 18.443

Dr. Kempthorne

Spring 2015

Outline

1. Regression Analysis II
   - Distribution Theory: Normal Regression Models
   - Maximum Likelihood Estimation
   - Generalized M Estimation

Marginal Distributions of Least Squares Estimates

Because

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

the marginal distribution of each $\hat{\beta}_j$ is:

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 C_{j,j})$$

where $C_{j.j} = j$th diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$

The Q-R Decomposition of $\mathbf{X}$

Consider expressing the $(n \times p)$ matrix $\mathbf{X}$ of explanatory variables as

$$\mathbf{X} = \mathbf{Q} \cdot \mathbf{R}$$

where

$\mathbf{Q}$ is an $(n \times p)$ orthonormal matrix, i.e., $\mathbf{Q}^T\mathbf{Q} = I_p$.
$\mathbf{R}$ is a $(p \times p)$ upper-triangular matrix.

The columns of $\mathbf{Q} = [\mathbf{Q}_{[1]}, \mathbf{Q}_{[2]}, \ldots, \mathbf{Q}_{[p]}]$ can be constructed by performing the *Gram-Schmidt Orthonormalization* procedure on the columns of $\mathbf{X} = [\mathbf{X}_{[1]}, \mathbf{X}_{[2]}, \ldots, \mathbf{X}_{[p]}]$

If $$\mathbf{R} = \begin{bmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,p-1} & r_{1,p} \\ 0 & r_{2,2} & \cdots & r_{2,p-1} & r_{2,p} \\ 0 & 0 & \ddots & \vdots & \vdots \\ 0 & 0 & & r_{p-1,p-1} & r_{p-1,p} \\ 0 & 0 & \cdots & 0 & r_{p,p} \end{bmatrix}$$, then

- $\mathbf{X}_{[1]} = \mathbf{Q}_{[1]} r_{1,1}$
  $\Longrightarrow$
  $$\begin{aligned} r_{1,1}^2 &= \mathbf{X}_{[1]}^T \mathbf{X}_{[1]} \\ \mathbf{Q}_{[1]} &= \mathbf{X}_{[1]} / r_{1,1} \end{aligned}$$

- $\mathbf{X}_{[2]} = \mathbf{Q}_{[1]} r_{1,2} + \mathbf{Q}_{[2]} r_{2,2}$
  $\Longrightarrow$
  $$\begin{aligned} \mathbf{Q}_{[1]}^T \mathbf{X}_{[2]} &= \mathbf{Q}_{[1]}^T \mathbf{Q}_{[1]} r_{1,2} + \mathbf{Q}_{[1]}^T \mathbf{Q}_{[2]} r_{2,2} \\ &= 1 \cdot r_{1,2} + 0 \cdot r_{2,2} \\ &= r_{1,2} \quad (\text{known since } \mathbf{Q}_{[1]} \text{ specfied}) \end{aligned}$$

- With $r_{1,2}$ and $\mathbf{Q}_{[1]}$ specfied we can solve for $r_{2,2}$:
$\Longrightarrow$
$$\mathbf{Q}_{[2]} r_{2,2} = \mathbf{X}_{[2]} - \mathbf{Q}_{[1]} r_{1,2}$$
Take squared norm of both sides:
$$r_{2,2}^2 = \mathbf{X}_{[2]}^T \mathbf{X}_{[2]} - 2 r_{1,2} \mathbf{Q}_{[1]}^T \mathbf{X}_{[2]} + r_{1,2}^2$$
(all terms on RHS are known)

  With $r_{2,2}$ specified
$\Longrightarrow$
$$\mathbf{Q}_{[2]} = \frac{1}{r_{2,2}} \left[\mathbf{X}_{[2]} - r_{1,2} \mathbf{Q}_{[1]} \right]$$
- Etc. (solve for elements of $\mathbf{R}$, and columns of $\mathbf{Q}$)

With the Q-R Decomposition

$$\mathbf{X} = \mathbf{QR}$$
$$(\mathbf{Q}^T\mathbf{Q} = \mathbf{I}_p, \text{ and } \mathbf{R} \text{ is } p \times p \text{ upper-triangular})$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{y}$$
$$(\text{plug in } \mathbf{X} = \mathbf{QR} \text{ and simplify})$$

$$Cov(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2\mathbf{R}^{-1}(\mathbf{R}^{-1})^T$$

$$H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{QQ}^T$$
$$(\text{giving } \hat{\mathbf{y}} = \mathbf{Hy} \text{ and } \hat{\epsilon} = (\mathbf{I}_n - \mathbf{H})\mathbf{y})$$

More Distribution Theory

Assume $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\{\epsilon_i\}$ are i.i.d. $N(0, \sigma^2)$, i.e.,

$$\begin{array}{rcl} \boldsymbol{\epsilon} & \sim & N_n(\mathbf{0}_n, \sigma^2\mathbf{I}_n) \\ \text{or } \mathbf{y} & \sim & N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n) \end{array}$$

Theorem* For any $(m \times n)$ matrix $\mathbf{A}$ of rank $m \leq n$, the random normal vector $\mathbf{y}$ transformed by $\mathbf{A}$,

$$\mathbf{z} = \mathbf{A}\mathbf{y}$$

is also a random normal vector:

$$\mathbf{z} \sim N_m(\boldsymbol{\mu}_{\mathbf{z}}, \boldsymbol{\Sigma}_{\mathbf{z}})$$

where $\quad \boldsymbol{\mu}_{\mathbf{z}} = \mathbf{A}E(\mathbf{y}) = \mathbf{A}\mathbf{X}\boldsymbol{\beta}$,

and $\quad \boldsymbol{\Sigma}_{\mathbf{z}} = \mathbf{A}Cov(\mathbf{y})\mathbf{A}^T = \sigma^2\mathbf{A}\mathbf{A}^T$.

Earlier, $\mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ yields the distribution of $\hat{\boldsymbol{\beta}} = \mathbf{A}\mathbf{y}$

With a different definition of $\mathbf{A}$ (and $\mathbf{z}$) we give an easy proof of:

Theorem For the normal linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

$$\mathbf{X} \ (n \times p) \text{ has rank } p \text{ and}$$
$$\boldsymbol{\epsilon} \sim N_n(\mathbf{0}_n, \sigma^2 \mathbf{I}_n).$$

(a) $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ and $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ are independent r.v.s

(b) $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$

(c) $\sum_{i=1}^n \hat{\epsilon}_i^2 = \hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}} \sim \sigma^2\chi^2_{n-p}$ (Chi-squared r.v.)

(d) For each $j = 1, 2, \dots, p$

$$\hat{t}_j = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} C_{j,j}} \sim t_{n-p} \ (t\text{− distribution})$$

where
$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^n \hat{\epsilon}_i^2$$
$$C_{j,j} = [(\mathbf{X}^T\mathbf{X})^{-1}]_{j,j}$$

Proof: Note that (d) follows immediately from (a), (b), (c)

Define $\mathbf{A} = \begin{bmatrix} \mathbf{Q}^T \\ \mathbf{W}^T \end{bmatrix}$, where

- $\mathbf{A}$ is an $(n \times n)$ orthogonal matrix (i.e. $\mathbf{A}^T = A^{-1}$)
- $\mathbf{Q}$ is the column-orthonormal matrix in a Q-R decomposition of $\mathbf{X}$

Note: $\mathbf{W}$ can be constructed by continuing the *Gram-Schmidt Orthonormalization* process (which was used to construct $\mathbf{Q}$ from $\mathbf{X}$) with $\mathbf{X}^* = [\ \mathbf{X} \mid \mathbf{I}_n\ ]$.
Then, consider

$$\mathbf{z} = \mathbf{A}\mathbf{y} = \begin{bmatrix} \mathbf{Q}^T\mathbf{y} \\ \mathbf{W}^T\mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{z_Q} \\ \mathbf{z_W} \end{bmatrix} \quad \begin{matrix} (p \times 1) \\ (n-p) \times 1 \end{matrix}$$

The distribution of $\mathbf{z} = \mathbf{A}\mathbf{y}$ is $N_n(\boldsymbol{\mu}_\mathbf{z}, \boldsymbol{\Sigma}_\mathbf{z})$
where

$$\begin{aligned}
\boldsymbol{\mu}_\mathbf{z} &= [\mathbf{A}][\mathbf{X}\beta] = \begin{bmatrix} \mathbf{Q}^T \\ \mathbf{W}^T \end{bmatrix} [\mathbf{Q} \cdot \mathbf{R} \cdot \beta] \\
&= \begin{bmatrix} \mathbf{Q}^T\mathbf{Q} \\ \mathbf{W}^T\mathbf{Q} \end{bmatrix} [\mathbf{R} \cdot \beta] \\
&= \begin{bmatrix} \mathbf{I}_p \\ \mathbf{0}_{(n-p)\times p} \end{bmatrix} [\mathbf{R} \cdot \beta] \\
&= \begin{bmatrix} \mathbf{R} \cdot \beta \\ \mathbf{0}_{(n-p)\times p} \end{bmatrix}
\end{aligned}$$

$$\boldsymbol{\Sigma}_\mathbf{z} = \mathbf{A} \cdot [\sigma^2 \mathbf{I}_n] \cdot \mathbf{A}^T = \sigma^2[\mathbf{A}\mathbf{A}^T] = \sigma^2 \mathbf{I}_n$$
$$\text{since } \mathbf{A}^T = \mathbf{A}^{-1}$$

Thus $z = \begin{pmatrix} \mathbf{z_Q} \\ \mathbf{z_W} \end{pmatrix} \sim N_n\left[\begin{pmatrix} \mathbf{R}\beta \\ \mathbf{O}_{n-p} \end{pmatrix}, \sigma^2\mathbf{I}_n\right]$

$\Longrightarrow$

$$\mathbf{z_Q} \sim N_p[(\mathbf{R}\beta), \sigma^2\mathbf{I}_p]$$
$$\mathbf{z_W} \sim N_{(n-p)}[(\mathbf{O}_{(n-p)}, \sigma^2\mathbf{I}_{(n-p)}]$$

and $\mathbf{z_Q}$ and $\mathbf{z_W}$ are independent.

The Theorem follows by showing

(a*) $\hat{\beta} = \mathbf{R}^{-1}\mathbf{z_Q}$ and $\hat{\epsilon} = \mathbf{W}\mathbf{z_W}$,
(i.e. $\hat{\beta}$ and $\hat{\epsilon}$ are functions of different independent vecctors).

(b*) Deducing the distribution of $\hat{\beta} = \mathbf{R}^{-1}\mathbf{z_Q}$,
applying Theorem* with $\mathbf{A} = \mathbf{R}^{-1}$ and "$\mathbf{y}$" $= \mathbf{z_Q}$

(c*) $\hat{\epsilon}^T\hat{\epsilon} = \mathbf{z_W}^T\mathbf{z_W}$
$= $ sum of $(n-p)$ squared r.v's which are i.i.d. $N(0, \sigma^2)$.
$\sim \sigma^2\chi^2_{(n-p)}$, a scaled Chi-Squared r.v.

Proof of (a*)

$\hat{\beta} = \mathbf{R}^{-1}\mathbf{z_Q}$ follows from

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{y} \quad \text{and}$$

$$\mathbf{X} = \mathbf{QR} \text{ with } \mathbf{Q} : \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_p$$

$$\begin{aligned}
\hat{\epsilon} &= \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\beta} = \mathbf{y} - (\mathbf{QR}) \cdot (\mathbf{R}^{-1}\mathbf{z_Q}) \\
&= \mathbf{y} - \mathbf{Q}\mathbf{z_Q} \\
&= \mathbf{y} - \mathbf{Q}\mathbf{Q}^T\mathbf{y} = (\mathbf{I}_n - \mathbf{Q}\mathbf{Q}^T)\mathbf{y} \\
&= \mathbf{W}\mathbf{W}^T\mathbf{y} \;\; (\text{since } \mathbf{I}_\mathrm{n} = \mathbf{A}^\mathrm{T}\mathbf{A} = \mathbf{Q}\mathbf{Q}^\mathrm{T} + \mathbf{W}\mathbf{W}^\mathrm{T}) \\
&= \mathbf{W}\mathbf{z_W}
\end{aligned}$$

Outline

Maximum-Likelihood Estimation

Consider the normal linear regression model:

$$\begin{aligned} \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}&, \text{ where } \{\epsilon_i\} \text{ are i.i.d. } N(0, \sigma^2), \text{ i.e.,} \\ \boldsymbol{\epsilon} &\sim N_n(\mathbf{0}_n, \sigma^2 \mathbf{I}_n) \\ \text{or } \mathbf{y} &\sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n) \end{aligned}$$

Definitions:

- The **likelihood function** is
  $$L(\boldsymbol{\beta}, \sigma^2) = p(\mathbf{y} \mid \mathbf{X}, \mathbf{B}, \sigma^2)$$
  where $p(\mathbf{y} \mid \mathbf{X}, \mathbf{B}, \sigma^2)$ is the joint probability density function (pdf) of the conditional distribution of $\mathbf{y}$ given data $\mathbf{X}$, (known) and parameters $(\boldsymbol{\beta}, \sigma^2)$ (unknown).
- The **maximum likelihood** estimates of $(\boldsymbol{\beta}, \sigma^2)$ are the values maximizing $L(\boldsymbol{\beta}, \sigma^2)$, i.e., those which make the observed data $\mathbf{y}$ most likely in terms of its pdf.

Because the y_i are independent r.v.'s with $y_i \sim N(\mu_i, \sigma^2)$ where $\mu_i = \sum_{j=1}^{p} \beta_j x_{i,j}$,

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2) &= \prod_{i=1}^{n} p(y_i \mid \beta, \sigma^2) \\ &= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \sum_{j=1} \beta_j x_{i,j})^2} \right] \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\sigma^2 \mathbf{I}_n)^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})} \end{aligned}$$

The maximum likelihood estimates $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ maximize the log-likeliood function (dropping constant terms)

$$\begin{aligned} logL(\boldsymbol{\beta}, \sigma^2) &= -\frac{n}{2} log(\sigma^2) - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\sigma^2 \mathbf{I}_n)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= -\frac{n}{2} log(\sigma^2) - \frac{1}{2\sigma^2} Q(\boldsymbol{\beta}) \end{aligned}$$

where $Q(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ ("Least-Squares Criterion"!)

- The OLS estimate $\hat{\boldsymbol{\beta}}$ is also the ML-estimate.
- The ML estimate of σ^2 solves

$$\frac{\partial log\ L(\hat{\boldsymbol{\beta}}, \sigma^2)}{\partial(\sigma^2)} = 0 \text{ ,i.e., } -\frac{n}{2}\frac{1}{\sigma^2} - \frac{1}{2}(-1)(\sigma^2)^{-2} Q(\hat{\boldsymbol{\beta}}) = 0$$

$$\implies \hat{\sigma}^2_{ML} = Q(\hat{\boldsymbol{\beta}})/n = (\sum_{i=1}^{n} \hat{\epsilon}_i^2)/n \quad (biased!)$$

Outline

1. Regression Analysis II
   - Distribution Theory: Normal Regression Models
   - Maximum Likelihood Estimation
   - Generalized M Estimation

Generalized M Estimation

For data $\mathbf{y}$, $\mathbf{X}$ fit the linear regression model

$$\mathbf{y}_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i,\ i = 1, 2, \ldots, n.$$

by specifying $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ to minimize

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^n h(y_i, \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2)$$

The choice of the function $h(\ )$ distinguishes different estimators.

(1) Least Squares: $h(y_i, \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) = (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2$

(2) Mean Absolute Deviation (MAD): $h(y_i, \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) = |y_i - \mathbf{x}_i^T \boldsymbol{\beta}|$

(3) Maximum Likelihood (ML): Assume the y_i are independent with pdf's $p(y_i \mid \boldsymbol{\beta}, \mathbf{x}_i, \sigma^2)$,

$$h(y_i, \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) = -log\ p(y_i \mid \boldsymbol{\beta}, \mathbf{x}_i, \sigma^2)$$

(4) Robust $M-$Estimator: $h(y_i, \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) = \chi(y_i - \mathbf{x}_i^T \boldsymbol{\beta})$

$\chi(\ )$ is even, monotone increasing on $(0, \infty)$.

(5) Quantile Estimator: For $\tau : 0 < \tau < 1$, a fixed *quantile*

$$h(y_i, \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) = \begin{cases} \tau|y_i - \mathbf{x}_i^T \boldsymbol{\beta}|, & if\, y_i \geq \mathbf{x}_i \boldsymbol{\beta} \\ (1-\tau)|y_i - \mathbf{x}_i^T \boldsymbol{\beta}|, & if\, y_i < \mathbf{x}_i \boldsymbol{\beta} \end{cases}$$

- E.g., $\tau = 0.90$ corresponds to the 90th quantile / upper-decile.
- $\tau = 0.50$ corresponds to the *MAD* Estimator