Gaussian Linear Models

MIT 18.655

Dr. Kempthorne

Spring 2016

Outline

1. Gaussian Linear Models
   - Linear Regression: Overview
   - Ordinary Least Squares (OLS)
   - Distribution Theory: Normal Regression Models
   - Maximum Likelihood Estimation
   - Generalized M Estimation

General Linear Model: For each case i, the conditional distribution $[y_i \mid x_i]$ is given by

$$y_i = \hat{y}_i + \epsilon_i$$

where

- $\hat{y}_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{i,p} x_{i,p}$
- $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T$ are p regression parameters (constant over all cases)
- ϵ_i Residual (error) variable (varies over all cases)

Extensive breadth of possible models

- Polynomial approximation ($x_{i,j} = (x_i)^j$, explanatory variables are different powers of the same variable $x = x_i$)
- Fourier Series: ($x_{i,j} = sin(jx_i)$ or $cos(jx_i)$, explanatory variables are different sin/cos terms of a Fourier series expansion)
- Time series regressions: time indexed by i, and explanatory variables include lagged response values.

Note: *Linearity* of $\hat{y}_i$ (in regression parameters) maintained with non-linear x.

Steps for Fitting a Model

(1) Propose a model in terms of
- Response variable Y (specify the scale)
- Explanatory variables $X_1, X_2, \ldots X_p$ (include different functions of explanatory variables if appropriate)
- Assumptions about the distribution of ϵ over the cases

(2) Specify/define a criterion for judging different estimators.

(3) Characterize the best estimator and apply it to the given data.

(4) Check the assumptions in (1).

(5) If necessary modify model and/or assumptions and go to (1).

Specifying Estimator Criterion in (2)

- Least Squares
- Maximum Likelihood
- Robust (Contamination-resistant)
- Bayes (assume β_j are r.v.'s with known *prior* distribution)
- Accommodating incomplete/missing data

Case Analyses for (4) Checking Assumptions

- Residual analysis
  - Model errors ϵ_i are unobservable
  - Model residuals for fitted regression parameters $\tilde{\beta}_j$ are:

$$e_i = y_i - [\tilde{\beta}_1 x_{i,1} + \tilde{\beta}_2 x_{i,2} + \cdots + \tilde{\beta}_p x_{i,p}]$$

- Influence diagnostics (identify cases which are highly 'influential'?)
- Outlier detection

Outline

Ordinary Least Squares Estimates

Least Squares Criterion: For $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$, define

$$\begin{aligned} Q(\boldsymbol{\beta}) &= \sum_{i=1}^{N} [y_i - \hat{y}_i]^2 \\ &= \sum_{i=1}^{N} [y_i - (\beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{i,p} x_{i,p})]^2 \end{aligned}$$

Ordinary Least-Squares (OLS) estimate $\hat{\boldsymbol{\beta}}$: minimizes $Q(\boldsymbol{\beta})$.

Matrix Notation

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{p,n} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

Solving for OLS Estimate $\hat{\beta}$

$$\hat{\mathbf{y}} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \mathbf{X}\beta \text{ and}$$

$$\begin{aligned} Q(\beta) &= \textstyle\sum_{i=1}^n (y_i - \hat{y}_i)^2 = (\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}}) \\ &= (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) \end{aligned}$$

OLS $\hat{\beta}$ solves $\dfrac{\partial Q(\beta)}{\partial \beta_j} = 0, \quad j = 1, 2, \ldots, p$

$$\begin{aligned} \frac{\partial Q(\beta)}{\partial \beta_j} &= \frac{\partial}{\partial \beta_j}\left(\textstyle\sum_{i=1}^n [y_i - (x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots x_{i,p}\beta_p)]^2\right) \\ &= \textstyle\sum_{i=1}^n 2(-x_{i,j})[y_i - (x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots x_{i,p}\beta_p)] \\ &= -2(\mathbf{X}_{[j]})^T(\mathbf{y} - \mathbf{X}\beta) \quad \text{where } \mathbf{X}_{[j]} \text{ is the } j\text{th column of } \mathbf{X} \end{aligned}$$

Solving for OLS Estimate $\hat{\beta}$

$$\frac{\partial Q}{\partial \beta} = \begin{bmatrix} \frac{\partial Q}{\partial \beta_1} \\ \frac{\partial Q}{\partial \beta_2} \\ \vdots \\ \frac{\partial Q}{\partial \beta_p} \end{bmatrix} = -2 \begin{bmatrix} \mathbf{X}_{[1]}^T(\mathbf{y} - \mathbf{X}\beta) \\ \mathbf{X}_{[2]}^T(\mathbf{y} - \mathbf{X}\beta) \\ \vdots \\ \mathbf{X}_{[p]}^T(\mathbf{y} - \mathbf{X}\beta) \end{bmatrix} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta)$$

So the OLS Estimate $\hat{\beta}$ solves the **"Normal Equations"**

$$\begin{aligned} \mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) &= \mathbf{0} \\ \iff \quad \mathbf{X}^T\mathbf{X}\hat{\beta} &= \mathbf{X}^T\mathbf{y} \\ \implies \quad \hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \end{aligned}$$

N.B. For $\hat{\beta}$ to exist (uniquely)

$(\mathbf{X}^T\mathbf{X})$ must be invertible

$\iff$ $\mathbf{X}$ must have Full Column Rank

(Ordinary) Least Squares Fit

OLS Estimate:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Fitted Values:

$$\hat{\mathbf{y}} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} x_{1,1}\hat{\beta}_1 + \cdots + x_{1,p}\hat{\beta}_p \\ x_{2,1}\hat{\beta}_1 + \cdots + x_{2,p}\hat{\beta}_p \\ \vdots \\ x_{n,1}\hat{\beta}_1 + \cdots + x_{n,p}\hat{\beta}_p \end{pmatrix}$$

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

Where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the $n \times n$ "Hat Matrix"

(Ordinary) Least Squares Fit

The Hat Matrix $\mathbf{H}$ projects R^n onto the column-space of $\mathbf{X}$

Residuals: $\hat{\epsilon}_i = y_i - \hat{y}_i$, $i = 1, 2, \dots, n$

$$\hat{\boldsymbol{\epsilon}} = \begin{pmatrix} \hat{\epsilon}_1 \\ \hat{\epsilon}_2 \\ \vdots \\ \hat{\epsilon}_n \end{pmatrix} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$$

Normal Equations: $\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\beta}) = \mathbf{X}^T\hat{\epsilon} = \mathbf{0}_p = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$

N.B. The Least-Squares Residuals vector $\hat{\epsilon}$ is orthogonal to the column space of $\mathbf{X}$

Outline

Normal Linear Regression Models

Distribution Theory

$$\begin{aligned} Y_i &= x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots x_{i,p}\beta_p + \epsilon_i \\ &= \mu_i + \epsilon_i \end{aligned}$$

Assume $\{\epsilon_1, \epsilon_2, \ldots, \epsilon_n\}$ are i.i.d $N(0, \sigma^2)$.

$$\Longrightarrow [Y_i \mid x_{i,1}, x_{i,2}, \ldots, x_{i,p}, \boldsymbol{\beta}, \sigma^2] \sim N(\mu_i, \sigma^2),$$

independent over $i = 1, 2, \ldots n$.

Conditioning on X, $\boldsymbol{\beta}$, and σ^2

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \sim N_n(\mathbf{O}_n, \sigma^2 \mathbf{I}_n)$$

Distribution Theory

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = E(\mathbf{Y} \mid \mathbf{X}, \beta, \sigma^2) = \mathbf{X}\beta$$

$$\mathbf{\Sigma} = Cov(\mathbf{Y} \mid \mathbf{X}, \beta, \sigma^2) = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}_n$$

That is, $\mathbf{\Sigma}_{i,j} = Cov(Y_i, Y_j \mid \mathbf{X}, \beta, \sigma^2) = \sigma^2 \times \delta_{i,j}$.

Apply Moment-Generating Functions (MGFs) to derive

- Joint distribution of $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$
- Joint distribution of $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)^T$.

MGF of Y

For the n-variate r.v. $\mathbf{Y}$, and constant $n-$vector $\mathbf{t} = (t_1, \ldots, t_n)^T$,

$$
\begin{aligned}
M_{\mathbf{Y}}(\mathbf{t}) &= E(e^{\mathbf{t}^T\mathbf{Y}}) = E(e^{t_1 Y_1 + t_2 Y_2 + \cdots t_n Y_n}) \\
&= E(e^{t_1 Y_1}) \cdot E(e^{t_2 Y_2}) \cdots E(e^{t_n Y_n}) \\
&= M_{Y_1}(t_1) \cdot M_{Y_2}(t_2) \cdots M_{Y_n}(t_n) \\
&= \prod_{i=1}^n e^{t_i \mu_i + \frac{1}{2} t_i^2 \sigma^2} \\
&= e^{\sum_{i=1}^n t_i \mu_i + \frac{1}{2} \sum_{i,k=1}^n t_i \mathbf{\Sigma}_{i,k} t_k} = e^{\mathbf{t}^T \mathbf{u} + \frac{1}{2} \mathbf{t}^T \mathbf{\Sigma} \mathbf{t}}
\end{aligned}
$$

$\Longrightarrow \mathbf{Y} \sim N_n(\boldsymbol{\mu}, \mathbf{\Sigma})$

Multivariate Normal with mean $\boldsymbol{\mu}$ and covariance $\mathbf{\Sigma}$

MGF of $\hat{\beta}$

For the p-variate r.v. $\hat{\beta}$, and constant $p-$vector $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)^T$,

$$M_{\hat{\beta}}(\boldsymbol{\tau}) = E(e^{\boldsymbol{\tau}^T\hat{\boldsymbol{\beta}}}) = E(e^{\tau_1\hat{\beta}_1+\tau_2\hat{\beta}_2+\cdots\tau_p\beta_p})$$

Defining $\mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ we can express

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

and

$$\begin{aligned} M_{\hat{\beta}}(\boldsymbol{\tau}) &= E(e^{\boldsymbol{\tau}^T\hat{\boldsymbol{\beta}}}) \\ &= E(e^{\boldsymbol{\tau}^T\mathbf{AY}}) \\ &= E(e^{\mathbf{t}^T\mathbf{Y}}), \quad \text{with } \mathbf{t} = \mathbf{A}^T\boldsymbol{\tau} \\ &= M_{\mathbf{Y}}(\mathbf{t}) \\ &= e^{\mathbf{t}^T\mathbf{u}+\frac{1}{2}\mathbf{t}^T\boldsymbol{\Sigma}\mathbf{t}} \end{aligned}$$

MGF of $\hat{\beta}$

For

$$\begin{aligned} M_{\hat{\beta}}(\boldsymbol{\tau}) &= E(e^{\boldsymbol{\tau}^T \hat{\beta}}) \\ &= e^{\mathbf{t}^T \mathbf{u} + \frac{1}{2}\mathbf{t}^T \boldsymbol{\Sigma} \mathbf{t}} \end{aligned}$$

Plug in:

$$\begin{aligned} \mathbf{t} &= \mathbf{A}^T \boldsymbol{\tau} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{\tau} \\ \boldsymbol{\mu} &= \mathbf{X}\beta \\ \boldsymbol{\Sigma} &= \sigma^2 \mathbf{I}_n \end{aligned}$$

Gives:

$$\begin{aligned} \mathbf{t}^T \boldsymbol{\mu} &= \boldsymbol{\tau}^T \beta \\ \mathbf{t}^T \boldsymbol{\Sigma} \mathbf{t} &= \boldsymbol{\tau}^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T [\sigma^2 \mathbf{I}_n] \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{\tau} \\ &= \boldsymbol{\tau}^T [\sigma^2 (\mathbf{X}^T\mathbf{X})^{-1}] \boldsymbol{\tau} \end{aligned}$$

So the MGF of $\hat{\beta}$ is

$$M_{\hat{\beta}}(\boldsymbol{\tau}) = e^{\boldsymbol{\tau}^T \beta + \frac{1}{2}\boldsymbol{\tau}^T [\sigma^2 (\mathbf{X}^T\mathbf{X})^{-1}] \boldsymbol{\tau}}$$

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

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

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

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

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

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

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

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

Theorem For the normal linear regression model

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

where

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

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

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

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

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

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

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

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

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

$\Longrightarrow$

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

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:

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

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

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
  $$\frac{\partial log\ L(\hat{\boldsymbol{\beta}}, \sigma^2)}{\partial(\sigma^2)} = 0 \text{ ,i.e., } -\frac{n}{2}\frac{1}{\sigma^2} - \frac{1}{2}(-1)(\sigma^2)^{-2}Q(\hat{\boldsymbol{\beta}}) = 0$$
  $$\Longrightarrow \hat{\sigma}^2_{ML} = Q(\hat{\boldsymbol{\beta}})/n = (\textstyle\sum_{i=1}^{n} \hat{\epsilon}_i^2)/n \quad (biased!)$$

Outline

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

(2) Least Absolue Deviation (LADE): $h(y_i, \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) = |y_i - \mathbf{x}_i^T \boldsymbol{\beta}|$

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

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

Laplace (LADE); Gauss and Legendre (LSE)

(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*

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

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