MTH499/599 Lecture Notes 04

Donghui Yan

Department of Math, Umass Dartmouth

December 23, 2015

Outline

- Multivariate linear regression
- Properties of the OLS estimate

Review on the simple linear model

- The simple linear model is specified as

$$Y = \beta_0 + \beta_1 X + \epsilon$$

  where ϵ is random error, and β_0, β_1 are constants
- Model assumption
  - $\mathbb{E}(Y|X) = \beta_0 + \beta_1 X$ (linear model)
  - $\mathbb{E}\epsilon = 0, Var(\epsilon) = \sigma^2$ (Homoscedasticity)
  - $\epsilon \sim \mathcal{N}(0, \sigma^2)$ (normality)
- The least square formulation

$$\arg \min_{\{\beta_0,\beta_1\}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

Multivariate linear model

- Given a sample $(\boldsymbol{X}_1, Y_1), (\boldsymbol{x}_2, Y_2), ..., (\boldsymbol{X}_n, Y_n)$, where $\boldsymbol{X}_i = (X_{i1}, X_{i2}, ..., X_{i(p-1)}) \in \mathbb{R}^{p-1}$ and $Y_i \in \mathbb{R}$
- The multivariate linear model is specified as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i(p-1)} + \epsilon_i$$

  where ϵ is random error, and β_i's are constants
- Model assumption
  - $\mathbb{E}(Y_i | \boldsymbol{X}_i) = \beta_0 + \beta_1 X_{i1} + ... + \beta_{p-1} X_{i(p-1)}$ (linear model)
  - $\mathbb{E}\epsilon_i = 0, Var(\epsilon_i) = \sigma^2$ (Homoscedasticity)
  - $\epsilon \sim \mathcal{N}(0, \sigma^2)$ (normality)
- The least square formulation

$$\arg \min_{\{\beta_0, ..., \beta_{p-1}\}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1} X_{i(p-1)})^2.$$

The matrix version of OLS

- Given a sample $(\boldsymbol{X}_1, Y_1), ..., (\boldsymbol{X}_n, Y_n)$, introduce notation

$$\boldsymbol{Y} = [Y_1, Y_2, ..., Y_n]^T,\ \boldsymbol{\beta} = [\beta_0, \beta_1, ..., \beta_{p-1}]^T,$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1(p-1)} \\ 1 & X_{21} & X_{22} & \cdots & X_{2(p-1)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n(p-1)} \end{bmatrix}$$

- The least square formulation becomes

$$\arg\min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \triangleq \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}).$$

The toy example in matrix notation

The data was given as the following sample

$$\cup_{i=1}^{8}\{(X_i, Y_i)\} =$$
$$\{(6,6),(5,9),(4,8),(3,10),(2,11),(2,12),(1,11),(1,13)\}$$

$$\begin{aligned}
\boldsymbol{\beta} &= [\beta_0, \beta_1]^T, \\
\boldsymbol{Y} &= [6,9,8,10,11,12,11,13]^T, \\
\boldsymbol{X} &= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 6 & 5 & 4 & 3 & 2 & 2 & 1 & 1 \end{bmatrix}^T.
\end{aligned}$$

A little matrix algebra

- Let $\mathbf{X}$ be a vector. Then

$$Cov(\mathbf{X}) = \mathbb{E}[(\mathbf{X} - \mathbb{E}(X))(\mathbf{X} - \mathbb{E}(X))^T]$$

- $\mathbf{A}$ and $\mathbf{B}$ constant matrices, $\mathbf{c}$ and $\mathbf{d}$ constant vectors. Then

$$\begin{aligned} Cov(\mathbf{A}\mathbf{x_1} + \mathbf{c}, \mathbf{B}\mathbf{x_2} + \mathbf{d}) &= \mathbf{A} Cov(\mathbf{x_1}, \mathbf{x_2}) \mathbf{B}^T \\ &\triangleq \mathbf{A} < \mathbf{x_1}, \mathbf{x_2} > \mathbf{B}^T \end{aligned}$$

- Let matrix $\boldsymbol{W}$ be symmetric. Then

$$\frac{\partial}{\partial \boldsymbol{s}}(Y - \boldsymbol{A}\boldsymbol{s})^T \boldsymbol{W} (Y - \boldsymbol{A}\boldsymbol{s}) = -2\boldsymbol{A}^T \boldsymbol{W} (Y - \boldsymbol{A}\boldsymbol{s}).$$

The matrix version of OLS

- Taking partial derivative of $\mathcal{L}(\hat{\beta})$ and setting to 0 yields

$$0 = \partial \mathcal{L}(\hat{\beta}) / \partial \hat{\beta} = -\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\beta}),$$
$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\hat{\beta},$$

  Thus

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

- The fitted value $\hat{\mathbf{y}}$ can be expressed as

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

  - $\mathbf{H} \triangleq \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called a hat matrix.

The toy example in matrix notation (R code)

```
> x<-matrix(0,8,2);
> x[,1]<-1; x[,2]<-c(6,5,4,3,2,2,1,1);
> x
     [,1] [,2]
[1,]    1    6
[2,]    1    5
[3,]    1    4
[4,]    1    3
[5,]    1    2
[6,]    1    2
[7,]    1    1
[8,]    1    1
> y<-matrix(c(6,9,8,10,11,12,11,13),8,1);
> y
     [,1]
[1,]    6
[2,]    9
[3,]    8
[4,]   10
[5,]   11
[6,]   12
[7,]   11
[8,]   13
> solve(t(x) %*% x) %*% t(x) %*% y;
       [,1]
[1,] 13.375
[2,] -1.125
```

Example (auto MPG in city)

- Data taken from UC Irvine machine learning repository
- # observations: 392
- Nine variables
  - Response variable: mpg
  - Predicator variables
    - # cylinders
    - Displacement
    - Horsepower
    - Weight
    - Acceleration
    - Model year
    - Origin
    - Car name.

Example (auto MPG in city)

- The firs few lines of the data looks like

mpg	cylinders	displacement	horsepower	weight	acceleration	modelyear	origin	carname
18.0	8	307.0	130.0	3504.	12.0	70	1	"chevrolet chevelle malibu"
15.0	8	350.0	165.0	3693.	11.5	70	1	"buick skylark 320"
18.0	8	318.0	150.0	3436.	11.0	70	1	"plymouth satellite"
16.0	8	304.0	150.0	3433.	12.0	70	1	"amc rebel sst"
17.0	8	302.0	140.0	3449.	10.5	70	1	"ford torino"
15.0	8	429.0	198.0	4341.	10.0	70	1	"ford galaxie 500"
14.0	8	454.0	220.0	4354.	9.0	70	1	"chevrolet impala"
14.0	8	440.0	215.0	4312.	8.5	70	1	"plymouth fury iii"
14.0	8	455.0	225.0	4425.	10.0	70	1	"pontiac catalina"
15.0	8	390.0	190.0	3850.	8.5	70	1	"amc ambassador dpl"
15.0	8	383.0	170.0	3563.	10.0	70	1	"dodge challenger se"
14.0	8	340.0	160.0	3609.	8.0	70	1	"plymouth 'cuda 340"
15.0	8	400.0	150.0	3761.	9.5	70	1	"chevrolet monte carlo"
14.0	8	455.0	225.0	3086.	10.0	70	1	"buick estate wagon (sw)"
24.0	4	113.0	95.00	2372.	15.0	70	3	"toyota corona mark ii"
22.0	6	198.0	95.00	2833.	15.5	70	1	"plymouth duster"
18.0	6	199.0	97.00	2774.	15.5	70	1	"amc hornet"
21.0	6	200.0	85.00	2587.	16.0	70	1	"ford maverick"
27.0	4	97.00	88.00	2130.	14.5	70	3	"datsun pl510"

Regression output of the auto MPG example

```
> tmp<-read.table("autompg.Data", header=TRUE);
> y<-tmp[,1];
> n<-nrow(tmp); p<-8;
> x<-matrix(0,nrow(tmp),(p-1));
> for(i in 1:(p-1)) { x[,i]<-tmp[,(i+1)];}
> mylm<-lm(y ~ x);
> summary(mylm);

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5903 -2.1565 -0.1169  1.8690 13.0604

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.218435   4.644294  -3.707  0.00024 ***
x1           -0.493376   0.323282  -1.526  0.12780
x2            0.019896   0.007515   2.647  0.00844 **
x3           -0.016951   0.013787  -1.230  0.21963
x4           -0.006474   0.000652  -9.929  < 2e-16 ***
x5            0.080576   0.098845   0.815  0.41548
x6            0.750773   0.050973  14.729  < 2e-16 ***
x7            1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,	Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

```
> tmp<-read.table("autompg.Data", header=TRUE);
> Y<-tmp[,1];
> X<-matrix(0,nrow(tmp),p);
> X[,1]<-1;
> for(i in 2:p) { X[,i]<-tmp[,i];}
> ##Regression estimates by matrix
> solve(t(X) %*% X) %*% t(X) %*% Y
             [,1]
[1,] -17.218434622
[2,]  -0.493376319
[3,]   0.019895644
[4,]  -0.016951144
[5,]  -0.006474043
[6,]   0.080575838
[7,]   0.750772678
[8,]   1.426140495
>
> ##The first few lines of X looks like
> X
      [,1] [,2]  [,3] [,4] [,5] [,6] [,7] [,8]
 [1,]    1    8 307.0  130 3504 12.0   70    1
 [2,]    1    8 350.0  165 3693 11.5   70    1
 [3,]    1    8 318.0  150 3436 11.0   70    1
 [4,]    1    8 304.0  150 3433 12.0   70    1
 [5,]    1    8 302.0  140 3449 10.5   70    1
 [6,]    1    8 429.0  198 4341 10.0   70    1
 [7,]    1    8 454.0  220 4354  9.0   70    1
 [8,]    1    8 440.0  215 4312  8.5   70    1
 [9,]    1    8 455.0  225 4425 10.0   70    1
[10,]    1    8 390.0  190 3850  8.5   70    1
```

Properties of the hat matrix

- The hat matrix is symmetric and idempotent, i.e., $H^2 = H$

$$\begin{aligned} H^2 &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \cdot \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = H \end{aligned}$$

- For matrices A and B, $trace(AB) = trace(BA)$
- $trace(H) = p$

$$\begin{aligned} trace(H) &= trace(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) \\ &= trace((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}) \\ &= trace(\mathbf{I}_p) \\ &= p. \end{aligned}$$

Properties of the hat matrix

- The diagonals of hat matrix, $0 \leq h_{ii} \leq 1$

Proof.

Consider the $i-th$ element along the diagonal of H. Since $H^2 = H$, we have

$$h_{ii} = \sum_{j=1}^{n} h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2,$$

implying that

$$h_{ii}^2 \leq h_{ii}.$$

Thus the result follows. $\square$

Properties of OLS estimate

- The least square estimate is unbiased.

Proof.

$$
\begin{aligned}
\mathbb{E}(\hat{\beta}) &= \mathbb{E}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}(\mathbf{y}) \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta \\
&= \beta.
\end{aligned}
$$

□

Properties of OLS estimate (continued)

- The variance of $\hat{\beta}$ is given by $(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$.

Proof.

$$
\begin{aligned}
Var(\hat{\beta}) &= Var((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) \\
&= Cov((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T < \mathbf{y}, \mathbf{y} > \left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right]^T \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2.
\end{aligned}
$$

$\square$

Hypothesis testing on OLS estimate

- If σ^2 is known, then $\hat{\beta} \sim \mathcal{N}(0, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$ thus a normal test otherwise a t_{n-p-1}-test under $H_0 : \boldsymbol{\beta} = \mathbf{0}$.
  - $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, a linear combination of i.i.d. normal r.v.'s thus follows a normal distribution
  - If σ^2 is known, the testing statistic $T = \hat{\beta}/SD(\hat{\beta}) \sim \mathcal{N}(0,1)$
  - Otherwise, σ^2 is replaced by

    $$\hat{\sigma}^2 = SSE/(n-p-1) \sim \chi_{n-p-1}$$

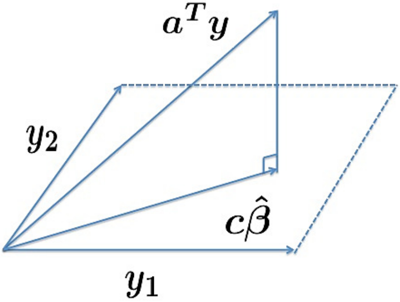
    thus

    $$T = \hat{\beta}/SD(\hat{\beta}) \sim t_{n-p-1}.$$

The Gauss-Markov Theorem

Theorem (Gauss-Markov)

Consider linear model $\mathbf{y} = \mathbf{X}\beta + \epsilon$ *with* $\mathbb{E}(\epsilon) = 0$ *and* $Var(\epsilon) = \sigma^2\mathbf{I}$. *Then the OLS estimate* $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ *is the Best Linear Unbiased Estimate (BLUE) of* β.



The Gauss-Markov Theorem

Proof. Consider the linear parameter of interest $\mathbf{c}\beta$.

Let $\mathbf{a}^T\mathbf{y} \triangleq \bar{\beta}$ be an unbiased estimate of $\mathbf{c}\beta$ (since any estimate would be a linear combination of $\mathbf{y}$'s components. Unbiasedness of $\mathbf{a}^T\mathbf{y}$ implies that

$$\mathbb{E}(\mathbf{a}^T\mathbf{y}) = \mathbf{a}^T X\beta = \mathbf{c}\beta$$

for all β. Thus $\mathbf{a}^T X = \mathbf{c}^T$.

The Gauss-Markov Theorem (continued)

Write $\bar{\beta}$ as $\bar{\beta} = \left(\mathbf{a}^T\mathbf{y} - \mathbf{c}\hat{\beta}\right) + \mathbf{c}\hat{\beta}$. Easily we can verify

$$\begin{aligned}
& Cov(\mathbf{a}^T\mathbf{y} - \mathbf{c}\hat{\beta}, \mathbf{c}\hat{\beta}) \\
= \ & Cov(\mathbf{a}^T\mathbf{y}, \mathbf{c}\hat{\beta}) - Cov(\mathbf{c}\hat{\beta}, \mathbf{c}\hat{\beta}) \\
= \ & Cov(\mathbf{a}^T\mathbf{y}, \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) - \mathbf{c}Var(\hat{\beta})\mathbf{c}^T \\
= \ & \mathbf{a}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{a}\sigma^2 - \mathbf{a}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{a}\sigma^2 \\
= \ & 0.
\end{aligned}$$

Thus variance decomposition

$$Var(\mathbf{a}^T\mathbf{y}) = Var(\mathbf{a}^T\mathbf{y} - \mathbf{c}\hat{\beta}) + Var(\mathbf{c}\hat{\beta})$$

and the result follows. $\square$