STATS 413 Applied Regression Analysis

Lecture 1-2

Gongjun Xu

Department of Statistics
University of Michigan

Regression Analysis

- Science and engineering
- Social Science
- Finance
- Epidemiology
- Psychology and Education
- ...
- Goal: to infer relationships between a response variable and one or more other variables from *data*.

History

Regression to the mean

- Sir Francis Galton, 19th century
  Studied the relation between heights of parents and children and noted that the children regressed to the population mean
- Karl Pearson, late 19th century
  studied $n = 1375$ heights of mothers in the United Kingdom under the age of 65 and one of their adult daughters over the age of 18

sample size

Example1: heights data

Karl Pearson studied heights of mothers and their adult daughters.
Data *Heights* on `http://users.stat.umn.edu/~sandy/alr4ed/data/`

X: Mheight (Mother's Height)	Y: Dheight (Daughter's Height)
$\mathbf{x}_1 = 59.7$	$y_1 = 55.1$
$\mathbf{x}_2 = 58.2$	$y_2 = 56.5$
...	...
60.6	56
60.7	56.8
61.8	56
55.5	57.9
55.4	57.1
56.8	57.6
57.5	57.2
57.3	57.1
...	...

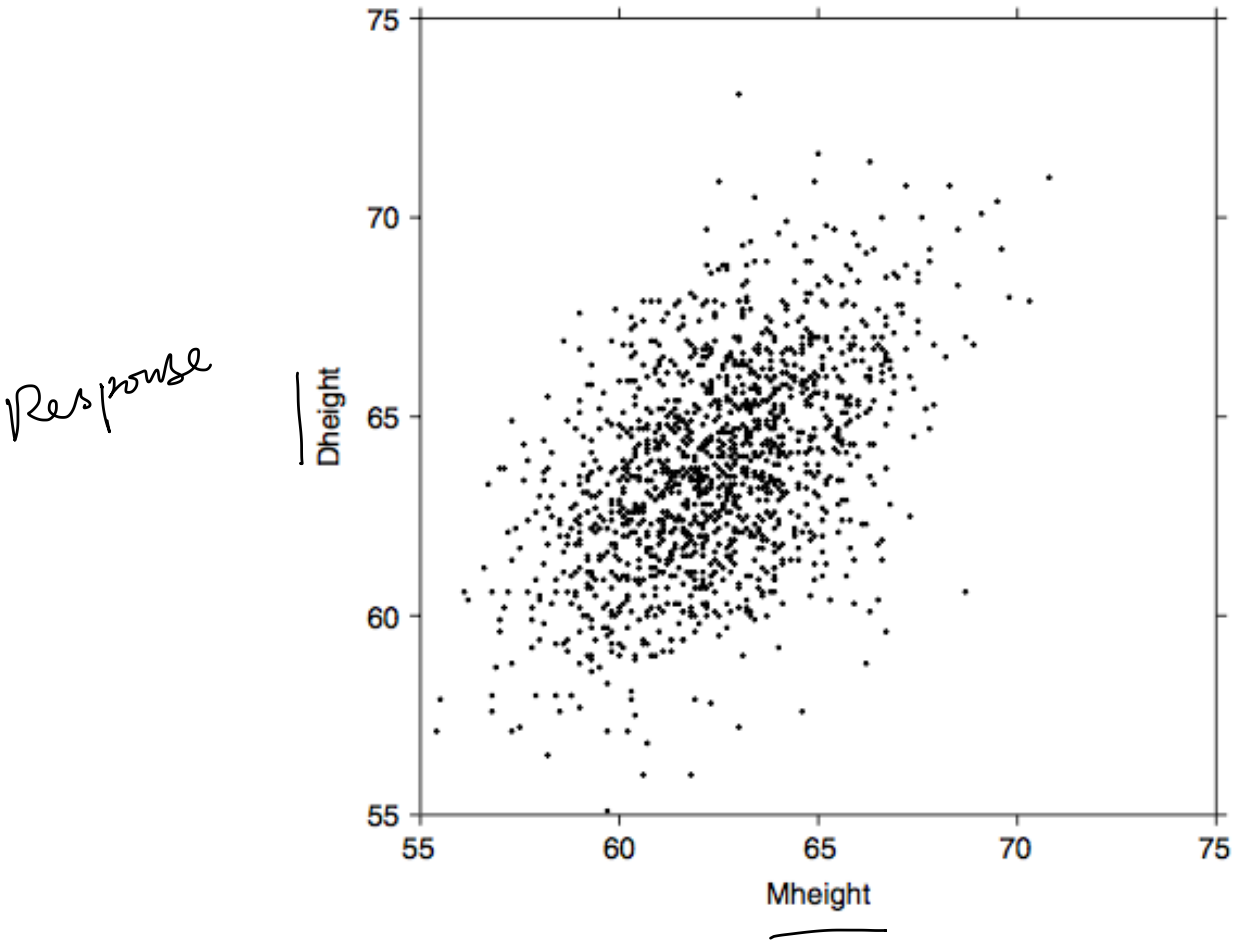
Data Matrix

1375×2

$n = 1375$

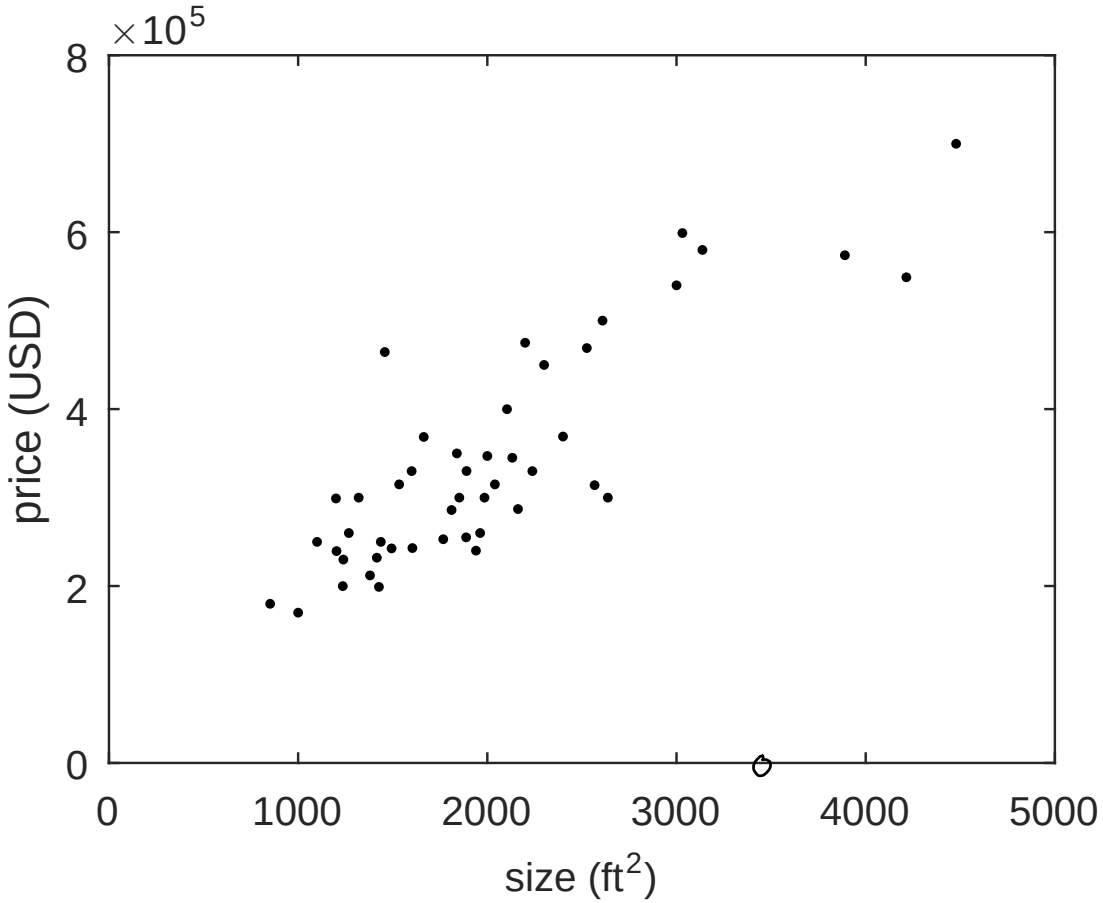
Example 1: scatterplot

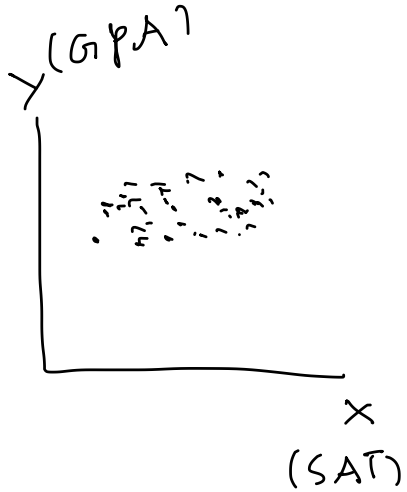
Figure: Scatterplot of mothers' and daughters' heights in the Pearson's data.



Example 2: House price example

A dataset consisting of the prices and square footage of 47 houses





Questions can be addressed with regression analysis

- **point estimation:** investigate *quantitative* questions
  - Instead of a yes or no question, we wish to infer the exact relationship between the response (e.g., price of a house) and the covariate (e.g, its size).
  - What is the slope of **the line of best fit** through the data?
- **hypothesis testing:** investigate *qualitative* questions
  - "guess" an answer and check whether the data supports the "guess"
  - Do taller mothers tend to have taller daughters?
  - Does the price of a house depend on its size?
- **prediction:** predict the value of the response
  - What is the price of a 3500ft^2 house?
  - called **supervised learning** in machine learning

Regression Analysis

Goal: to infer relationships between a response variable (denoted by Y) and one or more other variables (denoted by X) from *data*.
Data: values $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ of (X, Y) observed on each of n units or cases.

- $\mathbf{x}_i \in \mathbb{R}^d$: explanatory variables, features, predictors, covariates
- $y_i \in \mathbb{R}$: dependent variable, response variable, outcome variable, label
- n: sample size

Regression: Y continous (or approx.), Classification: Y discrete

To introduce the Linear Regression Model, we first consider the case when there is only one predictor X. This is also called simple linear regression model.

model: (Statistical) represents the data generating process. it specifies a set of statistical assumptions on the generation of some observed data from a larger population.

population (model)
data Estimation inference.

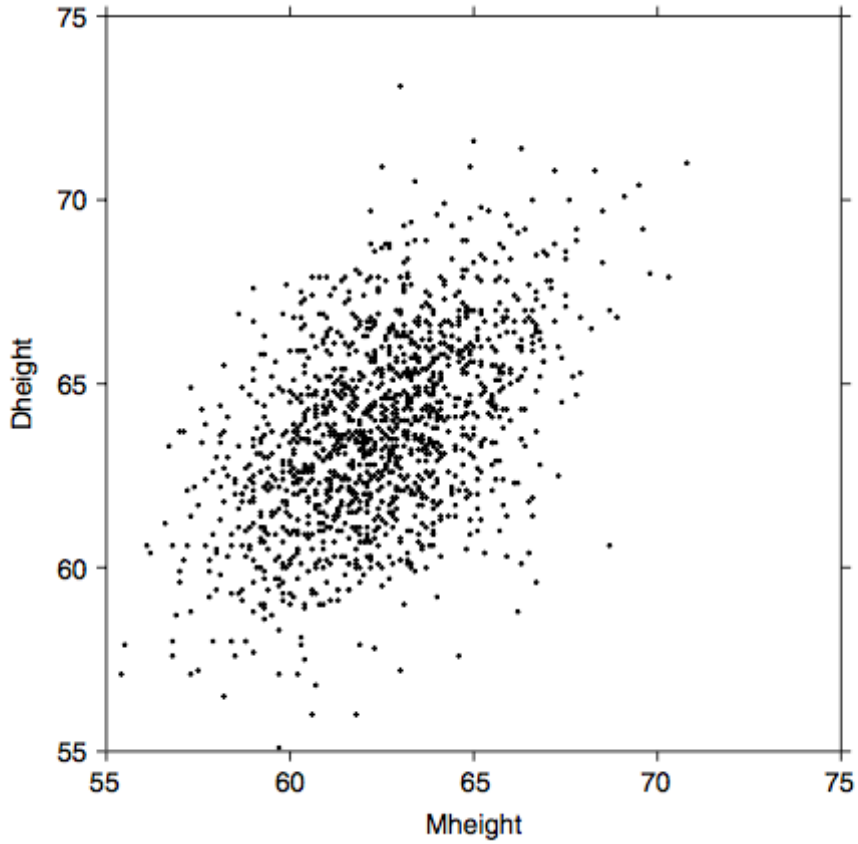
Example1: heights data

Karl Pearson studied heights of mothers and their adult daughters.
Data *Heights* on `http://users.stat.umn.edu/~sandy/alr4ed/data/`

X: Mheight (Mother's Height)	Y: Dheight (Daughter's Height)
$\mathbf{x}_1 = 59.7$	$y_1 = 55.1$
$\mathbf{x}_2 = 58.2$	$y_2 = 56.5$
...	...
60.6	56
60.7	56.8
61.8	56
55.5	57.9
55.4	57.1
56.8	57.6
57.5	57.2
57.3	57.1
...	...

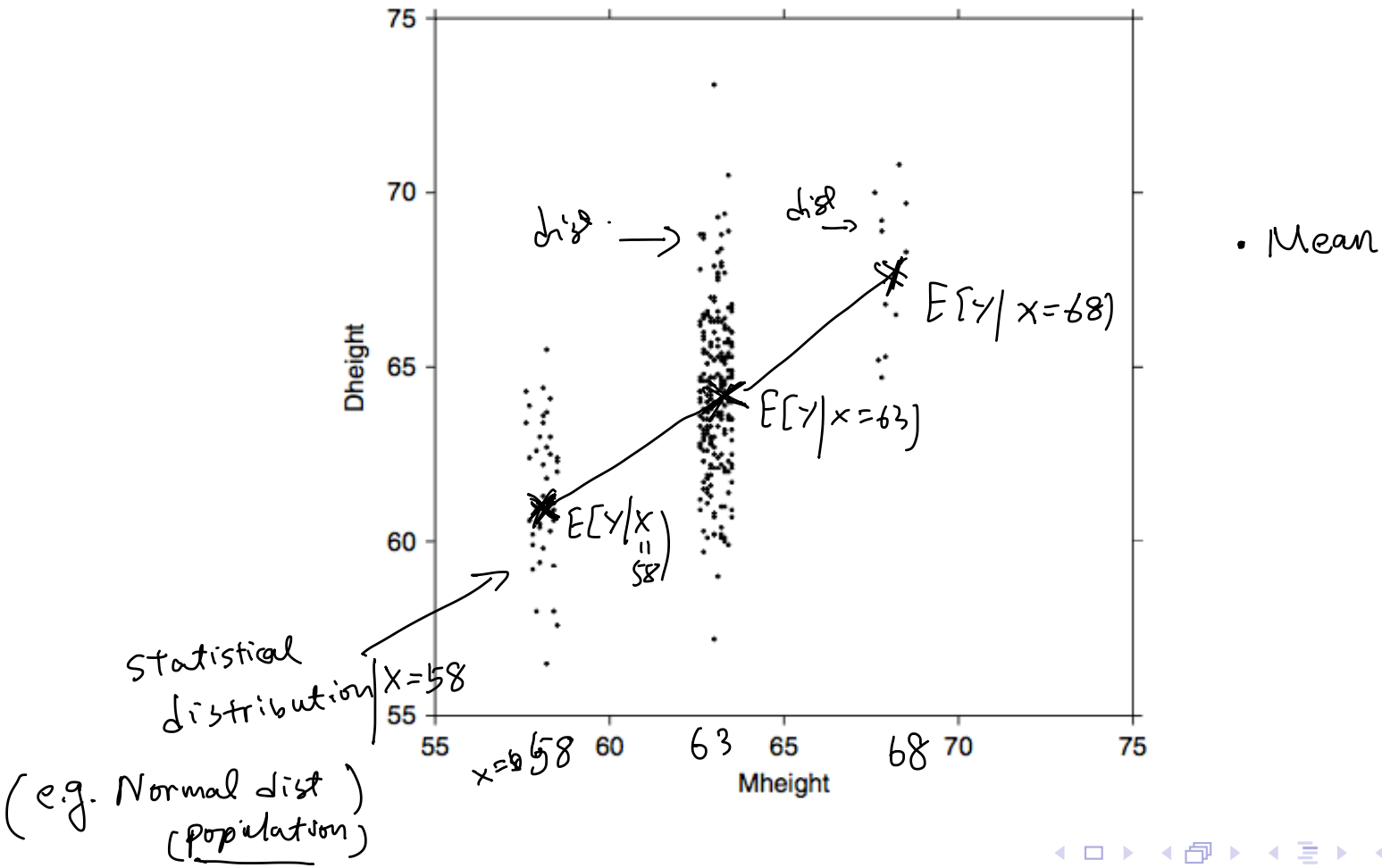
Example1: scatterplot

Figure: Scatterplot of mothers and daughters heights in the Pearson's data.



A closer look at Example 1

Figure: Scatterplot of mothers and daughters heights in the Pearson's data.



Linear Regression Formula (1): mean function.

- Our interest focuses on how the distribution of Y changes as X is varied. One characteristic of the distribution of Y given X is the mean function: mean/Expectation

$$E[Y|X = x] = \text{ a function that depends on the value of x}$$

- Linear Regression assumes that the mean function is a straight line, i.e.,

$$E[Y|X = x] = \beta_0 + \beta_1 x$$

(β_0: intercept, β_1: slope)

β_0, β_1 are regression parameters.

β_0, β_1: regression coefficients; model parameters (unknown quantities characterize our model)

- Example:

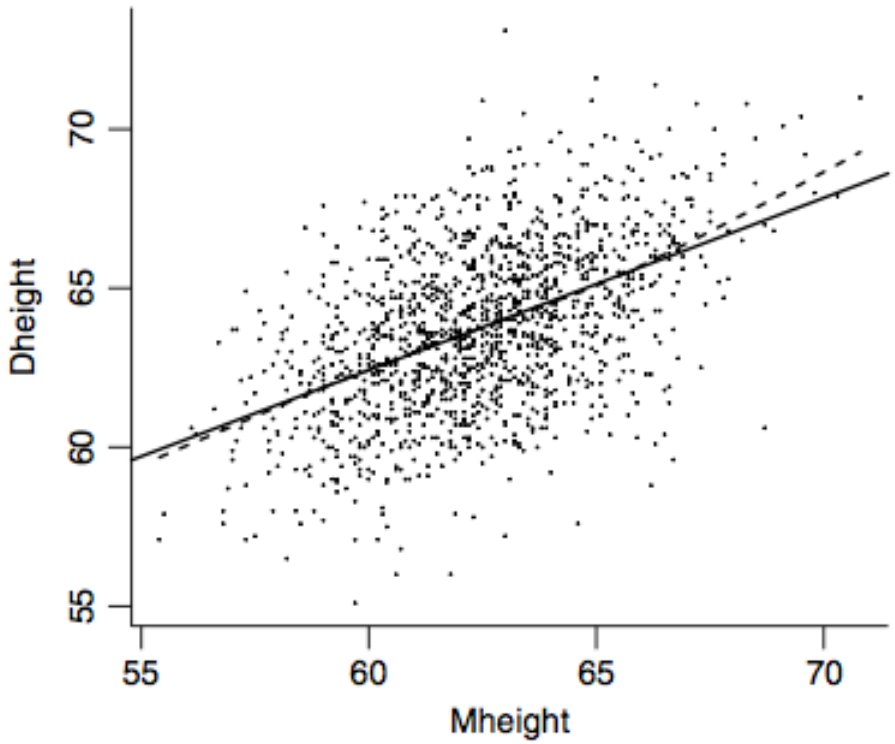
$$E[Dheight|Mheight = x] = \beta_0 + \beta_1 x$$

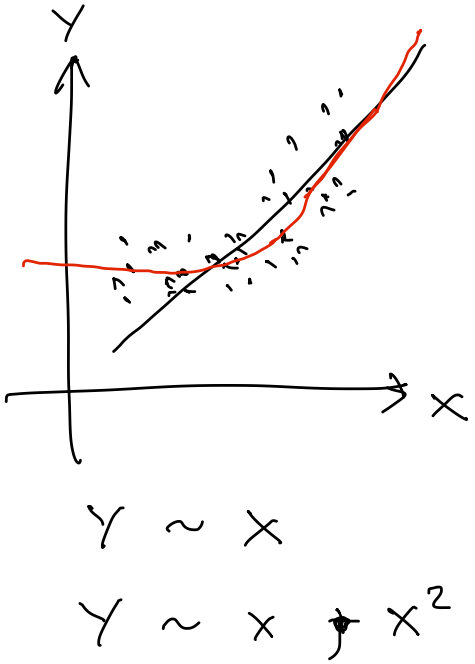
- Why linear? How to fit this line?
  Ordinary least squares (OLS) estimation (Lecture)

Example1: regression line

Why linear? (good interpretation)

Figure: Scatterplot of mothers' and daughters' heights in the Pearson's data.



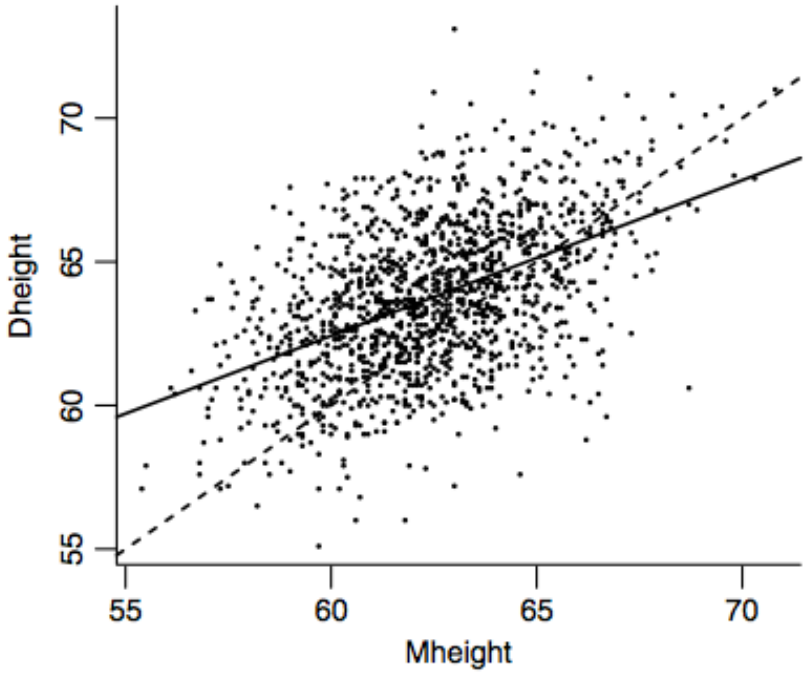


solid line is the OLS line; dashed line is "smooth" estimates of $E[Y|X = x]$.

Example1: regression line

How to fit this straight line? – ordinary least squares estimation

Figure: Scatterplot.



Positive association.

Linear Regression Formula (2): Variance

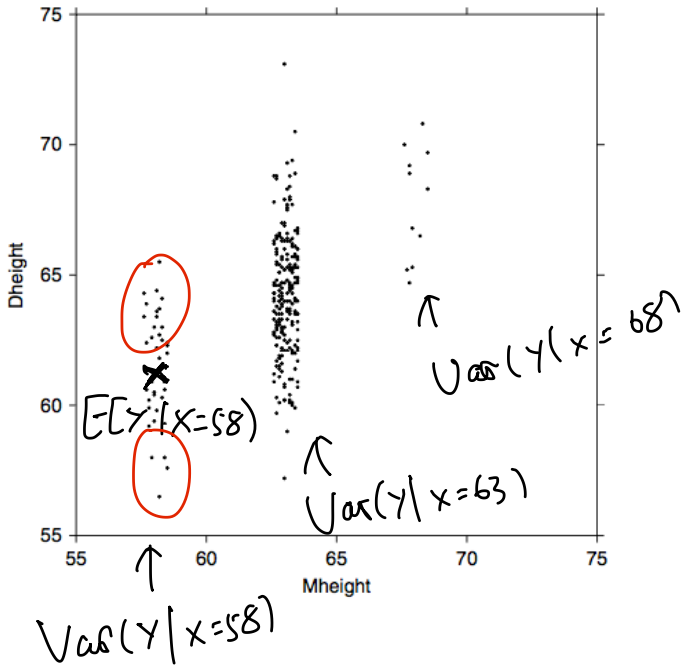
Besides mean function, another aspect of the distribution of Y given X is the variance function:

$$Var(Y|X = x) = \sigma^2.$$

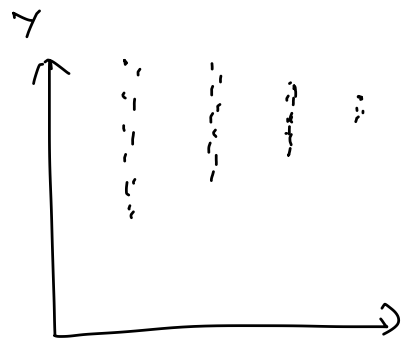
constant does not depend on X

(σ^2: model parameter)

Figure: Scatterplot.



E.g.



$Var(Y|X=x) = \sigma^2(x)$

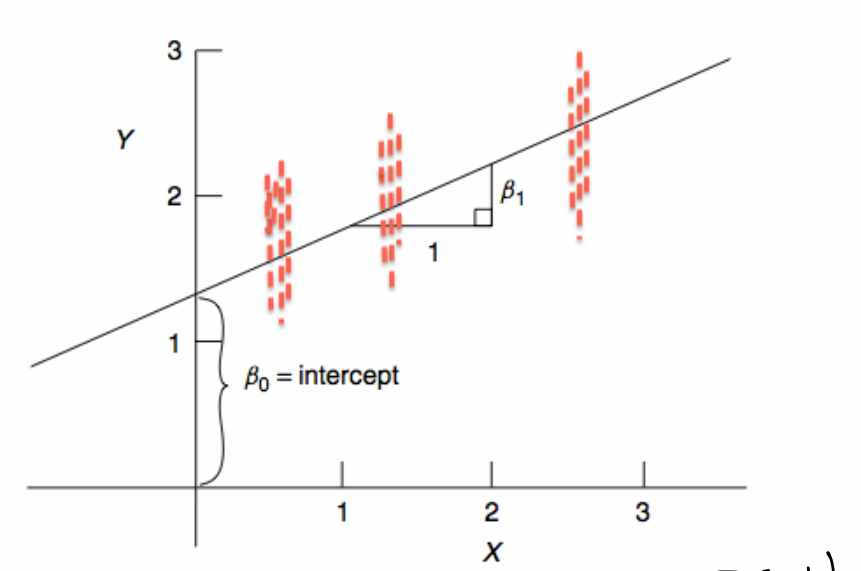
Weighted least Square (later lecture).

Linear Regression Formula

The simple linear regression model consists of the mean function and the variance function

$$E[Y|X = x] = \beta_0 + \beta_1 x \quad \& \quad Var(Y|X = x) = \sigma^2.$$

(x, Y)



$\beta_0, \beta_1, \sigma^2$

model parameters.

$E[Y|X=0] = \beta_0$

The intercept parameter β_0 is the expected value of the response when the predictor $x = 0$. The slope parameter β_1 gives the change in the expected value when the predictor x increases by 1 unit.

$$\begin{cases} E[Y|X] = \beta_0 + \beta_1 x & \cdots (1) \\ E[Y|X+1] = \beta_0 + \beta_1 (x+1) & \cdots (2) \end{cases}$$

Statistical error e

- To account for each observation's difference between the observed response and the expected value, statisticians have invented a quantity called a **statistical error**.
- For the ith observation, the statistical error e_i is defined by

$$e_i \triangleq y_i - E(Y|X = x_i) = y_i - \beta_0 - \beta_1 x_i$$

or implicitly by

$$y_i = E(Y|X = x_i) + e_i = \beta_0 + \beta_1 x_i + e_i.$$

- e_i corresponds to the vertical distance between the point y_i and the mean function $E(Y|X = x_i)$.

Linear Regression Model

- Simple Linear Regression Model

For a random pair (X, Y)

mean function $E[Y|X=x] = \beta_0 + \beta_1 x$

$E[Y - \beta_0 - \beta_1 x | X = x] = 0$

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

Variance function

$Var(Y|X=x) = \sigma^2$

$Var(Y - \beta_0 - \beta_1 x | X = x) = \sigma^2$

e

with $E[e|X = x] = 0$ and $Var(e|X = x) = \sigma^2$.

- For n independent observations $i = 1, \cdots, n$:

n: sample size

e.g. $n = 375$ Height Example

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

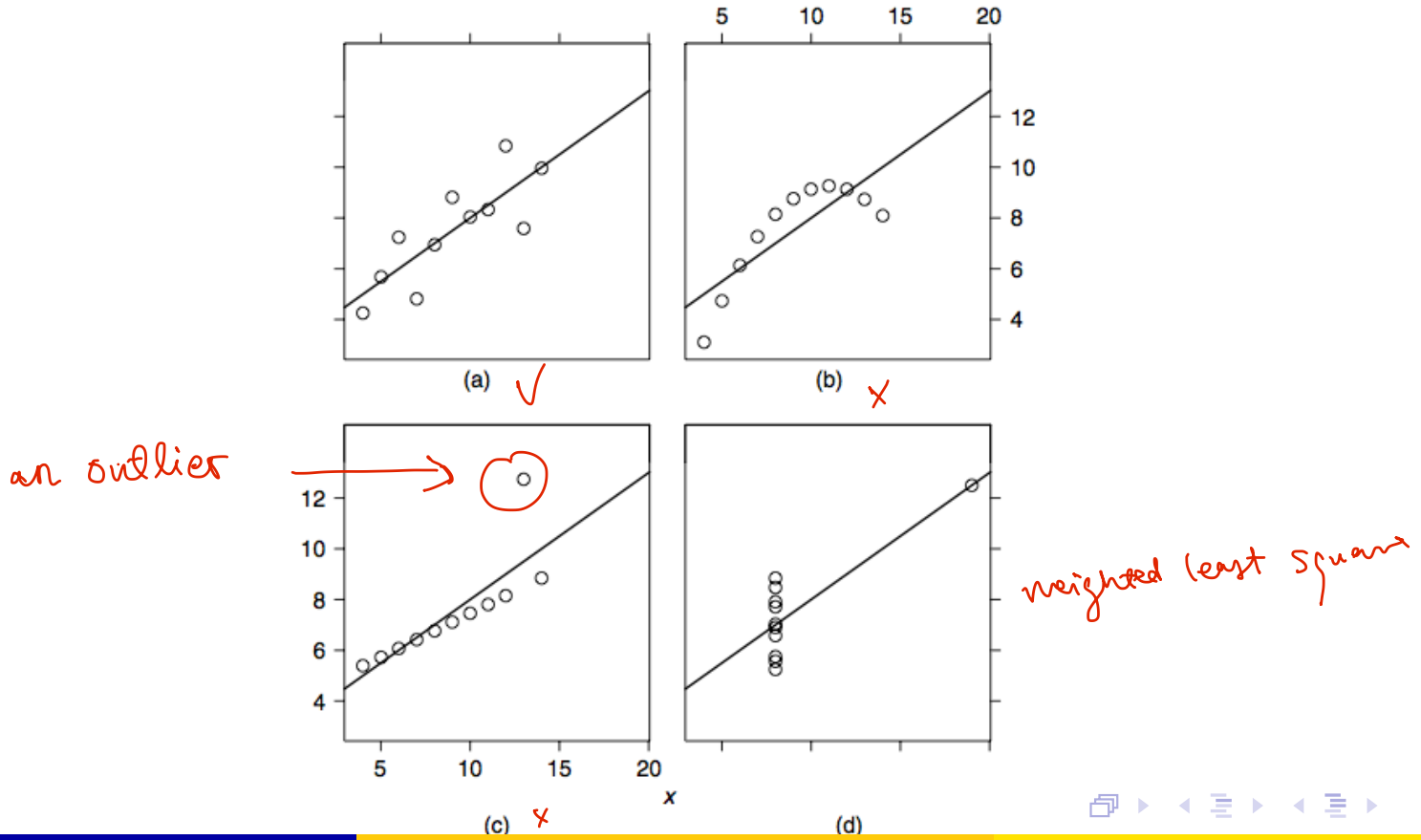
for $i = 1, \cdots, n$, $E[e_i|x_i] = 0$ and $Var(e_i|x_i) = \sigma^2$. The e_i's are independent and often assumed to be normally distributed.

- $\beta_0, \beta_1, \sigma^2$ are unknown parameters that need to be estimated from data.

First step of data analysis: summary graph

Check if the linear regression model assumptions are satisfied using summary graph

Figure: Examples



Multiple Regression

- We have considered one predictor case – simple linear regression.
- Often we have more than 1 predictor, and the response variable (Y) is best understood as a function of multiple predictors $(X_1, X_2, \cdots, X_p)$.
- Examples
  - **Spam fltering**: regress the probability of an email being a spam message against thousands of input variables, such as email address, xtc-dh4r-e4xxa-dd-x76oi@..., email subject,TOP SALE ...)
  - **Income prediction**: may depend on Gender, Education Level, Age, ...

Multiple Regression

- Multiple Regression generalizes the simple linear regression mean function

$$E[Y|X_1 = x_1] = \beta_0 + \beta_1 x_1$$

to

$$E[Y|X_1 = x_1, X_2 = x_2, \cdots, X_p = x_p] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

The main idea in adding $X_2, \ldots, X_p$ is to explain the part of Y that has not already been explained by X_1.

- Equivalently, for $i = 1, \cdots, n$,

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p} + e_i.$$

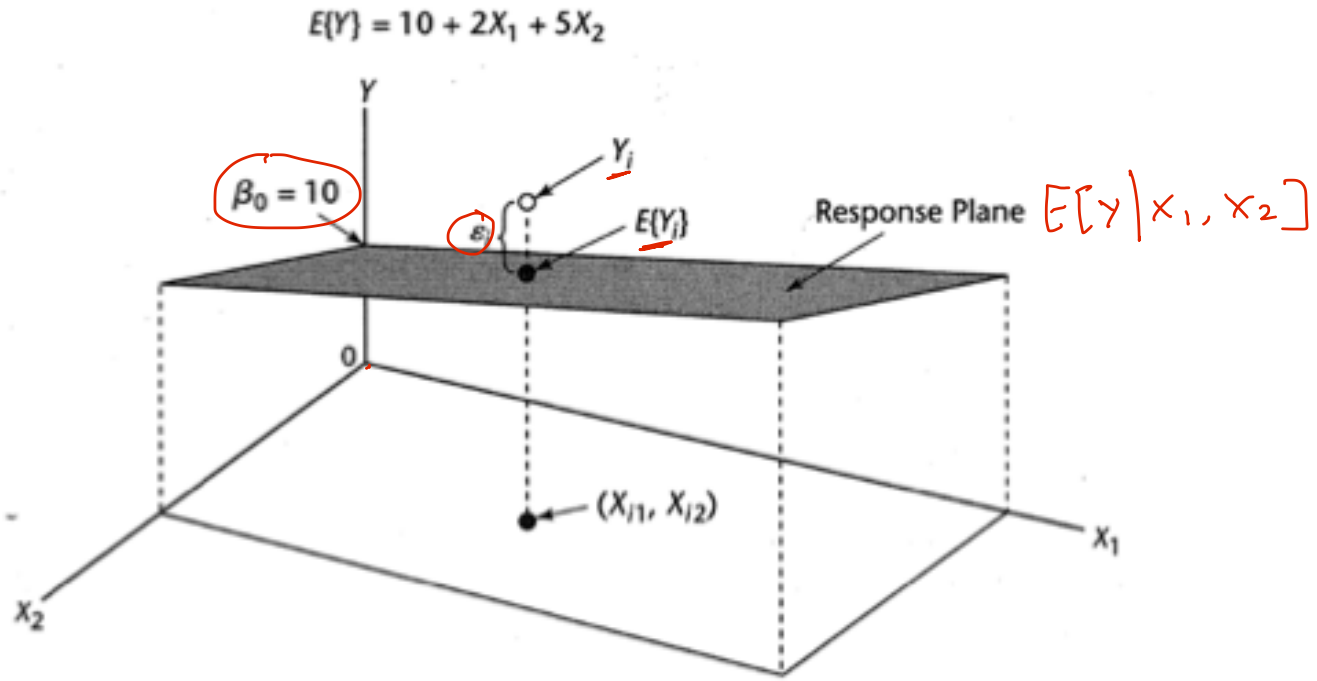
$X_{i,1}, \cdots, X_{i,p}$ are the values of p predictor variables for the i-th subject.

Multiple Regression p=2

- When $p = 2$, it becomes

$$E[Y|X_1 = x_1, X_2 = x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

- An example: $E[Y|X_1 = x_1, X_2 = x_2] = 10 + 2x_1 + 5x_2$.



Multiple Regression p=2

- An example: $E[Y|X_1 = x_1, X_2 = x_2] = 10 + 2x_1 + 5x_2$.
- β_0 is the intercept when both x_1 and x_2 are zero;
- β_1 indicates the change in the mean response EY per unit increase in x_1 when x_2 is held constant. β_2 -vice versa.
- Fix $x_2 = 3$, then

$$E[Y|x] = 10 + 2x_1 + 5 \times 3 = 25 + 2x_1$$

becomes a simple linear function.

- If we take $x_2 = x_1^2$, the model becomes polynomial regression.

Multiple Regression

From the initial collection of potential predictors, we have computed a set of $p+1$ regressors, including an intercept, $X = (1, X_1, \ldots, X_p)^T$. The mean function and variance function for multiple linear regression are

$$\begin{cases} \mathrm{E}(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = \beta^T X \\ \mathrm{Var}(Y|X) = \sigma^2 \end{cases}$$

Both the βs and σ^2 are unknown parameters that are to be estimated.

Let $\beta_{(p+1)\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}_{(p+1)\times 1}$.

from observed Data.

n independent and identically distributed (i.i.d) observations

(Y_i, X_i) $i = 1, \cdots, n$

$X_i = (1, X_{i,1}, \cdots X_{i,p})^T$

Multiple Regression Matrix Notation

- We write $Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p} + e_i, i = 1, \cdots, n$ as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

with

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}_{n\times(p+1)} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}_{(p+1)\times 1} \quad \text{and} \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}_{n\times 1}$$

Assumption of e

① Assumption $E[\underset{n\times 1}{e}|X] \triangleq \begin{pmatrix} E(e_1|X) \\ E(e_2|X) \\ \vdots \\ E(e_n|X) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$

② Conditional on X, Covariance matrix of $\underset{n\times 1}{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}_{n\times 1}$, denoted by $\mathrm{Var}(\underset{n\times 1}{e}|X)$ or $\mathrm{Cov}(\underset{n\times 1}{e}|X)$, is a $n \times n$ matrix with the (i,j)th entry being $\mathrm{Cov}(e_i, e_j|X)$. (row, column; covariance of e_i and e_j.)

Define the unobservable random vector of errors $\mathbf{e}$ elementwise by $e_i = y_i - \mathrm{E}(Y|X = \mathbf{x}_i) = y_i - \mathbf{x}_i'\boldsymbol{\beta}$, and $\mathbf{e} = (e_1, \ldots, e_n)'$ (transpose). The assumptions concerning the e_is are summarized in matrix form as

$$\underset{n\times 1}{\mathrm{E}(\mathbf{e}|X)} = \underset{n\times 1}{\mathbf{0}} \quad \underbrace{\mathrm{Var}(\mathbf{e}|X)}_{n\times n} = \sigma^2 \mathbf{I}_n$$

$\mathbf{I}_n$: $n\times n$ identity matrix.

The assumption ②: $\mathrm{Var}(e|X) = \sigma^2 I_n$ means that
- when $i = j \le n$, $\mathrm{Cov}(e_i, e_i|X) = \mathrm{Var}(e_i|X) = \sigma^2$
- when $i \neq j$, $\mathrm{Cov}(e_i, e_j|X) = 0$.

where $\mathrm{Var}(\mathbf{e}|X)$ means the covariance matrix of $\mathbf{e}$ ($n\times 1$ vector) for a fixed value of X, $\mathbf{I}_n$ is the $n \times n$ matrix with ones on the diagonal and zeroes everywhere else, and $\mathbf{0}$ is a matrix or vector of zeroes of appropriate size. If we add the assumption of normality, we can write

$$\underset{n\times 1}{(\mathbf{e}|X)} \sim \mathrm{N}(\underbrace{\mathbf{0}}_{\text{mean}}, \underbrace{\sigma^2\mathbf{I}_n}_{\text{covariance matrix}})$$

← n-dimensional Normal dist. (Gaussian)

$\Leftrightarrow$ $e_1, \ldots, e_n | X$ are independent, and for $i = 1, \ldots, n$

$$e_i|X \sim N(0, \sigma^2).$$

Multiple Regression Example

Example: The goal of this example is to understand how fuel consumption varies over the 50 United States and the District of Columbia.

Table 1.1 Variables in the Fuel Consumption Data[a]

`Drivers`	Number of licensed drivers in the state
`FuelC`	Gasoline sold for road use, thousands of gallons
`Income`	Per person personal income for the year 2000, in thousands of dollars
`Miles`	Miles of Federal-aid highway miles in the state
`Pop`	2001 population age 16 and over
`Tax`	Gasoline state tax rate, cents per gallon
`Fuel`	1000 × `Fuelc/Pop`
`Dlic`	1000 × `Drivers/Pop`
log(`Miles`)	Natural logarithm of `Miles`

[a]All data are for 2001, unless otherwise noted. The last three variables do not appear in the data file, but are computed from the previous variables, as described in the text.

Figure: Data *Fuel2001* on `http://users.stat.umn.edu/~sandy/alr4ed/data/`

Multiple Regression Example

$$
\begin{aligned}
&Y(\textit{State Fuel Consumption}) \\
&= \beta_0 + \beta_1 X_1(\textit{GasTax}) + \beta_2 X_2(\textit{Divers/Population}_{>16} * 1000) \\
&\quad + \beta_3 X_3(\textit{Income/person}) + \beta_4 X_4(\textit{logMills of Fed. highway}) + e(\textit{Error}).
\end{aligned}
$$

$$
\mathbf{X} = \begin{pmatrix}
1 & 18.00 & 1031.38 & 23.471 & 16.5271 \\
1 & 8.00 & 1031.64 & 30.064 & 13.7343 \\
1 & 18.00 & 908.597 & 25.578 & 15.7536 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
1 & 25.65 & 904.894 & 21.915 & 15.1751 \\
1 & 27.30 & 882.329 & 28.232 & 16.7817 \\
1 & 14.00 & 970.753 & 27.230 & 14.7362
\end{pmatrix}
\qquad
\mathbf{Y} = \begin{pmatrix}
690.264 \\
514.279 \\
621.475 \\
\vdots \\
562.411 \\
581.794 \\
842.792
\end{pmatrix}
$$

51×5

The first row of $\mathbf{X}$ is $\mathbf{x}_1' = (1, 18.00, 1031.38, 23.471, 16.5271)'$, and the first row of $\mathbf{Y}$ is $y_1 = 690.264$, an ordinary number or scalar. The regressors in $\mathbf{X}$ are in the order intercept, `Tax`, `Dlic`, `Income`, and finally, log(`Miles`). The matrix $\mathbf{X}$ is 51×5 and $\mathbf{Y}$ is 51×1.

Examination of the summary graph

More than two variables? Use Scatterplot Matrix ("pairs" function in R).

```
pairs(~Tax+Dlic+Income+log(Miles)+Fuel,data=fuel2001).
```

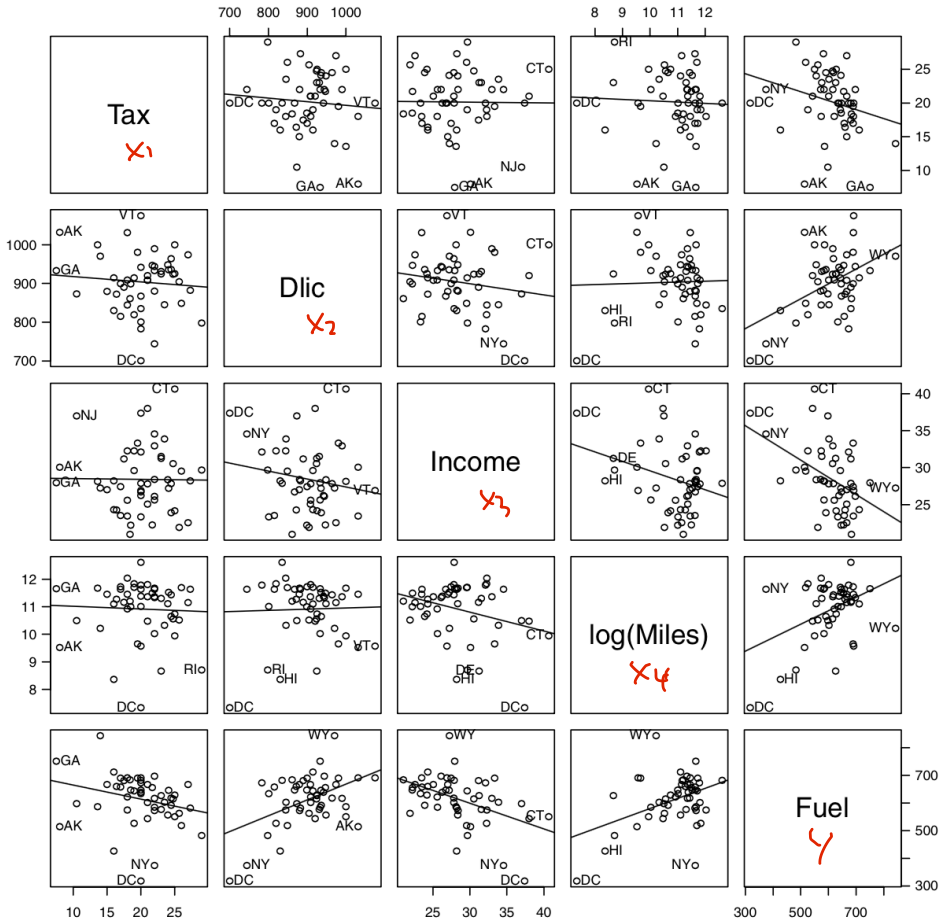


Figure 3.4 Scatterplot matrix for the fuel data.

Ordinary Least Squares Estimators

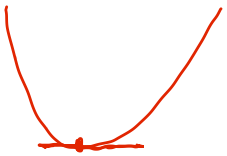
- Ordinary Least Squares, or OLS.
- The estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)^\top$ are chosen to minimize the residual sum of squares (RSS):

objective function

$$J(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i - \boldsymbol{\beta}^\top \mathbf{x}_i]^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p})]^2.$$

- Computation: "lm()" function in R.

from calculus, the minimum of a (smooth) function is achieved at a point where the gradient vanishes ($=0$)



To get $\hat{\beta}$, we find it from $\nabla_{\beta_{(p+1)\times 1}} J(\beta)_{(p+1)\times 1} = 0$ (next lecture)