Linear regression

Hastie, Tibshirani, Friedman Ch 6-7

Kevin Murphy Ch. 7

CS 6140

Machine Learning

Professor Olga Vitek

January 19, 2017

Generative vs discriminative models

- Goal: predict Y
  - Bayes rule:

$$p(Y|\mathbf{X}) = \frac{p(Y) \cdot p(\mathbf{X}|Y)}{p(\mathbf{X})}$$

- Generative classifiers
  - Specify prior probability of $p(Y)$
  - Assume conditional distribution $p(\mathbf{X}|Y)$
  - Use Bayes rule to derive the posterior $p(Y|\mathbf{X})$
  - **Example:** Linear discriminant analysis

- Discriminative classifiers
  - Estimate the posterior the posterior $p(Y|\mathbf{X})$
  - Do not assume the distribution on $\mathbf{X}$
  - **Example:** Y continuous: linear regression

Linear regression with two predictors

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i; \qquad i = 1, \ldots, n$$

- β_0 is the intercept
- β_1 and β_2 are the regression coefficients
- Meaning of regression coefficients
  - β_1 describes change in <u>mean response</u> per unit increase in X_1 when X_2 is held constant
  - β_2 describes change in <u>mean response</u> per unit increase in X_2 when X_1 is held constant

- Variables X_1 and X_2 are **additive**.
- Same change in X_1 for all X_2.
- The response surface is a plane.

Interaction model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

- Meaning of parameters:
  - Change in X_1 when $X_2 = x_2$

$$\begin{aligned} \Delta Y &= (\beta_0 + \beta_1(X_1+1) + \beta_2 x_2 + \beta_3 (X_1+1) x_2) - \\ &\quad (\beta_0 + \beta_1 X_1 + \beta_2 x_2 + \beta_3 X_1 x_2) \\ &= \beta_1 + \beta_3 x_2 \end{aligned}$$

  - Change in X_2 when $X_1 = x_1$

$$\Delta Y = \beta_2 + \beta_3 x_1$$

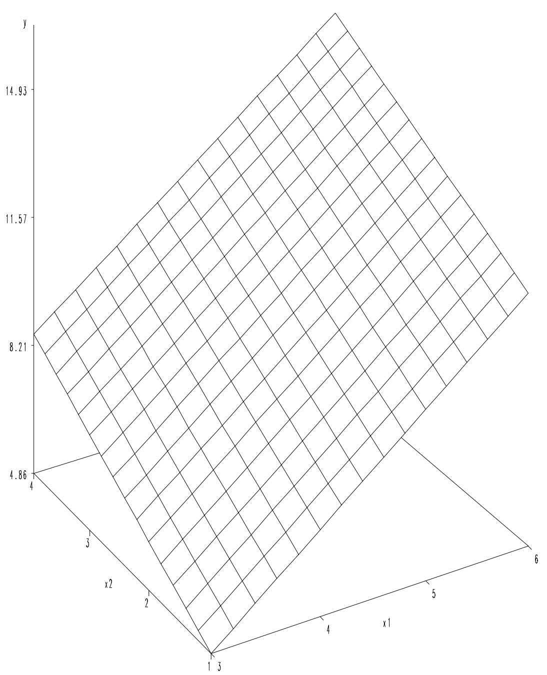
- Rate of change due to one variable affected by the other

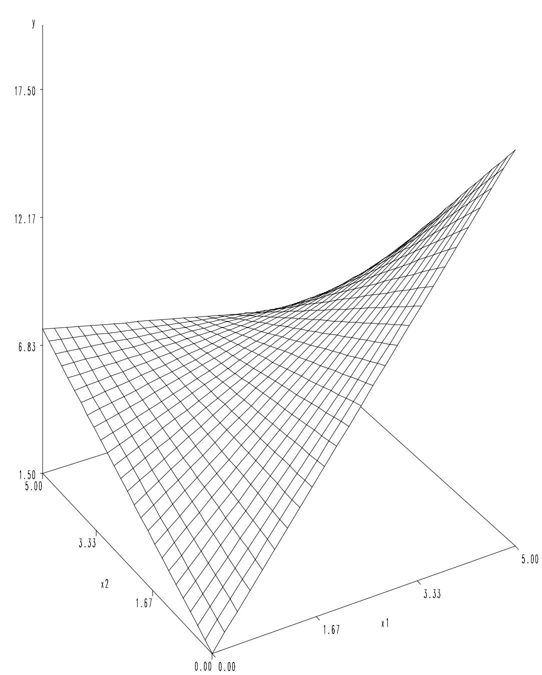
Additive vs interaction model

$$\widehat{Y}_i = -2.79 + 2.14X_{i1} + 1.21X_{i2}$$

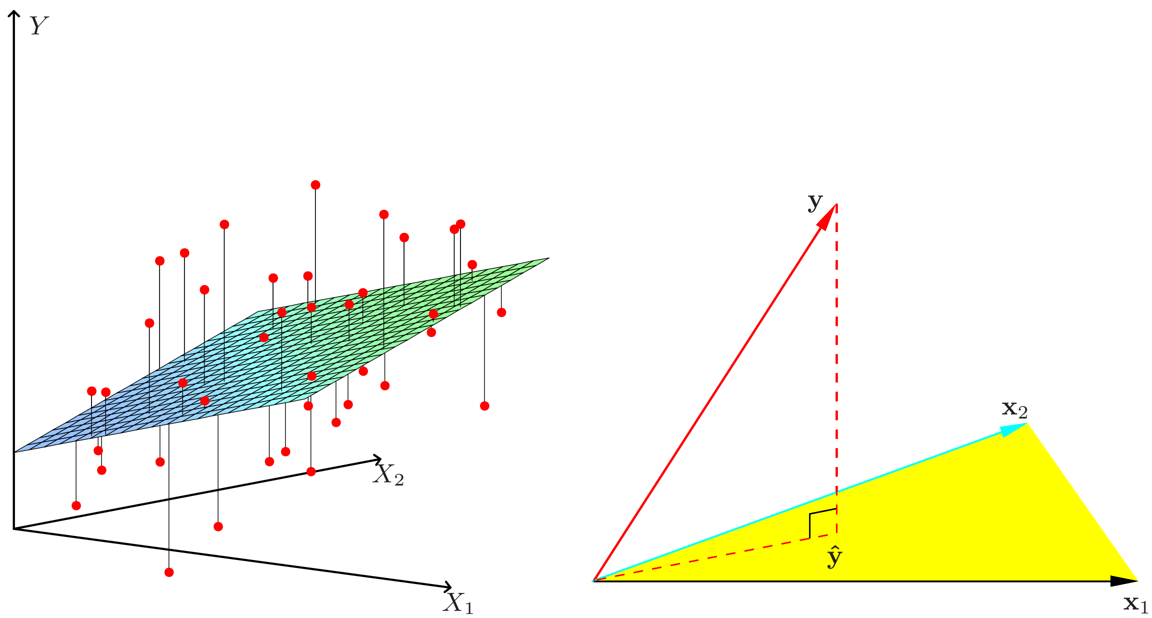
versus

$$\widehat{Y}_i = 1.5 + 3.2X_{i1} + 1.2X_{i2} - .75X_{i1}X_{i2}$$

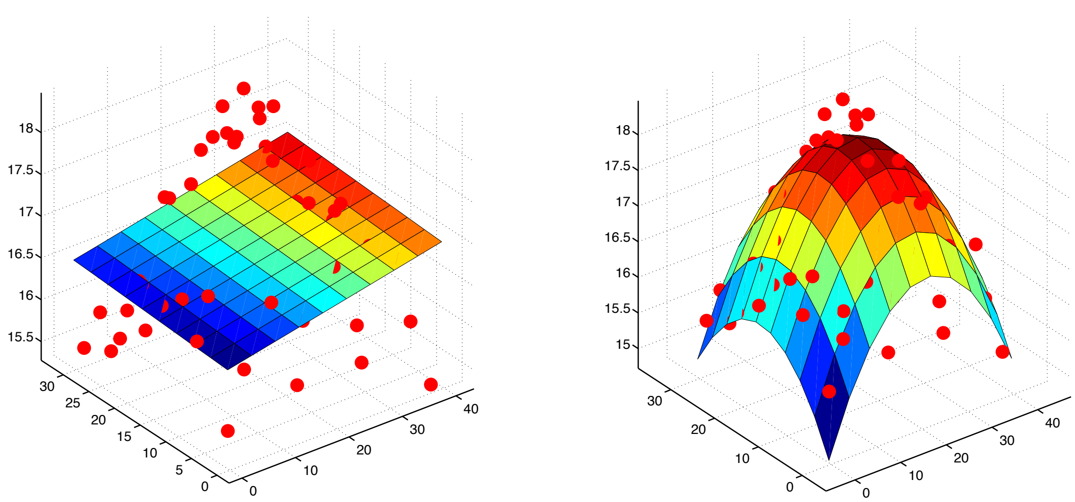




Linear regression with two predictors



Hastie, Tibshirani, Friedman, Fig 3.1 and 3.2



K. Murphy, Fig 7.1

Polynomial regression and transformations

- Polynomial regression:

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i \\
&= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i
\end{aligned}
$$

where $X_{i2} = X_i^2$.

  - this is a linear model because it is a linear function of parameters β

- Transformations

$$
log(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i
$$

$$
Y_i = \frac{1}{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i}
$$

  - this is a linear model on the $log(Y_i)$ scale

General linear regression in matrix terms

- As an equation

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

- As an array

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1\ p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2\ p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n\ p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \cdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

- In matrix notation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

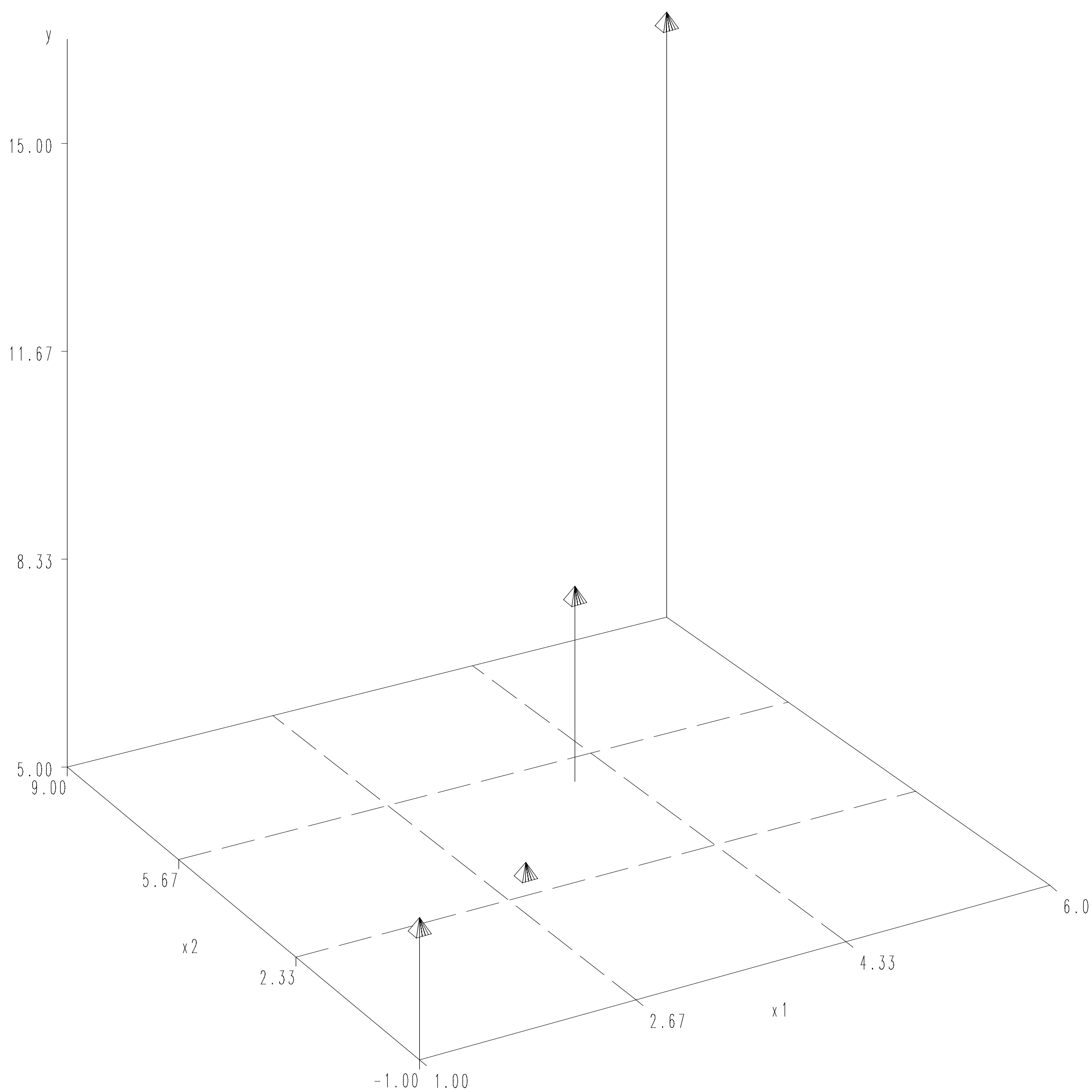
Estimation of regression coefficients

- Objective function: least squares
  - find $\hat{\beta}$ to minimize
    $$\sum_{i=1}^{N}(y_i - x_i'\beta)^2 = (\mathbf{Y} - \mathbf{X}\widehat{\beta})'(\mathbf{Y} - \mathbf{X}\widehat{\beta})$$
  - Quadratic objective function $\Rightarrow$
    its minimum always exists, but may not be unique

- Finding estimates
  - Differentiating wrt β:
  - Normal equations $\mathbf{X}'(y - \mathbf{X}\beta) = 0 \Rightarrow \widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

- Fitted values define a (hyper)plane
  - $\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{HY}$
  - Residuals: $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$

Multicollinearity



Qualitative predictors

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

- Let $X_2 = 1$ if case from Massachusetts
- Meaning of parameters:
  - Case from Massachusetts ($X_2 = 1$):

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 1 + \beta_3 X_1 (1) \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 \end{aligned}$$

  - Case from other location ($X_2 = 0$)

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 0 + \beta_3 X_1 (0) \\ &= \beta_0 + \beta_1 X_1 \end{aligned}$$

- Have <u>two</u> regression lines
- β_2 and β_3 quantify the differences

Two groups: Wrong coding

- Assume an additive model with two groups

- Wrong approach: add both indicators

$$X_2 = \begin{cases} 1 \text{ , if stock firm} \\ 0 \text{ , otherwise} \end{cases} \qquad X_3 = \begin{cases} 1 \text{ , if mutual fund} \\ 0 \text{ , otherwise} \end{cases}$$

  - the model below is wrong

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

- The corresponding design matrix
  - 4 data points (first 2 from stock firm, last 2 from mutual fund)

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & 1 & 0 \\ 1 & X_{21} & 1 & 0 \\ 1 & X_{31} & 0 & 1 \\ 1 & X_{41} & 0 & 1 \end{pmatrix}$$

  - this model creates fully collinear columns in the design matrix $\mathbf{X}$ (R will drop the first)

Two groups: Correct coding

- Correct approach 1:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

  - interpretation:

$$\begin{array}{rl} E\{Y_i\} = \beta_0 + \beta_1 X_{i1} & \quad \text{if mutual fund} \\ E\{Y_i\} = (\beta_0 + \beta_2) + \beta_1 X_{i1} & \quad \text{if stock firm} \end{array}$$

  - Mutual fund is the reference group
  - β_2: the deviation of the intercept of the stock firm from the reference

- The corresponding design matrix:

  - 4 data points (first 2 from stock firm, last 2 from mutual fund)

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & 1 \\ 1 & X_{21} & 1 \\ 1 & X_{31} & 0 \\ 1 & X_{41} & 0 \end{pmatrix}$$

Three groups: Wrong coding

- Extend the indicator

$$X_2 = \begin{cases} 0 \text{ , if mutual fund} \\ 1 \text{ , if stock firm} \\ 2 \text{ , if foreign firm} \end{cases}$$

- The model below is still appropriate

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

  - interpretation: enforces an equal change in $E\{Y\}$ for each extra indicator

$$\begin{aligned} E\{Y_i\} &= \beta_0 + \beta_1 X_{i1} && \text{if mutual fund} \\ E\{Y_i\} &= (\beta_0 + \beta_2) + \beta_1 X_{i1} && \text{if stock firm} \\ E\{Y_i\} &= (\beta_0 + 2\beta_2) + \beta_1 X_{i1} && \text{if foreign firm} \end{aligned}$$

- The corresponding design matrix:

  - 6 data points (first 2 from mutual fund, 2 from stock, 2 foreign)

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & 0 \\ 1 & X_{21} & 0 \\ 1 & X_{31} & 1 \\ 1 & X_{41} & 1 \\ 1 & X_{41} & 2 \\ 1 & X_{41} & 2 \end{pmatrix}$$

Three groups: Correct coding

- First option:

$$X_2 = \begin{cases} 1 \text{ , if stock firm} \\ 0 \text{ , otherwise} \end{cases} \qquad X_3 = \begin{cases} 1 \text{ , if foreign firm} \\ 0 \text{ , otherwise} \end{cases}$$

- The model below contains two indicators

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

  - interpretation:

$$\begin{array}{rl} E\{Y_i\} = \beta_0 + \beta_1 X_{i1} & \text{if mutual fund} \\ E\{Y_i\} = (\beta_0 + \beta_2) + \beta_1 X_{i1} & \text{if stock firm} \\ E\{Y_i\} = (\beta_0 + \beta_3) + \beta_1 X_{i1} & \text{if foreign firm} \end{array}$$

  - mutual fund is the reference
  - β_2 and β_3 are deviations of the intercepts from the reference
  - also more flexibility in presence of interactions X_1X_2 and X_1X_3
  - the number of indicators is always one less than the number of groups

Normal Error Model

- **The least square estimates of the parameters do not require the assumption of Normality**

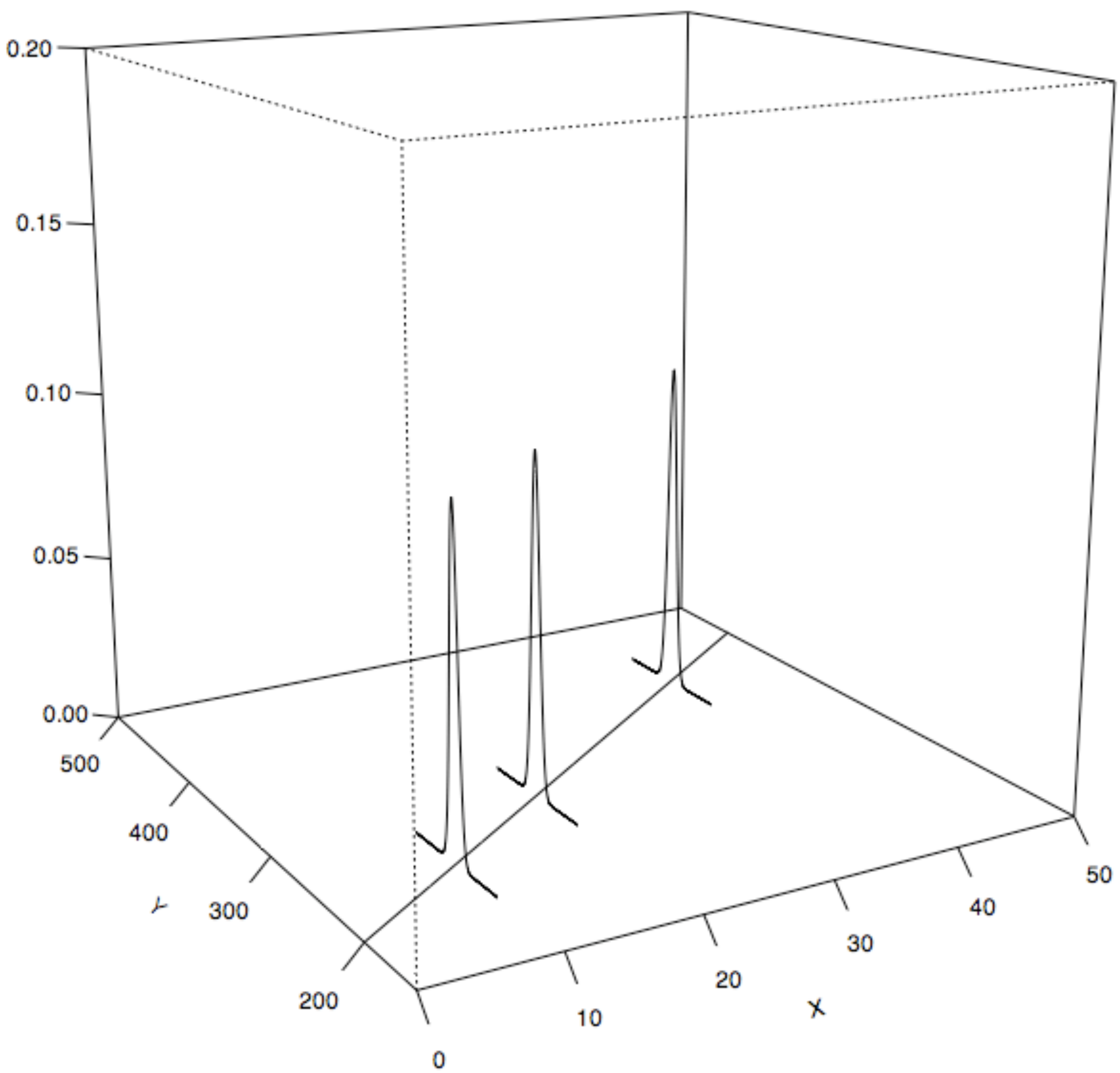
- Normal error assumption greatly simplifies the theory of analysis

- Normality is used to construct confidence intervals / perform hypothesis tests follow known distributions (e.g., t, F)

- While not always true in practice, most inference only sensitive to large departures from normality

Normal Error regression model

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$



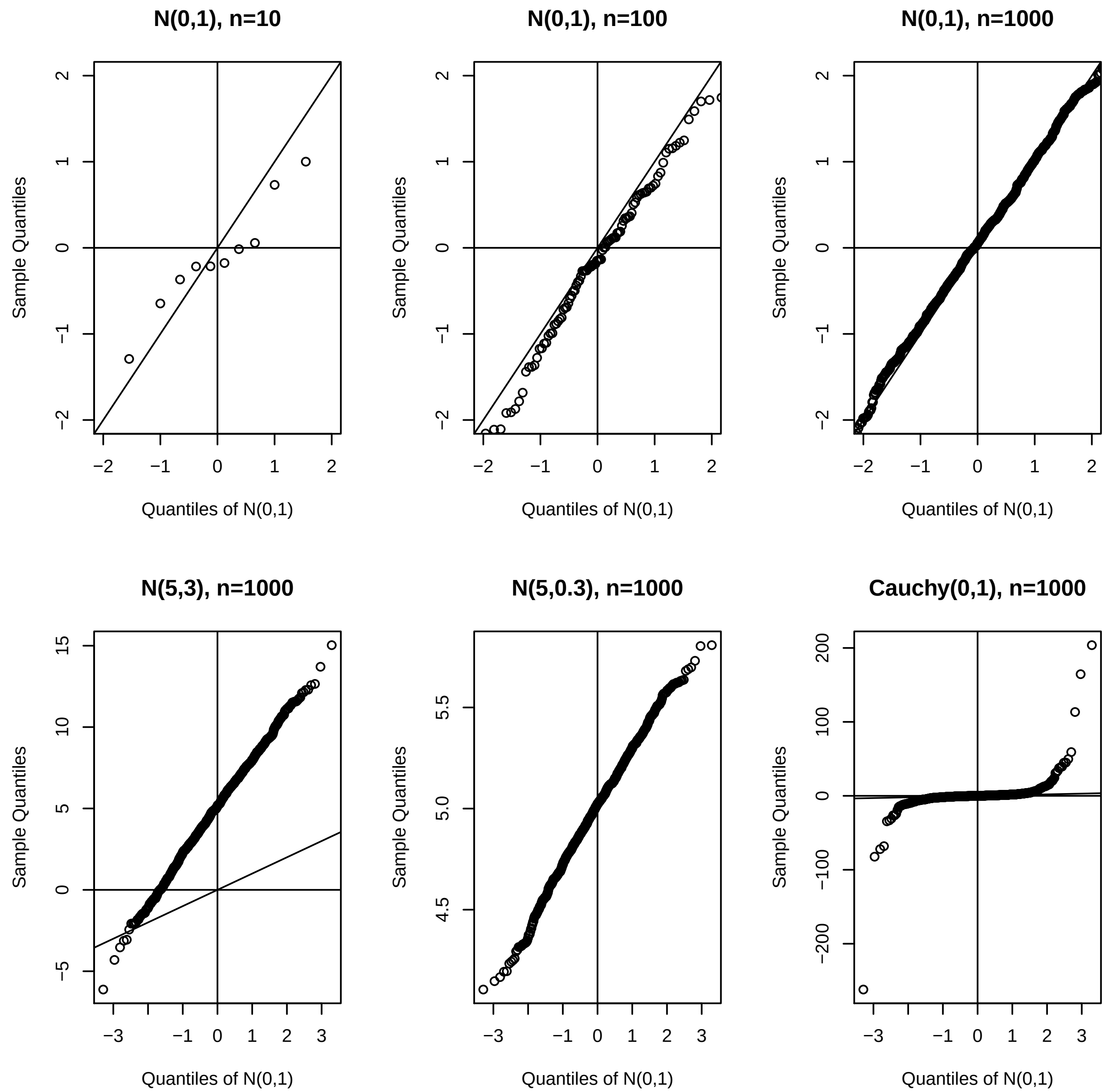
Normal Error regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- β_0 is the intercept
- β_1 in the slope
- ε_i is the i^{th} random error term
  - $\varepsilon_i \sim \mathrm{N}(0, \sigma^2) \longleftarrow$ **NEW**
  - Uncorrelated $\longrightarrow$ independent error terms
- Defines distribution of Y: $p(Y|\mathbf{X})$

$$Y_i \sim \mathrm{N}(\beta_0 + \beta_1 X_i, \sigma^2)$$

Assessing Normality: Quantile-quantile plot



Can be used with any other distribution

Example

Height of 11 women

i	Observed height	Adj. percentile $100(i-\frac{1}{2})/11)$	z	Sample quantiles
1	61.0	4.55	-1.69	60.6
2	62.5	13.64	-1.10	62.3
3	63.0	22.73	-0.75	63.4
4	64.0	31.82	-0.47	64.1
5	64.5	40.91	-0.23	64.8
6	65.0	50.00	0.00	65.5
7	66.5	59.09	0.23	66.2
8	67.0	68.18	0.47	66.9
9	68. 0	77.27	0.75	67.6
10	68.5	86.36	1.10	68.7
11	70.5	95.45	1.69	70.4

QQplot: plot Observed height vs sample quantiles

Sample quantiles $= x + Z \cdot \widehat{\sigma} + \widehat{\mu}$

```
> ?qqplot
> ?qqnorm
```

Maximum Likelihood Estimation

- Assumption of Normality gives us more choices of methods for parameter estimation

$$Y_i \sim \mathsf{N}(\beta_0 + \beta_1 X_i, \sigma^2)$$
$$\downarrow$$
$$f_i = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2\right\}$$

- Likelihood function $L = f_1 \times f_2 \times \cdots \times f_n$ (i.e. the joint probability distribution of the observations, viewed as function of parameters)

- Find β_0, β_1 and σ^2 which maximizes L

- Obtain same estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$

- A slightly smaller estimate of σ^2