Lecture 14
Multiple Linear Regression and Logistic Regression

Fall 2013

Prof. Yao Xie, yao.xie@isye.gatech.edu

H. Milton Stewart School of Industrial Systems & Engineering

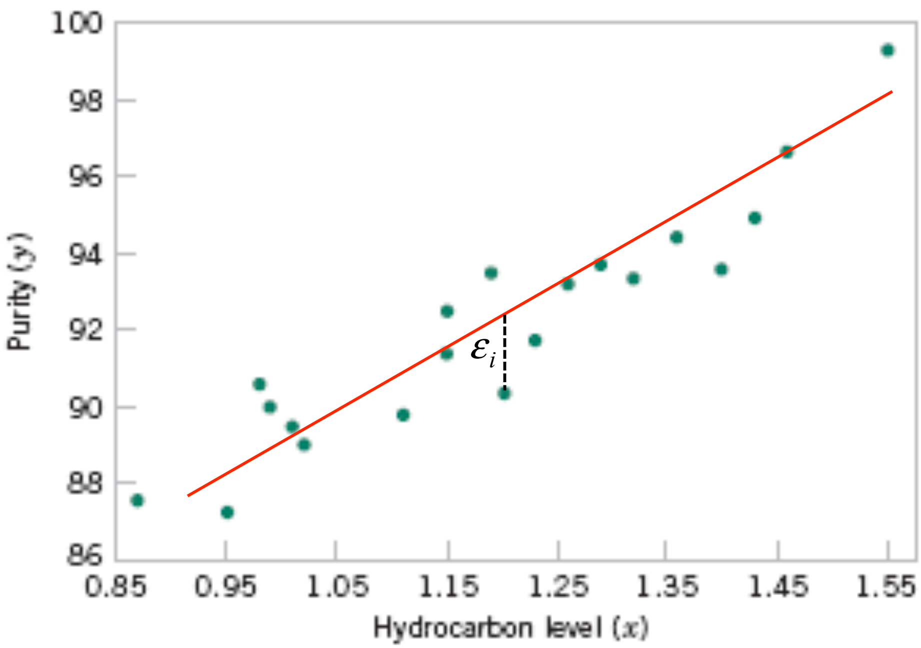
Georgia Tech

Outline

- Multiple regression
- Logistic regression

Simple linear regression

Based on the scatter diagram, it is probably reasonable to assume that the mean of the random variable Y is related to X by the following simple linear regression model:



Response

Regressor or Predictor

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1, 2, \cdots, n$$

$$\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$$

Intercept

Slope

Random error

where the slope and intercept of the line are called regression coefficients.

•The case of simple linear regression considers a single regressor or predictor x and a dependent or response variable Y.

Multiple linear regression

- Simple linear regression: one predictor variable x
- Multiple linear regression: multiple predictor variables $x_1, x_2, \ldots, x_k$
- Example:
  - simple linear regression

    property tax = a*house price + b
  - multiple linear regression

    property tax = a_1*house price + a_2*house size + b
- Question: how to fit multiple linear regression model?

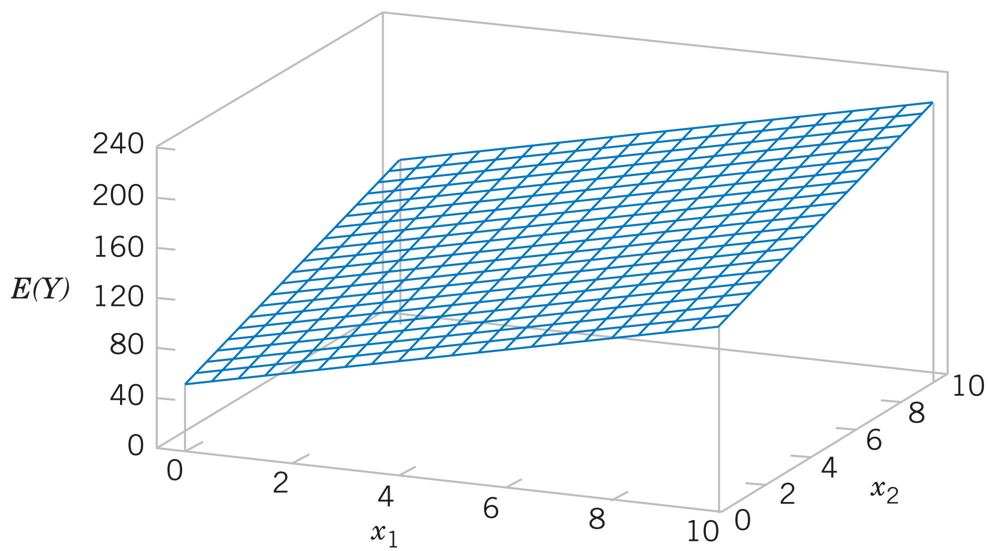
Multiple linear regression model

- Multiple linear regression model with two regressors (predictor variables)

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

dependent variable **independent**

response **regressor variables**



(a)

$$E(Y) = 50 + 10x_1 + 7x_2$$

More complex models can still be analyzed using multiple linear regression

- Cubic polynomial

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$$

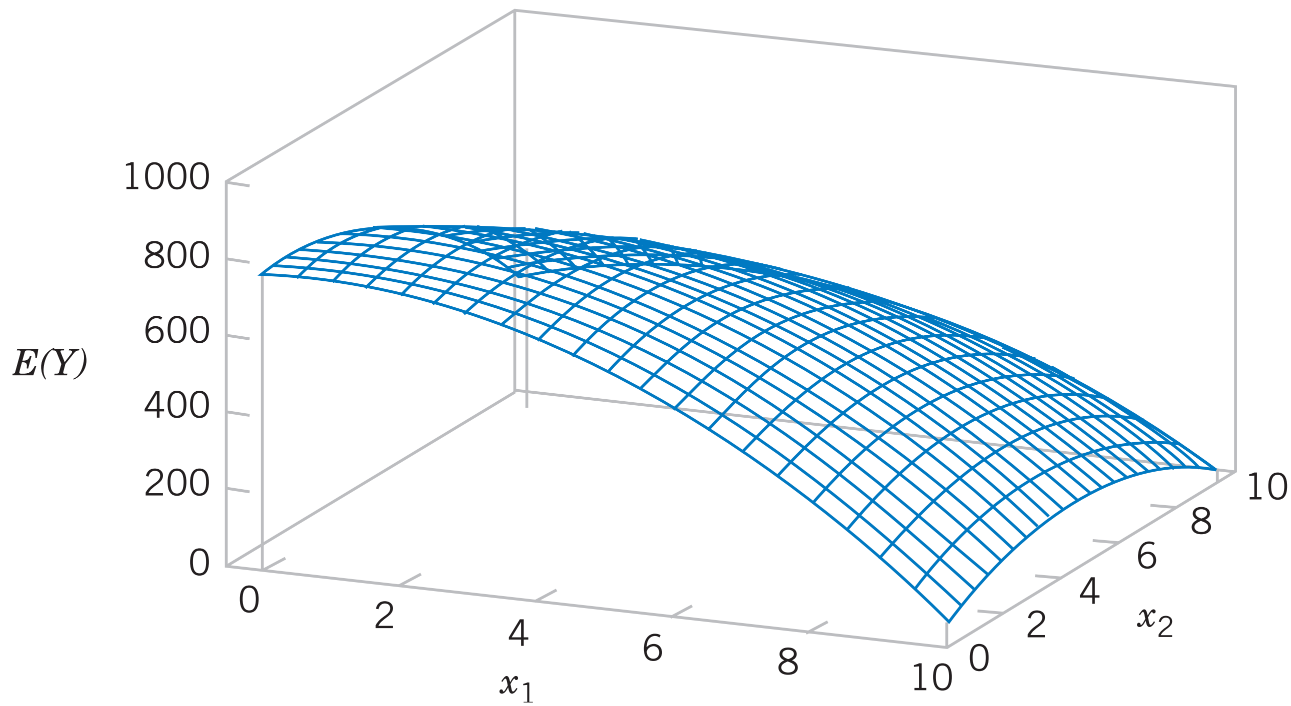
$$\text{let } x_1 = x, x_2 = x^2, x_3 = x^3$$

- Interaction effect

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$

$$\text{let } x_3 = x_1 x_2 \text{ and } \beta_3 = \beta_{12}$$

In general, **any regression model that is linear in parameters** (the β's) **is a linear regression model, regardless of the shape of the surface that it generates.**



$$E(Y) = 800 + 10x_1 + 7x_2 - 8.5x_1^2 - 5x_2^2 + 4x_1x_2$$

Data for multiple regression

Table 12-1 Data for Multiple Linear Regression

y	x_1	x_2	$\ldots$	x_k
y_1	x_{11}	x_{12}	$\ldots$	x_{1k}
y_2	x_{21}	x_{22}	$\ldots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	x_{n1}	x_{n2}	$\ldots$	x_{nk}

Data $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i), \qquad i = 1, 2, \ldots, n \quad \text{and} \quad n > k$

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \\
&= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i \qquad i = 1, 2, \ldots, n
\end{aligned}
$$

Least square estimate of coefficients

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

Set derivatives to 0

$$\left. \frac{\partial L}{\partial \beta_0} \right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) = 0$$

$$\left. \frac{\partial L}{\partial \beta_j} \right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) x_{ij} = 0$$

Normal equations

$$\begin{array}{lllll}
n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} & + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} & + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} & = & \sum_{i=1}^{n} y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 & + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2} & + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1} x_{ik} & = & \sum_{i=1}^{n} x_{i1} y_i \\
\vdots \qquad\quad \vdots & \vdots & \vdots \qquad \vdots & & \vdots \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik} x_{i1} & + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik} x_{i2} & + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 & = & \sum_{i=1}^{n} x_{ik} y_i
\end{array}$$

k+1 normal equations, k+1 coefficients to be determined — can be uniquely determined

Matrix form for multiple linear regression

- Write multiple regression as

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

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Matrix normal equation

- Least square function $L = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$
- Coefficient satisfies $\frac{\partial L}{\partial \boldsymbol{\beta}} = \mathbf{0}$
- Normal equation $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1} y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik} y_i \end{bmatrix}$$

Fitted model

- Fitted model

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

- Residual $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$

- Estimator of variance

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - p} = \frac{SS_E}{n - p}$$

Use R for multiple linear regression

- Fit model using

  lm(response ~ explanatory_1 + explanatory_2 + … + explanatory_p)

- Example

Ex. Data was collected on 100 houses recently sold in a city. It consisted of the sales price (in $), house size (in square feet), the number of bedrooms, the number of bathrooms, the lot size (in square feet) and the annual real estate tax (in $).

Read data

```
> Housing = read.table("C:/Users/Martin/Documents/W2024/housing.txt", header=TRUE)
> Housing
    Taxes Bedrooms Baths  Price Size   Lot
1    1360        3   2.0 145000 1240 18000
2    1050        1   1.0  68000  370 25000
…..
99   1770        3   2.0  88400 1560 12000
100  1430        3   2.0 127200 1340 18000
```

Suppose we are only interested in a subset of variables

We want to fit a linear regression model

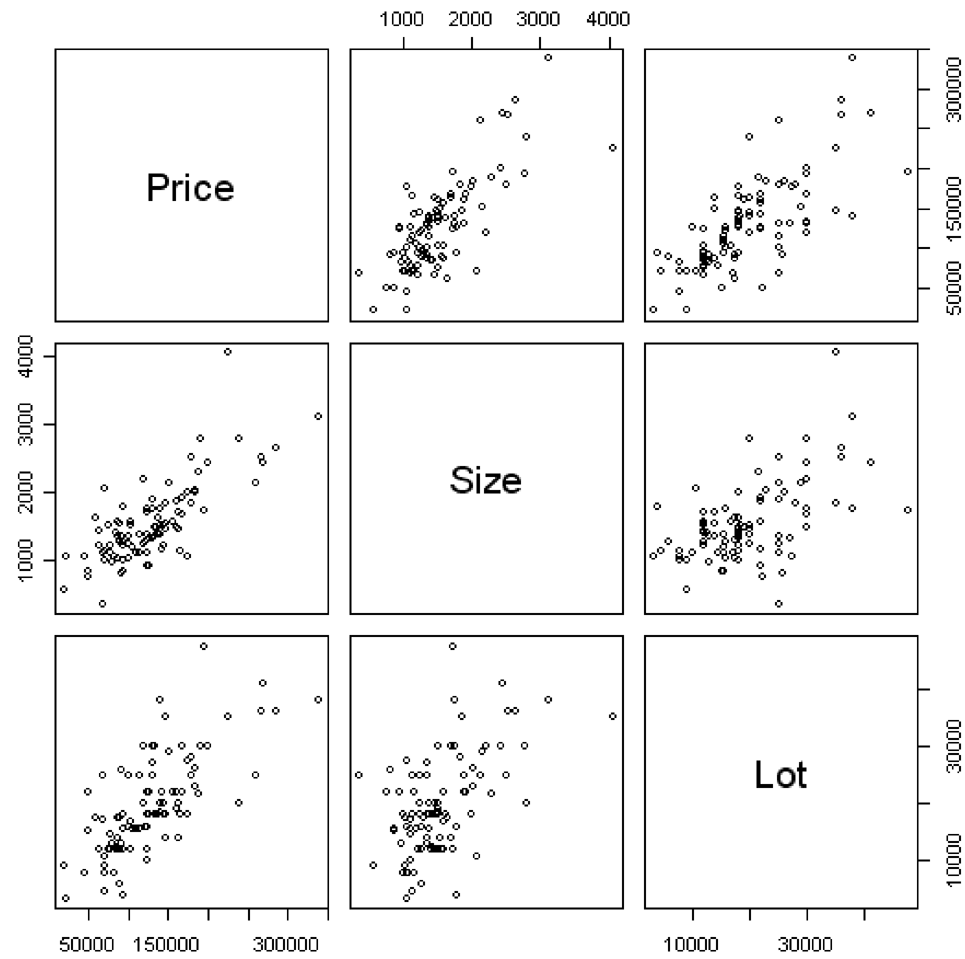
response variable: price

predictor variables: size, lot

Create multiple scatter plot

- scatter plots of all pair-wise combinations of variables we are interested in

```
> myvars = c("Price", "Size", "Lot")
> Housing2 = Housing[myvars]
> plot(Housing2)
```



Fit model

```
> results = lm(Price ~ Size + Lot, data=Housing)
> results

Call:
lm(formula = Price ~ Size + Lot, data = Housing)
Coefficients:
(Intercept)        Size          Lot
 -10535.951       53.779        2.840
```

$$\hat{y} = -10536 + 53.8x_1 + 2.8x_2$$

```
> summary(results)
Call:
lm(formula = Price ~ Size + Lot, data = Housing)
Residuals:
   Min      1Q     Median   3Q     Max
-81681 -19926   2530  17972  84978
Coefficients:
              Estimate     Std. Error    t value    Pr(>|t|)
(Intercept) -1.054e+04   9.436e+03   -1.117     0.267
Size         5.378e+01   6.529e+00    8.237     8.39e-13 ***
Lot          2.840e+00   4.267e-01    6.656    1.68e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 30590 on 97 degrees of freedom
Multiple R-squared: 0.7114,     Adjusted R-squared: 0.7054
F-statistic: 119.5 on 2 and 97 DF,  p-value: < 2.2e-16
```

Introduction to logistic regression

- linear regression:
  - response variable y is **quantitative (real value)**

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- logistic regression:
  - response variable Y only takes two values, {0, 1}

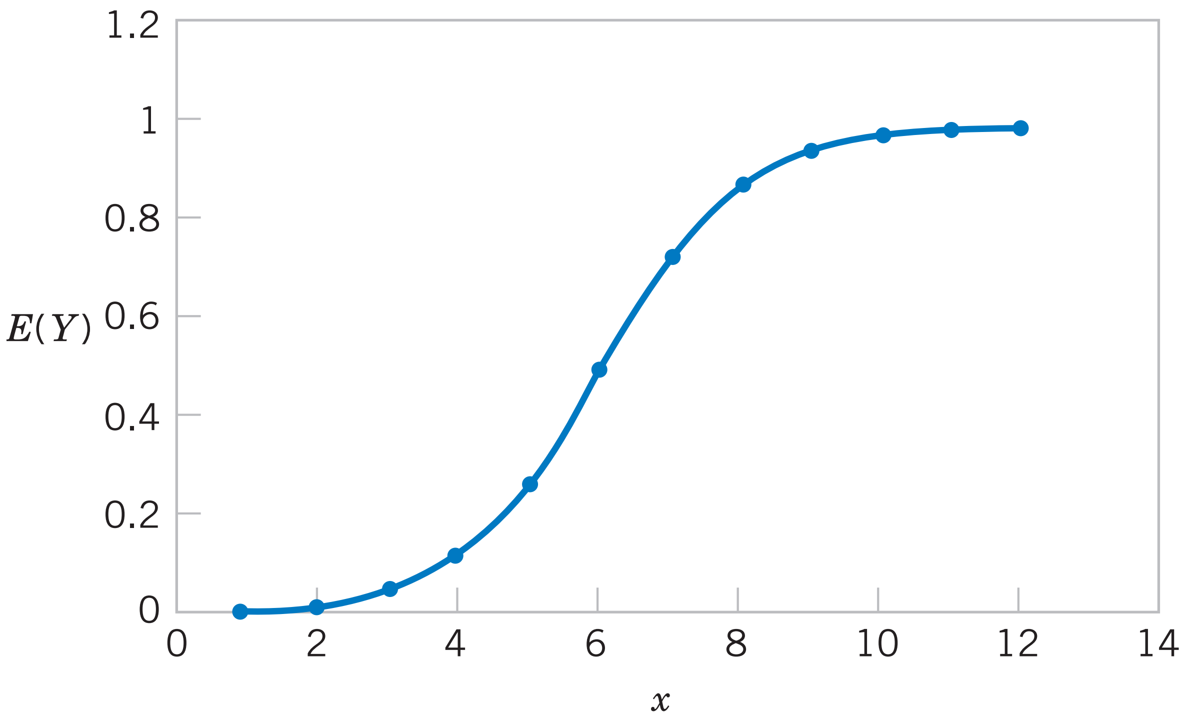
Y_i is a **Bernoulli random variable** with probability distribution

Y_i	Probability
1	$P(Y_i = 1) = \pi_i$
0	$P(Y_i = 0) = 1 - \pi_i$

Logistic response function

$$E(Y_i) = 1(\pi_i) + 0(1 - \pi_i) = \pi_i$$

logit response function, $$E(Y) = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 x)]}$$

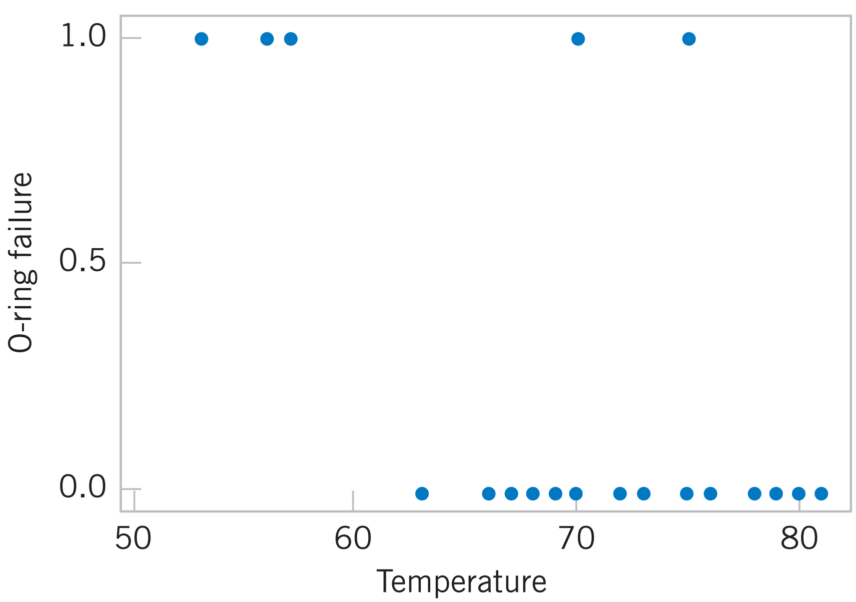


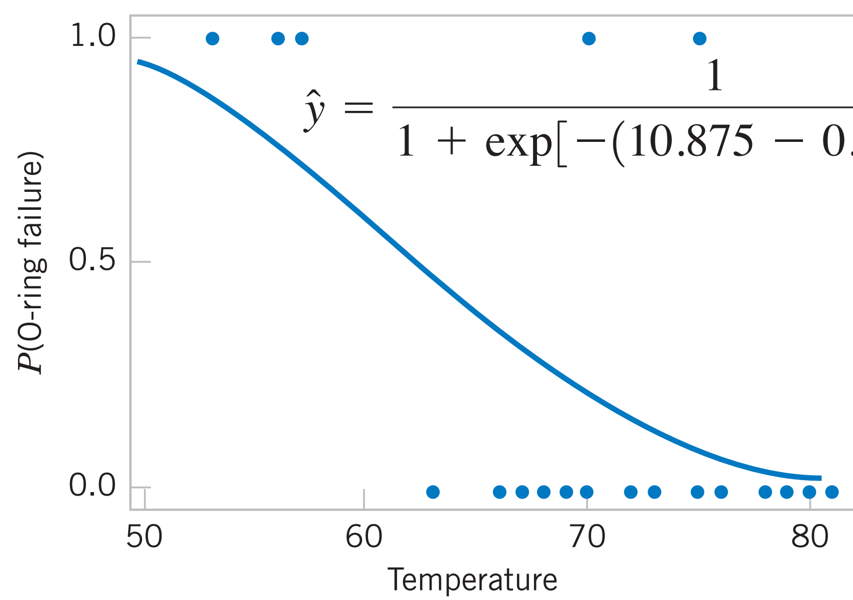
$$E(Y) = 1/(1 + e^{-6.0 - 1.0x})$$

Example

- Failure of machine vs temperature

Temperature	O-Ring Failure	Temperature	O-Ring Failure	Temperature	O-Ring Failure
53	1	68	0	75	0
56	1	69	0	75	1
57	1	70	0	76	0
63	0	70	1	76	0
66	0	70	1	78	0
67	0	70	1	79	0
67	0	72	0	80	0
67	0	73	0	81	0





$$\hat{y} = \frac{1}{1 + \exp[-(10.875 - 0.17132x)]}$$