Chapter 1
Introduction
Sir Francis Galton (1822-1911)

- Galton was a polymath who made important contributions in many fields of science, including meteorology (the anti-cyclone and the first popular weather maps), statistics (regression and correlation), psychology (synesthesia), biology (the nature and mechanism of heredity), and criminology (fingerprints).
- He first introduced the use of questionnaires and surveys for collecting data on human communities.
Karl Pearson (1857 - 1936)

- student of Francis Galton
- He has been credited with establishing the discipline of mathematical statistics, and contributed significantly to the field of biometrics, meteorology, theories of social Darwinism and eugenics
- Founding chair of department of Applied Statistics in University of London (1911), the first stat department in the world!
- Founding editor of *Biometrika*
Incomplete Data

- Due to no direct measurement
- Due to refusal / Don’t know / not available
- Due to uncertainty in the measurement
- Due to design
- Due to self-selection
Example 1: No direct measurement

- A study of managers of Iowa farmer cooperatives \((n = 98)\)
- Five variables
  - \(x_1\): Knowledge (knowledge of the economic phase of management directed toward profit-making in a business and product knowledge)
  - \(x_2\): Value Orientation (tendency to rationally evaluate means to an economic end)
  - \(x_3\): Role Satisfaction (gratification obtained by the manager from performing the managerial role)
  - \(x_4\): Past Training (amount of formal education)
  - \(y\): Role performance
- We are interested in estimating parameters in the regression model

\[
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon
\]
### Example 1 (Cont’d)

<table>
<thead>
<tr>
<th>Measure</th>
<th>No. of Items</th>
<th>Mean</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$ Knowledge</td>
<td>26</td>
<td>1.38</td>
<td>0.6096</td>
</tr>
<tr>
<td>$x_2$ Value orientation</td>
<td>30</td>
<td>2.88</td>
<td>0.6386</td>
</tr>
<tr>
<td>$x_3$ Role satisfaction</td>
<td>11</td>
<td>2.46</td>
<td>0.8002</td>
</tr>
<tr>
<td>$x_4$ Past training</td>
<td>1</td>
<td>2.12</td>
<td>1.0000</td>
</tr>
<tr>
<td>$y$ Role performance</td>
<td>24</td>
<td>0.0589</td>
<td>0.8230</td>
</tr>
</tbody>
</table>
Example 1 (Cont’d)

- Ordinary least squares method

\[ \hat{Y} = -0.9740 + 0.2300X_1 + 0.1199X_2 + 0.0560X_3 + 0.1099X_4 \]
\[ \text{SE} = (0.0535) (0.0356) (0.0375) (0.0392) \]

- Errors-in-variable estimates

\[ \hat{Y} = -1.1828 + 0.3579X_1 + 0.1549X_2 + 0.0613X_3 + 0.0715X_4 \]
\[ \text{SE} = (0.1288) (0.0794) (0.0510) (0.0447) \]

Reference:
## Example 2. Asthma Study Data (Pigott, 2001)

### Variable descriptions

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
<th>Possible values</th>
<th>Mean</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthma belief</td>
<td>Level of confidence</td>
<td>1 = little confidence  5 = lots of confidence</td>
<td>4.057</td>
<td>154</td>
</tr>
<tr>
<td>Group</td>
<td>Treatment or control</td>
<td>0 = treatment  1 = control</td>
<td>0.558</td>
<td>154</td>
</tr>
<tr>
<td>Symsev</td>
<td>Severity of asthma symptoms in 2 weeks</td>
<td>0 = no symptoms  3 = severe symptoms</td>
<td>0.235</td>
<td>141</td>
</tr>
<tr>
<td>Reading</td>
<td>Standardized state reading test scores</td>
<td>Grade equivalent scores, from 1.10 to 8.10</td>
<td>3.443</td>
<td>79</td>
</tr>
<tr>
<td>Age</td>
<td>Ranging from 8 to 14</td>
<td></td>
<td>10.586</td>
<td>152</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td>0 = Male  1 = Female</td>
<td>0.442</td>
<td>154</td>
</tr>
<tr>
<td>Allergy</td>
<td>No. of allergies</td>
<td>Range from 0 to 7</td>
<td>2.783</td>
<td>83</td>
</tr>
</tbody>
</table>
## Missing Data Patterns

<table>
<thead>
<tr>
<th>Symsev</th>
<th>Reading</th>
<th>Age</th>
<th>Allergy</th>
<th># of cases</th>
<th>% of cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>19</td>
<td>12.3</td>
</tr>
<tr>
<td>M</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td>O</td>
<td>M</td>
<td>O</td>
<td>O</td>
<td>54</td>
<td>35.1</td>
</tr>
<tr>
<td>O</td>
<td>O</td>
<td>O</td>
<td>M</td>
<td>56</td>
<td>36.4</td>
</tr>
<tr>
<td>M</td>
<td>M</td>
<td>O</td>
<td>O</td>
<td>9</td>
<td>5.8</td>
</tr>
<tr>
<td>M</td>
<td>O</td>
<td>O</td>
<td>M</td>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td>O</td>
<td>M</td>
<td>O</td>
<td>M</td>
<td>10</td>
<td>6.5</td>
</tr>
<tr>
<td>O</td>
<td>O</td>
<td>M</td>
<td>M</td>
<td>2</td>
<td>1.3</td>
</tr>
<tr>
<td>M</td>
<td>M</td>
<td>O</td>
<td>M</td>
<td>2</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>154</td>
<td>100.0</td>
</tr>
</tbody>
</table>
Results (CC: Complete Case, ML: Maximum Likelihood)

<table>
<thead>
<tr>
<th>Variable</th>
<th>CC analysis</th>
<th></th>
<th>ML analysis</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>SE</td>
<td>B</td>
<td>SE</td>
</tr>
<tr>
<td>Intercept</td>
<td>4.617</td>
<td>0.838</td>
<td>4.083</td>
<td>0.362</td>
</tr>
<tr>
<td>Trt group</td>
<td>-0.550</td>
<td>0.276</td>
<td>-0.132</td>
<td>0.112</td>
</tr>
<tr>
<td>Symsev</td>
<td>-0.315</td>
<td>0.161</td>
<td>-0.480</td>
<td>0.144</td>
</tr>
<tr>
<td>Reading</td>
<td>0.409</td>
<td>0.096</td>
<td>0.218</td>
<td>0.039</td>
</tr>
<tr>
<td>Age</td>
<td>-0.211</td>
<td>0.115</td>
<td>-0.089</td>
<td>0.043</td>
</tr>
<tr>
<td>Gender</td>
<td>0.198</td>
<td>0.189</td>
<td>0.084</td>
<td>0.104</td>
</tr>
<tr>
<td>Allergy</td>
<td>-0.005</td>
<td>0.057</td>
<td>0.063</td>
<td>0.029</td>
</tr>
</tbody>
</table>

Reference:
Example 3: 2009 Local Area Labor Force survey in Korea.

- Large scale survey with about \( n = 157K \) sample households.
- Obtain the employment status: Employed, Unemployed, Not in labor force.
- To obtain response, interviewers visit the sample households up to four times. That is, the current rule allows for three follow-ups.
Realized Responses from the Korean LF survey data

<table>
<thead>
<tr>
<th>status</th>
<th>t=1</th>
<th>t=2</th>
<th>t=3</th>
<th>t=4</th>
<th>No response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employment</td>
<td>81,685</td>
<td>46,926</td>
<td>28,124</td>
<td>15,992</td>
<td></td>
</tr>
<tr>
<td>Unemployment</td>
<td>1,509</td>
<td>948</td>
<td>597</td>
<td>352</td>
<td>32,350</td>
</tr>
<tr>
<td>Not in LF</td>
<td>57,882</td>
<td>32,308</td>
<td>19,086</td>
<td>10,790</td>
<td></td>
</tr>
</tbody>
</table>
Example 3 (Cont’d)

<table>
<thead>
<tr>
<th>First Response at $t$-th visit</th>
<th>No Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t = 1$</td>
<td>$t = 2$</td>
</tr>
<tr>
<td>Response Rate (%)</td>
<td>42.94</td>
</tr>
<tr>
<td>Ave. Unemp. Rate (%)</td>
<td>1.81</td>
</tr>
<tr>
<td>Response Rate (%)</td>
<td>9.85</td>
</tr>
</tbody>
</table>

Response propensity seems to be correlated with the unemployment rate.

Reference:
Measurement error: Age Heaping example

Bangladesh Age Clumping Display

% of children in 1 month age groups
Korean Longitudinal Study of Aging (KLoSA) data (http://www.kli.re.kr/klosa/en/about/introduce.jsp)

Original sample measures height and weight from survey questions (N=9,842)

A validation sample (n=505) is randomly selected from the original sample to obtain physical measurement for the height and weight.
Measurement error: BMI data example (Cont’d)
**Planned missingness: NRI example**

National Resources Inventory  
(http://www.nrcs.usda.gov/wps/portal/nrcs/main/national/technical/nra/nri/)

<table>
<thead>
<tr>
<th></th>
<th>1997</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Planned missingness: Split questionnaire design

<table>
<thead>
<tr>
<th>Pattern</th>
<th>$x$</th>
<th>$y_1$</th>
<th>$y_2$</th>
<th>$y_3$</th>
<th>Cost</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>$c_1$</td>
<td>$n_1$</td>
</tr>
<tr>
<td>2</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td>$c_2$</td>
<td>$n_2$</td>
</tr>
<tr>
<td>3</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>$c_3$</td>
<td>$n_3$</td>
</tr>
<tr>
<td>4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>$c_4$</td>
<td>$n_4$</td>
</tr>
<tr>
<td>5</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>$c_5$</td>
<td>$n_5$</td>
</tr>
<tr>
<td>6</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>$c_6$</td>
<td>$n_6$</td>
</tr>
<tr>
<td>7</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>$c_7$</td>
<td>$n_7$</td>
</tr>
</tbody>
</table>

Reference:
Will generally use this notation throughout

\[ Y = \text{outcome or dependent variable} \]
\[ X = \text{covariate or vector of covariates} \]
\[ R = \text{response indicator for } Y \]
\[ = 1 \text{ if } Y \text{ observed, } 0 \text{ if missing} \]
Simulate observations from a normal distribution

## 5 observations from \( N(0,1) \)
```r
> rnorm(n=5, mean=0, sd=1)
[1] -0.27961336 0.88267457 0.01061641 -0.08252131 0.61003977
```
```r
> z = rnorm(n=5, mean=0, sd=1)
> z
[1] 0.6741197 -0.3814703 1.4246447 0.2252487 -0.1592414
```
```r
> zbar = mean(z)
> zbar
[1] 0.3566603
```

## 30 observations from \( N(3,5^2) \)
```r
> y = rnorm(n=30, mean=3, sd=5)
```
Simulating data in R

Summarize results of 100 simulations

```r
### Simulate 5 observations 100 times
> results = matrix(0, nrow=100, ncol=2)
> colnames(results) = c("Mean", "SD")

> for (i in 1:100)
>   { z = rnorm(n=5, mean=0, sd=1)
>     results[i,1] = mean(z)
>     results[i,2] = sd(z) }

### Print results
> results[1:5,]

   Mean      SD
[1,] -0.0804799  0.804498
[2,]  0.4280679  0.401783
[3,]  0.8633050  1.729228
[4,] -0.5392521  1.138942
[5,] -0.0793508  0.615434
```

J. Kim (ISU)
Simulating data in R

```r
> results[1:5,]
   Mean       SD
[1,] -0.08047987 0.8044978
[2,]  0.42806792 0.4017826
[3,]  0.86330499 1.7292280
[4,] -0.53925212 1.1389417
[5,] -0.07935075 0.6154337

### calculate mean of individual sample means and SD’s
> apply(results, 2, mean)
   Mean       SD
0.03208639 0.95688116

### standard deviation of individual sample means and SD’s
> apply(results, 2, sd)
   Mean       SD
0.4985703  0.3696412
```
Simulating binary data in R

Use command `rbinom`

```r
### Simulate 10 binary observations having P(R=1) = .30

> R = rbinom(n=10, size=1, prob=.30)
> R
[1] 0 1 0 1 0 0 1 0 1 1
> mean(R)
[1] 0.5

> R = rbinom(n=10, size=1, prob=.30)
> R
[1] 0 1 0 1 0 0 0 0 1 0
> mean(R)
[1] 0.3
```
1. Generate the ‘full data’ – in this case a sample of continuous outcomes $Y$

2. Generate the response indicators $R$ – the *missing data mechanism*
   - Have to determine $P(R = 1)$
   - Can allow $P(R = 1)$ to depend on $Y$
Simulating incomplete data in R

**Example 1.** Random deletion, or missing (completely) at random.

\[
Y \sim N(0, 1) \\
R \sim \text{Ber}(0.5)
\]

**Example 2.** Deletion depends on \( Y \) such that lower values of \( Y \) are more likely to be observed. This is missing *not* at random.

\[
Y \sim N(0, 1) \\
R \sim \text{Ber}\{q(Y)\}
\]

where the function \( q(Y) \) is given by

\[
q(Y) = \frac{1}{1 + \exp(Y)}
\]
Simulating incomplete data in R

Probability of response as a function of $Y$

![Graph showing the probability of response as a function of $Y$.]
A more general form of missing data mechanism

Can introduce a parameter that governs degree of dependence on $Y$

$$q(\alpha Y) = \frac{1}{1 + \exp(\alpha Y)}$$

- When $\alpha = 0$, response probability does not depend on $Y$.
- For $\alpha \neq 0$, response probability depends on $Y$.
- Magnitude of $\alpha$ governs degree of dependence.
Different missing data mechanisms

The full-data model here is

\[ Y \sim N(0, 1) \]
\[ R \sim \text{Ber}\{q(\alpha Y)\} \]

where

\[ q(\alpha Y) = \frac{1}{1 + \exp(\alpha Y)} \]
Different missing data mechanisms

These plots represent $\alpha = -3$, $\alpha = 0$, $\alpha = 1$
## Example 2: nonrandom deletion

\[
Y = \text{rnorm}(n = 100, \text{mean}=0, \text{sd}=1)
\]

\[
q.Y = \frac{1}{1 + \exp(Y)}
\]

\[
R = \text{rbinom}(n = 100, \text{size}=1, \text{prob}=q.Y)
\]

\[
\text{Fulldata} = \text{cbind}(Y,R)
\]

\[
Y.\text{obs} = \text{Fulldata}[R==1,1]
\]

```r
Y.\text{obs}
```

```r
mean(Y)
```

```r
mean(Y.\text{obs})
```

```r
mean(R)
```
R Code for simulation

## Simulate the process in example #2 1000 times
results = matrix(0, nrow=1000, ncol=3)
summary = matrix(0, nrow=1, ncol=3)
labels = c("mean of Y", "mean of Y.obs", "mean of R")

colnames(results) = labels
colnames(summary) = labels

# alpha controls whether R depends on Y
alpha = 1
for (i in 1:1000)
{
Y = rnorm(n = 100, mean=0, sd=1)
q.Y = 1 / ( 1 + exp( alpha*Y ) )
R = rbinom(n = 100, size=1, prob=q.Y)
Fulldata = cbind(Y,R)
Y.obs = Fulldata[R==1,1]
results[i,] = c( mean(Y), mean(Y.obs), mean(R) )
}

summary = apply(results, 2, mean)
summary
ALPHA = -3
> summary
  mean of Y mean of Y.obs mean of R
0.0005652042  0.6911873446  0.4994900000

ALPHA = 0
> summary
  mean of Y mean of Y.obs mean of R
-0.001543965  0.001200788  0.501350000

ALPHA = 1
> summary
  mean of Y mean of Y.obs mean of R
-0.0004493881  -0.4136889588  0.4999100000