• Mostly we have studied the behavior of a single random variable.
• Often, however, we gather data on two random variables.

**Response Variable (Y):** Measures the major outcome of interest in the study (also called the *dependent* variable).

**Independent Variable (X):** Another variable whose value explains, predicts, or is associated with the value of the response variable (also called the *predictor* or the *regressor*).

• We wish to determine: Is there a relationship between the two r.v.’s?
• Can we use the values of one r.v. to predict the other r.v.?

**Observational Studies vs. Designed Experiments**

• In observational studies, we simply measure or observe both variables on a set of sampled individuals.

• In a designed experiment, we manipulate the predictors (*factors*), setting them at specific values of interest. We then observe what values of the response correspond to the fixed predictor values.
Example 1 (Table 6.1): We observe the Rockwell Hardness \((X)\) and Young’s modulus \((Y)\) for seven high-density metals. The resulting data were:

\[
\begin{array}{ccccccc}
X & 41 & 41 & 44 & 40 & 43 & 15 & 40 \\
Y & 310 & 340 & 380 & 317 & 413 & 62 & 119
\end{array}
\]

Example 2 (Table 6.3): A chemical engineering class studied the effect of the reflux ratio \((X)\) on the ethanol concentration \((Y)\) of an ethanol-water distillation. For a variety of settings of the reflux ratio, the ethanol concentration was measured:

\[
\begin{array}{cccccc}
X & 20 & 30 & 40 & 50 & 60 \\
Y & 0.446 & 0.601 & 0.786 & 0.928 & 0.950
\end{array}
\]

We assume there is random error in the observed response values, implying a probabilistic relationship between the 2 variables.

• Often we assume a straight-line relationship between two variables.
• This is known as simple linear regression.

\[
Y_i = \beta_0 + \beta_1 x_i + \epsilon_i
\]

\[
Y_i = \text{ith response value} \quad \beta_0 = \text{Intercept of regression line} \\
x_i = \text{ith predictor value} \quad \beta_1 = \text{slope of regression line} \\
\epsilon_i = \text{ith random error component}
\]
• We assume the random errors $\varepsilon_i$ have mean 0 (and variance $\sigma^2$), so that $E(Y) = \beta_0 + \beta_1 x$.

• Typically, in practice, $\beta_0$ and $\beta_1$ are unknown parameters. We estimate them using the sample data.

Fitting the Model (Least Squares Method)

• If we gather data $(X_i, Y_i)$ for several individuals, we can use these data to estimate $\beta_0$ and $\beta_1$ and thus estimate the linear relationship between $Y$ and $X$.

• First step: Decide if a straight-line relationship between $Y$ and $X$ makes sense.

Plot the bivariate data using a scatter plot.

R code:
```r
> x <- c(20, 30, 40, 50, 60)
> y <- c(.446, .601, .786, .928, .950)
> plot(x,y,pch=19)
```

• Once we settle on the “best-fitting” regression line, its equation gives a predicted $Y$-value for any new $X$-value.
• How do we decide, given a data set, which line is the best-fitting line?

Note that usually, no line will go through all the points in the data set.

For each point, the residual =
(Some positive residuals, some negative residuals)

We want the line that makes these errors as small as possible (so that the line is “close” to the points).

**Least-squares method:** We choose the line that minimizes the sum of all the squared residuals ($SS_{\text{res}}$).

$SS_{\text{res}} =$

**Least squares prediction equation:**

$\hat{Y} = b_0 + b_1 X$

where $b_0$ and $b_1$ are the estimates of $\beta_0$ and $\beta_1$ that produce the best-fitting line in the least squares sense.
Formulas for $b_0$ and $b_1$:

Estimated slope and intercept:

$$b_1 = \frac{SS_{xy}}{SS_{xx}} \quad \text{and} \quad b_0 = \bar{Y} - b_1 \bar{X}$$

where $SS_{xy} = \sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}$ and

$$SS_{xx} = \sum X_i^2 - \frac{(\sum X_i)^2}{n}$$

and $n = \text{the number of observations}$.

Example (see Table 6.4):

$$\sum Y_i = \sum X_i^2 =$$

$$\sum X_i = \sum X_iY_i =$$

$$SS_{xy} =$$

$$SS_{xx} =$$

R code:

```r
> x <- c(20, 30, 40, 50, 60)
> y <- c(0.446, 0.601, 0.786, 0.928, 0.950)
> lm(y ~ x)
```
Derivation of Formulas for $b_0$ and $b_1$:

Recall that $SS_{res} =$

To minimize the $SS_{res}$ with respect to $b_0$ and $b_1$:
Interpretations:

Slope:

Intercept:

Example:

Avoid extrapolation: predicting/interpreting the regression line for X-values outside the range of X in the data set.
Model Assumptions

• Recall model equation: \( Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \)

• To perform inference about our regression line, we need to make certain assumptions about the random error component, \( \varepsilon_i \). We assume:

(1) The mean of \( \varepsilon_i \) is 0. (In the long run, the values of the random errors average zero.)
(2) The variance of the probability distribution of \( \varepsilon_i \) is constant for all values of \( X \). We denote the variance of \( \varepsilon_i \) by \( \sigma^2 \).
(3) The probability distribution of \( \varepsilon_i \) is normal.
(4) The values of \( \varepsilon_i \) for any two observed \( Y \)-values are independent – the value of \( \varepsilon_i \) has no effect on the value of \( \varepsilon_j \) for the \( i \)th and \( j \)th \( Y \)-values.

Picture:

We will discuss later how to check these assumptions for a particular data set.
Estimating $\sigma^2$

Typically the error variance $\sigma^2$ is unknown.

An unbiased estimate of $\sigma^2$ is the mean squared residual ($MS_{\text{res}}$).

$$MS_{\text{res}} = \frac{SS_{\text{res}}}{(n-2)}$$

where $SS_{\text{res}} = SS_{yy} - b_1 SS_{xy}$

and $SS_{yy} = \sum Y_i^2 - \left(\frac{\sum Y_i}{n}\right)^2$

Note that an estimate of $\sigma$ is

$$\sqrt{MS_{\text{res}}} = \sqrt{\frac{SS_{\text{res}}}{n-2}}$$

Testing the Usefulness of the Model

For the SLR model, $E(Y) = \beta_0 + \beta_1 x$.

Note: $X$ is completely useless in helping to predict or explain $Y$ if and only if $\beta_1 = 0$.

So to test the usefulness of the model for predicting $Y$, we test:
If we reject $H_0$ and conclude $H_a$ is true, then we conclude that $X$ does provide information for the prediction of $Y$.

Recall that the estimate $b_1$ is a statistic that depends on the sample data.
This $b_1$ has a sampling distribution.

If our four SLR assumptions hold, the sampling distribution of $b_1$ is normal with mean $\beta_1$ and standard deviation which we estimate by

$$b_1 \sqrt{MS_{res} / SS_{xx}}$$

Under $H_0$: $\beta_1 = 0$, the statistic $b_1 \sqrt{MS_{res} / SS_{xx}}$ has a t-distribution with $n - 2$ d.f.
**Test about the Slope**

<table>
<thead>
<tr>
<th>One-Tailed Tests</th>
<th>Two-Tailed Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$: $\beta_1 = 0$</td>
<td>$H_0$: $\beta_1 = 0$</td>
</tr>
<tr>
<td>$H_a$: $\beta_1 &lt; 0$</td>
<td>$H_a$: $\beta_1 &gt; 0$</td>
</tr>
<tr>
<td>$H_a$: $\beta_1 \neq 0$</td>
<td>$H_a$: $\beta_1 \neq 0$</td>
</tr>
</tbody>
</table>

**Test statistic:** \[ t = \frac{b_1}{\sqrt{MS_{res} / SS_{xx}}} \]

**Rejection region:**
- $t < -t_{\alpha, n-2}$
- $t > t_{\alpha, n-2}$
- $t > t_{\alpha/2}$ or $t < -t_{\alpha/2}$

**P-value:**
- left tail area
- right tail area
- 2*(tail area outside $t$)
- outside $t$

**Example:** In the ethanol example, recall $b_1 = $
Is the real $\beta_1$ significantly greater than 0?
(Use $\alpha = .05$.)

Is the real $\beta_1$ significantly greater than 0? (Use $\alpha = .05$.)
A 100(1 – \(\alpha\))% Confidence Interval for the true slope \(\beta_1\) is given by:

where \(t_{\alpha/2}\) is based on \(n - 2\) d.f.

In our example, a 95% CI for \(\beta_1\) is:

**R code:**

```r
def x <- c(20, 30, 40, 50, 60)
def y <- c(.446, .601, .786, .928, .950)
def summary(lm(y ~ x))
def plot(x, y, pch=19); abline(lm(y ~ x))
```
Correlation

The scatterplot gives us a general idea about whether there is a linear relationship between two variables.

More precise: The coefficient of correlation (denoted \( r \)) is a numerical measure of the strength and direction of the linear relationship between two variables.

Formula for \( r \) (the correlation coefficient between two variables \( X \) and \( Y \)):

\[
    r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}
\]

Most computer packages will also calculate the correlation coefficient.

Interpreting the correlation coefficient:

- **Positive** \( r \) => The two variables are **positively associated** (large values of one variable correspond to large values of the other variable)
- **Negative** \( r \) => The two variables are **negatively associated** (large values of one variable correspond to small values of the other variable)
- \( r = 0 \) => **No linear association** between the two variables.

Note: \(-1 \leq r \leq 1\) always.
How far $r$ is from 0 measures the strength of the linear relationship:

- $r$ nearly 1 => Strong positive relationship between the two variables
- $r$ nearly -1 => Strong negative relationship between the two variables
- $r$ near 0 => Weak relationship between the two variables

Pictures:

Example (Rockwell hardness / Young’s modulus data):

```r
> rock <- c(41,41,44,40,43,15,40)
> young <- c(310,340,380,317,413,62,119)
> cor(rock, young)
[1] 0.7759845
```

Interpretation?
Notes: (1) Correlation makes no distinction between predictor and response variables.
(2) Variables must be numerical to calculate $r$.
(3) Correlation only measures the linear association between two variables, not any nonlinear relationship.

The square of the correlation coefficient is called the coefficient of determination, $R^2$.

**Interpretation:** $R^2$ represents the proportion of sample variability in $Y$ that is explained by its linear relationship with $X$.

$$R^2 = 1 - \frac{SS_{res}}{SS_{yy}} \quad (R^2 \text{ always between 0 and 1})$$

For the Rockwell hardness / Young’s modulus data example, $R^2 =$

**Interpretation:**

For the reflux ratio / ethanol concentration data example, $R^2 =$

**Interpretation:**
Estimation and Prediction with the Regression Model

Major goals in using the regression model:
(1) Determining the linear relationship between $Y$ and $X$ (accomplished through inferences about $\beta_1$)

(2) Estimating the mean value of $Y$, denoted $E(Y)$, for a particular value of $X$.
Example: Among all columns with reflux ratio 35 units, what is the estimated mean ethanol concentration?

(3) Predicting the value of $Y$ for a particular value of $X$.
Example: For a “new” column having reflux ratio 35 units, what is the predicted ethanol concentration?

• The point estimate for these last two quantities is the same; it is:

Example:

• However, the variability associated with these point estimates is very different.

• Which quantity has more variability, a single $Y$-value or the mean of many $Y$-values?
This is seen in the following formulas:

\[
100(1 - \alpha)\% \ \text{Confidence Interval} \ \text{for the mean value of} \ Y \ \text{at} \ X = x_0:
\]

where \( t_{\alpha/2} \) based on \( n - 2 \) d.f.

\[
100(1 - \alpha)\% \ \text{Prediction Interval} \ \text{for the an individual new value of} \ Y \ \text{at} \ X = x_0:
\]

where \( t_{\alpha/2} \) based on \( n - 2 \) d.f.

The extra “1” inside the square root shows the prediction interval is wider than the CI, although they have the same center.

Note: A “Prediction Interval” attempts to contain a random quantity, while a confidence interval attempts to contain a (fixed) parameter value.
The variability in our estimate of $E(Y)$ reflects the fact that we are merely estimating the unknown $\beta_0$ and $\beta_1$.

The variability in our prediction of the new $Y$ includes that variability, plus the natural variation in the $Y$-values.

Example (ethanol concentration data):  
95\% CI for $E(Y)$ with $X = 35$:  

```r  
> x <- c(20, 30, 40, 50, 60)  
> y <- c(.446, .601, .786, .928, .950)  
> predict(lm(y ~ x), data.frame(x = c(35)), interval="confidence", level=0.95)  
```

95\% PI for a new $Y$ having $X = 35$:  

```r  
> predict(lm(y ~ x), data.frame(x = c(35)), interval="prediction", level=0.95)  
```