• Mostly we have studied the behavior of a single random variable.
• Often, however, we gather data on two random variables.
• 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.?
• Often we assume a straight-line relationship between two variables.
• This is known as simple linear regression.

Probabilistic vs. Deterministic Models

If there is an exact relationship between two (or more) variables that can be predicted with certainty, without any random error, this is known as a deterministic relationship.

Examples:
In statistics, we usually deal with situations having random error, so exact predictions are not possible.

This implies a probabilistic relationship between the 2 variables.

Example: \( Y = \) breathalyzer reading
\( X = \) amount of alcohol consumed (fl. oz.)
• We typically assume the random errors balance out – they average zero.
• Then this is equivalent to assuming the mean of $Y$, denoted $E(Y)$, equals the deterministic component.

**Straight-Line Regression Model**

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

$Y = \text{response variable (dependent variable)}$

$X = \text{predictor variable (independent variable)}$

$\varepsilon = \text{random error component}$

$\beta_0 = \text{Y-intercept of regression line}$

$\beta_1 = \text{slope of regression line}$

Note that the deterministic component of this model is $E(Y) = \beta_0 + \beta_1 X$

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

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

**Predictor Variable (X):** Another variable whose value explains, predicts, or is associated with the value of the response variable.
Fitting the Model (Least Squares Method)

If we gather data \((X, Y)\) 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 scattergram (scatterplot).

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 error $= (\text{Some positive errors, some negative errors})$

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 errors (SSE).

Least squares regression line:

\[ \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \]

where \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) are the estimates of \( \beta_0 \) and \( \beta_1 \) that produce the best-fitting line in the least squares sense.
Formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$:

Estimated slope and intercept:

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \quad \text{and} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

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

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

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

Example (Table 11.1):

$Y = $

$X = $

$SS_{xy} = $

$SS_{xx} =$
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: \[ Y = \beta_0 + \beta_1 X + \varepsilon \]
To perform inference about our regression line, we need to make certain assumptions about the random error component, \( \varepsilon \). We assume:

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

Picture:
Estimating $\sigma^2$

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

An unbiased estimate of $\sigma^2$ is the mean squared error (MSE), also denoted $s^2$ sometimes.

$$\text{MSE} = \frac{\text{SSE}}{n-2}$$

where $\text{SSE} = \text{SS}_{yy} - \hat{\beta}_1 \text{SS}_{xy}$

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

Note that an estimate of $\sigma$ is

$$s = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n-2}}$$

Since $\varepsilon$ has a normal distribution, we can say, for example, that about 95% of the observed $Y$-values fall within $2s$ units of the corresponding values $\hat{Y}$. 
Testing the Usefulness of the Model

For the SLR model, \( Y = \beta_0 + \beta_1 X + \varepsilon \).

Note: \( X \) is completely useless in helping to predict \( Y \) 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 \( Y \).

Picture:
Recall that the estimate $\hat{\beta}_1$ is a statistic that depends on
the sample data.

This $\hat{\beta}_1$ has a sampling distribution.

If our four SLR assumptions hold, the sampling
distribution of $\hat{\beta}_1$ is normal with mean $\beta_1$ and standard
deviation which we estimate by

$$\frac{\hat{\beta}_1}{s / \sqrt{SS_{xx}}}$$

Under $H_0: \beta_1 = 0$, the statistic $\frac{\hat{\beta}_1}{s / \sqrt{SS_{xx}}}$ has a t-distribution with $n – 2$ d.f.

### Test for Model Usefulness

<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: $t = \frac{\hat{\beta}_1}{s / \sqrt{SS_{xx}}}$

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

P-value:
- Left tail area
- Right tail area
- $2*(\text{tail area outside } t)$
- Outside $t$
Example: In the drug reaction example, recall $\hat{\beta}_1 = 0.7$. Is the real $\beta_1$ significantly different from 0? (Use $\alpha = .05$.)
A 100(1 – α)% Confidence Interval for the true slope $\beta_1$ is given by:

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

In our example, a 95% CI for $\beta_1$ is:
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 $\Rightarrow$ Strong positive relationship between the two variables
- $r$ nearly -1 $\Rightarrow$ Strong negative relationship between the two variables
- $r$ near 0 $\Rightarrow$ Weak linear relationship between the two variables

Pictures:

Example (Drug/reaction time data):

Interpretation?
Notes: (1) Correlation makes no distinction between predictor and response variables.
(2) Variables must be numerical to calculate $r$.

Examples: What would we expect the correlation to be if our two variables were:
(1) Work Experience & Salary?

(2) Weight of a Car & Gas Mileage?

Some Cautions

Example:

<table>
<thead>
<tr>
<th>Speed of a car (X)</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mileage in mpg (Y)</td>
<td>24</td>
<td>28</td>
<td>30</td>
<td>28</td>
<td>24</td>
</tr>
</tbody>
</table>

Scatterplot of these data:

Calculation will show that $r = 0$ for these data.

Are the two variables related?
**Another caution:** Correlation between two variables does not automatically imply that there is a cause-effect relationship between them.

**Note:** The population correlation coefficient between two variables is denoted $\rho$. To test $H_0: \rho = 0$, we simply use the equivalent test of $H_0: \beta_1 = 0$ in the SLR model. If this null hypothesis is rejected, we conclude there is a significant correlation between the two variables.

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 $X$.

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

For the drug/reaction time 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 people with drug amount 3.5 mg, what is the estimated mean reaction time?

(3) Predicting the value of \( Y \) for a particular value of \( X \).
   Example: For a “new” individual having drug amount 3.5 mg, what is the predicted reaction time?

• 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:

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

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

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

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 (drug/reaction time data):
95% CI for $E(Y)$ with $X = 3.5$:

95% PI for a new $Y$ having $X = 3.5$: 