Chapter 13
Introduction to Linear Regression and Correlation Analysis

Chapter Goals
To understand the methods for displaying and describing relationship among variables
Methods for Studying Relationships

- **Graphical**
  - Scatterplots
  - Line plots
  - 3-D plots

- **Models**
  - Linear regression
  - Correlations
  - Frequency tables

Two Quantitative Variables

The *response variable*, also called the *dependent variable*, is the variable we want to predict, and is usually denoted by $y$.

The *explanatory variable*, also called the *independent variable*, is the variable that attempts to explain the response, and is denoted by $x$. 
YDI 7.1

<table>
<thead>
<tr>
<th>Response (y)</th>
<th>Explanatory (x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height of son</td>
<td></td>
</tr>
<tr>
<td>Weight</td>
<td></td>
</tr>
</tbody>
</table>

Scatter Plots and Correlation

- A scatter plot (or scatter diagram) is used to show the relationship between two variables.
- Correlation analysis is used to measure strength of the association (linear relationship) between two variables:
  - Only concerned with strength of the relationship
  - No causal effect is implied
Example

The following graph shows the scatterplot of Exam 1 score (x) and Exam 2 score (y) for 354 students in a class. Is there a relationship?

Scatter Plot Examples

Linear relationships

Curvilinear relationships
Scatter Plot Examples

(continued)

Correlation Coefficient

(continued)

- The population correlation coefficient $\rho$ (rho) measures the strength of the association between the variables.

- The sample correlation coefficient $r$ is an estimate of $\rho$ and is used to measure the strength of the linear relationship in the sample observations.
Features of $\rho$ and $r$

- Unit free
- Range between -1 and 1
- The closer to -1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker the linear relationship

Examples of Approximate $r$ Values

Tag with appropriate value:

-1, -.6, 0, +.3, 1
Earlier Example

![Scatter plot of Exam 1 vs Exam 2 with correlation coefficient](image)

<table>
<thead>
<tr>
<th></th>
<th>Exam1</th>
<th>Exam2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Correlation</td>
<td>1.400**</td>
<td>366</td>
</tr>
<tr>
<td>N</td>
<td>351</td>
<td>356</td>
</tr>
</tbody>
</table>

Correlation is significant at the 0.01 level.

YDI 7.3

What kind of relationship would you expect in the following situations:

- age (in years) of a car, and its price.
- number of calories consumed per day and weight.
- height and IQ of a person.
YDI 7.4

Identify the two variables that vary and decide which should be the independent variable and which should be the dependent variable. Sketch a graph that you think best represents the relationship between the two variables.

1. The size of a person’s vocabulary over his or her lifetime.
2. The distance from the ceiling to the tip of the minute hand of a clock hung on the wall.

Introduction to Regression Analysis

- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

**Dependent variable:** the variable we wish to explain

**Independent variable:** the variable used to explain the dependent variable
Simple Linear Regression Model

- Only **one** independent variable, $x$
- Relationship between $x$ and $y$ is described by a linear function
- Changes in $y$ are assumed to be caused by changes in $x$

Types of Regression Models

- **Positive Linear Relationship**
- **Negative Linear Relationship**
- **Relationship NOT Linear**
- **No Relationship**
Population Linear Regression

The population regression model:

\[ y = \beta_0 + \beta_1 x + \epsilon \]

Linear regression assumptions:

- Error values (\( \epsilon \)) are statistically independent
- Error values are normally distributed for any given value of \( x \)
- The probability distribution of the errors is normal
- The probability distribution of the errors has constant variance
- The underlying relationship between the \( x \) variable and the \( y \) variable is linear
Population Linear Regression

\[ y = \beta_0 + \beta_1 x + \varepsilon \]

Observed Value of \( y \) for \( x_i \)

Predicted Value of \( y \) for \( x_i \)

Intercept = \( \beta_0 \)

Random Error for this \( x \) value

Slope = \( \beta_1 \)

Estimated Regression Model

The sample regression line provides an estimate of the population regression line

\[ \hat{y}_i = b_0 + b_1 x \]

The individual random error terms \( e_i \) have a mean of zero
Earlier Example

Residual

A **residual** is the difference between the observed response $y$ and the predicted response $\hat{y}$. Thus, for each pair of observations $(x_i, y_i)$, the $i^{th}$ residual is

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1x_i)$$
Least Squares Criterion

- $b_0$ and $b_1$ are obtained by finding the values of $b_0$ and $b_1$ that minimize the sum of the squared residuals.

\[
\sum e^2 = \sum (y - \hat{y})^2 = \sum (y - (b_0 + b_1x))^2
\]

Interpretation of the Slope and the Intercept

- $b_0$ is the estimated average value of $y$ when the value of $x$ is zero.
- $b_1$ is the estimated change in the average value of $y$ as a result of a one-unit change in $x$. 
The Least Squares Equation

- The formulas for $b_1$ and $b_0$ are:

\[
b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}
\]

algebraic equivalent:

\[
b_1 = \frac{\sum xy - \sum x \sum y}{\sum x^2 - (\sum x)^2}
\]

and

\[
b_0 = \bar{y} - b_1 \bar{x}
\]

Finding the Least Squares Equation

- The coefficients $b_0$ and $b_1$ will usually be found using computer software, such as Excel, Minitab, or SPSS.

- Other regression measures will also be computed as part of computer-based regression analysis.
Simple Linear Regression Example

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)

- A random sample of 10 houses is selected
  - Dependent variable (y) = house price in $1000s
  - Independent variable (x) = square feet

Sample Data for House Price Model

<table>
<thead>
<tr>
<th>House Price in $1000s (y)</th>
<th>Square Feet (x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>245</td>
<td>1400</td>
</tr>
<tr>
<td>312</td>
<td>1600</td>
</tr>
<tr>
<td>279</td>
<td>1700</td>
</tr>
<tr>
<td>308</td>
<td>1875</td>
</tr>
<tr>
<td>199</td>
<td>1100</td>
</tr>
<tr>
<td>219</td>
<td>1550</td>
</tr>
<tr>
<td>405</td>
<td>2350</td>
</tr>
<tr>
<td>324</td>
<td>2450</td>
</tr>
<tr>
<td>319</td>
<td>1425</td>
</tr>
<tr>
<td>255</td>
<td>1700</td>
</tr>
</tbody>
</table>
The regression equation is:

\[
\text{house price} = 98.248 + 0.110 \times \text{(square feet)}
\]

Graphical Presentation

- House price model: scatter plot and regression line

\[
\text{Slope} = 0.110
\]

\[
\text{Intercept} = 98.248
\]

\[
\text{house price} = 98.248 + 0.110 \times \text{(square feet)}
\]
Interpretation of the Intercept, $b_0$

- $b_0$ is the estimated average value of $Y$ when the value of $X$ is zero (if $x = 0$ is in the range of observed $x$ values)
  - Here, no houses had 0 square feet, so $b_0 = 98.24833$ just indicates that, for houses within the range of sizes observed, $98,248.33$ is the portion of the house price not explained by square feet.

Interpretation of the Slope Coefficient, $b_1$

- $b_1$ measures the estimated change in the average value of $Y$ as a result of a one-unit change in $X$
  - Here, $b_1 = .10977$ tells us that the average value of a house increases by $.10977(1000) = 109.77$, on average, for each additional one square foot of size.
Least Squares Regression Properties

- The sum of the residuals from the least squares regression line is 0 \( \left( \sum (y - \hat{y}) = 0 \right) \)
- The sum of the squared residuals is a minimum \( \left( \text{minimized } \sum (y - \hat{y})^2 \right) \)
- The simple regression line always passes through the mean of the y variable and the mean of the x variable
- The least squares coefficients are unbiased estimates of \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \)

YDI 7.6

The growth of children from early childhood through adolescence generally follows a linear pattern. Data on the heights of female Americans during childhood, from four to nine years old, were compiled and the least squares regression line was obtained as \( \hat{y} = 32 + 2.4x \) where \( \hat{y} \) is the predicted height in inches, and \( x \) is age in years.

- Interpret the value of the estimated slope \( b_1 = 2.4 \).
- Would interpretation of the value of the estimated y-intercept, \( b_0 = 32 \), make sense here?
- What would you predict the height to be for a female American at 8 years old?
- What would you predict the height to be for a female American at 25 years old? How does the quality of this answer compare to the previous question?
The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable.

The coefficient of determination is also called R-squared and is denoted as $R^2$.

$$0 \leq R^2 \leq 1$$

Note: In the single independent variable case, the coefficient of determination is

$$R^2 = r^2$$

where:

$R^2 = \text{Coefficient of determination}$

$r = \text{Simple correlation coefficient}$
Examples of Approximate $R^2$ Values

- $R^2 = 0$
  - No linear relationship between $x$ and $y$: The value of $Y$ does not depend on $x$. (None of the variation in $y$ is explained by variation in $x$)
### SPSS Output

#### Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>.762a</td>
<td>.581</td>
<td>.528</td>
<td>41332232</td>
</tr>
</tbody>
</table>

* Predictors: (Constant), Square Feet

#### ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>18934.935</td>
<td>1</td>
<td>18934.935</td>
<td>11.085</td>
<td>.010a</td>
</tr>
<tr>
<td>Residual</td>
<td>13565.565</td>
<td>8</td>
<td>1708.196</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>32400.500</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Predictors: (Constant), Square Feet
** Dependent Variable: House Price

#### Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td>t</td>
<td>Sig</td>
</tr>
<tr>
<td>(Constant)</td>
<td>98.248</td>
<td>58.033</td>
<td>1.693</td>
<td>.129</td>
</tr>
<tr>
<td>Square Feet</td>
<td>.110</td>
<td>.033</td>
<td>.762</td>
<td>.329</td>
</tr>
</tbody>
</table>

* Dependent Variable: House Price

---

### Standard Error of Estimate

- The standard deviation of the variation of observations around the regression line is called the **standard error of estimate** \( s_\varepsilon \).

- The standard error of the regression slope coefficient (\( b_1 \)) is given by \( s_{b_1} \).
### SPSS Output

**Model Summary**

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.762</td>
<td>.581</td>
<td>.528</td>
<td>41.33032</td>
</tr>
</tbody>
</table>

- Predictors: (Constant), Square Feet

**Coefficients**

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Beta</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Constant)</td>
<td>98.248</td>
<td>58.033</td>
<td>1.693</td>
<td>.129</td>
</tr>
<tr>
<td>Square Feet</td>
<td>.110</td>
<td>.033</td>
<td>3.329</td>
<td>.010</td>
</tr>
</tbody>
</table>

- Dependent Variable: House Price

### Comparing Standard Errors

- **Variation of observed y values from the regression line**
  - Small $s_\varepsilon$
  - Large $s_\varepsilon$

- **Variation in the slope of regression lines from different possible samples**
  - Small $s_{b_1}$
  - Large $s_{b_1}$
Inference about the Slope: t Test

- t test for a population slope
  - Is there a linear relationship between x and y?
- Null and alternative hypotheses
  - $H_0: \beta_1 = 0$ (no linear relationship)
  - $H_1: \beta_1 \neq 0$ (linear relationship does exist)
- Test statistic
  - $t = \frac{b_1 - \beta_1}{s_{b_1}}$
  - d.f. = $n - 2$

where:
- $b_1$ = Sample regression slope coefficient
- $\beta_1$ = Hypothesized slope
- $s_{b_1}$ = Estimator of the standard error of the slope

Estimated Regression Equation:

\[
\text{house price} = 98.25 + 0.1098 \text{ (sq.ft.)}
\]

The slope of this model is 0.1098

Does square footage of the house affect its sales price?
Inferences about the Slope: t Test Example

Test Statistic: \( t = 3.329 \)

From Excel output:

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>98.24833</td>
<td>1.69296</td>
<td>0.12892</td>
</tr>
<tr>
<td>Square Feet</td>
<td>0.10977</td>
<td>3.32938</td>
<td>0.01039</td>
</tr>
</tbody>
</table>

Decision: Reject \( H_0 \)

Conclusion: There is sufficient evidence that square footage affects house price.

Regression Analysis for Description

Confidence Interval Estimate of the Slope:

\[
b_1 \pm t(1-\alpha/2)s_{b_1}
\]

Excel Printout for House Prices:

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>98.24833</td>
<td>1.69296</td>
<td>0.12892</td>
<td>-35.57729</td>
<td>232.07386</td>
</tr>
<tr>
<td>Square Feet</td>
<td>0.10977</td>
<td>3.32938</td>
<td>0.01039</td>
<td>0.03374</td>
<td>0.18580</td>
</tr>
</tbody>
</table>

At 95% level of confidence, the confidence interval for the slope is (0.0337, 0.1858)
Regression Analysis for Description

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>98.24833</td>
<td>58.03348</td>
<td>1.69296</td>
<td>-35.57720</td>
<td>232.07386</td>
</tr>
<tr>
<td>Square Feet</td>
<td>0.10977</td>
<td>0.03297</td>
<td>3.32938</td>
<td>0.01039</td>
<td>0.03374</td>
</tr>
</tbody>
</table>

Since the units of the house price variable is $1000s, we are 95% confident that the average impact on sales price is between $33.70 and $185.80 per square foot of house size.

This 95% confidence interval does not include 0.

Conclusion: There is a significant relationship between house price and square feet at the .05 level of significance.

Residual Analysis

- **Purposes**
  - Examine for linearity assumption
  - Examine for constant variance for all levels of x
  - Evaluate normal distribution assumption

- **Graphical Analysis of Residuals**
  - Can plot residuals vs. x
  - Can create histogram of residuals to check for normality
Residual Analysis for Linearity

- **Not Linear**
- **Linear**

Residual Analysis for Constant Variance

- **Non-constant variance**
- **Constant variance**
### Residual Output

<table>
<thead>
<tr>
<th>Predicted House Price</th>
<th>Residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 251.92316</td>
<td>-6.923162</td>
</tr>
<tr>
<td>2 273.87671</td>
<td>38.12329</td>
</tr>
<tr>
<td>3 284.85348</td>
<td>-5.853484</td>
</tr>
<tr>
<td>4 304.06284</td>
<td>3.937162</td>
</tr>
<tr>
<td>5 218.99284</td>
<td>-19.99284</td>
</tr>
<tr>
<td>6 268.38832</td>
<td>-49.38832</td>
</tr>
<tr>
<td>7 356.20251</td>
<td>48.79749</td>
</tr>
<tr>
<td>8 367.17929</td>
<td>-43.17929</td>
</tr>
<tr>
<td>9 254.6674</td>
<td>64.33264</td>
</tr>
<tr>
<td>10 284.85348</td>
<td>-29.85348</td>
</tr>
</tbody>
</table>

#### House Price Model Residual Plot

![Residual Plot](image_url)

- **Residuals**
- **Square Feet**

- **Predicted House Price**
- **Residuals**