Simple Linear Regression

Scatter Plots

- **Scatter plots** are used for numerical data consisting of paired observations taken from two numerical variables.

- One variable is measured on the vertical axis and the other variable is measured on the horizontal axis.
Scatter Plot Example

<table>
<thead>
<tr>
<th>Volume per day</th>
<th>Cost per day</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>125</td>
</tr>
<tr>
<td>26</td>
<td>140</td>
</tr>
<tr>
<td>29</td>
<td>146</td>
</tr>
<tr>
<td>33</td>
<td>160</td>
</tr>
<tr>
<td>38</td>
<td>167</td>
</tr>
<tr>
<td>42</td>
<td>170</td>
</tr>
<tr>
<td>50</td>
<td>188</td>
</tr>
<tr>
<td>55</td>
<td>195</td>
</tr>
<tr>
<td>60</td>
<td>200</td>
</tr>
</tbody>
</table>

Cost per Day vs. Production Volume

Class Exercise

2.36 The following is a set of data from a sample of \( n = 11 \) items:

\[
\begin{array}{cccccccccccc}
X & 7 & 5 & 8 & 3 & 6 & 10 & 12 & 4 & 9 & 15 & 18 \\
Y & 21 & 15 & 24 & 9 & 18 & 30 & 36 & 12 & 27 & 45 & 54 \\
\end{array}
\]

a. Construct a scatter plot.
b. Is there a relationship between \( X \) and \( Y \)? Explain.
Correlation vs. Regression

- A scatter plot (or scatter diagram) can be used to show the relationship between two numerical variables.

- Correlation analysis is used to measure strength of the association (linear relationship) between two variables.
  - Correlation is only concerned with strength of the relationship.
  - No causal effect is implied with correlation.
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 you 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 related to changes in X
Types of Relationships

Linear relationships

Curvilinear relationships

Strong relationships

Weak relationships
Types of Relationships

- No relationship

The Linear Regression Model

\[ Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \]

The population regression model:
The Linear Regression Model

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

Observed Value of Y for \( X_i \)

Predicted Value of Y for \( X_i \)

Intercept = \( \beta_0 \)

Random Error for this \( X_i \) value

Slope = \( \beta_1 \)

Linear Regression Equation

The simple linear regression equation provides an estimate of the population regression line

\[ \hat{Y}_i = a_0 + b_1 X_i \]

Estimated (or predicted) Y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

Value of X for observation i
The Least Squares Method

- $b_0$ and $b_1$ are obtained by finding the values of $b_0$ and $b_1$ that minimize the sum of the squared differences between $Y$ and $\hat{Y}$:

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$

Finding the Least Squares Equation

- The coefficients $a_0$ and $b_1$, and other regression results in this chapter, will be found using the following formulas:

$$\hat{Y}_i = a_0 + b_1 X_i$$

$$a = \frac{\Sigma y - b \cdot \Sigma x}{n}$$

$$b = \frac{n \Sigma (xy) - (\Sigma x)(\Sigma y)}{n \Sigma x^2 - (\Sigma x)^2}$$
Interpretation of the Intercept and the Slope

- $a_0$ is the estimated mean value of $Y$ when the value of $X$ is zero
- $b_1$ is the estimated change in the mean value of $Y$ for every one-unit change in $X$

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
### Linear Regression Example Data

<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>

<table>
<thead>
<tr>
<th>Square Feet (X)</th>
<th>House Price in $1000s (Y)</th>
<th>XY</th>
<th>X²</th>
</tr>
</thead>
<tbody>
<tr>
<td>1400</td>
<td>245</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1600</td>
<td>312</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1700</td>
<td>279</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1875</td>
<td>308</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1100</td>
<td>199</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1550</td>
<td>219</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2350</td>
<td>405</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2450</td>
<td>324</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1425</td>
<td>319</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1700</td>
<td>255</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Linear Regression Example

Scatterplot

- House price model: scatter plot

![Scatterplot](image)

Excel Output

The regression equation is:

$$\text{house price} = 98.24833 + 0.10977 (\text{square feet})$$
Linear Regression Example
Graphical Representation

- House price model: scatter plot and regression line

- Scatter plot with regression line.

- Slope: $0.10977$
- Intercept: $98.248$

- House price model: $Y = 98.24833 + 0.10977X$ (in thousands $\times 1000$s)

Interpretation of $b_0$

- $b_0$ is the estimated mean value of $Y$ when the value of $X$ is zero (if $X = 0$ is in the range of observed $X$ values).
- Because the square footage of the house cannot be 0, the Y intercept has no practical application.
Linear Regression Example

Interpretation of $b_1$

\[
\text{house price} = 98.24833 + 0.10977 \text{ (square feet)}
\]

- $b_1$ measures the mean change in the average value of $Y$ as a result of a one-unit change in $X$
- Here, $b_1 = 0.10977$ tells us that the mean value of a house increases by $0.10977(1000) = 109.77$, on average, for each additional one square foot of size

Linear Regression Example

Making Predictions

Predict the price for a house with 2000 square feet:

\[
\begin{align*}
\text{house price} &= 98.25 + 0.1098 \text{ (sq.ft.)} \\
&= 98.25 + 0.1098(2000) \\
&= 317.85
\end{align*}
\]

The predicted price for a house with 2000 square feet is $317.85(1000s) = 317,850$
Linear Regression Example  
Making Predictions

- When using a regression model for prediction, only predict within the relevant range of data

Relevant range for interpolation

Do not try to extrapolate beyond the range of observed X’s

Class Exercises

Applying the Concepts

10.4 The marketing manager of a large supermarket chain would like to use shelf space to predict the sales of pet food. A random sample of 12 equal-sized stores is selected, with the following results (stored in the file petfood.dat).

<table>
<thead>
<tr>
<th>Store</th>
<th>Shelf Space (X) (Feet)</th>
<th>Weekly Sales (Y) ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>160</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>220</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>140</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>190</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>240</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>260</td>
</tr>
<tr>
<td>7</td>
<td>15</td>
<td>230</td>
</tr>
<tr>
<td>8</td>
<td>15</td>
<td>270</td>
</tr>
<tr>
<td>9</td>
<td>15</td>
<td>280</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>260</td>
</tr>
<tr>
<td>11</td>
<td>20</td>
<td>290</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>310</td>
</tr>
</tbody>
</table>

- Construct a scatter plot.
- Interpret the meaning of the slope, $b_1$, in this problem.
- Predict the mean weekly sales (in hundreds of dollars) of pet food for stores with 8 feet of shelf space for pet food.
10.5 Circulation is the lifeblood of the publishing business. The larger the sales of a magazine, the more it can charge advertisers. Recently, a circulation gap has appeared between the publishers’ reports of magazines’ newsstand sales and subsequent audits by the Audit Bureau of Circulations. The data in the following table represent the reported and audited newsstand sales (in thousands) in 2001 for the following 10 magazines:

<table>
<thead>
<tr>
<th>Magazine</th>
<th>Reported (Y)</th>
<th>Audited (Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>YM</td>
<td>621.0</td>
<td>299.6</td>
</tr>
<tr>
<td>CosmoGirl</td>
<td>359.7</td>
<td>207.7</td>
</tr>
<tr>
<td>Baton</td>
<td>530.0</td>
<td>325.0</td>
</tr>
<tr>
<td>Playboy</td>
<td>-402.1</td>
<td>334.3</td>
</tr>
<tr>
<td>Esquire</td>
<td>703.5</td>
<td>48.6</td>
</tr>
<tr>
<td>Town/People</td>
<td>560.7</td>
<td>408.3</td>
</tr>
<tr>
<td>Move</td>
<td>120.5</td>
<td>91.2</td>
</tr>
<tr>
<td>Spic</td>
<td>50.6</td>
<td>91.1</td>
</tr>
<tr>
<td>Vogue</td>
<td>353.3</td>
<td>268.6</td>
</tr>
<tr>
<td>Elle</td>
<td>203.6</td>
<td>214.3</td>
</tr>
</tbody>
</table>


a. Construct Scatter Plot
b. Interpret the slope, b1 in this problem
c. Predict the mean audited newsstand sales for a magazine that reports newsstand sales of 400,000.

---

Measures of Variation

Total variation is made up of two parts:

\[
\text{SST} = \text{SSR} + \text{SSE}
\]

- **Total Sum of Squares**
- **Regression Sum of Squares**
- **Error Sum of Squares**

\[
\text{SST} = \sum (Y_i - \bar{Y})^2 \\
\text{SSR} = \sum (\hat{Y}_i - \bar{Y})^2 \\
\text{SSE} = \sum (Y_i - \hat{Y}_i)^2
\]

where:

- \(\bar{Y}\) = Mean value of the dependent variable
- \(Y_i\) = Observed values of the dependent variable
- \(\hat{Y}_i\) = Predicted value of Y for the given X value
Measures of Variation

- **SST** = total sum of squares
  - Measures the variation of the $Y_i$ values around their mean $Y$
- **SSR** = regression sum of squares
  - Explained variation attributable to the relationship between $X$ and $Y$
- **SSE** = error sum of squares
  - Variation attributable to factors other than the relationship between $X$ and $Y$

\[
\text{SST} = \sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2
\]

\[
\text{SSE} = \sum (Y_i - \hat{Y}_i)^2
\]

\[
\text{SSR} = \sum (\hat{Y}_i - \bar{Y})^2
\]
Linear Regression Example Data

<table>
<thead>
<tr>
<th>Square Feet (X)</th>
<th>House Price in $1000s (Y)</th>
<th>$Y_1$</th>
<th>SSR $(Y_1 - \hat{Y})^2$</th>
<th>SSE $(Y - \hat{Y})^2$</th>
<th>SST</th>
</tr>
</thead>
<tbody>
<tr>
<td>1400</td>
<td>245</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1600</td>
<td>312</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1700</td>
<td>279</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1875</td>
<td>308</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1100</td>
<td>199</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1550</td>
<td>219</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2350</td>
<td>405</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2450</td>
<td>324</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1425</td>
<td>319</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1700</td>
<td>255</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Class Exercises

Calculate the total variation in both of the exercises done above.
Coefficient of Determination, $r^2$

- 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 $r^2$.

$$
0 \leq r^2 \leq 1
$$

$r^2 = 1$

Perfect linear relationship between $X$ and $Y$:

100% of the variation in $Y$ is explained by variation in $X$.
Coefficient of Determination, $r^2$

When $0 < r^2 < 1$:
- Weaker linear relationships between $X$ and $Y$:
- Some but not all of the variation in $Y$ is explained by variation in $X$.

When $r^2 = 0$:
- No linear relationship between $X$ and $Y$:
- The value of $Y$ is not related to $X$.
- (None of the variation in $Y$ is explained by variation in $X$)
Linear Regression Example

Coefficient of Determination, \( r^2 \)

\[
\begin{align*}
\text{Multiple R} & = 0.76211 \\
\text{R Square} & = 0.58082 \\
\text{Adjusted R Square} & = 0.52842 \\
\text{Observations} & = 10
\end{align*}
\]

\[
\frac{\text{SSR}}{\text{SST}} = \frac{18934.9348}{32600.5000} = 0.58082
\]

58.08% of the variation in house prices is explained by variation in square feet.

Class Exercises

Calculate the coefficient of determination in both of the exercises done above.
# Assumptions of Regression

**L.I.N.E**

- **Linearity**
  - The relationship between X and Y is linear
- **Independence of Errors**
  - Error values are statistically independent
- **Normality of Error**
  - Error values are normally distributed for any given value of X
- **Equal Variance** (also called homoscedasticity)
  - The probability distribution of the errors has constant variance