Chapter 10
Regression
Scatter Diagrams

- A graph in which pairs of points, \((x, y)\), are plotted with \(x\) on the horizontal axis and \(y\) on the vertical axis.
- The explanatory variable is \(x\).
- The response variable is \(y\).
- One goal of plotting paired data is to determine if there is a linear relationship between \(x\) and \(y\).
Finding the Best Fitting Line

- One goal of data analysis is to find a mathematical equation that “best” represents the data.

- For our purposes, “best” means the line that comes closest to each point on the scatter diagram.
Not All Relationships are Linear

**FIGURE 10-2**

Scatter Diagrams with No Linear Correlation

(a)  

(b)
Correlation Strength

• Correlation is a measure of the *linear* relationship between two quantitative variables.
• If the points lie exactly on a straight line, we say there is *perfect correlation*.
• As the strength of the relationship decreases, the correlation moves from *perfect* to *strong* to *moderate* to *weak*. 
Correlation

- If the response variable, $y$, tends to increase as the explanatory variable, $x$, increases, the variables have *positive correlation*.
- If the response variable, $y$, tends to decrease as the explanatory variable, $x$, increases, the variables have *negative correlation*.
Correlation Examples

**FIGURE 10-4**

Scatter Diagrams with Moderate and Perfect Linear Correlation

- **Moderate linear correlation**
  - (a)
  - (b)

- **Perfect linear correlation**
  - (c)
  - (d)
Correlation Coefficient

• The mathematical measurement that describes the correlation is called the sample correlation coefficient.

• Denoted by $r$. 
Correlation Coefficient Features

1) $r$ has no units.
2) $-1 \leq r \leq 1$
3) Positive values of $r$ indicate a positive relationship between $x$ and $y$ (likewise for negative values).
4) $r = 0$ indicates no linear relationship.
5) Switching the explanatory variable and response variable does not change $r$.
6) Changing the units of the variables does not change $r$. 
There is no linear relation for the points of the scatter diagram.
$r = 1$ or $r = -1$

There is a perfect linear relation between $x$ and $y$ values; all points lie on the least-squares line.
The $x$ and $y$ values have a positive correlation. By this, we mean that large $x$ values are associated with large $y$ values, and small $x$ values are associated with small $y$ values. As we go from left to right, the least-squares line goes up.
-1 < r < 0

The x and y values have a negative correlation. By this, we mean large x values are associated with small y values, and small x values are associated with large y values.

As we go from left to right, the least-squares line goes down.
Developing a Formula for $r$

**FIGURE 10-6**

Patterns for Linear Correlation

(a) Positive linear correlation

(b) Negative linear correlation

(c) Little or no linear correlation
Developing a Formula for $r$

\[ r = \frac{1}{n - 1} \sum \frac{(y - \bar{y})}{s_y} \cdot \frac{(x - \bar{x})}{s_x} \]
A Computational Formula for $r$

Obtain a random sample of $n$ data pairs $(x, y)$. The data pairs should have a \textit{bivariate normal distribution}. This means that for a fixed value of $x$, the $y$ values should have a normal distribution (or at least a mound-shaped and symmetric distribution), and for a fixed $y$, the $x$ values should have their own (approximately) normal distribution.

1. Using the data pairs, compute $\Sigma x$, $\Sigma y$, $\Sigma x^2$, $\Sigma y^2$, and $\Sigma xy$.

2. With $n =$ sample size, $\Sigma x$, $\Sigma y$, $\Sigma x^2$, $\Sigma y^2$, and $\Sigma xy$, you are ready to compute the sample correlation coefficient $r$ using the computation formula

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

Be careful! The notation $\Sigma x^2$ means first square $x$ and then calculate the sum, whereas $(\Sigma x)^2$ means first sum the $x$ values, then square the result.
Critical Thinking: Population vs. Sample Correlation

• Beware: high correlations do not imply a cause and effect relationship between the explanatory and response variables!!

\[ r = \text{sample correlation coefficient computed from a random sample of } (x, y) \text{ data pairs.} \]

\[ \rho = \text{population correlation coefficient computed from all population data pairs } (x, y). \]
Critical Thinking: Lurking Variables

- A lurking variable is neither the explanatory variable nor the response variable.
- Beware: lurking variables may be responsible for the evident changes in $x$ and $y$. 


Correlations Between Averages

• If the two variables of interest are averages rather than raw data values, the correlation between the averages will tend to be higher!
Linear Regression

- When we are in the “paired-data” setting:
  - Do the data points have a linear relationship?
  - Can we find an equation for the best fitting line?
  - Can we predict the value of the response variable for a new value of the predictor variable?
  - What fractional part of the variability in $y$ is associated with the variability in $x$?
Least-Squares Criterion

• The line that we fit to the data will be such that the sum of the squared distances from the line will be minimized.

• Let $d =$ the vertical distance from any data point to the regression line.
Least-Squares Criterion

**FIGURE 10-10**

Least-Squares Criterion

\[ \Sigma d^2 \text{ is as small as possible} \]

Wolves Y

Caribou (100’s) X
Least-Squares Line

\[ \hat{y} = a + bx \]

- \( a \) is the \( y \)-intercept.
- \( b \) is the slope.
Finding the Regression Equation

Obtain a random sample of $n$ data pairs $(x, y)$, where $x$ is the explanatory variable and $y$ is the response variable. The data pairs should have a bivariate normal distribution. This means that for a fixed value of $x$, the $y$ values should have a normal distribution (or at least a mound-shaped and symmetric distribution), and for a fixed $y$, the $x$ values should have their own (approximately) normal distribution.

1. Using the data pairs, compute $\Sigma x$, $\Sigma y$, $\Sigma x^2$, $\Sigma y^2$, and $\Sigma xy$. Then compute the sample means $\bar{x}$ and $\bar{y}$. 
Finding the Regression Equation

2. With \( n = \) sample size, \( \Sigma x \), \( \Sigma y \), \( \Sigma x^2 \), \( \Sigma y^2 \), \( \Sigma xy \), \( \bar{x} \), and \( \bar{y} \), you are ready to compute the slope \( b \) and intercept \( a \) using the computation formulas:

\[
\text{Slope: } b = \frac{n \Sigma xy - (\Sigma x)(\Sigma y)}{n \Sigma x^2 - (\Sigma x)^2}
\]

\[
\text{Intercept: } a = \bar{y} - b \bar{x}
\]

Be careful! The notation \( \Sigma x^2 \) means first square \( x \) and then calculate the sum, whereas \( (\Sigma x)^2 \) means first sum the \( x \) values, then square the result.

3. The equation of the least-squares line computed from your sample data is

\[
\hat{y} = a + bx
\]
Using the Regression Equation

• The point \((\bar{x}, \bar{y})\) is always on the least-squares line.

• Also, the slope tells us that if we increase the explanatory variable by one unit, the response variable will change by the slope (increase or decrease, depending on the sign of the slope).
Critical Thinking: Making Predictions

• We can simply plug in $x$ values into the regression equation to calculate $y$ values.

Predicting $\hat{y}$ values for $x$ values that are between observed $x$ values in the data set is called **interpolation**.

Predicting $\hat{y}$ values of $x$ values that are beyond observed $x$ values in the data set is called **extrapolation**.

• Extrapolation can result in unreliable predictions.
Coefficient of Determination

- Another way to gauge the fit of the regression equation is to calculate the coefficient of determination, $r^2$.
- We can view the mean of $y$ as a baseline for all the $y$ values.
- Then:

  \[
  \text{Total deviation} = y - \bar{y}
  \]

  \[
  \text{Explained deviation} = \hat{y} - \bar{y}
  \]

  \[
  \text{Unexplained deviation} = y - \hat{y}
  \]
Explained and Unexplained Deviations

\[ \hat{y} = a + bx \]

Least square line

Unexplained deviation = \( y - \hat{y} \)

\( (x, \hat{y}) \)

\( (x, y) \)

\( (x, \bar{y}) \)

\( y - \bar{y} = \text{Total deviation} \)

\( \hat{y} - \bar{y} = \text{Explained deviation} \)

Base line
Coefficient of Determination

We algebraically manipulate the above equation to obtain:

\[
\begin{align*}
(y - \bar{y}) & = (\hat{y} - \bar{y}) + (y - \hat{y}) \\
\text{(Total deviation)} & = \text{(Explained deviation)} + \text{(Unexplained deviation)}
\end{align*}
\]

\[
\begin{align*}
\Sigma(y - \bar{y})^2 & = \Sigma(\hat{y} - \bar{y})^2 + \Sigma(y - \hat{y})^2 \\
\text{(Total variation)} & = \text{(Explained variation)} + \text{(Unexplained variation)}
\end{align*}
\]
Coefficient of Determination

If \( r \) is the correlation coefficient [see Equation (2)], then it can be shown that

\[
r^2 = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2} = \frac{\text{Explained variation}}{\text{Total variation}}
\]

\( r^2 \) is called the coefficient of determination.

Thus, \( r^2 \) is the measure of the total variability in \( y \) that is explained by the regression on \( x \).
Coefficient of determination $r^2$

1. Compute the sample correlation coefficient $r$ using the procedure of Section 10.1. Then simply compute $r^2$, the sample coefficient of determination.

2. The value $r^2$ is the ratio of explained variation over total variation. That is, $r^2$ is the fractional amount of total variation in $y$ that can be explained by using the linear model $\hat{y} = a + bx$.

3. Furthermore, $1 - r^2$ is the fractional amount of total variation in $y$ that is due to random chance or to the possibility of lurking variables that influence $y$. 
Inferences on the Regression Equation

- We can make inferences about the population correlation coefficient, $\rho$, and the population regression line slope, $\beta$.

<table>
<thead>
<tr>
<th>Sample Statistic</th>
<th>Population Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r$</td>
<td>$\rho$</td>
</tr>
<tr>
<td>$a$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>$b$</td>
<td>$\beta$</td>
</tr>
<tr>
<td>$\hat{y} = a + bx$</td>
<td>$y = \alpha + \beta x$</td>
</tr>
</tbody>
</table>
Inferences on the Regression Equation

• Inference requirements:
  – The set of ordered pairs \((x, y)\) constitutes a random sample from all ordered pairs in the population.
  – For each fixed value of \(x\), \(y\) has a normal distribution.
  – All distributions for all \(y\) values have equal variance and a mean that lies on the regression equation.
Testing the Correlation Coefficient

$H_0$: $x$ and $y$ have no linear correlation, so $\rho = 0$

$H_1$: $\rho > 0$

$H_1$: $\rho < 0$

$H_1$: $\rho \neq 0$
Testing the Correlation Coefficient

- We convert to a Student’s t distribution.
- The sample size must be $\geq 3$.

$$t = \frac{r \sqrt{n - 2}}{\sqrt{1 - r^2}} \quad \text{with d.f.} = n - 2$$
Testing Procedure

1. Use the null hypothesis $H_0: \rho = 0$. In the context of the application, state the alternate hypothesis ($\rho > 0$ or $\rho < 0$ or $\rho \neq 0$) and set the level of significance $\alpha$.

2. Obtain a random sample of $n \geq 3$ data pairs $(x, y)$ and compute the sample test statistic

   $$t = \frac{r\sqrt{n - 2}}{\sqrt{1 - r^2}}$$

   with degrees of freedom $d.f. = n - 2$.

3. Use a Student’s $t$ distribution and the type of test, one-tailed or two-tailed, to find (or estimate) the $P$-value corresponding to the test statistic.

4. Conclude the test. If $P$-value $\leq \alpha$, then reject $H_0$. If $P$-value $> \alpha$, then do not reject $H_0$.

5. State your conclusion in the context of the application.
Standard Error of Estimate

\[
\text{Standard error of estimate } = S_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}}
\]

where \( \hat{y} = a + bx \) and \( n \geq 3 \).

- As the points get closer to the regression line, \( S_e \) decreases.
- If all the points lie exactly on the line, \( S_e \) will be zero.
Computational Formula For $S_e$

1. Obtain a random sample of $n \geq 3$ data pairs $(x, y)$.
2. Use the procedures of Section 10.2 to find $a$ and $b$ from the sample least-squares line $\hat{y} = a + bx$.
3. The standard error of estimate is

$$S_e = \sqrt{\frac{\Sigma y^2 - a \Sigma y - b \Sigma xy}{n - 2}}$$
Confidence Intervals for $y$

- The population regression line:
  \[ y = \alpha + \beta x + \varepsilon \]
  Where $\varepsilon$ is random error

- Because of $\varepsilon$, for each $x$ value there is a corresponding distribution for $y$.

- Each $y$ distribution has the same standard deviation, estimated by $S_\varepsilon$.

- Each $y$ distribution is centered on the regression line.
Confidence Intervals for Predicted $y$

1. Obtain a random sample of $n \geq 3$ data pairs $(x, y)$.
2. Use the procedure of Section 10.2 to find $\hat{y} = a + bx$. You also need to find $\bar{x}$ from the sample data and the standard error of estimate $S_e$ using equation (8) of this section.
3. The $c$ confidence interval for $y$ for a specified value of $x$ is

$$\hat{y} - E < y < \hat{y} + E$$
Confidence Intervals for Predicted $y$

$$E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{n(x - \bar{x})^2}{n \Sigma x^2 - (\Sigma x)^2}}$$

$\hat{y} = a + bx$ is the predicted value of $y$ from the least-squares line for a specified $x$ value.

$c = \text{confidence level (} 0 < c < 1\text{)}$

$n = \text{number of data pairs (} n \geq 3\text{)}$

$t_c = \text{critical value from Student’s } t \text{ distribution for } c \text{ confidence level using } d.f. = n - 2$

$S_e = \text{standard error of estimate}$
Confidence Intervals for Predicted $y$

- The confidence intervals will be narrower near the mean of $x$.
- The confidence intervals will be wider near the extreme values of $x$. 
Confidence Intervals for Predicted $y$

**FIGURE 10-17**

95% Confidence Band for Predicted Values $\hat{y}$

- **Grams of CuSO₄**
- **°C**

Data points and trend line showing the confidence interval for predicted values.
Inferences About Slope $\beta$

- Recall the population regression equation: $y = \alpha + \beta x + \epsilon$
- Let $S_e$ be the standard error of estimate from the sample.
Inferences About Slope $\beta$

$$t = \frac{b - \beta}{S_e \sqrt{\sum x^2 - \frac{1}{n} (\sum x)^2}}$$

- Has a Student’s $t$ distribution with $n - 2$ degrees of freedom.
Testing $\beta$

For a statistical test of $\beta$

1. Use the null hypothesis $H_0$: $\beta = 0$. Use an alternate hypothesis $H_1$ appropriate to your application ($\beta > 0$ or $\beta < 0$ or $\beta \neq 0$). Set the level of significance $\alpha$.

2. Use the null hypothesis $H_0$: $\beta = 0$ and the values of $S_e$, $n$, $\Sigma x$, $\Sigma x^2$, and $b$ to compute the sample test statistic

$$t = \frac{b}{S_e} \sqrt{\frac{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2}{\Sigma x^2}} \quad \text{with d.f.} = n - 2$$

3. Use a Student’s $t$ distribution and the type of test, one-tailed or two-tailed, to find (or estimate) the $P$-value corresponding to the test statistic.

4. Conclude the test. If $P$-value $\leq \alpha$, then reject $H_0$. If $P$-value $> \alpha$, then do not reject $H_0$.

5. State your conclusion in the context of the application.
Finding a Confidence Interval for $\beta$

$$b - E < \beta < b + E$$

where

$$E = \frac{t_c S_e}{\sqrt{\sum x^2 - \frac{1}{n} (\sum x)^2}}$$

$c$ = confidence level $(0 < c < 1)$

$n$ = number of data pairs $(x, y)$, $n \geq 3$

$t_c$ = Student's $t$ distribution critical value for confidence level $c$ and

d.f. = $n - 2$

$S_e$ = standard error of estimate
Multiple Regression

• In practice, most predictions will improve if we consider additional data.

• Statistically, this means adding predictor variables to our regression techniques.
Multiple Regression Terminology

\[ y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k \]

- \( y \) is the response variable.
- \( x_1 \ldots x_k \) are explanatory variables.
- \( b_0 \ldots b_k \) are coefficients.
Regression Models –
A mathematical package that includes the following:

1) A collection of variables, one of which is the response, and the rest are predictors.
2) A collection of numerical values associated with the given variables.
3) Using software and the above numerical values, construct a least-squares regression equation.
Regression Models –
A mathematical package that includes the following:

4) The model includes information about the variables, coefficients of the equation, and various “goodness of fit” measures.

5) The model allows the user to make predictions and build confidence intervals for various components of the equation.
Coefficient of Multiple Determination

• How well does our equation fit the data?
  – Just as in linear regression, computer output will provide $r^2$.
• $r^2$ is a measure of the amount of variability in $y$ that is explained by the regression on all of the $x$ variables.
Predictions in the Multiple Regression Case

• Similar to the techniques described for linear regression, we can “plug in” values for all of the \( x \) variables and compute a predicted \( y \) value.

  – Predicting \( y \) for any \( x \) values outside the \( x \)-range is extrapolation, and the results will be unreliable.
Testing a Coefficient for Significance

• At times, one or more of the predictor variables may not be helpful in determining $y$.
• Using software, we can test any or all of the $x$ variables for statistical significance.
Testing a Coefficient for Significance

• The hypotheses will be:
  
  \[ H_0: \beta_i = 0 \]
  
  \[ H_1: \beta_i \neq 0 \]

  for the \( i^{th} \) coefficient in the model.

• If we fail to reject the null hypothesis for a given coefficient, we may consider removing that predictor from the model.
Confidence Intervals for Coefficients

- As in the linear case, software can provide confidence intervals for any $\beta_i$ in the model.