Chapter 9

Correlation and Regression

§ 9.1

Correlation

A correlation is a relationship between two variables. The data can be represented by the ordered pairs \((x, y)\) where \(x\) is the independent (or explanatory) variable, and \(y\) is the dependent (or response) variable.

A scatter plot can be used to determine whether a linear (straight line) correlation exists between two variables.

Example:

<table>
<thead>
<tr>
<th>(x)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>(y)</td>
<td>-4</td>
<td>-2</td>
<td>-1</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>
Linear Correlation

- As $x$ increases, $y$ tends to increase:
  - Negative Linear Correlation
  - No Correlation
  - Nonlinear Correlation

Correlation Coefficient

The **correlation coefficient** is a measure of the strength and the direction of a linear relationship between two variables. The symbol $r$ represents the sample correlation coefficient. The formula for $r$ is:

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

The range of the correlation coefficient is $-1$ to $1$. If $x$ and $y$ have a strong positive linear correlation, $r$ is close to 1. If $x$ and $y$ have a strong negative linear correlation, $r$ is close to $-1$. If there is no linear correlation or a weak linear correlation, $r$ is close to 0.

Linear Correlation

- $r = -0.91$: Strong negative correlation
- $r = 0.88$: Strong positive correlation
- $r = 0.42$: Weak positive correlation
- $r = 0.07$: Nonlinear Correlation
Calculating a Correlation Coefficient

Calculating a Correlation Coefficient

In Words
1. Find the sum of the x-values.
2. Find the sum of the y-values.
3. Multiply each x-value by its corresponding y-value and find the sum.
4. Square each x-value and find the sum.
5. Square each y-value and find the sum.
6. Use these five sums to calculate the correlation coefficient.

In Symbols
\[
\begin{align*}
\text{In Words} & \quad \text{In Symbols} \\
\text{1. } \Sigma x & \quad \Sigma x \\
\text{2. } \Sigma y & \quad \Sigma y \\
\text{3. } \Sigma xy & \quad \Sigma xy \\
\text{4. } \Sigma x^2 & \quad \Sigma x^2 \\
\text{5. } \Sigma y^2 & \quad \Sigma y^2 \\
\text{6. } \frac{\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{\Sigma x^2 - (\Sigma x)^2} \sqrt{\Sigma y^2 - (\Sigma y)^2}} & \quad \frac{\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{\Sigma x^2 - (\Sigma x)^2} \sqrt{\Sigma y^2 - (\Sigma y)^2}}
\end{align*}
\]

Correlation Coefficient

Example:
Calculate the correlation coefficient \(r\) for the following data.

<table>
<thead>
<tr>
<th>(x)</th>
<th>(y)</th>
<th>(x^2)</th>
<th>(y^2)</th>
<th>(xy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>1</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>16</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>25</td>
<td>4</td>
<td>10</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\Sigma x &= 15 \\
\Sigma y &= -3 \\
\Sigma xy &= 55 \\
\Sigma x^2 &= 55 \\
\Sigma y^2 &= 54 \\
\sum (55) &= 1.05 (15) = 1.05 \times \frac{55}{15} = 0.986 \\
\end{align*}
\]

There is a strong positive linear correlation between \(x\) and \(y\).

Correlation Coefficient

Example:
The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

a.) Display the scatter plot.
b.) Calculate the correlation coefficient \(r\).

<table>
<thead>
<tr>
<th>Hours, (x)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>3</th>
<th>5</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>7</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test score, (y)</td>
<td>96</td>
<td>85</td>
<td>82</td>
<td>74</td>
<td>95</td>
<td>68</td>
<td>76</td>
<td>84</td>
<td>58</td>
<td>65</td>
<td>75</td>
</tr>
</tbody>
</table>
Correlation Coefficient

Example continued:

<table>
<thead>
<tr>
<th>Hours, x</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>7</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test score, y</td>
<td>96</td>
<td>85</td>
<td>82</td>
<td>74</td>
<td>95</td>
<td>68</td>
<td>76</td>
<td>84</td>
<td>87</td>
<td>75</td>
</tr>
</tbody>
</table>

![Graph showing correlation between hours watching TV and test scores.]

Continued:

Correlation Coefficient

Example continued:

<table>
<thead>
<tr>
<th>Hours, x</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>7</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test score, y</td>
<td>96</td>
<td>85</td>
<td>82</td>
<td>74</td>
<td>95</td>
<td>68</td>
<td>76</td>
<td>84</td>
<td>87</td>
<td>75</td>
</tr>
<tr>
<td>Σx = 54</td>
<td>Σy = 908</td>
<td>Σxy = 3724</td>
<td>Σx² = 332</td>
<td>Σy² = 70836</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\rho = \frac{\sum xy - n \bar{x} \bar{y}}{\sqrt{\sum x^2 - n \bar{x}^2} \sqrt{\sum y^2 - n \bar{y}^2}} = \frac{12037241 - (54)(908)}{\sqrt{12037241 - (54)^2} \sqrt{70836 - (908)^2}} = -0.831
\]

There is a strong negative linear correlation. As the number of hours spent watching TV increases, the test scores tend to decrease.

Testing a Population Correlation Coefficient

Once the sample correlation coefficient \( r \) has been calculated, we need to determine whether there is enough evidence to decide that the population correlation coefficient \( \rho \) is significant at a specified level of significance.

One way to determine this is to use Table 11 in Appendix B.

If \( |r| \) is greater than the critical value, there is enough evidence to decide that the correlation coefficient \( \rho \) is significant.

<table>
<thead>
<tr>
<th>n</th>
<th>( \alpha = 0.05 )</th>
<th>( \alpha = 0.01 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.950</td>
<td>0.990</td>
</tr>
<tr>
<td>5</td>
<td>0.878</td>
<td>0.959</td>
</tr>
<tr>
<td>6</td>
<td>0.831</td>
<td>0.911</td>
</tr>
<tr>
<td>7</td>
<td>0.754</td>
<td>0.875</td>
</tr>
</tbody>
</table>

For a sample of size \( n = 6 \), \( \rho \) is significant at the 5% significance level, if \( |r| > 0.811 \).
Testing a Population Correlation Coefficient

Finding the Correlation Coefficient $\rho$

<table>
<thead>
<tr>
<th>In Words</th>
<th>In Symbols</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Determine the number of pairs of data in the sample.</td>
<td>Determine $n$.</td>
</tr>
<tr>
<td>2. Specify the level of significance.</td>
<td>Identify $\alpha$.</td>
</tr>
<tr>
<td>3. Find the critical value.</td>
<td>Use Table 11 in Appendix B.</td>
</tr>
<tr>
<td>4. Decide if the correlation is significant.</td>
<td>If $</td>
</tr>
<tr>
<td>5. Interpret the decision in the context of the original claim.</td>
<td></td>
</tr>
</tbody>
</table>

Testing a Population Correlation Coefficient

Example:
The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

The correlation coefficient $r \approx -0.831$. Is the correlation coefficient significant at $\alpha = 0.01$?

Testing a Population Correlation Coefficient continued:

<table>
<thead>
<tr>
<th>Example continued:</th>
<th>Appendix B: Table 11</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r = -0.831$</td>
<td></td>
</tr>
<tr>
<td>$n = 12$</td>
<td></td>
</tr>
<tr>
<td>$\alpha = 0.01$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$</td>
</tr>
</tbody>
</table>

Because, the population correlation is significant, there is enough evidence at the 1% level of significance to conclude that there is a significant linear correlation between the number of hours of television watched during the weekend and the scores of each student who took a test the following Monday.
Hypothesis Testing for $\rho$

A hypothesis test can also be used to determine whether the sample correlation coefficient $r$ provides enough evidence to conclude that the population correlation coefficient $\rho$ is significant at a specified level of significance.

A hypothesis test can be one tailed or two tailed.

| $H_0$: $\rho \geq 0$ (no significant negative correlation) | Left-tailed test |
| $H_0$: $\rho < 0$ (significant negative correlation) |
| $H_0$: $\rho \leq 0$ (no significant positive correlation) | Right-tailed test |
| $H_0$: $\rho > 0$ (significant positive correlation) |
| $H_0$: $\rho = 0$ (no significant correlation) | Two-tailed test |
| $H_0$: $\rho \neq 0$ (significant correlation) |

Hypothesis Testing for $\rho$

The $t$-Test for the Correlation Coefficient

A $t$-test can be used to test whether the correlation between two variables is significant. The test statistic is $r$ and the standardized test statistic

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

follows a $t$-distribution with $n-2$ degrees of freedom.

In this text, only two-tailed hypothesis tests for $\rho$ are considered.

Hypothesis Testing for $\rho$

Using the $t$-Test for the Correlation Coefficient $\rho$

1. State the null and alternative hypothesis.
2. Specify the level of significance. Identify $\alpha$.
3. Identify the degrees of freedom. d.f. = $n - 2$
4. Determine the critical value(s) and rejection region(s).

Use Table 5 in Appendix B.
Hypothesis Testing for $\rho$

Using the $t$-Test for the Correlation Coefficient $\rho$

In Words
5. Find the standardized test statistic.
6. Make a decision to reject or fail to reject the null hypothesis.
7. Interpret the decision in the context of the original claim.

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

If $t$ is in the rejection region, reject $H_0$. Otherwise fail to reject $H_0$.

Hypothesis Testing for $\rho$

Example:
The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

The correlation coefficient $r \approx -0.831$.

<table>
<thead>
<tr>
<th>Hours, $x$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>3</th>
<th>5</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>7</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test score, $y$</td>
<td>96</td>
<td>85</td>
<td>82</td>
<td>74</td>
<td>95</td>
<td>68</td>
<td>76</td>
<td>84</td>
<td>58</td>
<td>65</td>
<td>75</td>
</tr>
</tbody>
</table>

Test the significance of this correlation coefficient significant at $\alpha = 0.01$?

Continued:

Example continued:
$H_0: \rho = 0$ (no correlation)  $H_a: \rho \neq 0$ (significant correlation)

The level of significance is $\alpha = 0.01$.

Degrees of freedom are $d.f. = 12 - 2 = 10$.

The critical values are $-t_{0.005} = -3.169$ and $t_{0.005} = 3.169$.

The standardized test statistic is
$t = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}}$

The test statistic falls in the rejection region, so $H_0$ is rejected.

At the 1% level of significance, there is enough evidence to conclude that there is a significant linear correlation between the number of hours of TV watched over the weekend and the test scores on Monday morning.
Correlation and Causation

The fact that two variables are strongly correlated does not in itself imply a cause-and-effect relationship between the variables. If there is a significant correlation between two variables, you should consider the following possibilities.

1. Is there a direct cause-and-effect relationship between the variables? Does \( x \) cause \( y \)?
2. Is there a reverse cause-and-effect relationship between the variables? Does \( y \) cause \( x \)?
3. Is it possible that the relationship between the variables can be caused by a third variable or by a combination of several other variables?
4. Is it possible that the relationship between two variables may be a coincidence?

§ 9.2
Linear Regression

After verifying that the linear correlation between two variables is significant, next we determine the equation of the line that can be used to predict the value of \( y \) for a given value of \( x \). Each data point \( d_i \) represents the difference between the observed \( y \)-value and the predicted \( y \)-value for a given \( x \)-value on the line. These differences are called residuals.
Regression Line

A regression line, also called a line of best fit, is the line for which the sum of the squares of the residuals is a minimum.

The Equation of a Regression Line

The equation of a regression line for an independent variable $x$ and a dependent variable $y$ is

$$y = mx + b$$

where $y$ is the predicted $y$-value for a given $x$-value. The slope $m$ and $y$-intercept $b$ are given by

$$m = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$
$$b = \bar{y} - mx$$

where $\bar{y}$ is the mean of the $y$-values and $\bar{x}$ is the mean of the $x$-values. The regression line always passes through $(\bar{x}, \bar{y})$.

Example:

Find the equation of the regression line.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
<th>$xy$</th>
<th>$x^2$</th>
<th>$y^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-3</td>
<td>-3</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>-1</td>
<td>-2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>10</td>
<td>25</td>
<td>4</td>
</tr>
<tr>
<td>$\Sigma x = 15$</td>
<td>$\Sigma y = -1$</td>
<td>$\Sigma xy = 55$</td>
<td>$\Sigma x^2 = 55$</td>
<td>$\Sigma y^2 = 16$</td>
</tr>
</tbody>
</table>

$$m = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2} = \frac{559 - (15)(-1)}{555 - (15)^2} = \frac{60}{50} = 1.2$$

Continued:

$$b = \bar{y} - mx = -1 - (1.2)(\frac{15}{5}) = -3.8$$

The equation of the regression line is

$$\hat{y} = 1.2x - 3.8.$$
Regression Line

Example:
The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

<table>
<thead>
<tr>
<th>Hours, x</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>7</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test score, y</td>
<td>85</td>
<td>164</td>
<td>222</td>
<td>285</td>
<td>340</td>
<td>380</td>
<td>420</td>
<td>445</td>
<td>525</td>
<td>504</td>
</tr>
<tr>
<td>∑x</td>
<td>90</td>
<td>32</td>
<td>54</td>
<td>39</td>
<td>88</td>
<td>76</td>
<td>84</td>
<td>94</td>
<td>65</td>
<td>75</td>
</tr>
<tr>
<td>∑y</td>
<td>308</td>
<td>1178</td>
<td>1702</td>
<td>2030</td>
<td>2380</td>
<td>2460</td>
<td>2560</td>
<td>2630</td>
<td>2790</td>
<td>2730</td>
</tr>
<tr>
<td>∑xy</td>
<td>13724</td>
<td>14944</td>
<td>16164</td>
<td>16459</td>
<td>17260</td>
<td>17624</td>
<td>17780</td>
<td>17916</td>
<td>17280</td>
<td>17280</td>
</tr>
</tbody>
</table>

a.) Find the equation of the regression line.
b.) Use the equation to find the expected test score for a student who watches 9 hours of TV.

\[
m = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} = \frac{12(3724) - (54)(908)}{12(332) - (54)^2} = \frac{12(3724) - (54)(908)}{12(332) - (54)^2} = -4.067
\]

\[
b = \bar{y} - m\bar{x} = \frac{908}{12} - (-4.067) \cdot \frac{54}{12} = 93.97
\]

\[
y = -4.07x + 93.97
\]

Regression Line

Example continued:
Using the equation \(y = -4.07x + 93.97\), we can predict the test score for a student who watches 9 hours of TV.

\[
y = -4.07(9) + 93.97 = -4.07(9) + 93.97 = 57.34
\]

A student who watches 9 hours of TV over the weekend can expect to receive about a 57.34 on Monday’s test.
§ 9.3
Measures of Regression and Prediction Intervals

Variation About a Regression Line

To find the total variation, you must first calculate the total deviation, the explained deviation, and the unexplained deviation.

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

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

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

Total variation about a regression line is the sum of the squares of the differences between the \( y \)-value of each ordered pair and the mean of \( y \). The explained variation is the sum of the squares of the differences between each predicted \( y \)-value and the mean of \( y \). The unexplained variation is the sum of the squares of the differences between the \( y \)-value of each ordered pair and each corresponding predicted \( y \)-value.

Total variation = \( \sum (y_i - \bar{y})^2 \)

Explained variation = \( \sum (\hat{y}_i - \bar{y})^2 \)

Unexplained variation = \( \sum (y_i - \hat{y}_i)^2 \)

Total variation = Explained variation + Unexplained variation
**Coefficient of Determination**

The coefficient of determination \( r^2 \) is the ratio of the explained variation to the total variation. That is,

\[
r^2 = \frac{\text{Explained variation}}{\text{Total variation}}
\]

**Example:**

The correlation coefficient for the data that represents the number of hours students watched television and the test scores of each student is \( r = -0.831 \). Find the coefficient of determination.

\[
r^2 = (-0.831)^2 \\
\approx 0.691
\]

About 69.1% of the variation in the test scores can be explained by the variation in the hours of TV watched. About 30.9% of the variation is unexplained.

---

**The Standard Error of Estimate**

When a \( y \)-value is predicted from an \( x \)-value, the prediction is a point estimate. An interval can also be constructed.

The **standard error of estimate** \( s_x \) is the standard deviation of the observed \( y_i \)-values about the predicted \( \hat{y} \)-value for a given \( x_i \)-value. It is given by

\[
s_x = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n - 2}}
\]

where \( n \) is the number of ordered pairs in the data set.

The closer the observed \( y \)-values are to the predicted \( y \)-values, the smaller the standard error of estimate will be.

---

**Finding the Standard Error of Estimate**

**In Words**

1. Make a table that includes the column heading shown.
2. Use the regression equation to calculate the predicted \( y \)-values.
3. Calculate the sum of the squares of the differences between each observed \( y \)-value and the corresponding predicted \( y \)-value.
4. Find the standard error of estimate.

**In Symbols**

\[
x = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n - 2}}
\]
The Standard Error of Estimate

Example:
The regression equation for the following data is
\( \hat{y} = 1.2x - 3.8 \).
Find the standard error of estimate.

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>( \hat{y} )</th>
<th>( (y - \hat{y})^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>2.6</td>
<td>0.16</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1.4</td>
<td>0.16</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0.2</td>
<td>0.04</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1.0</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>2.2</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Unexplained variation

\[ s_x = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} \]

\[ = \sqrt{\frac{0.365}{3}} = 0.46 \]

The standard deviation of the predicted \( y \) value for a given \( x \) value is about 0.365.

Example continued:

The regression equation for the data that represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday is
\( \hat{y} = -4.07x + 93.97 \).
Find the standard error of estimate.

<table>
<thead>
<tr>
<th>Hours, ( x )</th>
<th>Test score, ( y )</th>
<th>( y_i - \hat{y}_i )</th>
<th>( (y_i - \hat{y}_i)^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>96</td>
<td>93.97</td>
<td>6.03</td>
</tr>
<tr>
<td>1</td>
<td>85</td>
<td>89.9</td>
<td>7.01</td>
</tr>
<tr>
<td>2</td>
<td>82</td>
<td>85.83</td>
<td>9.08</td>
</tr>
<tr>
<td>3</td>
<td>74</td>
<td>81.76</td>
<td>11.31</td>
</tr>
<tr>
<td>4</td>
<td>93</td>
<td>73.62</td>
<td>56.93</td>
</tr>
<tr>
<td>5</td>
<td>68</td>
<td>33.54</td>
<td>113.53</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hours, ( x )</th>
<th>Test score, ( y )</th>
<th>( y_i - \hat{y}_i )</th>
<th>( (y_i - \hat{y}_i)^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>76</td>
<td>73.62</td>
<td>3.96</td>
</tr>
<tr>
<td>6</td>
<td>84</td>
<td>69.55</td>
<td>20.17</td>
</tr>
<tr>
<td>7</td>
<td>65</td>
<td>65.48</td>
<td>2.97</td>
</tr>
<tr>
<td>8</td>
<td>75</td>
<td>53.27</td>
<td>129.43</td>
</tr>
<tr>
<td>9</td>
<td>50</td>
<td>90.83</td>
<td>90.83</td>
</tr>
<tr>
<td>10</td>
<td>65</td>
<td>107.74</td>
<td>113.53</td>
</tr>
</tbody>
</table>

Continued.

The standard deviation of the student test scores for a specific number of hours of TV watched is about 8.11.
Prediction Intervals

Two variables have a bivariate normal distribution if for any fixed value of \( x \), the corresponding values of \( y \) are normally distributed and for any fixed values of \( y \), the corresponding \( x \)-values are normally distributed.

A prediction interval can be constructed for the true value of \( y \).

Given a linear regression equation \( \hat{y} = mx + b \) and \( x_0 \), a specific value of \( x \), a prediction interval for \( y \) is:
\[
\hat{y} - E < y < \hat{y} + E
\]
where
\[
E = t_{\alpha/2} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x_i^2 - \sum (x_i - \bar{x})^2}}
\]
The point estimate is \( \hat{y} \) and the margin of error is \( E \). The probability that the prediction interval contains \( y \) is \( c \).

Prediction Intervals

Construct a Prediction Interval for \( y \) for a Specific Value of \( x \)

In Words

1. Identify the number of ordered pairs in the data set \( n \) and the degrees of freedom.
\( \text{d.f.} = n - 2 \)
2. Use the regression equation and the given \( x \)-value to find the point estimate \( \hat{y} \).
\( \hat{y} = mx_i + b \)
3. Find the critical value \( t_{\alpha/2} \) that corresponds to the given level of confidence \( c \). Use Table 5 in Appendix B.

Prediction Intervals

Construct a Prediction Interval for \( y \) for a Specific Value of \( x \)

In Words

4. Find the standard error of estimate \( s_e \).
\[
s_e = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n - 2}}
\]
5. Find the margin of error \( E \).
\[
E = t_{\alpha/2} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x_i^2 - \sum (x_i - \bar{x})^2}}
\]
6. Find the left and right endpoints and form the prediction interval.
Left endpoint: \( \hat{y} - E \) Right endpoint: \( \hat{y} + E \)
Interval: \( \hat{y} - E < y < \hat{y} + E \)
Prediction Intervals

Example:  
The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

<table>
<thead>
<tr>
<th>Hours, $x$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>7</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test score, $y$</td>
<td>96</td>
<td>85</td>
<td>82</td>
<td>74</td>
<td>95</td>
<td>68</td>
<td>76</td>
<td>84</td>
<td>58</td>
<td>65</td>
</tr>
</tbody>
</table>

\[
\hat{y} = -4.07x + 93.97 \quad x = 8.11
\]

Construct a 95% prediction interval for the test scores when 4 hours of TV are watched.

Example continued:  
Construct a 95% prediction interval for the test scores when the number of hours of TV watched is 4.

There are $n - 2 = 12 - 2 = 10$ degrees of freedom.

The point estimate is

\[
\hat{y} = -4.07x + 93.97 = -4.07(4) + 93.97 = 77.69.
\]

The critical value $t_{0.025} = 2.228$, and $s_x = 8.11$.

\[
\hat{y} - E < y < \hat{y} + E
\]

\[
77.69 - 8.11 = 69.58 \quad 77.69 + 8.11 = 85.8
\]

You can be 95% confident that when a student watches 4 hours of TV over the weekend, the student’s test grade will be between 69.58 and 85.8.

§ 9.4

Multiple Regression
Multiple Regression Equation

In many instances, a better prediction can be found for a dependent (response) variable by using more than one independent (explanatory) variable.

For example, a more accurate prediction of Monday’s test grade from the previous section might be made by considering the number of other classes a student is taking as well as the student’s previous knowledge of the test material.

A multiple regression equation has the form

\[ \hat{y} = b + m_1x_1 + m_2x_2 + \ldots + m_kx_k \]

where \( x_1, x_2, \ldots, x_k \) are independent variables, \( b \) is the y-intercept, and \( y \) is the dependent variable.

* Because the mathematics associated with this concept is complicated, technology is generally used to calculate the multiple regression equation.

Predicting \( y \)-Values

After finding the equation of the multiple regression line, you can use the equation to predict \( y \)-values over the range of the data.

Example:
The following multiple regression equation can be used to predict the annual U.S. rice yield (in pounds).

\[ \hat{y} = 859 + 5.76x_1 + 3.82x_2 \]

where \( x_1 \) is the number of acres planted (in thousands), and \( x_2 \) is the number of acres harvested (in thousands). (Source: U.S. National Agricultural Statistics Service)

a.) Predict the annual rice yield when \( x_1 = 2758 \), and \( x_2 = 2714 \).  

b.) Predict the annual rice yield when \( x_1 = 3581 \), and \( x_2 = 3021 \).  

Example continued:

a.) \[ \hat{y} = 859 + 5.76x_1 + 3.82x_2 \]

\[ = 859 + 5.76(2758) + 3.82(2714) \]

\[ = 859 + 15,967.28 + 10,392.78 \]

\[ = 27,112.56 \]

The predicted annual rice yield is 27,112.56 pounds.

b.) \[ \hat{y} = 859 + 5.76x_1 + 3.82x_2 \]

\[ = 859 + 5.76(3581) + 3.82(3021) \]

\[ = 859 + 33,025.78 \]

\[ = 33,025.78 \]

The predicted annual rice yield is 33,025.78 pounds.