Chapter 4

Correlation and Linear Regression

We look at scatter diagrams, linear correlation and regression for paired (bivariate) quantitative data sets.

4.1 Looking at Scatterplots

Scatter diagram is graph of paired sampled data.

Exercise 4.1 (Looking at Scatterplots)

1. Scatter Diagram: Reading Ability Versus Brightness.

<table>
<thead>
<tr>
<th>brightness, x</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>ability to read, y</td>
<td>70</td>
<td>70</td>
<td>75</td>
<td>88</td>
<td>91</td>
<td>94</td>
<td>100</td>
<td>92</td>
<td>90</td>
<td>85</td>
</tr>
</tbody>
</table>

Figure 4.1 (Scatter Diagram, Reading Ability Versus Brightness)

Import chapter4.reading.brightness text file into R: Environment panel, Import Dataset.

data.reading <- chapter4.reading.brightness; attach(data.reading); head(data.reading)
plot(brightness,reading,pch=16,col="red",xlab="Brightness, x",ylab="Reading Ability, y")
(a) There are (circle one) 10 / 20 / 30 data points.
One particular data point is (circle one) (70, 75) / (75, 2) / (2, 70).
Data point (9,90) means (circle one)
  i. for brightness 9, reading ability is 90.
  ii. for reading ability 9, brightness is 90.

(b) Reading ability positively / not / negatively associated to brightness.
As brightness increases, reading ability (circle one) increases / decreases.

(c) Association linear / nonlinear (curved) because straight line cannot be drawn on graph where all points of scatter fall on or near line.

(d) “Reading ability” is response / explanatory variable and “brightness” is response / explanatory variable because reading ability depends on brightness, not the reverse.
Sometimes it is not so obvious which is response variable and which is explanatory variable. For example, it is not immediately clear which is explanatory variable and response variable for a scatter plot of husband’s IQ scores and wife’s IQ scores. If you were interested in knowing husband’s IQ score, given the wife’s IQ score, say, then wives’s IQ score would be explanatory variable and husband’s IQ score would be response variable.

(e) Scatter diagrams drawn for quantitative data, not qualitative data because (circle one or more)
  i. qualitative data has no order,
  ii. distance between qualitative data points is not meaningful.

(f) Another ten individuals sampled gives same / different scatter plot. Data here is a sample / population. Data here is observed / known.

2. Scatter Diagram: Grain Yield (tons) versus Distance From Water (feet).

<table>
<thead>
<tr>
<th>dist, x</th>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>45</th>
<th>50</th>
<th>70</th>
<th>80</th>
<th>100</th>
<th>120</th>
<th>140</th>
<th>160</th>
<th>170</th>
<th>190</th>
</tr>
</thead>
<tbody>
<tr>
<td>yield, y</td>
<td>500</td>
<td>590</td>
<td>410</td>
<td>470</td>
<td>450</td>
<td>480</td>
<td>510</td>
<td>450</td>
<td>360</td>
<td>400</td>
<td>300</td>
<td>410</td>
<td>280</td>
<td>350</td>
</tr>
</tbody>
</table>

Figure 4.2 (Scatter Diagram, Grain Yield Versus Distance from Water)
Section 1. Looking at Scatterplots (lecture notes 4)

Import chapter4.grain.water text file into R: Environment panel, Import Dataset.

```r
data.grain <- chapter4.grain.water; attach(data.grain); head(data.grain)
plot(distance,yield,pch=16,col="red",xlab="Distance from Water, x",ylab="Grain Yield, y")
```

(a) Scatter diagram has **pattern** / **no pattern** (randomly scattered) with
(choose one) **positive** / **negative** association,
which is (choose one) **linear** / **nonlinear**, that is a
(choose one) **weak** / **moderate** / **strong** (non)linear relationship,
where grain yield is (choose one) **response** / **explanatory** variable.

(b) Review. Second random sample would be **same** / **different** scatter plot of
(distance, yield) points. Any statistics calculated from second plot would
be **same** / **different** from statistics calculated from first plot.

3. Scatter Diagram: Pizza Sales ($1000s) versus Student Number (1000s).

<table>
<thead>
<tr>
<th>student number, x</th>
<th>pizza sales, y</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>58</td>
</tr>
<tr>
<td>6</td>
<td>105</td>
</tr>
<tr>
<td>8</td>
<td>88</td>
</tr>
<tr>
<td>8</td>
<td>118</td>
</tr>
<tr>
<td>12</td>
<td>117</td>
</tr>
<tr>
<td>16</td>
<td>137</td>
</tr>
<tr>
<td>20</td>
<td>157</td>
</tr>
<tr>
<td>20</td>
<td>169</td>
</tr>
<tr>
<td>22</td>
<td>149</td>
</tr>
<tr>
<td>26</td>
<td>202</td>
</tr>
</tbody>
</table>

Import chapter4.pizza.students text file into R: Environment panel, Import Dataset.

```r
data.pizza <- chapter4.pizza.students; attach(data.pizza); head(data.pizza)
plot(students,sales,pch=16,col="red",xlab="Number of Students, x (1000s)",ylab="Pizza Sales, y ($1000)")
```

Scatter diagram has **pattern** / **no pattern** (randomly scattered) with
(choose one) **positive** / **negative** association,
which is (choose one) **linear** / **nonlinear**, that is a
(choose one) **weak** / **moderate** / **strong** (non)linear relationship,
where student number is (choose one) **response** / **explanatory** variable.

4. More Scatter Diagrams

![Figure 4.3 (More Scatter Diagrams)](image)

Describe each scatter plot.

(a) Scatter diagram (a) has **pattern** / **no pattern** (randomly scattered).
(b) Scatter diagram (b) has pattern / no pattern (randomly scattered) with (choose one) positive / negative association, which is (choose one) linear / nonlinear, that is a (choose one) weak / moderate / strong (non)linear relationship.

(c) Scatter diagram (c) has pattern / no pattern (randomly scattered) with (choose one) positive / negative association, which is (choose one) linear / nonlinear, that is a (choose one) weak / moderate / strong (non)linear relationship.

4.2 Assigning Roles to Variables in Scatterplots

Material covered in the previous section.

4.3 Understanding Correlation

Linear correlation is a measure of linearity of a scatter plot, calculated using:

\[ r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{\sum z_x z_y}{n - 1} = \frac{\sum (x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y} \]

**Exercise 4.3 (Understanding Correlation)**

1. **Linear Correlation Coefficient: Using R.**

Linear correlation coefficient statistic, \( r \), measures linearity of scatter diagram. The larger \(|r|\), the closer \( r \) is to \( \pm 1 \), the more linear the scatterplot.

(a) **Reading ability versus brightness**

<table>
<thead>
<tr>
<th>brightness, ( x )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>reading ability, ( y )</td>
<td>70</td>
<td>70</td>
<td>75</td>
<td>88</td>
<td>91</td>
<td>94</td>
<td>100</td>
<td>92</td>
<td>90</td>
<td>85</td>
</tr>
</tbody>
</table>

In this case, \( r \approx \) (circle one) 0.704 / 0.723 / 0.734.

`attach(data.reading)`
`cor(reading,brightness)`

So, association between reading ability and brightness is (circle one) positive strong linear

(b) **Grain yield versus distance from water**

<table>
<thead>
<tr>
<th>dist, ( x )</th>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>45</th>
<th>50</th>
<th>70</th>
<th>80</th>
<th>100</th>
<th>120</th>
<th>140</th>
<th>160</th>
<th>170</th>
<th>190</th>
</tr>
</thead>
<tbody>
<tr>
<td>yield, ( y )</td>
<td>500</td>
<td>590</td>
<td>410</td>
<td>470</td>
<td>450</td>
<td>480</td>
<td>510</td>
<td>450</td>
<td>360</td>
<td>400</td>
<td>390</td>
<td>410</td>
<td>280</td>
<td>350</td>
</tr>
</tbody>
</table>
In this case, $r \approx (\text{circle one}) -0.724 / -0.785 / -0.950$.

So, association between grain yield and distance from water is (circle one)
- positive strong linear
- negative moderate linear
- positive moderate linear

(c) Annual pizza sales versus student number

<table>
<thead>
<tr>
<th>student number, $x$</th>
<th>2 6 8 8 12 16 20 20 22 26</th>
</tr>
</thead>
<tbody>
<tr>
<td>pizza sales, $y$</td>
<td>58 105 88 118 117 137 157 169 149 202</td>
</tr>
</tbody>
</table>

In this case, $r \approx (\text{circle one}) 0.724 / 0.843 / 0.950$.

So, association between pizza sales and student number is (circle one)
- positive strong linear
- negative moderate linear
- positive moderate linear

2. Linear correlation coefficient: understanding.

![Scatter Diagrams and Possible Correlation Coefficients](image)

Figure 4.4 (Scatter Diagrams and Possible Correlation Coefficients)

Match correlation coefficients with scatter plots.

(a) scatter diagram (a): $r = -0.7 / r = 0 / r = 0.3$

(b) scatter diagram (b): $r = -0.7 / r = 0.1 / r = 1$
(c) scatter diagram (c): $r = -0.7$ / $r = 0$ / $r = 0.7$

(d) scatter diagram (d): $r = -0.7$ / $r = 0$ / $r = 0.7$

When $r \neq 0$, $x$ and $y$ are linearly related to one another. If $r = 0$, $x$ and $y$ are nonlinearly related to one another, which often means diagram (a) or sometimes means diagram (d) where positive and negative associated data points cancel one another out. Always show scatter diagram with correlation $r$.

3. Linear Correlation Coefficient: Properties (Reading Ability Versus Brightness).

<table>
<thead>
<tr>
<th>brightness, $x$</th>
<th>1 2 3 4 5 6 7 8 9 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>reading ability, $y$</td>
<td>70 70 75 88 91 94 100 92 90 85</td>
</tr>
</tbody>
</table>

(a) As brightness increases, reading ability increases / decreases because $r \approx 0.704$ is positive.

(b) The more positive $r$ is (the closer $r$ is to 1), the (circle one)
   i. more linear the scatter plot.
   ii. steeper the slope of the scatter plot.
   iii. larger the reading ability value.
   iv. brighter the brightness.

(c) If 0.5 is added to all $x$ values, 1 becomes 1.5, 2 becomes 2.5 and so on, $r$ changes from 0.704 to 0.892.

(d) The $r$–value calculated after accidentally reversing point (1,70) with point (70,1) equals / does not equal $r$ value before reversing this point.

(e) True / False The $r$–value remains same whether or not brightness is measured in foot candles or lumens.

(f) Ability to read and brightness are mistakenly reversed:

<table>
<thead>
<tr>
<th>ability to read, $y$</th>
<th>70 70 75 88 91 94 100 92 90 85</th>
</tr>
</thead>
<tbody>
<tr>
<td>brightness, $x$</td>
<td>1 2 3 4 5 6 7 8 9 10</td>
</tr>
</tbody>
</table>

The $r$ value (circle one) remains unchanged / changes.

(g) Compare original scatter diagram with one without outlier (7,130).

<table>
<thead>
<tr>
<th>brightness, $x$</th>
<th>1 2 3 4 5 6 7 8 9 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>ability to read, $y$</td>
<td>70 70 75 88 91 94 100 92 90 85</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>brightness, $x$</th>
<th>1 2 3 4 5 6 7 8 9 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>ability to read, $y$</td>
<td>70 70 75 88 91 94 130 92 90 85</td>
</tr>
</tbody>
</table>
The correlation coefficient is (circle one) resistant / sensitive to outliers.

(h) Identify statistical items in example.

<table>
<thead>
<tr>
<th>terms</th>
<th>reading/lighting example</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) population</td>
<td>(a) all reading/brightness levels</td>
</tr>
<tr>
<td>(b) sample</td>
<td>(b) correlation of 10 reading/brightness levels, $r$</td>
</tr>
<tr>
<td>(c) statistic</td>
<td>(c) correlation of all reading/brightness levels, $\rho$</td>
</tr>
<tr>
<td>(d) parameter</td>
<td>(d) 10 reading/brightness levels</td>
</tr>
</tbody>
</table>

Notice population parameter for linear correlation coefficient is $\rho$.

(i) Brightness increase causes / is associated with reading ability increase.

4. Linear Correlation Coefficient: Formulas.

<table>
<thead>
<tr>
<th>circumference, $x$</th>
<th>2.1</th>
<th>1.7</th>
<th>1.1</th>
<th>1.5</th>
<th>2.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>height, $y$</td>
<td>40</td>
<td>37</td>
<td>35</td>
<td>36</td>
<td>42</td>
</tr>
<tr>
<td>$z_x = \frac{x - \bar{x}}{s_x}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$z_y = \frac{y - \bar{y}}{s_y}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Since $n = 5 / 10 / 15$

$\bar{x} = 1.82 / 1.83 / 1.84$

$s_x = 0.57 / 0.61 / 0.67$

$\bar{y} = 28 / 38 / 48$

$s_y = 2.92 / 3.02 / 3.12$

and so $r = \frac{\sum z_x z_y}{n-1} \approx 0.560 / 0.621 / 0.984$

and $r^2 = 0.314 / 0.723 / 0.968$. 
4.4 Lurking Variables and Causation

In general, although two variables may be highly correlated, this does not necessarily mean that an increase (or decrease) in one variable causes an increase or decrease in other variable. It may be that a third variable, a lurking variable, which may explain the correlation between the two variables.

Exercise 4.4 (Lurking Variables and Causation)

1. Linear correlation coefficient: correlation, not causation (chimpanzees).
   In a study of chimpanzees it was found there was a positive correlation between tallness and intelligence. Circle true or false:
   
   True / False Taller chimpanzees were also more intelligent, on average.
   
   True / False Intelligent chimpanzees were also taller, on average.
   
   True / False The data show that tallness causes intelligence.
   
   True / False The data show that intelligence causes tallness.

   It may be that the chimpanzees were bred for both intelligence and tallness: breeding is a lurking variable which may explain the correlation between intelligence and tallness.

2. Association or causation: reading ability versus brightness.

   ![Figure 4.6 (Reading ability versus brightness)](image-url)
Participants were asked to read a text aloud as brightness levels were increased. Reading ability score was based on how well the participants were able to read the text. Experiments noticed, in particular, reading ability changed after brightness levels changed. Recall, linear correlation between reading ability and brightness is $r \approx 0.7$. Assume this correlation has been confirmed many times in a number of studies conducted by a number of different people.

(a) *Is association strong?*
Since linear correlation $r \approx 0.7 < 0.8$, association is **strong** / **not strong**
Since association (measured by correlation, $r < 0.8$) is **moderate**, this indicates there is just an association, **not** causation relationship between brightness and reading ability.

(b) *Is association consistent?*
Since *many* studies done by *different* people have shown $r \approx 0.7$, association is **consistent** / **not consistent**
Since association is consistent, this indicates that brightness is not merely associated with reading ability, but that a change in brightness *causes* a change in reading ability.

(c) *Is higher dose associated with stronger response?*
Brighter is associated with **lower** / **higher** reading ability
Since (within reason, for brightness levels between 0 and 8) brighter is associated with higher reading ability, this indicates brightness is not merely associated with reading ability, but that a change in brightness *causes* a change in reading ability.

(d) *Does (alleged) cause precede response in time?*
Change in brightness **precedes** / **unrelated** to change in reading ability
Yes, reading ability changed *after* brightness changed.

(e) *Is (alleged) cause plausible?*
Brightness change **plausibly** / **implausibly** causes reading ability change
Yes, it seems plausible that changes in brightness would cause changes in reading ability, but not vice-versa.

(f) Relationship between reading ability and brightness is (choose one) **association** / **causation**
It is most likely causation because four of the five criterion indicate this.

3. **Association or causation: grain yield versus distance from water.**
Recall, linear correlation between grain yield and distance from water is $r \approx -0.8$. Assume this correlation appears in only one study.

(a) *Is association strong?*

strong / not strong

Since association (measured by correlation, $r \approx -0.8$) is (negatively) strong, this indicates there is a causation relationship.

(b) *Is association consistent?*

consistent / not consistent

Only one study has been done.

(c) *Is higher dose associated with stronger response?*

Further from water associated with lower / higher yield

Grain yield clearly lower far away from water.

(d) *Does (alleged) cause precede response in time?*

Change in distance precedes / unrelated grain yield

Not clear from study, so correct answer is “unknown”.

(e) *Is (alleged) cause plausible?*

Distance change plausibly / implausibly causes grain yield change

Yes, it seems plausible that change in distance from water would cause change in grain yield.

(f) Relationship between grain yield and distance is (choose one)

association / causation

Hard to say if causation because only three of the five criterion indicate causation, although it seems reasonable to think there is causation here, but, really, more study is required.

4. *Response often influenced by more than one variable.*

(a) *Reading ability versus brightness.*
Different types of observed association are given in figure where, notice, response is often influenced by more than one variable. Dotted lines are observed association; solid arrows are possible cause-and-effect links.

A. brightness change causes / does not cause reading ability change previous analysis seems to indicate causation relationship between (explanatory variable) brightness and (response variable) reading ability

B. eyesight confounded / not confounded with brightness (explanatory variable) brightness and (lurking variable) eyesight are confounded because they are positively associated with one another and with (response) reading ability and it is not clear what contributions each has to (response) reading ability.

(b) Grain yield versus distance from water:

A. distance from water change causes / does not cause yield change not clear, but previous analysis indicates possibility of causation relationship between (explanatory variable) distance from water and (response variable) grain yield
B. soil type confounded / not confounded with water distance
(explanatory variable) distance from water and (lurking variable) soil type are confounded because they are positively associated with one another and with (response) grain yield and it is not clear what contributions each has to (response) grain yield.

B’ humidity confounded / not confounded with water distance
(explanatory variable) distance from water and (lurking variable) humidity are confounded because they are positively associated with one another and with (response) grain yield and it is not clear what contributions each has to (response) grain yield.

4.5 The Linear Model
Material is covered in next section.

4.6 Correlation and the Line
We fit a least-squares regression line to the scatterplot,

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

where \( b_1 \) is slope and \( b_0 \) is \( y \)-intercept, where

\[ b_1 = r \cdot \frac{s_y}{s_x}, \quad b_0 = \bar{y} - b_1 \bar{x} \]

where \( s_x \), \( s_y \) and \( r \) are standard deviation of \( x \), standard deviation of \( y \) and correlation respectively. If the predictor, \( x \), and response variable, \( \hat{y} \), are standardized \( z_x = \frac{x - \bar{x}}{s_x} \) and \( z_y = \frac{y - \bar{y}}{s_y} \), then \( s_{z_x} = s_{z_y} = 1 \) and \( \bar{z}_x = \bar{z}_y = 0 \), so the regression line becomes

\[ \hat{z}_y = r z_x \]

Exercise 4.6 (Correlation and the Line)
1. Reading ability versus brightness.
   Create scatter diagram, calculate least-squares regression line and superimpose line on scatter diagram.

<table>
<thead>
<tr>
<th>brightness, ( x )</th>
<th>1 2 3 4 5 6 7 8 9 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>reading ability, ( y )</td>
<td>70 70 75 88 91 94 100 92 90 85</td>
</tr>
</tbody>
</table>
Figure 4.10 (Least–squares Line, reading ability versus brightness)

(a) Least squares line output from R is

\[ \hat{y} = 72.2 + 2.418x \]

(b) Least-squares regression line. Choose two.

\[ \hat{y} = 47.04x + 2.944 \]

(c) Slope and \( y \)-intercept of least-squares regression line, \( \hat{y} = 2.418x + 72.2 \).

Slope is \( b_1 = (\text{circle one}) \ 72.2 / 2.418 \).

Slope, \( b_1 = 2.418 \), means, on average, reading ability increases 2.418 units for an increase of one unit of brightness.

The \( y \)-intercept is \( b_0 = (\text{circle one}) \ 72.2 / 2.418 \).

The \( y \)-intercept, \( b_0 = 72.2 \), means average reading ability is 72.2, if brightness is zero.
(d) **Prediction.**
At brightness $x = 6.5$, predicted reading ability is
\[ \hat{y} \approx 2.418x + 72.2 = 2.418(6.5) + 72.2 \approx 84.9 / 85.5 / 87.9. \]
\[
predict(model.reading,list(brightness=6.5))
\]

![Figure 4.11 (Least–Squares Line: Prediction)](image)

(e) **More Prediction.**
At brightness $x = 5.5$, $\hat{y} \approx 2.418(5.5) + 72.2 \approx 84.9 / 85.5 / 87.6$.
At brightness $x = 7.5$, $\hat{y} \approx 2.418(7.5) + 72.2 \approx 84.9 / 89.5 / 90.4$.
\[
predict(model.reading,list(brightness=c(5.5,7.5)))
\]

(f) **Residual.**
At brightness $x = 7$, $\hat{y} \approx 2.418(7) + 72.2 \approx 87.9 / 89.1 / 120.6$.
*Observed* value, $y = 100$ compared to predicted $\hat{y} = 89.1$; difference between two is *residual:*
\[ y - \hat{y} = 100 - 89.1 = (\text{circle one}) 9.2 / 10.9 / 12.6. \]
\[
residuals(model.reading)
\]

![Figure 4.12 (Least–Squares Line: Residual)](image)

Residual for $x = 7$ is vertical distance between observed $(7,100)$ and predicted $(7, 89.1)$ on least-squares regression line.
(g) More Residuals.
At brightness $x = 8$, $y - \hat{y} \approx 92 - 91.5 = -0.5 / 0.5 / 1.5$.
At brightness $x = 3$, $y - \hat{y} \approx 75 - 79.5 = -4.5 / -4.5 / -1.5$.
There are (circle one) 1 / 5 / 10 residuals on scatter diagram.

2. Grain yield (tons) versus distance from water (feet)

<table>
<thead>
<tr>
<th>dist, x</th>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>45</th>
<th>50</th>
<th>70</th>
<th>80</th>
<th>100</th>
<th>120</th>
<th>140</th>
<th>160</th>
<th>170</th>
<th>190</th>
</tr>
</thead>
<tbody>
<tr>
<td>yield, y</td>
<td>500</td>
<td>590</td>
<td>410</td>
<td>470</td>
<td>450</td>
<td>480</td>
<td>510</td>
<td>450</td>
<td>360</td>
<td>400</td>
<td>300</td>
<td>410</td>
<td>280</td>
<td>350</td>
</tr>
</tbody>
</table>

\[ \hat{y} = -1.06x + 515.45 \]

Figure 4.13 (Least–squares regression, grain yield versus distance)

(a) The least–squares line is (circle one)
\[ \hat{y} = 515.45 - 1.56x \]
\[ \hat{y} = 535.45 - 2.56x \]
\[ \hat{y} = 515.45 - 1.06x. \]

(b) Slope and y-intercept.
Slope is $b_1$ = (circle one) 515.45 / −1.06.
Slope, $b_1 = -1.06$, means, on average, grain yield decreases 1.06 tons for an increase of one foot away from water.

The $y$–intercept is $b_0$ = (circle one) 515.45 / −1.06.
The $y$–intercept, $b_0 = 515.45$, means average grain yield is 515.45 at water’s edge.

(c) Prediction.
At distance $x = 100$,
\[ \hat{y} = -1.06x + 515.45 = -1.06(100) + 515.45 = 400 / 407.3 / 409.5. \]
At distance $x = 165$,
\[ \hat{y} = -1.06x + 515.45 = -1.06(165) + 515.45 = 340.5 / 367.0 / 404.8 \]
Chapter 4. Describing the Relation Between Two Variables (lecture notes 4)

predict(model.grain,list(distance=c(100,155)))

(d) Residual.
At distance $x = 100$,
$y - \hat{y} \approx 360 - 409.5 = -49.5 / -36.5 / -25.5$.
At distance $x = 140$,
$y - \hat{y} \approx 300 - 367 = -67 / -55 / -25$.

residuals(model.grain)

(e) Review. Second random sample gives same / different scatter diagram.
Statistics calculated from second plot same / different from statistics calculated from first plot. So, slope, $b_1$, and $y$-intercept, $b_0$, and predicted values, $\hat{y} = b_1 x + b_0$, all statistics / parameters.

(f) Identify statistical items in example.

<table>
<thead>
<tr>
<th>terms</th>
<th>grain yield/water example</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) popluation</td>
<td>(a) all (yield, distance) amounts</td>
</tr>
<tr>
<td>(b) sample</td>
<td>(b) $b_0, b_1, \hat{y}$</td>
</tr>
<tr>
<td>(c) statistics</td>
<td>(c) $\alpha, \beta, \mu_x$</td>
</tr>
<tr>
<td>(d) parameters</td>
<td>(d) 14 (yield, distance) amounts</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>circumference, $x$</th>
<th>2.1</th>
<th>1.7</th>
<th>1.1</th>
<th>1.5</th>
<th>2.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>height, $y$</td>
<td>40</td>
<td>37</td>
<td>35</td>
<td>36</td>
<td>42</td>
</tr>
<tr>
<td>$z_x = \frac{x - \bar{x}}{s_x}$</td>
<td>0.46</td>
<td>-0.20</td>
<td>-1.18</td>
<td>-0.52</td>
<td>1.44</td>
</tr>
<tr>
<td>$z_y = \frac{y - \bar{y}}{s_y}$</td>
<td>0.69</td>
<td>-0.34</td>
<td>-1.03</td>
<td>-0.69</td>
<td>1.37</td>
</tr>
</tbody>
</table>

where $n = 5$, $\bar{x} = 1.82$, $s_x = 0.61$, $\bar{y} = 38$, $s_y = 2.92$ and $r = 0.984$.

(a) If a tree is $z_x = 1.5$ standard deviations above mean circumference, how many standard deviations above mean height, $z_y$, does the standardized regression predict?

$\hat{z}_y = r \cdot z_x = 0.984 \times 1.5 = (\text{choose one}) 1.476 / 1.500 / 1.678$.

(b) Since $r = 0.984$, $z_y$ is less / same / more standard deviations above the $\hat{y}$ than $z_x$ is above $\bar{x}$. This is an example of the regression to the mean.

(c) If a tree is $z_x = 1.5$ standard deviations above mean circumference, what is the predicted height? (Recall $\hat{z}_y = \frac{z_y}{s_y}$, so $y = s_y \cdot z_y + \bar{y}$.)

So $\hat{y} = s_y \cdot \hat{z}_y + \bar{y} = 2.92 \cdot 1.476 + 38 = 42.3 / 43.8 / 44.2$. 
4.7 Regression to the Mean

Material is covered in previous section.

4.8 Checking the Model

Use scatterplot and residual plot to assess the “fit” of least-squares line to data:

- **Linearity assumption/condition**: whether there are any patterns, other than linearity, in residuals,
- **Equal spread condition**: whether variance of residuals against exploratory variable is constant or not,
- **Outlier condition**: whether there are outliers,
- **Coefficient of determination, \( R^2 \)**: a measure of proportion of scatter explained by least-squares regression

**Exercise 4.8 (Checking the Model)**


<table>
<thead>
<tr>
<th>brightness, x</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>ability to read, y</td>
<td>70</td>
<td>70</td>
<td>75</td>
<td>88</td>
<td>91</td>
<td>94</td>
<td>100</td>
<td>92</td>
<td>90</td>
<td>85</td>
</tr>
<tr>
<td>predicted, ( \hat{y} )</td>
<td>74.6</td>
<td>77.0</td>
<td>79.5</td>
<td>81.9</td>
<td>84.3</td>
<td>86.7</td>
<td>89.1</td>
<td>91.5</td>
<td>94.0</td>
<td>96.4</td>
</tr>
<tr>
<td>residual, ( y - \hat{y} )</td>
<td>4.6</td>
<td>7.0</td>
<td>4.5</td>
<td>6.1</td>
<td>6.7</td>
<td>7.3</td>
<td>10.9</td>
<td>0.5</td>
<td>4.0</td>
<td>8.6</td>
</tr>
</tbody>
</table>

![Figure 4.14 (Scatterplot, residual plot, boxplot of residuals)](image-url)
Chapter 4. Describing the Relation Between Two Variables (lecture notes 4)

```r
data.reading <- chapter4.reading.brightness; attach(data.reading); head(data.reading)
par(mfrow=c(1,3))
plot(brightness,reading,pch=16,col="red",xlab="Brightness, x",ylab="Reading Ability, y")
model.reading <- lm(reading~brightness); model.reading
abline(model.reading,col="black")
plot(brightness,residuals(model.reading),pch=16,col="red",xlab="Brightness, x",ylab="Residuals")
abline(h=0,lty=2,col="black")
boxplot(residuals(model.reading),col="red",ylab="Residuals")
par(mfrow=c(1,1))
```

(a) **Linearity assumption/condition?**

According to either scatter diagram or residual plot, there is a / is no pattern (around line): points are curved.

(b) **Equal spread condition?**

According to residual plot, residuals vary -10 and 10 over entire range of brightness; that is, data variance is constant / variable.

(c) **Outliers?**

According to boxplot of residuals, there are / are no outliers

No outliers “•”s in boxplot.

(d) **Coefficient of determination, \( R^2 \).**

Since \( R^2 = (\text{choose one}) \ 0.496 / 0.523 / 0.539 \), least-squares line explains 49.6% of variability in reading ability.

\[
\text{cor(data.reading)}^2
\]

2. **Grain yield versus distance from water.** Consider scatter diagram and residual plot. Do not use StatCrunch, use plots and regressions given below instead.

```r
data.grain <- chapter4.grain.water; attach(data.grain); head(data.grain)
par(mfrow=c(1,3))
plot(distance,yield,pch=16,col="red",xlab="Distance from water, x",ylab="Yield, y")
model.grain <- lm(yield~distance); model.grain
abline(model.grain,col="black")
plot(distance,residuals(model.grain),pch=16,col="red",xlab="Yield, x",ylab="Residuals")
abline(h=0,lty=2,col="black")
boxplot(residuals(model.grain),col="red",ylab="Residuals")
par(mfrow=c(1,1))
```

Figure 4.15 (Scatterplot, residual plot, boxplot of residuals)
(a) **Linearity assumption/condition?**
According to either scatter diagram or residual plot, there is a / is no pattern (around line).

(b) **Equal spread condition?**
According to residual plot, residuals vary -85 and 85 over entire range of distances; that is, data variance is constant / variable.

(c) **Outliers?**
According to boxplot of residuals, there are / are no outliers.

(d) **Coefficient of determination** $R^2$.
Since $R^2 = 0.496 / -0.616 / 0.616$, least-squares line explains 61.0% of variability in grain yield.

3. **Understanding coefficient of determination, $R^2$.**
   True / False

![Figure 4.16 (Understanding Coefficient of Determination)](image)

If the average of the $y$ variable, $\bar{y}$, is a kind of baseline and since

$$
\frac{(y - \bar{y})}{\text{total deviation}} = \frac{(\hat{y} - \bar{y})}{\text{explained deviation}} + \frac{(y - \hat{y})}{\text{unexplained deviation}}
$$

then taking sum of squares over all data points

$$
\sum(y - \bar{y})^2 = \sum(\hat{y} - \bar{y})^2 + \sum(y - \hat{y})^2
$$

and so coefficient of determination is a measure of proportion of scatter diagram explained by least-squares line.

$$
R^2 = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2} = \frac{\text{explained variation}}{\text{total variation}}
$$
4.9 Variation in the Model and $R^2$

Material covered in previous section.

4.10 Reality Check: Is the Regression Reasonable?

Material covered in previous section.

4.11 Nonlinear Relationships

When fitting nonlinear scatterplots, two possible solutions are to fit a nonlinear model to the data or fit a linear model to transformed-to-linear data.

Exercise 4.11 (Nonlinear Relationships)

Consider table of residuals, scatter diagram and residual plot.

| brightness, x | 1 2 3 4 5 6 7 8 9 10 |
| ability to read, y | 70 70 75 88 91 94 100 92 90 85 |

![Figure 4.17 (Scatterplot, nonlinear fit, transformed data)]

1. A linear model (black line) does / does not fit the reading data in (a).

2. The nonlinear model (curved line) fits the data in (b) better / worse than the linear model fits the data in (a).

3. The linear model fits the transformed data in (c) better / worse than the linear model fits the original (untransformed) data in (a).
4. The statistical analyses for either quadratic fit case in (b) or transformed data case in (c) are easier / harder to interpret than the linear fit to the untransformed data in (a).

data.reading <- chapter4.reading.brightness; attach(data.reading); head(data.reading)
par(mfrow=c(1,3))
model.reading1 <- lm(reading~brightness) # original data, fit
plot(brightness,reading,pch=16,col="red",xlab="Brightness, x",ylab="Reading Ability, y",main="Linear fit")
abline(model.reading1,col="black")
model.reading2 <- lm(reading~brightness + I(brightness^2)) # original data, quadratic fit
plot(brightness,reading,pch=16,col="red",xlab="Brightness, x",ylab="Reading Ability, y",main="Quadratic fit")
x <- 1:10; y <- predict(model.reading2,list(brightness=x)); lines(x,y,col="black")
reading.transform <- reading # leave reading data alone
brightness.transform <- 53.8 + 11.6*brightness - 0.8*brightness^2 # transform brightness data
model.reading3 <- lm(reading.transform~brightness.transform)
plot(brightness.transform,reading,pch=16,col="red",xlab="transform Brightness",ylab="Reading Ability",main="Transformed data")
cor(reading.transform,brightness.transform)
par(mfrow=c(1,1))