**Chapters 7-10: Lines in 2D (Regression and Correlation)**

- Vertical Lines
- Horizontal Lines
- Oblique lines
- Increasing/Decreasing
- Slope of a line
- Intercept
- \( Y = \alpha X + \beta \), in general.

**Math Equation for the Line?**

- Draw the following lines:
  - \( Y = 2X + 1 \)
  - \( Y = -3X - 5 \)
- Line through \((X_1, Y_1)\) and \((X_2, Y_2)\).
- \[ \frac{Y-Y_1}{Y_2-Y_1} = \frac{X-X_1}{X_2-X_1}. \]

**Approaches for modeling data relationships**

**Regression and Correlation**

- There are random and nonrandom variables
- Correlation applies if both variables \((X/Y)\) are random (e.g., We saw a previous example, systolic vs. diastolic blood pressure \( \text{SISVOL/DIAVOL} \) and are treated symmetrically.
- Regression applies in the case when you want to single out one of the variables (response variable, \( Y \)) and use the other variable as predictor (explanatory variable, \( X \)), which explains the behavior of the response variable, \( Y \).

**Causal relationship?**

- infant death rate (per 1,000) in 14 countries

Strong evidence (linear pattern) of death rate increase with increasing level of breastfeeding (BF)?

Naïve conclusion: Breastfeeding is bad? But high rates of BF is associated with lower access to H2O.

**Regression relationship = trend + residual scatter**

- Regression is a way of studying relationships between variables (random/nonrandom) for predicting or explaining behavior of 1 variable (response) in terms of others (explanatory variables or predictors).
**Trend (does not have to be linear) + Scatter (could be of any type/distribution)**

(b) Oxygen uptake

<table>
<thead>
<tr>
<th>Ventilation (l/min)</th>
<th>Oxygen uptake (ml/min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>2000</td>
</tr>
<tr>
<td>2000</td>
<td>3000</td>
</tr>
<tr>
<td>3000</td>
<td>4000</td>
</tr>
</tbody>
</table>

**Trend + scatter (fetus liver length in mm)**

Change of scatter with age

(c) Liver lengths

<table>
<thead>
<tr>
<th>Gestational age (wk)</th>
<th>Liver length (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>150</td>
</tr>
<tr>
<td>20</td>
<td>200</td>
</tr>
<tr>
<td>25</td>
<td>250</td>
</tr>
<tr>
<td>30</td>
<td>300</td>
</tr>
<tr>
<td>35</td>
<td>350</td>
</tr>
<tr>
<td>40</td>
<td>400</td>
</tr>
</tbody>
</table>

**Trend + scatter**

Dotted curves (confidence intervals) represent the extent of the scatter.

(a) Scatter plot

Displacement versus weight for 74 models of automobile.

(b) With trend plus scatter

**Looking vertically**

Flutter line gives better prediction, since it approx. goes through the middle of the Y-range, for each fixed x-value (vertical line)

(a) Which line?

(b) Flutter line gives better predictions.

**Outliers – odd, atypical, observations (errors, B, or real data, A)**

Scatter plot from the heart attack data.

**A weak relationship**

58 abused children are rated (by non-abusive parents and teachers) on a psychological disturbance measure.

How do we quantify weak vs. strong relationship?

<table>
<thead>
<tr>
<th>Parent's rating</th>
<th>Teacher's rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td>80</td>
<td>80</td>
</tr>
</tbody>
</table>
5. When can you reliably conclude that changes in $X$ cause the changes in $Y$? (Only when controlled randomized experiments are used – levels of $X$ are randomly distributed to available experimental units, or experimental conditions need to be identical for different levels of $X$, this includes time.

Essential Points

1. What essential difference is there between the correlation and regression approaches to a relationship between two variables? (In correlation independent variables; in regression, response var depends on explanatory variable.)

2. What are the most common reasons why people fit regression models to data? (predict $Y$ or unravel reasons/causes of behavior.)

3. Can you conclude that changes in $X$ caused the changes in $Y$ seen in a scatter plot if you have data from an observational study? (No, there could be lurking variables, hidden effects/predictors, also associated with the predictor $X$, itself, e.g., time is often a lurking variable, or may be that changes in $Y$ cause changes in $X$, instead of the other way around.

Correlation Coefficient

\[ R(X, Y) = \frac{1}{N-1} \frac{N}{N} \sum_{k=1}^{N} \left( \frac{x_k - \mu_x}{\sigma_x} \right) \left( \frac{y_k - \mu_y}{\sigma_y} \right) \]

Example:

\[
\begin{array}{c|c|c|c|c|c}
\text{Gender} & \text{Height} & \text{Weight} & x - \mu_x & y - \mu_y & (x - \mu_x)(y - \mu_y) \\
\hline
1 & 175 & 80 & 4.87 & 30 & 21.62
2 & 170 & 64 & 0.67 & 1 & 0.67
3 & 180 & 67 & 1.67 & 1 & 1.67
4 & 165 & 60 & -0.67 & 1 & -0.67
5 & 175 & 70 & -1.67 & 1 & -1.67
6 & 170 & 50 & -0.67 & 1 & -0.67
\hline
\text{Total} & 950 & 332 & 0 & 216 & 216.3334 & 185.0
\end{array}
\]

\[ R(X, Y) = \frac{1}{N-1} \frac{N}{N} \sum_{k=1}^{N} \left( \frac{x_k - \mu_x}{\sigma_x} \right) \left( \frac{y_k - \mu_y}{\sigma_y} \right) \]

\[
\begin{align*}
\mu_x &= \frac{966}{6} = 161 \text{ cm}, & \mu_y &= \frac{332}{6} = 55 \text{ kg}, \\
\sigma_x &= \sqrt{\frac{216}{5}} = 6.573, & \sigma_y &= \sqrt{\frac{215.3}{5}} = 6.563
\end{align*}
\]

\[ Corr(X, Y) = R(X, Y) = 0.904 \]

A note of caution!

In observational data, strong relationships are not necessarily causal. It is virtually impossible to conclude a cause-and-effect relationship between variables using observational data!
Correlation Coefficient - Properties

Correlation measures linear association, NOT an association in a non-linear fashion. So, Corr(X, Y) could be misleading for X & Y related in a non-linear fashion.

6. If the experimenter has control of the levels of X used, how should these levels be allocated to the available experimental units?

At random! Example, testing hardness of concrete, Y, based on levels of cement, X, incorporated. Factors affecting Y: amount of H2O, ratio stone-chips to sand, drying conditions, etc. To prevent uncontrolled differences in batches of concrete in confounding our impression of cement effects, we should choose which batch (H2O levels, sand, dry-conditions) gets what amount of cement specified (y=ax+5, a=?).

8. People fit theoretical models to data for three main purposes.

a. To test the model, itself, by checking if the data is reasonably close agreement with the relationship predicted by the model.

b. Assuming the model is correct, to test if theoretically specified values of a parameter are consistent with the data (y=2x+1 vs. y=2.1x-0.9).

c. Assuming the model is correct, to estimate unknown constants in the model so that the relationship is completely specified (y=ax+5, a=?)

7. What theories can you explore using regression methods?

Prediction, explanation/causation, testing a scientific hypothesis/mathematical model:

a. Hooke’s spring law: amount of stretch in a spring, Y, is related to the applied weight X by Y=ax+b, X, a, b are spring constants.

b. Theory of gravity: force of gravity F between 2 objects is given by F = \( \alpha D^2 \), where D = distance between objects, \( \alpha \) is a constant related to the masses of the objects and \( \beta = 2 \), according to the inverse square law.

c. Economic production function: \( Q = \alpha L^\beta K^\gamma \), Q=production, L=quantity of labor, K=capital, a, b, \( \beta, \gamma \) are constants specific to the market studied.
Trend and Scatter - Computer timing data

- The major components of a regression relationship are trend and scatter around the trend.
- To investigate a trend – fit a math function to data, or smooth the data.
- Computer timing data: a mainframe computer has X users, each running jobs taking Y min time. The main CPU swaps between all tasks. Y* is the total time to finish all tasks. Both Y and Y* increase with increase of tasks/users, but how?

<table>
<thead>
<tr>
<th>X</th>
<th>Y*</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>6.6</td>
<td>9.9</td>
</tr>
<tr>
<td>50</td>
<td>14.9</td>
<td>17.8</td>
</tr>
<tr>
<td>60</td>
<td>18.4</td>
<td>18.4</td>
</tr>
<tr>
<td>45</td>
<td>12.4</td>
<td>16.5</td>
</tr>
<tr>
<td>40</td>
<td>7.9</td>
<td>11.9</td>
</tr>
<tr>
<td>10</td>
<td>0.9</td>
<td>5.5</td>
</tr>
<tr>
<td>30</td>
<td>5.5</td>
<td>11.1</td>
</tr>
<tr>
<td>20</td>
<td>2.7</td>
<td>5.5</td>
</tr>
</tbody>
</table>

X = Number of terminals: 40 50 60 45 40 10 30 20
Y* = Total Time (mins): 6.6 14.9 18.4 12.4 7.9 0.9 5.5 2.7
Y = Time Per Task (sec): 9.9 17.8 18.4 16.5 11.9 5.5 11.1 8.1

X = Number of terminals: 50 30 65 40 65 65
Y* = Total Time (mins): 12.6 6.7 23.6 9.2 20.2 21.4
Y = Time Per Task (sec): 15.1 13.3 21.8 13.8 18.6 19.8

Trend and Scatter - Computer timing data

Equation for the straight line – linear/affine function

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

\( \beta_0 \) = Intercept (the y-value at x=0)
\( \beta_1 \) = Slope of the line (rise/run), change of y for every unit of increase for x.

The quadratic curve

\[ y = \beta_0 + \beta_1 x + \beta_2 x^2 \]

\( \beta_2 \) positive \( \beta_2 \) negative

The exponential curve, \( y = a e^{bx} \)

<table>
<thead>
<tr>
<th>b positive</th>
<th>b negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>a</td>
</tr>
</tbody>
</table>

Used in population growth/decay models.
Effects of changing $x$ for different functions/curves

A straight line changes by a fixed amount with each unit change in $x$.

An exponential changes by a fixed percentage with each unit change in $x$.

Creatine kinase concentration in patient’s blood

You should not let the questions you want to ask be dictated by the tools you know how to use.

Here $Y$=creatinine kinase concentration in blood for a set of heart attack patients vs. the time, $X$.

No symmetry so $X^2$ models won’t work!

To tell whether a trend is exponential ….

check whether a plot of $\log(y)$ versus $x$ has a linear trend.

Comments

1. In statistics what are the two main approaches to summarizing trends in data? (model fitting, smoothing – done by the eye!)
2. In $y = 5x + 2$, what information do the 5 and the 2 convey? (slope, y-intercept)
3. In $y = 7 + 5x$, what change in $y$ is associated with a 1-unit increase in $x$? with a 10-unit increase? (-5; -50)
   How about for $y = 7 - 5x$, (-5; -50)

Fitting a line through the data

Choosing the “best-fitting” line

The idea of a residual or prediction error

Show the Regression-Line Simulation Applet: RegressionApplet.html

$x_i$  $y_i$  $\hat{y}_i$  $
Data point $(x_i, y_i)$

Observed $y_i$

Predicted $\hat{y}_i$

Residual $u = y_i - \hat{y}_i$

Predicted Value $\hat{y}_i$
Least squares criterion

**Least squares criterion:** Choose the values of the parameters to minimize the sum of squared prediction errors (or sum of squared residuals),

\[ \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \ldots + (y_n - \hat{y}_n)^2 \]

For each point \( P_1(x_1, y_1), P_2(x_2, y_2), \ldots, P_n(x_n, y_n) \).

\[ \hat{y} = \beta_0 + \beta_1 x \]

**The least squares line:**

**Computer timings data – linear fit**

**Adding the least squares line**

**Least-squares line:**

\[ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \]

**Computer-timings data with least-squares line.**

**Figure 12.3.2**

Two lines on the computer-timings data.


**TABLE 12.3.1 Prediction Errors**

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>y - \hat{y}</th>
<th>\hat{y} - \hat{y}</th>
<th>3 + 0.25x</th>
<th>7 + 0.15x</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>9.90</td>
<td>13.00</td>
<td>-3.10</td>
<td>13.00</td>
<td>-3.10</td>
</tr>
<tr>
<td>50</td>
<td>17.80</td>
<td>15.50</td>
<td>2.30</td>
<td>14.50</td>
<td>3.30</td>
</tr>
<tr>
<td>60</td>
<td>18.40</td>
<td>18.00</td>
<td>0.40</td>
<td>16.00</td>
<td>2.40</td>
</tr>
<tr>
<td>45</td>
<td>16.50</td>
<td>14.25</td>
<td>2.25</td>
<td>13.75</td>
<td>2.75</td>
</tr>
<tr>
<td>40</td>
<td>11.90</td>
<td>13.00</td>
<td>-1.10</td>
<td>13.00</td>
<td>-1.10</td>
</tr>
<tr>
<td>10</td>
<td>5.50</td>
<td>5.50</td>
<td>0.00</td>
<td>8.50</td>
<td>-3.00</td>
</tr>
<tr>
<td>30</td>
<td>11.00</td>
<td>10.50</td>
<td>0.50</td>
<td>11.50</td>
<td>-0.50</td>
</tr>
<tr>
<td>20</td>
<td>8.10</td>
<td>8.00</td>
<td>0.10</td>
<td>10.00</td>
<td>-1.00</td>
</tr>
<tr>
<td>50</td>
<td>15.10</td>
<td>15.50</td>
<td>-0.40</td>
<td>14.50</td>
<td>0.60</td>
</tr>
<tr>
<td>30</td>
<td>13.30</td>
<td>10.50</td>
<td>2.80</td>
<td>11.50</td>
<td>1.80</td>
</tr>
<tr>
<td>65</td>
<td>21.80</td>
<td>19.25</td>
<td>2.55</td>
<td>16.75</td>
<td>5.05</td>
</tr>
<tr>
<td>40</td>
<td>13.80</td>
<td>13.00</td>
<td>0.80</td>
<td>13.00</td>
<td>0.80</td>
</tr>
<tr>
<td>30</td>
<td>13.30</td>
<td>10.50</td>
<td>2.80</td>
<td>11.50</td>
<td>1.80</td>
</tr>
<tr>
<td>65</td>
<td>18.60</td>
<td>19.25</td>
<td>-0.65</td>
<td>16.75</td>
<td>1.85</td>
</tr>
<tr>
<td>65</td>
<td>19.80</td>
<td>19.25</td>
<td>0.55</td>
<td>16.75</td>
<td>3.05</td>
</tr>
</tbody>
</table>

Sum of squared errors 37.46 90.36

Some Minitab regression output

The regression equation is

\[ \text{timeper} = 3.05 + 0.26 \times \text{nterm} \]

Predictor Coef St.dev Const 3.050 .26034 xnterm 0.26034 ...
Review

1. The least-squares line \( \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \) passes through the points \((x = 0, \hat{y} = \gamma)\) and \((x = \tau, \hat{y} = \eta)\). Supply the missing values.

\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \tau)(y_i - \eta)}{\sum_{i=1}^{n} (x_i - \tau)^2}, \quad \hat{\beta}_0 = \eta - \hat{\beta}_1 \tau
\]

Review

1. X={-1, 2, 3, 4}, Y={0, -1, 1, 2}, \( \beta_0 = 0.5 \)

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>( x - \tau )</th>
<th>( y - \hat{y} )</th>
<th>( (x - \tau)^2 )</th>
<th>( (y - \hat{y})^2 )</th>
<th>( (x - \tau)(y - \hat{y}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>0</td>
<td>-3</td>
<td>-0.5</td>
<td>9</td>
<td>0.25</td>
<td>1.5</td>
</tr>
<tr>
<td>2</td>
<td>-1</td>
<td>0</td>
<td>-1.5</td>
<td>0</td>
<td>2.25</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
<td>1</td>
<td>0.25</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
<td>1.5</td>
<td>4</td>
<td>2.25</td>
<td>3</td>
</tr>
</tbody>
</table>

\( \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \tau)(y_i - \hat{y}_i)}{\sum_{i=1}^{n} (x_i - \tau)^2} \)

\( \hat{\beta}_0 = \hat{y} - \hat{\beta}_1 \tau \)

Review

1. What are the quantities that specify a particular line?
2. Explain the idea of a prediction error in the context of fitting a line to a scatter plot. To what visual feature on the plot does a prediction error correspond?
3. What property is satisfied by the line that fits the data best in the least-squares sense?
4. The least-squares line \( \hat{y} = \hat{\beta}_0 + \hat{\beta}_x x \) passes through the points \((x = 0, \hat{y} = \gamma)\) and \((x = \tau, \hat{y} = \eta)\). Supply the missing values.

\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \tau)(y_i - \hat{y}_i)}{\sum_{i=1}^{n} (x_i - \tau)^2}, \quad \hat{\beta}_0 = \hat{y} - \hat{\beta}_1 \tau
\]

RMS Error for regression

- Error = Actual value – Predicted value

\( Y = \beta_0 + \beta_1 X \)

The RMS error for the regression line \( Y = \beta_0 + \beta_1 X \) is

\[
\sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}
\]

where \( \hat{y}_i = \beta_0 + \beta_1 x_i \), \( 1 \leq i \leq N \)

Compute the RMS Error for this regression line

- Error = Actual value – Predicted value

\( X, Y \)

\( 0, 2, 4, 6, 8 \)

\( 0, 10, 20, 30 \)

\( 1, 9, 2, 15, 3, 12, 4, 19, 5, 11, 6, 20, 7, 22, 8, 18, 19, 27, 20, 33, 36, 39 \)

\( Y = \beta_0 + \beta_1 X \)

The RMS error for the regression line \( Y = \beta_0 + \beta_1 X \) is

\[
\sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}
\]

where \( \hat{y}_i = \beta_0 + \beta_1 x_i \), \( 1 \leq i \leq N \)
Compute the RMS Error for this regression line

- Error = Actual value – Predicted value
- The RMS Error for the regression line $Y = \beta_0 + \beta_1 \, X$ is
  \[ \sqrt{\frac{(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2}{n-2}} \]
  where $\hat{y}_i = \beta_0 + \beta_1 \, x_i$, $1 \leq i \leq n$.
- First compute the LS linear fit (estimate $\hat{\beta}_0$, $\hat{\beta}_1$).
- Then compute the individual errors.
- Finally compute the cumulative RMS measure.

Compute the R{MS Error} for this regression line

- First compute the LS linear fit (estimate $\hat{\beta}_0$, $\hat{\beta}_1$).
- Then compute the individual errors.
- Finally compute the cumulative RMS measure.

Compute the RMS Error for this regression line

- Then compute the individual errors
  \[ (y_i - \hat{y}_i)^2, \quad \text{where} \quad \hat{y}_i = \beta_0 + \beta_1 \, x_i, \quad 1 \leq i \leq n \]
- Finally compute the cumulative RMS measure.
  \[ \sqrt{\frac{(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2}{n-2}} \]
  where $\hat{y}_i = \beta_0 + \beta_1 \, x_i$, $1 \leq i \leq 5$.

Compute the RMS Error for this regression line

- The RMS Error for the regression line $Y = \beta_0 + \beta_1 \, X$ says how far away from the (model/predicting) regression line is each observation.
- Observe that the SD(Y) is also a RMS Error measure of another specific line – horizontal line through the average of the Y values. This line may also be taken for a regression line, but often it’s not the best linear fit.

Plotting the Residuals

- The Residuals=Observed –Predicted for the regression line $Y = \beta_0 + \beta_1 \, X$ (just like the error).
- Residuals average to zero, mathematically, and the regression line for the residuals is a horizontal line through $y=0$.

Plotting the Residuals – patterns?

- The Residuals=Observed –Predicted for the regression line $Y = \beta_0 + \beta_1 \, X + U$ should show no clear trend or pattern, for our linear model to be a good and useful approximation to the unknown process.
Is there always an X Y relationship? Linear Relationship?

(a) 1000 data points with no relationship between X and Y

Random samples from these 1000 data points

Random samples from these 1000 data points

(b) 12 random samples each of size 20

Review

1. Describe a fundamental difference between the way regression treats data and the way correlation treats data.
2. What is the correlation coefficient intended to measure?
3. For what shape(s) of trend in a scatter plot does it make sense to calculate a correlation coefficient?
4. What is the meaning of a correlation coefficient of $r = +1$? $r = -1$? $r = 0$?

Summary

Lines in the Plane

- Draw the following lines:
  - $Y = 3.4X + 13$
  - $Y = -3X - 5.7$
  - Line through $(X_1, Y_1)$ and $(X_2, Y_2)$.
  - $(Y - Y_1) / (Y_2 - Y_1) = (X - X_1) / (X_2 - X_1)$.
Relationships between quantitative variables should be explored using scatter plots.

- Usually the $Y$ variable is continuous (or behaves like one in that there are few repeated values)
- and the $X$ variable is discrete or continuous.

Regression singles out one variable ($Y$) as the response and uses the explanatory variable ($X$) to explain or predict its behavior.

Correlation treats both variables symmetrically as random.

In practical problems, regression models may be fitted for any of the following reasons:

- To understand a causal relationship better. Ex?
- To find relationships which may be causal. Ex?
- To make predictions. Ex?
  - But be cautious about predicting outside the range of the data
- To test theories. Ex?
- To estimate parameters in a theoretical model.

In observational data, strong relationships are not necessarily causal.

We can only have reliable evidence of causation from controlled, randomized, designed experiments.

Be aware of the possibility of lurking variables which may effect both $X$ and $Y$.

The two main approaches to summarizing trends in data are using smoothing and fitting models (e.g., regression lines).

The least-squares criterion for fitting a mathematical curve is to choose the values of the parameters (e.g. $\beta_0$ and $\beta_1$) to minimize the sum of squared prediction errors, $\sum(y_i - \hat{y}_i)^2$.

We fit the linear relationship $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

The slope $\hat{\beta}_1$ is the change in $\hat{y}$ associated with a one-unit increase in $x$.

Least-squares estimates

- The least-squares estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$, are chosen to minimize $\sum(y_i - \hat{y}_i)^2$.
- The least-squares regression line is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

The $i$-th residual (or prediction error) is $y_i - \hat{y}_i = \text{observed} - \text{predicted}$.

An outlier is a data point with an unexpectedly large residual (positive or negative).
The correlation coefficient $r$ is a measure of linear association with $-1 \leq r \leq 1$.

- If $r = 1$, then $X$ and $Y$ have a perfect positive linear relationship.
- If $r = -1$, then $X$ and $Y$ have a perfect negative linear relationship.
- If $r = 0$, then there is no linear relationship between $X$ and $Y$.
- Correlation does not necessarily imply causation.

Correlation coefficient

Recall the correlation coefficient...

\[
R(X;Y) = \text{Corr}(X;Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n}xy - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}
\]

Misuse of the correlation coefficient

Some patterns with $r = 0$

- Correlation is invariant w.r.t. linear transformations of $X$ or $Y$.
- Correlation is Associative.
- Correlation measures linear association, NOT an association in general!!!
**Correlation does not necessarily imply causation.**

**Linear Regression**

- Regression relationship = trend + residual scatter
  \[ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \text{Err} \]
- Trend = best linear fit line (LS)
  \[ \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \]
- Scatter = residual (prediction) error
  \[ \text{Err} = \text{Obs} - \text{Pred} \]

**Textbook vs. Lecture Notation …**

1. Note that there is a slight difference in the formula for the slope of the Least-Squares Best-Linear Fit line:

\[
\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}
\]

**Redo the problem from last time using:**

1. \( X = \{-1, 2, 3, 4\}, \ Y = \{0, -1, 1, 2\} \)

<table>
<thead>
<tr>
<th>( x )</th>
<th>( y )</th>
<th>( x-\bar{x} )</th>
<th>( y-\bar{y} )</th>
<th>( (x-\bar{x})(y-\bar{y}) )</th>
<th>( (x-\bar{x})^2 )</th>
<th>( (y-\bar{y})^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>0</td>
<td>-3</td>
<td>-0.5</td>
<td>1.5</td>
<td>9</td>
<td>0.25</td>
</tr>
<tr>
<td>2</td>
<td>-1</td>
<td>0</td>
<td>-1.5</td>
<td>2.25</td>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
<td>0.25</td>
<td>1</td>
<td>0.25</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
<td>1.5</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

\[ \hat{\beta}_1 = \text{Cov}(X,Y) \times \frac{\text{Var}(Y)}{\text{Var}(X)} \]

\[ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \]
Properties of Linear Regression

- Linear Fit that minimizes the sum-square error of:
  \[ \hat{y} = \beta_0 + \beta_1 x \]
  \[ \text{obs. vs. predicted values: } \sum (y_i - \hat{y}_i)^2 \]

- Properties of Linear Regression

\[ \hat{y} = \beta_0 + \beta_1 x \]

- The points \((x = 0, \ y = \hat{y})\) and \((x = \bar{x}, \ y = \bar{y})\) lie on the LS line.
- RMS error – indicates how far are typical points from the regression line (up/down)

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

where \(n = \beta + \beta x\), \(1 \leq k \leq 5\)