Chapter 12: 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.

Chapter 12: Lines in 2D

- 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 $SISVOL/DIAVOL$) 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, $Y$.

Looking vertically

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

Correlation Coefficient

Correlation coefficient ($-1 \leq R \leq 1$): a measure of linear association, or clustering around a line of multivariate data.

Relationship between two variables ($X, Y$) can be summarized by: ($\mu_X, \sigma_X$), ($\mu_Y, \sigma_Y$) and the correlation coefficient, $R$. $R = 1$, perfect positive correlation (straight line relationship), $R = 0$, no correlation (random cloud scatter), $R = -1$, perfect negative correlation.

Computing $R(X,Y)$: (standardize, multiply, average)

$$R(X,Y) = \frac{1}{N-1} \sum_{k=1}^{N} \left( \frac{x_k - \mu_X}{\sigma_X} \right) \left( \frac{y_k - \mu_Y}{\sigma_Y} \right) X = \{x_1, x_2, \ldots, x_N\}, \ Y = \{y_1, y_2, \ldots, y_N\}, \ \mu_X, \sigma_X, \mu_Y, \sigma_Y$$
Correlation Coefficient

Example:

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

<table>
<thead>
<tr>
<th>Student Height</th>
<th>Weight</th>
<th>R(X,Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>55</td>
<td>0.87</td>
</tr>
<tr>
<td>2</td>
<td>66</td>
<td>0.78</td>
</tr>
<tr>
<td>3</td>
<td>77</td>
<td>0.69</td>
</tr>
<tr>
<td>4</td>
<td>88</td>
<td>0.59</td>
</tr>
<tr>
<td>5</td>
<td>99</td>
<td>0.49</td>
</tr>
</tbody>
</table>

Total: 260.2 0 = 0.98

Correlation Coefficient - Properties

Correlation is invariant w.r.t. linear transformations of X or Y

\[
R(aX + b, cY + d) = \frac{\sum (aX_k + b - \mu_{aX} + b)(cY_k + d - \mu_{cY} + d)}{\sigma_{aX} \sigma_{cY}}
\]

Correlation Coefficient - Properties

Correlation is Associative

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

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

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 = Number of terminals:</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>45</th>
<th>40</th>
<th>10</th>
<th>30</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y* = Total Time (mins):</td>
<td>6.6</td>
<td>14.9</td>
<td>18.4</td>
<td>12.4</td>
<td>7.9</td>
<td>0.9</td>
<td>5.5</td>
<td>2.7</td>
</tr>
<tr>
<td>X = Number of terminals:</td>
<td>50</td>
<td>30</td>
<td>65</td>
<td>40</td>
<td>65</td>
<td>65</td>
<td>65</td>
<td>65</td>
</tr>
<tr>
<td>Y = Total Time (mins):</td>
<td>12.6</td>
<td>6.7</td>
<td>23.6</td>
<td>9.2</td>
<td>20.2</td>
<td>21.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X = Time Per Task (secs):</td>
<td>15.1</td>
<td>13.3</td>
<td>21.8</td>
<td>13.8</td>
<td>18.6</td>
<td>19.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
We want to find reasonable models (descriptions) for these 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 idea of a residual or prediction error

\[ u = y - \hat{y} \]

Least squares criterion

Choose the values of the parameters to minimize the sum of squared prediction errors (or sum of squared residuals),

\[ \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]
The least squares line

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

Choose line with smallest sum of squared prediction errors

\[ \text{Min } \sum (y_i - \hat{y}_i)^2 \]

Its parameters are denoted:

- Intercept: \( \hat{\beta}_0 \)
- Slope: \( \hat{\beta}_1 \)

Prediction errors

\[ (x_i, y_i) \]

Least-squares line

Choose line with smallest sum of squared prediction errors

\[ \text{Min } \sum (y_i - \hat{y}_i)^2 \]

Its parameters are denoted:

- Intercept: \( \hat{\beta}_0 \)
- Slope: \( \hat{\beta}_1 \)

Prediction errors

\[ (x_i, y_i) \]

Computer timings data – linear fit

\[ \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n} (y_i - \bar{y})^2 \]

\[ \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \]

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

Computer timings data

Table 12.3.1 Prediction Errors

<table>
<thead>
<tr>
<th>( x )</th>
<th>( y )</th>
<th>( \bar{y} )</th>
<th>( \bar{x} )</th>
<th>( \hat{y} )</th>
<th>( y - \hat{y} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>9.00</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>-2.00</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>2.00</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>13.00</td>
<td>13.00</td>
<td>0.00</td>
<td>13.00</td>
<td>0.00</td>
</tr>
<tr>
<td>65</td>
<td>18.60</td>
<td>10.50</td>
<td>2.80</td>
<td>11.50</td>
<td>1.85</td>
</tr>
<tr>
<td>65</td>
<td>15.10</td>
<td>19.25</td>
<td>-0.65</td>
<td>16.75</td>
<td>5.05</td>
</tr>
</tbody>
</table>

Computer timings data

\[ \text{Sum of squared errors } 37.46 \]

\[ \text{Sum of squared errors } 90.36 \]

Adding the least squares line

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

\[ (x_i, y_i) \]


1. The least-squares line \( \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \) passes through the points \( (x = 0, y = ?) \) and \( (x = \bar{x}, y = ?) \). Supply the missing values.
Hands – on worksheet!

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>1</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>-1</td>
<td>0</td>
<td>-2</td>
<td>2</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

\[
\hat{\beta}_0 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \quad \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}
\]

Hands – on worksheet!

1. $X = \{-1, 2, 3, 4\}, \ Y = \{0, -1, 1, 2\}, \ \bar{x} = 2, \ \bar{y} = 0.5$

<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>-1</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>-1</td>
<td>0</td>
<td>-2</td>
<td>2</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

\[
\hat{\beta}_0 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \quad \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}
\]

Fitting a line through the data

Show the Regression-Line Simulation Applet
RegressionApplet.html

(a) The data
(b) Which line?

The simple linear model

When $X = x$, $Y \sim \text{Normal}(\mu_Y, \sigma_Y)$ where $\mu_Y = \beta_0 + \beta_1 x$, OR when $X = x$, $Y = \beta_0 + \beta_1 x + U$, where $U \sim \text{Normal}(0, \sigma_U)$

Random error

Data generated from $Y = 6 + 2x + \text{error (U)}$

Dotted line ———— is true line and solid line ———— is the data-estimated LS line. Note differences between true $\beta_0 = 6, \beta_1 = 2$ and their estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.
Data generated from $Y = 6 + 2x + \text{error(U)}$

Sample 3: $\beta_0 = 7.38, \beta_1 = 2.10$

Sample 4: $\beta_0 = 7.92, \beta_1 = 1.59$

Sample 5: $\beta_0 = 9.14, \beta_1 = 1.13$

Combined: $\beta_0 = 7.44, \beta_1 = 1.70$

Data generated from $Y = 6 + 2x + \text{error(U)}$

Recall the correlation coefficient…

Another form for the correlation coefficient is:

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

Misuse of the correlation coefficient

Some patterns with $r = 0$

Linear Regression

Regression relationship = trend + residual scatter

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \text{Err}$

Trend=best linear fit Line (LS)

$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$; $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Scatter = residual (prediction) error

$\text{Err} = \text{Obs-Pred}$

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

Another Notation for the Slope of the LS line

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

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

$\beta_1 = \tau - \beta_0 x$

$\hat{\beta}_1 = \text{Corr}(X;Y) \times \frac{SD(Y)}{SD(X)}$; $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
Another Notation for the Slope of the LS line

\[ \hat{\beta}_1^{\text{New}} = \text{Corr}(X,Y) \frac{SD(Y)}{SD(X)} \]

\[
= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \times \left( \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{N-1} \right)^{1/2}
\]

Course Material Review

1. Part I
2. Data collection, surveys.
3. Experimental vs. observational studies
4. Numerical Summaries (5-number-summary)
5. Binomial distribution (prob’s, mean, variance)
6. Probabilities & proportions, independence of events and conditional probabilities
7. Normal Distribution and normal approximation

Course Material Review – cont.

1. Part II
2. Central Limit Theorem – sampling distribution of \( \bar{X} \)
3. Confidence intervals and parameter estimation
4. Hypothesis testing
5. Paired vs. Independent samples
6. Chi-Square (\( \chi^2 \)) Goodness-of-fit Test
7. Analysis Of Variance (1-way-ANOVA, one categorical var.)
8. Correlation and regression
9. Best-linear-fit, least squares method