Recap

• (Ch 2) Visualizing and summarizing relationships in data
  • Time series data
  • Scatter plots

Today

• (Ch 2) Visualizing and summarizing relationships in data
  • The correlation coefficient
  • Prediction
Correlation coefficient

Given a data set \( \{(x, y)\} \) consisting of items \((x_1, y_1), \ldots, (x_N, y_N)\),

- Standardize the data

\[
\hat{x}_i = \frac{x_i - \text{mean}(\{x\})}{\text{std}(\{x\})} \quad \text{and} \quad \hat{y}_i = \frac{y_i - \text{mean}(\{y\})}{\text{std}(\{y\})}
\]

- The correlation coefficient is the mean of \(\hat{x}_i \hat{y}_i\)

\[
\text{corr}(\{(x, y)\}) = \frac{1}{N} \sum_{i=1}^{N} \hat{x}_i \hat{y}_i
\]
Correlation, more precisely

In a data set \{ (x, y) \} consisting of items \( (x_1, y_1), \ldots, (x_N, y_N) \)

- we say \( x \) and \( y \) have **positive correlation** if \( \text{corr}(\{(x, y)\}) > 0 \)
- we say \( x \) and \( y \) have **negative correlation** if \( \text{corr}(\{(x, y)\}) < 0 \)
- we say \( x \) and \( y \) have **zero correlation** if \( \text{corr}(\{(x, y)\}) = 0 \)
Properties of the correlation coefficient

- The correlation coefficient is symmetric

\[ \text{corr}((x, y)) = \text{corr}((y, x)) \]

- Translating the data does not change the correlation coefficient
- Scaling the data may change the sign of the correlation coefficient

\[ \text{corr}((ax + b, cy + d)) = \text{sign}(ac) \text{corr}((x, y)) \]
Bounds on the correlation coefficient

The correlation coefficient takes values between -1 and 1 inclusive

$$\text{corr}((x, y)) = 1 \text{ if and only if } \hat{x}_i = \hat{y}_i$$

$$\text{corr}((x, y)) = -1 \text{ if and only if } \hat{x}_i = -\hat{y}_i$$
Prediction

From *Spurious Correlations* by Tyler Vigen
Using correlation to predict

- Given a correlated data set \( \{(x, y)\} \), we can predict a value \( y_0^p \) that goes with a given \( x_0 \).

- In standard coordinates \( \{ (\hat{x}, \hat{y}) \} \), we can predict a value \( \hat{y}_0^p \) that goes with a given \( \hat{x}_0 \).
Linear predictor and its error

• We will assume that our predictor is linear
  \[ \hat{y}^p = a\hat{x} + b, \text{ where } a \text{ & } b \text{ are constants} \]

• We denote the prediction of \( \hat{y}_i \) at each \( \hat{x}_i \) in the data set as \( \hat{y}^p_i \)
  \[ \hat{y}^p_i = a\hat{x}_i + b \]

• The error in the prediction of \( \hat{y}_i \) is denoted \( u_i \)
  \[ u_i = \hat{y}_i - \hat{y}_i^p = \hat{y}_i - a\hat{x}_i - b \]
The mean of the prediction error should be 0

\[ 0 = \text{mean}\{ \varepsilon_i \} \]
\[ = \text{mean}\{ \{ \hat{y}_i - a\hat{x}_i - b \} \} \]
\[ = \text{mean}\{ \{ \hat{y}_i \} \} - a\text{mean}\{ \{ \hat{x}_i \} \} - b = -b \]

because \{\hat{x}_i\} and \{\hat{y}_i\} are in standard coordinates
$b = 0$

So the linear predictor becomes $\hat{y} = a \hat{x}$
The variance of the error should be minimal

$$\begin{align*}
\text{var}(\{u_i\}) &= \text{mean} \left( \left\{ \left( u_i - \text{mean}(\{u_i\}) \right)^2 \right\} \right) \\
&= \text{mean}(\{u_i^2\}) \\
&= \text{mean}(\{(\hat{y}_i - \hat{y})^2\}) \\
&= \text{mean}(\{(\hat{y}_i - a\hat{x}_i)^2\})
\end{align*}$$
\[
= \text{mean} \left( \left\{ \hat{y}_i - 2a \hat{x}_i \hat{y}_i + a^2 \hat{x}_i^2 \right\} \right)
\]
\[
= \text{mean} \left( \left\{ \hat{x}_i^2 \right\} \right) - 2a \text{mean} \left( \left\{ \hat{x}_i \hat{y}_i \right\} \right) + a^2 \text{mean} \left( \left\{ \hat{x}_i^2 \right\} \right)
\]
\[
= \text{var} \left( \left\{ \hat{y}_i \right\} \right) - 2a \text{corr} \left( \left\{ \hat{x}_i \hat{y}_i \right\} \right) + a^2 \text{var} \left( \left\{ \hat{x}_i^2 \right\} \right)
\]
\[
= 1 - 2ar + a^2
\]

let this be \( r \)
\[ \frac{d}{da} \left( \text{var}(\hat{Y}_u|\beta) \right) = -2r + 2a \]

Setting this to 0 gives that \( a = r \) minimizes the variance of error

So \[ \hat{y} = \hat{r} x \] is the linear predictor
Prediction formulas

• In standard coordinates
  \[ \hat{y}_0^p = r \hat{x}_0 \text{ where } r = \text{corr}((x, y)) \]

• In original coordinates
  \[ \frac{y_0^p - \text{mean}\{y\}}{\text{std}\{y\}} = r \left( \frac{x_0 - \text{mean}\{x\}}{\text{std}\{x\}} \right) \]
Root-mean-square (RMS) prediction error

Recall that

$$ \text{var}(\{u\}) = \text{mean}(\{u^2\}) $$

$$ = 1 - 2a + a^2 $$

$$ = 1 - 2r^2 + r^2 \quad \text{since } a = r $$

$$ = 1 - r^2 $$

So RMS error = \( \sqrt{\text{mean}(\{u^2\})} = \sqrt{1-r^2} \)