Lecture 23
Linear Regression
The Correlation Coefficient $r$

- Measures linear association
- Based on standard units
- $-1 \leq r \leq 1$
  - $r = 1$: scatter is perfect straight line sloping up
  - $r = -1$: scatter is perfect straight line sloping down
- $r = 0$: No linear association; uncorrelated
Definition of $r$

Correlation Coefficient ($r$) =

| average of | product of | $x$ in standard units | and | $y$ in standard units |

Measures how clustered the scatter is around a straight line

(Demo)
Prediction
Prediction

If we have a line describing the relation between two variables, we can make predictions.
Prediction

- **Problem:** given a known $x$ value, predict $y$, where both are in standard units
- **Solution:**
  - Compute $r$
  - Predict that $y = r \times x$
- Why is that a line?
Algebra review:

Equation of a Line

In general:

\[ y = a \times x + b \]

(a is slope, b is intercept)
Prediction

- **Problem**: given a known $x$ value, predict $y$, where both are in standard units

- **Solution**:
  - Compute $r$
  - Predict that $y = r \times x$

- Why is that a line?
- Why use *that* line?

(Demo)
Prediction

- **Problem:** given a known $x$ value, predict $y$, where both are in standard units
- **Solution:**
  - Compute $r$
  - Predict that $y = r \times x$
- Why is that a line?
- Why use *that* line?
  - It is a version of the graph of averages, smoothed to a line (Demo)
**Prediction**

- **Predict** $y = r \times x$ (in standard units)
- If $r = .75$ and $x$ is 2 std above mean, then prediction for $y$ is 1.5 std above mean
- So $y$ predicted to be **closer to its mean** than $x$ is

- “Regression to the mean”
  - Children with exceptionally tall parents tend not to be as tall
  - Galton called it “regression to mediocrity” (Demo)
Linear Regression

(Demo)
Equation for regression line

\[(y \text{ in } su) = r \times (x \text{ in } su)\]
Equation for regression line

\[(y \text{ in } su) = r \times \frac{x - \text{mean(all } x)}{\text{std(all } x)}\]
Equation for regression line

\[
\frac{y - \text{mean(all}\ y)}{\text{std(all}\ y}} = r \times \frac{x - \text{mean(all}\ x)}{\text{std(all}\ x}}
\]
Equation for regression line

\[
\frac{y - \text{mean(all } y)}{\text{std(all } y)} = r \times \frac{x - \text{mean(all } x)}{\text{std(all } x)}
\]

Do some algebra to put that in the form \( y = \text{slope} \times x + \text{intercept} \)...
Slope and Intercept

\[ y = \text{slope} \times x + \text{intercept} \]

\[ \text{slope of the regression line} = r \cdot \frac{\text{SD of y}}{\text{SD of x}} \]

\[ \text{intercept of the regression line} = \text{average of y} - \text{slope} \times \text{average of x} \]

(Demo)
Regression Line

Standard Units

Original Units

\[(0, 0)\]

1

\[r\]

\[(\text{Average } x, \text{Average } y)\]

\[r \times \text{SD } y\]

\[\text{SD } x\]
Abuses of $r$

- Summarizing non-linear data with $r$
- Eliminating outliers to “improve” $r$
- Drawing conclusions about individuals based on data about groups (*ecological* correlations)
- Jumping to conclusions about causality
Correlation is not causation

Figure 1. Correlation between Countries’ Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.
Quantifying Error
Error in Prediction

● How good is the regression line at making predictions?
  ○ Hard to say for unknown data
  ○ But easy for data we already have

● \( \text{error} = \text{actual value} - \text{prediction} \)
Error in Prediction

- How good is the regression line at making predictions?
  - Hard to say for unknown data
  - But easy for data we already have

- \textbf{error} = \textbf{actual value} − \textbf{prediction}
- \textbf{RMSE} = \textbf{root mean square error}
  
  \[ \text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\text{error}_i)^2}{n}} \]

- \textbf{RMSE} = \textbf{root mean square of deviation from prediction}
  
  \[ \text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\text{deviation}_i)^2}{n}} \]
RMSE

RMSE = root mean square error

\[ \text{RMSE} = \text{std}(y) \times \sqrt{1 - r^2} \]

- If \( r = 1 \), what is RMSE? 0
- If \( r = 0 \), what is RMSE? \( \text{std}(y) \)

Compare regression line to other lines using RMSE... (Demo)
SciPy function `minimize(f)` returns the value $x$ that produces the minimum output $f(x)$ from $f$.

- Also works for functions that make multiple arguments.
- How to use to find best line:
  - Write function `rmse(a, b)` that returns the RMSE for line with slope $a$ and intercept $b$.
  - Call `minimize(rmse)` and get output array $[a_0, b_0]$.
  - $a_0$ is slope and $b_0$ intercept of line that minimizes RMSE.
Regression line

- Regression line has the minimum RMSE of all lines

Names:
- Regression line
- Least squares line
- “Best fit” line