Analysis of Dependent Variables: Correlation and Simple Regression

Zacariah Labby, PhD, DABR
Asst. Prof. (CHS), Dept. of Human Oncology
University of Wisconsin – Madison

Conflicts of Interest

None to disclose
Purpose

• Review basic statistics and identify appropriate use of statistics related to analyzing simple relationships between two variables:
  • Correlation statistics
  • Linear regression and model fitting
Correlation: Review of Terminology

• **Dependent vs. Independent Variables**
  • Standard plot: X is Independent Y is Dependent

• **Linear vs. Monotonic**
  • Linear: increase in X leads to proportional increase in Y
  • Monotonic: increase in X leads to some increase in Y

---

Correlation: Review of Terminology

• **Variable Type**
  • Continuous
    • Example: Ionization chamber charge collected vs. Dose delivered
  • Discrete
    • Example: Number of patients seen vs. Calendar year
  • Ordinal
    • Example: Severity of normal tissue toxicity vs. Prescription Level
  • Categorical
    • Example: RECIST response classification vs. Radiologist Observer
Correlation: Metrics of Interest

• Four big categories of data
  • Continuous
  • Discrete
  • Ordinal
  • Categorical

Three major correlation metrics

- Pearson’s $r$
- Spearman’s $\rho$
- Fleiss’ $\kappa$
**Correlation: Metrics of Interest**

- Four big categories of data
  - Continuous
  - Discrete
  - Ordinal
  - Categorical

- Three major correlation metrics
  - Pearson’s $r$
  - Spearman’s $\rho$
  - Fleiss’ $\kappa$

**Correlation: Pearson’s $r$**

- “Linear” or “Product-Moment” correlation
- Applies **only** to continuous data
- Parametric correlation
  - Tendency of dependent variable to **increase linearly** with the independent variable

- **Key Point:**
  - There is an assumed form to the relationship
  - Linear, and therefore also monotonic
Correlation: Spearman’s $\rho$

- “Rank” correlation
- Applies to continuous, discrete, or ordinal data
- Non-parametric correlation
- Tendency of dependent variable to increase with the independent variable
- Key Point:
  - There is no assumed relationship, only monotonicity
- Math: Pearson’s $r$ of rank-transformed data
Correlation: Spearman’s $\rho$

Raw: (0,0)  
Rank: (1,1)

Raw: (0.05,0.0025)  
Rank: (2,2)

(X,Y) pairs
Correlation: Spearman’s $\rho$

- Raw: $(1,1)$, Rank: $(20,20)$
- Raw: $(0.05,0.0025)$, Rank: $(2,2)$
- Raw: $(0,0)$, Rank: $(1,1)$

Pearson’s $r$ of rank-transformed data: 1.00

$r = 1.00$  \( \rho = 1.00 \)

$r = 0.97$  \( \rho = 0.97 \)

$r = 0.76$  \( \rho = 0.90 \)
Correlation: Spearman’s $\rho$

Continuous variables; “When one goes up, does the other (reliably) go down?”
Continuous variables; “When one goes up, does the other (reliably) go down?”

**Correlation: Which Metric?**

Answer: Spearman’s \( \rho \)

**Correlation: Fleiss’ \( \kappa \)**

- Categorical correlation
- Applies only to categorical data
  - Categorical data could be inherently ordinal
- Non-parametric correlation
  - How well do independent categories sort dependent categories?
- Math: number of dependent-independent pairs in agreement over the number expected by chance alone.
Correlation: Fleiss’ $\kappa$

- Example:
  - 5 radiologists contour tumors in
  - 31 patients
  - Response classification from baseline to post-chemo CT scans

<table>
<thead>
<tr>
<th>Response Classification</th>
<th>Obs. 1</th>
<th>Obs. 2</th>
<th>Obs. 3</th>
<th>Obs. 4</th>
<th>Obs. 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Progressive Disease</td>
<td>6</td>
<td>11</td>
<td>7</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td>Stable Disease</td>
<td>17</td>
<td>10</td>
<td>19</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>Partial Response</td>
<td>7</td>
<td>10</td>
<td>5</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Complete Response</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

$\kappa = 0.64$


Correlation vs. Agreement

- Quick tangent...

Important question:
Do you already know that the two variables will be correlated?

Example: Tumor volumes as assessed by Physician vs. Algorithm
Correlation vs. Agreement

- Especially with implicit independent variables (i.e., the true value remains unknown), correlation isn’t as meaningful
- **Correlation** is only the strength of a relationship between two variables
- **Agreement** is the actual 1:1 accuracy

Correlation vs. Agreement

- Absolute agreement vs. Relative agreement
  - Absolute: plot raw differences
  - Relative: plot log differences

\[ \ln \left( \frac{x}{y} \right) = \ln x - \ln y \]

- Get mean, SD of log-transformed data, then apply exponential to get relative agreement bounds


SIMPLE MODELING
Correlation vs. Agreement vs. Modeling

• Correlation: Strength of relationship
• Agreement: Accuracy of 1:1 match
• Modeling: Quantifying the relationship

• Rules of Modeling:
  1. Prefer model with $n-1$ parameters to $n$
  2. Prefer model with $k-1$ independent variables to $k$
  3. Prefer linear model to curved model


Simple Linear Regression

• Linear regression is linear in the coefficients, not necessarily in the independent variable

  $$y = \alpha + \beta x$$

• Linear:

  $$y = \alpha + \beta x^2$$

• Not Linear:

  $$y = \alpha + e^{\beta x}$$
Simple Linear Regression

• Sources of Variance in the data

• Your model: \( y = \alpha + \beta x \)
• Reality: \( y_i = \alpha + \beta x_i + \epsilon_i \)

• Variance in \( y \) can be explained by
  • Variance in \( x \)
  • Residual uncertainty (called \( \epsilon \))

Random (residual) error from fit for each \( x_i \)
Simple Linear Regression

• Sources of Variance in the data
  • Explained Sum of squared errors (ESS)
  • Residual Sum of squared errors (RSS)
  • Total sum of squared errors (TSS)

\[
(f(x_i) - \bar{y})^2 + (y_i - f(x_i))^2 = (y_i - \bar{y})^2
\]

• Coefficient of Determination: ESS/TSS
  • Proportion of total variation in \( y \) explained by the model

\[
R^2 = \frac{ESS}{TSS}
\]

• Pearson’s correlation coefficient
  • \( r = \sqrt{R^2} \)

• Drive home: Correlation quantifies strength of relationship, not relationship itself
Simple Linear Regression

• Making predictions
  • From analysis, derive best-fit values of fit $\hat{\alpha}$, $\hat{\beta}$, etc.
  • Predict new values according to $y_{new} = \hat{\alpha} + \hat{\beta}x_{new}$
• However, models have uncertainty!
  • Variance estimates can be provided for $\hat{\alpha}$, $\hat{\beta}$, etc. (e.g., $\hat{\sigma}_\alpha$)

Simple Linear Regression

• Confidence Bands
  • Variance associated with mean predicted response
    $\text{Var}(\hat{\alpha} + \hat{\beta}x_{new})$
• Prediction Bands
  • Variance associated with single new prediction
  • Takes into account residual errors in linear model
    $\text{Var}(\hat{\alpha} + \hat{\beta}x_{new}) + \hat{\sigma}_\epsilon^2$

(in some ways, this is like the difference between standard deviation and standard error)
Simple Linear Regression

Example

• Task: Department administrator asks you to figure out the relationship between patient census and required RadTech hours.

• Question 1: what kind of relationship would we expect?
  • Probably Linear with some residual uncertainty

• Question 2: which correlation metric would you use?
  • Pearson’s $r$

• Question 3: how would you quantify the relationship?
  • Simple linear regression
Example

• Regression model

\[ \text{Staffing Requirements} = \hat{\alpha} + \hat{\beta} \text{ Patient Workload} \]
\[ \hat{\alpha} = \text{Fixed Staff Overhead} \]
\[ \hat{\beta} = \text{Scalable Coefficient} \]

• Predict tomorrow’s RadTech staffing level if you know the patient workload
  • Could staff at the upper 95% prediction band?

A plug for “R”…

• R is a free software package for data analysis and is very common in the statistics community.

• Good text for learning R and basic stats:
A plug for “R”…

model = lm(y ~ x)
conf.bands = predict(model, interval = 'conf')
pred.bands = predict(model, interval = 'pred')

Use your Biostatisticians

• Many large centers have at least one biostatistician on staff
• In many centers, free consultations for
  • Experimental design
  • Simple clinical trials
  • Data analysis questions
• Prevent headaches and lost costs for rework and rejected papers