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

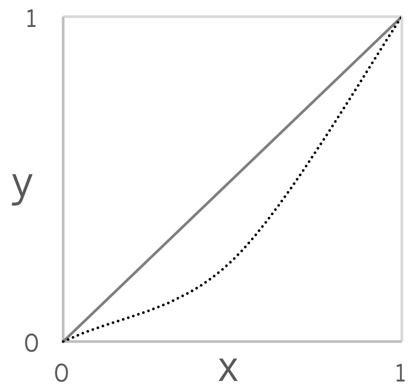
- Review basic statistics and identify appropriate use of statistics related to analyzing simple relationships between two variables:
  - Correlation statistics
  - Linear regression and model fitting

STATISTICS OF CORRELATION

WISCONSIN UNIVERSITY OF WISCONSIN–MADISON

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



WISCONSIN UNIVERSITY OF WISCONSIN–MADISON

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

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Correlation: Metrics of Interest

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

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Correlation: Metrics of Interest

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

Three major correlation metrics

Pearson's r

Spearman's ρ

Fleiss' κ

Correlation: Metrics of Interest

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

Three major correlation metrics

Pearson's r

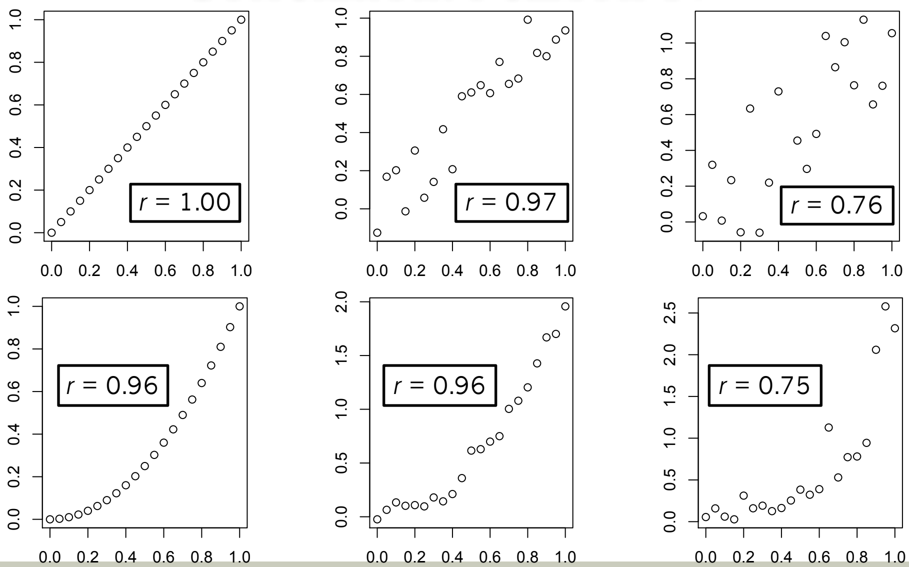
Spearman's ρ

Fleiss' κ

Correlation: Pearson's r

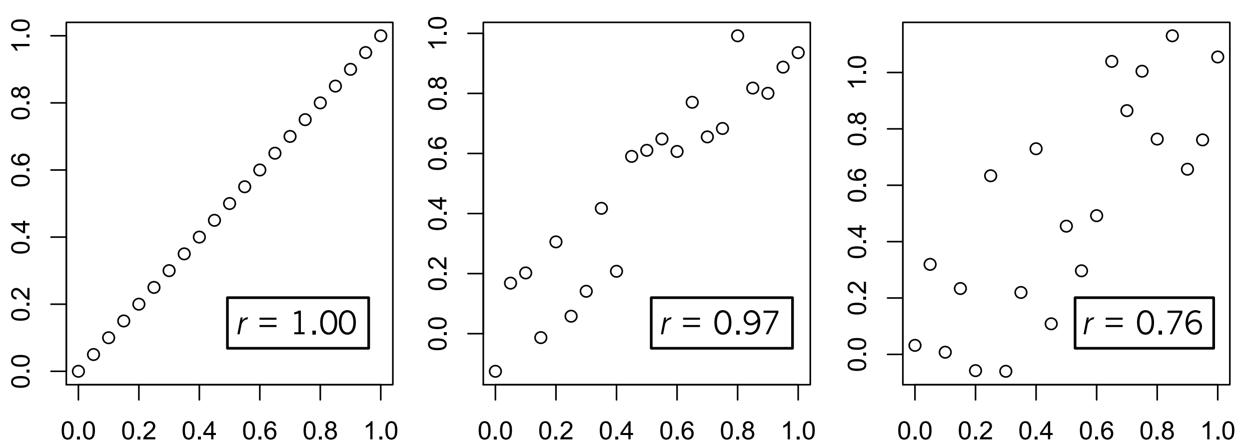
- "Linear" or "Product-Moment" correlation
- Applies only to continuous data
- Parametric correlation
  - Tendency of dependent variable to increase linearly with the independent variable
- Key Point:
  - There is an assumed form to the relationship
    - Linear, and therefore also monotonic

Correlation: Pearson's *r*



Aug 1, 2016 Labby - AAPM 2016 11

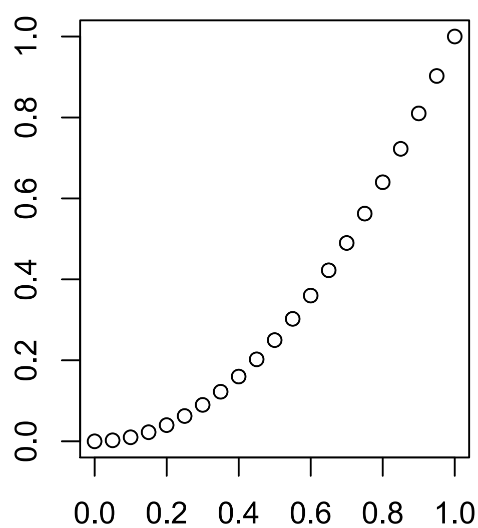
Correlation: Pearson's *r*



Correlation: Spearman's ρ

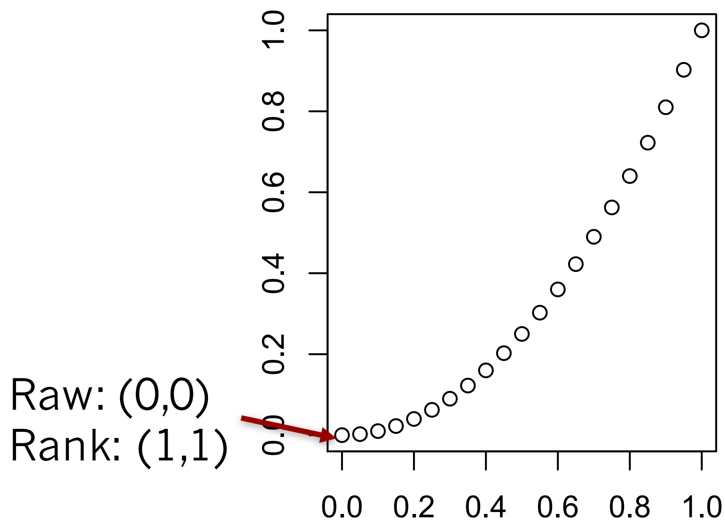
- "Rank" correlation
- Applies to continuous, discrete, or ordinal data
- Non-parametric correlation
  - Tendency of dependent variable to <u>increase</u> with the independent variable
- Key Point:
  - There is no assumed relationship, only monotonicity
- Math: Pearson's *r* of <u>rank-transformed data</u>

Correlation: Spearman's ρ



WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

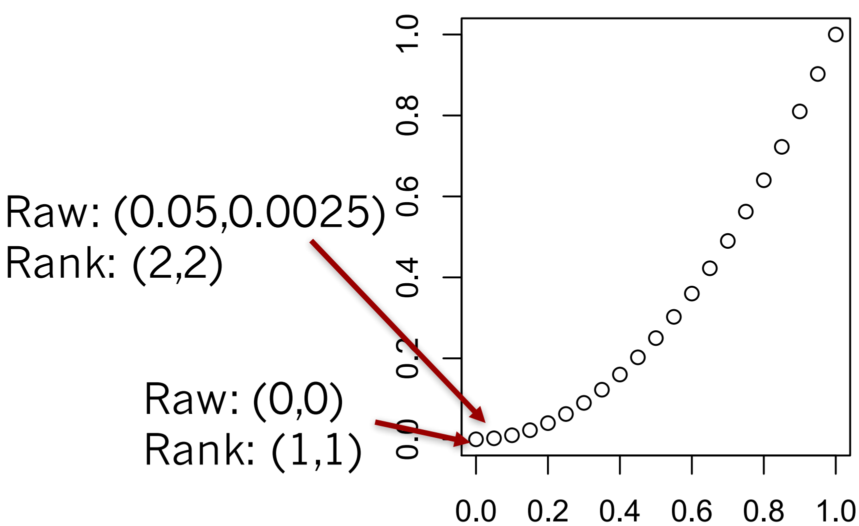
Correlation: Spearman's ρ



(X,Y) pairs

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Correlation: Spearman's ρ



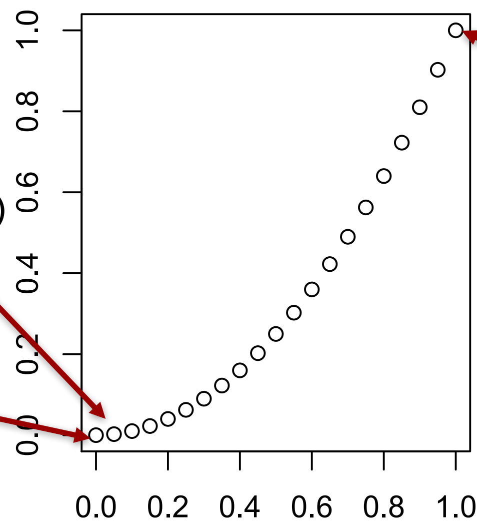
(X,Y) pairs

WISCONSIN

Correlation: Spearman's ρ

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

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

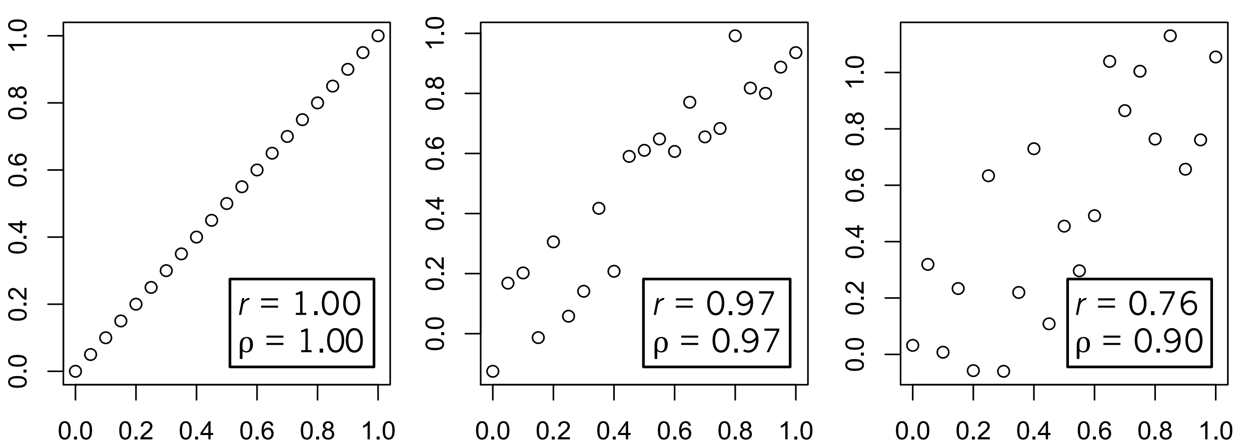


Raw: (1,1)
Rank: (20,20)

Pearson's r of rank-transformed data: 1.00

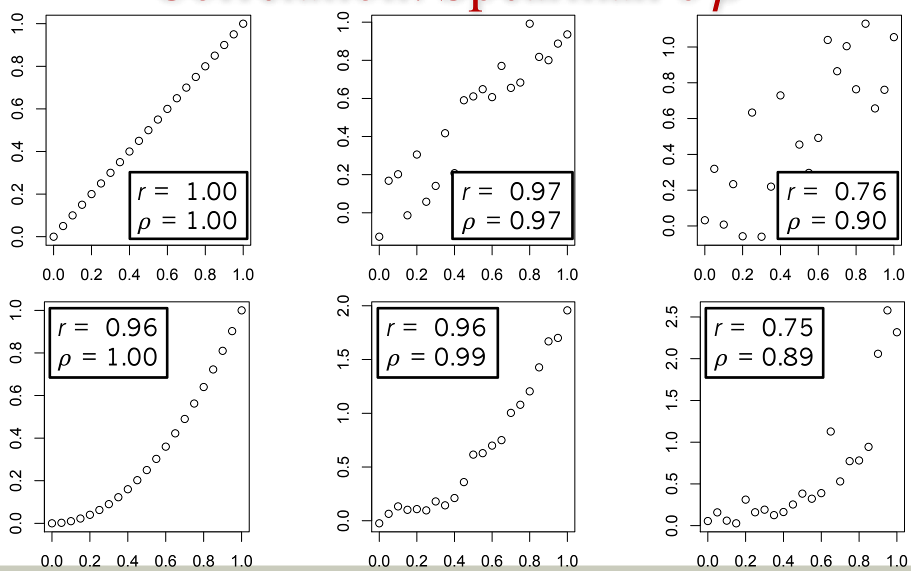
WISCONSIN

Correlation: Spearman's ρ



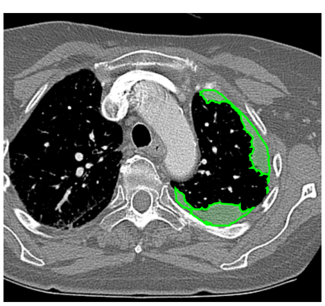
WISCONSIN

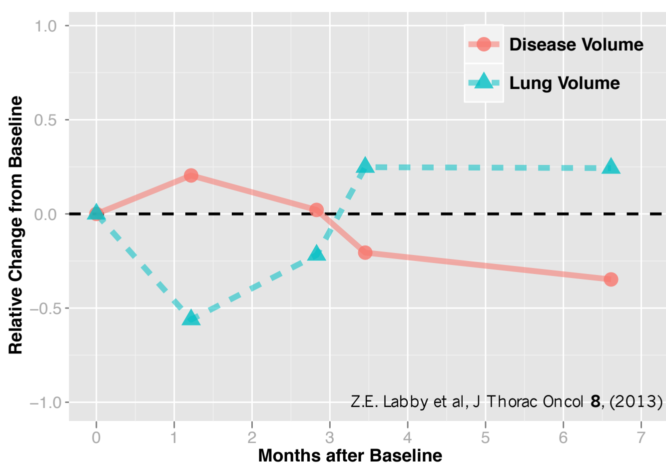
Correlation: Spearman's ρ



WISCONSIN

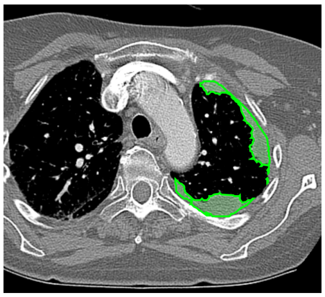
Correlation: Which Metric?

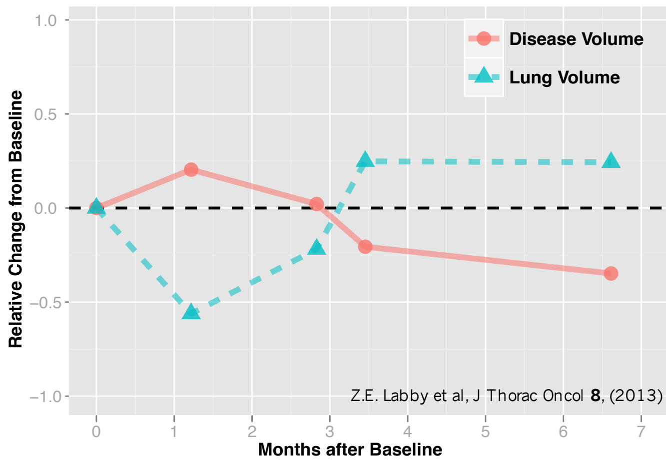




Continuous variables; "When one goes up, does the other (reliably) go down?"

Correlation: Which Metric?





Answer:
Spearman's ρ

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

Correlation: Fleiss' κ

- 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' κ

- Example:
  - 5 radiologists contour tumors in
  - 31 patients
  - Response classification from baseline to post-chemo CT scans
    - Progressive Disease
    - Stable Disease
    - Partial Response
    - Complete Response

	Obs. 1	Obs. 2	Obs. 3	Obs. 4	Obs. 5
Progression	6	11	7	11	14
Stable	17	10	19	15	9
Partial	7	10	5	4	8
Complete	1	0	0	1	0

$\kappa = 0.64$

Landis and Koch, *Biometrics*, **33**,159–174 (1977)

Correlation vs. Agreement

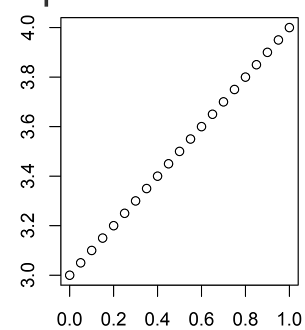
- Quick tangent…

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

Example: Tumor volumes as assessed by Physician vs. Algorithm

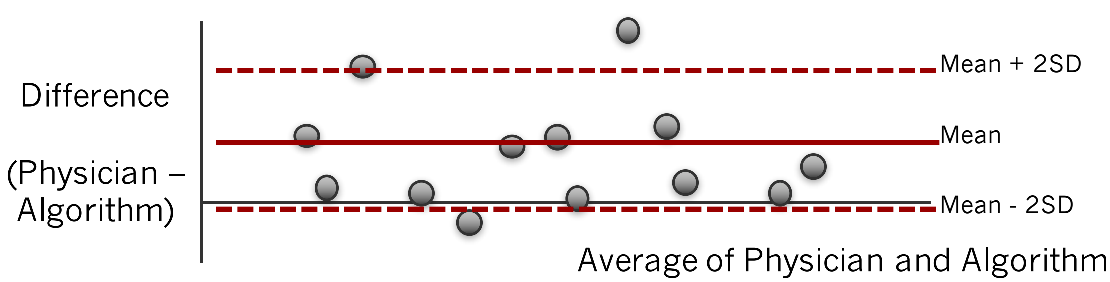
Correlation vs. Agreement

- 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



Bland and Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," Lancet **327**, 307 (1986).

Correlation vs. Agreement



Bland and Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," Lancet **327**, 307 (1986).

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

Bland and Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," Lancet **327**, 307 (1986).

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

Crawley, Statistics: An Introduction using R, Wiley (2005)

Simple Linear Regression

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

- Linear:

$$\boldsymbol{y} = \alpha + \beta\boldsymbol{x}$$

$$\boldsymbol{y} = \alpha + \beta\boldsymbol{x}^2$$

- Not Linear: $\boldsymbol{y} = \alpha + e^{\beta\boldsymbol{x}}$

Simple Linear Regression

- I was going to put the math here, but…

Simple Linear Regression

- Sources of Variance in the data

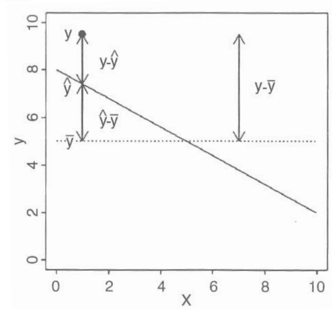
- Your model: $\boldsymbol{y} = \alpha + \beta \boldsymbol{x}$
- Reality: $y_i = \alpha + \beta x_i + \epsilon_i$

Random (residual) error from fit for each x_i

- Variance in $\boldsymbol{y}$ can be explained by
  - Variance in $\boldsymbol{x}$
  - Residual uncertainty (called $\boldsymbol{\epsilon}$)

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)



$$(f(x_i) - \bar{y})^2 + \left(y_i - f(x_i)\right)^2 = (y_i - \bar{y})^2$$

- Coefficient of Determination: ESS/TSS
  - Proportion of total variation in ***y*** explained by the model

Simple Linear Regression

- Coefficient of Determination: ESS/TSS
  - Proportion of total variation in ***y*** explained by the model
- Has another name…R-squared!

$$R^2 = \frac{ESS}{TSS}$$

- 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 <u>mean</u> predicted response

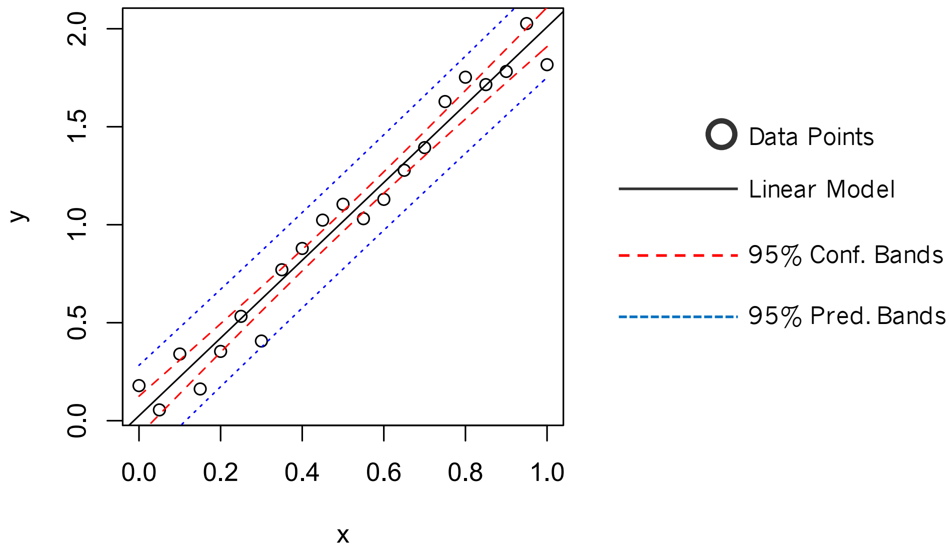
$$\mathrm{Var}\left(\hat{\alpha} + \hat{\beta}x_{new}\right)$$

- Prediction Bands
  - Variance associated with single new prediction
  - Takes into account residual errors in linear model

$$\mathrm{Var}\left(\hat{\alpha} + \hat{\beta}x_{new}\right) + \hat{\sigma}_{\epsilon}^{2}$$

(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?

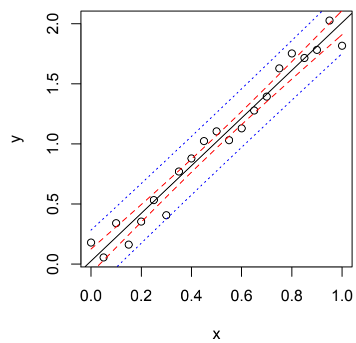
A plug for "R"…

- 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:
- Statistics: An Introduction using R by Michael J. Crawley, published 2005 by John Wiley & Sons, Ltd

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