BMTRY 701
Biostatistical Methods II

Elizabeth Garrett-Mayer, PhD
Associate Professor
Director of Biostatistics, Hollings Cancer Center
garrettm@musc.edu

Biostatistical Methods II

- **Description**: This is a one-semester course intended for graduate students pursuing degrees in biostatistics and related fields such as epidemiology and bioinformatics. Topics covered will include linear, logistic, poisson, and Cox regression. Advanced topics will be included, such as ridge regression or hierarchical linear regression if time permits. Estimation, interpretation, and diagnostic approaches will be discussed. Software instruction will be provided in class in R and Stata. **Students will be evaluated via homeworks (55%), two in-class exams (35%) and class participation (10%).** This is a four credit course.

- **Textbook**: Applied Linear Statistical Models. Kutner, Nachtsheim, Neter and Li. McGraw-Hill, Fifth Edition

- **Prerequisites**: Biometry 700

- **Course Objectives:** Upon successful completion of the course, the student will be able to
  - Apply, interpret and diagnose linear regression models
  - Apply, interpret and diagnose logistic, poisson and Cox regresssion models

Biostatistical Methods II

Instructor: Elizabeth Garrett-Mayer

Website: http://people.musc.edu/~elg26/teaching/methods2.2009/methods2.2009.htm

Contact Info: Hollings Cancer Center, Rm 118G

garrettm@musc.edu
(preferred mode of contact is email)

792-7764

Time: Mondays and Wednesdays, 1:30-3:30

Location: Cannon Place, Room 305V

Biostatistical Methods II

- Lecture schedule is on the website
- First time teaching this class
  - syllabus is a 'work in progress'
  - timing of topics subject to change
  - lectures may appear on website last-minute
- Computing
  - R
  - Stata
  - integrated into lecture time
- Homeworks, articles, datasets will also be posted to website
  - some/most problems will be from textbook
  - some datasets will be from textbook CD
- If you want printed versions of lectures:
  - download and print prior to lecture; OR
  - work interactively on your laptop during class
- We will take a break about halfway through each lecture

Expectations (from R. Carter)

- Academic
  - Participate in class discussions
  - Invest resources in YOUR education
  - Complete homework assignments on time
  - The results of the homework should be communicated so that a person knowledgable in the methodology could reproduce your results.
  - Create your own study groups
    - Challenge one another
    - everyone needs to contribute
    - you may do homeworks together, but everyone must turn in his/her own homework.
    - written sections of homework should be ‘independently’ developed
- General
  - Be on time to class
  - Be discrete with interruptions (pages, phones, etc.)
  - Do NOT turn in raw computer output

Other Expectations

- Methods I!
- You should be very familiar with
  - confidence intervals
  - hypothesis testing
    - t-tests
    - Z-tests
  - graphical displays of data
  - exploratory data analysis
    - estimating means, medians, quantiles of data
    - estimating variances, standard deviations

About the instructor

- B.A. from Bowdoin College, 1994
  - Double Major in Mathematics and Economics
  - Minor in Classics
- Ph.D. in Biostatistics from Johns Hopkins, 2000
  - Dissertation research in latent class models, Adviser Scott Zeger
- Assistant Professor in Oncology and Biostatistics at JHU, 2000-2007
- Taught course in Statistics for Psychosocial Research for 8 years
- Applied Research Areas:
  - oncology
- Biostats Research Areas:
  - latent variable modeling
  - class discovery in microarray data
  - methodology for early phase oncology clinical trials
- Came to MUSC in Feb 2007

Computing

- Who knows what?
- Who WANTS to know what?
- Who will bring a laptop to class?
- What software do you have and/or prefer?
- Should we get a lab classroom?

Regression

- Purposes of Regresssion
  1. Describe association between Y and X's
  2. Make predictions:
     - Interpolation: making prediction within a range of X's
     - Extrapolation: making prediction outside a range of X's
  3. To "adjust" or "control for" confounding variables
- What is "Y"?
  - an outcome variable
  - 'dependent' variable
  - 'response'
- Type of regression depends on type of Y
  - continuous (linear regression)
  - binary (logistic regression)
  - time-to-event (Cox regression)
  - rare event or rate (poisson regression)

Some motivating examples

- Example 1: Suppose we are interested in studying the relationship between fasting blood glucose (FBG) levels and the number hours per day of aerobic exercise. Let Y denote the fasting blood glucose level
  - Let X denote the number of hours of exercise
  - One may be interested in studying the relationship of Y and X
- Simple linear regression can be used to quantify this relationship

Some motivating examples

- Example 2: Consider expanding example 1 to include other factors that could be related FBG.
- Let X1 denote hours of exercise
- Let X2 denote BMI
- Let X3 indicate if the person has diabetes
- . . . (other covariates possible)
- One may be interested in studying the relationship of all X′s on Y and identifying the “best” combination of factors
- Note: Some of the X′s may correlated (e.g., exercise and bmi)
- Multiple (or multivariable, not multivariate) linear regression can be used to quantify this relationship

Some motivating examples

- Example 3: Myocardial infarction (MI, heart attack) is often a life-altering event
- Let Y denote the occurrence (Y = 1) of an MI after treatment, let Y = 0 denote no MI
- Let X1 denote the dosage of aspirin taken
- Let X2 denote the age of the person
- . . . (other covariates possible)
- One may be interested in studying the relationship of all X′s on Y and identifying the “best” combination of factors
- Multiple LOGISTIC regression can be used to quantify this relationship

More motivating examples

- Example 4: This is an extension of Ex 3: Myocardial infarction. Let the interest be now on when the first
- MI occurs instead of “if” one occurs.
- Let Y denote the occurrence (Y = 1) of an MI after treatment, let Y = 0 denote no MI observed
- Let Time denote the length of time the individual is observed
- Let X1 denote the dosage of aspirin taken
- . . . (other covariates possible)
- Survival Analysis (which, in some cases, is a regression model) can be used to quantify this relationship of aspirin on MI

More motivating examples

- Example 5: Number of cancer cases in a city
- Let Y denote the count (non-negative integer value) of cases of a cancer in a particular region of interest
- Let X1 denote the region size in terms of “at risk” individuals
- Let X2 denote the region
- . . . (other covariates possible)
- One may be interested in studying the relationship of the region on Y while adjusting for the population at risk sizes
- POISSON regression can be used to quantify this relationship

Brief Outline

- Linear regression: half semester (through spring break)
- Logistic regression
- Poisson regression
- Cox regression (survival)
- Hierarchical regression or ridge regression?

Linear Regression

- Outcome is a CONTINUOUS variable
- Assumes association between Y and X is a ‘straight line’
- Assumes relationship is ‘statistical’ and not ‘functional’
  - relationship is not perfect
  - there is ‘error’ or ‘noise’ or ‘unexplained variation’
- Aside:
  - I LOVE graphical displays of data
  - This is why regression is especially fun
  - there are lots of neat ways to show your data
  - prepare yourself for a LOT of scatterplots this semester

Graphical Displays

- Scatterplots: show associations between two variables (usually)
- Also need to understand each variable by itself
- Univariate data displays are important
- Before performing a regresssion, we should
  - identify any potential skewness
  - outliers
  - discreteness
  - multimodality
- Top choices for univariate displays
  - boxplot
  - histogram
  - density plot
  - dot plot

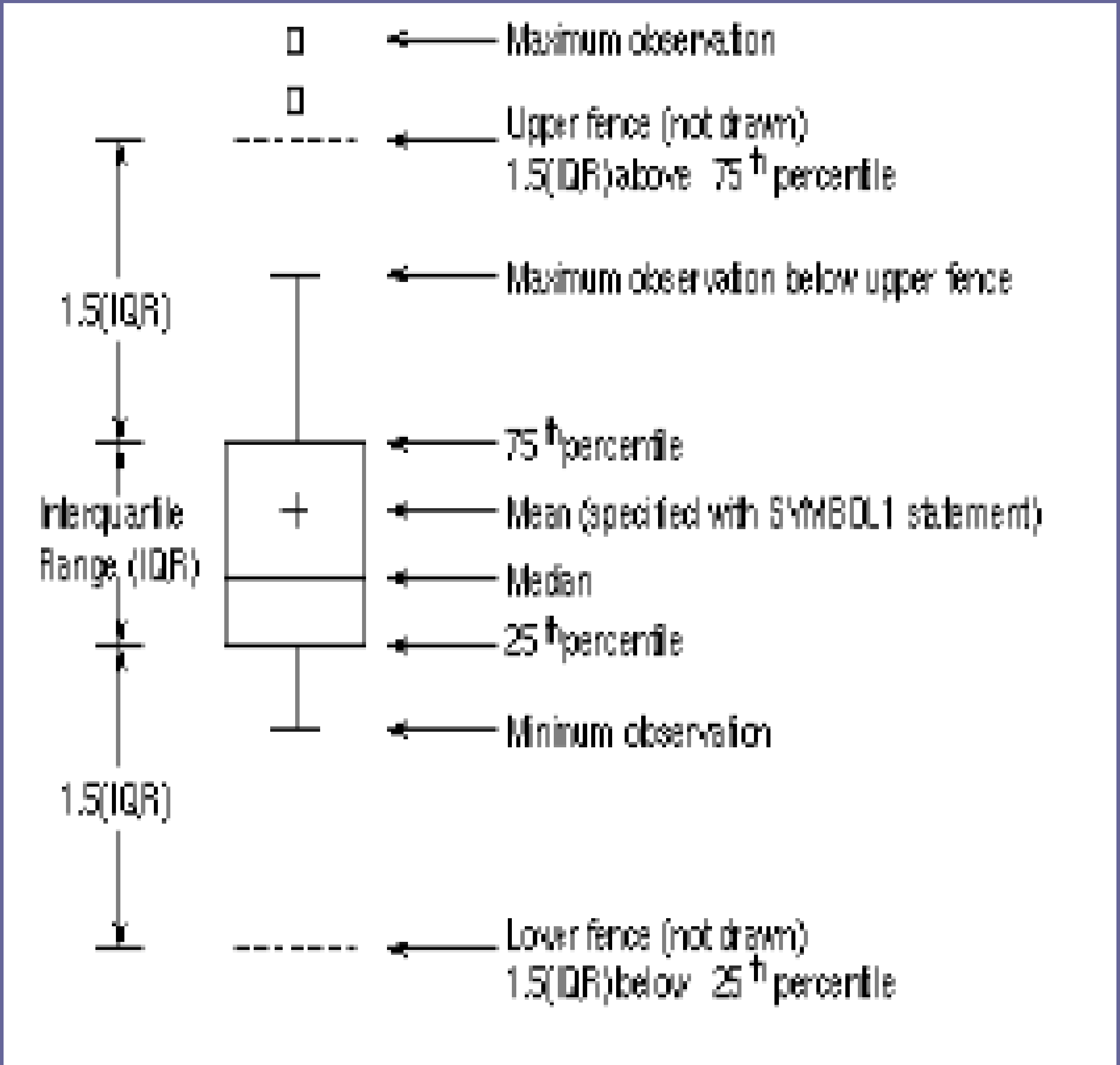
Linear regression example

- The authors conducted a pilot study to assess the use of toenail arsenic concentrations as an indicator of ingestion of arsenic-containing water. Twenty-one participants were interviewed regarding use of their private (unregulated) wells for drinking and cooking, and each provided a sample of water and toenail clippings. Trace concentrations of arsenic were detected in 15 of the 21 well-water samples and in all toenail clipping samples.
- Karagas MR, Morris JS, Weiss JE, Spate V, Baskett C, Greenberg ER. Toenail Samples as an Indicator of Drinking Water Arsenic Exposure. Cancer Epidemiology, Biomarkers and Prevention 1996;5:849-852.

Purposes of Regression

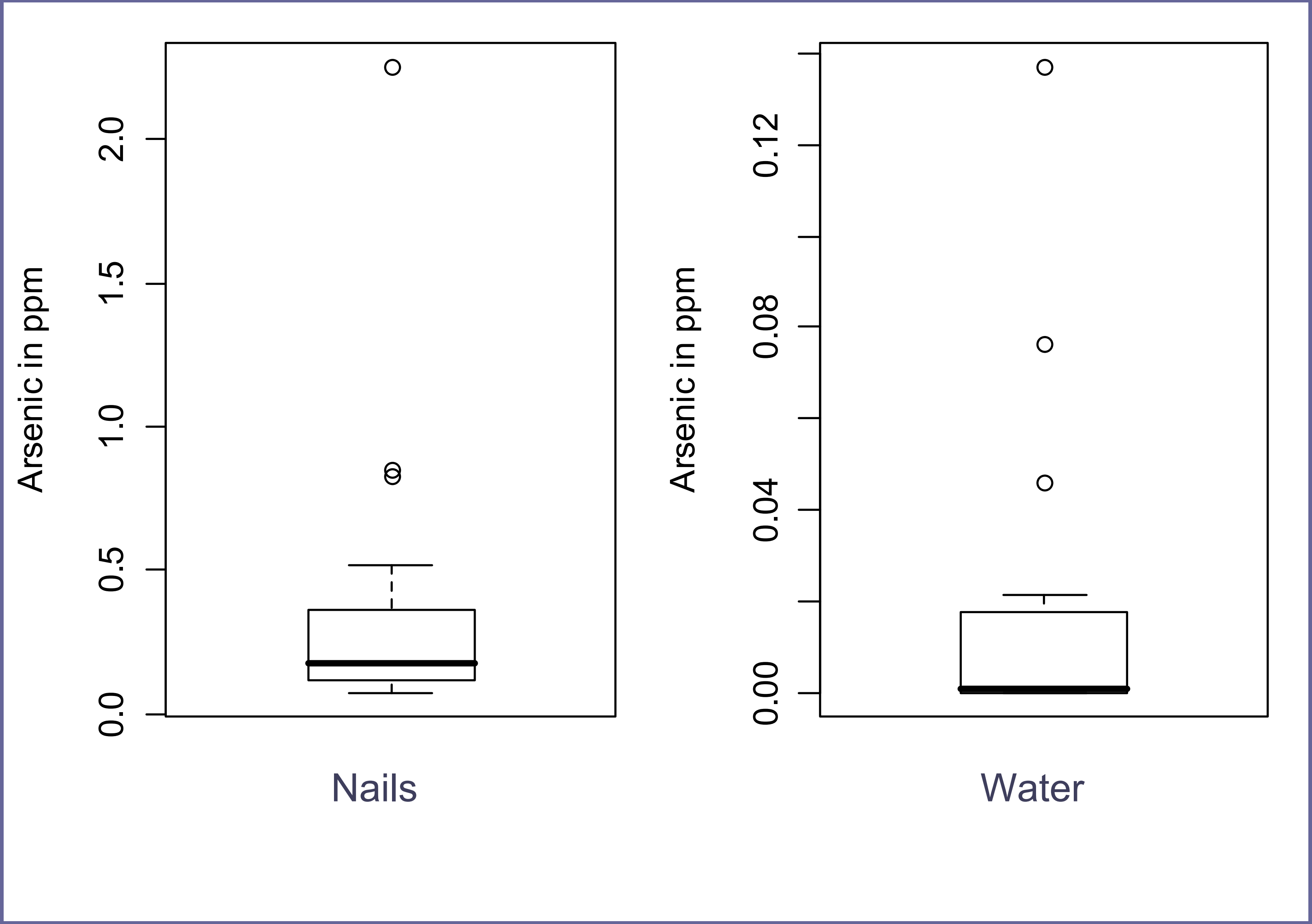
1. Describe association
   - hypothesis: as arsenic in well water increases, level of arsenic in nails also increases.
   - linear regression can tell us
     - how much increase in nail level we see on average for a 1 unit increase in well water level of arsenic
2. Predict
   - linear regresssion can tell us
     - what level of arsenic we would expect in nails for a given level in well water.
     - how precise our estimate of arsenic is for a given level of well water
3. Adjust
   - linear regression can tell us
     - what the association between well water arsenic and nail arsenic is adjusting for other factors such as age, gender, amount of use of water for cooking, amount of use of water for drinking.

Boxplot



Maximum observation
Upper fence (not drawn)
1.5(IQR) above 75 th percentile
Maximum observation below upper fence
75 th percentile
Mean (specified with SYMBOL1 statement)
Median
25 th percentile
Minimum observation
Lower fence (not drawn)
1.5(IQR) below 25 th percentile
1.5(IQR)
Interquartile Range (IQR)
1.5(IQR)

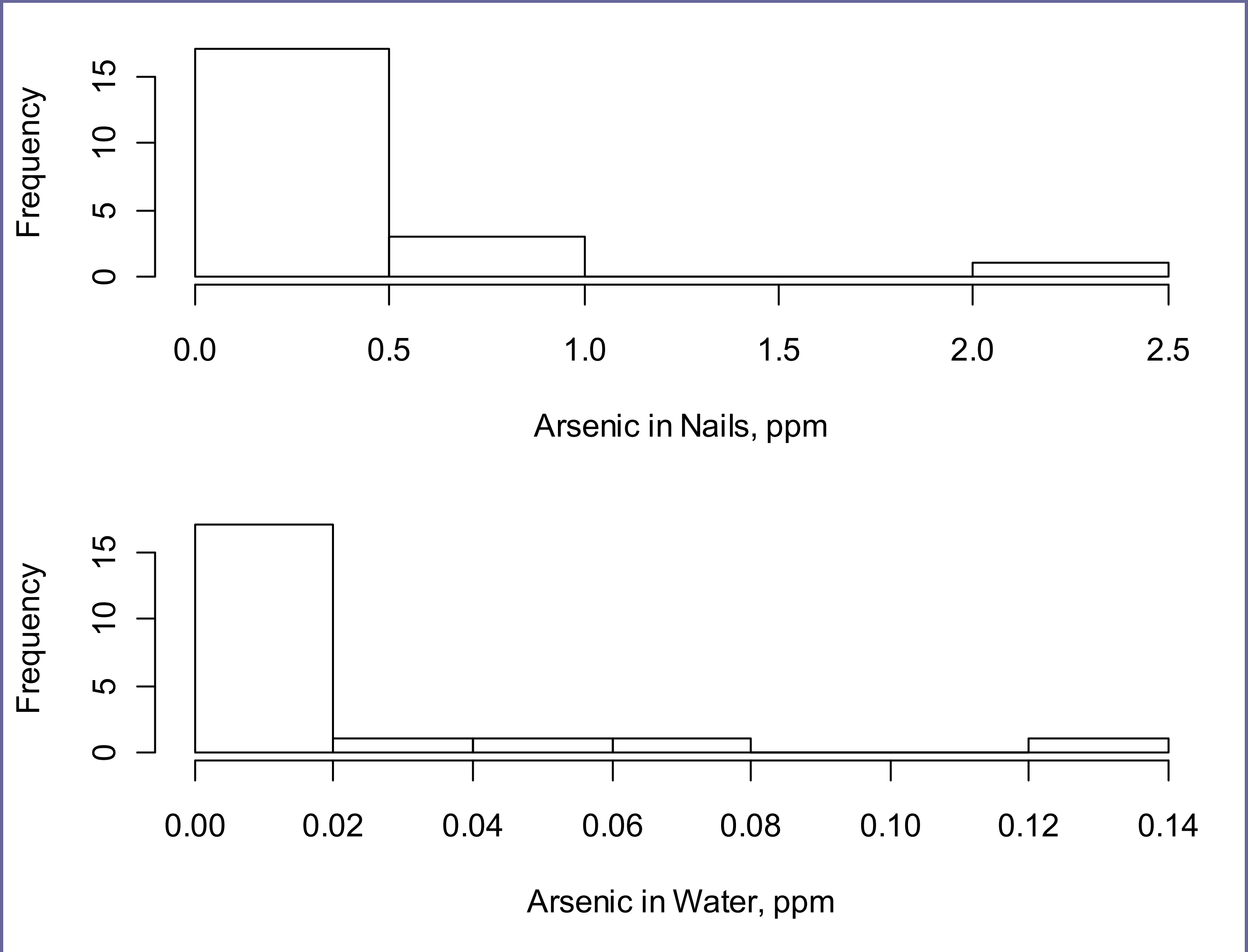
Graphical Displays



Histogram

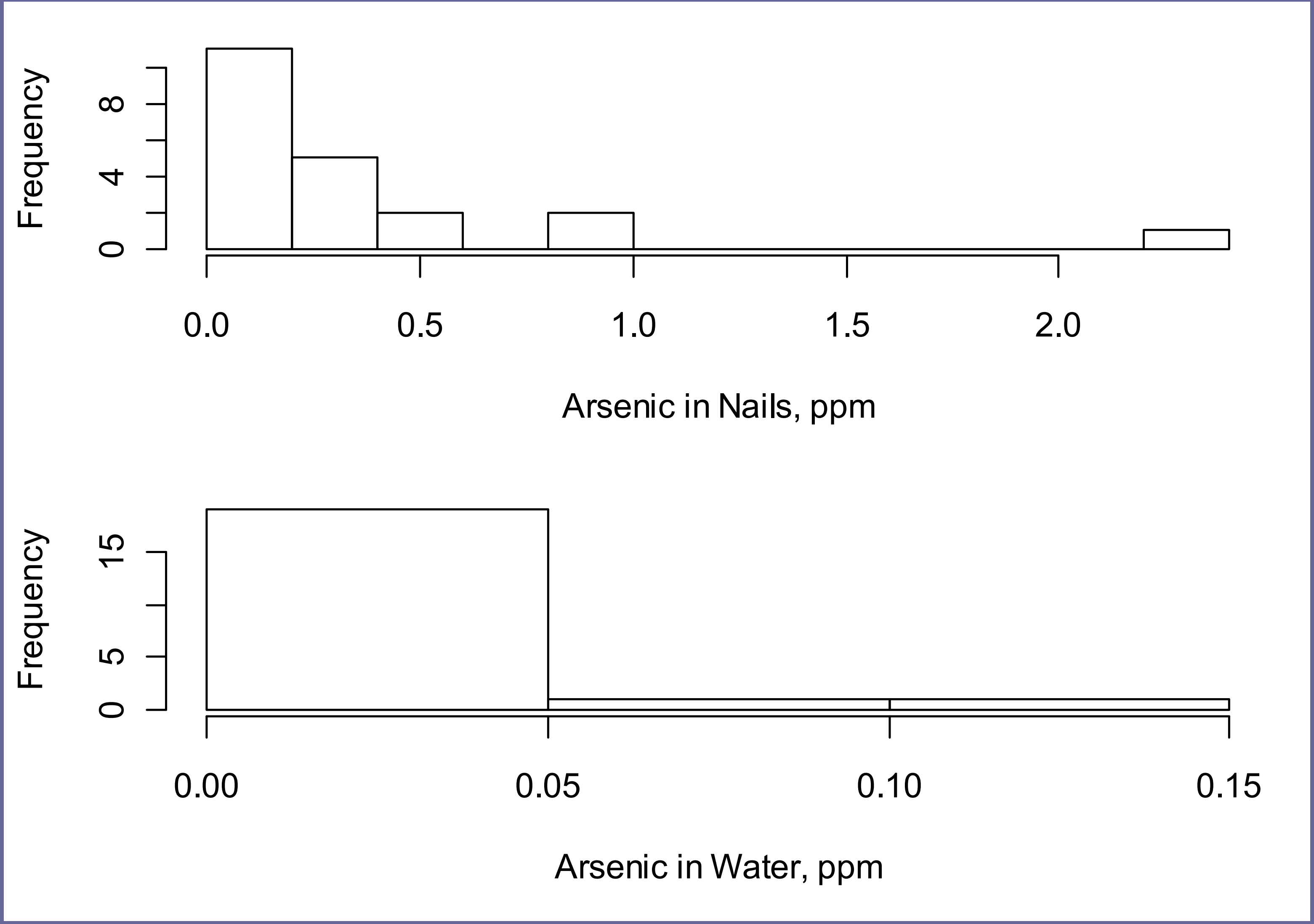
- Bins the data
- x-axis represents variable values
- y-axis is either
  - frequency of occurrence
  - percentage of occurence
- Visual impression can depend on bin width
- often difficult to see details of highly skewed data

Histogram



Frequency
0
5
10
15
0.0
0.5
1.0
1.5
2.0
2.5
Arsenic in Nails, ppm
Frequency
0
5
10
15
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Arsenic in Water, ppm

Histogram

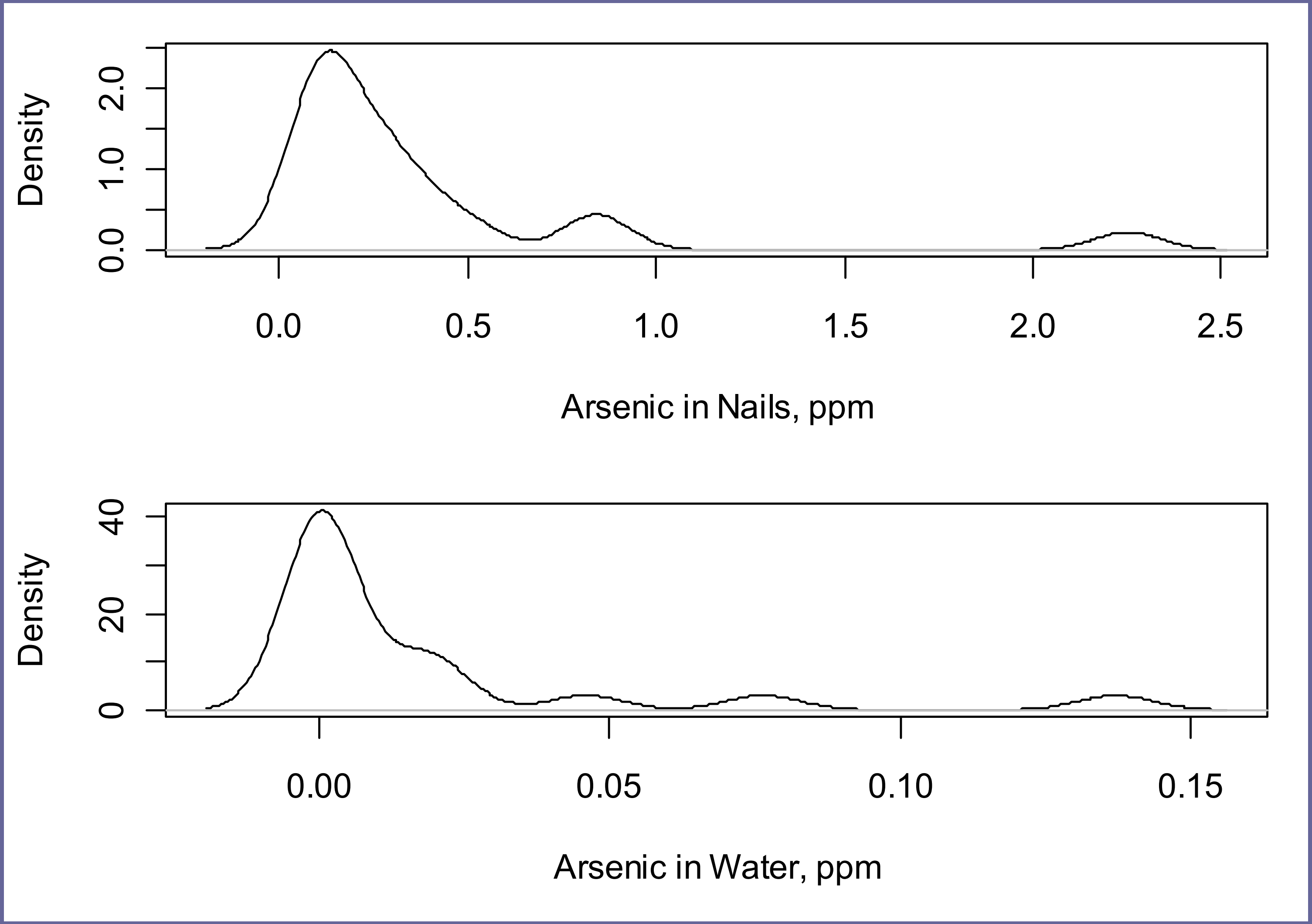


Frequency
0
4
8
0.0
0.5
1.0
1.5
2.0
Arsenic in Nails, ppm
Frequency
0
5
15
0.00
0.05
0.10
0.15
Arsenic in Water, ppm

Density Plot

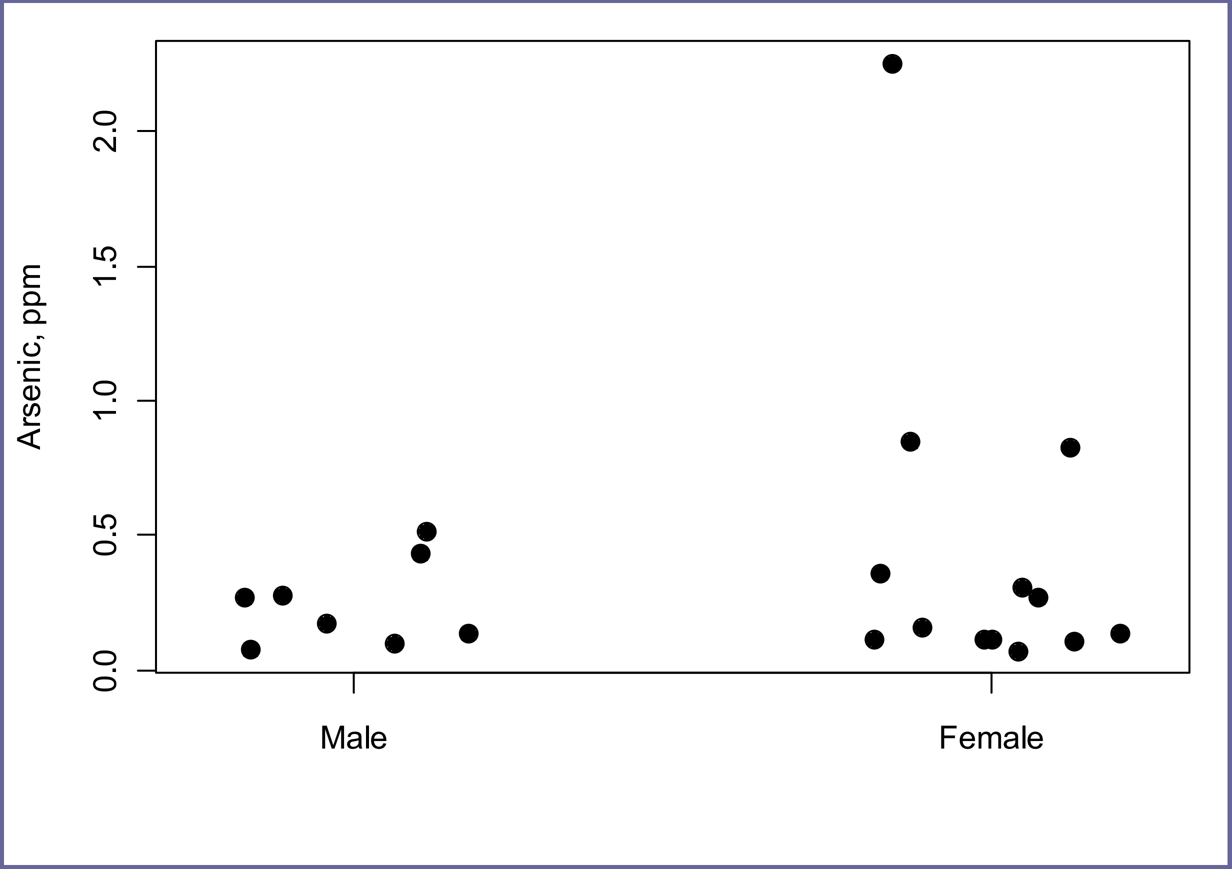
- Smoothed density based on kernel density estimates
- Can create similar issues as histogram
  - smoothing parameter selection
  - can affect inferences
- Can be problematic for ‘ceiling’ or ‘floor’ effects

Density Plot

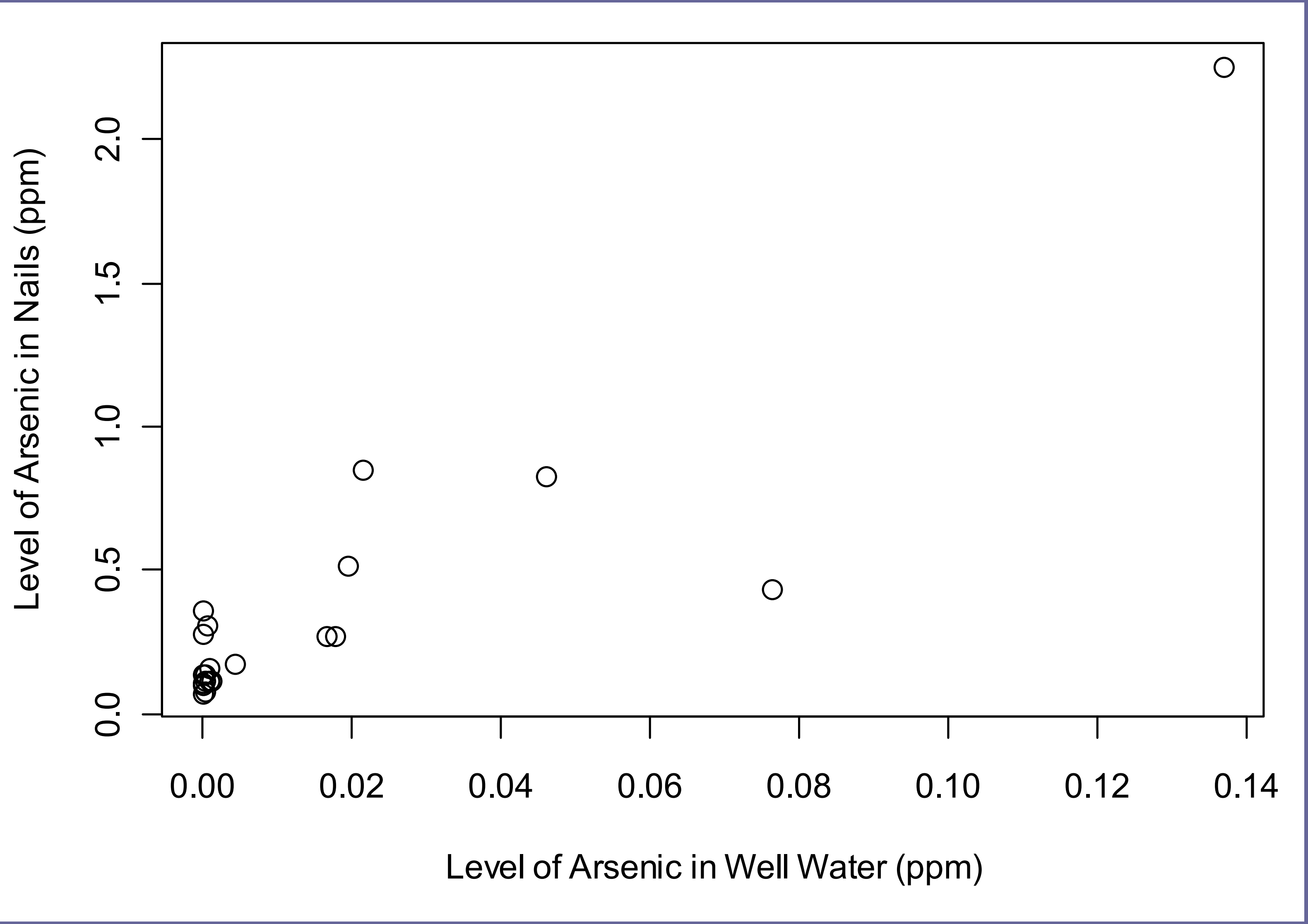


Dot plot

- My favorite for
  - small datasets
  - when displaying data by groups



And…the scatterplot



Measuring the association between X and Y

- Y is on the vertical
- X predicts Y
- Terminology:
  - "Regress Y on X"
  - Y: dependent variable, response, outcome
  - X: independent variable, covariate, regressor, predictor, confounder
- Linear regression → a straight line
  - important!
  - this is key to ***linear*** regression

Simple vs. Multiple linear regresssion

- Why ‘simple’?
  - only one “x”
  - we’ll talk about multiple linear regression later…
- Multiple regression
  - more than one “X”
  - more to think about: selection of covariates
- Not linear?
  - need to think about transformations
  - sometimes linear will do reasonably well

Association versus Causation

- Be careful!
- Association $\neq$ Causation
- Statistical relationship does not mean X causes Y
- Could be
  - X causes Y
  - Y causes X
  - something else causes both X and Y
  - X and Y are spuriously associated in your sample of data
- Example: vision and number of gray hairs

Basic Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Y_i is the value of the response variable in the ith individual
- β_0 and β_1 are parameters
- X_i is a known constant; the value of the covariate in the *ith* individual
- ε_i is the random error term
- Linear in the parameters
- Linear in the predictor

Basic Regression Model

- NOT linear in the parameters:

$$Y_i = \beta_0 + X_i^{\beta_1} + \varepsilon_i$$
$$Y_i = \beta_0 + \beta_1 \beta_2 X_i + \varepsilon_i$$

- NOT linear in the predictor:

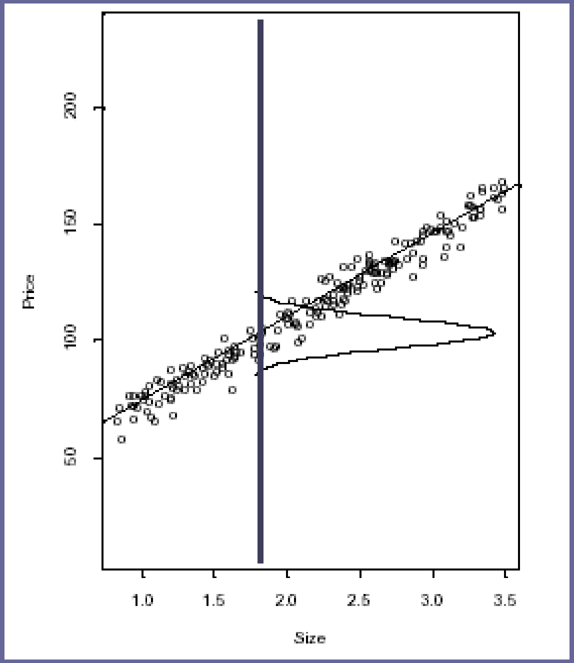
$$\log(Y_i) = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$Y_i^2 = \beta_0 + \beta_1 X_i + \varepsilon_i$$

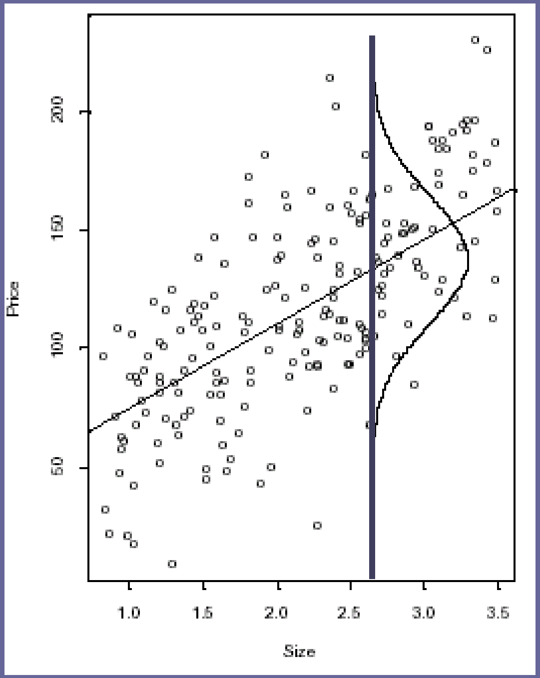
Model Features

- Y_i is the sum of a constant piece and a random piece:
  - $\beta_0 + \beta_1 X_i$ is constant piece (recall: x is treated as constant)
  - ε_i is the random piece
- Attributes of error term
  - mean of residuals is 0: $E(\varepsilon_i) = 0$
  - constant variance of residuals : $\sigma^2(\varepsilon_i) = \sigma^2$ for all i
  - residuals are uncorrelated: $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ for all i, j; $i \neq j$
- Consequences
  - Expected value of response
    - $E(Yi) = \beta_0 + \beta_1 X_i$
    - $E(Y) = \beta_0 + \beta_1 X$
  - Variance of $Y_i = \sigma^2$
  - Y_i and Y_j are uncorrelated

Probability Distribution of Y

- For each level of X, there is a probability distribution of Y
- The means of the probability distributions vary systematically with X.





Parameters

- β_0 and β_1 are referred to as "regression coefficients"
- Remember y = mx+b?
- β_1 is the slope of the regression line
  - the expected increase in Y for a 1 unit increase in X
  - the expected difference in Y comparing two individuals with X's that differ by 1 unit
  - Expected? Why?

Parameters

- β_0 is the intercept of the regression line
  - The expected value of Y when X = 0
  - Meaningful?
    - when the range of X includes 0, yes
    - when the range of X excluded 0, no
  - Example:
    - Y = baby's weight in kg
    - X = baby's height in cm
    - β_0 is the expected weight of a baby whose height is 0 cm.

SENIC Data

- Will be used as a recurring example
- SENIC = Study on the Efficacy of Nosocomial Infection Control
- The primary objective of the SENIC Project was to determine whether infection surveillance and control programs have reduced the rates of nosocomial (hospital-acquired) infection in the United States hospitals.
- This data set consists of a random sample of 113 hospitals selected from the original 338 hospitals surveyed.
- Each line of the data set has an ID number and provides information on 11 other variables for a single hospital.
- The data used here are for the 1975-76 study period.

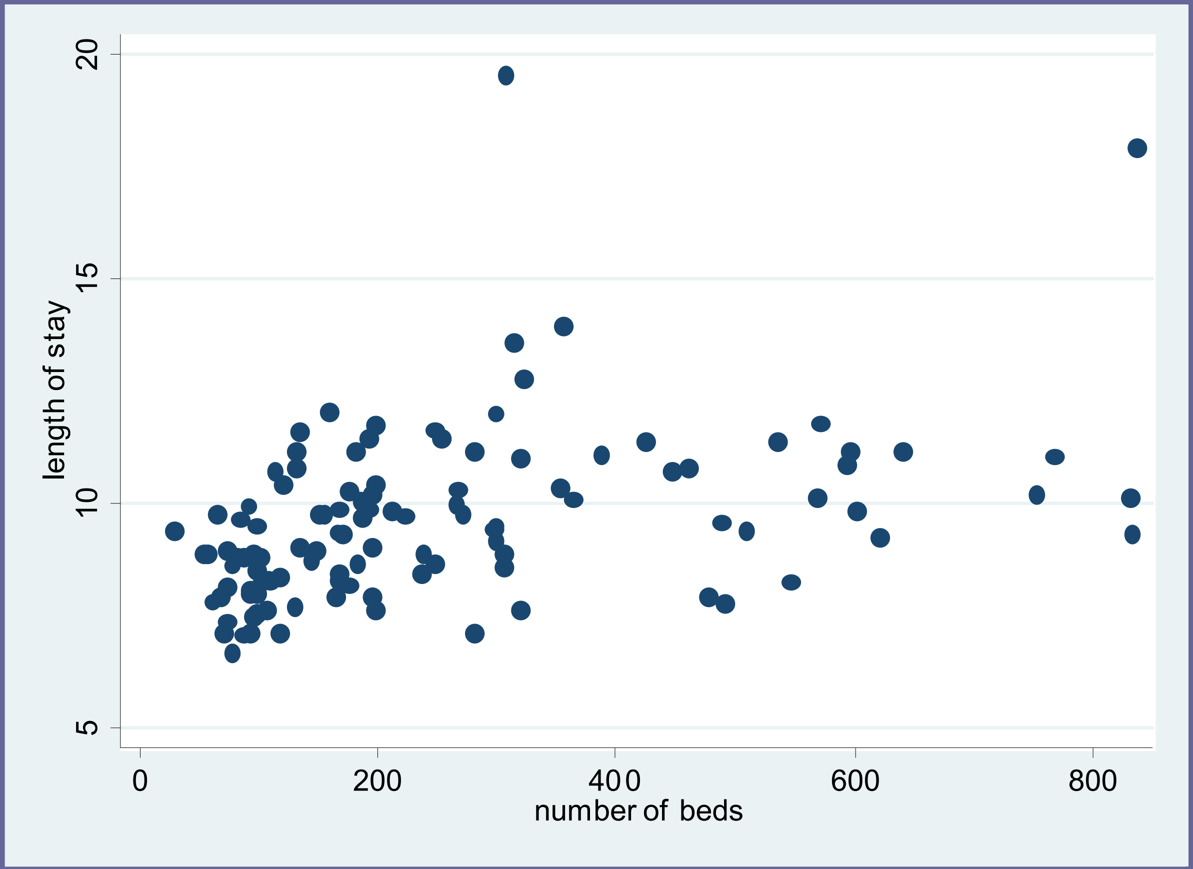
SENIC Data

ID	ID Number	1-113
LOS	Length of stay	Average length of stay of all Patients in hospital (in days)
AGE	Age	Average age of patients (in years)
INFRISK	Risk of infection	Average estimated probability of Acquiring infection in hospital (in percent)
CULT	Routine culturing ratio	Ratio of number of cultures performed to number of patients without signs or symptoms of hospital-acquired infection, times 100
XRAY	Routine chest X-ray ratio	Ratio of number of X-rays performed to number of patients without signs or symptoms of pneumonia, times 100
BEDS	Number of beds	Average number of beds in hospital in study period
MEDSCHL	Medical school affiliation	1 = Yes, 2 = No
REGION	Region	Geographic region, where: 1 = NE, 2 = NC, 3 = S, 4 = W
CENSUS	Average daily census	Average number of patients in hospital per day during study period
NURSE	Number of nurses	Average number of full-time equivalent registered and licensed practical nurses during study period (number full-time plus one half the the number part-time)
FACS	Available facilities and services	Percent of 35 potential facilities and services that are provided by the hospital

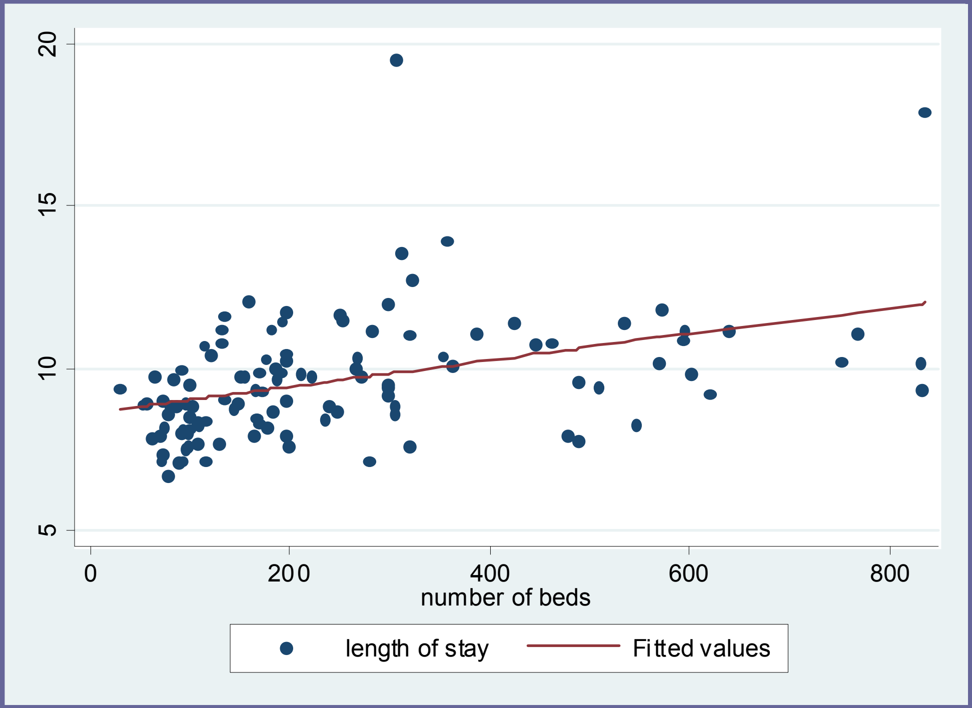
SENIC Simple Linear Regression Example

- Hypothesis: The number of beds in a given hospital is associated with the average length of stay.
- Y = ?
- X = ?
- Scatterplot

. scatter los beds



Stata Regression Results




```
. regress los beds

      Source |       SS       df       MS              Number of obs =     113
-------------+------------------------------           F(  1,   111) =   22.33
       Model |  68.5419355     1  68.5     355           Prob > F      =  0.0000
    Residu   |  340.668443   111  3.0      08            R-squared     =  0.1675
-------------+------------------------------           Adj R-squared =  0.1600
       To    |  409.210379   112  3.6536641            Root MSE      =  1.7519

------------------------------------------------------------------------------
         los |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        beds |   .0040566   .0008584     4.73   0.000     .0023556    .0057576
       _cons |   8.625364   .2720589    31.70   0.000     8.086262    9.164467
------------------------------------------------------------------------------
```

$\hat{\beta}_1$ → .0040566

$\hat{\beta}_0$ → 8.625364

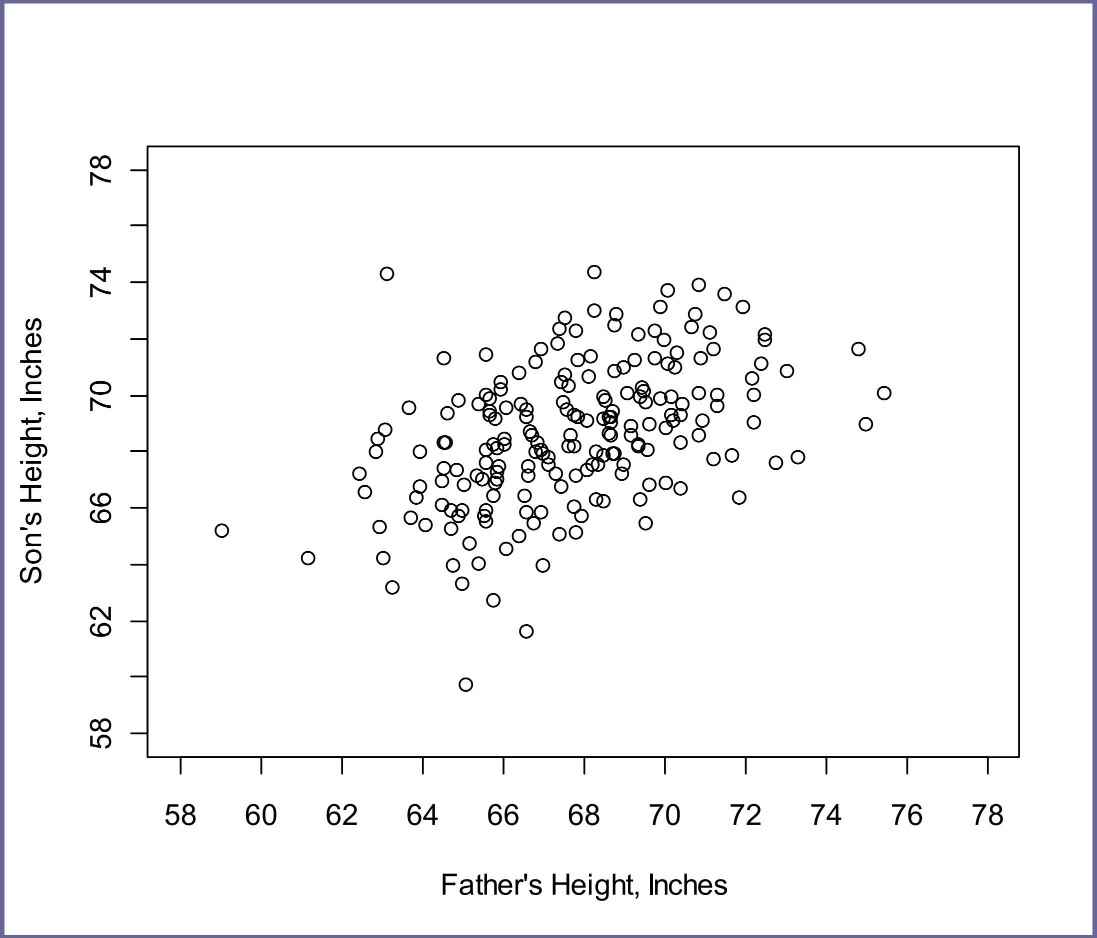
$\hat{\sigma}$ → 1.7519

Another Example: Famous data

- Father and sons heights data from Karl Pearson (over 100 years ago in England)
- 1078 pairs of fathers and sons
- Excerpted 200 pairs for demonstration
- Hypotheses:
  - there will be a positive association between heights of fathers and their sons
  - very tall fathers will tend to have sons that are shorter than they are
  - very short fathers will tend to have sons that are taller than they are

Scatterplot of 200 records of father son data

```
plot(father, son, xlab="Father's Height, Inches", ylab="Son's Height, Inches",
        xaxt="n",yaxt="n",ylim=c(58,78), xlim=c(58,78))
axis(1, at=seq(58,78,2))
axis(2, at=seq(58,78,2))
```



Regression Results

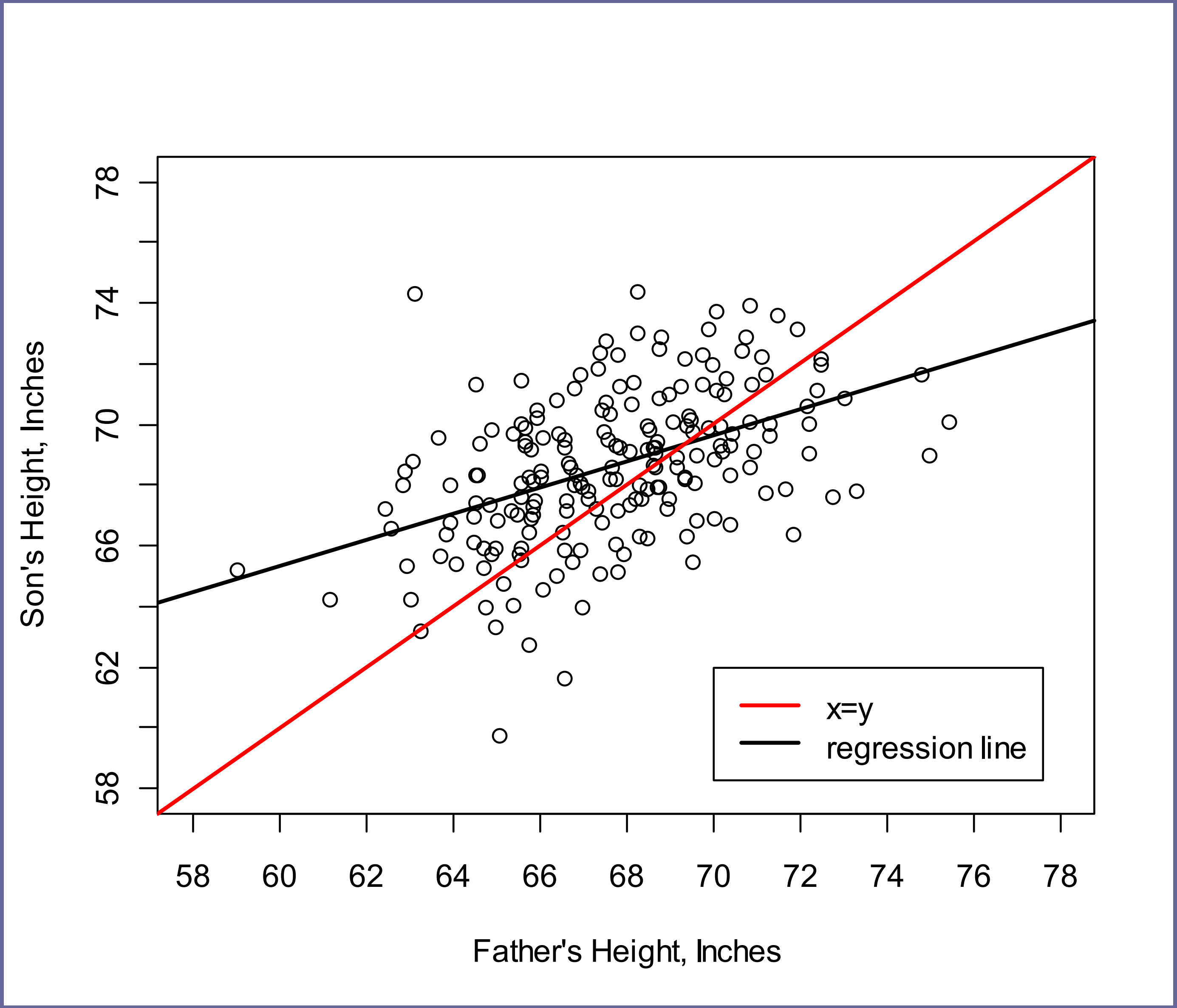
```
> reg <- lm(son~father)
> summary(reg)

Call:
lm(formula = son ~ father)

Residuals:
     Min       1Q   Med                3Q      Max
-7.72874 -1.39750 -0.04029  1.51871  7.66058

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.47177    3.96188   9.963  < 2e-16 ***
father       0.43099    0.05848   7.369 4.55e-12 ***
---
Signi  codes:  0 '***' 0.001 '**' 0.  '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.233 on 198 degrees of freedom
Multiple R-squared: 0.2152,     Adjusted R-squared: 0.2113
F-statistic: 54.31 on 1 and 198 DF,  p-value: 4.549e-12
```

$\hat{\beta}_0$

$\hat{\beta}_1$

$\hat{\sigma}$

This is where the term "regression" came from



$$\bar{x} = 67.7$$

$$\bar{y} = 68.6$$

Aside: Design of Studies

- Does it matter if the study is randomized? observational?
- Yes and no
- Regression modeling can be used regardless
- The model building will often depend on the nature of the study
- Observational studies:
  - adjustments for confounding
  - often have many covariates as a result
- Randomized studies:
  - adjustments may not be needed due to randomization
  - subgroup analyses are popular and can be done via regression

Estimation of the Model

- The Method of Least Squares
- Intuition: we would like to minimize the residuals:

$$\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$$

- Minimize/maximize: how to do that?
- Can we minimize the sum of the residuals?

Least Squares

- Minimize the distance between the fitted line and the observed data
- Take absolute values?
- Simpler? Square the errors.
- LS estimation:
  - Minimize Q:

$$Q = \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Least Squares

- Derivation
- Two initial steps: reduce the following

$$\sum_{i=1}^{N}(X_i - \overline{X})^2$$

$$\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})$$

Least Squares

$$Q = \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Least Squares