

---

# Thought leaders in data science and analytics: Linear Regression

**James G. Shanahan<sup>1</sup>**

***<sup>1</sup>Independent Consultant***

***EMAIL: James\_DOT\_Shanahan\_AT\_gmail\_DOT\_com***

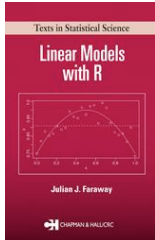
**I 296A UC Berkeley**

**Lecture 3 , Wednesday February 1, 2012**

# General Course References (Advanced)

- **R**

- Practical Regression and Anova using R, <http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf>, by JJ Faraway (please download PDF)



- John Fox (2010), Sage, [An R and S-PLUS Companion to Applied Regression](#) (second edition, PDFs)
  - [Preface to the book](#), [Chapter 1 - Getting Started With R](#) (PDFs available)
  - [Chapter 6 - Diagnosing Problems in Linear and Generalized Linear Models](#)
- The R book, by Michael J. Crawley, Wiley 2009

- **Linear Regression**

- Analyzing Multivariate Data by James Lattin, J. Douglas Carroll, Paul E. Green. Thompson 2003. ISBN: 0-534-349749
- *Introduction to Linear Regression Analysis*. D. Montgomery, E. Peck. GG Vining (4<sup>th</sup> Edition)

- **Data mining**

- TSK [Introduction to Data Mining](#), Pang-ning Tan, Michael Steinbach, Vipin Kumar. Addison Wesley 2005. ISBN: 0-321-32136-7

- **Machine Learning, probability theory**

- Duda, Hart, & Stork (2000). Pattern Classification. <http://rii.ricoh.com/~stork/DHS.html>
- Modern Multivariate Statistical Techniques: Regression, Classification, and manifold Learning, Alan Julian Izenman, Springer, 2008, ISBN 978-0-387-78188-4
- *Pattern Recognition and Machine Learning*, Christopher M. Bishop, Springer
- Elements of Machine Learning, Friedman et al., 2009, Download from here <http://www-stat.stanford.edu/~tibs/ElemStatLearn/download.html>

- **General AI**

- [Artificial Intelligence: A Modern Approach](#) (Third edition) by Stuart Russell and Peter Norvig.

# Lecture Outline

---

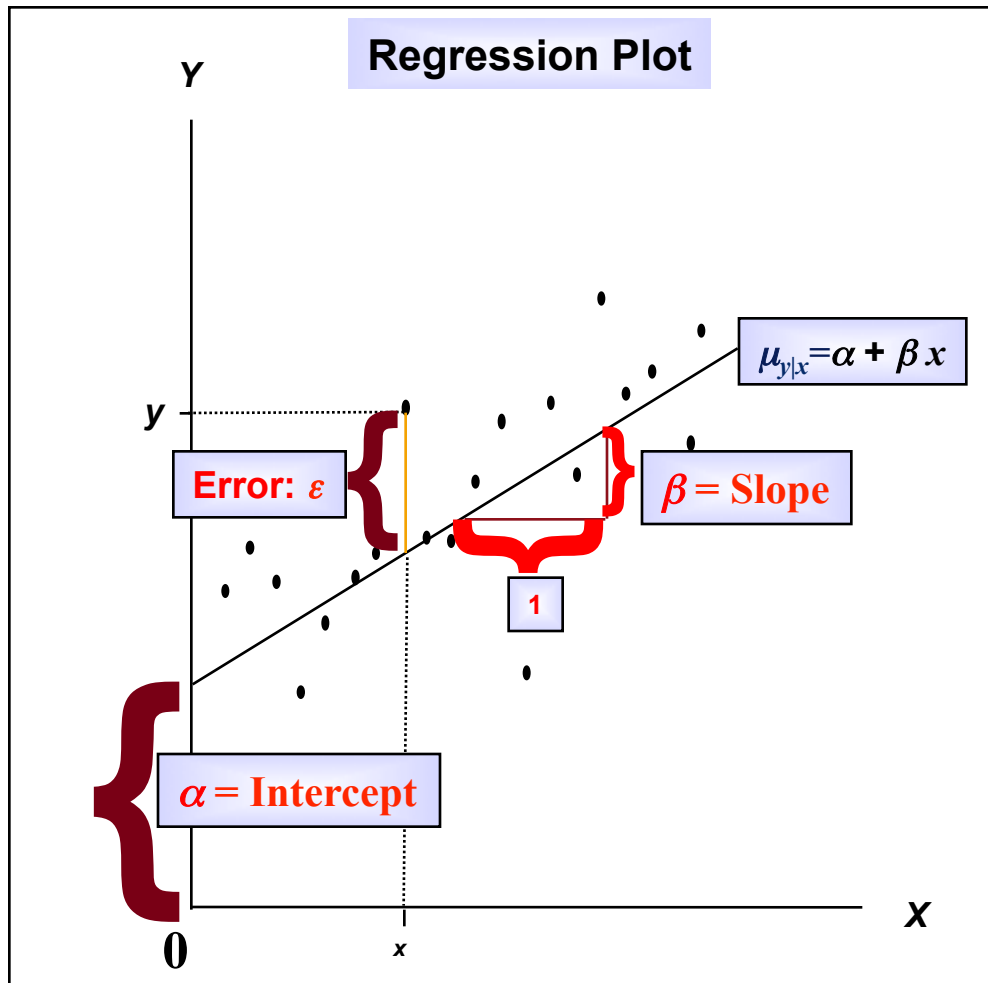
- **Linear Regression: a brief intro**
- **A quick statistics review**
  - Mean, expected value, variance, stdev, quantiles, stats in R
- **Locally Weighted Linear Regression**
- **Exploratory Data Analysis**
- **Simple Linear Regression**
  - Normal Equations
  - Closed form Solution
  - Variance of the estimators
- **Good model?**

# Regression and Model Building

---

- **Regression analysis is a statistical technique for investigating and modeling the relationship between variables.**
  - Assume two variables,  $x$  and  $y$ . Model relationship as  $y \sim x$  (aka  $y = f(x)$ ) as a linear relationship
    - $y = \beta_0 + \beta_1 x$
  - Not a perfect fit generally; Account for difference between model prediction and the actual target value as a statistical error  $\varepsilon$ 
    - $y = \beta_0 + \beta_1 x + \varepsilon$  #This is a linear regression model
  - This error  $\varepsilon$  maybe made up of the effects of other variables, measurement errors and so forth
  - Customarily  $x$  is called the independent variable (aka predictor or regressor) and  $y$  the dependent variable (aka response variable)
  - Simple linear regression involves only one regressor variable
  - Suppose we can fix the value of  $x$  and observe the corresponding value of the response  $y$ . Now if  $x$  is fixed, the random component  $\varepsilon$  determines the properties of  $y$

# Simple Linear Regression Model



The simple linear regression model posits an exact linear relationship between the expected or average value of Y, the dependent variable Y, and X, the independent or predictor variable:

$$\mu_{y|x} = \alpha + \beta x$$

Actual observed values of Y ( $y$ ) differ from the expected value ( $\mu_{y|x}$ ) by an unexplained or random error ( $\epsilon$ ):

$$\begin{aligned} y &= \mu_{y|x} + \epsilon \\ &= \alpha + \beta x + \epsilon \end{aligned}$$

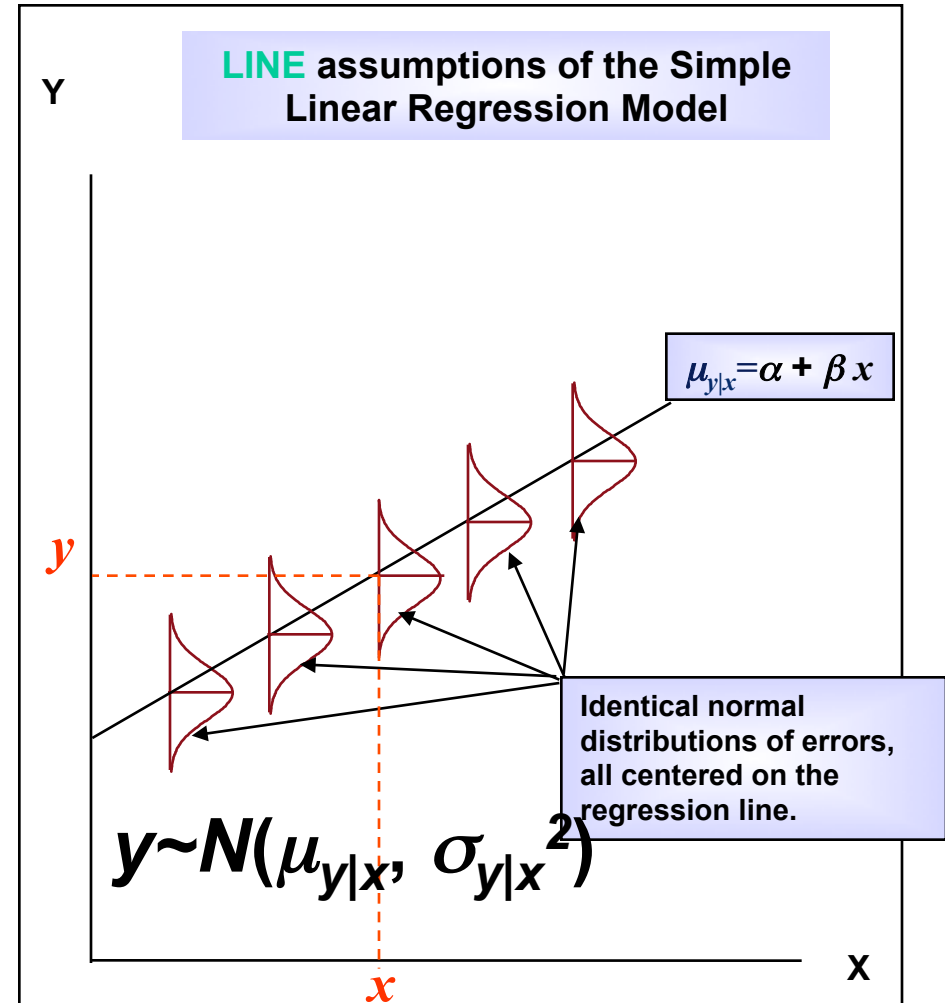
# $\varepsilon$ determines the properties of the response $y$

---

- Suppose we can fix the value of  $x$  and observe the corresponding value of the response  $y$ . Now if  $x$  is fixed, the random component  $\varepsilon$  determines the properties of  $y$ .
- Suppose the mean and variance of  $\varepsilon$  are 0 and  $\sigma^2$ , respectively. Then the mean response at any value of the regressor variable ( $x$ ) is
  - $E(y|x) = \mu_{y|x} = E(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1 x$
- The variance of  $y$  given any value  $x$  is
  - $\text{Var}(y|x) = \sigma_{y|x}^2 = \text{Var}(\beta_0 + \beta_1 x + \varepsilon) = \sigma^2$
- The variability of  $y$  at a particular value of  $x$  is determined by the variance of the error component of the model  $\sigma^2$ . This implies that there is a distribution of  $y$  values at each  $x$  and the variance of this distribution is the same at each  $x$
- Small  $\sigma^2$  implies the observed values  $y$  will fall close to the line.

# Assumptions of the Simple Linear Regression Model

- The relationship between  $X$  and  $Y$  is a straight-Line (linear) relationship.
- The values of the independent variable  $X$  are assumed fixed (not random); the only randomness in the values of  $Y$  comes from the error term  $\varepsilon$ .
- The errors  $\varepsilon$  are uncorrelated (i.e. Independent) in successive observations. The errors  $\varepsilon$  are Normally distributed with mean 0 and variance  $\sigma^2$  (Equal variance). That is:  $\varepsilon \sim N(0, \sigma^2)$



# Example

---

- Let  $y$  be a student's college achievement, measured by his/her **GPA**. This might be a function of several variables:
  - $x_1$  = rank in high school class
  - $x_2$  = high school's overall rating
  - $x_3$  = high school GPA
  - $x_4$  = SAT scores
- We want to predict  $y$  using knowledge of  $x_1$ ,  $x_2$ ,  $x_3$  and  $x_4$ .



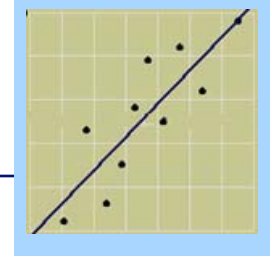
# Some Questions

---

- **Which of the independent variables are useful and which are not?**
- **How could we create a prediction equation to allow us to predict  $y$  using knowledge of  $x_1$ ,  $x_2$ ,  $x_3$  etc?**
- **How good is this prediction?**

**We start with the simplest case, in which the response  $y$  is a function of a single independent variable,  $x$ .**

# A Simple Linear Model



- We use the equation of a line to describe the relationship between  $y$  and  $x$  for a sample of  $n$  pairs,  $(x, y)$ .
- If we want to describe the relationship between  $y$  and  $x$  for the whole population, there are two models we can choose

• **Deterministic Model:**  $y = \beta_0 + \beta_1 x$

• **Probabilistic Model:**

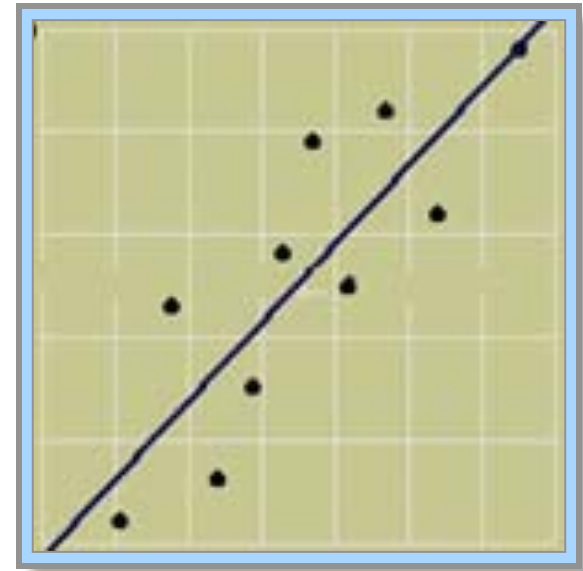
–  $y = \text{deterministic model} + \text{random error}$

$$-y = \beta_0 + \beta_1 x + \varepsilon$$

# A Simple Linear Model

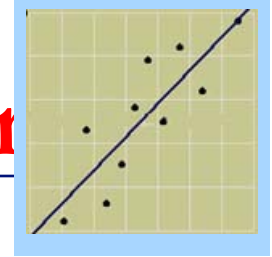
---

- **Since the measurements that we observe do not generally fall exactly on a straight line, we choose to use:**
- **Probabilistic Model:**
  - $y = \beta_0 + \beta_1 x + \varepsilon$
  - $E(y) = \beta_0 + \beta_1 x$

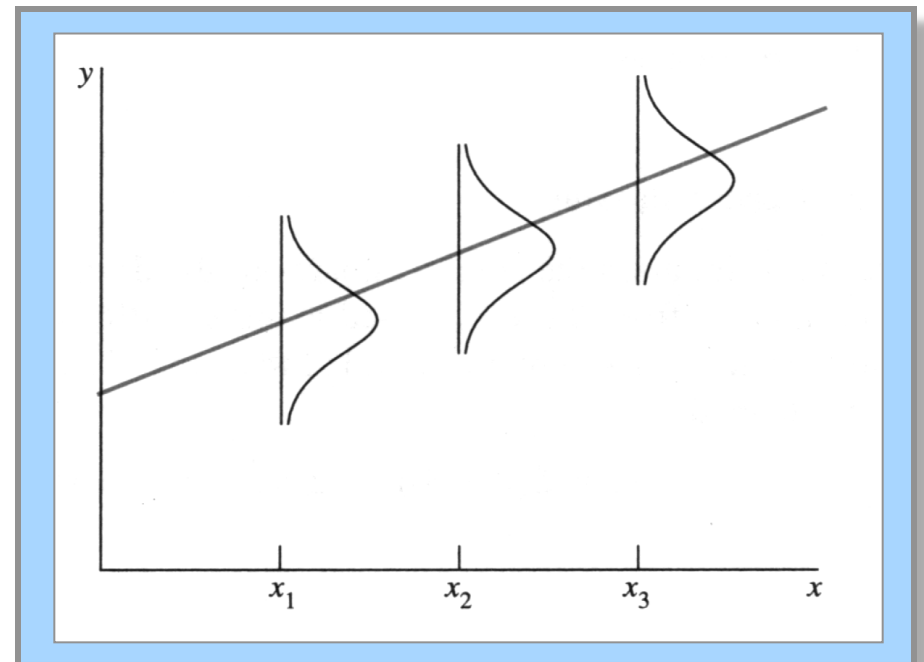


**Points deviate from the line of means by an amount  $\varepsilon$  where  $\varepsilon$  has a normal distribution with mean 0 and variance  $\sigma^2$ .**

# The Random Error



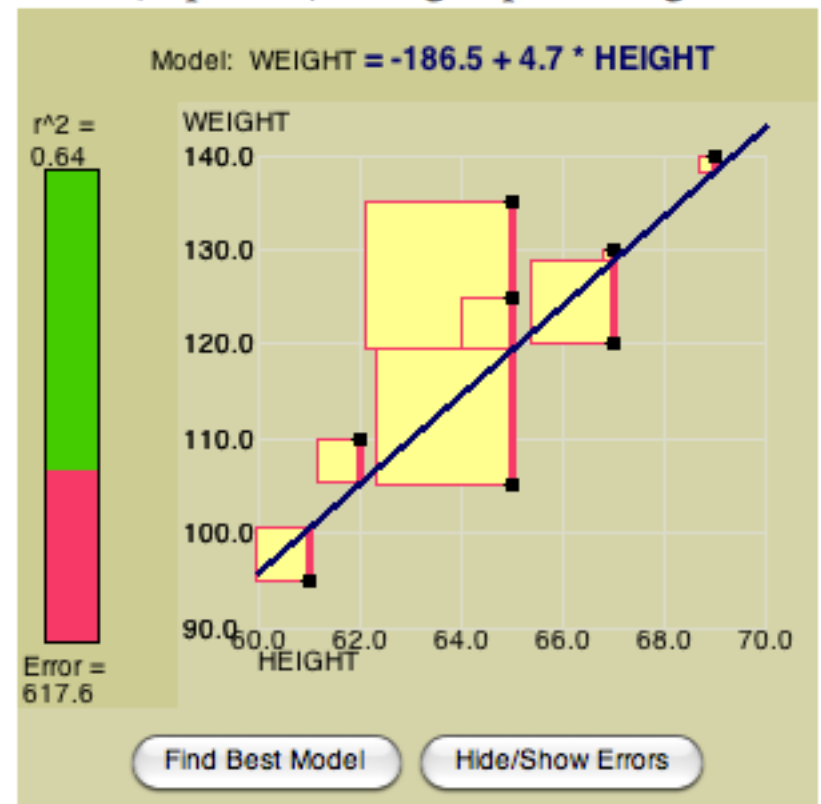
- The line of means,  $E(y) = \alpha + \beta x$ , describes average value of  $y$  for any fixed value of  $x$ .
- The population of measurements is generated as  $y$  deviates from the population line by  $\varepsilon$ . We estimate  $\alpha$  and  $\beta$  using sample information.



# Linear Regression App

- <http://www.duxbury.com/authors/mcclellandg/tiein/johnson/reg.htm>
- **Play with App to see the relationship between  $R^2$  and the error**

The data in the applet correspond to Table 3-12 and Figure 3-22 on pp. 164-165 in *Just the Essentials of Elementary Statistics*. The substantive question is the relationship between HEIGHT (in inches) and WEIGHT (in pounds) for a group of college women.



# Simple Linear Regression in R

<http://www-stat.stanford.edu/~jtaylo/courses/stats203/R/introduction/introduction.R.html>

```
heights.table <- read.table('http://www-stat.stanford.edu/~jtaylo/courses/stats203/data/heights.table', header=T, sep=',')
```

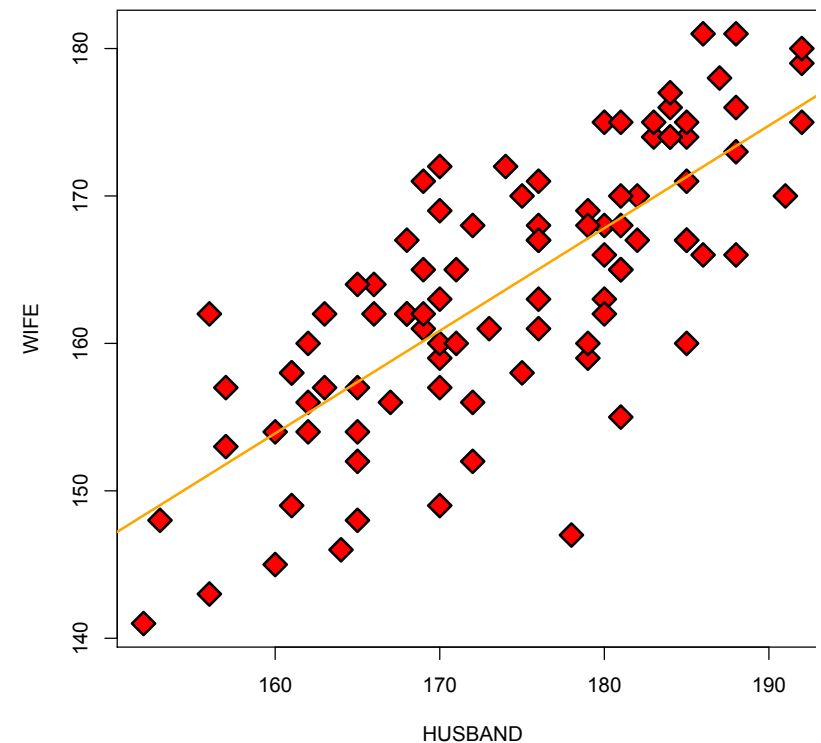
```
attach(heights.table)
```

```
# wife's height vs. husband's height  
plot(heights.table, pch=23, bg='red', cex=2, lwd=2)
```

```
# Fit model
```

```
wife.lm <- lm(WIFE ~ HUSBAND)  
print(summary(wife.lm))
```

```
# with fitted line  
plot(heights.table, pch=23, bg='red', cex=2, lwd=2)  
abline(wife.lm$coef, lwd=2, col='orange')
```



# R Example: Simple Linear Regression

- **### Download the data and tell R where to find the variables by attaching it**

```
heights.table <- read.table('http://www-stat.stanford.edu/~jtaylor/courses/stats203/data/heights.table',  
header=T, sep=',')  
attach(heights.table)
```

```
# wife's height vs. husband's height  
plot(heights.table, pch=23, bg='red', cex=2, lwd=2)
```

```
# Fit model  
wife.lm <- lm(WIFE ~ HUSBAND)  
print(summary(wife.lm))
```

```
# with fitted line  
plot(heights.table, pch=23, bg='red', cex=2, lwd=2)  
abline(wife.lm$coef, lwd=2, col='orange')
```

**### Some other aspects of R**

```
# Take a look at the variable names  
names(heights.table)  
# Estimate beta.1 using S_xx and S_yx
```

```
num <- cov(HUSBAND, WIFE) # = S_xx / (n-1)  
den <- var(HUSBAND) # = S_yx / (n-1)  
print(num/den)  
# Get predicted values (Y.hat)
```

```
wife.hat <- predict(wife.lm)  
# Two different ways of getting residuals  
wife.resid1 <- WIFE - predict(wife.lm)  
wife.resid2 <- resid(wife.lm)
```

```
# Computing sample variance by hand
```

[http://www-stat.stanford.edu/  
~jtaylor/courses/stats203/R/  
introduction/introduction.R.html](http://www-stat.stanford.edu/~jtaylor/courses/stats203/R/introduction/introduction.R.html)

# Residuals

---

- **# Get predicted values (Y.hat)**

```
wife.hat <- predict(wife.lm)
```

```
# Two different ways of getting residuals
```

```
wife.resid1 <- WIFE - predict(wife.lm)
```

```
wife.resid2 <- resid(wife.lm)
```

```
# Computing sample variance by hand
```

```
husband.var <- sum((HUSBAND - mean(HUSBAND))^2) / (length(HUSBAND) - 1)  
print(c(var(HUSBAND), husband.var))
```

```
# Estimating sigma.sq
```

```
S2 <- sum(resid(wife.lm)^2) / wife.lm$df
```

```
print(sqrt(S2))
```

```
print(sqrt(sum(resid(wife.lm)^2) / (length(WIFE) - 2)))
```

```
print(summary(wife.lm)$sigma)
```

```
# What else is in summary(wife.lm)?
```

```
print(names(summary(wife.lm)))
```



# Linear Regression in R : WWW

---

- [R Homepage](#)
- [R Download Page](#)
- [Using R in Statistics](#)
  
- **Dataframes, distributions etc. in R**
  
- <http://msenux.redwoods.edu/math/R/>
  
- [Linear Regression in R](#)

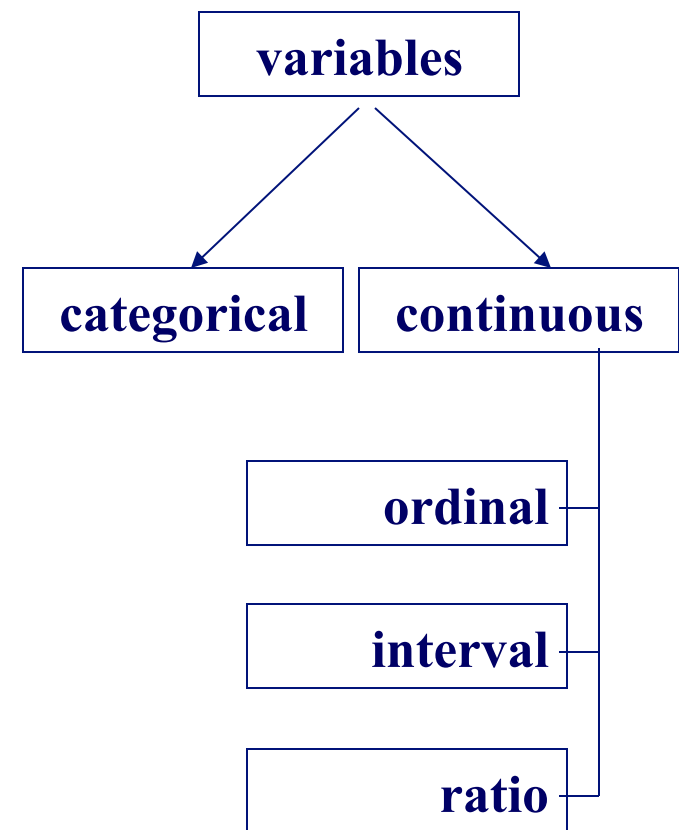
# Lecture Outline

---

- **Linear Regression: a brief intro**
- **A quick statistics review**
  - Mean, expected value, variance, stdev, quantiles, stats in R
- **Locally Weighted Linear Regression**
- **Exploratory Data Analysis**
- **Simple Linear Regression**
  - Normal Equations
  - Closed form Solution
  - Variance of the estimators
- **Good model?**

# Scales of Measurement

- All measurement in science was conducted using four different types of scales that he called "nominal", "ordinal", "interval" and "ratio"
- In general, many unobservable psychological qualities (e.g., extraversion), are measured on interval scales
- We will mostly concern ourselves with the simple categorical (nominal) versus continuous distinction (ordinal, interval, ratio)
- Check out
  - [http://en.wikipedia.org/wiki/Level\\_of\\_measurement](http://en.wikipedia.org/wiki/Level_of_measurement)



# Summarizing Data

---

- Data are a **bunch of values** of one or more **variables**.
- A **variable** is something that has different values.
  - Values can be **numbers** or **names**, depending on the variable:
    - **Numeric**, e.g. weight
    - **Counting**, e.g. number of injuries
    - **Ordinal**, e.g. competitive level (values are numbers/names)
    - **Nominal**, e.g. sex (values are names)
  - When values are **numbers**, visualize the **distribution** of all values in **stem and leaf plots** or in a frequency histogram.
    - Can also use **normal probability plots** to visualize how well the values fit a normal distribution.
  - When values are **names**, visualize the frequency of each value with a **pie chart** or a just a list of values and frequencies.

- 
- **A statistic is a number summarizing a bunch of values.**
    - Simple or univariate statistics summarize values of one variable.
    - Effect or outcome statistics summarize the relationship between values of two or more variables.
  - **Simple statistics for numeric variables...**
    - Mean: the average
    - Standard deviation: the typical variation
    - Standard error of the mean: the typical variation in the mean with repeated sampling
      - **Multiply by  $\sqrt{\text{sample size}}$**  to convert to standard deviation.
    - Use these also for counting and ordinal variables.
    - Use median (middle value or 50th percentile) and quartiles (25th and 75th percentiles) for grossly non-normally distributed data.
    - Summarize these and other simple statistics visually with box and whisker plots.

- **Simple statistics for nominal variables**

- Frequencies, proportions, or odds.
- Can also use these for ordinal variables.

- **Effect statistics...**

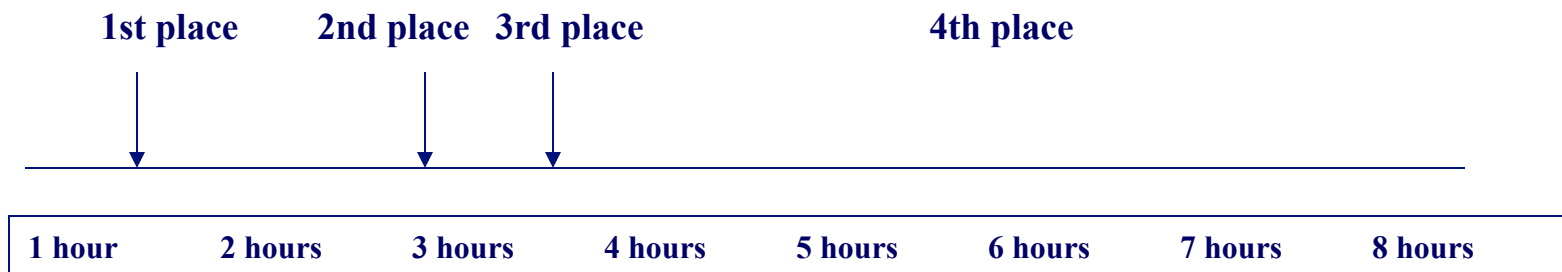
- Derived from statistical model (equation) of the form Y (dependent) vs X (predictor or independent).
- Depend on type of Y and X . Main ones:

Y	X	Model/Test	Effect statistics
numeric	numeric	regression	slope, intercept, correlation
numeric	nominal	t test, ANOVA	mean difference
nominal	nominal	chi-square	frequency difference or ratio
nominal	numeric	categorical	frequency ratio per...

# Ordinal Measurement

---

- **Ordinal: Designates an ordering; quasi-ranking**
  - Does not assume that the intervals between numbers are equal.
  - finishing place in a race (first place, second place)



# Interval and Ratio Measurement

---

- **Interval: designates an equal-interval ordering**
  - The distance between, for example, a 1 and a 2 is the same as the distance between a 4 and a 5
  - Example: Common IQ tests are assumed to use an interval metric
- **Ratio: designates an equal-interval ordering with a true zero point (i.e., the zero implies an absence of the thing being measured)**
  - Example: number of intimate relationships a person has had
    - 0 quite literally means *none*
    - a person who has had 4 relationships has had twice as many as someone who has had 2



# Statistics: Enquiry to the unknown

<b>Population</b>	<b>Sample</b>
Parameter	Estimate

**Parameter** A parameter is a value, usually unknown (and which therefore has to be estimated), used to represent a certain population characteristic. For example, the population mean is a parameter that is often used to indicate the average value of a quantity.

Within a population, a parameter is a fixed value which does not vary. Each sample drawn from the population has its own value of any statistic that is used to estimate this parameter. For example, the mean of the data in a sample is used to give information about the overall mean in the population from which that sample was drawn.

**Statistic:** A statistic is a quantity that is calculated from a sample of data. It is used to give information about unknown values in the corresponding population. For example, the average of the data in a sample is used to give information about the overall average in the population from which that sample was drawn.

It is possible to draw more than one sample from the same population and the value of a statistic will in general vary from sample to sample. For example, the average value in a sample is a statistic. The average values in more than one sample, drawn from the same population, will not necessarily be equal.

# Estimate the population mean

---

**Population height mean = 160 cm**  
**Standard deviation = 5.0 cm**

```
ht <- rnorm(10, mean=160, sd=5)
mean(ht)
```

```
ht <- rnorm(10, mean=160, sd=5)
mean(ht)
```

```
ht <- rnorm(100, mean=160, sd=5)
mean(ht)
```

```
ht <- rnorm(1000, mean=160, sd=5)
mean(ht)
```

```
ht <- rnorm(10000, mean=160, sd=5)
mean(ht)
hist(ht)
```

**The larger the sample, the more accurate the estimate is!**

# Estimate the population proportion

---

Population proportion of males = 0.50  
Take n samples, record the number of k males  
`rbinom(n, k, prob)`

```
males <- rbinom(10, 10, 0.5)
males
mean(males)
```

```
males <- rbinom(20, 100, 0.5)
males
mean(males)
```

```
males <- rbinom(1000, 100, 0.5)
males
mean(males)
```

**The larger the sample, the more accurate the estimate is!**

# Summary of Continuous Data

---

- **Measures of central tendency:**
  - Mean, median, mode
- **Measures of dispersion or variability:**
  - Variance, standard deviation, standard error
  - Interquartile range

## R commands

```
length(x) , mean(x) , median(x) , var(x) , sd(x)  
summary(x) , quantile(x)
```

```
full.deciles<-quantile(x,probs=seq(0,1,by=.1))
```

```
# now we're interested in each 10% cutoff, not just the quarters
```

# R example

---

```
height <- rnorm(1000, mean=55, sd=8.2)
```

```
mean(height)
```

```
[1] 55.30948
```

```
median(height)
```

```
[1] 55.018
```

```
var(height)
```

```
[1] 68.02786
```

```
sd(height)
```

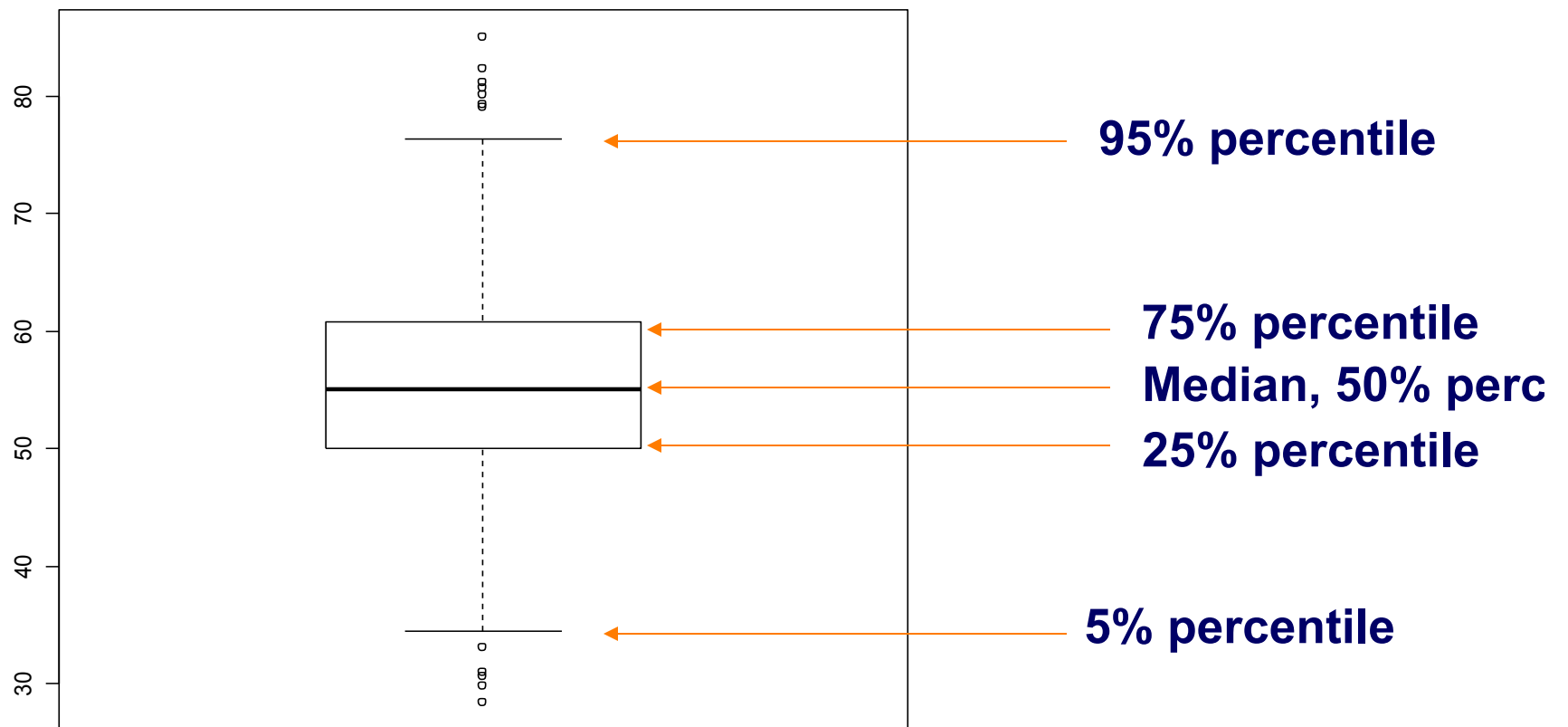
```
[1] 8.2479
```

```
summary(height)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
28.34	49.97	55.02	55.31	60.78	85.05

# Graphical Summary: Box plot

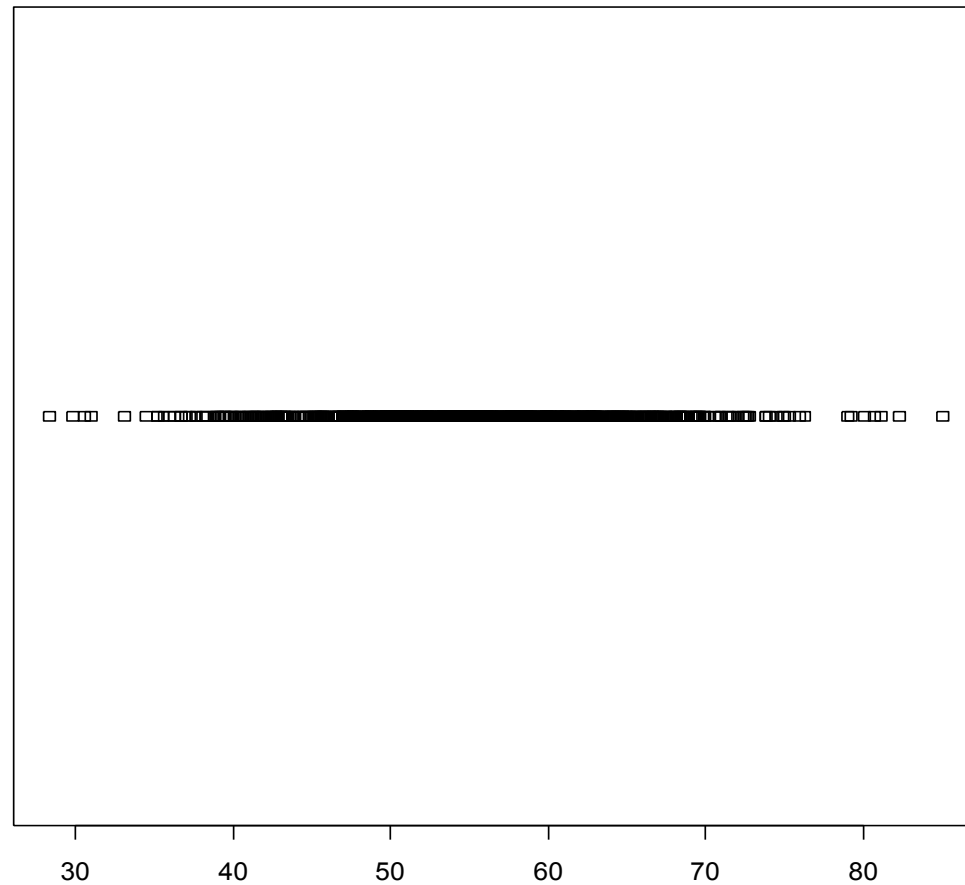
boxplot (height)



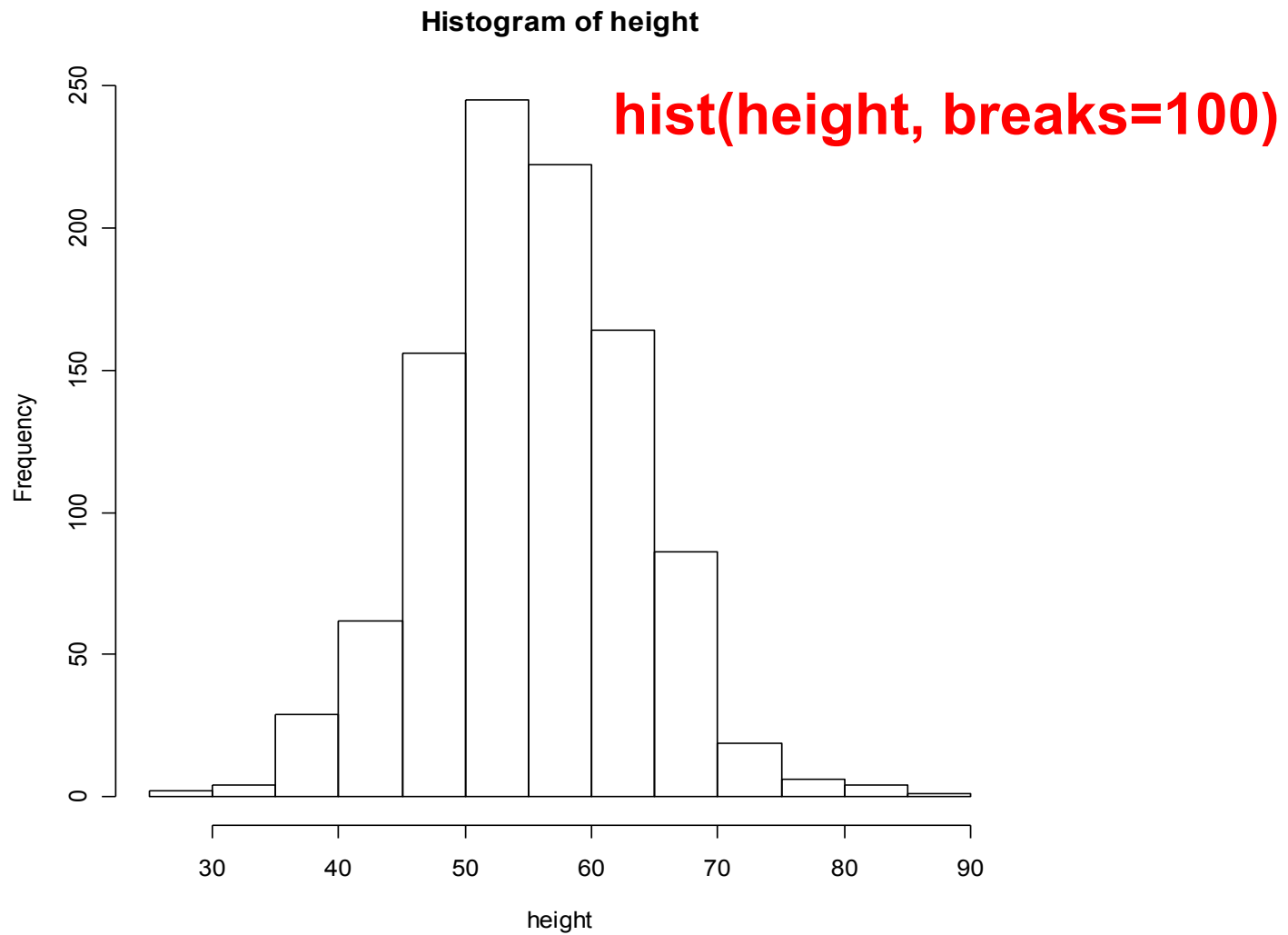
# Strip chart

---

**stripchart(height)**



# Histogram





# Expected Value (weighted average)

---

- **Definition (informal)**
  - The expected value of a random variable  $X$  is the weighted average of the values that  $X$  can take on, where each possible value is weighted by its respective probability.
  - The expected value of a random variable  $X$  is denoted by  $E(X)$  and it is often called the expectation of or the mean of  $X$ .
- **In probability theory, the expected value (or expectation, or mathematical expectation, or mean, or the first moment) of a random variable is the weighted average of all possible values that this random variable can take on.**
  - The weights used in computing this average correspond to the probabilities in case of a discrete random variable, or densities in case of a continuous random variable.
  - From a rigorous theoretical standpoint, the expected value is the integral of the random variable with respect to its probability measure.

# Expected Value for Discrete Variable

When  $X$  is a discrete random variable having support  $R_X$  and probability mass function  $p_X(x)$ , the formula for computing its expected value is a straightforward implementation of the informal definition given above: the expected value of  $X$  is the weighted average of the values that  $X$  can take on (the elements of  $R_X$ ), where each possible value  $x \in R_X$  is weighted by its respective probability  $p_X(x)$ .

**Definition** Let  $X$  be a discrete random variable with support  $R_X$  and probability mass function  $p_X(x)$ . The expected value of  $X$  is:

$$E[X] = \sum_{x \in R_X} xp_X(x)$$

provided that:

$$\sum_{x \in R_X} |x|p_X(x) < \infty$$

$$E_{P(x)}[X]$$

The symbol

$$\sum_{x \in R_X}$$

indicates summation over all the elements of the support  $R_X$ . So, for example, if

$$R_X = \{1, 2, 3\}$$

then:

$$\sum_{x \in R_X} xp_X(x) = 1 \cdot p_X(1) + 2 \cdot p_X(2) + 3 \cdot p_X(3)$$

# Expected Value wrt

Suppose **random variable**  $X$  can take value  $x_1$  with probability  $p_1$ , value  $x_2$  with probability  $p_2$ , and so on, up to value  $x_k$  with probability  $p_k$ . Then the **expectation** of this random variable  $X$  is defined as

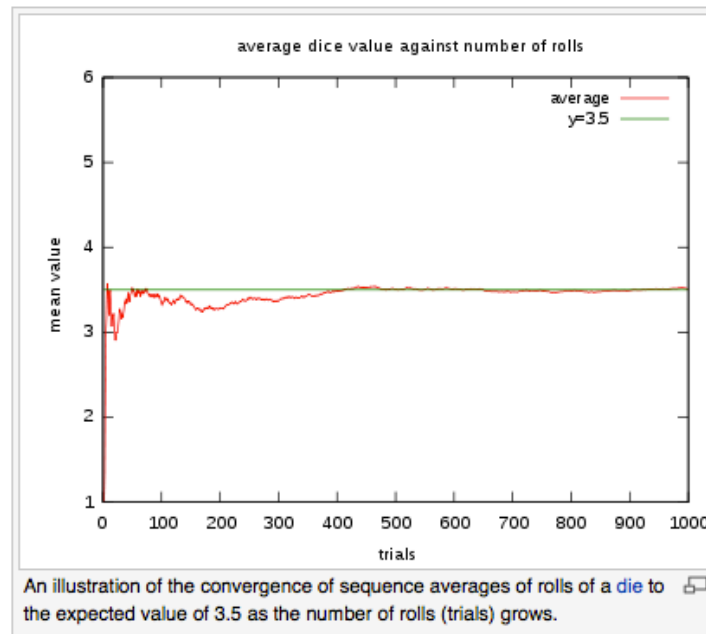
$$E[X] = x_1p_1 + x_2p_2 + \dots + x_kp_k .$$

Since all probabilities  $p_i$  add up to one:  $p_1 + p_2 + \dots + p_k = 1$ , the expected value can be viewed as the **weighted average**, with  $p_i$ 's being the weights:

$$E[X] = \frac{x_1p_1 + x_2p_2 + \dots + x_kp_k}{p_1 + p_2 + \dots + p_k} .$$

If all outcomes  $x_i$  are equally likely (that is,  $p_1 = p_2 = \dots = p_k$ ), then the weighted average turns into the simple **average**. This is intuitive: the expected value of a random variable is the average of all values it can take; thus the expected value is what you expect to happen *on average*. If the outcomes  $x_i$  are not equiprobable, then the simple average ought to be replaced with the weighted average, which takes into account the fact that some outcomes are more likely than the others. The intuition however remains the same: the expected value of  $X$  is what you expect to happen *on average*.

**Example 1.** Let  $X$  represent the outcome of a roll of a six-sided **die**. More specifically,  $X$  will be the number of pips showing on the top face of the **die** after the toss. The possible values for  $X$  are 1, 2, 3, 4, 5, 6, all equally likely (each having the probability of  $\frac{1}{6}$ ). The expectation of  $X$  is



# More Generally..

When  $X$  is an absolutely continuous random variable with probability density function  $f_X(x)$ , the formula for computing its expected value involves an integral, which can be thought of as the limiting case of the summation  $\sum_{x \in \mathcal{R}_X} xp_X(x)$  found in the discrete case above.

**Definition** Let  $X$  be an absolutely continuous random variable with probability density function  $f_X(x)$ . The expected value of  $X$  is:

$$E[X] = \int_{-\infty}^{\infty} xf_X(x)dx$$

In general, if  $X$  is a random variable defined on a probability space  $(\Omega, \Sigma, P)$ , then the expected value of  $X$ , denoted by  $E[X]$ ,  $\langle X \rangle$ ,  $\bar{X}$  or  $\mathbf{E}[X]$ , is defined as Lebesgue integral

$$E[X] = \int_{\Omega} X \, dP = \int_{\Omega} X(\omega) P(d\omega)$$

When this integral exists, it is defined as the expectation of  $X$ . Note that not all random variables have a finite expected value, since the integral may not converge absolutely; furthermore, for some it is not defined at all (e.g., Cauchy distribution). Two variables with the same probability distribution will have the same expected value, if it is defined.

It follows directly from the discrete case definition that if  $X$  is a constant random variable, i.e.  $X = b$  for some fixed real number  $b$ , then the expected value of  $X$  is also  $b$ .

**Berkeley** The expected value of an arbitrary function of  $X$ ,  $g(X)$ , with respect to the probability density function  $f(x)$  is given by the inner product of  $f$  and  $g$ :

# Variance

---

In probability theory and statistics, the variance is a measure of how far a set of numbers are spread out from each other. It is one of several descriptors of a probability distribution, describing how far the numbers lie from the mean (expected value).

If a random variable  $X$  has the expected value (mean)  $\mu = E[X]$ , then the variance of  $X$  is given by:

$$\text{Var}(X) = E[(X - \mu)^2].$$

That is, the variance is the expected value of the squared difference between the variable's realization and the variable's mean. This definition encompasses random variables that are discrete, continuous, or neither (or mixed). It can be expanded as follows:

$$\begin{aligned}\text{Var}(X) &= E[(X - \mu)^2] \\ &= E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 \\ &= E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - \mu^2 \\ &= E[X^2] - (E[X])^2.\end{aligned}$$

A mnemonic for the above expression is "mean of square minus square of mean". The variance of random variable  $X$  is typically designated as  $\text{Var}(X)$ ,  $\sigma_X^2$ , or simply  $\sigma^2$  (pronounced "sigma squared").

# Variance of a Fair Dice

---

A six-sided **fair die** can be modelled with a discrete random variable with outcomes 1 through 6, each with equal probability  $\frac{1}{6}$ . The expected value is  $(1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5$ . Therefore the variance can be computed to be:

$$\begin{aligned}\sum_{i=1}^6 \frac{1}{6}(i - 3.5)^2 &= \frac{1}{6} \sum_{i=1}^6 (i - 3.5)^2 = \frac{1}{6} ((-2.5)^2 + (-1.5)^2 + (-0.5)^2 + 0.5^2 + 1.5^2 + 2.5^2) \\ &= \frac{1}{6} \cdot 17.50 = \frac{35}{12} \approx 2.92.\end{aligned}$$

# Standard Deviation

---

- Standard deviation is a widely used measure of variability or diversity used in statistics and probability theory. It shows how much variation or "dispersion" there is from the average (mean, or expected value). A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data points are spread out over a large range of values.
- The standard deviation of a statistical population, data set, or probability distribution is the square root of its variance. It is algebraically simpler though practically less robust than the average absolute deviation.<sup>[1][2]</sup>
- A useful property of standard deviation is that, unlike variance, it is expressed in the same units as the data.

# Implications of the mean and SD

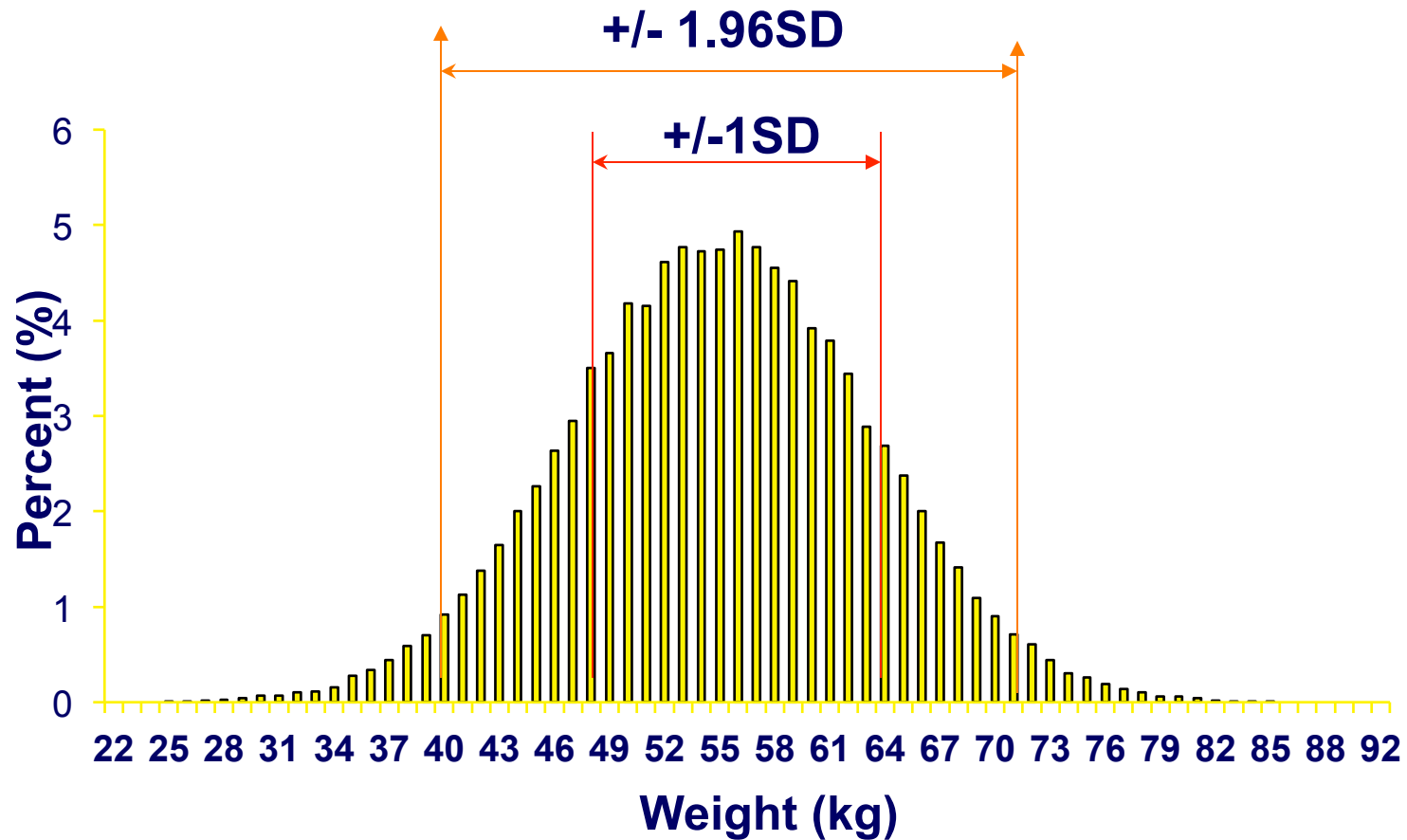
---

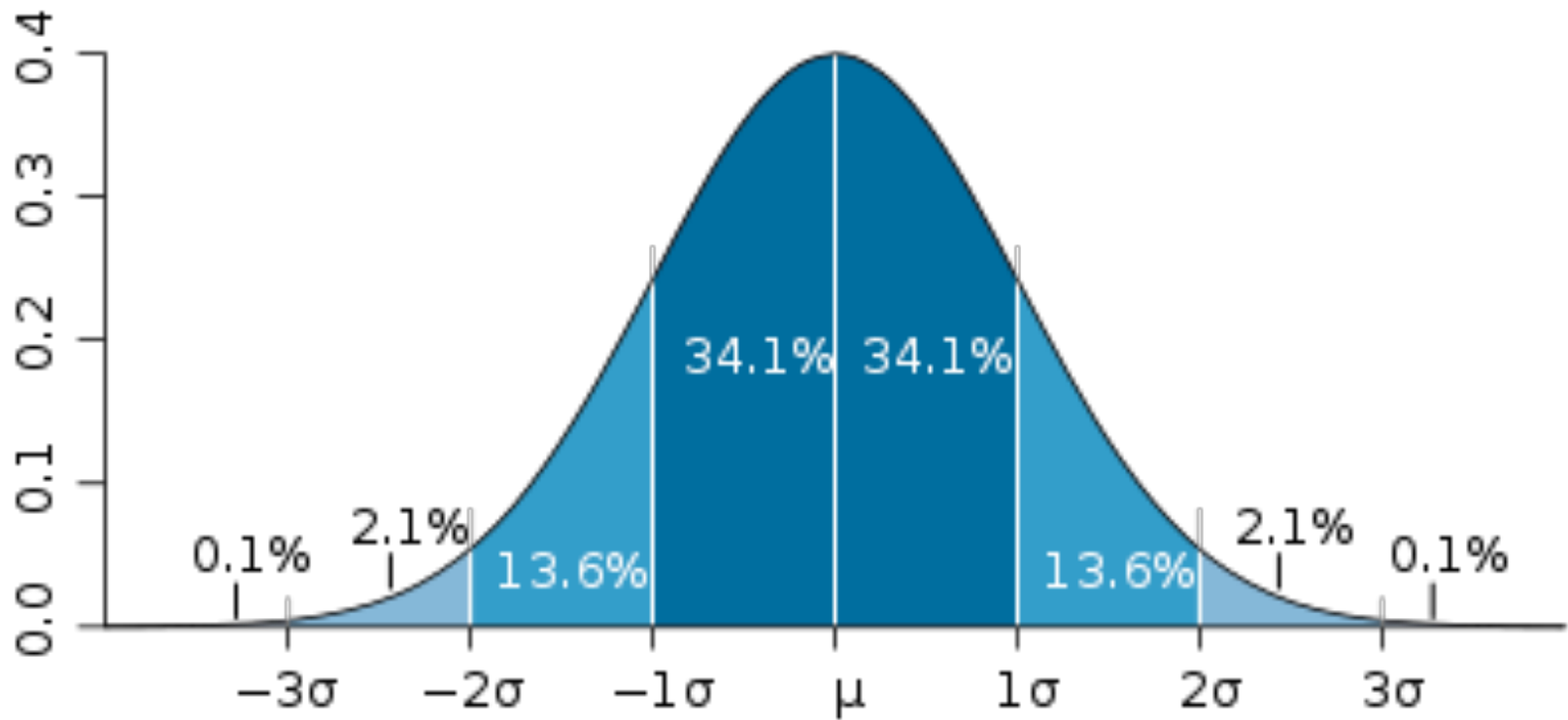
- *“In the Vietnamese population aged 30+ years, the average of weight was 55.0 kg, with the SD being 8.2 kg.”*
- What does this mean?
- 68% individuals will have height between  $55 \pm 8.2 \times 1 = 46.8$  to  $63.2$  kg
- 95% individuals will have height between  $55 \pm 8.2 \times 1.96 = 38.9$  to  $71.1$  kg



# Implications of the mean and SD

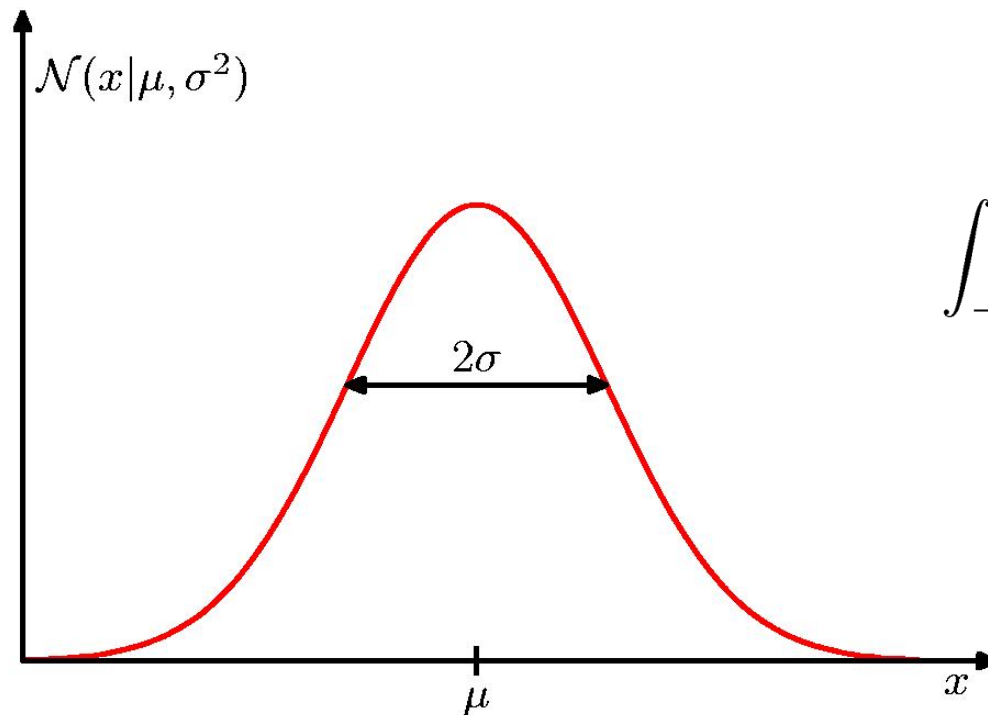
- The distribution of weight of the entire population can be shown to be:





# The Gaussian Distribution

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}$$



$$\mathcal{N}(x|\mu, \sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = 1$$

# Gaussian Mean and Variance

---

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x \, dx = \mu$$

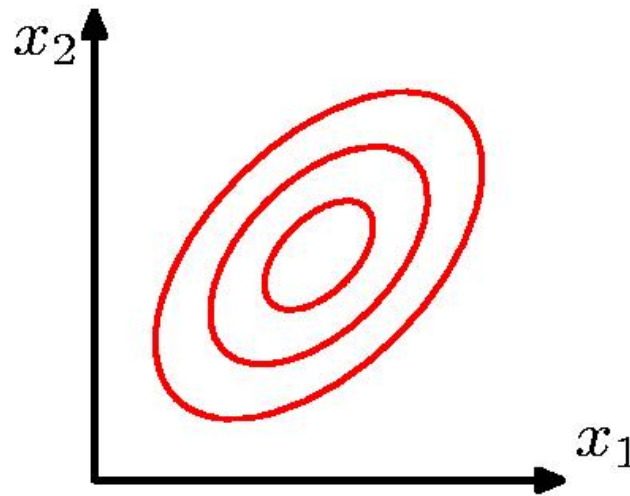
$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x^2 \, dx = \mu^2 + \sigma^2$$

$$\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

# The Multivariate Gaussian

---

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right\}$$



# Distributions in R

- <http://msenux.redwoods.edu/math/R/StandardNormal.php>

## *The Probability Density Function*

The *probability density function* for the normal distribution having mean  $\mu$  and standard deviation  $\sigma$  is given by the function in Figure 1.

### ***The Normal Probability Density Function***

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

**Figure 1.** *The probability density function for the normal distribution.*

If we let the mean  $\mu = 0$  and the standard deviation  $\sigma = 1$  in the probability density function in Figure 1, we get the probability density function for the *standard normal distribution* in Figure 2.

### ***The Standard Normal Probability Density Function***

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

**Figure 2.** *The probability density function for the standard normal distribution has mean  $\mu = 0$  and standard deviation  $\sigma = 1$ .*

---

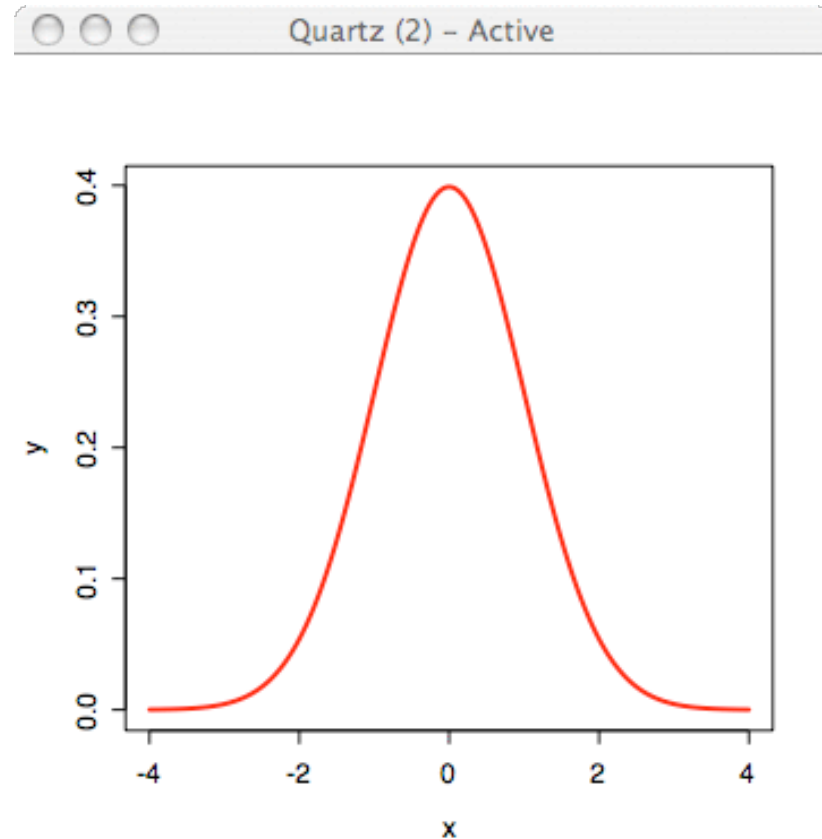
```
x=seq(-4,4,length=200)
y=1/sqrt(2*pi)*exp(-x^2/2)
plot(x,y,type="l",lwd=2,col="red")
```

- If you'd like a more detailed introduction to plotting in R, we refer you to the activity [Simple Plotting in R](#).
- However, these commands are simply explained.
  - The command `x=seq(-4,4,length=200)` produces 200 equally spaced values between -4 and 4 and stores the result in a vector assigned to the variable `x`.
  - The command `y=1/sqrt(2*pi)*exp(-x^2/2)` evaluates the probability density function of Figure 2 at each entry of the vector `x` and stores the result in a vector assigned to the variable `y`.
  - The command `plot(x,y,type="l",lwd=2,col="red")` plots `y` versus `x`, using:
    - a solid line type (**`type="l"`**) --- that's an "el", not an I (eye) or a 1 (one),
    - a line width of 2 points (**`lwd=2`**), and
    - uses the color red (**`col="red"`**).

# Standard Normal Distribution

```
x=seq(-4,4,length=200)  
y=1/sqrt(2*pi)*exp(-x^2/2)  
plot(x,y,type="l",lwd=2,col="red")
```

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$



**The bell-shaped curve of the standard normal distribution.**



# dnorm () as a An Alternate Approach

---

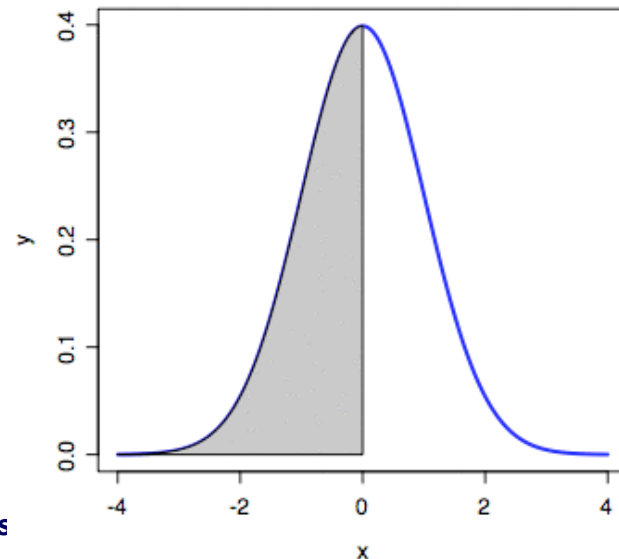
- An Alternate Approach
- The command dnorm can be used to produce the same result as the probability density function of Figure 2.
- Indeed, the "d" in dnorm stands for "density." Thus, the command dnorm is designed to provide values of the probability density function for the normal distribution.

```
x=seq(-4,4,length=200)  
y=dnorm(x,mean=0,sd=1)  
plot(x,y,type="l",lwd=2,col="red")
```

```
x=seq(-4,4,length=200)  
y=1/sqrt(2*pi)*exp(-x^2/2)  
plot(x,y,type="l",lwd=2,col="red")
```

# Area Under the PDF

- Like all probability density functions, the standard normal curves possess two very important properties:
  1. The graph of the probability density function lies entirely above the x-axis. That is,  $f(x) \geq 0$  for all  $x$ .
  2. The area under the curve (and above the x-axis) on its full domain is equal to 1.
- The probability of selecting a number between  $x = a$  and  $x = b$  is equal to the area under the curve from  $x = a$  to  $x = b$ .



# pnorm()

---

- If the total area under the curve equals 1, then by symmetry one would expect that the area under the curve to the left of  $x = 0$  would equal 0.5.
- R has a command called pnorm (the "p" is for "probability") which is designed to capture this probability (area under the curve).

```
pnorm(0, mean=0, sd=1)  
[1] 0.5
```

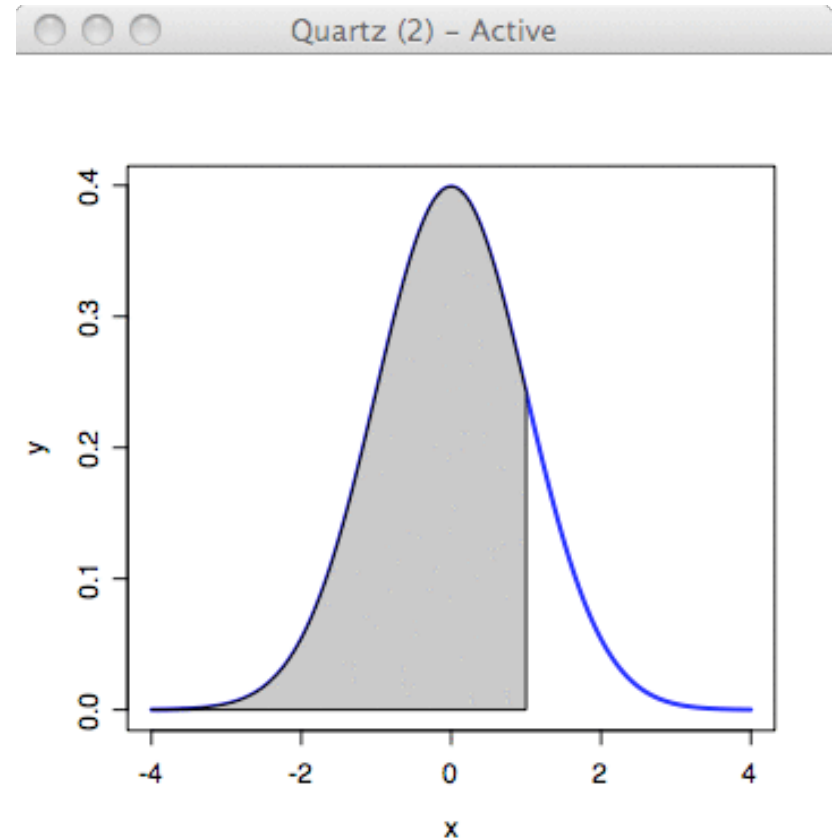
- Note that the syntax is strikingly similar to the syntax for the density function. The command `pnorm(x, mean = , sd = )` will find the area under the normal curve to the left of the number  $x$ . Note that we use `mean=0` and `sd=1`, the mean and density of the **standard normal distribution**.

# polygon()

```
x=seq(-4,4,length=200) > y=dnorm(x)
plot(x,y,type="l", lwd=2, col="blue")
x=seq(-4,1,length=200)
y=dnorm(x)
polygon(c(-4,x,1),c(0,y,0),col="gray")
```

For help on the polygon command enter

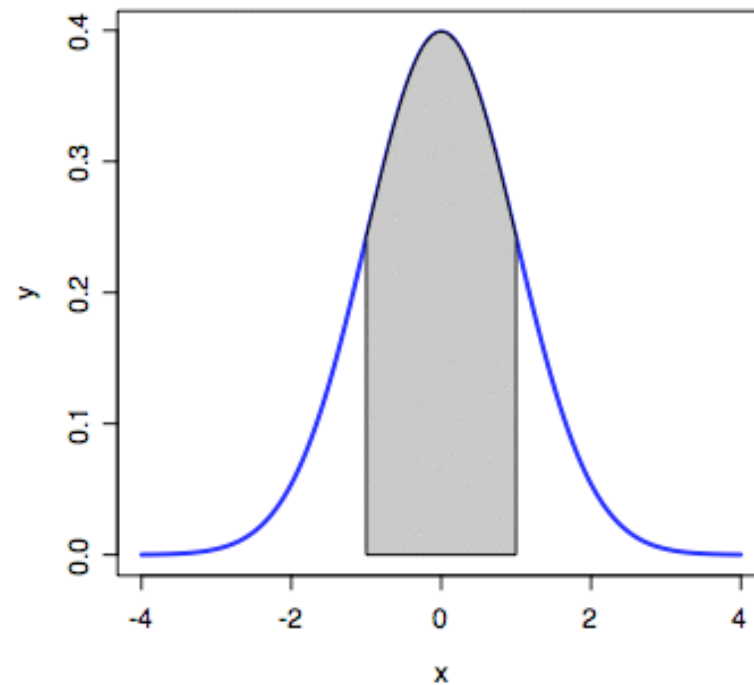
- ?polygon
- and read the resulting help file.
- However, the basic idea is pretty simple.
- In the syntax `polygon(x,y)`, the argument `x` contains the `x`-coordinates of the vertices of the polygon you wish to draw.
- Similarly, the argument `y` contains the `y`-coordinates of the vertices of the desired polygon.



# Exercise

---

- Plot this graph
- What is the area of the shaded area?



# Exercise Solution

---

## 68%-95%-99.7% Rule

The 68% - 95% - 99.7% is a rule of thumb that allows practitioners of statistics to estimate the probability that a randomly selected number from a normal distribution will fall within a certain number of standard deviations from the mean. Let  $\mu$  be the mean and  $\sigma$  be the standard deviation of a normal distribution. Then, the probability that a randomly selected number from the distribution will fall within 1 standard deviation of the mean is 68%, within 2 standard deviations is 95%, and within 3 standard deviations is 99.7%.

$x = \mu + z\sigma$   
 $y = \mu - z\sigma$   
probability  
 $x = \mu + z\sigma$   
probability

probability  
[1]

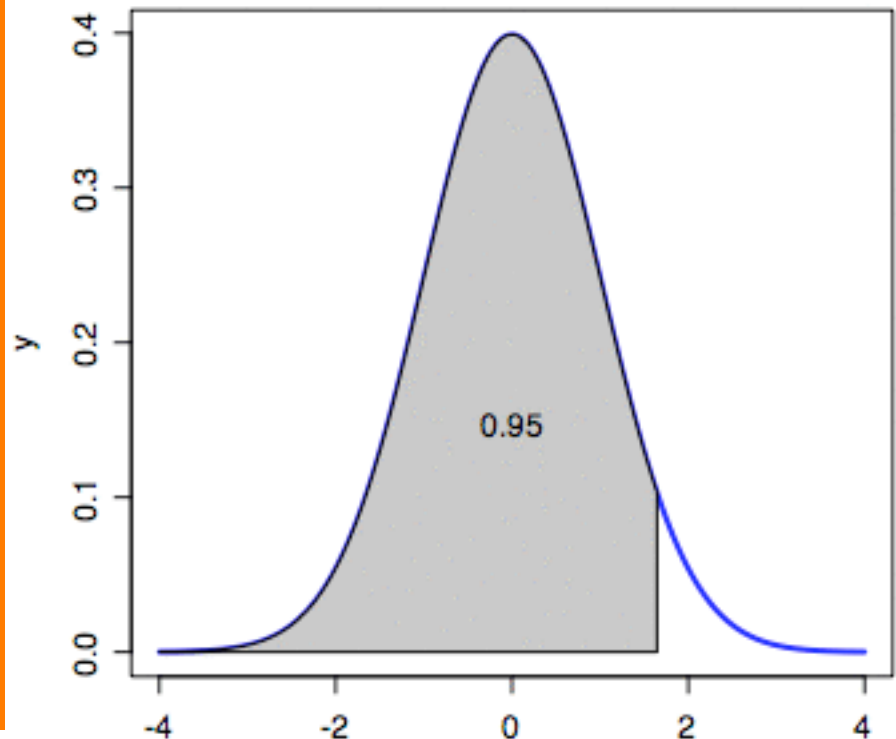
# Quantiles

- Sometimes the opposite question is asked. That is, suppose that the area under the curve to the left of some unknown number is known. What is the unknown number?
- For example, suppose that the area under the curve to the left of some unknown  $x$ -value is 0.85, as shown in Figure

To find the unknown value of  $x$  we use R's `qnorm` command (the "q" is for "quantile").

```
> qnorm(0.95,mean=0,sd=1)  
[1] 1.644854
```

Hence, there is a 95% probability that a random number less than or equal to 1.644854 is chosen from the standard normal distribution.



# pnorm() vs qnorm()

---

- In a sense, R's `pnorm` and `qnorm` commands play the roles of inverse functions.
- On one hand, the command `pnorm` is fed a number and asked to find the probability that a random selection from the standard normal distribution falls to the left of this number.
- On the other hand, the command `qnorm` is given the probability and asked to find a limiting number so that the area under the curve to the left of that number equals the given probability.

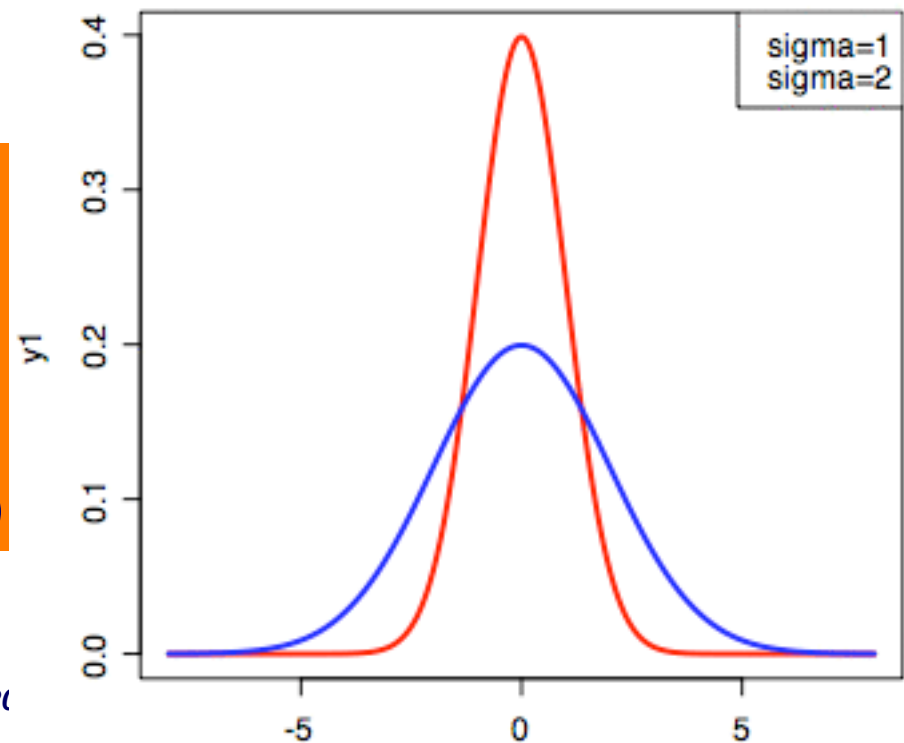


# The Standard Deviation

- The standard deviation represents the "spread" in the distribution. With "spread" as the interpretation, we would expect a normal distribution with a standard deviation of 2 to be "more spread out" than a normal distribution with a standard deviation of 1.
- Let's simulate this idea in R.

```
x=seq(-8,8,length=500)
y1=dnorm(x,mean=0,sd=1)
plot(x,y1,type="l",lwd=2,col="red")

y2=dnorm(x,mean=0,sd=2)
lines(x,y2,type="l",lwd=2,col="blue")
```



# Lecture Outline

---

- **Linear Regression: a brief intro**
- **A quick statistics review**
  - Mean, expected value, variance, stdev, quantiles, stats in R
- **Locally Weighted Linear Regression**
- **Exploratory Data Analysis**
- **Simple Linear Regression**
  - Normal Equations
  - Closed form Solution
  - Variance of the estimators
- **Good model?**

# Kernel Density Estimation

In statistics, **kernel density estimation** is a **non-parametric** way of estimating the **probability density function** of a **random variable**. Kernel density estimation is a fundamental data smoothing problem where inferences about the **population** are made, based on a finite data **sample**. In some fields such as **signal processing** and **econometrics** it is also known as the **Parzen–Rosenblatt window method**. It was developed by **Emanuel Parzen** and **Murray Rosenblatt**, who are usually credited with creating it in its current form.<sup>[1][2]</sup>

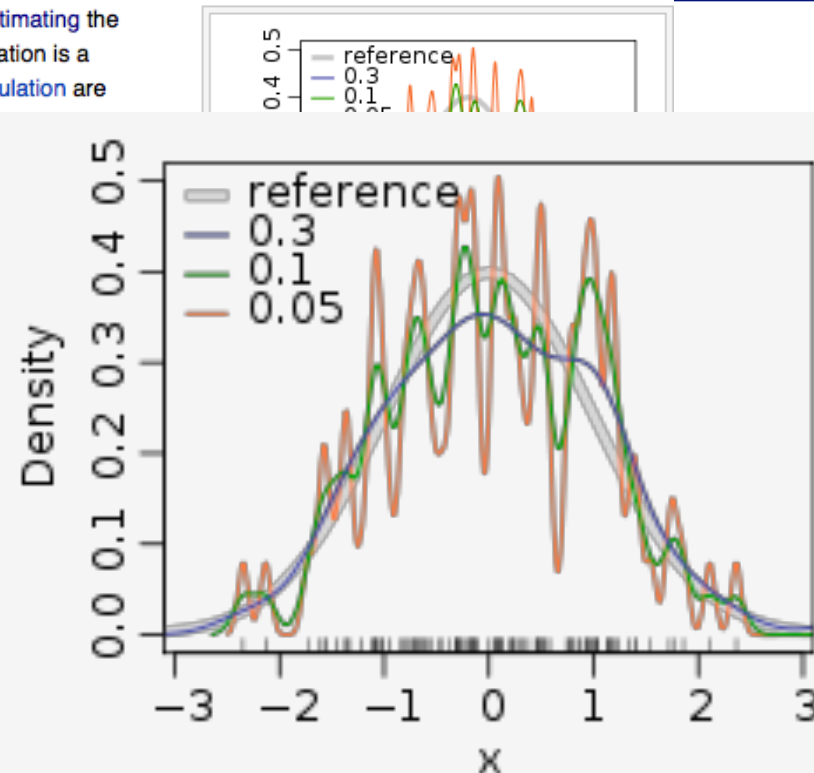
Contents [hide]	
1	Definition
1.1	Relation to the characteristic function density estimator
2	Bandwidth selection
3	Practical estimation of the bandwidth
4	Statistical implementation
4.1	Example in Matlab/octave
4.2	Example in R
5	See also
6	External links
7	References

## Definition

Let  $(x_1, x_2, \dots, x_n)$  be an **iid** sample drawn from some distribution with density function  $f$ . Its **kernel density estimator** is

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - x_i) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right),$$

where  $K(\cdot)$  is the **kernel** — a symmetric but not necessarily positive function that integrates to one — and  $h > 0$  is a **smoothing parameter** called the **bandwidth**. A kernel with subscript  $h$  is called the **scaled kernel** and defined as  $K_h(x) = 1/h K(x/h)$ . Intuitively one wants to choose  $h$  as small as the data allows, however there is always a trade-off between the bias of the estimator and its variance; more on the choice of bandwidth later. A range of **kernel functions** are commonly used: **uniform**, **triangular**, **biweight**, **triweight**, **Epanechnikov**, **normal**, and others. The Epanechnikov kernel is optimal in a minimum variance sense,<sup>[3]</sup> though the loss of efficiency is small for the kernels listed previously,<sup>[4]</sup> and due to its convenient mathematical properties, the normal kernel is often used  $K(x) = \phi(x)$ , where  $\phi$  is the **standard normal density function**.



# Parametric vs. Non-Parametric ML Algorithms

---

- **Parametric ML Algorithms (e.g., OLS, Decision Trees; SVMs)**
  - The linear regression algorithm that we saw earlier is known as a parametric learning algorithm, because it has a fixed, finite number of parameters (the  $W_i$ 's), which are fit to the data.
  - Once we've fit the  $W_i$ 's and stored them away, we no longer need to keep the training data around to make future predictions.
- **Non-Parametric (lowess()); knn; some flavours SVMs)**
  - In contrast, to make predictions using locally weighted linear regression, we need to keep the entire training set around.
  - The term “non-parametric” (roughly) refers to the fact that the amount of stuff we need to keep in order to represent the hypothesis/model grows linearly with the size of the training set.

# Locally Weighted Linear Regression

---

## Non-parametric approach

- **Locally Weighted (Linear) Regression (LWR):**
  - k-NN forms local approximation for each query point  $x_q$
  - Why not form an explicit approximation  $f^{\wedge}(x)$  for region surrounding  $x_q$ 
    - Fit linear function to k nearest neighbors
    - Fit quadratic, ...
    - Thus producing "piecewise approximation" to  $f$ 
      - Minimize error over k nearest neighbors of  $x_q$
      - Minimize error entire set of examples, weighting by distances
      - Combine two above
- **Non-parametric approach**

# Locally Weighted Linear Regression

---

- **Local linear function:**

$$f(x) = w_0 + w_1 a_1(x) + \dots + w_n a_n(x)$$

- **Error criterions:**

$$E_1(x_q) \equiv \frac{1}{2} \sum_{x \in k\_nearest\_nbrs\_of\_x_q} (f(x) - \hat{f}(x))^2$$

$$E_2(x_q) \equiv \frac{1}{2} \sum_{x \in D} (f(x) - \hat{f}(x))^2 K(d(x^i, x))$$

**Binary Neighbors  
With OLS**



**Weighted Neighbors  
With weighted OLS**



- **Combine  $E_1(x_q)$  and  $E_2(x_q)$**

$$E_3(x_q) \equiv \frac{1}{2} \sum_{x \in k\_nearest\_nbrs\_of\_x_q} (f(x) - \hat{f}(x))^2 K(d(x_q, x))$$

$$k(d(x, x^i) = w^i = \exp\left(-\frac{(x^i - x)^2}{2\tau}\right) \text{ where } \tau \text{ is the bandwidth parameter}$$

# Locally Weighted Linear Regression

## How it works

$$E_3(x_q) \equiv \frac{1}{2} \sum_{x \in k\_nearest\_nbrs\_of\_x_q} (f(x) - \hat{f}(x))^2 K(d(x_q, x)) \sum_{x \in k\_nearest\_nbrs\_of\_x_q} w_k (f(x) - \beta^T x_i)^2 \rightarrow \min$$

- For each point  $(x_k, y_k)$  compute  $w_k$
- Let  $WX = \text{Diag}(w_1, w_2, \dots, w_n)X$

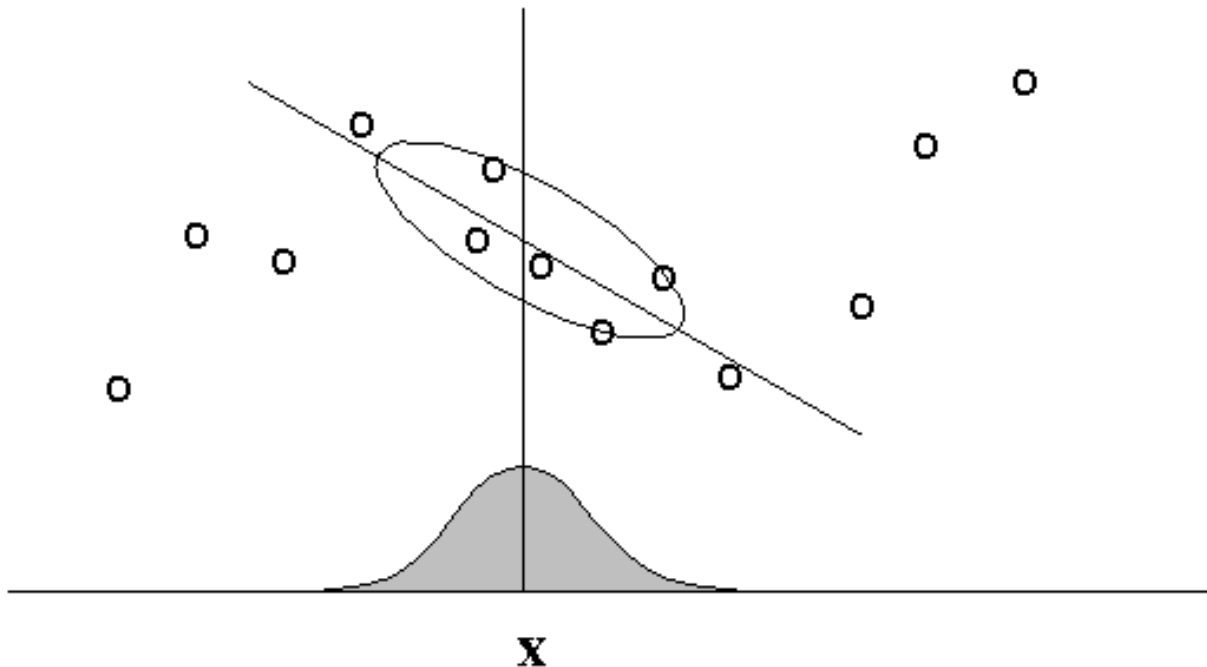
$$\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1D} \\ x_{21} & x_{22} & \dots & x_{2D} \\ \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & \dots & x_{ND} \end{bmatrix} \rightarrow \begin{bmatrix} w_1 & w_1 x_{11} & w_1 x_{12} & \dots & w_1 x_{1D} \\ w_2 & w_2 x_{21} & w_2 x_{22} & \dots & w_2 x_{2D} \\ \vdots & \vdots & \vdots & & \vdots \\ w_N & w_N x_{N1} & w_N x_{N2} & \dots & w_N x_{ND} \end{bmatrix}$$

- Let  $WY = \text{Diag}(w_1, w_2, \dots, w_n)Y$
- $\beta = (WX^T WX^{-1})(WX^T WY)$

# Kernel regression

---

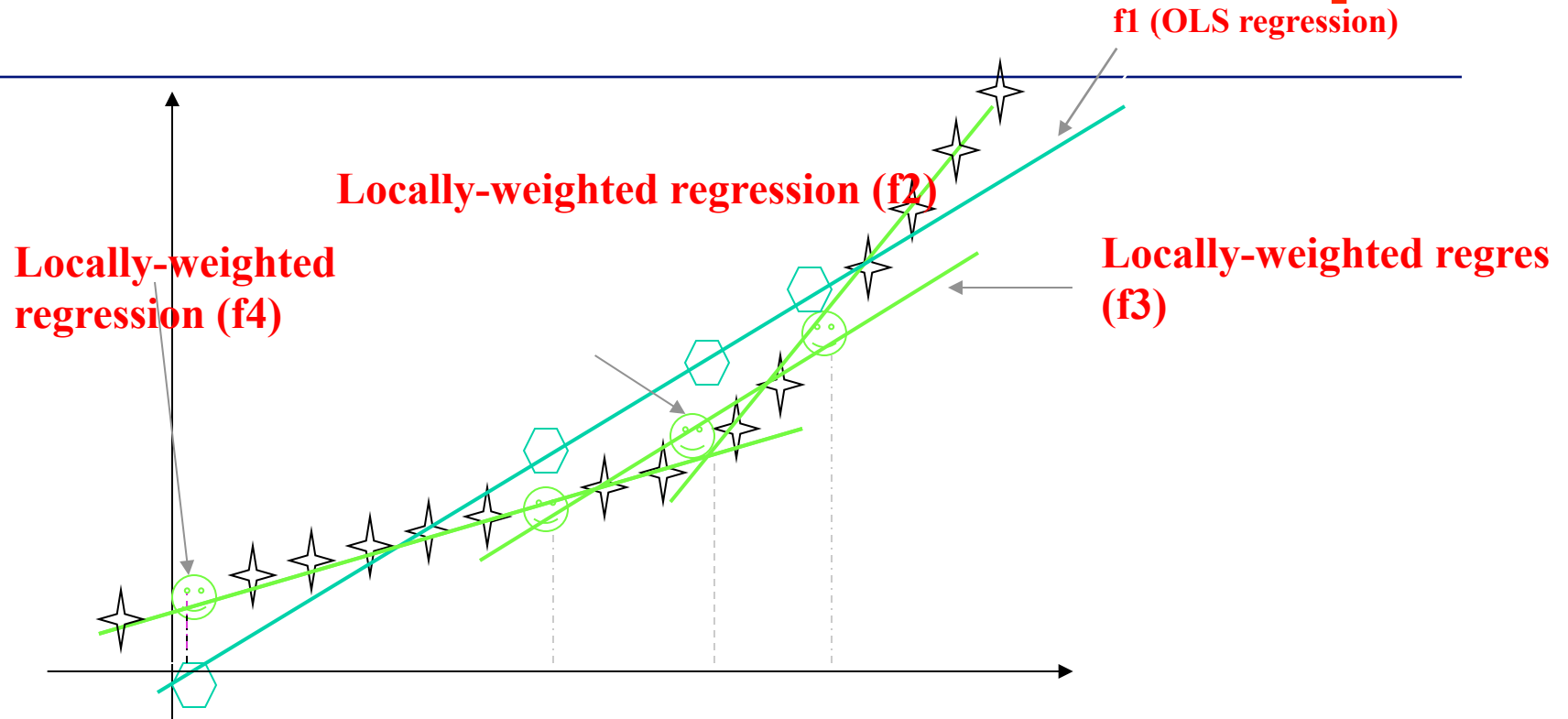
- *aka* locally weighted regression, locally linear regression, LOESS, ...



**Figure 2:** In locally weighted regression, points are weighted by proximity to the current  $x$  in question using a kernel. A regression is then computed using the weighted points.



# LWR Example

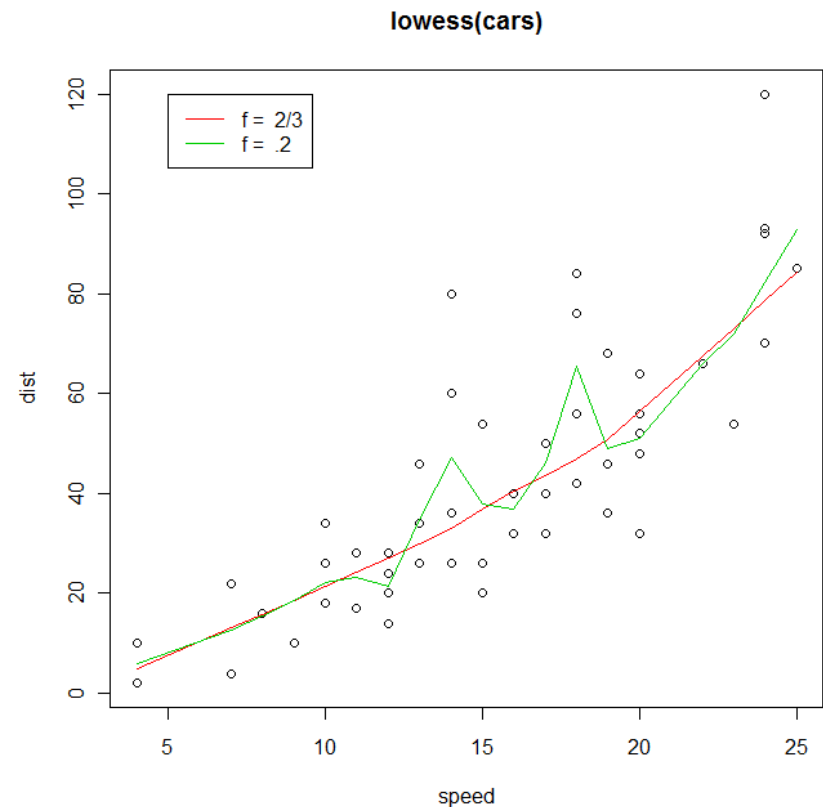


- ★ Training data
- ⬡ Predicted value using simple regression
- 😊 Predicted value using locally weighted (piece-wise) regression

[Yike Guo, Advanced Knowledge Management, 2000]

# LWR in R: `lowess()` or `loess()`

Note  $f$  is the smoother span. This gives the proportion of points in the plot which influence the smooth at each value. Larger values give more smoothness.



```
library(cars)
```

```
# formula method
```

```
plot(dist ~ speed, data=cars, main = "lowess(cars)")
```

```
lines(lowess(dist ~ speed, data=cars), col = 2)
```

```
lines(lowess(dist ~ speed, data=cars, f=.2), col = 3)
```

```
legend(5, 120, c(paste("f = ", c("2/3", ".2"))), lty = 1, col = 2:3)
```

# Lowess and Scatterplot Examples

```
example.lowess = function(){
```

## #EXAMPLE 1

```
library(cars)
# formula method
plot(dist ~ speed, data=cars, main = "lowess(cars)")
lines(lowess(dist ~ speed, data=cars), col = 2)
lines(lowess(dist ~ speed, data=cars, f=.2), col = 3)
legend(5, 120, c(paste("f = ", c("2/3", ".2"))), lty = 1, col = 2:3)
```

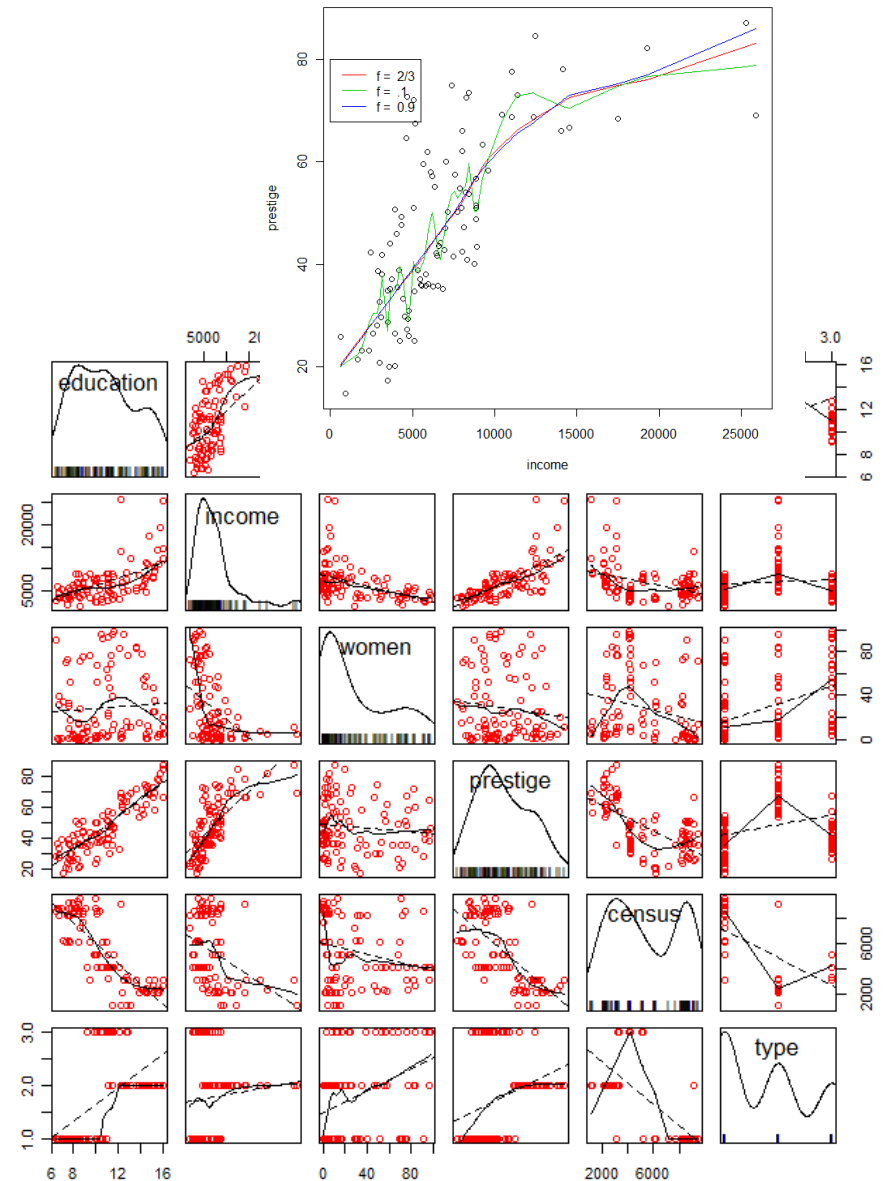
## #EXAMPLE 2

```
library(car)
attach( Prestige )
plot( income , prestige )
#click on examples to see lables; right click and select STOP to
identify( income, prestige, rownames(Prestige), xpd = T)
```

```
lines ( lowess( income, prestige) col=2)      # use the defaults
lines ( lowess( income, prestige, f = 1/10), c=3) # use smaller sp
lines ( lowess( income, prestige, f = 9/10), col=4) # use larger sp
legend(5, 80, c(paste("f = ", c("2/3", ".1", 0.9))), lty = 1, col = 2:4)
```

## #EXAMPLE 3

```
# robust fits for all pairs of variables
#excellent way to examine pairs of variables
?scatterplot.matrix
scatterplot.matrix( Prestige )
scatterplot.matrix( Prestige , span= .1)
detach( Prestige )
```



# LWR Examples

---

- **Loess examples**
  - <http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-nonparametric-regression.pdf>
  - <http://wiki.math.yorku.ca/images/a/a5/Math6630Fox-Chap18.R>

# Lecture Outline

---

- **Linear Regression: a brief intro**
- **A quick statistics review**
  - Mean, expected value, variance, stdev, quantiles, stats in R
- **Locally Weighted Linear Regression**
- **Exploratory Data Analysis**
- **Simple Linear Regression**
  - Normal Equations
  - Closed form Solution
  - Variance of the estimators
- **Good model?**

# Exploratory Data Analysis: rug, density

<http://socserv.socsci.mcmaster.ca/jfox/Books/Companion/scripts/chap-3.R>

```
## All R Companion to Applied Regression, Second Edition ##  
## Script for Chapter 3 ##  
## ##  
## John Fox and Sanford Weisberg ##  
## Sage Publications, 2011 ##  
##-----##
```

```
options(show.signif.stars=FALSE)
```

```
library(car)  
head(Prestige) # first 6 rows  
with(Prestige, hist(income))  
with(Prestige, hist(income, breaks="FD", col="gray"))  
box()
```

```
args(hist.default)
```

```
with(Prestige, {  
  hist(income, breaks="FD", freq=FALSE, ylab="Density")  
  lines(density(income), lwd=2)  
  lines(density(income, adjust=0.5), lwd=1)  
  rug(income)  
  box()  
})
```

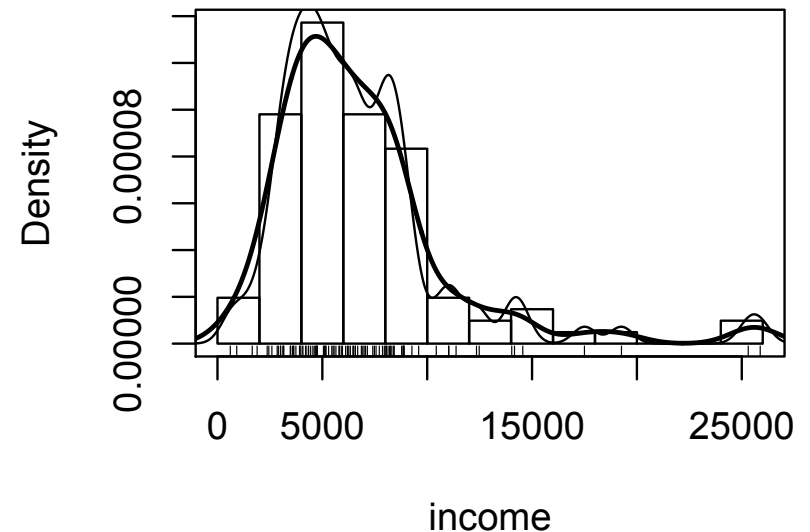
```
with(Prestige, qqPlot(income, labels=row.names(Prestige), id.n=3))
```

Berkeley I 296 Stat (184) for Reproducible Thought Leaders © 2011 James G. Shanahan

James.Shanahan\_AT\_gmail.com

70

Histogram of income



Chapter 3 CAR

# Q-Q plot: comparing two distributions by plotting their quantiles

---

- In statistics, a Q-Q plot<sup>[1]</sup> ("Q" stands for quantile) is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. First, the set of intervals for the quantiles are chosen. A point (x,y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). Thus the line is a parametric curve with the parameter which is the (number of the) interval for the quantile.
- If the two distributions being compared are similar, the points in the Q-Q plot will approximately lie on the line  $y = x$ . If the distributions are linearly related, the points in the Q-Q plot will approximately lie on a line, but not necessarily on the line  $y = x$ . Q-Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.
- A Q-Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. Q-Q plots can be used to compare collections of data, or theoretical distributions.

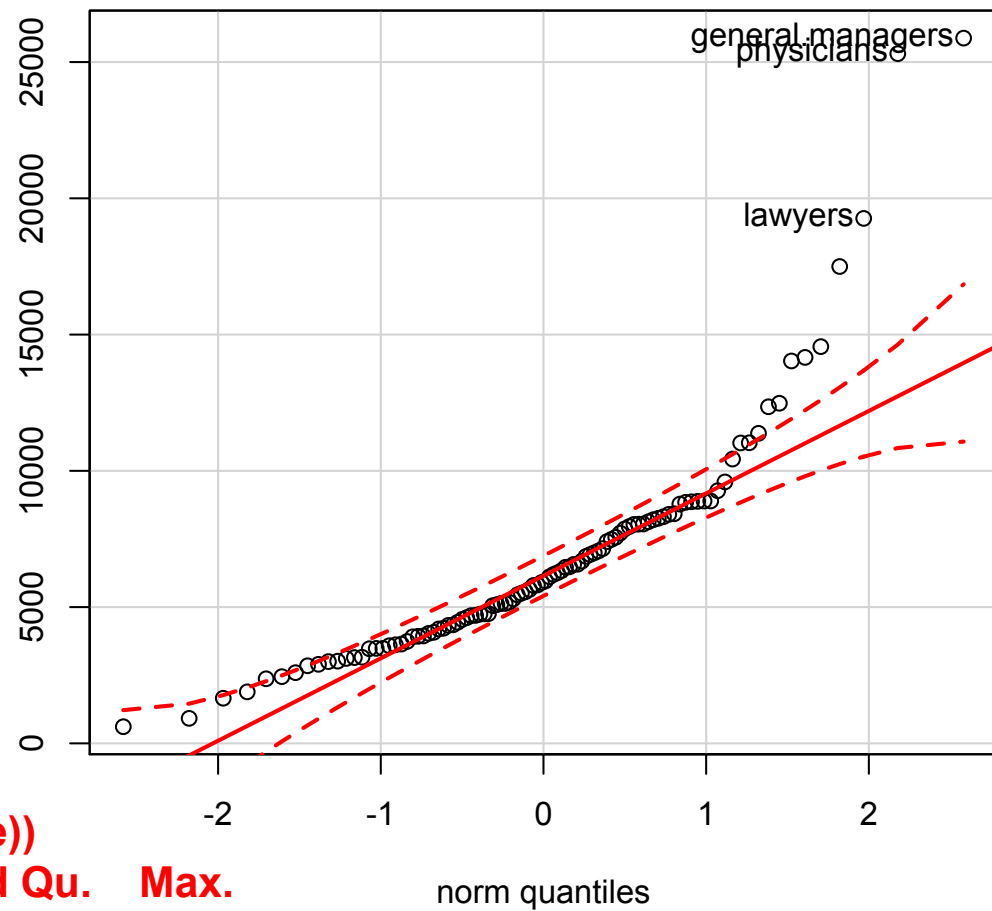
# Compare data to theoretical dist

```
with(Prestige, qqPlot(income, labels=row.names(Prestige), id.n=3))
```

Show 3 most extreme data and corresponding label

SEE:

[http://en.wikipedia.org/wiki/Q-Q\\_plot](http://en.wikipedia.org/wiki/Q-Q_plot)



```
> with(Prestige, summary(income))
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
------	---------	--------	------	---------	------

611	4106	5930	6798	8187	25880
-----	------	------	------	------	-------



# Identify outliers

```
> with(Prestige, qqPlot(income, labels=row.names  
(Prestige), id.n=3))
```

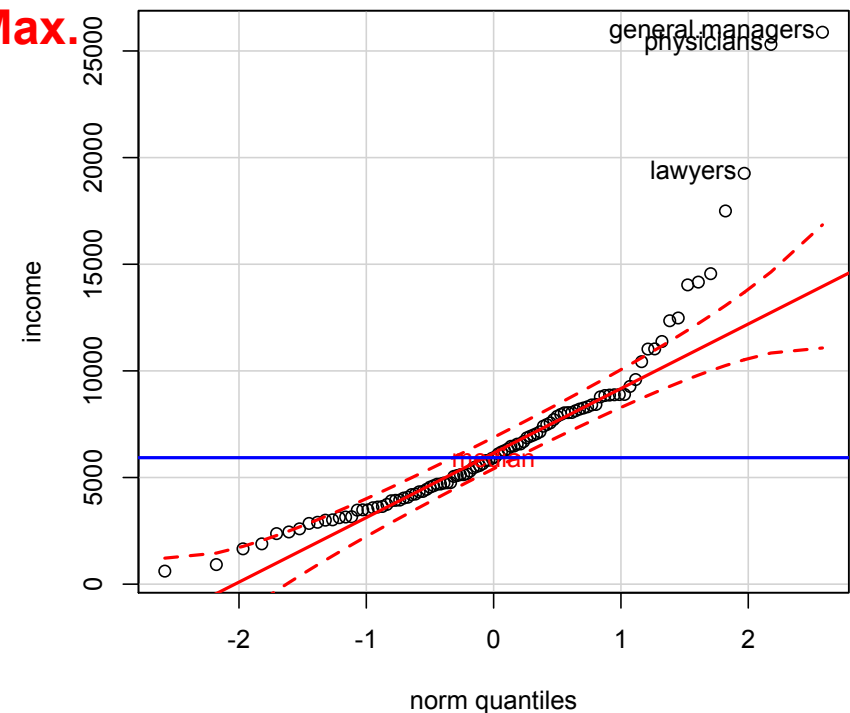
```
[1] "general.managers" "physicians" "lawyers"
```

```
> with(Prestige, summary(income))
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
611	4106	5930	6798	8187	25880

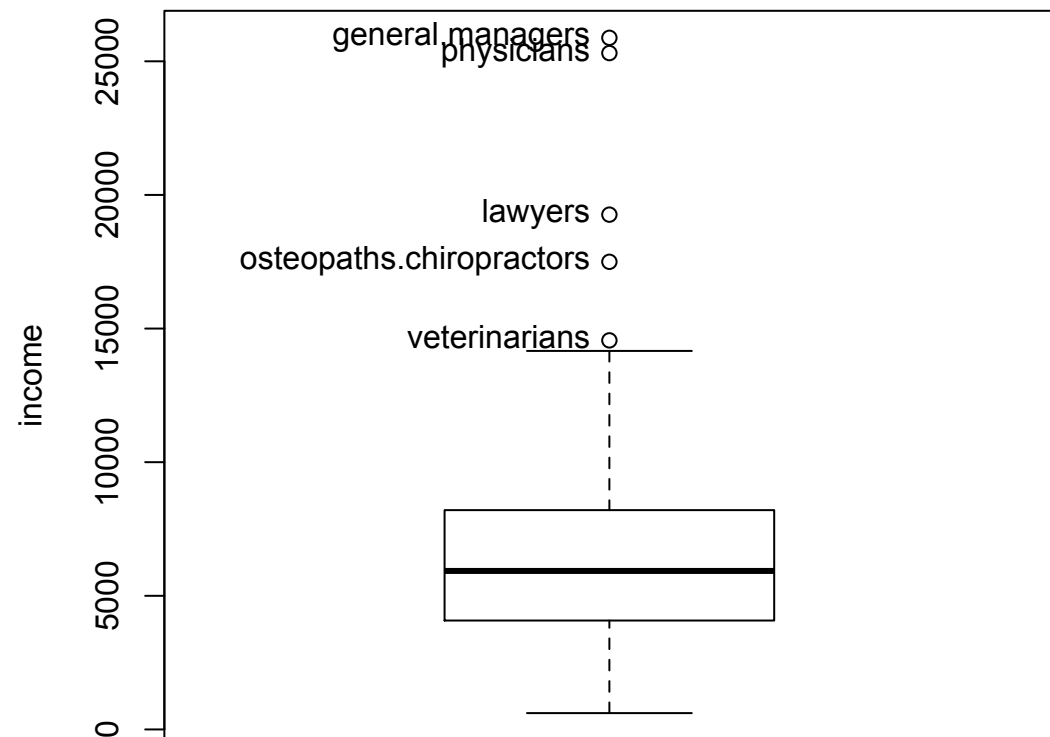
```
> text(0, 5930, "median", col="red")
```

```
> abline(5930,0, lwd=2,col="blue")
```



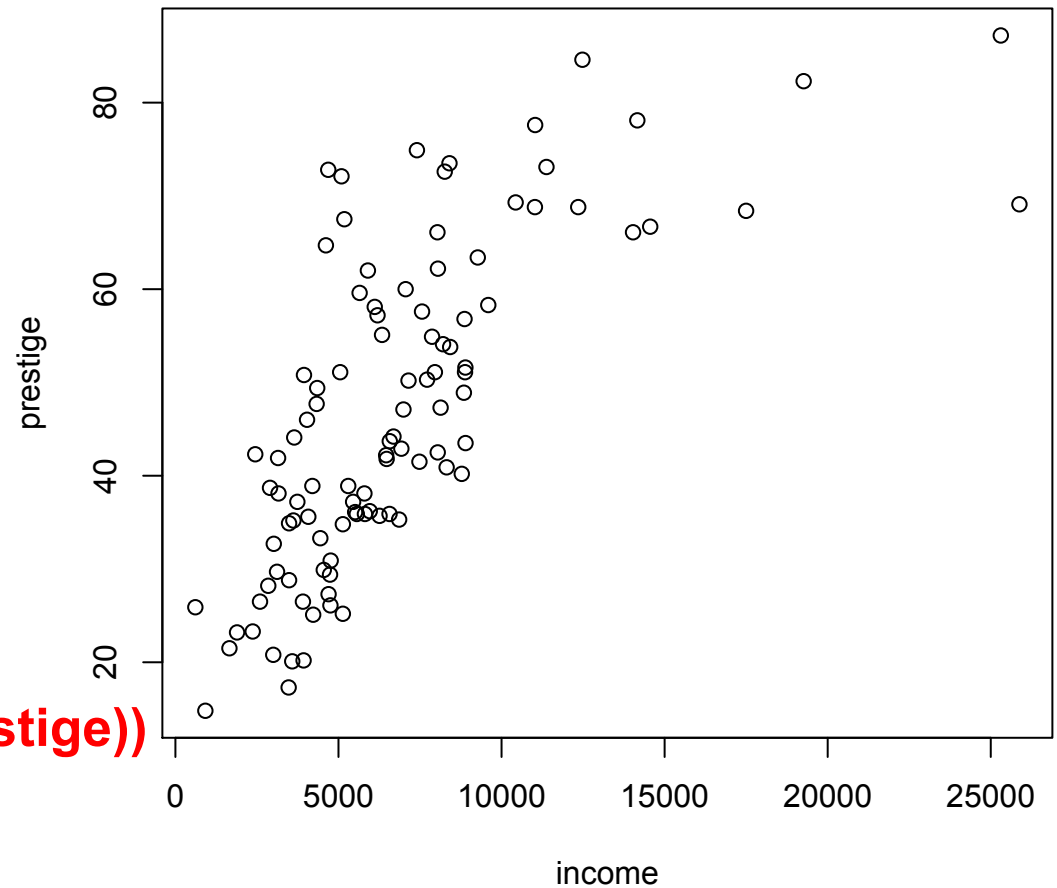
# Boxplot

- **Boxplot(~ income, data=Prestige)**



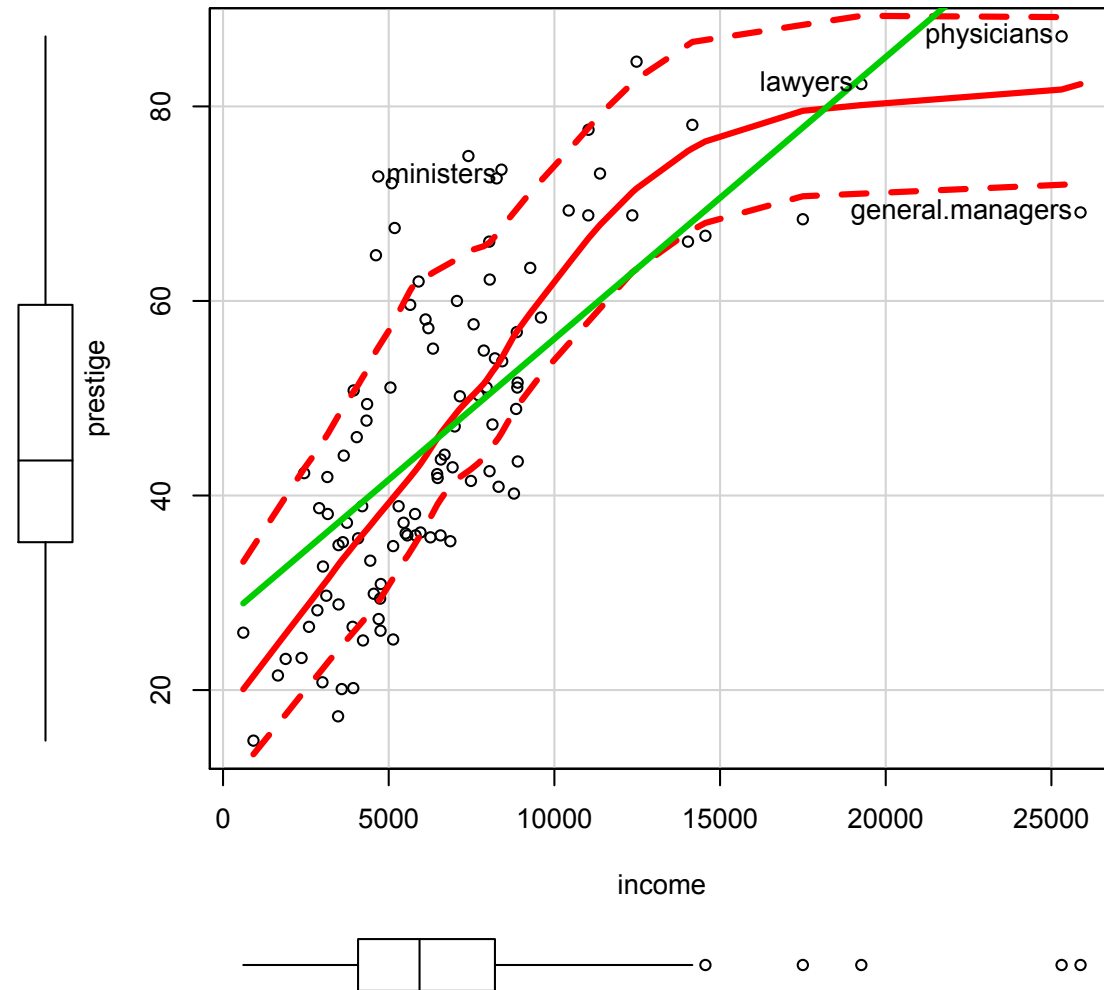
# Scatterplots

- Plot two quantitative variables
- Core to understanding regression analysis



`with(Prestige, plot(income, prestige))`

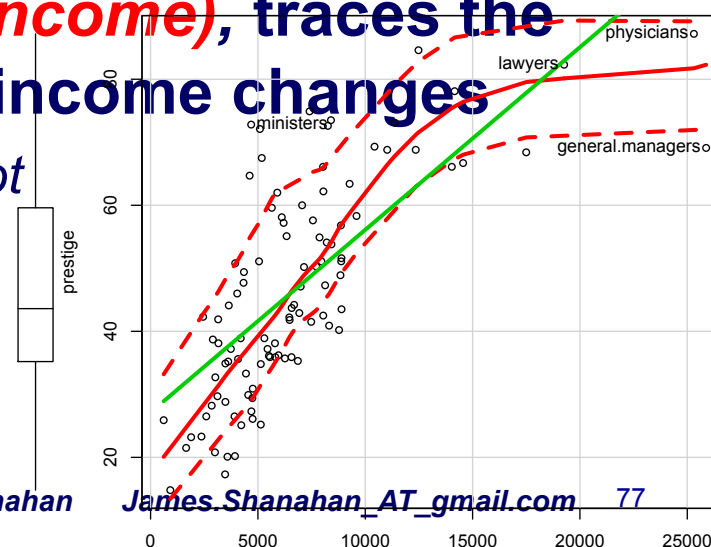
# Scatterplots



**scatterplot(prestige ~ income, span=0.6, lwd=3, id.n=4, data=Prestige)**

# Regression to mean; var

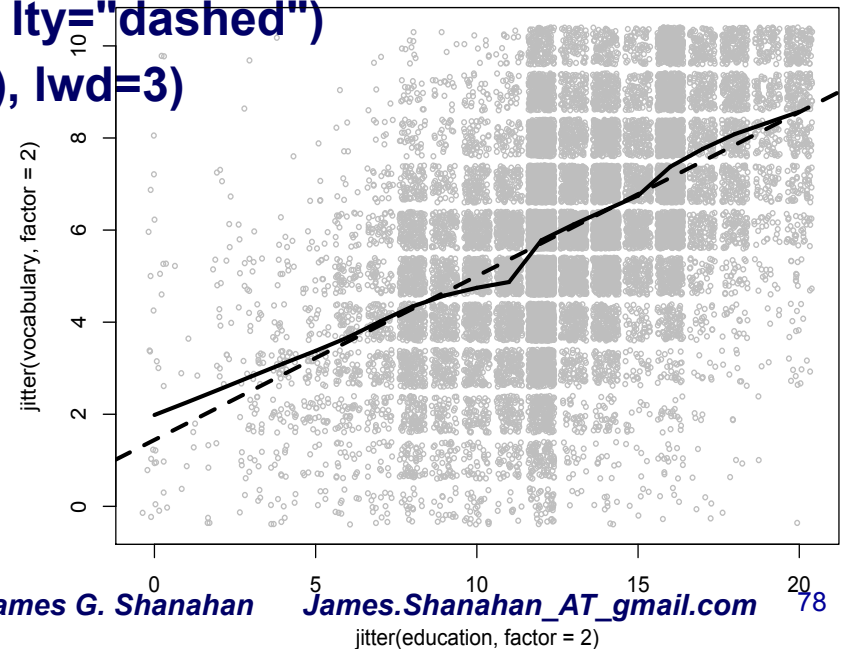
- Visualizing the conditional distributions of prestige given values of income
  - As income increase so does prestige
  - But after 10000 the value stays fixed at around 80
- **Q: is the prestige independent of income?**
- **$E(\text{prestige}|\text{income})$**  represents the mean value of prestige as the value of income varies
  - Known as conditional mean function or the regression function
- **The variance function,  $\text{Var}(\text{prestige}|\text{income})$ , traces the conditional variability in prestige as income changes**
  - That is the spread in vertical strips in the plot



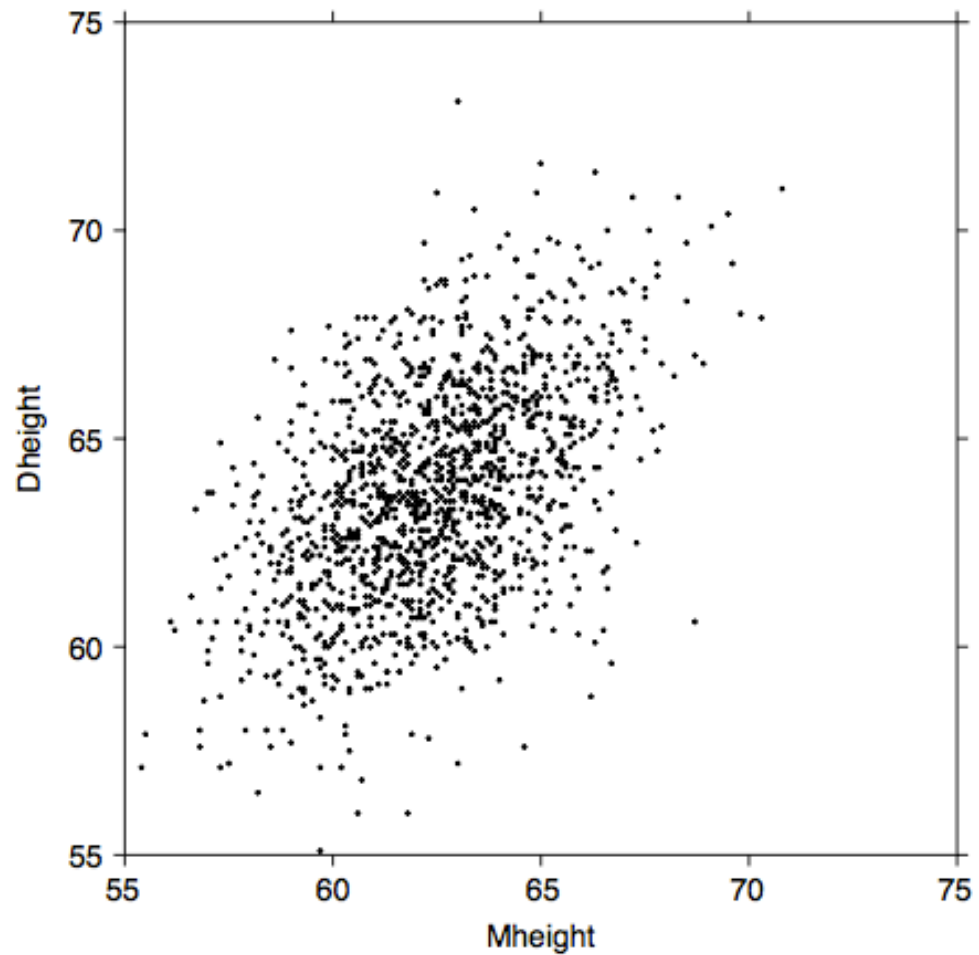
# Jitter the data so we can see it

```
head(Vocab)
nrow(Vocab)
plot(vocabulary ~ education, data=Vocab)
plot(jitter(vocabulary) ~ jitter(education), data= Vocab)
plot(jitter(vocabulary, factor=2) ~ jitter(education, factor=2),
     col="gray", cex=0.5, data=Vocab)
```

```
with(Vocab, {
  abline(lm(vocabulary ~ education), lwd=3, lty="dashed")
  lines(lowess(education, vocabulary, f=0.2), lwd=3)
})
```



# ScatterPlot with jitter



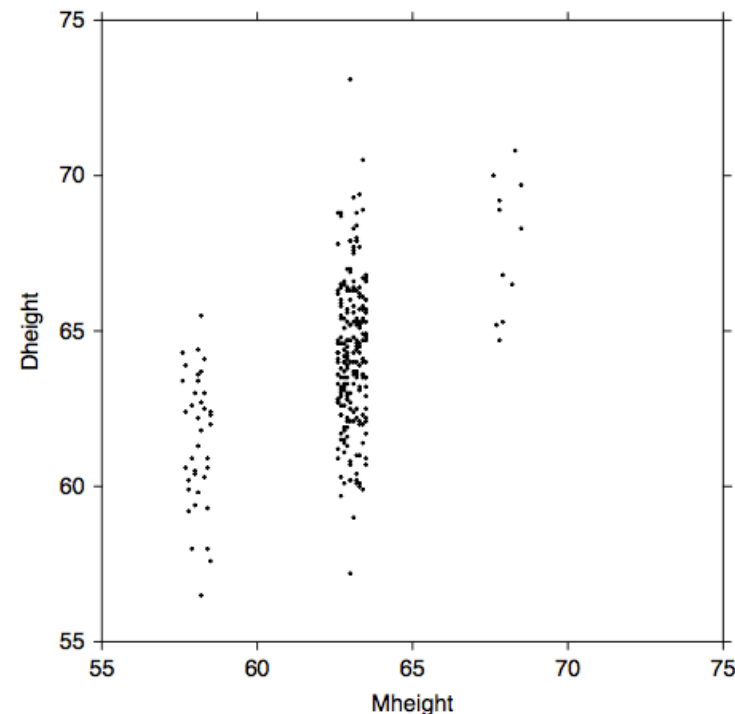
**FIG. 1.1** Scatterplot of mothers' and daughters' heights in the Pearson and Lee data. The original data have been jittered to avoid overplotting, but if rounded to the nearest inch would return the original data provided by Pearson and Lee.

# E[Daughterheight | MotherHeight]

- What we mean by this is shown in Figure 1.2, in which we show only points corresponding to mother–daughter pairs with *Mheight* rounding to either 58, 64 or 68 inches. We see that within each of these three strips or *slices*, even though the number of points is different within each slice, (a) the mean of *Dheight* is increasing from left to right, and (b) the vertical variability in *Dheight* seems to be more or less the same for each of the fixed values of *Mheight*.

E[Daughterheight | MotherHeight]

Mean is increasing  
Variance is similar





# Forbes: pressure and temperature

---

- **19<sup>th</sup> century data miner**
- **In an 1857 article, a Scottish physicist named James D. Forbes discussed a series of experiments that he had done concerning the relationship between atmospheric pressure and the boiling point of water.**
- **He knew that altitude could be determined from atmospheric pressure, measured with a barometer, with lower pressures corresponding to higher altitudes. In the middle of the nineteenth century, barometers were fragile instruments, and Forbes wondered if a simpler measurement of the boiling point of water could substitute for a direct reading of barometric pressure.**

# Residuals are quite big

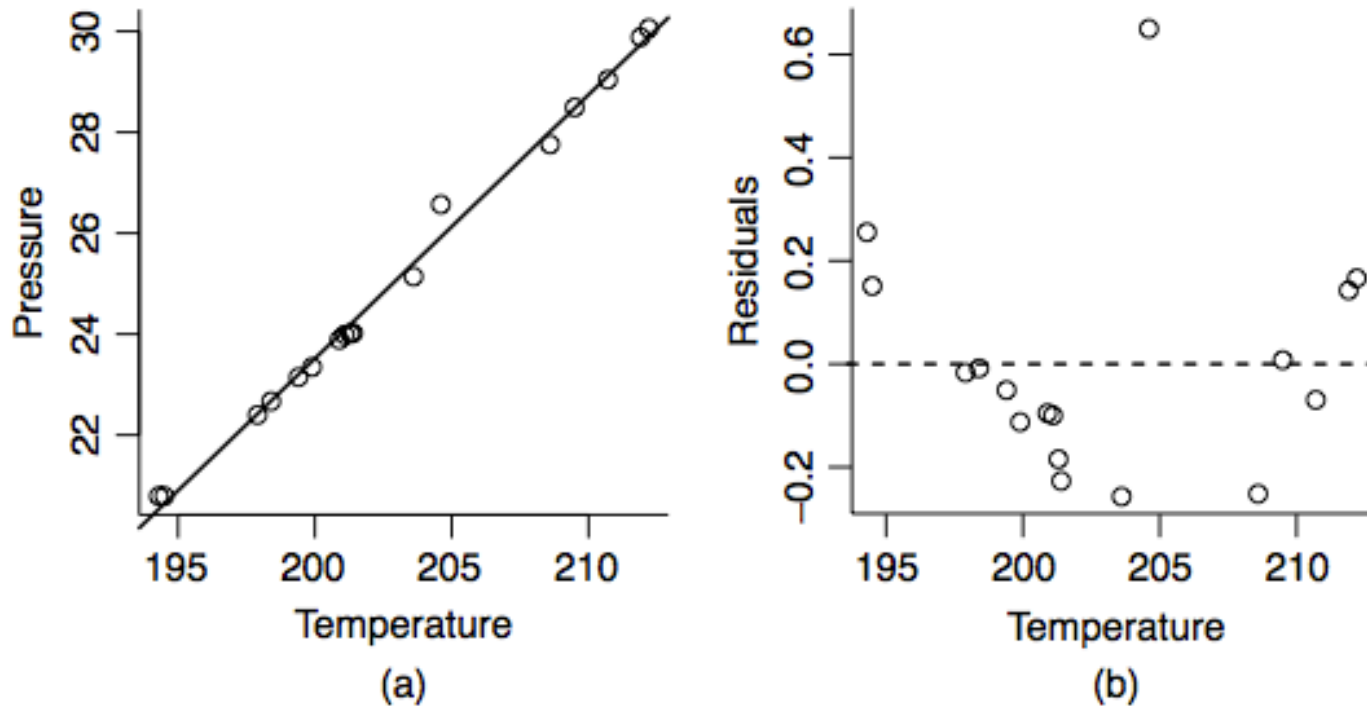
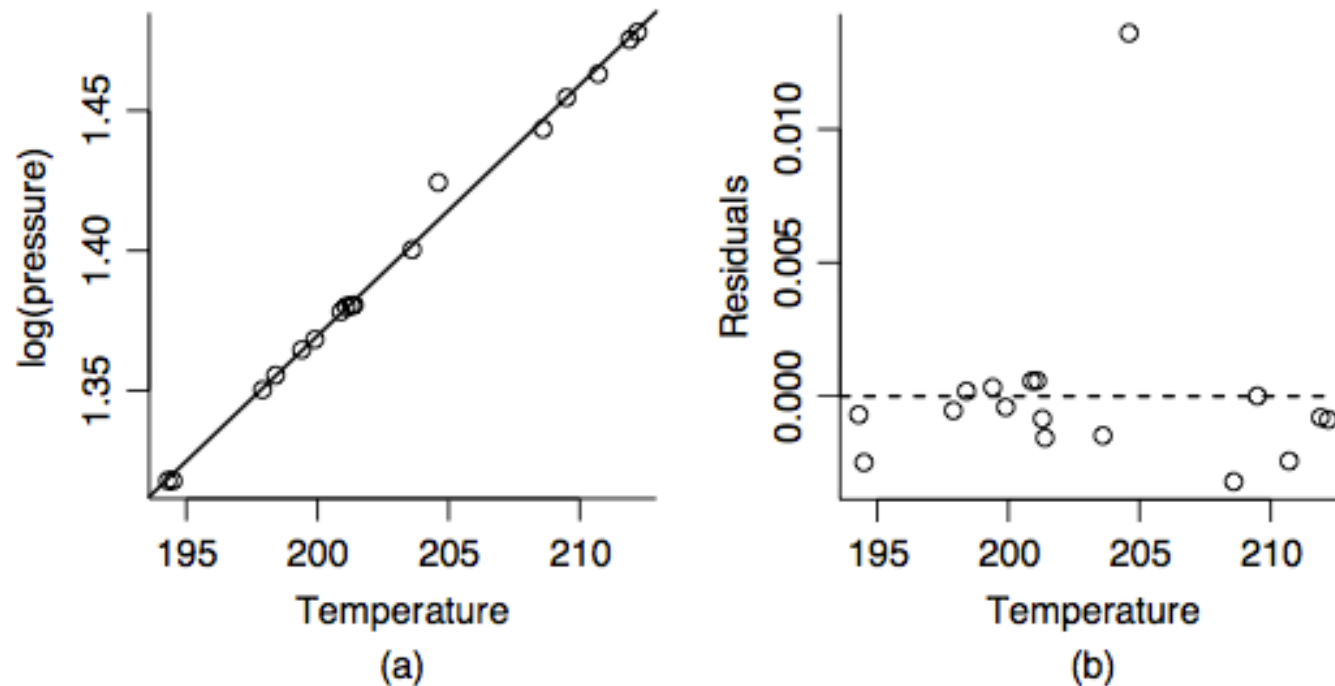


FIG. 1.3 Forbes data. (a) *Pressure* versus *Temp*; (b) *Residuals* versus *Temp*.

# Straight line: reasonable summary



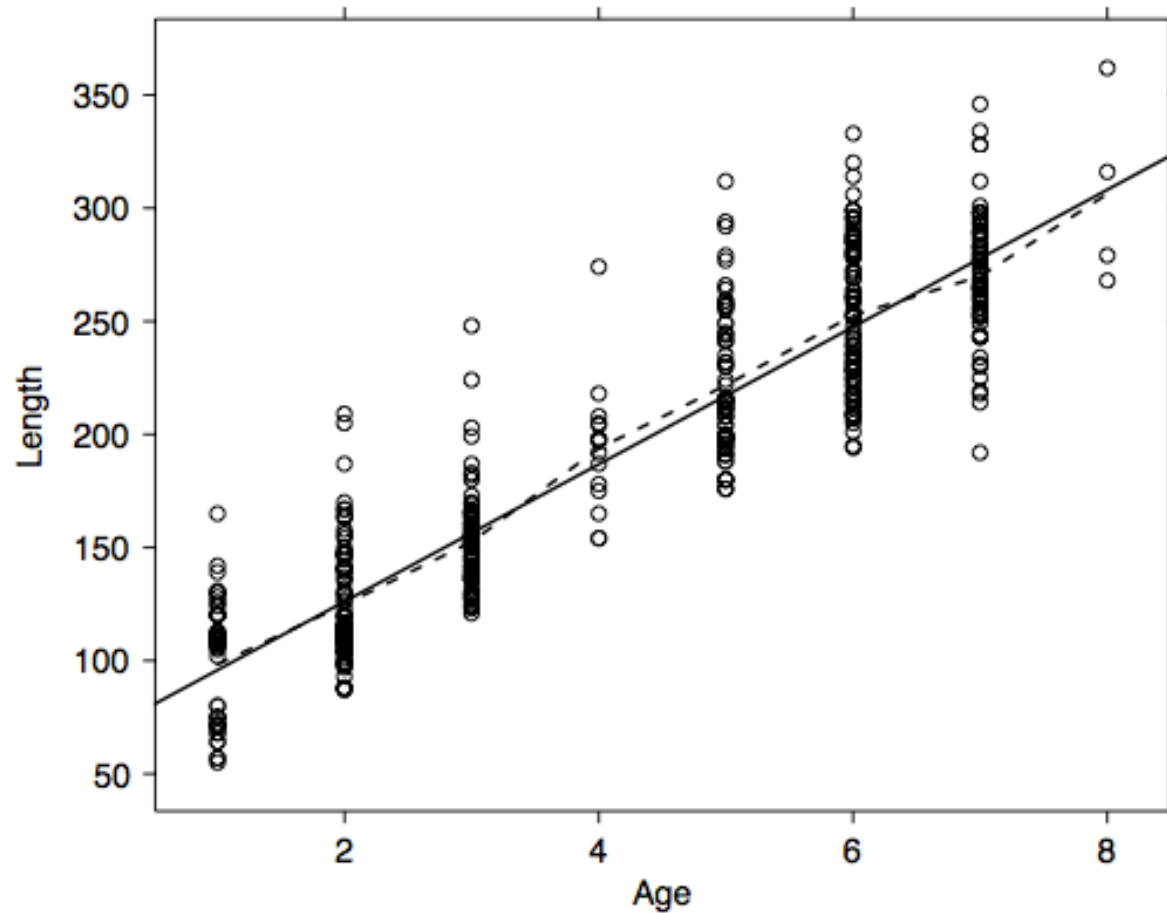
**FIG. 1.4** (a) Scatterplot of Forbes' data. The line shown is the OLS line for the regression of  $\log(\text{Pressure})$  on  $\text{Temp}$ . (b) Residuals versus  $\text{Temp}$ .

# Lecture Outline

---

- **Linear Regression: a brief intro**
- **A quick statistics review**
  - Mean, expected value, variance, stdev, quantiles, stats in R
- **Locally Weighted Linear Regression**
- **Exploratory Data Analysis**
- **Simple Linear Regression**
  - Normal Equations
  - Closed form Solution
  - Standard Error
  - Variance of the estimators
- **Good model?**

# Mean Functions



**FIG. 1.5** *Length* (mm) versus *Age* for West Bearskin Lake smallmouth bass. The solid line shown was estimated using ordinary least squares or OLS. The dashed line joins the average observed length at each age.

# Mean Functions

- Imagine a generic summary plot of Y versus X. Our interest centers on how the distribution of Y changes as X is varied. One important aspect of this distribution is the *mean function*, which we define by  $E(Y|X = x) = a$  function that depends on the value of x

$E(Y|X = x) = a$  function that depends on the value of x

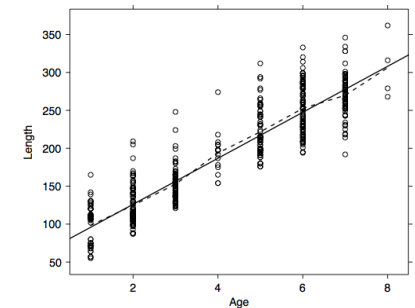


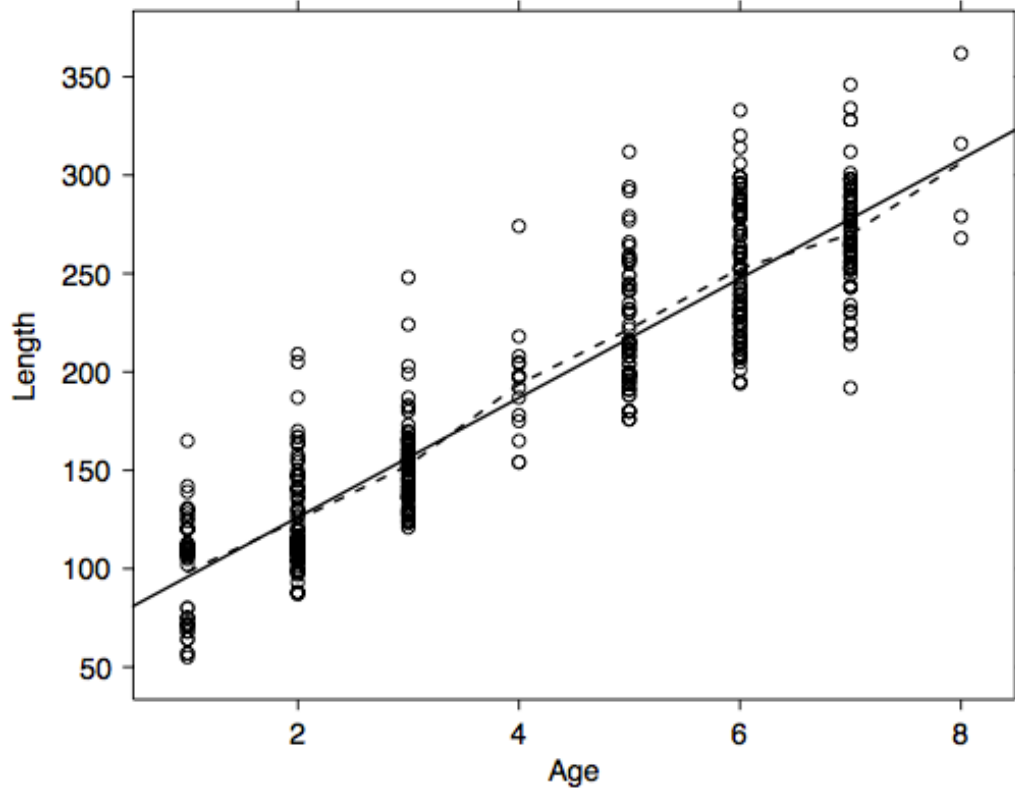
FIG. 1.5 Length (mm) versus Age for West Bearskin Lake smallmouth bass. The solid line shown was estimated using ordinary least squares or OLS. The dashed line joins the average observed length at each age.

- We read the left side of this equation as “the expected value of the response when the predictor is fixed at the value  $X = x$ ;”
  - The right side of (1.1) depends on the problem. For example, in the heights data, we might believe that

$$E(Dheight|Mheight = x) = \beta_0 + \beta_1 x$$

# Specifying the mean function

- Different ways:
  - Ordinary least squared:  $E(Dheight|Mheight = x) = \beta_0 + \beta_1x$
  - Nonparametric estimated mean function (Loess, locally weighted)



Surprisingly, the straight line and the dashed lines that join the within-age means appear to agree very closely, and we might be encouraged to use the straight-line mean function to describe these data.

*Any thoughts on OLS versus Nonparametric estimated mean function?*

1.5 Length (mm) versus Age for West Bearskin Lake smallmouth bass. The solid line shown is estimated using ordinary least squares or OLS. The dashed line joins the average observed length

# Variance Function

- Another characteristic of the distribution of the response given the predictor is the *variance function*, defined by the symbol  $\text{Var}(Y|X = x)$  and in words as the variance of the response distribution given that the predictor is fixed at  $X = x$ .
- For example, we can see that the variance function for  $Dheight|Mheight$  is approximately the same for each of the three values of  $Mheight$  shown in the graph.

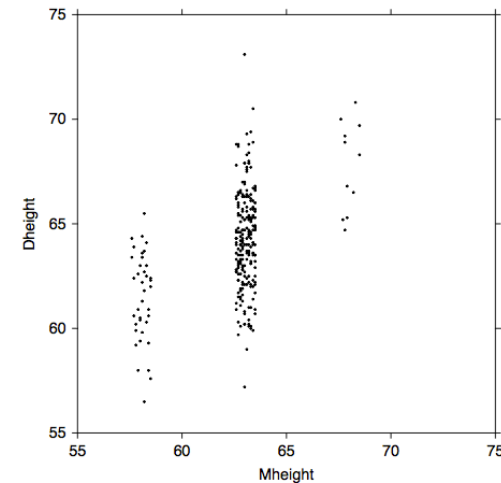


FIG. 1.2 Scatterplot showing only pairs with mother's height that rounds to 58, 64 or 68 inches.



# Assumptions of Linear Regression

---

- **A frequent assumption in fitting linear regression models is that the variance function is the same for every value of  $x$ . This is usually written as**
  - $\text{Var}(Y|X = x) = \sigma^2$
  - where  $\sigma^2$  (read “sigma squared”) is a generally unknown positive constant

# Simple Linear Regression

---

- **The *simple linear regression model* consists of the mean function and the variance function**
  - $E(Y|X = x) = \beta_0 + \beta_1 x$
  - $\text{Var}(Y|X = x) = \sigma^2$
- **The parameters in the mean function are**
  - the intercept  $\beta_0$ , which is the value of  $E(Y|X = x)$  when  $x$  equals zero,
  - and the slope  $\beta_1$ , which is the rate of change in  $E(Y|X = x)$  for a unit change in  $X$ ;
  - By varying the parameters, we can get all possible straight lines. In most applications, parameters are unknown and must be estimated using data.
  - The variance function is assumed to be constant, with a positive value  $\sigma^2$  that is usually unknown.

- `install.packages("alr3")`
- `library("alr3")`
- For R users, scripts can be obtained while you are running R and also connected to the internet. To get the script for Chapter 2 for this primer, for example, you could type
- To get the script for Chapter 2 of the text, use
  - `alrWeb(script = 'chapter2')`

# Applied Linear Regression, Third edition (Chapter 2)

```
# Applied Linear Regression, Third edition  
# Chapter 2  
# October 14, 2004; revised January 2011 for alr3 Version 2.0, R only
```

```
# Fig. 2.1 in the new edition  
# R only
```

```
x <- c(0, 4)  
y <- c(0, 4)  
plot(x, y, type="n", xlab="Predictor = X", ylab="E(Y|X=x)")  
abline(.8, 0.7)  
x<-c(2, 3, 3)  
y<-c(2.2, 2.2, 2.9)  
lines(x, y)  
lines(c(0, 0), c(0, .8), lty=2)  
lines(c(0, 4), c(0, 0), lty=2)  
text(3.05, 2.5, expression(beta[1] == Slope), adj="right")  
text(.05, .4, expression(beta[0] == Intercept), adj=0)  
text(2.5, 1.8, "1")
```

alrWeb(script = 'chapter2')  
<http://www.stat.umn.edu/alr/Links/scripts/chapter2.R>

# Simple Linear Regression

---

- **The *simple linear regression model* consists of the mean function and the variance function**
  - $E(Y|X = x) = \beta_0 + \beta_1 x$
  - $\text{Var}(Y|X = x) = \sigma^2$
- **The parameters in the mean function are**
  - the intercept  $\beta_0$ , which is the value of  $E(Y|X = x)$  when  $x$  equals zero,
  - and the slope  $\beta_1$ , which is the rate of change in  $E(Y|X = x)$  for a unit change in  $X$ ;
  - By varying the parameters, we can get all possible straight lines. In most applications, parameters are unknown and must be estimated using data.
  - The variance function is assumed to be constant, with a positive value  $\sigma^2$  that is usually unknown.

# Assumptions of Linear Regression

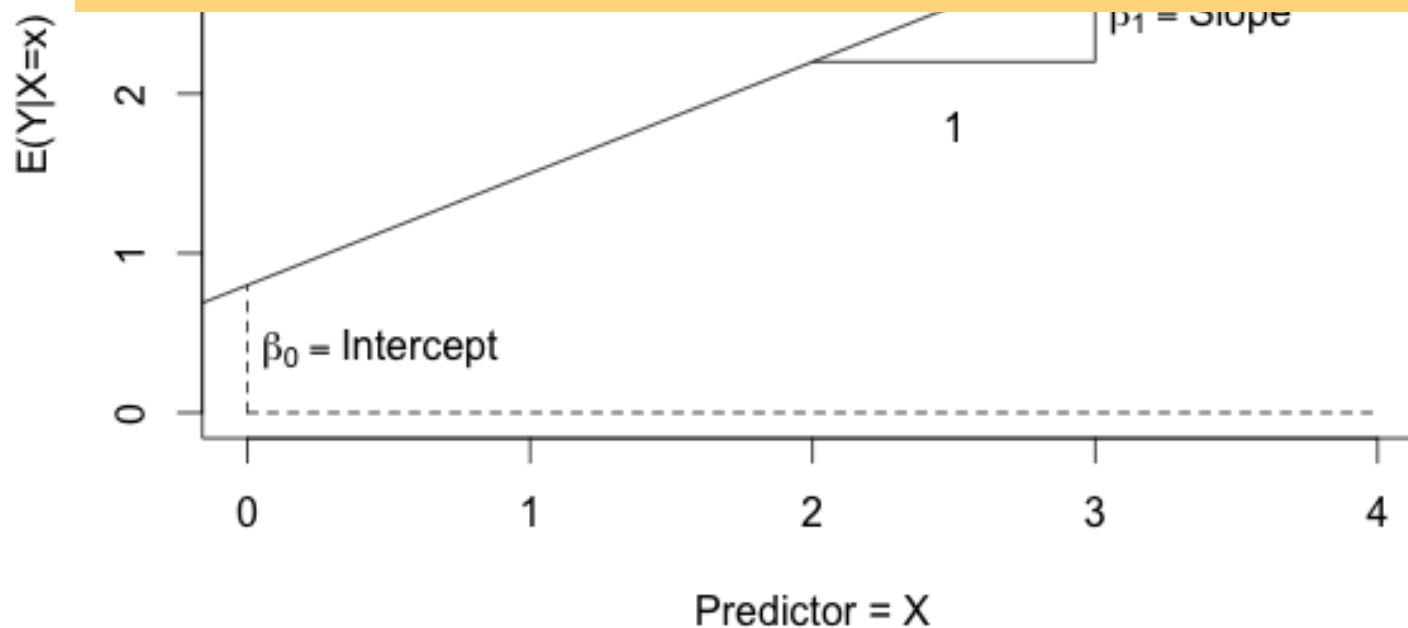
---

- Each example is independent of every other example
- Predictors can be numerical, qualitative, or ordinal
- Additional regressor variables can be generated using interactions
- The dependence of the response on the predictors is through the conditional expected value
  - $E(Y|X = x) = \beta_0 + \beta_1 x$
  - $\text{Var}(Y|X = x) = \sigma^2$       #conditional variance

```

x <- c(0, 4)
y <- c(0, 4)
plot(x, y, type="n", xlab="Predictor = X", ylab="E(Y|X=x)")
abline(.8, 0.7)
x<-c(2, 3, 3)
y<-c(2.2, 2.2, 2.9)
lines(x, y)
lines(c(0, 0), c(0, .8), lty=2)
lines(c(0, 4), c(0, 0), lty=2)
text(3.05, 2.5, expression(beta[1] == Slope), adj=0)
text(.05, .4, expression(beta[0] == Intercept), adj=0)
text(2.5, 1.8, "1")

```



# Choose parameters that minimize RSS

---

*Parameters* are unknown quantities that characterize a model. *Estimates of parameters* are computable functions of data and are therefore *statistics*. To keep this distinction clear, parameters are denoted by Greek letters like  $\alpha$ ,  $\beta$ ,  $\gamma$  and  $\sigma$ , and estimates of parameters are denoted by putting a “hat” over the corresponding Greek letter. For example,  $\hat{\beta}_1$ , read “beta one hat,” is the estimator of  $\beta_1$ , and  $\hat{\sigma}^2$  is the estimator of  $\sigma^2$ . The *fitted value* for case  $i$  is given by  $\widehat{E}(Y|X = x_i)$ , for which we use the shorthand notation  $\hat{y}_i$ ,

$$\hat{y}_i = \widehat{E}(Y|X = x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (2.2)$$

Although the  $e_i$  are not parameters in the usual sense, we shall use the same hat notation to specify the residuals: the residual for the  $i$ th case, denoted  $\hat{e}_i$ , is given by the equation

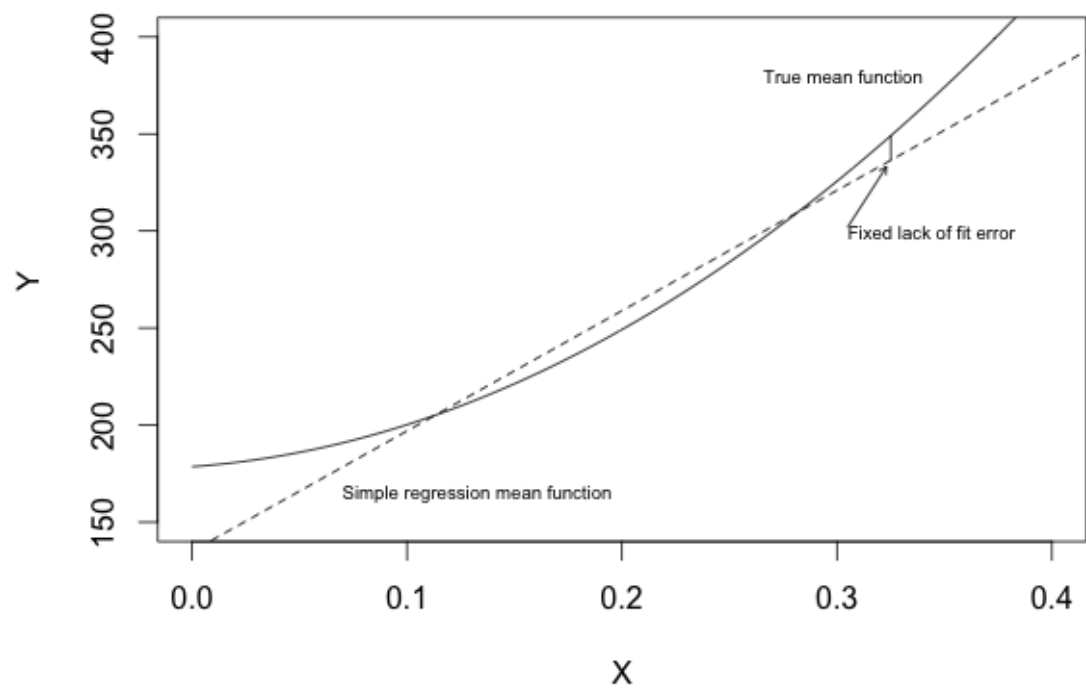
$$\hat{e}_i = y_i - \widehat{E}(Y|X = x_i) = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \quad i = 1, \dots, n \quad (2.3)$$

which should be compared with the equation for the statistical errors,

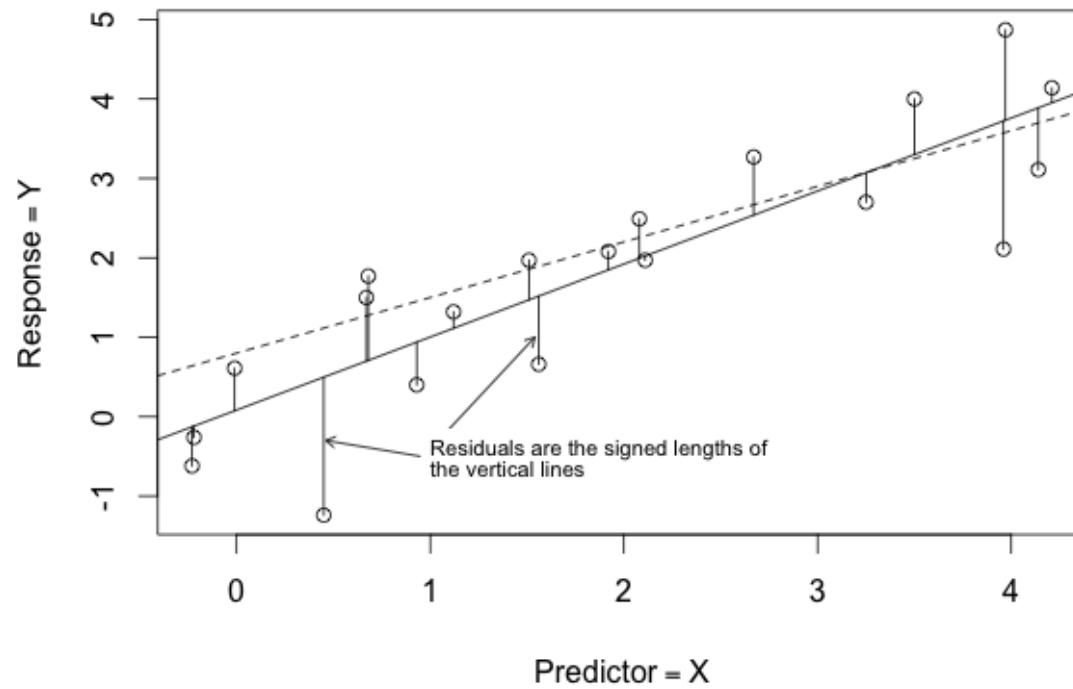
**Error**

$$e_i = y_i - (\beta_0 + \beta_1 x_i) \quad i = 1, \dots, n$$





# Residuals



# Standardizing the data

---

## Standardizing normal random variables

[edit]

As a consequence of property 1, it is possible to relate all normal random variables to the standard normal. For example if  $X$  is normal with mean  $\mu$  and variance  $\sigma^2$ , then

$$Z = \frac{X - \mu}{\sigma}$$

has mean zero and unit variance, that is  $Z$  has the standard normal distribution. Conversely, having a standard normal random variable  $Z$  we can always construct another normal random variable with specific mean  $\mu$  and variance  $\sigma^2$ :

$$X = \sigma Z + \mu.$$

This "standardizing" transformation is convenient as it allows one to compute the pdf and especially the cdf of a normal distribution having the table of pdf and cdf values for the standard normal. They will be related via

$$F_X(x) = \Phi\left(\frac{x - \mu}{\sigma}\right), \quad f_X(x) = \frac{1}{\sigma} \phi\left(\frac{x - \mu}{\sigma}\right).$$

**In R for the help page try  
?scale**

- All least squares computations for simple regression depend only on averages, sums of squares and sums of cross-products. Definitions of the quantities used are given in Table 2.1. Sums of squares and cross-products have been centered by subtracting the average from each of the values before squaring or taking cross-products.

**TABLE 2.1** Definitions of Symbols<sup>a</sup>

Quantity	Definition	Description
$\bar{x}$	$\sum x_i/n$	Sample average of $x$
$\bar{y}$	$\sum y_i/n$	Sample average of $y$
$SXX$	$\sum (x_i - \bar{x})^2 = \sum (x_i - \bar{x})x_i$	Sum of squares for the $x$ 's
$SD_x^2$	$SXX/(n - 1)$	Sample variance of the $x$ 's
$SD_x$	$\sqrt{SXX/(n - 1)}$	Sample standard deviation of the $x$ 's
$SYY$	$\sum (y_i - \bar{y})^2 = \sum (y_i - \bar{y})y_i$	Sum of squares for the $y$ 's
$SD_y^2$	$SYY/(n - 1)$	Sample variance of the $y$ 's
$SD_y$	$\sqrt{SYY/(n - 1)}$	Sample standard deviation of the $y$ 's
$SXY$	$\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i - \bar{x})y_i$	Sum of cross-products
$s_{xy}$	$SXY/(n - 1)$	Sample covariance
$r_{xy}$	$s_{xy}/(SD_x SD_y)$	Sample correlation

<sup>a</sup>In each equation, the symbol  $\sum$  means to add over all the  $n$  values or pairs of values in the data.

# OLS Closed Form

---

The OLS estimators are those values  $\beta_0$  and  $\beta_1$  that minimize the function<sup>1</sup>

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2 \quad (2.4)$$

When evaluated at  $(\hat{\beta}_0, \hat{\beta}_1)$ , we call the quantity  $RSS(\hat{\beta}_0, \hat{\beta}_1)$  the *residual sum of squares*, or just *RSS*.

The least squares estimates can be derived in many ways, one of which is outlined in Appendix A.3. They are given by the expressions

$$\hat{\beta}_1 = \frac{SXY}{SXX} = r_{xy} \frac{SD_y}{SD_x} = r_{xy} \left( \frac{SYY}{SXX} \right)^{1/2} \quad (2.5)$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

The several forms for  $\hat{\beta}_1$  are all equivalent.

**Since OLS minimizes (2.4), it will always fit at least as well as, and generally better than, the true mean function (actual function); OLS model is biased by data..**

# Closed form solution to OLS

How do we minimize (3.2)? Denote by  $\mathbf{X}$  the  $N \times (p + 1)$  matrix with each row an input vector (with a 1 in the first position), and similarly let  $\mathbf{y}$  be the  $N$ -vector of outputs in the training set. Then we can write the residual sum-of-squares as

$\beta$  is  $W$  in our notation

$$\text{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta). \quad (3.3)$$

This is a quadratic function in the  $p + 1$  parameters. Differentiating with respect to  $\beta$  we obtain

$$\begin{aligned} \frac{\partial \text{RSS}}{\partial \beta} &= -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) \\ \frac{\partial^2 \text{RSS}}{\partial \beta \partial \beta^T} &= 2\mathbf{X}^T\mathbf{X}. \end{aligned} \quad (3.4)$$

Assuming (for the moment) that  $\mathbf{X}$  has full column rank, and hence  $\mathbf{X}^T\mathbf{X}$  is positive definite, we set the first derivative to zero

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$$

to obtain the unique solution

$\beta$  is computed directly in closed form

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \quad (3.6)$$

[Friedman et al. 2001]  
James.Shanahan\_AT\_gmail.com 102

# Closed form solution to OLS

- To minimize  $J$  (aka RSS), we set its derivatives to zero, and obtain the normal equations:
  - $X^T X W = X^T y$
  - Thus the value of  $W$  that minimizes  $J(W)$  is give in closed form

$$\begin{aligned}\nabla J_{W_j}(W) &= \frac{\partial}{\partial W_j} J(W) = \frac{\partial}{\partial W_j} \left( \frac{1}{2} (f_W(x) - y)^2 \right) \\ &= 2 * \frac{1}{2} (f_W(x) - y) \frac{\partial}{\partial W_j} (f_W(x) - y) \\ &= (f_W(x) - y) \frac{\partial}{\partial W_j} \left( \left( \sum_{i=0}^n w_i x_i \right) - y \right) \\ &= (f_W(x) - y) x_j \quad \text{for each } j \text{ in } 1:n \\ &= (XW - Y)^T X \quad \text{overall and in terms of data} \\ &= X^T X W - X^T Y = 0 \\ &= X^T X W = X^T Y \quad \text{Normal Equations} \\ &= W = (X^T X)^{-1} X^T Y\end{aligned}$$

- For a full derivation see: <http://www.stanford.edu/class/cs229/notes/cs229->

notes1.pdf

# Normal Equations → Closed Form Soln. to OLS

---

- **An alternative is to performing the minimization explicitly and without resorting to an iterative algorithm**
  - In this method, we will minimize RSS by explicitly taking its derivatives with respect to the  $\beta_j$ 's (sometimes written as  $W$ , the weight vector), and setting them to zero.
  - Do this via calculus with matrices.
- **Gradient descent gives another way of minimizing  $RSS(\beta)$ . [Discussed next lecture]**



# Closed form solution to OLS

- To minimize RSS, we set its derivatives to zero, and obtain the normal equations:

$$- X^T X W = X^T y$$

*RSS = Variance of  $\varepsilon$*

$$\begin{aligned} 0 &= \frac{\partial \sum \hat{\varepsilon}_i^2}{\partial W} = \frac{\partial (y_i - XW)^2}{\partial W} & 0 &= \frac{\partial \sum \hat{\varepsilon}_i^2}{\partial \hat{\beta}_1} = \frac{\partial \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_1} \\ &= -2X(y - \hat{\beta}_0 - \hat{\beta}_1 x_i) & &= -2 \sum x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\ &= -2 \sum x_i (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i) & &= -2 \sum x_i (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i) \end{aligned}$$

For another derivation see:

<http://www.stanford.edu/class/cs229/notes/cs229-notes1.pdf>

# Derivation of Parameter Equations

---

- **An Alternative Derivation** treating the y-intercept and the variable coefficients separately; here we represent  $W$  as  $\beta$ .
- **Goal: Minimize squared error (WRT to the y-intercept)**

$$0 = \frac{\partial \sum \hat{\varepsilon}_i^2}{\partial \hat{\beta}_0} = \frac{\partial \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_0}$$

$$= \sum -2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$= -2(n\bar{y} - n\hat{\beta}_0 - n\hat{\beta}_1 \bar{x})$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

# OLS Via Gradient Descent: The Gradient

$$W_{j,t+1} = W_{j,t} - \alpha * \nabla J_{w_j}(W_{j,t})$$

- In order to implement this algorithm, we have to work out what is the partial derivative term at time t on the right hand side  $\nabla f_{w_j}(W) = dF(W)/dw_j$ .
- Assume we have only one training example (x, y), so that we can drop the sum in the definition of J.

$$\begin{aligned} \nabla J_{W_j}(W) &= \frac{\partial}{\partial W_j} J(W) &= \frac{\partial}{\partial W_j} \left( \frac{1}{2} (f_W(x) - y)^2 \right) && \text{Use chain rule } df/du * du/dx \\ &= 2 * \frac{1}{2} (f_W(x) - y) \frac{\partial}{\partial W_j} (f_W(x) - y) && \text{Assume a single training example} \\ &= (f_W(x) - y) \frac{\partial}{\partial W_j} \left( \left( \sum_{i=0}^n w_i x_i \right) - y \right) && \text{For a single } w_j \\ &= (f_W(x) - y) x_j \end{aligned}$$

**Recall**

$$\begin{aligned} \frac{\partial}{\partial W_j} \left( \left( \sum_{i=0}^n w_i x_i \right) - y \right) &= \frac{\partial}{\partial W_j} (w_0 x_0 + w_1 x_1 + \dots + w_j x_j + \dots + w_n x_n) \\ &= 0 + 0 + \dots + x_j + \dots + 0 \end{aligned}$$

# Forbes Model

---

Using Forbes' data, we will write  $\bar{x}$  to be the sample mean of *Temp* and  $\bar{y}$  to be the sample mean of *Lpres*. The quantities needed for computing the least squares estimators are

$$\begin{aligned}\bar{x} &= 202.95294 & SXX &= 530.78235 & SXY &= 475.31224 \\ \bar{y} &= 139.60529 & SYY &= 427.79402\end{aligned}\tag{2.6}$$

The quantity *SYY*, although not yet needed, is given for completeness. In the rare instances that regression calculations are not done using statistical software or a statistical calculator, intermediate calculations such as these should be done as accurately as possible, and rounding should be done only to final results. Using (2.6), we find

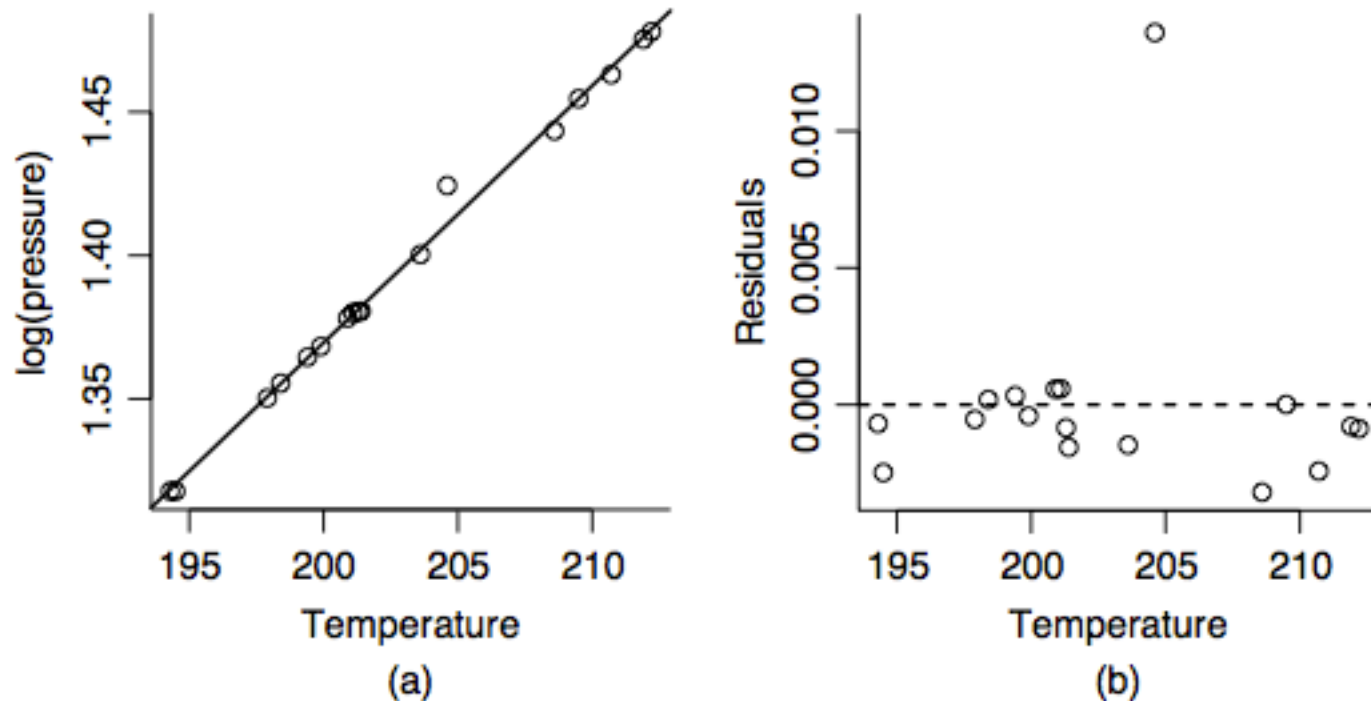
$$\begin{aligned}\hat{\beta}_1 &= \frac{SXY}{SXX} = 0.895 \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} = -42.138\end{aligned}$$

The estimated line, given by either of the equations

$$\begin{aligned}\hat{E}(Lpres|Temp) &= -42.138 + 0.895Temp \\ &= 139.606 + 0.895(Temp - 202.953)\end{aligned}$$

# Forbes Model

Using Forbes' data, we will write  $\bar{x}$  to be the sample mean of  $Temp$  and  $\bar{y}$  to be



**FIG. 1.4** (a) Scatterplot of Forbes' data. The line shown is the OLS line for the regression of  $\log(Pressure)$  on  $Temp$ . (b) Residuals versus  $Temp$ .

$$\begin{aligned}\hat{E}(Lpres|Temp) &= -42.138 + 0.895Temp \\ &= 139.606 + 0.895(Temp - 202.953)\end{aligned}$$

# Lecture Outline

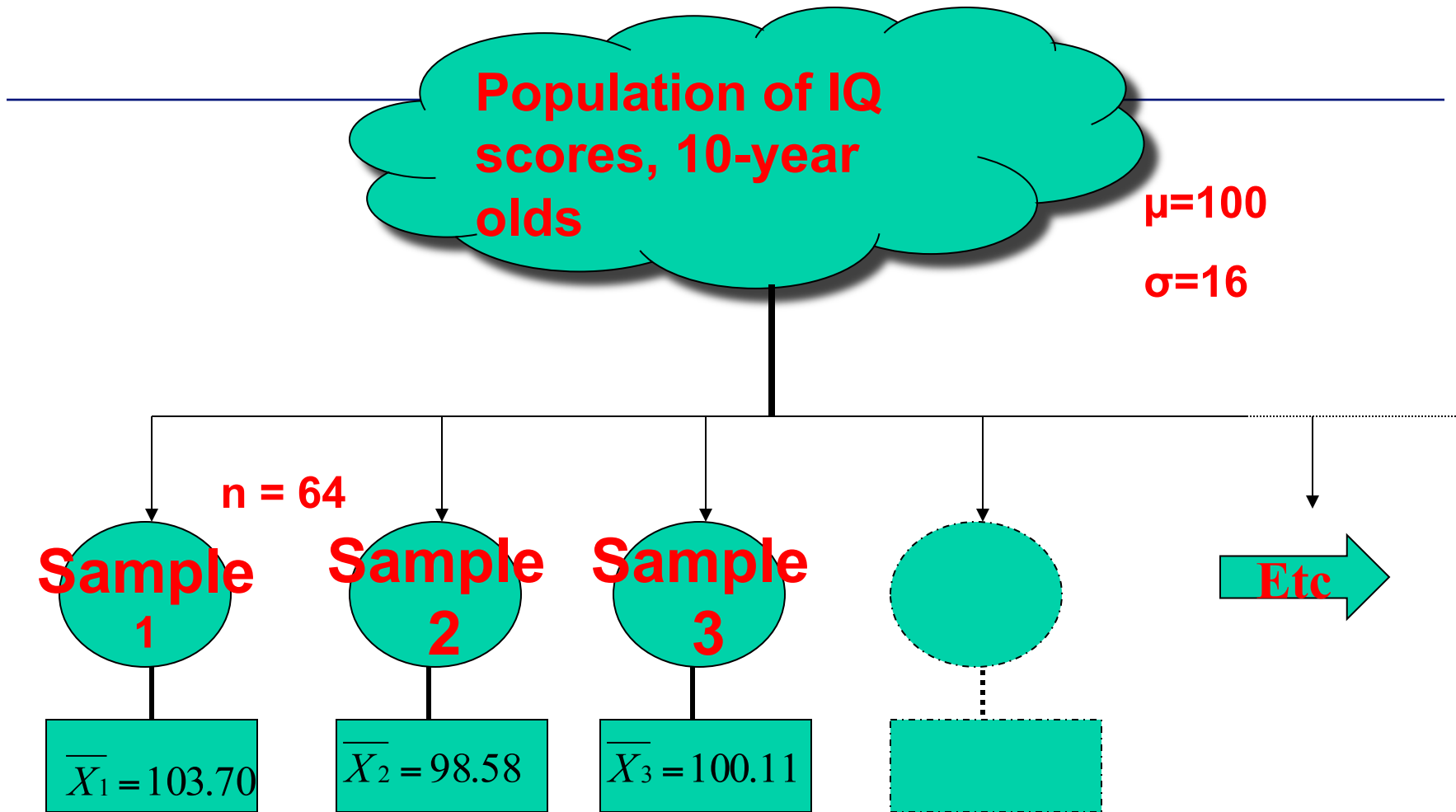
---

- **Linear Regression: a brief intro**
- **A quick statistics review**
  - Mean, expected value, variance, stdev, quantiles, stats in R
- **Locally Weighted Linear Regression**
- **Exploratory Data Analysis**
- **Simple Linear Regression**
  - Normal Equations
  - Closed form Solution
  - Standard Error
  - Variance of the estimators
- **Good model?**

# Standard Error

---

- The standard error is the standard deviation of the sampling distribution of a statistic.<sup>[1]</sup>
- The term may also be used to refer to an estimate of that standard deviation, derived from a particular sample used to compute the estimate.
- For example, the sample mean is the usual estimator of a population mean. However, different samples drawn from that same population would in general have different values of the sample mean. The standard error of the mean (i.e., of using the sample mean as a method of estimating the population mean) is the standard deviation of those sample means over all possible samples (of a given size) drawn from the population. Secondly, the standard error of the mean can refer to an estimate of that standard deviation, computed from the sample of data being analyzed at the time.



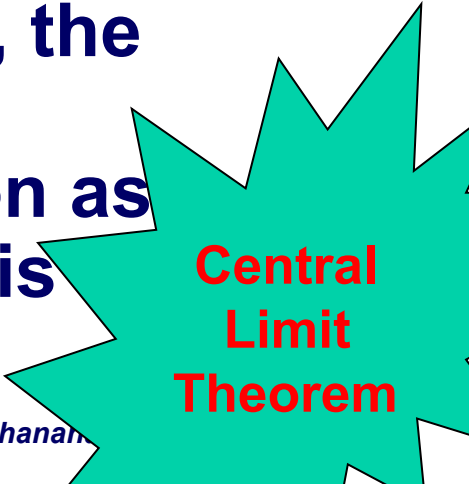
Is sample 2 a likely representation of our population?



# Distribution of Sample Means

---

1. The mean of a sampling distribution is identical to mean of raw scores in the population ( $\mu$ )
2. If the population is Normal, the distribution of sample means is also Normal
3. If the population is not Normal, the distribution of sample means approaches Normal distribution as the size of sample on which it is based gets larger



**Central  
Limit  
Theorem**

# Standard Error of the Mean

---

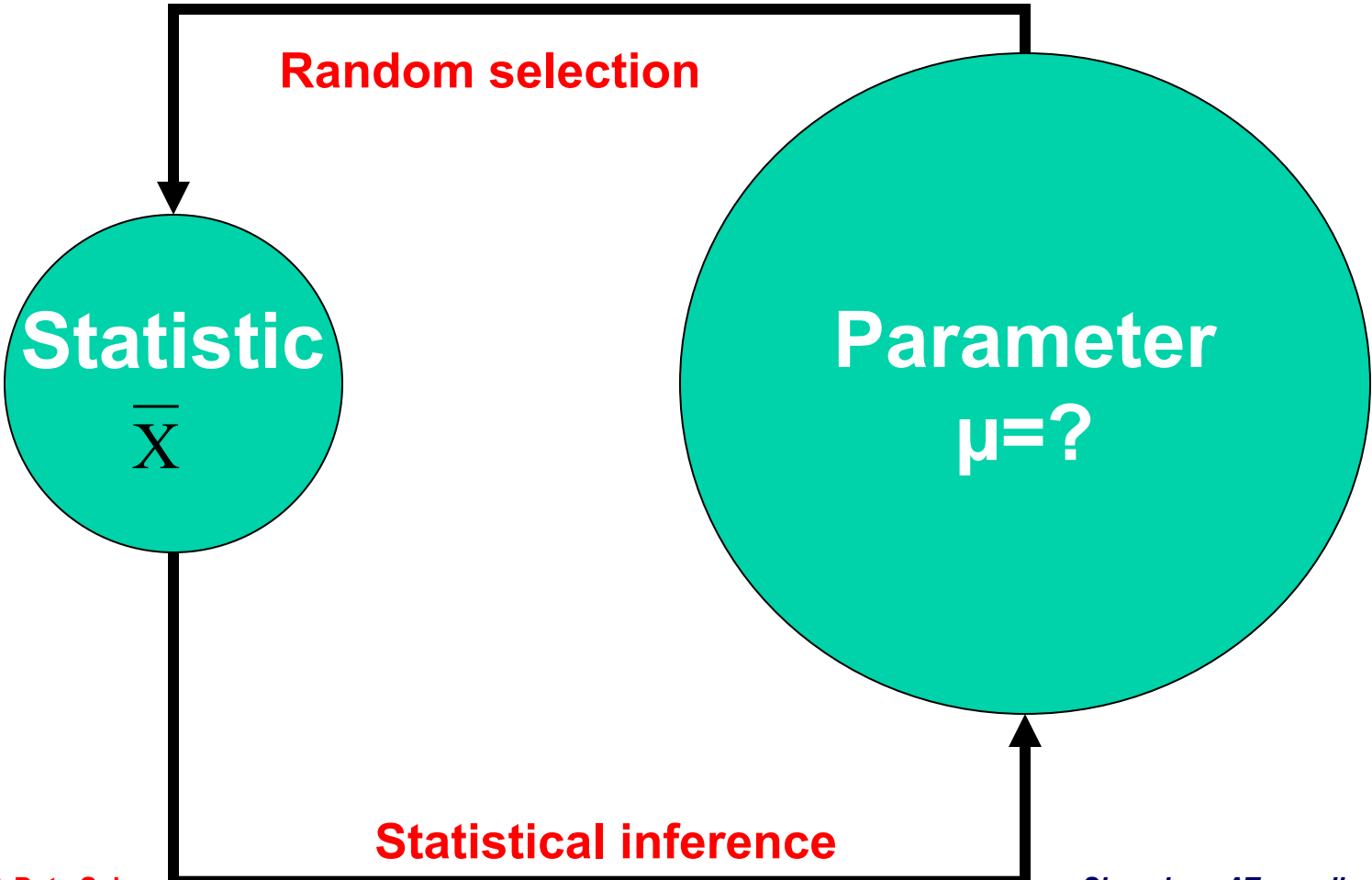
- The standard deviation of means in a sampling distribution is known as the **standard error of the mean**.
- It can be calculated from the standard deviation of observations

$$S_{\bar{X}} = \frac{s}{\sqrt{n}} \quad \boxed{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$$

3. The larger our sample size, the smaller our standard error

Sample of observations

Entire population of observations



# Estimation Procedures

---

- **Point estimates**

- For example mean of a sample of 25 patients
  - No information regarding probability of accuracy
- Interval estimates
- Estimate a range of values that is likely
  - Confidence interval between two limit values
    - The degree of confidence depends on the probability of including the population mean  $\mu$

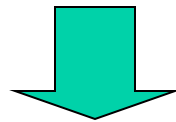
$$95\% \text{ CI} = \bar{X} \pm 1.96 \sigma_{\bar{x}}$$

$$99\% \text{ CI} = \bar{X} \pm 2.58 \sigma_{\bar{x}}$$

# When Sample size is small ...

---

$$95\% \text{ CI} = \bar{X} \pm 1.96 \sigma_{\bar{x}}$$



$$95\% \text{ CI} = \bar{X} \pm t S_{\bar{x}}$$

**A constant from  
Student t Distribution  
that depends on confidence  
interval and sample size**

# HYPOTHESIS TESTING

---

- Hygiene procedures are effective in preventing cold.
- State 2 hypotheses:
- Null:  $H_0$  : Hand-washing has no effect on bacteria counts.
- Alternative:  $H_a$  : Hand-washing reduces bacteria.
- The null hypothesis is assumed true: i.e., the defendant is assumed to be innocent.

# $\epsilon$ determines the properties of the response $y$

---

- Suppose we can fix the value of  $x$  and observe the corresponding value of the response  $y$ . Now if  $x$  is fixed, the random component  $\epsilon$  determines the properties of  $y$ .
- Suppose the mean and variance of  $\epsilon$  are 0 and  $\sigma^2$ , respectively. Then the mean response at any value of the regressor variable ( $x$ ) is

- $E(y|x) = \mu_{y|x} = \epsilon \sim N(0, \sigma^2)$

$$E(y|x) = \mu_{y|x} = E(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x$$

- The variance of
  - $\text{Var}(y|x) = \sigma^2$
- The variability of the response is constant, that is, the variance of the response is the same at all values of  $x$ . This means that there is a distribution of  $y$  for each value of  $x$ , and the variance of this distribution is constant.

Now if  $x$  is fixed, the random component  $\epsilon$  determines the properties of  $y$ .

$$\text{Var}(y|x) = \sigma_{y|x}^2 = \text{Var}(\beta_0 + \beta_1 x + \epsilon) = \sigma^2$$

- Small  $\sigma^2$  implies the observed values  $y$  will fall close to the line.

# Estimating Variance based on Residual

---

Since the variance  $\sigma^2$  is essentially the average squared size of the  $e_i^2$ , we should expect that its estimator  $\hat{\sigma}^2$  is obtained by averaging the squared residuals.

Since the variance  $\sigma^2$  is essentially the average squared size of the  $e_i^2$ , we should expect that its estimator  $\hat{\sigma}^2$  is obtained by averaging the squared residuals. Under the assumption that the errors are uncorrelated random variables with zero means and common variance  $\sigma^2$ , an unbiased estimate of  $\sigma^2$  is obtained by dividing  $RSS = \sum \hat{e}_i^2$  by its *degrees of freedom* (df), where residual df = number of cases minus the number of parameters in the mean function. For simple regression, residual df =  $n - 2$ , so the estimate of  $\sigma^2$  is given by

$$\hat{\sigma}^2 = \frac{RSS}{n - 2} \quad (2.7)$$

This quantity is called the *residual mean square*. In general, any sum of squares divided by its df is called a mean square. The residual sum of squares can be computed by squaring the residuals and adding them up. It can also be computed from the formula (Problem 2.9)

$$RSS = SYY - \frac{SXY^2}{SXX} = SYY - \hat{\beta}_1^2 SXX \quad (2.8)$$



# Standard Error : Same units as response variable

---

Using the summaries for Forbes' data given at (2.6), we find

$$\begin{aligned}RSS &= 427.79402 - \frac{475.31224^2}{530.78235} \\ &= 2.15493\end{aligned}\tag{2.9}$$

$$\sigma^2 = \frac{2.15493}{17 - 2} = 0.14366\tag{2.10}$$

The square root of  $\hat{\sigma}^2$ ,  $\hat{\sigma} = \sqrt{0.14366} = 0.37903$  is often called the *standard error of regression*. It is in the same units as is the response variable.

# Lecture Outline

---

- **Linear Regression: a brief intro**
- **A quick statistics review**
  - Mean, expected value, variance, stdev, quantiles, stats in R
- **Locally Weighted Linear Regression**
- **Exploratory Data Analysis**
- **Simple Linear Regression**
  - Normal Equations
  - Closed form Solution
  - Standard Error
  - Variance of the estimators
- **Good model?**

# Good model

---

- **Lower residual standard error is better**
- **More to come on this front next class**

---

- **End of Lecture**

# Guidelines for Homework

---

- **These exercises are OPTIONAL.**
- **GENERAL Guidelines for Homework**
  - Paste each question into your manuscript and then provide your solution
  - Please provide explanations, code, graphs captions, and cross references in a PDF report (that should read like research paper.
  - Don't forget to put your name, email and date of submission on each report.
  - In addition, please provide R code in separate file. Please comment your so that I or anybody else can understand it and please cross reference code with problem numbers and descriptions
  - Please create a separate driver function for each exercise or exercise part (and comment!)
  - If you have questions please raise them in class or via email or during office hours
  - Homework is due on Tuesday, February 21 of the following week by 5PM.
  - Please submit your homework by email to: [James.Shanahan@gmail.com](mailto:James.Shanahan@gmail.com) with the subject **“Berkeley I 296A”**
  - Have fun!

No grades  
If subject  
line is not  
correct

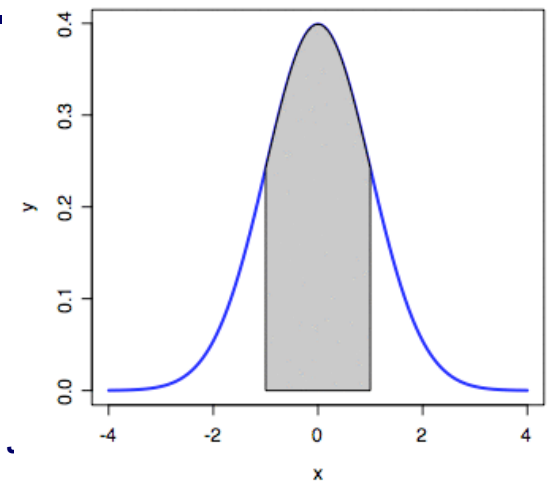
# Exercise 1

---

- **What is the difference between Parametric and Non-Parametric machine learning algorithms?**
- **Define the expected value for a discrete variable and give an example. Calculate the expected for your example and the variance.**

## Exercise 2

- The 68% - 95% - 99.7% is a rule of thumb that allows practitioners of statistics to estimate the probability that a randomly selected number from the standard normal distribution occurs within 1, 2, and 3 standard deviations of the mean at zero.
- What is the probability that a randomly selected number from the standard normal distribution occurs within one standard deviation of the mean? This probability is represented by the area under the standard normal curve between  $x = -1$  and  $x = 1$ , pictured below.
- Plot this graph
- What is the area of the shaded area?



# Exercise 2 Solution

---

## 68%-95%-99.7% Rule

The 68% - 95% - 99.7% is a rule of thumb that allows practitioners of statistics to estimate the probability that a randomly selected number from a normal distribution is within 1, 2, or 3 standard deviations of the mean. Let  $\mu$  be the mean and  $\sigma$  be the standard deviation of the distribution. Then the probability that a randomly selected number from the distribution is within 1 standard deviation of the mean is 68%, within 2 standard deviations is 95%, and within 3 standard deviations is 99.7%.

Fig. 1.1.1: The 68%-95%-99.7% Rule

$x = \mu + \sigma$   
 $y = \mu - \sigma$   
probability  
 $x = \mu + 2\sigma$   
 $y = \mu - 2\sigma$

probability  
[1]



# Exercise 3

**2.1. Height and weight data** The table below and in the data file `htwt.txt` gives  $Ht$  = height in centimeters and  $Wt$  = weight in kilograms for a sample of  $n = 10$  18-year-old girls. The data are taken from a larger study described in Problem 3.1. Interest is in predicting weight from height.

• **ff**

$Ht$	$Wt$
169.6	71.2
166.8	58.2
157.1	56.0
181.1	64.5
158.4	53.0
165.6	52.4
166.7	56.8
156.5	49.2
168.1	55.6
165.3	77.8

- 2.1.1. Draw a scatterplot of  $Wt$  on the vertical axis versus  $Ht$  on the horizontal axis. On the basis of this plot, does a simple linear regression model make sense for these data? Why or why not?
- 2.1.2. Show that  $\bar{x} = 165.52$ ,  $\bar{y} = 59.47$ ,  $SXX = 472.076$ ,  $SYY = 731.961$ , and  $SXY = 274.786$ . Compute estimates of the slope and the intercept for the regression of  $Y$  on  $X$ . Draw the fitted line on your scatterplot.
- 2.1.3. Obtain the estimate of  $\sigma^2$  and find the estimated standard errors of  $\hat{\beta}_0$  and  $\hat{\beta}_1$ . Also find the estimated covariance between  $\hat{\beta}_0$  and  $\hat{\beta}_1$ . Compute the  $t$ -tests for the hypotheses that  $\beta_0 = 0$  and that  $\beta_1 = 0$  and find the appropriate  $p$ -values using two-sided tests.
- 2.1.4. Obtain the analysis of variance table and  $F$ -test for regression. Show numerically that  $F = t^2$ , where  $t$  was computed in Problem 2.1.3 for testing  $\beta_1 = 0$ .

## Exercise 4: LWR

---

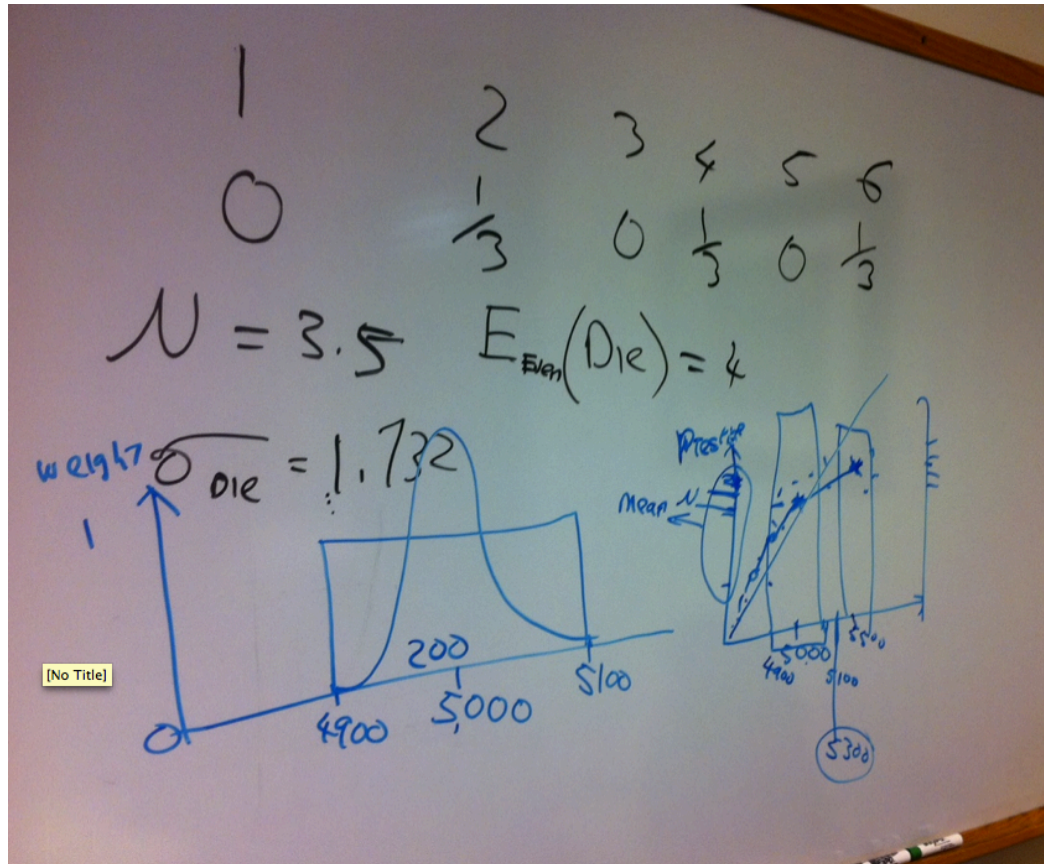
- In R locally weighted (linear) regression is available via `lowess()`; using `data(airquality)`
- Apply `lowess()` to ozone data set (available in R)
- Construct the `lowess` model for following formula “ozone ~ temp” for different  $f$ 's (0.01, 0.1, 0.3, 0.5, 1); comment on your results
- Comment on the computational requirements (memory and CPU) for LOESS models versus a linear regression model when it comes to
  - training a model
  - Classification of a new example (please write out the classification rule for Loess and and for linear regression)

# Exercise 5

---

- **Using the Davis dataset**
  - `library(car); head(Davis) # examine first 6 rows`
- **Tasks**
  - Build a linear regression model `weight ~ repwt`
  - Predict weight from reported weight of men and women engaged in regular exercise
  - What is the dimensionality of this data set
  - Compute the summary stats (using `summary()` command)
  - Comment on these stats
  - Use the summary command on you built linear regression model. Comment on the residuals and the Residual standard error. How is the residual standard error calculated ? (calculate this yourself and show the code)
  - Explain the Residual standard error wrtt problem in layperson's english
  - Plot scattorplots `weight ~ repwt`
    - `scatterplot(weight ~ repwt, span=0.6, lwd=3, id.n=4, data=Davis)`
    - Does everything look okay here? Comment on your findings. Take action and rebuild a new model and compare to the original model. How have things changed?

# Exercise 6: Whiteboard



# Exercise 6

---

- Visualizing the conditional distributions of prestige given values of income using LOESS, or LOWESS (locally weighted scatterplot smoothing)
- **Calculate the  $E(\text{prestige}|\text{income})$** , the mean value of prestige given income the following income values, and the variance ( $\text{prestige}|\text{income}$ ) using a LOWESS (using your own implementation)
  - Assume the neighborhood on the regressor is defined as follows
    - if (value is in the interval  $\text{income} = \pm 100$ )  $\text{weight}_{\text{UNIFORM}} = 1$  else 0
    - Weight  $\text{norm}(\text{value}, \sigma = 50)$  (NOTE standard deviation is 50)
    - Plot the resultant mean functions and calculate the residual standard error. Contrast this to ordinary least squared comment on your findings
  - Scatterplot the data, for Income = 5000 plot the active points for the for  $\text{weight}_{\text{UNIFORM}}$  in red big dots (twice the default size). Plot the mean value prediction using Lowess  $\text{weight}_{\text{UNIFORM}}$ ; Label using a text pointer; plot the 95% confidence interval using Lowess  $\text{weight}_{\text{UNIFORM}}$ ; label using a text pointer
    - Repeat this for incomes = 10,000 (plot utilised points in green stars) and 15,000 (plot utilised points in brown x s)
    - Remember to include a legend in the graph

# Guidelines for Homework

---

- **GENERAL Guidelines for Homework**

- Paste each question into your manuscript and then provide your solution
- Please provide explanations, code, graphs captions, and cross references in a PDF report (that should read like research paper.
- Don't forget to put your name, email and date of submission on each report.
- In addition, please provide R code in separate file. Please comment your so that I or anybody else can understand it and please cross reference code with problem numbers and descriptions
- Please create a separate driver function for each exercise or exercise part (and comment!)
- If you have questions please raise them in class or via email or during office hours

– Homework is due on Tuesday, February 21 of the following week by 5PM.

– Please submit your homework by email to: [James.Shanahan@gmail.com](mailto:James.Shanahan@gmail.com)

with the subject **“Berkeley I 296A”**

– Have fun!

No grades  
If subject  
line is not  
correct

---

- **END**