STA302 METHODS OF DATA ANALYSIS 1

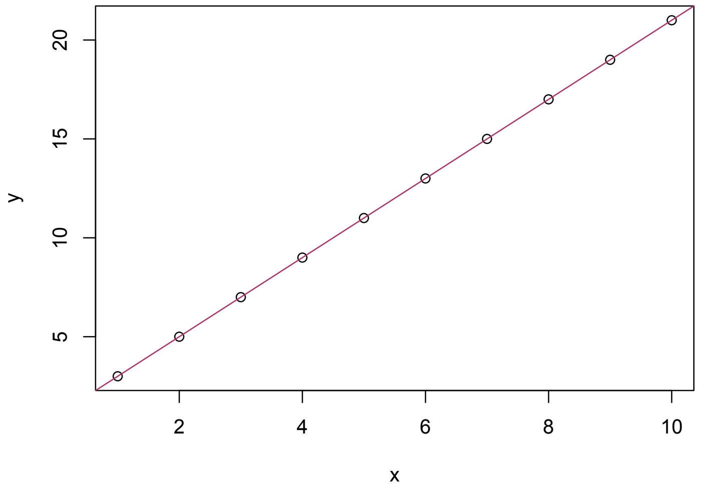
MODULE 1: SIMPLE LINEAR REGRESSION MODELS AND BASICS

PROF. KATHERINE DAIGNAULT

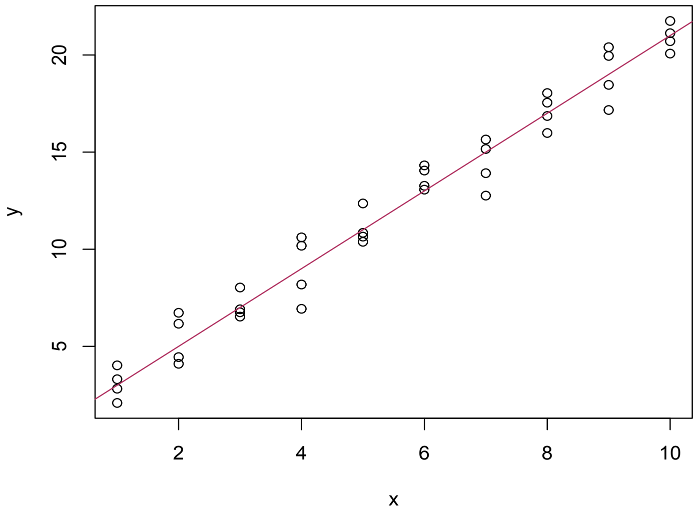
MODULE 1 OUTLINE

1. Estimation Basics and Simple Linear Regression Notation
2. Ordinary Least Squares Estimation Process
3. Interpretation of Simple Linear Regression Estimates
4. Application Example

STATISTICAL RELATIONSHIPS



- Linear relationship: $y_i = ax_i + b$, with slope a and intercept b
- Functional relationship: all y_i are determined by the line



- Statistical relationship: all y follow overall trend but with error/deviations
  - Written as $y_i = ax_i + b + e_i$, where e_i is the difference between each y_i and the trend $ax_i + b$

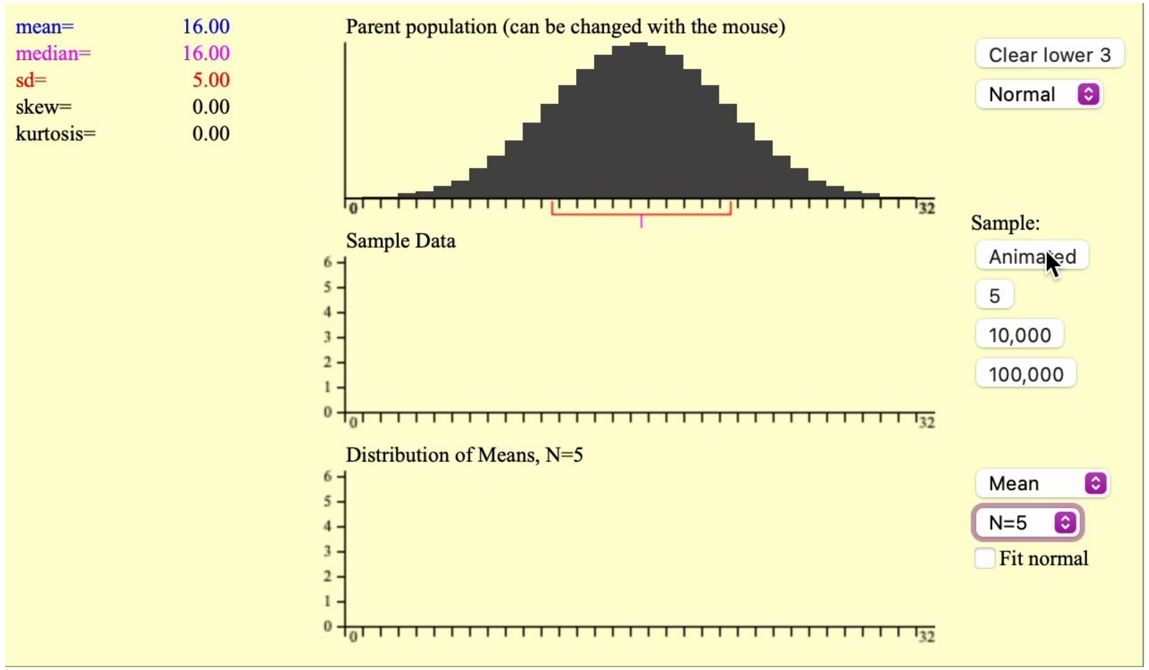
ESTIMATION OF A MEAN

Inference on the Population Mean

- Population, e.g., $Y \sim N(\mu, \sigma^2)$, σ^2 unknown
- Sample data, e.g., $y_1, \dots, y_n$
- Estimate μ with $\bar{y} = {\sum_{i=1}^{n} y_i}/_{n}$, and σ^2 with sample variance $s^2 = {\sum_{i=1}^{n} (y_i - \bar{y})^2}/_{n-1}$
- Variability due to different possible samples is represented with a sampling distribution

$$\frac{\bar{y} - \mu}{s/\sqrt{n}} \sim T_{n-1}$$

https://onlinestatbook.com/stat_sim/sampling_dist/



mean= 16.00
median= 16.00
sd= 5.00
skew= 0.00
kurtosis= 0.00

Sample: Animated

Mean

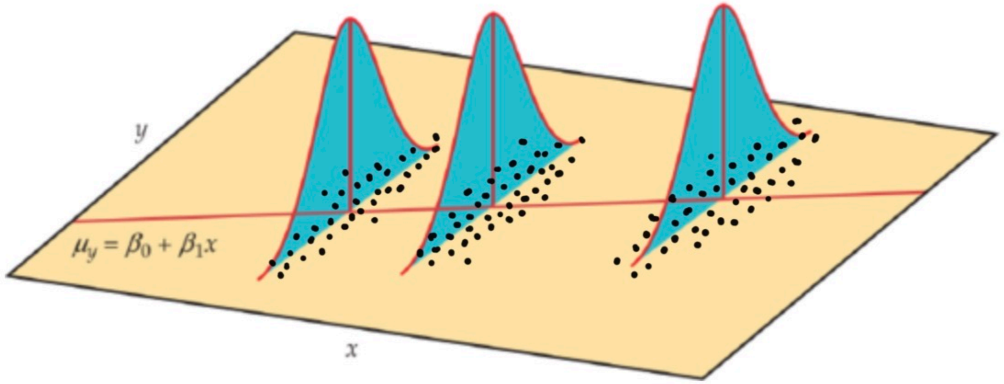
N=5

Parent population (can be changed with the mouse)
Clear lower 3
Normal
0
32
Sample Data
6
5
4
3
2
1
0
5
10,000
100,000
Distribution of Means, N=5
□ Fit normal

ESTIMATION OF A TREND

Inference on a Linear Trend

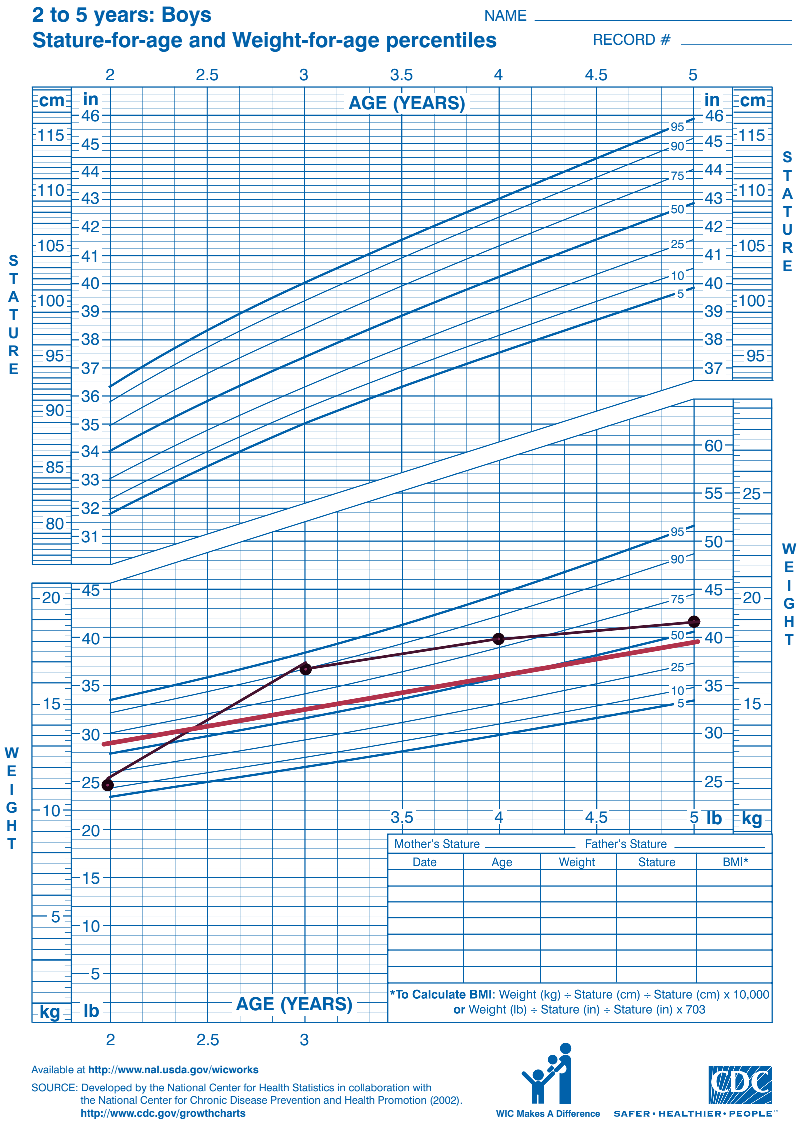
- Population trend $Y = \beta_0 + \beta_1 X + \varepsilon,$ a statistical relationship
  - Y is the *random* response variable
  - X is the *fixed* predictor variable
  - ε is the *random* error, given by $\varepsilon = Y - \beta_0 - \beta_1 X$
- Functional part: $E(Y|X) = \beta_0 + \beta_1 X$
- Sample data: pairs $(x_1, y_1), \ldots, (x_n, y_n)$
- Need to estimate β_0 and β_1 from the sample to estimate these means/the trend



CC BY-NC-SA 3.0 image by Diane Kiernan in Natural Resources Biometrics

WHY CONDITIONAL MEANS?

- Growth charts display distribution of weight or height conditional on age
- A single child may not follow a linear growth progression, e.g.,
  - At age 2, weighs 25lbs
  - At age 3, weighs 37lbs
  - At age 4, weighs 40lbs
  - At age 5, weighs 42lbs
- But for an **<u>average</u>** child, as their age increases, their weight increases relatively constantly



Available at http://www.nal.usda.gov/wicworks
SOURCE: Developed by the National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion (2002).
http://www.cdc.gov/growthcharts

WIC Makes A Difference
CDC
SAFER • HEALTHIER • PEOPLE™

ERRORS AND VARIATION

- Population errors (unknown): $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$
- Total error in the population trend is

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)\ ]^2 = \sum_{i=1}^{n} (Y_i - E(Y_i\ |X_i))^2$$

  - Like sample standard deviation numerator $\sum_{i=1}^{n} (y_i - \bar{y})^2$
- Estimated trend: $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{e}_i$
  - Sample errors are residuals: $\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{y}_i$
  - Fitted values are the estimated means: $\hat{y}_i$
- Measure total error around estimated trend using Residual Sum of Squares

$$RSS = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

From Sheather's A Modern Approach to Regression with R

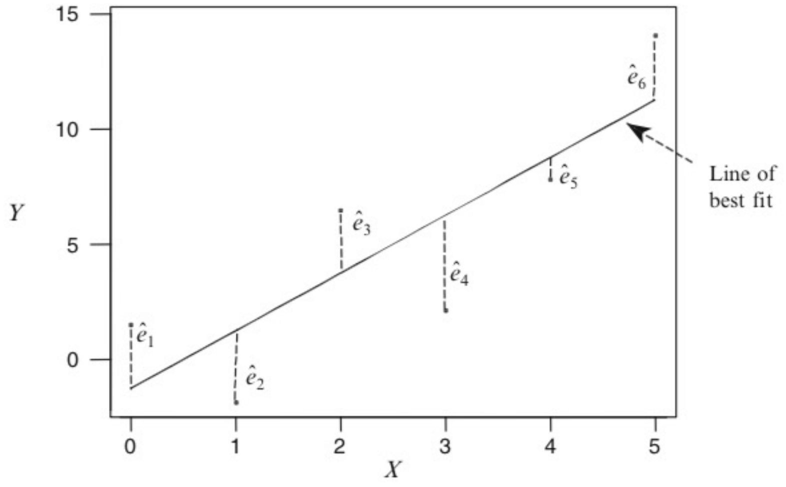


Figure 2.2 A scatter plot of data with a line of best fit and the residuals identified

MODULE 1 OUTLINE

1. Estimation Basics and Simple Linear Regression Notation
2. Ordinary Least Squares Estimation Process
3. Interpretation of Simple Linear Regression Estimates
4. Application Example

MINIMIZING RESIDUAL SUM OF SQUARES

- “Line of best fit” should fit snuggly among data points
  - Snugly = least amount of distance or error
- Instead of minimizing individual errors, minimize the Residual Sum of Squares, i.e. total variation around the line in the data,

$$RSS = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i \ - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

  - This RSS is called our estimating equation – used for estimating the unknown population trend
- We use this equation to find estimates for β_0 and β_1 that make the RSS as small as possible
  - Same as finding the line that makes all residuals as small as possible

WHY DO WE USE A SQUARE?

- Prevents cancellation of positive (above the line) with negative (below the line) residuals
- Easier to work with algebraically (e.g. taking derivatives)
- Penalizes distant points more than closer points
  - Reinforces we want a snug line
- Geometry of the linear algebra:
  - All possible lines/models in model space along with predictor information
  - Response information in different space
  - Model/line that is shortest distance from all points = orthogonal projection from y to the model space

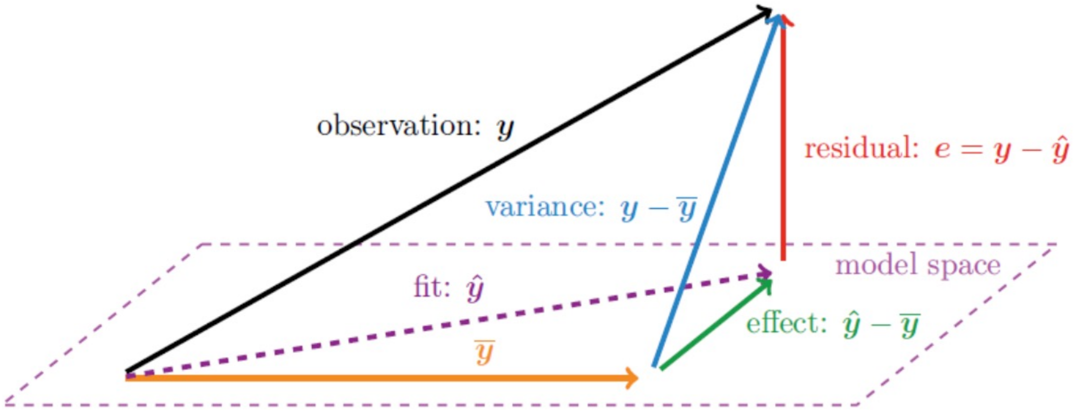

observation: y
residual: $e = y - \hat{y}$
variance: $y - \overline{y}$
model space
fit: $\hat{y}$
effect: $\hat{y} - \overline{y}$
$\overline{y}$

Figure 6.2. Generic linear model.

From Foundations and Application of Statistics by Pruim (2011)

ORDINARY LEAST SQUARES STEPS

Least Squares Procedure

1) Set up the estimating equation for given model with parameters present

2) Take partial derivatives of your estimating equation with respect to each unknown parameter

3) Set each derivative to 0 to obtain score equation

4) Rearrange equations to solve for each unknown parameter.

$$RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\frac{\partial RSS}{\partial \beta_0} = \cancel{-}2\sum_{i=1}^{n}(y_i - \boxed{\hat{\beta}_0 - \hat{\beta}_1 x_i)} = 0$$

$$\Longrightarrow \sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

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

$$\frac{\partial RSS}{\partial \beta_1} = \cancel{-}2\sum_{i=1}^{n} x_i(y_i - \boxed{\hat{\beta}_0 - \hat{\beta}_1 x_i) =} 0$$

$$\Longrightarrow \sum_{i=1}^{n} x_i y_i = \boxed{\hat{\beta}_0} \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

$$\Longrightarrow \sum_{i=1}^{n} x_i y_i = (\bar{y} - \boxed{\hat{\beta}_1 \bar{x})} n\bar{x} + \boxed{\hat{\beta}_1 \sum_{i=1}^{n} x_i^2}$$

$$\Longrightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

LEAST SQUARES ESTIMATES OF COEFFICIENTS

Estimate of Simple Linear Regression Slope

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Choice depends on available information
- Denominator represents total variation in predictor (SXX)
  - Equivalent to numerator in the sample variance of X
- Numerator used in determining sample covariance (SXY)
  - Deviations away from mean of x and mean of y

Estimate of Simple Linear Regression Intercept

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

- Requires estimate of slope to compute
- ”no relationship” is a horizontal line at $\hat{\beta}_0 = \bar{y}$

Important Notes

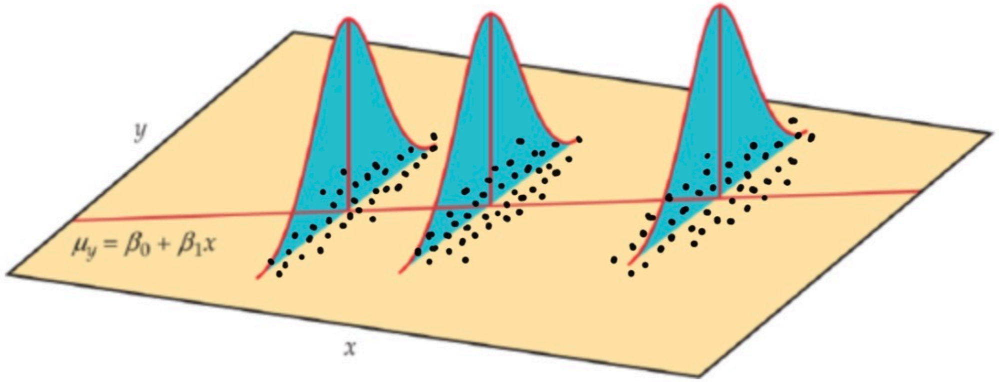
- Formulae ONLY used for simple linear models (1 predictor)
- If a predictor of a different form is used (e.g. x^2), need appropriate values or to rederive the formulae.

MODULE 1 OUTLINE

1. Estimation Basics and Simple Linear Regression Notation
2. Ordinary Least Squares Estimation Process
3. Interpretation of Simple Linear Regression Estimates
4. Application Example

INTERPRETATION OF COEFFICIENTS

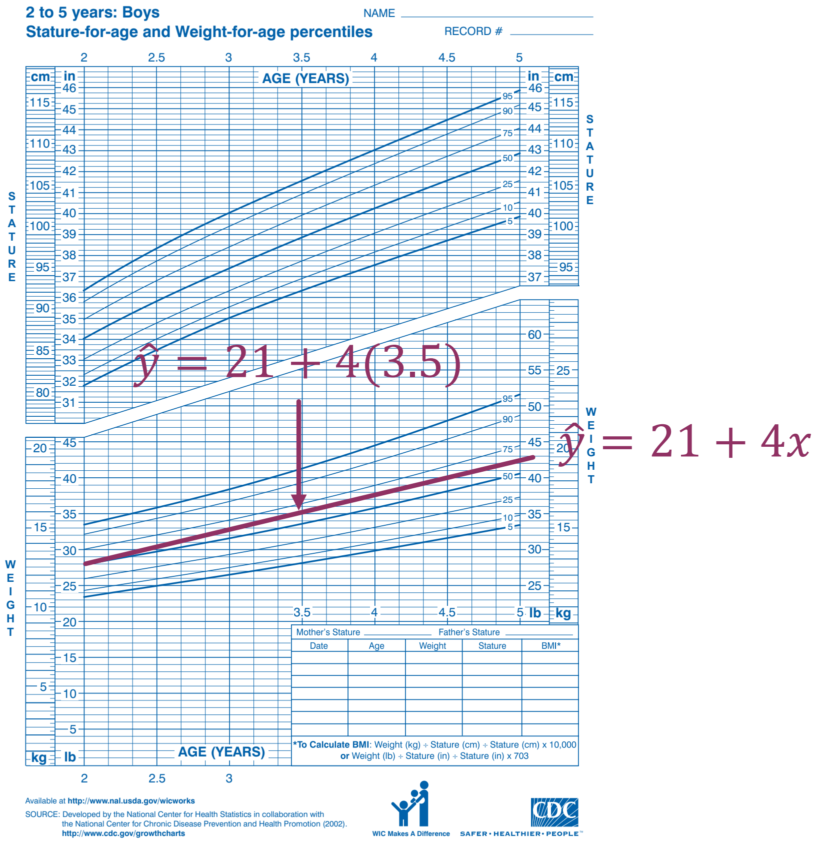
- Key to interpretation: $E(Y|X) = \beta_0 + \beta_1 X$, i.e. we are estimating means
- Intercept: $\hat{\beta}_0$ is the mean/average response when the predictor is zero.
  - Should always consider whether this is meaningful/realistic
- Slope: $\hat{\beta}_1$ is the change in the mean/average/expected response for a one-unit increase in the value of the predictor.
  - NOT the same as all responses change in value by $\hat{\beta}_1$ when predictor increases by one-unit



CC BY-NC-SA 3.0 image by Diane Kiernan in Natural Resources Biometrics

USING THE ESTIMATED SIMPLE LINEAR RELATIONSHIP

- Estimated simple linear relationship between age (X, years) and weight (Y, lbs) is $\widehat{E(Y|X)} = 21 + 4x$
  - 21 pounds is the average weight for a newborn boy – meaningful?
  - 4 pounds is the expected change in weight for a boy whose age increases by 1 year
- Use estimated trend to predict the average weight for a boy of a specific age
  - $\hat{y} = \widehat{E(Y|X = 3.5)} = 21 + 4(3.5) = 35$
  - We can call $\hat{y}$ a predicted/fitted value
  - In this case, it is the estimated value of our population conditional mean at X = 3.5



2 to 5 years: Boys
Stature-for-age and Weight-for-age percentiles

*To Calculate BMI: Weight (kg) ÷ Stature (cm) ÷ Stature (cm) x 10,000
or Weight (lb) ÷ Stature (in) ÷ Stature (in) x 703

Available at http://www.nal.usda.gov/wicworks

SOURCE: Developed by the National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion (2002).
http://www.cdc.gov/growthcharts

WHAT MAKES A RELATIONSHIP LINEAR?

- “Linear” in linear regression refers to the coefficients/parameters, not the predictor.
  - The relationship between Y and X may not appear linear
  - But $y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$ IS linear as the mean change in Y is constant for unit increases in X^2
- Other linear relationships:

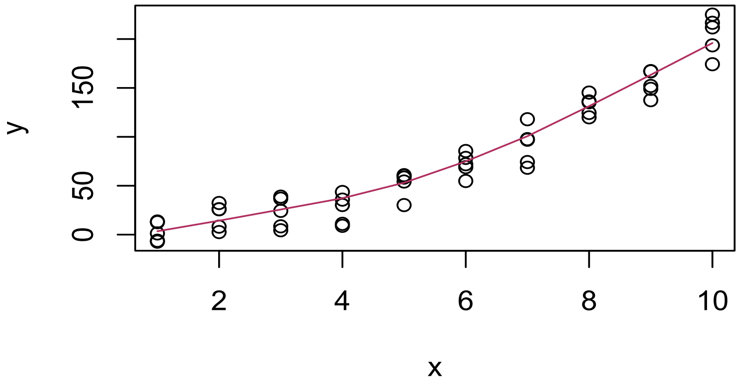
$$y_i = \beta_0 + \beta_1 \sin x_i + \varepsilon_i \quad \text{or} \quad y_i = \beta_0 + \beta_1 \mathbb{I}(answer = yes) + \varepsilon_i$$

- Non-linear relationships:

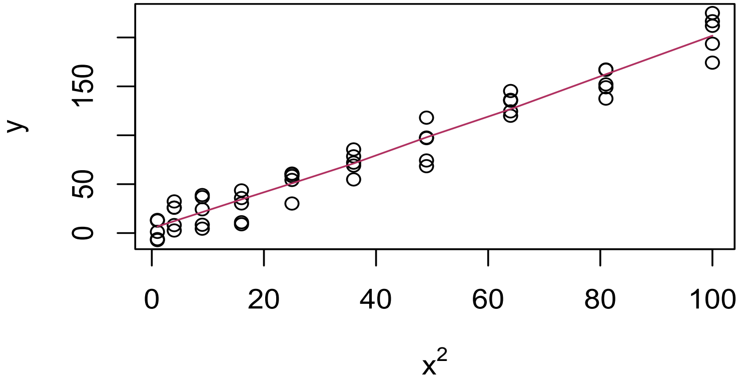
$$y_i = \log(\beta_0 + \beta_1 x_i + \varepsilon_i) \;\; \text{or} \;\; y_i = \beta_0 e^{\beta_1 x_i} + \varepsilon_i$$

- Any relationship that is a linear combination of the coefficients is a linear relationship

Plot of Y versus X



Plot of Y versus X^2



MODULE 1 OUTLINE

1. Estimation Basics and Simple Linear Regression Notation
2. Ordinary Least Squares Estimation Process
3. Interpretation of Simple Linear Regression Estimates
4. Application Example

WORKED EXAMPLE (BY HAND)

Students in a statistics class claimed that doing the homework had not helped them prepare for the midterm exam. The exam score (Y, out of 100) and the averaged homework score (X, out of 100) for 18 students in the class are collected, with summaries presented below:

$$\sum_{i=1}^{18} x_i y_i = 81195, \qquad \sum_{i=1}^{18} x_i^2 = 80199,$$

$$\bar{x} = 58.056, \qquad \bar{y} = 61.389$$

Are the students right?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$= \frac{81195 - 18(58.056)(61.389)}{80199 - 18(58.057)^2} = 0.8726$$

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

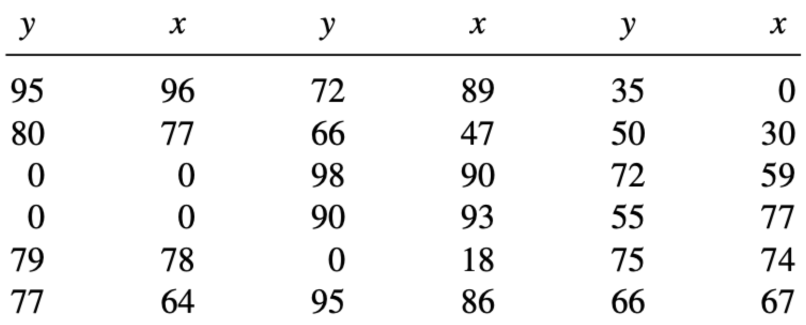
$$= 61.389 - 0.8726(58.056) = 10.73$$

The estimated linear relationship between exam score (Y) and average homework score (X) is

$$\hat{y}_i = 10.73 + 0.8726x_i$$

WORKED EXAMPLE (USING R)

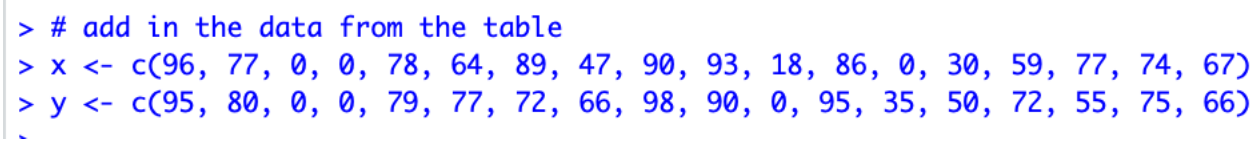
Students in a statistics class claimed that doing the homework had not helped them prepare for the midterm exam. The exam score (Y, out of 100) and the averaged homework score (X, out of 100) for 18 students in the class are displayed below. Are the students right?



y	x	y	x	y	x
95	96	72	89	35	0
80	77	66	47	50	30
0	0	98	90	72	59
0	0	90	93	55	77
79	78	0	18	75	74
77	64	95	86	66	67

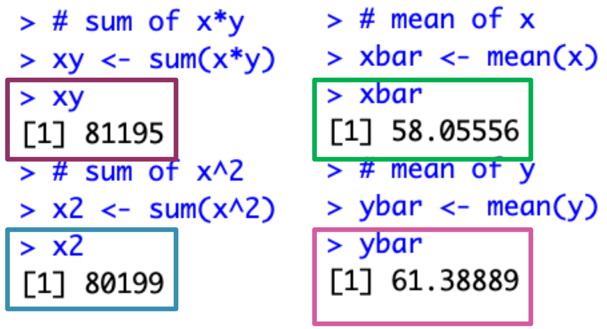
From Linear Models in Statistics by Rencher

Manually load in the data:



```
> # add in the data from the table
> x <- c(96, 77, 0, 0, 78, 64, 89, 47, 90, 93, 18, 86, 0, 30, 59, 77, 74, 67)
> y <- c(95, 80, 0, 0, 79, 77, 72, 66, 98, 90, 0, 95, 35, 50, 72, 55, 75, 66)
```

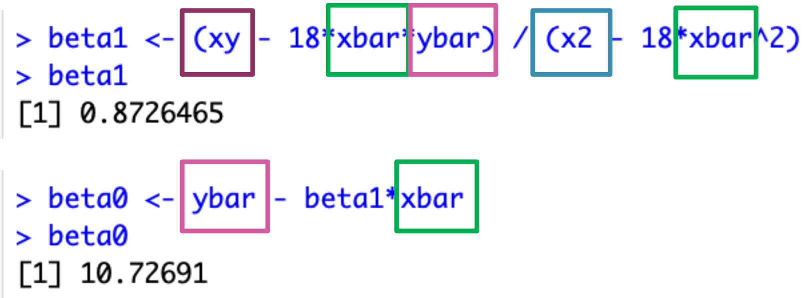
Get summary values:



```
> # sum of x*y
> xy <- sum(x*y)
> xy
[1] 81195
> # sum of x^2
> x2 <- sum(x^2)
> x2
[1] 80199
```

```
> # mean of x
> xbar <- mean(x)
> xbar
[1] 58.05556
> # mean of y
> ybar <- mean(y)
> ybar
[1] 61.38889
```

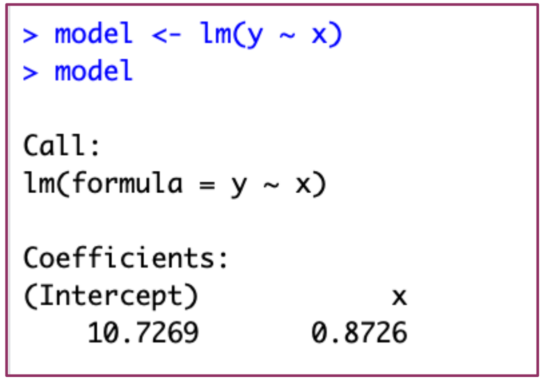
Use formulae as before:



```
> beta1 <- (xy - 18*xbar*ybar) / (x2 - 18*xbar^2)
> beta1
[1] 0.8726465

> beta0 <- ybar - beta1*xbar
> beta0
[1] 10.72691
```

Or, more simply…



```
> model <- lm(y ~ x)
> model

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    10.7269       0.8726
```

ADVICE FOR SIMILAR PROBLEMS

- Consider the values provided in the question and choose formulae accordingly
  - E.g. $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ and there are other equivalent formulae too
  - If model is using predictor in a different way (e.g. squared), be sure to use correct values
- Generally, no need to compute the coefficients by hand in R – just use the lm() function
  - But useful to double check hand calculations sometimes
- Always be careful reading the word problem itself
  - Important information can be contained in words instead of summations or formulae (e.g. stating that "the mean score was…").
- If information doesn't at first appear to match terms in formulae, work backwards
  - Look at formula terms and think about if the necessary information is hidden there in a different way
    - Useful as we move forward in the course and learn how the same information is used in multiple places and ways

MODULE TAKE-AWAYS

1. What does a simple linear model represent/estimate?

2. What are the components of a simple linear model? Which are known/unknown and fixed/random?

3. What are the steps involved in the Least Squares estimation process?

4. How do we interpret the values we estimate for the coefficients in the simple linear model?

5. How do we compute estimates for the coefficients by hand and with R?