Stat 201:
Introduction to Statistics

Standard 11: Variable Association – Linear Regression

From *Naked Statistics: Regression Analysis*

- "Regression analysis allows us to quantify the relationship between a particular variable and an outcome that we care about while controlling for other factors.

- "Most of the studies that you read about in the newspaper are based on regression analysis."

From *Naked Statistics: Regression Analysis*

- “Personal computing has made the mechanics of regression analysis almost effortless. The problem is that the mechanics of regression analysis are not the hard part; the hard part is determining which variables ought to be considered in the analysis and how that can best be done. “
- “Regression analysis is like one of those fancy power tools. It is relatively easy to use, but hard to use well – and potentially dangerous when used improperly.”

From *Naked Statistics: Regression Analysis*

- "First, our sample has to be representative of the population that we care about. A study of 2,000 young children in Sweden will not tell us much about the best policies for early childhood education in rural Mexico. And second, there will be variation from sample to sample. If we do multiple studies of children and child care, each study will produce slightly different findings, even if the methodologies are all sound and similar."
- "The good news is that if we have a large representative sample and solid methodology, the relationship we observe for our sample data is not likely to deviate wildly from the population."

From *Naked Statistics: Regression Analysis*

- "At its core, regression analysis seeks to find the 'best fit' for a linear relationship between two variables"
- "Regression analysis enables us to go one step further and 'fit a line' that best describes a linear relationship between the two variables"
- "Ordinary least squares gives us the best description of a linear relationship between two variables."

From *Naked Statistics: Regression Analysis*

- "The sign (positive or negative) on the coefficient for an independent variable tells us the direction of its association with the dependent variable (x.)"
- "How big is the observed effect between the independent variable and the dependent variable? Is it of a magnitude that matters?"
- "Is the observed result an aberration based on a quirky sample of data, or does it reflect a meaningful association that is likely to be observed for the population as a whole."

From *Naked Statistics: Regression Analysis*

- "Our basic regression analysis produces one other statistic of note: the R Squared, which is a measure of the total amount of variation explained by the regression equation."

- "The R Squared tells us how much of that variation around the mean is associated with differences in the independent variable alone.

Association of Variables – Two Categorical Variables

- **Response Variable –** this is our dependent variable, the outcome variable on which comparisons are made
- **Explanatory Variable –** this is our independent variable, the groups to be compared with respect to values on the response variable
- **Think "we use the explanatory variable to EXPLAIN what's going on with the response variable."**

Response Variable	this is our dependent variable, the outcome variable on which comparisons are made
Explanatory Variable	this is our independent variable, the groups to be compared with respect to values on the response variable

Let's Review Lines

- A **line** is the shortest distance between two points. It has no curve, no thickness and it extends both ways indefinitely.

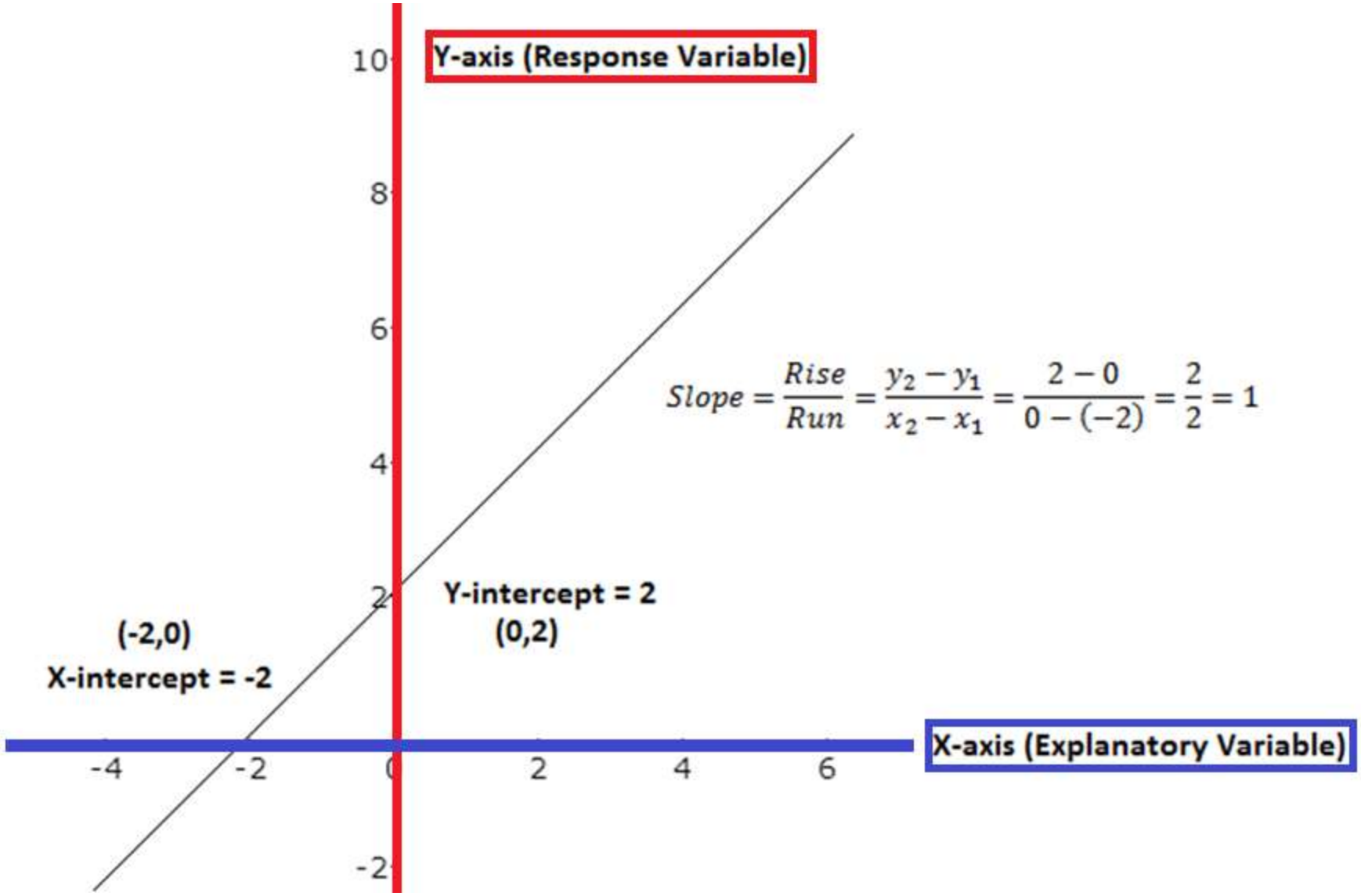
- The equation has the following forms
  - **Slope Intercept:** $y = m * x + b$
  - **Your book uses:** $y = a * x + b$

  - **Point-Slope:** $(y - y_1) = m * (x - x_1)$

Lines

- The **y-axis** runs vertically where x=0

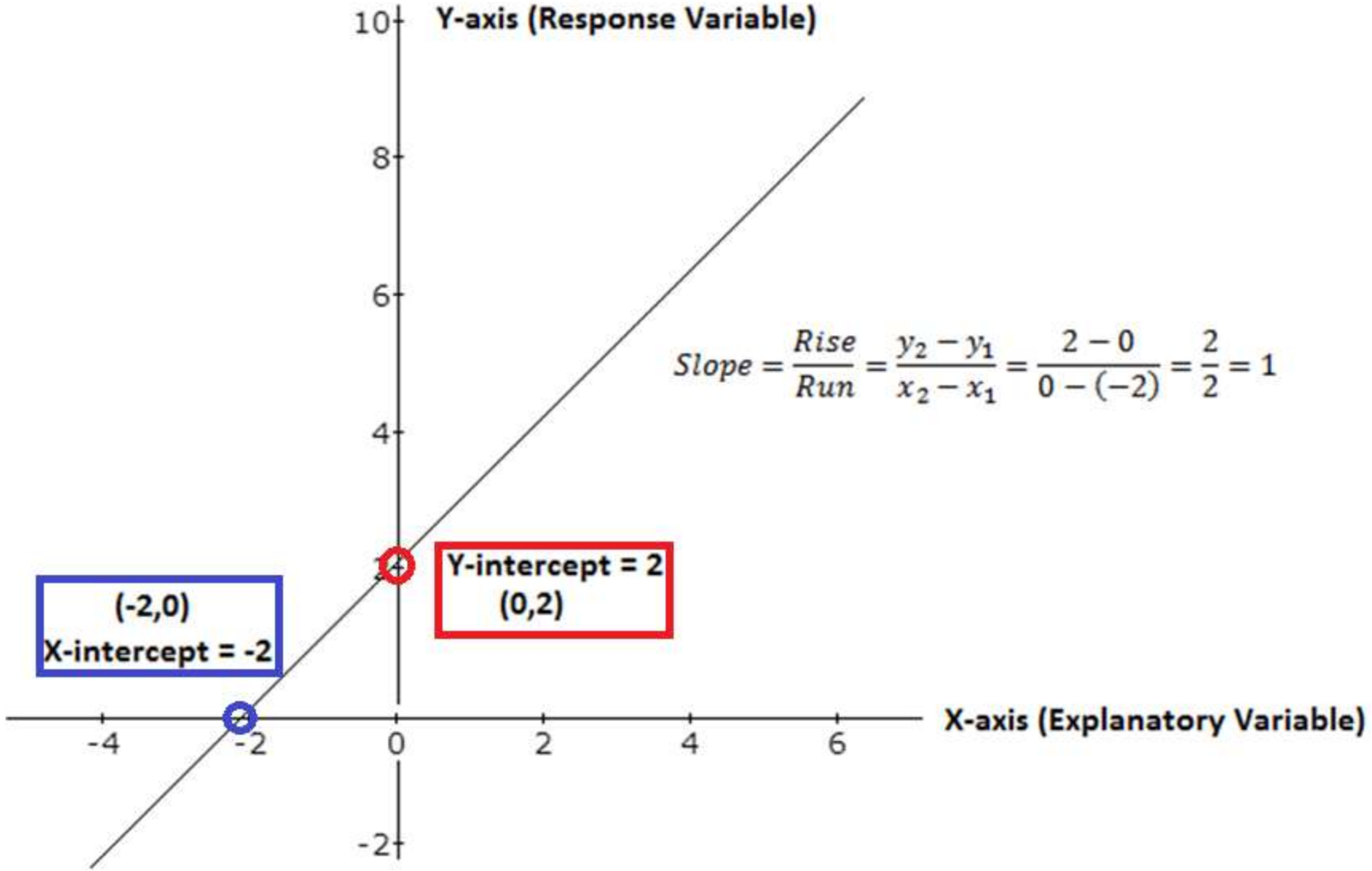
- The **x-axis** runs horizontally where y=0.


Y-axis (Response Variable)
10
8
6
4
2
-2
-4
-2
0
2
4
6
$Slope = \frac{Rise}{Run} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{2-0}{0-(-2)} = \frac{2}{2} = 1$
Y-intercept = 2
(0,2)
(-2,0)
X-intercept = -2
X-axis (Explanatory Variable)

Lines

- The **Y-intercept** is where the line crosses the y-axis and can be found by plugging in x=0 → y=a(0)+b=b. So, b is the Y-intercept.
  - This is important because it is the value the dependent (response) variable takes when the independent (explanatory) variable is zero

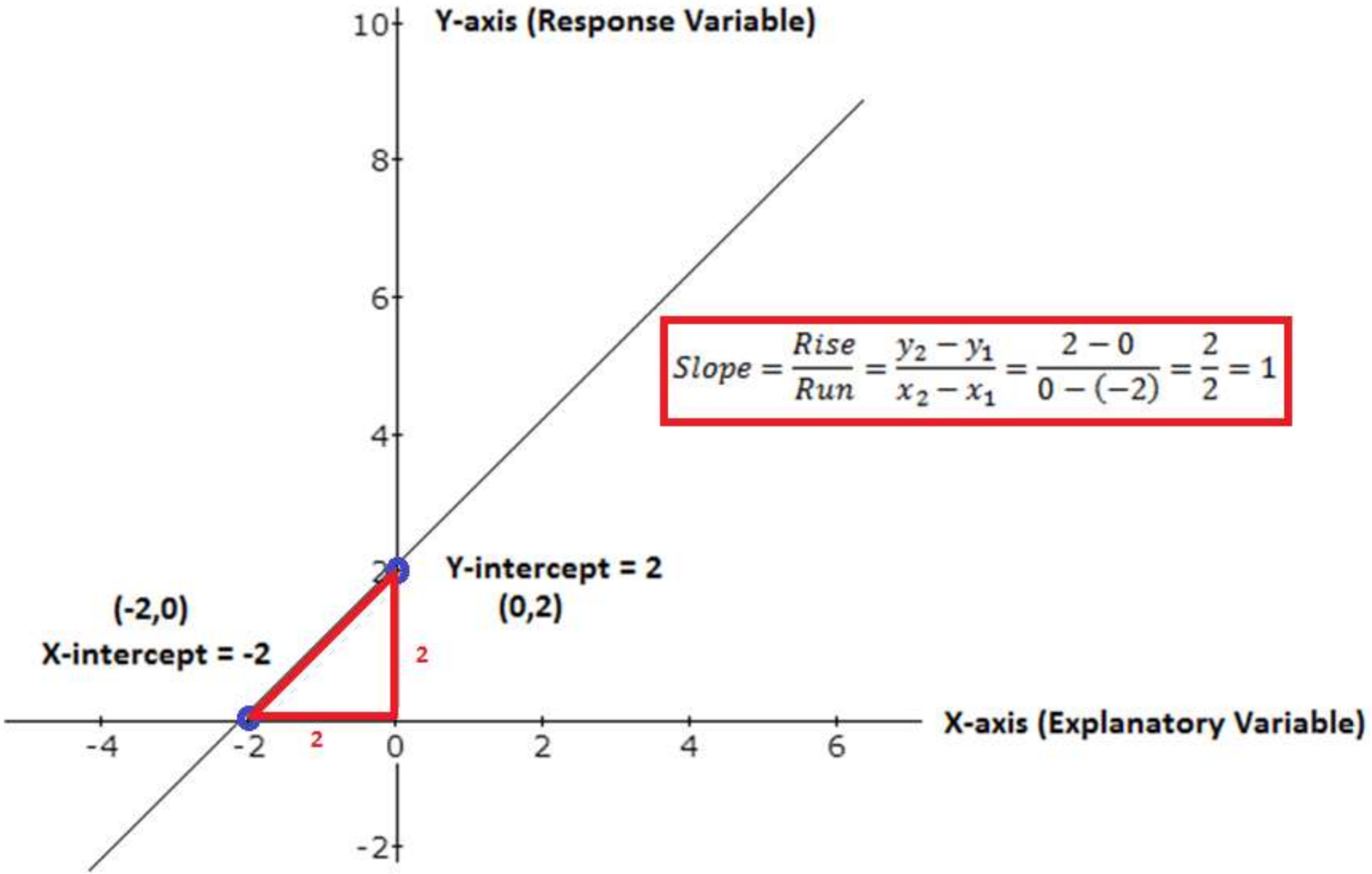
- The **X-intercept** is where the line crosses the x-axis and can be found by plugging in y=0 → 0=ax+b →ax=(-b)→x=(-b)/a.


Y-axis (Response Variable)
10
8
6
4
2
-2
-4
-2
0
2
4
6
X-axis (Explanatory Variable)
$Slope = \frac{Rise}{Run} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{2-0}{0-(-2)} = \frac{2}{2} = 1$
Y-intercept = 2
(0,2)
(-2,0)
X-intercept = -2

Lines

- The **slope (a)** is a measurement of how the line changes; it is the number that multiplies x. It can be thought of as **the change in y for every unit change in x**,
  - ie. the change in y for every increase of one in x.

- It can be calculated using any two points on the line (x_1, y_1) and (x_2, y_2) as below, but it is given by the "a" term in the equation for the line.
  - $Slope = a = \frac{Rise}{Run} = \frac{y_2 - y_1}{x_2 - x_1}$


Y-axis (Response Variable)
10
8
6
4
2
-2
-4
-2
0
2
4
6
X-axis (Explanatory Variable)
$Slope = \frac{Rise}{Run} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{2-0}{0-(-2)} = \frac{2}{2} = 1$
Y-intercept = 2
(0,2)
(-2,0)
X-intercept = -2
2
2

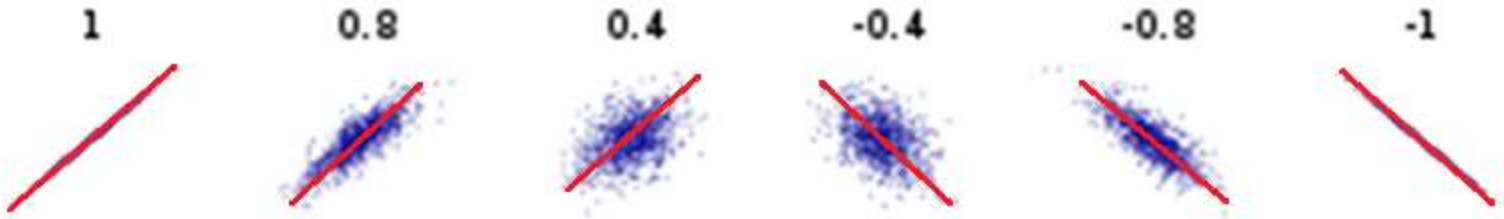
Regression – Making Lines Useful!

- Unlike the lines we learned in math, our data won't fit the line exactly
  - Math: deterministic model
  - Stats: probabilistic model

- **Regression Line –** predicts the value for the response variable y as a straight line function of the value of x, the explanatory variable, with some random error

Regression – Making Lines Useful!

- **Regression Line –** we make our regression line so that it best fits our data – unlike math it usually isn't a perfect



- $\hat{y} = a * x + b$

Regression – Making Lines Useful!

- $\boldsymbol{\hat{y} = a * x + b}$
  - $\boldsymbol{b}$ is the intercept – when x=0
    - This is important because it is the expected value of y when x=0
  - a is the slope of the line
    - This is important because it is the amount that $\hat{y}$ changes when x increases by one unit
  - $\boldsymbol{\hat{y}}$ is the predicted value for some x
  - **Residual** = (the real y) – $\hat{y}$

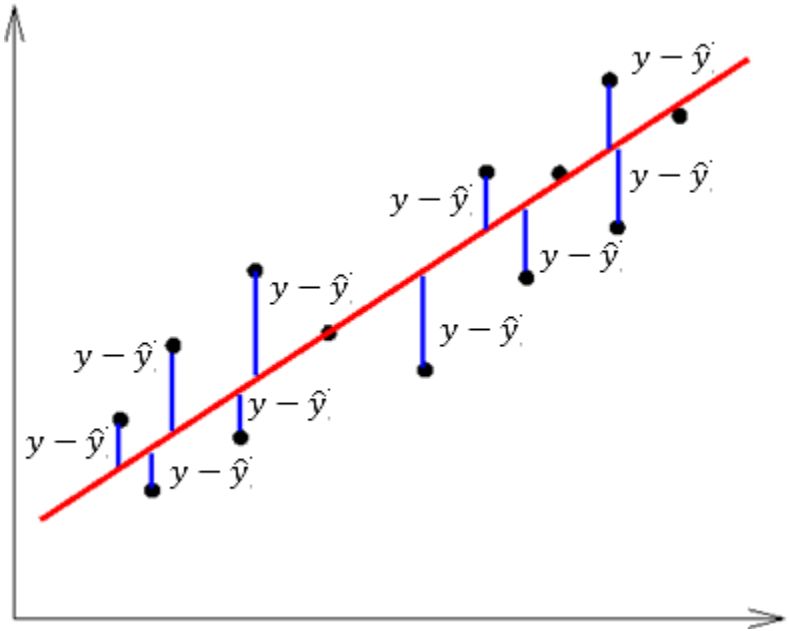
Regression – A Way to Find It

- **Least Squares** is the most popular method
  - It returns the line that has the smallest value for the residual sum of squares in using:

- $\hat{y} = a * x + b$
- Residual Sum of Squares = $\sum(y - \hat{y})^2$

Regression – A Way to Find It

- We don't just draw the 'best-fit line' like we might have before this class
- Least squares gives us the solution where the total length of blue lines the smallest



Regression – Least Squares

- We find a and b such that we minimize the sum of squared errors. These estimates are called the ordinary least squares estimators – we leave this up to software.

- In simple regression
  - $\hat{\boldsymbol{y}} = a + b * x$
  - $a = y\ intercept = \hat{y} - b * x$
  - $b = slope = r\ * \left(\frac{s_y}{s_x}\right)$

When Can We Use Regression?

- $\boldsymbol{R^2}$, given in the regression output, gives the percent of variation in y explained by x
- $\boldsymbol{R^2}$, given in the regression output, gives the percent of variation in the response variable, y, explained by the explanatory variable, x.

- **Note:** $R^2 = r^2$
- **Note:** $r = \sqrt{R^2}$

When Can We Use Regression?

- The **scatterplot** must show a fairly linear relationship
  - A rule of thumb is to look for a coefficient of correlation, **r > .7** or **r < -.7**
  - **Equivalently,** a rule of thumb is to look for a coefficient of correlation, R^2 **> .49**

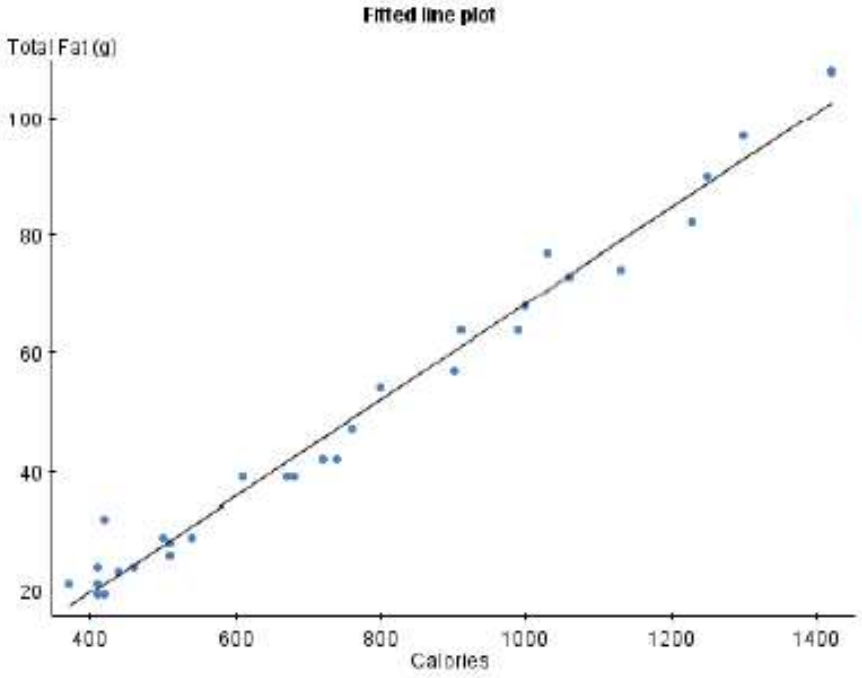
Comparing Two Quantitative Variables

Scatterplot	let the **response** variable be the y variable and the **explanatory** variable be the x variable and plot the points
Regression Line	predicts the value for the response variable y as a straight line function of the value of x, the explanatory variable $\hat{y} = a * x + b$
Intercept	a this is the expected value of y when x is zero
Slope	b this is the amount that $\hat{y}$ changes by when x increases by one unit

Comparing Two Quantitative Variables

$\boldsymbol{R^2 = r^2}$	gives the percent of variation in y explained by x
$\boldsymbol{r = \sqrt{R^2}}$	measures the **LINEAR** relationship between x and y
Estimate $\hat{y}$ for a given x	Plug x into the regression equation $\boldsymbol{\hat{y} = b_1 * x + b_0}$
Residual	**Residual = (the real y) – $\hat{y}$**

Regression Example: Fast Food Items



Simple linear regression results:

Dependent Variable: Total Fat (g)

Independent Variable: Calories

Total Fat (g) = -12.907254 + 0.081350215 Calories **Regression Equation**

Sample size: 32

R (correlation coefficient) = 0.9894 **r**

R-sq = 0.9789471 R^2

Estimate of error standard deviation: 3.7394521

Regression Example

- Visual Output
  - The points are fit by the line very well – the distance between the points and the line are very small
- Numerical Output
  - $\widehat{Fat} = .081350215 * (calories) - 12.907254$
  - r = .9894
    - r = .9894 > .7 → We have enough correlation to make a good regression line
  - $R^2 = .9789471$

Regression Example

- $\widehat{\boldsymbol{Fat}} = \mathbf{.081350215} * (\boldsymbol{calories}) - \mathbf{12.907254}$

  - $\boldsymbol{b}$= -12.907254 is the intercept
    - The expected grams of fat of in a fast food item with 0 calories is negative 12.907254
    - Here, interpreting the intercept doesn't make sense
  - $\boldsymbol{a}$= .081350215 is the slope of the line
    - For every additional calorie in a fast food item we expect the grams of fat to increase by .081350215 on average

Regression Example

- $\widehat{\boldsymbol{Fat}} = \mathbf{.081350215} * (\boldsymbol{calories}) - \mathbf{12.907254}$

  - $\boldsymbol{R^2} = .9789471$
    - 97.89% of the variation in fat in fast food items is explained by calories
  - $r = \sqrt{R^2} = \sqrt{.9789} = .9894$
    - Since r is very close to one we have a **very strong** positive correlation

Regression Example

- How many grams of fat **do you expect** a hamburger with 1000 calories to have?
  - Plug in 1000 for calories and see what fat is

$$\widehat{Fat} = .081350215 * (calories) - 12.907254$$
$$= .081350215 * (1000) - 12.907254$$
$$= 68.4 \text{ grams}$$

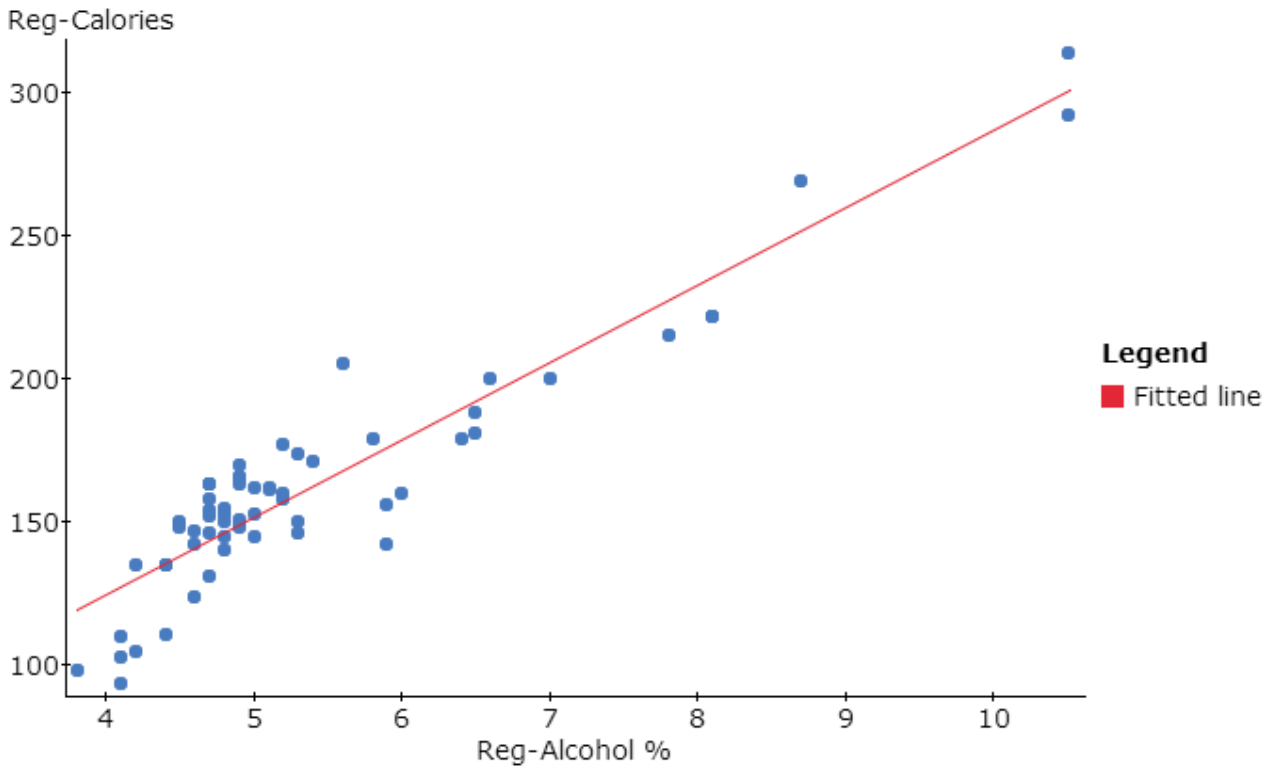
- If the true value is 70, **what is the residual**?

Residual = (the real y) – $\hat{y} = 70 - 68.4 = 1.6$

Regression Example 2

- In creating beer yeast and sugar react to create alcohol – the idea being, the more sugar and yeast you add the more alcohol the batch yields. It would then make sense that the more alcohol in the beer the more carbohydrates there are thus more calories – but who are we to make such assertions? Let us show it statistically.

Regression Example 2



- The points are fit by the line very well – the distance between the points and the line are very small

Regression Example 2

Simple linear regression results:

Dependent Variable: Reg-Calories

Independent Variable: Reg-Alcohol %

Reg-Calories = 16.374148 + 27.003873 Reg-Alcohol % **Regression Equation**

Sample size: 61

R (correlation coefficient) = 0.93198924 r

R-sq = 0.86860395 R^2

Estimate of error standard deviation: 14.762875

Regression Example 2

- The computer will give us output similar to the table on the previous slide. The best fit regression line, below, that the computer found can be interpreted by humans!

- (Calories) = 16.3741+ 27 * (Alcohol %)

Regression Example 2

- (Calories) = 16.3741+ 27.0039 $*$ (Alcohol %)
  - **b =** 16.3741 is the intercept
    - The expected number of calories of a beer with 0% alcohol is 16.3741. Unlike the previous example, interpreting the intercept does make sense here.
  - $\boldsymbol{a}$ =27.0039 is the slope of the line
    - For every additional percent alcohol in beer we expect the number of calories to increase by 27.0039 on average.

Regression Example

- (Calories) = 16.3741+ 27.0039 $*$ (Alcohol %)
  - $\boldsymbol{R^2} = .8686$
    - 86.86% of the variation in calories in beer is explained by alcohol
  - $r = \sqrt{R^2} = \sqrt{.8686} = .9320$
    - Since r is very close to one we have a **very strong** positive correlation

Regression Example 2

- If we wanted to **estimate** the calories of my favorite beer, Rogue Dead Guy, we can plug in its alcohol percentage into the equation to find an estimate of the calories.
- If the Alcohol % = 6.5% for Rogue Dead Guy we can plug it in to find the estimated calories of Rogue Dead Guy.

$$(Calories) = 16.37 + 27 * (Alcohol\%)$$
$$= 16.37 + 27 * (6.5)$$
$$= 191.87$$

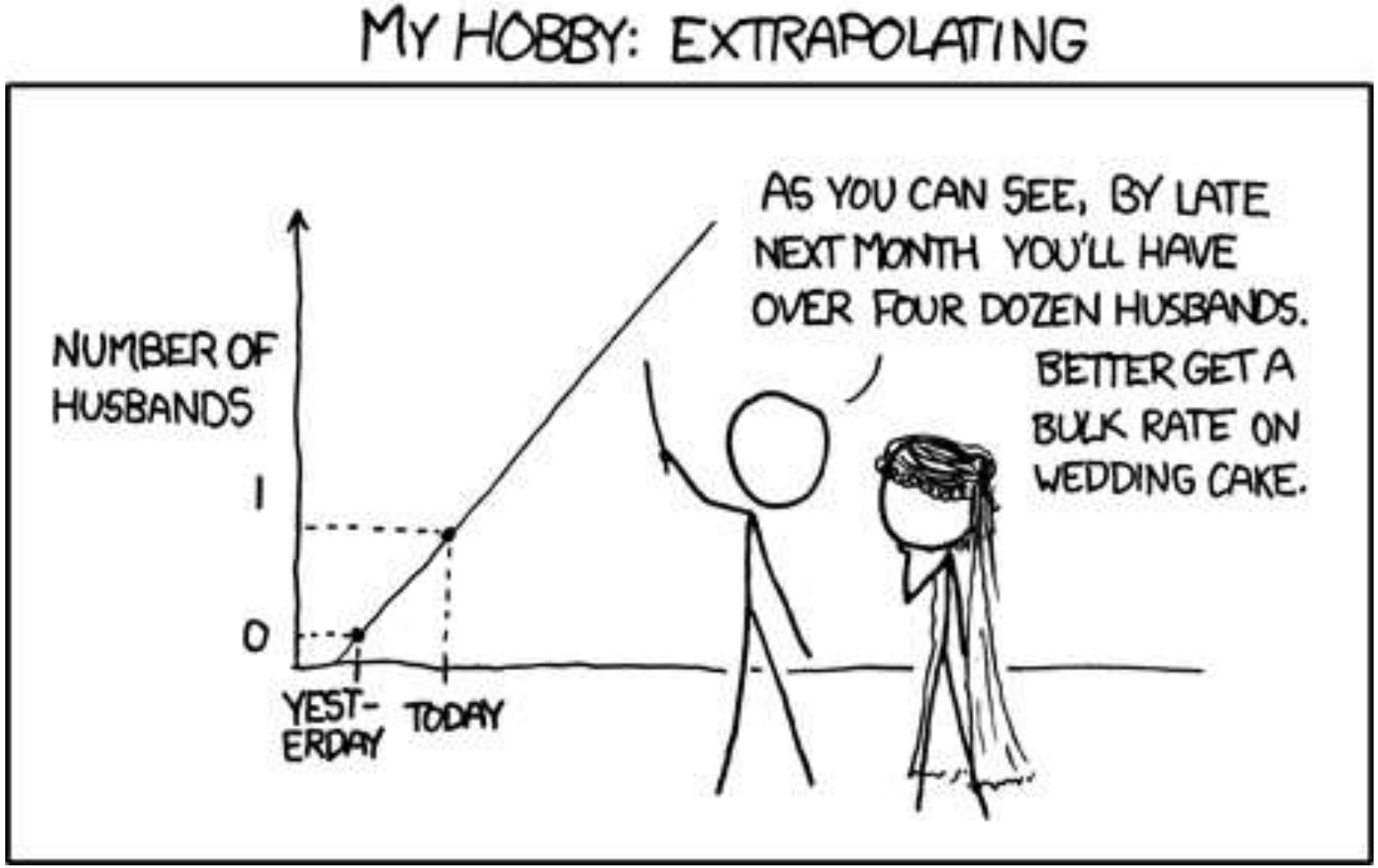
Regression Example 2

- So, the estimated amount of calories for Rogue Dead Guy is 191.87.
- In fact, the actual amount of calories is 250 so our estimate isn't very good, though it might be better than a random guess.
- **The residual** is the difference
  - true – estimate = 250-191.87=58.13.

Regressions – Problems

- **Extrapolation** – We don't want to predict using x values different than the known data
- **Influential Outliers** – a single point can really change the fit of the regression line – always check for stray points in the scatterplot
- **Correlation does not imply causation** – wait for it
- **Lurking Variables** – a variable that we don't look at that causes the correlation (hot summers)

Regressions – Problems



Credit: XKCD

Regression in your TI Calculator

- Example and Steps:
  - https://www.youtube.com/watch?v=nw6GOUtC2jY

Regression in your TI Calculator

- Set up – this only needs to be done once
  1. Press 2^{nd}
  2. Press 0
  3. Scroll down using ↓ to ‘DiagnosticOn’
  4. Hit ENTER
  5. Hit ENTER

Regression in your TI Calculator

- Input-
  1. Press STAT
  2. Press ENTER with ‘Edit’ highlighted
  3. Enter the X data into the L1 column
  4. Enter the Y data into the L2 column
  5. Press STAT
  6. Press → to CALC
  7. Press ENTER with ‘4:LinReg(ax+b)’ highlighted
  8. Press 2^{nd}
  9. Press 1
  10. Press ,
  11. Press 2^{nd}
  12. Press 2
  13. Press ENTER

Regression in your TI Calculator

- Output-
  1. Y = ax+b
     - This is the form of our regression equation
  2. a is our slope of the regression line
  3. b is the intercept of the regression line
  4. $r^2 = R^2$
  5. r is our correlation coefficient

Regression

- **StatCrunch Commands**

Stat→Regression→Simple Linear → Select your explanatory variable as x → Select your response variable as y → Compute!