STAT 201 Chapter 3

Association and Regression

Association of Variables – Two Categorical Variables

- **Response Variable (dependent variable):** the outcome variable whose variation is being studied
- **Explanatory Variable (independent variable):** the groups to be compared with respect to values on the response variable
- Example 1:
  - **Response:** Survival Status; **Explanatory:** Smoking Status
- Example 2:
  - **Response:** Happiness Level; **Explanatory:** Income Level

Definition

- **Association:** An association exists between two variables if a particular value for one variable is more likely to occur with certain values of the other variable

- **Contingency table:** A display for two categorical variables. Its rows list the categories of one variable and its columns list categories of the other variable.

Contingency Table: Example

- Two Variables
  - Would you keep or turn in a $100 if you found it on the library floor?
  - Do you recycle (cans / bottles)?

	Keep It	Turn It In	Total
No Recycle	17	8	25
Recycle	30	34	64
Total	47	42	89

Contingency Table: Example

Counts

	Keep It	Turn It In	Total
No Recycle	17	8	25
Recycle	30	34	64
Total	47	42	89

Percent

	Keep It	Turn It In	Total
No Recycle	17/89	8/89	8/89
Recycle	30/89	34/89	64/89
Total	47/89	42/89	89/89

Conditional Percent

	Keep It	Turn It In	Total
No Recycle	17/25	8/25	25/25
Recycle	30/64	34/64	64/64

Contingency Table: Example

Counts

	Keep It	Turn It In	Total
No Recycle	17	8	25
Recycle	30	34	64
Total	47	42	89

Percent

	Keep It	Turn It In	Total
No Recycle	19.1%	8.989%	8.99%
Recycle	33.71%	38.2%	71.91%
Total	52.81%	47.19%	100%

Conditional Percent

	Keep It	Turn It In	Total
No Recycle	68%	32%	100%
Recycle	46.88%	53.13%	100%

Contingency Table: Example

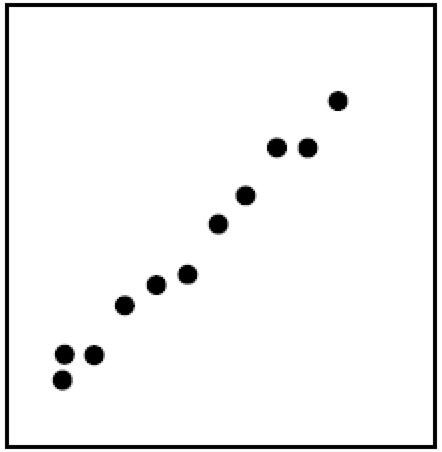
	Keep It	Turn It In	Total
No Recycle	68%	32%	100%
Recycle	46.88%	53.13%	100%
Total	52.81%	47.19%	100%

- With conditional percentage contingency table, does there appear to be an association between recycling and turning in money found on the floor?

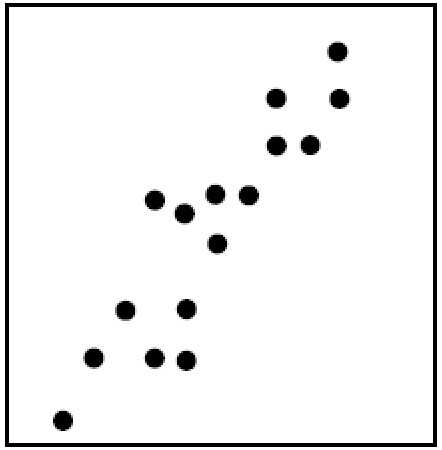
- Yes – it appears that a larger percent of people who do the recycle are willing to turn the $100 in compared to those that keep it

What About Two Quantitative Variables?

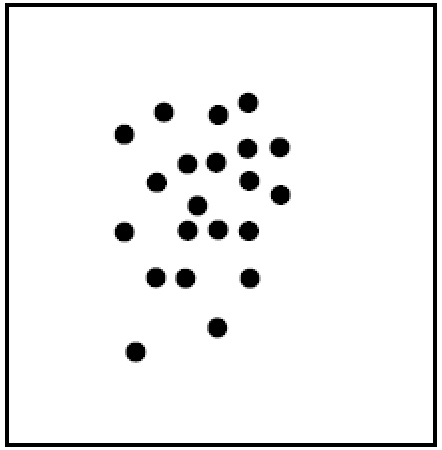
- We use a **scatterplot** to examine the association between the two quantitative variables
- To form a **scatterplot** we let the **response** variable be the Y variable and the **explanatory** variable be the X variable and just plot the points in coordinate system



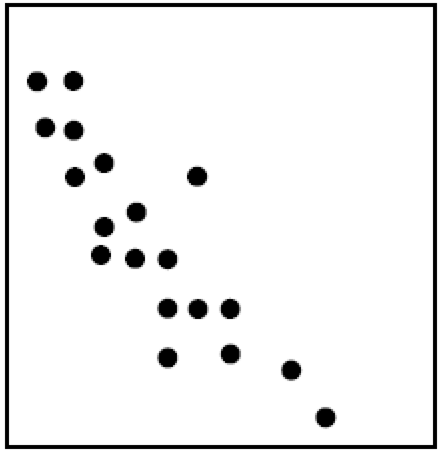
Strong positive correlation



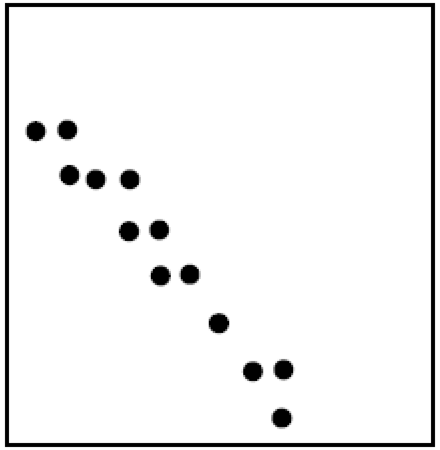
Moderate positive correlation



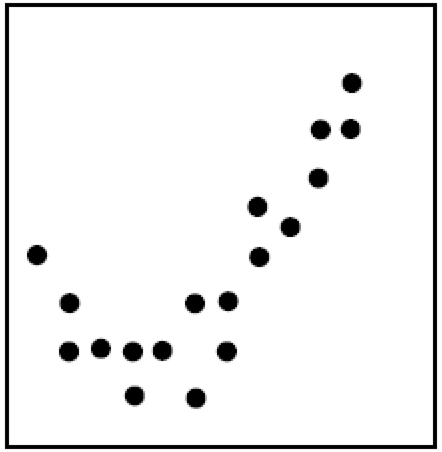
No correlation



Moderate negative correlation



Strong negative correlation



Curvilinear relationship

Coefficient of Correlation

- **Coefficient of correlation,** denoted by letter **r**, measures the **LINEAR** relationship between X and Y [linear, linear, linear!!!]

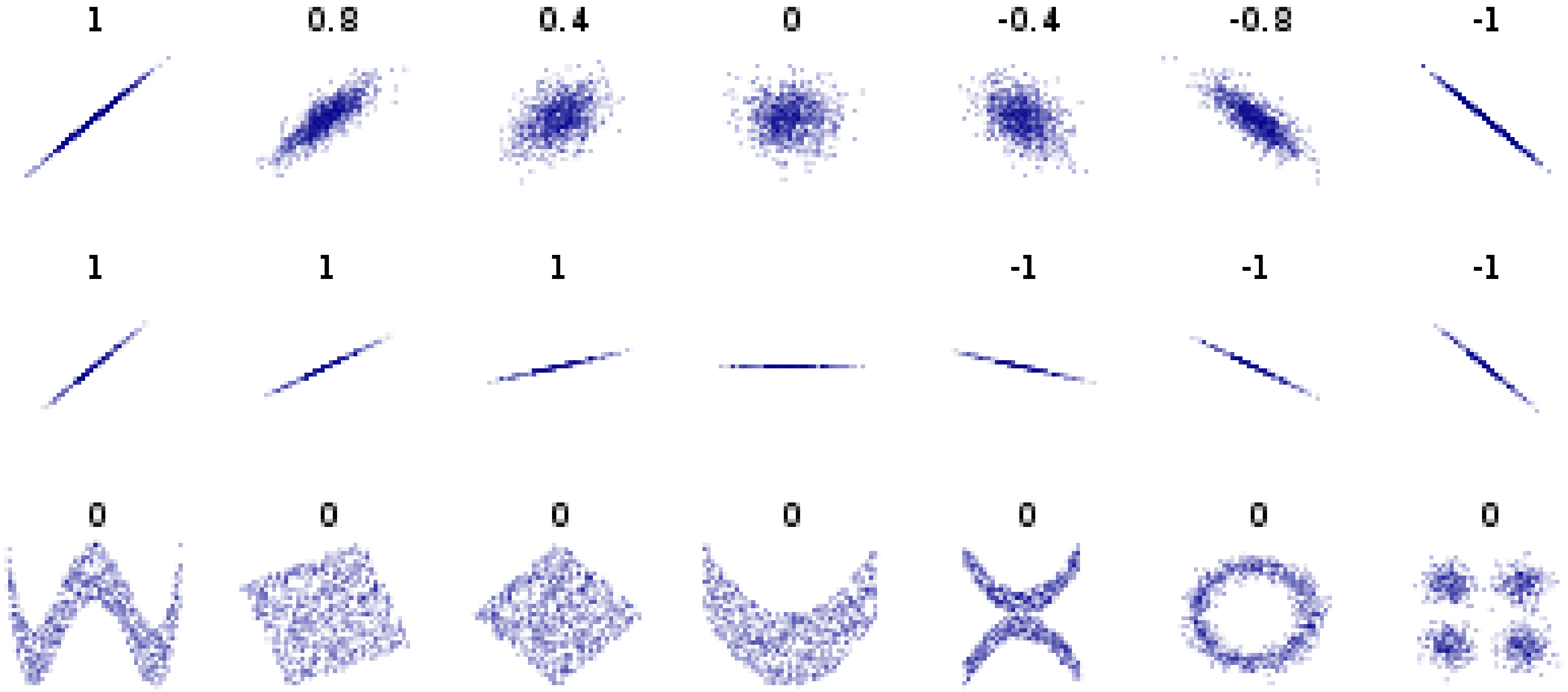
- **r > 0** indicates a positive linear correlation
- **r < 0** indicates a negative linear correlation
- **r=1** indicates a perfect positive linear correlation
- **r=-1** indicates a perfect negative linear correlation
- **r=0** indicates no linear correlation

Properties

- $-1 \leq r \leq 1$
- The closer r is to 1, the stronger the evidence for positive linear correlation
- The closer r is to -1, the stronger the evidence for a negative linear correlation
- The closer r is to 0, the weaker the evidence for linear correlation

- **r** is affected by outliers, so we have to be careful

Coefficient of Correlation: Example



Lines

- A line is the shortest distance between two points. It has no curve, no thickness and it extends to both negative and positive infinitely.

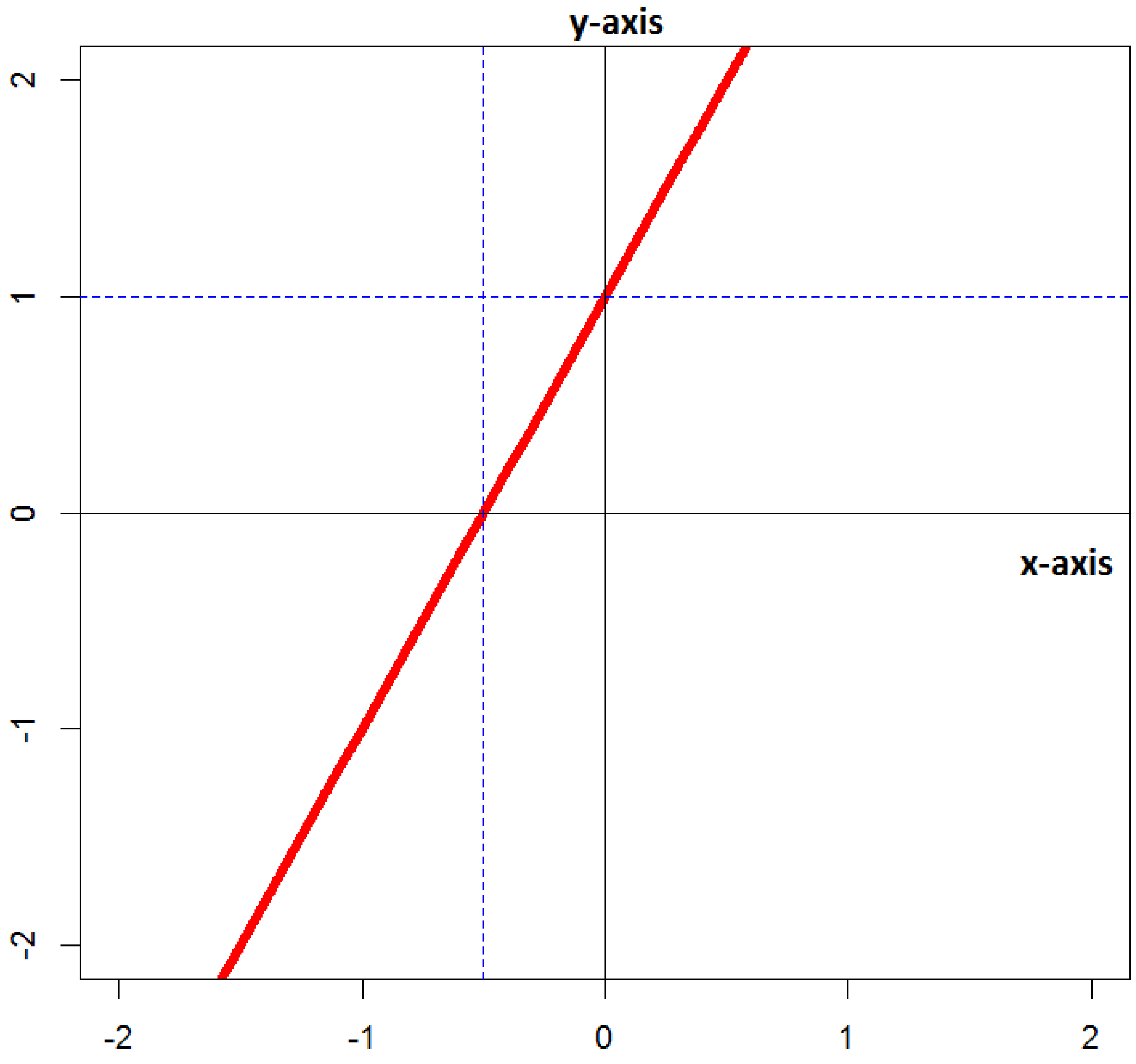
- The equation has the form

$$y = a + bx$$

- a is called intercept. It is the value of y when x is zero.
- b is called slope. When x increases/decreases one unit, y would increase/decrease b units. It measures how the line changes.

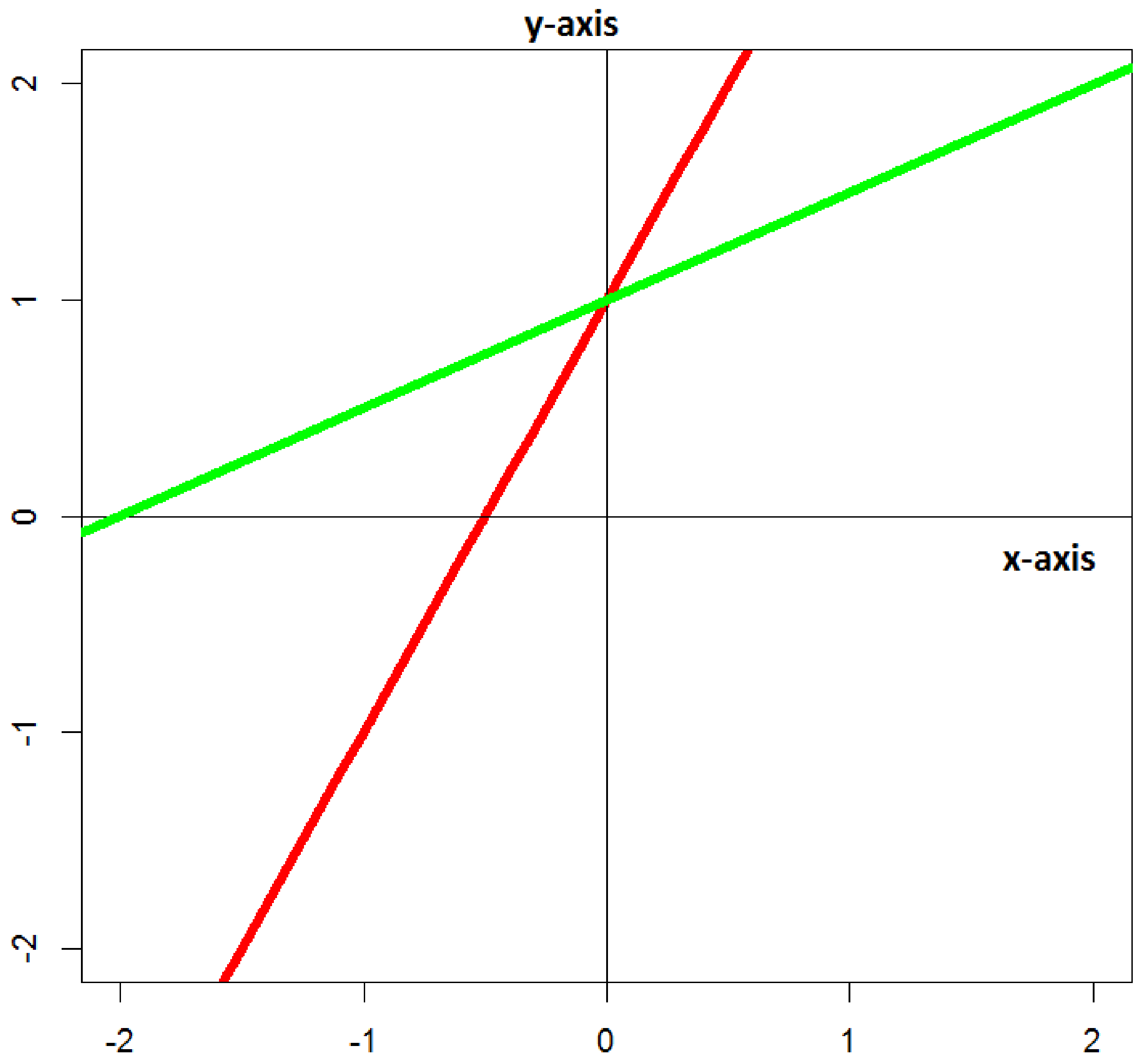
Example

- $y = a + bx$
- $y = 1 + 2x$
- a =1 is the intercept, the value of y when x=0
- $b = 2$ is the slope, when x increases/decreases one unit, the value of y increases/decreases two units.



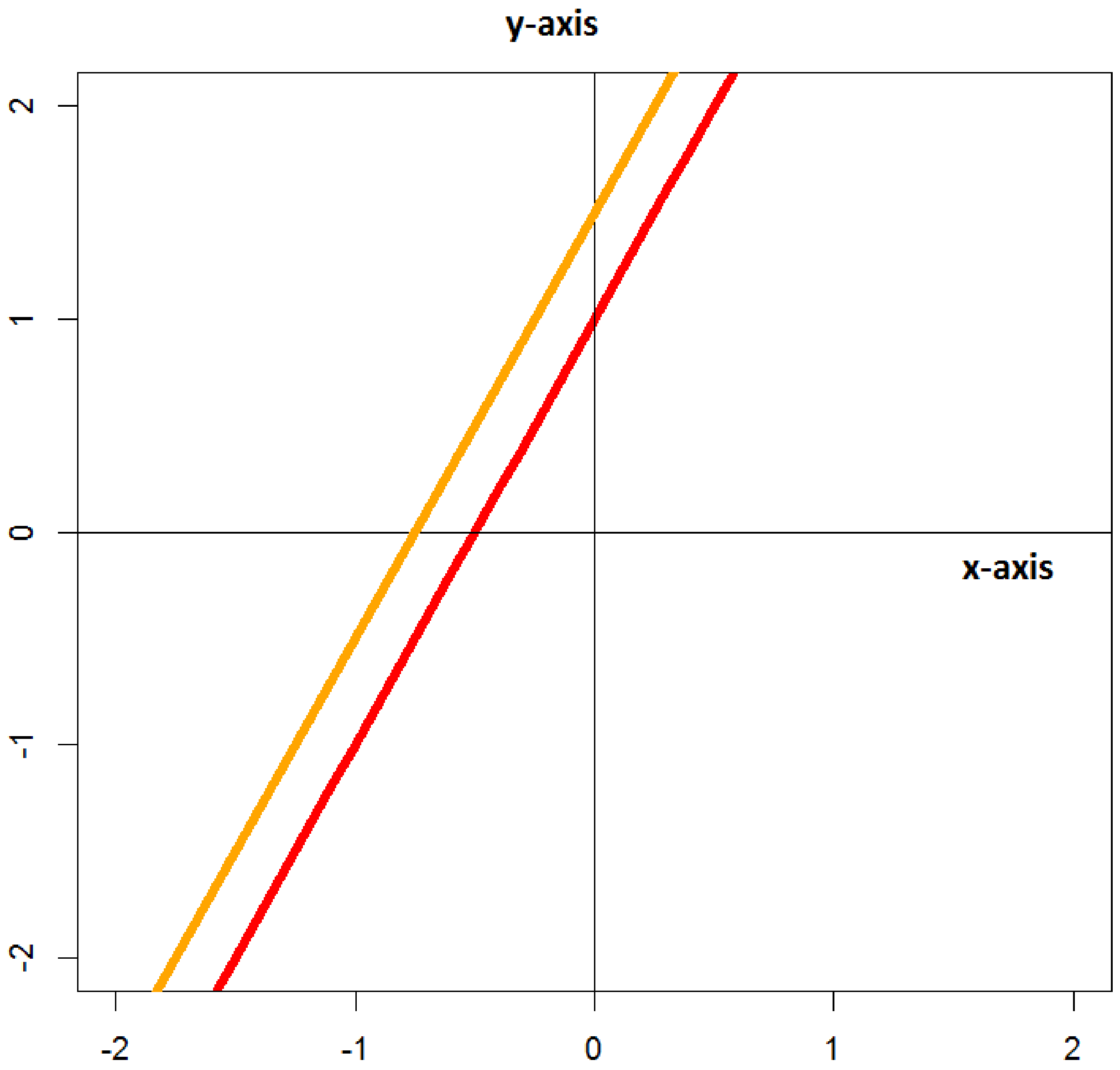
Example

- I add a green line
- $y = 1 + 0.5x$
- Compare to the red line $y = 1 + 2x$, the slope changes from 2 to 0.5.
- Can you see the difference?



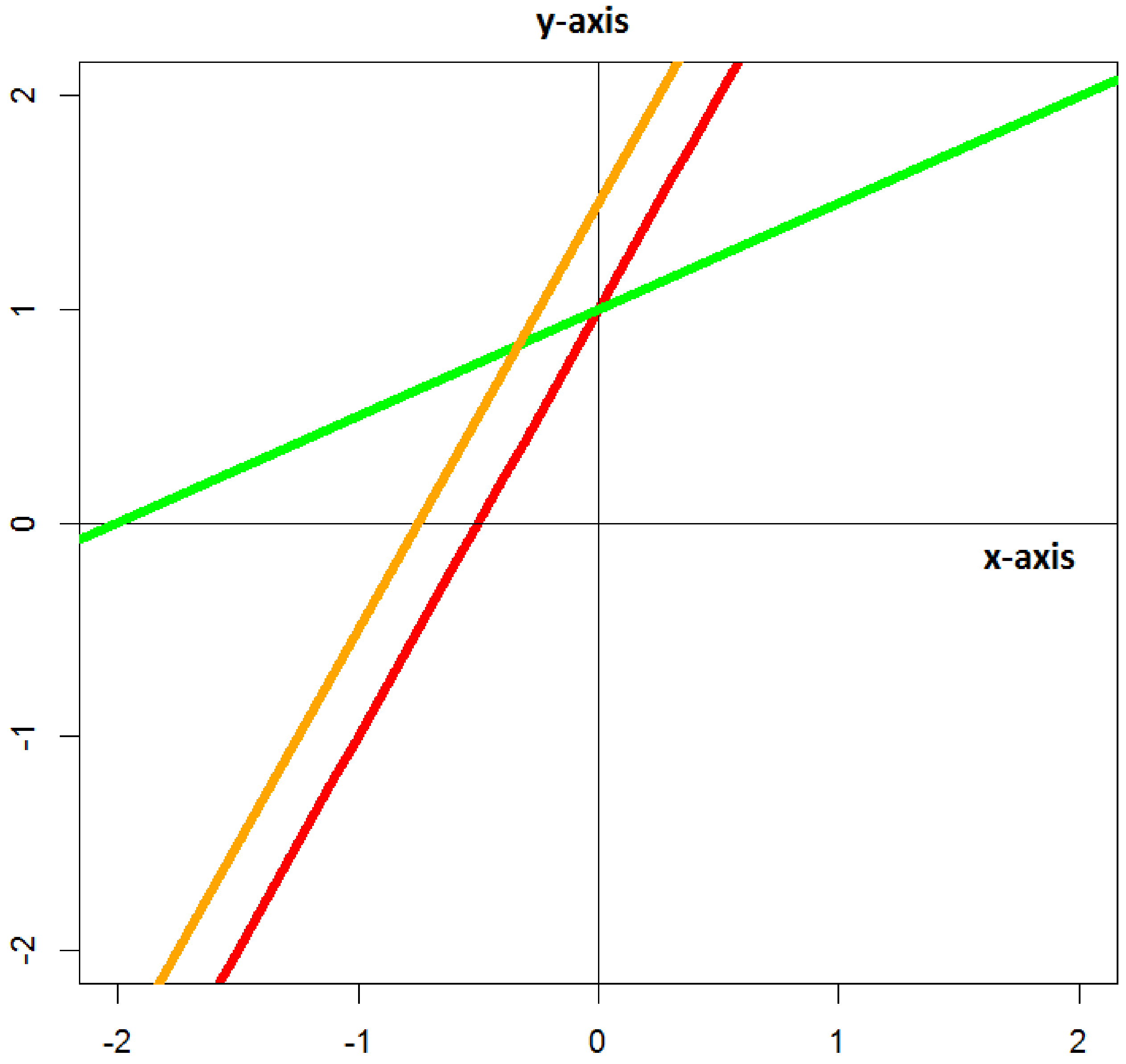
Example

- I add a orange line
- $y = 1.5 + 2x$
- Compare to the red line $y = 1 + 2x$, the intercept changes from 1 to 1.5.
- Can you see the difference?



Example

- Red line $y = 1 + 2x$
- Green line $y = 1 + 0.5x$
- Orange line $y = 1.5 + 2x$



Regression

- Regression analysis is a statistical method for estimating the relationships among two **quantitative variables**.

- **Regression Line –** predicts the value of y (response variable), as a straight line function of the value of x (explanatory variable)

Regression

- Why do we need regression?

- Let's consider the situation that we want to know % of fat in Shiwen's body. It's hard to measure % of fat, instead we can collect his blood pressure easily. If we know the regression line between % of fat and blood pressure, we can use Shiwen's blood pressure to estimate his % of fat.

- E.g. (% of fat) = -120+ (blood pressure)

Regression line

- $\hat{\boldsymbol{y}} = \boldsymbol{b_0} + \boldsymbol{b_1 x}$
  - $\boldsymbol{b_0}$ is the intercept (the value of $\hat{y}$ when x=0)
  - $\boldsymbol{b_1}$ is the slope of the line (the amount that $\hat{y}$ changes when x increases by one unit)
  - $\hat{\boldsymbol{y}}$ is the predicted value

  - **Residual** = (the real y) – $\hat{y}$

Regression R^2

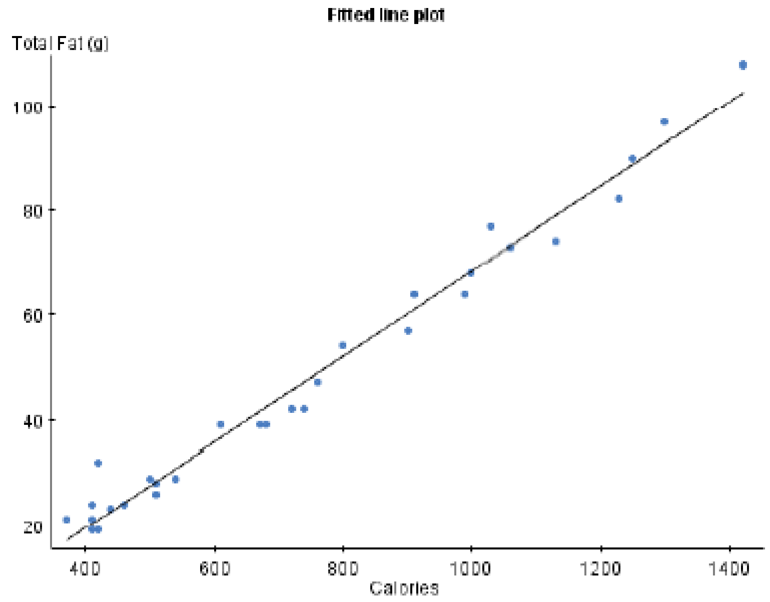
- $\boldsymbol{R^2}$, given in the regression output, gives the percent of variation in $\hat{y}$ explained by x

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

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**

Regression: Calories and Fat in Hamburger



Simple linear regression results:
Dependent Variable: Total Fat (g)
Independent Variable: Calories
Total Fat (g) = -12.907254 + 0.081350215 Calories
Sample size: 32
R (correlation coefficient) = 0.9894
R-sq = 0.9789471
Estimate of error standard deviation: 3.7394521

Parameter estimates:

Parameter	Estimate
Intercept	-12.907254
Slope	0.081350215

Regression: Example

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

Regression: Example

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

$$\widehat{Fat} = -12.9 + 0.0814 * (calories)$$
$$= -12.9 + 0.0814 * (1000)$$
$$= 68.5$$

Regression: Example

- 2) Write a thorough interpretation of the slope.
  - $\widehat{Fat} = -12.9 + 0.0814 * (calories)$
  - Here, the slope is .0814
    - So, with every unit increase in calories we expect a .0814 unit increase in grams of fat on average

Regression: Example

- 3) Write a thorough interpretation of regression R^2
  - R^2 = .9789 → 97.89% of the variation in fat is explained by calories

Regression: Example

- 4) Write a thorough interpretation of coefficient of correlation **r**

$r = \sqrt{R^2} = \sqrt{.9789} = .9894$ → since r is very close to one, Fat and Calories have a **very strong** positive linear correlation

Regression: Example

- 5) The hamburger with 1000 calories actually has 68 grams of fat. What is the residual?
  - Residual = (the real y) – $\hat{y}$
  - Residual = 68 – 68.5 = - 0.5

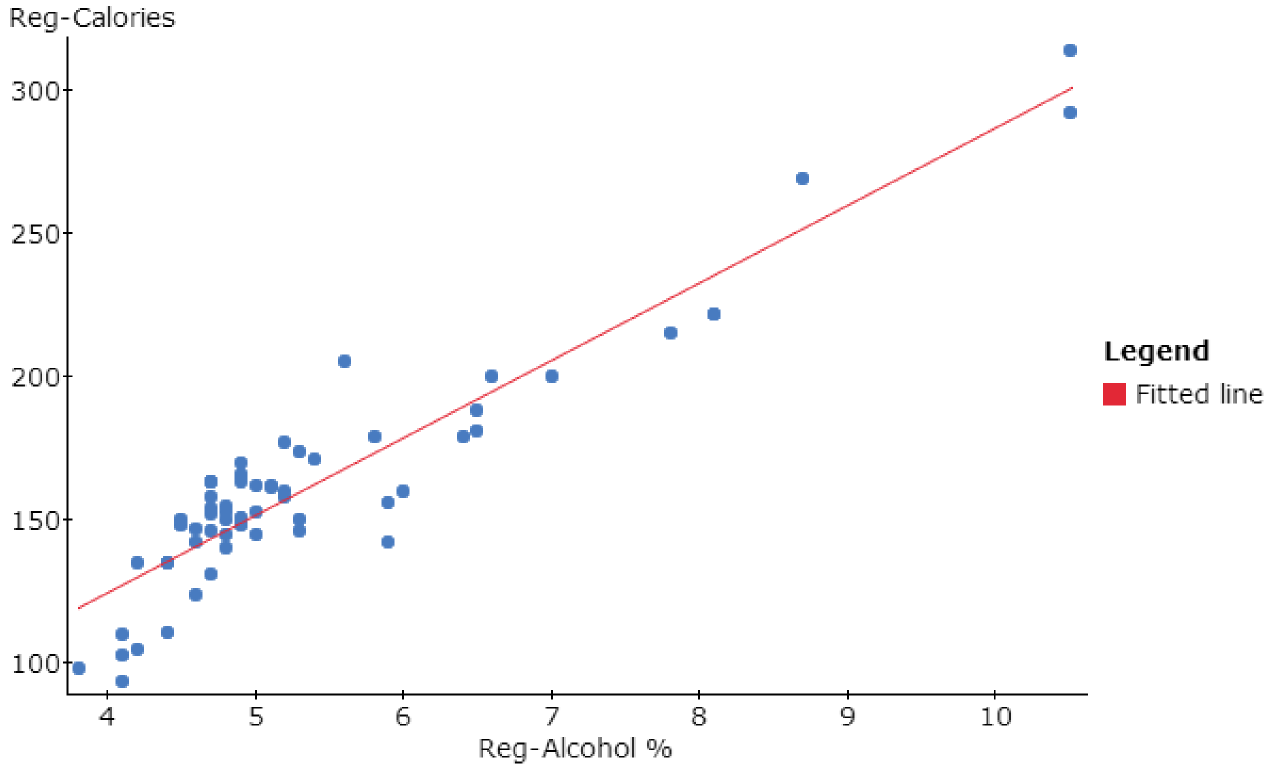
Regression: Beer Example

- In creating beer, yeast and sugar react to create alcohol. The more sugar and yeast you add the more alcohol level.

- **“It would make sense that the more alcohol in the beer, the more carbohydrates, so that more calories.”**

- Do you agree my statements? Let’s show it statistically.

Regression: Beer Example



- Here, we see a positive correlation. (moderate or strong?)

Regression: Beer Example

Simple linear regression results:

Dependent Variable: Reg-Calories

Independent Variable: Reg-Alcohol %

Reg-Calories = 16.374148 + 27.003873 Reg-Alcohol %

Sample size: 61

R (correlation coefficient) = 0.93198924

R-sq = 0.86860395

Estimate of error standard deviation: 14.762875

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-Value
Intercept	16.374148	7.6283163	$\neq 0$	59	2.1464957	0.036
Slope	27.003873	1.3673519	$\neq 0$	59	19.749029	<0.0001

Regression: Beer Example

- The regression line is
- (Calories) = 16.37+ 27 * (Alcohol %)
- The intercept is 16.37 and the slope is 27
- Regression R^2 is 0.87
- Correlation Coefficient **r** is 0.93

- **Can you interpret them? Write it down!**

Regression: Beer Example

- **Intercept:** when the alcohol percentage is 0, we expect the bear has calories 16.37.
- **Slope:** for each percentage increase in alcohol level there is an increase of 27 calories on average
- **Regression R^2:** 86.86% of the variation in calories is explained by alcohol percentage
- **Correlation Coefficient r:** there is a strong linear relationship between alcohol percentage and calories. (it confirms our visual observation!)

Regression: Beer Example

- If we wanted to **estimate** the calories of Rogue Dead Guy Ale, we can plug in its alcohol percentage into the equation to find an estimate of the calories.
- Rogue Dead Guy Ale has alcohol % to be 6.6%. we can plug it in to find the estimated calories of a bottle of Rogue Dead Guy Ale.

$$(Calories) = 16.37 + 27 * (Alcohol\%)$$
$$= 16.37 + 27 * (6.6)$$
$$= 194.57$$

Regression: Beer Example

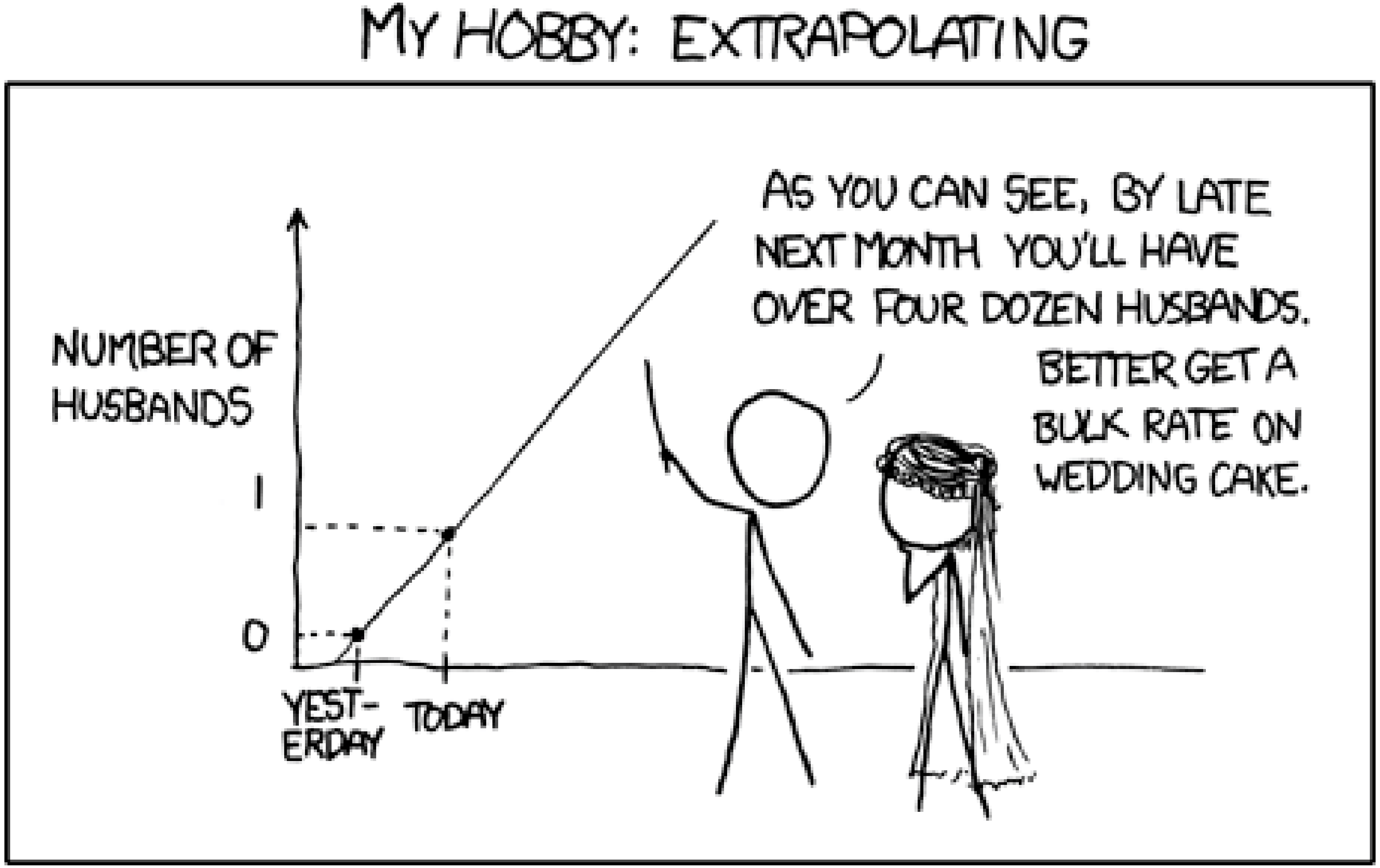
- So, the estimated amount of calories for Rogue Dead Guy is 194.57.
- In fact, the actual amount of calories is 198 so our estimate isn't perfect, but very close!

- **The residual** is the difference
  - Real value – $\hat{y}$(estimated value) = 198-194.57=3.43.

Something you need to be careful

- **Extrapolation**: we don't want to predict using x values different than the known data

Extrapolation



Something you need to be careful

- **Lurking Variables:** a variable that we don't look at that causes the correlation.
- **Confounding:** a study occurs when the effects of two or more variables are mixed together. It is often caused by a lurking variable.

Confounding and Lurking Variable

- It’s hot outside
- Where do you go? Beach! Swimming in the sea!
- What do you eat? Ice cream!
- We had a hot summer, so there would be more swimming and eating ice cream, and thus, more drowning deaths.
- If someone wasn’t carful and claims …
- The increase sales of ice cream causes drowning deaths.
- What do you say?

Something you need to be careful

- **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**