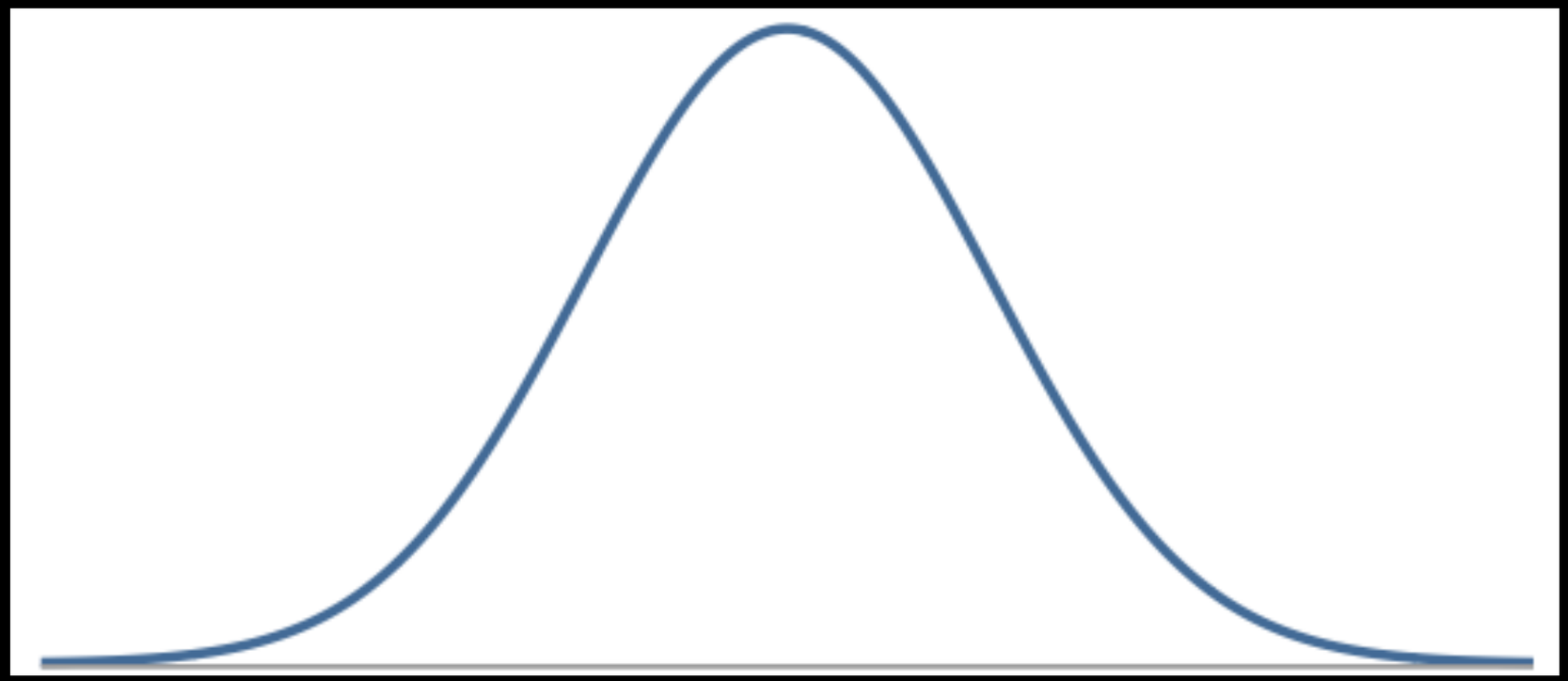
Welcome to Week #10!

Recall - where we are/what we have covered so far:

The Normal distribution

In Chapter 3, we look at the Normal distribution. The Normal distribution is the most famous continuous distribution.



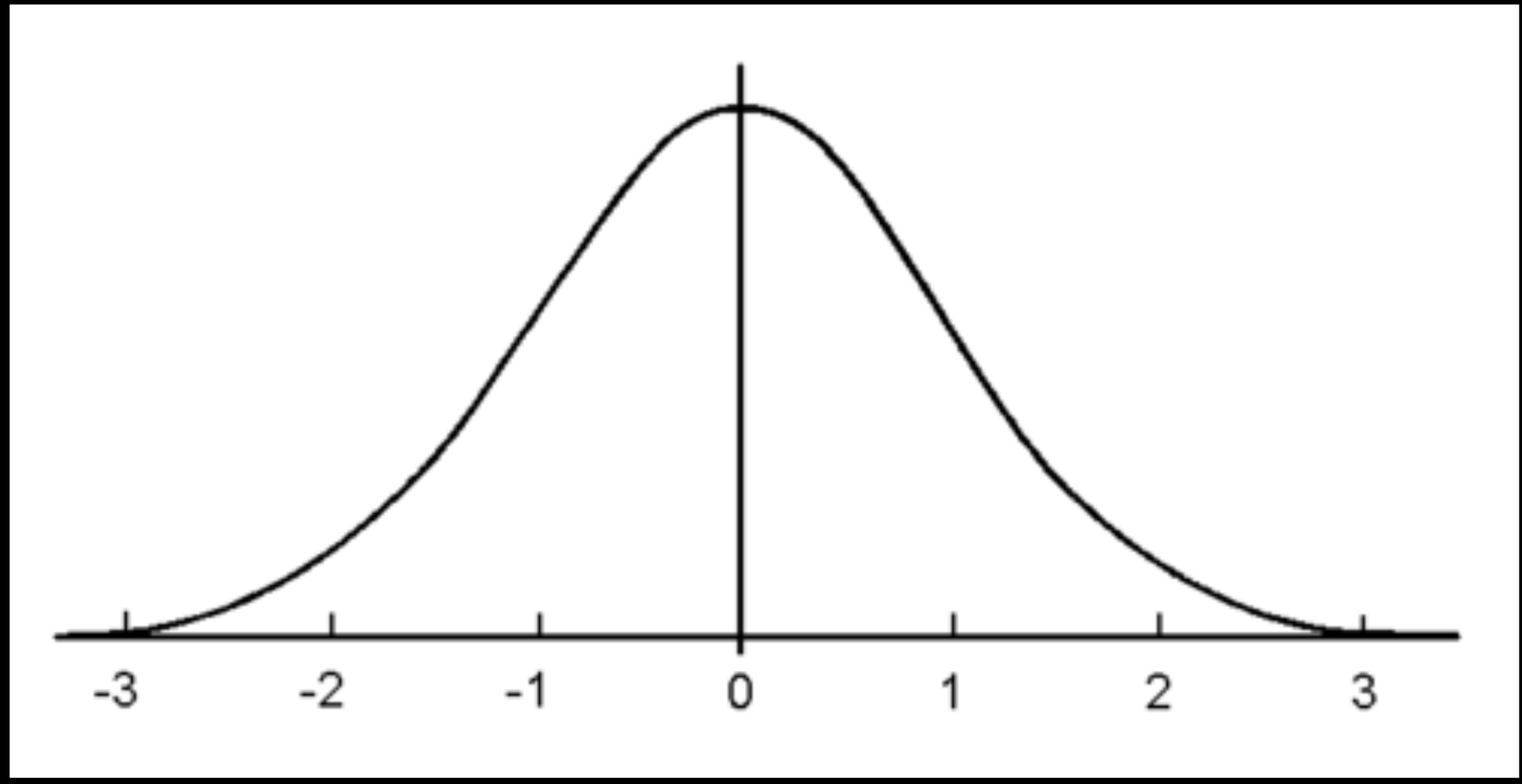
To find areas under curves, we generally use a table or technology (i.e. calculator, stat program, etc.).

The Standard Normal Curve

$$Z = \frac{(\text{observation} - \text{mean})}{SD}$$

What units are on the horizontal axis?
Z-scores!

A way to compare normal distributions

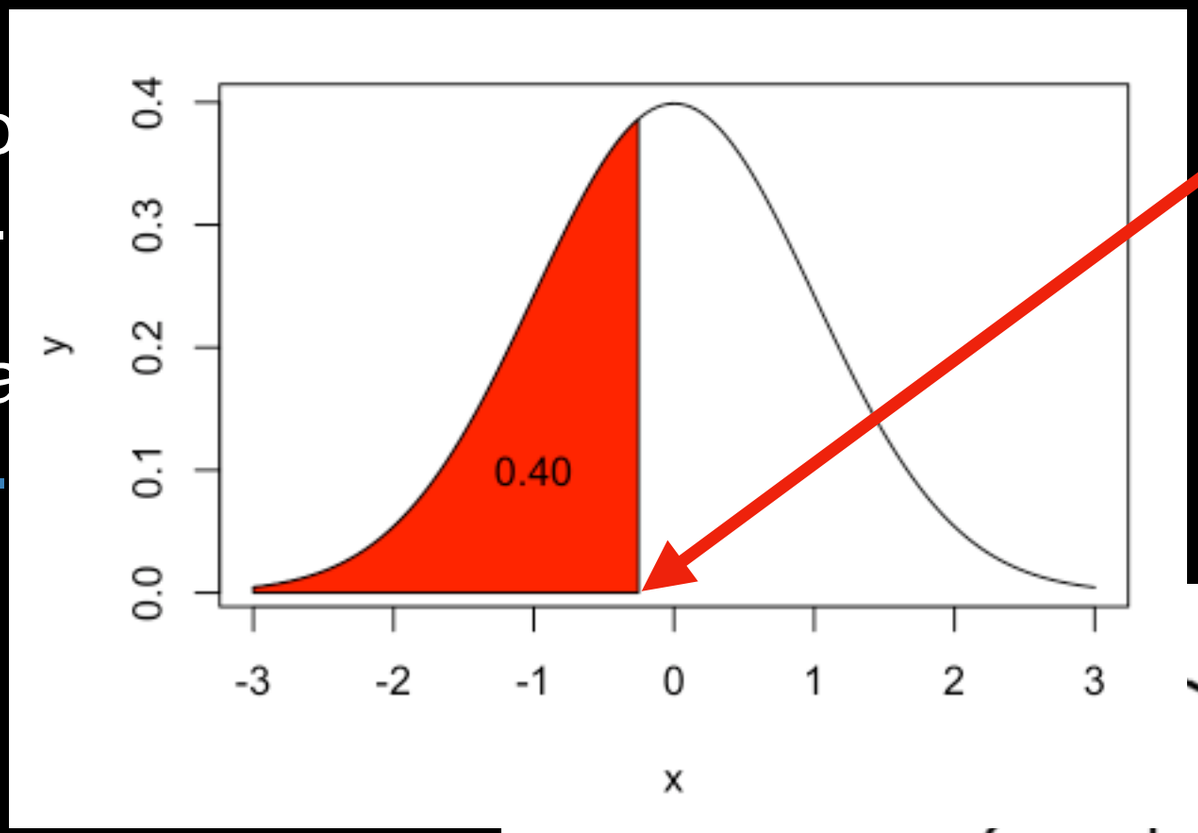


Finding percentiles from the standard normal curve

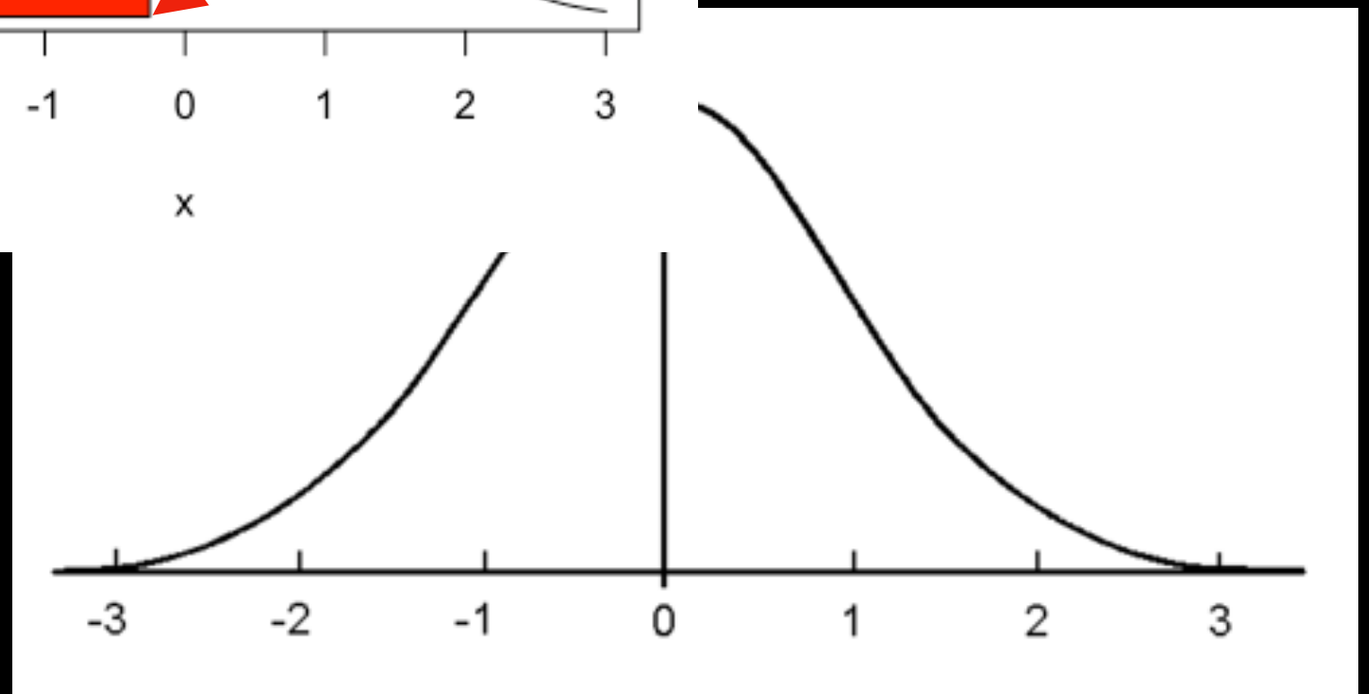
What Z-score corresponds to the 50th percentile?
i.e. P(Z < ?) = 0.5 Z =

What Z-score co
i.e. P(Z < ?) = 0.

What Z-score ha
i.e. P(Z < ?) = 0.



What is this number such that red area = 0.40 (40%)



Summary: Set the hypothesis

The general outline of the process:

1. Set the hypotheses.
   For a single proportion this will look like:
   H_0: μ = null value
   H_A: μ < or > or ≠ null value
2. Check assumptions and conditions
3. Calculate a test statistic and a p-value
4. Make a decision, and interpret it in context

- If p-value < α, reject H_0,
  there is sufficient evidence for [H_A]

- If p-value > α, do not reject H_0,
  there is not sufficient for evidence for [H_A]

Anatomy of a test statistic

The general form of a test statistic is

$$\frac{\text{point estimate} - \text{null value}}{\text{SE of point estimate}}$$

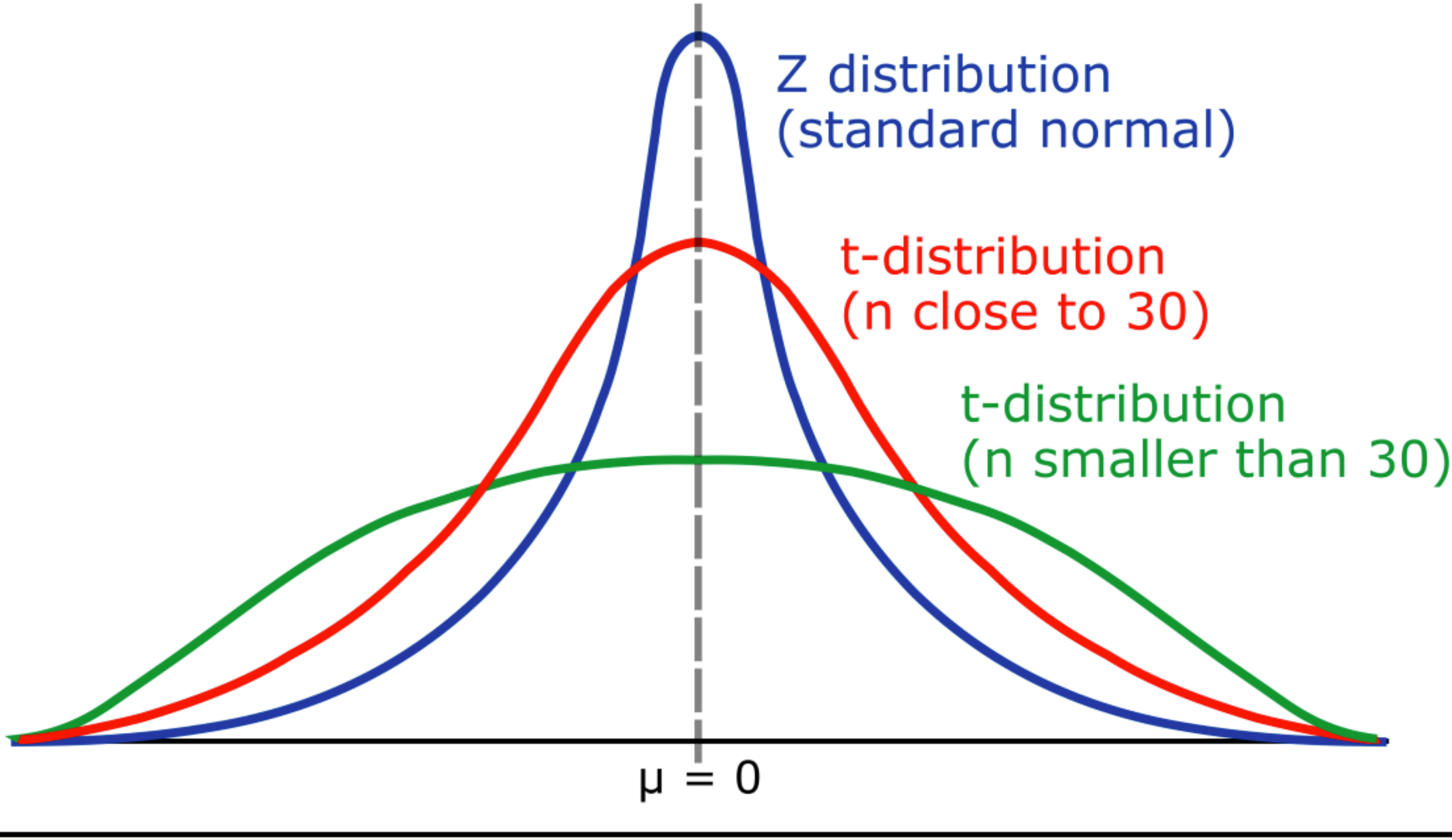
Only tricks are:
(1) picking what the point and null values are based on our hypotheses

(2) what the form of the standard errors is based on what our underlying distribution looks like (normal, t-distribution, $\chi^{2)}$

This construction is based on

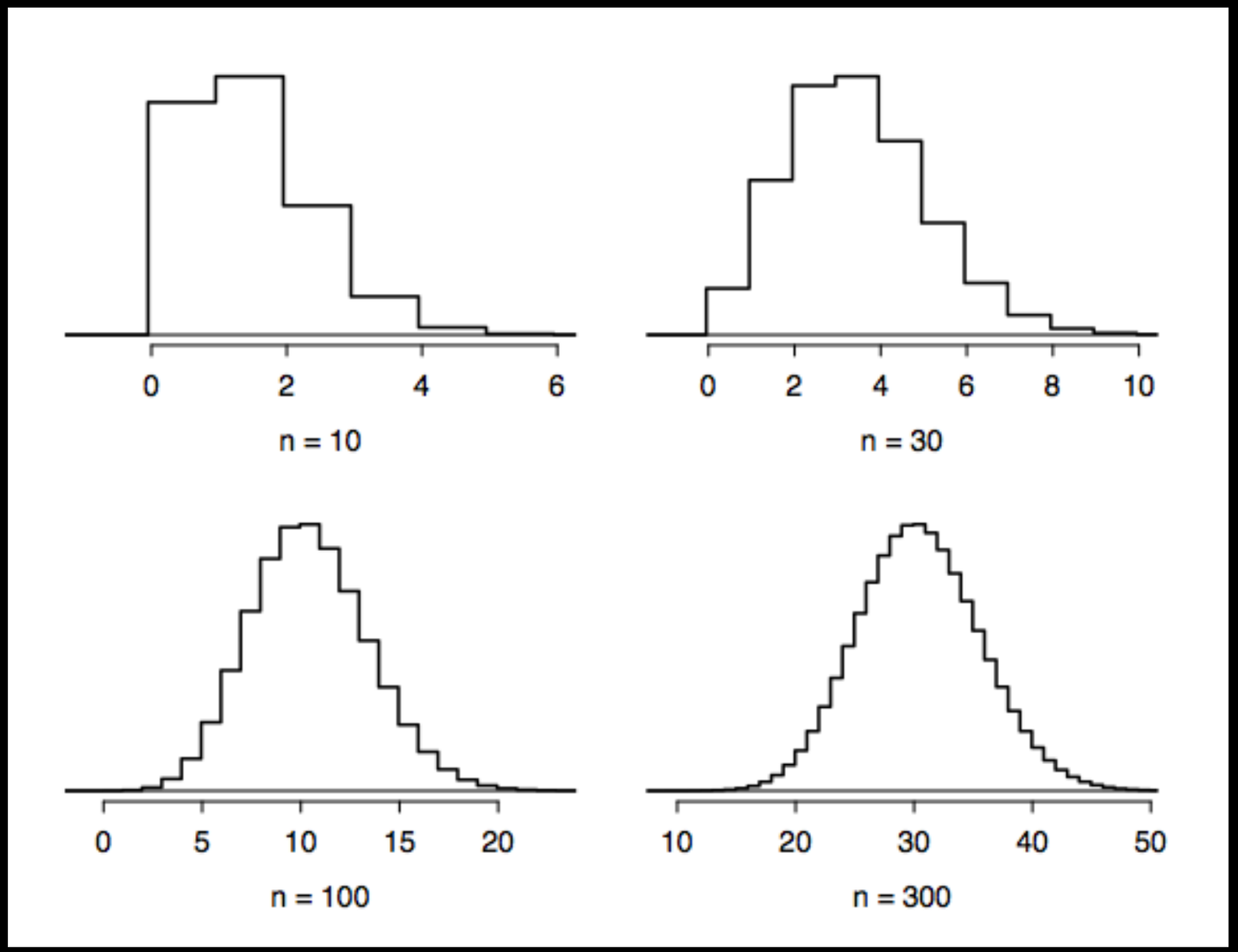
- identifying the difference between a point estimate and an expected value if the null hypothesis was true, and
- standardizing that difference using the standard error of the point estimate.

These two ideas will help in the construction of an appropriate test statistic for count data.


Z distribution
(standard normal)
t-distribution
(n close to 30)
t-distribution
(n smaller than 30)
μ = 0

Distributions of number of successes

Hollow histograms of samples from the binomial model where **p = 0.10** and n = 10, 30, 100, and 300.
What happens as n increases?



Note: the scales on the histograms are different!

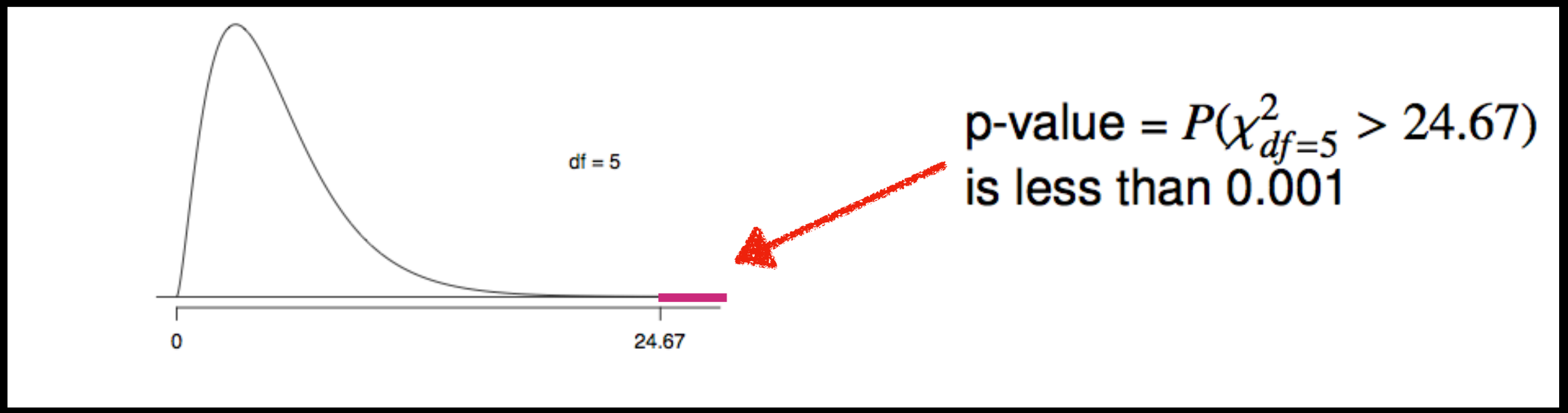
See this applet with sliders for n and p to see how shape binomial distribution changes as n and p change:

http://www.stat.berkeley.edu/~stark/Java/Html/BinHist.htm

χ^2 & ANOVA - For more complex datasets

Finding a p-value for a chi-square test

The p-value for a chi-square test is defined as the tail area above the calculated test statistic.



(more on how to do this in R in a few slides)

p-value < 0.05 (our typical level of significance)

Reject H_0, the data provide convincing evidence that the dice are biased.

z/t test vs. ANOVA - Method

Z or T test

Compute a test statistic
(a ratio).

$$z/t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE(\bar{x}_1 - \bar{x}_2)}$$

ANOVA

Compute a test statistic
(a ratio).

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$

mean square between groups (MSG)

mean square error (MSE)

Large test statistics lead to small p-values.

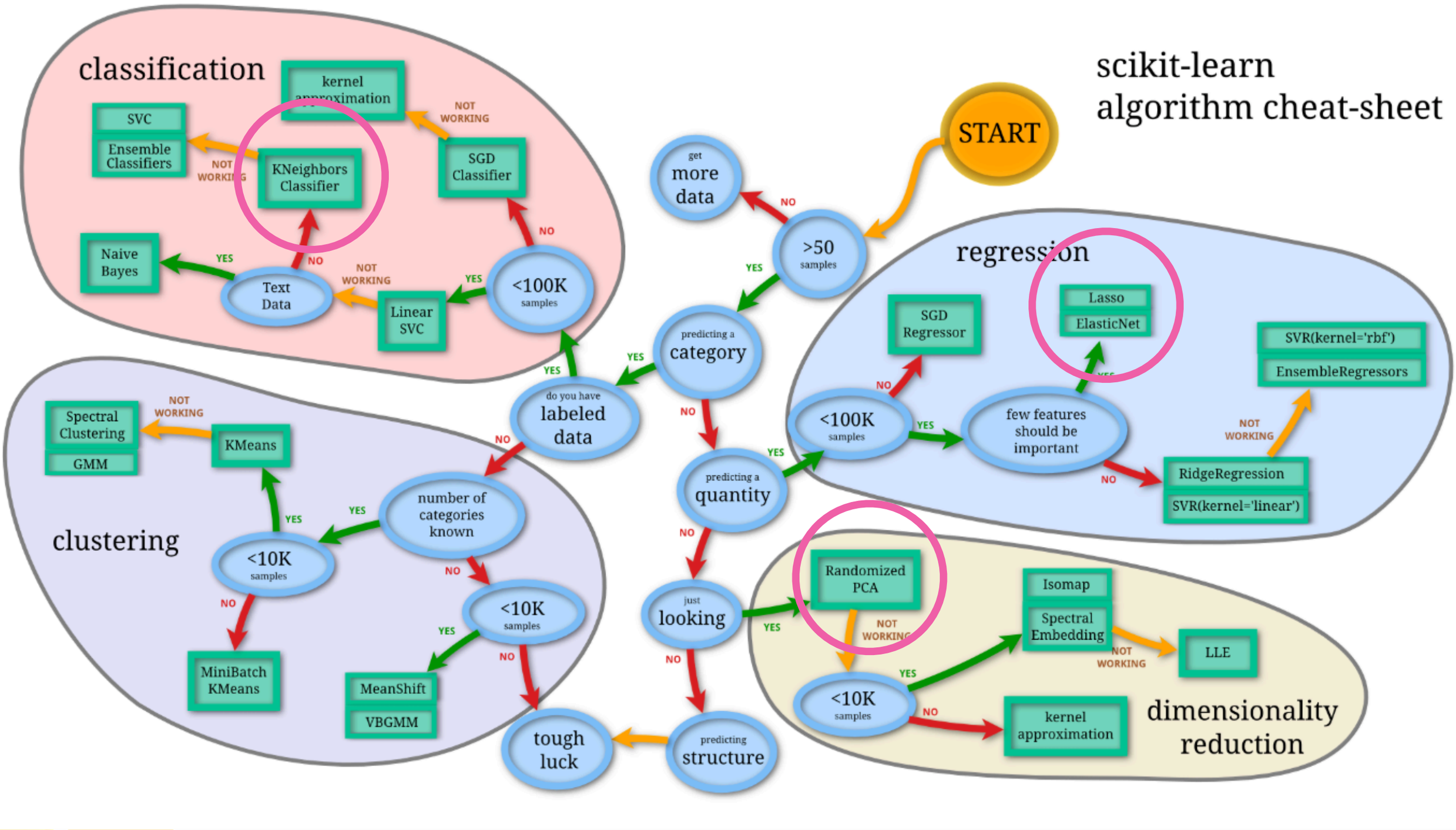
If the p-value is small enough H_0 is rejected, we conclude that the population means are not equal.

In order to be able to reject H_0, we need a small p-value, which requires a large F statistic.

In order to obtain a large F statistic, variability between sample means needs to be greater than variability within sample means.

Linear Regression: Beginning ML

Linear Regression: Beginning ML - Where are we going with this?

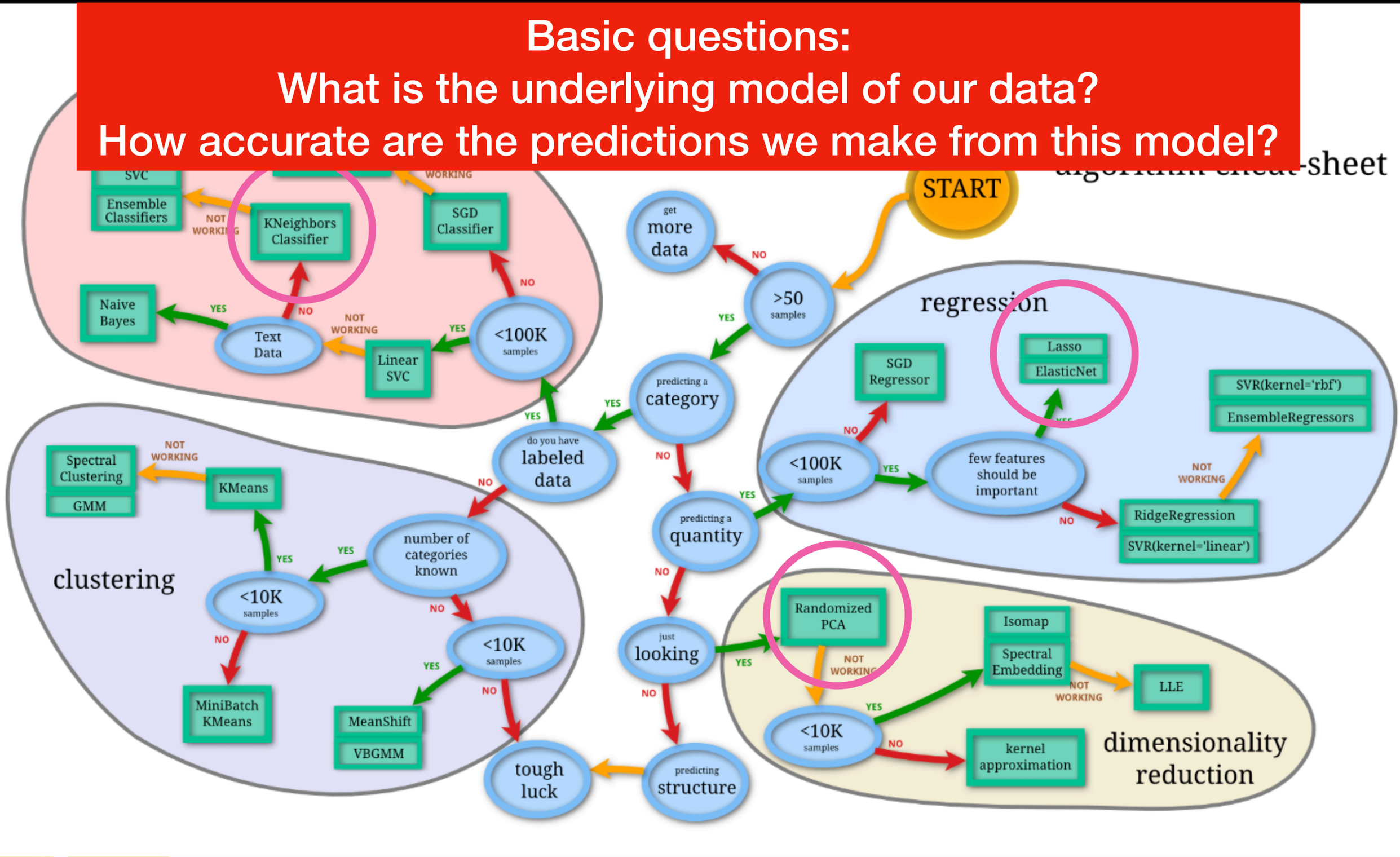


Linear Regression: Beginning ML - Where are we going with this?

Basic questions:

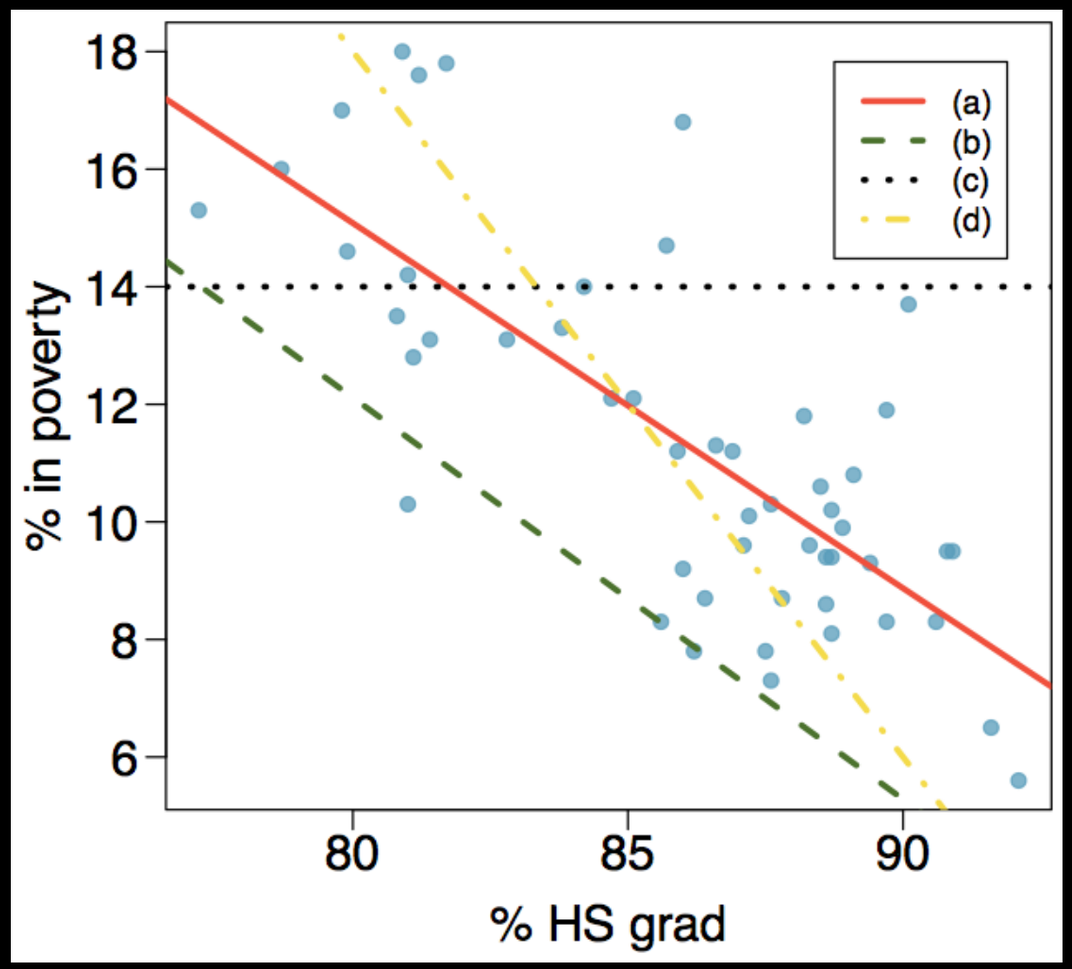
What is the underlying model of our data?

How accurate are the predictions we make from this model?



Poverty vs. HS graduate rate

The scatterplot below shows the relationship between HS graduate rate in all 50 US states and DC and the percent of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Goal #1: to fit a line to this relationship

Goal #2: Quantify how "good" this fit is

Quantifying the relationship

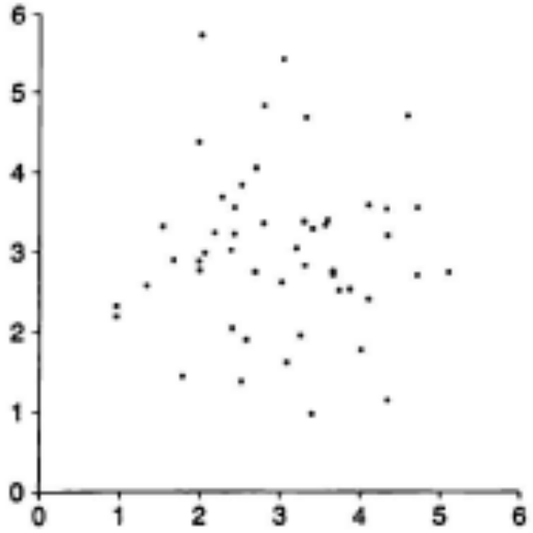
Correlation (R) describes the strength of the linear association between two variables.

It takes values between -1 (perfect negative) and +1 (perfect positive).

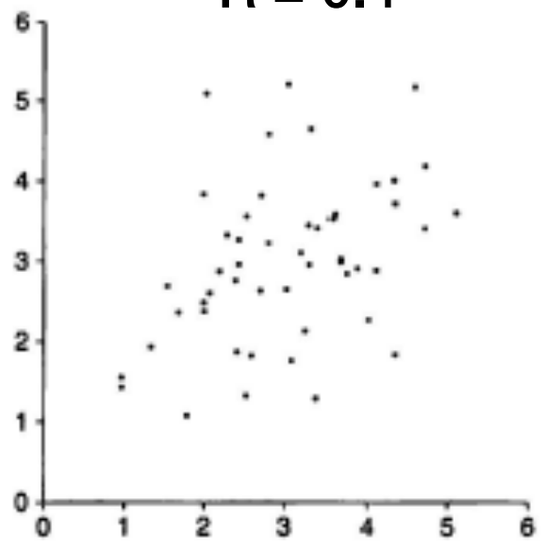
A value of 0 indicates no *linear* association.

(difference in x-values from the x-sample mean) X (difference in y-values from the y-sample mean)/[(variability of x) X (variability of y)]

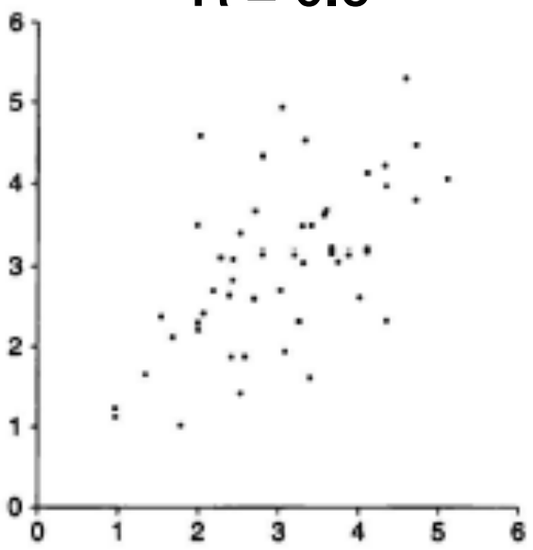
R = 0



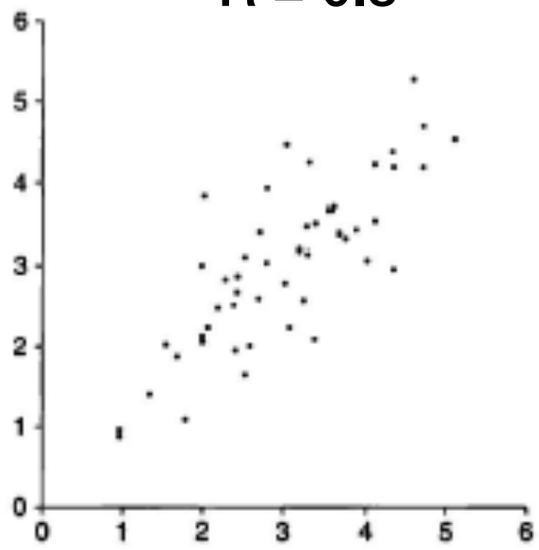
R = 0.4



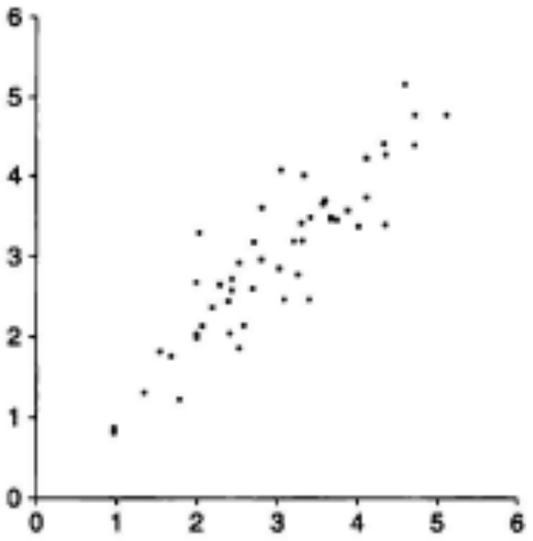
R = 0.6



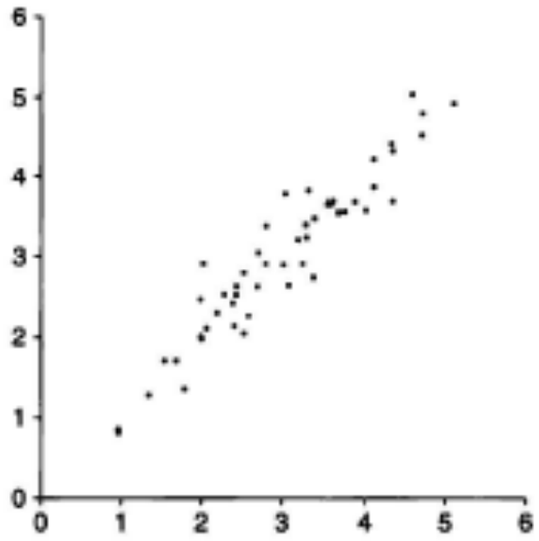
R = 0.8



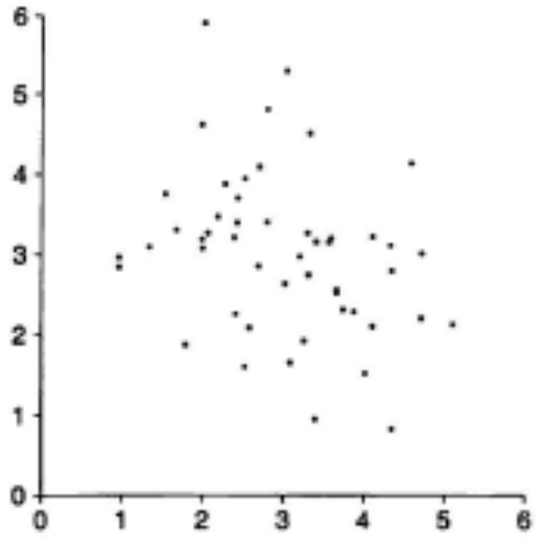
R = 0.9



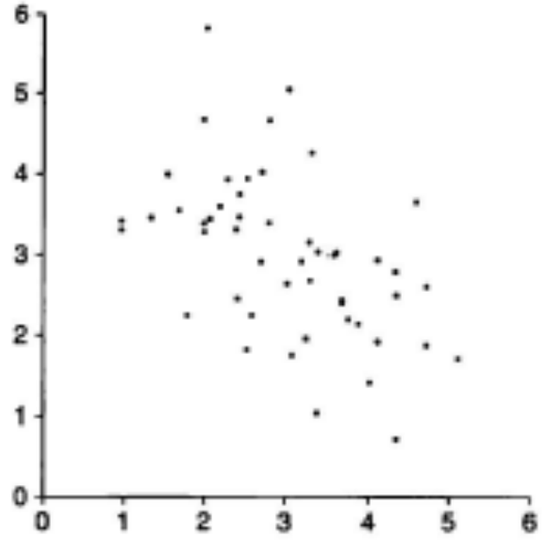
R = 0.95



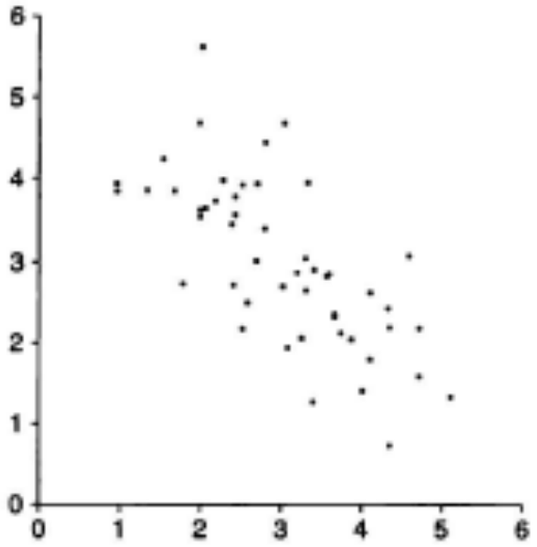
R = -0.3



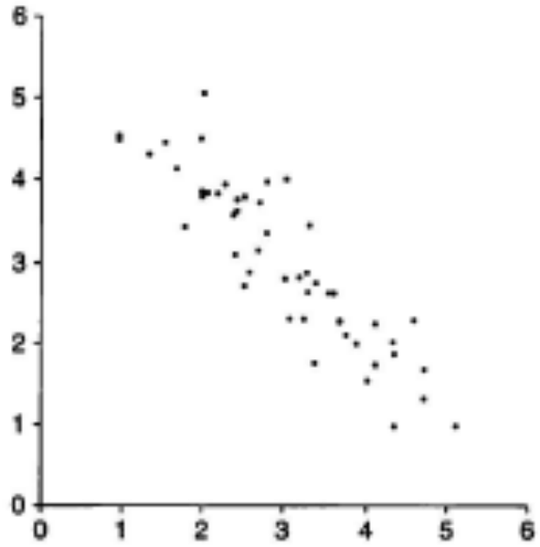
R = -0.5



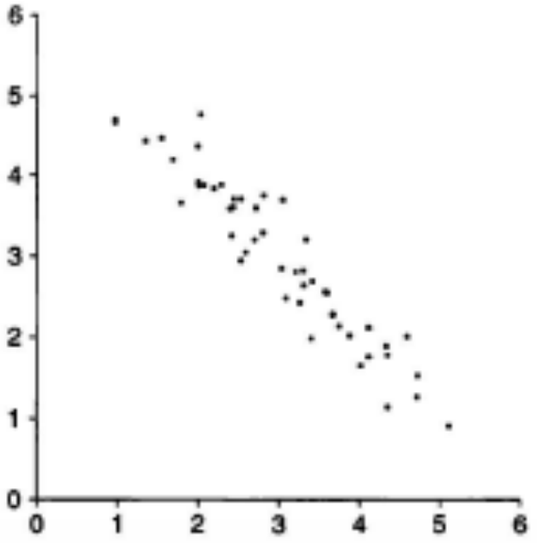
R = -0.7



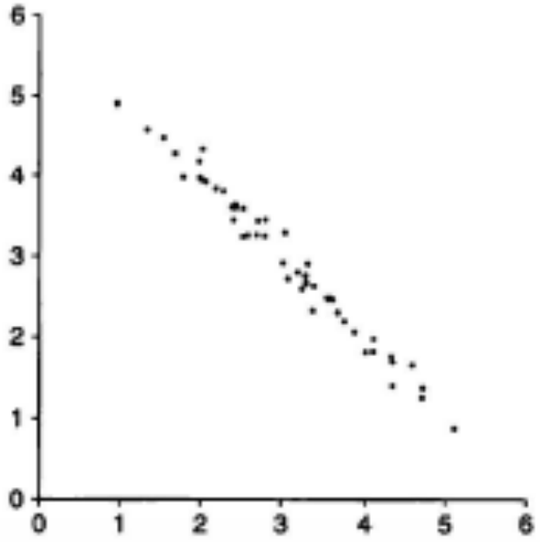
R = -0.9



R = -0.95



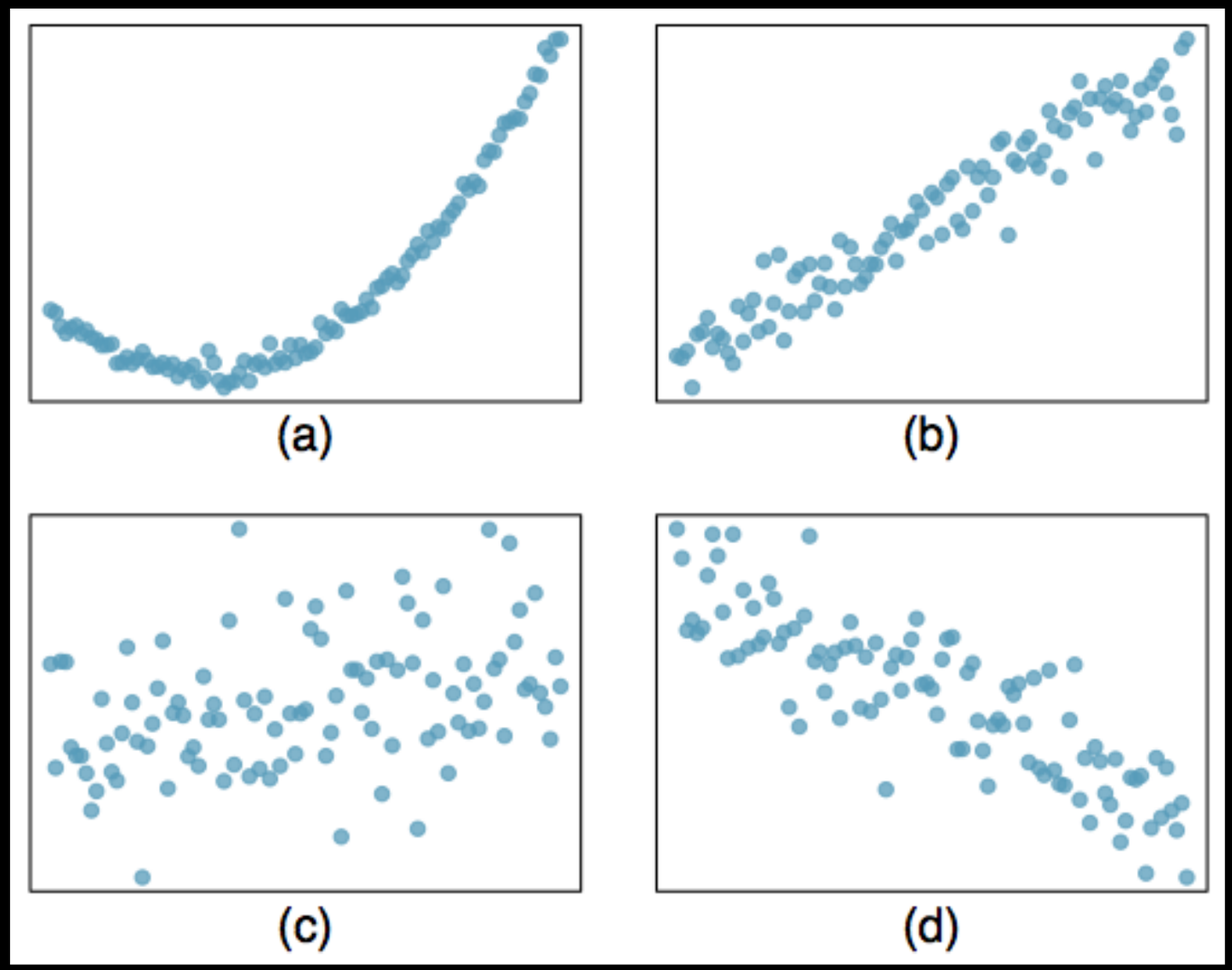
R = -0.99



Assessing the correlation

Which of the following is has the strongest correlation, i.e. correlation coefficient closest to +1 or -1?

(b) → correlation means linear association

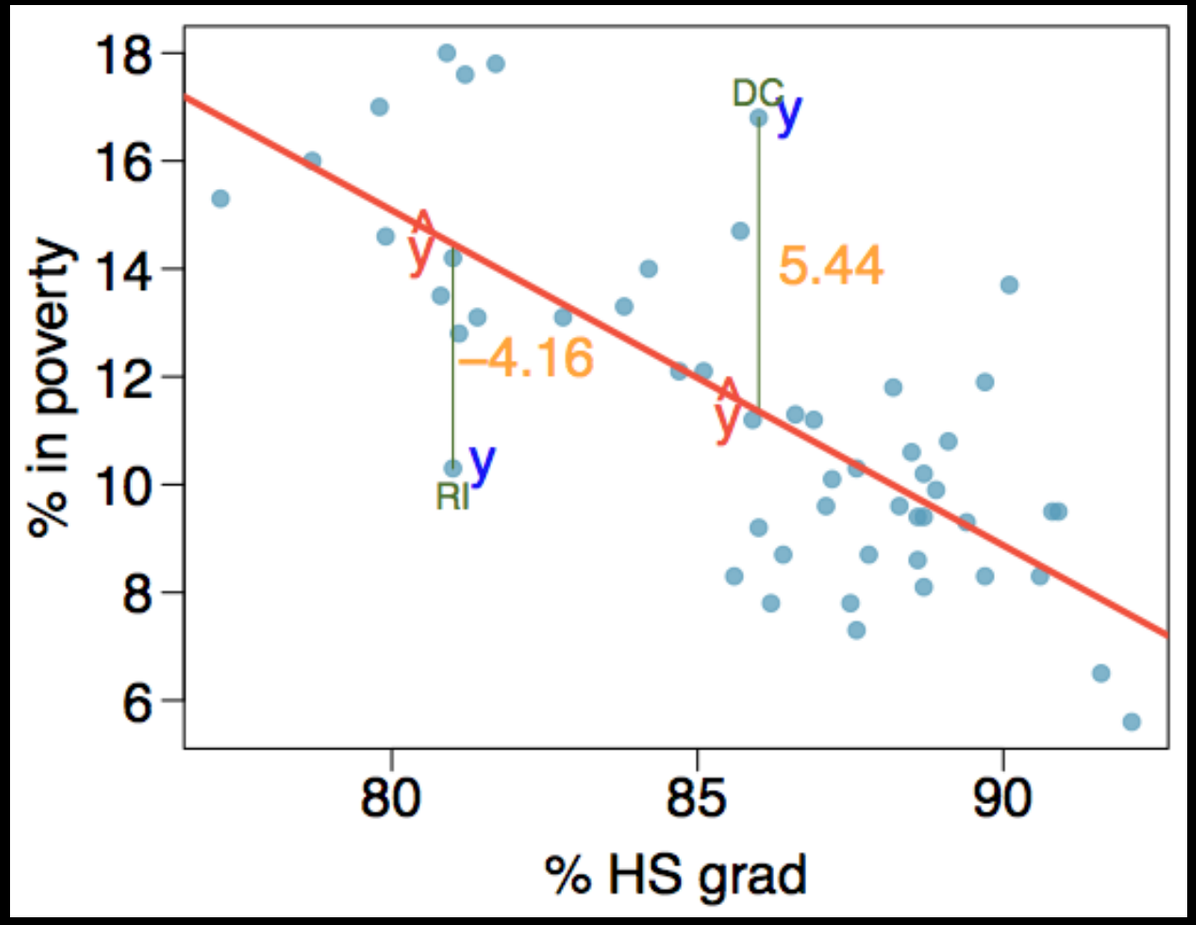


Residuals

Residuals are the leftovers from the model fit: Data = Fit + Residual

Aka a residual is the difference between the observed (y_i) and predicted $\hat{y}_i$.

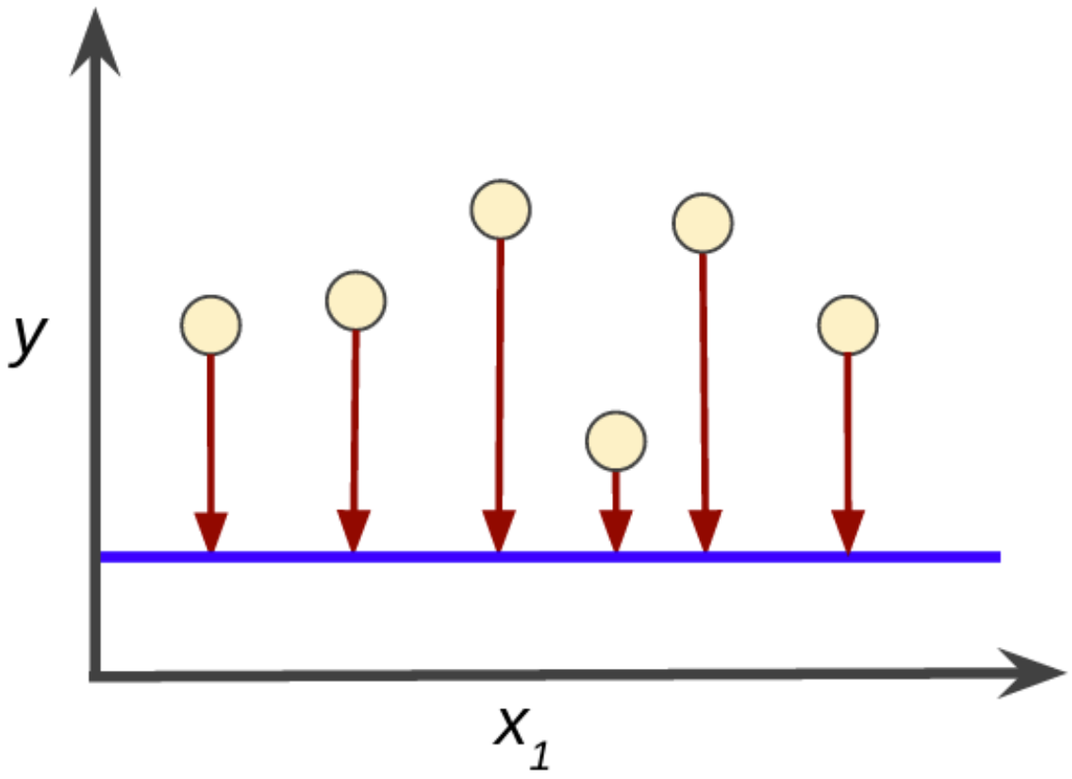
$e_i = y_i - \hat{y}_i$

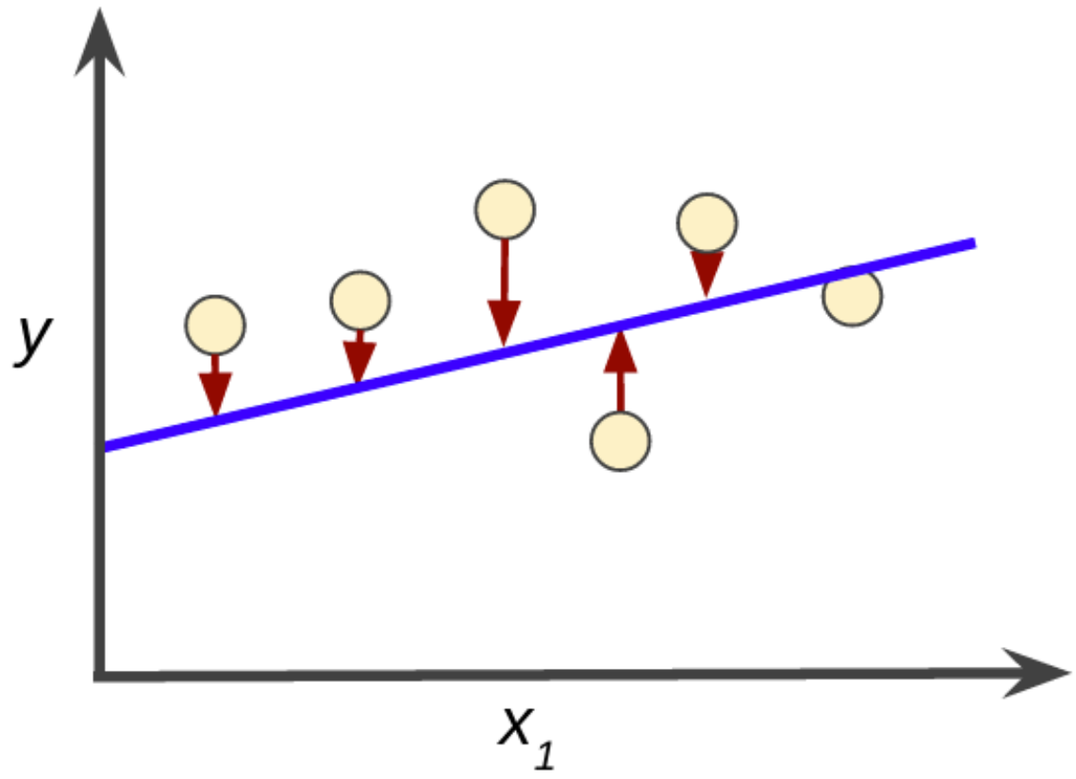


Here is a depiction of the residuals - how far each point is from our fitted line.

What is the “best” line?

$$e_i = y_i - \hat{y}_i$$





The one that minimizes the residuals.

In practice we minimize a function.

The precise function that we minimize is based on our residuals, and called the *Loss Function*.

A measure for the best line

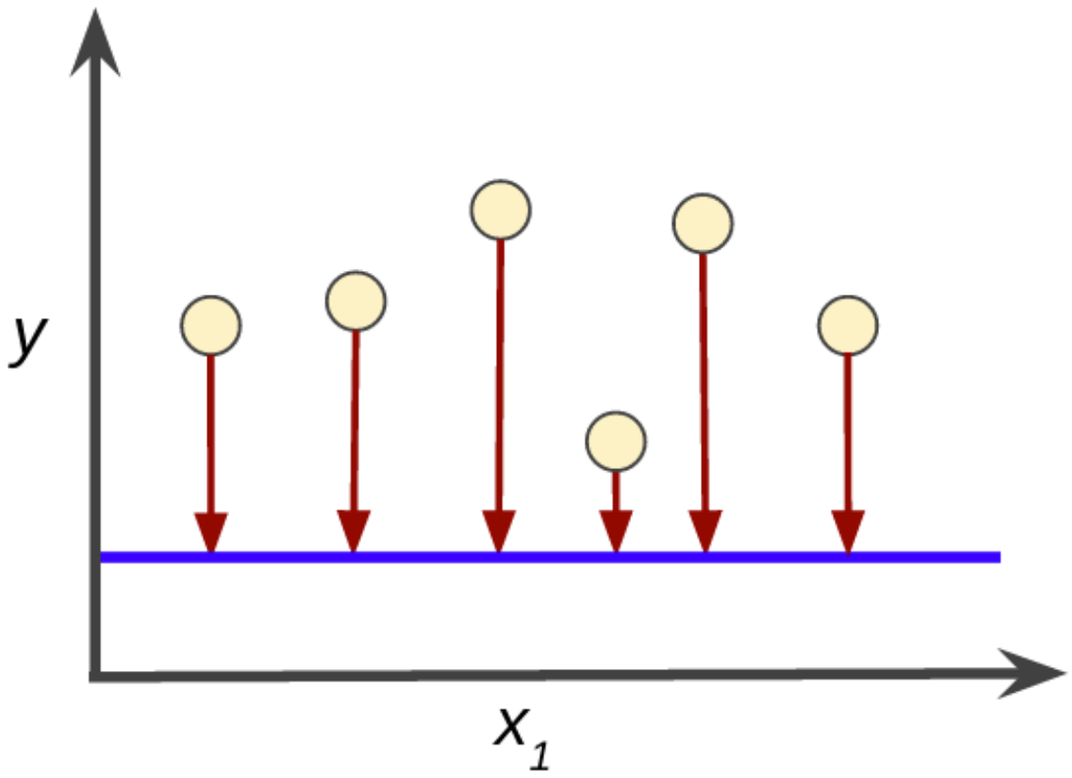
We want a line that has small residuals

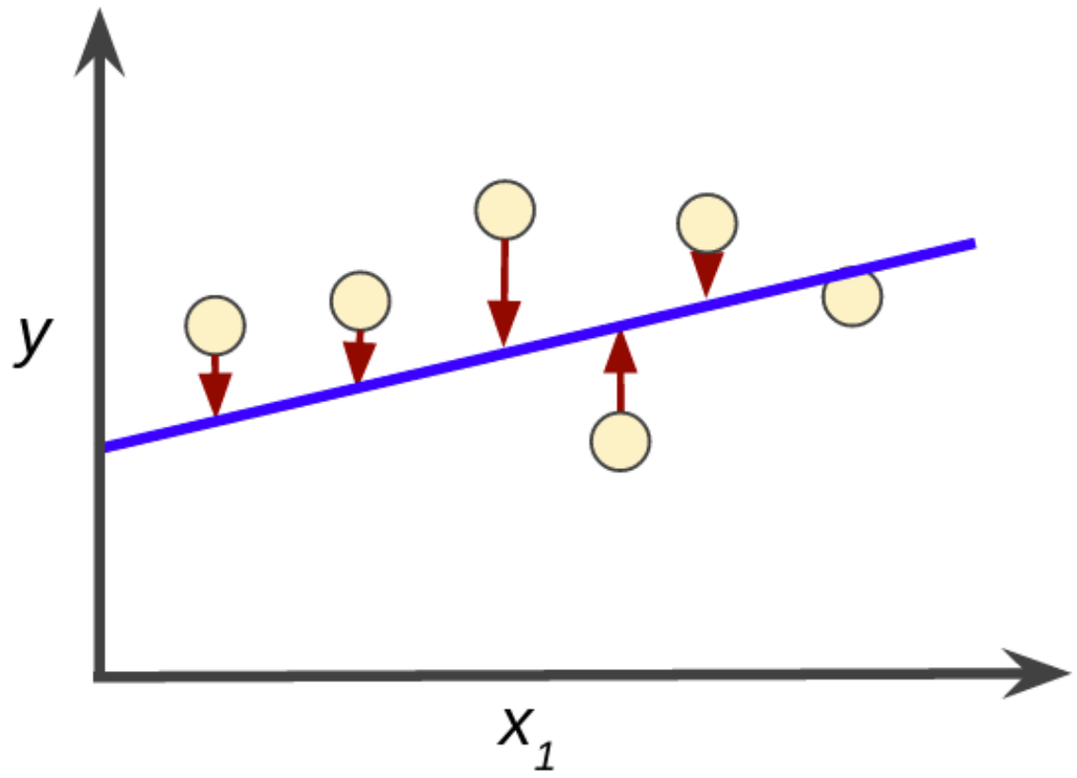
Option 1: Minimize the sum of magnitudes (absolute values) of residuals

$|e_1| + |e_2| + \ldots + |e_n|$

Option 2: Minimize the sum of squared residuals -- least squares

$e_1^2 + e_2^2 + \ldots + e_n^2$





A measure for the best line

We want a line that has small residuals

Option 1: Minimize the sum of magnitudes (absolute values) of residuals

$|e_1| + |e_2| + \ldots + |e_n|$

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

Option 2: Minimize the sum of squared residuals -- least squares

$e_1^2 + e_2^2 + \ldots + e_n^2$

FYI: this is a typical calculus optimization problem - finding the minimum of squared residuals

Why least squares?

- Most commonly used
- Easier to compute by hand and using software
- In many applications, a residual twice as large as another is usually more than twice as bad

A measure for the best line

We want a line that has small residuals

Option 1: Minimize the sum of magnitudes (absolute values) of residuals

$|e_1| + |e_2| + \ldots + |e_n|$

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

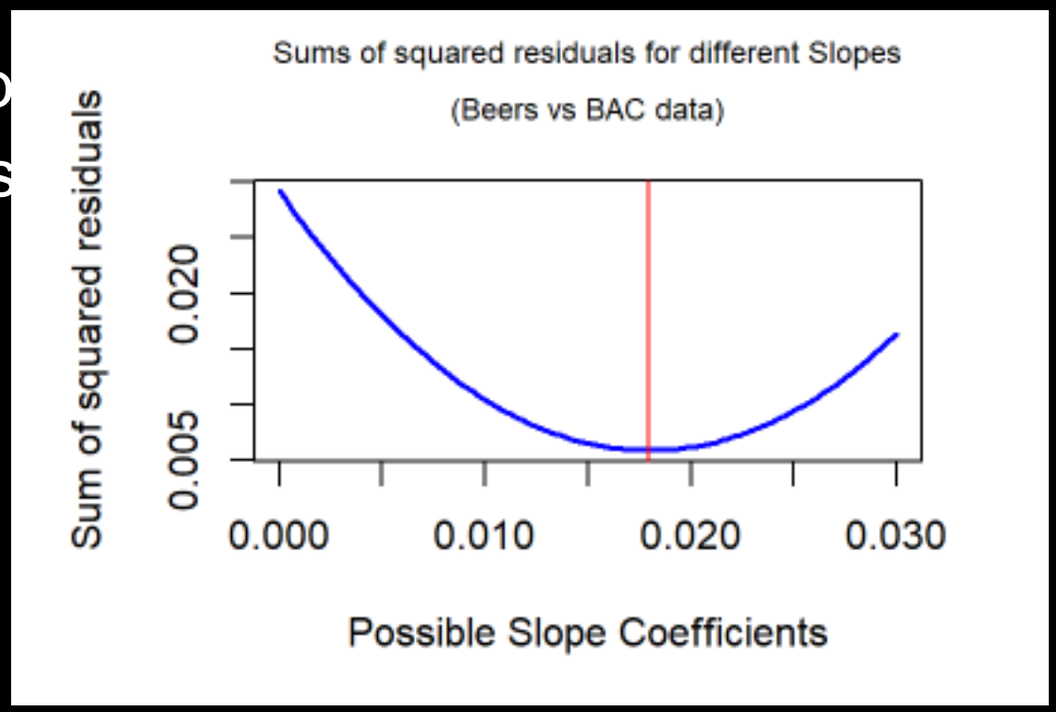
Option 2: Minimize the sum of squared residuals -- least squares

$e_1^2 + e_2^2 + \ldots + e_n^2$

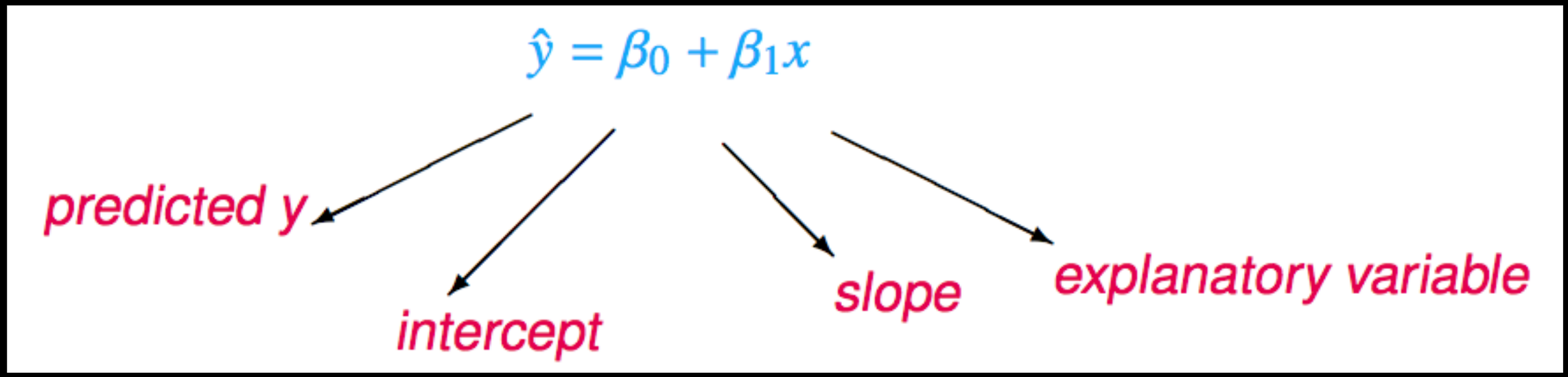
FYI: this is a typical calculus optimization problem - finding the minimum of squared residuals

Why least squares?

- Most commonly used
- Easier to compute by hand and using so
- In many applications, a residual twice as more than twice as bad



The least squares line: What are we actually fitting?



Intercept Notation

- Parameter: β_0
- Point estimate: b_0

Slope Notation

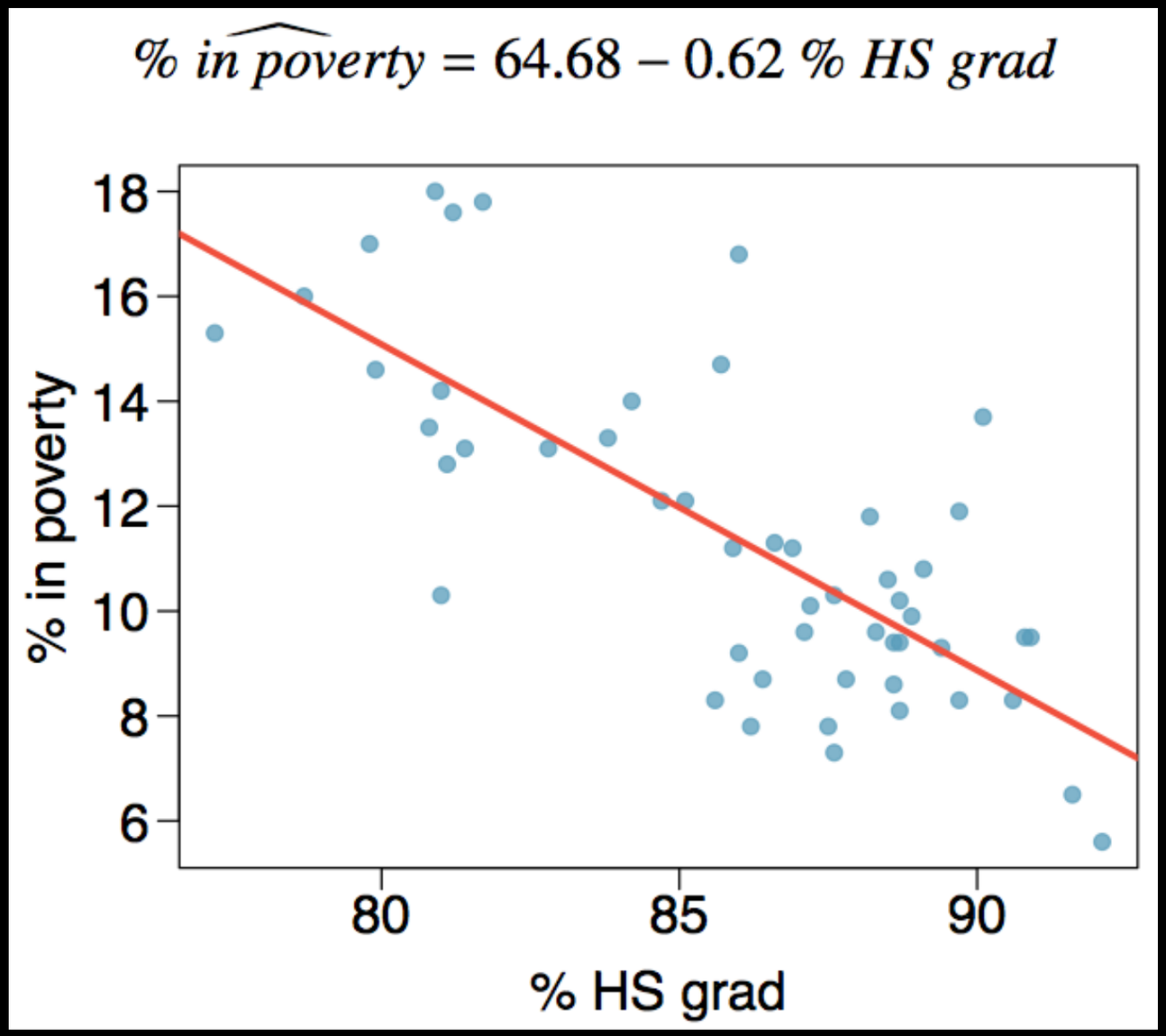
- Parameter: β_1
- Point estimate: b_1

sometimes you'll see β or b interchanged - "b" is technically for our point (sample) estimates

^ - means sample
no "^" means population

Regression line for our example

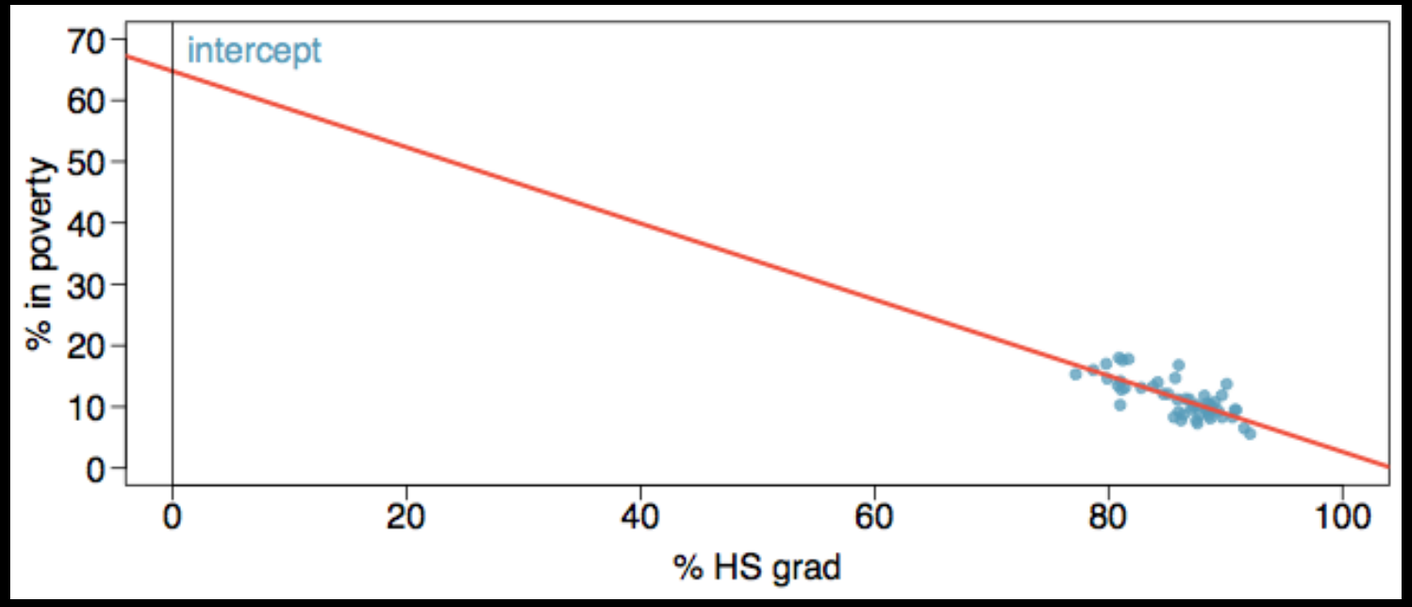
The scatterplot below shows the relationship between HS graduate rate in all 50 US states and DC and the percent of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Intercept

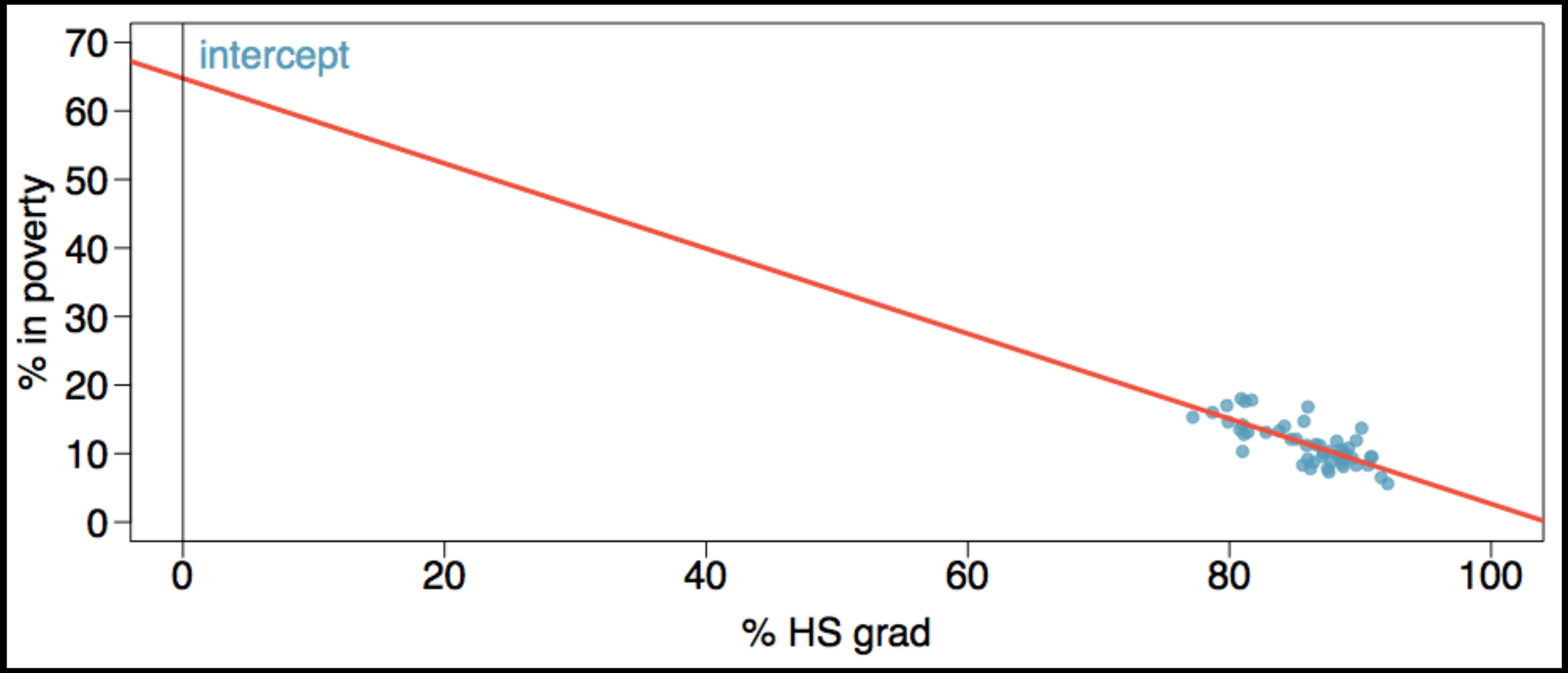
The intercept is where the regression line intersects the y-axis. The calculation of the intercept uses the fact the a regression line always passes through ($\bar{x}$, $\bar{y}$).

$$b_0 = \bar{y} - b_1 \bar{x}$$



More on the intercept

Since there are no states in the dataset with no HS graduates, the intercept is of no interest, not very useful, and also not reliable since the predicted value of the intercept is so far from the bulk of the data.



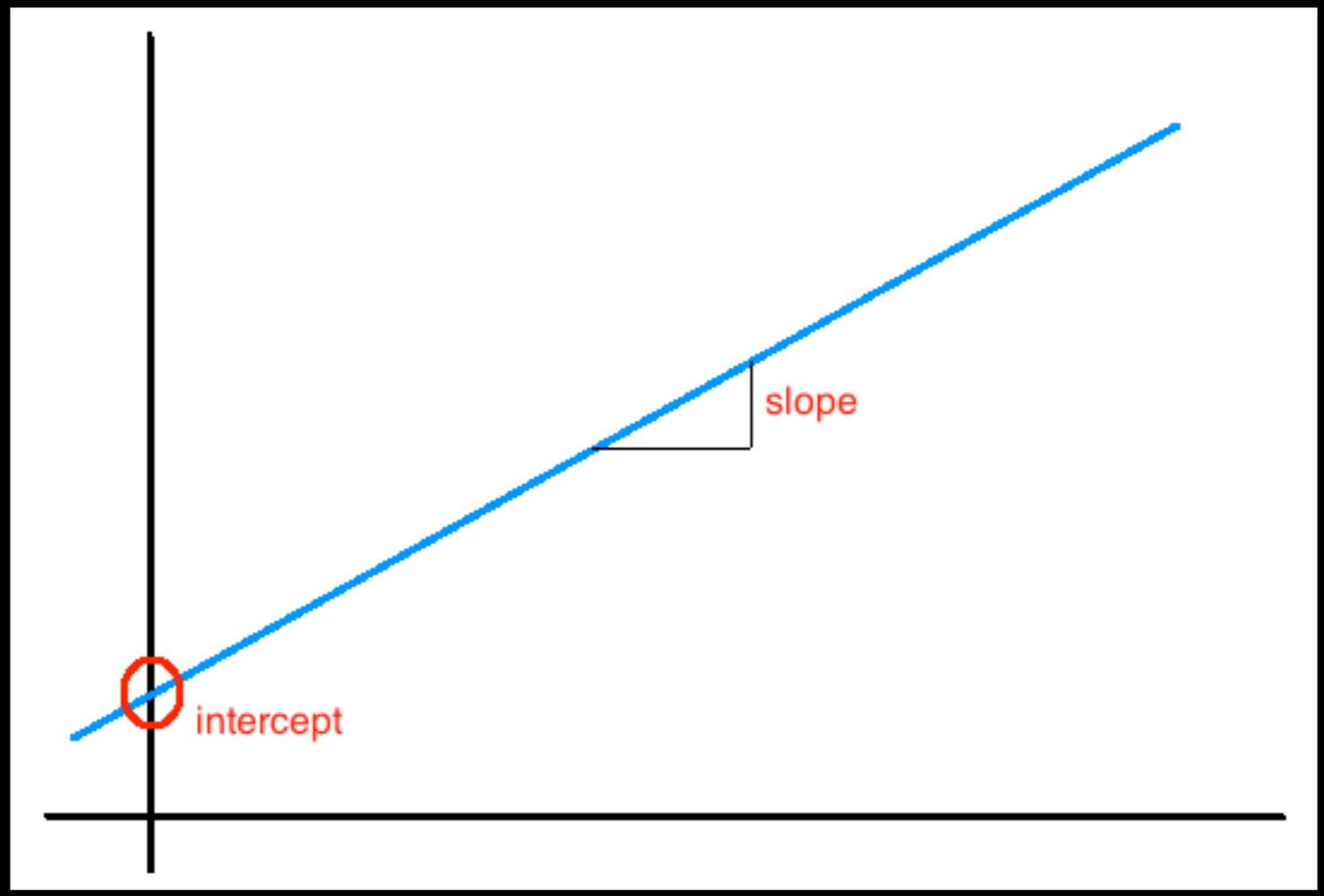
Interpretation of slope and intercept

Slope

For each unit in x, y is expected to increase / decrease on average by the slope.

Intercept

When x = 0, y is expected to equal the intercept.

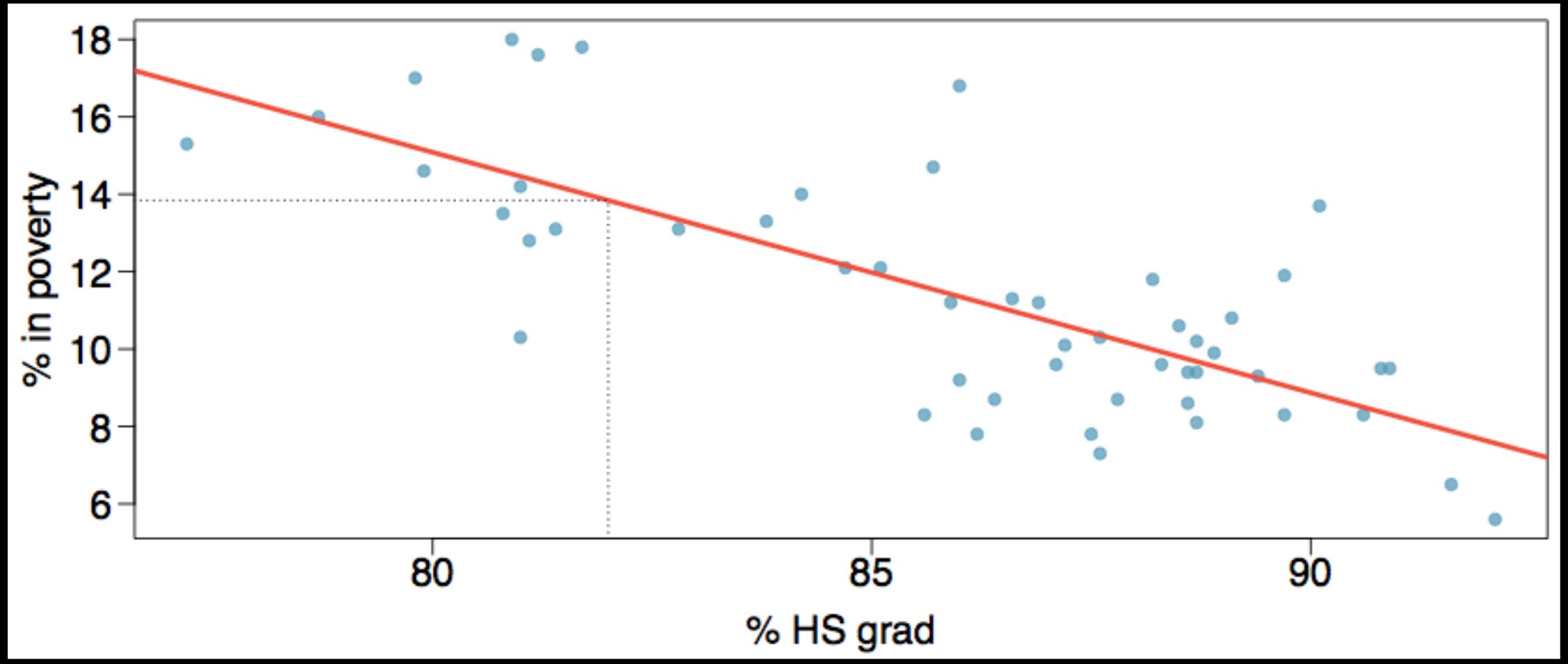


Note: These statements are not causal, unless the study is a randomized controlled experiment.

Prediction

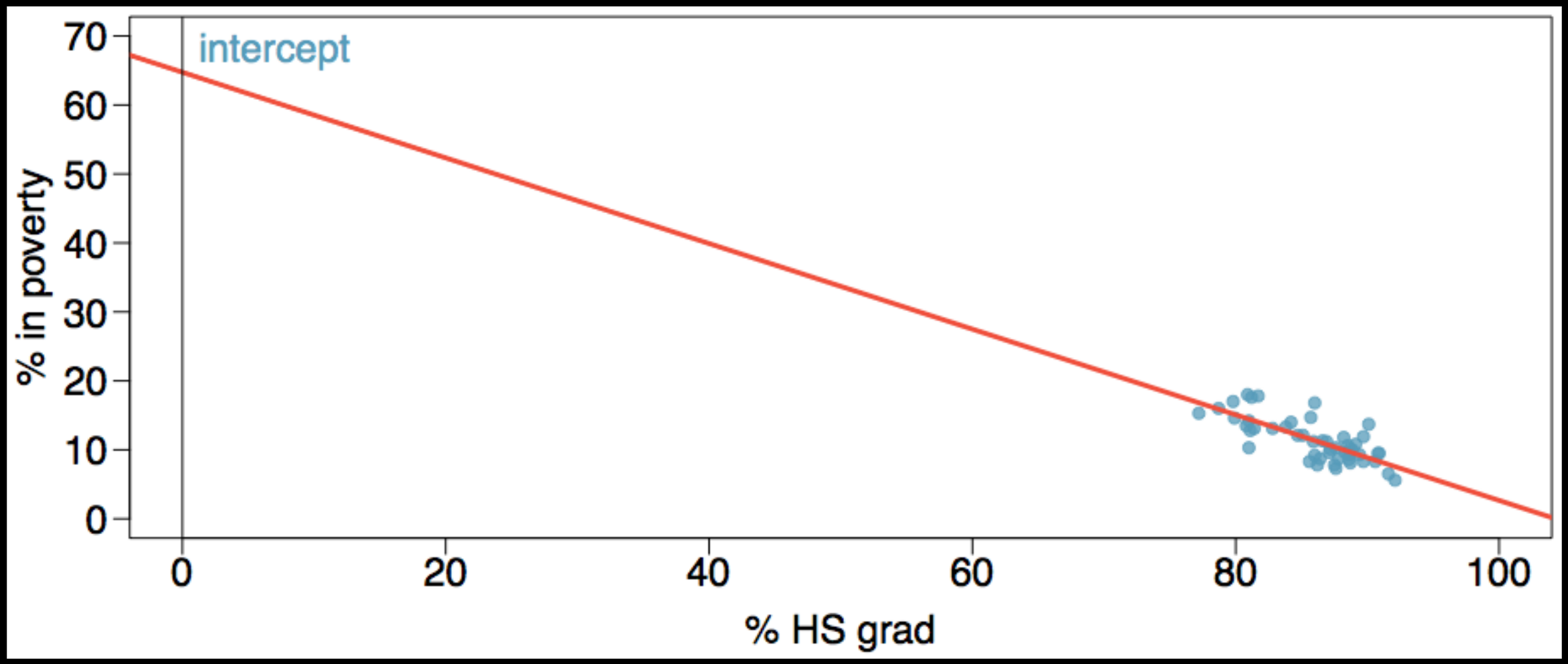
Using the linear model to predict the value of the response variable for a given value of the explanatory variable is called prediction (or interpolation), simply by plugging in the value of x in the linear model equation.

There will be some uncertainty associated with the predicted value.

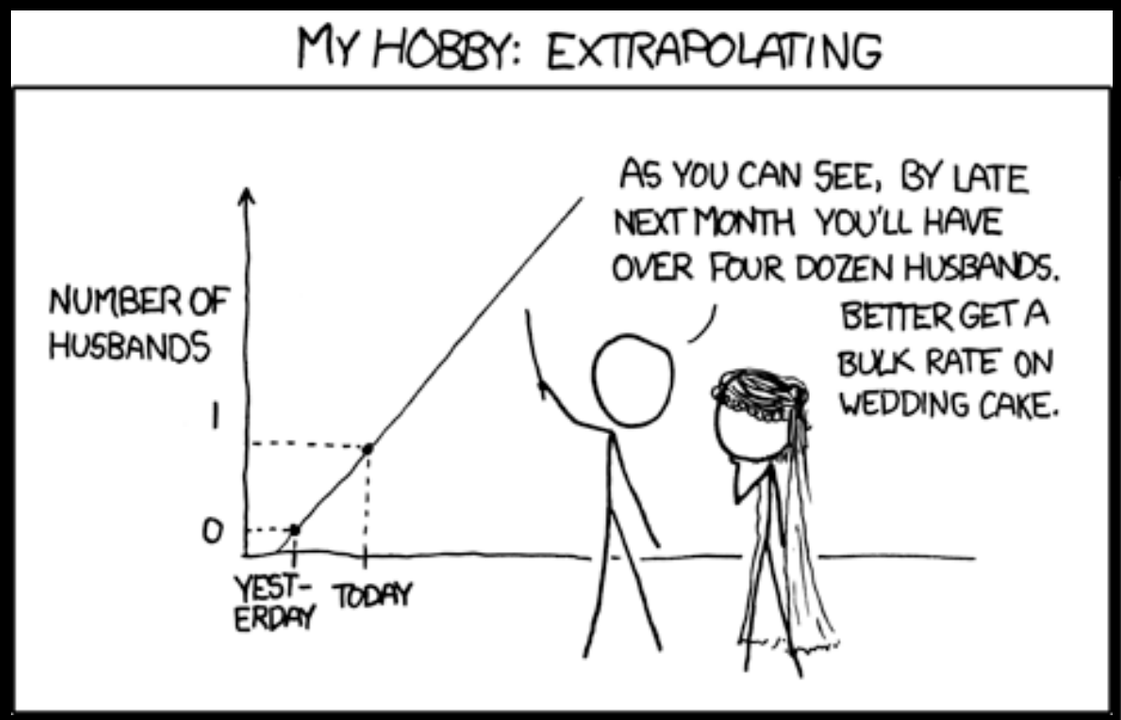


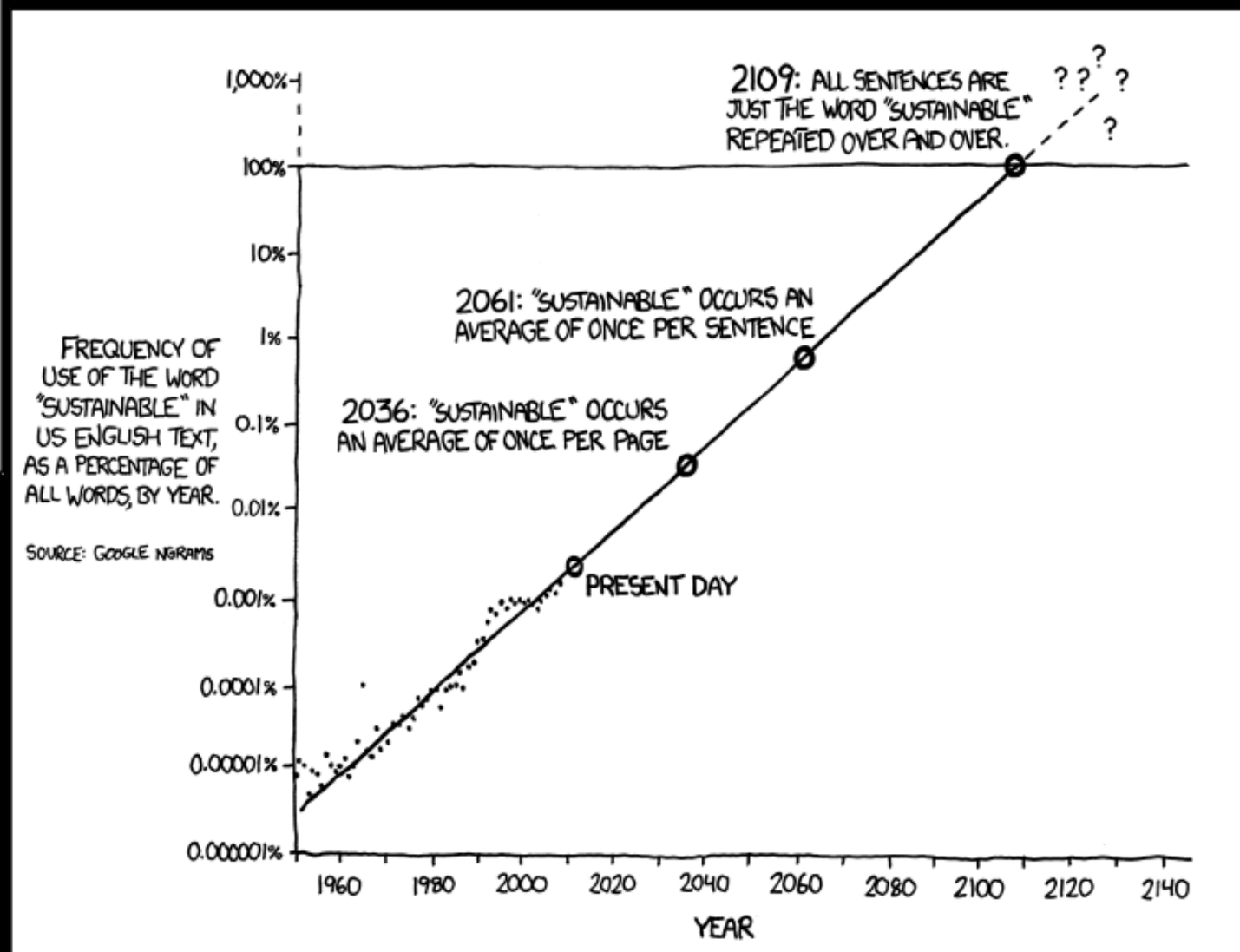
Extrapolation

Applying a model estimate to values outside of the realm of the original data is called extrapolation.
Sometimes the intercept might be an extrapolation.

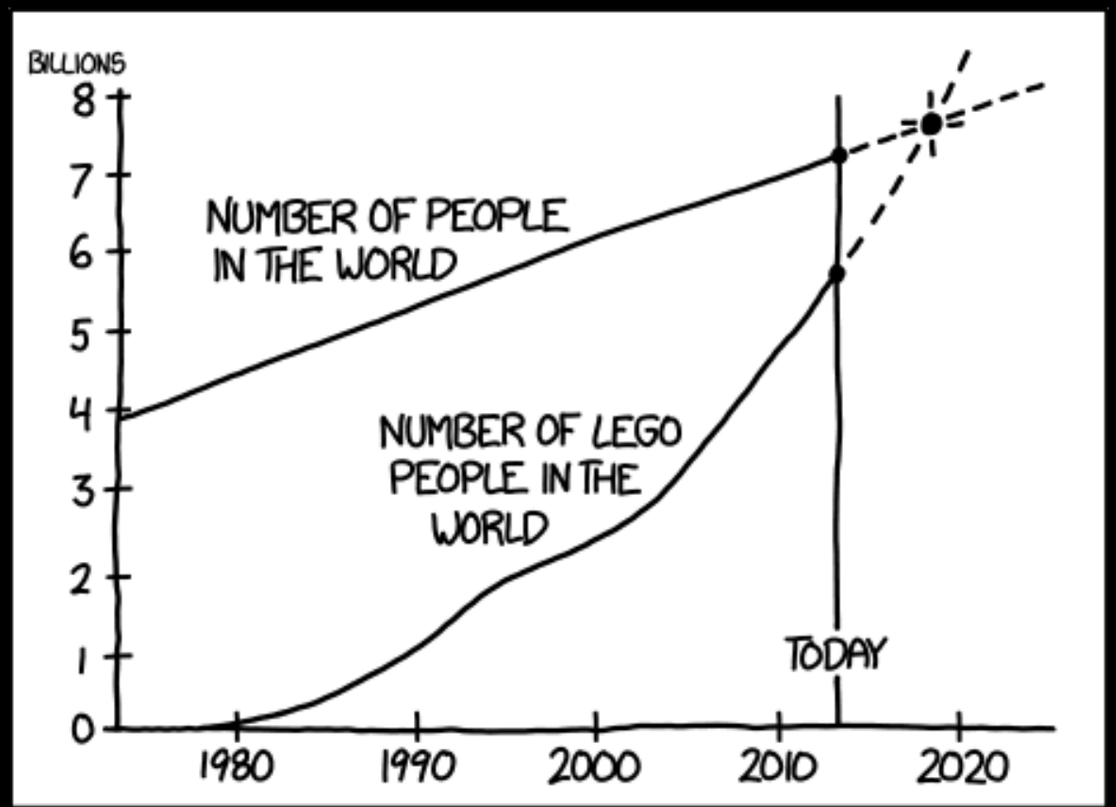


Extrapolation: Here there be Dragons





THE WORD "SUSTAINABLE" IS UNSUSTAINABLE.



BY 2019, HUMANS WILL BE OUTNUMBERED.

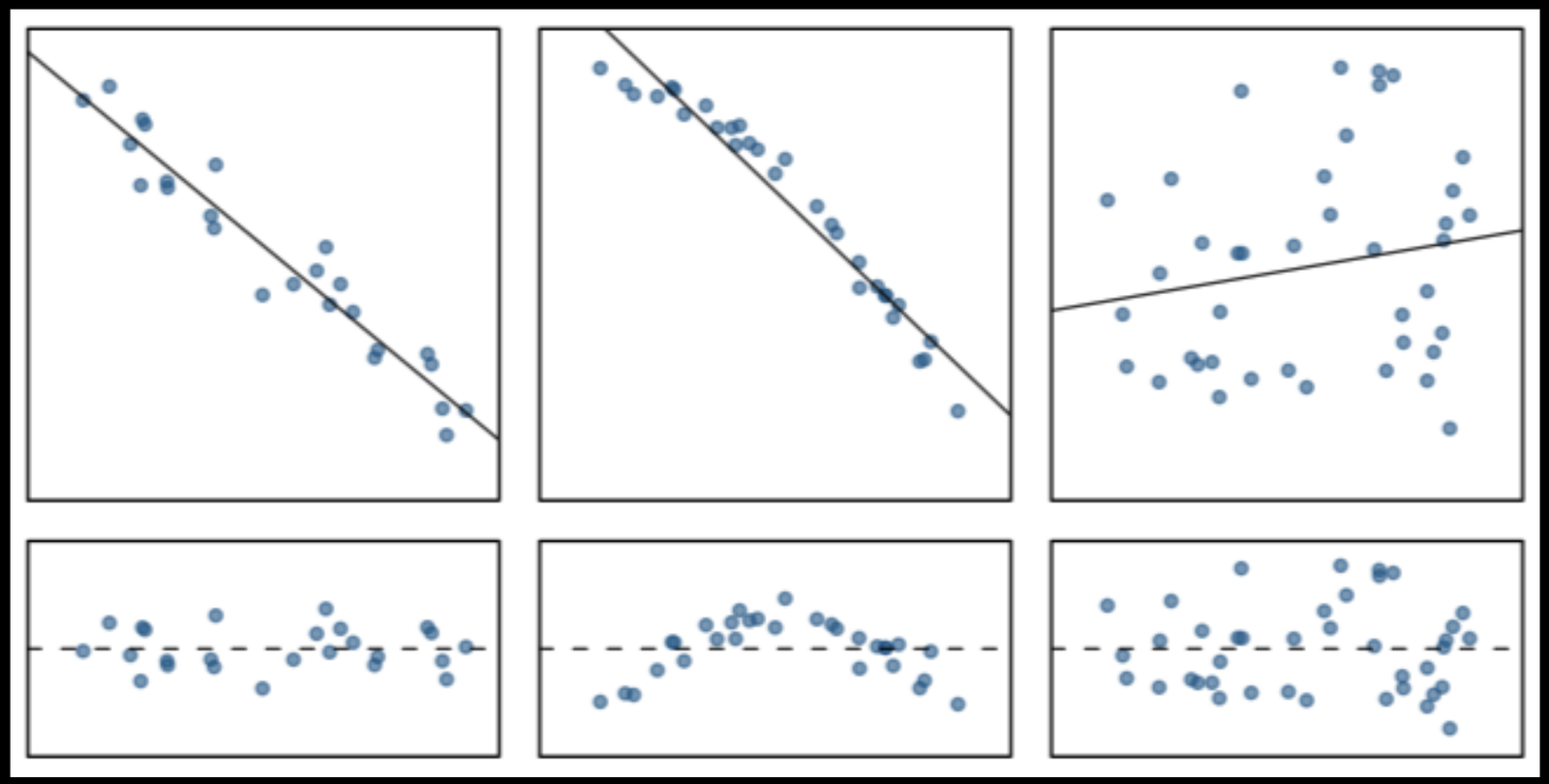
www.xkcd.com

Conditions to using Linear Regression on Data

Conditions: (1) Linearity

The relationship between the explanatory and the response variable should be linear.
Methods for fitting a model to non-linear relationships exist, but we will not go into them in detail.

Check using a scatterplot of the data, or a residuals plot.

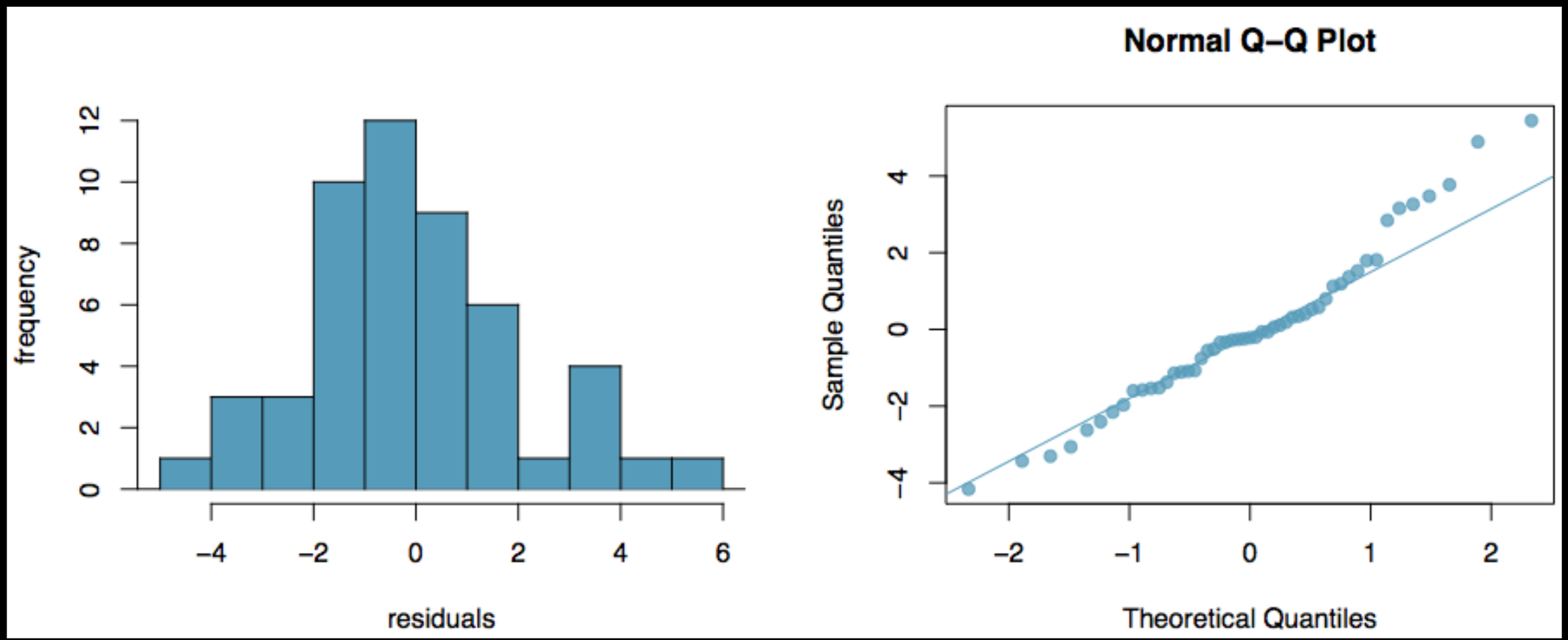


Conditions:
(2) Nearly normal residuals

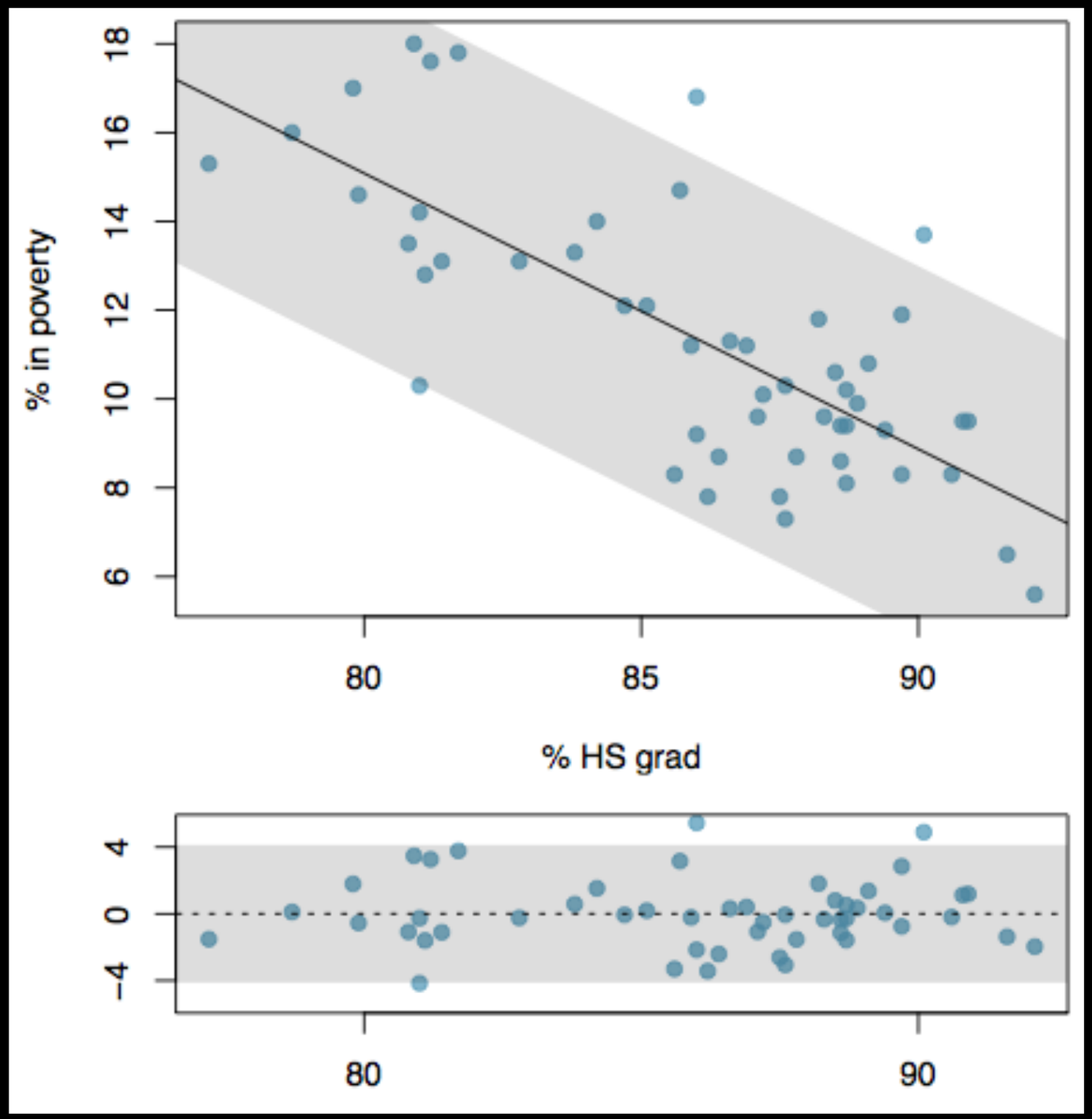
The residuals should be nearly normal.

This condition may not be satisfied when there are unusual observations that don't follow the trend of the rest of the data.

Check using a histogram or normal probability plot of residuals.



Conditions: (3) Constant variability

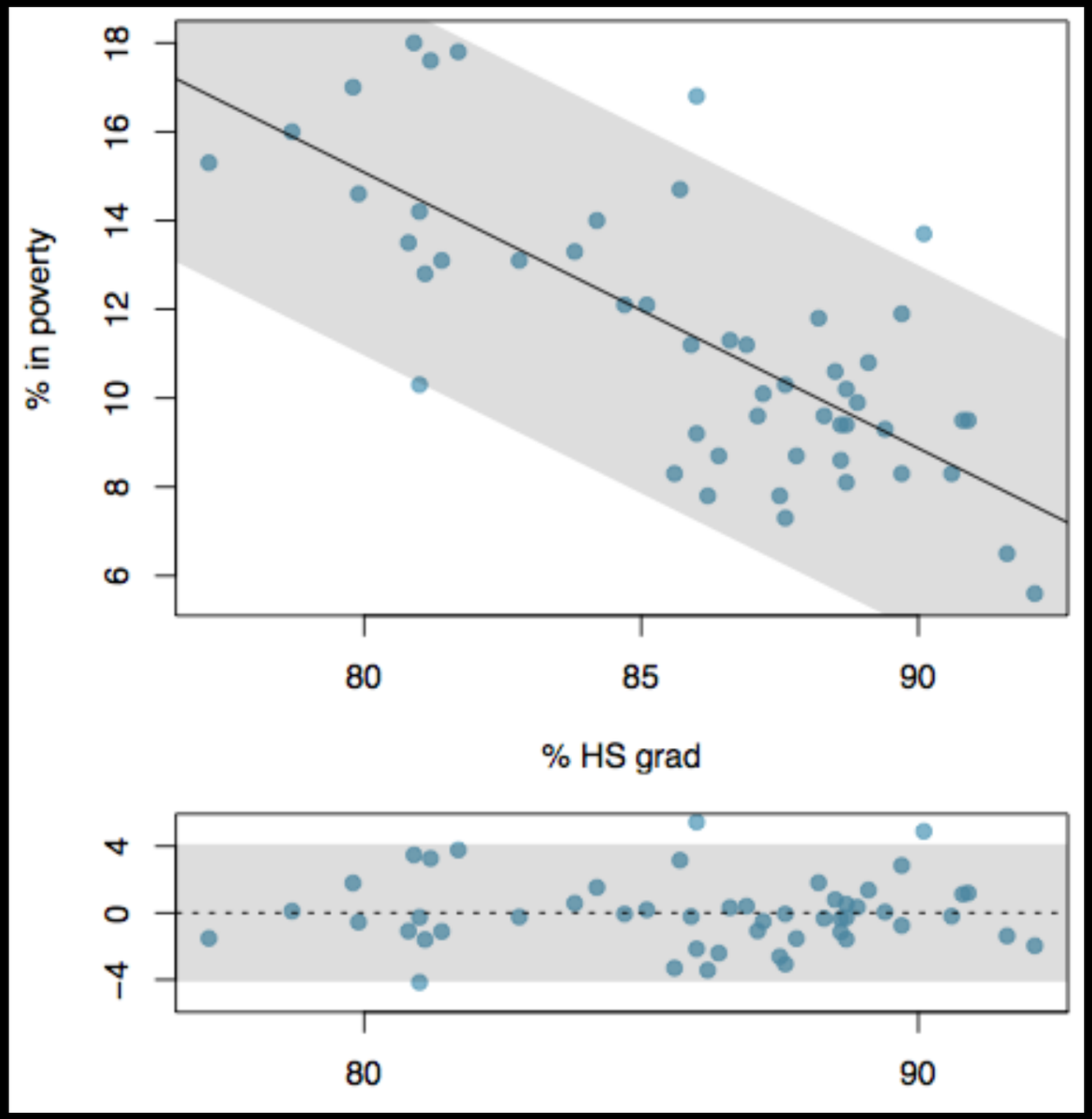


The variability of points around the least squares line should be roughly constant.

This implies that the variability of residuals around the 0 line should be roughly constant as well.

Also called homoscedasticity. Check using a histogram or normal probability plot of residuals.

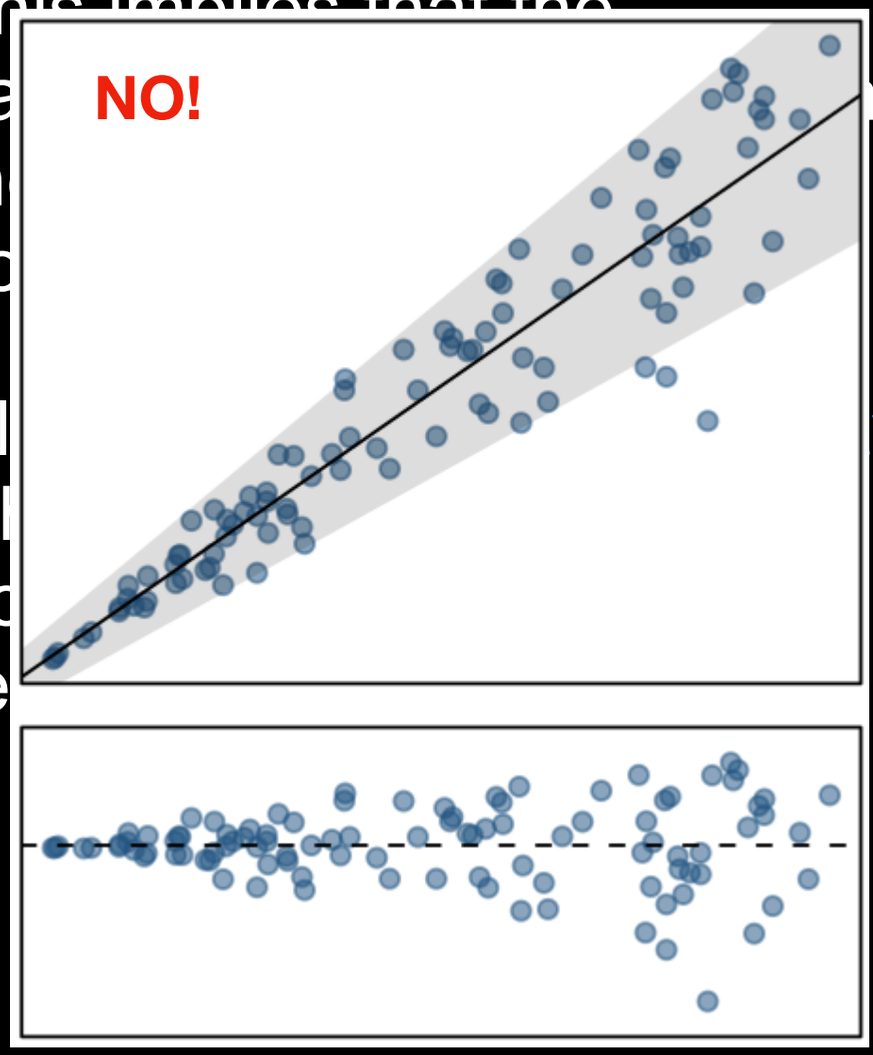
Conditions: (3) Constant variability



The variability of points around the least squares line should be roughly constant.

This implies that the va d th co

Al y. Cl no re



Conditions: (4) Independence

Random sampling from the population?

Sequential data?

A measure of strength of our fit: R^2

The strength of the fit of a linear model is most commonly evaluated using R^2.

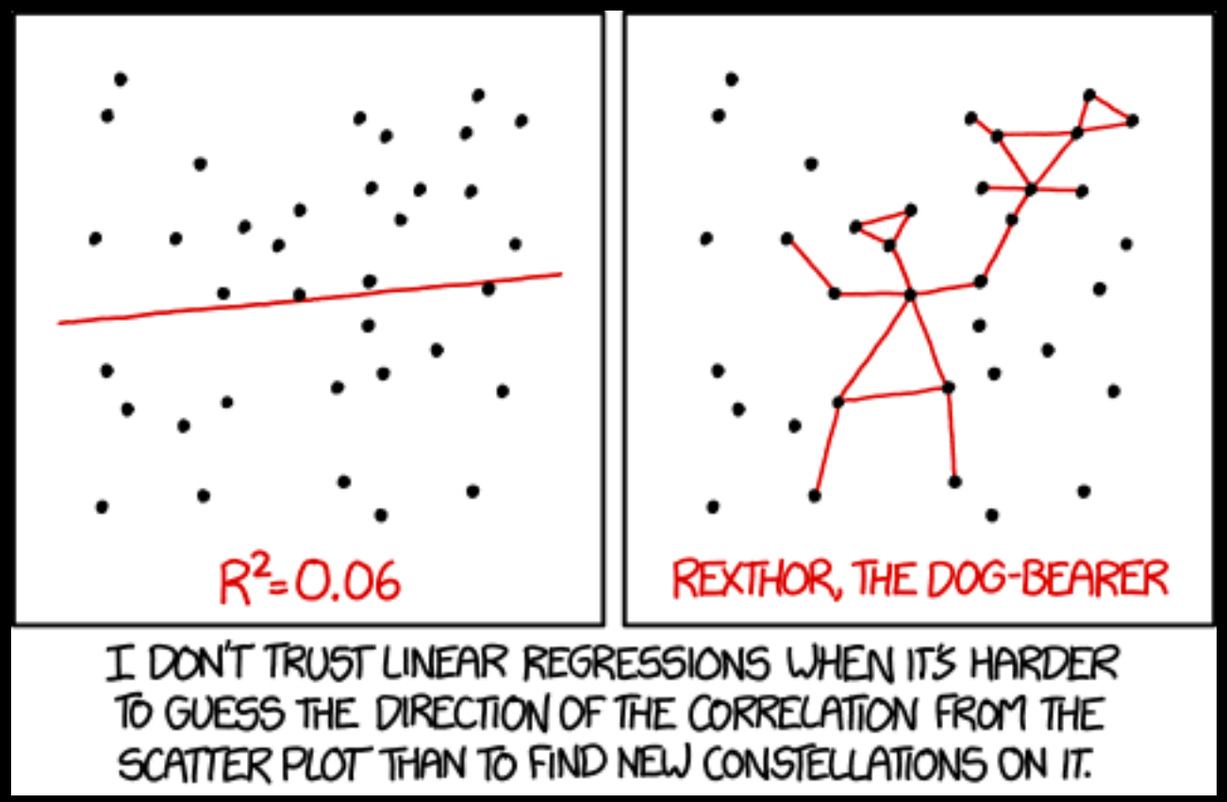
R^2 is calculated as the square of the correlation coefficient.

It tells us what percent of variability in the response variable is explained by the model.

The remainder of the variability is explained by variables not included in the model or by inherent randomness in the data.

In practice: R^2 is just a measure of how the var(dataset) compares to var(residuals after linear fit).

Interpreting values for R^2 is a matter of practice



www.xkcd.com

How this process works in practice:

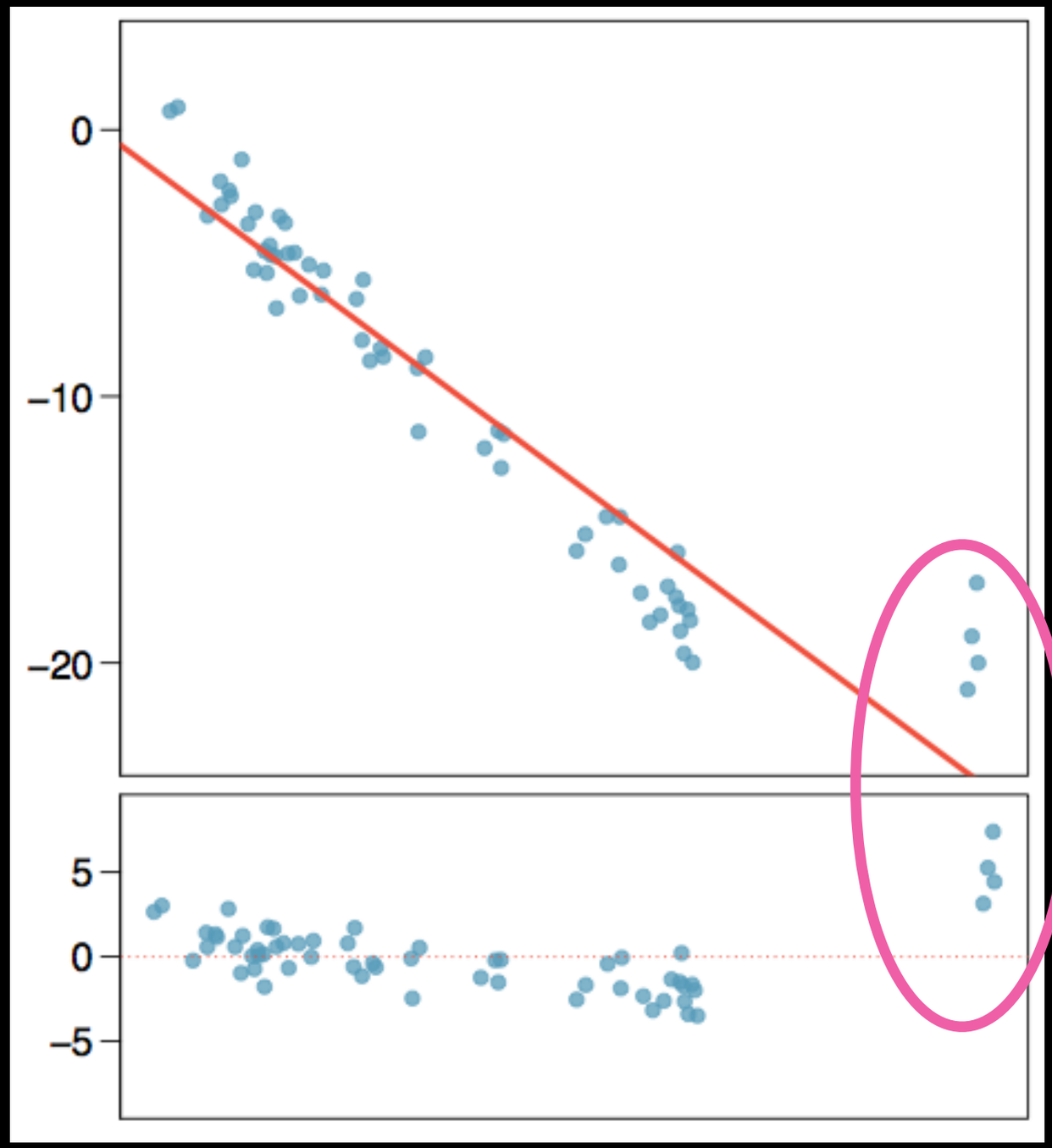
R example

Types of outliers in linear regression

Types of outliers

How do outliers influence the least squares line in this plot?

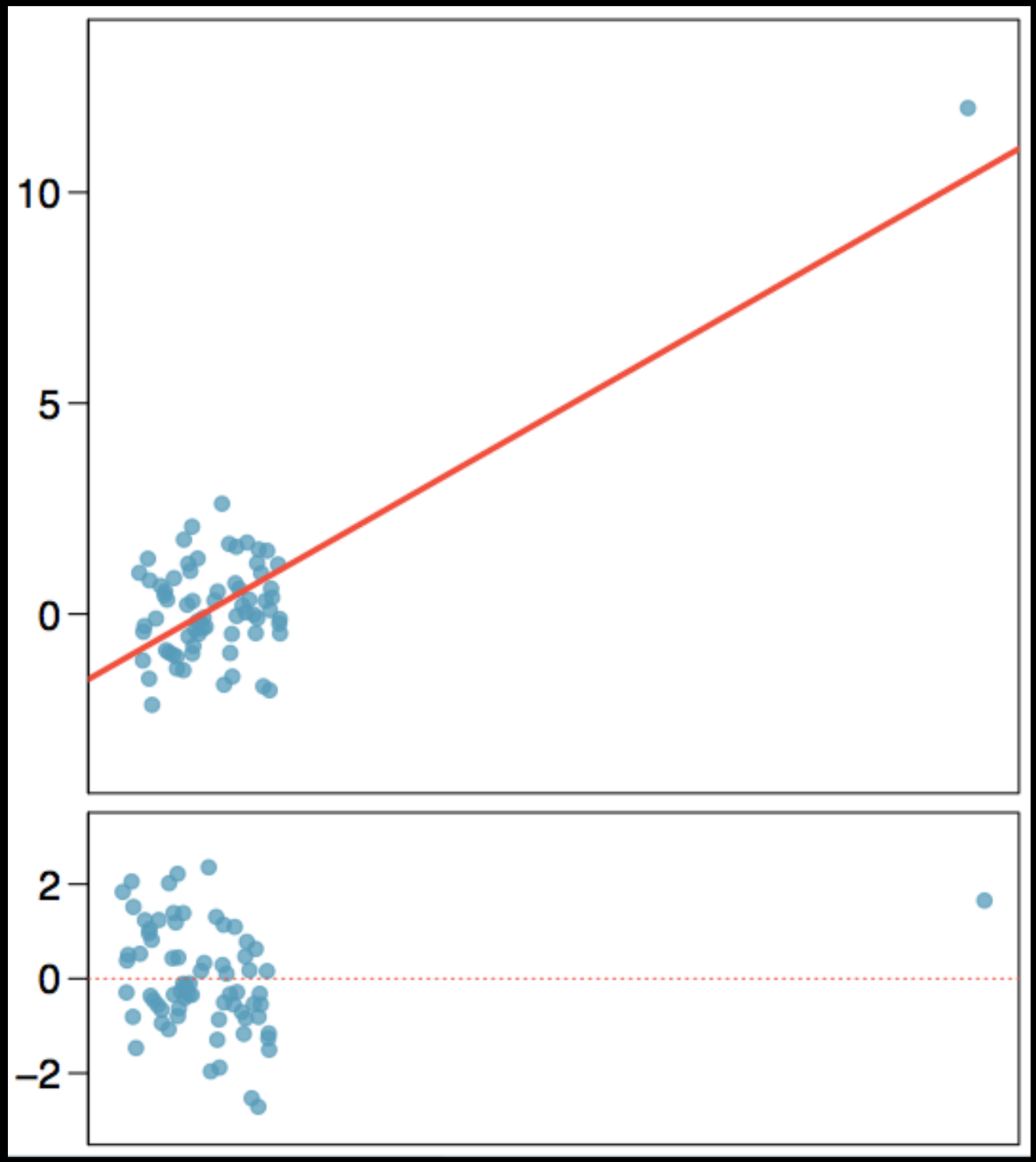
To answer this question think of where the regression line would be with and without the outlier(s). Without the outliers the regression line would be steeper, and lie closer to the larger group of observations. With the outliers the line is pulled up and away from some of the observations in the larger group.



Types of outliers

How do outliers influence the least squares line in this plot?

Without the outlier there is no evident relationship between x and y.



Some terminology

Outliers are points that lie away from the cloud of points.

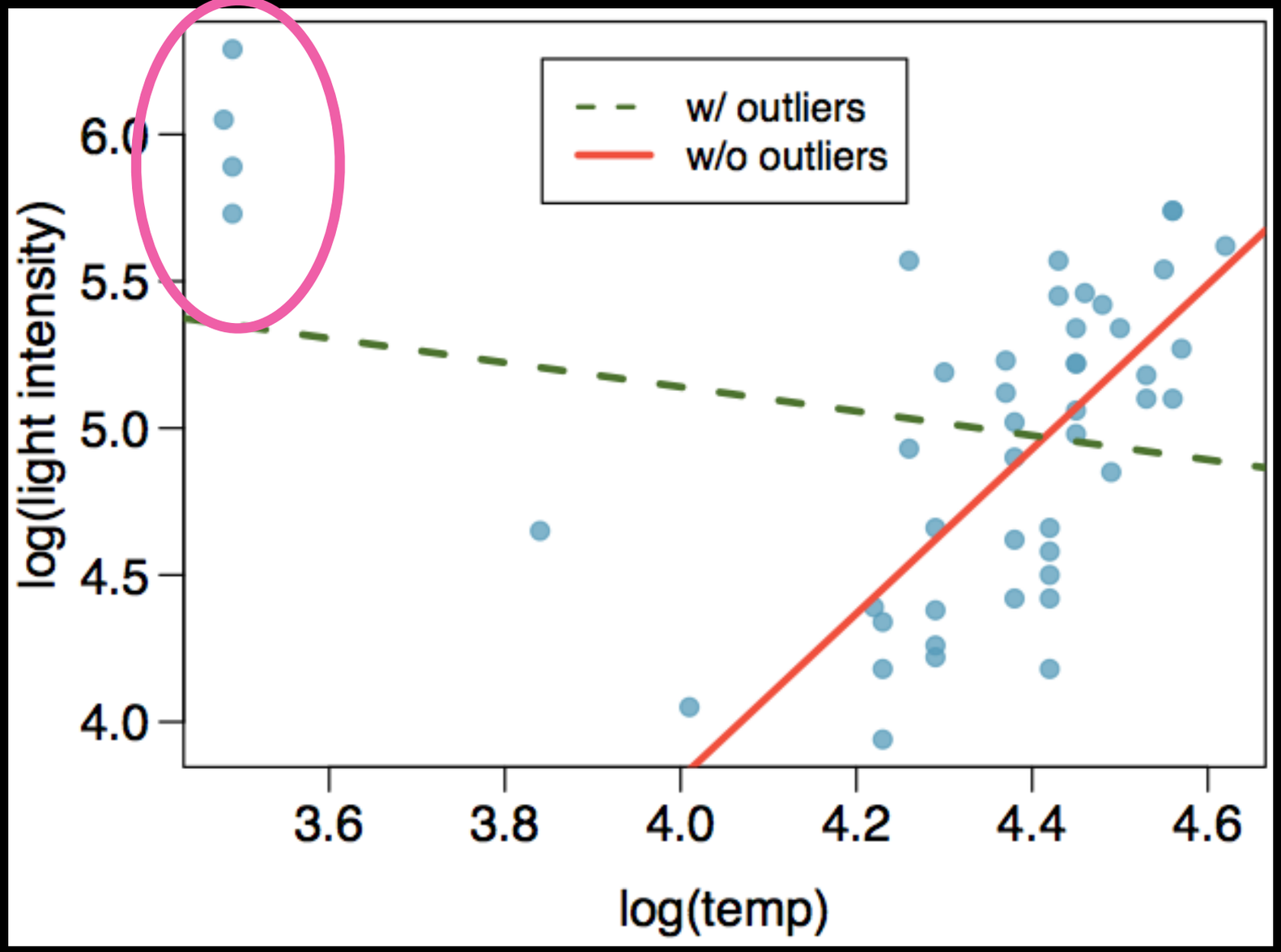
Outliers that lie horizontally away from the center of the cloud are called high leverage points.

High leverage points that actually influence the slope of the regression line are called influential points.

In order to determine if a point is influential, visualize the regression line with and without the point. Does the slope of the line change considerably? If so, then the point is influential. If not, then it's not an influential point.

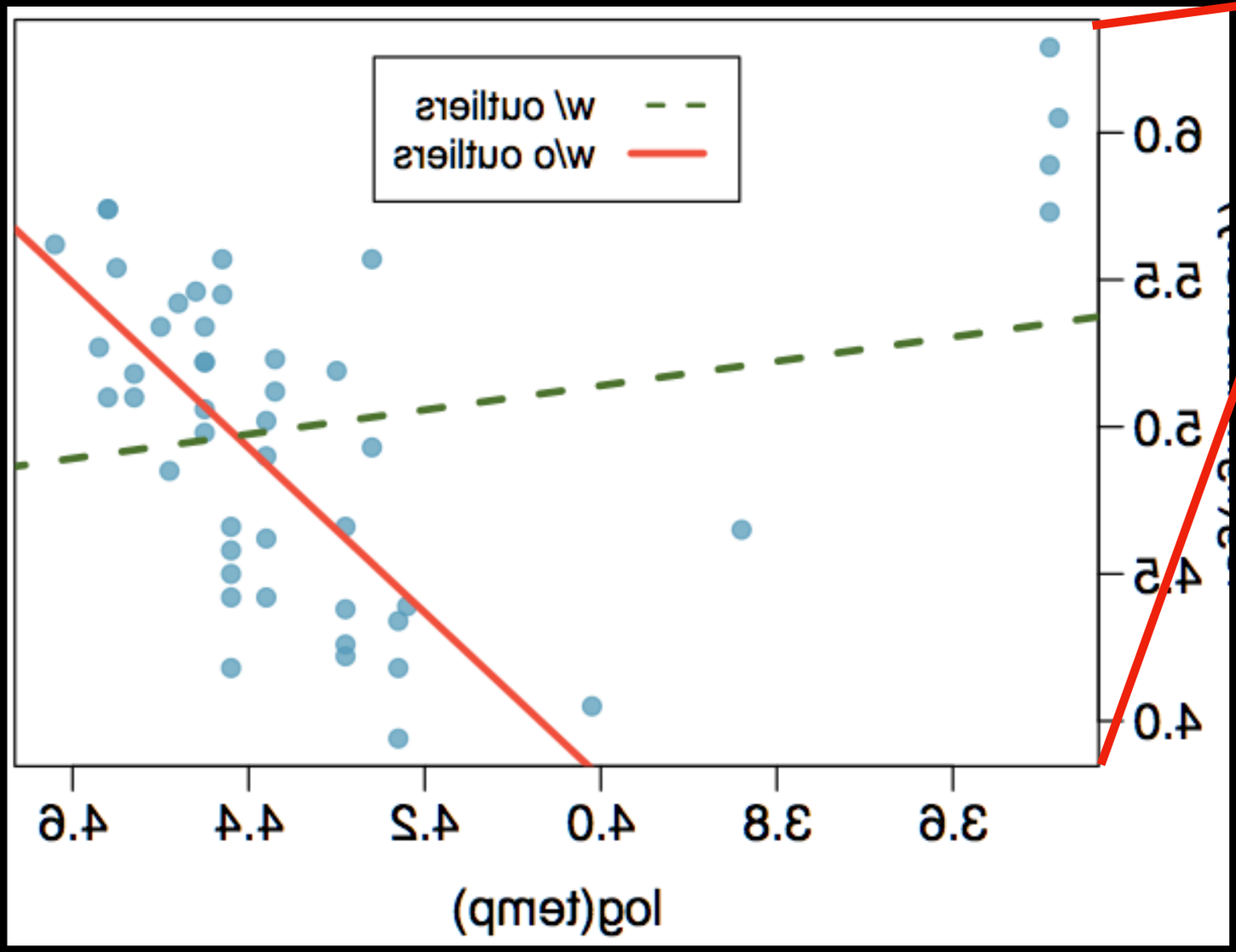
Influential points

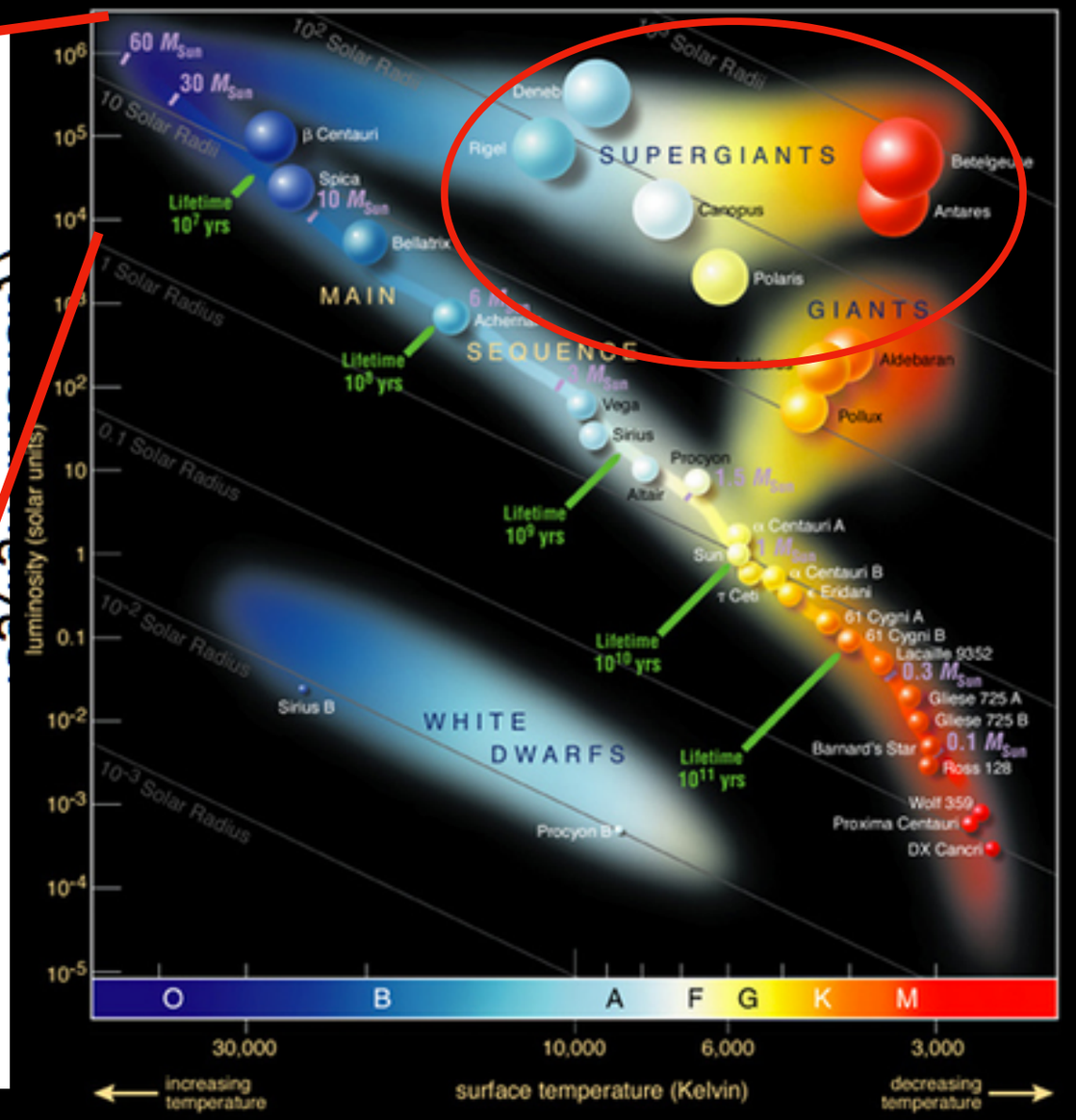
Data are available on the log of the surface temperature and the log of the light intensity of 47 stars in the star cluster CYG OB1.



Influential points

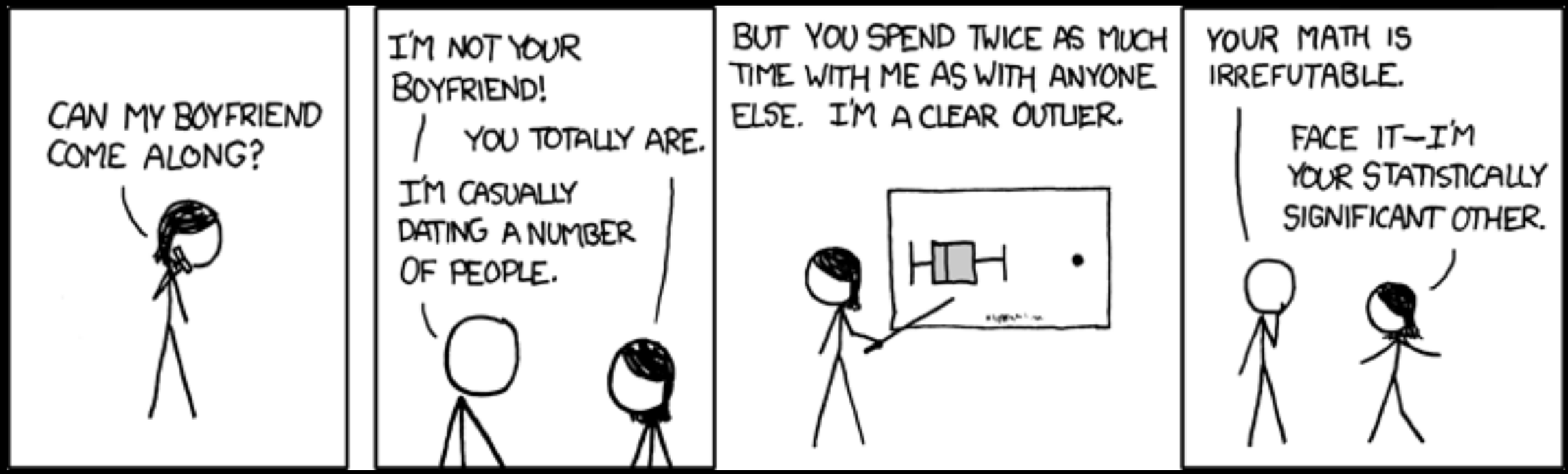
Data are available on the log of the surface temperature and the log of the light intensity of 47 stars in the star cluster CYG OB1.





Be wary of throwing out outliers…

Depending on the application, outliers might be the most interesting points!



Looking at outliers with R

Inference for Linear Regression

p-values for Linear Regression

What's really going on here? Just the same calculations we've been doing the past few weeks!

p-value < 0.05
so we can reject H_0

```
> summary(myLine)

Call:
lm(formula = BAC ~ Beers, data = BB)

Residuals:
      Min        1Q    Median        3Q       Max
-0.027118 -0.017350  0.001773  0.008623  0.041027

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.012701   0.012638  -1.005    0.332
Beers        0.017964   0.002402   7.480 2.97e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02044 on 14 degrees of freedom
Multiple R-squared:  0.7998,    Adjusted R-squared:  0.7855
F-statistic: 55.94 on 1 and 14 DF,  p-value: 2.969e-06
```

H_0: There is no relation between Beers and BAC - slope = 0
H_A: There is a relationship between Beers and BAC - slope != 0

p-values for Linear Regression

What's really going on here? Just the same calculations we've been doing the past few weeks!

```
> summary(myLine)

Call:
lm(formula = BAC ~ Beers, data = BB)

Residuals:
      Min        1Q    Median        3Q       Max
-0.027118 -0.017350  0.001773  0.008623  0.041027

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.012701   0.012638  -1.005    0.332
Beers        0.017964   0.002402   7.480 2.97e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02044 on 14 degrees of freedom
Multiple R-squared:  0.7998,    Adjusted R-squared:  0.7855
F-statistic: 55.94 on 1 and 14 DF,  p-value: 2.969e-06
```

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

H_0: There is no relation between Beers and BAC - slope = 0
H_A: There is a relationship between Beers and BAC - slope != 0

p-values for Linear Regression

What's really going on here? Just the same calculations we've been doing the past few weeks!

```
> summary(myLine)

Call:
lm(formula = BAC ~ Beers, data = BB)

Residuals:
      Min        1Q    Median        3Q       Max
-0.027118 -0.017350  0.001773  0.008623  0.041027

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.012701   0.012638  -1.005    0.332
Beers        0.017964   0.002402   7.480 2.97e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02044 on 14 degrees of freedom
Multiple R-squared:  0.7998,    Adjusted R-squared:  0.7855
F-statistic: 55.94 on 1 and 14 DF,  p-value: 2.969e-06
```

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

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

how different are the individual variances from the common variance?

H_0: There is no relation between Beers and BAC - slope = 0
H_A: There is a relationship between Beers and BAC - slope != 0

p-values for Linear Regression

What's really going on here? Just the same calculations we've been doing the past few weeks!

```
> summary(myLine)

Call:
lm(formula = BAC ~ Beers, data = BB)

Residuals:
      Min        1Q    Median        3Q       Max
-0.027118 -0.017350  0.001773  0.008623  0.041027

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.012701   0.012638  -1.005    0.332
Beers        0.017964   0.002402   7.480 2.97e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02044 on 14 degrees of freedom
Multiple R-squared:  0.7998,    Adjusted R-squared:  0.7855
F-statistic: 55.94 on 1 and 14 DF,  p-value: 2.969e-06
```

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

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

how different are the individual variances from the common variance?

H_0: There is no relation between Beers and BAC - slope = 0
H_A: There is a relationship between Beers and BAC - slope != 0

p-values for Linear Regression

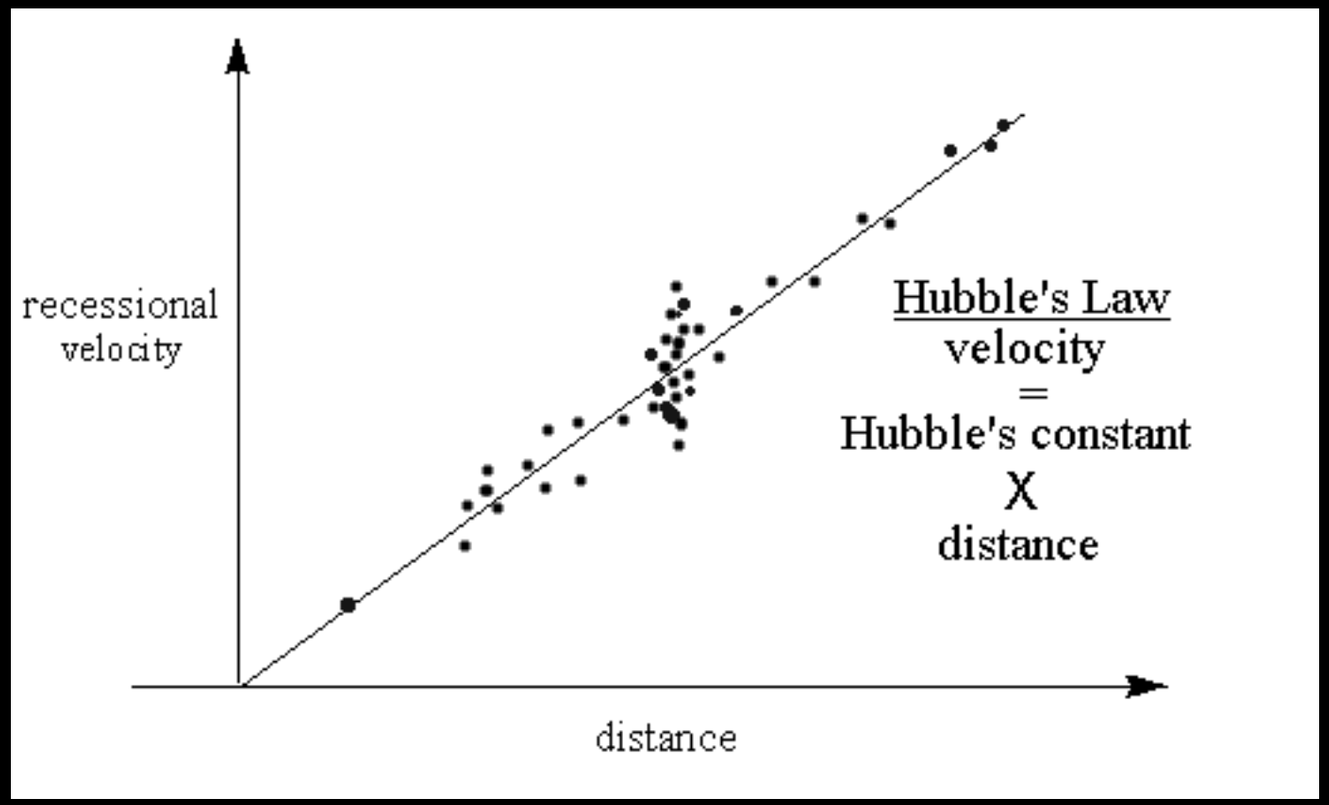
Caution: Don't carelessly use the p-value from regression output

The last column in regression output often lists p-values for one particular hypothesis: a two-sided test where the null value is zero. If your test is one-sided and the point estimate is in the direction of H_A, then you can halve the software's p-value to get the one-tail area. If neither of these scenarios match your hypothesis test, be cautious about using the software output to obtain the p-value.

H_0: There is no relation between Beers and BAC - slope = 0
H_A: There is a relationship between Beers and BAC - slope != 0

Put it all together with an example: Hubble's Law

Hubble's Law



In essence: galaxies that are further away from us look like they are moving away from us faster than those near by

Hubble's Law

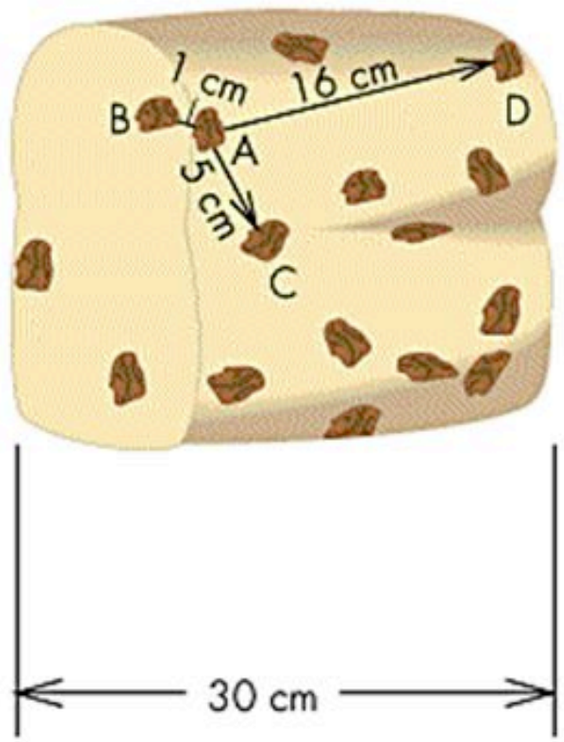
Hubble Expansion

Hubble's Law can be thought of two ways:

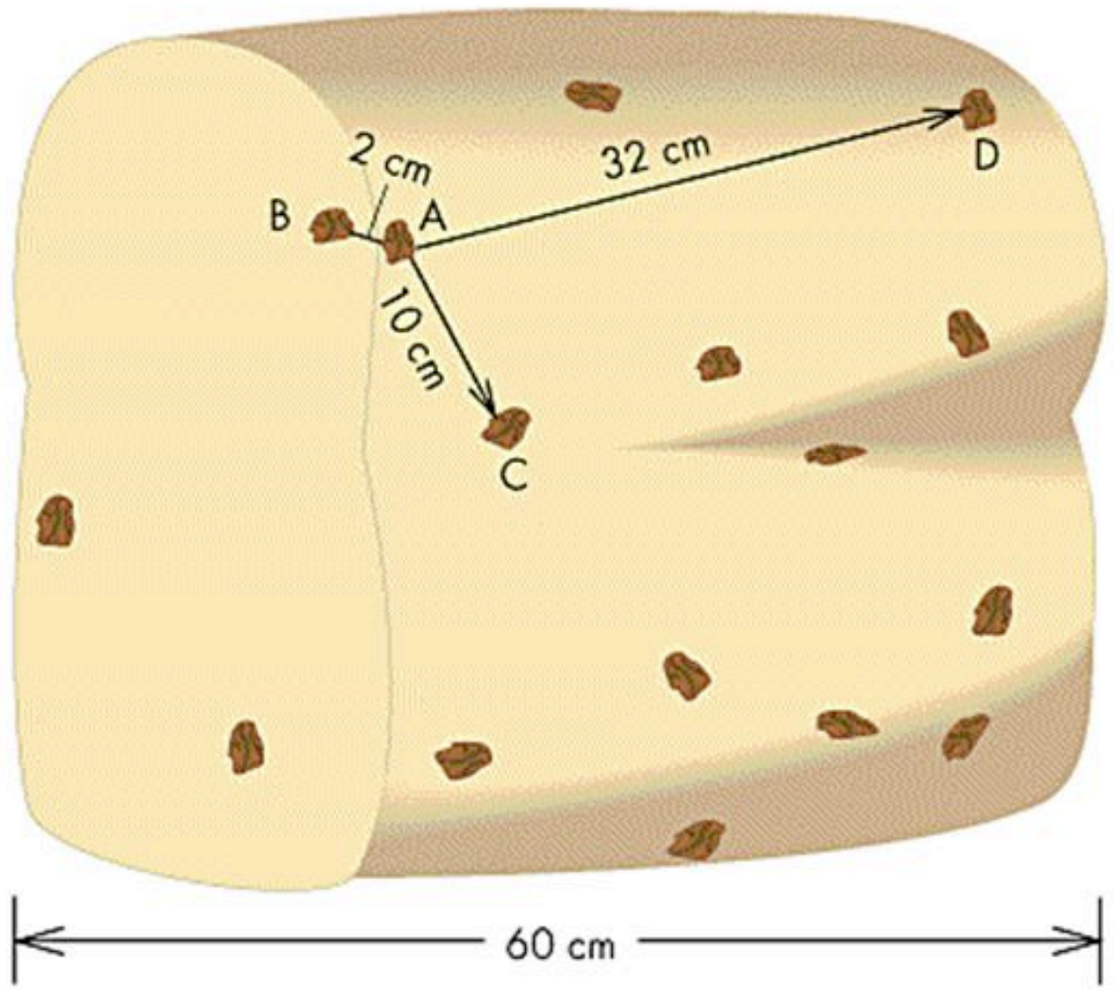
- All galaxies are flying apart from each other
- The space between the galaxies is expanding

There is no special place in the universe

- It is meaningless to ask "where is it expanding from"
- All observers see the same thing



A Raisin bread dough before rising



B Raisin bread dough after rising

https://slideplayer.com/user/3798162/

Hubble's Law in R!

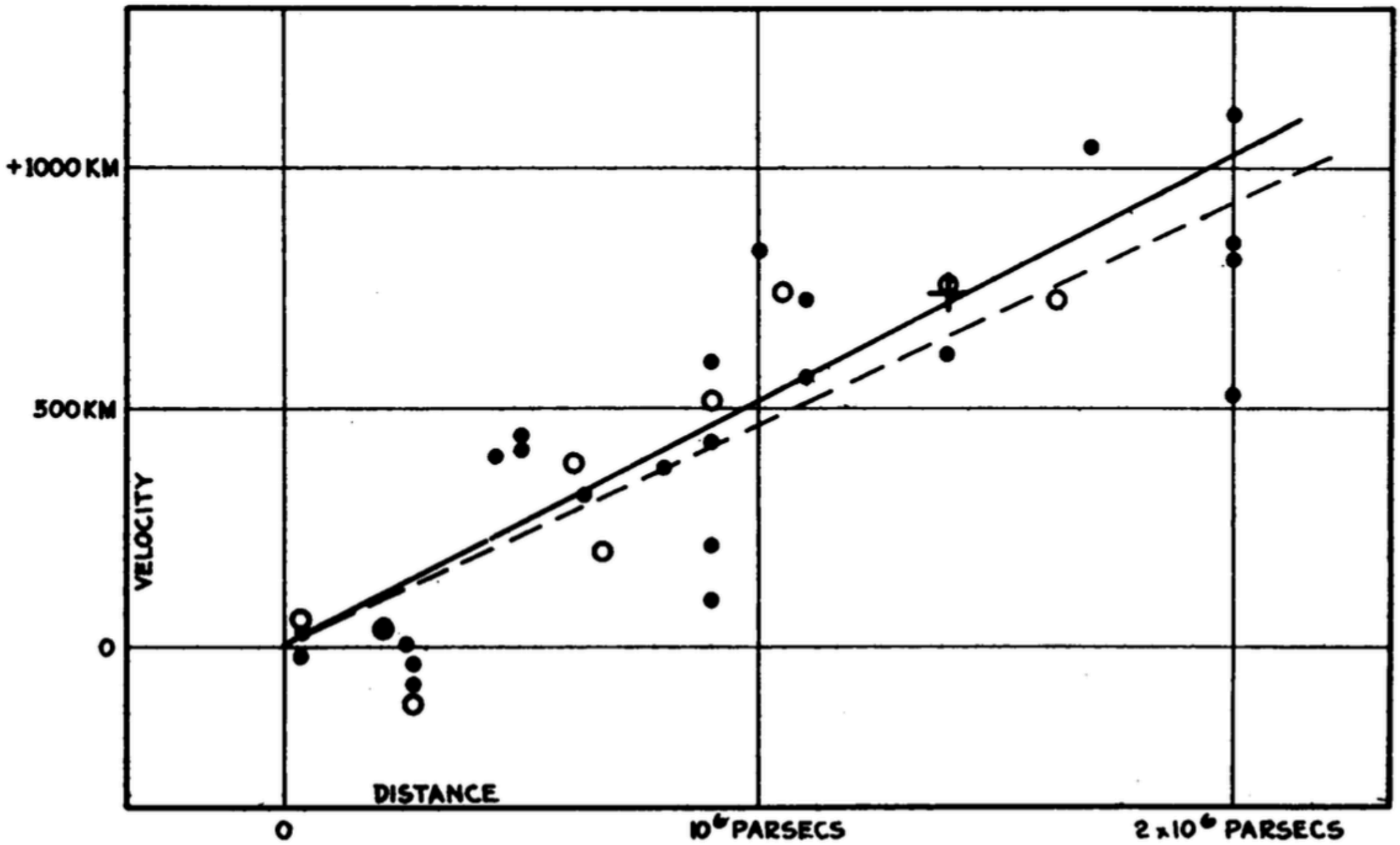


FIGURE 1

Velocity-Distance Relation among Extra-Galactic Nebulae.

Hubble 1929