Misconceptions, Problems, and Fallacies in Correlational Analysis

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Introduction to Linear Regression

Introduction

In this module, we discuss some common problems and fallacies regarding correlation coefficients and their interpretation:

1. Interpreting a Correlation
2. No Correlation vs. No Relation
3. Perfect Correlation vs. Equivalence
4. Combining Populations, and Ignoring Explanatory Variables
5. Restriction of Range
6. Prediction vs. Explanation: The Shrinkage Problem
7. The Third Variable Fallacy
8. Correlation and Causality

Interpreting a Correlation

If correlations have a familiar and well-behaved distribution, such as the bivariate normal form, then there is a well-defined relationship between the shape of the scatterplot and the correlation coefficient.

Many undergraduate courses begin with a demonstration that connects a series of scatterplots with the corresponding correlations.

Interpreting a Correlation

Some Typical Bivariate Normal Scatterplots

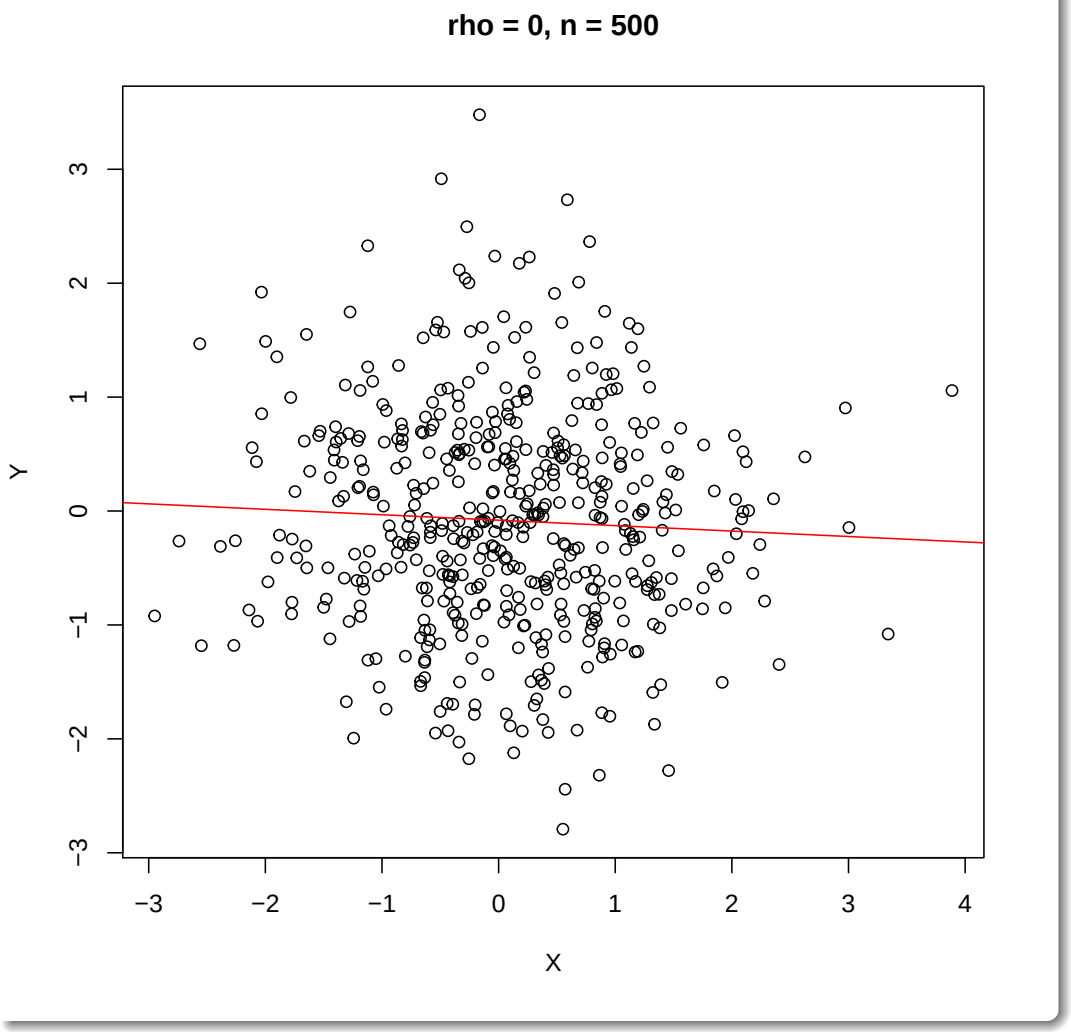
Example (Some Typical Scatterplots)

Let's examine some bivariate normal scatterplots in which the data come from populations with means of 0 and variances of 1. These will give you a feel for how correlations are reflected in a bivariate normal scatterplot.

Interpreting a Correlation

Some Typical Bivariate Normal Scatterplots

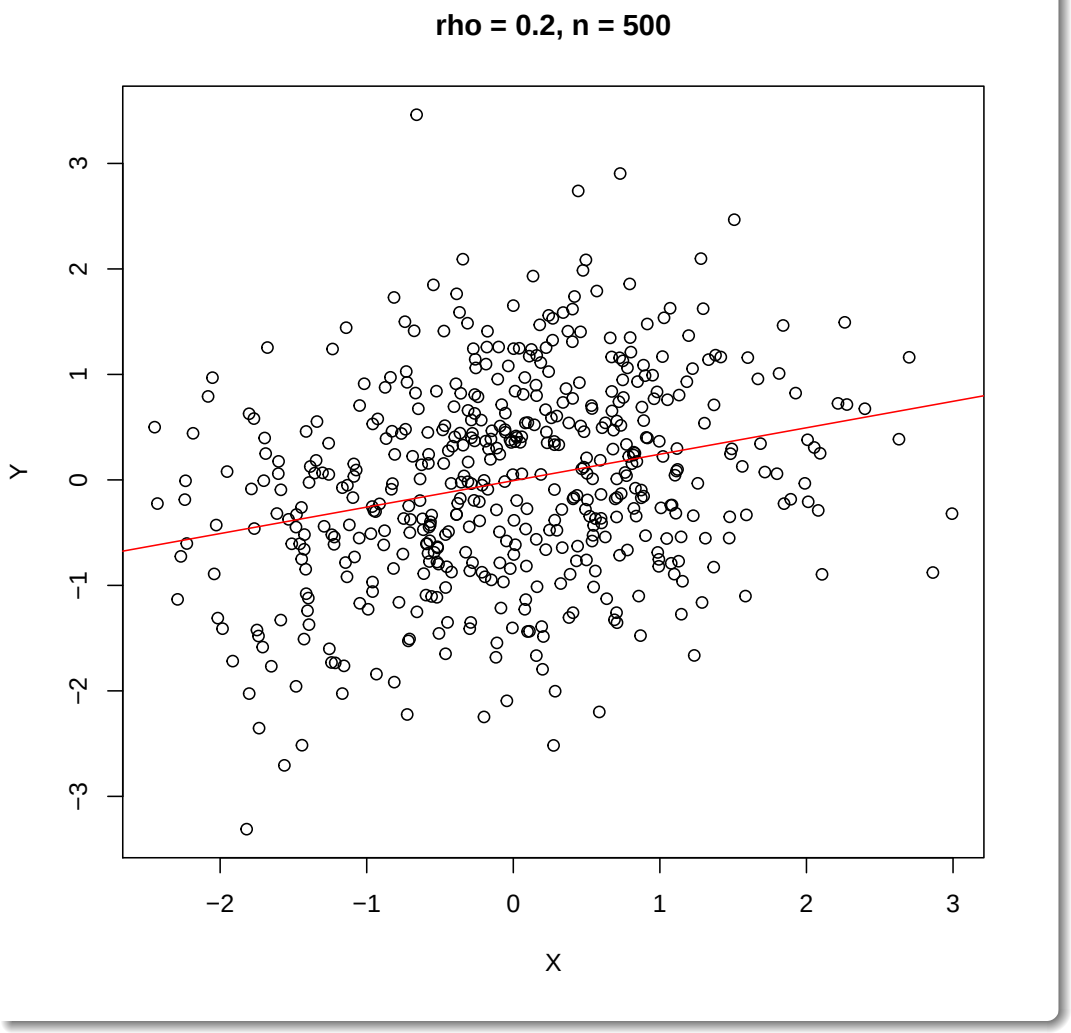
Example (Some Typical Scatterplots)



Interpreting a Correlation

Some Typical Bivariate Normal Scatterplots

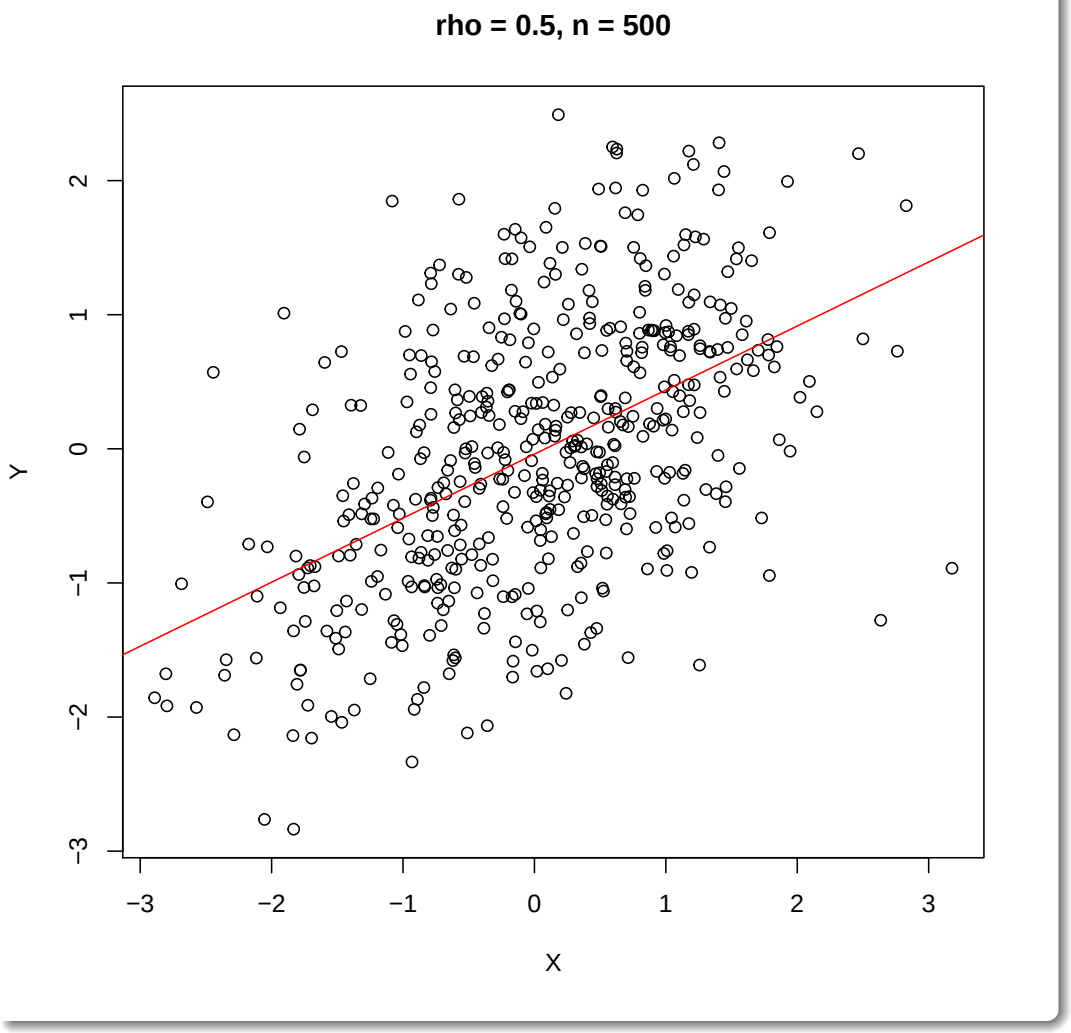
Example (Some Typical Scatterplots)



Interpreting a Correlation

Some Typical Bivariate Normal Scatterplots

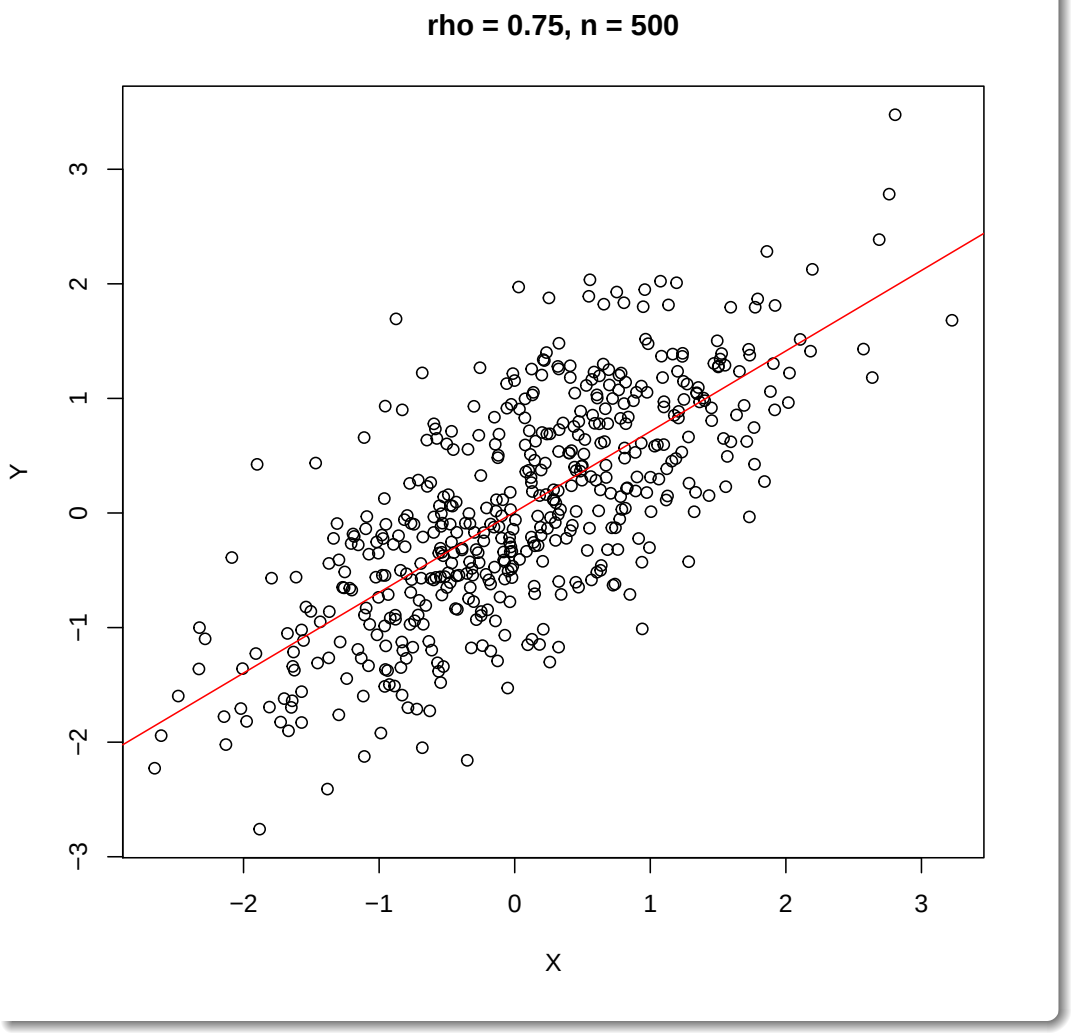
Example (Some Typical Scatterplots)



Interpreting a Correlation

Some Typical Bivariate Normal Scatterplots

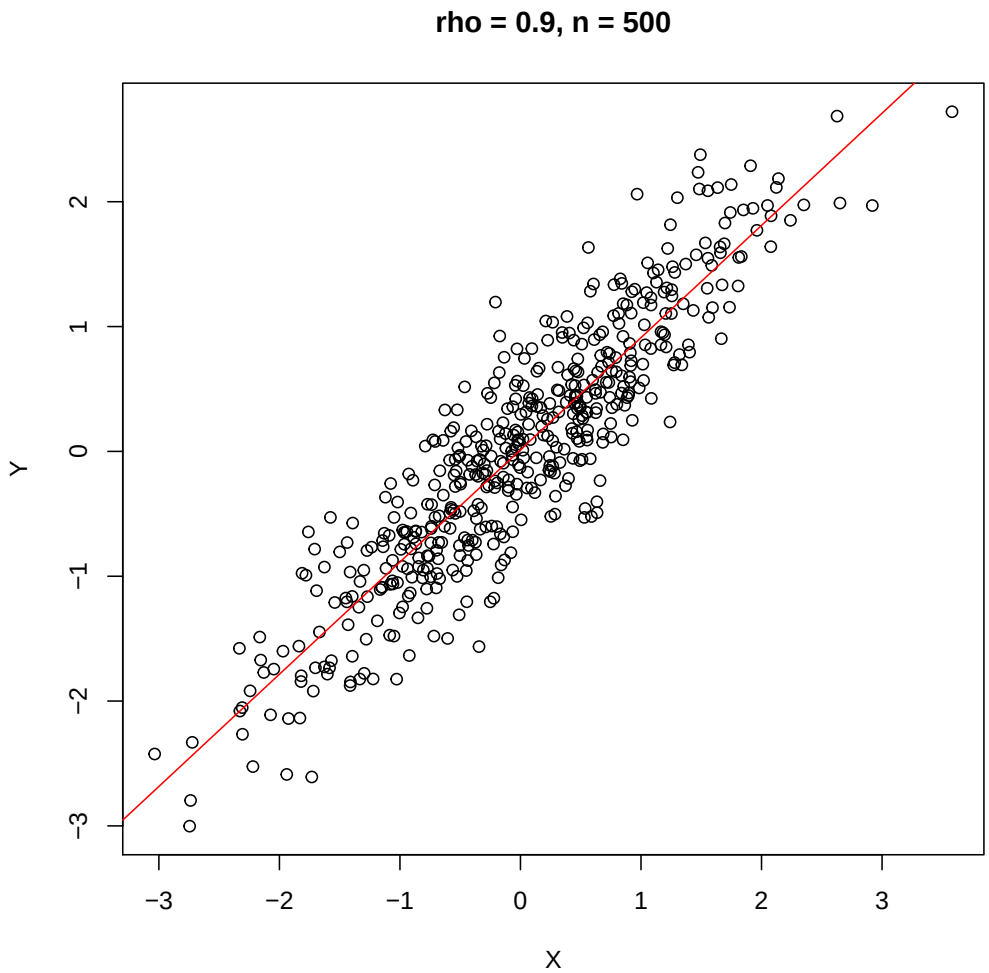
Example (Some Typical Scatterplots)



Interpreting a Correlation

Some Typical Bivariate Normal Scatterplots

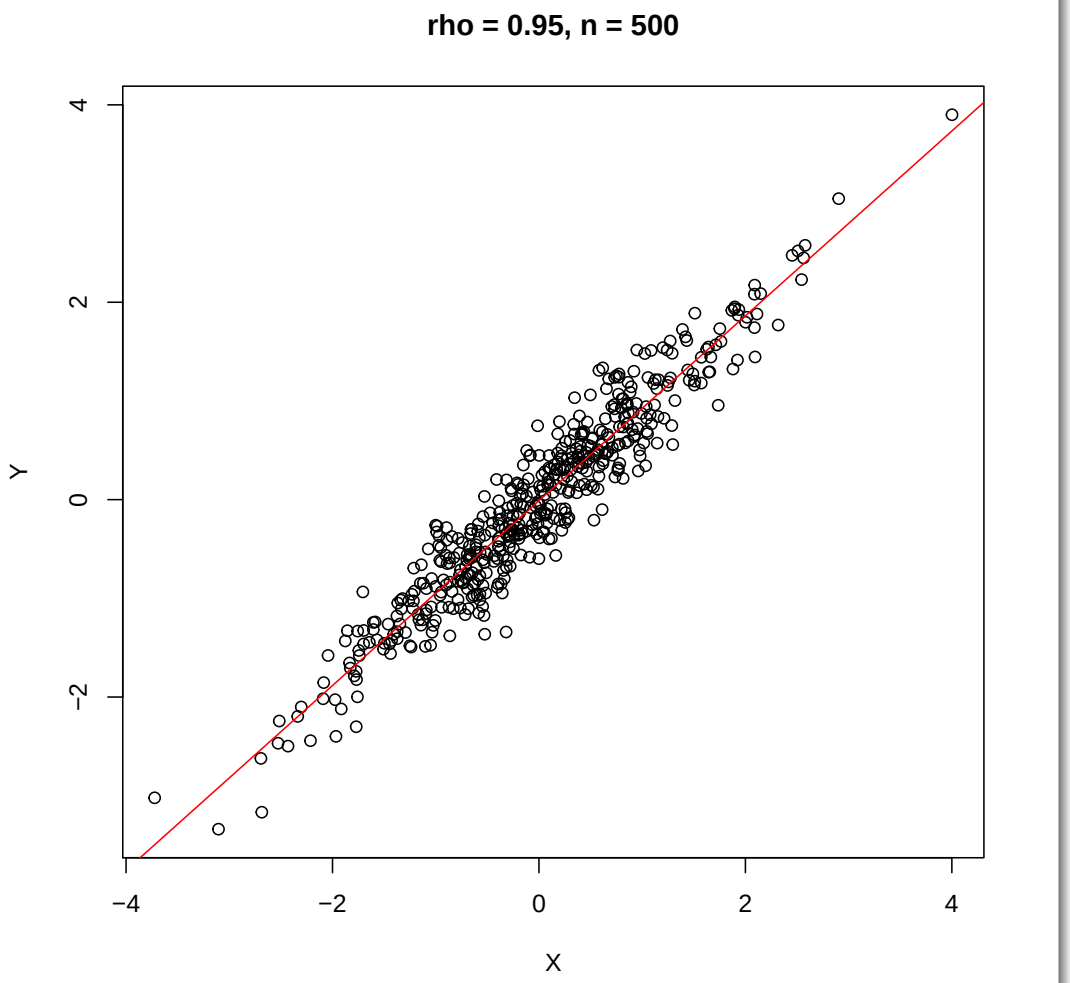
Example (Some Typical Scatterplots)



Interpreting a Correlation

Some Typical Bivariate Normal Scatterplots

Example (Some Typical Scatterplots)



Interpreting a Correlation

Some Typical Bivariate Normal Scatterplots

The problem with such examples is that people somethimes generalize from them incorrectly.

While a particular shape of scatterplot is connected with a certain level of correlation (circular scatterplots tend to reflect near-zero correlation, for example), the converse need not be true.

For example, a particular level of correlation may be connected with an infinity of different-shaped scatterplots.

Consequently, to understand the relationship between two variables, you must always examine the scatterplot.

This point is dramatized in the famous example known as Anscombe's Quartet.

Interpreting a Correlation

Anscombe's Quartet

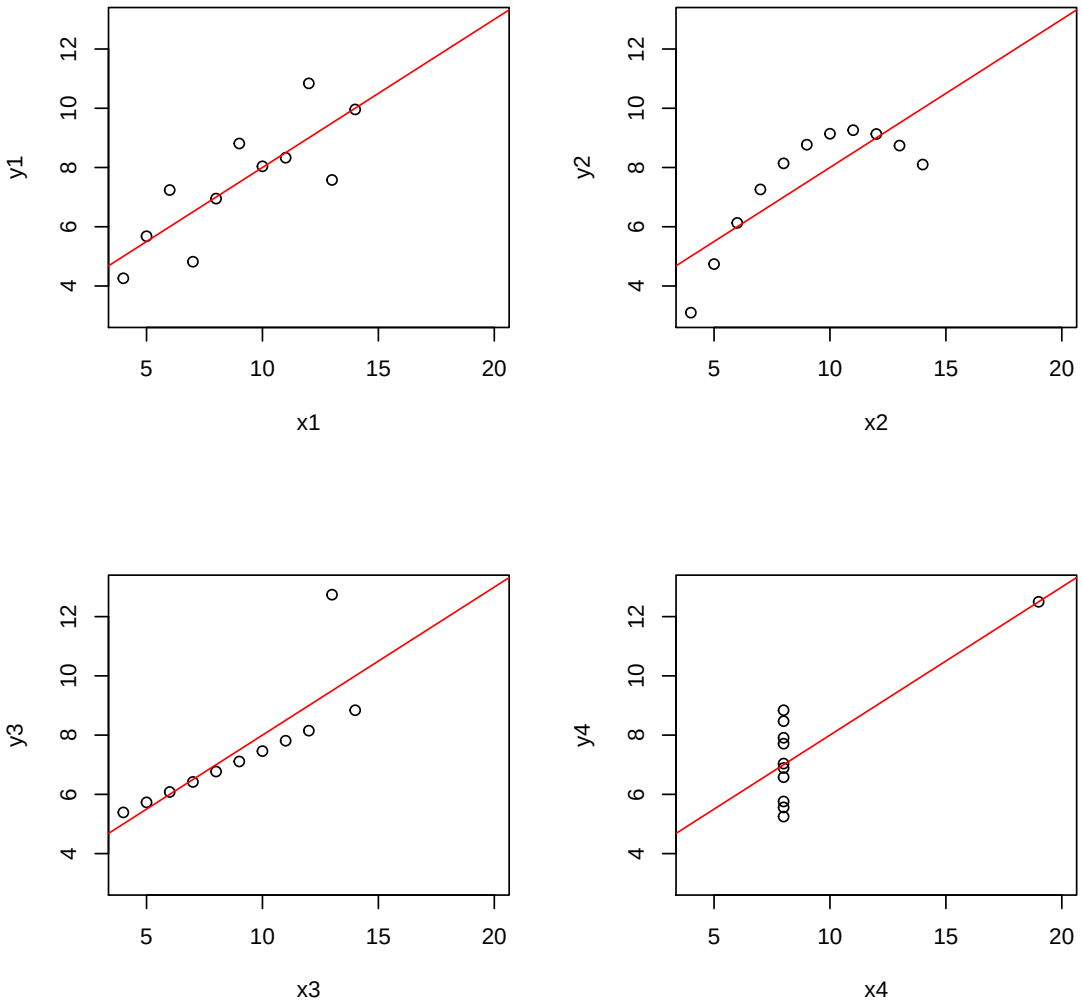
Anscombe presented 4 data sets that all have exactly the same means, variances, correlations, covariances, and regression lines.

Yet only one of the examples looks like the classic "bivariate normal with error" form.

Let's examine the scatterplots.

Interpreting a Correlation

Anscombe's Quartet



No Correlation vs. No Relation

In the previous subsection, we saw how many different scatterplots can support the idential correlation (and also support the same linear regression line).

The point is, the linear regression line is typically plotted with the assumption that an underlying linear model is correct.

If that assumption is correct, we might say that "all bets are off" regarding the relationship underlying a correlation.

No Correlation vs. No Relation

A classic example of this is embodied in a "trick" exam question that has annoyed a generation of students.

In this question, students are manipulated into agreeing with the statement that "a zero correlation between X and Y implies no relationship between X and Y."

This statement is incorrect. It is true that if there is no underlying relationship between X and Y (i.e., they are independent), then there will be no correlation between them.

However, a relationship can be perfect (but nonlinear) and the correlation coefficient can be zero.

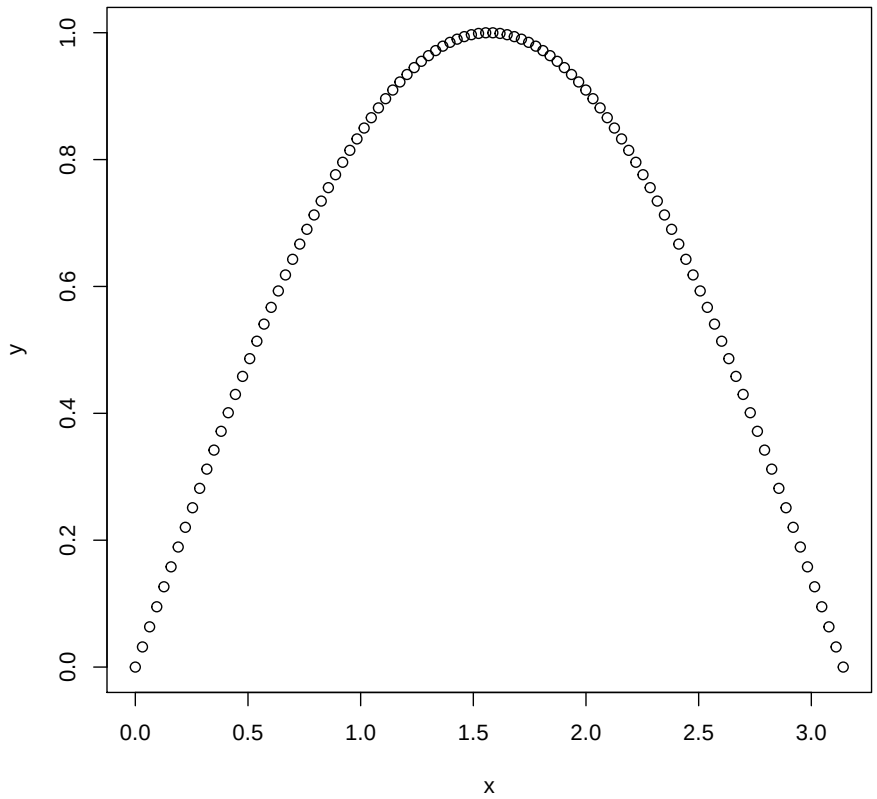
No Correlation vs. No Relation

Suppose, for example, the variables are X and $Y = \sin(X)$, and we have data for values of x ranging from 0 to π. The relationship between X and Y is perfect, but nonlinear

On the next slide, we plot the relationship and compute the correlation, which is within rounding error of zero.

No Correlation vs. No Relation

```
> x <- seq(from = 0, to = pi, length.out = 100)
> y <- sin(x)
> plot(x, y)
```




```
> cor(x, y)
[1] -1.423e-17
```

Perfect Correlation vs. Equivalence

It is not uncommon in the social sciences and education literature to see a high correlation between two variables to be used as justification for their "equivalence."

It is important to realize that measures of two completely different quantities can be correlated perfectly without the quantities being equivalent.

Here is an example.

Perfect Correlation vs. Equivalence

Height and Weight on the Planet Zorg

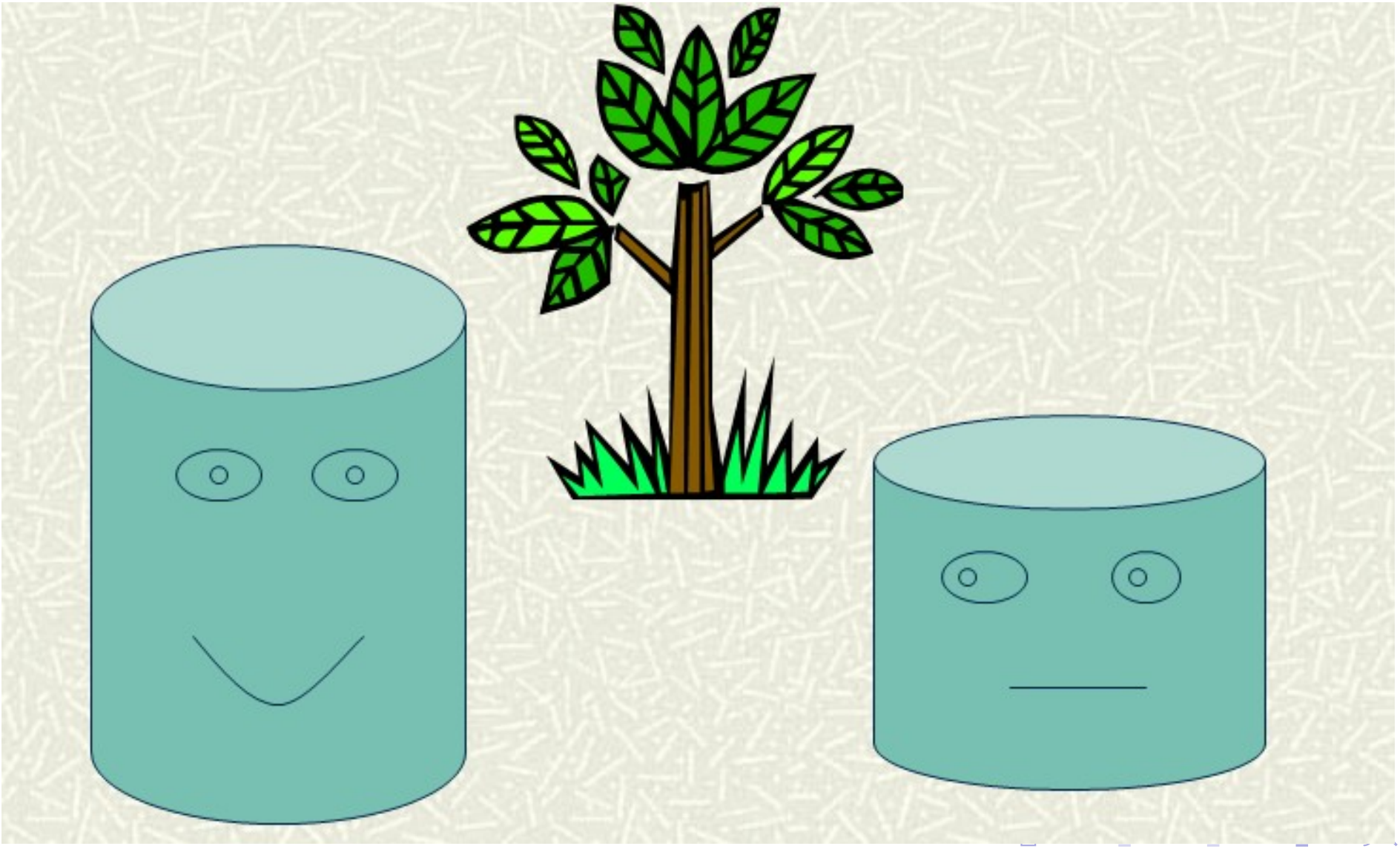
On the planet Zorg, Zorgians are perfect cylinders of the same width and density.

So on Zorg, height and weight are perfectly correlated.

Does this mean that height and weight are the same thing?

Perfect Correlation vs. Equivalence

Height and Weight on the Planet Zorg



Perfect Correlation vs. Equivalence

Height and Weight on the Planet Zorg

No, height and weight are not the same thing. Consider the following.

An initial evaluation of weight among Zorgians found that Zorgians above a certain "cutoff weight" had a high probability of dying before their 22nd birthday.

Although the data from the probe were sketchy, experts suggested that Zorgian physiology might be similar to earthlings, and that cardiac issues might be involved.

It turned out that this conclusion was completely wrong. It wasn't the Zorgian weights that were causing their early demise, it was their heights.

Perfect Correlation vs. Equivalence

Height and Weight on the Planet Zorg

Zorgians had a ceremony called "running the temple."

In this ceremony, young Zorgians around the age of 21–22, as a rite of passage to adulthood, ran through the doors of the temple while passersby showered them with flowers while simultaneously trying to hit them with switches made from branges of a local tree.

Unfortunately, the temple doors had sharp overhangs, and the taller Zorgians frequently suffered injuries by slicing off their tops (Zorgians have no head) and bleeding to death.

Perfect Correlation vs. Equivalence

Height and Weight on the Planet Zorg

This example highlights the deep conceptual confusion among those who seek to establish the existence and validity of constructs simply through correlation.

Combining Populations, and Ignoring Explanatory Variables

As we saw in a previous lecture, combining intact subgroups into one data set can result in correlations that are misleading, because they do not apply within any subgroup.

Restriction of Range

Regression can be used to describe the relationships in existing data.

It can also be used to produce a formula for predicting future performance from current data.

For example, we might analyze the relationship between SAT math scores and success in first year engineering at Vanderbilt. After gathering a fair amount of data, we might then use the regression formula to predict next year's performance at Vanderbilt from SAT data on this year's applicants.

Suppose we did that. There is a catch. We will select next year's students by using the formula and a "cutoff" score.

As a result, we will not have data next year for many of the applicants, because they did not gain admission to Vanderbilt.

Restriction of Range

When data are restricted on the dependent variable, the correlation between the predictor and the dependent variable tends to be an underestimate of the strength of the relationship that would have been evident had the data not been restricted

This is known as the restriction of range problem.

We can demonstrate it with an artificial data set.

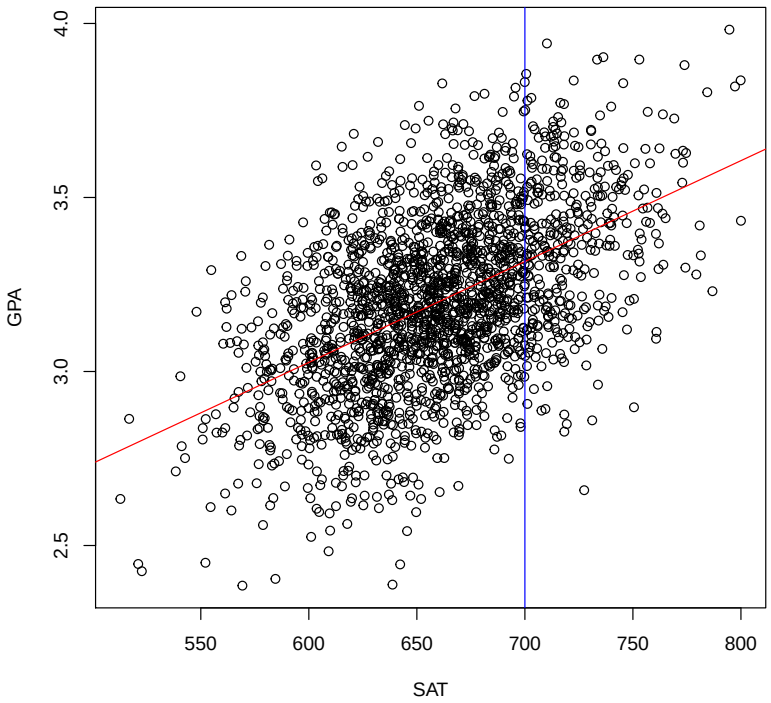
Restriction of Range

Suppose 2000 students apply to Vanderbilt Engineering, and the true relationship between SAT math and Engineering GPA shows a correlation of 0.51.

The plot below shows what would have been observed had all the applicants been admitted to Vanderbilt.

However, suppose that only those with SAT scores above 700 were admitted.

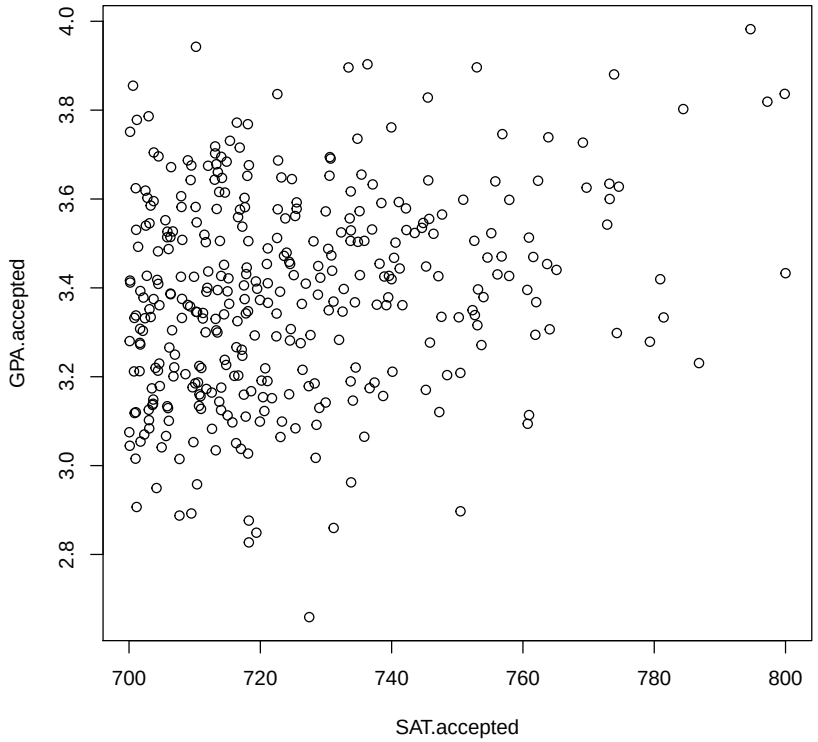
Then the data set would be reduced to those students to the right of the blue line.



Restriction of Range

Here is the scatterplot for only those applicants to the right of the blue line. The correlation is reduced from 0.51 to 0.24.

```
> SAT.accepted <- SAT[SAT > 700]
> GPA.accepted <- GPA[SAT > 700]
> plot(SAT.accepted, GPA.accepted)
```




```
> cor(SAT.accepted, GPA.accepted)
[1] 0.2408
```

Prediction vs. Explanation: The Shrinkage Problem

A common way that people can be deceived by a regression analysis is to assume that a multiple regression equation (especially one derived from a small sample using a large number of predictors) that works well in a current sample will work equally as well in a future sample.

Generally, the multiple correlation coefficient will "shrink" substantially if the current equation is applied to future values of X with the expectation of predicting future values of Y.

Most regression programs print an "adjusted R^2" value to attempt to adjust for shrinkage.

The Third Variable Fallacy

Often people assume, sometimes almost subconsciously, that when two variables correlate highly with a third variable, they correlate highly with each other.

Actually, if r_{XW} and r_{YW} are both .7071, r_{XY} can vary anywhere from 0 to 1.

Only when r_{XW} and/or r_{YW} become very high does the correlation between X and Y become highly restricted.

Correlation and Causality

Correlation is not causality. This is a standard adage in textbooks on statistics and experimental design, but it is still forgotten on occasion.

In an earlier lecture, we examined an example: the correlation between number of fire trucks sent to a fire and the dollar damage done by the fire.

This correlation can be elevated, because size of the fire directly affects damage done by the fire, and the number of trucks sent to the fire.