M7S1 - Correlation

Professor Jarad Niemi

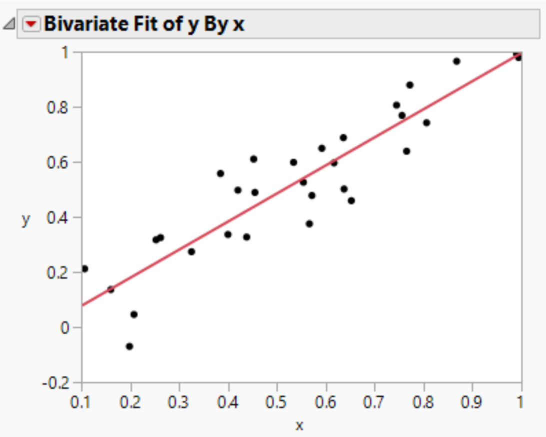
STAT 226 - Iowa State University

November 15, 2018

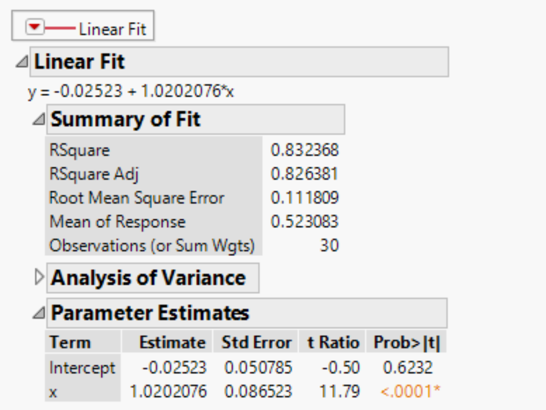
Overview

- Part 1 - Data and Probability
- Part 2 - Inferential Statistics (for a single variable)
    - Confidence intervals
    - Pvalues (hypothesis tests)
- Part 3 - Regression
    - Linear relationship between two variables: explanatory and response variables
    - Scatterplot
    - Fitting a line: intercept and slope
    - Confidence intervals and tests for the intercept and slope

Regression in JMP



Regression in JMP (cont.)



Linear Fit

y = -0.02523 + 1.0202076*x

Summary of Fit

RSquare	0.832368
RSquare Adj	0.826381
Root Mean Square Error	0.111809
Mean of Response	0.523083
Observations (or Sum Wgts)	30

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.02523	0.050785	-0.50	0.6232
x	1.0202076	0.086523	11.79	<.0001*

Outline

- Statistics for a single quantitative variable:
    - Location: mean, median, quartiles
    - Spread: standard deviation, variance, IQR
- Statistics for two quantitative variables:
    - Same statistics for each variable individually
    - Linear relationship: covariance, correlation

Association

Definition

Two variables are associated if certain values of one variable tend to occur often with certain values of a second variable.

Examples:

- height and weight of a person
- assessed value and sale price of a home
- quarterly profit and share price

These relationships won't be exact as there is always variation.

Explanatory vs response variable

Definition

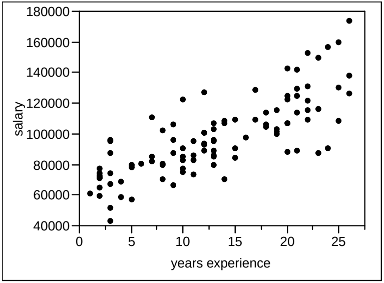
The response variable (or dependent variable) is the outcome of interest and is often denoted using the letter y. The explanatory variable (or independent variable) is the variable that explains (some of the) changes is the response variable and is often denoted using the letter x.

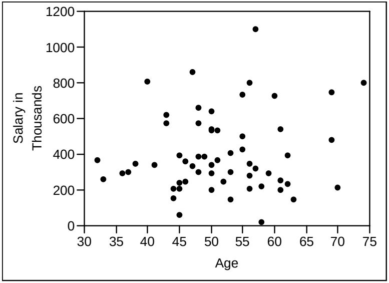
Examples:

Explanatory	Response
assessed value of a home	selling price of a home
years of education	starting salary

Scatterplot

When constructing a scatterplot, the explanatory variable is on the x-axis and the response variable is on the y-axis.



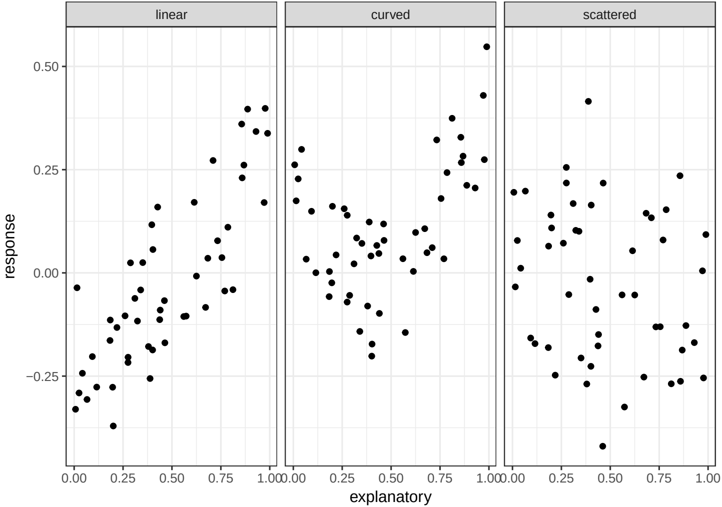


Scatterplots (cont.)

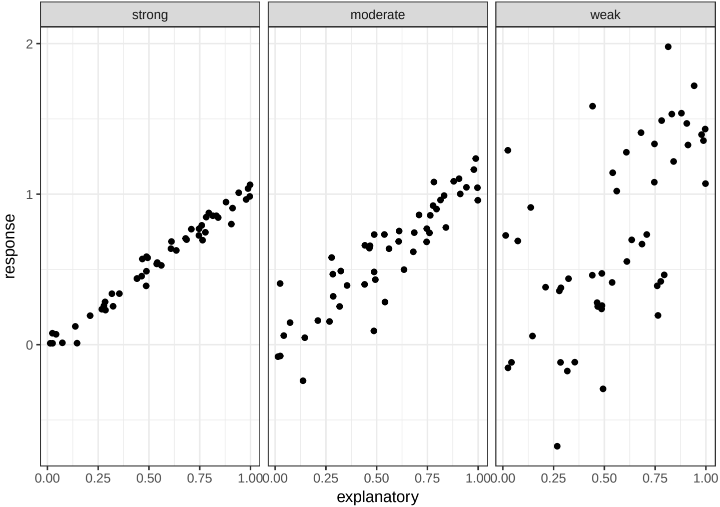
When looking at a scatterplot consider these 4 features:

- Form:
  - Linear
  - Curved
  - Scattered
- Direction:
  - Positive association
  - Negative association
- Strength:
  - Weak
  - Moderate
  - Strong
- (Possible) Outliers

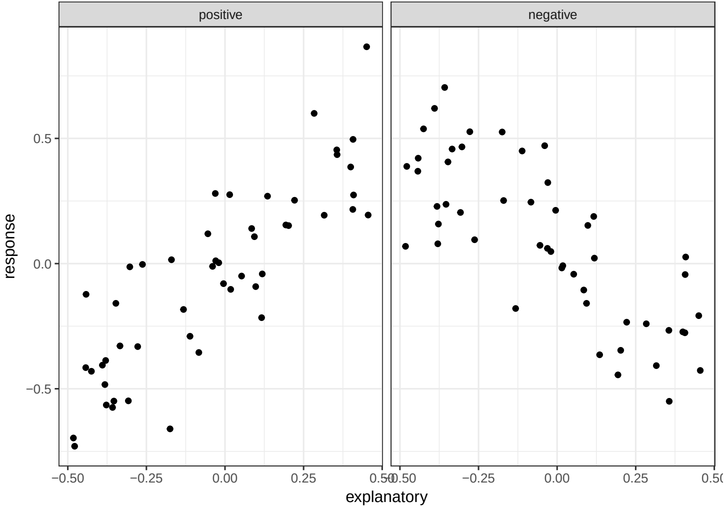
Form


linear
curved
scattered
response
0.50
0.25
0.00
−0.25
0.00
0.25
0.50
0.75
1.00
explanatory

Strength

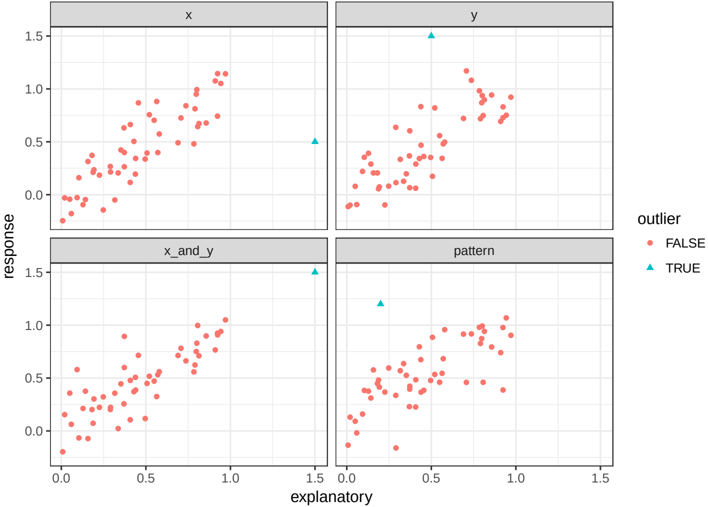

strong
moderate
weak
response
2
1
0
0.00
0.25
0.50
0.75
1.00
explanatory

Direction


positive
negative
response
0.5
0.0
−0.5
−0.50
−0.25
0.00
0.25
0.50
explanatory

Outliers

Observation(s) that differ from the pattern:



Correlation

Definition

For two variables x and y, the sample covariance is

$$s^2_{x,y} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

and the sample correlation (coefficient) is the sample covariance divided by the product of the sample standard deviations, i.e.

$$r = \frac{s^2_{x,y}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

where

- s_x is the sample standard deviation for the variable x and
- s_y is the sample standard deviation for the variable y.

Interpreting a correlation

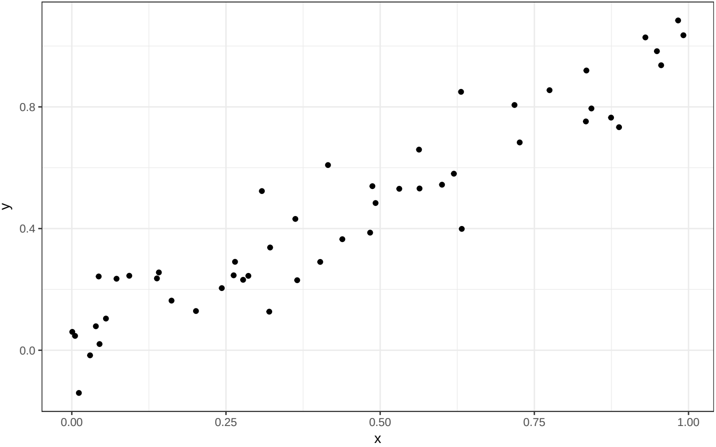
The sample correlation is a measure of the strength and direction of a linear relationship between two variables.

- Direction:
    - $r < 0$ indicates a negative direction
    - $r > 0$ indicates a positive direction
- Strength:
    - $r = 0$ indicates not linearly related
    - $0 < |r| \leq 0.3$ indicates weak strength
    - $0.4 < |r| \leq 0.7$ indicates moderate strength
    - $0.7 < |r| \leq 1$ indicates strong strength
    - $r = 1$ indicates a perfect, positive linear relationship
    - $r = -1$ indicates a perfect, negative linear relationship

Notes:

- sample correlation has no units
- sample correlation is easily influenced by outliers

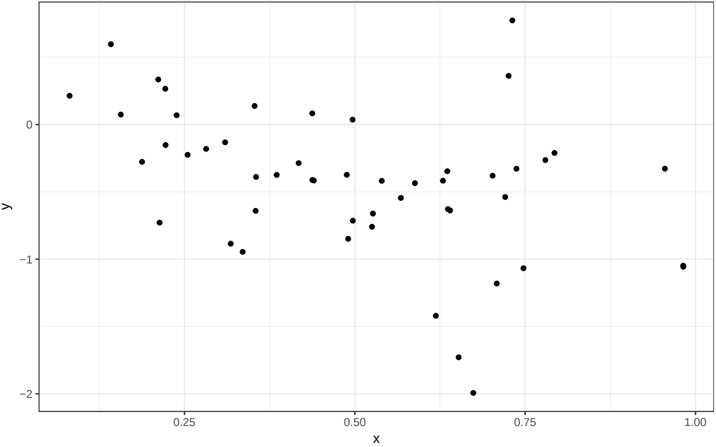
Guess sample correlation r



```
cor(x,y)
```

```
[1] 0.9452718
```

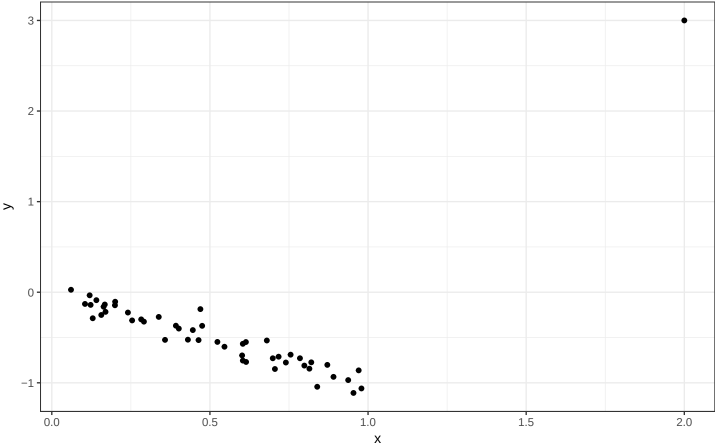
Guess sample correlation r



```
cor(x,y)

[1] -0.4030499
```

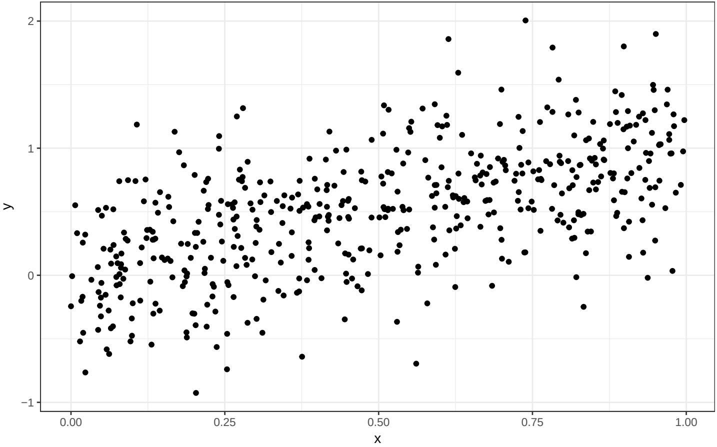
Guess sample correlation r



```
cor(x,y)
```

```
[1] 0.1209367
```

Guess sample correlation r



```r
cor(x,y)
```

```
[1] 0.5884374
```

Guess the correlation

For an additional practice guessing the correlation, see this shiny app
`http://shiny.stat.calpoly.edu/Corr_Reg_Game/`