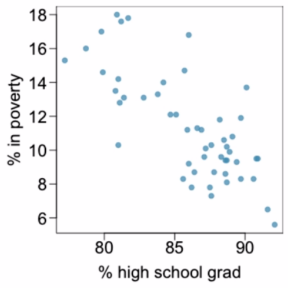
Poverty vs. high school graduation



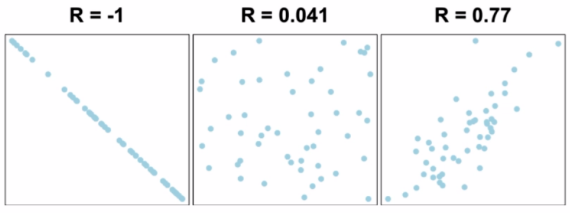
Scatterplot: for each US state and DC, percent of residents living in poverty vs high school graduation rate

Response variable:
% in poverty
Explanatory variable:
% high school grad
Relationship:
linear, negative, moderately strong

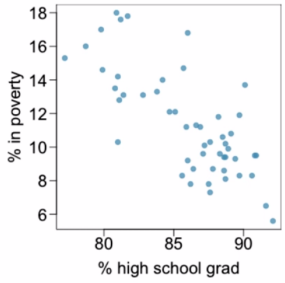
Correlation

Strength of linear association between 2 variables

-1 perfect negative association
0 no association
+1 perfect positive association



Example



Based on the plot, which of the following is the best guess for the correlation between % in poverty and % high school grad?

```
cor(poverty$Graduates, poverty$Poverty)
[1] -0.7468583
```

Don't rely on summary statistics

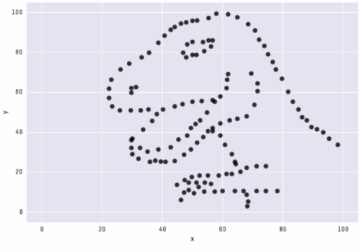
Picture a data set with the following statistics:

- mean x: 54.26
- mean y: 47.83
- sd x: 16.76
- sd y: 26.93
- R: -0.06

What does this data set look like?

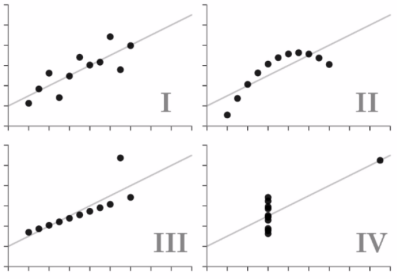
Don't rely on summary statistics

- mean x: 54.26
- mean y: 47.83
- sd x: 16.76
- sd y: 26.93
- R: -0.06



Don't rely on summary statistics

- mean x: 9
- mean y: 7.5
- sd x: 3.32
- sd y: 2.03
- R: 0.816

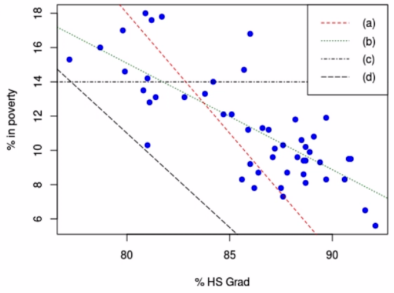


Fitting a line by least squares regression

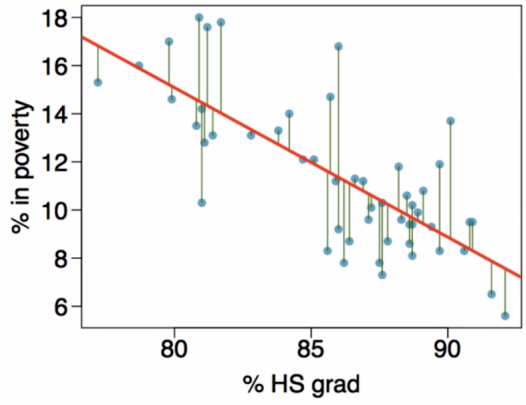
Example

Which of the lines appears to best fit the linear relationship between percent in poverty and percent HS grad?

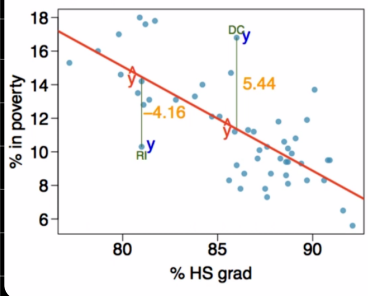
Choose one.



Data = Model + Residuals



Residuals



Residual: difference between observed y_i and predicted $\hat{y}_i$

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

% in poverty in DC is 5.44% > predicted

% in poverty in RI is 4.16% < predicted

A measure of the best line

Option 1: Minimize sum of residual magnitudes

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

Option 2: Minimize sum of squared residuals

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

"Least squares" is most common

- Makes the minimization math easier
- Penalizes larger errors more heavily

Calculating b_0 and b_1

Slope:

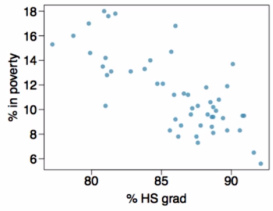
$b_1 = s_y / s_x \times R$

$b_1 = 3.1 / 3.73 \times -0.75$
$= -0.62$

Intercept:

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

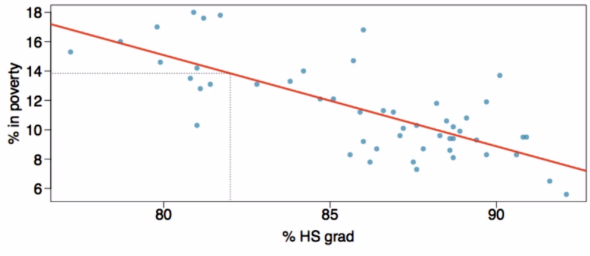
$b_0 = 11.35 - (-0.62)\ 86.01$
$= 64.68$



	% HS grad (x)	% in poverty (y)
mean	$\bar{x} = 86.01$	$\bar{y} = 11.35$
sd	$s_x = 3.73$	$s_y = 3.1$
	correlation	$R = -0.75$

Prediction

Plug x into $\hat{y} = b_0 + b_1 x$



$\hat{y}(82) = 64.68 - 0.62 * 82 = 13.84$

Residuals

Residuals measure the "error" in the model.

In order to find the best model, we should try to minimize the errors.

Usually, a change to the model will increase some residuals and decrease others; how best to minimize error overall?

Least Squares Model

$y = \beta_0 + \beta_1 x$ + "error"

Where:

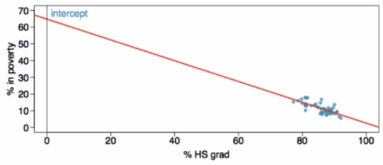
- x is the explanatory variable
- y is the response variable
- β_0 is the intercept
- β_1 is the slope
- Error assumed to be normally distributed, mean 0

For the population parameter, we use β_0 and β_1

For the calculated fit, we use b_0 and b_1 :

$\hat{y} = b_0 + b_1 x$

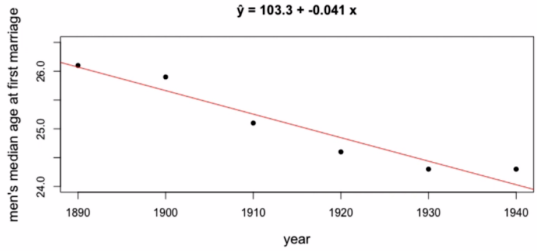
Interpreting b_0 and b_1



$b_1 = -0.62$ means:
For a 1% gain in graduation %, we would expect a 0.62% drop in the % in poverty

$b_0 = 64.68$ means:
For a state with 0% high school graduation, we would expect 64.68% in poverty

Example

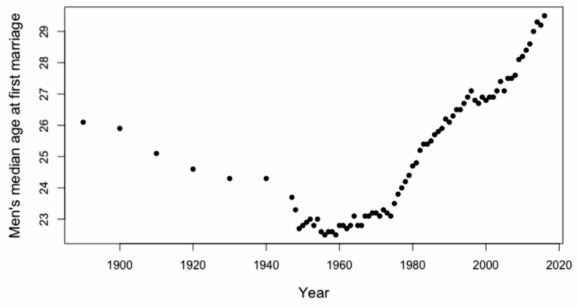


Extrapolate men's median age at first marriage in

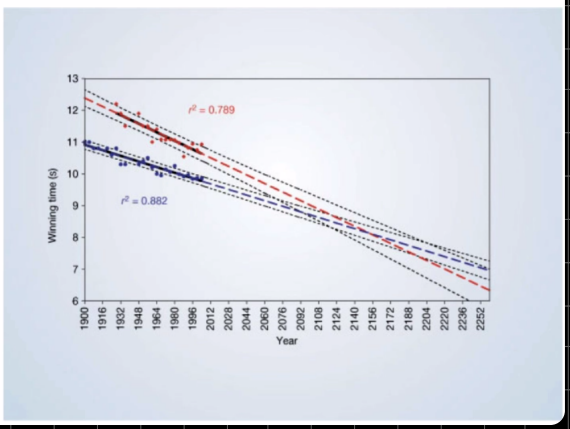
1950 1970 1990 2010

woman run faster in future?

Dangers of Extrapolation



Example of Extrapolation



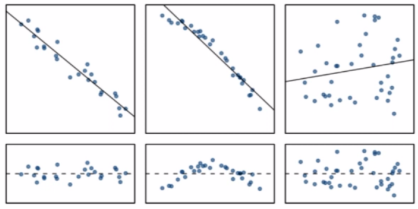
Conditions for least squares line

1. Linearity
2. Nearly normal residuals
3. Constant variability

Condition: Linearity

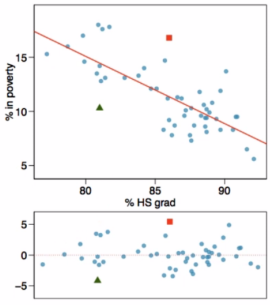
Relationship between the explanatory and the response variable should be linear

Check scatter plot and residuals plot



↑ not linear

Anatomy of a residuals plot



▲ *RI:*

$\%\ HS\ grad = 81 \qquad \%\ in\ poverty = 10.3$

$\widehat{\%\ in\ poverty} = 64.68 - 0.62 * 81 = 14.46$

$e = \%\ in\ poverty - \widehat{\%\ in\ poverty}$

$= 10.3 - 14.46 = -4.16$

■ *DC:*

$\%\ HS\ grad = 86 \qquad \%\ in\ poverty = 16.8$

$\widehat{\%\ in\ poverty} = 64.68 - 0.62 * 86 = 11.36$

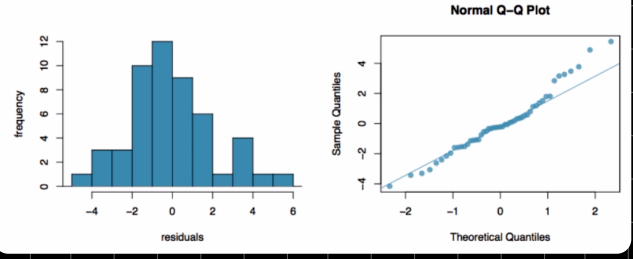
$e = \%\ in\ poverty - \widehat{\%\ in\ poverty}$

$= 16.8 - 11.36 = 5.44$

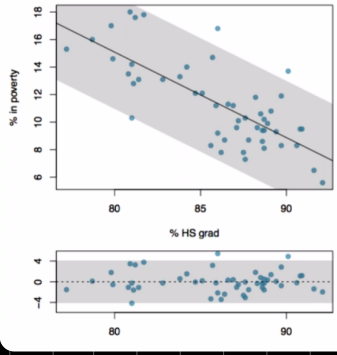
Condition: nearly normal residuals

Residuals should not be skewed

Check histogram, normal Q-Q plot of residuals



Condition: constant variability



Variability around least squares line should be roughly constant

Check histogram, normal Q-Q plot of residuals

R^2

R^2, the square of the correlation coefficient:

- measures strength of fit of a linear model
- represents percent of variability in response variable that is explained by the model

Remainder of the variability is explained by

- variables not included in the model
- inherent randomness in the data

Types of Outliers

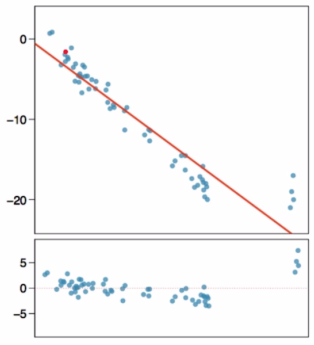
Regression lines: outlier influence

Imagine the line that would be drawn without outliers

- Would it look any different?
- Would it fit non-outlier points better?

For example, on the right:

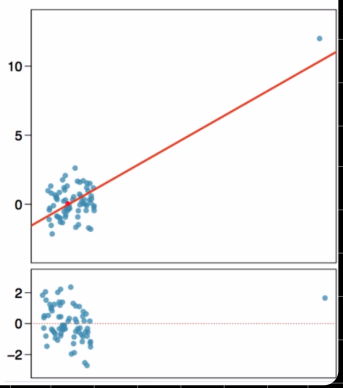
- The outliers have decreased the slope
- The fit on the non-outlier points could be better



Regression lines: outlier influence

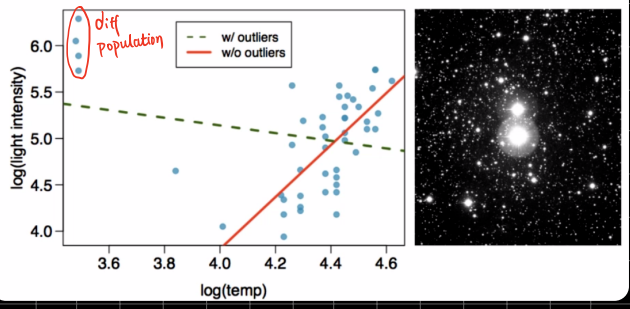
How has the outlier influenced the regression line?

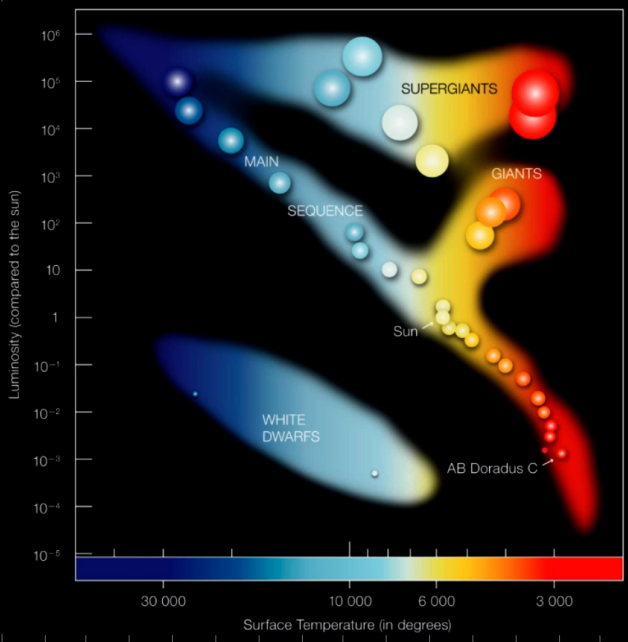
Without the outlier, there would be no visible association



Lines with and without outliers

Surface temperature vs light intensity of the 47 stars in the star cluster CYG OB1

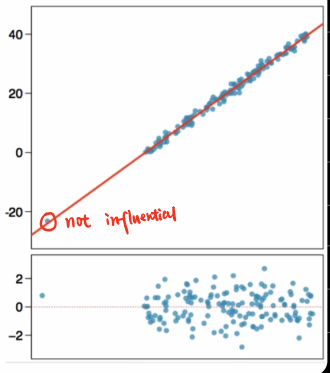




Leverage and Influence

A high leverage point has an extreme value of the explanatory variable. An influential point is a high leverage point that greatly changes the slope of the regression line.

The point at the lower left of the plot is high leverage but not influential.



Example

The lowest point on the plot is an outlier.

Is it influential?

Does it have high leverage?