Chapter 8: Regression

Math 140 · Fall '21

Based on content in OpenIntro Stats, 4th Ed

Hitchman

November 10, 2021

Section 8.1
Correlation between two numerical variables

Examining two numerical variables

Suppose two numerical variables describe a population

- people, described by height and weight
- countries, described by per capita income and life expectancy of its citizens
- movies in a theatre, described by proceeds in the first week and proceeds in the second week.

Linear Association?

Does one variable "explain" or cause changes in the other?

- the bigger the car, the worse the mileage?
- An *explanatory variable* (x) is a variable that attempts to explain observed outcomes.
- A *response variable* (y) measures an outcome of a study.
- *Linear regression* is the statistical method for fitting a line to data where the relationship between two variables, x and y, can be modeled by a straight line with some error:

$$y = \beta_0 + \beta_1 x + \epsilon.$$

Game plan

To study the relationship between 2 numerical variables:

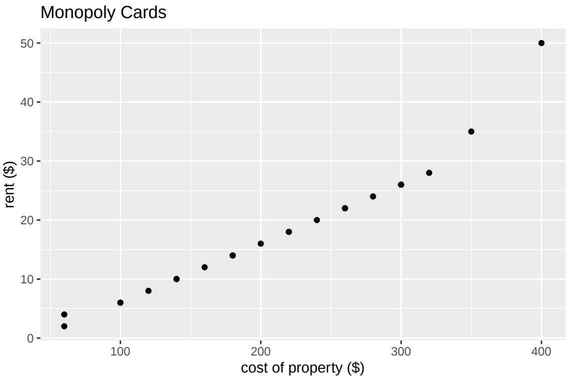
- graph the data (scatterplot!)
- look for the overall shape, pattern, and deviations from the pattern
- add numerical descriptions of specific aspects of the data.

Monopoly

Question: What is the relationship between the cost to buy a property, and the rent you collect if you own it?

x - cost of property (explanatory variable)

y - rent (response variable)

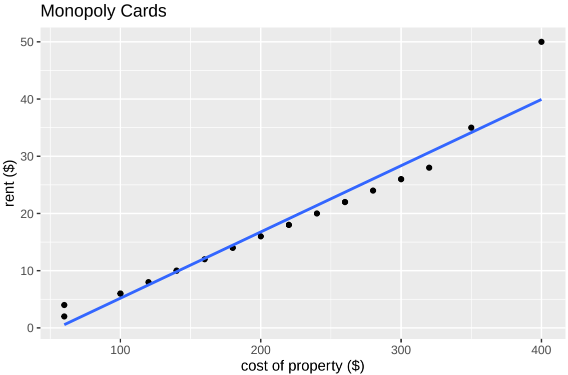


Qualitative Discussion of Association

- *Direction*
  - positive (as x increases so does y), or
  - negative (as x increases, y decreases)
- *Form*
  - Linear
  - Curved
  - Haphazard
- *Strength*
  - The tendency for points to stick close to the form

Monopoly Example

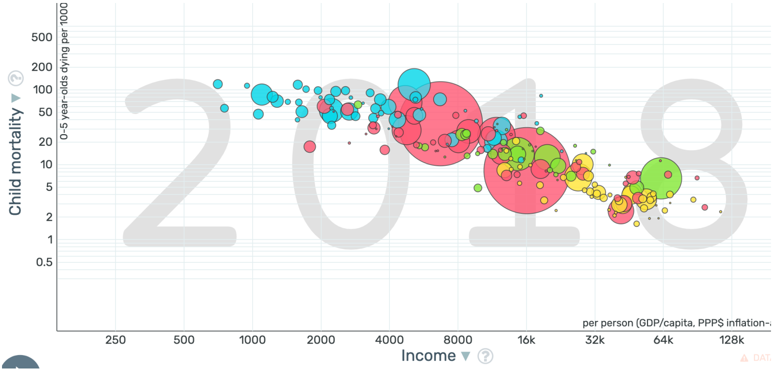
Direction - Positive! As cost increases so does the rent collected.
Form - Linear model seems reasonable.
Strength - It looks like the points hug the line pretty snugly, with the exception of the extremely cheap properties and the most expensive one - Boardwalk. Those extremes represent clear deviations from the form.



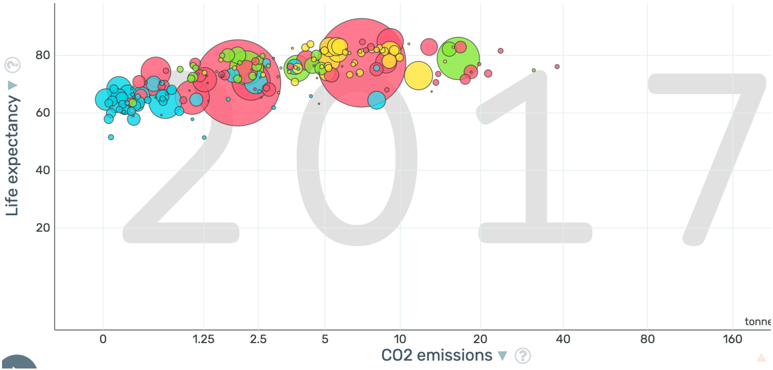
Gapminder

We can find excellent animations of real world data on Gapminder

Per capita income vs child mortality

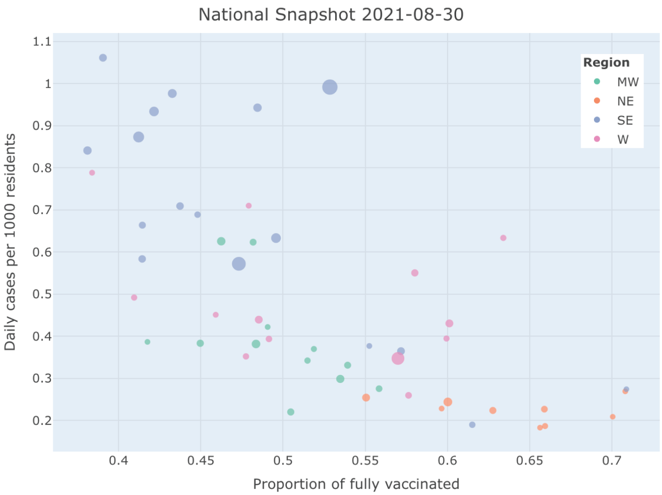


CO_2 emissions vs Life Expectancy



COVID-19

Daily new cases per 1000 residents vs vax rates, late August, 2021



Scatterplots Summary

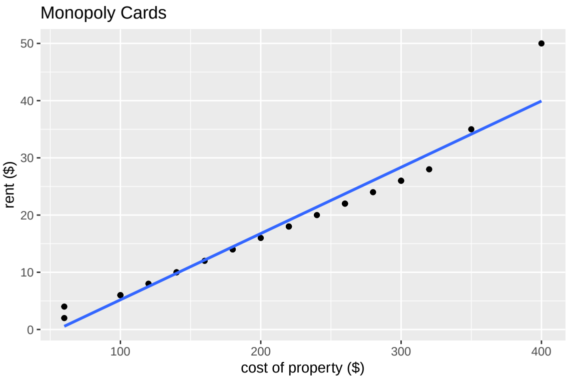
Scatterplots helps us

- Detect overall trends or patterns
- Observe deviations from that pattern (outliers)
- Get a feel for the strength and direction for the relationship between two variables.

Monopoly Example

If we believe a linear model is a good fit for the relationship between two variables, we can use the line to make predictions.

If we know x can we predict y?



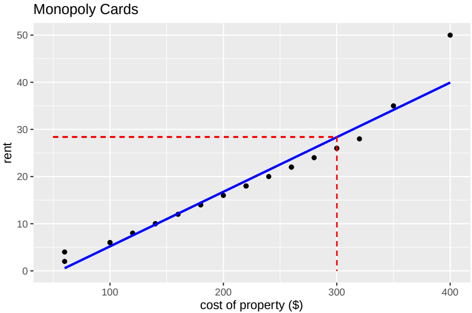
Monopoly Example

The equation of this line is $\hat{y} = -6.389 + 0.116x$

We can plug in a value of x to get a predicted value of y, which we denote as $\hat{y}$.

For instance, the linear model predicts that a property selling for $x =$\$300 has rent price equal to

$$\hat{y} = -6.389 + 0.116 \cdot 300 = \$28.4.$$



Residual

We note that Monopoly does have a property that sells for \$300, Pacific Ave, and the rent for this property is \$26.
So, Pacific Ave contributes the data point (300,26) to the scatter plot, and the linear model predicts a rent price equal to $\hat{y} = \$28.4$.

residual

The residual of the ith observation (x_i, y_i) is the difference of the observed response (y_i) and the predicted response $(\hat{y}_i)$ predicted by the model:

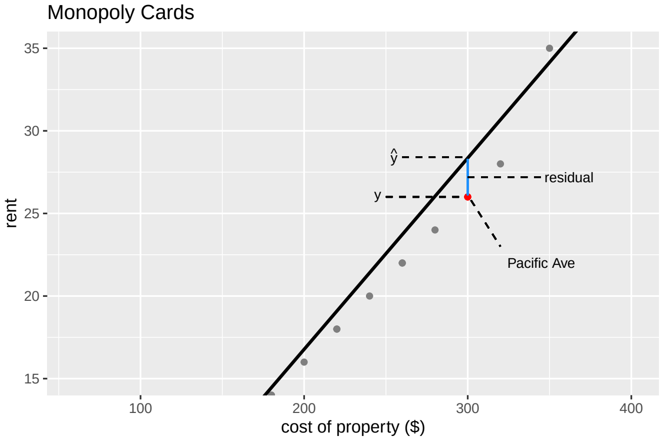
$$e_i = y_i - \hat{y}_i$$

We typically identify $\hat{y}_i$ by plugging x_i into the model.

"Residual equals observed minus predicted"

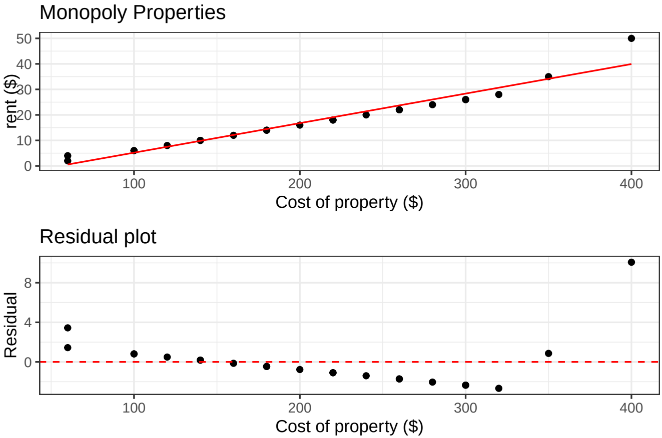
Residual value for Pacific Avenue

The residual for Pacific Ave, is $26 - 28.4 = -2.4$. This residual is negative because the data point for Pacific Ave is below the best-fit line!



Residual plots

A *residual plot* plots each data point's x-coordinate against that point's residual value.



A few more residual plots

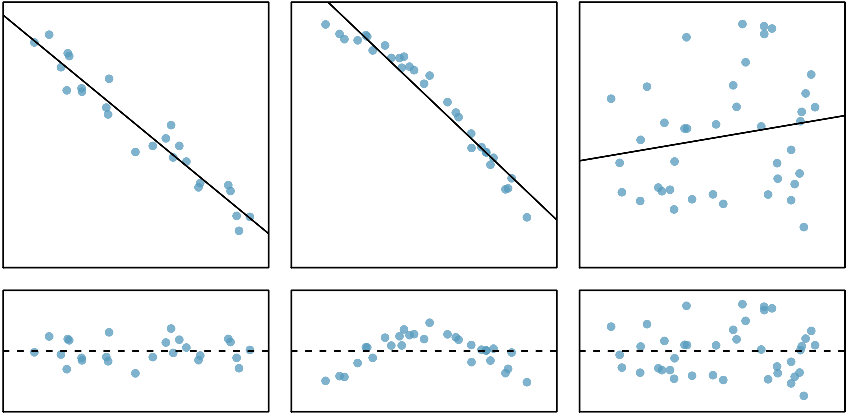


Figure 8.8, p. 310

Residual plots to check the model

- One purpose of residual plots is to identify characteristics or patterns still apparent in data after fitting a model.
- A clear pattern in a residual plot suggests a linear model might not be the best model for the relationship between x and y.

A numerical description of association

Correlation Coefficient - a number (r) between -1 and 1 that measures the strength and direction of the **linearity** of the relationship.

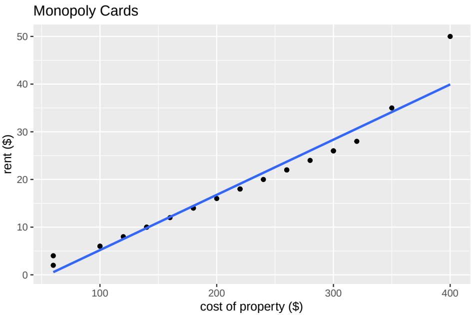
- If the direction is negative, $r < 0$
- if the direction is positive, $r > 0$
- the closer the points hug a single line, the closer r gets to ± 1.
- If there is really no linear form of any kind, $r \approx 0$.

Guess the correlation

Guess the correlation

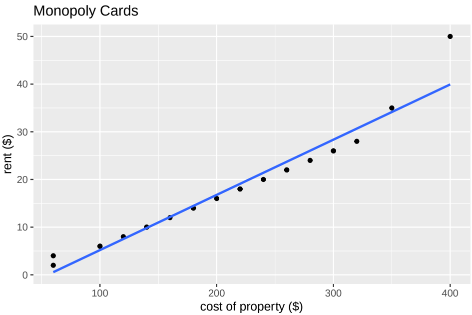
Monopoly Data

Recall, $x =$ cost of property, and $y =$ rent. The correlation coefficient for these data:



Monopoly Data

Recall, $x =$ cost of property, and $y =$ rent. The correlation coefficient for these data:




```
> cor(x,y)
[1] 0.9710672
```

Facts about r

- r does not depend on units.
- It only measures the strength of a linear relationship.
- r is strongly affected by outliers.
- $r = 1$ means all data literally lie on a single line with positive slope.
- $r = -1$ means all data literally lie on a single line with negative slope.
- r is the same whether we regress x on y or y on x.

```
> cor(x,y)
[1] 0.9710672
> cor(y,x)
[1] 0.9710672
```

Formula for r

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right),$$

where

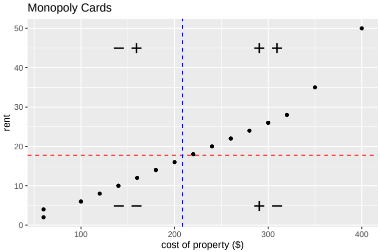
- n = number of observations
- (x_i, y_i) is our notation for a typical observation
- $\overline{x}$ is the sample mean of the x values
- $\overline{y}$ is the sample mean of the y values
- s_x is the standard deviation of the x values
- s_y is the standard deviation of the y values

Making sense of the formula

We can mark our plot into quarters by using a vertical line through the value of $\overline{x}$ on the x-axis, and a horizontal line through the value of $\overline{y}$ on the y-axis.

Quadrant signs $\leftrightarrow$ signs of $(x_i - \overline{x})/s_x$ term and $(y_i - \overline{y})/s_y$ for a point in that quadrant.

The quadrants marked $++$ and $--$ will contribute positively to the sum for r; the other quadrants contribute negatively to the sum.



Correlation is not Causation

http://www.tylervigen.com/spurious-correlations

Correlation is not Causation

Not all correlations are so obviously “spurious”!

- Does sleeping with a night-light as a young child lead to myopia (nearsightedness)?
- CNN article on a 1999 study
- A later study
- "[M]yopia is genetic, and nearsighted parents more frequently placed nightlights in their children's rooms.

Correlation is not Causation

Not all correlations are so obviously "spurious"!

- Does sleeping with a night-light as a young child lead to myopia (nearsightedness)?
- CNN article on a 1999 study
- A later study
- "[M]yopia is genetic, and nearsighted parents more frequently placed nightlights in their children's rooms.
- **HDL cholesterol**. This "good" cholesterol is associated with lower rates of heart disease. But heart-disease drugs that raise HDL cholesterol are ineffective. Why? It turns out that while HDL cholesterol is a byproduct of a healthy heart, it doesn't actually cause heart health.

https://qz.com/316826/mathematicians-have-finally-figured-out-how-to-tell-correlation-from-causation/

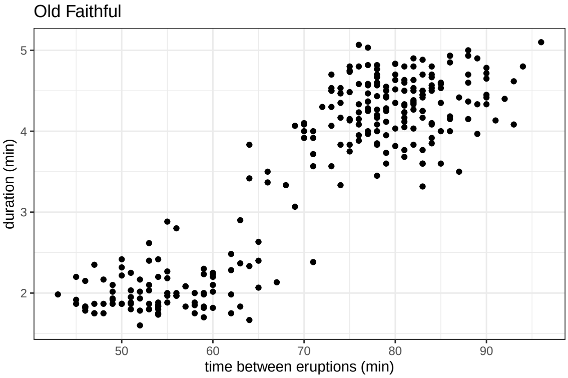
Section 8.2
Least Squares Regression

Least-Squares Regression

https://www.yellowstonepark.com/things-to-do/geysers-hot-springs/about-old-faithful/

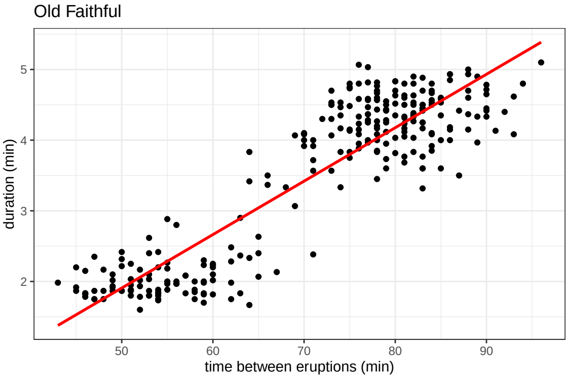
Least-Squares Regression Line

- The *Least-Squares Regression Line* is the line that minimizes the sum of the squares of the vertical distances between the data points and the line.



Least-Squares Regression Line

- The *Least-Squares Regression Line* is the line that minimizes the sum of the squares of the vertical distances between the data points and the line.

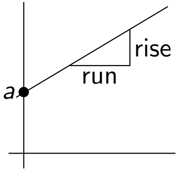


Lines - Quick Review

- A non-vertical line is determined by two features:
  - y-intercept: value at which the line intersects the y-axis
  - slope: rise/run

Lines - Quick Review

- A non-vertical line is determined by two features:
  - y-intercept: value at which the line intersects the y-axis
  - slope: rise/run
- The slope-intercept equation of a line: $y = a + bx$
  - $a = y$-intercept
  - $b = \text{slope} = \text{rise/run}$



Lines - Quick Review

desmos.com - A great online graphing calculator

Picturing the Least-Squares Regression Line

desmos example!

Determining the Least-squares regression line

Given n points of the form (x_i, y_i) we need to know:

- $\overline{x}$ - the mean of the x_i
- s_x - the standard deviation of the x_i
- $\overline{y}$ - the mean of the y_i
- s_y - the standard deviation of the y_i
- r - the correlation coefficient of the scatter plot

The equation of the least-squares regression line

has the form

$$y = a + bx$$

where the slope is

$$b = r\frac{s_y}{s_x}$$

and the y-intercept is

$$a = \overline{y} + b\overline{x}.$$

Regression in RStudio

The command `lm`($y \sim x$) will produce the slope and y-intercept of the least-squares regression line where x is the "predictor" variable and y is the "response".
(`lm` is short for "linear model")
For the Old Faithful data:

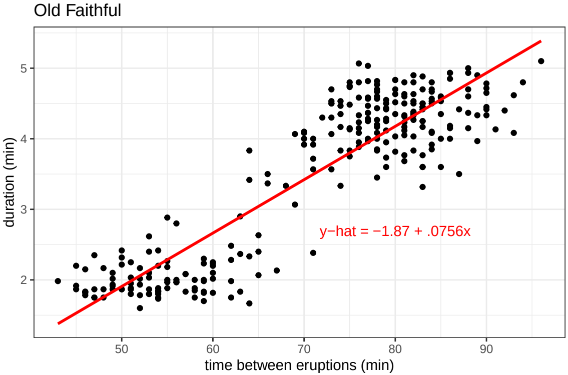
```
> lm(faithful$eruptions~faithful$waiting)

Call:
lm(formula = faithful$eruptions ~ faithful$waiting)

Coefficients:
     (Intercept)  faithful$waiting
        -1.87402           0.07563
```

Example: Old Faithful

So the least-squares regression line has equation

$$\hat{y} = -1.87 + .0756x.$$

- The *Least-Squares Regression Line* is the line that minimizes the sum of the squares of the vertical distances between the data points and the line.



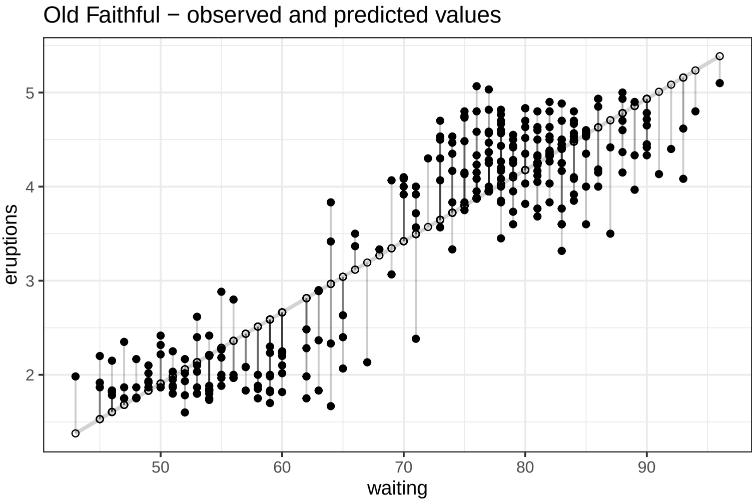
Adding Predicted and Residual Values to Data

```
of <- faithful
fit <- lm(eruptions~waiting, data=of)
of$predicted <- predict(fit)
of$residual <- residuals(fit)
```

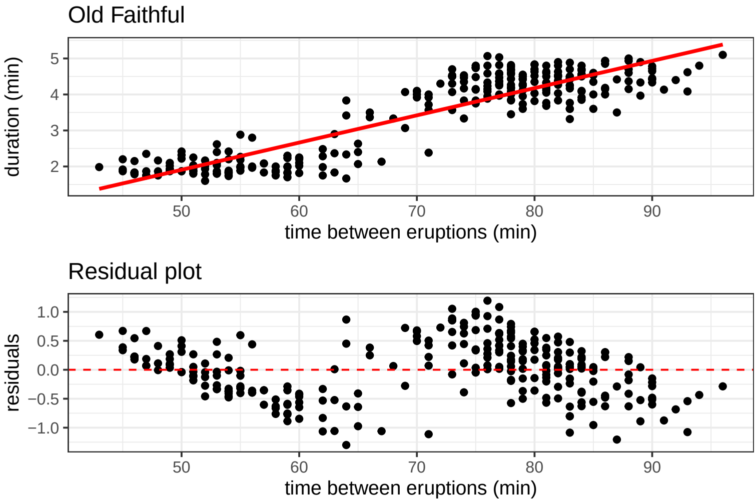
The Data

i	waiting (x_i)	eruptions (y_i)	predicted ($\hat{y}_i$)	residual (e_i)
1	79	3.60	4.10	-0.50
2	54	1.80	2.21	-0.41
3	74	3.33	3.72	-0.39
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
271	46	1.82	1.60	0.21
272	74	4.47	3.72	0.74

Example: Old Faithful



Example: Old Faithful



Interpreting Slope of least-squares regression line

slope

The *slope*, b, in

$$\hat{y} = a + bx$$

indicates the predicted change in y if x is increased one unit.

- The Old Faithful linear model $\hat{y} = -1.87 + .0756x$ predicts eruption duration ($\hat{y}$) based on the wait time between eruptions (x).
- The slope for this model is .0756.
- So, the model predicts that for each minute of wait time between eruptions, the duration of the eruption will increase by .0756 minutes (4.5 seconds).

Interpreting the y-intercept of least-squares regression line

y-intercept

The *y-intercept*, a, in

$$\hat{y} = a + bx$$

indicates the predicted value of y if $x = 0$, but this predicted value only makes sense if the linear model is reasonable all the way to $x = 0$.

- The Old Faithful linear model $\hat{y} = -1.87 + .0756x$ predicts eruption duration ($\hat{y}$) based on the wait time between eruptions (x).
- The y-intercept for this model is -1.87.
- Thus, the model predicts that if the wait time between eruptions is $x = 0$ minutes the new eruption will last $\hat{y} = -1.87$ minutes, which is clearly a meaningless statement.
- Note: the values of x in the data range from 43 to 96 minutes, $x = 0$ is well outside the range of x-values used to build the model.

When those blizzards hit the East Coast this winter, it proved to my satisfaction that global warming was a fraud. That snow was freezing cold. But in an alarming trend, temperatures this spring have risen. Consider this: On February 6th it was 10 degrees. Today it hit almost 80. At this rate, by August it will be 220 degrees. So clearly folks the climate debate rages on.

Stephen Colbert April 6th, 2010

Extrapolation

- Applying a model estimate to values outside of the realm of the original data is called *extrapolation*.
- If we extrapolate, we are making an unreliable bet that the approximate linear relationship will be valid in places where it has not been analyzed.

Using r^2 to describe the strength of a linear fit

- Recall, the correlation coefficient r is between -1 and 1:

$$-1 \leq r \leq 1$$

- If we square this value we get a number between 0 and 1:

$$0 \leq r^2 \leq 1.$$

- r^2 is a standard value used in statistics to indicate the strength of a linear fit for the relationship between x and y

What does r^2 measure?

What does r^2 measure?

The value r^2 of a linear model describes the amount of variation in the response that is explained by the least squares line.

Let's consider the Old Faithful Data:

```
> cor(of$eruptions, of$waiting)
[1] 0.9008112
> cor(of$eruptions, of$waiting)^2
[1] 0.8114608
```

Using r^2 to describe the strength of a linear fit

- For the Old Faithful data, the variance of the response variable is

$$s^2_{\text{eruptions}} = \texttt{var(of\$eruptions)=1.3027}$$

- Applying the least squares line reduces the uncertainty in predicting duration, and variation in the residuals describes how much variation **remains** after using the model:

$$s^2_{\text{res}} = \texttt{var(of\$residuals)=0.2456}$$

- So we have a reduction of

$$\frac{1.3027 - 0.2456}{1.3027} = 0.8115,$$

or about 81.15% in the data's variation by using waiting time to predict eruption duration with the linear model. This corresponds exactly to the r^2 value.

Four Conditions for the Least Squares Line

Linearity

The data should show a linear trend. If there is a nonlinear trend, an advanced regression method from another book or later course should be applied.

Nearly normal residuals

Generally, the residuals must be nearly normal. When this condition is not met, it is usually because of outliers or influential data points. An example of a residual of potential concern is shown in Figure 8.12

Four Conditions for the Least Squares Line

Constant variability

The variability of points around the least squares line remains roughly constant as x changes. An example of non-constant variability is shown in the third panel of Figure 8.12 of the text, which represents the most common pattern observed when this condition fails: the variability of y is larger when x is larger.

Independent observations

Be cautious about applying regression to time series data, which are sequential observations in time such as a stock price each day. Such data may have an underlying structure that should be considered in a model and analysis. An example of a data set where successive observations are not independent is shown in the fourth panel of Figure 8.12. There are also other instances where correlations within the data are important, which is further discussed in Chapter 9.