Introduction to Statistics I

Instructor: Jodin Morey moreyj@lemoyne.edu

Previous Lecture

- Graphically display relationships between two quantitative vars
- Scatterplots: Direction, Strength, Form

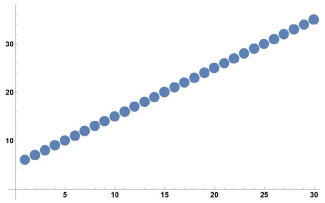
Topic 27/28: Correlation Coefficient & Least Squares Regression

Correlation Coefficient (r)

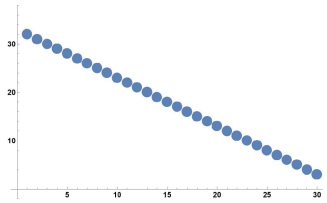
Measures degree to which two quantitative vars are *associated*.

Correlation coefficient describes direction & strength of a linear relationship.

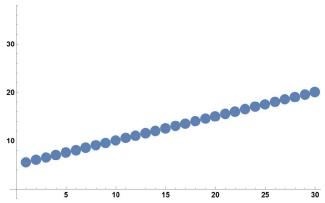
Correlation r can take values from -1 to 1. Measures how close data follows a straight line.



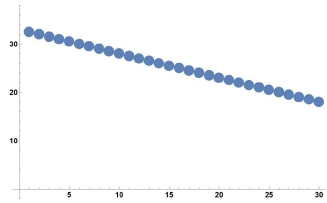
$r = 1$



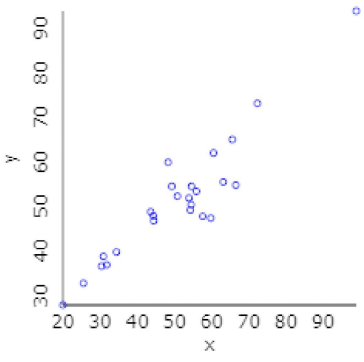
$r = -1$



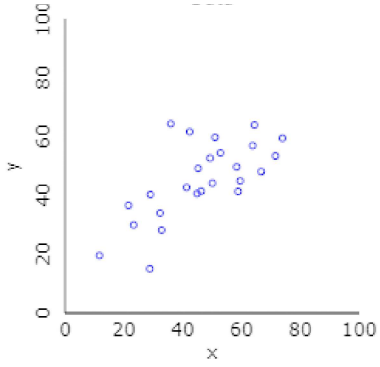
$r = 1$



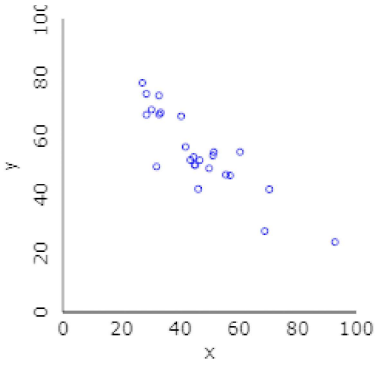
$r = -1$



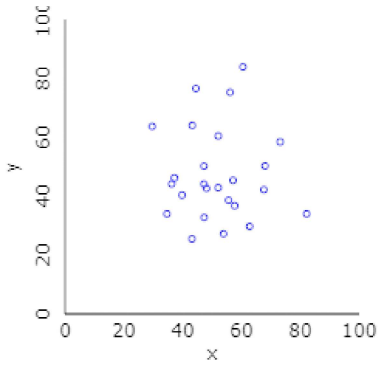
$r = 0.934$



$r = 0.684$

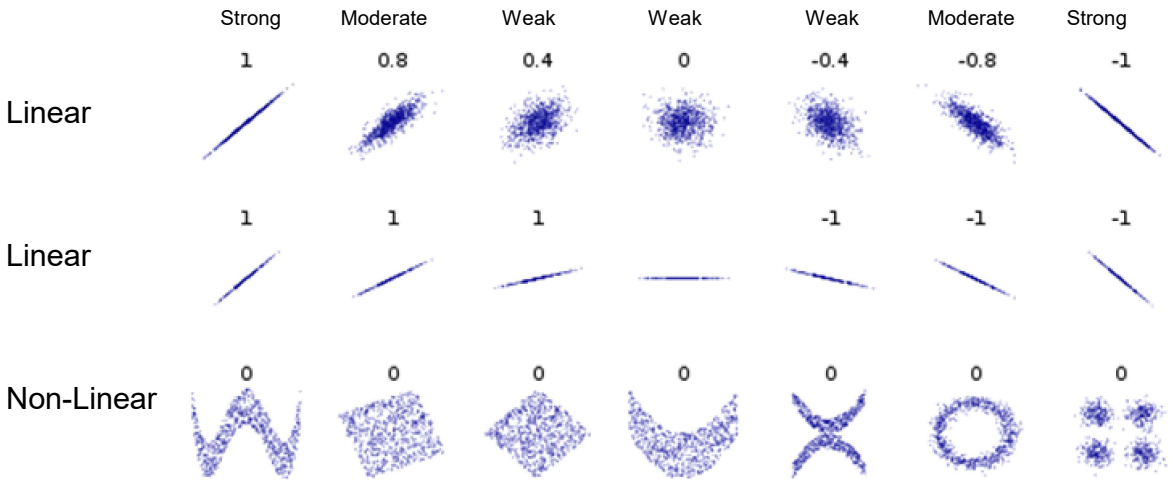


$r = -0.867$



$r = -0.035$

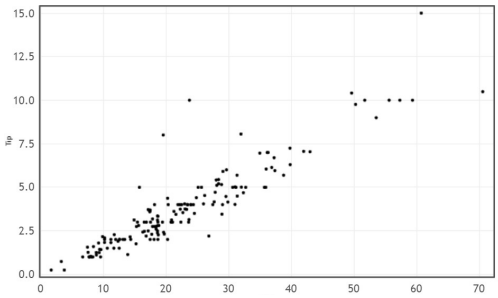
In case you wondered: $r := \frac{1}{n-1}\sum_{i-1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_i}\right)$. (!!!)

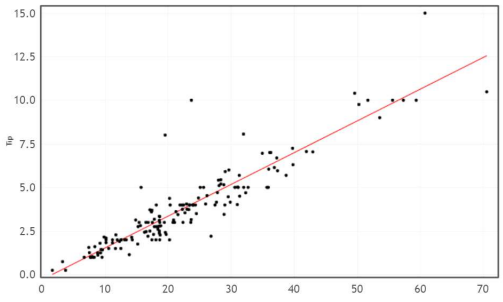


r values

Least Squares Regression Line

Bill vs Tip





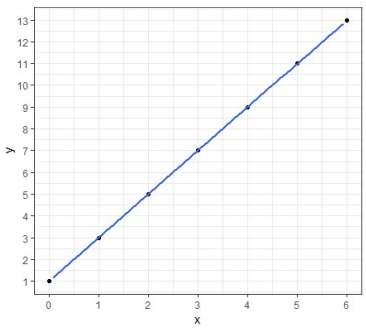
Fitting a Line

Suppose we have two vars: x - explanatory var, and y - response var.

And suppose the scatterplot indicates a linear relationship.

How to describe this relationship with a line?

Line Review



$y = 1 + 2x$

What's the height of the line (y) at $x = 2$?

$y = 1 + 2(2)$

$y = 1 + 4$

$y = 5$

Equation of a line describes the relationship between x & y.

For any x value, we can find matching y value.

Vocab

Equation for a line has form: $y = a + bx$.

Solutions to a line are pairs of numbers, denoted (x,y).

b is the *slope*; it describes how much y changes each time x increases by 1.

a is the *y-intercept*. The line's y-value when $x = 0$.

If two equations with this form differ from each other, they describe different lines.

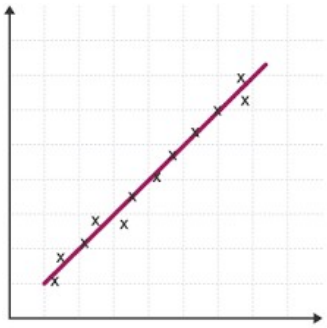
Fitting

Fitting a line to a graph, means finding the "best" line equation.

Let's fit a line to a graph (follow QR code)

Graph compares foot length (in cm) to height (in inches).

How to define "best"?



fitting line to data

bit.ly/introstatsdata

Applet: Fitting Line to Data

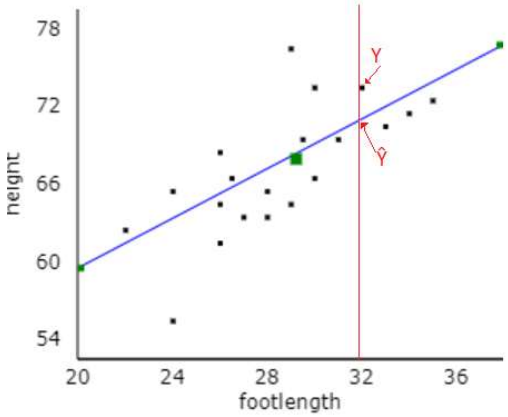
Let's suppose we have pts (**observations**) (x,y) and we fit a line: $\hat{y} = a + bx$

We use symbol $\hat{y}$, because $\hat{y}$ is height of the **line**, not the pt's height y.

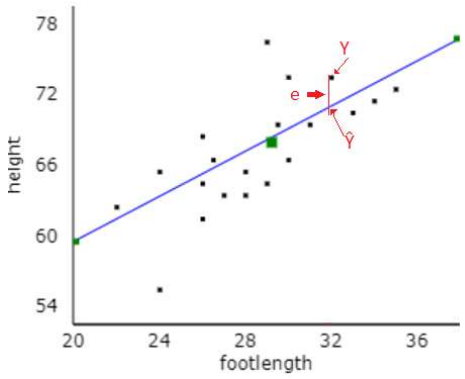
Let $y = 74$ be the pt's value when $x = 32$ (see graph).

$\hat{y} = 70$ is value of fitted line at $x = 32$, called the "fitted value."

What's the "best" line? It mimimize distances between y and $\hat{y}$.



Let's calculate distance (e) for each pt: $e = y - \hat{y}$.



e (error) is also called the **residual** for each observation.

Residual is the observed y minus the fitted (or **predicted**) $\hat{y}$.

Best line should have smallest total residual lengths.

Some residuals are positive - when y is above line.
Some residuals are negative - when y is below line.

To avoid cancelation of pos/neg values, we make all residuals positive. How?

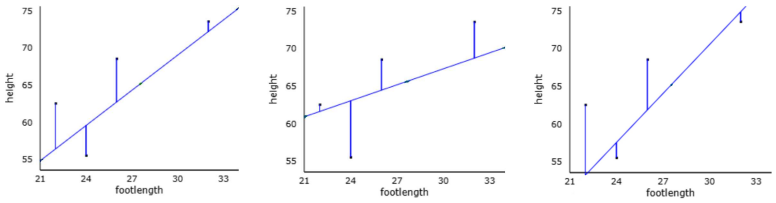
We square the residuals: $e^2 = (y - \hat{y})^2$.

We want the line with smallest **S**um of **S**quared **E**rrors (or residuals), SSE. $(e_1^2 + e_2^2 + \ldots + e_n^2)$

The line that minimizes SSE is called the **Least Squares Regression Line**.

Example: say we're comparing lines y_1, y_2, y_3 as possible fits for some data.

Let them have residuals $\{e_{11}, e_{12}, e_{13}, e_{14}\}$, $\{e_{21}, e_{22}, e_{23}, e_{24}\}$, and $\{e_{31}, e_{32}, e_{33}, e_{34}\}$.



y_1 y_2 y_3

Minimizing the SSE means determining which is least:

$e_{11}^2 + e_{12}^2 + e_{13}^2 + e_{14}^2$, $e_{21}^2 + e_{22}^2 + e_{23}^2 + e_{24}^2$, or $e_{31}^2 + e_{32}^2 + e_{33}^2 + e_{34}^2$.

Least squares line is the "best" line that describes relationship between x/y because it has smallest error.

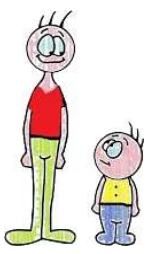
Foot Length & Height

Let x be **foot length** (cm). Let y be **height** (inches).

Least squares regression line: $\hat{y} = 38.30 + 1.03x$.

What does line tell us about relationship between foot length & height?

foot length



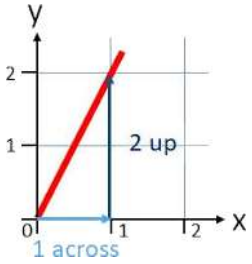
height

Line has two parts: y-intercept and slope.

Slope: 1.03. y-intercept: 38.30.

Interpreting Slope

Slope measures change in $\hat{y}$ for each unit change in x.



e.g., slope of 2

Increase x by 1, then $\hat{y}$ changes by amount of slope.

For $\hat{y} = 38.30 + 1.03x$, $b = 1.03$.

For each additional *cm* that we increase the foot length,
we expect predicted height to increase by 1.03 *inches*.

Interpreting Intercept

y-intercept is the predicted value of $\hat{y}$ when x is at 0.

Won't always make sense, depending on what x is measuring.

For $\hat{y} = 38.30 + 1.03x$, a = ??

$a = 38.30$.

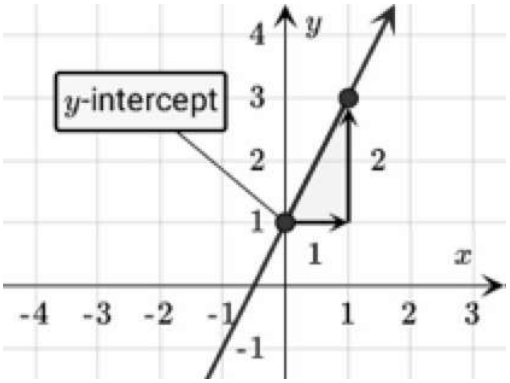


image shows a y-intercept of $a = 1$

This predicts a person with foot length of 0 cms (!?!) will be 38.30 inches tall (3 ft 2.3 inch).
So, when $x = 0$, line $\hat{y}$ predicts y to be a.

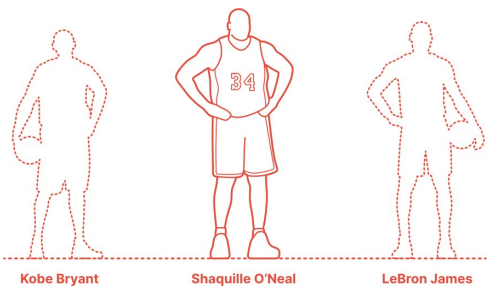
Using a Line to Make Predictions

Least squares regression lines allow us to make predictions for y,
given *unobserved* values of x. (preferably near the observed pts)

For any value x, we find the fitted value for y by plugging x into the line equation.

Foot Length and Height Prediction

Let's predict Shaquille O'Neal's height.
Shaq wears size 22 shoe, which is 41 cms long. Height prediction?

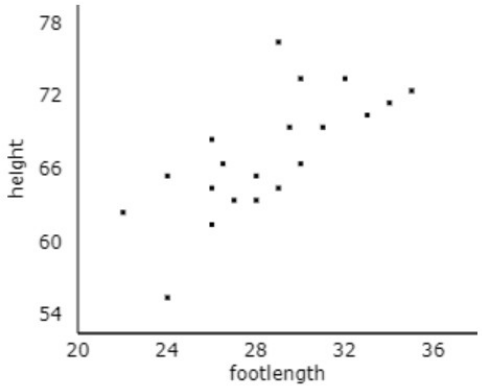


$\hat{y} = 38.30 + 1.03(41) = 38.30 + 1.03(41)$

$= 38.30 + 42.23 = 80.53$ in (or 6 ft and 8.53 in).

Shaq is actually 7 ft 1 in, or 85 in tall. What's the residual for this prediction? ...

$e = y - \hat{y} = 85 - 80.53 = 4.47$ in.



Recall the training data: Note that Shaq's foot length is not in this range.

How well do we expect the least squares regression line to predict the height of someone like Shaq?

Extrapolation

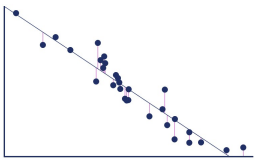
Extrapolation is making a prediction for a value of x well outside of what we've observed.

Predictions that are extrapolated should be viewed with suspicion.

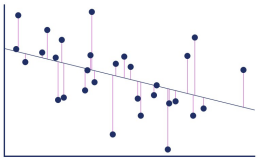
There's no guarantee a relationship that's observed over some range of values continues in same way forever.

Coefficient of Determination (CoD)

What *proportion* of data's *variability* can be explained by it's regression line?



Lots



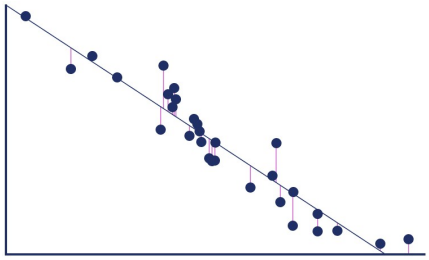
Not so much

The CoD (similar to r) measures how well a regression line fits data.

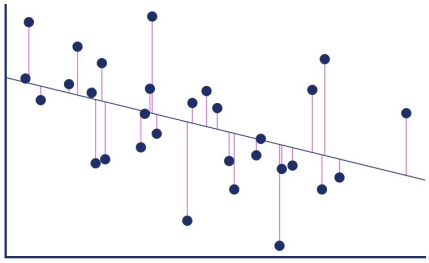
If CoD is near 1, line does good job of predicting y, and is a "good fit."
If CoD is near 0, line does poor job of predicting y, and is a "poor fit."

CoD is calculated as the square of the correlation coefficient: r^2.

Recall: **correlation coefficient** (r) is between –1 and 1, and it describes direction & strength of a linear relationship.



CoD is 0.9



CoD is 0.2

CoD: proportion of variability in y that can be explained by the regression line.

CoD is useful for comparing different **models**
(different models using different choices of explanatory var. Maybe nose-length predicts height better?).

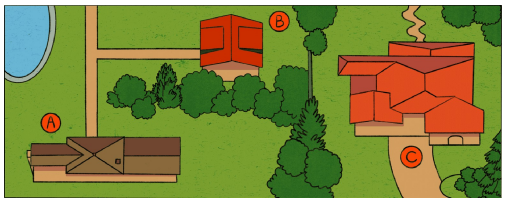
Back to Foot Length & Height: Corr coeff between foot length and height is: $r = 0.71$.

So CoD is: ??

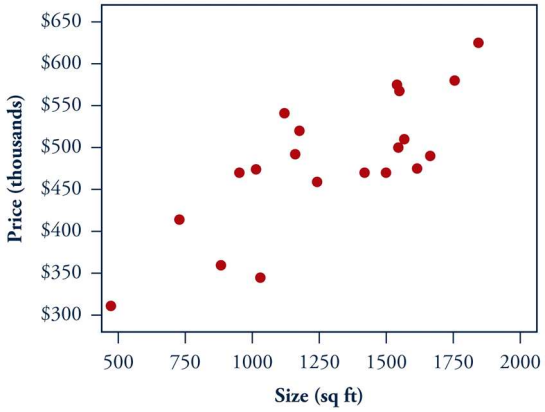
$r^2 = (0.71)^2 = 0.506$. So, ... ?

So, 50.6% of variability in height can be explained by regression line with foot length.
The remainder is explained by other factors (diet, genetics, ... , nose length?).

Linear Regression Example: Home Prices



Understanding the relationship between a home's sale price and sqr footage.



x is **size** (sq ft), y is **price**.

Recall: Positive, moderately strong, linear relationship.

Turns out, correlation coefficient is ??

$r = 0.780$.

If y-intercept/slope is $a = 265,217.8$, and $b = 168.6$, what is the equation for the line?

$\hat{y} = 265,217.8 + 168.6x$

Cost of house with 0 sq ft? (empty plot of land)?

$a = 265,217.80$ ($)

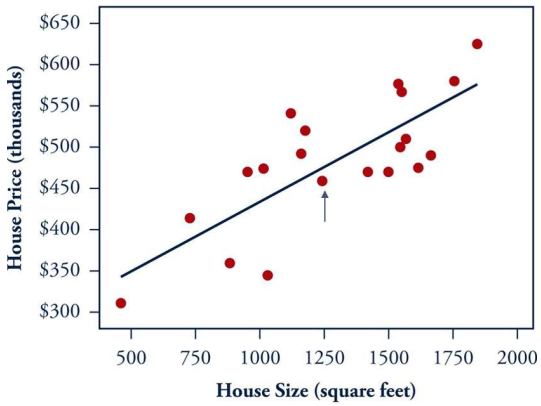
If I add an extra 100 sq ft to my house, the predicted price increases by ?

$100 \times 168.6 = \$16,860$.

Let's predict the price of 845 Pearl Drive: 1,242 sq ft.

$\hat{y} = 265,217.8 + 168.6(1,242) = \$474,613.$

What's the residual for 845 Pearl Dr?



845 Pearl Drive sold for $\$459,000$

$e = y - \hat{y} = \$459,000 - \$474,613 = -\$15,613.$

Regression line is an overestimate or underestimate for this house price?

Overestimate.

If $r = 0.780$, what proportion of variability in housing price can be explained by house size? ...

CoD: $r^2 = (0.780)^2 = 0.6084.$

House size explains 60.84% of variability in house prices. How much variability is explained by other factors?

$1 - 0.6084 = 0.3916.$ So, 39.16%.

If we plan to build a house that has 3000 sq ft, can we use this to predict how much it will sell for?

Activity: 27-3

What did we learn?

- Correlation Coefficient (r)
- Fit least squares line to data by minimizing SSE: $e^2 = (y - \hat{y})^2$
- Interpret slope.
- Interpret intercept
- Make prediction
- Calculate residual
- Be careful of extrapolations.
- Assess fit using CoD: r^2
- Interpret CoD: r^2