Introduction to Statistics I

Textbook: Elementary Statistics (4th Edition, by Navidi and Monk), and Workshop Statistics (4th Edition by Rossman and Chance).

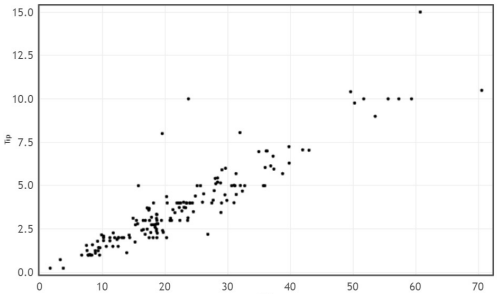
Previous Lecture

- Graphically display relationships between two quantitative vars
- Scatterplots: Direction, Strength, Form
- Correlation Coefficient (r)

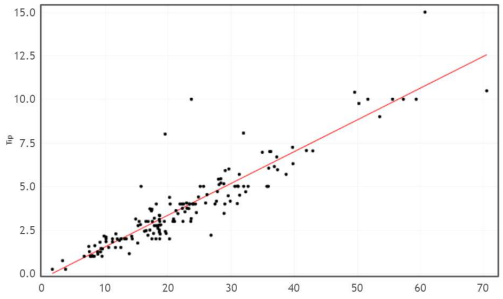
§4.2: Least Squares Regression Line

Least Squares Regression Line

Bill vs Tip



Bill



Bill (w/regression line)

Fitting a Line

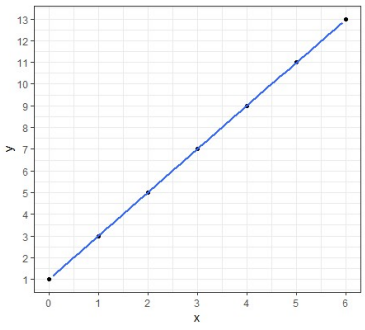
Suppose we have two vars: x - explanatory var, and y - response var.

And suppose the scatterplot indicates a linear relationship.

How to describe this relationship with a line?

Line Review

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



$y = 1 + 2x$

$y = 1 + 2(2)$

$y = 1 + 4$

$y = 5$

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

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

Vocab

The *Equation* for every line can take the form: $y = b_0 + b_1 x$.

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

b_1 is the *slope*; it describes how much y changes each time x increases by 1.

b_0 is the *y-intercept*. It is the line's y-value when $x = 0$.

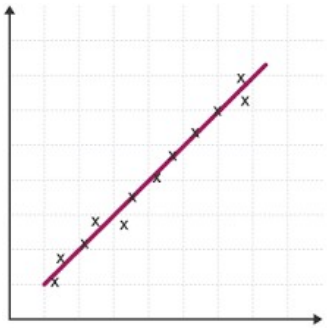
If two equations with this form have a different value for b_0 or b_1, they describe different lines.

Fitting

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

Let's fit a line to a graph (go to GoogleDoc, Applets: Fitting Line to Data)

The graph compares foot length (in cm) to height (in inches).
How to define "best"?



fitting line to data

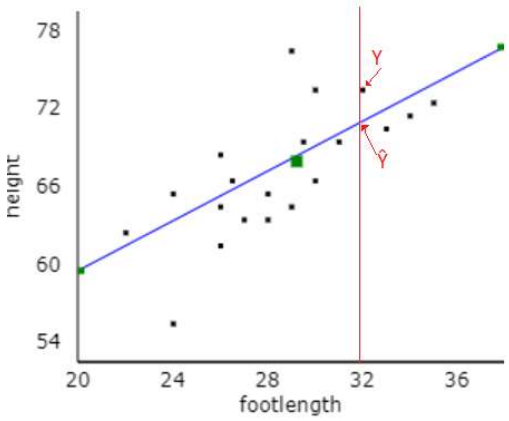
bit.ly/introstatsdata

Applets: Fitting Line to Data

Let's suppose we have pts (**observations**) (x,y) and we fit a line: $\hat{y} = b_0 + b_1 x$.
We use symbol $\hat{y}$, because $\hat{y}$ is height of the **line**, not a data 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 *mimimizes* distances between y and $\hat{y}$.

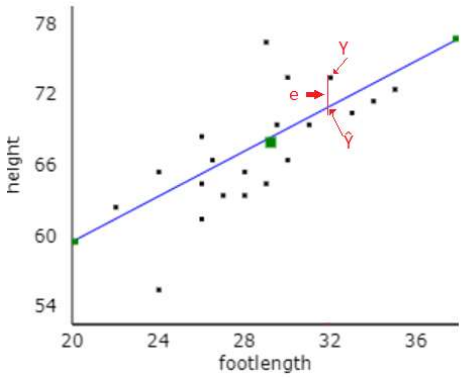


We calculate distance (e) for each pt as: $e = y - \hat{y}$.

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

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

The best line should have smallest total residual length.



Some residuals are positive - when y is above line.

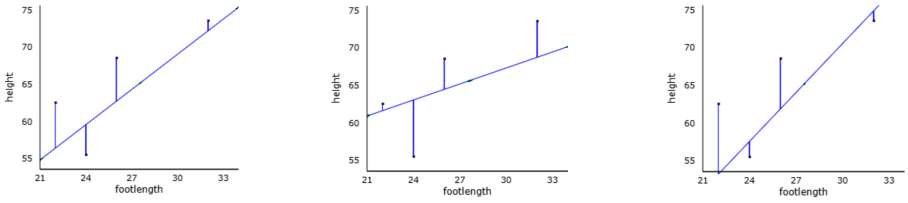
Some residuals are negative - when y is below line.

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

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

Thus, we want the line with smallest **sum of squared residuals**. $(e_1^2 + e_2^2 + \ldots + e_n^2)$

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



y_1 y_2 y_3

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

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

Minimizing the sum of squared residuals means determining which is least:

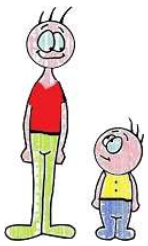
$e_{11}^2 + e_{12}^2 + e_{13}^2 + e_{14}^2$, $e_{21}^2 + e_{22}^2 + e_{23}^2 + e_{24}^2$, or $e_{31}^2 + e_{32}^2 + e_{33}^2 + e_{34}^2$.

The least squares regression line is the "best" line that describes relationship between x/y because it has smallest residuals.

Foot Length & Height

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

foot length



height

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

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

A line has two parts: y-intercept and slope.

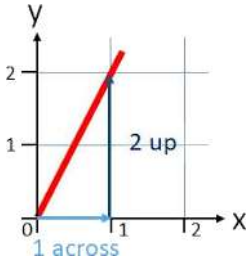
Slope: 1.03. y-intercept: 38.30.

Interpreting Slope

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

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

For $\hat{y} = 38.30 + 1.03x$, $b_1 = 1.03$.



e.g., slope of 2

prediction

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

Interpreting Intercept

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

It won't always make practical sense, depending on what x is measuring.

For $\hat{y} = 38.30 + 1.03x$, $b_0 = ??$

$$b_0 = 38.30.$$

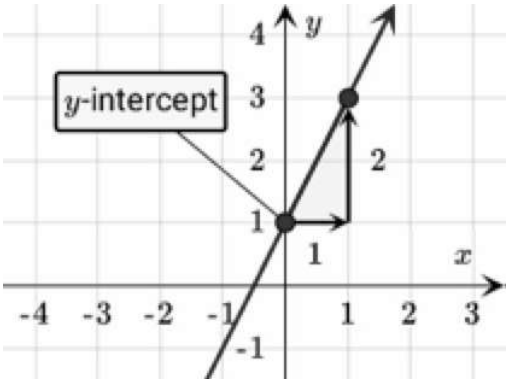


Image shows a y-intercept of $b_0 = 1$

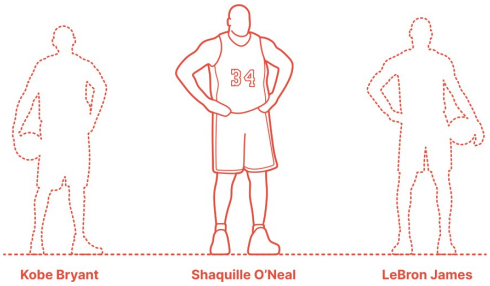
prediction

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 b_0.

Using a Line to Make General Predictions

Least squares regression lines allow us to make predictions for y,
given *unobserved* values of x.

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 for a 41 cms long foot. Height prediction?

prediction

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

$$= 38.30 + 42.23 = 80.53 \text{ in (or 6 ft and 8.53 in).}$$

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

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

Our prediction was a bit off, can you speculate why our regression line performed poorly?

What did we learn?

- Calculate residual $e = y - \hat{y}$ between data and regression line $\hat{y}$
- Fit *least squares line* to data by minimizing SSE: $e_{11}^2 + e_{12}^2 + \ldots + e_n^2$
- Interpret slope/intercept
- Make predictions