

13.6 - Linear regression
(sometimes called "best fit lines")

Present Thursday:

8.2 #47c, and 8.3 #41, 70*

HW 13.6 #1-3, 11, 12*, 13

(* select 'best fit line' by eye; choose points to find m and b)

HW 13.6 #21-24, 25-28

(use a computer / calculator to find the least squares regression line and the coefficient of correlation r).

Overheard: "75% of all statistics are made up on the spot."

No Homework Presentations Monday.
(Project 2 Due / Presentations)

After reviewing the correct solution (below), write your score on the back of your quiz.

0 = no progress at all; just rewrote problem

0.5 = false start, not based on relevant principles

1 = false start, but sustained effort with some relevant principles

1.5 = significant mistake(s), or significant misunderstanding(s)

2 = mistake near the end or could not finish; also excessive reliance on calculator or 'brute force' methods

2.5 = trivial mistake (e.g. arithmetic error), but work is mostly correct

3 = correct answer and work

Gina and Harriet compete in the 110 meter hurdles. Gina's average time in her last 10 races was 18.4 seconds. Harriet's average time was 17.2 seconds. Suppose the appropriate standard deviation for the confidence interval is 0.5 seconds.

(a) Complete the following confidence statement: We can be 90% confident that the true difference between Gina's time and Harriet's time is... (between 0.38 sec and 2.02 sec.)

Avg difference = 1.2 sec = \bar{x} .

look up z in Table 10, using $A = 0.45$. We find $z \approx 1.645$.

$\bar{x} - (z)(s) = 1.2 - 1.645(0.5) \approx 0.38$ sec. $\bar{x} + (z)(s) = 1.2 + 1.645(0.5) \approx 2.02$ sec.

(b) Is there sufficient evidence to conclude that Harriet is faster than Gina? Explain briefly.

Yes, because the confidence interval does not include 0 sec.

Statistics and Probability

8.SP

- Construct and interpret scatter plots for bivariate measurement data to investigate patterns of association between two quantities. Describe patterns such as clustering, outliers, positive or negative association, linear association, and nonlinear association.
- Know that straight lines are widely used to model relationships between two quantitative variables. For scatter plots that suggest a linear association, informally fit a straight line, and informally assess the model fit by judging the closeness of the data points to the line.
- Use the equation of a linear model to solve problems in the context of bivariate measurement data, interpreting the slope and intercept. For example, in a linear model for a biology experiment, interpret a slope of 1.5 cm/hr as meaning that an additional hour of sunlight each day is associated with an additional 1.5 cm in mature plant height.
- Understand that patterns of association can also be seen in bivariate categorical data by displaying frequencies and relative frequencies in a two-way table. Construct and interpret a two-way table summarizing data on two categorical variables collected from the same subjects. Use relative frequencies calculated for rows or columns to describe possible association between the two variables. For example, collect data from students in your class on whether or not they have a curfew on school nights and whether or not they have assigned chores at home. Is there evidence that those who have a curfew also tend to have chores?

Finding a Linear Model

Here are data on SES and WKCE from four Wisconsin schools.

Percent Free & Reduced Lunch: 31 54 21 14
Percent Above Proficient on WKCE: 68 45 72 81

1. Find a linear model for this data (choose points thoughtfully).

$$m = \frac{\Delta y}{\Delta x} = \frac{68-45}{31-54} = \frac{23}{-23} = -1$$

$$y = -1x + b \text{ (I'll use } (31, 68)\text{):}$$

$$68 = -1(31) + b \Rightarrow b = 99$$

2. Interpret the slope as a rate of change.

"For every 1 point increase in the FRL rate, the percent proficient..." the % passing falls by 1 percent.

Finding a Linear Model

Here are data on SES and WKCE from four Wisconsin schools.

Percent Free & Reduced Lunch: 31 54 21 14

Percent Above Proficient on WKCE: 68 45 72 81

3. Predict the proficiency level at a school with a 40% FRL rate.

(Un)related question: should teacher compensation be tied to student achievement on standardized tests?

4. Predict the WKCE proficiency level in an affluent district where no students qualify for a free or reduced lunch.

(Is the y-intercept meaningful in this case?)

Statistics and Probability

8.SP

3. Use the equation of a linear model to solve problems in the context of bivariate measurement data, interpreting the slope and intercept. For example, in a linear model for a biology experiment, interpret a slope of 1.5 cm/hr as meaning that an additional hour of sunlight each day is associated with an additional 1.5 cm in mature plant height.

Each student is given a bean sprout to care for, and each decides how many hours per day to let their plant out of the dark closet where they are kept. Plants are measured once at the end of 8 weeks.

A linear model obtained from the class data might be:

$$y = 1.5x + 4.5,$$

where y is the final height in centimeters and x is the number of hours of daylight the plant received per day.

Interpret the slope, and interpret the y-intercept.

Statistics and Probability

8.SP

3. Use the equation of a linear model to solve problems in the context of bivariate measurement data, interpreting the slope and intercept. For example, in a linear model for a biology experiment, interpret a slope of 1.5 cm/hr as meaning that an additional hour of sunlight each day is associated with an additional 1.5 cm in mature plant height.

Each student is given a bean seed to plant and care for. The bean plants are placed near a window and are measured once each week for 8 weeks.

A linear model obtained from the class data might be:

$$y = 3x - 2.4,$$

where y is the height in centimeters after x weeks of growth.

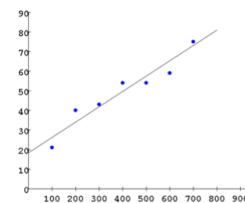
Interpret the slope, and interpret the y-intercept.

Often, it is sufficient to just estimate the 'best fit line' by eye.

1. Choose a couple of points to use to compute the slope.

(Choose thoughtfully).

2. Using the value for m and any appropriate point (x, y) , use algebra to solve the equation $y = mx + b$ for b .



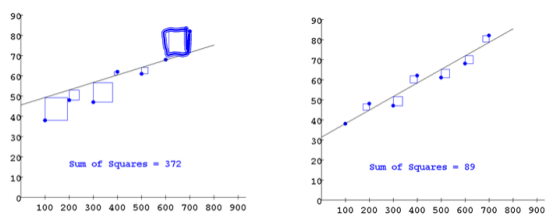
<<http://hspm.sph.sc.edu/Courses/J716/demos/LeastSquares/LeastSquaresDemo.html>>

Least Squares Regression

The most commonly used linear regression line (or "best fit line") is found by minimizing the sum of the squares of the vertical distances (the residuals) from each point to the line.

The resulting line is called the least squares regression line (LSR).

<<http://hspm.sph.sc.edu/Courses/J716/demos/LeastSquares/LeastSquaresDemo.html>>

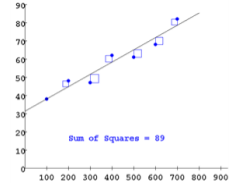


Least Squares Regression Calculations

The coefficients a and b of the least squares regression line $y' = ax + b$ can be computed from the (x,y) data points using the following formulas:

$$a = \frac{n \sum xy - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \frac{\sum y - a(\sum x)}{n}$$

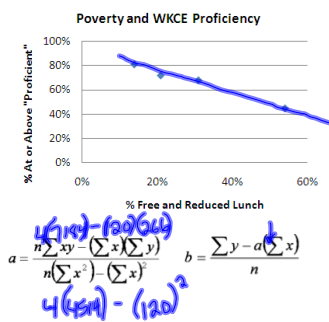


Their derivation is beyond the scope of this course, but applying them is a good exercise in mathematical literacy!

Least Squares Regression Calculations

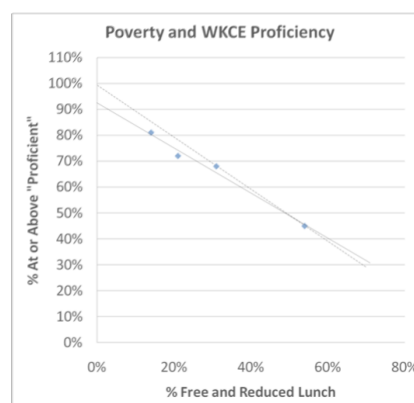
Percent Free & Reduced Lunch: 31 54 21 14
 Percent Scoring Above Proficient: 68 45 72 81

x	y	xy	x ²
31	68	2108	961
54	45	2430	2916
21	72	1512	441
14	81	1134	196
120	266	7184	4514



Use Excel to find the LSR (in Excel, it's a linear "trendline")

Comparing linear models: "Eye" vs. "LSR"



Handwritten red equations:

$$y = -1x + 99$$

$$y = -.87x + 92$$

The coefficients a and b of the least squares regression line $y' = ax + b$ can be computed from the (x,y) data points using the following formulas:

$$a = \frac{n\sum xy - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \quad b = \frac{\sum y - a(\sum x)}{n}$$

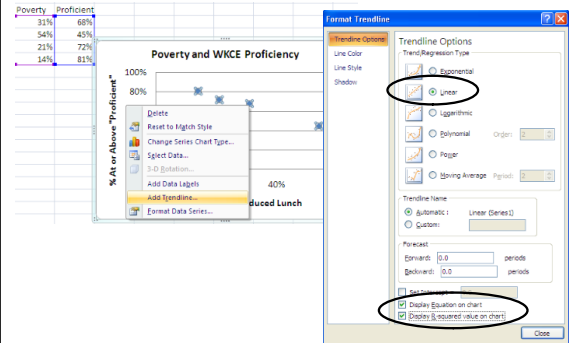
We can use a graphing calculator to greatly simplify this work!

On a TI-83:

1. Enter the data in L1 and L2. Use [STAT][Edit...]
2. To get two-variable statistics, use [STAT][CALC][2-Var Stats]
3. Or, to just go straight to the regression line, use [STAT][CALC][LinReg(ax+b)]

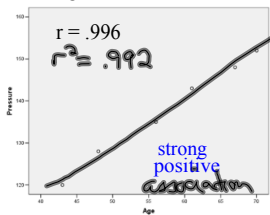
Note: To see the value of r as part of the [LinReg(ax+b)] output, you may have to go to the [CATALOG] and run the [DiagnosticsOn] command.

Microsoft Excel (and other programs) can calculate regression lines too! Just create a scatterplot of the xy data, right-click on the points on the graph, and turn on the Trendline.

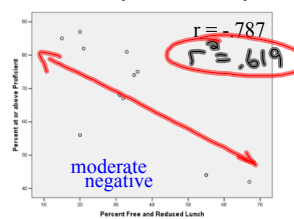


Examples of Correlation Coefficients

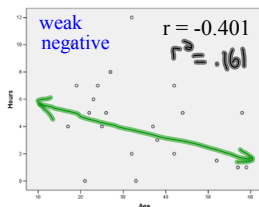
Age and Blood Pressure



Poverty & Proficiency



Age & Hours of Exercise/Week



positive
negative

strong
moderate
weak

The value of r may range from -1 to 1 . It is a measure of how well the linear model fits the data set. The closer to 1 or -1 , the better the fit!