Math 243

4.2: Least-Squares Regression

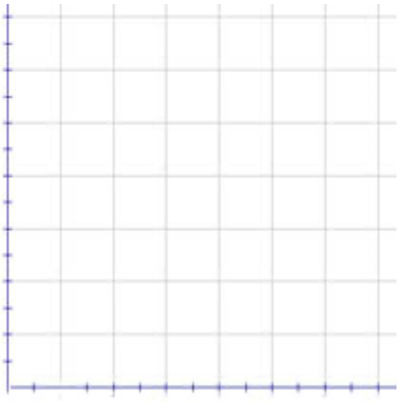
Finding the Least-Squares regression line.

A little Algebra Line Review:

The Slope-Intercept form of a linear equation is
$y = mx + b$ where
$m =$ slope = $\frac{rise}{run}$
$b =$ y-intercept.

Try It:

a) Graph $y = \frac{1}{2}x + 2$



b) Find the slope of the line that passes through the points (2,6) and (5,0)

Now Find the equation of the line that passes through these points.

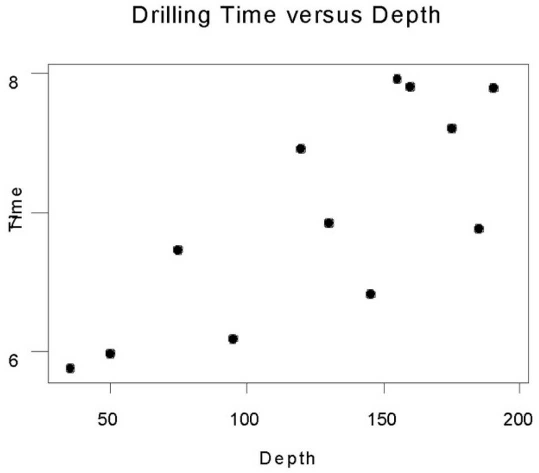
Method 1: using the slope-intercept form $y = mx + b$

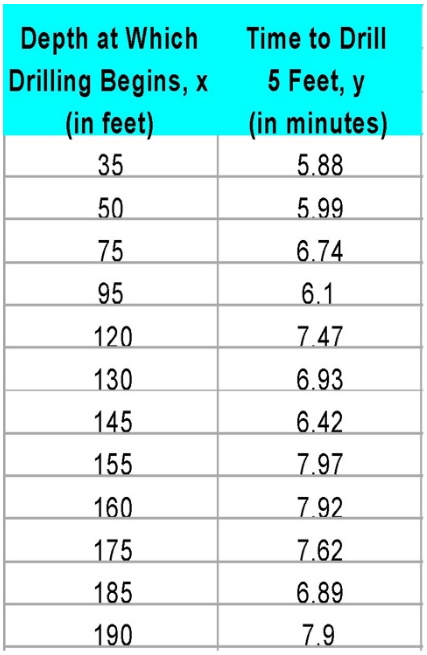
Method 2: using the point-slope form $y - y_1 = m(x - x_1)$

Notice that in part (a) we started with an equation and ended up with a line. Our goal in statistical "regression" is to reverse this and find the equation of the "best line" from the coordinates of thei points.

Lines from Data Points

Consider the scatter plot of the drilling data below. There is a slight positive correlation among the points. Draw a line through the middle of the points that passes through 2 points, then find the equation of this line.





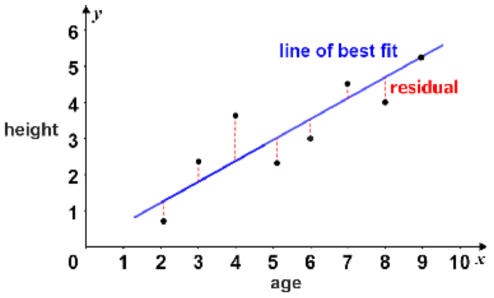
Depth at Which Drilling Begins, x (in feet)	Time to Drill 5 Feet, y (in minutes)
35	5.88
50	5.99
75	6.74
95	6.1
120	7.47
130	6.93
145	6.42
155	7.97
160	7.92
175	7.62
185	6.89
190	7.9

Finding the "Best" Line for the data

Our goal is to find the equation of a line that is the closest to all the points *on average.*

If $f(x) = mx + b$ is the line of best fit for some points, then the **Residual** for a given point is

$$Residual = Observed\ y - predicted\ y$$



Try it: Use the equation we found above to find the residual for the point (95,6.1) in the drilling times graph.

The Least-Squares Regression line On the TI-84

Try It: Use your Calculator to find the equation of the linear regression model for the drilling data.

Key Steps:

- [STAT] → [Edit]
- Enter data into L1 and L2
- [STAT]→[CALC]
- Choose LinReg{ax+b}
- Store this in equation Y_1 by selecting $[2nd] \rightarrow [Var] \rightarrow [Y - vars] \rightarrow Y_1$ **before hitting enter**. This will store your equation into Y_1

If you want to see the graph, do these steps

- [2nd]→[StatPlot]→ [Plot 1] → turn this on and select scatterplot
- [Zoom]→[ZoomStat]

a. What is the equation for the Linear regression model?

b. What is the correlation coefficient?

c. Is the linear regression model a good fit for the data?

d. How long should the drilling take if it starts at 100 ft.?

e. If the drilling took 7 minutes, approximate the depth it started at.

The Math Behind the Calculator

This line will minimize the sum of the squared residuals.

Here is the formula used to find it (you don't need to memorize this)

$$\hat{y} = mx + b$$
$$m = r\left(\frac{s_y}{s_x}\right)$$
$$b = \bar{y} - m\bar{x}$$

$$r = \frac{\Sigma\left[\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)\right]}{n-1}$$

$\hat{y}$= regression model equation; s_y= Standard deviation y-values;
s_x= Standard deviation x-values; $\bar{y}$ and $\bar{x}$ are the means of the x's and y's
Note: $\hat{y}$ is read "y-hat" and $\bar{y}$ is read "y-bar"