# Lecture Notes & Examples 3.2 Part A and Part B

## Section 3.2 - Least-Squares Regression

In the previous section we examined scatterplots for linear relationships. Correlation measures the direction and strength of these relationships. When the plot shows a linear relationship, we would like to summarize the overall pattern by drawing a line on the scatterplot. This is called a **Regression Line**. In order to do this we must have an *explanatory* and a *response variable*.

**Definition:** A **Regression line** is a line that describes how a response variable y changes as an explanatory variable x changes. We often use a regression line to predict the value of y for a given value of x.

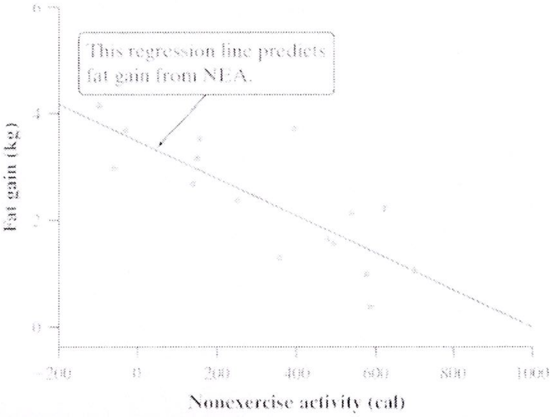
**Example – Does Fidgeting Keep You Slim? Regression Lines as Models.**

Some people don't gain weight even when they overeat. Perhaps fidgeting and other "nonexercise activity" (NEA) explains why—some people may spontaneously increase nonexercise activity when fed more. Researchers deliberately overfed 16 healthy young adults for 8 weeks. They measured fat gain (in kilograms) as the response variable and change in energy use (in calories) from activity other than deliberate exercise—fidgeting, daily living, and the like—as the explanatory variable. Here are the data:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| NEA change (cal): | −94 | −57 | −29 | 135 | 143 | 151 | 245 | 355 |
| Fat gain (kg): | 4.2 | 3.0 | 3.7 | 2.7 | 3.2 | 3.6 | 2.4 | 1.3 |
| NEA change (cal): | 392 | 473 | 486 | 535 | 571 | 580 | 620 | 690 |
| Fat gain (kg): | 3.8 | 1.7 | 1.6 | 2.2 | 1.0 | 0.4 | 2.3 | 1.1 |

Do people with larger increases in NEA tend to gain less fat? The figure below is a scatterplot of these data. The plot shows a moderately strong, negative linear association between NEA change and fat gain with no outliers. The correlation is r= −0.7786. The line on the plot is a regression line for predicting fat gain from change in NEA

$\widehat{\text{fat gain}} = f(\text{NEA})$



1. **Interpreting a Regression Line** – A regression line is a *model* for the data, much like the density curves we considered in Chapter 2. It gives a compact mathematical description of the relationship between the response variable y and the explanatory variable x.

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

In this equation:

$\hat{y}$ (read y-hat) is the **predicted value** of the response variable y for a given value of the explanatory variable x.

b is the slope (rate of change), the amount by which y is *predicted* to change when x increases by one unit.

a is the y-intercept, the *predicted* value of y when x = 0.

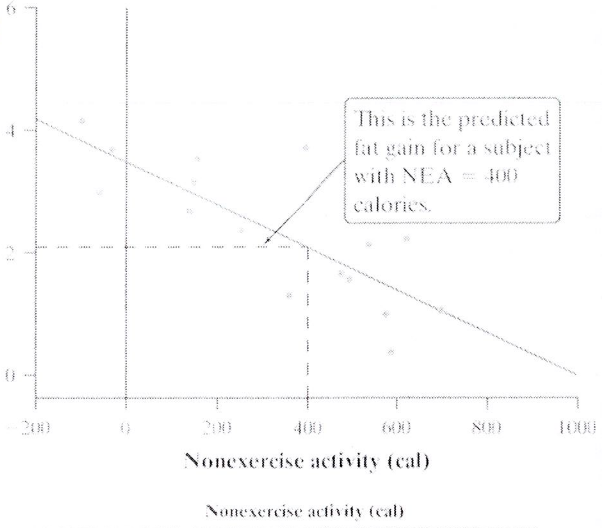
**Note: on the AP Exam formula sheet the regression equation is written $\hat{y} = b_0 + b_1x$ .**
**(Regardless of notation, the coefficient with x is the slope.)**

$\hat{y} = a + bx$

**Example:** The regression line for the figure to the right is $\widehat{fat\ gain}$ = 3.505 – 0.00344(NEA change)
Identify the slope and y intercept of the regression line. $b: \frac{\Delta \text{ fat gain kg}}{\Delta \text{ NEA cal}}$
Interpret each value in context.

slope b: -0.00344 tells us amt of fat gained is predicted to go down by -0.00344 kg for each added calorie of NEA.

y intercept a= 3.505 kg ; this is the fat gain estimated if NEA does not change (is 0 cal) when a person overeats.



1. **Prediction** – The regression line can be used to predict the response variable $\hat{y}$ for a specific value of the explanatory variable x.

**Example: Predict the fat gain if a person's NEA increases by $\overset{x}{400}$ calories.**

$\widehat{\text{fat gain}} = 3.505 - 0.00344\ (400) \approx 2.13$ kg

**Predict the fat gain for someone whose NEA increases by 1500 calories?**

$\widehat{\text{fat gain}} = 3.505 - 0.00344\ (1500) = -1.66$ kg

Looking at figure above, an NEA increase of 1500 calories is far outside the set of x values for our data. We can't say whether increases this large ever occur, or whether the relationship remains linear at such extreme values.
Predicting fat gain when NEA increases by 1500 cal is an extrapolation of the relationship beyond what the data show.

**Definition: Extrapolation** is the use of a regression line for prediction far outside the interval of values of the explanatory variable x used to obtain the line. Such predictions are often not accurate.

**Don't make predictions using values of x that are much larger or much smaller than those that actually appear in your data.**

**CHECK YOUR UNDERSTANDING**

Some data were collected on the weight of a male white laboratory rat for the first 25 weeks after its birth. A scatterplot of the weight (in grams) and time since birth (in weeks) shows a fairly strong, positive linear relationship. The linear regression equation $\widehat{weight} = 100 + 40(time)$models the data fairly well.

$m = \frac{\Delta \text{weight (g)}}{\Delta \text{time (weeks)}}$

1. What is the slope of the regression line? Explain what it means in context.

Slope is 40. We predict that the rat will gain 40 grams of weight per week.

2. What's the y intercept? Explain what it means in context.

y intercept is 100. We expect the birth rate of the rat to be 100grams.

3. Predict the rat's weight after 16 weeks. Show your work.

$\widehat{weight} = 100 + 40(16) = 740$ grams.

4. Should you use this line to predict the rat's weight at age 2 years? Use the equation to make the prediction and think about the reasonableness of the result. (There are 454 grams in a pound.)

2 years = 2(52) = 104 weeks.

$\widehat{weight} = 100 + 40(104) = \frac{4,260 \text{ g}}{1} \cdot \frac{1 \text{ pound}}{454 \text{ g}} \approx 9.4$ pounds

This would be a very large rat! This is unreasonable and is the result of extrapolation.

**3. Residuals and the Least-Squares Regression Line** – In most cases, no line will pass exactly through all the points in a scatterplot. The predicted values (y-hat) will not be the actual values of the response variable y. *A good regression line makes the vertical distance between the actual points from the line as small as possible*. Look at the following example describing the relationship between body weight and backpack weight for a group of 8 hikers.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Body weight (lb): | 120 | 187 | 109 | 103 | 131 | 165 | 158 | 116 |
| Backpack weight (lb): | 26 | 30 | 26 | 24 | 29 | 35 | 31 | 28 |

The figure below shows a scatterplot of the data with a regression line added. The prediction errors are marked as bold segments in the graph. These vertical deviations represent "leftover" variation in the response variable after fitting the regression line. For that reason, they are called residuals.

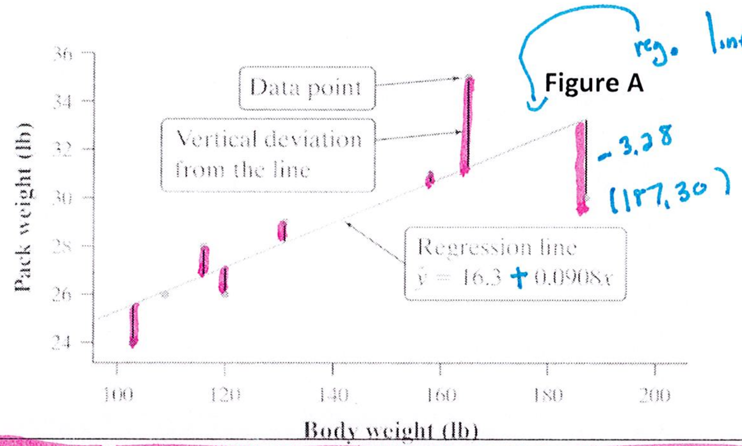


**Figure A**

reg. line over predicts pack weight at 187 pound body weigh

**Definition:** A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line. That is

$$\text{Residual} = \text{Observed } y - \text{Predicted } y = y - \hat{y}$$

**Example**: Find and interpret the residual for the hiker who weighed 187 pounds.

$\widehat{\text{Pack weight}} = 16.3 + 0.0908(187) = 33.28$ pounds

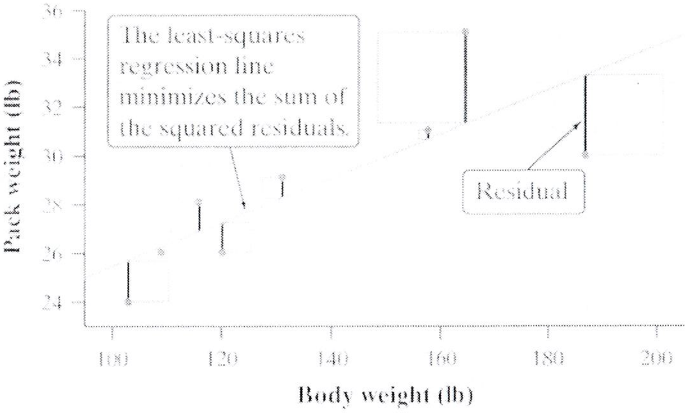
Residual $= 30 - 33.28 = -3.28$ pounds

Regression line over predicts this hiker's pack weight by 3.28 pounds

**AP EXAM TIP** There's no firm rule for how many decimal places to show for answers on the AP exam. Our advice: Give your answer correct to two or three nonzero decimal places. Exception: If you're using one of the tables in the back of the book, give the value shown in the table.

The line shown in **Figure A** above makes the residuals for the 8 hikers "as small as possible." But what does that mean? Maybe this line minimizes the sum of the residuals. Actually, if we add up the prediction errors for all 8 hikers, the positive and negative residuals cancel out. That's the same issue we faced when we tried to measure deviation around the mean. We'll solve the current problem in much the same way: by squaring the residuals. The regression line we want is the one that minimizes the sum of the squared residuals. That's what the line shown in the above figure does for the hiker data, which is why we call it the least-squares regression line.

**Least-squares regression line** - The least-squares regression line of y on x is the line that makes the sum of the squared residuals as small as possible. The figure at the right gives a geometric interpretation of the least-squares idea for the hiker data. The least-squares regression line shown minimizes the sum of the squared prediction errors, 30.90. No other regression line would give a smaller sum of squared residuals



Pack weight = 16.3 + 0.0908 Body weight; $r^2 = 0.63$
Sum of squares = 30.90

**Technology** – The least-squares regression line can be found by using your graphing calculator. Details are listed
on p. 171 of the text

**CHECK YOUR UNDERSTANDING**

It's time to practice your calculator regression skills. Using the familiar hiker data in the table below, calculate the least-squares regression line on your calculator. You should get as the equation $\hat{y} = 16.3 + 0.0908x$of the regression line.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $L_1$ | Body weight (lb): | 120 | 187 | 109 | 103 | 131 | 165 | 158 | 116 |
| $L_2$ | Backpack weight (lb): | 26 | 30 | 26 | 24 | 29 | 35 | 31 | 28 |

$\boxed{\text{STAT}}$ → Calc #8 (Linear Reg (a+bx)) $L_1$, $L_2$, y

$\hat{y} = 16.3 + .0908x$

r: .795

only need y if want to graph regression line on scatterplot

↙ $\boxed{\text{VARS}}$ → YVARS #1: Function
#1: $y_1$

$\boxed{\text{Zoom}}$ #9: STAT
To Show Scatterplot with regression line

**4. Regression to the Mean**

We can use technology to find the equation of the least-squares regression line, and it is possible to calculate the equation of the least-squares regression line using only the means and standard deviations of the two variables and their correlations.

**Definition:** **Equation of the least-squares regression line**

We have data on an explanatory variable *x* and a response variable *y* for *n* individuals. From the data, calculate the means and standard deviations of the two variables and their correlation. The least squares regression line is the line $\hat{y} = a + bx$ with

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

**And y-intercept** $a = \bar{y} - b\bar{x}$

**AP EXAM TIP** The formula sheet for the AP exam uses different notation for these equations:

$b_1 = r\frac{s_y}{s_x}$ and $b_0 = \bar{y} - b_1\bar{x}$

That's because the least-squares line is written as $\hat{y} = b_0 + b_1 x$.

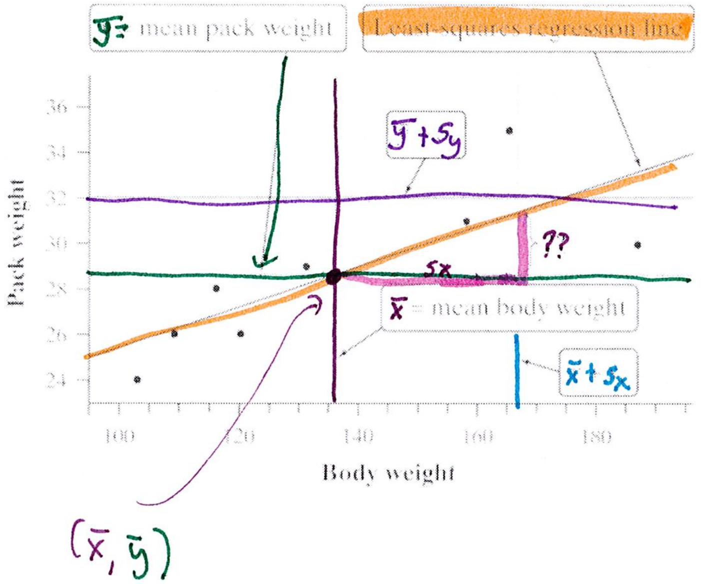
We prefer our simpler versions without the subscripts.

**What does the slope of the least-squares line tell us?**

The figure below shows the regression line in black for the hiker data.
We have added four more lines to the graph:

- a vertical line at the mean body weight $\bar{x}$
- a vertical line at $\bar{x}$ + $s_x$ (one standard deviation above the mean body weight)
- a horizontal line at the mean pack weight $\bar{y}$
- a horizontal line at $\bar{y}$+ $s_y$ (one standard deviation above the mean pack weight)

- Note that the regression line passes through ($\bar{x}$ , $\bar{y}$) as expected.



From the graph, the slope of the line is:

$$b = \text{slope} = \frac{\text{change in y}}{\text{change in x}} = \frac{??}{s_x}$$

From the definition box, we know that the slope is

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

Setting the two formulas equal to each other, we have

$$b = r\frac{s_y}{s_x} = \frac{??}{s_x} \Rightarrow ?? = r\cdot s_y$$

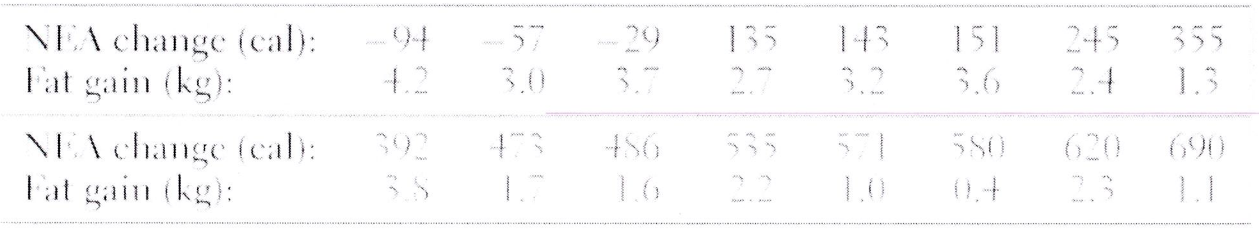
So, the unknown distance **??** above must be equal to r · $s_y$. In other words, **for an increase of one standard deviation in the value of the explanatory variable x, the least-squares regression line predicts an increase of $\underbrace{\text{r standard deviations}}_{r.s_y}$ in the response variable y**

There is a close connection between correlation and the slope of the least-squares line. The slope is

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

This equation says that along the regression line, **a change of one standard deviation in x corresponds to a change of r standard deviations in y.** When the variables are perfectly correlated (r = 1 or r = −1), the change in the predicted response $\hat{y}$ is the same (in standard deviation units) as the change in x. Otherwise, because $-1 \leq r \leq 1$, the change in $\hat{y}$ is less than the change in x. **As the correlation grows less strong, the prediction moves less in response to changes in x.**

**Example – Fat Gain and NEA Calculating the least-squares regression line**
Refer to the data from the example below:



| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| NEA change (cal): | −94 | −57 | −29 | 135 | 143 | 151 | 245 | 355 |
| Fat gain (kg): | 4.2 | 3.0 | 3.7 | 2.7 | 3.2 | 3.6 | 2.4 | 1.3 |
| NEA change (cal): | 392 | 473 | 486 | 535 | 571 | 580 | 620 | 690 |
| Fat gain (kg): | 3.8 | 1.7 | 1.6 | 2.2 | 1.0 | 0.4 | 2.3 | 1.1 |

The mean and standard deviation of the 16 changes in NEA are calories $\bar{x} = 324.8$ (cal) and $s_x = 257.66$ cal. For the 16 fat gains, the mean and standard deviation are $\bar{y} = 2.388$ and $s_y = 1.1389$ kg. The correlation between fat gain and NEA change is r = −0.7786.

(a) Find the equation of the least-squares regression line for predicting fat gain from NEA change. Show your work.

$b = r\frac{s_y}{s_x} = -0.7786\left(\frac{1.1389}{257.66}\right) = -.00344$ kg/cal ;

$\boxed{b = -.00344 \text{ kg/cal}}$

we know regression line goes through $(\bar{x}, \bar{y})$ which is (324.8, 2.388)

$a = \bar{y} - b\bar{x}$.

$a = 2.388 - (-.00344)(324.8)$

$\boxed{a = 3.505 \text{ kg}}$

$\boxed{\hat{y} = 3.505 - .00344x}$

(b) What change in fat gain does the regression line predict for each additional 257.66 cal of NEA? Explain.

this is $s_x$

a change in one standard deviation in x corresponds to a change of r standard deviations in y.

$r \cdot s_y = -0.7786(1.1389) = -.8867$ kg

The regression line predicts a decrease of -.8867 kg in fat gain for an additional 257.66 cal of NEA.