Math 183
Statistical Methods

Eddie Aamari
S.E.W. Assistant Professor

eaamari@ucsd.edu
math.ucsd.edu/~eaamari/
AP&M 5880A

Today: Chapter 7 (continued)

- Interpret the slope and intercept of a linear regression
- Formulas for a linear model
- Conditions required to do regression
- Examples where regression is useful

Finding Regression Lines

Last class: we introduced

- Correlation coefficient (R or r)
- The idea of regression line
  (or "line of best fit, "least square line", ...)

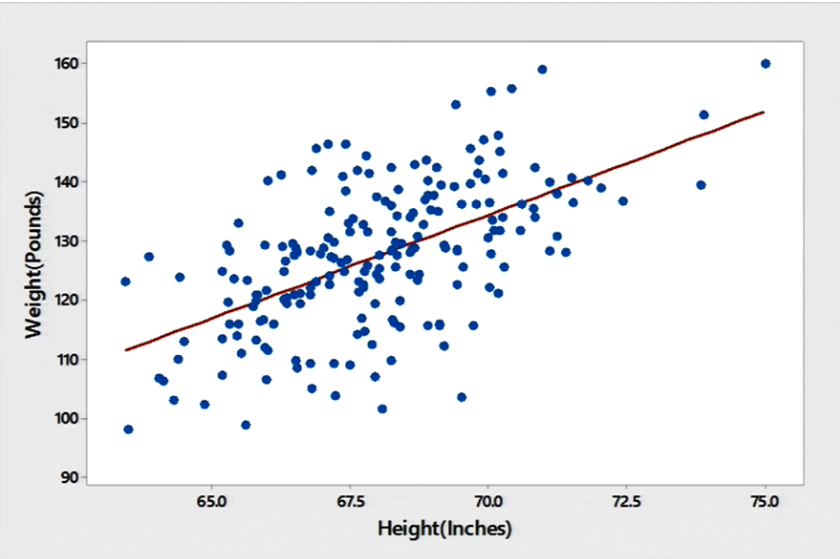
Notation:

- Regression Line: $\hat{y} = b_0 + b_1 \cdot x$

Interpretation:

- Intercept b_0: This is the predicted value for y when $x = 0$.
- Slope b_1: Measures the steepness of the regression line.
  It says how much y changes for each 1 unit change of x.

Our First Regression Equation



The intercept suggests that a 0 inch tall person should weigh -111 lbs.
This makes no real-world sense, but is a theoretical starting point for the model.

The slope suggests that for every inch increase in height, we expect a person to be about 3.5 lbs heavier.
Similarly, for every inch decrease in height we expect a decrease in 3.5 lbs.

$$\widehat{\text{Weight}} = -111 + 3.51 \times \text{Height}$$

$$\text{Slope} = \frac{\Delta y}{\Delta x}$$

$$3.5 \text{ lbs/inch} = \frac{3.5 \text{ lbs}}{1 \text{ inch}}$$

Why Build A Model?

– Perhaps y is really hard or expensive to measure, but well associated with x which is easy to measure.

– Perhaps y can only be measured after the fact (e.g. damage done by a tornado), but you need a sense for this before the fact.

– A model allows you to move from your data set to the larger universe of possibilities

– Parts of a model might answer questions you have about an issue (e.g. slope of height-weight graph gives the "weight of an inch of a person")

U.S. Navy

TABLE 1
MAXIMUM WEIGHT FOR HEIGHT SCREENING TABLE

Men Maximum Weight (pounds)	Member's Height (inches) (fractions rounded up to nearest whole inch)	Women Maximum Weight (pounds)
97	51	102
102	52	106
107	53	110
112	54	114
117	55	118
122	56	123
127	57	127
131	58	131
136	59	136
141	60	141
145	61	145
150	62	149
155	63	152
160	64	156
165	65	160
170	66	163
175	67	167
181	68	170
186	69	174

We see from this chart that every inch of height for a male equals about 5 or 6 pounds, and every inch for a female weighs about 3 or 4 pounds. This is exactly the slope of the regression line!

Calculating The Regression Equation

Regression Line: $\hat{y} = b_0 + b_1 \cdot x$

$$b_1 = R \cdot \frac{s_y}{s_x}$$

We see that:

- R gets the correct sign on the slope
- s_y/s_x gets the correct units on the slope

After calculating b_1 you get

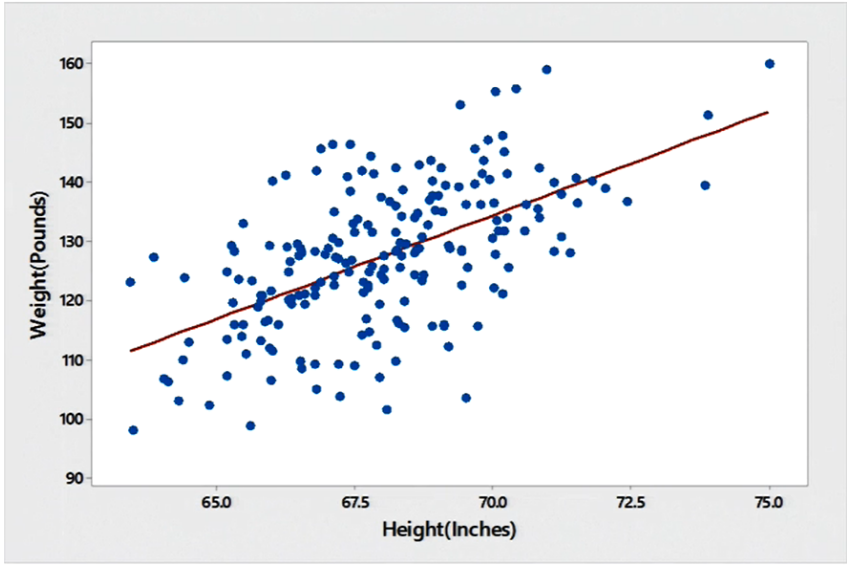
$$b_0 = \bar{y} - b_1 \bar{x}$$

This formula holds because the regression line always passes through $(\bar{x}, \bar{y})$.

Using the Regression Model

$$\widehat{\text{Weight}} = -111 + 3.51 \times \text{Height}$$

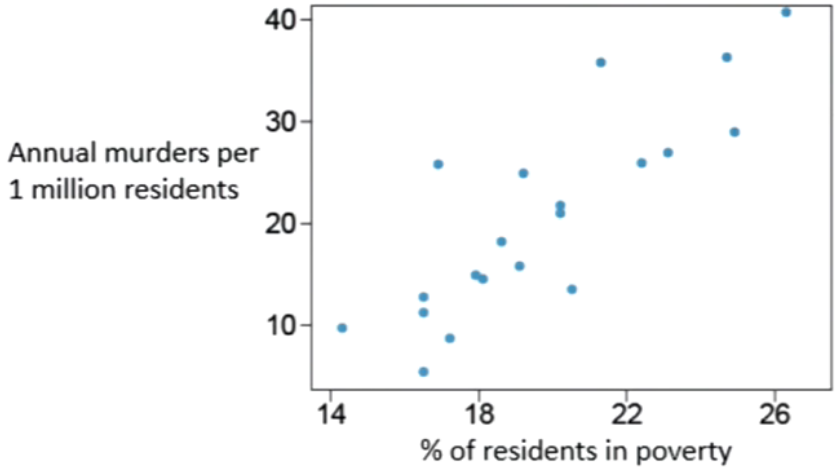
Try an example:
Convert your height to inches, see what the model predicts.
What is the residual based on your actual weight?



My data: 190cm converts to 75 inches.
The associated predicted weight is $-111 + 3.5 \cdot 75 = 151.5$.

My actual weight is 202 lbs, so the residual is $202 - 151.5 = 50.5$ lbs.
So my data point lies above the regression line (since residual >0).
The model (strongly) under-predicted.

Your Turn!



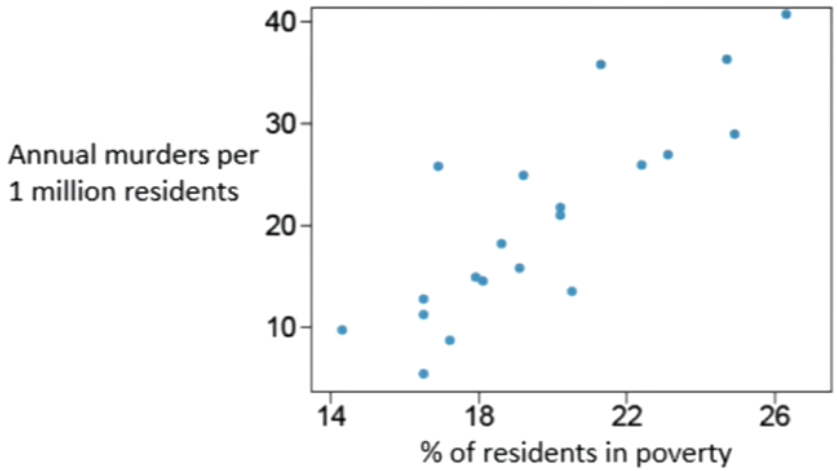
What might each dot represent?

1. A person in the U.S.
2. A small town in the U.S.
3. A metropolitan area in the U.S.
4. One of America's 20 richest cities

Answer: 3.

Individuals don't have poverty rates, so 1. is wrong.
Small towns don't have a million people, so the y axis wouldn't make sense in 2.
Rich cities have low poverty rates, so the x axis wouldn't make sense in 4.

Your Turn!

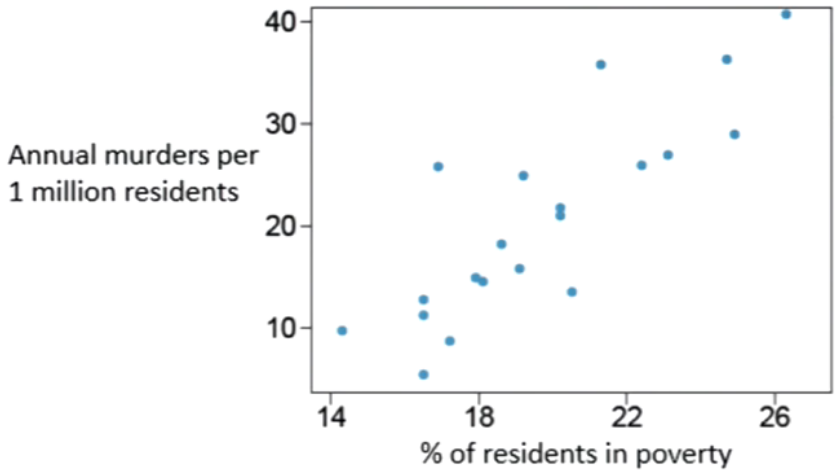


Guess the correlation coefficient for this scatterplot.

1. $R \simeq 0$
2. $R \simeq 0.25$
3. $R \simeq 0.55$
4. $R \simeq 0.85$
5. $R \simeq 1$

Answer: 4. The actual value is $R = 0.84$.

Your Turn!



You are told the regression line is

$$\text{Annual murder rate/million people} = -30 + 2.6 \cdot \text{Poverty Rate}.$$

What annual murder rate (per million people) do we expect in a city with a 20% poverty rate?

1. 4
2. 12
3. 22
4. 31

Answer: 3., since $-30 + 2.6 \cdot 20 = 20$.

Your Turn!

Which statements are true? Recall that the prediction is

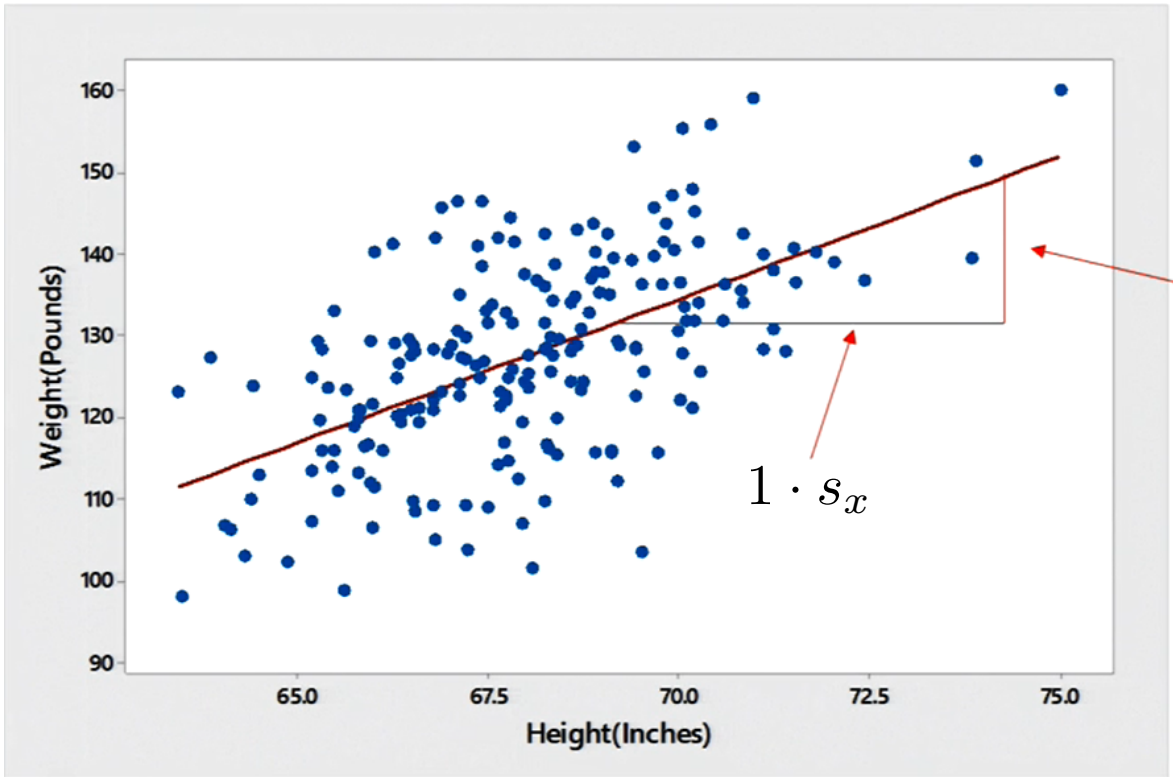
$$\text{Annual murder rate/million people} = -30 + 2.6 \cdot \text{Poverty Rate}.$$

1. A city with no poverty would have a murder rate of -30 people/million.
2. For every 1 unit increase in poverty, 2.6 more people will be murdered per year (for each million people in the city).
3. If you want to know the murder rate (per million people) of any city in the U.S., plug in the poverty rate into this equation.
4. The best values to plug in for the poverty rate are vetween 14 and 26.
5. The only values we may plug in for the poverty rate are between 14 and 26.

1. True. That's the interpretation of the intercept.
2. True. That's the interpreation of the slope.
3. False. Our prediction may only be valid for big cities.
4. True. Since most of the data used to build the model are between 14 and 26, we get the best results in this range.
5. False. Too strong language to be true.

More on the Slope

$$
\begin{aligned}
b_1 &= R \cdot \frac{s_y}{s_x} \\
&= \frac{R \cdot s_y}{1 \cdot s_x} \\
&= \frac{\Delta y}{\Delta x}
\end{aligned}
$$



In other words:

- If you're 1·SE above the mean height, you'll be R·SD's above the mean weight.
- If you're 2·SE above the mean height, you'll be $2R$·SD's above the mean weight.

Regression to the Mean

Recall that

$$-1 \leq R \leq 1.$$

Hence, moving 1SD form the mean of the x-variable takes us less than 1·SD (precisely R·SD) from the mean in the y-variable.
So, the world of x-values gets compressed (SD-wise) as the linear model converts them over to y-predictions.

The phrase "regression to the mean" is used to describe this phenomenon, and is where the term "linear regression" comes from.

Regression to the Mean: Examples

You give a class of students Test 1 today and Test 2 tomorrow. The tests cover the same material.
You make a scatterplot of scores and fit a regression model. You notice that all the high scorers (say 2·SD's above the Test 1 mean) didn't stand out as much on the second test (they will only be $2R$·SD's above the Test 2 mean). Their excellence seemed to regress some!

You look at the batting averages of all basketball players last year and this year. You notice the really bad players (3·SD's below last year's mean batting average) seem to do a little better this year (only $3R$·SD's below this year's mean).

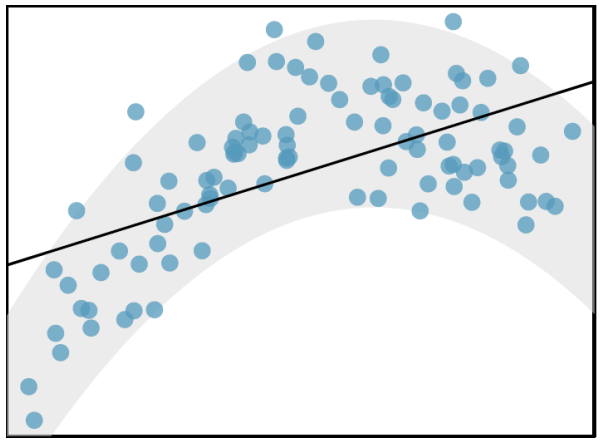
Why this occurs: Being exceptional on one measure (say, the x measure) requires exceptionalism and luck. If you focus on these people, you are focusing on the who had both exceptionalism and luck (on the x measure).
When you look at them on the other measure (y axis), they are still exceptional, but probably won't have the luck this time around.

Conditions for Creating a Regression Model

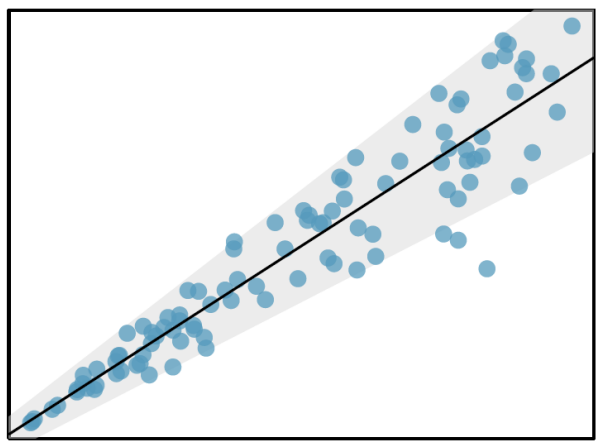
Four conditions must be met to create a linear regression:

- Graph looks roughly linear
- The histogram of residuals is nearly normal
- Constant variability around the regression line
- Independent observations in the scatterplot

Times series data often violate this last condition.



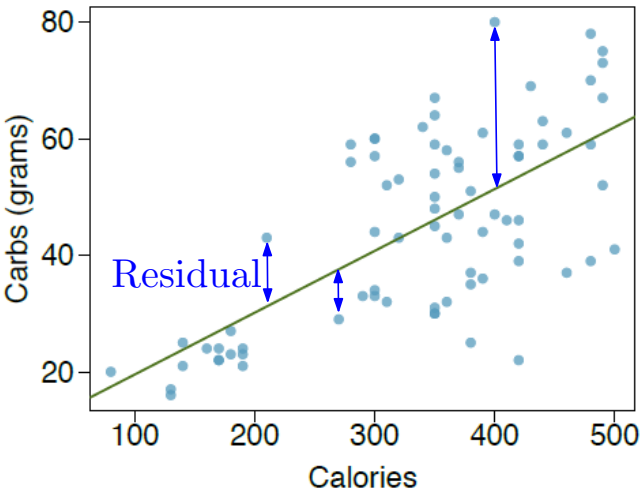
Not linear

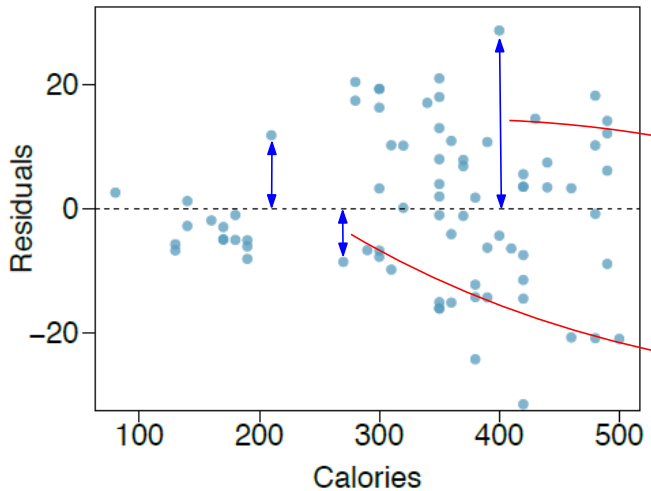


Non-constant variability

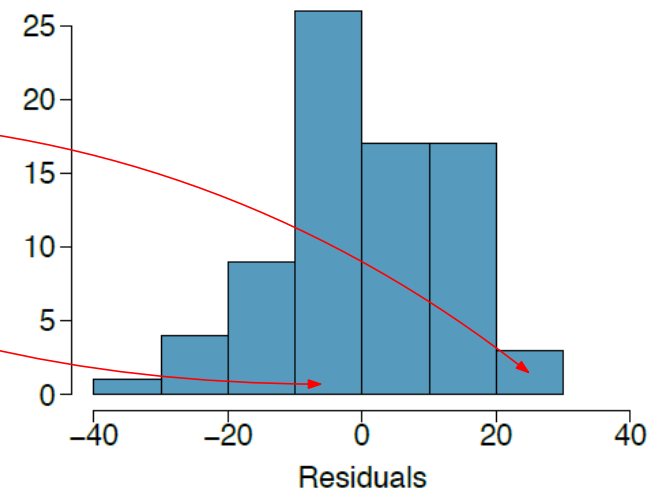
The Residual Plot and the Residual Histogram

Below is regression that predicts the number of carbs in a Starbucks item based on its calorie count.





Residual plot:
Imagine "substracting away" the regression line from the scatterplot to the left



Residual histogram:
Consider all the residuals and make a histogram of them

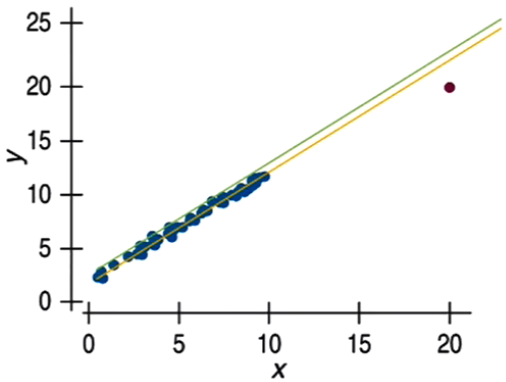
- Condition that variability is constant checked in the first 2 plots.
- Condition that residual histogram checked in the third one.

Here, it seems like the variability is not constant!

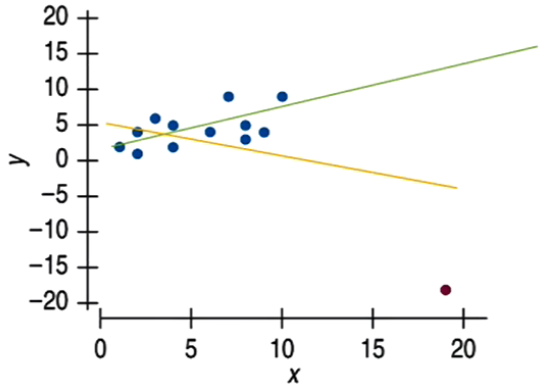
Shoud I Run a Regression With an Outlier?

For each graph,

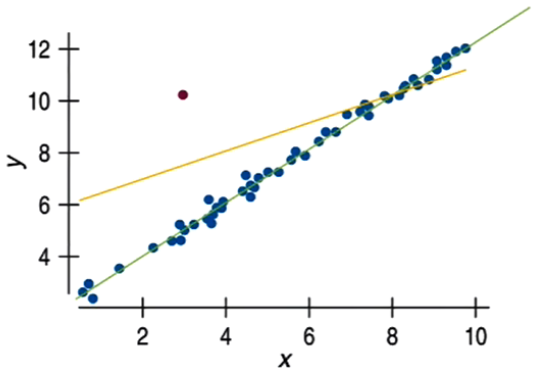
- the orange line is the regression line using all the point (including the outlier)
- the green line is the regression line when the using outlier is excluded.



High leverage
Low influence



High leverage
High influence



Low leverage
Moderate influence

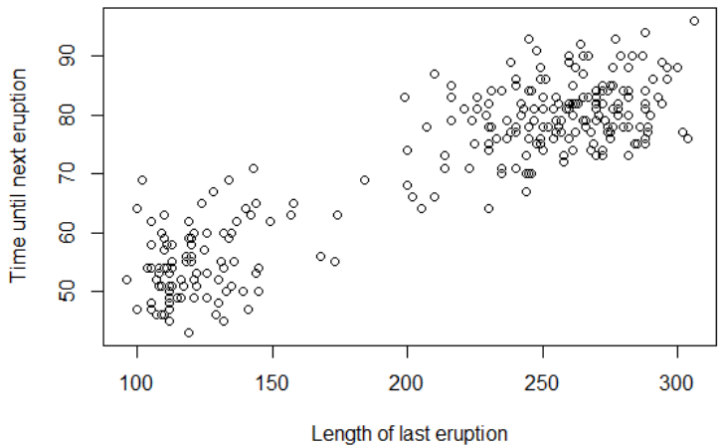
Two important ideas:

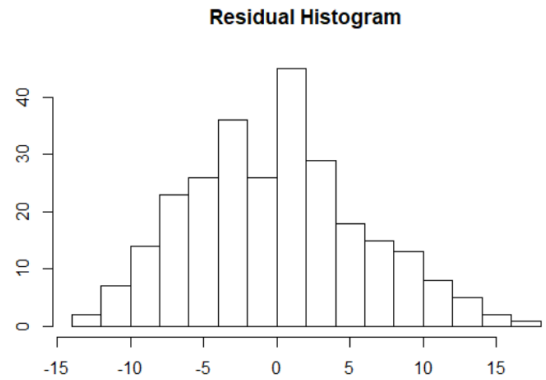
- A high leverage point is one where x is far from the mean of the x values.
- A high influence point is one that gives a significantly different slope for the regression line when it is included, versus excluded, for an analysis.

Predicting Old Faithful

Can we predict how long it will be until the next eruption of Old Faithful (a geyser) based on how long the current eruption lasted?

Duration	Interval
216	79
108	54
200	74
137	62
272	85
173	55
282	88





Source: R. Hutchinson, a geologist at Yellowstone.

We definitely meet the conditions for doing linear regression!

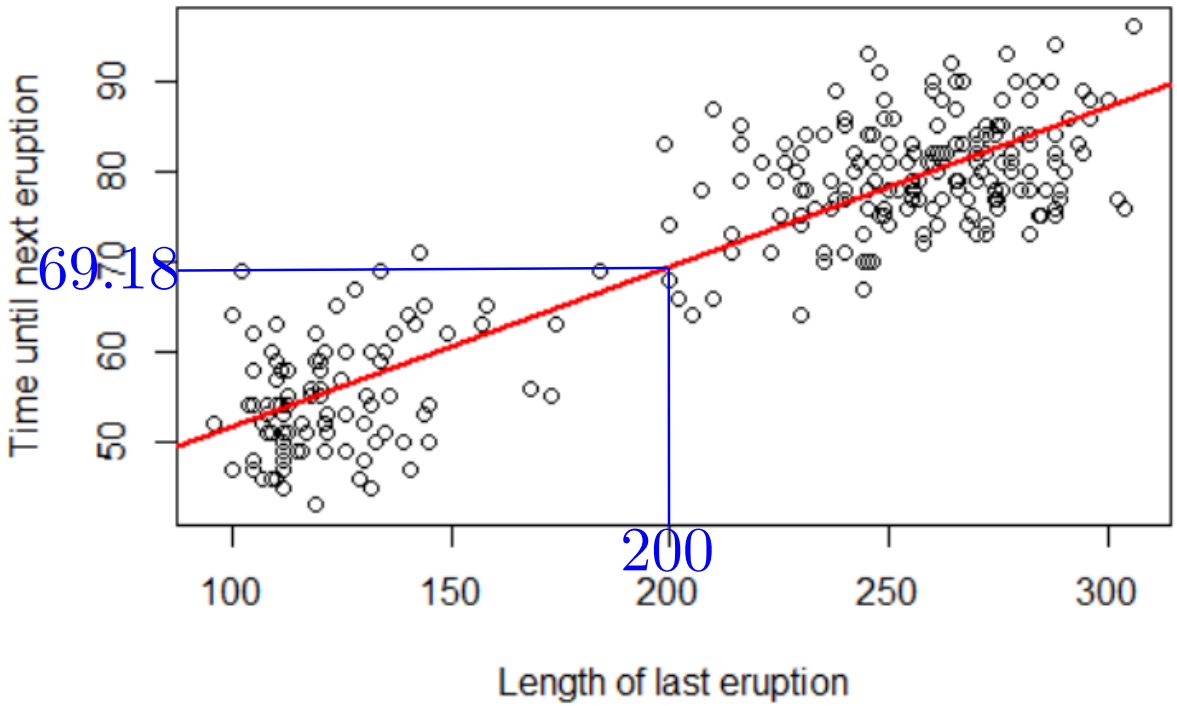
Finding the Regression Line in R

Part of the printout from R is:

```
> model = lm(Interval~Duration, data= oldfaith)
> summary(model)

Call:
lm(formula = Interval ~ Duration, data = oldfaith)

Residuals:
     Min       1Q   Median       3Q      Max
-12.3337  -4.5250   0.0612   3.7683  16.9722

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.987808   1.181217   28.77   <2e-16 ***
Duration     0.176863   0.005352   33.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- The estimate for the intercept is $b_0 = 33.98$.
- The estimate for the coefficient on the (Duration) is $b_1 = 0.176$.

Statistical Tourism



$$b_0 = 33.98 \text{ and } b_1 = 0.176.$$

Example: If I just saw an eruption that last 200 seconds, I expect to have to wait

$$33.98 + 0.176 \cdot 200 = 69.18 \text{ seconds}$$

before the next eruption starts.

More About Correlation: The R^2 Statistic

Recall that the correlation is defined as

$$R = \frac{1}{n-1} \sum_{(x,y) \text{ pairs}} \frac{(x-\bar{x})}{s_x} \frac{(y-\bar{y})}{s_y}.$$

One can actually show that

$$R^2 = \frac{\sum_y (\hat{y} - \bar{y})^2}{\sum_y (y - \bar{y})^2}.$$

For a given linear model, R^2 is the proportion of the variation in the y-variable that is accounted for (or explained by) the variation of the x-variable.

R^2 is called the **Coefficient of determination** of the data set.
(or just the "R-squared statistic")

More About Correlation: The R^2 Statistic

Similarly, one has

$$R^2 = 1 - \frac{\sum_y (y - \hat{y})^2}{\sum_y (y - \bar{y})^2} = 1 - \frac{s_e^2}{s_y^2},$$

where s_e is the standard deviation of the residuals.

We see that the smaller the residuals, the larger R^2.

The R^2 Statistic: Examples

A different portion of the Old Faithful print is:

```
Residual standard error: 6.004 on 268 degrees of freedom
Multiple R-squared:  0.8029,    Adjusted R-squared:  0.8022
F-statistic:  1092 on 1 and 268 DF,  p-value: < 2.2e-16
```

Hence, 80.29% of how long we must wait is completely determined by how long the last eruption lasted!

As another example, the R^2 in the height-weight regression is 0.67.
So, 67% of the variability in weights is simply because of height differences.