**Chapter 2**

**2.1-2.4:** Sections 1-4 of chapter 2 are considered pre-material for the course. You are responsible for

all of the material in the section. The purpose of these sections is to gain a basic understanding

of ways to describe a data set. Here we will discuss a few of the concepts in the pre-material but

not all of the concepts.

* Proportion: A proportion can be calculated from categorical data, when there are two options. It tells us relative to the total how often one of the options occurs. Proportions will always be between 0 and 1. -Categorical Variables
* Mean: The mean is a measure of centrality. It tells us where the center of the data is located. -Quantitative Variables
* Standard Deviation: The standard deviation tells us about the spread of the data. It looks at the average distance from each data point to the mean of the data set.

Key Words:

Proportion: Percentage

Mean: Average

Standard Deviation: Variance, or spread of data

Vocabulary is very important in this course! It is almost like learning a new language.

**Example:** With your group decide if each of the following refers to a proportion, mean or standard deviation.

* On average drivers in KY drive 4 miles over the speed limit when driving on the interstate. Mean
* 14% of people with blue eyes will develop cataracts. Proportion
* Compared to adults, the number of cavities a child has is much more varied from child to child. Standard Deviation

Throughout the rest of the course we will rely on some common notation. It is important that you memorize and understand the notation.

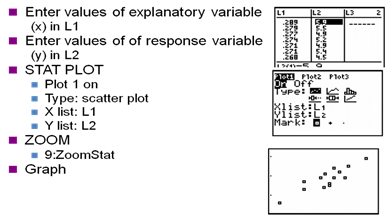
|  |  |  |
| --- | --- | --- |
|  | Statistic  (Calculated from Sample) | Parameter  (Calculated from Population) |
| Mean | (x-bar) | µ (mu) |
| Proportion | (p-hat) |  |
| Difference in Means |  |  |
| Difference in Proportions |  |  |
| Standard Deviation | s | σ (sigma) |

**2.5: Two Quantitative Variables: Scatterplot and Correlation**

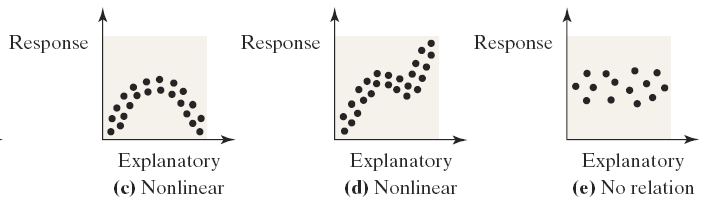
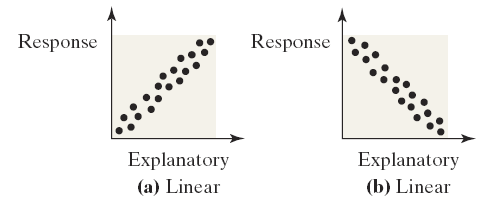
Often researchers would like to compare two variables to see if there is a relationship

between them.

* **Response variable** (Dependent Variable)
  + the outcome variable on which comparisons are made
* **Predictor variable** (Independent variable)
  + defines the groups to be compared with respect to values on the response variable
  + One variable (the explanatory or predictor variable) could be used to explain the other (the response or dependent variable)
* On way to visualize this type of data is by using a scatterplot
* **Scatterplot: Graphical display of relationship between two quantitative variables:**
  + **Horizontal Axis: *Predictor variable*, x**
  + **Vertical Axis: *Response variable*, y**
* **Create a Scatter Diagram use TI 83 or 84**

****

* Various Types of Relations in a Scatter Diagram

****

Scatterplots can help us visualize and describe associations. We can describe the direction of an association follows:

* A positive association means that values of one variable tend to be higher when values of the other variable are higher. In this case the scatterplot will go up to the right.
* A negative association means that values of one variable tend to be lower when values of the other variable are higher. In this case the scatterplot will go down to the right.

Two variables are not associated if knowing the value of one variable does not give you any information about the value of the other variable. In this case the scatterplot has no pattern

We can also describe the strength of an association.

* When the points are clustered close together we say the association is strong.
* When the points are clustered far apart we say the association is weak.



**Example**  With your group describe the direction and strength of each of the following scatterplots.

* 
* 
* **Correlation Coefficient *r***



***Example : Positive/Negative Associations***

In each case, do you expect a positive or negative association between the two quantitative variables?

a). Number of years of education and annual salary, for US adults

*Positive: people with more education generally make higher salaries*

b). Age and maximum running speed, for adults.

*Negative: Older people (such as 80-year-olds) tend to run slower than younger people (such as 20-year-olds).*

c). Age and maximum running speed, for children.

*Positive: Older children (such as 12-year-olds) tend to run faster than younger children (such as 4-year-olds).*

d). Age of the husband and age of the wife, for married couples

*Positive: 80-year-olds are more likely to be married to other 80-year-olds than to 20-year-olds. Ask the students what a negative relationship between these two variables would mean!*

Associations can take different shapes. When the underlying shape is curved we call the association curvilinear. When the underlying shape is a straight line we call the association linear.

***The Correlation coefficient (r***) measures the strength and direction of the linear association between x and y

**Properties of the Linear Correlation Coefficient**

1. The correlation coefficient *r* is always -1 ≤ *r* ≤ 1.
2. When *r* = +1, a perfect positive relationship exists between *x* and *y*.
3. Values of *r* near +1 indicate a positive relationship between *x* and *y.*
   * The closer *r* gets to +1, the stronger the evidence for a positive relationship.
   * The variables are said to be **positively associated.**
   * As *x* increases, *y* tends to increase.
4. When *r* = -1, a perfect negative relationship exists between *x* and *y.*
5. Values of *r* near -1 indicate a negative relationship between *x* and *y.*
   * The closer *r* gets to -1, the stronger the evidence for a negative relationship.
   * The variables are said to be **negatively associated.**
   * As *x* increases, *y* tends to decrease.
6. Values of *r* near 0 indicate there is no linear relationship between *x* and *y.*
   * The closer *r* gets to 0, the weaker the evidence for a linear relationship.
   * The variables are **not linearly associated.**
   * A nonlinear relationship may exist between *x* and *y*.

7. Correlation has a unitless measure - does not depend on the variables’ units

8. Two variables have the same correlation no matter which is treated as the response

Variable, in another word, the correlation between *X* and *Y* is the same as the correlation between *Y* and *X*



Correlation is not resistant to outliers

Correlation only measures strength of linear relationship. *r* = 0 means no *linear* association. The variables could still be otherwise

Correlation not the Causation. When data are observational, we cannot claim a causal relation exists between two variables

**Use TI-83 calculator:**

**First you need turn the diagnostics on by selecting the catalog(2nd 0). Scroll down**

**And select Diagnostic0n. Hit Enter twice to activate diagnostics.**

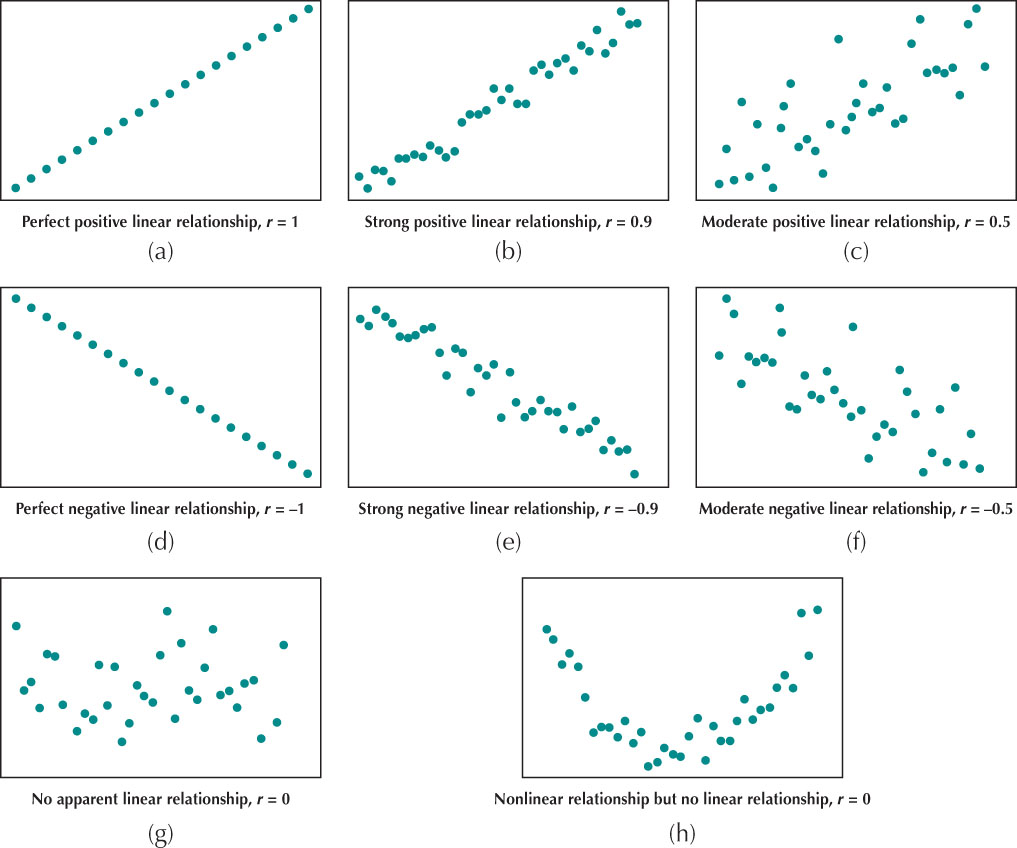
****Enter the *x* values in L1, *y* values in L2

STAT

🡪CALC

8🡫LinReg(a+bx)

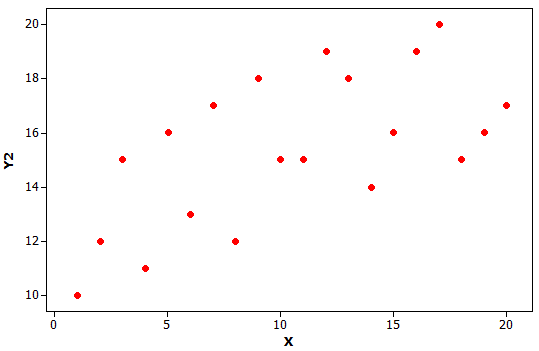
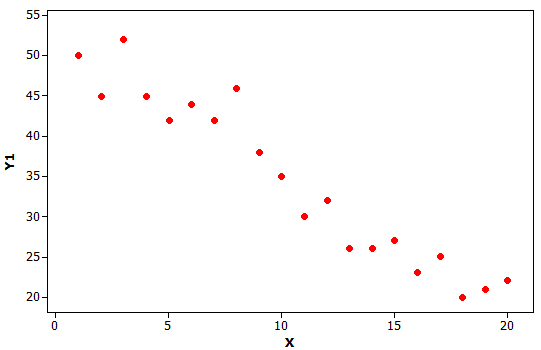
Enter



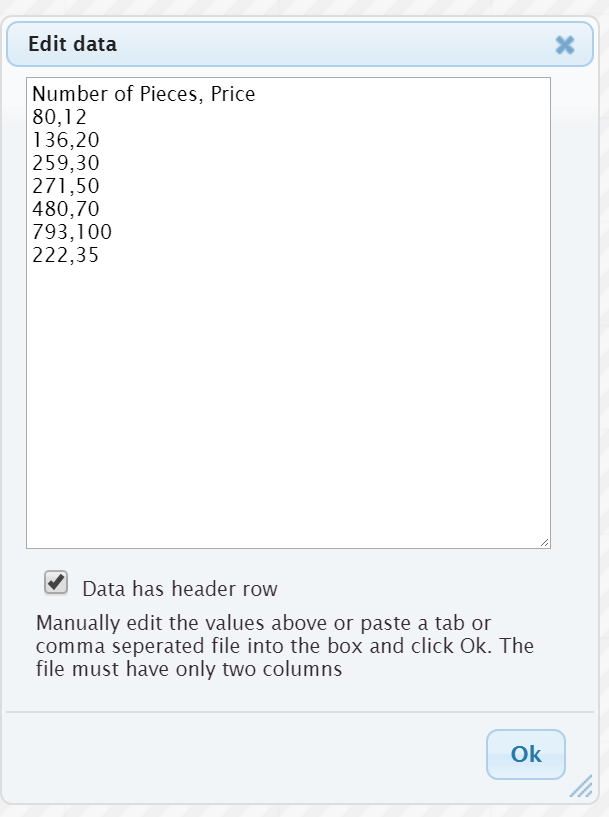
***Example 2: Scatterplots and Correlation***

For each of the scatterplots, which of the following correlation values most closely approximates the correlation of the data shown?

–1 -0.9 -0.5 0 0.5 0.9 1

*Left scatterplot: about 0.5 Right scatterplot about -0.9*

For this course we will use StatKey to find the regression line and r. In order to find a regression line in StatKey complete the following steps:

* Go to <http://www.lock5stat.com/StatKey/>
* Click on 2 quantitative variables
* Click on Edit data
* Enter the data (The image shows what data should look like)
* Click OK
* Click on Show regression line.

When using StatKey you always want to do some quick checks to make sure that the data was entered correctly.

* Looking at the plot are the explanatory and response variable correct? If they are switched you can click on “switch variables” to fix.
* Look at the sample size. Does the sample size match the number of data points in your data? You should have 7 data points.

***Example 3: Correlation and Outliers***

Use technology to find the correlation of the following dataset with n = 6 data values.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *x* | 1 | 2 | 3 | 4 | 5 | 15 |
| *y* | 5 | 8 | 6 | 4 | 9 | 50 |

*This is a good time to get the students comfortable using technology to find the correlation. The correlation here is r = 0.964.*

Now find the correlation for the same dataset, but with the outlier at (10, 50) removed (leaving n = 5 points). Comment on the effect of the outlier on the correlation.

*The correlation without the outlier is r = 0.305. Notice the very strong effect of the outlier on the correlation! You might want to draw the scatterplot with and without the outlier to show the strong visual effect it has as well.*

***Example 4: TV and Life Expectancy***

For a sample of 20 countries, we have values for the average life expectancy (in years) and the prevalence of television sets (number of TVs per 1000 people). The results are displayed in the scatterplot below.



a). Does there appear to be an association between the two variables? If so, is it positive or negative?

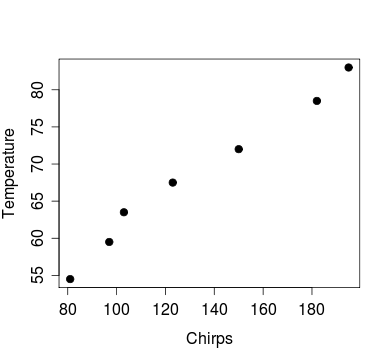
*There appears to be a strong positive association between life expectancy and TVs for countries.*

b). Based on these results, comment on a proposal to send more TVs to countries with lower life expectancy to help people in those countries live longer.

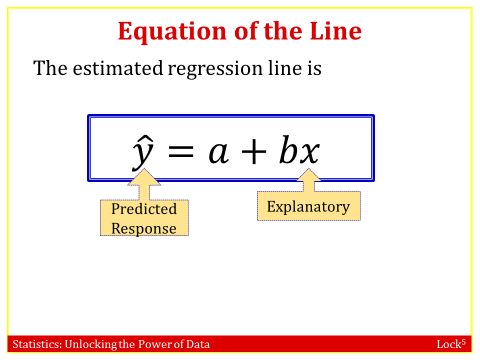
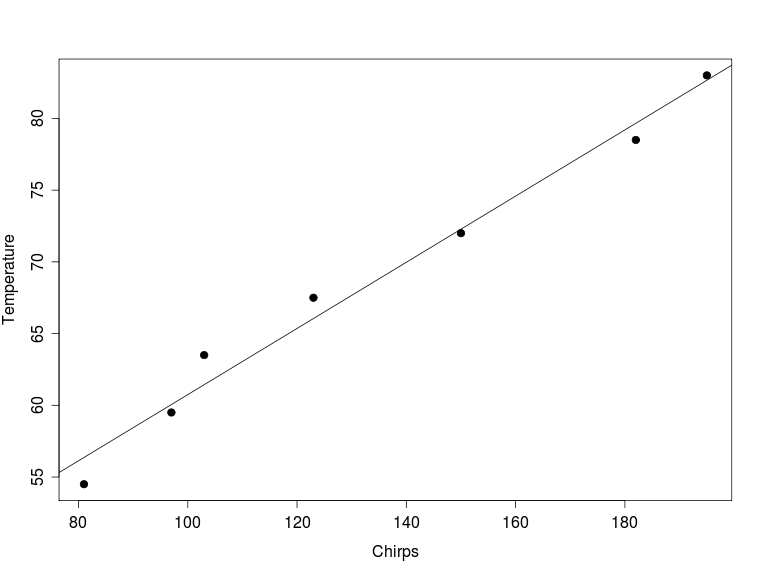
*Remember: correlation does not imply causation! We can’t infer that adding more TVs will cause people to live longer. Most likely, wealthier countries have better systems for health care and more TVs than poorer countries, so wealth of the county is a confounding variable that influences both of the other variables.*

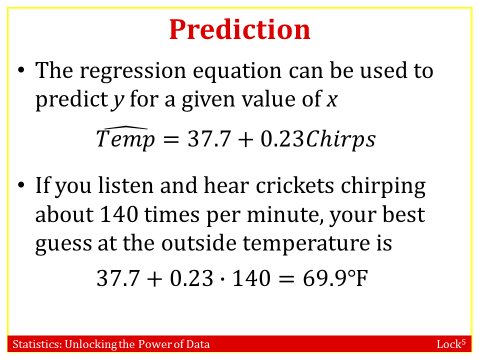
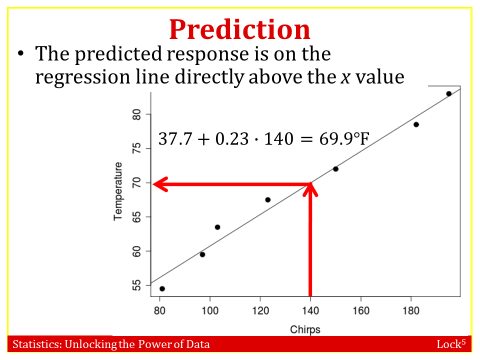
2.6: **Two Quantitative Variables: Linear Regression**

**Crickets and Temperature**



*Goal: Find a straight line that best fits the data in a scatterplot*



* **Finding the Equation of the Regression Line and use the line to make predictions**

Enter the *x* values in L1, *y* values in L2

STAT

🡪CALC

8🡫LinReg(a+bx)

Enter

Scroll down until you find the values for *a* and *b*

* **A regression line** is a straight line that best fits the data in a scatterplot. The equation for this line can be written as:



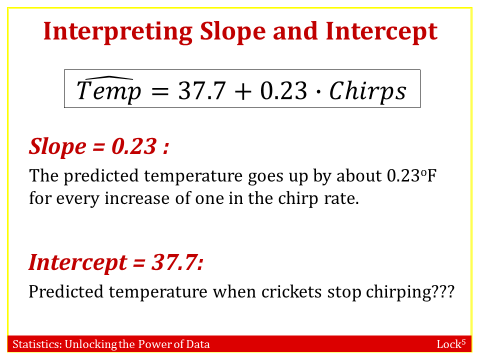
The regression equation can be used to predict *y* for a given value of *x.*

* **The y-intercept** of the regression line is denoted by *a*
  + The predicted value for y when x = 0
  + To interpret the *y*-intercept, we must first ask two questions:

**1.** Is 0 a reasonable value for the explanatory variable?   
**2.** Do any observations near *x* = 0 exist in the data set?

* **The slope** of the regression line is denoted by *b*

Slope: measures the change in the predicted variable (y) for a 1 unit increase in the explanatory variable in (x)



How do we choose the regression line? We need a way to compare the data we observe and the line that we summarize the data with.

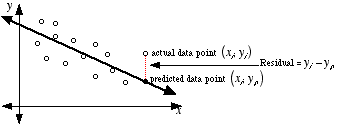
* The *observed response value* *y*, is the response value observed for a particular data point
* The *predicted response value* , is the response value that would be predicted for a given *x* value, based on a model
* The best fitting line is that which makes the predicted values closest to the actual values
* The residual gives us a way to compare the observed y’s to the predicted y’s. The ***residual*** for each data point is

*Residual = observed – predicted* =

* **Residuals**

Measures the size of the prediction errors, the vertical distance between the point and the regression line. The residual is [positive](http://www.mathwords.com/p/positive_number.htm) if the data point is above the graph. The residual is [negative](http://www.mathwords.com/n/negative_number.htm) if the data point is below the graph. The residual is 0 only when the graph passes through the data point.

* + Each observation has a residual
  + Calculation for each residual: *observed – predicted* =
  + The residual is also the vertical distance from each point to the line

A large residual indicates an unusual observation 

* Want to make all the residuals as small as possible.

**Example:** With your group answer the following:

* Using the cricket data how many residuals could you compute?

 7, because there are 7 data points

* What is the predicted temperature when the crickets chirp 81 times per minute?

* What is the residual when the crickets chirp 81 times per minute? Recall the observed data point is (81, 54.5).

Residual = 54.5-56.33=-1.83

The goal is to make all the residuals as small as possible. We can do this by drawing the least squares line, which minimizes the sum of squared residuals.

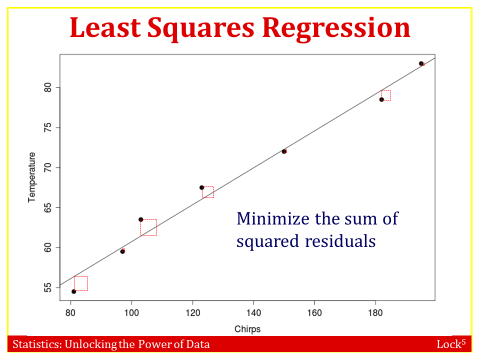
****

* **Residual sum of squares**:

**The least squares regression line** is the line that minimizes the vertical distance between the points and their predictions, i.e., minimizes the sum of squared residuals.

* “least squares line” = “regression line”

The area of each square represents the square of the residual, since the residual is the vertical distance from each point to the line.

* Do not use the regression equation or line to predict outside the range of *x* values available in your data (do not extrapolate!)
* If none of the *x* values are anywhere near 0, then the intercept is meaningless!
* Computers will calculate a regression line for any two quantitative variables, even if they are not associated or if the association is not linear
* The regression line/equation should only be used if the association is approximately linear
* Watch out for outliers that don’t fit the pattern or can greatly influence the line.
* Even a strong linear fit does not (necessarily) imply a cause/effect relationship.
* Correlation not the Causation. When data are observational, we cannot claim a causal relation exists between two variables
* ALWAYS PLOT YOUR DATA!

You will not solve this by hand instead you will use a computer. StatKey will find the line for you.

Example:

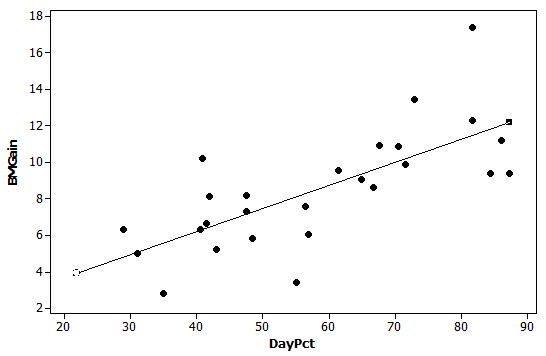
****

The monthly insurance payment is $200 for those who have 2 accidents, what is the residual? How would you interpret this residual?

Residual== 200-216.75 = -16.75. The monthly insurance payment is $16.75 below average for those who have 2 accidents

Example: Time of Eating and Weight Gain

The time of day in which calories are consumed can affect weight gain. At least, that appears to be true in mice. Mice normally eat all their calories at night, but when mice ate some of their calories during the day (when mice are supposed to be sleeping), they gained more weight even though all the mice ate the same total amount of calories. The scatterplot shows the percent of calories eaten during the day, DayPct, and body mass gain in grams, BMGain, for a study involving 27 mice. The regression equation to predict body mass gain from percent of calories eaten during the day is

a). Circle the dot that has the largest positive residual.

b). Put a square around the box that has the most negative residual.

c). What is the predicted body mass gain for a mouse that eats 50% of its calories during the day?

, so a mouse that eats 50% of its calories during the day is predicted to gain 7.46 grams.

d). Find the residual for the mouse who ate 48.3% of its calories during the day and gained 5.82

grams.

*We first find the predicted body mass gain:*

*The residual is then: Residual = Observed – Predicted = 5.82 – 7.24 = –1.42.*

*Have them find this point and its corresponding residual on the scatterplot.*

e). Interpret the slope of the regression line in context.

*The slope is 0.127. When a mouse eats one more percent of its calories during the day, its predicted body mass gain goes up by 0.127 grams.*

f). Interpret the intercept of the line in context, if it makes sense to do so.

*The intercept is 1.11. A mouse who eats 0% of its calories during the day (and all of them at night when a mouse normally eats all its food) is predicted to gain 1.11 grams. The intercept does make sense in this context.*