Chapter 7

Linear Regression

Homework

p195 1-9, 11, 13, 15, 17, 19,
21, 26, 28, 29, 32, 35, 37

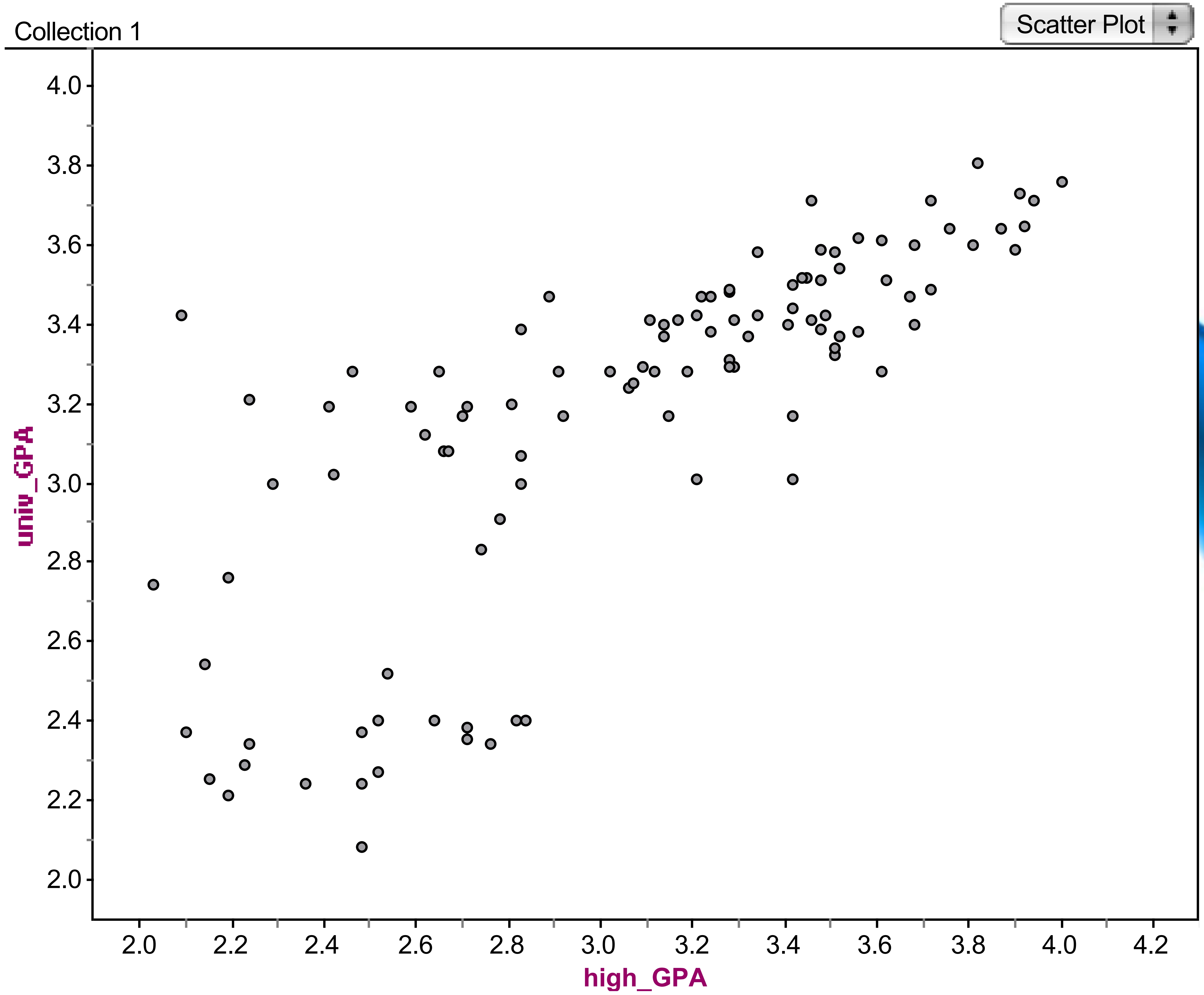
Objectives

Determine Least Squares Regression Line (LSRL) describing the association of two variables.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

GPA Example

- The following case study "SAT and College GPA" contains high school and university grades for 105 computer science majors at a state school.
  - Describe the relationship.



Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

The Linear Model

- The correlation in the GPA example is 0.7795.
  - The strength of the correlation coefficient suggests a linear association between these two variables, but the correlation coefficient tells us nothing about the association itself.
  - What we need to describe the linear relationship between two quantitative variables is a model of that association.
  - That model is a linear equation that quantifies the relationship or pattern in the co-variability of those variables.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

The Linear Model

- The **linear model** is nothing more complicated than a **linear equation** of a straight line through the data.

- That model (equation) will have a slope and intercept to help us understand and describe the relationship between the two variables.

- It is unlikely in the extreme that the points in a **scatterplot** of the variables data will actually form a nice, straight line. But a nice straight line might suffice in describing the overall pattern of the scatter plot with the slope and intercept.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Regression

- If the relationship between variables is strong, we know that the variables tend to cluster to a line.
- That means we can get some idea about the behavior of one (response) variable by observing the behavior of another (predictor) variable.
- We can then make predictions about the response variable by knowing a value of the predictor variable.
- What we need to do is find the equation of the line that best fits the data. That, of course, begs the question what do we mean by "best fit"?
- That "line of best fit" is called the regression line or....

Least Squares Regression Line (LSRL)

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Residuals

- First know this, **the model will be wrong**.
  - What that means is the data points will rarely actually fall on the line. The model will predict a different value than the observed data, but it may be close and the model will probably be of great interpolative value.
  - Similar to the way data values fall above and below the mean, some points will be above the line and some points will be below the prediction model (line).
  - The estimate made from a model is the **predicted value. ...**

- "y-hat" (denoted as $\hat{y}$)

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Residuals

- The difference between the observed value (Y) and its associated predicted value ($\hat{y}$) is called the residual.

- To find the residuals, we subtract the predicted value from the observed value:

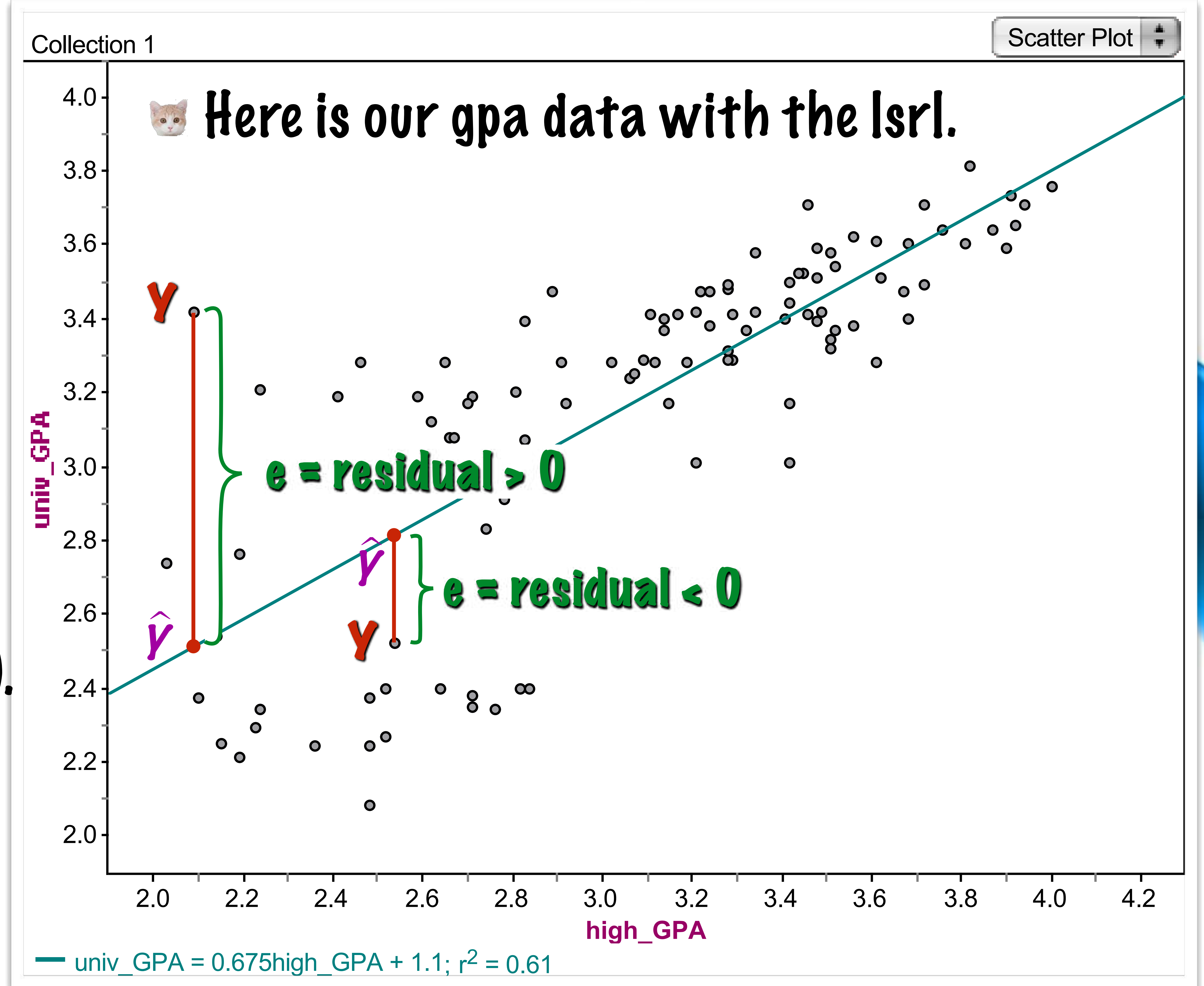
residual = observed - predicted

$$e = Y - \hat{Y}$$

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Residuals

- A positive residual means the predicted value is less than the actual value (an underestimate).
  - In the figure, the predicted college gpa is 2.5, while the true value of college gpa is 3.42, and the residual is 0.92.
- A negative residual means the predicted value is greater than the actual value (an overestimate).
  - In the figure, the estimated college gpa is 2.80, while the true value of college gpa is 2.52, thus the residual is -0.28.



"Best Fit" Means Least Squares

- Some residuals are positive, others are negative, and, on average, they balance each other out. This is very similar to the calculation for standard deviation.

- Like we did with deviations, we square the residuals and add the squares, resulting in the "**sum of squared residuals**" (This is also called "sum of squared errors" or SSE).

  - The smaller the sum (**least squares**), the better the fit.

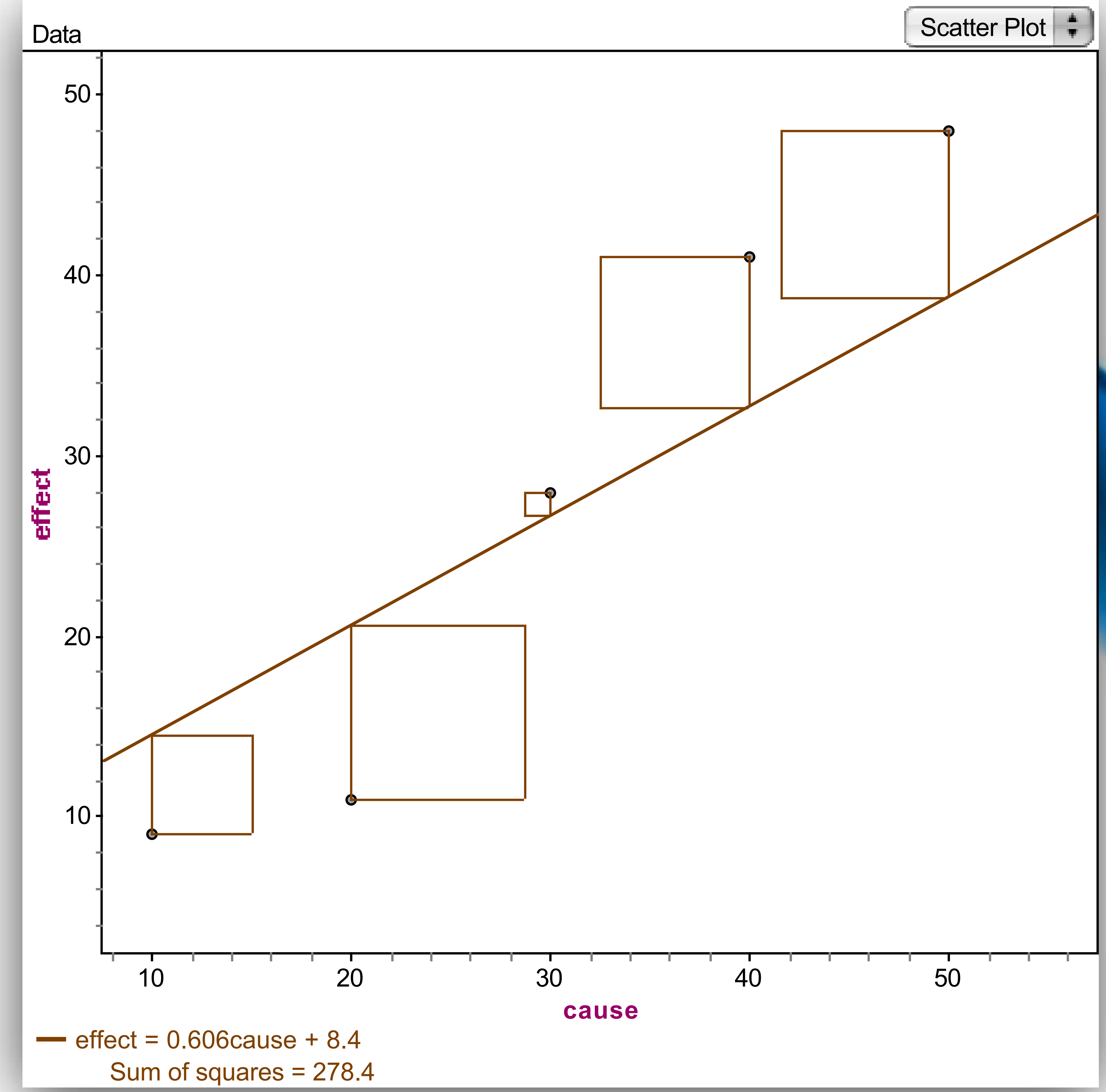
  - The **line of best fit** is the line for which the sum of the squared residuals is smallest, the **least squares** line.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Least Squares

- This shows a least squares residual line and visually demonstrates the methodology of "least squares".
  - The line that minimizes the residuals also minimizes the area of the squares.

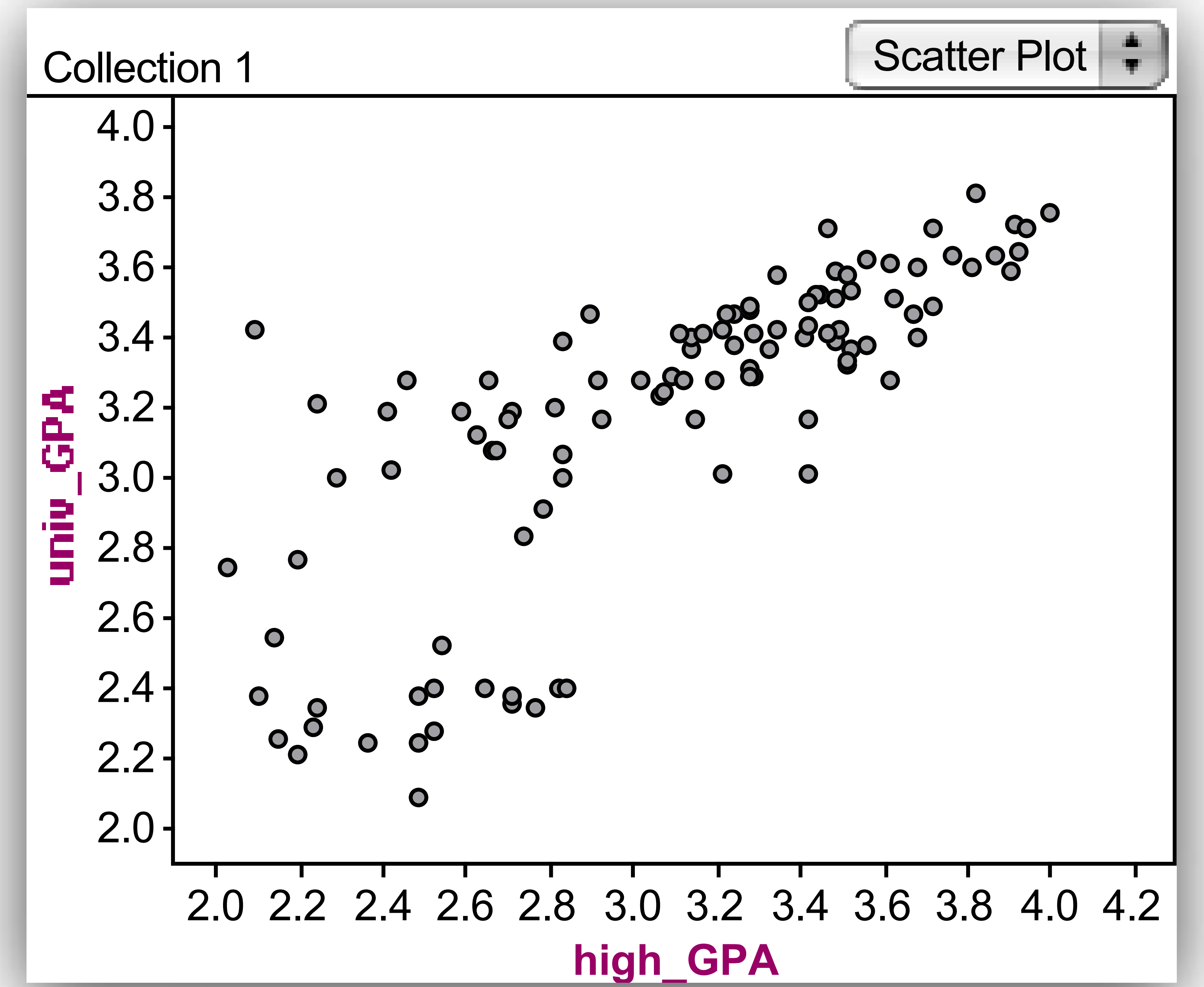
https://www.desmos.com/calculator/zvrc4lg3cr

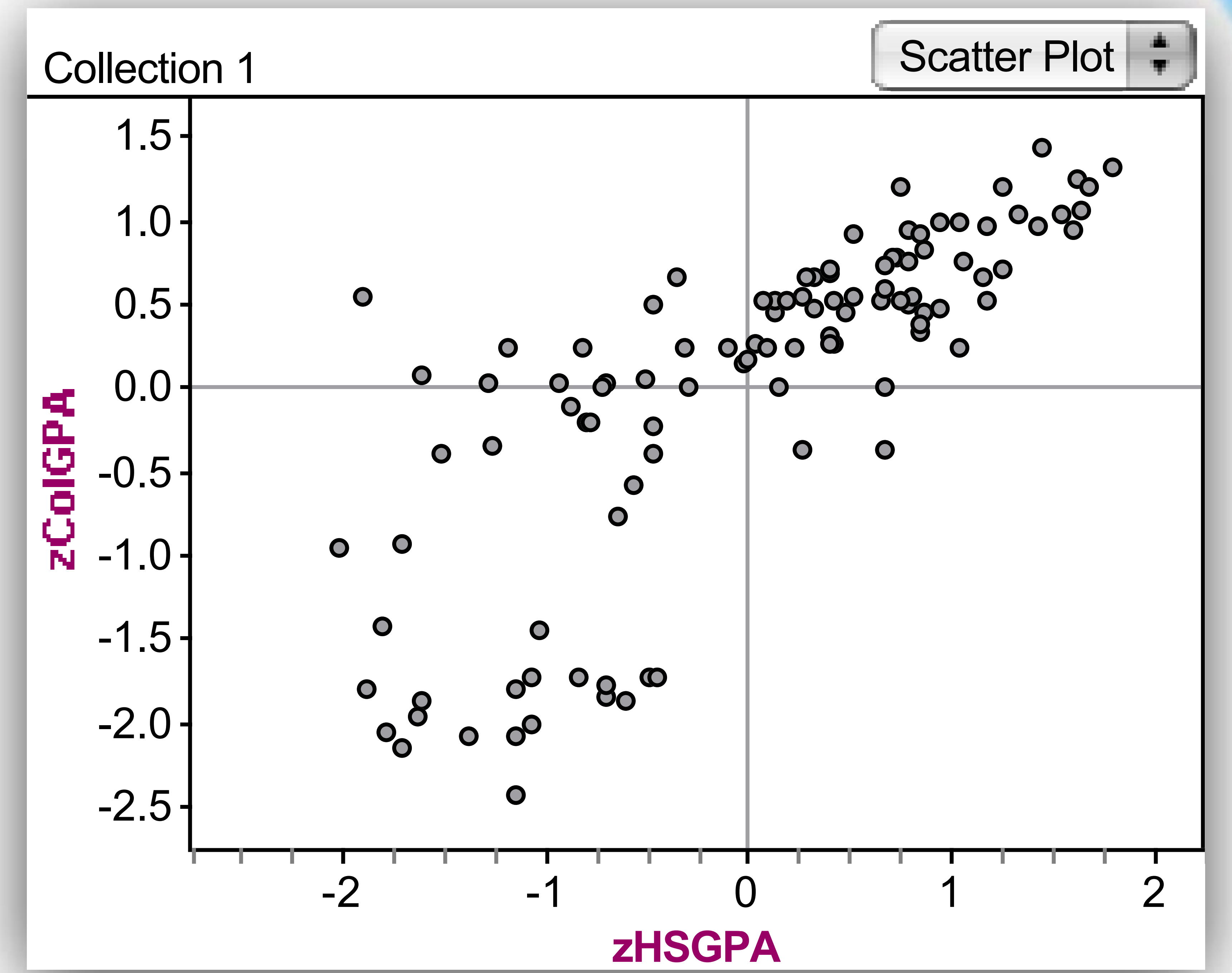


Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Correlation and the LSRL

- This figure shows the scatterplot of *z*-scores for the variables ColGPA and HSGPA along with the scatterplot of raw scores.
  - What do you notice about the distributions.



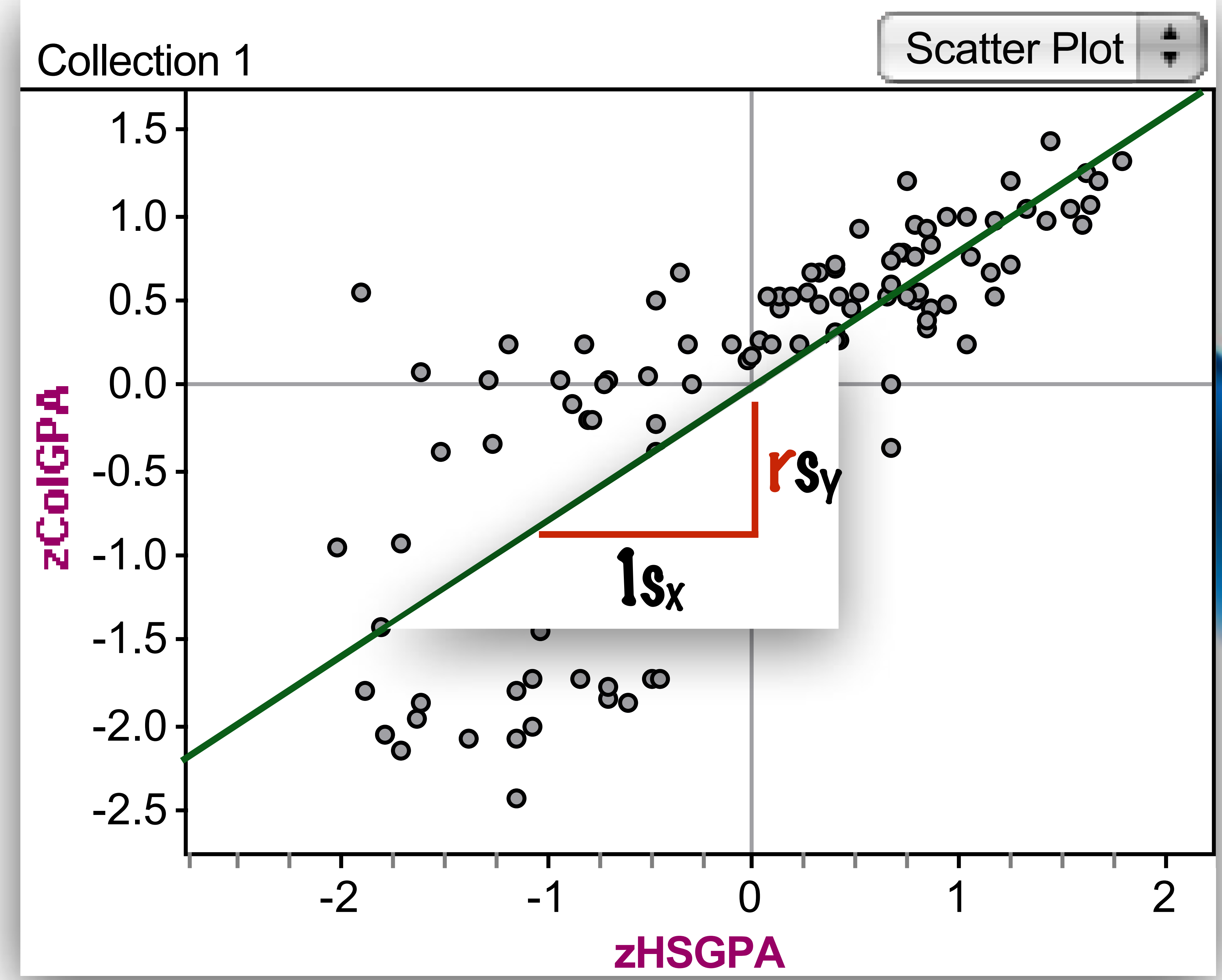


- Remember: changing scales does not change r.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Correlation and the LSRL

- If an individual have an average HSGPA, we would expect about an average ColGPA as well.
- That suggests the point consisting of the mean for each variable in on the LSRL.
- That would tell us that moving one standard deviation away from the mean in x moves us r standard deviations away from the mean in y.



rs_y

$1s_x$

$$r = \frac{1}{n-1}\sum z_x z_y$$

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Let Me Repeat That

- The correlation coefficient, r, is the slope of the line of best fit **for Z-scores**.
  - That means that if the data have been changed to z-scores, the line best fitting that data will have a slope of r.

However, if the data is **NOT** re-expressed as z-scores, the slope of the lsrl **IS NOT r.**

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

So What?

- A slope of r for the line of best fit for z-scores means that a change of 1 unit in z_x predicts a change of **r** units in $\hat{z}_y$.

- In terms of raw values x and y, a change in 1 standard deviation in x **predicts** a change of **r** **standard deviations** in $\hat{y}$

- Now we can express the **slope** of the lsrl in terms of **r**.

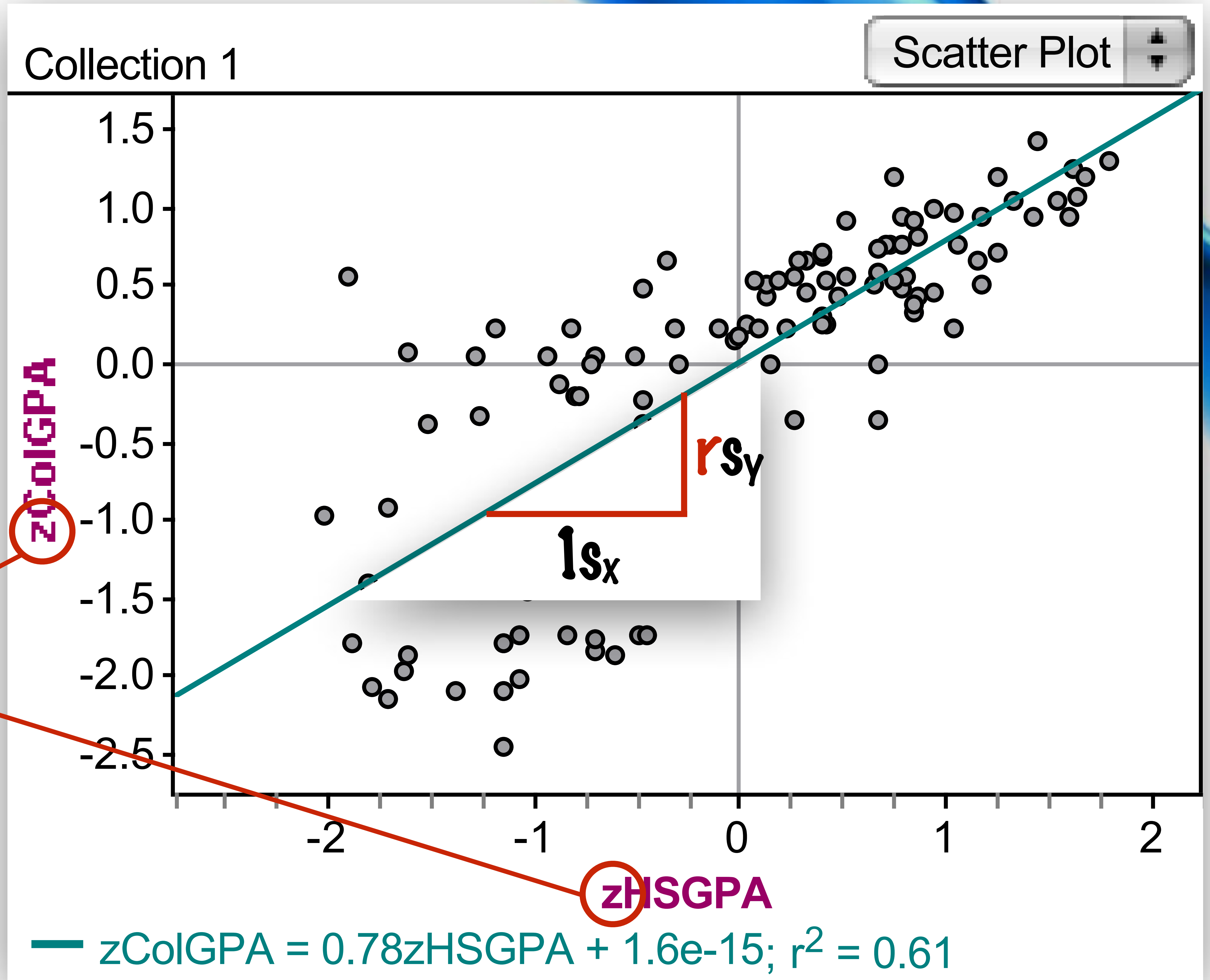
$$b_1 = \frac{r s_y}{s_x} = r \frac{s_y}{s_x}$$

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Correlation and the LSRL

- So now you know that a line that models our scatterplot of **z-scores** goes through the mean z-score values, the origin (0, 0),

- Since the standard deviation of z-scores is 1, The slope of that line from z scores is **r**.



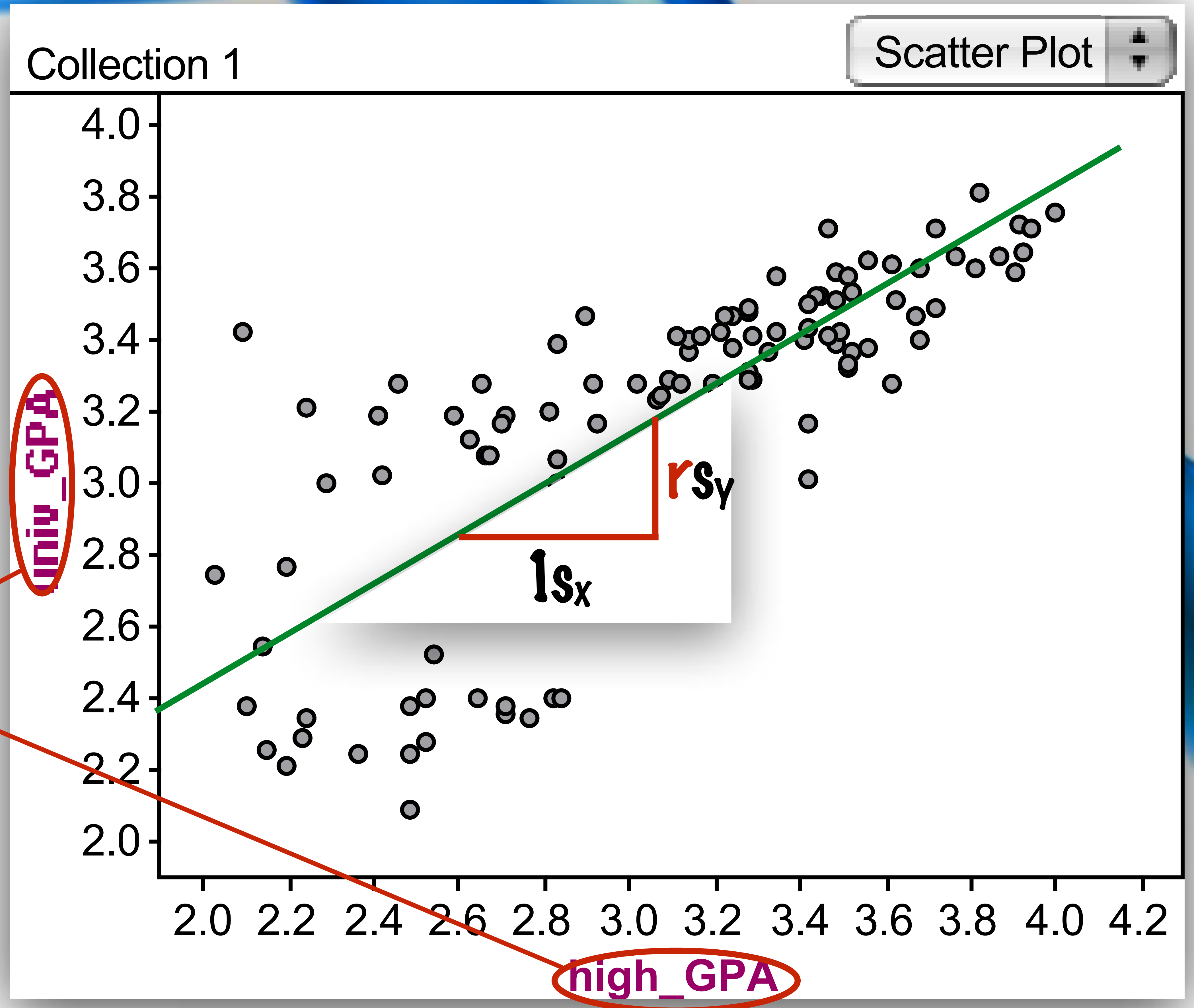
Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Correlation and the LSRL

- So now you know that a line that models our scatterplot of **raw scores** goes through the **mean raw scores**, $(\bar{x}, \bar{y})$.

- The standard deviation of raw scores are s_x, s_y and the slope of that line is …

$$b_1 = r\frac{s_y}{s_x}$$

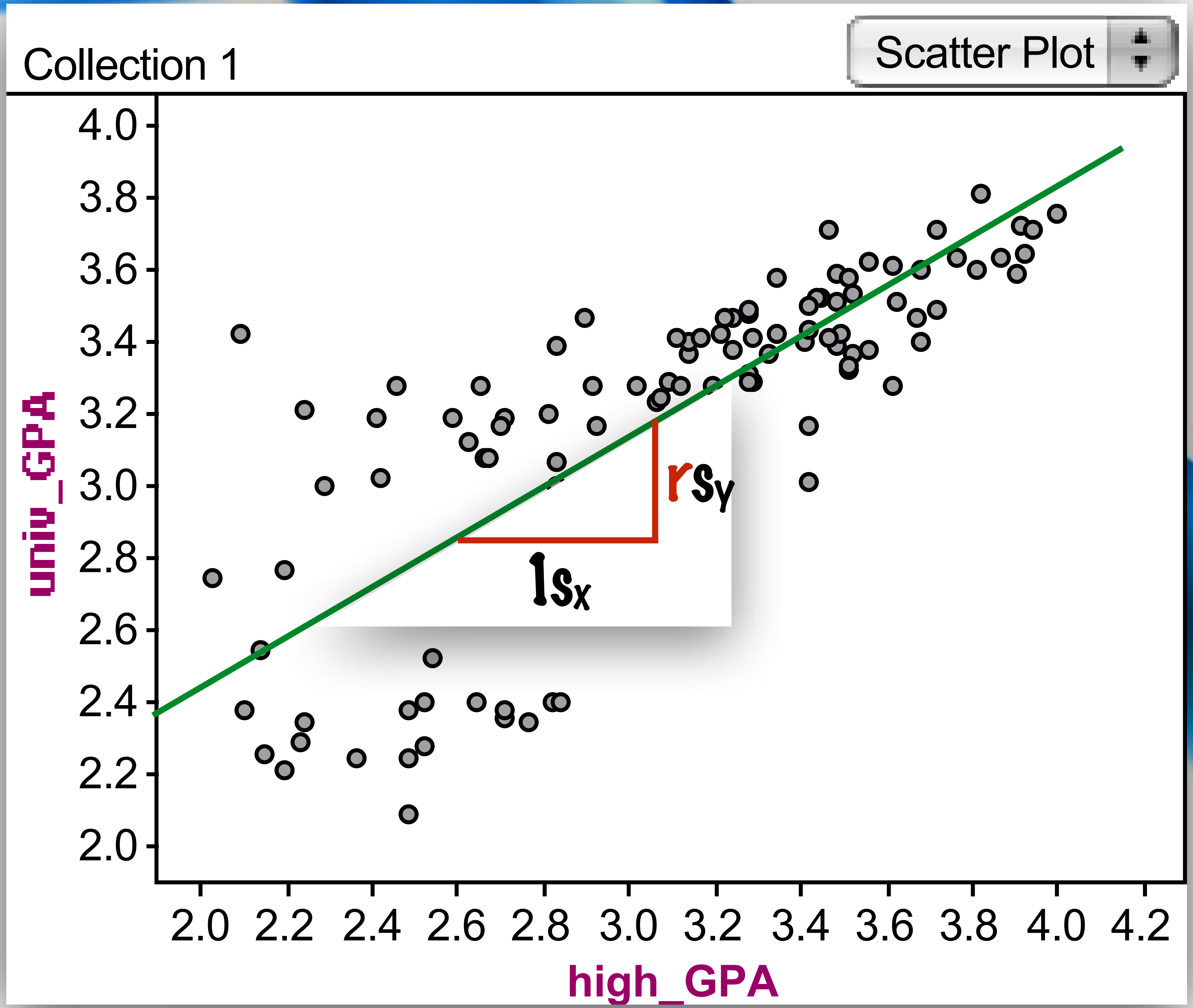


Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Correlation and the LSRL

- Thus, moving any number of standard deviations away from the mean in x moves us r times that number of standard deviations away from the mean in y.

- $|r|$ cannot be bigger than 1, so each predicted y tends to be closer to its mean than its corresponding x.



- This property of the linear model is called **regression to the mean**; the line is called the **regression line**.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Definition

- A regression line is a line that models how changes in an explanatory variable (x) **predict** changes in a response variable ($\hat{y}$).

- Thus, the regression line is used to **predict** the value of $\hat{y}$ for a given value of x.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

The Regression Line in Real Units.

- Remember from Algebra that a straight line can be written as: $y = mx + b$
  - In Statistics we use a slightly different notation:

$$\hat{y} = b_1 x + b_0 \quad \text{or} \quad \hat{y} = b_0 + b_1 x$$

- We write $\hat{y}$ to indicate that the points that satisfy this equation are our **predicted** values, not the actual data values (y).
- The model that is a straight line are our **predictions.**
- The more closely the model fits the data, the closer the data values will fit around the line of best fit (lsrl).

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

The Regression Line in Real Units.

- We write b_1 and b_0 for the slope and y-intercept of the line.
  - b_1 is the slope of the regression line, which tells us how much the response variable, $\hat{y}$, changes with a one unit change in the predictor variable, x.
  - b_0 is the y-intercept, which tells where the line crosses (intercepts) the vertical (y) axis. That tells us the predicted value of the response variable when the value of the predictor variable is 0.
- You want to remember these definitions. You will be asked to repeat them, repeatedly.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

The Real World

- Your TI-84, and many texts, use

$$\hat{y} = a + bx \quad \text{or} \quad \hat{y} = ax + b$$

for the regression equation.

- It is only important that you keep in mind which value is the **slope** of the regression line and which value is the **y-intercept**.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Reporting Slope and Intercept

- When interpreting the slope of the regression model the sentence frame you will use is:

Unit

Explanatory Variable

- For every additional _____ in ______ , there is a predicted ________ of ______ _____ in the ________.

Slope

Units

Response Variable

- Do not be creative or clever. Do not attempt to impress me as a vocabulist. Just use some variation of the given sentence frame or risk not getting credit.

Reporting Slope and Intercept

- **The y intercept is the predicted value of the response variable when the value of the predictor variable is 0.**

- Sometimes this is a meaningless statistic, and sometimes it is meaningful.
- What is the cost of renting a car when the number of miles driven is 0?
  - This may make sense since you may rent the car but never use it.
- What is the weight of a person that is 0" tall?
  - This definitely is nonsense.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

LSRL in real units

- In our model, we have a slope (b_1)
  - The slope is built from the correlation and the standard deviations:

$$b_1 = r\frac{s_y}{s_x}$$

  - Our slope is in **units of y** (response variable) **per unit of x** (predictor variable).
  - The size of the slope is determined by the units of the variables.
  - The size of the slope is **NOT** a measure of strength or significance.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

LSRL in real units

- In our model, we also have an intercept (b_0)
  - The intercept is built from the data **means** and the slope:

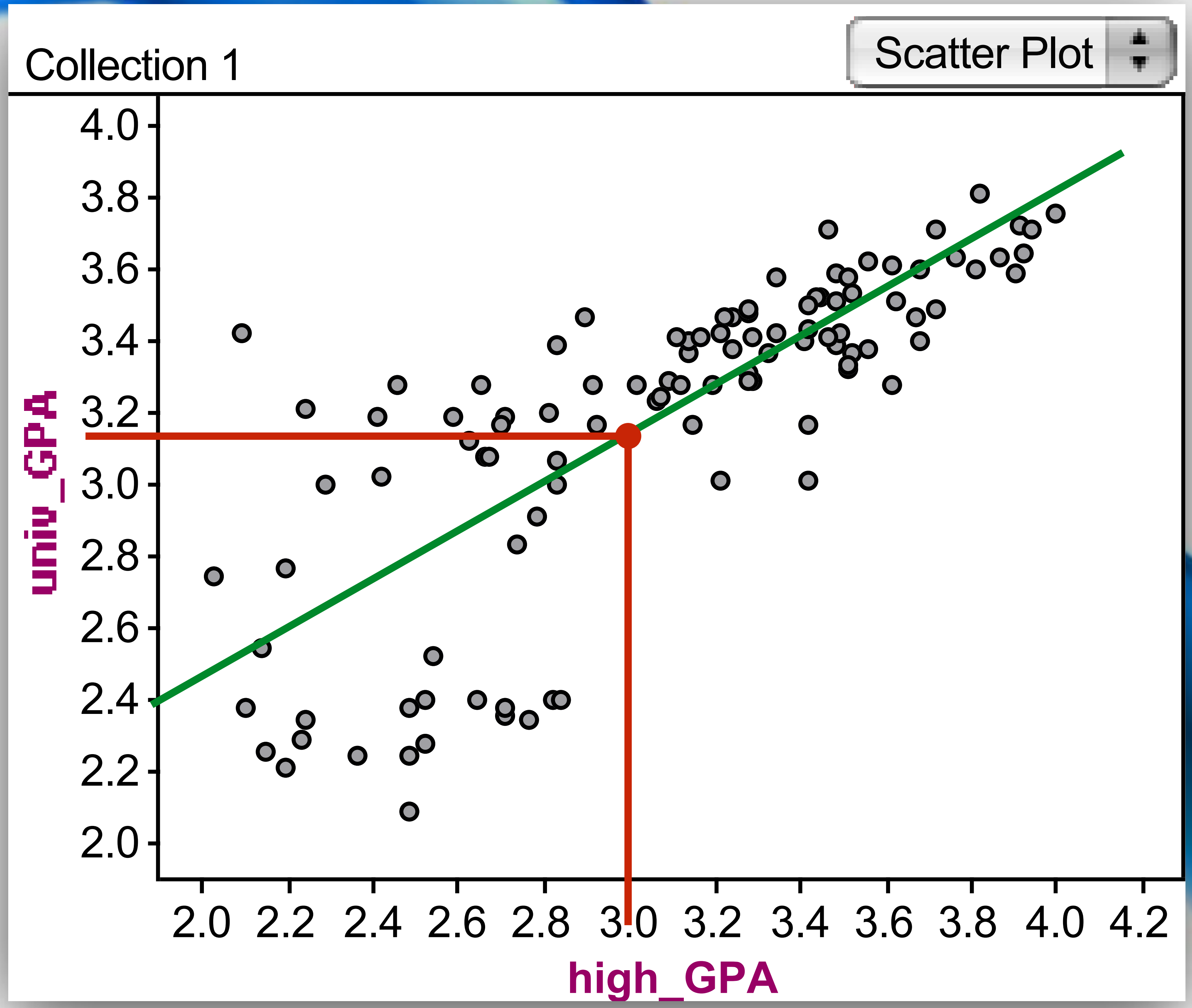
$$b_0 = \bar{y} - b_1\bar{x}$$

  - Obviously, our intercept is always in units of y.
  - Remember this, you will be asked to find the regression model from computer output of the standard deviations given with r.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

High School vs College GPA

- The regression line for the HSGPA and College GPA seems to fit pretty well.
  - The model is: $\widehat{ColGPA} = 1.1 + 0.67\,HSGPA$
  - The **predicted** college GPA for student with a high school GPA of 3.0 is 1.1 + 0.67(3.0) = 3.1 1.



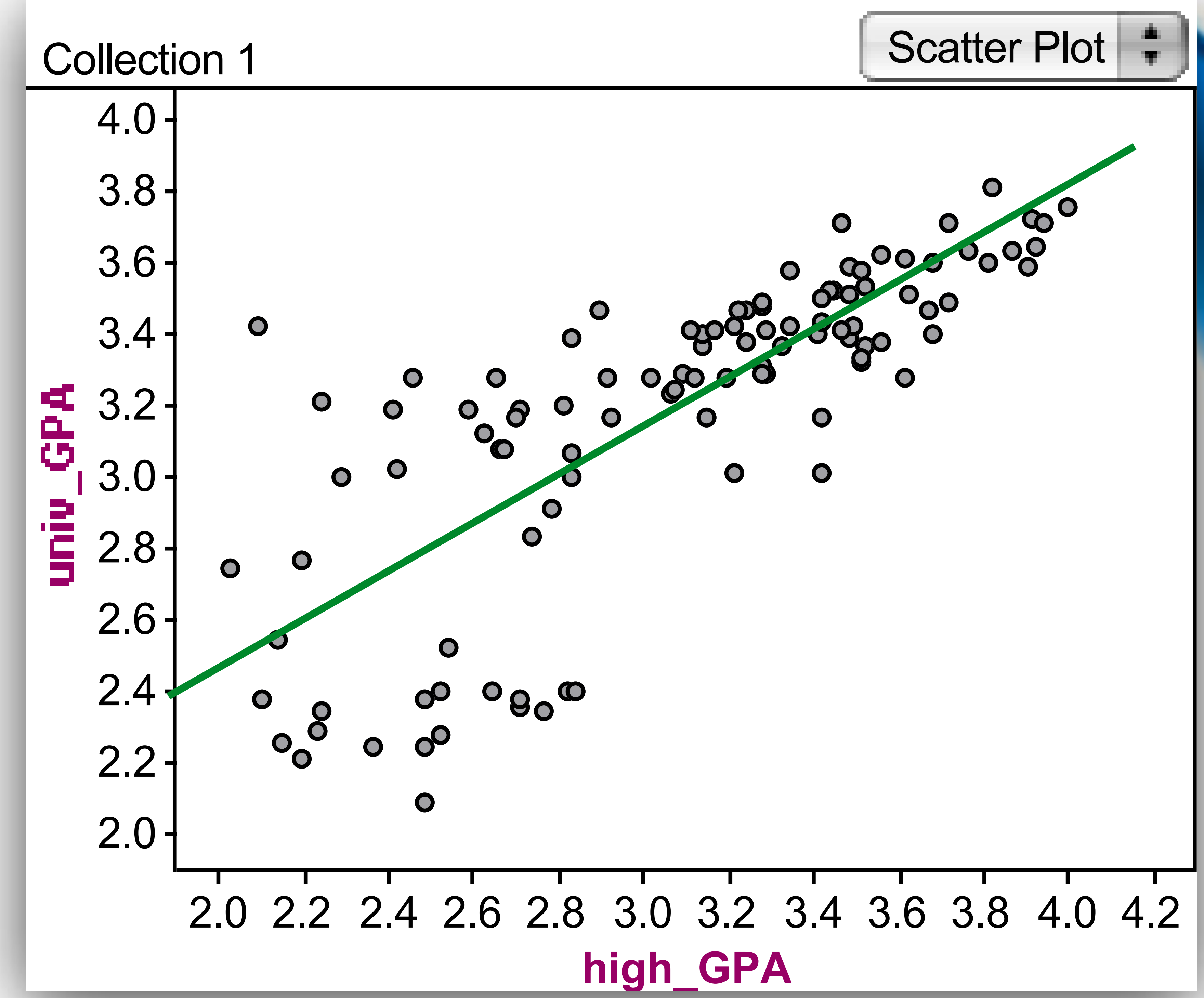
- For every increase of 1 point in high school GPA, there is a predicted increase of 0.67 points in the college GPA.

- We predict a college GPA of 1.1 when the high school GPA is 0.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

High School vs College GPA

- Since regression and correlation are from the same math, we must check the same conditions for regressions as we did for correlations:
    - Quantitative Variables
    - Sufficiently Linear
    - No significant Outliers
- Our data is quantitative, the pattern in the scatter plot is sufficiently linear and there are some data points that could be outliers such as (2.1, 3.4) or (2.5, 2.1), but I do not think they significantly affect our model.
- But there is a problem with the data!



Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Another Example

- Suppose we collected the weight of a male white lab rat for the first 25 weeks after its birth.

  - A scatterplot of the weight (grams) and time since birth (weeks) shows a fairly strong, positive linear relationship. The linear regression equation models the data fairly well.

$$\widehat{weight} = 100 + 40(time)$$

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

- The slope of 40 indicates that as the rat **ages one week** the model **predicts** an **increase** in weight of **40 grams**.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Growth of a rat

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

$$\widehat{weight} = 100 + 40(time)$$

  - The y-intercept is 100 grams which is the predicted weight of the rat at 0 weeks (birth weight).

- 3. Predict the rat's weight after 16 weeks.

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

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Growth of a rat

- 4. Should you use this line to predict the rat's weight at age 2 years?

$$\widehat{weight} = 100 + 40(time)$$

- Use the equation to make the prediction and think about the reasonableness of the result. There are 52 weeks in a year and 454 grams in a pound.

$$\widehat{weight} = 100 + 40(104) = 4260 \text{ grams} \approx 9.4 \text{ pounds}$$

- The result is not reasonable and highlights the danger in extrapolating a regression line beyond the data.

- W e l l, maybe not

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

TI-84

- Enter the following data into the calculator and find r, r^2, and the formula for the least squares regression line, then describe your results.

Body weight	120	187	109	103	131	165	158	116
Backpack weight	26	30	26	24	29	35	31	28

STAT ➢ CALC

4:LinReg(ax+b)
8:LinReg(a+bx)

XList:
YList:
FreqList:
Store RegEQ:
Calculate

- $r = .794692667$
- $r^2 = .6315364361$
- $\hat{y} = 16.26492733 + .0907994319x$

$$\widehat{backpack} = 16.2649 + .0908(body\ weight)$$

- There is a moderately strong, positive, linear relationship between body weight and weight of the backpack. As Body weight increases, backpack weight tends to increase. About 63% of the variability in backpack weight is accounted for by body weight. For every 1 pound increase in body weight the pack increases by .09 pounds and a 0 pound person is predicted to carry a 16.3 pound backpack.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

By calculation

Body weight	120	187	109	103	131	165	158	116
Backpack weight	26	30	26	24	29	35	31	28

- Now find the linear regression equation by hand.

$$b_1 = r\frac{s_y}{s_x}$$

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

- r = .794692667, s_x = 30.29586252, s_y = 3.461523199

$$b_1 = r\frac{s_y}{s_x} = .794692667\frac{3.461523199}{30.29586252} = .0907994318$$

$$b_0 = \bar{Y} - b_1\bar{X} = 28.625 - .0907994318(136.125) = 16.26492735$$

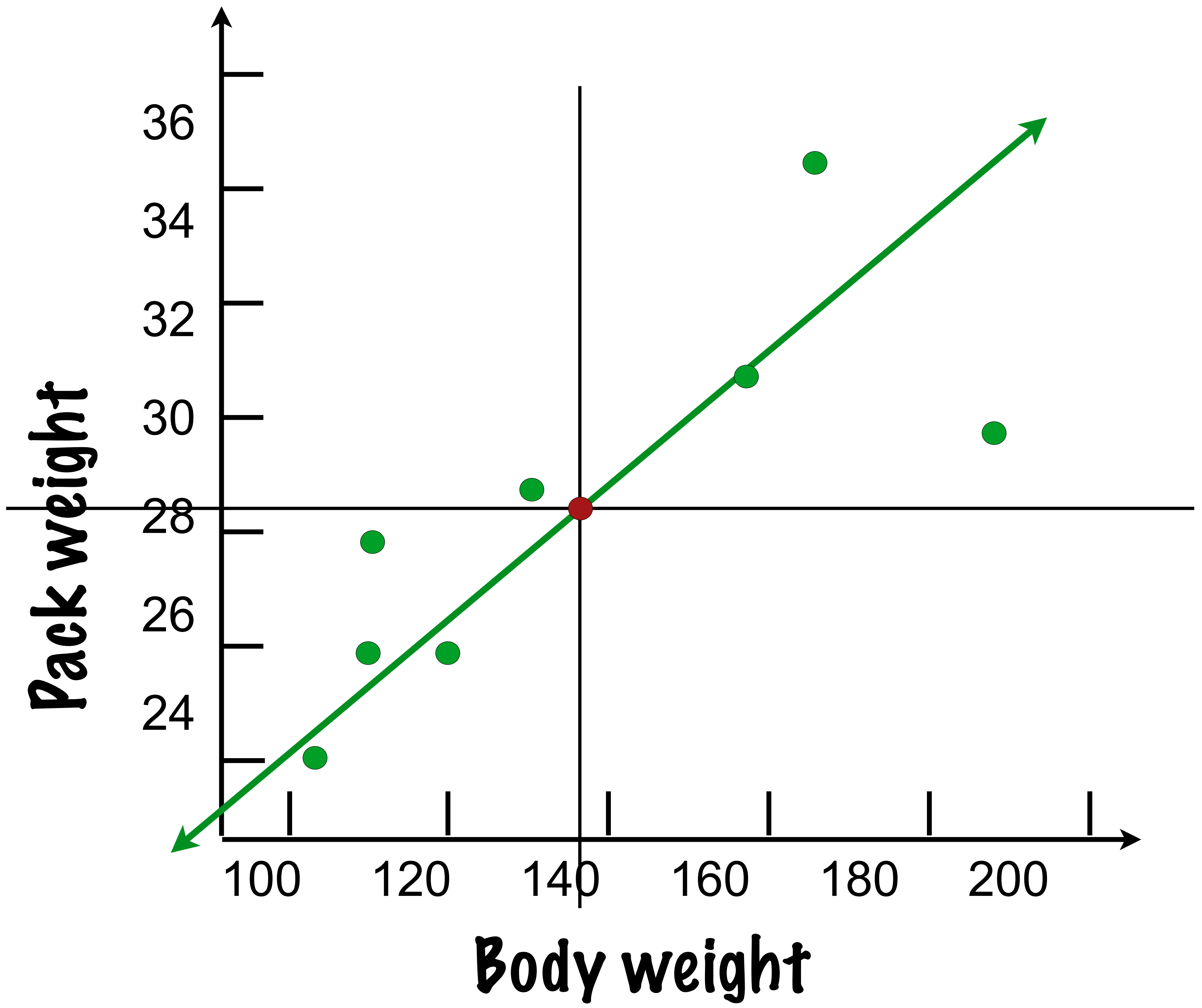
$$\widehat{backpack} = 16.2649 + .0908(body\ weight)$$

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Draw a picture

Body weight	120	187	109	103	131	165	158	116
Backpack weight	26	30	26	24	29	35	31	28

- Plot the points and draw the least squares regression line.



$$\widehat{backpack} = 16.2649 + .0908(body\ weight)$$

Not the reciprocal

- Remember that the regression model gives a prediction for the value of the response variable.

- We cannot use the same regression model to predict a value of the explanatory variable from the response variable. That requires a new model.

- The new model requires that we switch the roles of the variables.

$$b_1 = r\,\frac{s_x}{s_y}$$

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

- It is unlikely you will be asked to do this reversal, but it is important that you understand that the lsrl is not a two way model. The model is based on minimizing the residuals in the y variable, not in the x variable.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Not the reciprocal

- Reversing the roles of explanatory and response variables

$$b_1 = r\frac{s_X}{s_Y} = .794692667\frac{30.29586252}{3.461523199} = 6.955290605$$

$$b_0 = \bar{X} - b_1\bar{Y} = 136.125 - 6.955290605(28.625) = -62.97019357$$

$$\widehat{bodywt} = -62.97019357 + 6.955290605(pack\ weight)$$

$$\widehat{bodywt} = -62.97019357 + 6.955290605(27) \approx 124.823$$

- So if we find a pack weighing 27 lbs, we would predict a body weight of 124.8 lbs.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

What do the residuals have to say?

- The linear model assumes that the relationship between the two variables is a perfectly straight line.
  - The residuals are the part of the data that **has not** been accounted for by the model.

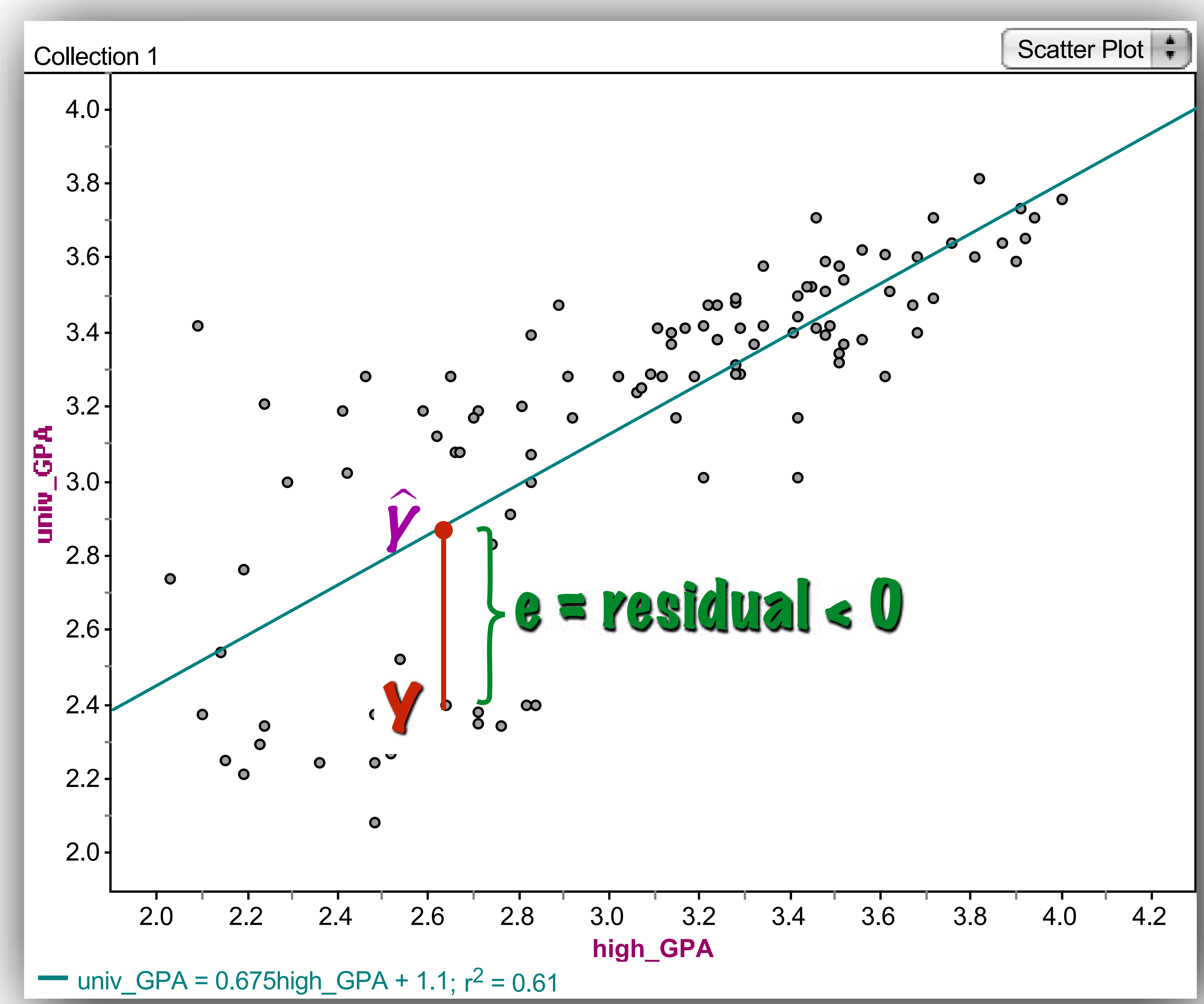
Data = Model + Residual

or (equivalently)

Residual = Data - Model

Or, in symbols,

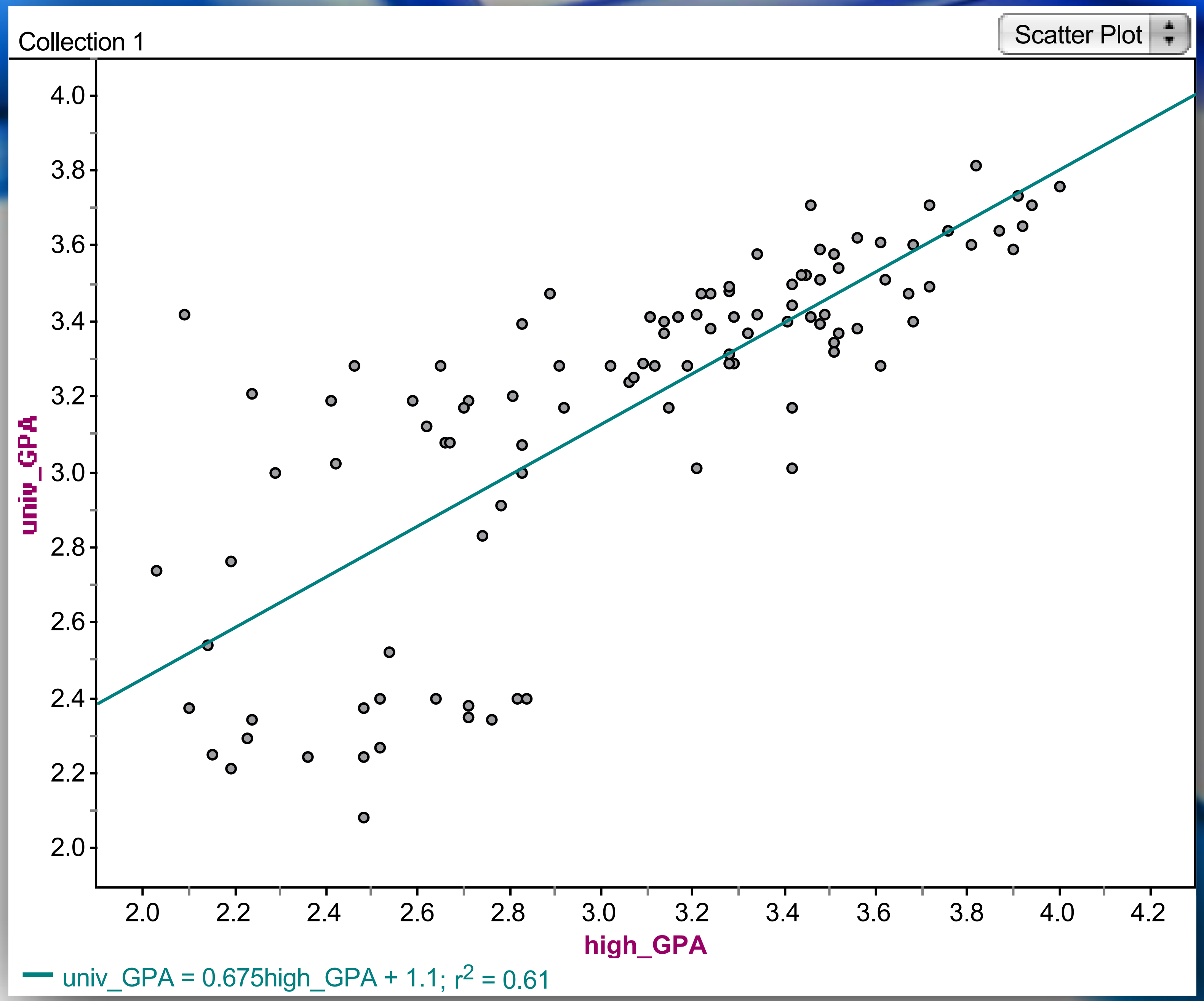
$$e = y - \hat{y}$$



Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Residuals Speak

- Residuals are a good indication of how well the model will predict response variable values. Does the model have value? Does the model make sense?
  - When a regression model is of value, there should be **nothing** of interest in the residuals.
  - In our GPA model, we note that the points above (positive residual) and points below (negative residual) are about even. That suggests the sum of the residuals should be 0.
  - Additionally, we hope that most points are near our LSRL so the residuals are small. Large residuals should be more infrequent, suggesting a unimodal and symmetric distribution of the residuals..
  - After we fit a regression model, we usually plot the residuals in the hope of finding...



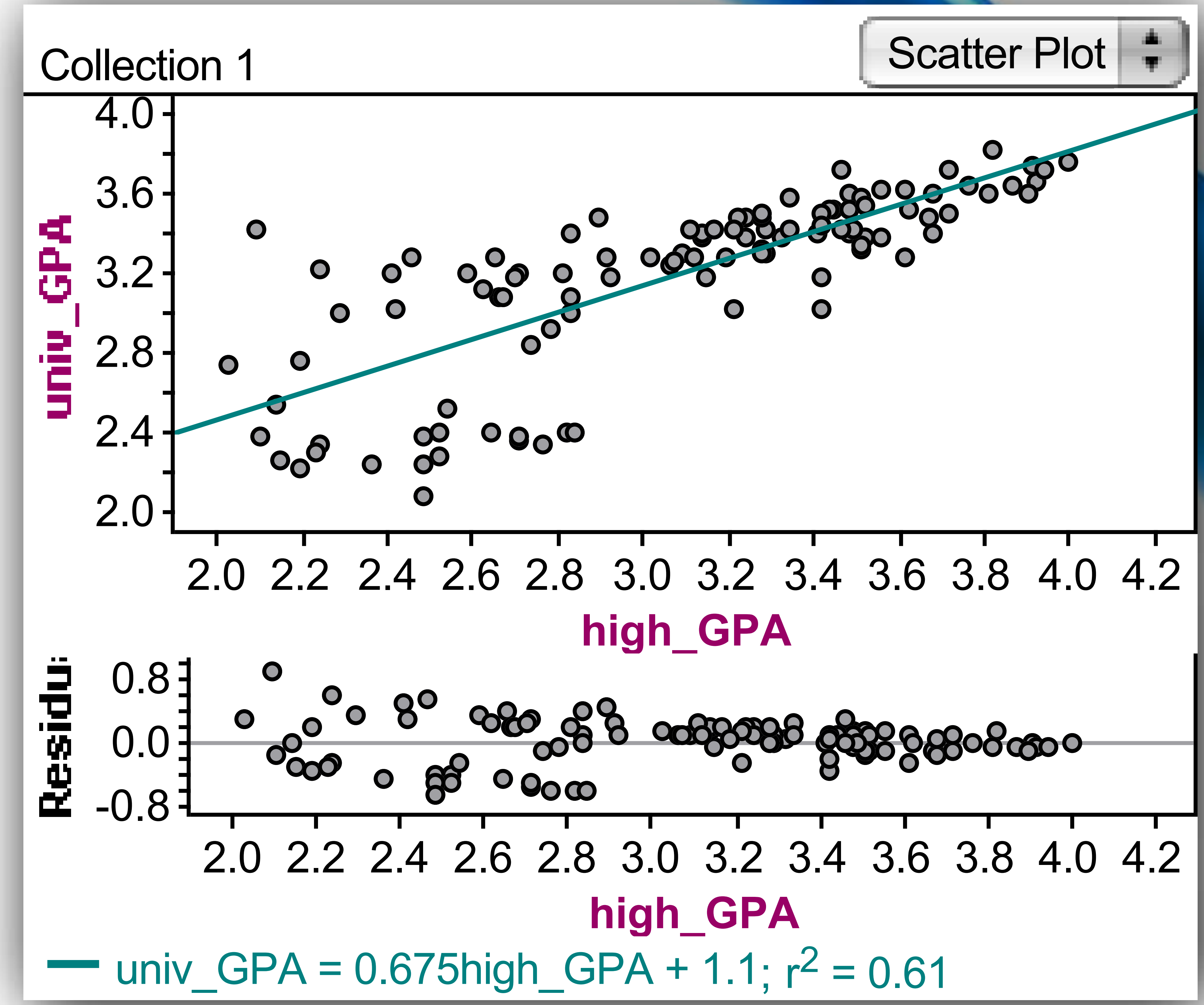
nothing.

Residual Plot

- A **residual plot** is a scatterplot of the residuals. The residual plot is the final test of how well our model represents the data.

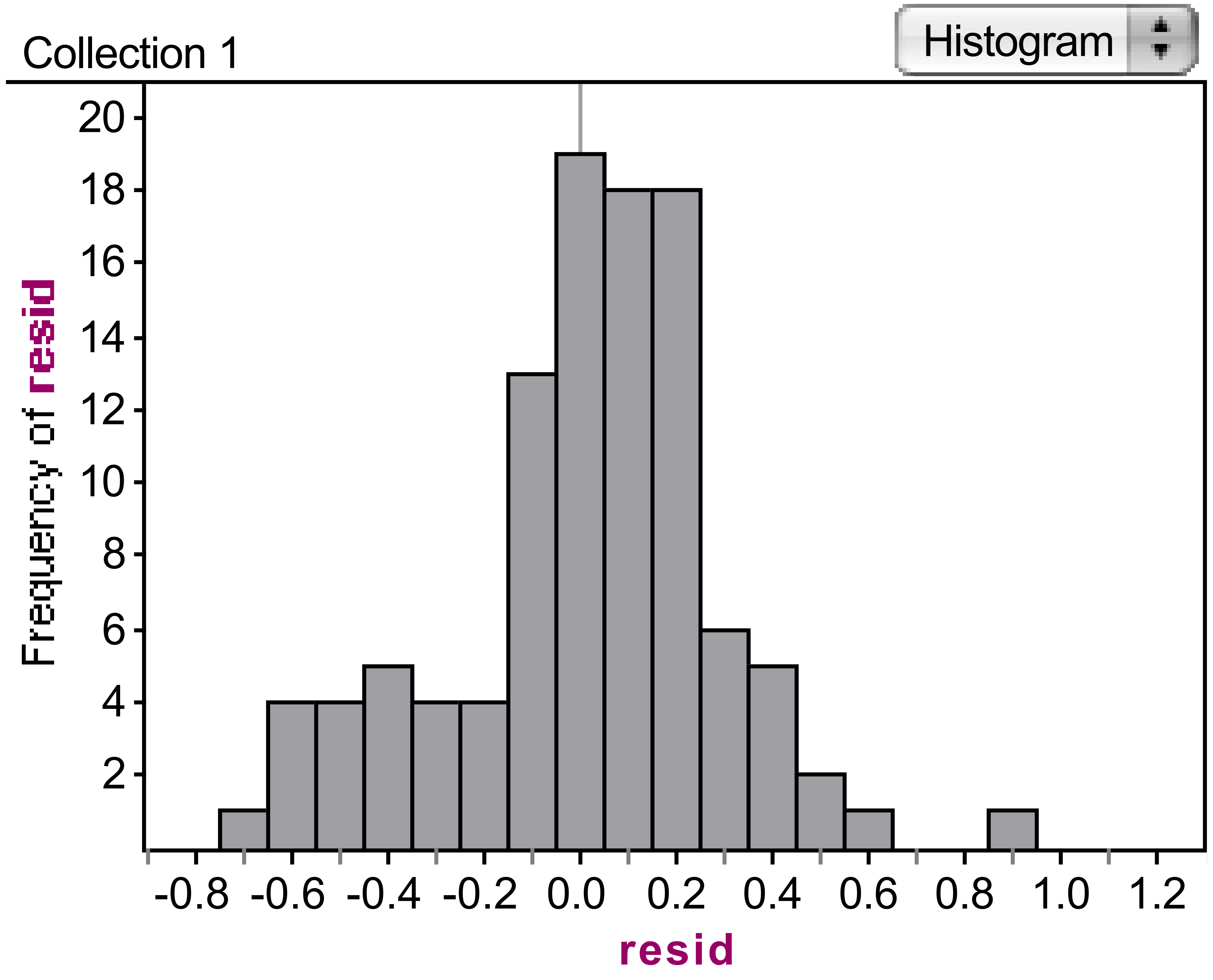
- There should no pattern to the residual scatterplot. If you do find a pattern in the residuals, there may be a problem with your model.

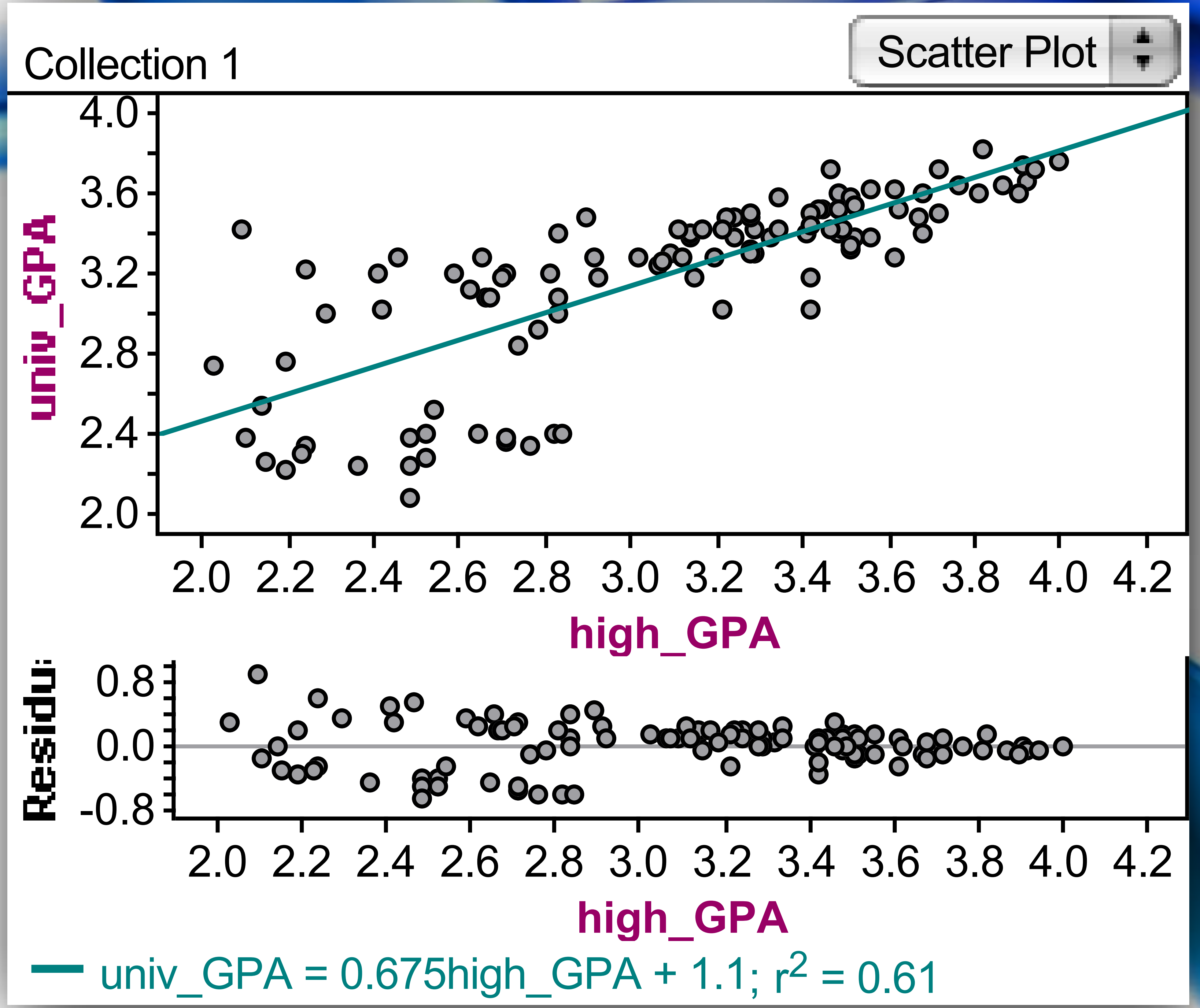
- A pattern in the residuals suggests a systematic relationship between the variables that has not been accounted for by our model.



Residuals

- A histogram of the residuals should appear unimodal and symmetric.





- The histogram of our gpa residuals appears sufficiently unimodal and symmetric. The model appears to meet all our requirements.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

TI-84

- Define L_3 as the predicted values from the regression equation. (L_1 and L_2 should still be the body weights and pack weights)
- Define L_4 as the observed y-value (L_2) minus the predicted y-value (L_3).
- Make sure all other plots are turned off. Choose Plot2 dot plot (first choice) with L_1 as x and L_4 as y. ZoomStat to see the residual plot.
- Just for snicks and giggles, lets look at the statistics of the residuals (L_4).
  - The mean should be 0, any difference is due to rounding (remember what I said about rounding too soon?).
  - What do you think s_x tells us? (s_x = 2.101)

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

TI-84

- Finally, we can find the residual plot by using a list created by the calculator called (not surprisingly) **resid**.

- Each time you ask for a regression analysis the calculator calculates the residuals automatically and stores them in a reserved calculator list 'resid'.

- We cannot manipulate or find the statistics for the 'resid' list but we can use it to create a residual plot.

- x list = L_1 or wherever you put the explanatory variable.

- y list = 2nd STAT (List) 7:Resid ENTER

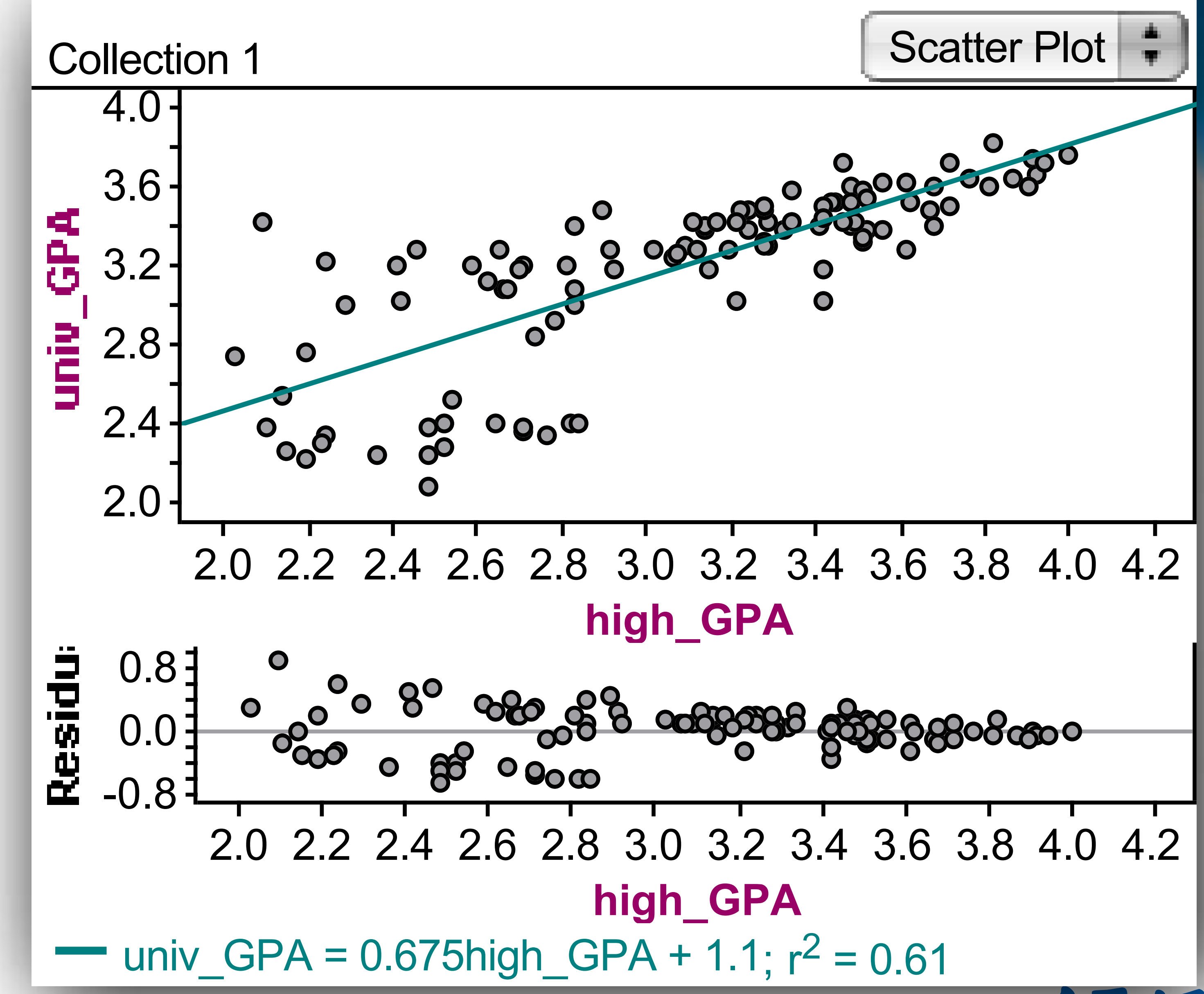
The Residual Standard Deviation

- The standard deviation of the residuals, s_e, measures the variability of the residuals or how the points vary above and below the regression line.
- For the GPA data, s_e, is .2801. Our data has a typical deviation of .28 from the predicted value.
- Check to make sure the residual plot has a consistent amount of scatter throughout the distribution. Check the **Equal Variance Assumption** with what your book calls the "**Does the Plot Thicken? Condition**".
- Our data does display changes in the variability of the residuals. It appears the model is not as accurate at lower HSGPA levels. So we note that in our conclusion.

Collection 1

	resid
	0.0116819
	105
	0.280099
	0.0273349
	0

S1 = mean ( )
S2 = count ( )
S3 = stdDev ( )
S4 = stdError ( )
S5 = count (missing ( ))



Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Another example

- Calculate and plot the residuals for our example of Body and Backpack weights.

$$\widehat{backpack} = 16.3 + .09(body\ weight)$$

Body weight	Pack weight
120	26
187	30
109	26
103	24
131	29
165	35
158	31
116	28

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

The Residual Standard Deviation

- We can estimate the SD of the residuals, s_e, using:

$$s_e = \sqrt{\frac{\Sigma e^2}{n-2}} \qquad s_e = \sqrt{\frac{\Sigma\left(y-\hat{y}\right)^2}{n-2}}$$

- look familiar? This is a modification resulting from two variables being involved.

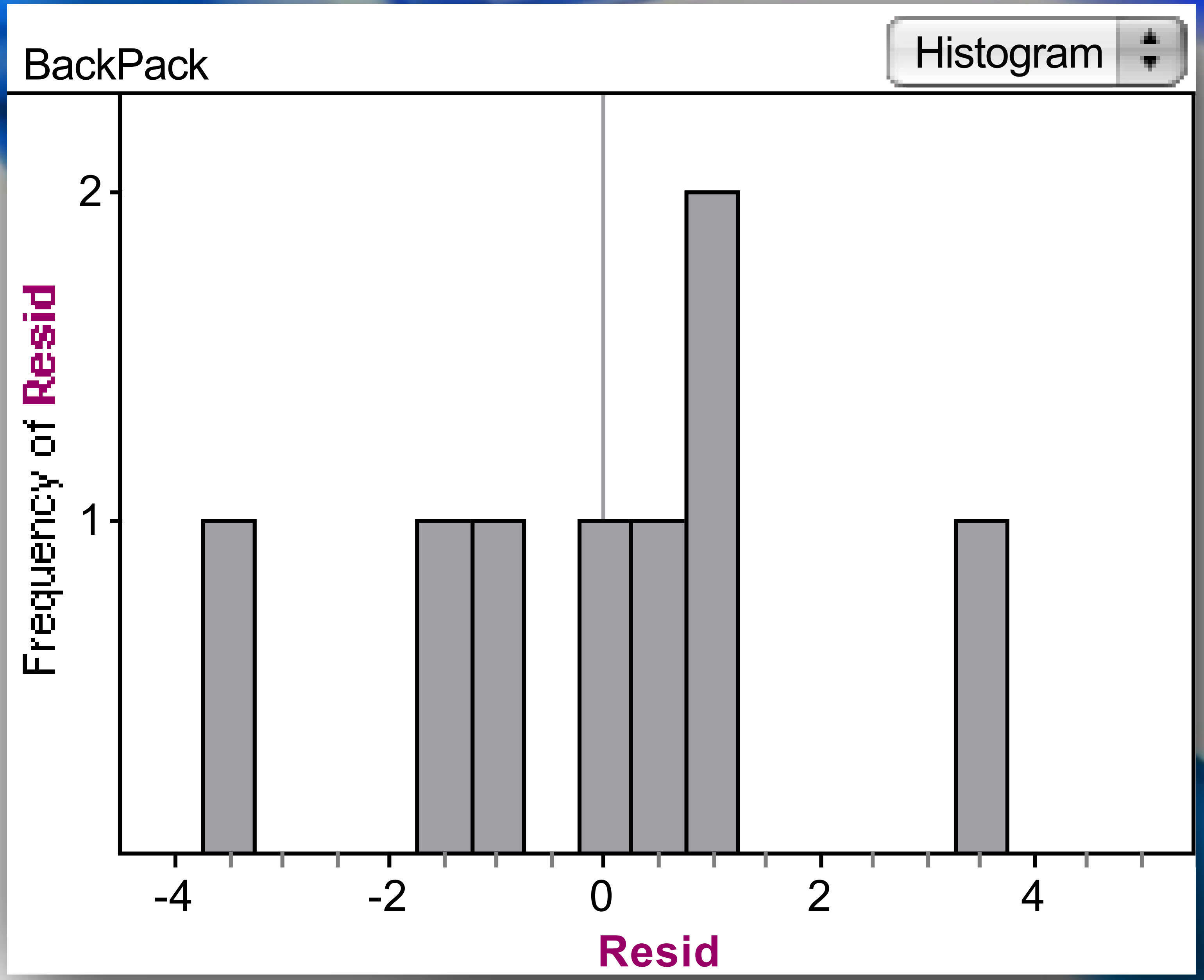
- but we will let the calculator do that for us if we ever need to do the calculations.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

The Residual Standard Deviation

- We do not need to subtract the mean because the mean of the residuals is 0. $\bar{e} = 0$

- Make a histogram (or normal probability plot) of the residuals of the backpack data. It should look unimodal and roughly symmetric. (Does it?)



- Then we can apply the 68-95-99.7 Rule to see how well the regression model describes the data.
- Do 95% of our residuals fall within 2 standard deviations of 0?

- 100% of our residuals fall within 2 standard deviations of 0.

BackPack

Resid	-0.03515
	8
	2.10119
$S_e = 2.1$	0.742881
	0

S1 = mean ( )
S2 = count ( )
S3 = stdDev ( )
S4 = stdError ( )
S5 = count (missing ( ))

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

The coefficient of determinations: r^2

- The coefficient of determination r^2 is the portion of the variation in the values of y that is **accounted for** by the least-squares regression model.

- We can calculate r^2 using the formula: $R^2 = 1 - \frac{SSE}{SST}$

  - SSE = Sum of Squares Error and SST = Sum of Squares Total

    - Where $SSE = \sum_{i=1}^{n}\left(Y - \hat{Y}\right)^2$ $SST = \sum_{i=1}^{n}\left(Y - \bar{Y}\right)^2$

  - R^2 is 1 – the proportion of variation not accounted for by the model out of the total variation in the response variable. Thus the variation that is accounted for.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

The coefficient of determinations: r^2

- The variation in the residuals is one big factor in assessing how well the model fits.
  - In our GPA data the variation in the residuals was .2801.
  - The variation in the response variable was .4472.
  - That is comforting, suggesting the model has less variability than the original data.

Collection 1

	resid
	0.0116819
	105
	0.280099
	0.0273349
	0

S1 = mean ( )
S2 = count ( )
S3 = stdDev ( )
S4 = stdError ( )
S5 = count (missing ( ))

Collection 1

univ_GPA	3.17286
	105
	0.447194
	0.0436416
	0

S1 = mean ( )
S2 = count ( )
S3 = stdDev ( )
S4 = stdError ( )
S5 = count (missing ( ))

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

The coefficient of determinations: r^2

- If the correlation were 1.0 and the model predicted the college GPAs perfectly, the residuals would all be zero and have no variation.
  - What we found for the correlation was 0.7795 – not quite perfection.
  - But we did see that the model residuals had less variation than total college GPA.
  - We can determine how much of the variation is accounted for by the model and how much is left in the residuals.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

The coefficient of determinations: r^2

- The squared correlation, r^2, gives the fraction of the data's variance accounted for by the model.
  - Thus, $1 - r^2$ is the fraction of the original variance left in the residuals.
  - $1 - r^2$ is the coefficient of non-determination.
  - For the GPA data, $r^2 = .7915^2 = .6076$, Thus 61% of the variability in college GPA is accounted for by variability in high school GPA, and 39% of the variability in college GPA has been left in the residuals.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

The coefficient of determinations: r^2

- All regression analyses include this statistic, although by tradition, it is written R^2.
- An R^2 of 0 means that none of the variance in the data is in the model; all of it is still in the residuals.
- When interpreting a regression model you must always correctly interpret R^2.
- Let me say that again.
- Anytime you are working with a correlation and regression model you must always, every single time, correctly interpret R^2.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Let Me Repeat That

- When discussing R^2 in the model this is the sentence frame you will use:

$R^2 x100$

Response Variable

- ____% of the variation in _____ is accounted for by variation in the ______.

Explanatory Variable (or model)

- As I have indicated previously, do not be creative or clever. Just use the sentence structure or risk not getting credit.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

How Big Should R^2 Be?

- R^2 is always between 0% and 100%. What makes a "good" R^2 value depends on the kind of data you are analyzing and on what you intend to do with it.

- The standard deviation of the residuals is a good indicator of the validity of the regression model by telling us how much our actual data deviates from the predicted values.

- When reporting a regression model you must report Pearson's r, the slope and intercept of the model, and R^2. This will give the audience a complete picture of the value of the model.

Reporting R^2

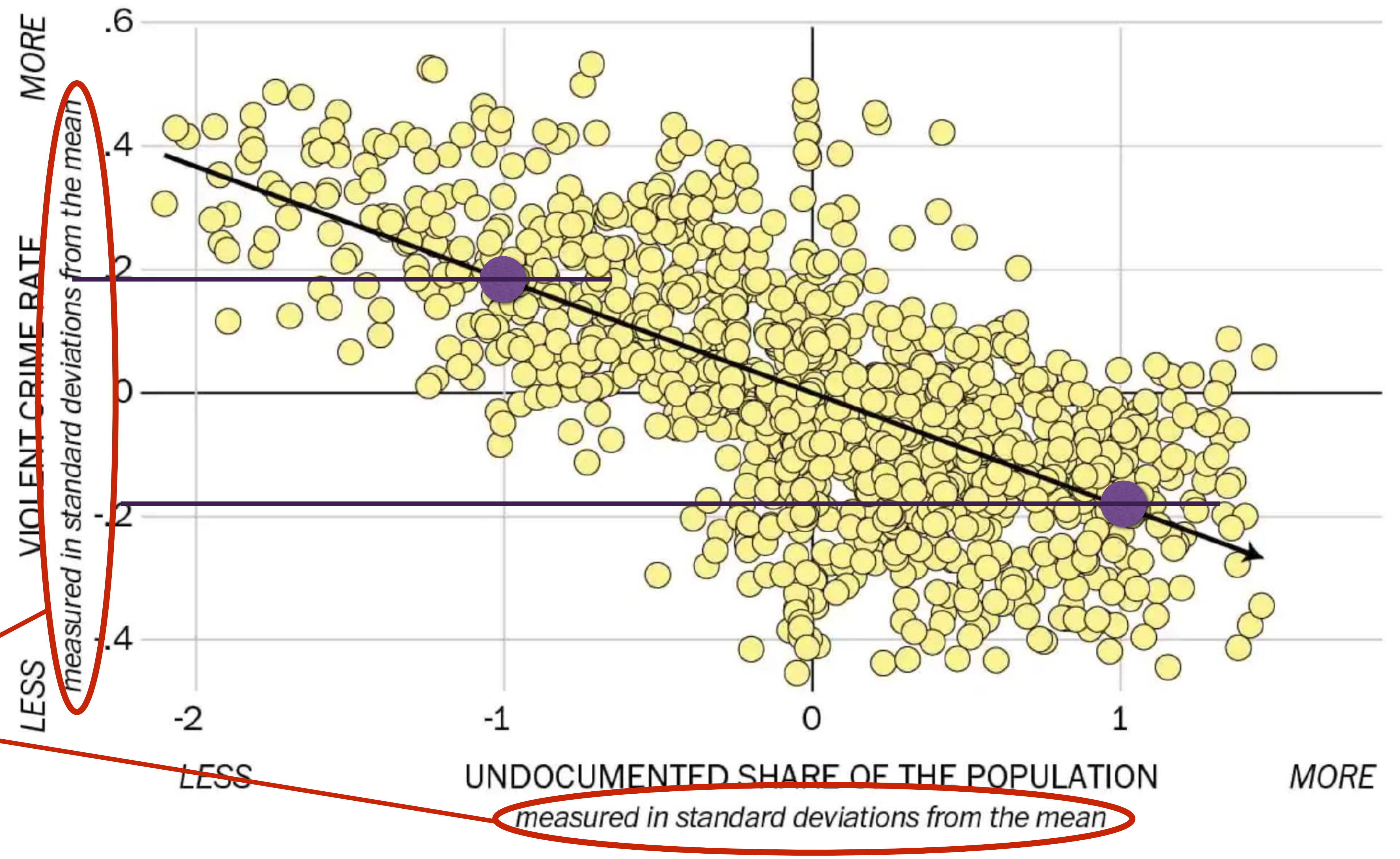
- Along with the slope and intercept for the least squares regression line, you should always report R^2 so that readers can judge for themselves how successful the regression is at fitting the data.

- In interpreting and analyzing a regression model you now have several items that must be included. Scatter plot of data, description of distribution, r, R^2, interpretation of R^2, the model itself, description of model (interpretation of slope and intercept), scatterplot of residuals, histogram of residuals (and/or normal probability plot).

Real Example

- Here is some recent data. Note scale along both axes.
  - Describe the relationship.
  - Negative, moderately weak, linear, no outliers. As undocumented share of population increases, violent crime tends to decrease.
  - Estimate r.
  - Since these are z-scores, the slope = r.
    - (-1,.2) and (1, -.2)
    - r = -.2

State-level correlation between levels of undocumented immigrants (bottom axis) and violent crime measures (left axis), 1990 through 2014. Each circle represents one state observed in one year.



Source: Light & Miller, 2018 WAPO.ST/WONKBLOG

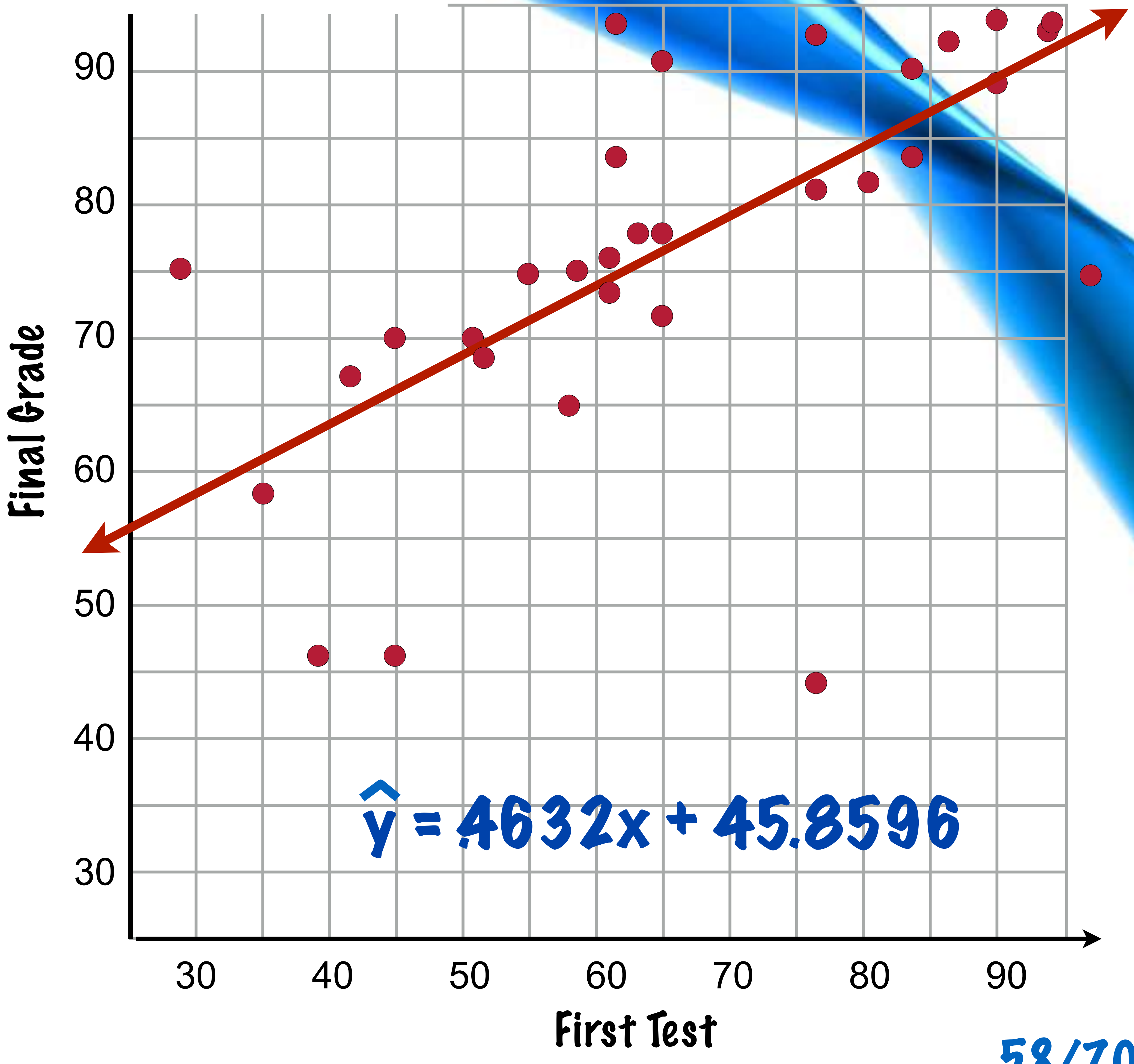
Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

How well does the first test predict final grade?

- This is actual data, taken from an actual statistics class in a previous year. Find a linear model.
- Describe the distribution.

Grade First Test	Final Grade
94	93
90	87
42	67
39	46
97	75
65	91
65	78
84	84
77	44
45	46
55	75
65	72
51	70
87	93
84	90

Grade First Test	Final Grade
52	69
90	94
35	58
61	83
58	65
45	70
61	76
61	73
67	78
58	75
61	93
77	81
29	75
81	82
94	94
77	93



Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

How well does the first test predict final grade?

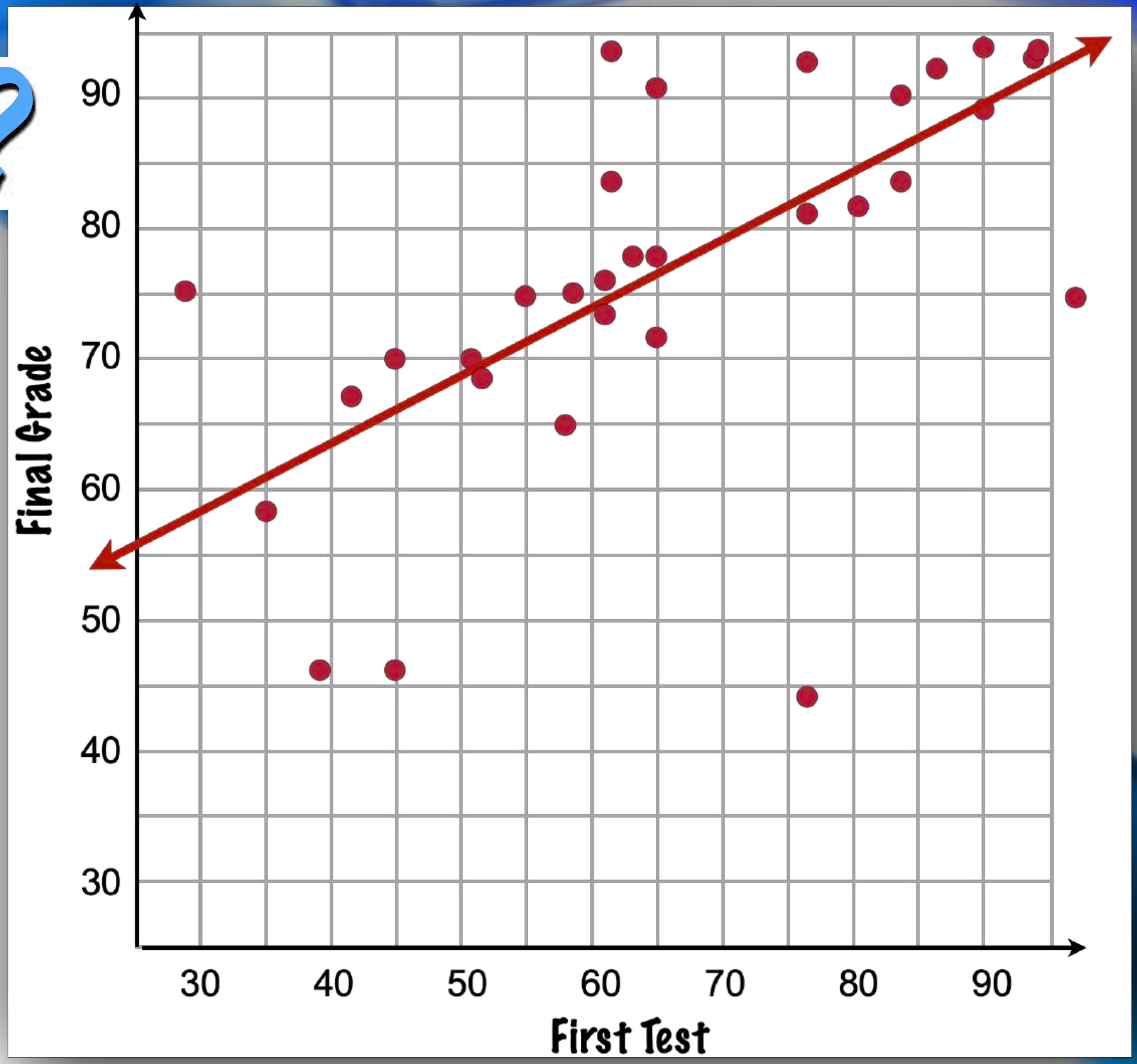
$\hat{y} = .4632x + 45.8596$

$r = .6161$, $r^2 = .3796$

$\hat{y}$ = predicted final grade x = score on first test

Predicted Final Grade = .4632(score on first test) + 45.8596

$$\widehat{FinalGrade} = .4632(firsttest) + 45.8956$$



There is a moderate, positive, linear relationship between score on first test and final grade. As scores on first test increase, scores on final grade tend to increase. About 37.96% of the variability in final grade is accounted for by score on first test. There are distinct outliers at (77, 44), (29, 76), and (96, 75). For every 1 percentage point increase in first test score the final grade is predicted to increase by .4632 points and a 0 on first test predicts a 45.90% final grade.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Interpreting regression output

cohol

Interpreting regression output

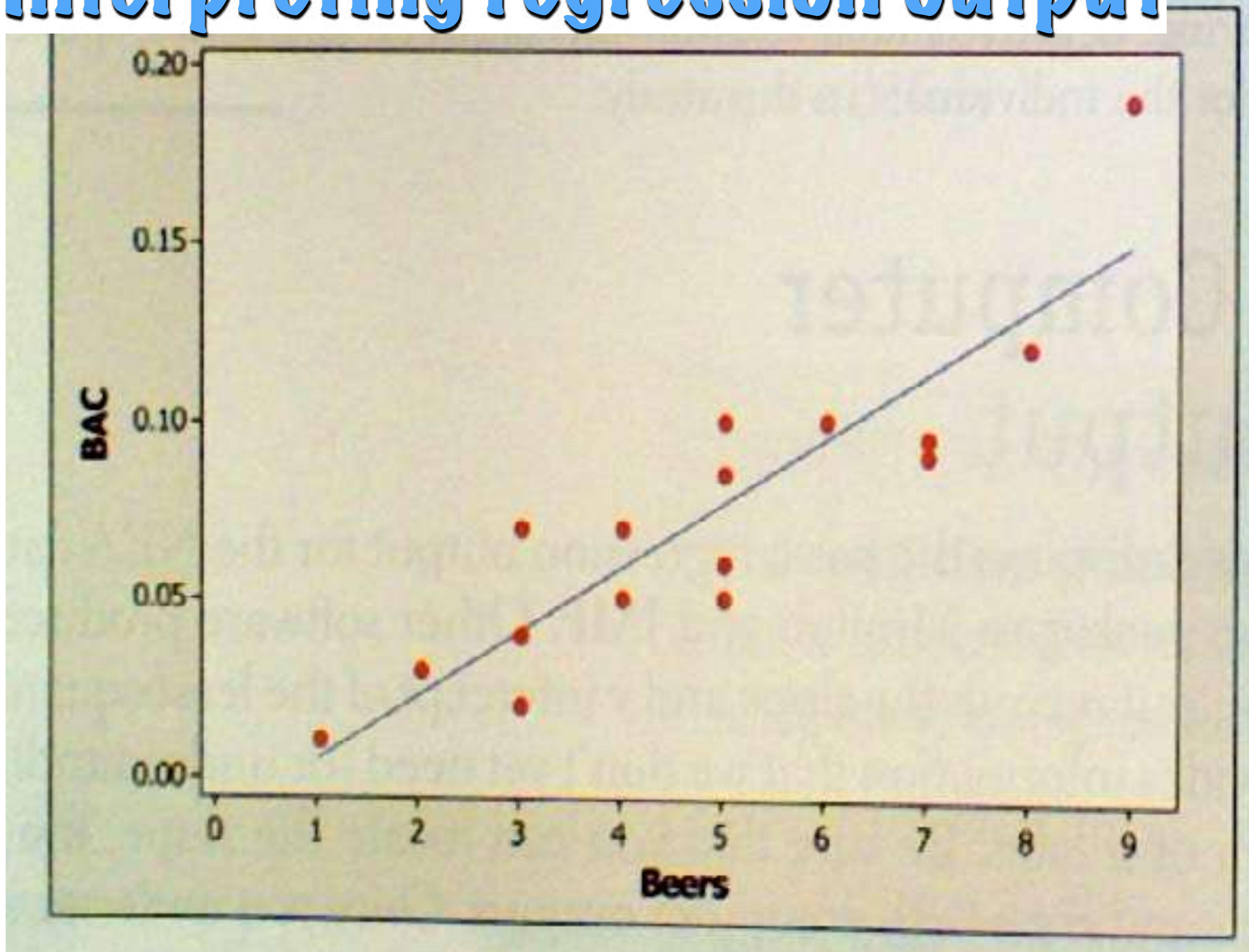
How well does the number of beers a person drinks predict his or her blood alcohol content (BAC)? Sixteen volunteers with an initial BAC of 0 drank a randomly assigned number of cans of beer. Thirty minutes later, a police officer measured their BAC. Least-squares regression was performed on the data. A scatterplot with the regression line added, a residual plot, and some computer output from the regression are shown below.

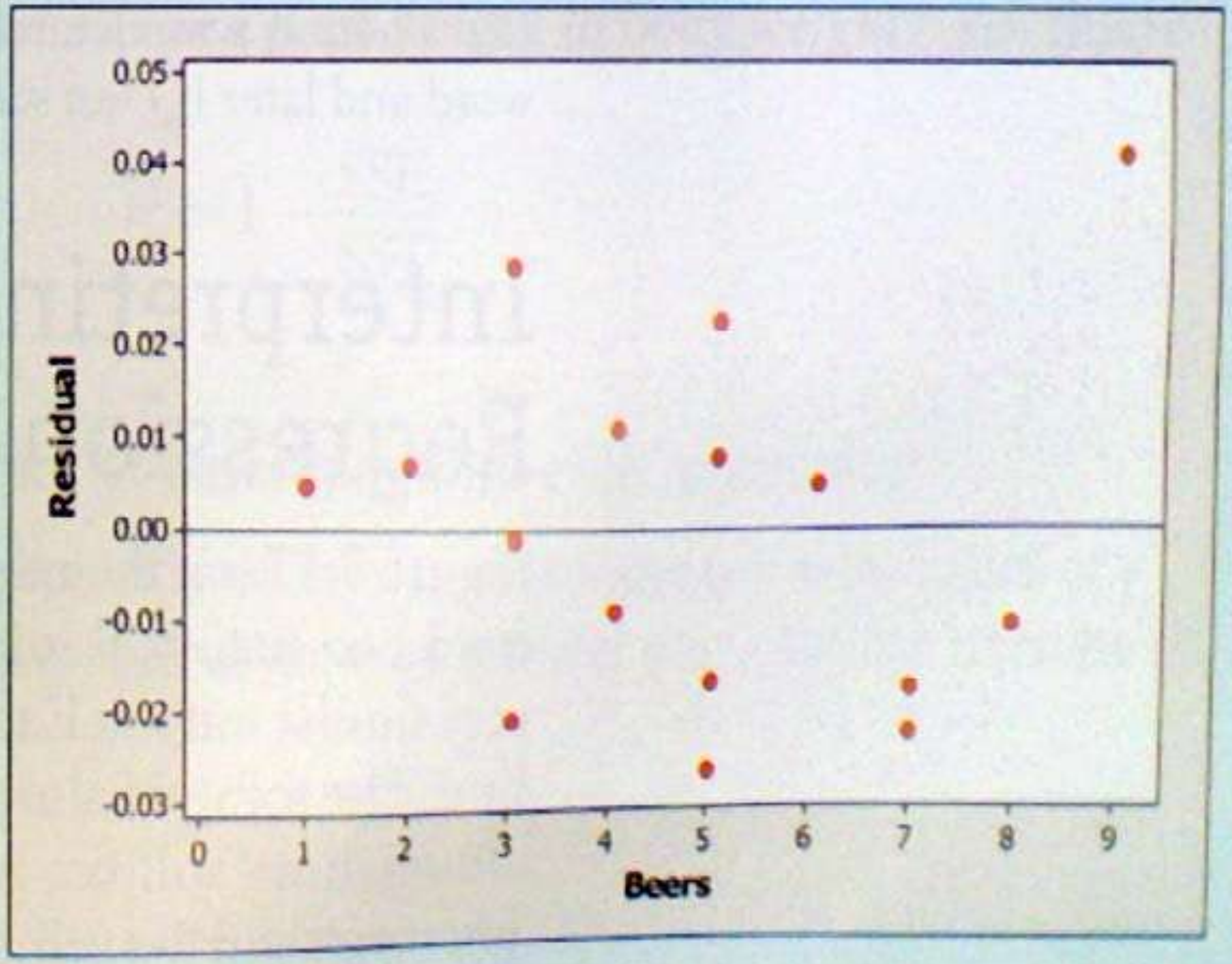
```
Dependent variable is:  BAC
No Selector
R squared = 80%
R squared (adjusted) = 78.6%
s = 0.0204 with
16 - 2 = 14 degrees of freedom
```

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	-0.012701	0.0126	-1.00	0.3320
Beers	0.017964	0.0024	7.48	≤0.0001

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Interpreting regression output





- Strong, positive, linear model, with one possible outlier at about (9, .19) as the number of beers increases blood alcohol trends to increase.

- No pattern in residuals. We are good to go.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Interpreting regression output

Dependent variable is: BAC — response variable

No Selector

R squared = 80% — r^2

~~R squared (adjusted) = 78.6%~~

s = 0.0204 with — s_e for residuals

16 - 2 = 14 degrees of freedom

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant (y-intercept)	-0.012701	0.0126	1.00	0.3320
Beers (explanatory variable)	0.017964 (slope)	0.0024	7.48	$\le$0.0001

Chapter 26

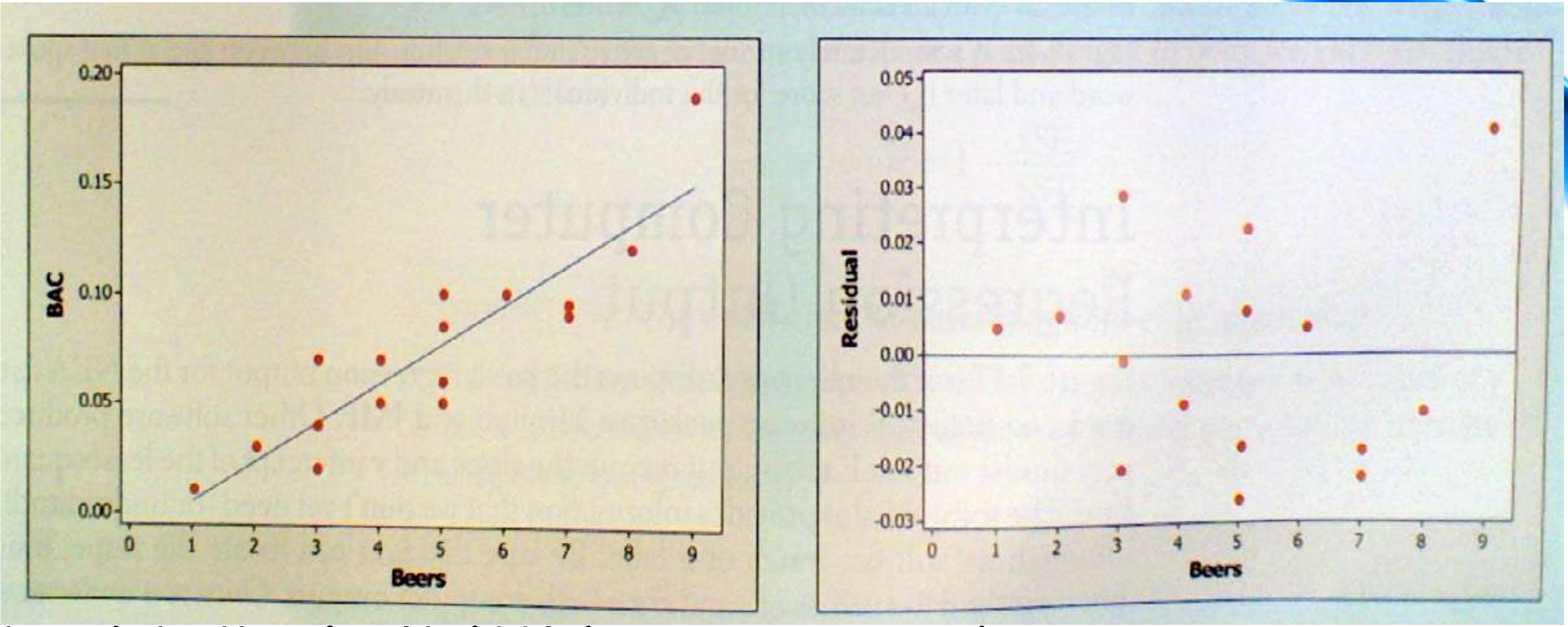
- The correlation would be. $r = \sqrt{.80} \approx \pm .894$

- The equation is $\widehat{BAC} = -0.0127 + .017964(beers)$
- For every additional beer consumed, the model predicts a BAC increase of about 0.018 BAC, 80% of the variability in BAC is accounted for by the model.
- A linear model does seem appropriate for this data.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Interpreting regression output

- The residuals are between -.03 and .03 except for the subject drinking 9 beers. On average, predictions of BAC using the regression line would be off by about s = 0.02 for the 16 people in the study.



- Given the legal limit for BAC of 0.08 that error may be too much.

Assumptions and Conditions

- **Quantitative Variables Condition:**
  - Regression can only be done on two quantitative variables (and not two categorical variables), so you must check and report on the quantitative condition.
- **Sufficiently Linear Condition:**
  - A linear model describes a relationship between the variables that is linear. If the data is not linear, a linear model is not appropriate.
  - A scatterplot will indicate that the assumption is reasonable.
  - If the data is not linear, you are done.
- Some nonlinear relationships can be saved by re-expressing the data to make the scatterplot more linear. We will revisit this at year end.

Assumptions and Conditions

- It is a necessary to check linearity again after computing the regression model when we can examine the residuals for any pattern that may have been missed in the scatterplot.
  - Ironically, we must run the model first to be completely certain the model is worth running.
- **Equal Variance Assumption Condition:**
  - Check the residual plot and the standard deviation of the residuals to summarize the scatter. The residuals should have the same spread throughout the domain of the predictor variable. Check for **changes in the spread** of the residual scatterplot.
- **Outlier Condition:**
  - Watch out for outliers.
  - Outlying points can dramatically change the correlation and the regression model.
  - In extreme cases outliers can even change the direction of the slope, completely misinforming the reader about the relationship between the variables.

Outliers and Influential Observations

- An outlier is an observation that lies outside the overall pattern of the other observations.
  - Points that are outliers in the y direction but not in the x direction of a scatterplot have large residuals. Outliers in the x direction may have undue influence on the model and, thus, not have large residuals.
- An observation is influential if removing it would significantly change the results. Points that are outliers in the x direction of a scatterplot are often influential for the least-squares regression line.

Reality Check: Is the Regression Reasonable?

- Statistics do not come from out of the blue. Statistics are based on data.
  - The results of a statistical analysis should make sense, be logical.
    - If the results are surprising, then either you have made a new discovery and added to the human lexicon about the world or your analysis is wrong.
- When you perform a regression, think about the slope and intercept and consider if the relationship makes sense.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

CCRs

- Don't fit a straight line to a nonlinear relationship.
  - Examine extraordinary points (y-values that appear far away from the linear pattern or extreme x-values).
  - **Do not extrapolate beyond the data—the linear model may no longer hold outside of the range of the data.**
  - DO NOT infer causality; that the predictor variable **causes** the response variable to change simply because there is a good linear model for the relationship.
  - **Association is not causation.**
- R^2 is a valuable indicator, and must be included in any regression, but look at the entirety of the data and model. Do not judge a model based on R^2 alone.

Summary

- A **regression line** is a straight line that describes how a response variable y changes as an explanatory variable x changes.

    - You can use a regression line to predict the value of $\hat{y}$ for any value of x by substituting the x into the equation of the line.

    - The **slope b_1** of a regression line $\hat{y} = b_0 + b_1x$ is the rate at which the predicted response y changes along the line as the explanatory variable x changes.

    - The **y-intercept b_0** of a regression line $\hat{y} = b_0 + b_1x$ is the predicted response y when the explanatory variable x = 0. This prediction is of no statistical use unless x can actually take values near 0.

Objective: Students will determine the line of best fit (Least Squares Regression Line) and describe the meaning.

Summary

- Avoid extrapolation.
  - Do not use the model for values of the predictor variable outside the range of the data from which the line was calculated.
- The most common method of fitting a line to a scatterplot is least squares. The **least-squares regression line** is the straight line $y = b_0 + b_1x$ that minimizes the sum of the squares of the vertical distances of the observed points from the line.
- The least-squares regression line of y on x is the line with slope $b_1 = r(s_y/s_x)$ and intercept $b_0 = \bar{y} - b_1\bar{x}$. This line always passes through the point $(\bar{x}, \bar{y})$.
- Most of all, be careful not to conclude that there is a cause-and-effect relationship between two variables just because they are strongly associated.