Applied Statistics: Linear Regression

4/13/2022

Plan for Today

- What’s Upcoming
- Graph Discussion: Regression Line
- Residuals
- Linear regression activity w/ data

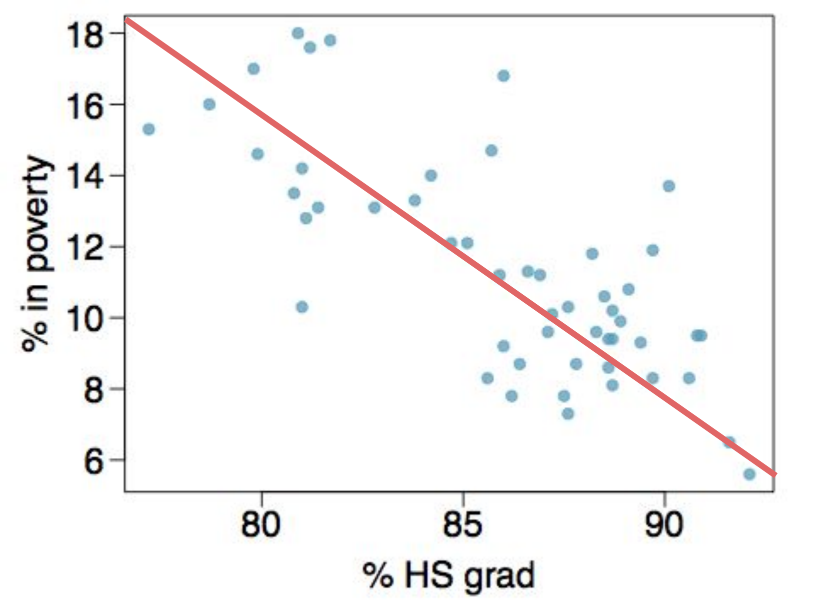
What’s upcoming

- Reading Assignment 4 Writeup – due Wednesday, April 13.
- Daily Survey 32 – Due Thursday.
- No class on Monday!
- Week 12 Homework – Due next **Tuesday**, April 19.
- Project 2 – Due next Wednesday, April 20.

Linear Regression: Information We Can Find

- **Scatterplot:** Visual representation of relationship
- **Correlation coefficient *R*:** Measure of how linearly related the two variables are and whether the relationship is positive/negative
- **Equation of regression line:** Of the form $y = mx + b$, describes relationship and lets us make predictions
- **Coefficient of determination $\boldsymbol{R^2}$**: we'll get to this today!

Poverty & HS Graduation Rate



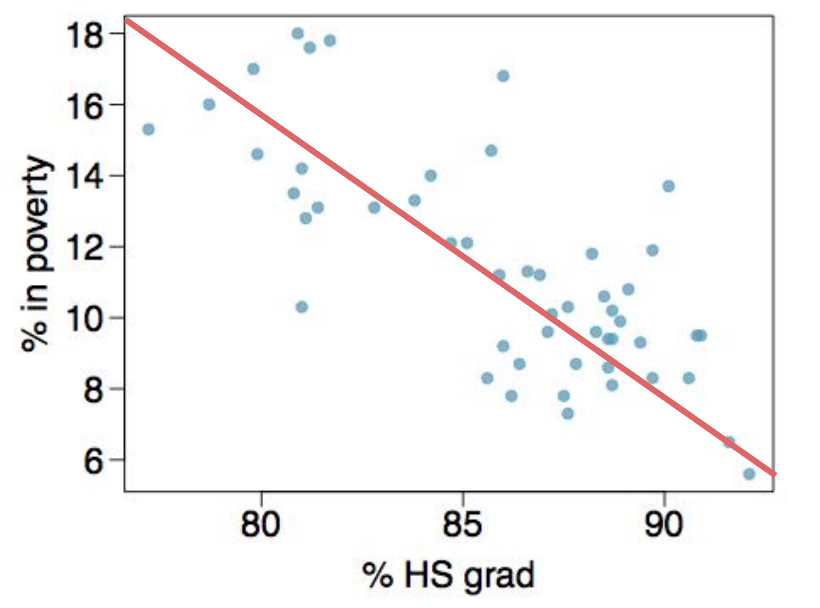
$$\widehat{y} = -0.62x + 64.68$$

Slope of least-squares line:

y-intercept:

Prediction of % in poverty when HS graduation is 85%?

Poverty & HS Graduation Rate



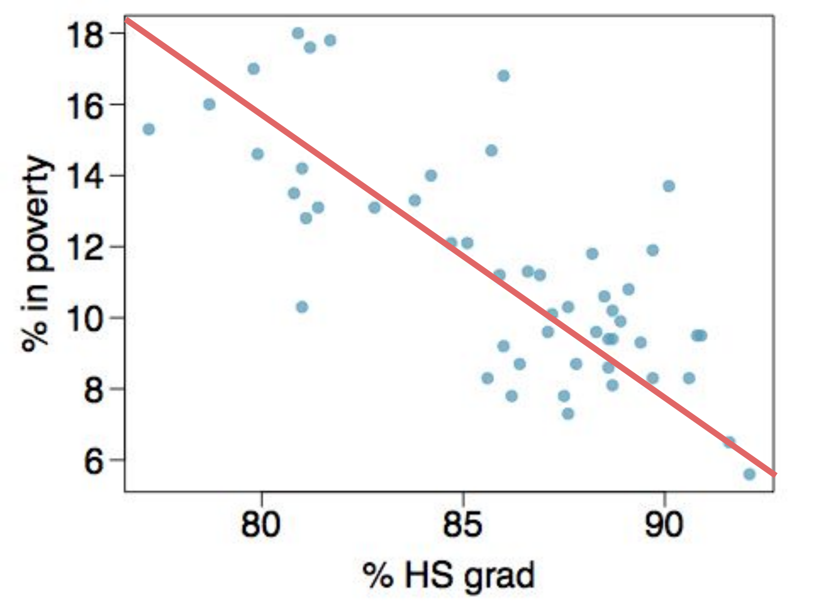
$$\widehat{y} = -0.62x + 64.68$$

Slope of least-squares line: -0.62 % in poverty per % HS grad

y-intercept:

Prediction of % in poverty when HS graduation is 85%?

Poverty & HS Graduation Rate



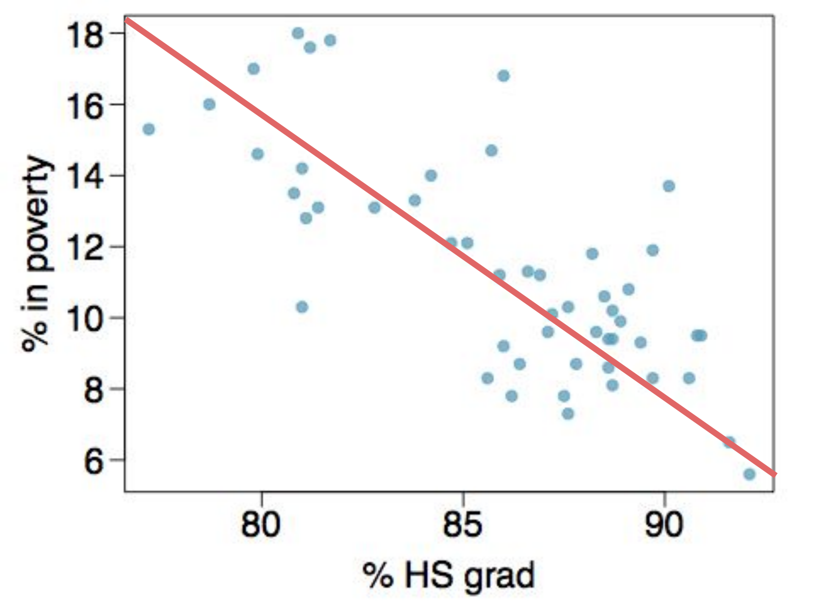
$$\widehat{y} = -0.62x + 64.68$$

Slope of least-squares line: -0.62 % in poverty per % HS grad

y-intercept: 64.68% in poverty with 0% HS grad

Prediction of % in poverty when HS graduation is 85%?

Poverty & HS Graduation Rate



$$\widehat{y} = -0.62x + 64.68$$

Slope of least-squares line: -0.62 % in poverty per % HS grad

y-intercept: 64.68% in poverty with 0% HS grad

Prediction of % in poverty when HS graduation is 85%?

-0.62(85) + 64.68 = 11.98 % in poverty

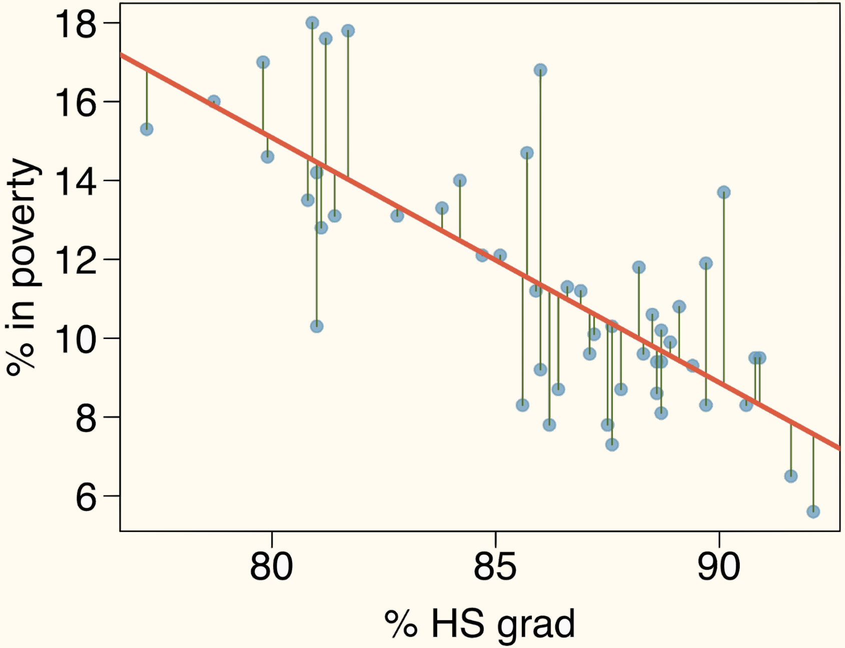
Reminder: Correlation ≠ Causation

Correlation lets us describe a relationship and make predictions but not say that one thing causes another.

Residuals

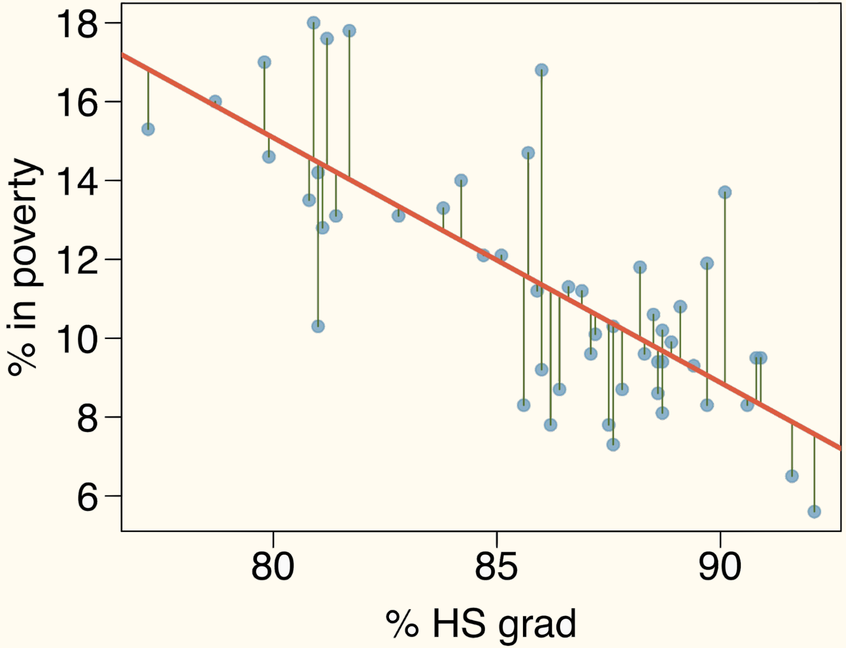
Residuals are the leftovers from the model fit:

Data = Fit + Residual



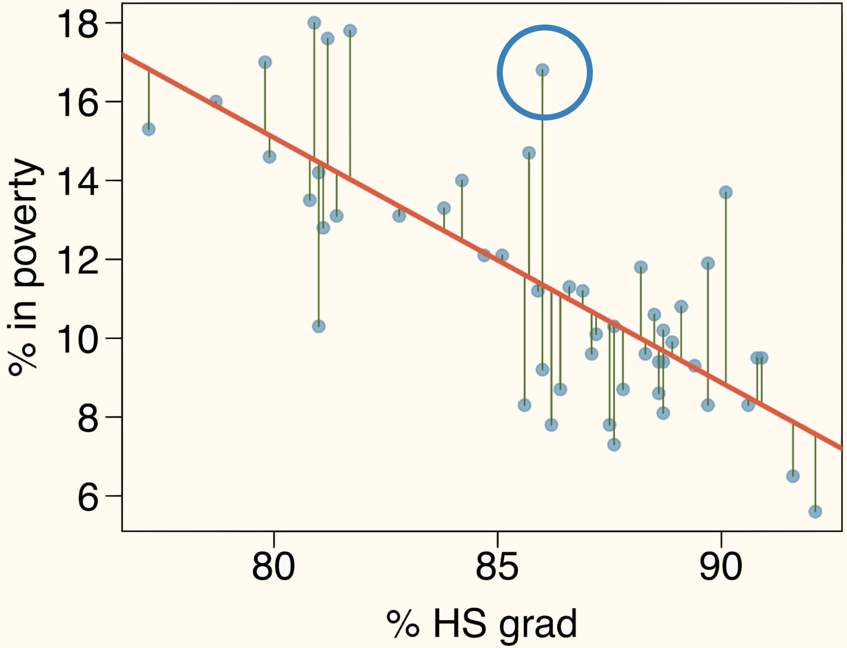
Residuals

In other words, they are the difference between what is observed and what is predicted.



Residuals

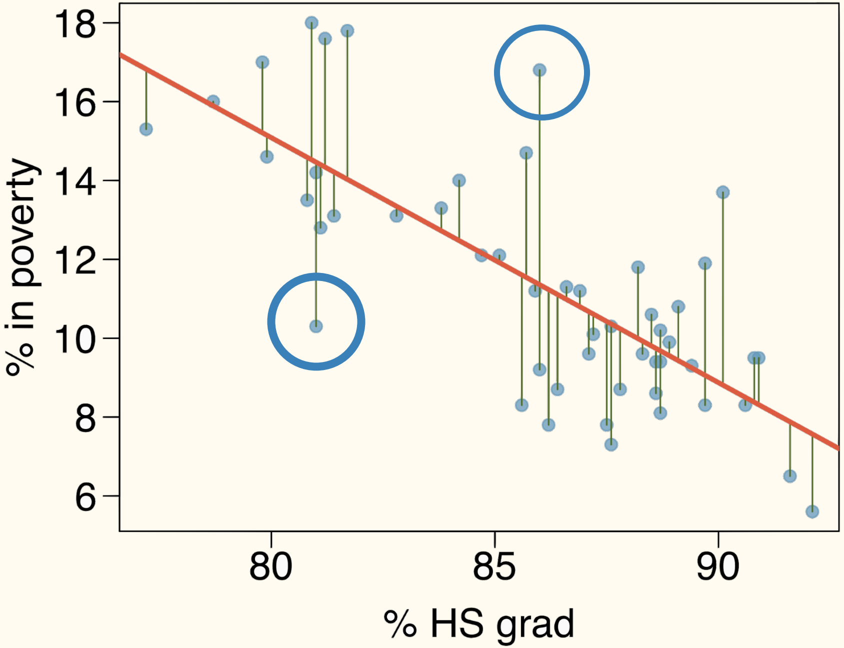
In other words, they are the difference between what is observed and what is predicted.



% living in poverty in DC is 5.44% more than predicted.

Residuals

In other words, they are the difference between what is observed and what is predicted.



% living in poverty in DC is 5.44% more than predicted.

% living in poverty in RI is 4.16% less than predicted.

Deciding if we should use linear regression

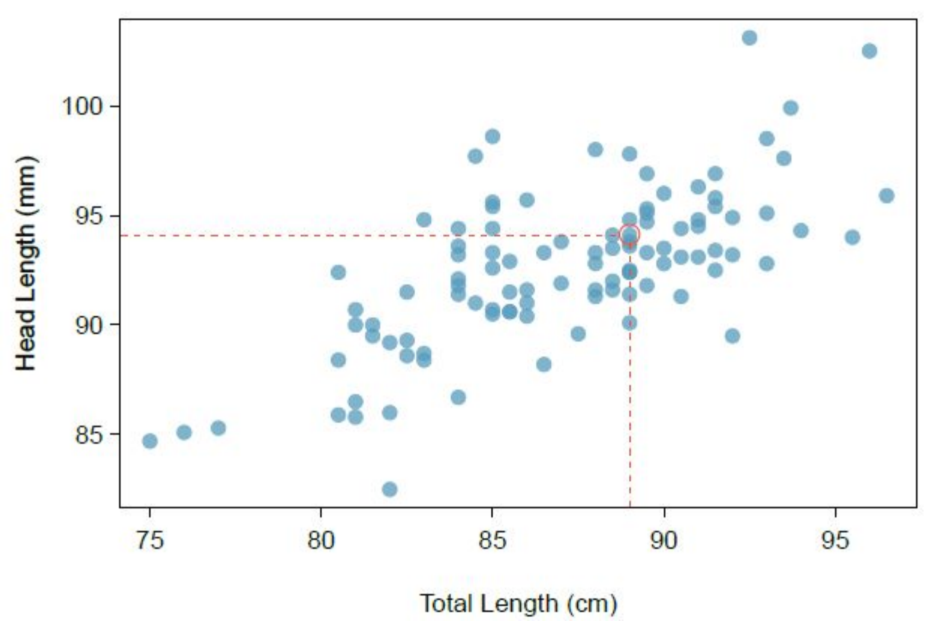
- Does a linear relationship look reasonable?

- Is there a pattern in the residuals?
    - Think of tipping the graph over until the best fit line is horizontal

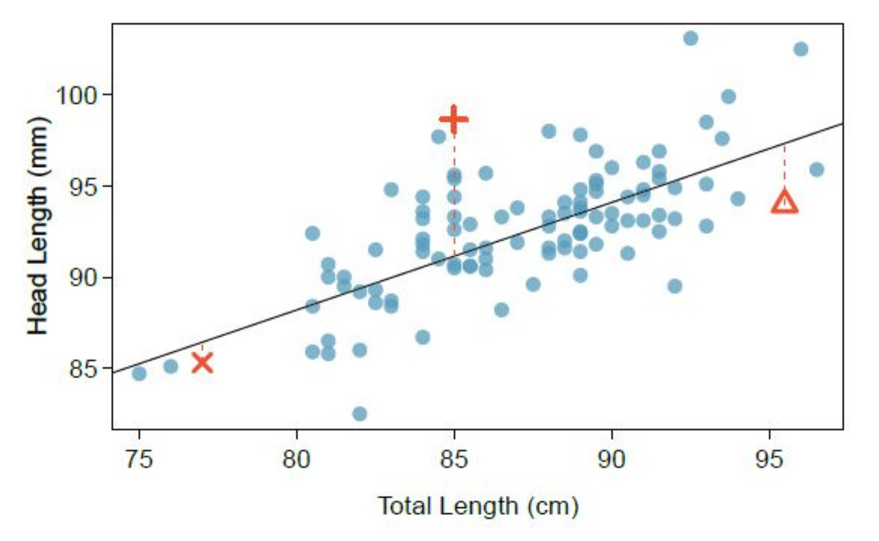
- Are the residuals approximately normal, without outliers?

- Are the data points independent?

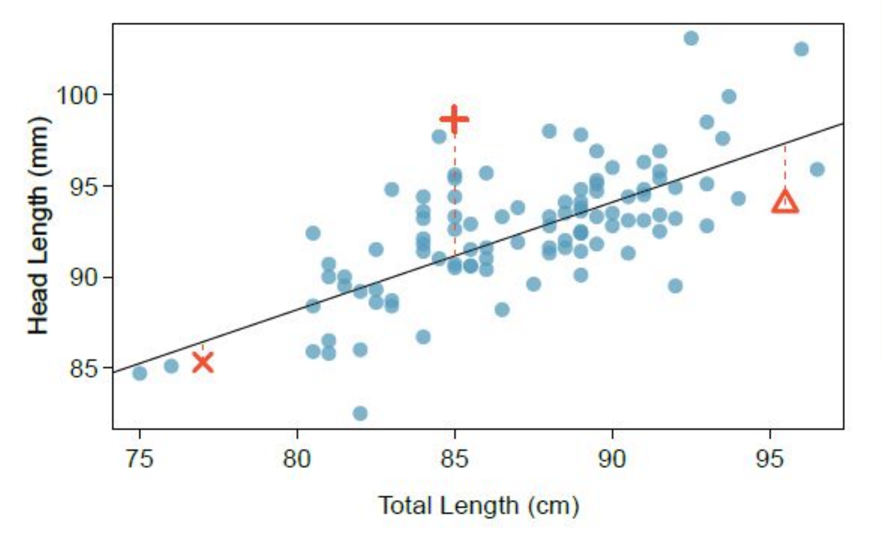
Possum Head Length vs. Total Length

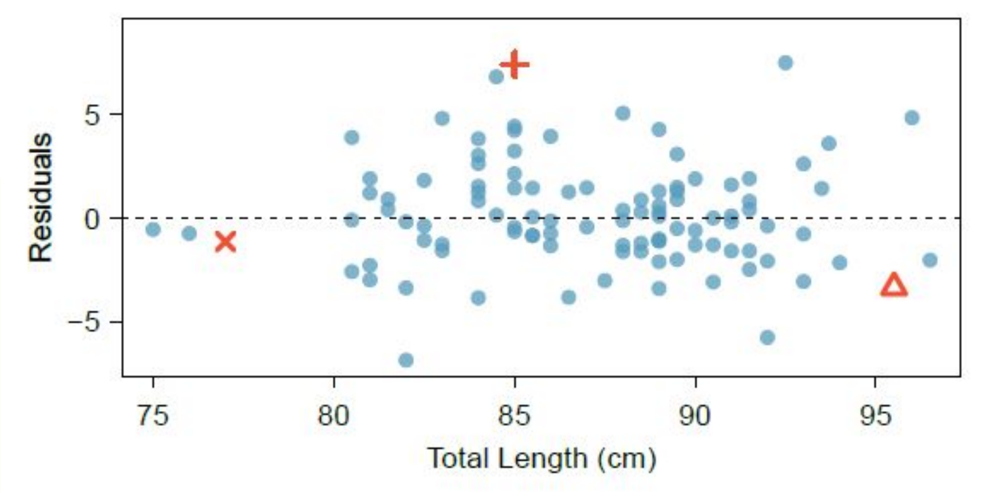


Possum Head Length vs. Total Length

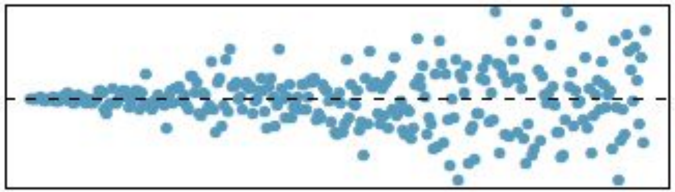


Possum Head Length vs. Total Length

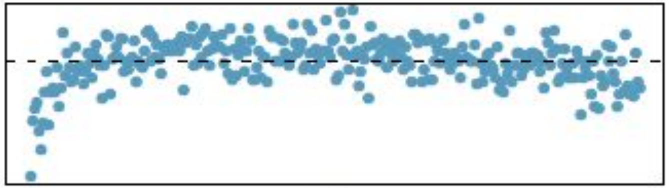




Do these residuals have patterns?



(a)



(b)

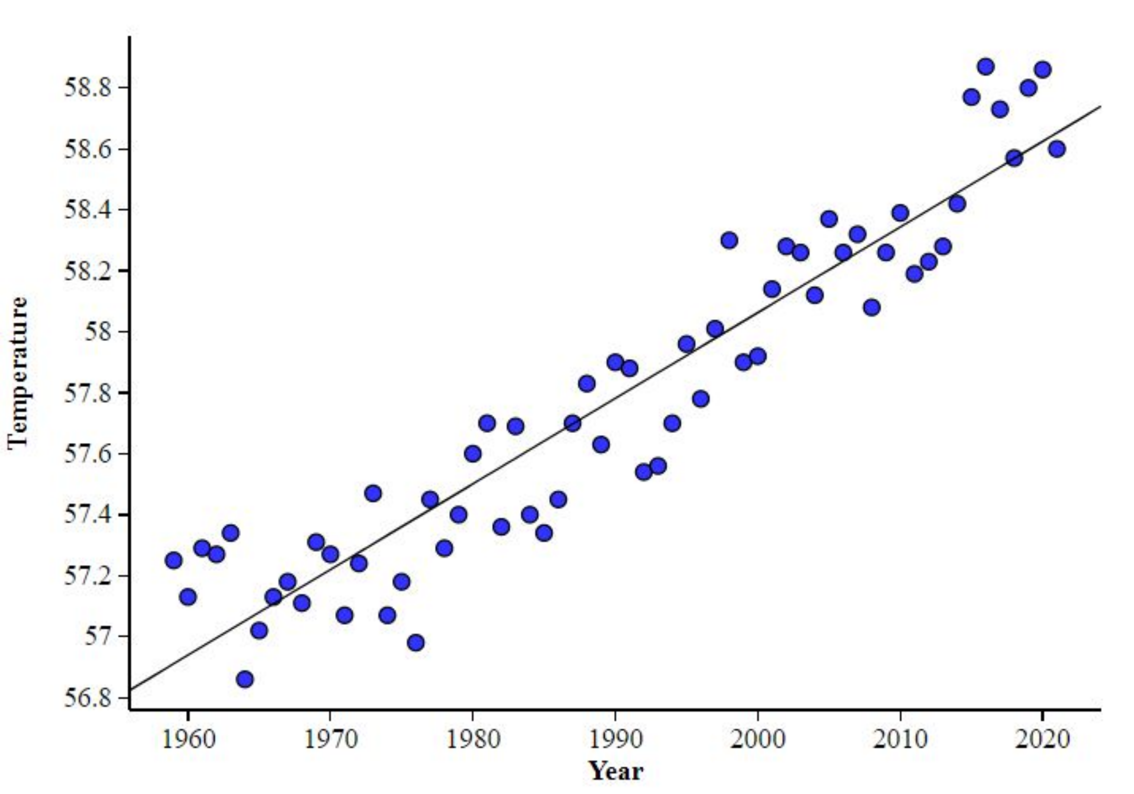
Activity, Part 1

You'll be looking at how global average temperature has changed over time as an example of finding scatterplots, regression lines, and R values for real data.

Do this individually if possible, but please feel free to ask each other for help as you go!

If online: You'll need to make your own copy of the questions as well as the data.

Temperature vs. Time

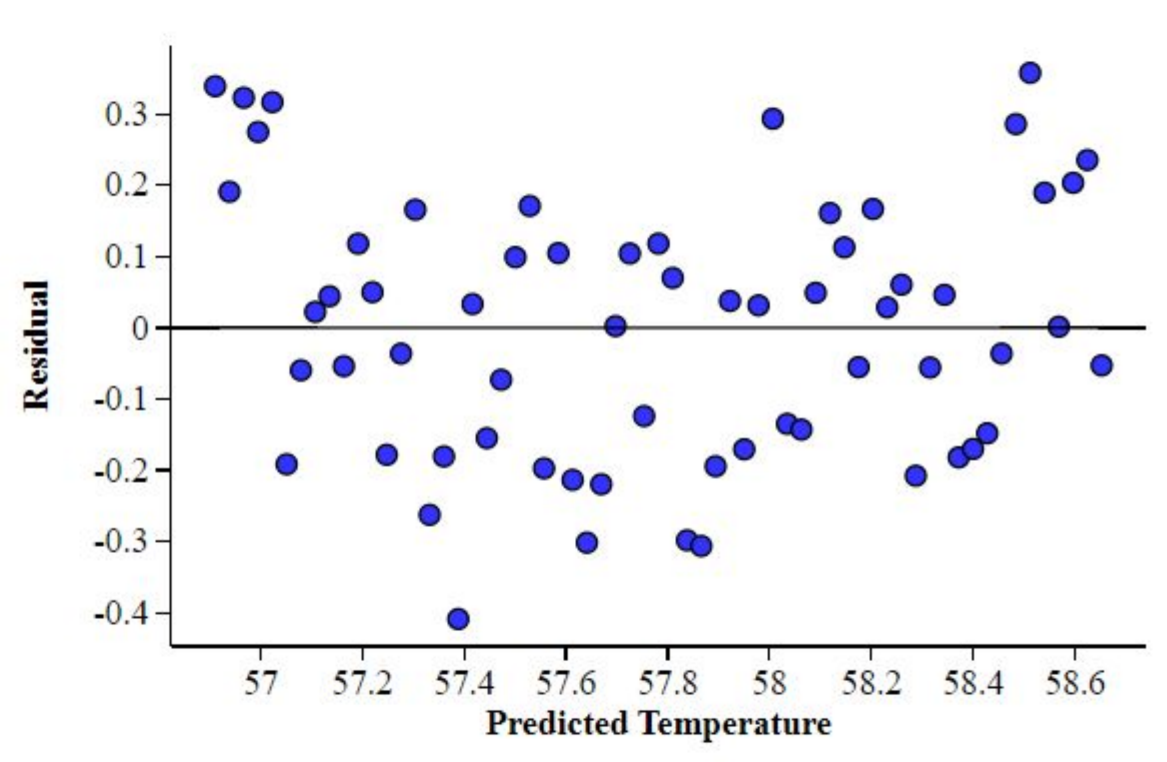


Regression line:
$\hat{y} = 1.88 + 0.028x$

R^2 value: 0.887

R value: 0.942

Temperature vs. Time: Residual Plot



R^2: Coefficient of Determination

In the case of linear regression, R^2 is the square of the correlation coefficient R.

But it can be used more generally, too! The closer it is to 1, the better the explanatory variable(s) predict the response variable in our model.

The most official/correct description is that R^2 tells us what percent of the variation in the response variable is explained by variation in the explanatory variable.

Project 3

- Use either a provided dataset or find/collect your own data and explore potential linear relationships between variables.
- Will be posted on Blackboard tomorrow
- Due Wednesday, May 4 (last day of our class, before finals period)

Friday

- More practice with regression and correlation.