Linear Regression
Part 2: Residuals and Errors

INFO-1301, Quantitative Reasoning 1

University of Colorado Boulder

April 21, 2017

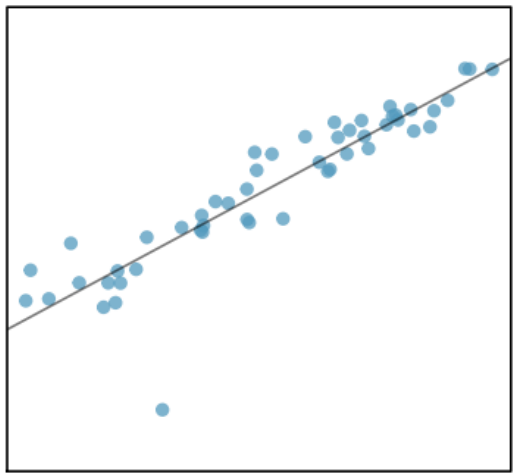
Prof. Michael Paul

Fitting Linear Functions

Where does a linear function such as
“y = 9.607x + 111.958” come from?

Want to pick slope and y-intercept (y=*mx*+*b*) such that the line is as close as possible to the true data points

- Want to minimize distance from each point to the line
- We’ll be more concrete today



Fitting Linear Functions

The process of picking the parameters of a function (e.g., *m* and *b*) to make it is close as possible to a set of data points is **regression**

If the function is linear (i.e., a line) then this is **linear regression**

Statistical software such as MiniTab Express can perform linear regression automatically

Residuals

The **residual** of a point (x_i, y_i) is the difference between the true y_i value and the value you estimated based on your best-fit line:

$$e_i = y_i - (mx_i + b)$$

Also referred to as the **error** of your line at that point

The *size* of a residual is its absolute value: $|e_i|$

Residuals

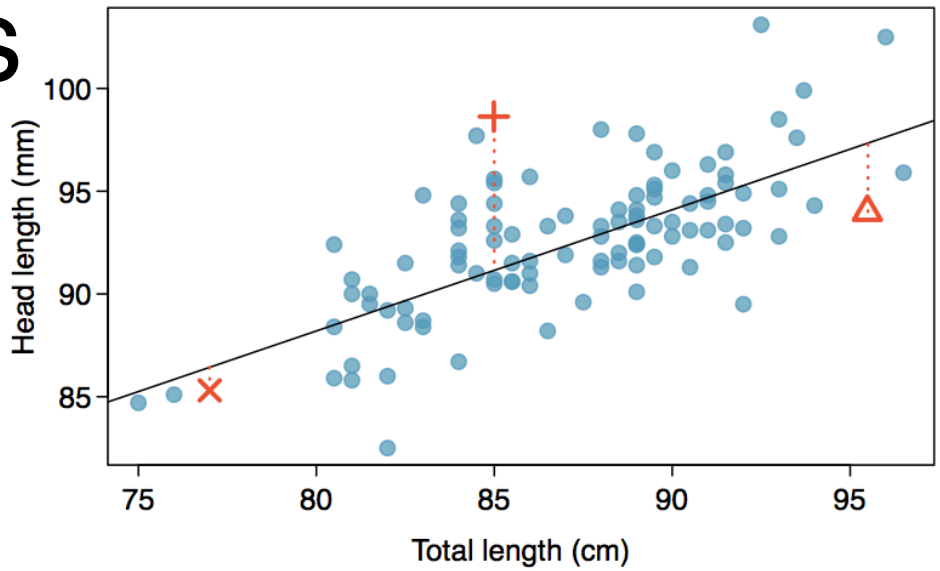
The average residual size can tell you the average error you will make if you estimate new data points (e.g., interpolation or extrapolation)

This is only true if the new data points follow the same pattern as the data you originally observed

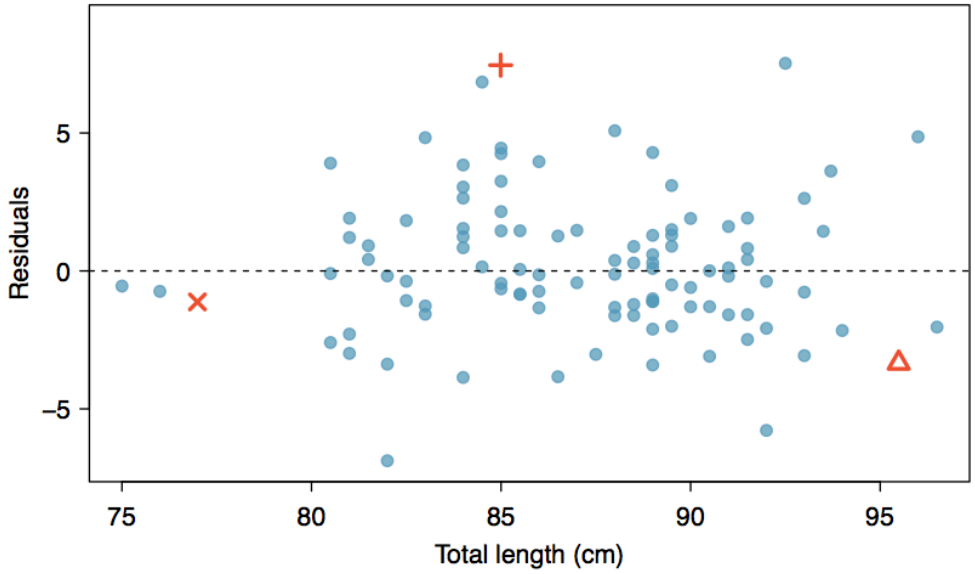
- More likely to be true for interpolation than extrapolation

Residual Plots

Original data:

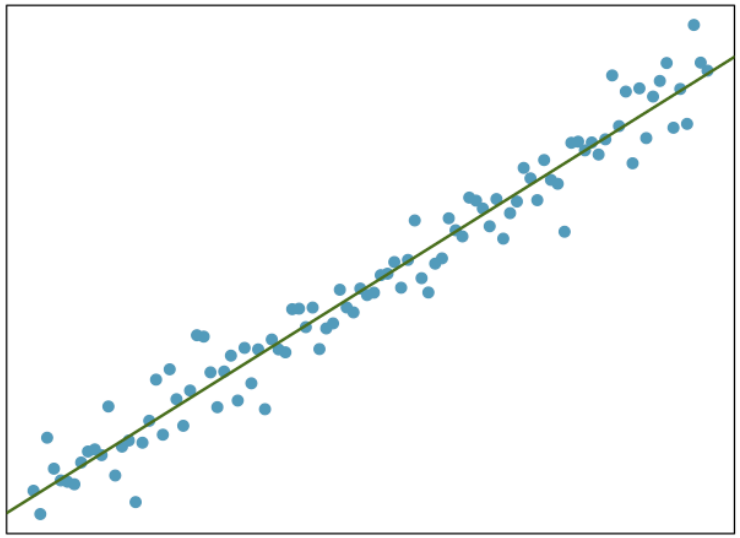


Residuals:

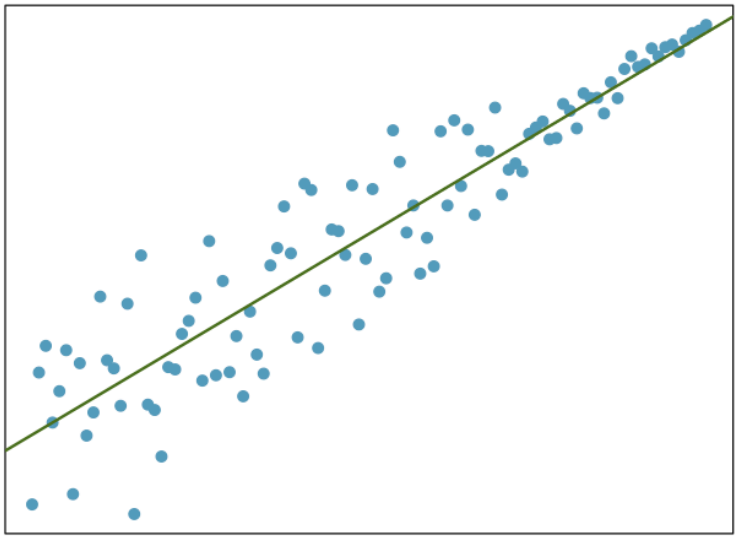


Practice

7.1 Visualize the residuals. The scatterplots shown below each have a superimposed regression line. If we were to construct a residual plot (residuals versus x) for each, describe what those plots would look like.



(a)



(b)

Practice

(7.21) Suppose we fit a regression line to predict the shelf life of an apple based on its weight. For a particular apple, we predict the shelf life to be 4.6 days. The apple's residual is -0.6 days. Did we over or under estimate the shelf-life of the apple? Explain your reasoning.

Fitting Linear Functions

How to choose *m* and *b*? Pick them so that the residuals are as small as possible.

- Could minimize the absolute value of the residuals
- More common: minimize the square of the residuals
  - “least squares regression”
  - This will favor solutions where no residual is especially large (outliers penalized more)

Root Mean Squared Error

$MSE = (\Sigma_i e_i^2)/n$ n is the number of points

$RMSE = \sqrt{(MSE)}$

Example: residuals are 1, 2, -1, 2, -2

- $RMSE = [[(1)^2+(2)^2+(-1)^2+(2)^2+(-2)^2]/5]^{.5}$

  $= [[1+4+1+4+4]/5]^{.5}$

  $= [14/5]^{.5}$

  $= 1.67$

Root Mean Squared Error

- Generally:
  - 68% of residuals are within 1 RSME of line.
  - 95% of residuals are within 2 RSME of line.
- Look familiar?
- In “nice” cases, residuals form a normal distribution around the least square line
  - The mean residual is 0.
  - The standard deviation is the RMSE.
- Can use Z-table to estimate probability of having an errors of a certain size.

Conditions for Least Squares Regression

- Eyeball that the data is linear (fits a line)
- Random residuals not too far from line (outliers can be a problem)
- The size of the residuals roughly constant (not those that get larger at one end)
- No repeating patterns in the data (e.g. time series)

Conditions for Least Squares Regression

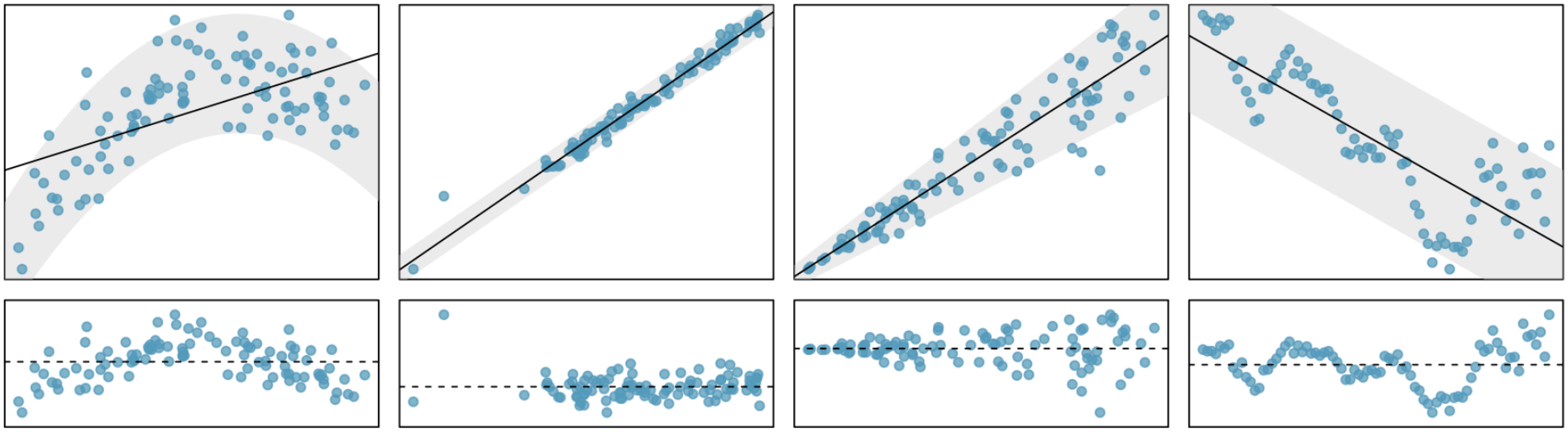
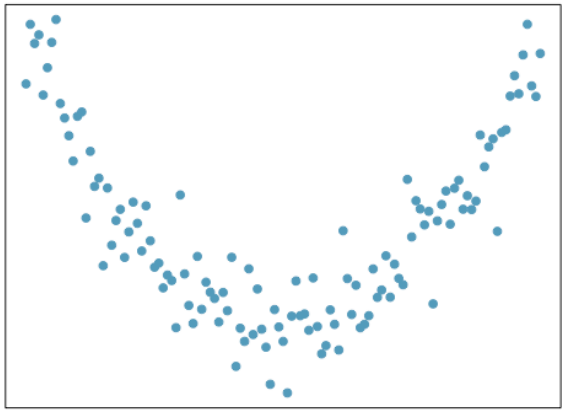


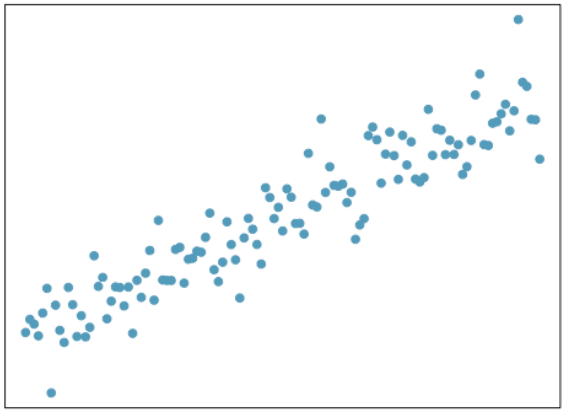
Figure 7.13, page 342

Practice

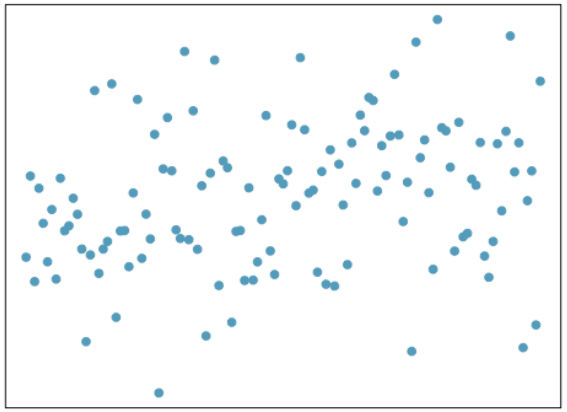
7.3 Identify relationships, Part I. For each of the six plots, identify the strength of the relationship (e.g. weak, moderate, or strong) in the data and whether fitting a linear model would be reasonable.



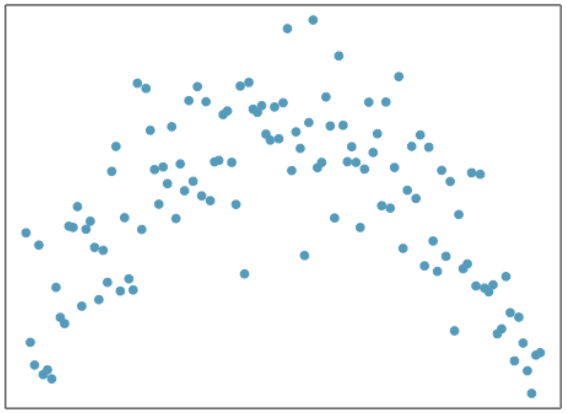
(a)



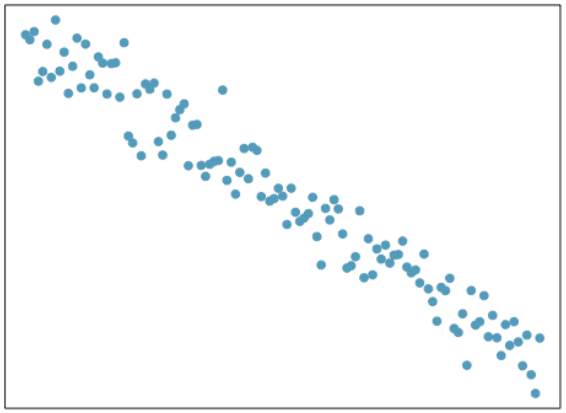
(b)



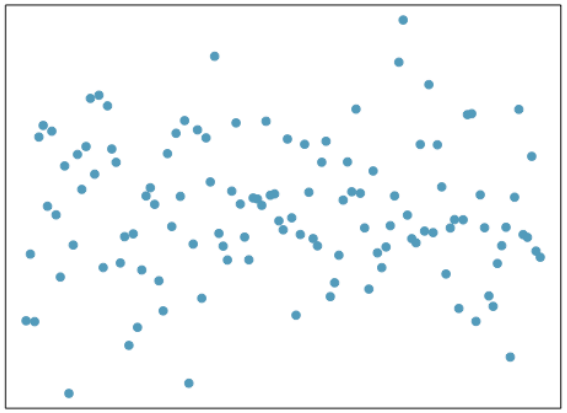
(c)



(d)



(e)



(f)

Practice

7.29 Murders and poverty, Part I. The following regression output is for predicting annual murders per million from percentage living in poverty in a random sample of 20 metropolitan areas.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-29.901	7.789	-3.839	0.001
poverty%	2.559	0.390	6.562	0.000

$s = 5.512$ $R^2 = 70.52\%$ $R^2_{adj} = 68.89\%$

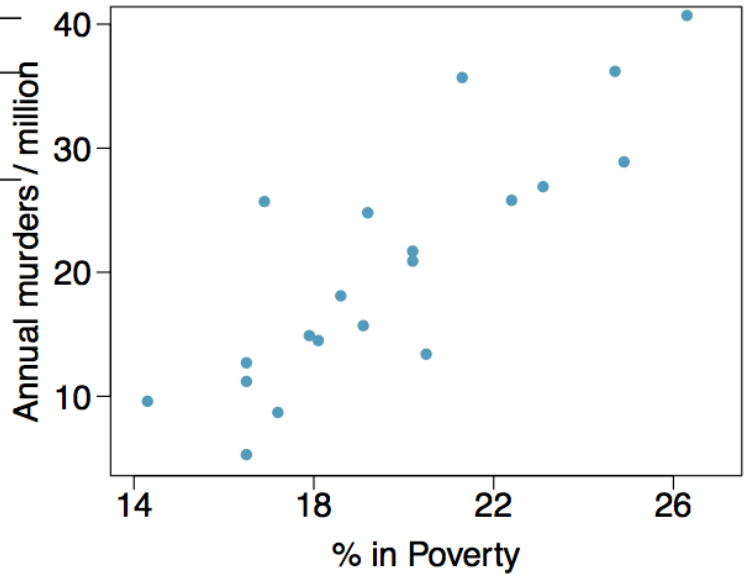
(a) Write out the linear model.

(b) Interpret the intercept.

(c) Interpret the slope.

(d) Interpret R^2.

(e) Calculate the correlation coefficient.



More Practice

7.39 Urban homeowners, Part II.
Exercise 7.33 gives a scatterplot displaying the relationship between the percent of families that own their home and the percent of the population living in urban areas. Below is a similar scatterplot, excluding District of Columbia, as well as the residuals plot. There were 51 cases.

(a) For these data, $R^2 = 0.28$. What is the correlation? How can you tell if it is positive or negative?

(b) Examine the residual plot. What do you observe? Is a simple least squares fit appropriate for these data?