Topic 7:
Linear Regression

STOR 155: Introduction to Data Models and Inference

Dr. Teressa Bergland

Fall 2025

Announcements

Course PSAs:

- Homework 6 due **TODAY, Thursday 2/5 on WebAssign**
- Homework 7 due **Tuesday 2/10**
- **Midterm Exam 1 on Thursday, February 12**
  - 100 points
  - 70-80% multiple choice or similar short-form (true/false, matching, etc)
  - 20-30% open-ended problems, will require showing work/justification
  - Calculator? **YES!** TI-84 or comparable is OK
  - Please do not bring a calculator that can access the internet
  - Formula sheet? **YES!** Handwritten, single-sided, 8.5x11 or smaller
  - Please put your name on your sheet – it will be collected with the exam

Practice Question #1

i	x_i	y_i	$(x_i - \bar{x})/s_x$	$(y_i - \bar{y})/s_y$	product
1	0	5	–1.1	+1.17	–1.28
2	2	2	***T***	–0.83	+0.25
3	3	2	+0.1	–0.83	–0.08
4	6	4	+1.3	+0.65	+0.65
	$\bar{x} = 2.75$	$\bar{y} = 3.25$			Sum = –0.46
	$s_x = 2.5$	$s_y = 1.5$			***r*** =

1. What is ***T*** ?
   A +0.3
   B –0.3
   C –0.75
   D 0.75
   E None of the above.

Practice Question #2

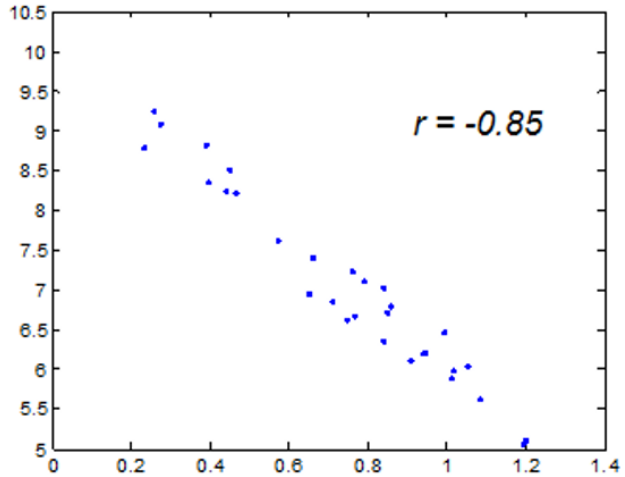
The two scatterplots show the same data, except that the point at *A* has been added to the dataset at the bottom. What can you say about the correlation coefficient of the dataset with *A* added?

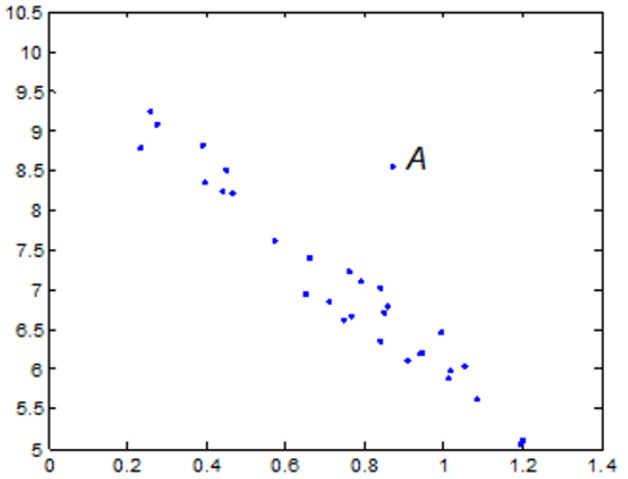
A It is greater than –0.85.

B It is less than –0.85.

C It equals –0.85.

D It is impossible to tell without computing the correlation.





Intro to Linear Regression

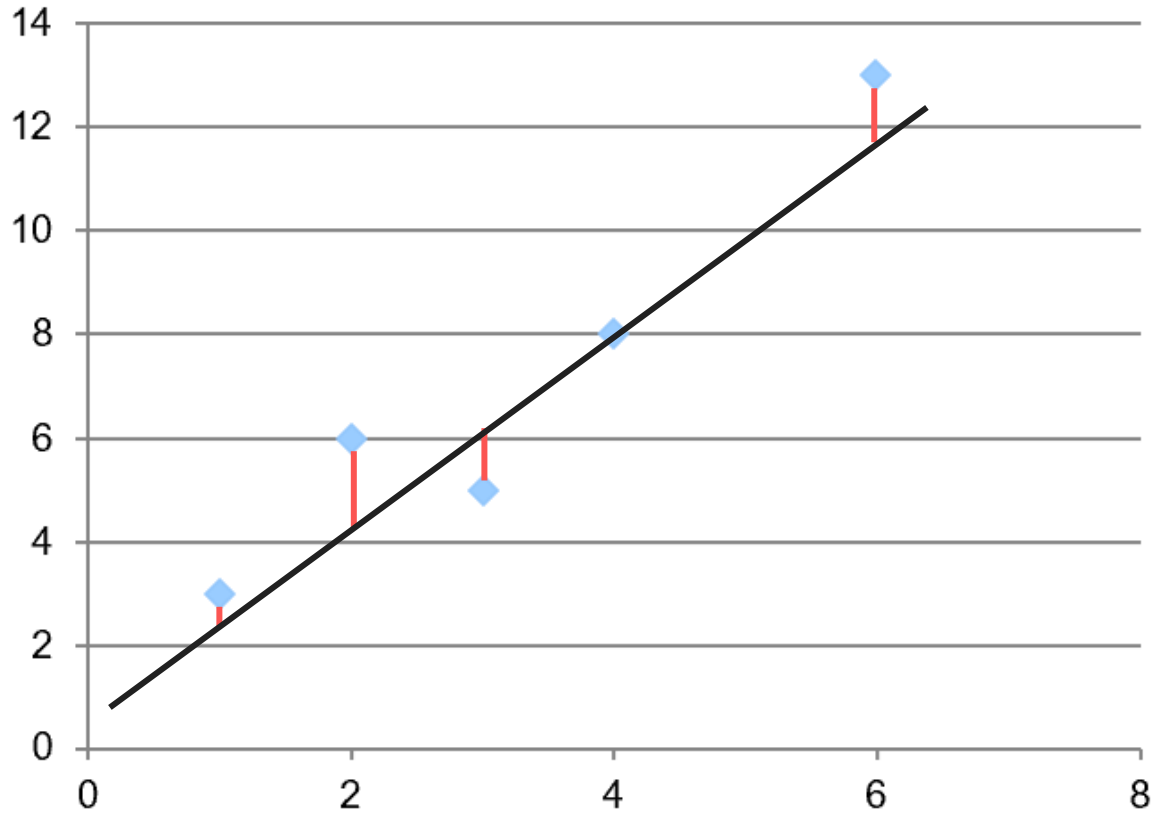
A regression line describes...

Unlike with correlation, it matters...

Standard equation of a line

Small Example

X	Y
1	3
2	6
3	5
4	8
6	13



Residuals and Regression Lines

Remember our notation!

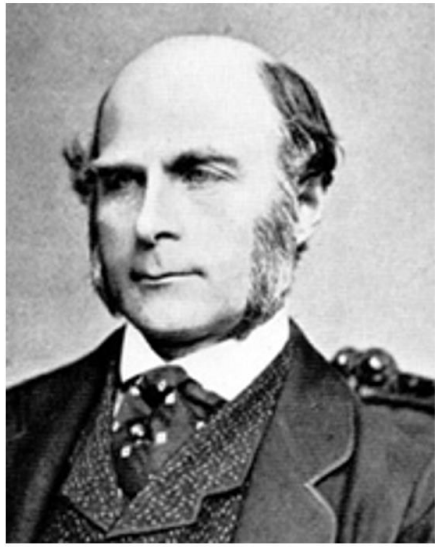
A **residual** is…

Mathematically, our line should…

What values of b_0, b_1 will accomplish this?

Math History Trivia: “Regression”

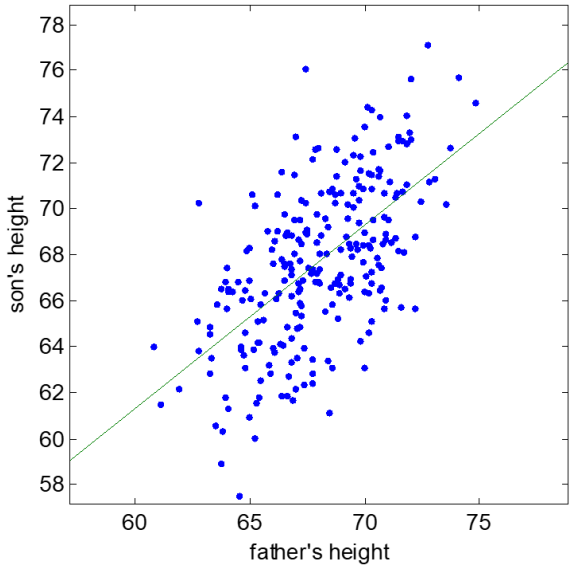
The term was coined by Francis Galton (1822 - 1911) in a study on the heights of fathers and sons. He found that fathers taller than average would more likely have shorter sons, and fathers shorter than average would more likely have taller sons. Galton called this “regression towards the mean height.” (His 1886 paper was titled “Regression towards mediocrity in hereditary stature.”)



Galton’s data looked something like this:

For fathers taller than average, the least-squares line predicts a shorter height for their sons, and vice-versa. That is, the slope b_1 of the line is less than 1.

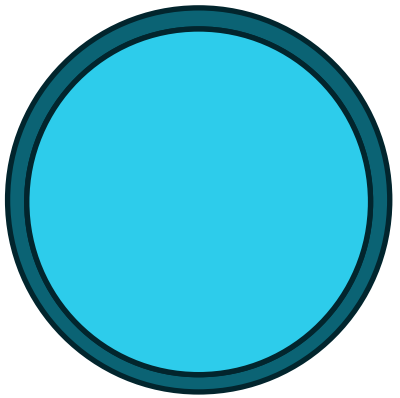
The name “regression line” stuck, even though it’s only a “regression” when the units are the same on both axes and slope is less than 1.



Let's Practice! Constructing and Using Lines

Suppose we're trying to use consumer ratings of various products (X) to predict consumer opinions on the corresponding companies (Y).

$$\bar{x} = 7, \bar{y} = 5$$
$$s_x = 4, s_y = 2$$
$$r = 0.85$$

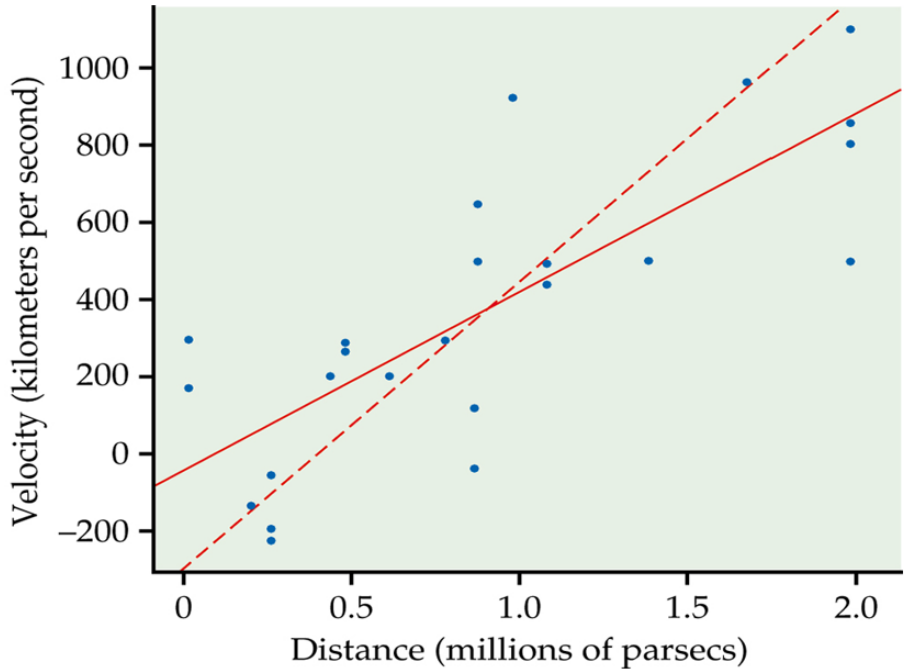


Compute a regression line for predicting Y from X.

What is the residual for the point (6,6)? How does the line's estimate for consumer opinion (Y) compare to the true value for this company?

Switching X and Y

Will switching the explanatory and response variables change the regression line?



Solid line: prediction of velocity for a given distance.

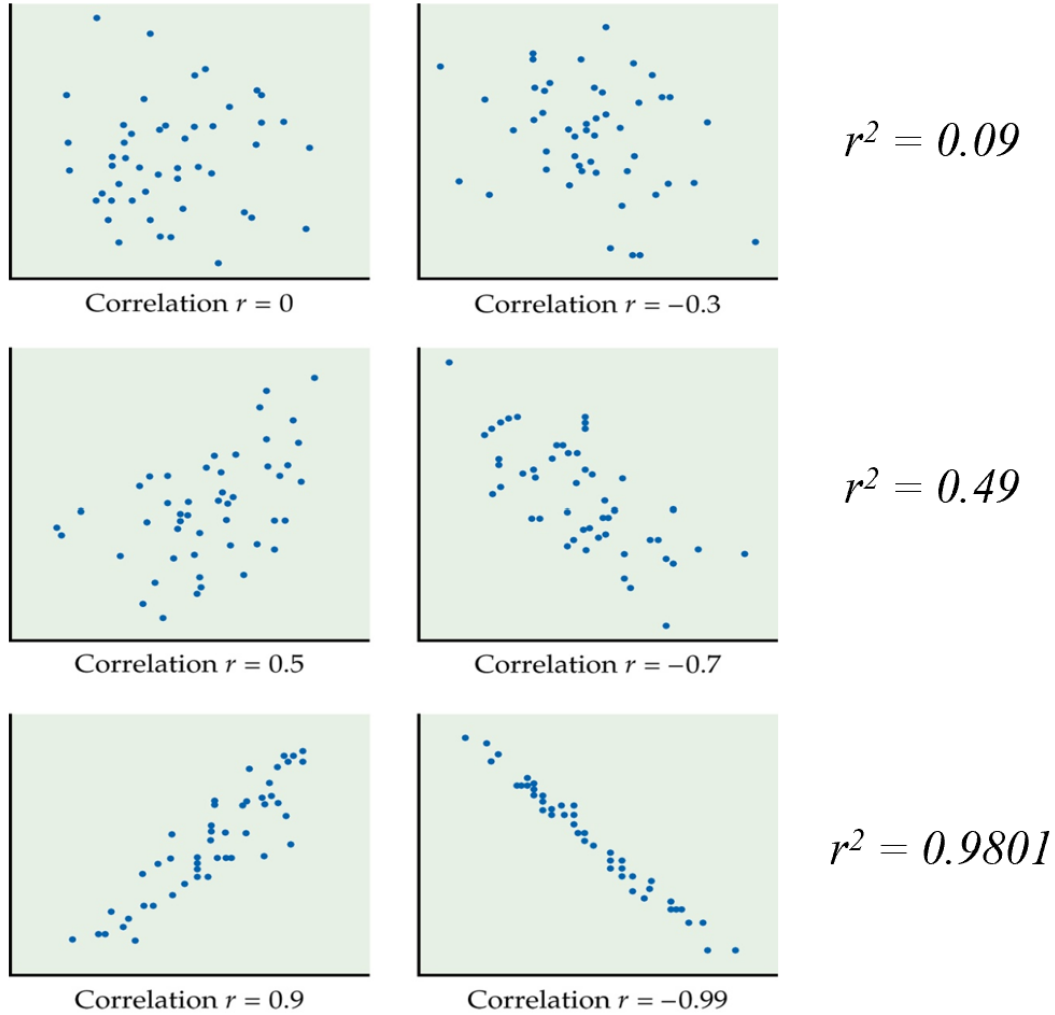
Dashed line: prediction of distance for a given velocity.

Correlation and Regression

The value r^2

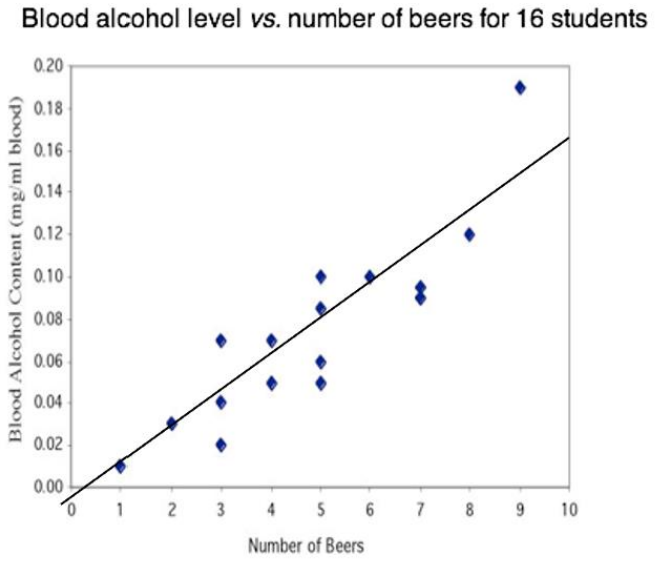
A rule of thumb

$r = 0.5$ vs. $r = -0.7$



Prediction Limitations

Student	Beers	BAC
1	5	0.1
2	2	0.03
3	9	0.19
6	7	0.095
7	3	0.07
9	3	0.02
11	4	0.07
13	5	0.085
4	8	0.12
5	3	0.04
8	5	0.06
10	5	0.05
12	6	0.1
14	7	0.09
15	1	0.01
16	4	0.05



Regression equation:

$\hat{y} = -0.0127 + 0.018x$

Extrapolation

vs.

Interpolation

Let's Practice! Sodium and Blood Pressure

Researchers measured the two variables below for a large number of subjects:

X = sodium intake in mg per day

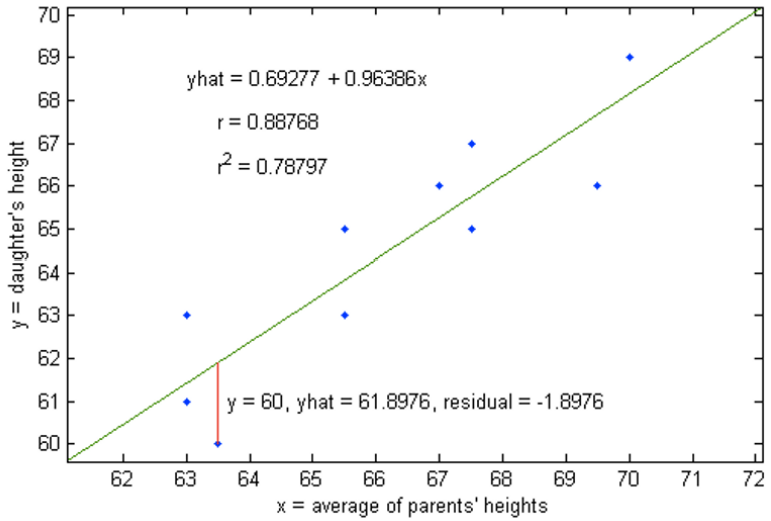
Y = systolic blood pressure

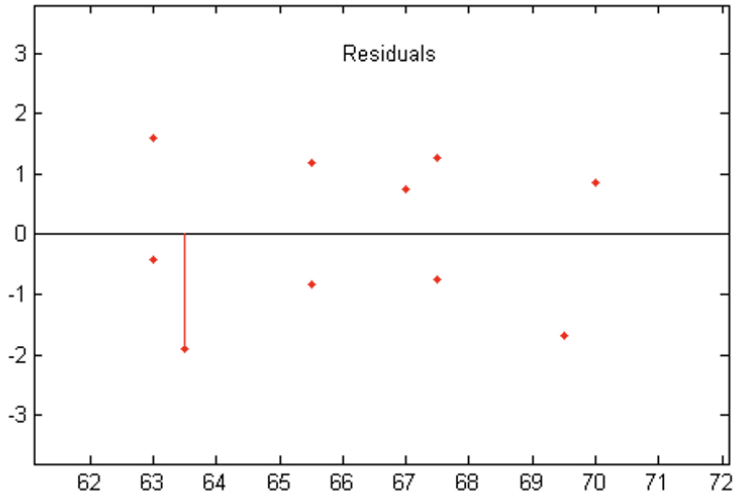
The least-squares regression line for predicting Y given X is:

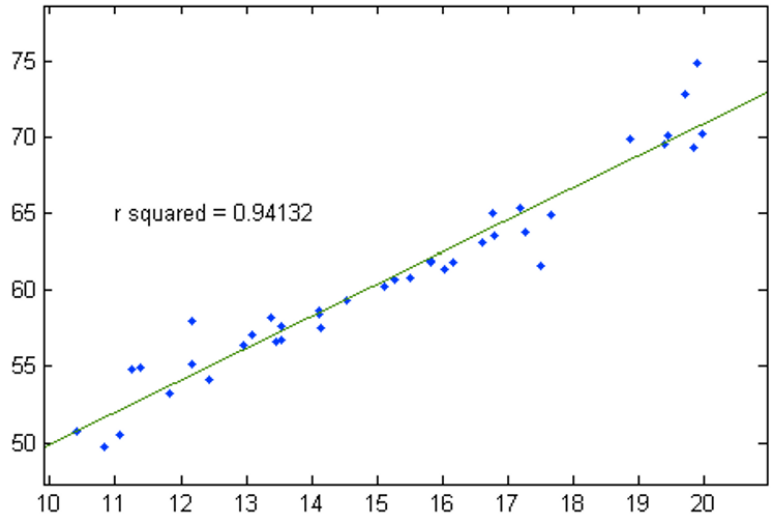
$$\hat{y} = -15.4 + 2.3x$$

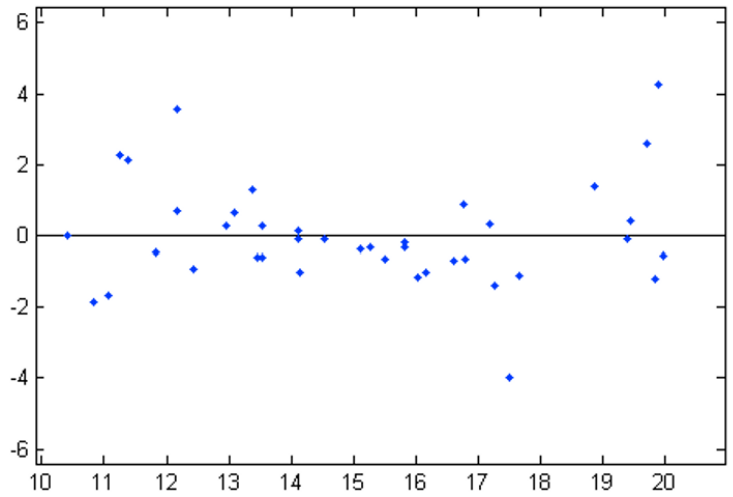
What is the predicted systolic blood pressure for someone whose sodium intake is 60 mg per day? **(Interpretation practice)**

Residual Plots

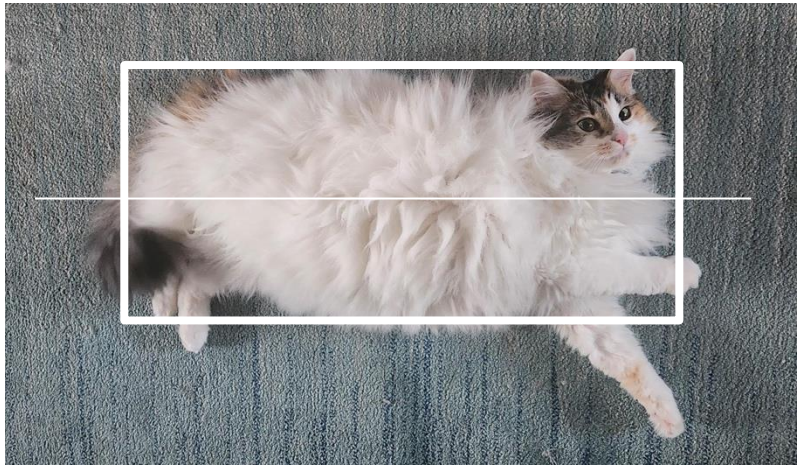




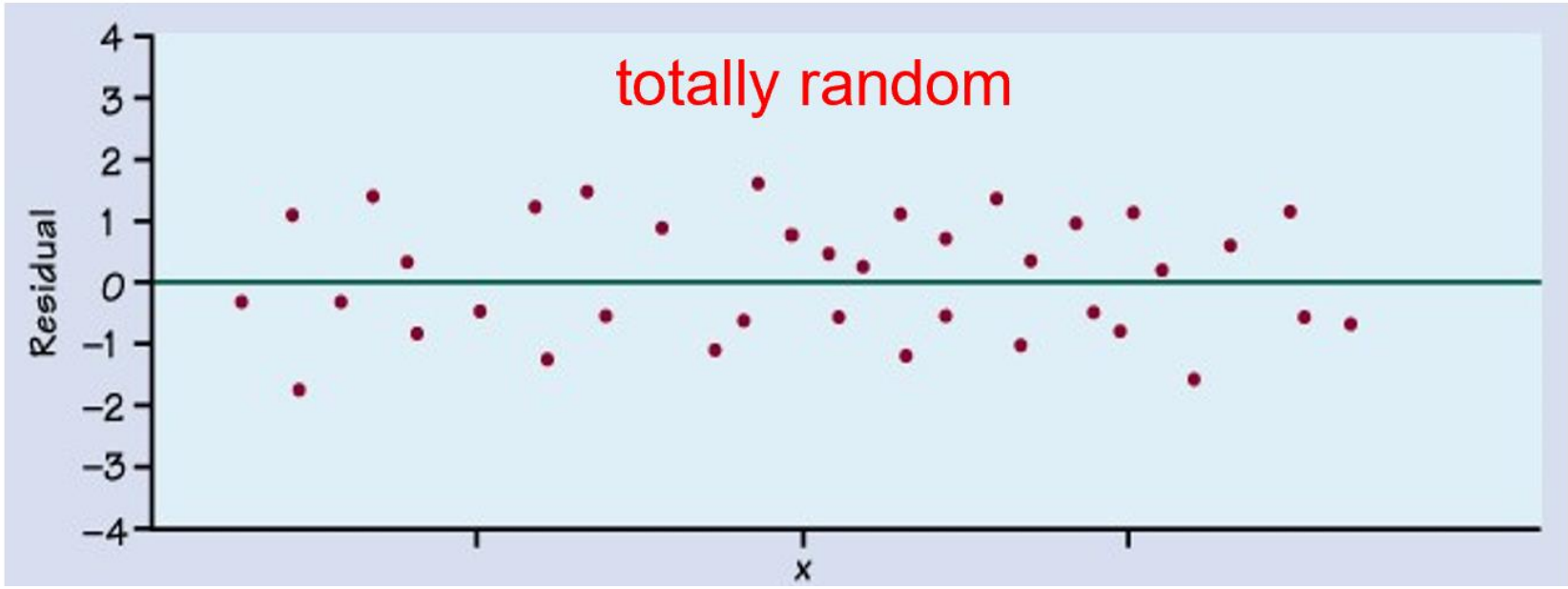




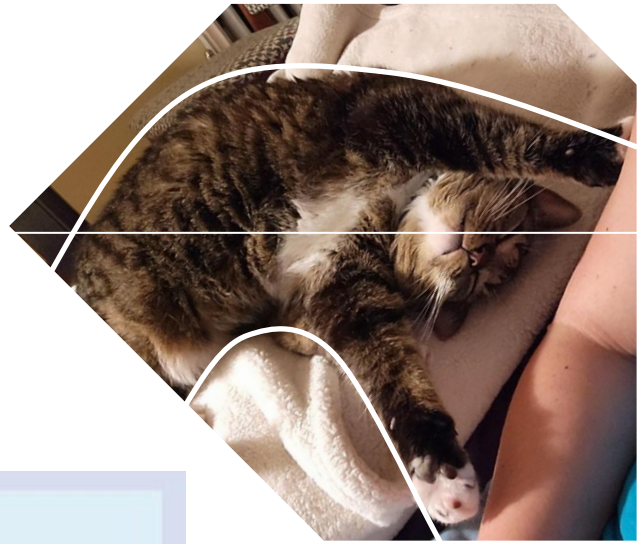
Residual Plots



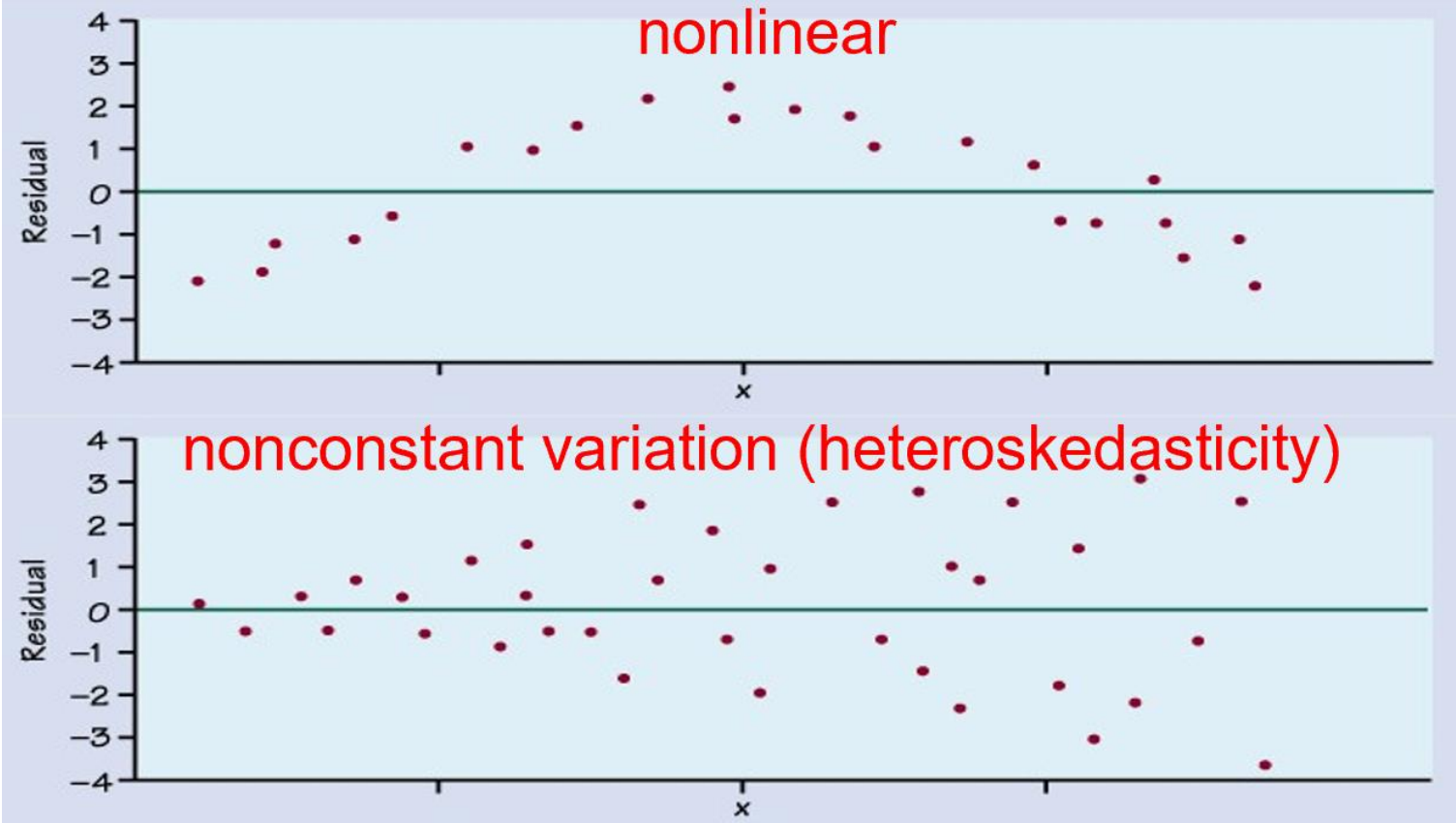
“Ideal” style of residual plot:



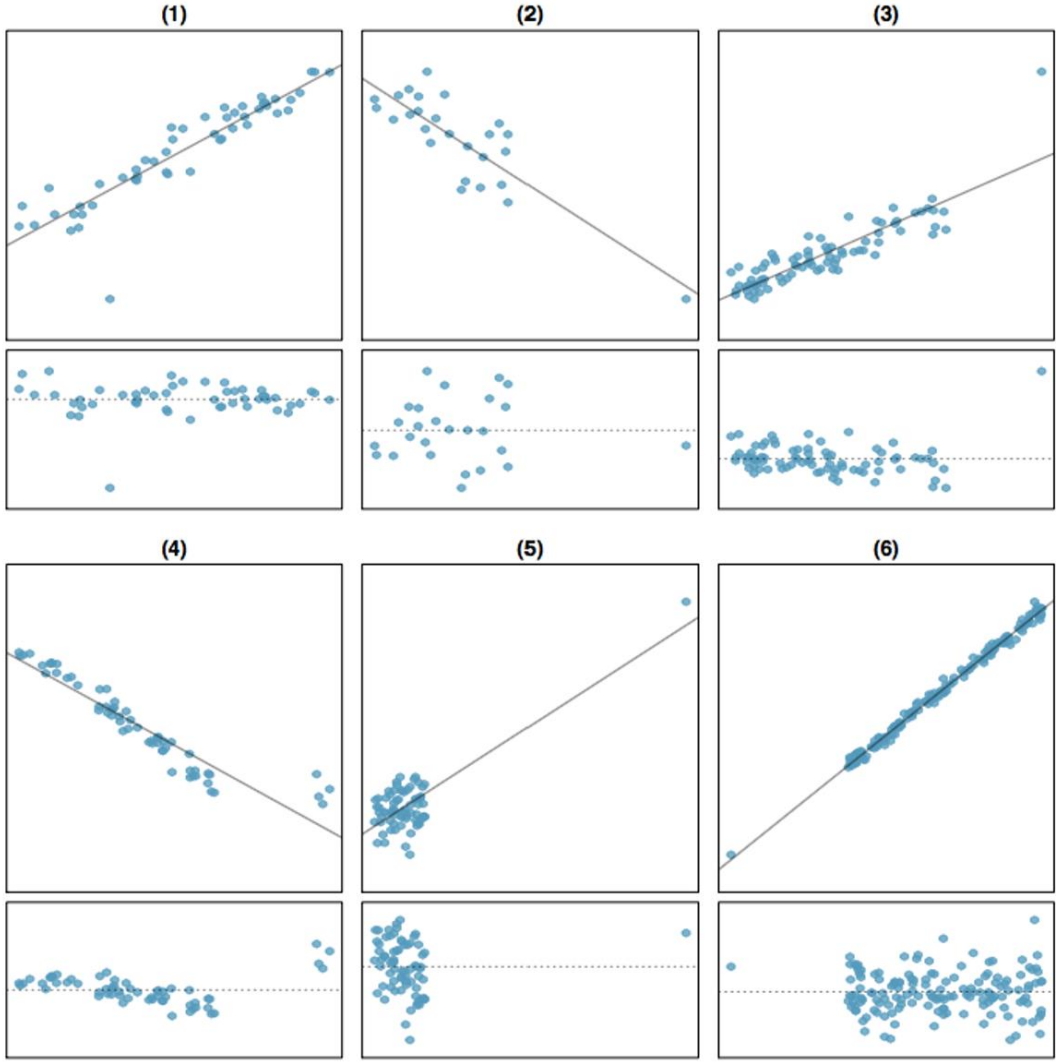
Residual Plots



Not-so-ideal residual plots:



Outliers and Residual Plots



Importance of Plotting Data

All have correlation $r = .816$,
regression equation $\hat{y} = 3 + 0.5x$

Data Set A

x	10	8	13	9	11	14	6	4	12	7	5
y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68

Data Set B

x	10	8	13	9	11	14	6	4	12	7	5
y	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74

Data Set C

x	10	8	13	9	11	14	6	4	12	7	5
y	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73

Data Set D

x	8	8	8	8	8	8	8	8	8	8	19
y	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

Source: Frank J. Anscombe, "Graphs in statistical analysis," The American Statistician, 27 (1973), pp. 17–21.