Stat 13, Intro. to Statistical Methods for the Life and Health Sciences.

1. Extra credit points.
2. Calculating correlation.
3. Linear regression.
4. Slope of the regression line.
5. Goodness of fit.

Read ch10.

Remember, **no lecture Fri Mar10.**
Hw4 is due Fri Mar10 at 2pm by email to statgrader or statgrader2.
10.1.8, 10.3.14, 10.3.21, and 10.4.11.
http://www.stat.ucla.edu/~frederic/13/W23 .

Remember, **no lecture Fri Mar10.**

1. Extra-credit points.

Find a website or article with dataset on detailed **spatial-temporal point process** data on each patient in some region with some contagious disease.
If you find one, email me at frederic@stat.UCLA.edu by Mar20.
It must have a specific, distinct location and time for each patient,
not just a total number of patients on each day in each city. Not cholera or cancer.
Not mosquitos or fish. No attempts after Mar20. No partial credit.
The dataset must have ≥ 40 pts., points at ≥ 10 distinct locations and ≥ 10 distinct times, and < 3 points at any given space-time location.

Unique point process dataset, downloadable or in a table in the paper ---------- 5%.
Non-unique point process dataset, downloadable or in a table ..... 2%.
Unique point process figure ..... 3%.
Non-unique point process figure ..... 2%.
INCORRECT ATTEMPT / NOT POINT PROCESS DATA ..... -0.1%.

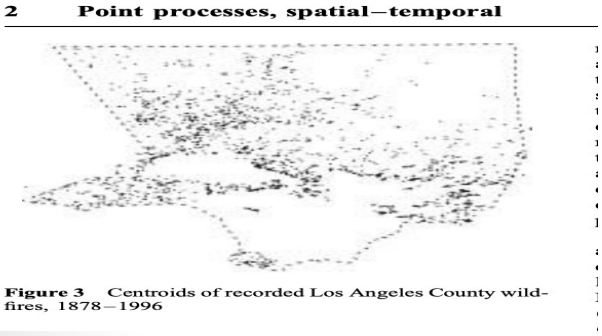


Figure 3 Centroids of recorded Los Angeles County wildfires, 1878–1996

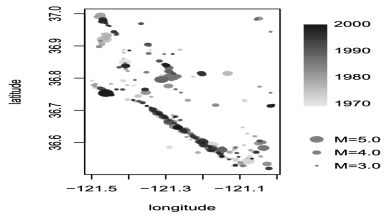


Figure 1: Locations, times and magnitudes of moderate-sized ($M \geq 3.5$) earthquakes in Bear Valley, CA, between 1970 and 2000.

Table 1. Shallow Shocks (M ≥ 6.0) in OFF Tohoku Area for 1885–1980

NO	YEAR	MO	DY	HR	MN	MAG	C	NO	YEAR	MO	DY	HR	MN	MAG	C	NO	YEAR	MO	DY	HR	MN	MAG	C
1	1885	2	9	2	0	6.0	0	84	1908	1	15	21	56	6.9	0	167	1923	5	31	14	55	6.2	1
2	1885	6	11	9	20	6.9	0	85	1908	1	18	1	5	6.0	0	168	1923	6	2	2	24	7.3	0
3	1885	7	29	5	30	6.0	0	86	1908	2	5	21	7	6.0	0	169	1923	6	2	5	14	7.1	2
4	1885	10	30	20	30	6.2	0	87	1908	6	27	23	21	6.1	0	170	1923	6	7	2	36	6.2	2
5	1885	12	7	13	2	6.3	0	88	1908	11	22	16	15	6.4	0	171	1923	9	2	18	49	6.3	2
6	1885	12	19	18	26	6.0	2	89	1909	9	17	4	39	6.8	0	172	1923	11	18	5	40	6.3	1
7	1886	4	13	5	44	6.3	0	90	1910	1	22	8	25	6.0	0	173	1923	12	27	23	39	6.4	0
8	1886	7	2	12	33	6.3	2	91	1910	5	9	18	53	6.0	1	174	1924	2	3	7	25	6.3	0
9	1887	5	29	0	50	6.4	0	92	1910	5	10	22	56	6.1	0	175	1924	5	31	21	2	6.3	1
10	1887	5	29	1	10	6.2	2	93	1910	5	12	12	22	6.0	2	176	1924	5	31	21	4	6.4	1
11	1888	2	5	0	50	7.1	0	94	1910	10	13	23	56	6.3	0	177	1924	8	6	23	22	6.3	0
12	1888	11	24	2	3	6.5	0	95	1912	1	4	4	4	6.1	0	178	1924	8	15	3	2	7.1	0
13	1889	3	31	6	42	6.6	0	96	1912	1	9	6	21	6.1	0	179	1924	8	15	8	27	6.7	2
14	1890	11	17	9	31	6.3	0	97	1912	6	8	13	41	6.6	0	180	1924	8	17	10	45	6.3	2
15	1891	4	7	9	49	6.7	0	98	1912	12	9	8	50	6.6	0	181	1924	8	17	11	10	6.6	2
16	1891	5	5	8	16	6.2	0	99	1913	2	20	17	58	6.9	0	182	1924	8	25	23	31	6.7	2
17	1891	7	21	20	19	7.0	0	100	1913	5	22	5	36	6.1	1	183	1925	2	7	2	11	6.0	0
18	1892	10	22	19	9	6.0	0	101	1913	5	29	19	14	6.4	0	184	1925	4	20	5	24	6.3	0
19	1894	2	25	4	18	6.8	0	102	1913	10	3	9	17	6.1	1	185	1925	6	2	14	18	6.4	0
20	1894	3	14	18	15	6.0	2	103	1913	10	11	18	10	6.9	0	186	1925	11	10	23	44	6.0	0
21	1894	8	29	19	55	6.6	0	104	1913	10	13	2	5	6.6	2	187	1926	4	7	4	33	6.3	0
22	1894	11	28	1	5	7.1	0	105	1914	2	7	15	50	6.8	0	188	1926	5	27	4	45	6.4	0
23	1894	12	1	18	37	6.3	0	106	1914	12	26	3	18	6.1	0	189	1926	9	5	0	37	6.8	0
24	1896	1	9	22	17	7.5	0	107	1915	3	9	0	29	6.8	0	190	1926	10	3	17	25	6.4	0
25	1896	1	10	5	52	6.0	2	108	1915	4	6	5	25	6.0	1	191	1926	10	19	9	29	6.2	0
26	1896	1	10	11	25	6.3	0	109	1915	4	6	14	32	6.2	0	192	1926	11	11	12	1	6.1	2
27	1896	2	23	19	42	6.1	2	110	1915	4	25	2	9	6.4	0	193	1927	1	18	6	58	6.4	0
28	1896	3	6	23	52	6.0	2	111	1915	5	28	2	26	6.0	2	194	1927	3	16	15	52	6.4	2
29	1896	4	11	23	0	6.0	2	112	1915	6	5	6	59	6.7	0	195	1927	7	30	23	18	6.4	0
30	1896	6	15	19	32	8.5	0	113	1915	7	9	7	21	6.4	0	196	1927	8	6	6	12	6.7	0
31	1896	6	16	4	16	7.5	2	114	1915	10	13	6	30	6.8	0	197	1927	9	30	16	38	6.3	0
32	1896	6	16	8	1	7.5	2	115	1915	10	14	4	43	6.2	2	198	1928	5	27	18	50	7.0	0
33	1896	7	29	17	44	6.1	2	116	1915	10	15	1	28	6.1	2	199	1928	5	29	0	35	6.7	2
34	1896	8	1	11	49	6.5	0	117	1915	10	15	3	40	6.3	2	200	1928	6	1	22	12	6.5	2
35	1896	9	5	23	7	6.5	2	118	1915	10	16	1	55	6.0	2	201	1928	6	2	7	6	6.0	2
36	1897	2	20	5	50	7.4	0	119	1915	10	17	0	21	6.1	2	202	1928	8	1	4	28	6.1	2
37	1897	2	20	8	47	7.0	2	120	1915	11	1	16	24	7.5	0	203	1929	3	15	10	57	6.0	2
38	1897	3	27	19	49	6.3	2	121	1915	11	1	16	50	6.7	2	204	1929	4	1	5	17	6.3	0
39	1897	5	23	21	22	6.9	2	122	1915	11	1	18	1	7.0	2	205	1929	4	16	9	53	6.3	0
40	1897	7	22	18	31	6.8	0	123	1915	11	2	0	43	6.2	2	206	1929	5	31	9	10	6.1	0
41	1897	7	29	22	45	6.0	2	124	1915	11	4	12	13	6.4	2	207	1929	6	27	1	49	6.1	0
42	1897	8	5	9	10	7.7	0	125	1915	11	18	13	4	7.0	2	208	1929	8	29	3	51	6.3	0

2. Calculating correlation, r.

ρ = rho = correlation of the population.
Suppose there are N people in the population,
X = temperature, Y = heart rate,
the mean and sd of temp in the pop. are μ_x and σ_x,
and the pop. mean and sd of heart rate are μ_y and σ_y.

$$\rho = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i-\mu_x}{\sigma_x}\right)\left(\frac{y_i-\mu_y}{\sigma_y}\right).$$

Given a sample of size n, we estimate ρ using

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{s_x}\right)\left(\frac{y_i-\overline{y}}{s_y}\right).$$

This is in Appendix A.

3. Linear Regression

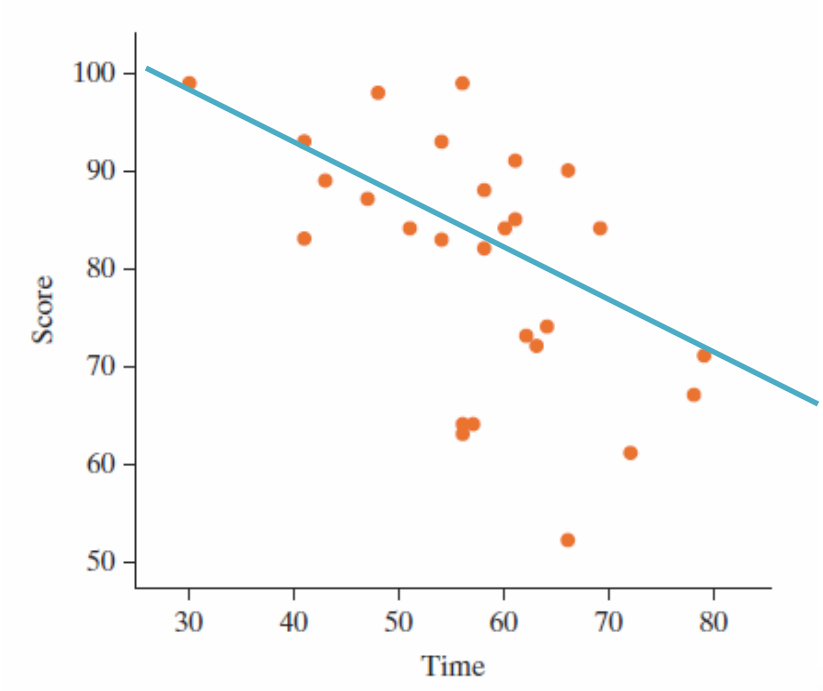
Section 10.3

Introduction

- If we decide an association is linear, it is helpful to develop a mathematical model of that association.
- Helps make predictions about the response variable.
- The ***least-squares regression line*** is the most common way of doing this.

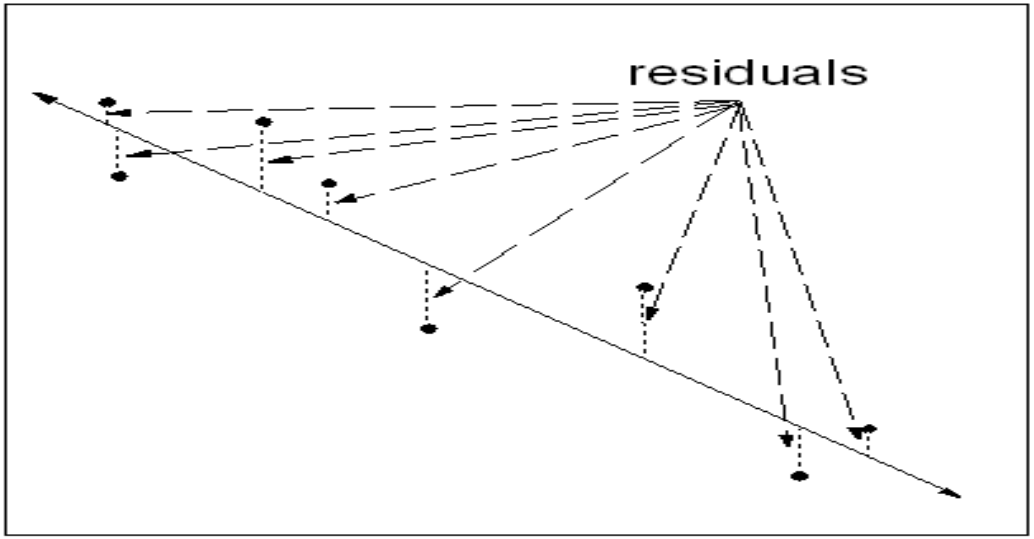
Introduction

- Unless the points are perfectly linearly alligned, there will not be a single line that goes through every point.



Introduction

- We want a line that minimizes the vertical distances between the line and the points
  - These distances are called **residuals.**
  - The line we will find actually minimizes the sum of the squares of the residuals.
  - This is called a **least-squares regression line**.

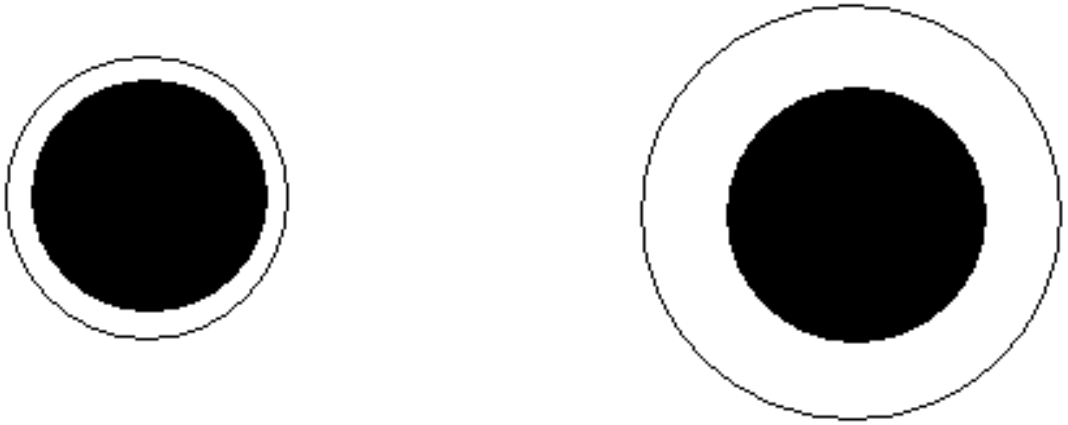


Are Dinner Plates Getting Larger?

Example 10.3

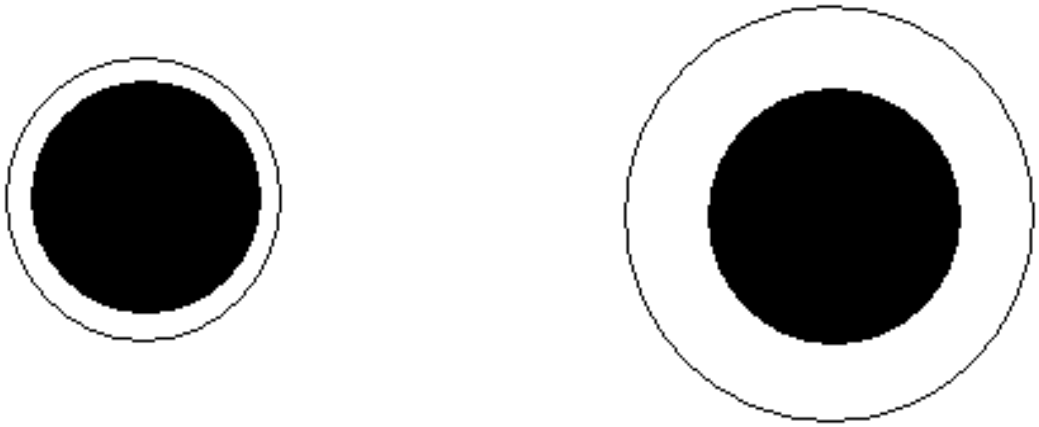
Growing Plates?

- There are many recent articles and TV reports about the obesity problem.
- One reason some have given is that the size of dinner plates are increasing.
- Are these black circles the same size, or is one larger than the other?



Growing Plates?

- They appear to be the same size for many, but the one on the right is about 20% larger than the left.



- This suggests that people will put more food on larger dinner plates without knowing it.
- There is name for this phenomenon: *Delboeuf illusion*.

Growing Plates?

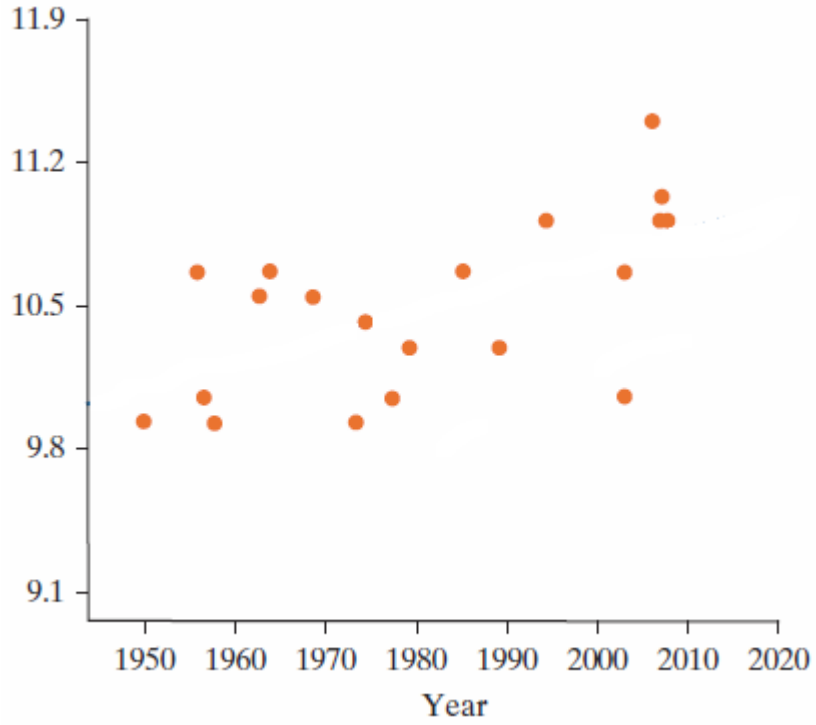
- Researchers gathered data to investigate the claim that dinner plates are growing
- American dinner plates sold on ebay on March 30, 2010 (Van Ittersum and Wansink, 2011)
- Year manufactured and diameter are given.

TABLE 10.1 Data for size (diameter, in inches) and year of manufacture for 20 American-made dinner plates

Year	1950	1956	1957	1958	1963	1964	1969	1974	1975	1978
Size	10	10.75	10.125	10	10.625	10.75	10.625	10	10.5	10.125
Year	1980	1986	1990	1995	2004	2004	2007	2008	2008	2009
Size	10.375	10.75	10.375	11	10.75	10.125	11.5	11	11.125	11

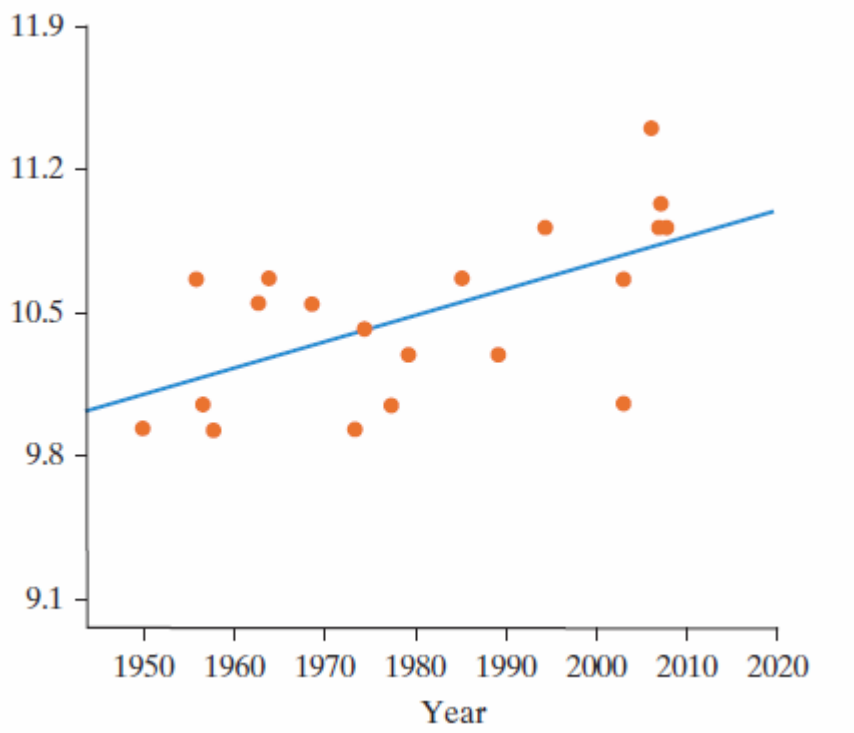
Growing Plates?

- Both year (explanatory variable) and diameter in inches (response variable) are quantitative.
- Each dot in this scatterplot represents one plate.



Growing Plates?

- The association appears to be roughly linear.
- The least squares regression line is added.
- The line slopes upward, but is the slope significant?



Regression Line

The regression equation is $\hat{y} = a + bx$:

- *a* is the *y*-intercept
- *b* is the slope
- *x* is a value of the explanatory variable
- $\hat{y}$ is the predicted value for the response variable

- For a specific value of *x*, the corresponding distance $y - \hat{y}$ (or actual – predicted) is a residual

Regression Line

- The least squares line for the dinner plate data is $\hat{y} = -14.8 + 0.0128x$
- Or $\widehat{\text{diameter}} = -14.8 + 0.0128(\text{year})$
- This allows us to predict plate diameter for a particular year.

Slope

$$\hat{y} = -14.8 + 0.0128x$$

- What is the predicted diameter for a plate manufactured in 2000?
  - -14.8 + 0.0128(2000) = 10.8 in.
- What is the predicted diameter for a plate manufactured in 2001?
  - -14.8 + 0.0128(2001) = 10.8128 in.
- How does this compare to our prediction for the year 2000?
  - 0.0128 larger
- Slope *b* = 0.0128 means that diameters are predicted to increase by 0.0128 inches per year on average

Slope

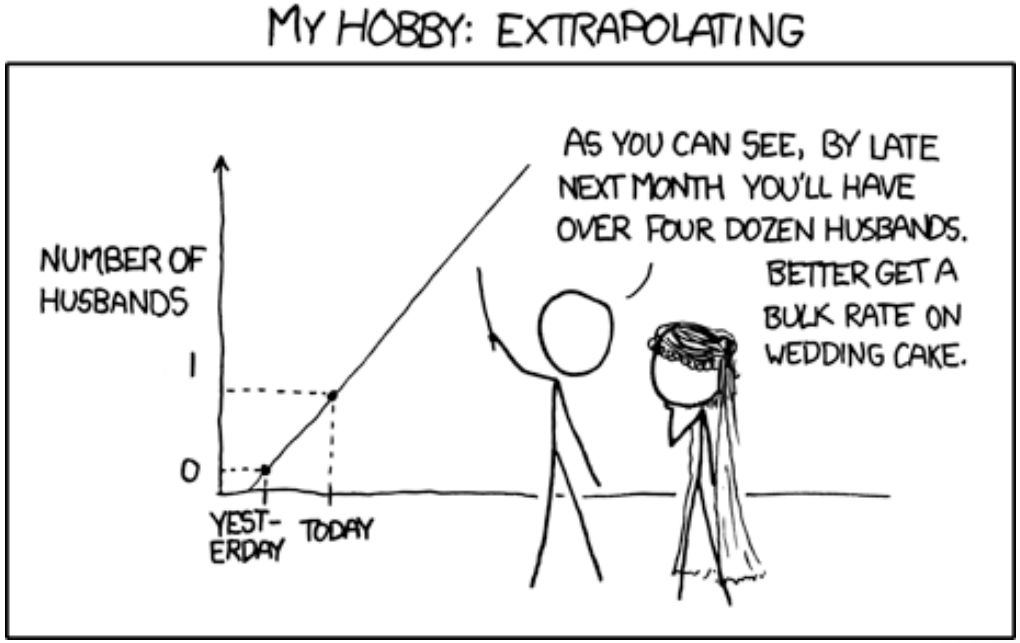
- Slope is the predicted change in the response variable for one-unit change in the explanatory variable.
- Both the slope and the correlation coefficient for this study were positive.
  - The slope is 0.0128
  - The correlation is 0.604
- The slope and correlation coefficient will always have the same sign.

y-intercept

- The *y*-intercept is where the regression line crosses the *y*-axis. It is the predicted response when the explanatory variable equals 0.
- We had a *y*-intercept of -14.8 in the dinner plate equation. What does this tell us about our dinner plate example?
  - Dinner plates in year 0 would be predicted to be -14.8 inches???
- How can it be negative?
  - The equation works well within the range of values given for the explanatory variable, but fails outside that range.
- Our equation should only be used to predict the size of dinner plates from about 1950 to 2010.

Extrapolation

- Predicting values for the response variable for values of the explanatory variable that are outside of the range of the original data is called ***extrapolation***.



Coefficient of Determination

- While the intercept and slope have meaning in the context of year and diameter, remember that the correlation does not. It is just 0.604.
- However, the square of the correlation (coefficient of determination or r^2) does have meaning.
- $r^2 = 0.604^2 = 0.365$ or 36.5%
- 36.5% of the variation in plate size (the response variable) can be explained by its linear association with the year (the explanatory variable).

Learning Objectives for Section 10.3

- Understand that one way a scatterplot can be summarized is by fitting the best-fit (least squares regression) line.
- Be able to interpret both the slope and intercept of a best-fit line in the context of the two variables on the scatterplot.
- Find the predicted value of the response variable for a given value of the explanatory variable.
- Understand the concept of residual and find and interpret the residual for an observational unit given the raw data and the equation of the best fit (regression) line.
- Understand the relationship between residuals and strength of association and that the best-fit (regression) line this minimizes the sum of the squared residuals.

Learning Objectives for Section 10.3

- Find and interpret the coefficient of determination (r^2) as the squared correlation and as the percent of total variation in the response variable that is accounted for by the linear association with the explanatory variable.
- Understand that extrapolation is when a regression line is used to predict values outside of the range of observed values for the explanatory variable.
- Understand that when slope = 0 means no association, slope < 0 means negative association, slope > 0 means positive association, and that the sign of the slope will be the same as the sign of the correlation coefficient.
- Understand that influential points can substantially change the equation of the best-fit line.

4. slope of regression line.

- Suppose $\hat{y}$ = a + bx is the regression line.
- The slope b of the regression line is b = r $\frac{s_y}{s_x}$.

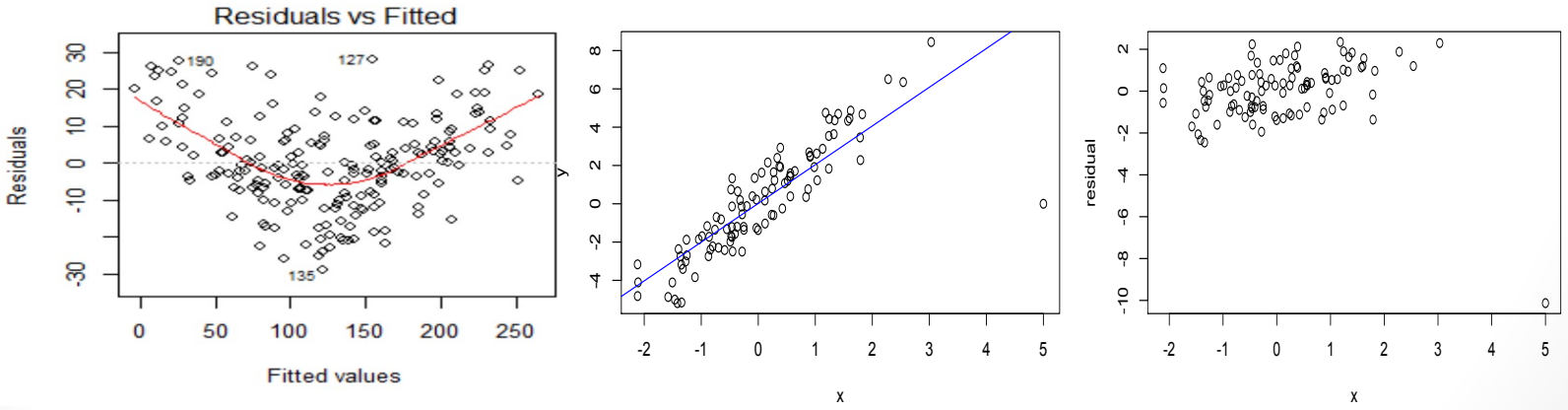
This is usually the thing of primary interest to interpret, as the predicted increase in y for every unit increase in x.

- Beware of assuming causation though, esp. with observational studies. Be wary of extrapolation too.
- The intercept a = $\overline{y}$ - b $\overline{x}$.
- The SD of the residuals is $\sqrt{1-r^2}\, s_y$.

This is a good estimate of how much the regression predictions will typically be off by.

5. How well does the line fit?

- r^2 is a measure of fit. It indicates the amount of scatter around the best fitting line.
- $\sqrt{1-r^2}\, s_y$ is useful as a measure of how far off predictions would have been on average.
- Residual plots can indicate curvature, outliers, or heteroskedasticity.



- Note that regression residuals have mean zero, whether the line fits well or poorly.

- Heteroskedasticity: when the variability in y is not constant as x varies.