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

1. Calculating correlation.
2. Linear regression, growing plates example.
3. Slope of regression line.
4. y-intercept of regression line.
5. Extrapolation.
6. r^2.
7. How well does the line fit?
8. Common problems with regression.

There will be no lecture or OH Mon Mar11.
Read ch9.
Hw4 is due Mon Mar11 1159pm by email to statgrader or statgrader2.
10.1.8, 10.3.14, 10.3.21, and 10.4.11. See day14 notes for the problems.

http://www.stat.ucla.edu/~frederic/13/W24 .

For the final exam, bring a pencil or pen, and a calculator. On the final exam, you cannot use computers or ipads or phones or anything that can surf the web or do email.

1. 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.

2. Linear Regression

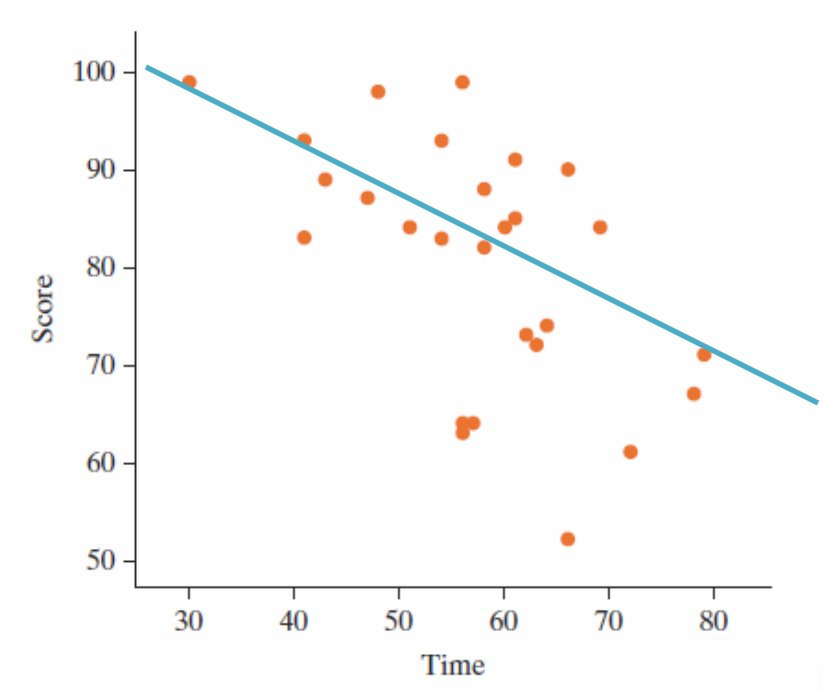
Section 10.3

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.

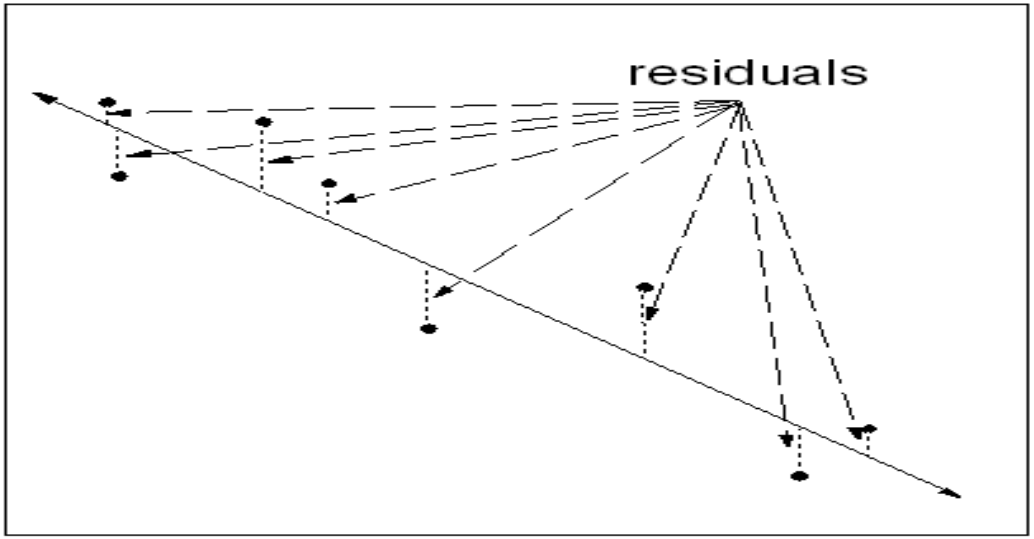
Introduction

- 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**.

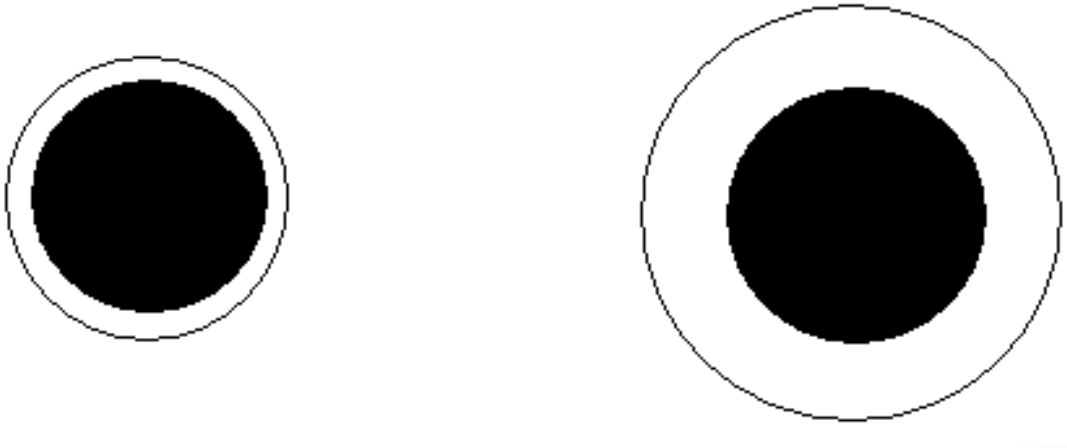


Are Dinner Plates Getting Larger?

Example 10.3

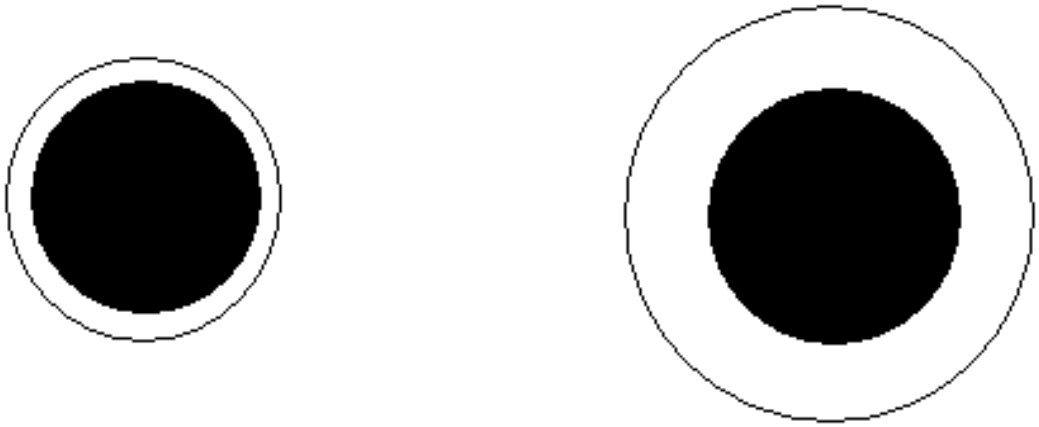
Growing Plates?

- 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?

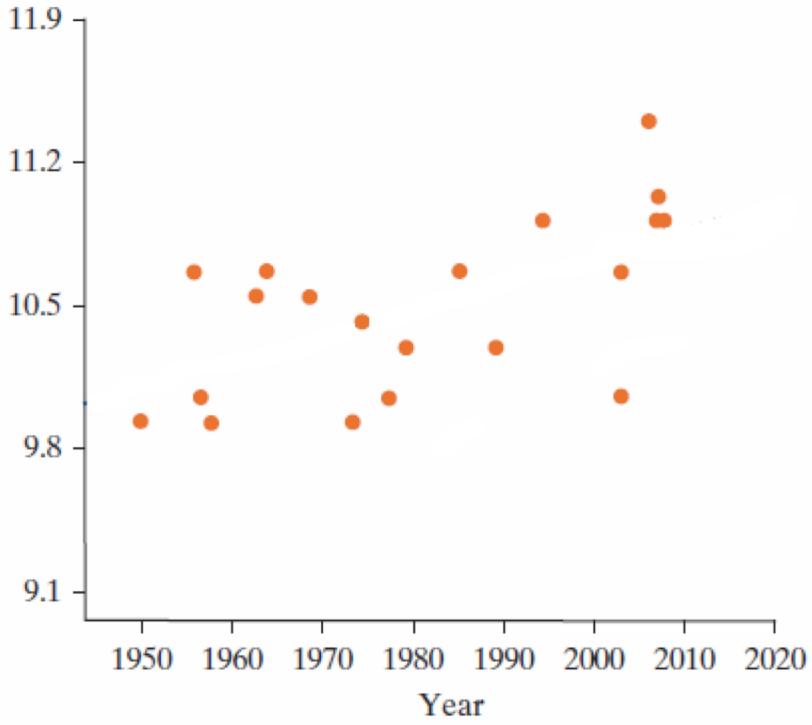
- Researchers gathered data to investigate the claim that dinner plates are growing
- American dinner plates sold on ebay on March 30, 2010 (Van Ittersum and Wansink, 2011)
- Year manufactured and diameter are given.

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

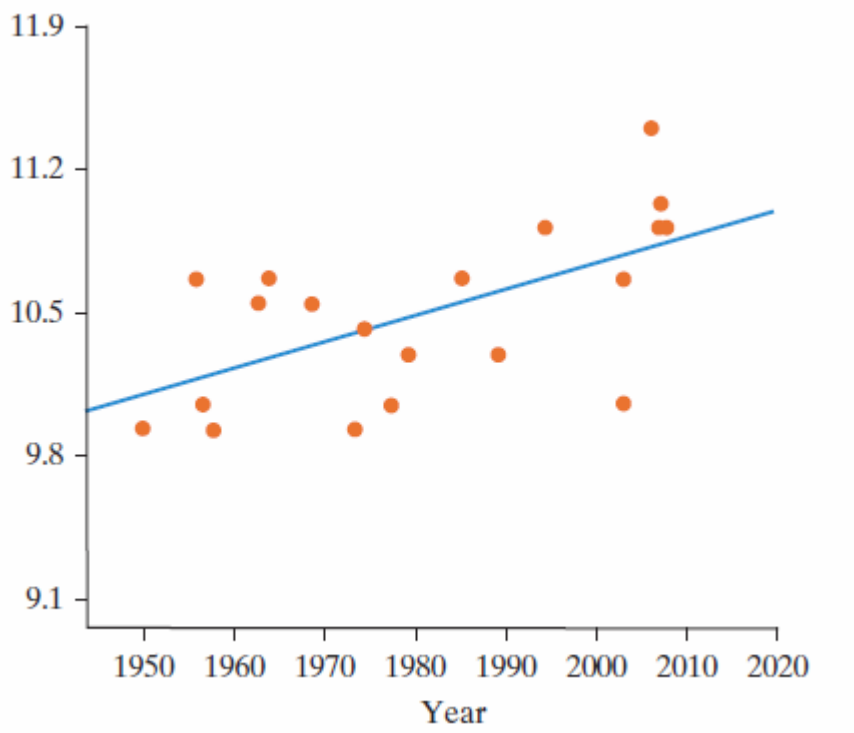
Growing Plates?

- 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.

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 $\sqrt{1-r^2}\, s_y$.

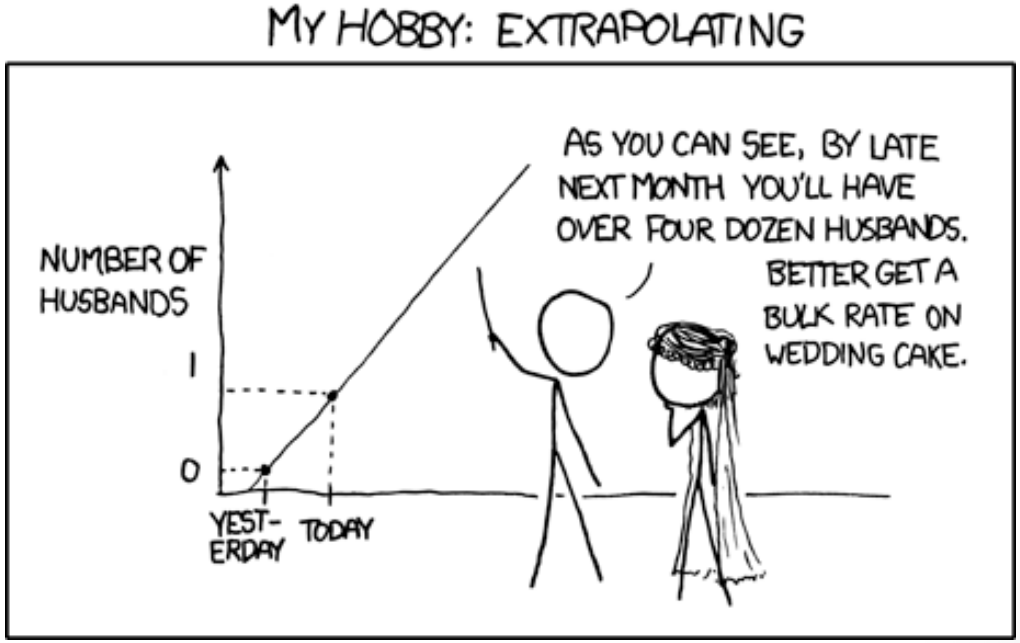
This is a good estimate of how much the regression predictions will typically be off by.

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***.



r^2

- 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

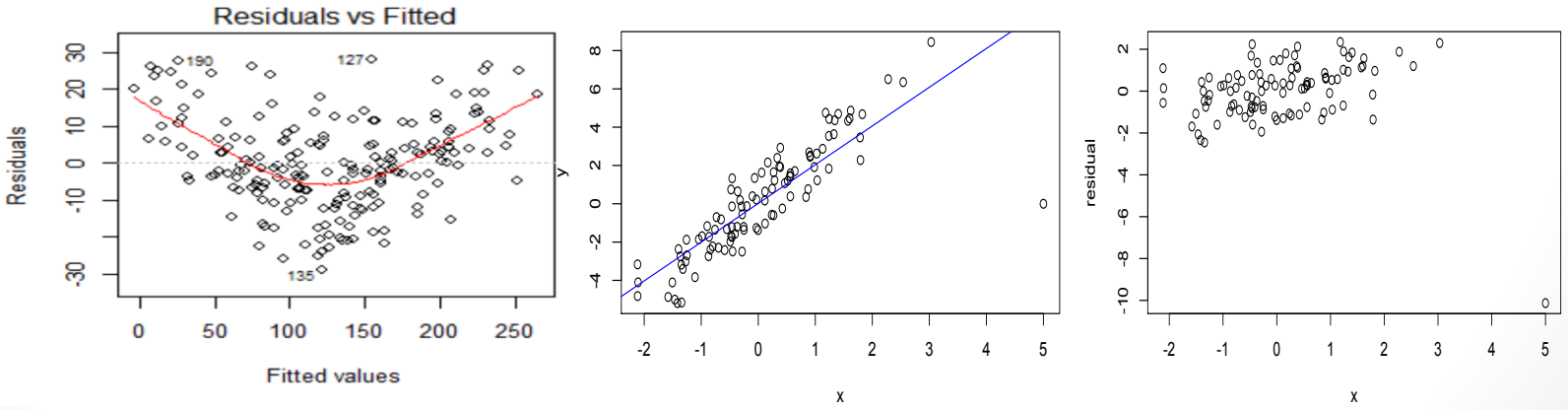
- Understand that one way a scatterplot can be summarized is by fitting the best-fit (least squares regression) line.
- Be able to interpret both the slope and intercept of a best-fit line in the context of the two variables on the scatterplot.
- Find the predicted value of the response variable for 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.

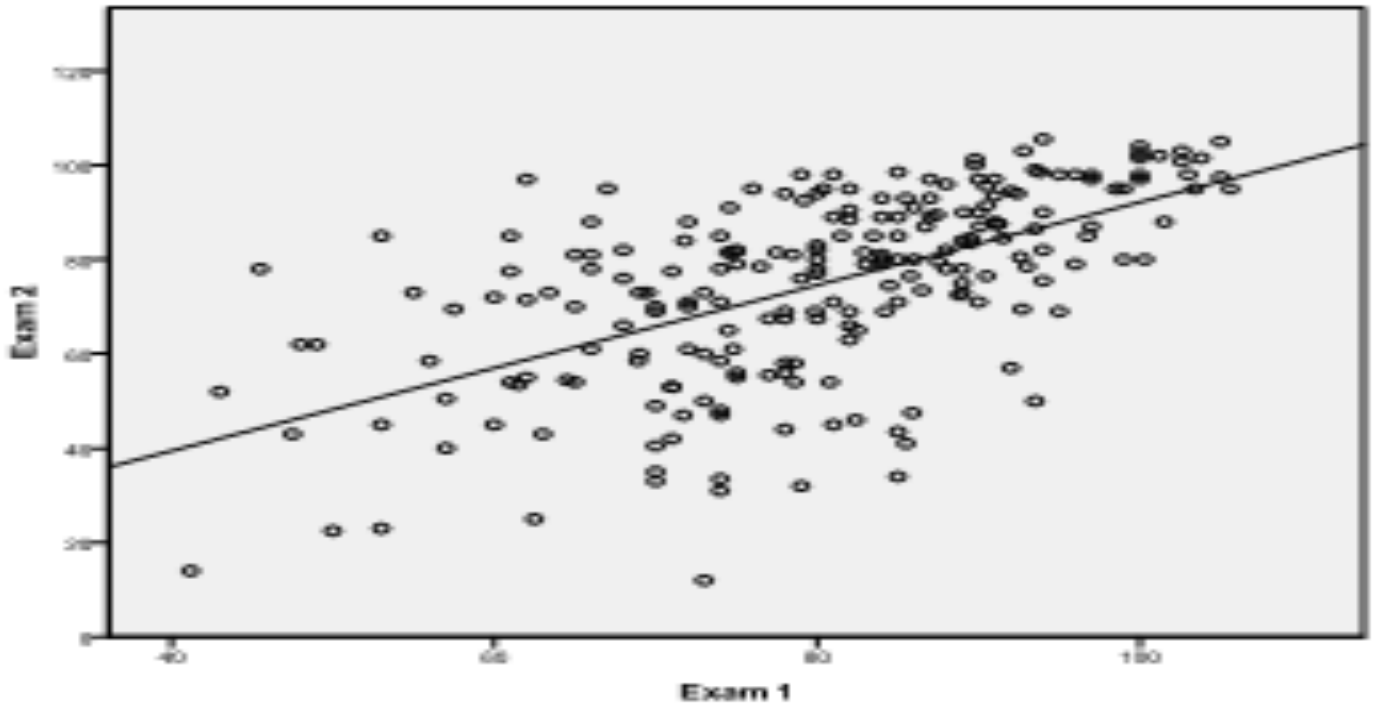
7. How well does the line fit?

- 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 regression line fits well or poorly.

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

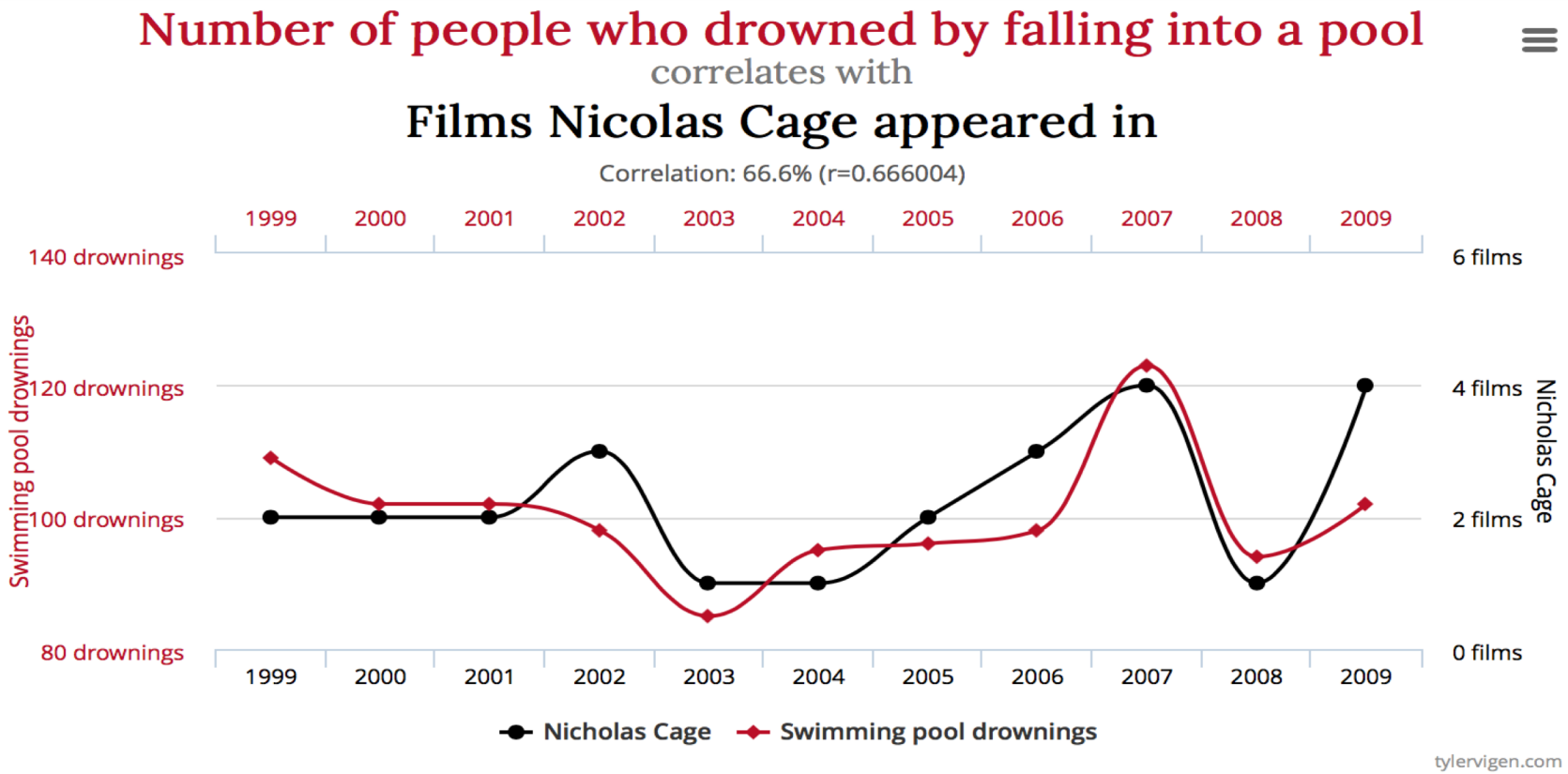


(b)

8. Common problems with regression.

- a. Correlation is not causation.

ESPECIALLY WITH OBSERVATIONAL DATA!



Data sources: Centers for Disease Control & Prevention and Internet Movie Database

Common problems with regression.

Common problems with regression.

Holmes and Willett (2004) reviewed all prospective studies on fat consumption and breast cancer with at least 200 cases of breast cancer. "Not one study reported a significant positive association with total fat intake.... Overall, no association was observed between intake of total, saturated, monounsaturated, or polyunsaturated fat and risk for breast cancer."

They also state "The dietary fat hypothesis is largely based on the observation that national per capita fat consumption is highly correlated with breast cancer mortality rates. However, per capita fat consumption is highly correlated with economic development. Also, low parity and late age at first birth, greater body fat, and lower levels of physical activity are more prevalent in Western countries, and would be expected to confound the association with dietary fat."

Common problems with regression.

- b. Extrapolation.

Common problems with regression.

- b. Extrapolation.
- Often researchers extrapolate from high doses to low.

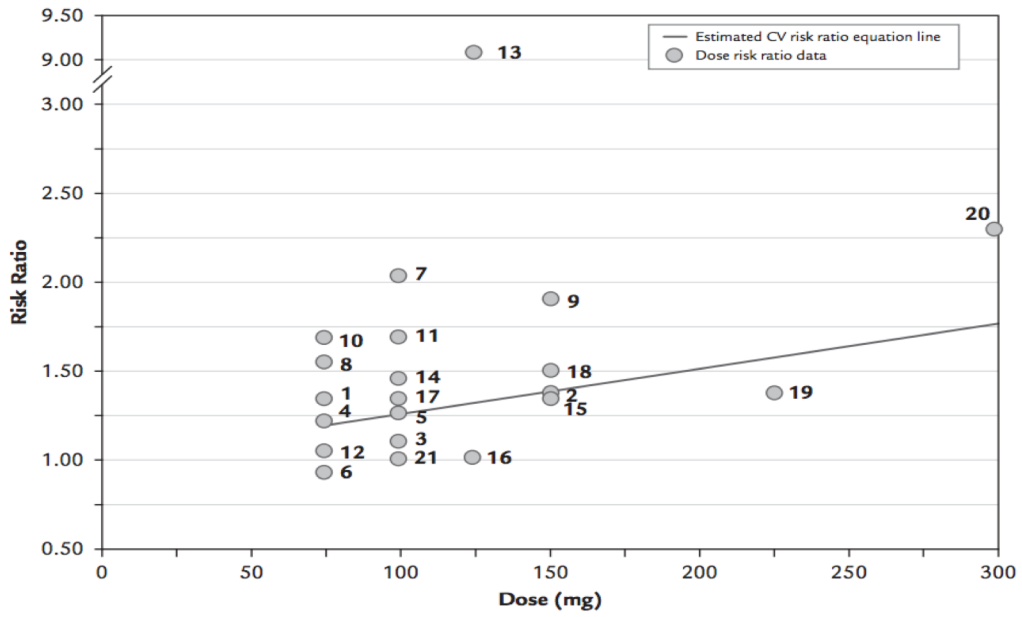


Figure 4. Relationship between diclofenac daily dose and the estimated risk ratio of a cardiovascular event. Numbers correspond to the observations in Table III.

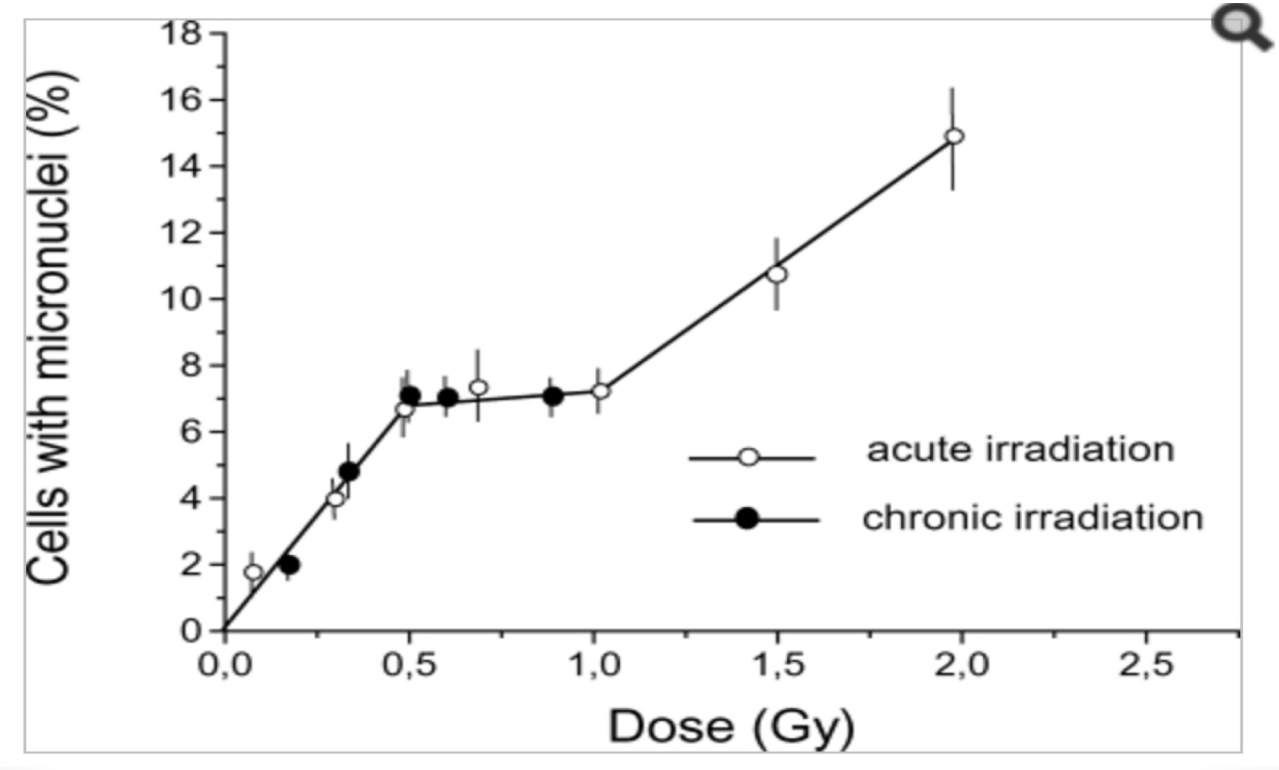
Common problems with regression.

- b. Extrapolation.

The relationship can be nonlinear though.

Researchers also often extrapolate from animals to humans.

Zaichkina et al. (2004) on hamsters



Common problems with regression.

- c. Curvature.

The best fitting line might fit poorly. Port et al. (2005).

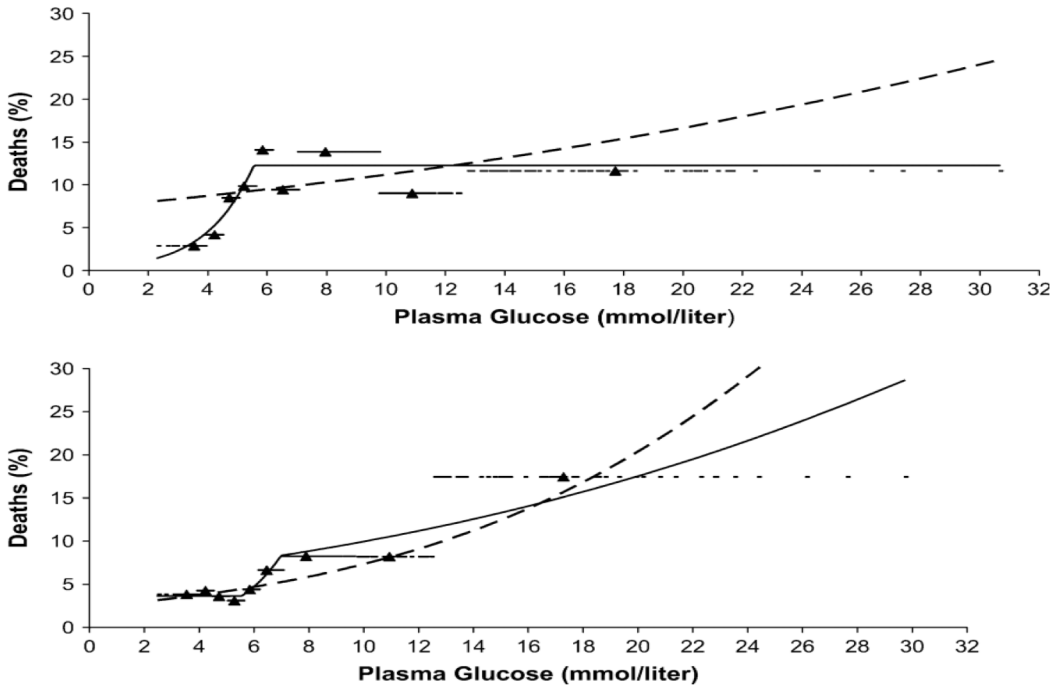
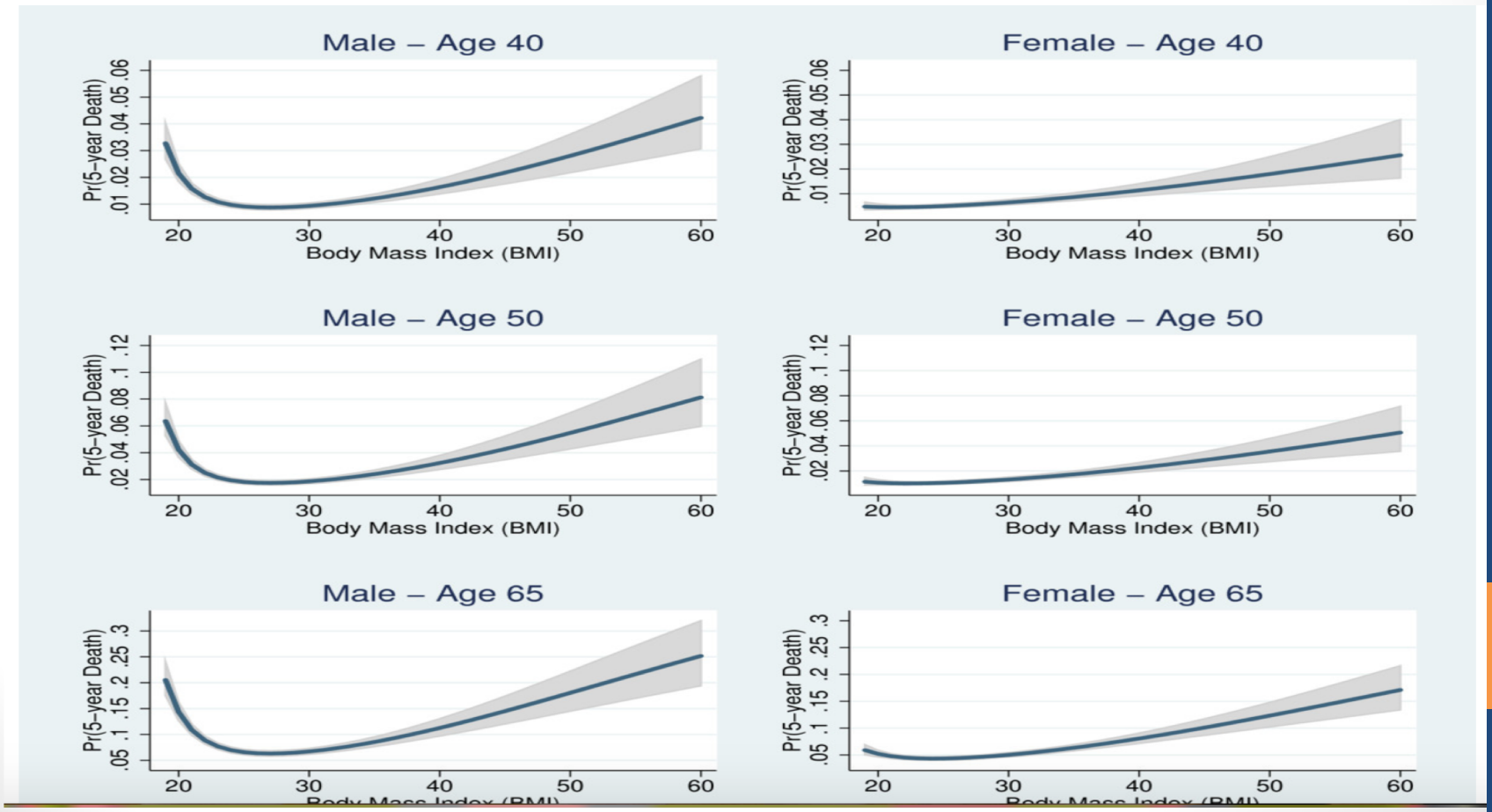


FIGURE 4. Adjusted 2-year rates of death from all causes for men (upper panel) and women (lower panel) separately, by glucose level, predicted by three models, Framingham Heart Study, 1948–1978. Linear model (dashed curve); optimal spline models (solid curve). The horizontal dashed

Common problems with regression.

- c. Curvature.

The best fitting line might fit poorly. Wong et al. (2011).



Common problems with regression.

- d. Statistical significance.

Could the observed correlation just be due to chance alone?