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

1. Extra credit points.
2. Linear regression.
3. Extrapolation.
4. R^2, or coefficient of determination.
5. Slope of the regression line.
6. 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.

On 10.4.11, the histogram shows simulated SLOPES of the regression line, under Ho.
And it gives you the mean and SD of these slopes. Use that.
Also, the blue line in the first plot in 10.4.11 slopes upward but the slope is -0.9658, which is negative. I will tell the grader to accept it if you use either 0.9658 or -0.9658 as the observed slope.
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%.

2 **Point processes, spatial—temporal**

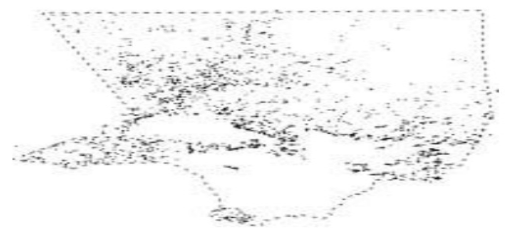


Figure 3 Centroids of recorded Los Angeles County wildfires, 1878—1996

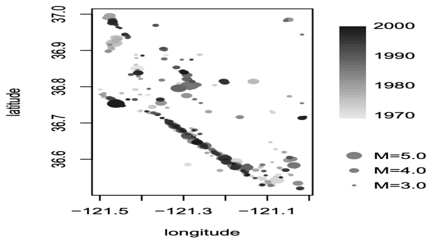


Figure 1: Locations, times and magnitudes of moderate-sized ($M \geq 3.5$) earthquakes in Bear Valley, CA, between 1970 and 2000.

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
- *ŷ* is the predicted value for the response variable

- For a specific value of *x*, the corresponding distance $\text{y} - \hat{y}$ (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

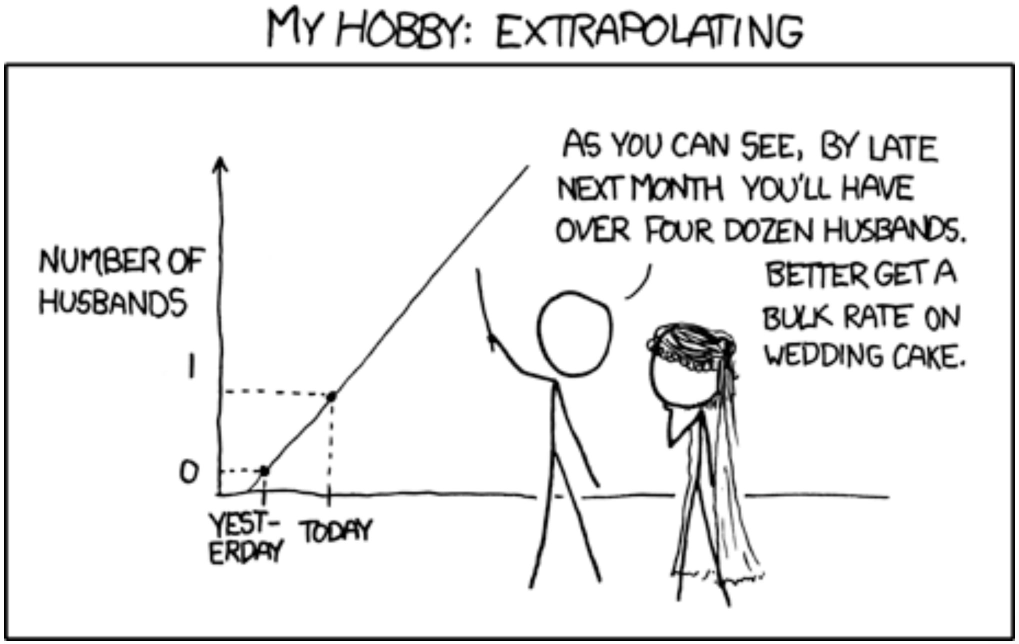
- Slope is the predicted change in the response variable for one-unit change in the explanatory variable.
- Both the slope and the correlation coefficient for this study were positive.
    - The slope is 0.0128
    - The correlation is 0.604
- The slope and correlation coefficient will always have the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


MY HOBBY: EXTRAPOLATING
NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS. BETTER GET A BULK RATE ON WEDDING CAKE.

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.

5. Slope of regression line.

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

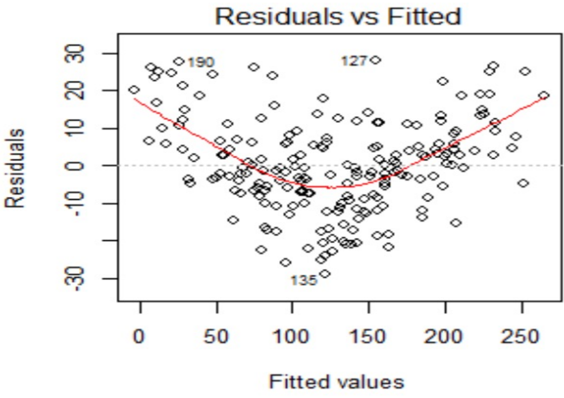
This is usually the thing of primary interest to interpret, as the predicted increase in y for every unit increase in x.

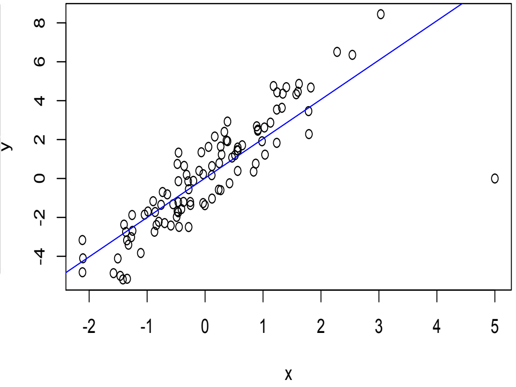
- Beware of assuming causation though, esp. with observational studies. Be wary of extrapolation too.
- The intercept a = $\bar{y}$ - b $\bar{x}$.
- The SD of the residuals is $\sqrt{1-r^2}\, s_y$.

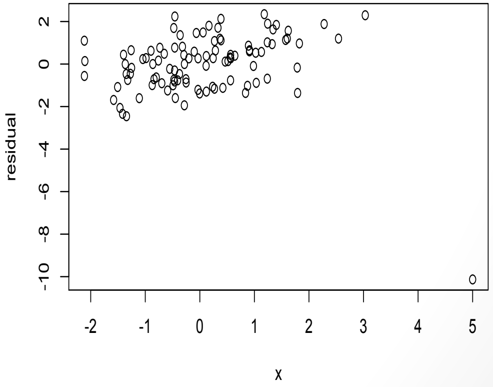
This is a good estimate of how much the regression predictions will typically be off by.

6. 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.

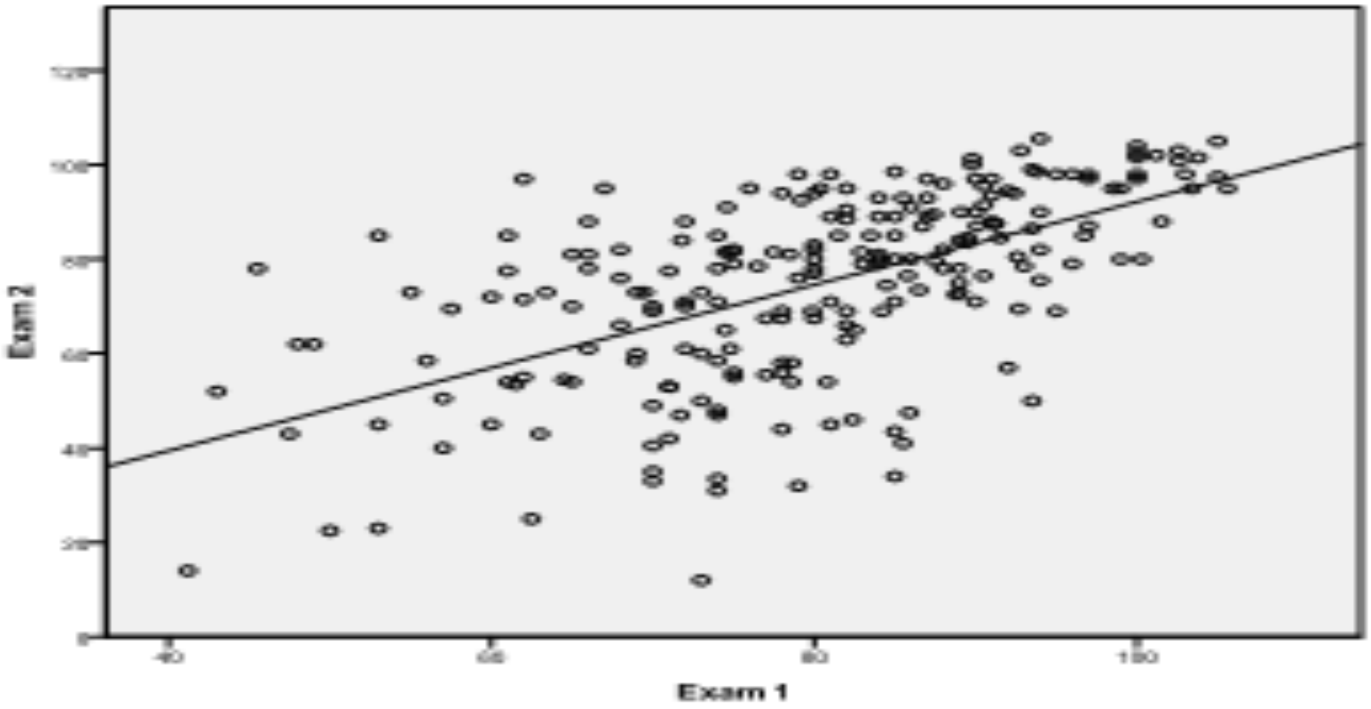






- Note that regression residuals have mean zero, whether the regression line fits well or poorly.

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



(b)

7. Common problems with regression.

- a. Correlation is not causation.

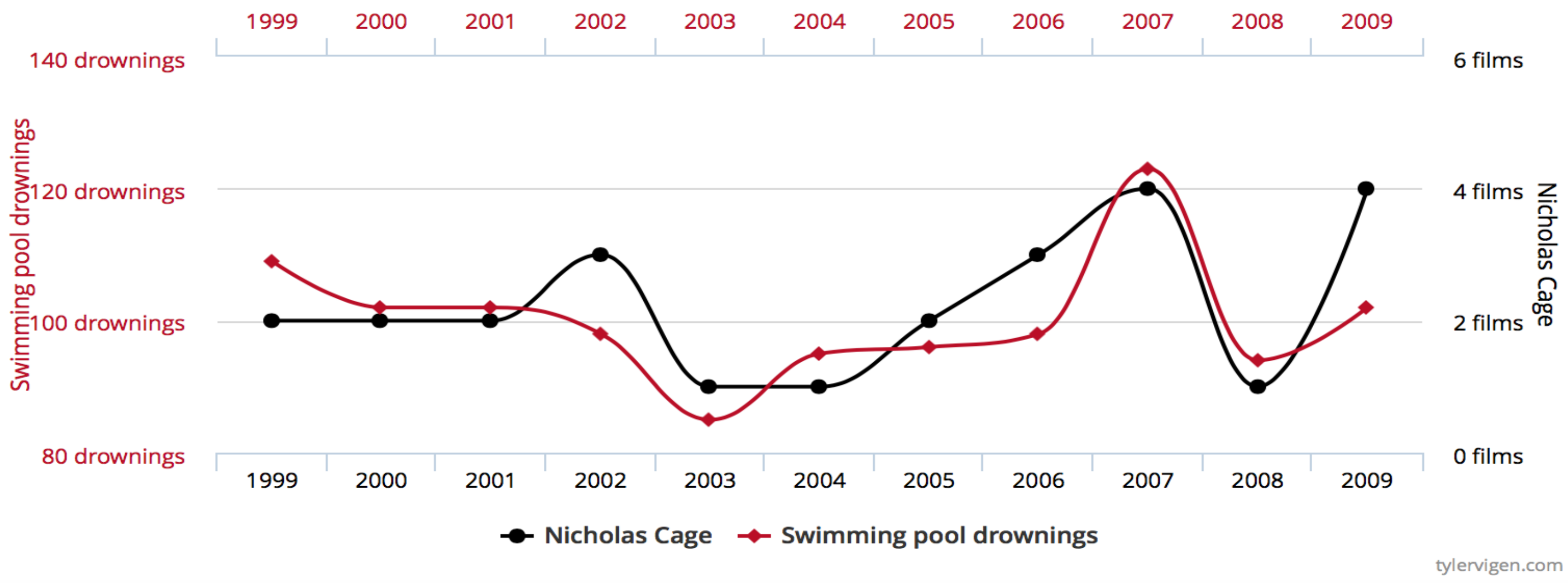
ESPECIALLY WITH OBSERVATIONAL DATA!

Number of people who drowned by falling into a pool

correlates with

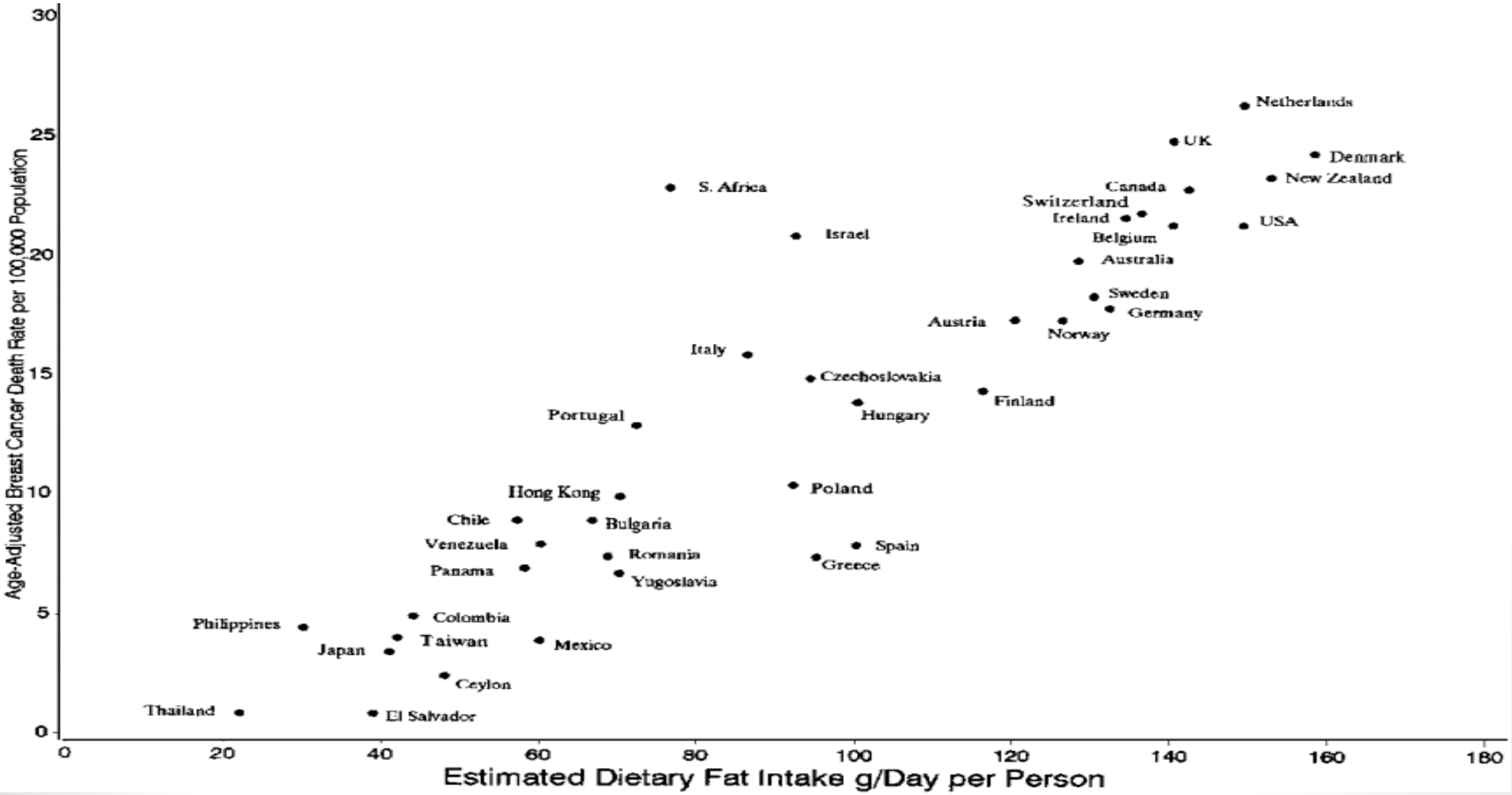
Films Nicolas Cage appeared in

Correlation: 66.6% (r=0.666004)



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

Common problems with regression.



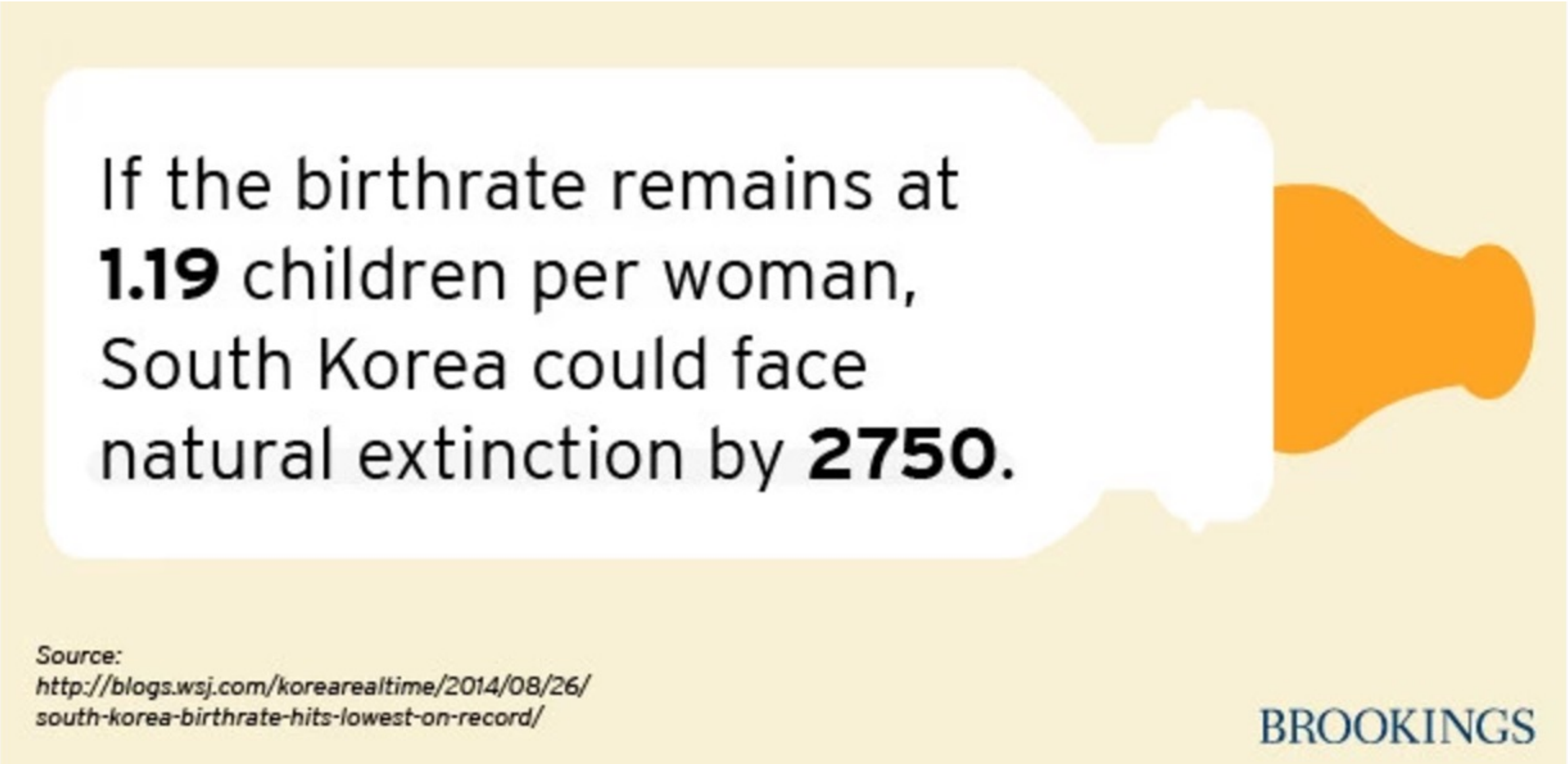
Common problems with regression.

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.



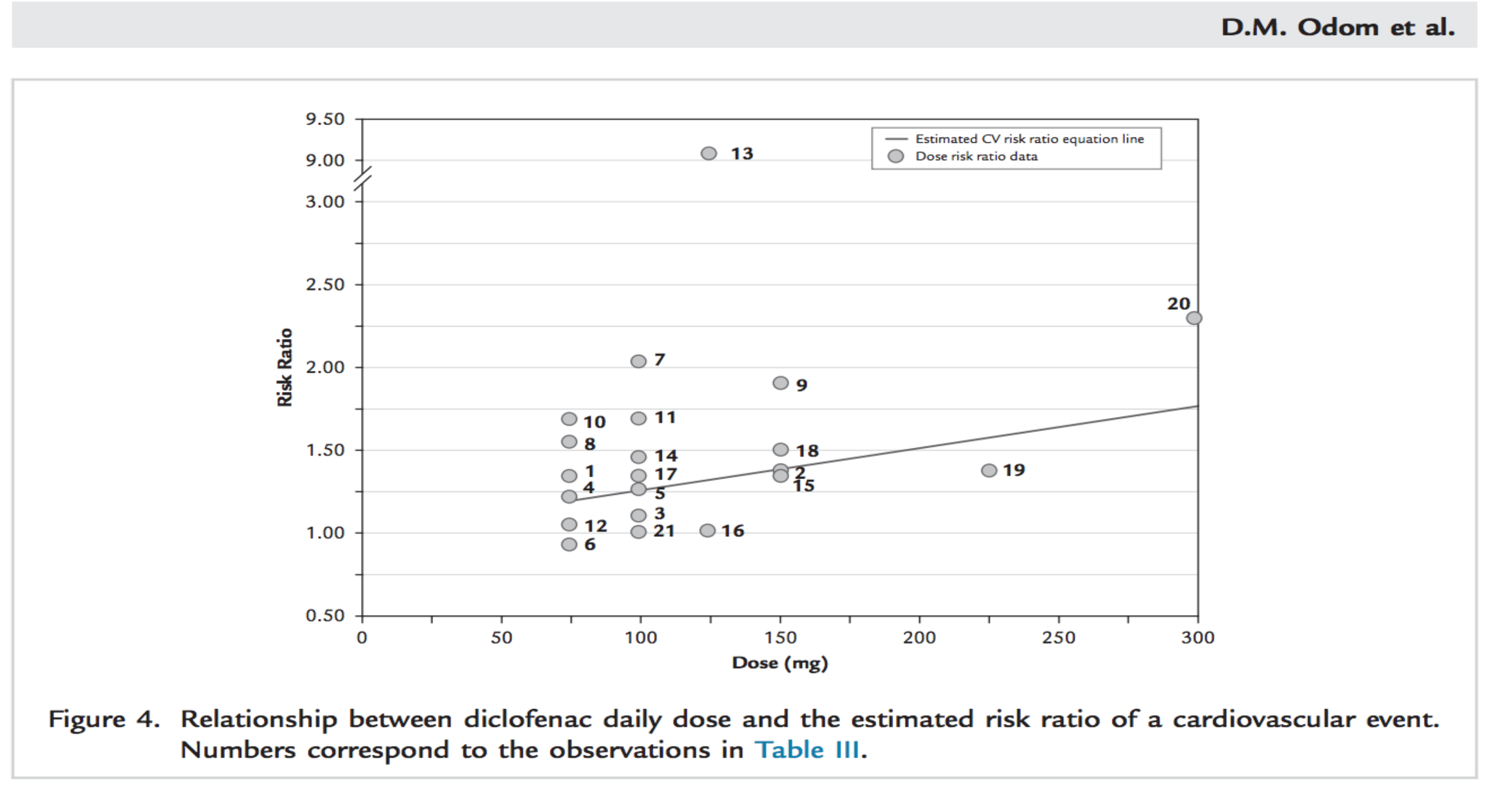
If the birthrate remains at **1.19** children per woman, South Korea could face natural extinction by **2750**.

Source:
http://blogs.wsj.com/korearealtime/2014/08/26/
south-korea-birthrate-hits-lowest-on-record/

BROOKINGS

Common problems with regression.

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



D.M. Odom et al.

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

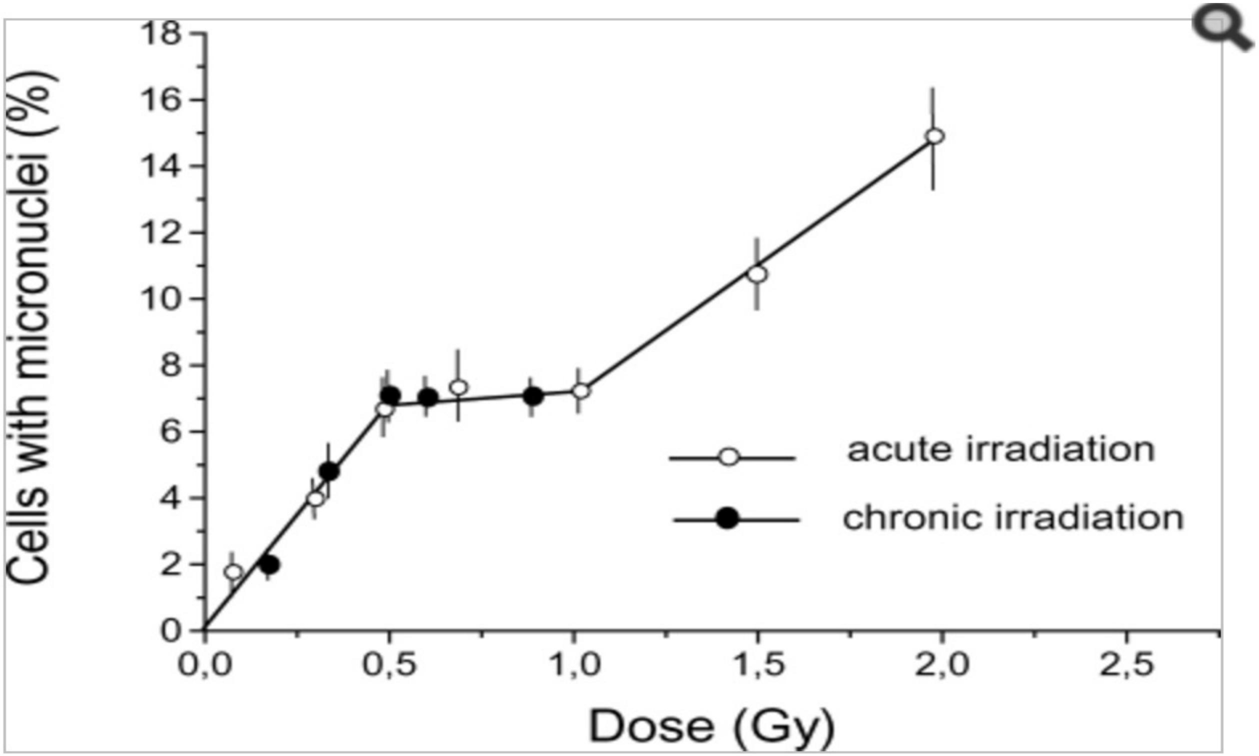
Common problems with regression.

- b. Extrapolation.

The relationship can be nonlinear though.

Researchers also often extrapolate from animals to humans.

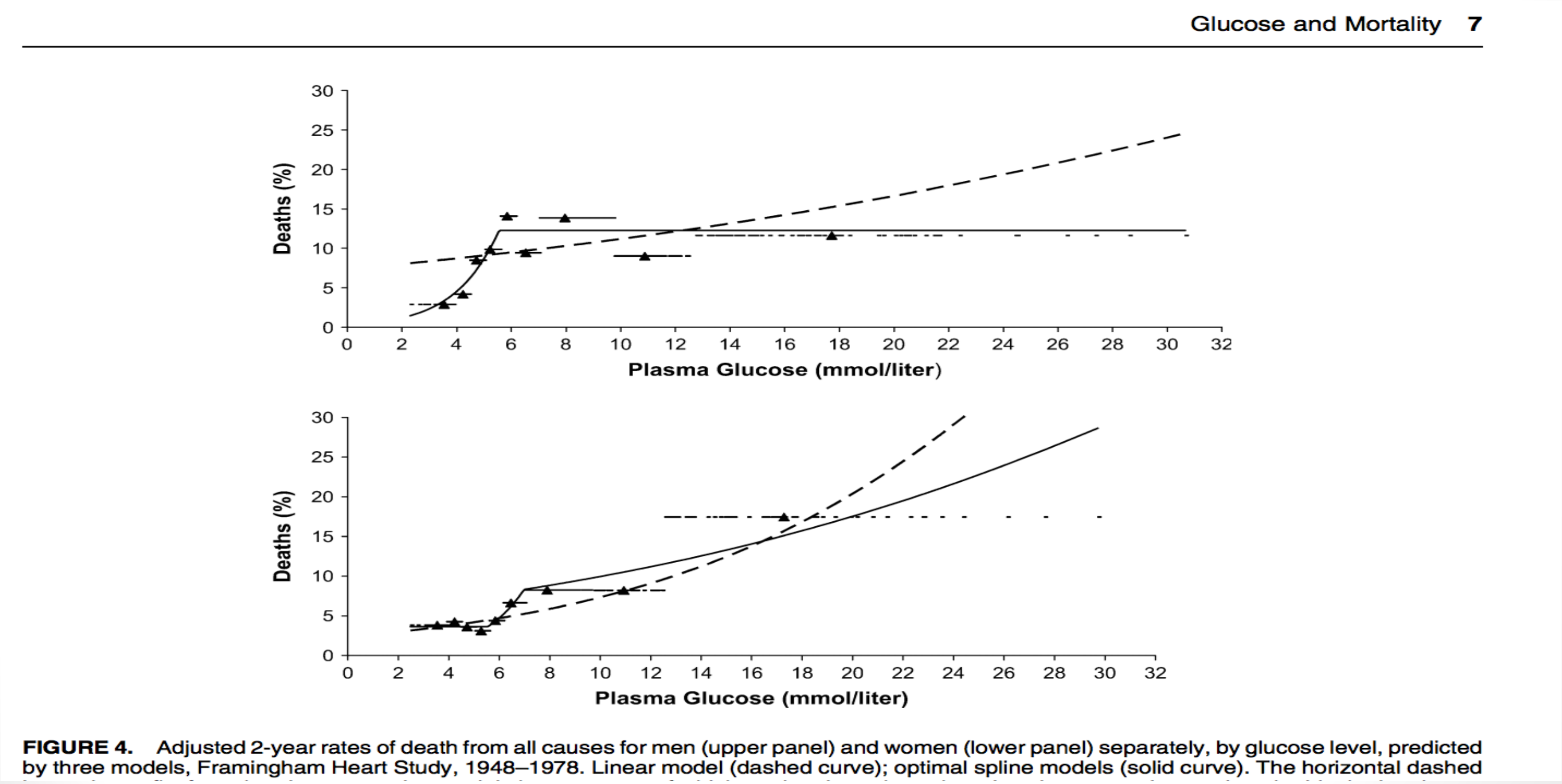
Zaichkina et al. (2004) on hamsters



Common problems with regression.

- c. Curvature.

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



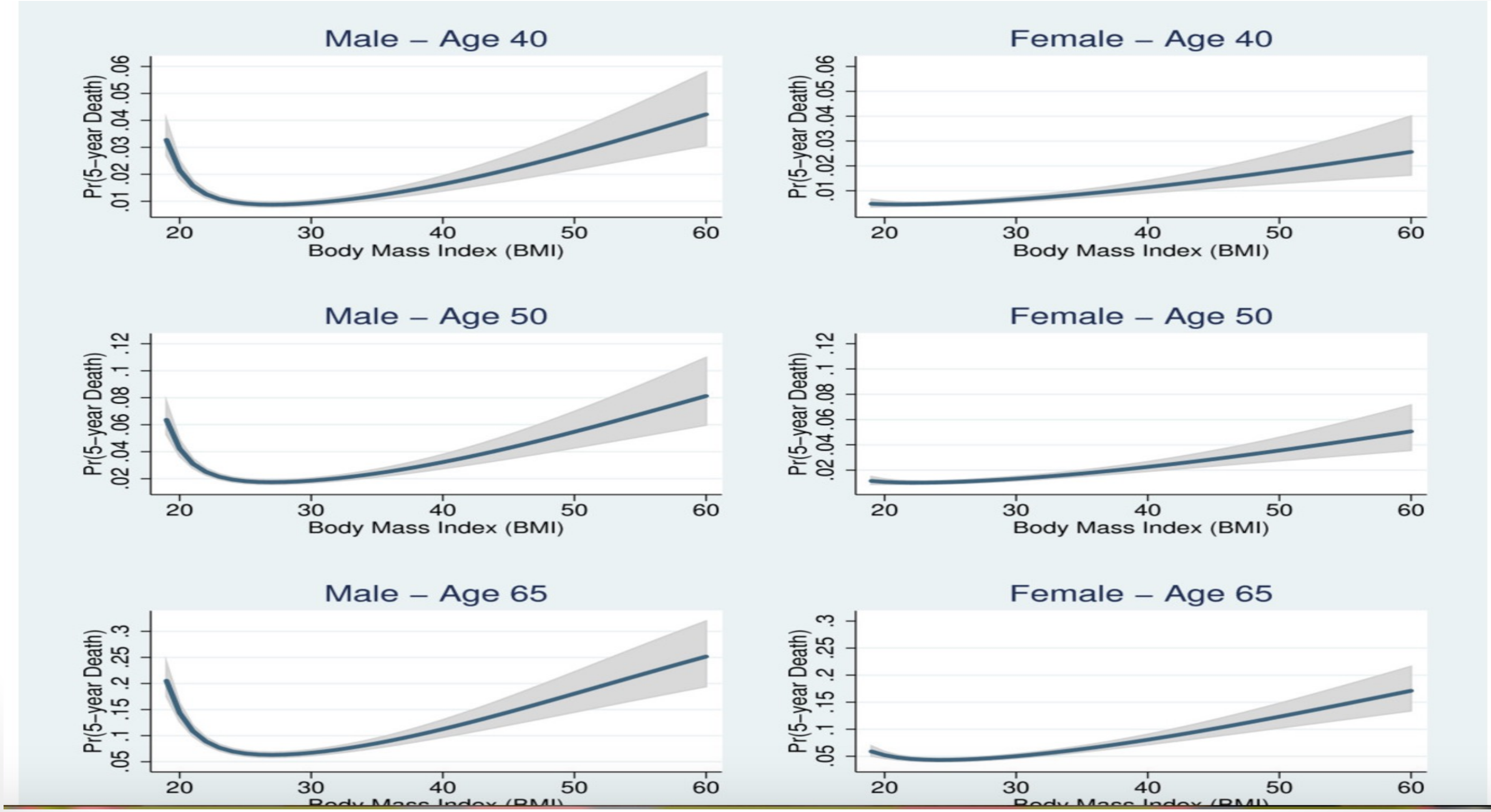
Glucose and Mortality 7

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.

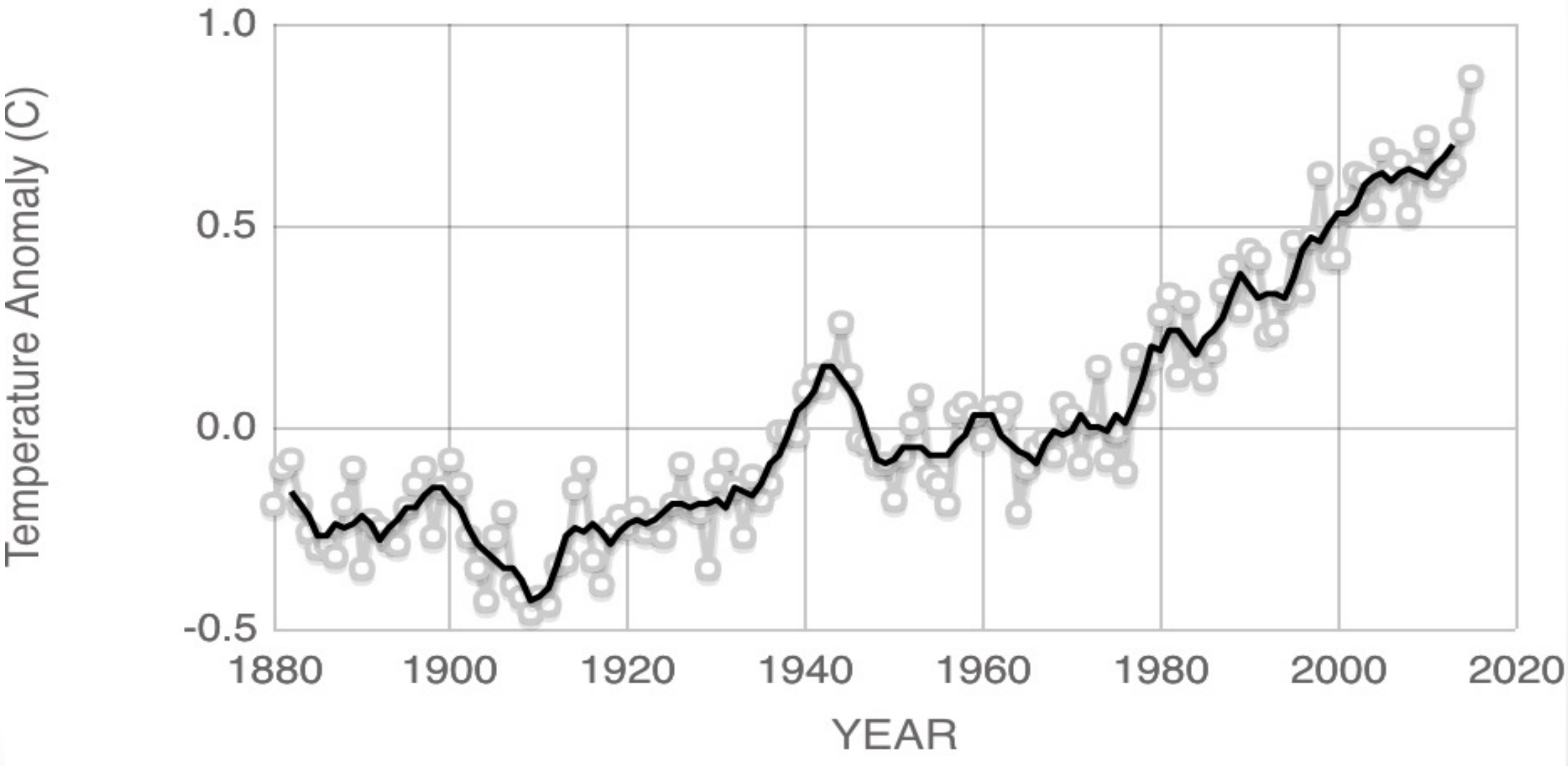
The best fitting line might fit poorly. Wong et al. (2011).



Common problems with regression.

- d. Statistical significance.

Could the observed correlation just be due to chance alone?



Source: climate.nasa.gov