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

1. Regression line and dinner plates example, continued.
2. Extrapolation.
3. Goodness of fit.
4. Common problems with regression.
5. Testing the slope of the regression line.

NO LECTURE OR Office Hour TUESDAY JUNE 2!!!

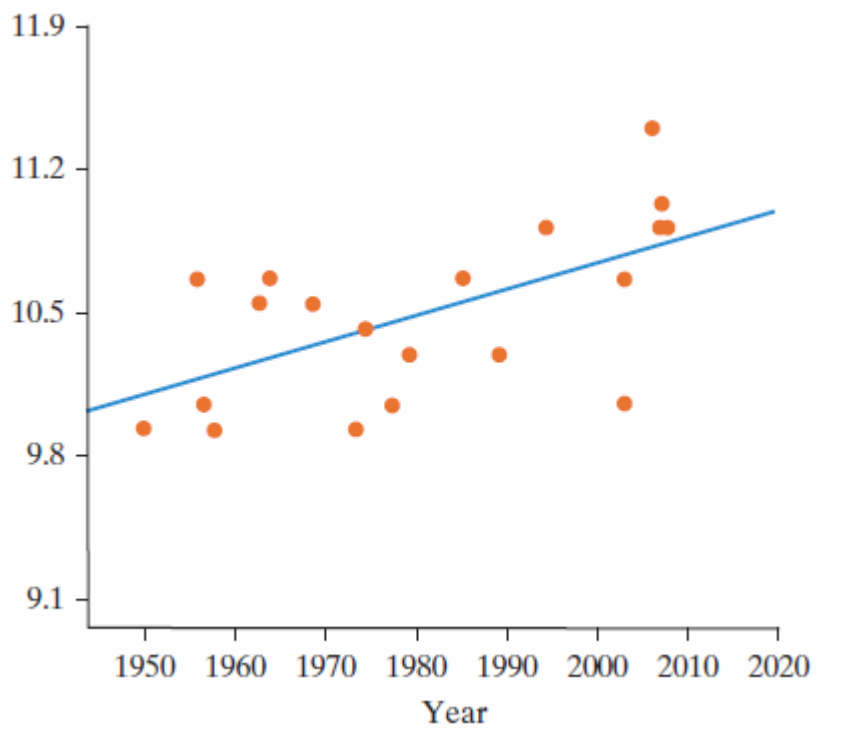
Read ch7 and 10.
The course website is http://www.stat.ucla.edu/~frederic/13/S26 .
HW4 is due Fri, Jun5, 1159pm. 10.1.8, 10.3.14, 10.3.21, and 10.4.11.

The final is Tue Jun9, 1130am-230pm, FOWLER A103b, and
will be on ch 1-7, 10, and 1 question on ch9.
Bring a PENCIL or pen and any books or notes you want.
No electronic devices allowed, not even a calculator.

If you cannot take it because of an emergency or other health reason,
then you will get an incomplete in the course and need to arrange to take the
Summer or Fall Stat 13 final.

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

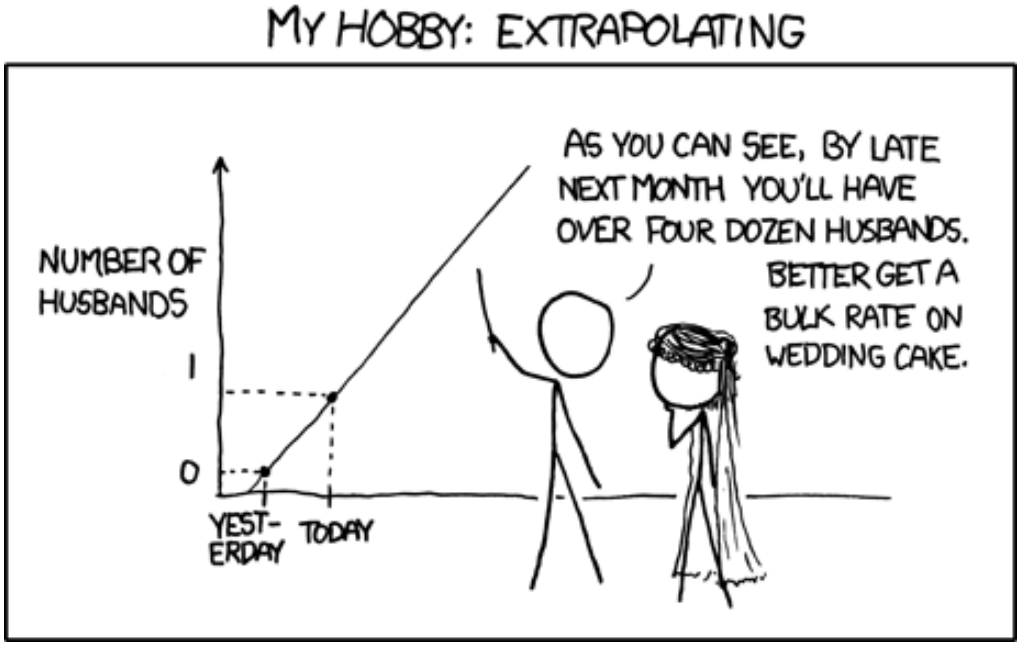
This is a good estimate of how much the regression predictions will typically be off by.

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.

2. 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

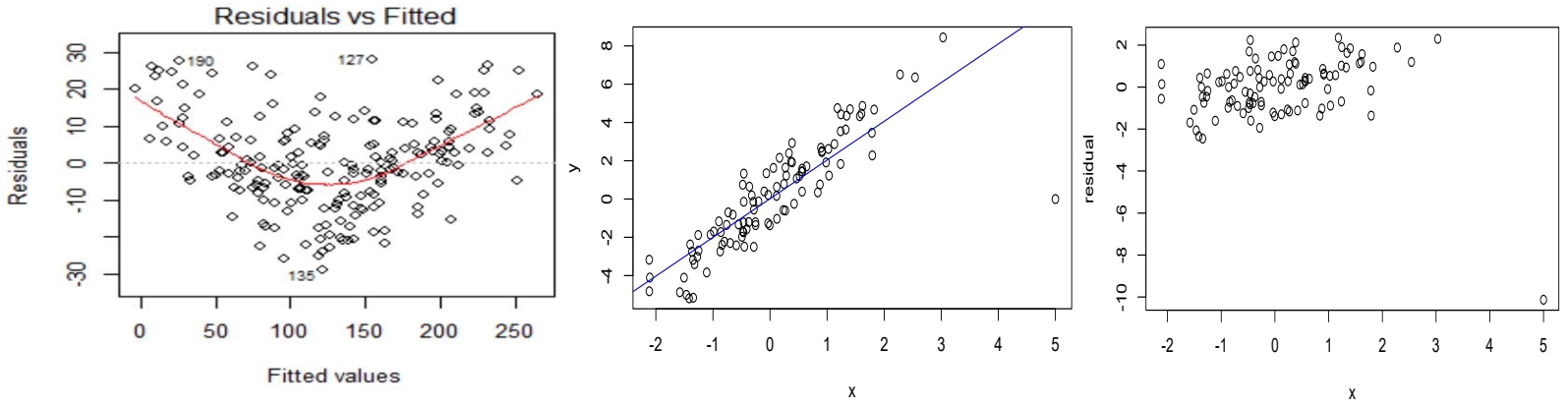
- Understand that one way a scatterplot can be summarized is by fitting the best-fit (least squares regression) line.
- Be able to interpret both the slope and intercept of a best-fit line in the context of the two variables on the scatterplot.
- Find the predicted value of the response variable for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

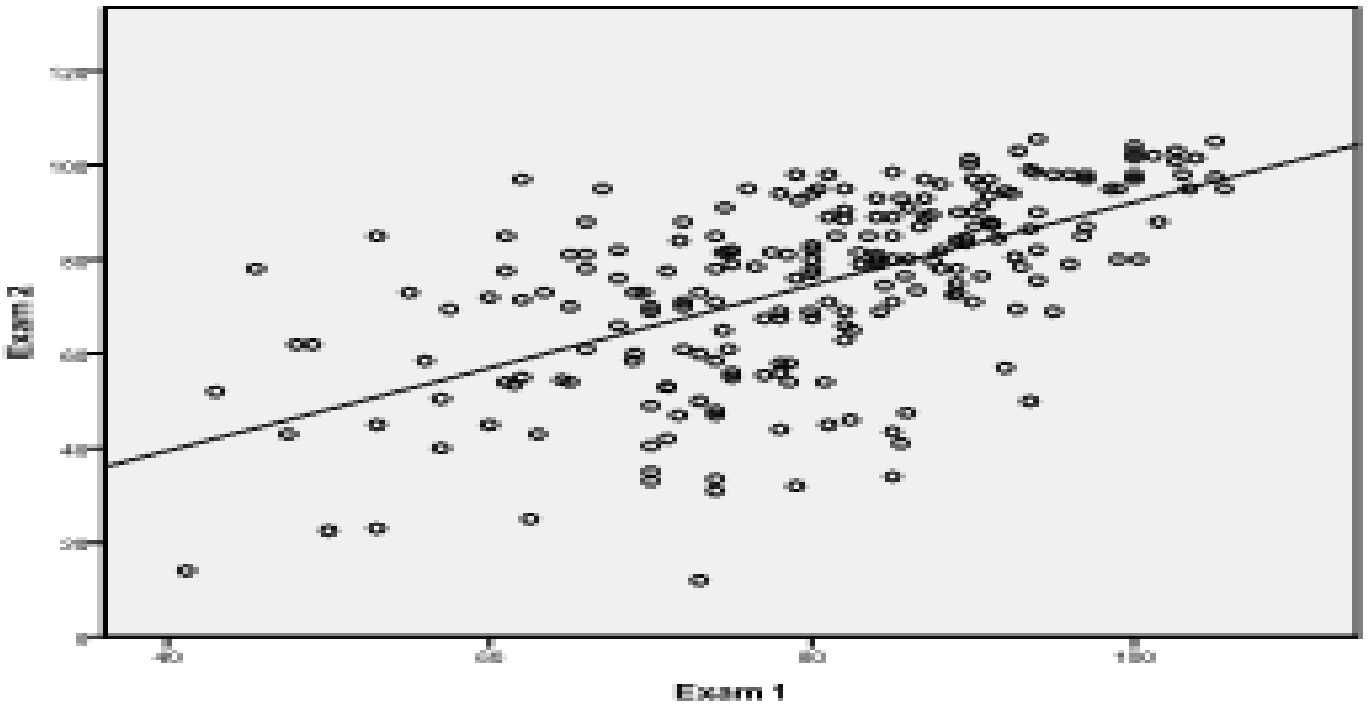
3. How well does the line fit?

- 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.

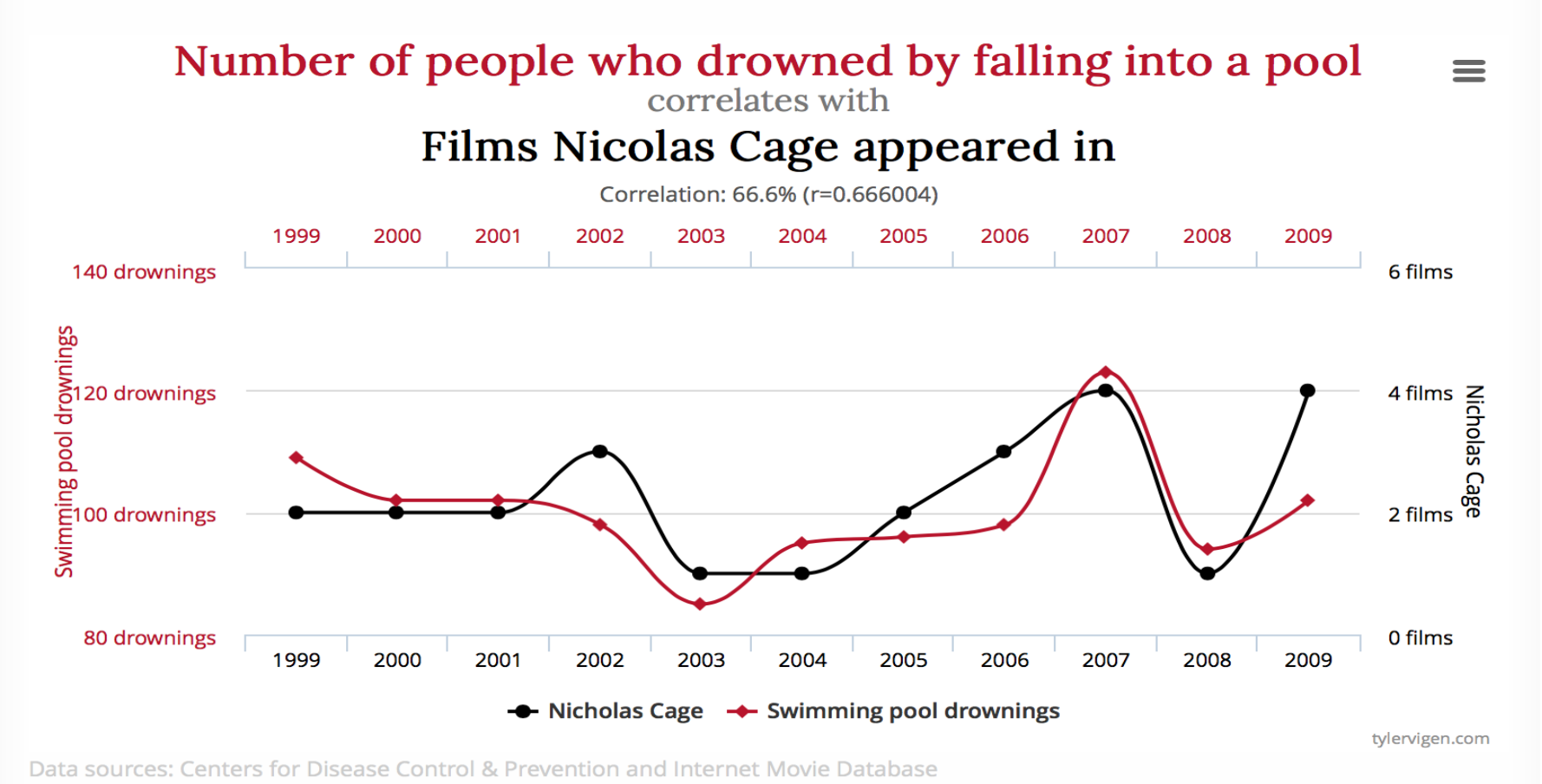


(b)

4. Common problems with regression.

- a. Correlation is not causation.

ESPECIALLY WITH OBSERVATIONAL DATA!



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.

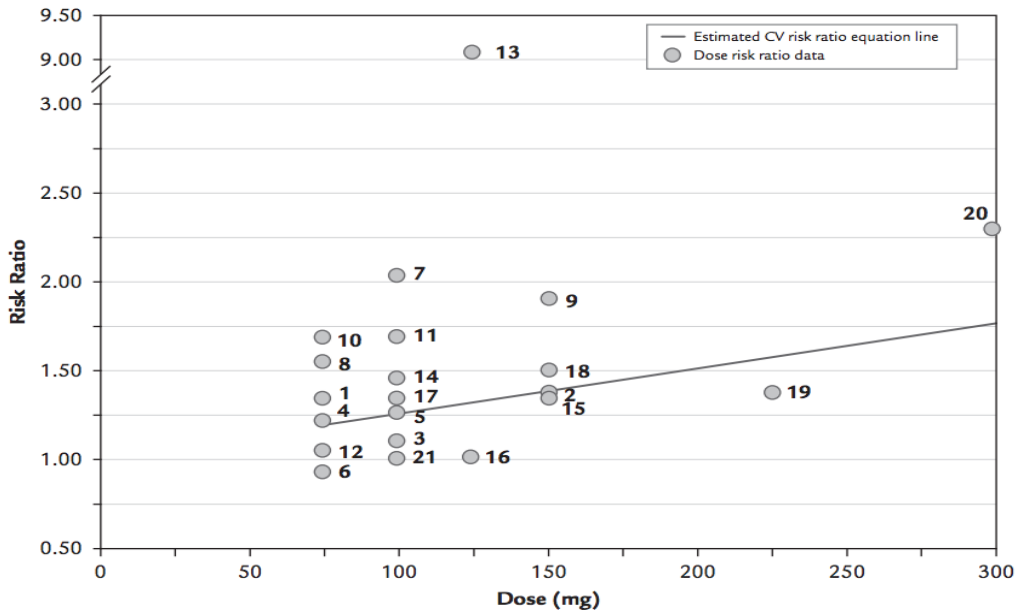


Figure 4. Relationship between diclofenac daily dose and the estimated risk ratio of a cardiovascular event. Numbers correspond to the observations in Table III.

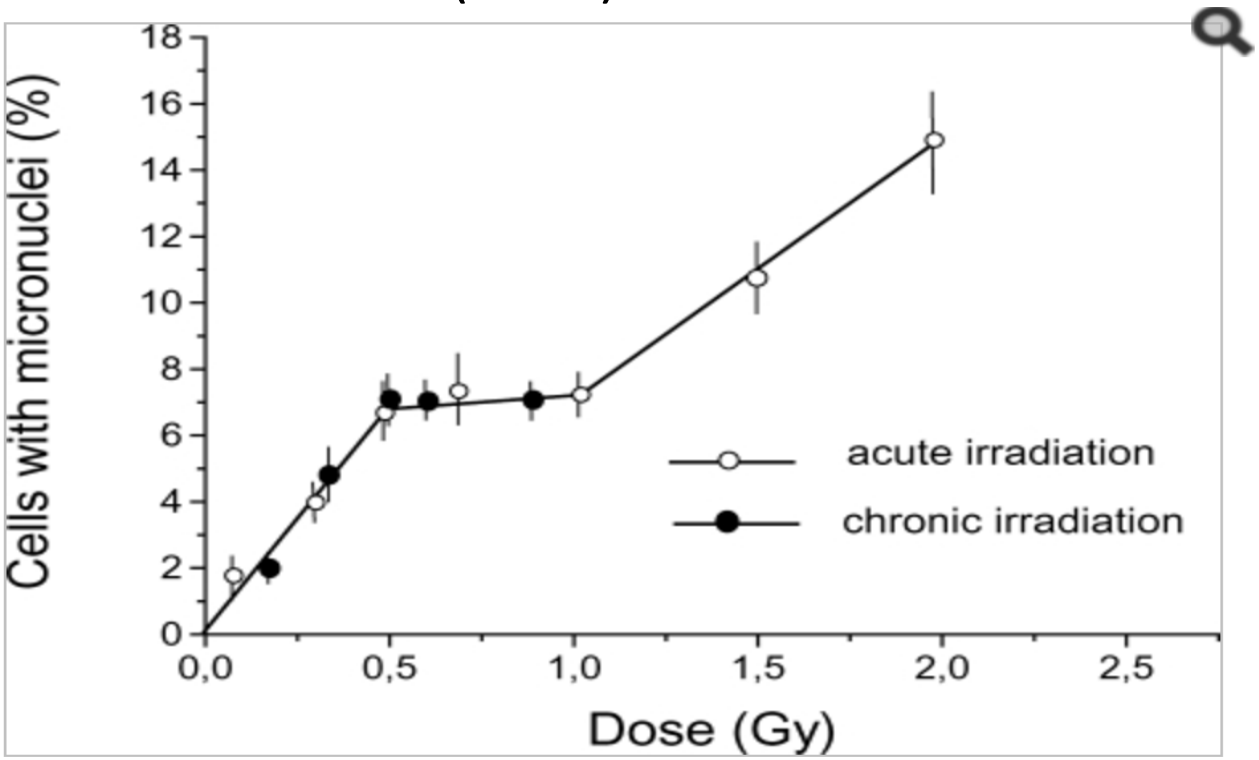
Common problems with regression.

- 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).

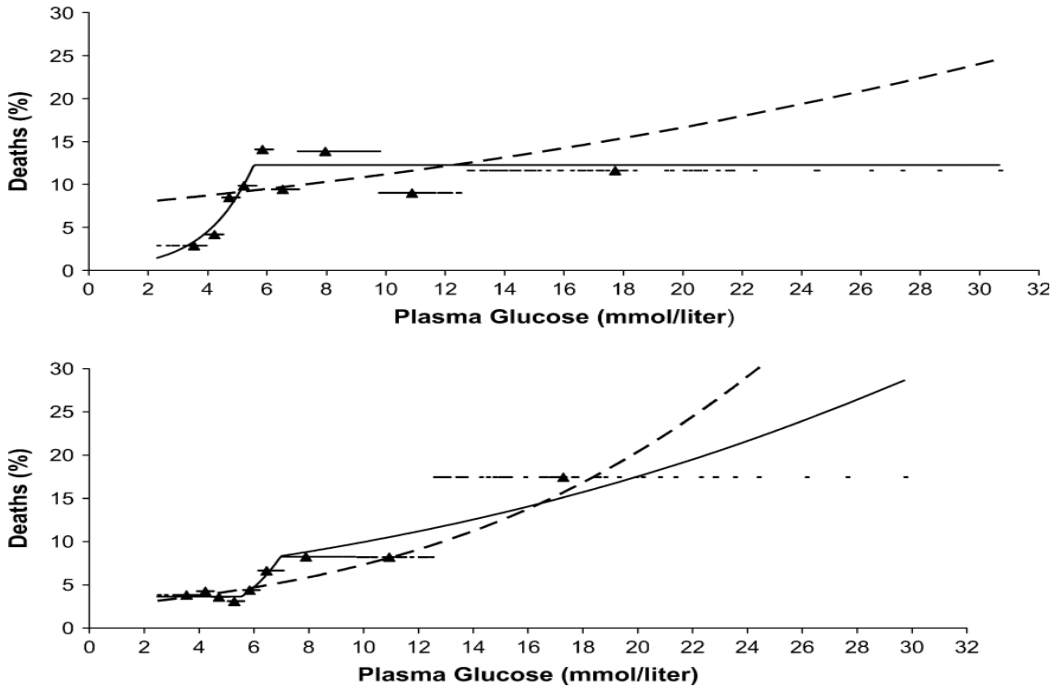
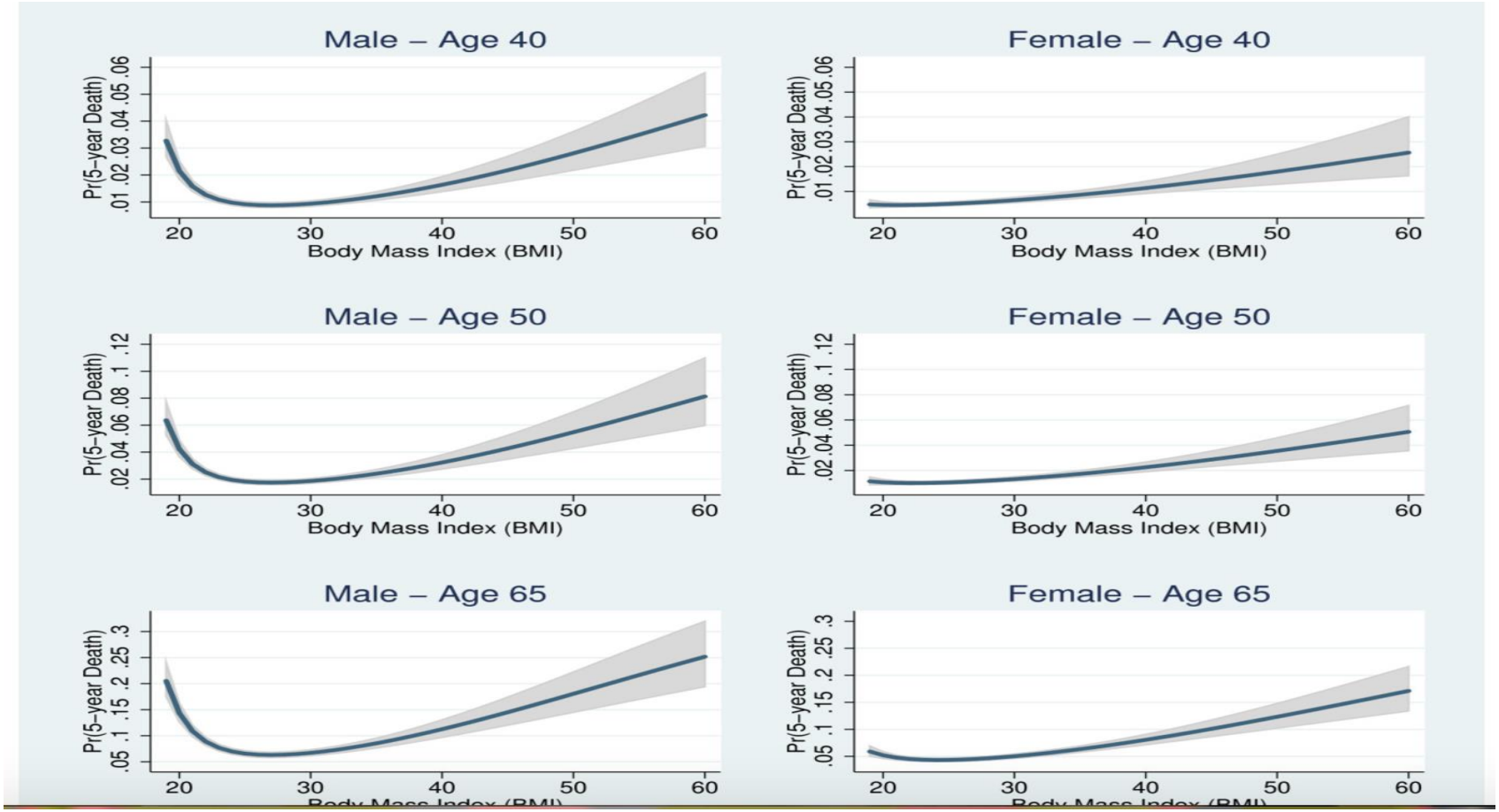


FIGURE 4. Adjusted 2-year rates of death from all causes for men (upper panel) and women (lower panel) separately, by glucose level, predicted by three models, Framingham Heart Study, 1948–1978. Linear model (dashed curve); optimal spline models (solid curve). The horizontal dashed

Common problems with regression.

- c. Curvature.

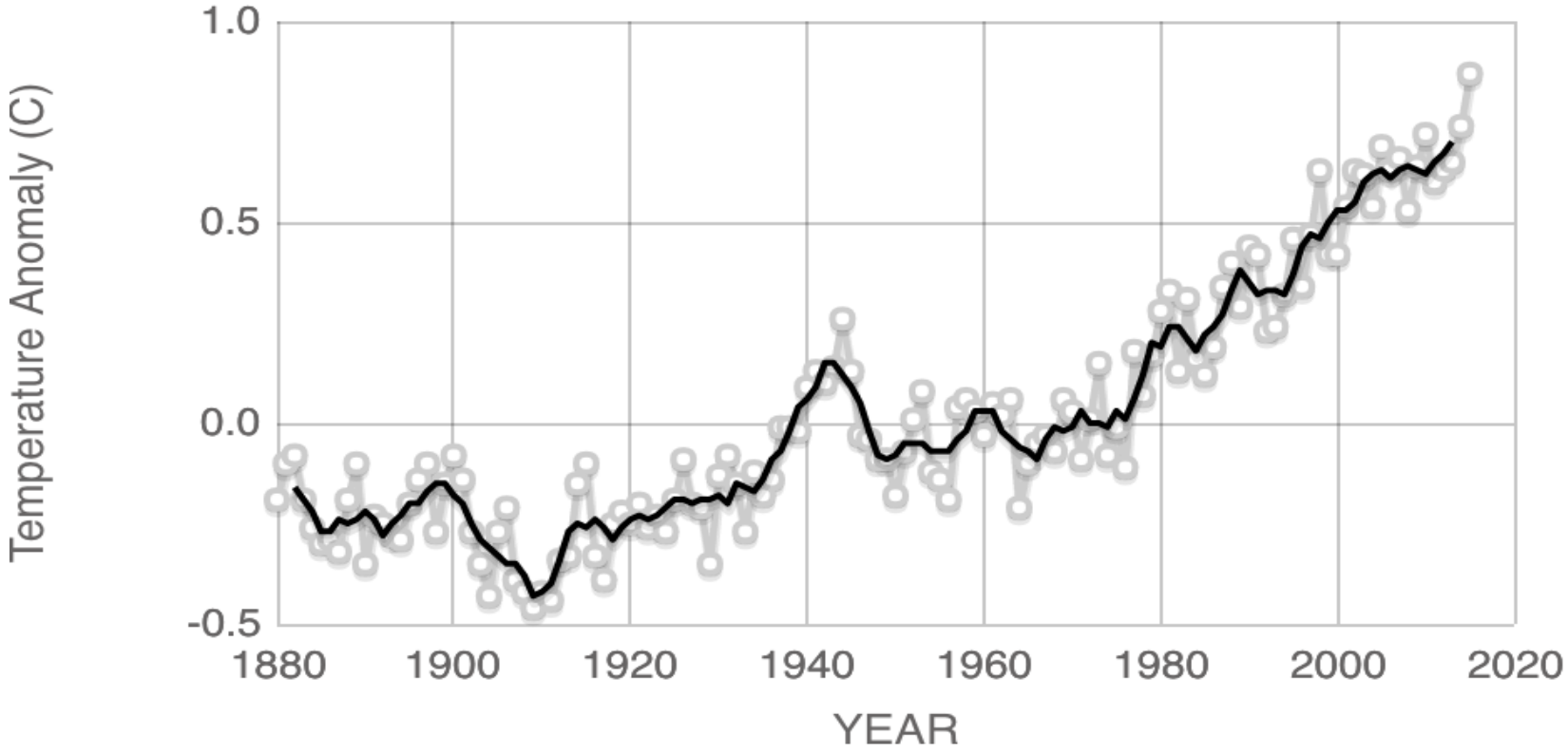
The best fitting line might fit poorly. Wong et al. (2011).



Common problems with regression.

- d. Statistical significance.

Could the observed correlation just be due to chance alone?



Source: climate.nasa.gov

5. Inference for the Regression Slope: Theory-Based Approach

Section 10.5

Do students who spend more time in non-academic activities tend to have lower GPAs?

Example 10.4

Do students who spend more time in non-academic activities tend to have lower GPAs?

- The subjects were 34 undergraduate students from the University of Minnesota.
- They were asked questions about how much time they spent in activities like work, watching TV, exercising, non-academic computer use, etc. as well as what their current GPA was.
- We are going to test to see if there is a significant negative association between the number of hours per week spent on nonacademic activities and GPA.

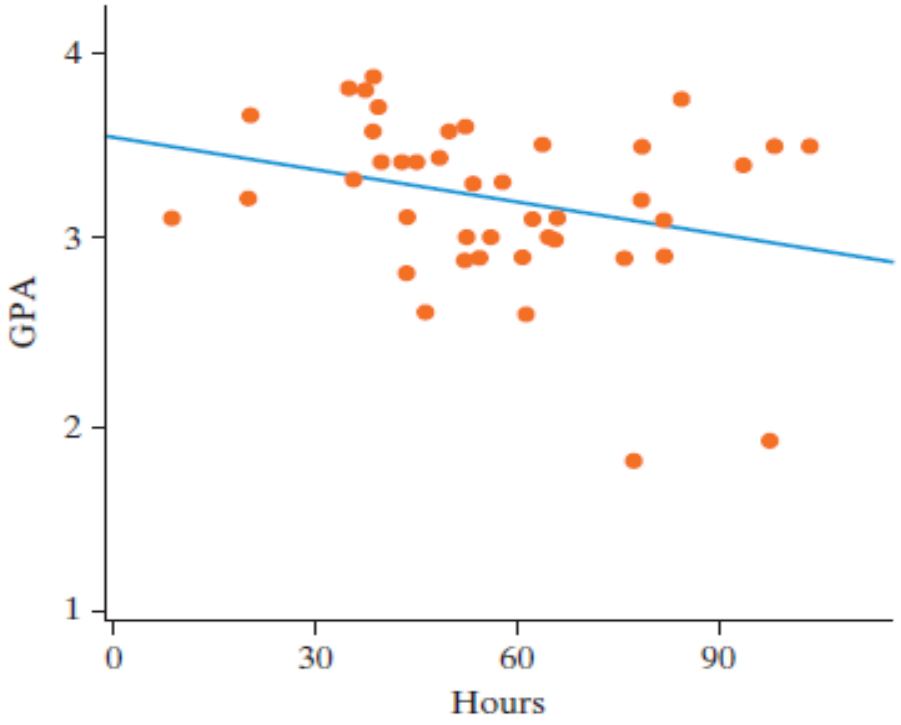
Hypotheses

- Null Hypothesis: There is no association between the number of hours students spend on nonacademic activities and student GPA in the population.

- Alternative Hypothesis: There is a negative association between the number of hours students spend on nonacademic activities and student GPA in the population.

Descriptive Statistics

- $\widehat{\text{GPA}} = 3.60 - 0.0059(\text{nonacademic hours}).$
- Is the slope significantly different from 0?



Shuffle to Develop Null Distribution

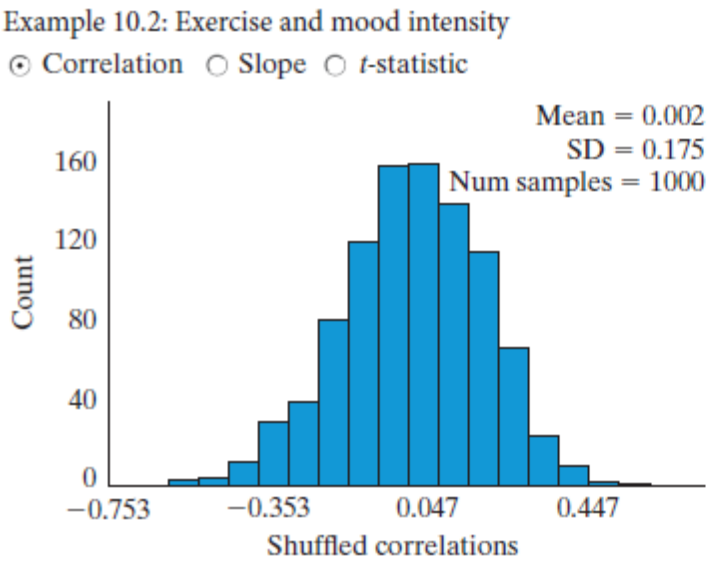
- We are going to shuffle just as we did with correlation to develop a null distribution.
- The only difference is that we will be calculating the slope each time and using that as our statistic.
- **a test of association based on slope is equivalent to a test of association based on a correlation coefficient.**

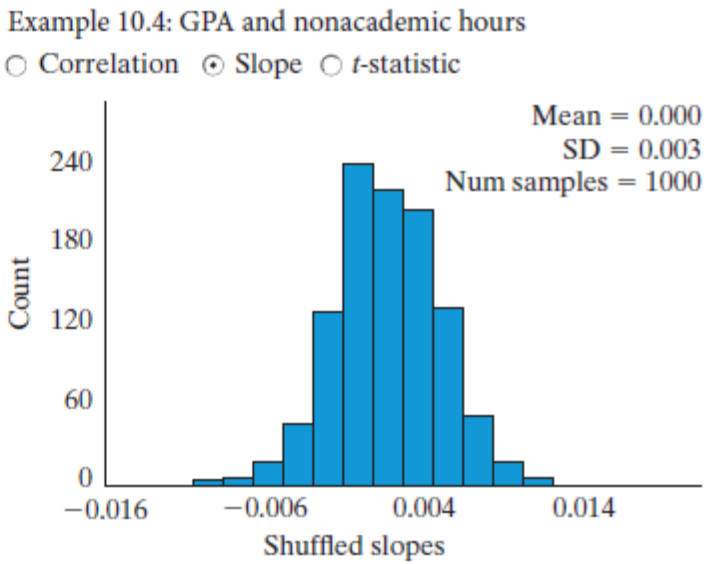
Beta vs Rho

- Testing the slope of the regression line is equivalent to testing the correlation (same p-value, but obviously different confidence intervals since the statistics are different)
- Hence these hypotheses are equivalent.
  - H_o: $\beta = 0$ H_a: $\beta < 0$ (Slope)
  - H_o: $\rho = 0$ H_a: $\rho < 0$ (Correlation)
- Sample slope (b) Population (β: beta)
- Sample correlation (r) Population (ρ: rho)
- When we do the theory based test, we will be using the *t*-statistic which can be calculated from either the slope or correlation.

Introduction

- Our null distributions are again bell-shaped and centered at 0 (for either correlation or slope as our statistic).





The book on p549 finds a p value of 3.3% by simulation.

Validity Conditions

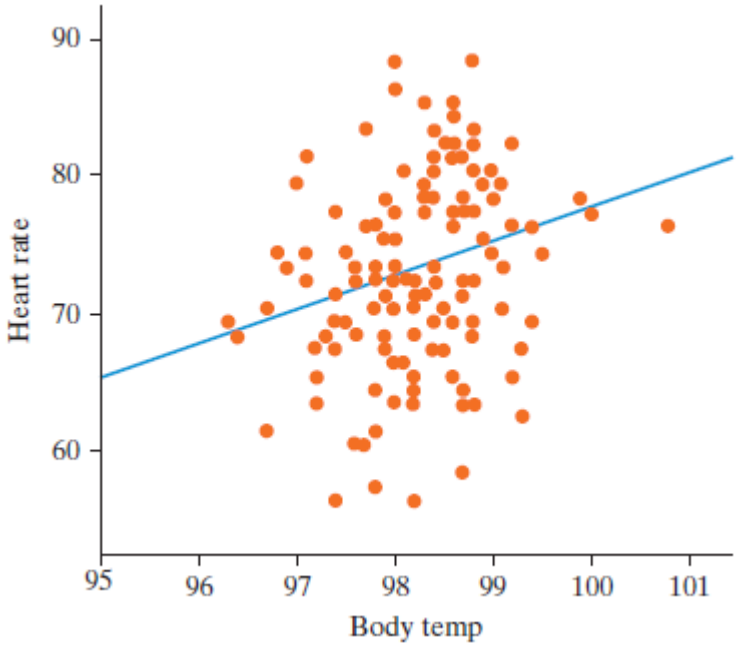
- Under the usual conditions: relationship is linear, observations are iid, both populations are normally distributed, and data are homoskedastic, theory-based inference for correlation or slope of the regression line uses the *t*-distribution.
- We could use simulations or the theory-based methods for the slope of the regression line.
- We would get exactly the same p-value if we used correlation as our statistic.

Predicting Heart Rate from Body Temperature

Example 10.5A

Heart Rate and Body Temp

- Earlier we looked at the relationship between heart rate and body temperature with 130 healthy adults
- $\text{Predicted Heart Rate} = -166.3 + 2.44(\text{Temp})$
- r = 0.257

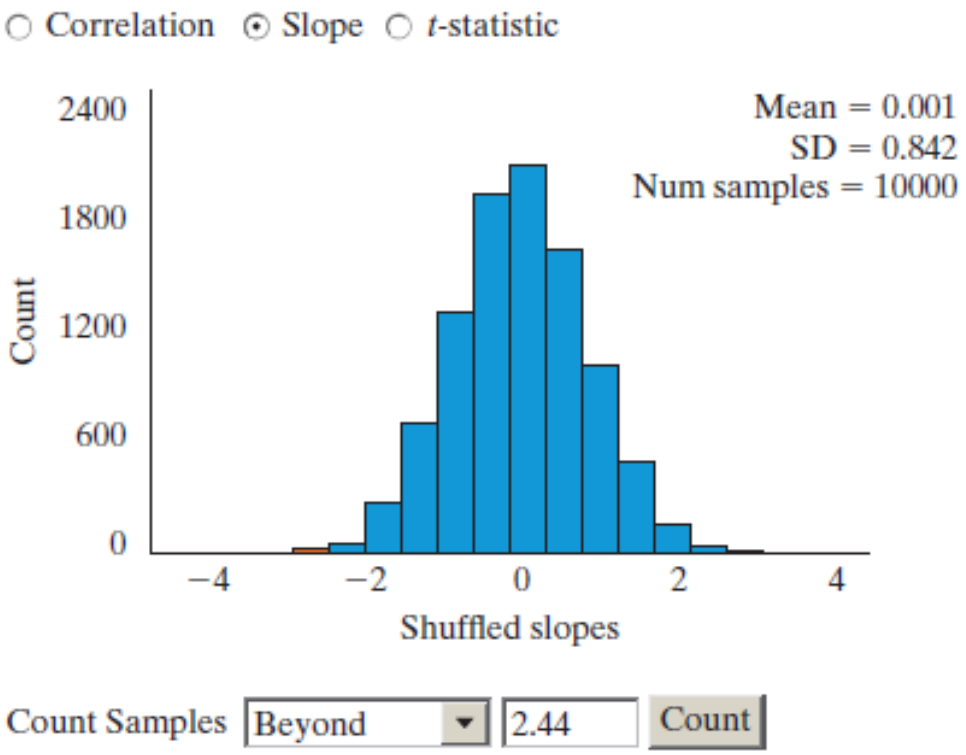


Heart Rate and Body Temp

- We tested to see if we had convincing evidence that there is a positive association between heart rate and body temperature in the population using a simulation-based approach. (We will make it 2-sided this time.)
- **Null Hypothesis**: There is no association between heart rate and body temperature in the population. $\beta = 0$
- **Alternative Hypothesis**: There is an association between heart rate and body temperature in the population. $\beta \neq 0$

Heart Rate and Body Temp

We get a very small p-value (0.0036). Anything as extreme as our observed slope of 2.44 happening by chance is very rare.



Heart Rate and Body Temp

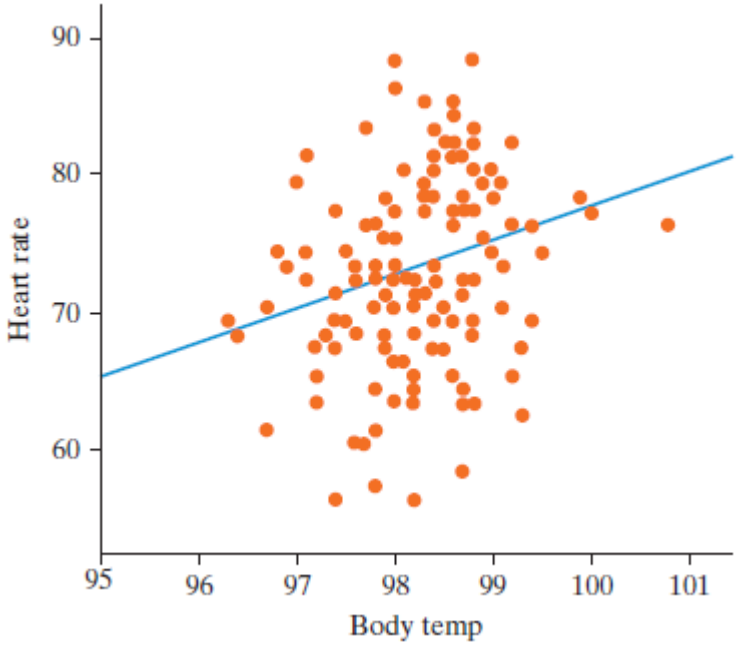
- We can also approximate a 95% confidence interval

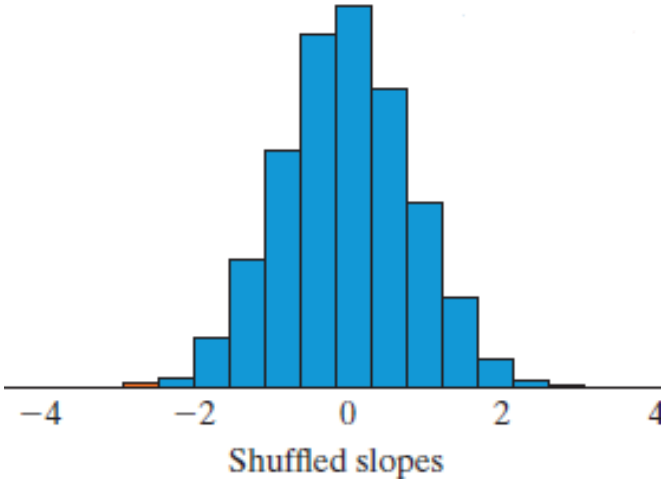
  observed statistic $\pm$ multiplier $\times$ SE

  $2.44 \pm 1.96 \times 0.842 =$ 0.790 to 4.09
- When both variables are normally distributed (scatterplot is elliptical), use the t-multiplier instead of 1.96, but when n is large it makes very little difference.
- This means we are 95% confident that, in the population of healthy adults, each 1° increase in body temp is associated with an increase in heart rate of between 0.790 to 4.09 beats per minute.

Heart Rate and Body Temp

- The theory-based approach should work well since the distribution of the slopes has a nice bell shape
- Also check the scatterplot





Heart Rate and Body Temp

- We will use the t-statistic to get our theory-based p-value.
- We will find a theory-based confidence interval for the slope.
- On p554, the book notes the formula $t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$.
- Here the t statistic is 2.97.
- The p-value is 0.36%. So the correlation is statistically significantly greater than zero.

Smoking and Drinking

Example 10.5B

Validity Conditions

Remember our validity conditions for theory-based inference for slope of the regression equation.

1. The scatterplot should follow a linear trend.

2. The observations should be iid.

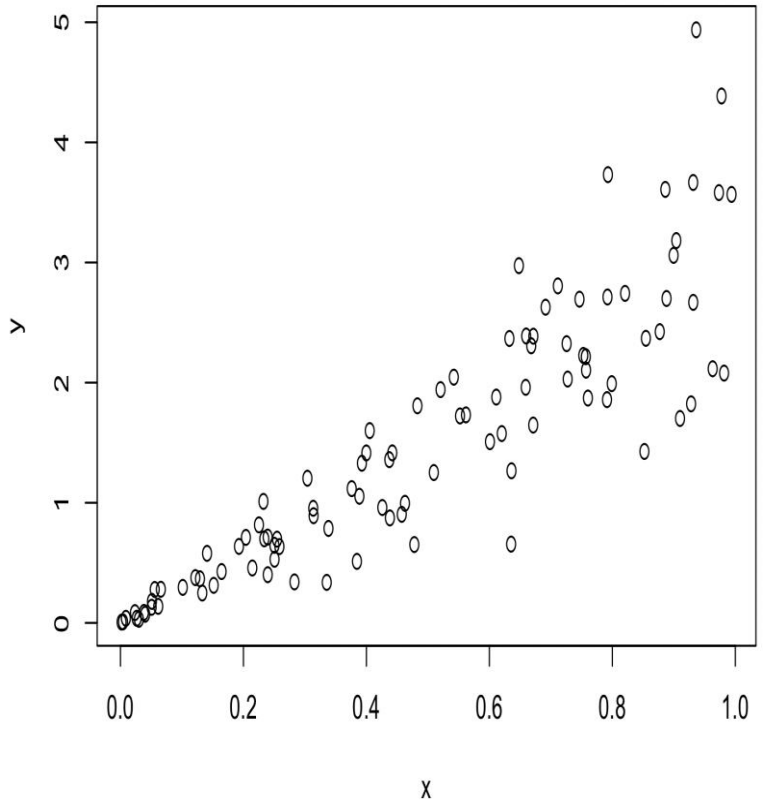
3. For the t-test, both variables should be normal.

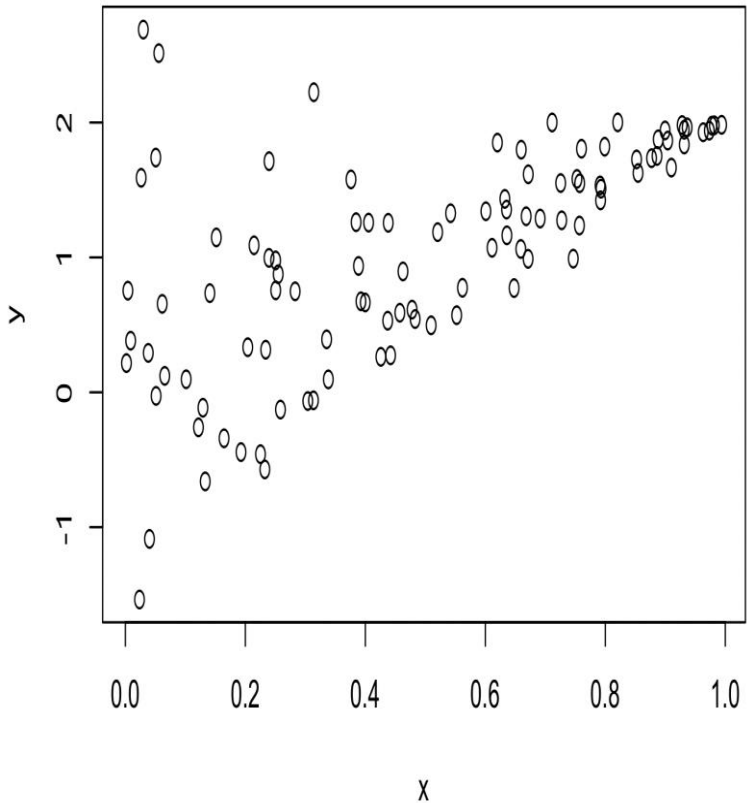
In particular, there should be approximately the same number of points above and below the regression line (symmetry).

4. The variability of vertical slices of the points should be similar. This is called homoskedasticity.

Validity Conditions

- Let's look at some scatterplots that do not meet the requirements.



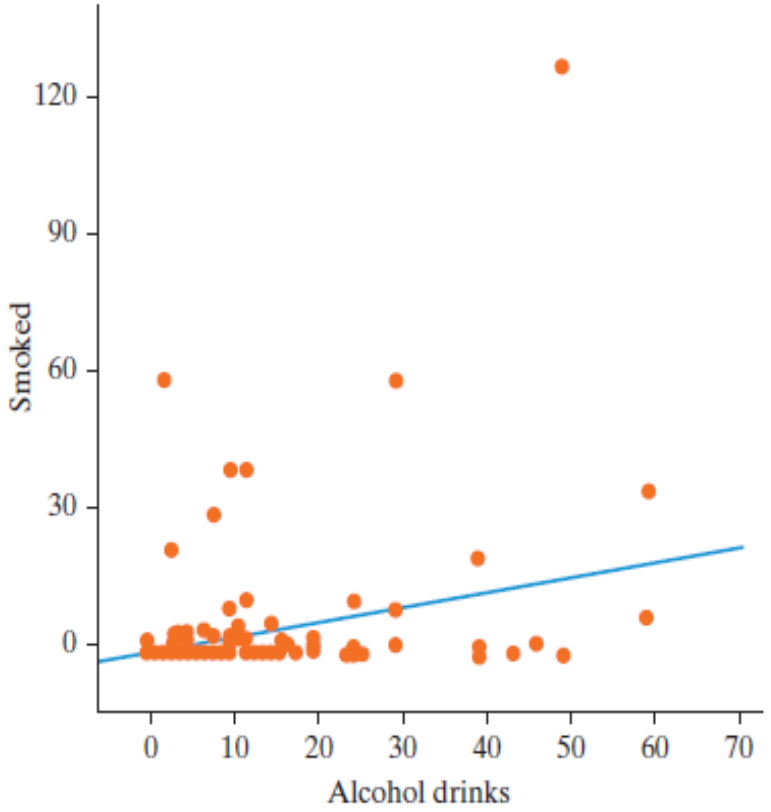


Smoking and Drinking

The relationship between number of drinks and cigarettes per week for a random sample of students at Hope College.

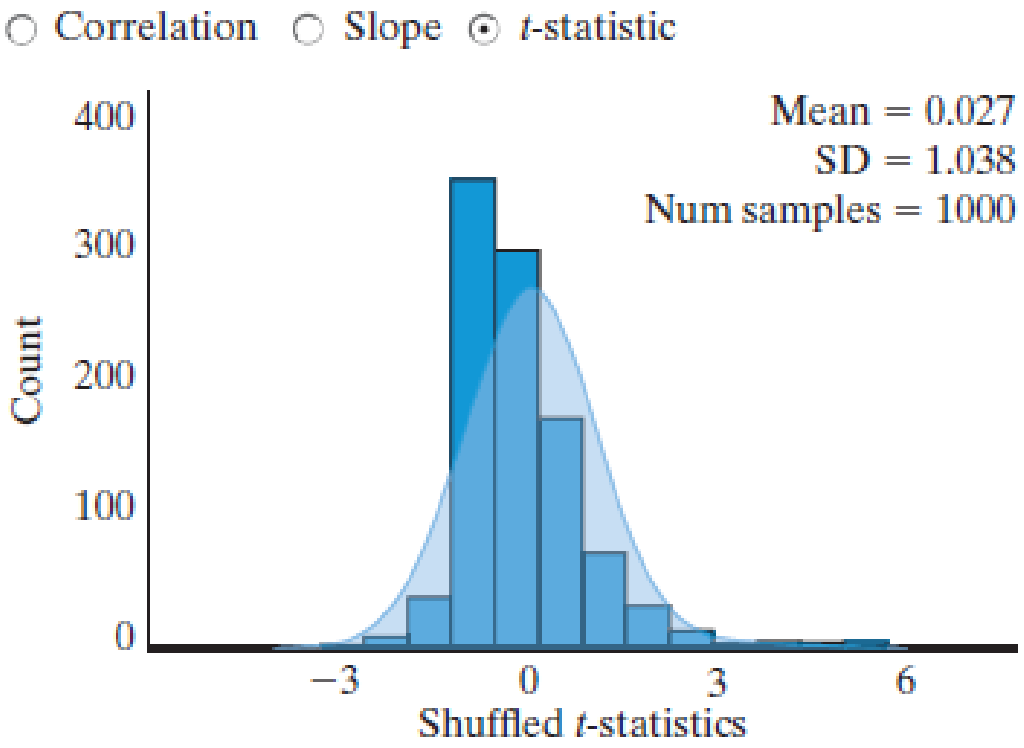
The dot at (0,0) represents 524 students

Are the conditions met? Hard to say. The book says no.



Smoking and Drinking

- When the conditions are not met, applying simulation-based inference is preferable to theory-based t-tests and CIs.



Validity Conditions

- What do you do when validity conditions aren't met for theory-based inference?
  - Use the simulated-based approach.
- Another strategy is to "transform" the data on a different scale so conditions are met.
  - The logarithmic scale is common.
- One can also fit a different curve, not necessarily a line.