Linear Methods for Regression

Victoria Stodden

Statistical Learning

Examples of learning problems (from ESL p. 1):

- Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack. The prediction is to be based on demographic, diet and clinical measurements for that patient.

- Predict the price of a stock in 6 months from now, on the basis of company performance measures and economic data.

- Identify the numbers in a handwritten ZIP code, from a digitized image.

- Estimate the amount of glucose in the blood of a diabetic person, from the infrared absorption spectrum of that person's blood.

- Identify the risk factors for prostate cancer, based on clinical and demographic variables.

Supervised Learning

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.007	0.43	0.11	0.18	0.42	0.29	0.01

TABLE 1.1. (ESL p. 2) Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between **spam** and **email**.

Supervised Learning

- input: pre-set or measured variable, a predictor or independent variable. Usually denoted *X.*

- output: what we would like to predict, and influenced by the input variables. Also called response, predicted or dependent variable. Usually denoted *Y*.

- Recall: different types of variables: qualitative, quantitative.

- *Regression* is a techniques that uses inputs to predict quantitative output.

- *Classification* predicts qualitative output.

Linear Regression: Overview

2.3.1 *Linear Models and Least Squares*

The linear model has been a mainstay of statistics for the past 30 years and remains one of our most important tools. Given a vector of inputs $X^T = (X_1, X_2, \ldots, X_p)$, we predict the output Y via the model

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j. \tag{2.1}$$

The term $\hat{\beta}_0$ is the intercept, also known as the *bias* in machine learning. Often it is convenient to include the constant variable 1 in X, include $\hat{\beta}_0$ in the vector of coefficients $\hat{\beta}$, and then write the linear model in vector form as an inner product

$$\hat{Y} = X^T \hat{\beta}, \tag{2.2}$$

Fitting the Linear Model

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2. \tag{2.3}$$

$\mathrm{RSS}(\beta)$ is a quadratic function of the parameters, and hence its minimum always exists, but may not be unique. The solution is easiest to characterize in matrix notation. We can write

$$\mathrm{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta), \tag{2.4}$$

where $\mathbf{X}$ is an $N \times p$ matrix with each row an input vector, and $\mathbf{y}$ is an N-vector of the outputs in the training set. Differentiating w.r.t. β we get the *normal equations*

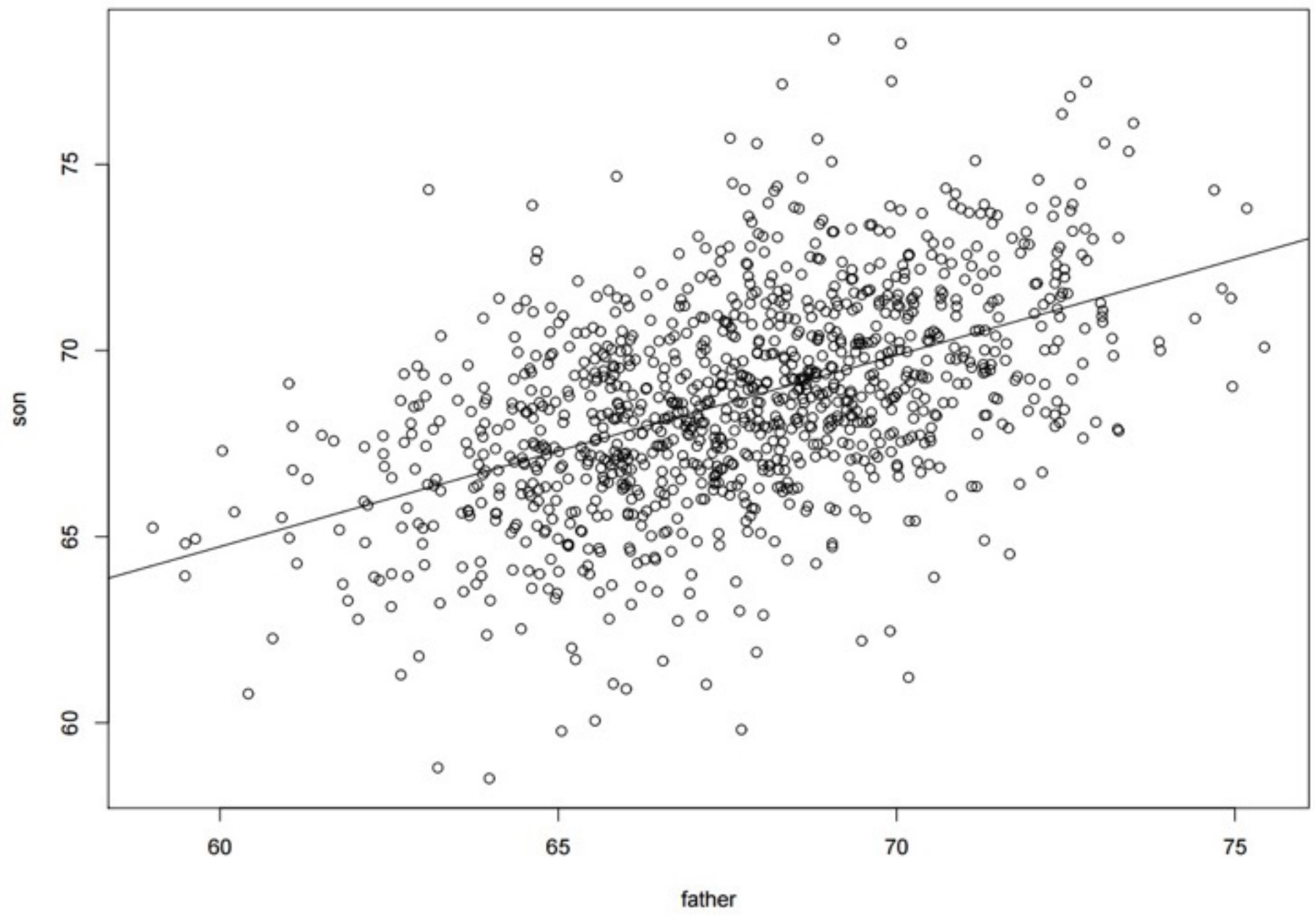
$$\mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta) = 0. \tag{2.5}$$

If $\mathbf{X}^T\mathbf{X}$ is nonsingular, then the unique solution is given by

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \tag{2.6}$$

and the fitted value at the ith input x_i is $\hat{y}_i = \hat{y}(x_i) = x_i^T \hat{\beta}$. At an arbitrary input x_0 the prediction is $\hat{y}(x_0) = x_0^T \hat{\beta}$.

Example: Predicting son's height from the father's height



credit: Kathryn Roeder

Building the Foundations

1. Randomness in Data
2. Distributions of Random Variables
3. Hypothesis Testing

We will take what we need from these topics, but note our coverage won't be comprehensive. The Supplemental Reading OS3 may be helpful (chapters 3,4, and 7).

Randomness in Data

Example: Two books are assigned for a statistics class: a textbook and a study guide. The university bookstore determined 20% of enrolled students do not buy either book, 55% buy the textbook only, and 25% buy both books, and these percentages are relatively constant from one term to another. If there are 100 students enrolled, how many books should the bookstore expect to sell to this class?

Around 20 students will not buy either book (0 books total), about 55 will buy one book (55 books total), and approximately 25 will buy two books (totaling 50 books for these 25 students). The bookstore should expect to sell about 105 books for this class.

Question: Would you be surprised if the bookstore sold slightly more or less than 105 books?

Random Variables

Such randomness or variability in observed data is expressed using the following paradigm:

The data we observe are drawn *noisily* or *with error* from some underlying (and unobserved) true *distribution*.

The bakes error into our modeling setup, since it appears right in our assumption about data generation.

A measurement drawn with error from an underlying distribution is called a *random variable.*

Random Variables 2

More generally:

Random variable: A random process or variable with a numerical outcome.

Random variables are usually labeled X or Y, and their actual data values, the draws from the distribution, labeled x or y.

Random variables can be *discrete* or *continuous*.

Here's how to tell the difference. Can the variable take on only discrete values, such as integers, or can it take on any value, such as the reals? (See e.g. http://www.mathsisfun.com/sets/number-types.html)

Expected Value

Expected value of a Discrete Random Variable

If X takes outcomes x1, ..., xk with probabilities P(X = x1), ..., P(X = xk), the expected value of X is the sum of each outcome multiplied by its corresponding probability:

$$E(X) = x1*P(X = x1)+\cdots+xk*P(X = xk)$$

The Greek letter μ may be used in place of the notation E(X).

Remember that for any observed value of X, such as x, it is not random (for E(x) doesn't make sense).

Variability of Random Values

If X takes outcomes x1, ..., xk with probabilities P(X = x1), ..., P(X = xk) and expected value μ = E(X), then the variance of X, denoted by Var(X) or the symbol σ^2, is

$$\sigma^2 = (x1 - \mu)^2 * P(X = x1) + \cdots + (xk - \mu)^2 * P(X = xk)$$

Linear Combinations of Random Variables

A linear combination of two random variables X and Y describes a combination aX + bY, for some constants a and b.

Note that E(aX + bY) = a*E(X) + b*E(Y) and

Var(aX+bY)=a2*Var(X) + b2*Var(Y)

More on Distributions

- We can imagine a random variable X takes on the value x with some probability.

- This probability won't be the same for all x, some are typically more likely than others.

- Imagine that X can only take on values that fall within some range, like heights for example. Heights won't be negative nor will they be above, say, 20 feet.

- Some heights are much more like than other. 5'5'' is much more likely than 6'6''.

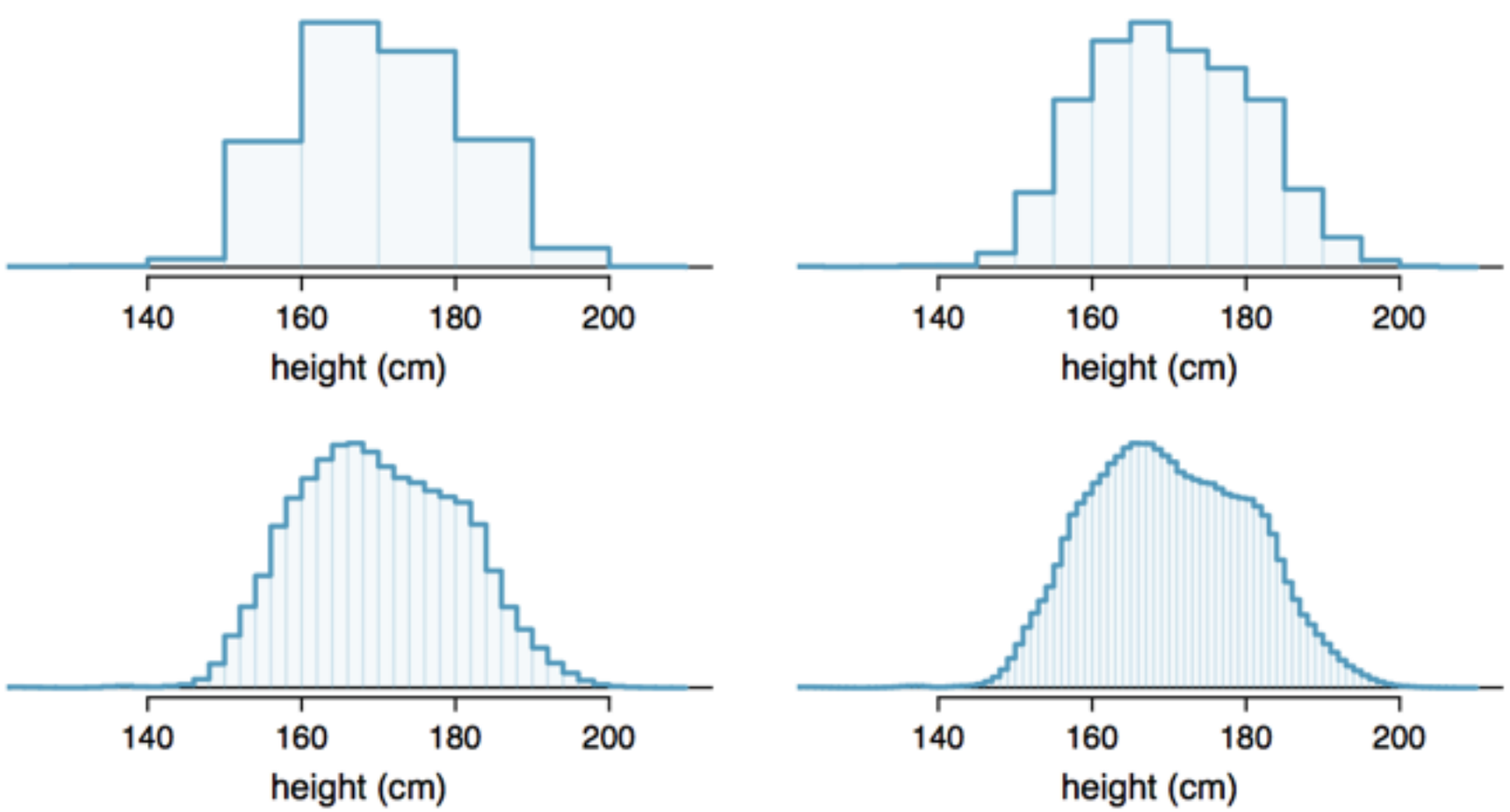


Figure 2.26: Four hollow histograms of US adults heights with varying bin widths.

Example 2.88 What proportion of the sample is between 180 cm and 185 cm tall (about 5'11" to 6'1")?

We can add up the heights of the bins in the range 180 cm and 185 and divide by the sample size. For instance, this can be done with the two shaded bins shown in Figure 2.27. The two bins in this region have counts of 195,307 and 156,239 people, resulting in the following estimate of the probability:

$$\frac{195307 + 156239}{3{,}000{,}000} = 0.1172$$

This fraction is the same as the proportion of the histogram's area that falls in the range 180 to 185 cm.

The Normal Distribution

- There several distributions that characterize commonly observed histograms that they have names.

- The more common is the Normal Distribution.

- Many variables are nearly normal, but none are exactly normal. Thus the normal distribution, while not perfect for any single problem, is very useful for a variety of problems. We will use it in data exploration and to solve important problems in statistics.

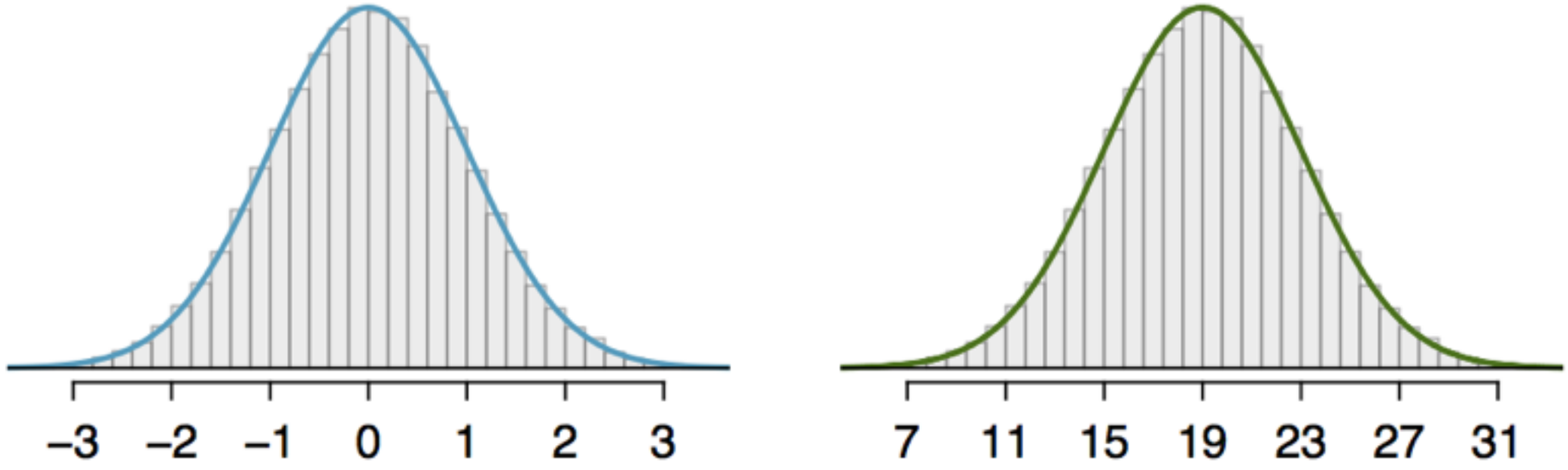


Figure 3.2: Both curves represent the normal distribution, however, they differ in their center and spread. The normal distribution with mean 0 and standard deviation 1 is called the **standard normal distribution**.

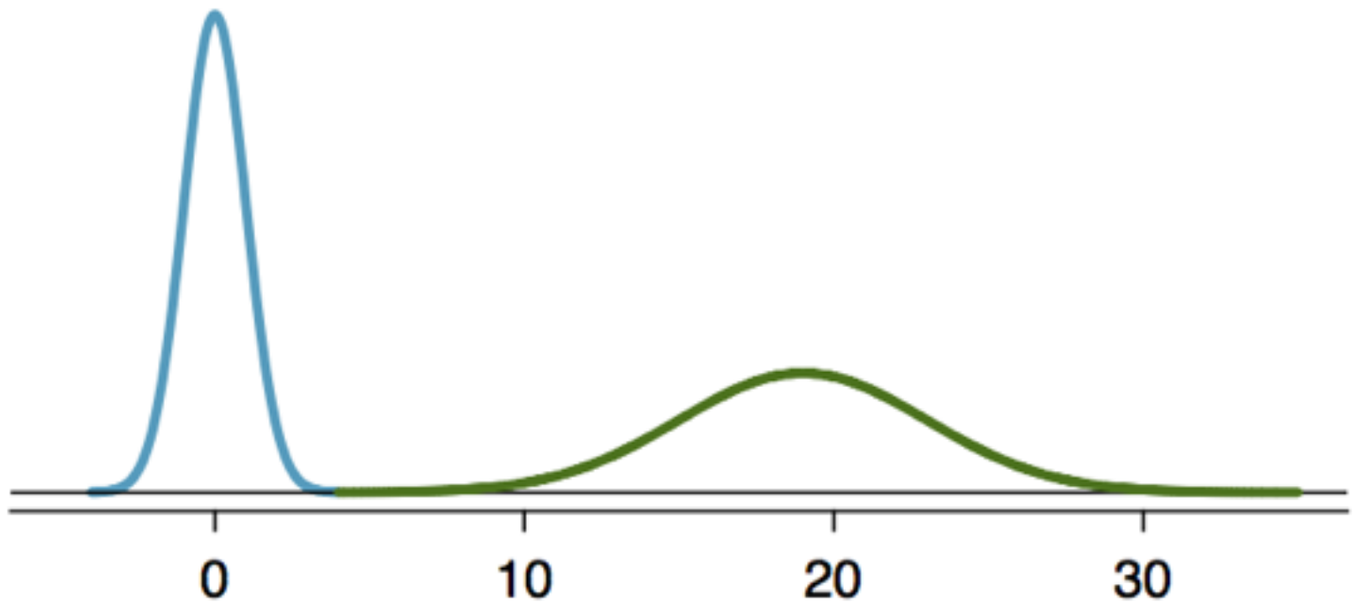


Figure 3.3: The normal models shown in Figure 3.2 but plotted together and on the same scale.

Parametrizing the Normal Distribution

If a normal distribution has mean μ and standard deviation σ, we may write the distribution as N(μ,σ). The two distributions in Figure 3.3 can be written as

$$N(\mu=0,\sigma=1) \text{ and } N(\mu=19,\sigma=4)$$

Because the mean and standard deviation describe a normal distribution exactly, they are called the distribution's parameters.

We can *standardize* all normally distributed data to be from N(0,1) via the transformation $Z = (x - \mu) / \sigma$.

The 68-95-99.7 Rule

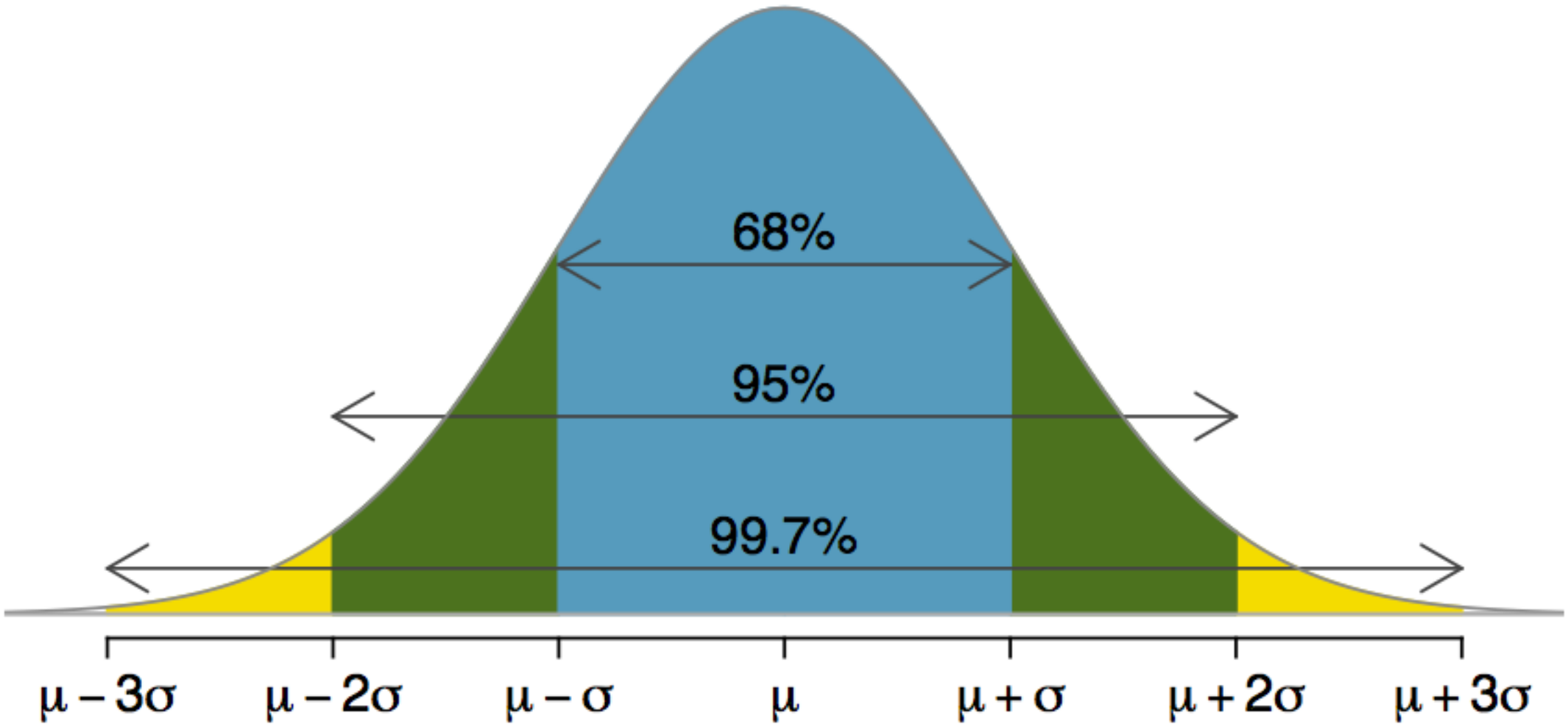


Figure 3.9: Probabilities for falling within 1, 2, and 3 standard deviations of the mean in a normal distribution.

Point Estimates

- An intuitive way to estimate a characteristic of the population under study, for example its mean, is to calculate the sample mean (i.e the mean of the observed data).

- Such as calculated mean is called a *point estimate* of the true mean.

- Estimates generally vary from one sample to another, and *sampling variation* the point estimate may be close, but it will not be exactly equal to the parameter.

- We can quantify this estimation error by finding the *standard error* of an estimate.

Confidence Intervals

SE(estimate of the mean) = σ*(standard deviation of the individual observations) / $\sqrt{n}$

Confidence Interval = point estimate ± 2 × SE

Note that different point estimate will have different calculations of SE.

This means we can define a *sampling distribution* for point estimate. Suppose you draw many samples from the population distribution and calculate the point estimate for the mean each time. They will vary, and so trace out an empirical distribution.

Hypothesis Testing

The null hypothesis (H0) often represents either a skeptical perspective or a claim to be tested. The alternative hypothesis (HA) represents an alternative claim under consideration and is often represented by a range of possible parameter values.

H0: The average days per week that UIUC students lifted weights was the same for 2011 and 2013.

HA: The average days per week that UIUC students lifted weights was different for 2013 than in 2011.

Hypothesis Testing and Confidence Intervals

- The difference in H0 and HA could be due to sampling variation.

- As we just saw, confidence intervals are a way to find a range of plausible values for the population mean.

Example 4.19 In the sample of 100 students from the 2013 YRBSS survey, the average number of days per week that students lifted weights was 2.78 days with a standard deviation of 2.56 days (coincidentally the same as days active). Compute a 95% confidence interval for the average for all students from the 2013 YRBSS survey. You can assume the conditions for the normal model are met.

The general formula for the confidence interval based on the normal distribution is

$$\bar{x} \pm z^\star SE_{\bar{x}}$$

We are given $\bar{x}_{13} = 2.78$, we use $z^\star = 1.96$ for a 95% confidence level, and we can compute the standard error using the standard deviation divided by the square root of the sample size:

$$SE_{\bar{x}} = \frac{s_{13}}{\sqrt{n}} = \frac{2.56}{\sqrt{100}} = 0.256$$

Entering the sample mean, $z^\star$, and the standard error into the confidence interval formula results in (2.27, 3.29). We are 95% confident that the average number of days per week that all students from the 2013 YRBSS lifted weights was between 2.27 and 3.29 days.

Because the average of all students from the 2011 YRBSS survey is 3.09, which falls within the range of plausible values from the confidence interval, we cannot say the null hypothesis is implausible. That is, we fail to reject the null hypothesis, H_0.

Decision Errors

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	okay	Type 1 Error
	H_A true	Type 2 Error	okay

Linear Regression Model

2 variables whose relationship can be modeled as a linear combination:

$$y = \beta 0 + \beta 1 x$$

This says that the true underlying model that is generating the data we observe, has that specific mathematical form.

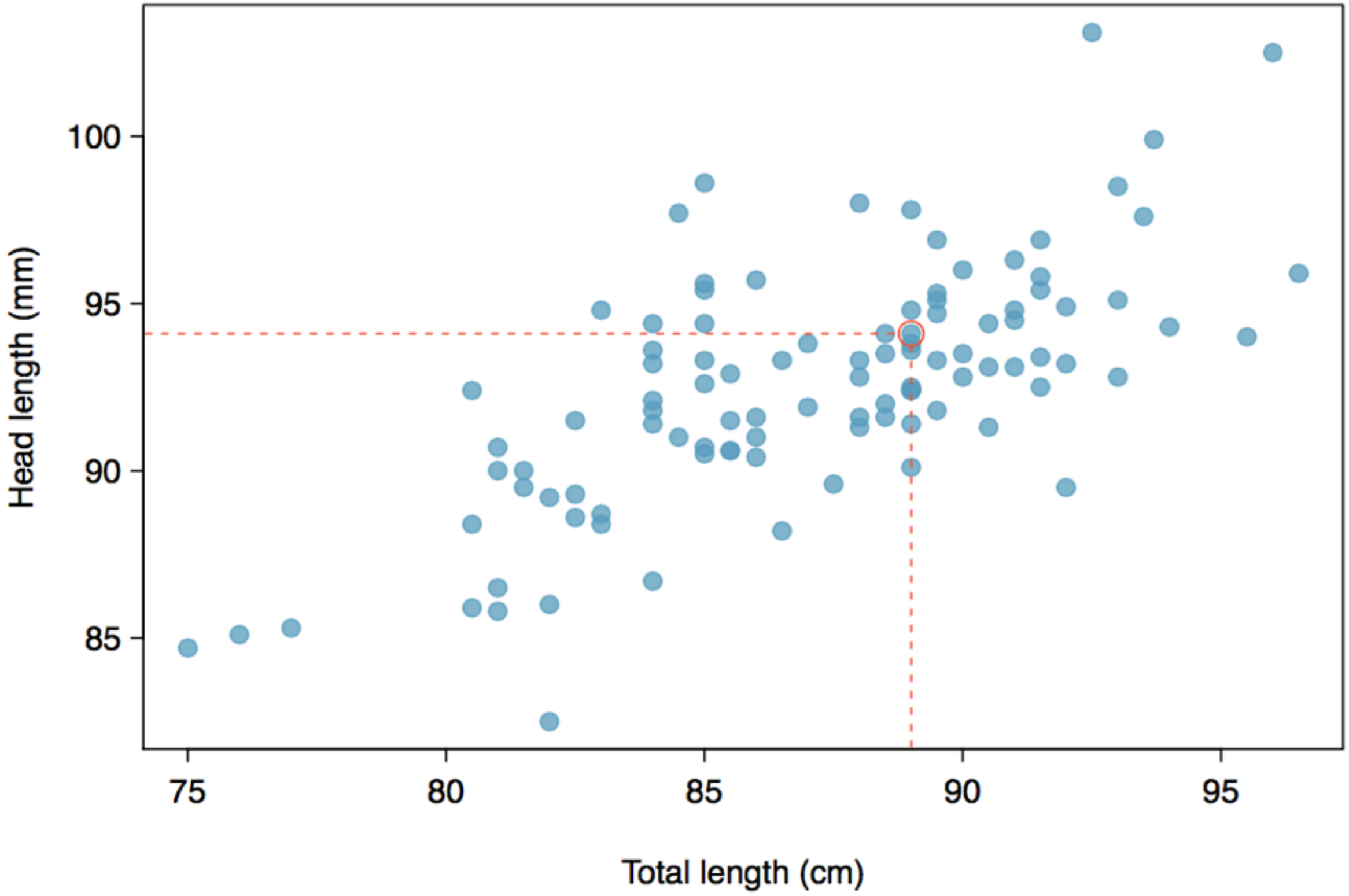


Figure 7.4: A scatterplot showing head length against total length for 104 brushtail possums. A point representing a possum with head length 94.1mm and total length 89cm is highlighted.

Residuals

How well does our guess at the true underlying model actually fit the data?

Residuals are the leftover variation in the data after accounting for the model fit: Data = Fit + Residual

Each observation will have a residual. If an observation is above the regression line, then its residual, the vertical distance from the observation to the line, is positive. Observations below the line have negative residuals. One goal in picking the right linear model is for these residuals to be as small as possible.

Residuals

- The residual of the ith observation (xi, yi) is the difference of the observed response (yi) and the response we would predict based on the model fit (yi):

- $e_i = y_i - \hat{y}_i$
  We typically identify $\hat{y}_i$ by plugging xi into the model.

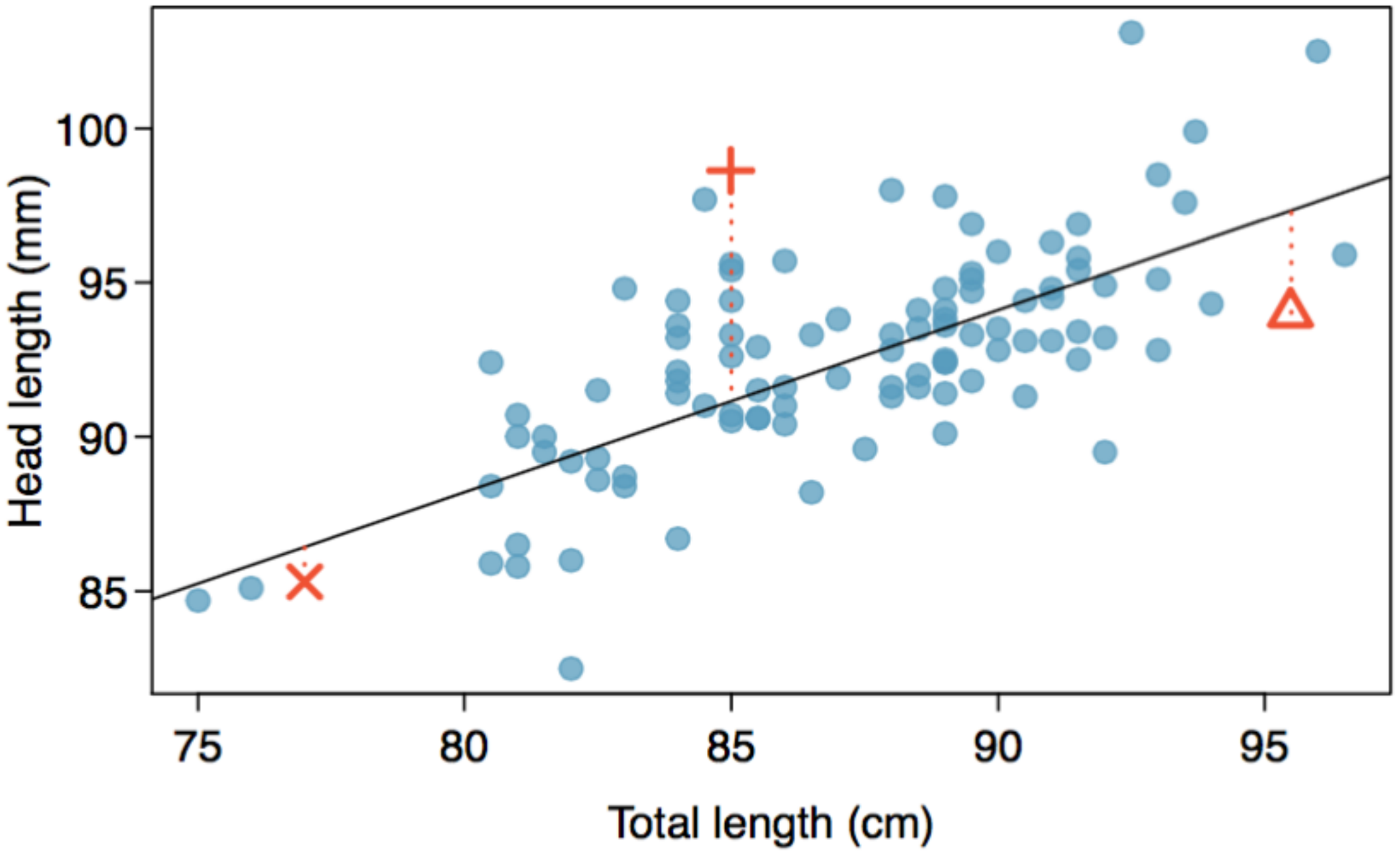


Figure 7.7: A reasonable linear model was fit to represent the relationship between head length and total length.

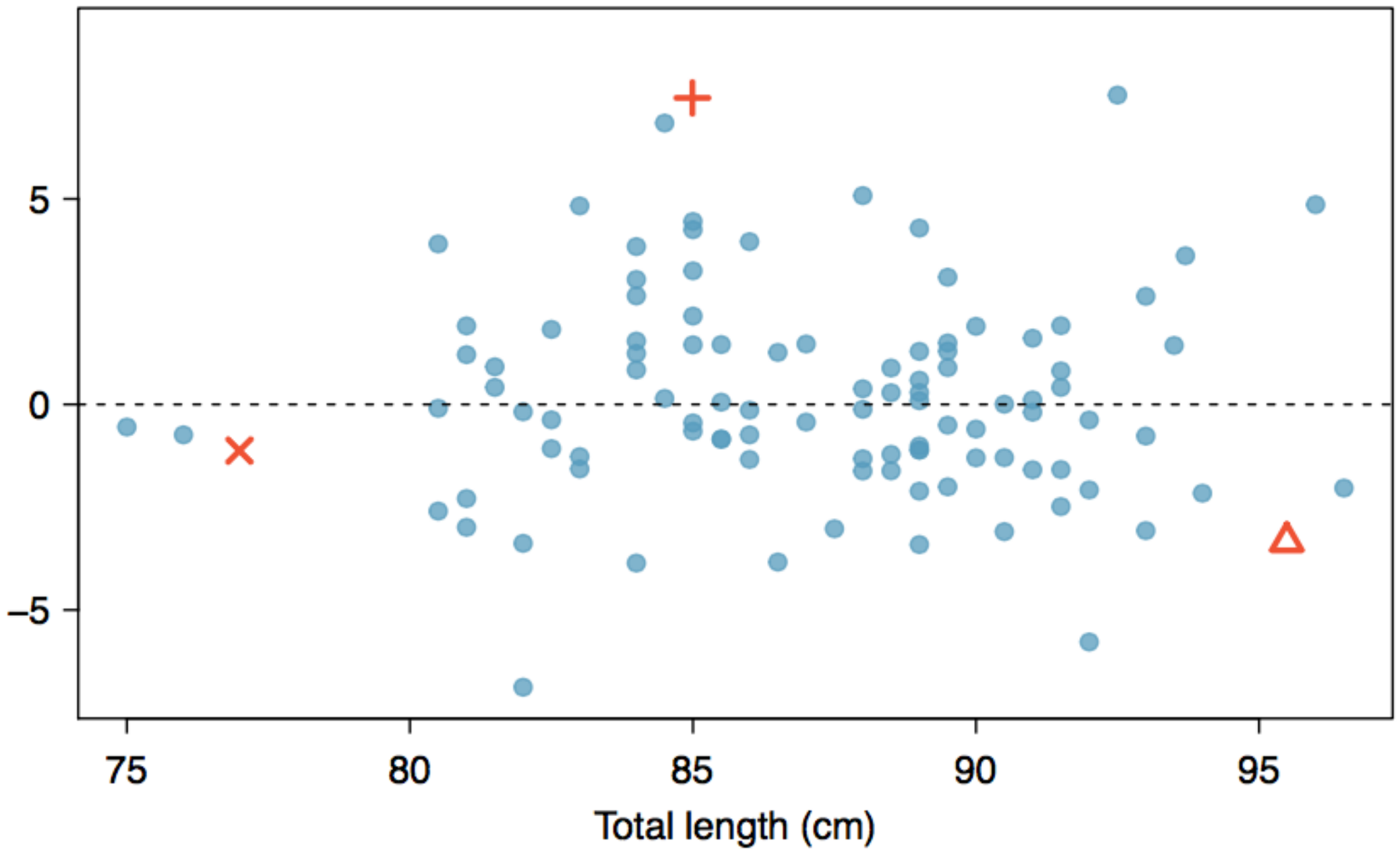
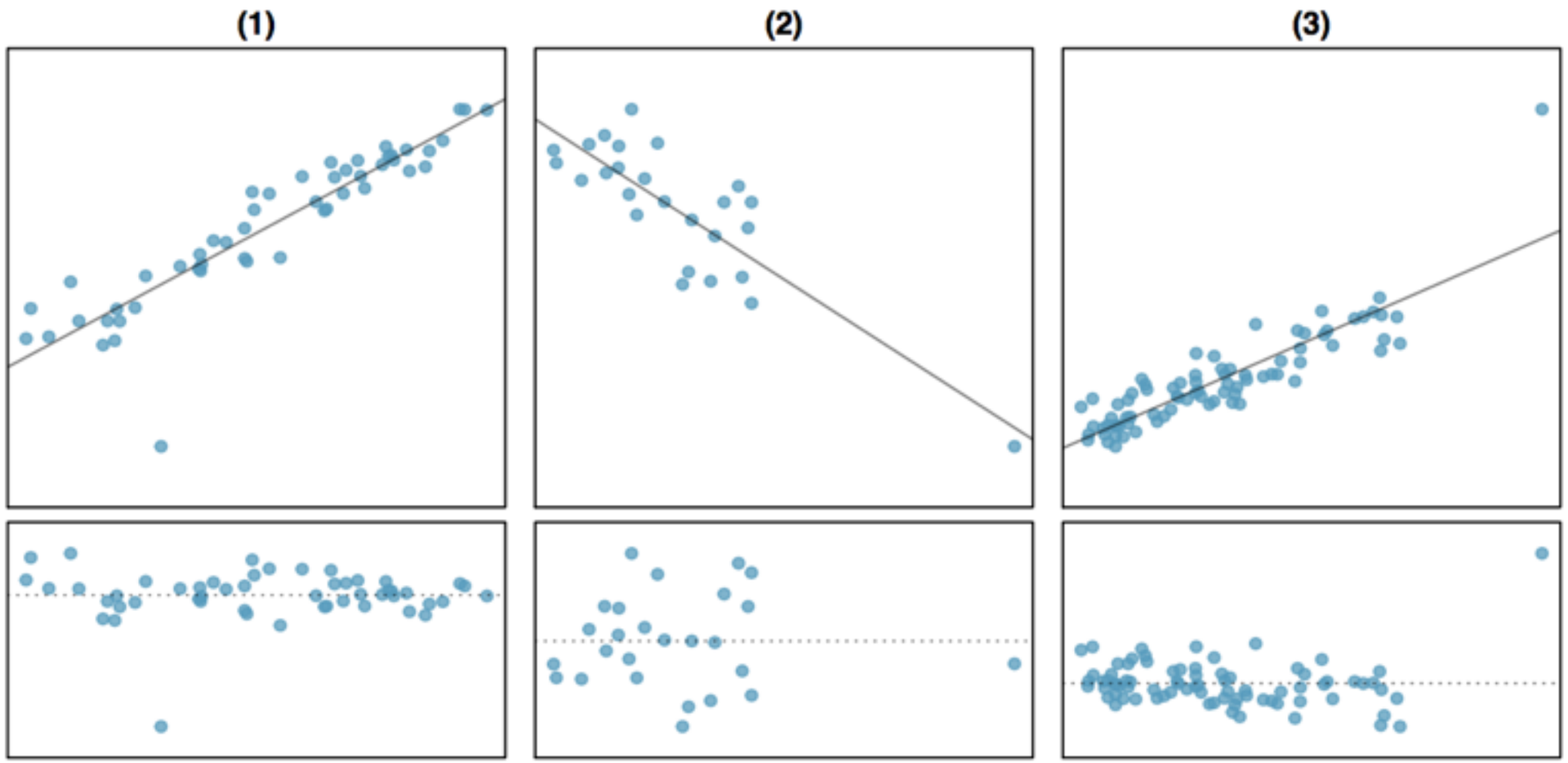


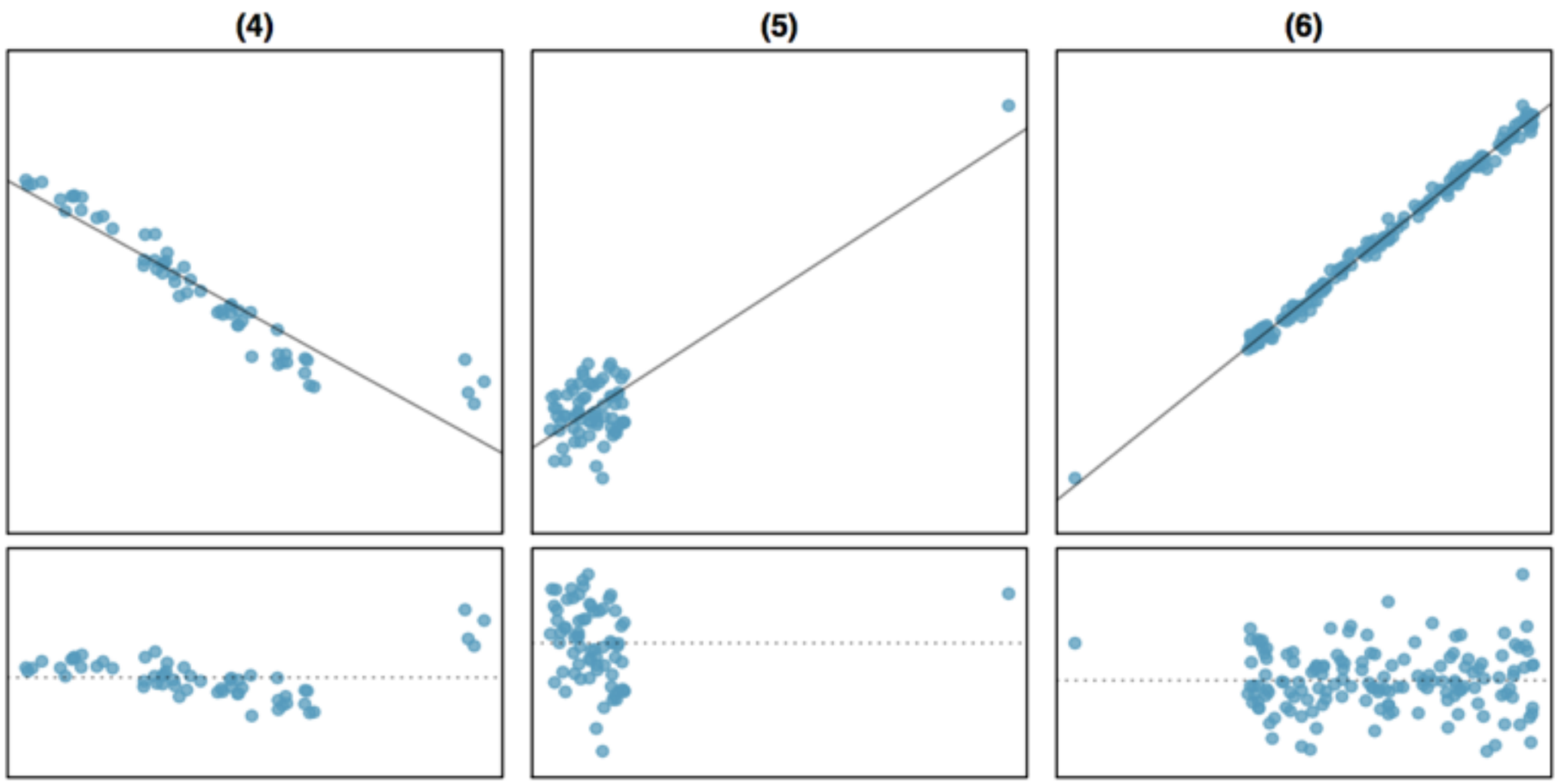
Figure 7.8: Residual plot for the model in Figure 7.7.

Outliers

Leverage: Points that fall horizontally away from the center of the cloud tend to pull harder on the line, so we call them points with high leverage.

Least squares regression is particularly vulnerable to leverage points because the distance to the best-fit line is squared what calculating residuals, amplifying the influence of the farthest points.


(1)
(2)
(3)


(4)
(5)
(6)

Correlation

Correlation: a measure of the strength of a linear relationship between two random variables.

Correlation always takes values between -1 and 1.

We denote the correlation by R.

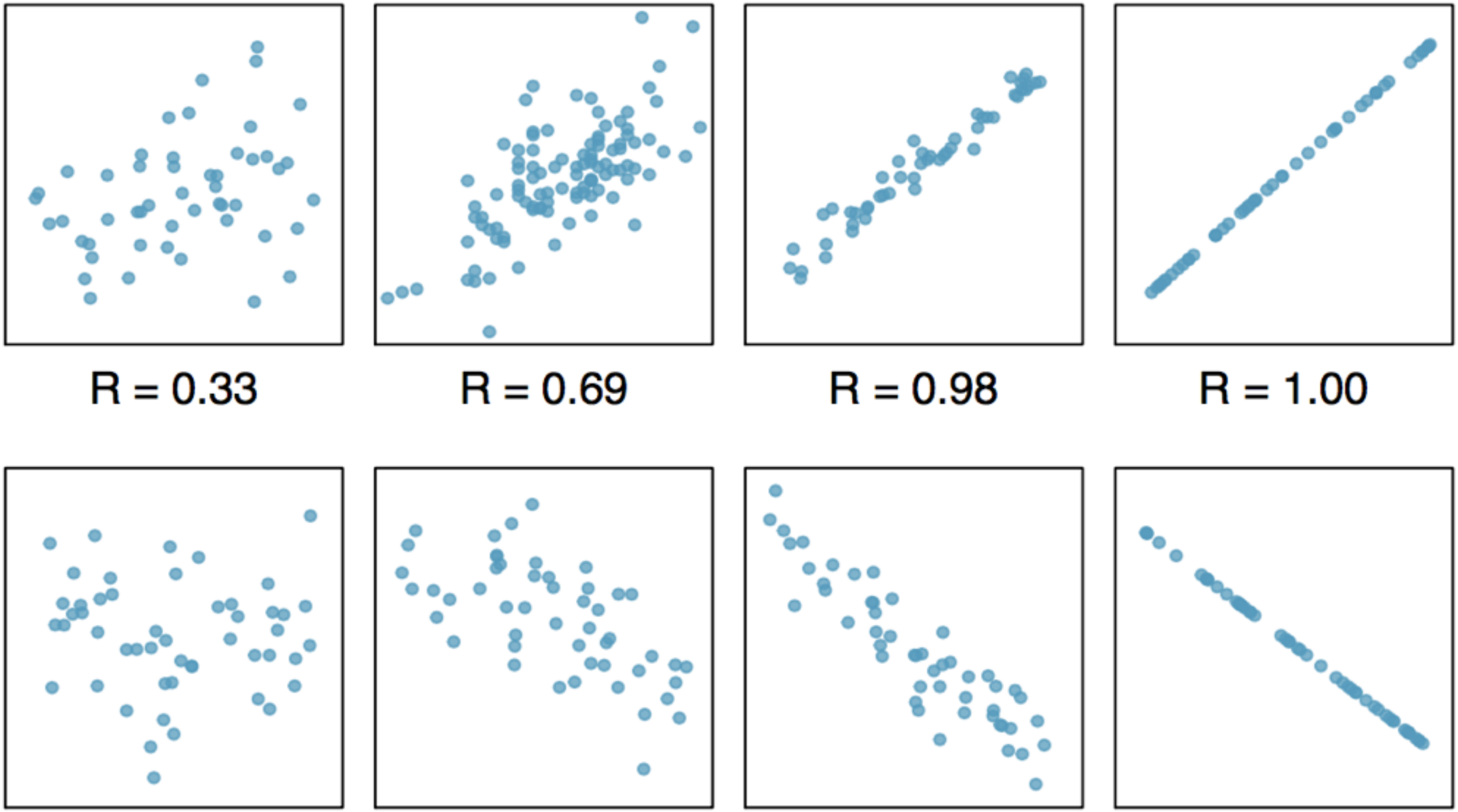


Figure 7.10: Sample scatterplots and their correlations. The first row shows variables with a positive relationship, represented by the trend up and to the right. The second row shows variables with a negative trend, where a large value in one variable is associated with a low value in the other.

Choosing the Regression Line

Given some data, there are clearly many lines that could be drawn through the scatterplot. How do we decide which one to choose?

Fitting the line means using the data to determine estimate for $\beta 0$ and $\beta 1$ in $y = \beta 0 + \beta 1x$

We choose the lines that makes the residuals the smallest.

Fitting the Line

We choose β0 and β1 such that this sum is minimized: (e1)^2+(e2)^2⋯+(en)^2, where e1 is the residual associated with the first observation x1, e2 with the second observation x2, etc.

why do we square the residuals? we need some way to control for the fact some residuals are above the line (positive) and some are below the line (negative) and they would cancel out when summed. squaring is why this method is sometimes called “least squares” or “sum of squares” regression.

Criteria for Linear Regression

When fitting a least squares line, we generally require:

- Linearity. The data should show a linear trend.

- Nearly normal residuals. Generally the residuals must be nearly normal. When this condition is found to be unreasonable, it is usually because of outliers or concerns about influential points.

- Constant variability. The variability of points around the least squares line remains roughly constant.

- Independent observations. Be cautious about applying regression to time series data, which are sequential observations in time such as a stock price each day. Such data may have an underlying structure that should be considered in a model and analysis.

Interpreting a Regression Model

The slope describes the estimated difference in the y variable if the explanatory variable x for a case happened to be one unit larger. The intercept describes the average outcome of y if x = 0 and the linear model is valid all the way to x = 0, which in many applications is not the case.

Examples

Dangers of Extrapolation

"[I]n an alarming trend, temperatures this spring have risen. Consider this: On February 6th it was 10 degrees. Today [April 6] it hit almost 80. At this rate, by August it will be 220 degrees. So clearly folks the climate debate rages on."

Stephen Colbert

Trying it out in R!

download effort.dat from http://data.princeton.edu/wws509/datasets/

read it into R! then try these commands. What is the output saying?

> lmfit = lm(change ~ setting)

> lmfit

> summary(lmfit)

> lmfit.multiple = lm(change ~ setting + effort)

See 4.1 and 4.2 from http://data.princeton.edu/R/linearModels.html

Example: R output

In America's two-party system, one political theory suggests the higher the unemployment rate, the worse the President's party will do in the midterm elections.

To assess the validity of this claim, we can compile historical data and look for a connection. We consider every midterm election from 1898 to 2010, with the exception of those elections during the Great Depression. Figure 7.20 shows these data and the least-squares regression line:

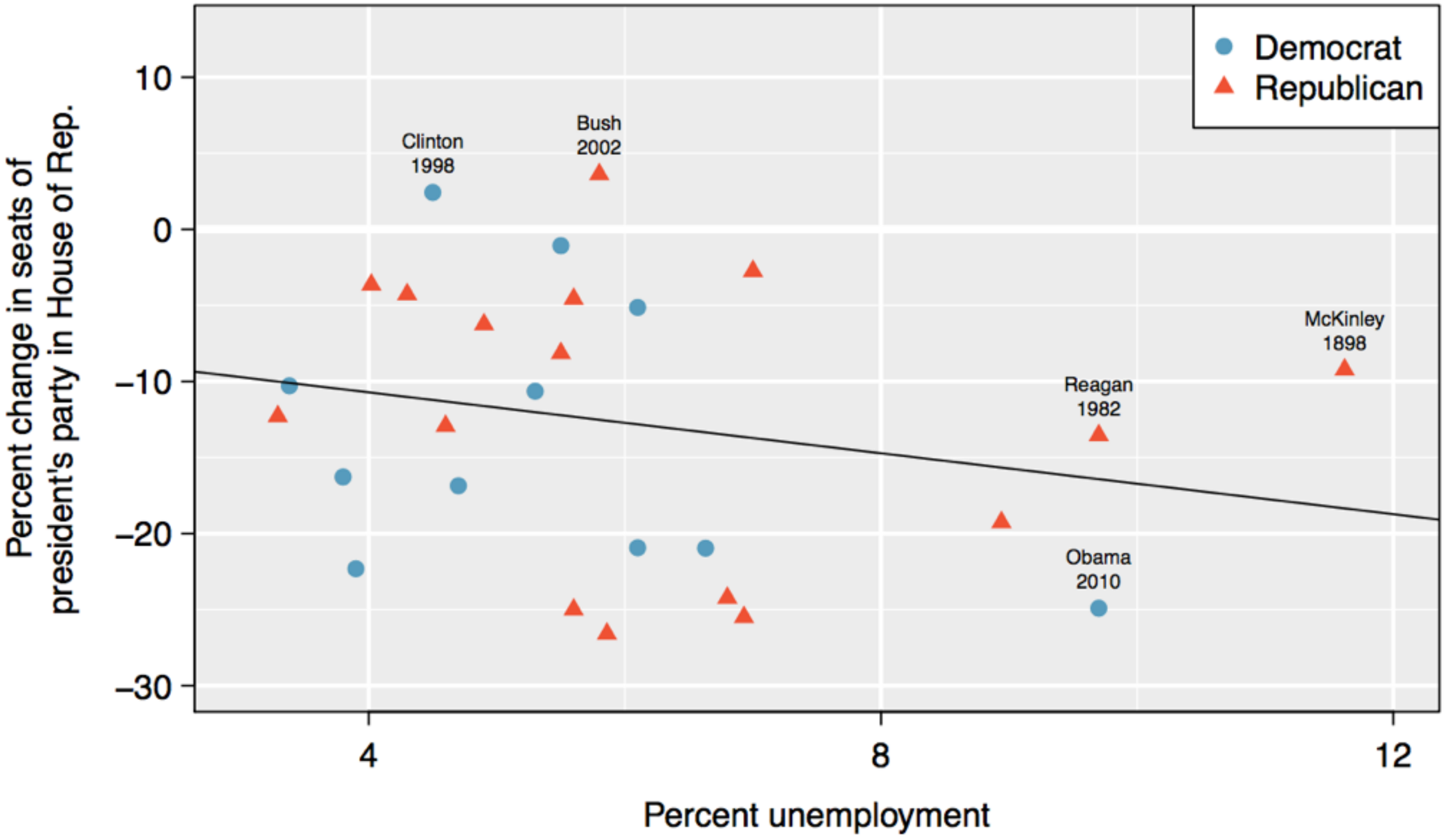
% change in House seats for President's party
= –6.71 – 1.00 × (unemployment rate)

See also https://www.youtube.com/watch?v=depiT-hTaGA

What's the hypothesis test?

H0: $\beta 1 = 0$. The true linear model has slope zero.

HA: $\beta 1 < 0$. The true linear model has a slope less than zero. The higher the unemployment, the greater the loss for the President's party in the House of Representatives.


Democrat
Republican
Clinton
1998
Bush
2002
McKinley
1898
Reagan
1982
Obama
2010
10
0
−10
−20
−30
4
8
12
Percent change in seats of president's party in House of Rep.
Percent unemployment

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-6.7142	5.4567	-1.23	0.2300
unemp	-1.0010	0.8717	-1.15	0.2617
				$df = 25$

Table 7.21: Output from statistical software for the regression line modeling the midterm election losses for the President's party as a response to unemployment.

The entries in the first column represent the least squares estimates, b0 and b1, and the values in the second column correspond to the standard errors of each estimate.

For the hypothesis we are considering, the null value for the slope is 0, so we can compute the test statistic using the Z score formula:

Z = (estimate – null value)/SE = (–1.0010 – 0)/0.8717 = –1.15

Do we reject our null hypothesis? No.. (why? a cutoff for significance is about 2, and |-1.15| < 2)

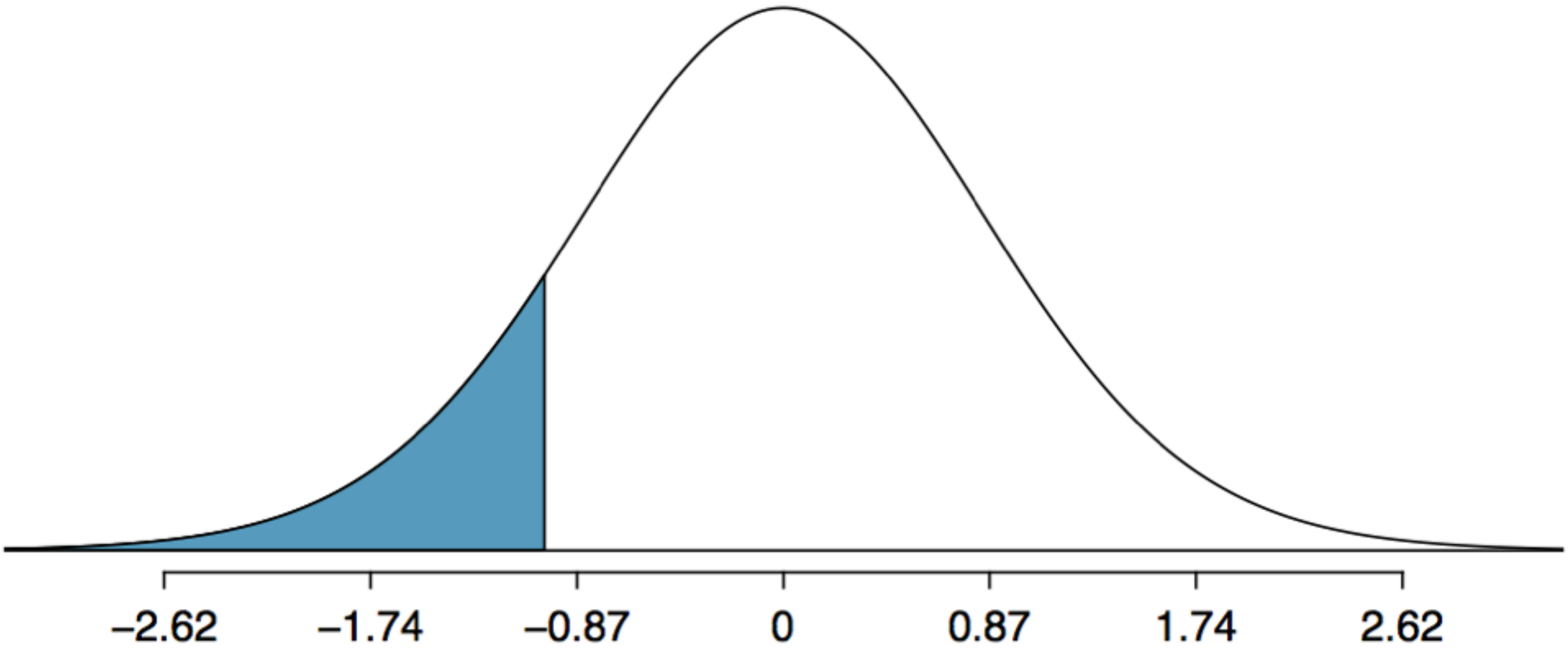


Figure 7.22: The distribution shown here is the sampling distribution for b_1, if the null hypothesis was true. The shaded tail represents the p-value for the hypothesis test evaluating whether there is convincing evidence that higher unemployment corresponds to a greater loss of House seats for the President's party during a midterm election.

More Examples in R

- https://www.r-bloggers.com/r-tutorial-series-simple-linear-regression/

- http://www.cyclismo.org/tutorial/R/linearLeastSquares.html

- http://www.ats.ucla.edu/stat/stata/output/reg_output.htm