Fitting a Line, Residuals, and Correlation

August 27, 2019

Fitting a Line to Data

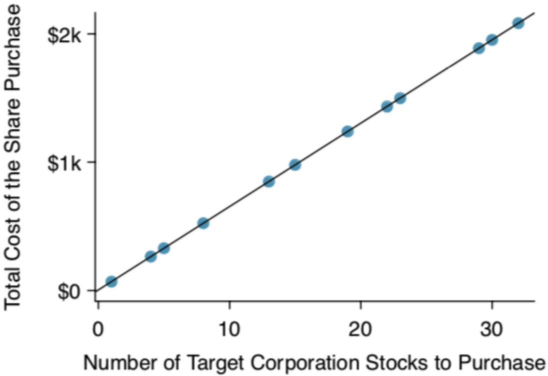
In this section, we will talk about fitting a line to data.

- Our hypothesis testing framework allowed us to examine one variable at a time.
- Linear regression will allow us to look at relationships between two (or more) variables.

Fitting a Line to Data

- We discussed relationships between two variables when we looked at scatterplots.
- We thought some about correlations and the strength of those relationships.
- This section will help us to formalize some of those concepts.

Fitting a Line to Data



This relationship can be modeled perfectly with a straight line:

$$y = 5 + 64.96x$$

Fitting a Line to Data

When we can model a relationship *perfectly*,

$$y = 5 + 64.96x,$$

we know the exact value of y just by knowing the value of x.

However, this kind of perfect relationship is pretty unrealistic... it's also pretty uninteresting.

Linear Regression

Linear regression takes this idea of fitting a line and allows for some error:

$$y = \beta_0 + \beta_1 x + \epsilon$$

- β_0 ("beta 0")and β_1 are the model's parameters.
- The error is represented by ϵ.

Linear Regression

- The parameters β_0 and β_1 are estimated using data.
- We denote these point estimates by b_0 and b_1.

Linear Regression

For a regression line

$$y = \beta_0 + \beta_1 x + \epsilon$$

we make predictions about y using values of x.

- y is called the **response variable**.
- x is called the **predictor variable**.

Linear Regression

When we find our point estimates b_0 and b_1, we usually write the line as

$$\hat{y} = b_0 + b_1 x$$

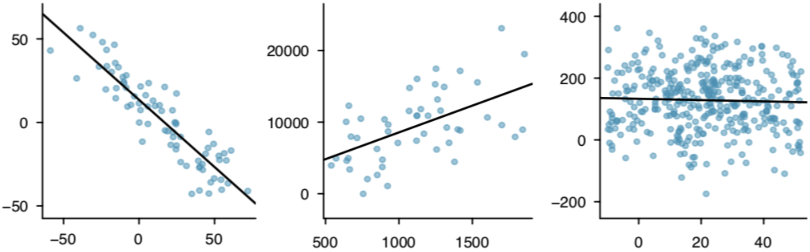
We drop the error term because it is a random, unknown quantity. Instead we focus on $\hat{y}$, the predicted value for y.

Linear Regression

As with any line, the intercept and slope are meaningful.

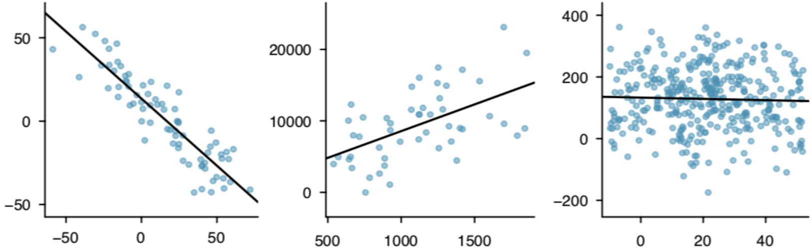
- The slope β_1 is the change in y for every one-unit change in x.
- The intercept β_0 is the predicted value for y when $x = 0$.

Clouds of Points



In all 3 datasets, finding the linear trend may be useful! This is true despite the points sometimes falling somewhat far from the line.

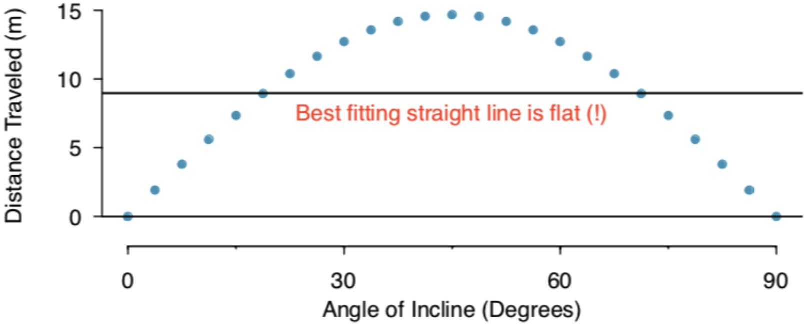
Clouds of Points



Think of this like the 2-dimensional version of a point estimate.

- The line gives our best estimate of the relationship.
- There is some variability in the data that will impact our confidence in our estimates.
- The true relationship is unknown.

Linear Trends



Sometimes, there is a clear relationship but linear regression will not work! We can use slightly more advanced models for these settings (but we'll leave that for STAT 100B).

Prediction

Often, when we build a regression model our goal is prediction.

- We want to use information about the predictor variable to make predictions about the response variable.

Example: Possum Head Lengths

Remember our brushtail possums?

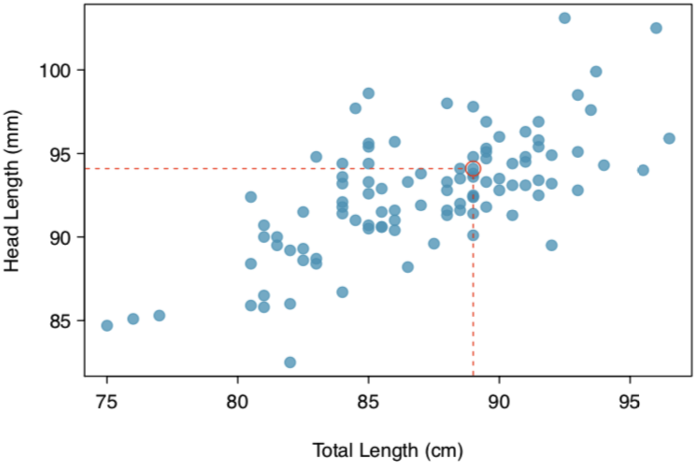
Example: Possum Head Lengths

Researchers captured 104 brushtail possums and took a variety of body measurements on each before releasing them back into the wild.

We consider two measurements for each possum:

- total body length.
- head length.

Example: Possum Head Lengths



Example: Possum Head Lengths

- The relationship isn't perfectly linear.
- However, there does appear to be a linear relationship.
- We want to try to use body length to predict head length.

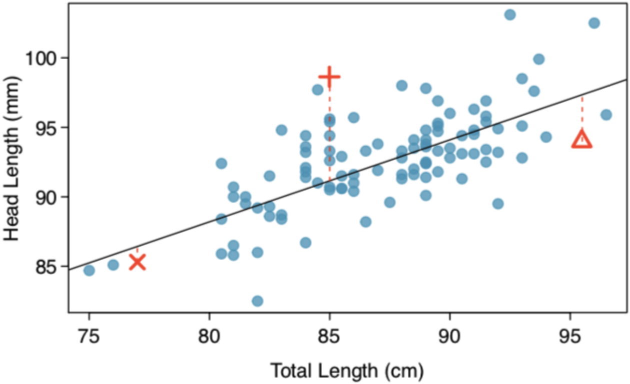
Example: Possum Head Lengths

The textbook gives the following linear relationship:

$$\hat{y} = 41 + 0.59x$$

As always, the hat denotes an estimate of some unknown true value.

Example: Possum Head Lengths



Suppose we wanted to predict the head length for a possum with a body length of 80 cm.

Example: Possum Head Lengths

We could try to do this using the scatterplot, but since the relationship isn't perfectly linear it's difficult to estimate.

With a regression line, we can instead calculate this mathematically:

$$\begin{aligned} \hat{y} &= 41 + 0.59x \\ &= 41 + 0.59 \times 80 \\ &= 88.2 \end{aligned}$$

Example: Possum Head Lengths

This estimate should be thought of as an average.

The regression equation predicts that, *on average*, possums with total body length 80 cm will have a head length of 88.2 mm.

Example: Possum Head Lengths

If we had more information (other variables), we could probably get a better estimate.

We might be interested in including

- sex
- region
- diet

or others.

Absent addition information, 88.2 mm is a reasonable prediction.

Residuals

Residuals are the leftover variation in the data after accounting for model fit:

$$\text{data} = \text{prediction} + \text{residual}$$

Each observation will have its own residual.

Residuals

Formally, we define the residual of the ith observation (x_i, y_i) as the difference between observed (y_i) and expected $(\hat{y}_i)$:

$$e_i = y_i - \hat{y}_i$$

We denote the residuals by e_i and find $\hat{y}$ by plugging in x_i.

Residuals

If an observation lands above the regression line,

$$e_i = y_i - \hat{y}_i > 0.$$

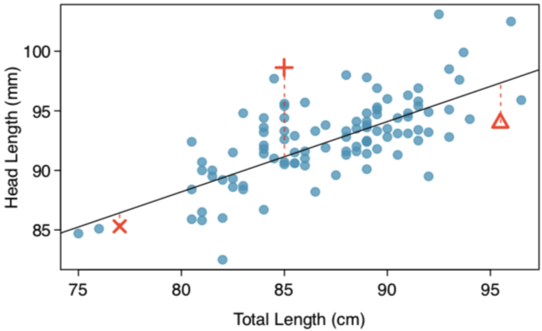
If below,

$$e_i = y_i - \hat{y}_i < 0.$$

Residuals

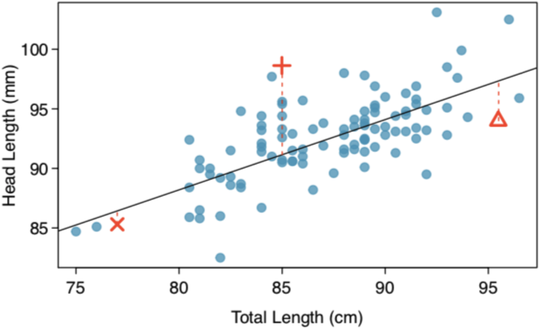
When we estimate the parameters for the regression, our goal is to get our residuals as close to 0 as possible.

Example: Possum Head Lengths



The residual for each observation is the vertical distance between the line and the observation.

Example: Possum Head Lengths



- $\times$ has a residual of about -1
- $+$ has a residual of about 7
- $\triangle$ has a residual of about -4

Example: Possum Head Lengths

The scatterplot is nice, but a calculation is always more precise. Let's find the residual for the observation $(77.0, 85.3)$.

Example: Possum Head Lengths

The predicted value $\hat{y}$ is

$$\begin{aligned}\hat{y} &= 41 + 0.59x \\ &= 41 + 0.59 \times 77.0 \\ &= 86.4\end{aligned}$$

Example: Possum Head Lengths

Then the residual is

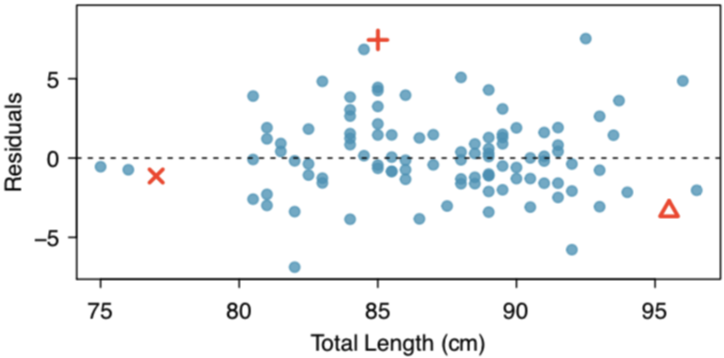
$$\begin{aligned} e &= y - \hat{y} \\ &= 85.3 - 86.4 \\ &= -1.1 \end{aligned}$$

So the model over-predicted head length by 1.1mm for this particular possum.

Residual Plots

- Our goal is to get our residuals as close as possible to 0.
- Residuals are a good way to examine how well a linear model fits a data set.
- We can examine these quickly using a residual plot.

Residual Plots



- Residual plots show the x-values plotted against their residuals.
- Essentially we've titled and re-scaled the scatterplot so that the regression line is horizontal at 0.

Residual Plots

- We use residual plots to identify characteristics or patterns.
- These are things that are still apparent event after fitting the model.
- Obvious patterns suggest some problems with our model fit.

Residual Plots

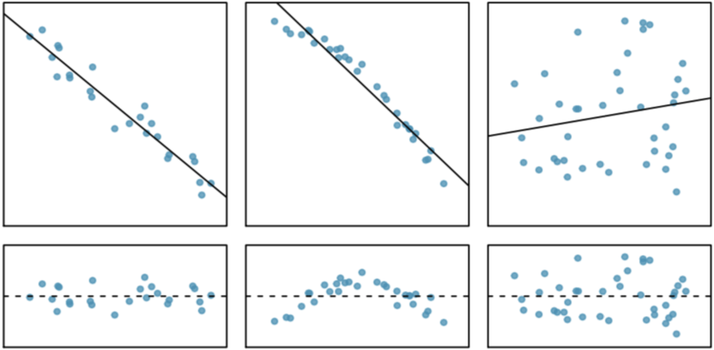


Figure 8.8: Sample data with their best fitting lines (top row) and their corresponding residual plots (bottom row).

Correlation

We've talked about the strength of linear relationships, but it would be nice to formalize this concept.

The **correlation** between two variables describes the strength of their linear relationship. It always takes values between -1 and 1.

Correlation

We denote the correlation (or correlation coefficient) by R:

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \times \frac{y_i - \bar{y}}{s_y} \right)$$

where s_x and s_y are the respective standard deviations for x and y.

Correlation

Correlations

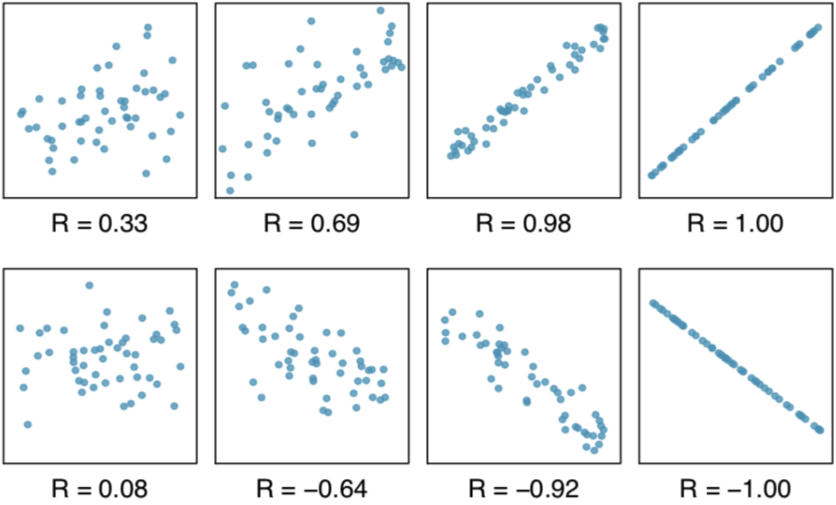
- Close to -1 suggest strong, negative linear relationships.
- Close to +1 suggest strong, positive linear relationships.
- Close to 0 have little-to-no linear relationship.

Correlation

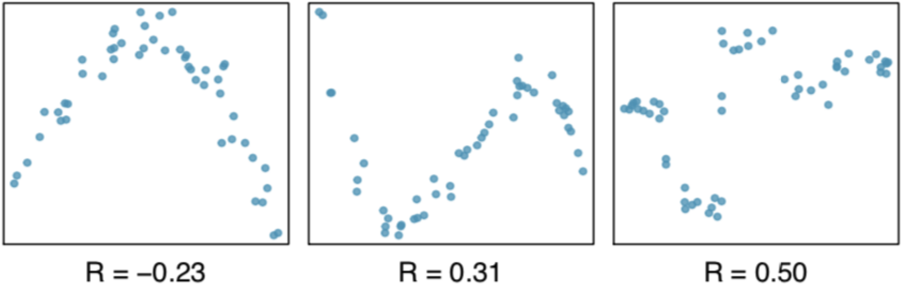
Note: the sign of the correlation will match the sign of the slope!

- If $R < 0$, there is a downward trend and $b_1 < 0$.
- If $R > 0$, there is an upward trend and $b_1 > 0$.
- If $R \approx 0$, there is no relationship and $b_1 \approx 0$.

Correlation



Correlations only represent *linear* trends!



Clearly there are some strong relationships here, but they are not ones we can represent well using a correlation coefficient.

Finding the Best Line

We want a line with small residuals, but if we minimize

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i)$$

we will get very large negative residuals!

Finding the Best Line

As with the standard deviation, we will use squares to shift the focus to magnitude:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

and will find the β estimates that minimize this. This is called the **Least Squares Criterion**.

Finding the Best Line

We often call this approach **least square regression**. To fit this line, we want

- Linearity. The data should show a linear trend.
- Nearly normal residuals. The residuals should be well-approximated by a normal distribution.
- Constant variability. As we move along x, the variability around the regression line should stay constant.
- Independent observations. This will apply to random samples.

Finding the Least Squares Line

We want to estimate β_0 and β_1 in the equation

$$y = \beta_0 + \beta_1 x + \epsilon$$

by minimizing $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

Finding the Least Squares Line

This turns out to be remarkably straightforward! The slope can be estimated as

$$b_1 = \frac{s_y}{s_x} R$$

and the intercept by

$$b_0 = \bar{y} - b_1 \bar{x}$$

Finding the Least Squares Line

Although these formulas are easy to write out, they can be cumbersome to work through.

We usually use a computer to find the equation for a least squares linear regression line!

Extrapolation

- When we make predictions, we simply plug in values of x to estimate values of y.
- However, this has limitations!
- We don't know how the data outside of our limited window will behave.

Extrapolation

Applying a model estimate for values outside of the data's range for x is called **extrapolation**.

- The linear model is only an approximation.
- We don't know anything about the relationship outside of the scope of our data.
- Extrapolation assumes that the linear relationship holds in places where it has not been analyzed.

Extrapolation

When those blizzards hit the East Coast this winter, it proved to my satisfaction that global warming was a fraud. That snow was freezing cold. But in an alarming trend, temperatures this spring have risen. Consider this: On February 6^{th} *it was 10 degrees. Today it hit almost 80. At this rate, by August it will be 220 degrees. So clearly folks the climate debate rages on.*

Stephen Colbert
April 6th, 2010[12]

Using R^2 to Describe Strength of Fit

We've evaluated the strength of a linear relationship between two variables using the correlation coefficient R.

However, it is also common to use R^2. This helps describe how closely the data cluster around a linear fit.

Using R^2 to Describe Strength of Fit

Suppose $R^2 = 0.62$ for a linear model. Then we would say

- About 62% of the data's variability is accounted for using the linear model.

And yes, R^2 is the square of the correlation coefficient R!

What's good?

So what is a good or a bad fit?
This will depend a lot on what field you are in!

However, for the purpose of this class, we will use a GPA system:

- $R^2 \geq 0.9$ is an A fit.
- $0.8 \leq R^2 < 0.9$ is a B fit.
- $0.7 \leq R^2 < 0.8$ is a C fit.
- $0.6 \leq R^2 < 0.7$ is a D fit.
- $R^2 < 0.6$ is an F fit.