Fitting a Line, Residuals, and Correlation

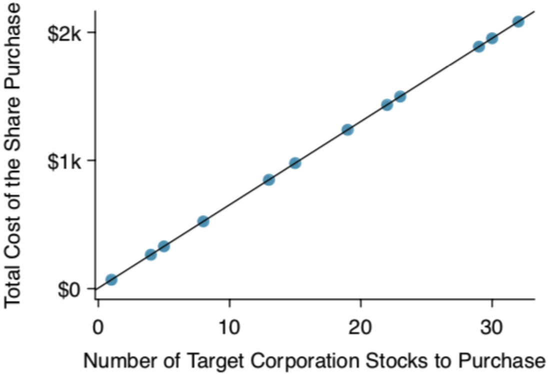
October 28, 2019

Fitting a Line to Data

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

- Linear regression will allow us to look at relationships between two (or more) variables.
- This is a bit like ANOVA, but now we will be able to *predict* outcomes.

Fitting a Line to Data



This relationship can be modeled perfectly with a straight line:

$$y = 5 + 64.96x$$

I.e., x and y are perfectly correlated.

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 and β_1 are the model's parameters.
- The error is represented by ϵ.

- The parameters β_0 and β_1 are estimated using data.
- We denote these point estimates by b_0 and b_1.
    - ...or sometimes $\hat{\beta}_0$ and $\hat{\beta}_1$

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

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

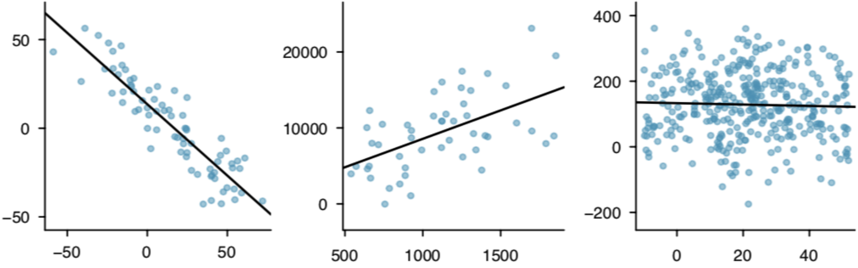
We drop the error term because it is a random, unknown quantity. Instead we focus on $\hat{y}$, the predicted value for y.

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

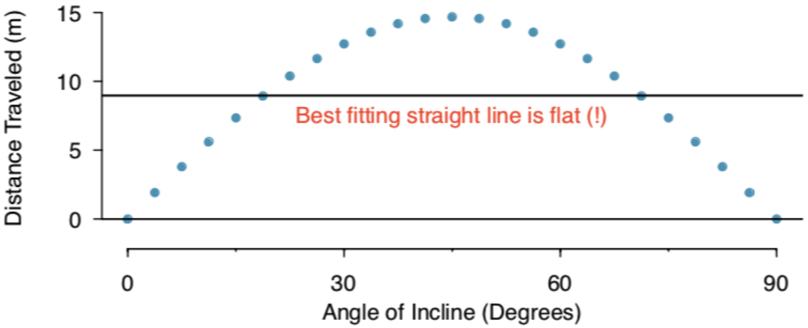


Clouds of Points

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 simple linear regression won't work! We will talk about this later in the term.

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?

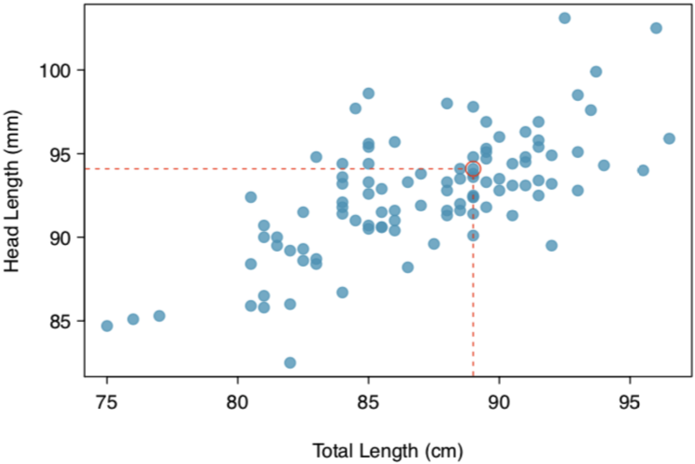
Example: Possum Head Lengths

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.

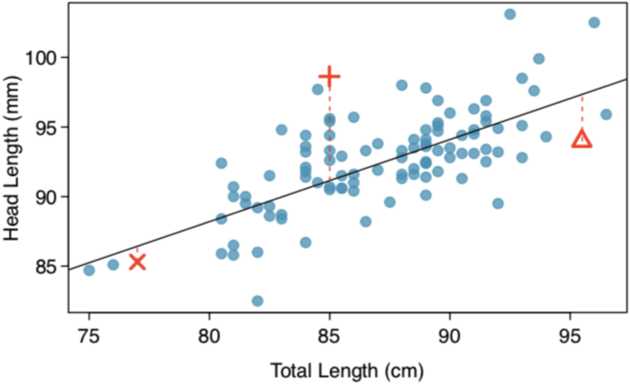
Example: Possum Head Lengths

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



Predict the head length for a possum with a body length of 80 cm.

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, our prediction is a reasonable estimate.

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,

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

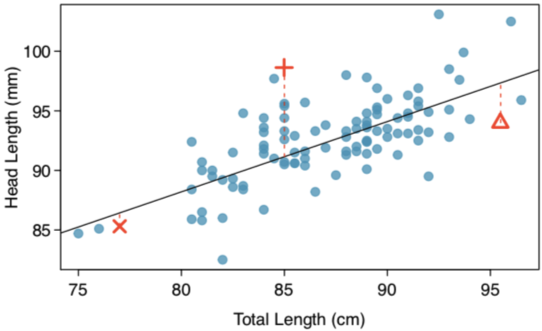
If below,

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

Residuals

When we estimate the parameters for the regression, our goal is to get each residual 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.

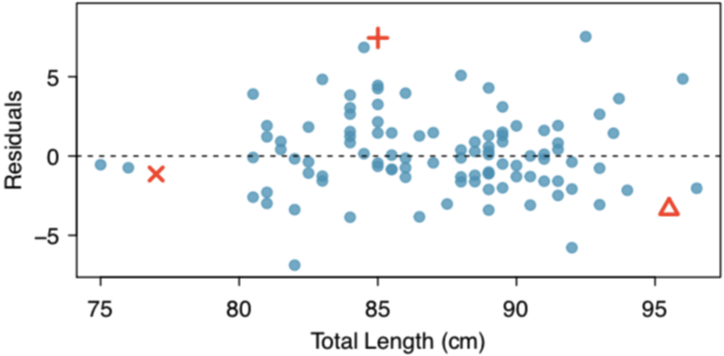
Example: Possum Head Lengths

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

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.

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

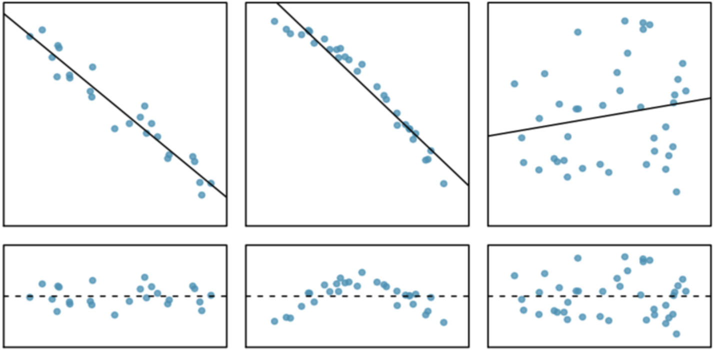


Figure 8.8: Sample data with their best fitting lines (top row) and their corresponding residual plots (bottom row).

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

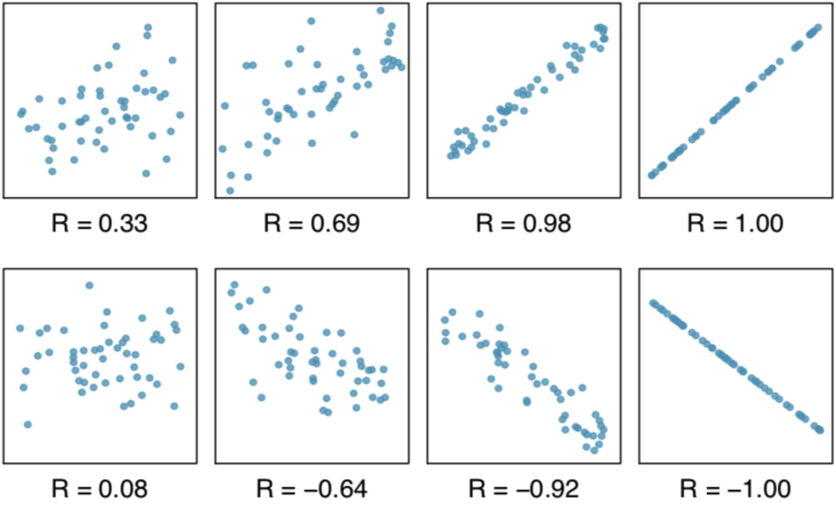
- Close to -1 suggest strong, negative linear relationships.
- Close to +1 suggest strong, positive linear relationships.
- Close to 0 have little-to-no linear relationship.

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!