Lecture 9

Simple linear regression

Dr. Franco Ubaudi

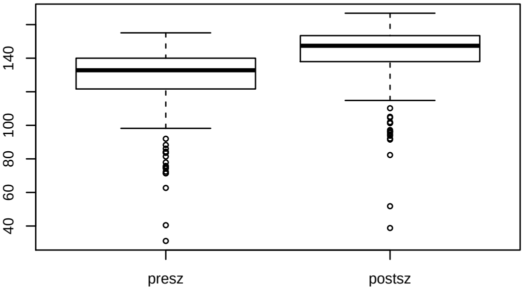
The Nature of Data
Western Sydney University

Spring 2021

Outline

- Straight lines
  - Slope and Intercept
  - Straight line parameters
- Simple Linear Regression:
  - Least-Squares
  - Residuals
  - Where Pearson correlation fits in
- Hypothesis testing linear model parameters
- Confidence Intervals of linear model parameters
- Evaluating Linear Models
  - Residual Sum of Squares
  - R-squared
- Using Linear Models
  - Prediction
  - Confidence Intervals for the mean of a prediction

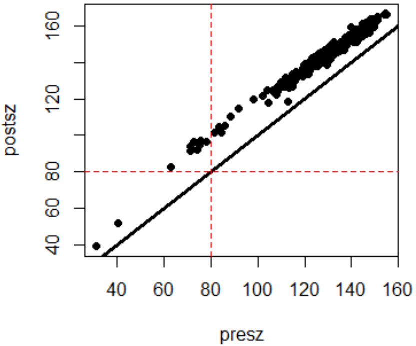
Crabs moulting



Is the difference in before and after moulting size:

- a constant? (e.g. increase of 5mm)
- or a percentage increase? (e.g. 10%)
- or a combination of the two?

Crabs moulting



The solid line shows where pre moult = post moult.

It appears that the points are just lifted up by a constant amount.

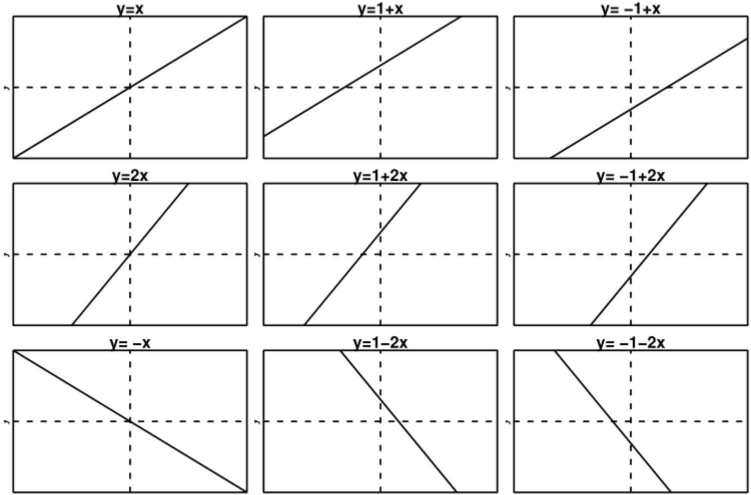
Straight Lines

$$y = a + bx$$

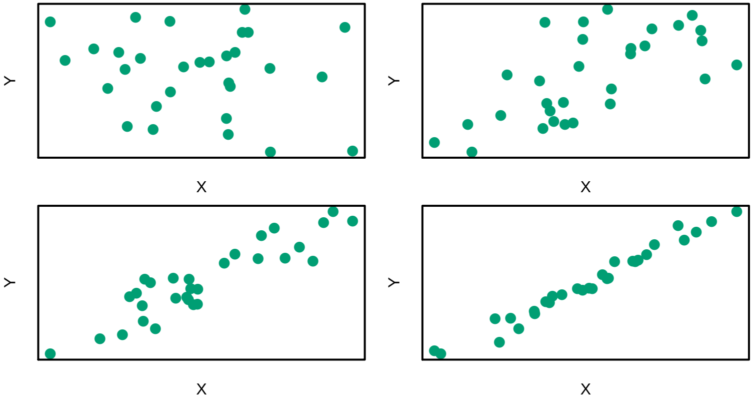
is a straight line relationship defined by x and y

with constants a (intercept) and b (slope).

Straight Lines



Straight Line Parameters



How to estimate a and b? $y = a + bx$

Simple Linear Regression

Simple linear regression is curve fitting:
estimate parameters for a best-fit.

- $y = a + bx$
- y is the dependent variable / response
- x is the independent variable / predictor

The line is a linear model for predicting y given x

Simple Linear Regression

Given points: (x_i, y_i) $i = 1, \ldots, n$

A perfect fit would allow us to write $y = a + bx$ for some a and b.

In practice, we have error around the line:

$$y = a + bx + \varepsilon$$

$$y_i = a + bx_i + \varepsilon_i$$

How to choose the "best" a and b?

Which is the best line?

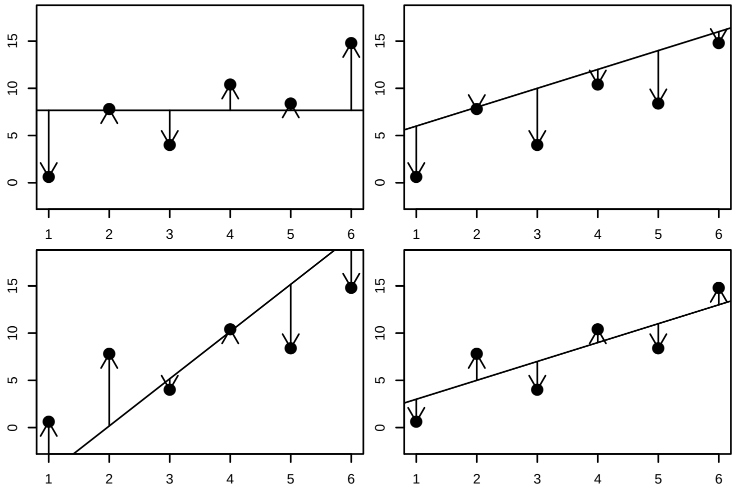


Figure: Difference in size after moulting

Least-Squares

$y = a + bx$

fitted value $\hat{y}_i = a + bx_i$

Let $e_i = y_i - \hat{y}_i$ be the residuals – difference between observed y_i and predicted by the line

Let RSS $= \sum_{i=1}^{n} e_i^2 =$ the squared distance of the line to all the observations – called the residual sum of squares.

Least-Squares

The best-fit in the "least-squares" sense is the a and b the minimise the RSS.

$$
\begin{aligned}
RSS &= \sum_{i=1}^{n} e_i^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n} (y_i - (a + bx_i))^2.
\end{aligned}
$$

Taking partial-derivatives of RSS w.r.t. a and b and solving equations at 0 gives us the answer.

Least-Squares

Define

$$COV(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

where $\bar{y}$ and $\bar{x}$ are the mean of y and x.

The RSS is minimised when $\hat{b} = \frac{COV(x,y)}{COV(x,x)}$ and $\hat{a} = \bar{y} - \hat{b}\bar{x}$.

Note the Pearson correlation

$$r = \frac{COV(x,y)}{\sqrt{COV(x,x)COV(y,y)}}.$$

Moulting crabs

Define $SS(X, Y) = nCOV(x, y)$

For the crab moulting data with `presz` as x, `postsz` as y

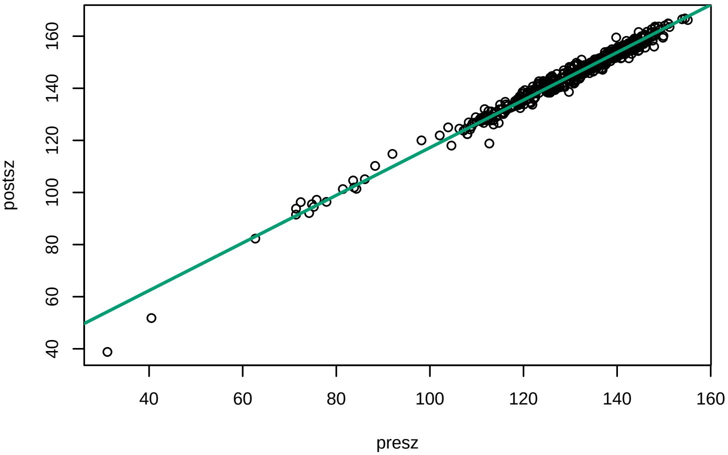
$$
\begin{array}{llll}
\bar{x} & = 129.21 & SS_{XX} & = 118542.69 \\
\bar{y} & = 143.9 & SS_{YY} & = 100957.55 \\
n & = 472 & SS_{XY} & = 108343.84
\end{array}
$$

So that

$$\hat{b} = 108343.84/118542.69 = 0.914$$

$$\hat{a} = 143.9 - 0.91 * 129.21 = 25.803$$

Moulting crabs



$$\hat{b} = 0.914 \qquad \hat{a} = 25.803$$

Slope and Intercept

The slope represents the amount by which y increases for every unit increase in x.

Crabs slope $= 0.914$

On average, for every mm the crab is larger pre moulting, it is 0.914 mm larger post moulting.

The intercept is the value of y when x is zero.

However, the post moult size of such a hypothetical crab of pre-moult size 0 would be 25.803mm.

Slope and Intercept

The formula $y = \hat{a} + \hat{b}x$ gives a prediction of y given x.

E.g., Using $\hat{a} = 25.803$mm and $\hat{b} = 0.914$, for a crab of pre-moult size $x = 120$mm the predicted post moult size is $\hat{y} = 25.803 + 0.91 * 120 = 135.48$mm.

That slope $\hat{b} < 1$ suggests that larger crabs grow by a smaller amount on average than the smaller crabs.

Is $\hat{b}$ or even $\hat{a}$, a sampling issue or true of the (crab) population?

Hypothesis testing

Using the permutation approach we can look for evidence against the hypothesis that $b = 0$.

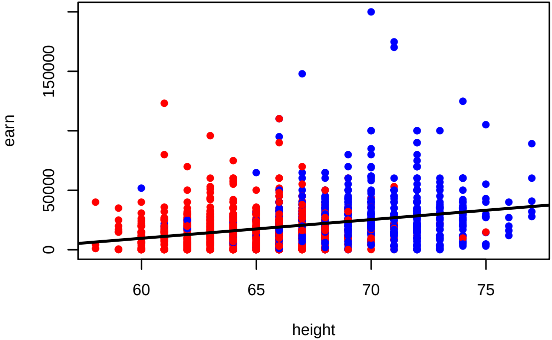
For the heights and earnings data a slope of zero would mean heights do not affect earnings.

Given $\quad earnings = a + b \times height$

if $\quad b = 0$

$$earnings = a$$

Hypothesis testing



$\hat{b} = 1571.05$, $\hat{a} =$ -84633.92.

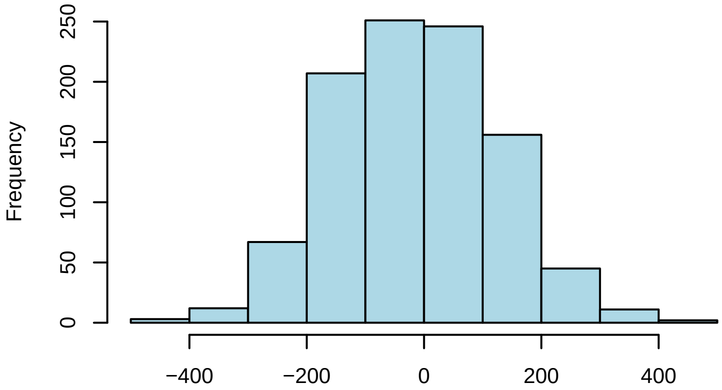
Each inch taller, seems to predict \$1,571 higher salary.

To see if population slope $\neq 0$, can use the same permutation strategy as for correlation.

Hypothesis testing

Problem

Below is the distribution of $\hat{b}$ when $b = 0$. Recall that the data slope is $\hat{b} = 1571.05$. What is the p value and the conclusion of the test ($H_0 : b = 0$, $H_A : b \neq 0$)?



Crabs moulting

Here interested in testing if the population slope $= 1$

$$\begin{aligned} postsz &= a + b\,presz \\ \implies postsz - presz &= a + (b-1)presz \end{aligned}$$

So we fit $y = postsz - presz$ against $x = presz$ and test its slope against zero

$$\begin{aligned} y &= postsz - presz \\ b' &= b - 1 \\ \implies y &= a + b' \times presz \end{aligned}$$

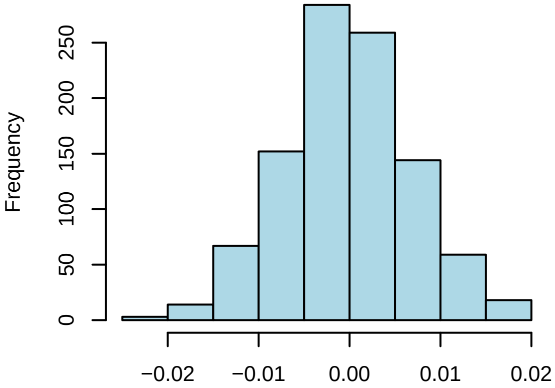
Crabs moulting

- Observed Slope = -0.086

Crabs moulting

Problem

Below shows the distribution of $\hat{b} - 1$, when $b = 1$. Given that $\hat{b} - 1 = -0.086$ in the data, what is the p value and the conclusion of the hypothesis test ($H_0 : b - 1 = 0$, $H_A : b - 1 \neq 0$)?



Confidence Intervals

The hypothesis test showed strong evidence that the change in size during moult is not simply a constant. The evidence suggests that larger crabs grow by a smaller amount.

To find a confidence interval for the slope we can use the bootstrapping.

It is important that we sample pairs (x_i, y_i) of points to keep the relationship intact.

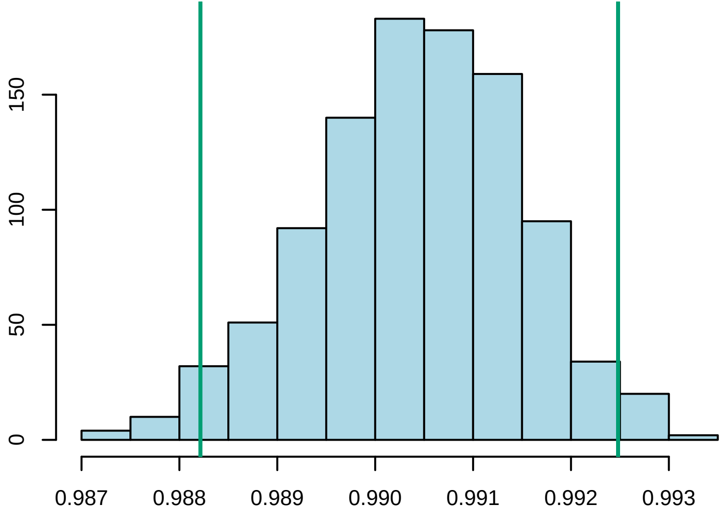
So simply,

- sample with replacement from the pairs of points
- compute the slope
- repeat, and use the bootstrapped slopes to find a confidence interval in the usual way

There are other (in some senses better) approaches, but this is the simplest.

Confidence Intervals

- observed slope = 0.914
- 95% confidence interval is 0.885, 0.943



Residuals - diagnostics

Recall model

$$y_i = a + bx_i + \varepsilon_i.$$

With estimated $\hat{a}$ and $\hat{b}$, we can compute **fitted values** for each pair.

$$\hat{y}_i = \hat{a} + \hat{b}x_i$$

The difference between the fitted values and the observed value is called the residual

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

Residuals - diagnostics

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

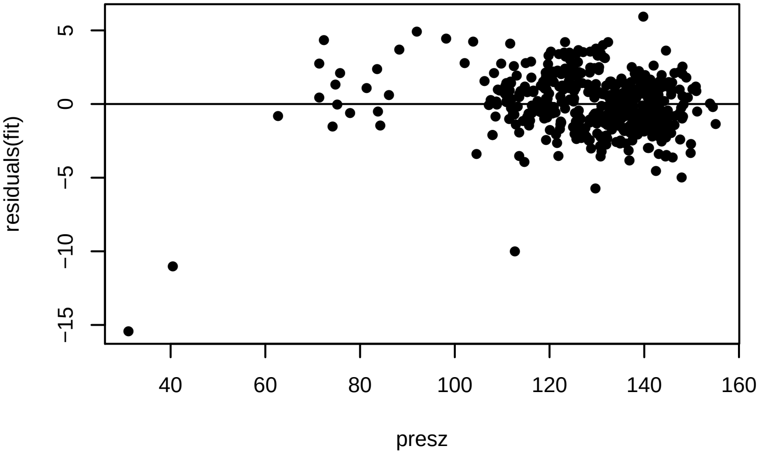
Residuals should be "noise" – more or less random points.

If not, the model is probably not appropriate.

To check, plot them against their corresponding x values.

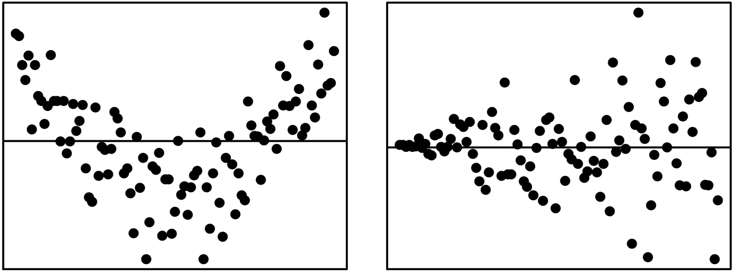
We are looking for any systematic variation.

Residuals - diagnostics



For the crab moult data there are a few rather large negative residuals, but otherwise no particular pattern.

Typical problems found in residuals



- The left panel: the true model is NOT a straight line.
- The right panel: residuals **fan out** to the right. This indicates variability dependent on x, and simple least squares is not appropriate.

Residual Sum of Squares

Recall we minimised

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

The residual variance can be estimated by:

$$s^2 = \frac{RSS}{n-2}$$

R-squared

R^2: another important goodness-of-fit measure represents the proportion of variation in y explained by regression on x.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

In simple linear regression, $R^2 = r^2$ where r is the Pearson correlation.

R-squared

For the crab moult data,

- The sample size is $n = 472$.
- The RSS is 1935.09
- The variance of the residuals is $s^2 = 4.1172$
- The Total sum of Squares is $SS_{Total} = 100957.55$
- Thus the R^2 is 0.9808

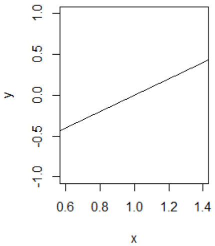
R^2 are often expressed as a percentage.

R-squared

People often interpret the R^2 as a measure of quality of the model.

But a model with a low R^2 may still be useful.

If the slope is significantly different from zero, the regression contains some predictive ability.



Prediction

Given an estimated slope $\hat{a}$ and intercept$\hat{b}$, we can compute a fitted $\hat{y}$ for any x:

$$\hat{y} = \hat{a} + \hat{b}x$$

x does not have to be **in** the original data

$\hat{y}$ is a *prediction* of the expected value of y at that value of x.

Predicted post moult sizes at pre moult sizes of 120, 140 and 160mm.

$$\begin{aligned} 25.803 + 0.914 * 120 &= 135.48\text{mm} \\ 25.803 + 0.914 * 140 &= 153.76\text{mm} \\ 25.803 + 0.914 * 160 &= 172.04\text{mm} \end{aligned}$$

Prediction

```
df <- read.csv('crabsmolt.csv')

m <- lm(postsz ~ presz, df)
predict(m,
        newdata = data.frame(presz = c(0,
                                       120,
                                       140,
                                       160)))
```

```
        1         2         3         4
 25.80258 135.47835 153.75765 172.03694
```

Confidence Interval for the mean of a predicted value

Again we can use bootstrapping to find a confidence interval for the predicted mean.

- Generate a bootstrap sample of pairs of data
- Fit the regression
- Make the prediction
- Repeat many times and construct an interval

For the crab moulting data at 120 the actual prediction is 135.48mm. A 95% confidence interval is 135.11, 135.87 mm.

Summary

- Simple linear regression fits the model $y = a + bx$ to the data by computing estimates of a and b.
- The best line is determined by a least-squares between the model line and the data.
- We can test if the slope $b = 0$ or any other value (e.g. $b = 1$).
- We can also compute the confidence interval for b.
- Examining residuals shows if the model is appropriate.
- R^2 measures the goodness of model fit.
- We can use the model to compute the expected y for a given x and provide a confidence interval.