Linear Regression
Chp 7

Chris Hallstrom

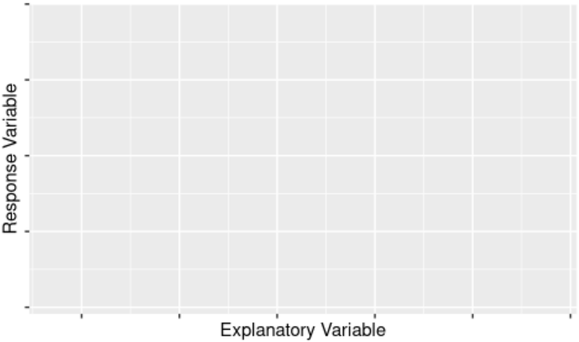
University of Portland

In groups

- Practice from last time: §7.5 (3, 7, 9, 19)
- Practice for next time: §7.5 ()

Associations between two numerical variables

- Explanatory variable: x, independent (a variable you think might be related to response)
- Response variable: y, dependent (a variable you want to understand)



Data: possum

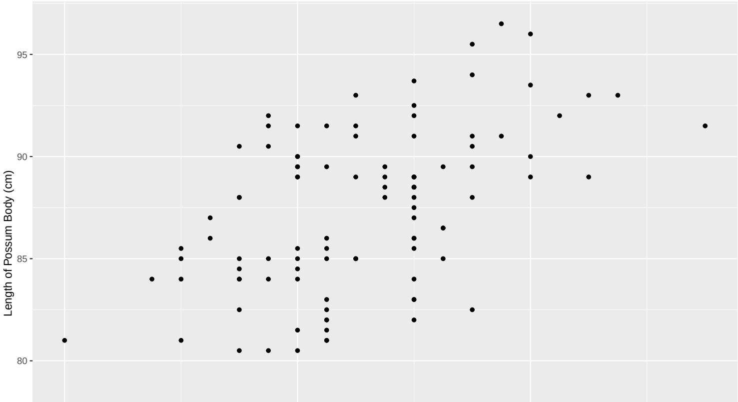
```
data(possum)

glimpse(possum)
```

```
Rows: 104
Columns: 8
$ site    <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
$ pop     <fct> Vic, Vic, Vic, Vic, Vic, Vic, Vic, Vic, Vi
$ sex     <fct> m, f, f, f, f, f, m, f, f, f, f, f, m, m, m
$ age     <int> 8, 6, 6, 6, 2, 1, 2, 6, 9, 6, 9, 5, 5, 3, 5
$ head_l  <dbl> 94.1, 92.5, 94.0, 93.2, 91.5, 93.1, 95.3, 9
$ skull_w <dbl> 60.4, 57.6, 60.0, 57.1, 56.3, 54.8, 58.2, 5
$ total_l <dbl> 89.0, 91.5, 95.5, 92.0, 85.5, 90.5, 89.5, 9
$ tail_l  <dbl> 36.0, 36.5, 39.0, 38.0, 36.0, 35.5, 36.0, 3
```

Possums: ‘tail_l’ vs ‘total_l’

Do possums’ tail length and total body length seem to be associated? Might knowing one help us predict the other?

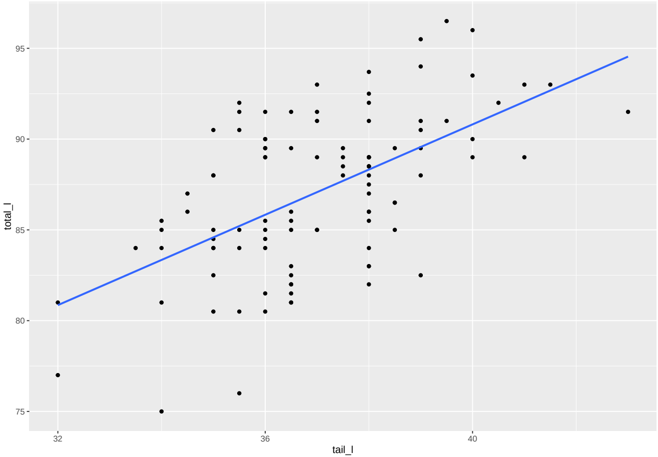
```
ggplot(data = possum, aes(y = total_l, x = tail_l)) +
  geom_point() +
   labs(x = "Length of Possum Tail (cm)",
       y = "Length of Possum Body (cm)")
```



Linear model: "line of best fit"

```r
ggplot(data = possum, aes(y = total_l, x = tail_l)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Finding the “least squares” regression line

Recall: a line has the equation

$$y = mx + b$$

where $m =$ slope and $b = y$-intercept. Here, we want a line of the form

$$\text{Total body length} = b_0 + b_1 \cdot (\text{Tail length})$$

How could you use this line to predict the total body length of a possum with tail 41.6 cm long?

Properties

The **least squares regression line** has the following properties:

- Minimizes (sum of squared) distance between data points and line
- The *residuals* balance out above and below line
- The point $(\bar{x}, \bar{y})$ always lies on line (though it's not necessarily a data point!)

Correlation coefficient, *r*

Also called the Pearson Product-Moment Correlation, here's how *r* is calculated:

$$r(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

This is a good computation to have R do for us!

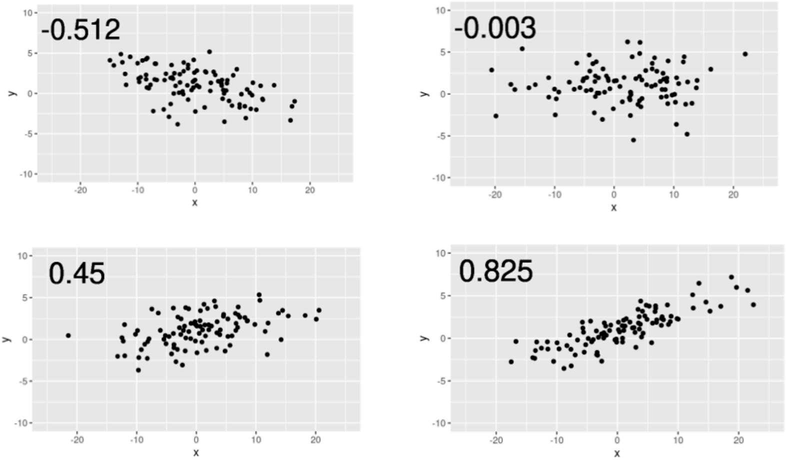
Correlation coefficient, r

```
possum |>
  summarize(N = n(),
        r = cor(tail_l, total_l,
                use = "pairwise.complete.obs"))
```

```
# A tibble: 1 x 2
      N     r
  <int> <dbl>
1   104 0.566
```

What does this number mean?

Visualizing the correlation coefficient, r



What would a correlation of $r = -0.998$ look like? How about $r = 1$?

Finding the "least squares" regression line

$$y = b_0 + b_1 \cdot x$$

First we find the slope:

$$b_1 = r\left(\frac{s_y}{s_x}\right)$$

where:

- r = correlation coefficient
- s_y = standard deviation of y
- s_x = standard deviation of x

Next we use the fact that $(\bar{x}, \bar{y})$ lies on our line to solve for the intercept, b_0:

$$\bar{y} = b_0 + b_1 \cdot \bar{x}$$

Example (continued)

To model the association between `tail_l` and `total_l`, we need summary statistics to find the slope and intercept of the linear regression:

```
possum |>
  summarize(
    mean_x = mean(tail_l),
    sd_x = sd(tail_l),
    mean_y = mean(total_l),
    sd_y = sd(total_l),
    r = cor(tail_l, total_l)
    )
```

```
# A tibble: 1 x 5
  mean_x  sd_x mean_y  sd_y     r
   <dbl> <dbl>  <dbl> <dbl> <dbl>
1   37.0  1.96   87.1  4.31 0.566
```

What line does this produce?

Example: Possum regression line

First, find the slope:

$$b_1 = r \cdot \left(\frac{s_y}{s_x}\right) = .566\left(\frac{4.31}{1.96}\right) = 1.24$$

Next, use $(\bar{x}, \bar{y})$ to find the intercept:

$$\bar{y} = b_0 + b_1 \cdot \bar{x} \quad \Longrightarrow \quad 87.1 = b_0 + 1.24(37.0)$$

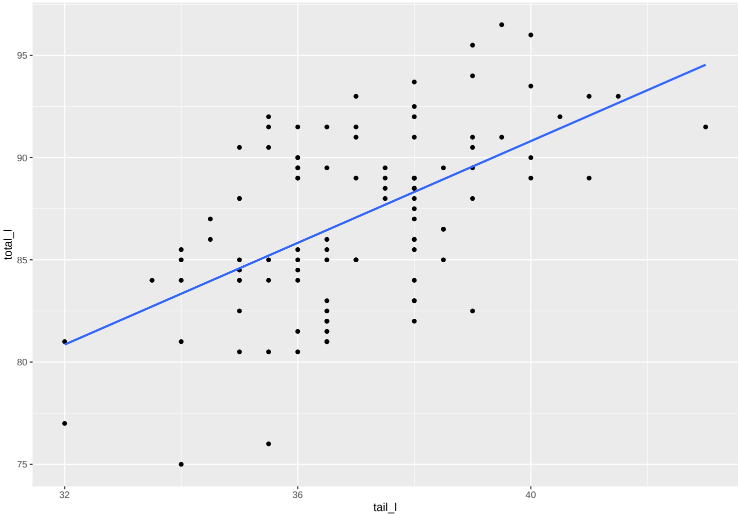
$$\Longrightarrow\ b_0 = 87.1 - 1.24(37.0) = 41.2$$

Thus: $y = 41.2 + 1.24x$ is the linear model describing these two possum variables.

Example: `Possum` regression line

$$y = 41.2 + 1.24x$$

```
`geom_smooth()` using formula = 'y ~ x'
```



Example: `Possum` regression line

Alternately, we can find the entire **linear model** in one go:

```r
library(tidymodels)
```

```
-- Attaching packages ------------------------------------

v broom        1.0.6     v rsample      1.2.1
v dials        1.2.1     v tune         1.2.1
v infer        1.0.7     v workflows    1.1.4
v modeldata    1.3.0     v workflowsets 1.1.0
v parsnip      1.2.1     v yardstick    1.3.1
v recipes      1.0.10

-- Conflicts ---------------------------------------------
x scales::discard() masks purrr::discard()
x dplyr::filter()   masks stats::filter()
x recipes::fixed()  masks stringr::fixed()
x dplyr::lag()      masks stats::lag()
x yardstick::spec() masks readr::spec()
x recipes::step()   masks stats::step()
```

Practice