Parallel slopes linear regression

INTERMEDIATE REGRESSION IN R

Richie Cotton
Curriculum Architect at DataCamp


INTRODUCTION TO
REGRESSION IN R

From simple regression to multiple regression

Multiple regression is a regression model with more than one explanatory variable.

More explanatory variables can give **more insight** and **better predictions**.

The course contents

Chapter 1

- "Parallel slopes" regression

Chapter 2

- Interactions
- Simpson's Paradox

Chapter 3

- More explanatory variables
- How linear regression works

Chapter 4

- Multiple logistic regression
- The logistic distribution
- How logistic regression works

The fish dataset

mass_g	length_cm	species
242.0	23.2	Bream
5.9	7.5	Perch
200.0	30.0	Pike
40.0	12.9	Roach

- Each row represents a fish
- `mass_g` is the response variable
- 1 numeric, 1 categorical explanatory variable

One explanatory variable at a time

```
mdl_mass_vs_length <- lm(mass_g ~ length_cm, data = fish)
```

```
Call:
lm(formula = mass_g ~ length_cm, data = fish)

Coefficients:
(Intercept)    length_cm
     -536.2         34.9
```

- 1 intercept coefficient
- 1 slope coefficient

```
mdl_mass_vs_species <- lm(mass_g ~ species + 0, data = fish)
```

```
Call:
lm(formula = mass_g ~ species + 0, data = fish)

Coefficients:
speciesBream  speciesPerch   speciesPike  speciesRoach
       617.8         382.2         718.7         152.0
```

- 1 intercept coefficient for each category

Both variables at same time

```
mdl_mass_vs_both <- lm(mass_g ~ length_cm + species + 0, data = fish)
```

```
Call:
lm(formula = mass_g ~ length_cm + species + 0, data = fish)

Coefficients:
   length_cm  speciesBream  speciesPerch   speciesPike  speciesRoach
       42.57       -672.24       -713.29      -1089.46       -726.78
```

- 1 slope coefficient
- 1 intercept coefficient for each category

Comparing coefficients

```
coefficients(mdl_mass_vs_length)
```

```
(Intercept)   length_cm
     -536.2        34.9
```

```
coefficients(mdl_mass_vs_species)
```

```
speciesBream speciesPerch  speciesPike speciesRoach
       617.8        382.2        718.7        152.0
```

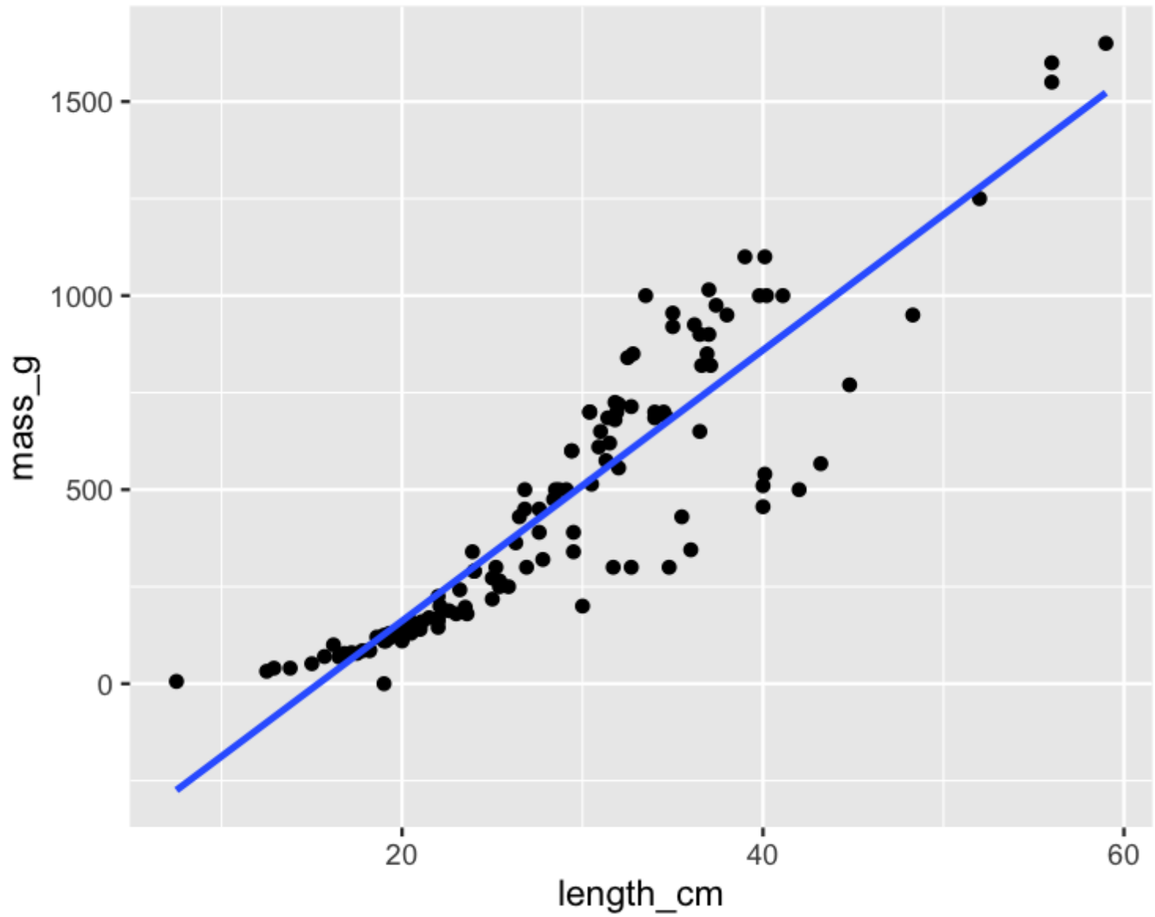
```
coefficients(mdl_mass_vs_both)
```

```
length_cm speciesBream speciesPerch  speciesPike speciesRoach
    42.57      -672.24      -713.29     -1089.46      -726.78
```

Visualization: 1 numeric explanatory var

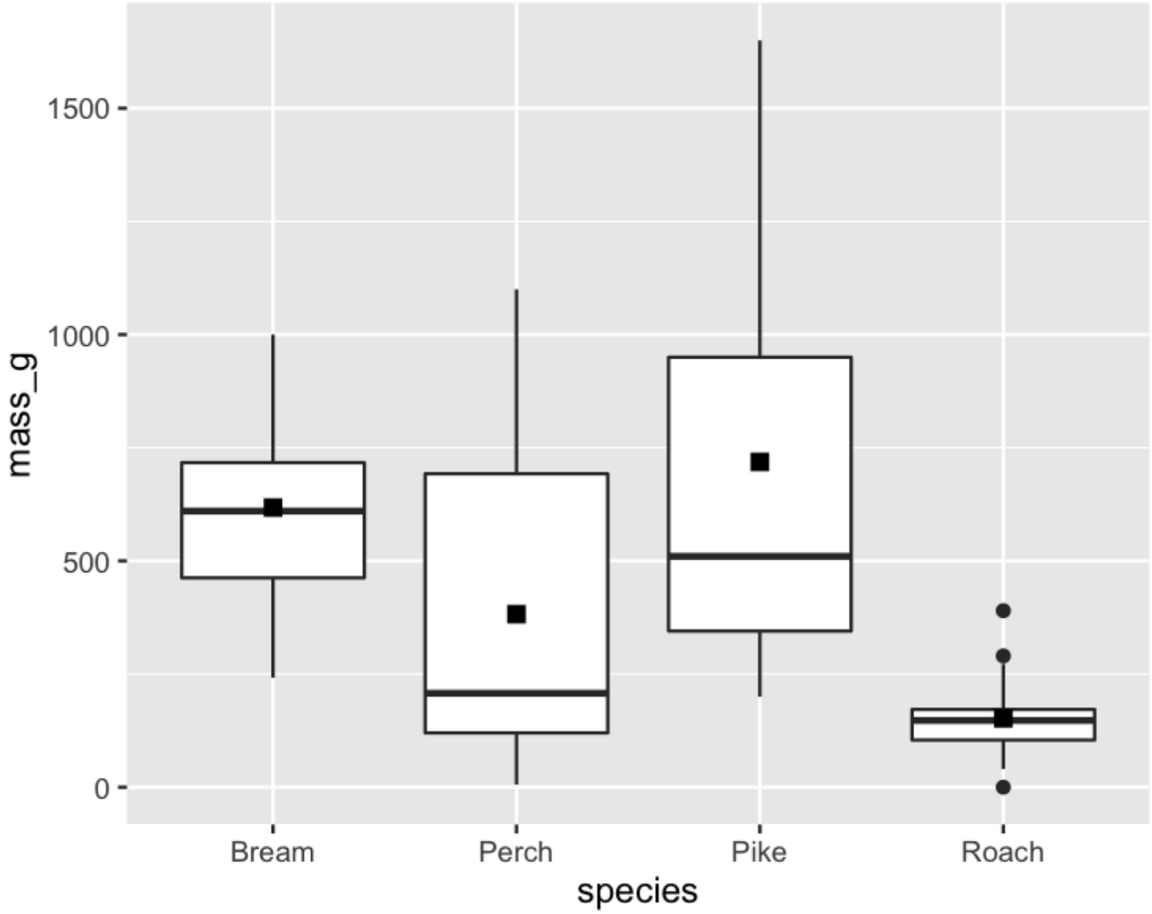
```
library(ggplot2)

ggplot(fish, aes(length_cm, mass_g)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



Visualization: 1 categorical explanatory var

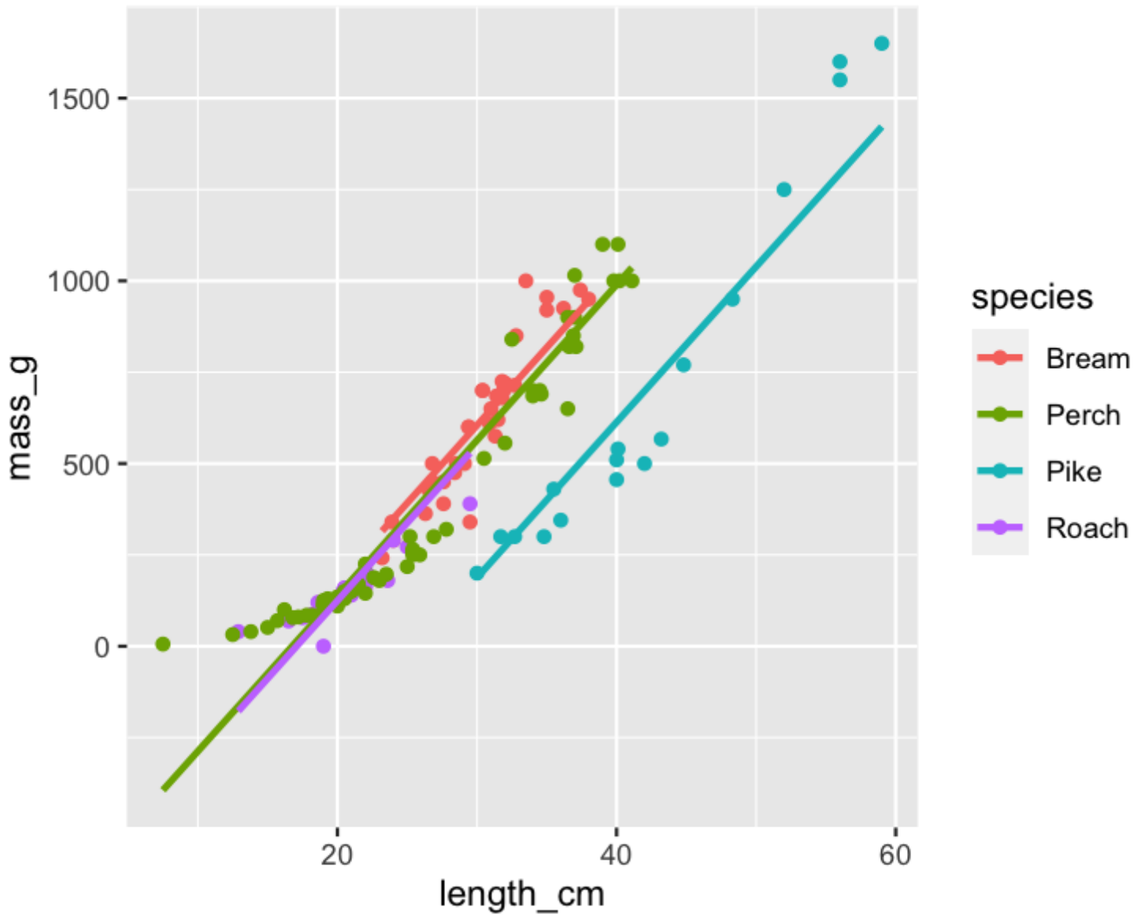
```
ggplot(fish, aes(species, mass_g)) +
  geom_boxplot() +
  stat_summary(fun.y = mean, shape = 15)
```



Visualization: both explanatory vars

```
library(moderndive)

ggplot(fish, aes(length_cm, mass_g, color = species)) +
  geom_point() +
  geom_parallel_slopes(se = FALSE)
```



Let's practice!

INTERMEDIATE REGRESSION IN R

Predicting parallel slopes

INTERMEDIATE REGRESSION IN R

Richie Cotton
Curriculum Architect

The prediction workflow 1

```
library(dplyr)

explanatory_data <- tibble(
  length_cm = seq(5, 60, 5)
)
```

```
glimpse(explanatory_data)
```

```
Rows: 12
Columns: 1
$ length_cm <dbl> 5, 10, 15, 20, 25, 30, 35, 40.
```

```
library(dplyr)
library(tidyr)
explanatory_data <- expand_grid(
  length_cm = seq(5, 60, 5),
  species = unique(fish$species)
)
```

```
glimpse(explanatory_data)
```

```
Rows: 48
Columns: 2
$ length_cm <dbl> 5, 5, 5, 5, 10, 10, 10, 10, 1.
$ species   <chr> "Bream", "Roach", "Perch", "P.
```

The prediction workflow 2

```
library(dplyr)

explanatory_data <- tibble(
  length_cm = seq(5, 60, 5)
)
```

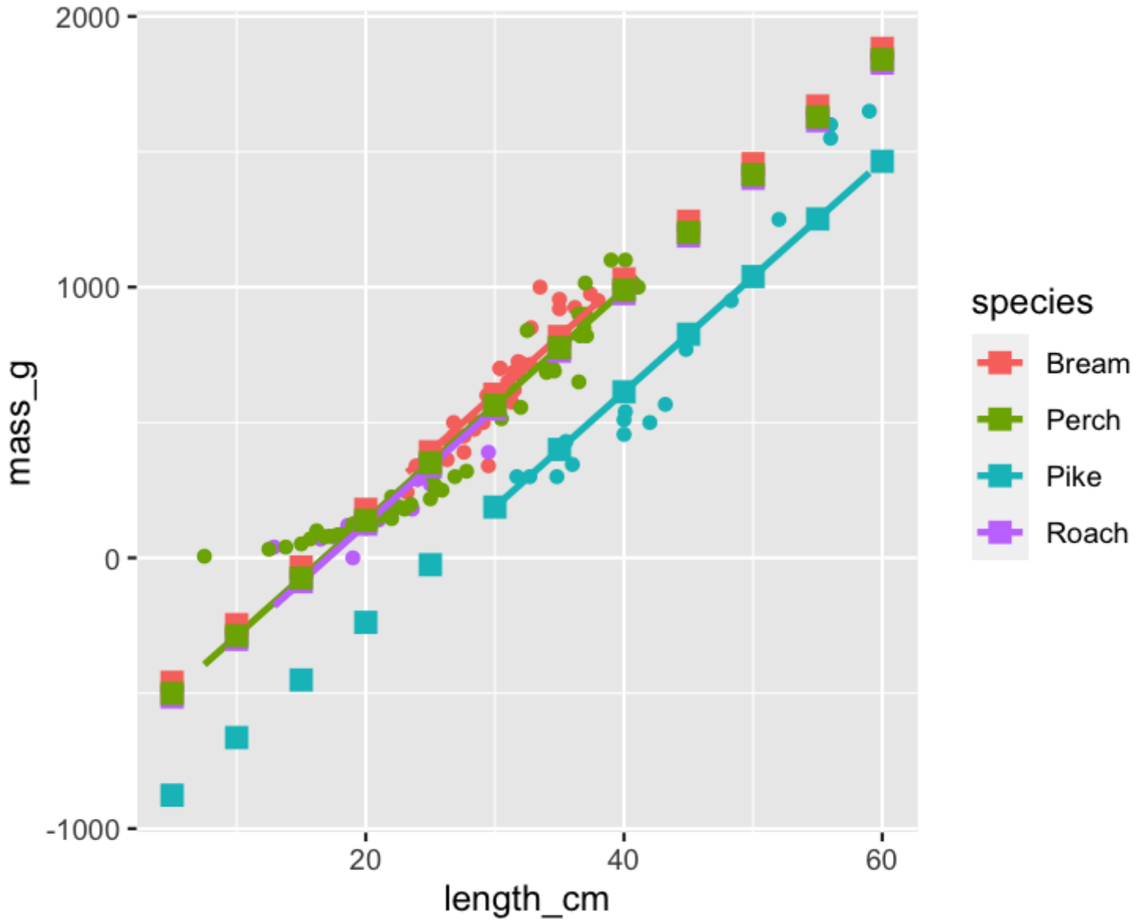
```
library(dplyr)
library(tidyr)
explanatory_data <- expand_grid(
  length_cm = seq(5, 60, 5),
  species = unique(fish$species)
)
```

```
prediction_data <- explanatory_data %>%
  mutate(
    mass_g = predict(
      mdl_mass_vs_length, explanatory_data
    )
  )
```

```
prediction_data <- explanatory_data %>%
  mutate(
    mass_g = predict(
      mdl_mass_vs_both, explanatory_data
    )
  )
```

Visualizing the predictions

```
library(ggplot2)
library(moderndive)
ggplot(fish, aes(length_cm, mass_g, color = species)) +
  geom_point() +
  geom_parallel_slopes(se = FALSE) +
  geom_point(
    data = prediction_data,
    size = 3, shape = 15
  )
```



Manually calculating predictions

```
coeffs <- coefficients(mdl_price_vs_length
```

```
(Intercept)   length_cm
     -536.2         34.9
```

```
intercept <- coeffs[1]
slope <- coeffs[2]
```

```
explanatory_data %>%
  mutate(
    mass_g = intercept + slope * length_cm
  )
```

length_cm	mass_g
5	-361.73
10	-187.23
15	-12.74
20	161.76
25	336.26
30	510.75

Coefficients for parallel slopes

```
coefficients(mdl_mass_vs_both)
```

```
length_cm speciesBream speciesPerch  speciesPike speciesRoach
    42.57      -672.24      -713.29     -1089.46      -726.78
```

```
slope <- coeffs[1]
intercept_bream <- coeffs[2]
intercept_perch <- coeffs[3]
intercept_pike <- coeffs[4]
intercept_roach <- coeffs[5]
```

Choosing an intercept with ifelse()

```
explanatory_data %>%
  mutate(
    intercept = ifelse(
      species == "Bream",
      intercept_bream,
      ifelse(
        species == "Perch",
        intercept_perch,
        ifelse(
          species == "Pike",
          intercept_pike,
          intercept_roach
        )
      )
    )
  )
```

case_when()

```
dataframe %>%
  mutate(
    case_when(
      condition_1 ~ value_1,
      condition_2 ~ value_2,
      # ...
      condition_n ~ value_n
    )
  )
```

Choosing an intercept with case_when()

```
explanatory_data %>%
  mutate(
    intercept = case_when(
      species == "Bream" ~ intercept_bream,
      species == "Perch" ~ intercept_perch,
      species == "Pike" ~ intercept_pike,
      species == "Roach" ~ intercept_roach
    )
  )
```

The final prediction step

```
explanatory_data %>%
  mutate(
    intercept = case_when(
      species == "Bream" ~ intercept_bream,
      species == "Perch" ~ intercept_perch,
      species == "Pike" ~ intercept_pike,
      species == "Roach" ~ intercept_roach
    ),
    mass_g = intercept + slope * length_cm
  )
```

```
# A tibble: 48 x 4
   length_cm species intercept mass_g
       <dbl> <chr>       <dbl>  <dbl>
 1         5 Bream       -672. -459.
 2         5 Roach       -727. -514.
 3         5 Perch       -713. -500.
 4         5 Pike       -1089. -877.
 5        10 Bream       -672. -247.
 6        10 Roach       -727. -301.
 7        10 Perch       -713. -288.
 8        10 Pike       -1089. -664.
 9        15 Bream       -672.  -33.7
10        15 Roach       -727.  -88.2
# ... with 38 more rows
```

Compare to predict()

```
predict(mdl_mass_vs_both, explanatory_data)
```

```
         1          2          3          4
-459.39910 -513.93503 -500.45009 -876.61328
         5          6          7          8
-246.55633 -301.09226 -287.60732 -663.77051
# ...
```

```
# A tibble: 48 x 4
   length_cm species intercept mass_g
       <dbl> <chr>       <dbl>  <dbl>
 1         5 Bream       -672. -459.
 2         5 Roach       -727. -514.
 3         5 Perch       -713. -500.
 4         5 Pike       -1089. -877.
 5        10 Bream       -672. -247.
 6        10 Roach       -727. -301.
 7        10 Perch       -713. -288.
 8        10 Pike       -1089. -664.
 9        15 Bream       -672.  -33.7
10        15 Roach       -727.  -88.2
# ... with 38 more rows
```

Let's practice!

INTERMEDIATE REGRESSION IN R

Assessing model performance

INTERMEDIATE REGRESSION IN R

Richie Cotton
Curriculum Architect at DataCamp

Model performance metrics

- *Coefficient of determination (R-squared)*: how well the linear regression line fits the observed values.
  - Larger is better.
- *Residual standard error (RSE)*: the typical size of the residuals.
  - Smaller is better.

Getting the coefficient of determination

```
library(dplyr)
library(broom)
```

```
mdl_mass_vs_length %>%
  glance() %>%
  pull(r.squared)
```

```
0.8226
```

```
mdl_mass_vs_species %>%
  glance() %>%
  pull(r.squared)
```

```
0.7163
```

```
mdl_mass_vs_both %>%
  glance() %>%
  pull(r.squared)
```

```
0.9694
```

Adjusted coefficient of determination

- More explanatory variables increases R^2.
- Too many explanatory variables causes overfitting.
- *Adjusted coefficient of determination* penalizes more explanatory variables.
- $\bar{R}^2 = 1 - (1 - R^2)\frac{n_{obs} - 1}{n_{obs} - n_{var} - 1}$
- Penalty is noticeable when R^2 is small, or n_{var} is large fraction of n_{obs}.
- In `glance()`, it's the `adj.r.squared` element.

Getting the adjusted coefficient of determination

```
library(dplyr)
library(broom)
```

```
mdl_mass_vs_length %>%
  glance() %>%
  select(r.squared, adj.r.squared)
```

```
  r.squared adj.r.squared
      <dbl>         <dbl>
1    0.8226        0.8212
```

```
mdl_mass_vs_species %>%
  glance() %>%
  select(r.squared, adj.r.squared)
```

```
  r.squared adj.r.squared
      <dbl>         <dbl>
1    0.7163        0.7072
```

```
mdl_mass_vs_both %>%
  glance() %>%
  select(r.squared, adj.r.squared)
```

```
  r.squared adj.r.squared
      <dbl>         <dbl>
1    0.9694        0.9682
```

Getting the residual standard error

```
library(dplyr)
library(broom)
```

```
mdl_mass_vs_length %>%
  glance() %>%
  pull(sigma)
```

```
152.1
```

```
mdl_mass_vs_species %>%
  glance() %>%
  pull(sigma)
```

```
313.6
```

```
mdl_mass_vs_both %>%
  glance() %>%
  pull(sigma)
```

```
103.4
```

Let's practice!

INTERMEDIATE REGRESSION IN R