UNIVERSITY of WEST FLORIDA

Interaction Terms in Regression Models

Statistics for Data Science II

Introduction

Recall interactions from two-way ANOVA:

> The relationship between the outcome and one predictor depends on the level of another predictor.

Interactions work (and are specified) the same way in regression.

The usual caveats apply:

> We do not want to load models with too many interactions.
>
> We favor simplicity over interactions that do not add much to the predictive power of the model.
>
> We do not want higher than two-way interactions unless necessary.

Types of Interactions

The easiest interactions to deal with are categorical $\times$ continuous interactions.

- The continuous predictor automatically is assigned to the x-axis when graphing.

We can also deal with categorical $\times$ categorical interactions.

- If one categorical variable is ordinal, that should be assigned to the x-axis when graphing.

Finally, we can also have categorical $\times$ categorical interactions.

- This can be tricky because we do not have easily-defined levels. Either can be assigned to the x-axis when graphing.

Modeling

We will construct what is called a hierarchical well-formulated (HWF) model.

This means that when a higher-order interaction term is included in the model, all lower-order terms are also included.

e.g., when a two-way interaction is included, we also include the corresponding main effects.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

e.g., when a three-way interaction is included, we also include the corresponding main effects and two-way interactions.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \beta_7 x_1 x_2 x_3$$

Modeling

Example:

Recall the data from the `palmerpenguin` package.

Let us now consider the following interactions:

- flipper length and species
- species and sex
- flipper length and body mass

For simplicity / example's purpose, let us consider them one at a time.

Modeling: Continuous × Categorical

Example:

First, let's look at the interaction between flipper length (continuous) and species (categorical).

```
m1 <- lm(bill_length_mm ~ flipper_length_mm + species_Chinstrap +
           species_Gentoo + species_Chinstrap:flipper_length_mm +
           species_Gentoo:flipper_length_mm, data = data)

coefficients(m1)
```

```
##                          (Intercept)                    flipper_length_mm
##                             13.03578                              0.13565
##                       species_Chinstrap                       species_Gentoo
##                             -7.44240                             -33.52367
## flipper_length_mm:species_Chinstrap    flipper_length_mm:species_Gentoo
##                              0.08516                              0.17763
```

Modeling: Continuous $\times$ Categorical

Example:

The model is

$$\hat{y} = 13.04 + 0.14\text{flipper} - 7.44\text{Chinstrap} - 33.52\text{Gentoo} + 0.09(\text{flipper} \times \text{Chinstrap}) + 0.18(\text{flipper} \times \text{Gentoo})$$

We can separate this into different models for species:

$$\text{Chinstraps (Chinstrap} = 1, \text{ Gentoo} = 0) : \hat{y} = 5.59 + 0.22\text{flipper}$$
$$\text{Gentoos (Chinstrap} = 0, \text{ Gentoo} = 1) : \hat{y} = -20.49 + 0.31\text{flipper}$$
$$\text{Adelies (Chinstrap} = 0, \text{ Gentoo} = 0) : \hat{y} = 13.04 + 0.14\text{flipper}$$

Modeling: Continuous $\times$ Categorical

Example:

We can see that the slope describing the relationship between bill length and flipper length depend on the species of penguin:

Adelie: $0.14 \times$ flipper

Chinstrap: $0.22 \times$ flipper

Gentoo: $0.31 \times$ flipper

Thus, Adelies have the smallest slope of the three species, while Gentoos have the steepest slope of the three species.

Modeling: Categorical × Categorical

Example:

Let us now look at the interaction between species and sex.

```
m2 <- lm(bill_length_mm ~ sex_male + species_Chinstrap +  species_Gentoo +
         species_Chinstrap:sex_male + species_Gentoo:sex_male, data = data)

coefficients(m2)
```

```
##                 (Intercept)                    sex_male
##                     37.2575                      3.1329
##           species_Chinstrap                species_Gentoo
##                      9.3160                      8.3063
## sex_male:species_Chinstrap    sex_male:species_Gentoo
##                      1.3877                      0.7771
```

Modeling: Categorical $\times$ Categorical

Example:

The model is

$$\hat{y} = 37.26 + 3.13\text{male} + 9.32\text{Chinstrap} + 8.31\text{Gentoo} + 1.39\text{male} \times \text{Chinstrap} + 0.78\text{male} \times \text{Gentoo}$$

We can separate this into different models for species:

$$\begin{aligned}
\text{Chinstrap } (\text{Chinstrap} = 1,\ \text{Gentoo} = 0) &: \hat{y} = 46.57 + 4.52\text{male} \\
\text{Gentoos } (\text{Chinstrap} = 0,\ \text{Gentoo} = 1) &: \hat{y} = 45.56 + 3.91\text{male} \\
\text{Adelies } (\text{Chinstrap} = 0,\ \text{Gentoo} = 0) &: \hat{y} = 37.26 + 3.13\text{male}
\end{aligned}$$

Males, on average, have larger bill sizes than females. The difference is largest in Chinstraps and smallest in Adelies.

Modeling: Categorical $\times$ Categorical

Example:

The model is

$$\hat{y} = 37.26 + 3.13\text{male} + 9.32\text{Chinstrap} + 8.31\text{Gentoo} + 1.39\text{male} \times \text{Chinstrap} + 0.78\text{male} \times \text{Gentoo}$$

We can also separate this into different models for sex:

$$\text{Males (male} = 1) : \hat{y} = 40.39 + 10.70\text{Chinstrap} + 9.08\text{Gentoo}$$
$$\text{Females (male} = 0) : \hat{y} = 37.26 + 9.32\text{Chinstrap} + 8.31\text{Gentoo}$$

Chinstraps and Gentoos both have longer bill lengths than Adelies; we can also see that males have longer bill lengths than females.

Modeling: Continuous × Continuous

Example:

Finally, let's look at the interaction between flipper length and body mass.

```
m3 <- lm(bill_length_mm ~ flipper_length_mm + body_mass_g + flipper_length_mm:body_mass_g,
         data = data)

coefficients(m3)
```

```
##                   (Intercept)             flipper_length_mm
##                    -2.675e+01                     3.394e-01
##                   body_mass_g flipper_length_mm:body_mass_g
##                     5.864e-03                    -2.587e-05
```

Modeling: Continuous $\times$ Continuous

Example:

The model is

$$\hat{y} = -26.75 + 0.34\text{flipper} + 0.01\text{body mass} - 0.00003(\text{flipper} \times \text{body mass})$$

Both flipper length and body mass have positive relationships with bill length, however, as the other variable increases, the slope decreases.

As body mass increases, bill length increases. However, as flipper length increases, the slope for body mass decreases slightly.

As flipper length increases, bill length increases. However, as body mass increases, the slope for flipper length decreases slightly.

Modeling: Continuous $\times$ Continuous

Example:

If we are interested, we can plug in values (my choices: 25th percentile, median, 75th percentile) for either continuous predictor to construct models as we did for categorical variables.

For flipper size,

$$\begin{aligned}
\text{25th percentile (flipper} = 190 \text{ mm)} &: \hat{y} = 37.73 + 0.001\text{body} \\
\text{median (flipper} = 197 \text{ mm)} &: \hat{y} = 40.11 + 0.0008\text{body} \\
\text{75th percentile (flipper} = 213 \text{ mm)} &: \hat{y} = 45.54 + 0.0004\text{body}
\end{aligned}$$

Modeling: Continuous $\times$ Continuous

Example:

If we are interested, we can plug in values (my choices: 25th percentile, median, 75th percentile) for either continuous predictor to construct models as we did for categorical variables.

For body mass,

$$\begin{aligned}
\text{25th percentile (body mass} = 3550 \text{ g)} &: \hat{y} = -5.94 + 0.25\text{flipper} \\
\text{median (body mass} = 4050 \text{ g)} &: \hat{y} = -3.01 + 0.24\text{flipper} \\
\text{75th percentile (body mass} = 4750 \text{ g)} &: \hat{y} = 1.10 + 0.22\text{flipper}
\end{aligned}$$

Testing

Largely, testing works the same way as we've tested predictors before.

If we have a categorical variable with more than 2 levels in an interaction term, we must use a multiple partial F test to determine if the interaction is significant.

Otherwise, we will use the t-test output from the `summary()` function to determine if interactions are significant.

Testing

Example:

Let's go back to our first model: bill length as a function of flipper length, species, and the interaction between flipper length and species.

```
summary(m1)[4]
```

```
## $coefficients
##                                          Estimate Std. Error t value   Pr(>|t|)
## (Intercept)                             13.03578    6.18399  2.1080 0.03579261
## flipper_length_mm                        0.13565    0.03251  4.1726 0.00003864
## species_Chinstrap                       -7.44240   10.56827 -0.7042 0.48179619
## species_Gentoo                         -33.52367    9.92017 -3.3793 0.00081446
## flipper_length_mm:species_Chinstrap      0.08516    0.05450  1.5627 0.11909443
## flipper_length_mm:species_Gentoo         0.17763    0.04828  3.6793 0.00027343
```

We need to use a multiple partial F to determine if the interaction is significant.

Testing

Example:

```
full <- lm(bill_length_mm ~ flipper_length_mm + species_Chinstrap +  species_Gentoo + species_Chinstrap:flipper_length_mm +
           species_Gentoo:flipper_length_mm, data = data)

reduced <- lm(bill_length_mm ~ flipper_length_mm + species_Chinstrap + species_Gentoo, data = data)

anova(reduced, full)
```

```
## Analysis of Variance Table
##
## Model 1: bill_length_mm ~ flipper_length_mm + species_Chinstrap + species_Gentoo
## Model 2: bill_length_mm ~ flipper_length_mm + species_Chinstrap + species_Gentoo +
##     species_Chinstrap:flipper_length_mm + species_Gentoo:flipper_length_mm
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1    329 2220
## 2    327 2132  2      88.3 6.77 0.0013 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The interaction (overall) is significant ($p = 0.001$). Now, we can look at the individual terms

Testing

Example:

```
summary(m1)[4]
```

```
## $coefficients
##                                            Estimate Std. Error t value   Pr(>|t|)
## (Intercept)                                13.03578    6.18399  2.1080 0.03579261
## flipper_length_mm                           0.13565    0.03251  4.1726 0.00003864
## species_Chinstrap                          -7.44240   10.56827 -0.7042 0.48179619
## species_Gentoo                            -33.52367    9.92017 -3.3793 0.00081446
## flipper_length_mm:species_Chinstrap         0.08516    0.05450  1.5627 0.11909443
## flipper_length_mm:species_Gentoo            0.17763    0.04828  3.6793 0.00027343
```

Flipper length slopes are different between Gentoos and Adelies ($p < 0.001$), but not Chinstraps and Adelies ($p = 0.119$).

Testing

Example:

Let's now look at the second model: bill length as a function of sex, species, and the interaction between sex and species.

```
summary(m2)[4]
```

```
## $coefficients
##                                Estimate Std. Error t value   Pr(>|t|)
## (Intercept)                     37.2575     0.2710 137.473 2.300e-291
## sex_male                         3.1329     0.3833   8.174  6.636e-15
## species_Chinstrap                9.3160     0.4808  19.377  3.059e-56
## species_Gentoo                   8.3063     0.4073  20.393  3.179e-60
## sex_male:species_Chinstrap       1.3877     0.6799   2.041  4.206e-02
## sex_male:species_Gentoo          0.7771     0.5721   1.358  1.753e-01
```

Again, we must use a multiple partial F to determine if the interaction is significant.

Testing

Example:

```
full <- lm(bill_length_mm ~ sex_male + species_Chinstrap +  species_Gentoo + species_Chinstrap:sex_male +
           species_Gentoo:sex_male, data = data)

reduced <- lm(bill_length_mm ~ sex_male + species_Chinstrap + species_Gentoo, data = data)

anova(reduced, full)
```

```
## Analysis of Variance Table
##
## Model 1: bill_length_mm ~ sex_male + species_Chinstrap + species_Gentoo
## Model 2: bill_length_mm ~ sex_male + species_Chinstrap + species_Gentoo +
##     species_Chinstrap:sex_male + species_Gentoo:sex_male
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1    329 1778
## 2    327 1753  2      24.5 2.28    0.1
```

The interaction (overall) is not significant ($p = 0.100$). We have no reason to look at the individual terms.

Testing

Example:

Let's now look at the third model: bill length as a function of flipper length, body mass, and the interaction between flipper length and body mass.

```
summary(m3)[4]
```

```
## $coefficients
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -2.675e+01  2.115e+01  -1.265 0.206774
## flipper_length_mm               3.394e-01  1.069e-01   3.175 0.001642
## body_mass_g                     5.864e-03  4.844e-03   1.210 0.226967
## flipper_length_mm:body_mass_g -2.587e-05  2.342e-05  -1.105 0.270144
```

Because we have a continuous $\times$ continous interaction term, we do not need to use a multiple partial F test – we can just use the t-test above.

The interaction between flipper length and body mass is not significant ($p = 0.270$).

Testing

Recall that we do not interpret main effects when interaction terms are included in the model.

In our first example, we would not be able to interpret main effects because the interaction term was significant.

In our second and third example, we would be able to interpret the main effects because the interaction terms were not significant.

> Because we are interested in parsimony, we would remove the interaction term(s) from the models and obtain new estimates for the main effects.

Visualization

Let us consider visualizing the models we have already constructed:

- flipper length, species, interaction
  - flipper length on x-axis, lines for species
- sex, species, interaction
  - neither of the predictors make sense on the x-axis
- flipper length, body mass, interaction
  - flipper length on x-axis, body mass held constant
  - body mass on x-axis, flipper length held constant

Visualization

Example:

Let's build a visualization for the first model. This requires predicted values.

Flipper length is the only continuous variable, so it will be on the x-axis.

We will plot a line for each of the species. This requires separate columns for prediction purposes.

```
data <- data %>% mutate(
          m1c = 5.59 + 0.22*data$flipper_length_mm,
          m1g = -20.49 + 0.31*data$flipper_length_mm,
          m1a = 13.04 + 0.14*data$flipper_length_mm
        )
```

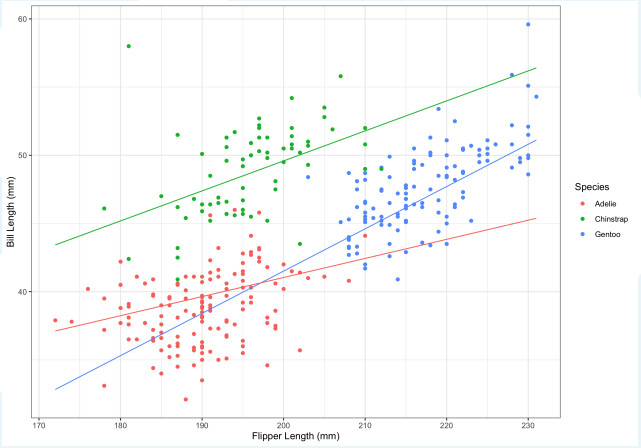
Visualization

Example:

```
p1 <- ggplot(data, aes(x=flipper_length_mm, y=bill_length_mm, color=species)) +
        geom_point() +
        geom_line(aes(y = m1c), color = "#00BA38", linetype = "solid") +
        geom_line(aes(y = m1g), color = "#619CFF", linetype = "solid") +
        geom_line(aes(y = m1a), color = "#F8766D", linetype = "solid") +
        xlab("Flipper Length (mm)") +
        ylab("Bill Length (mm)") +
        scale_color_discrete(name = "Species") +
        theme_bw()
```

Visualization

Example:



Visualization

Example:

Let's now build a visualization for the third model.

Flipper length and body mass are both continuous variables. For demonstration purposes we will build two visualizations: one with flipper length on the x-axis and another with body mass on the x-axis.

We again need predicted values,

```
data <- data %>% mutate(
          m3f25 = 37.73 + 0.001*body_mass_g,
          m3f50 = 40.11 + 0.0008*body_mass_g,
          m3f75 = 45.54 + 0.0004*body_mass_g,
          m3bm25 = -5.94 + 0.25*flipper_length_mm,
          m3bm50 = -3.01 + 0.23*flipper_length_mm,
          m3bm75 = 1.10 + 0.22*flipper_length_mm
        )
```

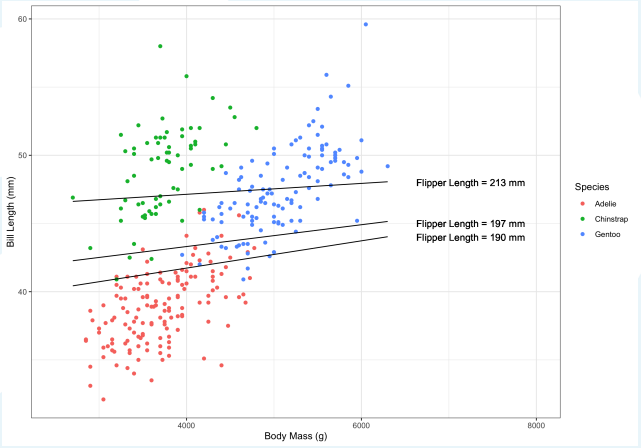
Visualization

Example:

```
p2 <- ggplot(data, aes(x=body_mass_g, y=bill_length_mm, color=species)) +
        geom_point() +
        geom_line(aes(y = m3f25), color = "black", linetype = "solid") +
        geom_text(aes(x = 7250, y = 44, label = "Flipper Length = 190 mm"), color="black", show.legend = FALSE) +
        geom_line(aes(y = m3f50), color = "black", linetype = "solid") +
        geom_text(aes(x = 7250, y = 45, label = "Flipper Length = 197 mm"), color="black", show.legend = FALSE) +
        geom_line(aes(y = m3f75), color = "black", linetype = "solid") +
        geom_text(aes(x = 7250, y = 48, label = "Flipper Length = 213 mm"), color="black", show.legend = FALSE) +
        xlab("Body Mass (g)") + xlim(2500,8000) +
        ylab("Bill Length (mm)") +
        scale_color_discrete(name = "Species") +
        theme_bw()
```

Visualization

Example:



Visualization

Example:

```
p3 <- ggplot(data, aes(x=flipper_length_mm, y=bill_length_mm, color=species)) +
        geom_point() +
        geom_line(aes(y = m3bm25), color = "black", linetype = "solid") +
        geom_text(aes(x = 240, y = 51, label = "Body Mass = 3550 g"), color="black", show.legend = FALSE) +
        geom_line(aes(y = m3bm50), color = "black", linetype = "solid") +
        geom_text(aes(x = 240, y = 50, label = "Body Mass = 4050 g"), color="black", show.legend = FALSE) +
        geom_line(aes(y = m3bm75), color = "black", linetype = "solid") +
        geom_text(aes(x = 240, y = 52, label = "Body Mass = 4750 g"), color="black", show.legend = FALSE) +
        xlab("Flipper Length (mm)") + xlim(170, 250) +
        ylab("Bill Length (mm)") +
        scale_color_discrete(name = "Species") +
        theme_bw()
```

Visualization

Example: