UNIVERSITY *of* WEST FLORIDA

Review of General Linear Models

Statistics for Data Science II

Introduction

Recall the general linear model,

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

This is a multiple regression model because it has multiple predictors (x_i).

A special case is simple linear regression, when there is a single predictor.

β_0 is the y-intercept, or the average outcome (y) when all $x_i = 0$.

β_i is the slope for predictor i and describes the relationship between the predictor and the outcome, after adjusting (or accounting) for the other predictors in the model.

Constructing the Model in R

We will use the lm() function to construct the linear model,

```
m <- lm([outcome] ~ [pred1] + [pred2] + [pred3] + ...,
        data = [dataset])
```

Then we run the model results through the summary() function to obtain information about the model,

```
summary(m)
```

Constructing the Model in R

Example

Consider the data from the palmerpenguin package. Let's create a dataset with the variables body_mass_g, bill_length_mm, and flipper_length_mm.

```
data <- as_tibble(penguins %>% select(body_mass_g,
                                      bill_length_mm,
                                      flipper_length_mm))
head(data)
```

```
## # A tibble: 6 x 3
##   body_mass_g bill_length_mm flipper_length_mm
##         <int>          <dbl>             <int>
## 1        3750           39.1               181
## 2        3800           39.5               186
## 3        3250           40.3               195
## 4          NA           NA                  NA
## 5        3450           36.7               193
## 6        3650           39.3               190
```

Constructing the Model in R

```r
m1 <- lm(bill_length_mm ~ body_mass_g + flipper_length_mm,
         data = data)
summary(m1)
```

```
##
## Call:
## lm(formula = bill_length_mm ~ body_mass_g + flipper_length_mm,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8064 -2.5898 -0.7053  1.9911 18.8288
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -3.4366939  4.5805532  -0.750    0.454
## body_mass_g        0.0006622  0.0005672   1.168    0.244
## flipper_length_mm  0.2218655  0.0323484   6.859 3.31e-11 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.124 on 339 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.4329, Adjusted R-squared:  0.4295
## F-statistic: 129.4 on 2 and 339 DF,  p-value: < 2.2e-16
```

Interpretation of Slope

We want to put the slope into perspective for whoever we are collaborating with.

Basic interpretation: for every 1 [units of x_i] increase in [x_i], [y] [increases or decreases] by $\left[\left|\hat{\beta}_i\right|\right]$ [units of y].

We say that y is decreasing if $\hat{\beta}_0 < 0$ and y is increasing if $\hat{\beta}_0 > 0$.

We can also scale our interpretations. e.g.,

> For every 7 [units of x_i] increase in [x_i], [y] [increases or decreases] by $\left[7 \times \left|\hat{\beta}_i\right|\right]$ [units of y].

Interpretation of Slope

Example:

```
coefficients(m1)
```

```
##       (Intercept)       body_mass_g flipper_length_mm
##     -3.4366939266      0.0006622186      0.2218654584
```

For a 1 gram increase in body mass, we expect bill length to increase by 0.0007 mm.

For a 1000 gram increase in body mass (i.e., 1 kg or $\sim$ 2.2 lbs), we expect bill length to increase by 0.66 mm.

For a 1 mm increase in flipper length, we expect bill length to increase by 0.22 mm.

Confidence Intervals for β_i

Recall confidence intervals – they allow us to determine how “good’ ’ our estimation is.

In general CIs will take the form

$$\text{point estimate} \ \pm \ \text{margin of error},$$

where the margin of error is a critical value (e.g., $t_{1-\alpha/2}$) multiplied by the standard error of the point estimate.

Recall that the standard error accounts for the sample size.

In R, we will run the model results through the `confint()` function.

```
confint(m)
```

Confidence Intervals for β_i

Example:

```
confint(m1)
```

```
##                            2.5 %      97.5 %
## (Intercept)        -1.244658e+01 5.573192182
## body_mass_g        -4.534709e-04 0.001777908
## flipper_length_mm   1.582365e-01 0.285494420
```

We have the following CIs:

95% CI for β_{mass} is (-0.0005, 0.0018)
95% CI for β_{flipper} is (0.1582, 0.2855)

Confidence Intervals for β_i

We can change the confidence level by specifying the `level`.

Example:

```
confint(m1, level=0.99)
```

```
##                           0.5 %      99.5 %
## (Intercept)       -1.530220e+01 8.428814954
## body_mass_g       -8.070812e-04 0.002131518
## flipper_length_mm  1.380697e-01 0.305661191
```

```
confint(m1, level=0.8914)
```

```
##                          5.43 %     94.57 %
## (Intercept)       -1.080570e+01 3.932309891
## body_mass_g       -2.502813e-04 0.001574718
## flipper_length_mm  1.698246e-01 0.273906301
```

Significant Regression Line

Hypotheses

$$H_0: \; \beta_1 = \ldots = \beta_k = 0$$
$$H_1: \text{at least one } \beta_i \neq 0$$

Test Statistic and p-Value

F_0 and p from `summary()` (last line)

Rejection Region

Reject H_0 if $p < \alpha$

Significant Regression Line

```
m1 <- lm(bill_length_mm ~ body_mass_g + flipper_length_mm,
         data = data)
summary(m1)
```

```
##
## Call:
## lm(formula = bill_length_mm ~ body_mass_g + flipper_length_mm,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8064 -2.5898 -0.7053  1.9911 18.8288
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -3.4366939  4.5805532  -0.750    0.454
## body_mass_g        0.0006622  0.0005672   1.168    0.244
## flipper_length_mm  0.2218655  0.0323484   6.859 3.31e-11 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.124 on 339 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.4329, Adjusted R-squared:  0.4295
## F-statistic: 129.4 on 2 and 339 DF,  p-value: < 2.2e-16
```

Significant Regression Line

Hypotheses

$$H_0: \ \beta_{\text{mass}} = \beta_{\text{flipper}} = 0$$
$$H_1: \text{at least one } \beta_i \neq 0$$

Test Statistic and p-Value

$$F_0 = 129.4 \ (p < 0.001)$$

Rejection Region

Reject H_0 if $p < \alpha$; $\alpha = 0.05$

Conclusion/Interpretation

Reject H_0. There is sufficient evidence to suggest that at least one slope is non-zero.

Significant Predictors of y

Hypotheses

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0$$

Test Statistic and p-Value

t_0 and p from `summary()` (last two columns)

Rejection Region

Reject H_0 if $p < \alpha$

Significant Predictors of y

```
m1 <- lm(bill_length_mm ~ body_mass_g + flipper_length_mm,
         data = data)
summary(m1)
```

```
##
## Call:
## lm(formula = bill_length_mm ~ body_mass_g + flipper_length_mm,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8064 -2.5898 -0.7053  1.9911 18.8288
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -3.4366939  4.5805532  -0.750    0.454
## body_mass_g        0.0006622  0.0005672   1.168    0.244
## flipper_length_mm  0.2218655  0.0323484   6.859 3.31e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.124 on 339 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.4329, Adjusted R-squared:  0.4295
## F-statistic: 129.4 on 2 and 339 DF,  p-value: < 2.2e-16
```

Significant Predictors of y

Hypotheses

$$H_0 : \beta_{\text{mass}} = 0$$
$$H_1 : \beta_{\text{mass}} \neq 0$$

Test Statistic and p-Value

$$t_0 = 1.168 \ (p = 0.244)$$

Rejection Region

Reject H_0 if $p < \alpha$; $\alpha = 0.05$

Conclusion / Interpretation

Fail to reject H_0. There is not sufficient evidence to suggest that body mass significantly predicts bill length.

Significant Predictors of y

Hypotheses

$$H_0 : \ \beta_{\text{flipper}} = 0$$
$$H_1 : \ \beta_{\text{flipper}} \neq 0$$

Test Statistic and p-Value

$$t_0 = 6.859 \ (p < 0.001)$$

Rejection Region

Reject H_0 if $p < \alpha$; $\alpha = 0.05$

Conclusion / Interpretation

Reject H_0. There is sufficient evidence to suggest that flipper length significantly predicts bill length.

Visualizing the Data

We can construct basic scatterplots to try to visualize the relationships*.

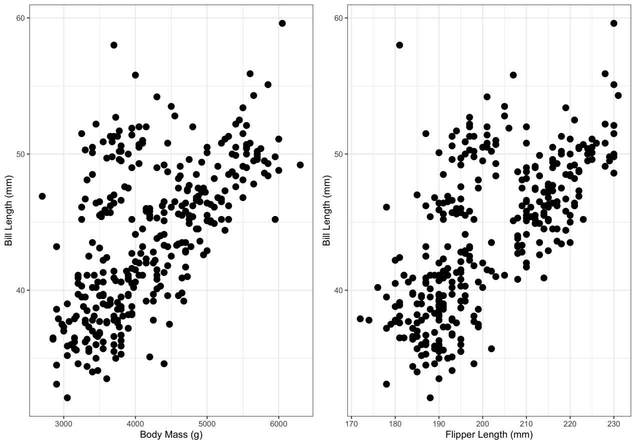
Example:

```
p1 <- data %>% ggplot(aes(x = body_mass_g, y = bill_length_mm)) +
               geom_point(size=3) +
               ylab("Bill Length (mm)") +
               xlab("Body Mass (g)") +
               theme_bw()
```

```
p2 <- data %>% ggplot(aes(x = flipper_length_mm, y = bill_length_mm)) +
               geom_point(size=3) +
               ylab("Bill Length (mm)") +
               xlab("Flipper Length (mm)") +
               theme_bw()
```

Visualizing the Data

Example:



Visualizing the Model

We can construct predicted values to overlay the resulting regression line.

> To do this, we must pick one predictor to vary. All other predictors must be held constant in order to overlay a regression line.
>
> First, we will plug in the average flipper length and let body mass vary (`p_mass`).
>
> Then, we will plug in the average body mass and let flipper length vary (`p_flip`).

Example:

```
c1 <- coefficients(m1)

data <- data %>%
  mutate(p_mass = c1[1] + c1[2]*body_mass_g + c1[3]*mean(data$flipper_length_mm, na.rm = TR
         p_flip = c1[1] + c1[2]*mean(data$body_mass_g, na.rm = TRUE) + c1[3]*flipper_length
```

Visualizing the Model

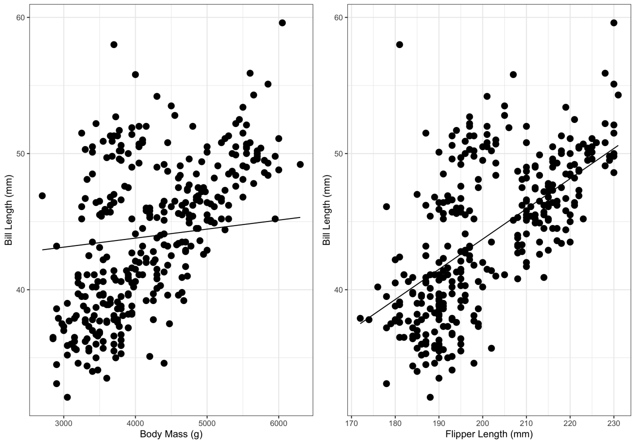
Example:

```
p3 <- data %>% ggplot(aes(x = body_mass_g, y = bill_length_mm)) +
               geom_point(size=3) +
               geom_line(aes(y = p_mass)) +
               ylab("Bill Length (mm)") +
               xlab("Body Mass (g)") +
               theme_bw()
```

```
p4 <- data %>% ggplot(aes(x = flipper_length_mm, y = bill_length_mm)) +
               geom_point(size=3) +
               geom_line(aes(y = p_flip)) +
               ylab("Bill Length (mm)") +
               xlab("Flipper Length (mm)") +
               theme_bw()
```

Visualizing the Data

Example:


Bill Length (mm)
Body Mass (g)
Bill Length (mm)
Flipper Length (mm)