M08Guided

Alanna Hazlett

2024-03-25

peguins dataset from palmer penguins, we focus on exploring the relationship between body mass (y) and bill. depth (x1) of three species of penguins.

Problem 1

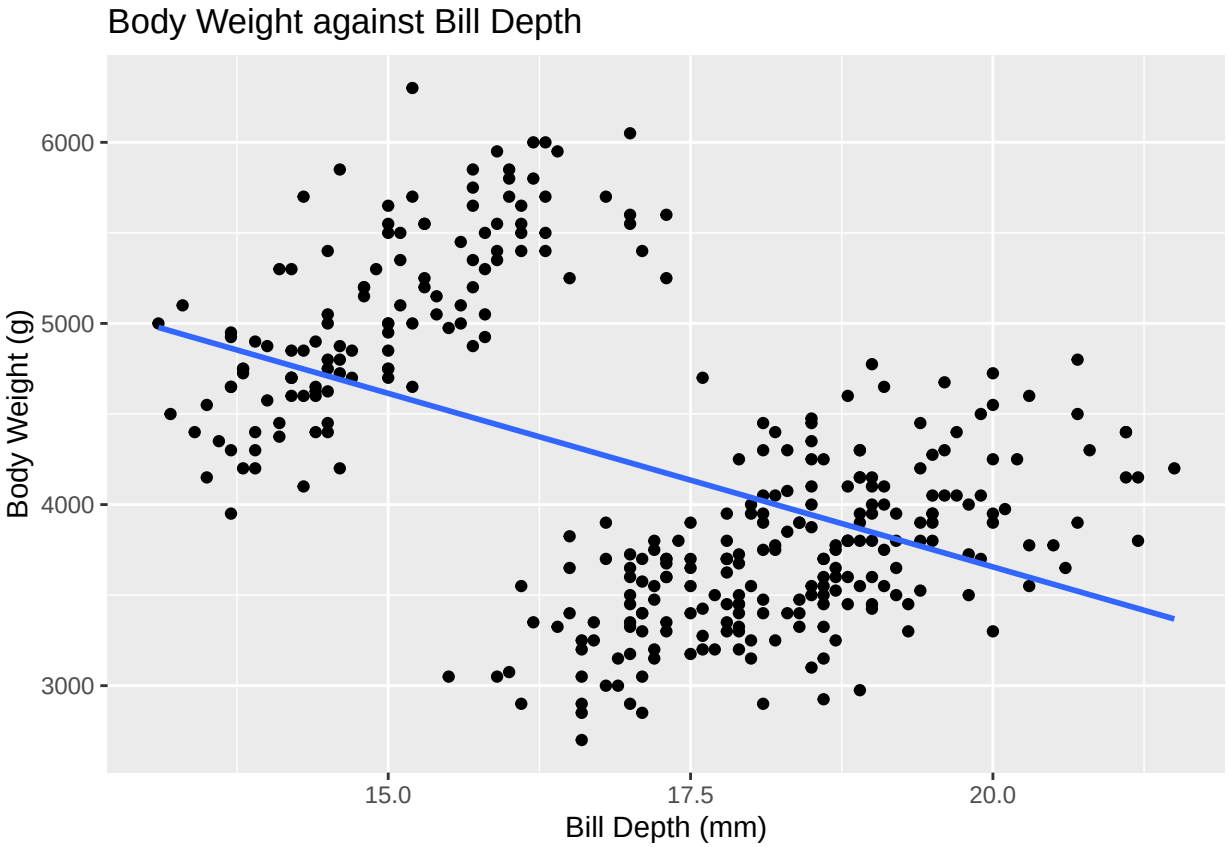
Create a scatterplot of the body mass against the bill depth of the penguins. How would you describe the relationship between these two variables?

```
Data<-penguins
ggplot(Data, aes(x=bill_depth_mm,y=body_mass_g))+
  geom_point()+
  geom_smooth(method=lm, se=FALSE)+
  labs(x="Bill Depth (mm)", y="Body Weight (g)", title = "Body Weight against Bill Depth")
```

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

```
## Warning: Removed 2 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```



Despite our estimated regression line showing a negative relationship, this looks like two distinct groupings of data that have positive linear relationships.

Problem 2

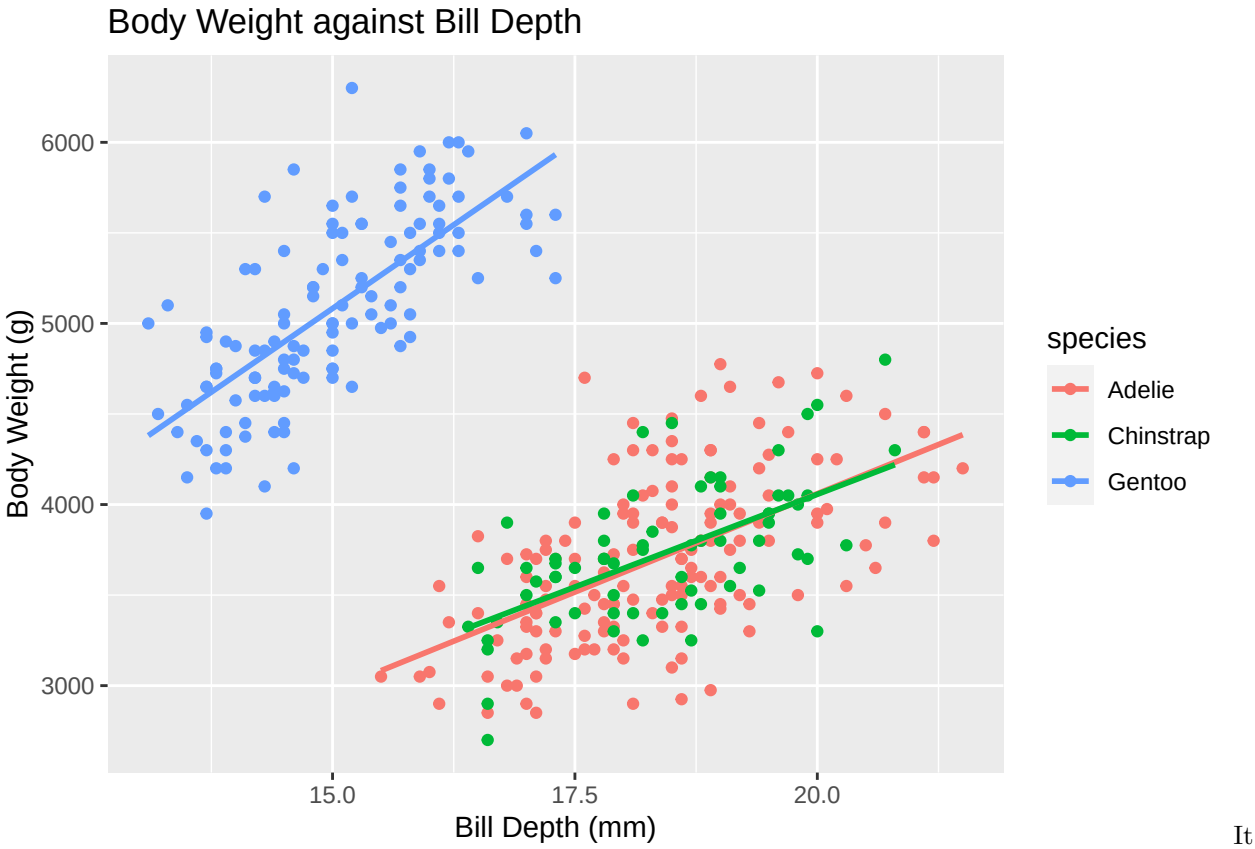
Create the same scatterplot but now with different colored plots for each species. Also be sure to overlay separate regression lines for each species. How would you now describe the relationship between the variables?

```
ggplot(Data, aes(x=bill_depth_mm,y=body_mass_g, color=species))+
  geom_point()+
  geom_smooth(method=lm, se=FALSE)+
  labs(x="Bill Depth (mm)", y="Body Weight (g)", title = "Body Weight against Bill Depth")
```

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

```
## Warning: Removed 2 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```



It does appear we have multiple positively linear relationships, however we have 3 not 2.

Problem 3

Create a regression with interaction between bill depth and species, where I1 and I2 are indicator variables, where I1 = 1 for Chinstrap penguins and 0 otherwise, and I2 = 1 for Gentoo penguins and 0 otherwise. Write the estimated regression equation.

```
#Check that the predictor variable is in the correct order
levels(Data$species)
```

```
## [1] "Adelie"    "Chinstrap" "Gentoo"
```

```
result.inter<-lm(body_mass_g ~ bill_depth_mm*species,data=Data)
summary(result.inter)
```

```
##
## Call:
## lm(formula = body_mass_g ~ bill_depth_mm * species, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -845.89 -254.74  -28.46  228.01 1161.41
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                      -283.28     437.94  -0.647   0.5182
## bill_depth_mm                     217.15      23.82   9.117   <2e-16 ***
## speciesChinstrap                  247.06     829.77   0.298   0.7661
## speciesGentoo                    -175.71     658.43  -0.267   0.7897
## bill_depth_mm:speciesChinstrap    -12.53      45.01  -0.278   0.7809
## bill_depth_mm:speciesGentoo       152.29      40.49   3.761   0.0002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 354.9 on 336 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.807,  Adjusted R-squared:  0.8041
## F-statistic:   281 on 5 and 336 DF,  p-value: < 2.2e-16
```

The model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 I_1 + \beta_3 I_2 + \beta_4 x_1 I_1 + \beta_5 x_1 I_2 + \in$$

$$y = -283.28 + 217.15 x_1 + 247.06 I_1 + -175.71 I_2 + -12.53 x_1 I_1 + 152.29 x_1 I_2$$

Problem 4

Carry out the relevant hypothesis test to see if the interaction terms can be dropped. What is the conclusion?

$$H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

$$H_a : at\ least\ one\ coefficient \neq 0$$

```
result.reduced<-lm(body_mass_g~bill_depth_mm+species,data=Data)
anova(result.reduced,result.inter)
```

```
## Analysis of Variance Table
##
## Model 1: body_mass_g ~ bill_depth_mm + species
## Model 2: body_mass_g ~ bill_depth_mm * species
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1    338 44399670
## 2    336 42325191  2   2074479 8.2342 0.0003227 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The general linear F test is significant (p-value < 0.05), so we need to use the full model with the interactive effects.

Problem 5

Based on your answer in part 4, write out the estimated regression equations relating body mass and bill depth, for each species of the penguins.
The regression equations are:
For Adelie:

$$E\{Y\} = -283.28 + 217.15 x_1 + 247.06\ (0) + -175.71\ (0) + -12.53 x_1\ (0) + 152.29 x_2\ (0)$$

$$E\{Y\} = -283.28 + 217.15 x_1$$

For Chinstrap:

$$E\{Y\} = -283.28 + 217.15x_1 + 247.06\ (1) + -175.71\ (0) + -12.53x_1\ (1) + 152.29x_2\ (0)$$
$$E\{Y\} = -36.22 + 204.62x_1$$

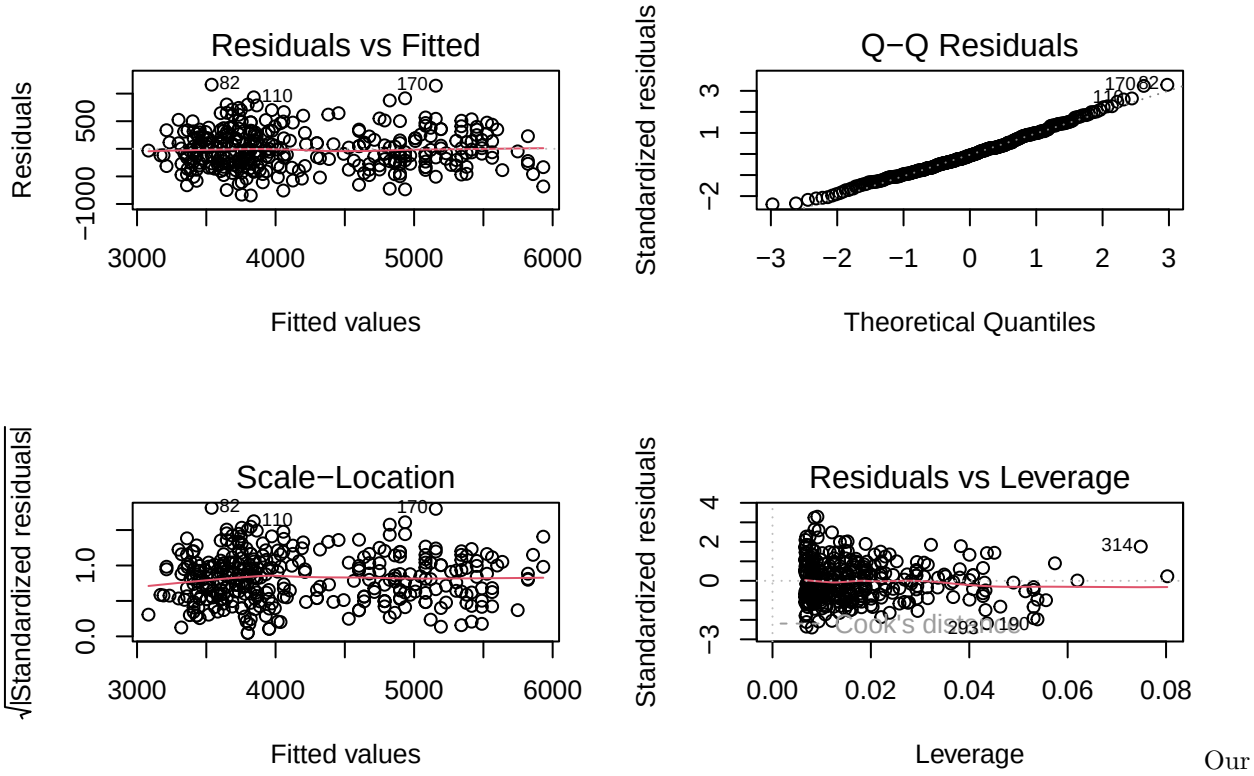
For Gentoo:

$$E\{Y\} = -283.28 + 217.15x_1 + 247.06\ (0) + -175.71\ (1) + -12.53x_1\ (0) + 152.29x_2\ (1)$$
$$E\{Y\} = -458.99 + 217.15x_1 + 152.29x_2$$

Problem 6

Assess if the regression assumptions are met, for the model you will recommend to use (based on part 4).

```
par(mfrow=c(2,2))
plot(result.inter)
```



Our regression assumptions are met.

Problem 7

Briefly explain if we can conduct pairwise comparisons for the difference in mean body mass among all pairs of species for given values bill depth:
(a) Adelie and Chinstrap
(b) Adelie and Gentoo
(c) Chrinstrap and Gentoo
If we are able to, conduct Tukey's multiple comparisons and contextually interpret the results of these hypothesis tests.

Yes, we can.

```
pairwise<-multcomp::glht(result.inter,linfct=mcp(species="Tukey"))
```

```
## Warning in mcp2matrix(model, linfct = linfct): covariate interactions found --
## default contrast might be inappropriate
```

```
summary(pairwise)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = body_mass_g ~ bill_depth_mm * species, data = Data)
##
## Linear Hypotheses:
##                          Estimate Std. Error t value Pr(>|t|)
## Chinstrap - Adelie == 0    247.1      829.8   0.298    0.952
## Gentoo - Adelie == 0      -175.7      658.4  -0.267    0.961
## Gentoo - Chinstrap == 0   -422.8      859.3  -0.492    0.874
## (Adjusted p values reported -- single-step method)
```

Test Statistic:

$$t = \frac{estimate}{se(estimate)}$$

Critical Value:

```
qt(1-0.05/(2*3), 344-6)
```

```
## [1] 2.405955
```

All 3 t statistics $<$ critical value, so there is a not a significant difference in body mass for the species, when controlling bill depth.