Homework 8: Multiple Regression

STAT 242: Intermediate Statistics

SOLUTIONS

The code below just loads some packages and makes it so that enough digits are printed that you won't get confused by rounding errors.

```
library(dplyr) # functions like summarize
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
library(ggplot2) # for making plots
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.0.5
```

```
library(egg)
options("pillar.sigfig" = 10) # print 10 significant digits in summarize output
```

Problem 1: Crabs (Adapted from Sleuth 3 exercise 9.17)

The description below comes from our book:

As part of a study of the effects of predatory intertidal crab species on snail populations, researchers measured the mean closing forces (in newtons) and the propodus heights (in mm) of the claws on several crabs of three species. (Data from S. B. Yamada and E. G. Boulding, "Claw Morphology, Prey Size Selection and Foraging Efficiency in Generalist and Specialist Shell-Breaking Crabs," *Journal of Experimental Marine Biology and Ecology, 220 (1998): 191-211.) Here we will examine the relationship between closing force (our response variable) and species and propodus height (explanatory variables). The following code reads the data in.

```
library(Sleuth3)
```

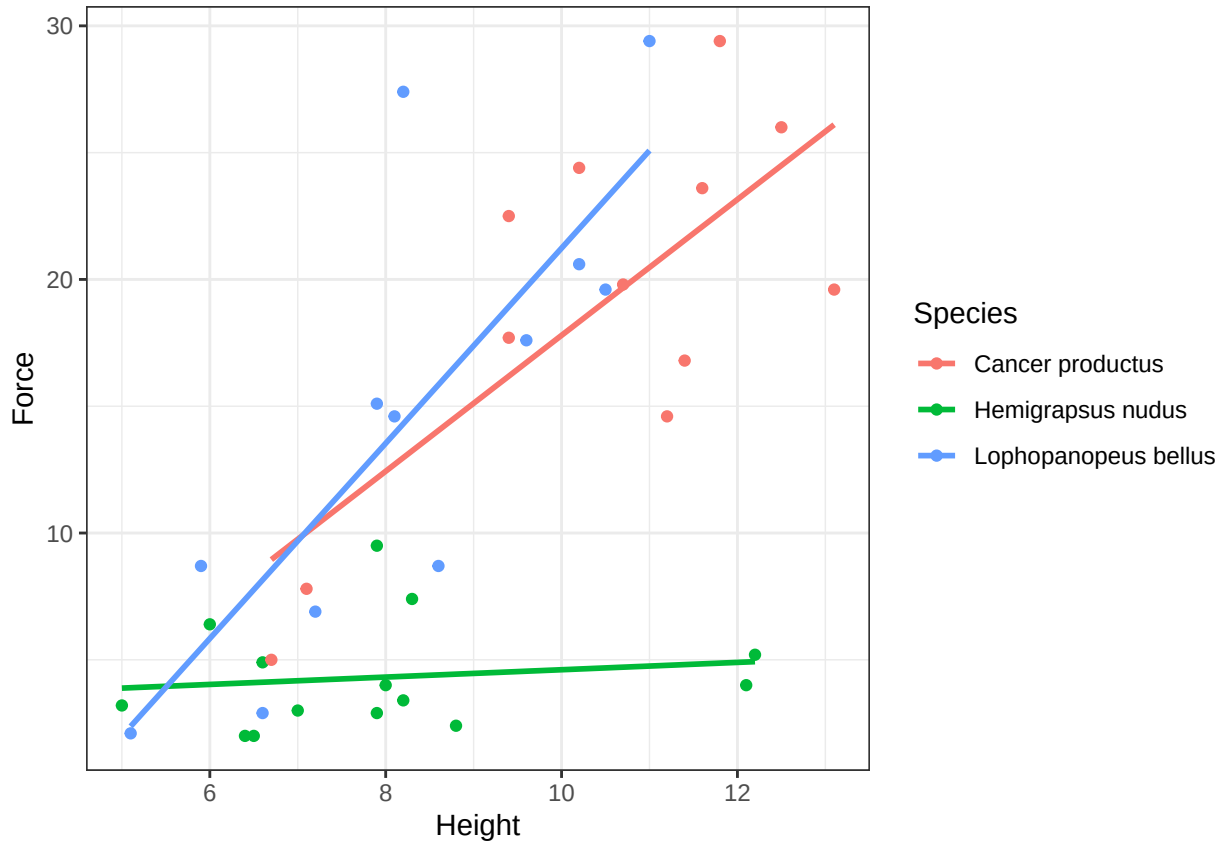
```
## Warning: package 'Sleuth3' was built under R version 4.0.3
```

```
crabs <- ex0722
```

(a) Create an appropriate plot of the data involving all three variables. Does it appear that an additive model or a model with interactions between species and height would be more appropriate?

```
ggplot(data=crabs, aes(x=Height, y=Force, color=Species)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw()
```

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

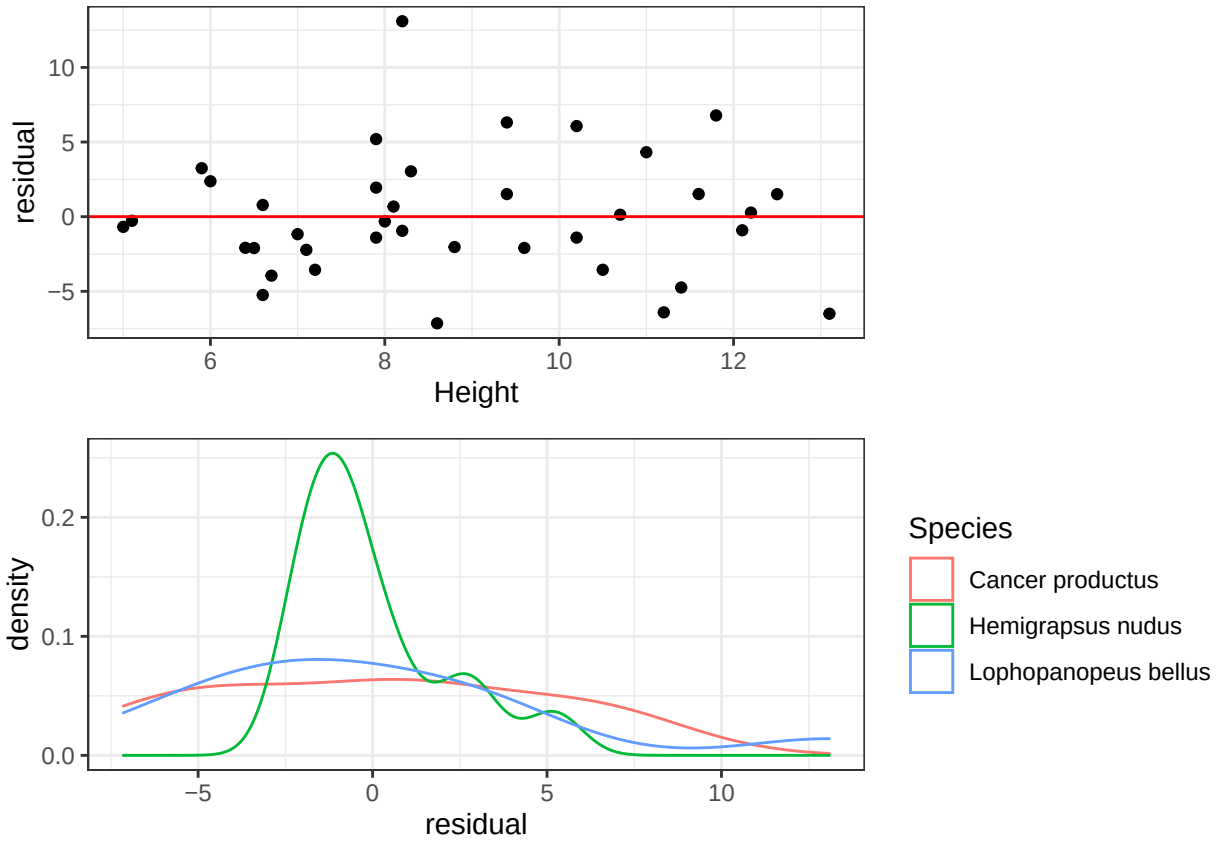


Based on the plot above, it appears that a model with interactions will be more appropriate. If a model an additive model was sufficient, I would expect to see three (roughly) parallel lines. These three lines look like they have different slopes (and different intercepts), so I will need the model with interactions to accommodate the different linear relationships for the three species.

(b) Fit a multiple regression model to the data, allowing for different slopes for the different species. In this model, use the original Height and Force variables as explanatory and response variables, respectively. Create residual diagnostic plots of your model fit and calculate the standard deviation of the residuals within each group. Discuss any conditions for the regression model that are not satisfied.

```
## Fit multiple regression model with interactions (different slope for each species)
lm_diff_slopes <- lm(Force ~ Height*Species, data=crabs)

## Create residual diagnostic plots for your model fit
```

```
crabs <- crabs %>% mutate(
  residual = residuals(lm_diff_slopes)
)

p1 <- ggplot(data=crabs, aes(x=Height, y=residual)) +
        geom_point() +
        geom_hline(yintercept=0, color="red") +
        theme_bw()

p2 <- ggplot(data=crabs, aes(x=residual, color=Species)) +
        geom_density() +
        theme_bw()

ggarrange(p1, p2)
```




```
## Calculate standard deviation of residuals within each group (with in each Species).
crabs %>%
  group_by(Species) %>%
  summarize(
    std_dev=sd(residual)
  )
```

```
## # A tibble: 3 x 2
##   Species                 std_dev
```

```
##   <fct>                     <dbl>
## 1 Cancer productus     4.824340538
## 2 Hemigrapsus nudus    2.167889758
## 3 Lophopanopeus bellus 5.353284342
```

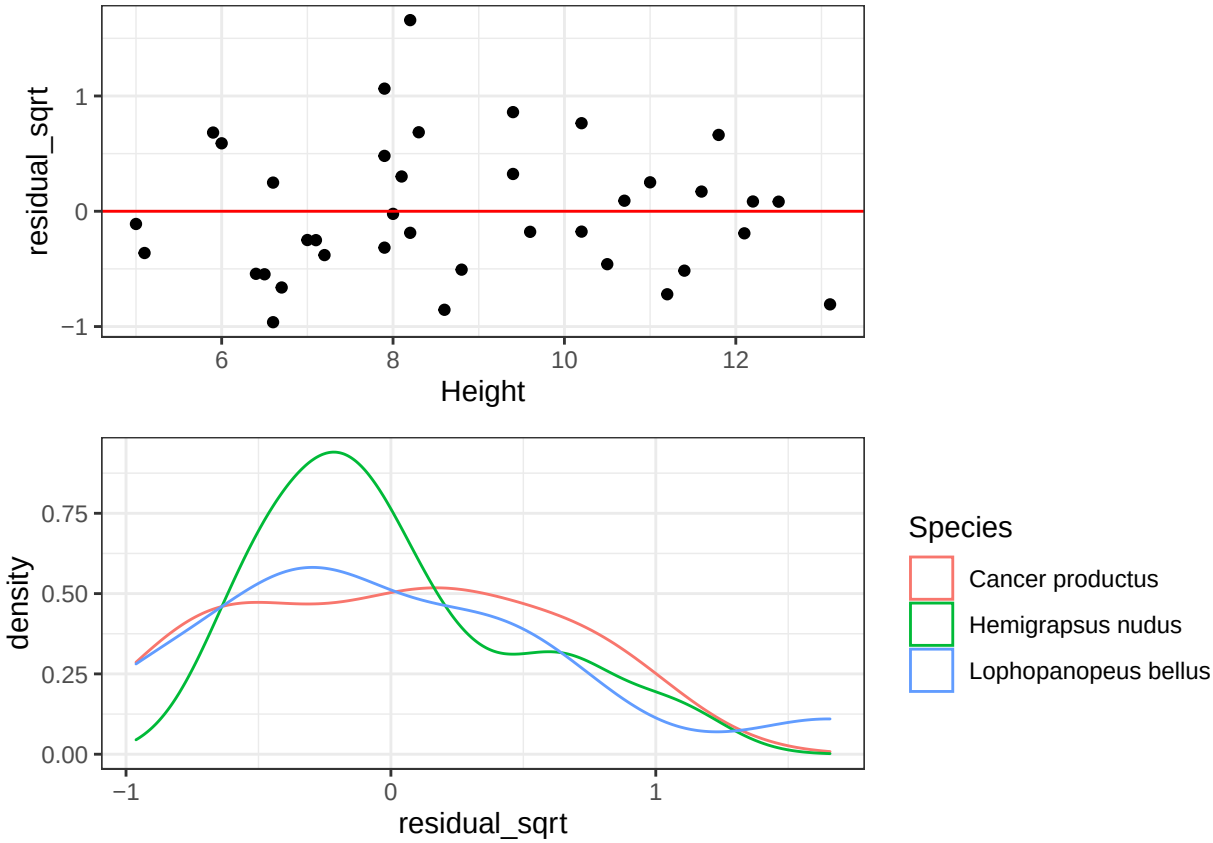
We would be checking the same conditions as we checked for the simple linear regression model (with a little more complexity).

- Linearity: Looking at the residual plot (residual vs. Height), there are no obvious patterns in the residuals. I feel good about linearity here.
- Independence: For the problem description as it stands, we do not know enough about whether this is satisfied. If the crabs are randomly sampled from the population of crabs (involving these three species), then they will be independent. If all of the crabs come from a particular area, they may not be independent.
- Normal residuals: This seems okay as a function of Height, since the residual plot reveals no patterns. It is not bad as a function of species, either, although there is some right skew present in the distributions of the residuals.
- Equal variance: This is a problem - the different species clearly have different standard deviations, and the standard deviation for Lophopanopeus bellus is more than twice that of Hemigrapsus nudus. We will need to consider a transformation. Note that as a function of Height, this assumption may be okay, although there is an apparent vertical outlier around Height=8.
- Outliers: there is at least one potential outlier around Height=8. There is also right skew present for the residual densities for the different groups, so there may be outliers present here, too. We should consider a transformation.

(c) Find a set of transformations of the data so that the conditions of the multiple regression model are better satisfied (Note: I think you can do well enough with transformations of the response variable only). Verify that you have succeed by discussing residual diagnostic plots and standard deviations of the residuals across the different species. Recreate your plot of the data from part (a), but with your transformed variables this time.

```
## Right-skewed, so step down the ladder:

## ---- Sqrt transformation:
crabs <- crabs %>% mutate(
  Force_sqrt = sqrt(Force)
)

## Fit multiple regression model with interactions (different slope for each species)
lm_diff_slopes_sqrt <- lm(Force_sqrt ~ Height*Species, data=crabs)

## Create residual diagnostic plots for your model fit
crabs <- crabs %>% mutate(
  residual_sqrt = residuals(lm_diff_slopes_sqrt)
)

p1 <- ggplot(data=crabs, aes(x=Height, y=residual_sqrt)) +
        geom_point() +
        geom_hline(yintercept=0, color="red") +
        theme_bw()

p2 <- ggplot(data=crabs, aes(x=residual_sqrt, color=Species)) +
```

```
        geom_density() +
        theme_bw()

ggarrange(p1, p2)
```




```
## Calculate standard deviation of residuals within each group (with in each Species).
crabs %>%
  group_by(Species) %>%
  summarize(
    std_dev=sd(residual_sqrt)
  )
```

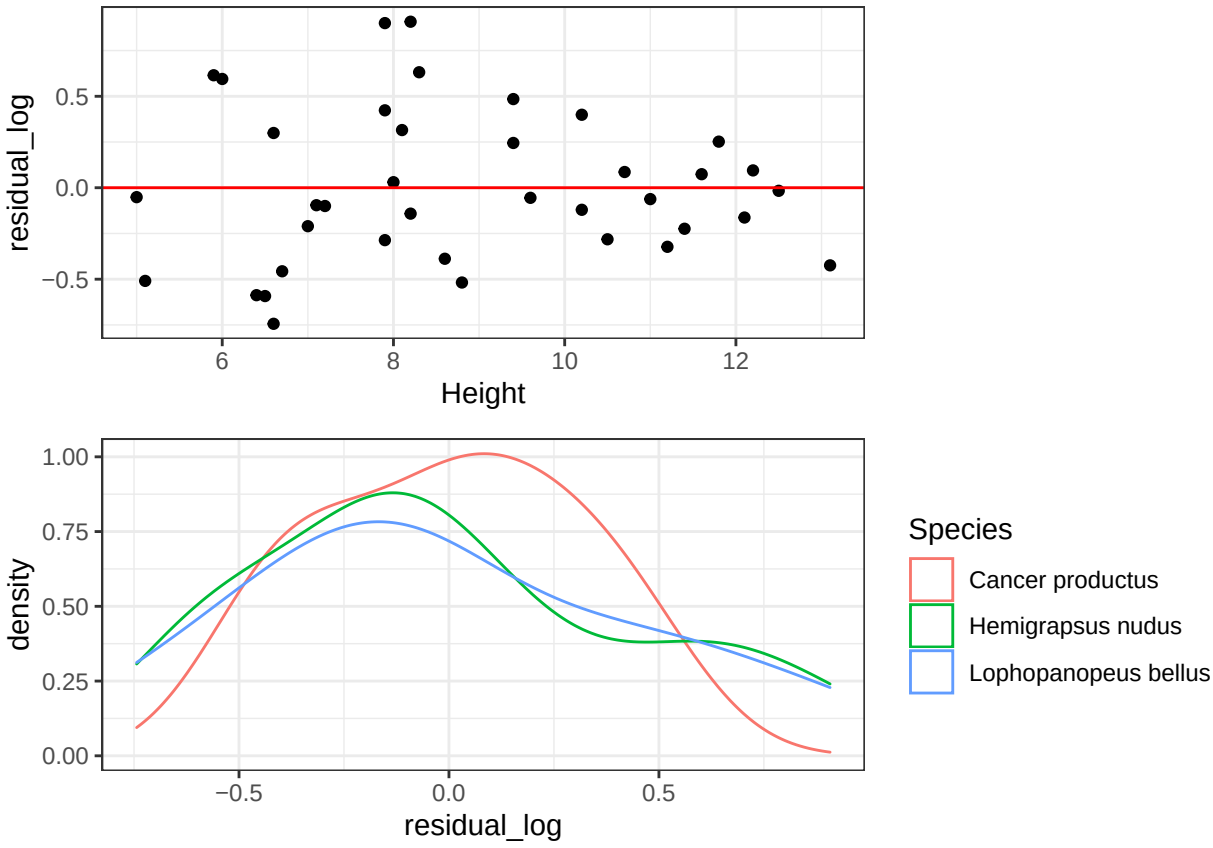
```
## # A tibble: 3 x 2
##   Species                   std_dev
##   <fct>                       <dbl>
## 1 Cancer productus     0.5913709475
## 2 Hemigrapsus nudus    0.4893722944
## 3 Lophopanopeus bellus 0.7265158084
```

```
## ---- log transformation:
crabs <- crabs %>% mutate(
  Force_log = log(Force)
)
```

```r
## Fit multiple regression model with interactions (different slope for each species)
lm_diff_slopes_log <- lm(Force_log ~ Height*Species, data=crabs)
summary(lm_diff_slopes_log)
```

```
##
## Call:
## lm(formula = Force_log ~ Height * Species, data = crabs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.74379 -0.28537 -0.05864  0.28745  0.90762
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         0.67009    0.72122   0.929  0.35979
## Height                              0.20838    0.06806   3.062  0.00443 **
## SpeciesHemigrapsus nudus            0.30888    0.87196   0.354  0.72548
## SpeciesLophopanopeus bellus        -1.31355    0.94795  -1.386  0.17543
## Height:SpeciesHemigrapsus nudus    -0.16126    0.09072  -1.778  0.08499 .
## Height:SpeciesLophopanopeus bellus  0.16313    0.09978   1.635  0.11188
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4486 on 32 degrees of freedom
## Multiple R-squared:  0.7794, Adjusted R-squared:  0.7449
## F-statistic: 22.61 on 5 and 32 DF,  p-value: 1.195e-09
```

```r
## Create residual diagnostic plots for your model fit
crabs <- crabs %>% mutate(
  residual_log = residuals(lm_diff_slopes_log)
)

p1 <- ggplot(data=crabs, aes(x=Height, y=residual_log)) +
        geom_point() +
        geom_hline(yintercept=0, color="red") +
        theme_bw()

p2 <- ggplot(data=crabs, aes(x=residual_log, color=Species)) +
        geom_density() +
        theme_bw()

ggarrange(p1, p2)
```




```
## Calculate standard deviation of residuals within each group (with in each Species).
crabs %>%
  group_by(Species) %>%
  summarize(
    std_dev=sd(residual_log)
  )
```

```
## # A tibble: 3 x 2
##   Species                  std_dev
##   <fct>                      <dbl>
## 1 Cancer productus     0.3137798207
## 2 Hemigrapsus nudus    0.4642135380
## 3 Lophopanopeus bellus 0.4819051524
```

Although the square root transformation is probably fine, the log transformation is better. The standard deviations for the three species are essentially the same, and the densities for the residuals are very similar. There are no apparent outliers in the residual plot (for Height). Also, the interpretations of our results on the original scale (when we transform back from the log) will make more sense.

(d) Write down the model you fit in part (c). This should not involve any numbers.

- Y: log Force
- X_1: indicator for Hemigrapsus nudus species; 1 if H. nudus, 0 otherwise
- X_2: indicator for Lophopanopeus bellus species; 1 if L. bellus, 0 otherwise

- X_3: Height
- X_4: $x_1 \times$ Height
- X_5: $x_2 \times$ Height

For the i^{th} individual ($i = 1, ..., 38$),

$$\mu(Y_i|\mathbf{X}_i) = \beta_0 + \beta_1(SpeciesHemigrapsus\_nudus) + \beta_2(SpeciesLophopanopeus\_bellus) + \beta_3(Height) + \beta_4(SpeciesHemigrapsus\_nudus \times Height) + \beta_5(SpeciesLophopanopeus\_bellus \times Height) + \epsilon_i$$

(e) Write down the equation for the estimated population mean (transformed) force as a function of species indicator variables and propodus height.

$$\hat{\mu}(Y_i|\mathbf{X}_i) = 0.670 + 0.309(SpeciesHemigrapsus\_nudus) - 1.314(SpeciesLophopanopeus\_bellus) + 0.208(Height) - 0.161(SpeciesHemigrapsus\_nudus \times Height) + 0.163(SpeciesLophopanopeus\_bellus \times Height)$$

(f) Write down the equation for the estimated mean (transformed) forces as a function of propodus height, in the population of Lophopanopeus bellus crabs. Group together like terms so you have a single intercept and slope.

$$\begin{aligned}\hat{\mu} &= 0.670 + 0.309(0) - 1.314(1) + 0.208(Height) - 0.161(0 \times Height) + 0.163(1 \times Height) \\ &= -0.643 + 0.840(Height)\end{aligned}$$

(g) What is the estimated change in (transformed) claw closing force that is associated with a 1 mm increase in propodus height, in the population of Cancer productus crabs? Just writing down a number is good enough.

0.208 log Newtons/mm

(h) What is the estimated change in (transformed) claw closing force that is associated with a 1 mm increase in propodus height, in the population of Hemigrapsus Nudus crabs? Just writing down a number is good enough.

0.047 log Newtons/mm

(i) Find and interpret a 95% confidence interval for the difference between the change in population mean (transformed) claw closing force that is associated with a 1 mm increase in propodus height in the populations of Hemigrapsus Nudus crabs and Cancer productus crabs. (That sentence was a lot to take in. I'm looking for a confidence interval for the difference between the population quantities from parts h and g.) Your answer should include a couple of sentences describing interpretation in context.

```
confint(lm_diff_slopes_log)
```

```
##                                         2.5 %     97.5 %
## (Intercept)                        -0.79899192 2.13917015
## Height                              0.06974908 0.34700719
## SpeciesHemigrapsus nudus           -1.46723696 2.08500663
## SpeciesLophopanopeus bellus        -3.24447037 0.61736621
## Height:SpeciesHemigrapsus nudus    -0.34604414 0.02353366
## Height:SpeciesLophopanopeus bellus -0.04011936 0.36638320
```

We are 95% confident that the difference between the change in mean log claw closing force associated with a 1 mm increase in propodus height in the populations of Hemigrapsus Nudus and Cancer productus is between -0.346 and 0.024 log Newtons. For 95% of samples from a similar population, the true difference in the change in mean log claw closing force associated with the two populations would be in the corresponding interval.

You are not asked to do this, but you could transform back to the original scale. Then, assuming you used a log transformation, the interpretation is as follows. We are 95% confident that the mean claw closing force associated with a 1 mm increase in propodus height in the population of Hemigraphsus Nudus crabs is between 0.707 and 1.024 times the mean claw closing force associated with a 1 mm increase in propodus height in the population of Cancer productus crabs.

(j) Conduct a test of the claim that the slopes of lines describing the relationship between propodus height and (transformed) closing force is the same in the populations of crabs of all three species. State your null and alternative hypotheses in terms of model parameters, the p-value for the test, and your conclusion in context.

This is an F test.

$H_0 : \beta_4 = 0$ and $\beta_5 = 0$ H_A : at least one of these parameters is not equal to 0

```r
## Need a reduced model for this F test - should have parallel lines
lm_parallel_log <- lm(Force_log ~ Species + Height, data=crabs)

## Use the anova function to conduct the F test between
## lm_parallel_log (reduced model) and lm_diff_slopes_log (full model)
anova(lm_parallel_log, lm_diff_slopes_log)
```

```
## Analysis of Variance Table
##
## Model 1: Force_log ~ Species + Height
## Model 2: Force_log ~ Height * Species
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     34 8.8395
## 2     32 6.4390  2    2.4005 5.9648 0.006285 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is strong evidence (p-value = 0.006) that the slopes of the lines describing the relationship between propodus height and log closing force are not the same in the populations of crabs of all three species; at least one species requires a different slope.

(k) Although you had R do the calculation of the test statistic and the p-value for the test in part (j), you should know how that statistic was calculated. Describe how to calculate the test statistic for your test from part (j) in a paragraph or so. Include a discussion of how the degrees of freedom for the statistic are found. Does a large value of the statistic offer strong or weak evidence against the null hypothesis? Why?

$$F = \frac{RSS_{Extra}/df_{Extra}}{RSS_{Full}/df_{Full}}$$

$df_{Extra} = df_{Reduced} - df_{Full} = (38-4) - (38-6) = 2$; $df_{Full} = 38 - 6 = 32$

A large value offers strong evidence against the null hypothesis because this indicates that the RSS for the reduced model is much larger than that for the full model, so it does not sufficiently explain the variability in the response.

(l) How were the β coefficients in your models above estimated? You can answer in just a sentence or two.

The coefficients are estimated by minimizing the residual sum of squares.