HW8

Sections 9.1, 9.2, 9.3.2, 9.3.3, 9.3.4, 10.1, 10.2.2, 10.2.3, 10.3

Your Name Here

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

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 hreee 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).

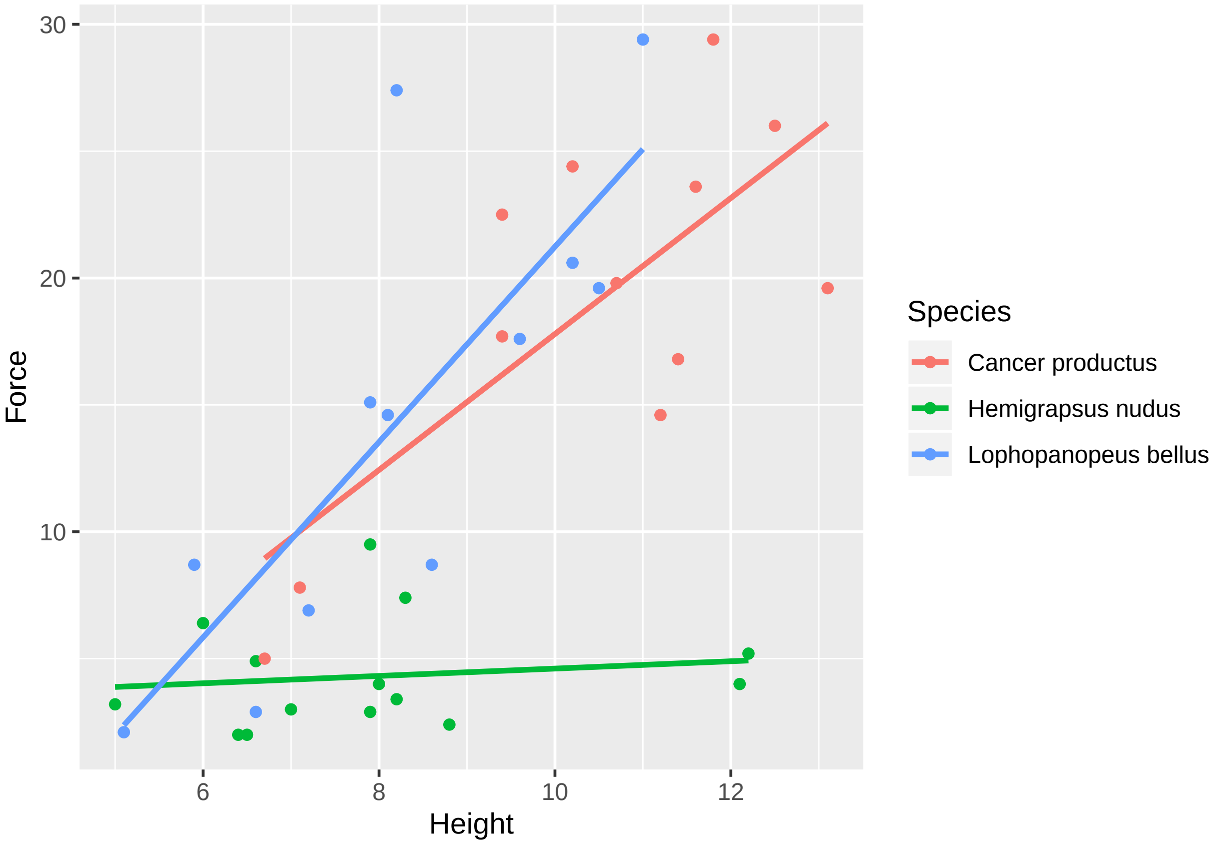
The following code reads the data in.

```
crabs <- read_csv("http://www.evanlray.com/data/sleuth3/ex0722_crabs.csv")
```

```
## Parsed with column specification:
## cols(
##   Force = col_double(),
##   Height = col_double(),
##   Species = col_character()
## )
```

(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?

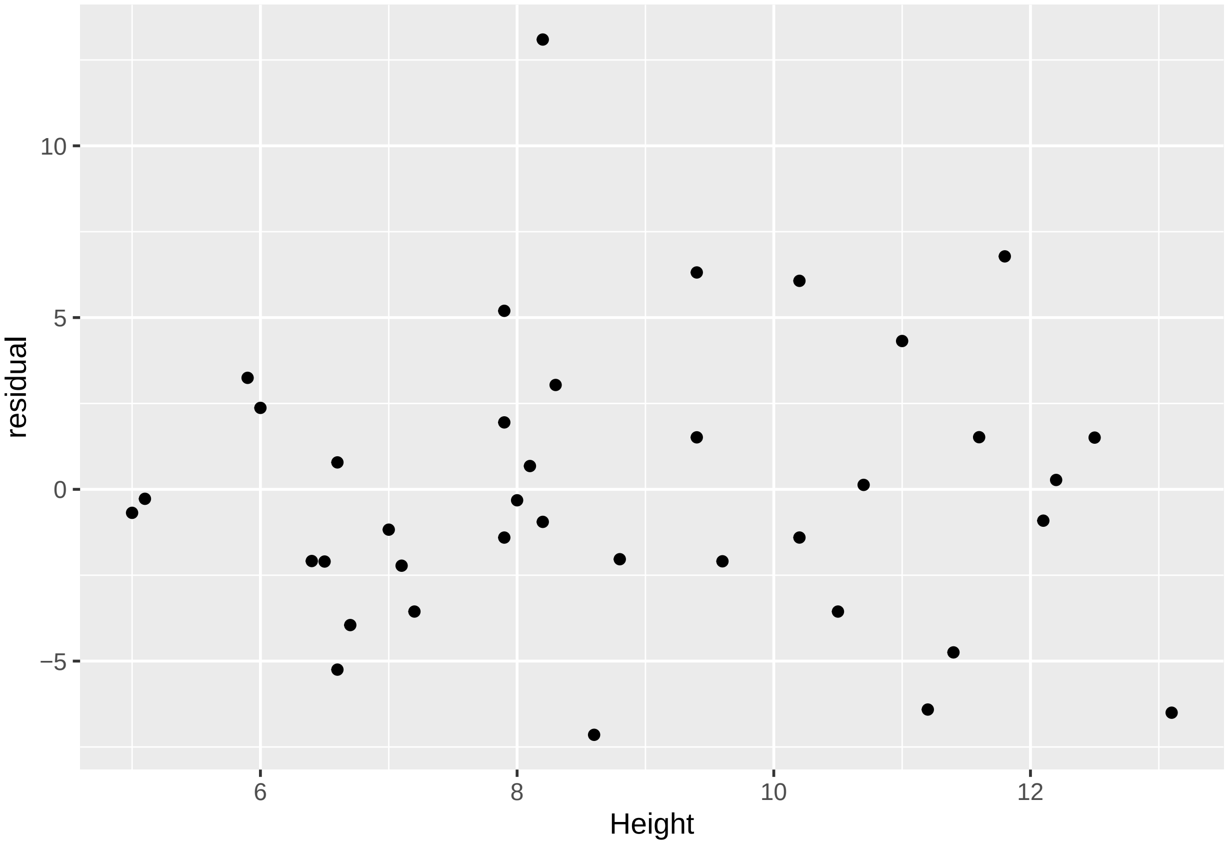
```
ggplot(data = crabs, mapping = aes(x = Height, y = Force, color = Species)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



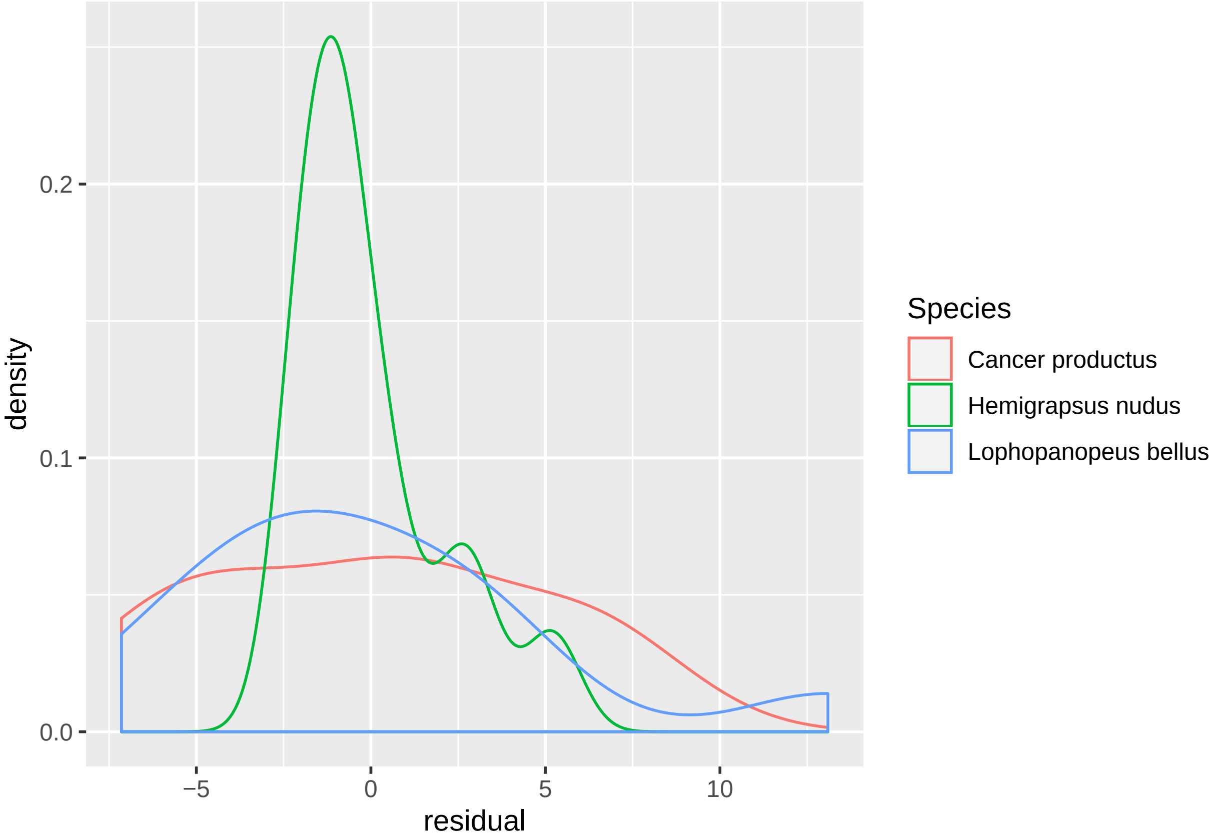
(b) Fit a multiple regression model to the data, allowing for different slopes for the different species. In this model, use the original Height and Force variables as explanatory variables. Create residual diagnostic plots of your model fit and calculate the standard deviation of the residuals within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

```r
lm_fit <- lm(Force ~ Height * Species, data = crabs)
crabs <- crabs %>%
  mutate(
    residual = residuals(lm_fit)
  )

ggplot(data = crabs, mapping = aes(x = Height, y = residual)) +
  geom_point()
```



```r
ggplot(data = crabs, mapping = aes(x = residual, color = Species)) +
  geom_density()
```



```
crabs %>%
  group_by(Species) %>%
  summarize(sd(residual))
```

```
## # A tibble: 3 x 2
##   Species              `sd(residual)`
##   <chr>                         <dbl>
## 1 Cancer productus        4.824340538
## 2 Hemigrapsus nudus       2.167889758
## 3 Lophopanopeus bellus    5.353284342
```

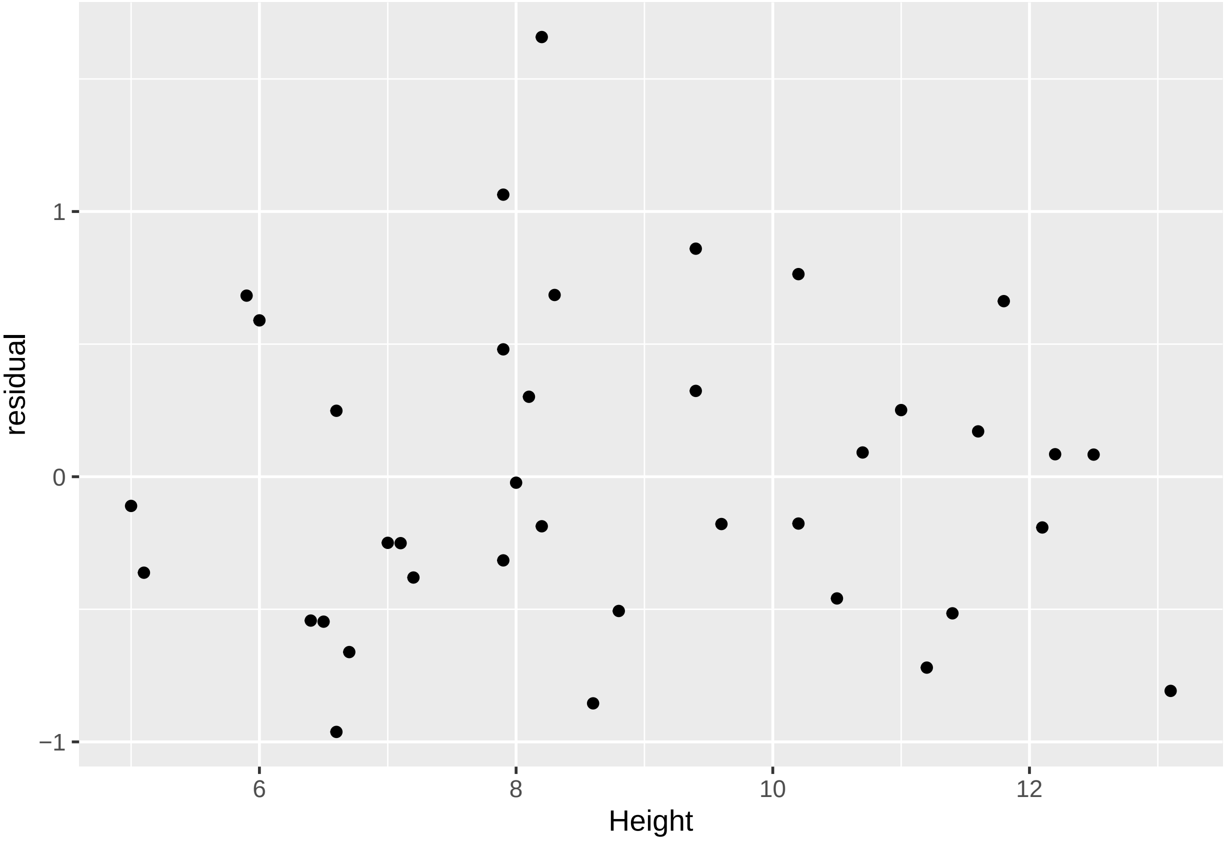
The main problem I see is that the standard deviation of the residuals is smaller for the Hemigrapsus nudus species than for the other two species. It does also look like the standard deviation of residuals may be slightly smaller for small values of Height than for large values of Height. I always have doubts about independence: were they careful in selecting the crabs for the sample "randomly"?

(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.

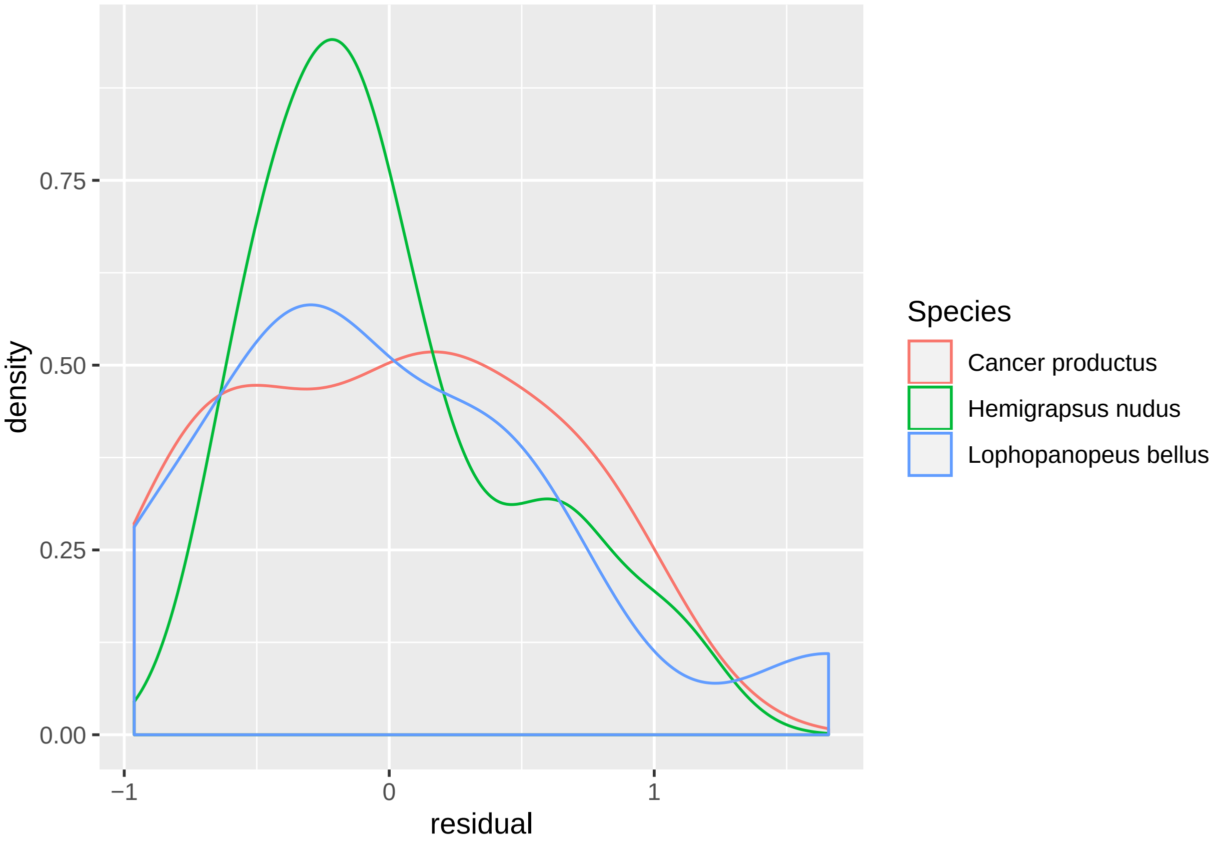
```
crabs <- crabs %>%
  mutate(
    sqrt_Force = sqrt(Force)
  )
```

```
lm_fit <- lm(sqrt_Force ~ Height * Species, data = crabs)
crabs <- crabs %>%
  mutate(
    residual = residuals(lm_fit)
  )

ggplot(data = crabs, mapping = aes(x = Height, y = residual)) +
  geom_point()
```



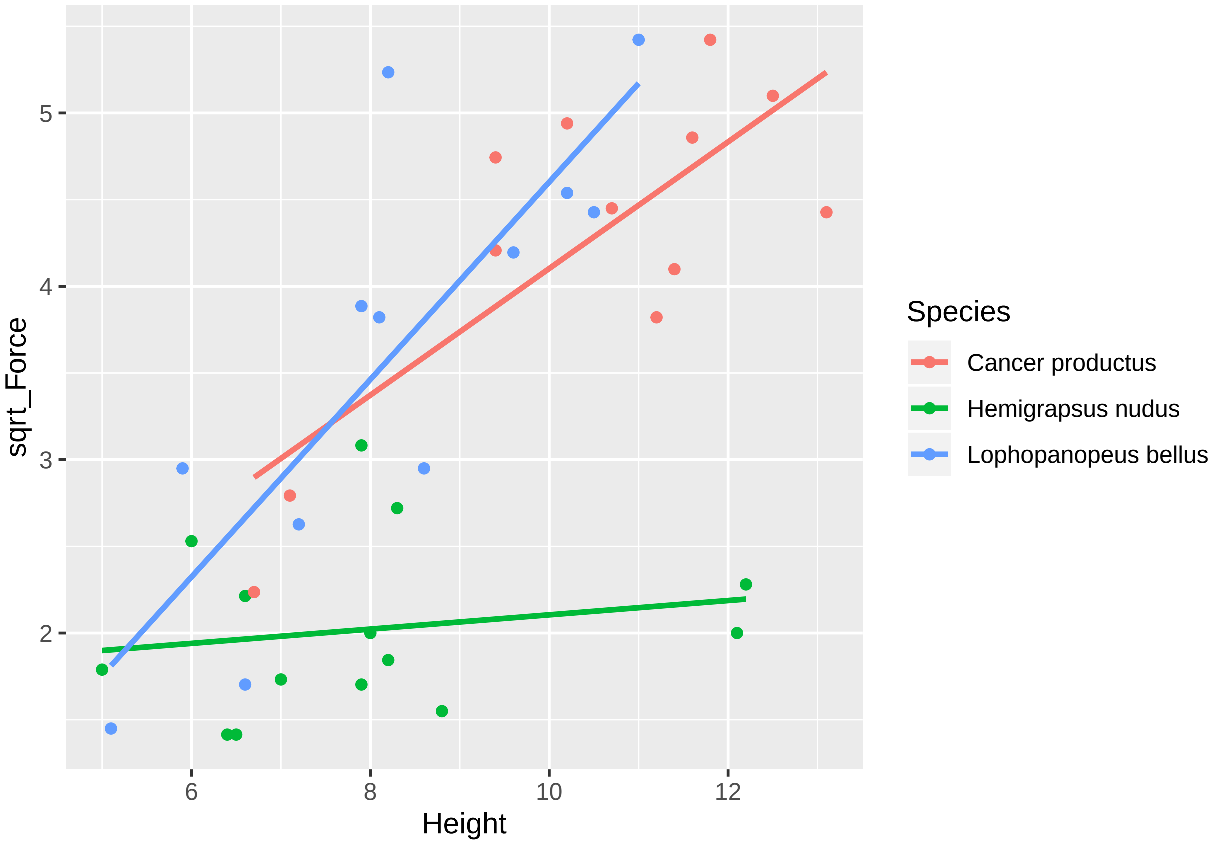
```
ggplot(data = crabs, mapping = aes(x = residual, color = Species)) +
  geom_density()
```



```
crabs %>%
  group_by(Species) %>%
  summarize(sd(residual))
```

```
## # A tibble: 3 x 2
##   Species              `sd(residual)`
##   <chr>                         <dbl>
## 1 Cancer productus       0.5913709475
## 2 Hemigrapsus nudus      0.4893722944
## 3 Lophopanopeus bellus   0.7265158084
```

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

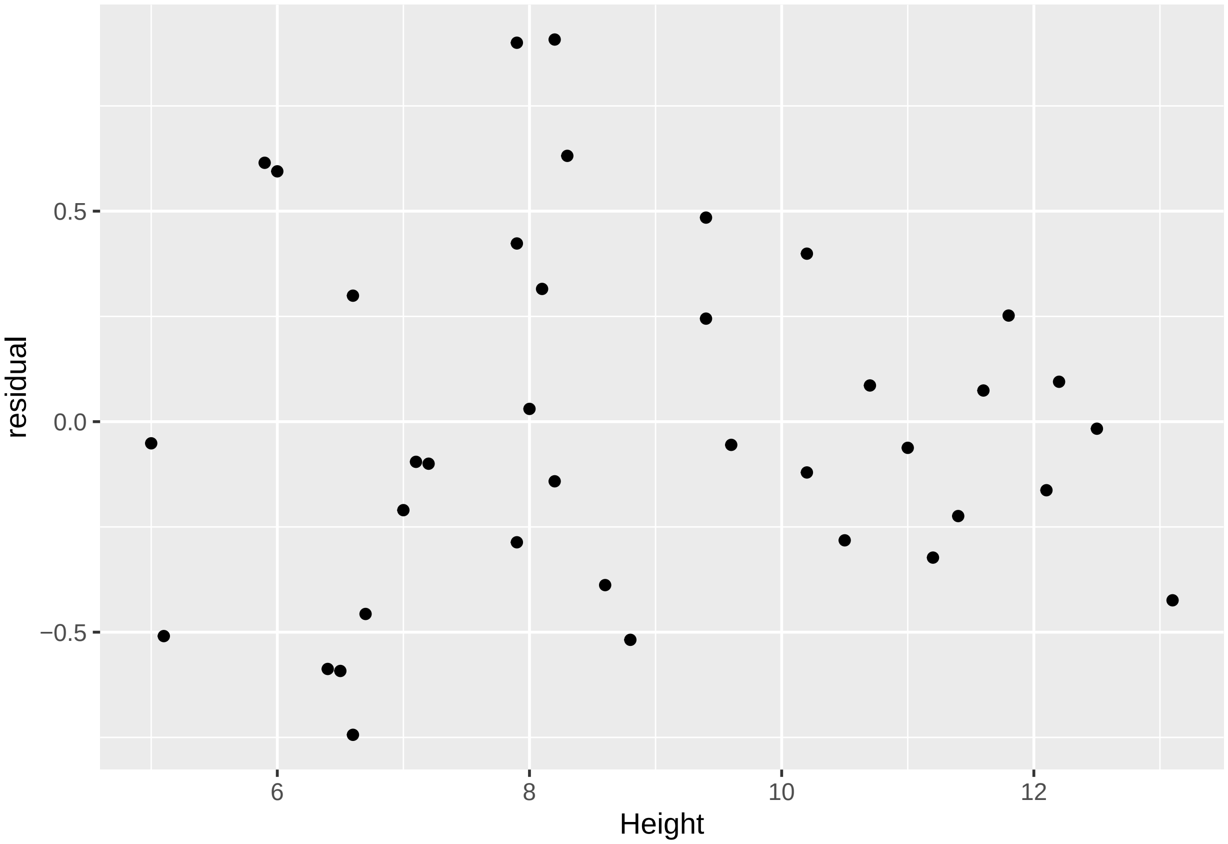


A square root transformation has helped, but the situation is not perfect. The standard deviation of residuals for the Hemigrapsus nudus group is smaller than for the other two groups, and the ratio of the largest standard deviation to the smallest is about 1.5. Really this is probably good enough, but you could also keep looking.

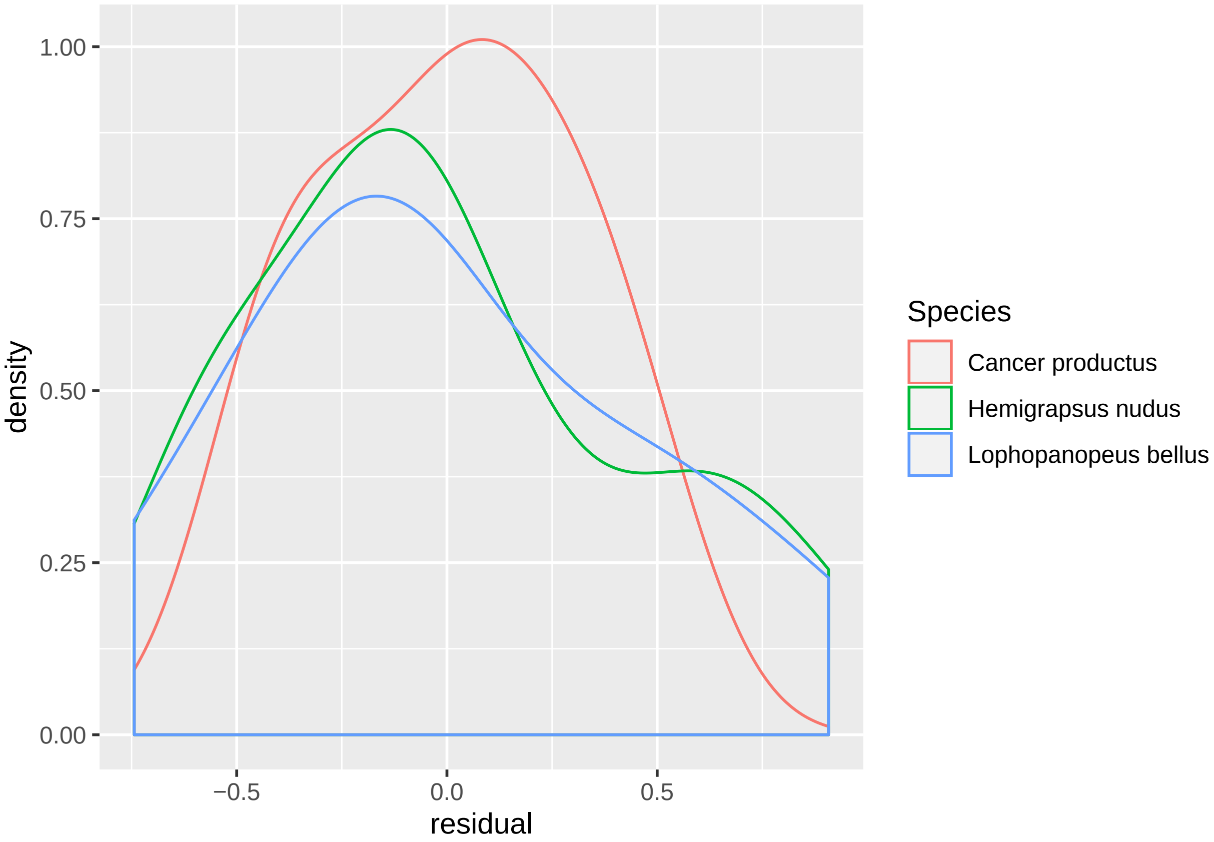
```
crabs <- crabs %>%
  mutate(
    log_Force = log(Force)
  )

lm_fit <- lm(log_Force ~ Height * Species, data = crabs)
crabs <- crabs %>%
  mutate(
    residual = residuals(lm_fit)
  )

ggplot(data = crabs, mapping = aes(x = Height, y = residual)) +
  geom_point()
```



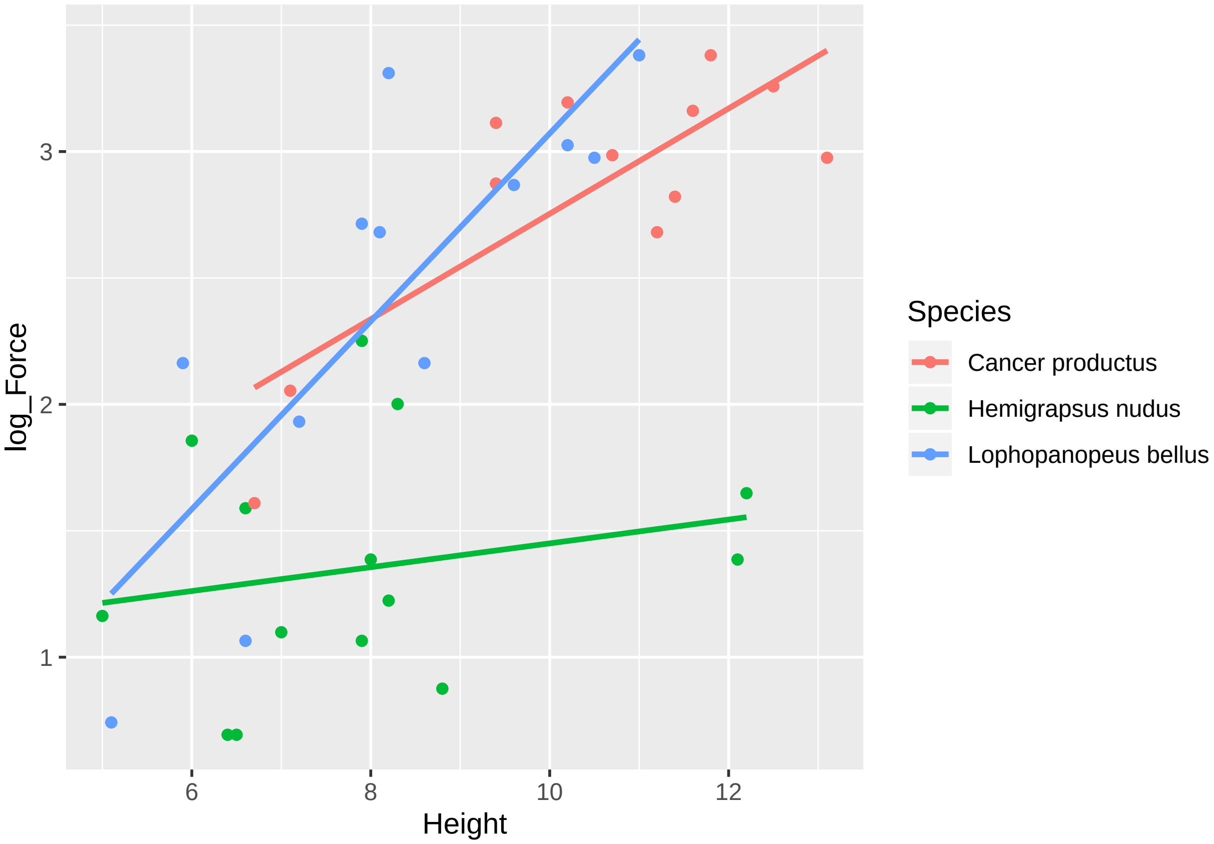
```
ggplot(data = crabs, mapping = aes(x = residual, color = Species)) +
  geom_density()
```



```
crabs %>%
  group_by(Species) %>%
  summarize(sd(residual))
```

```
## # A tibble: 3 x 2
##   Species              `sd(residual)`
##   <chr>                         <dbl>
## 1 Cancer productus       0.3137798207
## 2 Hemigrapsus nudus      0.4642135380
## 3 Lophopanopeus bellus   0.4819051524
```

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



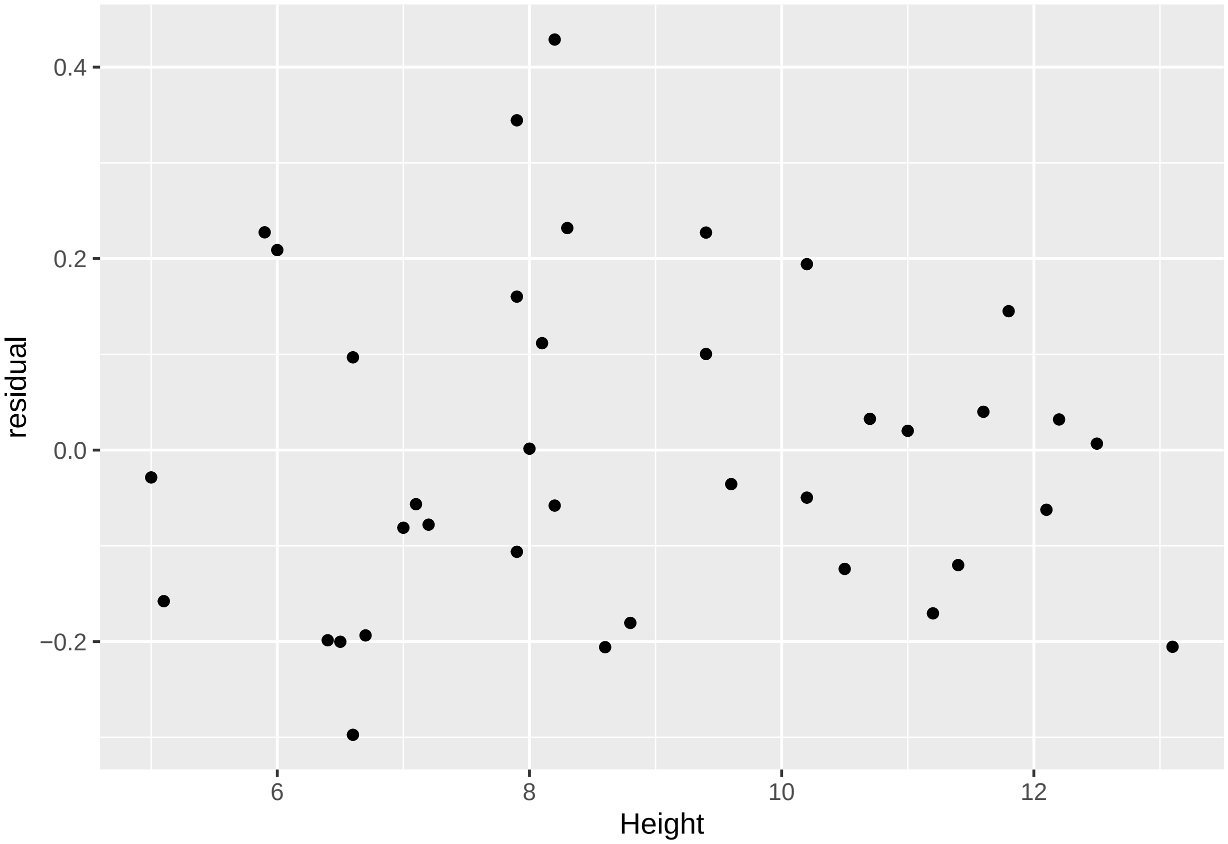
Now the standard deviation of residuals is smaller for the Cancer productus species than for the other two species. The ratio of the largest and smallest standard deviations is still about 1.5.

What if we try something in between?

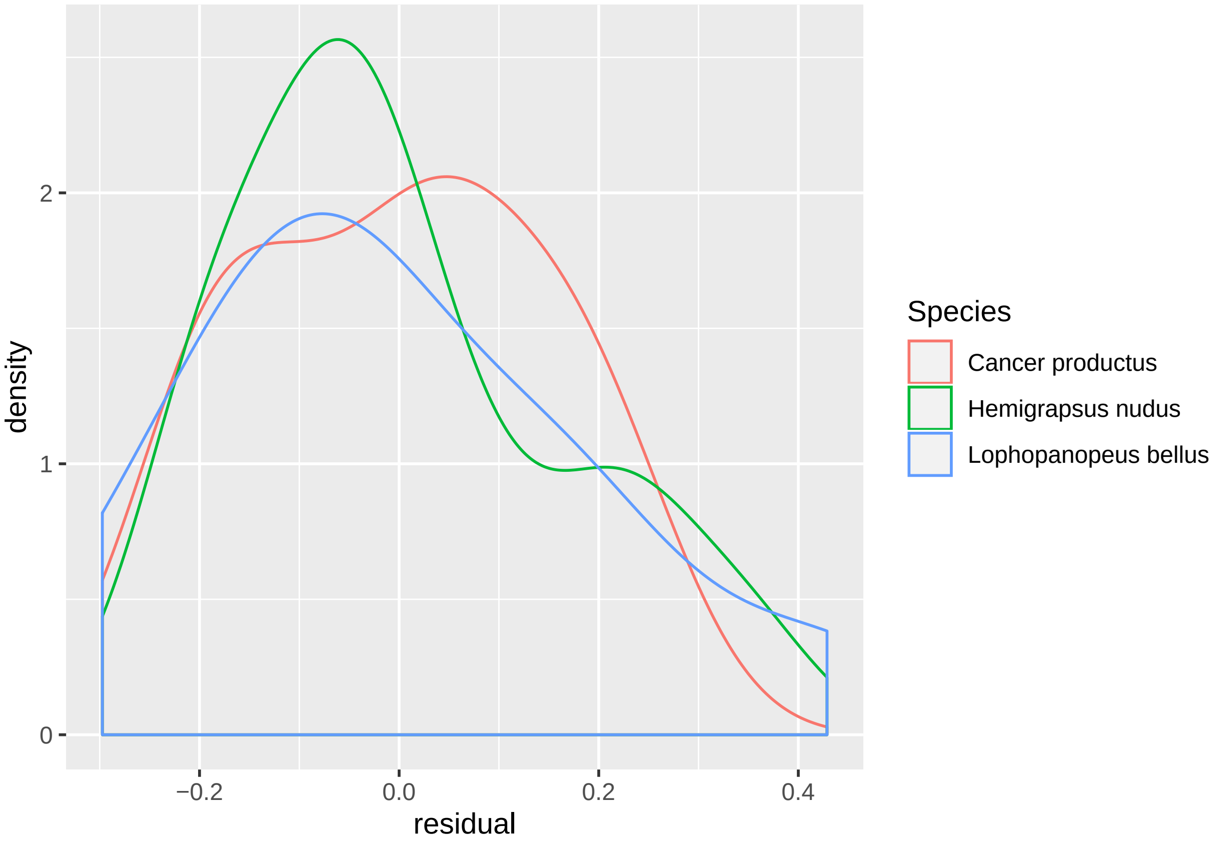
```
crabs <- crabs %>%
  mutate(
    Force_0.25 = Force^0.25
  )

lm_fit <- lm(Force_0.25 ~ Height * Species, data = crabs)
crabs <- crabs %>%
  mutate(
    residual = residuals(lm_fit)
  )

ggplot(data = crabs, mapping = aes(x = Height, y = residual)) +
  geom_point()
```



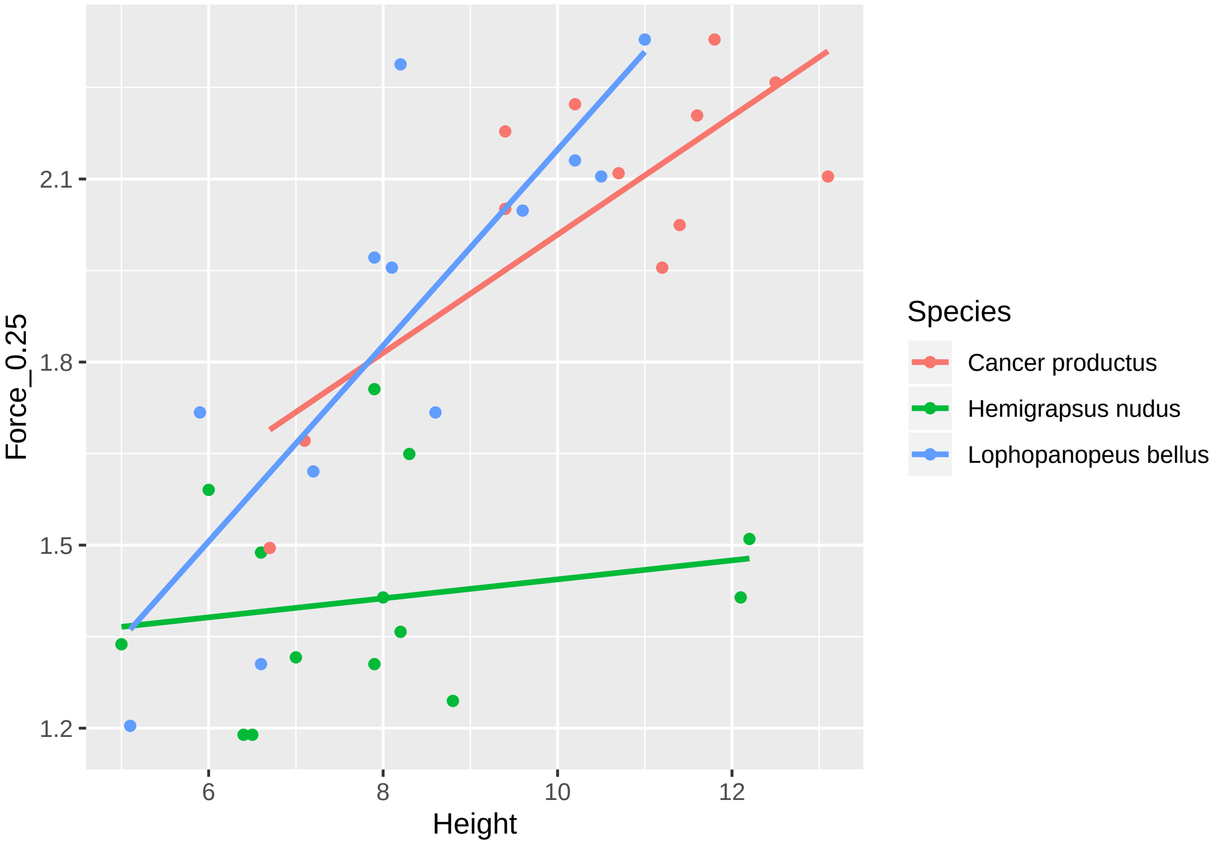
```
ggplot(data = crabs, mapping = aes(x = residual, color = Species)) +
  geom_density()
```



```
crabs %>%
  group_by(Species) %>%
  summarize(sd(residual))
```

```
## # A tibble: 3 x 2
##   Species              `sd(residual)`
##   <chr>                         <dbl>
## 1 Cancer productus       0.1506257697
## 2 Hemigrapsus nudus      0.1674519562
## 3 Lophopanopeus bellus   0.2033986264
```

```
ggplot(data = crabs, mapping = aes(x = Height, y = Force_0.25, color = Species)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



In terms of constant standard deviations across the three species and across values of Height, this transformation is best. There are some slight indications of non-linearity, but not too serious. The distributions of residuals are also skewed slightly right, but not too seriously. This is good enough.

```
summary(lm_fit)
```

```
##
## Call:
## lm(formula = Force_0.25 ~ Height * Species, data = crabs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29738 -0.12313 -0.03207  0.10885  0.42879
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           1.03919    0.29389   3.536  0.00126 **
## Height                                0.09697    0.02773   3.497  0.00141 **
## SpeciesHemigrapsus nudus              0.24906    0.35531   0.701  0.48839
## SpeciesLophopanopeus bellus          -0.49601    0.38628  -1.284  0.20834
## Height:SpeciesHemigrapsus nudus      -0.08141    0.03697  -2.202  0.03497 *
## Height:SpeciesLophopanopeus bellus    0.06351    0.04066   1.562  0.12816
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1828 on 32 degrees of freedom
## Multiple R-squared:  0.7986, Adjusted R-squared:  0.7671
## F-statistic: 25.38 on 5 and 32 DF,  p-value: 2.881e-10
```

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

$Y_i = \beta_0 + \beta_1 Height + \beta_2 SpeciesHemigrapsusnudus + \beta_3 SpeciesLophopanopeusbellus + \beta_4 Height * SpeciesHemigrapsusnudus + \beta_5 Height * SpeciesLophopanopeusbellus \; \varepsilon_i \sim \text{Normal}(0, \sigma)$

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

$\hat{\mu} = 1.039 + 0.097 Height + 0.249 SpeciesHemigrapsusnudus - 0.496 SpeciesLophopanopeusbellus - 0.081 Height * SpeciesHemigrapsusnudus + 0.064 Height * SpeciesLophopanopeusbellus$

(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.

$\hat{\mu} = (1.039 - 0.496) + (0.097 + 0.064) Height$

(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.

0.097

(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.

(0.097 - 0.081)

(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_fit)
```

```
##                                           2.5 %       97.5 %
## (Intercept)                          0.44055501  1.637818617
## Height                               0.04048431  0.153463459
## SpeciesHemigrapsus nudus            -0.47468225  0.972811747
## SpeciesLophopanopeus bellus         -1.28283427  0.290814927
## Height:SpeciesHemigrapsus nudus     -0.15671119 -0.006112945
## Height:SpeciesLophopanopeus bellus  -0.01931645  0.146328159
```

We are 95% confident that in the population of Hemigrapsus nudus crabs, a 1 mm increase in propodus height is associated with a change in mean claw closing force that is between 0.157 and 0.006 units smaller than the corresponding change 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.

$H_0 : \beta_4 = \beta_5 = 0$

H_A : At least one of β_4 and β_5 is not equal to 0.

```
reduced_lm_fit <- lm(Force_0.25 ~ Height + Species, data = crabs)
anova(reduced_lm_fit, lm_fit)
```

```
## Analysis of Variance Table
##
## Model 1: Force_0.25 ~ Height + Species
## Model 2: Force_0.25 ~ Height * Species
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     34 1.5584
## 2     32 1.0692  2   0.48926 7.3217 0.002408 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for the test is 0.0024. The data provide strong evidence against 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.

(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?

The test is an F test, based on the extra sum of squares for a comparison of a full model that allows for different slopes for all three species with a reduced model that has the same slope for all three species. We calculate the residual sum of squares for both models. The extra sum of squares is calculated as the difference between the residual sum of squares for the model with the same slope for all species and the residual sum of squares for the model with different slopes. The larger this extra sum of squares is, the stronger the indication that including different slopes is necessary.

The F statistic is calculated as

$$F = \frac{\text{(Extra Sum of Squares)/(Extra df)}}{\text{(Full Model Sum of Squares)/(Full df)}}$$

Since the extra sum of squares appears in the numerator of this calculation, a larger extra sum of squares means that the F statistic is larger; so, a large value of the F statistic offers strong evidence against the null hypothesis.

The degrees of freedom for the extra sum of squares is calculated as the difference in degrees of freedom for the reduced model and the degrees of freedom for the full model. The degrees of freedom for each model is calculated as the sample size minus the number of parameters for the mean.

(l) How were the β coefficients in your models above estimated? You can answer in just a sentence or two. I talked about this for about 3 minutes on Wed., Oct 16 and I just want you to remind yourself of this important idea that we have not spent much time on.

The β coefficients in a linear regression model are estimated by minimizing the sum of squared residuals.

Problem 2: Natal Dispersal Distances of Mammals (Sleuth3 problem 11.24)

Quote from the book:

> Natal dispersal distances are the distances that juvenile animals travel from their birthplace to their adult home. An assessment of the factors affecting dispersal distances is importan for understanding population spread, recolonization, and gene flow – which are central issues for conservation of many vertebrate species. For example, an understanding of dispersal distances will help to identify which species in a community are vulnerable to the loss of connectedness of habitat. To further the understanding of determinants of natal dispersal distances, researchers gathered data on body mass, diet type (herbivore, omnivore, or carnivore), and maximum natal dispersal distance for various mammals. . . . Analyze the data to describe the distribution of maximum dispersal distance as a function of body mass and diet type. Write a summary of statistical findings.

The following R code reads in the data.

```r
dispersion <- read_csv("http://www.evanlray.com/data/sleuth3/ex1124_natal_dispersion.csv")
```

```
## Parsed with column specification:
## cols(
##   Species = col_character(),
##   BodyMass = col_double(),
##   MaxDist = col_double(),
##   Type = col_character()
## )
```

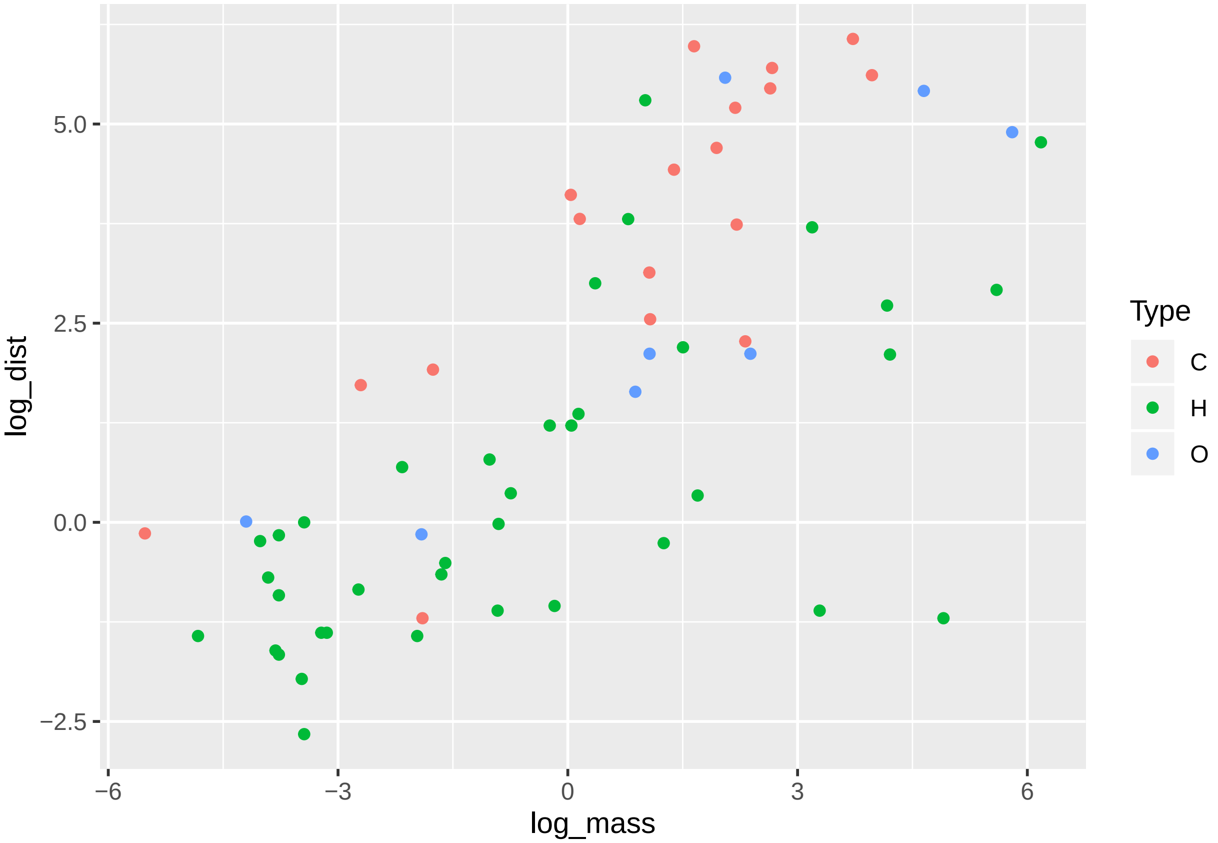
```r
head(dispersion)
```

```
## # A tibble: 6 x 4
##   Species                   BodyMass       MaxDist Type
##   <chr>                        <dbl>         <dbl> <chr>
## 1 Didelphis virginianus   2.415        5.150000000 O
## 2 Phascogale tapotafa     0.17200000   6.8         C
## 3 Trichosurus vulpecula   2.93        12.8         C
## 4 Sorex araneus           0.004        0.87        C
## 5 Scapanus townsendii     0.148        0.86        O
## 6 Ursus americanus      104.45       225           O
```

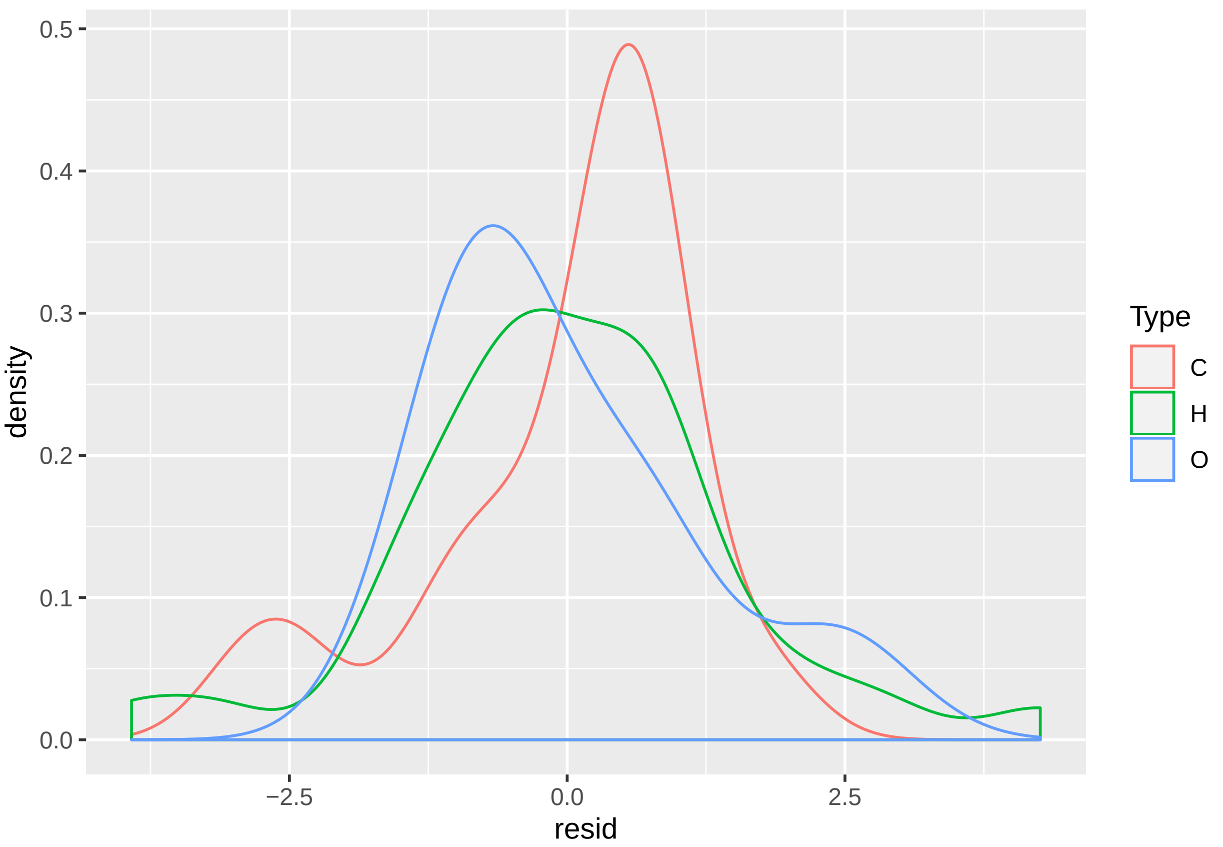
Here are things I will be looking more in more detail:

1. You will need to find a data transformation. Please justify your choice of transformation with a brief discussion of plots of the transformed data and residuals. Remember not to obsess about finding a perfect transformation; a good-enough transformation can be found on the steps of the ladder of powers.
2. The model you use should be justified (i.e., only allow for different slopes for the different diet types if the data indicate different slopes are necessary).
3. Your scientific conclusions should discuss, in context, conclusions that can be drawn about the associations between body mass, diet type, and natal dispersal distance. It would be good to discuss confidence intervals for these effects.

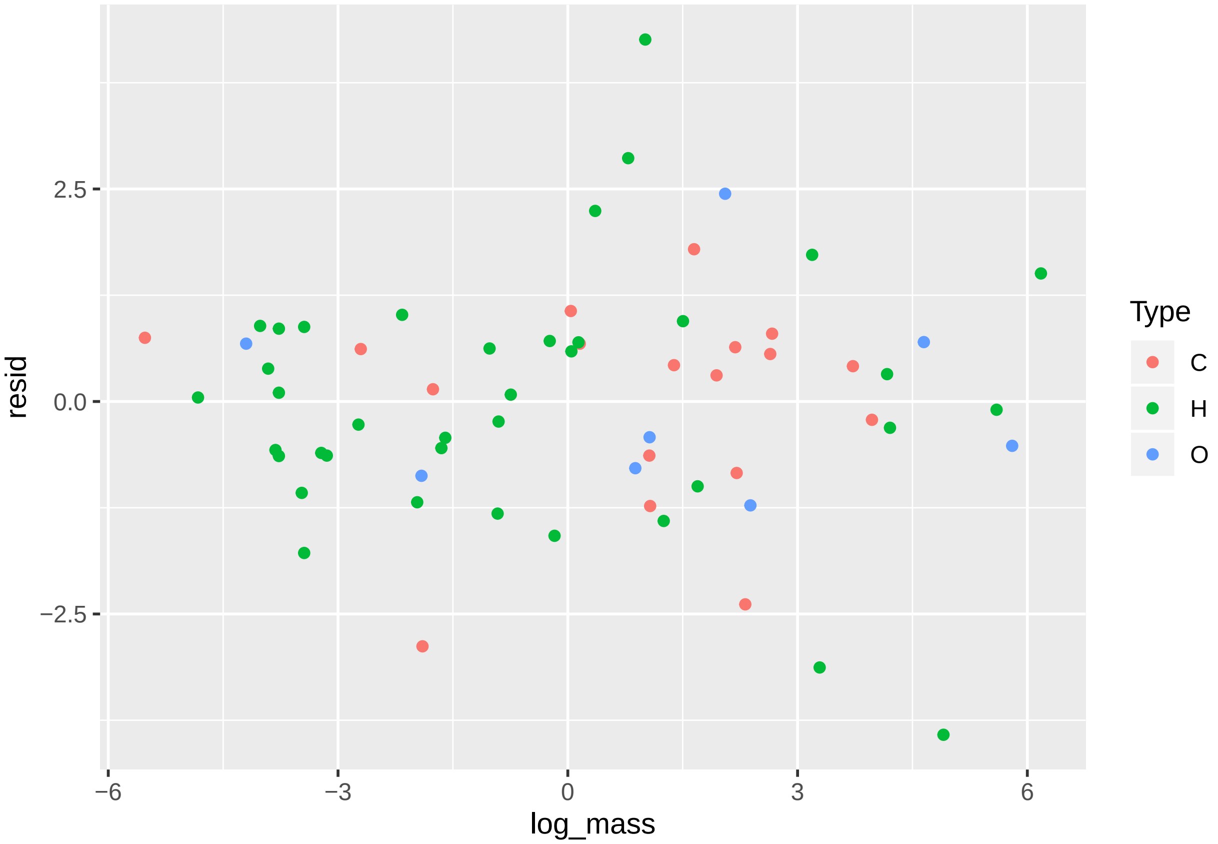
```
dispersion <- dispersion %>%
  mutate(
    log_dist = log(MaxDist),
    log_mass = log(BodyMass)
  )
ggplot(data = dispersion, mapping = aes(x = log_mass, y = log_dist, color = Type)) +
  geom_point()
```



```
lm_fit_different_slopes <- lm(log_dist ~ log_mass * Type, data = dispersion)
dispersion <- dispersion %>%
  mutate(
    resid = residuals(lm_fit_different_slopes)
  )
ggplot(data = dispersion, mapping = aes(x = resid, color = Type)) +
  geom_density()
```



```
ggplot(data = dispersion, mapping = aes(x = log_mass, y = resid, color = Type)) +
  geom_point()
```



```
dispersion %>%
  group_by(Type) %>%
  summarize(
    sd(resid)
  )
```

```
## # A tibble: 3 x 2
##   Type  `sd(resid)`
##   <chr>       <dbl>
## 1 C     1.193941844
## 2 H     1.499160872
## 3 O     1.209951114
```

The plots and summaries calculated above show that after applying a log transformation to the distances and the masses, the residuals follow a distribution that is approximately normal across all three groups, and is also fairly consistent in terms of standard deviation across all three groups and the range of values for mass. Additionally, the scatter plot of the transformed data shows approximately linear relationships between log mass and log distance within each group. As always, I find it difficult to assess the condition of independence. I could imagine that some of these species might be closely related, and might therefore have similar residuals around their respective means.

```
lm_fit_same_slopes <- lm(log_dist ~ log_mass + Type, data = dispersion)
anova(lm_fit_same_slopes, lm_fit_different_slopes)
```

```
## Analysis of Variance Table
##
## Model 1: log_dist ~ log_mass + Type
## Model 2: log_dist ~ log_mass * Type
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     60 124.46
## 2     58 117.64  2    6.8233 1.6821 0.1949
```

The data do not offer evidence that different slopes are required, so we will proceed with the model that uses the same slope for all three diet types.

```r
summary(lm_fit_same_slopes)
```

```
##
## Call:
## lm(formula = log_dist ~ log_mass + Type, data = dispersion)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3565 -0.6167 -0.0683  0.9041  4.1340
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.18440    0.34365   9.267 3.52e-13 ***
## log_mass     0.51061    0.06346   8.046 4.09e-11 ***
## TypeH       -2.53660    0.42128  -6.021 1.13e-07 ***
## TypeO       -1.16589    0.61282  -1.903   0.0619 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.44 on 60 degrees of freedom
## Multiple R-squared:  0.6878, Adjusted R-squared:  0.6721
## F-statistic: 44.05 on 3 and 60 DF,  p-value: 3.573e-15
```

```r
confint(lm_fit_same_slopes)
```

```
##                  2.5 %     97.5 %
## (Intercept)  2.4970089  3.8717975
## log_mass     0.3836742  0.6375539
## TypeH       -3.3792858 -1.6939169
## TypeO       -2.3917000  0.0599264
```

```r
lm_fit_mass_only <- lm(log_dist ~ log_mass, data = dispersion)
anova(lm_fit_mass_only, lm_fit_same_slopes)
```

```
## Analysis of Variance Table
##
## Model 1: log_dist ~ log_mass
## Model 2: log_dist ~ log_mass + Type
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     62 201.00
## 2     60 124.46  2    76.535 18.448 5.694e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The data indicate that an increase in log mass of one unit is associated with an increase in log natal distance of between 0.384 and 0.638 units (with 95% confidence), holding fixed the diet type.

Additionally, there is strong evidence of a difference in log natal distance between mammals with carnivorous diets and herbivorous diets, holding fixed the animal's mass. We are 95% confident that the mean log distance traveled is between about 3.38 and 1.69 units lower for herbivorous mammals than for carnivorous mammals, at a fixed body size. This suggests that carnivorous mammals tend to travel larger distances than herbivorous mammals.

There is not strong evidence of a difference in log natal distances between carnivorous and omnivorous animals; our 95% confidence interval for this difference at a fixed mass is [-2.37, 0.06]. However, the estimated difference for carnivorous mammals and omnivorous mammals is -1.16; when compared with the estimated difference of -2.54 for carnivorous mammals and herbivorous mammals, a consistent story emerges that mammals with more carnivorous diets tend to have larger natal distances than mammals with more herbivorous diets, on average.