Crab claw size and closing force. Problem 7.25, 10.9, and 10.10
Regression for all species at once, i.e., include dummy variables for *lb* and *cp*:

Where *lb* is the indicator variable for species 2, *L. bellus,* (i.e., *lb* = 1 for spp 2 and 0 for the other 2 species); *cp* is the indicator variable for species 3, *C. productus,* (i.e., *cp* = 1 for spp 3 and 0 for the other 2 species)

1.) Fit the full model, i.e., allow for separate regression lines (different slopes and intercepts) for each species

MODEL:
$$\hat{\mu}[\log(\text{force}) \mid \log(\text{height}), \text{species}] =$$
$$ß_0 + ß_1 \log(\text{height}) + ß_2 \text{lb} + ß_3 \text{cp} + ß_4 \text{lb} \times \log(\text{height}) + ß_5 \text{cp} \times \log(\text{height})$$

Thus, the regression model for *only* species 1, *H. nudus,* is (i.e., *lb* = 0, and *cp* = 0):

$$\hat{\mu}[\log(\text{force}) \mid \log(\text{height}), \text{species} = 1] = ß_0 + ß_1 \log(\text{height})$$

And, the regression model for *only* species 2, *L. bellus,* is (i.e., *lb* = 1, and *cp* = 0):

$$\hat{\mu}[\log(\text{force}) \mid \log(\text{height}), \text{species} = 2] =$$
$$ß_0 + ß_1 \log(\text{height}) + ß_2 + ß_4 \log(\text{height}) = (ß_0 + ß_2) + (ß_1 + ß_4) \log(\text{height})$$

And, the regression model for *only* species 3, *C. productus,* is (i.e., *lb* = 0, and *cp* = 1):

$$\hat{\mu}[\log(\text{force}) \mid \log(\text{height}), \text{species} = 3] =$$
$$ß_0 + ß_1 \log(\text{height}) + ß_3 + ß_5 \text{cp} \times \log(\text{height}) = (ß_0 + ß_3) + (ß_1 + ß_5) \log(\text{heigh t})$$

S-plus output:

```
      *** Linear Model ***

Call: lm(formula = log.force ~ log.height + lb + cp + log.height:lb + log.height:cp, data =
Ex0725, na.action = na.omit)

Coefficients:
               Value Std. Error t value Pr(>|t|)
  (Intercept)  0.5191  1.0001     0.5191  0.6073
   log.height  0.4083  0.4868     0.8387  0.4079
           lb -4.2992  1.5283    -2.8131  0.0083
           cp -2.4864  1.7606    -1.4123  0.1675
log.height:lb  2.5653  0.7354     3.4885  0.0014
log.height:cp  1.6601  0.7889     2.1043  0.0433

Residual SE: 0.4329 on 32 degrees of freedom
Multiple R-Squared: 0.7945
F-stat: 24.75 on 5 and 32 df, the p-value is 3.935e-010

Analysis of Variance Table

Terms added sequentially (first to last)
             Df Sum of Sq  Mean Sq  F Value      Pr(F)
   log.height  1  12.79793 12.79793 68.28824 0.00000000
           lb  1   2.80348  2.80348 14.95905 0.00050716
           cp  1   5.20634  5.20634 27.78041 0.00000904
log.height:lb  1   1.55457  1.55457  8.29501 0.00703654
log.height:cp  1   0.82985  0.82985  4.42797 0.04330026
    Residuals 32   5.99713  0.18741
```

2.) Can we use a simpler model and assume that the slope is the same for all 3 species? I.e., we want to see if the following reduced model would be OK (e.g., see exercise 10.10, pg. 284):

MODEL: $\hat{\mu}[\log(\text{force}) \mid \log(\text{height}), \text{species}] = \ß_0 + \ß_1 \log(\text{height}) + \ß_2 \text{lb} + \ß_3 \text{cp}$

We are not including the interaction terms *lb*log(height)* and *cp*log(height)* because we want to know if we can leave these out, i.e., assume equal slopes for all 3 species. Note that we are still allowing for the intercept to vary for the different species. To test if we can assume equal slopes, our null hypothesis is:

Ho: $\beta_4=\beta_5=0$ (*these betas refer to the full model parameterization*)
Ha: at least one of β_4 or β_5 is different from zero

Remember, go back to the full model and look at the regression equations for each species. β_4 is the difference in slopes between species 1 and 2 and β_5 is the difference in slopes between species 1 and 3.

We cannot use a t-test to test the above null hypothesis; it involves more than 2 parameters. So, we must use an F-test, i.e., the extra SS F-test.

To conduct the F-test, we must also run the reduced model:

S-plus output for the reduced model:

```
      *** Linear Model ***
Call: lm(formula = log.force ~ log.height + lb + cp, data = Ex0725, na.action = na.omit)

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) -2.0557  0.7480    -2.7483  0.0095
 log.height  1.6703  0.3608     4.6296  0.0001
         lb  0.9925  0.1960     5.0647  0.0000
         cp  1.0143  0.2207     4.5956  0.0001

Residual SE: 0.4965 on 34 df
Multiple R-Squared: 0.7129
F-stat: 28.14 on 3 and 34 df, the p-value is 2.489e-009

Analysis of Variance Table

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value       Pr(F)
log.height  1  12.79793 12.79793 51.91515 0.000000025
        lb  1   2.80348  2.80348 11.37240 0.001872254
        cp  1   5.20634  5.20634 21.11965 0.000057088
 Residuals 34   8.38155  0.24652
```

Thus, the Extra SS F-test is given by:

$$\text{F-stat} = \frac{\frac{\text{RSS(reduced)} - \text{RSS(full)}}{\#\text{ parameters in Ho}}}{\frac{\text{RSS(full)}}{\text{df for RSS (full)}}} = \frac{\frac{(8.38155 - 5.99713)}{2}}{\frac{5.99713}{32}} = \frac{1.19221}{0.18741} = 6.3615$$

And, using the S-plus command line, the p-value is `1-pf(6.3615, 2, 32) = 0.00471956`, which is much less than 0.05. Thus, we reject the null hypothesis; at least one of β_4 or β_5 is different from zero. So, we should not use the simpler model, we should model the slopes for each species separately.

3.) We could have conducted the above F-test just by looking at the sequential SS ANOVA table for the full model, which was given by:

Analysis of Variance Table

```
Terms added sequentially (first to last)
              Df Sum of Sq  Mean Sq  F Value      Pr(F)
   log.height  1  12.79793 12.79793 68.28824 0.00000000
           lb  1   2.80348  2.80348 14.95905 0.00050716
           cp  1   5.20634  5.20634 27.78041 0.00000904
log.height:lb  1   1.55457  1.55457  8.29501 0.00703654
log.height:cp  1   0.82985  0.82985  4.42797 0.04330026
    Residuals 32   5.99713  0.18741
```

And, since the parameters of interest (β_4 and β_5 i.e., the interactions `log.height:lb and log.height:cp`) are LAST in the ANOVA table, we can simply add up their corresponding SS to get the extra SS that is accounted for by modeling the slope separately for each species. Thus, the F-stat is:

$$\text{F-stat} = \frac{\dfrac{\text{Sum of the SS for the parameters of interst}}{\#\text{parameters in Ho}}}{\dfrac{\text{RSS (full)}}{\text{df for RSS (full)}}} = \frac{\dfrac{1.55457 + 0.82985}{2}}{\dfrac{5.99713}{32}} = \frac{1.19221}{0.18741} = 6.3615$$

Notice, this is the **same** F-stat that we got using the other method above, thus our p-value and conclusions are the same.

4.) Let's try another example. Suppose we want to test if we can use a simpler model and assume that the intercept is the same for all 3 species (but in this case we'll allow slopes to vary between species). I.e., we want to see if the following reduced model would be OK:

MODEL:

$$\hat{\mu}[\log(\text{force}) | \log(\text{height}), \text{species}] = \beta_0 + \beta_1 \log(\text{height}) + \beta_2 \text{lb} \times \log(\text{height}) + \beta_3 \text{cp} \times \log(\text{height})$$

So, we have 2 ways that we can test this, either run this reduced model in S-plus and compute the Extra SS F-test as in 2.) above, or use the Sequential SS ANOVA table from the full model, as in 3.) above. Let's use the sequential SS ANOVA table from the full model. The output in 1.) above for the full mode is:

```
      *** Linear Model ***

Call: lm(formula = log.force ~ log.height + lb + cp + log.height:lb + log.height:cp, data =
Ex0725, na.action = na.omit)
```

Analysis of Variance Table

```
Terms added sequentially (first to last)
              Df Sum of Sq  Mean Sq  F Value      Pr(F)
   log.height  1  12.79793 12.79793 68.28824 0.00000000
           lb  1   2.80348  2.80348 14.95905 0.00050716
```

```
           cp  1   5.20634  5.20634 27.78041 0.00000904
log.height:lb  1   1.55457  1.55457  8.29501 0.00703654
log.height:cp  1   0.82985  0.82985  4.42797 0.04330026
    Residuals 32   5.99713  0.18741
```

Can we use this ANOVA table to test the hypothesis that the intercepts do not differ between species? Where,
Ho: $\beta_2=\beta_3=0$ (***where these betas refer to the* ***full model parameterization*** ****)
Ha: at least one of β_2 or β_3 is different from zero

NO! We can't use the above ANOVA table because the variables (i.e., *lb* and *cp*) that correspond to β_2 and β_3 were NOT fit last. So, let's re-run the full model, and retype the model equation such that these variables are fit last:

S-plus output:

```
       *** Linear Model ***

Call: lm(formula = log.force ~ log.height + cp.log.height + lb.log.height + lb + cp, data =
Ex0725, na.action = na.omit)

Terms added sequentially (first to last)
              Df Sum of Sq  Mean Sq  F Value     Pr(F)
   log.height  1  12.79793 12.79793 68.28824 0.0000000
cp.log.height  1   1.72312  1.72312  9.19435 0.0047837
lb.log.height  1   7.15879  7.15879 38.19844 0.0000006
           lb  1   1.13853  1.13853  6.07507 0.0192639
           cp  1   0.37380  0.37380  1.99456 0.1675162
    Residuals 32   5.99713  0.18741
```

F-stat is: $$\frac{\frac{1.13853+0.37380}{2}}{\frac{5.99713}{32}}=\frac{0.756165}{0.18741}=4.0348$$

And, using the S-plus command line, the p-value is `1-pf(4.0348, 2, 32) = 0.0273749`, which is less than 0.05. Thus, we reject the null hypothesis; at least one of β_2 or β_3 is different from zero. So, we should not use the simpler model, we should model the intercepts for each species separately.

5.) One last example. We could have started our by asking if a simple regression model that assumed equal slopes and equal intercepts was appropriate, i.e.:

MODEL: $\hat{\mu}[\log(\text{force})|\log(\text{height}),\text{species}]=\beta_0+\beta_1\log(\text{height})$

We would have compared this reduced model to the full model in 1.) and using the sequential SS ANOVA table we could have tested the hypothesis:

Ho: $\beta_2=\beta_3=\beta_4=\beta_5=0$ (***where these betas refer to the* ***full model parameterization*** ****)
Ha: at least one of β_2 or β_3 or β_4 or β_5 is different from zero

Again, the Sequential SS ANOVA table from the full model was:

S-plus output:

```
      *** Linear Model ***

Call: lm(formula = log.force ~ log.height + cp.log.height + lb.log.height + lb + cp, data =
Ex0725, na.action = na.omit)

Terms added sequentially (first to last)
              Df Sum of Sq  Mean Sq  F Value     Pr(F)
   log.height  1  12.79793 12.79793 68.28824 0.0000000
cp.log.height  1   1.72312  1.72312  9.19435 0.0047837
lb.log.height  1   7.15879  7.15879 38.19844 0.0000006
           lb  1   1.13853  1.13853  6.07507 0.0192639
           cp  1   0.37380  0.37380  1.99456 0.1675162
    Residuals 32   5.99713  0.18741
```

(note, all of the variables that correspond to the parameters in the null hypothesis are last). And the F-stat is:

$$\text{F-stat is: } \frac{\frac{1.72312 + 7.15879 + 1.13853 + 0.37380}{4}}{\frac{5.99713}{32}} = \frac{5.19712}{0.18741} = 27.7313$$

And, using the S-plus command line, the p-value is `1-pf(27.7313, 2, 32) = 1.030959e-007`, which is much less than 0.05. Thus, we reject the null hypothesis; at least one of β_2 or β_3 or β_4 or β_5 is different from zero. So, we should not use the simpler model, we should model the each species separately.

NOTE: You should feel comfortable conducting the Extra SS F-test as presented in the 2 methods here. I.e., you need to know how to conduct the Extra SS F-test by:

1.) running the full and reduced model and getting the F-stat based on the residual SS (RSS in the above example) from both the reduced and full model.
2.) Running the full model and calculating the F-stat based on the sequential SS ANOVA table, making sure that the variable are fit in the correct order.