Unit 12 HW Solutions

## Question 1 (53 points total)

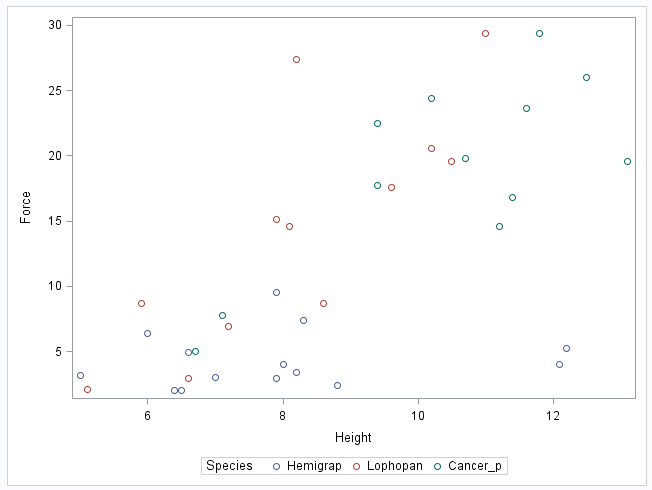
As part of a study of the effects of predatory intertidal crab species on snail populations, researchers measured the mean closing forces and the propodus heights of the claws on several crabs of three species. The crab data will be in your files repository.

* Use alpha = 0.05, where necessary.
* Use SAS and provide relevant code and output.

### Part A (9 points)

Step 1: Use the code from Dr. McGee’s lecture to plot a scatter plot of claw closing force (response variable) versus propodus height (explanatory variable), with different plotting symbols (or colors) to distinguish the three different crab species. Judging from an initial visual assessment of the scatterplots, you may apply a transformation and replot in this step. If a transformation is necessary, you only need to provide the scatterplot for the most visually satisfying transformation for now (but still provide a scatterplot of original data). You will formally assess the fit of the model in Step 4.

proc sgplot data = crab;  
scatter x = height y = force / group = species;  
run;



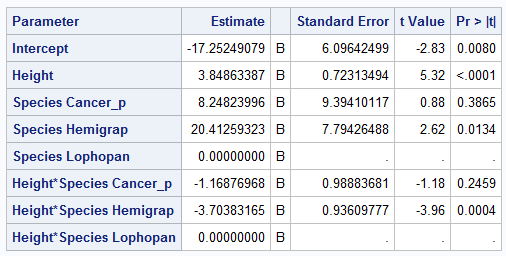
**There is no evidence that a transformation is necessary from the scatterplot.**

### Part B (9 points total)

Step 2: Build a model. (Simply write an appropriate equation as was shown in class.) This model should allow for separate fits (separate lines) for each crab species and should also allow for each line to have its own slope. Use lopho crab as the reference. (This is the default if the data is in alphabetical order.)

### Part C (10 points total)

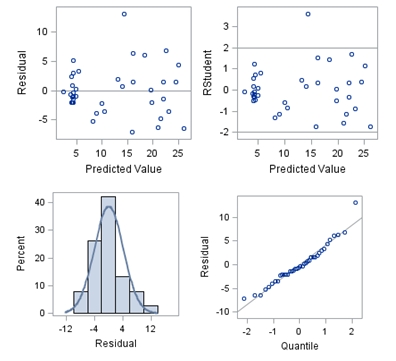
Step 3: Fit the model. (Fill in the relevant betas in your equation for step 2.) Make sure you provide relevant code and the table of parameter estimates as well.



**As intercept and slopes for cancer productus and lophopanopeus bellus are not significantly different (p-values = 0.3865, 0.2459, respectively), the next step could be to rerun the data by combining those two into the same group. We will omit that step here, as we will learn more about these procedures in the coming weeks.**

### Part D (10 points total)

Step 4: Provide a residual plot, studentized residual plot, histogram of residuals, and q-q plot of residuals to provide evidence of the appropriateness of the model. Provide a short one- or two-sentence discussion of EACH plot.



**(2.5 points) Residual Plot: The residual plot resembles somewhat of a random scatter of points around the 0 line, although there is a slight suspicion of non-constant variance judging from the dense cloud around the predicted value of 5.**

**(2.5 points) Studentized Residual Plot: This plot is very similar to the residual plot, although this plot identifies potential outlying observations. This plot identifies a potentially very outlying point with a predicted value of 15. This may provide some evidence against the normality assumption and this point should be examined further.**

**(2.5 points) Histogram of Residuals: The histogram of residuals displayed does not provide strong evidence that the residuals are not normally distributed.**

**(2.5 points) Q-Q Plot of Residuals: The Q-Q Plot of residuals provides no evidence against the residuals being normally distributed.**

**The model is a reasonable fit without transformations, although transformations may be investigated to handle the possible problem with equal standard deviations.**

### Part E (9 points total)

Step 5: If the fit assessed in Step 4 is sufficient, interpret each coefficient in the model.

**(1.5 points) : The intercept in this model provides an estimate (-17.25) of the force of the lopho crab (reference crab) with a height of zero. Of course, this is extrapolation and does not have a clear, practical meaning.**

**(1.5 points) : This is the adjustment of the intercept for a cancer crab with respect to a lopho crab. For a height of zero, the cancer crab has an estimated mean force 8.25 units more than the lopho crab.**

**(1.5 points) : This is the adjustment of the intercept for a hemi crab with respect to a lopho crab. For a height of zero, the hemi crab has an estimated mean force 20.41 units more than the lopho crab.**

**(1.5 points) : For each 1 unit increase in the height of a lopho crab, the estimated / predicted claw force increases 3.85 units.**

**(1.5 points) : For each 1 unit increase in the height of a cancer crab, the estimated / predicted claw force decreases 1.17 units from the change with the lopho crab.**

**(1.5 points) : For each 1 unit increase in the height of a hemi crab, the estimated / predicted claw force decreases 3.7 units from the change with the lopho crab.**

### Part F (6 points total)

Provide three individual regression equations (one for each crab species).

**(2 points) Lopho crab (cancer = 0 and hemi = 0):**

**(2 points) Cancer crab (cancer = 1 and hemi = 0):**

**(2 points) Hemi crab (cancer = 0 and hemi = 1):**

## Question 2 (47 points total)

Read the introduction to the Mammal Brain Weight data that starts on page 239 (Section 9.1.2). Download the Brain data set from 2DS. We would like to see if gestation length and litter size are associated with brain weight after controlling for different body sizes. That is, we already know that brain size is related to body weight; therefore, we don’t want body size to be a confounding variable. We would like to measure the association of the other variables after taking into account the body size.

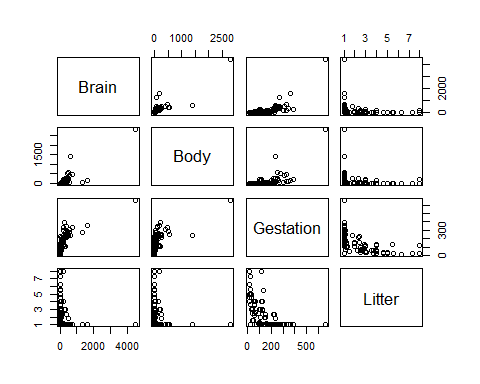
Answer this question by performing an analysis by following the 5 steps laid out in the problem above. Remember in step 2 to only include the terms that will help you answer this question of interest (QOI).

* Use alpha = 0.05, where necessary.
* Use R and provide relevant code and output.

**(9 points) Step 1: Scatterplots**

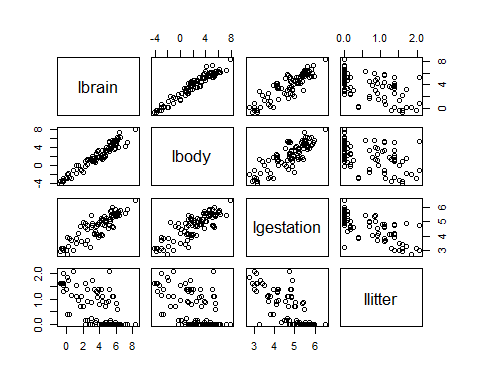
**Original data:**

brain <- read.csv("C:/Users/Charles/Documents/SMU/Online Teaching/MSDS 6371 - Statistical Foundations for Data Science/UNIT 12/Brain.csv")  
  
pairs (~ Brain + Body + Gestation + Litter, data = brain)



**Try a transformation with every relevant variable logged.**

lbrain <- log(brain$Brain)  
lbody <- log(brain$Body)  
lgestation <- log(brain$Gestation)  
llitter <- log(brain$Litter)  
pairs(~ lbrain + lbody + lgestation + llitter, data = brain)



**The log transformed variables look to visually satisfy assumptions of linear trend, constant variance and conditional normality. Independence will be assumed.**

**(9 points) Step 2: Build a model**

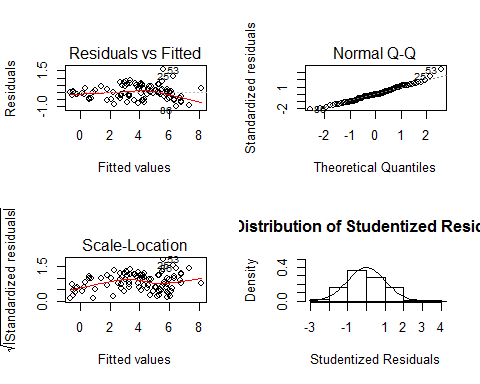
**(10 points) Step 3: Fit the model**

brain.lm <- lm(lbrain ~ lbody + lgestation + llitter, data = brain)  
summary(brain.lm)

##   
## Call:  
## lm(formula = lbrain ~ lbody + lgestation + llitter, data = brain)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.95415 -0.29639 -0.03105 0.28111 1.57491   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.85482 0.66167 1.292 0.19962   
## lbody 0.57507 0.03259 17.647 < 2e-16 \*\*\*  
## lgestation 0.41794 0.14078 2.969 0.00381 \*\*   
## llitter -0.31007 0.11593 -2.675 0.00885 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4748 on 92 degrees of freedom  
## Multiple R-squared: 0.9537, Adjusted R-squared: 0.9522   
## F-statistic: 631.6 on 3 and 92 DF, p-value: < 2.2e-16

**(10 points) Step 4: Provide a residual plot, studentized residual plot, histogram of residuals, and q-q plot of residuals to provide evidence of the appropriateness of the model.**

##Three of the graphs can be obtained by simply  
##plotting the model fit  
par(mfrow=c(2,2))  
plot(brain.lm, which=c(1:3))  
  
##Histogram with normal curve  
##Store studentized residuals  
studresbrain <- rstudent(brain.lm)  
  
##Histogram  
hist(studresbrain, freq=FALSE, main="Distribution of Studentized Residuals",  
xlab="Studentized Residuals", ylab="Density", ylim=c(0,0.5))  
  
##Create range of x-values for normal curve  
xfit2 <- seq(min(studresbrain)-1, max(studresbrain)+1, length=40)  
  
##Generate values from the normal distribution at the specified values  
yfit2 <- (dnorm(xfit2))  
  
##Add the normal curve  
lines(xfit2, yfit2, ylim=c(0,0.5))



**(2.5 points) Residual Plot: The residual plot resembles somewhat of a random scatter of points around the 0 line although there is a slight suspicion of non-constant variance, as we see a slight funnel shape as the predicted value increases. This is not too serious, and we will proceed with caution.**

**(2.5 points) Studentized Residual Plot: This plot is very similar to the Residual plot although this plot identifies potential outlying observations. This plot identifies two potentially very outlying points with a predicted value of around 6. This may provide some evidence against the normality assumption. Further analysis may be necessary on these points.**

**(2.5 points) Q-Q Plot of Residuals: The QQ Plot of residuals displayed below provides little evidence that the residuals are not normally distributed.**

**(2.5 points) Histogram of Residuals: The histogram of residuals displayed does not provide strong evidence that the residuals are not normally distributed.**

**(9 points) Step 5: Interpretation**

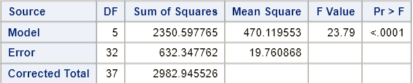
**Given that the model looks to fit (given the analysis in Step 4), we now move to interpret the parameter estimates found in Step 3. Both log gestation and log litter were found to be significant when log body was included in the model. This indicates that each doubling of the gestation period, holding all other variables constant, results in a multiplicative change in the median of the brain size. That translates to a 34% increase in the median of the brain size with every doubling of gestation period, holding the weight and litter size constant. In addition, holding all other variables constant, a doubling of the litter size is associated with a multiplicative decrease in the median of brain size. That equates to a 19% decrease in the median of brain size when the litter size is doubled, holding the gestation period and body size constant.**

## Bonus (+6 points total)

### Part A (+1)

How many degrees of freedom were used to estimate the error term (MSE) in question 1?

**From the regression output:**



**DF = 32**

### Part B (+1)

What is the estimate of the MSE in question 1?

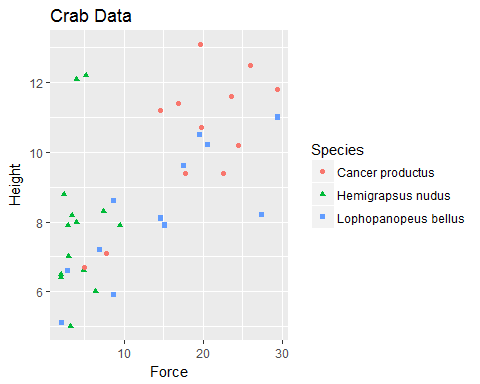
**MSE = 19.76**

### Part C (+1)

Repeat 1(a) in R.

*Note: there are multiple ways to do this. We have presented a ggplot approach. As long as your content matches, it should be counted as correct.*

library(ggplot2)  
  
crab <- read.csv("C:/Users/Charles/Documents/SMU/Online Teaching/MSDS 6371 - Statistical Foundations for Data Science/UNIT 12/Crab17.csv")  
  
ggplot(crab, aes(Force, Height, color=Species, shape=Species)) + geom\_point() + ggtitle("Crab Data")



### Part D (+1)

Repeat 1(c) in R

##Set the lopho as the reference level  
crab <- within(crab, Species <- relevel(Species, ref="Lophopanopeus bellus"))  
  
crab.lm <- lm(Force ~ Height\*Species, data=crab)  
summary(crab.lm)

##   
## Call:  
## lm(formula = Force ~ Height \* Species, data = crab)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.1458 -2.0996 -0.5013 1.8407 13.0937   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -17.2525 6.0964 -2.830 0.007977 \*\*   
## Height 3.8486 0.7231 5.322 7.78e-06 \*\*\*  
## SpeciesCancer productus 8.2482 9.3941 0.878 0.386476   
## SpeciesHemigrapsus nudus 20.4126 7.7943 2.619 0.013369 \*   
## Height:SpeciesCancer productus -1.1688 0.9888 -1.182 0.245929   
## Height:SpeciesHemigrapsus nudus -3.7038 0.9361 -3.957 0.000396 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.445 on 32 degrees of freedom  
## Multiple R-squared: 0.788, Adjusted R-squared: 0.7549   
## F-statistic: 23.79 on 5 and 32 DF, p-value: 6.419e-10

**The regression equation should be the same as in the first question.**

### Part E (+1)

How many degrees of freedom were used to estimate the error term (MSE) in question 2?

**DF = 92**

### Part F (+1)

What is the estimate of the error (MSE) in question 2?