**UNIT 12 HW**

1. 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.

1. 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.

Code:

SYMBOL1 V = 'Hemigrapsus\_nudus' C = 'Red';

SYMBOL2 V = 'Lophopanopeus\_bellus' C = 'Black';

SYMBOL3 V = 'Cancer\_productus' C = 'BLUEVIOLET';

Title 'Height and Claw Force of Different Species Crab';

proc gplot data = work.crab;

plot Force \* Height = Species;

run;

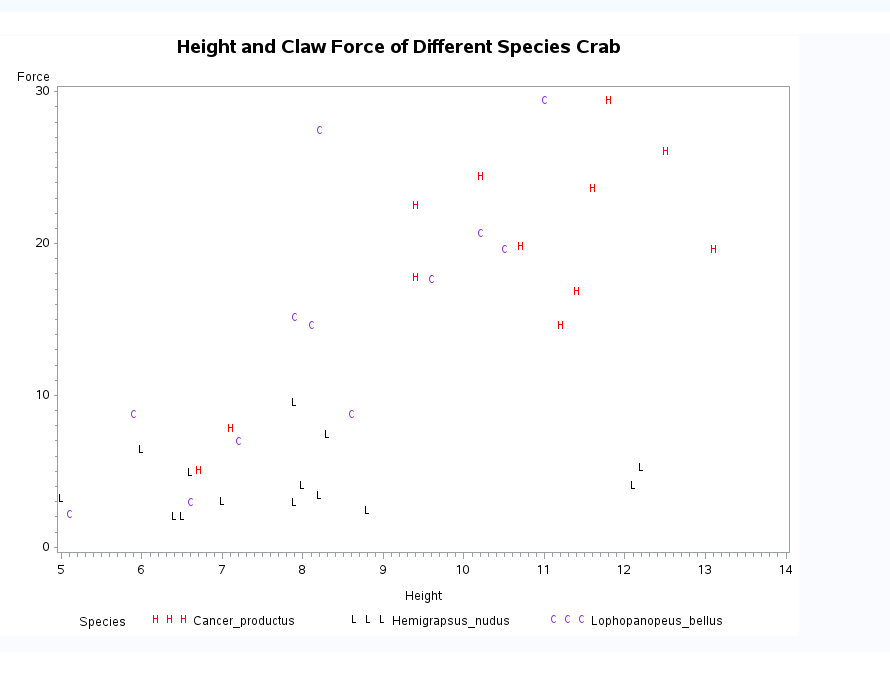
quit;

proc glm data = work.crab;

class Species (ref = "Lophopanopeus\_bellus");

model Force = Height \* Species / solution clparm;

run;



There appears to be a linear relationship with different slopes. The control group (L) is on the bottom and has a different slope than the other two. The other two seem close, but may be different.

1. 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.)

U{Force | Species, Height} = Bo +

B1 \* Hemigrapsus\_nudus

+ B2 \* Height

+ B3 Hemigrapsus\_nudus \* Height

+ B4 \* Cancer\_productus

+ B5 (Cancer\_productus \* Height)

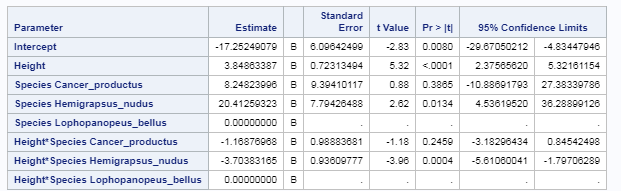
1. Step 3: Fit the model. (Fill in the relevant betas in your equation for step 2.) Make sure you provide relevant code and the table of parameter estimates as well.

proc glm data = work.crab plots = ALL;

class Species (ref = "Lophopanopeus\_bellus");

model Force = Height | Species / solution;

run;



U{Force | Species, Height} = -17.25 +

20.41 \* Hemigrapsus\_nudus

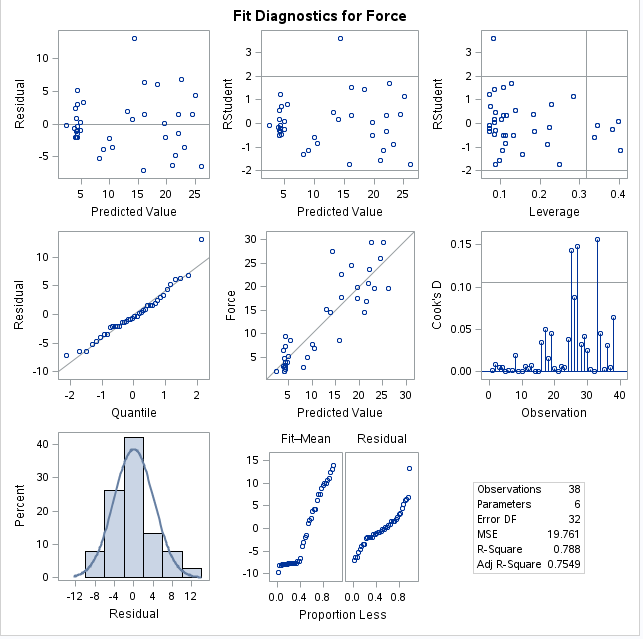
+ 3.85 \* Height

- 3.70 \* Hemigrapsus\_nudus \* Height

+ 8.25 \* Cancer\_productus

-1.17 (Cancer\_productus \* Height)

1. 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.

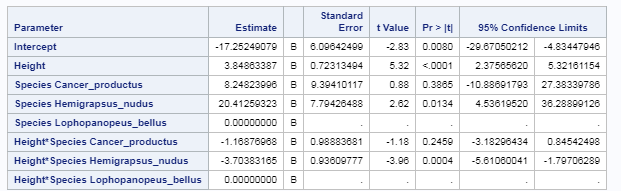


Normality: Judging from scatter plot, qq-plot, and histogram of residuals, there is some slight right skew, but the not strong evidence against normality.

Linear trend: It is tough to check linearity in multiple dimensions.

Equal SD: There is little evidence from the residual plots that there is unequal variance given the appearance of a random cloud in the residual plots.

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



The Species Cancer\_productus is predicted to have a force of 8.24 more than the Lophopanopeus\_bellus crab considering the same height. (p-value .39) A 95% confidence interval for this estimate is (-10.89,27.38)

The Hemigrapsus\_nudus is predicted to have a force of 20.41 more than the Lophopanopeus\_bellus crab considering the same height as well. (p-value .013) A 95% confidence interval for this estimate is (-4.53,36.28)

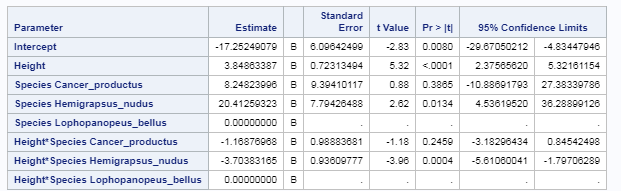
There wasn’t any good examples on how to interpret the other parameters, but will give this a shot.

The mean height of the reference crab is 2.84 with a p-value of 5.32 and a 95% confidence interval of (2.37,5.32)

The Cancer\_productus crab’s height is .14 of the reference crab with a p-value of .25 and a 95% confidence interval of (-3.18,.85)

The Hemigrapsus\_nudus crab’s height is 2.68 of the reference crab with a p-value of .25 and a 95% confidence interval of (-5.61,-1.79)

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



U{Force | Lophopanopeus, Height} = -17.25 +

+ 3.85 \* Height

U{Force | Cancer\_productus, Height} = -9 +

+ 3.84 \* Height

-1.17 \* (Cancer\_productus \* Height)

U{Force | Hemigrapsus, Height} =

3.17 +

3.84 \* Height

-3.70 Hemigrapsus\_nudus \* Height

1. 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.

Brain Weight = B0 \* Gestatation length + B1 \* Litter Size + B2 \* Body

1. Step 1: Use the code from Dr. McGee’s lecture to plot a scatter plot

ScatterPlot:

Here we don’t have a categorical variable like above. Not sure how the data should be partitioned to show a scatter plot. So, I’m going to show a scatter matrix instead.

Code:

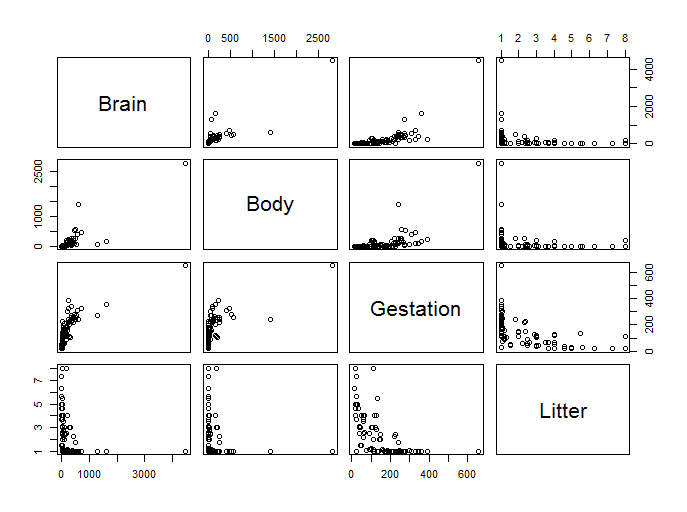
library(xlsx)

setwd("C:/Users/Marin Family/Desktop/Statistical Foundations for Data Science/Unit 12")

Brain <- read.xlsx(file = "Brain\_2\_2\_2.xlsx",sheetName = "Brain")

Brain$Body <- NULL

plot(Brain)



When looking at the scatterplots, there seem to be issues with linearity. Will try doing a log transformation of variables.

Code:

library(xlsx)

library(olsrr)

library(sqldf)

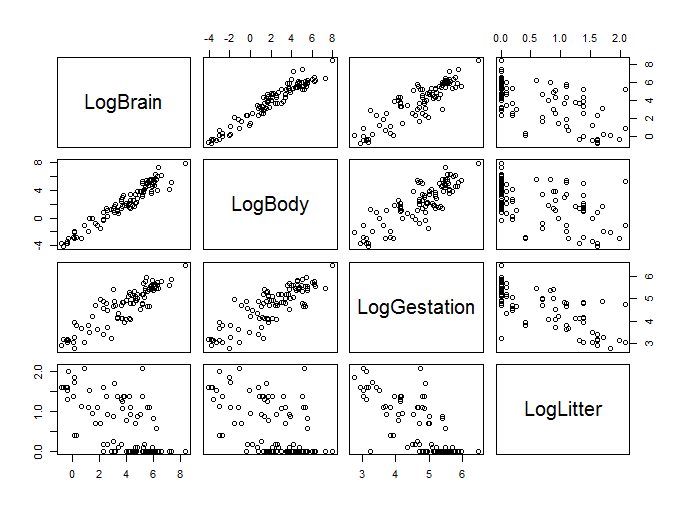
setwd("C:/Users/Marin Family/Desktop/Statistical Foundations for Data Science/Unit 12")

Brain <- read.xlsx(file = "Brain\_2\_2\_2.xlsx",sheetName = "Brain")

BrainLog <- sqldf("select Species, log(Brain) as LogBrain, log(Body) as LogBody, log(Gestation) as LogGestation, log(Litter)

as LogLitter from Brain")

plot(BrainLog)



Linearity seems to be better.

Step 2: Build a model. (Simply write an appropriate equation as was shown in class.)

Predlog(Brain) = B0 + B1log(Gestation) + B2log(Litter) + b3 (Body)

1. 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

Code:

BrainLog <- sqldf("select Species, log(Brain) as LogBrain, log(Body) as LogBody, log(Gestation) as LogGestation, log(Litter)

as LogLitter from Brain")

model <- lm(formula = LogBrain ~ LogBody + LogLitter + LogGestation, data = BrainLog )

summary(model)

Call:

lm(formula = LogBrain ~ LogBody + LogLitter + LogGestation, data = BrainLog)

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

LogBody 0.57507 0.03259 17.647 < 0.0000000000000002 \*\*\*

LogLitter -0.31007 0.11593 -2.675 0.00885 \*\*

LogGestation 0.41794 0.14078 2.969 0.00381 \*\*

---

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

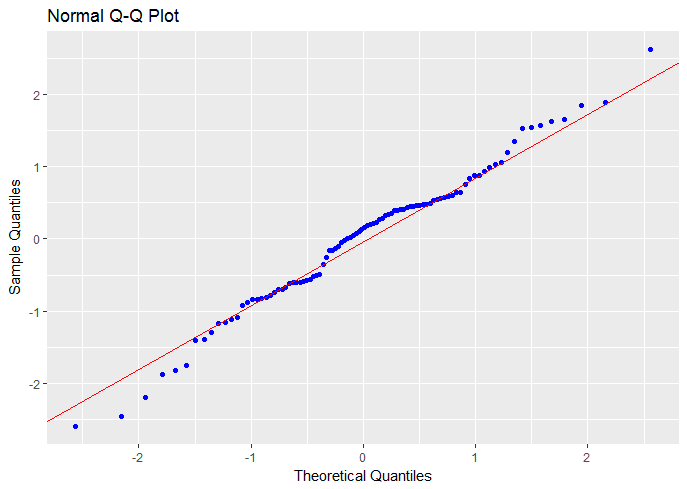
F-statistic: 631.6 on 3 and 92 DF, p-value: < 0.00000000000000022

1. 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.

Code:

#qqplot

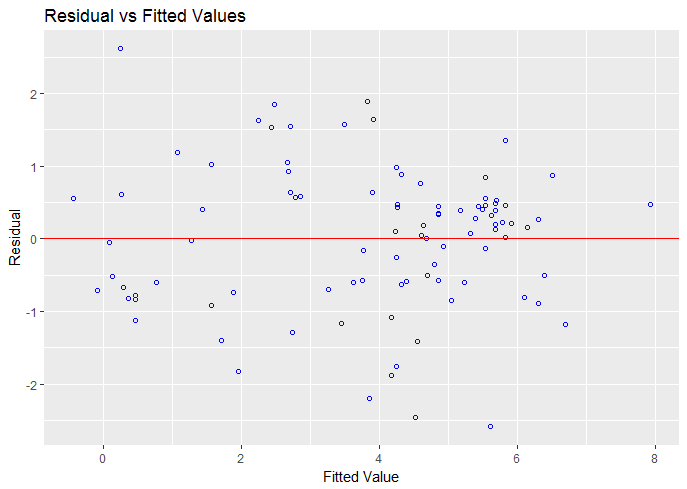
ols\_rsd\_qqplot(model)



QQPlot looks normal.

##residual vs fitted test

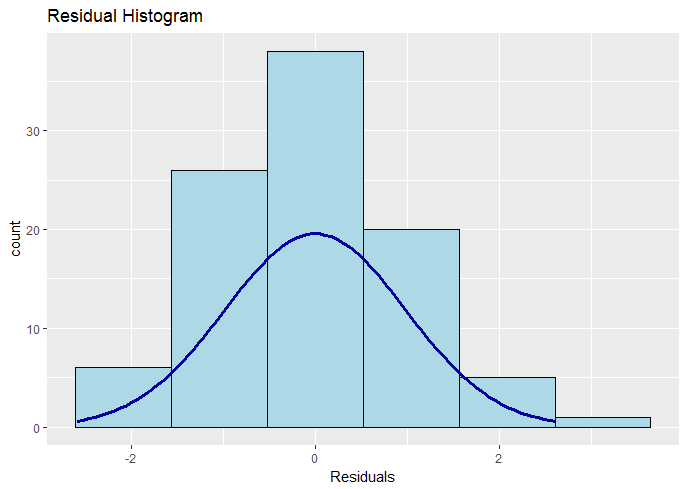
ols\_rvsp\_plot(model)



There seems to be equal variance given there is evidence of a random cloud.

##histogram

ols\_rsd\_hist(model)



Assumptions of normality look good.

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

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.85482 0.66167 1.292 0.19962

LogBody 0.57507 0.03259 17.647 < 0.0000000000000002 \*\*\*

LogLitter -0.31007 0.11593 -2.675 0.00885 \*\*

LogGestation 0.41794 0.14078 2.969 0.00381 \*\*

A change in Brain size is a 10-fold increase. Holding body mass constant, it is estimated that a 10-fold increase in Litter size is associated with a (10**-.31007**= 02.63) 263% increase in the median Brain size of these mammal species (p-value = **.004**). A 95% confidence interval for the multiplicative decrease is (10**-.54**,10**-.080**) = (.29,.83) which equates to an estimated decrease between 29% and 67%.

A change in Brain size is a 10-fold increase. Holding body mass constant, it is estimated that a 10-fold increase in Gestation time is associated with a (10**.42**= 0.49) 51% decrease in the median Brain size of these mammal species (p-value = **0.008**). A 95% confidence interval for the multiplicative decrease is (10**.13**,10**.70**) = (1.37,4.98) .

1. Provide individual regression equations

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -225.29213 83.05875 -2.712 0.00797 \*\*

Body 0.98588 0.09428 10.457 < 0.0000000000000002 \*\*\*

Litter 27.64864 17.41429 1.588 0.11579

Gestation 1.80874 0.35445 5.103 0.00000179 \*\*\*

Predlog(Brain) = -225 + .98 log(gestation) + 27.65 log(Litter)

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

5

1. What is the estimate of the MSE in question 1?

470.12

1. Repeat 1(a) in R.
2. Repeat 1(c) in R.
3. How many degrees of freedom were used to estimate the error term (MSE) in question 2?
4. What is the estimate of the error (MSE) in question 2?