|  |
| --- |
| title: "Statistical Foundations for Data Science - Assignment8 " |
| author: "Travis Deason" |
| date: "Oct 22nd 2017" |
| output: word\_document |

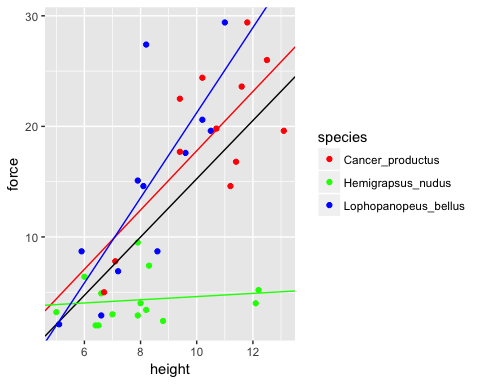
# MSDS 6371 UNIT 10 HW

* These are the same data from last week’s HW. Now, we are going to use them for simple linear regression.

## # 2.

## [1] Hemigrapsus\_nudus Lophopanopeus\_bellus Cancer\_productus   
## Levels: Cancer\_productus Hemigrapsus\_nudus Lophopanopeus\_bellus

## a.  
# find unique sepcies values  
specs <- unique(crabs$species)  
## b. Generate model  
model <- lm(force~height, data=crabs)  
model\_cancer <- lm(force~height, data=subset(crabs, species == 'Cancer\_productus'))  
model\_hemi <- lm(force~height, data=subset(crabs, species == 'Hemigrapsus\_nudus'))  
model\_lapho <- lm(force~height, data=subset(crabs, species == 'Lophopanopeus\_bellus'))  
coeff <- coefficients(model)  
group.colors <- c(Hemigrapsus\_nudus = "green",  
 Lophopanopeus\_bellus = "blue",  
 Cancer\_productus ="red")  
ggplot(crabs, aes(x=height, y=force, colour=species)) +  
 geom\_point() +  
 # plot regression for all species  
 scale\_colour\_manual(values=setNames(group.colors, specs)) +  
 # plot regression for all species  
 geom\_abline(intercept= coeff[1],1, slope= coeff[2],1) +  
 # plot regression for cancer  
 geom\_abline(intercept= model\_cancer$coefficients[1],1,   
 slope= model\_cancer$coefficients['height'],1, color='red') +  
 # plot regression for hemi  
 geom\_abline(intercept= model\_hemi$coefficients[1],1,   
 slope= model\_hemi$coefficients['height'],1, color='green') +  
 # plot regression for lapho  
 geom\_abline(intercept= model\_lapho$coefficients[1],1,   
 slope= model\_lapho$coefficients['height'],1, color='blue')



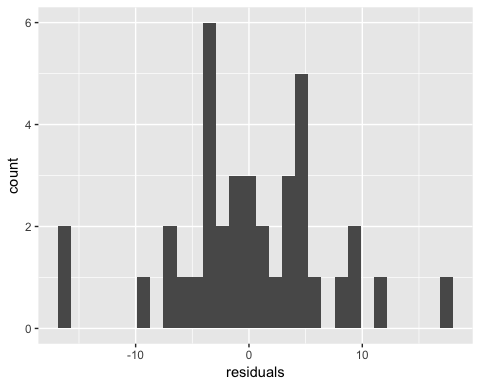
print('all crabs')

## [1] "all crabs"

print(summary(model))

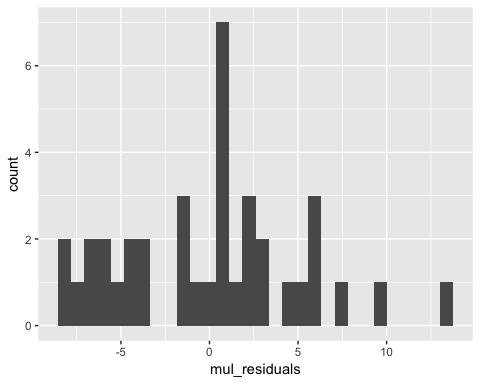
##   
## Call:  
## lm(formula = force ~ height, data = crabs)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.7945 -3.8113 -0.2394 4.1444 16.8814   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11.0869 4.6224 -2.399 0.0218 \*   
## height 2.6348 0.5089 5.177 8.73e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.892 on 36 degrees of freedom  
## Multiple R-squared: 0.4268, Adjusted R-squared: 0.4109   
## F-statistic: 26.8 on 1 and 36 DF, p-value: 8.73e-06

crabs$residuals <- model$residuals  
ggplot(crabs, aes(x=residuals)) + geom\_histogram()

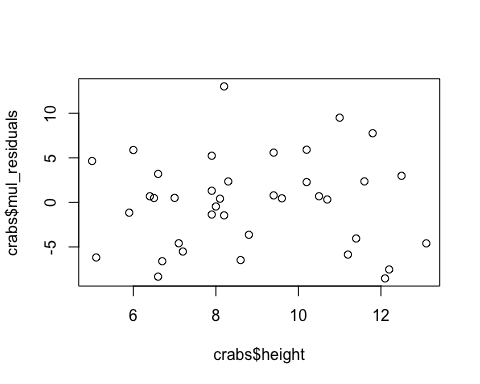


# the residuals seem to be bimodal, which could be an indication that multipule linear regression is required in this model with a dummy variable for the forms of species

crabs$cancer <- crabs$species == 'Cancer\_productus'  
crabs$hemi <- crabs$species == 'Hemigrapsus\_nudus'  
crabs$lapho <- crabs$species == 'Lophopanopeus\_bellus'  
  
multi\_model <- lm(force~height +cancer + hemi + lapho, data=crabs)  
  
crabs$mul\_residuals <- multi\_model$residuals  
ggplot(crabs, aes(x=mul\_residuals)) + geom\_histogram()



plot(crabs$height, crabs$mul\_residuals)



summary(multi\_model)

##   
## Call:  
## lm(formula = force ~ height + cancer + hemi + lapho, data = crabs)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.5316 -4.4548 0.4752 2.8254 13.0153   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.7549 4.0710 -0.431 0.669132   
## height 1.9682 0.4570 4.307 0.000133 \*\*\*  
## cancerTRUE 0.1693 2.4020 0.070 0.944209   
## hemiTRUE -9.5292 2.1106 -4.515 7.24e-05 \*\*\*  
## laphoTRUE NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.352 on 34 degrees of freedom  
## Multiple R-squared: 0.6735, Adjusted R-squared: 0.6447   
## F-statistic: 23.38 on 3 and 34 DF, p-value: 2.152e-08

* For the Crabs dataset, using dummy variables to substitute for species, our model gives us an adjusted R squared of .6447, and the p-value for height is .000133, .944209 for the cancer dummy variable, .000724 for the hemi dummy variable. This indicates that the Lophopanopeus\_bellus species dummy variable added the most value to to model; while the other crab types showed a similarly strong linear relationship between height and claw force. This indicates that there is a strong linear relationship between height and force for most crab types, but this is not true for Lophopanopeus\_bellus crabs. The model, with 34 degrees of freedom shows that for every 1 cm of additional height, claw force increases by 1.9N. for the Hemigrapsus\_nudus there is an offset of 9.52 Newtons (decreased force), an additional .1693N for the Cancer\_productus, and the Lophopanopeus\_bellus are represented by setting all dummy values to zero. The Intercept value for this model is -1.75 N which would imply a crab with no height would have a clamp force of -1.75N. This interpetation is impossible because a crab could not have a height of zero, and it is far out of sample.

## # 2.

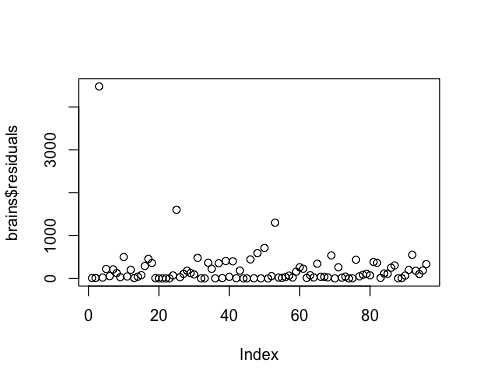
library(reshape2)  
brains$gest\_norm <- (brains$gestation - min(brains$gestation)) /   
 max(brains$gestation) - min(brains$gestation)  
  
brains$litter\_norm <- (brains$litter - min(brains$litter)) /   
 max(brains$litter) - min(brains$litter)  
  
#brain\_melt <- melt(brains, id.vars = "brain")  
#brain\_melt <- subset(brain\_melt, variable == 'litter\_norm' | variable == 'gest\_norm')  
  
#brain\_melt$value <- sapply(brain\_melt$value, as.numeric)  
  
br\_model <- lm(brain~gestation+litter, data=brains)  
brains$resid <- br\_model$residuals  
summary(br\_model)

##   
## Call:  
## lm(formula = brain ~ gestation + litter, data = brains)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -870.74 -133.59 -23.31 108.06 2235.38   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -617.8022 109.0185 -5.667 1.62e-07 \*\*\*  
## gestation 4.2411 0.3935 10.778 < 2e-16 \*\*\*  
## litter 84.5180 24.3414 3.472 0.000785 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 330.4 on 93 degrees of freedom  
## Multiple R-squared: 0.5842, Adjusted R-squared: 0.5752   
## F-statistic: 65.32 on 2 and 93 DF, p-value: < 2.2e-16

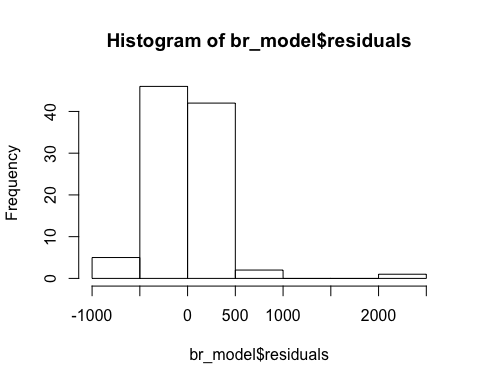
#all\_model <- lm(brain~gest\_norm + litter\_norm, data=brains)  
gest\_model <- lm(brain~gestation, data=brains)  
litter\_model <- lm(brain~litter, data=brains)  
  
summary(br\_model)

##   
## Call:  
## lm(formula = brain ~ gestation + litter, data = brains)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -870.74 -133.59 -23.31 108.06 2235.38   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -617.8022 109.0185 -5.667 1.62e-07 \*\*\*  
## gestation 4.2411 0.3935 10.778 < 2e-16 \*\*\*  
## litter 84.5180 24.3414 3.472 0.000785 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 330.4 on 93 degrees of freedom  
## Multiple R-squared: 0.5842, Adjusted R-squared: 0.5752   
## F-statistic: 65.32 on 2 and 93 DF, p-value: < 2.2e-16

group.colors <- c(litter\_norm = "red",  
 gest\_norm = "blue")  
  
plot(x=brains$brain, y=brains$residuals)



hist(br\_model$residuals)



* For the model on the dataset, the adjusted Rsquared value generated by a multipule linear regression on brain weight verses litter size and getstation time is .5752; which means that 57.52 percent of the variance in brain weight can be accounted for by the realtionship with litter size and gestation time.
* The coefficents in this model are 4.2411 for gestation time in weeks with a t-value of 10.778 and a p value of ~0. This means that for each additional week of gestation time, there isa n estimated increase of brain mass of 4.2411 grams. For the litter size the coefficent is 84.5180. This means that for each additional offspring brought in a litter, there is an additonal 84 grams of brain mass. This relationship has a t-value of 3.472 and a p value of .000785. There are 93 degrees of freedom in this model, and both coefficents are significaent at a 95 percent confidence level. The residuals are normally distributed enough that the model can be considered statistically significent, and the MSE is 18.176 grams