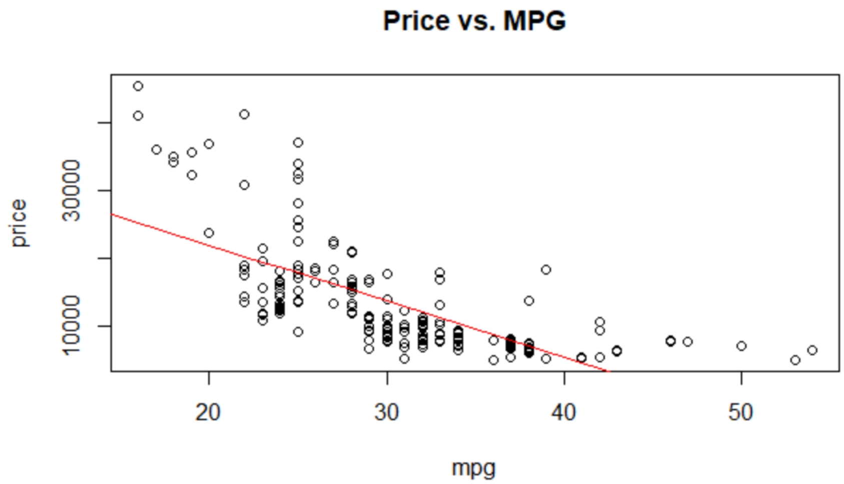
MATH 3060 Homework

March 19, 2022

Project 1: Cars

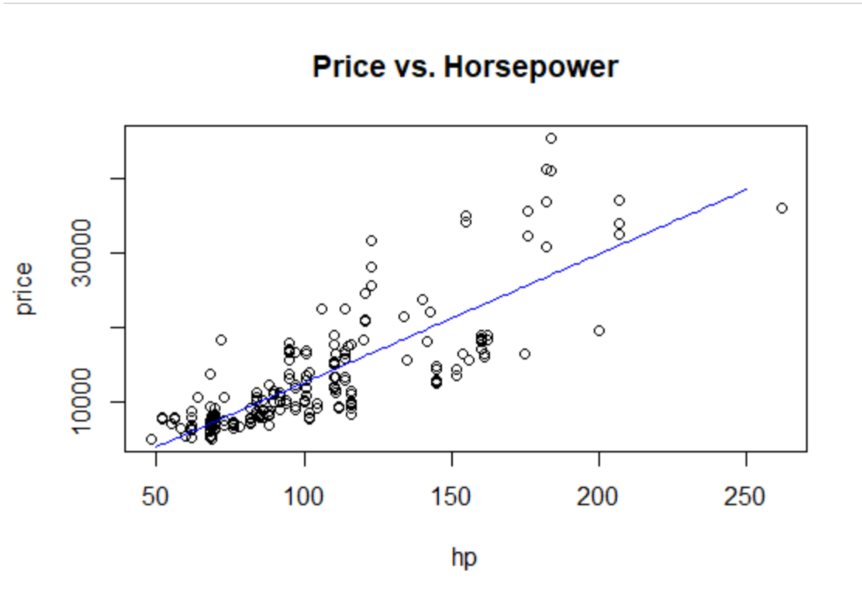
Simple linear regression: Regress price on highway.mpg.




```
Residuals:
   Min     1Q Median     3Q    Max 
 -8647  -3411  -1102   1092  20970 

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 38423.31    1843.39   20.84   <2e-16 ***
X1                NA         NA      NA       NA    
X2           -821.73      58.65  -14.01   <2e-16 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5653 on 199 degrees of freedom
Multiple R-squared:  0.4966,	Adjusted R-squared:  0.4941 
F-statistic: 196.3 on 1 and 199 DF,  p-value: < 2.2e-16
```

Using linear regression to regress price on highway miles per gallon, a statistical significance is shown, due to the low p-value of 2.2e-16. Both coefficients, b0 and b1, are also statistically significant because of their low p-value of 2e-16. The R-squared value of 0.4966 shows that ~50% of the variance can be explained by the regression line. A car's predicted mpg is equal to 38423.31 - 821.73x .

Simple linear regression: Regress price on horsepower

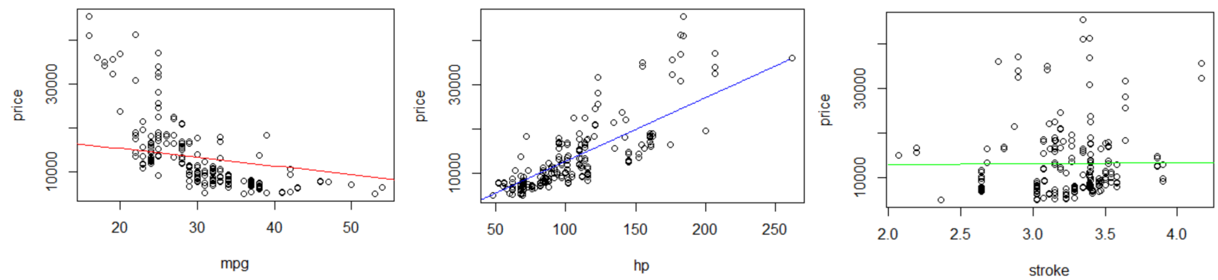



```
Residuals:
     Min       1Q   Median       3Q      Max 
-10140.1  -2295.3   -441.9   1813.7  18315.9 

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4597.56     972.81  -4.726 4.32e-06 ***
X1                NA         NA      NA       NA    
X2            172.18       8.85  19.455  < 2e-16 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4677 on 199 degrees of freedom
Multiple R-squared:  0.6554,	Adjusted R-squared:  0.6537 
F-statistic: 378.5 on 1 and 199 DF,  p-value: < 2.2e-16
```

Using linear regression to regress price on horsepower, a statistical significance is also shown, due to the low p-value of 2.2e-16. Both coefficients, b0 and b1, are also statistically significant because of their low p-value of 4.32 e-06 and 2e-16. The data in this model with horsepower is a better fit than the previous model on mpg. The R-squared value of 0.6554 shows that ~66% of the variance can be explained by the regression line. A car's predicted horsepower is equal to -4597.56 + 172.18x .

Multiple regression: Regress price on highway.mpg, horsepower and stroke.

Price vs. MPG, HP, and Stroke




```
Residuals:
    Min      1Q  Median      3Q     Max 
-8924.2 -2495.5  -439.5  1768.1 17661.6 

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3693.38    5025.98   0.735   0.4633    
X1                NA         NA      NA       NA    
X2           -198.12      84.65  -2.341   0.0203 *  
X3            143.25      15.33   9.345   <2e-16 ***
X4            255.64    1054.28   0.242   0.8087    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4675 on 193 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.6657,	Adjusted R-squared:  0.6605 
F-statistic: 128.1 on 3 and 193 DF,  p-value: < 2.2e-16
```

Using multiple linear regression to regress price on mpg, hp and stroke, the model is still shown to be statistically significant, owing to its p-value of 2.2e-16. The R2 of 0.6657 shows that ~66% of the variance can be explained by the regression line (slightly better than the previous model).

The p-value of b0 is not significant at 0.4633 >> 0.05. The coefficient of b3 (stroke) is also not significant, due to its p-value of 0.8087 >> 0.05, as well as the flatness of the slope in the graph above.

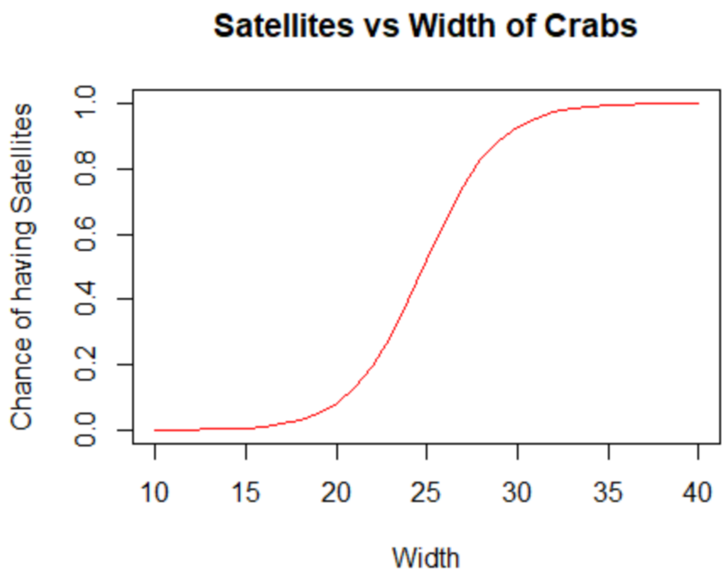
The p-value of b1 (mpg), while still significant at 0.02 << 0.05, is closer to the alpha value. (Is there another variable affecting mpg?)

The B2 coefficient(HP) has the lowest p-value of the coefficients at 2e-16, which shows a strong significance.

Analysis

The model for Price vs. Horsepower is, in my opinion, the best statistical model. First, the p-value for its coefficients are the smallest out of all the models. Also, while the multiple regression model has a slightly higher $R2$, it also has coefficients that are not significant ($B0$ and stroke). Finally, the $R2$ for Price vs. mpg is low, showing a lesser goodness-of-fit.

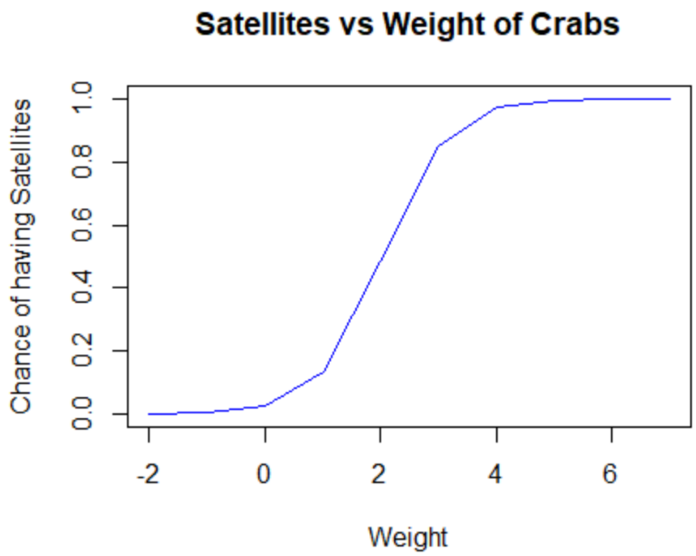
Project 2: Crabs




```
Call:
glm(formula = y ~ width, family = binomial())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0281  -1.0458   0.5480   0.9066   1.6942

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.3508     2.6287  -4.698 2.62e-06 ***
width         0.4972     0.1017   4.887 1.02e-06 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be
 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 194.45  on 171  degrees of freedom
AIC: 198.45

Number of Fisher Scoring iterations: 4
```

A logistic model was used to analyze the impact of width on the number of satellites on female crabs. The p-value of the coefficients are significant at 2.62e-06 and 1.02e-06, which shows that width does influence the number of satellites. The positive value of the width coefficient means that an increase in width increases the chance of a crab having satellites. The residual deviance is less than the null deviance at 194.45 < 225.76, and the p-value of the model is small at 2.199438e-08, which means that the model fits the data well and the overall model is statistically significant.

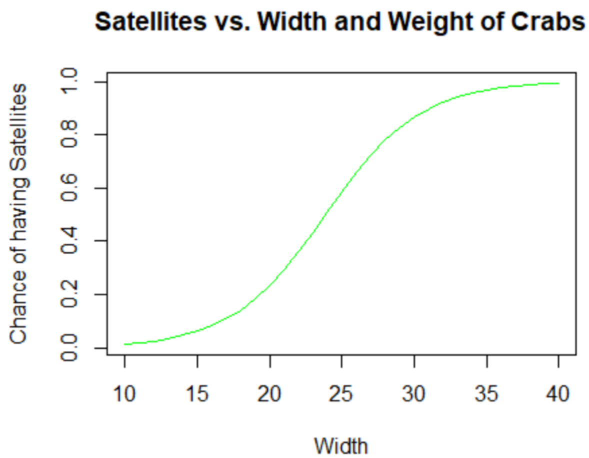
Perform a logistic regression to investigate the impact of weight.

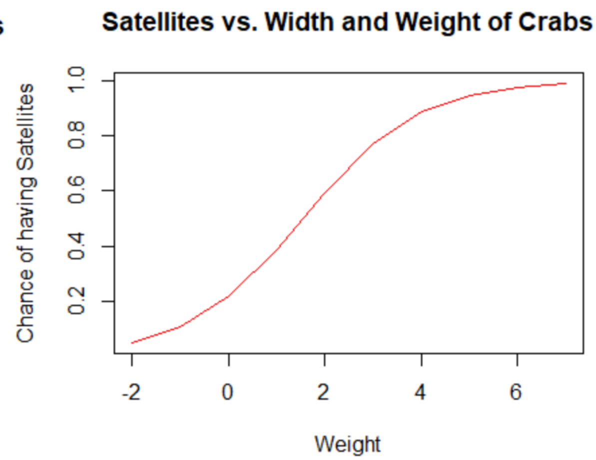



```
Call:
glm(formula = y ~ weight, family = binomial())

Deviance Residuals:
    Min       1Q    Median       3Q      Max
-2.1108  -1.0749   0.5426   0.9122   1.6285

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.6947     0.8802  -4.198 2.70e-05 ***
weight        1.8151     0.3767   4.819 1.45e-06 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be
 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 195.74  on 171  degrees of freedom
AIC: 199.74

Number of Fisher Scoring iterations: 4
```

Looking at the logistic model for the impact of weight on the number of satellites, the p-values for both coefficients are significant at 2.70e-05 and 1.45e-06, which shows that weight does influence the chance of satellites. The positive value of weight shows that an increase in weight correlates to a higher chance of satellites. The residual deviance is smaller than the null deviance at 195.74 < 225.76, and the p-value of the model is low at 4.276131e-08. This model also fits the data well and the model is statistically significant.

Perform a logistic regression to investigate the impact of width and weight.






```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1127  -1.0344   0.5304   0.9006   1.7207

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.3547     3.5280  -2.652  0.00801 **
width         0.3068     0.1819   1.686  0.09177 .
weight        0.8338     0.6716   1.241  0.21445
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 192.89  on 170  degrees of freedom
AIC: 198.89

Number of Fisher Scoring iterations: 4
```

Applying logistic regression to 2 variables (width and weight), the positive value of both the width and weight coefficients show an association between an increase in width and weight, and an increase in the chance of satellites. However, both variables are shown to have a high p-value (0.09177 > 0.05 for width, 0.21445 > 0.05 for weight). This shows that the 2 coefficients are not significant. The residual deviance 192.89 is lower than the null deviance of 225.76. The p-value for the model is low at 7.284004e-08, which shows that the overall model is significant.

Analysis

In terms of the best overall model, the model investigating the impact of both width and weight is the only model which has 2 coefficients with a p-value above 0.05. Because of the insignificance of the coefficients, this model is not the best statistical model.

While the model investigating width, and the model investigating weight, are both statistically significant, I would probably choose the model investigating width as the best model. First, the p-value of the coefficients for the width model are slightly smaller, which shows a slightly better statistical significance. Second, it has the smallest p-value out of all the models at 2.199438e-08, which means it is the most statistically significant model. Finally, the AIC for the width model is the smallest at 198.45, which is another indication that it is the best model fit.

Script

```
carPrice = read.csv('C:\\datasets\\CarPrice.csv')
stroke = carPrice$stroke
hp = carPrice$horsepower
mpg = carPrice$highway.mpg
price = carPrice$price

#Project 1
#1. Regress price on highway.mpg
plot(mpg, price, main="Price vs. MPG")
Y = cbind(price)
X = cbind(c(1),c(mpg))
LM = lm(Y~X); summary(LM)
b0 = LM$coefficients[1]; b0
b1 = LM$coefficients[3]; b1
xseq = seq(0,60)
points(xseq, b0+b1*xseq, type='l', col='red')

#2. Regress price on horsepower
Y = cbind(price)
X = cbind(c(1),c(hp))
LM = lm(Y~X); summary(LM)
plot(hp, price, main="Price vs. Horsepower")
b0 = LM$coefficients[1]; b0
b1 = LM$coefficients[3]; b1
xseq = seq(50,250)
points(xseq, b0+b1*xseq, type='l', col='blue')

#3. Regress price on highway.mpg, horsepower and stroke.
Y = cbind(price)
X = cbind(c(1),c(mpg),c(hp),c(stroke))
LM = lm(Y~X); summary(LM)
b0 = LM$coefficients[1]; b0
b1 = LM$coefficients[3]; b1
b2 = LM$coefficients[4]; b2
b3 = LM$coefficients[5]; b3

plot(mpg, price)
xseq = seq(1,60)
points(xseq,b0+b1*xseq+b2*mean(hp)+b3*mean(na.omit(stroke)),type="l",col="red")
```

```
plot(hp, price)
xseq = seq(40,260)
points(xseq,b0+b1*mean(mpg)+b2*xseq+b3*mean(na.omit(stroke)),type="l",col="blue")

plot(stroke, price)
xseq = seq(2,5)
points(xseq,b0+b1*mean(mpg)+b2*mean(hp)+b3*xseq,type="l",col="green")


#Project 2
crabs = read.table("http://www.stat.ufl.edu/???aa/cat/data/Crabs.dat", header = TRUE)
y = crabs$y
width = crabs$width
weight = crabs$weight

#1. Perform a logistic regression to investigate the impact of width
GLM = glm(y ~ width, family=binomial()); summary(GLM)
b0 = GLM$coefficients[1]; b1 = GLM$coefficients[2]
width_range = seq(10, 40)
plot(
 width_range, 1/(1+exp(-(b0+b1*width_range)))
 , type='l'
 , col='red'
 , main='Satellites vs Width of Crabs'
 , xlab='Width'
 , ylab='Chance of having Satellites')
p_value = 1-pchisq(225.76-194.45, df=1); p_value

#2. Perform a logistic regression to investigate the impact of weight.
GLM = glm(y ~ weight, family=binomial()); summary(GLM)
b0 = GLM$coefficients[1]; b1 = GLM$coefficients[2]
weight_range = seq(-2, 7)
plot(
 weight_range
 , 1/(1+exp(-(b0+b1*weight_range)))
 , type='l'
 , col='blue'
 , main='Satellites vs Weight of Crabs'
 , xlab='Weight'
 , ylab='Chance of having Satellites'
 )
```

```
p_value = 1-pchisq(225.76-195.74, df=1); p_value

#3. Perform a logistic regression to investigate the impact of width and weight.
GLM = glm(y ~ width+weight, family=binomial()); summary(GLM)
b0 = GLM$coefficients[1]; b1 = GLM$coefficients[2]; b2 = GLM$coefficients[3]
width_seq = seq(10, 40)
weight_seq = seq(-2, 7)
plot(
 width_seq
 , 1/(1+exp(-(b0+b1*width_seq+b2*mean(weight))))
 , type="l"
 , col="green"
 , main='Satellites vs. Width and Weight of Crabs'
 , xlab='Width'
 , ylab='Chance of having Satellites'
 )

plot(weight, y, main='Satellites vs. Width and Weight of Crabs')
plot(
 weight_seq
 , 1/(1+exp(-(b0+b1*mean(width)+b2*weight_seq)))
 , type="l"
 , col="red"
 , main='Satellites vs. Width and Weight of Crabs'
 , xlab='Weight'
 , ylab='Chance of having Satellites'
 )
p_value = 1-pchisq(225.76-192.89, df=2); p_value
```