M349R (Unique 54230)

**Instructor:** Gustavo Cepparo

Project 2

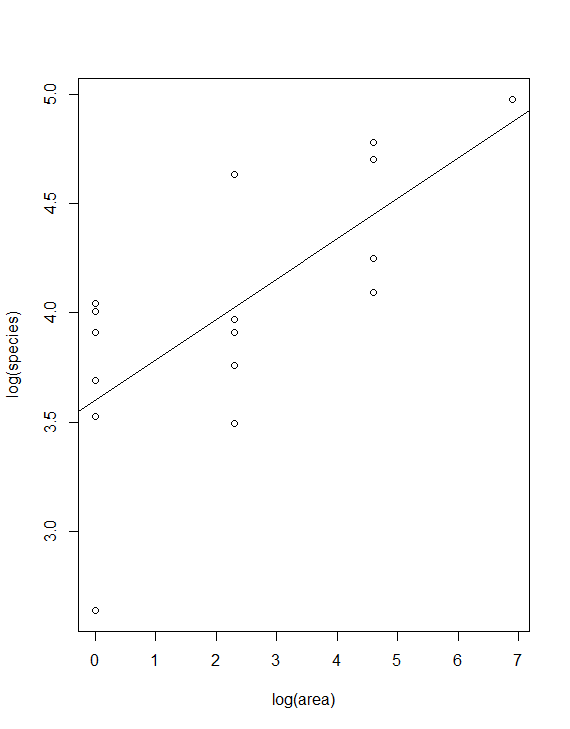
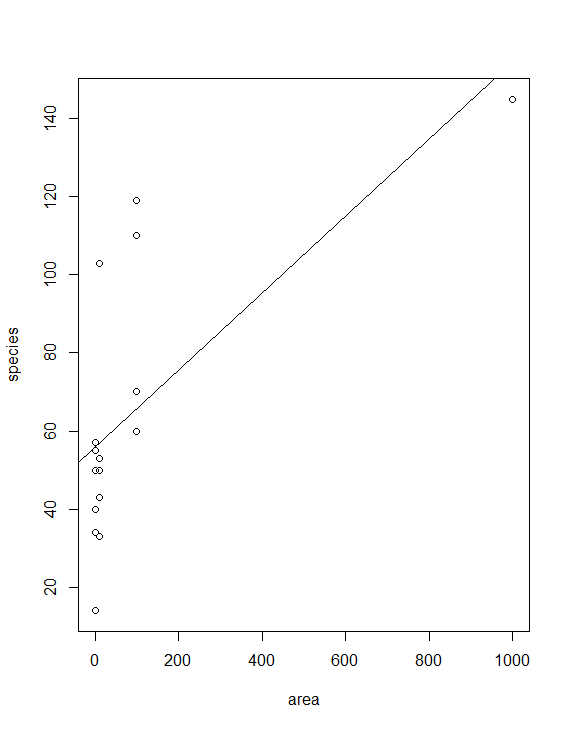
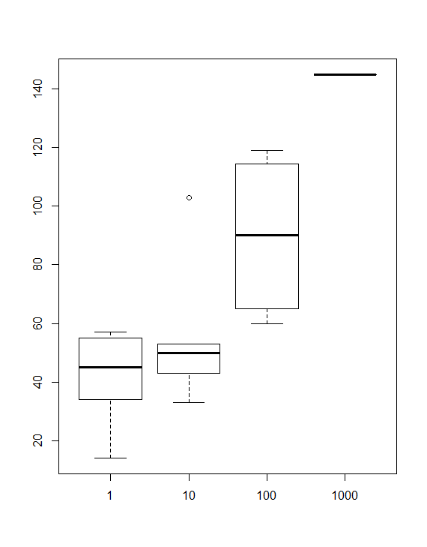
Juan Acosta

JA45384

Fall 2018

**Problem 1**

Looking a the data, there appears to be some form of relationship.

Yet a scatterplot of the original data does not show a strong linear relationship. Applying logarithm to both Species and Area, we get a more linear relationship.

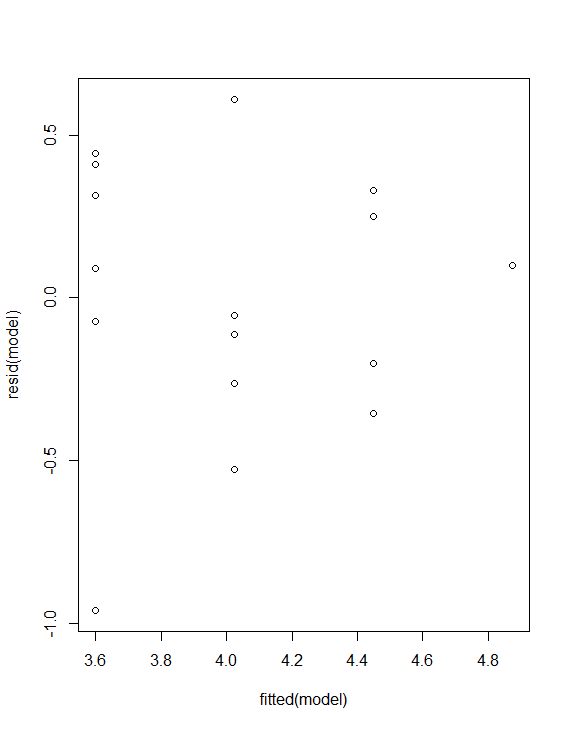
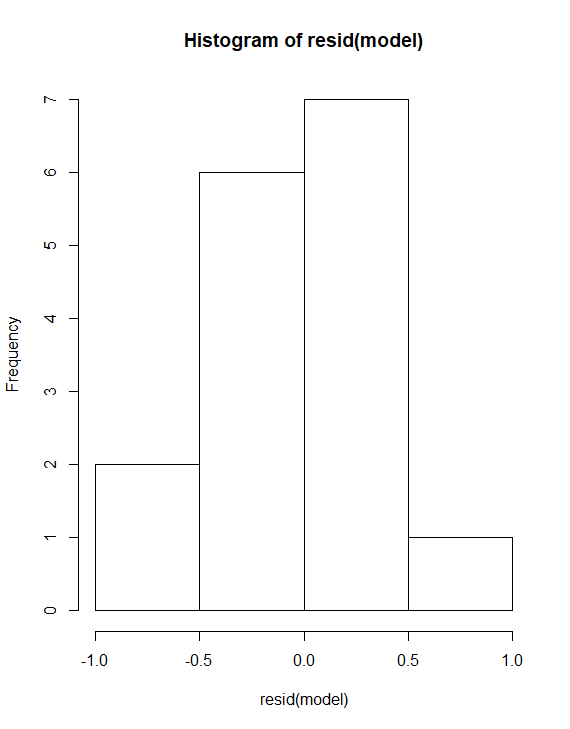
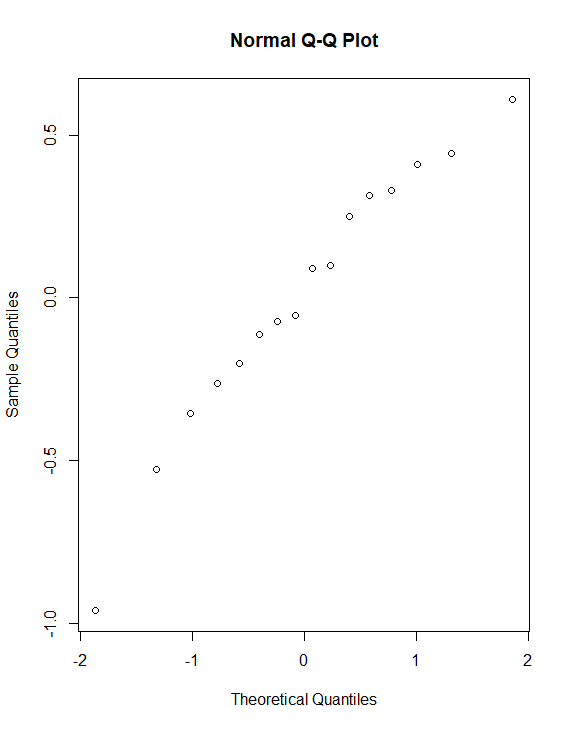
The regression line from the right is derived from the linear model

log(Species) = 3.5987 + 0.1848log(Area)

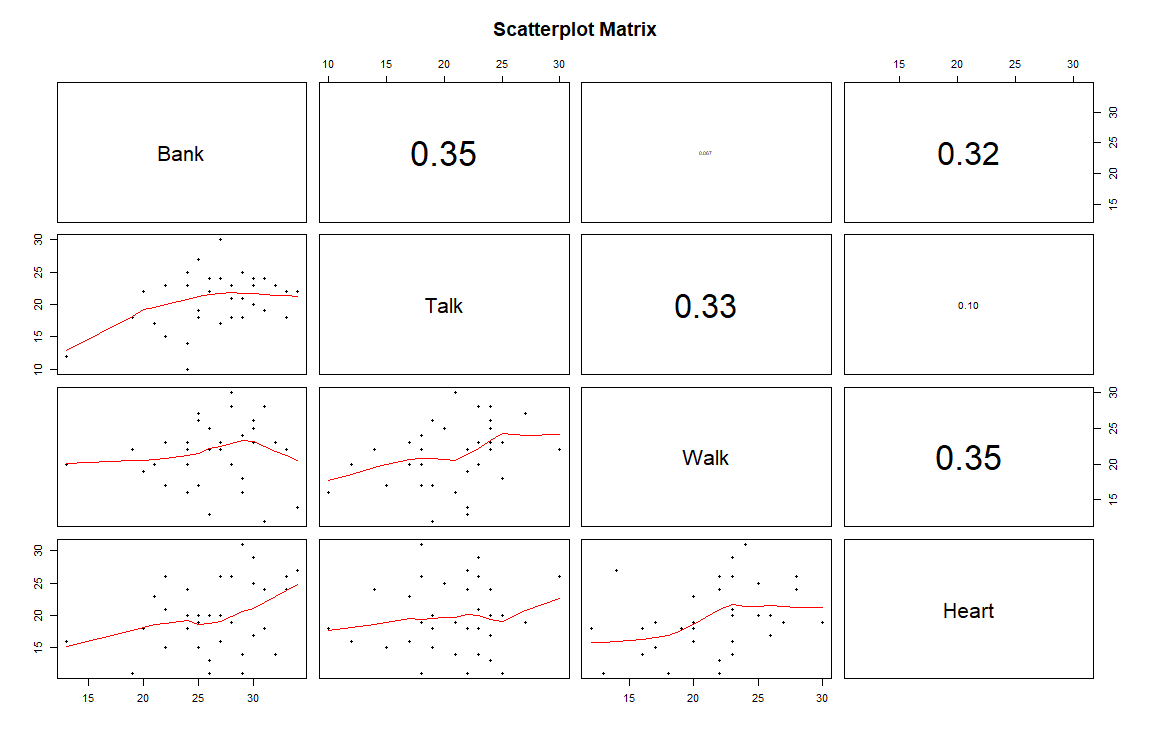
(0.1540) (0.0488)

R2 = 0.5055 RMSE = 0.421 n = 16

The Lack of fit F test produces an F statistic of 0.1054 and a p-value of 0.9008. Therefore we do not reject the model. And we can conclude that there is an effect of Area on number of Species. Yet the logarithmic transformation lets us know that there is a decrease in the ratio at which number of species increases compared to area.

**Problem 2**

a)

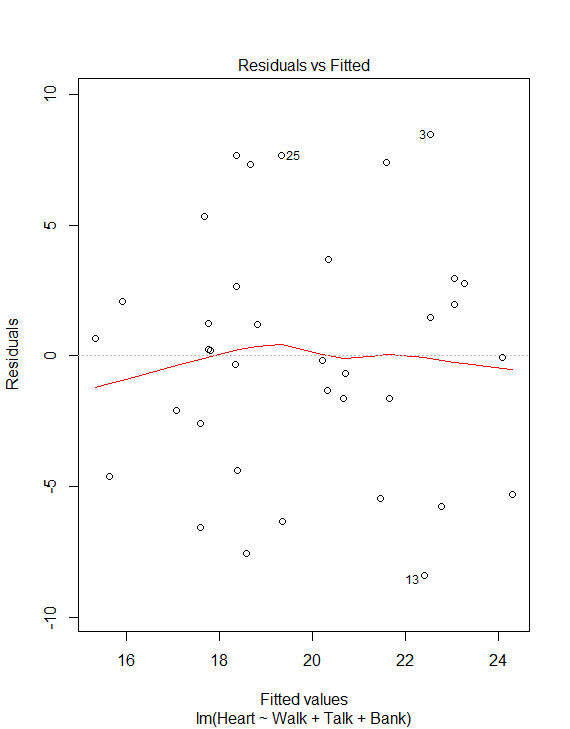
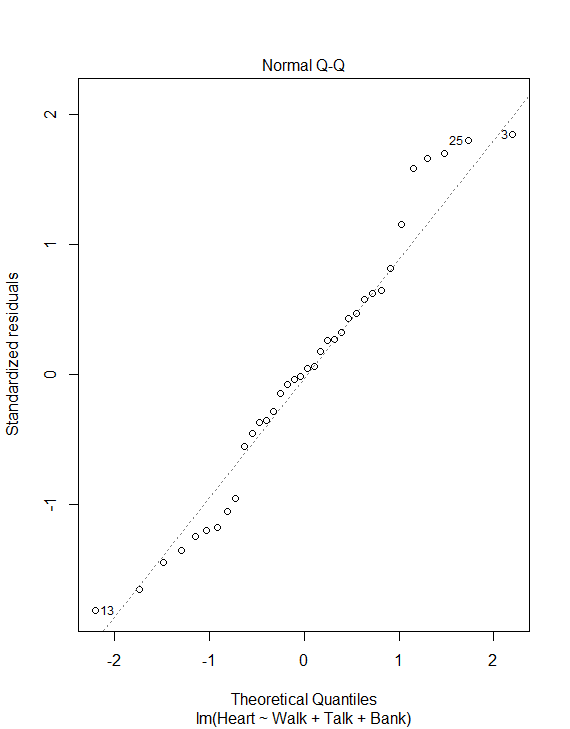
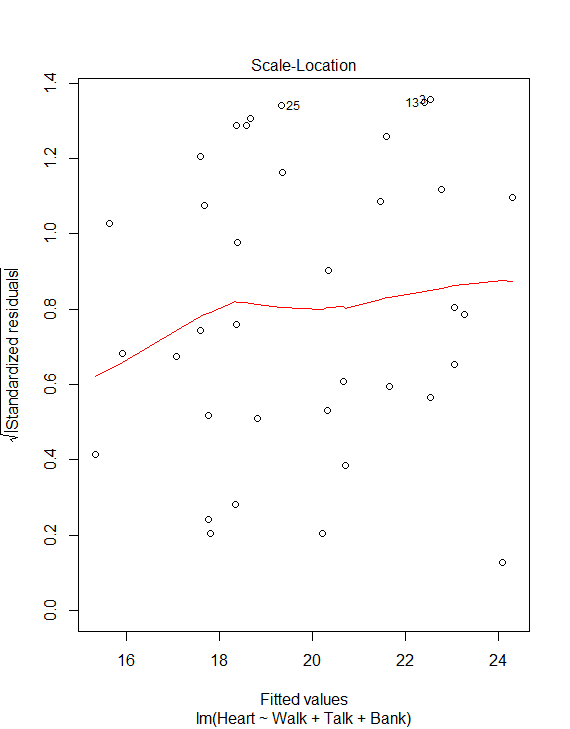
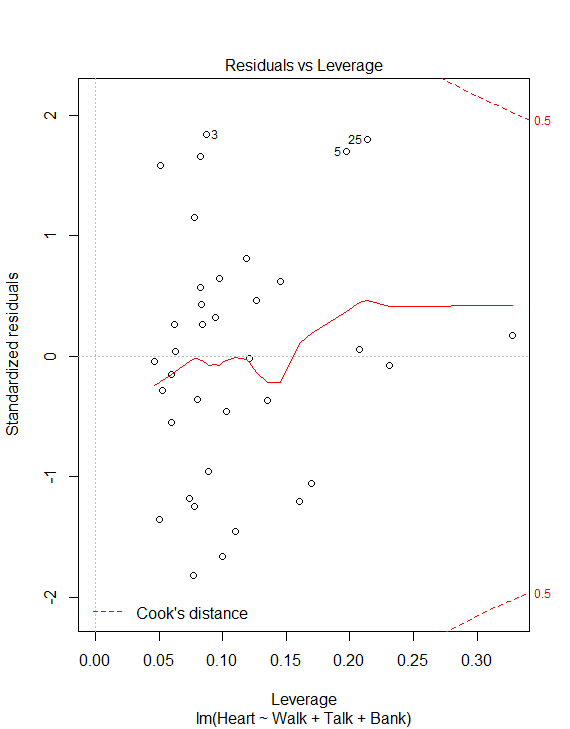
b) The Linear model is

Heart = 3.1787 + 0.4516 Walk – 0.1796 Talk + 0.4052 Bank

(6.3369) (0.2009) (0.2222) (0.1971)

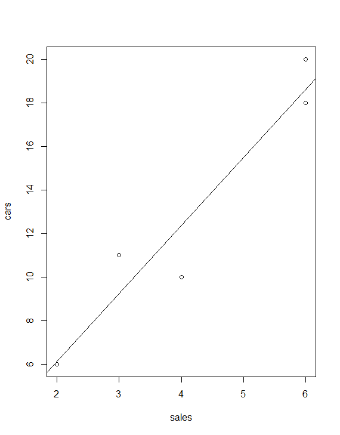
R2 = 0.2236 RMSE = 4.805 n = 36

c) The Lack of fit F test produces an F statistic of 1.8744 and a p-value of 0.5294. Therefore we do not reject the model. Yet we have a low R2 and the residual plots show some signs of clustering.

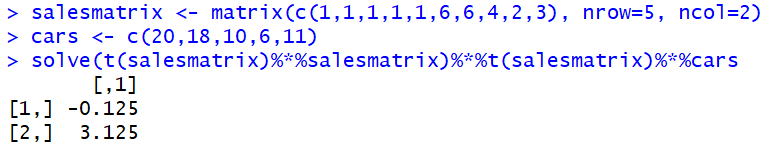
   

d) The model defined above in b presents estimates for B0, B1, B2, and B3. Yet from the standard errors we can see that B0 and B2 are not statistically significant, while B1 and B3 are. Therefore, we can trust the values for B1 and B3 to state the estimated on the average effect of Walk and Bank on Heart, yet we shouldn’t make them for Talk. Moreover, since the intercept is not statistically significant we should consider forcing the intercept to 0, and tweak the model further.

**Problem 3**

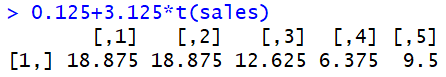
a,c)

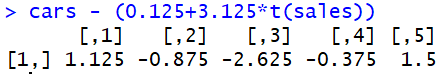
The few amount of data points makes any assessment of the form of this data highly subjective. Having said that, we could argue that the data is slightly linear.

b)

d) The average number of cars sold when five salespeople are kept on the showroom each day for a week is 15.5, with a 95% confidence interval of (8.68679, 22.31321)

e) A PI to predict the number of cars sold when five salesperson are kept on the showroom for the week would be (12.45304, 18.54696)

f)



g)

h) A 95% confidence interval for the slope is (1.4216, 4.8283). This means that for an alpha of 0.05, we reject the null that the intercept is equal to cero.

i) I would conclude that the number of sales people within the sample range could have a positive correlation with cars sold. Despite this model being statistically insignificant due to small sample size, it is clear that it has practical significance.

j) I wouldn’t rely on this model due to its lack of statistical significance, yet I would consider it due to its practical significance.

**Problem 4**

a) The explanatory variable are the number of manatees seen by the helicopter and the response variable are those seen by the airplane.

b) The linear model is

Airplane = -7.4632 + 1.1483 Helicopter

(12.4957) (0.2311)

R2 = 0.7553 RMSE = 13.13 n = 10

The slope implies that we would expect to see from the airplane, on the average 1.1483 as many manatees as on the helicopter

The intercept is outside the interval of observations and provides no useful information. Additionally, it is statistically insignificant.

The R2 implied that the value of Helicopter describes 75.58% of the variability of Airplane

c) For H0: slope = 0 we have an F-statistic of 24.692 and a P-value of 0.001094. therefore we can conclude that at a significance level of 5% the slope is significantly different from 0.

For H0: slope = 1 we have an F-statistic of 0.4118 and a P-value of 0.539 therefore at a 5% significance level we cannot conclude that the slope is significantly different from 1.

A slope of 1 makes more sense in this scenario since one could expect that Airplane and Helicopter would be co-linear

d) The linear model is

Airplane = 1.01813 Helicopter

(0.07401)

R2 = 0.9546 RMSE = 12.66 n = 10

Conducting a linear hypothesis on the slope being 1, we obtain an F-statistic of 0.06 and a P-value of 0.812 therefore at a 5% significance level we cannot conclude that the slope is significantly different from 1.

**Problem 5**

a) The simple regression model of colGPA on ACT

colGPA = 2.40298 + 0.02706 ACT

(0.26420) (0.01086)

R2 = 0.04275 RMSE = 0.3656 n = 141

The variables seem to be positively correlated. The intercept does not provide useful information as it is outside of the range of observations. If ACT increases by 5 points, colGPA is expected to increase on the average 0.1353.

b) colGPA = 2.9983 for ACT=22

CI = (2.921904, 3.07487)

PI = (2.271547, 3.725226)

The 95% confidence interval tells us that we are 95% certain that the true GPA mean for students with an ACT score of 22 is inside the CI

The 95% PI tells us that a given student with an ACT = 22, we are 95% certain that their GPA will be within the PI

c) As determined by our model, our coefficient of determination (R2) is equal to 0.04275. Which means that 4.275% of the variation of colGPA is explained by ACT scores.

d) The 95% confidence interval would be

CI = (0.3272095, 0.4142771)

Then since 1.28\*0.4142771 = 0.53 > 0.5 Therefore, since the threshold was exceeded, we can conclude that ACT is not an adequate predictor for colGPA.

e.i) The regression model

colGPA = 1.41543 + 0.48243 hsGPA

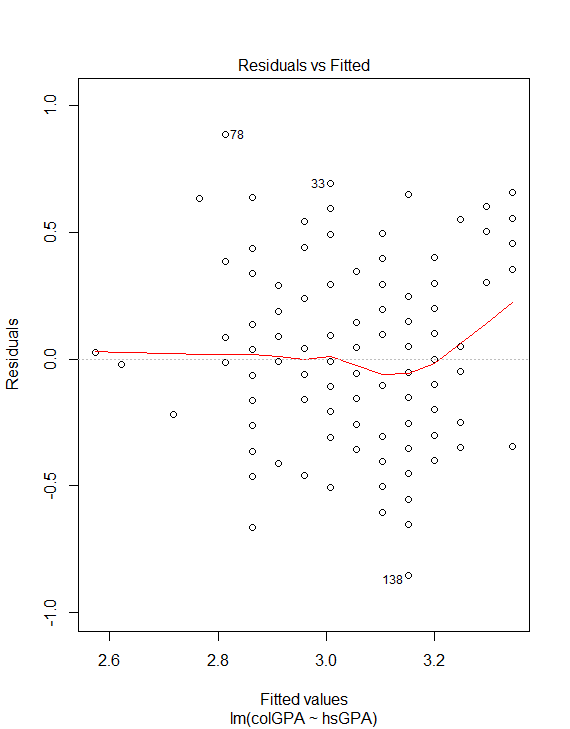
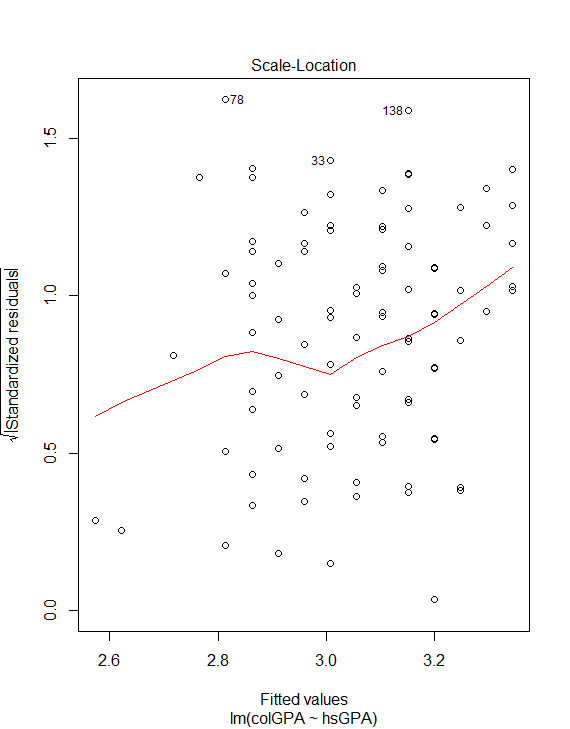
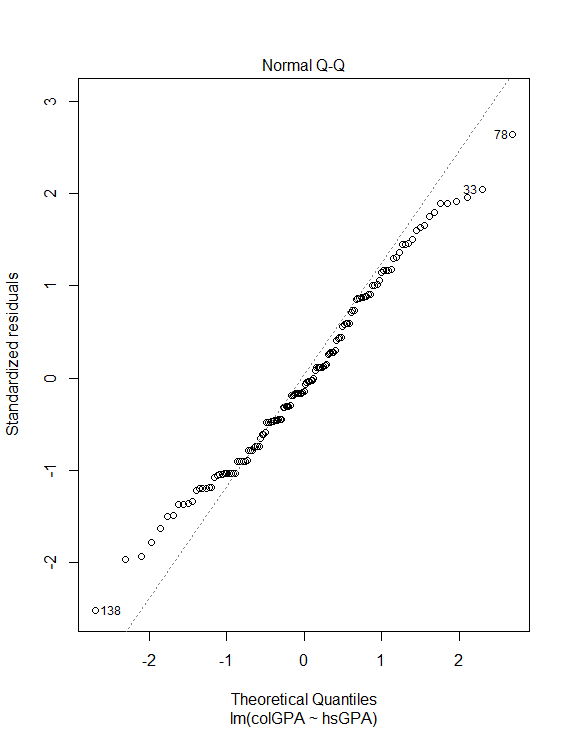
(0.30694) (0.08983)

R2 = 0.1719 RMSE = 0.34 n = 141

We can understand the coefficient of hsGPA to mean that on average when hsGPA increases by 1 point, colGPA will be expected to increase approximately 0.48243 points.

e.ii) Reading the results from the regression output we can understand the T-statistic of 5.371 and a P-value of 3.21e-07 as the H0: slope = 0. Therefore, we can conclude that the coefficient is significantly different from 0.

e.iii)

The residual plots look slightly right skewed yet random enough to not be an indicator of any issues.

e.iv) The 95% confidence interval would be

CI = (0.3042977, 0.3852686)

Then since 1.28\*0.3852686 = 0.49 < 0.5 Therefore, since the threshold was not exceeded, we can conclude that hsGPA is an adequate predictor for colGPA.

e.v) hsGPA seems like a better predictor since it has a higher R2. Moreover, the coefficient for ACT in our original regression was barely significant, while the coefficient of hsGPA is considerably larger than its calculated error.

**Code**

**Problem 1**

area <- data1$Area

species <- data1$Species

areagroup <- data1$Areagroup

boxplot(species ~ area)

model <- lm(species ~ area)

plot(area,species)

abline(model)

summary(model)

model <- lm(log(species) ~ log(area))

plot(log(area), log(species))

abline(model)

summary(model)

fit <- lm(log(species) ~ areagroup)

summary(fit)

anova(model)

anova(model, fit)

plot(resid(model)~log(area))

plot(resid(model)~fitted(model))

hist(resid(model))

qqnorm(resid(model))

**Problem 2**

library(gtable)

library(grid)

library(gridExtra)

data = data2

p1 <- qplot(seq\_along(data$Heart), data$Heart)

p2 <- qplot(seq\_along(data$Walk), data$Walk)

p3 <- qplot(seq\_along(data$Talk), data$Talk)

p4 <- qplot(seq\_along(data$Bank), data$Bank)

grid.arrange(p1, p2, p3, p4, nrow=2)

panel.cor <- function(x, y, digits=2, prefix="", cex.cor, ...)

{

usr <- par("usr"); on.exit(par(usr))

par(usr = c(0, 1, 0, 1))

r <- abs(cor(x, y))

txt <- format(c(r, 0.123456789), digits=digits)[1]

txt <- paste(prefix, txt, sep="")

if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)

text(0.5, 0.5, txt, cex = cex.cor \* r)

}

# Plot #2: same as above, but add loess smoother in lower and correlation in upper

pairs(~Bank+Talk+Walk+Heart, data=data2,

lower.panel=panel.smooth, upper.panel=panel.cor,

pch=20, main="Scatterplot Matrix")

model <- lm(Heart ~ Walk+Talk+Bank, data=data2)

summary(model)

plot(model)

pureErrorAnova(model)

**Problem 3**

sales <- c(6,6,4,2,3)

salesmatrix <- matrix(c(1,1,1,1,1,6,6,4,2,3), nrow=5, ncol=2)

cars <- c(20,18,10,6,11)

plot(sales, cars)

model <- lm(cars ~ sales)

summary(model)

abline(model)

w <- solve(t(salesmatrix)%\*%salesmatrix)%\*%t(salesmatrix)%\*%cars

w

newdata = data.frame(sales=5)

predict(model, newdata, interval="predict")

predict(model, newdata, interval="confidence")

confint(model)

0.125+3.125\*t(sales)

cars - (0.125+3.125\*t(sales))

mean((cars - (0.125+3.125\*t(sales)))^2)

**Problem 4**

library(car)

air <- data4$Airplane

heli <- data4$Helicopter

plot(air, heli)

model <- lm(air ~ heli)

abline(model)

summary(model)

linearHypothesis(model, "heli = 0")

linearHypothesis(model, "heli = 1")

model <- lm(air ~ -1+heli)

summary(model)

linearHypothesis(model, "heli = 1")

**Problem 5**

model <- lm(colGPA ~ ACT, data=data5)

summary(model)

new <- data.frame(ACT=22)

predict(model, new, se.fit=TRUE,interval="confidence")

predict(model, new, se.fit=TRUE,interval="prediction")

sqrt(139\*(.3656)^2/qchisq(.025, 139))

sqrt(139\*(.3656)^2/qchisq(.975, 139))

model2<-lm(colGPA ~ hsGPA, data=data5)

summary(model2)

min(na.omit(data5$hsGPA))

1.28\*sqrt(139\*(0.34)^2/qchisq(.025, 139))

sqrt(139\*(0.34)^2/qchisq(.975, 139))