# Task 1: Hypothesis

We want to know whether morning routine affects the amount that an individual exercises. Average daily commute, wake up time, in terms of hours after midnight, and whether an individual has breakfast. We expect individuals to exercise more if they wake up earlier, have breakfast, and have a shorter average commute time.

# Task 2: Model

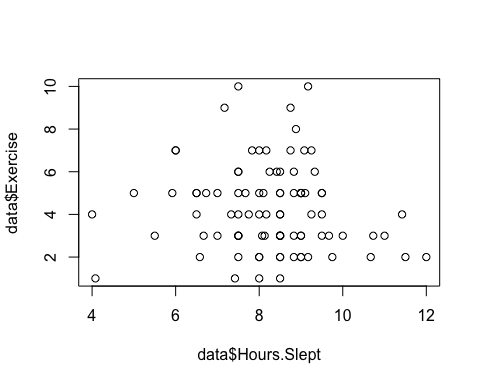
1. Adjust columns to make them the appropriate units and classes.

# adjust columns  
data$avg.commute<- (data$Commute.today + data$Commute.yesterday)/ 2 #average commute

data <- data[!is.na(data$avg.commute),] #remove nas from average commute

1. Explore variables + Check for correlations or interactions in the predictors:

* Plot exercise against the continuous predictors Hours.Slept and avg.commute

A graph of numbers and dots

Description automatically generated with medium confidence

#Check if continuous variables are correlated  
cor.test(data$Hours.Slept,data$avg.commute)   
# p = 0.5643, cor = -0.06157292. Not correlated.

#Is there a potential relationship between the categorical and continuous predictors?  
print(aggregate(data$Hours.Slept, by=list(data$Breakfast), mean, na.rm =T))

## Group.1 x  
## 1 No 7.903333  
## 2 Yes 8.359130

#Seems like people who eat breakfast sleep longer.

#Distribution of response variable

hist(data$Exercise)



# The response is not normal, but it is count data so should be modeled in a poisson distribution

1. Create the MAM

* The first, most complex model was created and reviewed with an analysis of deviance:

msi1<- glm(Exercise~ Hours.Slept + Hours.Slept\*Breakfast+ I(Hours.Slept^2) + avg.commute, data=data, family= poisson)

anova(msi1, test="Chisq")# p value of average commute = 0.68261 and is the highest p value so we can remove average commute as a predictor.

* Average commute was removed, so now the two final models will confirm whether an interaction term is needed:

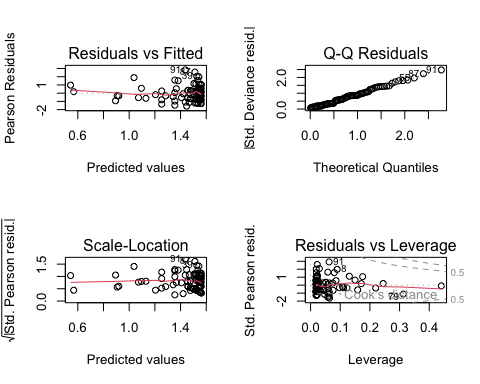
m.s<- glm(Exercise~Hours.Slept + Breakfast+ I(Hours.Slept^2), data=data, family= poisson)  
# no interaction term.

m.s.i<- glm(Exercise~Hours.Slept+ Hours.Slept\*Breakfast + I(Hours.Slept^2), data=data, family= poisson)  
# interaction term.

* summary() results from two remaining models, and results from analysis of deviance:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Models | AIC | Residual Deviance | Degrees of Freedom | Deviance | Pr(>Chi) |
| Hours.Slept + Breakfast + I(Hours.Slept^2) | 374.53 | 76.877 | 86 |  |  |
| Hours.Slept + Hours.Slept\*Breakfast + I(Hours.Slept^2) | 372.53 | 72.879 | 85 | 3.9976 | 0.04557 |

* Checking diagnostic plots



The diagnostic plots only look ok for the QQ Normal plot. The variances do not look evenly distributed, however, the diagnostic plots for glm are harder to interpret. Therefore, we will still use the more complex model as it is the best fit for the data. So this is the final model:

glm(Exercise~Hours.Slept\*Breakfast + I(Hours.Slept^2), data = data, family = poisson)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients | Estimate | Standard Error | Z Value | P(>|z|) |
| Intercept | 0.14819 | 1.07038 | 0.138 | 0.8899 |
| Hours.Slept | 0.48511 | 0.27184 | 1.785 | 0.0743 |
| BreakfastYes | -1.31249 | 0.77351 | -1.697 | 0.0897 |
| I(Hours.Slept^2) | -0.04178 | 0.01812 | -2.306 | 0.0211\* |
| Hours.Slept:BrekafastYes | 0.18981 | 0.09693 | 1.958 | 0.0502 |

# Task 3: Results

There was a significant quadratic relationship between the the number of times and individual exercised and hours slept (b±SE = -0.04178 ±0.01812; Z-Value 1,85 = -2.3066; p =0.0211). Note that the parameter estimate is on the log scale. While the interaction term of Breakfast was not significant in the model prediction, it did significantly improve the model fit when tested in an analysis of deviance (χ2 1 = 3.9976 , p = 0.04557).

# Task 4: Plot

First create a new data fram for the predicted values to ensure a smooth line in the final plot.

newdata.Y <- data.frame(Breakfast=rep("Yes", 100),  
 Hours.Slept=seq(min(data$Hours.Slept[data$Breakfast=="Yes"]),  
 max(data$Hours.Slept[data$Breakfast=="Yes"]),  
 length.out=100))  
  
newdata.N <- data.frame(Breakfast=rep("No", 100),  
 Hours.Slept=seq(min(data$Hours.Slept[data$Breakfast=="No"]),  
 max(data$Hours.Slept[data$Breakfast=="No"]),length.out=100))  
  
predicted.Y2 <- predict(m.s.i, newdata.Y, type='response')  
predicted.N2 <- predict(m.s.i, newdata.N, type='response')

Now plot using the new data:

plot(Exercise ~ Hours.Slept, data=data, pch=NA, xlab="Hours of Sleep", ylab="Exercise")  
points(Exercise ~ Hours.Slept, data=data[data$Breakfast=="Yes", ], pch=19, col="blue")   
points(Exercise ~ Hours.Slept, data=data[data$Breakfast=="No", ], pch=19, col="red")  
lines(predicted.Y2[order(newdata.Y$Hours.Slept)] ~  
 sort(newdata.Y$Hours.Slept), lwd=1.5, col="blue")  
lines(predicted.N2[order(newdata.N$Hours.Slept)] ~  
 sort(newdata.N$Hours.Slept), lwd=1.5, col="red")  
legend(x="topleft", legend=c("Yes", "No"), pch=19,  
col=c("blue", "red"), lwd=c(1,1), title="Eats Breakfast", cex=0.8)

