Problem Sheet 1

2024-10-08

# Task 1 : Hypothesis

## Import data

setwd("/Users/carlaleone/Desktop/Exeter/Problem Sheet")  
data <- read.table("BIOM4025\_data\_2024.csv", header=TRUE, sep = ",",   
 stringsAsFactors = FALSE)

## 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 are not are the factors describing morning routine. 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.

# add other columns  
data$avg.commute<- (data$Commute.today + data$Commute.yesterday)/ 2 #average commute of individual  
class(data$Waking.up.time) #character class, so need to convert it to numeric

## [1] "character"

data$wakeup<- as.POSIXct(paste("2024-01-01", data$Waking.up.time), format = "%Y-%m-%d %H:%M") #convert character to date and time format  
data$hours.midnight<- as.numeric(format(data$wakeup, "%H")) + as.numeric(format(data$wakeup, "%M")) / 60 #convert date and time to numeric  
# are there any missing values?  
is.na(data$avg.commute)

## [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE  
## [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE  
## [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

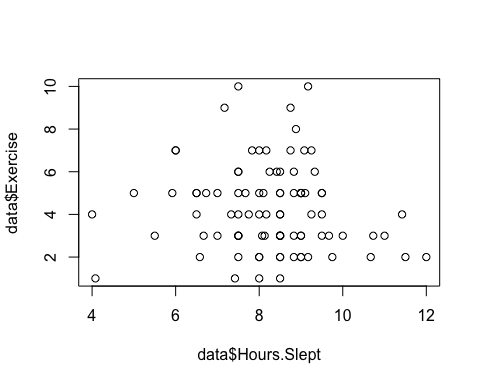
data <- data[!is.na(data$avg.commute),] #remove na from average commute column

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

# Plot Exercise against hours slept  
mean(data$Hours.Slept)

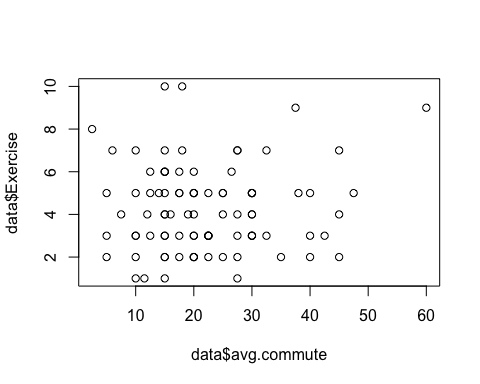
## [1] 8.252778

# mean hours slept = 8.252778  
plot(data$Exercise~data$Hours.Slept)



#The data seems to be centered around the mean, so I would expect the model to have a curved line and peak at the mean.Therefore, I will consider adding a square term.

# Plot Exercise against average commute  
plot(data$Exercise~data$avg.commute)



# Again,no obvious relationships from the data, but could be leaning towards a positive slope.

#Check if continuous variables are correlated  
cor.test(data$Hours.Slept,data$avg.commute)

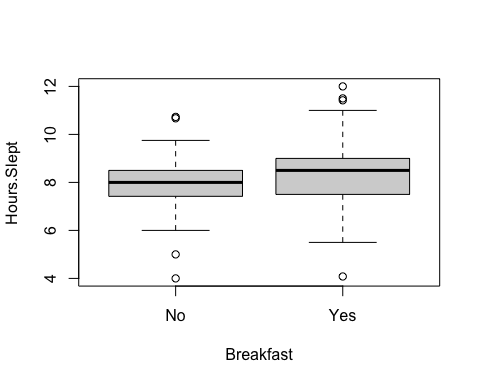
##   
## Pearson's product-moment correlation  
##   
## data: data$Hours.Slept and data$avg.commute  
## t = -0.5787, df = 88, p-value = 0.5643  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.2652814 0.1473977  
## sample estimates:  
## cor   
## -0.06157292

#not correlated  
#p = 0.5643, cor = -0.06157292

#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

# Visualize the table with a box plot  
boxplot(Hours.Slept~Breakfast, data=data)



#Seems like people who eat breakfast sleep longer.

hist(data$Exercise)



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

1. Create the model

* Starting with the most complex model

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

##   
## Call:  
## glm(formula = Exercise ~ Hours.Slept + Hours.Slept \* Breakfast +   
## I(Hours.Slept^2) + avg.commute, family = poisson, data = data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.069518 1.088053 0.064 0.9491   
## Hours.Slept 0.493655 0.273179 1.807 0.0708 .  
## BreakfastYes -1.332075 0.774392 -1.720 0.0854 .  
## I(Hours.Slept^2) -0.042225 0.018163 -2.325 0.0201 \*  
## avg.commute 0.002164 0.004701 0.460 0.6452   
## Hours.Slept:BreakfastYes 0.190818 0.096922 1.969 0.0490 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 84.306 on 89 degrees of freedom  
## Residual deviance: 72.669 on 84 degrees of freedom  
## AIC: 374.32  
##   
## Number of Fisher Scoring iterations: 5

anova(msi1, test="Chisq")

## Analysis of Deviance Table  
##   
## Model: poisson, link: log  
##   
## Response: Exercise  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 89 84.306   
## Hours.Slept 1 0.8624 88 83.444 0.35307   
## Breakfast 1 2.8629 87 80.581 0.09065 .  
## I(Hours.Slept^2) 1 3.7039 86 76.877 0.05429 .  
## avg.commute 1 0.1672 85 76.710 0.68261   
## Hours.Slept:Breakfast 1 4.0411 84 72.669 0.04441 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# we can remove average commute

* Is the interaction term significant? First, model without interaction

m.s<- glm(Exercise~Hours.Slept + Breakfast+ I(Hours.Slept^2), data=data, family= poisson)  
# without interaction  
summary(m.s)

##   
## Call:  
## glm(formula = Exercise ~ Hours.Slept + Breakfast + I(Hours.Slept^2),   
## family = poisson, data = data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.24702 1.08274 -0.228 0.8195   
## Hours.Slept 0.44903 0.27135 1.655 0.0980 .  
## BreakfastYes 0.19209 0.12886 1.491 0.1360   
## I(Hours.Slept^2) -0.03089 0.01687 -1.831 0.0671 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 84.306 on 89 degrees of freedom  
## Residual deviance: 76.877 on 86 degrees of freedom  
## AIC: 374.53  
##   
## Number of Fisher Scoring iterations: 5

# AIC = 374.53  
# No overdispersion

Model with interaction

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

##   
## Call:  
## glm(formula = Exercise ~ Hours.Slept + Hours.Slept \* Breakfast +   
## I(Hours.Slept^2), family = poisson, data = data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|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:BreakfastYes 0.18981 0.09693 1.958 0.0502 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 84.306 on 89 degrees of freedom  
## Residual deviance: 72.879 on 85 degrees of freedom  
## AIC: 372.53  
##   
## Number of Fisher Scoring iterations: 5

#AIC = 372.53  
# No overdispersion

Compare the two models in an analysis of deviance:

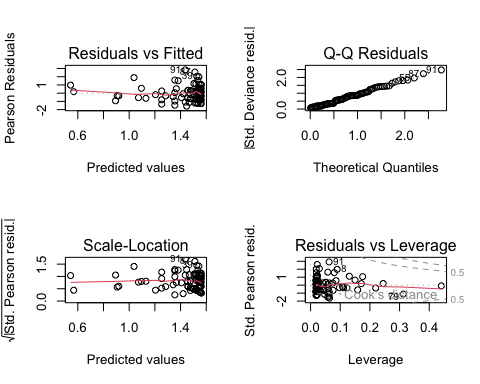
anova(m.s,m.s.i,test ="Chisq")

## Analysis of Deviance Table  
##   
## Model 1: Exercise ~ Hours.Slept + Breakfast + I(Hours.Slept^2)  
## Model 2: Exercise ~ Hours.Slept + Hours.Slept \* Breakfast + I(Hours.Slept^2)  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 86 76.877   
## 2 85 72.879 1 3.9976 0.04557 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The more complex model, with the interaction term has a lower residual deviance and is a significantly better fit of the data p = 0.04557. It is more complex, but it is also better at modelling the data and has a lower AIC. Both models have a residual deviance below the residual degrees of freedom, meaning they are not overdispersed.

* Checking diagnostic plots

par(mfrow=c(2,2))  
plot(m.s.i)



anova(m.s.i, test="Chisq")

## Analysis of Deviance Table  
##   
## Model: poisson, link: log  
##   
## Response: Exercise  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 89 84.306   
## Hours.Slept 1 0.8624 88 83.444 0.35307   
## Breakfast 1 2.8629 87 80.581 0.09065 .  
## I(Hours.Slept^2) 1 3.7039 86 76.877 0.05429 .  
## Hours.Slept:Breakfast 1 3.9976 85 72.879 0.04557 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

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)

# 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 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)

