What influences co-operative decision making?

Dr. Ulrik Beierholm, Associate Professor of Psychology, Durham University, UK

Answer 1:

```
# Load the data from the CSV file
data <- read.csv("PDPMetaUni.csv")
```

Answer 2:

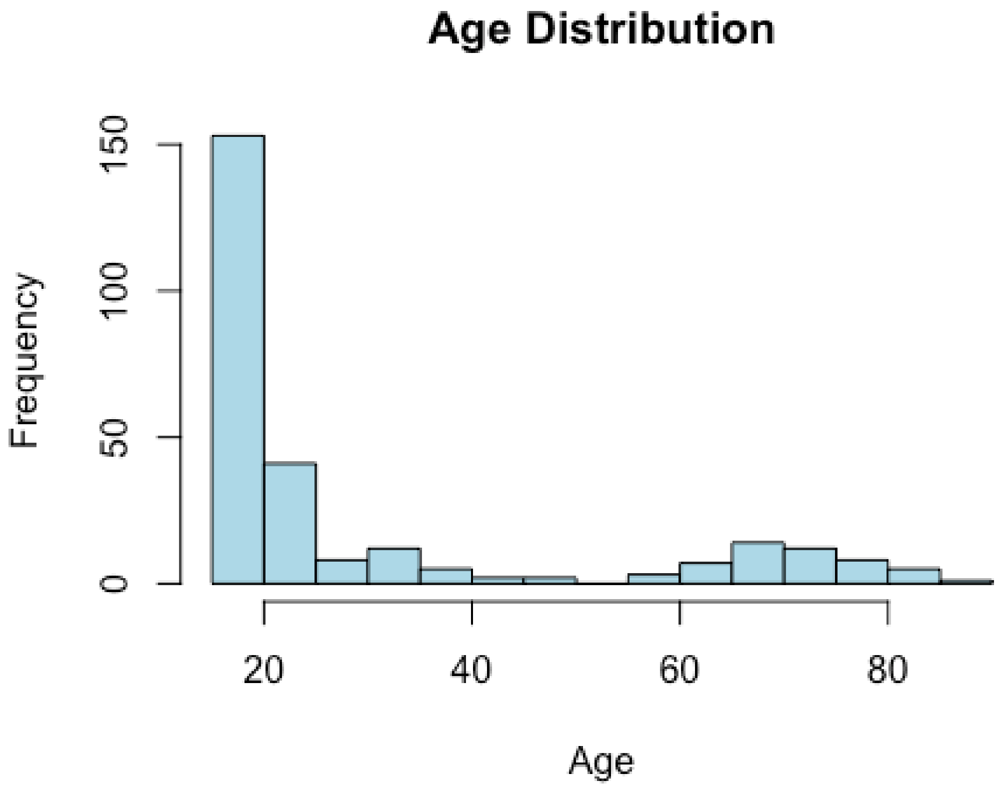
```
# Display the first few rows of the data
head(data)

##   Choice TypeHum Anxiety Avoidance Cooperativeness  Age Gender Subject
## 1      2       1    2.25      3.25              35 19.0      2       1
## 2      1       1    3.50      3.50              35 19.0      1       2
## 3      2       1    2.00      2.00              29 20.0      1       3
## 4      1       1    1.58      1.54              37 19.0      1       4
## 5      2       1    2.75      3.54              26 26.0      1       5
## 6      1       1    1.25      2.25              39 18.9      1       6

# Summary of the data
summary(data)

##      Choice          TypeHum          Anxiety         Avoidance    
##  Min.   :1.000   Min.   :0.0000   Min.   :1.000   Min.   :1.030  
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:1.940   1st Qu.:2.010  
##  Median :1.000   Median :0.0000   Median :2.750   Median :2.670  
##  Mean   :1.188   Mean   :0.4539   Mean   :2.829   Mean   :2.749  
##  3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:3.587   3rd Qu.:3.498  
##  Max.   :2.000   Max.   :1.0000   Max.   :5.590   Max.   :5.340  
##                                                                  
##  Cooperativeness      Age            Gender         Subject      
##  Min.   : 0.00   Min.   :18.00   Min.   :0.000   Min.   :  1.00  
##  1st Qu.: 0.00   1st Qu.:19.00   1st Qu.:1.000   1st Qu.: 71.25  
##  Median : 0.00   Median :20.00   Median :1.000   Median :141.50  
##  Mean   :10.73   Mean   :30.62   Mean   :1.238   Mean   :141.50  
##  3rd Qu.:31.00   3rd Qu.:32.00   3rd Qu.:2.000   3rd Qu.:211.75  
##  Max.   :41.00   Max.   :86.00   Max.   :2.000   Max.   :282.00  
##                  NA's   :9       NA's   :43
```

Answer 3:

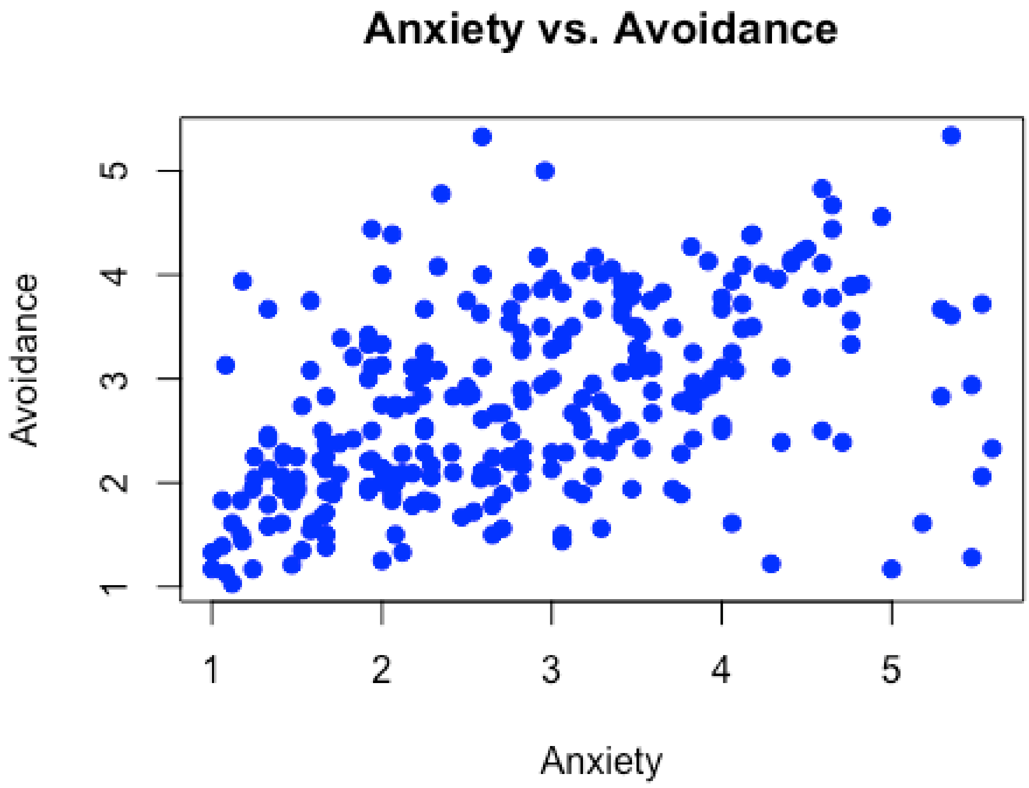
```
#Clean data
data$Choice01<-ifelse(data$Choice==1,1,0)
#Remove NaN data
dataClean<-data[!is.nan(data$Age),]
```

Answer 4:

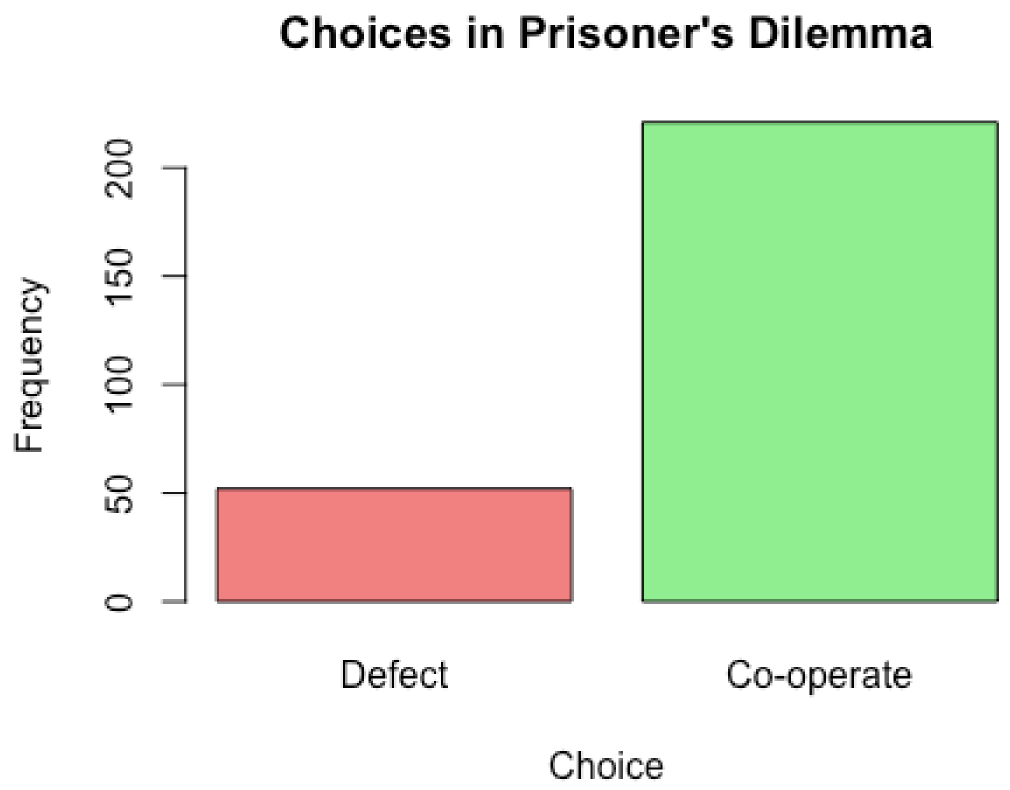
```
# Plot Age distribution
hist(dataClean$Age, main="Age Distribution", xlab="Age", col="lightblue",
border="black")
```




```
# Plot Anxiety vs Avoidance
plot(dataClean$Anxiety, dataClean$Avoidance, main="Anxiety vs. Avoidance",
xlab="Anxiety", ylab="Avoidance", col="blue", pch=19)
```




```r
# Bar plot for Choice
barplot(table(dataClean$Choice01), main="Choices in Prisoner's Dilemma", xlab="Choice", ylab="Frequency", col=c( "lightcoral","lightgreen"), names.arg=c("Defect","Cooperate"))
```



Correlation

The scatter plot for Anxiety vs. Avoidance seems to show that the two are correlated. Will this be a problem?

Answer 5:

```
cor.test(dataClean$Anxiety,dataClean$Avoidance)

##
##  Pearson's product-moment correlation
##
## data:  dataClean$Anxiety and dataClean$Avoidance
## t = 8.6087, df = 271, p-value = 6.133e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3647512 0.5517644
## sample estimates:
##       cor
## 0.4634019
```

Ok, correlation is not too high. Although significant, it should not create problems with collinearity in our model.

Summative Statistics

Calculating the mean and standard deviation for key variables such as Age, Anxiety, and Avoidance provides a concise summary of the data. These statistics help us understand our variables in terms of their central tendency and variability.

Answer 6:

```
# Mean and standard deviation of Age
mean_age <- mean(dataClean$Age)
sd_age <- sd(dataClean$Age)

# Mean and standard deviation of Anxiety
mean_anxiety <- mean(dataClean$Anxiety)
sd_anxiety <- sd(dataClean$Anxiety)

# Mean and standard deviation of Avoidance
mean_avoidance <- mean(dataClean$Avoidance)
sd_avoidance <- sd(dataClean$Avoidance)

# Print the results
cat("Mean Age:", mean_age, "\nSD Age:", sd_age, "\n")

## Mean Age: 30.61512
## SD Age: 20.13245

cat("Mean Anxiety:", mean_anxiety, "\nSD Anxiety:", sd_anxiety, "\n")

## Mean Anxiety: 2.828278
## SD Anxiety: 1.114453

cat("Mean Avoidance:", mean_avoidance, "\nSD Avoidance:", sd_avoidance, "\
n")

## Mean Avoidance: 2.766264
## SD Avoidance: 0.919936
```

Answer 7:

```
# Logistic regression model with Choice as the outcome variable
model1 <- glm(Choice01 ~  Age + Anxiety + Avoidance + TypeHum, data=dataCl
ean, family=binomial)

# Summary of the model
summary(model1)

##
## Call:
## glm(formula = Choice01 ~ Age + Anxiety + Avoidance + TypeHum,
##     family = binomial, data = dataClean)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.038177   0.604301   1.718   0.0858 .
## Age         0.008536   0.008415   1.014   0.3104
```

```
## Anxiety      0.024572   0.157854   0.156   0.8763
## Avoidance   -0.079142   0.189344  -0.418   0.6760
## TypeHum      0.808679   0.335621   2.409   0.0160 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 265.85  on 272  degrees of freedom
## Residual deviance: 258.44  on 268  degrees of freedom
## AIC: 268.44
##
## Number of Fisher Scoring iterations: 4
```

The only variable that has a significant effect on the model (apart from the intercept) is TypeHum. That the estimate for TypeHum is positive indicates that participants are more likely to want to co-operate with a human, as opposed to a computer.

Answer 8:

```
# Logistic regression model with Choice as the outcome variable
model2 <- glm(Choice01 ~  Age + TypeHum, data=dataClean, family=binomial)

# Summary of the model
summary(model2)

##
## Call:
## glm(formula = Choice01 ~ Age + TypeHum, family = binomial, data = dataC
lean)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.890167   0.308177   2.888  0.00387 **
## Age         0.008466   0.008420   1.006  0.31465
## TypeHum     0.807850   0.335007   2.411  0.01589 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 265.85  on 272  degrees of freedom
## Residual deviance: 258.62  on 270  degrees of freedom
## AIC: 264.62
##
## Number of Fisher Scoring iterations: 4
```

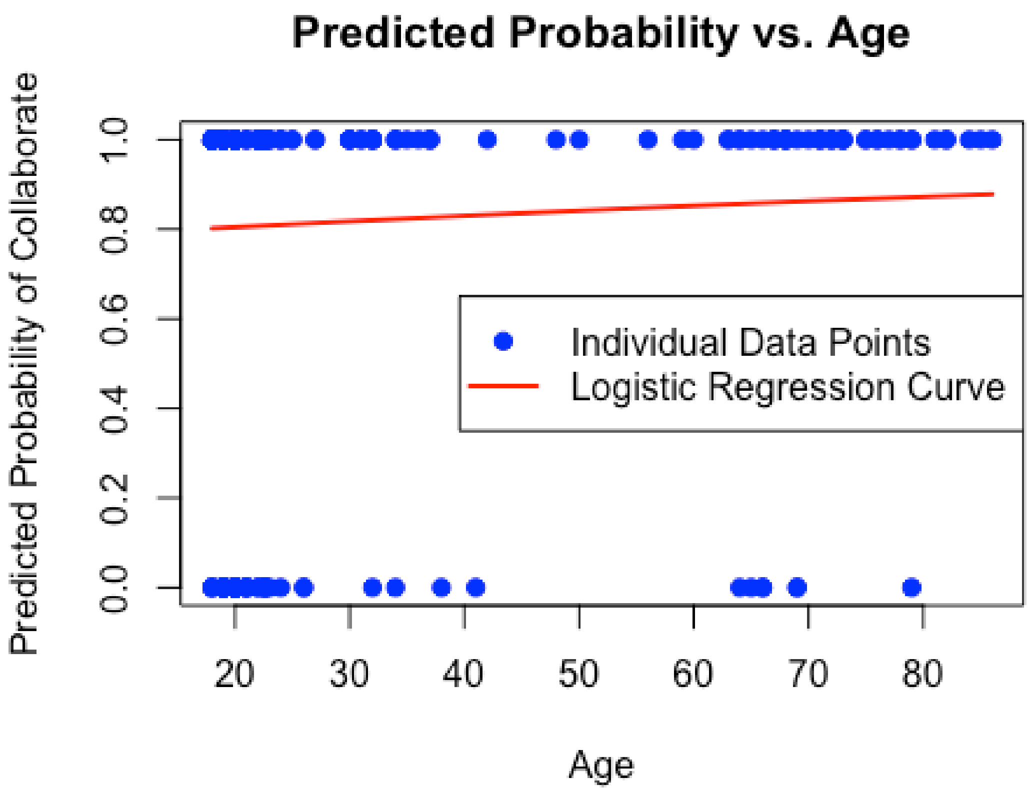
Comparing the two models, we can see that model2 has a lower AIC value, and may be better to trust going forwards. Potentially we could continue to try different models (e.g. different combinations of variables) until we find the model with the smallest AIC value. This will however lead to overfitting on this data set and make us less able to generalise to larger datasets (some researchers in data modeling split their data into training and testing datasets to test for this, but that is a discussion for another day!).

Graphical representation

It is useful to also look at the graphical representation of the model.

Answer 9:

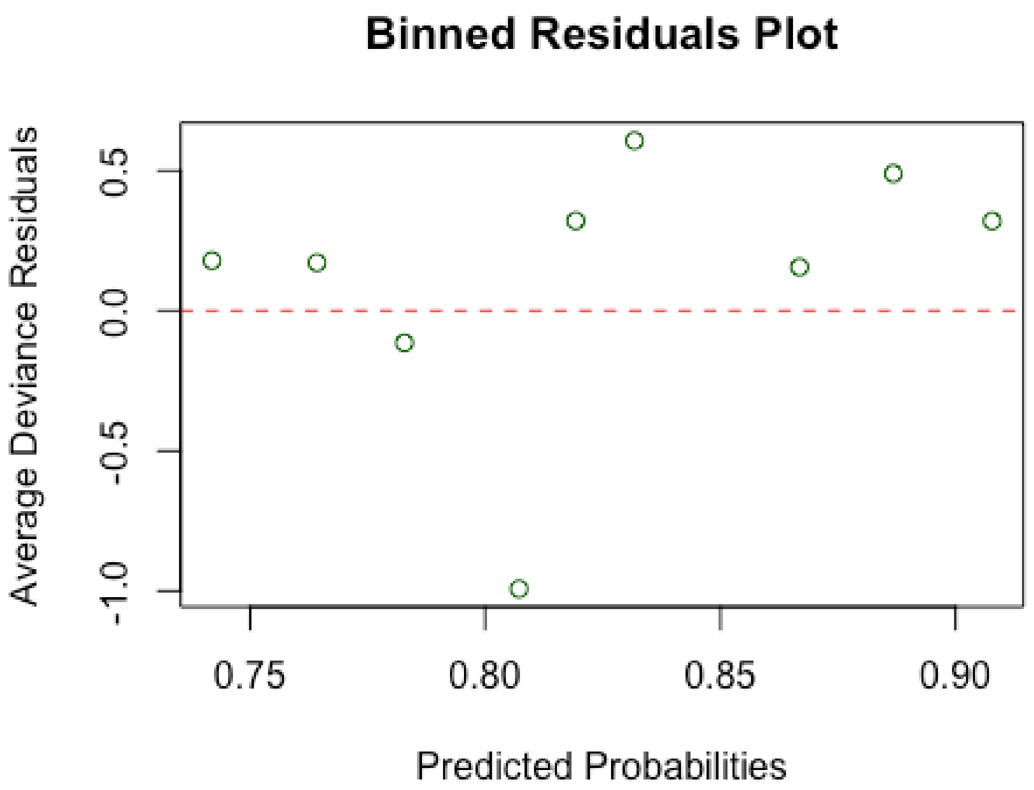
```
# Plot the individual data points

plot(dataClean$Age, dataClean$Choice01, main="Predicted Probability vs. Ag
e", xlab="Age", ylab="Predicted Probability of Collaborate", col="blue", p
ch=19)

# Add the logistic regression curve
age_seq <- seq(min(dataClean$Age), max(dataClean$Age), length.out=100)
predicted_curve <- predict(model2, newdata=data.frame(Age=age_seq, Gender=
mean(dataClean$Gender), Anxiety=mean(dataClean$Anxiety), Avoidance=mean(da
taClean$Avoidance), TypeHum=mean(dataClean$TypeHum)), type="response")
lines(age_seq, predicted_curve, col="red", lwd=2)

# Add a legend
legend("right", legend=c("Individual Data Points", "Logistic Regression Cu
rve"), col=c("blue", "red"), pch=c(19, NA), lty=c(NA, 1), lwd=c(NA, 2))
```



The plot confirms what we already saw in the model: there is a small tendency for participants to collaborate more when they get older.

Answer 10:

```r
# Calculate the predicted probabilities
predicted_probs <- predict(model2, type="response")

# Define the number of bins
num_bins <- 10

# Create bins
bins <- cut(predicted_probs, breaks=num_bins, labels=FALSE)

# Calculate the average residuals within each bin
binned_residuals <- tapply(residuals(model2, type="deviance"), bins, mean)

# Calculate the midpoints of each bin
bin_midpoints <- tapply(predicted_probs, bins, mean)

# Plot the binned residuals
plot(bin_midpoints, binned_residuals, main="Binned Residuals Plot", xlab="
Predicted Probabilities", ylab="Average Deviance Residuals", col="darkgree
n")
abline(h=0, col="red", lty=2)
```



Binned residuals plots are useful for diagnosing issues with the model fit, such as systematic deviations or non-linearity. They provide a visual way to check if the model's predictions are consistent across different probability ranges and if the model is appropriately capturing the data. In this case, it looks like there is no systematic problem with the model. Note that we chose to use the deviance residuals as they are often more useful for logistic models.