R-DPD-Age-Estimation-Analysis

Zheng Wang

Install and load the dataset

```
options(repos = c(CRAN = "https://cran.rstudio.com/"))
install.packages('dafs')

## Warning: unable to access index for repository
https://cran.rstudio.com/src/contrib:
##   cannot open URL 'https://cran.rstudio.com/src/contrib/PACKAGES'

## Warning: package 'dafs' is not available for this version of R
##
## A version of this package for your version of R might be available
elsewhere,
## see the ideas at
## https://cran.r-project.org/doc/manuals/r-patched/R-
admin.html#Installing-packages

## Warning: unable to access index for repository
https://cran.rstudio.com/bin/windows/contrib/4.3:
##   cannot open URL
'https://cran.rstudio.com/bin/windows/contrib/4.3/PACKAGES'

library(dafs)

## Warning: package 'dafs' was built under R version 4.3.3

## Loading required package: s20x

## Warning: package 's20x' was built under R version 4.3.3

library(ggplot2)

data(dpd.df)
```

First of all, explore the data and check for any missing values: There is no any missing value in dataset, so we can start.

```
View(dpd.df)
summary(dpd.df)

##      sample           age         sex        molar          dpd.ratio
##  Min.   : 1.00   Min.   :15.00   F:13   Min.   :16.00   Min.
:0.620
##  1st Qu.: 6.25   1st Qu.:21.75   M: 9   1st Qu.:27.00   1st
Qu.:1.087
##  Median :11.50   Median :38.50          Median :37.00   Median
:1.265
```

```
##  Mean   :11.50   Mean   :41.23          Mean   :36.09   Mean
:1.427
##  3rd Qu.:16.75   3rd Qu.:60.00          3rd Qu.:46.75   3rd
Qu.:1.587
##  Max.   :22.00   Max.   :73.00          Max.   :48.00   Max.
:3.050
##     est.age       assoc.error      real.error
##  Min.   :24.00   Min.   :14.60   Min.   :-29.7000
##  1st Qu.:33.88   1st Qu.:14.60   1st Qu.: -7.2250
##  Median :37.70   Median :14.70   Median :  0.1500
##  Mean   :41.15   Mean   :14.91   Mean   :  0.8045
##  3rd Qu.:44.60   3rd Qu.:14.88   3rd Qu.:  8.4500
##  Max.   :75.70   Max.   :16.80   Max.   : 26.0000
```

```r
dim(dpd.df)
```

```
## [1] 22  8
```

```r
sum(is.na(dpd.df))
```

```
## [1] 0
```

```r
colSums(is.na(dpd.df))
```

```
##      sample         age         sex       molar   dpd.ratio
est.age
##           0           0           0           0           0
0
## assoc.error  real.error
##           0           0
```

Gender count(How many male individuals are there in this study? How many female individuals are there in this study?)

```r
summary(dpd.df$sex)
```

```
##  F  M
## 13  9
```

```r
print(paste("Male individuals are there in this study is:",
table(dpd.df$sex)['M']))
```

```
## [1] "Male individuals are there in this study is: 9"
```

```r
print(paste("Female individuals are there in this study is:",
table(dpd.df$sex)['F']))
```

```
## [1] "Female individuals are there in this study is: 13"
```

Mean and median age for female.(What is the mean and median age of the female individuals in this study?)

```
Fe_data <- subset(dpd.df, sex == "F")
Ma_data <- subset(dpd.df, sex == "M")
mean_Fe_age <- mean(Fe_data$age, na.rm = TRUE)
median_Fe_age <- median(Fe_data$age, na.rm = TRUE)
print(paste("The mean and median age of the female individuals are:",
mean_Fe_age, "and", median_Fe_age))
```

```
## [1] "The mean and median age of the female individuals are:
40.3076923076923 and 37"
```

Age range check(What is the range for the age of the male individuals in this study? What is the range for the age of the female individuals in this study?)

```
print(paste("The female age range is:", paste(range(Fe_data$age),
collapse = " - ")))
```

```
## [1] "The female age range is: 15 - 73"
```

```
print(paste("The male age range is:", paste(range(Ma_data$age),
collapse = " - ")))
```

```
## [1] "The male age range is: 17 - 61"
```

maximum DPD ratio(What is the maximum DPD ratio for the male individuals in this study? What is the maximum DPD ratio for the female individuals in this study?)

```
max_Ma_DPD_ratio <- max(Ma_data$dpd.ratio,na.rm = TRUE)
max_Fe_DPD_ratio <- max(Fe_data$dpd.ratio,na.rm = TRUE)

print(paste("The maximum DPD ratio for the male individuals is:",
paste(max_Ma_DPD_ratio)))
```

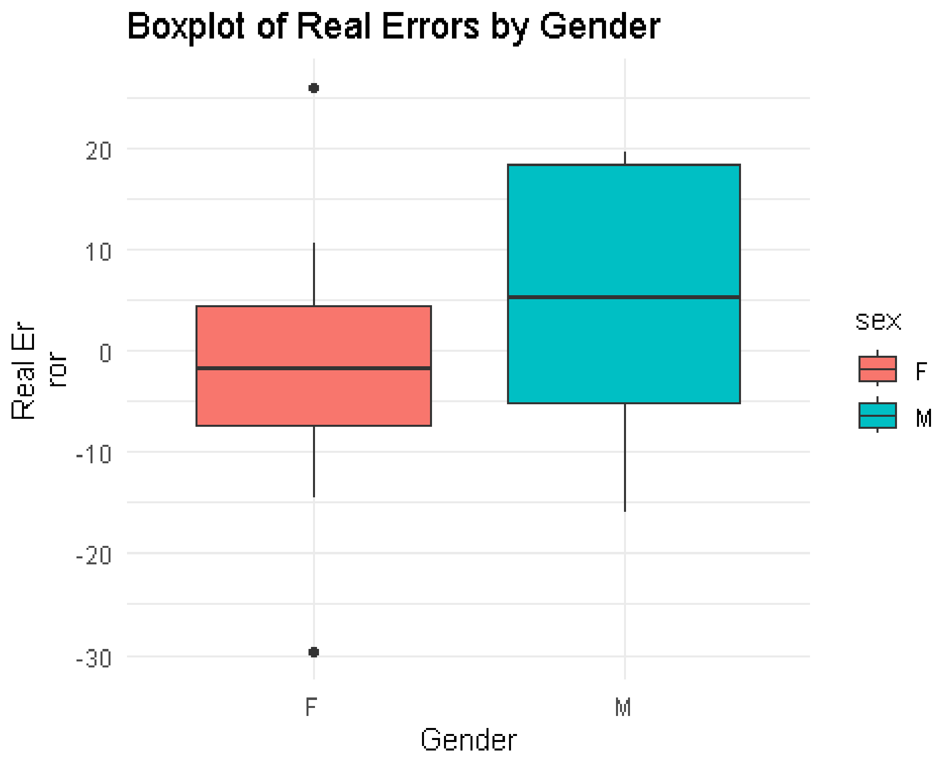
```
## [1] "The maximum DPD ratio for the male individuals is: 1.69"
```

```
print(paste("the maximum DPD ratio for the female individuals is:",
paste(max_Fe_DPD_ratio)))
```

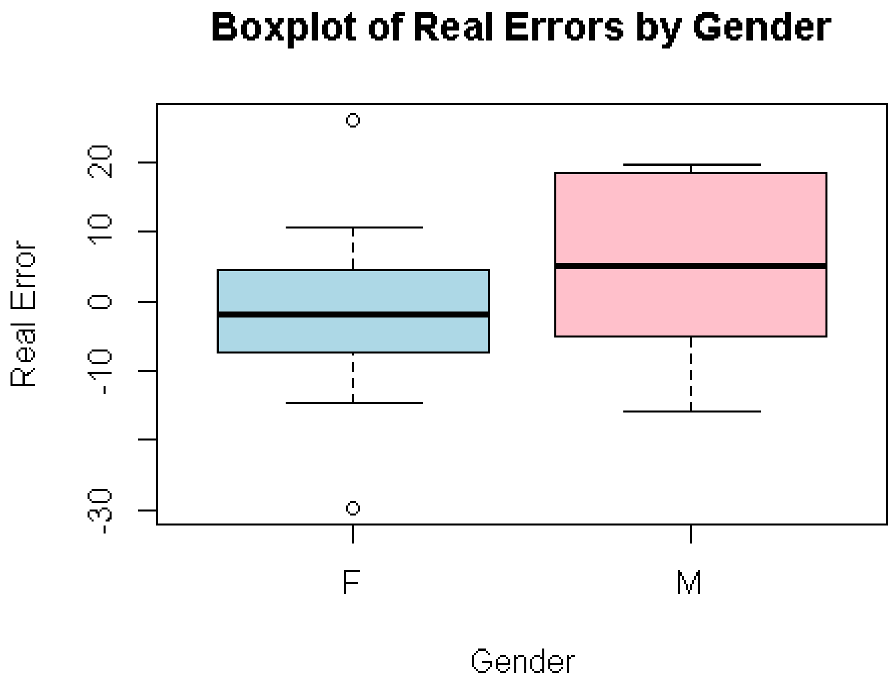
```
## [1] "the maximum DPD ratio for the female individuals is: 3.05"
```

To see whether there is a difference between the actual errors for the female individuals and the actual errors for the male individuals, create a box and whisker plot comparing the actual errors obtained for female individuals with the actual errors obtained for male individuals.

```
ggplot(dpd.df, aes(x = sex, y = real.error, fill = sex)) +
 geom_boxplot() +
 labs(title = "Boxplot of Real Errors by Gender", x = "Gender", y =
"Real Er
ror") +
 theme_minimal()
```



```r
boxplot(real.error ~ sex, data = dpd.df,
 main = "Boxplot of Real Errors by Gender",
 xlab = "Gender",
 ylab = "Real Error",
 col = c("lightblue", "pink"))
```



the Boxplot

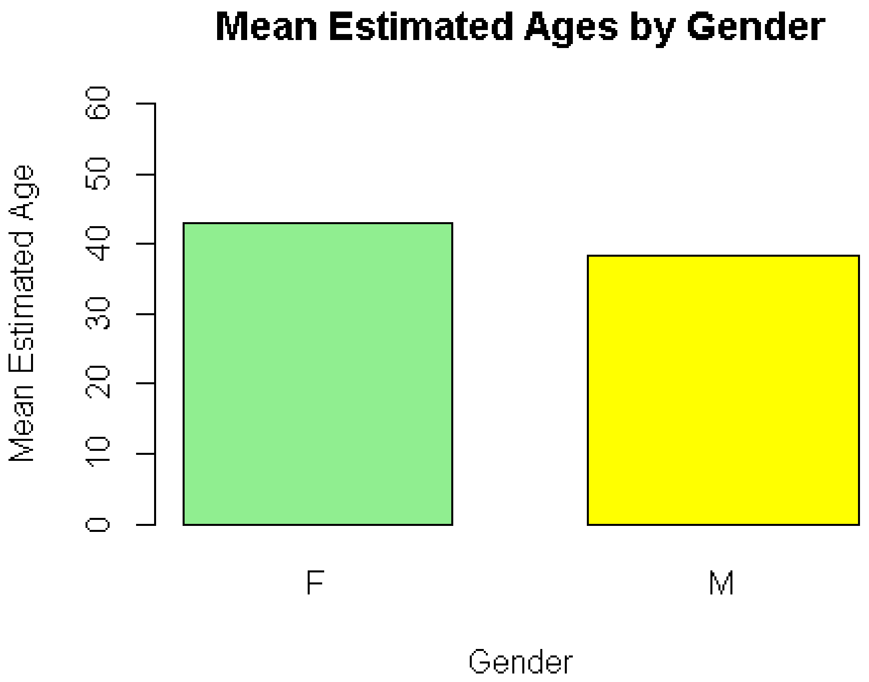
Comparing Real Errors for Male (M) and Female (F) Individuals: 1. Median: The median for male individuals (M) is higher than for female individuals (F), indicating that males tend to have higher real errors compared to females. 2. IQR: The IQR for male individuals is wider than that for female individuals, meaning that there is more variability in the real errors among males than females. 3. Whiskers: For the range within 1.5 times the IQR, males extend -16 to 19, female -14 to 19. 4. Outliers: No outlier for the male but there are 2 outliers for females. One is higher than 25 and the other one is around -30.

IQR(What is the interquartile range for the actual error for female individuals? What is the interquartile range for the actual error for male individuals?)

```
IQR_Fe <- IQR(Fe_data$real.error)
IQR_Ma <- IQR(Ma_data$real.error)
print(paste("Interquartile Range for Female Individuals:", IQR_Fe))

## [1] "Interquartile Range for Female Individuals: 11.9"

print(paste("Interquartile Range for Male Individuals:", IQR_Ma))

## [1] "Interquartile Range for Male Individuals: 23.6"
```

Bar plot of mean estimated age by gender(To compare the estimated ages for each gender, create a bar plot showing the mean estimated ages for each gender.)

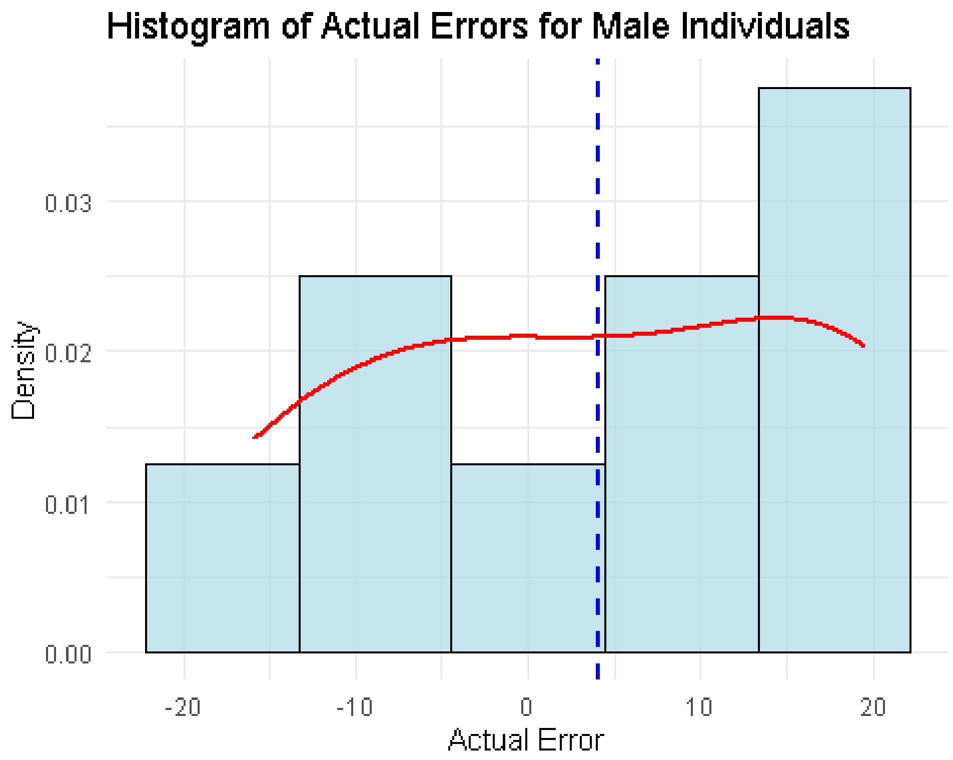
```
mean_est_age <- aggregate(est.age ~ sex, data = dpd.df, FUN = mean)
barplot(mean_est_age$est.age,
        names.arg = mean_est_age$sex,
        col = c("lightgreen", "yellow"),
        main = "Mean Estimated Ages by Gender",
        xlab = "Gender", ylab = "Mean Estimated Age",
        width = 0.5, space = 0.5, ylim = c(0, 60))
```



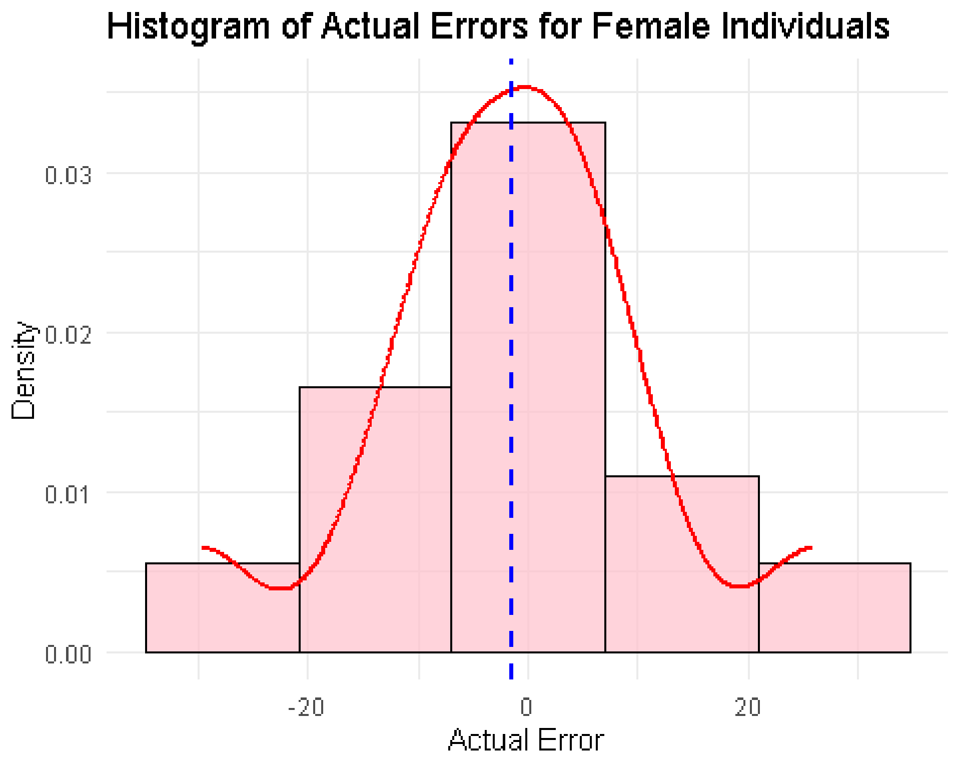
Represent the actual errors for the female individuals and the actual errors for the male individuals as separate histograms. In each histogram, plot a line representing the density and a vertical line marking the mean value of the actual errors.

Histograms with density + mean line

```
mean_Ma_re.er <- mean(Ma_data$real.error)
mean_Fe_re.er <- mean(Fe_data$real.error)

ggplot(Ma_data, aes(x = real.error)) +
  geom_histogram(aes(y = after_stat(density)), bins = 5, fill =
"lightblue", color = "black", alpha = 0.7) +
  geom_density(color = "red", linewidth = 1) +
  geom_vline(aes(xintercept = mean_Ma_re.er), color = "blue", linetype
= "dashed", linewidth = 1) +
  labs(title = "Histogram of Actual Errors for Male Individuals", x =
"Actual Error", y = "Density") +
  theme_minimal()
```



```
ggplot(Fe_data, aes(x = real.error)) +
  geom_histogram(aes(y = after_stat(density)), bins = 5, fill = "pink",
color = "black", alpha = 0.7) +
  geom_density(color = "red", linewidth = 1) +
  geom_vline(aes(xintercept = mean_Fe_re.er), color = "blue", linetype
= "dashed", linewidth = 1) +
  labs(title = "Histogram of Actual Errors for Female Individuals", x =
"Actual Error", y = "Density") +
  theme_minimal()
```



Histogram of Actual Errors for Male Individuals 1. The histogram shows that the actual errors for male individuals are not symmetrically distributed. 2. The actual errors range from approximately -20 to +20. 3. The red density curve does not show a distinct peak. 4. The blue dashed line marks the mean of the actual errors around 4.

Histogram of Actual Errors for Female Individuals 1. The histogram and the red density curve suggest that the actual errors for female individuals are distributed symmetrically around zero. 2. Unlike the male error distribution, this plot shows a distinct peak in the density curve at zero. 3. Most of the errors fall within the range of -20 to +20. 4. The blue dashed line indicates the mean error, which is very close to zero.

Advanced Analysis Simple linear regression

```
Fe_data <- subset(dpd.df, sex == "F")
Ma_data <- subset(dpd.df, sex == "M")

model_simple <- lm(age ~ dpd.ratio, data = dpd.df)
summary(model_simple)

##
## Call:
## lm(formula = age ~ dpd.ratio, data = dpd.df)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max 
## -29.8056  -8.6384  -0.0186   8.4226  25.8918 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   10.826      7.778   1.392 0.179272    
## dpd.ratio     21.301      5.011   4.251 0.000391 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 14.34 on 20 degrees of freedom
## Multiple R-squared:  0.4746, Adjusted R-squared:  0.4484 
## F-statistic: 18.07 on 1 and 20 DF,  p-value: 0.0003914
```

Intercept (10.83): When the DPD ratio is zero (which may not be realistic), the predicted age would be approximately 10.83 years. However, this value is not statistically significant (p = 0.179), meaning it's not reliable.

DPD ratio coefficient (21.30): For every 1 unit increase in dpd.ratio, the predicted age increases by approximately 21.3 years. his relationship is statistically significant (p < 0.001).

R-squared (0.475): About 47.5% of the variation in age can be explained by the DPD ratio. This is a moderate level of explanatory power.

Residual standard error (14.34): On average, the predictions are off by about 14 years, which indicates there is still room to improve the model.

F-statistic (18.07, p < 0.001): The model as a whole is statistically significant, meaning DPD ratio is a useful predictor of age.

Add gender as a categorical predictor

```
model_gender <- lm(age ~ dpd.ratio + sex, data = dpd.df)
summary(model_gender)
```

```
## 
## Call:
## lm(formula = age ~ dpd.ratio + sex, data = dpd.df)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -27.3317  -8.3321   0.4596   9.7681  29.0355 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    6.575      8.618   0.763 0.454850    
## dpd.ratio     22.271      5.056   4.405 0.000304 ***
## sexM           7.003      6.275   1.116 0.278304    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.25 on 19 degrees of freedom
## Multiple R-squared:  0.507,  Adjusted R-squared:  0.4551
## F-statistic: 9.768 on 2 and 19 DF,  p-value: 0.001209
```

Coefficients: Intercept (6.575, p = 0.455): Not statistically significant. When dpd.ratio = 0 and sex = F (reference group), the predicted age is ~6.58 years — but this value isn't reliable.

DPD ratio (22.271, p = 0.0003) A 1-unit increase in dpd.ratio leads to an increase of ~22.3 years in predicted age. Highly statistically significant → DPD ratio remains a strong predictor.

sex M (7.003, p = 0.278) Being male increases predicted age by about 7 years, but this effect is not statistically significant. The gender effect does not significantly improve the model on its own.

Model Fit: R-squared = 0.507 The model explains 50.7% of the variation in age — a slight improvement over the simple model (which had 47.5%).

Adjusted R-squared = 0.455 Adjusted for the number of predictors. Still better than before (was 0.448).

F-statistic = 9.768 (p = 0.0012) The overall model is statistically significant → the predictors together explain age variation well.

conclusion : Adding gender to the model slightly improves age prediction, but gender itself is not a significant predictor (p = 0.278). The DPD ratio remains the key variable, showing a strong and significant relationship with age. Overall, the model performs a bit better than the simple one.

Add interaction term between dpd.ratio and sex(Test if the slope of dpd.ratio is different for male and female.)

```
model_interaction <- lm(age ~ dpd.ratio * sex, data = dpd.df)
summary(model_interaction)
```

```
##
## Call:
## lm(formula = age ~ dpd.ratio * sex, data = dpd.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.961  -7.796  -1.872   9.824  28.595
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)        8.355      8.773   0.952 0.353528
## dpd.ratio         21.096      5.173   4.078 0.000706 ***
## sexM            -24.921     31.507  -0.791 0.439264
## dpd.ratio:sexM   24.343     23.545   1.034 0.314892
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.23 on 18 degrees of freedom
## Multiple R-squared:  0.5346, Adjusted R-squared:  0.457
## F-statistic: 6.892 on 3 and 18 DF,  p-value: 0.00275
```

Coefficients Interpretation: Intercept (8.36, p = 0.356) Not statistically significant. Represents the predicted age when dpd.ratio = 0 and sex = F.

dpd.ratio (21.10, p < 0.001) For females, each 1 unit increase in dpd.ratio increases age by ~21.1 years. This is a strong and significant predictor.

sexM (-24.92, p = 0.439) Being male decreases baseline age prediction by ~24.9 years when dpd.ratio = 0. Not statistically significant.

dpd.ratio:sexM (24.34, p = 0.315) The interaction is not significant, meaning that the effect of dpd.ratio on age does not significantly differ between males and females.

Model Fit: R-squared = 0.534 → Explains 53.4% of age variation (slightly better than previous models)

Adjusted R-squared = 0.457 → About the same as previous model with only sex and dpd.ratio

F-statistic = 6.89 (p = 0.0027) → Model overall is statistically significant

Summary: This model shows that dpd.ratio is still the only significant predictor of age. Although the interaction term (dpd.ratio × sex) was included, it does not significantly improve the model, and gender differences in the effect of dpd.ratio are not statistically meaningful in this dataset.

Try a Nonlinear (Quadratic) Model(Sometimes age increases with dpd.ratio nonlinearly.)

```
model_poly <- lm(age ~ poly(dpd.ratio, 2), data = dpd.df)
summary(model_poly)

##
## Call:
## lm(formula = age ~ poly(dpd.ratio, 2), data = dpd.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.824  -5.584  -0.105   9.055  25.528
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           41.227      3.081  13.379 4.04e-11 ***
## poly(dpd.ratio, 2)1   60.962     14.453   4.218 0.000466 ***
## poly(dpd.ratio, 2)2  -12.013     14.453  -0.831 0.416212
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.45 on 19 degrees of freedom
## Multiple R-squared:  0.4931, Adjusted R-squared:  0.4397
## F-statistic:  9.24 on 2 and 19 DF,  p-value: 0.001574
```

This model fits a 2nd-degree polynomial (quadratic curve) to allow for a nonlinear relationship between dpd.ratio and age. Coefficients: Intercept (41.23, p < 0.001) Statistically significant. Represents the baseline average age when centered/polynomial terms are zero.

poly(dpd.ratio, 2)1 = 60.96 (p < 0.001) The linear term is statistically significant, indicating a strong upward trend between dpd.ratio and age.

poly(dpd.ratio, 2)2 = -12.01 (p = 0.416) The quadratic term is not statistically significant, suggesting no strong evidence of curvature (nonlinearity) in this relationship.

Model Fit: R-squared = 0.493 Explains 49.3% of the variation in age — similar to the simple linear model (47.5%).

Adjusted R-squared = 0.439 Slightly lower than the linear model with gender.

F-statistic = 9.24, p = 0.0016 The model is statistically significant overall.

Conclusion: Although the quadratic model is statistically significant overall, the nonlinear (curved) component is not significant, meaning a straight-line model fits nearly as well. Therefore, a simple linear regression (with or without gender) is more interpretable and performs just as well.

Include Additional Variable: molar(Add molar count to account for biological variation.)

```
model_molar <- lm(age ~ dpd.ratio + molar + sex, data = dpd.df)
summary(model_molar)

##
## Call:
## lm(formula = age ~ dpd.ratio + molar + sex, data = dpd.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.361 -10.091  -1.356  10.570  27.939
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.6343    15.0162   1.108 0.282549
## dpd.ratio    22.0161     5.1091   4.309 0.000422 ***
## molar        -0.2409     0.2932  -0.822 0.422094
## sexM          4.5560     6.9947   0.651 0.523052
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.38 on 18 degrees of freedom
## Multiple R-squared:  0.5248, Adjusted R-squared:  0.4456
## F-statistic: 6.626 on 3 and 18 DF,  p-value: 0.003293
```

Coefficients: Intercept (16.63, p = 0.283) Not statistically significant. Represents the baseline age when all predictors = 0.

DPD ratio (22.02, p = 0.00042) Highly significant. For every 1 unit increase in DPD ratio, age increases by ~22 years. → This confirms DPD ratio is a strong and consistent predictor.

Molar (-0.24, p = 0.422) Not significant. The number of molars has no meaningful impact on predicted age in this model.

SexM (4.56, p = 0.523) Not significant. Being male is associated with a +4.56 year change in predicted age, but this is not statistically reliable.

Model Fit: R-squared = 0.5248 Explains 52.5% of age variability. Slightly better than the linear model without molar.

Adjusted R-squared = 0.4456 About the same as the dpd.ratio + sex model, meaning the molar variable adds little.

F-statistic = 6.626 (p = 0.00329) The model is statistically significant overall.

Conclusion : This multivariate model shows that DPD ratio remains the only significant predictor of age. Neither molar count nor gender significantly improve the model. While the model explains a bit more variance, the added variables do not meaningfully enhance prediction accuracy.

Compare Model Performance

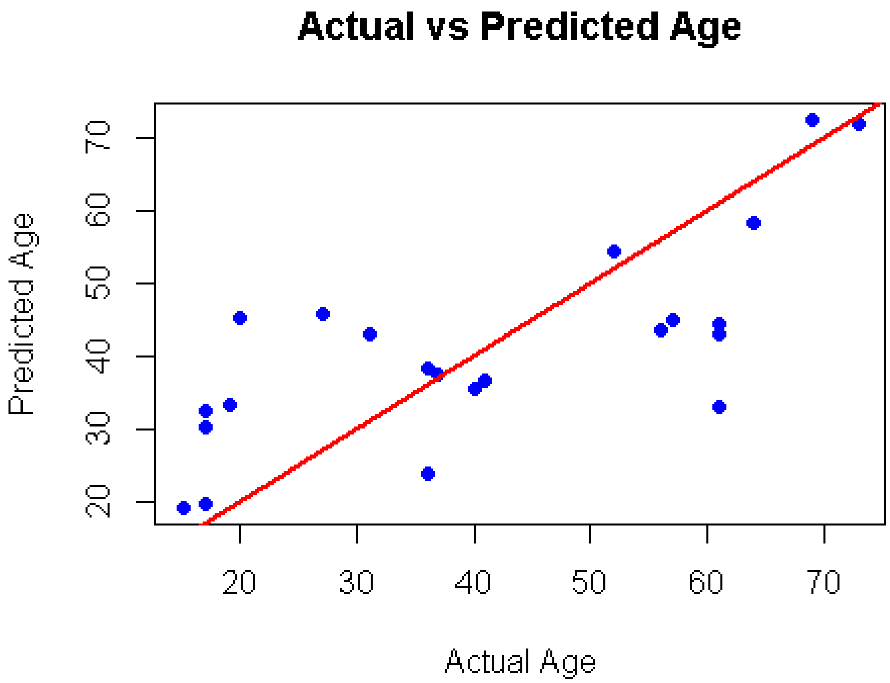
```
AIC(model_simple, model_gender, model_interaction, model_poly,
model_molar)
```

```
##                   df      AIC
## model_simple       3 183.5149
## model_gender       4 184.1179
## model_interaction  5 184.8488
## model_poly         4 184.7292
## model_molar        5 185.3081
```

The simplest model (age ~ dpd.ratio) performs the best, achieving the lowest AIC and strong statistical significance. Adding gender, molar count, or polynomial complexity does not improve model performance, and in fact introduces unnecessary complexity.

Visualize Actual vs Predicted Age

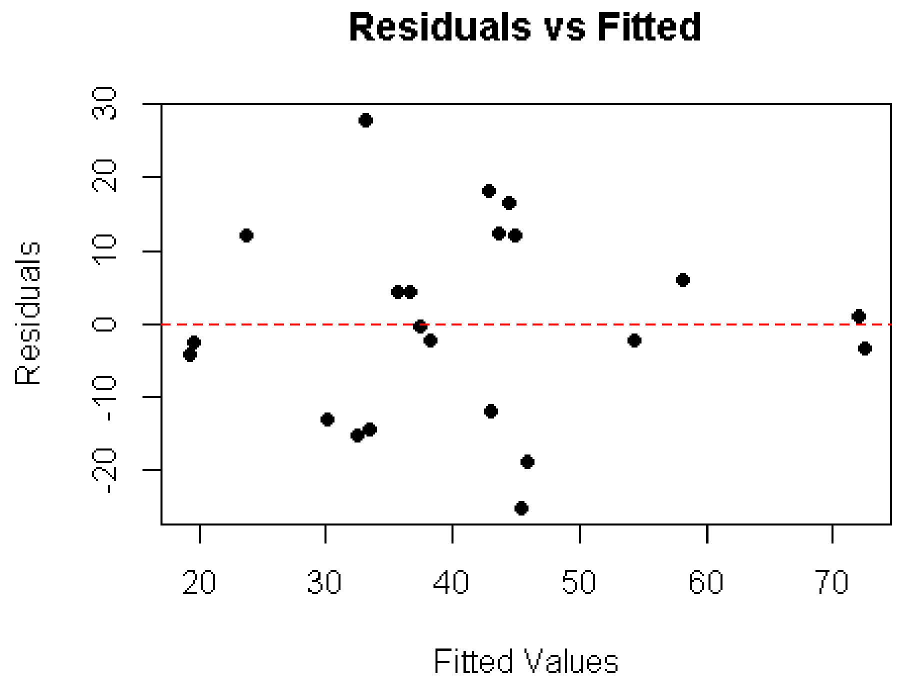
```
plot(dpd.df$age, fitted(model_molar),
     xlab = "Actual Age", ylab = "Predicted Age",
     main = "Actual vs Predicted Age",
     pch = 19, col = "blue")
abline(0, 1, col = "red", lwd = 2)
```



The model captures the main trend between actual and predicted age, but some variability remains—especially in the younger and older age ranges. This confirms your earlier finding: while the model is statistically significant, prediction uncertainty (residual error) is still considerable, and could potentially be reduced with more data or refined features.

Residual Plot Check

```
plot(fitted(model_molar), resid(model_molar),
     xlab = "Fitted Values", ylab = "Residuals",
     main = "Residuals vs Fitted",
     pch = 19)
abline(h = 0, col = "red", lty = 2)
```



The residuals appear roughly randomly scattered, which supports the validity of the model assumptions. There's no strong evidence of nonlinearity or major model violations. However, a few large residuals suggest some predictions are off by 20+ years, which may warrant further investigation or model refinement with more features or a larger dataset.