SYDE 780 Assignment 3: R Module on statistical analyses

Winter 2025

Read in the dataset

```
#Read in data
  fileToImport = "../dataset.csv"
  Data.RSA = read.csv(fileToImport, header=TRUE) #n=1071
```

Data analysis

Objectively supporting our hypothesis

Student's t-tests

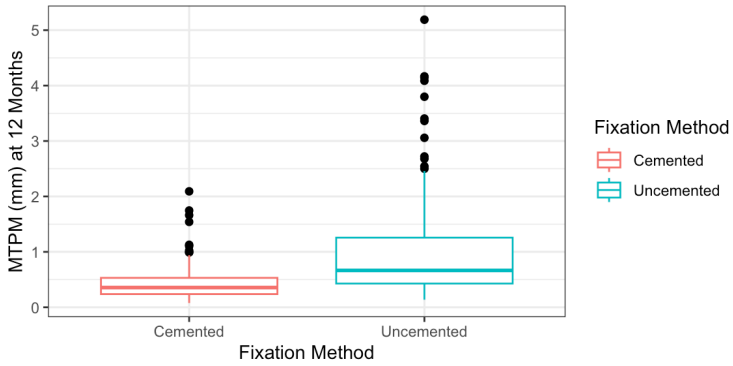
Objective statistical tests enable us to objectively support or reject our hypotheses (without these, we are just guessing).

For example, we can ask "is there a statistical difference in MTPM between cemented and un-cemented fixation methods at 1 year (i.e., 12 months) follow-up?", and objectively state if there is, or is not.

To do this, we will start by plotting MTPM by fixation at 1 year using a box & whisker plot.

Note that the line `Data.RSA[which(Data.RSA$Months==12), ]` in the code below pulls all rows when the column `Months==12`. This is because we only want to compare the one-year data. Run this line in the console by yourself to test it.

```
#Box plot of fixation methods at 1 year (12 month) follow-up post-TKA
  ggplot(data=Data.RSA[which(Data.RSA$Months==12), ], aes(x=Fixation, y=MTPM, color=Fixation)) +
    theme_bw() +
    geom_boxplot(outlier.colour="black", outlier.shape=16, outlier.size=2, notch=FALSE) +
    labs(x="Fixation Method", y="MTPM (mm) at 12 Months", color = "Fixation Method")
```

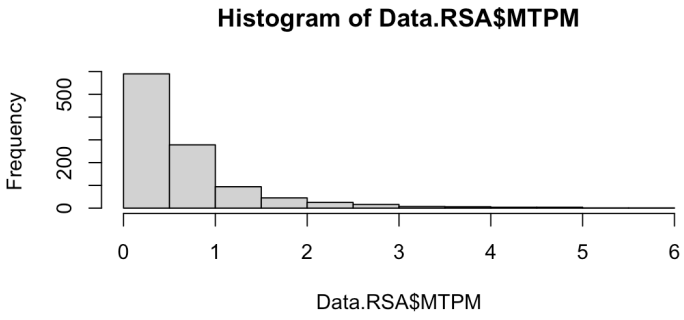


From the image above, the means look like they could be different between fixation methods, but there is some overlap in the boxes. If we want to objectively determine if MTPM between fixation methods **is** different, we need to conduct a statistical hypothesis test.

We will use an **un-paired Student's t-test**. Student's t-tests assume the response variable (in this case MTPM) is normally distributed (if you are un-sure what a "normal" distribution looks like, please look it up!), and both groups (cemented vs. uncemented) have equal variances. The test is invalid if these assumptions are not met.

Let's at least check the first assumption:

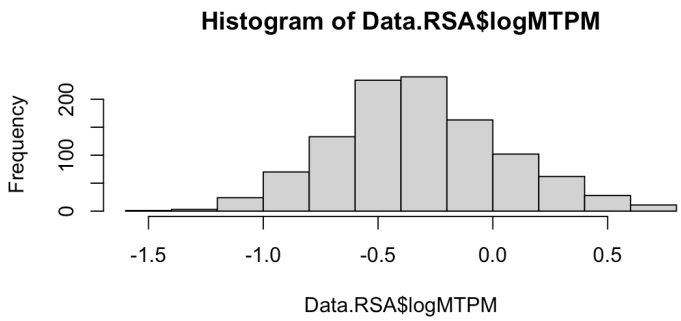
```
#Plot the distribution of MTPM using a histogram
  hist(Data.RSA$MTPM)
```



This doesn't look "normally distributed". Let's perform a $\log_{10}$ transformation on MTPM.

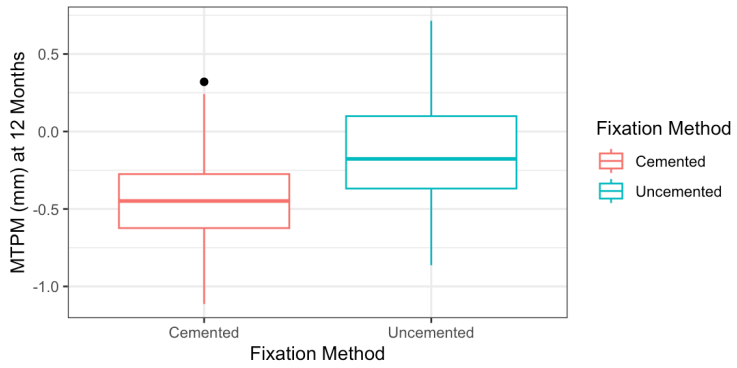
```
#Create a log variable
  Data.RSA$logMTPM = log10(Data.RSA$MTPM)

#Plot the distribution of our new response variable, logMTPM
  hist(Data.RSA$logMTPM)
```



This looks much better, and will satisfy our first assumption. We can move on, using $\log_{10}(MTPM)$ as our variable of interest. First, let's re-plot the box-plots using the $\log_{10}$ transformed MTPM value.

```
#Box plot of fixation methods at 1 year (12 month) follow-up post-TKA
  ggplot(data=Data.RSA[which(Data.RSA$Months==12), ], aes(x=Fixation, y=logMTPM, color=Fixation)) +
    theme_bw() +
    geom_boxplot(outlier.colour="black", outlier.shape=16, outlier.size=2, notch=FALSE) +
    labs(x="Fixation Method", y="MTPM (mm) at 12 Months", color = "Fixation Method")
```



It still looks like the mean MTPM between cemented and un-cemented fixation methods at 1 year could be different, so let's move on to verify this statistically.

Our hypothesis for the t-test is as follows:

$H_0: \overline{x_c} = \overline{x_{uc}}$ (the mean MTPM of cemented implants is equal to the mean MTPM of uncemented implants, our **null hypothesis**)

$H_1: \overline{x_c} \neq \overline{x_{uc}}$ (the mean MTPM of cemented implants is equal to the mean MTPM of uncemented implants, our **alternative hypothesis**)

To determine this, we calculate a t-statistic:

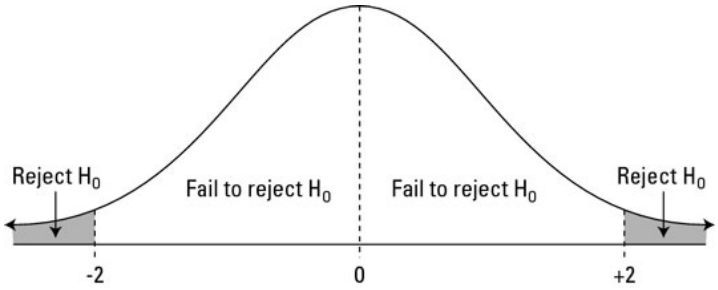
$$t = \frac{\overline{x_c} - \overline{x_{uc}}}{\frac{\sqrt{S^2}}{n_c} + \frac{\sqrt{S^2}}{n_{uc}}}$$

- n reflects the sample size of the cemented and uncemented groups
- S^2 is an estimate of the pooled variance between the two groups, calculated as follows:

$$S^2 = \frac{\sum(x-\overline{x_c})^2 + \sum(x-\overline{x_{uc}})^2}{n_c + n_{uc} - 2}$$

From where that t-statistic value falls on x-axis of the t-distribution (figure below), we can determine if we reject or accept the null hypothesis. Our null hypothesis would be 0 on the t-distribution figure (no difference in means), and the t-statistic we compute corresponds to how far our test lands from this null hypothesis. When the t-statistic falls above or below the critical value (2 in the figure) we "Reject the H_0", and accept the alternative hypothesis, i.e., we conclude that a difference between the means exists.

From the t-statistic, we can also determine a corresponding p-value (probability value, you may remember probability distribution tables from a statistics course). In most cases, significance (α) is set to 0.05, such that p-values <0.05 are considered significant. This corresponds to being in the critical range of the t-distribution (the "Reject the H_0" shaded area). This alpha value means there is a 5% risk of concluding that a difference exists when there is no actual difference (the Type I error rate).



We will conduct this test using the `t.test` function in R.

Note that the line `Data.RSA[which(Data.RSA$Months==12 & Data.RSA$Fixation=="Cemented"), ]$MTPM` pulls all rows when column `Months==12` and column `Fixation==Cemented`. This is because we only want to compare the one-year data between the cemented and uncemented groups. Run this line by yourself to test it.

```
#Create a numeric list of MTPM values for cemented (x_c) and un-cemented (c_uc) conditions.
  x_c = Data.RSA[which(Data.RSA$Months==12 & Data.RSA$Fixation=="Cemented"), ]$logMTPM
  x_uc = Data.RSA[which(Data.RSA$Months==12 & Data.RSA$Fixation=="Uncemented"), ]$logMTPM

#t-test of MTPM by fixation method at 12 months
  t.test(x_c, x_uc, paired=FALSE, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  x_c and x_uc
## t = -8.0874, df = 267, p-value = 2.144e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3835949 -0.2333897
## sample estimates:
##  mean of x  mean of y
## -0.4416438 -0.1331515
```

Conclusion: We reject our null hypothesis and accept the alternative hypothesis. Uncemented implants demonstrated more MTPM at 12 months relative to cemented implants at 12 months post-TKA ($p<0.001$).

Note that in the previous steps we created two vectors of data (logMTPM at 12 months for cemented, logMTPM at 12 months for uncemented)

There is an alternative way to set up a t-test.

First subset the data to include only 12 month data:

```
    Data.RSA.12mon<-Data.RSA[which(Data.RSA$Months==12 ), ]
```

Next, perform the t-test by indicating: t.test(DependentVariable ~ IndependentData, data = DataSetName):

```
#Alternative form for performing a t-test of MTPM by fixation method at 12 months
  t.test(logMTPM~Fixation, data = Data.RSA.12mon, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  logMTPM by Fixation
## t = -8.0874, df = 267, p-value = 2.144e-14
## alternative hypothesis: true difference in means between group Cemented and group Uncemented is not equal to 0
## 95 percent confidence interval:
##  -0.3835949 -0.2333897
## sample estimates:
##   mean in group Cemented mean in group Uncemented
##               -0.4416438               -0.1331515
```

Note that the options for “paired” and “var.equal” are optional (these will be default values if not specified). If you don’t specify them, you may get slightly different results, since the default options are paired = FALSE and var.equal = TRUE.

You can check if you have equal variances using:

```
var.test(logMTPM~Fixation, data = Data.RSA.12mon)
```

```
##
##  F test to compare two variances
##
## data:  logMTPM by Fixation
## F = 0.55391, num df = 140, denom df = 127, p-value = 0.0006814
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3932508 0.7779811
## sample estimates:
## ratio of variances
##          0.5539138
```

Therefore, could also run without the assumption of equal variances which used the Welch t-test that doesn't require equal variances:

```
  t.test(logMTPM~Fixation, data = Data.RSA.12mon)
```

```
##
##  Welch Two Sample t-test
##
## data:  logMTPM by Fixation
## t = -7.9756, df = 233.32, p-value = 6.777e-14
## alternative hypothesis: true difference in means between group Cemented and group Uncemented is not equal to 0
## 95 percent confidence interval:
##  -0.3846982 -0.2322863
## sample estimates:
##   mean in group Cemented mean in group Uncemented
##               -0.4416438               -0.1331515
```

Compare your results.

Non-parmetric equivalent to the Student's t-test: Mann–Whitney U test

We had to transform MTPM to logMTPM to satisfy the requirements for normally distributed data for the t-test. The non-parametric equivalent to the t-test is the Mann–Whitney U test (also called the Mann-Whitney-Wilcoxon Test). Run the Mann Whitney U Test and compare the results to the t-test.

```
#Perform a Mann-Whitney-Wilcoxon Test for on the non-paramteric data
  wilcox.test(MTPM ~ Fixation, data=Data.RSA.12mon)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  MTPM by Fixation
## W = 4531.5, p-value = 1.8e-12
## alternative hypothesis: true location shift is not equal to 0
```

Additional Analyses For completeness, repeat the analyses using an Analysis of Variance (ANOVA)

```
model1<-aov(logMTPM~Fixation, data = Data.RSA.12mon) #ANOVA
summary(model1)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Fixation      1  6.385   6.385   65.41 2.14e-14 ***
## Residuals   267 26.065   0.098
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Assignment Question

Question 1

a. Compare the conclusions from the t-test and Mann Whitney U test.

b. Of these two test, which test would you choose to use?

c. For the tests using logMPTM (t-test, ANOVA), how do the conlusions compare?

d. What is the relationship between these 2 tests?

e. Search for how to implement a Bayesian t-test in R and implement it. Provide an interpretation of the result of the test

a. Compare the conclusions from the t-test and Mann Whitney U test.

- The t-test rejected the null hypothesis with p=2.144e-14
- The Mann Whitney U test also rejects the null hypothesis, with p=1.8e-12
- Both tests reject the null hypothesis and accept the alternate hypothesis that there is a difference between the MPTM means depending on the fixture type. Their p-values vary slightly but both are very small and statistically significant (<0.001)

b. Of these two test, which test would you choose to use?

- I would use the Mann-Whitney U test as it is designed for metrics with non-parametric distributions such as MTPM. Using the t-test required first applying a logarithmic transformation which complicates the measurement as I would no longer be directly observing the

MTPM. This could potentially make measures and differences appear deceivingly small, especially when looking at graphs because the overall distribution would be compressed.

c. For the tests using logMPTM (t-test, ANOVA), how do the conclusions compare?

- The t-test concludes there is a difference between the 2 fixtures based on log(MPTM) (p=6.77e-14)
- The ANOVA test agrees there is a difference between the 2 fixtures based on log(MPTM) (p<0.001)

d. What is the relationship between these 2 tests?

- Based on the lecture notes, the t-test is a special case of ANOVA where there are only 2 groups. Since there are only 2 fixture types, then the two tests are equivalent in this analysis.

e. Search for how to implement a Bayesian t-test in R and implement it. Provide an interpretation of the result of the test

```
# install.packages("BayesFactor")
library(BayesFactor)
```

```
## Loading required package: coda
```

```
## Loading required package: Matrix
```

```
## ************
## Welcome to BayesFactor 0.9.12-4.7. If you have questions, please contact Richard Morey (richarddmorey@gmail.com).
##
## Type BFManual() to open the manual.
## ************
```

```
bayes_t_test <- ttestBF(formula = logMTPM ~ Fixation, data = Data.RSA.12mon)
bayes_t_test
```

```
## Bayes factor analysis
## --------------
## [1] Alt., r=0.707 : 267925807951 ±0%
##
## Against denominator:
##   Null, mu1-mu2 = 0
## ---
## Bayes factor type: BFindepSample, JZS
```

- The Bayes Factor (BF) is 2.67e11% (>100%) which is substantially in favor of the alternate hypothesis that the log(MTPM) mean values vary between the 2 fixation types

Assignment Question

Question 2

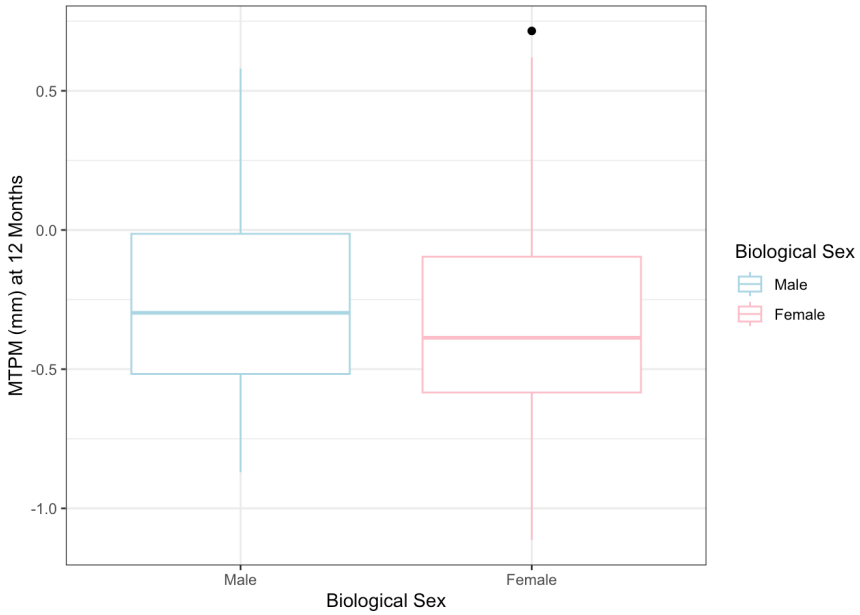
a. Is there a difference in MTPM between males and females at 1 year (12 months) follow-up? Plot log MTPM by sex and support your conclusion statisticallly.

b. BONUS: Is there a difference in MTPM between prosthesis type at 1 year (12 months) follow-up? Plot log MTPM at 12 months by prosthesis and support your conclusion statisticallly.

a. Is there a difference in MTPM between males and females at 1 year (12 months) follow-up? Plot log MTPM by sex and support your conclusion statistically

```
#Box plot of biological sex versus MPTM at 1 year (12 month) follow-up post-TKA
Data.RSA$Sex <- factor(Data.RSA$Sex, levels = c(1, 2), labels = c("Male", "Female"))

ggplot(data=Data.RSA[which(Data.RSA$Months==12), ],
       aes(x=factor(Sex), y=logMTPM, color=factor(Sex))) +
  theme_bw() +
  geom_boxplot(outlier.colour="black", outlier.shape=16, outlier.size=2, notch=FALSE) +
  labs(x="Biological Sex", y="MTPM (mm) at 12 Months", color="Biological Sex") +
  scale_color_manual(values = c('lightblue', 'pink'))
```



- The Log(MPTM) v Biological Sex boxplots appear to be normally distributed so compare the 2 groups via a t-test:
  - $H_0: \mu_M = \mu_F$
  - $H_1: \mu_M \neq \mu_F$

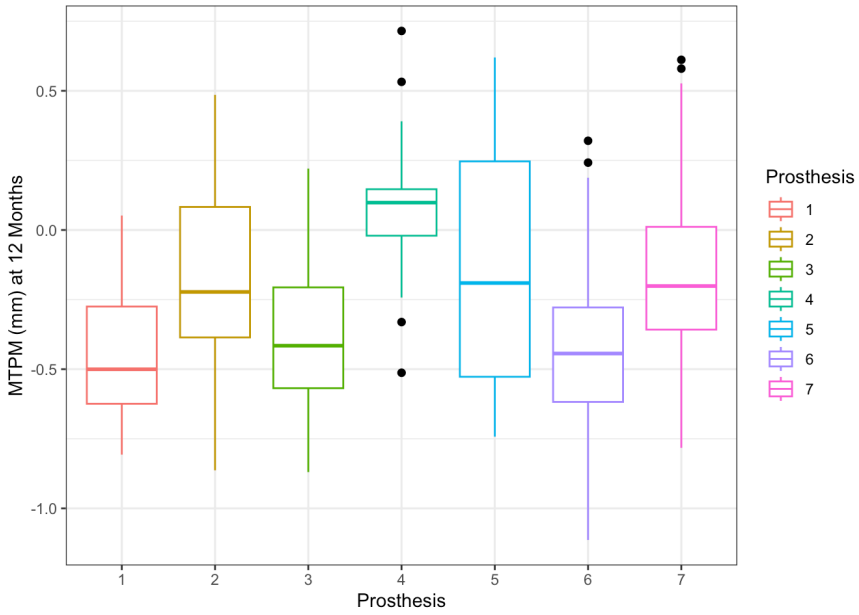
```
# t-test for biological sex v log(MPTM)
t.test(logMTPM~Sex, data = Data.RSA.12mon)
```

```
##
##  Welch Two Sample t-test
##
## data:  logMTPM by Sex
## t = 1.2423, df = 253.93, p-value = 0.2153
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  -0.0304596  0.1345455
## sample estimates:
## mean in group 1 mean in group 2
##       -0.263897       -0.315940
```

- Fail to reject the null hypothesis ($p = 0.2153 > 0.05$)
- Biological sex does not have a significant impact on Log(MPTM) at the 12-month followup mark.

b. BONUS: Is there a difference in MTPM between prosthesis type at 1 year (12 months) follow-up? Plot log MTPM at 12 months by prosthesis and support your conclusion statistically

```
#Box plot of prosthesis type versus MPTM at 1 year (12 month) follow-up post-TKA
ggplot(data=Data.RSA[which(Data.RSA$Months==12), ],
       aes(x=factor(Prosthesis), y=logMTPM, color=factor(Prosthesis))) +
  theme_bw() +
  geom_boxplot(outlier.colour="black", outlier.shape=16, outlier.size=2, notch=FALSE) +
  labs(x="Prosthesis", y="MTPM (mm) at 12 Months", color="Prosthesis")
```



- I would have liked to directly plot the p-values and statistical comparisons on the graph but it needs an additional library… instead I'll use ANOVA since the data appears normally distributed and there are multiple groups to compare

```
aov_prosthesis <- aov(logMTPM~factor(Prosthesis), data = Data.RSA.12mon)
summary(aov_prosthesis)
```

```
##                     Df Sum Sq Mean Sq F value   Pr(>F)
## factor(Prosthesis)   6  7.315  1.2192   12.71 1.38e-12 ***
## Residuals          262 25.135  0.0959
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The p-value of 1.38e-12<0.001 indicates strong support for the alternate hypothesis
- Reject the null hypothesis and accept that alternate hypothesis that prosthesis type affects log(MTPM) at the 12-month follow up mark.

```
TukeyHSD(aov_prosthesis)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = logMTPM ~ factor(Prosthesis), data = Data.RSA.12mon)
##
## $`factor(Prosthesis)`
##             diff          lwr         upr     p adj
## 2-1  0.274678703  0.097770279  0.45158713 0.0001252
## 3-1  0.045777695 -0.160596144  0.25215153 0.9945966
## 4-1  0.525408473  0.272069153  0.77874779 0.0000001
## 5-1  0.308862216  0.055522895  0.56220154 0.0063764
## 6-1 -0.005818366 -0.180875609  0.16923888 0.9999999
## 7-1  0.285057389  0.101736253  0.46837853 0.0001215
## 3-2 -0.228901008 -0.441443435 -0.01635858 0.0255694
## 4-2  0.250729770 -0.007659342  0.50911888 0.0636984
## 5-2  0.034183513 -0.224205600  0.29257263 0.9997088
## 6-2 -0.280497069 -0.462785742 -0.09820840 0.0001501
## 7-2  0.010378686 -0.179860009  0.20061738 0.9999984
## 4-3  0.479630778  0.200241257  0.75902030 0.0000134
## 5-3  0.263084521 -0.016305000  0.54247404 0.0798279
## 6-3 -0.051596061 -0.262600163  0.15940804 0.9908616
## 7-3  0.239279694  0.021370695  0.45718869 0.0210575
## 5-4 -0.216546258 -0.532218801  0.09912629 0.3931010
## 6-4 -0.531226839 -0.788352064 -0.27410161 0.0000001
## 7-4 -0.240351084 -0.503172283  0.02247011 0.0980010
## 6-5 -0.314680582 -0.571805807 -0.05755536 0.0060727
## 7-5 -0.023804827 -0.286626025  0.23901637 0.9999684
## 7-6  0.290875755  0.102357297  0.47939421 0.0001421
```

- Analyze pair-wise prosthesis comparisons by checking the "diff" value, and confirm statistical significance using the "p adj" value (p adj < 0.05). There are 11 statistically significant comparisons

Investigate multi-variable associations

Linear Regression

So far, we've only addressed relationships between features and MPTM that are univariate. The next step would be to consider what features in combination contribute to MPTM. We can assess this using multi-variable regression approaches.

Remember $y = mx + b$ or $E(Y) = \beta_0 + \beta_1 X$?

This is a linear model equation, where:

- $E(Y)$ = response variable, the expected value of Y (such as MPTM or $\log_{10}(MPTM)$)
- β_0 = the intercept
- β_1 = a coefficient (slope) associated with a feature
- X = the independent variable, value of the feature

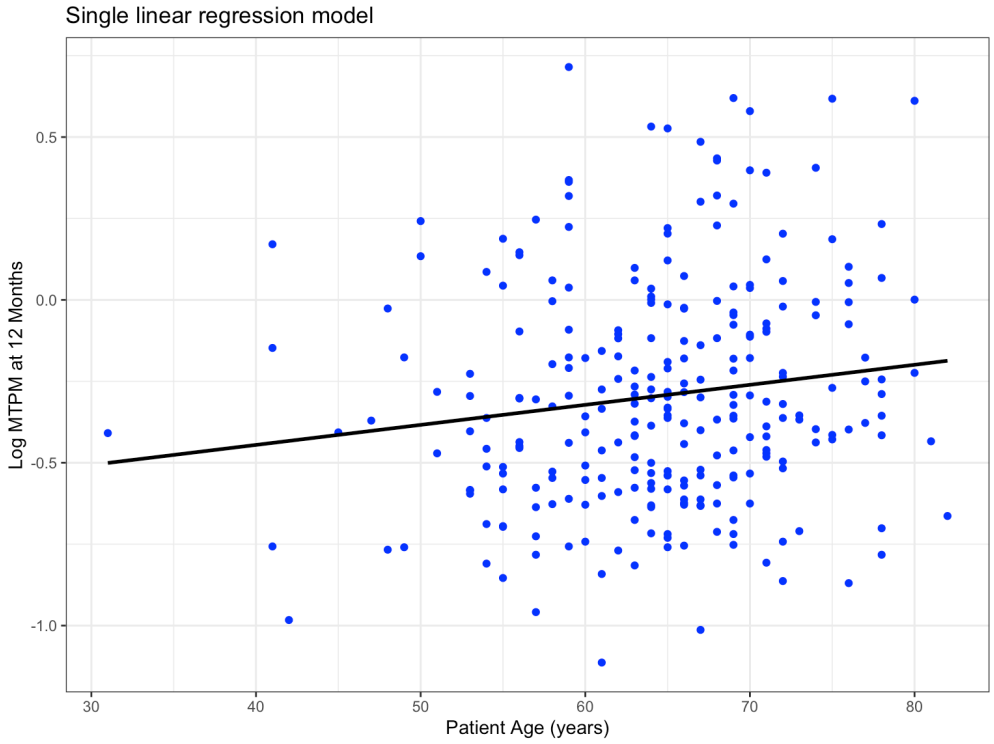
Let's consider one variable in a linear model equation to start. Setting patient age to x we get:

$$\log_{10}(MPTM) = \beta_0 + \beta_1 Age$$

Because there are only two features (MTPM and Age), we can plot this relationship in two dimensions:

```
#Plot the relationship
  ggplot(data=Data.RSA[which(Data.RSA$Months==12), ], aes(x=Age, y=logMTPM)) +
    theme_bw() +
    geom_point(color='blue') +
    geom_smooth(method='lm',se=FALSE, color='black') +
    labs(x="Patient Age (years)", y="Log MTPM at 12 Months", title = "Single linear regression model")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Linear models (lm) compute a regression line (black) by solving for the intercept, β_0, and slope β_1 that results in the least residual sum of squares (RSS) error between the regression line, which represents the expected value of MTPM, or $E(Y_i)$, and the observed value in the dataset Y_i.

$$RSS = \sum_{i=1}^{n} (y_i - E(Y_i))^2$$

We can use R and the `lm` function to solve our linear regression model.

```
#Compute linear model with log MTPM and age
  linearmodel1 <-lm(formula = logMTPM ~ Age, data=Data.RSA[which(Data.RSA$Months==12), ], na.action=na.omit)
  summary(linearmodel1)
```

```
##
## Call:
## lm(formula = logMTPM ~ Age, data = Data.RSA[which(Data.RSA$Months ==
##     12), ], na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79746 -0.24766 -0.03924  0.21677  1.04344
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.691309   0.170613  -4.052 6.67e-05 ***
## Age          0.006152   0.002627   2.342   0.0199 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3451 on 267 degrees of freedom
## Multiple R-squared:  0.02012,    Adjusted R-squared:  0.01645
## F-statistic: 5.483 on 1 and 267 DF,  p-value: 0.01994
```

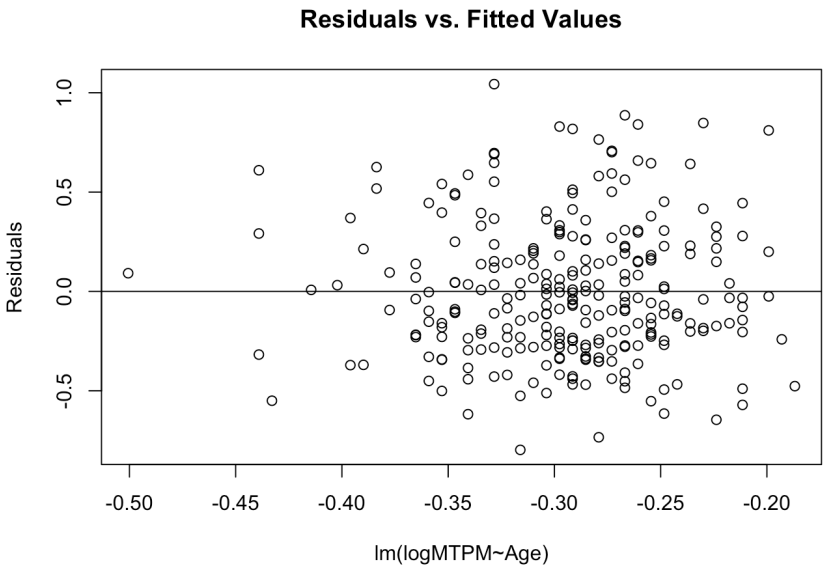
The computed linear model is as follows:

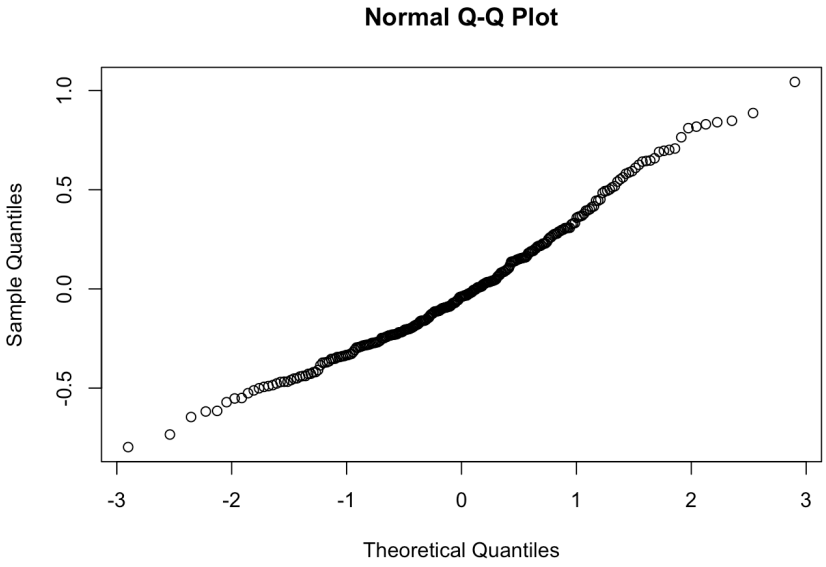
$$\log_{10}(MPTM) = -0.691 + 0.006Age$$

The model is significant (p=0.020)

As the coefficient (slope), β_1, for feature Age is positive, older age is associated with a greater MTPM. We can confirm this from the plot above. Error can be incorporated by the standard error (typically represented using confidence intervals), around the estimates for β_0 and β_1.

Note: It's also important to always check the "residuals" of any linear model. The residuals of our model are already available in our linear model under `linearmodel$residuals` .

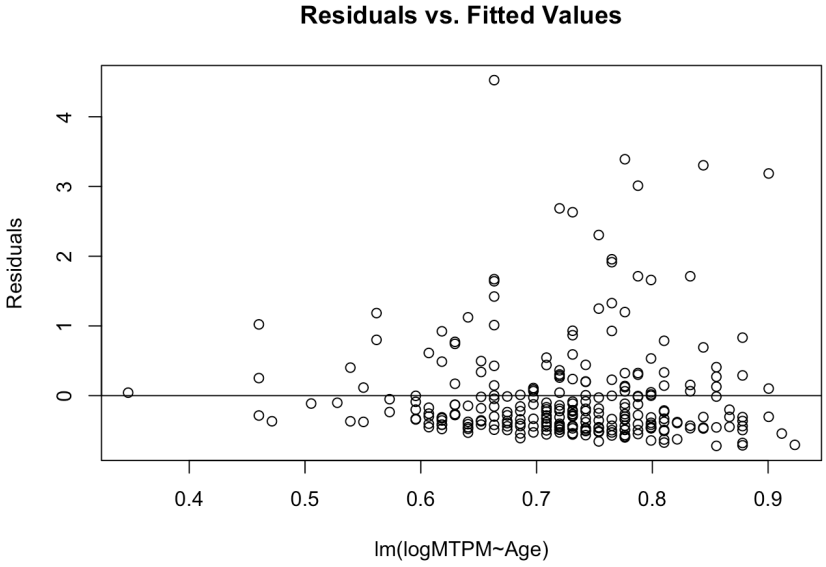


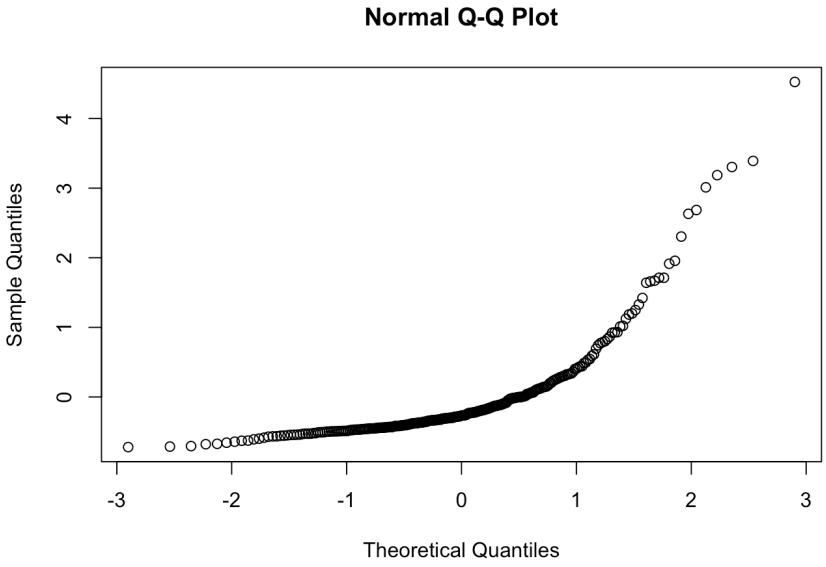


Our residuals are fairly evenly distributed above and below the zero line, and the Q-Q plot follows an approximate straight line, so we can say this model is valid.

In contrast, we can check the residuals if we didn't perform a $\log_{10}$ transformation on MTPM, below. As you can see, there is clustering of the residuals below the zero line, and the Q-Q plot is far from straight. Like t-tests, linear models require a normally distributed response variable (y), and the assumptions of the model are violated when this requirement is not met. This poor model quality is reflected in the residual and Q-Q plots.

```
##
## Call:
## lm(formula = MTPM ~ Age, data = Data.RSA[which(Data.RSA$Months ==
##     12), ], na.action = na.omit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7203 -0.4397 -0.2681  0.1155  4.5256
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.002799   0.376281  -0.007   0.9941
## Age          0.011291   0.005794   1.949   0.0524 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7611 on 267 degrees of freedom
## Multiple R-squared:  0.01402,    Adjusted R-squared:  0.01033
## F-statistic: 3.798 on 1 and 267 DF,  p-value: 0.05237
```





Assignment Question

Question 3

a. Compute the linear model, and present equation for log MTPM and fixation. Discuss if the model is significant and which fixation method results in a greater MTPM. You do not need to plot the residuals and Q-Q plot as part of this question

b. Compare the t-statistic results in your model to those previously calculated using Student's t-tests.

a. Compute the linear model, and present equation for log MTPM and fixation. Discuss if the model is significant and which fixation method results in a greater MTPM. You do not need to plot the residuals and Q-Q plot as part of this question

```
#Compute linear model with log MTPM and fixation type
  linearmodel1 <-lm(formula = logMTPM ~ Fix, data=Data.RSA[which(Data.RSA$Months==12), ], na.action=na.omit)
  summary(linearmodel1)
```

```
##
## Call:
## lm(formula = logMTPM ~ Fix, data = Data.RSA[which(Data.RSA$Months ==
##     12), ], na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73013 -0.22143 -0.01305  0.19752  0.84824
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.75014    0.05943 -12.622  < 2e-16 ***
## Fix          0.30849    0.03814   8.087 2.14e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3124 on 267 degrees of freedom
## Multiple R-squared:  0.1968, Adjusted R-squared:  0.1938
## F-statistic: 65.41 on 1 and 267 DF,  p-value: 2.144e-14
```

- The equation is $\log_{10}(MPTM) \approx -0.750 + 0.308 \times \text{Fixation}$ where cemented=1 and uncemented=2
- The model is significant with p=2.14e-14 < 0.05
- Uncemented fixtures have higher Log(MTPM) as indicated by the positive slope from cemented to uncemented fixtures. However, the model is incorrect as it models categorical explanatory variables (ie fixture type) using numeric values. That is, each fixture type is given an ID (1 or 2) which contextually shouldn't impact the MPTM, but ultimately do change the modeled MPTM value. Visually, the graph would essentially be the previously modeled box plots (except with the raw datapoints) and a line of best fit drawn between the mean of the 2 groups.

b. Compare the t-statistic results in your model to those previously calculated using Student's t-tests.

- From the Student's t-tests: $t = -7.9756,\ p = 6.777 \times 10^{-14}$. This means that mean log(MTPM) values are significantly ***different*** between fixture types.
- From the model: $t_{\beta_0} = -12.622$ and $t_{\beta_1} = 8.087$ and both coefficients are statistically significant according to their p-values. This means that the fixture type significantly ***influences*** log(MTPM)

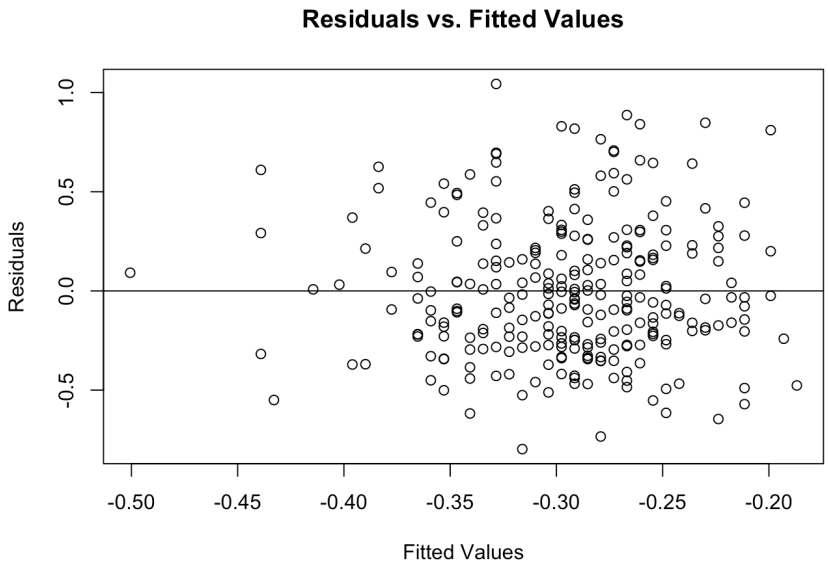
We needed to transform the dependent variable (MTPM) to log(MTPM) to meet the requirements of the linear model (aka general linear model) (normally distributed continuous outcome variable). However, we could also a *generalized* linear model that can accommodate other distributions of dependent variables.

Let's first check how a generalized linear model and also be used to fit a (general) linear model. Note in R "lm" is used for (general) linear model and "glm" is used for generalized linear model. Here we specify "gaussian" for the family, to tell the model the we want to model the dependent variable as normally distributed.

```
model3<-glm(logMTPM~Age, family = gaussian(), data = Data.RSA.12mon)
summary(model3)
```

```
##
## Call:
## glm(formula = logMTPM ~ Age, family = gaussian(), data = Data.RSA.12mon)
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.691309   0.170613  -4.052 6.67e-05 ***
## Age          0.006152   0.002627   2.342   0.0199 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1190898)
##
##     Null deviance: 32.450  on 268  degrees of freedom
## Residual deviance: 31.797  on 267  degrees of freedom
## AIC: 194.98
##
## Number of Fisher Scoring iterations: 2
```

```
  plot(model3$fitted.values, model3$residuals, ylab="Residuals", xlab="Fitted Values", main="Residuals vs. Fitted
Values")
  abline(0, 0)
```

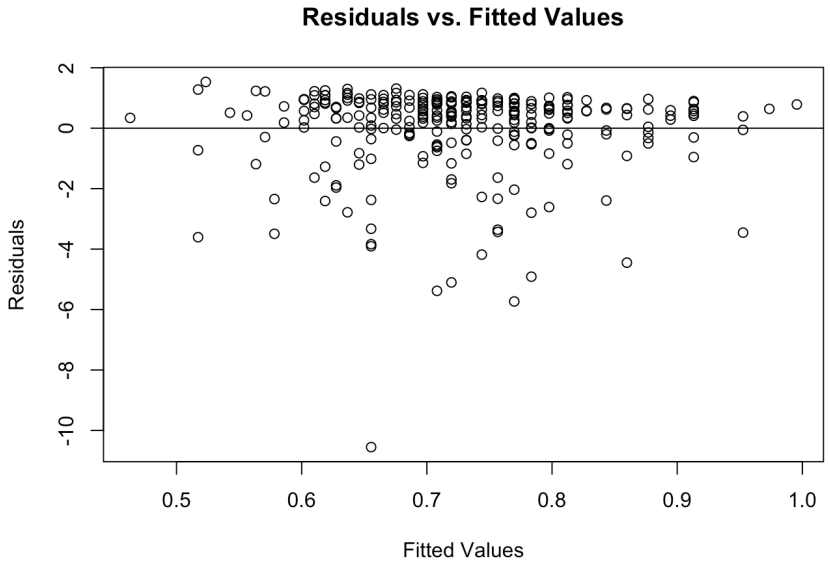


Looking at our original histogram of MTPM, another distribution that might model the dependent variable is a gamma distribution (since it is also bounded at zero). Let's try it and check the residuals plot.

```
model4<-glm(MTPM~Age, family = Gamma(), data = Data.RSA.12mon)
summary(model4)
```

```
##
## Call:
## glm(formula = MTPM ~ Age, family = Gamma(), data = Data.RSA.12mon)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.86224    0.76025   3.765 0.000205 ***
## Age         -0.02266    0.01137  -1.992 0.047380 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 1.0764)
##
##     Null deviance: 192.17  on 268  degrees of freedom
## Residual deviance: 187.81  on 267  degrees of freedom
## AIC: 338.34
##
## Number of Fisher Scoring iterations: 6
```

```
#Residuals plot
  plot(model4$fitted.values, model4$residuals, ylab="Residuals", xlab="Fitted Values", main="Residuals vs. Fitted
Values")
  abline(0, 0)
```



Assignment Question

Question 4

Do you think you have a better model using MTPM and a gamma distribution? Why or why not?

4. Do you think you have a better model using MTPM and a gamma distribution? Why or why not?

- Empirically looking at the residual plots, the GLM with Log(MPTM) fits better than the gamma distribution with MPTM because the residuals appear more evenly distributed above and below the line (though, there may be bias due to the difference in scales). The GLM model also appears to have better homoscedasticity according to the distribution of residuals.
- With consideration for the p-values, the Gamma distribution has $p = 0.0473 \approx 0.05$ and the GLM model has $p = 0.0199$. With a smaller P-value, the GLM model provides stronger evidence that age is correlated with MPTM
- The smaller sum of residuals and AIC also favor the GLM model

Assignment Question

Question 5

In the following multi-variable model, which features have effects on log MTPM? Indicate how you know this.

$$\log_{10}(MPTM) = \beta_0 + \beta_1 Age + \beta_2 BMI + \beta_3 TibialArea + \beta_4 Fixation + \beta_5 Sex + \beta_7 SmokingStatus$$

```
#Compute multi-variable linear model
  linearmodel <-lm(formula = logMTPM ~ Age + BMI + TibiaArea + factor(Fixation) + factor(SexDesc) + factor(Smok
e),
      data=Data.RSA[which(Data.RSA$Months==12), ], na.action=na.omit)
  summary(linearmodel)
```

```
##
## Call:
## lm(formula = logMTPM ~ Age + BMI + TibiaArea + factor(Fixation) +
##     factor(SexDesc) + factor(Smoke), data = Data.RSA[which(Data.RSA$Months ==
##     12), ], na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80345 -0.19619 -0.01959  0.19662  0.86943
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -0.866084   0.306828  -2.823  0.00513 **
## Age                          0.004704   0.002713   1.733  0.08421 .
## BMI                          0.005604   0.003573   1.568  0.11801
## TibiaArea                   -0.002343   0.007055  -0.332  0.74009
## factor(Fixation)Uncemented   0.298893   0.039032   7.658 3.85e-13 ***
## factor(SexDesc)Male          0.016702   0.058432   0.286  0.77523
## factor(Smoke)1               0.018162   0.062253   0.292  0.77072
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3116 on 257 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.2031, Adjusted R-squared:  0.1845
## F-statistic: 10.92 on 6 and 257 DF,  p-value: 7.916e-11
```

5. In the following multi-variable model, which features have effects on log MTPM? Indicate how you know this

- By observing the **Pr(>|t|)** values, the following explanatory variables are significantly correlated to Log(MPTM): age [moderate] (p=0.08), BMI [very weak] (p=0.12), and fixation [strong] (p<0.01)
- Putting aside statistical significance, all of the features technically affect log(MTPM) in the proposed model, but the inclusion of redundant variables weakens the strength of the model and this shows through the noticeable difference between the adjusted r^2 compared to the multiple r^2 value. The 'adjustment' is made to discount for the number of explanatory variables, as otherwise adding redundant variables would likely cause r^2 to increase regardless of the number or explanatory strength of the variables.

Model Reduction

Although the model above is valid, we typically only want use models that are the most frugal (i.e., we don't want added complexity from non-significant features).

In statistics this is referred to as the most "parsimonious" model. In computer science, this is referred to as the "Occam's razor" principal.

Backwards elimination is one approach to linear regression model reduction, where we remove features one-by-one that are the least significant, and therefore do not significantly contribute to our response (MTPM), until we have our final frugal model.

Assignment Question

Question 6
Using the model above and a backwards elimination step-wise approach, remove features that are not significant to in the model, and present your final model equation.

6. Using the model above and a backwards elimination step-wise approach, remove features that are not significant to in the model, and present your final model equation

- Iteratively remove variables starting from the highest p value until all variables are statistically significant at $p < 0.05$ while observing the general model strength with respect to sum of residuals and adjusted r^2:

```
# Remove variables with p>0.20: TibiaArea, SexDesc, Smoke
linearmodel2 <- lm(formula = logMTPM ~ Age + BMI + factor(Fixation),data=Data.RSA[which(Data.RSA$Months==12), ],
na.action=na.omit)

summary(linearmodel2)
```

```
##
## Call:
## lm(formula = logMTPM ~ Age + BMI + factor(Fixation), data = Data.RSA[which(Data.RSA$Months ==
##     12), ], na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80645 -0.19677 -0.01939  0.19720  0.86802
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -0.968426   0.232028  -4.174 4.07e-05 ***
## Age                         0.005324   0.002558   2.082   0.0383 *
## BMI                         0.005744   0.003499   1.641   0.1019
## factor(Fixation)Uncemented  0.305492   0.038321   7.972 4.68e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3106 on 265 degrees of freedom
## Multiple R-squared:  0.2123, Adjusted R-squared:  0.2034
## F-statistic: 23.81 on 3 and 265 DF,  p-value: 1.121e-13
```

- Multiple R-squared: 0.2031 -> 0.2123
- Adjusted R-squared: 0.1845 -> 0.2031
- Residual standard error: 0.3116 –> 0.3106

```
# Remove remaining variables with p>0.10: BMI
linearmodel3 <- lm(formula = logMTPM ~ Age + factor(Fixation), data=Data.RSA[which(Data.RSA$Months==12), ], na.ac
tion=na.omit)

summary(linearmodel3)
```

```
##
## Call:
## lm(formula = logMTPM ~ Age + factor(Fixation), data = Data.RSA[which(Data.RSA$Months ==
##     12), ], na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.75483 -0.20434 -0.01039  0.19500  0.87293
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -0.682880   0.154035  -4.433 1.36e-05 ***
## Age                         0.003800   0.002391   1.589    0.113
## factor(Fixation)Uncemented  0.300853   0.038339   7.847 1.04e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3116 on 266 degrees of freedom
## Multiple R-squared:  0.2043, Adjusted R-squared:  0.1983
## F-statistic: 34.15 on 2 and 266 DF,  p-value: 6.281e-14
```

- Multiple R-squared: 0.2123 -> 0.2043
- Adjusted R-squared: 0.2031 -> 0.1983
- Residual standard error: 0.3106 -> 0.3116

```
# Remove remaining variables with p>0.05:
linearmodel4 <- lm(formula = logMTPM ~ factor(Fixation), data=Data.RSA[which(Data.RSA$Months==12), ], na.action=n
a.omit)

summary(linearmodel4)
```

```
##
## Call:
## lm(formula = logMTPM ~ factor(Fixation), data = Data.RSA[which(Data.RSA$Months ==
##     12), ], na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73013 -0.22143 -0.01305  0.19752  0.84824
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -0.44164    0.02631 -16.785  < 2e-16 ***
## factor(Fixation)Uncemented  0.30849    0.03814   8.087 2.14e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3124 on 267 degrees of freedom
## Multiple R-squared:  0.1968, Adjusted R-squared:  0.1938
## F-statistic: 65.41 on 1 and 267 DF,  p-value: 2.144e-14
```

- Multiple R-squared: 0.2043 -> 0.1968
- Adjusted R-squared: 0.2031 -> 0.1938
- Residual standard error: 0.3106 -> 0.3124
- Removal of explanatory variables on the basis of p-values (p<0.05) leaves only fixation type remaining. Ideally I would have incorporated additional variables in the initial model, such as prosthesis type, to have a higher r^2 value and stronger model.
- Final model: $\log_{10}(MPTM) = \beta_0 + \beta_1 Fixation$

Longitudinal Data Analysis

So far we have ignored the repeated measures in this dataset and only examine implant migration (MTPM) at one time point (12 months). To make full use of this longitudinal dataset, we can use an approach that accounts for the correlation between repeated measures on the same participants. We'll use generalized estimating equations (GEE) to see if there are differences between fixation when looking at all longitudinal data, and see if other variables are significantly associated with migration. Another way to think of this is that any potential difference of Fixation, if significant, would be there after accounting for differences due to the other variables (controlling for those variables).

To fit a GEE, the formula is similar to previous models, with "id" showing how repeated measurements are labelled, "family" for the distribution of the dependent variable, and "corstr" defining the correlation structure. Here we chose autoregressive ("corstr ="AR-M", Mv = 1") which means that the correlation between measures varies relative to how close the measurements are together.

```
library(gee)

gee_model1 <-gee(logMTPM~Months + Age + BMI + factor(Sex) + factor(Fixation), data = Data.RSA, id = Subject, na.a
ction = na.omit, family = gaussian, corstr = "AR-M", Mv = 1)
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
##               (Intercept)                    Months
##              -1.094860344                0.006757734
##                       Age                       BMI
##               0.006152273                0.004224595
##          factor(Sex)Female factor(Fixation)Uncemented
##              -0.001327955                0.316099008
```

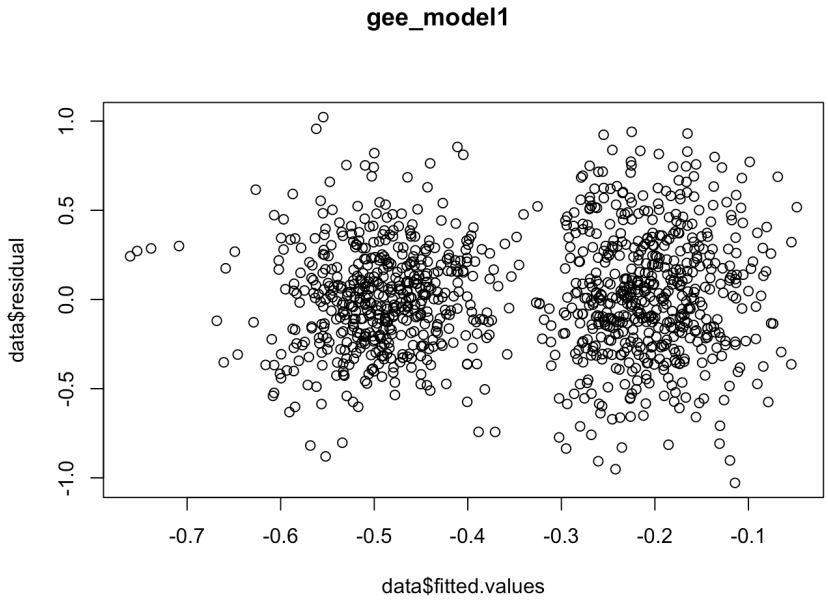
```
summary(gee_model1)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Identity
##  Variance to Mean Relation: Gaussian
##  Correlation Structure:     AR-M , M = 1
##
## Call:
## gee(formula = logMTPM ~ Months + Age + BMI + factor(Sex) + factor(Fixation),
##     id = Subject, data = Data.RSA, na.action = na.omit, family = gaussian,
##     corstr = "AR-M", Mv = 1)
##
## Summary of Residuals:
##         Min          1Q      Median          3Q         Max
## -1.02825150 -0.20097246  0.01209127  0.23238348  1.02239491
##
##
## Coefficients:
##                                Estimate   Naive S.E.    Naive z  Robust S.E.
## (Intercept)                -1.052619080 0.2119476323 -4.9664111 0.2184825080
## Months                      0.004971358 0.0007678968  6.4739928 0.0006106318
## Age                         0.005243994 0.0023523190  2.2292869 0.0024456512
## BMI                         0.005455366 0.0032065552  1.7013167 0.0032808099
## factor(Sex)Female          -0.017385133 0.0367830638 -0.4726396 0.0380496041
## factor(Fixation)Uncemented  0.290120709 0.0355311341  8.1652533 0.0363842754
##                              Robust z
## (Intercept)                -4.8178643
## Months                      8.1413350
## Age                         2.1442116
## BMI                         1.6628107
## factor(Sex)Female          -0.4569071
## factor(Fixation)Uncemented  7.9737938
##
## Estimated Scale Parameter:  0.108218
## Number of Iterations:  3
##
## Working Correlation
##           [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] 1.0000000 0.8424238 0.7096779 0.5978495 0.5036427
## [2,] 0.8424238 1.0000000 0.8424238 0.7096779 0.5978495
## [3,] 0.7096779 0.8424238 1.0000000 0.8424238 0.7096779
## [4,] 0.5978495 0.7096779 0.8424238 1.0000000 0.8424238
## [5,] 0.5036427 0.5978495 0.7096779 0.8424238 1.0000000
```

If you look at the results, you will see that you are given estimates for each independent variable, but not confidence intervals or p-values. To calculate these, use the function below. Plotting the residuals is included in this function as well, because plotting residuals is always a good idea.

You can now look at the CIs and p-values for each dependent variable, and the residuals for the model.

```
gee_model1_results<-geeCI(gee_model1)
```



```
gee_model1_results
```

```
##                            Estimate gee_model1   Naive.S.E.         lwrCI
## (Intercept)                       -1.052619080 0.2119476323 -1.4680364397
## Months                             0.004971358 0.0007678968  0.0034662805
## Age                                0.005243994 0.0023523190  0.0006334486
## BMI                                0.005455366 0.0032065552 -0.0008294824
## factor(Sex)Female                 -0.017385133 0.0367830638 -0.0894799385
## factor(Fixation)Uncemented         0.290120709 0.0355311341  0.2204796859
##                                       uprCI p.value    Naive.z  Robust.S.E.
## (Intercept)                     -0.637201721  0.0000 -4.9664111 0.2184825080
## Months                           0.006476436  0.0000  6.4739928 0.0006106318
## Age                              0.009854539  0.0320  2.2292869 0.0024456512
## BMI                              0.011740214  0.0964  1.7013167 0.0032808099
## factor(Sex)Female                0.054709672  0.6477 -0.4726396 0.0380496041
## factor(Fixation)Uncemented       0.359761732  0.0000  8.1652533 0.0363842754
##                                   Robust.z   lwrCIRobust  uprCIRobust
## (Intercept)                     -4.8178643 -1.4808447960 -0.624393365
## Months                           8.1413350  0.0037745198  0.006168197
## Age                              2.1442116  0.0004505174  0.010037470
## BMI                              1.6628107 -0.0009750215  0.011885753
## factor(Sex)Female               -0.4569071 -0.0919623574  0.057192091
## factor(Fixation)Uncemented       7.9737938  0.2188075291  0.361433889
```

The residuals don't look great as they are not evenly distributed along the x-axis.

Let's try another model where we don't transform the MTPM variable and instead use a model for a Gamma distribution (instead of a guassian, aka normal distribution)

```
gee_model2 <-gee(MTPM~Months + Age + BMI + factor(Sex) + factor(Fixation), data = Data.RSA, id = Subject, na.acti
on = na.omit, family = Gamma(link=identity), corstr = "AR-M", Mv = 1)
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
##               (Intercept)                    Months
##              -0.002310280                 0.008577139
##                       Age                       BMI
##               0.002888694                 0.004314287
##         factor(Sex)Female factor(Fixation)Uncemented
##              -0.004232645                 0.528412673
```

```
summary(gee_model2)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Identity
##  Variance to Mean Relation: Gamma
##  Correlation Structure:     AR-M , M = 1
##
## Call:
## gee(formula = MTPM ~ Months + Age + BMI + factor(Sex) + factor(Fixation),
##     id = Subject, data = Data.RSA, na.action = na.omit, family = Gamma(link = identity),
##     corstr = "AR-M", Mv = 1)
##
## Summary of Residuals:
##        Min         1Q     Median         3Q        Max
## -0.9005287 -0.2969776 -0.1236838  0.1651525  4.8777462
##
##
## Coefficients:
##                               Estimate  Naive S.E.    Naive z  Robust S.E.
## (Intercept)                 0.037992297 0.267814566  0.1418605 0.1880166036
## Months                      0.006761456 0.001090215  6.2019500 0.0009816184
## Age                         0.001588859 0.002944260  0.5396462 0.0024440802
## BMI                         0.006691061 0.004404115  1.5192748 0.0044536070
## factor(Sex)Female          -0.033804982 0.053868081 -0.6275513 0.0422523917
## factor(Fixation)Uncemented  0.457883765 0.068994793  6.6364974 0.0632858135
##                               Robust z
## (Intercept)                 0.2020688
## Months                      6.8880703
## Age                         0.6500845
## BMI                         1.5023913
## factor(Sex)Female          -0.8000726
## factor(Fixation)Uncemented  7.2351723
##
## Estimated Scale Parameter:  0.8628628
## Number of Iterations:  13
##
## Working Correlation
##           [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] 1.0000000 0.8944655 0.8000686 0.7156338 0.6401097
## [2,] 0.8944655 1.0000000 0.8944655 0.8000686 0.7156338
## [3,] 0.8000686 0.8944655 1.0000000 0.8944655 0.8000686
## [4,] 0.7156338 0.8000686 0.8944655 1.0000000 0.8944655
## [5,] 0.6401097 0.7156338 0.8000686 0.8944655 1.0000000
```

```
gee_model2_results<-geeCI(gee_model2)
```


```
gee_model2_results
```

```
##                             Estimate gee_model2  Naive.S.E.        lwrCI
## (Intercept)                        0.037992297 0.267814566 -0.486924253
## Months                             0.006761456 0.001090215  0.004624636
## Age                                0.001588859 0.002944260 -0.004181891
## BMI                                0.006691061 0.004404115 -0.001941004
## factor(Sex)Female                 -0.033804982 0.053868081 -0.139386421
## factor(Fixation)Uncemented         0.457883765 0.068994793  0.322653971
##                                       uprCI p.value    Naive.z  Robust.S.E.
## (Intercept)                     0.562908846  0.8399  0.1418605 0.1880166036
## Months                          0.008898277  0.0000  6.2019500 0.0009816184
## Age                             0.007359608  0.5156  0.5396462 0.0024440802
## BMI                             0.015323126  0.1330  1.5192748 0.0044536070
## factor(Sex)Female               0.071776457  0.4237 -0.6275513 0.0422523917
## factor(Fixation)Uncemented      0.593113560  0.0000  6.6364974 0.0632858135
##                                   Robust.z  lwrCIRobust uprCIRobust
## (Intercept)                      0.2020688 -0.330520247 0.406504840
## Months                           6.8880703  0.004837484 0.008685428
## Age                              0.6500845 -0.003201539 0.006379256
## BMI                              1.5023913 -0.002038009 0.015420130
## factor(Sex)Female               -0.8000726 -0.116619670 0.049009706
## factor(Fixation)Uncemented       7.2351723  0.333843571 0.581923960
```

Residuals look even worse! We'll stick with the transformed logMTPM for now.

The residual plots in both cases seem to have 2 clusters in them. This suggests that we may have 2 underlying distributions that are preventing a good model fit when combined together. This is also supported by the results showing that "Fixation" was highly signficiant and had a large estimate. Thinking back to the original plots that you made as well, it may be that cemented and uncemented implants need different models, so let's look at them separately.

```
#Cemented implants only:
Data.RSA.Cemented <- Data.RSA[which(Data.RSA$Fixation == "Cemented"), ] #only cemented implants

#Uncemented implants only:
Data.RSA.Uncemented <- Data.RSA[which(Data.RSA$Fixation == "Uncemented"), ] #only cemented implants
```

Perform the analysis on cemented implants alone:

```
gee_model_cemented1 <-gee(logMTPM~Months + Age + BMI + factor(Sex), data = Data.RSA.Cemented, id = Subject, na.ac
tion = na.omit, family = gaussian, corstr = "AR-M", Mv = 1)
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

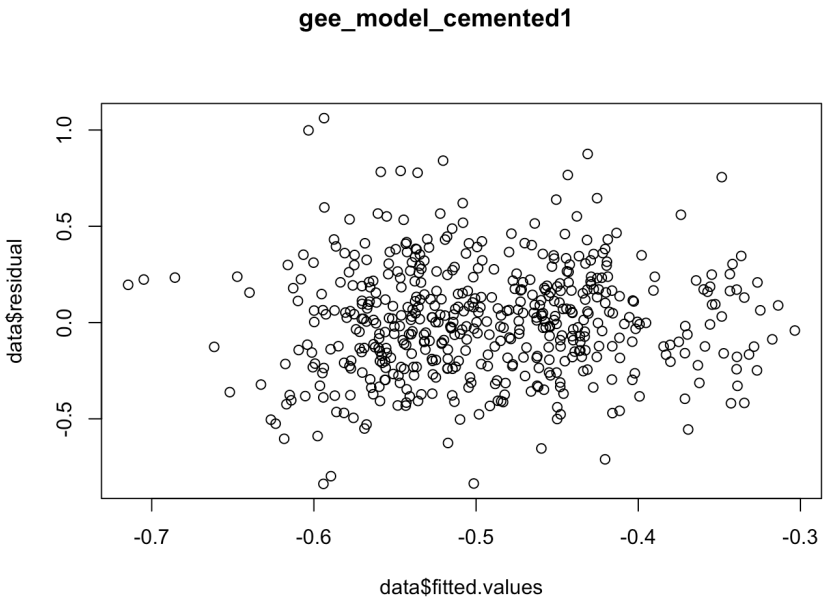
```
## running glm to get initial regression estimate
```

```
##       (Intercept)            Months               Age               BMI
##      -0.794453310       0.008618116       0.003500664       0.001264819
## factor(Sex)Female
##      -0.084976875
```

```
summary(gee_model_cemented1)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Identity
##  Variance to Mean Relation: Gaussian
##  Correlation Structure:     AR-M , M = 1
##
## Call:
## gee(formula = logMTPM ~ Months + Age + BMI + factor(Sex), id = Subject,
##     data = Data.RSA.Cemented, na.action = na.omit, family = gaussian,
##     corstr = "AR-M", Mv = 1)
##
## Summary of Residuals:
##          Min           1Q       Median           3Q          Max
## -0.837657379 -0.183845974  0.006696694  0.183585116  1.061636905
##
##
## Coefficients:
##                        Estimate  Naive S.E.    Naive z  Robust S.E.   Robust z
## (Intercept)       -0.806432922 0.233676702 -3.4510626 0.2410488708 -3.3455163
## Months             0.006407142 0.001073972  5.9658355 0.0009159812  6.9948403
## Age                0.003230634 0.002591279  1.2467334 0.0027325798  1.1822651
## BMI                0.002737074 0.003473512  0.7879844 0.0035167276  0.7783013
## factor(Sex)Female -0.087663269 0.044997521 -1.9481800 0.0439981685 -1.9924300
##
## Estimated Scale Parameter:  0.0827166
## Number of Iterations:  3
##
## Working Correlation
##           [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] 1.0000000 0.7828412 0.6128404 0.4797567 0.3755733
## [2,] 0.7828412 1.0000000 0.7828412 0.6128404 0.4797567
## [3,] 0.6128404 0.7828412 1.0000000 0.7828412 0.6128404
## [4,] 0.4797567 0.6128404 0.7828412 1.0000000 0.7828412
## [5,] 0.3755733 0.4797567 0.6128404 0.7828412 1.0000000
```

```
gee_model_cemented1_results<-geeCI(gee_model_cemented1)
```




```
gee_model_cemented1_results
```

```
##                   Estimate gee_model_cemented1  Naive.S.E.        lwrCI
## (Intercept)                       -0.806432922 0.233676702 -1.264439259
## Months                             0.006407142 0.001073972  0.004302156
## Age                                0.003230634 0.002591279 -0.001848273
## BMI                                0.002737074 0.003473512 -0.004071011
## factor(Sex)Female                 -0.087663269 0.044997521 -0.175858409
##                            uprCI p.value    Naive.z  Robust.S.E.   Robust.z
## (Intercept)       -0.3484265855  0.0008 -3.4510626 0.2410488708 -3.3455163
## Months             0.0085121276  0.0000  5.9658355 0.0009159812  6.9948403
## Age                0.0083095398  0.2371  1.2467334 0.0027325798  1.1822651
## BMI                0.0095451583  0.4364  0.7879844 0.0035167276  0.7783013
## factor(Sex)Female  0.0005318714  0.0463 -1.9481800 0.0439981685 -1.9924300
##                     lwrCIRobust  uprCIRobust
## (Intercept)       -1.278888709 -0.333977136
## Months             0.004611819  0.008202465
## Age               -0.002125223  0.008586490
## BMI               -0.004155712  0.009629860
## factor(Sex)Female -0.173899679 -0.001426859
```

Don't forget you can search "gee" in the help window to get more information

Assignment Question

Question 7

a. Repeat the GEE analysis on the uncemented group and generate the output and residuals plot.

b. Provide your assessment of the residuals plots for the cemented only GEE and uncemented only GEE

c. You can also use alternative correlation structures. Try replacing the whole '*AR-M, Mv = 1'* with 'exchangeable' and then 'independence'

d. Which final model(s) would you choose to look at the associations of age, sex, and BMI on longitudinal migration for the cemented and uncemented groups?

a. Repeat the GEE analysis on the uncemented group and generate the output and residuals plot.

```
gee_model_uncemented1 <-gee(logMTPM~Months + Age + BMI + factor(Sex), data = Data.RSA.Uncemented, id = Subject, n
a.action = na.omit, family = gaussian, corstr = "AR-M", Mv = 1)
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

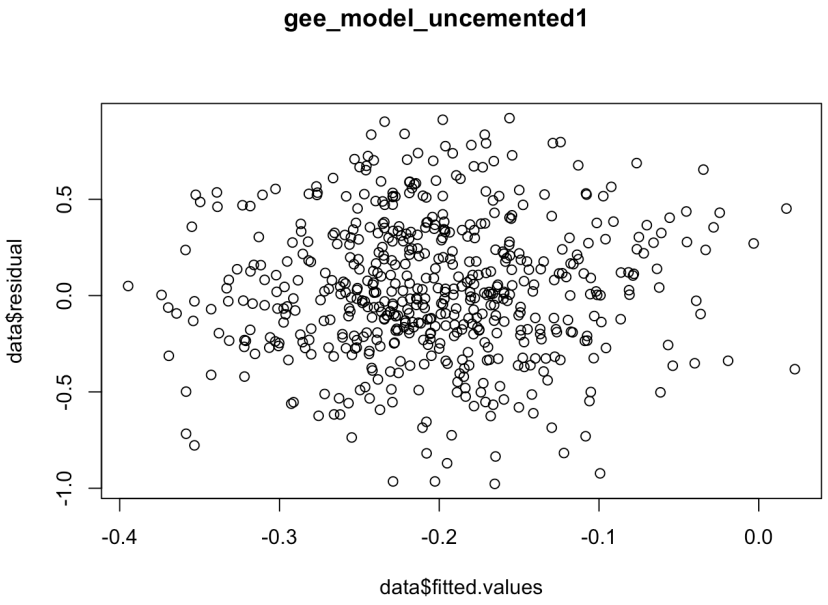
```
## running glm to get initial regression estimate
```

```
##       (Intercept)            Months               Age               BMI
##      -1.130673039       0.004984237       0.008734702       0.009515082
## factor(Sex)Female
##       0.061409712
```

```
summary(gee_model_uncemented1)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Identity
##  Variance to Mean Relation: Gaussian
##  Correlation Structure:     AR-M , M = 1
##
## Call:
## gee(formula = logMTPM ~ Months + Age + BMI + factor(Sex), id = Subject,
##     data = Data.RSA.Uncemented, na.action = na.omit, family = gaussian,
##     corstr = "AR-M", Mv = 1)
##
## Summary of Residuals:
##         Min          1Q      Median          3Q         Max
## -0.97740259 -0.19796465  0.01483376  0.28120074  0.92125377
##
##
## Coefficients:
##                       Estimate  Naive S.E.    Naive z  Robust S.E.   Robust z
## (Intercept)       -1.030509869 0.365646634 -2.8183218 0.3598731107 -2.8635367
## Months             0.003481498 0.001064707  3.2699136 0.0007792358  4.4678367
## Age                0.006953841 0.004108559  1.6925257 0.0041488480  1.6760895
## BMI                0.009865163 0.005687351  1.7345796 0.0058274993  1.6928638
## factor(Sex)Female  0.044242848 0.057084456  0.7750419 0.0599796608  0.7376308
##
## Estimated Scale Parameter:  0.1280133
## Number of Iterations:  3
##
## Working Correlation
##           [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] 1.0000000 0.8803311 0.7749829 0.6822416 0.6005985
## [2,] 0.8803311 1.0000000 0.8803311 0.7749829 0.6822416
## [3,] 0.7749829 0.8803311 1.0000000 0.8803311 0.7749829
## [4,] 0.6822416 0.7749829 0.8803311 1.0000000 0.8803311
## [5,] 0.6005985 0.6822416 0.7749829 0.8803311 1.0000000
```

```
gee_model_uncemented1_results<-geeCI(gee_model_uncemented1)
```




```
gee_model_uncemented1_results
```

```
##                   Estimate gee_model_uncemented1  Naive.S.E.        lwrCI
## (Intercept)                         -1.030509869 0.365646634 -1.747177271
## Months                               0.003481498 0.001064707  0.001394674
## Age                                  0.006953841 0.004108559 -0.001098934
## BMI                                  0.009865163 0.005687351 -0.001282045
## factor(Sex)Female                    0.044242848 0.057084456 -0.067642686
##                           uprCI p.value    Naive.z  Robust.S.E.   Robust.z
## (Intercept)       -0.313842467  0.0042 -2.8183218 0.3598731107 -2.8635367
## Months             0.005568323  0.0000  3.2699136 0.0007792358  4.4678367
## Age                0.015006616  0.0937  1.6925257 0.0041488480  1.6760895
## BMI                0.021012370  0.0905  1.7345796 0.0058274993  1.6928638
## factor(Sex)Female  0.156128381  0.4607  0.7750419 0.0599796608  0.7376308
##                    lwrCIRobust uprCIRobust
## (Intercept)       -1.735861166 -0.32515857
## Months             0.001954196  0.00500880
## Age               -0.001177901  0.01508558
## BMI               -0.001556736  0.02128706
## factor(Sex)Female -0.073317288  0.16180298
```

b. Provide your assessment of the residuals plots for the cemented only GEE and uncemented only GEE

- **Empirical observations:** The GEE models for both fixture types appear to fit well as the residuals seem evenly distributed above and below the y=0 line and there is heteroscedasticity
- **Statistical observation**: The residuals are contained well with the interquartile range being ~0.367 and ~0.479 for cemented and uncemented fixtures, respectively. Coupled with the small median residual values, this implies the models can accurately fit the majority of the dataset.
    - There may be overfitting due to the presence of explanatory variables with low z-scores. A z-score of $|z| = 1.96$ corresponds to the 95% confidence level, but only 'months' and 'BMI' are statistically significant for uncemented fixtures, and only 'months' for cemented fixtures. This inconsistency also suggests that separate fine-tuning may be required for each model in terms of variable selection.

c. You can also use alternative correlation structures. Try replacing the whole '*AR-M, Mv = 1'* with 'exchangeable' and then 'independence'

```
gee_model_uncemented_exc <-gee(logMTPM~Months + Age + BMI + factor(Sex), data = Data.RSA.Uncemented, id = Subjec
t, na.action = na.omit, family = gaussian, corstr = "exchangeable")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

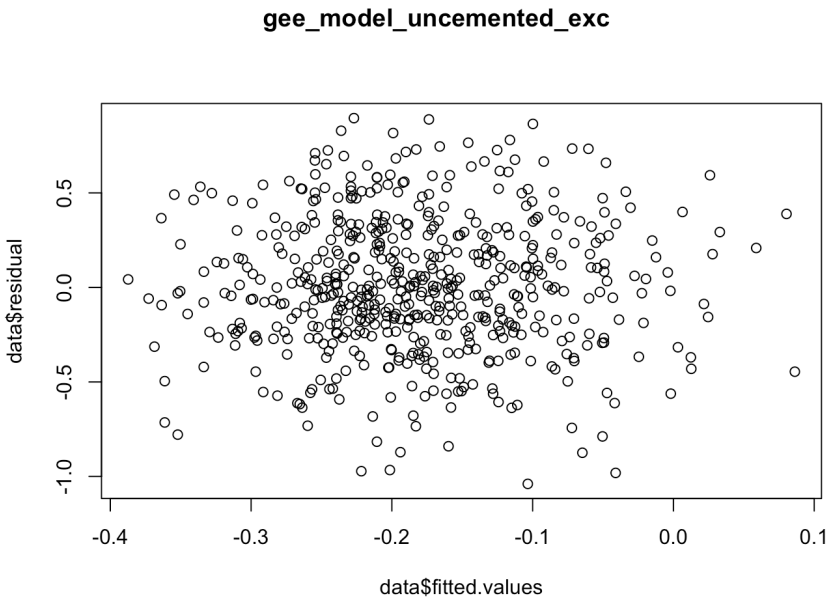
```
## running glm to get initial regression estimate
```

```
##       (Intercept)            Months               Age               BMI
##      -1.130673039       0.004984237       0.008734702       0.009515082
## factor(Sex)Female
##       0.061409712
```

```
summary(gee_model_uncemented_exc)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Identity
##  Variance to Mean Relation: Gaussian
##  Correlation Structure:     Exchangeable
##
## Call:
## gee(formula = logMTPM ~ Months + Age + BMI + factor(Sex), id = Subject,
##     data = Data.RSA.Uncemented, na.action = na.omit, family = gaussian,
##     corstr = "exchangeable")
##
## Summary of Residuals:
##          Min           1Q       Median           3Q          Max
## -1.039222190 -0.231289477 -0.006045717  0.266436254  0.895619887
##
##
## Coefficients:
##                       Estimate   Naive S.E.    Naive z  Robust S.E.   Robust z
## (Intercept)       -1.063360571 0.3755314662 -2.8316151 0.3494523817 -3.0429341
## Months             0.006155820 0.0007281958  8.4535228 0.0008821803  6.9779616
## Age                0.007212564 0.0042211222  1.7086841 0.0040005233  1.8029052
## BMI                0.010278060 0.0058353396  1.7613474 0.0058491176  1.7571984
## factor(Sex)Female  0.039785397 0.0588160998  0.6764372 0.0587885770  0.6767539
##
## Estimated Scale Parameter:  0.127134
## Number of Iterations:  3
##
## Working Correlation
##           [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] 1.0000000 0.8535693 0.8535693 0.8535693 0.8535693
## [2,] 0.8535693 1.0000000 0.8535693 0.8535693 0.8535693
## [3,] 0.8535693 0.8535693 1.0000000 0.8535693 0.8535693
## [4,] 0.8535693 0.8535693 0.8535693 1.0000000 0.8535693
## [5,] 0.8535693 0.8535693 0.8535693 0.8535693 1.0000000
```

```
gee_model_uncemented_exc_results<-geeCI(gee_model_uncemented_exc)
```




```
gee_model_uncemented_exc_results
```

```
##                   Estimate gee_model_uncemented_exc   Naive.S.E.        lwrCI
## (Intercept)                              -1.063360571 0.3755314662 -1.799402244
## Months                                    0.006155820 0.0007281958  0.004728556
## Age                                       0.007212564 0.0042211222 -0.001060835
## BMI                                       0.010278060 0.0058353396 -0.001159205
## factor(Sex)Female                         0.039785397 0.0588160998 -0.075494158
##                           uprCI p.value    Naive.z  Robust.S.E.   Robust.z
## (Intercept)       -0.327318897  0.0023 -2.8316151 0.3494523817 -3.0429341
## Months             0.007583084  0.0000  8.4535228 0.0008821803  6.9779616
## Age                0.015485964  0.0714  1.7086841 0.0040005233  1.8029052
## BMI                0.021715326  0.0789  1.7613474 0.0058491176  1.7571984
## factor(Sex)Female  0.155064953  0.4986  0.6764372 0.0587885770  0.6767539
##                      lwrCIRobust  uprCIRobust
## (Intercept)       -1.7482872388 -0.378433902
## Months             0.0044267467  0.007884893
## Age               -0.0006284615  0.015053590
## BMI               -0.0011862101  0.021742331
## factor(Sex)Female -0.0754402137  0.155011008
```

```
gee_model_cemented_exc <-gee(logMTPM~Months + Age + BMI + factor(Sex), data = Data.RSA.Cemented, id = Subject, n
a.action = na.omit, family = gaussian, corstr = "exchangeable")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

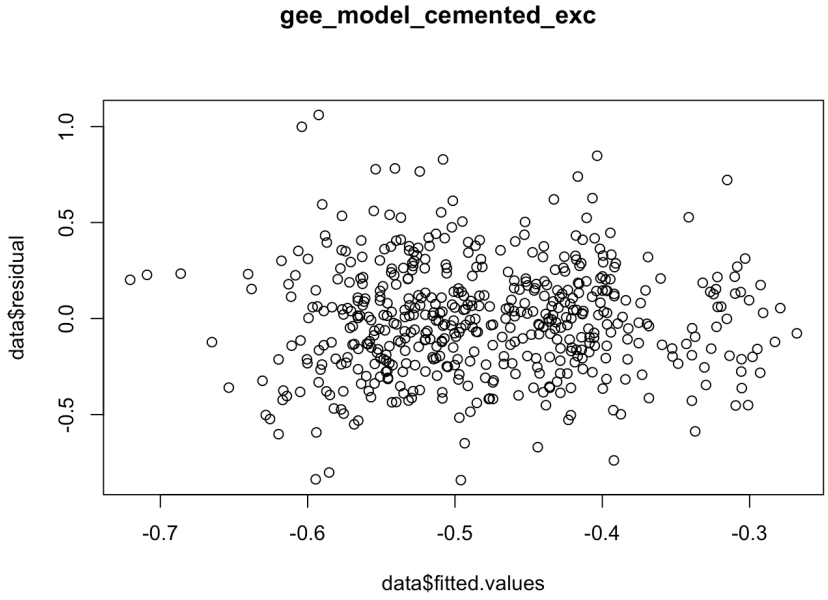
```
## running glm to get initial regression estimate
```

```
##       (Intercept)            Months               Age               BMI
##      -0.794453310       0.008618116       0.003500664       0.001264819
## factor(Sex)Female
##      -0.084976875
```

```
summary(gee_model_cemented_exc)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Identity
##  Variance to Mean Relation: Gaussian
##  Correlation Structure:     Exchangeable
##
## Call:
## gee(formula = logMTPM ~ Months + Age + BMI + factor(Sex), id = Subject,
##     data = Data.RSA.Cemented, na.action = na.omit, family = gaussian,
##     corstr = "exchangeable")
##
## Summary of Residuals:
##          Min           1Q       Median           3Q          Max
## -0.841244065 -0.206190958 -0.001924565  0.174477763  1.060423945
##
##
## Coefficients:
##                      Estimate   Naive S.E.    Naive z  Robust S.E.   Robust z
## (Intercept)       -0.816364392 0.2447705587 -3.3352230 0.2375258086 -3.4369503
## Months             0.007625283 0.0008238053  9.2561712 0.0009600538  7.9425580
## Age                0.003365700 0.0026998318  1.2466334 0.0027079439  1.2428989
## BMI                0.002889674 0.0036381769  0.7942643 0.0035717735  0.8090306
## factor(Sex)Female -0.093595239 0.0463260366 -2.0203593 0.0425056440 -2.2019485
##
## Estimated Scale Parameter:  0.0824316
## Number of Iterations:  3
##
## Working Correlation
##           [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] 1.0000000 0.7492096 0.7492096 0.7492096 0.7492096
## [2,] 0.7492096 1.0000000 0.7492096 0.7492096 0.7492096
## [3,] 0.7492096 0.7492096 1.0000000 0.7492096 0.7492096
## [4,] 0.7492096 0.7492096 0.7492096 1.0000000 0.7492096
## [5,] 0.7492096 0.7492096 0.7492096 0.7492096 1.0000000
```

```
gee_model_cemented_exc_results<-geeCI(gee_model_cemented_exc)
```




```
gee_model_cemented_exc_results
```

```
##                   Estimate gee_model_cemented_exc   Naive.S.E.        lwrCI
## (Intercept)                          -0.816364392 0.2447705587 -1.296114687
## Months                                0.007625283 0.0008238053  0.006010625
## Age                                   0.003365700 0.0026998318 -0.001925970
## BMI                                   0.002889674 0.0036381769 -0.004241153
## factor(Sex)Female                    -0.093595239 0.0463260366 -0.184394271
##                           uprCI p.value    Naive.z  Robust.S.E.   Robust.z
## (Intercept)       -0.336614097  0.0006 -3.3352230 0.2375258086 -3.4369503
## Months             0.009239942  0.0000  9.2561712 0.0009600538  7.9425580
## Age                0.008657371  0.2139  1.2466334 0.0027079439  1.2428989
## BMI                0.010020501  0.4185  0.7942643 0.0035717735  0.8090306
## factor(Sex)Female -0.002796208  0.0277 -2.0203593 0.0425056440 -2.2019485
##                    lwrCIRobust  uprCIRobust
## (Intercept)       -1.281914977 -0.350813807
## Months             0.005743578  0.009506989
## Age               -0.001941870  0.008673271
## BMI               -0.004111002  0.009890350
## factor(Sex)Female -0.176906302 -0.010284177
```

```
gee_model_uncemented_ind <-gee(logMTPM~Months + Age + BMI + factor(Sex), data = Data.RSA.Uncemented, id = Subjec
t, na.action = na.omit, family = gaussian, corstr = "independence")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

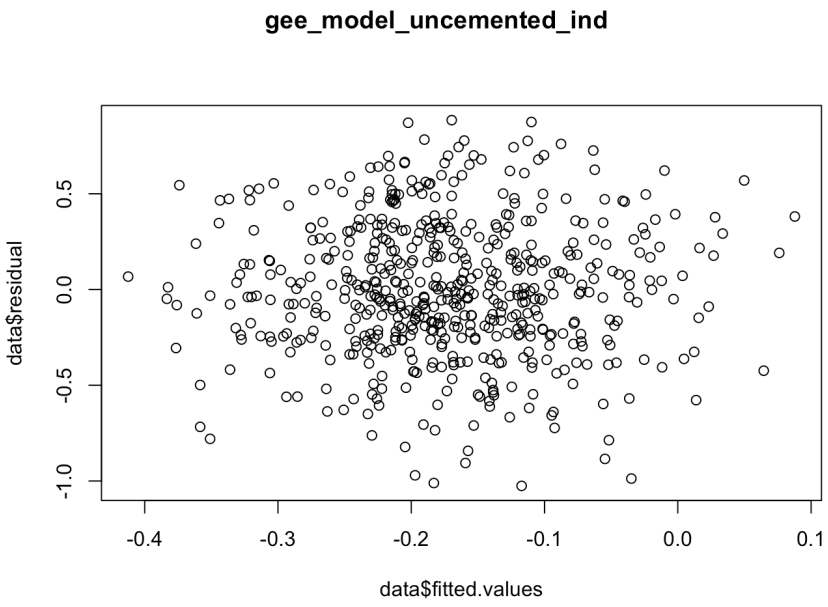
```
## running glm to get initial regression estimate
```

```
##       (Intercept)            Months               Age               BMI
##      -1.130673039       0.004984237       0.008734702       0.009515082
## factor(Sex)Female
##       0.061409712
```

```
summary(gee_model_uncemented_ind)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Identity
##  Variance to Mean Relation: Gaussian
##  Correlation Structure:     Independent
##
## Call:
## gee(formula = logMTPM ~ Months + Age + BMI + factor(Sex), id = Subject,
##     data = Data.RSA.Uncemented, na.action = na.omit, family = gaussian,
##     corstr = "independence")
##
## Summary of Residuals:
##        Min         1Q     Median         3Q        Max
## -1.0254721 -0.2330880 -0.0232837  0.2487793  0.8847244
##
##
## Coefficients:
##                       Estimate  Naive S.E.   Naive z Robust S.E.  Robust z
## (Intercept)       -1.130673039 0.200157341 -5.648921 0.365352469 -3.094746
## Months             0.004984237 0.001836328  2.714242 0.001117100  4.461765
## Age                0.008734702 0.002241661  3.896532 0.004126390  2.116790
## BMI                0.009515082 0.003114761  3.054835 0.006130884  1.551992
## factor(Sex)Female  0.061409712 0.030678732  2.001703 0.060564545  1.013955
##
## Estimated Scale Parameter:  0.1267272
## Number of Iterations:  1
##
## Working Correlation
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    0    0    0    0
## [2,]    0    1    0    0    0
## [3,]    0    0    1    0    0
## [4,]    0    0    0    1    0
## [5,]    0    0    0    0    1
```

```
gee_model_uncemented_ind_results<-geeCI(gee_model_uncemented_ind)
```




```
gee_model_uncemented_ind_results
```

```
##                   Estimate gee_model_uncemented_ind  Naive.S.E.        lwrCI
## (Intercept)                              -1.130673039 0.200157341 -1.522981427
## Months                                    0.004984237 0.001836328  0.001385035
## Age                                       0.008734702 0.002241661  0.004341047
## BMI                                       0.009515082 0.003114761  0.003410151
## factor(Sex)Female                         0.061409712 0.030678732  0.001279396
##                          uprCI p.value   Naive.z Robust.S.E.  Robust.z
## (Intercept)       -0.73836465  0.0020 -5.648921 0.365352469 -3.094746
## Months             0.00858344  0.0000  2.714242 0.001117100  4.461765
## Age                0.01312836  0.0343  3.896532 0.004126390  2.116790
## BMI                0.01562001  0.1207  3.054835 0.006130884  1.551992
## factor(Sex)Female  0.12154003  0.3106  2.001703 0.060564545  1.013955
##                     lwrCIRobust  uprCIRobust
## (Intercept)       -1.8467638790 -0.414582199
## Months             0.0027947217  0.007173753
## Age                0.0006469773  0.016822426
## BMI               -0.0025014511  0.021531615
## factor(Sex)Female -0.0572967973  0.180116221
```

```
gee_model_cemented_ind <-gee(logMTPM~Months + Age + BMI + factor(Sex), data = Data.RSA.Cemented, id = Subject, n
a.action = na.omit, family = gaussian, corstr = "independence")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

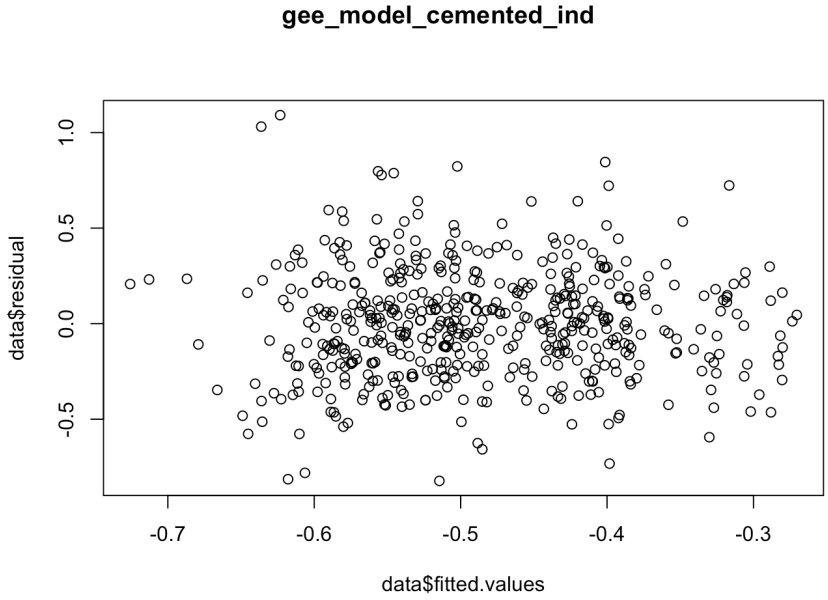
```
## running glm to get initial regression estimate
```

```
##       (Intercept)            Months               Age               BMI
##      -0.794453310       0.008618116       0.003500664       0.001264819
## factor(Sex)Female
##      -0.084976875
```

```
summary(gee_model_cemented_ind)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Identity
##  Variance to Mean Relation: Gaussian
##  Correlation Structure:     Independent
##
## Call:
## gee(formula = logMTPM ~ Months + Age + BMI + factor(Sex), id = Subject,
##     data = Data.RSA.Cemented, na.action = na.omit, family = gaussian,
##     corstr = "independence")
##
## Summary of Residuals:
##          Min           1Q       Median           3Q          Max
## -0.822721554 -0.203114406  0.001742283  0.178976330  1.091173996
##
##
## Coefficients:
##                      Estimate  Naive S.E.    Naive z Robust S.E.   Robust z
## (Intercept)       -0.794453310 0.139217473 -5.7065632 0.252506430 -3.1462696
## Months             0.008618116 0.001556620  5.5364286 0.001247648  6.9074927
## Age                0.003500664 0.001560017  2.2439906 0.002797331  1.2514303
## BMI                0.001264819 0.002061316  0.6135976 0.003681893  0.3435239
## factor(Sex)Female -0.084976875 0.027925269 -3.0430101 0.045588091 -1.8640148
##
## Estimated Scale Parameter:  0.0821903
## Number of Iterations:  1
##
## Working Correlation
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    0    0    0    0
## [2,]    0    1    0    0    0
## [3,]    0    0    1    0    0
## [4,]    0    0    0    1    0
## [5,]    0    0    0    0    1
```

```
gee_model_cemented_ind_results<-geeCI(gee_model_cemented_ind)
```




```
gee_model_cemented_ind_results
```

```
##                   Estimate gee_model_cemented_ind  Naive.S.E.         lwrCI
## (Intercept)                          -0.794453310 0.139217473 -1.0673195563
## Months                                0.008618116 0.001556620  0.0055671406
## Age                                   0.003500664 0.001560017  0.0004430302
## BMI                                   0.001264819 0.002061316 -0.0027753609
## factor(Sex)Female                    -0.084976875 0.027925269 -0.1397104021
##                           uprCI p.value    Naive.z Robust.S.E.   Robust.z
## (Intercept)       -0.521587063  0.0017 -5.7065632 0.252506430 -3.1462696
## Months             0.011669091  0.0000  5.5364286 0.001247648  6.9074927
## Age                0.006558298  0.2108  2.2439906 0.002797331  1.2514303
## BMI                0.005304998  0.7312  0.6135976 0.003681893  0.3435239
## factor(Sex)Female -0.030243348  0.0623 -3.0430101 0.045588091 -1.8640148
##                    lwrCIRobust  uprCIRobust
## (Intercept)       -1.289365912 -0.299540707
## Months             0.006172727  0.011063505
## Age               -0.001982104  0.008983432
## BMI               -0.005951692  0.008481329
## factor(Sex)Female -0.174329533  0.004375783
```

d. Which final model(s) would you choose to look at the associations of age, sex, and BMI on longitudinal migration for the cemented and uncemented groups?

- Difference in correlation structure (**corstr**) with respect to observations (ie. log(MTPM) values at (primarily) different follow-up months):
  - **AR-M**: Non-linear correlation (weakens with time)
  - **Exchangeable**: Constant correlation
  - **Independence**: No correlation
- **Empirical observations**: It's difficult to empirically tell which residual plot is the best fit given the abundance of data and their similar structures
- **Intuition**: Considering the definitions of each **corstr** value, I believe the performance would be ranked as AR-M > Exchangeable > Independence as the log(MTPM) values are heavily affected by 'months' as found by previous analysis. On a case-by-case level:
  - **Independence**: Invalid assumption as MTPM is known to increase with time due to impant wear-and-tear and the bone weakening
  - **Exchangeable & AR-M**: The relationship between MTPM and the dependent variables (primarily time) is unlikely to be constant over a long time-period
- **Statistical observations**: Based on the absolute median residual values (smaller values are better):
  - **Cemented**: ***Independence (0.0017)*** > Exchangeable (0.0019) > AR-M (0.0067)
  - **Uncemented**: ***Exchangeable (0.0060)*** > AR-M (0.015) > Independence (0.023)
- Purely based on the median residual values, I would use independence to model the cemented GEE and exchangeable for the uncemented GEE model
  - The favorable outcome of 'independence' is surprising for cemented. Either the code or model may have issues that may be causing this