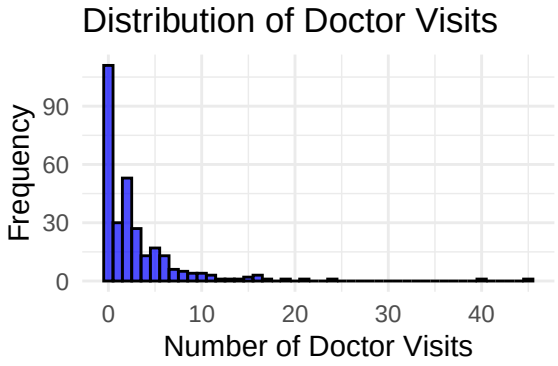
STAT7111 & STAT8111 Generalized Linear Models Assignment 3

Umut Demirhan - Student ID: 46739106

02 November, 2023

Question 1: Analyzing Patients with Primary Biliary Cirrhosis Data

A: Data Visualization and Preliminary Observations

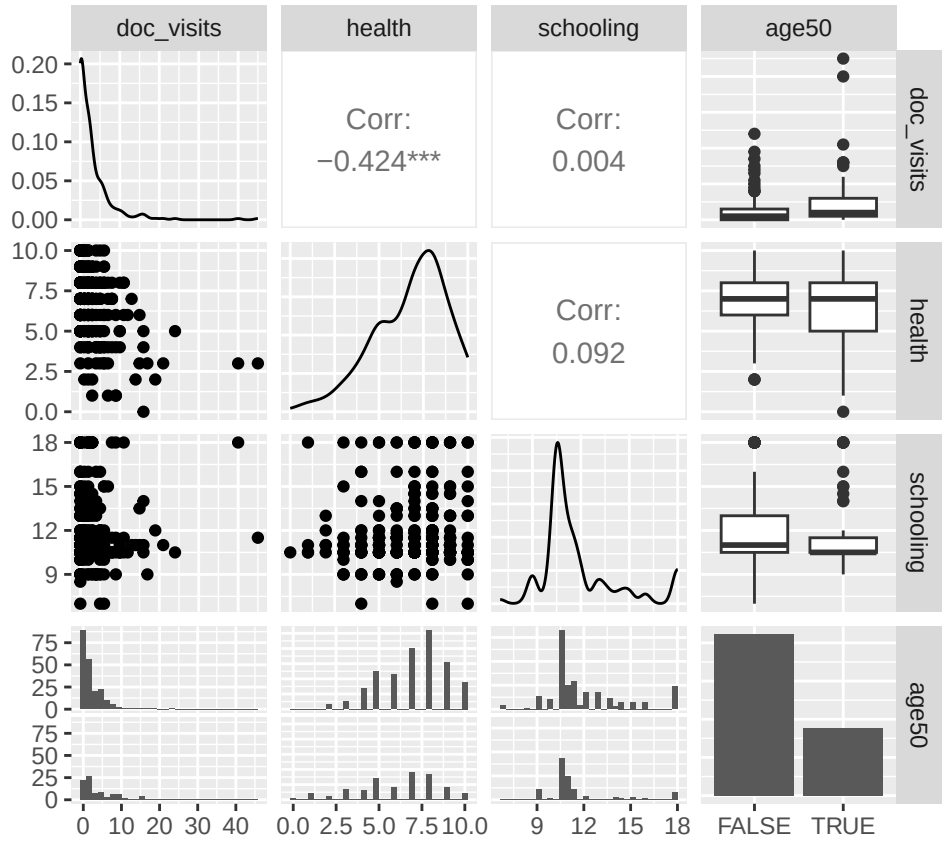


The histograms show a spike at zero, it indicates that many individuals did not visit a doctor at all, suggesting the potential need for a zero-inflated model,so ZINBI looks a suitable option.

B:Modeling the Number of Doctor Visits:

Table 1: Correlation Matrix

	doc_visits	health	schooling	age50
doc_visits	1.000	-0.424	0.004	0.227
health	-0.424	1.000	0.092	-0.212
schooling	0.004	0.092	1.000	-0.117
age50	0.227	-0.212	-0.117	1.000



Comments on initial visualisations The correlation matrix indicates a noticeable negative relationship between health satisfaction and doctor visits, suggesting that individuals with better health satisfaction scores visit doctors less frequently. This trend aligns with expectations, as healthier individuals might require fewer medical consultations. Meanwhile, individuals over 50 tend to visit doctors more often, as indicated by the positive correlation with doctor visits, perhaps due to age-related health concerns. The years of schooling seem to have minimal linear association with both health satisfaction and doctor visits.The correlation values among the covariates are all below 0.3, indicating that multicollinearity is likely not a concern for this model.

Examining the scatter and box plots, a downward trend in doctor visits with increasing health satisfaction is anticipated. The box plots show that older individuals have more doctor visits and slightly lower health satisfaction. Any trend with schooling appears weak, confirming the minimal correlation values in the matrix.

There are potential outliers that can be removed; However, as it is not specifically asked in the question, I will omit this operation from my analysis.

```
# Fit a ZINB model
zinb_model <- gamlss(doc_visits ~ health + factor(age50) + schooling,
                     family = ZINBI, data = doc_data)
```

```
## GAMLSS-RS iteration 1: Global Deviance = 1260.307
## GAMLSS-RS iteration 2: Global Deviance = 1251.634
## GAMLSS-RS iteration 3: Global Deviance = 1251.632
## GAMLSS-RS iteration 4: Global Deviance = 1251.632
```

```r
summary(zinb_model)
```

```
## ******************************************************************
## Family:  c("ZINBI", "Zero inflated negative binomial type I")
##
## Call:  gamlss(formula = doc_visits ~ health + factor(age50) + schooling,
##     family = ZINBI, data = doc_data)
##
## Fitting method: RS()
##
## ------------------------------------------------------------------
## Mu link function:  log
## Mu Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.528482   0.452353   5.590 5.17e-08 ***
## health             -0.266296   0.038570  -6.904 3.08e-11 ***
## factor(age50)TRUE   0.420869   0.178190   2.362   0.0188 *
## schooling           0.003518   0.028767   0.122   0.9028
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  log
## Sigma Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.2162     0.1265   1.709   0.0885 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Nu link function:  logit
## Nu Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -36.04    5773.50  -0.006    0.995
##
## ------------------------------------------------------------------
## No. of observations in the fit:  300
## Degrees of Freedom for the fit:  6
##       Residual Deg. of Freedom:  294
##                       at cycle:  4
##
## Global Deviance:     1251.632
##             AIC:     1263.632
##             SBC:     1285.855
## ******************************************************************
```

Comments on initial model The model suggests that health satisfaction scores and being over 50 are significant predictors of the number of doctor visits. As health scores increase, the expected number of doctor visits decreases. On the other hand, individuals over 50 years of age are likely to have more doctor visits than their younger counterparts. The variable schooling doesn't seem to play a significant role in predicting doctor visits in this model.

```r
# Check the significance of covariates
drop_results <- drop1All(zinb_model)

# Display using kable
kable(drop_results, caption="Significance of Covariates", digits=3, align='c')
```

Table 2: Significance of Covariates

	Df	AIC	LRT	Pr(Chi)
	NA	1263.632	NA	NA
health	1	1313.784	52.152	0.000
factor(age50)	1	1268.258	6.626	0.010
schooling	1	1261.645	0.013	0.909

Comments on the Results: Similar to initial models result;

- health: Highly significant (p-value < 0.001). Health scores notably influence doctor visits.
- factor(age50): Significant (p-value $= 0.01$). Being over 50 impacts the frequency of doctor visits.
- schooling: Not significant (p-value $= 0.909$). Schooling doesn't significantly affect doctor visits.

```r
# Stepwise selection based on GAIC
final_model <- stepGAIC(zinb_model)
```

```
## Distribution parameter:  mu
## Start:  AIC= 1263.63
##  doc_visits ~ health + factor(age50) + schooling
##
##                 Df    AIC
## - schooling      1 1261.6
## <none>             1263.6
## - factor(age50)  1 1268.3
## - health         1 1313.8
##
## Step:  AIC= 1261.64
##  doc_visits ~ health + factor(age50)
##
##                 Df    AIC
## <none>             1261.6
## - factor(age50)  1 1266.3
## - health         1 1311.8
```

```r
# Summary of the final model
summary(final_model)
```

```
## ******************************************************************
## Family:  c("ZINBI", "Zero inflated negative binomial type I")
##
## Call:  gamlss(formula = doc_visits ~ health + factor(age50), family = ZINBI,
##     data = doc_data, trace = FALSE)
##
```

```
## Fitting method: RS()
##
## ------------------------------------------------------------------
## Mu link function:  log
## Mu Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.56961    0.29148   8.816  < 2e-16 ***
## health            -0.26623    0.03861  -6.896 3.23e-11 ***
## factor(age50)TRUE  0.42067    0.17844   2.358    0.019 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  log
## Sigma Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.2165     0.1263   1.714   0.0875 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Nu link function:  logit
## Nu Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -36.04    5773.50  -0.006    0.995
##
## ------------------------------------------------------------------
## No. of observations in the fit:  300
## Degrees of Freedom for the fit:  5
##       Residual Deg. of Freedom:  295
##                       at cycle:  3
##
## Global Deviance:     1251.645
##             AIC:     1261.645
##             SBC:     1280.164
## ******************************************************************
```

Table 3: Comparison of AIC and BIC for Initial and Final Models

Model	df	AIC	BIC
final_model	5	1261.645	1285.855
zinb_model	6	1263.632	1280.164

Comments on final model The stepwise selection process based on GAIC suggests that removing the "schooling" variable yields the best model fit in terms of AIC.The final model retains the variables "health" and "factor(age50)". Both "health" and "factor(age50)" remain significant in the final model, consistent with the initial model's findings. The AIC of the final model (1261.645) is slightly lower than the initial model's AIC (1263.6), indicating a marginally better fit after dropping "schooling". Although the improvement is minimal, it's evident that "schooling" doesn't contribute significantly to the prediction of doctor visits, and thus its exclusion leads to a more parsimonious model without sacrificing explanatory power. However, BIC, which places a stricter penalty on model complexity, slightly favors the initial zinb_model.

Final Model Given the zero-inflated negative binomial distribution for the observations $i = 1, \ldots, n$:

$$Y_i \sim \text{ZINB}(\mu_i, \sigma, \pi_i)$$

The model can be described as:

Link function for (μ):

$$\log(\mu_i) = 2.56961 - 0.26623 \times \text{health} + 0.42067 \times \text{factor(age50)}$$

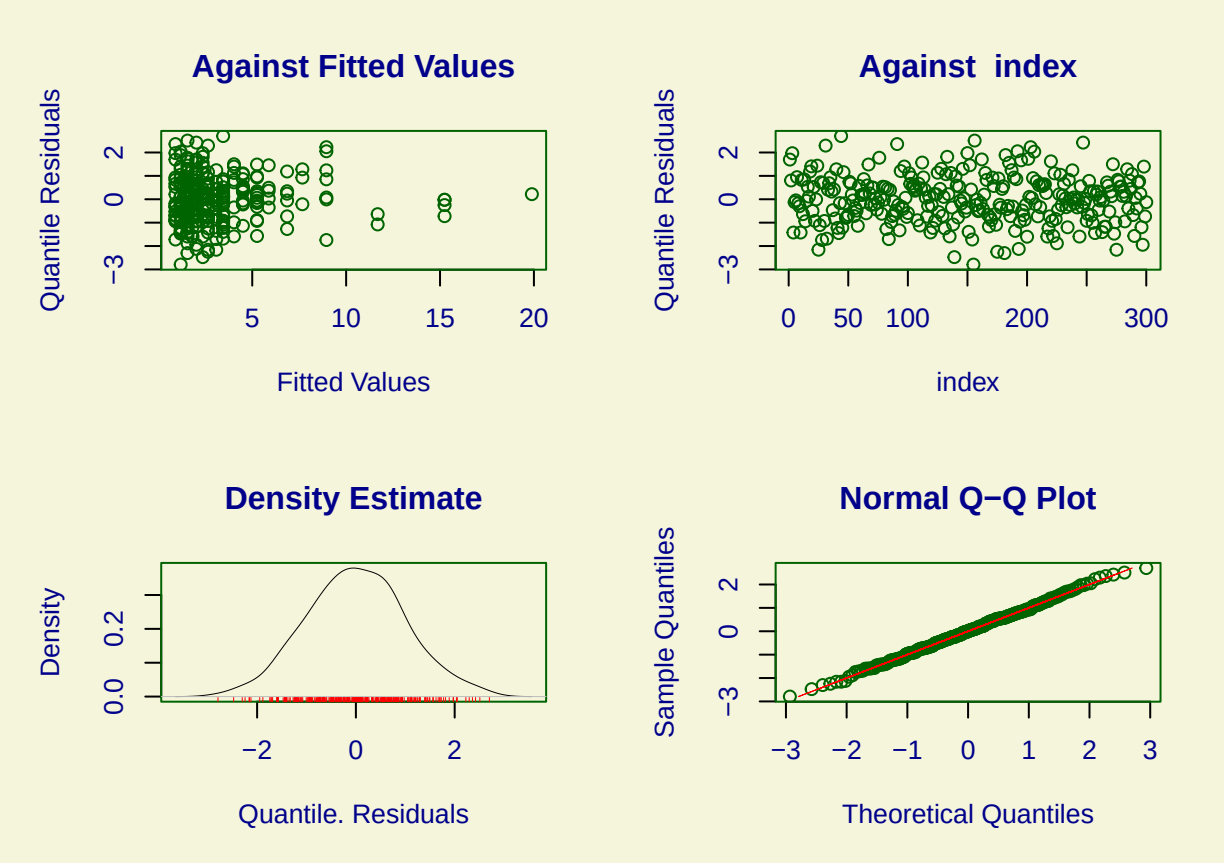
Link function for (σ):

$$\log(\sigma) = 0.2165$$

Link function for (π) (with logit link):

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = -36.04$$

Model Diagnostic




```
## ****************************************************************
##   Summary of the Randomised Quantile Residuals
##                            mean   =  0.003609857
##                        variance   =  0.9948354
##                coef. of skewness  =  0.03538392
##                coef. of kurtosis  =  2.857892
## Filliben correlation coefficient  =  0.9993178
## ****************************************************************
```

Comments on model diagnostic Upon visualizing the model diagnostics for the final_model, the plots indicate that the model assumptions are largely met. The residuals appear to be distributed randomly around zero, with no discernible patterns, which suggests that the model is capturing the underlying patterns in the data. Overall, the diagnostic plots support the conclusion that the final_model provides a good fit to the data and that the model assumptions are reasonably satisfied.

Interpretation of Model Parameters:

1. **Mean of Negative Binomial Component (μ)**:
   - **Intercept (Baseline)**: When all variables are zero, the expected number of doctor visits is $\exp(2.56961) \approx 13.06$.
   - **Health**: For each one-unit increase in the health score, the expected number of doctor visits decreases by approximately 23% [$\exp(-0.26623) \approx 0.77$, or a 23% decrease].
   - **Age50**: Being 50 or older is associated with an increase in expected doctor visits by a multiplicative factor of $\exp(0.42067) \approx 1.52$, or a 52% increase, compared to younger individuals.
2. **Overdispersion Parameter (σ)**:
   - The estimate is $\exp(0.2165) \approx 1.24$, indicating that there is approximately 24% more variability in the number of doctor visits than would be expected under a Poisson distribution (which assumes equal mean and variance).
3. **Zero-Inflation Parameter (π)**:
   - The estimate is extremely low (close to zero), suggesting that the excess zeros in the data are not significantly accounted for by zero-inflation, and are more likely due to the natural variability in doctor visits.

Question 2:

A: Compute the AIS code and give the corresponding frequency table.

```
# Importing the data
crash <- read.csv("crash.csv")
# Creating AIC based on the introductions
crash$AIS <- cut(crash$head_ic, breaks = c(135, 520, 900, 1255, 1575, 1860, Inf),
                 labels = c(1,2,3,4,5,6), right = FALSE)
# Calculate the missing values for each column and the total number of rows
missing_values_per_column <- sapply(crash, function(x) sum(is.na(x)))
missing_summary <- data.frame(t(missing_values_per_column))
missing_summary$TotalEntries <- nrow(crash)
kable(missing_summary, caption = "Summary of Missing Values for Each Column and Total Entries")
```

Table 4: Summary of Missing Values for Each Column and Total Entries

dp	weight	head_ic	AIS	TotalEntries
0	0	12	12	352

```r
# Frequency table before combining
initial_table <- table(crash$AIS)
# Combining categories 5 and 6 as they have less then 20 values
crash$AIS <- as.character(crash$AIS)
crash$AIS[crash$AIS %in% c("5", "6")] <- "5+"
# Generating a new frequency table
table_combined <- as.data.frame(table(crash$AIS))
colnames(table_combined) <- c("AIS Code", "Frequency")
# Creating a table in a nice format
kable(table_combined, caption = "Frequency Table", col.names = c("AIS Code", "Frequency"))
```

Table 5: Frequency Table

AIS Code	Frequency
1	60
2	148
3	72
4	30
5+	30

Comments on missing values and the frequency table

- There are 12 missing values. No ned to treat them before modelling as they will be ignored when fitting the model anyway. They do not affect the proportion of the categories of AIS much either.
- From the frequency table, categories 5 and 6 were sensibly combined into a new category "5+" as both had fewer than 20 observations each.The total observations for the combined "5+" category equals 30 (which is the sum of categories 5 and 6 from the initial table).

B: Fit an ordinal regression model

```r
# Fit the Cumulative Logit Model
model_ord <- vglm(ordered(AIS) ~ factor(dp) + weight,
family = cumulative(parallel = TRUE), data = crash)
summary(model_ord)
```

```
##
## Call:
## vglm(formula = ordered(AIS) ~ factor(dp) + weight, family = cumulative(parallel = TRUE),
##     data = crash)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept):1    -0.7838649  0.4898795  -1.600  0.10957
## (Intercept):2     1.3287791  0.4910795   2.706  0.00681 **
## (Intercept):3     2.4911537  0.5055276   4.928 8.31e-07 ***
## (Intercept):4     3.3134829  0.5250469   6.311 2.78e-10 ***
## factor(dp)Passen  0.8685650  0.2045149   4.247 2.17e-05 ***
## weight           -0.0004361  0.0001597  -2.730  0.00633 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]),
## logitlink(P[Y<=3]), logitlink(P[Y<=4])
##
## Residual deviance: 943.1875 on 1354 degrees of freedom
##
## Log-likelihood: -471.5937 on 1354 degrees of freedom
##
## Number of Fisher scoring iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Exponentiated coefficients:
## factor(dp)Passen           weight
##         2.383488         0.999564
```

Cumulative Logit Model Equation For $i = 1, \ldots, 352$ and $j = 1, 2, 3, 4$, the first model equation is:

$$j = 1 : \ln\left(\frac{\hat{\gamma}_{1i}}{1 - \hat{\gamma}_{1i}}\right) = -0.7839 + 0.8686 \times x_{i1} - 0.000436 \times x_{i2}$$

Where:

- x_{i1} represents the dummy variable for `dp`, coded as 1 for "Passenger" and 0 for "Driver".
- x_{i2} is the weight of the car.
- $\hat{\gamma}_{ji}$ represents the estimated probability that an observation falls in or below category j for the ith observation.

Interpratation of the Parameters The model parameters from the `vglm` output can be interpreted as follows:

- **Intercepts**: Baseline log-odds of being at or below each AIS injury severity category for a driver with a zero-weight car.
- **factor(dp)Passenger** ($\beta_1 = 0.8686$): The log-odds increase by 0.8686 for passengers compared to drivers. The odds ratio is calculated as $e^{0.8686}$, which is approximately 2.38. This means passengers are 2.38 times more likely to be at or below a given AIS category compared to drivers.
- **Weight** ($\beta_2 = -0.000436$): For each additional pound of car weight, the log-odds decrease by 0.000436. The odds ratio is calculated as $e^{-0.000436}$, which is approximately 0.999564. This indicates a very slight decrease in the odds of being at or below a given AIS category with each additional pound, showing a marginal protective effect of increased car weight on injury severity.

These interpretations consider the impact of being a passenger versus a driver and the effect of car weight on injury severity.

C: Fit a nominal regression model

```r
# Set AIS code = 1 as the reference category
crash$AIS <- relevel(factor(crash$AIS), ref = "1")
# Fit the multinomial logistic regression model
mcrash <- vglm(AIS ~ factor(dp) + weight, family = multinomial(refLevel = 1), data = crash)
summary(mcrash)
```

```
##
## Call:
## vglm(formula = AIS ~ factor(dp) + weight, family = multinomial(refLevel = 1),
##     data = crash)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept):1       2.1204281  0.8024382   2.642  0.00823 **
## (Intercept):2       1.5774830  0.9275440   1.701  0.08900 .
## (Intercept):3      -2.4722408  1.1429361  -2.163  0.03054 *
## (Intercept):4      -2.5128223  1.1162273  -2.251  0.02437 *
## factor(dp)Passen:1 -0.5010468  0.3204467  -1.564  0.11791
## factor(dp)Passen:2 -1.5123195  0.3750566  -4.032 5.52e-05 ***
## factor(dp)Passen:3 -2.1005086  0.5387538  -3.899 9.67e-05 ***
## factor(dp)Passen:4 -0.5739474  0.4609891  -1.245  0.21312
## weight:1           -0.0003182  0.0002633  -1.208  0.22695
## weight:2           -0.0002296  0.0003079  -0.746  0.45583
## weight:3            0.0008602  0.0003553   2.421  0.01546 *
## weight:4            0.0007064  0.0003442   2.053  0.04010 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: log(mu[,2]/mu[,1]), log(mu[,3]/mu[,1]),
## log(mu[,4]/mu[,1]), log(mu[,5]/mu[,1])
##
## Residual deviance: 916.8306 on 1348 degrees of freedom
##
## Log-likelihood: -458.4153 on 1348 degrees of freedom
##
## Number of Fisher scoring iterations: 5
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Reference group is level  1  of the response
```

Model Equation

The regression equation for the outcome AIS code = 2 compared to AIS code = 1:

$$\ln\left(\frac{\hat{\pi}_{2i}}{\hat{\pi}_{1i}}\right) = 1.5775 - 1.5123 \times \text{dp\_Passenger} - 0.0002296 \times \text{weight}$$

The interpretation of these coefficients:

- **Intercept (1.5775)**: Represents the log odds of the outcome being AIS code = 2 (as opposed to AIS code = 1) when the dummy is in the driver's seat (`dp` = Driver) and the weight is zero. It's important to note that a weight of zero is more of a theoretical reference point rather than a realistic scenario.
- **dp_Passenger (-1.5123)**: If a dummy is in the passenger seat, as opposed to being a driver, the log odds of observing AIS code = 2 (relative to AIS code = 1) decreases by 1.5123. In terms of odds, this coefficient implies: $e^{-1.5123} \approx 0.22$. This is roughly a 78% decrease in the odds of observing AIS code = 2 when compared to AIS code = 1 for a passenger dummy versus a driver.
- **Weight (-0.0002296)**: For every unit increase in weight, the log odds of the outcome being AIS code = 2 (relative to AIS code = 1) decreases by 0.0002296. This implies: $1 - e^{-0.0002296} \approx 0.023\%$. So, there's a negligible decrease in the odds with each unit increase in weight.

Question 3:

A: A marginal (GEE) model for modelling the mean of the Children's heights (Ht) across

age (Agedays) with Sex as adjusted variable

```
load("data.child.Rdata")
model_gee <- geeglm(Ht ~ Agedays + factor(Sex), data = data.child, id = Id, corstr = "exchangeable")
summary(model_gee)
```

```
##
## Call:
## geeglm(formula = Ht ~ Agedays + factor(Sex), data = data.child,
##     id = Id, corstr = "exchangeable")
##
##  Coefficients:
##                Estimate   Std.err     Wald Pr(>|W|)
## (Intercept)  6.642e+01 8.954e-02 550184.9   <2e-16 ***
## Agedays      4.508e-02 6.999e-05 414932.8   <2e-16 ***
## factor(Sex)1 1.536e+00 1.230e-01    156.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)    7.136  0.1455
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.2055 0.01483
## Number of clusters:   602  Maximum cluster size: 7
```

The GEE model indicates that as children age, their height increases by approximately 0.04508 cm per day. Furthermore, female children are, on average, 1.53640 cm taller than male counterparts when considering the same age. The chosen exchangeable correlation structure suggests a consistent correlation between repeated height measurements over time.

B: The distributional assumption

In fitting the GEE model using the **geeglm** function with the default **family = gaussian** argument (as evident from the summary which mention the "Link = identity"), the distributional assumption made is that the response variable, children's height (Ht), follows a **Gaussian (or Normal) distribution**. This implies that the residuals from the model are assumed to be normally distributed.

C: Fitted Model Equation

Given the output from the GEE model, the equation representing the mean height of children across age, adjusted for sex, is as follows:

$$\mu_{ij} = 66.41754 + 0.04508 \times \text{Agedays}_{ij} + 1.53640 \times \text{Sex}_i$$

Where:

- μ_{ij} represents the expected height for the i^{th} child at the j^{th} visit.
- Agedays_{ij} is the age in days of the i^{th} child at the j^{th} visit.
- Sex_i denotes the gender of the i^{th} child (0 for male, 1 for female).

Interpretation:

- The intercept, $\beta_0 = 66.41754$, is the estimated mean height (in cm) of male children (since Sex is 0 for male) when age is 0 days.
- The coefficient $\beta_1 = 0.04508$ implies that for each additional day in age, we expect the height to increase by approximately 0.04508 cm, keeping sex constant.
- The coefficient $\beta_2 = 1.53640$ suggests that, on average, female children have an estimated height that's 1.53640 cm greater than male children, when age is held constant.

D: Estimated Correlation Matrix for the Whole Model

The estimated correlation for any pair of observations within the same child is given by:

$$\alpha = 0.2055$$

Thus, the estimated exchangeable correlation matrix:

$$\begin{bmatrix} 1 & 0.2055 & 0.2055 & 0.2055 & 0.2055 & 0.2055 & 0.2055 \\ 0.2055 & 1 & 0.2055 & 0.2055 & 0.2055 & 0.2055 & 0.2055 \\ 0.2055 & 0.2055 & 1 & 0.2055 & 0.2055 & 0.2055 & 0.2055 \\ 0.2055 & 0.2055 & 0.2055 & 1 & 0.2055 & 0.2055 & 0.2055 \\ 0.2055 & 0.2055 & 0.2055 & 0.2055 & 1 & 0.2055 & 0.2055 \\ 0.2055 & 0.2055 & 0.2055 & 0.2055 & 0.2055 & 1 & 0.2055 \\ 0.2055 & 0.2055 & 0.2055 & 0.2055 & 0.2055 & 0.2055 & 1 \end{bmatrix}$$

This matrix reflects the exchangeable correlation assumption, with all pairs of observations within the same child having a consistent correlation of $\alpha = 0.2055$, irrespective of the time difference between visits.

E: Effect of Sex and Age on Babies' Heights

- **Age (`Agedays`)**: The coefficient estimate for age is 0.04508, which is highly significant with a p-value $<$ 2e-16. This suggests that for each additional day of age, the expected height of the baby increases by **0.04508 cm**, holding the sex constant.
- **Sex (`factor(Sex)1`)**: The coefficient for sex, where a code of 1 represents females, is 1.53640, which is also highly significant with a p-value $<$ 2e-16. This suggests that, on average, female babies are 1.53640 cm taller than male babies, holding age constant. In summary:
- Age has a significant and positive linear effect on the height of babies. As babies grow older by one day, their height is expected to increase by 0.04508 cm.
- Female babies are, on average, 1.53640 cm taller than male babies at the same age.

F: Suggesting a New Correlation Structure:

Considering the nature of the data - repeated measurements of children's heights at distinct intervals - an appropriate correlation structure, besides the exchangeable structure, could be the AR(1) or first-order autoregressive structure. This structure assumes that the correlation between observations decreases as the time gap between them increases. In the context of this data, it means that heights recorded at visits closer in time would be more correlated than those farther apart.

Reasons for suggesting AR(1):

- **Nature of Growth:** Children's growth might be more influenced by their most recent measurements rather than earlier ones. Thus, measurements taken closer together might be more correlated.
- **Decreasing Correlation over Time:** As time progresses, the influence of external factors (like diet, environment, activities, etc.) might differ, leading to a decreasing correlation between height measurements as the time gap increases. Let's fit the model using this correlation structure and compare it to the previous model:

```
# Fitting the GEE model with AR(1) correlation structure
model_gee_ar1 <- geeglm(Ht ~ sqrt(Agedays) + factor(Sex), data = data.child, id = Id, corstr = "ar1")
summary(model_gee_ar1)
```

```
##
## Call:
## geeglm(formula = Ht ~ sqrt(Agedays) + factor(Sex), data = data.child,
##     id = Id, corstr = "ar1")
##
##  Coefficients:
##               Estimate  Std.err   Wald Pr(>|W|)
## (Intercept)   56.00546  0.09896 320298   <2e-16 ***
## sqrt(Agedays)  1.50324  0.00285 277943   <2e-16 ***
## factor(Sex)1   1.62778  0.12523    169   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = ar1
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)     3.11   0.132
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.884 0.00627
## Number of clusters:   602  Maximum cluster size: 7
```

To compare the two models, checking the QIC (Quasi Information Criterion). Lower QIC values indicate better-fitting models:

```
# Extracting QIC values
QIC(model_gee_ar1,model_gee)
```

```
##                 QIC  QICu Quasi Lik   CIC params  QICC
```

```
## model_gee_ar1 13134 13117     -6556 11.15      3 13134
## model_gee     30080 30077    -15036  4.61      3 30080
```

- QIC for the AR(1) model: 31712
- QIC for the Exchangeable model: 30080

Lower QIC values indicate better model fit, and thus, the exchangeable model has a better fit for the data as compared to the AR(1) model.

The results suggest that the exchangeable correlation structure fits the data better than the AR(1) structure. Although it's reasonable to think that height measurements closer in time might be more correlated due to the nature of growth and external influences, the data does not seem to support this. Instead, it aligns more with the exchangeable structure, which assumes a consistent correlation between repeated height measurements over time.

G: Linear Mixed Model with Random Intercept

$$Ht_{ij} = \beta_0 + b_i + \beta_1 \times Agedays_{ij} + \beta_2 \times Sex_i + \epsilon_{ij}$$

with:

- $b_{0i} \sim N(\beta_0, \nu^2)$
- $\epsilon_{ij} \sim N(0, \tau^2)$

Where:

- Ht_{ij} is the height of the i^{th} child at the j^{th} visit.
- β_0 is the fixed effect intercept representing the average height across all children when $Agedays = 0$ and $Sex = 0$.
- b_i is the random intercept for the i^{th} child, which captures the child-specific deviation from the overall mean height. $b_i \sim N(0, \sigma_b^2)$, where σ_b^2 is the variance of the random intercepts.
- β_1 and β_2 are the fixed effect coefficients for `Agedays` and `Sex`, respectively.
- $Agedays_{ij}$ is the age in days of the i^{th} child at the j^{th} visit.
- Sex_i is the sex of the i^{th} child.
- ϵ_{ij} is the random error term associated with the j^{th} visit of the i^{th} child, assumed to be normally distributed with mean 0 and variance σ^2.

H: Showing that a linear mixed model with random intercept leads to an identical correlation

structure

The linear mixed-effects model is defined as:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp} + b_i + \epsilon_{ij}$$

where: - Y_{ij} is the observed value for the ith cluster and jth measurement, - $\beta_0, \beta_1, \ldots, \beta_p$ are the fixed effect coefficients, - b_i is the random intercept for the ith cluster, assumed to be $N(0, \nu^2)$, - ϵ_{ij} is the random error term, assumed to be $N(0, \tau^2)$ and independent of b_i.

The variance and covariance are then:

$$Var(Y_{ij}) = Var(b_i) + Var(\epsilon_{ij}) = \nu^2 + \tau^2$$

$$Cov(Y_{ij}, Y_{ik}) = E[(Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})] = \nu^2$$

This results in the variance-covariance matrix V_i for the ith cluster:

$$V_i = \begin{pmatrix} \nu^2 + \tau^2 & \nu^2 & \dots & \nu^2 \\ \nu^2 & \nu^2 + \tau^2 & \dots & \nu^2 \\ \vdots & \vdots & \ddots & \vdots \\ \nu^2 & \nu^2 & \dots & \nu^2 + \tau^2 \end{pmatrix}$$

The corresponding correlation matrix R_i showing an exchangeable structure is:

$$R_i = \begin{pmatrix} 1 & \frac{\nu^2}{\nu^2+\tau^2} & \dots & \frac{\nu^2}{\nu^2+\tau^2} \\ \frac{\nu^2}{\nu^2+\tau^2} & 1 & \dots & \frac{\nu^2}{\nu^2+\tau^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\nu^2}{\nu^2+\tau^2} & \frac{\nu^2}{\nu^2+\tau^2} & \dots & 1 \end{pmatrix}$$

This exchangeable correlation structure implies that all pairs of observations within the same cluster have the same correlation coefficient, which is a feature of the random intercept model.

I: Linear mixed model with a random intercept

```
# Running the linear mixed model with random intercept for the child
lmm_model <- lmer(Ht ~ Agedays + factor(Sex) + (1|Id), data = data.child)
summary(lmm_model)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Ht ~ Agedays + factor(Sex) + (1 | Id)
##    Data: data.child
##
## REML criterion at convergence: 19914
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -2.940 -0.701  0.139  0.750  2.444
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Id       (Intercept) 1.47     1.21
##  Residual             5.67     2.38
## Number of obs: 4214, groups:  Id, 602
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)  6.64e+01   1.01e-01   659.0
## Agedays      4.51e-02   1.62e-04   277.7
## factor(Sex)1 1.54e+00   1.23e-01    12.5
##
## Correlation of Fixed Effects:
##             (Intr) Agedys
```

```
## Agedays     -0.503
## factor(Sx)1 -0.611  0.000
```

```r
# Given results:
sigma2_b <- 1.47    # Variance of the random intercept
sigma2 <- 5.67      # Residual variance
# Compute the correlation
rho <- sigma2_b / (sigma2_b + sigma2)
rho
```

```
## [1] 0.206
```

Given the results from the mixed model, the estimated correlation between different time points for the same child, assuming an exchangeable correlation structure, can be calculated as:

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}$$

Where:

- σ_b^2 is the variance of the random intercepts, which is 1.47 in our case.
- σ^2 is the residual variance, which is 5.67 for this dataset.

Substituting in the given values:

$$\rho = \frac{1.47}{1.47 + 5.67} = 0.206$$

The estimated exchangeable correlation matrix for the repeated measurements within the same child would then be:

$$\begin{bmatrix} 1 & 0.206 & \cdots & 0.206 \\ 0.206 & 1 & \cdots & 0.206 \\ \vdots & \vdots & \ddots & \vdots \\ 0.206 & 0.206 & \cdots & 1 \end{bmatrix}_{7\times 7}$$

Where:

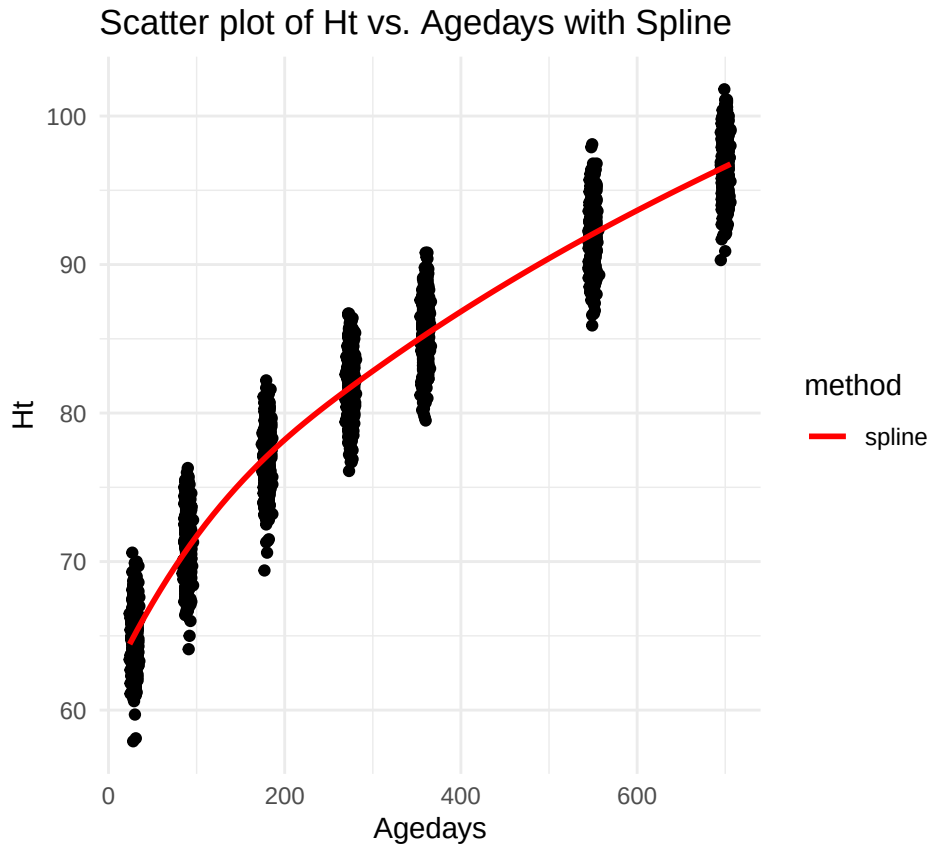
- The diagonal elements are all 1, indicating perfect correlation of a measurement with itself.
- The off-diagonal elements all have a value of 0.206, indicative of the correlation between different time points for the same child.

Remark:

The estimated correlation of 0.206 between the heights at different age days for the same child is significantly positive, indicating that heights on different days for a child are positively correlated. This emphasizes the importance of considering the correlation among repeated measures on the same subject in the analysis. The matrix showcases the inherent relationship and potential dependencies among the repeated measures within each child.

J: Scatter plot of Ht vs. Agedays with Spline

```r
ggplot(data.child, aes(x = Agedays, y = Ht)) +
  geom_point(color = "black") +
  geom_smooth(aes(colour = "spline"), method = lm, formula = y ~ splines::bs(x, 4), se = FALSE) +
  scale_color_manual(name = "method", values = c(spline = "red")) +
  theme_minimal() +
  labs(title = "Scatter plot of Ht vs. Agedays with Spline")
```



The scatter plot visualizes the relationship between Agedays (age in days) and Ht (height) for a given dataset. The individual data points are shown as black dots, indicating the exact height measurements for various ages. A red spline traces a curve through the data, indicating a general trend of increasing height with age. This spline provides a smoothed representation of the central tendency of height across age days. The trajectory suggests a consistent growth in height over the age span depicted.

K:Scatterplot of 'Ht' vs 'Agedays'

Based on the scatterplot of Ht vs Agedays, a strict linear growth model appears inadequate. The data shows a slight curvature, indicating the growth rate isn't constant across all ages. The spline fit further underscores this variation. Thus, a model with more flexibility, like a spline-based model, might better capture the relationship between Ht and Agedays.

L:Finding the appropriate form for Agedays in the model.

To determine the best form for Agedays in the model, I would use the square root transformation as I believe, it would effectively address the deceleration in height growth over time. This transformation counters the

non-linear pattern seen in early childhood, where growth is rapid and then slows, resulting in a more linear relationship.

AI Use Acknowledgement

I have used Chat GPT and Grammarly to revise my writing, and debugging.