STAT7111 & STAT8111 Generalized Linear Models Assignment 2

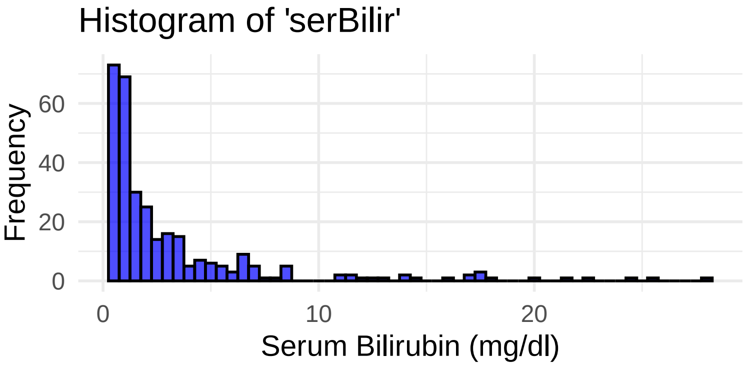
Umut Demirhan - Student ID: 46739106

06 October, 2023

Question 1: Analyzing Patients with Primary Biliary Cirrhosis Data

The dataset pbc.csv contains baseline data on 312 patients with primary biliary cirrhosis who were about to undergo a clinical trial for a treatment at the Mayo Clinic. I aim to study the relationship between serum bilirubin, considered as a strong indicator of disease progression, and other variables.

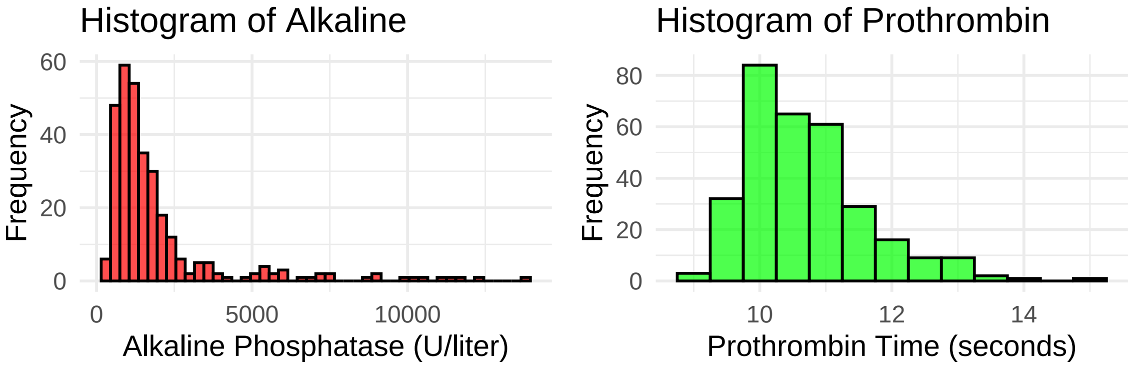
A: Examination of Variable `serBilir`



The variable `serBilir` displays a positively skewed, continuous, and non-negative distribution, which leans towards the selection of Gamma or Inverse Gaussian distributions. Mathematically: - For Gamma: $X \sim \Gamma(\alpha, \beta)$ - For Inverse Gaussian: $X \sim IG(\mu, \lambda)$

B: Examination of Variables `alkaline` and `prothrombin`

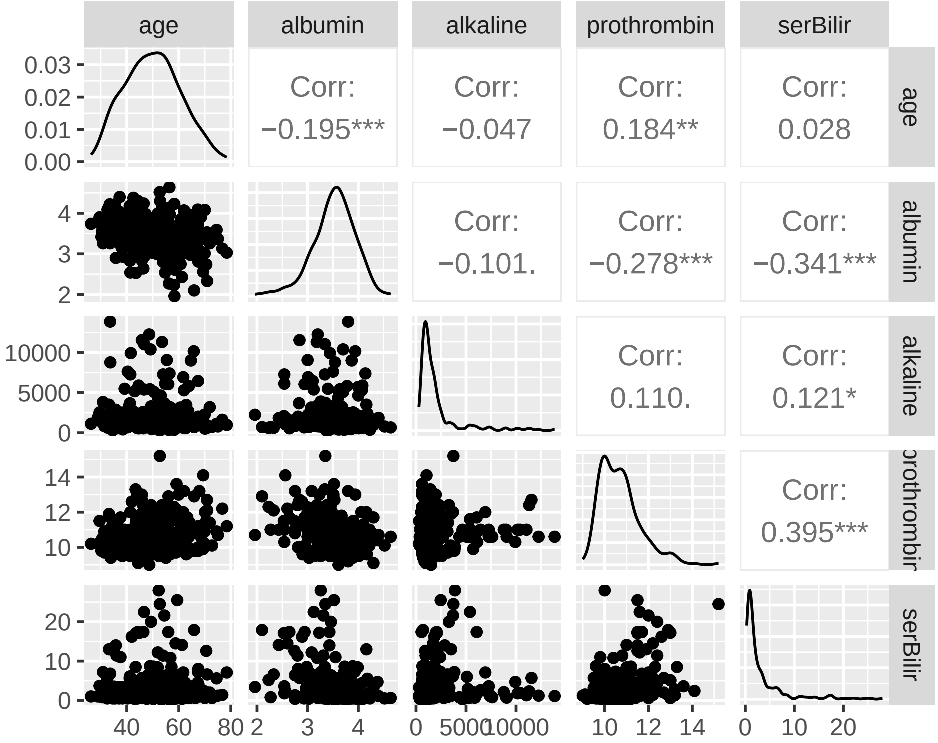
Graphical Examination of `alkaline` and `prothrombin`



- Examining the histogram for `alkaline`, I observe a pronounced right-skewness. The high variance, outliers, and skewed nature of `alkaline` suggest a necessity for a transformation to stabilize the variance and to achieve a more symmetrical, Gaussian-like distribution of the data. The `logalkaline` variable, obtained by $\log_e$(alkaline), minimizes the impact of extreme values and facilitates improved modeling.
- Moderate skewness and high-end values in `prothrombin` data might impact linear modeling. Logarithmic transformation is employed to enhance modeling by addressing skewness and stabilizing variance, expressed as logprothrombin $= \log_e$(prothrombin).

```
# Creating log-transformed variables
pbc$logalkaline <- log(pbc$alkaline)
pbc$logprothrombin <- log(pbc$prothrombin)
pbc$logserBilir <- log(pbc$serBilir)
```

C: Investigating Continuous Covariates and `serBilir`



Interpretation of Scatter Plots and Collinearity Investigation

- **Age and serBilir:** Scatter and correlation $r = 0.0279$ suggest a non-linear, weak relationship.
- **Albumin and serBilir:** Observed non-linear pattern in scatter; weak negative linear relationship $r = -0.341$.
- **Alkaline and serBilir:** Non-linear pattern with $r = 0.121$ indicates a weak positive relationship.
- **Prothrombin and serBilir:** Scatter implies exponential relationship; moderate positive correlation $r = 0.395$.

Collinearity Notes: No severe multicollinearity observed from correlation matrix. Moderate correlations warrant cautious interpretation of model coefficients.

D: Frequency Tables for Discrete Variables

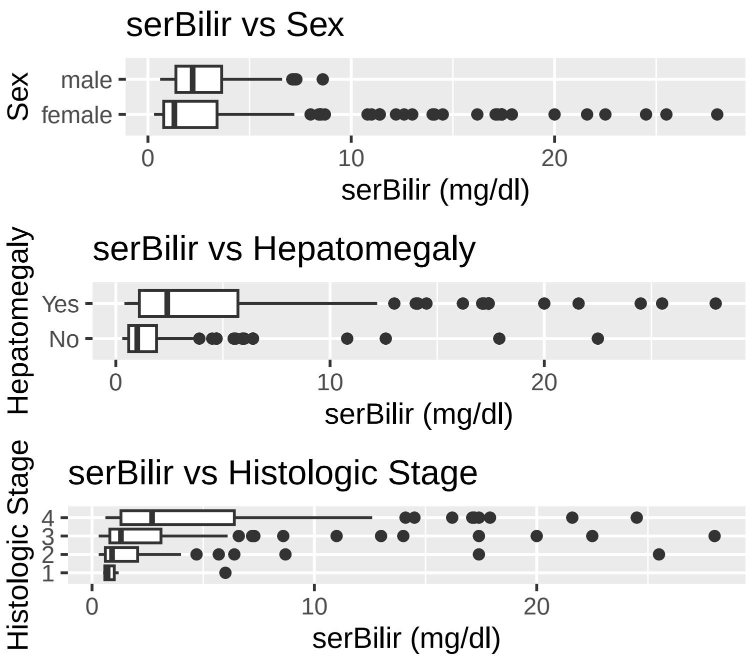
```r
# Discrete variables
discrete_vars <- c("sex", "hepatomegaly", "histologic")

# Generate and print frequency tables for each discrete variable in one line
for(var in discrete_vars){
  cat(paste("\nFrequency Table for '", var, "': ", sep=""), "\n")
  print(table(pbc[[var]]), quote = FALSE)
}
```

```
##
## Frequency Table for ’sex’:
##
## female   male
##    276     36
##
## Frequency Table for ’hepatomegaly’:
##
##  No Yes
## 152 160
##
## Frequency Table for ’histologic’:
##
##   1   2   3   4
##  16  67 120 109
```

- **Sex:** Predominantly female subjects (276 females vs. 36 males), indicating a substantial gender imbalance in the dataset.
- **Hepatomegaly:** There is a balanced distribution between subjects with hepatomegaly (`Yes` = 160) and those without (`No` = 152), allowing comparative analysis between these groups.
- **Histologic:** Most subjects are categorized in the moderate to severe histologic stages (2-4), with the least number of subjects in stage 1, suggesting a prevalence of moderate to severe conditions in the study population.

E: Investigation of Discrete Covariates as Predictors of serBilir



Interpretation of Boxplots for `serBilir` by Sex, Hepatomegaly, and Histologic Stage

- **Sex:** Notable differences in `serBilir` between sexes may suggest sex as a potential predictor, albeit with observed imbalance.
- **Hepatomegaly:** Disparities in `serBilir` levels between groups hint at hepatomegaly being a predictive factor.
- **Histologic Stage:** The trend of increasing `serBilir` with advancing histologic stages implies its predictive relevance.

F: Single-variable regressions of serBilir against each covariates

```r
# Initializing a dataframe to store results
regression_summary <- data.frame( Covariate = character(),Estimate = numeric(),
  Std.Error = numeric(),tValue = numeric(),Pr = numeric(),
  stringsAsFactors = FALSE)
# List of covariates
covariates <- c("age", "sex", "hepatomegaly", "albumin", "logalkaline",
                "logprothrombin", "histologic")
# Loop through each covariate and perform single-variable regression
for (covariate in covariates) {
  # Formulating model string
  model_string <- paste("serBilir ~", covariate)
  # Fitting the model
  model <- glm(model_string, family = Gamma(link = "log"), data = pbc)
  # Extracting summary statistics
  summary_stat <- summary(model)$coefficients[2,]
  # 2 corresponds to the covariate's row in the summary
  # Adding the result to the dataframe
  regression_summary <- rbind(regression_summary,
                              data.frame(Covariate = covariate,
                                         Estimate = summary_stat["Estimate"],
                                         Std.Error = summary_stat["Std. Error"],
                                         tValue = summary_stat["t value"],
                                         Pr = summary_stat["Pr(>|t|)"]))}
# Print the summary table
knitr::kable(regression_summary, caption = "Summary of Single-variable Regressions
             using Gamma Distribution with Log Link.", row.names = FALSE)
```

Table 1: Summary of Single-variable Regressions using Gamma Distribution with Log Link.

Covariate	Estimate	Std.Error	tValue	Pr
age	0.0038578	0.0074628	0.5169379	0.6055681
sex	-0.1270190	0.2450581	-0.5183219	0.6046034
hepatomegaly	0.8876389	0.1523794	5.8251907	0.0000000
albumin	-1.0156585	0.1749913	-5.8040523	0.0000000
logalkaline	0.5183143	0.1019231	5.0853483	0.0000006
logprothrombin	5.1822019	0.8532929	6.0731804	0.0000000
histologic	0.4187978	0.0934948	4.4793698	0.0000105

G: Forward model selection using AIC

```
# Extracting covariates that were significant at the 20% level from the univariate analysis
significant_covariates <- regression_summary$Covariate[regression_summary$Pr < 0.2]
# Creating a formula for the full model using significant covariates
full_model_formula <- as.formula(paste("serBilir ~", paste(significant_covariates,
                                                           collapse = " + ")))
# Starting with a null model
null_model <- glm(serBilir ~ 1, family = Gamma(link = "log"), data = pbc)
# Using step function for forward selection
final_model <- step(null_model,
                    scope = list(lower = null_model, upper = glm(full_model_formula,
                    family = Gamma(link = "log"), data = pbc)),
                    direction = "forward", trace = 1)
```

```
# Displaying the final model summary
summary(final_model)
```

```
##
## Call:
## glm(formula = serBilir ~ logprothrombin + hepatomegaly + logalkaline +
##     albumin, family = Gamma(link = "log"), data = pbc)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8294  -0.8373  -0.4013   0.2283   2.6858
##
```

```
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -7.99862    1.89689  -4.217 3.27e-05 ***
## logprothrombin   3.21106    0.69612   4.613 5.84e-06 ***
## hepatomegalyYes  0.51800    0.11849   4.372 1.69e-05 ***
## logalkaline      0.42250    0.07828   5.397 1.35e-07 ***
## albumin         -0.56930    0.14274  -3.988 8.32e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for Gamma family taken to be 0.9603474)
##
##     Null deviance: 373.71  on 311  degrees of freedom
## Residual deviance: 229.08  on 307  degrees of freedom
## AIC: 1193.9
##
## Number of Fisher Scoring iterations: 7
```

H: Final model equation

The model for $\log_e(\text{serBilir})$ is:

$\log_e(\text{serBilir}) = -7.99862 + 3.21106 \cdot X_2 + 0.51800 \cdot I_1 + 0.42250 \cdot \log_e(X_3) - 0.56930 \cdot \log_e(X_4) + \varepsilon$

Where:

- $log_e(X_2)$: prothrombin.
- I_1: hepatomegaly (1=yes).
- $log_e(X_3)$: alkaline.
- X_4: albumin.
- ε: Error term.

I: Interpretation of the Final Model

- **Hepatomegaly**: Associated with an increase in serum bilirubin of $e^{0.51800} - 1$ (approximately 68%), holding other variables constant and statistically significant.
- **Albumin**: A one-unit increase is linked to a serum bilirubin decrease of $e^{-0.56930} - 1$ (approximately 43%), controlling for other factors and is statistically significant.
- **logAlkaline**: A 1% rise relates to a 0.423% increase in serum bilirubin, ceteris paribus, and is highly significant.
- **logProthrombin**: A 1% increase is correlated with a 3.21% ascent in serum bilirubin, with other variables held constant, and is highly significant.
- **Intercept**: The significant intercept lacks practical interpretation, indicating a near-zero expected serum bilirubin for a hypothetical patient with all variables at baseline levels.

J: Characteristics of Patients with Elevated Levels of Serum Bilirubin

- **Presence of Hepatomegaly**: Patients with hepatomegaly are more likely to have elevated serum bilirubin levels, showing approximately a 68% increase compared to those without hepatomegaly.
- **Lower Albumin Levels**: Lower levels of albumin are associated with elevated levels of serum bilirubin. A one-unit increase in albumin corresponds to an approximate 43% decrease in serum bilirubin, suggesting that those with lower albumin levels are more prone to increased serum bilirubin levels.

- **Higher Alkaline Levels**: Elevated levels of alkaline are associated with elevated serum bilirubin. A 1% increase in alkaline is associated with a 0.42250% increase in serum bilirubin.
- **Higher Prothrombin Levels**: Elevated levels of prothrombin are also linked to elevated serum bilirubin. A 1% increase in prothrombin levels results in a 3.21106% increase in serum bilirubin levels.

In conclusion, patients with hepatomegaly, lower levels of albumin, and higher levels of alkaline and prothrombin are typically characterized by elevated levels of serum bilirubin. The significance and magnitude of these variables in the model underscore their importance in assessing the levels of serum bilirubin in patients.

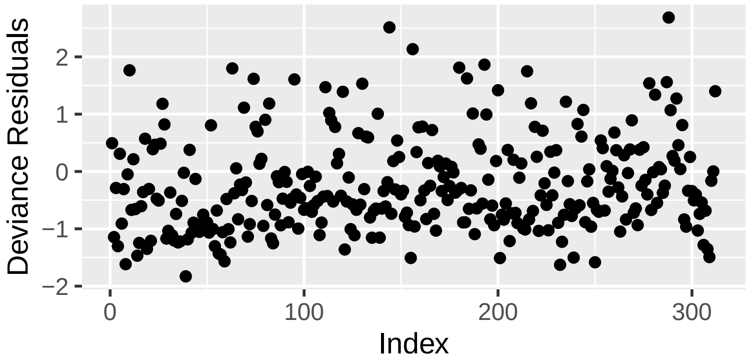
K: Model checking

```r
# calculating scaled deviance
scaled_deviance <- final_model$deviance / summary(final_model)$dispersion
cat("The Scaled Deviance is:", scaled_deviance,"\n")
```

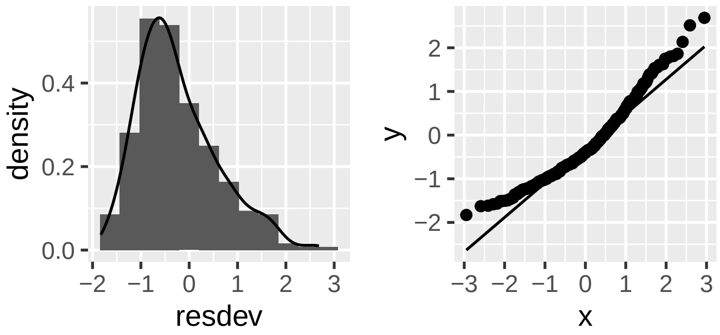
Scaled Deviance

```
## The Scaled Deviance is: 238.54
```

The scaled deviance obtained from our model is 238.54, with 306 degrees of freedom in the residuals. Generally, if a model fits well, I might expect the scaled deviance to be approximately equal to the residual degrees of freedom, suggesting that the model is not over-dispersed and provides a suitable fit to the data.

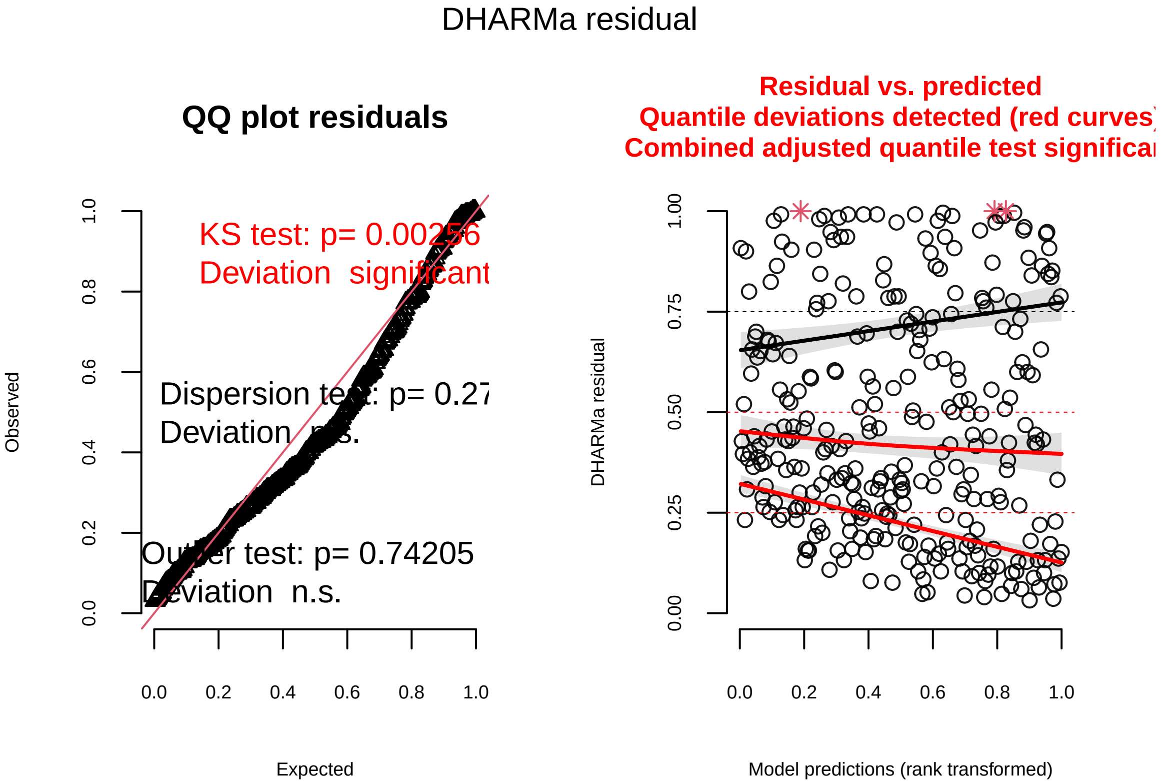


No obvious pattern observed above. Residuals are randomly scattered around the zero line without displaying any apparent pattern. Some points may be considered far away from the zero line and could be influential points or outliers. It will be investigated further.



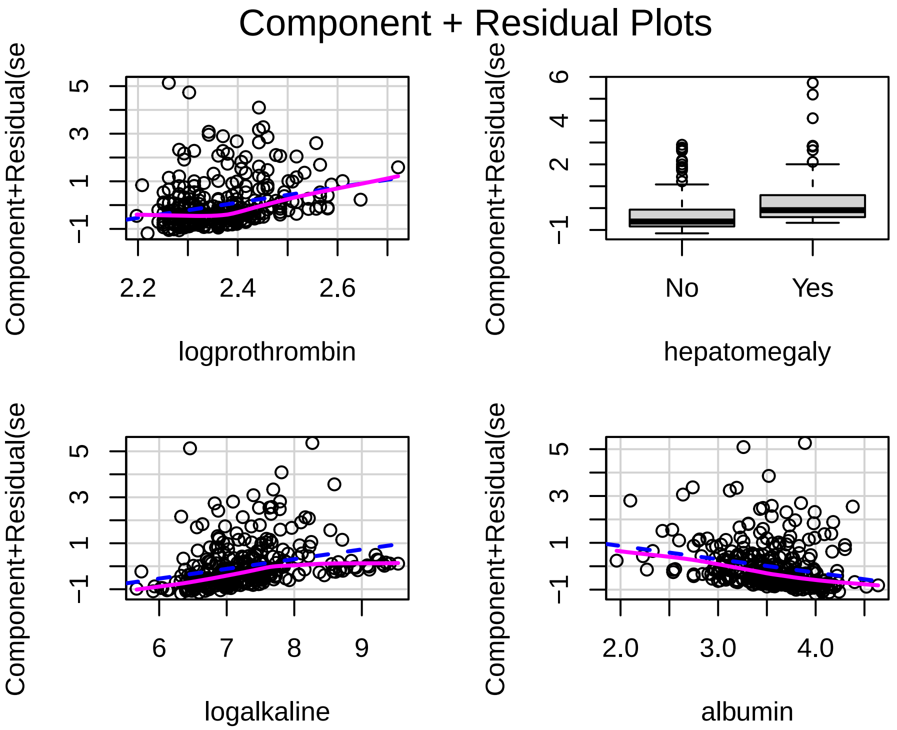
- The slightly right-skewed histogram suggests occasional underpredictions by the model, indicating potential non-normality in the residuals.
- Deviations in the tails of the Q-Q plot signal non-normality in the residuals, implying potential outlier effects or heavy-tailed distributions not captured by the model.

DHARMa Diagnostic Plots and Tests



- **QQ Plot**
  - **KS Test**: $p = 0.00256$ suggests significant deviation and potential issues with model fit.
  - **Dispersion Test**: $p = 0.272$ shows no concern for overdispersion.
  - **Outlier Test**: $p = 0.74205$ indicates no significant outliers affecting model fit.
- **Residuals vs Predictions**
  - Notable quantile deviations and a significant combined adjusted quantile test suggest model fit issues across various predicted values.

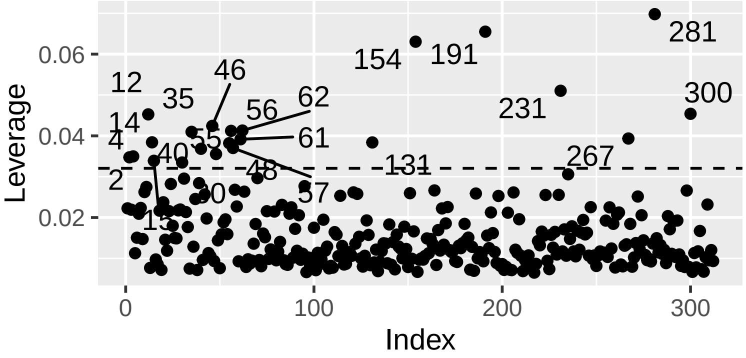
Partial residual plots

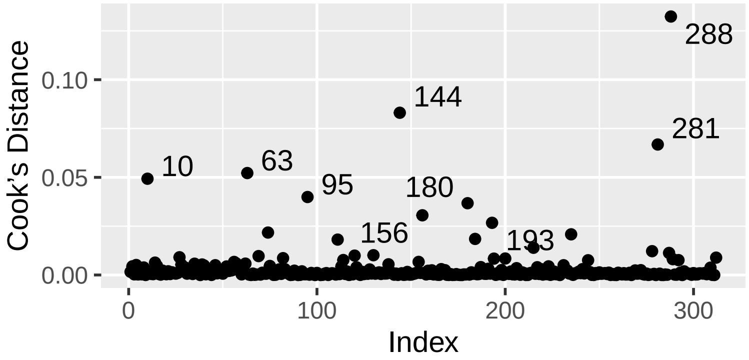


The broken and solid lines relatively align for all continuous variables, suggesting a reasonably linear relationship with the response when accounting for other predictors.hepatomegaly: The boxplot indicates that there is more variation and higher values for hepatomegalyYes compared to hepatomegalyNo, implying a notable difference in the response for these two groups.The assumptions of linearity seem largely met for continuous predictors, and there's a substantial distinction between levels of hepatomegaly with respect to the response variable.

Cook's distances and Leverage statistics

```
## Leverage cutoff: 0.03205128
```





Several observations exhibit high leverage, exceeding the calculated leverage cutoff of 0.03846154, necessitating further scrutiny to determine if they also possess high Cook's Distance values (beyond a cutoff value of "1"). While no observations demonstrate substantially large Cook's Distance values (using a cutoff of 1), suggesting an absence of highly influential points, the potential removal of observation "288" might enhance the model. This implies that, although some data points have notable influence on the model, they do not drastically alter the overall model fit and predictions. Nevertheless, a cautious approach entails investigating these points further to validate model robustness.

L) Model fitting with Inverse Gaussian distribution and log link

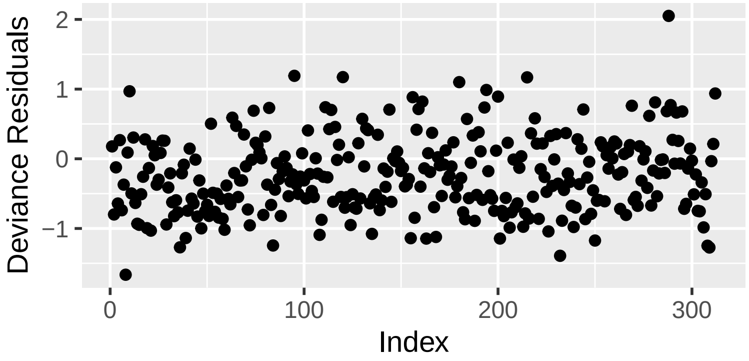
```
# Extracting covariates that were significant at the 20% level from the univariate analysis
significant_covariates <- regression_summary$Covariate[regression_summary$Pr < 0.2]
# Creating a formula for the full model using significant covariates
full_model_ig_formula <- as.formula(paste("serBilir ~", paste(significant_covariates,
                                                               collapse = " + ")))
# Starting with a null model
null_model_ig <- glm(serBilir ~ 1, family = inverse.gaussian(link = "log"), data = pbc)
# Using step function for forward selection
final_model_ig <- step(null_model_ig,
                    scope = list(lower = null_model_ig, upper =
                    glm(full_model_ig_formula, family = Gamma(link = "log"), data = pbc)),
                    direction = "forward", trace = 1)
```

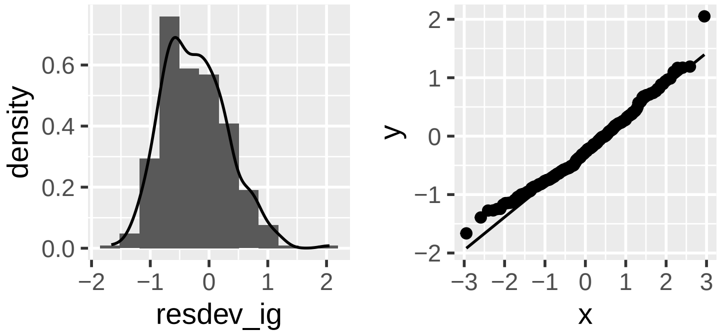
```
# Displaying the final model summary
summary(final_model_ig)
```

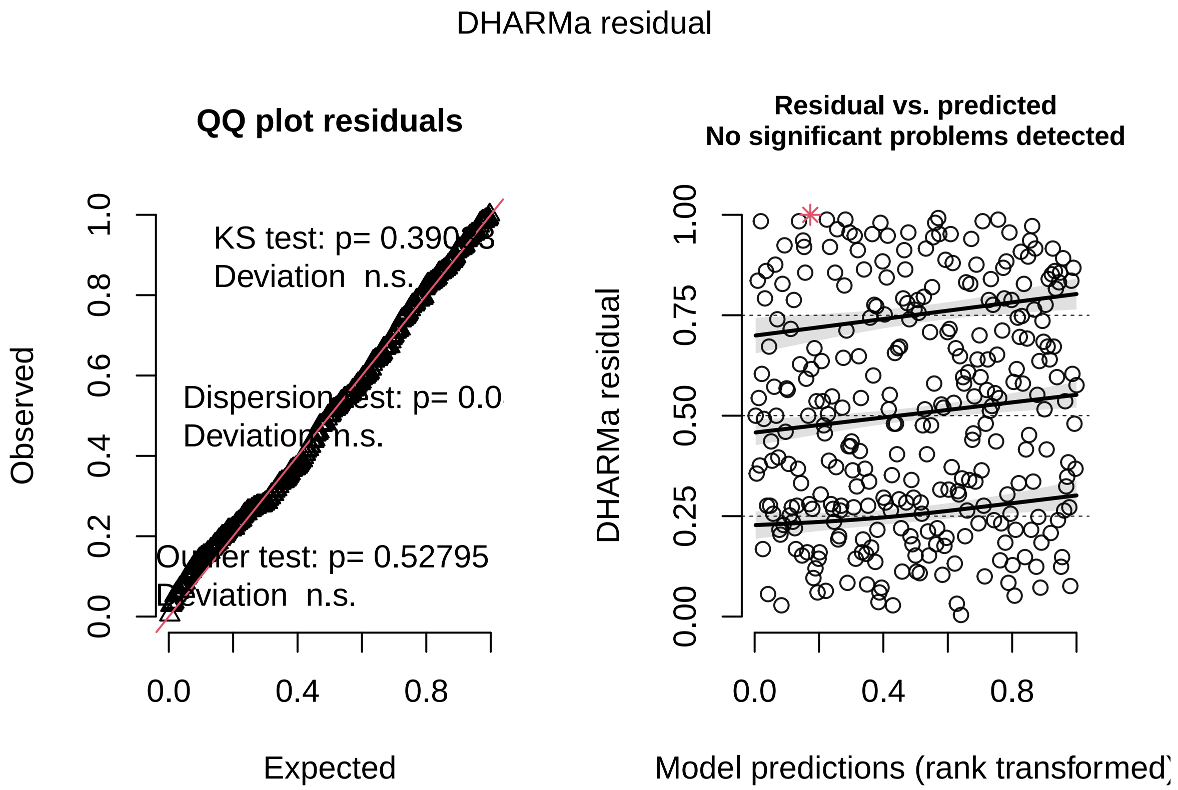
```
##
## Call:
## glm(formula = serBilir ~ hepatomegaly + logprothrombin + logalkaline +
##     albumin + histologic, family = inverse.gaussian(link = "log"),
##     data = pbc)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6636  -0.6407  -0.2677   0.1176   2.0504
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -7.69490    2.00623  -3.835 0.000152 ***
## hepatomegalyYes  0.56393    0.12834   4.394 1.54e-05 ***
## logprothrombin   2.13917    0.75210   2.844 0.004752 **
## logalkaline      0.58462    0.09218   6.342 8.13e-10 ***
## albumin         -0.41035    0.14528  -2.825 0.005045 **
## histologic       0.16239    0.06471   2.510 0.012601 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for inverse.gaussian family taken to be 0.4204193)
##
##     Null deviance: 169.28  on 311  degrees of freedom
## Residual deviance: 111.83  on 306  degrees of freedom
## AIC: 1112.2
##
## Number of Fisher Scoring iterations: 24
```

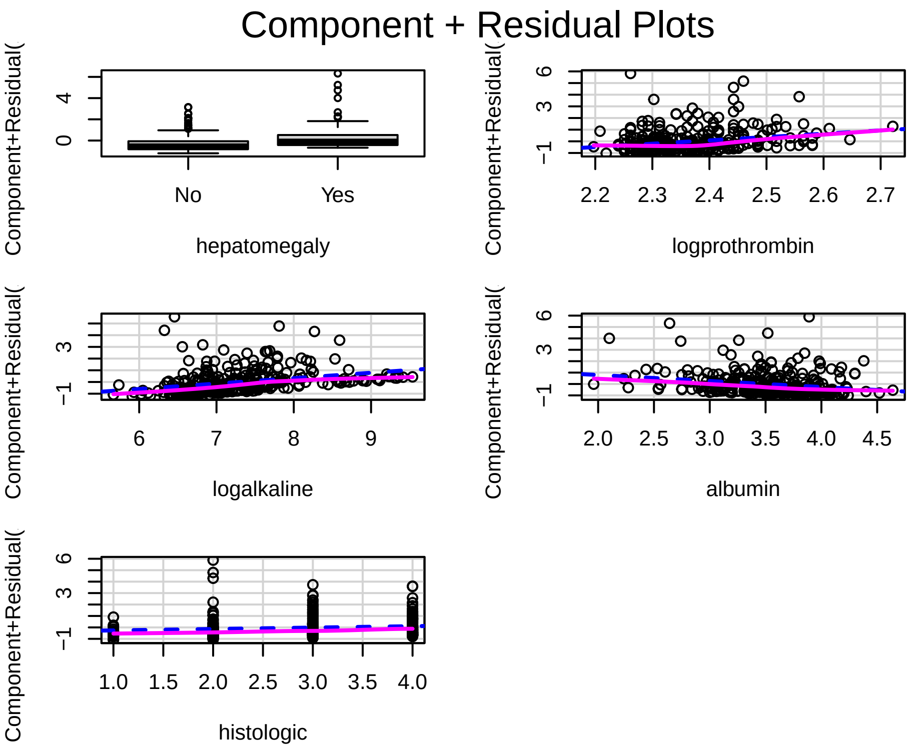
```
# calculating scaled deviance
scaled_deviance_ig <- final_model_ig$deviance / summary(final_model_ig)$dispersion
cat("The Scaled Deviance is:", scaled_deviance_ig,"\n")
```

```
## The Scaled Deviance is: 265.9984
```

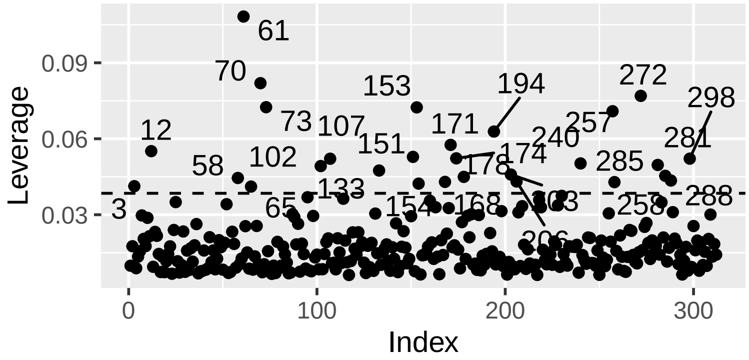


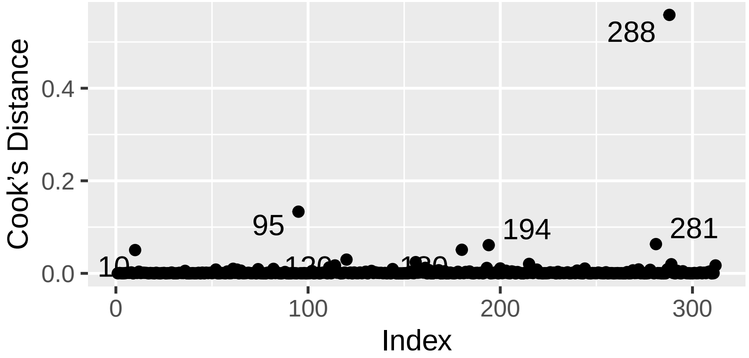






```
## Leverage cutoff: 0.03846154
```



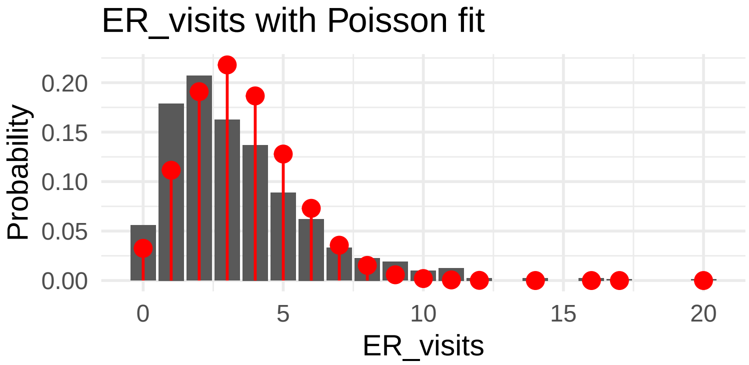


Model Diagnostic Summary

The inverse Gaussian model, exhibiting a higher scaled deviance (265.9984) compared to the Gaussian model (238.54),and displayed several promising diagnostic attributes.The deviance residuals showed desirable random dispersion, and the histogram indicated less right skewness in residual distribution. The Q-Q plot suggested a potentially better normal approximation of residuals, supported by non-significant DHARMa diagnostic tests. However, attention is needed for observations with leverage surpassing the 0.03846 cutoff, even though no major concerns were spotted in Cook's distance or partial residual plots.

Question 2: Analyzing Heart Disease Data

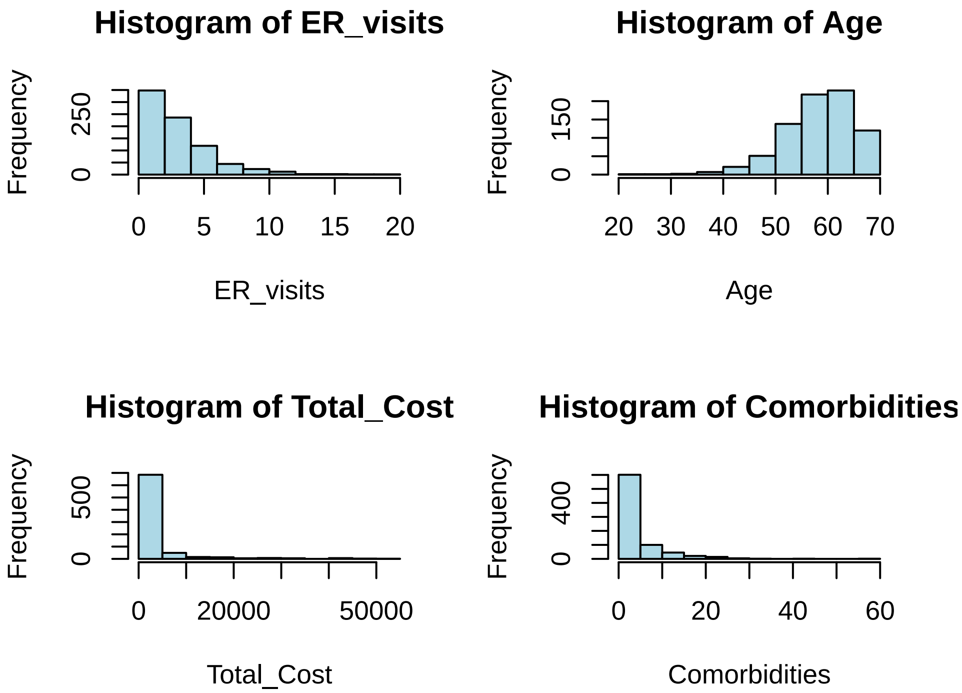
A: Examination of Variable ER_visits



```
## Mean of ER_visits: 3.425127
## Variance of ER_visits: 6.956267
```

Overdispersion Identified: The number of ER visits with a Poisson fit is higher than the number of ER visits without a Poisson fit. This means that there are more ER visits than would be expected under a Poisson distribution.The variance of ER visits (6.96) notably exceeds its mean (3.43), highlighting overdispersion and suggesting the possible inadequacy of Poisson regression for modeling this data. Alternative models like Negative Binomial Regression may be more apt due to its capability to handle overdispersion.

B: Graphical Examination of ER_visits, Age, Total_cost, and Comorbidities


Histogram of ER_visits

Histogram of Age

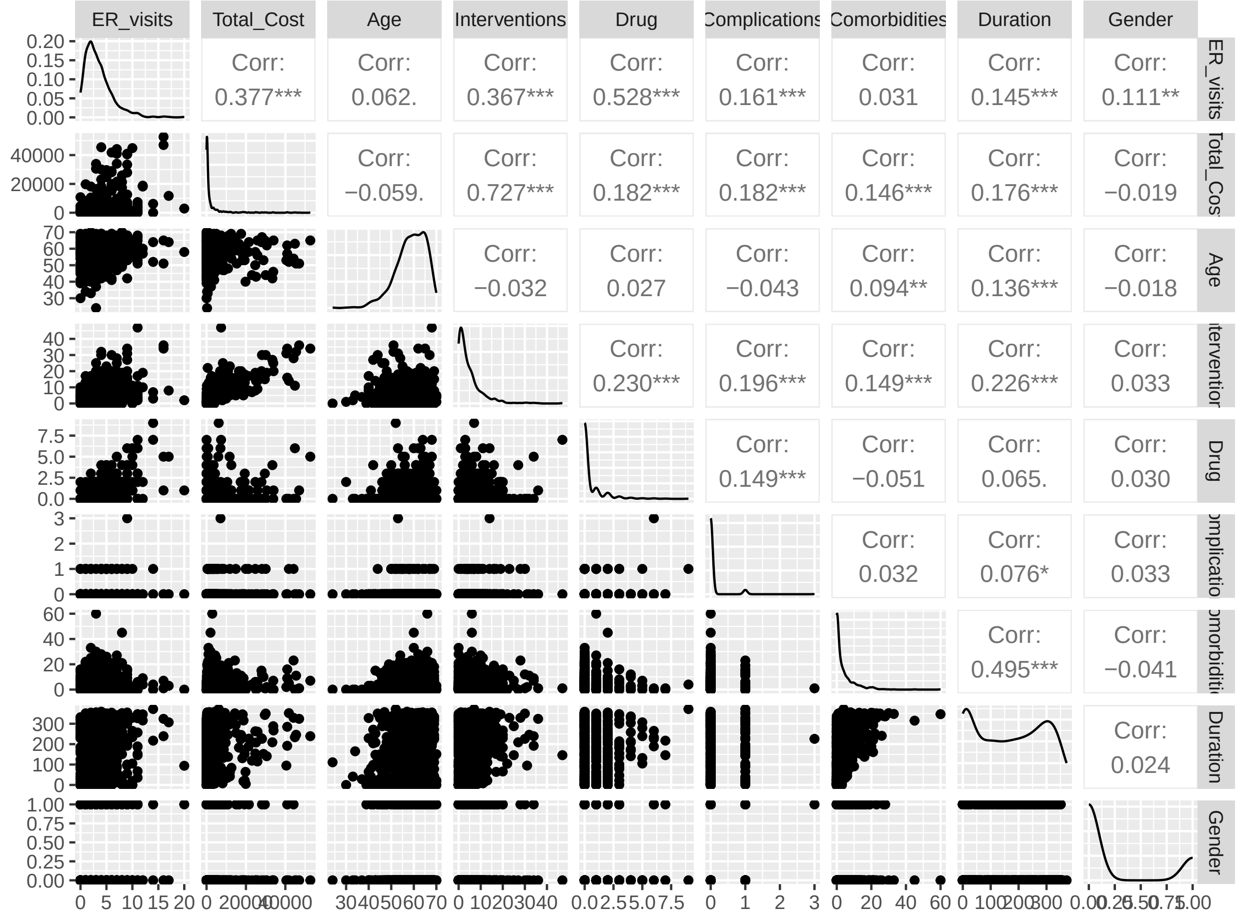
Histogram of Total_Cost

Histogram of Comorbidities

```
# Log transformations
# Add a small constant before taking log
heart_data$logtotal_cost <- log(heart_data$Total_Cost + 1)
heart_data$logcomorbidities <- log(heart_data$Comorbidities + 1)
```

- ER_visits is gradually decreasing frequency,no need to apply transformations to the dependent variable in a count regression context. This is especially true for Poisson or negative binomial regressions where the dependent variable is expected to be a count.
- Age may not need transformation despite left-skewness as older ages are naturally more prevalent.
- Total_Cost is right-skewed with a long tail; Log transformation applied to address right skewness and linearise its relationship with response variables.
- Comorbidities is also exhibits a long tail; log transformation can manage skewness and potentially stabilize variance when used in regression analyses.

C: Investigating Continuous Covariates



Interpretation of Scatter Plots and Collinearity Investigation

- **ER_visits and Total_Cost:** The correlation coefficient of $r = 0.377$ alongside scatter plot visualization might suggest a mild positive linear relationship.
- **ER_visits and Age:** A weak correlation of $r = 0.062$ and scatter plot might hint at a very subtle or potentially non-existent linear relationship between ER visits and age.However, categorizing age variable might improve the relationship (young & elder etc.)
- **ER_visits and Interventions:** The moderate correlation $r = 0.367$ coupled with scatter plot assessment might imply a relatively weak positive linear relationship.
- **ER_visits and Drug:** Noticing a moderately strong correlation of $r = 0.528$ and a scatter plot could reveal a reasonably discernible positive linear relationship.
- **Gender and ER_visits:** The correlation of 0.111 indicates a slight positive linear relationship between gender and ER visits.
- **Collinearity Notes:** Noticing some moderate correlations (e.g., Total_Cost and Interventions $r = 0.727$), caution should be taken to avoid multicollinearity issues in the regression model. The association between Comorbidities and Duration ($r = 0.495$) also stands out, inviting a deeper dive during modeling.

D) Using Drug as the only covariate, fit a Poisson GLM on ER_visits.

```r
# Fit a Poisson GLM
poisson_model <- glm(ER_visits ~ Drug, family = poisson, data = heart_data)
summary(poisson_model)
```

```
##
## Call:
## glm(formula = ER_visits ~ Drug, family = poisson, data = heart_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0741  -1.1951  -0.3376   0.5768   5.8803
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.0834     0.0217   49.93   <2e-16 ***
## Drug          0.2348     0.0114   20.60   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1485.0  on 787  degrees of freedom
## Residual deviance: 1169.5  on 786  degrees of freedom
## AIC: 3382.9
##
## Number of Fisher Scoring iterations: 5
```

Discussion: The max deviance residual is quite large, indicating potential outliers or misfit. Residual Deviance relative to the degrees of freedom suggests overdispersion (1169.5/786), violating the equidispersion assumption of Poisson regression.

Alternative Model Recommendation: Given the indication of overdispersion, Negative Binomial Regression model accommodates it by introducing an additional parameter. Negative Binomial regression adjusts for overdispersion, providing more robust standard errors and inferences.Improved Fit: It may provide a better fit and more reliable estimates if the data violates Poisson's assumptions.

E: Single-variable regressions of ER_visits against each covariates

```r
# Initializing a dataframe to store results
regression_summary_nb <- data.frame(Covariate = character(), Estimate = numeric(),
                                    Std.Error = numeric(), zValue = numeric(), Pr = numeric(),
                                    stringsAsFactors = FALSE)
# List of covariates
covariates_p <- c("Total_Cost","logtotal_cost","logcomorbidities","Age",
    "Gender","Interventions","Drug","Complications","Comorbidities","Duration")
# Loop through each covariate and perform single-variable regression
for (covariate in covariates_p) {
  # Formulating model string
  model_string_p <- paste("ER_visits ~", covariate)
  # Fitting the model
  model_p <- glm.nb(model_string_p, data = heart_data)
  # Extracting summary statistics
  summary_stat_p <- summary(model_p)$coefficients[2,]
  # Adding the result to the dataframe
  regression_summary_nb <- rbind(regression_summary_nb,
                                data.frame(Covariate = covariate,
                                           Estimate = summary_stat_p["Estimate"],
                                           Std.Error = summary_stat_p["Std. Error"],
                                           zValue = summary_stat_p["z value"],
                                           Pr = summary_stat_p["Pr(>|z|)"]))}
# Print the summary table
kable(regression_summary_nb, caption = "Summary of Single-variable Regressions using
      Negative Binomial Distribution with Log Link.", row.names = FALSE)
```

Table 2: Summary of Single-variable Regressions using Negative Binomial Distribution with Log Link.

Covariate	Estimate	Std.Error	zValue	Pr
Total_Cost	0.0000300	0.0000030	9.9336607	0.0000000
logtotal_cost	0.1179146	0.0132247	8.9162250	0.0000000
logcomorbidities	0.0221598	0.0256960	0.8623838	0.3884764
Age	0.0074515	0.0039303	1.8959233	0.0579702
Gender	0.1928216	0.0604898	3.1876689	0.0014342
Interventions	0.0385982	0.0038411	10.0486812	0.0000000
Drug	0.2532490	0.0170298	14.8709737	0.0000000
Complications	0.3906757	0.0931705	4.1931254	0.0000275
Comorbidities	0.0038202	0.0043312	0.8820299	0.3777607
Duration	0.0009366	0.0002153	4.3509671	0.0000136

F: Backward model selection using AIC

```r
# Full model with variables significant at the 20% level in the univariate analyses
full_model_nb <- glm.nb(ER_visits ~ Total_Cost + Age + Gender + Interventions +
                          Drug + Complications + Duration,
                        data = heart_data)
# Backward model selection using AIC with step function
final_model_nb <- step(full_model_nb, direction = "backward", trace = TRUE)
```

```r
# Displaying the summary of the final model
summary(final_model_nb)
```

```
##
## Call:
## glm.nb(formula = ER_visits ~ Total_Cost + Age + Gender + Interventions +
##     Drug + Duration, data = heart_data, init.theta = 11.93881587,
##     link = log)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -2.4356  -0.9369  -0.1988   0.4722   4.5078
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   4.579e-01  2.009e-01   2.279 0.022665 *
## Total_Cost    1.620e-05  3.801e-06   4.262 2.03e-05 ***
## Age           7.192e-03  3.388e-03   2.123 0.033773 *
## Gender        1.876e-01  5.090e-02   3.685 0.000229 ***
## Interventions 1.140e-02  4.961e-03   2.298 0.021571 *
## Drug          2.120e-01  1.611e-02  13.163  < 2e-16 ***
## Duration      2.941e-04  1.930e-04   1.524 0.127563
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(11.9388) family taken to be 1)
##
##     Null deviance: 1152.6  on 787  degrees of freedom
## Residual deviance:  820.5  on 781  degrees of freedom
## AIC: 3236.6
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  11.94
##           Std. Err.:  2.57
##
##  2 x log-likelihood:  -3220.647
```

G: Final model equation

Given the results of the final model, the equation to predict the expected number of emergency room (*ER_visits*) visits using a Negative Binomial regression model is expressed as:

$$\log(ER\_visits) = \beta_0 + \beta_1 \times Total\_Cost + \beta_2 \times Age + \beta_3 \times Gender + \beta_4 \times Interventions + \beta_5 \times Drug + \beta_6 \times Duration + \varepsilon$$

where:

- $\log(\cdot)$ denotes the natural logarithm.
- $\beta_0 = 0.4579$ is the intercept.
- $\beta_1 = 1.620 \times 10^{-5}$ is the coefficient for $Total\_Cost$ (the total cost of claims by subscriber in dollars).
- $\beta_2 = 0.007192$ is the coefficient for Age (age of the subscriber in years).
- $\beta_3 = 0.1876$ is the coefficient for $Gender$ (1 if the subscriber is male, and 0 otherwise).
- $\beta_4 = 0.01140$ is the coefficient for $Interventions$ (total number of interventions or procedures carried out).
- $\beta_5 = 0.2120$ is the coefficient for $Drug$ (number of tracked drugs prescribed).
- $\beta_6 = 0.0002941$ is the coefficient for $Duration$ (number of days of duration of treatment condition).
- ε: Error term.

H: Interpretation of the Final Model

- **Total_Cost**: A one-unit increase (i.e., one-dollar increase in total claims) is associated with an increase in ER visits of $e^{1.620\times10^{-5}} - 1$ (approximately 0.0016% or about a 0.16% increase for every

100 dollar increase), holding all other variables constant. This relationship is statistically significant.

- **Age**: A one-year increase in age is linked with an increase in ER visits of $e^{0.007192} - 1$ (approximately 0.72%), controlling for other factors, and is statistically significant.
- **Gender**: Being male (Gender = 1) is associated with an increase in ER visits of $e^{0.1876} - 1$ (approximately 20.6%), ceteris paribus, and is highly significant.
- **Interventions**: A one-unit increase in the total number of interventions or procedures is correlated with an increase in ER visits of $e^{0.01140} - 1$ (approximately 1.15%), with all else held constant, and is statistically significant.
- **Drug**: A one-unit increase in the number of tracked drugs prescribed is associated with an increase in ER visits of $e^{0.2120} - 1$ (approximately 23.6%), holding all other variables constant, and is statistically significant.
- **Duration**: A one-day increase in the duration of treatment condition is linked to an increase in ER visits of $e^{0.0002941} - 1$ (approximately 0.029%), controlling for other factors. However, note that this variable was not statistically significant in the final model.
- **Intercept**: The model's intercept of 0.4579 would theoretically represent the log count of ER visits for a female (Gender = 0) with all numerical predictors being zero and categorical variables at their reference levels. However, since it is not practically possible for many of these variables (like Age, Total_Cost, etc.) to be zero, the intercept does not have a tangible interpretation without considering the context of the other variables.

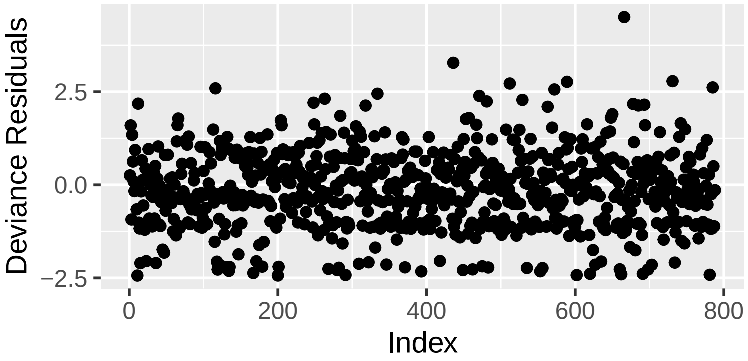
I: Model checking

```r
# calculating scaled deviance
scaled_deviance <- final_model_nb$deviance / summary(final_model_nb)$dispersion
cat("The Scaled Deviance is:", scaled_deviance,"\n")
```

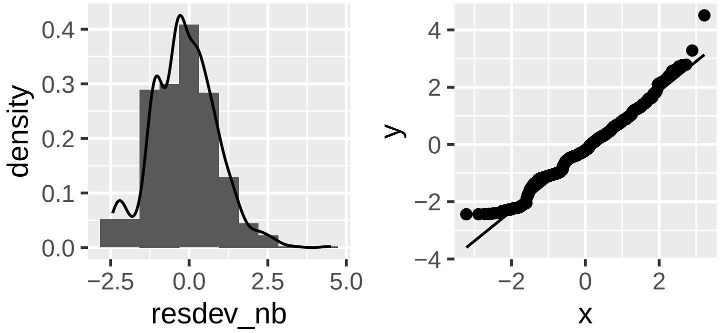
Scaled Deviance

```
## The Scaled Deviance is: 820.5005
```

The scaled deviance from our adjusted Negative Binomial model is 820.5005 with 781 residual degrees of freedom. Ideally, these two values should be fairly close if the model fits well, suggesting that our model might not provide a perfect fit to the data.However it is still quite good and it may offer useful insights.

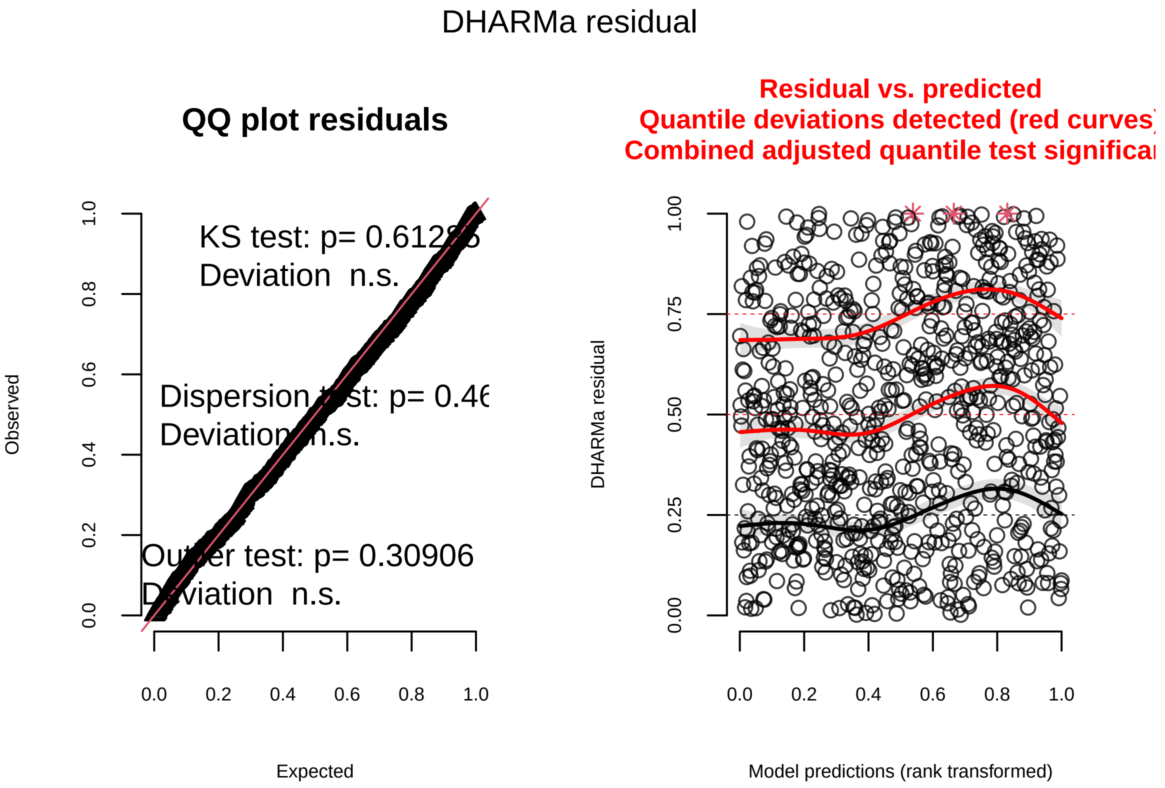


The observed residuals exhibit no clear pattern, scattering arbitrarily around the zero line without revealing any noticeable trend.



- The mildly right-skewed histogram indicates sporadic underestimations from the negative binomial model, hinting at a potential departure from residual normality.
- Tail divergences in the Q-Q plot highlight the presence of non-normal residuals in the negative binomial model, suggesting the possibility of unaccounted outlier influence or distributions with heavy tails.

DHARMa Diagnostic Plots and Tests

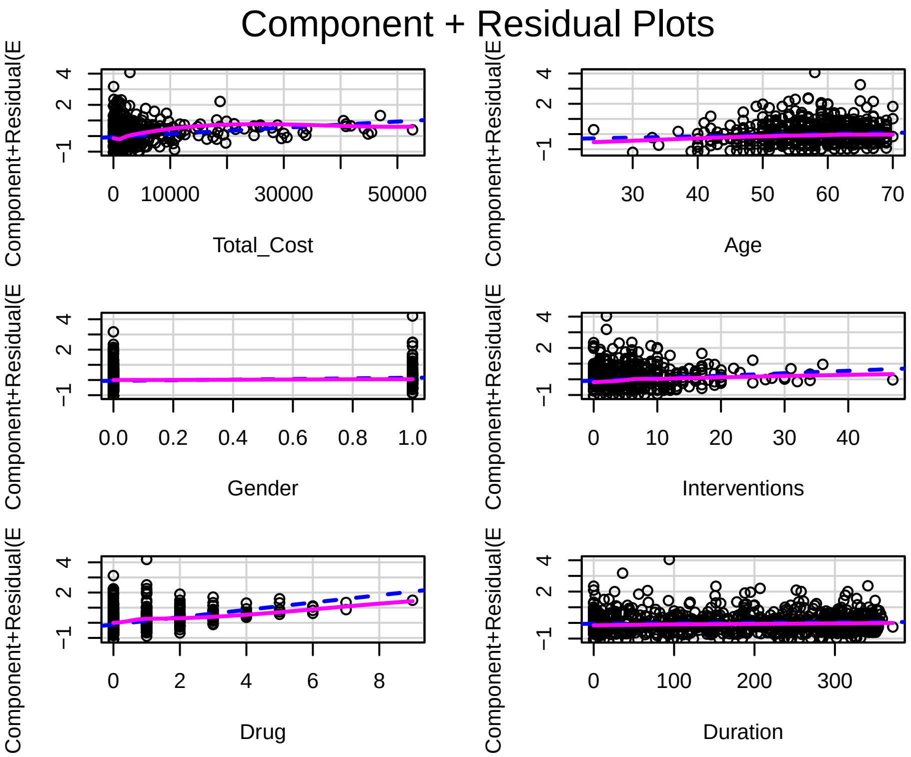


- **QQ Plot**
  - **KS Test**: $p = 0.613$ suggests that the model's predicted values do not significantly deviate from the expected distribution, presenting no evident concerns regarding the goodness-of-fit of the model.

  - **Dispersion Test**: $p = 0.464$ does not present any issues related to overdispersion.
  - **Outlier Test**: $p = 0.309$, there doesn't appear to be any significant outliers impacting the model fit.

- **Residuals vs Predictions**
  - Prominent deviations in the quantiles, along with a substantial combined adjusted quantile test, indicate potential problems with the model fit across a range of predicted values.

Partial residual plots

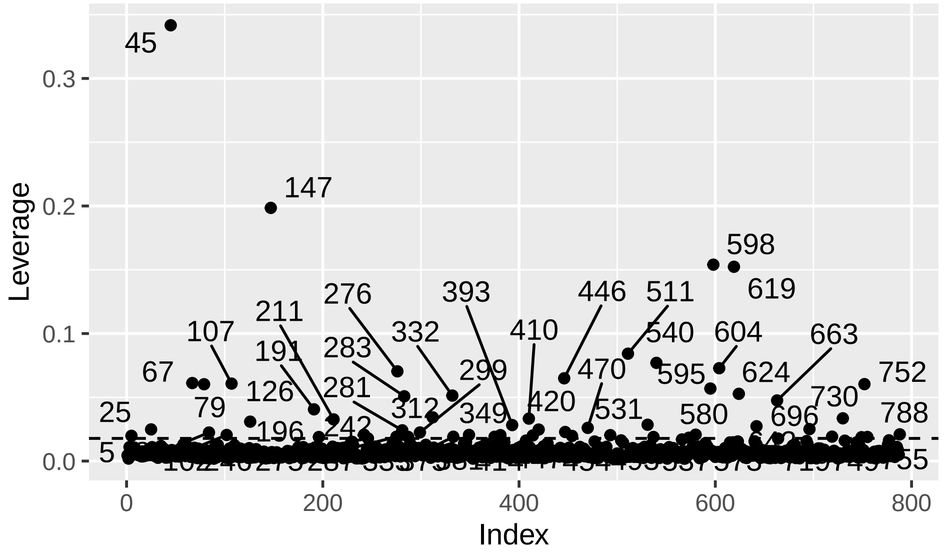


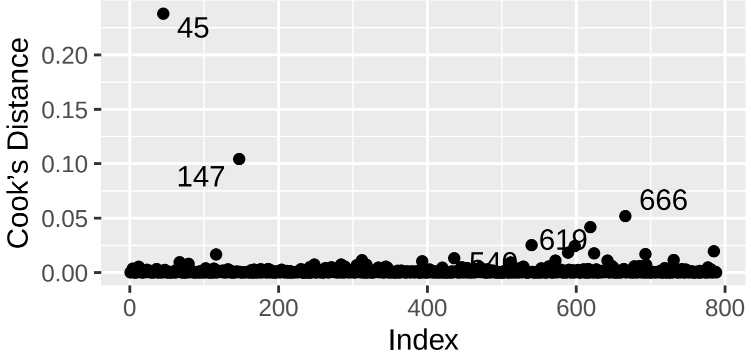
The alignment of broken and solid lines in the partial residual plots for all variables indicates a generally linear relationship with the response, while holding other predictors constant. The assumptions of linearity appear largely satisfied.

```
# Calculating and plotting leverage
hatvalues <- hatvalues(final_model_nb)  # Calculate hat values (leverage)
cutoff_leverage_nb <- 2*length(coef(final_model_nb))/length(final_model_nb$fitted.values)  # Typical le
cat("Leverage cutoff:", cutoff_leverage_nb,"\n")
```

Cook's distances and Leverage statistics

```
## Leverage cutoff: 0.0177665
```





Numerous observations exhibit high leverage, surpassing the established leverage cutoff of 0.0178. Despite this, no observations present large Cook's distance values (using a threshold of 1), indicating the absence of points with substantial influence on the model. This suggests that while certain data points have pronounced leverage, they don't significantly impact the model's fit and predictive capability. Observations 45 and 147 may be treated carefully.

Question 3

A: Logistic Distribution PDF

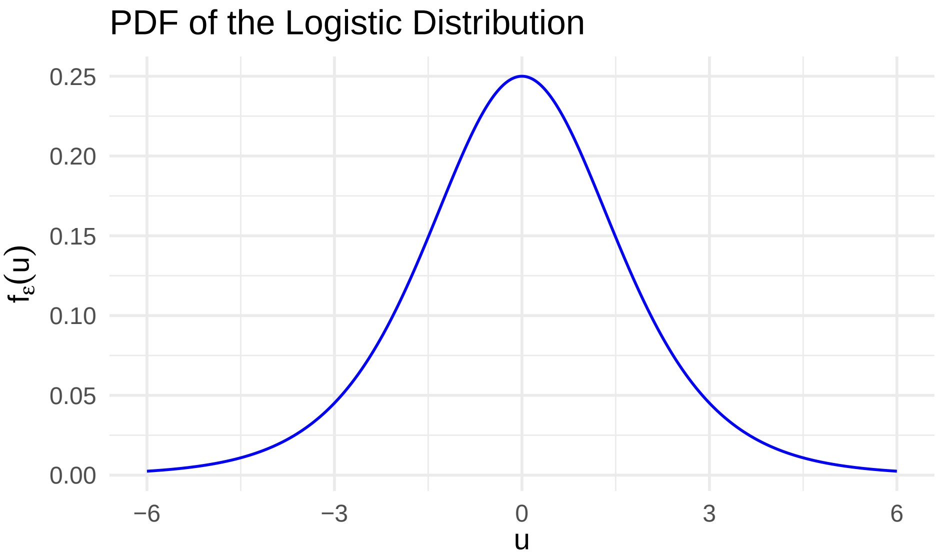
I first consider that the noise component ϵ_i is a (standard) logistic distribution where the cumulative distribution function $F_{\epsilon_i}(u) = P(\epsilon_i \leq u)$ is defined via,$F_\epsilon(u) = \frac{1}{1+e^{-u}}$

The PDF (probability density function) of the logistic distribution can be derived from the CDF (cumulative distribution function) as follows: $f_\epsilon(u) = \frac{d}{du} F_\epsilon(u)$ Where: $F_\epsilon(u) = \frac{1}{1+e^{-u}}$

Calculating the derivative with respect to u yields the PDF: $f_\epsilon(u) = \frac{e^{-u}}{(1+e^{-u})^2}$

Let's plot the PDF of the logistic distribution:

```r
# Define the function for the pdf of logistic distribution
logistic_pdf <- function(u) {
  exp(-u) / (1 + exp(-u))^2}
# Create a sequence of u values
u <- seq(-6, 6, by=0.01)
# Create a data frame for plotting
df <- data.frame(u = u, pdf = logistic_pdf(u))
# Create the ggplot object
p <- ggplot(df, aes(x = u, y = pdf)) +
  geom_line(color = "blue") +
  labs(title = "PDF of the Logistic Distribution",
       x = expression(u),
       y = expression(f[epsilon](u))) +
  theme_minimal()
# Display the plot
print(p)
```



B: Symmetry of Logistic Distribution

Given the cumulative distribution function (CDF) of the logistic distribution as:

$$F_\varepsilon(u) = \frac{1}{1 + e^{-u}}$$

To establish the symmetry about 0, I need to demonstrate that:

$$F_\varepsilon(u) = 1 - F_\varepsilon(-u)$$

I'll substitute the expression for $F_\varepsilon(u)$ into the above equation and verify whether the relation holds:

$$\frac{1}{1 + e^{-u}} = 1 - \frac{1}{1 + e^{u}}$$

Computing the right-side expression:

$$\frac{1}{1+e^{-u}} = \frac{1+e^u-1}{1+e^u} = \frac{e^u}{1+e^u}$$

So now I have:

$$\frac{1}{1+e^{-u}} = \frac{e^u}{1+e^u}$$

To check the equality, let's work with the expression on the right:

$$1 = (1+e^{-u}) \times \frac{e^u}{1+e^u}$$

$$1 = \frac{e^u+1}{1+e^u}$$

Since the numerator and the denominator are the same, the fraction equals 1:

$$1 = 1$$

Thus, successfully demonstrated the symmetry of the logistic distribution about 0, since

$$F_\varepsilon(u) = 1 - F_\varepsilon(-u)$$

C: Derivation of Probabilistic Relation

Having the latent variable representation:

$$\Psi_i = x_i^\top \beta + \varepsilon_i$$

And:

$$Y_i = \begin{cases} 1, & \text{if } \Psi_i \geq 0 \\ 0, & \text{if } \Psi_i < 0 \end{cases}$$

To find the probability that $Y_i = 1$ given $X_i = x_i, \beta$, I express this probability in terms of the latent variable:

$$P(Y_i = 1 | X_i = x_i, \beta) = P(\Psi_i \geq 0 | X_i = x_i, \beta)$$

Substituting the expression for Ψ_i:

$$P(Y_i = 1 | X_i = x_i, \beta) = P(x_i^\top \beta + \varepsilon_i \geq 0 | X_i = x_i, \beta)$$

Rearranging to isolate the noise term:

$$P(Y_i = 1 | X_i = x_i, \beta) = P(\varepsilon_i \geq -x_i^\top \beta)$$

Given that ε_i follows a logistic distribution, I use the CDF to express this probability:

$$P(Y_i = 1|X_i = x_i, \beta) = 1 - F_\varepsilon(-x_i^\top \beta)$$

Where the CDF of the logistic distribution is defined as:

$$F_\varepsilon(u) = \frac{1}{1 + e^{-u}}$$

Substituting this definition and simplifying:

$$P(Y_i = 1|X_i = x_i, \beta) = 1 - \frac{1}{1 + e^{x_i^\top \beta}} = \frac{1}{1 + e^{-x_i^\top \beta}}$$

D: Defining a GLM and Specifying the Link Function

A Generalized Linear Model (GLM) is defined as a model where the random component is a member of the exponential family, and the systematic component is linear in the parameters. The link function provides the relationship between the linear predictor and the mean of the distribution function.

Given (2):

$$P(Y_i = 1|X_i = x_i, \beta) = \frac{1}{1 + e^{-x_i^\top \beta}}$$

This implies a relationship between the linear predictor (the systematic component) and the mean of the distribution function (the expected value of the response variable). Here, the linear predictor is:

$$\eta_i = x_i^\top \beta$$

And the mean of the Bernoulli distributed response variable Y_i is:

$$\mu_i = E[Y_i|X_i = x_i] = P(Y_i = 1|X_i = x_i, \beta)$$

The relationship between η_i and μ_i is established by the link function $g(\cdot)$, such that:

$$g(\mu_i) = \eta_i$$

In this case, from (2), I can identify that the link function $g(\cdot)$ is the logit function, defined as:

$$g(\mu) = \log\left(\frac{\mu}{1 - \mu}\right)$$

Because if I take the logit of the probability in (2), I have:

$$g(\mu_i) = \log\left(\frac{\mu_i}{1 - \mu_i}\right) = \log\left(\frac{P(Y_i = 1|X_i = x_i, \beta)}{1 - P(Y_i = 1|X_i = x_i, \beta)}\right) = x_i^\top \beta = \eta_i$$

Thus, with the Bernoulli distribution for the response variable and the logit as the link function, defining a logistic regression model, which is a special case of a GLM.

E: Probability Expression for Normal Distribution Noise

Given the latent variable Ψ_i and considering the noise component ε_i to be normally distributed, i.e., $\varepsilon_i \sim \mathcal{N}(0,1)$

let's derive the probability expression.

(1): $\Psi_i = x_i^\top \beta + \varepsilon_i$ with,

$$Y_i = \begin{cases} 1, & \text{if } \Psi_i \geq 0 \\ 0, & \text{if } \Psi_i < 0 \end{cases}$$

find the expression for: $P(Y_i = 1|X_i = x_i, \beta)$

Using (1), express the probability as: $P(Y_i = 1|X_i = x_i, \beta) = P(\Psi_i \geq 0|X_i = x_i, \beta)$

Substituting the expression for Ψ_i from (1):

$$P(Y_i = 1|X_i = x_i, \beta) = P(x_i^\top \beta + \varepsilon_i \geq 0|X_i = x_i, \beta)$$

Now, rearranging to isolate the noise term:

$$P(Y_i = 1|X_i = x_i, \beta) = P(\varepsilon_i \geq -x_i^\top \beta)$$

Given that ε_i follows a normal distribution, use the Cumulative Distribution Function (CDF) of the standard normal distribution, denoted as Φ, to express this probability:

$$P(Y_i = 1|X_i = x_i, \beta) = P(\varepsilon_i \geq -x_i^\top \beta) = 1 - \Phi(-x_i^\top \beta)$$

Where Φ is defined as:

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{z^2}{2}} dz$$

Utilize the symmetry property of the standard normal distribution: $1 - \Phi(u) = \Phi(-u)$

Hence, the expression becomes:

$$P(Y_i = 1|X_i = x_i, \beta) = \Phi(x_i^\top \beta)$$

This gives the probability of class membership with the noise term ε_i following a standard normal distribution.

F: Identifying GLM and Specifying the Link Function

In part e), derived the probability expression:

$$P(Y_i = 1|X_i = x_i, \beta) = \Phi(x_i^\top \beta)$$

With:

- **Random Component:** Y_i, representing a binary outcome and hence, following a Bernoulli distribution.
- **Systematic Component:** The linear predictor, which is given as $x_i^\top \beta$.
- **Link Function:** Utilizing the cumulative distribution function (CDF) of the standard normal distribution, Φ, the link function in this context is identified as the **probit link function**.

Therefore, formalizing this in the context of a Generalized Linear Model (GLM), I express:

$$\Phi^{-1}(P(Y_i = 1|X_i = x_i, \beta)) = x_i^\top \beta$$

Here, Φ^{-1} denotes the probit function, which is the inverse of Φ, acting as the link that correlates the mean of the random component, $\mu_i = E[Y_i]$, with the linear predictor in the systematic component.

Thus, the generalized linear model (GLM) identified from part e) employs the **probit link function**, associating the linear predictor with the Bernoulli-distributed response variable through the CDF of the standard normal distribution.

AI Use Acknowledgement

I have used Chat GPT and Grammarly to revise my writing, debugging and creating Latex. It helped make the report more professional and concise and allowed me to use my time more effectively.