Abstract

This paper sought to build an appropriate generalized linear model to determine the covariates that are associated with whether a high risk neonatal surviving infant will be re-admitted to a neonatal unit within a year based on questionnaire data. The objective of the project was to find a model that only includes statistically significant predictors. Exploratory data analysis included scatterplots, histograms, and the computation of Kendall's Tau. The model was built by first simulating forward model selection for the main effects models via likelihood ratio tests, and then considering all potential interaction terms afterwards. The final model was a binomial generalized linear model with a complementary log-log link function consisting of the covariates los, emp.f, sex, and accom. The model was examined for multicollinearity, unexpected characteristics of the residuals, influential points, and goodness of fit using the Hosmer-Lemeshow test. The results overall suggested a good fit and that model assumptions were reasonable, however there seemed to be many influential points.

Introduction

The Langley study provided a dataset to identify characteristics of high risk neonatal survivors, including variables such the employment status of the mother (Langley et al, 2002). The response variable, re.ad, represented whether an infant will be re-admitted to a neonatal unit within a year. This paper will attempt to build a model to determine which covariates are associated with this response variable. The variables in the dataset include the provision of community neonatal services, the size of the neonatal unit, the gestation period, the birthweight in kg, whether the father or partner of the mother is employed, and whether the mother is employed. Other variables also include the age mothers left full time education, whether the infant is re-admitted to a neonatal unit within a year, the total length of the stay in the hospital in log(days) during an initial admission, the sex of the baby, as well as whether the mother owns the accommodation she occupies. This paper will explore the capabilities of modelling the response variable using generalized linear models. Generalized linear model can provide the additional flexibility compared to regular linear regression models, including transformations on the response variable made via a link function. Link functions allow the model's predicted values to match the properties of re-ad. The paper will cover the methods for exploratory analysis, model building, and model assessment.

Methods

Visualisations for Exploratory Analysis

Histograms and scatterplots were created for the response variable along with several other variables provided in the dataset. Histograms were created to depict the types of values of the response variable and the potential covariates. Scatterplots provided a preliminary understanding of the relationships between the response variable and the potential covariates. Exploratory analysis informed whether a generalized linear model is suitable for the data and which family would be appropriate for model building.

Significance Testing and Kendall's Tau for Exploratory Analysis

After choosing a model type, a model consisting of one variable was created for each variable to provide an understanding of which variables would be significant in the model. This was accomplished by inspecting whether the p-values for each covariate < 0.05. One may expect one or more of the significant variables to appear while building the model. Hence, Kendall's Tau was also computed to inspect for associations between some of the variables. Kendall's Tau was used over other correlation measures since most variables in the dataset were categorical and had tied values (Kendall's Rank Correlation, 2019). That is, the variables have the same values. The computation of Kendall's Tau indicated pairs of variables that had a strong association with each other, and hence could potentially cause multicollinearity if both were included in the model.

Building the Main Effects Model

The method for model building first focused on determining the most appropriate main effects model by comparing the results of simulating forward model selection (Lee, 2019). There were multiple possible criteria to decide which model was considered preferable. While AIC was seen to be effective for predictive accuracy and BIC was known to be used for finding a simple interpretable model, the likelihood ratio test was identified to be the most useful for assessing whether a given effect or set of effects is significant (Brewer & Butler, 2017). Since the aim of the project was to determine the covariates associated with the re-admission to the hospital, the likelihood ratio tests were used as the basis to add or drop variables. It was noted for implementation that Chi-squared tests are equivalent to the likelihood ratio test in the context of generalized linear models according to R documentation (add1: Add or Drop

All Possible Single Terms to a Model, 2019). The method of forward selection started with the empty model and added the covariate with the smallest statistically significant p-value until all remaining variables not in the model were statistically insignificant. That is, the process terminated when all other variables had p-values greater than 0.05. Although the likelihood ratio tests took precedence in model selection criterion, the add1 function in R also provided AICs by default. Since BICs provide a measure of simplicity and ease of explanation in models compared to AIC, the BICs were provided as a secondary metric to the p-value instead of the default AICs. Analysis would be conducted on whether the model building decisions made using the likelihood ratio test conflicted with the measures of BIC.

Consideration of Interaction Terms
After building the main effects model, each interaction term was tested for significance as an addition to the current model individually. If there were no significant p-values for interaction terms, no interaction terms were added. If any interaction terms were significant without compromising the statistical significance of the main effects, then they were added into the model and the rest of the second-order interaction terms were examined again. For example, even if los and bwt were statistically significant in the main effects model, but the addition of the interaction term los:bwt resulted in one of these three effects possessing a p-value less than 0.05, then los:bwt would not have been added.

Choice of Link Function
After adding all main effects and interaction terms, the resulting deviances were compared for different link functions. The binomial family of generalized linear models could use a logit, probit, complementary log-log, or cauchit link function. Four different models with the same main effects and interaction terms were fitted with these four link functions. Since all the models here had the same number of parameters, the deviance was compared between the four models to determine which link function would be the best fit.

Model Selection
Based on the previous steps, a final model was selected with the chosen link function. The summary statistics for the model were examined to develop and understanding of the covariates' association with the response variable and ensure all covariates in the model were significant.

Examination for Multicollinearity
The examination for multicollinearity involved examining Kendall's Tau values that were computed as part of exploratory analysis. The model was examined to identify whether covariates in the model were correlated. If two covariates in the model were heavily correlated, then the model was to be re-examined for multicollinearity.

Examination of Residuals
Pearson as well as deviance residuals were examined for their mean and variance. These residuals were compared to the expected value of one and zero respectively. The residuals were plotted against the covariates in the model. A smoothed mean was plotted to verify whether the linearity assumption between the link function and continuous covariates is reasonable. The means of the Pearson residuals were inspected for different values of the categorical covariates in the model.

Examination of Influence Measures
Influence measures were computed to identify influential points and the number of influential points were computed for these influence measurements. The influential points were determined based on the values of their standardized residuals, dffit residuals, Cook's distances, leverages, and covariance ratios. Standardized residuals were considered abnormally large if they exceeded 3 or were less than -3. The dffit residual described the impact on the model if the point were removed. A point was considered influential according to the dffit residual if the dffit residual was greater than $2\sqrt{p/n}$.
A point was influential according to Cook's distance if Cook's distance was greater than $4/(n-p)$. A point had high leverage if it exceeded $2(p/n)$. A point was influential according to the covariance ratio if the covariance ratio exceeded $1 \pm \frac{3p}{n-p}$.

Application of the Hosmer-Lemeshow Test
The Hosmer-Lemeshow test was considered useful for logistic regression and the assessment of goodness-of-fit (Lee, 2019). It involved splitting the observations into G number of groups, which was

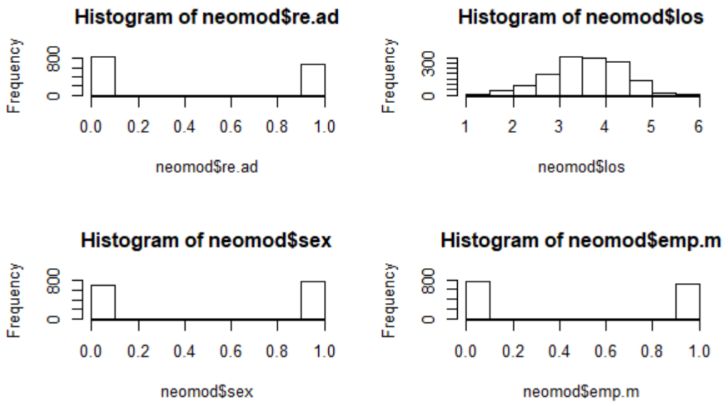
typically five or ten groups. The general guideline for selecting the number of groups was choosing $G > p + 1$ where p is the number of parameters and G is the number of groups. If $\chi^2_{HL} > \chi^2_{G-2}(0.95)$, meaning the p-value < 0.05, then this measure indicated that the model has poor fit.

Results

Visualizations for Exploratory Analysis

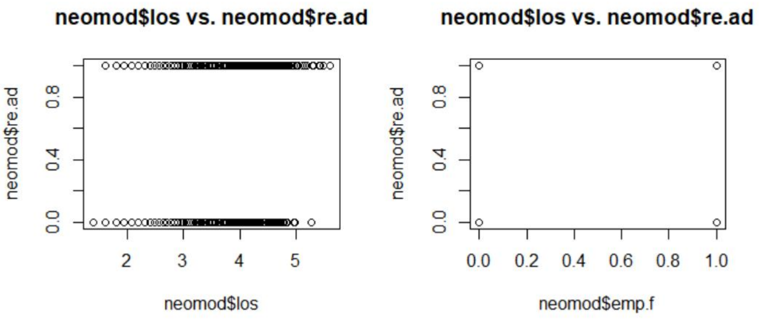
The visualizations in this section include four histograms and two scatterplots. The full code can be found in Appendix A.

The following histograms display the frequency of the observations in four of the variables in the dataset, including the response variable re.ad.



As seen in the top left plot, the response variable re.ad only seemed to take on values of zero or one. Some of the covariates in the data such as sex or emp.m were categorical, while los was continuous.

The following two scatter plots show bwt vs. re.ad and los vs re.ad.



The scatterplots confirmed that re.ad only took values of zero or one, and from the scatterplots, it became evident that a regular multiple linear regression model would not be suitable for the data.

From the above visualizations, the response variable re.ad was identified to be binary as its values took on one of {0,1}. The model for binary data was identified to be a special case of the binomial family. Additionally, the logit link function was the canonical link function for the binomial generalized linear model and provides ease of interpretation (Lee, 2019). Therefore, a binomial family with a logit link function was deemed appropriate.

Generalized linear models worked with the binomial distribution as it was a member of the exponential family (Lee, 2019), If Y was a random variable for a binomial distribution with probability of success π and n trials, then its p.d.f. could be written as

$$f(y, \pi) = \exp\,(y \log \frac{\pi}{1-\pi} + n \log(1-\pi) + \log \binom{n}{y}$$

This equation was further rewritten as a member of the exponential family of distributions

$$f(y;\theta,\varphi) = \exp\left\{\frac{(y\theta - b(\theta))}{a(\varphi)} + c(y,\varphi)\right\}$$

where $\theta = log\frac{\pi}{1-\pi}, a(\varphi) = \varphi = 1,\ b(\theta) = -n\log\{1+\exp(\theta)), \text{and } c(y,\varphi) = \log\binom{n}{y}$.

The canonical link function for the binomial model or logit function, which was defined as $g(x) = \log(\frac{x}{1-x})$. The logit function was the default choice and the decision regarding the choice of link function would have been finalized after adding all significant variables.

Significance Testing and Kendall's Tau for Exploratory Analysis

Ten models were constructed with each of the variables in the dataset using a binomial and default logit link function. A summary of the variables and their p-values can be found in the table below. Refer to Appendix B1 for the R code.

Variable	p-value from fitting
cns	0.599818
size	0.196907
gest	8.624e-10
bwt	1.149e-06
emp.f	6.329e-05
emp.m	0.001555
edu	0.708204
los	<2.2e-16
sex	0.0129
accom	2.683e-05

The variables gest, bwt, emp.f, emp.m, los, sex, and accom seemed to be statistically significant. Then, correlation tests were conducted to compute Kendall's tau between these variables. While most of the Kendall's tau values computed for variables resulted in a statistically significant p-value, suggesting that there was a relationship between most covariates, most of the relationships were not identified to be very strong since most of the actual estimates of correlation were not very large.

The following table shows some notable relationships for Kendall's Tau. Refer to Appendix C for the R code for computing Kendall's Tau.

Variables	Value of Kendall's Tau Estimate
Emp.f and accom	0.3839189
Emp.m and accom	0.2391108
Bwt and los	-0.5577403
Gest and los	-0.6227063
Gest and bwt	0.6473643

The most significant relationships were identified to be between the variable los, gest, and bwt. Hence one may need to take extra precautions if there was a combination of these three variables in the main effects model to be built.

Building the Main Effects Model

The result of forward selection was re.ad ~ los + emp.f + sex + accom.

The summary of the results can be found below. Refer to Appendix D for the full execution of the method.

Step	Model	Variable Added	p-Value of Variable Added	BIC
1 (start)	re.ad~1	los	< 2.2e-16	2049.7
2	re.ad~los	emp.f	0.0002667	1970.0
3	re.ad~los+emp.f	sex	0.006008	1964.0
4	re.ad~los+emp.f+sex	accom	0.02189	1963.8
5 (stop)	re.ad~los+emp.f+sex+accom			1965.8

It was noted that the outcomes of using the likelihood ratio test directly conflicted with the measure of BIC. In particular, the BIC was lower for the model without accom. Hence, the resulting model was not the best model according to BIC. However, not including the variable accom would have meant leaving out a statistically significant variable, as it had a p-value of $0.02 < 0.05$.

Consideration of Interaction Terms

Since there were four covariates in the main effects model, there were 6 possible interaction terms. Hence 6 models were constructed, each with one of the 6 possible interaction terms. The p-value was assessed, and the results showed that none of the interactions were worth adding to model. Specifically, the interaction emp.f:sex appeared significant, however the p-value of sex was compromised as it increased to 0.8463. The other interactions los:accom, sex:accom, emp.f:sex, los:emp.f, and sex:los were all found to be insignificant with p-value < 0.05. Hence, no interaction terms were added to the model. Refer to Appendix E for the implementation and summary statistics of the models.

Choice of Link Function

The table below shows the resulting deviance of using each of the link functions in the binomial family. For the full summary statistics of the models with each of the four link functions, refer to Appendix F.

Link function	Deviance
Logit	1929.3
Probit	1929.9
Complementary log-log	1925.2
Cauchit	1926.5

Since the models had the same number of parameters, the deviances were comparable. The deviance among the link functions were quite similar, but the complementary log-log link function had the lowest deviance. Switching from the logit to the complementary log-log link would decrease the deviance by 4.3. Hence, a complementary log-log link was used for the final model. The complementary log-link function specifies that $\log\{-\log(1-\pi_i)\} = \eta_i$ where η_i is the linear predictor.

Model Selection

The final model was a binomial generalized linear model with a complementary log-log link function. The covariates in the model were los, emp.f, sex, and accom. These four covariates represented the total length of the stay in the hospital in log(days) during the initial admission, the employment status of the mother, sex of the baby, and whether the occupier of the accommodation was the owner. The following table summarizes the covariates, their estimates of beta, as well as the p-value of the covariates. Refer to Appendix F for the full summary.

Covariate	Estimate	p-value
Intercept	-2.09213	<2e-16
los	0.52118	<2e-16
emp.f	-0.32757	0.00938
sex	0.22112	0.00608
accom	-0.21767	0.02312

The complementary log-log link function satisfies the following:

$$\pi_i = 1 - e^{-e^{\eta_i}}$$

where $\eta_i = \beta_0 + \beta_1 \times los + \beta_2 \times emp.f + \beta_3 \times sex + \beta_4 \times accom$

Also, with a complementary log-log link function,

$$e^{\beta_j} = \frac{\log(1-\pi_1)}{\log(1-\pi_0)}$$

where π_1 is the probability of not being re-admitted, π_0 is the probability of re-admitted.

β_j could be thought of as a log-hazard ratio where π_i represents the hazard of an event. e^{β_j} is the ratio of the log of the probability of being re-admitted to the log of the probability of not being re-admitted.

Using the estimate of beta for los, $e^{0.52118} = 1.68401$, which is the hazard ratio. That is, the log of re-admission would have increased about 168% per unit increase in los, which was the log of the length of the initial stay in days. The result is 1.68 * log(days), which can be rewritten as $\log(days^{1.68})$ in the linear predictor.

Using the estimate of beta for emp.f, $e^{-0.32757} = 0.72067$. Hence the log of the probability of re-admission would have decreased about 72% if the mother is employed instead of unemployed.

Similarly, using the estimate for sex, $e^{0.22112} = 1.24747$. Then the log of probability of re-admission would have increased by 125% if the infant was male instead of female.

For accom, $e^{-0.21767} = 0.804391$. The log of the probability of re-admission would have decreased about 80% if the mother was the owner of the accommodation instead of not being the owner.

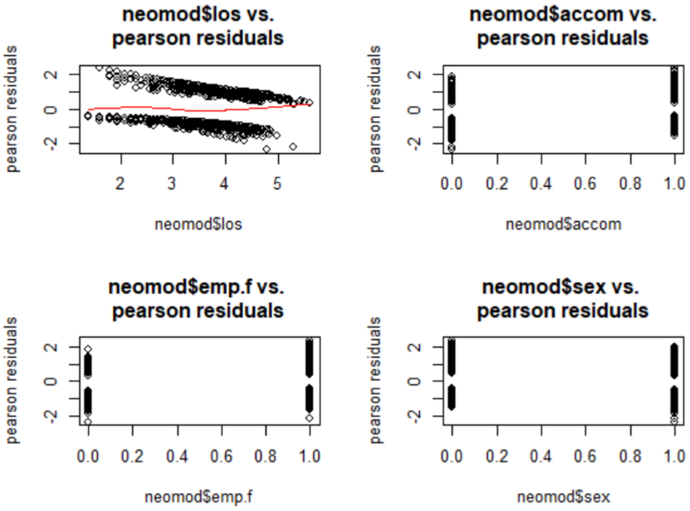
Examination for Multicollinearity

The most notable association between covariates within the proposed model had a Kendall's Tau estimate of 0.38 between emp.f and accom. While there could have been a real some relationship between the employment status of a mother and whether they possess ownership of their accommodation, the estimate of 0.38 did not seem strong enough to pose any multicollinearity issues within the model. The other associations between covariates in the model were identified to be smaller and hence did not appear to cause any issues with multicollinearity. Refer to Appendix C for the remaining correlations.

Examination of Residuals

The model's residuals were expected to have an approximate mean of zero and variance of one. Deviance residuals had a mean of -0.03682001 and variance of 1.293322. Pearson residuals had a mean of 0.002195945 and variance of 1.005389. Hence, their means and variances were quite similar, and both seemed to mostly align with the expectation. Refer to Appendix G for the full calculations.

The plots below show the Pearson residuals plotted against the covariates used in the model.



The first plot in the top left hand corner displayed a scatterplot of the continuous variable los against the Pearson residuals. The smoothed mean, shown by the red line, appeared relatively constant although there was some minimal variation with los. However, overall the variation of residuals with los seems too insignificant to warrant a transformation to the los covariate. Hence no transformation needed to be applied to los and the linearity assumption between the complementary log-log link function and covariate los seemed to hold reasonably.

The categorical variables, accom, emp.f, and sex all took on values of either zero or one. It appears that there might have been a slight difference in the means of the residuals for different levels. To provide a quantitative measure of this difference, the following table shows for every covariate the absolute difference between the mean of the Pearson residuals when the value of the covariate equals one and the mean of the Pearson residuals when the value of the covariate equals zero. Refer to Appendix G for the full calculation.

Covariate	Difference in Means for Covariate Values of 0 and 1
accom	0.003085645
emp.f	0.0009267997
sex	0.004080338

The difference in means seemed to be very small as they were approximately zero. Hence, there seemed to be no need to take alternative measures to improve the model fit. The linearity assumption seemed to overall hold for the model's covariates.

Examination of Influence Measures

The influence measures detected many influential points, although the exact number varied greatly depending on the measure used. There were no abnormally large standardized residuals and seven points were influential according to the dffit residual. However, according to Cook's distance, 35 points were influential. More importantly, 235 points had high leverage and 120 points were influential according to the covariance ratio. Some of the influential points included the 56th, 218th and 242nd observations. However, the data did not seem to have any evidently erroneous entries and the residuals do not seem abnormally high. For example, there did not seem to be any extreme values for continuous covariates such as los. Since the measures of influence gave a very different number of influential points and the data did not appear to have any evidently erroneous entries, all points were retained in the dataset. Refer to Appendix H for the R code that computed these measures.

Application of the Hosmer-Lemeshow Test

The results of the Hosmer-Lemeshow test indicated no evidence suggesting that the model was ill-fitting. Since there were $p = 4$ parameters in the model, $p + 1 = 5$. Since the general practice was to use either $G = 5$ or $G = 10$ groups and the guideline for choosing G was $G > p + 1$, the test was conducted with $G = 10$ groups. The result from final model returned a p-value of $0.33 > 0.05$. Hence, there was no evidence against the hypothesis that the model was well-fitting, however it was noted that this did not necessarily mean that this model is the best. Refer to Appendix I for the full calculation.

Discussion

One concerning aspect of the model was the presence of many influential points. Further investigation could be conducted solely on the influential points. If there was a pattern in the influential points, the model may need to be revised. 235 high-leverage points is a large number of points in comparison to the 1488 datapoints in the original data. However by keeping these points, it might have been possible that the model was too heavily influenced by 16% of the data, but it also seemed impractical to remove 16% of the data that do not necessarily consist of outliers or improperly recorded datapoints.

Depending on the purpose of the exploration, alternative models could have been fitted. If the purpose was to predict whether an infant will be re-admitted to the hospital within the next year using future data, a variable such as los could only be collected post-partum. The prediction was time-specific as it describes the probability of re-admission within a year. If there was to be decision-making, it will be subject to time constraints and so los might not have been the best choice of variable in terms of availability. While this covariate was statistically significant in the model, based on Kendall's Tau it possessed a relationship with gest and bwt as well. Hence if los was not available, then perhaps gest or bwt might have been a good proxy variable instead.

Alternative models could also have been fitted through different methods. While the method used in this exploration simulated forward selection, different orders to stepwise selection could have impacted the result. Using backwards or bidirectional elimination would have impacted the order in which variables were added or removed, as well as which covariates would be in the final model. Another source of model deviation was the criterion used for model selection. In this exploration, it was the likelihood ratio test, which was equivalent to performing the Chi-squared test for generalized linear models. However, using AIC or BIC could have provided alternative results. Notably, the inclusion of the covariate accom was a decision that conflicted with the measures of BIC for the models. BIC suggested that the model should not have accom. Performing the entire procedure using BIC would have for certain led to a different result.

Limits to the conclusion included generalization to the greater population. Errors in the sampling process or data collection could have skewed the result. For example, the data may be susceptible to sampling error. The study indicated that questionnaires were sent to the parents of 3367 eligible infants and excluded any cases of multiple births, infants who had died or had severe congenital abnormalities. Additionally, not all eligible participants responded and postal questionnaires were used. Therefore, it was possible that the characteristics of the excluded neonatal survivors differ from the recorded data at hand. While there were 1488 datapoints in the dataset, however, conducting the same method on a larger dataset may have resulted in different covariates being significant instead. Hence, the conclusions generated by this report could only be applied to a subset of high-risk neonatal survivors from regions with similar characteristics as the original study location, England and Wales (Langley et al, 2002). Furthermore, even if the true underlying distribution was the same as the model, that is, the model determined by this exercise was the "perfect" model, there would have still been variation or error in the results.

Stepwise model selection was heavily criticized as the procedure was easy to explain and compute but needed to be applied with caution (Smith, 2012). In this exploration, some precautions were taken to scrutinize the validity of the model, including conducting a variety of exploratory analysis and examining the results against model diagnostics. However, these tests might not have yielded a perfect result and other tools may have differing results as well. One limitation of stepwise selection methods was that it only searched about ten percent of all possible models. Since there were ten variables, there were about $2^{10}=1024$ possible main effects models but stepwise procedures would only have searched about $10^{2}=100$ of them (Lee, 2019). The method used in this exploration added the most significant variable at each step but did not consider adding variables in different orders. It was possible that adding covariates in an alternate order could have lea to a completely different model. Other models could be also have been adequate or better than the one found in this exploration. While simulating forward model selection did not provide an exhaustive search of all possible models, it did attempt to only include the most significant variables in an efficient manner. Notably, one of the benefits of this method was that it terminated rather quickly in the fifth step in this scenario.

Conclusion

This paper concluded that a parsimonious model for the response variable re.ad would be a generalized linear model of the binomial family with a complementary log-log link function containing the covariates los, emp.f, sex, and accom. These four covariates were found to be significant and associated with the response variable re.ad. Multicollinearity did not appear to be an issue, while residuals appeared to conform well to their expected means and variances. The linearity assumptions between the complementary log-log link function and covariates seemed reasonable as well. The results of the Hosmer-Lemeshow test also proved satisfactory. However, some concern should be given to the influence measures as it appeared that there were many influential points. Having a high number of points that had high leverage may have led the model to conform to the characteristics of only a small proportion of the actual data.

References

add1: Add or Drop All Possible Single Terms to a Model. (2019). Retrieved from R Package Documentation: https://rdrr.io/r/stats/add1.html

Brewer, M. J., & Butler, A. (2017, June 16). *Model Selection and the Cult of AIC.* Retrieved from UAB Barcelona: http://sct.uab.cat/estadistica/sites/sct.uab.cat.estadistica/files/uab_june_2017.pdf

Kendall's Rank Correlation. (2019). Retrieved from StatisticsDirect Limited: https://www.statsdirect.com/help/nonparametric_methods/kendall_correlation.htm

Lee, C. (2019). MATH452/552 Generalized Linear Models **-** Lancaster University. Lancaster University.

Smith, M. K. (2012, January 20). *Problems with Stepwise Model Selection Procedures*. Retrieved from Department of Mathematics at the University of Texas at Austin: https://web.ma.utexas.edu/users/mks/statmistakes/stepwise.html

Langley D, Hollis S, Friede T*, et al*. Impact of community neonatal services: a multicentre survey. *Archives of Disease in Childhood - Fetal and Neonatal Edition* 2002; **87:** F204-F208.

Appendix

Appendix A – Visualisations for Exploratory Analysis

```
> # Data for GLM
> neomod<-read.table(dir)
> # Code categorical variables as factors
> neomod$gest<-as.factor(neomod$gest)
> neomod$edu<-as.factor(neomod$edu)
>
> # Exploratory analysis
> # Histograms
> par(mfrow=c(2,2))
> hist(neomod$re.ad)
> hist(neomod$los)
> hist(neomod$sex)
> hist(neomod$emp.m)
>
> # Scatterplots
> par(mfrow=c(1,2))
> plot(neomod$los,neomod$re.ad,main="neomod$los vs. neomod$re.ad")
> plot(neomod$emp.f,neomod$re.ad,main="neomod$los vs. neomod$re.ad")
```

Appendix B – Significance Testing for Exploratory Analysis

```
> # Test all covariates individually
> model1<-glm(re.ad~cns,family=binomial(link="logit"),data=neomod)
> model2<-glm(re.ad~size,family=binomial(link="logit"),data=neomod)
> model3<-glm(re.ad~gest,family=binomial(link="logit"),data=neomod)
> model4<-glm(re.ad~bwt,family=binomial(link="logit"),data=neomod)
> model5<-glm(re.ad~emp.f,family=binomial(link="logit"),data=neomod)
> model6<-glm(re.ad~emp.m,family=binomial(link="logit"),data=neomod)
> model7<-glm(re.ad~edu,family=binomial(link="logit"),data=neomod)
> model8<-glm(re.ad~los,family=binomial(link="logit"),data=neomod)
> model9<-glm(re.ad~sex,family=binomial(link="logit"),data=neomod)
> model10<-glm(re.ad~accom,family=binomial(link="logit"),data=neomod)
>
> summary(model1)

Call:
glm(formula = re.ad ~ cns, family = binomial(link = "logit"),
    data = neomod)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.090  -1.090  -1.067   1.268   1.292

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.20972    0.07092  -2.957  0.00311 **
cns         -0.05497    0.10479  -0.525  0.59986
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2042.4  on 1487  degrees of freedom
Residual deviance: 2042.1  on 1486  degrees of freedom
AIC: 2046.1

Number of Fisher Scoring iterations: 3
```

```
> summary(model2)

Call:
glm(formula = re.ad ~ size, family = binomial(link = "logit"),
    data = neomod)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.098  -1.098  -1.039   1.259   1.323

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.33573    0.09426  -3.562 0.000368 ***
size         0.14590    0.11324   1.288 0.197606
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2042.4  on 1487  degrees of freedom
Residual deviance: 2040.7  on 1486  degrees of freedom
AIC: 2044.7

Number of Fisher Scoring iterations: 4

> summary(model3)

Call:
glm(formula = re.ad ~ gest, family = binomial(link = "logit"),
    data = neomod)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3172  -1.0101  -0.9538   1.3545   1.4188

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.1603     0.2838   0.565   0.5720
gest2         0.1624     0.3013   0.539   0.5898
gest3        -0.5675     0.2984  -1.902   0.0572 .
gest4        -0.6878     0.3044  -2.260   0.0238 *
gest5        -0.7120     0.3200  -2.225   0.0261 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2042.4  on 1487  degrees of freedom
Residual deviance: 1994.2  on 1483  degrees of freedom
AIC: 2004.2

Number of Fisher Scoring iterations: 4

> summary(model4)

Call:
glm(formula = re.ad ~ bwt, family = binomial(link = "logit"),
    data = neomod)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2396  -1.1090  -0.8987   1.2200   1.7374

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.29914    0.12292   2.434   0.0149 *
bwt         -0.31963    0.06741  -4.742 2.12e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2042.4  on 1487  degrees of freedom
Residual deviance: 2018.8  on 1486  degrees of freedom
AIC: 2022.8

Number of Fisher Scoring iterations: 4

> summary(model5)

Call:
glm(formula = re.ad ~ emp.f, family = binomial(link = "logit"),
    data = neomod)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-1.331  -1.049  -1.049   1.311   1.311

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3552     0.1582   2.246   0.0247 *
emp.f        -0.6645     0.1677  -3.963 7.39e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2042.4  on 1487  degrees of freedom
Residual deviance: 2026.4  on 1486  degrees of freedom
AIC: 2030.4

Number of Fisher Scoring iterations: 4

> summary(model6)

Call:
glm(formula = re.ad ~ emp.m, family = binomial(link = "logit"),
    data = neomod)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-1.145  -1.145  -1.009   1.211   1.355

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.07756    0.07194  -1.078  0.28101
emp.m       -0.33141    0.10495  -3.158  0.00159 **
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2042.4  on 1487  degrees of freedom
Residual deviance: 2032.4  on 1486  degrees of freedom
AIC: 2036.4

Number of Fisher Scoring iterations: 4

> summary(model7)

Call:
glm(formula = re.ad ~ edu, family = binomial(link = "logit"), 
    data = neomod)

Deviance Residuals: 
   Min      1Q  Median      3Q     Max  
-1.112  -1.093  -1.040   1.265   1.322  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.28090    0.21780  -1.290    0.197
edu2         0.07838    0.22819   0.343    0.731
edu3        -0.05192    0.24160  -0.215    0.830
edu4         0.12474    0.27123   0.460    0.646

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2042.4  on 1487  degrees of freedom
Residual deviance: 2041.0  on 1484  degrees of freedom
AIC: 2049

Number of Fisher Scoring iterations: 4

> summary(model8)

Call:
glm(formula = re.ad ~ los, family = binomial(link = "logit"), 
    data = neomod)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.5826  -1.0656  -0.7815   1.1605   1.8718  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -2.64657    0.27664  -9.567   <2e-16 ***
los          0.67426    0.07533   8.951   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2042.4  on 1487  degrees of freedom
Residual deviance: 1955.4  on 1486  degrees of freedom
AIC: 1959.4

Number of Fisher Scoring iterations: 4
```

```
> summary(model9)

Call:
glm(formula = re.ad ~ sex, family = binomial(link = "logit"),
    data = neomod)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.129  -1.129  -1.023   1.226   1.340

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.37447    0.07709  -4.858 1.19e-06 ***
sex          0.26057    0.10496   2.483    0.013 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2042.4  on 1487  degrees of freedom
Residual deviance: 2036.2  on 1486  degrees of freedom
AIC: 2040.2

Number of Fisher Scoring iterations: 4

> summary(model10)

Call:
glm(formula = re.ad ~ accom, family = binomial(link = "logit"),
    data = neomod)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.236  -1.027  -1.027   1.336   1.336

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.1378     0.1031   1.336    0.181
accom        -0.5021     0.1198  -4.191 2.78e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2042.4  on 1487  degrees of freedom
Residual deviance: 2024.8  on 1486  degrees of freedom
AIC: 2028.8

Number of Fisher Scoring iterations: 4

>
```

Appendix C – Kendall's Tau for Exploratory Analysis

```
> # Kendall's Tau
> los_empf<-cor.test(neomod$los,neomod$emp.f,method="kendall")
> los_empf

                Kendall's rank correlation tau
```

```
data:  neomod$los and neomod$emp.f
z = -2.06, p-value = 0.0394
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau 
-0.0439813 

> 
> los_sex<-cor.test(neomod$los,neomod$sex,method="kendall")
> los_sex

	Kendall's rank correlation tau

data:  neomod$los and neomod$sex
z = -0.80086, p-value = 0.4232
alternative hypothesis: true tau is not equal to 0
sample estimates:
        tau 
-0.01709831 

> 
> los_accom<-cor.test(neomod$los,neomod$accom,method="kendall")
> los_accom

	Kendall's rank correlation tau

data:  neomod$los and neomod$accom
z = -3.349, p-value = 0.0008111
alternative hypothesis: true tau is not equal to 0
sample estimates:
        tau 
-0.07150021 

> 
> empf_accom<-cor.test(neomod$emp.f,neomod$accom,method="kendall")
> empf_accom

	Kendall's rank correlation tau

data:  neomod$emp.f and neomod$accom
z = 14.805, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau 
0.3839189 

> 
> empf_sex<-cor.test(neomod$emp.f,neomod$sex,method="kendall")
> empf_sex

	Kendall's rank correlation tau

data:  neomod$emp.f and neomod$sex
z = -0.047685, p-value = 0.962
alternative hypothesis: true tau is not equal to 0
sample estimates:
         tau 
-0.001236604 
```

```
>
> sex_accom<-cor.test(neomod$sex,neomod$accom,method="kendall")
> sex_accom

	Kendall's rank correlation tau

data:  neomod$sex and neomod$accom
z = -0.12652, p-value = 0.8993
alternative hypothesis: true tau is not equal to 0
sample estimates:
         tau 
-0.003280915 

>
> empf_bwt<-cor.test(neomod$emp.f,neomod$bwt,method="kendall")
> empf_bwt

	Kendall's rank correlation tau

data:  neomod$emp.f and neomod$bwt
z = 1.0588, p-value = 0.2897
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau 
0.02249681 

>
> empf_gest<-cor.test(neomod$emp.f,as.numeric(neomod$gest),method="kendall")
> empf_gest

	Kendall's rank correlation tau

data:  neomod$emp.f and as.numeric(neomod$gest)
z = 1.1229, p-value = 0.2615
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau 
0.02655374 

>
> empf_empm<-cor.test(neomod$emp.f,neomod$emp.m,method="kendall")
> empf_empm

	Kendall's rank correlation tau

data:  neomod$emp.f and neomod$emp.m
z = 6.9672, p-value = 3.234e-12
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau 
0.1806758 

>
> empm_sex<-cor.test(neomod$emp.m,neomod$sex,method="kendall")
> empm_sex

	Kendall's rank correlation tau
```

```
data:  neomod$emp.m and neomod$sex
z = -0.88904, p-value = 0.374
alternative hypothesis: true tau is not equal to 0
sample estimates:
        tau 
-0.02305514 

> 
> empm_los<-cor.test(neomod$emp.m,neomod$los,method="kendall")
> empm_los

	Kendall's rank correlation tau

data:  neomod$emp.m and neomod$los
z = -1.8878, p-value = 0.05906
alternative hypothesis: true tau is not equal to 0
sample estimates:
        tau 
-0.04030392 

> 
> empm_accom<-cor.test(neomod$emp.m,neomod$accom,method="kendall")
> empm_accom

	Kendall's rank correlation tau

data:  neomod$emp.m and neomod$accom
z = 9.2205, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau 
0.2391108 

> 
> gest_empf<-cor.test(neomod$emp.f,as.numeric(neomod$gest),method="kendall")
> gest_empf

	Kendall's rank correlation tau

data:  neomod$emp.f and as.numeric(neomod$gest)
z = 1.1229, p-value = 0.2615
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau 
0.02655374 

> 
> bwt_los<-cor.test(neomod$bwt,neomod$los,method="kendall")
> bwt_los

	Kendall's rank correlation tau

data:  neomod$bwt and neomod$los
z = -31.88, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau 
-0.5577403 
```

```
>
> gest_los<-cor.test(neomod$los,as.numeric(neomod$gest),method="kendall")
> gest_los

	Kendall's rank correlation tau

data:  neomod$los and as.numeric(neomod$gest)
z = -31.982, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
-0.6227063

>
> gest_bwt<-cor.test(neomod$bwt,as.numeric(neomod$gest),method="kendall")
> gest_bwt

	Kendall's rank correlation tau

data:  neomod$bwt and as.numeric(neomod$gest)
z = 33.41, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.6473643

>
```

Appendix D – Building the Main Effects Model

```
> # Simulate Forward Model Selection
> # Empty model
> empt_model<-glm(re.ad~1,data=neomod,family=binomial(link="logit"))
> # Full model
> full_model<-glm(re.ad~.,data=neomod, family=binomial(link="logit"))
>
> # Using LRT to add/drop variables
> n=nrow(neomod)
>
> # Forward Selection
>
add1(empt_model,data=neomod,family=binomial(link="logit"),scope=full_model,test="LRT",k=log(n))
Single term additions

Model:
re.ad ~ 1
       Df Deviance    AIC    LRT  Pr(>Chi)
<none>      2042.4 2049.7
cns     1   2042.1 2056.8  0.275  0.599818
size    1   2040.8 2055.4  1.665  0.196907
gest    4   1994.2 2030.8 48.188 8.624e-10 ***
bwt     1   2018.8 2033.4 23.661 1.149e-06 ***
emp.f   1   2026.4 2041.0 16.002 6.329e-05 ***
emp.m   1   2032.4 2047.0 10.012  0.001555 **
edu     3   2041.0 2070.2  1.389  0.708204
los     1   1955.4 1970.0 87.014 < 2.2e-16 ***
sex     1   2036.2 2050.8  6.181  0.012910 *
accom   1   2024.8 2039.4 17.630 2.683e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
>
add1(glm(re.ad~los,data=neomod,family=binomial(link="logit")),scope=full_model,test="LRT",k=log(
n))
Single term additions

Model:
re.ad ~ los
       Df Deviance    AIC     LRT  Pr(>Chi)
<none>      1955.4 1970.0
cns     1   1955.4 1977.3  0.0013 0.9709685
size    1   1954.2 1976.1  1.1721 0.2789647
gest    4   1945.2 1989.0 10.2345 0.0366581 *
bwt     1   1952.2 1974.1  3.2379 0.0719508 .
emp.f   1   1942.1 1964.0 13.2906 0.0002667 ***
emp.m   1   1947.6 1969.5  7.7945 0.0052405 **
edu     3   1954.1 1990.7  1.2746 0.7351804
sex     1   1947.9 1969.8  7.4648 0.0062919 **
accom   1   1943.1 1965.0 12.2755 0.0004589 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
add1(glm(re.ad~los+emp.f,data=neomod,family=binomial(link="logit")),scope=full_model,test="LRT",
k=log(n))
Single term additions

Model:
re.ad ~ los + emp.f
       Df Deviance    AIC    LRT Pr(>Chi)
<none>      1942.1 1964.0
cns     1   1942.1 1971.3 0.0000 0.996530
size    1   1940.4 1969.7 1.6678 0.196555
gest    4   1932.7 1983.8 9.4058 0.051718 .
bwt     1   1939.3 1968.5 2.8319 0.092410 .
emp.m   1   1937.3 1966.6 4.7676 0.029000 *
edu     3   1941.2 1985.0 0.9449 0.814573
sex     1   1934.6 1963.8 7.5479 0.006008 **
accom   1   1936.9 1966.1 5.2274 0.022234 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
add1(glm(re.ad~los+emp.f+sex,data=neomod,family=binomial(link="logit")),scope=full_model,test="
LRT",k=log(n))
Single term additions

Model:
re.ad ~ los + emp.f + sex
       Df Deviance    AIC    LRT Pr(>Chi)
<none>      1934.6 1963.8
cns     1   1934.5 1971.1 0.0132  0.90865
size    1   1932.9 1969.5 1.6300  0.20170
gest    4   1926.0 1984.5 8.5515  0.07334 .
bwt     1   1932.9 1969.4 1.6426  0.19997
emp.m   1   1930.0 1966.6 4.5263  0.03338 *
edu     3   1933.7 1984.8 0.8940  0.82687
accom   1   1929.3 1965.8 5.2545  0.02189 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> add1(glm(re.ad~los+emp.f+sex+accom,data=neomod,family=binomial(link="logit")),scope=full_model,test="LRT",k=log(n))
Single term additions

Model:
re.ad ~ los + emp.f + sex + accom
       Df Deviance    AIC    LRT Pr(>Chi)  
<none>      1929.3 1965.8                  
cns     1   1929.3 1973.1 0.0159  0.89960  
size    1   1927.6 1971.4 1.6963  0.19277  
gest    4   1920.9 1986.7 8.3805  0.07859 .
bwt     1   1927.6 1971.4 1.7024  0.19198  
emp.m   1   1926.3 1970.2 2.9826  0.08417 .
edu     3   1928.3 1986.8 0.9963  0.80215  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # Result: re.ad~los+emp.f+sex+accom
>
```

Appendix E – Consideration of Interaction Terms

```
> # 6 possible combinations (4 choose 2)
> int_1<-glm(re.ad~los+emp.f+sex+accom+emp.f:sex,data=neomod,family=binomial(link="logit"))
> int_2<-glm(re.ad~los+emp.f+sex+accom+sex:los,data=neomod,family=binomial(link="logit"))
> int_3<-glm(re.ad~los+emp.f+sex+accom+los:emp.f,data=neomod,family=binomial(link="logit"))
> int_4<-glm(re.ad~los+emp.f+sex+accom+emp.f:accom,data=neomod,family=binomial(link="logit"))
> int_5<-glm(re.ad~los+emp.f+sex+accom+sex:accom,data=neomod,family=binomial(link="logit"))
> int_6<-glm(re.ad~los+emp.f+sex+accom+los:accom,data=neomod,family=binomial(link="logit"))
>
> summary(int_1)

Call:
glm(formula = re.ad ~ los + emp.f + sex + accom + emp.f:sex, 
    family = binomial(link = "logit"), data = neomod)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.7277  -1.0412  -0.7528   1.1543   1.9949  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -1.99327    0.37483  -5.318 1.05e-07 ***
los          0.66160    0.07609   8.695  < 2e-16 ***
emp.f       -0.60600    0.26258  -2.308   0.0210 *  
sex          0.06376    0.32884   0.194   0.8463    
accom       -0.30839    0.13386  -2.304   0.0212 *  
emp.f:sex    0.26441    0.34832   0.759   0.4478    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2042.4  on 1487  degrees of freedom
Residual deviance: 1928.7  on 1482  degrees of freedom
AIC: 1940.7

Number of Fisher Scoring iterations: 4
```

```
> summary(int_2)

Call:
glm(formula = re.ad ~ los + emp.f + sex + accom + sex:los, family = binomial(link = "logit"),
   data = neomod)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.8366  -1.0305  -0.7807   1.1660   1.8856

Coefficients:
          Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6790     0.4554  -3.687 0.000227 ***
los          0.5446     0.1132   4.813 1.49e-06 ***
emp.f       -0.4774     0.1874  -2.547 0.010858 *
sex         -0.4739     0.5622  -0.843 0.399300
accom       -0.3128     0.1340  -2.334 0.019590 *
los:sex      0.2146     0.1532   1.400 0.161394
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 2042.4  on 1487  degrees of freedom
Residual deviance: 1927.3  on 1482  degrees of freedom
AIC: 1939.3

Number of Fisher Scoring iterations: 4

> summary(int_3)

Call:
glm(formula = re.ad ~ los + emp.f + sex + accom + los:emp.f,
   family = binomial(link = "logit"), data = neomod)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.8612  -1.0384  -0.7628   1.1648   1.9732

Coefficients:
          Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.6889     0.8343  -3.223 0.001270 **
los          0.8194     0.2249   3.644 0.000269 ***
emp.f        0.1651     0.8783   0.188 0.850855
sex          0.3031     0.1091   2.778 0.005474 **
accom       -0.3047     0.1339  -2.275 0.022900 *
los:emp.f   -0.1759     0.2387  -0.737 0.461232
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 2042.4  on 1487  degrees of freedom
Residual deviance: 1928.8  on 1482  degrees of freedom
AIC: 1940.8

Number of Fisher Scoring iterations: 4

> summary(int_4)
```

```
Call:
glm(formula = re.ad ~ los + emp.f + sex + accom + emp.f:accom, 
    family = binomial(link = "logit"), data = neomod)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.7603  -1.0426  -0.7553   1.1590   1.9933  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -2.1738     0.3460  -6.282 3.34e-10 ***
los           0.6652     0.0761   8.741  < 2e-16 ***
emp.f        -0.4049     0.2314  -1.750  0.08016 .  
sex           0.2998     0.1090   2.750  0.00596 ** 
accom        -0.1530     0.3651  -0.419  0.67521    
emp.f:accom  -0.1781     0.3922  -0.454  0.64979    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2042.4  on 1487  degrees of freedom
Residual deviance: 1929.1  on 1482  degrees of freedom
AIC: 1941.1

Number of Fisher Scoring iterations: 4

> summary(int_5)

Call:
glm(formula = re.ad ~ los + emp.f + sex + accom + sex:accom, 
    family = binomial(link = "logit"), data = neomod)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.7135  -1.0449  -0.7534   1.1590   2.0119  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -1.950099   0.347640  -5.610 2.03e-08 ***
los          0.660294   0.076104   8.676  < 2e-16 ***
emp.f       -0.469382   0.187040  -2.510  0.01209 *  
sex         -0.004941   0.214482  -0.023  0.98162    
accom       -0.525351   0.188461  -2.788  0.00531 ** 
sex:accom    0.409816   0.248947   1.646  0.09972 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2042.4  on 1487  degrees of freedom
Residual deviance: 1926.6  on 1482  degrees of freedom
AIC: 1938.6

Number of Fisher Scoring iterations: 4

> summary(int_6)
```

```
Call:
glm(formula = re.ad ~ los + emp.f + sex + accom + los:accom,
    family = binomial(link = "logit"), data = neomod)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.868  -1.037  -0.773   1.168   1.939

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.7709     0.5890  -4.704 2.55e-06 ***
los           0.8393     0.1528   5.493 3.95e-08 ***
emp.f        -0.4644     0.1885  -2.464  0.01375 *
sex           0.3055     0.1092   2.798  0.00513 **
accom         0.5496     0.6533   0.841  0.40020
los:accom    -0.2362     0.1761  -1.341  0.17994
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2042.4  on 1487  degrees of freedom
Residual deviance: 1927.5  on 1482  degrees of freedom
AIC: 1939.5

Number of Fisher Scoring iterations: 4

>
```

Appendix F – Choice of Link Function

```
> # Compare link functions
> main_model<-glm(re.ad~los+emp.f+sex+accom,binomial(link="logit"),data=neomod)
> summary(main_model)

Call:
glm(formula = re.ad ~ los + emp.f + sex + accom, family = binomial(link = "logit"),
    data = neomod)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7790  -1.0394  -0.7572   1.1572   1.9903

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.12704    0.33050  -6.436 1.23e-10 ***
los          0.66426    0.07605   8.735  < 2e-16 ***
emp.f       -0.46713    0.18704  -2.497  0.01251 *
sex          0.29927    0.10898   2.746  0.00603 **
accom       -0.30717    0.13383  -2.295  0.02172 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2042.4  on 1487  degrees of freedom
Residual deviance: 1929.3  on 1483  degrees of freedom
AIC: 1939.3

Number of Fisher Scoring iterations: 4
```

```
>
> main_model_2<-glm(re.ad~los+emp.f+sex+accom,binomial(link="probit"),data=neomod)
> summary(main_model_2)

Call:
glm(formula = re.ad ~ los + emp.f + sex + accom, family = binomial(link = "probit"),
    data = neomod)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7860  -1.0432  -0.7594   1.1580   2.0146

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.29631    0.20089  -6.453  1.1e-10 ***
los          0.40663    0.04582   8.874  < 2e-16 ***
emp.f       -0.29045    0.11508  -2.524  0.01161 *
sex          0.18080    0.06697   2.700  0.00694 **
accom       -0.18733    0.08267  -2.266  0.02346 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2042.4  on 1487  degrees of freedom
Residual deviance: 1929.9  on 1483  degrees of freedom
AIC: 1939.9

Number of Fisher Scoring iterations: 4

>
> main_model_3<-glm(re.ad~los+emp.f+sex+accom,binomial(link="cloglog"),data=neomod)
> summary(main_model_3)

Call:
glm(formula = re.ad ~ los + emp.f + sex + accom, family = binomial(link = "cloglog"),
    data = neomod)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9322  -1.0283  -0.7661   1.1629   1.9392

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.09213    0.24602  -8.504  < 2e-16 ***
los          0.52118    0.05628   9.260  < 2e-16 ***
emp.f       -0.32757    0.12610  -2.598  0.00938 **
sex          0.22112    0.08060   2.743  0.00608 **
accom       -0.21767    0.09583  -2.271  0.02312 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2042.4  on 1487  degrees of freedom
Residual deviance: 1925.2  on 1483  degrees of freedom
AIC: 1935.2

Number of Fisher Scoring iterations: 5
```

```
>
> main_model_4<-glm(re.ad~los+emp.f+sex+accom,binomial(link="cauchit"),data=neomod)
> summary(main_model_4)

Call:
glm(formula = re.ad ~ los + emp.f + sex + accom, family = binomial(link = "cauchit"),
    data = neomod)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7368  -1.0185  -0.7562   1.1621   1.8761

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.95215    0.31280  -6.241 4.35e-10 ***
los          0.59880    0.07541   7.940 2.02e-15 ***
emp.f       -0.39201    0.16441  -2.384  0.01711 *
sex          0.28431    0.09684   2.936  0.00333 **
accom       -0.28587    0.11554  -2.474  0.01336 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2042.4  on 1487  degrees of freedom
Residual deviance: 1926.5  on 1483  degrees of freedom
AIC: 1936.5

Number of Fisher Scoring iterations: 5

>
> # Choose final model as cloglog link
> main_model<-main_model_3
```

Appendix G – Residuals for Model Evaluation

```
> # Residuals
> dev_res<-residuals(main_model,type="deviance")
> mean(dev_res)
[1] -0.03682001
> var(dev_res)
[1] 1.293322
>
> pea_res<-residuals(main_model,type="pearson")
> mean(pea_res)
[1] 0.002195945
> var(pea_res)
[1] 1.005389
>
> # Examine for linear relationship between covariates and residuals
> par(mfrow=c(2,2))
> plot(neomod$los,pea_res,main="neomod$los vs.\npearson residuals",ylab="pearson residuals")
> lines(smooth.spline(neomod$los,pea_res),col=2)
> plot(neomod$accom,pea_res,main="neomod$accom vs.\npearson residuals",ylab="pearson
residuals")
> plot(neomod$emp.f,pea_res,main="neomod$emp.f vs. \npearson residuals",ylab="pearson residuals")
> plot(neomod$sex,pea_res,main="neomod$sex vs. \npearson residuals",ylab="pearson residuals")
>
> res_df<-data.frame(cbind(neomod$accom,neomod$emp.f,neomod$sex,pea_res))
> colnames(res_df)<-c("accom","emp.f","sex","pea_res")
>
> # Accom
> df_acc0<-res_df %>% filter(accom==0)
> mean(df_acc0$pea_res)
[1] -0.0001058469
> df_acc1<-res_df %>% filter(accom==1)
> mean(df_acc1$pea_res)
[1] 0.002979799
> # difference
> abs(mean(df_acc0$pea_res)-mean(df_acc1$pea_res))
[1] 0.003085645
>
> # Emp.f
> df_empf0<-res_df %>% filter(emp.f==0)
> mean(df_empf0$pea_res)
[1] 0.003019975
> df_empf1<-res_df %>% filter(emp.f==1)
> mean(df_empf1$pea_res)
[1] 0.002093175
> # difference
> abs(mean(df_empf0$pea_res)-mean(df_empf1$pea_res))
[1] 0.0009267997
>
> # Sex
> df_sex0<-res_df %>% filter(sex==0)
> mean(df_sex0$pea_res)
[1] 0.004364996
> df_sex1<-res_df %>% filter(sex==1)
> mean(df_sex1$pea_res)
[1] 0.0002846575
> # difference
> abs(mean(df_sex0$pea_res)-mean(df_sex1$pea_res))
[1] 0.004080338
>
```

Appendix H – Influence Measures for Model Evaluation

```
> # Outlier detection
> p<-4 # 4 parameters
>
> influence.measures(main_model)
> Influence measures of
         glm(formula = re.ad ~ los + emp.f + sex + accom, family = binomial(link = "cloglog"),     data
 = neomod) :

    dfb.1_   dfb.los dfb.emp. dfb.sex dfb.accm   dffit cov.r   cook.d     hat inf
1   1.43e-02 -0.023234  0.00412  0.0254  0.01423  0.0501 1.000 0.000560 0.00162
2   5.58e-02 -0.070181  0.00384  0.0229  0.01061  0.0820 0.998 0.002055 0.00285
3   3.44e-02 -0.031361 -0.00481 -0.0304 -0.01981 -0.0594 1.002 0.000724 0.00265
4  -1.75e-02  0.023667 -0.00219 -0.0133 -0.00693 -0.0336 1.004 0.000175 0.00206
5  -2.19e-02  0.027777 -0.00165 -0.0099 -0.00469 -0.0332 1.005 0.000164 0.00270
6  -1.41e-02  0.013090 -0.00280  0.0197 -0.00834 -0.0310 1.003 0.000151 0.00166
7  -8.79e-03  0.014767 -0.00284 -0.0175 -0.00988 -0.0339 1.003 0.000191 0.00159
8  -8.79e-03  0.014767 -0.00284 -0.0175 -0.00988 -0.0339 1.003 0.000191 0.00159
9  -1.41e-02  0.013090 -0.00280  0.0197 -0.00834 -0.0310 1.003 0.000151 0.00166
10  3.39e-02 -0.025964  0.03404 -0.0344 -0.05689  0.0771 1.002 0.001303 0.00394
11  2.15e-02 -0.023890  0.03270  0.0255 -0.05636  0.0726 1.003 0.001049 0.00422
12 -1.76e-02  0.019472 -0.02860 -0.0224  0.04920 -0.0630 1.004 0.000708 0.00421
13  1.60e-03  0.003306  0.00498 -0.0339  0.01632  0.0495 1.000 0.000516 0.00176
```

```
14  5.95e-03 -0.001631  0.00498 -0.0342  0.01606  0.0496 1.000 0.000532 0.00167
15  2.48e-02 -0.020906 -0.00445 -0.0280 -0.01787 -0.0512 1.001 0.000516 0.00218
16  6.49e-02 -0.027204 -0.05888 -0.0323 -0.01825  0.0900 1.007 0.001560 0.00699
17  9.49e-03 -0.015320 -0.00441  0.0296 -0.01516 -0.0461 1.002 0.000383 0.00222
18 -2.64e-02  0.034810  0.00473 -0.0309  0.01729  0.0579 1.004 0.000597 0.00359
19 -1.24e-02  0.018444 -0.00260 -0.0160 -0.00878 -0.0336 1.003 0.000183 0.00172
20 -2.16e-02  0.022754 -0.00171  0.0127 -0.00429 -0.0287 1.005 0.000120 0.00237
21  2.11e-03 -0.009259  0.00412  0.0255  0.01498  0.0450 1.000 0.000418 0.00150
22 -1.84e-02  0.025910  0.00485 -0.0321  0.01717  0.0542 1.003 0.000549 0.00275
23 -3.47e-02  0.033457  0.00369  0.0235  0.01571  0.0510 1.005 0.000430 0.00358
24 -7.90e-03  0.013838 -0.00290 -0.0179 -0.01014 -0.0341 1.002 0.000193 0.00157
25 -7.90e-03  0.013838 -0.00290 -0.0179 -0.01014 -0.0341 1.002 0.000193 0.00157
26 -1.29e-02  0.011610 -0.00291  0.0204 -0.00878 -0.0314 1.003 0.000157 0.00163
27 -9.19e-03  0.007131 -0.00322  0.0223 -0.01002 -0.0329 1.003 0.000176 0.00161
28  8.72e-05  0.005464 -0.00335 -0.0208 -0.01230 -0.0361 1.002 0.000227 0.00151
29 -9.19e-03  0.007131 -0.00322  0.0223 -0.01002 -0.0329 1.003 0.000176 0.00161
30 -4.08e-03 -0.007989 -0.03399  0.0333  0.05454 -0.0728 1.004 0.001010 0.00470
31 -1.91e-02  0.019425 -0.00222  0.0160 -0.00611 -0.0297 1.004 0.000133 0.00194
32  1.45e-02 -0.009858 -0.00403 -0.0252 -0.01568 -0.0437 1.002 0.000357 0.00178
33  1.45e-02 -0.009858 -0.00403 -0.0252 -0.01568 -0.0437 1.002 0.000357 0.00178
34  2.07e-02 -0.028674 -0.00499  0.0329 -0.01777 -0.0567 1.003 0.000610 0.00290
35  3.43e-02 -0.044732 -0.00561  0.0365 -0.02068 -0.0707 1.003 0.001008 0.00390
36 -1.64e-02  0.012100  0.00401  0.0252  0.01574  0.0440 1.002 0.000359 0.00186
37 -4.89e-02 -0.002657  0.07344  0.0394  0.02048 -0.1072 1.006 0.002473 0.00782
38 -2.16e-02  0.022754 -0.00171  0.0127 -0.00429 -0.0287 1.005 0.000120 0.00237
39 -1.57e-02  0.017331 -0.02954 -0.0232  0.05063 -0.0642 1.004 0.000747 0.00422
40 -6.42e-03  0.006706 -0.03364 -0.0267  0.05679 -0.0707 1.004 0.000961 0.00435
41 -1.10e-02  0.009375 -0.00307  0.0214 -0.00942 -0.0321 1.003 0.000166 0.00161
42 -1.03e-04 -0.000534 -0.03606 -0.0289  0.06037 -0.0754 1.004 0.001136 0.00454
43  1.29e-02 -0.019423 -0.00460  0.0306 -0.01598 -0.0492 1.002 0.000443 0.00241
44 -1.64e-02  0.015996 -0.00256  0.0182 -0.00740 -0.0303 1.003 0.000142 0.00175
45 -3.19e-02  0.030249  0.00376  0.0239  0.01580  0.0496 1.004 0.000414 0.00316
46 -5.50e-03  0.002654 -0.00349  0.0240 -0.01116 -0.0348 1.003 0.000202 0.00165
47  3.17e-02  0.005449 -0.05235 -0.0280 -0.01438  0.0766 1.009 0.000971 0.00804
48  2.50e-02  0.009399 -0.08392 -0.0307  0.05176  0.0964 1.010 0.001660 0.00972
49 -2.17e-02  0.035140  0.02804 -0.0266 -0.04309  0.0691 1.009 0.000762 0.00769
50 -3.54e-02  0.034317  0.00367  0.0234  0.01568  0.0514 1.005 0.000435 0.00370
 [ reached 'max' / getOption("max.print") -- omitted 1438 rows ]

> # Number of abnormally large standardized residuals
> sum(as.numeric(abs(residuals(main_model))>3)) # None, consistent with the plot
[1] 0
> # Number of points that are influential according to dffit residual
> sum(as.numeric(dffits(main_model)>2*sqrt(p/n))) #7
[1] 7
> # Number of influential points according to residual variance
> sum(as.numeric(abs(covratio(main_model))>(1+((3*p)/(n-p))))) #120
[1] 120
> # Number of influential points according to Cook's Distance
> sum(as.numeric(cooks.distance(main_model)>4/(n-p))) #35
[1] 35
> # Number of influential points according to high leverage
> sum(as.numeric((hatvalues(main_model)/(p/n))>2)) #235
[1] 235
>
```

Appendix I – Hosmer-Lemeshow Test for Model Assessment

```
> # Hosmer-Lemeshow with G=10
> group<-cut(fitted.values(main_model),breaks=quantile(fitted.values(main_model),
+                    seq(0,1,length.out=11)),include.lowest=TRUE)
> obser<-tapply(main_model$y,group,sum)
> pr<-tapply(fitted.values(main_model),group,mean)
> ng<-as.numeric(table(group))
> expec<-pr*ng
> HLcontr<-(obser-expec)^2/(ng*pr*(1-pr))
> HL<-sum(HLcontr)
> HL
[1] 9.086084
> 1-pchisq(HL,df=8)
[1] 0.3350862
```