SMM634 Analytics Methods for Business 2022/23

Final Group Coursework

Group 6

Linh Nguyen

Soumya Ogoti

Wenxu Tian

Aparna Viswanathan

Fan Xia

Question 1. (a) Justification of model choice

In the Visits dataset, the topic of interest is the number of doctor visits in the past two weeks which is count data. To model this, the first choice of model is Poisson regression, which fits a Poisson distribution to the data with a log link:

$$Visits_i \sim Poisson(\mu_i) \text{ with } V(\mu_i) = \mu_i$$

$$log(\mu_i) = \beta_0 + \beta_1 gender.male + \beta_2 age + \beta_3 income + \beta_4 illness + \beta_5 reduced + \beta_6 health + \beta_7 privatee.yes + \beta_8 freepoor.yes + \beta_9 freerepat.yes + \beta_{10} nchronic.yes + \beta_{11} lchronic.yes$$

This serves as a baseline model and its results would be used to inform the following analyses. Subsequently, analysing the first model's result suggests the presence of overdispersion which makes the estimated coefficients unreliable, a Quasi-Poisson model was selected to quantify overdispersion. The model achieves this by estimating a dispersion parameter Φ where

$$Visits_i \sim Poisson(\mu_i) \text{ with } V(\mu_i) = \Phi\mu_i$$

Since the Poisson distribution shows a poor fit and the mean and variance are not equal, it was logical to consider the Negative Binomial (NB) distribution - a more flexible distribution which does not impose the equality of mean and variance by introducing the shape parameter τ:

$$Visits_i \sim NB(\mu_i) \text{ with } V(y) = \mu + {}^{\mu}/_{\tau}$$

The results suggested that overdispersion has been addressed, but the model suffers from slight underdispersion and there are still uncaptured patterns in the residuals, the data was enriched by transforming the non-categorical variables into factors and the negative binomial model was refitted on the data. However, since only factorising the **`illness`** variable yielded a significant improvement indicated by the ANOVA test, the model's equation is now:

$$log(\mu_i) = \beta_0 + \beta_1 gender.male + \beta_2 age + \beta_3 income + \beta_4 reduced + \beta_5 health + \beta_6 privatee.yes + \beta_7 freepoor.yes + \beta_8 freerepat.yes + \beta_9 nchronic.yes + \beta_{10} lchronic.yes + \beta_{11} illness.1 + \beta_{12} illness.2 + \beta_{13} illness.3 + \beta_{14} illness.4 + \beta_{15} illness.5$$

Finally, since the coefficient for level 2 of **`illness`** being slightly higher than that of level 3 indicates the presence of nonlinearity in the data, a Generalised Additive Model using Thin Plate Regression Spline (TPRS) was fitted on the data, using the Negative Binomial distribution with a log link. This is advantageous because this model does not make assumptions on the functional shapes of the relations and without the shortcomings of using polynomial bases as well as categorising continuous variables in the groups. In this model, all the non-numerical variables were smoothened. The number of bases was set at 4 as a starting point and tuned using a smoothing parameter λ which was automatically selected with cross-validation. The model equation is now

$$log(\mu_i) = \beta_0 + \beta_1 gender.male + \beta_2 privatee.yes + \beta_3 freepoor.yes + \beta_4 freerepat.yes + \beta_5 nchronic.yes + \beta_6 lchronic.yes + s(age) + s(income) + s(illness) + s(reduced) + s(health)$$

in which, s is a smooth function which can be approximated by

$$f(x_i) = \sum_{k=1}^{d} \beta_k b_k(x_i)$$

b_k are the known basis functions, β_k unknown regression parameters, and d number of bases.

(b) Summaries and interpretation of the fitted models

Table 1. *Comparison of the fitted models*

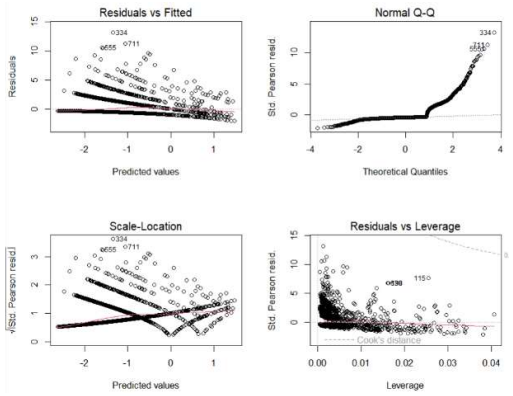
Model no.	Model description	Null deviance	Residual deviance	Dispersion parameter	AIC
1	Poisson GLM	5634.8	4380.1	-	6735.7
2	Quasi-Poisson GLM	5634.8	4380.1	1.32	-
3	Negative binomial GLM	3930.4	3029.8	0.93	6423.7
4	Negative binomial GLM - illness factorised	4020.7	3023.5	1.02	6357.9
5	Generalised additive model using TPRS	4143	3001.4	1.16	6249.7

Model 1: Poisson GLM

Table 2. *Model 1's summary*

```
Call:
glm(formula = "visits ~ .", family = "poisson", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9502  -0.6858  -0.5747  -0.4852   5.7055

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.941332   0.100992 -19.223  < 2e-16 ***
gendermale   -0.156490   0.056139  -2.788  0.00531 **
age           0.279123   0.165981   1.682  0.09264 .
income       -0.187416   0.085478  -2.193  0.02834 *
illness       0.186156   0.018263  10.193  < 2e-16 ***
reduced       0.126690   0.005031  25.184  < 2e-16 ***
health        0.030683   0.010074   3.046  0.00232 **
privateyes    0.126498   0.071552   1.768  0.07707 .
freepooryes  -0.438462   0.179799  -2.439  0.01474 *
freerepatyes  0.083640   0.092070   0.908  0.36365
nchronicyes   0.117300   0.066545   1.763  0.07795 .
lchronicyes   0.150717   0.082260   1.832  0.06692 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 5634.8  on 5189  degrees of freedom
Residual deviance: 4380.1  on 5178  degrees of freedom
AIC: 6735.7

Number of Fisher Scoring iterations: 6
```

Figure 1. *Residual analysis of Model 1*



The results indicate that **`illness`** and **`reduced`** are the most significant variables, followed by **`health`**, **`gender`**, **`income`** and **`freepoor`**. The remaining variables, except for **`freerepat`** are significant at 0.1 level. Holding other variables constant, the model estimates that on average, men have 14.45% fewer visits compared to women. The model's residual variance of 4380.1 being lower than the null deviance of 5634.8 means the current model is better than the null model, and the Chi-square test on 5178 degrees of freedom yields a p-value of 1 means we fail to reject the null hypothesis that the fitted model is reasonable. The residual vs fitted plots in Figure 1 shows that the residuals are not randomly scattered around zero, and there are patterns not captured by the model indicated by the line patterns. The Normal Q-Q plot suggests that the residuals are not normally distributed, which does not satisfy the model's assumption of normality. No high-leverage points can be identified based on the Cook's distance. Therefore, the coefficients' significances outputted by the model are not reliable as the model's assumptions are not satisfied, likely due to overdispersion.

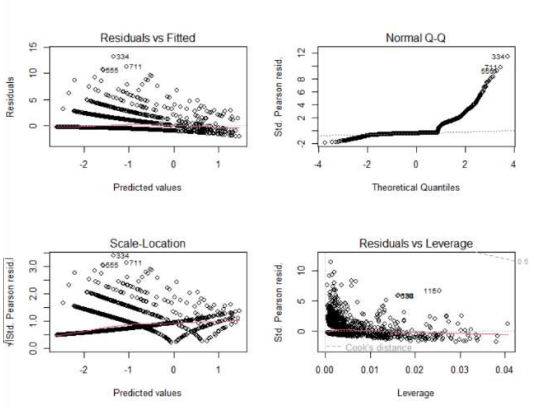
Model 2: Quasi-Poisson GLM

The second model is a quasi-Poisson model which accounts for overdispersion.

Table 3. *Model 2's summary*

```
Call:
glm(formula = "visits ~ .", family = "quasipoisson", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9502  -0.6858  -0.5747  -0.4852   5.7055

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.941332   0.116363 -16.683  < 2e-16 ***
gendermale   -0.156490   0.064683  -2.419  0.01558 *
age           0.279123   0.191244   1.460  0.14448
income       -0.187416   0.098488  -1.903  0.05711 .
illness       0.186156   0.021043   8.847  < 2e-16 ***
reduced       0.126690   0.005796  21.857  < 2e-16 ***
health        0.030683   0.011607   2.644  0.00823 **
privateyes    0.126498   0.082442   1.534  0.12500
freepooryes  -0.438462   0.207164  -2.116  0.03435 *
freerepatyes  0.083640   0.106083   0.788  0.43048
nchronicyes   0.117300   0.076674   1.530  0.12611
lchronicyes   0.150717   0.094780   1.590  0.11186
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 1.327571)

    Null deviance: 5634.8  on 5189  degrees of freedom
Residual deviance: 4380.1  on 5178  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

Figure 2. *Residual analysis of Model 2*



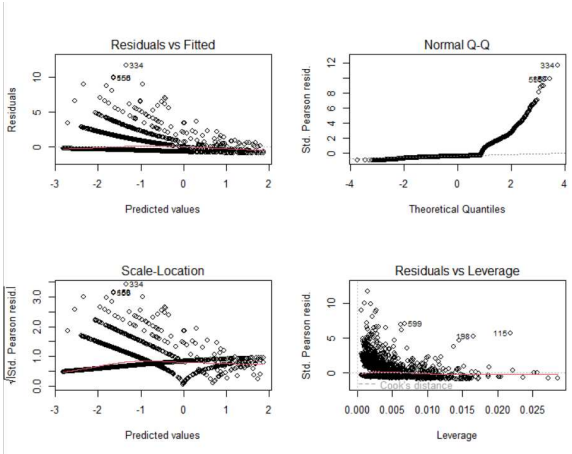
In this model, the dispersion parameter has been estimated as 1.327, suggesting overdispersion, that is for every unit increase in the mean, the variance increases by 1.327. Therefore, although **`illness`** and **`reduced`** are still 2 significant regressors, **`income`** is no longer significant at 0.05 level and the remaining variables that were significant at 0.1 level are no longer significant, after having adjusted the standard errors using the estimated dispersion parameter. Finally, since the models' coefficients did not change, the model's goodness of fit was not improved, as confirmed in Figure 2.

Model 3: Negative Binomial GLM

Table 4. *Model 3's summary*

```
Call:
glm.nb(formula = "visits ~ .", data = data, init.theta = 0.9301535861,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9630  -0.6355  -0.5279  -0.4411   4.0045

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.059804   0.122616 -16.799  < 2e-16 ***
gendermale   -0.216333   0.069681  -3.105  0.00191 **
age           0.331326   0.207755   1.595  0.11076
income       -0.156214   0.103907  -1.503  0.13273
illness       0.214937   0.023521   9.138  < 2e-16 ***
reduced       0.143729   0.007305  19.674  < 2e-16 ***
health        0.037535   0.013609   2.758  0.00581 **
privateyes    0.116379   0.085666   1.359  0.17430
freepooryes  -0.497256   0.210696  -2.360  0.01827 *
freerepatyes  0.145683   0.115851   1.258  0.20857
nchronicyes   0.097905   0.079153   1.237  0.21612
lchronicyes   0.183473   0.103176   1.778  0.07536 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.9302) family taken to be 1)

    Null deviance: 3930.4  on 5189  degrees of freedom
Residual deviance: 3029.8  on 5178  degrees of freedom
AIC: 6423.7
```

Figure 3. *Residual analysis of Model 3*



The magnitude and significance level of the variables that were significant at 0.05 level in the previous model only changed slightly. Interestingly, **`income`** which was significant at 0.1 level is now insignificant, while the reverse is true for **`lchronic`**. As indicated by the dispersion parameter of 0.93, although we have addressed the overdispersion issue, the model now suffers from slight underdispersion. The model's residual variance of 3029.8 being lower than the null

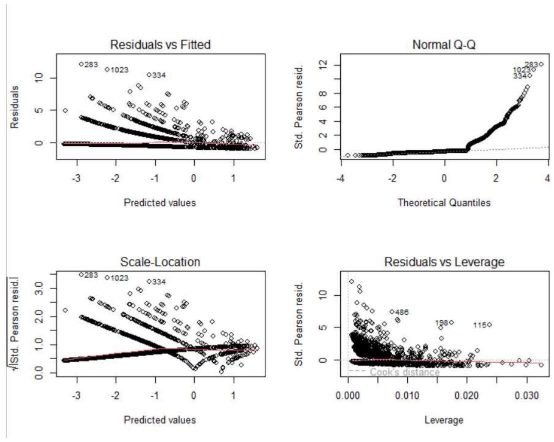
deviance of 3930.4 means the current model is better than the null model, and the Chi-square tests yield a p-value of 1 means we fail to reject the null hypothesis that the null and fitted models are reasonable. Compared to the Poisson model, the residual deviance and AIC has dropped by 31% and 4.6% respectively means this model provides a much better fit. Although the previously identified issues in the residual analysis still persist, the 0.5 Cook's distance line is no longer visible, indicating a slightly improved model robustness.

Model 4: Negative Binomial GLM – illness factorised

Table 5. *Model 4's summary*

```
Call:
glm.nb(formula = visits ~ gender + age + income + reduced + health +
    private + freepoor + freerepat + nchronic + lchronic + as.factor(illness),
    data = data, init.theta = 1.021667857, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8718  -0.6894  -0.5396  -0.3430   3.8413

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.652603   0.147525 -17.981  < 2e-16 ***
gendermale          -0.194151   0.068921  -2.817  0.00485 **
age                  0.501477   0.207020   2.422  0.01542 *
income              -0.153905   0.103010  -1.494  0.13515
reduced              0.135809   0.007157  18.975  < 2e-16 ***
health               0.037536   0.013303   2.822  0.00478 **
privateyes           0.101599   0.085502   1.188  0.23473
freepooryes         -0.478343   0.209788  -2.280  0.02260 *
freerepatyes         0.119424   0.114803   1.040  0.29822
nchronicyes          0.026141   0.078873   0.331  0.74032
lchronicyes          0.133869   0.101409   1.320  0.18681
as.factor(illness)1  0.982747   0.111796   8.791  < 2e-16 ***
as.factor(illness)2  1.248407   0.118507  10.534  < 2e-16 ***
as.factor(illness)3  1.206534   0.132104   9.133  < 2e-16 ***
as.factor(illness)4  1.325622   0.150688   8.797  < 2e-16 ***
as.factor(illness)5  1.392921   0.154086   9.040  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.0217) family taken to be 1)

    Null deviance: 4020.7  on 5189  degrees of freedom
Residual deviance: 3023.5  on 5174  degrees of freedom
AIC: 6357.9
```

Figure 4. *Residual analysis of Model 4*



By factorising **`illness`**, the dispersion parameter is now 1.02, which indicates little dispersion. The model's residual deviance and AIC have also slightly improved from the previous model (3029 to 3023 and 6423 to 6357 respectively). Interestingly, **`age`** is now statistically significant with the coefficient's estimated increasing from 0.33 to 0.5. All five levels of **`illness`** are also statistically significant and indicate that the more illness someone has in the last 2 weeks, the more visits they will make, which makes sense. For instance, the number of visits by people with 5 illnesses is 1.069 times the number of visits by people with 4 illnesses, holding other variables constant. This is also consistent with the positive coefficient estimated by the previous models. Interestingly, the coefficient for level 2 of **`illness`** being slightly higher than that of level 3 suggests the presence of nonlinearity in this variable. Finally, the ANOVA table below suggests that factorising the **`illness`** variable has significantly improved the model.

Table 6. *The ANOVA test shows that factorising illness has significantly improved the model*

```
                                                                                                  Model     theta
1            gender + age + income + illness + reduced + health + private + freepoor + freerepat + nchronic + lchronic 0.9301536
2 gender + age + income + reduced + health + private + freepoor + freerepat + nchronic + lchronic + as.factor(illness) 1.0216679
  Resid. df    2 x log-lik.   Test    df LR stat.      Pr(Chi)
1      5178       -6397.676
2      5174       -6323.948 1 vs 2     4 73.72776 3.663736e-15
```

Model 5: Generalised additive model using TPRS

Table 7. *Model 5's summary*

```
Family: Negative Binomial(1.165)
Link function: log

Formula:
visits ~ private + freepoor + freerepat + nchronic + lchronic +
    gender + s(illness, bs = "tp", k = 4) + s(age, bs = "tp",
    k = 4) + s(income, bs = "tp", k = 4) + s(reduced, bs = "tp",
    k = 4) + s(health, bs = "tp", k = 4)

Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.67815    0.08901 -18.853   <2e-16 ***
privateyes    0.13623    0.08585   1.587   0.1126
freepooryes  -0.52161    0.21164  -2.465   0.0137 *
freerepatyes  0.18233    0.11445   1.593   0.1111
nchronicyes   0.05369    0.07852   0.684   0.4941
lchronicyes   0.14484    0.09968   1.453   0.1462
gendermale   -0.14791    0.06838  -2.163   0.0305 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
             edf Ref.df  Chi.sq p-value
s(illness) 2.908  2.993  96.013 < 2e-16 ***
s(age)     1.001  1.002   9.244 0.00238 **
s(income)  1.991  2.388   7.346 0.03959 *
s(reduced) 2.825  2.973 473.194 < 2e-16 ***
s(health)  1.002  1.003   9.916 0.00165 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5. *Residual analysis of Model 5*

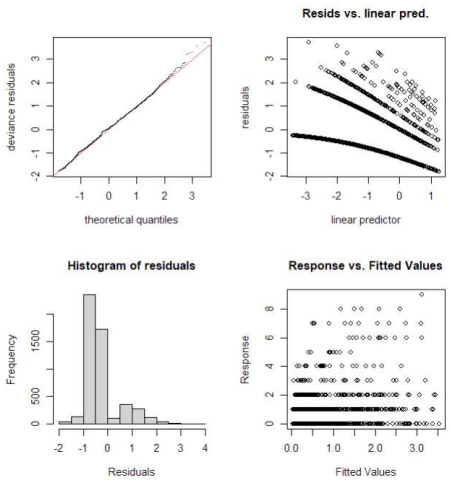
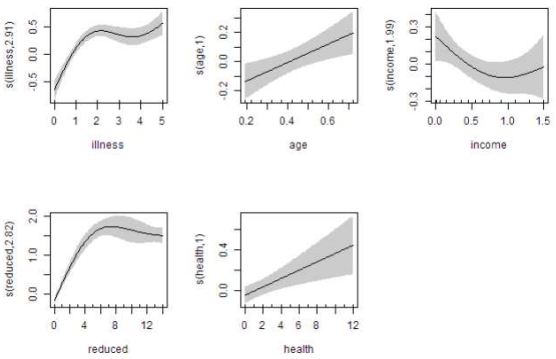


Figure 6. *Effects of interest*



The table indicates that the coefficients for the non-smoothed variables did not change drastically. The insignificant variables in the previous model remain insignificant, except for **`income`** which is now significant after being smoothened. This suggests the non-linearity effect of this variable. This is also confirmed by its effective degree of freedom (edf) of 1.99. Interestingly, the edf for **`illness`** and **`reduced`** are also approximately 2.9, indicating the added value of applying TPRS. This agrees with the previous model on the potential non-linearity of **`illness`**. In contrast, the edf of **`age`** and **`health`** being close to 1 suggests there is no need to transform these variables. The dispersion parameter is estimated at 1.16 which is slightly higher than in the previous model. The model has achieved an AIC and residual deviance of 3001.4 and 6249.7 respectively, the lowest among all the fitted models. Residuals-wise, the plots indicate a strong improvement now that the residuals are mostly normally distributed, although there are still uncaptured patterns as shown by the clear curves in the residuals vs linear predictor plot. This model can be considered the best achieved so far.

(c) Pros and cons of the analyses and more suitable model specifications

In this analysis, we have identified multiple reasonable models as indicated by analysing the residual deviance. The analysis followed an iterative approach, where the results of the previous models were used to inform the following model specifications. This has resulted in the final

model being the most adequate in terms of satisfaction with the model's assumptions and goodness of fit.

In the final model, the problem of overdispersion identified from the beginning has been mostly addressed, which means the variables' significances can be interpreted more reliably. The residuals also mostly follow a normal distribution. The nonlinearity of some explanatory variables was also captured by using TPRS, controlled by the smoothing factor which was automatically tuned. This resulted in **`age`** becoming significant in the final model. In terms of inference, the model coefficients and the shape of the smoothened variables also do not contradict common knowledge: The number of illnesses **`illness`**, days of reduced activity due to illness or injury **`reduced`**, age **`age`**, and general health questionnaire score **`health`** which all indicate deteriorating health conditions are positively associated with the number of doctor visits. The nonlinearity in the **`income`** variable can be explained as poverty being associated with poor health conditions, whereas doctor visits are more affordable for high-income people.

In terms of the disadvantages, the final model still highlights some uncaptured patterns in the data, as indicated in the residual plots. The model is also not parsimonious as it still includes some insignificant variables (e.g., **`nchronic`**, **`lchronic`**) and the **`age`** and **`health`** variables are reduced to simple lines. Furthermore, the analysis did not analyse variable interactions which could be meaningful in improving the model.

Moving forward, the uncaptured patterns in the data can potentially be addressed by enriching the dataset with more variables or exploring different variable transformations and variable interactions. In terms of model specification, since the dependent variable **`visits`** only has 10 levels and the dataset contains 5190 observations, a classification model such as logistic regression could potentially provide a better fit by modelling each level independently. Furthermore, because the dataset contains 4141 observations with zero visits, a zero-inflated model (e.g., zero-inflated negative binomial) could potentially offer improvement to the analysis by modelling the excess zeroes independently. However, these models assume that the zeroes are generated by a separate process, which might not be true. Finally, the final model could be simplified by removing the insignificant variables identified above, and keeping the **`age`** and **`health`** variables linear instead of smoothing them. By doing so, the effect of these variables can be interpreted directly by looking at their estimated coefficients.

Question 2.

(a)

The statistical representation of the model is as follows:

$P(\widehat{CertTaken = 1|DVRT, GenderFemale, Prestige}) = \Lambda (\beta_0 + \beta_1 \times DVRT + \beta_2 \times GenderFemale + \beta_3 \times Prestige)$

The estimated model is:

$P(\widehat{CertTaken = 1|DVRT, GenderFemale, Prestige}) = \Lambda (-7.486916 + 0.055578 \times DVRT + 0.496952 \times GenderFemale + 0.037782 \times Prestige)$

The estimated values of the model parameters (assuming a 0.05 significance level) are:

	Mean	Lower bound	Upper bound
Intercept	-7.48692	-8.4104100	-6.5634220
DVRT	0.055578	0.0472720	0.0638840
GenderFemale	0.496952	0.2791590	0.7147450
Prestige	0.037782	0.0301890	0.0453750

(b)

With the other variables remaining constant, the odds of gaining Leaving Certificates as a female is 1.6437 $(\exp(0.496952) = 1.6437)$ times the odds of gaining Leaving Certificates as a male.
The association between gender and gaining Leaving Certificates is statistically significant (p-value = 0.0225 < 0.05). This indicates that females, on average, have higher odds of gaining leaving certificate than males.

(c)

When gender is male, DVRT score is 100 and prestige score is 40, the probability of taking the Leaving Certificate is:

$$P(\widehat{CertTaken = 1|DVRT = 100, Prestige = 40, GenderFemale = 0}) = \frac{1}{1 + e^{-(-7.486916 + 100*0.055578 + 0*0.496952 + 40*0.037782)}} \approx 0.397$$

Therefore, the probability of taking a Leaving Certificate for the boy is approximately 0.397.

(d)

About the roles of gender, the result indicates that the odds of gaining Leaving Certificate as a female is 0.9951 **(**exp(-0.004924) = 0.9951**)** times the odds of gaining Leaving Certificate as a male.

The p-value associated with GenderFemale shows that gender does not have a statistically significant impact on CertTaken given the p-value being 0.98. However, school type appears to have a significant effect on CertTaken based on the >99% confidence interval. The result indicates that the odds of gaining Leaving Certificate in a vocational school is 0.0429 **(**exp(**-**3.149843) = 0.0429**)** times the odds of gaining Leaving Certificate in a secondary school. Thus, the variable Gender should be removed from this model.

(e)

The hypothesis being tested in this table is whether adding a new variable will improve the model fit. Precisely, it is:

H_0: The coefficients of the added variable(s) are zero

H_1: The coefficients of the added variable(s) are not zero

If the p-values are less than 0.05, then we reject the null hypothesis, assuming a 0.05 significance level. Therefore, the p-value of 2.731e-07 indicates that adding Gender and Prestige has significantly improved Model 1, and the p-value of < 2.2e-16 indicates that SchoolType has significantly improved Model 2. These additions have resulted in the reductions of variance of 30.227 and 97.187, respectively. It also suggests that the improvement to the model made by adding SchoolType is much larger than adding Gender and Prestige. Based on the above reasons, we can conclude that Model 3 is the best model among the three models and should be selected.

(f)

The “residuals versus linear predictors" plot shows 2 well-defined curves, signifying a poor model fit. This is because ideally, the residuals should randomly scatter around the zero horizontal line. It also means that there is still a pattern not captured by the current model. However, there are still cases where the average residuals along the vertical axis are close to zero, which lie at the left and right-most part of the plot. This indicates that the model works much better in extreme cases compared to anywhere else.

The 2-curve pattern can be explained by the first plot which plots the response variable against the linear predictor. It is clear that the linear predictor is not useful for separating the 2 classes **yes** and **no**, as almost all the value range of the linear predictor contains both classes **yes** and **no** (again, except for the two extremes in the value range of the linear predictor) Therefore, the model could be improved by introducing other explanatory variables.