# Generalized Linear Models - Assignment I

Vincent Buekers - r0754046

22nd April 2019

# Contents

# 1 Introduction

The data of interest contain information about student performance of secondary school students in Portugal, in particular about their performance on mathematics. 35 variables are recorded for a sample of 395 Portuguese students across two different schools. However, for the scope of this analysis only a selection of these variables is considered. For the first analysis, it is of interest to classify students to either pass or fail a math test, according to a set of predictors. As for the second analysis, absenteeism among students is investigated. Two well known members of the generalized linear model framework, namely logistic regression and Poisson regression, are used respectively. The statistical computing software R serves as the main data analysis tool, used to implement both frequentist and Bayesian approaches to the aforementioned statistical modelling techniques. The scripts for both analyses are added in appendix.

# 2 Analysis I

For each student in the data sample, the final decision on the math test is encoded as a binary variable. That is, students either pass or fail the test. Several candidate predictors are investigated in terms of their relationship to this score, namely: sex, age, Pstatus, Medu, famsup, paid, higher, internet and romantic (see documentation). Relationships between such binary outcomes and aforementioned covariates are typically modelled via logistic regression.

The data on the math test measured by the response variable ‘Score’ is binomially distributed, as is visualized in Figure 1 below. More specifically, 130 out of the 395 students have failed the test while the remaining 165 students managed to pass on the math test. This corresponds to 33% and 67% of students respectively. Hence the empirical distribution of the response variable can be summarized as $Score\ \tilde{}Bin(395, 0.67)$. Note that this is an asymmetric binomial distribution.

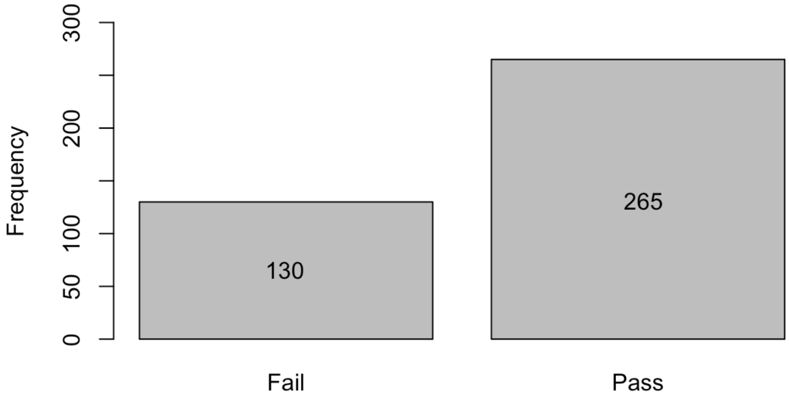


*Figure 1: Empirical Binomial distribution of the math test scores (Pass or Fail)*

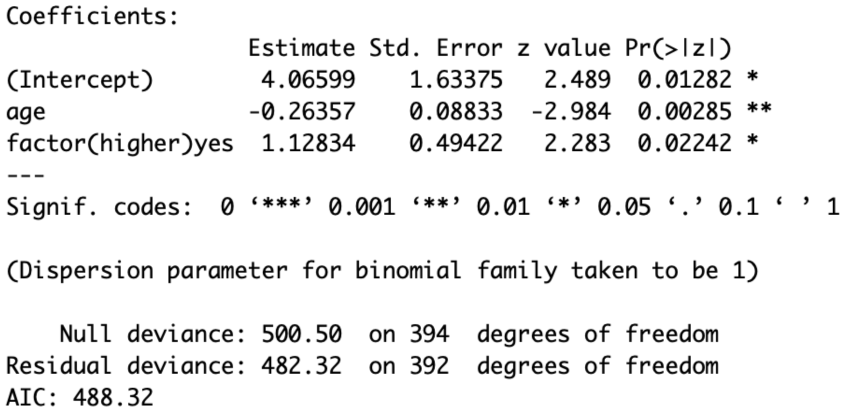
## 2.1 Frequentist Logistic Regression

In order to model the math test score, a logistic regression with the aforementioned covariates is conducted. Regardless of starting from the null model (intercept only) or the full model (including all candidate predictors), an identical resulting model is obtained based on the Chi-squared test statistic as well as the Likelihood-Ratio test. More specifically, by means of individual addition or deletion of covariates, both the variables age and higher are considered significant predictors with respect to the response variable Score.

Hence the probability to pass is modelled as:

$$\Pr(\text{Score} = Pass \mid \text{age, higher}) = \frac{\exp(\beta_0 + \beta_1 \text{age} + \beta_2 \text{higher})}{1 + \exp(\beta_0 + \beta_1 \text{age} + \beta_2 \text{higher})} \tag{1}$$

This model yields the results provided below in Figure 2. These results can be easily obtained by using the base R function glm, as is shown in Appendix I.



```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        4.06599    1.63375   2.489  0.01282 *
age               -0.26357    0.08833  -2.984  0.00285 **
factor(higher)yes  1.12834    0.49422   2.283  0.02242 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 500.50  on 394  degrees of freedom
Residual deviance: 482.32  on 392  degrees of freedom
AIC: 488.32
```

*Figure 2: Logistic Regression Output*

Before moving on to the interpretation of the coefficient estimates, several important aspects are first outlined in the following sections. The reason figure 2 is provided is for comparative purposes later on.

### 2.1.1 Link function

The link function of a logistic regression model determines the relationship between the linear predictor and the mean response. Two very commonly chosen link functions for binomial data are the logit link and the probit link. Both the logit and the probit link are appropriate for symmetric relationships, while the complementary log-log is better suited for relationships that do not correspond to a mirror image. Furthermore, also the regular log link and the Cauchy cumulative distribution function are offered in R. These alternatives, together with the complementary log-log, all result in practically equivalent residual deviances and AIC values. Yet the coefficient estimates differ substantially. For this application, the logit and probit link are examined in more detail.

In theory the coefficient estimates corresponding to the logit and probit link function should only differ up to a constant term, namely $\beta^L = \pi/\sqrt{3}\beta^P \approx 1.814\beta^P$.Upon evaluation of $\beta^L/\beta^P$, one should hence find values close to 1.814. In this setting, these ratio's evaluate to 1.667, 1.657 and 1.600 for $\beta_0$, $\beta_1$ and $\beta_2$ respectively. Although close, it is possible that the normal approximation to the binomial distribution is sub-optimal as the empirical data are asymmetric. Moreover, the canonical link (logit link) enjoys several properties such as residuals summing up to zero for models containing an intercept (which is the case in this application) and a unique maximum for the ML procedure. Hence, the subsequent analysis is carried out using the logit link function.

### 2.1.2 Goodness-of-Fit

To examine the goodness-of-fit, the observed frequencies can be compared with the estimated ones. Neither the Pearson test nor the deviance test can be used since not only categorical covariates are included in model (1), but also a continuous one (age). One can however use the Hosmer-Lemeshow statistic in such settings. Unfortunately, the Hosmer-Lemeshow function provided in the lecture notes reported an error (breaks not being unique). Consequently, the function hoslem.test() was used from the R package ResourceSelection. The Hosmer-Lemeshow test statistic with regards to (1) reported it to be a poor fit: $\chi^2 = 0.92275, df = 8, p - value = 0.9987$).

### 2.1.3 Quality of Prediction

In order to assess the predictive performance of the model, different measures are available in a logistic regression context. Among the various coefficients of determination, the Nagelkerke $R^2$ is commonly chosen. Note that the R package fmsb is used to compute it. The value of the Nagelkerke $R^2$ for model (1) is about 0.063. Hence, the predictive performance of this model is quite poor.

Aside from the different interpretations of the coefficient of determination, a measure known as the Concordance ($C$) is frequently reported as well. It gives insight into the proportion of pairs (any two outcomes) which have predictions in the same direction as the observed binary outcomes of the corresponding subjects. The value obtained for $C = 51.36\%$. Hence, little over half of the predictions correspond to the pairwise ordering of the labeled data. This is once again not very satisfactory.

As far as the actual classification based on the predictions is concerned, the model manages to correctly classify 271 out of the 395 observations. This information is summarized below in a so called "Confusion Matrix". The assigned classes are based on the cut-off value of 0.5 for a probability as predicted by the model discussed above.

| | Fail | Pass |
|---|---|---|
| Predicted Fail | 15 | 9 |
| Predicted Pass | 115 | 256 |

*Table 1: Confusion Matrix*

A popular measure to derive from such matrices is the Accuracy, which is defined as the number of true positives and true negatives divided by the total amount of observations. Hence, the accuracy for this model can be calculated as follows:

$$Accuracy = \frac{15 + 256}{15 + 9 + 115 + 256} = 0.6861$$

This is a very poor result since the binomial distribution of the test score has an empirical success probability of about 67%. In practice this would mean that when one predicts all students of this sample to pass, this would be an accurate prediction 67% of the students. Hence, the incremental predictive performance obtained by implementing the model as such is only very minor.

### 2.1.4 Residual Diagnostics & Outlier Detection

As outlined above, the model diagnostics are somewhat disappointing. A proper analysis of the residuals and possible outliers could give more insight in the cause of these inadequate results. Indeed, the residuals for the students older than 19 (ages 20, 21 and 22) deviated substantially as shown in the left plot of Figure 3 below. When removing these 5 observations, a much more satisfactory residual plot (Figure 3, right plot) is obtained. Logically, students that are already in their 20's are not quite representative of a high school environment whatsoever. Hence it should be justified to leave out these 5 observations.

Aside from the deviance residuals shown below, little else showed improvements for the logistic regression that excludes the outlying observations. As shown below in figure 4, the AIC and deviance residual decreased only slightly. There exists a rule-of-thumb that a preference between models is only justified for differences in AIC values exceeding 5. In this case the difference is 5.65, so by that rule this model is preferred. Moreover, the parameter estimates remain largely unchanged as do their significance levels. Consequently, it is not suprising that the quality of prediction does not improve as indicated by, among other measures, the $NagelkerkeR^2$. The $NagelkerkeR^2$ corresponding to the model without the outliers is 0.0545. As for the Concordance measure, this now evaluates to $C = 50.56\%$.

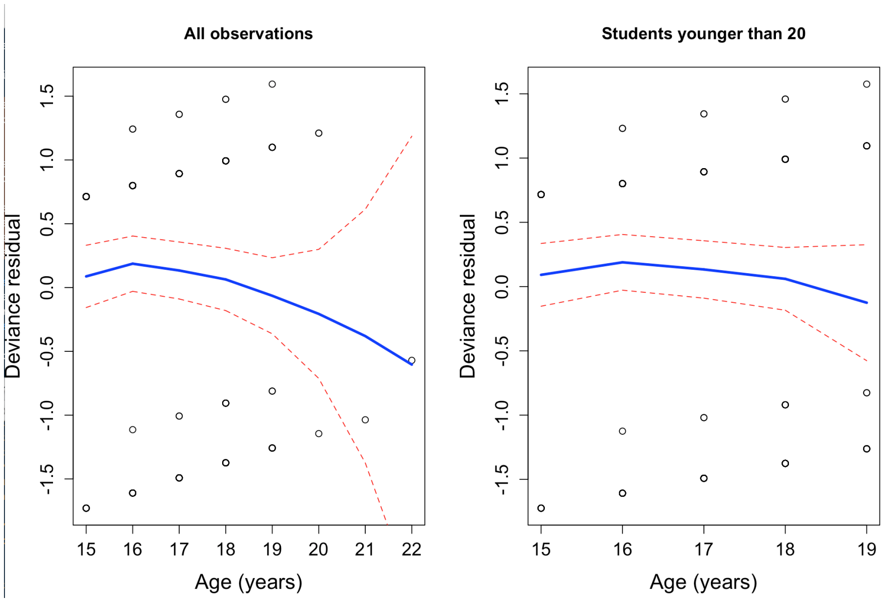
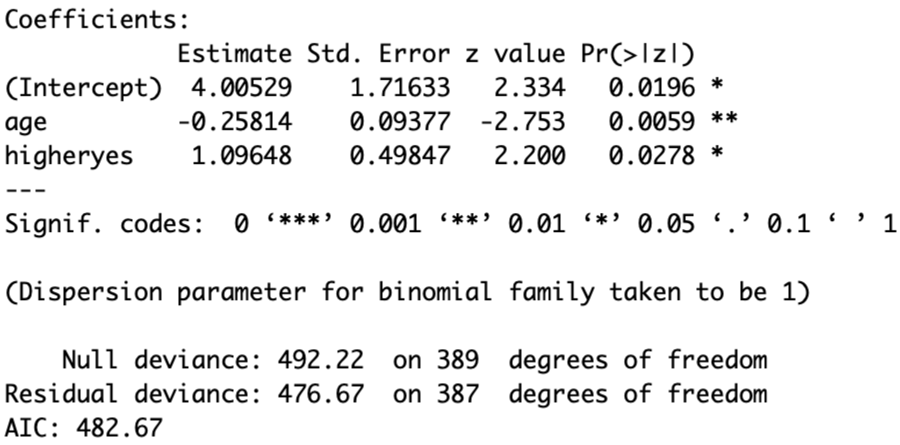


*Figure 3: Deviance Residuals. Left: model based on all observations. Right: Model using only the students younger than 20 years old*

### 2.1.5 Model Fit & Discussion

Finally, the interpretation of the results can be done for the model based on the teenage students (for which the output is provided in figure 4). The intercept coefficient is the mean response value when all regressors evaluate to 0. However, considering that the age of the students is included as one of the regressors, this is not very insightful. The values for $\beta_{age}$ and $\beta_{higher}$ are about -0.26 and 1.10 respectively. These values are essentially log odds ratios with respect to the test Score.



```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.00529    1.71633   2.334   0.0196 *
age         -0.25814    0.09377  -2.753   0.0059 **
higheryes    1.09648    0.49847   2.200   0.0278 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 492.22  on 389  degrees of freedom
Residual deviance: 476.67  on 387  degrees of freedom
AIC: 482.67
```

*Figure 4: Output of logistic regression without outliers*

Instead of interpreting the coefficient estimates as the log odds ratio, a more natural approach is to transform them into actual odds ratio's. Note that, for multiple regression, changes in the regressors affect the odds ratio in a multiplicative way. Consequently, the odds of passing on the math test seems to decrease as students get older, namely by $exp(-0.25814) = 0.773$ for growing 1 year older. On the other hand, given that students want to follow higher education, the odds ratio is nearly triple in comparison to those who don't (keeping age constant), since $exp(1.09648) = 2.994$.

Finally, the relationship is also displayed graphically in figure 5. The signature S-shape of the sigmoid function is hardly recognizable, which is not surprising given the poor model fit and low predictive performance.

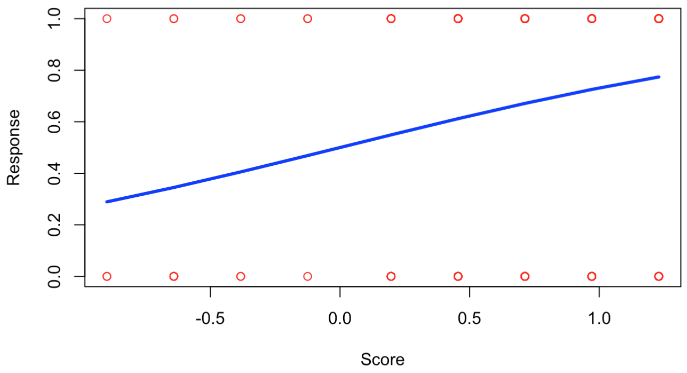
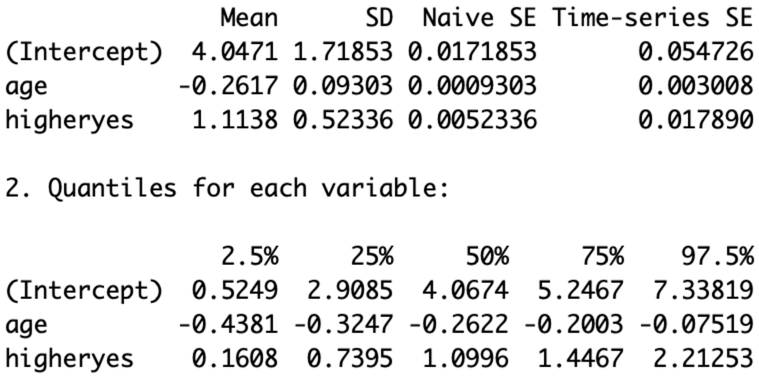


*Figure 5: Model fit for model without outliers*

## 2.2 Bayesian Logistic Regression

In contrast to the frequentist analysis outlined above, also the Bayesian approach to logistic regression is now considered. The model as specified in (1) is now regressed in a Bayesian manner, for which the R package MCMCpack has been used to do so. Note that the also here only the teenage students are taken into consideration. The output obtained from this Bayesian regression is summarized in Figure 6.



```
               Mean      SD  Naive SE Time-series SE
(Intercept)  4.0471 1.71853 0.0171853       0.054726
age         -0.2617 0.09303 0.0009303       0.003008
higheryes    1.1138 0.52336 0.0052336       0.017890

2. Quantiles for each variable:

               2.5%     25%     50%     75%    97.5%
(Intercept)  0.5249  2.9085  4.0674  5.2467  7.33819
age         -0.4381 -0.3247 -0.2622 -0.2003 -0.07519
higheryes    0.1608  0.7395  1.0996  1.4467  2.21253
```

*Figure 6: Output of Bayesian Logistic Regression*

The values obtained for the Bayesian logistic regression are approximately the same as the estimates of the frequentist logistic regression. Yet, the statistical interpretation is now different; rather than a point estimate for the parameters, a posterior probability distribution is obtained since the parameter is assumed to be random in a Bayesian context. The reported mean values are calculated based on a Monte Carlo sample. This sampling process in and of itself involves some degree of inaccuracy, which is represented by the Naive or the Time-series Standard error (preferred). These standard errors can be used to compute classical $100(1-\alpha)\%$ confidence intervals to reflect the precision of the posterior mean. Since the mean values of these posterior densities are the same as the values of the frequentist parameter estimates, the practical interpretation with respect to the high school students obviously does stay unchanged.

In order to verify whether or not the Gibbs sampling procedure has produced a sample of the posterior distribution, one can examine the stability, or rather the stationarity of the sampling iterations. Considering this iterative procedure, it is clear why the Time-series SE is preferred over the Naive one. As is shown below in figure 7, indeed the trace plots (left) indicates stationary iterations for the according posterior parameter distributions (right).

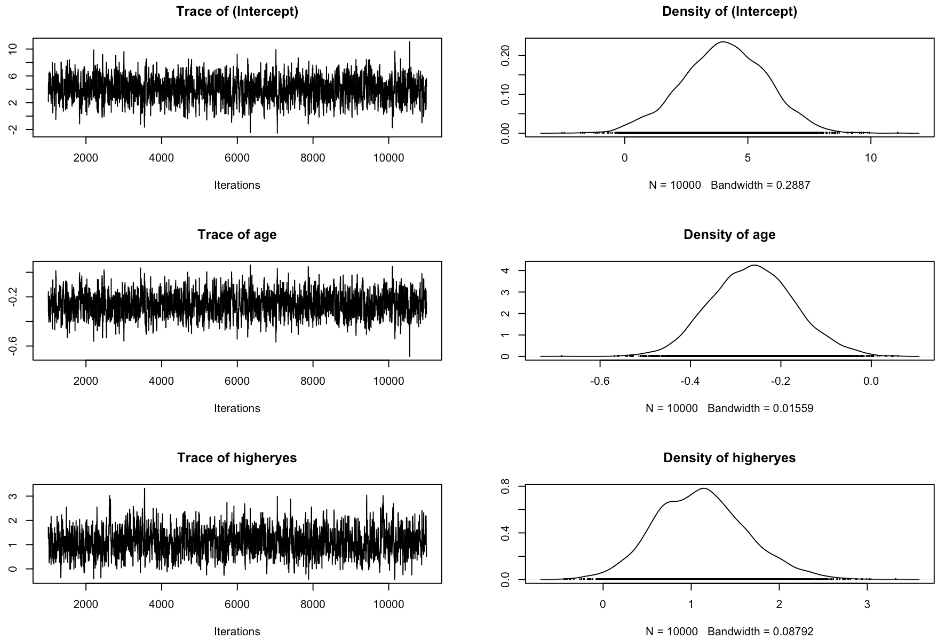


*Figure 7: Trace plots and posterior sample distributions for the parameters of the Bayesian logistic regression*

# 3 Analysis II

In this second analysis, a closer look is taken at the absenteeism among the students. The variable absences is measured by recording the amount of school absences per student. This is obviously a count variable, or rather the number of ”successes” that are observed. When relating this type of response to covariates, a commonly implemented method is the Poisson regression model. Both a Frequentist and Bayesian approach are carried out in the following sections.

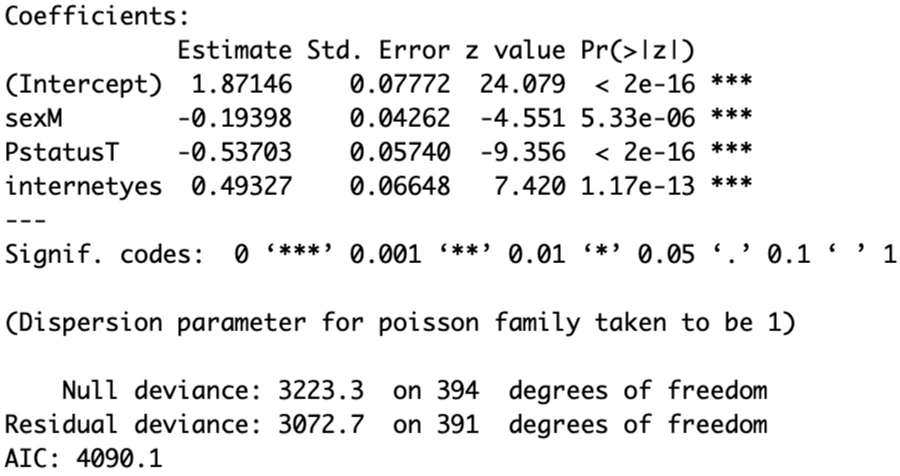
## 3.1 Frequentist Poisson Regression

The candidate predictors considered with respect to the amount of school absences are the following: sex, Pstatus, famsup and internet (see documentation). Consequently, a Poisson regression is performed with these covariates using the canonical link function (log link).

Based on both the Chi-squared and the likelihood-ratio test, the variable famsup is not a significant predictor w.r.t. the school absences. Hence, the resulting model can be summarized as follows.

$$\log(\text{Absences}) \ = \beta_0 + \beta_1 \text{sex} + \beta_2 \text{Pstatus} + \beta_3 \text{internet} \tag{2}$$

The results corresponding to this model are provided below in Figure 8. This can easily be done by using the base R function glm as shown in Appendix II.



```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.87146    0.07772  24.079  < 2e-16 ***
sexM        -0.19398    0.04262  -4.551 5.33e-06 ***
PstatusT    -0.53703    0.05740  -9.356  < 2e-16 ***
internetyes  0.49327    0.06648   7.420 1.17e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3223.3  on 394  degrees of freedom
Residual deviance: 3072.7  on 391  degrees of freedom
AIC: 4090.1
```

*Figure 8: Poisson Regression Results*

At first glance it seems that all considered predictors are highly significant. Yet when looking at the residual deviance, there is an obvious problem with this model. For an appropriate Poisson regression the value of the residual deviance should approximately be equal to the expected value of the Chi-squared distribution at convergence, which is in fact the degrees of freedom. Clearly, $D = 3072.7$ is not even remotely close to $n - p = 391$. This problem will be further examined in the following section and the interpretation of the results follows once it has been dealt with accordingly.

### 3.1.1 Overdispersion

As the residual deviance already indicated an issue with the Poisson model, it is necessary to investigate the overdispersion for model (2) as a sort of goodness-of-fit assessment. To that end, the histogram of the number of school absences and the corresponding Poisson fit is examined in Figure 9.

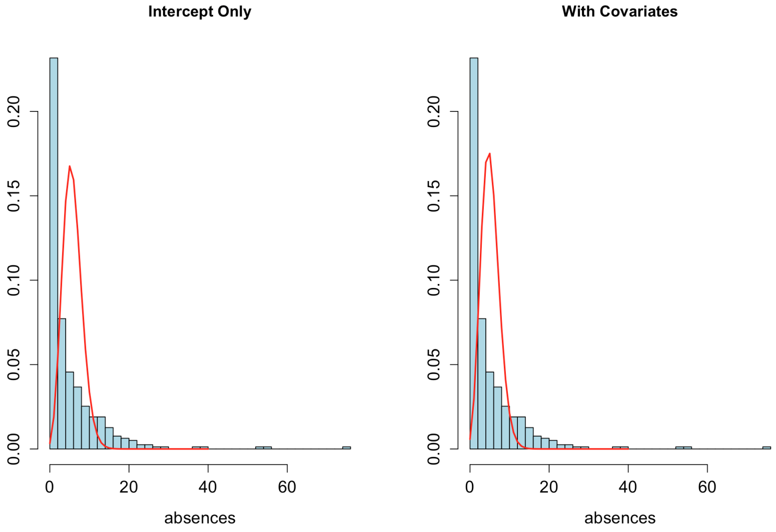
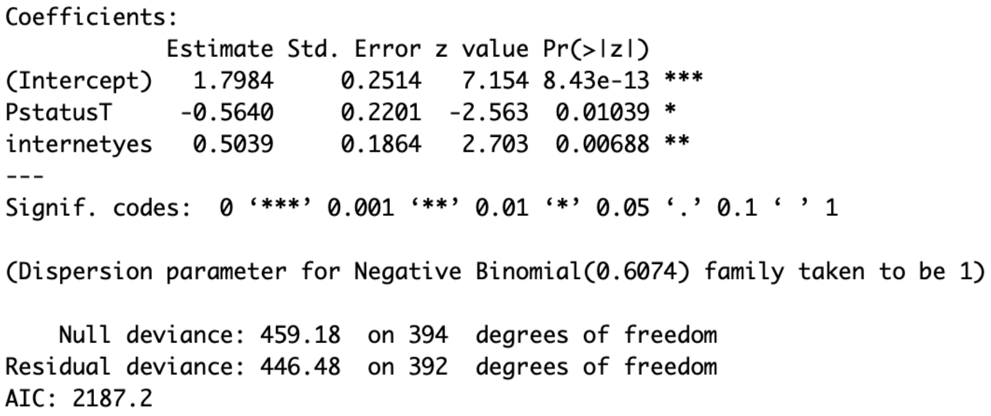


*Figure 9: Histogram and corresponding Poisson fit. Left: Model containing only an intercept. Right: Model with covariates as stated in (2).*

Clearly there is a problem with overdispersion in the model with or without the inclusion of the covariates. Hence, the inclusion of the predictors have not provided an adequate fit. A possible cause of this issue could be the fact that the data contains students from two different schools. It is not unlikely that the probability to be absent from school is different across these two schools. A possible way to address this problem could be to consider a negative binomial regression instead of a Poisson regression. This alternative approach is outlined in the following section.

## 3.2 Negative Binomial Regression

The negative binomial regression model is first considered with all the candidate predictors suggested in the documentation. Once again the redundant covariates, as indicated by the likelihood-ratio or the Chi-squared test statistic, are left out of the model. Now not only famsup, but also the factor sex is considered insignificant with respect to the school absences. The resulting model output is given below in figure 10.



```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)   1.7984     0.2514   7.154 8.43e-13 ***
PstatusT     -0.5640     0.2201  -2.563  0.01039 *  
internetyes   0.5039     0.1864   2.703  0.00688 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.6074) family taken to be 1)

    Null deviance: 459.18  on 394  degrees of freedom
Residual deviance: 446.48  on 392  degrees of freedom
AIC: 2187.2
```

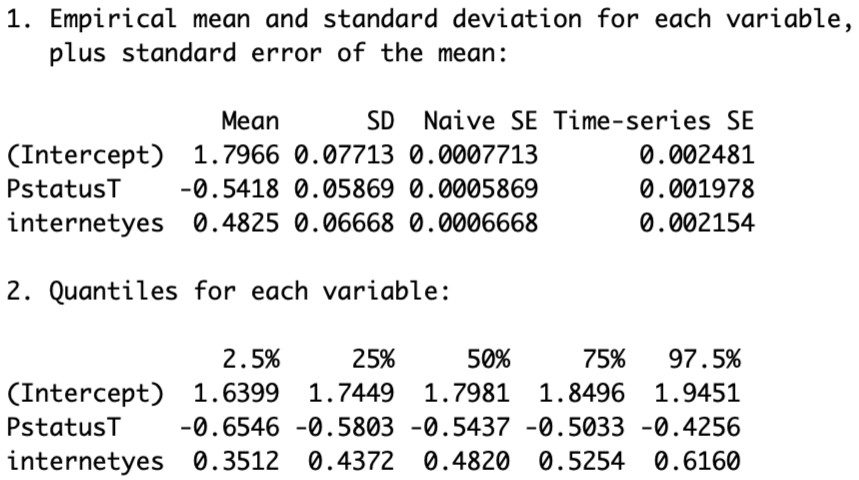
*Figure 10: Negative Binomial regression output*

In terms of the overdispersion, there is a considerable improvement since the residual deviance is now much closer to the degrees of freedom of the corresponding Chi-squared distribution (may it not be perfect yet). Furthermore, the factors Pstatus and internet are still significant yet not as extremely significant like with the Poisson regression. Note that also the AIC (=2187.2) is much lower in than that of the Poisson model (=4090.1).

As was the case with the logistic regression, it possibly makes more practical sense to apply a transformation to the coefficient estimates for the purpose of interpretation. The expected amount of school absences, for Pstatus = 0 and internet = 0, is exp(1.7984) = 6.04. Given that all other regressors are kept constant, the multiplicative change resulting from parents living together (as opposed to seperated), is exp(-0.5639607) = 0.57. So on average, parents living together tends to result in less absenteeism. On the other hand, having access to internet increases the expected amount of school absences by exp(0.5038519) = 1.6550841. Thus, students that can connect to the internet at home tend to be absent from school more on average.

## 3.3 Bayesian Poisson Regression

Considering the inapt nature of the Poisson regression in this context, the negative binomial regression ought to be a better reference as for which covariates to include in the Bayesian logistic regression. Hence, a Bayesian Poisson regression is now considered with the variables Pstatus and internet serving as regressors. This yields the output given below in Figure 11.



```
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

              Mean      SD  Naive SE Time-series SE
(Intercept)  1.7966 0.07713 0.0007713       0.002481
PstatusT    -0.5418 0.05869 0.0005869       0.001978
internetyes  0.4825 0.06668 0.0006668       0.002154

2. Quantiles for each variable:

              2.5%     25%     50%     75%   97.5%
(Intercept)  1.6399  1.7449  1.7981  1.8496  1.9451
PstatusT    -0.6546 -0.5803 -0.5437 -0.5033 -0.4256
internetyes  0.3512  0.4372  0.4820  0.5254  0.6160
```

*Figure 11: Bayesian Poisson Regression output*

The values for the Bayesian approach are very similar to the estimates of the negative binomial regression. Fortunately, this a reassurance that the frequentist Poisson regression is not suitable to model the school absences. Rather than estimates however, once again a posterior distribution for the different parameters is returned. The approximate posterior mean is again obtained via the Monte carlo sample mean, with its corresponding Monte Carlo error. Although the statistical meaning of the parameters is fundamentally different now, the practical interpretation of the effects with regards to the school absences remains the same of course.

Lastly it is also recommendable to verify whether the Gibbs sampling procedure has produced a stable approximation of the posterior distributions for the model parameters. Indeed, the sample procedure is stable across iterations for each of the model parameters, as is shown in the trace plots below (Figure 12, left plots).

Hence, the samples for the parameter estimates are likely a good approximation of the according posterior distribution.

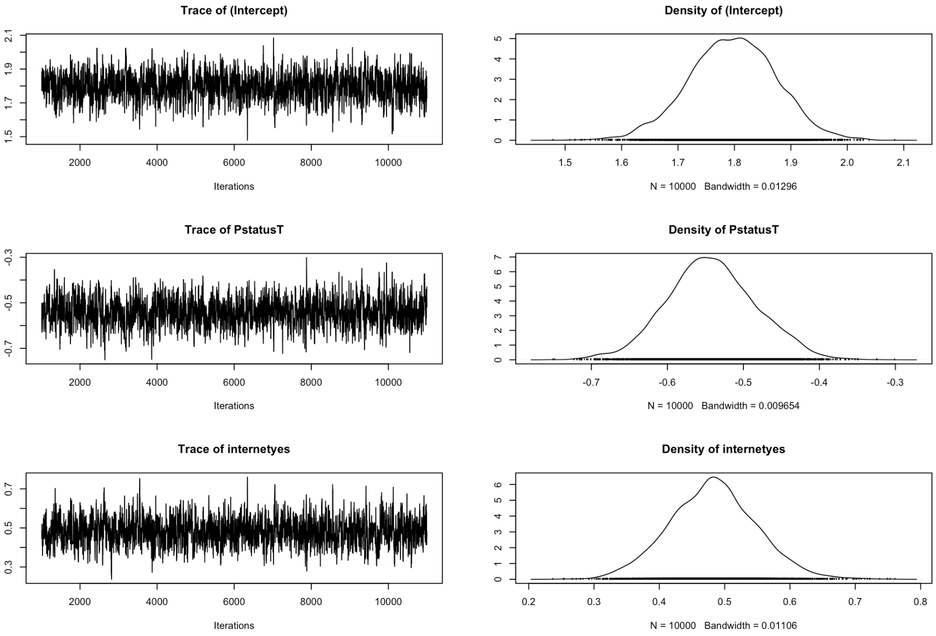


*Figure 12: Trace plots and Parameter distributions*

# 4 Appendix

## 4.1 Analysis I - R Script

```
setwd("~/Desktop/GLM/Assignment")
data = read.csv("student-mat.csv", header=T)

library(dplyr)

# subset variables of interest (for analysis I and II)
data.sub = data %>%
    select(sex, age, Pstatus, Medu, famsup, paid, higher,
    internet, romantic, Score, absences)
attach(data.sub)

str(data.sub)

# frequencies
table(data.sub$Score)

# Barplot of Score
barplot(table(data.sub$Score), ylim=c(0,300), ylab = "Frequency")
locator()
text(0.6955059, 65.69598, '130')
text(1.918371, 131.2172, '265')

# Code as 0/1
data.sub$Score = ifelse(data.sub$Score == "Pass", 1, 0)

# Frequentist logistic starting from null model
model1.logit = glm(Score ~ 1, family=binomial(link="logit"), data.sub)

# Add age & higher
add1(model1.logit ,~ . + age + paid + Medu + higher + romantic,
test="Chisq")
model1.logit = update(model1.logit, . ~ . + age)
add1(model1.logit ,~ . + paid + Medu + higher + romantic, test="Chisq")
model1.logit = update(model1.logit, . ~ . + factor(higher))
add1(model1.logit ,~ . + paid + Medu + romantic, test="Chisq")
# None of the candidate predictors are considered significant
additions
summary(model1.logit)

# Frequentist logistic starting from full model
model2.logit = glm(Score ~ age + paid + Medu + higher + romantic,
family= binomial(link= "logit"), data.sub)

# Remove paid, romantic and Medu
drop1(model2.logit, test="Chisq")
model2.logit = update(model2.logit, ~ . -paid)
```

```
drop1(model2.logit, test=”Chisq”)
model2.logit = update(model2.logit, ~ . -romantic)
drop1(model2.logit, test=”Chisq”)
model2.logit = update(model2.logit, ~ . -Medu)
# Same resulting model irrespective of starting from null or full
model specification
summary(model2.logit)

# Comparison with other link functions
model.probit = glm(Score ~ age + factor(higher),
family=binomial(link = ”probit”), data.sub)
summary(model.probit)

model.clog = glm(Score ~ age + factor(higher),
family=binomial(link = ”cloglog”), data.sub)
summary(model.clog)

model.log = glm(Score ~ age + factor(higher),
family=binomial(link = ”log”), data.sub)
summary(model.log)

model.cauchit = glm(Score ~ age + factor(higher),
family=binomial(link = ”cauchit”), data.sub)
summary(model.cauchit)

# Ratio should approximate 1.814 (=pi/sqrt(3))
coeff.logit = summary(model1.logit)$coefficients[1:3]
coeff.probit = summary(model.probit)$coefficients[1:3]
as.data.frame(coeff.logit / coeff.probit, row.names = c(”Beta_0”,
”Beta_1”, ”Beta_2”))

#--------------------
# Goodness-of-fit
#--------------------
# Hosmer-Lemeshow
# function provided in lecture notes reported error due to bug...
hosmerlem = function(y, yhat, g=10)
{
  cutyhat = cut(yhat,breaks = quantile(yhat, probs = seq(0,1, 1/g)), include.lowest=TRUE)
  obs = xtabs(cbind(1 - y, y) ~ cutyhat)
  expect = xtabs(cbind(1 - yhat, yhat) ~ cutyhat)
  chisq = sum((obs - expect)^2/expect)
  P = 1 - pchisq(chisq, g - 2)
  return(list(chisq=chisq,p.value=P))
}

hosmerlem(y=Score, yhat=fitted(model1.logit))

# Hence I will compare functions from different packages for consistency
library(ResourceSelection)
hoslem.test(data.sub$Score, fitted(model1.logit), g=10)
```

```
# Poor fit

library(generalhoslem)
logitgof(data.sub$Score, fitted(model1.logit), g = 10, ord = FALSE)
# reports df = 3, while it should be 8... Not sure if this is a reliable function

#-------------------
# Prediction quality
#-------------------
library(fmsb)
NagelkerkeR2(model1.logit)
# Very poor predictive quality

# Concordance measure
OptimisedConc = function(model)
  {
  Data = cbind(model$y, model$fitted.values)
  ones = Data[Data[,1] == 1,]
  zeros = Data[Data[,1] == 0,]
  conc=matrix(0, dim(zeros)[1], dim(ones)[1])
  disc=matrix(0, dim(zeros)[1], dim(ones)[1])
  ties=matrix(0, dim(zeros)[1], dim(ones)[1])
  for (j in 1:dim(zeros)[1])
  {
    for (i in 1:dim(ones)[1])
    {
      if (ones[i,2]>zeros[j,2])
      {conc[j,i]=1}
      else if (ones[i,2]<zeros[j,2])
      {disc[j,i]=1}
      else if (ones[i,2]==zeros[j,2])
      {ties[j,i]=1}
    }
  }
  Pairs = dim(zeros)[1]*dim(ones)[1]
  PercentConcordance = (sum(conc)/Pairs)*100
  PercentDiscordance = (sum(disc)/Pairs)*100
  PercentTied = (sum(ties)/Pairs)*100
  return(list("Percent␣Concordance" = PercentConcordance,
              "Percent␣Discordance" = PercentDiscordance,
              "Percent␣Tied" = PercentTied,
              "Pairs" = Pairs))
}

OptimisedConc(model1.logit)
# Concordance measure: C = 51.36%

#---------------------
# Residual diagnostics
#---------------------
# Deviance residuals
```

```
r.dev = residuals(model1.logit , type = "deviance")
summary(r.dev)
par(mfrow=c(1,2))
#hist(r.dev)

plot(data.sub$age,r.dev,xlab="Age␣(years)",ylab="Deviance
residual", main="All␣observations",
      cex.lab=1.5,cex.axis=1.3)
loess.dev <- loess(r.dev~data.sub$age)
lo.pred <- predict(loess.dev, se=T)

orderage <- order(data.sub$age)
lines(data.sub$age[orderage],lo.pred$fit[orderage],col="blue",w
=3)
lines(data.sub$age[orderage],lo.pred$fit[orderage]+2*lo.pred$so
erage], lty=2,col="red")
lines(data.sub$age[orderage],lo.pred$fit[orderage]-2*lo.pred$so
derage], lty=2,col="red")

# Remove influential observations
which(age==22 | age==21 | age==20)
data.no.outlier = data.sub[-which(age==22|age==21|age==20),]

model3.logit= glm(Score ~ age + higher,
family=binomial(link="logit"), data.no.outlier)
summary(model3.logit)
Odds.ratio = exp(summary(model3.logit)$coefficients[2:3])

hoslem.test(data.no.outlier$Score, fitted(model3.logit), g=10)

NagelkerkeR2(model3.logit)

OptimisedConc(model3.logit)

# Residual plot without outliers
r.dev = residuals(model3.logit , type = "deviance")
plot(data.no.outlier$age,r.dev, xlab="Age(years)",
      ylab="Deviance␣residual",
      main = "Students␣younger␣than␣20",
      cex.lab=1.5,cex.axis=1.3)
loess.dev <- loess(r.dev~data.no.outlier$age)
lo.pred <- predict(loess.dev, se=T)

orderage <- order(data.no.outlier$age)
lines(data.no.outlier$age[orderage],lo.pred$fit[orderage],col=blue",lwd=3)
lines(data.no.outlier$age[orderage],lo.pred$fit[orderage]+2*lop
ed$s[orderage],␣lty=2,col="red")
lines(data.no.outlier$age[orderage],lo.pred$fit[orderage]-2*lop
ed$s[orderage],␣lty=2,col="red")
```

```
# Graphical plot of logistic regression analysis
score1 = model1.logit$linear.predictors
phat1 = model1.logit$fitted.values

plot(score1,phat1,xlab="Score",ylab="Response",type="n",
     xlim=c(min(score1),max(score1)),ylim=c(0,1),
     cex.lab=1.5,cex.axis=1.3)
orderscore1 = order(score1)
points(score1, data.sub$Score, col="red",pch=1)
lines(score1[orderscore1], phat1[orderscore1],
col="blue",lwd=3)

# Without outliers
score3= model3.logit$linear.predictors
phat3 = model3.logit$fitted.values

plot(score3,phat3,xlab="Score",ylab="Response",type="n",
     xlim=c(min(score3),max(score3)),ylim=c(0,1))
orderscore3 = order(score3)
points(score3, data.no.outlier$Score, col="red",pch=1)
lines(score3[orderscore3], phat3[orderscore3],
col="blue",lwd=3)

#----------
# Bayesian
#----------
library(MCMCpack)

model1.bayes = MCMClogit(Score ~ age + higher, family=binomial,
data.no.outlier)
summary(model1.bayes)

plot(model1.bayes)

detach(data.sub)
```

### 4.2 Analysis II - R Script

```r
setwd("~/Desktop/GLM/Assignment")
data = read.csv("student-mat.csv", header=T)

library(tidyverse)

data.sub = data %>%
    select(sex, age, Pstatus, Medu, famsup, paid, higher,
    internet,romantic, Score, absences)
attach(data.sub)
#-------------
# Frequentist
#-------------
# Poisson regression
model0.ps = glm(absences ~ 1, family = poisson(link = "log"),
data.sub)
summary(model0.ps)

model1.ps = glm(absences ~ sex + Pstatus + famsup + internet,
                family = poisson(link = "log"), data.sub)
summary(model1.ps)

# Histgrams for null model and full model
par(mfrow=c(1,2))
hist(absences, nclas=30,col="light blue",prob=T,
      xlab="absences", ylab="",main='Intercept Only',
      cex.lab=1.3, cex.axis=1.3)
lines(0:40,dpois(0:40,exp(1.74202)),col="red",lwd=2)

hist(absences, nclas=30,col="light blue",prob=T,
      xlab="absences", ylab="",main='With Covariates',
      cex.lab=1.3, cex.axis=1.3)
lines(0:40,dpois(0:40,exp(1.86380 -0.19217 -0.53724 + 0.01445 +
0.49100)),col="red",lwd=2)

# Check covariate significance
drop1(model1.ps, . ~ ., test="LRT")
model2.ps = update(model1.ps, . ~ . - famsup)
summary(model2.ps)

# Negative Binomial regression
library(MASS)
model1.nb = glm.nb(absences ~ sex + famsup + Pstatus +
internet,
data = data.sub)
summary(model1.nb)

# Check covariate significance
drop1(model1.nb, . ~ ., test="LRT")
```

```
model2.nb = update(model1.nb, . ~ . - famsup)
drop1(model2.nb, . ~ ., test=”LRT”)
model2.nb = update(model2.nb, . ~ . - sex)
summary(model2.nb)

exp(summary(model2.nb)$coefficients[1:3])

#-------------
# Bayesian
#-------------
library(MCMCpack)
model2.bayes = MCMCpoisson(absences ~ Pstatus + internet,
data.sub)
summary(model2.bayes)

plot(model2.bayes)

detach(data.sub)
```