Part II: Generalized Linear Models

Load Packages

Again, we must load the packages that will be used in the first part of this workshop.

```r
library(pastecs, quietly = TRUE)
library(lm.beta,  quietly = TRUE)
library(lmtest,  quietly = TRUE)
library(foreign,  quietly = TRUE)
library(lattice,  quietly = TRUE)
library(lme4,  quietly = TRUE)
library(nlme,  quietly = TRUE)
library(survival,  quietly = TRUE)
library(dplyr,  quietly = TRUE)
library(ggfortify,  quietly = TRUE)
library(survminer,  quietly = TRUE)
library(rms,  quietly = TRUE)
library(MASS, quietly = TRUE)
library(pscl, quietly = TRUE)
```

A generalized linear model (GLM) has three components:

- a random component with mean μ. Generally, the random component is the response variable Y_i.
- a systematic component, η_i, that relates the relates the explanatory variables,

$$\eta_i = \sum_{j=i}^{n} \beta_j x_{ij}$$

- a link function that relates the mean of the random to the systematic component

$$g(\mu) = \eta_i$$

Logistic regression

Logistic regression is a GLM used the model binary (0 or 1) data. The response variable must be binary and is assumed to follow a bernoulli distribution.

That said, logistic regression has the following components:

- a response binary variable, Y_i, that follows a bernoulli distribution with mean π_i.
- a systematic component, η_i, that relates the relates the explanatory variables,

$$\eta_i = \sum_{j=1}^{n} \beta_j x_{ij}$$

- a link function that relates the mean of the random to the systematic component

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{j=i}^{n} \beta_j x_{ij}.$$

$\log(\pi_i/(1-\pi_i))$ is known as the log odds.

Logistic regression

Data

Using the iris data, we create binary data. We add the column `Sepal.Width_binary` to iris. If the `Sepal.Width` is greater than the median then the associated value in `Sepal.Width_binary` is 1. Otherwise, `Sepal.Width_binary` is 0.

```
data <- iris
data$Sepal.Width_binary <- ifelse(data$Sepal.Width
                                  >= median(data$Sepal.Widt
                                  1, 0)
```

Logistic regression

Logistic Regression with only the constant term

Fitting only a constant term, the systematic component is

$$\eta_i = \beta_0.$$

```
logit <- glm(Sepal.Width_binary ~ 1, data = data, family = "binomial")
summary(logit)
```

```
##
## Call:
## glm(formula = Sepal.Width_binary ~ 1, family = "binomial", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3911  -1.3911   0.9778   0.9778   0.9778
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.4895     0.1682    2.91  0.00361 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 199.22  on 149  degrees of freedom
## Residual deviance: 199.22  on 149  degrees of freedom
## AIC: 201.22
##
```

Logistic regression

Logistic Regression with only the constant term

```
p_avg <- mean(data$Sepal.Width_binary)
log_odds_avg <- log(p_avg/(1-p_avg))
print(log_odds_avg)
```

```
## [1] 0.4895482
```

Logistic regression

Logistic Regression with Species

Fitting the species term, the systematic component is

$$\eta_i = 1 + \beta_2 X_{1i} + \beta_3 X_{2i}.$$

where

$$X_{1i} = \begin{cases} 1 & \text{if } i\text{th data point is versicolor} \\ 0 & \text{otherwise} \end{cases},$$

$$X_{2i} = \begin{cases} 1 & \text{if } i\text{th data point is virginica} \\ 0 & \text{otherwise} \end{cases}.$$

Logistic regression

Logistic Regression with Species

```
logit <- glm(Sepal.Width_binary ~ as.factor(Species), data = data, family = "binomial")
summary(logit)
```

```
##
## Call:
## glm(formula = Sepal.Width_binary ~ as.factor(Species), family = "binomial",
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5373  -0.8782   0.2857   1.0438   1.5096
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    3.1781     0.7215   4.405 1.06e-05 ***
## as.factor(Species)versicolor  -3.9318     0.7826  -5.024 5.06e-07 ***
## as.factor(Species)virginica   -2.8553     0.7763  -3.678 0.000235 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 199.22  on 149  degrees of freedom
## Residual deviance: 147.51  on 147  degrees of freedom
## AIC: 153.51
##
## Number of Fisher Scoring iterations: 5
```

Logistic regression

Logistic Regression with Species

Let's compare the results to the average log odds of each Species group

```
log_odds_avg_fun <- function(data){
  p_avg <- mean(data)
  log_odds_avg <- log(p_avg/(1-p_avg))
  return(log_odds_avg)
}

tapply(data$Sepal.Width_binary,
       data$Species, log_odds_avg_fun)
```

```
##     setosa versicolor  virginica
##  3.1780538 -0.7537718  0.3227734
```

The intercept corresponds to the average log odds of setosa as we would expect. However, the other coefficients do not correspond to the average log odds of the other species. Why?

Logistic regression

Logistic Regression with Species

From the formula, $\eta_i = 1 + \beta_2 X_{2i} + \beta_3 X_{3i}$, the log odds of versicolor actually corresponds to $1 + \beta_2$. The log odds of versicolor actually corresponds to $1 + \beta_3$.

```
coefficients<-unname(coef(logit))
print(c(coefficients[1],coefficients[1]+coefficients[2],
        coefficients[1]+coefficients[3]))
```

```
## [1]  3.1780537 -0.7537718  0.3227734
```

Logistic regression

Logistic Regression with continuous variable

COMPLETE

Logistic regression

Logistic Regression with continuous variable, Sepal.Length

Fitting the species term, the systematic component is

$$\eta_i = \beta_3 X_{1i}.$$

where X_{1i} = Sepal.Length of the ith data point.

```r
logit <- glm(Sepal.Width_binary ~ Sepal.Length,
             data = data, family = "binomial")
summary(logit)
```

Logistic regression

Logistic Regression with continuous variable, Sepal.Length

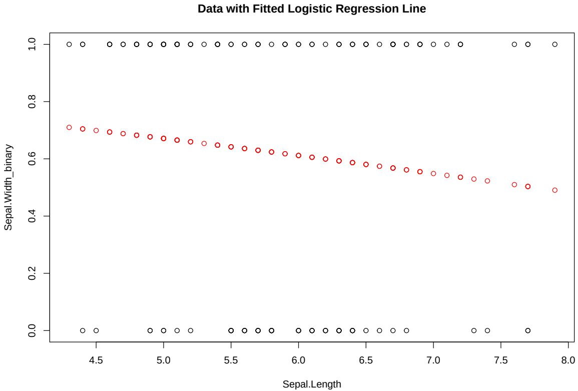
```
logit <- glm(Sepal.Width_binary ~ Sepal.Length,
             data = data, family = "binomial")
summary(logit)
```

```
##
## Call:
## glm(formula = Sepal.Width_binary ~ Sepal.Length, family = "binomial",
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5614  -1.3524   0.8883   0.9890   1.1936
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.0088     1.2176   1.650    0.099 .
## Sepal.Length  -0.2591     0.2050  -1.264    0.206
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 199.22  on 149  degrees of freedom
## Residual deviance: 197.61  on 148  degrees of freedom
## AIC: 201.61
##
## Number of Fisher Scoring iterations: 4
```

Logistic regression

Logistic Regression with continuous variable, Sepal.Length

```
plot(Sepal.Width_binary~Sepal.Length, data=data)
points(data$Sepal.Length[order(data$Sepal.Length)],
       logit$fitted[order(data$Sepal.Length)],  col="red")
title(main="Data with Fitted Logistic Regression Line")
```



Logistic regression

Logistic Regression with Species and Sepal.Length

Fitting the species term, the systematic component is

$$\eta_i = 1 + \beta_2 X_{1i} + \beta_3 X_{2i} + \beta_3 X_{3i}.$$

where

$$X_{1i} = \begin{cases} 1 & \text{if } i\text{th data point is versicolor} \\ 0 & \text{otherwise} \end{cases},$$

$$X_{2i} = \begin{cases} 1 & \text{if } i\text{th data point is virginica} \\ 0 & \text{otherwise} \end{cases}$$

and $X_{3i} =$ Sepal.Length of the ith data point.

Logistic regression

Logistic Regression with continuous variable, Sepal.Length

Fitting the logistic model accordingly,

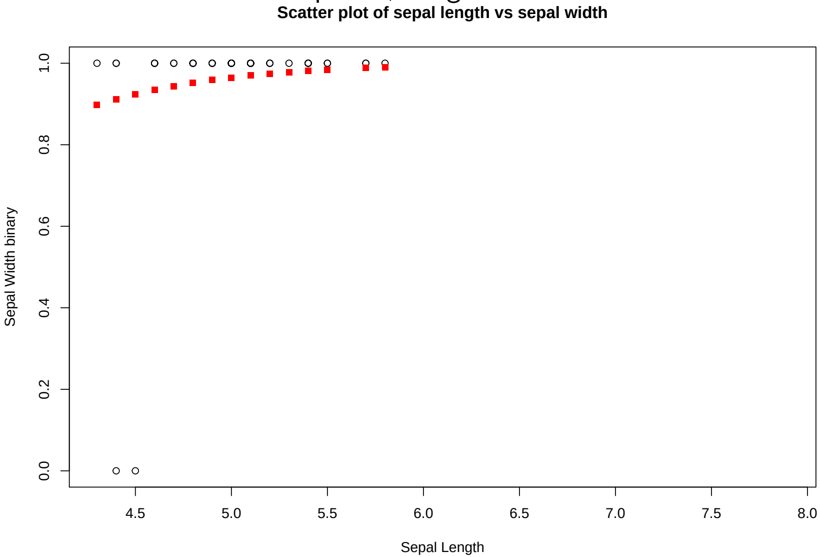
```
logit <- glm(Sepal.Width_binary ~ Species +Sepal.Length,
             data = data, family = "binomial")
summary(logit)
```

```
##
## Call:
## glm(formula = Sepal.Width_binary ~ Species + Sepal.Length, family = "binomial",
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2710  -0.7538   0.2472   0.7020   1.9477
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -4.7988     2.2981  -2.088 0.036784 *
## Speciesversicolor  -5.6936     0.9686  -5.878 4.16e-09 ***
## Speciesvirginica   -5.4812     1.0879  -5.039 4.69e-07 ***
## Sepal.Length        1.6219     0.4510   3.596 0.000323 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 199.22  on 149  degrees of freedom
## Residual deviance: 131.27  on 146  degrees of freedom
## AIC: 139.27
##
```

Logistic regression

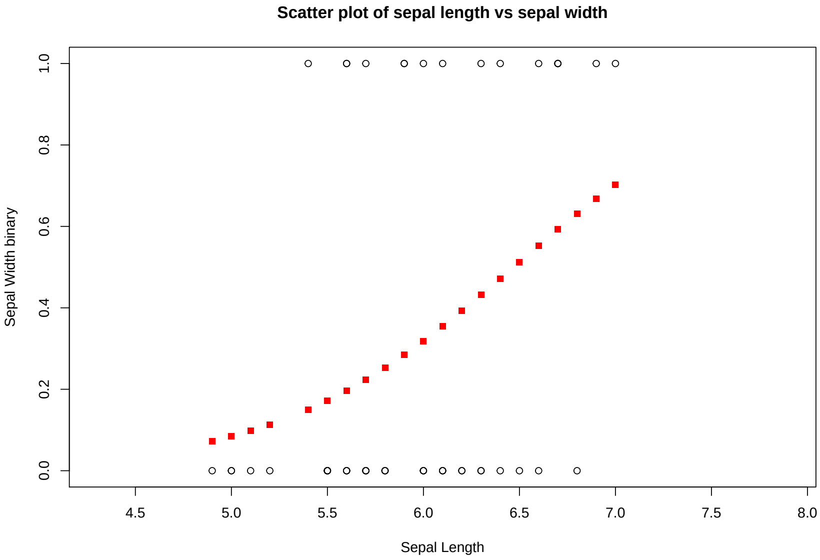
Logistic Regression with continuous variable, Sepal.Length

Plot the results for each species, we get that



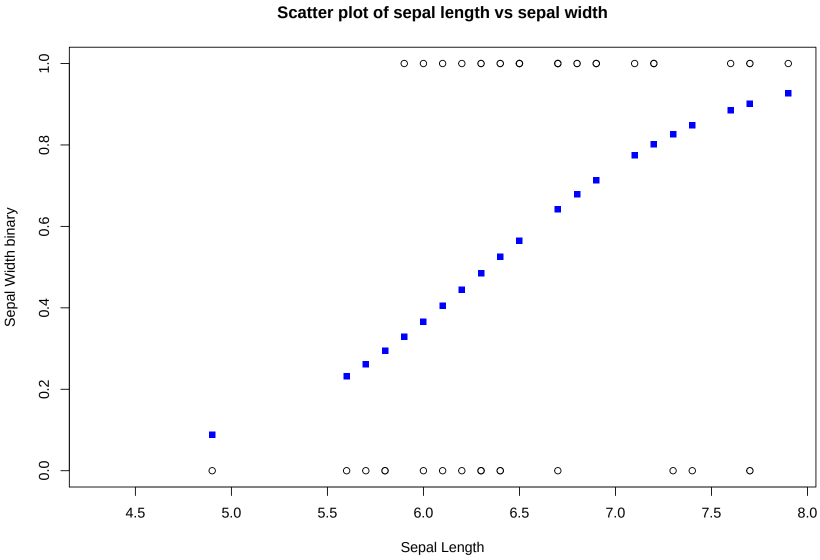
Logistic regression

Logistic Regression with continuous variable, Sepal.Length



Logistic regression

Logistic Regression with continuous variable, Sepal.Length



Logistic regression

Goodness of Fit

Deviance

For general linear models, we use *deviance* to the compare to two different models. Deviance is the difference in log likelihood of the models multipled by 2.

Logistic regression

Goodness of Fit

Saturated Model

Let's consider model in which each data point has its own mean and coefficients. This is called the saturated model. It basically replicates the data at hand.

Using deviance, we can compare our fitted model to a saturated model.

If the fitted model is behaves similiar to the saturated model, then the deviance can be well approximated by a chi-squared distribution with $m - n$ degrees of freedom. m is number of the data points and n is number of coefficients in our fitted model.

Logistic regression

Goodness of Fit

Saturated Model

This statistical property of the deviance allows us perform a hypothesis test

$$H_0: \text{ the fitted model is equivalent to the saturated model}$$

$$H_\alpha: \text{the fitted model is not equivalent to the saturated model}$$

Logistic regression

Goodness of Fit

Saturated Model

`logit$deviance` is the deviance between saturated model and fitted model.

`logit$df.residual` is equal to number of observations minus the number of coefficients in the fitted model.

Using this, we can calculate the p value for the hypothesis test above.

```r
p_value = pchisq(logit$deviance,
                 logit$df.residual, lower.tail = F)
print(p_value)
```

```
## [1] 0.8032738
```

Since the p value is greater than 0.05, we fail to reject the null hypothesis. (This is a good thing.)

Logistic regression

Goodness of Fit

Null Model

We can also use deviance to determine if our fitted model is better than the null model. The null model is a model with only a linear term.

Like above, we can design a hypothesis test comparing the null model to the fitted model.

Logistic regression

Goodness of Fit

Null Model

$$H_0 = \text{ the fitted model is equivalent to the null model}$$
$$H_\alpha = \text{ the fitted model is not equivalent to the null model}$$

In the limit of large data, it is known that the deviance follows a chi-squared distribution with parameter $n - 1$.

Logistic regression

Goodness of Fit

Null Model

`logit$deviance` is the deviance between saturated model and fitted model. `logit$df.residual` is equal to number of observations minus the number of coefficients in the fitted model.

`logit$null.deviance` is the deviance between saturated model and the null model. `logit$df.null` is the number of observations minus 1.

Logistic regression

Goodness of Fit

Null Model

Using this information, we can calculate the p value for the hypothesis test above.

```r
p_value = pchisq(logit$null.deviance-logit$deviance,
                 logit$df.null-logit$df.residual,
                 lower.tail = F)
print(p_value)
```

```
## [1] 1.173879e-14
```

Since the p value is less than one, we reject our null hypothesis. (This is a good thing.)

Logistic regression

Goodness of Fit

Anova

`anova` with argument `test="Chisq` allows us to compare change in deviance after sequencially adding terms our model.

```r
anova(logit,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Sepal.Width_binary
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                           149     199.22
## Species       2   51.709       147     147.51 5.910e-12 ***
## Sepal.Length  1   16.239       146     131.27 5.583e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Poisson Generalized Linear Model

A possion GLM is used to study *count* data (i.e. discrete numbers, $0, 1, 2, \cdots$). *Count* data describes the number of events that occur within a given time frame.

insert plot of poisson distribution here

Poisson Generalized Linear Model

A possion GLM is most useful when studying data in which the mean and variable are approximately equal. If they are not not equal, the standard error of the model terms must adjusted to account for the assumption violation.

Poisson Generalized Linear Model

Poisson Regression has the following components:

- response count variables, Y_i, that follows a Possion distribution with mean μ_i
- a systematic component, η_i, that relates the relates the explanatory variables, $\eta_i = \sum_{j=1}^{n} \beta_j x_{ij}$
- a link function, $log(\mu_i) = \sum_{j=1}^{n} \beta_j x_{ij}$

From Poisson regression, we learn the *mean* of each Y_i given the associated the explanatory variables.

Poisson Generalized Linear Model

Data

We will be consider the `bioChemists` data set in this section.

This data set contains number of articles produced by PhD biochemistry student during the last 3 years of their PhD.

```
attach(bioChemists)
summary(bioChemists)
```

```
##       art             fem           mar            kid5       
##  Min.   : 0.000   Men  :494   Single :309   Min.   :0.0000  
##  1st Qu.: 0.000   Women:421   Married:606   1st Qu.:0.0000  
##  Median : 1.000                             Median :0.0000  
##  Mean   : 1.693                             Mean   :0.4951  
##  3rd Qu.: 2.000                             3rd Qu.:1.0000  
##  Max.   :19.000                             Max.   :3.0000  
##       phd             ment       
##  Min.   :0.755   Min.   : 0.000  
##  1st Qu.:2.260   1st Qu.: 3.000  
##  Median :3.150   Median : 6.000  
##  Mean   :3.103   Mean   : 8.767  
##  3rd Qu.:3.920   3rd Qu.:12.000  
##  Max.   :4.620   Max.   :77.000
```

Poisson Generalized Linear Model

Data

The data set also contains demographic data associated with each student. data of the flower of certain plant species. The data set has five variables:

- *art* - number of articles produced by the student in the last 3 years of their PhD
- *fem* - gender
- *mar* - martial status
- *kid5* - number of children less than 5
- *phd* - pretige of PhD program
- *ment* - number of articles of the mentor in the last 3 years

Poisson Generalized Linear Model

Data

```
sapply(bioChemists, class)
```

```
##       art       fem       mar      kid5       phd      ment 
## "integer"  "factor"  "factor" "numeric" "numeric" "integer" 
```

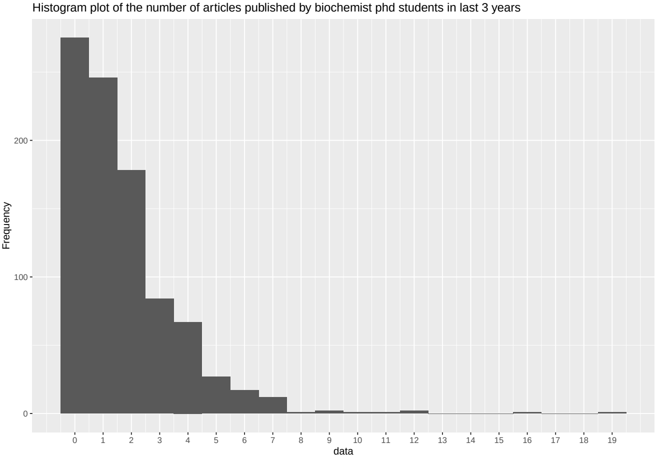
I convert bioChemists$kid5 from numeric to factor. This will be used later.

```
bioChemists$kid5 <- factor(bioChemists$kid5,
                            levels= unique(bioChemists$kid5),
                            labels= unique(bioChemists$kid5))
```

Poisson Generalized Linear Model

Data Visualization

Plotting the bar graph of `bioChemists$art`, we can see than the data looks Poisson-like since there is large number of observations at 0.



Poisson Generalized Linear Model

Data

We can “quantify” the Poission-ness by analyzing the mean and variance of the data.

```
mean(bioChemists$art)
```

```
## [1] 1.692896
```

```
var(bioChemists$art)
```

```
## [1] 3.709742
```

Although mean and variance are not equal, we will still fit it to Poisson distribution.

Poisson Generalized Linear Model

Possion Regression with constant term

To model only the constant term, I use the formula `art ~ 1`. This formula is equivalent to

$$\log \mu_i = \beta_0.$$

Poisson Generalized Linear Model

Possion Regression with constant term

```
poisson_model = glm(art ~ 1, family=poisson(link=log),data=bioChemists)
summary(poisson_model)
```

```
##
## Call:
## glm(formula = art ~ 1, family = poisson(link = log), data = bioChemists)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8401  -1.8401  -0.5770   0.2294   7.5677
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.52644    0.02541   20.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1817.4  on 914  degrees of freedom
## Residual deviance: 1817.4  on 914  degrees of freedom
## AIC: 3487.1
##
## Number of Fisher Scoring iterations: 5
```

Poisson Generalized Linear Model

Possion Regression with constant term

Note that the constant term is the log mean number of counts.

```r
print(coef(poisson_model))
```

```
## (Intercept)
##   0.5264408
```

```r
print(log(mean(bioChemists$art)))
```

```
## [1] 0.5264408
```

Goodness of fit

Saturated model

We can again compare the current model to the saturated model (best possible fit).

```
p_value = pchisq(poisson_model$deviance,
                 poisson_model$df.residual,
                 lower.tail = F)
print(p_value)
```

```
## [1] 3.304511e-62
```

Since our p value is less than 0.05, we reject the null hypothesis. The models are not equivalent.

Goodness of fit

Null model

We can also compare the current model to the null model (worst possible fit).

```
p_value = pchisq(poisson_model$null.deviance
                 -poisson_model$deviance,
                poisson_model$df.null
                -poisson_model$df.residual,
                lower.tail = F)
print(p_value)
```

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

We fail to reject the null hypothesis. This makes sense since the models are literally the same thing.

Possion Regression with martial status covariate

To model the martial status covariate, I use the formula `art ~ 1+mar`. This formula is equivalent to

$$\log \mu_i = \beta_0 + \beta_1 X_{1i}$$

where

$$X_{1i} = \begin{cases} 1 & \text{if mar} = \text{Married} \\ 0 & \text{otherwise} \end{cases} .$$

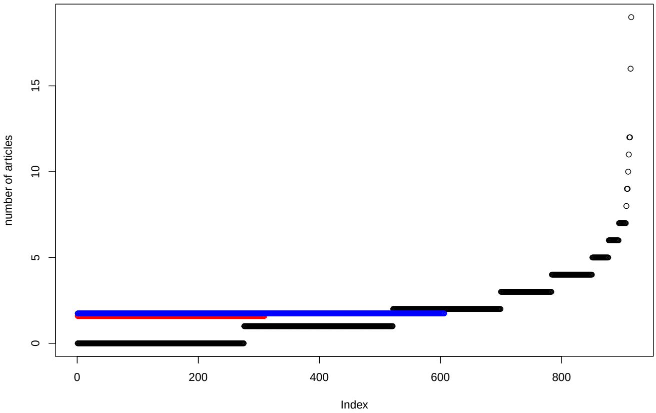
Possion Regression with martial status covariate

```
poisson_model = glm(art~1+mar , family=poisson(link=log),data=bioChemists)
summary(poisson_model)
```

```
##
## Call:
## glm(formula = art ~ 1 + mar, family = poisson(link = log), data = bioChemists)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8677  -1.7845  -0.5042   0.3107   7.4992
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.46514    0.04508  10.317   <2e-16 ***
## marMarried   0.09117    0.05458   1.671   0.0948 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1817.4  on 914  degrees of freedom
## Residual deviance: 1814.6  on 913  degrees of freedom
## AIC: 3486.3
##
## Number of Fisher Scoring iterations: 5
```

Possion Regression with martial status covariate

```
plot(bioChemists$art,ylab='number of articles',xlab = 'Index')
points(poisson_model$fitted[bioChemists$mar=='Single'],col="red")
points(poisson_model$fitted[bioChemists$mar=='Married'],col="blue")
```



Graphically, we can see than that martial status is not good indicator of number articles published.

Goodness of fit

Saturated model

We can again compare the current model to the saturated model (best possible fit).

```r
p_value = pchisq(poisson_model$deviance,
                 poisson_model$df.residual,
                 lower.tail = F)
print(p_value)
```

```
## [1] 4.731233e-62
```

Since our p value is less than 0.05, we reject the null hypothesis. The models are not equivalent and our model is a bad fit.

Goodness of fit

Null model

We can also compare the current model to the null model (worst possible fit).

```
p_value = pchisq(poisson_model$null.deviance-logit$deviance,
                 poisson_model$df.null-logit$df.residual, lower.tail = F)
print(p_value)
```

```
## [1] 1.016236e-70
```

Since our p value is less than 0.05, we reject the null hypothesis. The models are not equivalent. Though our current model does not capture much deviance, the current model captures much more variance than the null model.

Anova

```r
anova(poisson_model,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: art
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                   914     1817.4
## mar   1   2.8211       913     1814.6  0.09304 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Possion Regression with martial status and children covariate

To model the martial status and children as covariates, I use the formula `art ~ 1+mar + kid5`. This formula is equivalent to

$$\log \mu_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i}$$

where

$$X_{1i} = \begin{cases} 1 & \text{if the ith data point is married} \\ 0 & \text{otherwise} \end{cases},$$

$$X_{2i} = \begin{cases} 1 & \text{if the number of children of ith data point is 1} \\ 0 & \text{otherwise} \end{cases},$$

$$X_{3i} = \begin{cases} 1 & \text{if the number of children of ith data point is 2} \\ 0 & \text{otherwise} \end{cases}$$

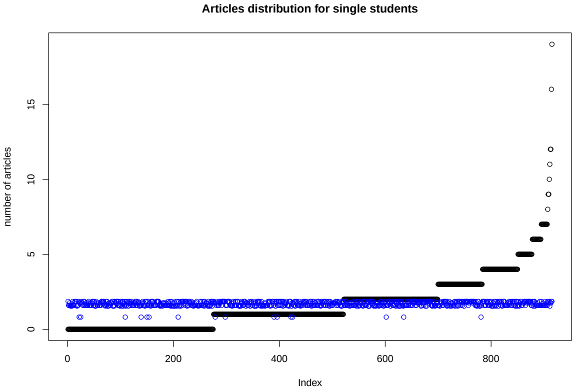
$$\text{and } X_{4i} = \begin{cases} 1 & \text{if the number of children of ith data point is 3} \\ 0 & \text{otherwise} \end{cases}.$$

Possion Regression with martial status and children covariate

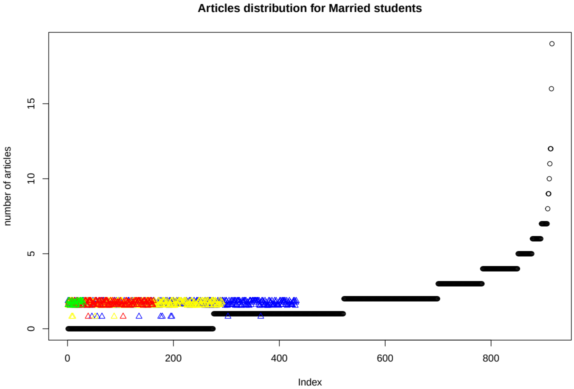
```
poisson_model = glm(art ~ 1 + kid5 + mar,
                    family=poisson(link=log),data=bioChemists)
summary(poisson_model)
```

```
##
## Call:
## glm(formula = art ~ 1 + kid5 + mar, family = poisson(link = log),
##     data = bioChemists)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9280  -1.7845  -0.5042   0.3518   7.3520
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.46514    0.04508  10.317   <2e-16 ***
## kid51       -0.05510    0.06907  -0.798   0.4250
## kid52       -0.18620    0.08960  -2.078   0.0377 *
## kid53       -0.82747    0.28067  -2.948   0.0032 **
## marMarried   0.15470    0.06235   2.481   0.0131 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1817.4  on 914  degrees of freedom
## Residual deviance: 1799.9  on 910  degrees of freedom
## AIC: 3477.7
##
## Number of Fisher Scoring iterations: 5
```

Possion Regression with martial status and children covariate



Possion Regression with martial status and children covariate



Graphically, we can see than that martial status and number of children is not good indicator of number articles published.

Goodness of Fit

Saturated model

We can again compare the current model to the saturated model (best possible fit).

```
p_value = pchisq(poisson_model$deviance,
                 poisson_model$df.residual,
                 lower.tail = F)
print(p_value)
```

```
## [1] 6.462874e-61
```

Since our p value is less than 0.05, we reject the null hypothesis. The models are not equivalent.

Goodness of fit

Null model

We can also compare the current model to the null model (worst possible fit).

```
p_value = pchisq(poisson_model$null.deviance
                 -poisson_model$deviance,
                poisson_model$df.null
                -poisson_model$df.residual,
                lower.tail = F)
print(p_value)
```

```
## [1] 0.001567133
```

Goodness of fit

Anova

We can also determine the model terms that cause a significance reduction in deviance.

```
anova(poisson_model,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: art
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                   914     1817.4
## kid5  3  11.3045       911     1806.1  0.01019 *
## mar   1   6.1638       910     1799.9  0.01304 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Possion Regression with continuous variables, mentor articles and martial status

To model the martial status and number of mentor articles as covariates, I use the formula `art ~ 1+mar + ment`. This formula is equivalent to

$$\log \mu_i = \beta_0 + \beta_1 X_{1i} + X_{2i}$$

where

$$X_{1i} = \begin{cases} 1 & \text{if the i data point is Married} \\ 0 & \text{otherwise} \end{cases}$$

and X_{2i} is the number of publications of the ith data point's mentor.

Possion Regression with continuous variables, mentor articles and martial status

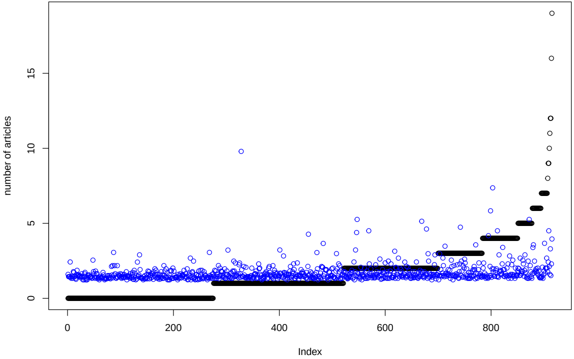
```r
poisson_model = glm(art ~ 1 + ment +mar,
                    family=poisson(link=log),data=bioChemists)
summary(poisson_model)
```

```
##
## Call:
## glm(formula = art ~ 1 + ment + mar, family = poisson(link = log),
##     data = bioChemists)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6086  -1.6317  -0.3608   0.5039   5.8942
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.210726   0.049847   4.227 2.36e-05 ***
## ment        0.025917   0.001915  13.530  < 2e-16 ***
## marMarried  0.075332   0.054643   1.379    0.168
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1817.4  on 914  degrees of freedom
## Residual deviance: 1667.6  on 912  degrees of freedom
## AIC: 3341.4
##
## Number of Fisher Scoring iterations: 5
```

Possion Regression with continuous variables, mentor articles and martial status

```
plot(bioChemists$art,ylab='number of articles',xlab = 'Index')

points(poisson_model$fitted,col="blue",pch=1)
```



Graphically, we can see than that martial status and number of children is not good indicator of number articles published.

Goodness of Fit

Saturated model

We can again compare the current model to the saturated model (best possible fit).

```r
p_value = pchisq(poisson_model$deviance,
                 poisson_model$df.residual,
                 lower.tail = F)
print(p_value)
```

```
## [1] 6.132629e-47
```

Since our p value is less than 0.05, we reject the null hypothesis. The models are not equivalent.

Goodness of fit

Null model

We can also compare the current model to the null model (worst possible fit).

```
p_value = pchisq(poisson_model$null.deviance
                 -poisson_model$deviance,
                poisson_model$df.null
                -poisson_model$df.residual,
                lower.tail = F)
print(p_value)
```

```
## [1] 2.993003e-33
```

Since our p value is less than 0.05, we reject the null hypothesis. The models are not equivalent.

Goodness of fit

Anova

We can also determine the model terms that cause a significance reduction in deviance.

```r
anova(poisson_model,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: art
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                   914     1817.4
## ment  1  147.860       913     1669.5   <2e-16 ***
## mar   1    1.918       912     1667.6   0.1661
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Log-Linear Regression

Log-linear models allow us to model asscociation between between two or more variables in contingency table. In a log-linear model, there are no well defined explanatory/response variables. This is because we are focused more on the *interaction* between two variables.

Log-Linear Regression

Contingency Table

Contingency table displays number of observations for a given combination of factors.

This definition is best represented by an example.

Log-Linear Regression

Contingency Table

One-Way Contingency Table

A one-way contingency table shows the counts according to one covariate.

```
table(art_relative=bioChemists$art_binary)
```

```
## art_relative
##   0   1
## 521 394
```

This one-way contigency table shows that:

- there are 521 biochemists with 1 or less papers
- there are 394 biochemists with greater than 1 papers.

Log-Linear Regression

Contingency Table

Two-Way Contingency Table

A two-way contingency table shows the counts according to two covariates.

```
table(art_relative=bioChemists$art_binary,ment=bioChemists$ment_binary)
```

```
##             ment
## art_relative   0   1
##            0 321 200
##            1 171 223
```

This two-way contigency table shows that:

- there are 321 biochemists with 1 or less papers and with a mentor that produced less than or equal to 6 papers
- there are 200 biochemists with 1 or less papers and with a mentor that produced more than 6 papers

Log-Linear Regression

Contingency Table

Two-Way Contingency Table

A two-way contingency table shows the counts according to two covariates.

```
table(art_relative=bioChemists$art_binary,ment=bioChemists$ment_binary)
```

```
##             ment
## art_relative   0   1
##            0 321 200
##            1 171 223
```

- there are 171 biochemists with more than 1 paper and with a mentor that produced less than or equal to 6 papers
- there are 200 biochemists with more than 1 paper and with a mentor that produced more than 6 papers

Log-Linear Regression

Contingency Table

Three-Way Contingency Table

A three-way contingency table shows the counts according to three covariates.

```r
table(art_relative=bioChemists$art_binary,ment=bioChemists$ment_binary,
      kid5=bioChemists$kid5)
```

```
## , , kid5 = 0
##
##             ment
## art_relative   0   1
##            0 208 128
##            1 116 147
##
## , , kid5 = 1
##
##             ment
## art_relative   0   1
##            0  66  46
##            1  38  45
##
## , , kid5 = 2
##
##             ment
## art_relative   0   1
##            0  39  21
```

Log-Linear Regression

Contingency Table

Three-Way Contingency Table

This three-way contigency table shows that:

- With no children,
  - there are 208 biochemists with 1 or less papers and with a mentor that produced less than or equal to 6 papers
  - there are 128 biochemists with 1 or less papers and with a mentor that produced more than 6 papers
  - there are 116 biochemists with more than 1 paper and with a mentor that produced less than or equal to 6 papers
  - there are 147 biochemists with more than 1 paper and with a mentor that produced more than 6 papers

Log-linear Regression for two way contingency table

For a two-way contingency table, log-linear GLMs have the following components:

- count response variables, Y_{ij}, which is the number of entries in the (i,j)th cell of the table. Y_{ij} follows a Possion distribution with mean μ_{ij}.
- a systematic component, η_i, that relates the relates the explanatory variables,

$$\eta_{ij} = \sum_{j=1}^{n} \beta_k X_{ijk}$$

- a link function that relates the mean of the random to the systematic component

$$\log \mu_{ij} = \sum_{k=1}^{n} \beta_k X_{ijk}$$

Independent Model for two-way contigency table

We use log-linear model to model the group mean count of each cell of the contingency table. Remember, using a log-linear model, our primary goal is to learn the interaction effects between covariates.

Again, we build the same two-way contingency table. We need to convert the contigency table in a form that is acceptable to `glm`.

```
contigency_table = table(art_relative=bioChemists$art_binar
                         ment=bioChemists$ment_binary)
contigency_table.df = as.data.frame(contigency_table)
```

Independent Model for two-way contigency table

```
print(contigency_table.df)
```

```
##   art_relative ment Freq
## 1            0    0  321
## 2            1    0  171
## 3            0    1  200
## 4            1    1  223
```

Independent Model for two-way contigency table

Assuming each number of articles and mentor do not affect each other, we build a model of the cell count that does not take into account interaction effects. Such a model is called the *independent* model.

To do this, we use formula `Freq ~ art_relative + ment`.

```
log_linear_model_int <- glm(Freq ~ art_relative + ment,
            data = contigency_table.df, family = poisson)
```

Independent Model for two-way contigency table

```
summary(log_linear_model_int)
```

```
##
## Call:
## glm(formula = Freq ~ art_relative + ment, family = poisson, data = contigency_table.df)
##
## Deviance Residuals:
##      1       2       3       4
##  2.385  -2.905  -2.713   2.923
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    5.63530    0.05347 105.392  < 2e-16 ***
## art_relative1 -0.27940    0.06676  -4.185 2.85e-05 ***
## ment1         -0.15111    0.06631  -2.279   0.0227 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 52.927  on 3  degrees of freedom
## Residual deviance: 30.035  on 1  degrees of freedom
## AIC: 65.008
##
## Number of Fisher Scoring iterations: 4
```

Independent Model for two-way contigency table

Goodness of fit

We compare the current model to the saturated model (best possible fit).

```
p_value = pchisq(log_linear_model_int$deviance,
                 log_linear_model_int$df.residual,
                 lower.tail = F)
print(p_value)
```

```
## [1] 4.243721e-08
```

Since our p value is less than 0.05, we reject the null hypothesis. The models are not equivalent.

Independent Model for two-way contigency table

Saturated Model for the two-way contingency table

Assuming each number of articles and mentor affect each other, we build a model of the cell count that takes into account all interaction effects. Such a model is called the *saturated* model. To do this, we use formula `Freq ~ art_relative*ment`.

```r
log_linear_model_sat <- glm(Freq ~ art_relative*ment,
            data = contigency_table.df,
            family = poisson)
```

Independent Model for two-way contigency table

Saturated Model for the two-way contingency table

```
summary(log_linear_model_sat)
```

```
##
## Call:
## glm(formula = Freq ~ art_relative * ment, family = poisson, data = contigency_table.df)
##
## Deviance Residuals:
## [1]  0  0  0  0
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          5.77144    0.05581 103.404  < 2e-16 ***
## art_relative1       -0.62978    0.09467  -6.652 2.89e-11 ***
## ment1               -0.47312    0.09008  -5.252 1.50e-07 ***
## art_relative1:ment1  0.73863    0.13582   5.438 5.38e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 5.2927e+01  on 3  degrees of freedom
## Residual deviance: 1.8874e-14  on 0  degrees of freedom
## AIC: 36.973
##
## Number of Fisher Scoring iterations: 2
```

Independent Model for two-way contigency table

Goodness of fit

We compare the current model to the saturated model (best possible fit).

```
p_value = pchisq(0,
                 log_linear_model_sat$df.residual,
                 lower.tail = F)
print(p_value)
```

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

We fail to reject the null hypothesis. This makes sense since the models are literally the same thing.

Independent Model for two-way contigency table

Model Comparison

We use `anova` with `test='Chisq'` to compare the independent and saturated model.

```r
anova(log_linear_model_int,log_linear_model_sat,
      test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: Freq ~ art_relative + ment
## Model 2: Freq ~ art_relative * ment
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1         1     30.035
## 2         0      0.000  1   30.035 4.244e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From `anova`, we can see that the saturated model provides a statisically significant result.

Independent Model for two-way contigency table

Anova

We use also `anova` to determine what caused the significant decrease in the deviance.

```r
anova(log_linear_model_sat,test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: Freq
##
## Terms added sequentially (first to last)
##
##
##                  Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                 3     52.927
## art_relative      1  17.6844         2     35.243 2.608e-05 ***
## ment              1   5.2082         1     30.035   0.02248 *
## art_relative:ment  1  30.0348         0      0.000 4.244e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adding `art_relative:ment` to the independent model caused significant decrease in deviance.

Independent Model for the three-way contingency table

Again, we build the same three-way contingency table. We need to convert the contigency table in a form that is acceptable to `glm`.

To create the *independent* model for the three-way contingency table, we use formula `Freq ~  art_relative + ment + kid5`.

```r
log_linear_model_int <- glm(Freq ~ art_relative + ment + kid5,
            data = contigency_table.df, family = poisson)
```

Independent Model for the three-way contingency table

```
summary(log_linear_model_int)
```

```
##
## Call:
## glm(formula = Freq ~ art_relative + ment + kid5, family = poisson,
##     data = contigency_table.df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4439  -1.4070  -0.1702   1.1974   2.4521
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    5.21164    0.05861  88.915  < 2e-16 ***
## art_relative1 -0.27940    0.06676  -4.185 2.85e-05 ***
## ment1         -0.15111    0.06631  -2.279   0.0227 *
## kid51         -1.12226    0.08245 -13.612  < 2e-16 ***
## kid52         -1.74130    0.10580 -16.459  < 2e-16 ***
## kid53         -3.62267    0.25331 -14.301  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 901.879  on 15  degrees of freedom
## Residual deviance:  36.651  on 10  degrees of freedom
## AIC: 131.03
##
## Number of Fisher Scoring iterations: 4
```

Goodness of fit

We compare the current model to the saturated model (best possible fit).

```
p_value = pchisq(log_linear_model_int$deviance,
                 log_linear_model_int$df.residual, lower.tail = F)
print(p_value)
```

```
## [1] 6.50121e-05
```

Since our p value is less than 0.05, we reject the null hypothesis. The models are not equivalent.

Saturated Model

To create the *saturated* model for the three-way contingency table, we use formula `Freq ~  art_relative*ment*kid5`.

```
log_linear_model_sat <- glm(Freq ~ art_relative*ment*kid5,
            data = contigency_table.df, family = poisson)
```

Saturated Model

```r
summary(log_linear_model_sat)
```

```
##
## Call:
## glm(formula = Freq ~ art_relative * ment * kid5, family = poisson,
##     data = contigency_table.df)
##
## Deviance Residuals:
##  [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)               5.33754    0.06934  76.979  < 2e-16 ***
## art_relative1            -0.58395    0.11588  -5.039 4.67e-07 ***
## ment1                    -0.48551    0.11234  -4.322 1.55e-05 ***
## kid51                    -1.14788    0.14128  -8.125 4.47e-16 ***
## kid52                    -1.67398    0.17450  -9.593  < 2e-16 ***
## kid53                    -3.25810    0.36029  -9.043  < 2e-16 ***
## art_relative1:ment1       0.72235    0.16746   4.314 1.61e-05 ***
## art_relative1:kid51       0.03188    0.23430   0.136    0.892
## art_relative1:kid52      -0.30703    0.31870  -0.963    0.335
## art_relative1:kid53      -1.49549    1.06697  -1.402    0.161
## ment1:kid51               0.12449    0.22251   0.559    0.576
## ment1:kid52              -0.13353    0.29305  -0.456    0.649
## ment1:kid53               0.01550    0.58105   0.027    0.979
## art_relative1:ment1:kid51 -0.19226    0.33686  -0.571    0.568
## art_relative1:ment1:kid52  0.49140    0.44529   1.104    0.270
## art_relative1:ment1:kid53  0.44080    1.36126   0.324    0.746
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
```

Goodness of fit

We compare the current model to the saturated model (best possible fit).

```r
p_value = pchisq(log_linear_model_sat$deviance,
                 log_linear_model_sat$df.residual, lower.tail = F)
print(p_value)
```

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

We fail to reject the null hypothesis. This makes sense since the models are literally the same thing.

Model Comparison

We use `anova` with `test='Chisq'` to compare the independent and saturated model.

```
anova(log_linear_model_int,log_linear_model_sat,test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: Freq ~ art_relative + ment + kid5
## Model 2: Freq ~ art_relative * ment * kid5
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        10     36.651
## 2         0      0.000 10   36.651 6.501e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From `anova`, we can see that the saturated model provides a statisically significant result.

Model Comparison

We use also `anova` to determine what caused the significant decrease in the deviance.

```
anova(log_linear_model_sat,test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: Freq
##
## Terms added sequentially (first to last)
##
##
##                         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                       15     901.88
## art_relative             1    17.68        14     884.20 2.608e-05 ***
## ment                     1     5.21        13     878.99   0.02248 *
## kid5                     3   842.34        10      36.65 < 2.2e-16 ***
## art_relative:ment        1    30.03         9       6.62 4.244e-08 ***
## art_relative:kid5        3     4.45         6       2.17   0.21665
## ment:kid5                3     0.19         3       1.97   0.97873
## art_relative:ment:kid5   3     1.97         0       0.00   0.57819
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adding `art_relative:ment` to the independent model caused significant decrease in deviance