SS9155 - Assignment 3 - 250620601

Ravin Lathigra

2019-01-28

R Packages & Libraries

```
library(corrplot)    #Visualize Correlation between variables
library(kableExtra)  #Style tables
library(tidyverse)   #contains ggplot2,dplyr,tidyr, readr,purr,tibble,stringr,forcats
library(formatR)     #Improve readability of code
library(e1071)       #Functions for latent class analysis, Fourier transform ect.
library(VIM)         #Knn
library(ggfortify)   #Add on to ggplot2 to allow for more plot types
library(Rtsne)       #Dimension reduction classification
library(caret)       #streamlined model development
library(RColorBrewer)#Control colours of visualizations
library(GGally)      #Contains ggpairs plots
library(lmtest)      #Test for linear assumptions
library(MASS)
library(faraway)
```

```
data("esoph")

c3_q1 <- esoph

str(c3_q1)
```

```
## 'data.frame':    88 obs. of  5 variables:
##  $ agegp    : Ord.factor w/ 6 levels "25-34"<"35-44"<..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ alcgp    : Ord.factor w/ 4 levels "0-39g/day"<"40-79"<..: 1 1 1 1 2 2 2 2 3 3 ...
##  $ tobgp    : Ord.factor w/ 4 levels "0-9g/day"<"10-19"<..: 1 2 3 4 1 2 3 4 1 2 ...
##  $ ncases   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ncontrols: num  40 10 6 5 27 7 4 7 2 1 ...
```

```
summary(c3_q1)
```

```
##    agegp          alcgp          tobgp        ncases         ncontrols
##  25-34:15   0-39g/day:23   0-9g/day:24   Min.   : 0.000   Min.   : 1.00
##  35-44:15   40-79    :23   10-19   :24   1st Qu.: 0.000   1st Qu.: 3.00
##  45-54:16   80-119   :21   20-29   :20   Median : 1.000   Median : 6.00
##  55-64:16   120+     :21   30+     :20   Mean   : 2.273   Mean   :11.08
##  65-74:15                                3rd Qu.: 4.000   3rd Qu.:14.00
##  75+  :11                                Max.   :17.000   Max.   :60.00
```

Chapter 3 | Question 1

Using data from the case-control study of oesophageal cancer in ille-et-Vilaine, France we will explore the relationships between the available predictors and the presence of oesophageal cancer. The data is sourced from the `datasets` package more specifically the `esoph` dataset.

Age Group [agegp] Alcohol Consumption [alcgp] Tobacco Consumption [tobgp] Number of Cases [ncases] Number of controls [ncontrols] (DELETE BEFORE KNITTING)

(a) Plot the proportion of cases against each predictor using the size of the point to indicate the number of subject as seen in Figure 2.7. Comment on the rela- tionships seen in the plots.

How does age relate to the presence of cancer?

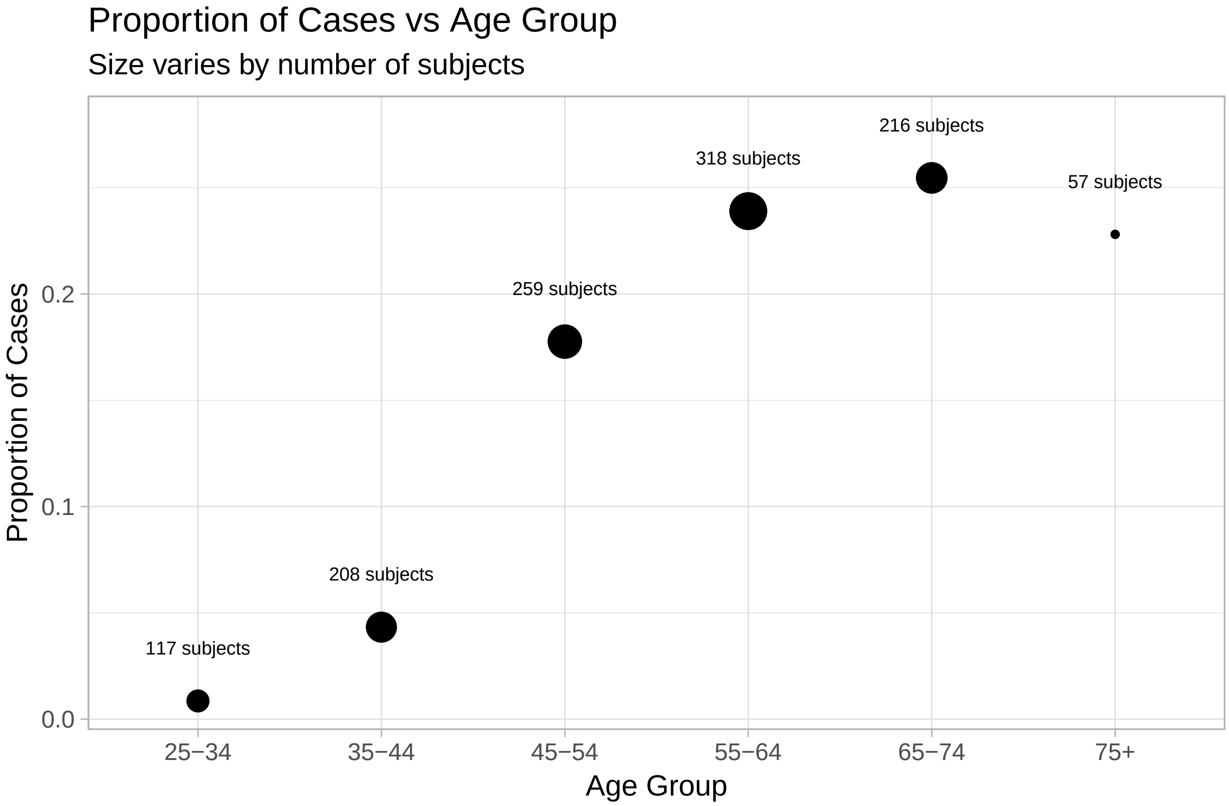


Figure 1.0 | Source: Esoph Dataset Breslow, N. E. and Day, N. E. (1980) Statistical Methods in Cancer Research.

`Figure 1.0` plots the proportion of cases of cancer to total subjects within each age group. The size of the points vary by the number of observations in a given age grouping. The plot demonstrates that as age increases, we see an increase in proportion of subjects with cancer. Additionally, it indicates that there are very few cases of cancer at younger ages otherwise known as sparse data.

How does alcohol consumption relate to the presence of cancer?

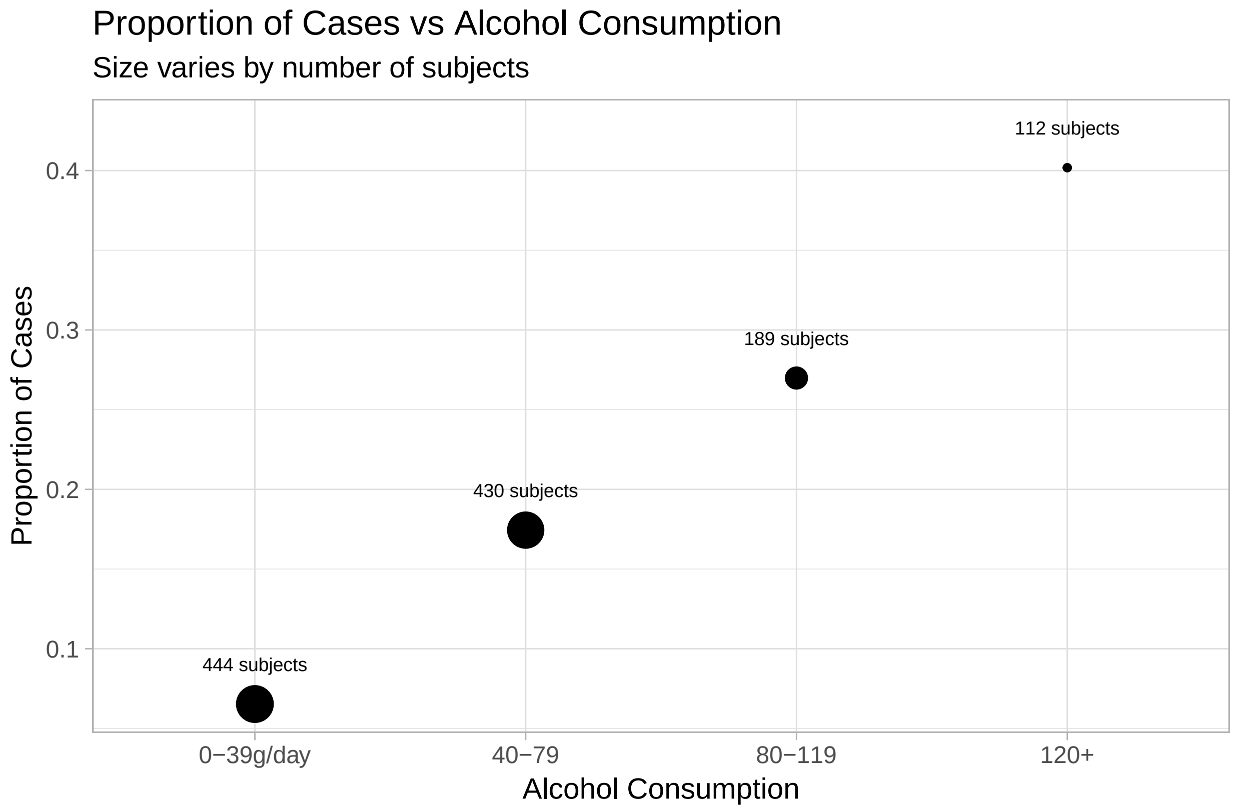


Figure 2.0 | Source: Esoph Dataset Breslow, N. E. and Day, N. E. (1980) Statistical Methods in Cancer Research.

`Figure 2.0` plots the proportion of cases of cancer to total subjects considerinng alcohol consumotion. The size of the points vary by the number of observations in a given consumption grouping. The plot demonstrates that as consumption increases, we see an increase in proportion of subjects with cancer. Additionally, the largest groupings (i.e. 0-39 g/day and 40-79 g/day) have the lowest proportion of observations that have concer.

How does tobacco consumption relate to the presence of cancer?

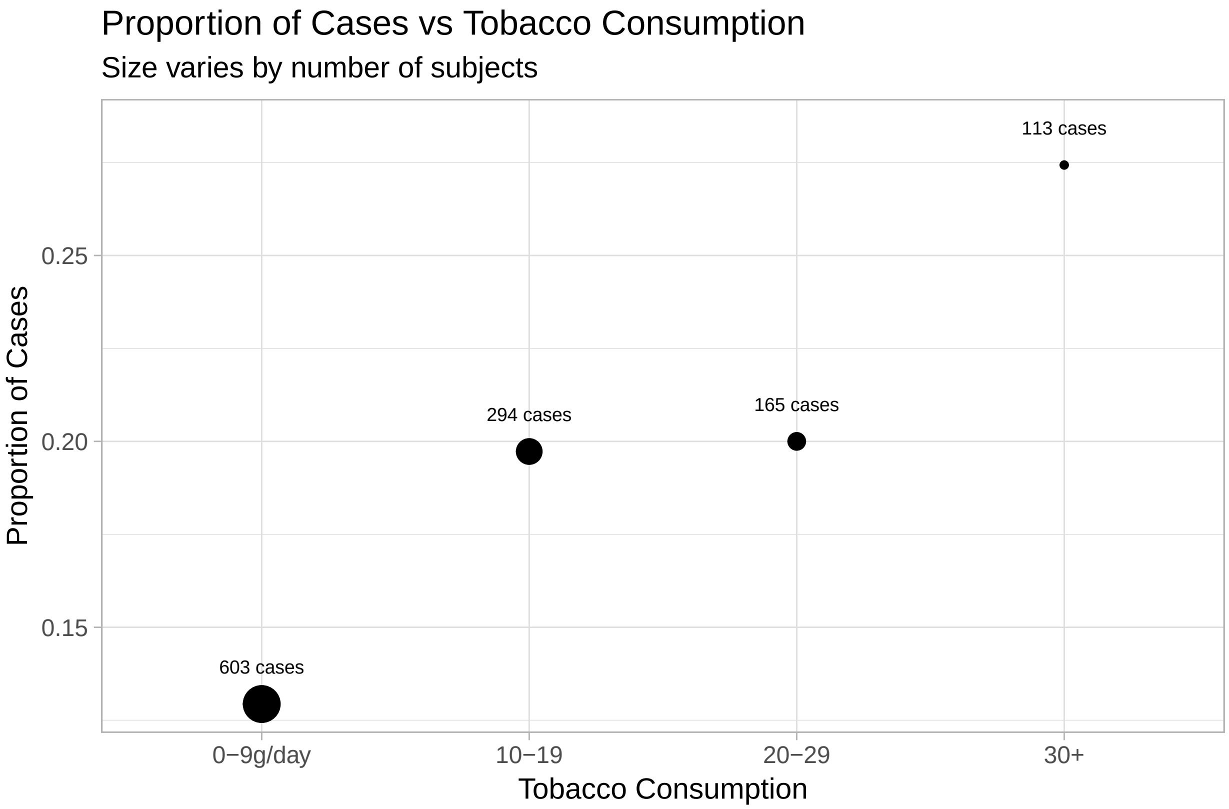


Figure 3.0 | Source: Esoph Dataset Breslow, N. E. and Day, N. E. (1980) Statistical Methods in Cancer Research.

`Figure 3.0` plots the proportion of cases of cancer to total subjects considerinng tobacco consumotion. The size of the points vary by the number of observations in a given consumption grouping. The plot demonstrates that as consumption increases, we see an increase in proportion of subjects with cancer, though there is a plateau between 10 and 29 g/day. Additionally, the largest groupings (i.e. 0-9 g/day) has the lowest proportion of observations that have concer.

(b) Fit a binomial GLM with interactions between all three predictors. Use AIC as a criterion to select a model using the step function. Which model is selected?

The following output provides a summary for the final model selected using AIC criterion. The available features included age group, alcohol consumption, tobacco consumption and interactions between these terms.

```
##
## Call:
## glm(formula = cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp,
##     family = "binomial", data = c3_q1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6891  -0.5618  -0.2168   0.2314   2.0642
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.77997    0.19796  -8.992  < 2e-16 ***
## agegp.L      3.00534    0.65215   4.608 4.06e-06 ***
## agegp.Q     -1.33787    0.59111  -2.263  0.02362 *
## agegp.C      0.15307    0.44854   0.341  0.73291
## agegp^4      0.06410    0.30881   0.208  0.83556
## agegp^5     -0.19363    0.19537  -0.991  0.32164
## alcgp.L      1.49185    0.19935   7.484 7.23e-14 ***
## alcgp.Q     -0.22663    0.17952  -1.262  0.20680
## alcgp.C      0.25463    0.15906   1.601  0.10942
## tobgp.L      0.59448    0.19422   3.061  0.00221 **
## tobgp.Q      0.06537    0.18811   0.347  0.72823
## tobgp.C      0.15679    0.18658   0.840  0.40071
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 227.241  on 87  degrees of freedom
## Residual deviance:  53.973  on 76  degrees of freedom
## AIC: 225.45
##
## Number of Fisher Scoring iterations: 6
```

The final model used only the main effects without interactions. More simply, the formula can be written as:

$$AICSelectedModel : (Ncases, Ncontrols) = Agegp + Alcgp + Tobgp$$

(c) All three factors are ordered and so special contrasts have been used appropriate for ordered factors involving linear, quadratic and cubic terms. Further simplification of the model may be possible by eliminating some of these terms. Use the unclass function to convert the factors to a numerical representation and check whether the model may be simplified.

To remove the effect of ordinal factors we can unclass out features. This provides a binomial model linear in its predictors. The table `Goodness of Fit - Unclass` shows the goodness of fit for this simplified model. Notice the Deviance is near the degrees of freedom which indicates that there is a sufficient fit. Furthermore inspecting the p-value (0.775) we gather that the fit is sufficient thus the model can be simplified.

```
##
## Call:
## glm(formula = cbind(ncases, ncontrols) ~ unclass(agegp) + unclass(tobgp) +
##     unclass(alcgp), family = "binomial", data = c3_q1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9586  -0.8555  -0.4358   0.3075   1.9349
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -5.59594    0.41540 -13.471  < 2e-16 ***
## unclass(agegp)  0.52867    0.07188   7.355 1.91e-13 ***
## unclass(tobgp)  0.27446    0.08074   3.399 0.000676 ***
## unclass(alcgp)  0.69382    0.08342   8.317  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 227.241  on 87  degrees of freedom
## Residual deviance:  73.959  on 84  degrees of freedom
## AIC: 229.44
##
## Number of Fisher Scoring iterations: 4
```

Table 1: Goodness of Fit - Unclass

Degrees.Of.Freedom	Deviance	P.value
84	73.959	0.775

(d) Use the summary output of the factor model to suggest a model that is slightly more complex than the linear model proposed in the previous question

The summary output for the factor model suggests that the only feature that benefited from higher degree representation was age group. Since the previous model showed that the model could be simplified it would be appropriate to develop a 3rd model that keeps the ordinal relationships for age groups but simplifies the other features. The following formula represents the proposed model:

$$ProposedModel :(Ncases, Ncontrols) = Agegp[Ordinal] + unclass(Alcgp) + unclass(Tobgp)$$

(e) Does your final model fit the data? Is the test you make accurate for this data?

To test if the final model fits the data we can inspect the deviance of the model. If the model is appropriate the deviance should follow a chi-squared distribution with n-p-1 degrees of freedom. We can perform a chi-squared test on the model to see if the deviance follows this distribution.

The table `Goodness of Fit - Proposed` shows the goodness of fit for the proposed model. Notice the Deviance is near the degrees of freedom which indicates that there is a sufficient fit. Furthermore inspecting the p-value (0.96) we gather that the fit is sufficient thus the model can be simplifiof the proposed model is sufficient. The test for fit is appropriate since there are no indications of overdispersion requiring estimation of dispersion parameter and an F-test to be performed.

```
##
## Call:
## glm(formula = cbind(ncases, ncontrols) ~ agegp + unclass(tobgp) +
##     unclass(alcgp), family = binomial(), data = c3_q1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7628  -0.6426  -0.2709   0.3043   2.0421
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -4.01097    0.31224 -12.846  < 2e-16 ***
## agegp.L         2.96113    0.65092   4.549 5.39e-06 ***
## agegp.Q        -1.33735    0.58918  -2.270  0.02322 *
## agegp.C         0.15292    0.44792   0.341  0.73281
## agegp^4         0.06668    0.30776   0.217  0.82848
## agegp^5        -0.20288    0.19523  -1.039  0.29872
## unclass(tobgp)  0.26162    0.08198   3.191  0.00142 **
## unclass(alcgp)  0.65308    0.08452   7.727 1.10e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

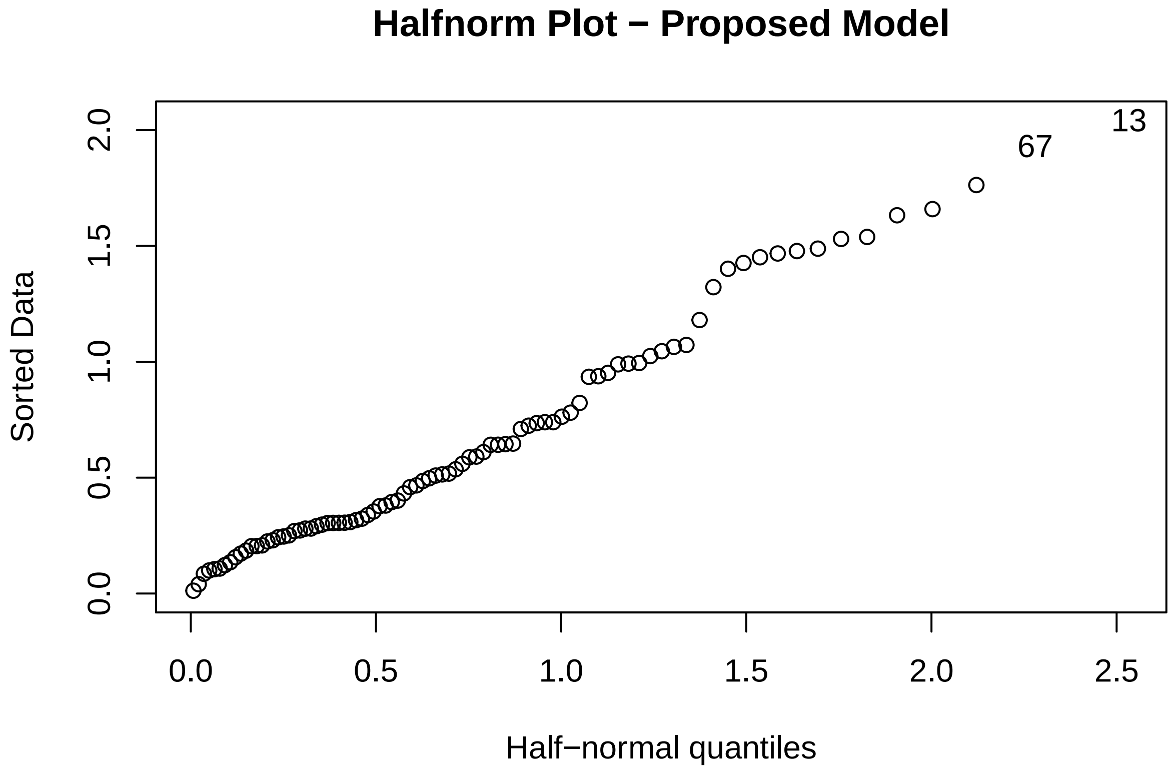
```
##
##     Null deviance: 227.241  on 87  degrees of freedom
## Residual deviance:  59.277  on 80  degrees of freedom
## AIC: 222.76
##
## Number of Fisher Scoring iterations: 6
```

Table 2: Goodness of Fit - Proposed

Degrees.Of.Freedom	Deviance	P.value
80	59.277	0.96

(f) Check for outliers in your final model

Inspecting a half norm plot we can visually identify outliers. The following plot suggests that there are 2 "outliers" though they are not extreme enough for them to be considered true outliers or perhaps better stated, influential observations.



(g) What is the predicted effect of moving one category higher in alcohol consumption?

Since we use the log-link function i.e.

$$log(odds) = \beta_0 + \sum \beta_i x_i$$

then;

$$Odds = e^{\beta_0 + \sum \beta_i x_i}$$

$$Odds = e^{\beta_0}....e^{\beta_n x_i}$$

To isolate the predicted effect of moving one class up in alcohol consumption we need to exponentiate the correseponding coefficient.

```
## [1] "The predicted effect of moving up 1 group in alcohol consumption is: 1.92"
```

(h) Compute a 95% confidence interval for this predicted effect.

Table `Predicted Effect` displays the 95% confidence interval for a 1 class increase in alcohol consumption.

```
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
## Waiting for profiling to be done...
```

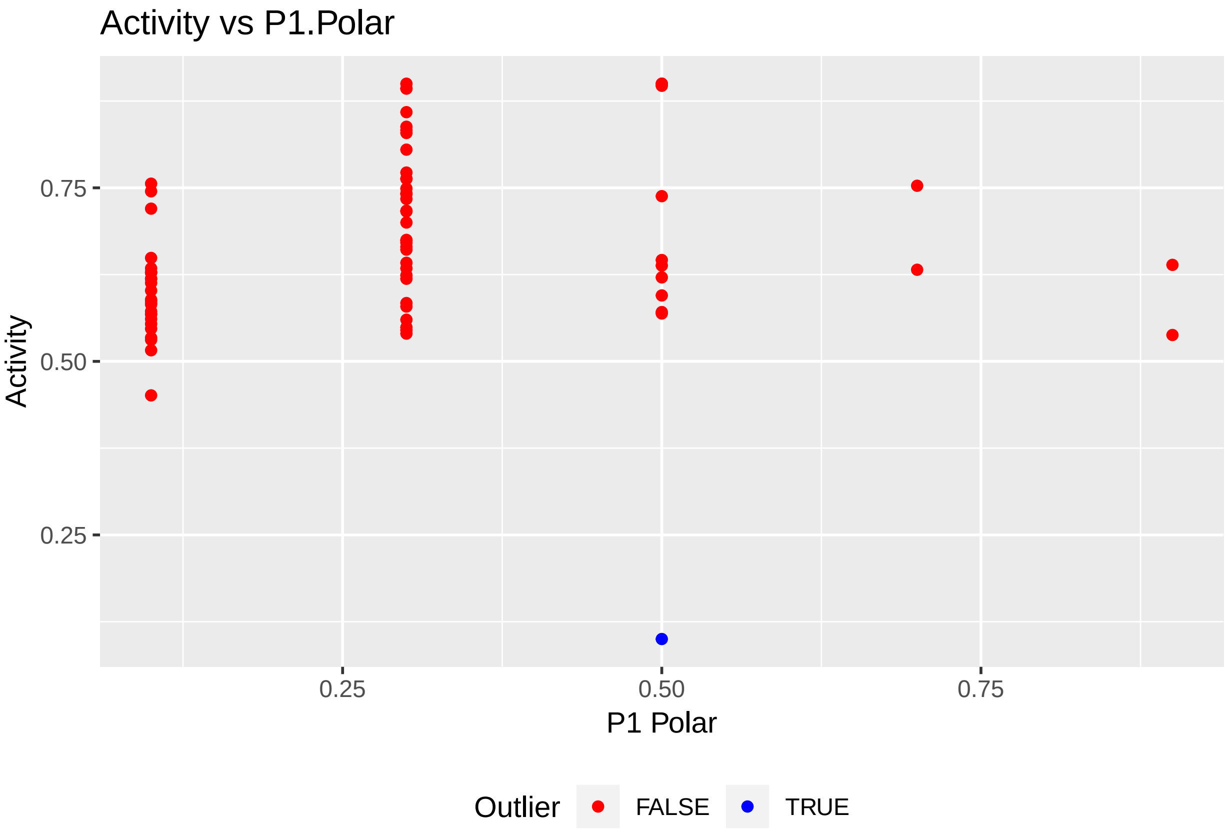
Table 3: Predicted Effect

<table>
<tr><th>Feature</th><th>Coefficient</th><th>Lower Bound</th><th>Upper Bound</th><th><b>Predicted Effect|Lower Bound</b></th><th><b>Predicted Effect|Up</b></th></tr>
<tr><td>unclass(alcgp)</td><td>0.65</td><td>0.49</td><td>0.82</td><td><b>1.63</b></td><td><b>2.27</b></td></tr>
</table>

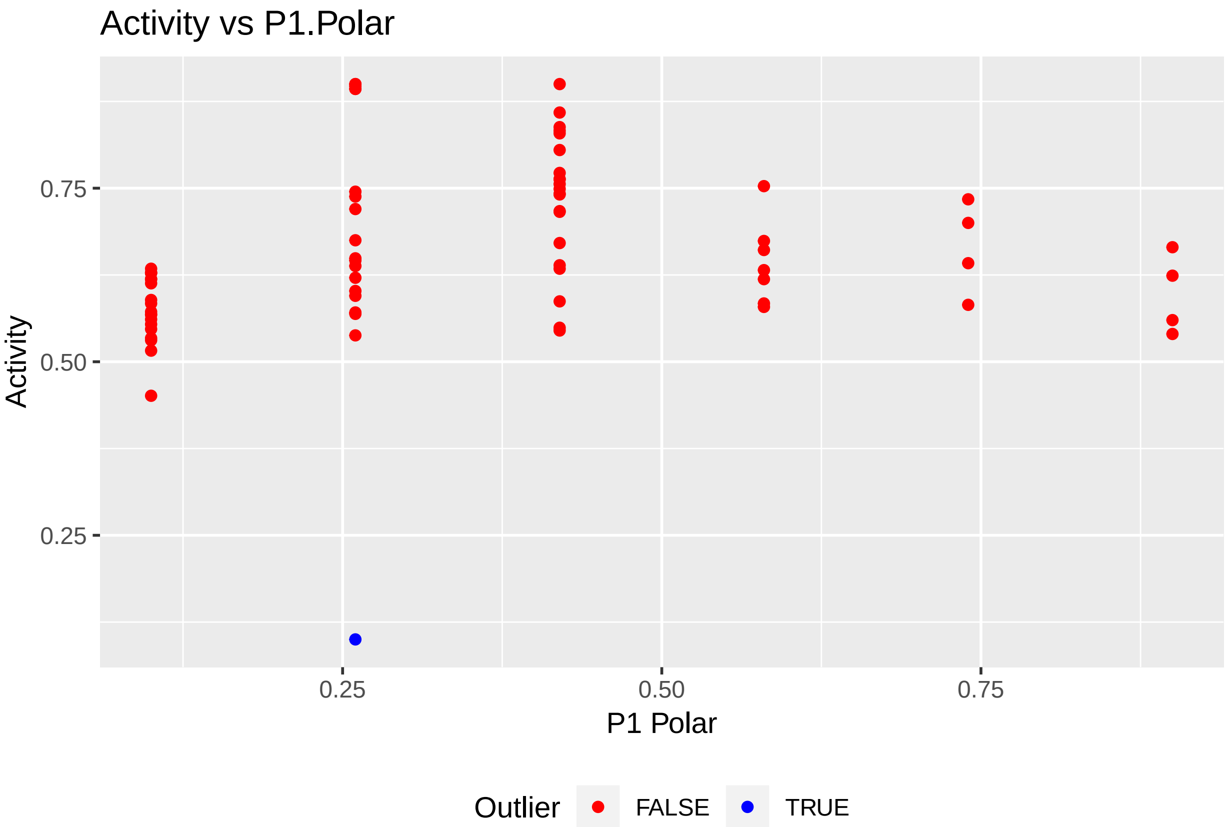
Chapter 3 | Question 4

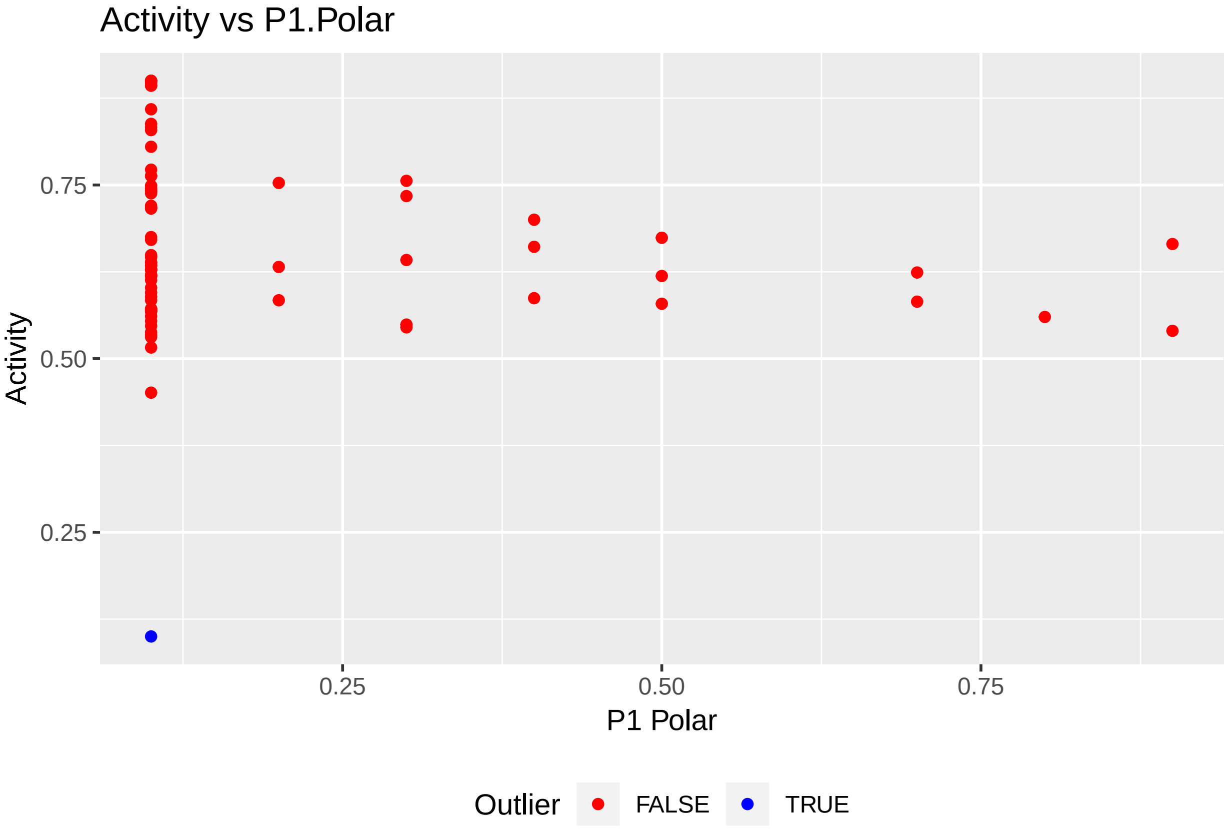
(a) Plot the activity (response) against the first three predictors. Are any outliers in the response apparent? Remove any such cases

The following plots, plot activity vs the first 3 predictors of the `pyrimidines` dataset namely p1.size, p1.polar and p1.flex. Outliers are identified in blue and were removed.



Activity vs P1.Polar

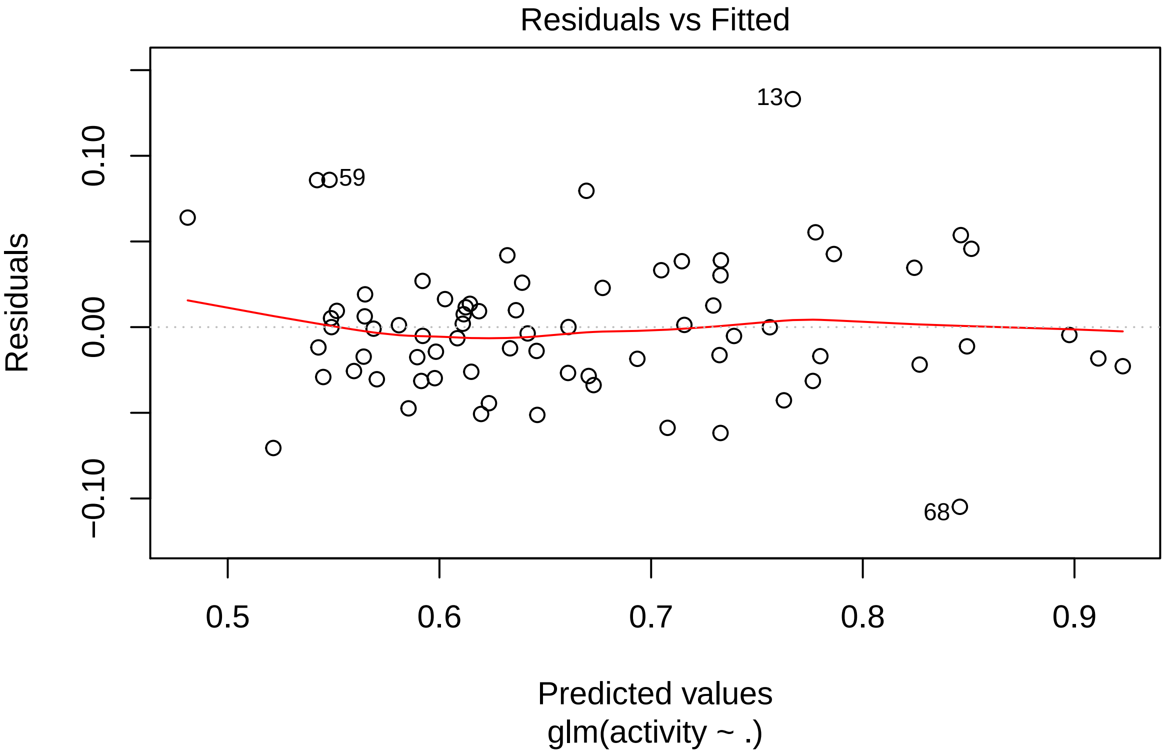


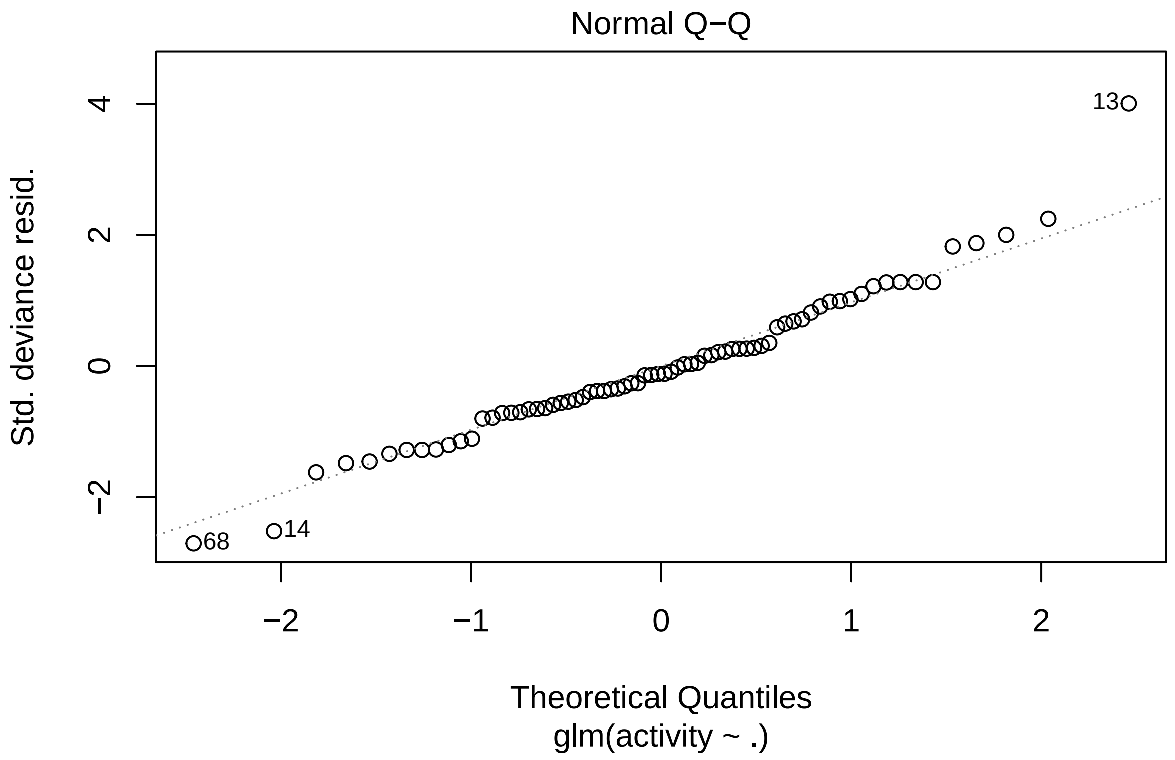


(b) Fit a Gaussian linear model for the response with all 26 predictors. How well does this model fit the

data in terms of R2? Plot the residuals against the fitted values. Is there any evidence of a violation of the standard assumptions?

The following plots, plots the residuals vs fitted values and Q-Q plot to assess normality of residuals for the gaussian linear model. The residual vs fitted model shows there is zero mean for residuals, the variance seems constant thought the q-q plaot suggests the residuals may not be normally distributed. A shapiro-wilkes test confirms that at a 10% significance level we reject the null hypothesis that the residuals are normally distributed, while a bp test shows that we fail to reject that their is a constant variance.





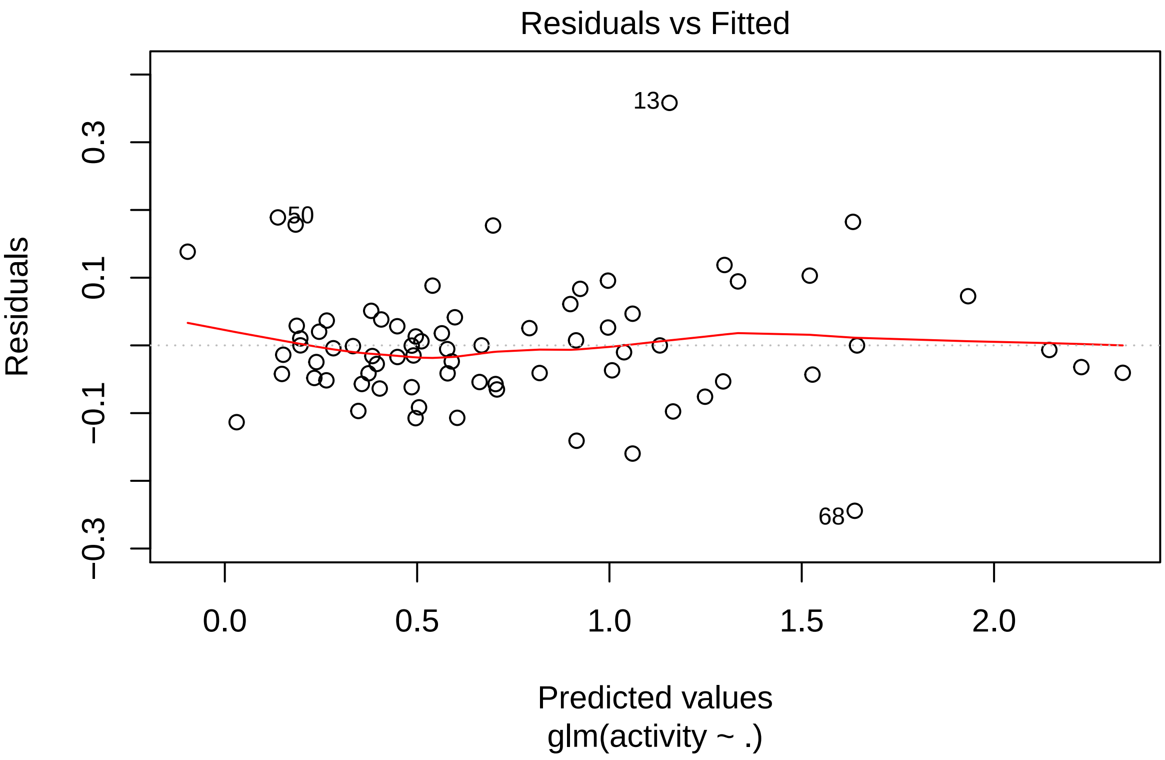
```
## [1] "The r-squared value for the gaussian linear model is:0.87438377322429"
```

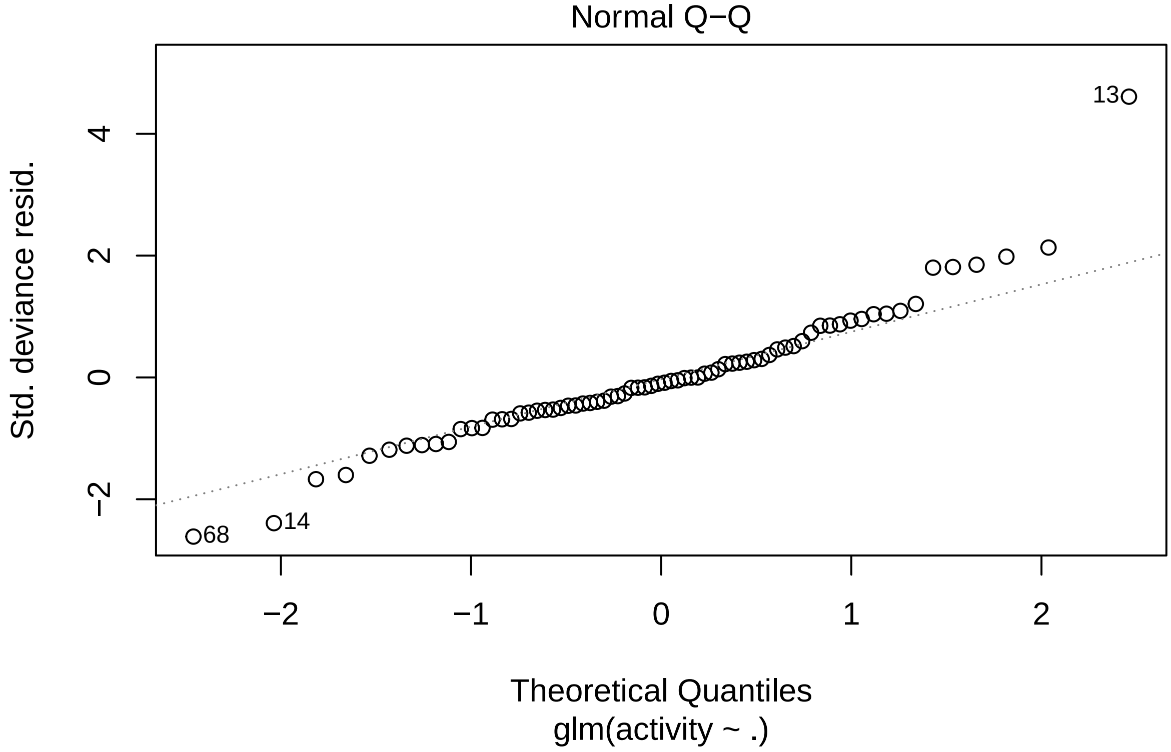
(c) Fit a quasi-binomial model for the activity response. Compare the predicted values for this model to those for the Gaussian linear model. Take care to compute the predicted values in the appropriate scale. Compare the fitted coefficients between the two models. Are there any substantial differences?

The plot `Gaussian and Quasibinomial Fitted Values` plots fitted values from the quasibinomial and gaussion models against eachother. Notice that they map approximately 1 to 1.

The table `Ratio of Model Coefficients` compares the ratios of coefficients of the 2 models sorted by absolute ratio. For the majority of the predictors the coefficients differ substantially.

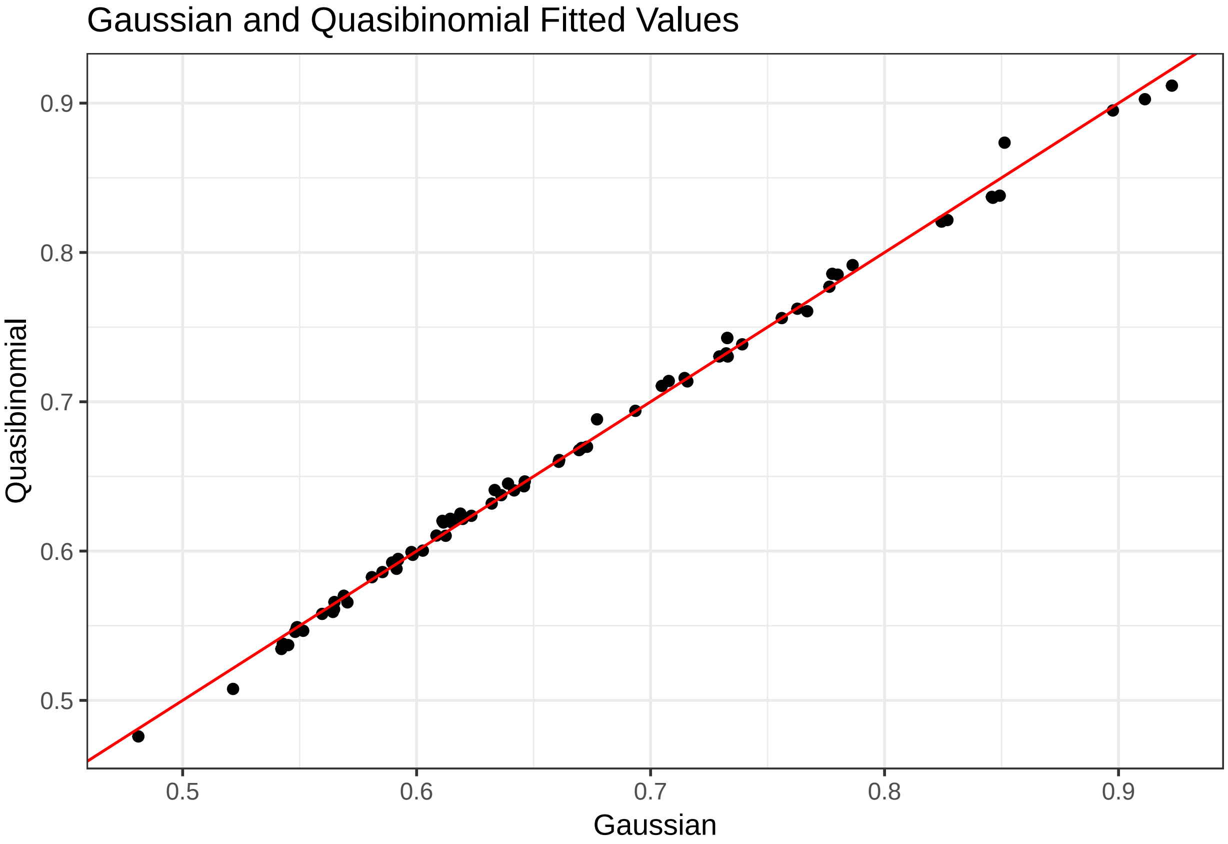
```
## [1] "Quasibinomial Psudo R^2: 0.87"
```





```
## [1] "The r-squared value for the gaussian linear model is:0.87438377322429"
```

```
##         1         2         3         4         5         6         7         8         9        10
## 0.5647202 0.9228129 0.7776576 0.5808790 0.5921090 0.5490000 0.7293934 0.6607505 0.6728256 0.5642200
##        11        12        13        14        15        16        17        18        19        20
## 0.5688931 0.5451172 0.7669289 0.5853934 0.5428326 0.7799423 0.6026488 0.6110314 0.6114755 0.8243498
##        21        22        23        24        25        26        27        28        29        30
## 0.5704242 0.9112677 0.8492213 0.8512885 0.7764176 0.5914189 0.5648656 0.8463005 0.8975819 0.6320833
##        31        32        33        34        35        36        37        38        39        40
## 0.6196773 0.6233959 0.6705152 0.7627319 0.5920347 0.6458849 0.5215374 0.5895248 0.7047895 0.5515199
##        41        42        43        44        45        46        47        48        49        50
## 0.7327827 0.6124136 0.5596260 0.5487898 0.6186974 0.6416793 0.7863450 0.5983417 0.6084689 0.5421965
##        51        52        53        54        55        56        57        58        59        60
## 0.6462356 0.6361466 0.4810687 0.6934960 0.5977897 0.6150495 0.6333631 0.6143962 0.5480722 0.7077747
##        61        62        63        64        65        66        67        68        69        70
## 0.6609509 0.6390671 0.7327827 0.6771014 0.7323142 0.7157076 0.7391581 0.8458564 0.6694378 0.7145278
##        71        72        73
## 0.7560737 0.7329674 0.8268597
```



(d) Fit a Gaussian linear model with the logit transformation applied to the re- sponse. Compare the coefficients of this model with the quasi-binomial model.

Table `Ratio of Model Coefficients | Logit and Quasibiniomial` shows the ratio of the coefficients of a quasinomial and gaussian model with logit response. Notice that the coefficients are much closer this time.

(e) Fit a Beta regression model. Compare the coefficients of this model with that of logit response regression model.

Table `Ratio of Model Coefficients | Logit and Beta` shows the ratio of the coefficients of a quasinomial and gaussian model with logit response. Notice that the coefficients are still close.

```
## Loading required package: nlme
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:dplyr':
##
```

Table 4: Ratio of Model Coefficients

Coefficient	ratio
p3.pi.doner	-22.3922630
p3.polarisable	14.7353659
p3.polar	6.5297363
p3.sigma	6.3381971
p2.polar	6.1535711
p3.h.doner	6.0926247
p3.h.acceptor	5.6763005
p2.h.acceptor	5.2587177
p2.flex	4.9938472
p2.size	4.9122031
p2.pi.doner	4.8665174
p1.sigma	4.7430927
p1.h.acceptor	4.7331615
p1.polar	4.7280665
p2.pi.acceptor	4.7011211
p1.polarisable	4.6279969
p2.polarisable	4.5572975
p1.flex	4.3884772
p1.size	4.3648930
p1.pi.acceptor	4.2762463
p1.h.doner	4.1915994
p3.size	4.0670318
p3.flex	3.7545360
p2.h.doner	3.4923142
p1.pi.doner	3.3112227
p2.sigma	-0.1661803
(Intercept)	0.0932964

```
##     collapse
```

```
## This is mgcv 1.8-27. For overview type 'help("mgcv-package")'.
```

```
##    (Intercept)       p1.polar        p1.size        p1.flex     p1.h.doner  p1.h.acceptor
##     0.06638845    -1.30075800     0.68199536    -0.89892434    -0.51958845     0.24389482
##    p1.pi.doner p1.pi.acceptor p1.polarisable       p1.sigma       p2.polar        p2.size
##     0.16747768     0.48565595     0.67591612     1.32438985    -0.25287951     0.76894030
##        p2.flex     p2.h.doner  p2.h.acceptor    p2.pi.doner p2.pi.acceptor p2.polarisable
##    -1.16337272     0.06296800     0.08760704    -0.01797100    -0.02046421     0.22679415
##       p2.sigma       p3.polar        p3.size        p3.flex     p3.h.doner  p3.h.acceptor
##     0.04654013    -0.64590357     1.53010260    -0.51692882     1.23817312    -1.44155832
##    p3.pi.doner p3.polarisable       p3.sigma
##     0.05549494     0.40111498     1.67938382
```

(F) What property of the response leads to the similarity of the models considered thus far in this question?

The similarities between the response variables is attributable to the fact they are valued between 0 and 1.

Table 5: Ratio of Model Coefficients | Logit and Quasibinomial

Coefficient	ratio
(Intercept)	-4.1998807
p2.h.acceptor	1.9594527
p2.pi.doner	1.8240392
p2.polarisable	1.7702302
p3.polar	1.3851018
p3.sigma	1.3189658
p1.pi.acceptor	1.1885191
p3.pi.doner	1.1638623
p1.polar	1.1370782
p1.polarisable	1.1344678
p1.h.acceptor	1.1228480
p1.size	1.1119473
p1.flex	1.0557509
p1.sigma	1.0415453
p3.h.doner	1.0213384
p3.h.acceptor	0.9952718
p3.polarisable	0.9789098
p2.flex	0.9367060
p2.size	0.9004419
p1.h.doner	0.8659983
p3.size	0.8564376
p2.polar	-0.7744854
p3.flex	0.7159891
p2.h.doner	0.6169101
p1.pi.doner	0.6047603
p2.pi.acceptor	0.2542903
p2.sigma	-0.0040024

Chapter 5 | Question 3

Examine the data for the period of operation 1960–1974 for ships constructed in the years 1975–1979. Why are there no damage incidents?

There are no damage incidents as the ships were built after the operation period.

(b) Make a two-way table that shows the rate of damage incidents per 1000 months of aggregate service classified by type and year of construction. Comment on the table.

```
##     year
## type    60    65    70    75
##    A  0.00  6.39  9.34  4.90
##    B  2.56  4.63  5.06  2.53
##    C  2.66  1.48  8.69  3.65
##    D  0.00  0.00 14.84  1.95
##    E  0.00 24.89  9.87  1.85
```

Ships of type “E” constructed in 1965 and ships lf type “D” constructed in 1970 have high service rates.

(c) Compute the rate of damage incidents per year for all cases where some service was recorded.

```
## # A tibble: 33 x 2
##    `service/12` `sum(incidents)`
```

Table 6: Ratio of Model Coefficients | Logit and Beta

Coefficient	ratio
(Intercept)	-5.6946501
p2.polarisable	1.6274798
p2.h.acceptor	1.4004847
p3.polar	1.3263492
p3.sigma	1.2828193
p3.polarisable	1.2283757
p1.pi.acceptor	1.1394310
p1.h.acceptor	1.1155135
p1.polar	1.1127827
p1.polarisable	1.1126660
p1.size	1.0742303
p1.sigma	1.0535564
p1.flex	1.0300576
p3.flex	0.9853167
p2.polar	-0.9718945
p2.flex	0.9254370
p3.size	0.9199708
p3.h.doner	0.9028809
p2.size	0.9006467
p1.h.doner	0.8894033
p3.h.acceptor	0.8882575
p1.pi.doner	0.6906701
p2.h.doner	0.6160763
p2.pi.doner	0.4740404
p3.pi.doner	0.1648155
p2.sigma	-0.1204011
p2.pi.acceptor	0.1126692

```
##          <dbl>           <int>
##  1        3.75               0
##  2        5.25               0
##  3        8.75               0
##  4       10.6                0
##  5       16                  0
##  6       20.9                0
##  7       22.8                1
##  8       24                  0
##  9       29.1                2
## 10       36.4                7
## # ... with 23 more rows
```

(d) Fit linear models with the observed rate of damage incidents as the response and the following three combinations of predictors: (i) All two-way interac- tions, (ii) main effects terms only, (iii) null (no predictors). Make sure year is treated as a factor rather than numerical variable. Which of these three models is preferred?

(e) Fit a Poisson response model for the number of incidents with the predictors: log of service, type, year and period. Test whether the parameter associated with the service term can be one. Explain why we are interested in such a test.

Table 7: Comparing Linear Models

Interaction Terms	Main Effects	Null
0.63	0.21	0

The following output shows the summary of poisson model without offset:

```
##
## Call:
## glm(formula = incidents ~ log(service) + type + year + period,
##     family = poisson, data = shipdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2355  -1.0345  -0.4454   0.6005   2.8353
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -8.616856   1.528004  -5.639 1.71e-08 ***
## log(service)  0.886469   0.099297   8.927  < 2e-16 ***
## typeB        -0.330248   0.261301  -1.264   0.2063
## typeC        -0.736295   0.341342  -2.157   0.0310 *
## typeD        -0.284220   0.291989  -0.973   0.3304
## typeE         0.335936   0.242645   1.384   0.1662
## year          0.035468   0.013802   2.570   0.0102 *
## period        0.022079   0.008114   2.721   0.0065 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 614.539  on 33  degrees of freedom
## Residual deviance:  58.114  on 26  degrees of freedom
## AIC: 171.98
##
## Number of Fisher Scoring iterations: 5
```

This output then shows the summary of poisson model with offset:

```
##
## Call:
## glm(formula = incidents ~ offset(log(service)) + type + year +
##     period, family = poisson, data = shipdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5348  -0.9319  -0.3686   0.4654   2.8833
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -10.079076   0.876149 -11.504  < 2e-16 ***
## typeB        -0.546090   0.178415  -3.061 0.002208 **
## typeC        -0.632631   0.329500  -1.920 0.054862 .
## typeD        -0.232257   0.287979  -0.807 0.419951
## typeE         0.405975   0.234933   1.728 0.083981 .  
## year          0.042247   0.012826   3.294 0.000988 ***
## period        0.023705   0.008091   2.930 0.003392 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 146.328  on 33  degrees of freedom
## Residual deviance:  59.375  on 27  degrees of freedom
## AIC: 171.24
## 
## Number of Fisher Scoring iterations: 5
```

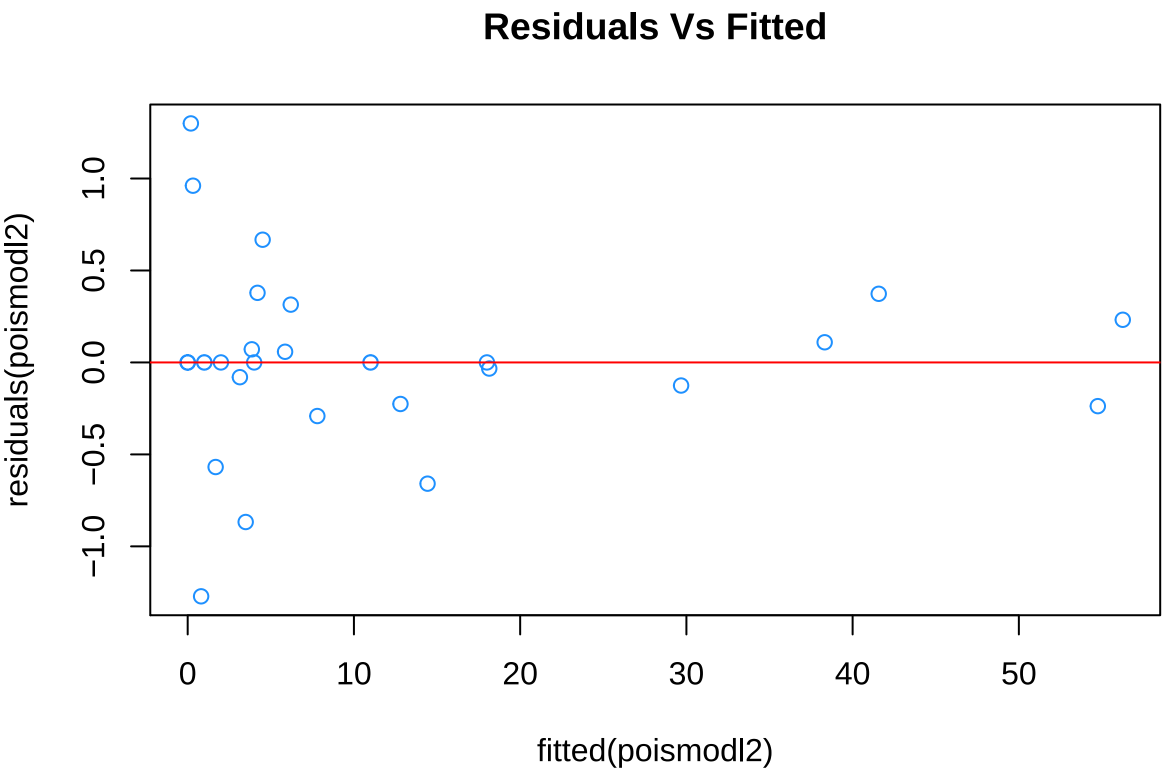
Introducing an offset simplified the model as the log(service) coefficient was set to 1 and the remaining coefficients remained similar as did the deviance.

(f) Fit the Poisson rate model with all two-way interactions of the three predictors. Does this model fit the data?

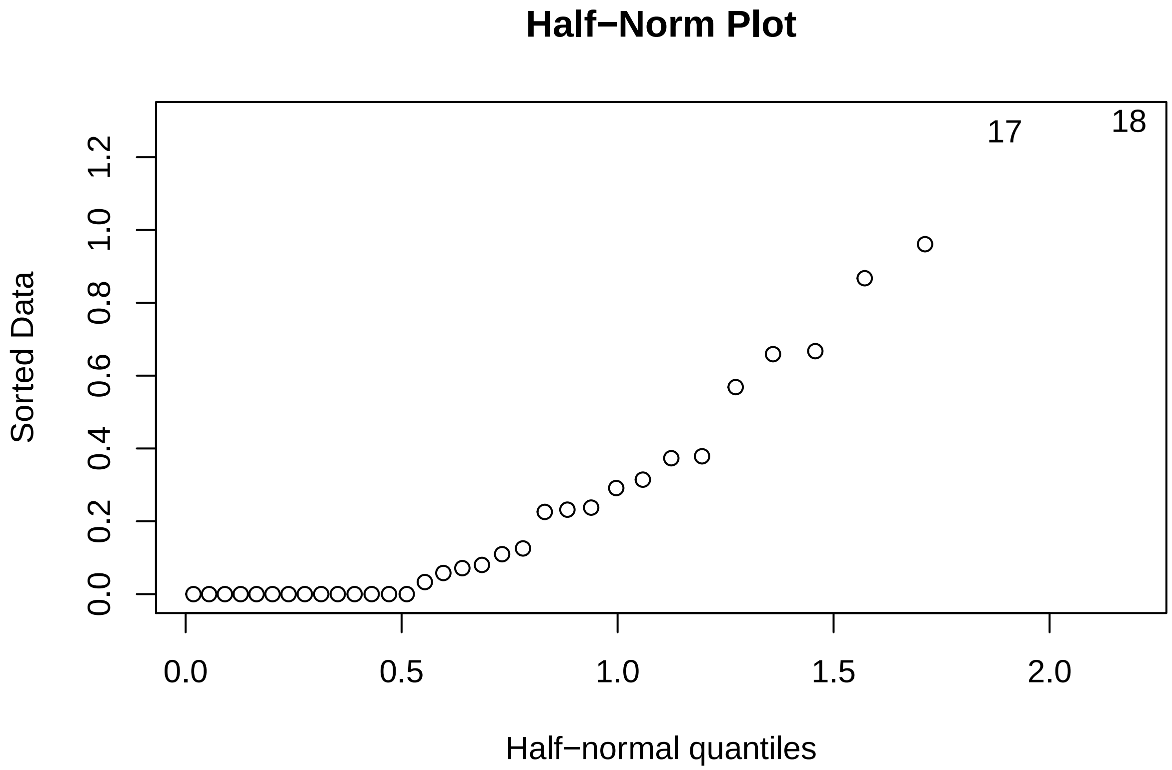
```
## 
## Call:
## glm(formula = incidents ~ offset(log(service)) + (type + factor(year) + 
##     period)^2, family = poisson, data = shipdata)
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -1.27162  -0.06844  -0.00003   0.06804   1.29954  
## 
## Coefficients: (1 not defined because of singularities)
##                         Estimate Std. Error z value Pr(>|z|)  
## (Intercept)           -2.749e+01  1.098e+04  -0.003   0.9980  
## typeB                  1.760e+01  1.098e+04   0.002   0.9987  
## typeC                  2.455e+01  1.098e+04   0.002   0.9982  
## typeD                 -1.178e+00  1.553e+04   0.000   0.9999  
## typeE                  3.641e-01  1.903e+04   0.000   1.0000  
## factor(year)65         2.081e+01  1.098e+04   0.002   0.9985  
## factor(year)70         2.060e+01  1.098e+04   0.002   0.9985  
## factor(year)75         1.918e+01  1.098e+04   0.002   0.9986  
## period                 3.989e-02  3.420e-02   1.166   0.2435  
## typeB:factor(year)65  -1.841e+01  1.098e+04  -0.002   0.9987  
## typeC:factor(year)65  -1.956e+01  1.098e+04  -0.002   0.9986  
## typeD:factor(year)65  -1.942e+01  1.902e+04  -0.001   0.9992  
## typeE:factor(year)65   2.506e-01  1.903e+04   0.000   1.0000  
## typeB:factor(year)70  -1.867e+01  1.098e+04  -0.002   0.9986  
## typeC:factor(year)70  -1.815e+01  1.098e+04  -0.002   0.9987  
## typeD:factor(year)70   1.052e+00  1.553e+04   0.000   0.9999  
## typeE:factor(year)70  -1.052e+00  1.903e+04   0.000   1.0000  
## typeB:factor(year)75  -1.880e+01  1.098e+04  -0.002   0.9986  
## typeC:factor(year)75  -1.734e+01  1.098e+04  -0.002   0.9987  
## typeD:factor(year)75  -3.900e-01  1.553e+04   0.000   1.0000  
## typeE:factor(year)75  -2.119e+00  1.903e+04   0.000   0.9999  
## typeB:period           7.117e-03  3.024e-02   0.235   0.8139  
## typeC:period          -1.001e-01  5.092e-02  -1.966   0.0493 *
## typeD:period           8.621e-03  5.814e-02   0.148   0.8821  
## typeE:period           1.038e-02  3.764e-02   0.276   0.7828
## factor(year)65:period -2.617e-02  2.021e-02  -1.295   0.1954
## factor(year)70:period -1.764e-02  2.406e-02  -0.733   0.4636
## factor(year)75:period         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 146.3283  on 33  degrees of freedom
## Residual deviance:   6.8565  on  7  degrees of freedom
## AIC: 158.72
##
## Number of Fisher Scoring iterations: 18
```

```
## [1] "Goodness of fit using Chi-squared test shows the model fit is sufficient P-value = 02"
```

(g) Check the residuals. Are there any outliers? Plot residuals against the fitted values. Investigate any unusual features of this plot.



The residual plot suggests that there may be outliers present in the data. We can explore this further by inspecting a half norm plot.



This plot confirms that there are a few outliers present in the data.

(h) Now fit the rate model with just the main effects and compare it to the interaction model. Which model is preferred?

The following summary output is for a model with just main effects.

```
##
## Call:
## glm(formula = incidents ~ offset(log(service)) + type + factor(year) +
##     period, family = poisson, data = shipdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6768  -0.8293  -0.4370   0.5058   2.7912
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -7.943769   0.561747 -14.141  < 2e-16 ***
## typeB           -0.543344   0.177590  -3.060  0.00222 **
## typeC           -0.687402   0.329044  -2.089  0.03670 *
## typeD           -0.075961   0.290579  -0.261  0.79377
## typeE            0.325579   0.235879   1.380  0.16750
## factor(year)65   0.697140   0.149641   4.659 3.18e-06 ***
## factor(year)70   0.818427   0.169774   4.821 1.43e-06 ***
## factor(year)75   0.453427   0.233170   1.945  0.05182 .
## period           0.025631   0.007885   3.251  0.00115 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
##     Null deviance: 146.328  on 33  degrees of freedom
## Residual deviance:  38.695  on 25  degrees of freedom
## AIC: 154.56
##
## Number of Fisher Scoring iterations: 5
```

Anova testing allows for the comparison models with and without the two-way interaction terms.

```
## Analysis of Deviance Table
##
## Model 1: incidents ~ offset(log(service)) + type + factor(year) + period
## Model 2: incidents ~ offset(log(service)) + (type + factor(year) + period)^2
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        25     38.695
## 2         7      6.856 18   31.839  0.02297 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The anova test yields a p-value that is small therefore we reject the null hypothesis in favour of the larger model with interaction terms.

(i) Fit quasi-Poisson versions of the two previous models and repeat the comparison.

```
## Analysis of Deviance Table
##
## Model 1: incidents ~ offset(log(service)) + type + factor(year) + period
## Model 2: incidents ~ offset(log(service)) + (type + factor(year) + period)^2
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        25     38.695
## 2         7      6.856 18   31.839  0.07408 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is now 0.074 therfore our action depends on significance level. At a 5% significance level so we fail to reject the main effect model in favor of the model with interaction terms.

(j) Interpret the coefficients of the main effects of the quasi-Poisson model. What factors are associated with higher and lower rates of damage incidents?

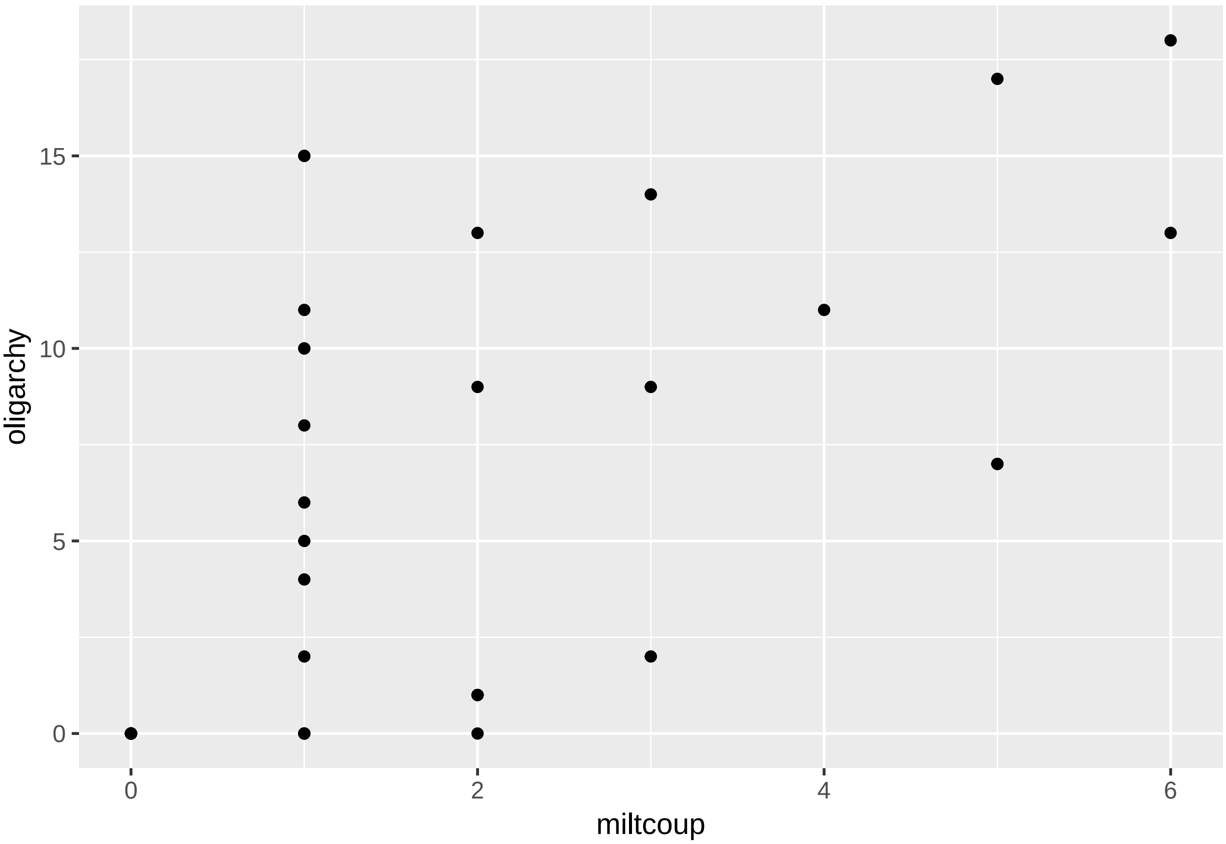
```
##
## Call:
## glm(formula = incidents ~ offset(log(service)) + type + factor(year) +
##     period, family = quasipoisson(), data = shipdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6768  -0.8293  -0.4370   0.5058   2.7912
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -7.94377    0.73049 -10.875 5.75e-11 ***
## typeB          -0.54334    0.23094  -2.353  0.02681 *
## typeC          -0.68740    0.42789  -1.607  0.12072
## typeD          -0.07596    0.37787  -0.201  0.84230
## typeE           0.32558    0.30674   1.061  0.29864
## factor(year)65  0.69714    0.19459   3.583  0.00143 **
## factor(year)70  0.81843    0.22077   3.707  0.00105 **
## factor(year)75  0.45343    0.30321   1.495  0.14733
## period          0.02563    0.01025   2.500  0.01935 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 1.691028)
##
##     Null deviance: 146.328  on 33  degrees of freedom
## Residual deviance:  38.695  on 25  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

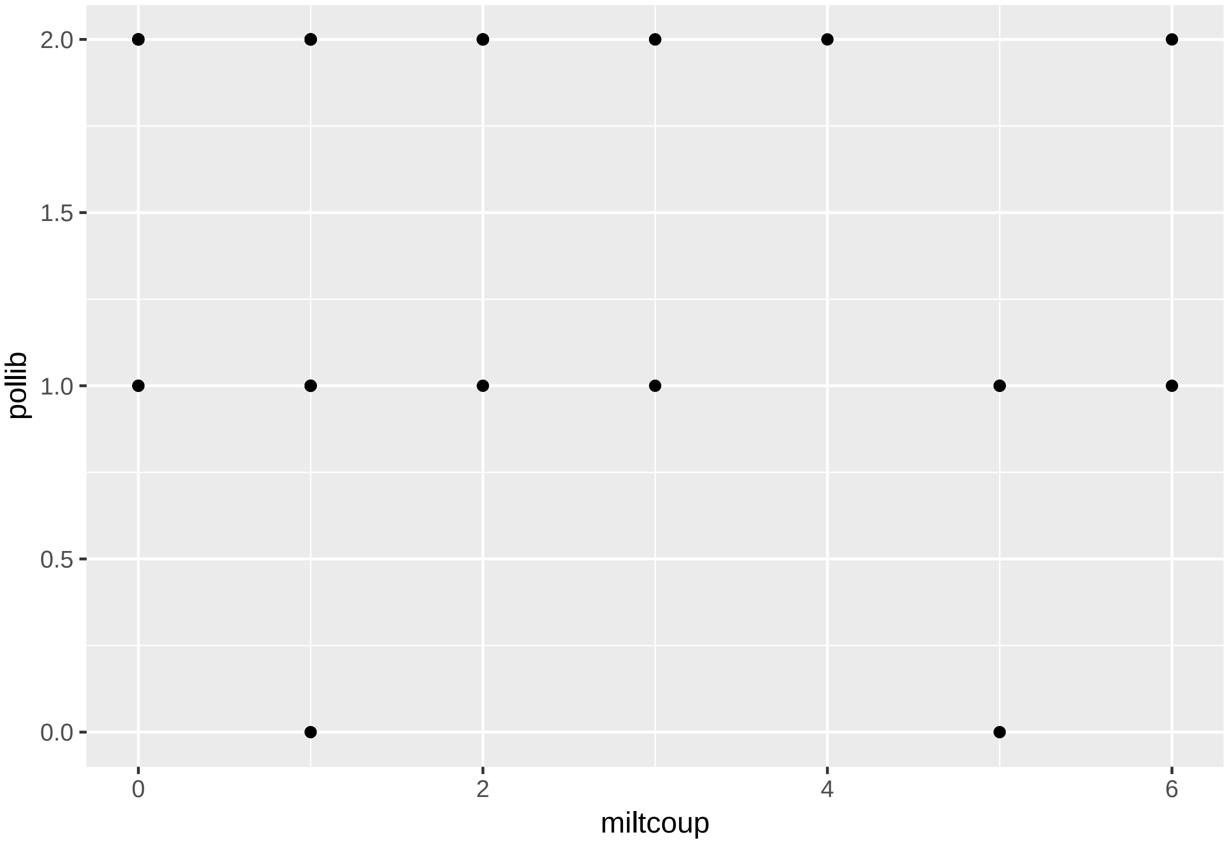
It is interesting to note that incidence rates are more sensitive to being built in 65/70 than 75. Also note ships of type C and B are likely to have lower incidence rates.

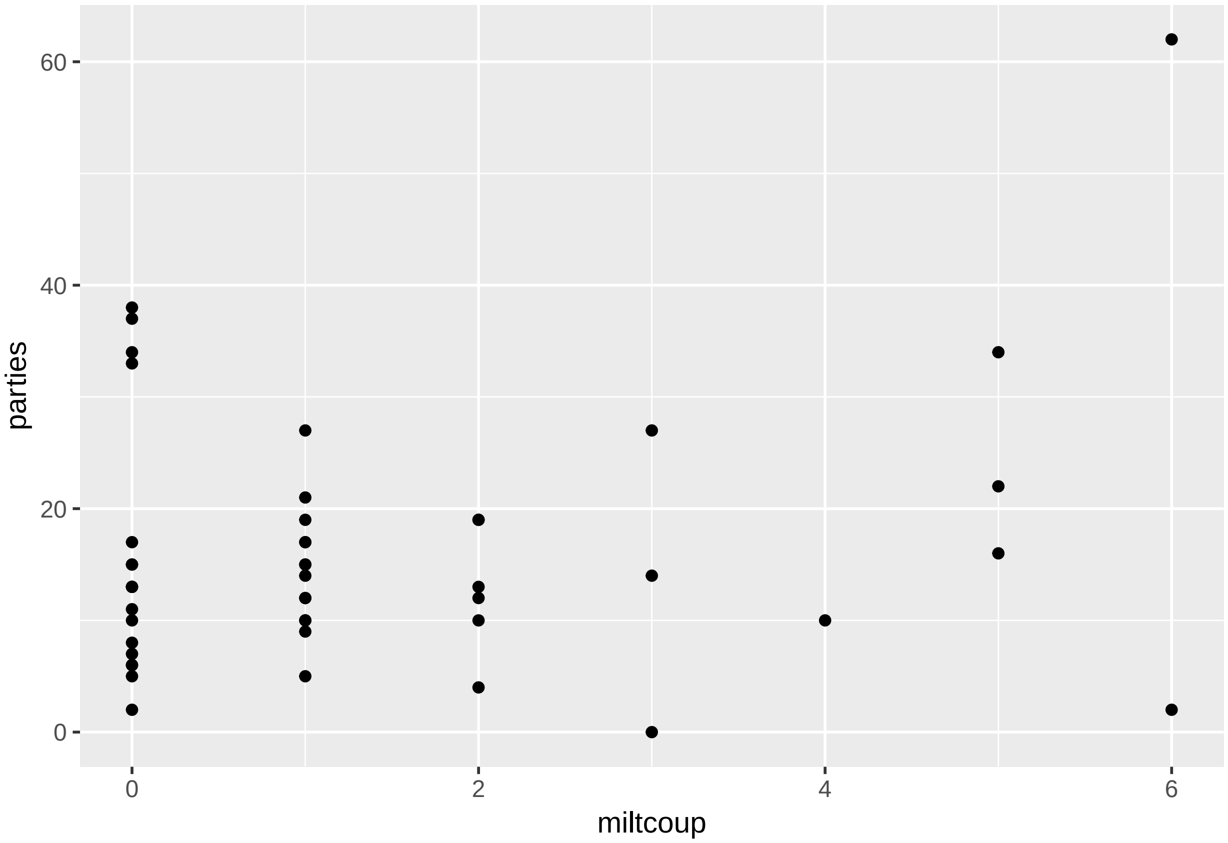
Chapter 5 | Question 3

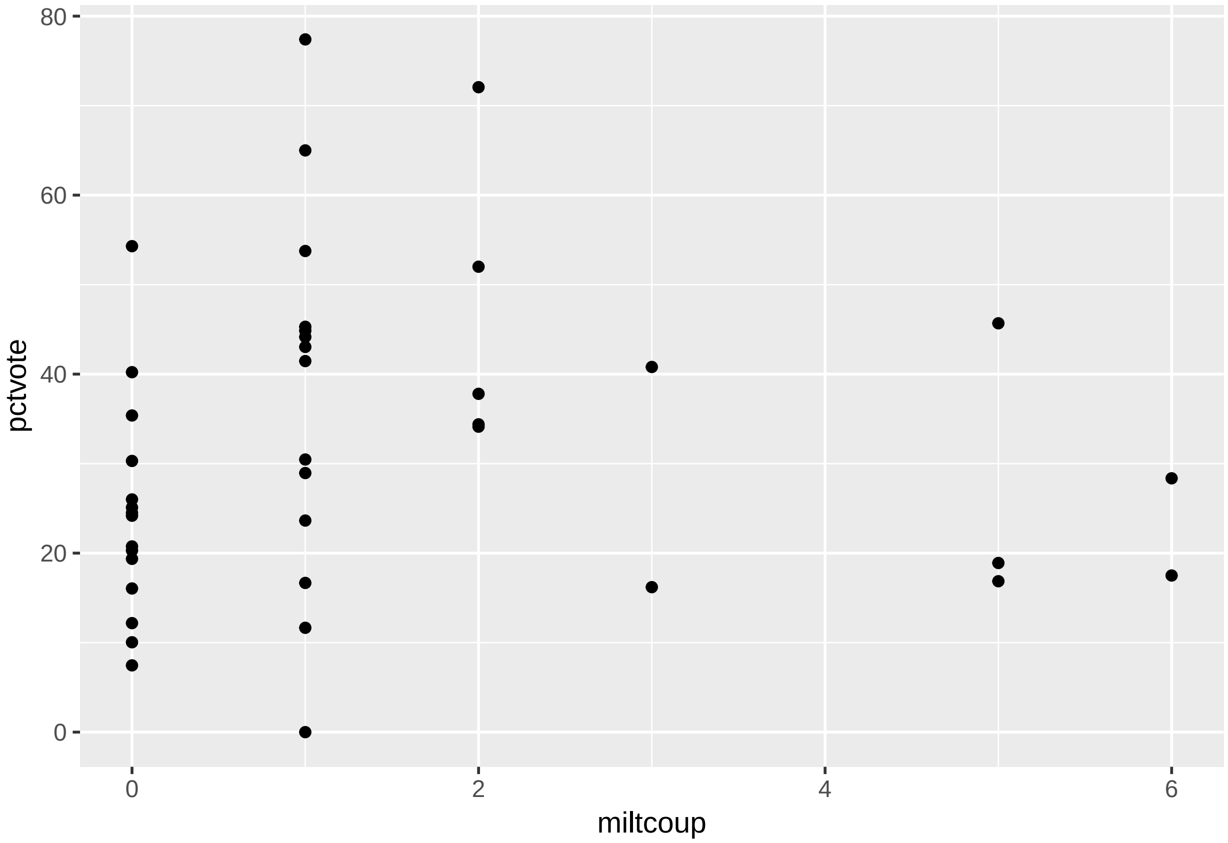
This dataset, Africa, gives information about the number of military coups in sub-Saharan Africa. These are plots of the interaction between the response variable and each of the predictors.

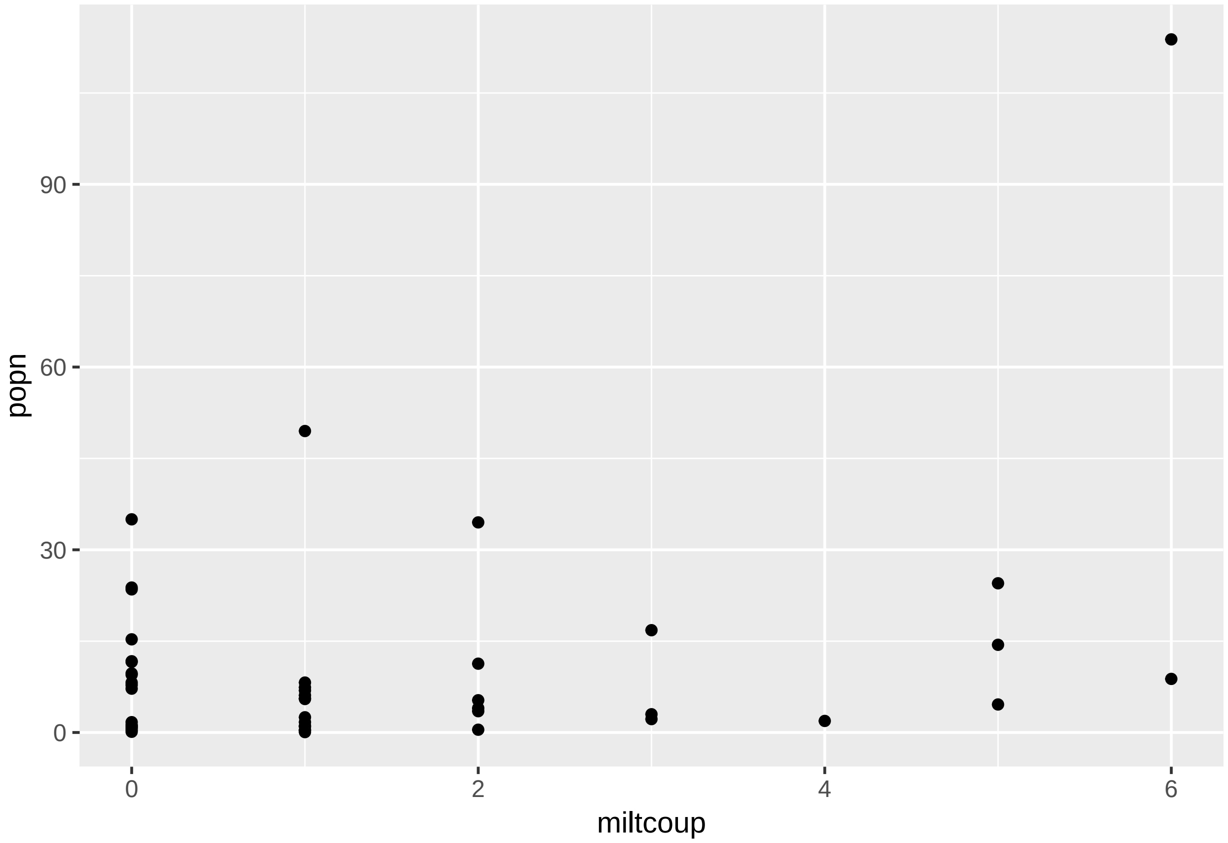
(a) Plot the response, the number of military coups against each of the other variables.)

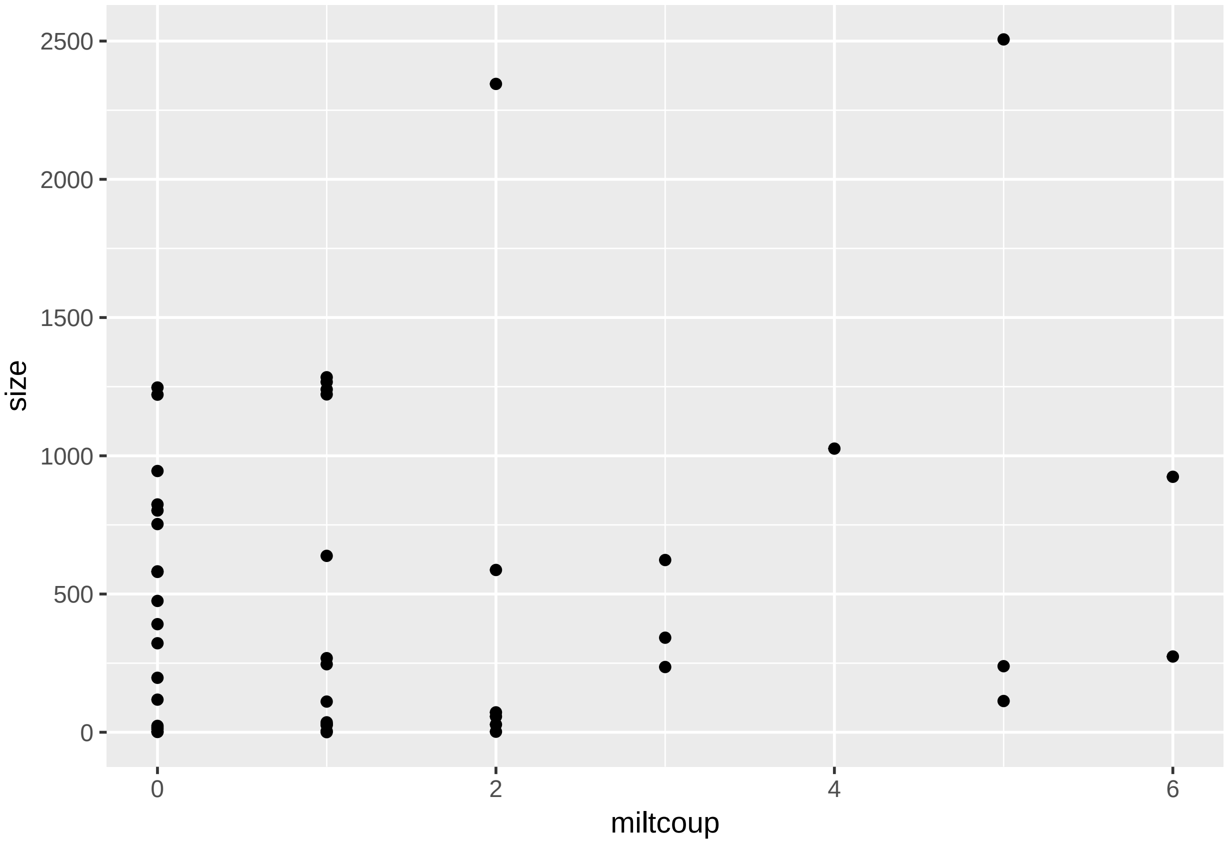


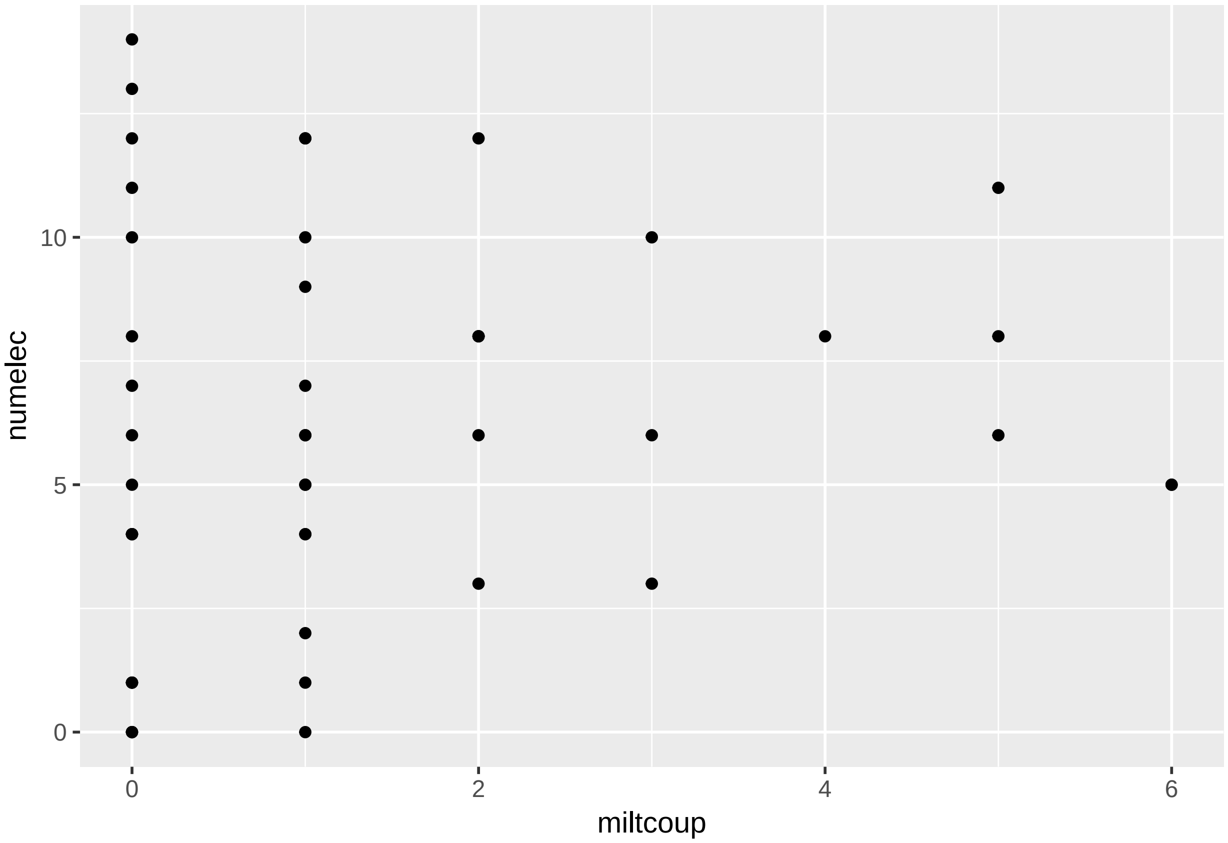

2.0
1.5
1.0
0.5
0.0
pollib
0
2
4
6
miltcoup

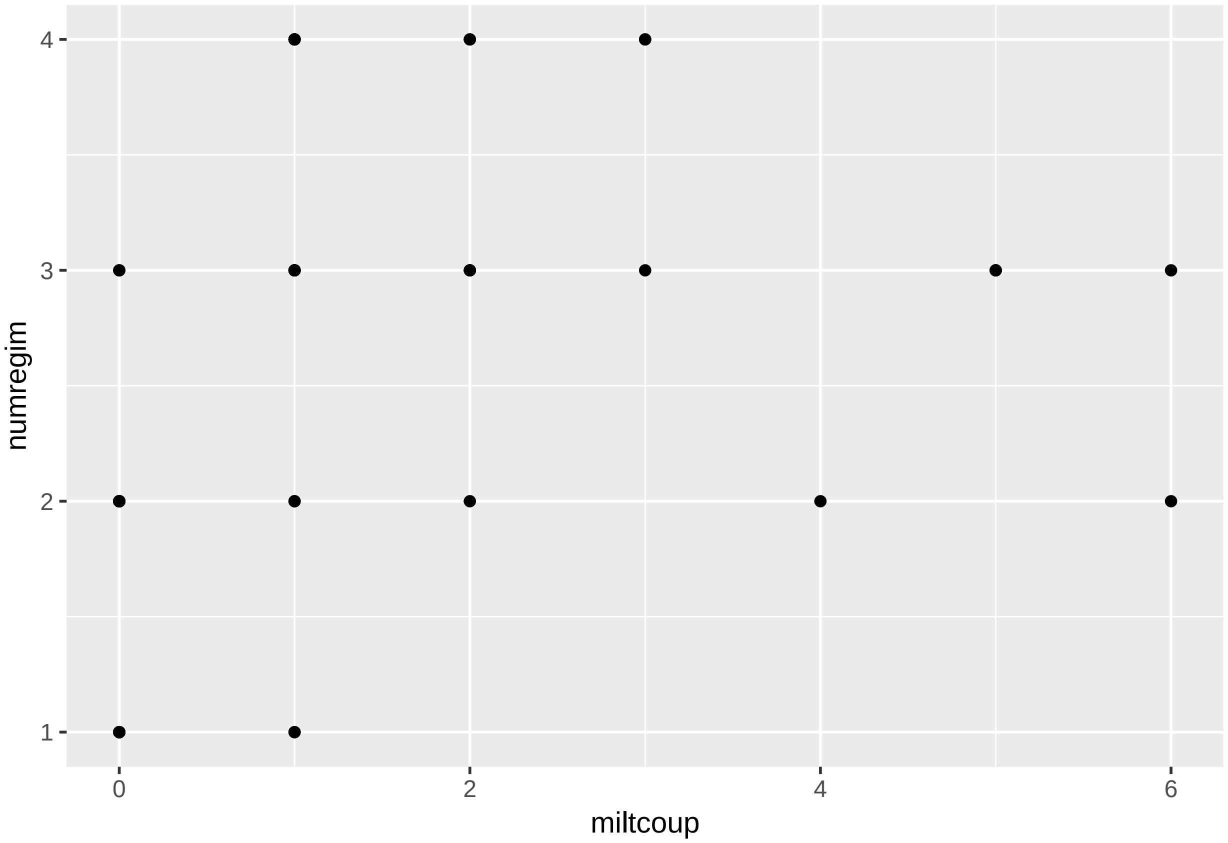

60
40
20
0
parties
0
2
4
6
miltcoup


pctvote
80
60
40
20
0
0
2
4
6
miltcoup


popn
90
60
30
0
0
2
4
6
miltcoup


size
2500
2000
1500
1000
500
0
0
2
4
6
miltcoup


numelec
10
5
0
0
2
4
6
miltcoup



(b) Use a stepwise AIC-based method to select a model that uses a smaller number of the available predictors.

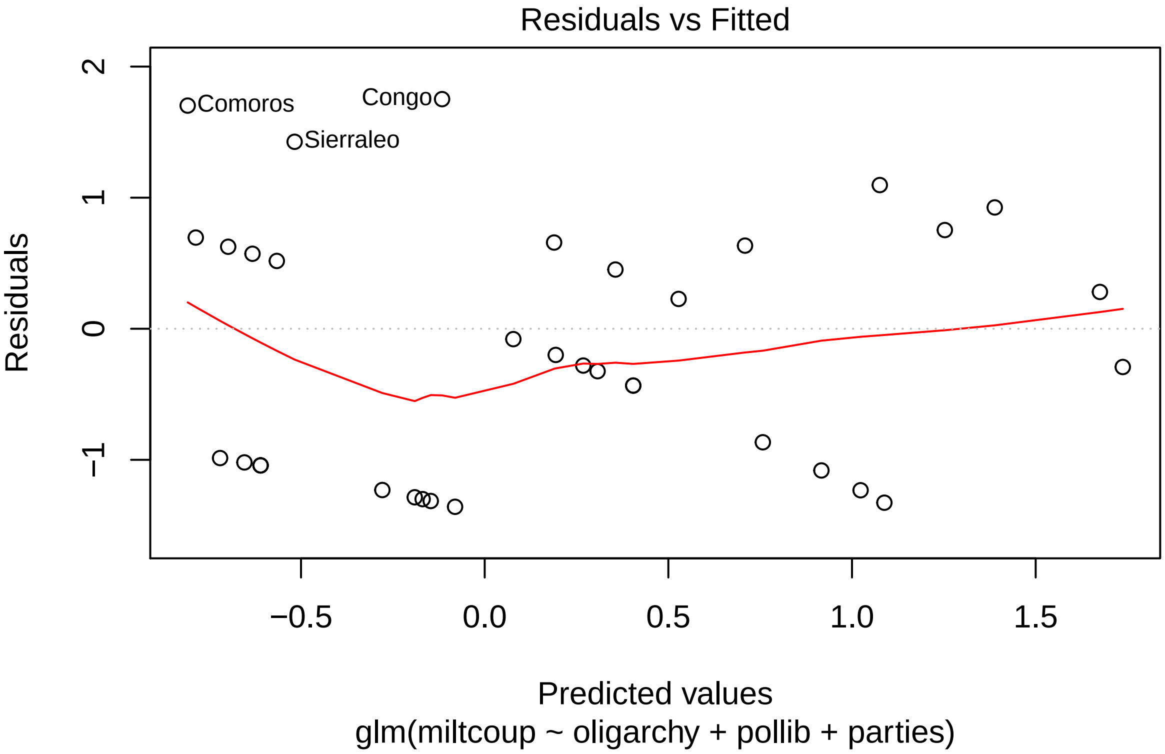
```
##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties, family = poisson,
##     data = africadata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3583  -1.0424  -0.2863   0.6278   1.7517
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.251377   0.372689   0.674  0.50000
## oligarchy    0.092622   0.021779   4.253 2.11e-05 ***
## pollib      -0.574103   0.204383  -2.809  0.00497 **
## parties      0.022059   0.008955   2.463  0.01377 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 65.945  on 35  degrees of freedom
## Residual deviance: 32.856  on 32  degrees of freedom
## AIC: 105.66
##
## Number of Fisher Scoring iterations: 5
```

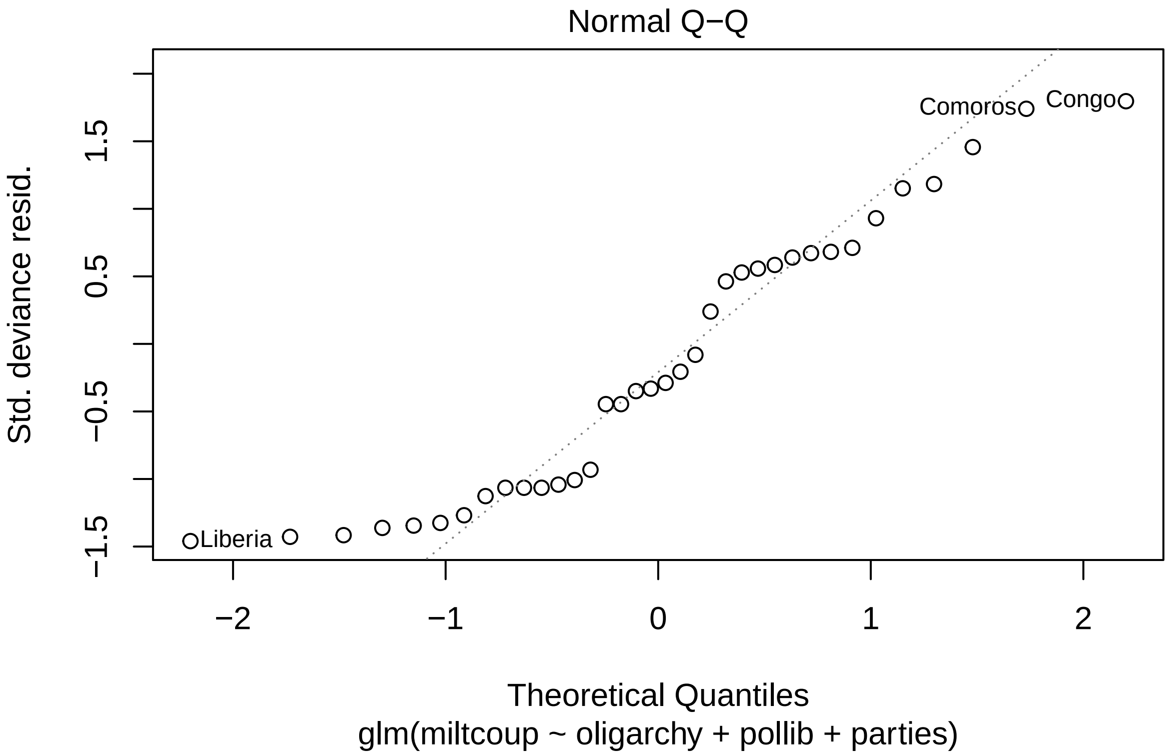
The following predictors are considered significant using AIC: + oligarchy + pollib + parties.

(c) Does the deviance of your selected model indicate a good fit to the data?

The deviance is near the degrees of freedom and the null deviance is reduced by about half. Therefore the deviance does indicate a good fit.

(d) Make a QQ plot of the residuals and comment. Plot the residuals against the fitted values and interpret the result. What is the source of the lines of points observed on this plot?





The QQ plot doesn't suggest normality.

The lines on residual vs fitted plot are caused by having a discrete response variable though continuous predictor.

(e) Give an interpretation of the coefficients of this plot

```
##
## Call:
## glm(formula = miltcoup ~ oligarchy + pollib + parties, family = poisson,
##     data = africadata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3583  -1.0424  -0.2863   0.6278   1.7517
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.251377   0.372689   0.674  0.50000
## oligarchy    0.092622   0.021779   4.253 2.11e-05 ***
## pollib      -0.574103   0.204383  -2.809  0.00497 **
## parties      0.022059   0.008955   2.463  0.01377 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 65.945  on 35  degrees of freedom
## Residual deviance: 32.856  on 32  degrees of freedom
## AIC: 105.66
##
## Number of Fisher Scoring iterations: 5
```

- If a country is liberal there is predicted to be less number of successful coups.
- Positive relationship b/w the number of years the country run by an oligarchy and the the number of military coups.

(f) Count the number of countries with each number of military coups. Compare this with the numbers predicted by the previous model. Is there any evidence of excess of countries with zero coups? Use a Chi-squared test as implemented in chisq.test().

```
## # A tibble: 6 x 2
##   miltcoup     n
##      <int> <int>
## 1        0    10
## 2        1    14
## 3        2     5
## 4        3     2
## 5        5     3
## 6        6     2
```

This is the tally of the number of countries who have each number of coups. (i.e 10 countries have no coups)

```
##     Benin   Burkina   Burundi  Cameroon Capeverde       CAR      Chad   Comoros     Congo CotedIvoi
## 2.9318214 5.3382421 1.6952785 0.8634359 0.4554195 2.0313567 2.9685943 0.4454836 0.8904739 0.9225076
##  Djibouti  Eqguinea  Ethiopia     Gabon     Ghana    Guinea GuineaBis     Kenya   Lesotho   Liberia
## 0.7568359 1.3598583 1.4983891 0.5678206 5.6810908 1.0811428 0.5314608 0.5433144 1.3078042 2.7823884
## Madagasca    Malawi      Mali   Namibia     Niger   Nigeria    Rwanda Seychelle Sierraleo   Somalia
## 1.2080535 0.4865769 2.5010376 0.5198659 2.1323164 4.0092040 1.2133657 0.4974294 0.5960416 1.4983891
##  S.Africa     Sudan Swaziland  Tanzania      Togo    Zambia
## 0.8445982 3.4996802 0.8266497 0.5433144 1.4274442 0.5433144
```

A difference is noted because of the predicted values, many are below 1 (number of coups) and note enough above 2.

```
## 
##  Chi-squared test for given probabilities
## 
## data:  africadata$miltcoup
## X-squared = 68.684, df = 35, p-value = 0.0005751
```

The p-value is very small so we reject the null hypothesis which states they are independent in favor of the alternative, this means the numbers are dependent.