ENV 710: Lab 9

Jiahuan Li

Spring 2023

Problem 1

Hypothesis

The null hypothesis for problem 1 is that there is no significant relationship between the predictors (aircraft type, bomb load, and aircrew experience) and the response variable (damage of attack aircraft during the Vietnam War). The alternative hypothesis is that there is a significant relationship between at least one of the predictors and the response variable.

Method

Data preview

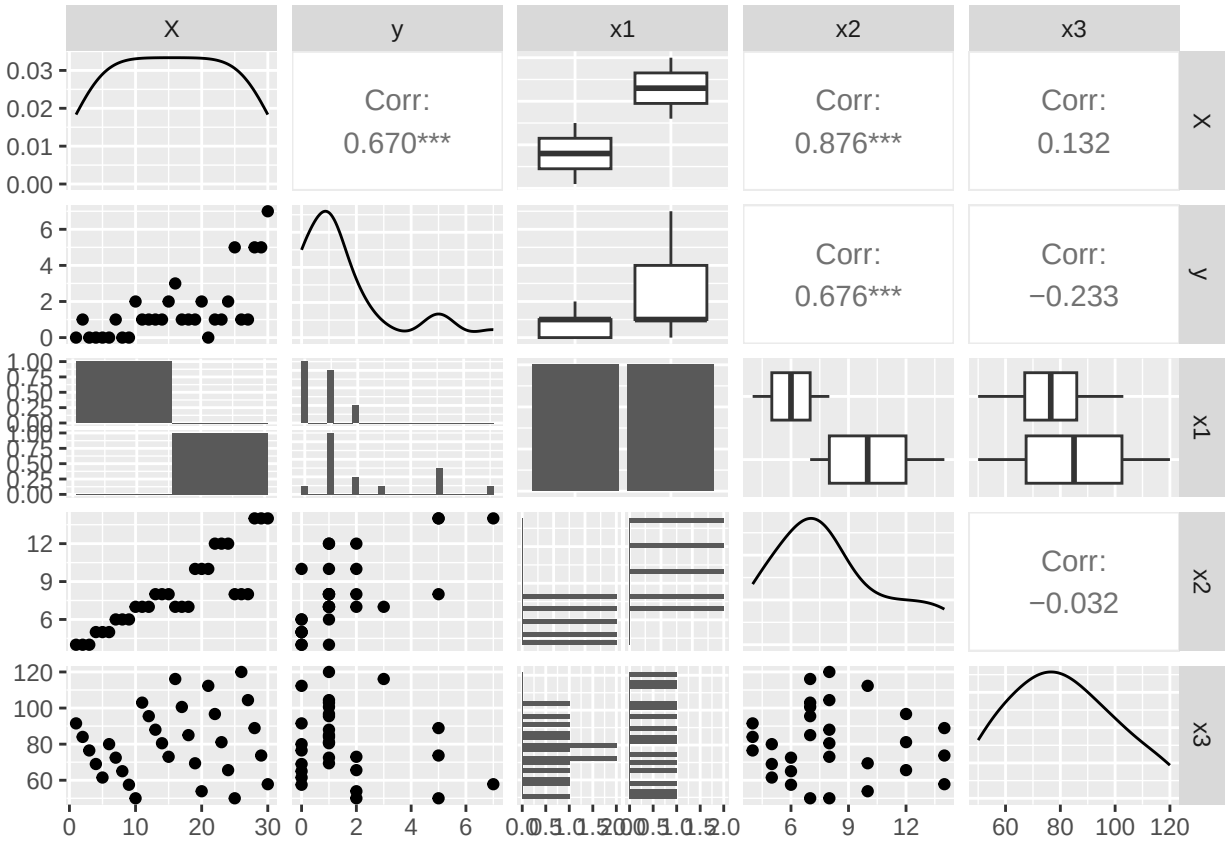

Figure 1: Preview of Dataset

```
##        X                y         x1           x2             x3
##  Min.   : 1.00   Min.   :0.000   0:15   Min.   : 4.0   Min.   : 50.00
##  1st Qu.: 8.25   1st Qu.:0.250   1:15   1st Qu.: 6.0   1st Qu.: 66.45
##  Median :15.50   Median :1.000          Median : 7.5   Median : 80.25
##  Mean   :15.50   Mean   :1.533          Mean   : 8.1   Mean   : 80.77
##  3rd Qu.:22.75   3rd Qu.:2.000          3rd Qu.:10.0   3rd Qu.: 94.50
##  Max.   :30.00   Max.   :7.000          Max.   :14.0   Max.   :120.00
```

From the plot and table, we can observe that the dependent variable varies from 1 to 30, which is appropriate because the dependent variable needs to be a non-negative integer for Poisson regression. It is also not normally distributed, which is also expected since the Poisson distribution is a discrete distribution. Furthermore, two of the independent variables, bomb load (in tons) and total months of aircrew experience, are continuous, and the other independent variable x1 is a categorical variable indicating the type of aircraft (A-4 or A-6). Moreover, there do not seem to be clear linear relationships between the dependent variable and the independent variables in the raw data - that is fine too, because that is not an assumption of Poisson regression.

Initial model

Poisson regression is utilized in this case since the response variable consists of count data. Therefore, the initial model encompassing all potential influencing variables is formulated as follows:

```
##
## Call:
## glm(formula = y ~ as.factor(x1) + x2 + x3, family = poisson,
##     data = air)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6418  -1.0064  -0.0180   0.5581   1.9094
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.406023   0.877489  -0.463   0.6436
## as.factor(x1)1  0.568772   0.504372   1.128   0.2595
## x2              0.165425   0.067541   2.449   0.0143 *
## x3             -0.013522   0.008281  -1.633   0.1025
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 53.883  on 29  degrees of freedom
## Residual deviance: 25.953  on 26  degrees of freedom
## AIC: 87.649
##
## Number of Fisher Scoring iterations: 5
```

The regression summary reveals that only the `x2` terms are statistically significant, suggesting that the model is likely not the minimum adequate model for the problem.

Model reduction

To identify the minimum adequate model, I constructed models for all possible combinations of variables and used LRT tests and AIC values to assess the performance of each model. The results are presented below:

```
## Likelihood ratio test
##
```

```
## Model 1: y ~ as.factor(x1) + x2 + x3
## Model 2: y ~ as.factor(x1) + x3
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -39.825
## 2   3 -42.944 -1 6.2386     0.0125 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Likelihood ratio test
##
## Model 1: y ~ as.factor(x1) + x2 + x3
## Model 2: y ~ x2 + x3
## Model 3: y ~ x2
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -39.825
## 2   3 -40.458 -1 1.2667     0.2604
## 3   2 -41.451 -1 1.9861     0.1588
```

```
## Likelihood ratio test
##
## Model 1: y ~ as.factor(x1) + x2 + x3
## Model 2: y ~ as.factor(x1) + x2
## Model 3: y ~ as.factor(x1)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -39.825
## 2   3 -41.165 -1 2.6811   0.101543
## 3   2 -45.990 -1 9.6492   0.001894 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the first LRT tests, it was found that removing `x2` from the model resulted in a significantly different reduced model compared to the original model, indicating that `x2` is an important variable in the model. Using the same method, results of all three tests jointly showed that only keeping the term `x2` in the regression is appropriate. As shown in the second test, the reduction process did not significantly impact the original model.

The minimum adequate model can also be justified using the Akaike Information Criterion.

```
##       df       AIC
## lm1    4  87.64922
## lm2.1  3  86.91589
## lm2.2  3  91.88781
## lm2.3  3  88.33037
## lm3.1  2  95.97952
## lm3.2  2  86.90196
## lm3.3  2 108.23330
```

We can find that the model with only `x2` as its predictor has the smallest AIC number (86.90196).

Minimum adequate model

```
summary(lm3.2)
```

```
##
## Call:
## glm(formula = y ~ x2, family = poisson, data = air)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9188  -1.0473  -0.1518   0.1650   2.6331
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.70097    0.50685  -3.356 0.000791 ***
## x2           0.23112    0.04677   4.942 7.72e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 53.883  on 29  degrees of freedom
## Residual deviance: 29.206  on 28  degrees of freedom
## AIC: 86.902
##
## Number of Fisher Scoring iterations: 5
```

Results

From the regression report, it can be concluded that bomb load has a significant positive impact on the damage of attack aircraft ($z = 4.942$, $p < 0.001$). The coefficient is 0.23112, which means that for a one-unit increase in tons of bomb load, the expected damage of attack aircraft will increase by a factor of 1.26 ($exp^{0.23112}$). And the intercept does not have a practical interpretation since it assumes a bomb load of zero, which is not possible.

The deviance residuals are a measure of how well the model fits the data. The residual deviance of 29.206 on 28 degrees of freedom, which are close, indicates that the model does not have a great concern about the over-dispersion issue. The null deviance is the deviance of a model with no predictors, and the residual deviance is the deviance of the fitted model. The difference between the two deviances represents the reduction in deviance achieved by the model. In this case, the reduction in deviance is 53.883 - 29.206 = 24.677, which indicates that the model is a significant improvement over the null model.

Assumptions check

```
## [1] -0.171638
```

The calculated mean error is very small (-0.1716) and close to 0 as expected, indicating that the model has a good fit to the data. Besides, the residuals are nearly independent and normally distributed as shown in the diagnosis plots. The plots also indicates that there is no overdispersion issue as 1) no more than 5% of the standardized residuals are greater than 2 in the scale-location plot; and 2) there are no outlier observations in the leverage plot. The dispersive test result also supported this with an overdispersion ratio of 1.037 less than 2.

```
##
##  Overdispersion test
##
## data:  lm3.2
## z = 0.11044, p-value = 0.456
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   1.036967
```

Then I tested the goodness of fit of the model using the chi-square goodness-of-fit test. I compared the

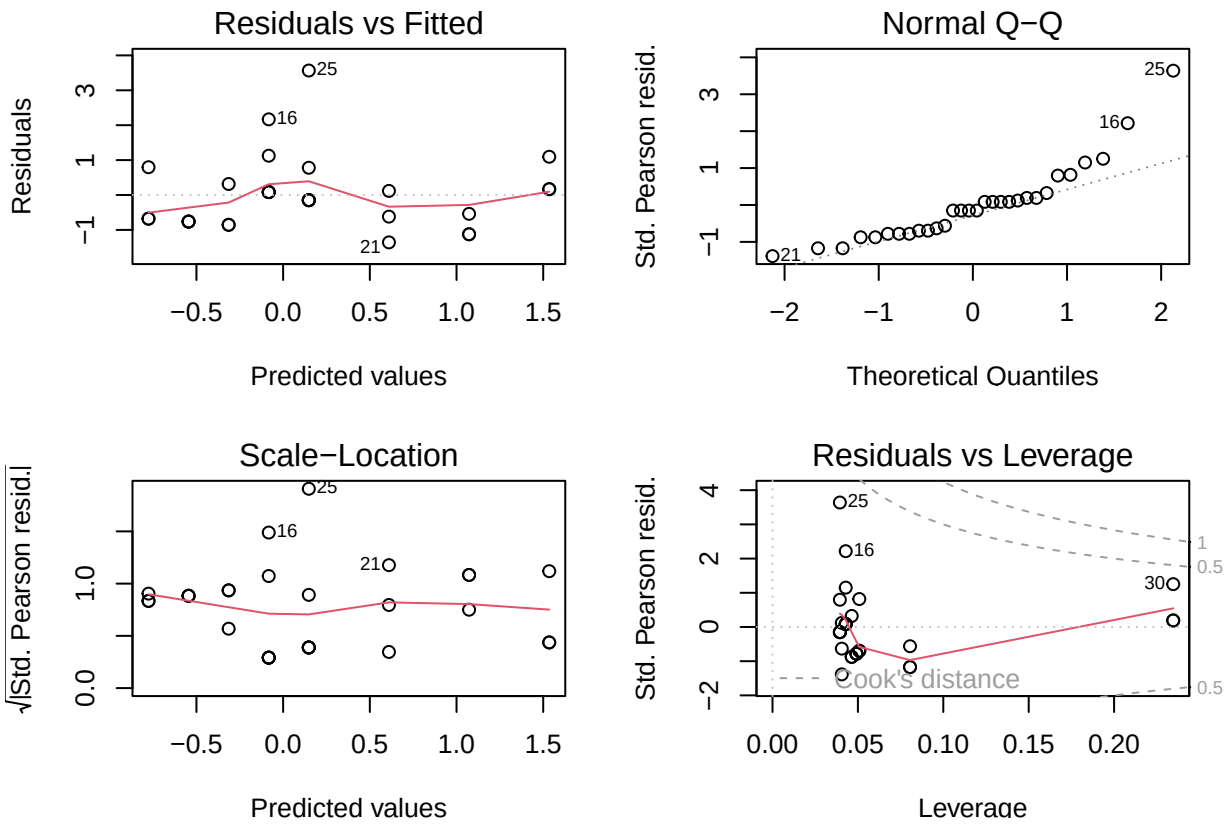

Figure 2: Assumption plots

residual deviance of the reduced model to a chi-square distribution with the same degrees-of-freedom. It returns a value of 0.40, which is not statistically significant and indicates the model is good.

The pseudo-R^2 test returns a value of 0.58, which is slightly less than the initial model with all the predictors (0.62). But the close and reletively high proportion still indicates the reduced model has a good ability to explain the variation in the dependent variable.

```
## VeallZimmermann
##       0.5771834
```

```
## VeallZimmermann
##       0.6165824
```

Visualization

The figure below displays the impact of bomb load on the damage of attack aircraft.

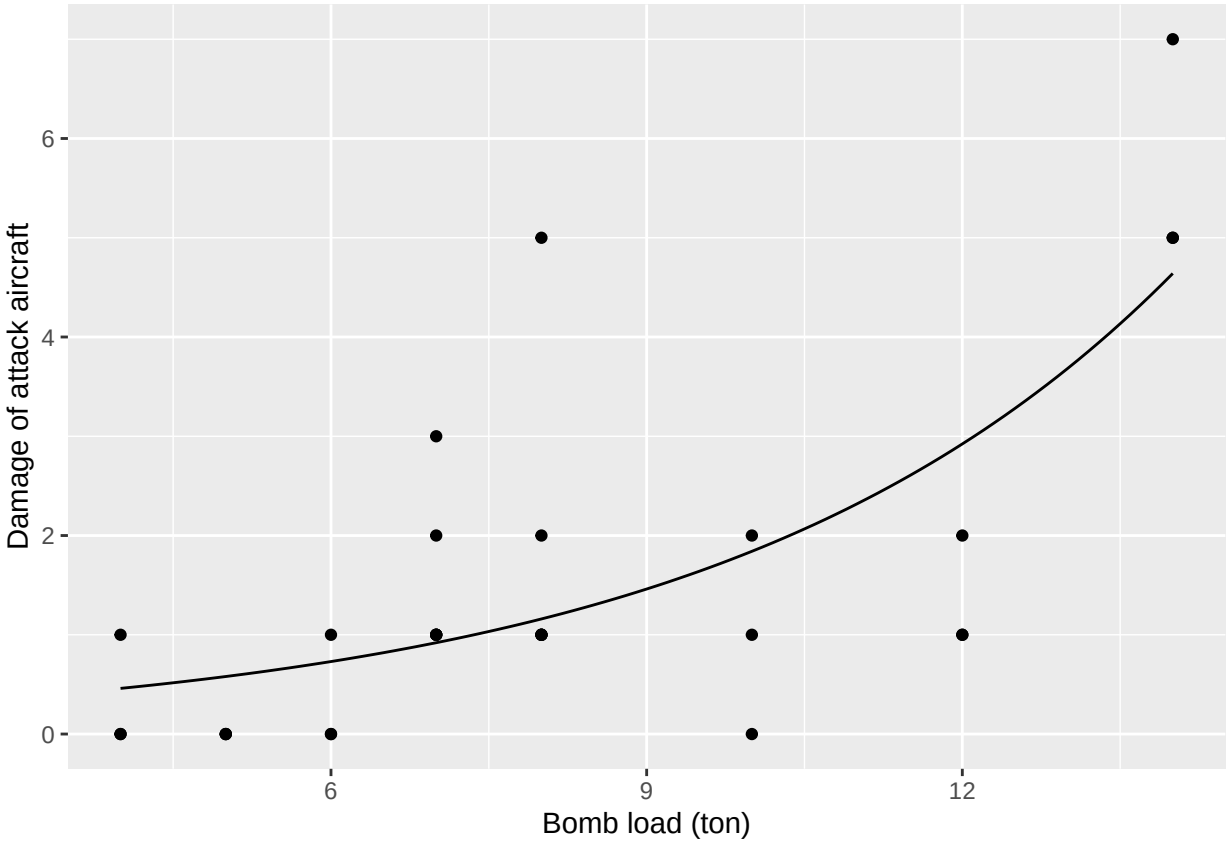

Figure 3: Impact of bomb load on the damage of attack aircraft

Problem 2

Hypothesis

The null hypothesis for this question is that there is no association between bird keeping and lung cancer, after accounting for age and smoking. The alternative hypothesis is that bird keeping is associated with a higher rate of lung cancer, after accounting for age and smoking.

Method

Data preview

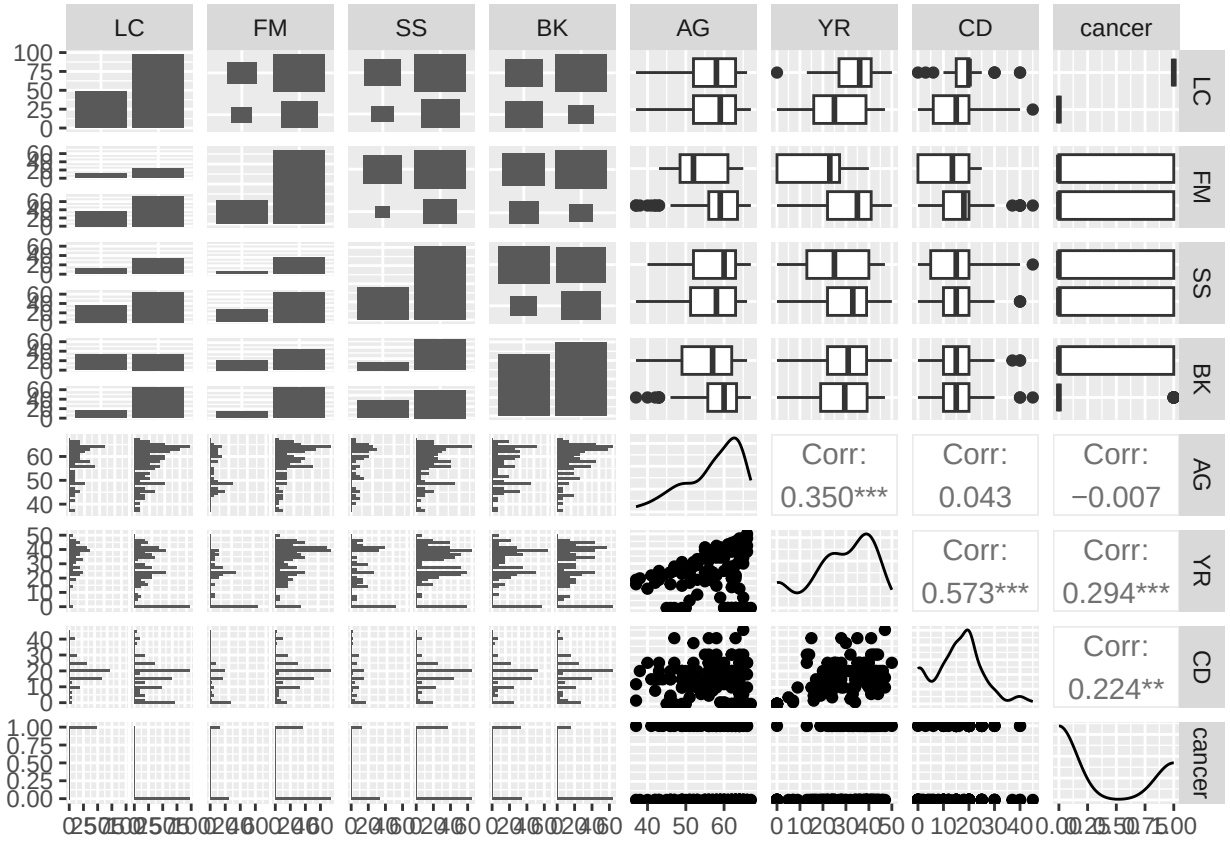

Figure 4: Preview of Dataset

From the plot, we can observe that there might be some interactions among the predictors. Thus, we begin our initial model include all the potential interactions.

Initial model

Logistic regression is utilized in this case. The initial model encompassing all potential influencing variables and interactions is formulated as follows:

```
##
## Call:
## glm(formula = cancer ~ factor(BK) * AG * YR, family = binomial,
##     data = case2002)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5998  -0.8306  -0.4445   0.9694   2.1910
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -3.538014   6.040581  -0.586    0.558
## factor(BK)NoBird       -5.765924  10.504118  -0.549    0.583
```

```
## AG                      0.020210   0.106945   0.189    0.850
## YR                      0.218653   0.234216   0.934    0.351
## factor(BK)NoBird:AG     0.073539   0.179511   0.410    0.682
## factor(BK)NoBird:YR     0.223348   0.404856   0.552    0.581
## AG:YR                  -0.002400   0.003914  -0.613    0.540
## factor(BK)NoBird:AG:YR -0.003726   0.006656  -0.560    0.576
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 187.14  on 146  degrees of freedom
## Residual deviance: 154.21  on 139  degrees of freedom
## AIC: 170.21
##
## Number of Fisher Scoring iterations: 5
```

All the predictors are not statistically significant in the report, suggesting that the model is not the minimum adequate model for the problem.

Model reduction

I used a stepwise variable selection method, removing predictors with higher p-values first, until the remaining predictors all had p-values below a certain threshold (e.g., 0.05). This approach helps to simplify the model and avoid overfitting, while retaining predictors that are most strongly associated with the response variable.

At last, only the predictor `BK` and `YR` are included in the reduced model.

```
##
## Call:
## glm(formula = cancer ~ factor(BK) + YR, family = binomial, data = case2002)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6093  -0.8644  -0.5283   0.9479   2.0937
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.70460    0.56267  -3.030 0.002450 **
## factor(BK)NoBird -1.47555    0.39588  -3.727 0.000194 ***
## YR                0.05825    0.01685   3.458 0.000544 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 187.14  on 146  degrees of freedom
## Residual deviance: 158.11  on 144  degrees of freedom
## AIC: 164.11
##
## Number of Fisher Scoring iterations: 4
```

```
## Likelihood ratio test
##
## Model 1: cancer ~ factor(BK) * AG * YR
## Model 2: cancer ~ factor(BK) + YR + factor(BK):AG + factor(BK):YR + AG:YR +
##     factor(BK):AG:YR
## Model 3: cancer ~ factor(BK) + YR + factor(BK):YR + YR:AG + factor(BK):YR:AG
```

```
## Model 4: cancer ~ factor(BK) + YR + YR:AG + factor(BK):YR:AG
## Model 5: cancer ~ factor(BK) + YR + YR:AG
## Model 6: cancer ~ factor(BK) + YR
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   8 -77.104
## 2   8 -77.104  0 0.0000    1.00000
## 3   6 -77.349 -2 0.4888    0.78316
## 4   5 -77.423 -1 0.1496    0.69894
## 5   4 -77.495 -1 0.1429    0.70537
## 6   3 -79.057 -1 3.1245    0.07712 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##      df      AIC
## lmm1  8 170.2085
## lmm2  8 170.2085
## lmm3  6 166.6974
## lmm4  5 164.8470
## lmm5  4 162.9899
## lmm6  3 164.1144
```

We can justify this model by referring to the LRT test and AIC number. Specifically, the AIC numbers for model 4,5, and 6 are very close with a gap less than 2. It is acceptable and, thus, the simplest model (model 6) should be our choice.

Results

The regression model is a logistic regression with cancer as the response variable and bird-keeping status (BK) and years smoked (YR) as predictors. The intercept (coefficient for the reference level of the factor variable BK , which is "Bird" in this case, and the numeric variable YR, which equals to 0 in this case) is -1.70460 in log-odds scale and 0.18 in odd scale, which is statistically significant ($p < 0.01$).

The hypothesis that bird keepers have higher rates of lung cancer than non-bird keepers is supported by the regression result. Specifically, changing from the bird keeper group to the non-bird keeper group will cause a 1.48 decrease in log-odds scale (z = -3.727, $p < 0.001$). It implies that the probability of getting lung cancer will decrease by 77%.

Besides, for a one-unit increase in smoking, the expected change in log-odds is 0.05825 ($z = 3.458$, $p < 0.001$), which corresponds to a change of 1.06 in the odds of survival. And the coefficient can be interpreted as: for each additional year smoked, the odds of getting lung cancer change by a factor of 1.06 (or increase by 6%), holding bird keeping status constant.

Assumptions check

To test the goodness of fit, `pchisq`, `hoslem`, and `PseudoR2` tests have been conducted. Values larger than 0.05 (a certain significant level) shown in the first two tests indicate that the null hypotheses (the model is good) cannot be rejected. The relatively large R2 value of the reduced model (0.29), with 0.33 for the original one, also reveals that the model is good.

```
## [1] 0.1990647
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  case2002$cancer, fitted(lmm6)
## X-squared = 7.5439, df = 8, p-value = 0.4792
```

```
## VeallZimmermann
##       0.2943823
```

```
## VeallZimmermann
##       0.3267523
```

```
## [1] -0.08686428
```

```
## factor(BK)         YR
##   1.030305   1.030305
```

The calculated mean error is very small (-0.087) and close to 0 as expected, indicating that the model has a good fit to the data. Furthermore, the variance inflation factors (VIF) are around 1.03 for BK and YR in this case, indicating that multicollinearity is not an issue. That is, the independent variables are not highly correlated with each other.

Also, the traditional diagnosis plots and DHARMa plots did not report potential conflicts with the model assumptions.

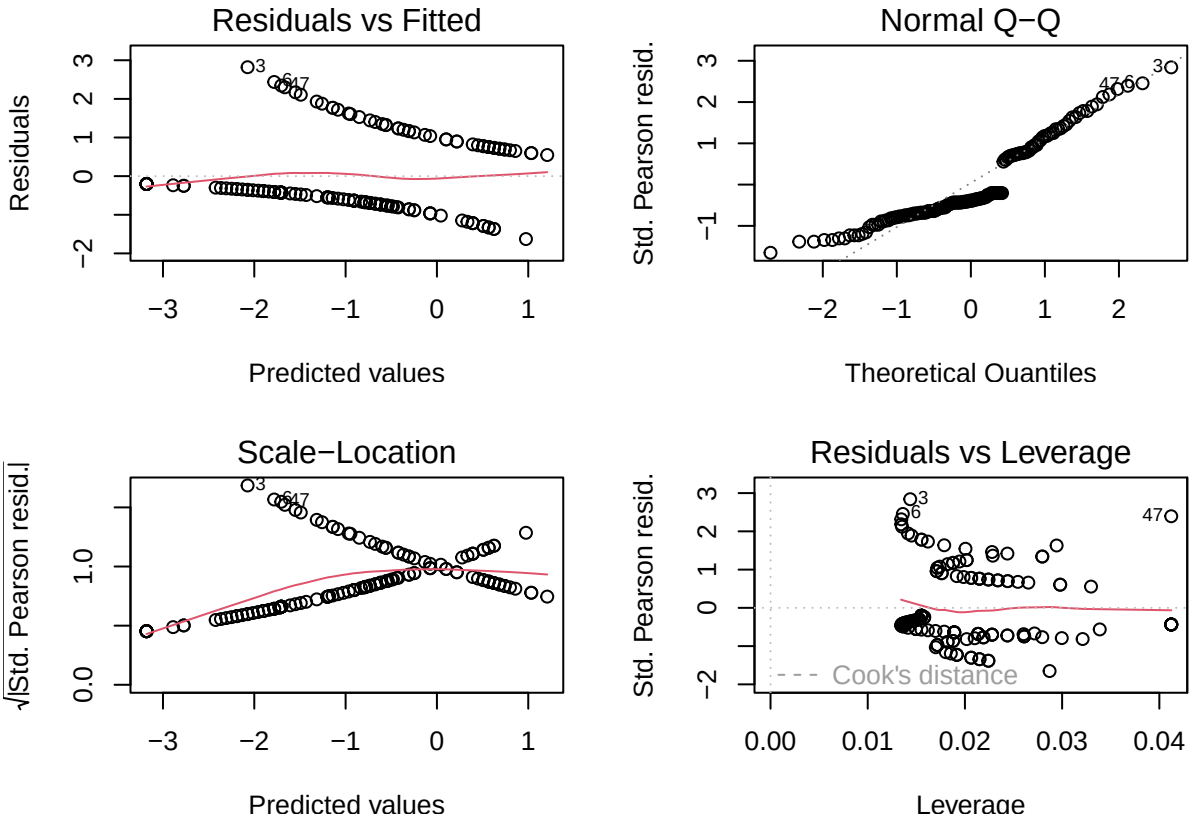

Figure 5: Assumption plots

Visualization

The figure displays the impacts of bird keeper and year of smoke on lung cancer rate.

Prediction

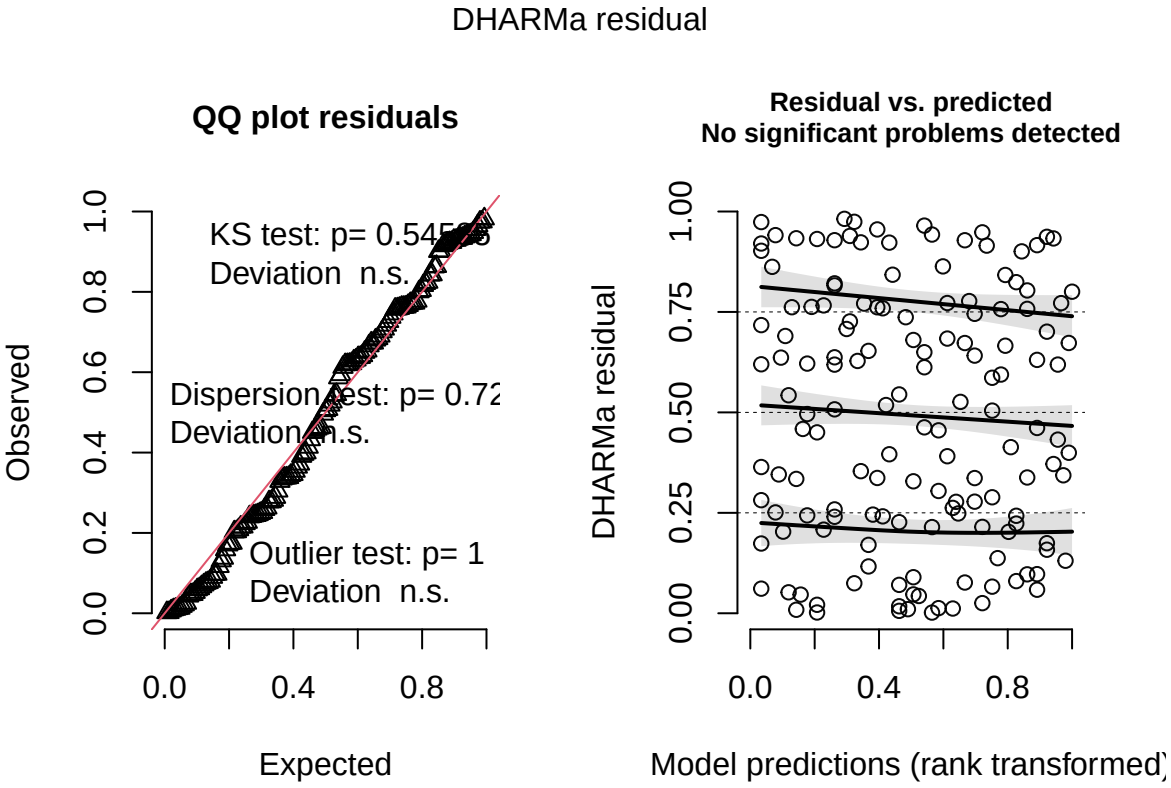

Figure 6: DHARMa plots

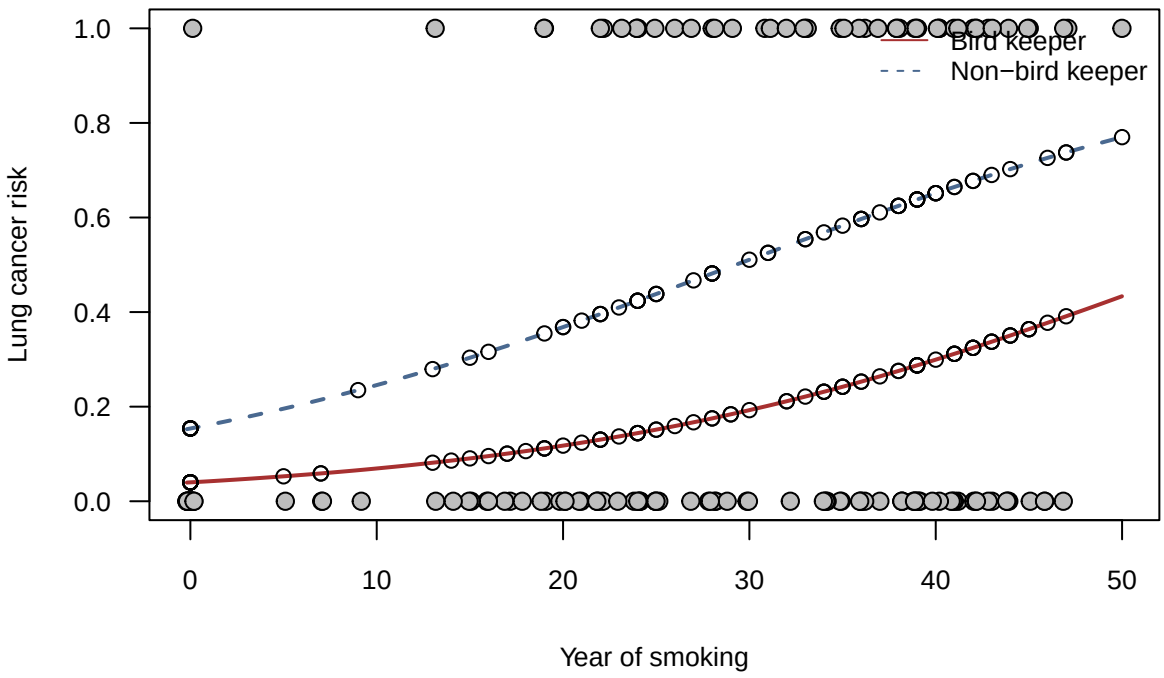

Figure 7: Impacts of bomb of bird keeper and year of smoke on lung cancer rate

```
inv.logit (coef(lmm6)[1] +coef(lmm6)[3]*32)
```

```
## (Intercept)
##   0.5397627
```

```
inv.logit(coef(lmm6)[1])
```

```
## (Intercept)
##   0.1538649
```

The probability of having lung cancer is 54.0% if you are a bird keeper and have smoked for 32 years. And the probability of the intercept is 15.4%, which means the individual is a bird keeper and have not smoked.

Problem 3

Hypothesis

We want to test whether well depth, distance from the mine, and site affect the proportion of contaminated wells at a site. H_0 : The proportion of contaminated wells is not affected by well depth, distance from the mine, and site. H_a : The proportion of contaminated wells is affected by at least one of the factors including well depth, distance from the mine, and/or site.

Method

Data preview

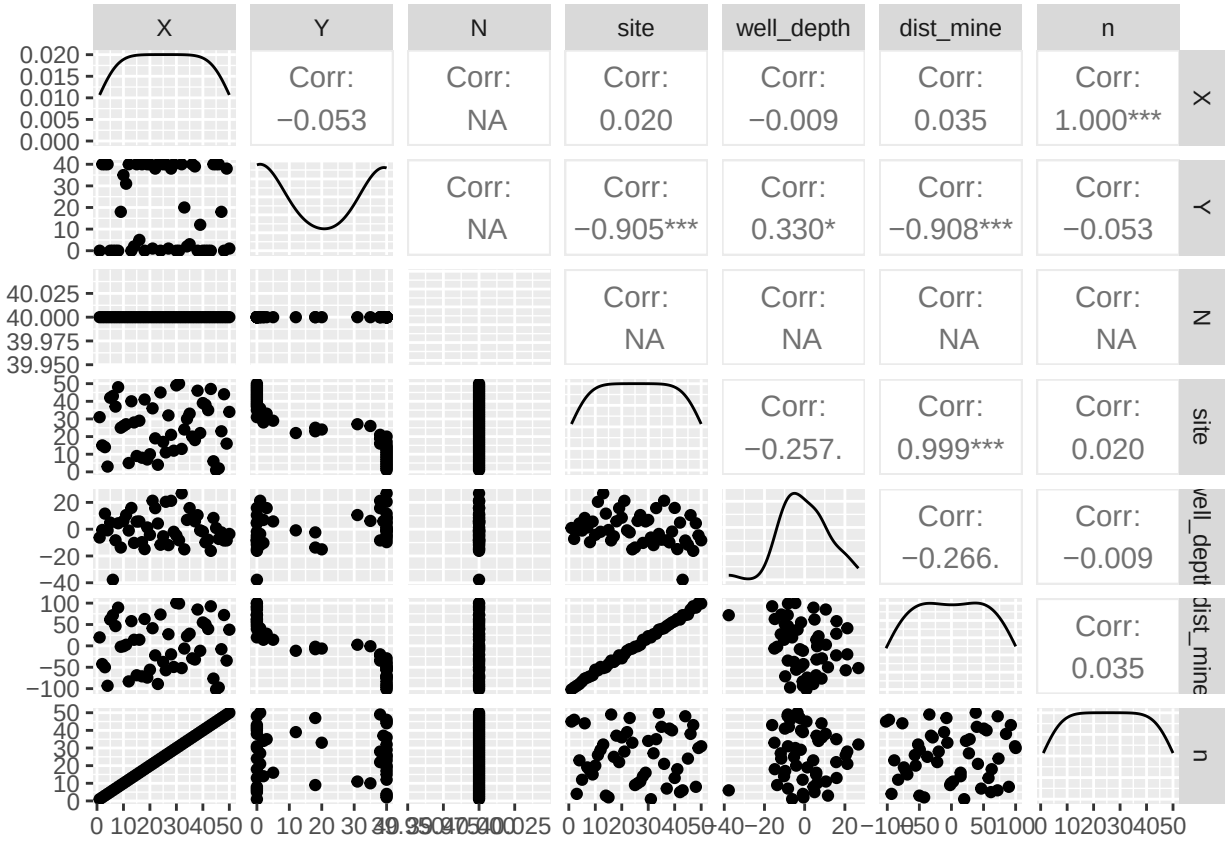

Figure 8: Preview of Dataset

Model

Because the site effect is considered in this model and the response can be summarized in two groups, a random effect logistic model should be established.

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: cbind(Y, N - Y) ~ well_depth + dist_mine + (1 | site)
##    Data: water
##
##      AIC      BIC   logLik deviance df.resid
##     99.8    107.4    -45.9     91.8       46
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -0.93994 -0.15155  0.00125  0.17362  1.23979
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  site   (Intercept) 0.705    0.8396
```

```
## Number of obs: 50, groups:  site, 50
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.22362    0.24606  -0.909  0.36346
## well_depth   0.06869    0.02245   3.059  0.00222 **
## dist_mine   -0.12855    0.01146 -11.215  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##            (Intr) wll_dp
## well_depth -0.122
## dist_mine   0.016 -0.114
```

There is no need to reduce the model since all the predictors are significant.

Results

The model has a log-likelihood of -45.9, an AIC of 99.8, and a BIC of 107.4. The AIC and BIC values indicate that the model is a good fit for the data. The scaled residuals have a mean of zero and a standard deviation of one, which indicates that the model assumptions are met. The response variable is a binomial variable that represents the number of contaminated wells out of the total number of tested wells at each site. The model includes three fixed effects: well_depth, dist_mine, and an intercept, and a random intercept for each site (grouped by the n variable).

The coefficient for well_depth is 0.06869 ($z = 3.059$, $p < 0.05$). This indicates that, holding the distance to mine constant, a one-unit increase in well depth is associated with a 0.06869 increase in the log-odds of a well being contaminated (increase by a factor of 1.07 in the odd scale). The coefficient for dist_mine is -0.12855 ($z = -11.215$, $p < 0.05$). This indicates that, holding well depth constant, a one-unit increase in the distance to mine is associated with a -0.12855 decrease in the log-odds scale, which is a 12% decrease in the odd scale. The correlation between well_depth and dist_mine is -0.122, which indicates that there is a weak negative correlation between the two predictor variables.

In conclusion, the model suggests that well_depth and dist_mine have significant effects on the proportion of contaminated wells at a site. Specifically, increasing well depth is associated with an increase in the probability of a well being contaminated, while increasing the distance to mine is associated with a decrease in the probability of a well being contaminated.

Assumptions check

```
## [1] 0.02519482
```

```
## well_depth  dist_mine
##   1.013283   1.013283
```

```
##                    R2m       R2c
## theoretical 0.9874380 0.9986876
## delta       0.9871449 0.9983911
```

The calculated mean error is very small (0.025) and close to 0, indicating that the model has a good fit to the data. Furthermore, the variance inflation factors (VIF) are around 1 for both well_depth and dist_mine in this case, indicating that multicollinearity is not an issue.

Besides, the values for R2m (0.987) and R2c (0.999) show that the fixed and random effects of the model explain a substantial proportion of the variation in the response variable. The differences between the full and null models are quite large, indicating that the model provides a good fit to the data.

We can also justify the goodness of the model from the DHARMa residuals plots. No conflicts with the assumptions are reported by the plots.

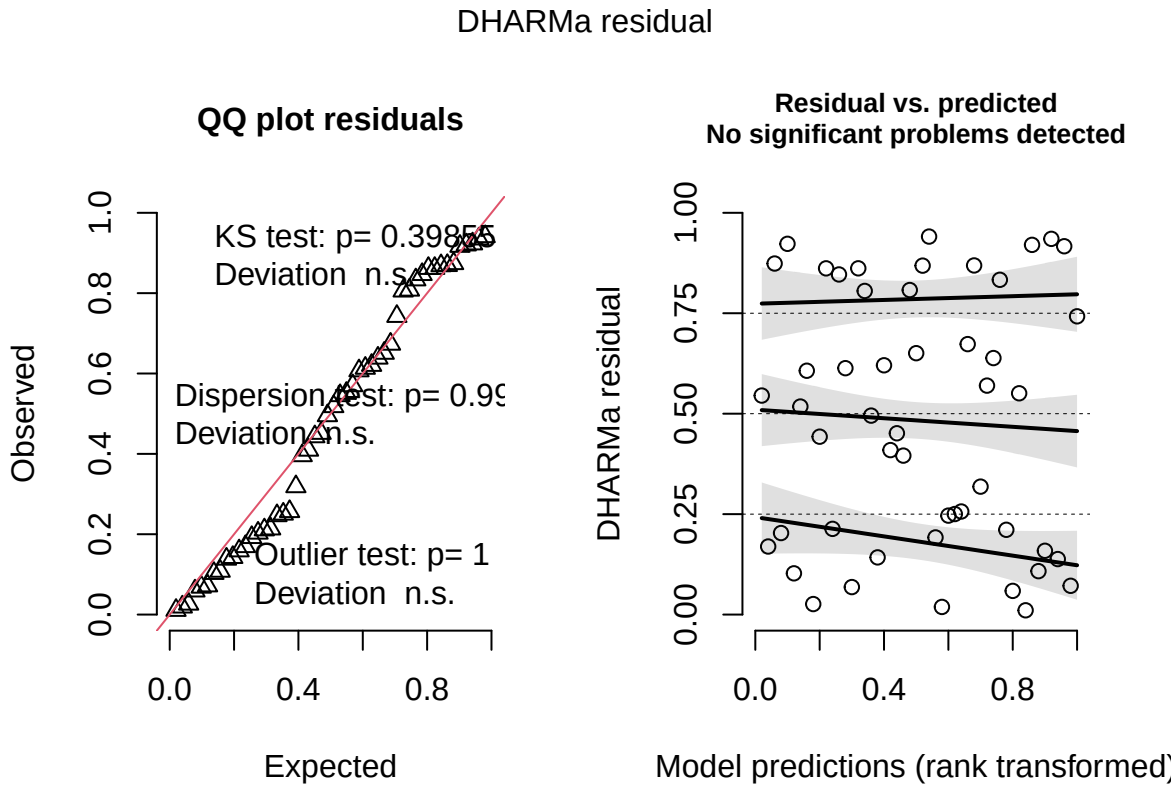

Figure 9: DHARMa plots

```
##
##  DHARMa zero-inflation test via comparison to expected zeros with
##  simulation under H0 = fitted model
##
## data:  simulationOutput
## ratioObsSim = 0.92721, p-value = 0.528
## alternative hypothesis: two.sided
```

The "ratioObsSim" value in the report indicates the ratio of observed zeros to expected zeros based on simulations under the null hypothesis that the model is correctly specified. A value close to 1 suggests a good fit between the observed and expected distributions of zeros. In this case, the value is 0.927, which indicates a slight deviation from the expected distribution of zeros. The "p-value" in the report represents the probability of observing a ratio as extreme or more extreme than the observed ratio, assuming that the model is correctly specified. In this case, the p-value is 0.528, which is greater than the commonly used significance level of 0.05. Therefore, we do not reject the null hypothesis that the model is correctly specified.

Overall, the results of the DHARMa zero-inflation test suggest that the model adequately accounts for zero-inflation in the data, and that the observed and expected distributions of zeros are similar. Thus, it is not an issue in this case. Also, the result of nonparametric dispersion test indicates the overdispersion issue does not need to be concerned in this model.

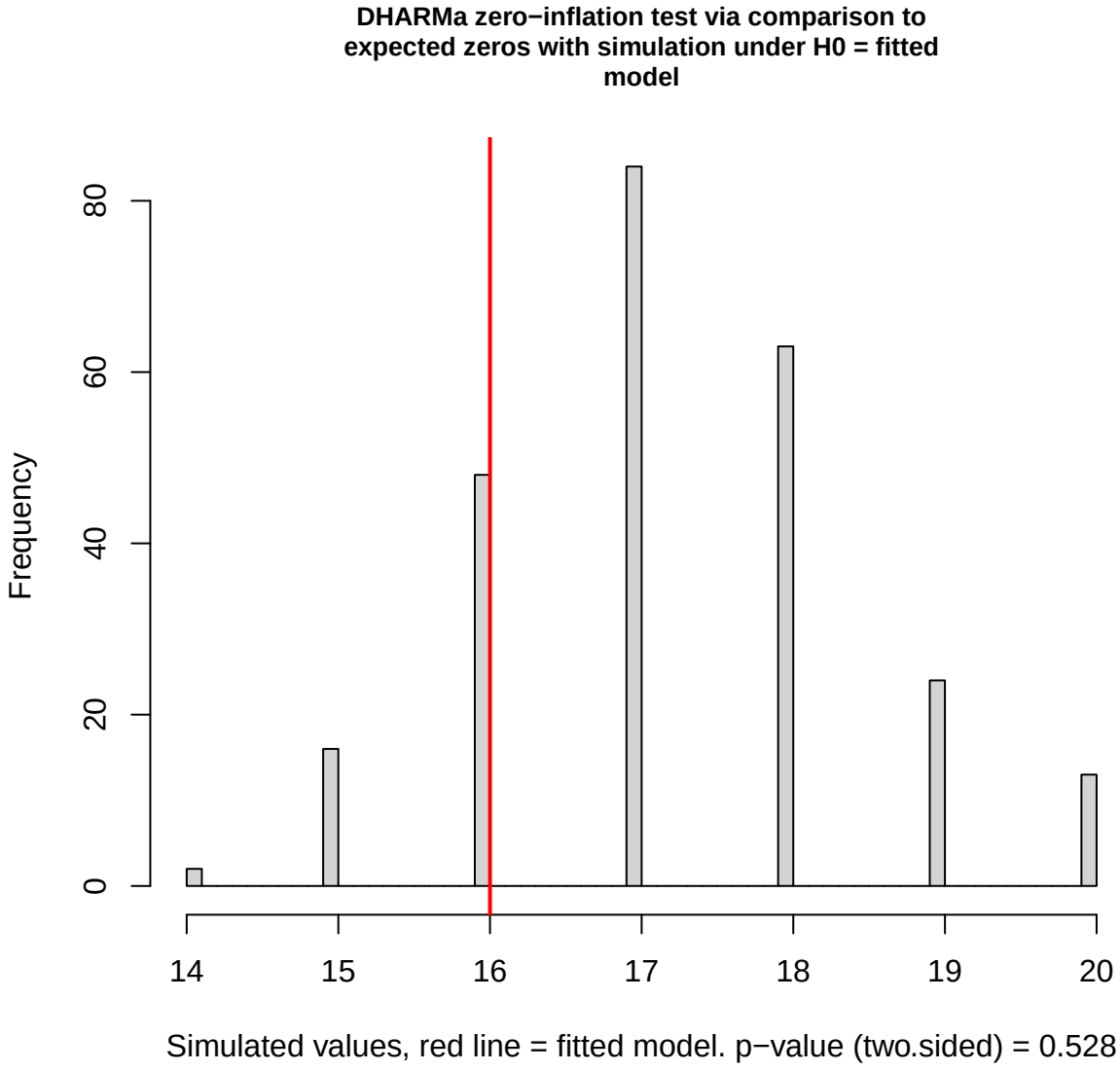

Figure 10: zero inflation test

Visualization

Since the proportion of random effect is small in this case, only the main fixed effects of the two predictors are visualized in the two graphs, holding the other variable constant as its mean value.

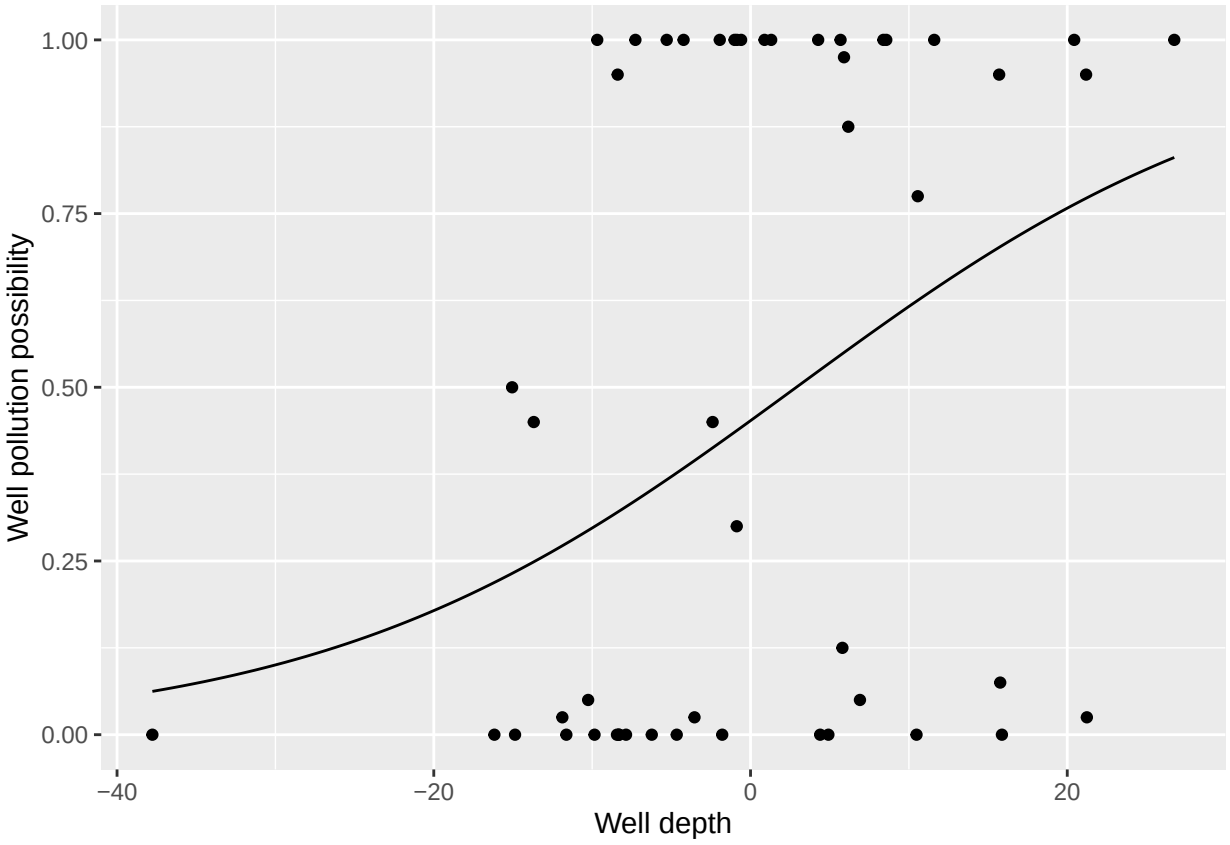

Figure 11: Impacts of well depth on well pollution

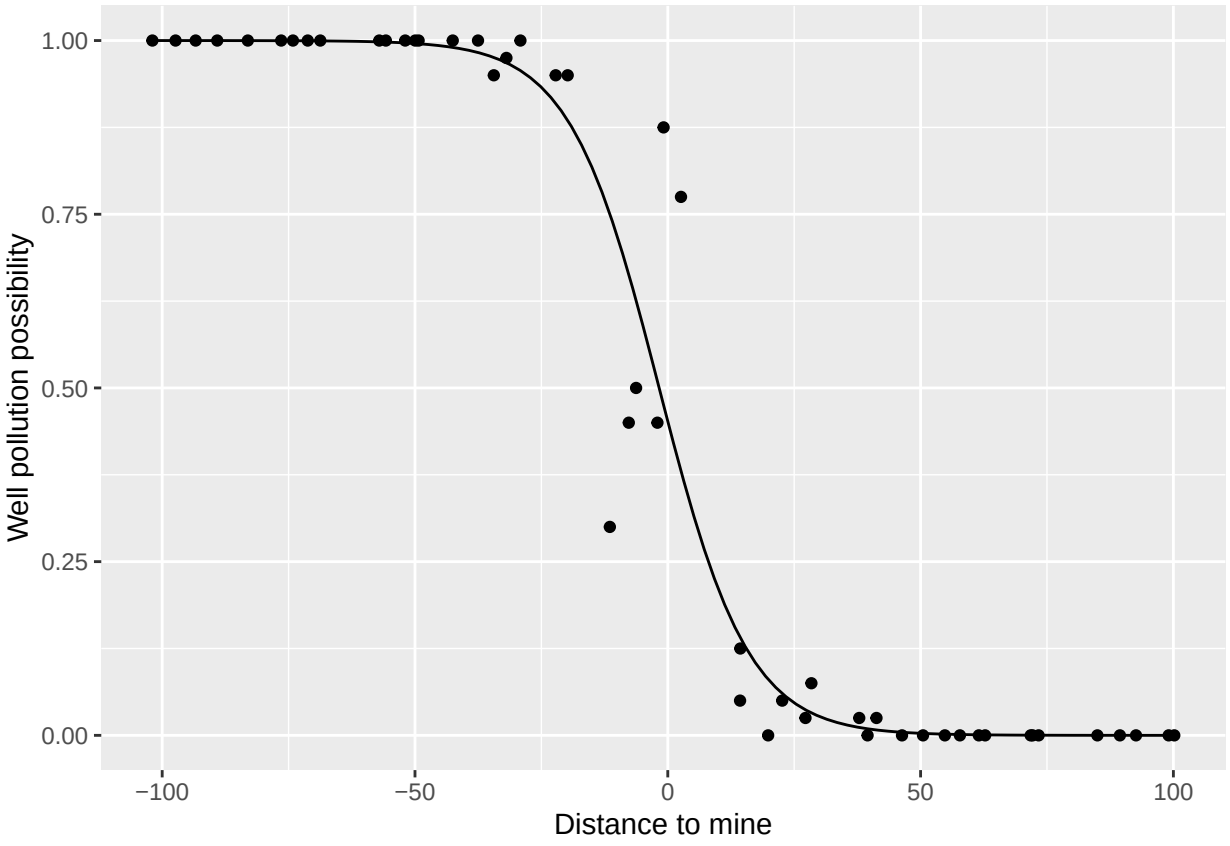

Figure 12: Impacts of distance to mine on well pollution

```r
knitr::opts_chunk$set(echo = FALSE, eval = TRUE, warning = F, message = FALSE)
pacman::p_load(ggplot2, GGally)
library(ggplot2)
library(GGally)
air <- read.csv("./labs/lab9 Generalized Linear Models/AircraftDat.csv")
# air <- read.csv("AircraftDat.csv")
air$x1 <- factor(air$x1)
ggpairs(air)
summary(air)
lm1 <- glm(y ~ as.factor(x1) + x2 + x3,
          data = air, family = poisson)
summary(lm1)
lm2.1<- update(lm1, ~.-as.factor(x1))

lm2.2 <- update(lm1, ~.-x2)

lm2.3 <- update(lm1, ~.-x3)

lm3.1 <- glm(y ~ as.factor(x1),
          data = air, family = poisson)

lm3.2 <- glm(y ~ x2,
          data = air, family = poisson)

lm3.3 <- glm(y ~ x3,
          data = air, family = poisson)

library(lmtest)

lrtest(lm1,lm2.2)
lrtest(lm1, lm2.1, lm3.2)
lrtest(lm1,lm2.3,lm3.1)
AIC(lm1,lm2.1,lm2.2,lm2.3,lm3.1,lm3.2,lm3.3)
summary(lm3.2)
mean(residuals(lm3.2))
par(mfrow=c(2,2), mar = c(3.8, 4, 3, 2))
plot(lm3.2)
library(AER)
dispersiontest(lm3.2)
pchisq(lm3.2$deviance, lm3.2$df.residual, lower.tail=F)
library(DescTools)
DescTools::PseudoR2(lm3.2, c("VeallZimmermann"))
DescTools::PseudoR2(lm1, c("VeallZimmermann"))

eq1 <- function(x)
{ exp(coef(lm3.2)[1] + coef(lm3.2)[2] * x)}

ggplot(air) +
geom_point(aes(x = x2, y = y)) +
stat_function(aes(x = x2), fun = eq1, geom = "line") +
labs(x = "Bomb load (ton)", y = "Damage of attack aircraft")
library(Sleuth3)
data(case2002)
```

```r
case2002$cancer <- ifelse(case2002$LC == "NoCancer", 0, 1)
ggpairs(case2002)
lmm1 <- glm(cancer ~ factor(BK)*AG*YR, family=binomial, data=case2002)
summary(lmm1)
lmm2 <- update(lmm1, ~.-AG)
summary(lmm2)
lmm3 <- update(lmm2, ~.-factor(BK):AG)
summary(lmm3)
lmm4 <- update(lmm3, ~.-factor(BK):YR)
summary(lmm4)
lmm5 <- update(lmm4, ~.-factor(BK):YR:AG)
summary(lmm5)
lmm6 <- update(lmm5, ~.-YR:AG)
summary(lmm6)
lrtest(lmm1, lmm2, lmm3, lmm4, lmm5, lmm6)
AIC(lmm1, lmm2, lmm3, lmm4, lmm5, lmm6)
library(ResourceSelection)
pchisq(lmm6$deviance, lmm6$df.residual, lower=F)
hoslem.test(case2002$cancer, fitted(lmm6))

DescTools::PseudoR2(lmm6, c("VeallZimmermann"))
DescTools::PseudoR2(lmm1, c("VeallZimmermann"))
mean(residuals(lmm6))
vif(lmm6)
par(mfrow=c(2,2), mar = c(3.8, 4, 3, 2))
plot(lmm6)
library(DHARMa)
lm6 <- simulateResiduals(fittedModel = lmm6)
plot(lm6)
library(boot)

jcoPalette <- c("#0073C2FF", "#EFC000FF", "#868686FF", "#CD534CFF",
"#7AA6DCFF", "#003C67FF", "#8F7700FF", "#3B3B3BFF", "#A73030FF",
"#4A6990FF")
# plot the data points
plot(jitter(case2002$YR), case2002$cancer, las=1, pch=21, cex=1.2, bg="grey",
     xlab = "Year of smoking", ylab = "Lung cancer risk", cex.axis = 0.8,
     cex.lab = 0.8)

# add curves for bird and no-bird
x <- seq(min(case2002$YR), max(case2002$YR), length = 50)
curve(expr = inv.logit(lmm6$coef[1] + lmm6$coef[2] + lmm6$coef[3]*x), add=T,
      lwd=2, col= jcoPalette[9], lty = 1)
curve(expr = inv.logit(lmm6$coef[1] + lmm6$coef[3]*x), add=T,
      lwd=2, col= jcoPalette[10], lty = 2)
points(case2002$YR, fitted(lmm6))

# add a legend
legend("topright", legend=c("Bird keeper", "Non-bird keeper"), lty=c(1, 2), col=c(jcoPalette[9], jcoPal
inv.logit (coef(lmm6)[1] +coef(lmm6)[3]*32)
inv.logit(coef(lmm6)[1])
water <- read.csv("./labs/lab9 Generalized Linear Models/water.csv")
ggpairs(water)
```

```
library(lme4)
llm <- glmer(cbind(Y, N-Y) ~ well_depth + dist_mine + (1|site), data = water, family = binomial)
summary(llm)
mean(residuals(llm))
vif(llm)

library(MuMIn)
MuMIn::r.squaredGLMM(llm)
library(DHARMa)
llm1 <- simulateResiduals(fittedModel = llm)
plot(llm1)
testZeroInflation(llm1)
llm2 <- glm(cbind(Y, N-Y) ~ well_depth + dist_mine, data = water, family = binomial)

eq3.dep <- function(x)
{ inv.logit(coef(llm2)[1] + coef(llm2)[2] * x + coef(llm2)[3] * mean(water$dist_mine))}

ggplot(water) +
geom_point(aes(x = well_depth, y = Y/N)) +
stat_function(aes(x = well_depth), fun = eq3.dep, geom = "line") +
labs(x = "Well depth", y = "Well pollution possibility")
eq3.dis <- function(x)
{ inv.logit(coef(llm2)[1] + coef(llm2)[2] * mean(water$well_depth) + coef(llm2)[3] * x)}

ggplot(water) +
geom_point(aes(x = dist_mine, y = Y/N)) +
stat_function(aes(x = dist_mine), fun = eq3.dis, geom = "line") +
labs(x = "Distance to mine", y = "Well pollution possibility")
```