**STAT34700 HW1**

**Problem 1**

**(a)**

**Model: Logistic Regression Model**

**Output:**

**The summary of the glm:**

> summary(lmod)

Call:

glm(formula = spikes ~ xN + yN, family = binomial, data = hipp)

Deviance Residuals:

Min 1Q Median 3Q Max

-0.5637 -0.3806 -0.2653 -0.1925 2.9658

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.20040 0.02814 -113.73 <2e-16 \*\*\*

xN -0.53624 0.04872 -11.01 <2e-16 \*\*\*

yN -1.24991 0.04635 -26.97 <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: 16306 on 41327 degrees of freedom

Residual deviance: 15332 on 41325 degrees of freedom

AIC: 15338

Number of Fisher Scoring iterations: 6

> sumary(lmod)

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.200397 0.028140 -113.732 < 2.2e-16

xN -0.536244 0.048724 -11.006 < 2.2e-16

yN -1.249913 0.046348 -26.968 < 2.2e-16

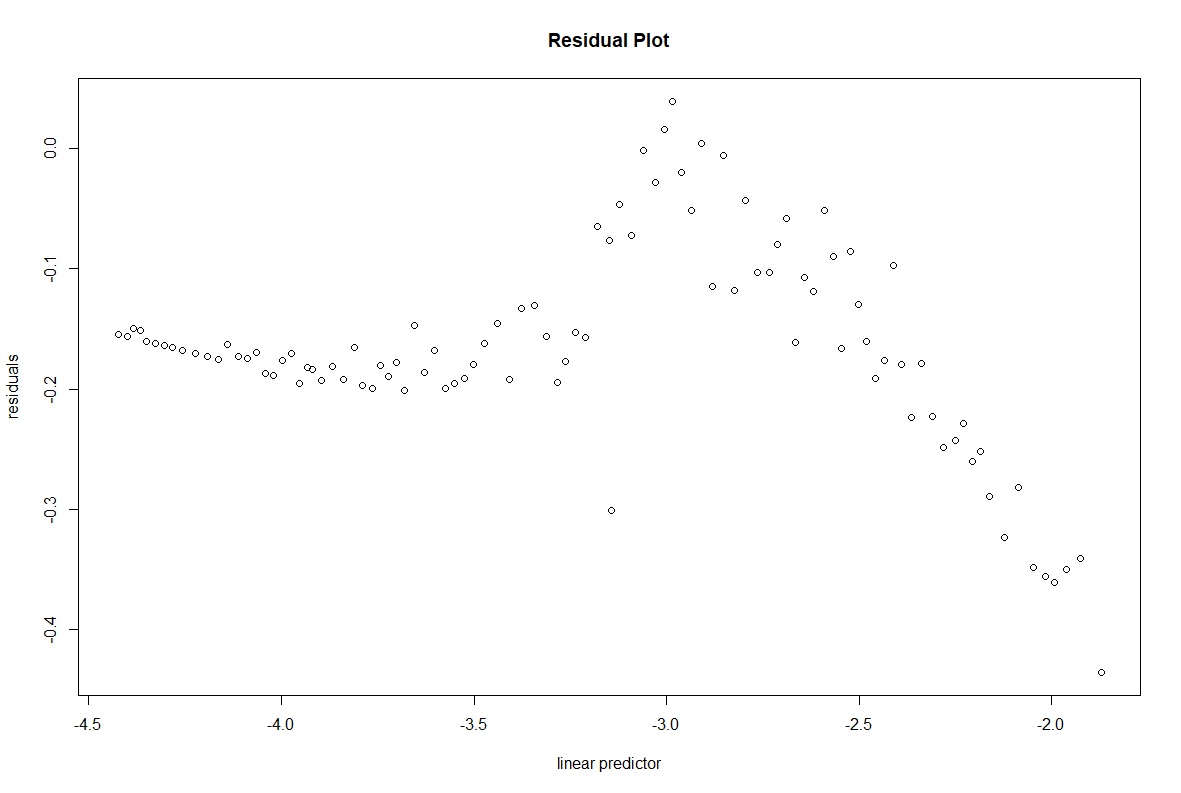
n = 41328 p = 3

Deviance = 15331.56093 Null Deviance = 16305.85484 (Difference = 974.29391)

**Fitted Model:**

**Check the model assumptions:**

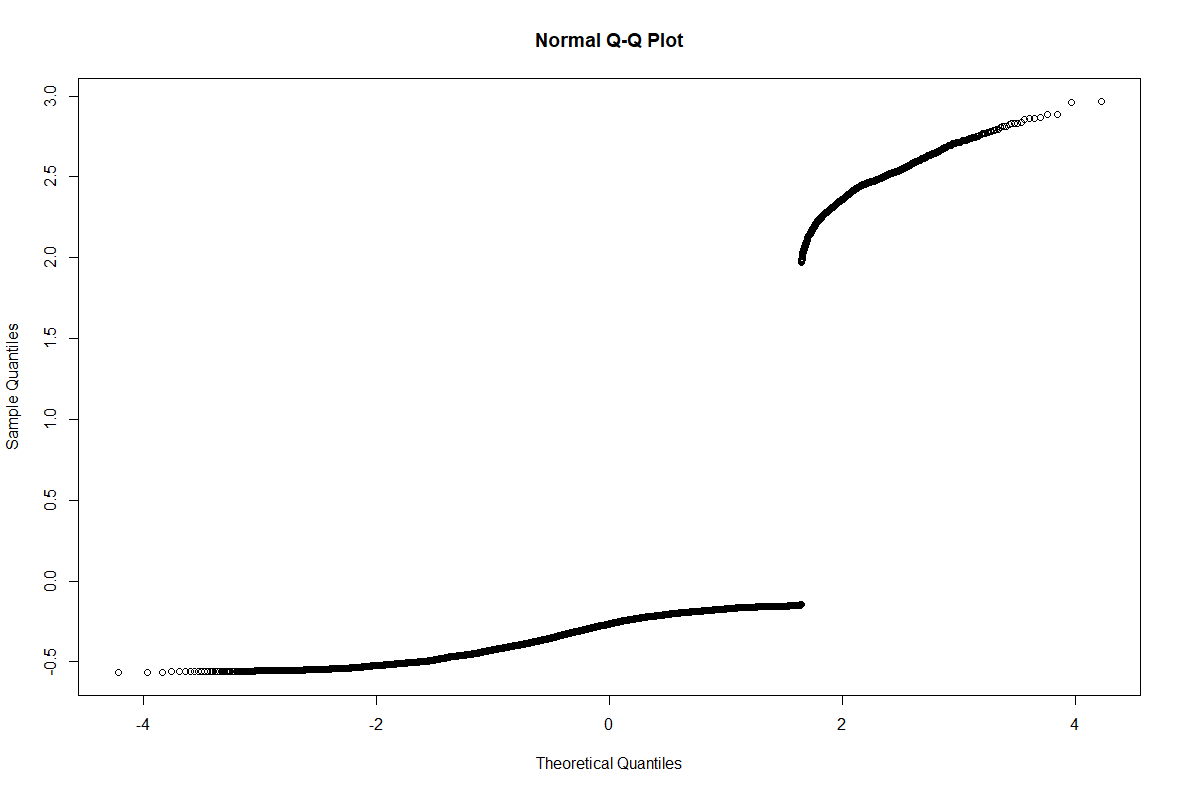
**1) Group the residuals into bins and plot residuals v.s. fitted linear predictors**



The residual plot shows that the variance of residuals is not constant so the constant variance

assumption is not satisfied in this model.

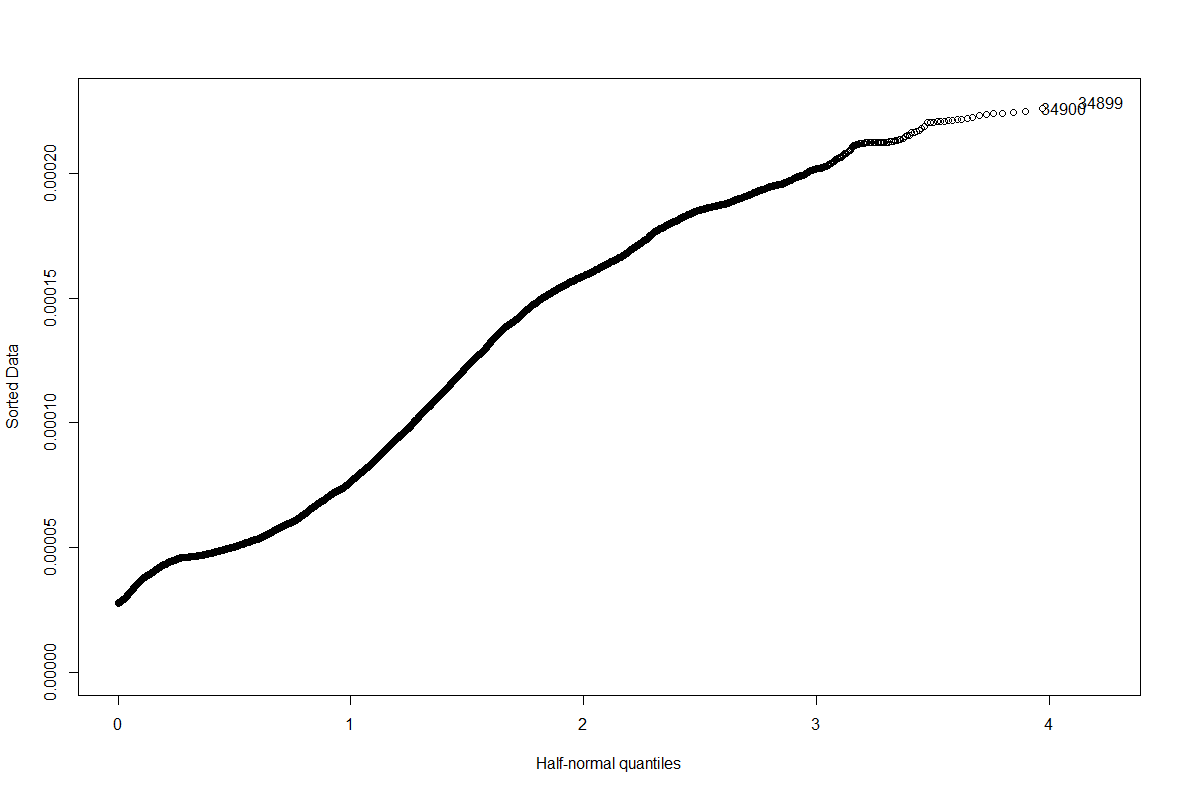
**2) QQ Plot**



We can see that the QQ plot is very far from the desired linear relationship. We see two

clusters of points corresponding to y = 0 and y = 1. But there is no reason to expect these residuals to be normally distributed so this does not raise any concern. The normality assumption is satisfied in this model.

**3) Half-normal Plot**



From the half-normal plot, we can see that leverage of the two points showed are not so extreme, compared to other points in the plot. Hence, we can say that there is no outlier in the model.

**(b)**

**Summary:** we can improve the fit by adding in the regression model to describe the distance between location coordinates and the origin.

**1) Improve the fit with interactions of xN and yN:**

**Model: Logistic Regression Model**

**Output:**

> summary(lmodinter)

Call:

glm(formula = spikes ~ xN + yN + xN:yN, family = binomial, data = hipp)

Deviance Residuals:

Min 1Q Median 3Q Max

-0.5573 -0.3821 -0.2676 -0.1907 2.9747

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.20264 0.02831 -113.113 <2e-16 \*\*\*

xN -0.57038 0.05637 -10.118 <2e-16 \*\*\*

yN -1.26036 0.04722 -26.690 <2e-16 \*\*\*

xN:yN -0.14722 0.12084 -1.218 0.223

---

Signif. codes: 0 ?\*\*?0.001 ?\*?0.01 ??0.05 ??0.1 ??1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 16306 on 41327 degrees of freedom

Residual deviance: 15330 on 41324 degrees of freedom

AIC: 15338

Number of Fisher Scoring iterations: 6

> anova(lmod, lmodinter, test = "Chi")

Analysis of Deviance Table

Model 1: spikes ~ xN + yN

Model 2: spikes ~ xN + yN + xN:yN

Resid. Df Resid. Dev Df Deviance Pr(>Chi)

1 41325 15332

2 41324 15330 1 1.4876 0.2226

**Comment:**

After adding the interactions of xN and yN, the AIC value of this model is still 15388. In

addition, from the Chi-square test of these two models, the p-value is 0.2226 which is greater

than significance level 0.05. Hence, the null hypothesis: model without interactions is better, cannot be rejected. The model with interactions cannot improve the result of fitting the data.

**2) Improve the fit with non-linear functions of xN and yN:**

**Model: Logistic Regression Model**

**Output:**

> summary(lmodimprove)

Call:

glm(formula = spikes ~ xN + yN + distance, family = binomial,

data = hipp)

Deviance Residuals:

Min 1Q Median 3Q Max

-0.9484 -0.3593 -0.1763 -0.0546 4.0611

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.53374 0.05560 -9.599 <2e-16 \*\*\*

xN -1.18748 0.07698 -15.425 <2e-16 \*\*\*

yN -3.00255 0.09926 -30.249 <2e-16 \*\*\*

distance -5.30315 0.13114 -40.438 <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: 16306 on 41327 degrees of freedom

Residual deviance: 13128 on 41324 degrees of freedom

AIC: 13136

Number of Fisher Scoring iterations: 8

> anova(lmod,lmodimprove,test="Chi")

Analysis of Deviance Table

Model 1: spikes ~ xN + yN

Model 2: spikes ~ xN + yN + distance

Resid. Df Resid. Dev Df Deviance Pr(>Chi)

1 41325 15332

2 41324 13128 1 2203.7 < 2.2e-16 \*\*\*

---

Signif. codes: 0 ?\*\*?0.001 ?\*?0.01 ??0.05 ??0.1 ??1

**Comments:**

To express the location more preciously by x and y coordinates, we can add a non-linear term of xN and yN: the distance between location and origin: After adding the non-linear function of xN and yN, the AIC value of this model becomes smaller.

In addition, from the Chi-square test of these two models, the p-value is smaller than significancelevel 0.05. Hence, the null hypothesis: model without non-linear function is better, can be

rejected. The model with non-linear function can improve the result of fitting the data.

**Final Model:**

**(c)**

**Summary:** If we sample the data every 20 point, the fitted coefficients will be slightly different

from the fitted coefficients for the original data.

**Output:**

> sumary(lmodsub)

Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.28177 0.24769 -1.1376 0.2553026

xN -1.36860 0.37681 -3.6321 0.0002811

yN -3.25326 0.49144 -6.6199 3.595e-11

distance -6.03964 0.63495 -9.5120 < 2.2e-16

n = 2067 p = 4

Deviance = 623.16103 Null Deviance = 800.81878 (Difference = 177.65775)

**Fitted Model:**

**Comments:**

Original data:

Sampled data:

There is slight difference between the fitted coefficients from two data sets.

**(d)**

**Summary:** If we sample the data using the method in part (d), the fitted coefficients for xN, yN

and distance will be slightly different from the fitted coefficients for the original data. The fitted coefficient of intercept changes a lot, compared to the original data.

The sampling method in part (d) is called retrospective sampling (case-control study) because theresponse is fixed and then the predictors are observed. The original sampling method is called

prospective sampling (cohort study) because the predictors are fixed and then the response is observed.

**Output:**

> summary(lmodtotal)

Call:

glm(formula = spikes ~ xN + yN + distance, family = binomial,

data = hipptotal)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.3443 -0.6706 0.1626 0.7302 3.3099

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 2.8433 0.5453 5.215 1.84e-07 \*\*\*

xN -1.5383 0.4963 -3.100 0.00194 \*\*

yN -3.1151 0.6128 -5.083 3.71e-07 \*\*\*

distance -6.1494 1.0282 -5.981 2.22e-09 \*\*\*

---

Signif. codes: 0 ?\*\*?0.001 ?\*?0.01 ??0.05 ??0.1 ??1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 277.26 on 199 degrees of freedom

Residual deviance: 178.17 on 196 degrees of freedom

AIC: 186.17

Number of Fisher Scoring iterations: 6

**Fitted Model:**

**Comments:**

Original data:

Sampled data:

There is slight difference between the fitted coefficients from two data set but the fitted coefficients of intercept are different a lot in these two data sets.

**(e)**

**Summary:** The history improves the prediction of the spikes relative to the location covariates. The history of spike’s activity is more recent, the absolute value of the coefficient is larger. As a result, most recent history contributes more to predict the response value.

**Output:**

> summary(lmodhis)

Call:

glm(formula = spikes ~ xN + yN + distance + spikes.hist.1 + spikes.hist.2 +

spikes.hist.3 + spikes.hist.4 + spikes.hist.5 + spikes.hist.6 +

spikes.hist.7 + spikes.hist.8 + spikes.hist.9 + spikes.hist.10 +

spikes.hist.11 + spikes.hist.12 + spikes.hist.13 + spikes.hist.14 +

spikes.hist.15 + spikes.hist.16 + spikes.hist.17 + spikes.hist.18 +

spikes.hist.19 + spikes.hist.20, family = binomial, data = hipp)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.7534 -0.3024 -0.1424 -0.0483 4.1250

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.460099 0.109071 -4.218 2.46e-05 \*\*\*

xN -1.224943 0.083156 -14.731 < 2e-16 \*\*\*

yN -3.070232 0.121159 -25.340 < 2e-16 \*\*\*

distance -5.597465 0.189732 -29.502 < 2e-16 \*\*\*

spikes.hist.1 -3.186323 0.187443 -16.999 < 2e-16 \*\*\*

spikes.hist.2 -2.038970 0.122471 -16.649 < 2e-16 \*\*\*

spikes.hist.3 -1.035182 0.096704 -10.705 < 2e-16 \*\*\*

spikes.hist.4 0.215254 0.079003 2.725 0.00644 \*\*

spikes.hist.5 1.002009 0.074394 13.469 < 2e-16 \*\*\*

spikes.hist.6 1.030072 0.078630 13.100 < 2e-16 \*\*\*

spikes.hist.7 0.765968 0.084295 9.087 < 2e-16 \*\*\*

spikes.hist.8 0.360102 0.087862 4.098 4.16e-05 \*\*\*

spikes.hist.9 0.092836 0.087357 1.063 0.28791

spikes.hist.10 0.056860 0.084617 0.672 0.50161

spikes.hist.11 0.008576 0.084865 0.101 0.91950

spikes.hist.12 0.056076 0.084678 0.662 0.50783

spikes.hist.13 0.166682 0.084327 1.977 0.04809 \*

spikes.hist.14 0.061277 0.087029 0.704 0.48138

spikes.hist.15 0.164836 0.084321 1.955 0.05060 .

spikes.hist.16 0.012462 0.085155 0.146 0.88365

spikes.hist.17 0.043672 0.083688 0.522 0.60178

spikes.hist.18 -0.048080 0.085139 -0.565 0.57226

spikes.hist.19 -0.034110 0.083827 -0.407 0.68407

spikes.hist.20 -0.034542 0.082368 -0.419 0.67495

---

Signif. codes: 0 ?\*\*?0.001 ?\*?0.01 ??0.05 ??0.1 ??1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 16306 on 41327 degrees of freedom

Residual deviance: 11862 on 41304 degrees of freedom

AIC: 11910

Number of Fisher Scoring iterations: 8

> anova(lmod,lmodhis,test="Chi")

Analysis of Deviance Table

Model 1: spikes ~ xN + yN

Model 2: spikes ~ xN + yN + distance + spikes.hist.1 + spikes.hist.2 +

spikes.hist.3 + spikes.hist.4 + spikes.hist.5 + spikes.hist.6 +

spikes.hist.7 + spikes.hist.8 + spikes.hist.9 + spikes.hist.10 +

spikes.hist.11 + spikes.hist.12 + spikes.hist.13 + spikes.hist.14 +

spikes.hist.15 + spikes.hist.16 + spikes.hist.17 + spikes.hist.18 +

spikes.hist.19 + spikes.hist.20

Resid. Df Resid. Dev Df Deviance Pr(>Chi)

1 41325 15332

2 41304 11862 21 3469.2 < 2.2e-16 \*\*\*

---

Signif. codes: 0 ?\*\*?0.001 ?\*?0.01 ??0.05 ??0.1 ??1

**Comments:**

Adding the 20ms history of spike activity as predictors in the regression model, we get a new

logistic regression model with 23 predictors. From the summary of new model, the history of

spike’s activity is more recent, the absolute value of the coefficient is larger. As a result, most

recent history contributes more to predict the response value.

Compared to the model in part (b), the AIC value is getting smaller. By performing the Chi-

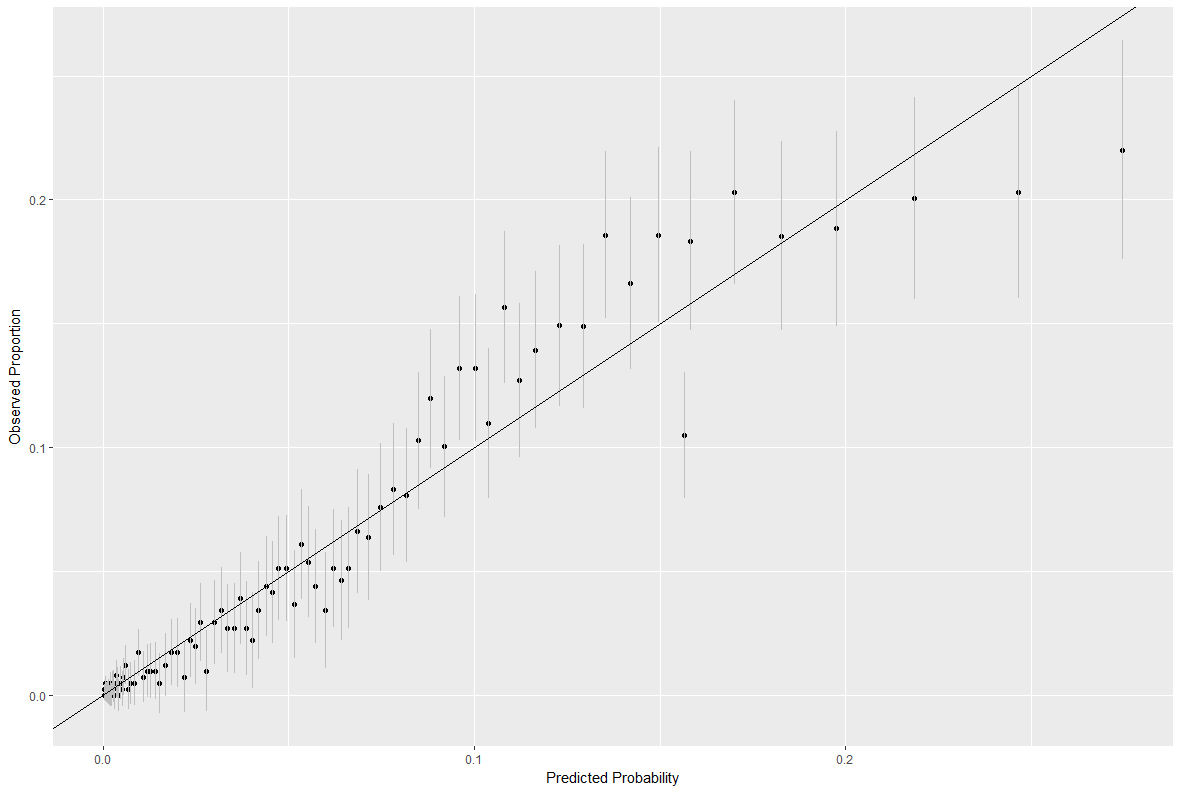
square test among these two models, the p-value is smaller than significance level 0.05. Hence,

the null hypothesis: model without history is better, can be rejected. The model with history

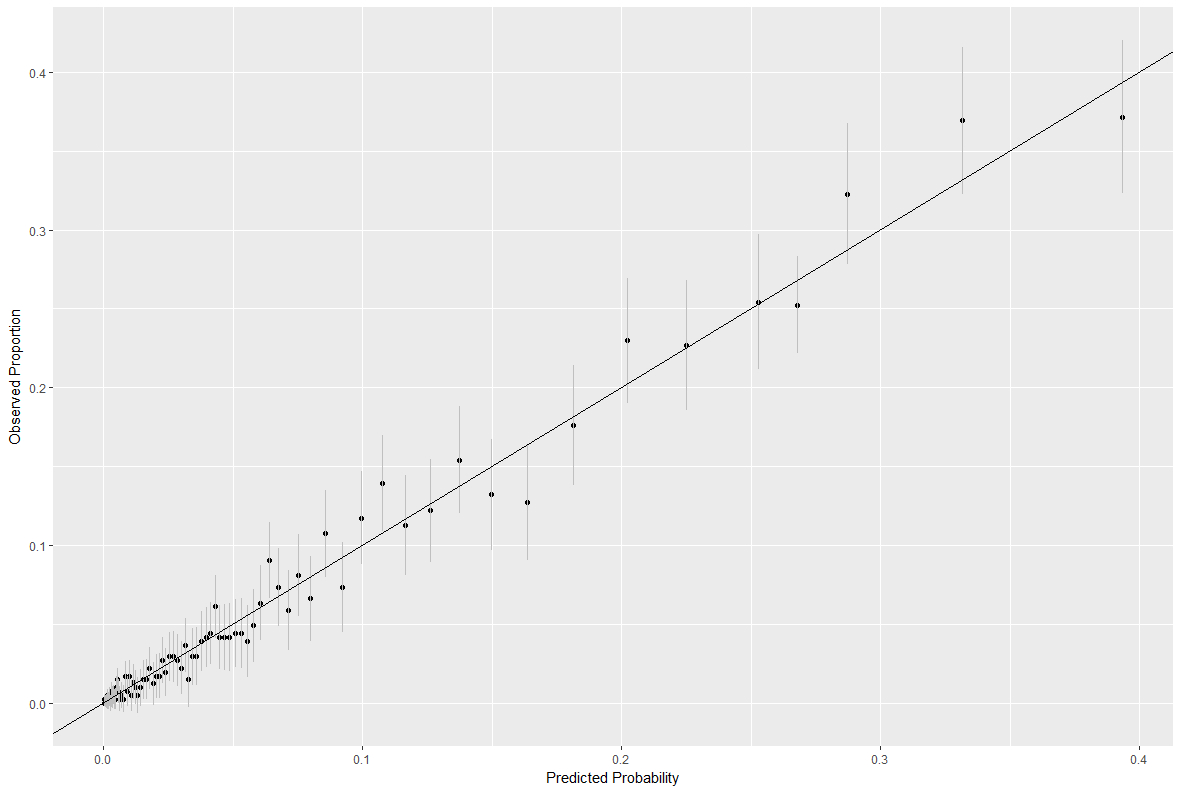
information can improve the result of fitting the data.

We can also plot the observed proportions against the predicted probabilities to see whether the history information can improve the prediction of the model.

**The plot of model without history information:**



**The plot of model with history information:**



**Comments:**

We can see that using the model with history information, the predicted probabilities are close to the observed proportions. However, using the model without history information, the predicted

probabilities are far away from the observed proportions. As a result, we can say the history

improves the prediction of the spikes relative to the location covariates.

**R Code**

**# Problem 1 ---------------------------------------------------------------**

install.packages("faraway")

library(faraway)

install.packages("dplyr")

library(dplyr)

install.packages("ggplot2")

library(ggplot2)

**#a**

#fit the logistic model

load("D:/Study/UChicago/Second Quarter/STAT34700/HW1/HIPP.RData")

hipp=data.frame(HIPP)

lmod <- glm(spikes ~ xN + yN, family = binomial, hipp)

summary(lmod)

**#Diagnostics**

**#Residual Plot**

linpred <- predict(lmod)

predprob <- predict(lmod, type="response")

hippre <- mutate(hipp, residuals=residuals(lmod), linpred=predict(lmod))

gdf <- group\_by(hippre, cut(linpred, breaks=unique(quantile(linpred,(1:100)/101))))

diagdf <- summarise(gdf, residuals=mean(residuals), linpred=mean(linpred))

plot(residuals ~ linpred, diagdf, xlab="linear predictor",main="Residual Plot")

**#QQ plot**

qqnorm(residuals(lmod))

**#Half-normal plot**

halfnorm(hatvalues(lmod))

**#b**

**#interactions**

lmodinter = glm(spikes ~ xN + yN + xN:yN , family = binomial, data=hipp)

sumary(lmodinter)

anova(lmod, lmodinter, test = "Chi")

**#non-linear function**

hipp$distance=sqrt((hipp$xN)^2+(hipp$yN)^2)

#use distance describe the location in the cage

lmodimprove=glm(spikes ~ xN + yN + distance , family = binomial, data=hipp)

summary(lmodimprove)

anova(lmod,lmodimprove,test="Chi")

**#c**

hippsub = hipp[1:NROW(hipp) %% 20 == 1, ]

lmodsub <- glm(spikes ~ xN + yN + distance, family = binomial, data=hippsub)

sumary(lmodsub)

**#d**

hipp1=subset(hippsub, spikes == 1)

hipp0=subset(hippsub, spikes == 0)

hipp0rs=hipp0[sample(1:nrow(hipp0), 100, replace=FALSE),]

hipptotal=rbind(hipp1,hipp0rs)

lmodtotal=glm(spikes ~ xN + yN + distance, family = binomial, data=hipptotal)

summary(lmodtotal)

**#e**

#fit the model with history information

lmodhis=glm(spikes ~ xN + yN + distance + spikes.hist.1 + spikes.hist.2 + spikes.hist.3

+ spikes.hist.4 + spikes.hist.5 + spikes.hist.6 + spikes.hist.7 + spikes.hist.8

+ spikes.hist.9 + spikes.hist.10 + spikes.hist.11 + spikes.hist.12 + spikes.hist.13

+ spikes.hist.14 + spikes.hist.15 + spikes.hist.16 + spikes.hist.17 + spikes.hist.18

+ spikes.hist.19 + spikes.hist.20, family = binomial, data=hipp)

summary(lmodhis)

anova(lmod,lmodhis,test="Chi")

**#plot of observed proportion vs predicted probability**

**#prediction with history information**

linpredhis <- predict(lmodhis)

predprobhis <- predict(lmodhis, type="response")

dfsm <- na.omit(hipp)

dfsm <- mutate(dfsm, predprobhis=predict(lmodhis,type="response"))

gdf <- group\_by(dfsm, cut(linpredhis, breaks=unique(quantile(linpredhis,(1:100)/101))))

hldf <- summarise(gdf, spikes=sum(spikes), ppred=mean(predprobhis), count=n())

hldf <- mutate(hldf, se.fit=sqrt(ppred\*(1-ppred)/count))

ggplot(hldf,aes(x=ppred,y=spikes/count,ymin=spikes/count-2\*se.fit,ymax=

spikes/count+2\*se.fit))+

geom\_point() + geom\_linerange(color=grey(0.75)) + geom\_abline(intercept=0,slope=1) +

xlab("Predicted Probability")+ylab("Observed Proportion")

**#prediction without history information**

linpred <- predict(lmodimprove)

predprob <- predict(lmodimprove, type="response")

hippsm <- na.omit(hipp)

hippsm <- mutate(hippsm, predprob=predict(lmodimprove,type="response"))

ghipp <- group\_by(hippsm, cut(linpred, breaks=unique(quantile(linpred,(1:100)/101))))

hlhipp <- summarise(ghipp, spikes=sum(spikes), ppred=mean(predprob), count=n())

hlhipp <- mutate(hlhipp, se.fit=sqrt(ppred\*(1-ppred)/count))

ggplot(hlhipp,aes(x=ppred,y=spikes/count,ymin=spikes/count-2\*se.fit,ymax=

spikes/count+2\*se.fit))+

geom\_point() + geom\_linerange(color=grey(0.75)) + geom\_abline(intercept=0,slope=1) +

xlab("Predicted Probability")+ylab("Observed Proportion")